Science at Extreme Scales: Where Big Data Meets Large-Scale Computing

Frank Jenko
jenko@ipp.mpg.de

Director, Max Planck Institute for Plasma Physics
Professor, The University of Texas at Austin
Professor, Technical University of Munich
Two recent waves of innovations affecting science (= main drivers of the expansion of the role of the mathematical sciences\textsuperscript{1}): High Performance Computing & Big Data

Currently, these themes are usually addressed rather independently – but they are intrinsically linked:

- **HPC needs Big Data** for dealing with increasingly large data sets
  - Communication bottleneck on the path to exascale computing
  - Develop novel ways of representing, reducing, reconstructing, and transferring huge amounts of data (*need new algorithms!*)

- **Big Data needs HPC** for analyzing increasingly large data sets
  - Data analytics becomes ever more compute-intensive

\textsuperscript{1}emphasized by the NRC
Only together can they pave the road towards a “predictive science.”

The fusion of HPC and Big Data is a new, emerging field with an endless number of potential applications and an enormous game changer potential.

The present Long Program aims at being a catalyst at this exciting frontier of science by bringing together leading innovators and pioneers from:

- Applied Mathematics & Statistics
- Computer Science & Large-Scale Computing
- Machine Learning & Big Data
- Domain Sciences
Organizing Committee

Frank Jenko, IPP/UT/TUM
Computational Plasma Physics & HPC

Hans Bungartz, TUM
CS & Applied Math

Tandy Warnow, UIUC
CS & Bioengineering

Joachim Buhmann, ETHZ
Machine Learning

Jeff Hittinger, LLNL
Applied Math

Claudia Draxl, HUB
Computational Materials Science

David Keyes, KAUST
Applied Math & HPC

Emmanuel Candès, Stanford
Mathematics and Statistics

Alan Lee, AMD
Corporate VP
Four Workshops

One theme, but looked at from four different perspectives...

Two workshops (WS1 & WS4) will be methods-based, emphasizing recent developments in mathematics & computer science regarding computing and data analytics (together).

Two additional workshops (WS2 & WS3) will be centered around (traditionally) compute-intense or data-driven application areas as they start to explore the complementary side.
Workshop I

Topic: Big Data Meets Large-Scale Computing

This workshop will bring together analysts and developers of data and computationally intensive applications interested in early exploitation of extreme-scale computing platforms to define common ground and seek new opportunities.

Examples of topics that will be discussed:
- requirements / relations of high-performance analytics and simulation
- scalable hierarchical algorithms for analytics and simulation
- detecting and exploiting data sparsity within large-scale data sets
- open problems, where no scalable methods yet exist

Organizing Committee: Keyes, Bungartz, Candès, Meila, Johnson
Classical HPC applications – usually based on numerically solving ODEs/PDEs – develop towards a data-centric approach. This includes:

- patient-specific simulations in medicine
- integrated analytics of experimental and simulation data in plasma physics
- learning from simulation data in materials science

Similar developments take place in many other domain sciences – including astrophysics & cosmology, weather prediction, climate research, and biophysics – and shall be explored in the present workshop.

We will discuss the question: What are the requirements, implications, opportunities, and limitations in this context?

Organizing Committee: Draxl, Jenko, Biros, Müller, Balaban
Typical data analytics applications, which are usually based much more on a statistical (or discrete) apparatus than on numerical computations, will develop in a direction with much more HPC relevance than today. This includes, in particular, bioinformatics and social sciences.

The computational challenges arising in this context go far beyond the “embarrassingly parallel” paradigm and will require more HPC topics to be addressed in large-scale data analytics.

As in Workshop II, but now starting from the Big Data perspective, we will discuss the question: What are the requirements, implications, opportunities, and limitations in this context?

Organizing Committee: Warnow, Buhmann, Chayes, Kumar, LeCun
Workshop IV

Topic: New Architectures and Algorithms

Physical limitations and consumer-driven markets are leading to disruptive changes in computer architectures (even in the near term):

- more on-node parallelism provided by lightweight cores
- more complex and deeper memory hierarchies

New architectures call for new algorithms; active research areas include:

- communication-avoiding algorithms
- data compression and variable precision
- multi-level iterative techniques
- randomized and asynchronous algorithms
- integration of data analysis with simulation

We will explore the nexus of algorithms, architectures, Big Data, and HPC.

Organizing Committee: Hittinger, Lee, Krause, Brown, Balcan
The ambitious goal of this Long Program is to foster – in a tangible and significant way – the “convergence” of Big Data and HPC.

This is also a response to a call by the participants of 6 workshops since 2013 on *Big Data and Extreme-Scale Computing (BDEC)*, supported, e.g., by the science agencies of the G-8 countries (*www.exascale.org*).

The relevant communities are currently compartmentalized. To make progress, applied mathematicians, statisticians, computer scientists, and domain scientists need to engage in an intense dialogue.

During this Long Program, we intend to foster fruitful conversations across a wide range of disciplines, also pulling into the view of the mathematics community a topic of groundbreaking potential.