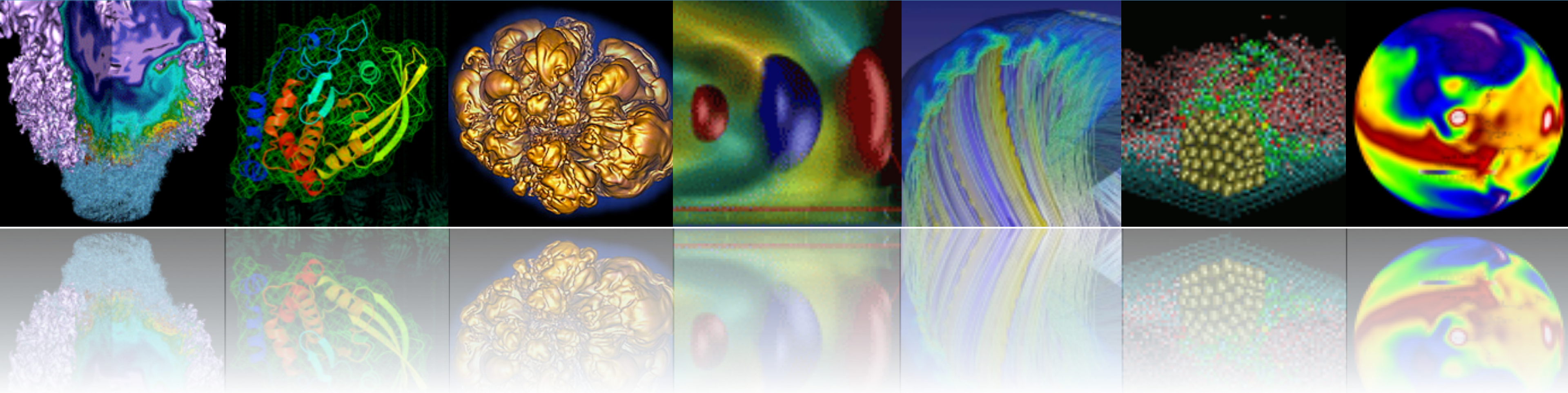


Big Data and Exascale
A Tale of Two Ecosystems

Kathy Yelick

Associate Laboratory Director of Computing Sciences

Lawrence Berkeley National Laboratory

EECS Professor, UC Berkeley

White House Announces the National Strategic Computing Initiative (NSCI)

THE WHITE HOUSE

Office of the Press Secretary

For Immediate Release July 29, 2015

EXECUTIVE ORDER

- - - - - - -

CREATING A NATIONAL STRATEGIC COMPUTING INITIATIVE

By the authority vested in me as President by the Constitution and the laws of the United States of America, and to maximize benefits of high-performance computing (HPC) research, development, and deployment, it is hereby ordered as follows:

Section 1. Policy. In order to maximize the benefits of HPC for economic competitiveness and scientific discovery, the United States Government must create a coordinated Federal strategy in HPC research, development, and deployment. Investment in HPC has contributed substantially to national economic prosperity and rapidly accelerated scientific discovery. Creating and deploying technology at the leading edge is vital to advancing my Administration's priorities and spurring innovation. Accordingly, this order establishes the National Strategic Computing Initiative (NSCI). The NSCI is a

Five goals:

1. **Create systems that can apply exaflops of computing power to exabytes of data.**
2. **Keep the United States at the forefront of HPC capabilities.**
3. **Improve HPC application developer productivity.**
4. **Make HPC readily available.**
5. **Establish hardware technology for future HPC systems.**

[DOE SC and NNSA] will execute a joint program focused on advanced simulation through a capable exascale computing …

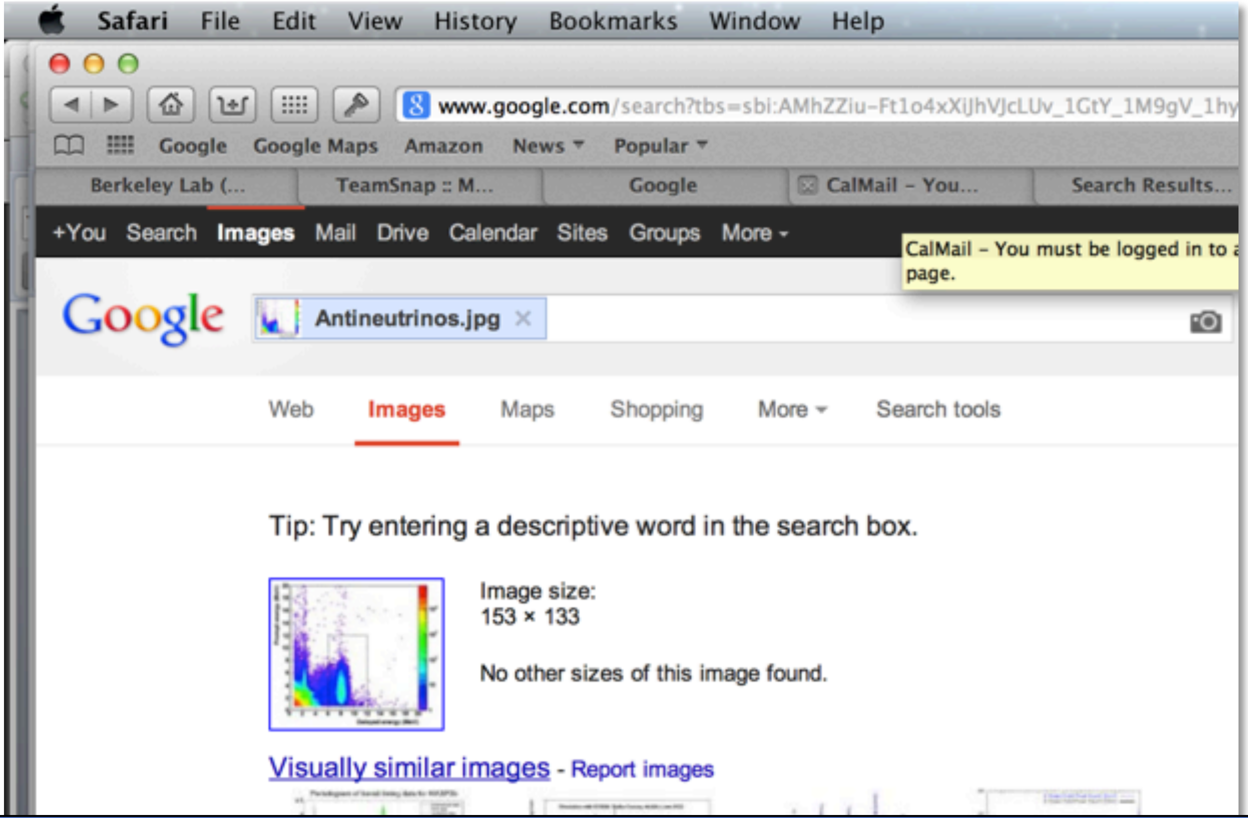
Big Data and HPC

Convergence in:

- **Science**
- **Algorithms**
- **Software**
- **Systems**

“Big Data” Changes Everything…What about Science?

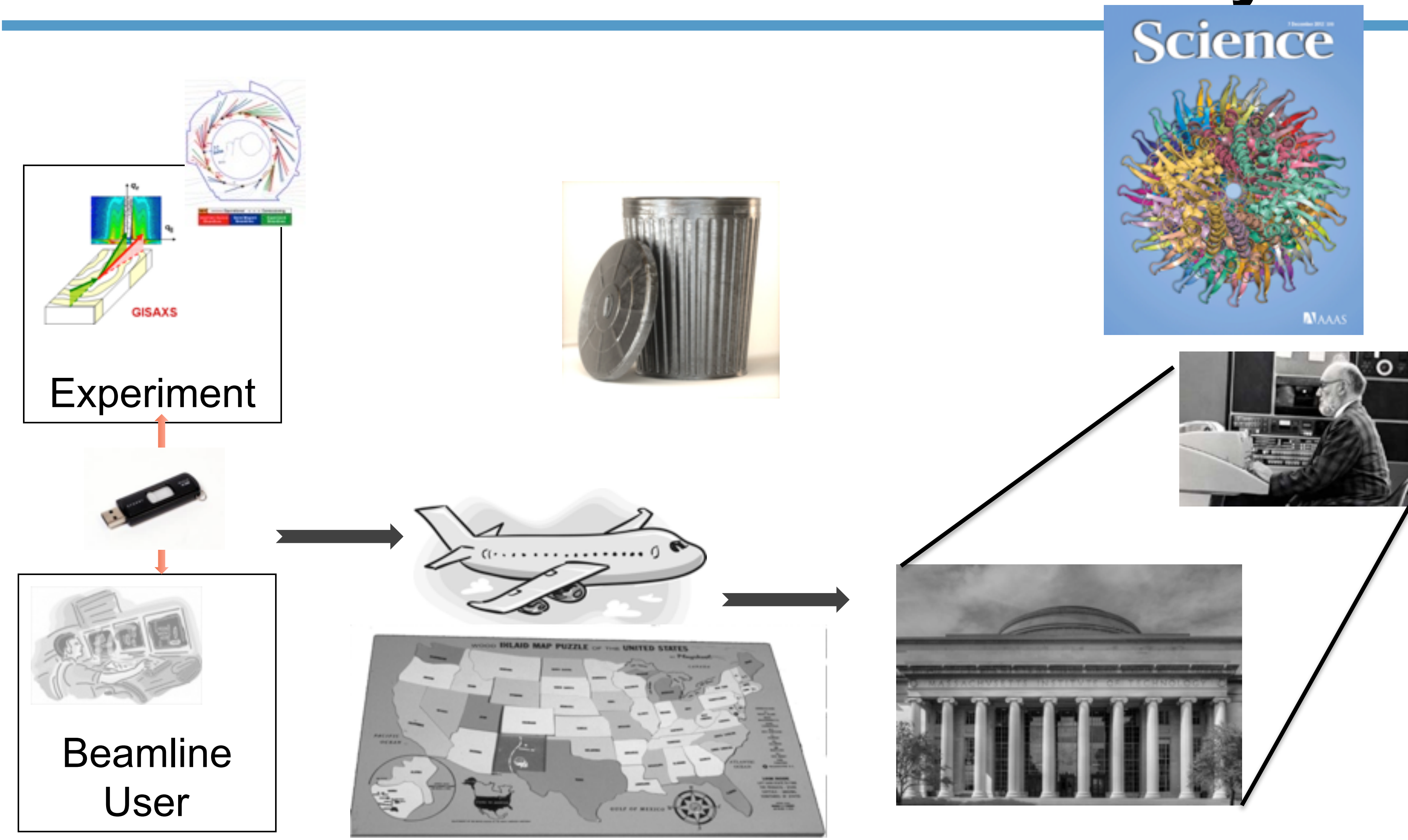
Transforming Science: Finding Data



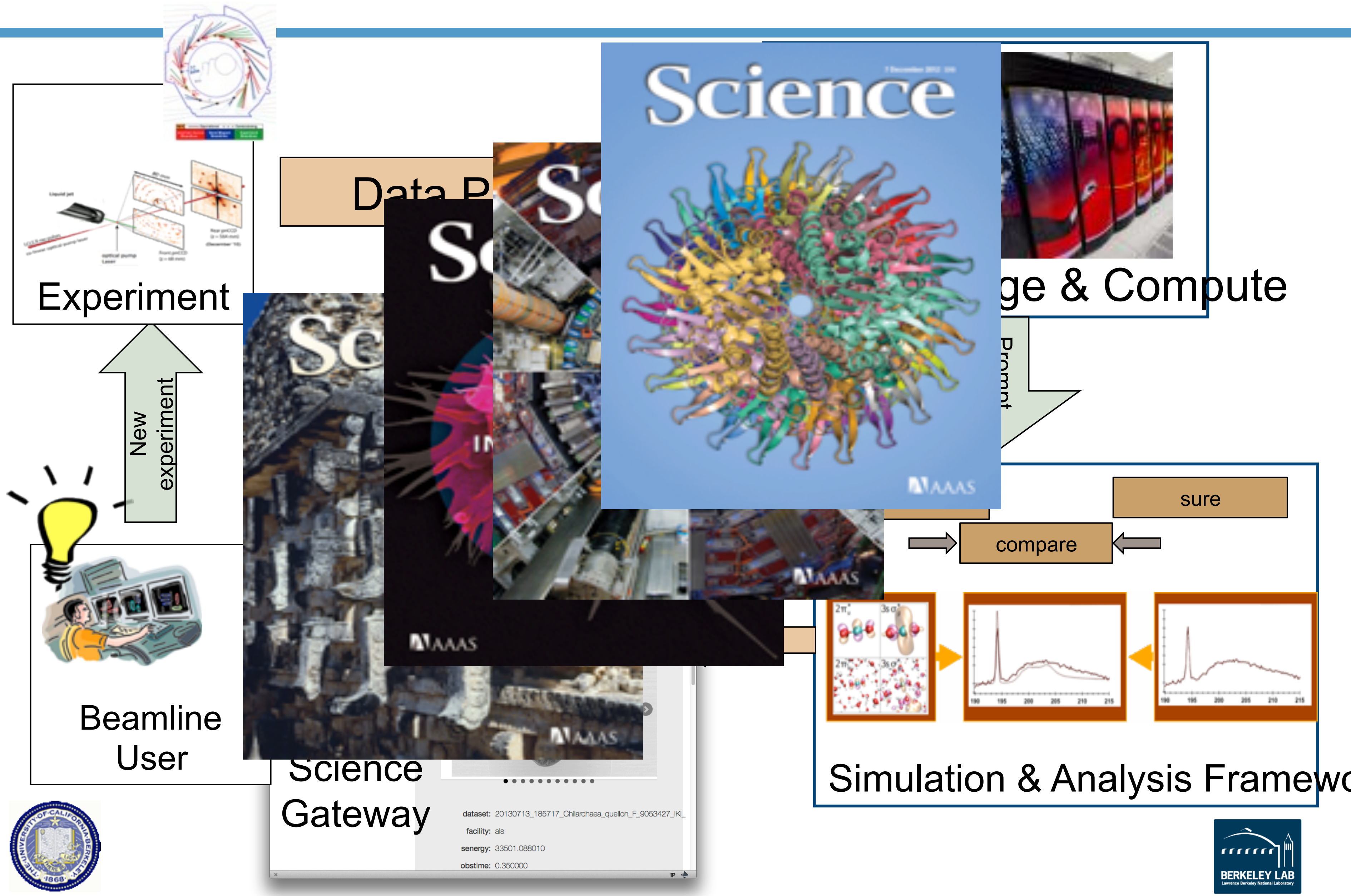
Computing Challenges:

- **Search for scientific data on the web**
- **Automated metadata annotation / feature identification**
- **Data: images, genomes, simulations, MRI, MassSpec,…**

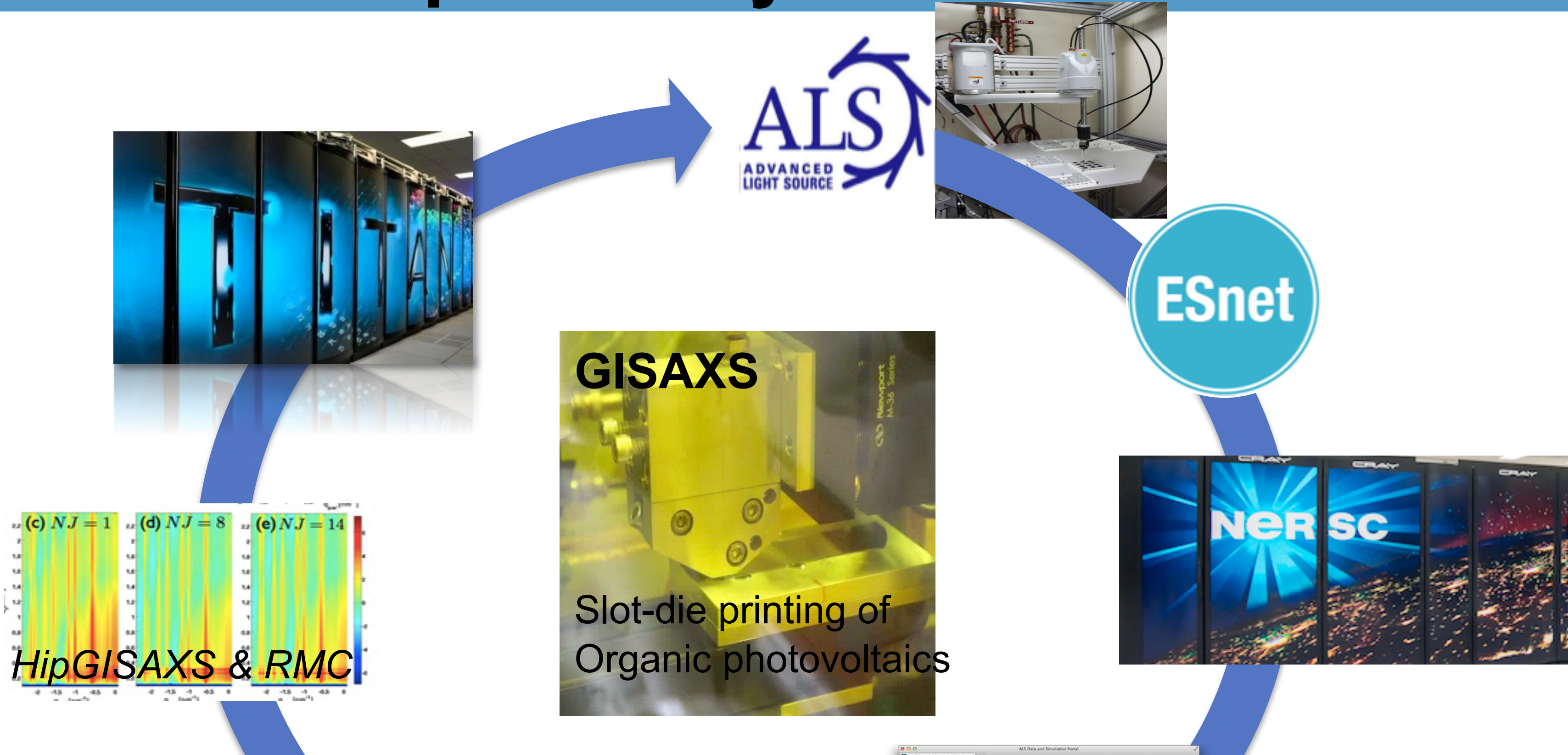
Scientific Workflow Today



The Future of Experimental Science



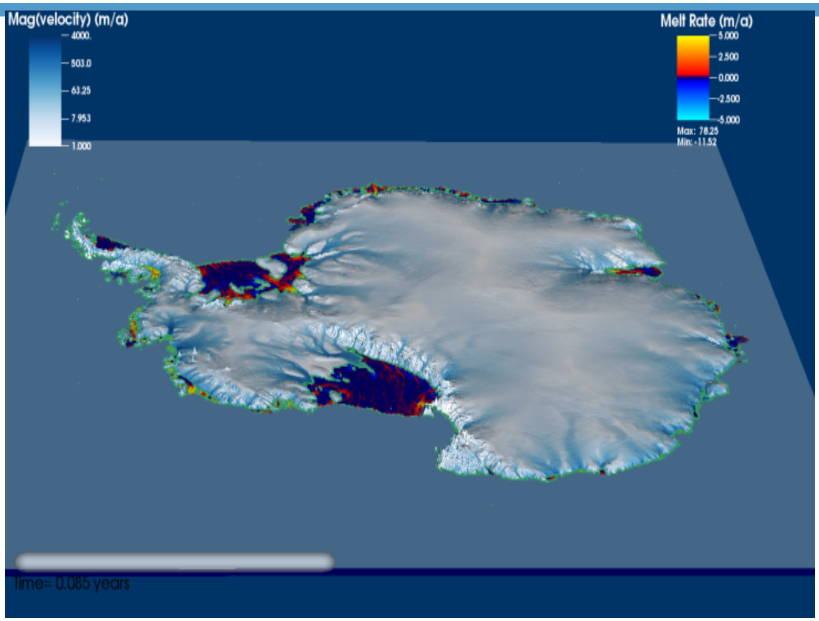
Transforming experimental science: "Superfacility" for Science



Computing Challenges:

- Robotics, Special purpose processors at experiments
- Mathematics / algorithm for real-time and offline analysis
- Massive numbers of simulations for inverse problems
- Networks and software for data movement, management

Scientific Discovery at the boundary of Simulation and Observation: Climate and microbes



New climate modeling methods, including AMR "Dycore" produce new understanding of ice

Genomes to watersheds Scientific Focus Area

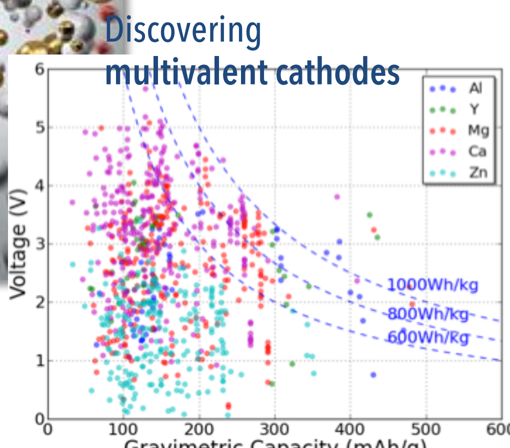
Computing Challenges:

- **Multimodal analysis from sensors, genomes, images…**
- **High performance methods and implementations**
- **Data-driven simulations to predict regional effects on environment and weather events**

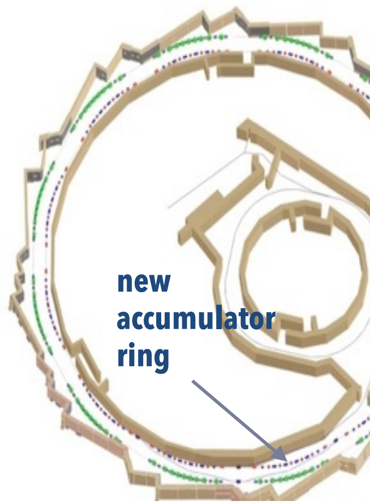
Science at the Boundary of Simulation and Observation: Understand and control energy

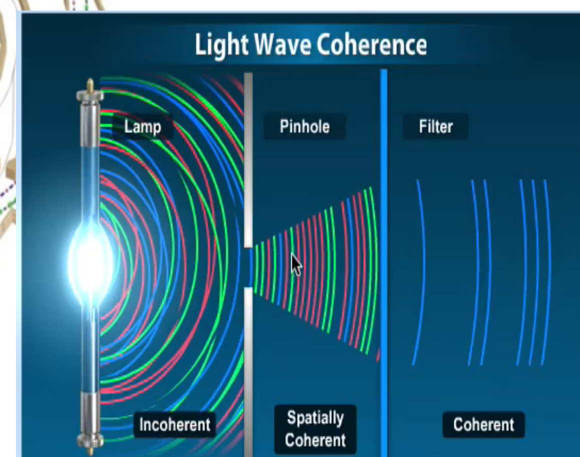
Materials Project

13,030 users hosted at NERSC with software co-developed by CRD



ALS-U Upgrade



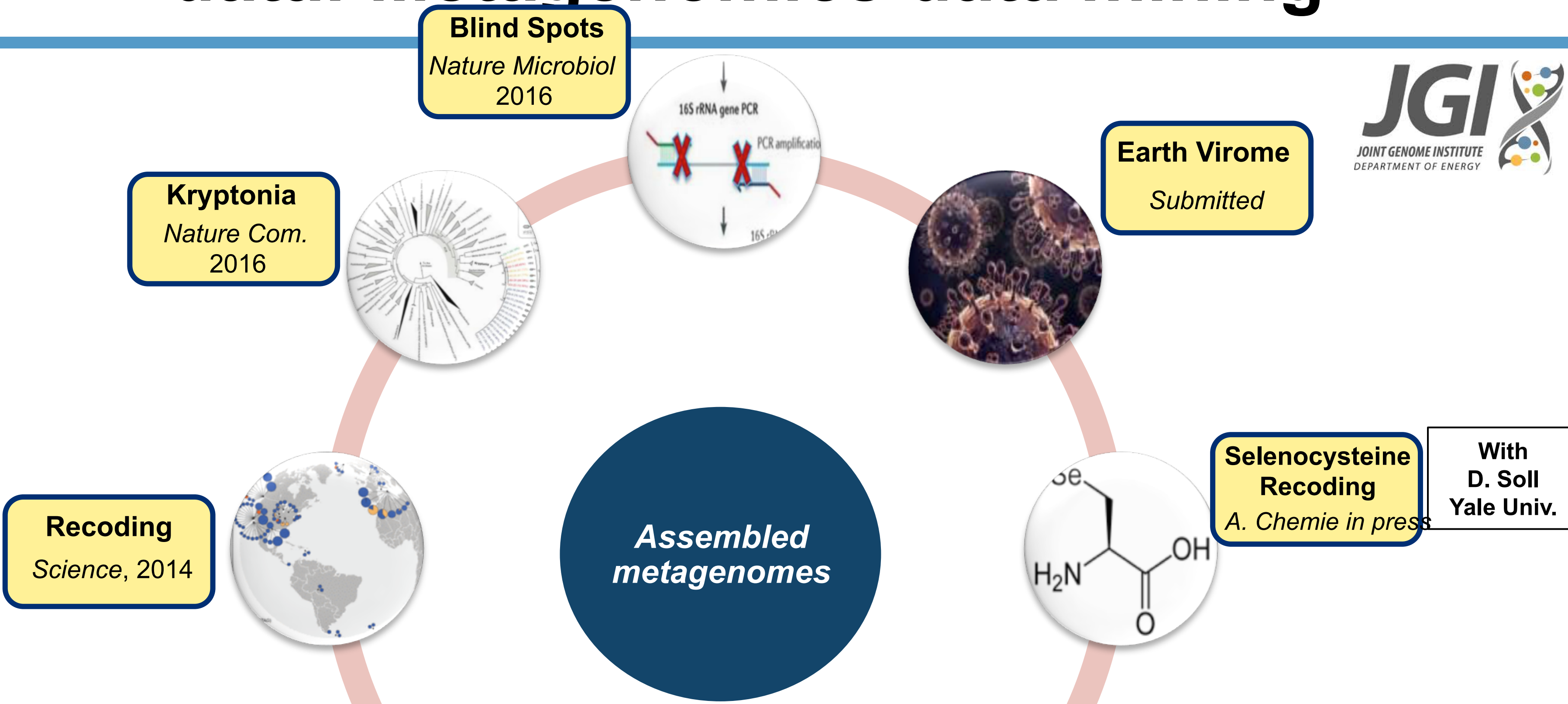


Computing Challenges:

- Machine learning on materials simulation data
- Analysis problems for experimental data (tomographic 3D reconstruction, x-ray scattering, etc.)
- Real-time job execution mixed with batch jobs

Finding structure and function in noisy data: Metagenomics data mining

JGI
JOINT GENOME INSTITUTE
DEPARTMENT OF ENERGY

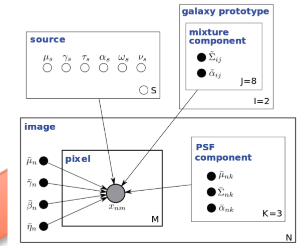


Computing Challenges:

- Distributed memory graph algorithms / hash tables
- Low latency interconnects; low overhead communication
- Algorithms to separate and assembly genomes
- Many-to-Many comparisons against databases

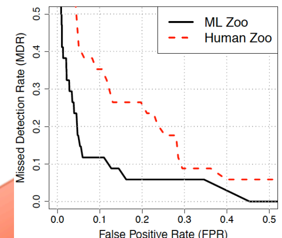
Finding smaller signals in noisy, biased data: Removing Systematic Bias in Cosmology

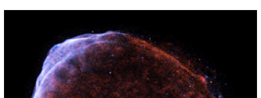
Graphical models



New simulation models and AMR code (Nyx)

Machine Learning





Crowd sourced

Filtered

Example: Astrophysicists discover early nearby

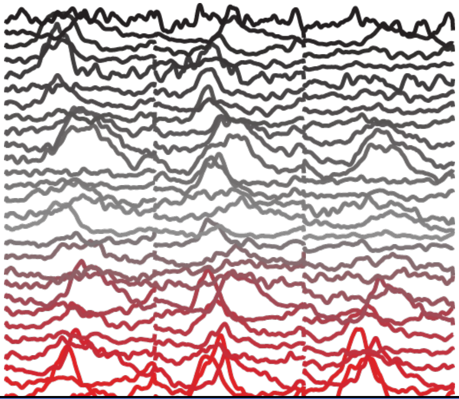
Computing Challenges:

- **Better machine learning for event detection**
- **Removing systematic bias in experimental data**
- **Simulations to interpret data; data constrain simulations**

Finding information across data modalities: Computing and the BRAIN Initiative

Function

dynamic data

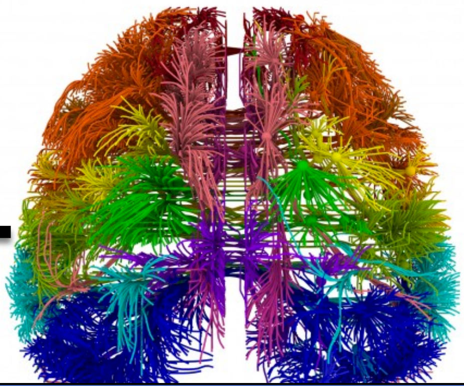


Theory & Models

abstractions

Structure

static data

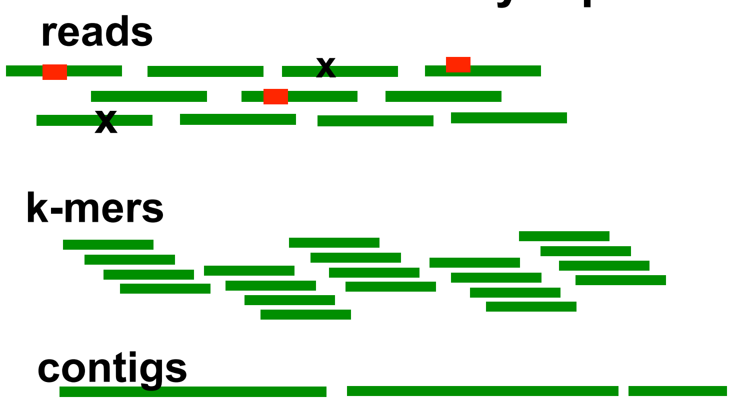


Computing Challenges:

- **Multimodal analysis (MRI, EM, CT, MS,…)**
- **Graph algorithms (irregular sparse matrices) at scale**

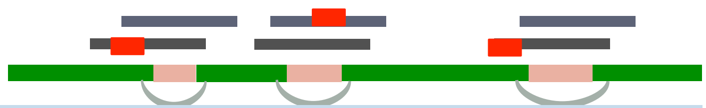
Languages for Random Access to Large Memory

Meraculous Assembly Pipeline



Human: 44 hours to 20 secs
Wheat: “doesn’t run” to 32 secs

Scaffolds using Scalable Alignment



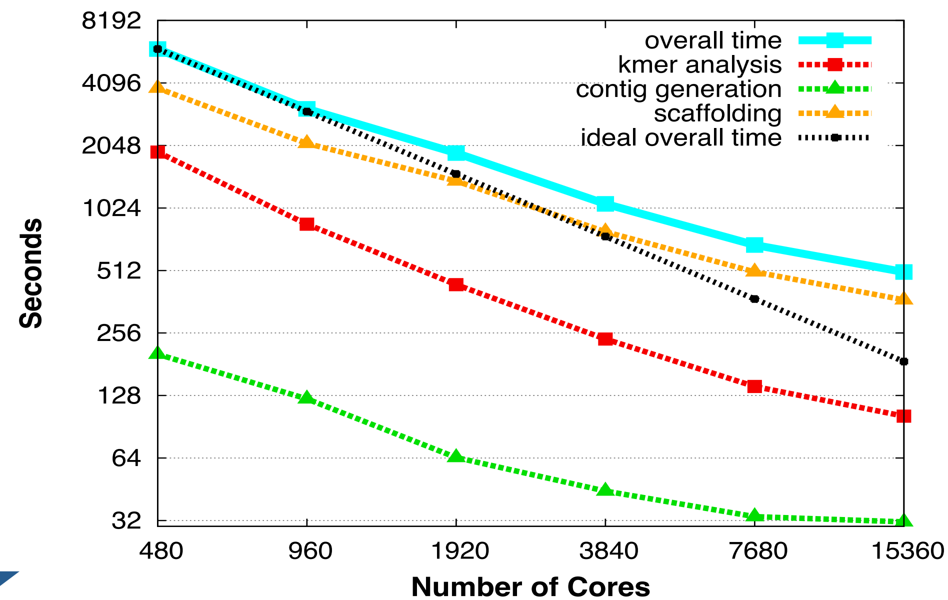
Combines with new algorithm to anchor 92% of wheat chromosome

Transforms process of discovery for de novo assembly

Perl to PGAS: Distributed Hash Tables

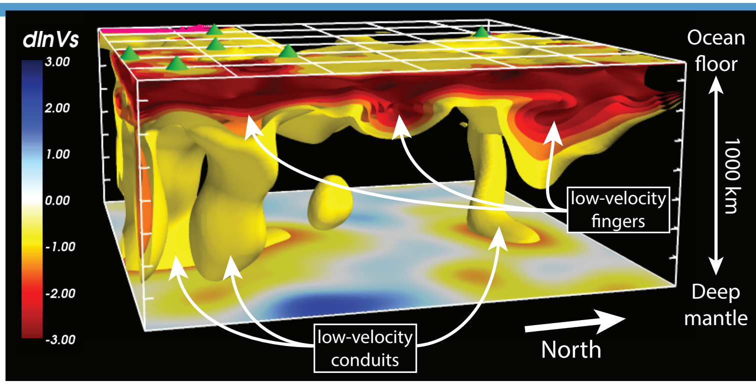
- Remote Atomics
- Dynamic Aggregation
- Software Caching (sometimes)
- Clever algorithms and data structures (bloom filters, locality-aware hashing)

→ **Hash Table with “tunable” runtime**

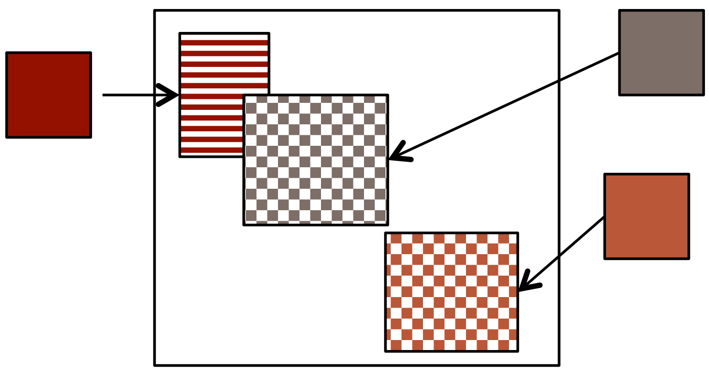


Evangelos Georganas, Aydin Buluc (MANTISSA), Lenny Oliker, Jarrod Chapman (JGI), Dan Rokhsar (JGI), Kathy Yelick

Languages for Irregular Access: Data Fusion in UPC++



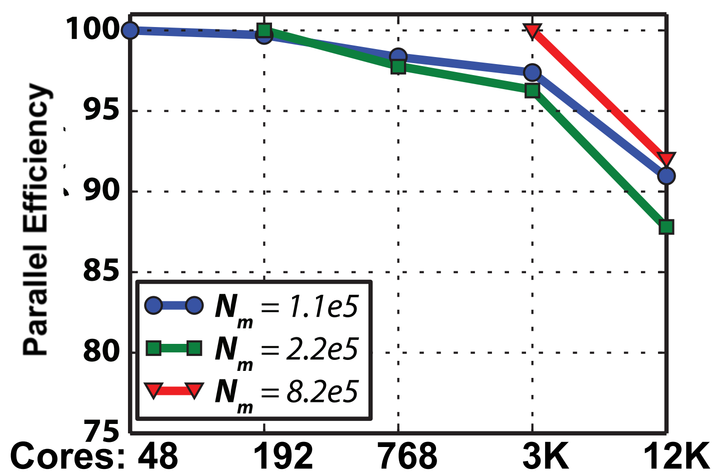
- Seismic modeling for energy applications “fuses” observational data into simulation
- With UPC++, can solve larger problems



Distributed Matrix Assembly

- Remote asyncs with user-controlled resource management
- Team idea to divide threads into injectors / updaters
- 6x faster than MPI 3.0 on 1K nodes

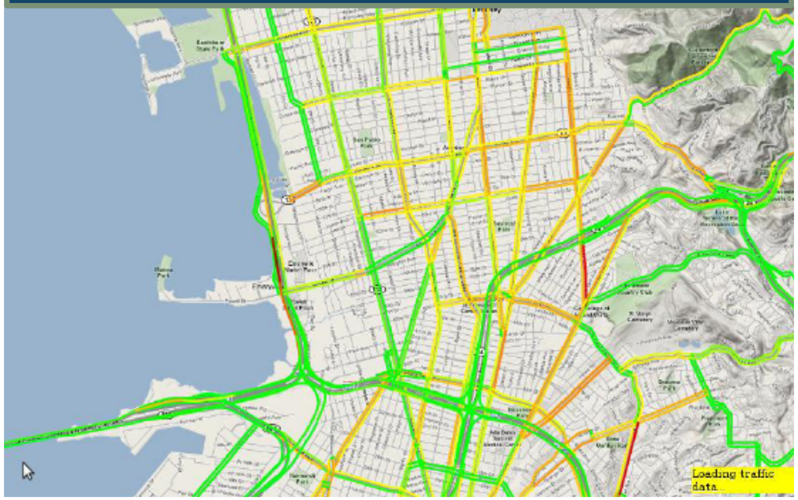
→ **Improving UPC++ team support**



French and Romanowicz use code with UPC++ phase to compute *first ever* whole-mantle global tomographic model using numerical seismic wavefield computations (F & R, 2014, GJI, extending F et al., 2013, Science). See F et al, IPDPS 2015 for parallelization overview.

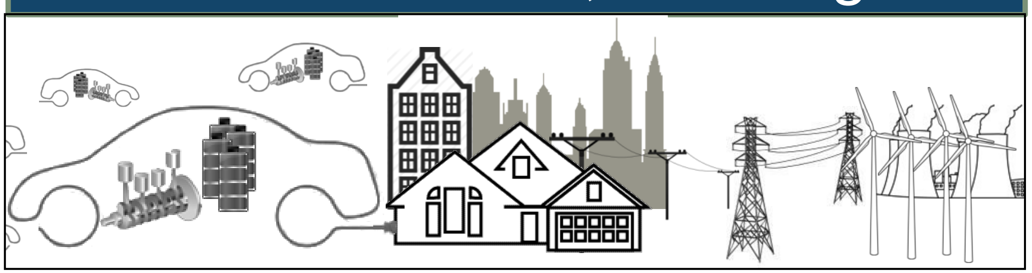
Science in embedded sensors: Internet of Things

Transportation Modeling

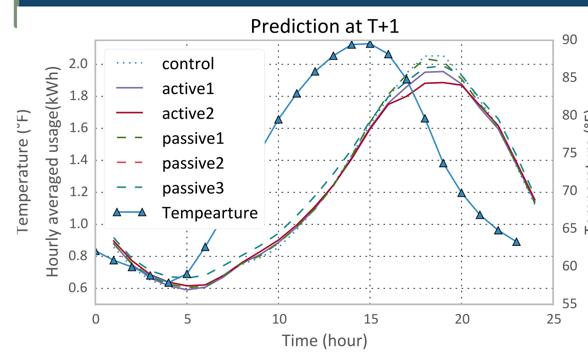


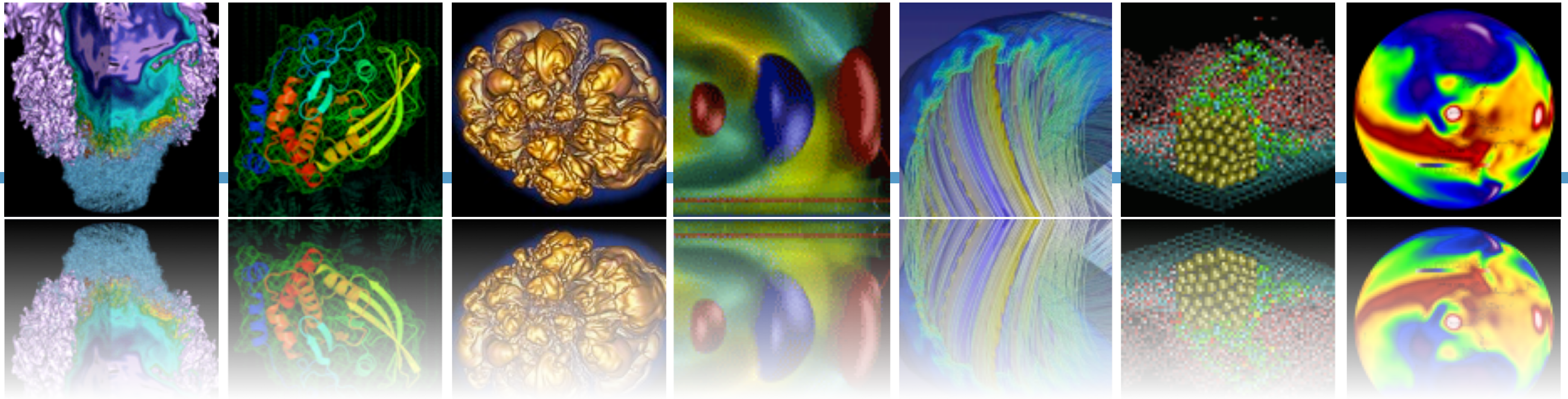
Power Grid Modeling

Scenario Prediction, Planning



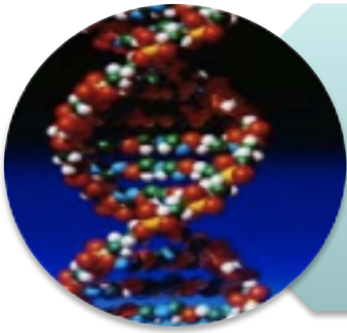
Decision Science



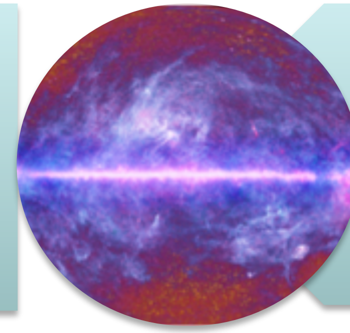


Science Data
Big (and Growing)

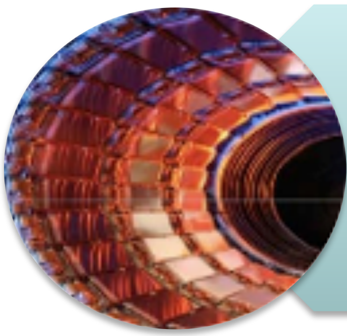
“Big Data” Challenges in Science
Volume, velocity, variety, and veracity



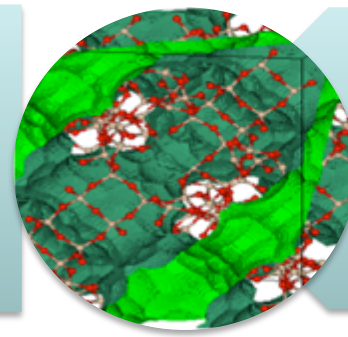
Biology
- *Volume:* Petabytes now; computation-limited
- *Variety*: multi-modal analysis on bioimages



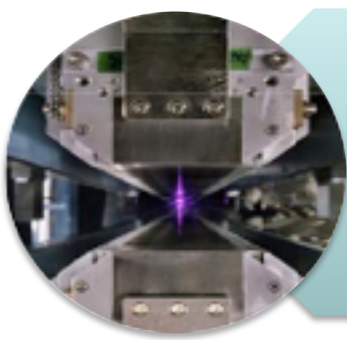
Cosmology / Astronomy:
- *Volume:* 1000x increase every 15 years
- *Variety:* combine data sources for accuracy



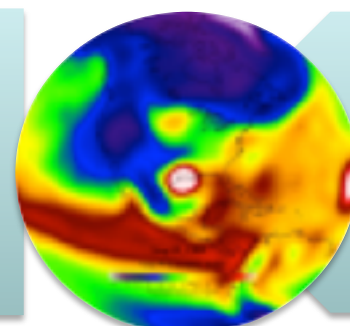
High Energy Physics
- *Volume*: 3-5x in 5 years
- *Velocity*: real-time filtering adapts to intended observation



Materials:
- *Variety:* multiple models and experimental data
- *Veracity:* quality and resolution of simulations

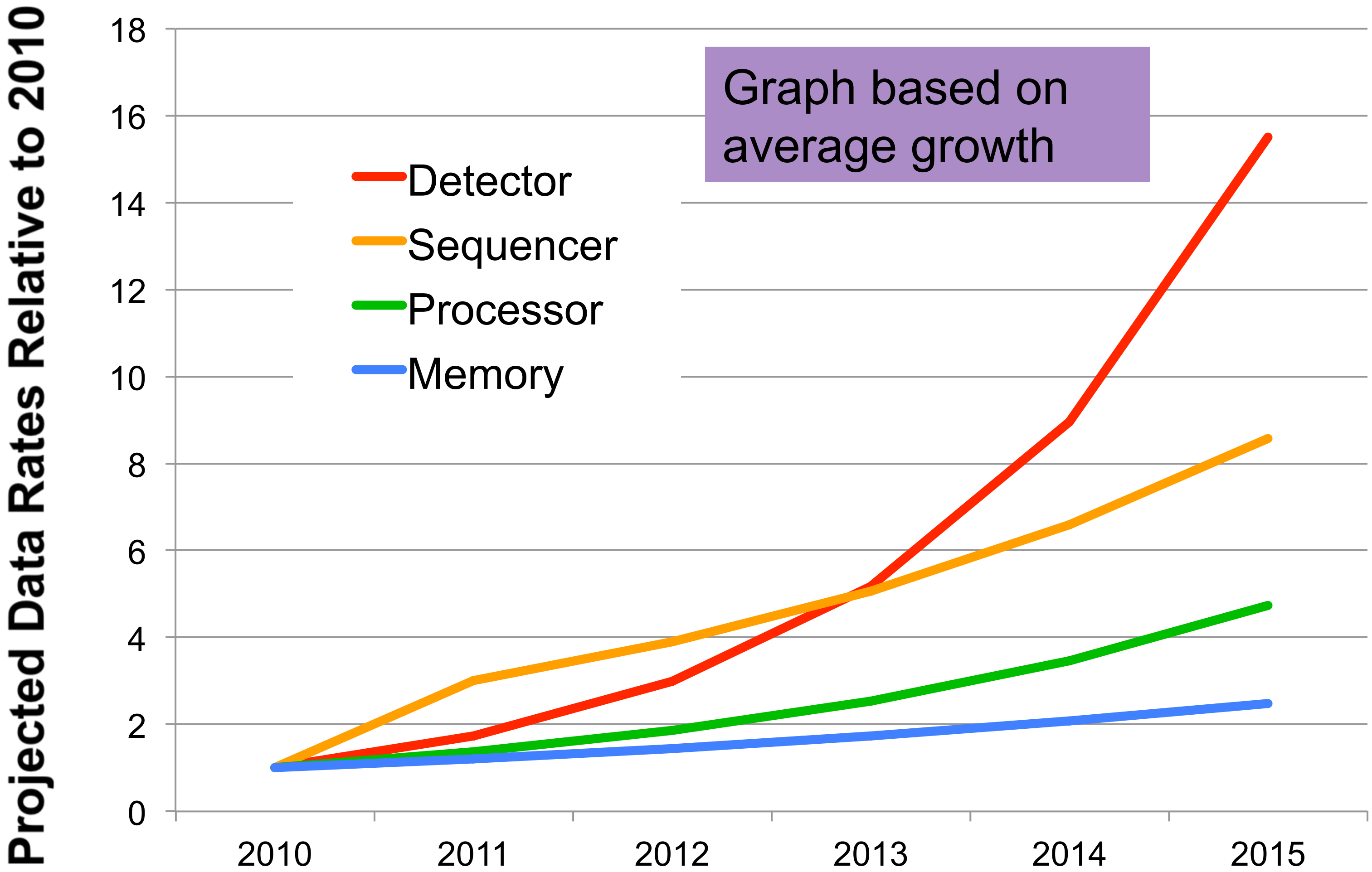


Light Sources
- *Velocity:* CCDs outpacing Moore’s Law
- *Veracity:* noisy data for 3D reconstruction

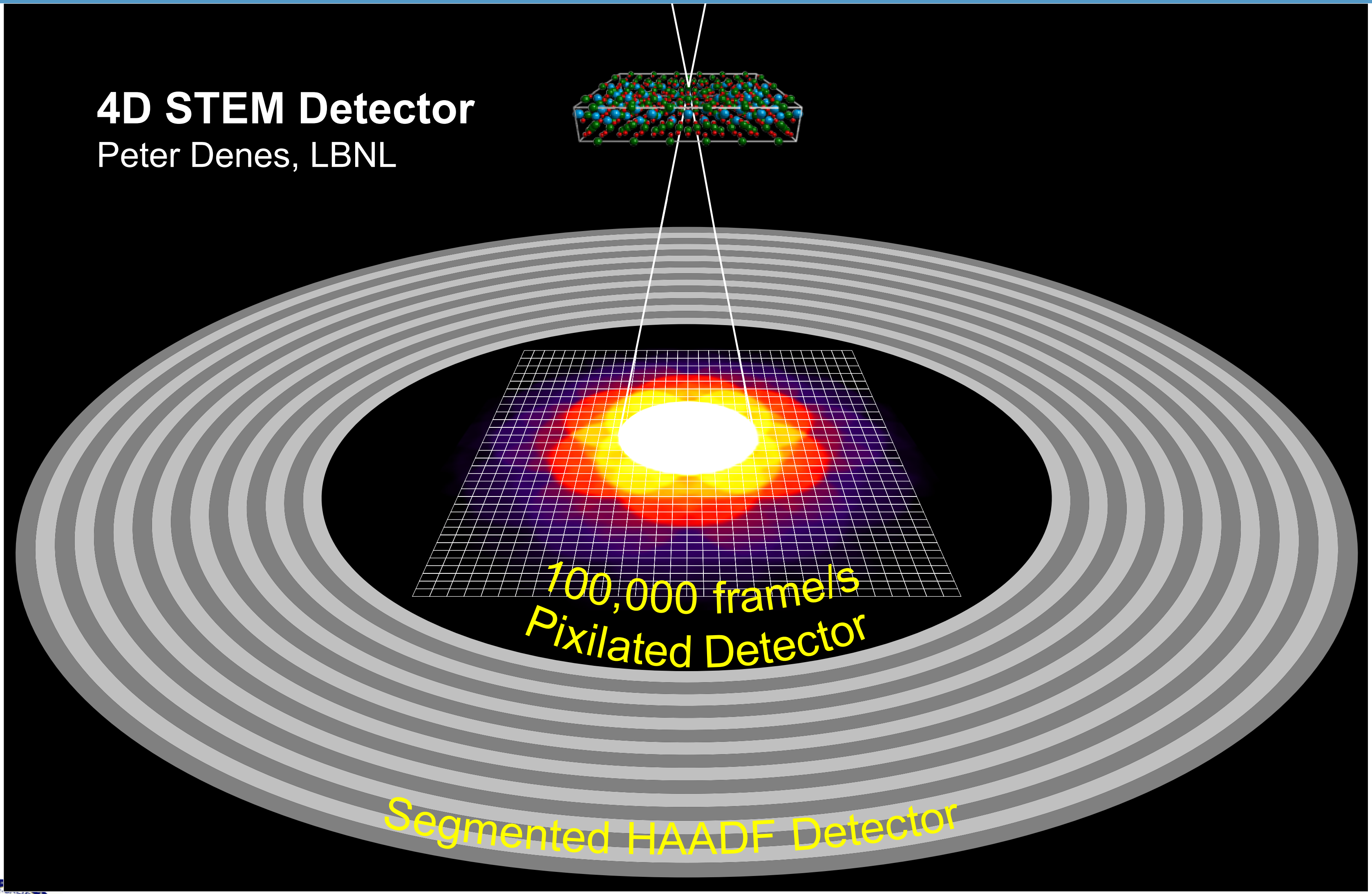


Climate
- *Volume:* Hundreds of exabytes by 2020
- *Veracity:* Reanalysis of 100-year-old sparse data

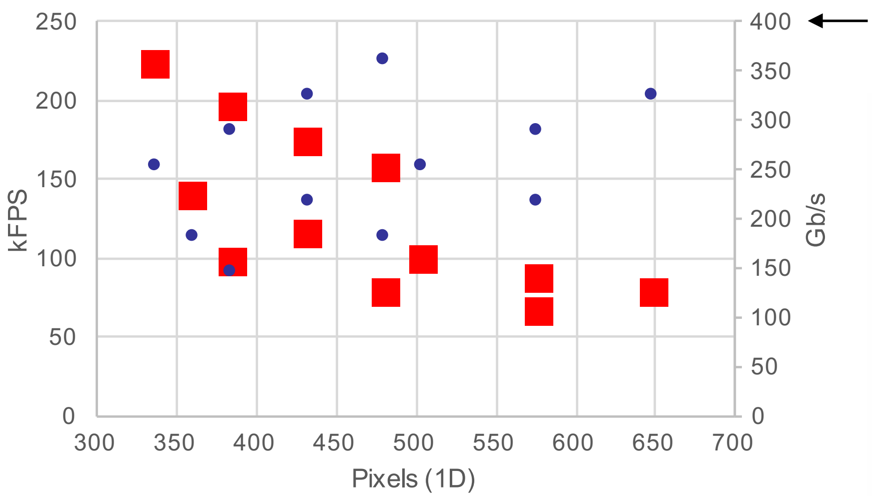
Data Growth is Outpacing Computing Growth



Measurement technology getting better; computation getting hardware



Superfacility for 100,000 FPS Detector



- Brocade: 400 Gb/s

- 100 kFPS → 10s of TB / hour
- Real time analysis:
  - Sparsification
  - Clustering
  - Dedicated network to NERSC

Brocade
130 Holger Way, San Jose, CA 95134
T. 408.333.8000 F. 408.333.8101
www.brocade.com

BROCADE

April 7, 2015

Mr. Brent Draney
LBL-NERSC
415-20th Street
Oakland, CA 94612

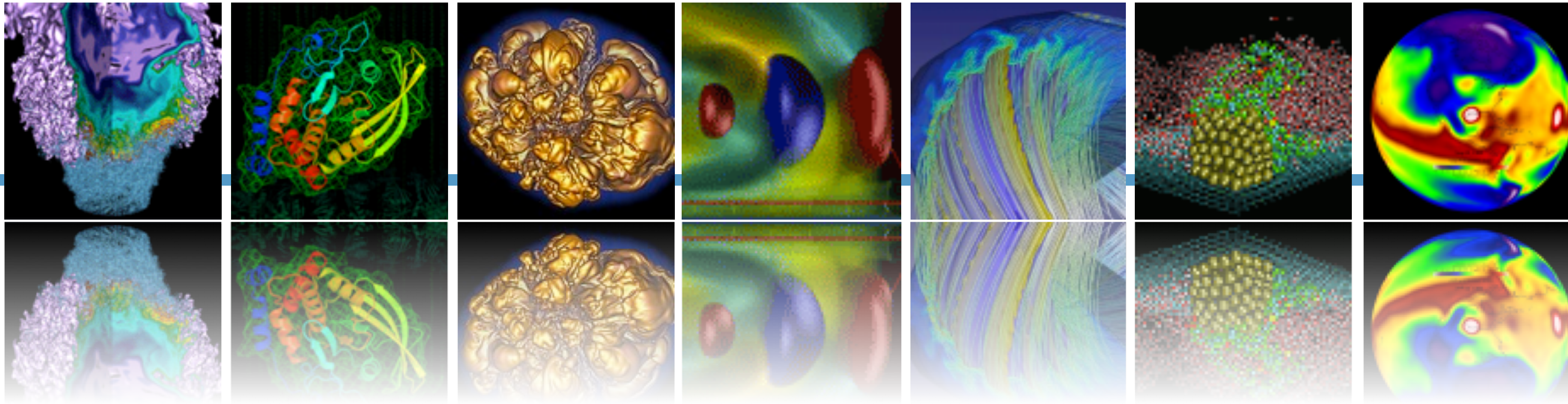
Dear Brent,

Brocade has a long history of innovation and collaboration in the high tech research community. Continuing this tradition, Brocade would be honored to partner with NERSC on the "Future Electron Scattering Project" by loaning switching hardware. Brocade agrees to loan a switching layer for the project which provides at least ten ports of 40 gig and 4 ports of 100G by Q42016.

Brocade understands that at the end of the project all equipment will be returned to Brocade.

Sincerely,

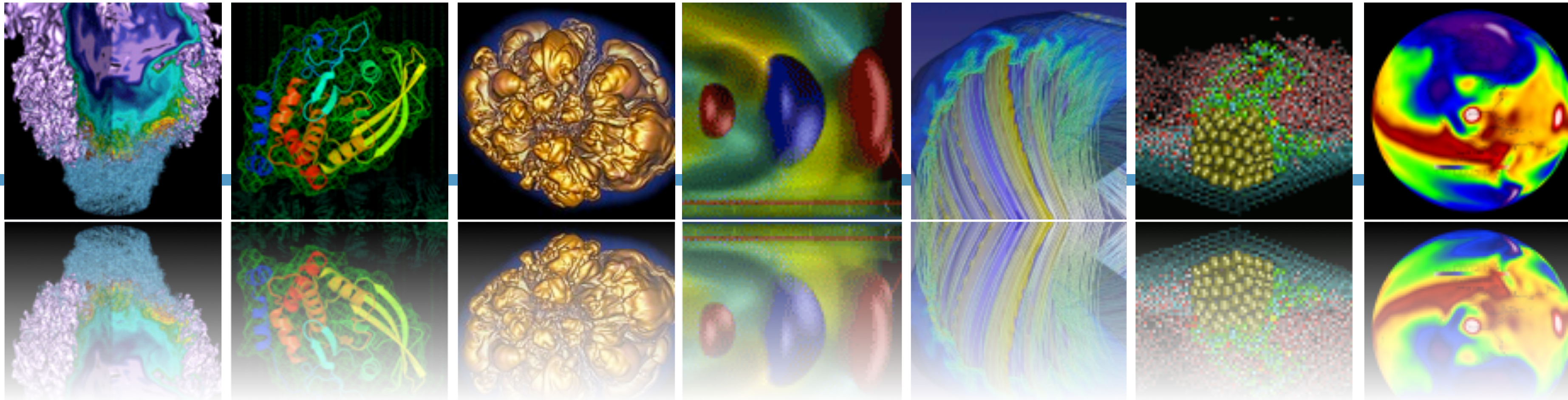
Michael Bushong
Vice President
Data Center Switching and Routing



Algorithms Convergence?

Analytics vs. Simulation Kernels:

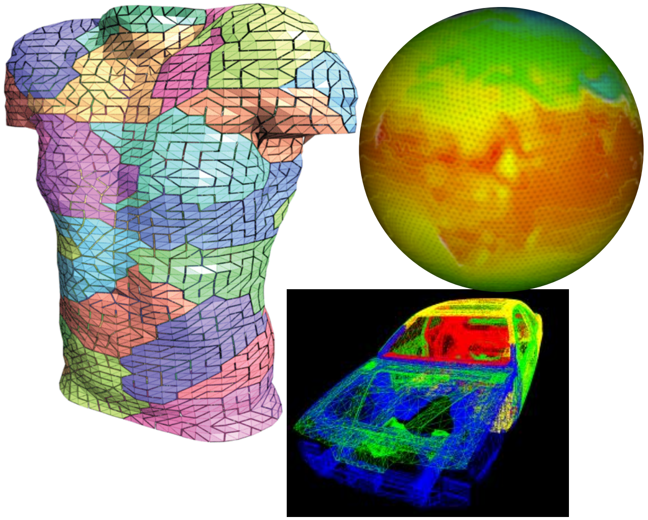
7 Giants of Data	7 Dwarfs of Simulation
Basic statistics	Monte Carlo methods
Generalized N-Body	Particle methods
Graph-theory	Unstructured meshes
Linear algebra	Dense Linear Algebra
Optimizations	Sparse Linear Algebra
Integrations	Spectral methods
Alignment	Structured Meshes



Software Convergence?

Data Analytics: Case for PGAS

More Regular



<u>Message Passing Programming</u>

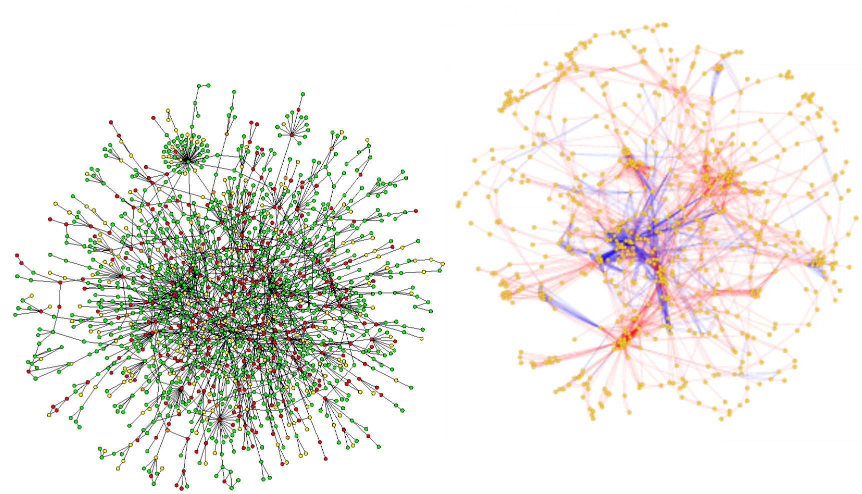
Divide up domain in pieces

Compute one piece

Send/Receive data from others

MPI, and many libraries

More Irregular



<u>Global Address Space Programming</u>

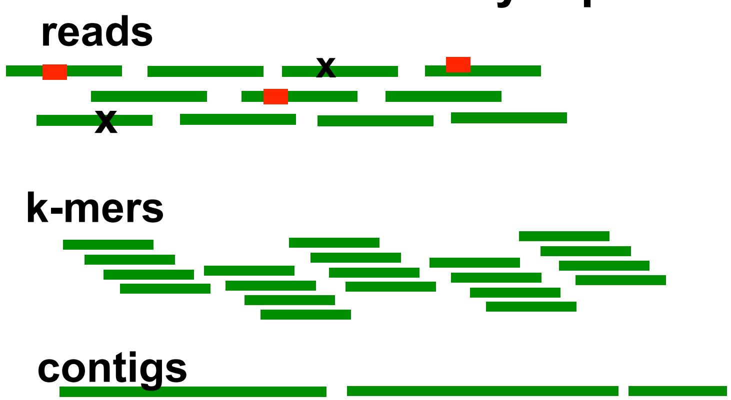
Each start computing

Grab whatever / whenever

UPC, CAF, X10, Chapel, GlobalArrays

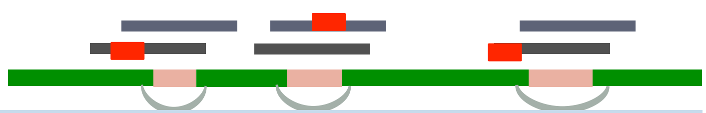
Languages for Random Access to Large Memory

Meraculous Assembly Pipeline



Human: 44 hours to 20 secs
Wheat: "doesn't run" to 32 secs

Scaffolds using Scalable Alignment



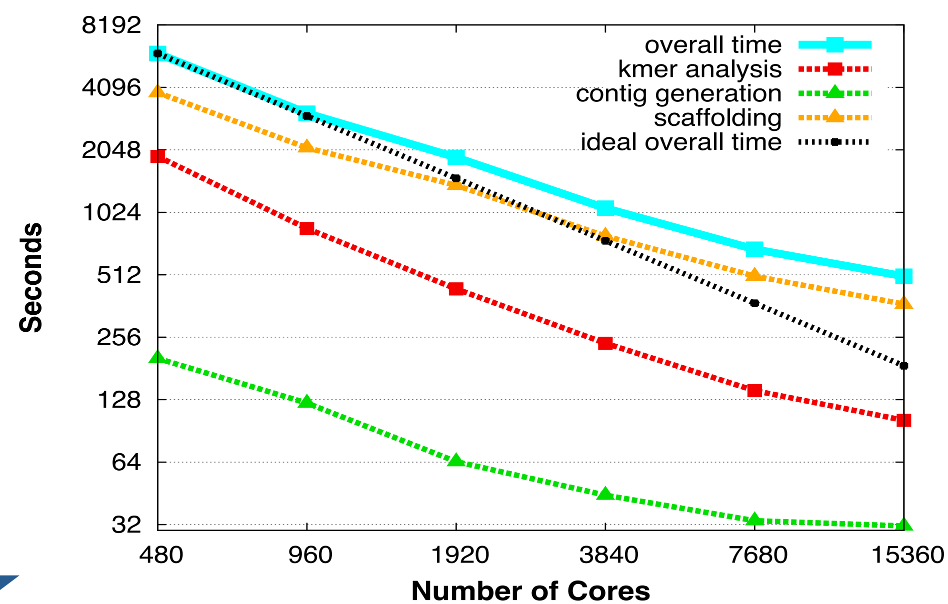
Combines with new algorithm to anchor 92% of wheat chromosome

Transforms process of discovery for de novo assembly

Perl to PGAS: Distributed Hash Tables

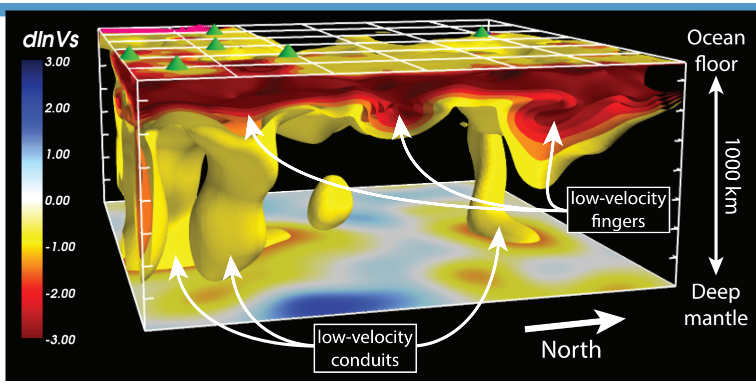
- Remote Atomics
- Dynamic Aggregation
- Software Caching (sometimes)
- Clever algorithms and data structures (bloom filters, locality-aware hashing)

→ Hash Table with "tunable" runtime

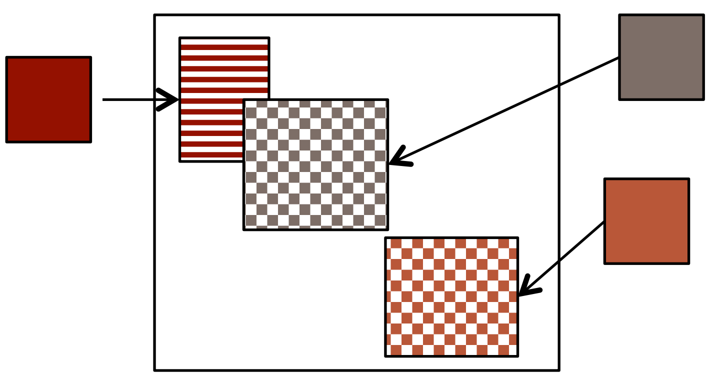


Evangelos Georganas, Aydin Buluc (MANTISSA), Lenny Oliker, Jarrod Chapman (JGI), Dan Rokhsar (JGI), Kathy Yelick

Languages for Irregular Access: Data Fusion in UPC++



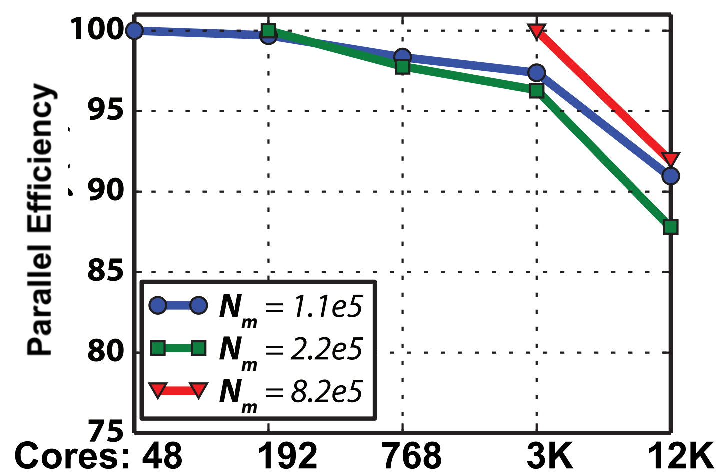
- Seismic modeling for energy applications "fuses" observational data into simulation
- With UPC++, can solve larger problems



Distributed Matrix Assembly

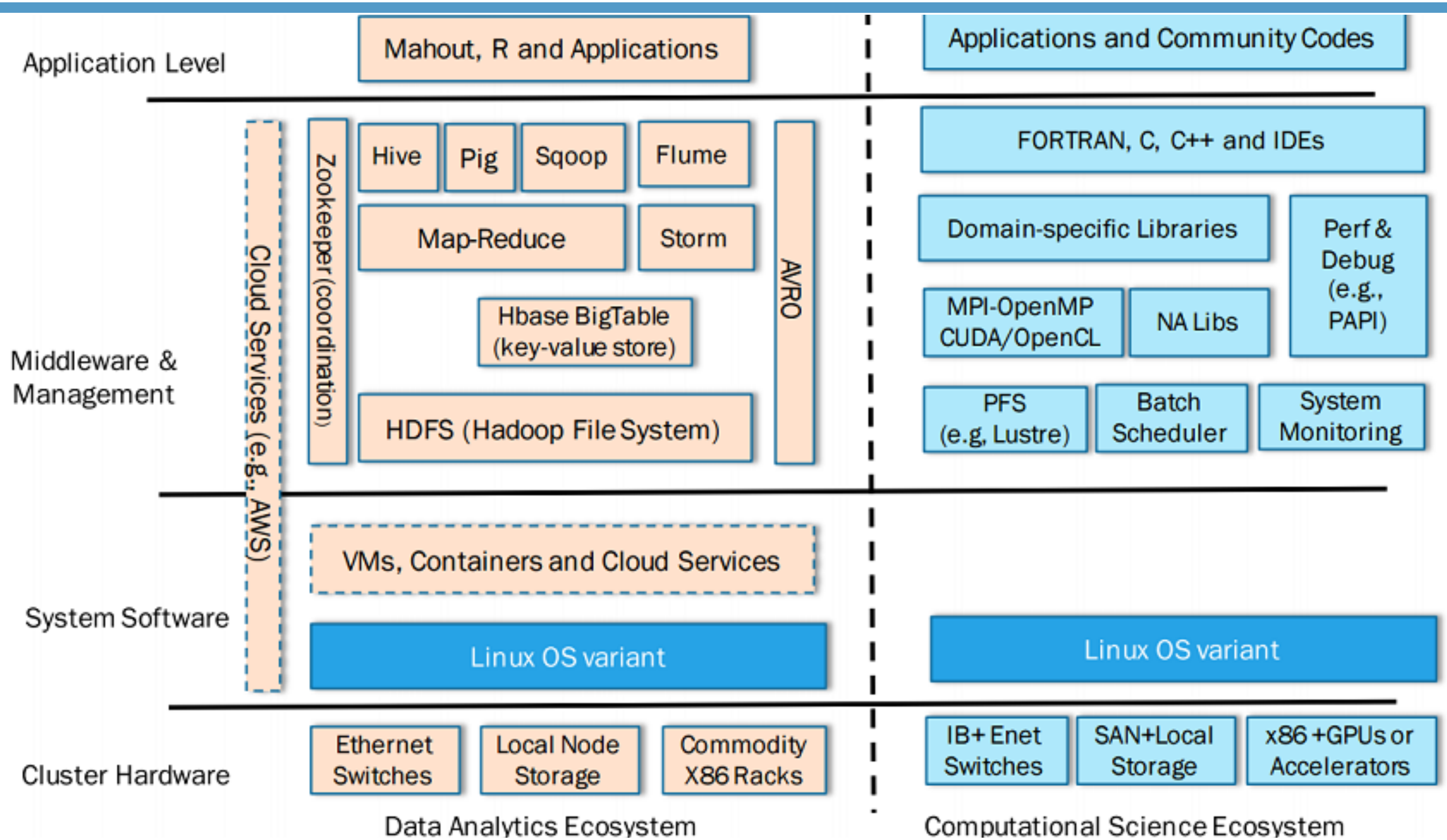
- Remote asyncs with user-controlled resource management
- Team idea to divide threads into injectors / updaters
- 6x faster than MPI 3.0 on 1K nodes

→ **Improving UPC++ team support**



French and Romanowicz use code with UPC++ phase to compute *first ever* whole-mantle global tomographic model using numerical seismic wavefield computations (F & R, 2014, GJI, extending F et al., 2013, Science). See F et al, IPDPS 2015 for parallelization overview.

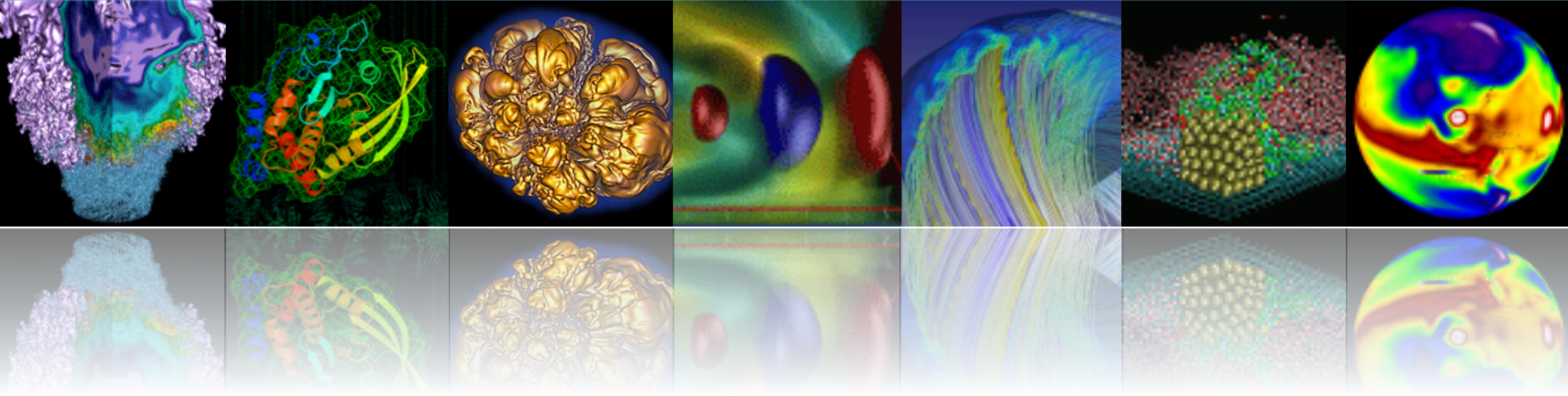
Divergent Ecosystems



Dongarra and Reed, HPCWire, 2016

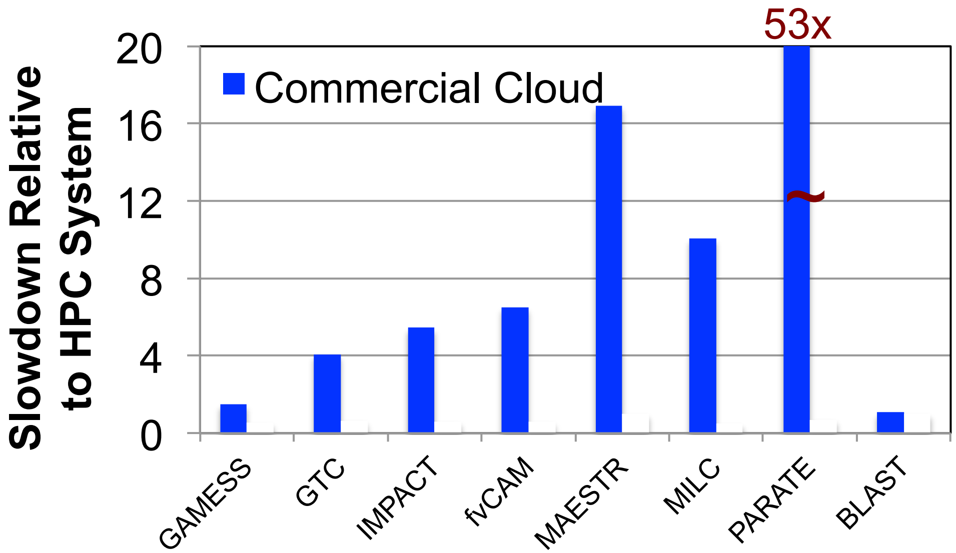
Two ecosystems

Data / Cloud + Analytics	HPC / Simulation
Commodity processors	Commodity processors (latest)
	Accelerators
DRAM	DRAM (+ NVRAM?)
Ethernet	Low latency / overhead interconnect
Local disk (+ NVRAM?)	Shared disk filesystem (+ NVRAM)
Low density (air cooled)	High density (liquid cooled)
<50% utilization (never wait)	>90%+ utilization (often wait)
Fault tolerant programming	After-the-fact checkpoint/restart
On-demand scheduling	Batch scheduling
Loosely coupled applications	Tightly coupled applications
Hadoop, SPARK,…	MPI, PGAS,…



System Convergence?

Myth: Supercomputers are Expensive, Clouds are Cheap

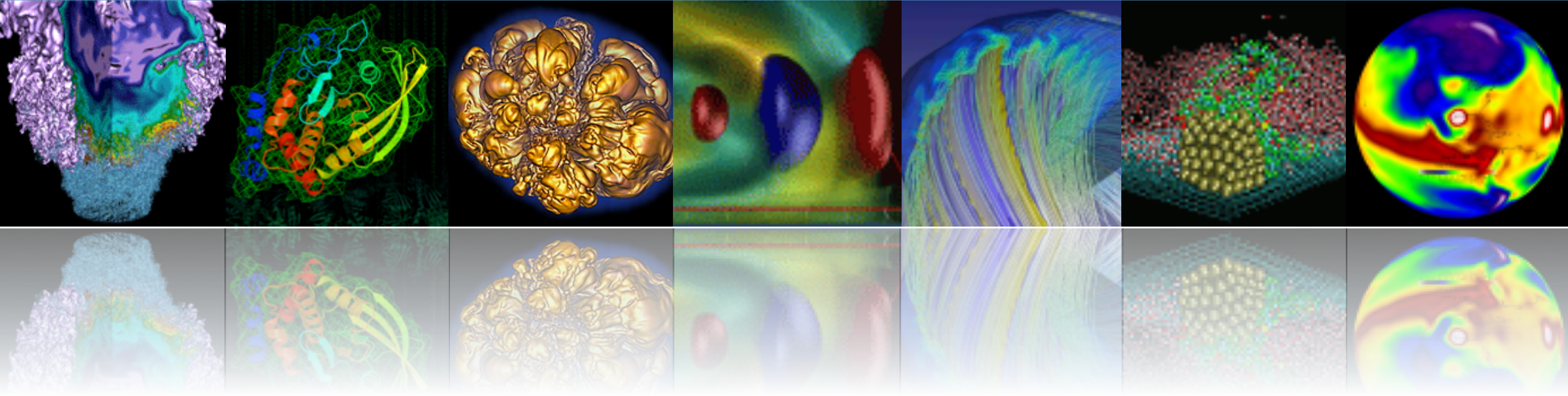


To buy raw NERSC core hours costs more than NERSC budget

- Even ignoring the measured performance slowdown
- **Doesn't include consulting staff, account management, licenses, bandwidth, software support: ~2/3 of NERSC's Budget**

Why?

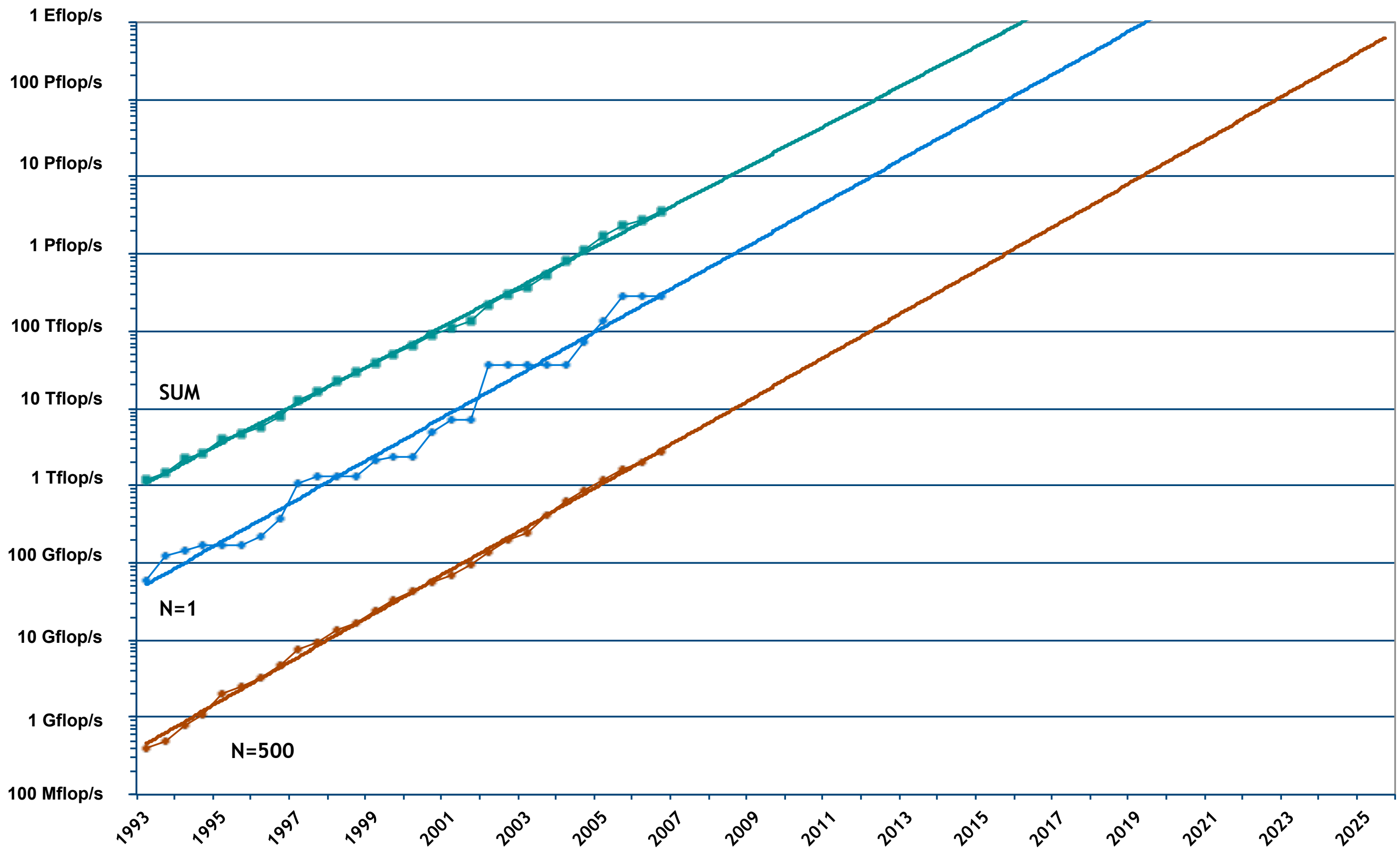
- NERSC cost/core hours dropped 10x (1000%) from 2007 to 2011, while Amazon pricing dropped 15% in the same period



What is Exascale about?
Real performance on real applications

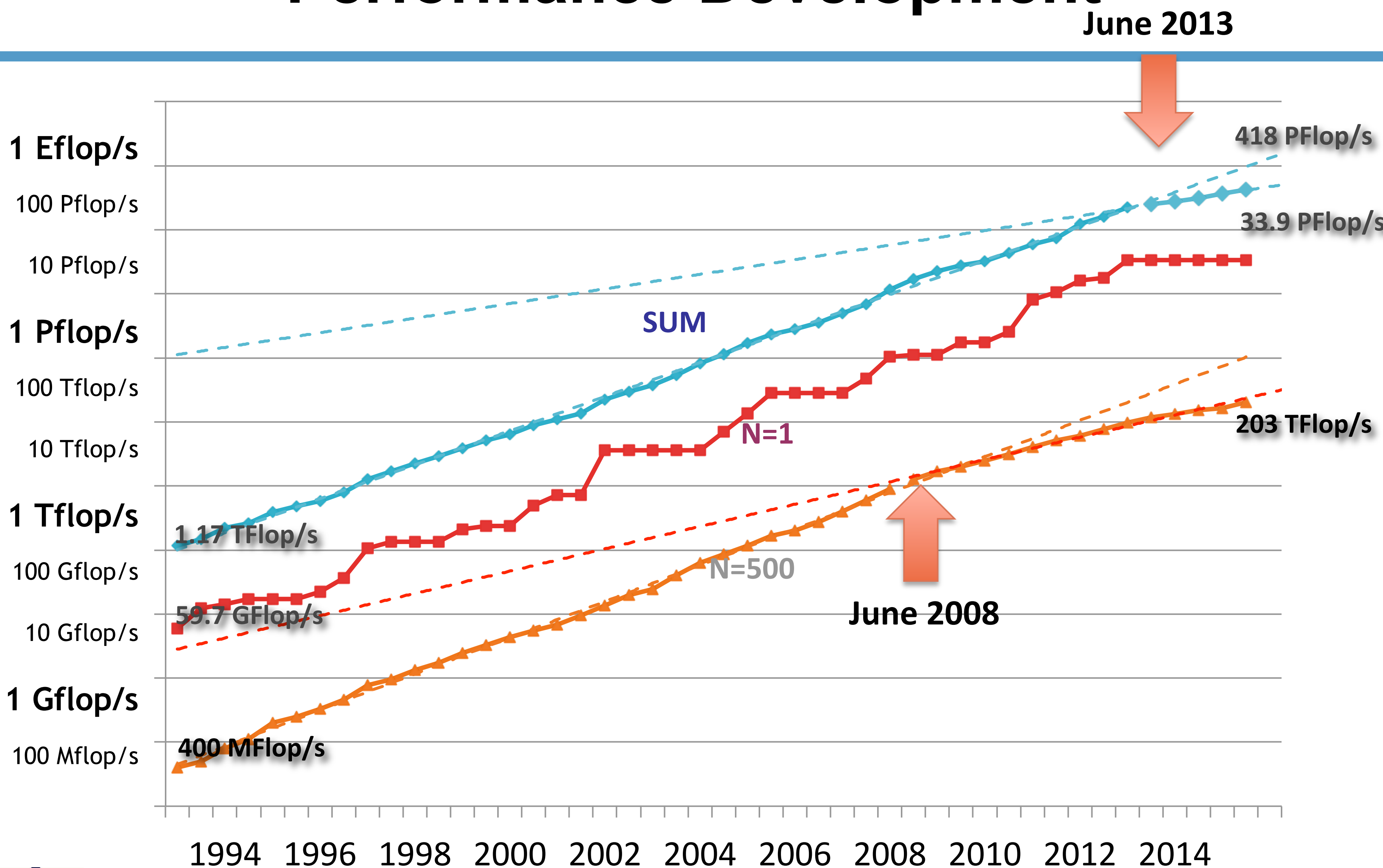
But let's try something easier:
HPL: High Performance LINPACK

TOP 500 Performance Projection - The Old Picture From 2007



Top500 (Slide from Horst Simon)

Performance Development



Top500 (Slide from Horst Simon)

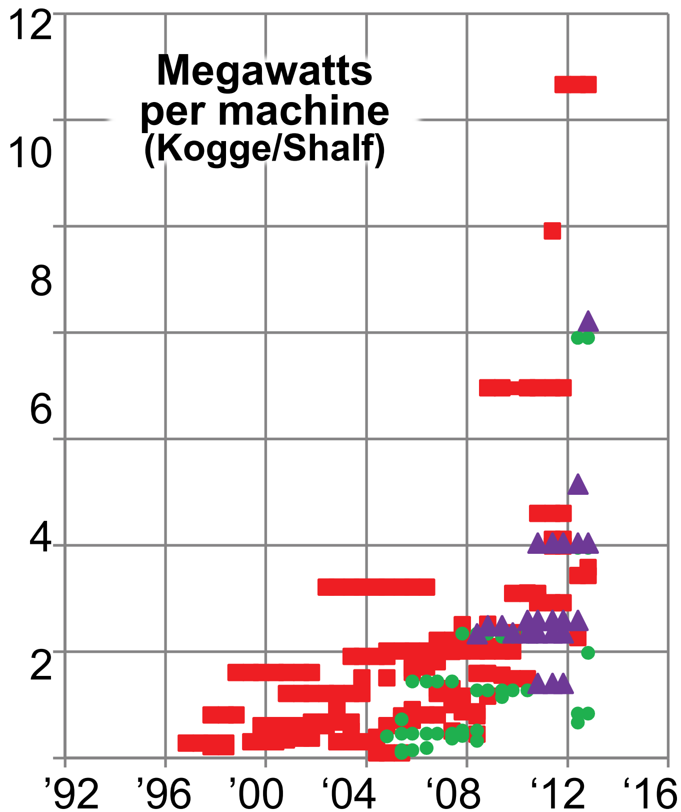
What Limits Computer Performance?

Computing is energy-constrained

At ~$1M per MW, energy costs are substantial

- 1 petaflop in 2008 used 3 MW
- 1 exaflop in 2018 at 200 MW "usual **chip** scaling"

Missing Tihanhe-2 at 18MW

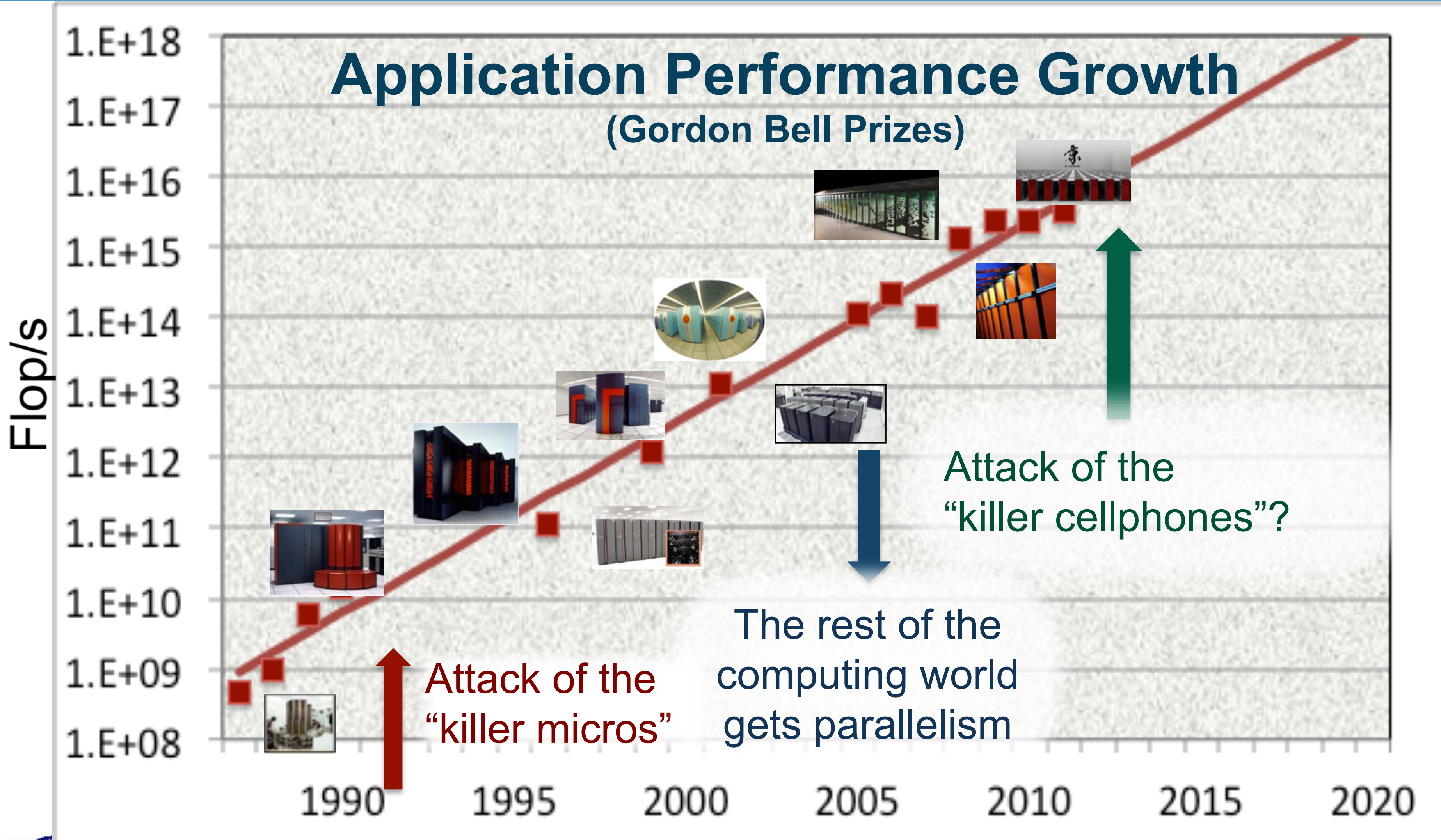


Goal: 1 Exaflop in 20 MW
= 20 pJ / operation

Note: The 20 pJ / operation is

- **Independent of machine size**
- **Independent of # cores used per application**
- **But "operations" need to be useful ones**

Computational Science has Moved through Difficult Technology Transitions



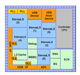
“Exascale” Challenges Affect Performance Growth at all Scales

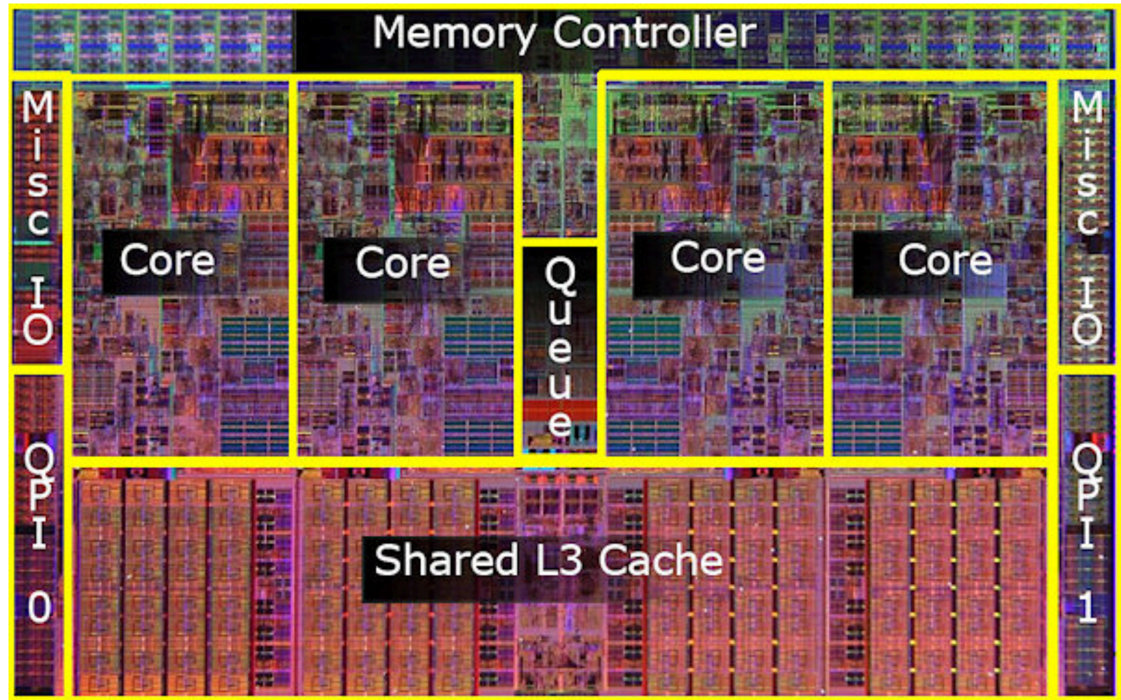
1) **Power** is the primary constraint
2) **Parallelism** (1000x today)
3) **Processor architecture** will change
4) **Data movement** dominates
5) **Memory** growth will not keep up
6) **Programming models** will change
7) **Algorithms** must adapt
8) **I/O** performance will not keep up
9) **Resilience** will be critical at this scale
10) **Interconnect bisection** must scale

- **These are all at the node levels**
- **Happening NOW!**
- **Emerging Programming solutions are**
  - **Hard to use**
  - **Non-portable**
  - **Non-durable**

Lightweight Cores are the Future

Cell phone processor (0.1 Watt, 4 Gflop/s)





Server processor (100 Watts, 50 Gflop/s)

- **Small, simple cores are energy and area efficient**
  - **10-100x more energy efficient**
- **Want to encourage "parallel thinking" in algorithms and software**

Take Home Message for Data and HPC (aka Analysis and Simulation)

- "Roofline" your code
- Understand motifs of your applications
- Question conventional wisdom: a system of type X is best
- Data is as important to science as business, society,…
- Clouds and HPC centers are optimized for different usage, but the underlying components are the same

Questions?