Program

Conference on Parallel Processing for Scientific Computing

February 12–15, 2020
Seattle, Washington, U.S.

Sponsored by the SIAM Activity Group on Supercomputing
This activity group provides a forum for computational mathematicians, computer scientists, computer architects, and computational scientists to exchange ideas on mathematical algorithms and computer architecture needed for high-performance computer systems. We promote the exchange of ideas by focusing on the interplay of analytical methods, numerical analysis, and efficient computation.

Workshop on Combinatorial Scientific Computing

February 11–13, 2020

*The **SIAM Workshop on Combinatorial Scientific Computing (CSC20)** is co-located with this conference (PP20). PP20 participants are invited to participate in the workshop (being held in the Quinault Ballroom, February 11–13, 2020). This workshop is sponsored by the SIAM Activity Group on Applied and Computational Discrete Algorithms.*

SIAM Events Mobile App
Scan the QR code with any QR reader and download the TripBuilder EventMobile™ app to your iPhone, iPad, iTouch or Android mobile device.
You can also visit *http://www.tripbuildermedia.com/apps/siamevents*

Society for Industrial and Applied Mathematics
3600 Market Street, 6th Floor
Philadelphia, PA 19104-2688 U.S.
Telephone: +1-215-382-9800 Fax: +1-215-386-7999
Conference E-mail: meetings@siam.org • Conference Web: *www.siam.org/meetings/*
Membership and Customer Service:
(800) 447-7426 (U.S. & Canada) or +1-215-382-9800 (worldwide)
https://www.siam.org/conferences/CM/Main/pp20
https://www.siam.org/conferences/CM/Main/csc20

Table of Contents

CSC Workshop

The ***SIAM Workshop on Combinatorial Scientific Computing (CSC20)*** is co-located with The ***SIAM Conference on Parallel Process for Scientific Computing (PP20)***. PP20 participants are invited to participate in the workshop, being held in the Quinault Ballroom, February 11–13, 2020.

More information is available on the CSC20 website *https://csc20.uni-jena.de/dates.html*.

Organizing Committee Co-chairs

George Biros
University of Texas at Austin, U.S.

Ulrike Meier Yang
Lawrence Livermore National Laboratory, U.S.

Organizing Committee

Olivier Beaumont
INRIA, France

Erik Boman
Sandia National Laboratories, U.S.

Aydin Buluc
Lawrence Berkeley National Laboratory, U.S.

Cris Cecka
NVIDIA Research & Stanford University, U.S.

Judith Hill
Oak Ridge National Laboratory, U.S.

Kirk Jordan
IBM T.J. Watson Research, U.S.

Axel Klawonn
University of Cologne, Germany

Petros Koumoutsakos
ETHZ, Switzerland

Maryam Mehri Dehnavi
University of Toronto, Canada

Boyana Norris
University of Oregon, U.S.

Rio Yokota
Tokyo Institute of Technology, Japan

Proceedings Paper Committee

Allison Baker
National Center for Atmospheric Research, U.S.

Costas Bekas
IBM, Switzerland

Erik Boman
Sandia National Laboratories, U.S.

Jed Brown
University of Colorado, U.S.

Aydin Buluc
Lawrence Berkeley National Laboratory, U.S.

Laura Grigori
INRIA, France

Judith Hill
Oak Ridge National Laboratory, U.S.

Kamer Kaya
Sabanci University, Turkey

Hatem Ltaief
King Abdullah University of Science and Technology, Saudi Arabia

Miriam Mehl
University of Stuttgart, Germany

Kengo Nakajima
University of Tokyo/RIKEN R-CCS, Japan

Yves Robert
ENS Lyon, France

Edgar Solomonik
University of Illinois, Urbana-Champaign, U.S.

Sivan Toledo
Tel-Aviv University, Israel

Gerhard Wellein
University of Erlangen, Germany

Rio Yokota
Tokyo Institute of Technology, Japan

Steering Committee

George Biros
University of Texas at Austin, U.S.

Matthias Bolten
Bergische Universität Wuppertal, Germany

Olaf Schenk
USI, Switzerland

Rich Vuduc
Georgia Tech, U.S.

Ulrike Meier Yang
Lawrence Livermore National Laboratory, U.S.

Conference Themes

Themes:

- Scalable Parallel Algorithms
- Large-Scale Parallel Applications
- Parallel Computer Systems
- Performance Analysis, Tuning, and Debugging
- Data Analysis and Visualization
- Fault-Tolerance
- Reproducibility
- Scientific Workflows
- Compilers and Programming Systems
- Education in HPC

Special Themes:

- Frontiers of Scientific Computing
- Algorithms and Applications on Multiphysics and Multiscale Computing
- Verification, Validation, and Uncertainty Quantification
- Post Moore Systems and Algorithms
- Integration of Scientific Computing and Machine Learning Methods

SIAM Registration Desk

Located in room 503 on the 5th floor, the registration desk is open at the following times:

Tuesday, February 11
7:00 a.m. – 3:30 p.m.

Wednesday, February 12
8:00 a.m. – 6:30 p.m.

Thursday, February 13
8:00 a.m. – 5:00 p.m.

Friday, February 14
8:00 a.m. – 5:00 p.m.

Saturday, February 15
8:00 a.m. – 2:00 p.m.

Hotel Address

Hyatt Regency Seattle

808 Howell Street, Seattle, WA 98101, U.S.

Hotel Telephone Number

To reach an attendee or leave a message, call +1-206-973-1234. If the attendee is a hotel guest, the hotel operator can connect you with the attendee's room.

Hotel Check-in and Check-out Times

Check-in time is 4:00 p.m.
Check-out time is 11:00 a.m.

Child Care

The Hyatt Regency Seattle recommends both Seattle Nanny Network, Inc. and Big Time Kid Care for attendees interested in child care services. Attendees are responsible for making their own child care arrangements.

Corporate Members and Affiliates

SIAM corporate members provide their employees with knowledge about, access to, and contacts in the applied mathematics and computational sciences community through their membership benefits. Corporate membership is more than just a bundle of tangible products and services; it is an expression of support for SIAM and its programs. SIAM is pleased to acknowledge its corporate members and sponsors. In recognition of their support, non-member attendees who are employed by the following organizations are entitled to the SIAM member registration rate.

Corporate/Institutional Members

The Aerospace Corporation
Air Force Office of Scientific Research
Amazon
Argonne National Laboratory
Bechtel Marine Propulsion Laboratory
The Boeing Company
CEA/DAM
Cirrus Logic
Department of National Defence (DND/CSEC)
DSTO- Defence Science and Technology Organisation, Edinburgh
Exxon Mobil
IDA Center for Communications Research, La Jolla
IDA Institute for Defense Analyses, Princeton
IDA Institute for Defense Analyses, Bowie, Maryland
Lawrence Berkeley National Laboratory
Lawrence Livermore National Labs
Lockheed Martin Maritime Systems & Sensors
Los Alamos National Laboratory
Max Planck Institute for Dynamics of Complex Technical Systems Magdeburg
Mentor Graphics
National Institute of Standards and Technology (NIST)
National Security Agency
Oak Ridge National Laboratory
Sandia National Laboratories
Schlumberger
Simons Foundation
United States Department of Energy
U.S. Army Corps of Engineers, Engineer Research and Development Center

List current as of January 2020.

Funding Agency

SIAM and the Organizing Committee extend their thanks and appreciation to National Science Foundation and the DOE for their support of this conference.

Join SIAM and save!

Leading the applied mathematics community

SIAM members save up to $130 on full registration for the SIAM Conference on Parallel Processing for Scientific Computing! Join your peers in supporting the premier professional society for applied mathematicians and computational scientists. SIAM members receive subscriptions to *SIAM Review*, *SIAM News* and *SIAM Unwrapped*, and enjoy substantial discounts on SIAM books, journal subscriptions, and conference registrations.

If you are not a SIAM member and paid the Non-Member rate to attend, you can apply the difference of $130 between what you paid and what a member paid towards a SIAM membership. Contact SIAM Customer Service for details or join at the conference registration desk.

If you are a SIAM member, it only costs $15 to join the SIAM Activity Group on Supercomputing (SIAG/SC). As a SIAG member, you are eligible for an additional $15 discount on this conference, so if you paid the SIAM member rate to attend the conference, you might be eligible for a free SIAG/Supercomputing membership. Check at the registration desk.

Students who paid the Student Non-Member Rate will be automatically enrolled as SIAM Student Members. Please go to *https://my.siam.org* to update your education and contact information in your profile.

If you attend a SIAM Academic Member Institution or are part of a SIAM Student Chapter you will be able to renew next year for free.

Join onsite at the registration desk, go to *https://www.siam.org/Membership/Join-SIAM* to join online or download an application form, or contact SIAM Customer Service:

Telephone: +1-215-382-9800 (worldwide); or 800-447-7426 (U.S. and Canada only)

Fax: +1-215-386-7999

Email: membership@siam.org

Postal mail: Society for Industrial and Applied Mathematics, 3600 Market Street, 6th floor, Philadelphia, PA 19104-2688 USA

Audio-Visual Set-Up in Meeting Rooms

SIAM does not provide computers for any speaker. When giving an electronic presentation, speakers must provide their own computers. SIAM is not responsible for the safety and security of speakers' computers.

A data (LCD) projector and screen will be provided in all technical session meeting rooms. The data projectors support both VGA and HDMI connections. Presenters requiring an alternate connection must provide their own adaptor.

All speakers should review SIAM's AV Policy.

Internet Access

Complimentary wireless internet access is available in the meeting space, lodging and public areas of the hotel. In addition, a limited number of email stations will also be available.

Registration Fee Includes

Admission to all technical sessions (SIAM Conference on Parallel Processing for Scientific Computing (PP20) and SIAM Workshop on Combinatorial Scientific Computing (CSC20)).

- Business Meeting (open to SIAG/SC members)
- Coffee breaks daily
- Room set-ups and audio/visual equipment
- Poster Session
- Welcome Reception

Job Postings

Please check at the registration desk for job posting availability, or visit *https://jobs.siam.org/.*

Poster Participant Information

The poster session is scheduled to be held on the fifth floor on Thursday, February 13, 2020, from 6:00 p.m. to 8:00 p.m.

Poster set-up time:

Poster boards will be available starting Wednesday, February 12, 2020 at 3:00 p.m.

All posters must be set-up by the beginning of the poster session which begins at 6:00 p.m. on Thursday, February 13, 2020.

Posters must be removed by:

Friday, February 14, 11:00 a.m.

SIAM Books and Journals

Please stop by the SIAM books table to browse and purchase our selection of textbooks and monographs. Enjoy discounted prices and free shipping. Complimentary copies of selected SIAM journals are available, as well. The books booth will be staffed from 9:00 a.m. through 5:00 p.m. The books table will open at 1:00 p.m. on Wednesday, February 12 and close at 12:00 p.m. on Saturday, February 15.

Table Top Displays

- Arm Ltd.
- Krell Institute
- SIAM
- Springer Nature

Conference Sponsors

2020 Conference Bag Sponsor

Name Badges

Please complete the emergency contact information on the back of your name badge.

Comments?

Comments are encouraged! Please send to: James Nagy, SIAM Vice President for Programs.

Social Events

Welcome Reception

Wednesday, February 12, 2020
6:00 p.m. – 8:00 p.m.

Business Meeting
(open to members of SIAM Activity Group on Supercomputing)

Thursday, February 13, 2020
5:15 p.m. – 6:00 p.m.
Complimentary beer and wine will be served.

Poster Session

Thursday, February 13, 2020
6:00 p.m. – 8:00 p.m.

Statement on Inclusiveness

As a professional society, SIAM is committed to providing an inclusive climate that encourages the open expression and exchange of ideas, that is free from all forms of discrimination, harassment, and retaliation, and that is welcoming and comfortable to all members and to those who participate in its activities. In pursuit of that commitment, SIAM is dedicated to the philosophy of equality of opportunity and treatment for all participants regardless of gender, gender identity or expression, sexual orientation, race, color, national or ethnic origin, religion or religious belief, age, marital status, disabilities, veteran status, field of expertise, or any other reason not related to scientific merit. This philosophy extends from SIAM conferences, to its publications, and to its governing structures and bodies. We expect all members of SIAM and participants in SIAM activities to work towards this commitment.

If you have experienced or observed behavior that is not consistent with the principles expressed above, you are encouraged to report any violation using the SIAM hotline, hosted by the third-party hotline provider, EthicsPoint. The information you provide will be sent to us by EthicsPoint on a totally confidential and anonymous basis if you should choose. You have our guarantee that your comments will be heard. Please submit reports at *http://www.siam.ethicspoint.com/.*

Please Note

SIAM is not responsible for the safety and security of attendees' belongings. Do not leave your property unattended. Additionally, please silence your devices.

Recording of Presentations

Audio and video recording of presentations is prohibited without the written permission of the presenter and SIAM.

Social Media

#SIAMPP20 ***@TheSIAMNews***

Changes to the Program

To reduce the environmental footprint of SIAM conferences, SIAM's Board of Directors has decided to eliminate paper copies of conference programs. A downloadable program PDF is available on the conference webpage under "Program and Abstracts"; however, attendees are strongly encouraged to utilize the Mobile App or visit *https://www.siam.org/conferences/cm/program/program-and-abstracts/pp20-program-abstracts* to view the Online Program Schedule. The Mobile App and Online Program Schedule contain the most up-to-date scheduling information.

SIAM Events Mobile App Powered by TripBuilder®

To enhance your conference experience, we're providing a state-of-the-art mobile app to give you important conference information right at your fingertips. With this TripBuilder EventMobile™ app, you can:

- Create your own custom schedule
- View sessions, speakers, exhibitors and more
- Take notes and export them to your email
- View award-winning TripBuilder recommendations for the meeting location
- Get instant alerts about important conference info

SIAM Events Mobile App

Scan the QR code with any QR reader and download the TripBuilder EventMobile™ app to your iPhone, iPad, iTouch or Android mobile device.
You can also visit *http://www.tripbuildermedia.com/apps/siamevents*

Invited Plenary Speakers

*** All Invited Plenary Presentations will take place in Elwha Ballroom ***

Wednesday, February 12

5:15 p.m. – 6:00 p.m.

IP1 Parallel Tomographic Reconstruction – Where Combinatorics Meets Geometry
Rob H. Bisseling, *Utrecht University, The Netherlands*

Thursday, February 13

8:30 a.m. – 9:15 a.m.

IP2 Accelerated-Node-Enabled Computational and Data Science: It's not just for Exascale
Douglas Kothe, *Oak Ridge National Laboratory, U.S.*

2:05 p.m. – 2:50 p.m.

IP3 Development of an Eigen-Analysis Engine for Large-Scale Simulation and Big Data Analysis
Tetsuya Sakurai, *University of Tsukuba, Japan*

Friday, February 14

2:05 p.m. – 2:50 p.m.

IP4 Modeling of Heterogeneous Computing Systems and Their Usages
Hyesoon Kim, *Georgia Institute of Technology, U.S.*

Saturday, February 15

8:30 a.m. – 9:15 a.m.

IP5 Methods and Models for Reducing Communication
Luke Olson, *University of Illinois at Urbana-Champaign, U.S.*

9:25 a.m. – 10:10 a.m.

IP6 Cognitive Discovery: Pushing the Frontier of Technical R&D with AI
Costas Bekas, *IBM Research - Zurich, Switzerland*

Prize Lectures

*** All Prize Lectures will take place in Elwha Ballroom ***

Friday, February 14

8:30 a.m. – 9:15 a.m.

SP1 SIAG Best Paper Prize: The BLIS Framework: Experiments in Portability

Robert A. van de Geijn, *The University of Texas at Austin, U.S.*

9:25 a.m. – 9:55 a.m.

SP2 SIAG/Supercomputing Early Career Prize

Scalable Algorithms for Tensor Computations

Edgar Solomonik, *University of Illinois at Urbana-Champaign, U.S.*

9:55 a.m. – 10:25 a.m.

SP3 SIAG/Supercomputing Career Prize

Ghosts of Parallel Computing: Past, Present, and Future

Steve Plimpton, *Sandia National Laboratories, U.S.*

Minitutorial

Wednesday, February 12

1:00 p.m. – 2:40 p.m.

MT1 Combinatorial Optimization on Quantum Computers

Room: 605

Quantum computing has the potential to provide speedups over classical state-of-the-art for some combinatorial optimization problems. Recent advances in both hardware and algorithm development have made it possible to solve small problems on modern quantum computers. Combinatorial optimization problems (especially NP-hard problems) are of particular interest, since for many of these problems best classical algorithms can not provide solutions of sufficient quality in reasonable time. One such problem is MaxCut on graphs. In this minitutorial, we will introduce the MaxCut problem and explain how it can be solved on IBM quantum computers available on the cloud today using the Qiskit framework. Our presentation will assume little to no prior knowledge of quantum computation. Moreover, we will provide examples of how more complicated problems can be solved using the QAOA (Quantum Approximate Optimization Algorithm). The tutorial will be interactive and will use Jupyter notebooks to explain step by step how to formulate and solve optimization problems on quantum computers.

Participants are encouraged to bring their own laptops.

The recommended way to run the hands-on Jupyter notebooks is using IBM Quantum Experience web interface. Please create an account at *https://quantum-computing.ibm.com/* before the tutorial (you might have to create an IBM ID if you do not have one). Additional instructions and the hands-on notebook are available at *https://github.com/rsln-s/SIAM_PP_20_minitutorial*

Speakers:

Yuri Alexeev, *Argonne National Laboratory, U.S.*

Ilya Safro, *Clemson University, U.S.*

Ruslan Shaydulin, *Clemson University, U.S.*

Panel

Thursday, February 13

9:25 a.m. – 10:25 a.m.

PD1 Is AI transforming HPC or HPC transforming AI?

Room: Elwha Ballroom

Chair:

Aparna Chandramowlishwaran, *University of California, Irvine, U.S.*

Panelists:

Srinivas Aluru, *Iowa State University, U.S.*

Tamara G. Kolda, *Sandia National Laboratories, U.S.*

Dong Li, *University of California, Merced, U.S.*

SIAM Activity Group on Supercomputing (SIAG/SC)

siam.org/Activity-Groups/SIAG/SC

A great way to get involved!

The SIAM Activity Group on Supercomputing provides a forum for computational mathematicians, computer scientists, computer architects, and computational scientists to exchange ideas on mathematical algorithms and computer architecture needed for high-performance computer systems.

ACTIVITIES INCLUDE

- Special Sessions at SIAM meetings
- Biennial conference
- SIAG/Supercomputing Career Prize
- SIAG/Supercomputing Early Career Prize
- SIAG/Supercomputing Best Paper Prize

BENEFITS OF SIAG/SC MEMBERSHIP

- Listing in the SIAG's online membership directory
- Additional $15 discount on registration at the SIAM Conference on Parallel Processing
- Eligibility for candidacy for SIAG/SC office
- Participation in the selection of SIAG/SC officers
- Electronic communications about recent developments in your specialty

ELIGIBILITY

- Be a current SIAM member.

COST

- $15 per year
- Student members can join two activity groups for free!

2020/2021 SIAG/SC OFFICERS

Chair
OLAF SCHENK
Università della Svizzera Italiana

Vice Chair
MATTHIAS BOLTEN
Universität Wuppertal

Program Director
KEITA TERANISHI
Sandia National Laboratories

Secretary
AMANDA RANDLES
Duke University

To Join:
my.siam.org

siam 2020 Conference on Parallel Processing for Scientific Computing

siam Society for Industrial and Applied Mathematics

Schedule


siam
2020
Conference on
Parallel Processing for
Scientific Computing

Tuesday, February 11

Registration

7:00 a.m.-3:30 p.m.

Room: 503

SIAM Workshop on Combinatorial Scientific Computing (CSC20)

csc20.uni-jena.de/program.html

8:30 a.m.-4:20 p.m.

Room: Quinault Ballroom

Wednesday, February 12

Registration

8:00 a.m.-6:30 p.m.

Room: 503

SIAM Workshop on Combinatorial Scientific Computing (CSC20)

csc20.uni-jena.de/program.html

9:15 a.m.-6:00 p.m.

Room: Quinault Ballroom

Wednesday, February 12

MT1

Combinatorial Optimization on Quantum Computers

1:00 p.m.-2:40 p.m.

Room: 605

Chair: Yuri Alexeev, Argonne National Laboratory, U.S.

Chair: Ilya Safro, Clemson University, U.S.

Chair: Ruslan Shaydulin, Clemson University, U.S.

Quantum computing has the potential to provide speedups over classical state-of-the-art for some combinatorial optimization problems. Recent advances in both hardware and algorithm development have made it possible to solve small problems on modern quantum computers. Combinatorial optimization problems (especially NP-hard problems) are of particular interest, since for many of these problems best classical algorithms can not provide solutions of sufficient quality in reasonable time. One such problem is MaxCut on graphs. In this minitutorial, we will introduce the MaxCut problem and explain how it can be solved on IBM quantum computers available on the cloud today using the Qiskit framework. Our presentation will assume little to no prior knowledge of quantum computation. Moreover, we will provide examples of how more complicated problems can be solved using the QAOA (Quantum Approximate Optimization Algorithm). The tutorial will be interactive and will use Jupyter notebooks to explain step by step how to formulate and solve optimization problems on quantum computers. Participants are encouraged to bring their own laptops. The recommended way to run the hands-on Jupyter notebooks is using IBM Quantum Experience web interface. Please create an account at *https://quantum-computing.ibm.com/* before the tutorial (you might have to create an IBM ID if you do not have one). Additional instructions and the hands-on notebook are available at *https://github.com/rsln-s/SIAM_PP_20_minitutorial*

Speakers:

Yuri Alexeev
Argonne National Laboratory, U.S.

Ilya Safro
Clemson University, U.S.

Ruslan Shaydulin
Clemson University, U.S.

Wednesday, February 12

MS1

Industrial Mathematical Software

1:00 p.m.-2:15 p.m.

Room: Elwha Ballroom

Mathematical software, which is essential to all areas of science, engineering, and machine learning, is continually being developed and improved to attack new applications and to keep up with breakthroughs in hardware technologies. In this minisymposium we highlight recent advances in industrial software, covering both CPUs and GPUs, as well as low-level optimizations through high-level software packages.

Organizer: Sarah Knepper
Intel Corporation, U.S.

Organizer: Pat Quillen
MathWorks, U.S.

1:00-1:20 Intel Software Solutions for Diverse Computing Architectures
Sarah Knepper and Shane Story, Intel Corporation, U.S.

1:25-1:45 Strategies for Exploiting Multicore in High-Level Software Packages
Pat Quillen and Christopher Turnes, MathWorks, U.S.

1:50-2:10 What's New in the Cuda Math Libraries
Timothy Costa and Harun Bayraktar, NVIDIA, U.S.

Wednesday, February 12

MS2

Meaningful Performance Indicators for Scientific Computing

1:00 p.m.-2:40 p.m.

Room: 501

It is essential be able to evaluate computational and data analysis systems during many stages of their life cycle, from design, to development, to deployment, and then life-cycle long system improvement. Therefore, the metrics and methods for consistent and meaningful performance evaluation is important for the scientific computing and data analysis. This symposium focuses on the performance evaluation for systems developed as the fundamental infrastructure of scientific research. A wide area related to the performance evaluation, such as benchmark design, initial and on-going performance metrics, comprehensive evaluation methodologies, productivity and reproducibility will be discussed.

Organizer: Miwako Tsuji
RIKEN Advanced Institute for Computational Science, Japan

Organizer: Mitsuhisa Sato
RIKEN Advanced Institute for Computational Science, Japan

1:00-1:20 Why Run Real Benchmarks? Can't I Just Run Linpack to Understand My System?
William T. Kramer, National Center for Supercomputing Applications and University of Illinois, U.S.

1:25-1:45 Evaluation of Large Scale Systems with Focus on Application Performance: the Benchmarking Perspective
Piotr Luszczek, University of Tennessee, Knoxville, U.S.

1:50-2:10 How to Monitor the Performance Evolution of a Large HPC System: a System and Application Points of View
Matthieu Hautreux, Marc Perache, and *Jean-Christophe Weill*, CEA, DAM, DIF, France

2:15-2:35 Performance Modeling of Applications by Benchmarks
Miwako Tsuji, RIKEN Advanced Institute for Computational Science, Japan

Wednesday, February 12

MS3

Experiences in Developing GPU Support for Department of Energy Math Libraries - Part I of II

1:00 p.m.-2:40 p.m.

Room: 502

For Part 2 see MS12

The FASTMath (Frameworks, Algorithms and Scalable Technologies for Mathematics) Institute is a R&D project funded by the SciDAC Program at the U.S. Department of Energy (DOE). The goal of FASTMath is to develop and deploy scalable mathematical algorithms and software tools for reliable simulation of complex physical phenomena and collaborating with DOE domain scientists to ensure the usefulness and applicability of the work in the project. The focus of FASTMath is strongly driven by the requirements of DOE application scientists who require fast, accurate, and robust forward simulation along with the ability to efficiently perform ensembles of simulations in optimization or uncertainty quantification studies. This minisymposium will present work by FASTMath participants focused on developing GPU support for their numerical mathematics libraries. Presentations will include discussion of different strategies as well as performance results.

Organizer: Carol S. Woodward
Lawrence Livermore National Laboratory, U.S.

Organizer: Ann S. Almgren
Lawrence Berkeley National Laboratory, U.S.

1:00-1:20 Implementation of Fast Eigensolvers on GPUs
Chao Yang, Lawrence Berkeley National Laboratory, U.S.

1:25-1:45 Strategies, Challenges, and Lessons Learned in Developing GPU Support for SUNDIALS
Cody J. Balos, David J. Gardner, and Carol S. Woodward, Lawrence Livermore National Laboratory, U.S.; Daniel R. Reynolds, Southern Methodist University, U.S.; Alan C. Hindmarsh, Lawrence Livermore National Laboratory, U.S.

continued on next page

Wednesday, February 12

MS3

Experiences in Developing GPU Support for Department of Energy Math Libraries - Part I of II

1:00 p.m.-2:40 p.m.

continued

1:50-2:10 PETSc's Accelerator Model and Algebraic Multigrid Work and Data Placement at Extreme-Scale

Mark Adams, Lawrence Berkeley National Laboratory, U.S.; Matthew G. Knepley, State University of New York at Buffalo, U.S.; Joseph Pusztay, University at Buffalo, U.S.

2:15-2:35 Investigating Quasi-Newton Outer Product Representations on GPUs

Alp Dener, Argonne National Laboratory, U.S.

Wednesday, February 12

MS4

Physical-Based Modeling and Machine Learning for Biological and Environmental Sciences - Part I of II

1:00 p.m.-2:40 p.m.

Room: 505

For Part 2 see MS13

Scientific computing is critical in understanding the biological, biogeochemical, and physical processes that span from molecular and genomics-controlled scales in biological systems, to the regional and global scales in climate and the earth system. The integration of physical models and machine learning play a significant role in both bio-science discovery and neural network based algorithms, where high-performance computing (HPC) is heavily used. Similarly, recent data science advances hold significant opportunity to transform our ability to rapidly use diverse datasets and physics-based models for predicting how earth systems respond to perturbations. On the other hand, the improved visualization and predictions in various domains of the biological and environmental systems translate to innovations of technologies and methodologies that may contribute to artificial intelligence. Those computationally intensive models usually rely on parallel processing and are impossible to implement without HPC clusters. In this minisymposium, we gather researchers from various fields in biological and environmental sciences to talk about the frontiers of scientific computing and machine learning in their domain work, which covers bioinformatics, hemodynamics and neuroscience in biological sciences, and atmosphere and hydrology research in environmental science. Relevant work that leverages PDE-constrained optimization, generative models, and deep learning using HPC will be discussed.

Organizer: Zhe Bai
Lawrence Berkeley National Laboratory, U.S.

Organizer: Zexuan Xu
Lawrence Berkeley National Laboratory, U.S.

1:00-1:20 High-Resolution Visualization of In Vivo Blood Flow from Low-Resolution MRI Scans using Computational Fluid Dynamics and Optimization

Matthew J. Zahr, University of Notre Dame, U.S.

1:25-1:45 Drug Response Prediction and Generative Models for Molecules

Fangfang Xia, Argonne National Laboratory, U.S.

1:50-2:10 Data-Driven Models of the Mouse Mesoscale Connectome: Network Structure and Functionality

Hannah Choi, University of Washington, U.S.

2:15-2:35 Scalable Deep Learning of Biological Dynamical Systems

Kathleen Champion, University of Washington, U.S.; *Bethany Lusch*, Argonne National Laboratory, U.S.; J. Nathan Kutz and Steven Brunton, University of Washington, U.S.

continued in next column

Wednesday, February 12

MS5

Parallel Simulation of Circuits, Devices, and Electromagnetics Environments Effects

1:00 p.m.-2:40 p.m.

Room: 507

This minisymposium will present the current research, development, and performance of parallel simulation tools for electrical systems and electromagnetic environments effects. These tools enable simulation of integrated circuits (Xyce), devices (Charon), and electromagnetics (EMPIRE/GEMMA) at unprecedented scales on advanced architectures. This collection of talks will present the unique needs and challenges in achieving efficient and scalable parallel performance for each of these simulation tools. This rich diversity permeates every layer of the simulation infrastructure from system assembly through the implementation of numerical methods for solving linear systems. Current results will be presented for each of the simulation tools and future research directions will be identified for addressing the needs of scaling, robustness, and efficiency on next-generation computing platforms.

Organizer: Heidi K. Thornquist
Sandia National Laboratories, U.S.

1:00-1:20 Toward Exascale Plasma Simulations using Particle in Cell (pic) Algorithms
Matthew Bettencourt, Sandia National Laboratories, U.S.

1:25-1:45 Gemma: A Sandia National Laboratories Electromagnetic Code for Heterogeneous Computer Architectures
Brian Zinser, Samuel Blake, Robert Pfeiffer, Andy Huang, John Himbele, Brian A. Freno, Vinh Dang, Joseph Kotulski, Sivasankaran Rajamanickam, William Johnson, Salvatore Campione, and William Langston, Sandia National Laboratories, U.S.

1:50-2:10 Technical Computer Aided Design Modeling of Semiconductor Devices in Parallel Computing Architectures
Lawrence C. Musson, Gary Hennigan, Jason Gates, Xujiao Gao, Mihai Negoita, and Andy Huang, Sandia National Laboratories, U.S.

2:15-2:35 Parallel Simulation of Large-Scale Integrated Circuits Using Xyce
Heidi K. Thornquist, Sandia National Laboratories, U.S.

continued in next column

Wednesday, February 12

MS6

High Performance Krylov Subspace Methods: Theory, Implementation, and Application - Part I of III

1:00 p.m.-2:40 p.m.

Room: 512

For Part 2 see MS16

Krylov subspace methods are widely used for solving very large (sparse) linear systems and eigenvalue problems. However, when these algorithms are implemented on parallel architectures in their original forms, communication (rather than floating point operations) becomes the dominant cost. To address this, many mathematically equivalent implementations, designed explicitly to reduce the time spent on communication, have been introduced. However on some problems these new formulations can behave (very) differently in finite precision than the original implementations. This makes the design of numerically stable variants for parallel architectures challenging. In this minisymposium we hear from researchers about the design and analysis of high performance Krylov subspace methods, practical concerns relating to implementing these methods on modern parallel architectures, and applications relevant today.

Organizer: Tyler Chen
University of Washington, U.S.

1:00-1:20 The Numerical Stability Analysis of Parallel Conjugate Gradient Methods: Historical Context and Methodology
Miro Rozloznik, Academy of Sciences of the Czech Republic, Prague, Czech Republic

1:25-1:45 Inexactness and Compression in Krylov Subspace Methods
Nick Schenkels and Emmanuel Agullo, Inria, France; Franck Cappello and Sheng Di, Argonne National Laboratory, U.S.; Luc Giraud, Inria, France; Xin Liang, Argonne National Laboratory, U.S.

continued on next page

Wednesday, February 12

MS6

High Performance Krylov Subspace Methods: Theory, Implementation, and Application - Part I of III

1:00 p.m.-2:40 p.m.

continued

1:50-2:10 Speculations on the Future of Krylov Methods using Results from Emerging Architectures

Hannah M. Morgan, Argonne National Laboratory, U.S.

2:15-2:35 Enlarged Krylov Methods and 2-Level Preconditionner for the Map-Making Problem in CMB Data Analysis

Thibault Cimic and Laura Grigori, Inria Paris, France

Wednesday, February 12

MS7

Advances in Parallel-in-Time Integration Methods

1:00 p.m.-2:40 p.m.

Room: 603

The rapid increase and availability of massively parallel computational resources has inspired new ways to parallelize numerical algorithms associated with evolutionary processes. The need for parallel-in-time integration methods is mainly being driven by changes in computer architectures where future speedups will be available through greater concurrency rather than reduced clock-speeds, which are stagnant. Parallel-in-time approaches introduce a new dimension of parallelism by distributing workload to multiple processors along the time domain of dynamical systems. This minisymposium presents recent advances in both theoretical and algorithmic aspects of parallel-in-time approaches for simulation and optimization with evolutionary processes, with applications to realistic scenarios in science and engineering.

Organizer: David J. Gardner
Lawrence Livermore National Laboratory, U.S.

Organizer: Stefanie Guenther
Lawrence Livermore National Laboratory, U.S.

1:00-1:20 Parallel-in-Time with Sundials and Xbraid

David J. Gardner and Robert Falgout, Lawrence Livermore National Laboratory, U.S.; Daniel R. Reynolds, Southern Methodist University, U.S.; Carol S. Woodward, Lawrence Livermore National Laboratory, U.S.

1:25-1:45 Parallel-in-Time Simulation of Electrical Powergrids with Unscheduled Events

Stefanie Guenther, Carol S. Woodward, and Robert Falgout, Lawrence Livermore National Laboratory, U.S.

1:50-2:10 A Layer-Parallel Approach for Training Deep Neural Networks

Eric C. Cyr, Sandia National Laboratories, U.S.; Stefanie Guenther, Lawrence Livermore National Laboratory, U.S.; Lars Ruthotto, Emory University, U.S.; Jacob B. Schroder, University of New Mexico, U.S.; Nicolas R. Gauger, Technische Universität Kaiserslautern, Germany

2:15-2:35 Multigrid Reduction in Time for the Shallow Water Equations using Asymptotic Techniques

Nicholas Abel, University of New Mexico, U.S.

continued in next column

Wednesday, February 12

MS8

Co-Design of Networking for Scientific HPC Applications

1:00 p.m.-2:40 p.m.

Room: 604

Modernizing scientific codes to run performantly on multiple emerging architectures is a task of great complexity and challenge for computational scientists today. We examine the intersection between multiple topic areas related to networking and scientific HPC applications; specifically, the evaluation of both network tapering, and new topologies, and the evaluation of on-node bottlenecks, that impact internode communication. This minisymposium will address strategies to act on issues that arise from networking, and evaluations of new topics, for some large codes today.

Organizer: Geoff Womeldorff
Los Alamos National Laboratory, U.S.

1:00-1:20 The Last Centimeter: Trials, Tribulations and Bottlenecks in Getting Data onto the Wire
Ian Karlin, Lawrence Livermore National Laboratory, U.S.

1:25-1:45 Network Bottleneck Handling in Multi-Threaded MPI Context
Julien C. Jaeger, CEA, France

1:50-2:10 A Tool (interconnect) is Only As Good As its User (the Network Stack): Solving Network Bottlenecks via Software Stack Simulation in the Structural Simulation Toolkit
Jeremiah Wilke, Sandia National Laboratories, U.S.

2:15-2:35 GPU Communication Considerations for a Modern Multi Physics Code
Michael Lang, Los Alamos National Laboratory, U.S.

Wednesday, February 12

MS9

Parallel Adaptive Multigrid - Part I of II

1:00 p.m.-2:40 p.m.

Room: 606

For Part 2 see MS19
The multigrid method is well-known as an efficient and highly-scalable solver for many applications. Parallel adaptive mesh refinement nicely integrates with the geometric multigrid method and parallel algebraic multigrid methods are established as state-of-the-art parallel sparse matrix solvers. This session is dedicated to advances for the multigrid method with a focus on adaptive and parallel simulations. It offers a platform to present applications that allow for profitable employment of adaptive mesh refinement and implementations for high-performance computing.

Organizer: Gillian Queisser
Temple University, U.S.

Organizer: Andreas Vogel
Ruhr-University Bochum, Germany

Organizer: Gabriel Wittum
King Abdullah University of Science & Technology (KAUST), Saudi Arabia

1:00-1:20 Parallel Geometric Multigrid for Continuum Models and Numerical Optimization
Andreas Vogel, Ruhr-University Bochum, Germany

1:25-1:45 Smooth Subdivision Multigrid and Large Scale Simulations in Life Science
Gillian Queisser, Temple University, U.S.

1:50-2:10 A Dimension Switching Multigrid Method and Applications
Qingguang Guan and Gillian Queisser, Temple University, U.S.

2:15-2:35 A Space-Time Semi-Geometric Multigrid Method for Electrophysiology
Patrick Zulian, Università della Svizzera italiana, Switzerland

Wednesday, February 12

CP1

UQ and Stochastic Processes

1:00 p.m.-2:15 p.m.

Room: 508

Chair: Daniele E. Schiavazzi, University of Notre Dame, U.S.

1:00-1:20 Toward a Predictive Model to Monitor the Balance Between Discretization and Rounding Errors in Hydrodynamic Simulations
William Weens, Alexandre Merasli, Thibaud Vazquez-Gonzalez, and R?mi Chauvin, CEA, France

1:25-1:45 The Validation of the Simulated HIRF Effects in Metallic Cases using Parallel FDTD Solver on a High Performance Computer
Yueqian Wu, Institute of Applied Physics and Computational Mathematics, China

1:50-2:10 A Scalable Explicit Finite Element Solver for Cardiovascular Models with Uncertain Material Properties
Xue Li and Daniele E. Schiavazzi, University of Notre Dame, U.S.

Wednesday, February 12

CP2

Efficient Methods for PDEs and IDEs

1:00 p.m.-2:15 p.m.

Room: 607

Chair: Ron Gonzales, Idaho State University, U.S.

1:00-1:20 High-Order Spatio-Temporally Parallel "Fast-Hybrid" Wave Equation Solver at $\mathcal{O}(1)$ Sampling Cost

Thomas Anderson and Oscar P. Bruno, California Institute of Technology, U.S.; Mark Lyon, University of New Hampshire, U.S.

1:25-1:45 High Order Scalable FFT-Krylov Subspace Methods for the 3D Convection Diffusion Equation

Ron Gonzales, Yun Teck Lee, and Yury A. Gryazin, Idaho State University, U.S.

1:50-2:10 An Efficient Method for Solving the Fractional Fredholm Integrodifferential Equation

Muhammed I. Syam, United Arab Emirates University, United Arab Emirates

Coffee Break

2:40 p.m.-3:10 p.m.

Room: Gallery - 5th Floor

Wednesday, February 12

MS10

Advances in Algorithms Exploiting Low Precision Floating-Point Arithmetic - Part I of II

3:10 p.m.-4:50 p.m.

Room: Elwha Ballroom

For Part 2 see MS21

In the June 2019 Top 500 list, 133 systems are accelerator-based, more than half of which support half precision floating-point arithmetic. Since low precision floating-point formats are increasingly being supported by hardware vendors, developing algorithms and software that can exploit these formats is of increasing importance. This minisymposium will give a broad overview of recent contributions in exploiting low precision arithmetic in scientific computing, and details of applications in which it can successfully be used.

Organizer: Nicholas J. Higham
The University of Manchester, UK

Organizer: Srikara Pranesh
The University of Manchester, UK

3:10-3:30 Low Precision Floating-Point Arithmetic for the Solution of Linear System of Equations

Nicholas J. Higham and *Srikara Pranesh*, The University of Manchester, UK

3:35-3:55 Performance and Accuracy of Mixed-Precision Matrix Factorizations with GPU Tensor Cores

Pierre Blanchard and Nicholas J. Higham, The University of Manchester, UK; Florent Lopez, University of Tennessee, Knoxville, U.S.; *Theo Mary* and Srikara Pranesh, The University of Manchester, UK

4:00-4:20 Opportunities for Multi Precision Computation in Memory Bound Applications

Nicholas J. Higham, The University of Manchester, UK; Craig Lucas, Numerical Algorithms Group Ltd, United Kingdom; Francoise Tisseur and *Mawussi Zounon*, The University of Manchester, UK

4:25-4:45 Addressing the Communication Bottleneck: Towards a Modular Precision Ecosystem for High Performance Computing

Hartwig Anzt, Karlsruhe Institute of Technology, Germany and University of Tennessee, U.S.; Terry Cojean and *Thomas Grützmacher*, Karlsruhe Institute of Technology, Germany

continued in next column

Wednesday, February 12

MS11

Formal (Mathematical) Methods Enabling Applications of Quantum Computers

3:10 p.m.-4:50 p.m.

Room: 501

What are quantum computers (QCs) today and what will they be in 5, 10, or 15 years? In what mathematical models should subject-matter experts formulate their problems so applications will benefit from QCs in a sustainable way? How will subject-matter experts develop applications for QCs? What are to be considered best practices in developing and programming quantum computing architectures? What opportunities for quasi-automatic transformation by new compiler-like tools can exploit the power of well-matched mathematical models? As the first generation of commercial QCs matures, the answers to such questions have become fundamentally important for the successful evolution of QCs to widespread industrial use. Whereas the development of previous-generation QCs relied heavily on hardware-focused heuristics, with this workshop, we aim to bring together leading experts working in areas of formal mathematical methods to discuss how these methods can address the challenges of developing next-generation and future QCs and developing QC applications that enable broad use.

Organizer: Faisal Shah Khan
Khalifa University of Science, Technology and Research, United Arab Emirates

Organizer: Travis Humble
Oak Ridge National Laboratory, U.S.

Organizer: Yaakov Weinstein
MITRE Corporation, U.S.

Organizer: Steve P. Reinhardt
Quantum Computing Inc., U.S.

3:10-3:30 Nash Embedding: A Road Map to Realizing Quantum Hardware
Faisal Shah Khan, Khalifa University of Science, Technology and Research, United Arab Emirates

3:35-3:55 GAMA: Quantum and Quantum-Inspired Algorithms for Non-Linear Integer Optimization
Sridhar Tayur, Carnegie Mellon University, U.S.

4:00-4:20 Adapting Engineering Design Practices to Mathematical Methods in Quantum Computing
Steven A. Bleiler and Marek Perkowski, Portland State University, U.S.

4:25-4:45 Commutative and Non Commutative Geometry for Noisy Intermediate-Scale Quantum (NISQ) Computations
Raouf Dridi, Carnegie Mellon University, U.S.

continued in next column

Wednesday, February 12

MS12

Experiences in Developing GPU Support for Department of Energy Math Libraries - Part II of II

3:10 p.m.-4:50 p.m.

Room: 502

For Part 1 see MS3
The FASTMath (Frameworks, Algorithms and Scalable Technologies for Mathematics) Institute is a R&D project funded by the SciDAC Program at the U.S. Department of Energy (DOE). The goal of FASTMath is to develop and deploy scalable mathematical algorithms and software tools for reliable simulation of complex physical phenomena and collaborating with DOE domain scientists to ensure the usefulness and applicability of the work in the project. The focus of FASTMath is strongly driven by the requirements of DOE application scientists who require fast, accurate, and robust forward simulation along with the ability to efficiently perform ensembles of simulations in optimization or uncertainty quantification studies. This minisymposium will present work by FASTMath participants focused on developing GPU support for their numerical mathematics libraries. Presentations will include discussion of different strategies as well as performance results.

Organizer: Carol S. Woodward
Lawrence Livermore National Laboratory, U.S.

Organizer: Ann S. Almgren
Lawrence Berkeley National Laboratory, U.S.

3:10-3:30 AMReX on GPUs: Strategies, Challenges and Lessons Learned
Kevin N. Gott, Lawrence Berkeley National Laboratory, U.S.

3:35-3:55 MFEM: Accelerating Efficient Solution of PDEs at Exascale
Yohann Dudouit, Lawrence Livermore National Laboratory, U.S.

continued on next page

Wednesday, February 12

MS12

Experiences in Developing GPU Support for Department of Energy Math Libraries - Part II of II

3:10 p.m.-4:50 p.m.

continued

4:00-4:20 Progress on the Development of Mesh-Based PIC for Fusion Codes

Cameron W. Smith, Gerrett Diamond, and Chonglin Zhang, Rensselaer Polytechnic Institute, U.S.; Eisung Yoon, Ulsan National Institute of Science and Technology, South Korea; Gopan Perumpilly, Onkar Sahni, and Mark S. Shephard, Rensselaer Polytechnic Institute, U.S.

4:25-4:45 Portable Performance for AMR on GPUs: The Proto Approach

Brian Van Straalen, Lawrence Berkeley National Laboratory, U.S.

Wednesday, February 12

MS13

Physical-Based Modeling and Machine Learning for Biological and Environmental Sciences - Part II of II

3:10 p.m.-4:50 p.m.

Room: 505

For Part 1 see MS4

Scientific computing is critical in understanding the biological, biogeochemical, and physical processes that span from molecular and genomics-controlled scales in biological systems, to the regional and global scales in climate and the earth system. The integration of physical models and machine learning play a significant role in both bio-science discovery and neural network based algorithms, where high-performance computing (HPC) is heavily used. Similarly, recent data science advances hold significant opportunity to transform our ability to rapidly use diverse datasets and physics-based models for predicting how earth systems respond to perturbations. On the other hand, the improved visualization and predictions in various domains of the biological and environmental systems translate to innovations of technologies and methodologies that may contribute to artificial intelligence. Those computationally intensive models usually rely on parallel processing and are impossible to implement without HPC clusters. In this mini-symposium, we gather researchers from various fields in biological and environmental sciences to talk about the frontiers of scientific computing and machine learning in their domain work, which covers bioinformatics, hemodynamics and neuroscience in biological sciences, and atmosphere and hydrology research in environmental science. Relevant work that leverages PDE-constrained optimization, generative models, and deep learning using HPC will be discussed.

Organizer: Zhe Bai
Lawrence Berkeley National Laboratory, U.S.

Organizer: Zezuan Xu
Lawrence Berkeley National Laboratory, U.S.

3:10-3:30 Infusing Physics and Domain Knowledge into ML and DL Models

Karthik Kashinath, Lawrence Berkeley National Laboratory, U.S.; Jinlong Wu, California Institute of Technology, U.S.; Chiyu Jiang, University of California, Berkeley, U.S.; Adrian Albert, Lawrence Berkeley National Laboratory, U.S.; Heng Xiao, Virginia Tech, U.S.; Philip Marcus, University of California, Berkeley, U.S.; Mr Prabhat, Lawrence Berkeley National Laboratory, U.S.

3:35-3:55 Use of Deep Neural Networks for Estimating Subsurface Property Field from Time-Lapse Geophysical Imaging

Xingyuan Chen, Erol Cromwell, and Tim Johnson, Pacific Northwest National Laboratory, U.S.; Glenn Hammond, Sandia National Laboratories, U.S.; Hongsheng Wang, Pacific Northwest National Laboratory, U.S.

4:00-4:20 Satellite Precipitation Estimation at Uci Chrs: Algorithm Development & Challenges

Phu Nguyen, Kuolin Hsu, Dan Braithwaite, and Soroosh Sorooshian, University of California, Irvine, U.S.

4:25-4:45 Application of Information Theory in Understanding Hydrological Interactions and Model Diagnostics

Edom Moges, University of California, Berkeley, U.S.

continued in next column

Wednesday, February 12

MS15

Nonlinear Preconditioning

3:10 p.m.-4:50 p.m.

Room: 508

Traditionally, the nonlinear systems arising from the discretization of nonlinear partial differential equations are solved by variants of Newton's method. Often, a Newton-Krylov approach in combination with suitable preconditioners (multigrid, domain decomposition) is used. If necessary, globalization techniques, e.g., trust region, line search, load stepping, etc. are applied additionally. Nonlinear preconditioning (or nonlinear elimination) is an alternative approach to improve the robustness and convergence properties of nonlinear solvers. But these methods also have a great potential to increase parallel scalability and to decrease time to solution. Here, one important property is localization of work and to decrease reduction of communication, which can lead to scalability up to more than a million parallel processes. Different approaches, including nonlinear Schwarz methods (ASPIN) and nonlinear FETI-DP/BDDC methods, are considered.

Organizer: Axel Klawonn
Universitaet zu Koeln, Germany

Organizer: Oliver Rheinbach
Technische Universität Bergakademie Freiberg, Germany

Organizer: Matthew G. Knepley
State University of New York at Buffalo, U.S.

3:10-3:30 Additive and Hybrid Nonlinear Two-Level Schwarz Methods and Energy Minimizing Coarse Spaces for Unstructured Grids
Martin Lanser, Universitaet zu Koeln, Germany

3:35-3:55 Detecting and Mitigating Stagnation in Nonlinear Multiphysics Problems
Matthew G. Knepley, State University of New York at Buffalo, U.S.

4:00-4:20 Globalization in Nonlinear Feti-Dp
Oliver Rheinbach, Technische Universität Bergakademie Freiberg, Germany

4:25-4:45 Nonlinear Feti-Dp and Bddc Methods
Axel Klawonn, Universitaet zu Koeln, Germany

Wednesday, February 12

MS16

High Performance Krylov Subspace Methods: Theory, Implementation, and Application - Part II of III

3:10 p.m.-4:50 p.m.

Room: 512

For Part 1 see MS6
For Part 3 see MS26

Krylov subspace methods are widely used for solving very large (sparse) linear systems and eigenvalue problems. However, when these algorithms are implemented on parallel architectures in their original forms, communication (rather than floating point operations) becomes the dominant cost. To address this, many mathematically equivalent implementations, designed explicitly to reduce the time spent on communication, have been introduced. However on some problems these new formulations can behave (very) differently in finite precision than the original implementations. This makes the design of numerically stable variants for parallel architectures challenging. In this minisymposium we hear from researchers about the design and analysis of high performance Krylov subspace methods, practical concerns relating to implementing these methods on modern parallel architectures, and applications relevant today.

Organizer: Tyler Chen
University of Washington, U.S.

3:10-3:30 Predict-and-Recompute Conjugate Gradient Variants
Tyler Chen, University of Washington, U.S.

3:35-3:55 Improving Attainable Accuracy of the Deep Pipelined Conjugate Gradient Algorithm
Jeffrey Cornelis, Siegfried Cools, and Wim Vanroose, University of Antwerp, Belgium

4:00-4:20 Implementation and Performance Evaluation of High Performance Conjugate Gradients on Containers
Neena Imam, Oak Ridge National Laboratory, U.S.

4:25-4:45 Stable One-Reduce Gram Schmidt Orthogonalization Algorithms
Katarzyna Swirydowicz and Stephen Thomas, National Renewable Energy Laboratory, U.S.; Julien Langou and Daniel Bielich, University of Colorado, Denver, U.S.

continued in next column

Wednesday, February 12

MS17

HPC Simulation of the Hydrological Cycle

3:10 p.m.-4:50 p.m.

Room: 603

Large-scale computer simulations are increasingly important in addressing research and operational water resource issues related to the hydrological cycle. These issues include improving our understanding of complex interactions affecting streamflow, quantifying the effects of climate and landcover change, analyzing the frequency of extreme events, predicting flood impacts, and designing future infrastructure. Such simulations face significant challenges such as observational data scarcity and uncertainty, surface and sub-surface heterogeneity, scale disparity, and complex topography. In order to address these challenges, increasingly sophisticated numerical models are being developed and, along with them, increasingly sophisticated and computationally intense numerical algorithms and software. The advantages of such models include increased predictive capabilities, but these advantages can only be transformed into practical policies or disaster mitigation strategies through the use of scalable high-performance computing. This minisymposium offers four presentations on four different software packages for hydrological simulation, illustrating the successes and challenges for HPC in this increasingly important area. up.

Organizer: Raymond J. Spiteri
University of Saskatchewan, Canada

3:10-3:30 The Canadian Hydrological Model
Raymond J. Spiteri, Christopher B. Marsh, and Kevin R. Green, University of Saskatchewan, Canada

3:35-3:55 Parallelization of Computations Across Hierarchical River Networks
Martyn P. Clark, University of Saskatchewan, Canada

4:00-4:20 Interactive Computing at Scale: Applications in Climate and Hydrologic Modeling
Joseph J. Hamman, NCAR Earth System Laboratory, U.S.

continued in next column

4:25-4:45 Parameter Inference for a Massively Parallel Global Hydrologic Model
Luis Samaniego, Maren Kaluza, Stephan Thober, Rohini Kumar, Robert Schweppe, and Oldrich Rakovec, Helmholtz Centre for Environmental Research - UFZ, Germany

Wednesday, February 12

MS18

Exploiting Task Parallelism in Exascale Computing Era

3:10 p.m.-4:50 p.m.

Room: 605

As we approach the age of exascale computing, developers of large scale scientific applications will be facing new challenges in terms of performance and portability. In this regard, task parallel computation paradigm presents a promising way forward as it allows programmers to express a high degree of parallelism, utilize advanced scheduling techniques based on the data dependency information available, and obtain architectural portability relatively easily through runtime support. This minisymposium will feature talks on recent developments in task-parallel runtimes, middlewares and applications to explore the advantages of the task-parallel computation paradigm, as well as to identify the challenges ahead for a successful transition to the exascale era.

Organizer: H. Metin Aktulga
Michigan State University, U.S.

Organizer: Umit V. Catalyurek
Georgia Institute of Technology, U.S.

3:10-3:30 DeepSparse: A Task-Parallel Framework for Sparse Solvers on Deep Memory Architectures
Md Afibuzzaman and Fazlay Rabbi, Michigan State University, U.S.; M. Yusuf Ozkaya and Umit V. Catalyurek, Georgia Institute of Technology, U.S.; *H. Metin Aktulga*, Michigan State University, U.S.

3:35-3:55 Exploiting Task Parallelism in an Exascale Ecosystem: Some Issues and Possible Solutions
Olivier Aumage, Inria, France

4:00-4:20 Novel Approaches to Optimize and Execute Task-Based, Irregular Applications on Extreme-Scale, Heterogeneous Systems using PaRSEC
Thomas Herault, University of Tennessee, Knoxville, U.S.

4:25-4:45 Asynchronous Programming in Modern C++: What Is An AMT and Why Do You Want One for Christmas?
Hartmut Kaiser, Louisiana State University, U.S.

Wednesday, February 12

MS19

Parallel Adaptive Multigrid - Part II of II

3:10 p.m.-4:50 p.m.

Room: 606

For Part 1 see MS9

The multigrid method is well-known as an efficient and highly-scalable solver for many applications. Parallel adaptive mesh refinement nicely integrates with the geometric multigrid method and parallel algebraic multigrid methods are established as state-of-the-art parallel sparse matrix solvers. This session is dedicated to advances for the multigrid method with a focus on adaptive and parallel simulations. It offers a platform to present applications that allow for profitable employment of adaptive mesh refinement and implementations for high-performance computing.

Organizer: Gillian Queisser
Temple University, U.S.

Organizer: Andreas Vogel
Ruhr-University Bochum, Germany

Organizer: Gabriel Wittum
King Abdullah University of Science & Technology (KAUST), Saudi Arabia

3:10-3:30 Space-Time Multilevel Monte Carlo with Application to Cardiac Electrophysiology
Rolf Krause, Università della Svizzera italiana, Switzerland

3:35-3:55 Enriched Finite Volume Methods - Taylored Test Spaces for Interface Problems
Gabriel Wittum, King Abdullah University of Science & Technology (KAUST), Saudi Arabia

4:00-4:20 Simulation of Density-Driven Flow in Aquifers with Phreatic Surface
Dmitry Logashenko and Gabriel Wittum, King Abdullah University of Science & Technology (KAUST), Saudi Arabia

4:25-4:45 Homogenized Modeling of Microscopic Anisotropic Diffusion for Effective Diffusivities in Stratum Corneum
Junxi Wang, Goethe University Frankfurt, Germany

Wednesday, February 12

MS20

Frameworks/Libraries for High-Performance Tensor Computations - Part I of II

3:10 p.m.-4:50 p.m.

Room: 607

For Part 2 see MS31

Tensors are higher dimensional analogs of matrices, and represent a key data abstraction for many applications in computational science and data science. While robust high-performance libraries for matrix computations are widely available for a broad range of hardware platforms, the same has not been true for tensor computations. However, there has been considerable recent activity in developing libraries/frameworks for efficient tensor computations on different target platforms. This minisymposium will bring together researchers from academia, industry, and research laboratories to present recent advances, for dense as well as sparse tensors.

Organizer: Ponnuswamy Sadayappan
Ohio State University, U.S.

3:10-3:30 Model-Driven Tile Optimization for Tensor Contractions
Ponnuswamy Sadayappan, Ohio State University, U.S.; Rui Li, University of Utah, U.S.; Atanas Rountev, Ohio State University, U.S.; Aravind Sukumaran-Rajam, Washington State University, U.S.; Fabrice Rastello, Inria, France; Richard Veras, Louisiana State University, U.S.; Tze Meng Low, Carnegie Mellon University, U.S.

3:35-3:55 TAMM: Tensor Algebra for Many-Body Methods
Sriram Krishnamoorthy, Pacific Northwest National Laboratory, U.S.; Erdal Mutlu, ; Ajay Panyala, Pacific Northwest National Laboratory, U.S.

4:00-4:20 Cutensor: A High-Performance Cuda Library for Tensor Primitives
Paul Springer, NVIDIA, U.S.

4:25-4:45 A Hybrid Analytical/ Machine-Learning Model to Optimize Tensor-Contractions on GPUs
Aravind Sukumaran-Rajam, Washington State University, U.S.; Jinsung Kim, Ohio State University, U.S.

Wednesday, February 12

CP3

Application - Part I of III

3:10 p.m.-4:25 p.m.

Room: 604

For Part 2 see CP5

Chair: Bo Peng, Pacific Northwest National Laboratory, U.S.

3:10-3:30 Large-Scale DPG Finite Element Simulation of a Nonlinear Multiphysics Fiber Amplifier Model
Stefan Henneking, University of Texas at Austin, U.S.; Jacob Grosek, Air Force Research Laboratory, U.S.; Leszek Demkowicz, University of Texas at Austin, U.S.

3:35-3:55 Approaching Exa-Scalable and Accurate Green's Function Coupled Cluster Calculations of DNA Fragments
Bo Peng, Ajay Panyala, Sriram Krishnamoorthy, and Karol Kowalski, Pacific Northwest National Laboratory, U.S.

4:00-4:20 A Scalable Parallel Contact Algorithm Based on Dynamic Ghost Reconstruction for Lagrangian Hydrodynamic Application
Li Liao, Institute of Applied Physics and Computational Mathematics, China

Intermission

4:50 p.m.-5:00 p.m.

Welcome Remarks

5:00 p.m.-5:15 p.m.

Room: Elwha Ballroom

Wednesday, February 12

IP1

Parallel Tomographic Reconstruction - Where Combinatorics Meets Geometry

5:15 p.m.-6:00 p.m.

Room: Elwha Ballroom

Chair: Erik G. Boman, Sandia National Laboratories, U.S.

Today, high-resolution tomographic reconstruction of 3D objects is within reach, but the associated data sets are huge and calling for parallel computation. A typical 3D reconstruction with 4k resolution already produces an image of 64 Gbytes. Tomographic reconstruction is often done using iterative algorithms that involve repeated sparse matrix-vector multiplication (SpMV). The matrix, however, may be too large to store, requiring Tbytes of memory, and hence each matrix row is recomputed upon use. In this talk, we present data partitioning methods for tomography matrices of increasing size. For small matrices, we can compute an optimal bipartitioning by an exact combinatorial method, as implemented in the packages MondriaanOpt and MP. This allows us to gauge the quality of medium-grain partitioning (default in the Mondriaan package), which is a heuristic combinatorial method that can handle larger problems. Medium-grain results in turn justified choosing row partitioning for the tomographic matrix-free SpMV. For this row partitioning, we developed a geometric recursive coordinate bisection algorithm with nearly the same output quality as combinatorial partitioning that can handle huge, matrix-free problems and is also faster. We conclude with showing an actual reconstruction that was written using Bulk, a modern C++ library for easy development of parallel programs in bulk-synchronous parallel style.

Rob H. Bisseling

Utrecht University, The Netherlands

Wednesday, February 12

Welcome Reception

6:00 p.m.-8:00 p.m.

Room: Gallery - 5th Floor

BoF - NSF CSSI and IP Meeting

6:00 p.m.-7:00 p.m.

Room: 507

Thursday, February 13

Registration

8:00 a.m.-5:00 p.m.

Room: 503

SIAM Workshop on Combinatorial Scientific Computing (CSC20)

csc20.uni-jena.de/program.html

8:30 a.m.-4:45 p.m.

Room: Quinault Ballroom

Thursday, February 13

IP2

Accelerated-Node-Enabled Computational and Data Science: It?s not just for Exascale

8:30 a.m.-9:15 a.m.

Room: Elwha Ballroom

Chair: Olaf Schenk, Università della Svizzera italiana, Switzerland

In just the past decade, heterogeneous node-based computing hardware and software architectures have moved from novelty to mainstream. Prototypical examples are the now ubiquitous “accelerators”, or GPU-based architectures designed to accelerate certain operations on data arranged in certain ways. This heterogeneity will soon be more extreme, where computing nodes will not just have GPU accelerators, but other application-specific accelerators such as those for key computational motifs, AI, and even quantum-based optimization. GPU-based accelerators on supercomputers in the US Department of Energy (DOE) began with the Roadrunner and Titan systems, then on to the current Summit and Sierra and planned Perlmutter systems, and finally for the first three US exascale systems (Aurora, Frontier, El Capitan). Accelerated node hardware and software architectures are not only here and now but represent our collective future. The US scientific community must be able to effectively exploit these architectures to address urgent problems of National importance. A key response by the DOE to this “call to arms” was the launching of the Exascale Computing Project in 2016, which is an aggressive RD&D project focused on delivery of mission critical applications, an integrated software stack, exascale hardware technology advances and all within the context of an accelerated-node co-design software and algorithm paradigm that is not just for exascale but here to stay for the foreseeable future.

Douglas Kothe
Oak Ridge National Laboratory, U.S.

Intermission

9:15 a.m.-9:25 a.m.

Thursday, February 13

PD1

Is AI transforming HPC or HPC transforming AI?

9:25 a.m.-10:25 a.m.

Room: Elwha Ballroom

Chair: Aparna Chandramowlishwaran, University of California, Irvine, U.S.

AlexNet in 2012 started a new era in computing where AI (especially deep learning) started to find applications in every conceivable domain in science and engineering. This transformation was ignited by decades-long advances in high-performance computing and with the emergence of GPU’s which are a natural fit for training deep neural networks. However, since 2012, both HPC and AI have been driving one another. On the one hand, HPC is pushing the limits of AI model complexity and depth of the neural networks for increased model accuracy while simultaneously reducing training time. On the other hand, AI is transforming the way we do science. The quest for knowledge used to begin with grand theories and physics- based models form the core of our understanding of science. However, data-driven models are beginning to outperform physics-based models in specific tasks although several challenges such as model interpretability and generalization remain to be addressed before such models become commonplace. Looking into the future, are the paths of AI and HPC forever intertwined? Will the rapidly growing area of “black box” AI models make existing physics-based models obsolete? Is the increasing trend toward domain-specific architectures and infrastructure that favor AI applications, a stimulus for this HPC transformation?

Srinivas Aluru
Iowa State University, U.S.

Tamara G. Kolda
Sandia national Laboratories, U.S.

Dong Li
University of California, Merced, U.S.

Coffee Break

10:25 a.m.-10:55 a.m.

Room: Gallery - 5th Floor

Thursday, February 13

MS21

Advances in Algorithms Exploiting Low Precision Floating-Point Arithmetic - Part II of II

10:55 a.m.-12:35 p.m.

Room: Elwha Ballroom

For Part 1 see MS10

In the June 2019 Top 500 list, 133 systems are accelerator-based, more than half of which support half precision floating-point arithmetic. Since low precision floating-point formats are increasingly being supported by hardware vendors, developing algorithms and software that can exploit these formats is of increasing importance. This minisymposium will give a broad overview of recent contributions in exploiting low precision arithmetic in scientific computing, and details of applications in which it can successfully be used.

Organizer: Nicholas J. Higham
The University of Manchester, UK

Organizer: Srikara Pranesh
The University of Manchester, UK

10:55-11:15 Recent Half Precision Developments in the MAGMA Library
Ahmad Abdelfattah, University of Tennessee, Knoxville, U.S.

11:20-11:40 Solving Neural ODEs using Fixed-Point Arithmetic with Stochastic Rounding
Mantas Mikaitis, The University of Manchester, UK

11:45-12:05 Reduced Precision in Weather Forecasting Models
Andrew McRae, University of Oxford, United Kingdom

12:10-12:30 Mixed Precision Numerical Techniques Accelerated with Tensor Cores and its Impact on Today’s Scientific Computing and Implications for Tomorrow’s Hardware Design
Azzam Haidar, Harun Bayraktar, and Timothy Costa, NVIDIA, U.S.

Thursday, February 13

MS22

Parallel Processing for Particle Codes - Part I of II

10:55 a.m.-12:35 p.m.

Room: 502

For Part 2 see MS33

A large number of industrial problems can be modelled using particle-based methods. The particle method based application software is one of the main development areas in exascale computing including DOE Exascale Computing Projects(ECP). The reason is that particle based methods provide extremely fine-grained parallelism and allow the exploitation of asynchronous parallelism. Efficient parallel particle method applications require the efficient implementation when computing with billions and trillions of particles. The main objective of this proposal is to getting together experts to specifically address issues on the following common topics for large scale application simulations, scalable distributed computing (domain decomposition, dynamic load balancing), optimised data mapping (structured and unstructured communication), efficient parallel I/O, nearest neighbour lists searching using tree algorithms(particular for particle refinement) and cell linked lists, particle-to-mesh, and mesh-to-particle interpolation, sparse linear solver for incompressible problems, particle based applications involving complex geometries, parallel-in-time for particle based methods, best practice using machine learning and future-proof programming models for various modern and future computing architectures.

Organizer: Debasmita Samaddar
UK Atomic Energy Authority, United Kingdom

Organizer: Xiaohu Guo
Science and Technology Facilities Council, United Kingdom

Organizer: Philipp Neumann
Helmut Schmidt University, Germany

Organizer: Ivo F. Sbalzarini
TU Dresden, Germany and MPI Molecular Cell Biology & Genetics, Germany

Organizer: Mikito Furuichi
JAMSTEC, Japan

continued in next column

10:55-11:15 Particle Sorting for Projection Based Particle Methods
Xiaohu Guo, Science and Technology Facilities Council, United Kingdom

11:20-11:40 Multi-Architecture Parallel Particle Simulations on HPC Systems
Pietro Incardona, TU Dresden, Germany; Ivo F. Sbalzarini, TU Dresden, Germany and MPI Molecular Cell Biology & Genetics, Germany

11:45-12:05 Extending Quantum Molecular Dynamics to the Exascale: Latte, Progress and BML Libraries
Christian Negre, Los Alamos National Laboratory, U.S.

12:10-12:30 Optimizing Molecular Dynamics Simulations with Dynamic Auto-Tuning
Fabio A. Gratl, Steffen Seckler, and Hans-Joachim Bungartz, Technische Universität München, Germany; Philipp Neumann, Helmut Schmidt University, Germany

Thursday, February 13

MS23

Advances and Challenges in Solvers on GPGPU-based High-Performance Computing Architectures - Part I of II

10:55 a.m.-12:35 p.m.

Room: 505

For Part 2 see MS34

As the high-performance computing community pushes towards the exascale horizon, power and heat considerations have driven the increasing importance and prevalence of fine-grained parallelism in new computer architectures. This is particularly evident in the widespread adoption of general purpose graphics processing unit (GPGPU) accelerators at high-performance computing centers; reflecting this trend, the announced plans for the first generation of exascale-class systems to be fielded in the United States all rely on GPGPUs to provide the bulk of their computational power. This trend presents several challenges for scientific algorithms and software that were originally designed with scalar processor architectures in mind: large kernel launch and host-device transfer latencies, need for explicit application management of memory hierarchies, many different vectorization strategies and programming models to choose from, etc. This minisymposium will bring together researchers working on adapting and optimizing existing solvers for GPGPUs as well as developing novel approaches that map naturally onto emerging GPGPU computing resources.

Organizer: Richard T. Mills
Oak Ridge National Laboratory, U.S.

Organizer: Karl Rupp
Technische Universit?t Wien, Austria

Organizer: Jed Brown
University of Colorado Boulder, U.S.

Organizer: Hong Zhang
Argonne National Laboratory, U.S.

Organizer: Matthew G. Knepley
State University of New York at Buffalo, U.S.

continued on next page

10:55-11:15 Vendor-Optimized vs. Portable Performance: Approaches to Get Both
Karl Rupp, Technische Universität Wien, Austria

11:20-11:40 The Sparse Matrix Vector Product on High-End GPU Clusters
Hartwig Anzt, Karlsruhe Institute of Technology, Germany and University of Tennessee, U.S.; Terry Cojean and *Yuhsiang M. Tsai*, Karlsruhe Institute of Technology, Germany

11:45-12:05 Optimization of SpMV on GPU for Iterative Solvers in PETSc
Hong Zhang, Argonne National Laboratory, U.S.

12:10-12:30 Using the PETSc/TAO ADMM Methods on GPUs
Hansol David Suh, Georgia Institute of Technology, U.S.; Alp Dener, Argonne National Laboratory, U.S.; Tobin Isaac, Georgia Institute of Technology, U.S.; *Todd Munson*, Argonne National Laboratory, U.S.

Thursday, February 13

MS24

Parallel Matrix Factorization Algorithms - Part I of III

10:55 a.m.-12:35 p.m.

Room: 507

For Part 2 see MS35

This minisymposium brings together research on parallel algorithms for matrix factorizations for sparse, dense, and structured methods. Communication-avoidance has motivated new algorithms for dense matrix factorizations that are faster and more parallelizable. These techniques are being adapted also by new algorithms for parallel sparse direct solvers and preconditioning via approximate sparse matrix factorization. Low rank and hierarchically low rank matrix representations provide a competing route for finding solutions to numerical PDEs, but face their own scalability challenges. For example, low-rank factorizations motivate the development of efficient parallel pivoted QR factorization. The minisymposium will encompass state-of-the-art in parallel algorithms for factorization of dense, sparse, and structured matrices, and other related topics. In particular, speakers will discuss communication-avoiding parallel algorithms for dense QR, sparse QR, and QR with randomized column pivoting. Further, talks will describe innovations in state-of-the-art parallel libraries for sparse direct solvers and their application as solvers for numerical PDEs and dynamic optimization.

Organizer: Edgar Solomonik
University of Illinois at Urbana-Champaign, U.S.

10:55-11:15 A Communication-Avoiding Sparse Direct Solver for CPU+GPU Platforms
Piyush Sao, Georgia Institute of Technology, U.S.; Xiaoye S. Li, Lawrence Berkeley National Laboratory, U.S.; Ramakrishnan Kannan, Oak Ridge National Laboratory, U.S.; *Richard Vuduc*, Georgia Institute of Technology, U.S.

11:20-11:40 Communication-Avoiding Sparse Direct Solvers for Linear Systems & Graph Problems
Piyush Sao and Ramakrishnan Kannan, Oak Ridge National Laboratory, U.S.; Prasun Gera and Richard Vuduc, Georgia Institute of Technology, U.S.

11:45-12:05 Parallel Direct Matrix Factorization Methods for Dynamic Optimization
Samah Karim and Edgar Solomonik, University of Illinois at Urbana-Champaign, U.S.; Ryne Beeson, CU Aerospace, U.S.

12:10-12:30 Efficient Sparse Triangular Solve
Sebastien Cayrols and Florent Lopez, University of Tennessee, Knoxville, U.S.; Iain Duff, Science & Technology Facilities Council, United Kingdom and CERFACS, Toulouse, France

continued in next column

Thursday, February 13

MS25

Progress and Challenges in Extreme Scale Computing and Big Data - Part I of II

10:55 a.m.-12:35 p.m.

Room: 508

For Part 2 see MS36

Extreme scale computing efforts have resulted in numerous advances for multicore, manycore and accelerator based scalable systems. In addition, large-scale applications must increasingly deal with data management and analysis as a first-class concern. Therefore, new applications often have to manage distributed and parallel computing, and have to manage workflows of different tasks (computing, data analytics, machine learning, visualization,...). In this MS, we present some of the latest works in scalable algorithms, programming paradigms, and libraries for next generation computing platforms. Furthermore, we discuss efforts to better incorporate data science concerns as a principle component of our scientific workflows.

Organizer: Kengo Nakajima
University of Tokyo, Japan

Organizer: Michael A. Heroux
Sandia National Laboratories, U.S.

Organizer: Serge Petiton
Universite de Lille 1, France

10:55-11:15 Exascale Node-Level Parallel Programming Environments: Overview and Deciding What's Right for You

Michael A. Heroux, Sandia National Laboratories, U.S.

11:20-11:40 On Performance Portability - Application and Illustration on MC Neutron Transport Application

Christophe Calvin, CEA Saclay, France; Emeric Brun and Tao Chang, CEA, DEN, SRMP, France

11:45-12:05 A Taxonomy of Distributed and Parallel Languages for High Performance Tasks-Based Multilevel Computing

Serge Petiton, Universite de Lille 1, France; Jerome Gurhem, Université Lille 1 and CNRS, France; Henri Calandra, Total, France

12:10-12:30 AMR Framework for Large-Scale Simulations on Multiple GPUs

Takashi Shimokawabe, University of Tokyo, Japan; Naoyuki Onodera, Japan Atomic Energy Agency, Japan

continued in next column

Thursday, February 13

MS26

High Performance Krylov Subspace Methods: Theory, Implementation, and Application - Part III of III

10:55 a.m.-12:35 p.m.

Room: 512

For Part 2 see MS16

Krylov subspace methods are widely used for solving very large (sparse) linear systems and eigenvalue problems. However, when these algorithms are implemented on parallel architectures in their original forms, communication (rather than floating point operations) becomes the dominant cost. To address this, many mathematically equivalent implementations, designed explicitly to reduce the time spent on communication, have been introduced. However on some problems these new formulations can behave (very) differently in finite precision than the original implementations. This makes the design of numerically stable variants for parallel architectures challenging. In this minisymposium we hear from researchers about the design and analysis of high performance Krylov subspace methods, practical concerns relating to implementing these methods on modern parallel architectures, and applications relevant today.

Organizer: Tyler Chen
University of Washington, U.S.

10:55-11:15 The Lanczos Method in Data Science: New Challenges and the Continued Importance of Stability

Christopher Musco, New York University, U.S.

11:20-11:40 Wasserstein Discriminant Analysis with Eigensolver

Hexuan Liu, University of Washington, U.S.

11:45-12:05 Polynomial Preconditioned GMRES in Trilinos: Practical Considerations for High-Performance Computing[1]

Jennifer A. Loe, Baylor University, U.S.; Heidi K. Thornquist and Erik G. Boman, Sandia National Laboratories, U.S.

[1]This presentation is included in the proceedings

continued on next page

12:10-12:30 Improving the Performance of GMRES with Mixed Precision
Neil Lindquist, University of Tennessee, U.S.

Thursday, February 13

MS27

Challenges in Parallel Adaptive Mesh Refinement - Part I of III

10:55 a.m.-12:35 p.m.

Room: 603

For Part 2 see MS37

Parallel adaptive mesh refinement (AMR) is a key technique when simulations are required to capture time-dependent and/or multiscale features. Frequent re-adaptation and repartitioning of the mesh during the simulation can impose significant overhead, particularly in largescale parallel environments. Further challenges arise due to the availability of accelerated or special-purpose hardware, and the trend toward hierarchical and hybrid compute architectures. Our minisymposium addresses algorithms, scalability, and software issues of parallel AMR.

Organizer: Michael Bader
Technische Universität München, Germany

Organizer: Martin Berzins
University of Utah, U.S.

Organizer: Carsten Burstedde
Universitaet Bonn, Germany

10:55-11:15 {AMR} in Core-Collapse Supernova Simulations
Anshu Dubey, Argonne National Laboratory, U.S.; J. Austin Harris, Oak Ridge National Laboratory, U.S.; Bronson Messer, Oak Ridge National Laboratory & University of Tennessee, U.S.

11:20-11:40 Asynchronous Task-Based {AMR} with Distributed Parallel Objects
James Bordner and Michael L. Norman, University of California, San Diego, U.S.

11:45-12:05 Scalable Space-Time Adaptivity for Simulations of Binary Black Hole Intermediate-Mass-Ratio-Inspirals
Hari Sundar, University of Utah, U.S.

12:10-12:30 Direct Parallel Visualization Using Forest-of-Octrees Meshes
Carsten Burstedde, Universitaet Bonn, Germany

Thursday, February 13

MS28

High-Performance Numerics and Model Development for Geophysical Systems - Part I of II

10:55 a.m.-12:10 p.m.

Room: 604

For Part 2 see MS38

There is a current explosion of interest in new numerical methods for the simulation of complex geophysical systems (atmosphere, ocean, land, ice) on next-generation supercomputers. To generate accurate predictions, models must be able to efficiently resolve multi-scale processes, surpass computational barriers for higher-resolution simulations, utilize scalable methods for emerging architectures, and provide actionable predictions driven by localized dynamics. To accomplish this, models require state-of-the-art numerical algorithms which can exploit current trends in supercomputing design. This minisymposium highlights some of the development challenges, with application areas relevant for large-scale atmospheric forecasting, ocean circulation, storm-surge modeling, and glaciology. Emphasis is placed on efficient solver algorithms, software frameworks, and application-driven model development.

Organizer: Thomas H. Gibson
Imperial College London, United Kingdom

Organizer: David Ham
Imperial College London, United Kingdom

Organizer: Christopher Eldred
CNRS and Grenoble University, France

10:55-11:15 NUMO: A Non-Hydrostatic Unified Model of the Ocean
Michal A. Kopera, Boise State University, U.S.

11:20-11:40 Recent Developments in Hybridisation Techniques for Finite Element Problems in Numerical Weather Prediction
Jack Betteridge, University of Bath, United Kingdom

continued on next page

Thursday, February 13

MS28

High-Performance Numerics and Model Development for Geophysical Systems - Part I of II

10:55 a.m.-12:10 p.m.

continued

11:45-12:05 Entropically Consistent Models of Geophysical Fluid Dynamics

Christopher Eldred, CNRS and Grenoble University, France; Thomas Dubos, Ecole Polytechnique, France; Francois Gay-Balmaz, Ecole Normale Superieure, France

Thursday, February 13

MS29

Recent Advances and Trends in Hybrid Quantum-Classical Algorithms - Part I of II

10:55 a.m.-12:35 p.m.

Room: 605

For Part 2 see MS39

Noisy Intermediate-Scale Quantum (NISQ) devices that are becoming available on the clouds are limited to between tens and hundreds of qubits. Recent experimental demonstrations suggest that in the next few years we will witness qubits with noise levels sufficiently low to solve computationally hard problems. Combinatorial optimization problems are considered one of the main candidates for demonstrating quantum advantage, i.e. solving a given problem faster or finding a better solution that classical state-of-the-art solvers. However, a limited number of qubits (and, thus, a limited number of corresponding optimization variables) presents a major obstacle driving the design of the novel algorithms efficiently using NISQ devices. Combining classical and quantum resources is one of the expedient answers that researchers suggest today to tackle real-life problems with existing quantum hardware. These hybrid algorithms attempt to take advantage of ?the best of both worlds?, leveraging the power of quantum computation while using a classical machine to address the limitations of NISQ devices. In this mini-symposium, we will present state of the art methods and ideas related to hybrid quantum-classical algorithms, demonstrate computational results, focus on combining NISQ devices and HPC, and discuss problems and obstacles related to these approaches.

Organizer: Ilya Safro
Clemson University, U.S.

Organizer: Yuri Alexeev
Argonne National Laboratory, U.S.

Organizer: Ruslan Shaydulin
Clemson University, U.S.

10:55-11:15 Hybrid Quantum-Classical Approaches to Exact and Approximate Optimization

Stuart Hadfield, NASA, U.S.

continued in next column

11:20-11:40 Scientific Computing Benchmarks for Quantum Computers

Travis Humble, Dmitry Liakh, and Alexander McCaskey, Oak Ridge National Laboratory, U.S.

11:45-12:05 Overview of Hybrid Quantum/Classical Methods for Quantum Annealing

Catherine McGeoch, D-Wave Systems, Inc., U.S.

12:10-12:30 Training Quantum Boltzmann Machines on Near-Term Quantum Devices

Nathan O. Wiebe, Microsoft Research, U.S.

Thursday, February 13

MS30

Accelerating Data Sparse Applications on Massively Parallel Systems - Part I of III

10:55 a.m.-12:35 p.m.

Room: 606

For Part 2 see MS40

With multiple monikers such as block low-rank, data sparse, or hierarchical matrices, this minisymposium features presentations on exploiting structure of matrix data for more efficient solvers in science and engineering. This low-rank approximation technique exploits the data sparsity of the application system matrix by compressing matrix entries. Often, the off-diagonal submatrices are amenable to compression and may be regulated with a user-defined accuracy ranges. During the actual solver phase, the underlying linear algebra operations are carried out on the compressed data format rather than the original full data. As a result, the arithmetic complexity of the new solver ends up being reduced to much lower levels. This allows to tackle much larger simulation problems that would otherwise be constrained by either their storage size of computational cost. The wide variety of solvers and compression techniques will be presented to showcase to broad applicability of the data-sparse techniques in modern scientific applications.

Organizer: Hatem Ltaief
King Abdullah University of Science & Technology (KAUST), Saudi Arabia

Organizer: Piotr Luszczek
University of Tennessee, Knoxville, U.S.

10:55-11:15 Extreme-Scale Task-Based Cholesky Factorization Toward Climate and Weather Prediction Applications
Qinglei Cao, University of Tennessee, Knoxville, U.S.; Yu Pei, University of Tennessee, U.S.; Kadir Akbudak and Aleksandr Mikhalev, King Abdullah University of Science & Technology (KAUST), Saudi Arabia; George Bosilca, University of Tennessee, Knoxville, U.S.; *Hatem Ltaief*, King Abdullah University of Science & Technology (KAUST), Saudi Arabia; Jack J. Dongarra, University of Tennessee and Oak Ridge National Laboratory, U.S.; David E. Keyes, King Abdullah University of Science & Technology (KAUST), Saudi Arabia

11:20-11:40 Evaluation of Programming Models to Address Load Imbalance on Distributed Multi-Core CPUs: A Case Study with Block Low-Rank Factorization
Yu Pei, University of Tennessee, U.S.; Ichitaro Yamazaki, Sandia National Laboratories, U.S.; George Bosilca, University of Tennessee, Knoxville, U.S.; Akihiro Ida, Kyoto University, Japan; Jack J. Dongarra, University of Tennessee and Oak Ridge National Laboratory, U.S.

11:45-12:05 Clustering Techniques and Hierarchical Matrix Formats for Scalable Kernel Ridge Regression
Xiaoye S. Li, Lawrence Berkeley National Laboratory, U.S.; Rebrova Elizaveta, University of California, Los Angeles, U.S.; Gustavo Chavez, Pieter Ghysels, and Yang Liu, Lawrence Berkeley National Laboratory, U.S.

12:10-12:30 Decoupling Structure Analysis in Hierarchical Matrix Approximations
Maryam Mehri Dehnavi, Kazem Cheshmi, and Bangtian Liu, University of Toronto, Canada

continued in next column

Thursday, February 13

MS31

Frameworks/Libraries for High-Performance Tensor Computations - Part II of II

10:55 a.m.-12:35 p.m.

Room: 607

For Part 1 see MS20

Tensors are higher dimensional analogs of matrices, and represent a key data abstraction for many applications in computational science and data science. While robust high-performance libraries for matrix computations are widely available for a broad range of hardware platforms, the same has not been true for tensor computations. However, there has been considerable recent activity in developing libraries/frameworks for efficient tensor computations on different target platforms. This minisymposium will bring together researchers from academia, industry, and research laboratories to present recent advances, for dense as well as sparse tensors.

Organizer: Ponnuswamy Sadayappan
Ohio State University, U.S.

10:55-11:15 ENSIGN: A Framework for High-performance Tensor Decompositions
Muthu M. Baskaran, Reservoir Labs, U.S.

11:20-11:40 The Design and Implementation of Large-Scale Tensor Decomposition in SPLATT
Shaden Smith, Intel, AI, U.S.

11:45-12:05 Structured Abstractions in MLIR: High-Level Infrastructure for Optimizing Matrix/Tensor Computations
Nicolas Vasilache and *Manesh Ravishankar*, Google Brain, U.S.

12:10-12:30 The Sparse Tensor Algebra Compiler
Fredrik Kjolstad, Stanford University, U.S.; Stephen Chou, Ryan Senanayake, and Peter Ahrens, Massachusetts Institute of Technology, U.S.

Thursday, February 13

CP4

Proceedings Papers - Part I of III

10:55 a.m.-12:10 p.m.

Room: 501

For Part 2 see CP7

Chair: Judith Hill, Oak Ridge National Laboratory, U.S.

10:55-11:15 General Memory-Independent Lower Bound for MTTKRP[2]

Grey Ballard, Wake Forest University, U.S.; *Kathryn Rouse*, Inmar, USA

11:20-11:40 Tensor Processing Units for Financial Monte Carlo[3]

Francois Belletti, Davis King, Kun Yang, Roland Nelet, Yusef Shafi, Yi-Fan Chen, and John Anderson, Google Research, U.S.

11:45-12:05 Fast Image Reconstruction at a Synchrotron Laboratory[4]

Eduardo Miqueles, Gilberto Martinez, Jr., and Patricio Guerrero, LNLS/CNPEM, Brazil

Lunch Break

12:35 p.m.-2:05 p.m.

Attendees on their own

[2] This presentation is included in the proceedings
[3] This presentation is included in the proceedings
[4] This presentation is included in the proceedings

Thursday, February 13

IP3

Development of an Eigen-Analysis Engine for Large-Scale Simulation and Big Data Analysis

2:05 p.m.-2:50 p.m.

Room: Elwha Ballroom

Chair: Axel Klawonn, Universitaet zu Koeln, Germany

Large-scale eigenvalue problems arise in a wide variety of scientific and engineering applications such as nano-scale materials simulation, vibration analysis of automobile components, data analysis, graph analysis, etc. In such computations, high-performance parallel eigensolvers are required to exploit distributed parallel computing environments. In this talk, we present a parallel eigensolver, the Sakurai-Sugiura method (SSM), for large-scale interior eigenvalue problems. This method is derived using numerical quadrature and has good parallel scalability. We also show a software package "z-Pares," which enables users to utilize a large number of computational resources because of its hierarchical parallel structure. While the presented technology is already well-established, its applications are yet to be fully explored. Hence this talk will also give an overview of challenging issues on the intersection of Big Data analysis and simulation that could be tackled with a scalable eigensolver.

Tetsuya Sakurai
University of Tsukuba, Japan

Coffee Break

2:50 p.m.-3:20 p.m.

Room: Gallery - 5th Floor

Thursday, February 13

MS32

Transparency, Reproducibility, Sustainability, and Security: The Four Pillars of the Next Generation Scientific Software Stack

3:20 p.m.-5:00 p.m.

Room: Elwha Ballroom

The software stacks used in computationally- and data-enabled research and discovery is increasingly dependent on transparency, reproducibility, sustainability, and security. Transparency exposes the salient computational aspects of the discovery workflow and permits reviewers and readers of published research findings to understand and assess the computational result. Reproducibility refers to the regeneration of computationally- and data-enabled findings on the same system or a different system. Sustainability goes to the ability to persistently reuse software. Open science needs assurance of computational systems that behave as intended, including data integrity, defending against assertions of data manipulation in an increasing polarized world, and support for work that is often extremely collaborative. To trust the scientific claims that result from these systems, we not only need to trust the scientific methodology but also trust the behavior of the underlying computational system itself, in other words the implementation in software and the software environment itself. Today, few guarantees are given regarding the computational infrastructure that supports scientific discovery and even fewer guarantees are given regarding security for openly shared data and codes that support computational reproducibility in the scientific realm. Researchers are increasingly sharing and reusing data and code, with very little guidance on appropriate security guarantees or standards

Organizer: Michela Taufer
University of Tennessee, U.S.

3:20-3:40 Transparency and Reproducibility: Case Studies, Formalisms, and Structured Guidance in Scientific Applications at Scale

Michela Taufer, University of Tennessee, U.S.

continued on next page

3:45-4:05 Transparency and Reproducibility: Case Studies, Formalisms, and Structured Guidance in Computational Social Science Applications
Victoria Stodden, University of Illinois at Urbana-Champaign, U.S.

4:10-4:30 Next Generation Cyberinfrastructure for Science: Cyberinfrastructure Center of Excellence Pilot for Large Facilities
Ewa Deelman, University of Southern California, U.S.

4:35-4:55 The Mission of Cybersecurity in Science: Productivity, Reproducibility, and Trust
Von Welch, Indiana University, U.S.

Thursday, February 13

MS33

Parallel Processing for Particle Codes - Part II of II

3:20 p.m.-5:00 p.m.

Room: 502

For Part 1 see MS22

A large number of industrial problems can be modelled using particle-based methods. The particle method based application software is one of the main development areas in exascale computing including DOE Exascale Computing Projects(ECP). The reason is that particle based methods provide extremely fine-grained parallelism and allow the exploitation of asynchronous parallelism. Efficient parallel particle method applications require the efficient implementation when computing with billions and trillions of particles. The main objective of this proposal is to getting together experts to specifically address issues on the following common topics for large scale application simulations, scalable distributed computing (domain decomposition, dynamic load balancing), optimised data mapping (structured and unstructured communication), efficient parallel I/O, nearest neighbour lists searching using tree algorithms(particular for particle refinement) and cell linked lists, particle-to-mesh, and mesh-to-particle interpolation, sparse linear solver for incompressible problems, particle based applications involving complex geometries, parallel-in-time for particle based methods, best practice using machine learning and future-proof programming models for various modern and future computing architectures.

Organizer: Debasmita Samaddar
UK Atomic Energy Authority, United Kingdom

Organizer: Xiaohu Guo
Science and Technology Facilities Council, United Kingdom

Organizer: Philipp Neumann
Helmut Schmidt University, Germany

Organizer: Ivo F. Sbalzarini
TU Dresden, Germany and MPI Molecular Cell Biology & Genetics, Germany

Organizer: Mikito Furuichi
JAMSTEC, Japan

3:20-3:40 Massively Parallel DEM Simulation and Stress Chain Characterization with over Billion Particles
Mikito Furuichi, JAMSTEC, Japan; Jian Chen, Natsuki Hosono, and Daisuke Nishiura, Japan Agency for Marine-Earth Science and Technology, Japan

3:45-4:05 Computing Particle Trajectories with Spectral Deferred Corrections
Krasymyr Tretiak, Kris Smedt, Jitse Niesen, and Steve Tobias, University of Leeds, United Kingdom; *Daniel Ruprecht*, Hamburg University of Technology, Germany

4:10-4:30 Parallel-in-Time and Aymptotic-Preserving Monte Carlo Methods for Particle-Based Systems in a Diffusive Scaling
Bert Mortier and Giovanni Samaey, KU Leuven, Belgium

4:35-4:55 Parallelisation of the SMARDDA-PFC Software
Huw Leggate, Dublin City University, Ireland; *Wayne Arter*, UK Atomic Energy Authority, United Kingdom

continued in next column

Thursday, February 13

MS34

Advances and Challenges in Solvers on GPGPU-Based High-Performance Computing Architectures - Part II of II

3:20 p.m.-5:00 p.m.

Room: 505

For Part 1 see MS23

As the high-performance computing community pushes towards the exascale horizon, power and heat considerations have driven the increasing importance and prevalence of fine-grained parallelism in new computer architectures. This is particularly evident in the widespread adoption of general purpose graphics processing unit (GPGPU) accelerators at high-performance computing centers; reflecting this trend, the announced plans for the first generation of exascale-class systems to be fielded in the United States all rely on GPGPUs to provide the bulk of their computational power. This trend presents several challenges for scientific algorithms and software that were originally designed with scalar processor architectures in mind: large kernel launch and host-device transfer latencies, need for explicit application management of memory hierarchies, many different vectorization strategies and programming models to choose from, etc. This minisymposium will bring together researchers working on adapting and optimizing existing solvers for GPGPUs as well as developing novel approaches that map naturally onto emerging GPGPU computing resources.

Organizer: Richard T. Mills
Oak Ridge National Laboratory, U.S.

Organizer: Karl Rupp
Technische Universität Wien, Austria

Organizer: Jed Brown
University of Colorado Boulder, U.S.

Organizer: Hong Zhang
Argonne National Laboratory, U.S.

Organizer: Matthew G. Knepley
State University of New York at Buffalo, U.S.

3:20-3:40 Current State and Future Goals for Multigrid-Preconditioned Linear Solvers on GPU-Based Supercomputers

Daniel Ibanez, Sandia National Laboratories, U.S.

3:45-4:05 MueLu's Algorithmic Performance on GPU

Luc Berger-Vergiat, Ray S. Tuminaro, Jonathan J. Hu, Chris Siefert, and Christian Glusa, Sandia National Laboratories, U.S.

4:10-4:30 Geometric and Algebraic Multigrid Solvers in {PETSc} on Many-{GPU} Supercomputer Architectures

Richard T. Mills, Oak Ridge National Laboratory, U.S.; Mark Adams, Lawrence Berkeley National Laboratory, U.S.; Hannah M. Morgan, Argonne National Laboratory, U.S.; Karl Rupp, Technische Universität Wien, Austria; Barry F. Smith, Argonne National Laboratory, U.S.

4:35-4:55 Towards Efficient Communication on Heterogeneous Architectures

Amanda Bienz, Luke Olson, and William D. Gropp, University of Illinois at Urbana-Champaign, U.S.

continued in next column

Thursday, February 13

MS35

Parallel Matrix Factorization Algorithms - Part II of III

3:20 p.m.-5:00 p.m.

Room: 507

For Part 1 see MS24
For Part 3 see MS44

This minisymposium brings together research on parallel algorithms for matrix factorizations for sparse, dense, and structured methods. Communication-avoidance has motivated new algorithms for dense matrix factorizations that are faster and more parallelizable. These techniques are being adapted also by new algorithms for parallel sparse direct solvers and preconditioning via approximate sparse matrix factorization. Low rank and hierarchically low rank matrix representations provide a competing route for finding solutions to numerical PDEs, but face their own scalability challenges. For example, low-rank factorizations motivate the development of efficient parallel pivoted QR factorization. The minisymposium will encompass state-of-the-art in parallel algorithms for factorization of dense, sparse, and structured matrices, and other related topics. In particular, speakers will discuss communication-avoiding parallel algorithms for dense QR, sparse QR, and QR with randomized column pivoting. Further, talks will describe innovations in state-of-the-art parallel libraries for sparse direct solvers and their application as solvers for numerical PDEs and dynamic optimization.

Organizer: Edgar Solomonik
University of Illinois at Urbana-Champaign, U.S.

3:20-3:40 Shifted CholeskyQR3 for High Performance Tall-Skinny QR Factorization

Takeshi Fukaya, Hokkaido University, Japan; Ramaseshan Kannan, Arup, Manchester, United Kingdom; Yuji Nakatsukasa, University of Oxford, United Kingdom; Yusaku Yamamoto, University of Electro-Communications, Japan; Yuka Yanagisawa, Waseda University, Japan

continued on next page

3:45-4:05 On Leveraging Communication-Optimal QR Factorization in Dense Symmetric Eigensolvers
Edward Hutter and Edgar Solomonik, University of Illinois at Urbana-Champaign, U.S.

4:10-4:30 A Matlab Package for Superfast Divide-and-Conquer Hermitian Eigenvalue Decompositions
Jianlin Xia and Xiaofeng Ou, Purdue University, U.S.

4:35-4:55 Parallel Butterfly-Based Sherman-Morrison-Woodbury Inversion
Yang Liu, Ghysels Pieter, and Xiaoye S. Li, Lawrence Berkeley National Laboratory, U.S.

Thursday, February 13

MS36

Progress and Challenges in Extreme Scale Computing and Big Data - Part II of II

3:20 p.m.-5:00 p.m.

Room: 508

For Part 1 see MS25

Extreme scale computing efforts have resulted in numerous advances for multicore, manycore and accelerator based scalable systems. In addition, large-scale applications must increasingly deal with data management and analysis as a first-class concern. Therefore, new applications often have to manage distributed and parallel computing, and have to manage workflows of different tasks (computing, data analytics, machine learning, visualization,?). In this MS, we present some of the latest works in scalable algorithms, programming paradigms, and libraries for next generation computing platforms. Furthermore, we discuss efforts to better incorporate data science concerns as a principle component of our scientific workflows.

Organizer: Kengo Nakajima
University of Tokyo, Japan

Organizer: Michael A. Heroux
Sandia National Laboratories, U.S.

Organizer: Serge Petiton
Universite de Lille 1, France

3:20-3:40 Innovative Methods for Scientific Computing in the Exascale Era by Integrations of (Simulation + Data + Learning)
Kengo Nakajima, University of Tokyo, Japan

3:45-4:05 Composition in Scientific and Engineering Applications: Lessons Learned from the Legion Programming System
Galen Shipman, Los Alamos National Laboratory, U.S.

4:10-4:30 Application Developers? Experiences with RAJA
David Beckingsale, Lawrence Livermore National Laboratory, U.S.

4:35-4:55 Toward AI-based Medical Data Analytics and Clinical Workflows
Weichung Wang, National Taiwan University, Taiwan

Thursday, February 13

MS37

Challenges in Parallel Adaptive Mesh Refinement - Part II of III

3:20 p.m.-5:00 p.m.

Room: 603

For Part 1 see MS27
For Part 3 see MS46

Parallel adaptive mesh refinement (AMR) is a key technique when simulations are required to capture time-dependent and/or multiscale features. Frequent re-adaptation and repartitioning of the mesh during the simulation can impose significant overhead, particularly in largescale parallel environments. Further challenges arise due to the availability of accelerated or special-purpose hardware, and the trend toward hierarchical and hybrid compute architectures. Our minisymposium addresses algorithms, scalability, and software issues of parallel AMR.

Organizer: Michael Bader
Technische Universität München, Germany

Organizer: Martin Berzins
University of Utah, U.S.

Organizer: Carsten Burstedde
Universitaet Bonn, Germany

3:20-3:40 Event-Driven, Stream-Based Amr Software
Tobias Weinzierl, Durham University, United Kingdom

3:45-4:05 An Evaluation of Asynchronous Task Execution Strategies in AMT AMR Approaches
Martin Berzins and Alan Humphrey, University of Utah, U.S.

4:10-4:30 Tasks Unlimited : Lightweight Task Offloading and Replication for Parallel Adaptive Mesh Refinement
Philipp Samfass, Technical University of Munich, Germany

4:35-4:55 On the Theory of Discrete, Adaptive Space Filling Curves
Johannes Holke, German Aerospace Center (DLR), Germany; Carsten Burstedde and David Knapp, Universitaet Bonn, Germany

Thursday, February 13

MS38

High-Performance Numerics and Model Development for Geophysical Systems - Part II of II

3:20 p.m.-5:00 p.m.

Room: 604

For Part 1 see MS28

There is a current explosion of interest in new numerical methods for the simulation of complex geophysical systems (atmosphere, ocean, land, ice) on next-generation supercomputers. To generate accurate predictions, models must be able to efficiently resolve multi-scale processes, surpass computational barriers for higher-resolution simulations, utilize scalable methods for emerging architectures, and provide actionable predictions driven by localized dynamics. To accomplish this, models require state-of-the-art numerical algorithms which can exploit current trends in supercomputing design. This minisymposium highlights some of the development challenges, with application areas relevant for large-scale atmospheric forecasting, ocean circulation, storm-surge modeling, and glaciology. Emphasis is placed on efficient solver algorithms, software frameworks, and application-driven model development.

Organizer: Thomas H. Gibson
Imperial College London, United Kingdom

Organizer: David Ham
Imperial College London, United Kingdom

Organizer: Christopher Eldred
CNRS and Grenoble University, France

3:20-3:40 CLIMA-atmos: a Non-Hydrostatic Model of the Atmosphere for Next-Generation Super-Computing Systems

Thomas H. Gibson, Imperial College London, United Kingdom; Maciej Waruszewski, Naval Postgraduate School, U.S.; Simon Bryne, California Institute of Technology, U.S.; Jeremy E. Kozdon, Francis X. Giraldo, and Lucas Wilcox, Naval Postgraduate School, U.S.; Valentin Churavy, Massachusetts Institute of Technology, U.S.

continued in next column

3:45-4:05 A Discontinuous Galerkin Method for Idealised Hurricane Storm Surge

Nicole Beisiegel, University College Dublin, Ireland; Jörn Behrens, Universit?t Hamburg, Germany; Cristóbal E. Castro, Universidad de Tarapaca, Chile

4:10-4:30 A Higher-Order Model for Glacier Flow via Spectral Semi-Discretization

Daniel Shapero, University of Washington, U.S.

4:35-4:55 A System for Programming Symbolic Manipulations Finite Element Problems in UFL: FML

David Ham, Imperial College London, United Kingdom; Jemma Shipton, University of Exeter, United Kingdom

Thursday, February 13

MS39

Recent Advances and Trends in Hybrid Quantum-Classical Algorithms - Part II of II

3:20 p.m.-5:00 p.m.

Room: 605

For Part 1 see MS29

Noisy Intermediate-Scale Quantum (NISQ) devices that are becoming available on the clouds are limited to between tens and hundreds of qubits. Recent experimental demonstrations suggest that in the next few years we will witness qubits with noise levels sufficiently low to solve computationally hard problems. Combinatorial optimization problems are considered one of the main candidates for demonstrating quantum advantage, i.e. solving a given problem faster or finding a better solution that classical state-of-the-art solvers. However, a limited number of qubits (and, thus, a limited number of corresponding optimization variables) presents a major obstacle driving the design of the novel algorithms efficiently using NISQ devices. Combining classical and quantum resources is one of the expedient answers that researchers suggest today to tackle real-life problems with existing quantum hardware. These hybrid algorithms attempt to take advantage of "the best of both worlds", leveraging the power of quantum computation while using a classical machine to address the limitations of NISQ devices. In this mini-symposium, we will present state of the art methods and ideas related to hybrid quantum-classical algorithms, demonstrate computational results, focus on combining NISQ devices and HPC, and discuss problems and obstacles related to these approaches.

Organizer: Ilya Safro
Clemson University, U.S.

Organizer: Yuri Alexeev
Argonne National Laboratory, U.S.

Organizer: Ruslan Shaydulin
Clemson University, U.S.

3:20-3:40 What Do QAOA Energies Reveal About Graphs?

Mario Szegedy, Alibaba Group, U.S.

continued on next page

3:45-4:05 Multilevel Hybrid Quantum-Classical Algorithms on Graphs
Ruslan Shaydulin, Clemson University, U.S.

4:10-4:30 Quantum Approximate Optimization with a Trapped-Ion Quantum Simulator
Guido Pagano, University of Maryland, U.S.

4:35-4:55 Fast quantum subroutines for the simplex method
Giacomo Nannicini, IBM T.J. Watson Research Center, U.S.

Thursday, February 13

MS40

Accelerating Data Sparse Applications on Massively Parallel Systems - Part II of III

3:20 p.m.-5:00 p.m.

Room: 606

For Part 1 see MS30
For Part 3 see MS48

With multiple monikers such as block low-rank, data sparse, or hierarchical matrices, this minisymposium features presentations on exploiting structure of matrix data for more efficient solvers in science and engineering. This low-rank approximation technique exploits the data sparsity of the application system matrix by compressing matrix entries. Often, the off-diagonal submatrices are amenable to compression and may be regulated with a user-defined accuracy ranges. During the actual solver phase, the underlying linear algebra operations are carried out on the compressed data format rather than the original full data. As a result, the arithmetic complexity of the new solver ends up being reduced to much lower levels. This allows to tackle much larger simulation problems that would otherwise be constrained by either their storage size of computational cost. The wide variety of solvers and compression techniques will be presented to showcase to broad applicability of the data-sparse techniques in modern scientific applications.

Organizer: Hatem Ltaief
King Abdullah University of Science & Technology (KAUST), Saudi Arabia

Organizer: Piotr Luszczek
University of Tennessee, Knoxville, U.S.

3:20-3:40 Semi-Automatic DAG Generation for H-Arithmetic
Ronald Kriemann, Max Planck Institute for Mathematics in the Sciences, Germany; Steffen Börm and Sven Christophersen, University of Kiel, Germany

3:45-4:05 Fast Algorithms for Hierarchical Matrices
George Biros, University of Texas at Austin, U.S.

4:10-4:30 Hierarchical Matrix Algorithms on Manycore Architectures
George M. Turkiyyah, American University of Beirut, Lebanon; Wajih Halim Boukaram and David E. Keyes, King Abdullah University of Science & Technology (KAUST), Saudi Arabia

4:35-4:55 Solving Acoustic Boundary Integral Equations using High Performance Tile Low-Rank LU Factorization
Rabab Omairy, Noha Harthi, Kadir Akbudak, Rui Chen, Hatem Itaief, Hakan Bagci, and David E. Keyes, King Abdullah University of Science & Technology (KAUST), Saudi Arabia

continued in next column

Thursday, February 13

CP5

Application - Part II of III

3:20 p.m.-5:00 p.m.

Room: 501

For Part 1 see CP3
For Part 3 see CP11

Chair: Barry C. White, US Army Engineer Research and Development Center, U.S.

3:20-3:40 Landing on Mars: Petascale Unstructured-Grid CFD Simulations on Summit

Eric Nielsen, Aaron Walden, Ashley Korzun, Bill Jones, and Jan-Renee Carlson, NASA Langley Research Center, U.S.; Pat Moran and Tim Sandstrom, NASA Ames Research Center, U.S.; Mohammad Zubair, Old Dominion University, U.S.

3:45-4:05 Multiscale Multiphysics Coupling of Hall-Effect Thruster, Plume, and Surface Charging Models for Spacecraft Integration

Samuel Araki, ERC Inc. and Air Force Research Laboratory , U.S.

4:10-4:30 Numerical Method and Parallelization for the Computation of Coherent Synchrotron Radiation

Boqian Shen, Rice University, U.S.

4:35-4:55 Virtual Environment for Sensor Performance Assessment (vespa) Geometry Engine

Barry C. White, Robert H Hunter, and Aaron Valoroso, US Army Engineer Research and Development Center, U.S.; Reena Patel and Jerrell Ballard, US Army Corps of Engineers, U.S.

Thursday, February 13

CP6

Multigrid and Preconditioning

3:20 p.m.-4:35 p.m.

Room: 512

Chair: Delyan Z. Kalchev, Lawrence Livermore National Laboratory, U.S.

3:20-3:40 Graph Based Algebraic Multigrid Method

Manan J. Shah and Dr. Sashikumaar Ganesan, Indian Institute of Science, Bangalore, India

3:45-4:05 ThunderEgg: A Multigrid Solver for Variable Coefficient Elliptic Problems on Adaptive Cartesian Meshes

Scott Aiton, Donna Calhoun, and Grady B. Wright, Boise State University, U.S.

4:10-4:30 Preconditioning Finite Element Equations via Nonconforming Reformulations

Delyan Z. Kalchev and Panayot Vassilevski, Lawrence Livermore National Laboratory, U.S.

Thursday, February 13

CP7

Proceedings Papers - Part II of III

3:20 p.m.-4:35 p.m.

Room: 607

For Part 1 see CP4
For Part 3 see CP10

Chair: Olivier Beaumont, Inria, France

3:20-3:40 FastSV: A Distributed-Memory Connected Component Algorithm with Fast Convergence[5]

Yongzhe Zhang, SOKENDAI, Japan; Ariful Azad, Indiana University, U.S.; Zhenjiang Hu, Peking University, China

3:45-4:05 A Projection-Based Data Partitioning Method for Distributed Tomographic Reconstruction[6]

Jan-Willem Buurlage and Willem Jan Palenstijn, Centrum Wiskunde & Informatica, The Netherlands; Rob H. Bisseling, Utrecht University, The Netherlands; Joost Batenburg, Centrum voor Wiskunde en Informatica (CWI), Netherlands

4:10-4:30 Two-Level Dynamic Load Balancing for High Performance Scientific Applications[7]

Ali Mohammed, Aurélien Cavelan, Florina M. Ciorba, and Ruben Cabezon, University of Basel, Switzerland; Ioana Banicescu, Mississippi State University, U.S.

Intermission

5:00 p.m.-5:15 p.m.

SIAG/SC Business Meeting

5:15 p.m.-6:00 p.m.

Room: Elwha Ballroom

Complimentary beer and wine will be served.

[5] This presentation is included in the proceedings
[6] This presentation is included in the proceedings
[7] This presentation is included in the proceedings

Thursday, February 13

PP1

Poster Session

6:00 p.m.-8:00 p.m.

Room: Foyer - 5th Floor

Adapting a Multibody System Simulator to Auto-Tuning Linear Algebra Routines

Jesús Cémara, University of Murcia, Spain; José-Carlos Cano, Universidad Politécnica de Cartagena, Spain; Javier Cuenca and Domingo Giménez, University of Murcia, Spain; Mariano Saura-Sánchez, Universidad Politécnica de Cartagena, Spain

A Multi-GPU Implementation of a Second Order Scheme to Simulate Landslide-Generated Tsunamis

Marc de la Asuncion and Manuel J. Castro, University of Malaga, Spain

Fast Generation of Extreme-Scale Matrices with Preassigned 2-Norm Condition Number

Massimiliano Fasi, University of Manchester, United Kingdom; Nicholas J. Higham, The University of Manchester, UK

Multicolor Block Gauss-Seidel Using Kokkos

Brian Kelley and Sivasankaran Rajamanickam, Sandia National Laboratories, U.S.

A Parallel Framework for Nonlinear Optimization

Mohan Krishnamoorthy, Argonne National Laboratory, U.S.

Massive Scaling of MASSIF: Algorithm Development for Hooke's Law Simulations on Distributed GPU Systems

Anuva Kulkarni, Carnegie Mellon University, U.S.; Jelena Kovacevic, New York University, U.S.; Franz Franchetti, Carnegie Mellon University, U.S.

Accelerating Alternating Least Squares for Tensor Decomposition by Pairwise Perturbation

Linjian Ma and Edgar Solomonik, University of Illinois at Urbana-Champaign, U.S.

continued in next column

Application of Projectile Physics and Variable Drag Implications in Determining Market Price Movements for Futures Derivatives

Leonard Mushunje, Midlands State University, Zimbabwe

Simulating Quantum Circuits with Tensor Network States

Yuchen Pang, Yiqing Zhou, Tianyi Hao, and Edgar Solomonik, University of Illinois at Urbana-Champaign, U.S.

Parallel Fenics Implementation of Block Solvers

Innokentiy Protasov, University of Houston, U.S.; Justin Chang and Jeffery M. Allen, National Renewable Energy Laboratory, U.S.

Kokkos Kernels : A Performance Portable Library for Linear Algebra and Graph Algorithms

Siva Rajamanickam, Christian Trott, Nathan Ellingwood, Kyungjoo Kim, Brian Kelley, Vinh Dang, Luc Berger-Vergiat, and Jeremiah Wilke, Sandia National Laboratories, U.S.

An Interface for Extreme-Scale Geometric Multigrid in PETSc

Johann Rudi, Argonne National Laboratory, U.S.

Planning Robot Manipulation Using Parallel-in-Time Integration

Daniel Ruprecht, Hamburg University of Technology, Germany; Wisdom Agboh, Oliver Grainger, and Mehmet Dogar, University of Leeds, United Kingdom

Massively-Parallel Computing of Multi-Channel 2D Convolution

Stanislav Sedukhin and Yoichi Tomioka, University of Aizu, Japan

Semi-Structured Hybrid Multigrid on Octree Meshes

Yu-Hsuan Shih and Georg Stadler, Courant Institute of Mathematical Sciences, New York University, U.S.

A Parallel Implementation of Non-Linear Least Squares Method for CP Decomposition for Tensors

Navjot Singh, University of Illinois at Urbana-Champaign, U.S.; Hongru Yang, University of Texas at Austin, U.S.; Linjian Ma and Edgar Solomonik, University of Illinois at Urbana-Champaign, U.S.

continued in next column

GPU-Accelerated Barycentric Cluster-Particle Treecodes

Leighton Wilson and Nathan Vaughn, University of Michigan, U.S.; Lei Wang, University of Wisconsin, Milwaukee, U.S.; Robert Krasny, University of Michigan, U.S.

Graphblas with Performance Semantics: Specification, Design, and Implementation

Albert-Jan N. Yzelman and Wijnand Suijlen, Huawei Technologies, France

Thursday, February 13

PP2

Poster Session

6:00 p.m.-8:00 p.m.

Room: Foyer - 5th Floor

Ginkgo - a Node-Level Sparse Linear Algebra Library for High Performance Computing

Hartwig Anzt, Karlsruhe Institute of Technology, Germany and University of Tennessee, U.S.; Terry Cojean, Thomas Gruetzmacher, Pratik Nayak, Tobias Ribizel, and Yushiang Tsai, Karlsruhe Institute of Technology, Germany

Solving Hyperbolic PDE Systems with ExaHyPE

Michael Bader, Leonhard Rannabauer, and Anne Reinarz, Technische Universit?t M?nchen, Germany; Michael Dumbser, University of Trento, Italy; Alice-Agnes Gabriel, Ludwig Maximilian University of Munich, Germany; Tobias Weinzierl, Durham University, United Kingdom

On the Implementation of an FEM FMM Integral Solver for MHD Magnetic Fields

K. Daniel Brauss, Francis Marion University, U.S.

Construction of High-Order Multirate Imex Integrators for Large-Scale Complex Multiphysics Applications

Rujeko Chinomona and Daniel R. Reynolds, Southern Methodist University, U.S.

High-Performance Implementation of Wavelet Transforms Using SIMD

Camille Coti, Universit? Paris 13, France; Joel Falcou, LRI, Universite Paris-Sud, Orsay, France; Basarab Matei, Universit? Paris 13, France

Comparison of a Randomized and a Deterministic Low-Rank Approximation Algorithm

Robert Ernstbrunner, Viktoria Mayer, and Wilfried N. Gansterer, University of Vienna, Austria

SEMANTICMODELS.JL: A Framework for Automatic Composition of Scientific Models Across Domains

Micah E. Halter, Kun Cao, and James Fairbanks, Georgia Tech Research Institute, U.S.

continued in next column

Parallelization of DD and QD High-Precision Arithmetic Operations

Hidehiko Hasegawa, University of Tsukuba, Japan; Hotaka Yagi and Emiko Ishiwata, Tokyo University of Science, Japan

Parallelized Dual-Stage Energy Minimization

Khaled A. Helal and Bardia Barabadi, University of Victoria, Canada; Matt Gara, 3vGeomatics, Canada; Amirali Baniasadi and Nikitas Dimopoulos, University of Victoria, Canada

Multi-Step Communication in Enlarged Krylov Subspace Solvers

Shelby Lockhart and Luke Olson, University of Illinois at Urbana-Champaign, U.S.

Greedy Algorithms for Neural Network Architecture Optimization

Massimiliano Lupo Pasini and Junqi Yin, Oak Ridge National Laboratory, U.S.; Ying Wai Li, Los Alamos National Laboratory, U.S.; Markus Eisenbach, Oak Ridge National Laboratory, U.S.

PCPATCH: Software for the Topological Construction of Multigrid Relaxation Methods

Lawrence Mitchell, Durham University, United Kingdom; Patrick E. Farrell, Imperial College London, United Kingdom; Matthew G. Knepley, State University of New York at Buffalo, U.S.; Florian Wechsung, New York University, U.S.

A Scalable Block Preconditioner for High-Order Hybridized Discontinuous Galerkin Methods Applied to Incompressible Resistive MHD

Sriramkrishnan Muralikrishnan, Paul Scherrer Institut, Switzerland; Tan Bui-Thanh, University of Texas at Austin, U.S.; John N. Shadid, Sandia National Laboratories, U.S.

Additively Damped AFAC Variants for High Order Discretisations

Charles D. Murray and Tobias Weinzierl, Durham University, United Kingdom

Adaptive Domain Decomposition Method for Saddle Point Problem in Matrix Form

Frederic Nataf, Université Pierre et Marie Curie, France

A Discrete Element Method using Triangulated Particles

Peter J. Noble and Tobias Weinzierl, Durham University, United Kingdom

continued in next column

Algorithm-Specific Checkpointing vs. Exact State Reconstruction for the Preconditioned Conjugate Gradient Method

Christina Pacher, Carlos Pachajoa, Markus Levonyak, and Wilfried N. Gansterer, University of Vienna, Austria

Hardware Agnostic Implementation of Cosmo-Eulag Dynamical Core for Regional Numerical Weather Prediction using GridTools Framework

Zbigniew P. Piotrowski and Adam Ryczkowski, Institute of Meteorology and Water Management, Poland

Tasktorrent: A Task-Based Distributed Runtime System

Yizhou Qian, Leopold Cambier, and Eric Darve, Stanford University, U.S.

EquationBC: Solving a Different PDE on Boundary

Koki Sagiyama, Imperial College London, United Kingdom; Lawrence Mitchell, Durham University, United Kingdom; David Ham, Imperial College London, United Kingdom

Stable Automatic Tuning Method for Performance Fluctuation and Evaluation

Naoto Seki, Toshiki Tabeta, Akihiro Fujii, and Teruo Tanaka, Kogakuin University, Japan

Using Quantized Integer in LU Factorization with Partial Pivoting

Yaohung M. Tsai, Piotr Luszczek, and Jack Dongarra, University of Tennessee, Knoxville, U.S.

Gemslr: a Multilevel Low-Rank Preconditioning and Solution Package

Tianshi Xu, University of Minnesota, Twin Cities, U.S.; Vasilis Kalantzis, IBM T.J. Watson Research Center, U.S.; Geoffrey Dillon, University of South Carolina, U.S.; Yuanzhe Xi, Emory University, U.S.; Ruipeng Li, Lawrence Livermore National Laboratory, U.S.; Yousef Saad, University of Minnesota, U.S.

Optimization of GPU Kernels for Sparse Matrix Computations in HYPRE

Chaoyu Zhang, Arkansas State University, U.S.; Ruipeng Li, Lawrence Livermore National Laboratory, U.S.

Proxy App for Mesh Relaxation via Machine Learning on Advanced Hardware

Kristofer Zieb, Lawrence Livermore National Laboratory, U.S.

Friday, February 14

Registration

8:00 a.m.-5:00 p.m.

Room: 503

Friday, February 14

SP1

SIAG Best Paper Prize: The BLIS Framework: Experiments in Portability

8:30 a.m.-9:15 a.m.

Room: Elwha Ballroom

Chair: George Biros, University of Texas at Austin, U.S.

Over the past decade, the BLAS-like Library Instantiation Software (BLIS) project has carefully revisited past progress on how to structure the implementation of the level-3 BLAS-like operations (matrix-matrix computations) in particular and all basic linear algebra operations in general. The paper "The BLIS Framework: Experiments in Portability" demonstrates how a refactoring of prior approaches yields a more flexible, more easily maintained, highly portable, yet high-performing and scalable software library. Casting level-3 BLAS functionality in terms of multiplication with submatrices was proposed in the works of Kågström, Ling, and Van Loan. This led to efforts to auto-generate and tune such as PHiPAC and ATLAS. A dramatic breakthrough came circa 2000 when Goto proposed "Goto's algorithm" (now at the heart of most high-performance BLAS) for implementing matrix-matrix multiplication. BLIS casts Goto's algorithm in terms of five portable loops (written in C99) around a "microkernel" that updates a small submatrix of C that fits in registers. It is only this microkernel that needs to be customized for a new architecture when implementing matrix multiplication. The refactoring exposed in BLIS drastically reduced the size, complexity, and number of assembly kernels necessary for supporting high-performance across all datatypes and level-3 operations. The prize-winning paper was coauthored with Tyler Smith, Bryan Marker, Tze Meng Low, Francisco Igual, Mikhael Smelyanskiy, Xianyi Zhang, Michael Kistler, Vernon Austel, John Gunnels, and Lee Killough.

Field G. Van Zee
Robert A. van de Geijn
The University of Texas at Austin, U.S.

Intermission

9:15 a.m.-9:25 a.m.

Friday, February 14

SP2

SIAG/Supercomputing Early Career Prize: Scalable Algorithms for Tensor Computations

9:25 a.m.-9:55 a.m.

Room: Elwha Ballroom

Chair: Matthias Bolten, University of Wuppertal, Germany

We give a high-level overview of a few recent algorithmic advances in the design of efficient algorithms for tensor computations. We highlight a communication-avoiding parallel algorithm for dense symmetric eigenvalue problems, which is part of a broader family of new matrix factorization algorithms with costs that attain communication lower bounds. One recently developed practical variant of these algorithms, a 3D parallel algorithm for CholeskyQR2, achieves speed-ups of up to 3.3X over the best existing libraries. These algorithmic innovations are extended to tensor operations and deployed as part of the Cyclops library. We describe novel parallel implementations of tensor decomposition and tensor completion methods using Cyclops. Finally, we introduce a new algorithm for tensor decomposition, pairwise perturbation, which approximates the alternating least squares procedure with asymptotically less cost.

Edgar Solomonik
University of Illinois at Urbana-Champaign, U.S.

Friday, February 14

SP3

SIAG/Supercomputing Career Prize: Ghosts of Parallel Computing: Past, Present, and Future

9:55 a.m.-10:25 a.m.

Room: Elwha Ballroom

Chair: Ulrike Meier Yang, Lawrence Livermore National Laboratory, U.S.

Receiving a career prize means you're old. So in the first part of my talk I'll highlight a couple themes from the past 30 years of supercomputing which have taken us from gigaflops to the threshold of exaflops. In the latter part I'll speculate on what comes next. Spoiler alert: it seems unlikely the youngest members of the audience will see another 9 orders-of-magnitude speed-up within their career horizons. What does that mean for SIAM members interested in supercomputing and the topics we research? My perspective here will be application-centric (really, what else matters for science), so there will hopefully be ideas to both agree and disagree with.

Steve Plimpton

Sandia National Laboratories, U.S.

Coffee Break

10:25 a.m.-10:55 a.m.

Room: Gallery - 5th Floor

Friday, February 14

MS41

Trilinos and Hardware Independent Computing (Kokkos)

10:55 a.m.-12:35 p.m.

Room: Elwha Ballroom

Organizer: Andreas Adelmann
Paul Scherrer Institut, Switzerland

10:55-11:15 Hardware Architecture Independent Adaptive Mesh Refinement Solver using Trilinos
Matthias Frey, Andreas Adelmann, and Uldis Locans, Paul Scherrer Institut, Switzerland

11:20-11:40 Utopia: a Performance-Portable C++ Library for Non-Linear Algebra
Nur A. Fadel, Swiss National Supercomputing Center, Switzerland; Patrick Zulian and Alena Kopanicakova, Universitá della Svizzera italiana, Switzerland; Andreas Fink and Daniel Ganellari, Swiss National Supercomputing Center, Switzerland; Rolf Krause, Universit? della Svizzera italiana, Switzerland

11:45-12:05 State of the Tpetra Linear Solver Stack
Christopher Siefert, Karen D. Devine, Mark Hoemmen, Jonathan J. Hu, and Brian Kelley, Sandia National Laboratories, U.S.

12:10-12:30 Scalable Geometric Search Algorithms in ArborX
Andrey Prokopenko, Damien Lebrun-Grandie, Bruno Turcksin, and Daniel Arndt, Oak Ridge National Laboratory, U.S.

Friday, February 14

MS42

Improving Productivity and Sustainability for Parallel Computing Software - Part I of II

10:55 a.m.-12:35 p.m.

Room: 502

For Part 2 see MS52

As we move toward exascale machines and beyond, we face a daunting task of providing software for emerging extreme-scale disruptive architectures that will allow applications to run with extreme performance, high scalability, and reliability. In addition, demand for crosscutting software that spans multiple domain-specific sciences, a clamor for greater reproducibility, and new opportunities for greatly improved simulation capabilities, especially through coupling of data and physics across scales, will result in software being extremely complex and challenging. While hardware remains a primary focus for next-generation exascale computing, the aforementioned challenges also demand large investments in scientific software development. In this minisymposium, we aim to cover lost ground and address growing technical, practical, and social challenges in software productivity, quality, and sustainability, with a goal of helping develop better and long-living, cutting-edge parallel computing software that can fulfill its designated roles in emerging exascale ecosystems. Along with sharing of diverse experiences by experts in the parallel computing field, the hope is that this minisymposium will motivate the community to realign and refocus on developer productivity and software sustainability---which are both urgent and essential---to build better scientific software.

Organizer: Rinku Gupta
Argonne National Laboratory, U.S.

Organizer: Judith Hill
Oak Ridge National Laboratory, U.S.

Organizer: Lois McInnes
Argonne National Laboratory, U.S.

10:55-11:15 Views on Software Sustainability from a Computing Facility Perspective
Judith Hill, Oak Ridge National Laboratory, U.S.

continued on next page

11:20-11:40 The Sustainability Lessons of the SLATE Project
Jakub Kurzak and *Mark Gates*, University of Tennessee, Knoxville, U.S.; Ali M. Charara, King Abdullah University of Science & Technology (KAUST), Saudi Arabia; Asim YarKhan and Jamie M. Finney, University of Tennessee, Knoxville, U.S.; Jack J. Dongarra, University of Tennessee and Oak Ridge National Laboratory, U.S.

11:45-12:05 FEniCSX: A Sustainable Future for the FEniCS Project
Michal Habera and Jack S. Hale, University of Luxembourg, Luxembourg; Chris Richardson, University of Cambridge, United Kingdom; Johannes Ring and Marie E. Rognes, Simula Research Laboratory, Norway; Nathan Sime, Carnegie Institution for Science, U.S.; Garth Wells, University of Cambridge, United Kingdom

12:10-12:30 Training and Best Practices to Develop Portable Yet Performant Code
Sunita Chandrasekaran, University of Delaware, U.S.

Friday, February 14

MS43

Parallel Sparse and Conventional FFTs, Applications and Implementation - Part I of II

10:55 a.m.-12:35 p.m.

Room: 505

For Part 2 see MS53

The fast Fourier Transform (FFT) is an algorithm used in a wide variety of applications, yet does not make optimal use of many current and emerging platforms such as many-core processors, GPUs, and distributed-memory systems. Hardware utilization performance on its own does not, however, imply optimal problem-solving. The purpose of this minisymposium is to enable an exchange of information between people working on FFT algorithms such as sparse and conventional FFTs, to those working on FFT implementations, in particular for parallel hardware.

Organizer: Daisuke Takahashi
University of Tsukuba, Japan

Organizer: Franz Franchetti
Carnegie Mellon University, U.S.

Organizer: Samar A. Aseeri
King Abdullah University of Science & Technology (KAUST), Saudi Arabia

Organizer: Benson K. Muite
University of Tartu, Estonia

10:55-11:15 The Design of FFTX
Franz Franchetti, Carnegie Mellon University, U.S.

11:20-11:40 3D FFTs on HPC Many-Core and Hybrid CPU-GPU Platforms: Applications in Materials and Chemistry Codes
Andrew M. Canning, Lawrence Berkeley National Laboratory, U.S.

11:45-12:05 Designing An Adaptable Framework for Highly Scalable Multidimensional Spectral Transforms
Dmitry Pekurovsky, University of California, San Diego, U.S.

12:10-12:30 Implementation of Parallel 3-D Real FFT with 2-D Decomposition on Intel Xeon Phi Clusters
Daisuke Takahashi, University of Tsukuba, Japan

Friday, February 14

MS44

Parallel Matrix Factorization Algorithms - Part III of III

10:55 a.m.-12:35 p.m.

Room: 507

For Part 2 see MS35

This minisymposium brings together research on parallel algorithms for matrix factorizations for sparse, dense, and structured methods. Communication-avoidance has motivated new algorithms for dense matrix factorizations that are faster and more parallelizable. These techniques are being adapted also by new algorithms for parallel sparse direct solvers and preconditioning via approximate sparse matrix factorization. Low rank and hierarchically low rank matrix representations provide a competing route for finding solutions to numerical PDEs, but face their own scalability challenges. For example, low-rank factorizations motivate the development of efficient parallel pivoted QR factorization. The minisymposium will encompass state-of-the-art in parallel algorithms for factorization of dense, sparse, and structured matrices, and other related topics. In particular, speakers will discuss communication-avoiding parallel algorithms for dense QR, sparse QR, and QR with randomized column pivoting. Further, talks will describe innovations in state-of-the-art parallel libraries for sparse direct solvers and their application as solvers for numerical PDEs and dynamic optimization.

Organizer: Edgar Solomonik
University of Illinois at Urbana-Champaign, U.S.

10:55-11:15 Fine-Grained Parallel Incomplete Factorizations
Edmond Chow, Georgia Institute of Technology, U.S.

11:20-11:40 Parallel Sparse Indefinite Solvers
Florent Lopez, University of Tennessee, Knoxville, U.S.; Iain Duff, Science & Technology Facilities Council, United Kingdom and CERFACS, Toulouse, France

11:45-12:05 Hierarchical Algorithms on Hierarchical Architectures
David E. Keyes, King Abdullah University of Science & Technology (KAUST), Saudi Arabia

continued on next page

Friday, February 14

MS44

Parallel Matrix Factorization Algorithms - Part III of III

10:55 a.m.-12:35 p.m.

continued

12:10-12:30 PartILUT - a Parallel Threshold Incomplete Factorization Preconditioner for Multicore and GPU

Hartwig Anzt, Karlsruhe Institute of Technology, Germany and University of Tennessee, U.S.; Edmond Chow, Georgia Institute of Technology, U.S.; Tobias Ribizel, Karlsruhe Institute of Technology, Germany; Goran Flegar, Universitat Jaume I, Spain; Jack J. Dongarra, University of Tennessee and Oak Ridge National Laboratory, U.S.

Friday, February 14

MS45

High Performance Computing in Scientific Applications

10:55 a.m.-12:35 p.m.

Room: 508

Modern high-performance computing (HPC) is being widely used to massively parallelize scientific applications such as those in high-energy physics, climate, weather, materials science, and computational chemistry, to name a few. Some of the applied mathematics problems being tackled include optimizing objective functions having multiple local minima, optimizing simulation-based objective functions, and optimizing functions that are composed of information from multiple sources, including functions over surrogate models. Devising highly scalable, parallel, and efficient solutions to these problems requires an interdisciplinary approach drawing on expertise from domain sciences, data science, applied mathematics, and HPC. This minisymposium brings together pioneers solving problems in these areas. They will discuss the current state of the art, domain science impact and challenges, modeling and optimization approaches, scalable parallel software frameworks, and potential future research directions. They will describe software tools, algorithms, and techniques applied to problems in high-energy physics, but we believe that the lessons learned can be applied to other scientific domains as well.

Organizer: Mohan Krishnamoorthy
Argonne National Laboratory, U.S.

Organizer: Tom Peterka
Argonne National Laboratory, U.S.

10:55-11:15 Revolutionizing HEP Data Storage using HPC Technologies

Saba Sehrish, Fermi National Accelerator Laboratory, U.S.

11:20-11:40 Software Infrastructure for Scalable Data Analysis on HPC Systems

Orcun Yildiz, Argonne National Laboratory, U.S.

11:45-12:05 A Scalable Framework for Simulating Particle Collisions at the Energy Frontier

Stefan Hoeche, Fermi National Accelerator Laboratory, U.S.; *Holger Schulz*, University of Cincinnati, U.S.

12:10-12:30 Surrogate Optimization for HPC Applications

Juliane Mueller, Lawrence Berkeley National Laboratory, U.S.

continued in next column

Friday, February 14

MS46

Challenges in Parallel Adaptive Mesh Refinement - Part III of III

10:55 a.m.-12:35 p.m.

Room: 603

For Part 2 see MS37

Parallel adaptive mesh refinement (AMR) is a key technique when simulations are required to capture time-dependent and/or multiscale features. Frequent re-adaptation and repartitioning of the mesh during the simulation can impose significant overhead, particularly in largescale parallel environments. Further challenges arise due to the availability of accelerated or special-purpose hardware, and the trend toward hierarchical and hybrid compute architectures. Our minisymposium addresses algorithms, scalability, and software issues of parallel AMR.

Organizer: Carsten Burstedde
Universitaet Bonn, Germany

Organizer: Michael Bader
Technische Universität München, Germany

Organizer: Martin Berzins
University of Utah, U.S.

10:55-11:15 AMReX: A Block-Structured AMR Software Framework for the Exascale

Ann S. Almgren, John B. Bell, Kevin N. Gott, Weiqun Zhang, and Andrew Myers, Lawrence Berkeley National Laboratory, U.S.

11:20-11:40 Low-Latency Mesh-Refinement Cycle Algorithms for Octrees

Hansol David Suh and Tobin Isaac, Georgia Institute of Technology, U.S.

11:45-12:05 AMR During PDE-Constrained Optimization using PETSc

Matthew G. Knepley, State University of New York at Buffalo, U.S.; *Tobin Isaac*, Georgia Institute of Technology, U.S.

12:10-12:30 An Asynchronous Algorithm for 2:1 Octree Balance

Hansol David Suh and Tobin Isaac, Georgia Institute of Technology, U.S.

Friday, February 14

MS47

Parallel-in-Time Integration Methods - Part I of II

10:55 a.m.-12:35 p.m.

Room: 604

For Part 2 see MS58

Driven by the rapid increase in core numbers in high-performance computing systems, parallel-in-time methods are a quickly growing field of study that promise parallelization beyond widely developed space-parallel algorithms. The mini symposium will feature presentations on recent developments with respect to algorithms, implementation and analysis and application of various types of algorithms offering concurrency along the time direction.

Organizer: Daniel Ruprecht
Hamburg University of Technology, Germany

Organizer: Andrew T. Clarke
University of Leeds, United Kingdom

10:55-11:15 Warmstarting PFASST Iterations

Sebastian Götschel, Zuse Institute Berlin, Germany

11:20-11:40 Multigrid Reduction in Time for High-Order Discretizations of Hyperbolic Problems

Hans De Sterck, University of Waterloo, Canada; *Robert D. Falgout*, Lawrence Livermore National Laboratory, U.S.; Stephanie Friedhoff, University of Wuppertal, Germany; Oliver A. Krzysik, Monash University, Australia; Scott Maclachlan, Memorial University, Newfoundland, Canada

11:45-12:05 Parallelizing Exponential Integrators with PFASST

Michael Minion, Lawrence Berkeley National Laboratory, U.S.; Tommaso Buvoli, University of California, Merced, U.S.

12:10-12:30 Applications of Parareal to Geo/astrophysical Fluid Dynamics: Convection and Magnetic Field Generation

Andrew T. Clarke, Chris Davies, and Steve Tobias, University of Leeds, United Kingdom; Daniel Ruprecht, Hamburg University of Technology, Germany

Friday, February 14

MS48

Accelerating Data Sparse Applications on Massively Parallel Systems - Part III of III

10:55 a.m.-12:35 p.m.

Room: 606

For Part 2 see MS40

With multiple monikers such as block low-rank, data sparse, or hierarchical matrices, this minisymposium features presentations on exploiting structure of matrix data for more efficient solvers in science and engineering. This low-rank approximation technique exploits the data sparsity of the application system matrix by compressing matrix entries. Often, the off-diagonal submatrices are amenable to compression and may be regulated with a user-defined accuracy ranges. During the actual solver phase, the underlying linear algebra operations are carried out on the compressed data format rather than the original full data. As a result, the arithmetic complexity of the new solver ends up being reduced to much lower levels. This allows to tackle much larger simulation problems that would otherwise be constrained by either their storage size of computational cost. The wide variety of solvers and compression techniques will be presented to showcase to broad applicability of the data-sparse techniques in modern scientific applications.

Organizer: Hatem Ltaief
King Abdullah University of Science & Technology (KAUST), Saudi Arabia

Organizer: Piotr Luszczek
University of Tennessee, Knoxville, U.S.

10:55-11:15 Exploiting Generic Tiled Algorithms Toward Scalable H-Matrices Factorizations on Top of Runtime Systems

Rocío Carratalá-Sáez, Universitat Jaume I, Spain; Mathieu Faverge, Bordeaux INP, Inria, LaBRI, France; Gregoire Pichon, Inria, France; Enrique S. Quintana-Orté, Universidad Politecnica de Valencia, Spain; Guillaume Sylvand, Airbus Group Innovations, France

continued on next page

Friday, February 14

MS48

Accelerating Data Sparse Applications on Massively Parallel Systems - Part III of III

10:55 a.m.-12:35 p.m.

continued

11:20-11:40 A Task-Based H-Matrix Solver for Distributed Machines Memory with Manycore Nodes
David Goudin and C?dric Augonnet, CEA/DAM, France; Matthieu Kuhn, CEA, DAM, DIF, France

11:45-12:05 Approximate and Exact Selection Algorithms on GPUs
Tobias Ribizel, Karlsruhe Institute of Technology, Germany

12:10-12:30 Hierarchical Techniques for Solvers with Regulated Accuracy
Micah Beck, University of Tennessee, U.S.

Friday, February 14

MS49

GPU Computing for Solving Large Scale Scientific Problems

10:55 a.m.-12:35 p.m.

Room: 607

Graphic processing units (GPUs) are becoming an ever more important aspect of the high performance computing landscape. This is particularly true as the majority of compute performance for a number of pre-exascale systems is due to GPUs. For these, and many other systems, optimal performance can only be achieved if GPUs are used efficiently. Nevertheless, in many application domains significant challenges to effectively use GPUs remain. To mention just a few: GPUs require very fine grained parallelism (the result is that algorithms which are most effective on CPU based systems are not necessarily the best choice for GPUs), application developers have to deal with less available memory and smaller caches, communication between host and device and between multiple devices can be a bottleneck, and obtaining good performance on multiple architectures without code duplication (i.e. performance portability). This minisymposium is dedicated to the discussion of how to overcome some of these difficulties. We aim to highlight successful strategies for incorporating GPU support into scientific codes and present novel research that will help to improve, both in terms of development time as well as in terms of performance, the development of the next generation of GPU enabled codes.

Organizer: Lukas Einkemmer
University of Innsbruck, Austria

Organizer: Lukas Einkemmer
University of Innsbruck, Austria

Organizer: Martina Prugger
Vanderbilt University, U.S.

10:55-11:15 GPU Computing for Solving Kinetic Equations
Lukas Einkemmer, University of Innsbruck, Austria

11:20-11:40 On the Use of GPU-Enabled Clusters in Cardiac Electrophysiology
Johannes Langguth, Kristian Hustad, Neringa Altanaite, and Xing Cai, Simula Research Laboratory, Norway

11:45-12:05 Using GPUs in Chemical Kinetics - a Stochastic Solver Within the Framework of PySB
Martina Prugger, Vanderbilt University, U.S.

12:10-12:30 GPU-Powered Particle-in-Cell Community Frameworks for Laser-Plasma Interaction
Axel Huebl, Lawrence Berkeley National Laboratory, U.S.; PIConGPU Collaboration, Helmholtz-Center Dresden-Rossendorf, Germany; WarpX Collaboration, Lawrence Berkeley National Laboratory, U.S.

continued in next column

Friday, February 14

CP8

Numerical Methods for Flow Simulations

10:55 a.m.-12:35 p.m.

Room: 501

Chair: Het Mankat, University of Texas at Dallas, U.S.

10:55-11:15 Efficient Vectorised Cuda Kernels for High-Order Finite Element Flow Solvers

Jan R. Eichstaedt and Joaquim Peiro, Imperial College London, United Kingdom; David Moxey, University of Exeter, United Kingdom

11:20-11:40 SMILU: a Staggered-Grid, Multi-Level ILU Preconditioner for Steady Fluid-Transport Problems

Jonas Thies, German Aerospace Center (DLR), Germany; Fred Wubs, University of Groningen, The Netherlands; Sven Baars, University of Groningen, Netherlands

11:45-12:05 MaMiCo: Scalable Coupling of Particle Ensembles to Transient Continuum Flow Simulations

Piet Jarmatz and Philipp Neumann, Helmut Schmidt University, Germany

12:10-12:30 The Multiscale Perturbation Method for Elliptic Equations

Het Y. Mankad, Alsadig Ali, and Felipe Pereira, University of Texas at Dallas, U.S.; Fabricio S. Sousa, Universidade de Sao Paulo, Brazil

Friday, February 14

CP9

Autotuning

10:55 a.m.-12:10 p.m.

Room: 512

Chair: Alexander Lindsay, Idaho National Laboratory, U.S.

10:55-11:15 Hybrid Empirical and Simulation-Based Autotuning of a Dense Linear Algebra Library for Heterogeneous Architectures

Jesús Cámara, University of Murcia, Spain; Emmanuel Agullo, Inria, France; Javier Cuenca and Domingo Gim?nez, University of Murcia, Spain

11:20-11:40 HMG: A Configurable High-Performance Multilevel Preconditioner

Fande Kong, Cody J. Permann, and Alexander Lindsay, Idaho National Laboratory, U.S.

11:45-12:05 Automatic Scaling for Improved Conditioning in the Multiphysics Object-Oriented Simulation Environment

Alexander Lindsay, Fande Kong, Cody J. Permann, Derek R. Gaston, and Richard Martineau, Idaho National Laboratory, U.S.

Friday, February 14

CP10

Proceedings Papers - Part III of III

10:55 a.m.-12:10 p.m.

Room: 605

For Part 2 see CP7

Chair: Aparna Chandramowlishwaran, University of California, Irvine, U.S.

10:55-11:15 Scalable Resilience Against Node Failures for Communication-Hiding Preconditioned Conjugate Gradient and Conjugate Residual Methods[8]

Markus Levonyak, Christina Pacher, and Wilfried N. Gansterer, University of Vienna, Austria

11:20-11:40 Leveraging One-Sided Communication for Sparse Triangular Solvers

Nan Ding, *Samuel Williams*, Yang Liu, and Xiaoye S. Li, Lawrence Berkeley National Laboratory, U.S.

11:45-12:05 Efficient Parallel Algorithms and Implementations for Sparse Triangular Solves on GPUs[9]

Chaoyu Zhang, Arkansas State University, U.S.; Ruipeng Li, Lawrence Livermore National Laboratory, U.S.

Lunch Break

12:35 p.m.-2:05 p.m.

[8] This presentation is included in the proceedings
[9] This presentation is included in the proceedings

Friday, February 14

IP4

Modeling of Heterogeneous Computing Systems and Their Usages

2:05 p.m.-2:50 p.m.

Room: Elwha Ballroom

Chair: Richard Vuduc, Georgia Institute of Technology, U.S.

The last decade has seen a paradigm shift in the architecture of computing platforms, with a trend toward combining general-purpose processors and specialized accelerators. For example, GPUs have made a significant impact in both the hardware industry and the application domain, as has been seen from the recent development of machine learning applications. Other platforms such as Processing-In-Memory and FPGA-based reconfigurable architectures have regained attention, and with these special accelerators, the computing platforms become more heterogeneous. These heterogeneous architectures are especially attractive because they can provide high performance and energy efficiency for both general-purpose applications and high-throughput applications. Thus, from IoT devices to server processors, heterogeneous architectures have become increasingly popular. However, these heterogeneous architectures introduce several new challenges, including programmability issues and designing the hardware architecture in a way that can maximally exploit the underlying heterogeneity. To address these issues, a wide variety of modeling has been used including analytical, regression based, and cycle-level. In this talk I will discuss how different modeling techniques have been addressed and how they can be helpful to guide architecture studies and software optimizations.

Hyesoon Kim

Georgia Institute of Technology, U.S.

Coffee Break

2:50 p.m.-3:20 p.m.

Room: Gallery - 5th Floor

Friday, February 14

MS50

Advances in Parallel Sparse Linear System Solver Stacks for Exascale - Part I of II

3:20 p.m.-5:00 p.m.

Room: Elwha Ballroom

For Part 2 see MS61

Our minisymposium presents advances in high-performance algorithms and GPU implementations in sparse linear system solver stacks including Trilinos-Belos and Muelu, Hypre-BoomerAMG, VT libParaNumAl high-order matrix-free libraries and fast direct methods based on low-rank approximation. Although the focus is on low communication algorithms and GPU implementations, we also consider Exascale applications and mini-apps that can demonstrate significant performance gains by employing these frameworks. For Krylov methods our focus is on recent advances in communication-avoiding and low-synch algorithms. Preconditioners, such as ILU will require fast and possibly iterative triangular solvers. Polynomial type preconditioners and AMG smoothers have also been advocated for GPU, along with sparse approximate inverse algorithms such as AINV. GPU acceleration is prompting new directions and a re-examination of the basic set-up and solution algorithms. For example, semi-structured AMG facilitates the local tuning of smoothers to improve convergence and execution speed. We bring together the architects of these important solver frameworks to explore algorithm choices, describe the latest performance results using these approaches and discuss future plans for the recently announced Exascale computers Aurora and Frontier.

Organizer: Stephen Thomas

National Renewable Energy Laboratory, U.S.

Organizer: Erik G. Boman

Sandia National Laboratories, U.S.

3:20-3:40 Low-Synchronization Orthogonalization Schemes for S-Step and Pipelined Krylov Solvers in Trilinos[10]

Ichitaro Yamazaki, Sandia National Laboratories, U.S.; Stephen Thomas, National Renewable Energy

Laboratory, U.S.; *Mark Hoemmen* and Erik G. Boman, Sandia National Laboratories, U.S.; Katarzyna Swirydowicz, National Renewable Energy Laboratory, U.S.; James Elliott, Sandia National Laboratories, U.S.

3:45-4:05 Progress on Trilinos-Muelu Smoothed Aggregation Amg for GPU

Jonathan J. Hu, Sandia National Laboratories, U.S.

4:10-4:30 Recent Development of Multigrid Solvers in HYPRE on Modern Heterogeneous Computing Platforms

Ruipeng Li, Bjorn Sjogreen, Ulrike Meier Yang, and Robert Falgout, Lawrence Livermore National Laboratory, U.S.

4:35-4:55 Enhancing Hypre's Semi-Structured Capabilities

Ulrike Meier Yang, Robert Falgout, and *Victor Paludetto Magri*, Lawrence Livermore National Laboratory, U.S.

continued in next column

[10] This presentation is included in the proceedings

Friday, February 14

MS51

Novel Computational Algorithms for Future Computing Platforms - Part I of III

3:20 p.m.-5:00 p.m.

Room: 501

For Part 2 see MS62

In the early 2000s, due to constraints on economical heat dissipation, clock speeds of single-core CPUs could no longer be increased, which marked the adoption of multi-core CPUs, together with a paradigm shift to algorithms specifically designed for multi-core architectures. About 15 years into this current architectural cycle and on its way to exascale performance, the computing industry finds itself at the confluence of technical difficulties that cast doubt on its ability to sustain this architectural model beyond the exascale capability. These difficulties are driving the hardware industry to develop application-specific chips and to look beyond silicon- based chips (e.g., quantum computing, physical annealing, neuromorphics, etc.), with a continued emphasis on raw processing power and emerging concerns about energy efficiency. Hardware specialization will likely redefine the way computational algorithms are developed over the next two decades for a wide range of important applications: large-scale PDE-based problems (CFD, wave propagation, subsurface modeling, etc.), artificial intelligence, computational chemistry, and optimization problems, to name a few. The pressure to decrease time to solution or improve simulation fidelity, or both, for these applications will continue unabated. This minisymposium provides a forum for sharing innovative ideas on algorithm development for leveraging future computing platforms.

Organizer: Arash Fathi
ExxonMobil Research and Engineering, U.S.

Organizer: Dimitar Trenev
ExxonMobil Research and Engineering

Organizer: Jason Riedy
Georgia Institute of Technology, U.S.

Organizer: Jeffrey Young
Georgia Institute of Technology, U.S.

Friday, February 14

MS51

Novel Computational Algorithms for Future Computing Platforms - Part I of III

3:20 p.m.-5:00 p.m.

continued

Organizer: Laurent White
ExxonMobil Research and Engineering

3:20-3:40 Scientific Computing in a Changing Landscape
Laurent White and Dimitar Trenev, ExxonMobil Research and Engineering; Arash Fathi, ExxonMobil Research and Engineering, U.S.

3:45-4:05 Data Movement Orchestration in Accelerator-Rich Systems
Andreas Gerstlauer, University of Texas at Austin, U.S.

4:10-4:30 Codesign Tradeoffs for Deep Learning Hardware Accelerators
Ardavan Pedram, Stanford University, U.S.

4:35-4:55 A Wafer-Scale Chip and System for Deep Neural Networks
Rob Schreiber, Cerebras, U.S.

Friday, February 14

MS52

Improving Productivity and Sustainability for Parallel Computing Software - Part II of II

3:20 p.m.-5:00 p.m.

Room: 502

For Part 1 see MS42

As we move toward exascale machines and beyond, we face a daunting task of providing software for emerging extreme-scale disruptive architectures that will allow applications to run with extreme performance, high scalability, and reliability. In addition, demand for crosscutting software that spans multiple domain-specific sciences, a clamor for greater reproducibility, and new opportunities for greatly improved simulation capabilities, especially through coupling of data and physics across scales, will result in software being extremely complex and challenging. While hardware remains a primary focus for next-generation exascale computing, the aforementioned challenges also demand large investments in scientific software development. In this minisymposium, we aim to cover lost ground and address growing technical, practical, and social challenges in software productivity, quality, and sustainability, with a goal of helping develop better and long-living, cutting-edge parallel computing software that can fulfill its designated roles in emerging exascale ecosystems. Along with sharing of diverse experiences by experts in the parallel computing field, the hope is that this minisymposium will motivate the community to realign and refocus on developer productivity and software sustainability---which are both urgent and essential---to build better scientific software.

Organizer: Rinku Gupta
Argonne National Laboratory, U.S.

Organizer: Judith Hill
Oak Ridge National Laboratory, U.S.

Organizer: Lois McInnes
Argonne National Laboratory, U.S.

continued on next page

Friday, February 14

MS52

Improving Productivity and Sustainability for Parallel Computing Software - Part II of II

3:20 p.m.-5:00 p.m.

continued

3:20-3:40 Productivity and Sustainability in a Community-Driven Software Ecosystem for Watershed Science

David Moulton, Los Alamos National Laboratory, U.S.; Scott Painter, Oak Ridge National Laboratory, U.S.; Sergi Molins, Lawrence Berkeley National Laboratory, U.S.; Xingyuan Chen, Pacific Northwest National Laboratory, U.S.; Reed M. Maxwell, Colorado School of Mines, U.S.; Laura Condon, University of Arizona, U.S.; Steve G. Smith, Lawrence Livermore National Laboratory, U.S.; Hai Ah Nam, Los Alamos National Laboratory, U.S.

3:45-4:05 Software Sustainability Lessons from the Fluid Dynamics Community

Kenneth Jansen, University of Colorado Boulder, U.S.

4:10-4:30 Challenges and Best Practices in the Computational Molecular Sciences

Benjamin P. Pritchard, Virginia Tech, U.S.

4:35-4:55 Experiences with Productivity and Software Sustainability on LCF Machines

Katherine Riley, Argonne National Laboratory, U.S.

Friday, February 14

MS53

Parallel Sparse and Conventional FFTs, Applications and Implementation - Part II of II

3:20 p.m.-5:00 p.m.

Room: 505

For Part 1 see MS43

The fast Fourier Transform (FFT) is an algorithm used in a wide variety of applications, yet does not make optimal use of many current and emerging platforms such as many-core processors, GPUs, and distributed-memory systems. Hardware utilization performance on its own does not, however, imply optimal problem-solving. The purpose of this minisymposium is to enable an exchange of information between people working on FFT algorithms such as sparse and conventional FFTs, to those working on FFT implementations, in particular for parallel hardware.

Organizer: Daisuke Takahashi
University of Tsukuba, Japan

Organizer: Franz Franchetti
Carnegie Mellon University, U.S.

Organizer: Samar A. Aseeri
King Abdullah University of Science & Technology (KAUST), Saudi Arabia

Organizer: Benson K. Muite
University of Tartu, Estonia

3:20-3:40 Implementing some Strategies to Reduce Communication Time of FFT Algorithm

Samar A. Aseeri and David E. Keyes, King Abdullah University of Science & Technology (KAUST), Saudi Arabia; Anando Chatterjee and Mahendra Verma, Indian Institute of Technology Kanpur, India

3:45-4:05 Fast Parallel Multidimensional FFT Using Advanced MPI

Lisandro Dalcin, King Abdullah University of Science & Technology (KAUST), Saudi Arabia; Mikael Mortensen, University of Oslo, Norway; David E. Keyes, King Abdullah University of Science & Technology (KAUST), Saudi Arabia

4:10-4:30 FFTW++: A Hybrid OpenMP/MPI Implementation of FFTs and Implicitly Dealiased Convolutions

John C. Bowman, University of Alberta, Canada; Malcolm Roberts, AMD, Canada

4:35-4:55 rocFFT: An Open-Source GPU FFT Library for Exascale Systems

Malcolm Roberts, Fei Zheng, and Bragadeesh Natarajan, AMD, Canada

continued in next column

Friday, February 14

MS54

Particle Methods: Algorithms and Software Technology for Exascale - Part I of III

3:20 p.m.-5:00 p.m.

Room: 507

For Part 2 see MS64

Computational algorithms that use particles are ubiquitous in high performance computing. In a given algorithm, particles may represent physical particles as in molecular dynamics algorithms, they may represent ensembles of physical particles such as in some variations of particle-in-cell methods, or they may represent a discrete computational element of the continuum in Lagrangian or hybrid Lagrangian-Eulerian computational schemes for fluids and mechanics. Within the Exascale Computing Project (ECP), a large project sponsored by the U.S. Department of Energy for the development of exascale hardware, software, and science applications, numerous groups are developing particle methods targeting forthcoming exascale machines with a focus on performance and portability. In this minisymposium we will present software libraries and frameworks currently under development in ECP that are designed for deploying particle methods at scale, we will discuss algorithmic developments and performance engineering of particle methods along with their unique challenges, and then finally we will cover several of the ECP applications themselves, many of which use the aforementioned software libraries and frameworks. ECP particle applications covered will include particle-in-cell algorithms, molecular dynamics, as well as other Lagrangian and hybrid Lagrangian-Eulerian schemes with specific applications to linear accelerator modeling, plasma physics, cosmology, and multi-phase flow.

Organizer: Stuart Slattery
Oak Ridge National Laboratory, U.S.

3:20-3:40 Cabana — a Co-Designed Library for Particle Applications
Stuart Slattery, Oak Ridge National Laboratory, U.S.

3:45-4:05 An Overview of Particles in Amrex, with Applications to Accelerator Modelling, Cosmology, and Multi-Phase Flow
Andrew Myers, Lawrence Berkeley National Laboratory, U.S.

4:10-4:30 Compatible Particle Discretization via Generalized Moving Least Squares
Nathaniel Trask, Sandia National Laboratories, U.S.

4:35-4:55 Rendezvous Algorithms for Large-Scale Particle Simulations
Steve Plimpton, Sandia National Laboratories, U.S.; Chris Knight, Argonne National Laboratory, U.S.

continued in next column

Friday, February 14

MS55

High-Performance Tensor Computation and Applications - Part I of III

3:20 p.m.-5:00 p.m.

Room: 508

For Part 2 see MS65

Tensors are higher order generalization of matrices that provide a natural way to represent a multi-relational dataset. Given a dataset encoded as a tensor, tensor decomposition serves as a promising analytics tool for mining this data to uncover hidden structure within the data's relations. This minisymposium explores efficient and scalable solutions for calculating tensor decomposition, as well as its application in data analytics across areas spanning signal processing, cybersecurity, machine learning, and beyond.

Organizer: Jee W. Choi
University of Oregon, U.S.

Organizer: Richard Vuduc
Georgia Institute of Technology, U.S.

Organizer: Eric Phipps
Sandia National Laboratories, U.S.

3:20-3:40 Tensor Decomposition for Malware Detection
Jee W. Choi, University of Oregon, U.S.

3:45-4:05 Sketching-Based Streaming Tensor Decompositions and Detection of Concept Drift in Unsupervised Exploratory Analysis
Evangelos Papalexakis, University of California, Riverside, U.S.

4:10-4:30 PASTA: A Parallel Sparse Tensor Algorithm Benchmark Suite
Jiajia Li, Pacific Northwest National Laboratory, U.S.; Mahesh Lakshminarasimhan, University of Utah, U.S.; Xiaolong Wu, Purdue University, U.S.; Ang Li, Pacific Northwest National Laboratory, U.S.; Catherine Olschanowsky, Boise State University, U.S.; Kevin Barker, Pacific Northwest National Laboratory, U.S.

4:35-4:55 Efficient Tensor Operations via Sketching
Yang Shi, Rakuten, Japan and University of California, Irvine, U.S.

Friday, February 14

MS56

HPC Aspects of Tsunami Simulation

3:20 p.m.-5:00 p.m.

Room: 512

To support hazard assessment and mitigation, large ensembles of tsuinami simulations are required, to cover parameter spaces, expose sensitivities and possibly exploit formal uncertatinty quantification techniques. Hence, we need to bring down the time to solution for individual tsunami simulations as far as possible, exploiting latest advances in numerics and (adaptive) algorithms and especially latest HPC architectures.

Organizer: Michael Bader
Technische Universit?t M?nchen, Germany

Organizer: Manuel J. Castro
University of Malaga, Spain

3:20-3:40 Optimising Time to Solution of Finite Volume and Discontinuous Galerkin Tsunami Models in Sam(oa)^2

Leonhard Rannabauer and Michael Bader, Technische Universit?t M?nchen, Germany

3:45-4:05 HPC Acceleration of the GeoClaw Software for Modeling Geohazards

Randall LeVeque, University of Washington, U.S.; Kyle T. Mandli, Columbia University, U.S.; Xinsheng Qin, University of Washington, Seattle, U.S.

4:10-4:30 A Fully Unspilt Wave Propagation Algorithm for Shallow Water Flows on GPUs

Donna Calhoun, Boise State University, U.S.

4:35-4:55 Improving Tsunami-HySEA as FTRT Simulator in the Framework of TEWS

Jose Manuel Gonzalez-Vida and Manuel J. Castro, University of Malaga, Spain; Jorge Macias Sanchez, Universidad de Málaga, Spain; Marc de la Asuncion, University of Malaga, Spain

Friday, February 14

MS57

Parallel Eigenvalue Algorithms for Physical Simulation - Part I of III

3:20 p.m.-5:00 p.m.

Room: 603

For Part 2 see MS67

Large scale eigenvalue computations are ubiquitous throughout scientific computation. However, the steep $O(N^3)$ scaling of traditional eigenvalue algorithms often lead to the solution of a particular eigenvalue problem becoming the computational bottleneck in many simulations of physical systems. Over the years, many important algorithmic developments have been made to allow leverage of the latest advances in massively parallel computing architectures to enable the simulation of large physical systems on the world's largest supercomputers. In this minisymposium, we examine several recent advances in parallel eigenvalue algorithms for eigenvalue problems which arise in scientific computation.

Organizer: Roel Van Beeumen
Lawrence Berkeley National Laboratory, U.S.

Organizer: David B. Williams-Young
Lawrence Berkeley National Laboratory, U.S.

Organizer: Chao Yang
Lawrence Berkeley National Laboratory, U.S.

3:20-3:40 Solving Large Eigenvalue Problems with the Parallel EVSL Package

Yousef Saad, University of Minnesota, U.S.; Ruipeng Li, Lawrence Livermore National Laboratory, U.S.; Yuanzhe Xi, Emory University, U.S.

3:45-4:05 Low-Rank Stopping Criteria for Block Parallel SVD

Steven Goldenberg, College of William & Mary, U.S.

4:10-4:30 Matrix Powers Kernels for Thick-Restart Lanczos with Explicit External Deflation

Ichitaro Yamazaki, Sandia National Laboratories, U.S.; Zhaojun Bai, University of California, Davis, U.S.; Ding Lu, University of Geneva, Switzerland; Chao-Ping Lin, University of California, Davis, U.S.; Jack J. Dongarra, University of Tennessee and Oak Ridge National Laboratory, U.S.

4:35-4:55 Parallel Shift-Invert Spectrum Slicing for Symmetric Self-Consistent Eigenvalue Computation

David B. Williams-Young, Lawrence Berkeley National Laboratory, U.S.

continued in next column

Friday, February 14

MS58

Parallel-in-Time Integration Methods - Part II of II

3:20 p.m.-5:00 p.m.

Room: 604

For Part 1 see MS47

Driven by the rapid increase in core numbers in high-performance computing systems, parallel-in-time methods are a quickly growing field of study that promise parallelization beyond widely developed space-parallel algorithms. The mini symposium will feature presentations on recent developments with respect to algorithms, implementation and analysis and application of various types of algorithms offering concurrency along the time direction.

Organizer: Daniel Ruprecht
Hamburg University of Technology, Germany

Organizer: Andrew T. Clarke
University of Leeds, United Kingdom

3:20-3:40 Performance Analysis and Benchmarking for pySDC
Robert Speck, Jülich Supercomputing Centre, Germany

3:45-4:05 Multigrid-in-Time SQP Methods for PDE-Constrained Optimization
Denis Ridzal and Eric C. Cyr, Sandia National Laboratories, U.S.

4:10-4:30 Multigrid for Shifted Systems Appearing in Parallel-in-Time Integration
Matthias Bolten, University of Wuppertal, Germany

4:35-4:55 Parallel-in-Time Simulation of the Schrödinger Equation
Hannah Rittich, Jülich Supercomputing Centre, Germany

Friday, February 14

MS59

Resilience and Fault Tolerance for Extreme Computing Systems - Part I of III

3:20 p.m.-5:00 p.m.

Room: 606

For Part 2 see MS70

The reliability of large scale computing systems has been a major concern of high performance computing due to the ever increasing complexity of hardware and software components combined with the tight power budget for the operations. Under such unreliable computing systems, it is essential to introduce failure mitigations at the runtime and application layers to complement the resilience at the hardware/system levels. Today, the major resilience scheme is coordinated checkpoint and restart (CR) that involves global coordination of processes and threads. This global recovery model entails inherent scalability issues to handle a large class of errors and failures, and thus alternative approaches would improve the scalability and reliability of computing systems and applications together. In this minisymposium, we will discuss the reliability issues of the large scale computing systems, application/algorithm based resilience and programming model support to enhance the reliability of application program executions. We will also cover the recent global CR techniques that mitigates the performance interference to applications from checkpointing operations.

Organizer: Keita Teranishi
Sandia National Laboratories, U.S.

Organizer: Christian Engelmann
Oak Ridge National Laboratory, U.S.

Organizer: George Bosilca
University of Tennessee, Knoxville, U.S.

3:20-3:40 The Resilience Problem in Extreme Scale Computing
Christian Engelmann, Oak Ridge National Laboratory, U.S.

3:45-4:05 Resilient Computation Patterns in Scientific Applications
Luanzheng Guo and Dong Li, University of California, Merced, U.S.; *Ignacio Laguna*, Lawrence Livermore National Laboratory, U.S.; Martin Schulz, Technische Universitaet Muenchen, Germany

4:10-4:30 Resilience for Large-Scale Iterative Linear Solvers
Wilfried N. Gansterer, Markus Levonyak, Carlos Pachajoa, and Christina Pacher, University of Vienna, Austria

4:35-4:55 The Mathematical Analysis of Faults and the Resilience of Applications
Laura Monroe, Los Alamos National Laboratory, U.S.

continued in next column

Friday, February 14

MS60

Low-Rank Compression-Based Fast Sparse Direct Solvers

3:20 p.m.-5:00 p.m.

Room: 607

Recent advances have drastically reduced the asymptotic complexity of computation and memory usage of sparse factorization based solvers by incorporating hierarchical rank-structured or other data-sparse matrix compression techniques. This has resulted in complex solvers that are highly scalable, purely algebraic and robust even for highly indefinite and ill-conditioned systems. In this minisymposium, we discuss algorithmic innovations, recent software releases and potential applications of these solvers. We hope to open a discussion on the relative merits of different approaches in terms of both their asymptotic and practical benefits and improve collaboration among mathematics researchers and application scientists in need of fast sparse direct solvers.

Organizer: Pieter Ghysels
Lawrence Berkeley National Laboratory, U.S.

Organizer: Yang Liu
Lawrence Berkeley National Laboratory, U.S.

3:20-3:40 Incorporating Hierarchical Matrix Compression and Butterfly Factorizations in a Multifrontal Lu Solver
Pieter Ghysels, Yang Liu, and Xiaoye S. Li, Lawrence Berkeley National Laboratory, U.S.

3:45-4:05 Using Block-Low Rank Techniques for Large Finite Element Industrial Applications
Francois-Henry Rouet, Livermore Software Technology Corporation, U.S.; Patrick R. Amestoy, Universit? of Toulouse, France; Alfredo Buttari, CNRS-IRIT, France; Jean-Yves L'Excellent, Inria-LIP-ENS Lyon, France; Theo Mary, The University of Manchester, UK

4:10-4:30 Fast H^2 Algorithms for Directly Solving General Sparse Matrices
Dan Jiao and Miaomiao Ma, Purdue University, U.S.

continued in next column

4:35-4:55 Recent Developments Around the Block Low-Rank PaStiX Solver
Mathieu Faverge, Bordeaux INP, Inria, LaBRI, France; Gregoire Pichon, Inria, France; Pierre Ramet, Université de Bordeaux, Inria, LaBRI, France

Friday, February 14

CP11

Applications - Part III of III

3:20 p.m.-5:00 p.m.

Room: 605

For Part 2 see CP5

Chair: David Day, Sandia National Laboratories, U.S.

3:20-3:40 Graph Partitioning for Computational Mechanics Simulations with GDSW
David Day, Sandia National Laboratories, U.S.

3:45-4:05 Scalable FFT-Krylov Subspace Method for Scattering Problem
Yun Teck Lee, Ron Gonzales, and Yuiry Gryazin, Idaho State University, U.S.

4:10-4:30 Parallel Sweeping Preconditioners For Domain Decomposition Methods Applied to the Helmholtz Equation
Ruiyang Dai and Jean François Remacle, Université Catholique de Louvain, Belgium; Chistophe Geuzaine, University of Liege, Belgium

4:35-4:55 Communication Performance Modeling of LAP3D and its Application in Performance Optimization
Hong Guo, Institute of Applied Physics and Computational Mathematics, China

Saturday, February 15

Registration

8:00 a.m.-2:00 p.m.

Room: 503

Saturday, February 15

IP5

Methods and Models for Reducing Communication

8:30 a.m.-9:15 a.m.

Room: Elwha Ballroom

Chair: Matthias Bolten, University of Wuppertal, Germany

Parallel communication costs can lead to significant performance limitations in many computational kernels. From structured operations across grids to sparse matrix computations to global reductions, understanding the roadblocks is critical in achieving the highest performance. In addition, new architectures and networks are exposing opportunities and challenges in organizing communication even in the most fundamental routines. In this talk we survey new approaches for modeling performance and offer several routes to reorganizing computation by increasing locality in a node aware fashion. As an example, algebraic multigrid methods are becoming robust solvers for a variety of problems. Yet the strong scaling limitations in the underlying sparse matrix routines can severely hinder performance. Likewise, a key feature of current architectures is node-level parallelism. Standard approaches to inter-process communication will send data regardless of the locations of send and receive processes. Yet, there are notable differences in the cost of intra- and inter-node communication. In response, communication can be reorganized to take advantage of the less costly intra-node communication, reducing both the number and size of inter-node messages. We will introduce some new models and new approaches to organizing communication that can significantly speed up solvers in this setting.

Luke Olson

University of Illinois at Urbana-Champaign, U.S.

Intermission

9:15 a.m.-9:25 a.m.

Saturday, February 15

IP6

Cognitive Discovery: Pushing the Frontier of Technical R&D with AI

9:25 a.m.-10:10 a.m.

Room: Elwha Ballroom

Chair: Ulrike Meier Yang, Lawrence Livermore National Laboratory, U.S.

Cognitive Discovery is an overarching framework that uses AI to achieve scientific knowledge extraction and representation, to intelligently design and guide simulations, in order to drastically accelerate the pace of scientific discovery. Cognitive Discovery targets to accelerate scientific workflows in technical disciplines and provide a new generation of tools. The workflows follow the cycle: a) Massive literature review in order to understand the problem at hand. Literature refers to all aspects such as mathematical modelling, solution methods, actual computer models and HPC deployment. b) Enrichment of literature data with experimental data and formation of hypotheses. c) Running simulations to test hypotheses and generate new knowledge in order to close any knowledge gaps. All three phases suffer today major disruptions. Simply put: the volume of new literature is exploding (e.g. roughly 450K new publications in materials science are published every year, tens of thousands of papers in numerical and HPC methods need to be reviewed). IoT advances as well as advances in measuring all aspects of HPC systems create an explosion of data. High fidelity models lead to massive configuration spaces the complexity of which clearly outpaces our capability to scale and efficiently run modern HPC systems. We will showcase how AI can help dramatically improve this setting and lead to a massive acceleration for scientific discovery.

Costas Bekas

IBM Research - Zurich, Switzerland

Coffee Break

10:10 a.m.-10:40 a.m.

Room: Gallery - 5th Floor

Saturday, February 15

MS14

Advanced Visualisation, Analysis, and Parallelisation Concepts for Multi-Scale CFD Simulations in Science and Engineering

10:40 a.m.-11:55 a.m.

Room: 607

Due to recent advances in supercomputing, more and more scientific questions - especially from the so-called emerging sciences such as medicine, sociology, biology, virology, chemistry, climate or geo-sciences - can be answered today using high-performance computing (HPC). Such questions could cover the structural analysis of buildings and constructions in a global context (e.g. earth quakes), the prediction of floods and flooding damages due to heavy rainfall, the thread of tsunamis on coastal regions, the risk analysis of pollutant diffusion in populated regions, the simulation of evacuation scenarios on the facility, urban quarter, and city scale, or the optimisation of traffic flow within entire cities during rush hour; just to name a few. On the other side, "high-performance computing must now assume a broader meaning, encompassing not only flops, but also the ability, for example, to efficiently manipulate vast and rapidly increasing quantities of both numerical and non-numerical data" (Kalil, Miller: Advancing U.S. Leadership in High-Performance Computing. The White House, 2015.). In this minisymposium, different aspects of multi-scale, multi-level, multi-physics applications from science and engineering should be addressed, dealing with topics "but not limited to" such as parallelisation strategies, advanced numerical algorithms, coupling interfaces, data orchestration, interaction concepts, (big) data exploration, or visual data analytics.

Organizer: Ralf-Peter Mundani
FHGR, Switzerland

Organizer: Jérôme Frisch
RWTH Aachen University, Germany

10:40-11:00 Large-Scale Multiphase Flow Simulations with AMR using Dynamic Load Balance
Takayuki Aoki, Tokyo Institute of Technology, Japan

11:05-11:25 Distributed Domain Generation for Large Scale CFD Applications
Christoph Ertl, Technische Universitaet Muenchen, Germany; Ralf-Peter Mundani, FHGR, Switzerland; Ernst Rank, Technische Universit?t M?nchen, Germany

11:30-11:50 Immersive Data Exploration and Analysis of Large Scale CFD Applications
Ralf-Peter Mundani, FHGR, Switzerland; Sari Ugurcan and Christoph Ertl, Technische Universitaet Muenchen, Germany

continued in next column

Saturday, February 15

MS61

Advances in Parallel Sparse Linear System Solver Stacks for Exascale - Part II of II

10:40 a.m.-12:20 p.m.

Room: Elwha Ballroom

For Part 1 see MS50

Our minisymposium presents advances in high-performance algorithms and GPU implementations in sparse linear system solver stacks including Trilinos-Belos and Muelu, Hypre-BoomerAMG, VT libParaNumAl high-order matrix-free libraries and fast direct methods based on low-rank approximation. Although the focus is on low communication algorithms and GPU implementations, we also consider Exascale applications and mini-apps that can demonstrate significant performance gains by employing these frameworks. For Krylov methods our focus is on recent advances in communication-avoiding and low-synch algorithms. Preconditioners, such as ILU will require fast and possibly iterative triangular solvers. Polynomial type preconditioners and AMG smoothers have also been advocated for GPU, along with sparse approximate inverse algorithms such as AINV. GPU acceleration is prompting new directions and a re-examination of the basic set-up and solution algorithms. For example, semi-structured AMG facilitates the local tuning of smoothers to improve convergence and execution speed. We bring together the architects of these important solver frameworks to explore algorithm choices, describe the latest performance results using these approaches and discuss future plans for the recently announced Exascale computers Aurora and Frontier.

Organizer: Stephen Thomas
National Renewable Energy Laboratory, U.S.

Organizer: Erik G. Boman
Sandia National Laboratories, U.S.

10:40-11:00 Ceed/vt LibParaNumAl High-Order Matrix-Free AMG on AMD-GPU
Noel A. Chalmers, AMD Research, U.S.

continued on next page

11:05-11:25 Asynchronous Jacobi-Richardson, and Gauss-Seidel Smoothers for Hypre One-Synch FGMRES-AMG
Stephen Thomas, Katarzyna Swirydowicz, Shreyas Ananthan, and Michael Sprague, National Renewable Energy Laboratory, U.S.

11:30-11:50 A General Parallel Sparsified Nested Dissection Algorithm using a Task-Based Runtime System
Leopold Cambier, Stanford University, U.S.

11:55-12:15 One-Sync CGS2 Algorithm in the Context of QR Factorization and Arnoldi Process
Daniel Bielich, University of Colorado, Denver, U.S.; Katarzyna Swirydowicz and Stephen Thomas, National Renewable Energy Laboratory, U.S.; Julien Langou, University of Colorado, Denver, U.S.

Saturday, February 15

MS62

Novel Computational Algorithms for Future Computing Platforms - Part II of III

10:40 a.m.-12:20 p.m.

Room: 501

For Part 1 see MS51
For Part 3 see MS72

In the early 2000s, due to constraints on economical heat dissipation, clock speeds of single-core CPUs could no longer be increased, which marked the adoption of multi-core CPUs, together with a paradigm shift to algorithms specifically designed for multi-core architectures. About 15 years into this current architectural cycle and on its way to exascale performance, the computing industry finds itself at the confluence of technical difficulties that cast doubt on its ability to sustain this architectural model beyond the exascale capability. These difficulties are driving the hardware industry to develop application-specific chips and to look beyond silicon- based chips (e.g., quantum computing, physical annealing, neuromorphics, etc.), with a continued emphasis on raw processing power and emerging concerns about energy efficiency. Hardware specialization will likely redefine the way computational algorithms are developed over the next two decades for a wide range of important applications: large-scale PDE-based problems (CFD, wave propagation, subsurface modeling, etc.), artificial intelligence, computational chemistry, and optimization problems, to name a few. The pressure to decrease time to solution or improve simulation fidelity, or both, for these applications will continue unabated. This minisymposium provides a forum for sharing innovative ideas on algorithm development for leveraging future computing platforms.

Organizer: Arash Fathi
ExxonMobil Research and Engineering, U.S.

Organizer: Dimitar Trenev
ExxonMobil Research and Engineering

Organizer: Jason Riedy
Georgia Institute of Technology, U.S.

Organizer: Jeffrey Young
Georgia Institute of Technology, U.S.

Organizer: Laurent White
ExxonMobil Research and Engineering

10:40-11:00 Quantum Computing -- Overview & Potential Applications
Gilad Ben-Shach, IBM Corporation, U.S.

11:05-11:25 Towards Optical Neural Networks and Annealing Machines at the Quantum Limit
Ryan Hamerly, Massachusetts Institute of Technology, U.S.

11:30-11:50 Quantum Circuit Synthesis using Linear Algebra and Optimization Algorithms
Marc Baboulin, University of Paris-Sud, France

11:55-12:15 Neuromorphic Computing: A Platform for Machine Learning and Beyond
Catherine D. Schumann, Oak Ridge National Laboratory, U.S.

continued in next column

Saturday, February 15

MS63

Data-Centric Operating Systems and Runtimes

10:40 a.m.-12:20 p.m.

Room: 505

Data-Centric Operating Systems and Runtimes are becoming increasingly important to maximize the efficiency of scalable parallel scientific computing applications. These new systems develop custom runtime and operating systems software to maximize the integration of heterogeneous coprocessors and I/O devices, using global system knowledge to efficiently schedule tasks between system components, making it possible to develop applications which fully utilize system resources without imposing an undue hardship on the application developer. The proposed minisymposium gathers 4 talks, covering systems software support for future heterogeneous architectures, specifically new Operating Systems, proving runtimes correct, the challenges of heterogeneity and future systems architectures and their systems software. The objective is to summarize the latest developments in data-centric operating systems and runtime research and how the capabilities they provide relate to the construction of scientific applications. The goal is to provide applications researchers a view of future systems software and how these new capabilities relate to the construction of scientific applications on future data-centric systems.

Organizer: Noah Evans
Sandia National Laboratories, U.S.

10:40-11:00 Oversubscription and Your Data, How User Level Scheduling Can Increase Data Flow
Noah Evans, Sandia National Laboratories, U.S.

11:05-11:25 Data-Centric Operating Systems
Arthur Mccabe, Oak Ridge National Laboratory, U.S.

11:30-11:50 Aml: Building Blocks for Memory Management
Swann Perarnau, Argonne National Laboratory, U.S.

11:55-12:15 Operating System Support for Intelligent Management of Heterogeneous Memory
Balazs Gerofi, RIKEN, Japan

Saturday, February 15

MS64

Particle Methods: Algorithms and Software Technology for Exascale - Part II of III

10:40 a.m.-12:20 p.m.

Room: 507

For Part 1 see MS54
For Part 3 see MS73
Computational algorithms that use particles are ubiquitous in high performance computing. In a given algorithm, particles may represent physical particles as in molecular dynamics algorithms, they may represent ensembles of physical particles such as in some variations of particle-in-cell methods, or they may represent a discrete computational element of the continuum in Lagrangian or hybrid Lagrangian-Eulerian computational schemes for fluids and mechanics. Within the Exascale Computing Project (ECP), a large project sponsored by the U.S. Department of Energy for the development of exascale hardware, software, and science applications, numerous groups are developing particle methods targeting forthcoming exascale machines with a focus on performance and portability. In this minisymposium we will present software libraries and frameworks currently under development in ECP that are designed for deploying particle methods at scale, we will discuss algorithmic developments and performance engineering of particle methods along with their unique challenges, and then finally we will cover several of the ECP applications themselves, many of which use the aforementioned software libraries and frameworks. ECP particle applications covered will include particle-in-cell algorithms, molecular dynamics, as well as other Lagrangian and hybrid Lagrangian-Eulerian schemes with specific applications to linear accelerator modeling, plasma physics, cosmology, and multi-phase flow.

Organizer: Stuart Slattery
Oak Ridge National Laboratory, U.S.

10:40-11:00 The HACC Code: Particle Methods for Large-Scale Cosmological Simulations
Salman Habib, Argonne National Laboratory, U.S.

continued in next column

11:05-11:25 MFIX-Exa: An Exascale CFD-DEM Model for Reactor Design Engineering
Jordan Musser and William Fullmer, National Energy Technology Laboratory, U.S.; Ann S. Almgren, Lawrence Berkeley National Laboratory, U.S.

11:30-11:50 Lagrangian Particle Methods for Exascale Simulation of Dilute Sprays in Combustion Systems
Wenjun Ge and Ramanan Sankaran, Oak Ridge National Laboratory, U.S.

11:55-12:15 Exascale Molecular Dynamics with Cabana: from Lennard-Jones to Neural Network Potentials
Samuel Reeve, Lawrence Livermore National Laboratory, U.S.; Saaketh Desai, Purdue University, U.S.; Jim Belak, Lawrence Livermore National Laboratory, U.S.

Saturday, February 15

MS65

High-Performance Tensor Computation and Applications - Part II of III

10:40 a.m.-11:55 a.m.

Room: 508

For Part 1 see MS55
For Part 3 see MS74

Tensors are higher order generalization of matrices that provide a natural way to represent a multi-relational dataset. Given a dataset encoded as a tensor, tensor decomposition serves as a promising analytics tool for mining this data to uncover hidden structure within the data's relations. This minisymposium explores efficient and scalable solutions for calculating tensor decomposition, as well as its application in data analytics across areas spanning signal processing, cybersecurity, machine learning, and beyond.

Organizer: Jee W. Choi
University of Oregon, U.S.

Organizer: Richard Vuduc
Georgia Institute of Technology, U.S.

Organizer: Eric Phipps
Sandia National Laboratories, U.S.

10:40-11:00 Scaling Up Streaming Tensor Decompositions
Shaden Smith, Intel, AI, U.S.

11:05-11:25 Hpc_td_tbd_battaglino
Casey Battaglino, Georgia Institute of Technology, U.S.

11:30-11:50 Streaming Cp Tensor Decompositions
Tamara G. Kolda, Sandia National Laboratories, U.S.

Saturday, February 15

MS66

New Approaches for Software Auto-Tuning and Accuracy Assurance - Part I of II

10:40 a.m.-12:20 p.m.

Room: 512

For Part 2 see MS75

Exascale computers are expected to be deployed in the next several years. Their architectures will be complex, building uponmulti-core units, and offering unprecedented levels of parallelism. It is expected that auto-tuning (AT) research and technology will continue building upon its proven success for delivering high performance on a variety of computer architectures, thus enabling optimized, high performance implementations of specific computations for those challenging architectures. While performance is a major goal, in many conventional numerical computations accuracy is not well-assured. Here, we can think of the Basic Linear Algebra Subprograms (BLAS), which are often used as kernels in more complex linear algebra computations. The BLAS are carefully optimized for speed (e.g. by computer vendors), but the accuracy of the results is somewhat neglected. In fact, ensuring the accuracy of the computational results of BLAS operations is recurrent and crucial problem. Accuracy assurance becomes more difficult for more complex algorithms, for example for eigensolvers, which are essential parts in a number of applications of interest. This minisymposium will discuss state-of-the-art technologies for AT, accuracy assurance, and innovative algorithms that target extreme levels of computing.

Organizer: Osni A. Marques
Lawrence Berkeley National Laboratory, U.S.

Organizer: Toshiyuki Imamura
RIKEN, Japan

Organizer: Takahiro Katagiri
Nagoya University, Japan

10:40-11:00 Autotuning Exascale Applications
Wissam M. Sid-Lakhdar, Yang Liu, Xiaoye S. Li, and Osni A. Marques, Lawrence Berkeley National Laboratory, U.S.; James Demmel, University of California, Berkeley, U.S.

11:05-11:25 Verified Numerical Computations for Eigenvalue Problems on Large-Scale Parallel Systems
Takeshi Terao and Katsuhisa Ozaki, Shibaura Institute of Technology, Japan; Takeshi Ogita, Tokyo Woman's Christian University, Japan

11:30-11:50 An Efficient Contour Integral Based Eigensolver for Surface Plasmon Simulations
Huang Tsung-Ming, National Taiwan Normal University, Taiwan; Weichung Wang, National Taiwan University, Taiwan; Wen-Wei Lin, National Chiao Tung University, Taiwan; William Liao, National Taiwan University, Taiwan

11:55-12:15 Minimal-Precision Computing for High-Performance, Energy-Efficient, and Reliable Computations
Daichi Mukunoki, RIKEN, Japan

continued in next column

Saturday, February 15

MS67

Parallel Eigenvalue Algorithms for Physical Simulation - Part II of III

10:40 a.m.-12:20 p.m.

Room: 603

For Part 1 see MS57
For Part 3 see MS76

Large scale eigenvalue computations are ubiquitous throughout scientific computation. However, the steep $O(N^3)$ scaling of traditional eigenvalue algorithms often lead to the solution of a particular eigenvalue problem becoming the computational bottleneck in many simulations of physical systems. Over the years, many important algorithmic developments have been made to allow leverage of the latest advances in massively parallel computing architectures to enable the simulation of large physical systems on the world's largest supercomputers. In this minisymposium, we examine several recent advances in parallel eigenvalue algorithms for eigenvalue problems which arise in scientific computation.

Organizer: Roel Van Beeumen
Lawrence Berkeley National Laboratory, U.S.

Organizer: David B. Williams-Young
Lawrence Berkeley National Laboratory, U.S.

Organizer: Chao Yang
Lawrence Berkeley National Laboratory, U.S.

10:40-11:00 Parallel Computation of Many Eigenvalues by S-Step Thick-Restart Lanczos Algorithm with Explicit External Deflation - II
Zhaojun Bai, University of California, Davis, U.S.; Jack J. Dongarra, University of Tennessee and Oak Ridge National Laboratory, U.S.; Chao-Ping Lin and Ding Lu, University of California, Davis, U.S.; Ichitaro Yamazaki, Sandia National Laboratories, U.S.

11:05-11:25 Eigensolvers for Ab Initio CI Calculations in Nuclear Physics
Pieter Maris, Iowa State University, U.S.

continued in next column

11:30-11:50 Adaptive Step Size Strategies for Line Search Methods and their Applications to Electronic Structure Calculations
Xiaoying Dai, Chinese Academy of Sciences, China

11:55-12:15 A Scalable Matrix-Free Eigensolver for Studying Many-Body Localization
Roel Van Beeumen, Lawrence Berkeley National Laboratory, U.S.; Gregory D. Meyer and Norman Y. Yao, University of California, Berkeley, U.S.; Chao Yang, Lawrence Berkeley National Laboratory, U.S.

Saturday, February 15

MS68

Advanced HPC Trends Oil and Gas Applications - Part I of II

10:40 a.m.-12:20 p.m.

Room: 604

For Part 2 see MS77

For decades, advances in high performance computing have led to more accurate, safer, and faster oil and gas exploration and production processes. Applications range from seismic imaging and reservoir simulations to seismic interpretation and digital rock physics. This evolution is paramount today to further develop production by exploring increasingly complex reservoirs and by enhancing the recovery ratios from existing fields. In fact, these requirements translate into larger volumes of data to process and models of higher fidelity in terms of physical formulation and space and time resolutions. To cope with these challenges, the oil and gas industry has to adapt constantly to a changing technology landscape in terms of algorithms, platforms, software, and tools. The move towards less synchrony at all system levels, the widening gap between cost of IO versus floating-point operations, and the recent advent of Deep Learning are a few of these important changes the oil and gas software industry has to consider in moving forward with exascale. This minisymposium is an opportunity to discuss today's and future trends defining oil and gas HPC applications design.

Organizer: Rached Abdelkhalak
King Abdullah University of Science & Technology (KAUST), Saudi Arabia

Organizer: Stefano Zampini
King Abdullah University of Science & Technology (KAUST), Saudi Arabia

10:40-11:00 Alleviating the Memory Pressure for Seismic Modeling and Imaging
Rached Abdelkhalak, Hatem Ltaief, and David E. Keyes, King Abdullah University of Science & Technology (KAUST), Saudi Arabia

continued on next page

11:05-11:25 Devito - DSL for Generating MPI and OpenMP Parallel Finite Difference Operators for Seismic Imaging
Gerard J Gorman, Fabio Luporini, and Rhodri Nelson, Imperial College London, United Kingdom

11:30-11:50 Porting Sewas Task-Based Seismic Code on Arm Platforms
Fabrice Dupros, ARM, France; Salli Moustafa, Aneo, France; Conrad Hillairet, ARM, France

11:55-12:15 A Library to Accelerate Stencil Codes on Vector Processors
Arihiro Yoshida, NEC Corporation, Japan

Saturday, February 15

MS69

The Many Faces of Simulation for HPC - Part I of II

10:40 a.m.-12:20 p.m.

Room: 605

For Part 2 see MS78

In the field of HPC research and development, simulation has mainly been used for the purpose of evaluating and comparing the performance of application implementations and of the algorithms therein. While this use remains critical, for good reasons, many other compelling use cases have emerged. These have often been made possible by recent advances in the simulation methodologies at the core of available simulation frameworks. Examples of new areas in which simulation has become a compelling proposition include debugging and verification, application/simulation co-design, and HPC education. In this multi-part minisymposium, we bring together researchers who have contributed to traditional and explored emerging uses of simulation of HPC systems and applications. The objective is for them to share their experiences, present recent results, identify areas of convergence, and discuss future directions.

Organizer: Rafael Ferreira da Silva
University of Southern California, U.S.

Organizer: Frédéric Suter
CNRS, France

10:40-11:00 The Many Faces of Simulation for HPC
Frédéric Suter, CNRS, France; Rafael Ferreira da Silva, University of Southern California, U.S.

11:05-11:25 Teaching Parallel and Distributed Computing Concepts in Simulation
Henri Casanova, University of Hawaii, U.S.; Ryan Tanaka, ; Rafael Ferreira da Silva, University of Southern California, U.S.

11:30-11:50 Fast and Faithful Performance Prediction of MPI Applications: the HPL Case Study
Tom Cornebize, Université Grenoble Alpes, France; Arnaud Legrand, CNRS, France; Franz Christian Heinrich, Inria, France

11:55-12:15 Power-Aware Scheduling with Slurm: Simulation and Practice
Tapasya Patki, Lawrence Livermore National Laboratory, U.S.

Saturday, February 15

MS70

Resilience and Fault Tolerance for Extreme Computing Systems - Part II of III

10:40 a.m.-12:20 p.m.

Room: 606

For Part 1 see MS59
For Part 3 see MS79

The reliability of large scale computing systems has been a major concern of high performance computing due to the ever increasing complexity of hardware and software components combined with the tight power budget for the operations. Under such unreliable computing systems, it is essential to introduce failure mitigations at the runtime and application layers to complement the resilience at the hardware/system levels. Today, the major resilience scheme is coordinated checkpoint and restart (CR) that involves global coordination of processes and threads. This global recovery model entails inherent scalability issues to handle a large class of errors and failures, and thus alternative approaches would improve the scalability and reliability of computing systems and applications together. In this minisymposium, we will discuss the reliability issues of the large scale computing systems, application/algorithm based resilience and programming model support to enhance the reliability of application program executions. We will also cover the recent global CR techniques that mitigates the performance interference to applications from checkpointing operations.

Organizer: Keita Teranishi
Sandia National Laboratories, U.S.

Organizer: Christian Engelmann
Oak Ridge National Laboratory, U.S.

Organizer: George Bosilca
University of Tennessee, Knoxville, U.S.

10:40-11:00 New Non-Blocking Extensions to the ULFM Proposal
George Bosilca and Aur?lien Bouteiller, University of Tennessee, Knoxville, U.S.; Nuria Losada, University of A Coru?a, Spain

continued on next page

Saturday, February 15

MS70

Resilience and Fault Tolerance for Extreme Computing Systems - Part II of III

10:40 a.m.-12:20 p.m.

continued

11:05-11:25 Performance Portable and Productive Resilience using Kokkos

Jeffery Miles and Nicholas Morales, Sandia National Laboratories, U.S.; Carson Mould, Georgia Institute of Technology, U.S.; Bogdan Nicolae, Argonne National Laboratory, U.S.; Keita Teranishi, Sandia National Laboratories, U.S.

11:30-11:50 Composing Asynchrony, Communication and Resilience

Sri Raj Paul, Georgia Institute of Technology, U.S.; Akihiro Hayashi, Rice University, U.S.; Nicole Slattengren and Hemanth Kolla, Sandia National Laboratories, U.S.; Seonmyeong Bak and Matthew Whitlock, Georgia Institute of Technology, U.S.; Jackson Mayo and Keita Teranishi, Sandia National Laboratories, U.S.; Vivek Sarkar, Georgia Institute of Technology, U.S.; Max Grossman, Rice University, U.S.

11:55-12:15 Resilience in the Context of GPUs: A Technique for Interrupting GPU Kernels

Max M. Baird, Heriot-Watt University, United Kingdom

Saturday, February 15

CP12 - Session Cancelled

Communication Performance

10:40 a.m.-12:20 p.m.

Lunch Break

12:20 p.m.-1:50 p.m.

Attendees on their own

Saturday, February 15

MS71

Toward Efficient Software Integration and Deployment

1:50 p.m.-3:30 p.m.

Room: Elwha Ballroom

With the increasing complexity and diversity of the software stack and system architecture of high performance computing (HPC) systems, the traditional HPC community is facing a huge productivity challenge in software building, integration and deployment for multiple exascale computing systems that will be deployed in year 2020 and after. Recently, this challenge has been addressed by new software build management tools such as Spack and Guix that enable seamless software building and integration. Despite these efforts, we are wanting in software testing, verification and deployment methodologies and tools to complement the capability of these new software integration tools. Effective automation of these processes will reduce the code development time and the system resource usage of exascale systems for preparing applications for future systems. In this minisymposium, we will discuss the recent efforts and techniques to improve software integration process, testing and deployment for HPC platforms. We will cover the issues on software development practice, use of containers and automation of performance testing and software building.

Organizer: Keita Teranishi
Sandia National Laboratories, U.S.

1:50-2:10 Extreme-Scale Scientific Software Stack (E4S)

Sameer Shende, University of Oregon, U.S.

2:15-2:35 XSDK: Toward Efficient and Interoperable Scientific Library Collection

Piotr Luszczek, University of Tennessee, Knoxville, U.S.; Satish Balay, Argonne National Laboratory, U.S.; Keita Teranishi and James Willenbring, Sandia National Laboratories, U.S.; Lois Curfman McInnes, Argonne National Laboratory, U.S.; Ulrike Meier Yang, Lawrence Livermore National Laboratory, U.S.; *Asim YarKhan*, University of Tennessee, Knoxville, U.S.

continued on next page

2:40-3:00 Towards Continuous Benchmarking: An Automated Performance Evaluation Framework for High Performance Software

Hartwig Anzt, Karlsruhe Institute of Technology, Germany and University of Tennessee, U.S.; *Terry Cojean*, Karlsruhe Institute of Technology, Germany; Goran Flegar, Universitat Jaume I, Spain; Pratik Nayak and Yuhsiang M. Tsai, Karlsruhe Institute of Technology, Germany

3:05-3:25 Faster Spack Package Manager Installations through Task Parallelism

Samuel Knight and Jeremiah Wilke, Sandia National Laboratories, U.S.; Todd Gamblin, Lawrence Livermore National Laboratory, U.S.

Saturday, February 15

MS72

Novel Computational Algorithms for Future Computing Platforms - Part III of III

1:50 p.m.-3:30 p.m.

Room: 501

For Part 2 see MS62

In the early 2000s, due to constraints on economical heat dissipation, clock speeds of single-core CPUs could no longer be increased, which marked the adoption of multi-core CPUs, together with a paradigm shift to algorithms specifically designed for multi-core architectures. About 15 years into this current architectural cycle and on its way to exascale performance, the computing industry finds itself at the confluence of technical difficulties that cast doubt on its ability to sustain this architectural model beyond the exascale capability. These difficulties are driving the hardware industry to develop application-specific chips and to look beyond silicon- based chips (e.g., quantum computing, physical annealing, neuromorphics, etc.), with a continued emphasis on raw processing power and emerging concerns about energy efficiency. Hardware specialization will likely redefine the way computational algorithms are developed over the next two decades for a wide range of important applications: large-scale PDE-based problems (CFD, wave propagation, subsurface modeling, etc.), artificial intelligence, computational chemistry, and optimization problems, to name a few. The pressure to decrease time to solution or improve simulation fidelity, or both, for these applications will continue unabated. This minisymposium provides a forum for sharing innovative ideas on algorithm development for leveraging future computing platforms.

Organizer: Arash Fathi
ExxonMobil Research and Engineering, U.S.

Organizer: Dimitar Trenev
ExxonMobil Research and Engineering

Organizer: Jason Riedy
Georgia Institute of Technology, U.S.

Organizer: Jeffrey Young
Georgia Institute of Technology, U.S.

Organizer: Laurent White
ExxonMobil Research and Engineering

1:50-2:10 Leveraging Random Walks and Neuromorphic Hardward to Solve Elliptical Integro-PDEs

Brad Aimone and *J. Darby Smith*, Sandia National Laboratories, U.S.

2:15-2:35 Adaptive Local Timestepping and its Parallelization

Max Bremer, University of Texas at Austin, U.S.; John Bachan, Lawrence Berkeley National Laboratory, U.S.; Cy Chan, Massachusetts Institute of Technology, U.S.; Clint Dawson, University of Texas at Austin, U.S.

2:40-3:00 The Rogues Gallery as a Testbed for Novel Algorithm Design for Future Architectures

Jeffrey Young and Jason Riedy, Georgia Institute of Technology, U.S.; Thomas M. Conte, ; Vivek Sarkar, Georgia Institute of Technology, U.S.

3:05-3:25 Design of New Streaming and Graph Analytics Algorithms for the Strider Architecture

Sriseshan Srikanth, Georgia Institute of Technology, U.S.

continued in next column

Saturday, February 15

MS73

Particle Methods: Algorithms and Software Technology for Exascale - Part III of III

1:50 p.m.-3:30 p.m.

Room: 507

For Part 2 see MS64

Computational algorithms that use particles are ubiquitous in high performance computing. In a given algorithm, particles may represent physical particles as in molecular dynamics algorithms, they may represent ensembles of physical particles such as in some variations of particle-in-cell methods, or they may represent a discrete computational element of the continuum in Lagrangian or hybrid Lagrangian-Eulerian computational schemes for fluids and mechanics. Within the Exascale Computing Project (ECP), a large project sponsored by the U.S. Department of Energy for the development of exascale hardware, software, and science applications, numerous groups are developing particle methods targeting forthcoming exascale machines with a focus on performance and portability. In this minisymposium we will present software libraries and frameworks currently under development in ECP that are designed for deploying particle methods at scale, we will discuss algorithmic developments and performance engineering of particle methods along with their unique challenges, and then finally we will cover several of the ECP applications themselves, many of which use the aforementioned software libraries and frameworks. ECP particle applications covered will include particle-in-cell algorithms, molecular dynamics, as well as other Lagrangian and hybrid Lagrangian-Eulerian schemes with specific applications to linear accelerator modeling, plasma physics, cosmology, and multi-phase flow.

Organizer: Stuart Slattery
Oak Ridge National Laboratory, U.S.

1:50-2:10 WarpX: Electromagnetic Particle-in-Cell with Adaptive Mesh Refinement for Advanced Particle Accelerators

Maxence Thevenet, Jean-Luc Vay, Weiqun Zhang, Rémi Lehe, Andrew Myers, Ann S. Almgren, and John B. Bell, Lawrence Berkeley National Laboratory, U.S.; David Grote, Lawrence Livermore National Laboratory, U.S.; Olga Shapoval, Axel Huebl, Jaehong Park, and Ligia Diana Amorim, Lawrence Berkeley National Laboratory, U.S.; Mark Hogan, SLAC National Accelerator Laboratory, U.S.; Lixin Ge and Cho-Kuen Ng, Stanford Linear Accelerator Center, U.S.

2:15-2:35 Particle-in-Cell Simulations at Exascale

Robert F. Bird, Los Alamos National Laboratory, U.S.

2:40-3:00 Taking the Plasma Physics Code XGC to Summit and Beyond with Kokkos/Cabana

Aaron Scheinberg and Stephane Ethier, Princeton Plasma Physics Laboratory, U.S.; Guangye Chen and Bob Bird, Los Alamos National Laboratory, U.S.; Stuart Slattery, Oak Ridge National Laboratory, U.S.; CS Chang, Princeton Plasma Physics Laboratory, U.S.

3:05-3:25 Pumipic: Infrastructure for Unstructured Mesh Pic on GPUs

Gerrett Diamond, Cameron W. Smith, Chonglin Zhang, Eisung Yoon, Gopan Perumpilly, Onkar Sahni, and Mark S. Shephard, Rensselaer Polytechnic Institute, U.S.

continued in next column

Saturday, February 15

MS74

High-Performance Tensor Computation and Applications - Part III of III

1:50 p.m.-3:30 p.m.

Room: 508

For Part 2 see MS65

Tensors are higher order generalization of matrices that provide a natural way to represent a multi-relational dataset. Given a dataset encoded as a tensor, tensor decomposition serves as a promising analytics tool for mining this data to uncover hidden structure within the data's relations. This minisymposium explores efficient and scalable solutions for calculating tensor decomposition, as well as its application in data analytics across areas spanning signal processing, cybersecurity, machine learning, and beyond.

Organizer: Jee W. Choi
University of Oregon, U.S.

Organizer: Richard Vuduc
Georgia Institute of Technology, U.S.

Organizer: Eric Phipps
Sandia National Laboratories, U.S.

1:50-2:10 Computing Generalized CP Decompositions on Emerging Parallel Architectures

Eric Phipps and Tamara G. Kolda, Sandia National Laboratories, U.S.

2:15-2:35 Communication Lower Bounds for Rectangular MTTKRPS

Grey Ballard, Wake Forest University, U.S.; Kathryn Rouse, Inmar, USA

2:40-3:00 Load Balancing Strategy of Parallel Performance Portable Sparse CP-APR Decomposition

Keita Teranishi and David S. Hollman, Sandia National Laboratories, U.S.; Jeremy Myers, The College of William & Mary, U.S.; Richard Barrett and Daniel M. Dunlavy, Sandia National Laboratories, U.S.

3:05-3:25 A Distributed Memory Generalized Sparse Tensor Decomposition

Karen D. Devine, Sandia National Laboratories, U.S.; Grey Ballard, Wake Forest University, U.S.; Tamara G. Kolda, Sandia National Laboratories, U.S.

Saturday, February 15

MS75

New Approaches for Software Auto-Tuning and Accuracy Assurance - Part II of II

1:50 p.m.-3:30 p.m.

Room: 512

For Part 1 see MS66

Exascale computers are expected to be deployed in the next several years. Their architectures will be complex, building uponmulti-core units, and offering unprecedented levels of parallelism. It is expected that auto-tuning (AT) research and technology will continue building upon its proven success for delivering high performance on a variety of computer architectures, thus enabling optimized, high performance implementations of specific computations for those challenging architectures. While performance is a major goal, in many conventional numerical computations accuracy is not well-assured. Here, we can think of the Basic Linear Algebra Subprograms (BLAS), which are often used as kernels in more complex linear algebra computations. The BLAS are carefully optimized for speed (e.g. by computer vendors), but the accuracy of the results is somewhat neglected. In fact, ensuring the accuracy of the computational results of BLAS operations is recurrent and crucial problem. Accuracy assurance becomes more difficult for more complex algorithms, for example for eigensolvers, which are essential parts in a number of applications of interest. This minisymposium will discuss state-of-the-art technologies for AT, accuracy assurance, and innovative algorithms that target extreme levels of computing.

Organizer: Toshiyuki Imamura
RIKEN, Japan

Organizer: Osni A. Marques
Lawrence Berkeley National Laboratory, U.S.

Organizer: Takahiro Katagiri
Nagoya University, Japan

1:50-2:10 Precision Tuning of the Arithmetic Units in Matrix Multiplication on FPGA
Toshiyuki Imamura and *Yiyu Tan*, RIKEN, Japan

2:15-2:35 Reproducible Linear Algebra from Application to Architecture
Jason Riedy, Georgia Institute of Technology, U.S.; James W. Demmel, University of California, Berkeley, U.S.; Peter Ahrens, Massachusetts Institute of Technology, U.S.

2:40-3:00 Stable Automatic Tuning Method for Performance Fluctuation
Naoto Seki, Toshiki Tabeta, Akihiro Fujii, and Teruo Tanaka, Kogakuin University, Japan

3:05-3:25 Search Space Reduction Through Analytical Modeling
Tze Meng Low, Carnegie Mellon University, U.S.

continued in next column

Saturday, February 15

MS76

Parallel Eigenvalue Algorithms for Physical Simulation - Part III of III

1:50 p.m.-3:30 p.m.

Room: 603

For Part 2 see MS67

Large scale eigenvalue computations are ubiquitous throughout scientific computation. However, the steep $O(N^3)$ scaling of traditional eigenvalue algorithms often lead to the solution of a particular eigenvalue problem becoming the computational bottleneck in many simulations of physical systems. Over the years, many important algorithmic developments have been made to allow leverage of the latest advances in massively parallel computing architectures to enable the simulation of large physical systems on the world's largest supercomputers. In this minisymposium, we examine several recent advances in parallel eigenvalue algorithms for eigenvalue problems which arise in scientific computation.

Organizer: Roel Van Beeumen
Lawrence Berkeley National Laboratory, U.S.

Organizer: David B. Williams-Young
Lawrence Berkeley National Laboratory, U.S.

Organizer: Chao Yang
Lawrence Berkeley National Laboratory, U.S.

1:50-2:10 A Parallel Strategy for Kohn-Sham Solver with GPU-Accelerated Nodes
Jean-Luc Fattebert, Luigi Capone, Massimiliano Lupo Pasini, and Bruno Turcksin, Oak Ridge National Laboratory, U.S.

2:15-2:35 Scalable Kohn-Sham Matrix Algebra Solutions with the ELSI Infrastructure
Volker Blum, Duke University, U.S.

2:40-3:00 Real-Space Parallel Eigensolvers for Electronic Structure Calculations: The Future
James R. Chelikowsky, University of Texas at Austin, U.S.

continued on next page

Saturday, February 15

MS76

Parallel Eigenvalue Algorithms for Physical Simulation - Part III of III

1:50 p.m.-3:30 p.m.

continued

3:05-3:25 Revisiting the Jacobi Method for Eigen Problems in Computational Chemistry

Hua Huang, Georgia Institute of Technology, U.S.

Saturday, February 15

MS77

Advanced HPC Trends for Oil and Gas Applications - Part II of II

1:50 p.m.-3:30 p.m.

Room: 604

For Part 1 see MS68

For decades, advances in high performance computing have led to more accurate, safer, and faster oil and gas exploration and production processes. Applications range from seismic imaging and reservoir simulations to seismic interpretation and digital rock physics. This evolution is paramount today to further develop production by exploring increasingly complex reservoirs and by enhancing the recovery ratios from existing fields. In fact, these requirements translate into larger volumes of data to process and models of higher fidelity in terms of physical formulation and space and time resolutions. To cope with these challenges, the oil and gas industry has to adapt constantly to a changing technology landscape in terms of algorithms, platforms, software, and tools. The move towards less synchrony at all system levels, the widening gap between cost of IO versus floating-point operations, and the recent advent of Deep Learning are a few of these important changes the oil and gas software industry has to consider in moving forward with exascale. This minisymposium is an opportunity to discuss today's and future trends defining oil and gas HPC applications design.

Organizer: Rached Abdelkhalak
King Abdullah University of Science & Technology (KAUST), Saudi Arabia

Organizer: Stefano Zampini
King Abdullah University of Science & Technology (KAUST), Saudi Arabia

1:50-2:10 Large Scale Inversion of CSEM Data in the Time Domain

Stefano Zampini, King Abdullah University of Science & Technology (KAUST), Saudi Arabia; George M. Turkiyyah, American University of Beirut, Lebanon; David E. Keyes, King Abdullah University of Science & Technology (KAUST), Saudi Arabia

2:15-2:35 Revisiting Seismic Data Management with Distributed Asynchronous Object Storage (DAOS)

Essam Algizawy, Brightskies Inc, Egypt; Philippe Thierry, Intel Corporation, France; Mohamad Chaarawi, Intel Corporation, U.S.; Johann Lombardi, Intel, France; Khaled El-Amrawi, Brightskies Inc, Egypt

2:40-3:00 Hiding I/O Latency in Large-Scale Scientific Computation Through Buffering and Prefetching on Advanced Storage Technologies

Tariq Alturkestani and David E. Keyes, King Abdullah University of Science & Technology (KAUST), Saudi Arabia

3:05-3:25 Distributed ML-Based Applications in Oil & Gas

Mauricio Araya-Polo and Denis Akhiyarov, Total E&P, U.S.

continued in next column

Saturday, February 15

MS78

The Many Faces of Simulation for HPC - Part II of II

1:50 p.m.-3:05 p.m.

Room: 605

For Part 1 see MS69

In the field of HPC research and development, simulation has mainly been used for the purpose of evaluating and comparing the performance of application implementations and of the algorithms therein. While this use remains critical, for good reasons, many other compelling use cases have emerged. These have often been made possible by recent advances in the simulation methodologies at the core of available simulation frameworks. Examples of new areas in which simulation has become a compelling proposition include debugging and verification, application/simulation co-design, and HPC education. In this multi-part minisymposium, we bring together researchers who have contributed to traditional and explored emerging uses of simulation of HPC systems and applications. The objective is for them to share their experiences, present recent results, identify areas of convergence, and discuss future directions.

Organizer: Rafael Ferreira da Silva
University of Southern California, U.S.

Organizer: Frédéric Suter
CNRS, France

1:50-2:10 Faithful Performance Prediction of a Dynamic Task-Based Runtime System, an Opportunity for Task Graph Scheduling

Samuel Thibault, LaBRI, France; Luka Stanisic, Inria Bordeaux Sud-Ouest, France; Arnaud Legrand, CNRS, France; Brice Videau, INRIA Grenoble Rh?ne-Alpes, France; Jean-François M?haut, Universite Joseph Fourier, France

2:15-2:35 New Horizons for Debugging Long-Running Parallel Programs: DMTCP and SimGrid

Gene Cooperman and Rohan Garg, Northeastern University, U.S.

2:40-3:00 Application-Simulation Co-Design for Performance and Correctness Evaluation

Luigi Genovese, CEA, France; Augustin Degomme, CEA Grenoble

Saturday, February 15

MS79

Resilience and Fault Tolerance for Extreme Computing Systems - Part III of III

1:50 p.m.-3:30 p.m.

Room: 606

For Part 2 see MS70

The reliability of large scale computing systems has been a major concern of high performance computing due to the ever increasing complexity of hardware and software components combined with the tight power budget for the operations. Under such unreliable computing systems, it is essential to introduce failure mitigations at the runtime and application layers to complement the resilience at the hardware/system levels. Today, the major resilience scheme is coordinated checkpoint and restart (CR) that involves global coordination of processes and threads. This global recovery model entails inherent scalability issues to handle a large class of errors and failures, and thus alternative approaches would improve the scalability and reliability of computing systems and applications together. In this minisymposium, we will discuss the reliability issues of the large scale computing systems, application/algorithm based resilience and programming model support to enhance the reliability of application program executions. We will also cover the recent global CR techniques that mitigates the performance interference to applications from checkpointing operations.

Organizer: Keita Teranishi
Sandia National Laboratories, U.S.

Organizer: Christian Engelmann
Oak Ridge National Laboratory, U.S.

Organizer: George Bosilca
University of Tennessee, Knoxville, U.S.

1:50-2:10 Modern Use of Checkpoint-Restart at Large Scale using VeloC

Bogdan Nicolae, Argonne National Laboratory, U.S.; Adam Moody and Kathryn Mohror, Lawrence Livermore National Laboratory, U.S.; Franck Cappello, Argonne National Laboratory, U.S.

2:15-2:35 Robust Dynamic Load Balancing of Scientific Applications against System Variability and Failures

Ali Mohammed, Aurélien Cavelan, and *Florina M. Ciorba*, University of Basel, Switzerland

2:40-3:00 Silent-Error Detection, Local Recovery, and Failure Masking in MPI-Based Solvers

Hemanth Kolla, *Jackson Mayo*, and Rob Armstrong, Sandia National Laboratories, U.S.

3:05-3:25 Space-Efficient Reed-Solomon Encoding to Detect and Correct Pointer Corruption

Scott Levy and Kurt Ferreira, Sandia National Laboratories, U.S.

continued in next column

Saturday, February 15

CP13

Accelerating Software with GPUs

1:50 p.m.-3:30 p.m.

Room: 502

Chair: Nathan Vaughn, University of Michigan, U.S.

1:50-2:10 Accelerating Jacobi Iteration on GPUs using Shared Memory
Mohammad S. Islam and Qiqi Wang, Massachusetts Institute of Technology, U.S.

2:15-2:35 GPU-Accelerated Barycentric Treecodes
Nathan Vaughn and Leighton Wilson, University of Michigan, U.S.; Lei Wang, University of Wisconsin, Milwaukee, U.S.; Robert Krasny, University of Michigan, U.S.

2:40-3:00 GPU Acceleration of a High-Order Arbitrary Lagrangian-Eulerian Code
Arturo Vargas, Ryan Bleile, Jean-Sylvain Camier, Veselin Dobrev, and Robert Rieben, Lawrence Livermore National Laboratory, U.S.

3:05-3:25 Multi GPU Algorithms for Hierarchical Matrix Computations
Wajih Halim Boukaram, King Abdullah University of Science & Technology (KAUST), Saudi Arabia; George M. Turkiyyah, American University of Beirut, Lebanon; David E. Keyes, King Abdullah University of Science & Technology (KAUST), Saudi Arabia

Saturday, February 15

CP14

HPC for Data Science and Large Graphs

1:50 p.m.-3:05 p.m.

Room: 505

Chair: Oded Green, NVIDIA and Georgia Institute of Technology, U.S.

1:50-2:10 Introduction to High Performance Computing for Decision Makers: Creating a One-Day Workshop for a Data Science Audience
Cindy Orozco Bohorquez, Stanford University, U.S.

2:15-2:35 HashGraph - Scalable Hash Tables using A Sparse Graph Data Structure
Oded Green, NVIDIA and Georgia Institute of Technology, U.S.

2:40-3:00 Scalable Triangle Counting on Distributed-Memory Systems
Seher Acer, Sandia National Laboratories, U.S.; Abdurrahman Yasar, Georgia Institute of Technology, U.S.; Sivasankaran Rajamanickam and Michael Wolf, Sandia National Laboratories, U.S.; Umit V. Catalyurek, Georgia Institute of Technology, U.S.

Saturday, February 15

CP15

Highly Parallel Algorithms

1:50 p.m.-3:05 p.m.

Room: 607

Chair: Sergey Khilkov, Keldysh Institute for Applied Mathematics, RAS, Russia

1:50-2:10 Implementing Randomized Asynchronous Linear System Solvers
Masha Sosonkina, Old Dominion University and Ames Laboratory/ DOE, U.S.; Erik Jensen, Old Dominion University, U.S.; Evan Coleman, Naval Surface Warfare Center, U.S.

2:15-2:35 Asynchronous Smoothers for Computational Fluid Dynamics on Graphics Processing Units
Aditya Kashi and Siva Nadarajah, McGill University, Canada

2:40-3:00 Prismatic Space-Time Tiling as an Approach to High Efficient Stencil Computations
Sergey Khilkov, Anastasia Perepelkina, and Vadim Levchenko, Keldysh Institute for Applied Mathematics, RAS, Russia

Abstracts


siam
2020
Conference on
Parallel Processing for
Scientific Computing

IP1

Parallel Tomographic Reconstruction - Where Combinatorics Meets Geometry

Today, high-resolution tomographic reconstruction of 3D objects is within reach, but the associated data sets are huge and calling for parallel computation. A typical 3D reconstruction with 4k resolution already produces an image of 64 Gbytes. Tomographic reconstruction is often done using iterative algorithms that involve repeated sparse matrix-vector multiplication (SpMV). The matrix, however, may be too large to store, requiring Tbytes of memory, and hence each matrix row is recomputed upon use. In this talk, we present data partitioning methods for tomography matrices of increasing size. For small matrices, we can compute an optimal bipartitioning by an exact combinatorial method, as implemented in the packages MondriaanOpt and MP. This allows us to gauge the quality of medium-grain partitioning (default in the Mondriaan package), which is a heuristic combinatorial method that can handle larger problems. Medium-grain results in turn justified choosing row partitioning for the tomographic matrix-free SpMV. For this row partitioning, we developed a geometric recursive coordinate bisection algorithm with nearly the same output quality as combinatorial partitioning that can handle huge, matrix-free problems and is also faster. We conclude with showing an actual reconstruction that was written using Bulk, a modern C++ library for easy development of parallel programs in bulk-synchronous parallel style.

Rob H. Bisseling
Utrecht University
Mathematical Institute
r.h.bisseling@uu.nl

IP2

Accelerated-Node-Enabled Computational and Data Science: Its not just for Exascale

In just the past decade, heterogeneous node-based computing hardware and software architectures have moved from novelty to mainstream. Prototypical examples are the now ubiquitous accelerators, or GPU-based architectures designed to accelerate certain operations on data arranged in certain ways. This heterogeneity will soon be more extreme, where computing nodes will not just have GPU accelerators, but other application-specific accelerators such as those for key computational motifs, AI, and even quantum-based optimization. GPU-based accelerators on supercomputers in the US Department of Energy (DOE) began with the Roadrunner and Titan systems, then on to the current Summit and Sierra and planned Perlmutter systems, and finally for the first three US exascale systems (Aurora, Frontier, El Capitan). Accelerated node hardware and software architectures are not only here and now but represent our collective future. The US scientific community must be able to effectively exploit these architectures to address urgent problems of National importance. A key response by the DOE to this call to arms was the launching of the Exascale Computing Project in 2016, which is an aggressive RDD project focused on delivery of mission critical applications, an integrated software stack, exascale hardware technology advances and all within the context of an accelerated-node co-design software and algorithm paradigm that is not just for exascale but here to stay for the foreseeable future.

Douglas Kothe
Oak Ridge National Laboratory
kothe@ornl.gov

IP3

Development of an Eigen-Analysis Engine for Large-Scale Simulation and Big Data Analysis

Large-scale eigenvalue problems arise in a wide variety of scientific and engineering applications such as nano-scale materials simulation, vibration analysis of automobile components, data analysis, graph analysis, etc. In such computations, high-performance parallel eigensolvers are required to exploit distributed parallel computing environments. In this talk, we present a parallel eigensolver, the Sakurai-Sugiura method (SSM), for large-scale interior eigenvalue problems. This method is derived using numerical quadrature and has good parallel scalability. We also show a software package "z-Pares," which enables users to utilize a large number of computational resources because of its hierarchical parallel structure. While the presented technology is already well-established, its applications are yet to be fully explored. Hence this talk will also give an overview of challenging issues on the intersection of Big Data analysis and simulation that could be tackled with a scalable eigensolver.

Tetsuya Sakurai
Department of Computer Science
University of Tsukuba
sakurai@cs.tsukuba.ac.jp

IP4

Modeling of Heterogeneous Computing Systems and Their Usages

The last decade has seen a paradigm shift in the architecture of computing platforms, with a trend toward combining general-purpose processors and specialized accelerators. For example, GPUs have made a significant impact in both the hardware industry and the application domain, as has been seen from the recent development of machine learning applications. Other platforms such as Processing-In-Memory and FPGA-based reconfigurable architectures have regained attention, and with these special accelerators, the computing platforms become more heterogeneous. These heterogeneous architectures are especially attractive because they can provide high performance and energy efficiency for both general-purpose applications and high-throughput applications. Thus, from IoT devices to server processors, heterogeneous architectures have become increasingly popular. However, these heterogeneous architectures introduce several new challenges, including programmability issues and designing the hardware architecture in a way that can maximally exploit the underlying heterogeneity. To address these issues, a wide variety of modeling has been used including analytical, regression based, and cycle-level. In this talk I will discuss how different modeling techniques have been addressed and how they can be helpful to guide architecture studies and software optimizations.

Hyesoon Kim
Georgia Institute of Technology
hyesoon@cc.gatech.edu

IP5

Methods and Models for Reducing Communication

Parallel communication costs can lead to significant per-

formance limitations in many computational kernels. From structured operations across grids to sparse matrix computations to global reductions, understanding the roadblocks is critical in achieving the highest performance. In addition, new architectures and networks are exposing opportunities and challenges in organizing communication even in the most fundamental routines. In this talk we survey new approaches for modeling performance and offer several routes to reorganizing computation by increasing locality in a node aware fashion. As an example, algebraic multigrid methods are becoming robust solvers for a variety of problems. Yet the strong scaling limitations in the underlying sparse matrix routines can severely hinder performance. Likewise, a key feature of current architectures is node-level parallelism. Standard approaches to inter-process communication will send data regardless of the locations of send and receive processes. Yet, there are notable differences in the cost of intra- and inter-node communication. In response, communication can be reorganized to take advantage of the less costly intra-node communication, reducing both the number and size of inter-node messages. We will introduce some new models and new approaches to organizing communication that can significantly speed up solvers in this setting.

Luke Olson
University of Illinois Urbana-Champaign
lukeo@illinois.edu

IP6

Cognitive Discovery: Pushing the Frontier of Technical RD with AI

Cognitive Discovery is an overarching framework that uses AI to achieve scientific knowledge extraction and representation, to intelligently design and guide simulations, in order to drastically accelerate the pace of scientific discovery. Cognitive Discovery targets to accelerate scientific workflows in technical disciplines and provide a new generation of tools. The workflows follow the cycle: a) Massive literature review in order to understand the problem at hand. Literature refers to all aspects such as mathematical modelling, solution methods, actual computer models and HPC deployment. b) Enrichment of literature data with experimental data and formation of hypotheses. c) Running simulations to test hypotheses and generate new knowledge in order to close any knowledge gaps. All three phases suffer today major disruptions. Simply put: the volume of new literature is exploding (e.g. roughly 450K new publications in materials science are published every year, tens of thousands of papers in numerical and HPC methods need to be reviewed). IoT advances as well as advances in measuring all aspects of HPC systems create an explosion of data. High fidelity models lead to massive configuration spaces the complexity of which clearly outpaces our capability to scale and efficiently run modern HPC systems. We will showcase how AI can help dramatically improve this setting and lead to a massive acceleration for scientific discovery.

Costas Bekas
IBM Research - Zurich
bekas.costas@gmail.com

SP1

SIAG Best Paper Prize: The BLIS Framework: Experiments in Portability

Over the past decade, the BLAS-like Library Instantiation Software (BLIS) project has carefully revisited past progress on how to structure the implementation of the level-3 BLAS-like operations (matrix-matrix computations) in particular and all basic linear algebra operations in general. The paper "The BLIS Framework: Experiments in Portability" demonstrates how a refactoring of prior approaches yields a more flexible, more easily maintained, highly portable, yet high-performing and scalable software library. Casting level-3 BLAS functionality in terms of multiplication with submatrices was proposed in the works of Kgstrm, Ling, and Van Loan. This led to efforts to auto-generate and tune such as PHiPAC and ATLAS. A dramatic breakthrough came circa 2000 when Goto proposed "Goto's algorithm" (now at the heart of most high-performance BLAS) for implementing matrix-matrix multiplication. BLIS casts Goto's algorithm in terms of five portable loops (written in C99) around a "microkernel" that updates a small submatrix of C that fits in registers. It is only this microkernel that needs to be customized for a new architecture when implementing matrix multiplication. The refactoring exposed in BLIS drastically reduced the size, complexity, and number of assembly kernels necessary for supporting high-performance across all datatypes and level-3 operations. The prize-winning paper was coauthored with Tyler Smith, Bryan Marker, Tze Meng Low, Francisco Igual, Mikhael Smelyanskiy, Xianyi Zhang, Michael Kistler, Vernon Austel, John Gunnels, and Lee Killough.

Field G. Van Zee
Department of Computer Sciences
The University of Texas at Austin
field@cs.utexas.edu

Robert A. van de Geijn
The University of Texas at Austin
Department of Computer Science
rvdg@cs.utexas.edu

SP2

SIAG/Supercomputing Early Career Prize: Scalable Algorithms for Tensor Computations

We give a high-level overview of a few recent algorithmic advances in the design of efficient algorithms for tensor computations. We highlight a communication-avoiding parallel algorithm for dense symmetric eigenvalue problems, which is part of a broader family of new matrix factorization algorithms with costs that attain communication lower bounds. One recently developed practical variant of these algorithms, a 3D parallel algorithm for CholeskyQR2, achieves speed-ups of up to 3.3X over the best existing libraries. These algorithmic innovations are extended to tensor operations and deployed as part of the Cyclops library. We describe novel parallel implementations of tensor decomposition and tensor completion methods using Cyclops. Finally, we introduce a new algorithm for tensor decomposition, pairwise perturbation, which approximates the alternating least squares procedure with asymptotically less cost.

Edgar Solomonik
University of Illinois at Urbana-Champaign
solomon2@illinois.edu

SP3

SIAG/Supercomputing Career Prize: Ghosts of

Parallel Computing: Past, Present, and Future

Receiving a career prize means you're old. So in the first part of my talk I'll highlight a couple themes from the past 30 years of supercomputing which have taken us from gigaflops to the threshold of exaflops. In the latter part I'll speculate on what comes next. Spoiler alert: it seems unlikely the youngest members of the audience will see another 9 orders-of-magnitude speed-up within their career horizons. What does that mean for SIAM members interested in supercomputing and the topics we research? My perspective here will be application-centric (really, what else matters for science), so there will hopefully be ideas to both agree and disagree with.

Steve Plimpton
Sandia National Laboratories
sjplimp@sandia.gov

CP1

A Scalable Explicit Finite Element Solver for Cardiovascular Models with Uncertain Material Properties

Simulation of cardiovascular flow under uncertainty is an extremely intensive computational task, requiring large meshes and hundreds to thousands of high-fidelity model solutions. In this context, we propose a novel approach for ensemble simulation, and demonstrate it to the segregated, explicit-in-time solution of blood flow in the thoracic aorta with random material properties, focusing on the implementation of fast matrix-vector products on CPUs and GPUs. Distributed CPU storage is achieved through METIS partitioning and using a sparse compressed row storage (CRS) format with dense blocks containing multiple material property realizations for a three-d.o.f.s shell finite element. We developed optimized Cython code for the sparse matrix-vector multiplication using MPI+openMP and compared it to the *mkl_cspblas_dcsrgemv* routine provided through the Intel MKL library. Our implementation achieves better performance on a wide range of mesh sizes, number of cores, and with/without multithreading. Our openCL-based GPU matrix-vector product achieves instead a 10-fold speed-up with respect to a naïve implementation, by using separate command-queues, overlapping the CPU to GPU data transfer with GPU kernel execution, and using page-locked CPU memory. Additional improvements are obtained by direct computation of local element matrices on the GPU. Ongoing work focuses on coupling our explicit structural solver with a variational multiscale finite element fluid solver.

Xue Li, Daniele E. Schiavazzi
University of Notre Dame
xli26@nd.edu, dschiavazzi@nd.edu

CP1

Toward a Predictive Model to Monitor the Balance Between Discretization and Rounding Errors in Hydrodynamic Simulations

In order to take advantage of HPC and exaflopic machines, a natural tendency is to scale up the simulations by increasing mesh sizes, and to rely blindly on the double floating-point precision. With more points in the mesh, the discretization error decreases, and simulations should deliver more precise results. However, increasing too much the number of points has negative consequences on the result quality. Numerically, the time step is constrained by the number of points: with more points in the mesh, simulations require more iterations, hence more operations. With more operations, the rounding errors accumulate and perturb the computation, preventing the simulation from converging to its solution. This competition between discretization and rounding errors shows that there is an optimal mesh size, where the cumulated (and unavoidable) rounding errors due to floating-point arithmetic are still less significant (in the mantissa) than the discretization error. The present study investigates hydrodynamic computation in 1, 2 and 3D and aims at building a predictive model that can determine this optimal size of the mesh. This predictive model is based on analysis tools on floating-point operations that must be inserted in simulation codes. We shall discuss how its results still depend on cases and schemes. These tools, the simulation and the results that helped to design the model are detailed in this work.

William Weens, Alexandre Merasli
CEA
william.weens@gmail.com, alexandre.merasli@cea.fr

Thibaud Vazquez-Gonzalez
CEA Bruyeres le Chatel
thibaud.vazquez-gonzalez@cea.fr

Rémi Chauvin
CEA
remi.chauvin@cea.fr

CP1

The Validation of the Simulated HIRF Effects in Metallic Cases using Parallel FDTD Solver on a High Performance Computer

To accurately predict High Intensity Radiated Field (HIRF) effects in various systems is highly required for most commercial as well as defense applications, which is mainly based on electromagnetic simulation. Further, the validation of the simulated HIRF effects at different levels is also required. However, it is often difficult and needs some appropriate techniques and facilities. Here, in-house developed parallel Finite-Difference Time-Domain (FDTD) solver is at first employed for simulating the HIRF effects of two metallic cases with multiple slots and apertures, which is carried out on a high performance computer with tens of thousands of processors. The uncertainty analysis of the incident plane wave angle related to the electric field distribution in the metallic cases is finished with the help of non-intrusive polynomial chaos (NIPC) method, and the feature selective validation (FSV) is performed to quantify the difference between the measured and simulated field results. It is shown that the results obtained from the FSV method provide high reliability and confidence for our numerical characterization of HIRF effects in the metallic cases and even other complex structures.

Yueqian Wu
Institute of Applied Physics and Computational Mathematics
icy_0313@hotmail.com

CP2

High-Order Spatio-Temporally Parallel "Fast-Hybrid' Wave Equation Solver at $\mathcal{O}(1)$ Sampling Cost

We propose and demonstrate a frequency/time hybrid integral-equation method for the time dependent wave

equation in two and three-dimensional spatial domains. Relying on Fourier Transformation in time, the method utilizes a fixed (time-independent) number of frequency-domain integral-equation solutions to evaluate, with superalgebraically-small errors, time domain solutions for arbitrarily long times. As a transform method, the required frequency-domain integral equation problems are completely decoupled and can be computed in an entirely parallel manner. The approach relies on two main elements, namely, 1) A smooth time-windowing methodology that enables accurate band-limited representations for arbitrary long time signals, and 2) A novel Fourier transform approach which, without causing spurious periodicity effects, delivers dispersionless spectrally accurate solutions. The algorithm can handle dispersive media, complex physical structures, it enables parallelization in time in a straightforward manner, and it allows for time leaping– that is, solution sampling at any given time T at $\mathcal{O}(1)$-bounded sampling cost, for arbitrarily large values of T, and without requirement of evaluation of the solution at intermediate times. The proposed frequency/time hybridization strategy provides significant advantages over other available alternatives such as volumetric discretization and convolution-quadrature approaches.

Thomas Anderson, Oscar P. Bruno
California Institute of Technology
tanderson@caltech.edu, obruno@caltech.edu

Mark Lyon
Mathematics & Statistics
University of New Hampshire
mlyonn@unh.edu

CP2

High Order Scalable FFT-Krylov Subspace Methods for the 3D Convection Diffusion Equation

In this presentation, we develop an efficient parallel iterative approach to the solution of the three-dimensional convection-diffusion equation with variable coefficients based on high resolution compact finite-difference approximation schemes. In our approach the resulting system is solved by a combination of a Krylov subspace-type method with a matching high order Fast Fourier Transform (FFT) preconditioner with coefficients depending only on one spatial variable. This preconditioner also includes the convection terms only in the spatial direction of the dependent coefficients. In the algorithms considered, the exact solution of the high order preconditioning system is based on a combination of the separation of variables technique and FFT-type methods developed in our previous publication. The resulting numerical methods allow efficient implementation on parallel computers. The results of implementation of these methods in OpenMP and MPI programming environments are presented. Numerical results on synthetic data confirm the high efficiency of the iterative algorithms.

Ron Gonzales, Yun Teck Lee, Yury A. Gryazin
Idaho State University
gonzrona@isu.edu, leeyunt@isu.edu, gryazin@isu.edu

CP3

Large-Scale DPG Finite Element Simulation of a Nonlinear Multiphysics Fiber Amplifier Model

Multi-mode fiber lasers in high-power operation suffer from undesired thermal coupling effects such as the transverse mode instability (TMI). The TMI is a major obstacle in power-scaling of large mode area, active gain, fiber amplifiers. A better understanding of these nonlinear coupling effects is beneficial in the design of new fibers. We present details on the implementation and numerical results for a high-fidelity fiber amplifier model. This model is based on the 3D vectorial time-harmonic Maxwell equations for two weakly coupled electromagnetic fields. Thermal effects are modeled via coupling with the heat equation. The high-frequency nature of the wave propagation problem requires the use of high-order discretizations to effectively counter numerical pollution. The discontinuous Petrov-Galerkin (DPG) finite element method provides a stable discretization with a built-in error indicator suitable for hp-adaptivity. For simulating a significant fiber length of more than one thousand wavelengths, a scalable parallel implementation is critical. For this, we have developed an MPI/OpenMP version of an hp-adaptive finite element software that supports high-order discretizations for complex multiphysics problems with variables of the entire H^1-H(curl)-H(div)-L^2 exact sequence, hybrid meshes with elements of all shapes, and anisotropic hp-refinements. We show scalability results, in the context of this particular fiber laser application, for modern manycore architectures.

Stefan Henneking
Oden Institute, UT Austin
stefan@ices.utexas.edu

Jacob Grosek
Air Force Research Laboratory, Kirtland Air Force Base
jacob.grosek.1@us.af.mil

Leszek Demkowicz
Institute for Computational Engineering and Sciences (ICES)
The University of Texas
leszek@ices.utexas.edu

CP3

A Scalable Parallel Contact Algorithm Based on Dynamic Ghost Reconstruction for Lagrangian Hydrodynamic Application

Parallel simulation of contact problem is very challenging due to the extremely complex behavior of the contact line motion on the boundary. A data decomposition based parallel algorithm is imposed to deal with contact with large elastic deformations including time dependent responses. Entity set technology and patch based slide line data structure are combined to support the parallel contact definition and continuously varying relation management. Dynamic ghost region reconstruction technique is designed to support patch based parallel contact search and computation. Check point technique is used to reduce communication overhead caused by reconstruction process. The algorithm is designed and implemented on JAUMIN which is a component based domain specific parallel programming framework. The implementation of the algorithm in a two dimentional Lagrangian elastic-plastic fluid mechanics application proves that it effectively balances the load of internal and contact force calculation to achieve good parallel efficiency.

Li Liao
Institute of Applied Physics and Computational Mathematics

liliao@iapcm.ac.cn

CP3

Approaching Exa-Scalable and Accurate Green's Function Coupled Cluster Calculations of DNA Fragments

Nucleobases are molecular building blocks of DNA. An accurate understanding of the photoelectron properties of these molecules is extremely important for unveiling the mechanism of the formation and damage of DNA chains. However, only limited knowledge in this area has been known. Ab initio calculations in various levels of theory have only been limited to single nucleic acid bases. When the system becomes larger and more quantum-mechanical meaningful, the obtained results will be greatly shifted by significant many-body effects. In this work, by developing Green's function coupled cluster theory and a more efficient tensor algebra library, we are able to compute the more accurate photoelectronic properties of a series structures ranging from single nucleobases up to three base pairs. The algorithm and library have been carefully designed for achieving exa-scaling performance. All the many-body calculations have been performed on Summit supercomputing facility. The obtained results of nucleobases and base pairs have exhibited significant improvements over single-particle results and great capability of obtaining full many-body spectra function. The results are so far the most accurate results for the most quantum-mechanical meaningful DNA fragments, and thus help us to unveil the photoelectronic properties change upon the structural change, calibrate the other cheaper theories, and obtain the exa-scaling experience of many-body calculations.

Bo Peng
Advanced Computing, Mathematics and Data Division
Pacific Northwest National Laboratory
peng398@pnnl.gov

Ajay Panyala
High Performance Computing Division
Pacific Northwest National Laboratory
ajay.panyala@pnnl.gov

Sriram Krishnamoorthy, Karol Kowalski
Pacific Northwest National Laboratory
sriram@pnnl.gov, karol.kowalski@pnnl.gov

CP4

Tensor Processing Units for Financial Monte Carlo

Monte Carlo methods are core to many routines in quantitative finance such as derivatives pricing, hedging and risk metrics. Unfortunately, Monte Carlo methods are very computationally expensive when it comes to running simulations in high-dimensional state spaces where they are still a method of choice in the financial industry. Recently, Tensor Processing Units (TPUs) have provided considerable speedups and decreased the cost of running Stochastic Gradient Descent (SGD) in Deep Learning. After having highlighted computational similarities between training neural networks with SGD and stochastic process simulation, we ask in the present paper whether TPUs are accurate, fast and simple enough to use for financial Monte Carlo. Through a theoretical reminder of the key properties of such methods and thorough empirical experiments we examine the fitness of TPUs for option pricing, hedging and risk metrics computation. We show in the following that Tensor Processing Units (TPUs) in the cloud help accelerate Monte Carlo routines compared to Graphics Processing Units (GPUs) which in turn decreases the cost associated with running such simulations while leveraging the flexibility of the cloud. In particular we demonstrate that, in spite of the use of mixed precision, TPUs still provide accurate estimators which are fast to compute. We also show that the Tensorflow programming model for TPUs is elegant, expressive and simplifies automated differentiation.

Francois Belletti
Google Research
belletti@google.com

Davis King, Kun Yang, Roland Nelet, Yusef Shafi, Yi-Fan Chen, John Anderson
Google
davisking@google.com, kuny@google.com, rnelet@google.com, yusef@google.com, yifanchen@google.com, janders@google.com

CP4

Fast Image Reconstruction at a Synchrotron Laboratory

In this work we present a fast GPU implementation for tomographic reconstruction of large datasets using data obtained at the brazilian synchrotron light source. The algorithm is distributed in a server with 4 GPUs through a fast pipeline implemented in C/CUDA programming language. Our algorithm is theoretically based on a recently discovered low complexity formula, computing the total volume within O(N3 logN) floating point operations; much less than traditional algorithms that operates within O(N4) FLOPS over an input data of size O(N3). The results obtained with real data indicate that a reconstruction can be achieved within orders of second, provided the data is transferred completely to the memory.

Eduardo Miqueles, Gilberto Martinez, Jr., Patricio Guerrero
LNLS/CNPEM
eduardo.miqueles@lnls.br, gilberto.martinez@lnls.br, patricio.guerrero@lnls.br

CP4

General Memory-Independent Lower Bound for MTTKRP

Our goal is to establish lower bounds on the communication required to perform the Matricized-Tensor Times Khatri-Rao Product (MTTKRP) computation on a distributed-memory parallel machine. MTTKRP is the bottleneck computation within algorithms for computing the CP tensor decomposition, which is an approximation by a sum of rank-one tensors and frequently used in multidimensional data analysis. The main result of this paper is a communication lower bound that generalizes previous results, tightening the bound so that it is attainable even when the tensor dimensions vary (the tensor is not cubical) and when the number of processors is small relative to the tensor dimensions. The attainability of the bound proves that the algorithm that attains it, which is based on a block distribution of the tensor and communicating only factor matrices, is communication optimal. The proof technique utilizes an established inequality that relates computations to data access as well as a novel approach based on convex

optimization.

Grey Ballard, Kathryn Rouse
Wake Forest University
ballard@wfu.edu, kathryn.rouse@inmar.com

CP5

Multiscale Multiphysics Coupling of Hall-Effect Thruster, Plume, and Surface Charging Models for Spacecraft Integration

Assessment of spacecraft integration is particularly important when designing spacecraft, as exhaust gases from propulsion devices can damage and contaminate important components of spacecraft such as the payload and solar array. When using a Hall-Effect Thruster (HET), ions originating from the thruster and created in the plume through a charge-exchange process can contribute to sputtering of spacecraft components even in regions with no direct line of sight of the thruster channel. The assessment of spacecraft is typically done with combinations of numerical models that simulate a HET device, plume region, and spacecraft surface charging. These models involve multiple timescales that are governed by ions, electrons, grid resolution, and electron motion within materials, and therefore coupling of the models has to be done with care for stable convergence of numerical solution. In addition, it is difficult to achieve good performance from naive coupling of different simulations due to communication. To overcome the difficulty, a unified framework across the three numerical models is currently under development, which will simplify data exchange and enable strong coupling between models.

Samuel Araki
ERC Inc. / Air Force Research Laboratory
samuel.j.araki@gmail.com

CP5

Landing on Mars: Petascale Unstructured-Grid CFD Simulations on Summit

A campaign to investigate the use of supersonic retropropulsion as a means to land payloads on Mars large enough to enable human exploration is presented. Simulations are performed on the worlds largest supercomputer, Summit, located at Oak Ridge National Laboratory. The engineering and computational challenges associated with retropropulsion aerodynamics and the need for large-scale resources like Summit are reviewed. For these simulations, a GPU implementation of NASA Langley Research Center's FUN3D flow solver is used. The development history, performance, and scalability are compared with those of contemporary HPC architectures. The use of an optimized GPU-accelerated CFD solver on Summit has enabled simulations well beyond conventional computing paradigms.

Eric Nielsen, Aaron Walden, Ashley Korzun, Bill Jones, Jan-Renee Carlson
NASA Langley
eric.j.nielsen@nasa.gov, aaron.c.walden@nasa.gov, ashley.m.korzun@nasa.gov, w.t.jones@nasa.gov, jan-renee.carlson@nasa.gov

Pat Moran, Tim Sandstrom
NASA Ames
patrick.moran@nasa.gov, timothy.a.sandstrom@nasa.gov

Mohammad Zubair
Old Dominion University
zubair@cs.odu.edu

CP5

Numerical Method and Parallelization for the Computation of Coherent Synchrotron Radiation

The purpose of this work is to develop and parallelize an accurate and efficient numerical method for the computation of synchrotron radiation from relativistic electrons in the near field. The high-brilliance electron beam and coherent short-wavelength light source provides a powerful method to understand the microscopic structure and dynamics of materials. Such a method supports a wide range of applications including matter physics, structural biology, and medicine development. To understand the interaction between the beam and synchrotron radiation an accurate and efficient numerical simulation is needed. With millions of electrons, the computational cost of the field would be large. Thus, multilevel parallelism and performance portability are desired since modern supercomputers are getting more complex and heterogeneous.

Boqian Shen
Rice University
boshen@rice.edu

CP5

Virtual Environment for Sensor Performance Assessment (vespa) Geometry Engine

For realistic synthetic imagery, radiative transfer methods coupled with large mesh geometry provide the most scientifically accurate way to model a scene. Radiative models typically use ray-tracing techniques to determine where radiative energy is coming from or moving to. This work presents an approach to making a ray query geometry engine that actively stores large scale terabyte geometry in out-of-core memory on parallel general purpose processors. Procedures for geometry distribution, structures for efficient ray-tracing, and the ray query API are discussed. Geometry distribution uses Morton codes and parallel sorting routines to create geometry scene-chunks that are distributed among processing nodes. Each scene chunk is then broken down using a bounding volume hierarchy (BVH) using axis-aligned bounding boxes (AABB). The BVH allows for efficient ray tracing of the geometry. The ray query API allows client-side programs, such as sensor models and radiative transfer models, which exist on the same high performance computer to efficiently identify intersected geometry given directed rays. The geometry has key values that can uniquely identify data from solver programs. Scalability, partition timing, and ray timing results will be presented.

Barry C. White, Robert H Hunter, Aaron Valoroso
US Army Engineer Research and Development Center
barry.c.white@usace.army.mil, robert.h.hunter@erdc.dren.mil, aaron.a.valoroso@erdc.dren.mil

Reena Patel, Jerrell Ballard
US Army Corps of Engineers
reena.r.petal@erdc.dren.mil, jerry.ballard@usace.army.mil

CP6

ThunderEgg: A Multigrid Solver for Variable Coefficient Elliptic Problems on Adaptive Cartesian

Meshes

ThunderEgg is a library that provides geometric multigrid preconditioners for variable coefficient elliptic PDEs on octree and quadtree based Cartesian adaptive meshes. ThunderEgg is an object-oriented C++ library designed for flexibility and to allow users to implement multigrid preconditioners for various problems on octree and quadtree adaptive meshes. In this talk, we will give an overview of the design principles and use of the ThunderEgg library. In addition, we will discuss numerical results from using this library to implement a variable coefficient elliptic multigrid preconditioned BiCGStab solver for the p4est based ForestClaw software package.

Scott Aiton, Donna Calhoun
Boise State University
scottaiton@u.boisestate.edu,
donnacalhoun@boisestate.edu

Grady B. Wright
Department of Mathematics
Boise State University, Boise ID
gradywright@boisestate.edu

CP6

Preconditioning Finite Element Equations via Nonconforming Reformulations

We present an approach for building optimal preconditioners for algebraic linear systems coming from conforming finite element discretizations of partial differential equations. The preconditioners utilize the framework of auxiliary or fictitious space methods. Namely, the main idea is to reformulate the original problem in a nonconforming setting, using discontinuous elements across subdomains, obtaining the element-by-element assembly property. This is achieved by introducing additional unknowns associated with the interfaces between subdomains, which decouples the subdomains. The continuity is enforced weakly via interface penalty terms. The resulting nonconforming formulation (or a Schur complement (reduced) version of it, that "condenses" the formulation only to the interfaces) is used to precondition the original problem. The element-by-element assembly property can be useful in "matrix-free" computations for high order discretizations, since it minimizes the coupling across subdomains and interfaces. Also, this provides a natural setting to apply element-based algebraic multigrid (AMGe) techniques for solving the nonconforming formulation. The resulting decoupling and locality can be beneficial for parallel computing. The approach allows for directly obtaining coarse nonconforming formulations, or their reduced (Schur complement) versions. This largely reduces the cost of the "condensation" step.

Delyan Z. Kalchev
University of Colorado Boulder
kalchev1@llnl.gov

Panayot Vassilevski
Lawrence Livermore National Laboratory
vassilevski1@llnl.gov

CP6

Graph Based Algebraic Multigrid Method

Algebraic Multigrid (AMG) method is widely used to solve large sparse linear systems which arise in various computational simulations. They are also used as preconditioners with Krylov subspace solvers like conjugate gradient method. With evolving multicore architectures, we face new challenges to achieve linear complexity for AMG. In this work, we analyze computational complexity of aggregation-based AMG on multicore architectures. We demonstrate that linear complexity of AMG can be achieved with graph based approaches. Our Primitive Graph based approach helps to achieve good strong scaling on multicore architectures till it achieves the bandwidth bound of multicore architecture. We also evaluate complexity of the coarse grid construction phase of AMG with graph based approaches for larger system sizes.

Manan J. Shah
Indian Institute of Science
Bangalore
shahjayant@iisc.ac.in

Dr. Sashikumaar Ganesan
Associate Professor, Dept. of Computational & Data Sciences
Indian Institute of Science, Bangalore, India
sashi@iisc.ac.in

CP7

A Projection-Based Data Partitioning Method for Distributed Tomographic Reconstruction

Tomography is a non-destructive technique for imaging the interior of a 3D object. We present an efficient data partitioning strategy for distributed tomographic reconstruction algorithms. Our novel partitioning method is a refinement of the previously published GRCB algorithm. Instead of taking as input a discrete set of lines corresponding to source–pixel pairs, the introduced algorithm works directly on the (cone-shaped) projections. We introduce a geometric characterization of communication volume, as well as a continuous model for load-balancing based on the varying line densities throughout the object volume. The resulting algorithm is orders of magnitude faster than the original algorithm while producing partitionings of similar quality. We introduce a novel communication data structure that can efficiently represent the communication metadata. An implementation on top of Bulk and the ASTRA toolbox is discussed. We provide experimental results of our method for various commonly used acquisition geometries. We achieve a speedup of 2.8x compared to ASTRA-MPI when using 32 GPUs to reconstruct an image for a circular-cone beam acquisition geometry.

Jan-Willem Buurlage
Utrecht University
janwillembuurlage@gmail.com

Willem Jan Palenstijn
Centrum Wiskunde & Informatica, The Netherlands
wjp@cwi.nl

Rob H. Bisseling
Utrecht University
Mathematical Institute
r.h.bisseling@uu.nl

Joost Batenburg
Centrum Wiskunde en Informatica

joost.batenburg@cwi.nl

CP7

Two-Level Dynamic Load Balancing for High Performance Scientific Applications

Scientific applications are often complex, irregular, and computationally-intensive. To accommodate their ever-increasing computational demands, the high performance computing (HPC) systems have become larger and more complex, offering increased hardware parallelism at multiple levels. Scientific applications need to exploit all multi-level hardware parallelism to harness the available computational power. The performance of applications executing on such HPC systems may adversely be affected by load imbalance at multiple levels, caused by problem, algorithmic, and systemic characteristics. Existing dynamic load balancing methods do not simultaneously address load imbalance at multiple software parallelism levels. This work investigates the joint impact of load imbalance on the performance of three scientific applications at the thread and process levels. We jointly apply and evaluate selected dynamic loop self-scheduling (DLS) techniques to both levels. We conduct an exhaustive set of experiments to assess and compare the combination of six DLS techniques at the thread level and eleven at the process level. The results show that improved overall application performance, by up to 21%, can only be achieved by jointly addressing load imbalance at both software parallelism levels. We offer insights into the performance of the selected DLS techniques and discuss the interplay of load balancing at the thread and process levels.

Ali Mohammed, Aurélien Cavelan, Florina M. Ciorba
University of Basel, Switzerland
ali.mohammed@unibas.ch, aurelien.cavelan@unibas.ch, florina.ciorba@unibas.ch

Ruben Cabezon
University of Basel
ruben.cabezon@unibas.ch

Ioana Banicescu
Dept. of Computer Science and Engineering
Mississippi State University
ioana@cse.msstate.edu

CP7

FastSV: A Distributed-Memory Connected Component Algorithm with Fast Convergence

This paper presents a new distributed-memory algorithm called FastSV for finding connected components in an undirected graph. Our algorithm simplifies the classic Shiloach-Vishkin algorithm and employs several novel and efficient hooking strategies for faster convergence. We map different steps of FastSV to linear algebraic operations and implement them with the help of scalable graph libraries. FastSV uses sparse operations to avoid redundant work and optimized MPI communication to avoid bottlenecks. The resultant algorithm shows high-performance and scalability as it can find the connected components of a hyperlink graph with over 134B edges in 30 seconds using 262K cores on a Cray XC40 supercomputer. FastSV outperforms the state-of-the-art algorithm by an average speedup of 2.21x (max 4.27x) on a variety of real-world graphs.

Yongzhe Zhang
SOKENDAI, Japan
zyz915@nii.ac.jp

Ariful Azad
Lawrence Berkeley National Laboratory
azad@iu.edu

Zhenjiang Hu
Peking University, China
huzj@pku.edu.cn

CP8

Efficient Vectorised Cuda Kernels for High-Order Finite Element Flow Solvers

Modern shared-memory systems like GPGPUs offer massively parallel compute power, while the gap between available FLOPS and memory bandwidth increases. This demands for parallel schemes with high arithmetic intensity, such as high-order finite element methods. In this work, we optimise the performance of elemental operators for matrix-free Newton-Krylov solvers. We consider triangular, quadrilateral, tetrahedral, prismatic and hexahedral elements from curved unstructured high-order meshes to maintain geometric flexibility for engineering applications. The tensor-product bases of all these spectral elements further allows the exploitation of the sum-factorisation approach that reduces the number of operations. We implement efficient CUDA kernels for all operators of advection-diffusion and Helmholtz equations for a range of polynomial orders. We place special emphasis on efficient data-structures by interleaving the data of a group of 32 elements in order to achieve contiguous memory access for the vectorised operations within a CUDA warp. We further investigate the benefit of increasing the arithmetic intensity by re-computing quadrature metrics and mappings within each kernel, instead of loading pre-computed data from global memory. We ultimately present the aggregate performance of our kernels in simple 2D and 3D advection-diffusion solvers.

Jan R. Eichstaedt
Department of Aeronautics
Imperial College London
jan.eichstaedt13@imperial.ac.uk

Joaquim Peiro
Dept of Aeronautics
Imperial College London, UK
j.peiro@imperial.ac.uk

David Moxey
College of Engineering, Mathematics and Physical Sciences
University of Exeter
d.moxey@exeter.ac.uk

CP8

MaMiCo: Scalable Coupling of Particle Ensembles to Transient Continuum Flow Simulations

Particle systems, such as molecular dynamics (MD) simulations, are an indispensable tool for a variety of scientific applications. However, the computational demand limits MD applicability to rather small spatial setups, even for massively parallel simulations on high-performance computing systems. One approach to extend this applicability to larger simulation domains is given by multiscale coupling schemes which link particle methods to continuum flow

simulations. Advanced coupling methods filter MD data to reduce thermal noise, improve stability and reduce computational cost. We present a new hydrodynamic noise filter based on the non-local means (NLM) algorithm, which was developed for image processing originally, and we compare it to conventional filtering based on Proper Orthogonal Decomposition (POD). This presentation aims to point out scalable ways to design, implement and execute molecular-continuum coupled systems, with focus on new parallelization paradigms and distributed data filtering and processing. We utilize and extend the Macro-Micro-Coupling tool (MaMiCo), a modular C++ framework for parallel coupling of arbitrary continuum and particle solvers. Our results suggest that in many cases the NLM noise filter in MaMiCo is able to improve signal-to-noise ratios of fluctuating MD data compared to POD filtering, and thus enables stable coupling on shorter time scales, while having only a small impact on simulation performance.

Piet Jarmatz
Helmut Schmidt University
jarmatz@hsu-hh.de

Philipp Neumann
Helmut Schmidt University, Germany
philipp.neumann@hsu-hh.de

CP8

The Multiscale Perturbation Method for Elliptic Equations

In the formulation of multiscale methods for second order elliptic equations that are based on domain decomposition procedures, typically the computational domain is decomposed into subdomains, and for each subdomain a set of multiscale basis functions is numerically constructed. Consider the application of such a method to solve a multiphase flow problem where through an operator splitting algorithm, the velocity-pressure and transport equations are solved sequentially. From one time step to the next the multiscale basis functions should be recomputed, because of the coupling of the underlying PDEs. Instead of recomputing all multiscale basis function every time step of a numerical solution, we propose the Multiscale Perturbation Method (MPM). In MPM an approximate solution of velocity and pressure for a new time is obtained by combining regular perturbation theory with multiscale basis functions computed in an earlier time. The solution obtained at the perturbation step is improved further by a local update step to give the final solution of the MPM. An efficient parallel algorithm is implemented in multi-core machines and the method also fits well in GPU clusters. Numerical experiments, where the perturbation theory results are compared with direct fine grid solutions, are presented and discussed.

Het Y. Mankad, Alsadig Ali, Felipe Pereira
The University of Texas at Dallas
hym130030@utdallas.edu, alsadig.ali@utdallas.edu, luisfelipe.pereira@utdallas.edu

Fabricio S. Sousa
Instituto de Ciencias Matematicas e de Computacao
Universidade de Sao Paulo
fsimeoni@icmc.usp.br

CP8

SMILU: a Staggered-Grid, Multi-Level ILU Preconditioner for Steady Fluid-Transport Problems

When computing steady state solutions of the incompressible Navier-Stokes equations of computational fluid dynamics, a common approach is to use "pseudo time stepping", which allows to control the conditioning of the arising linear systems, or even resort to explicit schemes, by using a small time step. An arguably more effective approach is to directly solve for steady states and step through parameter space instead using a homotopy or numerical continuation technique. This, however, leads to non-symmetric linear systems of saddle point type, for which it is difficult to find an appropriate preconditioner, especially if the Reynolds Number becomes large. In this talk we exploit structure-preserving properties of a two-level ILU method geared towards a particular discretization method (the Arakawa-C grid finite volume method) [Wubs & Thies, SIMAX (32), 2011]. We extend the scheme to a parallel fully coupled multi-level incomplete factorization for the 3D stationary incompressible Navier-Stokes equations by observing that the structure-preserving properties can be fully retained in a recursion. The main idea is to define an appropriate geometric domain decomposition for the staggered grid, leading to a robust method with computational complexity of O(N log N). We show results for the 3D lid-driven cavity benchmark, and investigate the performance of another recent development, the FROSch preconditioner available in Trilinos.

Jonas Thies
German Aerospace Center (DLR)
jonas.thies@dlr.de

Fred Wubs
University of Groningen
f.w.wubs@rug.nl

Sven Baars
Dept. of Mathematics, Univ. of Groningen, The Netherlands
s.baars@rug.nl

CP9

Hybrid Empirical and Simulation-Based Autotuning of a Dense Linear Algebra Library for Heterogeneous Architectures

Parallel numerical libraries for modern architectures can be finely optimized. However, this optimization process may require to carefully set up many parameters, which can be a pretty cumbersome task when done manually. For instance, modern dense linear algebra libraries often split the matrix in submatrices whose size may significantly impact performance. Similarly, the scheduling policy may be of importance, especially when dealing with heterogeneous architectures. Possibly, it may also be interesting to let computational units idle. In this work, we discuss how to automatically decide the selection of these parameters. The idea is that, during a preliminary phase at install time, well chosen tests are performed, so that, once the library has been installed, these parameters get automatically (and instantly) decided while achieving an overall performance close to optimum. We illustrate our discussion with the Chameleon dense linear algebra library running on top of the StarPU runtime system. We use the SimGrid simulator to predict the performance of time consuming executions (large matrices) and hence reduce the time spent in the autotuning phase.

Jesús Cámara

University of Murcia
jcamara@um.es

Emmanuel Agullo
INRIA
emmanuel.agullo@inria.fr

Javier Cuenca
Departamento de Ingeniería y Tecnología de Computadores
University of Murcia
jcuenca@um.es

Domingo Giménez
Departamento de Informática y Sistemas
University of Murcia
domingo@um.es

CP9

HMG: A Configurable High-Performance Multilevel Preconditioner

MOOSE is a general finite element framework for multiphysics simulations. Applications built upon MOOSE cover a broad range of engineering problems from many disciplines including nuclear energy, environmental science, and mining industries. Different sets of partial differential equations arise in each of these fields and there is no single preconditioning algorithm that works effectively for all these problems. In this work, we attempt to create a configurable preconditioner system where all aspects of the underlying algorithms can be configured and tuned for specific target applications. The numerical effectiveness of this system on modern supercomputers with tens of thousands of processor cores is demonstrated for multiple problems including neutron transport equations, Navier-Stokes equations, solid mechanics problems, and compressible flows.

Fande Kong
Idaho National Laboratory
fande.kong@inl.gov

Cody J. Permann
Advanced Modeling and Simulation
Idaho National Laboratory
cody.permann@inl.gov

Alexander Lindsay
Idaho National Laboratory
alexander.lindsay@inl.gov

CP9

Automatic Scaling for Improved Conditioning in the Multiphysics Object-Oriented Simulation Environment

How to improve non-linear/linear convergence is a frequent question of PDE software. One reason for poor convergence can be an ill-conditioned linear operator. This scenario commonly arises in multi-physics simulations when different physics have different relative scales. It can even arise in single-physics cases when hanging-node constraints or Dirichlet boundary conditions are enforced strongly. To combat poor conditioning, we present an automatic scaling system implemented in the Multiphysics Object-Oriented Simulation Environment (MOOSE) that ensures the maximum absolute value for a single physics variable diagonal is unity. Improved convergence results for arbitrary multi-physics simulations employing either an explicit matrix or matrix-free linear operator are shown. Combination of automatic scaling with MOOSEs automatic differentiation system for solution of thermo-mechanical problems is also demonstrated. Further research will focus on automatic scaling of saddle-point problems for application to mortar finite element problems like frictionless and frictional mechanical contact.

Alexander Lindsay, Fande Kong
Idaho National Laboratory
alexander.lindsay@inl.gov, fande.kong@inl.gov

Cody J. Permann
Advanced Modeling and Simulation
Idaho National Laboratory
cody.permann@inl.gov

Derek R. Gaston, Richard Martineau
Idaho National Laboratory
derek.gaston@inl.gov, richard.martineau@inl.gov

CP10

Scalable Resilience Against Node Failures for Communication-Hiding Preconditioned Conjugate Gradient and Conjugate Residual Methods

The observed and expected continued growth in the number of nodes in large-scale parallel computers gives rise to two major challenges: global communication operations are becoming major bottlenecks due to their limited scalability, and the likelihood of node failures is increasing. We study an approach for addressing these challenges in the context of solving large sparse linear systems. In particular, we focus on the pipelined preconditioned conjugate gradient (PPCG) method, which has been shown to successfully deal with the first of these challenges. In this paper, we address the second challenge. We present extensions to the PPCG solver and two of its variants which make them resilient against the failure of a compute node while fully preserving their communication-hiding properties and thus their scalability. The basic idea is to efficiently communicate a few redundant copies of local vector elements to neighboring nodes with very little overhead. In case a node fails, these redundant copies are gathered at a replacement node, which can then accurately reconstruct the lost parts of the solver's state. After that, the parallel solver can continue as in the failure-free scenario. Experimental evaluations of our approach illustrate on average very low runtime overheads compared to the standard non-resilient algorithms. This shows that scalable algorithmic resilience can be achieved at low extra cost.

Markus Levonyak, Christina Pacher
University of Vienna
markus.levonyak@univie.ac.at,
christina.pacher@univie.ac.at

Wilfried N. Gansterer
Department of Computer Science
University of Vienna
wilfried.gansterer@univie.ac.at

CP10

Leveraging One-Sided Communication for Sparse Triangular Solvers

In this paper, we implement and evaluate a one-sided

communication-based distributed-memory sparse triangular solve (SpTRSV). SpTRSV is used in conjunction with Sparse LU to affect preconditioning in linear solvers while one-sided communication paradigms enjoy higher effective network bandwidth and lower synchronization costs compared to their two-sided counterparts. We use a passive target mode in one-sided communication to implement a synchronization-free task queue to manage the messaging between producer-consumer pairs. Whereas some numerical methods lend themselves to simple performance analysis, the DAG-based computational graph of SpTRSV demands we construct a critical path performance model in order to assess our observed performance relative to machine capabilities. In alignment with our model, our foMPI-based one-sided implementation of SpTRSV reduces communication time by 1.5x to 2.5x and improves SpTRSV solver performance by up to 2.4x compared to the SuperLU_DIST's two-sided MPI implementation running on 64 to 4,096 processes on Cray supercomputers.

Nan Ding
lawrence berkeley national laboratory
nanding@lbl.gov

Samuel Williams, Yang Liu
Lawrence Berkeley National Laboratory
swwilliams@lbl.gov, liuyangzhuan@lbl.gov

Xiaoye S. Li
Computational Research Division
Lawrence Berkeley National Laboratory
xsli@lbl.gov

CP10

Efficient Parallel Algorithms and Implementations for Sparse Triangular Solves on GPUs

The sparse triangular matrix solve (SpTrSV) is an important computation kernel that is demanded by a variety of numerical methods such as the Gauss-Seidel iterations. However, developing efficient parallel algorithms for SpTrSV that are suitable for GPUs remains a challenging task due to the inherently sequential nature in the solve. In this paper, we revisit this problem by reviewing several parallel algorithms based on different task scheduling and different sparse matrix storage schemes, proposing modifications to the existing methods that can greatly improve the performance, and describing the implementations in details. Numerical results of Gauss-Seidel iterations with structured and unstructured matrices make evident the superiority of the proposed algorithms and implementations comparing with state-of-the-art methods in the literature.

Chaoyu Zhang
Arkansas State University
chaoyu.zhang@smail.astate.edu

Ruipeng Li
Center for Applied Scientific Computing
Lawrence Livermore National Laboratory
li50@llnl.gov

CP11

Parallel Sweeping Preconditioners For Domain Decomposition Methods Applied to the Helmholtz Equation

Various wave propagation problems can be modelled in the frequency domain by the Helmholtz equation. Efficiently solving the Helmholtz equation at high frequencies in complex geometrical and/or material configurations is the subject of intense research, with recent promising contributions focused on high-order finite element methods coupled with optimized Schwarz Domain Decomposition (DD) schemes. Sweeping preconditioners for DD have recently gained a lot of interest because of their fast convergence, promising a number of DD iterations that is quasi independent of the number of subdomains. These preconditioners however have two major drawbacks: they rely on intrinsically sequential operations and they are naturally only suited for layered-type domain decompositions. In this talk we will present a family of generalized sweeping preconditionners where sweeps can be done, in parallel, in several directions, for checkerboard-type domain decompositions. This contribution relies on the availability of accurate, high-order transmission conditions between the subdomains. Numerical results for high-frequency Helmholtz problems will be presented.

Ruiyang Dai
Universite catholique de Louvain (UCL)
Institute of Mechanics, Materials and Civil Engineering (iMM
ruiyang.dai@uclouvain.be

Jean François Remacle
Université catholique de Louvain
Institute of Mechanics, Materials and Civil Engineering
jean-francois.remacle@uclouvain.be

Chistophe Geuzaine
University of Liège
Electrical Engineering and Computer Science
cgeuzaine@uliege.be

CP11

Graph Partitioning for Computational Mechanics Simulations with GDSW

In the application of scalable parallel computing algorithms and hardware to computational mechanics problems for aerospace structures, typically the bottleneck is the iterative linear solver. In the case of the Generalized Dryja, Widlund, Schmidt (GDSW) overlapping Schwarz domain decomposition method with different levels of overlap, evidence will be presented that preconditioner quality and effectiveness, and hence overall simulation time, is very sensitive to the graph partition. Numerical case studies will be presented using graph partitioners that wrap the Zoltan, ParMetis and Chaco graph partitioners with a layer adding application specific geometric and topological information. Overall simulation time will be shown to best correlate with the load imbalance of the subdomain overlaps. This observation will be shown to be consistent with a rudimentary complexity analysis of GDSW. Modifications of the graph partitioner wrapper will be shown to significantly reduce simulation time by incrementally improving the overlap load balance.

David Day
Sandia National Laboratories
dmday@sandia.gov

CP11

Communication Performance Modeling of LAP3D

and its Application in Performance Optimization

LAP3D(Laser And Plasma 3 Dimension code) is a three-dimensional laser and plasma code for explorering the influence of various instabilities in LPI processes. In LAP3D, several physical processes are coupled together. Fine numerical simulation requires the grid to have a certain size, which leads to a large amount of computation and long calculation time. These become challenges that LAP3D faces. To address such challenges, the following work is carried out. Firstly, LAP3D is analyzed to get the key characteristics which determine the performance of the code. Secondly, a communication performance model is set up based on the key characteristics. Lastly, the model is applied to predict the LAP3D communication costs, to look for the shortest communication cost and to predict the parallel efficiency. In this talk, the communication performance model is tested. From the test results, the communication performance model effectively predicts the optimal communication time and the parallel efficiency of the code, supports it run efficiently on tens of thousands of cores and provides theoretical guidance for the code optimization.

Hong Guo
Institute of Applied Physics and Computational Mathematics
guo_hong@iapcm.ac.cn

CP11

Scalable FFT-Krylov Subspace Method for Scattering Problem

This talk is to present an efficient parallel implementation of the direct compact numerical solver for 3D Helmholtz equation for forward scattering problem on multicore computers. Forward scattering problems have many applications such as predicting the scattered electromagnetic field given the geometry and find many applications in computer-aided design. Hence a fast and efficient way to solve the forward scattering problem will impact areas, like high-speed circuits, integrated optics, antenna analysis, remote sensing, geophysical sensing, and inverse scattering. Our parallel algorithm will be used to computationally simulate data for the solution of the inverse problem of imaging mine-like targets. In our setting, land mines are modeled as small abnormalities embedded in an otherwise uniform media with an air-ground interface. The main challenge in this setting is the requirement of solving the Helmholtz equation for high frequencies which are time-consuming using standard direct solution techniques. In our approach, this system is solved by a combination of Krylov subspace-type method with a direct parallel FFT-type preconditioner. The resulting numerical method allows a natural and efficient implementation on parallel computers. Numerical results for realistic ranges of parameters in soil and mine-like targets confirm the high efficiency of the proposed parallel iterative algorithm.

Yun Teck Lee, Ron Gonzales, Yuiry Gryazin
Idaho State University
leeyunt@isu.edu, gonzrona@isu.edu, gryazin@isu.edu

CP13

Multi GPU Algorithms for Hierarchical Matrix Computations

We present scalable distributed memory algorithms for algebraic compression and matrix-vector products involving hierarchical matrices. Motivated by the limited memory of GPUs, our representation of hierarchical matrices uses nested bases for asymptotically optimal memory storage. The representation consists of two trees for row and column bases and a matrix tree for the low rank matrix blocks. The trees are flattened and distributed at every level on the GPUs. This allows upsweep and downsweep operations on the basis trees, which are the fundamental blocks of hierarchical matrix algorithms, to use efficient and parallel marshaling of linear algebra operations (GEMV, GEMM, QR, and RRQR) for batched execution. Parallelism can then achieved both within individual GPUs as well as among GPUs. Our algorithms structure the computations so that the data communicated between GPUs is limited to the data produced by the bases trees, reducing the volume of data movement necessary. The implementation uses NCCL and MPI for inter and intranode communications, and we present results on a V100 cluster.

Wajih Halim Boukaram
KAUST
wajihhalim.boukaram@kaust.edu.sa

George M. Turkiyyah
American University of Beirut
gt02@aub.edu.lb

David E. Keyes
KAUST
david.keyes@kaust.edu.sa

CP13

Accelerating Jacobi Iteration on GPUs using Shared Memory

Scientific simulations modeling important physical phenomena typically require solving linear systems of equations. This step often takes a vast amount of computational time to complete, and therefore presents a bottleneck in simulation work. Solving these linear systems efficiently requires the use of highly parallel hardware with high computational throughput, as well as the development of algorithms which respect the memory hierarchy of these hardware architectures. In this talk, we present an algorithm to accelerate Jacobi iteration for solving linear systems arising from structured problems, on graphics processing units (GPUs). Our algorithm takes advantage of the quick to access shared memory which resides in each streaming multiprocessor (SM). This is done using a domain decomposition approach, in which we partition our problem domain into subdomains whose information is stored in the shared memory of each SM. Jacobi iterations are performed internally within each SM's shared memory, avoiding the need to perform expensive global memory accesses every step. We test our algorithm on the linear systems arising from discretization of Poissons equation in 1D and 2D, and observe speedup in convergence using our shared memory approach compared to a traditional Jacobi implementation which only uses global memory on the GPU.

Mohammad S. Islam, Qiqi Wang
Massachusetts Institute of Technology
moislam@mit.edu, qiqi@mit.edu

CP13

GPU Acceleration of a High-Order Arbitrary Lagrangian-Eulerian Code

With the introduction of advanced architectures such as GPUs, a major effort has been set forth to develop al-

gorithms and implementations for hydrodynamics codes which can realize performance on next gen computer systems. In this talk, we present our two-fold approach in tailoring Blast, an arbitrary Lagrangian-Eulerian (ALE) multi-material hydrodynamics code developed at LLNL, for advanced architectures. A distinguishing feature of Blast is the choice of high order finite element methods leading to higher arithmetic intensity per data access; a trait favored by modern computing processors. Our first approach leverages partial assembly techniques and exploits the tensor product structure of quad/hex elements. Partial assembly decomposes global operators into a sequence of operators with cascading scope. Through this approach, the action of global operators only requires values at quadrature points enabling reduced storage and on the fly calculations. Second, to remain portable across computing platforms, our implementation builds on the RAJA programming model and Umpire resource manager developed at LLNL. Together RAJA and Umpire enable maintaining a single source code suitable for heterogenous computing systems. This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. LLNL-ABS-788139.

Arturo Vargas, Ryan Bleile
Lawrence Livermore National Lab
vargas45@llnl.gov, bleile1@llnl.gov

Jean-Sylvain Camier, Veselin Dobrev, Robert Rieben
Lawrence Livermore National Laboratory
camier1@llnl.gov, dobrev1@llnl.gov, rieben1@llnl.gov

CP13

GPU-Accelerated Barycentric Treecodes

This talk presents recent developments in the GPU implementations of the barycentric treecodes. Treecodes are fast summation techniques that accelerate the calculation of N-body interactions, for example charged particles interacting via the Coulomb or screened Coulomb kernel. They can also be applied to accelerate the discrete sums that arise in Green's function methods. Treecodes organize source particles into a hierarchy of clusters, typically an octree, then approximate the interactions between a target particle and clusters of source particles at various depths of the tree, reducing the complexity of the calculation from $O(N^2)$ to $O(N \log N)$. In particular, the barycentric treecodes use the barycentric form of an interpolating polynomial to approximate the particle-cluster interactions. This work uses the Lagrange and Hermite interpolating polynomials. These treecodes are capable of achieving very good accuracy by increasing the interpolation order. In contrast to previous approaches based on the Taylor expansions, this form of the particle-cluster approximation has a readily parallelizable structure which enables an efficient GPU implementation. I will discuss the approximation, demonstrate its parallel structure, present the GPU implementation, and show results from various test cases.

Nathan Vaughn, Leighton Wilson
University of Michigan
njvaughn@umich.edu, lwwilson@umich.edu

Lei Wang
University of Wisconsin, Milwaukee
wang256@uwm.edu

Robert Krasny
Department of Mathematics
University of Michigan
krasny@umich.edu

CP14

Scalable Triangle Counting on Distributed-Memory Systems

Triangle counting is a canonical graph-analysis problem arising in various network science applications such as spam detection, link recommendation and dense neighborhood graph discovery. It is also one of the IEEE HPEC Graph Challenge problems. In this talk, we describe a novel, hybrid parallel triangle counting algorithm based on a linear algebra formulation of this problem. In our algorithm, we exploit MPI and Cilk frameworks at the distributed-memory and shared-memory parallelism levels, respectively. For each of these parallelism levels, we use a different partitioning strategy tailored to the specific needs of that level. At the distributed-memory level, we partition the problem among MPI processes using 2D Cartesian block partitioning, which is commonly used for reducing communication overheads. At the shared-memory level, we use 1D block partitioning within each Cartesian block using Cilk programming model, which has been proven to be more efficient than OpenMP in the triangle counting context. Our algorithm demonstrates very good strong scaling behavior in almost all tested graphs. It also achieves the fastest time on the 1.4B edge real-world twitter graph, 3.217 seconds. Compared to the results of the previous years champion of the Graph Challenge, we demonstrate a speed up of 2.7x on this twitter graph using 5.6x fewer cores.

Seher Acer
Sandia National Laboratories
sacer@sandia.gov

Abdurrahman Yasar
Georgia Institute of Technology
School of Computational Science and Engineering
ayasar@gatech.edu

Sivasankaran Rajamanickam, Michael Wolf
Sandia National Laboratories
srajama@sandia.gov, mmwolf@sandia.gov

Umit V. Catalyurek
Georgia Institute of Technology
School of Computational Science and Engineering
umit@gatech.edu

CP14

HashGraph - Scalable Hash Tables using A Sparse Graph Data Structure

Hash tables are ubiquitous and used in a wide range of applications for efficient probing of large and unsorted data. If designed properly, hash-tables can enable efficient look ups in a constant number of operations. As data sizes continue to grow and data becomes less structured, the need for efficient and scalable hash table also grows. In this paper we introduce HashGraph, a new scalable approach for building hash tables that uses concepts taken from sparse graph representations—hence the name HashGraph. We show two different variants of HashGraph, a simple algorithm that outlines the method to create the hash-table and an advanced method that creates the hash

table in a more efficient manner. HashGraph shows a new way to deal with hash-collisions that does not use "open-addressing' or "chaining', yet has all the benefits of both these approaches. We show that HashGraph can deal with a large number of hash-values per entry without loss of performance as most open-addressing and chaining approaches have. Further, we show that HashGraph is indifferent to the load-factor. Given the above, HashGraph is extremely fast and outperforms several state of the art hash-table implementations. While HashGraph is not architecture dependent, using a NVIDIA GV100 GPU, HashGraph is anywhere from 2X-8X faster than cuDPP, WarpDrive, and cuDF. HashGraph is able to build a hash-table at a rate of 2.5 billion keys per second and can probe at nearly the same rate.

Oded Green
Georgia Institute of Technology
School of Computational Science and Engineering
ogreen@gatech.edu

CP14

Introduction to High Performance Computing for Decision Makers: Creating a One-Day Workshop for a Data Science Audience

The excitement around data science has percolated multiple places in our society. Therefore, it is not surprising that a data-science workshop called *Introduction to High Performance Computing* in Silicon Valley receives registrations from entry level engineers, department directors, entrepreneurs, undergraduate students, researchers and faculty. However, the diversity of backgrounds implies different student motivations and, together with the time constraint of a one-day workshop, it challenges the traditional approach to learn HPC. Students are not interested anymore in learning how to code prototypical algorithms using a particular library, they want to understand what HPC means to them and what it has to offer to their work. Two years ago, we created a summer workshop with the aim of giving participants the tools to analyze HPC applications without previous experience in coding. We took a top to bottom approach that puts hardware scalability as the main constraint, and OpenMP, MPI and Spark as tools to attenuate synchronization and memory demands. Instead of focusing on the correct syntax or usage, we expose the abstraction behind the tools using everyday analogies. With this approach we provide a common framework to understand HPC libraries that facilitates understanding of hardware challenges. As a result, we close the gap between decision makers and developers, empowering them to question the best use of the available resources for their application.

Cindy Orozco Bohorquez
Stanford University
orozcocc@stanford.edu

CP15

Asynchronous Smoothers for Computational Fluid Dynamics on Graphics Processing Units

Asynchronous iterations have started to regain interest over the past few years for fine-grain parallel solution of large sparse systems of equations. Recent efforts to use them for solving linear systems on modern parallel architectures include those by Anzt et al. ["A block-asynchronous relaxation method for graphics processing units", J. Parallel Distrib. Comput. (73), 2013] and Chow and Patel ["Fine-grained parallel incomplete LU factorization", SIAM J. Sci. Comput. (37:2), 2015]. However, application to specific domains has been lacking, especially in fluid dynamics. Hawkes et al. ["Chaotic linear-system solvers for unsteady CFD", VI International Conference on Computational Methods in Marine Engineering, 2015] demonstrated asynchronous linear iterations for incompressible flows in marine engineering, but on a traditional multi-core central processing unit (CPU). Our work aims to demonstrate the applicability of asynchronous Gauss-Seidel and incomplete LU factorization smoothers to multigrid solvers for compressible turbulent flows in external aerodynamics, implemented for graphics processing units (GPUs). Our previous work has shown that modifications to existing asynchronous iterations are needed to achieve parallel scalability for compressible flows. In this work, we detail the modifications needed to achieve good performance on GPUs. Two- and three-dimensional benchmark cases of external aerodynamics on multi-block structured grids will be used as test cases.

Aditya Kashi
PhD student
Department of Mechanical Engineering, McGill University
aditya.kashi@mail.mcgill.ca

Siva Nadarajah
McGill University
siva.nadarajah@mcgill.ca

CP15

Prismatic Space-Time Tiling as an Approach to High Efficient Stencil Computations

Low operation intensity is the main reason of low efficiency for HPC application. This problem proved to be quite hard. There are automatic solutions like polyhedral compilers [Bondhugula et al, Automatic Transformations for Communication-Minimized Parallelization and Communication-Minimized Parallelization and Locality Optimization in the Polyhedral Model] but they usually show lower performance than state-of-the-art codes. In this work we present a way to design scalable parallel applications that could utilize all levels of parallel execution for modern computational systems. Our approach is based on the same ideas as polyhedral optimization. Since we mostly deal with stencil computations and hand-tailor the code we can reduce the complicated theory to a much simpler one. Our approach also provides a way to quite accurately evaluate application performance [Levchenko et al, Locally Recursive Non-Locally Asynchronous Algorithms for Stencil Computation]. This fact is employed to find the best tiling parameters for computational systems. This approach was successfully applied for different physical problems, numerical methods and computational systems [Levchenko et al, LRnLA Lattice Boltzmann Method: A Performance Comparison of Implementations on GPU and CPU; Korneev et al, Numerical simulation of increasing initial perturbations of a bubble in the bubbleshock interaction problem; Zakirov et al, DiamondTorre Algorithm for High-Performance Wave Modeling].

Sergey Khilkov, Anastasia Perepelkina, Vadim Levchenko
Keldysh Institute of Applied Mathematics
khilkov.s@gmail.com, mogmi@narod.ru, lev@keldysh.ru

CP15

Implementing Randomized Asynchronous Linear

System Solvers

Asynchronous iterative methods present a mechanism to improve the performance of algorithms for highly parallel computational platforms by removing the overhead associated with synchronization among computing elements. This paper considers a class of asynchronous iterative linear system solvers that employ randomization to determine the component update orders, specifically focusing on the effects of drawing the order from non-uniform distributions. Results from shared-memory experiments with a two-dimensional finite-difference discrete Laplacian problem show that using distributions favoring the selection of components with a larger contribution to the residual may lead to faster convergence than selecting uniformly. Multiple implementations of the randomized asynchronous linear system solvers are considered and tested with various distributions and parameters. In the best case of parameter choices, average times for the normal and exponential distributions were, respectively, 13.3% and 17.3% faster than the performance with a uniform distribution, and were able to converge in approximately 10% fewer iterations than traditional stationary solvers.

Masha Sosonkina
Old Dominion University
Ames Laboratory/DOE
masha@scl.ameslab.gov

Erik Jensen
Old Dominion University
ejens005@odu.edu

Evan Coleman
Naval Surface Warfare Center,
Dahlgren Division
evanccoleman@gmail.com

MS1

Whats New in the Cuda Math Libraries

Todays fastest compute platforms are designed from the ground up to leverage the immense compute power of NVIDIA GPUs. As these platforms increase in scale and add specialized hardware, the CUDA Math Libraries are keeping up by constantly expanding, providing industry leading performance and coverage of common compute workflows across AI, ML and HPC. Major initiatives to support common workflows are: multi-GPU scalability, reduced and mixed precision computing, and libraries that allow kernel fusion and customizations. In this talk we review the latest developments in the CUDA Math Libraries including Tensor Core acceleration of HPC solvers without loss of accuracy, support for multiple GPUs in FFTs, BLAS and LAPACK routines, and the addition of new libraries with device function support and tensor linear algebra functionality.

Timothy Costa
Nvidia
tcosta@nvidia.com

Harun Bayraktar
NVIDIA
hbayraktar@nvidia.com

MS1

Intel Software Solutions for Diverse Computing Architectures

Today's applications are more diverse than ever, as is the hardware used to run and accelerate them. This causes a challenge for developers, as they must juggle multiple code bases, tools, programming languages, and workflows. The oneAPI initiative from Intel is designed to help overcome these challenges, by offering a unified programming model that allows acceleration of the code on various processing architectures. This new standards-based programming model will allow developers to run their workload on existing Intel Xeon systems and accelerate portions of it to specialized processors, including GPUs, FPGAs, and other accelerators. Included in the Intel oneAPI portfolio are powerful developer tools like Intel oneAPI Math Kernel Library.

Sarah Knepper, Shane Story
Intel Corporation
sarah.knepper@intel.com, shane.story@intel.com

MS1

Strategies for Exploiting Multicore in High-Level Software Packages

High-level software packages such as MATLAB® are typically designed to favor productivity over performance. However, customers still have the expectation that software will optimally exploit available hardware resources to extract the best possible performance during computation. This requirement can often prove challenging, as to serve a variety of customers, the software cannot in general be designed to target any one specific set of hardware. In this talk, we share our experience developing mathematical libraries used by MATLAB to deliver scalability and performance on a variety of commodity multicore processors. In particular, we focus on our strategy for implementing core algorithms for convolutional neural networks supporting applications in deep learning.

Pat Quillen, Christopher Turnes
MathWorks
pquillen@mathworks.com, cturnes@mathworks.com

MS2

Why Run Real Benchmarks? Can't I Just Run Linpack to Understand My System?

Understanding the true performance of systems, including complex computational and data analysis systems is increasingly difficult and getting more challenging as we experience a new explosion of architectural choices, development of new and more dynamic algorithms, convergence of Big Data and Extreme Computing and increasingly complex workflows. This talk will evaluate the current state of the challenges and approaches for system assessment - both initial and ongoing - that are influencing the HPC marketplace and investments. The talk will then explore how to improve our performance evaluation and system assessment state of the art and suggest ways to become more effective in understand productivity and achievable performance - not just isolated performance.

William T. Kramer
NCSA / University of Illinois
wtkramer@illinois.edu

MS2

Evaluation of Large Scale Systems with Focus on

Application Performance: the Benchmarking Perspective

With benchmark-driven performance evaluation, informed by a broad field experience of involvement with 4 widely known benchmarks in circulation, an opportunity arises to obtain meaningful measurements of a number of HPC metrics that relate to applications' behavior. The scope of these past benchmarks spans a number of computational patterns that are often directly observed in scientific codes on both past and modern hardware platforms. This may then advance the progress of co-design even for early hardware but still with meaningful insight into the performance space. This was often observed when careful consideration is given to how the benchmark metrics relate to the essential aspects of applications' rates of execution. The extensive adoption and longevity of these efforts combined with a large volume of results' data will provide a relevant background on the community process involved in benchmarking: from basic implementation, through the performance engineering, deployment efforts, and the eventual adoption for use in applications for proxy use or to be leveraged for code reuse in critical performance kernels. The process ends up being beneficial as benchmarks and applications provide feedback to each other as the early hardware platforms mature to production and the porting of software on the delivered system is tagged for general use.

Piotr Luszczek
University of Tennessee
luszczek@icl.utk.edu

MS2

Performance Modeling of Applications by Benchmarks

Widely used benchmarks, such as High Performance Linpack (HPL), no longer strongly correlate to real application performance. On the other hand, while using real applications or mini applications can give a direct estimation into application performance, more efforts have been required to port, to tune and to optimize them. In this work, we introduce a new methodology to estimate applications' performance by only performing simple benchmarks. We show and discuss some experimental results obtained several systems including X86, Arm, and so on.

Miwako Tsuji
RIKEN Advanced Institute for Computational Science
miwako.tsuji@riken.jp

MS2

How to Monitor the Performance Evolution of a Large HPC System : a System and Application Points of View

To measure the initial performance of a new HPC system, it is usual to use standard benchmarks, like TOP500, Graph500, on the whole system. During its lifetime, the system is evolving, the hardware is changing, and software, including firmware and processor microcode, also is evolving to correct bugs, security problems and to offer new features. So, inevitably HPC systems' performance is moving. While on one side, it is hardly the case that we can run the standard benchmarks again because it requires monopolizing the system for a too long time. On the other side, the standard benchmark does not represent very well the applications that are running on the system: users' feeling is very different from the benchmarkers' one. In this talk, we will introduce the solutions we are implementing that combine the system and the application points of view. For the system side, we will discuss benchmarks that enable us to quickly assess the proper functioning and performances of both individual nodes as well as the entire HPC system. We will describe the way those benchmarks are integrated into our large HPC cluster components lifecycle to capture potential issues before they are raised into production systems. For the application side, we will discuss how to choose performance indicators that are meaningful from an application user's point of view, and how to choose the right use cases. We will describe our solution to monitor the performance evolution.

Matthieu Hautreux, Marc Perache, Jean-Christophe Weill
CEA, DAM, DIF, F-91297 Arpajon, France
matthieu.hautreux@cea.fr, marc.perache@cea.fr, jean-christophe.weill@cea.fr

MS3

PETSc's Accelerator Model and Algebraic Multigrid Work and Data Placement at Extreme-Scale

This talk will discuss several new solvers being developed for fusion particle codes. We will discuss Structure preserving methods for the Vlasov-Maxwell-Fokker-Planck system, new Fokker-Planck collision operator in Landau form optimized for emerging architectures, 3D fluid preconditioners for implicit kinetic methods, and new GPU alagebraic multigird solvers in PETSc at scale on SUMMIT.

Mark Adams
Lawrence Berkeley National Laboratory
mfadams@lbl.gov

Matthew G. Knepley, Joseph Pusztay
University at Buffalo
Department of Computer Science and Engineering
knepley@gmail.com, josephpu@buffalo.edu

MS3

Strategies, Challenges, and Lessons Learned in Developing GPU Support for SUNDIALS

SUNDIALS is a suite of robust and scalable solvers for systems of ordinary differential equations, differential-algebraic equations, and nonlinear equations designed to be used on a variety of computing systems ranging from laptops to super computers. With GPUs providing most of the FLOPS available on top-end machines, such as Summit at Oak Ridge National Laboratory, there has been a large effort to develop GPU support in SUNDIALS. Encapsulation of parallelism within SUNDIALS lends itself to a framework where data is offloaded to the GPU and control is kept on the CPU. In this framework, we can support GPUs in SUNDIALS by providing interfaces to GPU-enabled vector and linear solver operations. Another avenue of support involves taking advantage of the fact that SUNDIALS packages often operate at a fine-grained level of an application's parallel decomposition. In this talk we will discuss the strategies, challenges, and the lessons learned in our ongoing effort.

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. LLNL-ABS-789680.

Cody J. Balos, David J. Gardner, Carol S. Woodward

Lawrence Livermore National Laboratory
balos1@llnl.gov, gardner48@llnl.gov, woodward6@llnl.gov

Daniel R. Reynolds
Southern Methodist University
Mathematics
reynolds@smu.edu

Alan C. Hindmarsh
Lawrence Livermore National Laboratory
hindmarsh1@llnl.gov

MS3

Investigating Quasi-Newton Outer Product Representations on GPUs

The slowdown of Moore's Law and the growth of compute-intensive workloads such as artificial intelligence has pushed the development of high-performance computing (HPC) towards accelerator-based systems. Among supercomputers in the Top500 list, the share of accelerator FLOPs has grown from 20% in 2010 to 76% in 2018. As part of this trend, many libraries and application codes for solving partial-differential-equations (PDEs) have been ported and adapted to run on accelerator hardware such as graphical processing units (GPUs). Consequently, it has now become imperative for PDE-constrained optimization algorithms to also live on and exploit the capabilities of the same hardware used by the underlying PDE solvers. In the present work, we launch an investigation into the use of quasi-Newton (QN) methods on GPUs. QN approximations are among the most popular gradient-based kernels for solving large-scale nonlinear systems of equations, and are widely used for both continuous optimization and for PDE solutions. We implement both matrix-free and compact dense representations of several QN methods in PETSc/TAO, and leverage PETSc data structure abstractions to profile QN performance on both CPUs and GPUs using MPI and ViennaCL backends.

Alp Dener
Argonne National Laboratory
Mathematics and Computer Science Division
adener@anl.gov

MS3

Implementation of Fast Eigensolvers on GPUs

We discuss the potential benefits and challenges of developing fast eigensolvers on mulit-GPU systems. We will first examine the implementation of spectrum slicing algorithms for solving symmetric eigenvalue problems in which each spectral slice is computed on a single or a group of GPUs using either a shift-invert subspace iteration or a polynomial filtering subspace iteration. We leverage the existing CUDA libraries for performing the matrix decomposition and matrix-matrix multiplications required in the shift-invert and polynomial filtering procedures. We discuss strategies for minimizing inter GPU communication. We then discuss challenges in implmenting an iterative many-body eigensolver in a legacy nuclear physics application code, and present some preliminary results.

Chao Yang
Lawrence Berkeley National Lab
cyang@lbl.gov

MS4

Data-Driven Models of the Mouse Mesoscale Connectome: Network Structure and Functionality

The complex connectivity structure unique to the brain network is believed to underlie its robust and efficient coding capability. The recent development of the structural mouse brain network available at the Allen Mouse Brain Connectivity Atlas, makes it possible to conduct in-depth analyses on connections between structure and computation in the brain network. The recent expansion of the Allen Mouse Brain Connectivity data includes cell-type and layer-specific cortical network, constructed from viral tracing experiments in Cre-transgenic mice. Using this large anatomical connectivity dataset, we developed an unsupervised method to find the hierarchical organization of the mouse cortical-thalamic network, based on the layer-specific connectivity. The implemented method discovers the hierarchy of the mouse brain areas based on their anatomical connectivity patterns, and provides a measure of hierarchy scores for different connectomes. The uncovered hierarchy provides insights into the direction of information flows in the mouse brain, which has been less well-defined compared to the primate brain. Furthermore, the newly introduced global hierarchy score of the mouse brain network which measures the self-consistency of the obtained hierarchy suggests that the mouse cortical-thalamic network has a relatively shallow but clear hierarchical organization.

Hannah Choi
University of Washington
hannahch@uw.edu

MS4

Scalable Deep Learning of Biological Dynamical Systems

The discovery of governing equations from scientific data has the potential to transform data-rich fields that lack models. Advances in sparse regression are currently enabling the tractable identification of both the structure and parameters of a nonlinear dynamical system from data. The resulting models have the fewest terms necessary to describe the dynamics, balancing model complexity with descriptive ability, and thus promoting interpretability and generalizability. This provides an algorithmic approach to Occam's razor for model discovery. However, this approach fundamentally relies on an effective coordinate system in which the dynamics have a simple representation. In this work, we design a custom autoencoder to discover a coordinate transformation into a reduced space where the dynamics may be sparsely represented. Thus, we simultaneously learn the governing equations and the associated coordinate system. The resulting modeling framework combines the strengths of deep neural networks for flexible representation and sparse identification of nonlinear dynamics (SINDy) for parsimonious models. It is the first method of its kind to place the discovery of coordinates and models on an equal footing. We demonstrate this approach on several example high-dimensional biological dynamical systems with low-dimensional behavior. We then discuss how this method scales with the size of the system and the size of the dataset.

Kathleen Champion
University of Washington, Department of Applied

Mathematics
kpchamp@uw.edu

Bethany Lusch
Argonne National Laboratory
blusch@anl.gov

J. Nathan Kutz
University of Washington, Seattle
Dept of Applied Mathematics
kutz@uw.edu

Steven Brunton
University of Washington
sbrunton@uw.edu

MS4

Drug Response Prediction and Generative Models for Molecules

Since the early wins in computer vision, deep learning has increasingly been applied to hard problems that have defied previous modeling efforts. This is particularly true in chemistry and drug development where there are dozens of efforts to replace the traditional drug development computational pipelines with machine learning based alternatives. In our work we are applying deep learning to the problem of predicting tumor drug response for both single drugs and drug combinations. We have developed drug response models for cell lines, patient derived xenograft models and orgnoids that are used in preclinical drug development. Our approaches leverage work on attention, weight sharing between closely related runs for accelerated training and active learning for prioritization of experiments. Our goal is a broad set of models that can be used to screen drugs during early stage drug development as well as predicting tumor response for pre-clinical study design. Results to date include response classifications that achieve $> 92\%$ balanced classification accuracy on a pan-cancer collection of tumor models and broad collection of drugs. Additionally, we present preliminary results in applying machine learning to de novo molecule generation and production of surrogate models for activity screening based on high quality simulation and experimental datasets.

Fangfang Xia
Argonne National Laboratory
fangfang@anl.gov

MS4

High-Resolution Visualization of In Vivo Blood Flow from Low-Resolution MRI Scans using Computational Fluid Dynamics and Optimization

Detailed in vivo imaging of the human body using magnetic resonance imaging (MRI) holds great potential for scientific discovery and impact in health care. However, current MRI scanners and flow reconstruction techniques are limited by a fundamental trade-off between resolution, image quality, and scan time. We propose a new method of flow reconstruction that aims to overcome this limitation by integrating computational fluid dynamics (CFD), numerical optimization, and uncertainty quantification into the MRI workflow. Our approach defines an optimization problem that aims to define the boundary conditions and material properties of a high-order CFD simulation that best describes data from low-resolution (fast) MRI scans. The resulting data-certified simulation is subsequently used for flow visualization and to accurately compute clinical biomarkers. To quantify the uncertainty in the reconstruction due to noise in the measurements, we adopt a Bayesian setting and estimate the posterior distribution using implicit sampling. The optimization and sampling procedures are accelerated using adaptive projection-based reduced-order models. We demonstrate the method reconstructs both external and in vivo flows more accurately than standard 4D flow MRI techniques.

Matthew J. Zahr
University of Notre Dame
mzahr@nd.edu

MS5

Toward Exascale Plasma Simulations using Particle in Cell (pic) Algorithms

With the trend of computing resources moving toward more "advanced architectures from more traditional architectures like Intel's Phi systems (Trinity), to GPUs systems (Sierra/Summit) to the first exascale Cray/Intel system (Aurora) the DOE is developing several new simulation tools in a variety of disciplines. This talk focuses Sandia National Laboratories' code EMPIRE, which is a plasma simulation suite designed for performance on these new systems. This talk will overview the many capabilities of the EMPIRE code but will focus on the unstructured mesh finite element particle in cell code, and the approach toward future proofing as computing moves toward exascale. Results will be shown for realistic geometries on full system scale of Trinity, Astra and Sierra systems. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Matthew Bettencourt
Sandia National Laboratories
mbetten@sandia.gov

MS5

Technical Computer Aided Design Modeling of Semiconductor Devices in Parallel Computing Architectures

Sandia National Laboratories technical computer aided design (TCAD) device simulator, Charon, is based on a multiphysics code for simulating general transport-reaction phenomena in semiconducting and insulating materials including the effects of heat generation and radiation. The semiconductor modeling capability in Charon was developed to work in a manner similar to other commercially available TCAD codes. What sets it apart is that it supports large-scale parallel execution through its use of decomposition and parallel solution tools available in the Trilinos suite of algorithms and enabling technologies. We will present the current state of TCAD modeling at large and Charons current parallel capabilities. Strong and weak scaling and grind times will be presented for a range of typical semiconductor devices. Plans for Charons future approaches to using next generation platforms will be discussed. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under

contract DE-NA0003525.

Lawrence C. Musson
Sandia National Laboratories
lcmusso@sandia.gov

Gary Hennigan
Sandia National Labs
glhenni@sandia.gov

Jason Gates
Computational Mathematics Department
Sandia National Laboratories
jmgate@sandia.gov

Xujiao Gao
Sandia Natl. Labs
xngao@sandia.gov

Mihai Negoita, Andy Huang
Sandia National Laboratories
mnegoit@sandia.gov, ahuang@sandia.gov

MS5

Parallel Simulation of Large-Scale Integrated Circuits Using Xyce

The Xyce Parallel Circuit Simulator is a SPICE-style analog circuit simulator that has been designed, from the ground up, to achieve scalable performance through distributed memory techniques. With heterogeneous computing architectures becoming the norm, Xyce is going through a reinvention of the underlying infrastructure that has provided scalable computation for the last 20 years. The infrastructure changes are wide-ranging, from device distribution and topology to the numerical methods that enable Xyce to perform time-domain and frequency-domain simulation. In this talk we will present the basics of implementing a SPICE-style analog circuit simulator and take a closer look at the essential parts that make it scalable in circuit dimension, as well as parallel performance. Some of the topics discussed will include techniques for dynamically distributing devices and the preconditioned linear solvers that are essential for large-scale circuit performance. With a nod to the past capabilities of the Xyce circuit simulator, this talk will provide a forward-looking vision of what is coming next. SNL is managed and operated by NTESS under DOE NNSA contract DE-NA0003525.

Heidi K. Thornquist
Sandia National Laboratories
hkthorn@sandia.gov

MS5

Gemma: A Sandia National Laboratories Electromagnetic Code for Heterogeneous Computer Architectures

Gemma aims to provide Sandia National Laboratories with a tool for efficiently solving frequency domain electromagnetic problems on heterogeneous computer architectures, including CPUs, GPUs, and MICs, using Sandia's Kokkos performance portability library. While Gemma will include alternative algorithms for select use cases, it will rely on the method of moments to accurately and reliably solve electromagnetic problems on the most powerful computers available. Through several case studies of simulating electromagnetic environment effects, this presentation will illustrate Gemma's approach to establishing confidence and trustworthiness in its solutions. SNL is managed and operated by NTESS under DOE NNSA contract DE-NA0003525.

Brian Zinser, Samuel Blake, Robert Pfeiffer, Andy Huang, John Himbele, Brian A. Freno, Vinh Dang, Joseph Kotulski, Sivasankaran Rajamanickam, William Johnson, Salvatore Campione, William Langston
Sandia National Laboratories
bzinser@sandia.gov, sablake@sandia.gov, rapfeif@sandia.gov, ahuang@sandia.gov, jjhimbe@sandia.gov, bafreno@sandia.gov, vqdang@sandia.gov, jdkotul@sandia.gov, srajama@sandia.gov, wajohns@sandia.gov, sncampi@sandia.gov, wllangs@sandia.gov

MS6

Enlarged Krylov Methods and 2-Level Preconditionner for the Map-Making Problem in CMB Data Analysis

The Cosmological Microwave Background (CMB) data analysis in astro-physics consists in filtering different noises in the electromagnetic radiation as a remnant from an early stage of the universe in Big Bang cosmology. In order to do so and due to a constant need for better sensitivity, the size of the data sets used to solve this problem grows at Moore's rate. Therefore suitable, accurate, massively paralellisable algorithms are needed. We focus at one particular problem for the CMB data analysis, called the map-making problem, where one standardly seeks for a maximum likelihood estimator of the signal by solving a generalized least square problem via the normal equation. Our aim is to develop numerical tools to solve this problem which includes a study on Krylov methods and preconditioning techniques. This work fits in the Big Bang from Big Data of the Cosmic Microwave Background project (B3DCMB).

Thibault Cimic
Inria, France
thibault.cimic@inria.fr

Laura Grigori
INRIA
France
Laura.Grigori@inria.fr

MS6

Speculations on the Future of Krylov Methods using Results from Emerging Architectures

Krylov methods are indispensable tools that are used to solve large, sparse systems of equations arising in many physical simulations. These methods spend considerable time in a few computational kernels such as vector dot products and AXPY operations. As such, it is crucial that we understand the execution of these operations in order to leverage good performance. However, as HPC architectures become more complex with the addition of compute units such as GPUs, this task becomes more complicated. In this talk, we detail PETSc benchmarking results from the CPU-GPU Summit computing system at the Oak Ridge Leadership Computing Facility. Through detailed study of the performance of the fundamental compute kernels, such as vector dot products, that are the building blocks of Krylov methods, we identify computational bottlenecks, guide the use of new systems, and perhaps inform the design of algorithms that we expect to run on hetero-

geneous architectures.

Hannah M. Morgan
Argonne National Laboratory
Mathematics and Computer Science
hmorgan@anl.gov

MS6

The Numerical Stability Analysis of Parallel Conjugate Gradient Methods: Historical Context and Methodology

Numerical stability issues have been studied since the formulation of the conjugate gradient method in the middle of the last century, with many remarkable results achieved since then. Inexact computations in conjugate gradient method, due to either floating point roundoff error or due to intentional relaxation motivated by savings, have two basic effects, namely, slowing down convergence and limiting attainable accuracy. Although the methodologies for their investigation are different, these phenomena are closely related and cannot be separated from one another. Recently, the issues of attainable accuracy and delayed convergence caused by inexact computations became of interest in relation to pipelined conjugate gradient methods and their generalizations. In this contribution we recall the related early results and developments in synchronization-reducing conjugate gradient methods, identify the main factors determining possible numerical instabilities, and present a methodology for the analysis and understanding of pipelined conjugate gradient methods. We derive an expression for the residual gap that applies to any conjugate gradient method variant that uses a particular auxiliary vector in updating the residual, including pipelined conjugate gradient methods, and show how this result can be used to perform a full-scale analysis for a particular implementation.

Miro Rozloznik
Czech Academy of Sciences
Prague, Czech Republic
miro@cs.cas.cz

MS6

Inexactness and Compression in Krylov Subspace Methods

Krylov subspace methods have been around since the second half of the 20th century and remain among the most widely used algorithms for solving large linear systems. In order to adapt to the ever more parallel architecture of computer systems and – soon – exascale systems they remain the subject of ongoing research. One recent development are so-called inexact Krylov subspace methods. These are inspired by the observation that the matrix vector product required in each iteration is often only calculated approximately [V. Simoncini and D. B. Szyld: Flexible inner-outer Krylov subspace methods. SIAM Journal on Numerical Analysis (2002)] [L. Giraud, S. Gratton and J. Langou: Convergence in backward error of relaxed GMRES. SIAM Journal on Scientific Computing (2007)]. At the same time, it has become clear that the main bottlenecks on HPC systems are the communication bandwidth and the storage capacity. Lossy data compression techniques have therefore gained a lot of interest [D. Tao et al.: Significantly improving lossy compression for scientific data sets based on multidimensional prediction and error-controlled quantization. IEEE IPDPS 2017] [S. Götschel and M. Weiser: Lossy Compression for Large Scale PDE Problems. arXiv preprint (2019)]. In this talk we will explore the connection between inexact Krylov subspace methods and lossy compression and discuss some practical strategies that allow us to reduce the memory footprint of these methods.

Nick Schenkels
Inria
nick.schenkels@inria.fr

Emmanuel Agullo
INRIA
emmanuel.agullo@inria.fr

Franck Cappello
ANL
cappello@mcs.anl.gov

Sheng Di
Argonne National Laboratory
sdi1@anl.gov

Luc Giraud
Inria
luc.giraud@inria.fr

Xin Liang
Argonne National Laboratory
xlian007@ucr.edu

MS7

A Layer-Parallel Approach for Training Deep Neural Networks

Deep neural networks are a powerful machine learning tool with the capacity to learn complex nonlinear relationships described by large data sets. Despite their success training these models remains a challenging and computationally intensive undertaking. In this talk we will present a new layer-parallel training algorithm that exploits a multigrid scheme to accelerate both forward and backward propagation. Introducing a parallel decomposition between layers requires inexact propagation of the neural network. The multigrid method used in this approach stiches these subdomains together with sufficient accuracy to ensure rapid convergence. We demonstrate an order of magnitude wall-clock time speedup over the serial approach, opening a new avenue for parallelism that is complementary to existing approaches.

Eric C. Cyr
Computational Mathematics Department
Sandia National Laboratotories
eccyr@sandia.gov

Stefanie Guenther
Lawrence Livermore National Laboratory
guenther5@llnl.gov

Lars Ruthotto
Department of Mathematics and Computer Science
Emory University
lruthotto@emory.edu

Jacob B. Schroder
Department of Mathematics and Statistics
University of New Mexico
jbschroder@unm.edu

Nicolas R. Gauger
TU Kaiserslautern
nicolas.gauger@scicomp.uni-kl.de

MS7

Parallel-in-Time with Sundials and Xbraid

Numerical methods for integrating evolutionary equations in time typically rely on sequential time marching schemes and parallelization in the spatial dimension. As gains in computational power shift from faster processors to massively parallel systems serial time integration becomes a bottleneck in parallel efficiency. In order to leverage the greater concurrency on these systems parallel-in-time methods introduce an additional dimension of parallelism by distributing the workload in time across multiple processors. Recent work on multigrid-reduction-in-time (MGRIT) has shown significant speedups over sequential time stepping and is a relatively non-intrusive approach. In this talk we will discuss the recent efforts to combine the adaptive-step explicit, implicit, and IMEX time integration methods from the SUNDIALS ARKode library with the XBraid MGRIT library.

David J. Gardner, Robert Falgout
Lawrence Livermore National Laboratory
gardner48@llnl.gov, falgout2@llnl.gov

Daniel R. Reynolds
Southern Methodist University
Mathematics
reynolds@smu.edu

Carol S. Woodward
Lawrence Livermore National Laboratory
woodward6@llnl.gov

MS7

Parallel-in-Time Simulation of Electrical Power-grids with Unscheduled Events

Numerical simulations of electrical power grids often require accounting for unscheduled events such as equipment limits, deadbands, and faults. Time-serial integration approaches typically capture those events at runtime, modifying grid components and their dynamics in a sequential step-by-step manner. However, some power grid simulation studies involve long time domains and various time scales, serial time integration becomes a bottleneck for fast and scalable simulation solvers. In order to reduce the time-to-solution of power grid simulations, we apply a time-parallel integration scheme based on multigrid reduction in time, which allows for concurrency along the time-domain. While time-parallel simulations for power grids show excellent scaling behavior if a priori knowledge of the event times is available, handling unscheduled events can impair the multigrid convergence. In this talk we will summarize approaches for capturing unscheduled events in a time-parallel multigrid scheme and present first results on regulation study problems that include unscheduled component limits and faults.

Stefanie Guenther, Carol S. Woodward, Robert Falgout
Lawrence Livermore National Laboratory
guenther5@llnl.gov, woodward6@llnl.gov, falgout2@llnl.gov

MS8

Network Bottleneck Handling in Multi-Threaded MPI Context

Nowadays, codes combine multiple parallel programming models. These models often follow the hybrid MPI+X pattern, with X being a threaded model, either directly in the code or in DSL abstraction layers. To offer a better support for hybrid programs, MPI provides different levels of thread support up to MPI_THREAD_MULTIPLE, which allows a program to call MPI functions in any thread of a process. This behavior has multiple impacts: first on the MPI library which has now to be thread-safe, then on the interconnect card and the network due to the flow of messages coming from all the threads. MPC implements its own MPI library with two flavors: process-based and thread-based. Having a thread-based MPI enforces to have an efficient thread-safe MPI library. Hence, MPC has to deal with all the bottlenecks induced by multi-threaded MPI. In this presentation, we will describe some of the features implemented in MPC to loosen these bottlenecks.

Julien C. Jaeger
CEA
julien.jaeger@cea.fr

MS8

The Last Centimeter: Trials, Tribulations and Bottlenecks in Getting Data onto the Wire

GPU accelerators are increasing node complexity. They increase the number of locations data can be sent from between nodes. New technologies, including GPUDirect and GPUDirect async increase the options of how to send messages. In this talk, we investigate how performance varies with messaging location, and the performance and coding tradeoffs presented by new technologies. We show proper configuration and usage are important and messaging bottlenecks on GPUs occur for different reasons than on CPUs.

Ian Karlin
Lawrence Livermore National Laboratory
karlin1@llnl.gov

MS8

A Tool (interconnect) is Only As Good As its User (the Network Stack): Solving Network Bottlenecks via Software Stack Simulation in the Structural Simulation Toolkit

Designing interconnects for future HPC systems now goes far beyond questions of topology and routing. High-performance "lossless" interconnects like Infiniband are increasingly incorporating features long popular in Ethernet like quality of service (QoS), congestion control, and software-defined networks (SDN). Combining these features with new system trends including heterogenous accelerators, further node disaggregation, and optical switches are creating a new and challenging network design space. Architecture simulation has been used successfully to predict and understand network performance. To simulate large systems, these simulations are often based on lightweight models or traces specifically designed for MPI. This co-design disconnect between production network

stacks and simulation limits collaboration between architecture and runtime teams. Tuning new interconnect features for maximum performance creates more and more burden on the network software stack, meaning simulation tools must adapt to develop network software as well as hardware. Here we present recent developments in compiler tools for the Structural Simulation Toolkit (SST) designed to bridge the gap between simulator-specific models and real applications. In particular, we address auto-skeletonization via compilers that transforms large codes spanning thousands of processes and PBs of memory into a single simulator process needing only GBs of memory to simulate.

Jeremiah Wilke
Sandia National Laboratories
jjwilke@sandia.gov

MS9

A Dimension Switching Multigrid Method and Applications

In this project, we developed an efficient MultiGrid Method using dimension switch technique. To reduce the computation cost for modeling signal propagation within and between neurons in 3D, we employed 1D partition to replace the coarse grid in 3D but keep the fine grid. This way balances both efficiency and accuracy of MultiGrid Method. We started from working on a perfect cylinder to represent the axon of a neuron, then extend the method to simulate signal transport in axons with general shapes. The ultimate goal is to couple the method into solving problems on whole neurons. Meanwhile, we considered the implementation of parallel computation on multiprocessors with distributed memory for the proposed method.

Qingguang Guan, Gillian Queisser
Temple University
qingguang.guan@temple.edu,
gillian.queisser@temple.edu

MS9

Smooth Subdivision Multigrid and Large Scale Simulations in Life Science

We introduce a smooth subdivision theory-based geometric multigrid method. While theory and efficiency of geometric multigrid methods rely on grid regularity, this requirement is often not directly fulfilled in applications where partial differential equations are defined on complex geometries. Instead of generating multigrid hierarchies with classical linear refinement, we propose the use of smooth subdivision theory for automatic grid hierarchy regularization within a geometric multigrid solver. This subdivision multigrid method is compared to the classical geometric multigrid method for two benchmark problems. Numerical tests show significant improvement factors for iteration numbers and solve times when comparing subdivision to classical multigrid. A second study focusses on the regularizing effects of surface subdivision refinement, using the Poisson-Nernst-Planck equations as a model problem. These techniques can be coupled to novel dimension-switching multigrid methods. Various large-scale life-science applications are presented.

Gillian Queisser
Temple University
gillian.queisser@temple.edu

MS9

Parallel Geometric Multigrid for Continuum Models and Numerical Optimization

Geometric multigrid is well-known for highly-scalable implementations on distributed machines by balancing the coarse grid hierarchy among the involved processes. It nicely interplays with adaptive mesh refinement to account for highly localized features and to speed up the computation by saving non-required degrees of freedom. We present an overview about our implementation in the massively parallel simulation framework UG4 (unstructured grids) focusing on distributed-memory architectures. The applicability for largest high-performance computing clusters is demonstrated on a variety of problems including PDE continuum model simulations and numerical optimization.

Andreas Vogel
Ruhr University Bochum, Germany
a.vogel@rub.de

MS10

Addressing the Communication Bottleneck: Towards a Modular Precision Ecosystem for High Performance Computing

Over the last years, we have observed a growing mismatch between the arithmetic performance of processors in terms of the number of floating point operations per second (FLOPS) on the one side, and the memory performance in terms of how fast data can be brought into the computational elements (memory bandwidth) on the other side. As a result, more and more applications can utilize only a fraction of the available compute power as they are waiting for the required data. With memory operations being the primary energy consumer, data access is pivotal also in the resource balance and the battery life of mobile devices. In this talk we will introduce a disruptive paradigm change with respect to how scientific data is stored and processed in computing applications. The goal is to 1) radically decouple the data storage format from the processing format; 2) design a "modular precision ecosystem' that allows for more flexibility in terms of customized data access; 3) develop algorithms and applications that dynamically adapt data access accuracy to the numerical requirements.

Hartwig Anzt
Steinbuch Centre for Computing
Karlsruhe Institute of Technology
hartwig.anzt@kit.edu

Terry Cojean, Thomas Grützmacher
Karlsruhe Institute of Technology
terry.cojean@kit.edu, thomas.gruetzmacher@kit.edu

MS10

Performance and Accuracy of Mixed-Precision Matrix Factorizations with GPU Tensor Cores

The emergence of new hardware supporting low precision floating-point formats, such as half precision, presents great potential for scientific computing. However, it also presents new challenges as algorithms must be adapted to exploit the hardware: in particular, the NVIDIA GPU Tensor Cores are a new type of mixed-precision, block fused

multiply-add units. We propose several matrix factorization algorithms exploiting such units and compare them in terms of performance and accuracy, both theoretically (via their rounding error analysis) and experimentally. Our numerical results using the GPU Tensor Cores give new insights into how to best implement matrix factorizations in a mixed precision setting.

Pierre Blanchard, Nicholas J. Higham
School of Mathematics
The University of Manchester
pierre.blanchard00@gmail.com,
nick.higham@manchester.ac.uk

Florent Lopez
University of Tennessee, Knoxville
flopez@icl.utk.edu

Theo Mary
University of Manchester
School of Mathematics
theo.mary@manchester.ac.uk

Srikara Pranesh
University of Manchester
srikara.pranesh@manchester.ac.uk

MS10

Low Precision Floating-Point Arithmetic for the Solution of Linear System of Equations

Motivated by the demand in machine learning, modern computer hardware is increasingly supporting reduced precision floating-point arithmetic, which provides advantages in speed, energy, and memory usage over single and double precision. Given the availability of such hardware, mixed precision algorithms that work in single or double precision but carry out part of a computation in half precision are now of great interest for general scientific computing tasks. GMRES-based iterative-refinement (GMRES-IR) is one such algorithm for the solution of linear system of equations, that has demonstrated 4 times speedup over state of the art solver using the tensor cores feature of NVIDIA V100. Furthermore it has achieved a performance of 445 petaflops at scale on the Summit machine that leads the June 2019 TOP500 list. In this work we extend the applicability of GMRES-IR to linear systems where the matrix is symmetric and positive definite. A major challenge is that the matrix can lose its positive definiteness upon conversion to half precision. To address this issue we propose a diagonal perturbation based algorithm and experimentally demonstrate the effectiveness of the half precision Cholesky factor in GMRES-IR. Since the matrix is symmetric and positive definite, the conjugate gradient method might seem preferable to GMRES. We demonstrate that for ill conditioned matrices GMRES is the better choice because of its backward stability.

Nicholas J. Higham
School of Mathematics
The University of Manchester
nick.higham@manchester.ac.uk

Srikara Pranesh
University of Manchester
srikara.pranesh@manchester.ac.uk

MS10

Opportunities for Multi Precision Computation in Memory Bound Applications

Initially introduced for data storage only, the half precision floating-point format has been quickly adopted for computation by the machine learning community. Latterly, the iterative refinement algorithm for solving systems of equations has been revisited to harness the performance gain of half precision. It is for such compute-intensive operations that half precision has gained popularity. However, with as few as 16 bits for floating-point number representation, half precision also has the advantage of accelerating data movement operations. It is therefore worthwhile evaluating its potential for improving the performance of memory bandwidth bound applications. In this talk, we firstly discuss various ways to exploit reduced precision in memory bandwidth bound applications, with a special focus on sparse linear algebra routines. Secondly, we will present a set of benchmarks to assess the gain of reduced precision in these applications, when the time to convert input data to the reduced precision format is considered. Finally, we will share the key lessons learned from re-engineering real-world application to exploit reduced precision.

Nicholas J. Higham
School of Mathematics
The University of Manchester
nick.higham@manchester.ac.uk

Craig Lucas
The Numerical Algorithms Group
craig.lucas@nag.co.uk

Francoise Tisseur
The University of Manchester
School of Mathematics
francoise.tisseur@manchester.ac.uk

Mawussi Zounon
The Numerical Algorithms Group (NAG)
mawussi.zounon@nag.co.uk

MS11

GAMA: Quantum and Quantum-Inspired Algorithms for Non-Linear Integer Optimization

We discuss two original approaches to solve non-linear integer optimization problems that arise in finance, cancer genomics and supply chain optimization. Our Graver Augmented Multi-seed Algorithm (GAMA) utilizes augmentation along Graver basis elements (a test set) from multiple initial feasible solutions. A hybrid quantum-classical approach (GAMA-Q) is tested on D-Wave. Our experiments suggest that with a modest increase in coupler precisionalong with near-term improvements in the number of qubits and connectivity that are expectedthe ability to outperform classical best-in-class algorithms is within reach. A (fully classical) approach (GAMA-C) is well suited for Cardinality Boolean Quadratic Problems (CBQP), Quadratic Semi- Assignment Problems (QSAP) and Quadratic Assignment Problems (QAP). We find that for several instances of practical relevance, GAMA-C vastly out-performs best-in-class commercial solvers in terms of time to find the optimal solution (by two or three orders

of magnitude).

Sridhar Tayur, Hedayat Alghassi
Carnegie Mellon University
stayur@cmu.edu, hedayata@gmail.com

MS11

Adapting Engineering Design Practices to Mathematical Methods in Quantum Computing

In realizing a quantum computation, frequent practice in Math and CS is top down, i.e. quantum computations are realized via the creation of a big unitary matrix, whereas in Engineering common practice is bottom up design, i.e. quantum computations are realized via the assembly of well vetted blocks, e.g. the adder, counter, or shifter. Incompatibility arises via the latters use of ancillary bits, which can make the blocks and/or their final assembly non-unitary. This is a reflection of why we dont see more quantum algorithms, namely that many classical algorithms are difficult to quantize as the matrix structures involved are non-unitary, e.g. finite element methods, graph coloring, or set covering. Our presentation will discuss an approach around these difficulties that is proving particularly useful in Oracle based solution algorithms, e.g. Grover, quantum walk, or phase estimation, noting that small quantum computers are easier to construct and control than large ones. Our approach uses a classical computer to organize and reorganize a family of small quantum computers into a variety of structures, each specialized to the problem at hand, e.g. adding more constraints to the oracle of a Grover type algorithm as the algorithm progresses. Our presentation closes with applications to quantum machine learning and CAD via the approachs implicit quantization of classical uses of decision trees and graphs.

Steven A. Bleiler
Portland State University
bleilers@pdx.edu

Marek Perkowski
Portland State University, U.S.
mperkows@ee.pdx.edu

MS11

Commutative and Non Commutative Geometry for Noisy Intermediate-Scale Quantum (NISQ) Computations

Algebraic geometry, differential geometry, and topos theory are mighty mathematical theories that are used and cherished by mathematicians, and computer scientists, for decades now. I will survey their recent use in quantum computationsparticularly in the circuit model, adiabatic quantum computing, and measurement-based quantum computingand the benefit of doing so. Particular attention will be given to the compiling problem on NISQ architectures, hardware design, and connection to the theory of programming languages.

Raouf Dridi
Carnegie Mellon University
dridi.raouf@gmail.com

MS11

Nash Embedding: A Road Map to Realizing Quantum Hardware

The non-Euclidean nature of the state-space of qubits (and qudits in general) gives rise to the problem of practically implementing quantum circuits in physical hardware which necessarily resides in the Euclidean space R3. On the other hand, the Euclidean nature of bits (and dits in general) makes the implementation of reversible circuits in physical hardware relatively straight forward. I offer here a road-map to solving this problem in which the Nash embedding theorem isometrically maps qubits into bits and a quantum circuit into an equivalent reversible one, followed by the embedding of the resulting reversible circuit into R3 as a hardware graph.

Faisal Shah Khan
Khalifa University
faisal.khan@ku.ac.ae

MS12

MFEM: Accelerating Efficient Solution of PDEs at Exascale

Traditional implementations of high-order finite element methods using sparse matrices are unsuited to achieve high performance on GPUs due to their low arithmetic intensity and high data movements. On the opposite side, a matrix free formulation of high-order finite element methods can achieve high arithmetic intensity and optimal data movement, making them ideal for GPUs. Finding good algorithms is only the beginning of the journey though. Careful implementation of the algorithms is required to fully exploit the raw power of GPUs. This talk will explore the journey from redesigning finite element algorithms to implementing them efficiently on GPUs to achieve peak performance in the MFEM finite element library.

Yohann Dudouit
LLNL
dudouit1@llnl.gov

MS12

AMReX on GPUs: Strategies, Challenges and Lessons Learned

AMReX is a software framework for building massively parallel block-structured AMR applications using mesh operations, particles, linear solvers and/or complex geometry. AMReX was originally designed to use MPI + OpenMP on multicore systems and recently has ported the majority of its features to GPU accelerators. AMReXs porting strategy has been designed to allow code teams without a heavy computer science background to port their codes efficiently and quickly with the software framework of their choosing, while minimizing impact to CPU performance or the scientific readability of the code. Further elements of this strategy include providing a clear and concise recommended strategy to application teams, supporting features that allow porting to GPUs in a piece-meal fashion as well as creating sufficiently general interfaces to facilitate adaptation to future changes without user intervention. This talk will give an overview of AMReX's GPU porting strategy to date. This includes a general overview of the porting philosophy and some specific examples that generated noteworthy lessons about porting a large-scale scientific framework. The discussion will also include the current status of AMReX applications that have begun to migrate to hybrid CPU/GPU systems, detail into GPU specific features that have given substantial performance

gains, issues with porting a hybrid C++/Fortran code to GPUs and an overview of the limitations of the strategy.

Kevin N. Gott
LBNL
kngott@lbl.gov

MS12

Progress on the Development of Mesh-Based PIC for Fusion Codes

An important class of multiscale simulation is the Particle-in-Cell (PIC) method in which particle motion is coupled to fields described by PDEs that are discretized over meshes of the domain of interest. The planned exascale computers will provide the computational power required for the PIC method to be effectively applied to a wide range of scientific and engineering problems. For example, PIC methods are central to many ITER fusion tokamak simulations. Due to their ability to deal with very general geometries and support general anisotropic mesh gradations, these simulation codes are increasingly employing unstructured mesh discretizations of the simulation complex reactor domains. The price that has to be paid when using unstructured meshes is the need to employ irregular data structures and core mesh level operations. The need to attain performance on GPUs has only increased the complexity of developing performant unstructured mesh methods. The presentation will give an overview of efforts underway for a set of structures and methods that will support the developers of unstructured mesh PIC codes. The status of the development of the versions of the two fusion plasma PIC codes, XGC edge plasma and GITR for impurity transport will be given.

Cameron W. Smith
Scientific Computation Research Center
Rensselaer Polytechnic Institute
smithc11@rpi.edu

Gerrett Diamond
Rensselaer Polytechnic Institute
diamog@rpi.edu

Chonglin Zhang
Scientific Computation Research Center Rensselaer Polytechni
zhangc20@rpi.edu

Eisung Yoon
Ulsan National Institute of Science and Technology
esyoon@unist.ac.kr

Gopan Perumpilly, Onkar Sahni
Rensselaer Polytechnic Institute
gopan.p@gmail.com, sahni@rpi.edu

Mark S. Shephard
Scientific Computing Research Center
Rensselaer Polytechnic Institute
shephard@rpi.edu

MS12

Portable Performance for AMR on GPUs: The Proto Approach

The Chombo Team has built and kept up a structured adaptive mesh refinement C++ package for 20 years now and we have made a good long run with MPI+OpenMP parallelism and a simple patch-based parallelism and ChomboFortran. Future exascale computing platforms are placing much more emphasis on accelerators. The programming environments for these heterogeneous computing platforms are still in a great state of flux so the Chombo team has been working on a new abstraction layer Proto to help provide portable high-performance AMR applications across a variety of computer architectures. The key abstractions are pointwise tensor functions and scalar stencil operations with optimized implementations across multiple programming models.

Brian Van Straalen
Lawrence Berkeley National Laboratory
Compuational Research Division
bvstraalen@lbl.gov

MS13

Use of Deep Neural Networks for Estimating Subsurface Property Field from Time-Lapse Geophysical Imaging

The hydrological and biogeochemical processes at the groundwater and river water interface are largely controlled by the exchange dynamics between the two water bodies. Accurate characterization of the heterogeneous permeability field at such interface is critical for modeling the bulk flow as well as the biogeochemical processes that are coupled with the flow. Taking advantage of the distinct conductivities in groundwater and rive water, time lapse electrical resistivity tomography (ERT) can provide rich spatial and temporal data for characterizing the permeability field, by imaging the change in subsurface electric conductivity driven by river water intrusion and retreat. In this study, we demonstrate the use of various deep neural networks including generative adversarial network for developing two-way mappings between the inputs to a system (i.e., heterogeneous permeability field and dynamic river stage etc) and system responses (i.e., the dynamic mixing of river water and groundwater imaged by electric resistivity tomography). By developing such two-way mappings, deep learning methods were able to assist both parameter estimation and surrogate model development that are essential components when developing physics-based predictive models for complex systems. While the new methods are powerful in capturing complex signatures and patterns, we will also share the computational and data challenges we dealt with during the training phase.

Xingyuan Chen, Erol Cromwell, Tim Johnson
Pacific Northwest National Laboratory
xingyuan.chen@pnnl.gov, erol.cromwell@pnnl.gov,
tj@pnnl.gov

Glenn Hammond
Sandia National Laboratories
gehammo@sandia.gov

Hongsheng Wang
Pacific Northwest National Laboratory
hongsheng.wang@pnnl.gov

MS13

Infusing Physics and Domain Knowledge into ML and DL Models

Simulating complex multi-scale physical systems often involves solving partial differential equations (PDEs) with

closures for the unresolved scales. Although the advancement of HPC has made resolving small-scale physics possible, such simulations are still very expensive. Therefore, reliable and accurate closure models for the unresolved physics remains an important requirement for many computational physics problems. Recently, generative adversarial networks (GANs), have shown promise in emulating complex systems without explicitly solving their governing PDEs. However, GANs are known to be difficult to train and to achieve convergence. We present approaches to enforcing constraints either from the training data or from the underlying physics of the system. We also show that such approaches can lead to better performance, measured by (i) the models ability to better emulate physical properties of the system, and (ii) reduced training time. We exemplify this approach on canonical turbulent flows of relevance to fluid and climate dynamics. Given the ever-growing high-fidelity simulation databases of physical systems, this work shows potential as an alternative to the explicit modeling of closures or parameterizations for unresolved physics, which are known to be a major source of uncertainty in simulating multi-scale physical systems such as turbulent flows or Earths weather and climate

Karthik Kashinath
Lawrence Berkeley National Laboratory
kkashinath@lbl.gov

Jinlong Wu
California Institute of Technology
jinlong@caltech.edu

Chiyu Jiang
University of California, Berkeley, U.S.
chiyu.jiang@berkeley.edu

Adrian Albert
Lawrence Berkeley National Laboratory
aalbert@lbl.gov

Heng Xiao
Dept. of Aerospace and Ocean Engineering, Virginia Tech
hengxiao@vt.edu

Philip Marcus
University of California at Berkeley
phil@cfd.me.berkeley.edu

Mr Prabhat
Lawrence Berkeley National Laboratory
prabhat@lbl.gov

MS13

Application of Information Theory in Understanding Hydrological Interactions and Model Diagnostics

Watershed model calibration tunes model parameters to information content found in inputs to fit observations. This tuning has mostly been done using different goodness of fit performance measures (e.g. NSCE, percent bias, etc.). However, such tuning is limited in quantifying and providing insights on how information flows among the various input and output variables and thus whether observed processes are adequately represented within the model. By considering output from uncalibrated and calibrated parameterizations of the National Hydrologic Model run with the Precipitation Runoff Modeling System (NHM-PRMS), this study demonstrated the utility of information theory (IT) in characterizing information flow during model calibration. The results in the model diagnostic suggest that the calibration process overfits the observed streamflow by poorly extracting the information content of input precipitation. On the other hand, by applying transfer entropy, a concept from information theory, the strength, memory length and factors controlling the interaction between streamflow and precipitation are demonstrated across 671 watersheds over the conterminous US. Here, it is found that precipitation can explain up to 50% of the dynamics in streamflow.

Edom Moges
University of California, Berkeley
edom.moges@berkeley.edu

MS13

Satellite Precipitation Estimation at Uci Chrs: Algorithm Development Challenges

Precipitation is a key variable in hydrological processes and varies within time and space. Understanding and monitoring precipitation is crucially important in the human society. Over the past two decades, Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks (PERSIANN) products, including PERSIANN, PERSIANN-CCS and PERSIANN-CDR, have been produced and incorporated in a wide range of studies and applications. While algorithms that exclusively use satellite infrared data as input are attractive owing to their rich spatiotemporal resolution and near-instantaneous availability, their sole reliance on cloud-top brightness temperature (Tb) readings causes over-predictions in wet regions and under-predictions in dry regionsthis is especially evident over the Western Contiguous United States (CONUS). We introduce a new algorithm, the Precipitation Estimations from Remotely Sensed Information using Artificial Neural Networks (PERSIANN) Dynamic-Infrared Rain rate model (PDIR), which utilizes climatological data to construct a dynamic (e.g. laterally shifting) Tb-rain rate relationship that better estimates precipitation totals and distributions, notably over the Western CONUS. This presentation also introduces the recently developed data and information systems by CHRS including RainSphere, iRain, and Data-Portal.

Phu Nguyen
University of California, Irvine
ndphu@uci.edu

Kuolin Hsu, Dan Braithwaite
UC Irvine
kuolinh@uci.edu, dbraithw@uci.edu

Soroosh Sorooshian
University of California, Irvine
soroosh@uci.edu

MS14

Large-Scale Multiphase Flow Simulations with AMR using Dynamic Load Balance

It is one of challenging topics to carry out large-scale simulations for multiphase flows such as gas-liquid or and solid-gas two-phase flows. We have developed a weakly compressible flow solver and applied AMR (Adaptive Mesh Refinement) fine meshes to the gas-liquid interfaces. Dynamic domain partitioning based on space-filling curves has been employed and a few simulations on liquid film dynamics

and tsunami flows with a lot of debris are carried out on 10-100 GPUs of the GPU supercomputer TSUBAME 3.0.

Takayuki Aoki
Tokyo Institute of Technology
taoki@gsic.titech.ac.jp

MS14

Distributed Domain Generation for Large Scale CFD Applications

One of the big challenges in numerical computing on modern high performance clusters for the simulation of real world phenomena is the efficient handling and management of the computational domain. Classical approaches may need to store the complete (logical) topology with each process / node in order to ensure global consistency. This entails significant memory requirements and expensive global communication overhead especially when the domain configuration is altered during run time. Our current work addresses this shortcoming by employing a local, i.e. decentral approach to domain organisation, where the essential idea is to limit the domain view of each participating unit to their direct neighbours. The new approach affects various parts of the computational pipeline significantly, e.g. the octree-based domain generation which had to be revamped from a central based one to support also decentral structures similarly. To this end, an algorithm has been devised which refines an input geometry on all processes up to a predetermined depth, before distributing the resulting leaf nodes of the geometry as starting points for a subsequent refinement on the respective processes. We will discuss the key points of this approach, including the distribution of the root tree among all processes, avoiding costly communication patterns for neighbourhood searches, and an efficient tree balancing algorithm for a decentral environment.

Christoph Ertl
Technical University of Munich
Chair for Computation in Engineering
christoph.ertl@tum.de

Ralf-Peter Mundani
FHGR, Switzerland
Centre for Data Analytics, Visualization and Simulation
ralf-peter.mundani@fhgr.ch

Ernst Rank
Computation in Engineering
Technische Universität München
ernst.rank@tum.de

MS14

Immersive Data Exploration and Analysis of Large Scale CFD Applications

Due to ever increasing advances in hardware, today's HPC systems allow for an interactive treatment of even complex problems stemming from domains like engineering, medicine, or geo-sciences. Such an approach not only bridges the gap between (often batch-oriented) HPC applications and real-time user interaction, it also paves the way for new types of interactive applications in order to obtain insight – not numbers! While those in situ approaches are on the rise, very often they suffer from hardware and/or algorithmic limitations hindering fast and efficient visual exploration in order to experience phenomena that would not be possible or accessible in reality. Within our research, various aspects of different CFD applications (e.g. floods, HVAC) are in the focus of interest. Those applications are often of large scale and, thus, forbid an interactive visual analysis due to their huge data advent. Key feature of our approach is a data exploration technique called sliding window that allows for online/offline visualisations even of huge data sets (up to hundreds of billions of unknowns) during run time. For better visual comprehension of the computed results an immersive virtual reality facility (CAVE) is coupled to the running application. While standing in the CAVE, users have the possibility moving around and zooming into the data, moving forward/backward in time, and manipulating simulation parameters (e.g. boundary conditions) to evaluate different scenarios.

Ralf-Peter Mundani
FHGR, Switzerland
Centre for Data Analytics, Visualization and Simulation
ralf-peter.mundani@fhgr.ch

Sari Ugurcan
Technical University of Munich
ugurcansari93@gmail.com

Christoph Ertl
Technical University of Munich
Chair for Computation in Engineering
christoph.ertl@tum.de

MS15

Detecting and Mitigating Stagnation in Nonlinear Multiphysics Problems

We examine some straightforward tools to understand stagnation of nonlinear iterations for coupled systems, which mainly amount to monitoring portions of the residual. Using this information, we try to guide the choice of nonlinear preconditioner. Some illustrative examples are presented using PETSc.

Matthew G. Knepley
University at Buffalo
Department of Computer Science and Engineering
knepley@gmail.com

MS16

Predict-and-Recompute Conjugate Gradient Variants

The standard implementation of the conjugate gradient algorithm suffers from communication bottlenecks on parallel architectures, due primarily to the two global reductions required every iteration. In this paper, we introduce several conjugate gradient variants, which decrease the runtime per iteration by overlapping global synchronizations, and in the case of our pipelined variants, matrix vector products. Through the use of a predict-and-recompute scheme, whereby recursively updated quantities are first used as a predictor for their true values and then recomputed exactly at a later point in the iteration, our variants are observed to have convergence properties nearly as good as the standard conjugate gradient problem implementation on every problem we tested. It is also verified experimentally that our variants do indeed reduce runtime per iteration in practice, and that they scale similarly to previously studied communication hiding variants. Finally, because our variants achieve good convergence without the use of any

additional input parameters, they have the potential to be used in place of the standard conjugate gradient implementation in a range of applications.

Tyler Chen
University of Washington
chentyl@uw.edu

MS16

Improving Attainable Accuracy of the Deep Pipelined Conjugate Gradient Algorithm

Conjugate Gradients (CG) is one of the most widely used algorithms for solving linear systems with a symmetric positive definite matrix. The classical formulation due to Hestenes and Stiefel requires two global reduction phases in each iteration, namely for the calculation of the dot-products. On HPC hardware this causes a bottleneck in the parallel scalability of the method. When executing CG on a large number of compute nodes the total runtime is dominated by the time it takes to communicate data. Moreover, during the global communication phase the cores stay idle waiting for the result of the dot-product. The main idea of the deep pipelined CG algorithm is to loosen data dependencies in the classical formulation of the method by introducing auxiliary variables, such that the cores can continue to perform useful computations during the global communication phase. These auxiliary variables allow us to overlap the global reduction with the computation of l sparse matrix vector products, where l is a parameter of the algorithm called the pipeline length. The way the auxiliary variables are computed has a huge impact on the maximal attainable accuracy of the method. In this talk we present two different ways to recursively calculate the auxiliary variables, equivalent in exact arithmetic, but with significantly different finite precision behaviour. We analyse the round-off error in the recurrence relations and present numerical experiments which illustrate the analysis.

Jeffrey Cornelis, Siegfried Cools, Wim Vanroose
University of Antwerp
jeffrey.cornelis@uantwerp.be, siegfried.cools@uantwerp.be, wim.vanroose@uantwerp.be

MS16

Stable One-Reduce Gram Schmidt Orthogonalization Algorithms

In many important algorithms (i.e. Krylov solvers, eigenvalue computations) orthogonality among multiple vectors is required. The orthogonality is usually achieved using one of the Gram-Schmidt variants or using Householder decomposition. The modified Gram-Schmidt algorithm can produce $O(\varepsilon)\kappa(A)$ orthogonal vectors and this is sufficient for the backward stability of the MGS-GMRES algorithm, (Paige, Rozložnik and Strakoš (2006)). The original MGS algorithm required $k-1$ inner products and a norm for each column or iteration. Iterated classical Gram-Schmidt with re-orthogonalization (CGS-2) requires 2 matrix-vector products and a norm, whereas Householder requires at least 4 reductions. Global reductions are very costly on modern large scale systems, hence minimizing the number of reductions is paramount for performance. It turns out that both modified Gram-Schmidt and CGS2 (and the corresponding Arnoldi-QR) can be reformulated to require only one reduction, while being equivalent to the original algorithm. During the talk, both theoretical and performance results will be presented.

Katarzyna Swirydowicz
NREL
National Renewable Energy Lab
katarzyna.swirydowicz@nrel.gov

Stephen Thomas
National Renewable Energy Laboratory
stephethomas@gmail.com

Julien Langou
University of Colorado Denver
julien.langou@ucdenver.edu

Daniel Bielich
University of Colorado at Denver
daniel.bielich@ucdenver.edu

MS17

Parallelization of Computations Across Hierarchical River Networks

In this presentation, we will highlight some key cyber-infrastructure challenges in continental-domain hydrological modeling. We will discuss 1) multi-scale continental-domain instantiations from the same geospatial framework, introducing novel approaches for adaptive nests; (2) methods in large-domain parameter estimation, focusing on the development of model workflows and computational infrastructure; and (3) parallelization of continental-domain models, with focus on efficient hierarchical/hybrid spatial decomposition strategies that are necessary to efficiently process connected river networks. We will summarize recent progress on each of these topics as well as outstanding research challenges.

Martyn P. Clark
National Center for Atmospheric Research
University of Colorado Boulder
martyn.clark@usask.ca

MS17

Interactive Computing at Scale: Applications in Climate and Hydrologic Modeling

Climate and hydrologic data produced by both simulation models and observational platforms is being made available to researchers at an overwhelming volume. While the availability of large volumes of climate and hydrologic data presents a myriad of new and exciting opportunities for ground breaking research, it also brings to the forefront significant computational challenges that need to be addressed before we can make the most of our big data for geoscientific research. In this presentation, I will outline some of the central computational challenges that face geoscientists today and how the Pangeo Project is working to address these. In particular, I will focus on the Pangeo Project is making interactive computing on very large datasets possible, through the use of open source software, cloud computing, and new data analysis techniques. Finally, I will highlight some of the scientific applications employing the these emerging tools.

Joseph J. Hamman
NCAR

jhamman@ucar.edu

MS17

Parameter Inference for a Massively Parallel Global Hydrologic Model

The Advanced Earth System Modelling Capacity (www.esm-project.net) aims at a more realistic description of the global hydrology at multiple scales. Two of the main challenges to achieve this goal are: (1) a parallel routing algorithm suitable for a high resolution global streamflow network and (2) a parallel multi-scale parameter estimation technique able to assimilate time series of observations and remotely sensed data gathered at multiple resolutions across the globe. Here we use a hybrid MPI-OpenMP parallelization approach in mHM (www.ufz.de/mhm) which recursively cuts off equal sized subdomains to gain computational load balancing while distributing them among all available nodes using several effective scheduling algorithms for the high-performance supercomputer JUWELS. Speedup tests carried out on a European data set [Samaniego et al. 2019 BAMS] indicate that nested multiscale simulations are possible only if the model exhibits a scale-invariant parameterization. In mHM, the parameter estimation is provided via the multiscale parameter regionalization technique [Samaniego et al 2010 WRR] and a parallelized global search algorithm [DDS, Tolson & Shoemaker 2007, WRR]. In this study, transfer-function parameters for mHM are estimated with thousands of global streamflow time series (GRDC), FLUXNET ET products, land surface temperature (MODIS) and the terrestrial total water storage anomaly (GRACE).

Luis Samaniego
Helmholtz Centre for Environmental Research, Germany
luis.samaniego@ufz.de

Maren Kaluza, Stephan Thober, Rohini Kumar, Robert Schweppe, Oldrich Rakovec
Helmholtz Centre for Environmental Research - UFZ
maren.kaluza@ufz.de, stephan.thober@ufz.de, rohini.kumar@ufz.de, robert.schweppe@ufz.de, oldrich.rakovec@ufz.de

MS17

The Canadian Hydrological Model

The Canadian Hydrological Model (CHM) is a C++ simulation code designed to increase future capacity for hydrological simulations over large spatial extents. CHM combines multi-scale unstructured spatial meshes with a plug-in architecture of modular process representations for developing and simulating hydrological processes. This presentation provides an overview of CHM and its capabilities, including recent technical advancements aimed at improving the efficiency and scalability of CHM. These advancements include: 1) Mesh organization for simplified element coupling, 2) Extension to distributed computing environments through use of the Message Passing Interface, and 3) Updated linear algebra solvers to efficiently deal with advection-diffusion differential equation modules at large spatial extents. Preliminary results have shown that these advancements have: 1) Significantly decreased the simulation run time for a fixed simulation size and 2) Vastly increased the spatial domain of a simulation without increasing simulation run time. For example, using the Snowcast domain (www.snowcast.ca) with a snowdrift-resolving mesh (approximately 100 m ridge scale) as a test, the speed of simulation decreased by more than an order of magnitude.

Raymond J. Spiteri
University of Saskatchewan, Canada
Department of Computer Science
spiteri@cs.usask.ca

Christopher B. Marsh, Kevin R. Green
Global Institute for Water Security
University of Saskatchewan
chris.marsh@usask.ca, kevin.green@usask.ca

MS18

DeepSparse: A Task-Parallel Framework for Sparse Solvers on Deep Memory Architectures

Data movement is an important performance bottleneck in large-scale sparse matrix computations, e.g., linear solvers, eigensolvers and graph analytics. We introduce a novel sparse solver framework, named DeepSparse, which adopts a fully integrated task-parallel approach. DeepSparse differs from existing work in that it adopts a holistic approach that targets all computational steps in a sparse solver rather than narrowing the problem into specific kernels (e.g., SpMVs, SpMMs). We present the implementation details of DeepSparse and demonstrate its merit in two commonly used eigensolvers, Lanczos and LOBPCG algorithms. We observe that DeepSparse achieves 2× - 16× fewer cache misses across different layers (L1, L2 and L3) and 2× - 3.9× reduction in execution time over implementations of the same solvers using optimized library function calls on traditional multicore architectures. We also discuss extensions to DeepSparse for facilitating effective memory management in multi-level memory hierarchies that is becoming prevalent in high-end systems as a result of the high-bandwidth device memory and DRAM coupling.

Md Afibuzzaman, Fazlay Rabbi
Michigan State University
afibuzza@msu..edu, rabbimd@msu.edu

M. Yusuf Ozkaya
Georgia Institute of Technology
myozka@gatech.edu

Umit V. Catalyurek
Georgia Institute of Technology
School of Computational Science and Engineering
umit@gatech.edu

H. Metin Aktulga
Michigan State University
hma@cse.msu.edu

MS18

Exploiting Task Parallelism in an Exascale Ecosystem: Some Issues and Possible Solutions

With the advent of the exascale era, the level of scale reached in terms of hardware parallelism, both within nodes and between nodes, is such that new issues have to be addressed by task-based parallel runtime system, to efficiently use computing resources and to orderly interact with the ecosystem. This talk will discuss some of these issues, ranging from parallelism expression, work mapping, dependence management, to interoperability issues, and introduce some possible solutions explored by INRIA Team

STORM in Bordeaux, France.

Olivier Aumage
INRIA, France
olivier.aumage@inria.fr

MS18

Novel Approaches to Optimize and Execute Task-Based, Irregular Applications on Extreme-Scale, Heterogeneous Systems using PaRSEC

In the context of the Epexa project, we research and test the hypothesis that a tight integration between the language, the compiler, and the runtime system enable developers to dramatically improve the performance of applications that target irregular problems on distributed memory, extreme-scale, and heterogeneous architectures. Specifically, we believe this can be achieved by through a software ecosystem that attacks the twin challenges of programmer productivity and portable performance for advanced scientific applications on massively-parallel, hybrid, many-core systems. The new ecosystem is focused upon high-performance implementations of irregular and dynamic computations that are poorly supported by current programming paradigms, while also enabling efficient execution of regular computations such as dense (multi)linear algebra.

Thomas Herault
Univ. of Tennessee - Knoxville
herault@icl.utk.edu

MS18

Asynchronous Programming in Modern C++: What Is An AMT and Why Do You Want One for Christmas?

With the advent of modern computer architectures characterized by many-core nodes, deep and complex memory hierarchies, heterogeneous subsystems, and power-aware components, it is becoming increasingly difficult to achieve best possible application scalability and satisfactory parallel efficiency. The community is experimenting with new programming models that rely on finer-grain parallelism, flexible and lightweight synchronization, combined with work-queue-based, message-driven computation. The recently growing interest in the C++ programming language increases the demand for libraries implementing those programming models for the language. We present a new asynchronous C++ parallel programming model that is built around lightweight tasks and mechanisms to orchestrate massively parallel and distributed execution. This model uses the concept of Futures to make data dependencies explicit, employs explicit and implicit asynchrony to hide latencies and to improve utilization, and manages finer-grain parallelism with a work-stealing scheduling system enabling automatic load balancing of tasks. We have implemented such a model as a C++ library exposing a higher-level parallelism API that is fully conforming to the existing C++11/14/17 standards and is aligned with the ongoing standardization work. This API and programming model has shown to enable writing highly efficient parallel applications for heterogeneous resources with excellent performance and scaling characteristics.

Hartmut Kaiser
Louisiana State University
hkaiser@cct.lsu.edu

MS19

Simulation of Density-Driven Flow in Aquifers with Phreatic Surface

Phreatic surface is a typical element of subsurface groundwater flow. It influences the dynamics of the liquid phase significantly and should not be neglected in the modeling. In particular, tracking of its motion is very important in simulations of the density-driven groundwater flow which is very sensitive to the problem setting. We present an approach for the simulation of the density-driven flow in aquifers with complicated layered geometry whose hydrogeological structure approximates real-world geological formations. The phreatic surface is modeled as a moving boundary of the saturated part of the domain. To resolve the geological layers by the discretization grid, we represent this surface by the level-set approach and discretize the boundary conditions on it by the ghost-fluid method. For the numerical solution of the model, we use a vertex-centered finite-volume discretization. The discretized equations are linearized in the Newton's method, in which the geometric multigrid solver with the ILU smoothing is used. The implementation is parallelized. Examples of the simulations are presented.

Dmitry Logashenko, Gabriel Wittum
Extreme Computing Research Center
King Abdullah University of Science and Technology
dmitry.logashenko@kaust.edu.sa,
gabriel.wittum@kaust.edu.sa

MS19

Homogenized Modeling of Microscopic Anisotropic Diffusion for Effective Diffusivities in Stratum Corneum

In pharmacology, mathematical models of molecular diffusion through the stratum corneum (SC) layer can be greatly used to predict transdermal delivery of drugs and to risk assessment of chemical exposures. The membrane microstructure is generally accepted as a periodic arrangement of corneocytes embedded in an anisotropic lipid matrix. We show how the effective diffusivity can be calculated using mathematical models from homogenization theory. Numerical results are expressed in the dimensionless effective diffusivity tensor $\bar{A}^{SC}$, which is a function of homogenization results χ_i in each field. χ_i is determined by the dimensionless parameters, λ and γ. λ is the degree of anisotropy in a lipid bilayer and γ is the degree of drug hydrophilicity between corneocyte and lipid phases. The effective diffusion tensors are finally calculated for a classical and a spatial "brick-and-mortar" structures and a tetrakaidekahedral-shaped cell. A comparison with a previous study confirms that transdermal and lateral diffusivities are strongly influenced by λ due to the barrier function of the lipid phase against transbilayer flow. The lateral diffusivities of poorly water-soluble drugs are limited by γ in contract to transdermal diffusivity. This presented method may more precisely and more intensively predict the drug permeability and evaluate the membrane microstructure.

Junxi Wang
Goethe-Center for Scientific Computing
Goethe-University Frankfurt

junxi.wang@gcsc.uni-frankfurt.de

MS20

TAMM: Tensor Algebra for Many-Body Methods

Tensor contractions constitute a computationally critical part of many quantum many-body methods. To enable performance-portable implementation of electronic structure methods on the upcoming exascale computers, we have been developing the Tensor Algebra Framework for Many-body Methods (TAMM) runtime framework. In this talk, I will present the design of TAMM and its approach to performance optimization and productive development of new methods. I will also present recent results in efficient implementations of Coupled Cluster methods using TAMM.

Sriram Krishnamoorthy
Pacific Northwest National Laboratory
sriram@pnnl.gov

Erdal Mutlu
Pacific Northwest National Lab
erdal.mutlu@pnnl.gov

Ajay Panyala
High Performance Computing Division
Pacific Northwest National Laboratory
ajay.panyala@pnnl.gov

MS20

Model-Driven Tile Optimization for Tensor Contractions

Data movement between the processor and memory hierarchy is a fundamental bottleneck that limits the performance of many applications on modern computer architectures. Tiling and loop permutation are key techniques for reducing data movement in the memory hierarchy. However, selecting effective tile-sizes and permutations of tiling loops is particularly challenging for tensor contractions due to the large number of loops. We describe an analytical model-driven approach to multi-level tile size optimization and permutation selection for tensor contractions.

Ponnuswamy Sadayappan
Ohio State University
saday@cs.utah.edu

Rui Li
University of Utah
lirui@cs.utah.edu

Atanas Rountev
Ohio State University
rountev@cse.ohio-state.edu

Aravind Sukumaran-Rajam
Washington State University
aravind_sr@outlook.com

Fabrice Rastello
INRIA
fabrice.rastello@inria.fr

Richard Veras
Louisiana State University
richard.m.veras@gmail.com

Tze Meng Low
Carnegie Mellon University
lowt@andrew.cmu.edu

MS20

A Hybrid Analytical/Machine-Learning Model to Optimize Tensor-Contractions on GPUs

Designing performance models for code generators The high popularity and demand for domain-specific computing and heterogeneous computing have increased the popularity of domain-specific optimizers and code generators. The performance of the generated code depends on the optimization parameters, and thus, selecting the right optimization configuration is of utmost importance to achieve good performance. Unfortunately, the space of valid optimization configurations is often exponential. As an example, consider Tensor Contractions (TC) – higher-dimensional analogs of matrix multiplication. TC can be implemented using (i) Direct approach (no transpose) and (ii) TTGT (transpose + GEMM). Along with these algorithmic choices, a TC code generator has to determine the optimal loop permutation and tile size choices. Building an efficient performance model helps to estimate the cost of each such configuration. Reasonably good performance models can be built using machine learning. However, using a pure machine learning based model potentially ignores any domain-specific information. In this talk, we present the design of an analytical model for tensor contractions and how such a model can be combined with machine learning models to select optimal or close to optimal optimization configuration for tensor contractions.

Aravind Sukumaran-Rajam
Washington State University
aravind_sr@outlook.com

Jinsung Kim
Ohio State University
kim.4232@osu.edu

MS21

Recent Half Precision Developments in the MAGMA Library

The huge advancements in machine learning algorithms and artificial intelligence applications have developed a significant demand for high performance low-precision arithmetic. This resulted in vendors providing hardware that is capable of performing native low-precision operations, such as the 16-bit floating point arithmetic (i.e. half precision) in modern GPUs. However, it has been shown that half precision can still be useful in other domains, such as mixed precision linear solvers. In this talk, we present the latest developments in the MAGMA library (Matrix Algebra for GPUs and Multicore Architectures) for dense linear solvers with mixed precision arithmetic and iterative refinement algorithms. We show that, under certain conditions, it is possible to exploit the high performance of half precision in solving linear systems, while maintaining the accuracy of higher precisions (32-bit or even 64-bit).

Ahmad Abdelfattah
University of Tennessee, Knoxville
ahmad@icl.utk.edu

MS21

Mixed Precision Numerical Techniques Acceler-

ated with Tensor Cores and its Impact on Todays Scientific Computing and Implications for Tomorrows Hardware Design

Double-precision-floating-point has been the-de-facto standard for doing scientific simulation for several decades. Problem complexity and the sheer magnitude of data coming from various instruments and sensors motivate researchers to mix and match various approaches to optimize compute resources, including different levels of floating-point precision. In recent years, the big bang for machine learning has focused significant attention on half-precision. We explored the possibility of using FP16/FP32-Tensor-Cores on NVIDIA-Volta-GPUs to accelerate one of the most common linear algebra routines without loss of accuracy. We achieved a 4x performance increase and 5x better energy efficiency versus the standard FP64 implementation while providing a solution with FP64 accuracy. We studied a plasma fusion application that simulates the instabilities that occur inside a plasma inside the International-Thermonuclear-Experimental-Reactor (ITER). We show that using our mixed precision solver that harnesses the FP16/FP32-Tensor-cores in Volta GPUs, it is possible to simulate the instability between plasma beams 3.5x faster.

Azzam Haidar, Harun Bayraktar
NVIDIA
azzamhaidar@nvidia.com, hbayraktar@nvidia.com

Timothy Costa
Nvidia
tcosta@nvidia.com

MS21

Reduced Precision in Weather Forecasting Models

Double-precision arithmetic has become standard in numerical codes. However, in real applications, most of this precision is unnecessary. This is particularly true in the field of weather and climate modelling, in which model error is large, and parameterisations of unresolved physical processes introduce reasonably big errors at the grid scale. Indeed, several operational weather prediction models now successfully use single-precision in some or all of their code. Using software emulation of reduced precision, we have lowered numerical precision even further than this in a range of problems, from toy ODE models to components of state-of-the-art forecast systems. I will summarise some of this work, including identifying workarounds and calculations that require higher precision, with an eye towards future hardware in which precision may be adjustable on a per-operation basis.

Andrew McRae
The University of Oxford
andrew.mcrae@physics.ox.ac.uk

MS21

Solving Neural ODEs using Fixed-Point Arithmetic with Stochastic Rounding

Increasingly more processors are designed without hardware support for floating-point arithmetic with an expectation that either fixed-point arithmetic will be enough for intended applications or that software floating-point will be feasible. In some specialized systems, even fixed-point support is highly limited due to reduced precision (8- or 16-bit) integer arithmetic units. The main reasons for doing this are potential improvements in energy efficiency and memory footprint and bandwidth. However, simply switching to low-precision data types and operations typically results in increased numerical errors. We investigate approaches to improving the accuracy of fixed-point arithmetic types, using examples in an important domain for numerical computation in neuroscience: the solution of Ordinary Differential Equations (ODEs). The Izhikevich neuron model is used to demonstrate that rounding has an important role in producing accurate spike timings from explicit ODE solution algorithms. In particular, fixed-point arithmetic with stochastic rounding consistently results in smaller errors compared to single-precision floating-point and fixed-point arithmetic with round-to-nearest across a range of neuron configurations and ODE solvers. The experiments were run on the SpiNNaker chip, an 18-core (ARM968) 32-bit integer processor.

Mantas Mikaitis
The University of Manchester
mantas.mikaitis@manchester.ac.uk

MS22

Optimizing Molecular Dynamics Simulations with Dynamic Auto-Tuning

Molecular Dynamics simulations are highly diverse by nature and require different optimization strategies depending on the scenario. For example, the simulation of a homogeneous gas poses different challenges than droplets. To make matters worse this optimum can change over the course of a simulation if said gas begins to form droplets. To tackle this challenge we published the C++ library project AutoPas, which acts as a node-level performance library for arbitrary N-Body simulations. It implements multiple types of particle containers, parallelization strategies, data layouts, and further optimization techniques. During runtime AutoPas then chooses the fastest combination of aforementioned aspects for an optimal time to solution of the pairwise force calculation in the system. Throughout the simulation, the library periodically reevaluates the current state of the simulation and is capable to adapt all of its internal algorithm configurations to sustain optimal performance for the whole duration of the simulation. In this talk, we present recent optimization options and parallelization strategies implemented in the library and their impact while integrated into the simulation software ls1-mardyn. The introduction of even more parameters and choices inevitably increases the search space for the auto-tuning process. Therefore techniques are discussed how to efficiently find a near-optimal configuration and first results are presented.

Fabio A. Gratl
Technical University Munchen, Germany
f.gratl@tum.de

Steffen Seckler
Technical University of Munich
Department of Informatics
steffen.seckler@tum.de

Hans-Joachim Bungartz
Technical University of Munich, Department of Informatics
Chair of Scientific Computing in Computer Science
bungartz@in.tum.de

Philipp Neumann

Helmut Schmidt University, Germany
philipp.neumann@hsu-hh.de

MS22

Particle Sorting for Projection Based Particle Methods

The projection-based particle method including incompressible SPH and moving particle semi-implicit both solve pressure Poisson equation using HelmHoltz-Leray decomposition and application of Chorins projection method. Comparing with weakly compressible SPH, projection-based particle methods generally have higher accuracy for pressure and volume conservation. Due to the Lagrangian nature of the particle methods, such as SPH or incompressible SPH, the pattern of data access and computation are unknown until the applications runtime. This often leads to poor temporal and spatial data accesses and insufficient usage of a memory hierarchy. Sorting algorithms are critical techniques to increase applications data locality. In this talk, we have investigated the effects of the particles sorting for projection based particle method, particularly studied the effects of natural order sorting, Hilbert and Morton space filling curve sorting. The data layout changes because of sorting and implementation procedure have been described in detail. We have found that the particles sorting does not only improve the general SPH kernels performance, but also has an influence on the sparse matrix structure of the pressure Poisson equation. We have also discussed and compared the performance difference between space filling curve sorting methods and natural order sorting both in serial and parallel with typical violent flow.

Xiaohu Guo
Hartree Centre
Science and Technology Facilities Council
xiaohu.guo@stfc.ac.uk

MS22

Multi-Architecture Parallel Particle Simulations on HPC Systems

Parallel numerical simulation using particles and/or meshes are common in both scientific and industrial applications. Popular numerical methods that fall into this category are Smoothed Particle Hydrodynamics (SPH), Vortex Methods (VM), Molecular Dynamics (MD), Discrete Element Methods (DEM), Finite-Difference Methods (FDM), Particle-in-Cell (PIC) codes, and Finite-Volume methods (FVM). Parallel implementation of all of these methods can be expressed as a set of common abstract data types and operators, as implemented in the OpenFPM middleware. OpenFPM is a scalable C++ framework for rapid and efficient development of particle- and particle-mesh simulations. It uses C++ Template Meta-Programming for compile-time code generation in order to allow arbitrary C++ objects to be particle or mesh node properties, and perform simulations in arbitrary-dimensional spaces. Originally developed for shared- and distributed-memory CPU computing, we here present an extension of OpenFPM to GPUs and multi-GPU setups. We demonstrate two different ways of implementing GPU computations (using kernels and using inline code) in OpenFPM without explicitly having to write any CUDA, OpenCL, or the like. This renders GPU computing much more accessible and enables rapid porting of existing CPU codes to GPUs and to using a mixture of CPU and GPU operations in a distributed environment.

Pietro Incardona
TU Dresden, Germany
incardon@mpi-cbg.de

Ivo F. Sbalzarini
TU Dresden, Computer Science
& MPI Molecular Cell Biology & Genetics
ivos@mpi-cbg.de

MS22

Extending Quantum Molecular Dynamics to the Exascale: Latte, Progress and BML Libraries

We have developed a modern framework to extend Quantum Molecular Dynamics (QMD) simulation to the exascale. The main deliverable consists of a flexible library ecosystem for Quantum Chemistry Applications adapted to pre-exascale architectures. This framework is composed by two libraries: BML which is a low-level API for linear algebra or mathematically related operations, and PROGRESS which consists of a collection of O(N) electronic structure solvers that rely entirely on BML. PROGRESS implements an electronic structure solver based on a recursive Fermi operator expansion (SP2) that exploits the underlying matrix sparsity. In order to achieve distributed memory parallelism, a graph-based approach where the calculations are distributed across multiple nodes following a predicted data dependency graph has been implemented. By combining the two approaches, practical QMD simulations of million atom biomolecular systems are within reach in the very near future.

Christian Negre
Los Alamos National Laboratory
cnegre@lanl.gov

MS23

The Sparse Matrix Vector Product on High-End GPU Clusters

Efficient processing of Irregular Matrices on SIMD-type architectures is a persistent challenge. Resolving it requires innovations in the development of data formats, computational techniques, and implementations that strike a balance between thread divergence, which is inherent for irregular matrices, and padding, which alleviates the performance-detrimental thread divergence but introduces artificial overheads. To this end, we address the challenge of designing high performance sparse matrix-vector product kernels for GPUs with a coordinate format suitable for unbalanced matrices. We also provide a hybrid algorithm that stores part of the matrix in SIMD-friendly ELL format. The ratio between the ELL- and the COO-part is determined using a theoretical analysis of the nonzeros-per-row distribution. For the over 2,800 test matrices available in the Suite Sparse matrix collection, we compare the performance against vender kernels providing the same functionality.

Hartwig Anzt
Steinbuch Centre for Computing
Karlsruhe Institute of Technology
hartwig.anzt@kit.edu

Terry Cojean, Yuhsiang M. Tsai
Karlsruhe Institute of Technology

terry.cojean@kit.edu, yu-hsiang.tsai@kit.edu

MS23

Using the PETSc/TAO ADMM Methods on GPUs

The Alternating Direction Method of Multipliers (ADMM) is an algorithm for structured numerical optimization problems that solves a sequence of subproblems that update the variables and the multipliers. The variables can typically be further subdivided and updated in parallel. In this talk, we explore using the ADMM code in the Toolkit for Advanced Optimization for solving structured regression problems on GPUs and provide some initial performance results.

Hansol David Suh
Georgia Tech
hsuh7@gatech.edu

Alp Dener
Argonne National Laboratory
Mathematics and Computer Science Division
adener@anl.gov

Tobin Isaac
Georgia Tech
tisaac@cc.gatech.edu

Todd Munson
Argonne National Laboratory
Mathematics and Computer Science Division
tmunson@mcs.anl.gov

MS23

Vendor-Optimized vs. Portable Performance: Approaches to Get Both

Scientific software should run at maximum performance on any hardware. Reaching this noble ideal can be simplified by reusing well-tuned low-level routines provided by hardware vendors. However, these vendor libraries only provide good performance on the vendor's hardware. As such, scientific software has to either interface different vendor libraries, or provide portable performance by itself. If no vendor-tuned implementations of a particular functionality exist, a strategy for achieving portable performance is required anyway. On the one hand, this talk surveys available strategies and experience on providing portable performance. On the other hand, a software library approach is presented that acts as an intermediate wrapper to provide a single, unified interface for several vendor-optimized linear algebra libraries. This resolves the burden of interfacing many different vendor libraries; instead, only the intermediate wrapper needs to be interfaced, while preserving the full performance benefit of vendor libraries tuned for the particular hardware available.

Karl Rupp
TU Wien
me@karlrupp.net

MS23

Optimization of SpMV on GPU for Iterative Solvers in PETSc

Large sparse systems resulted from the discretization of partial differential equations often need to be solved by iterative linear solvers due to scalability and memory constraints. Typically the computation of sparse matrix-vector (SpMV) products is the workhorse of iterative linear solvers. There are several optimized general-purpose SpMV implementations available, such as Intel MKL, CUSPARSE, CUSP, and ViennaCL. However, to use these libraries, explicit conversion between different matrix format may be required, which hampers the performance. Thus it is desirable and beneficial to have a native implementation of SpMV in the solver that can be optimized to exploit the characteristics of a hardware platform (e.g. memory bandwidth, thread-level parallelism) as well as the characteristics of the application problem (e.g. the block structure, sparsity pattern). In this talk, I will present our recent work on optimizing the Sliced Ellpack (SELL) matrix class in PETSc on NVIDIA GPUs. SELL was previously designed for multicore CPUs with long SIMD and tailored to the needs of PETSc iterative solvers and preconditioners. I will show how we extend SELL for better vectorization on GPU and compare the performance of SELL with other options supported by PETSc (ViennaCL and CUSPARSE).

Hong Zhang
Argonne National Laboratory
hongzhang@anl.gov

MS24

Efficient Sparse Triangular Solve

In the context of the solution of sparse system of equations using direct methods, many efforts have been dedicated to improve the performance of the factorization. However the triangular solve still suffers from a lack of optimisation. It is especially true when looking at preconditioned iterative methods such as the Preconditioned Conjugate Gradient. The application of the Block Jacobi preconditioner is performed at each iteration whereas the factorization of the diagonal block is done once. It often results a time to solve led by the application of the preconditioner. This phasis becomes critical when considering a large number of right-hand sides. In this talk, we present the parallelization of the triangular solve phasis using a task based programming model in a the sparse Chokesly solver called SpLLT [Duff, Hogg, and Lopez. Numerical Algebra, Control and Optimization. Volume 8, 235-237, 2018]. Using OpenMP runtime as well as BLAS3 operations, we show that this approach can outperform state-of-the-art solvers such as MKL Pardiso. We also interface SpLLT with Enlarged Conjugate Gradient [Grigori, and Tissot. Research Report RR-9023, Inria] where the number of directions of research is typically 10.

Sebastien Cayrols, Florent Lopez
University of Tennessee, Knoxville
sebastien.cayrols@gmail.com, flopez@icl.utk.edu

Iain Duff
Science & Technology Facilities Council, UK
and CERFACS, Toulouse, France
iain.duff@stfc.ac.uk

MS24

Parallel Direct Matrix Factorization Methods for Dynamic Optimization

Dynamic optimization problems are a special class of optimization problems that involve decision or control variables, and environment or state variables which vary with time. Their goal is to minimize/maximize some cost func-

tional which is a function of these variables, subject to a set of differential and algebraic constraints. The most common approach to solve these problems is a direct control transcription that replaces the differential equations with discretized approximations, hence transforming the dynamic optimization problem into a nonlinear programming problem (NLP). That NLP can then be solved with a dedicated NLP solver which more often than not utilizes Newtons method to compute the step at every iteration. Furthermore Newtons method at its core involves the solution of a symmetric indefinite linear system stemming from the KarushKuhnTucker (KKT) conditions. And for problems with a large number of variables, solving the resulting large sparse linear systems is the most expensive computational part. Thus it becomes crucial to use a parallel sparse direct solver that can take advantage of the specific sparsity patterns of these KKT systems.

Samah Karim, Edgar Solomonik
University of Illinois at Urbana-Champaign
swkarim2@illinois.edu, solomon2@illinois.edu

Ryne Beeson
CU Aerospace
beeson@cuaerospace.com

MS24

Communication-Avoiding Sparse Direct Solvers for Linear Systems & Graph Problems

Graph shortest path problems are min-plus semiring equivalent of solving a system of linear equations. In this talk, we show how to exploit graph sparsity in the Floyd-Warshall (Fw) algorithm for the all-pairs shortest path (Apsp) problem. (Fw) is an attractive choice for Apsp on high-performing systems due to its structural similarity to solving dense linear systems and matrix multiplication. However, if sparsity of the input graph is not properly exploited, (Fw) will perform unnecessary asymptotic work and thus may not be a suitable choice for many input graphs. To overcome this limitation, the key idea in our approach is to use the known algebraic relationship between Fw and Gauss-Jordan elimination, and import several algorithmic techniques from sparse Cholesky factorization, namely, fill-in reducing ordering, symbolic analysis, supernodal traversal, and elimination tree parallelism. When combined, these techniques reduce computation and communication, improve locality and enhance parallelism. We implement these ideas in an efficient shared memory parallel prototype that is orders of magnitude faster than an efficient multi-threaded baseline (Fw) that does not exploit sparsity. Our experiments suggest that (Fw) algorithm can be competitive with Dijkstra's algorithm (the algorithmic core of Johnson's algorithm) for several classes sparse graphs.

Piyush Sao
Oak Ridge National Laboratory
saopk@ornl.gov

Ramakrishnan Kannan
Oak Ridge National Laboratories
kannanr@ornl.gov

Prasun Gera, Richard Vuduc
Georgia Institute of Technology
prasun.gera@gatech.edu, richie@cc.gatech.edu

MS25

On Performance Portability - Application and Illustration on MC Neutron Transport Application

Today and future supercomputer architecture are more an more complex and heterogeneous. It is now quite impossible to optimize a full application on all the different architectures. Moreover, having portions of code dedicated to specific architecture is a major concern regarding maintenance and portability of the full application. Portability of performances is thus a way to explore for large scientific application, managing a tradeoff between code portability, maintenability and performances. We illustrate this approach on a Monte Carlo Neutron Transport code developped at CEA. A partial offload version of the code has been developed via several programming models (OpenMP thread, OpenMP offload, OpenACC and CUDA) on different architectures (x86, Power and GPU). Moreover we have tried to define and use preformance-portbaility metrics to compare the different implementations.

Christophe Calvin
CEA/DRF - Maison de la Simulation
christophe.calvin@cea.fr

Emeric Brun
CEA/DEN
emeric.brun@cea.fr

Tao Chang
CEA/DEN - Maison de la Simulation
tao.chang@cea.fr

MS25

Exascale Node-Level Parallel Programming Environments: Overview and Deciding What's Right for You

Accelerated architectures have been around in some form or another for decades. For example, Control Data and Cray vector systems were available nearly 40 years ago. In the recent past, Nvidia GPUs have been the most prominent accelerated devices, and now represent a mature software and hardware ecosystem that reliably delivers performance across a broad set of applications. In the coming phase of high-performance computing, accelerated architectures will represent a greater diversity of platforms. Vector processors based on Arm and its scalable vector extensions (SVE) represent a new approach to vectorization requiring significant algorithmic concurrency. In addition, new GPU accelerators are emerging which, while similar in concurrency strategy as Nvidia, represent new software stacks and important variations in processor details. The advent of these new platforms means that performance portability becomes increasingly important. In this presentation, we give a brief overview of publicly-available information about new highly-concurrent processors (vector and GPU) and discuss the programming model and environment options available to programmers, along with some guidance to help programmers make good choices.

Michael A. Heroux
Sandia National Laboratories
St. John's University

maherou@sandia.gov

MS25

A Taxonomy of Distributed and Parallel Languages for High Performance Tasks-Based Multilevel Computing

Extreme scale supercomputer programming often mix coarse-grained distributed programming and parallel programming. Task based programming paradigms based on graph of task/components are proposed to compute and exchange data on those large number of core machines. Scheduling these tasks, and associated communications, is critical to obtain performance and minimize energy consumptions. It is necessary to extract the most expertise from the users in order to optimize these criteria. Then, language describing these task graphs must be well-designed and adapted to end-user programming skills. Moreover, each task-component may have several levels of parallel programming itself. Therefore, we have to propose adapted interfaces between the task graph languages and the parallel languages used in the tasks. On the other hand, pre-existing parallel software may be used as tasks and tasks may be developed using different languages. In this talk, we present experiments and analyses of some block linear algebra methods using several languages proposed for such programming on a cluster of multi-core processors. We propose results with respect to several parameters such as : the number of tasks, the number of cores per tasks, the size of the matrices and the number of blocks. We propose a taxonomy tentative for such languages and we discuss how large applications may be developed using these programming paradigms, in particular in geoscience.

Serge Petiton
University Lille 1, Science and Technologies - CNRS
serge.petiton@univ-lille.fr

Jerome Gurhem
University of Lille, CNRS
jerome.gurhem@univ-lille.fr

Henri Calandra
Total
CSTJF, Geophysical Operations & Technology, R&D Team
henri.calandra@total.com

MS25

AMR Framework for Large-Scale Simulations on Multiple GPUs

Recently grid-based physical simulations with GPU require effective methods to adapt grid resolution to certain sensitive regions of simulations. An adaptive mesh refinement (AMR) method is one of the effective methods to compute certain local regions that demand higher accuracy with higher resolution. To develop the applications adopting AMR effectively with maintaining high performance on multiple GPUs, we are developing a block-based AMR framework for stencil applications written in C++ and CUDA. The programmer simply describes a C++11 lambda that updates a grid point, which is applied to the entire grids with various resolution over a tree-based AMR data structure effectively. The framework also provides the halo exchange between GPUs based on the temporal blocking method, which contributes to performance improvement. The framework-based application for compressible flow has demonstrated good weak scalability with 84% of the parallel efficiency on the TSUBAME3.0 supercomputer at Tokyo Institute of Technology. In this talk, we provide the programming model and implementation of the AMR framework for multiple GPUs, and show the computation results of the compressive fluid calculation based on the proposed AMR framework.

Takashi Shimokawabe
Information Technology Center
The University of Tokyo
shimokawabe@cc.u-tokyo.ac.jp

Naoyuki Onodera
Center for Computational Science & e-Systems
Japan Atomic Energy Agency
onodera.naoyuki@jaea.go.jp

MS26

Improving the Performance of GMRES with Mixed Precision

Krylov solvers, such as GMRES, are often memory bound on modern systems due to the gap in improvement between arithmetic performance and memory access performance. Reducing the precision of the preconditioner is a known approach to improve performance by reducing the pressure on the memory bandwidth. However, this idea has not been extended to use lower precision for other parts of the solver. By judiciously storing vectors in single precision, we are able to achieve the accuracy of a fully double precision GMRES implementation while reducing the memory traffic and footprint. This approach for mixed precision provides an avenue for improving the performance of GMRES, and possibly other solvers, without compromising the accuracy of the solution.

Neil Lindquist
University of Tennessee
nlindqu1@icl.utk.edu

MS26

Wasserstein Discriminant Analysis with Eigensolver

The original Wasserstein Discriminant Analysis paper uses automatic differentiation to compute the partial derivative of the optimization objective with respect to the subspace P, and then uses projected gradient descent to compute the optimal P. In practice, this approach often takes too many iterations to reach convergence when applied to real data and the unconverged solution does not serve as a good approximation to the true optimal. We instead relax the optimization problem to a generalized eigenvalue problem and solve the nonlinear generalized eigenvalue problem using the Self-consistent field iteration. The new solver, which usually converges within 10 iterations in practice, provides a huge speedup in convergence compared to the original solver and therefore aids in iteratively applying WDA to find the most discriminative subspace in an unsupervised setting.

Hexuan Liu
Applied Mathematics
University of Washington

hl65@uw.edu

MS26

Polynomial Preconditioned GMRES in Trilinos: Practical Considerations for High-Performance Computing

Polynomial preconditioners for Krylov solvers are well-known but are not normally used in large-scale software libraries or applications. This may be due to stability problems or complicated algorithms, especially for nonsymmetric matrices. We implement a new GMRES polynomial preconditioner in the software library Trilinos and demonstrate that it is stable and effective for parallel computing. We give several examples with GMRES and compose the polynomial with other preconditioners such as ILU(k) and algebraic multigrid. Trade-offs when selecting a polynomial degree and combining with other preconditioners are analyzed. We also discuss communication-avoiding properties of the polynomial preconditioner.

Jennifer A. Loe
Baylor University
jennifer_loe@baylor.edu

Heidi K. Thornquist
Sandia National Laboratories
hkthorn@sandia.gov

Erik G. Boman
Center for Computing Research
Sandia National Labs
egboman@sandia.gov

MS26

The Lanczos Method in Data Science: New Challenges and the Continued Importance of Stability

The Lanczos method is one of linear algebras most remarkable and enduring algorithms. Introduced nearly 70 years ago, this algorithm is uniquely multipurpose: it can be used to invert linear systems, to compute eigenvectors, and to approximately solve dozens of other important matrix problems. For many of these problems, the Lanczos method gives both state-of-the-art theoretical guarantees and is the method of choice in practice. I will discuss a resurgence of interest in applying the Lanczos method to new applications in machine learning and computational data science, including regularized function fitting, online eigenvector estimation, and spectral analysis. In data applications, Lanczos is especially powerful when composed with highly scalable stochastic optimization or randomized sketching methods. Combining Lanczos with these inherently noisy tools raises questions about the method's stability that tie back to decades old work on understanding how the algorithm performs in finite precision. I will discuss recent work on addressing some of these questions, as well as a number of interesting open problems.

Christopher Musco
Princeton University
cmusco@nyu.edu

MS27

Asynchronous Task-Based AMR with Distributed Parallel Objects

Cello is a highly scalable "array-of-octree" based adaptive mesh refinement (AMR) software framework, implemented using Charm++, an object-oriented message-driven parallel programming system. Enzo-E, being developed concurrently with Cello, is a branch of the ENZO astrophysics and cosmology application that has been modified to use the Cello scalable AMR framework. The Cello framework provides a scientific application with mesh adaptivity, data-driven ghost cell refresh, generic field and particle data types, and task-based asynchronous distributed computation on block data. In this presentation we describe Cello's distributed data structures and asynchronous algorithms, which include a revised buffered refresh scheme, and a recently-implemented domain-decomposition based scalable gravity solver. We also present parallel scaling results of Enzo-E simulations of cosmological structure formation, including more recent experiments with dynamic load balancing.

James Bordner
University of California, San Diego
jobordner@ucsd.edu

Michael L. Norman
University of California at San Diego
mlnorman@ucsd.edu

MS27

Direct Parallel Visualization Using Forest-of-Octrees Meshes

The visualization of simulation data is a primary scalability challenge arising in large-scale numerical simulation. Writing data to disk for subsequent post-processing is slow and often entirely impracticable. Interfacing and linking to third-party visualization libraries has become fairly popular, but usually requires to duplicate and reformat the simulation data for the library, often losing the advantage of native acceleration data structures. We are thus investigating parallel algorithms that natively support fast mesh refinement and partitioning as well as efficient in-situ visualization, in our case based on a forest-of-octrees design, and present the current state of our research.

Carsten Burstedde
Universität Bonn
burstedde@ins.uni-bonn.de

MS27

AMR in Core-Collapse Supernova Simulations

Successful simulations for scientific discovery on high-performance computing (HPC) platforms require careful planning, including verification of specific application configuration and runtime parameters, estimation of resource requirements, and steering and monitoring of the simulation. With rising heterogeneity in both platforms and solvers within applications, effective use of the available hardware is becoming more challenging on HPC platforms. Application configuration has to evaluate the mapping of computation to hardware resources with cost-benefit analysis that weighs the overhead of using the target resource (e.g. data movement) against its computational efficiency in addition to the other concerns of simulation planning. The runtime at the application level needs this information to orchestrate the task assignment and data movement between devices on the compute nodes, and between different nodes. We present a methodology for building component-based cost models that can help in resource and simulation planning, and illustrate the methodology through formu-

lation of a cost model for simulating thermonuclear supernovae using FLASH, a highly configurable adaptive mesh refinement based community code.

Anshu Dubey
Argonne National Laboratory
adubey@anl.gov

J. Austin Harris
Oak Ridge National Laboratory
harrisja@ornl.gov

Bronson Messer
Oak Ridge National Laboratory
& University of Tennessee
bronson@ornl.gov

MS27

Scalable Space-Time Adaptivity for Simulations of Binary Black Hole Intermediate-Mass-Ratio-Inspirals

We present a highly scalable framework that targets problems of interest to the numerical relativity and broader astrophysics communities. This framework combines a parallel octree-refined adaptive mesh with a wavelet adaptive multiresolution and a physics module to solve the Einstein equations of general relativity. The goal of this work is to perform advanced, massively parallel numerical simulations of intermediate-mass-ratio inspirals of binary black holes with mass ratios on the order of 100:1. These studies will be used to generate waveforms as used in the data analysis of the Laser Interferometer Gravitational-Wave Observatory and to calibrate semi-analytical approximate methods. Our framework consists of a distributed memory octree-based adaptive meshing framework in conjunction with a sophisticated code generator from symbolic expressions. In addition, high-levels of adaptivity are also required to ensure scalability when the mass ratios become large. The code generator makes our code portable across different architectures, including SIMD vectorization, OpenMP, and CUDA combined with efficient distributed memory adaptive data-structures. The equations corresponding to the target application are written in symbolic notation, and generators for different architectures can be added independently of the application.

Hari Sundar
School of Computing
University of Utah
hari@cs.utah.edu

MS28

Recent Developments in Hybridisation Techniques for Finite Element Problems in Numerical Weather Prediction

For problems in Numerical Weather Prediction (NWP), time to solution is a critical factor. Semi-implicit time-stepping methods can speed up geophysical fluid dynamics simulations by taking larger time-steps than explicit methods. This is possible because they treat the fast waves implicitly, and the time-step size is not restricted by the CFL condition for these waves. This method requires an expensive linear solve that must be performed at every time-step, however, an effective preconditioner can significantly reduce the computational cost of this solve, making a semi-implicit scheme faster overall. Finite element methods are often difficult to precondition due to the large number of coupled degrees of freedom. Hybridisation methods can eliminate this coupling and instead couple the equations to a smaller global system on the trace space, which is easier to precondition. This is achieved by considering variables which only lie on the facets of the mesh. The resultant trace system can be solved using multigrid instead of directly, allowing us to run high resolution simulations. Hybridisation is possible in both a conforming finite element setting and in a discontinuous Galerkin setting. We demonstrate the effectiveness of a multigrid preconditioner for a semi-implicit IMEX time-stepper. The method is implemented in the Slate language, which is part of the Firedrake project. Firedrake is a Python framework for solving finite element problems via code generation.

Jack Betteridge
Department of Mathematics
University of Bath
J.D.Betteridge@bath.ac.uk

MS28

Entropically Consistent Models of Geophysical Fluid Dynamics

The dynamics of real fluids are governed by both reversible (entropy-conserving, such advection) and irreversible (entropy-generating, such as dissipation) processes, both of which conserve energy. Underlying this is a geometric structure: a metriplectic formulation, which combines a Poisson bracket for the reversible dynamics with a metric/dissipation bracket for the irreversible dynamics. By discretizing this geometric structure directly using what are known as mimetic discretizations and therefore preserving its essential elements, it is possible to construct numerical models with many desirable properties. This talk will present one possible approach to this in the context of geophysical fluids: the nonhydrostatic fully compressible Euler equations with some typical subgrid turbulence parameterizations. Specifically, we use mimetic Galerkin differences coupled with a second-order, implicit energy-conserving metriplectic integrator. The resulting model conserves mass and energy to machine precision, while the reversible dynamics conserve entropy and the irreversible dynamics (parameterizations) generate it. Results using this model from planar versions of the commonly used DCMIP test suite will be shown. If time permits, there will be some discussion of future work on the extension of these ideas to multicomponent/multiphase fluids (including moisture), more sophisticated turbulence parameterizations and other areas of geophysical fluids/physics more generally.

Christopher Eldred
Univ. Grenoble Alpes, Inria, CNRS, LJK
chris.eldred@gmail.com

Thomas Dubos
LMD
Ecole Polytechnique
dubos@lmd.polytechnique.fr

Francois Gay-Balmaz
Ecole Normale Superieure
gaybalma@lmd.ens.fr

MS28

NUMO: A Non-Hydrostatic Unified Model of the

Ocean

Ice-sheet/ocean interaction in Greenland is one of the key outstanding challenges in climate modeling, yet present-day climate models are not able to resolve fine-scale processes in the fjords. This is due to orders of magnitude difference in spatial scales between the open ocean (1000km) and fjord (1km) as well as small-scale processes at the glacier terminus (1m), complicated bathymetry and coastline. The most recent computational studies of fjord circulation focused on simplified box-like domains (e.g., [Kimura et al., *The effect of meltwater plumes on the melting of a vertical glacier face*, 2014]). We develop the Non-hydrostatic Unified Model of the Ocean (NUMO) to address the need for hi-resolution simulations of ocean circulation in Greenland fjords which take into account complex geometry and the range of scales of motion. The model uses fully three-dimensional unstructured mesh and high-order element-based Galerkin methods to discretize the Boussinesq approximation of incompressible Navier-Stokes equations. In this talk, we will discuss the validation of the model on a range of test cases with increasing complexity and present recent developments and challenges. We will focus on the parallel performance of the model.

Michal A. Kopera
Boise State University
michalkopera@boisestate.edu

MS29

Hybrid Quantum-Classical Approaches to Exact and Approximate Optimization

Hybrid algorithms utilizing emerging quantum technologies offer new approaches to hard computational problems. A promising family of quantum heuristics for combinatorial optimization is the quantum approximate optimization algorithm, and more generally, the quantum alternating operator ansatz (QAOA). We show new results that relate the behavior of QAOA to the structure of the underlying cost function. We then discuss how our results may aid in the design of new quantum approaches to optimization, and may be extended to hybrid approaches to other scientific problems such as the electronic structure problem of quantum chemistry. Finally, we will address research challenges for hybrid algorithms with near-term quantum hardware.

Stuart Hadfield
NASA
stuart.hadfield@nasa.gov

MS29

Scientific Computing Benchmarks for Quantum Computers

The burgeoning interest in using quantum computers for modeling and simulation of physical systems raises the question of how to track progress toward the ultimate goal of scientific discovery with quantum computing. Using noisy, intermediate scale quantum computers as platforms for testing and evaluating, we prototype applications in chemistry and materials science to evaluate the accuracy and computational complexity underlying such approaches. We explore the trade-space of algorithm design and platform constraints to define metrics that offer insights into when quantum computers may be ready for performance-driven applications and what must be overcome to reach this capability. These metrics give rises to scientific computing benchmarks that may be used to evaluate both device-level characteristics and application-level behaviors for the next generation of quantum computers.

Travis Humble, Dmitry Liakh, Alexander McCaskey
Oak Ridge National Laboratory
humblets@ornl.gov, liakhdi@ornl.gov, mccaskeyaj@ornl.gov

MS30

Clustering Techniques and Hierarchical Matrix Formats for Scalable Kernel Ridge Regression

We present scalable and memory-efficient algorithm to approximate kernel methods for machining learning. Namely, we exploit the sub-block rank deficiency of the kernel matrices to build fast direct solvers to compute the regression weights during the training stage of the kernel methods. Our method results in optimal $O(r^2n)$ training time using hierarchical matrix algebra combined with the geometry information of the data, where r is the maximum off-diagonal rank. The accuracy of our method matches the state-of-the-art non-approximated kernel regression, and our parallel code can effectively work with datasets of millions of data points.

Xiaoye S. Li
Computational Research Division
Lawrence Berkeley National Laboratory
xsli@lbl.gov

Rebrova Elizaveta
University of California, Los Angeles
erebrova@umich.edu

Gustavo Chavez
Lawrence Berkeley National Laboratory
gichavez@lbl.gov

Pieter Ghysels
Lawrence Berkeley National Laboratory
Computational Research Division
pghysels@lbl.gov

Yang Liu
Lawrence Berkeley National Laboratory
liuyangzhuan@lbl.gov

MS30

Extreme-Scale Task-Based Cholesky Factorization Toward Climate and Weather Prediction Applications

Climate and weather can be predicted statistically via geospatial Maximum Likelihood Estimates (MLE), as an alternative to running large ensembles of forward models. The MLE-based iterative optimization procedure requires the solving of large-scale linear systems that performs a Cholesky factorization on a symmetric positive-definite covariance matrixa demanding dense factorization in terms of memory footprint and computation. We propose a novel solution to this problem: at the mathematical level, we reduce the computational requirement by exploiting the data sparsity structure of the matrix off-diagonal tiles by means of low-rank approximations; and, at the programming-paradigm level, we integrate PaRSEC, a dy-

namic, task-based runtime to reach unparalleled levels of efficiency for solving extreme-scale linear algebra matrix operations. The resulting solution leverages fine-grained computations to facilitate asynchronous execution while providing a flexible data distribution to mitigate load imbalance. Performance results are reported using 3D synthetic datasets up to 42M geospatial locations on 130, 000 cores, which represent a cornerstone toward fast and accurate predictions of environmental applications.

Qinglei Cao
University of Tennessee Knoxville
qcao3@vols.utk.edu

Yu Pei
University of Tennessee
ypei2@vols.utk.edu

Kadir Akbudak
Bilkent University
kadir.cs@gmail.com

Aleksandr Mikhalev
King Abdullah University of Science & Technology (KAUST)
aleksandr.mikhalev@kaust.edu.sa

George Bosilca
University of Tennessee Knoxville
bosilca@icl.utk.edu

Hatem Ltaief
King Abdullah University of Science & Technology (KAUST)
hatem.Ltaief@kaust.edu.sa

Jack J. Dongarra
University of Tennessee, Oak Ridge National Laboratory, USA
dongarra@icl.utk.edu

David E. Keyes
KAUST
david.keyes@kaust.edu.sa

MS30

Decoupling Structure Analysis in Hierarchical Matrix Approximations

Hierarchical matrix approximations have gained significant traction in the machine learning and scientific community as they exploit available low-rank structures in kernel methods to compress the kernel matrix. The resulting compressed matrix, HMatrix, is used to reduce the computational complexity of operations such as HMatrix-matrix multiplications with tuneable accuracy in an evaluation phase. Existing implementations of HMatrix evaluations do not preserve locality and often lead to unbalanced parallel execution with high synchronization. Also, current solutions require the compression phase to re-execute if the kernel method or the required accuracy change. In this work, we describe MatRox, a framework that uses novel structure analysis strategies, blocking and coarsen, with code specialization and a storage format to improve locality and create load-balanced parallel tasks for HMatrix-matrix multiplications. Modularization of the matrix compression phase enables the reuse of computations when there are changes to the input accuracy and the kernel function.

Maryam Mehri Dehnavi, Kazem Cheshmi, Bangtian Liu
University of Toronto
mmehride@gmail.com, kazem@cs.toronto.edu, bangtian@cs.toronto.edu

MS30

Evaluation of Programming Models to Address Load Imbalance on Distributed Multi-Core CPUs: A Case Study with Block Low-Rank Factorization

To minimize data movement, many parallel applications statically distribute computational tasks among the processes. However, modern simulations often encounters irregular computational tasks. As a result, load imbalance among the processes must be dealt with at the programming level. One critical application for many domains is the LU factorization of a large dense matrix stored in the Block Low-Rank format. Using the low-rank format can significantly reduce the cost of factorization in many scientific applications, including the boundary element analysis of electrostatic field. However, the partitioning of the matrix based on underlying geometry leads to different sizes of the matrix, thus load imbalance among the processes at each step of factorization. We use BLR LU factorization as a test case to study the programmability and performance of five different programming approaches: (1) flat MPI, (2) Adaptive MPI (3) MPI + OpenMP, (4) parameterized task graph, and (5) dynamic task discovery (DTD). The last two versions use a task-based system (PaRSEC) to express the algorithm. We first point out programming features needed to efficiently solve this category of problems, hinting at possible alternatives to the MPI+X programming paradigm. We then evaluate the programmability of the different approaches. Finally, we show the performance result on the Intel Haswell system and analyze the effectiveness of the implementations to address the load imbalance.

Yu Pei
University of Tennessee
ypei2@vols.utk.edu

Ichitaro Yamazaki
Sandia National Laboratories
iyamaza@sandia.gov

George Bosilca
University of Tennessee Knoxville
bosilca@icl.utk.edu

Akihiro Ida
Academic Center for Computing and Media Studies
Kyoto University
ida@media.kyoto-u.ac.jp

Jack J. Dongarra
University of Tennessee, Oak Ridge National Laboratory, USA
dongarra@icl.utk.edu

MS31

ENSIGN: A Framework for High-performance Tensor Decompositions

ENSIGN is a high-performance tensor decomposition framework developed by Reservoir Labs that is intended to transition tensor methods into a practical data sci-

ence tool for addressing modern data analytics problems at scale. The core of ENSIGN is a suite of decomposition routines built on specialized data structures that are optimized for large shared- and distributed-memory architectures. These optimizations include operation-minimal parallel computations leading to near-ideal scaling of computations, reduced communications and synchronizations, and in-memory computation that avoids writing data to disk throughout the workflow. Through these optimizations, ENSIGN has been able to effectively exploit the ultra-low latency and high-bandwidth benefits offered by the HPE Superdome Flex server, a high-end modular system offering large-scale in-memory computing, and has also demonstrated improved performance on a distributed cluster of Intel Xeon multi-core nodes. ENSIGN has been applied successfully to problems in cybersecurity and geospatial analytics, most recently assisting in the cybersecurity of SCinet 2019 where ENSIGN was able to uncover otherwise undetected patterns of distributed attacks and other malicious activity. With the addition of a Python-enabled in-memory workflow, ENSIGN can now be used in concert with other popular machine learning libraries.

Muthu M. Baskaran
Reservoir Labs
baskaran@reservoir.com

MS31

The Sparse Tensor Algebra Compiler

Tensor and Linear Algebra are powerful tools with applications in data analytics, machine learning, science, and engineering. The massive growth of data in these applications makes performance critical. For applications that use sparse tensors, where most components are zeros, programmers must choose between libraries with hand-optimized implementations of select operations and generalized software systems with poor performance. In this talk, I will present compiler abstractions and techniques that combine tensor expressions with specifications of sparse irregular tensor data structures to produce efficient parallel source code. I will discuss the three main problems of sparse tensor algebra compilation and show how their solutions can be combined to compile and optimize sparse tensor expressions. We have implemented these ideas in the TACO sparse tensor algebra compiler. It is the first compiler to generate sparse code for any basic tensor expression on many sparse tensor representations. The generated code matches or exceeds the performance of hand-optimized libraries while generalizing to any expression and many user-specified irregular data structures.

Fredrik Kjolstad
Stanford University
fredrikbk@gmail.com

Stephen Chou
Massachusetts Institute of Technology
s3chou@csail.mit.edu

Ryan Senanayake
MIT
rsen@mit.edu

Peter Ahrens
Massachusetts Institute of Technology
pahrens@mit.edu

Shoaib Kamil
Adobe Systems Inc
kamil@adobe.com

David Lugato
CEA
david.lugato@gmail.com

Saman Amarasinghe
Massachusetts Institute of Technology
saman@csail.mit.edu

MS31

Structured Abstractions in MLIR: High-Level Infrastructure for Optimizing Matrix/Tensor Computations

MLIR is a new SSA-based compiler infrastructure designed for extensibility. It allows one to easily define intermediate representations at various levels of abstraction from tensorflow ops operating on tensor values all the way down to control-flow graphs operating scalar and vector SSA values. This talk proposes a composable and structured abstraction for code generation in MLIR. This abstraction aims at building tractable end-to-end flows for high-performance optimization of matrix and tensor operations.

Nicolas Vasilache
Google Brain
ntv@google.com

Mahesh Ravishankar
Google
ravishankarm@google.com

MS31

The Design and Implementation of Large-Scale Tensor Decomposition in SPLATT

Tensor factorization is a powerful technique for analyzing multi-way data and has applications in fields such as cybersecurity, social network analysis, and health analytics. The tensors that arise in these domains are increasingly large, sparse, and high dimensional. The ubiquity of parallel processors and large-scale clusters motivates the development of scalable parallel approaches for sparse tensor computations. This talk presents an overview on the design of SPLATT, an open source toolkit for sparse tensor factorization that is used by academia, industry, and government. Topics include data structures for sparse tensors and flexible and incorporation of factorization constraints.

Shaden Smith
Intel AI
shadentsmith@gmail.com

MS33

Parallelisation of the SMARDDA-PFC Software

The talk describes work performed on the SMARDDA-PFC software under the SMARTEST project run by the Eurofusion High Level Support Team. SMARDDA-PFC is the computational kernel of the SMITER package for calculating power deposited by escaping particles onto the first wall of magnetically confined nuclear fusion devices. The magnetic field is sufficiently strong that particles may be assumed to follow fieldlines to intersection with what is normally a complex engineered geometry, thus much of

the execution cost lies in operations fundamental to ray-tracing. SMARDDA-PFC consists of 4 programs run sequentially, of which the final program powcal is the most computationally expensive and takes up between 50% and 80% of the runtime depending on the processor type used. The aim of this project was to identify bottlenecks in execution time and optimise the code to reduce run-time. It was determined that MPI would be most effective method for achieving large speedup. First the individual modules were parallelised, then to reduce delays caused by disk traffic a monolithic combined kernel was produced. The target of a 10 s turnround for processing a detailed 360-degree first wall model of the JET tokamak model (over 2 million triangles) was successfully met.

Huw Leggate
Dublin City University
huw.leggate@dcu.ie

Wayne Arter
United Kingdom Atomic Energy Authority, UK
wayne.arter@ukaea.uk

MS33

Massively Parallel DEM Simulation and Stress Chain Characterization with over Billion Particles

The efficient parallel implementation of the Discrete Element Method (DEM) is a still big challenge because it requires complex computations for the data stored at each contact points for pairwise tangential forces. Here, we present the overview of our parallel implementation of the DEM for a large parallel computer system. Our method utilizes the action-reaction law to save memory and halve the arithmetic costs. A load-balancer with the flexible domain decomposition is applied to manage the load-imbalances between MPI procs. The shape of subdomains is iteratively updated within the framework of an iterative non-linear solver. The parallel scaling test showed good scalabilities up to 2.4 billion particles. Our complex parallel implementation of the code is verified with a reproducibility test. We will also present the challenge for using modern manycore architectures such as GPGPU and PEZY-SCs. To understand the granular mechanisms in the simulation, we also consider new iterative method for large-scale stress chain analysis. These algorithms and code development enable us to perform the real scale granular simulations in various science and engineering fields.

Mikito Furuichi
Department of Mathematical Science and Advanced Technoloy
JAMSTEC
m-furuic@jamstec.go.jp

Jian Chen, Natsuki Hosono, Daisuke Nishiura
Center for Mathematical Science and Advanced Technology
Japan Agency for MarineEarth Science and Technology
jchen@jamstec.go.jp, natsuki.hosono@jamstec.go.jp, nishiura@jamstec.go.jp

MS33

Parallel-in-Time and Aymptotic-Preserving Monte Carlo Methods for Particle-Based Systems in a Diffusive Scaling

In many applications, such as plasma edge simulation of a nuclear fusion reactor, a coupled PDE/kinetic description is required, which is usually solved with a coupled finite-volume/Monte-Carlo method. The plasma in a fusion reactor, for instance, can usually be modeled using fluid equations (for mass, momentum and energy). However, the reactor also contains neutral (non-charged) particles (which are important in its operation), of which both the position and velocity distribution is important. This leads to a Boltzmann-type transport equation that needs to be discretized with a Monte Carlo method. One then obtains a coupled finite-volume/Monte-Carlo simulation, of which the results possess both a bias and a variance. In many relevant regimes, the simulation cost is dominated by the high collisionality of the neutral particles, which lead to an unacceptably small time step. However, in the limit of infinite collisionality, a macroscopic fluid equation arises, which is much easier to simulate. In this talk, we discuss how this limiting equation can be used to accelerate convergence, either by creating an asymptotic-preverving particle method base on the limiting fluid equation or by using a parallel-in-time method.

Bert Mortier
Department of Computer Science,
K. U. Leuven
bert.mortier@cs.kuleuven.be

Giovanni Samaey
Department of Computer Science, K. U. Leuven
giovanni.samaey@cs.kuleuven.be

MS33

Computing Particle Trajectories with Spectral Deferred Corrections

The Lorentz equations describe movement of charged particles in electric and magnetic fields and are widely used in plasma physics. Arguably the most popular numerical method to solve them is the Boris algorithm, a variant of the Verlet scheme. It is computationally cheap, has good conservation properties but is only second order accurate. The talk will introduce Boris-SDC, a high-order generalization of the Boris algorithm using spectral deferred corrections. Numerical examples will be shown, illustrating performance of Boris-SDC when tracking fast ions in fusion reactors as well as using it for particle pusing in particle-in-cell (PIC) codes.

Krasymyr Tretiak
School of Mechanical Engineering
University of Leeds
k.tretiak@leeds.ac.uk

Kris Smedt, Jitse Niesen
University of Leeds
mn12kms@leeds.ac.uk, jitse@maths.leeds.ac.uk

Steve Tobias
University of Leeds
UK
smt@maths.leeds.ac.uk

Daniel Ruprecht
School of Mechanical Engineering
University of Leeds

daniel.ruprecht@fu-berlin.de

MS34

MueLu's Algorithmic Performance on GPU

With the push of exascale computing, HPC architectures have evolved significantly in the last decade. But the change is not limited to hardware, designing scientific codes to take advantage of the power offered by GPUs is a significant challenge for multiple reasons. At least two come to mind: 1. the hardware is still evolving quickly and with it the run-times or language extensions needed to program this hardware, 2. the amount of parallelism provided by GPU is penalizing data movement and until recently code divergence within computational kernels. These challenges have created a need for upgraded scientific libraries that implement efficiently computational methods. Trilinos' MueLu package provides scalable distributed memory multigrid algorithms and has recently been rewritten in large parts to utilize GPUs. This talk will provide an overview of the implementation of MueLu's smoothed aggregation and geometric interpolation algorithms on GPU. It will also explain what parts of the code have been identified as performance bottlenecks. A comparison between the performance of multigrid on GPUs and on recent CPUs will be provided. Finally a path forward to eliminate the remaining roadblocks will be outlined.

Luc Berger-Vergiat
Sandia National Labs
lberge@sandia.gov

Ray S. Tuminaro
Sandia National Laboratories
Computational Mathematics and Algorithms
rstumin@sandia.gov

Jonathan J. Hu
Sandia National Laboratories
Livermore, CA 94551
jhu@sandia.gov

Chris Siefert, Christian Glusa
Sandia National Laboratories
csiefer@sandia.gov, caglusa@sandia.gov

MS34

Towards Efficient Communication on Heterogeneous Architectures

Emerging heterogeneous architectures are comprised of nodes that contain both CPUs and GPUs, with many paths for communication inbetween. As solvers are optimized for heterogeneous machines, inter-GPU communication becomes a dominant source of cost. However, communication costs are highly dependent on the relative locations of the sending and recieving GPUs, with intra-socket messages accelerated by GPU Direct RDMA, while communication across sockets yields larger costs. Furthermore, inter-node messages sent directly between GPUs with Cuda-Aware MPI require additional costs as they are injected into the network. In this talk, we present performance models for a Power9 system, displaying the variation in message costs. Furthermore, we present node-aware strategies to improve inter-GPU communication performance.

Amanda Bienz
University of Illinois at Urbana-Champaign
bienz2@illinois.edu

Luke Olson
University of Illinois Urbana-Champaign
lukeo@illinois.edu

William D. Gropp
University of Illinois at Urbana-Champaign
wgropp@illinois.edu

MS34

Current State and Future Goals for Multigrid-Preconditioned Linear Solvers on GPU-Based Supercomputers

We present our efforts over several years to evaluate the latest state-of-the-art multigrid-preconditioned linear solvers to support our MHD modeling efforts on GPGPU machines. We will give updates as to the capabilities and performance of several linear solvers in the context of a Poisson solve with a large range of material constants. We will advocate for our requirements including changing mesh topology, the ability to maintain a grid hierarchy despite changing matrix values, the need to build a multigrid hierarchy on the GPU, and the need for further established capability in preconditioning more complex equation systems such as curl-curl systems.

Daniel Ibanez
Sandia National Laboratories
daibane@sandia.gov

MS34

Geometric and Algebraic Multigrid Solvers in PETSc on Many-GPU Supercomputer Architectures

As the high-performance computing community pushes towards the exascale horizon, many supercomputing designs are relying increasingly heavily on general purpose graphics processing unit (GPGPU) accelerators to provide nearly all of their computational power. Simultaneously, scientific application teams are developing increasingly ambitious simulations, many of which—due to their sheer size—will require efficient multilevel methods that possess (near-)asymptotic computational complexity. In this talk, we will discuss ongoing work to adapt geometric and algebraic multigrid implementations in the open-source library PETSc (the Portable, Extensible Toolkit for Scientific Computation) to compute systems with many powerful GPGPUs on each node. We will present performance measurements from Summit, the IBM AC922 at the Oak Ridge Leadership Computing Facility, that contains more than 27,000 NVIDIA Volta GPGPUs and is currently one of the best available proxies for anticipated exascale system designs.

Richard T. Mills
Argonne National Laboratory
rtmills@anl.gov

Mark Adams
Lawrence Berkeley National Laboratory
mfadams@lbl.gov

Hannah M. Morgan
Argonne National Laboratory
Mathematics and Computer Science
hmorgan@anl.gov

Karl Rupp
TU Wien
me@karlrupp.net

Barry F. Smith
Argonne National Laboratory
bsmith@mcs.anl.gov

MS35

Shifted CholeskyQR3 for High Performance Tall-Skinny QR Factorization

The Cholesky QR algorithm is suitable for high performance computing of the QR factorization of a tall and skinny matrix but has a serious numerical instability. The CholeskyQR2 algorithm, that is Cholesky QR with reorthogonalization, has improved stability but still has the problem of breakdown for ill-conditioned matrices (i.e. when the condition number is larger than 10^8 when using double precision arithmetic). Recently, we proposed the shifted Cholesky QR algorithm, where we introduced a positive diagonal shift into the computed Gram matrix so as to avoid the breakdown of its numerical Cholesky factorization. Then, we use shifted Cholesky QR as a preconditioning step before CholeskyQR2, and finally we can obtain an accurate QR factorization even if the target matrix is ill-conditioned (up to about 10^{16}). We call the resulting algorithm shifted CholeskyQR3. In this talk, we will give an overview of shifted CholeskyQR3 algorithm and related theoretical results. Then, we will present detailed performance results on recent computer systems, which show the effectiveness of shifted CholeskyQR3. We will also mention extensions of shifted CholeskyQR3 such as the QR factorization in an oblique inner product space.

Takeshi Fukaya
Hokkaido University
fukaya@iic.hokudai.ac.jp

Ramaseshan Kannan
Arup, Manchester
ramaseshan.kannan@arup.com

Yuji Nakatsukasa
University of Oxford
nakatsukasa@maths.ox.ac.uk

Yusaku Yamamoto
The University of Electro-Communications, Japan
yusaku.yamamoto@uec.ac.jp

Yuka Yanagisawa
Waseda University
yuuka@aoni.waseda.jp

MS35

On Leveraging Communication-Optimal QR Factorization in Dense Symmetric Eigensolvers

Architectural trends over the past two decades have enlarged the algorithmic search space to include and prioritize algorithms that asymptotically decrease communication and synchronization along their critical path, even at the expense of extra computation. Recent theoretical developments in both numerically stable QR factorizations and dense symmetric eigensolvers focus on asymptotically decreasing the communication and synchronization, yet few have been shown to be practical and thus have not been incorporated into popular distributed-memory linear algebra libraries such as ScaLAPACK and Elemental. We investigate the performance and scalability of successive band reduction in dense symmetric eigensolvers that leverage communication-optimal QR factorizations , as determined by the feasible memory footprint. Specifically, we study the practical potential for bulge chasing more than one intermediate band (recently proven to incur the minimal amount of communication) using a new practical communication-optimal QR factorization algorithm.

Edward Hutter
University of Illinois at Urbana-Champaign
Department of Computer Science
hutter2@illinois.edu

Edgar Solomonik
University of Illinois at Urbana-Champaign
solomon2@illinois.edu

MS35

Parallel Butterfly-Based ShermanMorrisonWoodbury Inversion

Butterfly decompositions, originally inspired by fast Fourier transforms (FFT), are promising numerical linear algebra tools for constructing fast direct solvers and preconditioners for highly oscillatory integral and differential equations. Previously, butterfly extensions of the Hierarchical matrix and HODLR (hierarchically off-diagonal low-rank) matrix formats have been developed to accelerate compression and factorization of dense linear systems arising from integral equations. These solvers represent off-diagonal blocks as butterfly decompositions and leverage complicated hierarchical butterfly arithmetic to achieve $O(N\log^2 N)$ compression and $O(N^{1.5}\log N)$ factorization complexities. Here we further simplify the hierarchical representation via essentially representing the entire matrix as one butterfly decomposition plus manageable number of smaller blocks. This representation can be treated as butterfly generalization of HSS (hierarchically semi-separable) matrix format and reduce the compression complexity to $O(N\log N)$ compared to existing constructs. The essential operation for its factorization is the butterfly extension of the ShermanMorrisonWoodbury inversion formula, which can be computed recursively leveraging the randomized butterfly arithmetic. In addition, a distributed-memory parallelization of the proposed format has been developed and integrated into the software package ButterflyPACK.

Yang Liu, Ghysels Pieter
Lawrence Berkeley National Laboratory
liuyangzhuan@lbl.gov, pghysels@lbl.gov

Xiaoye S. Li
Computational Research Division
Lawrence Berkeley National Laboratory
xsli@lbl.gov

MS35

A Matlab Package for Superfast Divide-and-Conquer Hermitian Eigenvalue Decompositions

We discuss a Matlab package for the quick eigenvalue decompositions of some useful Hermitian matrices (tridiagonal, banded, diagonal plus low rank, Toeplitz, hierarchically semiseparable, etc.). The package is based on a superfast divide-and-conquer algorithm that uses structured matrices and the fast multiple method to accelerate all the

intermediate operations. The algorithm has nearly O(n) complexity for Hermitian matrices with small off-diagonal (numerical) ranks. The package tackles a series of challenges in the accuracy and efficiency of the structured accelerations of the divide-and-conquer algorithm. For dense rank structured matrices, it does not an extra tridiagonal reduction. The Matlab routine is significantly faster than the Matlab eig function when the matrix size increases.

Jianlin Xia, Xiaofeng Ou
Purdue University
xiaj@purdue.edu, ou17@purdue.edu

MS36

Application Developers Experiences with RAJA

Modern high-performance computing (HPC) hardware provides a wide range of heterogeneous execution resources, each of which requires a vendor-specific API or programming model. RAJA provides a rich and unified interface for writing both simple and complex numerical kernels that target these execution resources in a portable manner. Production scientific applications are typically characterized by their long lives and large code bases. Such applications often need to run efficiently on multiple architectures at any time, and are under continual development. Plus, they must be viable across multiple generations of HPC platforms. It is often the case that writing and maintaining multiple versions for different architectures is untenable. RAJA eases this transition by allowing iterative adoption with a low barrier to entry, requiring manageable disruption to existing application source code, and enabling systematic performance tuning of different computational kernels. In this talk, we explain the motivation behind RAJA, share simple examples and discuss approaches and benefits of RAJA adoption in production applications at Lawrence Livermore National Laboratory and ECP applications, showing how RAJA can be used to enable high-impact scientific calculations on large-scale HPC systems.

David Beckingsale
Lawrence Livermore National Laboratory
david@llnl.gov

MS36

Innovative Methods for Scientific Computing in the Exascale Era by Integrations of (Simulation+Data+ Learning)

Towards the end of Moore's law, we need to develop not only new hardware, but also new algorithms and applications. In this study, we propose an innovative method for computational science for sustainable promotion of scientific discovery by supercomputers in the Exascale Era by combining (Simulation + Data + Learning (S+D+L)), where ideas of data science and machine learning are introduced to computational science. The BDEC system (Big Data Extreme Computing), which is scheduled to be introduced to the Tokyo University in 2021, is a Hierarchical, Hybrid, Heterogeneous (h3) system, which consists of computing nodes for computational science and those for data science/machine learning. In this study, we consider the BDEC as the platform for integration of (S+D+L), develop an innovative software platform h3-Open-BDEC for integration of (S+D+L), and evaluate the effects of integration of (S+D+L) on the BDEC. The h3-Open-BDEC is designed for extracting the maximum performance of the supercomputers with minimum energy consumption focusing on (1) innovative method for numerical analysis with high-performance/high-reliability/power-saving based on the new principle of computing by adaptive precision, accuracy verification and automatic tuning, and (2) Hierarchical Data Driven Approach (hDDA) based on machine learning. Integration of (S+D+L) by h3-Open-BDEC enables significant reduction of computations and power consumption, compared to those by conventional simulations.

Kengo Nakajima
Information Technology Center, The University of Tokyo
RIKEN Center for Computational Science (R-CCS)
nakajima@cc.u-tokyo.ac.jp

MS36

Composition in Scientific and Engineering Applications: Lessons Learned from the Legion Programming System

Complexity of scientific and engineering software for HPC is often attributed to the requirements of performance and scalability on modern supercomputers. Another axis of complexity is a result of the complex compositions of models, libraries, frameworks, and utilities that make up some Scientific and Engineering software projects. This talk will focus on lessons learned from the Legion Programming System in building software compositions and how a strong data model coupled with apparently sequential semantics can be useful in managing complexity.

Galen Shipman
Los Alamos National Laboratory
gshipman@lanl.gov

MS36

Toward AI-based Medical Data Analytics and Clinical Workflows

We lay out our plan to build a platform called Artificial Intelligence for Medical Image Analysis (AIMIA). The AIMIA platform consists of Artificial Intelligent Engines (AI Engines) and Augmented Intelligence Workflows (AI Workflows). The AI Engines include high-performance algorithms and software modules aiming to extract insightful information from a large volume of medical image datasets accurately, efficiently, and robustly. In particular, the AI Engines include Image Processing, Quantitative Analytics, Deep Learning, Machine Learning, and High Dimensional Data Analysis Toolboxes to analyze medical images. By taking these algorithms and software modules as the building blocks, we further build up innovative AI Workflows in various clinical applications. AI Workflows examples include precision cancer treatments in a lung, hypopharyngeal, hepatocellular carcinoma, digital pathology whole slide image analysis for prostate cancers, pancreatic masses classification and detection, radiotherapy treatment planning in lung cancer, and psychiatric disorders phenotyping. These examples illustrate how we apply the AI Engines to configure AI Workflows in clinical medical cares and biomedical research. AIMIA is also a platform allowing interdisciplinary experts from academia and industry in medical, mathematical, statistical, computational, and information sciences to work together to ensure the research and development efforts can benefit the society broadly.

Weichung Wang
National Taiwan University
Institute of Applied Mathematical Sciences

wwang@ntu.edu.tw

MS37

On the Theory of Discrete, Adaptive Space Filling Curves

We present a newly developed, self-contained theory for discrete space-filling curves (SFCs). Mesh partitioning according to such SFCs has been established as a fast and reliable technique, in particular when combined with frequent adaptive mesh refinement (AMR) and coarsening. SFCs map the elements of a uniform or adaptive mesh onto a finite index set, thus providing a linear order of the elements. AMR operations change this order only locally. In addition to practical use in HPC, investigating the properties of SFCs and developing new constructions are subjects of many theoretical studies. The definition for discrete SFCs is usually stated as an iteration step in a sequence that converges to an analytical SFC, or provided in the language of L-systems. Both of these definitions, however, are not ideally suited to represent the complexity of arbitrarily refined adaptive meshes. To address this issue, we provide a set of self-contained concepts and definitions, introducing new underlying structures such as refinement spaces and refinement rules. Discrete SFCs map these refinement spaces to an index set and satisfy certain locality properties. Our construction is independent of any particular geometric embedding. To demonstrate the usefulness of this approach, we present as first application the cross-product binary operation between SFCs that allows us to construct a new SFC for prism elements.

Johannes Holke
German Aerospace Center (DLR)
johannes.holke@dlr.de

Carsten Burstedde, David Knapp
Universität Bonn
burstedde@ins.uni-bonn.de, s6daknap@uni-bonn.de

MS37

An Evaluation of Asynchronous Task Execution Strategies in AMT AMR Approaches

The uintah software has used AMR since 2005 and adopted an asynchronous task based approach since 2011. This approach has made it possible to solve AMR problems for PDE problems with steep gradients and more complex fluid-structure problems and to use AMR for the scalable solution of thermal radiation problems which have global connectivity across billions of mesh cells. The relative maturity of this technology allows the advantages and challenges of an asynchronous approach to be considered. It will be shown s that a combination of asynchronicity with over decomposition helps with scalability and decreases computation time. Finally, we consider how to extend this approach to many different architectures using an intermediate layer to drive performance portability libraries for present and future leading edge architectures.

Martin Berzins, Alan Humphrey
Scientific Computing and Imaging Institute
University of Utah
mb@sci.utah.edu, ahumphrey@sci.utah.edu

MS37

Tasks Unlimited : Lightweight Task Offloading and Replication for Parallel Adaptive Mesh Refinement

As we move towards exascale, new challenges arise: performance becomes less predictable due to variability at runtime and hard faults are more and more likely to occur. Besides, numerical algorithms such as the ADER-DG method pose challenges with respect to load balancing. In this work — in an effort to prepare AMR frameworks to future exascale systems — we propose to unleash tasks into distributed memory: i.e., tasks are no longer bound to an MPI process but they can dynamically migrate between processes. We discuss how this relaxed distributed tasking can be exploited for load balancing and fault tolerance.

Philipp Samfass
Technical University of Munich
samfass@in.tum.de

MS37

Event-Driven, Stream-Based Amr Software

Dynamically adaptive mesh refinement (AMR) for PDEs on spacetrees can be realised via stacks only, if we fuse the mesh traversal and data storage with space-filling curves (SFCs). While this yields quasi-optimal cache access characteristics due to the curve's Hölder continuity, we end up with a scheme that yields many memory moves and some inherent sequentiality even though SFC cuts empirically give us good domain decompositions. In this talk, we review stack-based multiscale AMR and derive a novel traversal paradigm: It analyses the grid structure on-the-fly and identifies sections along the SFC which remain invariant under the refinement criteria. A data movement graph along these sections then is assembled, on which we eliminate unnecessary data movements and find parallel data access patterns before we invoke the actual PDE (stencil) operations. The underlying algorithmic mindset resembles automata for LR(k) grammars. We end up with a scheme which is modest w.r.t. memory movements, identifies stencil evaluation concurrency on-the-fly, and does not impose any block/grid regularity constraints on the mesh.

Tobias Weinzierl
Durham University
tobias.weinzierl@durham.ac.uk

MS38

A Discontinuous Galerkin Method for Idealised Hurricane Storm Surge

In recent years, Discontinuous Galerkin (DG) discretisations of non-linear 2D shallow water equations have evolved to be a viable tool for many geophysical applications, including hurricane storm surge simulations. As these flood simulations are time critical and computational resources are limited, there is a growing demand for the increase of computational efficiency and reduced run time. For that reason we have developed a DG storm surge model on a triangular and dynamically adaptive mesh. The dynamic mesh refinement and coarsening are driven by physics-based refinement indicators capturing major model sensitivities. In general, non-uniform, dynamically adaptive meshes are a useful tool for reducing computational complexities for simulations that exhibit strongly localised features as is the case for hurricane storm surges. A thorough comparison with simulation results obtained on a uniform mesh can be difficult as commonly used metrics and measures such as the classical definition of convergence do not

apply without modification. To gain more insight into these adaptive meshes and to build a theoretical framework for their assessment, we present and discuss a number of mesh metrics that we apply to our triangular adaptive mesh and that demonstrate the benefits. We acknowledge funding by the Irish Research Council under the research project GOIPD/2018/248 and Proyecto Mayor UTA 8718-16, Universidad de Tarapaca.

Nicole Beisiegel
University College Dublin
School of Mathematical Science & Statistics
nicole.beisiegel@ucd.ie

Jörn Behrens
Dept. of Mathematics, University of Hamburg
joern.behrens@uni-hamburg.de

Cristóbal E. Castro
Escuela Universitaria de Ingeneria Mecanica
Universidad de Tarapaca, Arica
ccastro@academicos.uta.cl

MS38

CLIMA-atmos: a Non-Hydrostatic Model of the Atmosphere for Next-Generation Super-Computing Systems

The Climate Modeling Alliance (CliMA) is an ambitious project with the goal to deliver actionable predictions for the future state of the Earth's climate. The project is a coalition of scientists, engineers, and applied mathematicians from Caltech, MIT, the Naval Postgraduate School, and NASAs Jet Propulsion Laboratory. Leveraging recently advances in computational and data science, the model known as "CLIMA" is an open-source Earth-system framework built in the Julia programming language. In this talk, we present the atmospheric dynamical core of CLIMA: CLIMA-atmos. The model utilizes a non-hydrostatic formulation of the governing equations and state-of-the-art high-order discontinuous Galerkin discretizations on tensor-product grids. We provide an overview of the dynamical core and highlight its role in the CLIMA Earth system model.

Thomas H. Gibson
Department of Mathematics
Imperial College London
gibsonthomas1120@hotmail.com

Maciej Waruszewski
Naval Postgraduate School
maciej.waruszewski@fuw.edu.pl

Simon Bryne
California Institute of Technology
simonbyrne@caltech.edu

Jeremy E. Kozdon
Department of Applied Mathematics
Naval Postgraduate School
jekozdon@nps.edu

Francis X. Giraldo
Naval Postgraduate School
fxgirald@nps.edu

Lucas Wilcox
Department of Applied Mathematics
Naval Postgraduate School
lwilcox@nps.edu

Valentin Churavy
Massachusetts Institute of Technology
vchuravy@mit.edu

MS38

A System for Programming Symbolic Manipulations Finite Element Problems in UFL: FML

The Firedrake system provides a highly automated mechanism for solving a variational problem once the user has decided on spatial and temporal discretisations. But what happens if the user wants to change those choices, for example to switch timesteppers or to change between continuous and discontinuous formulations? Parametrising a variational problem over function space is straightforward UFL, the symbolic language used by Firedrake, and UFLs symbolic manipulation features enable manipulations such as replacing the test function to implement SUPG. What is missing are the mechanisms required to programmatically reason about which transformations should be applied and to which terms in the forms.Form Manipulation Labelling (FML) is a new Firedrake feature, currently being piloted in the Gusto atmospheric core toolkit, which addresses this lacuna. FML enables terms in forms to be associated with key-value pairs which it calls labels. Reasoning about symbolic manipulations is performed by applying successive filter passes in which labels are added to, or removed from, terms on the basis of form features or the presence and value of other labels. Ultimately, symbolic transformations are also applied as form filters to, for example, drop all terms with a particular label, or perform a replacement on terms with another label.This presentation will describe the design and features of FML, and present some examples, drawn from Gusto, of its capabilities.

David Ham
Department of Mathematics
Imperial College London
david.ham@imperial.ac.uk

Jemma Shipton
University of Exeter
j.shipton@exeter.ac.uk

MS38

A Higher-Order Model for Glacier Flow via Spectral Semi-Discretization

Mathematical models of glacier flow velocity have historically been divided into 2D models that are easy to solve but only applicable over particular stress regimes, and 3D models that are applicable everywhere but expensive computationally. In this talk I will describe a model of glacier flow that can capture all stress regimes but that is much more tractable computationally. Three tricks are involved in defining the model: terrain-following coordinates; vertical semi-discretization; and convex variational principles. I will describe how this model was implemented using the package Firedrake and, finally, I will show some results for real glaciers.

Daniel Shapero
University of Washington

shapero@uw.edu

MS39

Fast quantum subroutines for the simplex method

We propose quantum subroutines for the simplex method that avoid classical computation of the basis inverse. For a well-conditioned $m \times n$ constraint matrix with at most d_c nonzero elements per column, at most d nonzero elements per column or row of the basis, and optimality tolerance ϵ, we show that pricing can be performed in time $\tilde{O}(\frac{1}{\epsilon}\sqrt{n}(d_c n + d^2 m))$, where the $\tilde{O}$ notation hides polylogarithmic factors. If the ratio n/m is larger than a certain threshold, the running time of the quantum subroutine can be reduced to $\tilde{O}(\frac{1}{\epsilon} d\sqrt{d_c} n\sqrt{m})$. Classically, pricing would require $O(d_c^{0.7} m^{1.9} + m^{2+o(1)} + d_c n)$ in the worst case. We also show that the ratio test can be performed in time $\tilde{O}(\frac{t}{\delta} d^2 m^{1.5})$, where t, δ determine a feasibility tolerance; classically, this requires $O(m^2)$ in the worst case. For well-conditioned sparse problems the quantum subroutines scale better in m and n, and may therefore have a worst-case asymptotic advantage. This asymptotic speedup does not depend on the data being available in some "quantum form': the input of our quantum subroutines is the natural classical description of the problem, and the output is the index of the variables that should leave or enter the basis.

Giacomo Nannicini
IBM T.J. Watson
nannicini@us.ibm.com

MS39

Quantum Approximate Optimization with a Trapped-Ion Quantum Simulator

Quantum computers and simulators may offer significant advantages over their classical counterparts, providing insights into quantum many-body systems and possibly solving exponentially hard problems, such as optimization and satisfiability. We report the first implementation of a shallow-depth Quantum Approximate Optimization Algorithm (QAOA) using an analog quantum simulator to estimate the ground state energy of the transverse field Ising model with tunable long-range interactions. First, we exhaustively search the variational control parameters to approximate the ground state energy with up to 40 trapped-ion qubits. We then interface the quantum simulator with a classical algorithm to more efficiently find the optimal set of parameters that minimizes the resulting energy of the system. We finally sample from the full probability distribution of the QAOA output with single-shot and efficient measurements of every qubit.

Guido Pagano
University of Maryland
pagano@umd.edu

MS39

Multilevel Hybrid Quantum-Classical Algorithms on Graphs

Quantum optimization methods have the potential to deliver computational speed-ups over classical state-of-the-art methods. However, in the near term the size of the problems that can be directly tackled using these quantum solvers is constrained by the size of NISQ hardware, which is expected to remain limited. To address these limitations, we propose a family of methods that utilize quantum optimization solvers within a problem decomposition scheme. We present Quantum Local Search (QLS) that integrates quantum optimization solvers into an iterative improvement scheme. We extend this approach to multilevel paradigm (ML-QLS) and demonstrate its potential on Graph Partitioning and Network Community Detection problems.

Ruslan Shaydulin
Clemson University
rshaydu@g.clemson.edu

MS39

What Do QAOA Energies Reveal About Graphs?

Quantum Approximate Optimization Algorithm (QAOA) is a hybrid classical-quantum algorithm to approximately solve NP optimization problems such as MAX-CUT. We describe a new application area of QAOA circuits: graph structure discovery. We omit the time-consuming parameter-optimization phase and utilize the dependence of QAOA energy on the graph structure for randomly or judiciously chosen parameters to learn about graphs. We show that the layer-one QAOA energy for the MAX-CUT problem for three regular graphs carries exactly the information: (# of vertices, # of triangles). We have calculated our explicit formulas by developing the notion of the U-polynomial of a graph G. Many of our discoveries can be interpreted as computing U(G) under various restrictions. The most basic question when comparing the structure of two graphs is if they are isomorphic or not. We find that the QAOA energies separate all non-isomorphic three-regular graphs up to size 18, all strongly regular graphs up to size 26 and the Praust and the smallest Miyazaki examples. We observe that the QAOA energy values can be also used as a proxy to how much graphs differ. Unfortunately, we have also found a sequence of non-isomorphic pairs of graphs, for which the energy gap seems to shrink at an exponential rate as the size grows.

Mario Szegedy
Alibaba, Inc. China
mario.szegedy@alibaba-inc.com

MS40

Solving Acoustic Boundary Integral Equations using High Performance Tile Low-Rank LU Factorization

We design and develop a new numerical solver for non-symmetric matrices based on a fast direct LU factorization on parallel systems. The LU factorization is the first and most time-consuming step toward solving systems of linear equations in the context of analyzing acoustic scattering from large 3D objects. The matrix equation is obtained by discretizing the boundary integral of the exterior Helmholtz problem using a higher-order Nystrom scheme. The main idea is to exploit the inherent data sparsity of the matrix operator by performing local approximation while capturing the most significant information. In particular, the proposed LU-based solver leverages the Tile Low-Rank (TLR) data compression format as implemented in the Hierarchical Computations on Manycore Architectures (HiCMA) library to decrease the cubic complexity of "classical' dense direct solvers. This requires to taskify the underlying boundary integral kernels to expose fine-grained computations. We then employ the dynamic runtime system StarPU to orchestrate the computational tasks onto the underlying hardware resources. The resulting asyn-

chronous execution permit to weaken the artifactual synchronization points, while mitigating the overhead of data motion, especially when tackling large-scale problems. The new TLR HiCMA LU factorization outperforms the state-of-the-art dense factorizations on various parallel systems.

Rabab Alomairy
KAUST
rabab.omairy@kaust.edu.sa

Noha Harthi
KAUST, Saudia Arabia
noha.harthi@kaust.edu.sa

Kadir Akbudak
Bilkent University
kadir.cs@gmail.com

Rui Chen
King Abdullah University of Science and Technology
rui.chen@kaust.edu.sa

Hatem Itaief
KAUST, Saudia Arabia
hatem.itaief@kaust.edu.sa

Hakan Bagci
King Abdullah University of Science and Technology
hakan.bagci@kaust.edu.sa

David E. Keyes
KAUST
david.keyes@kaust.edu.sa

MS40

Semi-Automatic DAG Generation for H-Arithmetic

A new method for generating task graphs for H-matrix arithmetic is presented which is based on the standard recursive functions of sequential H-arithmetic and the block index set information of the involved matrix blocks. With the new DAG generation algorithm, the process of implementing H-arithmetic on many-core CPUs is much simplified compared to previous approaches while preserving the high parallel efficiency. Various options for memory or runtime optimization are presented together with numerical examples.

Ronald Kriemann
Scientific Computing, Max Planck Institute for Mathematics i
rok@mis.mpg.de

Steffen Börm, Sven Christophersen
University of Kiel
Germany
boerm@math.uni-kiel.de, christophersen@math.uni-kiel.de

MS40

Hierarchical Matrix Algorithms on Manycore Architectures

In this talk we describe parallel algorithms of core linear algebra operations on hierarchical matrices optimized for GPUs and multi GPU systems. We consider hierarchical matrix-vector multiplication operations, low rank updates and recompression, construction of a hierarchical matrix approximation of matrices that may be available only via matrix-vector products, as well matrix multiplication through randomized sampling. Our algorithms rely on flattening the trees of the hierarchical representations to allow level by level distribution and processing, and the effective use of batched dense linear algebra kernels QR and randomized SVD kernels that we have also developed.Results on single and multi GPU systems show the high performance of our implementations on both compute bound and memory bound operations.

George M. Turkiyyah
American University of Beirut
gt02@aub.edu.lb

Wajih Halim Boukaram, David E. Keyes
KAUST
wajihhalim.boukaram@kaust.edu.sa, david.keyes@kaust.edu.sa

MS41

Utopia: a Performance-Portable C++ Library for Non-Linear Algebra

We introduce Utopia, a C++ library designed for parallel non-linear solution strategies that is performance-portable by virtue of using Kokkos. Scientific-computing libraries require continuous development and updating to support new hardware architectures, new numerical methods, and new programming models. We wish to avoid such changes to libraries and their dependencies requiring modifications in user and application code that depend on them. State-of-the-art software achieves this by leveraging high-level programming interfaces that provide abstractions for the economical expression of complex numerical procedures. These interfaces separate the model from the computation, hiding low-level implementation details from the code of applications, such as non-linear solution algorithms. Utopia has multiple interoperable back-ends, such as PETSc and Trilinos, which allows users to choose between the algorithms offered by the two libraries without modifying their code. Users can also seamlessly and transparently run on all hardware architectures supported by the chosen back end, which is Kokkos when using Trilinos. A wide range of tensor manipulation abstractions are accessible via Utopia's front-end, which facilitates implementation of custom algorithms that can use both PETSc or Trilinos back-ends. Several solvers have been implemented on top of Utopia's front-end and are provided out of the box, including multigrid and recursive multilevel trust-region strategies.

Nur A. Fadel
CSCS - Swiss National Supercomputing Centre
nur.fadel@cscs.ch

Patrick Zulian
Università della Svizzera italiana, Switzerland
patrick.zulian@usi.ch

Alena Kopanicakova
Università della Svizzera italiana
kopana@usi.ch

Andreas Fink, Daniel Ganellari
CSCS - Swiss National Supercomputing Centre
andreas.fink@cscs.ch, daniel.ganellari@cscs.ch

Rolf Krause
Università della Svizzera italiana, Switzerland
rolf.krause@usi.ch

MS41

Hardware Architecture Independent Adaptive Mesh Refinement Solver using Trilinos

In this talk we present a hardware architecture independent implementation of an adaptive mesh refinement Poisson solver that is integrated into the electrostatic Particle-In-Cell beam dynamics code OPAL. The Poisson solver is solely based on second generation Trilinos packages to ensure the desired hardware portability. We show CPU-benchmarks and report our experience running the code on GPUs.

Matthias Frey
Paul Scherrer Institut (PSI), Switzerland
matthias.frey@psi.ch

Andreas Adelmann, Uldis Locans
Paul Scherrer Institut
andreas.adelmann@psi.ch, uldis.locans@psi.ch

MS41

Scalable Geometric Search Algorithms in ArborX

A recently developed HPC spatial indexing library ArborX implements geometric search algorithms, such as k-nearest neighbors and radius search, using a bounding volume hierarchy (BVH). ArborX is implemented using MPI+Kokkos programming model. In this talk, we focus on the distributed component of the library, and discuss the existing challenges and algorithms. We demonstrate the scalability of the library for scientific computing applications.

Andrey Prokopenko, Damien Lebrun-Grandie, Bruno Turcksin, Daniel Arndt
Oak Ridge National Laboratory
prokopenkoav@ornl.gov, lebrungrandt@ornl.gov, turcksinbr@ornl.gov, arndtd@ornl.gov

MS41

State of the Tpetra Linear Solver Stack

The Trilinos/Tpetra linear solver stack, including MueLu, Ifpack2, Belos and Amesos2 offers the potential to efficiently solve large, sparse linear systems on both conventional (e.g. CPU systems) and advanced (e.g. many-core or GPU systems) architectures. Built using components from Kokkos and KokkosKernels, offers portability across a wide range of architectures. This talk will give an overview of the capabilities available in the Tpetra linear solver stack and present performance results on both CPU and GPU architectures.

Christopher Siefert, Karen D. Devine, Mark Hoemmen
Sandia National Laboratories
csiefer@sandia.gov, kddevin@sandia.gov, mhoemme@sandia.gov

Jonathan J. Hu
Sandia National Laboratories
Livermore, CA 94551
jhu@sandia.gov

Brian Kelley
Sandia National Laboratories
bmkelle@sandia.gov

MS42

Training and Best Practices to Develop Portable Yet Performant Code

As architectures evolve, it is only becoming increasingly important to develop performance portable software. Such diverse architectures require their own code optimization strategies, while the application developers prefer a write-once code development strategy in which a single code will execute efficiently on all targeted architectures. In addition to considering the underlying hardware, a software solution—or let's call it a programming mode—is also expected to address requirements of applications and their algorithms. The programming language that implements the model should provide the right abstractions to improve the productivity of scientific developers. Programmers often resort to a trade-off between achieving portability and high performance. Why? The issue is two-fold. Adequate application parallelism will not be exposed to the hardware architecture if the algorithm is structured in a way that limits the level of concurrency that a programming model can benefit from. Secondly, such a single code representation is possible only if the programming abstractions are carefully crafted for the programming models to provide informative hints to the compilers to generate optimized code across platforms. This talk will discuss this art of developing a single source code base that demonstrates the best performance and at the same time does not lose portability.

Sunita Chandrasekaran
University of Delaware
schandra@udel.edu

MS42

The Sustainability Lessons of the SLATE Project

The Software for Linear Algebra Targeting Exascale (SLATE) project is a US Department of Energy (DOE)-funded effort to replace the legacy libraries LAPACK and ScaLAPACK with a modern software package that targets both GPU-accelerated and homogeneous multi-core systems—and will be capable of reaching exascale performance. The development of SLATE faces numerous risks due to the uncertainty of the target hardware and the rapid evolution of the software stack. This presentation highlights the software engineering practices meant to minimize the impact of changes in technology and to maximize SLATE's sustainability in the long run.

Jakub Kurzak
University of Tennessee, Knoxville
kurzak@icl.utk.edu

Mark Gates
Computer Science Department
University of Tennessee
mgates3@icl.utk.edu

Ali M. Charara
University of Tennessee, Knoxville
University of Tennessee, Knoxville
ali.charara@kaust.edu.sa

Asim YarKhan
Innovative Computing Laboratory

University of Tennessee
yarkhan@icl.utk.edu

Jamie M. Finney
University of Tennessee, Knoxville
jmfinney@icl.utk.edu

Jack J. Dongarra
University of Tennessee, Oak Ridge National Laboratory, USA
dongarra@icl.utk.edu

MS42

FEniCSX: A Sustainable Future for the FEniCS Project

The FEniCS Project was founded in 2003, and is open source software for the automatic solution of partial differential equations using the finite element method. Only a few individuals involved in the early days are still active in 2019, and the scientific software landscape has changed immeasurably in that time. Nonetheless, the FEniCS Project is still actively maintained, widely used, and is currently undergoing a complete overhaul in a project we call FEniCSX. In this talk I will highlight four of the most impactful changes with respect to the sustainability of the FEniCS Project:

- *Formalisation of a governance structure within NumFOCUS.* By joining NumFOCUS, the FEniCS Project has gained access to a huge variety of legal, administrative and fundraising opportunities.
- *Sustainable pathways for bringing in new contributors.* Examples include access to the Google Summer of Code programme, which has lead to some of the most significant new features (e.g. complex number support).
- *The FEniCSX technical redevelopment.* FEniCSX is a complete redevelopment of the FEniCS Project's software components, with a strong focus on simplicity, extensibility and standards compliance.
- *Reducing time burden on core developers with software as a service.* We used to administer many services e.g. continuous integration, ourselves. Wherever possible, we are now using external third party services.

Michal Habera
University of Luxembourg
michal.habera@uni.lu

Jack S. Hale
University of Luxemburg
jack.hale@uni.lu

Chris Richardson
University of Cambridge
chris@bpi.cam.ac.uk

Johannes Ring, Marie E. Rognes
Simula Research Laboratory
johannr@simula.no, meg@simula.no

Nathan Sime
Department of Terrestrial Magnetism
Carnegie Institution for Science
nsime@carnegiescience.edu

Garth Wells
Department of Engineering
University of Cambridge
gnw20@cam.ac.uk

MS42

Views on Software Sustainability from a Computing Facility Perspective

As large-scale computing resources become more complex in many aspects—including the node architecture, the memory hierarchy, the interconnect, and the tiered storage systems—software portability between architectures at different facilities and between generations of machines within a facility can be a challenge for application developers. Since the transition from homogeneous to heterogeneous node architectures, the Oak Ridge Leadership Computing Facility (OLCF) has been collaborating with application developers to ensure that their developers employ best practices for portability and maximize the application developer productivity to ensure the sustainability of software applications. During this talk, I will discuss the productivity lessons learned and best practices developed by OLCF computational scientists and application developers over the past ten years. Additionally, I will also discuss the broader DOE Exascale Computing Project (ECP) productivity project, IDEAS-ECP, as the DOE applications community looks ahead to the sustainability challenges in the exascale era.

Judith Hill
Oak Ridge National Laboratory
hilljc@ornl.gov

MS43

3D FFTs on HPC Many-Core and Hybrid CPU-GPU Platforms: Applications in Materials and Chemistry Codes

First principles electronic structure calculations based on a plane wave expansion of the wavefunctions are the most commonly used approach for electronic structure calculations in materials, chemistry and nanoscience applications. In this approach the electronic wavefunctions are expanded in Fourier components and 3D FFTs are used to construct the charge density in real space. Due to the large amount of communications required in 3D parallel FFTs the scaling of these application codes on large parallel machines depends critically on having a 3D FFT that scales efficiently to large processor counts. In order to avoid latency issues as well as reduce communications as much as possible it is necessary to have a well designed 3D FFT that can exploit the specifics of the FFTs required for plane wave codes. We present research into the best methods of performing these types of parallel 3D FFTs for materials science and chemistry codes on exascale type architectures such as many core CPUs as well as CPU-GPU nodes. I will also discuss the development of 3D FFTs for plane wave codes in the context of the FFTX project.

Andrew M. Canning
Lawrence Berkeley National Laboratory
acanning@lbl.gov

MS43

The Design of FFTX

We present the design of FFTX and provide a first look at SpectralPack. These software packages are developed as

part of the DOE ExaScale effort by LBL, Carnegie Mellon University, and SpiralGen, Inc. We aim at translating the LAPACK/BLAS approach from the numerical linear algebra world to the spectral algorithm domain. FFTX is extending and updating FFTW for the exascale era and beyond while providing backwards compatibility. SpectralPack captures higher level spectral algorithms and their variants, including convolutions, Poisson solvers, correlations, and numerical differentiation approaches that translate to FFT calls. The SPIRAL code generation and autotuning system–now available as open source under a BSD/Apache license–underpins the effort to provide performance portability.

Franz Franchetti
Department of Electrical and Computer Engineering
Carnegie Mellon University
franzf@ece.cmu.edu

MS43

Designing An Adaptable Framework for Highly Scalable Multidimensional Spectral Transforms

Since multidimensional Fast Fourier Transforms (FFTs) is a ubiquitous algorithm in many areas of High Performance Computing, it is important to maximize its access for multiple use cases in a consistent way, without sacrificing scalable performance. P3DFFT++ was conceived as a universal toolbox for spectral transforms on pre-Exascale and Exascale platforms. It is a unique adaptable software framework encompassing a wide range of options for using FFTs and other spectral-like transforms in three and more dimensions. It includes user-defined data layout and distribution schemes, a variety of transform types, transpose utilities, as well as derivative and convolution operations. The framework is modular and flexible for various use cases as well as architectures, including heterogeneous platforms. In addition, it includes an autotuning mechanism for maximizing performance. This talk provides the background of this work, touches on P3DFFT++ design choices and discusses performance considerations. A typical use scenario is demonstrated.

Dmitry Pekurovsky
San Diego Supercomputer Center
University of California, San Diego
dmitry@sdsc.edu

MS43

Implementation of Parallel 3-D Real FFT with 2-D Decomposition on Intel Xeon Phi Clusters

In this talk, we propose an implementation of a parallel 3-D real fast Fourier transform (FFT) with 2-D decomposition on Intel Xeon Phi clusters. The proposed implementation of the parallel 3-D real FFT is based on the conjugate symmetry property of the discrete Fourier transform (DFT) and the row-column FFT algorithm. We vectorized FFT kernels using the Intel Advanced Vector Extensions 512 (Intel AVX-512) instructions. Performance results of parallel 3-D real FFTs on Intel Xeon Phi clusters are reported. We successfully achieved a level of performance over 10 TFlops on 2048 nodes of Fujitsu PRIMERGY CX1640 M1 cluster for an 8192^3-point FFT.

Daisuke Takahashi
Center for Computational Sciences
University of Tsukuba
daisuke@cs.tsukuba.ac.jp

MS44

PartILUT - a Parallel Threshold Incomplete Factorization Preconditioner for Multicore and GPU

The explosion of hardware-parallelism inside a single node asks for a shift in the programming paradigms and disruptively-different algorithm designs that allow to exploit the compute power available in new hardware technology. We propose a parallel algorithm for computing a threshold incomplete LU (ILU) factorization. The main idea is to interleave an element-parallel fixed-point iteration that approximates an incomplete factorization for a given sparsity pattern with a procedure that adjusts the pattern to the problem characteristics. We describe and test a strategy for identifying nonzeros to be added and nonzeros to be removed from the sparsity pattern. The resulting pattern may be different and more effective than that of existing threshold ILU algorithms. Also in contrast to other parallel threshold ILU algorithms, much of the new algorithm has fine-grained parallelism. Most notably, it is the first GPU-capable algorithm for computing a threshold ILU.

Hartwig Anzt
Steinbuch Centre for Computing
Karlsruhe Institute of Technology
hartwig.anzt@kit.edu

Edmond Chow
School of Computational Science and Engineering
Georgia Institute of Technology
echow@cc.gatech.edu

Tobias Ribizel
KIT
tobias.ribizel@student.kit.edu

Goran Flegar
Universitat Jaume I
flegar@uji.es

Jack J. Dongarra
University of Tennessee, Oak Ridge National Laboratory, USA
dongarra@icl.utk.edu

MS44

Fine-Grained Parallel Incomplete Factorizations

In the fine-grained parallel method for computing incomplete factorizations, a small number of highly parallel fixed-point iterations are used. We present some recent developments that improve the generality and robustness of the method. We also discuss issues in using a parallel iterative method for solving with the sparse triangular factors.

Edmond Chow
School of Computational Science and Engineering
Georgia Institute of Technology
echow@cc.gatech.edu

MS44

Hierarchical Algorithms on Hierarchical Architectures

Some algorithms achieve optimal arithmetic complexity

but have low arithmetic intensity. Others possess high arithmetic intensity but lack optimal complexity. A special group of algorithms, Fast Multipole and its H-matrix generalizations, realizes a combination of optimal complexity and high intensity. Hierarchically low-rank linear algebra is bringing about a renaissance in linear algebra, offering data sparsity to problems formally defined as dense, and thus significantly increasing the range of problem sizes that can be accommodated in (among others) integral equations, covariance matrices in statistics, and Hessians in optimization. Implemented with task-based dynamic runtime systems, these hierarchical methods also have potential for relaxed synchrony, which is important for future energy-austere architectures, since there may be significant nonuniformity in processing rates of different cores even if task sizes can be controlled. We describe modules of KAUST's Hierarchical Computations on Manycore Architectures (HiCMA) software toolkit that illustrate these features and are intended as building blocks of more sophisticated applications, such as matrix-free higher-order methods in optimization. HiCMA's target is hierarchical algorithms on emerging architectures, which have hierarchies of their own that generally do not align with those of the algorithm. Some modules of this open source project have been adopted in the software libraries of major vendors.

David E. Keyes
KAUST
david.keyes@kaust.edu.sa

MS44

Parallel Sparse Indefinite Solvers

Many applications in science and engineering require the solution of large sparse linear systems of equations. For solving such problems, direct methods are frequently employed because of their robustness, accuracy and usability as black-box solvers. As modern architectures become more and more complex, with an increasing number of cores per chip, a deeper memory hierarchy and the integration of accelerators such as GPUs, it becomes all the more challenging to exploit the potential performance of such machines for sparse matrix factorization algorithms especially in the context of symmetric indefinite systems. Although significant efforts has gone into positive-definite systems, little progress has been reported in the much harder indefinite case. One major advance for tackling these problems is the design of the APTP (a posteriori threshold pivoting) strategy that has been implemented in the SSIDS solver and proven to be both efficient on multicore architectures compared to the state-of-the-art direct solvers. In this talk, we present the DAG-based solver SpLDLT that relies on a APTP strategy and uses the StarPU runtime system for implementing it parallel version. We show the benefits of our approach for exploiting heterogeneity in the the context of GPU-accelerated multicore systems.

Florent Lopez
University of Tennessee, Knoxville
flopez@icl.utk.edu

Iain Duff
Science & Technology Facilities Council, UK
and CERFACS, Toulouse, France
iain.duff@stfc.ac.uk

MS45

Surrogate Optimization for HPC Applications

High performance computing is crucial in many applications important to the Department of Energy for simulating complex physical phenomena, including climate sciences, high energy physics, and combustion research. These simulations usually contain parameters that determine how well the simulation represents reality. Optimizing these parameters is a computationally expensive task as it may take several minutes to hours to run the simulations with a given parameter set. Efficient optimization algorithms that do not rely on derivative information of the simulation objective function are needed. In this talk, we present an overview of surrogate model algorithms, which are commonly used to tackle these types of black-box expensive optimization problems. We discuss the importance of taking different problem characteristics into account when formulating the optimization problem and during the algorithm development and the potential impact on parallelizing these methods.

Juliane Mueller
Lawrence Berkeley National Lab
julianemueller@lbl.gov

MS45

A Scalable Framework for Simulating Particle Collisions at the Energy Frontier

Computer simulations of high-energy particle collisions exhibit an exponential scaling with the number of observed objects in the final state. Applications to experiments like ATLAS and CMS at the Large Hadron Collider (LHC) are presently constrained by the limited computing resources on the worldwide LHC computing grid. This talk discusses a novel simulation framework, designed for computations at scale on DOE supercomputing facilities. We exemplify how these simulations can be used to obtain predictions for Drell-Yan type processes, which constitute the most relevant irreducible backgrounds in searches for phenomena beyond the Standard Model of particle physics.

Stefan Hoeche
Fermi National Accelerator Laboratory
shoeche@fnal.gov

Holger Schulz
University of Cincinnati, U.S.
hschulz@fnal.gov

MS45

Revolutionizing HEP Data Storage using HPC Technologies

In this talk, we will describe the two approaches we have been investigating to effectively bring HPC technologies to the existing HEP (non-traditional) data processing. Our focus is on both performance improvement of the processing and physics-developer/analyzer time to develop these processing tasks. HEP data processing is inherently shaped by the tools and resources that have been available to the HEP developers. More access and availability of the HPC resources and technologies are providing HEP developers with opportunities to use modern and high performance infrastructures. We are helping to bridge the gap between

current HEP processing and HPC tools. HEP data processing currently revolves around processing a large number of small files to satisfy grid computing environment constraints. Each file can be processed independently allowing for file-granularity parallelism. Our first approach uses MPI parallel IO with HDF5. Use of HDF5 allows us to do parallel reads and achieve parallelism finer than the granularity of files. Our second approach involves the use of object stores, thus removing the concept of files entirely. With either approach, the end user code does not need to change with the back-end storage we choose use.

Saba Sehrish
Fermi National Accelerator Laboratory
ssehrish@fnal.gov

MS45

Software Infrastructure for Scalable Data Analysis on HPC Systems

Today's scientific applications can run on hundreds of thousands of processors and produce massive amounts of data. While modern HPC systems promise such large scales, scalable software infrastructure is required to employ these systems efficiently for distributed data analysis at scale. In this talk, we will introduce the software tools that constitute this software infrastructure, and present how we apply these tools to the science problems in high-energy physics. These tools include a library for development of scalable parallel algorithms, an in situ middleware for parallel dataflow coupling and composition of scientific workflows, and a distributed workflow system to allow those workflows to run at several independent systems in a wide area. We hope that the lessons learned will increase understanding and motivate further research into scalable data analysis on HPC systems.

Orcun Yildiz
Argonne National Laboratory
oyildiz@anl.gov

MS46

AMReX: A Block-Structured AMR Software Framework for the Exascale

AMReX is a software framework for the development of block-structured AMR algorithms on current and future architectures. AMR reduces the computational cost and memory footprint compared to a uniform mesh while preserving the essentially local descriptions of different physical processes in complex multiphysics algorithms. AMReX supports a number of different time-stepping strategies and spatial discretizations, and incorporates data containers and iterators for mesh-based fields, particle data and irregular embedded boundary (cut cell) representations of complex geometries. Current AMReX applications include accelerator design, additive manufacturing, astrophysics, combustion, cosmology, microfluidics, materials science and multiphase flow. In this talk I will focus on AMReX's strategy for balancing readability, usability, maintainability and performance across multiple applications and architectures.

Ann S. Almgren
Lawrence Berkeley National Laboratory
asalmgren@lbl.gov

John B. Bell
CCSE
Lawrence Berkeley Laboratory
jbbell@lbl.gov

Kevin N. Gott
LBNL
kngott@lbl.gov

Weiqun Zhang
Lawrence Berkeley National Laboratory
Center for Computational Sciences and Engineering
weiqunzhang@lbl.gov

Andrew Myers
Lawrence Berkeley National Lab
atmyers@lbl.gov

MS46

An Asynchronous Algorithm for 2:1 Octree Balance

Adaptive mesh refinement (AMR) is a key technology in many simulations, allowing for locally increased resolution without the cost of a globally refined regular mesh. The p4est library, which takes a forest-of-quadtrees (2D) or forest-of-octrees (3D) approach to AMR, has been demonstrated to be highly scalable for large scale applications. In particular, its implementation of the mesh adaptivity cycle—all steps between marking cells for coarsening or refinement and readiness to resume computation—has demonstrated parallel scalability beyond 100,000 MPI processes.Previous studies of the performance of p4est have foregrounded weak scalability, with a large number of octree leaves per MPI process. This work focuses on the strong scalability of the mesh adaptivity cycle. Improving the strong scalability and reducing the latency of the mesh adaptivity cycle reduces the time-to-solution of existing simulations and allows for more frequent adaptivity within runtime constraints.We have identified the communication pattern of p4est's *2:1-balance* algorithm—which enforces a size constraint between adjacent leaves in a partitioned octree by refining those that are too coarse—as a latency bottleneck. We compare the existing "two round" algorithm for this problem to a new "local-sort' algorithm and a simple global sort algorithm, comparing on the TACC Stampede2 supercomputer.

Hansol David Suh, Tobin Isaac
Georgia Tech
hsuh7@gatech.edu, tisaac@cc.gatech.edu

MS47

Applications of Parareal to Geo/astrophysical Fluid Dynamics: Convection and Magnetic Field Generation

The precise mechanisms responsible for the natural dynamos in the Earth and Sun are still not fully understood. Numerical simulations which couple the flow of a conducting fluid with magnetic effects, using the equations of magnetohydrodynamics (MHD), are used to investigate the dynamo effect. These simulations are extremely computationally intensive, and are carried out in parameter regimes many orders of magnitude away from real conditions. Parallelization in space is a common strategy to speed up simulations on high performance computers (HPCs), but eventually a scaling limit is reached due to increasing overheads from communication. Additional directions of parallelization are desirable to improve utilisation of HPCs. We have

used the spectral code Dedalus to study the performance of the parallel-in-time algorithm Parareal when applied to simulations of geo/astrophysical fluid dynamics. Specifically, we look at the kinematic dynamo problem, which concentrates on magnetic field generation, and Rayleigh-Bénard convection, in which the characteristics of the fluid flow are investigated. Results for kinematic dynamo studies show that parareal offers increased performance for both the Roberts and Galloway-Proctor dynamos, when compared to parallel in space methods alone. Galloway-Proctor simulations showed speed ups of 300 when using 1600 cores. Early Rayleigh Bénard convection results shows speed ups are possible up to at least $Ra = 10^6$

Andrew T. Clarke, Chris Davies
University of Leeds
scatc@leeds.ac.uk, c.davies@leeds.ac.uk

Steve Tobias
University of Leeds
UK
smt@maths.leeds.ac.uk

Daniel Ruprecht
School of Mechanical Engineering
University of Leeds
daniel.ruprecht@fu-berlin.de

MS47

Multigrid Reduction in Time for High-Order Discretizations of Hyperbolic Problems

Because computer clock speeds are no longer increasing, the sequential time marching approach used in science simulation codes is becoming a bottleneck. Parallel time integration is a way of creating concurrency in a simulation that can be exploited to remove this bottleneck and provide speed ups, sometimes dramatic. The multigrid reduction in time (MGRIT) approach applies existing knowledge and expertise in parallel spatial multigrid methods to the time dimension. The MGRIT method is designed to be as non-intrusive as possible and to take advantage of existing simulation codes and techniques as much as possible. This has worked well for parabolic equations, but parallel-in-time methods for advection-dominated or purely hyperbolic problems has proven to be difficult to develop. In this talk, we will present progress on the development of a non-standard coarse grid operator for MGRIT that has been demonstrated to produce scalable results for scalar advection problems discretized with either explicit or implicit high-order (up to order 5) Runge-Kutta methods. Our initial proof-of-principle approach computes coarse-grid operators by solving a least-squares problem based on existing MGRIT theory. We will also present newer more practical approaches under development and discuss various insights we have learned for solving hyperbolic problems parallel in time.

Hans De Sterck
University of Waterloo
hans.desterck@uwaterloo.ca

Robert D. Falgout
Lawrence Livermore National Laboratory
rfalgout@llnl.gov

Stephanie Friedhoff
University of Wuppertal
friedhoff@uni-wuppertal.de

Oliver A. Krzysik
Monash University
oliver.krzysik@monash.edu

Scott Maclachlan
Department of Mathematics and Statistics
Memorial University of Newfoundland
smaclachlan@mun.ca

MS47

Warmstarting PFASST Iterations

Adjoint gradient computation for optimization problems governed by parabolic PDEs requires one solve of the state equation and one backward-in-time solve of the adjoint equation, which makes the iterative optimization process extremely costly. With today's modern computers, the time-to-solution can be decreased through massive parallelization, which is traditionally done in the spatial dimensions. In addition, time-parallel methods have received increasing interest in recent years. Iterative multilevel schemes such as PFASST (Parallel Full Approximation Scheme in Space and Time) are currently state of the art and can achieve significant parallel efficiency. As PFASST is based on spectral deferred correction methods for the time integration, their iterative nature allows reusing previously computed solutions in the optimization loop to reduce SDC iterations required for solving state and adjoint equations at the cost of additional storage. We investigate the the impact of warmstarting PFASST iterations on the parallel performance, and discuss the influence of inexact storage of solutions. This is joint work with Michael Minion (Lawrence Berkeley National Lab).

Sebastian Götschel
Zuse Institute Berlin
goetschel@zib.de

MS47

Parallelizing Exponential Integrators with PFASST

A new approach for parallel-in-time methods combining exponential integrators and the PFASST method will be introduced. Exponential integrators for ODEs use the matrix exponential operator to integrate a linear part of the equation exactly while nonlinear terms are typically included using an operator splitting approach. A serial exponential integrator due to Tommaso Buvoli that treats nonlinear terms to arbitrarily high order by an extension of Spectral Deferred Corrections (SDC) will be discussed followed by extensions to multi-level SDC and PFASST. Preliminary numerical results on the parallel performance of the PFASST algorithm using exponential SDC sweepers and comparisons with more traditional parallel-in-time methods based on parareal and PFASST will then be presented.

Michael Minion
Lawrence Berkeley National Lab
mlminion@lbl.gov

Tommaso Buvoli
Department of Applied Mathematics, University of Washington
Seattle, WA 98195 USA

tbuvoli@ucmerced.edu.

MS48

Hierarchical Techniques for Solvers with Regulated Accuracy

By regulating the accuracy of the operations in a parallel solver it is possible to drastically increase the resulting performance without much reduction in the quality of the solution. One such accuracy regulation may occur is the use of mixed precision while another is a class of approximation techniques that take advantage of the structure in the originating domain. When used in the context of modern parallel hardware, these methods allow to take full advantage of the available resources and increase both the available performance peak as well as increase the optimality of the communication.

Micah Beck
University of Tennessee
Electrical Engineering and Computer Science
mbeck@utk.edu

MS48

Exploiting Generic Tiled Algorithms Toward Scalable H-Matrices Factorizations on Top of Runtime Systems

Hierarchical matrices (H-matrices) have become important in applications where accuracy can be reduced to decrease to a logarithmic order both the execution time and memory consumption. It happens for instance when solving Boundary Element Methods (BEM) problems. However the natural hierarchical structure of the H-Matrices makes it more difficult to efficiently parallelize with modern programming paradigm such as task-based implementations. We discuss in this presentation how we can combine, Chameleon, a tiled dense linear algebra software relying on sequential task-based algorithms and runtime systems such as StarPU, and Hmat-oss, a library focused on providing a set of sequential algorithms for H-algebra operations. We will discuss the limitations in terms of H-matrices structure and memory compression that are imposed by the use of a tiled algorithm, and we will show the performance that can be brought by such a generic solution with respect to more advanced implementation fully exploiting the hierarchical data structure.

Rocío Carratalá-Sáez
Universitat Jaume I
rcarrata@uji.es

Mathieu Faverge
Bordeaux INP - Labri - Inria
mathieu.faverge@inria.fr

Gregoire Pichon
INRIA
gregoire.pichon@inria.fr

Enrique S. Quintana-Ortí
Universitat Politécnica de València
quintana@disca.upv.es

Guillaume Sylvand
Airbus Group Innovations
guillaume.sylvand@airbus.com

MS48

A Task-Based H-Matrix Solver for Distributed Machines Memory with Manycore Nodes

Hierarchically compressed matrices (H-matrices) are well-known to significantly boost the performance of codes based on the Boundary Element Method (BEM), not only in terms of execution time, but also regarding memory consumption. There are however only few available implementations of direct solvers based Cholesky or LU decompositions) for hierarchically compressed systems, especially on distributed machines equipped with manycore nodes such as Intel KNL boards. This talk will therefore describe our task-based implementation of a H-matrix direct solver targeting super-computers with many-core nodes. A significant effort was devoted to the careful design of a task-based programming model suitable for such irregular hierarchical workloads in a distributed environment. The first challenge was to efficiently interleave tasks and asynchronous MPI operations, which we solved by the means ofinterruptibletasks. The second difficulty was to design a synchronization mechanism to efficiently and easily implement recursive compute kernels that access hierarchical pieces of data.By introducing hierarchical RW-dependencies into our task-based model, we made it possible to seamlessly design such kernels by letting the runtime system automatically derive task dependencies. Thanks to these techniques, we reached a parallel efficiency of 70% on our direct H-matrix solver when solving a problem with 4.4 million unknowns over 380 Intel KNLs (24320 cores) of the CEA's machine.

David Goudin, Cédric Augonnet, Matthieu Kuhn
CEA/DAM
david.goudin@cea.fr, cedric.augonnet@cea.fr, matthieu.kuhn@cea.fr

MS49

GPU Computing for Solving Kinetic Equations

Graphic processing units (GPUs) have emerged as a viable platform to solve large scale scientific problems. However, they also pose significant challenges both in terms of designing suitable algorithms and in implementing them efficiently. In this talk we will consider the solution of kinetic equations using the SLDG code. Kinetic models are extensively used in plasma physics. They are posed in an up to six-dimensional phase space, which makes their numerical solution extremely expensive. Traditional numerical methods have a number of downsides, such as the global data interchange inherent in these methods, that makes them ill-suited for GPUs. We describe how GPU support was incorporated into the SLDG code. Both in terms of constructing suitable algorithms for the massively parallel systems targeted (in our case we use a semi-Lagrangian discontinuous Galerkin scheme) and how to implement them efficiently on modern computer hardware. We also touch on some software architecture aspects. In particular, we will discuss the trade-off between code specific to a given hardware platform (to increase performance) and the generic implementation of the algorithm (to facilitate code reuse).

Lukas Einkemmer
University of Innsbruck

lukas.einkemmer@uibk.ac.at

MS49

GPU-Powered Particle-in-Cell Community Frameworks for Laser-Plasma Interaction

In the context of laser-particle acceleration, the electromagnetic particle-in-cell codes PIConGPU and WarpX are presented. Novel developments and workflows that enable high-resolution, fast turn-around computations on manycore-powered, leadership-scale supercomputers are essential to make optimal use of upcoming Exascale machines. Both codes' software libraries and abstractions are build on top of a generalized, single-source programming model (Alpaka) or parallel-for/-reduce based kernels. While PIConGPU is designed on top of modular, single-purpose libraries, WarpX's core routines are constructed on top of a more monolithic dependency, AMReX, which is a widely used adaptive-mesh refinement framework. Both particle-in-cell codes share the same challenges for handling PByte-scale data workflows on pre-Exascale machines. A common, open data format for particle and mesh data (openPMD) avoids duplicating I/O efforts and allows to reuse scalable data workflows with common libraries. In production runs, close bindings to scripting languages and the Jupyter platform can provide efficient control of simulations, with the goal of fast turn-arounds and good applicability to experiments. We present standardization efforts and prototypes of both communities with emphasis on reproducibility and flexibility.

Axel Huebl
Lawrence Berkeley National Laboratory (LBNL), USA
Helmholtz-Zentrum Dresden-Rossendorf (HZDR), Germany
axelhuebl@lbl.gov

PIConGPU Collaboration
Helmholtz-Zentrum Dresden-Rossendorf
picongpu@hzdr.de

WarpX Collaboration
Lawrence Berkeley National Laboratory
warpx@lbl.gov

MS49

On the Use of GPU-Enabled Clusters in Cardiac Electrophysiology

Recent advances in personalized arrhythmia risk prediction show that computational models can provide not only safer but also more accurate results than invasive procedures. However, detailed organ-scale simulations of calcium handling and electrical signal transmission in the human heart require the stochastic simulation of a large number of ion channels in each cell, which consumes immense processing power for the simulation of a single heartbeat, thereby creating the need for large scale parallel implementations. Fortunately, the cost of computing has fallen dramatically in the past decade. A prominent reason for this is the recent introduction of manycore processors such as GPUs, which by now power the majority of the worlds leading supercomputers. These devices owe their success to the fact that they are optimized for massively parallel workloads, such as applying similar ODE kernel computations to millions of mesh elements in scientific computing applications. In this talk, we present codes for solving such cardiac models on structured and unstructured meshes, and discuss the challenges involved in modernizing these codes to run on heterogeneous supercomputers. We focus on the interaction between OpenMP, MPI, and CUDA in such computations, as well as optimizations to communication and vector processing, and discuss the challenges involved in building the scalable simulation codes needed to close the gap to accurate real-time computations.

Johannes Langguth, Kristian Hustad, Neringa Altanaite
Simula research laboratory
langguth@simula.no, kghustad@simula.no, altanaite@gmail.com

Xing Cai
Simula Research Laboratory
1325 Lysaker, Norway
xingca@simula.no

MS49

Using GPUs in Chemical Kinetics - a Stochastic Solver Within the Framework of PySB

Chemical kinetics is the field of study that focuses on modeling the dynamics of chemical reactions. This dynamics is described by the chemical master equation (CME). The computational effort of solving the full master equation is usually avoided by approximating the system using a set of ODE that supposedly captures the underlying dynamics of the system. Recent technology enables us to study biological cells with single-cell resolution - a level of detail, where the influence of low concentration species can no longer be neglected and therefore the approximation of ODEs is no longer sufficient. In this talk we introduce the implementation of a stochastic simulation algorithm (SSA) introduced by Gillespie to solve the CME using GPUs. This solver is part of the PySB framework - a Python based environment that enables the user to encode the chemical network using rules that resemble the standard notation of chemical reactions.

Martina Prugger
Vanderbilt University
martina.prugger@vanderbilt.edu

MS50

Low-Synchronization Orthogonalization Schemes for S-Step and Pipelined Krylov Solvers in Trilinos

We investigate two single-reduce orthogonalization schemes for both *s*-step and pipelined GMRES. The first is based on classical Gram Schmidt with reorthogonalization (CGS2), and the second on inverse compact *WY* modified Gram Schmidt (MGS). Standard iterated CGS2 requires three global reductions. In standard MGS, the number of global reductions is proportional to the number of vectors against which we are orthogonalizing. In both cases, we can reduce this to a single global reduction, including reorthogonalization for accuracy. Our implementation is based on Trilinos software components, and therefore, is portable to different machine architectures with a single code base. We first demonstrate solver performance on the Intel Haswell nodes of the NERSC Cori Supercomputer. For these experiments, we integrated our solvers into `Nalu-wind`, a computational fluid dynamics application. At each time step, Nalu uses GMRES with a smoothed aggregation algebraic multigrid (SA-AMG) preconditioner to solve a pressure Poisson linear system. By combining *s*-step GMRES with low-synchronization orthogonalization algorithms, we reduced Nalu's total

GMRES solve time by a factor of 1.4×. We also benchmarked the single-reduce orthogonalization schemes on the ORNL Summit supercomputer on both NVIDIA V100 GPUs and IBM Power9 CPUs.

Ichitaro Yamazaki
Sandia National Laboratories
iyamaza@sandia.gov

Stephen Thomas
National Renewable Energy Laboratory
stephethomas@gmail.com

Mark Hoemmen
Sandia National Laboratories
mhoemme@sandia.gov

Erik G. Boman
Center for Computing Research
Sandia National Labs
egboman@sandia.gov

Katarzyna Swirydowicz
NREL
National Renewable Energy Lab
katarzyna.swirydowicz@nrel.gov

James Elliott
Sandia National Laboratories
jjellio@sandia.gov

MS50

Recent Development of Multigrid Solvers in HYPRE on Modern Heterogeneous Computing Platforms

Modern many-core processors such as the graphics processing units (GPUs) are becoming an integral part of many high performance computing systems nowadays. These processors yield enormous raw processing power in the form of massive SIMD parallelism. Accelerating multigrid methods on GPUs has drawn a lot of research attention in recent years. For instance, in recent releases of the HYPRE package, the structured multigrid solvers (SMG, PFMG) have full GPU-support for both the setup and the solve phases, whereas the algebraic multigrid (AMG) solver, namely BoomerAMG, has only its solve phase been ported and the setup can still be computed on CPUs only. In this talk, we will provide an overview of the available GPU-acceleration in HYPRE and present our current work on the algorithms in the AMG setup that are suitable for GPUs including the parallel coarsening algorithms, the interpolation methods and the triple-matrix multiplications. The recent results as well as the future work will also be included.

Ruipeng Li, Bjorn Sjogreen
Center for Applied Scientific Computing
Lawrence Livermore National Laboratory
li50@llnl.gov, sjogreen2@llnl.gov

Ulrike Meier Yang, Robert Falgout
Lawrence Livermore National Laboratory
yang11@llnl.gov, falgout2@llnl.gov

MS50

Enhancing Hypre's Semi-Structured Capabilities

Multigrid methods are well suited to large massively parallel computer architectures, because they are mathematically optimal and display excellent parallelization properties. Since current architecture trends are favoring regular compute patterns to achieve high performance, the ability to express structure has become much more important. An alternative to standard sparse matrix classes expressed with row and column indices is a semi-structured matrix class that is primarily described in terms of stencils and logically rectangular grids is. The definition of semi-structured rectangular matrices, which are needed as prolongation operators in algebraic multigrid, is however nontrivial. We will discuss our efforts on the new semi-structured matrix class and a semi-structured algebraic multigrid solver built upon it.

Ulrike Meier Yang, Robert Falgout,
Victor Paludetto Magri
Lawrence Livermore National Laboratory
yang11@llnl.gov, falgout2@llnl.gov, paludettomag1@llnl.gov

MS51

Data Movement Orchestration in Accelerator-Rich Systems

With the end of traditional technology scaling, accelerator-rich computer architectures have emerged as a primary means to achieve continued performance gains. A wide variety of specialized accelerators have been developed for a range of applications, including linear algebra, scientific computing and machine learning. While it is relatively well understood how to accelerate computations, integration of a large number of heterogeneous accelerators into an overall system architecture has become the main challenge. Specifically, if not carefully orchestrated, the costs of offloading computations, feeding accelerators and moving data between different computational elements on or off chip can quickly outweigh the benefits of acceleration. In this talk, we will discuss data movement challenges as well various software and hardware solutions to minimize offloading and data movement costs. This includes an outlook on how future large-scale system architectures that incorporate near-memory processing to exploit data and compute locality will have to be treated as distributed systems both from a programming and runtime management perspective.

Andreas Gerstlauer
University of Texas at Austin
gerstl@ece.utexas.edu

MS51

Codesign Tradeoffs for Deep Learning Hardware Accelerators

Today the key challenge in improving performance is how to leverage transistors when they cannot all be used at the same time due to power/area restrictions. Modern architectures exploit specialization and application specific accelerators to overcome such limitations and to achieve orders of magnitude better performance while staying under power/area budget. In this talk I will introduce the archi-

tectural concepts to consider while codesigning hardware accelerators for deep learning. I will also review some of the deep learning kernels from the performance and hardware design point of view.

Ardavan Pedram
Stanford University
perdavan@stanford.edu

MS51

A Wafer-Scale Chip and System for Deep Neural Networks

The compute requirement for training neural networks is growing at a steep exponential pace that far exceeds the now-slowing growth of single chip performance. We badly need more compute for AI. So, what comes after Moores Law? If we cannot endlessly pack more transistors into one square millimeter, perhaps we should pack more square millimeters into one chip. Large chips are faster because on-chip memory is so fast it can alleviate or remove the memory wall, and on-chip interprocessor communication is so fast it can alleviate or remove network bandwidth and latency as performance limiters. That idea, using the whole silicon wafer for a single processor, was tried in the 1980s, and it failed then. That was then. Now, Cerebras Systems, a venture capital funded startup company, has built a system designed around a chip 56 times larger than the largest GPU: a wafer-scale chip. The Cerebras Wafer-Scale Engine has an architecture optimized for neural network training and inference. It enables unprecedented performance at greater power efficiency, and opens the door to scale and methods not practical on today's machines. I will give an overview of the system and how it is used for training neural networks.

Rob Schreiber
Cerebras, U.S.
rob.schreiber@cerebras.net

MS51

Scientific Computing in a Changing Landscape

In the early 2000s, due to constraints on economical heat dissipation, clock speeds of single-core CPUs could no longer be increased, which marked the adoption of multi-core CPUs, together with a paradigm shift to algorithms specifically designed for parallel architectures. About 15 years into this architectural cycle and on its way to exascale performance, the computing industry finds itself at the confluence of technical difficulties that cast doubt on its ability to sustain this architectural model beyond the exascale capability. These difficulties are driving the hardware industry and computational scientists to develop application-specific chips and to look beyond silicon-based technology (e.g., quantum computing, physical annealing, neuromorphics, etc.), with a continued emphasis on raw processing power while addressing concerns about energy efficiency. Hardware specialization will likely redefine the development of computational algorithms over the next two decades for a wide range of important applications: large-scale PDE-based problems, artificial intelligence, computational chemistry, and optimization problems, to name just a few. The pressure to decrease time to solution or improve simulation fidelity, or both, for these applications will continue unabated. Sitting between hardware designers and application users, the computational science engineering community will play a pivotal role to commercialize future computing technologies.

Laurent White, Dimitar Trenev, Arash Fathi
ExxonMobil Research and Engineering
laurent.white@exxonmobil.com, dimitar.trenev@exxonmobil.com, Arash.fathi@utexas.edu

MS52

Experiences with Productivity and Software Sustainability on LCF Machines

The Argonne Leadership Computing Facility was created 13 years ago to provide access to computational and data resources that high-impact science and engineering researchers would not be able to access anywhere else. To perform this mission, the ALCF has deployed four supercomputers that were among the largest systems of their time. Readiness to use these platforms is challenging. Application teams need the agility to adapt to new bleeding-edge hardware, develop new science capability, and manage large science campaigns that generate complex datasets. Through these 13 years, there have been patterns for the approach of some of the most success research groups and their application codes. I will discuss how projects have sustained their applications and evolved their software engineering and management practices, and I will touch on practices that were not as effective. Practices are also evolving as hardware complexity is increasing. I will discuss how projects are collaborating with the ALCF to improve their software engineering practices and how the ALCF is changing to better support these needs.

Katherine Riley, Anshu Dubey
Argonne National Laboratory
riley@alcf.anl.gov, adubey@anl.gov

MS52

Software Sustainability Lessons from the Fluid Dynamics Community

Fluid dynamics is a broad field spanning a large number of science and engineering problem domains that are critical to a wide variety of important applications. To support the fluid dynamics research and applications community, we are executing a community-driven conceptualization of a future institute Fluid Dynamics Software Infrastructure (FDSI) to broadly develop, share, and apply computational tools for the generation and analysis of fluid dynamics data from both experimental and computational sources. Though we originally envisioned the institutes name to be Computational Fluid Dynamics Software Infrastructure, it became apparent in dialog with our community that having the first word in the institute be computational left experimental data (and more importantly experimental researchers) less than equally important. The primary objective is to facilitate the sharing of computational tools and data resources through a rich and extensible set of software components that can be applied with a wide range of existing fluid dynamics analysis tools. Essential to the success of this objective is to sustainably develop software components to analyze data from a wide range of sources. Two workshops have already been held and additional workshops are planned. In this talk we will summarize the results of those workshops and discuss the ongoing plans to design FDSI.

Kenneth Jansen
University of Colorado at Boulder

kenneth.jansen@colorado.edu

MS52

Productivity and Sustainability in a Community-Driven Software Ecosystem for Watershed Science

Through its Science Focus Area (SFA) projects the Subsurface and Biogeochemical Research program within the U.S. Department of Energy is tightly integrating observations, experiments, and modeling to advance a systems-level understanding of how watersheds function, and to translate that understanding into advanced science-based models of watershed systems. To broaden the impact of the existing SFAs, the IDEAS-Watersheds project builds on the success of a synergistic family of IDEAS projects initiated in 2014. Specifically, it strives to improve watershed modeling capacity by increasing software development productivity through an agile approach to creating a sustainable, reliable, parallel, software ecosystem with interoperable components. In this talk we highlight the unique structure of the IDEAS-Watersheds project and describe how it addresses many challenges of software development with interdisciplinary teams. Specifically, it is organized around six Research Activities that develop concrete use cases to drive advances in our parallel software ecosystem. These use cases are designed to balance advancing software design and development practices for parallel architectures with the model and algorithmic development needed to address scientific challenges. In addition, we use a co-funding model to create a team of early career researchers that will be trained in modern software engineering, while acting as liaisons that contribute to shared deliverables.

David Moulton
Los Alamos National Laboratory
Applied Mathematics and Plasma Physics
moulton@lanl.gov

Scott Painter
Oak Ridge National Laboratory
Environmental Sciences Division
paintersl@ornl.gov

Sergi Molins
Lawrence Berkeley National Laboratory
smolins@lbl.gov

Xingyuan Chen
Pacific Northwest National Laboratory
xingyuan.chen@pnnl.gov

Reed M. Maxwell
Department of Geology and Geological Engineering and IGWMC
Colorado School of Mines
rmaxwell@mines.edu

Laura Condon
University of Arizona
Department of Hydrology and Atmospheric Sciences
lecondon@email.arizona.edu

Steve G. Smith
Lawrence Livermore National Lab
smith84@llnl.gov

Hai Ah Nam
Los Alamos National Laboratory
hnam@lanl.gov

MS52

Challenges and Best Practices in the Computational Molecular Sciences

The area of computational molecular science (CMS) encompasses a diverse range of fields, including quantum chemistry, molecular dynamics, and materials science. All of these fields rely on software and computing resources to carry out simulations. Like many scientific fields, CMS faces many challenges related to software development. Some challenges arise from the mismatch of incentives between the desire for novel, career-progressing research and the need for reliable, well-written software. Other obstacles are purely technical – not only does CMS contain large legacy codebases that are still widely used, there also exists a sizable but fragmented software ecosystem. While many of these difficulties are similar to those found in other scientific fields, CMS has its own unique challenges. As the HPC community rapidly moves to heterogeneous architectures, some fields of CMS risk being left behind due to the inability to adapt not only code but algorithms and methods to the new HPC resources. As part of an investigation into the HPC field, the Molecular Sciences Software Institute (MolSSI) has conducted many interviews with researchers, developers, and users of HPC resources. This talk will review the results of these interviews, as well as what we consider to be the overarching themes we see in the community. What we consider to be best practices, and how these relate to HPC software development, will be also be discussed.

Benjamin P. Pritchard
Virginia Tech University
bpp4@vt.edu

MS53

Implementing some Strategies to Reduce Communication Time of FFT Algorithm

This work evaluates the impact of job placement, grid ordering and Slurm scheduling techniques to the performance of FFT Kampur library on a Dragonfly network cluster, i.e. Shaheen II. The evaluations demonstrate that configuration case when all processors on different blades which have physical all-to-all connections are utilized for execution the total time reduces to half. Further configurations were tested as well and will be demonstrated in this talk.

Samar A. Aseeri
Computational Scientist at KAUST
samar.aseeri@kaust.edu.sa

David E. Keyes
KAUST
david.keyes@kaust.edu.sa

Anando Chatterjee, Mahendra Verma
Indian Institute of Technology Kanpur
anandogc@iitk.ac.in, mkv@iitk.ac.in

MS53

FFTW++: A Hybrid OpenMP/MPI Implementation of FFTs and Implicitly Dealiased Convolutions

Originally developed as a C++ interface to FFTW, the

FFTW++ library has been extended to provide state-of-the-art implicit dealiasing for efficiently computing Fourier-based convolutions on serial and parallel architectures. For 1D inputs, the memory usage of implicitly dealiased and conventional zero-padded convolutions are identical. However, for large problems, implicit dealiasing is faster. In higher dimensions, the decoupling of work buffers results in significant memory savings, yielding better data locality and performance. FFTW++ uses a general multithreaded implementation of implicit dealiasing that accepts an arbitrary number of input and output vectors. Thread-safe Hermitian convolutions are implemented to avoid loop dependencies. An extended data layout enhances cache efficiency. Implicit dealiasing of higher-dimensional convolutions over distributed memory benefits significantly from the reduction of communication costs associated with its smaller memory footprint. It provides a natural way of overlapping communication with FFT computation. Our implementation relies on an adaptive matrix transposition algorithm optimized for distributed networks of multicore processors. Shared memory parallelism between the cores of a single processor is become increasing advantageous as the number of cores per processor rises. FFTW++ has been designed to exploit hybrid OpenMP/MPI parallelism and general slab-like and pencil-like decompositions.

John C. Bowman
Dept. of Mathematical and Statistical Sciences
University of Alberta
bowman@ualberta.ca

Malcolm Roberts
AMD
malcolm.roberts@amd.com

MS53

Fast Parallel Multidimensional FFT Using Advanced MPI

We present a new method for performing global multidimensional array redistributions required in the implementation of parallel fast Fourier (or similar) transforms. Our method grounds on subarray datatypes and generalized all-to-all scatter/gather from the MPI-2 standard to describe global/local array layouts and perform communication in a single collective call, thus effectively eliminating local "transpose" steps. We provide a set of compact, self-contained, high-level routines using pure MPI. These routines can be used on top of any sequential or shared-memory multidimensional FFT library to easily provide a simple and performant distributed memory implementation. A series of strong and weak scaling tests confirm our method can match and often surpass in performance other well-established libraries like MPI-FFTW, P3DFFT, and 2DECOMP&FFT.

Lisandro Dalcin
Centro Int. de Métodos Computacionales en Ingeniería
dalcinl@gmail.com

Mikael Mortensen
University of Oslo
Department of Mathematics, Division of Mechanics
mikaem@math.uio.no

David E. Keyes
KAUST
david.keyes@kaust.edu.sa

MS53

rocFFT: An Open-Source GPU FFT Library for Exascale Systems

rocFFT is an open-source software FFT library which is part of AMD's Radeon Open Compute Platform (ROCm). rocFFT is written in the HIP programming language, and designed for distributed computing with GPU-enabled MPI communications. In this presentation, I will introduce rocFFT and talk about how we are adapting the project to work on exascale systems.

Malcolm Roberts, Fei Zheng, Bragadeesh Natarajan
AMD
malcolm.roberts@amd.com, fei.zheng@amd.com, bragadeesh.natarajan@amd.com

MS54

An Overview of Particles in Amrex, with Applications to Accelerator Modelling, Cosmology, and Multi-Phase Flow

AMReX is an ECP-funded software framework for developing massively parallel block-structured adaptive mesh refinement (AMR) applications on current and upcoming supercomputing architectures. It provides the basis for the temporal and spatial discretization strategy for a number of applications - including five of the ECP application development projects - spanning fields such as accelerator design, astrophysics, combustion, cosmology, microfluidics, materials science, and multiphase flows. In addition to the conventional representation of variables on a cell, edge, face, or node-centered mesh, AMReX also provides support for particle data. Particles introduce additional irregularity and complexity to the way data is stored and operated on, requiring special attention in the presence of the dynamically changing hierarchical mesh structure, particularly when sub-cycling in time. In this talk, I will give an overview of the particle capabilities provided by AMReX and how those capabilities are used by application codes, with a particular focus on recent work towards supporting hybrid CPU/GPU platforms. Topics will include data layout, the parallel communication of particle data, including both redistribution and ghost exchange, neighbor list construction for particle-particle collisions, particle-mesh operations, and the parallel reduction of particle data.

Andrew Myers
Computational Research Division
Lawrence Berkeley National Laboratory
atmyers2@gmail.com

MS54

Rendezvous Algorithms for Large-Scale Particle Simulations

Rendezvous algorithms are useful communication patterns when you have data to send to processors who need it, but the senders don't know who the receivers are, or vice versa. The idea is to define an intermediate decomposition where data from various processors can "rendezvous", and the owning rendezvous processors can be identified by both senders and receivers. Originally designed for interpolating between overlaid grids with independent decompositions [Plimpton, Hendrickson, Stewart, "A Parallel Rendezvous Algorithm for Interpolation Between Multiple

Grids", JPDC, 64, 266-276 (2004)], we've recently realized they are also useful for operations needed at large scale in two of our particle codes: LAMMPS for molecular dynamics and SPARTA for DSMC modeling. They are particularly useful for infrequent setup operations which require data to move around the machine in arbitrary ways, but can leverage large bisection bandwidths. Brute-force alternatives e.g. circulating data in a ring to all processors so they can find what they need often work fine when processor counts are small, but can become significant bottlenecks at scale, for billions of particles and/or millions of MPI tasks. We describe how the rendezvous algorithm works, how it can be implemented simply as a flexible tool, and illustrate some dramatic performance improvements it offers. We also point out similarities with the MapReduce paradigm popularized by Google and Hadoop.

Steve Plimpton
Sandia National Laboratories
sjplimp@sandia.gov

Chris Knight
Argonne National Laboratory
knightc@anl.gov

MS54

Compatible Particle Discretization via Generalized Moving Least Squares

While an important tool well suited for particular applications, particle/meshfree methods for solving PDEs have lagged behind their mesh-based counterparts due to the loss of mathematical structures that occur when one gives up a mesh. In this talk we present results from the Compadre project at Sandia (COMpatible PArticle DiscREtization). We introduce a series of new techniques related to: conservative discretization, variational discretizations, surface PDEs and scientific machine learning. Additionally, we introduce the Compadre toolkit - a Trilinos-based package for particle discretizations utilizing Kokkos kernels for hardware acceleration.

Nathaniel Trask
Sandia National Laboratory
nat.trask@gmail.com or natrask@sandia.gov

MS55

Tensor Decomposition for Malware Detection

In this talk, we will discuss our experience in applying tensor decomposition for malware detection. We compare using n-gram and information gain with SVM to using n-gram and other features with information gain and then applying tensor decomposition as an unsupervised method. We show that tensor decomposition demonstrates comparable performance to using SVM, and may be usable with much fewer labeled data compared traditional ML methods.

Jee W. Choi
University of Oregon
jeec@uoregon.edu

MS55

PASTA: A Parallel Sparse Tensor Algorithm Benchmark Suite

Tensor computations present significant performance challenges that impact a wide spectrum of applications ranging from machine learning, healthcare analytics, social network analysis, data mining to quantum chemistry and signal processing. Efforts to improve the performance of tensor computations include exploring data layout, execution scheduling, and parallelism in common tensor kernels. This work presents a benchmark suite for arbitrary-order sparse tensor kernels using state-of-the-art tensor formats: coordinate (COO) and hierarchical coordinate (HiCOO) on CPUs and GPUs. It presents a set of reference tensor kernel implementations that are compatible with real-world tensors and power law tensors extended from synthetic graph generation techniques. We also propose Roofline performance models for these kernels to provide insights of computer platforms from sparse tensor view. This benchmark suite is publicly available: https://gitlab.com/tensorworld/pasta.

Jiajia Li
Pacific Northwest National Laboratory
Pacific Northwest National Laboratory
jiajia.li@pnnl.gov

Mahesh Lakshminarasimhan
University of Utah
maheshlakshminar@u.boisestate.edu

Xiaolong Wu
Electrical and Computer Engineering
Purdue University
xu1565@purdue.edu

Ang Li
Pacific Northwest National Laboratory
ang.li@pnnl.gov

Catherine Olschanowsky
Boise State University
catherineolschan@boisestate.edu

Kevin Barker
Pacific Northwest National Laboratory
kevin.barker@pnnl.gov

MS55

Sketching-Based Streaming Tensor Decompositions and Detection of Concept Drift in Unsupervised Exploratory Analysis

Tensors and tensor decompositions have been very popular and effective tools for analyzing multi-aspect data in a wide variety of fields, ranging from Psychology to Chemometrics, and from Signal Processing to Data Mining and Machine Learning. Using tensors in the era of big data presents us with a rich variety of applications, but also poses great challenges, especially when it comes to scalability and efficiency. In this talk, we touch upon the problem of streaming tensor decomposition, where new data "slices" are added to our existing data. An instance of this problem is the case of a time-evolving graph, where each time snapshot, a new adjacency matrix is appended to our tensor. We discuss algorithmic approaches that leverage sketching for updating the already-computed decomposition with the new data, and we present novel results on capturing and alleviating concept drift in such streaming cases.

Evangelos Papalexakis
University of California, Riverside

epapalex@cs.ucr.edu

MS55

Efficient Tensor Operations via Sketching

Sketching is a randomized dimensionality-reduction method that aims to preserve relevant information in large-scale datasets. Count sketch is a simple popular sketch which uses a randomized hash function to achieve compression. In this talk, we propose a novel extension known as Higher-order Count Sketch (HCS). While count sketch uses a single hash function, HCS uses multiple (smaller) hash functions for sketching. HCS reshapes the input (vector) data into a higher-order tensor and employs a tensor product of the random hash functions to compute the sketch. This results in an exponential saving (with respect to the order of the tensor) in the memory requirements of the hash functions, under certain conditions on the input data. Furthermore, when the input data itself has an underlying structure in the form of various tensor representations such as the Tucker decomposition, we obtain significant advantages. We derive efficient (approximate) computation of various tensor operations such as tensor products and tensor contractions directly on the sketched data. Thus, HCS is the first sketch to fully exploit the multi-dimensional nature of higher-order tensors. We apply HCS to tensorized neural networks where we replace fully connected layers with sketched tensor operations. We achieve nearly state of the art accuracy with significant compression on the image classification benchmark.

Yang Shi
Rakuten
UC Irvine
shiy4@uci.edu

MS56

A Fully Unspilt Wave Propagation Algorithm for Shallow Water Flows on GPUs

The focus of this talk is on a GPU implementation of the fully multi-dimensional patch solver based on the wave propagation algorithm (R. J. LeVeque, Clawpack). Our CUDA implementation is designed for use on small, fixed size patches (32x32) used in a larger the block-based adaptive code ForestClaw (D. Calhoun and C. Burstedde). To update patches on the GPU, we batch-process O(1000) patches per kernel call. Each patch is assigned a single CUDA thread block, eliminating the need for syncing between blocks. By redesigning the WPA, we are able to completely update the solution on each patch in a single batch kernel call. To avoid branch divergence, special attention is given to the implementation of wave limiters. Resulting time using the GPU is about 5-7x speedup over a single CPU. We will demonstrate our algorithm using examples from the shallow water wave equations implemented in ForestClaw.

Donna Calhoun
Boise State University
donnacalhoun@boisestate.edu

MS56

Improving Tsunami-HySEA as FTRT Simulator in the Framework of TEWS

In the framework of Tsunami Early Warning Systems (TEWS) is being very important the new FTRT (faster than real time) tsunami computations provided by numerical models as Tsunami-HySEA. To have available greatly improved and highly efficient computational methods are the first raw ingredient to achieve extremely fast and effective calculations for these kinds of hazards. HPC facilities have the role to bring this efficiency to a maximum possible while drastically reducing computational times. In this work, we present the improvements achieved in the Tsunami-HySEA model in the Pilot PD2 in the context of the ChEESE (Center of Excellence (CoE) in the domain of Solid Earth (SE)) European Research Project. This pilot aim is to increase the size of the problems by increasing spatial resolution and/or producing longer simulations while still computing FTRT for TEWS including inundation for a particular target coastal zone. Some examples of the achieved progresses will be shown. Acknowledgements: This research has been partially supported by ChEESE project (EU Horizon 2020, grant agreement N 823844).

Jose Manuel Gonzalez-Vida
University of Malaga, Spain
vida@anamat.cie.uma.es

Manuel J. Castro
Dpt. Análisis Matemático
University of Málaga
castro@anamat.cie.uma.es

Jorge Macias Sanchez
Universidad de Malaga
Malaga, Spain
jmacias@uma.es

Marc de la Asuncion
University of Malaga, Spain
marcah@uma.es

MS56

HPC Acceleration of the GeoClaw Software for Modeling Geohazards

The open source GeoClaw software is widely used for modeling geophysical hazards, such as tsunamis, storm surge, and overland flooding, by solving the two dimensional shallow water equations (SWE). Adaptive mesh refinement (AMR) is incorporated to efficiently handle a wide range of length scales, e.g., from tsunami propagation across the ocean to onshore inundation modeling at the scale of meters in a single simulation. OpenMP is supported, and together with AMR allows rapid simulation of many real-world problems on shared memory machines. Large-scale problems can benefit from further acceleration of the code, and we will report on some recent improvements of the software. GPU acceleration has been incorporated in order to rapidly integrate the SWE on each rectangular grid patch, while AMR management and communication between patches is performed on the CPU. The use of AMR adds several challenges to the implementation, and we discuss some issues relating to accelerating AMR codes more generally. GeoClaw has also been containerized using Docker to simplify use on cloud computing platforms. We will report some efforts to select optimal hardware configurations in this setting to minimize expense for large runs or for ensembles of many runs, and will show some timing comparisons on large-scale simulations.

Randall LeVeque
University of Washington
Applied Math

rjl@uw.edu

Kyle T. Mandli
Columbia University
Applied Physics and Applied Mathematics
kyle.mandli@columbia.edu

Xinsheng Qin
University of Washington, Seattle
xsqin@uw.edu

MS57

Low-Rank Stopping Criteria for Block Parallel SVD

The singular value decomposition (SVD) is one of the most commonly used low rank approximation techniques due to its optimality for all Schatten p-norms. However, most SVD algorithms only provide a residual stopping criteria at best, and at worst, stop based purely on iterations. Additionally, they require practitioners to determine the ideal rank, often without prior spectral information. We propose new stopping criteria for parallel block Lanczos methods and show their performance in both synthetic and real-world applications.

Steven Goldenberg
College of William & Mary
sgoldenberg@email.wm.edu

MS57

Solving Large Eigenvalue Problems with the Parallel EVSL Package

This talk will discuss how filtering techniques for eigenvalue problems can be put to work to implement 'spectrum slicing' strategies, i.e., strategies that extract slices of the spectrum independently. The presentation will begin with an overview of filtering strategies, whether polynomial or rational. Polynomial filtering can be particularly efficient for standard problems and in situations where the matrix-vector product operation is inexpensive and a large number of eigenvalues is sought. Rational filtering can be competitive in other situations and when a good parallel direct solver is available. We will also present practical implementation issues in a scalar and parallel environment and provide some details on the related EVSL package. Finally, we will discuss a collaborative effort that involves the use of EVSL to solve large eigenvalue problems that arise when computing the normal modes of the Earth.

Yousef Saad
University of Minnesota
saad@umn.edu

Ruipeng Li
Center for Applied Scientific Computing
Lawrence Livermore National Laboratory
li50@llnl.gov

Yuanzhe Xi
University of Minnesota
yxi26@emory.edu

MS57

Parallel Shift-Invert Spectrum Slicing for Symmetric Self-Consistent Eigenvalue Computation

The central importance of large scale eigenvalue problems in scientific computation necessitates the development massively parallel algorithms for their solution. Recent advances in dense numerical linear algebra have enabled the routine treatment of eigenvalue problems with dimensions on the order of hundreds of thousands on the world's largest supercomputers. In cases where dense treatments are not feasible, Krylov subspace methods offer an attractive alternative due to the fact that they do not require storage of the problem matrices. However, demonstration of scalability of either of these classes of eigenvalue algorithms on computing architectures capable of expressing excessive parallelism is non-trivial due to communication requirements and serial bottlenecks, respectively. In this work, we introduce the SISLICE method: a parallel shift-invert algorithm for the solution of the symmetric self-consistent field (SCF) eigenvalue problem. The SISLICE method drastically reduces the communication requirement of current parallel shift-invert eigenvalue algorithms through various shift selection and migration techniques based on density of states estimation and k-means clustering, respectively. This work demonstrates the robustness and parallel performance of the SISLICE method on a representative set of SCF eigenvalue problems and outlines research directions which will be explored in future work.

David B. Williams-Young
Lawrence Berkeley National Laboratory
dbwy@lbl.gov

MS57

Matrix Powers Kernels for Thick-Restart Lanczos with Explicit External Deflation

There are continual and compelling needs for computing many eigenpairs of very large Hermitian matrix in physical simulations and data analysis. Though the Lanczos method is effective for computing a few eigenvalues, it can be expensive for computing a large number of eigenvalues. This in fact is true for most of codes developed in the past even when vast computing resources are available. To improve the performance of the Lanczos method, we are developing a new TRLan eigensolver. It is an s-step thick-restart Lanczos (s-step TRLan) combined with an explicit external deflation (EED). The s-step Lanczos method can achieve an order of s reduction in data movement while the EED enables to compute eigenpairs in batches along with a number of other advantages. We will first present an overall design of the new eigensolver, and then focus on a specialized matrix powers kernel (MPK) for s-step Lanczos to reduce both communication and computation costs by taking advantage of sparse-plus-low-rank property. Performance results will be presented to demonstrate the potential of the special MPK and the new TRLan eigensolver.

Ichitaro Yamazaki
Sandia National Laboratories
iyamaza@sandia.gov

Zhaojun Bai
Departments of Computer Science and Mathematics
University of California, Davis
bai@cs.ucdavis.edu

Ding Lu
University of Geneva
ding.lu@unige.ch

Chao-Ping Lin
University of California, Davis
cplin@ucdavis.edu

Jack J. Dongarra
University of Tennessee, Oak Ridge National Laboratory, USA
dongarra@icl.utk.edu

MS58

Multigrid for Shifted Systems Appearing in Parallel-in-Time Integration

Implicit parallel time-intgration methods in many cases require to solve shifted linear systems. The shift depends on the basic time-integration method, e.g., in contour integration, and influences the spectral properties of the system. To solve the arising systems effectively, multigrid methods can be used. In the case of structured discretizations the arising matrices allow for a detailed analysis of the solver and thus its improvement. Further, the presence of structure can be exploited to obtain efficient, scalable implementations of the spatial solver. In this talk the appearing discrete operators and their spectral properties will be presented. Based on this, the analysis and design of suitable multigrid methods to takle the arising linear systems will be discussed.

Matthias Bolten
Bergische Universität Wuppertal
Fakultät für Mathematik und Naturwissenschaften
bolten@math.uni-wuppertal.de

MS58

Multigrid-in-Time SQP Methods for PDE-Constrained Optimization

Parallel solution of optimal control problems modeled by time-dependent PDEs is limited by the serial bottleneck of forward and backward (adjoint) time integration. To enable scalability in the time dimension, we introduce a multigrid-in-time solver for a special type of optimality systems, which maintains the coupling of forward and backward time integration in each time subdomain. The solver is used as the main computational kernel in composite-step sequential quadratic programming (SQP) methods for nonlinear optimization with inequality constraints, where the constraints are handled through an augmented-Lagrangian approach. We examine the weak and strong scaling of the multigrid-in-time SQP approach on optimal control problems in electromagnetics and fluid dynamics applications.

Denis Ridzal
Sandia National Laboratories
dridzal@sandia.gov

Eric C. Cyr
Computational Mathematics Department
Sandia National Laboratotories
eccyr@sandia.gov

MS58

Parallel-in-Time Simulation of the Schrdinger Equation

We discuss the applicability of parallel-in-time integration methods to the Schrdinger equation. Modern supercomputers consist of millions of cores. Hence, to use these machines efficiently for numerical simulations, implementations need to be designed to have minimal serial parts. Many current parallel solvers for time-dependent partial differential equations compute a sequence of time-steps. Here, each core computes the solution for a portion of the spatial domain, which can be done in parallel for each time-step. Each time-step, however, is computed sequentially, often limit the scalability of the implementation. Parallel-in-time integrators avoid this limit by allowing to compute multiple time-steps in parallel. While this approach works particularly well for diffusive problems, parallel-in-time integration of oscillatory problems, like the Schrdinger equation, is more difficult. The Schrdinger equation describes the time evolution of quantum mechanical systems. To understand atomic and molecular phenomena it is often useful to inspect these time evolutions. Hence, efficient ways of solving the equation are needed. Being essentially the heat equation in imaginary time, it is similarly easy to implement. It is, however, more challenging to obtain parallel speedup in time due to its oscillatory nature. Solving the Schrdinger equation hopefully provides insight on how to solve other oscillatory problems more efficiently as well.

Hannah Rittich
Forschungszentrum Jülich
h.rittich@fz-juelich.de

MS58

Performance Analysis and Benchmarking for pySDC

The parallel full approximation scheme in space and time (PFASST) allows to integrate multiple time-steps simultaneously. Based on iterative spectral deferred correction (SDC) methods, PFASST uses a space-time hierarchy with various coarsening strategies to maximize parallel efficiency. In numerous studies, this approach has been used on up to 448K cores and coupled to space-parallel solvers which use finite differences, spectral methods or even particles for discretization in space. However, since the integration of SDC or PFASST into an existing application code is not straightforward and the potential gain is typically uncertain, we have developed the Python prototyping framework pySDC. While it allows to rapidly test new ideas and to implement first toy problems more easily, it can also be used to run space-time parallel tests and applications using mpi4py. In this talk, we examine pySDCs performance on an HPC cluster and demonstrate the application of the "Scalable Performance Measurement Infrastructure for Parallel Codes" (Score-P) for analyzing the performance of our code. We highlight Python-, MPI- and PinT-specific aspects of our results and show the benefits of a structured benchmarking workflow.

Robert Speck
Juelich Supercomputing Centre
Forschungszentrum Juelich GmbH
r.speck@fz-juelich.de

MS59

The Resilience Problem in Extreme Scale Computing

Resilience is one of the critical challenges of extreme-scale high-performance computing (HPC) systems, as component counts increase, individual component reliability decreases, and software complexity increases. Building a reliable supercomputer that achieves the expected performance within a given cost budget and providing efficiency

and correctness during operation in the presence of faults, errors, and failures requires a full understanding of the resilience problem. This talk provides an overview of recent achievements in developing a taxonomy, catalog and models that capture the observed and inferred fault, error, and failure conditions in current supercomputers and in extrapolating this knowledge to future-generation systems.

Christian Engelmann
Oak Ridge National Laboratory
engelmannc@ornl.gov

MS59

Resilience for Large-Scale Iterative Linear Solvers

As HPC systems grow in scale to meet increased computational demands, the incidence of faults in a given window of time is expected to grow. This issue is addressed by the scientific community with research on solutions at every computational layer. In this presentation, we focus on algorithm-based strategies for tolerating node failures in the preconditioned conjugate gradient (PCG) method for solving large sparse linear systems. Our approach is based on exact state reconstruction, so that the solver state can be fully reconstructed if a node fails unexpectedly. In particular, we show how to support recovery from *multiple* simultaneous or overlapping failures of several nodes for general sparsity patterns of the system matrix and how to efficiently recover from node failures *without* the use of extra spare nodes, i. e., without any overhead in terms of available hardware. We also investigate the influence of the preconditioner on a trade-off between load-balancing and communication cost in the recovery phase. Our analyses and experimental evaluations of these fault tolerant PCG algorithms illustrate that the price to be paid for the strongly improved resilience is usually small and thus this approach turns out to be a promising way to make an important iterative linear solver fault tolerant. Generalizations to other solvers are possible.

Wilfried N. Gansterer
Department of Computer Science
University of Vienna
wilfried.gansterer@univie.ac.at

Markus Levonyak, Carlos Pachajoa, Christina Pacher
University of Vienna
markus.levonyak@univie.ac.at, carlos.pachajoa@univie.ac.at, christina.pacher@univie.ac.at

MS59

Resilient Computation Patterns in Scientific Applications

As HPC systems scale in size and computational power, the danger of silent errors grows dramatically. Consequently, applications running on HPC systems need to exhibit resilience to such errors. Previous work has found that, for certain codes, this resilience can come for free, i.e., some applications are naturally resilient, but few studies have shown the code patterns (i.e., combinations or sequences of computations) that make an application naturally resilient. In this talk, we present FlipTracker, a framework designed to extract these patterns using fine-grained tracking of error propagation and resilience properties. Using this framework, we present a set of computation patterns that are responsible for making representative HPC applications naturally resilient to errors. This not only enables a deeper understanding of resilience properties of these codes, but also can guide future application designs towards patterns with natural resilience.

Luanzheng Guo
University of California, Merced
lguo4@ucmerced.edu

Dong Li
the Department of Electrical Engineering and Computer Scienc
University of California, Merced
dli35@ucmerced.edu

Ignacio Laguna
Lawrence Livermore National Laboratory
lagunaperalt1@llnl.gov

Martin Schulz
TU-Munich
schulzm@in.tum.de

MS59

The Mathematical Analysis of Faults and the Resilience of Applications

As the post-Moores-Law era advances, faults are expected to increase in number and in complexity on emerging novel devices. This will happen on exascale and post-exascale architectures due to smaller feature sizes, and also on new devices with unusual fault models. Attention to error-correction and resilience will thus be needed in order to use such devices effectively. Known mathematical error-correction methods may not suffice under these conditions, and an ad hoc approach will not cover the cases likely to emerge, so mathematical approaches will be essential. We will discuss the mathematical underpinnings behind such approaches, illustrate with examples, and emphasize the interdisciplinary approaches that combine experimentation, simulation, mathematical theory and applications that will be needed for success.

Laura Monroe
Los Alamos National Laboratory
lmonroe@lanl.gov

MS60

Recent Developments Around the Block Low-Rank PaStiX Solver

Recent works in sparse direct solvers have multiplied the solutions to exploit data-sparsity additionally to sparse structure in these solvers. Among these solution, the PaStiX solver exploits block low-rank structures to reduce both its memory consumption and time to solution. In this talk, we will discuss recent changes made to the solver in this context such as the study of suitable low-rank compression kernels, adapted ordering heuristics, or scheduling strategies through the use of runtime systems. Performance results of the latest release including the presented features will be shown and discussed.

Mathieu Faverge
Bordeaux INP - Labri - Inria
mathieu.faverge@inria.fr

Gregoire Pichon
INRIA
gregoire.pichon@inria.fr

Pierre Ramet
Bordeaux University - INRIA
pierre.ramet@inria.fr

MS60

Incorporating Hierarchical Matrix Compression and Butterfly Factorizations in a Multifrontal Lu Solver

In this talk we discuss recent improvements to the fast direct solvers and rank structured preconditioners implemented in STRUMPACK, a parallel multifrontal LU solver. New frontal matrix compression schemes including HOD-LR (Hierarchically Off-Diagonal Low-Rank) and HOD-BF (Hierarchically Off-Diagonal Butterfly) have been implemented. For HOD-LR, a blocked and hierarchical version of the adaptive cross approximation algorithm is used which offers better performance, robustness and scalability. For HOD-BF, the off-diagonal blocks are compressed as butterfly decompositions, which decompose a matrix as a product of sparse factors. For 3D high frequency Helmholtz problems, HOD-BF can reduce the matrix ranks even taking into account near-field interactions and hence keeps the overall sparse solver quasi-linear. We also show how the robustness and efficiency of the compression can be drastically improved by using the graph of the original sparse matrix. The new algorithms, all purely algebraic and implemented in parallel using OpenMP and MPI, are benchmarked on some large scale test problems from DOE applications.

Pieter Ghysels
Lawrence Berkeley National Laboratory
Computational Research Division
pghysels@lbl.gov

Yang Liu
Lawrence Berkeley National Laboratory
liuyangzhuan@lbl.gov

Xiaoye S. Li
Computational Research Division
Lawrence Berkeley National Laboratory
xsli@lbl.gov

MS60

Fast H^2 Algorithms for Directly Solving General Sparse Matrices

In this work, we take advantage of the sparse linear algebra, and meanwhile develop fast $\mathcal{H}^2$-algorithms to accelerate all the dense matrix computations incurred during the direct solution procedure. It is challenging to develop an $\mathcal{H}^2$-based direct solution for sparse matrices since to take advantage of the zeros in the original matrix, we cannot treat the entire sparse matrix as an $\mathcal{H}^2$-matrix and directly invert or factorize it. If we do so, many more fill-ins will be introduced in the L and U factors, which makes the resultant direct solver slower than prevailing direct sparse solvers such as multifrontal methods. However, leveraging the framework of the multifrontal solver and accelerating all the dense frontal matrix computations is difficult, since every node in the elimination tree has its own structure, and it is difficult to communicate between different $\mathcal{H}^2$-matrices while keeping the computation nested across the elimination tree. In this work, we successfully overcome this challenge and develop a series of fast $\mathcal{H}^2$ algorithms to reduce the complexity of directly solving a sparse matrix. Numerical results will be shown to demonstrate its performance.

Dan Jiao
Electrical and Computer Engineering
Purdue University
djiao@purdue.edu

Miaomiao Ma
Purdue University
ma307@purdue.edu

MS60

Using Block-Low Rank Techniques for Large Finite Element Industrial Applications

Block-Low Rank matrices (BLR) have attracted a lot of attention in recent years. BLR matrices are dense matrices with low-rank blocks, and they are a subset of the widely-known family of hierarchical matrices (H-matrices); they can be used to design fast solvers and preconditioners. They are easy to work with and have been shown to be applicable to a wide variety of problems. BLR techniques have been implemented in different parallel sparse direct solvers, MUMPS among others; they are used to approximate and accelerate the dense kernels that appear at each step of the sparse factorization. In this presentation, we investigate the use of BLR techniques for problems arising from finite element industrial applications. We use the multiphysics code LS-DYNA and a set of large real-life problems from different applications: implicit structural mechanics (linear and nonlinear, static and dynamic), heat transfer, incompressible fluid flow, and electromagnetics. For each of these problems, LS-DYNA uses MUMPS (and its BLR feature) as a preconditioner for iterative solvers. We investigate different algorithmic parameters of BLR techniques (in particular the compression threshold) and compare BLR against different preconditioners (incomplete factorization, algebraic multigrid...). We demonstrate that BLR techniques are very robust, a key requirement for industrial applications.

Francois-Henry Rouet
Livermore Software Technology Corp.
fhrouet@lstc.com

Patrick R. Amestoy
ENINPT-IRIT, Université of Toulouse, France
patrick.amestoy@mumps-tech.com

Alfredo Buttari
CNRS-IRIT-Université de Toulouse, France
alfredo.buttari@enseeiht.fr

Jean-Yves L'Excellent
INRIA-LIP-ENS Lyon
jean-yves.l.excellent@ens-lyon.fr

Theo Mary
University of Manchester
School of Mathematics
theo.mary@manchester.ac.uk

MS61

One-Sync CGS2 Algorithm in the Context of QR Factorization and Arnoldi Process

The number of global reductions is an important metric for the parallel scalability of Krylov iterative methods. We

focus on the Arnoldi-QR algorithm for nonsymmetric matrices. The underlying orthogonalization scheme is "left-looking" and "sees" columns one at a time. Thus, at least one global reduction is required per iteration. A stable method for orthogonalizing the Krylov vectors during the Arnoldi process is the classical Gram Schmidt algorithm with reorthogonalization (CGS2), requiring three reductions per step. A new variant of Arnoldi-CGS2 that requires only one reduce has been derived. Stability and strong-scaling results are presented for finding eigenvalue-pairs of a nonsymmetric matrix. A preliminary attempt to derive a similar algorithm (one reduction per Arnoldi iteration with a robust orthogonalization scheme) was presented by Hernandez et al. 2007 [1] but their method lacks numerically stability; while our new method, after extensive experiments, is much more stable and accurate. Our algorithm can also be implemented in the context of a QR factorization (as opposed to Arnoldi), and we explain the method in this context as well. [1] V. Hernandez, J.E. Roman, A. Tomas, Parallel Arnoldi eigensolvers with enhanced scalability via global communications rearrangement. Parallel Computing 33 (2007) 521540.

Daniel Bielich
University of Colorado at Denver
daniel.bielich@ucdenver.edu

Katarzyna Swirydowicz
NREL
National Renewable Energy Lab
katarzyna.swirydowicz@nrel.gov

Stephen Thomas
National Renewable Energy Laboratory
stephethomas@gmail.com

Julien Langou
University of Colorado Denver
julien.langou@ucdenver.edu

MS61

A General Parallel Sparsified Nested Dissection Algorithm using a Task-Based Runtime System

We propose a new algorithm for the fast solution of large, sparse linear systems. It is based on nested dissection, sparsification and low-rank compression. After eliminating all interiors at a given level of the elimination tree, the algorithm sparsifies all separator. This reduces the size of the separators but without introducing any fill-in, at the expense of a small and controllable approximation error. We present the algorithm as well as a parallel version using TaskTorrent. TaskTorrent is a task-based distributed runtime system. It is lightweight, portable (using C++ threads and MPI) and requires only minor changes to the user code. We then demonstrate that a version using orthogonal factorization and block-diagonal scaling takes fewer CG iterations to converge than previous similar algorithms on various kinds of problems. Furthermore, this algorithm is provably guaranteed to never break down and the matrix stays symmetric positive-definite throughout the process. The factorization time is roughly O(N) and the number of iterations grows slowly with N. We finally show the scalability of the algorithm as a function of the number of cores. In particular, we show that the sparsification step increases the concurrency of the algorithm. By moving most of the computation towards the leaves in the sparsification step, the algorithm avoids the typical bottleneck of direct methods at the root of the tree. This significantly speeds up solving large sparse linear systems.

Leopold Cambier
Stanford University
Institute for Computational and Mathematical Engineering
lcambier@stanford.edu

MS61

Asynchronous Jacobi-Richardson, and Gauss-Seidel Smoothers for Hypre One-Synch FGMRES-AMG

We present recent advances in low-sync Krylov iterative solvers and algebraic multigrid (AMG) preconditioners for the *hypre* and Trilinos linear system solver stacks. The single-reduce orthogonalization algorithms derived by Świrydowicz et al (2019) maintain the numerical stability of the original Gram-Schmidt and GMRES algorithms . These have been extended to pipelined and s-step GMRES by Yamazaki et al (2020). Asynchronous Jacobi-Richardson and Gauss-Seidel smoothers avoid sparse triangular solves and are combined with a low-sync FGMRES-AMG solver. The set-up time for computing these smoothers on the NVIDIA Volta V100 GPU is minimal and can be combined with matrix assembly on the GPU. The multi-MPI and multi-GPU implementation of the Nalu-Wind FGMRES-AMG solver results in significant reductions in the compute time per solver iteration and extends strong scaling roll-off by at least 4x on the NREL Eagle and ORNL Summit supercomputers.

Stephen Thomas
National Renewable Energy Laboratory
stephethomas@gmail.com

Katarzyna Swirydowicz
NREL
National Renewable Energy Lab
katarzyna.swirydowicz@nrel.gov

Shreyas Ananthan, Michael Sprague
National Renewable Energy Laboratory
shreyas.ananthan@nrel.gov, michael.a.sprague@nrel.gov

MS62

Quantum Circuit Synthesis using Linear Algebra and Optimization Algorithms

Programming a quantum co-processor relies on the description of a quantum circuit which performs a series of elementary operations acting on the quantum memory to generate the final desired state. We propose two methods that optimize the synthesis of quantum circuits in two ways. The first method is based on Householder transformations and minimizes the classical resources required (flops, time) for the synthesis. The second method minimizes for some type of circuits the quantum resources (number of gates, number of qubits) using a numerical optimization algorithm. We present experimental results in the simulation of circuit synthesis using current CPU and GPU devices and we address the tradeoff between the two types of resources optimization.

Marc Baboulin
University of Paris-Sud

marc.baboulin@lri.fr

MS62

Quantum Computing – Overview Potential Applications

Quantum Computing leverages the physics of quantum mechanics to enable a new approach to computation. Recently, the technology has moved from academic research labs to the cloud, where people can access real quantum hardware from around the globe. Although quantum computers are not useful for business applications today, there is hope that continued research will enable them to help solve complex problems in the future. This talk will cover an introduction to what quantum computing is (and isn't), an exploration of potential applications, and an overview of what IBM is doing to help drive the field forward.

Gilad Ben-Shach
IBM
gilad@ibm.com

MS62

Towards Optical Neural Networks and Annealing Machines at the Quantum Limit

Progress in deep learning has relied on exponential growth of available compute power, but is facing a resource crunch as electronic systems become limited by energy consumption, interconnect bottlenecks, and the end of Moore's Law. Optics offers an intriguing alternative due to potential orders-of-magnitude advantages in latency, throughput, and energy efficiency at the bottleneck tasks for deep learning. This talk reviews the field of optical neural networks (ONNs) and introduces our new approach based on optical homodyne detection. Designing an ONN that can benefit from the advantages of optics involves maximizing the amount of optical parallelism so that a computation is not bottlenecked by I/O costs. In conventional ONNs, the optical parallelism is greatly limited by available chip area, but that the homodyne scheme circumvents this limitation by trading off spatial and time complexity. We analyze the energy consumption of ONNs and introduce a "standard quantum limit' for ONNs, set by quantum fluctuations in coherent states. This limit, which can be as low as 50 zJ/op, is more than 6 orders of magnitude beyond the current state-of-art and suggests that performance below the thermodynamic (Landauer) limit is theoretically possible in ONNs.

Ryan Hamerly, Alexander Sludds, Liane Bernstein
MIT
rhamerly@mit.edu, asludds@mit.edu, lbern@mit.edu

Marin Soljacic
Massachusetts Institute of Technology
soljacic@mit.edu

Dirk Englund
MIT
englund@mit.edu

MS62

Neuromorphic Computing: A Platform for Machine Learning and Beyond

Neuromorphic computing systems provide a promising approach for performing machine learning tasks (both training and inference) for the future of computing. They are also a compelling platform for performing neuroscience simulation. However, neuromorphic computers also have promise in performing non-machine learning, non-neuroscience tasks. Neuromorphic systems are a novel compute platform that perform specific types of computation in a massively parallel and often event-driven way, with collocated processing and memory, and often implement some form of stochasticity in the way computation is performed that provide noise. These characteristics can be leveraged to perform other types of computation, including certain graph algorithms. In this talk, the various use cases of neuromorphic systems, as well as how those use cases leverage the unique properties of neuromorphic computers, will be presented.

Catherine D. Schumann
ORNL
schumancd@ornl.gov

MS64

Lagrangian Particle Methods for Exascale Simulation of Dilute Sprays in Combustion Systems

The Lagrangian-Eulerian (LE) method is widely used for simulations of fuel sprays in turbulent combustion because of its discrete treatment of the spray droplets. We present a performance-portable library, Grit, that has been developed to simulate dilute spray evolution in preparation for exascale capable computers. The tracking of the evaporating fuel droplets represented by the Lagrangian particles are coupled to direct numerical simulations (DNS) of turbulent flows on Eulerian meshes. Grit employs the Message Passing Interface (MPI) for distributed memory parallelism and Kokkos programming model for on-node shared memory parallelism. We present details of the software implementation for the Lagrangian physics kernels and numerical operators, along with performance on the pre-exascale Summit system.

Wenjun Ge
Oak Ridge National Laboratory
gew1@ornl.gov

Ramanan Sankaran
Center for Computational Sciences
Oak Ridge National Laboratory
sankaranr@ornl.gov

MS64

The HACC Code: Particle Methods for Large-Scale Cosmological Simulations

The coming generation of cosmological surveys will provide a wealth of information regarding a number of mysteries – the nature of dark energy and dark matter, and the origin of the primordial fluctuations from which all structure in the Universe came to be. The interpretation of the observations requires a predictive computational capability that can simultaneously cover a large dynamic range in space, time, and density. Modern particle methods for the Vlasov-Poisson system of equations, as well as for an additional gasdynamic component, provide high performance on next-generation platforms as well as good control on accuracy. In this talk, I will describe the algorithms underlying HACC (Hardware/Hybrid Accelerated Cosmology Code) and describe its capabilities for carrying out extreme-scale cosmological simulations on a number of

different computational platforms.

Salman Habib
Argonne National Laboratory
habib@anl.gov

MS64

MFIX-Exa: An Exascale CFD-DEM Model for Reactor Design Engineering

Commercial deployment of gas-solids reactors requires an understanding of how to scale laboratory designs of multiphase systems to industrial sizes. However, the direct scale up of such reactors is known to be unreliable, and the current approach necessitates building and testing physical systems at increasingly larger intermediate scales. CFD-DEM (computational fluid dynamics discrete element method) offers an accurate way to model gas-solids flows and could reduce the number of physical devices that need to be built and tested thereby helping control costs and reduce risk. Because CFD-DEM tracks individual particles and resolves particle-particle-wall collisions, simulations are computationally expensive, limiting applications to small reactors containing, at most, tens of millions of particles. To address the computational and data management challenges of modeling industrially relevant devices with CFD-DEM, a new code called MFIX-Exa is being developed by migrating the core hydrodynamic models from the widely used, open-source code MFIX into the AMReX framework. Parallel performance of the code has been improved by employing a dual grid approach for fluid and particle load balancing, modernizing the fluid solution algorithm, and extending kernels to run on GPU. To handle complex reactor geometries, AMReX embedded boundary (EB) data structures and iterators were incorporated and an algorithm to address particle collisions with EB walls was implemented.

Jordan Musser
National Energy Technology Laboratory (NETL)
jordan.musser@netl.doe.gov

William Fullmer
National Energy Technology Laboratory
Leidos Research Support Team
william.fullmer@netl.doe.gov

Ann S. Almgren
Lawrence Berkeley National Laboratory
asalmgren@lbl.gov

MS64

Exascale Molecular Dynamics with Cabana: from Lennard-Jones to Neural Network Potentials

In order to effectively drive scientific discovery with tools like classical molecular dynamics in the current high-performance computing era, performance portable solutions are required. To achieve this, we use the Co-design center for Particle Applications (CoPA) Cabana particle library which, i) is built on Kokkos for on-node parallelism on various hardware, ii) provides performant particle-centric functionality for both MPI communication and neighbor search, and iii) enables optimization of data structure for a given architecture through arrays-of-structs-of-arrays (AoSoA), intermediate between arrays-of-structs and structs-of-arrays. To explore the performance of Cabana we develop the CabanaMD proxy app and test it with both the Lennard-Jones model, the near 100 year-old benchmark kernel, and a newly developed neural network potential, a kernel with near-quantum level accuracy and significantly higher computational cost. With these substantially different kernels we focus on the impact of algorithmic, data layout/access, and communication decisions on performance and scaling across hardware, including many-core CPU, among them ARM and AMD, and NVIDIA GPU. Work performed under the auspices of the U.S. DOE by LLNL under contracts DE-AC52-07NA27344 and supported by the Exascale Computing Project (17-SC-20-SC), a collaborative effort of the U.S. DOE Office of Science and the NNSA.

Samuel Reeve
Lawrence Livermore National Laboratory
reeve5@llnl.gov

Saaketh Desai
Purdue University
desai61@purdue.edu

Jim Belak
Lawrence Livermore National Laboratory
belak@llnl.gov

MS65

Streaming Cp Tensor Decompositions

We talk about ways of creating locality for randomized tensor decomposition methods, including specialized techniques for sampling large-scale sparse tensors.

Tamara G. Kolda
Sandia National Laboratories
tgkolda@sandia.gov

MS66

Minimal-Precision Computing for High-Performance, Energy-Efficient, and Reliable Computations

We introduce a joint research project with RIKEN, Sorbonne University, and University of Tsukuba for developing a minima-precision computing scheme – to obtain the floating-point result with the precision requested by users with the minimal-precision use – for high-performance (speed and energy) and reliability (accuracy and reproducibility). This is a system-level proposal involving both hardware and software stack by combining (1) a precision-tuning method through numerical validation by Discrete Stochastic Arithmetic (DSA), (2) arbitrary-precision arithmetic libraries, (3) fast and accurate numerical libraries, and (4) Field-Programmable Gate Array (FPGA) with high-level synthesis. As a result, it achieves the following goals: (a) Reliable: the precision of the result is validated by DSA. It meets the demands for accurate and reproducible computation. (b) General: it is applicable for any floating-point computations. It contributes to low development cost and sustainability. (c) Comprehensive: our system covers from the precision-tuning to the execution and combines heterogeneous hardware and hierarchal software stack. (d) High-performance: it utilizes fast numerical libraries and hardware acceleration with FPGA and GPU. (e) Realistic: it can be realized by combining available technologies. Although it is still in development, this talk introduces that the system could be constructed by combining already available technologies and that many of

them are developed by us.

Daichi Mukunoki
RIKEN, Japan
daichi.mukunoki@riken.jp

MS66
Autotuning Exascale Applications

This presentation summarizes the main features of GPTune, an autotuning framework that relies on multitask and transfer learning to find optimal performance parameters of a kernel or an application that are treated as a black-box functions. We show that the framework can be more suitable than some state-of-the-art frameworks for the tuning of any general application and, in particular, expensive large-scale applications. To illustrate, we provide comparative results with state-of-the-art autotuning techniques, for a set of solvers and linear algebra computations used by a range of applications. GPTune is part of the xSDK effort supported by DOE's Exascale Computing Project (ECP).

Wissam M. Sid-Lakhdar, Yang Liu
Lawrence Berkeley National Laboratory
wissam@lbl.gov, liuyangzhuan@lbl.gov

Xiaoye S. Li
Computational Research Division
Lawrence Berkeley National Laboratory
xsli@lbl.gov

Osni A. Marques
Lawrence Berkeley National Laboratory
Berkeley, CA
oamarques@lbl.gov

James Demmel
UC Berkeley
demmel@eecs.berkeley.edu

MS66
Verified Numerical Computations for Eigenvalue Problems on Large-Scale Parallel Systems

Eigensolvers are essential parts in a number of applications of interest. We focus on both standard and generalized symmetric eigenvalue problems. We assume that all approximate eigenpairs are given. We propose a verification method that provides error bounds of the approximate eigenvalues on the basis of the Gershgorin circle theorem. Since the main cost of the verification method is devoted to matrix multiplication, scalability of the proposed method is expected to be very high on large-scale parallel systems. We will show scalability of the verification method on the RIKEN K computer and the FUJITSU Supercomputer PRIMEHPC FX100. In addition, comparison of the error bounds of given approximate eigenvalues will be provided. It is confirmed that there is no multiple eigenvalue for test matrices in ELSES Matrix Library, which is a collection of matrix data generated by a quantum mechanical nanomaterial simulator.

Takeshi Terao
Shibaura Institute of Technology
nb17105@shibaura-it.ac.jp

Katsuhisa Ozaki
Shibaura Institute of Technology / JST
ozaki@sic.shibaura-it.ac.jp

Takeshi Ogita
Tokyo Woman's Christian University
ogita@lab.twcu.ac.jp

MS66
An Efficient Contour Integral Based Eigensolver for Surface Plasmon Simulations

Numerical simulations play a significant role for studying the properties of surface plasmon. The surface plasmon problem is first modelled by the Maxwell equations, and the equations is then discretized by the widely-used Yees scheme. After applying certain similarity transformations to the discretized system, the original simulation problem becomes a clustered non-Hermitian eigenvalue problem. An efficient contour integral (CI) based eigensolver is developed to overcome the difficulties of applying current existing methods to solve eigenvalues in particular designated regions for this problem. This efficient method combines the contour integral, the fast matrix-vector multiplication and efficient linear system solver. The numerical results can show the efficiency of solving linear systems and eigenvalues with the efficient CI eigensolver.

Huang Tsung-Ming
National Taiwan Normal University
min@ntnu.edu.tw

Weichung Wang
National Taiwan University
Institute of Applied Mathematical Sciences
wwang@ntu.edu.tw

Wen-Wei Lin
National Chiao Tung University
wwlin@math.nctu.edu.tw

William Liao
National Taiwan University
r05246012@ntu.edu.tw

MS67
Parallel Computation of Many Eigenvalues by S-Step Thick-Restart Lanczos Algorithm with Explicit External Deflation - II

There are continual and compelling needs for computing many eigenpairs of very large Hermitian matrix in physical simulations and data analysis. Though the Lanczos method is effective for computing a few eigenvalues, it can be expensive for computing a large number of eigenvalues. This in fact is true for most of codes developed in the past even when vast computing resources are available. To improve the performance of the Lanczos method, we are developing a new TRLan eigensolver. It is an s-step thick-restart Lanczos (s-step TRLan) combined with an explicit external deflation (EED). The s-step Lanczos method can achieve an order of s reduction in data movement while the EED enables to compute eigenpairs in batches along with a number of other advantages. In this talk, we will focus on stability analysis of the new TRLan eigensolver and show the parallel computation of many eigenvalues for physical simulations in electronic structure calculations and dynamics analysis of supramolecular systems.

Zhaojun Bai

Departments of Computer Science and Mathematics
University of California, Davis
bai@cs.ucdavis.edu

Jack J. Dongarra
University of Tennessee, Oak Ridge National Laboratory, USA
dongarra@icl.utk.edu

Chao-Ping Lin, Ding Lu
University of California, Davis
cplin@ucdavis.edu, dinlu@ucdavis.edu

Ichitaro Yamazaki
Sandia National Laboratories
iyamaza@sandia.gov

MS67

Adaptive Step Size Strategies for Line Search Methods and their Applications to Electronic Structure Calculations

In this talk, we will introduce an adaptive step size strategy for a class of line search methods for orthogonality constrained minimization problems, which avoids the classic backtracking procedure. We prove the convergence of the line search methods equipped with our adaptive step size strategy under some mild assumptions.We then apply the adaptive algorithm to electronic structure calculations, which show that our strategy is efficient and recommended. This is a joint work with Liwei Zhang and Aihui Zhou.

Xiaoying Dai
Institute of Computational Mathematics
Chinese Academy of Sciences, Beijing, China
daixy@lsec.cc.ac.cn

MS67

Eigensolvers for Ab Initio CI Calculations in Nuclear Physics

I describe a recently developed block iterative method to efficiently obtain the lowest eigenpairs of large sparse symmetric matrices that arise in ab initio configuration interaction (CI) calculations for atomic nuclei. The dimensions of these matrices can be in the (tens of) billions, and one therefore needs efficient algorithms and implementations for large-scale high-performance computing platforms. Rapid convergence of the block iterative method is achieved by a suitable choice of starting guesses of the eigenvectors and the construction of an effective distributed preconditioner. The use of a block method leads to greater concurrency and increased arithmetic intensity of the computation, and allows us to take advantage of e.g. the vector units on intel's xeon phi processors. I also discuss the implementation details that are critical to achieving high performance on massively parallel multi-core supercomputers, and demonstrate that this block iterative solver is two to three times faster than the Lanczos algorithm for problems of moderate sizes on a Cray XC40 system.

Pieter Maris
Iowa State University
pmaris@iastate.edu

MS67

A Scalable Matrix-Free Eigensolver for Studying Many-Body Localization

We present a scalable and matrix-free eigensolver for studying nearest-neighbor Heisenberg spin chain plus random on-site disorder models that undergo a many-body localization (MBL) transition. This type of problem is computationally challenging because its dimension grows exponentially with the physical system size, and the solve must be iterated many times to average over different configurations of the random disorder. For each eigenvalue problem, eigenvalues from different regions of the spectrum and their corresponding eigenvectors need to be computed. Traditionally, the interior eigenstates for a single eigenvalue problem are computed via the shift-and-invert Lanczos algorithm. Due to the extremely high memory footprint of the LU factorizations, this technique is not well suited for large number of spins L, e.g., one needs thousands of compute nodes on modern high performance computing infrastructures to go beyond $L = 24$. We propose a new matrix-free approach that does not suffer from this memory bottleneck and even allows for simulating spin chains up to $L = 24$ spins on a single compute node. We discuss the OpenMP and hybrid MPI–OpenMP implementations of matrix-free block matrix-vector operations that are the key components of the new approach. The efficiency and effectiveness of the proposed algorithm is demonstrated by computing eigenstates in a massively parallel fashion, and analyzing their entanglement entropy to gain insight into the MBL transition.

Roel Van Beeumen
Lawrence Berkeley National Laboratory
rvanbeeumen@lbl.gov

Gregory D. Meyer, Norman Y. Yao
University of California, Berkeley
gregory.meyer@berkeley.edu, norman.yao@berkeley.edu

Chao Yang
Lawrence Berkeley National Lab
cyang@lbl.gov

MS68

Alleviating the Memory Pressure for Seismic Modeling and Imaging

We revisit the high performance implementations of the first order and second order formulations of the 3D acoustic wave equation using two cache blocking techniques: spatial blocking (SB) and multicore wavefront diamond temporal blocking (MWD-TB) techniques. Solving the acoustic wave equation is cornerstone of seismic modeling, seismic imaging and seismic inversion. Various hardware platforms are considered to assess the performance impact of each optimization technique.

Rached Abdelkhalak
KAUST
rached.abdelkhalak@kaust.edu.sa

Hatem Ltaief
King Abdullah University of Science & Technology (KAUST)
hatem.Ltaief@kaust.edu.sa

David E. Keyes
KAUST

david.keyes@kaust.edu.sa

MS68

Porting Sewas Task-Based Seismic Code on Arm Platforms

The production of reliable three-dimensional images of the subsurface remains a major challenge in the oil and gas industry and strongly relies on the efficient exploitation of supercomputers. In recent years, heterogeneous hardware has gained a lot of traction, requiring a profound shift on the way numerical applications are implemented. In this paper, we highlight the key role of high-level programming models and efficient runtime systems to design next generation seismic wave propagation codes. In our case, the application dataflow is built on top of PaRSEC, a generic task-based runtime system. We discuss the results obtained with the implementation of a finite-differences numerical scheme on various platforms including AArch64 architectures.

Fabrice Dupros
ARM
Fabrice.Dupros@arm.com

Salli Moustafa
Aneo, France
siam@siam.org

Conrad Hillairet
ARM
conrad.hillairet@arm.com

MS68

Devito - DSL for Generating MPI and OpenMP Parallel Finite Difference Operators for Seismic Imaging

Devito is an innovative Python project based on domain-specific languages (DSL) and compiler technology for rapid implementation of high-performance, structured-grid solvers for partial differential equations (PDEs). In FWI/RTM for example, Devito is used to implement efficient, full-fledged production-grade wave propagators. In this article, we introduce Distributed-memory parallelism (DMP) in Devito. The key innovations are the abstractions provided to the user and the compiler-based implementation approach, which we consider invaluable for long-term sustainable software to replace (partly or fully) obsolete, impenetrable, hardly extendable and often inefficient legacy code. Our approach allows users to think as sequentially as possible all major aspects of DMP are handled by Devito, including domain decomposition, data accessing (through distributed NumPy arrays), and optimized communications. We provide compelling evidence demonstrating the efficiency of our technology at all levels of abstraction. In particular, we show the results of strong scaling experiments on single-node multi-socket as well as multi-node systems, using production-level seismic inversion propagators employed in FWI/RTM.

Gerard J Gorman
Department of Earth Science and Engineering
Imperial College London
g.gorman@imperial.ac.uk

Fabio Luporini
Department of Computing
Imperial College London
f.luporini12@imperial.ac.uk

Rhodri Nelson
Imperial College, United Kingdom
rhodri.nelson@imperial.ac.uk

MS68

A Library to Accelerate Stencil Codes on Vector Processors

One of most costly computational kernels in seismic imaging is "stencil code", which iteratively updates each element in a multidimensional grid by referring to the neighbor elements. Since it frequently appears also in other wide variety of scientific simulations, image processing, deep learning, and so on, it is quite an important issue to make stencil codes perform faster. As a solution, we developed a library that can accelerate stencil codes on vector processors, which is one of most suitable hardware architectures for processing a massive amount of data with high performance. Stencil codes load a value of each element several times while they store a new value once. Hence, for gaining performance, it is required to reduce memory load access. We optimized stencil code computation by a combination of two techniques. The first one is loop blocking, which is a well-known technique to utilize a cache efficiently. The second one is a unique technique using vector registers like a high-speed cache by elaborate loop structure deformation. Every vector register can have hundreds of element values and can be accessed much faster than a cache. The library running on vector processors shows significant performance compared to stencil codes running on commonly used scalar processors.

Arihiro Yoshida
NEC Corporation
a-yoshida@ap.jp.nec.com

MS69

Teaching Parallel and Distributed Computing Concepts in Simulation

Teaching topics related to high performance, parallel and/or distributed computing in a hands-on manner is challenging, especially at introductory, undergraduate levels. There is a participation challenge due to the need to secure access to a platform on which students can learn via hands-on activities, which is not always possible. There are also pedagogic challenges as particular platforms provided to students impose constraints on which learning objectives can be achieved hands-on. These challenges become steeper as the topics being taught target more heterogeneous, more distributed, and/or larger platforms, as needed to prepare students for using and developing Cyberinfrastructure. In this presentation we will elaborate on the above challenges, and report on successful experiences with using simulation as a pedagogic tool in the classroom for teaching the above topics. These experiences relied on hands-on, simulation-driven, pedagogic modules developed with the SimGrid and the WRENCH simulation frameworks. These modules can be integrated piecemeal into existing university courses, and we will explain how they have contributed to students achieving learning objectives in specific undergraduate and graduate computer science courses taught at the University of Hawai'i at Manoa.

Henri Casanova
University of Hawai'i
henric@hawaii.edu

Ryan Tanaka, Rafael Ferreira da Silva
University of Southern California
tanaka@isi.edu, rafsilva@isi.edu

MS69

Fast and Faithful Performance Prediction of MPI Applications: the HPL Case Study

Finely tuning MPI applications (number of processes, granularity, collective operation algorithms, topology and process placement) is critical to obtain good performance on supercomputers. With a rising cost of modern supercomputers, running parallel applications at scale solely to optimize their performance is extremely expensive. Having inexpensive but faithful predictions of expected performance could be a great help for researchers and system administrators. The methodology we propose captures the complexity of adaptive applications by emulating the MPI code while skipping insignificant parts. We demonstrate its capability with High Performance Linpack (HPL), the benchmark used to rank supercomputers in the TOP500 and which requires a careful tuning. We explain (1) how we both extended the SimGrid's SMPI simulator and slightly modified the open-source version of HPL to allow a fast emulation on a single commodity server at the scale of a supercomputer and (2) how to model the different components (network, BLAS, ...) of the system. We show that a careful modeling of both spatial and temporal node variability allows us to obtain predictions within a few percents of real experiments.

Tom Cornebize
Université Grenoble Alpes, France
tom.cornebize@inria.fr

Arnaud Legrand
CNRS
arnaud.legrand@inria.fr

Franz Christian Heinrich
Inria
franz-christian.heinrich@inria.fr

MS69

Power-Aware Scheduling with Slurm: Simulation and Practice

Power management is one of the key research challenges on the path to exascale. Supercomputers today are designed to be worst-case power provisioned, leading to two main problems – limited application performance and underutilization of procured power. In this talk, we will present a practical and simulation perspective on the development and evaluation of a low-overhead resource manager targeted at future power-constrained clusters. Our approach is based on an adaptive policy, which derives job-level power bounds in a fair-share manner and supports overprovisioning and power-aware backfilling.

Tapasya Patki
Lawrence Livemore National Laboratory
patki1@llnl.gov

MS69

The Many Faces of Simulation for HPC

In this talk, we will recall how simulation has been widely used in the field of HPC research and development, to evaluate and compare the performance of application implementations and of the algorithms therein. We will provide a rapid survey of the tools developed to this end over the last decade. Then we will explain how recent advances in the simulation methodologies at the core of available simulation frameworks open the way for the study of other compelling use cases beyond performance analysis.

Frédéric Suter
CC-IN2P3 / CNRS
frederic.suter@cc.in2p3.fr

Rafael Ferreira da Silva
University of Southern California
rafsilva@isi.edu

MS70

Resilience in the Context of GPUs: A Technique for Interrupting GPU Kernels

In computing, resilience is a way of maintaining software operability in the face of unanticipated faults that challenge normal execution. Resilience strategies are typically employed in systems comprised of many computers as it is more likely to suffer faults when computing at this scale. In recent years, these systems have been massively augmented with graphics process- ing units (GPUs) to accelerate tasks. GPUs present a new challenge to resilience by being specialized pieces of hardware which do not follow the traditional execution model relied upon by established strategies. Further, new or retrofitted strategies quickly become obsolete as they typically rely on undocumented behaviour which often later changes due to the rapidly evolving GPU landscape. This work proposes a new approach to checkpointing MPI applications that spend the majority of their execution time on the GPU. It is possible to take snapshots of data residing on GPUs without waiting for kernels to complete. The proposed technique is implemented in the context of FTI, a state-of-the-art high performance fault tolerance library. The result is an elegant solution for developing resilient MPI applications where kernels run longer than the mean time between hardware failures.

Max M. Baird
Heirott-Watt University, The United Kingdom
mmb1@hw.ac.uk

MS70

New Non-Blocking Extensions to the ULFM Proposal

This talk presents new asynchronous extensions to the User Level Failure Mitigation (ULFM) MPI Standard draft proposal. The ULFM proposal, under evaluation by the MPI Forum's Fault Tolerance Working Group, support the continued operation of MPI programs after permanent failures have impacted the execution. The key principle is that no MPI call (point-to-point, collective, RMA, IO, ...) can block indefinitely after a failure, but must either succeed or raise an MPI error. The new extensions will allow applications to construct non-blocking faster-changing worlds, and fulfill the resilient needs or more dynamic application.

George Bosilca
University of Tennessee Knoxville
bosilca@icl.utk.edu

Aurélien Bouteiller
University of Tennessee, Knoxville, US
bouteill@icl.utk.edu

Nuria Losada
University of A Coruña
nuria.losada@udc.es

MS70

Performance Portable and Productive Resilience using Kokkos

Performance portable abstractions, such as Kokkos, RAJA, and HPX are becoming more and more prevalent among HPC applications. Kokkos in particular not only defines performance portable execution abstractions, but also performance portable data abstractions, such as the `Kokkos::View`. We propose a framework for both implicit and explicit checkpointing of `Kokkos::View` in addition to a resilient version of `Kokkos::parallel_for` to provide redundant computation for reliability. Our framework provides a natural API for checkpointing that allows applications already using Kokkos for performance portability to implement fault tolerance with minimal and non-intrusive code changes. In the presentation, we will discuss the implementation of the resilience Kokkos extension and a couple of application use cases.

Jeffery Miles
Sandia National Laboratories, New Mexico
jsmiles@sandia.gov

Nicholas Morales
Sandia National Laboratories, California
nmmoral@sandia.gov

Carson Mould
Georiga Institute of Technology
cmould@sandia.gov

Bogdan Nicolae
MCS Division
Argonne National Laboratory
bnicolae@anl.gov

Keita Teranishi
Sandia National Laboratories
Livermore, California
knteran@sandia.gov

MS70

Composing Asynchrony, Communication and Resilience

Resilience is an imminent issue for next-generation systems due to projected increases in soft/transient failures. Much of the work for resilience has been focused on traditional bulk-synchronous parallel programming models. We believe that Asynchronous Many-Task (AMT) models are better suited to enabling resilience since they provide explicit abstractions of data and tasks which contribute to more scalable and portable approaches. In this talk, we introduce a comprehensive approach to enabling resiliency in AMT programming models, along with integrating the capability to perform asynchronous inter-node communication. Specifically, we make the following contributions:

1. programming model extensions to enable resilience techniques from past work (task replay, task replication, ABFT) to be applied to AMT applications along with asynchronous communication
2. unified execution framework that supports arbitrary composition of these extensions and non-resilient tasks
3. implementation of our approach as extensions to the Habanero C/C++ library

Our results show that we can enable resilience in AMT programs with low overheads, and we demonstrate the ability to combine different resilience schemes along with communication.

Sri Raj Paul
College of Computing
Georgia Institute of Technology
sriraj@gatech.edu

Akihiro Hayashi
Rice University, USA
ahayashi@rice.edu

Nicole Slattengren
Sandia National Labs
nlslatt@sandia.gov

Hemanth Kolla
Sandia National Laboratories
hnkolla@sandia.gov

Seonmyeong Bak
Georiga Institute of Technology
sbak5@gatech.edu

Matthew Whitlock
College of Computing
Georiga Institute of Technology
mwhitlo@sandia.gov

Jackson Mayo
Sandia National Laboratories
jmayo@sandia.gov

Keita Teranishi
Sandia National Laboratories
Livermore, California
knteran@sandia.gov

Vivek Sarkar
College of Computing
Georgia Institute of Technology
vivek.sarkar@cc.gatech.edu

Max Grossman
Rice University, Houston, Texas
jmaxg3@gmail.com

MS71

Towards Continuous Benchmarking: An Automated Performance Evaluation Framework for High Performance Software

We present an automated performance evaluation framework that enables an automated workflow for testing and performance evaluation of software libraries. The Continuous Integration (CI) framework employed for the Ginkgo library is extended to a Continuous Benchmarking framework (CB) which automatically runs benchmark tests on

predetermined HPC systems, stores the state of the machine and the environment along with the compiled binaries, and collect results in a publicly accessible performance data repository. The Ginkgo performance explorer (GPE) can be used to retrieve the performance data from the repository, and visualizes it in a web browser. The combination of Continuous Integration, Continuous Benchmarking, and the Ginkgo Performance Explorer creates a workflow which enables performance reproducibility and software sustainability.

Hartwig Anzt
Steinbuch Centre for Computing
Karlsruhe Institute of Technology
hartwig.anzt@kit.edu

Terry Cojean
Karlsruhe Institute of Technology
terry.cojean@kit.edu

Goran Flegar
Universitat Jaume I
flegar@uji.es

Pratik Nayak, Yuhsiang M. Tsai
Karlsruhe Institute of Technology
pratik.nayak@kit.edu, yu-hsiang.tsai@kit.edu

MS71

Faster Spack Package Manager Installations through Task Parallelism

Package managers, containers, automated testing, and Continuous Integration (CI), are becoming an essential part of HPC development workflows. These automated tools often require re-compiling entire software stacks from source and spend a significant amount of time and computing resources just for compilation and integration. However, large software stacks such as those deployed on HPC clusters can have complex combinatorial dependencies, and may take a system several days to compile. Despite the use of simple parallelization (such as 'make -j'), build execution time often do not scale with system resources. For such cases, it is possible to improve overall installation time by compiling parts of software stack independently, each scheduled on a subset of available cores. We apply malleable-task scheduling algorithms to better exploit available parallelism in build system workflows and improve stack build time overall. Using a prototype implementation in the Spack package manager, malleable-task scheduling can improve build times by more than 2x.

Samuel Knight, Jeremiah Wilke
Sandia National Laboratories
sknigh@sandia.gov, jjwilke@sandia.gov

Todd Gamblin
Lawrence Livermore National Laboratory
gamblin2@llnl.gov

MS71

Extreme-Scale Scientific Software Stack (E4S)

The DOE Exascale Computing Project (ECP) Software Technology focus area is developing an HPC software ecosystem that will enable the efficient and performant execution of exascale applications. Through the Extreme-scale Scientific Software Stack (E4S) [https://e4s.io], it is developing a comprehensive and coherent software stack that will enable application developers to productively write highly parallel applications that can portably target diverse exascale architectures. E4S provides both source builds through the Spack platform and a set of containers that feature a broad collection of HPC software packages. E4S exists to accelerate the development, deployment, and use of HPC software, lowering the barriers for HPC users. It provides container images, build manifests, and turn-key, from-source builds of popular HPC software packages developed as Software Development Kits (SDKs). This effort includes a broad range of areas including programming models and runtimes (MPICH, Kokkos, RAJA, OpenMPI), development tools (TAU, HPCToolkit, PAPI), math libraries (PETSc, Trilinos), data and visualization tools (Adios, HDF5, Paraview), and compilers (LLVM), all available through the Spack package manager. It will describe the community engagements and interactions that led to the many artifacts produced by E4S. It will introduce the E4S containers are being deployed at the HPC systems at DOE national laboratories using Singularity, Shifter, and Charliecloud container runtimes.

Sameer Shende
Dept. of Computer and Information Science
University of Oregon
sameer@cs.uoregon.edu

MS71

XSDK: Toward Efficient and Interoperable Scientific Library Collection

With the development of increasingly complex architectures and software due to multiphysics modeling, and the coupling of simulations and data analytics, applications increasingly require the combined use of software packages developed by diverse, independent teams throughout the HPC community. The Extreme-scale Scientific Software Development Kit (xSDK) is being developed to provide such an infrastructure of independent mathematical libraries to support rapid and efficient development of high-quality applications. This presentation will introduce the xSDK, its history and discuss in more detail the development and impact of the xSDK community policies, which were defined to achieve improved code quality and compatibility across xSDK member packages and constitute an integral part of the xSDK.

Piotr Luszczek
University of Tennessee
luszczek@icl.utk.edu

Satish Balay
Argonne National Laboratory
balay@anl.gov

Keita Teranishi
Sandia National Laboratories
Livermore, California
knteran@sandia.gov

James Willenbring
Sandia National Laboratories
jmwille@sandia.gov

Lois Curfman McInnes
Mathematics and Computer Science Division
Argonne National Laboratory
curfman@mcs.anl.gov

Ulrike Meier Yang
Lawrence Livermore National Laboratory
yang11@llnl.gov

Asim YarKhan
Innovative Computing Laboratory
University of Tennessee
yarkhan@icl.utk.edu

MS72

Adaptive Local Timestepping and its Parallelization

Systems of conservation laws model a large cross section of scientific applications of interest. Explicit timestepping schemes must satisfy the famous CFL stability condition. However, in the presence of large variations of wavespeed or mesh size, existing timestepping schemes can be overly conservative. Local timestepping is an adaptive timestepping technique, whereby regions of the mesh stably advance with varying timestep sizes. Verifying that correctness is enforced in a distributed context requires evaluating a task graph that cannot be determined statically. We propose a novel adaptive local timestepping scheme that has been formulated as a discrete event simulation that guarantees the CFL constraint is satisfied. Important features of this work are (1) a formal correctness proof using loop invariants that guarantees the correctness of this algorithm in a highly asynchronous execution context, and (2) the use of an optimistic (i.e. speculative) discrete event simulator to efficiently parallelize the algorithm. Preliminary scaling results will outline the performance compared to a flat MPI implementation using synchronous timestepping. We expect the techniques used to develop this timestepping scheme to have broad crossover to algorithm design in highly asynchronous settings.

Max Bremer
University of Texas at Austin
max@oden.utexas.edu

John Bachan
Lawrence Berkeley National Laboratory
jdbachan@lbl.gov

Cy Chan
Future Technologies Group
Lawrence Berkeley National Laboratory
cychan@lbl.gov

Clint Dawson
Institute for Computational Engineering and Sciences
University of Texas at Austin
clint@ices.utexas.edu

MS72

Leveraging Random Walks and Neuromorphic Hardward to Solve Elliptical Integro-PDEs

The bulk of recent research into neuromorphic hardware applications has focused on whether it is suitable for use with emerging AI algorithms. The field of neuromorphic computing must determine whether hardware will primarily be used for algorithms that brains implement or if hardware can have a broader impact outside of AI applications. We explore the latter, providing a framework for numerically solving certain PDEs on large-scale spiking neuromorphic platforms. In previous work, we have demonstrated two efficient and parallelizable random walk algorithms for spiking circuits. Random walk solutions to PDEs, including results such as the Feynman-Kac formula, are ideal applications of highly efficient and parallel algorithms. In this talk we showcase the relation between a class of elliptical integro-PDEs and jump diffusion SDEs, developing a general algorithm for implementation on neuromorphic hardware. We demonstrate proof of concept on three applications: basic diffusion, Black-Scholes option pricing models, and particle transport. Finally we discuss the scaling implications of our algorithm on existing spiking neuromorphic platforms.

Brad Aimone, J. Darby Smith
SNL
jbaimon@sandia.gov, jsmit16@sandia.gov

MS72

Design of New Streaming and Graph Analytics Algorithms for the Strider Architecture

A vertex-centric, push-based implementation of a graph analytics algorithm is equivalent to a sparse reduction problem, where the algorithm decides the associative reduction operator. Obtaining high performance from these problems is particularly challenging as they suffer from minimal spatio-temporal locality, depending upon the input. By automatically and intelligently marshaling an input sparse data stream using specialized hardware, the Strider architecture is able to maximize bandwidth utilization from on-chip as well as off-chip memory subsystems, while also performing parallel reduction. As a result, Strider is able to significantly and efficiently accelerate SpMV and SpGEMM based graph analytics problems. Furthermore, as Strider treats the input as a dynamic sparse data stream with no pre-ordering assumed, it is able to handle both static and evolving graphs. This talk presents an overview of the Strider architecture, and the design of graph analytics algorithms to best leverage such an architecture.

Sriseshan Srikanth, Thomas M. Conte
Georgia Institute of Technology
seshan@gatech.edu, conte@gatech.edu

MS72

The Rogues Gallery as a Testbed for Novel Algorithm Design for Future Architectures

The Rogues Gallery is a new testbed deployment at Georgia Tech for understanding next-generation hardware with a focus on unorthodox and uncommon technologies for what is commonly referred to as post-Moore computing. This testbed project was originally initiated in 2017 in response to IEEE Rebooting Computing efforts and initiatives for new hardware designs. The Gallery's focus is to acquire new and unique hardware (the rogues) from vendors, research labs, and start-ups and to make this hardware widely available to students, faculty, and industry collaborators within a managed data center environment. By exposing students and researchers to this set of unique hardware, we hope to foster cross-cutting discussions about hardware and software design that will drive future performance improvements in computing long after the Moore's Law era of cheap transistors ends. We present highlights of the first two years of post-Moore era research with the Rogues Gallery. Specifically, we focus on how the Rogues Gallery has supported new algorithm and application development efforts in the areas of near-memory,

neuromorphic, and quantum computing with some brief lessons learned on how we can better support future research in these unique research areas.

Jeffrey Young
Georgia Institute of Technology
jyoung9@gatech.edu

Jason Riedy
Georgia Institute of Technology
School of Computational Science and Engineering
jason.riedy@cc.gatech.edu

Thomas M. Conte
Georgia Institute of Technology
conte@gatech.edu

Vivek Sarkar
College of Computing
Georgia Institute of Technology
vivek.sarkar@cc.gatech.edu

MS73

Particle-in-Cell Simulations at Exascale

Current and future exascale systems are introducing unparalleled levels of heterogeneity, with planned Department of Energy (DoE) alone offering at least three radically different hardware configurations. This presents a challenge for all codes, especially particle simulations which rely on high levels of performance and large particle counts to achieve the desired level of scientific fidelity. Increasingly, we are seeing a shift from code developers deploying platform specific optimizations to a focus on achieving portable-performance as code developers need to be able to run on all available platforms. In this work we present a performance-portable version of the VPIC which aims to deliver high-performance on current and up-coming DoE systems through the deployment of the Kokkos programming model. This gives us a single, easy to read, code base that is not obfuscated by platform specific optimizations. In so doing, we aim to allow code users and physicists to focus on scientific research, and removing the focus on code development. As part of this work we offer a study into performance-portability across different hardware, and demonstrate the codes ability to scale successfully across both Summit and Sierra. Finally, we offer some broader picture insights for Particle-in-Cell codes as a whole as we head into the exascale era.

Robert F. Bird
LANL
LANL
bird@lanl.gov

MS73

Pumipic: Infrastructure for Unstructured Mesh Pic on GPUs

Designing an efficient infrastructure for unstructured mesh particle-in-cell, PIC, simulations running on GPUs requires precise approaches to promote performance and ease of development for implementing physical simulations. This talk presents the PUMIPic library that provides mesh-centric data structures and operations for unstructured mesh PIC simulations. PUMIPic uses a GPU based unstructured mesh library, Omega, with a generalized particle data structure designed to support a range of PIC simulations. The design of both structures has been strongly influenced by need for performant parallel execution of PIC operations that involve irregular data dependencies when performing the PIC operations that involve mesh-article interactions. To ensure the ability to scale to any desired problem size, both the mesh and particle data can be distributed. Even with the mesh-based PUMIPic structures that tie the particles to the mesh, supporting the distributed mesh complicates the algorithms needed to scale on todays heterogeneous systems. The methods being developed to minimize data movement/communication and support load effective balancing will be discussed. Performance results will be presented for execution on the Summit system at Oak Ridge National Laboratory.

Gerrett Diamond
Rensselaer Polytechnic Institute
diamog@rpi.edu

Cameron W. Smith
Scientific Computation Research Center
Rensselaer Polytechnic Institute
smithc11@rpi.edu

Chonglin Zhang
Scientific Computation Research Center Rensselaer Polytechni
zhangc20@rpi.edu

Eisung Yoon, Gopan Perumpilly, Onkar Sahni
Rensselaer Polytechnic Institute
yoone@rpi.edu, gopan.p@gmail.com, sahni@rpi.edu

Mark S. Shephard
Rensselaer Polytechnic Institute
Scientific Computation Research Center
shephard@rpi.edu

MS73

Taking the Plasma Physics Code XGC to Summit and Beyond with Kokkos/Cabana

XGC is a particle-in-cell code that models plasma physics in fusion devices. To achieve legible and sustainable source code that runs with optimized performance across platforms, we adapted XGC to utilize Kokkos, a portability framework that manages data layout and execution strategies, and Cabana, a library built on Kokkos within the ECP-CoPA project with a focus on optimizations specific to particle codes. To further complicate things, XGC is a Fortran code, while Kokkos and Cabana are in C++. Building additional interfaces to make use of these libraries thus posed an additional challenge. Here we summarize our experience adopting them. We report optimized performance on Summit and on Cori supercomputers, demonstrating a viable path for Fortran codes seeking cross-platform performance. However, we also point to the limitations of this strategy and present our progress towards implementing a C++ version of XGC in preparation for future architectures.

Aaron Scheinberg, Stephane Ethier
Princeton Plasma Physics Laboratory
ascheinb@pppl.gov, ethier@pppl.gov

Guangye Chen, Bob Bird
Los Alamos National Laboratory
gchen@lanl.gov, bird@lanl.gov

Stuart Slattery

Computational Sciences and Engineering Division
Oak Ridge National Laboratory
slatterysr@ornl.gov

CS Chang
PPPL
cschang@pppl.gov

MS73

WarpX: Electromagnetic Particle-in-Cell with Adaptive Mesh Refinement for Advanced Particle Accelerators

Turning the current experimental plasma accelerator state-of-the-art from a promising technology into mainstream scientific tools depends critically on high-performance, high-fidelity modeling of complex processes that develop over a wide range of space and time scales. As part of the U.S. Department of Energys Exascale Computing Project (ECP), WarpX[1] is an electromagnetic Particle-In-Cell (PIC) code for plasma dynamics, with specific features for plasma-based particle accelerators. It combines adaptive mesh refinement from the ECP co-design center AMReX with advanced algorithms, including a Lorentz boosted frame technique and spectral solvers based on local FFTs. WarpX has been used for production runs on CPU, and has been ported to GPU with significant speedup. The general organization and main features of the code will be presented. When running on a GPU-accelerated architecture, the whole PIC loop is executed on the GPU to minimize host-device communications. Main challenges and strategy for porting the code to GPU will be addressed, with special focus on particle handling. [1] [J.-L.Vay, Warp-X: A new exascale computing platform for beamplasma simulations, Nuclear Instruments and Methods in Physics Research Section A, 909, p. 476]

Maxence Thevenet
Lawrence Berkeley National Laboratory
mthevenet@lbl.gov

Jean-Luc Vay
Lawrence Berkeley National Laboratory
Berkeley, CA
JLVay@lbl.gov

Weiqun Zhang
Lawrence Berkeley National Laboratory
Center for Computational Sciences and Engineering
weiqunzhang@lbl.gov

Rémi Lehe
Lawrence Berkeley National Laboratory
ATAP
rlehe@lbl.gov

Andrew Myers
Lawrence Berkeley National Lab
atmyers@lbl.gov

Ann S. Almgren
Lawrence Berkeley National Laboratory
asalmgren@lbl.gov

John B. Bell
CCSE
Lawrence Berkeley Laboratory
jbbell@lbl.gov

David Grote
Lawrence Livermore National Laboratory
dpgrote@lbl.gov

Olga Shapoval
LBNL
OShapoval@lbl.gov

Axel Huebl
Lawrence Berkeley National Laboratory (LBNL), USA
Helmholtz-Zentrum Dresden-Rossendorf (HZDR), Germany
axelhuebl@lbl.gov

Jaehong Park
Lawrence Berkeley National Laboratory
ATAP
jaehongpark@lbl.gov

Ligia Diana Amorim
Lawrence Berkeley National lLaboratory
ldianaamorim@lbl.gov

Mark Hogan
SLAC National Accelerator Laboratory
mjhogan@stanford.edu

Lixin Ge, Cho-Kuen Ng
Stanford Linear Accelerator Center
lge@slac.stanford.edu, cho@slac.stanford.edu

MS74

Communication Lower Bounds for Rectangular MTTKRPS

Our goal is to establish lower bounds on the communication required to perform the Matricized-Tensor Times Khatri-Rao Product (MTTKRP) computation on a distributed-memory parallel machine. MTTKRP is the bottleneck computation within algorithms for computing the CP tensor decomposition, which is an approximation by a sum of rank-one tensors and frequently used in multidimensional data analysis. We will present a communication lower bound that generalizes previous results, tightening the bound so that it is attainable even when the tensor dimensions vary (the tensor is not cubical) and when the number of processors is small relative to the tensor dimensions. We'll also describe the communication-optimal algorithm that attains the lower bound. Finally, we'll give highlights from the proof, which includes a novel approach based on convex optimization.

Grey Ballard, Kathryn Rouse
Wake Forest University
ballard@wfu.edu, kathryn.rouse@inmar.com

MS74

A Distributed Memory Generalized Sparse Tensor Decomposition

The generalized canonical polyadic (GCP) low-rank tensor decomposition method of Hong, Kolda and Duersch (SIAM Review, 2019) allows use of a variety of loss functions to enable decomposition of tensors representing binary or count data. Kolda and Hong (2019) propose solving the GCP problem using a stochastic gradient descent (SGD) method; implementations of SGD-GCP are available in TensorToolbox (for Matlab) and Genten (for GPU

and multicore processors). In this talk, we present a distributed memory implementation of SGD-GCP. We discuss sampling strategies for SGD in a parallel environment, and present results for large tensors from the FROSTT tensor repository.

Karen D. Devine
Sandia National Laboratories
kddevin@sandia.gov

Grey Ballard
Wake Forest University
ballard@wfu.edu

Tamara G. Kolda
Sandia National Laboratories
tgkolda@sandia.gov

MS74

Computing Generalized CP Decompositions on Emerging Parallel Architectures

In this talk we describe the computation of generalized canonical polyadic (GCP) tensor decompositions of large, sparse tensors on emerging manycore architectures such as multicore CPUs, Intel Xeon Phi, and Nvidia GPUs. In particular we describe the implementation of numerical sampling and gradient kernels arising in the computation of GCP decompositions using stochastic gradient descent algorithms using a software tool called Kokkos that enables a single software implementation of these kernels that is portable to and performant on a variety of contemporary manycore architectures. Performance of GCP decompositions on several data sets and hardware platforms will be presented. All of the numerical methods described have been encapsulated into a publicly available software tool called GenTen.

Eric Phipps
Sandia National Laboratories
Center for Computing Research
etphipp@sandia.gov

Tamara G. Kolda
Sandia National Laboratories
tgkolda@sandia.gov

MS74

Load Balancing Strategy of Parallel Performance Portable Sparse CP-APR Decomposition

Sparse CP-APR tensor decomposition, extended from nonnegative matrix factorization, enables effective analysis of large tensors, representing count (non-negative and discrete) data. Despite the effectiveness, its parallel implementation poses several challenges because of the increasing diversity of parallel computing architecture and irregularity of sparsity patterns of real application data. We have addressed these issues using Kokkos parallel programming model to implement sparse CP-APR code for multiple HPC platforms. However, Kokkos default load balancing and dynamic scheduling capability is lacking some flexibility to mitigate the extreme irregularity found in the tensors from real applications. In this work, we will present our load balancing strategies of our parallel sparse CP-APR decomposition and discuss the performance and the issues of high performance parallel programming model and library support for the emerging applications.

Keita Teranishi
Sandia National Laboratories
Livermore, California
knteran@sandia.gov

David S. Hollman
Sandia National Labs
dshollm@sandia.gov

Jeremy Myers
The College of WIlliam and Mary
jmmyers01@email.wm.edu

Richard Barrett
Sandia National Laboratories
rfbarre@sandia.gov

Daniel M. Dunlavy
Sandia National Laboratories
Computer Science and Informatics
dmdunla@sandia.gov

MS75

Search Space Reduction Through Analytical Modeling

As architectures become more complex, and more tuning factors are introduced, the time required to obtain a fast implementation via auto-tuning is expected to continue to increase. In this talk, we focused on using analytical modeling techniques to model the hardware in order to reduce the search space so that auto-tuning can be employed on parameters and hardware features that are harder to model.

Tze Meng Low
Carnegie Mellon University
lowt@andrew.cmu.edu

MS75

Reproducible Linear Algebra from Application to Architecture

All computing must be parallel to take advantage of modern systems like multicore processors, GPUs, and distributed systems. Results that are not bit-wise reproducible introduce doubt on many levels. Sometimes that is appropriate. Reproducibility limitations occur because underlying libraries do not specify their reproducibility requirements. New advances in interfaces, algorithms, and architectures allow selecting among those requirements in the future. This talk covers many of the upcoming options and their trade-offs.

Jason Riedy
Georgia Institute of Technology
School of Computational Science and Engineering
jason.riedy@cc.gatech.edu

James W. Demmel
University of California
Division of Computer Science
demmel@berkeley.edu

Peter Ahrens
Massachusetts Institute of Technology

pahrens@mit.edu

MS75

Stable Automatic Tuning Method for Performance Fluctuation

Automatic tuning (AT) is a technique to search for optimum parameter value settings of a user program. The execution time of a program depends on the value of a given parameter, but execution time for a single parameter setting can vary from one run to the next. We call this fluctuation. In our tests, the fluctuations were found to depend on the execution condition of the computing environment. The bad influence of fluctuation on parameter optimization must be mitigated. Thus, our goal is set to consider fluctuations and realize a stable AT. We researched the iterative one-dimensional (1D) search that can search good parameter with a little search cost. The iterative 1D search is based on an approximation function (d-Spline). The d-Spline function was chosen because it flexibly follows the measured data with small calculation cost. To realize a stable AT, we extend the conventional iterative 1D search. The proposed method calculates the improvement rate of the estimated best parameter after each 1D parameter search, and starts the multiple-time measurements of a single parameter value after the estimated best parameter is expected to nearly optimum. Moreover, after the iterative 1D search is ended, it keeps and updates average execution time of each parameter setting, and choses the parameter setting with shortest execution time. In our numerical evaluation, the proposed method significantly reduced the total execution time in comparison with conventional methods.

Naoto Seki, Toshiki Tabeta, Akihiro Fujii, Teruo Tanaka
Kogakuin University
em18006@ns.kogakuin.ac.jp, em19009@ns.kogakuin.ac.jp, fujii@cc.kogakuin.ac.jp, teru@cc.kogakuin.ac.jp

MS75

Precision Tuning of the Arithmetic Units in Matrix Multiplication on FPGA

Data precision affects memory demand and computation performance. However, current computer systems only support limited data formats, such as single precision, double precision, and users can not customize their own data format. This results in low efficiency of computation resources. In this research, the precision tuning of data in matrix multiplication on FPGA is investigated, and its impacts on system performance are evaluated, including computation performance, energy efficiency, and hardware resource utilization.

Toshiyuki Imamura
RIKEN Center for Computational Science
imamura.toshiyuki@riken.jp

Yiyu Tan
RIKEN Advanced Institute for Computational Science
tan.yiyu@riken.jp

MS76

Scalable Kohn-Sham Matrix Algebra Solutions with the ELSI Infrastructure

This talk describes the open-source infrastructure 'ELSI' (https:// wordpress.elsi-interchange.org/ and Ref. [1]), which provides simple access to state-of-the art solutions to the Kohn-Sham eigenvalue problem for different codes and solvers using a single uniform interface. ELSI provides solutions ranging from simple serial to large-scale massively parallel execution, with efficient matrix conversion between dense and sparse matrix formats. Supported solvers include ELPA (massively parallel $O(N^3)$ eigenproblem solutions), PEXSI ($O(N^2)$ density-matrix based solutions including for metallic systems), NTPoly (O(N) density matrix purification), and several further, specialized solvers. ELSI is a cross-code development, now used in production versions of FHI-aims, Siesta, DFTB+, and DGDFT; additionally, ELSI is part of the broader 'Electronic Structure Library' Bundle of open-source libraries for electronic structure theory. Different solvers have different use scenarios in terms of system size, system type and parallelism, assessed in a comprehensive set of benchmarks in this talk. Finally, we outline a new reverse communication interface (RCI) enabling the facile, efficient implementation of different iterative solver strategies aimed at plane wave basis sets, led by ELSI project members Yingzhou Li and Jianfeng Lu (Duke University). [1] V. Yu *et al.*, Comput. Phys. Commun. **222**, 267 (2018).

Volker Blum
Duke University
Department of Mechanical Engineering and Materials Science
volker.blum@duke.edu

MS76

Real-Space Parallel Eigensolvers for Electronic Structure Calculations: The Future

Two physical ingredients, pseudopotentials and density functional theory, are widely used and accepted in electronic structure computations for a wide variety of materials applications. If we wish to address large, complex systems, the implementation of these ingredients on high performance computational platforms is vital. Real space grid methods offer a compelling vehicle for such computations. These methods are mathematically robust, very accurate and well suited for modern, massively parallel computing resources. I will illustrate the utility of these methods as implemented in the PARSEC code. Key algorithms in this code include subspace filtering based on Chebyshev polynomials, spectrum slicing for added level of parallelism, Cholesky QR algorithms to improve the performance of orthogonalization, and a 2D partition of the wave functions for efficient matrix-vector operations. Applications will be illustrated for nanostructures containing tens of thousands of atoms.

James R. Chelikowsky
Institute for Computational Engineering and Sciences
University of Texas at Austin
jrc@utexas.edu

MS76

A Parallel Strategy for Kohn-Sham Solver with GPU-Accelerated Nodes

We consider the problem of solving the Kohn-Sham eigenvalue problem using a very accurate numerical discretization such as finite differences, finite elements or plane waves (pseudo-spectral) on a distributed memory architecture with very fast nodes, such as GPU-accelerated nodes. Traditionally, wave functions would be distributed among nodes, while matrices expressing operators in the subspace

spanned by trial wave functions would be distributed using a library such as ScaLapack. We investigate how the increasing performance gap between floating point operations and off-node communications affects this strategy. We use a proxy-app to investigate performance and develop an effective strategy for strong scaling of problems with O(500-10,000) wave functions on the Summit supercomputer at the Oak Ridge Leadership Computing Facility (OLCF). Research sponsored by the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory (ORNL), managed by UT-Battelle, LLC for the U. S. Department of Energy under Contract No. De-AC05-00OR22725. This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

Jean-Luc Fattebert, Luigi Capone
Oak Ridge National Laboratory
fattebertj@ornl.gov, caponel@ornl.gov

Massimiliano Lupo Pasini
Computational Sciences and Engineering Division
Oak Ridge National Laboratory
lupopasinim@ornl.gov

Bruno Turcksin
Oak Ridge National Laboratory
turcksinbr@ornl.gov

MS76

Revisiting the Jacobi Method for Eigen Problems in Computational Chemistry

Jacobi methods for symmetric eigenvalue problems have attracted attention because of their inherent parallelism. In this work, we study the application of Jacobi methods in electronic structure calculations. In particular, we consider how the number of Jacobi sweeps could be reduced in certain situations. We also discuss the potential regimes for which the Jacobi algorithm could be preferred over the QR algorithm, in terms of problem size and number of processing units.

Hua Huang
Georgia Institute of Technology
huangh223@gatech.edu

MS77

Hiding I/O Latency in Large-Scale Scientific Computation Through Buffering and Prefetching on Advanced Storage Technologies

Simulation systems often produce a massive amount of data that cannot fit a single compute node, and they also require to read back these data for computation. As a result, I/O data movement can be a bottleneck in large-scale simulations. Advances in memory architecture have made it feasible and affordable to include multiple heterogeneous storage media in a single compute node. A typical workstation nowadays contains several Hard Disk Drives (HDDs) and Solid State Drives (SSDs) while an advanced one might contain a high-throughput Non-Volatile Memory (NVMe). However, while adding additional and faster storage media increases I/O bandwidth, it pressures the Central Processing Unit (CPU) as it becomes responsible for managing and moving data between these layers of storage. Simulation systems are thus vulnerable to being blocked by I/O operations. The system, Multilayer Layer Buffer System (MLBS), proposed in this talk, demonstrates a general and versatile method for overlapping I/O with computation that helps to ameliorate the strain on the processors through asynchronous access, while minimizing the impact on the computational kernel.

Tariq Alturkestani
KAUST, Saudia Arabia
tariq.alturkistani@kaust.edu.sa

David E. Keyes
KAUST
david.keyes@kaust.edu.sa

MS77

Distributed ML-Based Applications in Oil Gas

ML-based applications in Oil Gas are maturing fast, and with that the need to access and train on large datasets is becoming a must. We will introduce example applications, distributing learning frameworks and how these frameworks help our applications to scale and solve large datasets. The examples will range from core plugs to reservoir level.

Mauricio Araya-Polo
Total, U.S.
mauricio.araya@total.com

Denis Akhiyarov
Total, Inc., U.S.
denis.akhiyarov@total.com

MS77

Large Scale Inversion of CSEM Data in the Time Domain

Controlled-source electromagnetic (CSEM) time-domain data acquired by seabed receivers provide an imaging modality that can be used, either alone or jointly with seismic data, for resolving resistive bodies in increasingly complex geological settings and improving identification of hydrocarbon reserves. There are however a number of computational challenges that have to be overcome in order to develop parallel scalable inversion algorithms for CSEM data, particularly in the time domain. These include the need for efficient and robust models for the propagation of electromagnetic disturbances in a reservoir, governed by Maxwell equations in the low-frequency regime. Implicit time integration schemes are necessary for insuring stable simulations, and the required linear solves at every time step represent a key challenge for scalability. On the optimization side, there is a need for efficient gradient and Hessian-vector products needed for a Krylov solver, which require adjoint solves as well as incremental forward and incremental adjoint solves for their evaluation. Regularization of the optimization problem must allow non-smooth solutions and promote consistency between vertical and horizontal resistivities. This may be done through Total Variation (TV) vector versions (VTV). In this talk, we describe parallel algorithms for addressing these challenges and show illustrative results on three-dimensional problems.

Stefano Zampini
King Abdullah University of Science and Technology
stefano.zampini@kaust.edu.sa

George M. Turkiyyah
American University of Beirut
gt02@aub.edu.lb

David E. Keyes
KAUST
david.keyes@kaust.edu.sa

MS78

New Horizons for Debugging Long-Running Parallel Programs: DMTCP and SimGrid

A confluence of recent advances in two independent domains has serendipitously opened up new horizons for debugging long-running programs. The two independent domains are: (a) checkpoint-restart; and (b) process simulation through stateless model checking. In each case, the recent advances concern extensions to better support parallel and distributed computation (e.g., MPI). The speaker's own team supports a long-running project, DMTCP (Distributed MultiThreaded CheckPointing), for checkpoint-restart. A new approach, called split processes, allows DMTCP to now checkpoint under one implementation of MPI, and then restart under another implementation of MPI. In particular, the second implementation of MPI can in fact be the smpi/SimGrid simulator, which targets the production code of an MPI application. Thus, a new collaboration was born, which now allows simulation and debugging of a long-running parallel program. SimGrid, in combination with DMTCP, can now simulate a restarted MPI application. That restarted application is based on a checkpoint that occurred before a crash or failed assert. SimGrid can then produce an execution trace showing a deterministic schedule from the time of the checkpoint to the crash or failed assert. This new collaboration originated soon after the recent extension of SimGrid to support distributed simulations of MPI applications. But then a simpler approach was found, and therein lies an interesting story.

Gene Cooperman
Khoury College of Computer Sciences
Northeastern University, Boston, MA
gene@ccs.neu.edu

Rohan Garg
Northeastern University
Boston, MA
rhn.grg@gmail.com

MS78

Application-Simulation Co-Design for Performance and Correctness Evaluation

For large-scale HPC applications, the search space of possible optimization techniques is usually vast. It is common to try several algorithmic alternatives to relax synchronizations to allow applications to opportunistically use resources, or to rely on automatic scheduling and load-balancing techniques. However, (1) the respective efficiency of such approaches is difficult to assess and highly depends on the target system; and (2) the corresponding implementations are quite complex and thus error-prone. It is not uncommon in dynamic and highly optimized approaches to obtain efficient implementations that suffer from rare non-deterministic deadlocks or failstop errors which are extremely difficult to narrow and debug. Additionally, direct experiments provide only limited experimental control, hindering the reproducibility of experiments. Simulation is an appealing alternative to study the performance and the correctness of such systems. In this talk, we explain how partially porting the BigDFT application on top of SimGrid simulation toolkit led us to very promising results. We present the co-design approach in which modifications were done to the scientific application and SimGrid to enable and ease its simulation in various experimental conditions. If successful, this approach will represent a radical epistemic shift; the developer will be able to test and validate the performance of an exascale run without the need for a huge amount of computational resources.

Luigi Genovese
Institut Nanosciences et Cryogénie, CEA
luigi.genovese@cea.fr

Augustin Degomme
CEA Grenoble
augustin.degomme@cea.fr

MS78

Faithful Performance Prediction of a Dynamic Task-Based Runtime System, an Opportunity for Task Graph Scheduling

Obtaining the maximum performance of heterogeneous machines is challenging as it requires to carefully offload computations and manage data movements between different processing units. Task-based runtime systems allow to abstract the question and rely on opportunistic scheduling algorithms. The problem then gets shifted to choosing a task granularity and task graph structure, and optimizing scheduling strategies. Trying different combinations of these different alternatives is however a challenge by itself. Indeed, getting accurate measurements requires reserving the target system for the whole duration of experiments. Furthermore, observations are limited to the few available systems at hand and may be difficult to generalize. In this talk, I will describe how we have crafted a coarse-grain hybrid simulation/emulation of StarPU, a dynamic runtime for hybrid architectures, over SimGrid, a versatile simulator for distributed systems. This approach allows to obtain performance predictions of classical dense linear algebra kernels accurate within a few percents and in a matter of seconds, which allows both runtime and application designers to quickly decide which optimization to enable or whether it is worth investing in higher-end GPUs or not. Additionally, it allows to conduct robust and extensive scheduling studies in a controlled environment whose characteristics are very close to real platforms while having reproducible behavior.

Samuel Thibault
University of Bordeaux
samuel.thibault@labri.fr

Luka Stanisic
INRIA Grenoble Rhône-Alpes
luka.stanisic@inria.fr

Arnaud Legrand
CNRS
arnaud.legrand@inria.fr

Brice Videau
INRIA Grenoble - Rhône-Alpes
brice.videau@imag.fr

Jean-François Méhaut
Université Joseph Fourier, Grenoble
jean-francois.mehaut@imag.fr

MS79

Robust Dynamic Load Balancing of Scientific Applications against System Variability and Failures

Exploiting the full computational power of HPC systems for extreme-scale scientific applications is a complex challenge due to load imbalance, system variability, fail-stop failures and silent data corruptions (SDCs). Dynamic load balancing (DLB) improves the performance of such applications, typically via the use of self-scheduling-based techniques. However, the robustness of self-scheduling-based techniques against variability in and failures of system components remains an open problem. Existing fault-tolerant self-scheduling techniques need fault-detection mechanisms to trigger the rescheduling of tasks whose execution failed. In this talk, we present a *robust dynamic load balancing* (rDLB) approach for the robust self-scheduling of scientific applications with independent tasks on HPC systems under system variability and/or failures. We also present a *selective particle replication* (SPR) approach for detecting SDCs in particle-based scientific applications. The rDLB proactively reschedules already allocated tasks, requires no detection of system variability or failures, and is integrated into an MPI-based DLB library. The experiments show that rDLB boosts application robustness against system variability and fail-stop failures by up to 30% compared to the case without rDLB. The SPR replicates the computation and data of a few carefully selected particles and achieves detection rates of 91-99.9%, no false-positives, at an overhead of 1-10%.

Ali Mohammed, Aurélien Cavelan, Florina M. Ciorba
University of Basel, Switzerland
ali.mohammed@unibas.ch, aurelien.cavelan@unibas.ch, florina.ciorba@unibas.ch

MS79

Space-Efficient Reed-Solomon Encoding to Detect and Correct Pointer Corruption

Concern about memory errors has been widespread in HPC for decades. These concerns have led to significant research on detecting and correctingmemory errors to improve performance and to provide strong guarantees about the correctness of the memory contents of scientific simulations. However, power concerns and changes in memory architectures threaten the continued viability of current approaches to protecting memory (e.g., Chipkill). Returning to a less protective error-correcting code, e.g., SECDED, may increase the frequency of memory errors, including silent data corruption (SDC). SDC has the potential to silently cause applications to produce incorrect results and mislead domain scientists. In this presentation, we provide a detailed description of how we can exploit unnecessary pointer bits to store Reed-Solomon parity symbols. We evaluate the performance impacts of this approach and examine the effectiveness of the approach against corruption. Our results demonstrate that encoding and decoding is fast (less than 45 us per event) and that the protection it provides is robust (the rate of miscorrection is less than 5% even for significant corruption).

Scott Levy, Kurt Ferreira
Sandia National Laboratories
sllevy@sandia.gov, kbferre@sandia.gov

MS79

Silent-Error Detection, Local Recovery, and Failure Masking in MPI-Based Solvers

Efficient local detection of silent data corruption in parallel scientific computations, using "physics-based checksums", enables straightforward recovery via checkpoint/restart. We discuss advantages of solver implementations that permit purely local recovery (restarting only processes with errors) using an extension of the Fenix fault tolerance library. These techniques can improve scalability and reduce failure delays, offering some of the benefits of task-based programming models within the familiar SPMD setting. [SNL is managed and operated by NTESS under DOE NNSA contract DE-NA0003525.]

Hemanth Kolla, Jackson Mayo, Rob Armstrong
Sandia National Laboratories
hnkolla@sandia.gov, jmayo@sandia.gov, rob@sandia.gov

MS79

Modern Use of Checkpoint-Restart at Large Scale using VeloC

This talk presents recent advances of VeloC, a multi-level checkpoint/restart runtime for high performance computing applications that takes advantage of and delivers high performance and scalability for modern, complex heterogeneous storage hierarchies without sacrificing ease of use and flexibility. First, the talk introduces the need for checkpointing at Exascale and associated challenges. Next, it highlights the key features of VeloC: (1) exposes a simple application-level API to checkpoint and restart HPC applications based on either protecting application data structures directly or managing application-defined checkpoint files; (2) hides the complexity of interacting with the storage hierarchy (burst buffers, etc.) of current and future HPC systems; (3) has a modular design that facilitates flexibility in choosing the resilience strategies and mode of operation (synchronous or asynchronous), while being highly customizable with additional post-processing modules; (3) supports use cases beyond fault tolerance: suspend-resume jobs over multiple reservations, revisit previous states (e.g. adjoint computations), etc. Furthermore, it provides a brief overview of the techniques introduces by VeloC to implements these features. Finally, the talk concludes with examples of and early results with HPC applications that aim for Exascale and use VeloC for checkpointing, as well as exploratory scenarios where checkpointing is of interest for large-scale deep learning.

Bogdan Nicolae
MCS Division
Argonne National Laboratory
bnicolae@anl.gov

Adam Moody
Lawrence Livermore National Laboratory, USA
moody20@llnl.gov

Kathryn Mohror
Lawrence Livermore National Laboratory
mohror1@llnl.gov

Franck Cappello
ANL

cappello@mcs.anl.gov

PP1

Adapting a Multibody System Simulator to Auto-Tuning Linear Algebra Routines

Multibody systems (MBS) consist of a set of bodies interconnected through mechanical joints which allow combined movements among them. Computational kinematics studies the movement of MBS from different approaches. One recently developed modular approach divides a MBS into a set of modules whose kinematics can be solved in a hierarchical order to analyse the complete MBS. Each module is solved efficiently using a block of linear algebra routines which organize their computation in tasks which can be carried out sequentially or simultaneously in different ways to reduce the simulation time in contemporary standard computational nodes (multicore CPU+multiGPU). This particular modular approach, implemented in a multibody system simulator (MBSS) can be applied to other problems which admit a hierarchical solution of modules defined by specific blocks of linear algebra routines, although it needs to be adapted to determine the appropriate block size in algorithms by blocks and the assignation of tasks to the computing units. This poster presents the structure of the MBSS and includes possible modifications to adapt it to help in the auto-tuning of linear algebra routines. Examples and results with Strassen's matrix multiplication and an LU factorization by blocks are shown.

Jesús Cámara
University of Murcia
jcamara@um.es

José-Carlos Cano
Politechnic University of Cartagena
josecarlos.canol@um.es

Javier Cuenca
Departamento de Ingeniería y Tecnología de Computadores
University of Murcia
jcuenca@um.es

Domingo Giménez
Departamento de Informática y Sistemas
University of Murcia
domingo@um.es

Mariano Saura-Sánchez
Politechnic University of Cartagena
msaura.sanchez@upct.es

PP1

Fast Generation of Extreme-Scale Matrices with Preassigned 2-Norm Condition Number

Generating benchmark problems to test numerical linear algebra algorithms at scale presents new challenges. The test problems must be inexpensive to generate, yet large enough to saturate the computational capabilities of the supercomputer being tested. In many cases, being able to tune some parameters of the test problem may also be required. When assessing the performance of iterative solvers for linear systems, for instance, one typically desires to generate square matrices with a preassigned 2-norm condition number. For small-scale problems, such a matrix can be constructed from its singular value decomposition, as it is well known that the 2-norm condition number is the ratio of the largest to the smallest singular value. This approach, however, requires a dense matrix-matrix multiplication, an operation that has cubic cost and is communication-intensive in a distributed setting, and can quickly become too costly as the size of the problem grows. We propose a new method to generate extremely large matrices with preassigned 2-norm condition number. In order to produce a matrix of order n, our technique requires only $\mathcal{O}(n^2)$ floating point operations and minimal communication between the nodes in the cluster.

Massimiliano Fasi
The University of Manchester
School of Mathematics
massimiliano.fasi@manchester.ac.uk

Nicholas J. Higham
School of Mathematics
The University of Manchester
nick.higham@manchester.ac.uk

PP1

Multicolor Block Gauss-Seidel Using Kokkos

The classical Gauss-Seidel method is a very common preconditioner, but it is inherently sequential. Gauss-Seidel may be adapted into a distributed algorithm by coloring each processor's subdomain so that all processors sharing a color can update the global residual in parallel. This can be further improved by partitioning each local subdomain before coloring [Saad 99]. The Gauss-Seidel method may also be parallelized on a shared memory system by coloring individual rows. This approach achieves high performance on GPUs and multicore CPUs but converges more slowly than the classical method [Deveci 16]. We present a shared-memory version of multicolor block Gauss-Seidel using Kokkos that allows the user to manage the tradeoff between parallelism and convergence by tuning the block size.

Brian Kelley, Sivasankaran Rajamanickam
Sandia National Laboratories
bmkelle@sandia.gov, srajama@sandia.gov

PP1

A Parallel Framework for Nonlinear Optimization

Optimizations performed in todays scientific applications have to contend with finding the global optimum for highly nonlinear and nonconvex functions. Even when the functions are differentiable, the optimization techniques developed so far either find a local optimum, or do not scale well in terms of dimension for the problem of finding the global optimum. This problem is further compounded when closed form expressions for these functions are unavailable, thereby needing expensive simulations to find their value at the current iterate. In this poster, we will present a preliminary framework to bridge this gap. We show that by extending the existing gradient-based and gradient-free approaches to make use of HPC infrastructure, it is possible to efficiently find the global optimum for these highly nonlinear and nonconvex functions. Additionally, we will show how we use this framework to solve parameter inversion problems in high-energy physics where optimization needs to be performed over high-dimensional nonlinear and nonconvex functions. We hope that this effort will broaden the understanding of how to make use of state-of-the-art HPC infrastructure to find the global op-

timum in ever increasing complexity of optimization problems and will encourage further research into developing such techniques.

Mohan Krishnamoorthy
Argonne National Laboratory
mkrishnamoorthy@anl.gov

PP1

Massive Scaling of MASSIF: Algorithm Development for Hooke's Law Simulations on Distributed GPU Systems

Seven out of the top ten supercomputers are powered by Graphic Processing Units (GPUs) and can help solve computationally intensive problems in various fields. However, adapting legacy Fortran code with large working sets and low arithmetic intensity to GPUs with small on-chip memories is a challenge due to high memory requirements and communication patterns in these simulation algorithms traditionally designed to run on CPUs or CPU clusters. One such simulation is Micromechanical Analysis of Stress-Strain Inhomogeneities with Fourier transforms (MASSIF), an iterative Hooke's law partial differential equation solver. In this work, we describe algorithm development for porting MASSIF to heterogeneous platforms. We propose a domain decomposition method with multi-resolution octree-based adaptive sampling on domain-local results to enable computation within GPU memory constraints, and ease of accumulation of results over distributed GPUs. A first-order performance model of our method estimates that compression and multi-resolution sampling strategies can enable domain computation within GPU memory constraints for 3D grids larger than those simulated by the current state-of-the-art Fortran MPI implementation. Theoreticalmodel evaluation on (1)Single node, multi-GPU (2)DGX2 workstation (3)Summit (ORNL) provides insight into design requirements for further scalability. We also discuss potential for cross-platform high performance implementations using emerging FFT APIs like FFTX.

Anuva Kulkarni
Carnegie Mellon University
anuvak@andrew.cmu.edu

Jelena Kovacevic
New York University
jelenak@nyu.edu

Franz Franchetti
Carnegie Mellon University
franzf@cmu.edu

PP1

Accelerating Alternating Least Squares for Tensor Decomposition by Pairwise Perturbation

The alternating least squares algorithm for CP and Tucker decomposition is dominated in cost by the tensor contractions necessary to set up the quadratic optimization subproblems. We introduce a novel family of algorithms that uses perturbative corrections to the subproblems rather than recomputing the tensor contractions. This approximation is accurate when the factor matrices are changing little across iterations, which occurs when alternating least squares approaches convergence. We provide a theoretical analysis to bound the approximation error. Our numerical experiments demonstrate that the proposed pairwise perturbation algorithms are easy to control and converge to minima that are as good as alternating least squares. We tested our algorithm on 1, 16 and 256 Intel KNL nodes of the Stampede2 supercomputer. The performance of one pairwise perturbation approximation step is 7.8-10.5X faster than one state of the art ALS step, and the overall performance of the new algorithms shows improvements of 1.3-2.8X with respect to alternating least squares approaches for various model tensor problems and real datasets.

Linjian Ma
University of Illinois at Urbana Champaign
lma16@illinois.edu

Edgar Solomonik
University of Illinois at Urbana-Champaign
solomon2@illinois.edu

PP1

Application of Projectile Physics and Variable Drag Implications in Determining Market Price Movements for Futures Derivatives

This particular study took an econophysics route to explain the market behaviour for futures contracts in terms of prices and its market life span. We used projectile motion models defined under two distinct conditions (perfect/horizontal and imperfect/ drag implication) based on Newtons and Galileos laws of motion. Despite that it was more theoretical we managed to derive the futures price functions and the results showed that futures prices depends largely on market forces of demand and supply and underlying assets price behavior. Also, we managed to find the terminal prices for the securities given the initial prices, which is a worrying matter to the trading parties. From the performance comparison of the two models used, results suggested that futures price function from a drag variable is more powerful in modelling the price behavior for options than the one sorely controlled by market demand and supply forces. However, it should not be carelessly taken that the projectile models used are much good at price motions/movements within the market from time to time with a stunted ability to capture in other facts of interest such as volatility coefficients which paves a research way for other scholars.

Leonard Mushunje
stusent at
Midlnads state university
leonsmushunje@gmail.com

PP1

Simulating Quantum Circuits with Tensor Network States

One challenge in simulating quantum circuits is that the memory necessary to store the qubits grows exponentially in general as the number of qubits grows. On the other hand, tensor networks provide a compact way to approximately represent quantum states with specific entanglement patterns. We develop a parallel quantum circuit simulator based on tensor network states like Projected Entangled Pair States (PEPS), and compare it with the standard state vector approach. These simulators are implemented using Cyclops Tensor Framework (CTF) that supports efficient parallel numerical operations for tensors in distributed memory. The benchmark we use includes random quantum circuits, which are considered to be generally hard to simulate. We investigate various aspects of

these two simulating methods such as speed, memory cost, accuracy, scaling, etc. and discuss the tradeoff between them.

Yuchen Pang, Yiqing Zhou, Tianyi Hao
Department of Computer Science
University of Illinois at Urbana-Champaign
yuchenp2@illinois.edu, yiqing2@illinois.edu,
tianyih2@illinois.edu

Edgar Solomonik
University of Illinois at Urbana-Champaign
solomon2@illinois.edu

PP1

Parallel Fenics Implementation of Block Solvers

The FEniCS Project is a popular and sophisticated open-source software library for conducting general-purpose finite element simulations. Once the variational formulations are assembled, FEniCS relies on other scientific libraries like PETSc to solve the discretized linear and nonlinear systems of equations. However, when it comes to mixed finite element formulations, neither parallel direct solvers nor simple implementations of Krylov methods with standard preconditioning techniques will be sufficient to solve large-scale problems. Instead, one needs to split them into blocks and apply said preconditioning techniques to them individually. This paper presents an overview of an open-source software library called pFibs: a parallel FEniCS implementation of Block Solvers, which provides a FEniCS interface to PETSc's composable block solver capabilities. The software enables users to perform nested block preconditioning, provide their own variational form for the preconditioner, and interface directly with Dolfin-Adjoint. We outline the essential pieces of information PETSc requires from the FEniCS Project, give examples on how to utilize each feature in pFibs, and present some scalability results. Our examples have shown that pFibs is indeed a viable way to solve mixed finite element formulations constructed from the FEniCS Project.

Innokentiy Protasov
University of Houston
iprotasov@uh.edu

Justin Chang
NREL
justin.chang@nrel.gov

Jeffery M. Allen
University of Colorado at Boulder
Jeffery.Allen@nrel.gov

PP1

Kokkos Kernels : A Performance Portable Library for Linear Algebra and Graph Algorithms

Kokkos Kernels is a library in the Kokkos ecosystem for performance portable sparse/dense linear algebra and graph kernels. Both the performance and the portability aspects are key to the stakeholders of the product. The two main focuses of the project are performance portability and delivering robust software to computational science and engineering applications. This poster outlines the key functionality in the library along with performance of newly developed kernels on CPUs and GPUs.

Siva Rajamanickam, Christian Trott, Nathan Ellingwood,
Kyungjoo Kim, Brian Kelley, Vinh Dang
Sandia National Laboratories
srajama@sandia.gov, crtrott@sandia.gov,
ndellin@sandia.gov, kyukim@sandia.gov,
bmkelle@sandia.gov, vqdang@sandia.gov

Luc Berger-Vergiat
Sandia National Labs
lberge@sandia.gov

Jeremiah Wilke
Sandia National Laboratories
jjwilke@sandia.gov

PP1

An Interface for Extreme-Scale Geometric Multigrid in PETSc

Multigrid methods have been effective iterative solvers for linear and nonlinear systems in a wide range of large-scale applications, due to exhibiting a computational complexity that scales linearly (or close to linearly) with respect to the degrees of freedom. While algebraic multigrid methods profit from being relatively simple to incorporate into existing domain-specific programs, they come with challenges regarding parallel scalability. Geometric multigrid methods access domain-specific knowledge directly and thereby avoid certain parallel scalability bottlenecks of algebraic variants. However, the programming effort required by geometric methods can be prohibitive. Since, the PETSc library already provides scalable algorithms for a large community of developers and domain-specialists, it is natural to extend PETSc by geometric multigrid capabilities. We propose a multigrid interface in PETSc based on hybrid spectral–geometric–algebraic multigrid methods that have demonstrated scalability to 1.6M compute cores. We focus on the geometric component utilizing the adaptive mesh refinement library, p4est, where the parallel distribution of adapted meshes is based on space-filling curves and an octree topology; resulting in efficient parallel mesh coarsening, refinement, and partitioning. The interface to geometric multigrid aims to be accessible to domain-specialists while providing parallel scalability to leadership-class computing systems.

Johann Rudi
Argonne National Laboratory
Mathematics and Computer Science
jrudi@anl.gov

PP1

Planning Robot Manipulation Using Parallel-in-Time Integration

Planning actions performed by a robot to manipulate objects requires many rapid solves of systems of ordinary differential equations. In many cases, these planning times can be substantial and thus the robot "freezes", sometimes for minutes, while it "thinks". We explore the feasibility of using parallel-in-time integration algorithms to reduce planning times. Examples will include both simulations and experiments with a real robot pushing an object around an obstacle into a target zone.

Daniel Ruprecht
School of Mechanical Engineering
University of Leeds
daniel.ruprecht@fu-berlin.de

Wisdom Agboh
School of Computing
University of Leeds
scwca@leeds.ac.uk

Oliver Grainger
School of Mechanical Engineering
University of Leeds
mn17omg@leeds.ac.uk

Mehmet Dogar
School of Computing
University of Leeds
m.r.dogar@leeds.ac.uk

PP1

Massively-Parallel Computing of Multi-Channel 2D Convolution

Convolutional neural networks (CNNs) are among the most powerful and widely used algorithms for computer vision applications. The primary computationally-demanding parts of CNNs are the convolutional layers, which convolve multi-channel images with multiple kernels. As a result, the different acceleration techniques for fast and efficient CNN computing are rapidly growing in popularity. In this work, the extreme acceleration technique for massively-parallel fine-grained computing of multi-channel 2D convolutions is proposed and evaluated. This technique uses a 4D processing space to speed-up an implementation of multi-dimensional convolution which is originally represented in the 6D index/iteration space. Unlike other algorithms in the existing CNN accelerators, the proposed algorithm computes the required convolution by parallel execution on each time-step all the permissible multiply-accumulate (MAC) operations combined with a systolic-like massively-parallel data movement to resolve data dependencies. A required by multi-channel convolution data reduction is computed by proper chaining of MAC operations. Our implementation radically reduces the number of time-steps needed to compute a convolutional layer. Under real-time constrains to execute a given CNN model, such deep reduction of the time-steps can be effectively used for lowering operating frequency in the algorithms hardware implementation and, as a consequence, diminishing the power and energy consumption.

Stanislav Sedukhin, Yoichi Tomioka
University of Aizu
sedukhin@u-aizu.ac.jp, ytomioka@u-aizu.ac.jp

PP1

Semi-Structured Hybrid Multigrid on Octree Meshes

We present a multigrid method that combines the flexibility of unstructured octree meshes with the efficiency of structured grid computations. While octrees allow for flexible adaptive mesh refinement, matrix-free finite element computations on non-conforming adaptive meshes require a large number of conditional statements and indirect memory access. This is due to the need to gather contributions from neighboring element at coarse-fine interfaces to shared unknowns. We thus combine non-conforming adaptivity with large regular mesh patches, which reuse data structures generated for high-order elements. We perform several levels of geometric multigrid on these meshes, and use algebraic multigrid as coarse grid solver. We present convergence and scalability results for Poisson's equation and linear elasticity with strongly varying coefficients.

Yu-Hsuan Shih
New York University
yhs264@nyu.edu

Georg Stadler
Courant Institute for Mathematical Sciences
New York University
stadler@cims.nyu.edu

PP1

A Parallel Implementation of Non-Linear Least Squares Method for CP Decomposition for Tensors

The Canonical Polyadic decomposition (CPD) is used to approximate high dimensional data in various fields including Quantum Chemistry and Machine Learning. Alternating Least Squares (ALS) algorithm is usually regarded as the workhorse algorithm for CPD, however, ALS may exhibit slow or no convergence when the CP rank is greater than the dimensions of the Tensor or the factor matrices of CPD have high collinearity. For the above-mentioned scenarios, the Non-linear least squares (NLS) methods are considered superior to ALS but are not scalable for large tensors. Damped Gauss Newton algorithm is a well-known method to solve the NLS problem iteratively, where in each iteration, a system of equations of the approximate Hessian and the gradient of the residual function is solved. We explore parallelism in this method for CPD by realizing the approximate Hessian as tensor operators. This formulation is useful when the system of equations is solved by using Conjugate Gradient algorithm which would require tensor contractions in each iteration, allowing a parallel tensor contraction library to exploit parallelism and enabling NLS to outperform ALS in terms of time and accuracy for computing CPD for large high CP rank tensors.

Navjot Singh
University of Illinois at Urbana Champaign
navjot2@illinois.edu

Hongru Yang
University of Texas at Austin
hy6385@cs.utexas.edu

Linjian Ma
University of Illinois at Urbana Champaign
lma16@illinois.edu

Edgar Solomonik
University of Illinois at Urbana-Champaign
solomon2@illinois.edu

PP1

GPU-Accelerated Barycentric Cluster-Particle Treecodes

We present GPU-accelerated implementations of cluster-particle versions of the kernel independent Barycentric Lagrange Treecode (BLTC) and the weakly kernel dependent Barycentric Hermite Treecode (BHTC). In the cluster-particle form, hierarchical octtrees are built on the target sites instead of the source particles. Particle-cluster treecodes are typically superior to cluster-particle treecodes, except when the number of targets is greater than the number of sources. The simple structure of the cluster-particle approximations in these treecodes allow for

efficient GPU parallelization. We implement the treecodes with OpenACC and OpenMP and test on NVIDIA GPUs. We demonstrate the treecode on several test cases consisting of $10^5 - 10^7$ randomly distributed particles with randomly distributed charges, interacting via the Coulomb and Yukawa kernels. We show strong GPU speedup, both for a single GPU versus a single CPU and for a complete GPU node versus a complete CPU node.

Leighton Wilson, Nathan Vaughn
University of Michigan
lwwilson@umich.edu, njvaughn@umich.edu

Lei Wang
University of Wisconsin, Milwaukee
wang256@uwm.edu

Robert Krasny
Department of Mathematics
University of Michigan
krasny@umich.edu

PP1

Graphblas with Performance Semantics: Specification, Design, and Implementation

We present a hybrid shared- and distributed-memory implementation of the GraphBLAS. Different from other recently published work, we assign performance semantics to each GraphBLAS primitive: the maximum amount of work, intra-process data movement, inter-process communication, and any extra memory required. Furthermore, we define only a small set of primitives is allowed to make system calls such as for memory allocations or I/O. Such performance semantics must balance between allowing a back-end sufficient freedom to choose the best implementation on a given system, while remaining useful enough for an algorithm designer to optimise her end of the design space. We follow by introducing three back-ends that comply to the defined performance semantics: a sequential reference implementation, a shared-memory parallel one using OpenMP, and a distributed-memory parallel one using the Lightweight Parallel Foundations (LPF). The latter requires a secondary back-end for intra-node computations; a composition with the OpenMP back-end thus naturally yields a fully hybrid GraphBLAS implementation. We will demonstrate some of our C++ interface, discuss key implementation choices, and show attained performance of classical graph algorithms on scales varying from phones to clusters. Finally, through LPF and using Spark as an example, we demonstrate how our hybrid back-end transparently integrates with auxiliary parallel Big Data framework in a low-maintenance fashion.

Albert-Jan N. Yzelman, Wijnand Suijlen
Huawei Technologies France
albertjan.yzelman@huawei.com, wijnand.suijlen@huawei.com

PP1

A Multi-GPU Implementation of a Second Order Scheme to Simulate Landslide-Generated Tsunamis

Landslide-HySEA is a model to simulate tsunamis generated by landslides. The evolution of the landslide, the propagation of the generated tsunami and the inundation of the coastal areas are performed in a single code. It uses a second order HLL well-balanced path-conservative MUSCL-Hancock scheme that only requires one time step to evolve in time. The numerical scheme is able to preserve the stationary solutions corresponding to water at rest and it is positive preserving with the usual 0.5 CFL condition. Landslide-HySEA has been implemented in multi-GPU, and a load balancing algorithm is applied taking into account the presence of water and granular material in the domain. It also supports time series, asynchronous writing of the resulting NetCDF files, and resuming a stored simulation. Additional techniques have been implemented to increase the efficiency, such as overlapping MPI communications with GPU computations, and asynchronous memory transfers between CPU and GPU memory. Several test cases are presented along with scaling measures. Acknowledgements: This research has been partially supported by ChEESE project (EU Horizon 2020, grant agreement N 823844).

Marc de la Asuncion
University of Malaga, Spain
marcah@uma.es

Manuel J. Castro
Dpt. Análisis Matemático
University of Málaga
castro@anamat.cie.uma.es

PP2

Ginkgo - a Node-Level Sparse Linear Algebra Library for High Performance Computing

With the rise of manycore accelerators like GPUs, there exists an increasing demand for linear algebra libraries that can efficiently exploit the concurrency and performance available in a single compute node. At the same time, more and more application projects move towards an object-oriented software design based on C++ for both efficiency and ease of use. In the Ginkgo software effort, we design and develop a next-generation sparse linear algebra library able to run on multi- and manycore architectures. The library design is guided by combining ecosystem extensibility with heavy, architecture-specific kernel optimization using the platform-native languages CUDA (NVIDIA GPUs), HIP (AMD GPUs), or OpenMP (Intel/AMD multicore).

Hartwig Anzt
Steinbuch Centre for Computing
Karlsruhe Institute of Technology
hartwig.anzt@kit.edu

Terry Cojean, Thomas Gruetzmacher, Pratik Nayak, Tobias Ribizel, Yushiang Tsai
Karlsruhe Institute of Technology
terry.coean@kit.edu, thomas.gruetzmacher@kit.edu, pratik.nayak@kit.edu, tobias.ribizel@kit.edu, yushiang.tsai@kit.edu

PP2

Solving Hyperbolic PDE Systems with ExaHyPE

We present latest features and results for ExaHyPE, an engine to solve hyperbolic PDE systems. ExaHyPE employs high-order discontinuous Galerkin with ADER time stepping and a-posteriori Finite-Volume limiting. It builds on the Peano framework, which features tree-structured Cartesian meshes and shared- and distributed-memory parallelism using MPI and Intel TBB. ExaHyPE provides role-oriented code generation utilities that offer tailored views

for application, algorithm and optimisation experts. Application experts can quickly realise solver via providing only PDE-specific implementation, such as for conservative fluxes, non-conservative products, source terms, etc. We present several use cases for the engine, partiocularly focusing on computational seismology. Here, we show results for a curvilinear model for seismic wave propagation and for a diffused interface model - both able to treat complex topographies with almost no effort on meshing.

Michael Bader, Leonhard Rannabauer, Anne Reinarz
Technical University of Munich
bader@in.tum.de, rannaba@in.tum.de, reinarz@in.tum.de

Michael Dumbser
University of Trento
michael.dumbser@unitn.it

Alice-Agnes Gabriel
LMU Munich
gabriel@geophysik.uni-muenchen.de

Tobias Weinzierl
Durham University
tobias.weinzierl@durham.ac.uk

PP2

On the Implementation of an FEM FMM Integral Solver for MHD Magnetic Fields

Magnetohydrodynamics (MHD) is the study of the movement of conducting fluids in the presence of electromagnetic fields. MHD phenomena range from industrial processes like aluminum casting to the formation and collapse of stars. We use a velocity-current formulation to model MHD phenomena. Essential to our model is the Biot-Savart Law, which determines magnetic fields due to current densities. Direct numerical integration of the Biot-Savart Law using N quadrature points results in $O(N^2)$ floating point operations to determine the system matrix of the discretized weak formulation. We discuss the implementation of the fast multipole method (FMM) coupled with the finite element method (FEM) to overcome this bottleneck and reduce the operation count to $O(N)$.

K. Daniel Brauss
Francis Marion University
dbrauss@fmarion.edu

PP2

Construction of High-Order Multirate Imex Integrators for Large-Scale Complex Multiphysics Applications

Multiphysics simulations for example in models of the early universe or in climate modeling tend to be large-scale and involve complex dynamics. Part of these complex dynamics can be disparities in temporal scales for different physical processes. In addition to robust and high performing frameworks, such multiphysics simulations can greatly benefit from high-order multirate schemes in terms of accuracy and efficiency. The focus of this study is to construct (the first of their kind) high-order multirate time integrators that evolve the slow dynamics with an implicit-explicit (IMEX) scheme and the fast dynamics with either an implicit or explicit time integrator. These new multirate integrators are designed for problems that have a structure similar to advection-diffusion-reaction systems. The slow dynamics (advection and diffusion) which also tend to be the most communication intensive are evolved with a small number of large time steps. Furthermore, the advection is treated in an explicit fashion and the diffusion in an implicit fashion. The fast dynamics (reactions) are evolved with a smaller time step. Our hypothesis is that our high-order multirate integrators will achieve the same accuracy as other integration schemes faster while taking large time steps for the slow dynamics. Computation and communication costs are therefore greatly reduced for the slow dynamics.

Rujeko Chinomona
Southern Methodist University
rchinomona@smu.edu

Daniel R. Reynolds
Southern Methodist University
Mathematics
reynolds@smu.edu

PP2

High-Performance Implementation of Wavelet Transforms Using SIMD

This poster presents an implementation of wavelet transforms taking advantage of vector registers featured by modern architectures. We are presenting quickly the ideas behind wavelet transforms, and some leads on how they can be useful for high-performance computing. Then we are presenting how data is handled in the matrices and, consequently, in the memory. We present two algorithmic approaches and how register-blocking can minimize data movements while using a reasonable number of registers. We are comparing and analyzing the performance obtained by these approaches and the various options proposed by SIMD instruction sets on two wavelet transforms (Haar and Daubechies 4) on two recent CPUs. We limited ourselves to these two transforms and these two CPUs in order to keep the poster readable. We chose these two transforms because they exhibit relevant and significant algorithmic features (folding with Db4 no folding with Haar). We conclude that the algorithmic optimizations allowed by the SIMD registers and associated instruction sets provide very significant performance improvements.

Camille Coti
LIPN, CNRS UMR 7030, Université Paris 13, Sorbonne Paris C
camille.coti@lipn.univ-paris13.fr

Joel Falcou
LRI, Universite Paris-Sud, Orsay, France
falcou@lri.fr

Basarab Matei
LIPN, CNRS UMR 7030, Université Paris 13, Sorbonne Paris Cit
basarab.matei@lipn.univ-paris13.fr

PP2

Comparison of a Randomized and a Deterministic Low-Rank Approximation Algorithm

Low-rank approximations of large sparse matrices are important in many scientific applications. We compared a deterministic algorithm for computing a truncated LU factorization with tournament pivoting, proposed by Grigori et al. (2018), with the adaptive randomized range finder, proposed by Halko et al. (2011). The latter is a randomized

algorithm which computes an approximate basis for the range of the input matrix. We developed optimized implementations of these algorithms in order to evaluate their accuracy and runtime behavior. Moreover, we compared parallelization properties of randomized low-rank approximations to deterministic algorithms by using shared-memory parallelization. For the adaptive randomized range finder, we implemented a blocked version that utilizes sparse embedding matrices to reduce communication. Moreover, we used the Block Classical Gram-Schmidt algorithm in combination with Tall Skinny QR factorization for orthogonalization, which is known to be communication-optimal. For the randomized algorithm the runtime can be estimated well based on the problem size due to the inevitable use of dense linear algebra, since the resulting low-rank approximation is usually quite dense. In the deterministic algorithm, after each iteration the input matrix is updated by removing the portion approximated so far. This potentially produces fill-in on the matrix used in successive iterations, which can have large impact on the runtime but is hard to predict.

Robert Ernstbrunner
Department of Computer Science
University of Vienna, sterreich
robert.ernstbrunner@univie.ac.at

Viktoria Mayer
University of Vienna
viktoria.mayer@univie.ac.at

Wilfried N. Gansterer
Department of Computer Science
University of Vienna
wilfried.gansterer@univie.ac.at

PP2

SEMANTICMODELS.JL: A Framework for Automatic Composition of Scientific Models Across Domains

Scientific progress comes from adapting and extending models from prior work to address new problems. However, this ideal workflow is difficult primarily because of the informal or inconsistent representation of models. This problem is exacerbated in fields outside of computer science where scientific models are formulated in informal languages. We propose `SemanticModels.jl`, a category theory-based framework for defining meta-modeling tasks such as model augmentation and model selection along with semantic information extraction. We illustrate the major features of `SemanticModels.jl` including representating models as wiring diagrams, extending and composing models with algebraic operations, and generating executable code of the resulting models. These features are demonstrated by constructing a model of mosquito borne illness in humans along with predator-prey dynamics between mosquitos and birds to study the effects of predator species on the control of mosquito borne illnesses. Through the careful application of category theoretic ideas to scientific computing, general patterns in scientific modeling languages and frameworks can be axiomatized and these formalizations can be exploited to improve scientific computing research and development processes.

Micah E. Halter, Kun Cao, James Fairbanks
Georgia Tech Research Institute
micah.halter@gtri.gatech.edu, kun.cao@gtri.gatech.edu, james.fairbanks@gtri.gatech.edu

PP2

Parallelization of DD and QD High-Precision Arithmetic Operations

To reduce rounding errors in floating-point arithmetic, the use of high-precision arithmetic is effective. DD (Double-Double) arithmetic and QD (Quad-Double) arithmetic are simple way to achieve a quasi quadruple precision and a quasi octuple precision arithmetic operations. DD and QD arithmetics use a combination of double precision floating-point number operations only without using any additional hardwares, however enormous computation time is required. Our team developed MuPAT, an open-source interactive Multiple Precision Arithmetic Toolbox for MATLAB using DD and QD arithmetics, and tried to reduce its computation times by using parallel processing. Since the order of computation in DD and QD arithmetics cannot be changed, we considered processing multiple data simultaneously using data-level parallelism. It is applicable to vector operations, matrix-vector operations, and matrix-matrix operations. We reduce computation time for heavier DD and QD arithmetic operations by using AVX2 and OpenMP. This poster shows a comparison of execution time for these parallelized high-precision arithmetics and a possiblity to perform more accurate computation without additional execution time by using parallelization.

Hidehiko Hasegawa
University of Tsukuba, Japan
hasegawa@slis.tsukuba.ac.jp

Hotaka Yagi
Tokyo University of Science
1419521@ed.tus.ac.jp

Emiko Ishiwata
Department of Mathematical Information Science,
Tokyo University of Science, Japan
ishiwata@rs.tus.ac.jp

PP2

Parallelized Dual-Stage Energy Minimization

InSAR time-series derived from space-borne radar satellite data can be effectively used to identify potential geotechnical risks such as landslides and abnormal subsidence. Typically, time-series analysis at full or high resolution relies on phase unwrapping to generate accurate deformation maps over large regions. A popular formulation for unwrapping congruently is based on global optimization under the L1 norm. This formulation has the benefit of being robust to large scale errors at the cost of resource-intensive computation. As the volume of available datasets grows, more potential geotechnical risks can be caught and addressed; however, this large volume of data requires processing, with globally optimal phase unwrapping as a potential bottleneck. Previous work on this subject has focused on solving this globally optimal formulation using primal methods. In this work, we present a scalable alternative by adopting a specialized primal-dual approach, in which the problem is subdivided and parallelized into smaller problems. We show careful experiments and benchmarks to show the efficiency of this approach.

Khaled A. Helal, Bardia Barabadi
University of Victoria
khaledkelnay@uvic.ca, bardiabarabadi@uvic.ca

Matt Gara
3vGeomatics
mgara@3vgeomatics.com

Amirali Baniasadi, Nikitas Dimopoulos
University of Victoria
amirali@ece.uvic.ca, nikitas@ece.uvic.ca

PP2

Multi-Step Communication in Enlarged Krylov Subspace Solvers

Enlarged Krylov subspace methods aim to address the performance constraints of classical Krylov solvers at scale by reducing the number of iterations to convergence and thus the number of synchronization points and steps of communication, overall. This iteration reduction, however, comes at the cost of more on-node computation and higher communication costs per iteration as a result of expanding the Krylov subspace, and thus increasing single vector computations to multiple vector computations. This poster addresses the higher communication costs resulting from larger messages being injected into the network. To mitigate these costs at scale within algebraic multigrid solvers, multi-step communication techniques have been proposed and shown significant speedups over standard communication strategies. Here, we provide an analysis of these multi-step communication techniques within the context of enlarged Krylov subspace solvers, focusing on the performance of multiple vector SpMVs which represent the key computation kernel in these solvers. Our results show that multiple node-aware (multi-step) communication strategies provide significant performance improvements over standard communication strategies independent of the sparsity of the accompanying matrix and number of vectors being communicated.

Shelby Lockhart
University of Illinois at Urbana-Champaign
sll2@illinois.edu

Luke Olson
University of Illinois Urbana-Champaign
lukeo@illinois.edu

PP2

Greedy Algorithms for Neural Network Architecture Optimization

In this project we aim to develop a new algorithm for the optimization of neural network architectures. The motivation at the core of this project is twofold. First, there is a need from domain scientists to easily interpret predictions returned by the statistical model and this tends to be unpractical when neural networks with complex structures are deployed. Secondly, there is a need to facilitate the use of neural networks in situations of compute/memory limitations. We address these demands by identifying a neural network that attains a prescribed accuracy with a minimal complexity. We propose a novel approach that is iterative in nature and that searches over sets of neural networks with incremental complexity. At each iteration the number of hidden layers is fixed and a random search is performed over all the hyper parameters other than the number of hidden layers. This method has appealing properties for algorithmic and computational scalability. Moreover, the random nature of the algorithm at each iteration still guarantees a thorough exploration of the hyper parameter space. Numerical experiments are shown to compare our method with other hyper parameter optimization algorithms. Preliminary results show that our approach can significantly improve the accuracy attained by the selected model as well as it can reduce the computational time to complete the optimization search.

Massimiliano Lupo Pasini
Computational Sciences and Engineering Division
Oak Ridge National Laboratory
lupopasinim@ornl.gov

Junqi Yin
Oak Ridge National Laboratory
yinj@ornl.gov

Ying Wai Li
Los Alamos National Laboratory
yingwaili@lanl.gov

Markus Eisenbach
Oak Ridge National Laboratory
Oak Ridge, TN 37831, USA
eisenbachm@ornl.gov

PP2

PCPATCH: Software for the Topological Construction of Multigrid Relaxation Methods

Effective relaxation methods are necessary for good multigrid convergence. For many equations, standard Jacobi and Gauß–Seidel are inadequate, and more sophisticated space decompositions are required; examples include problems with semidefinite terms or saddle point structure. In this poster we present a unifying software abstraction, PCPATCH, for the topological construction of space decompositions for multigrid relaxation methods. Space decompositions are specified by collecting topological entities in a mesh (such as all vertices or facets) and applying a construction rule (such as taking all degrees of freedom in the cells around each entity). The software is implemented in PETSc and facilitates the elegant expression of a wide range of schemes merely by varying solver options at runtime. In turn, this allows for the very rapid development of fast solvers for difficult problems.

Lawrence Mitchell
Durham University
lawrence.mitchell@durham.ac.uk

Patrick E. Farrell
Mathematical Institute
University of Oxford
patrick.farrell@maths.ox.ac.uk

Matthew G. Knepley
University at Buffalo
Department of Computer Science and Engineering
knepley@gmail.com

Florian Wechsung
New York University
wechsung@cims.nyu.edu

PP2

A Scalable Block Preconditioner for High-Order Hybridized Discontinuous Galerkin Methods Ap-

plied to Incompressible Resistive MHD

We propose a scalable block preconditioning strategy for linear systems arising from high-order hybridized discontinuous Galerkin discretization of incompressible visco-resistive magnetohydrodynamics (MHD) equations. Incompressible resistive MHD presents several challenges in terms of nonlinearity, coupled fluid and magnetic physics, incompressibility constraints in both velocity and magnetic fields to name a few. The preconditioner uses a least squares commutator approximation for the inverse of the Schur complement and algebraic multigrid with GMRES smoother or the multilevel preconditioner for the approximate inverse of the nodal block. For several 2D and 3D transient examples from MHD, including, but not limited to the island coalescence problem at high Lundquist numbers the preconditioner is robust. We show strong and weak scalability of the block preconditioner up to 8192 cores in the Stampede2 supercomputer.

Sriramkrishnan Muralikrishnan
Laboratory for scientific computing and modeling
Paul Scherrer Institute
sriramkrishnan.muralikrishnan@psi.ch

Tan Bui-Thanh
The University of Texas at Austin
tanbui@ices.utexas.edu

John N. Shadid
Sandia National Laboratories
Albuquerque, NM
jnshadi@sandia.gov

PP2

Additively Damped AFAC Variants for High Order Discretisations

Multigrid algorithms are among the best solvers for elliptic partial differential equations. However, small coarse grid problems create a bottleneck that limits multiplicative multigrid's concurrency. Additive multigrid decouples coarse grid solves from those on finer levels. This increases the concurrency but may reduce stability as the number of grid levels grows. Duplicated corrections across levels tend to make iterations overshoot. Instead of damping coarse grids - which harms convergence - we follow up on the idea of AFACx and introduce an additional, level-specific damping equation that approximates corrections from the next coarsest level to reduce overcorrection. Notably, this additional correction space uses smoothed transfer operators. Stability is further increased through BoxMG transfer operators and arbitrary dynamic AMR is achieved through the use of HTMG and FAS. We increase the arithmetic load slightly yet obtain a stabilised additive, i.e. parallelisable multigrid code. Time-to-solution is improved further as we don't assemble the full matrix and instead hold approximations to the matrix within the mesh. The multigrid scheme kicks off with low accuracy stencils guiding the smoother in the right direction, while tasks in the background improve stencil quality. This scheme makes the solver of relevance for complex material configurations or high order discretisations where the construction of proper stencils, in particular in an AMR context, is not cheap.

Charles D. Murray, Tobias Weinzierl
Durham University
c.d.murray@durham.ac.uk, tobias.weinzierl@durham.ac.uk

PP2

Adaptive Domain Decomposition Method for Saddle Point Problem in Matrix Form

Convergence of domain decomposition methods rely heavily on the efficiency of the coarse space used in the second level. The GenEO coarse [Spillane N., Dolean V., Hauret P., Nataf F., Pechstein C. and Scheichl R., Abstract robust coarse spaces for systems of PDEs via generalized eigenproblems in the overlaps, Numer. Math., 2014], [Dolean, V. and Jolivet, P. and Nataf, F., An Introduction to Domain Decomposition Methods: algorithms, theory and parallel implementation, SIAM, 2015] space has been shown to lead to a robust two-level Schwarz preconditioner which scales well over thousands of cores. The robustness is due to its good approximation properties for problems with highly heterogeneous parameters. It is available in the finite element package FreeFem++ [Hecht F., New development in FreeFem++, J. Numer. Math., 2012] and as a standalone library in HPDDM [Jolivet, P. and Nataf, F., HPDDM: High-Performance Unified framework for Domain Decomposition methods, MPI-C++ library, https://github.com/hpddm/hpddm, 2014] as well as a PETSc preconditioner. We introduce here its extension for solving saddle point problems defined as a block two by two matrix. The algorithm does not require any knowledge of the constrained space. We assume that all sub matrices are sparse and that the diagonal blocks are the sum of positive semi definite matrices. This enables the design of adaptive coarse space for DD methods in the spirit of the GenEO method.

Frederic Nataf
Laboratoire J.L. Lions
nataf@ann.jussieu.fr

PP2

A Discrete Element Method using Triangulated Particles

Discrete Element Methods (DEM) simulate the interaction of large numbers of rigid, incompressible objects with each other. Mainstream DEM codes focus on analytical shapes to streamline the identification of contacts between objects. This step dominates the simulation time. We manage to support triangulated particles with a wide variety of sizes in an efficient DEM code due to a combination of several new algorithmic ideas. First, we sieve the objects into a cascade of finer and finer adaptive Cartesian grids, which provide natural cut-off radii to search for collisions. Intelligent inter-grid transfer operations and neighbouring cell searches allow us to reduce the number of geometric checks and to parallelise both within one grid resolution and between meshes. Second, we model objects as hierarchy of shapes (level of detail). Collision tests are done with as low detail and cost as possible before we fall back to high resolution comparisons. Third, we run all comparisons as combination of geometric checks and a distance minimisation algorithm. The latter yields high flop rates yet breaks down for ill-posed geometric constellations. Only in this latter case, we fall back to robust geometric checks, which do not vectorise. Finally, we phrase our calculations in different floating point precisions and gradually step from smaller float memory footprints with high vector concurrency to high precision yet slow calculations.

Peter J. Noble

University of Durham
peter.j.noble@durham.ac.uk

Tobias Weinzierl
Durham University
tobias.weinzierl@durham.ac.uk

PP2

Algorithm-Specific Checkpointing vs. Exact State Reconstruction for the Preconditioned Conjugate Gradient Method

As current and future computing clusters grow in scale, the likelihood of hardware failures increases. This has motivated the development of a number of strategies to enable HPC applications to recover from node failures. We experimentally evaluate different strategies for protecting the Preconditioned Conjugate Gradient (PCG) solver against node failures. The exact state reconstruction (ESR) approach exploits inherent redundancies of the sparse matrix-vector product to store copies of the local search direction on neighbouring nodes (Chen 2011, Pachajoa et al. 2019). After a node failure, this information can be used to reconstruct the lost parts of the solvers state. We compare this approach to two algorithm-specific checkpointing strategies for PCG, a disk-storage-based approach and an in-memory buddy-checkpointing approach. Our experiments investigate the performance of these schemes both for single and multiple overlapping node failures. ESR and in-memory checkpointing are both shown to be very efficient, with the superiority of one or the other mostly depending on the given scenario, and they both outperform the disk-based checkpointing approach by a wide margin.

Christina Pacher, Carlos Pachajoa, Markus Levonyak
University of Vienna
christina.pacher@univie.ac.at,
carlos.pachajoa@univie.ac.at,
markus.levonyak@univie.ac.at

Wilfried N. Gansterer
Department of Computer Science
University of Vienna
wilfried.gansterer@univie.ac.at

PP2

Hardware Agnostic Implementation of Cosmo-Eulag Dynamical Core for Regional Numerical Weather Prediction using GridTools Framework

One of the major computational challenges for the developers of weather prediction models is to efficiently exploit modern supercomputing architectures. This opportunity is particularly attractive due to the higher memory bandwidth and lower total cost of ownership of an energy-efficient cluster. While implementations of particular components of the weather frameworks for manycore architectures were reported decade ago, it is uncommon to exploit these ports in the operational service. A notable example is the COSMO framework for regional weather prediction in Europe. In this case, operational computations have been performed for several years using fat-node GPU solution. This was enabled thanks to the rewritten or adapted codebase using family of DSLs embedded in C++, namely: STELLA and GridTools to compute atmospheric dynamics, as well as OpenACC directives to accelerate physics. To address numerical difficulties influencing operational COSMO high resolution runs over topography of the Alps, novel dynamical core based on EULAG research model for multiscale flows was implemented within the COSMO. To enable efficient COSMO-EULAG computations on GPUs and multicore processors, EULAG components were ported into the recently announced initial stable Gridtools release. Within this poster presentation we document the process of porting of legacy Fortran 2003 codebase to the C++/GridTools, along with software engineering, energetic and performance considerations on CPU and GPU.

Zbigniew P. Piotrowski, Adam Ryczkowski
Institute of Meteorology and Water Management
zbigniew.piotrowski@imgw.pl,
adam.ryczkowski@imgw.pl

PP2

Tasktorrent: A Task-Based Distributed Runtime System

We propose Tasktorrent, a task-based distributed runtime system that emphasizes an asynchronous style of programming. Tasktorrent is lightweight and uses MPI as its backend for communication. Dependencies of tasks in Tasktorrent can be conveniently expressed with a parametrized task graph. We compare Tasktorrent with other runtime systems including Legion, Starpu and multithreaded Lapack on dense Cholesky factorization. We show the scalability of Tasktorrent as a function of number of cores. Tasktorrent delivers excellent performance and scalability while not being intrusive and allowing maximum reuse of the existing code base. Its in particular best suited for porting legacy codes to modern parallel clusters.

Yizhou Qian
Institute for Computational and Mathematical Engineering,
Stanford University
adncat@stanford.edu

Leopold Cambier
Stanford University
Institute for Computational and Mathematical Engineering
lcambier@stanford.edu

Eric Darve
Stanford University
darve@stanford.edu

PP2

EquationBC: Solving a Different PDE on Boundary

Firedrake (https://www.firedrakeproject.org/download.html) is a finite element analysis software for solutions of partial differential equations that takes a high-level mathematical representation of a problem and generates/compiles efficient low-level codes that run in parallel. In this poster presentation we introduce a recently added feature of Firedrake to apply boundary conditions in equation form using *EquationBC* class. Designed to be a generalisation of DirichletBC class in Firedrake for applying Dirichlet boundary conditions, EquationBC class allows users to force degrees of freedom on boundary to satisfy some partial differential equations. This generalisation becomes useful when, e.g., solving problems that have distinct physics in the domain and on the boundary. Examples include ocean wave modelling in which we solve the Navier-Stokes equation in the domain and the wave equation on the boundary. Here, we will illustrate the

concept and the use of EquationBC using simple examples.

Koki Sagiyama
Imperial College London
k.sagiyama@imperial.ac.uk

Lawrence Mitchell
Durham University
lawrence.mitchell@durham.ac.uk

David Ham
Department of Mathematics
Imperial College London
david.ham@imperial.ac.uk

PP2

Stable Automatic Tuning Method for Performance Fluctuation and Evaluation

Automatic tuning (AT) is a technique to search for optimum parameter value settings of a user program. The execution time for a single parameter setting can vary from one run to the next. We call this fluctuation. In our tests, the fluctuations were found to depend on the execution condition of the computing environment. The bad influence of fluctuation on parameter optimization must be mitigated. Thus, our goal is set to consider fluctuations and realize a stable AT. We researched the iterative one-dimensional (1D) search that can search good parameter with a little search cost. The iterative 1D search is based on an approximation function (d-Spline). The d-Spline function was chosen because it flexibly follows the measured data with small calculation cost. To realize a stable AT, we extend the conventional iterative 1D search. The proposed method calculates the improvement rate of the estimated best parameter after each 1D parameter search, and starts the multiple-time measurements of a single parameter value after the estimated best parameter is expected to nearly optimum. Moreover, after the iterative 1D search is ended, it keeps and updates average execution time of each parameter setting, and choses the parameter setting with shortest execution time. In our numerical evaluation, the proposed method significantly reduced the total execution time in comparison with conventional methods. The contents of this poster will be presented at MS75.

Naoto Seki, Toshiki Tabeta, Akihiro Fujii, Teruo Tanaka
Kogakuin University
em18006@ns.kogakuin.ac.jp, em19009@ns.kogakuin.ac.jp, fujii@cc.kogakuin.ac.jp, teru@cc.kogakuin.ac.jp

PP2

Using Quantized Integer in *LU* Factorization with Partial Pivoting

Quantization is a common technique to speed the deep learning inference. It is using integers with a shared scalar to represent a set of equally spaced numbers. The quantized integer method has shown great success in compressing the deep learning models, reducing the computation cost without losing too much accuracy. New application specific hardware and specialized CPU extension instructions like `Intel AVX-512 VNNI` are providing capabilities for us to do integer `MADD` (multiply and add) efficiently. In this poster, we would like to show our preliminary results of using quantization integers for *LU* factorization with partial pivoting. Using `Int32`, the backward error can outperform single precision. However, quantized integer has the similar issue of limited range as `FP16` that it would not work directly for large matrices because of big numbers would occur in factored U. We will show some possible solutions to it and how we would like to apply this quantized integer technique to other numerical linear algebra applications.

Yaohung M. Tsai
Innovative Computing Laboratory
University of Tennessee
ytsai2@icl.utk.edu

Piotr Luszczek
University of Tennessee
luszczek@icl.utk.edu

Jack Dongarra
Innovative Computing Laboratory
University of Tennessee
dongarra@cs.utk.edu

PP2

Gemslr: a Multilevel Low-Rank Preconditioning and Solution Package

This poster summarizes the development and implementation of GeMSLR (Generalized Multilevel Schur complement Low-Rank), a distributed-memory preconditioner for the solution of large and sparse (non)symmetric linear systems of equations. The GeMSLR preconditioner is purely algebraic and is based on a multilevel reordering of the original set of equations/variables. The reordering is implemented by hierarchically ordering the interface degrees of freedom at each level and several reordering schemes are available. At each given level, GeMSLR decouples the solution of the current linear system into one associated with the interior variables and another associated with the interface ones. The first subproblem is block-diagonal and solved in parallel by applying some form of ILU preconditioning. The recursive nature of the preconditioner appears on the second subproblem where the Schur complement linear system is preconditioned by the interface coupling matrix. The latter is applied by descending to the next level until the last level is reached. In the latter case, the user can choose to use either Block Jacobi acceleration or redundantly solve the problem by (I)LU. Low-rank correction terms can be added at each level to further enhance robustness, and these are applied using the Woodbury formula. GeMSLR is implemented in MPI. We demonstrate the potential of GeMSLR by presenting numerical tests performed on several 2D and 3D problems, and both strong and weak scaling is discussed.

Tianshi Xu
Department of Computer Science and Engineering
University of Minnesota, Twin Cities
xuxx1180@umn.edu

Vasilis Kalantzis
Thomas J. Watson Research Center
IBM Research
vkal@ibm.com

Geoffrey Dillon
Department of Mathematics
University of South Carolina
dillong@mailbox.sc.edu

Yuanzhe Xi
University of Minnesota

yxi26@emory.edu

Ruipeng Li
Center for Applied Scientific Computing
Lawrence Livermore National Laboratory
li50@llnl.gov

Yousef Saad
University of Minnesota
saad@umn.edu

PP2

Optimization of GPU Kernels for Sparse Matrix Computations in HYPRE

The acceleration of sparse matrix computations on GPUs can significantly enhance the performance of iterative methods for solving linear systems. In this work, we consider the kernels of Sparse Matrix Vector Multiplications (SpMV), Sparse Triangular Matrix Solves (SpTrSv) and Sparse Matrix Matrix Multiplications (SpMM), which are often demanded by Algebraic Multigrid (AMG) solvers. With the CUDA and the hardware support of the Volta GPUs on Sierra, the existing kernels should be further optimized to fully take the advantage of the new hardware, and the optimizations have shown significant performance improvement. The presented kernels have been put in HYPRE for solving large scale linear systems on HPC equipped with GPUs. These shared-memory kernels for single GPU are the building blocks of distributed matrix operations required by the solver across multiple GPUs and compute nodes. The implementations of these kernels in Hypre and the code optimizations will be discussed.

Chaoyu Zhang
Arkansas State University
chaoyu.zhang@smail.astate.edu

Ruipeng Li
Center for Applied Scientific Computing
Lawrence Livermore National Laboratory
li50@llnl.gov

PP2

Proxy App for Mesh Relaxation via Machine Learning on Advanced Hardware

With the doubling of transistors no longer a reality for CPUs, the advent of application specific hardware is becoming the new standard. As it is unreasonable to port large codes to new hardware with no promise of performance gain, the idea of a mini/proxy app is essential. Numerous mini/proxy apps have been developed that capture the approximate behavior of specific portions of the code essential for radiation transport and hydrodynamics. Whether that be memory impact specifically, say for cross section retrieval (Quicksilver or XSBench), or the behavior of a full hydrocode, but on a much smaller scale (LULESH). As machine learning (ML) comes into vogue, along with hardware specific to its application, the development of a mini/proxy app capturing its potential changes to program performance becomes necessary to evaluate memory, accuracy, and runtime needs of an application. This work presents early results developed to detect mesh tangling in hydrodynamic simulations through the use of machine learning. Portability is the key focus of this application, as such, performance on GPUs, as well as new ML-specific hardware is presented.

Kristofer Zieb
Lawrence Livermore National Laboratory
zieb1@llnl.gov

Notes

Organizer and Speaker Index

Italicized names indicate session organizers

Italicized names indicate session organizers

Italicized names indicate session organizers

T

V

W

X

Y

Z

PP20 Conference Budget

Conference Budget

SIAM Conference on Parallel Processing
February 12 - 15, 2020
Seattle, WA

Expected Paid Attendance: 400

Revenue		
Registration Income		$144,600
	Total	$144,600
Expenses		
Printing		$1,100
Organizing Committee		$3,800
Invited Speakers		$9,000
Food and Beverage		$27,000
AV Equipment and Telecommunication		$25,000
Advertising		$11,600
Proceedings		$10,000
Conference Labor (including benefits)		$57,758
Other (supplies, staff travel, freight, misc.)		$14,200
Administrative		$18,292
Accounting/Distribution & Shipping		$9,959
Information Systems		$19,840
Customer Service		$6,169
Marketing		$11,749
Office Space (Building)		$6,626
Other SIAM Services		$6,922
	Total	$239,015
Net Conference Expense		-$94,415
Support Provided by SIAM		$94,415
		$0

Estimated Support for Travel Awards not included above:

Early Career and Students	24	$19,750

Hyatt Regency Seattle - Floor Plan