IPAM / KI-Net Workshop

January 30 – February 3, 2017

Big Data Meets Computation

Predictive Plasma Simulations at the Exascale: Scalable Algorithms for HPC and Data Analytics

Frank Jenko

University of California, Los Angeles
Max Planck Institute for Plasma Physics

I greatfully acknowledge inspiring discussions with
Stan Osher, Russ Caflisch, Jeff Hittinger, and Hans Bungartz

Scope of this workshop (I)

Two recent waves of innovations affecting science (= main drivers of the expansion of the role of the mathematical sciences[1]):

High Performance Computing & Big Data

[1]emphasized by the NRC

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

Scope of this workshop (II)

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 Workshop 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

Timeliness of this workshop

The ambitious goal of this Workshop is to foster the "convergence" of Big Data and HPC.

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

Basic idea: We must begin to systematically map out and account for the ways in which the major issues associated with Big Data intersect with, impinge upon, and potentially change the international plans that are now being laid for achieving exascale computing.

Comparing "Numerically Intensive" and "Data Intensive" High Performance Computing

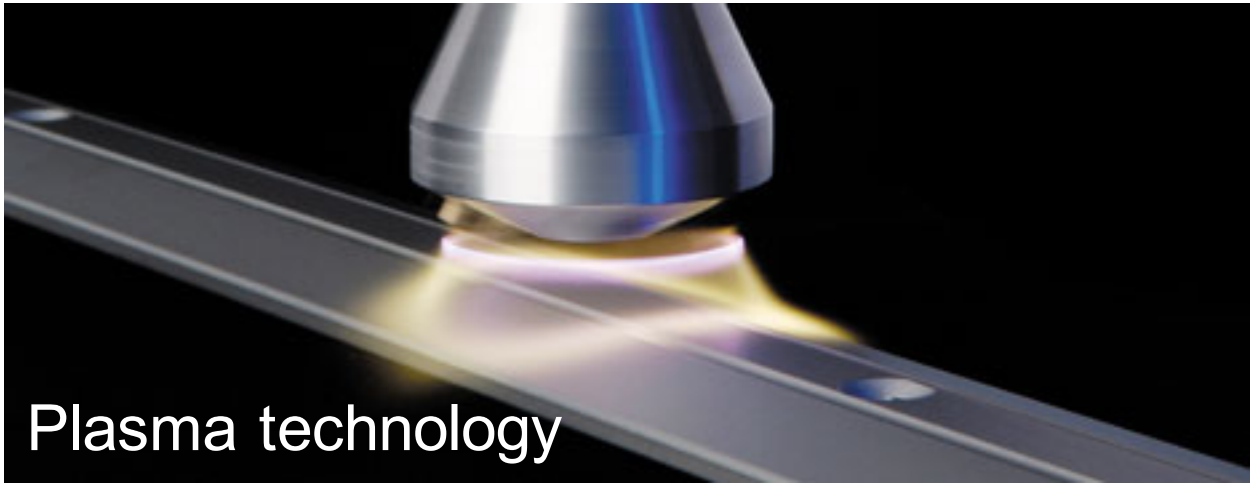
- Both numerically intensive (NI) and data intensive (DI) approaches share the common challenge of gaining scientific insights, making prediction, and quantifying uncertainty
  - NI primarily through first principles models
  - DI primarily through statistical models
- Disclaimer: these labels are imperfect; the right labels are a "work in progress"

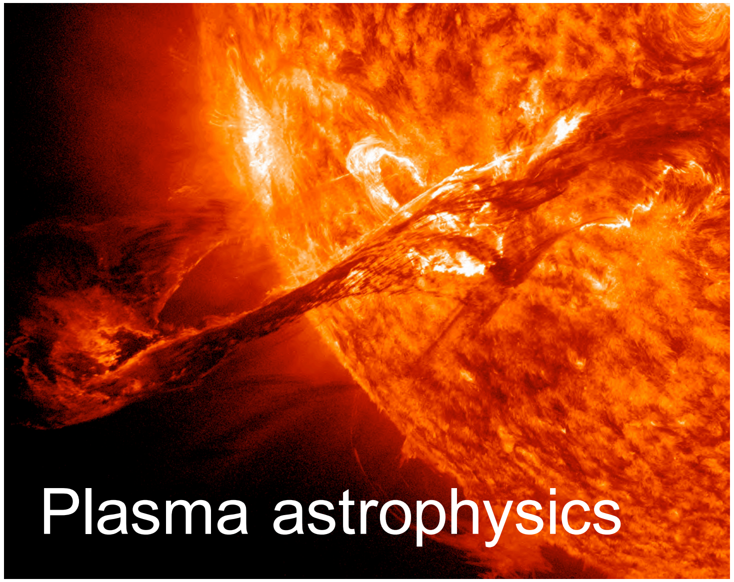
D. Keyes

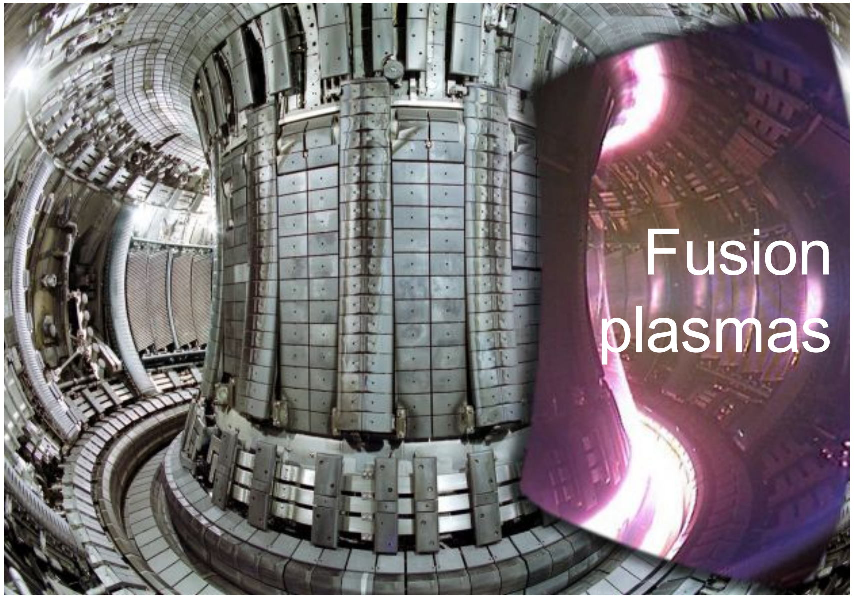
“Big Data Meets Computation”

From the Perspective of Plasma Physics

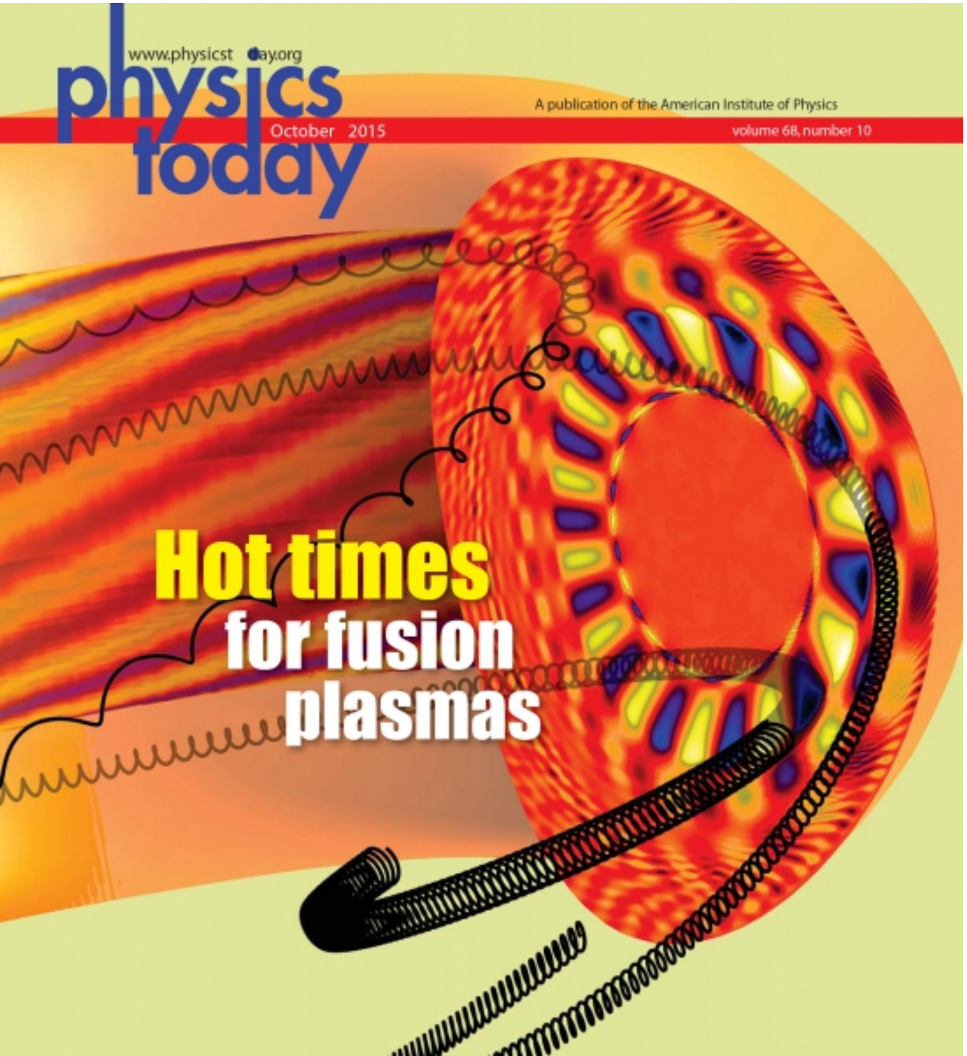
Our plasma universe

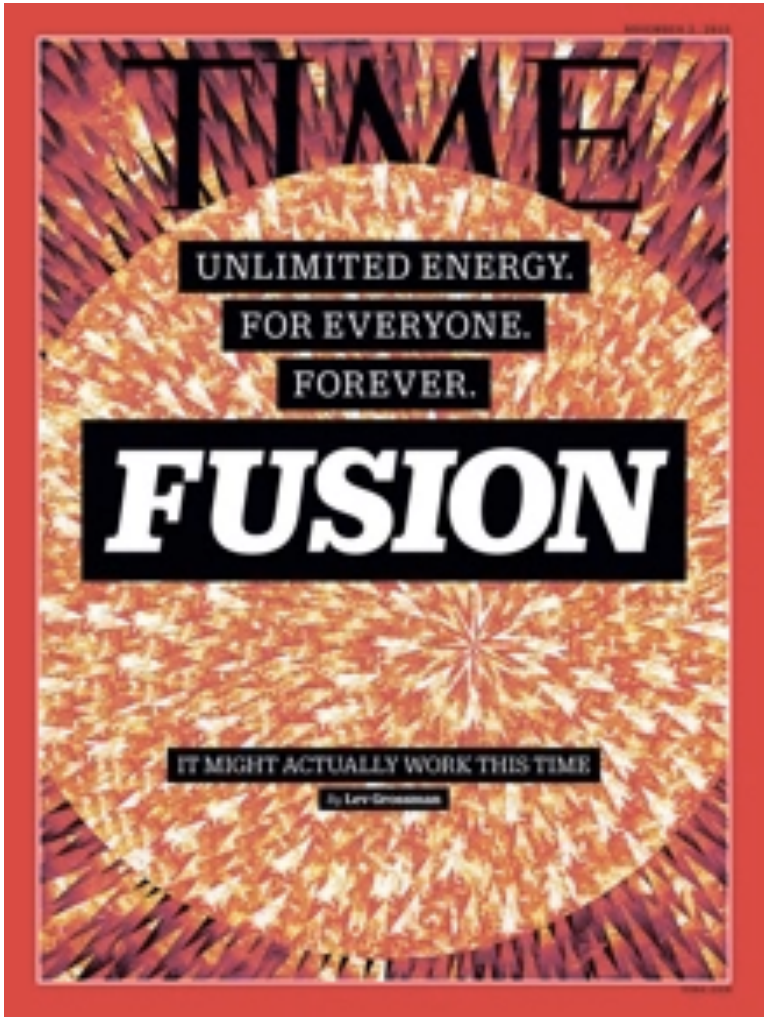




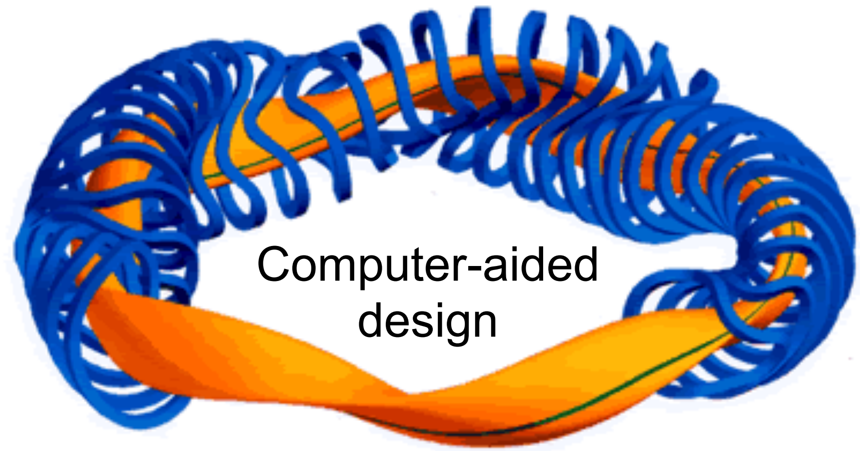


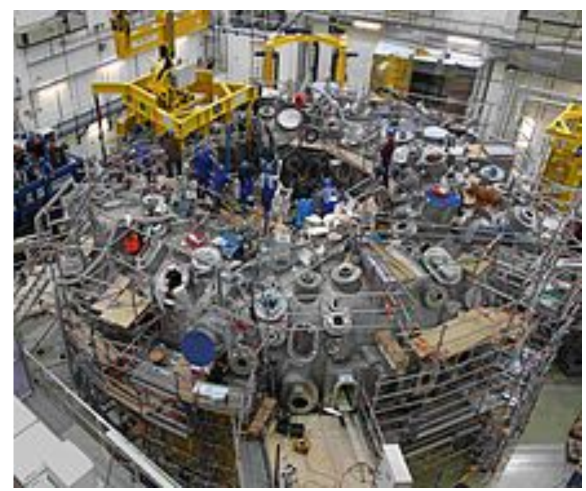
Fusion energy research: Sun in a (magnetic) bottle



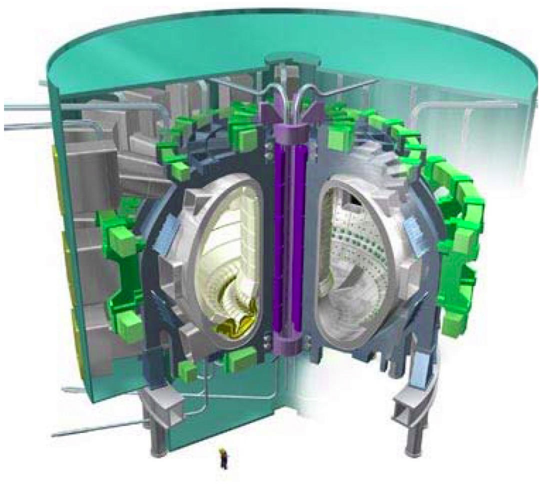


Chancellor Angela Merkel starts up Wendelstein 7-X on February 3, 2016



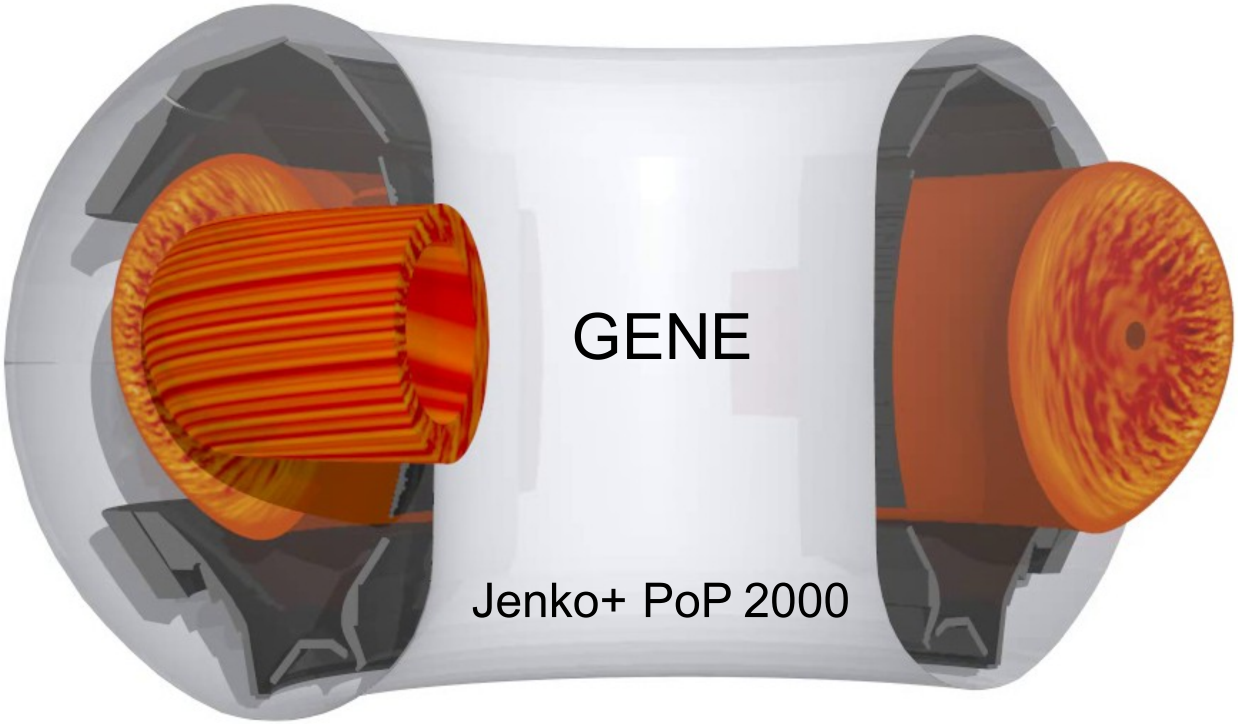


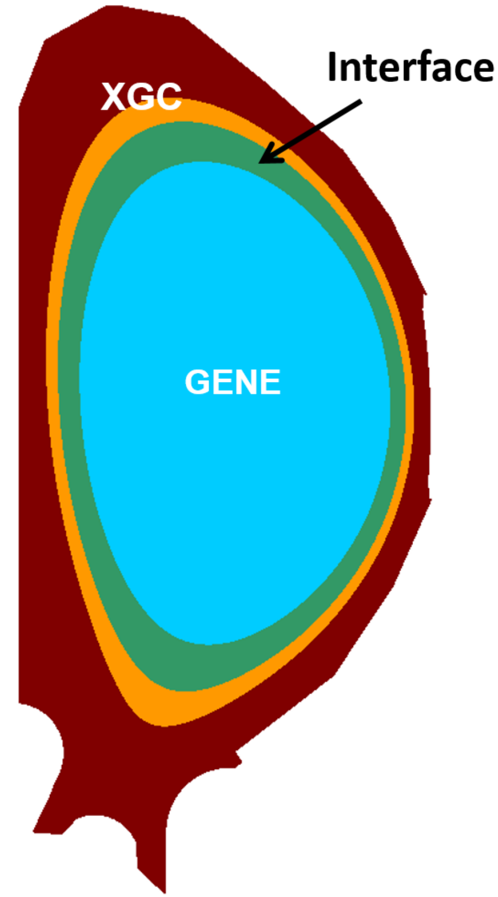
ITER construction site (a global project)



Towards whole-device modeling at exascale

4-year, $10M project (since 10/1/2016)





ECP
EXASCALE COMPUTING PROJECT

Overall goal:
From post-diction to prediction

Many multi-scale, multi-physics problems to solve

Pre-exascale computations with GENE

- Developed by an international team of physicists, applied mathematicians, and computer scientists led by FJ
- Comprehensive physics & flexible geometries (unique feature; ranging from flux-tube tokamaks to full-torus stellarators…)
- Open source: http://genecode.org
- World-wide user base from ~40 scientific institutions (including all U.S. labs and major research universities active in fusion research)
- Output to date: 150+ papers (20+ PRLs)
- Scales well on many leading HPC systems

GENE on top-level HPC resources

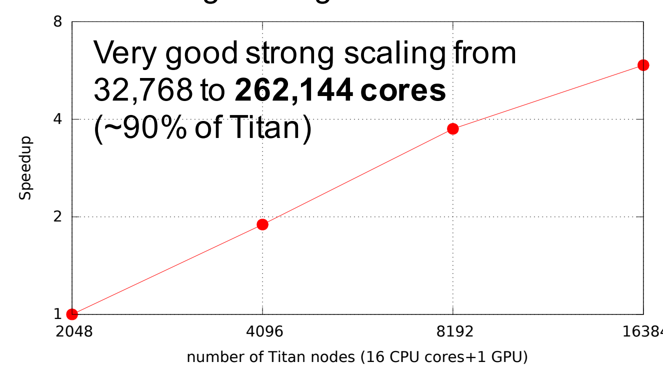
Ranked #1 out of 68 proposals in PRACE Early Access Call (2010)

PRACE

U.S. DEPARTMENT OF ENERGY INCITE LEADERSHIP COMPUTING

First grid-based (gyro-)kinetic code to receive an INCITE Award (2016)

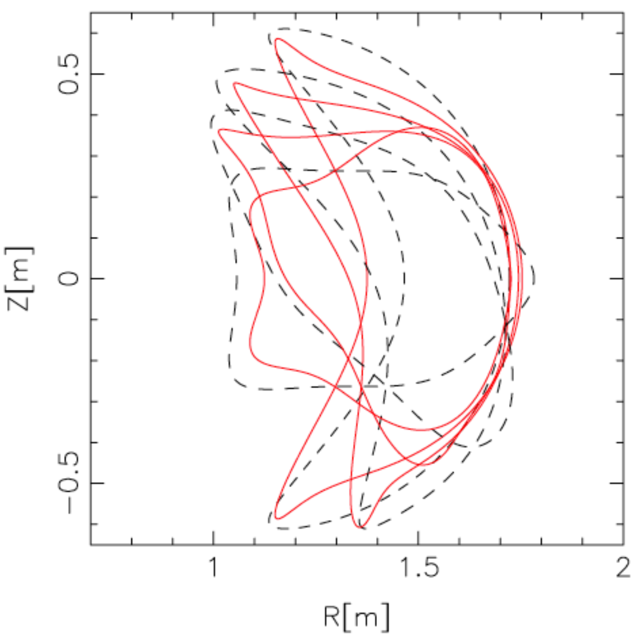
Strong scaling of GENE on Titan



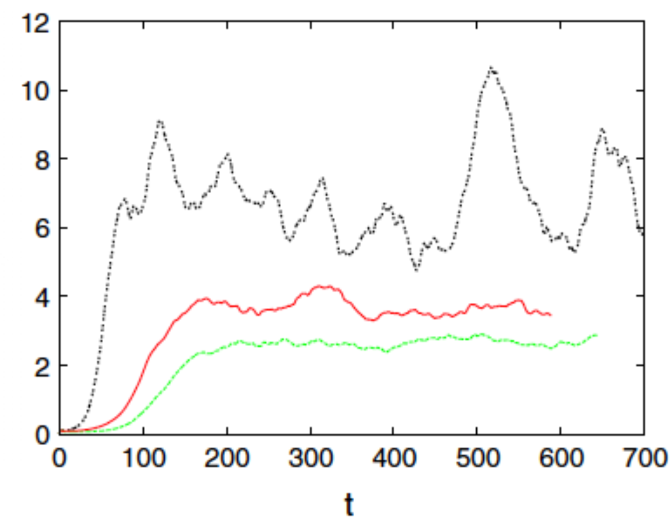
Gyrokinetic Electromagnetic Numerical Experiment

Optimized design of fusion experiments

Proof-of-principle: NCSX-like geometries, optimized for turbulent transport via the generation of successive variations of magnetohydrodynamic equilibria, using simple “cost functions” and *ab initio* plasma turbulence simulations



PRL **105**, 095004 (2010)



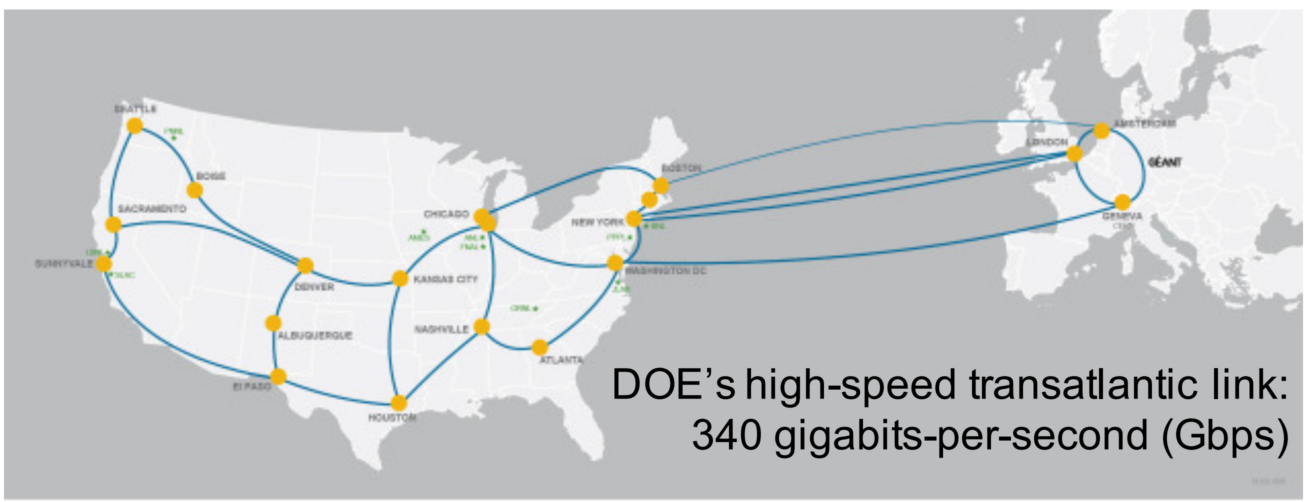
For many other aspects, see: http://irfm.cea.fr/TMFDPVA15

Big Data Meets Computation: Some Frontiers

Handling large amounts of scientific data

Scientific data from experiments and simulations

- Present-day datasets can easily be in the TB…PB range
- This number will continue to grow rapidly
- Need to develop novel ways of representing, reducing, reconstructing, and transferring large datasets



ESnet
ENERGY SCIENCES NETWORK

Sunway TaihuLight

- ~10M cores
- ~0.1 Eflop/s
- ~1 PB memory

Scalable lossy compression of scientific data

Data compression must be ***lossy*** and ***scalable***

Treatment of ***scientific*** data is still in its infancy

Compression factors of key compressors (Cappello & Di 2016)

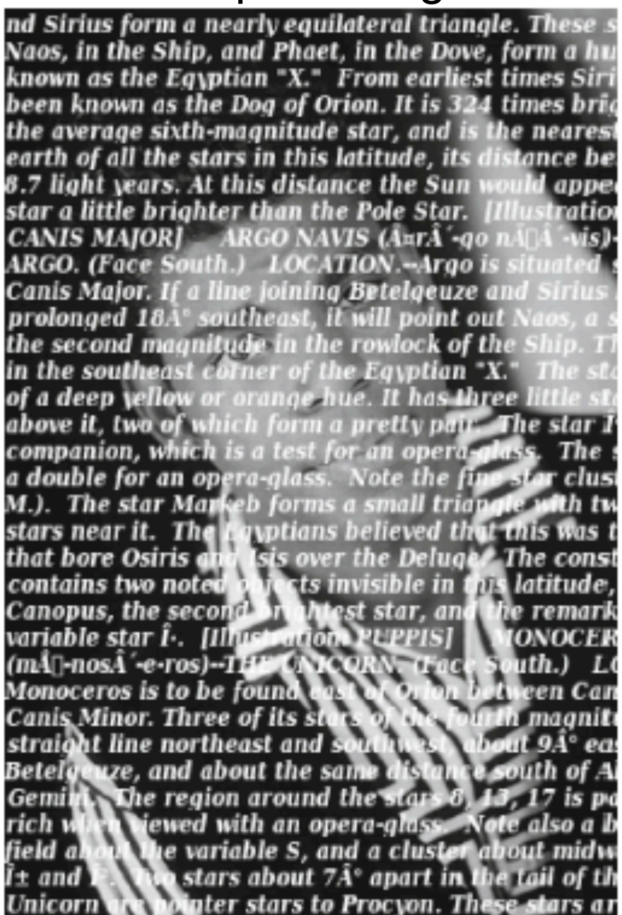
Benchmark	SZ	ZFP	ZFP+Gzip	ISA	ISA+Gzip	SSEM[a]	FPZIP-40[b]	Gzip	FPC[c]
Blast2	110	6.48	36.2	4.56	46.2	39.7	22.9	77	11.4
Sedov	7.44	4.42	5.47	4.42	7.44	17	3.43	3.13	1.9
BlastBS	3.26	3.48	3.65	4.43	5.06	8.45	2.43	1.24	1.29
Eddy	8.13	2.5	2.61	4.34	5.18	N/A	2.56	5.5	3.89
Vortex	13.6	4.45	4.77	4.43	4.72	12	3.35	2.23	2.34
BrioWu	71.2	8.1	43.4	5	57.4	35.7	21.9	73	8.5
GALLEX	183.6	36.7	92.7	4.89	33.6	82.4	20.35	34.7	11.37
MacLaurin	116	10.2	14	4.1	5.47	7.44	3.84	2.03	2.08
Orbit	433	31.7	89	4.96	8.43	11.7	3.9	1.8	1.86
ShafranovShock	48	3.68	8.75	4.24	12.2	20.3	19.9	28	7.33
CICE	5.43	2.11	2.16	4.19	4.46	3.83	2.3	2.6	2.67
ATM	3.95	2.3	2.75	3.1	3.7	1.82	1.04	1.36	N/A
Hurricane	1.63	1.19	1.2	2.57	2.65	1.11	2.07	1.16	N/A

Carry out data analytics directly on compressed data?

Fast decompression of scientific data

Inpainting with Deep Neural Networks (Köhler+ 2014)

corrupted image



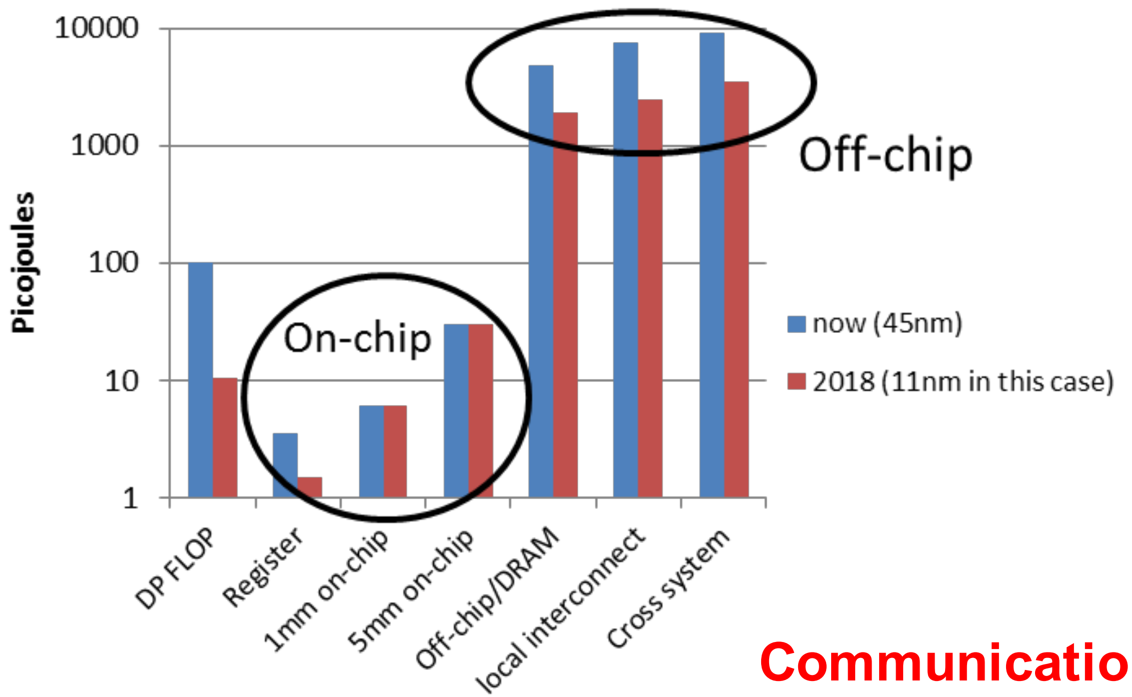
reconstruction

Key question:

Can similar techniques be applied to *scientific* data?

Minimizing data motion in simulations

The energy required to **move data around** accounts for a significant portion of the energy consumption of modern supercomputers.



Traditional approach:

Minimize #operations

Pre-exascale era:

Minimize data motion

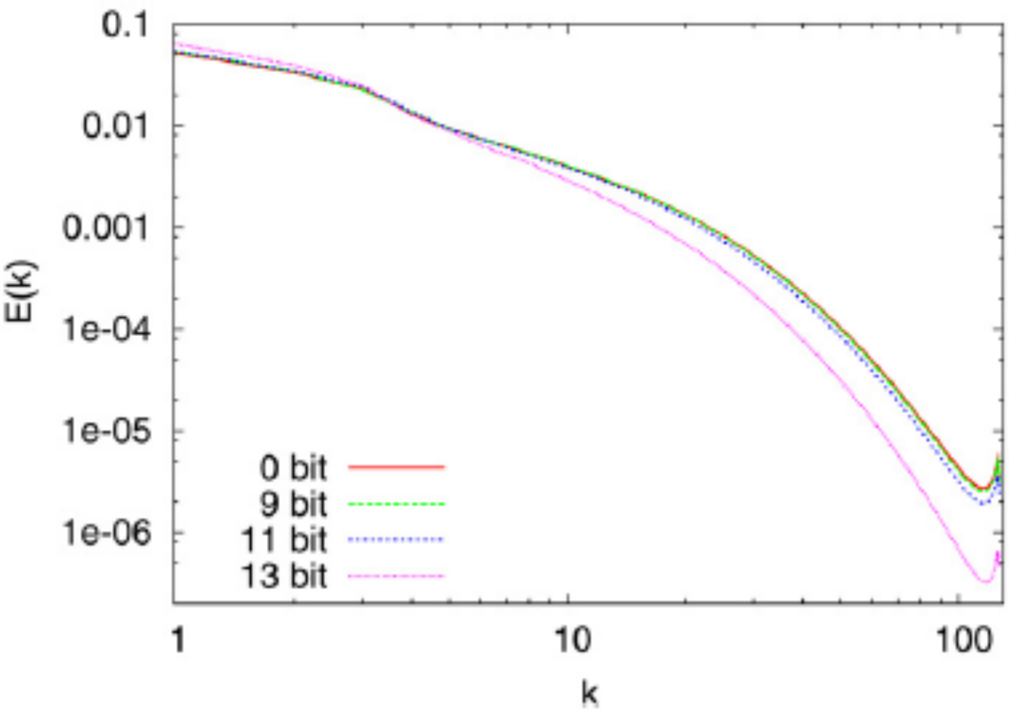
Communication-avoiding algorithms

Working with variable precision

Impact of the floating-point precision and interpolation scheme on the results of DNS of turbulence by pseudo-spectral codes

Holger Homann, Jürgen Dreher, Rainer Grauer *

CPC 2007



Example:

Turbulent energy spectra for simulations w/ single precision reduced by several bits; *statistical* properties tend to be pretty robust

Approximate computing

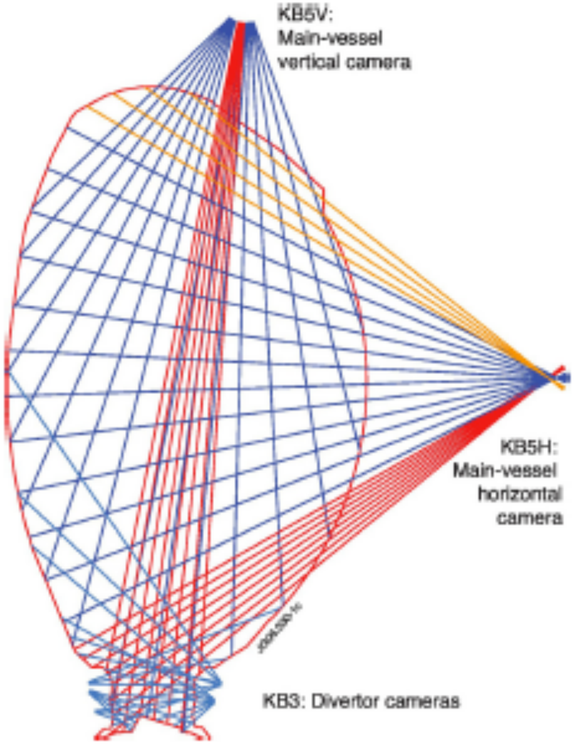
Balancing accuracy and efficiency (closely related to *resilience*)

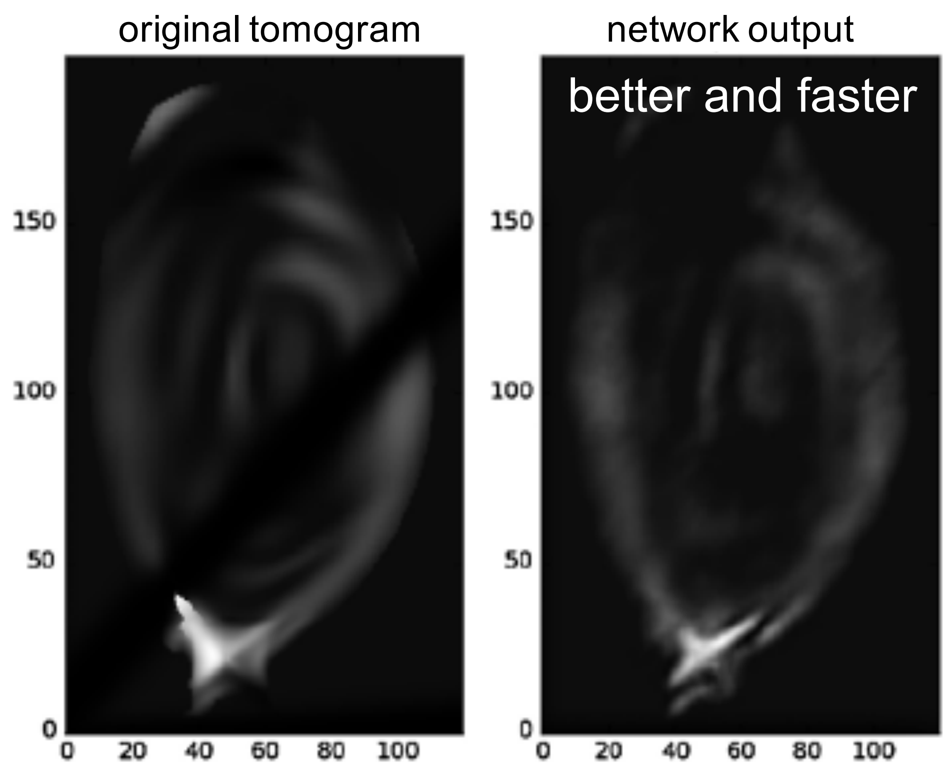
Quality metric(s)	Corresponding applications/kernels
Relative difference/ error from standard output	Fluidanimate, blackscholes, swaptions (PARSEC), Barnes, water, Cholesky, LU (Splash2), vpr, parser (SPEC2000), Monte Carlo, sparse matrix multiplication, Jacobi, discrete Fourier transform, MapReduce programs (e.g., page rank, page length, project popularity, and so forth), forward/inverse kinematics for 2-joint arm, Newton-Raphson method for finding roots of a cubic polynomial, n-body simulation, adder, FIR filter, conjugate gradient
PSNR and SSIM	H.264 (SPEC2006), x264 (PARSEC), MPEG, JPEG, rayshade, image resizer, image smoothing, OpenGL games (e.g., Doom 3)
Pixel difference	Bodytrack (PARSEC), eon (SPEC2000), raytracer (Splash2), particle filter (Rodinia), volume rendering, Gaussian smoothing, mean filter, dynamic range compression, edge detection, raster image manipulation
Energy conservation across scenes	Physics-based simulation (e.g., collision detection, constraint solving)
Classification/clustering accuracy	Ferret, streamcluster (PARSEC), k-nearest neighbor, k-means clustering, generalized learning vector quantization (GLVQ), MLP, convolutional neural networks, support vector machines, digit classification
Correct/incorrect decisions	Image binarization, jmeint (triangle intersection detection), ZXing (visual bar code recognizer), finding Julia set fractals, jMonkeyEngine (game engine)
Ratio of error of initial and final guess	3D variable coefficient Helmholtz equation, image compression, 2D Poisson's equation, preconditioned iterative solver
Ranking accuracy	Bing search, supervised semantic indexing (SSI) document search

S. Mittal

Inverse problems and deep learning

Plasma tomography: Use DNNs to reconstruct cross-section from projections





Deep learning for plasma tomography using the bolometer system at JET

Francisco A. Matos[a], Diogo R. Ferreira[a,*], Pedro J. Carvalho[b], JET Contributors[1]

Real-time modeling via neural networks

Letter Nucl. Fusion 2015

Real-time capable first principle based modelling of tokamak turbulent transport

J. Citrin[1,2], S. Breton[2], F. Felici[3], F. Imbeaux[2], T. Aniel[2], J.F. Artaud[2],
B. Baiocchi[4], C. Bourdelle[2], Y. Camenen[5] and J. Garcia[2]

Nonlinear multivariate regression of simulation data with a NN

- Proof-of-principle: input layer size N=5; 2 hidden layers of 40 neurons each
- ~5 orders of magnitude faster than conventional (reduced) transport models
- Simulates a 300 s ITER discharge in ~10 s
- First-principles based simulations would require ~$10^{8\text{-}9}$ core-hours
- In practice, training set size limits N to **N_{lim} ~ 10**; use experimental data

Modeling of transport phenomena in tokamak plasmas with neural networks

O. Meneghini,[1,a)] C. J. Luna,[2] S. P. Smith,[3] and L. L. Lao[3]
[1]*Oak Ridge Associated Universities, 120 Badger Ave, Oak Ridge, Tennessee 37830, USA*
[2]*Arizona State University, 411 N. Central Ave, Phoenix, Arizona 85004, USA*
[3]*General Atomics, San Diego, California 92186-5608, USA*

Phys. Plasmas 2014

Based on experimental data

N~20; 3 hidden layers of 30 neurons each

Real-time control via neural networks

Most critical problem for MFE: *avoid/mitigate large-scale major disruptions*

- Approach: Use of big-data-driven statistical/machine-learning predictions for the occurrence of disruptions in JET
- Current Status: ~ 6+ years of R&D results (led by JET) using SVM-based ML on *zero-D* time trace data executed on modern clusters yielding ~ *reported success rates ranging from 80 up to 90% for JET, BUT > 98% with false alarm rate < 3% actually needed for ITER* (Reference – P. DeVries, et al., June 2015)
- **PPPL Team Goals include:**

(i) improve physics fidelity via development of new *ML multi-D, time-dependent software including better classifiers;*

(ii) develop *"portable"* predictive software beyond JET to other devices and eventually ITER; and

(iii) enhance execution speed of disruption analysis for very large datasets via *development & deployment of advanced ML software via SVM (Support Vector Machine) & DRNN (Deep Recurrent Neural Network) methods*

W. Tang

IPAM Long Program (Fall 2018)

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

Organizing committee

Frank Jenko, UCLA/IPP
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

Bridging scientific fields

Applied Mathematics & Statistics

Computer Science & Large-Scale Computing

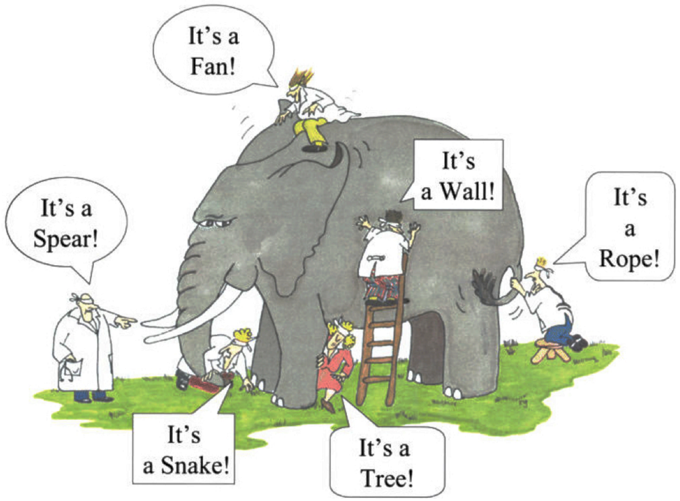
Machine Learning & Big Data

Domain Sciences

- Materials Science
- Astrophysics & Cosmology
- Plasma Physics
- High-Energy Physics
- Weather & Climate
- Geosciences
- Biophysics
- Bioinformatics & Genomics
- …

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

Workshop II

Topic: **HPC-Driven Applications Go Big Data**

Classical HPC applications – usually based on numerically solving ODEs/PDEs – develop towards a data-centric approach.

This includes:

- patient-specific simulations in medicine
- data analytics of experimental/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?

Workshop III

Topic: **Big-Data-Driven Applications Go HPC**

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?

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.

Some key questions regarding BG & HPC

How to handle large *scientific* datasets from experiments and / or simulations?

How to find an optimal balance between accuracy and efficiency in large-scale simulations?

How to apply ML techniques to equation-based sciences?