


Future Outlook for Advanced Computing

A photograph of a woman standing in the center of a long aisle in a server room. The aisle is lined with rows of black server racks on both sides. The racks have some green indicator lights. The ceiling has a grid of recessed lights. The woman is wearing a dark suit and a colorful scarf.

Dona L. Crawford
Associate Director Computation
Lawrence Livermore National Laboratory

Presented to
SLURM User Group Meeting
September 18, 2013

LLNL-PRES-643810

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

 Lawrence Livermore National Laboratory

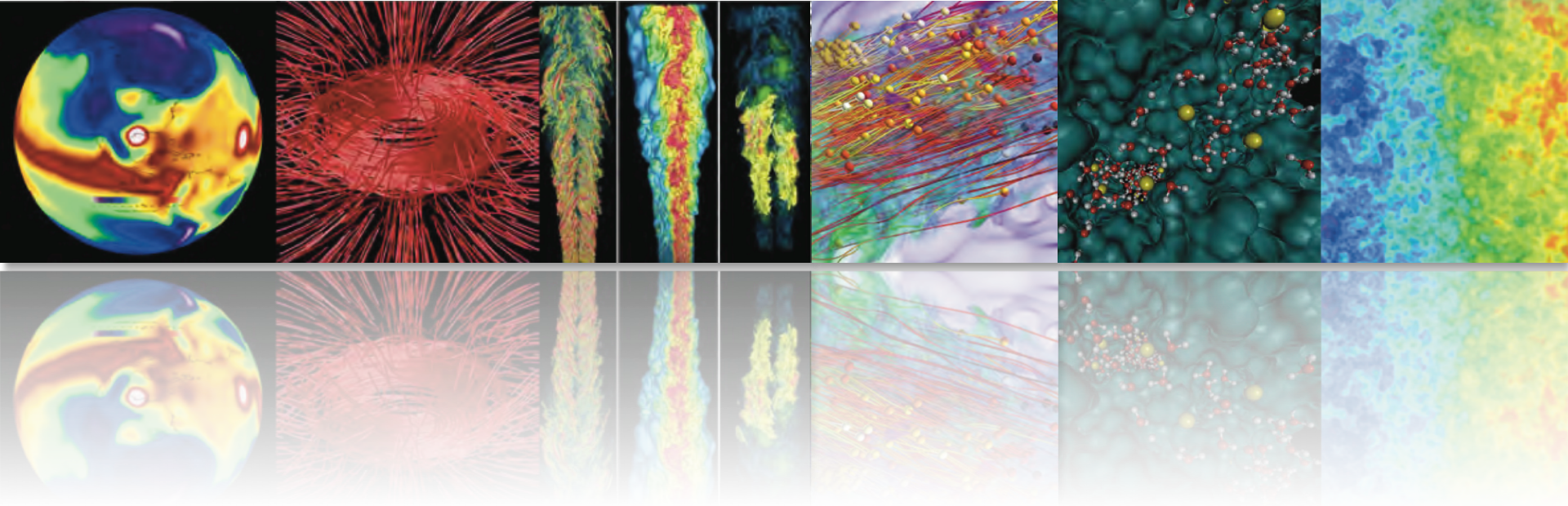
Acknowledgements

A host of people too numerous to mention from:

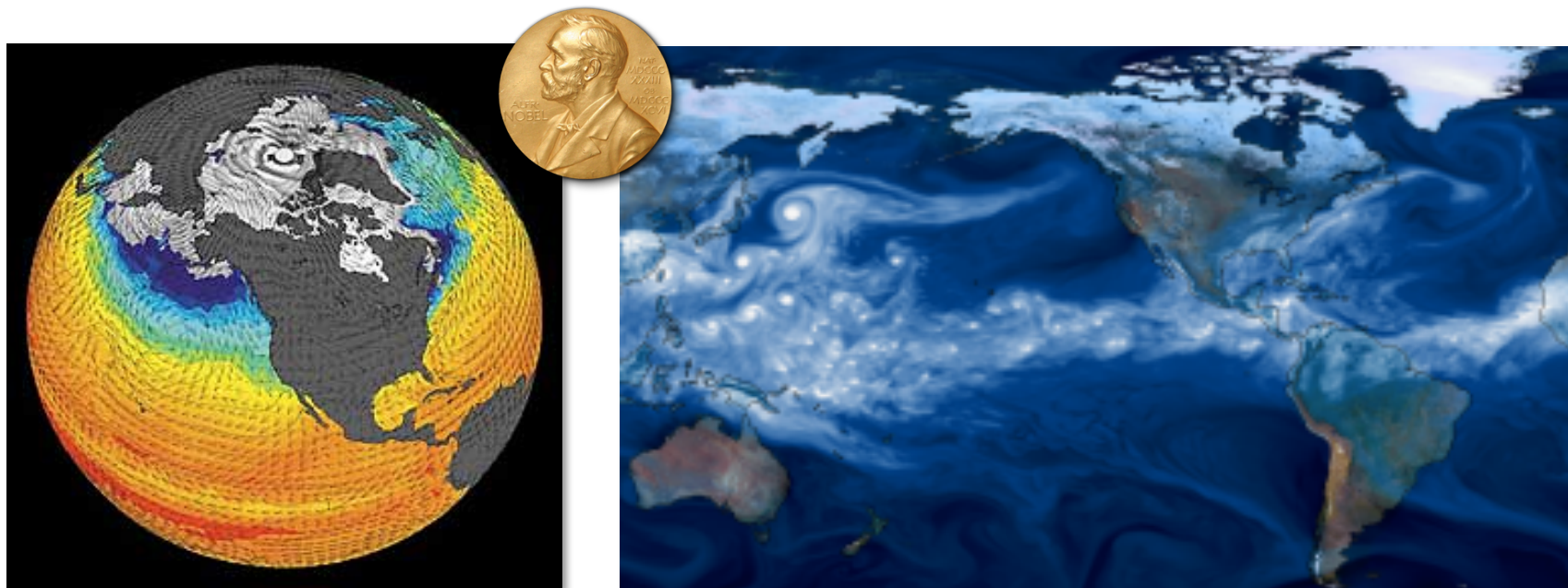
- DOE Office of Science and Advanced Scientific Computing Research
- DOE NNSA and Advanced Simulation and Computing
- Argonne National Laboratory
- Lawrence Berkeley National Laboratory
- Lawrence Livermore National Laboratory
- Los Alamos National Laboratory
- Pacific Northwest National Laboratory
- Oak Ridge National Laboratory
- Sandia National Laboratories
- Computing Vendors/Hardware Suppliers

“The Opportunities and Challenges of Exascale Computing,” Summary Report of the Advanced Scientific Computing Advisory Committee (ASCAC) Subcommittee, Fall 2010

Need for Advanced Computing



Climate change analysis



Simulations

- Cloud resolution, quantifying uncertainty, understanding tipping points, etc., will drive climate to exascale platforms
- New math, models, and systems support will be needed

Extreme data

- “Reanalysis” projects need 100× more computing to analyze observations
- Machine learning and other analytics are needed today for petabyte data sets
- Combined simulation/observation will empower policy makers and scientists

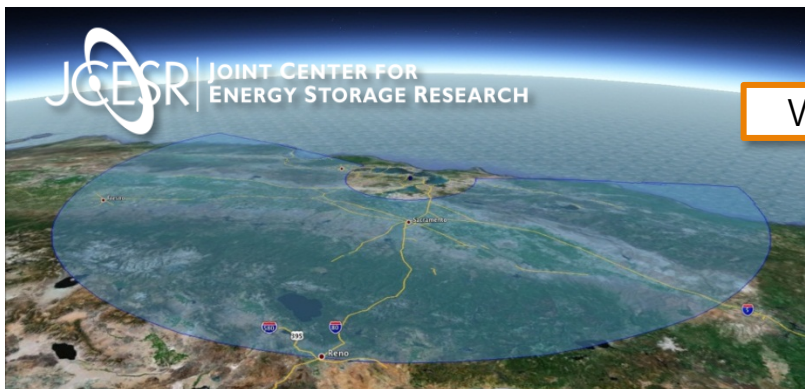
Materials genome

Computing 1000× today

- Key to DOE's Energy Storage Hub
- Tens of thousands of simulations used to screen potential materials
- Need more simulations and fidelity for new classes of materials, studies in extreme environments, etc.

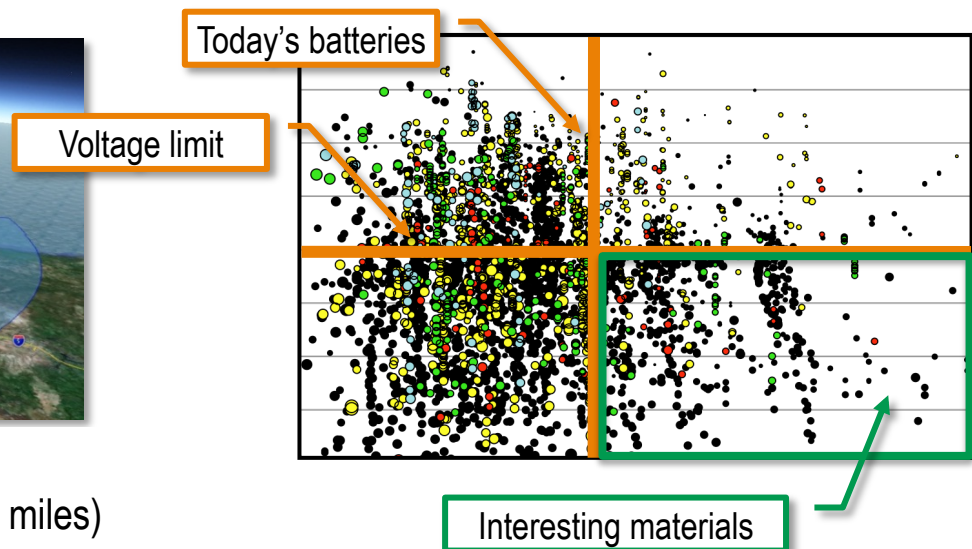
Data services for industry and science

- Results from tens of thousands of simulations web-searchable
- Materials Project launched in October 2012, now has >3,000 registered users
- Increase U.S. competitiveness; cut in half 18 year time from discovery to market



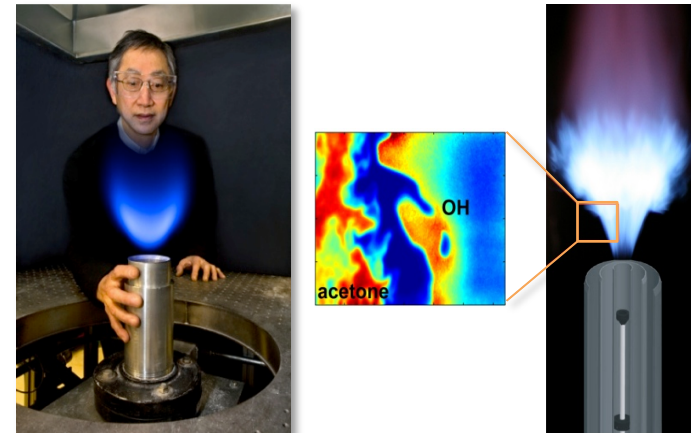
By 2018:

- Increase energy density (70 miles → 350 miles)
- Reduce battery cost per mile (\$150 → \$30)



Combustion simulations

- Transportation accounts for just under 40% of energy consumption in the US
- Goal: 50% improvement in engine efficiency
- Design challenge:
 - understand new combustion modes, control ignition timing, rate of pressure rise, etc.
- Science challenge:
 - low-temperature ignition kinetics and coupling with turbulence are poorly understood
- Approach:
 - perform multi-scale high-fidelity simulations to predict behavior under engine relevant conditions



High-fidelity combustion simulations require exascale computing to predict the behavior of alternative fuels in novel fuel-efficient, clean engines, and so facilitate design of optimal combined engine-fuel system

Multi-scale modeling and simulation of combustion at exascale is key for prediction and design

■ Engine design and optimization:

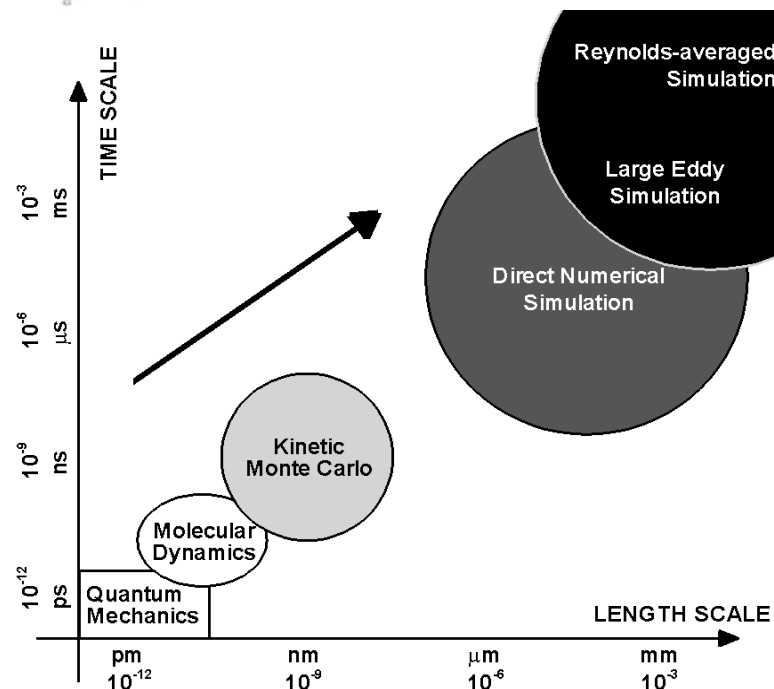
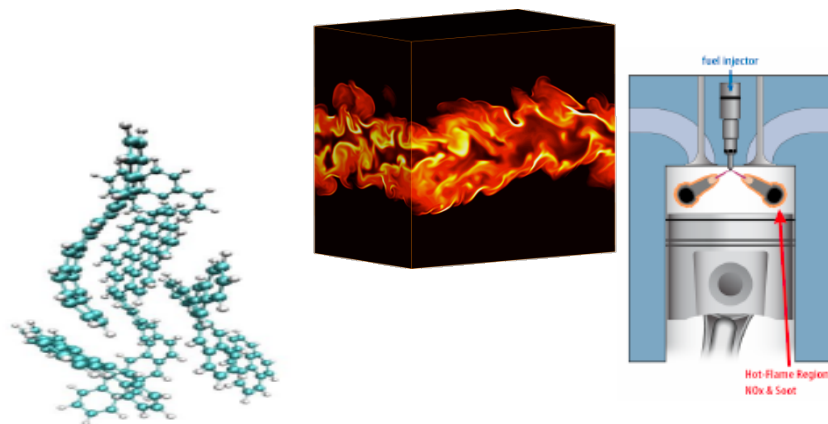
- Design and optimization of diesel/HCCI engines, in geometry with surrogate large-molecule fuels representative of actual

■ Direct numerical simulation:

- DNS of turbulent jet flames at engine pressures (30-50 atm) with iso-butanol (50-100 species) for diesel or HCCI engine thermochemical conditions

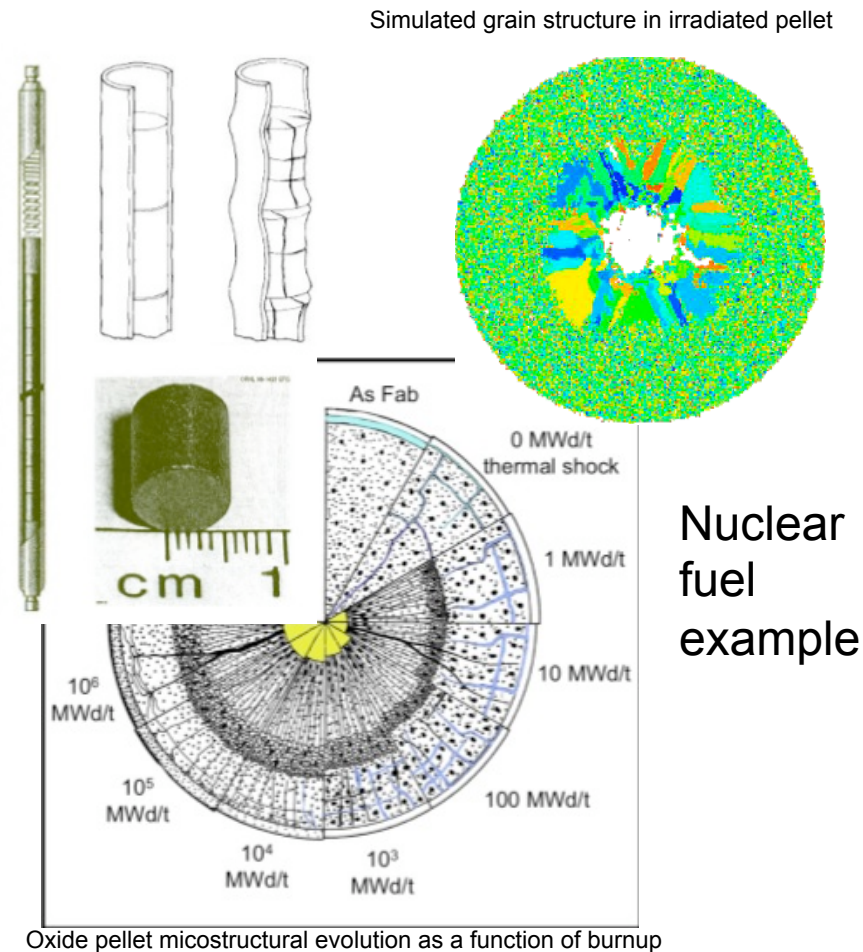
■ Molecular scale:

- Full-scale MD with on-the-fly ab-initio force field. Study combined physical and chemical processes affecting nanoparticle and soot formation processes. Size of the problem: >1000 molecules



Computational modeling is a critical component of U.S. Nuclear Energy strategy

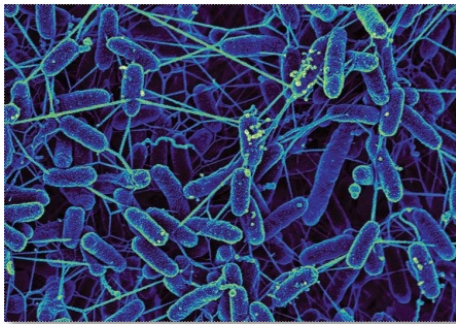
- Improved GEOMETRIC fidelity
 - Sub-10 μm resolution
 - UQ methodologies
 - 3D phenomena, microstructure evolution, material failure
 - Improved lower-length scale fidelity
- Improved NUMERICAL fidelity
 - Bridging vastly different time and length scales with multi-physics phenomena
 - Bubble/fission fragment interactions (MD)
 - Upscale oxide and metal models into pellet simulations
- Improved PHYSICS fidelity
 - Fission gas bubble formation, transport, and release
 - Fuel chemistry and phase stability
 - Fuel-cladding mechanical interaction
 - Thermal hydraulics, turbulence, and coolant flow in pin assembly \rightarrow effect on fuel and clad evolution



3D predictive simulations of fuel pin behavior from microstructure evolution will require exascale resources.

DOE Systems Biology Knowledgebase: KBase

- Integration and modeling for predictive biology
- Knowledgebase enabling predictive systems biology
 - Powerful modeling framework
 - Community-driven, extensible and scalable open-source software and application system
 - Infrastructure for integration and reconciliation of algorithms and data sources
 - Framework for standardization, search, and association of data
 - Resource to enable experimental design and interpretation of results



Microbes



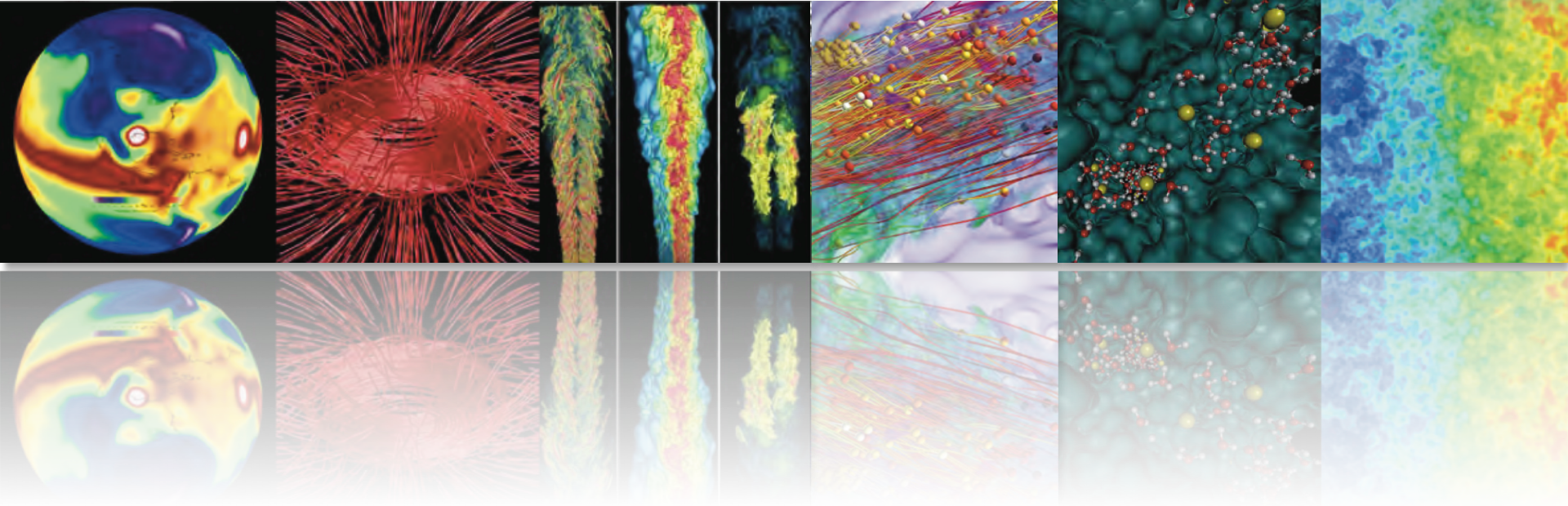
Communities



Plants



Advanced Computing is Critical to a Nation

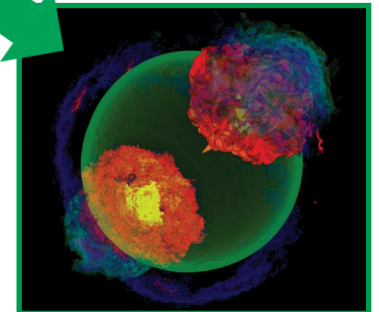
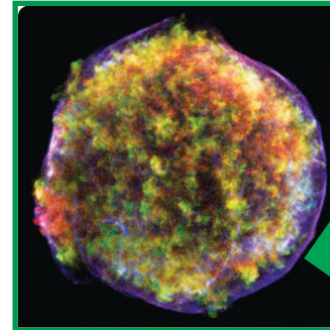


Advanced computing is driven by three national priorities

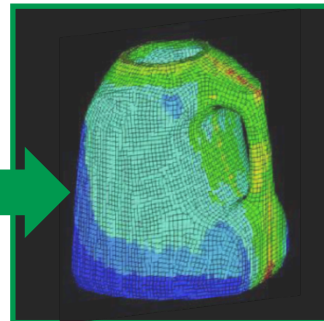
National security



Scientific leadership



Economic competitiveness



Competitive advantage demands advanced computing resources

- Digital design and prototyping through advanced computing enables rapid delivery of new products to market by minimizing the need for expensive, dangerous, and/or inaccessible testing
- Potential key differentiator for innovation
- Shrinks time to insight and shrinks time to solution



China – Investing for world leadership



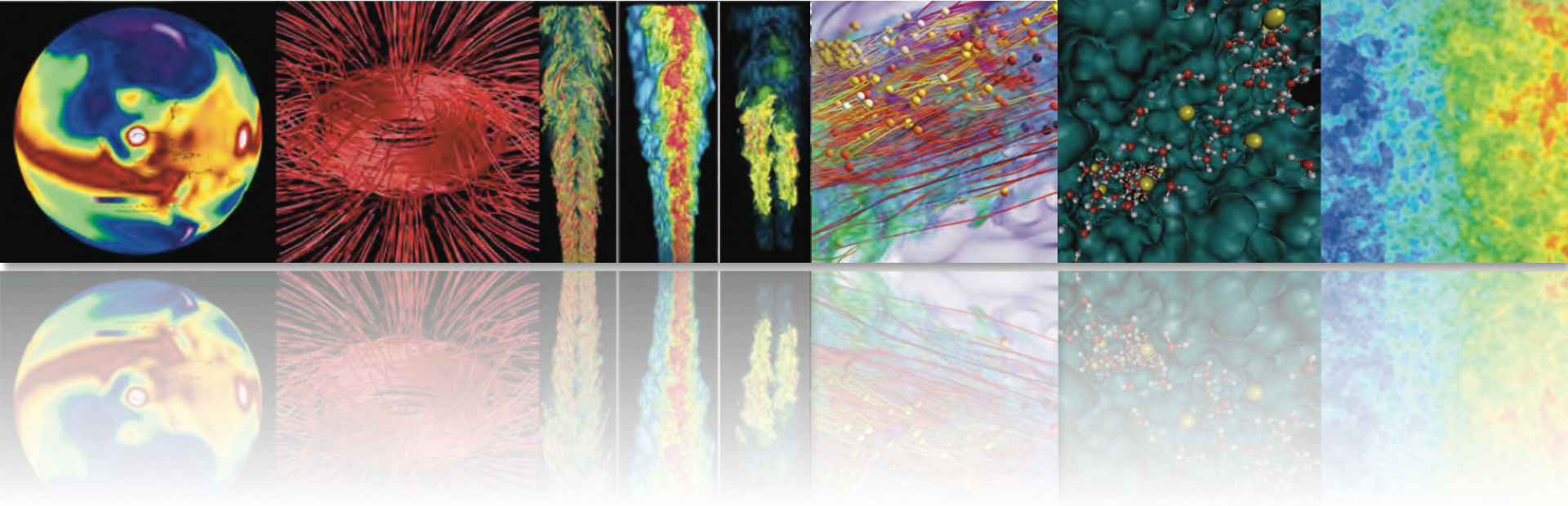
- Building out a substantial HPC infrastructure, including companies
- Developing indigenous HPC capabilities
- Software development currently lags hardware; this is being addressed

EU – seeking to establish an indigenous computer industry



- EU buys 30% of the HPC systems
- Seeking to develop HPC industry
- Initiated \$63M in platform exascale R&D
- Partnership for Advanced Computing in Europe (PRACE) provides coordinated HPC infrastructure

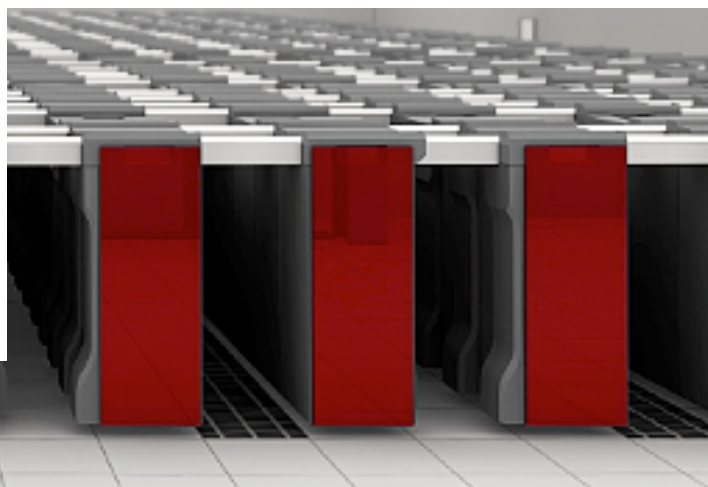
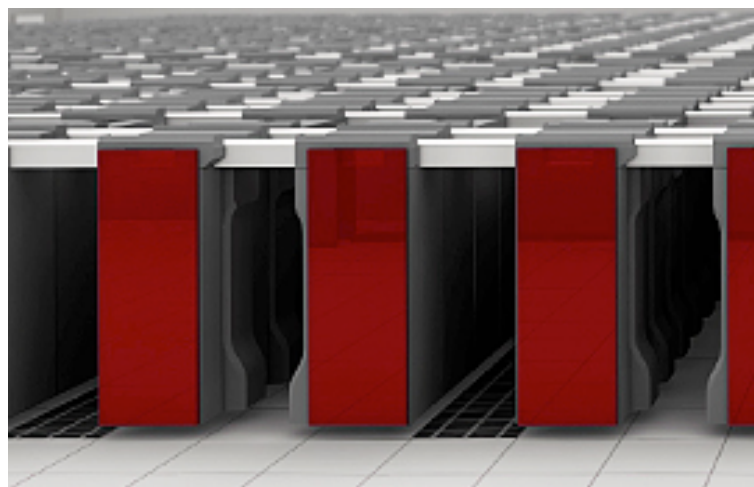
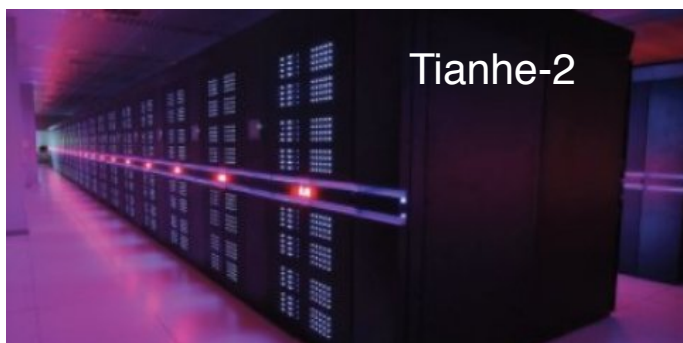
Challenges of Advanced Computing



Machines with 100,000+ cores are the current state of the art in advanced computing

What science can they enable?

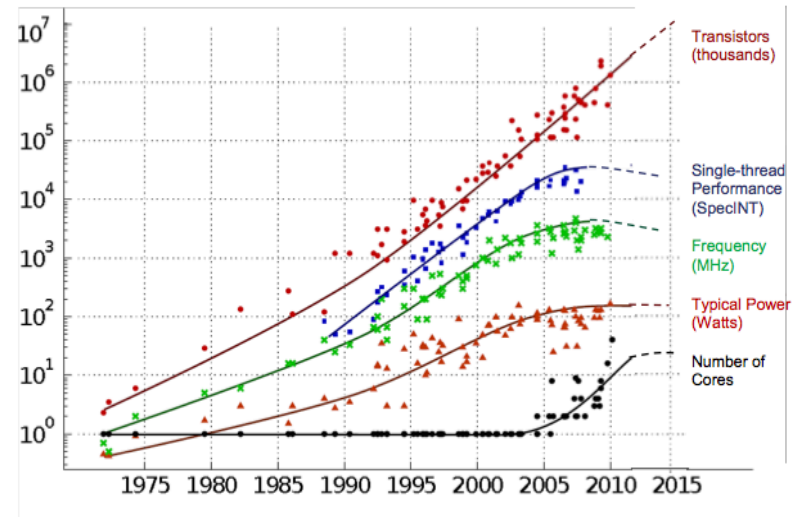
How does one get the most out of these significant resources?



Barriers to Exascale

- Power consumption
 - Goal is a factor of **5** from industry Business as Usual (BAU)
- Memory and storage bandwidth
 - Goal is a factor of **5** from industry BAU
- Reliability and resilience
 - Goal is a factor of **10** from industry BAU
- Scalability of systems software
 - Goal is a factor of **100** from industry BAU
- Programming models and environments
 - Goal is a factor of **10** productivity over today's mixed models while increasing parallelism in applications
 - by a factor of **1000**

35 YEARS OF MICROPROCESSOR TREND DATA



Original data collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond and C. Batten
Dotted line extrapolations by C. Moore

Power Consumption

■ Barriers

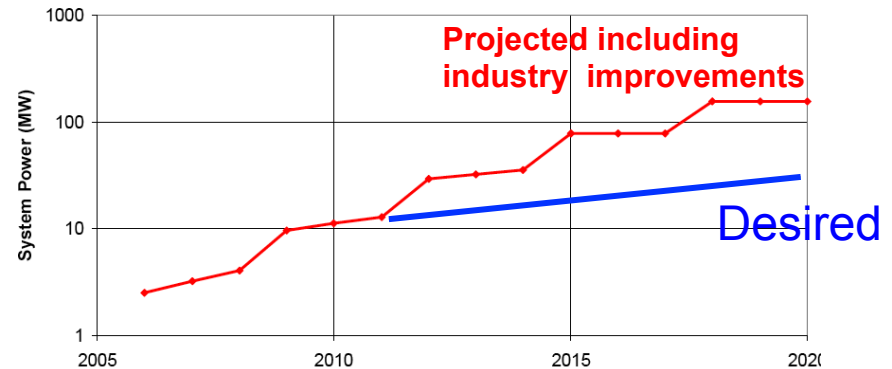
- Power is leading design constraint for computing technology
- Target ~20MW, estimated > 100MW required for Exascale systems (DARPA, DOE)
- Efficiency is industry-wide problem (IT technology >2% of US energy consumption and growing)

■ Technical Focus Areas

- Energy efficient hardware building blocks (CPU, memory, interconnect)
- Novel cooling and packaging
- Si-Photonic Communication
- Power Aware Runtime Software and Algorithms

■ Technical Gap

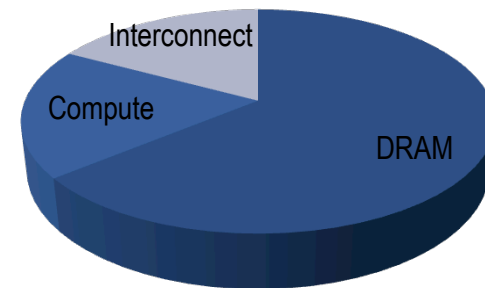
- Need **5X** improvement in power efficiency over projections that include technological advancements



Possible Leadership class power requirements

From Peter Kogge (on behalf of Exascale Working Group), "Architectural Challenges at the Exascale Frontier", June 20, 2008

Projected Power Usage



System memory dominates energy budget

Memory and Storage Bandwidth

■ Barriers

- Per-disk performance, failure rates, and energy efficiency no longer improving
- Linear extrapolation of DRAM vs. Multi-core performance means the height of the memory wall is accelerating
- Off-chip bandwidth, latency throttling delivered performance

■ Technical Focus Areas

- *Efficient Data Movement*
 - Photonic DRAM interfaces
 - Optical interconnects / routers
 - Communications optimal algorithms
- *New Storage Approaches*
 - Non-volatile memory gap fillers
 - Advanced packaging (chip stacking)
 - Storage efficient programming models (Global Address Space)

■ Technical Gap

- Need **5X** improvement in memory access speeds to keep current balance with computation

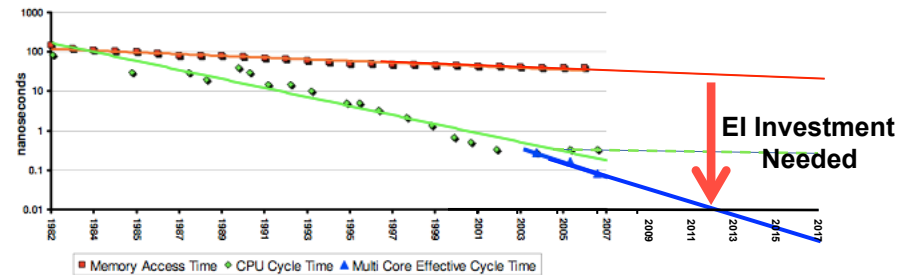
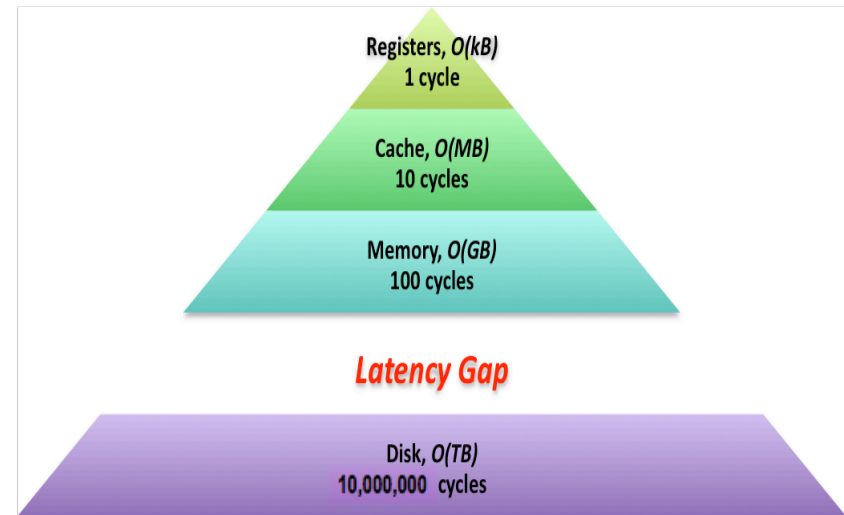


Figure 6.12: CPU and memory cycle time trends.



Reliability and Resilience

Barriers

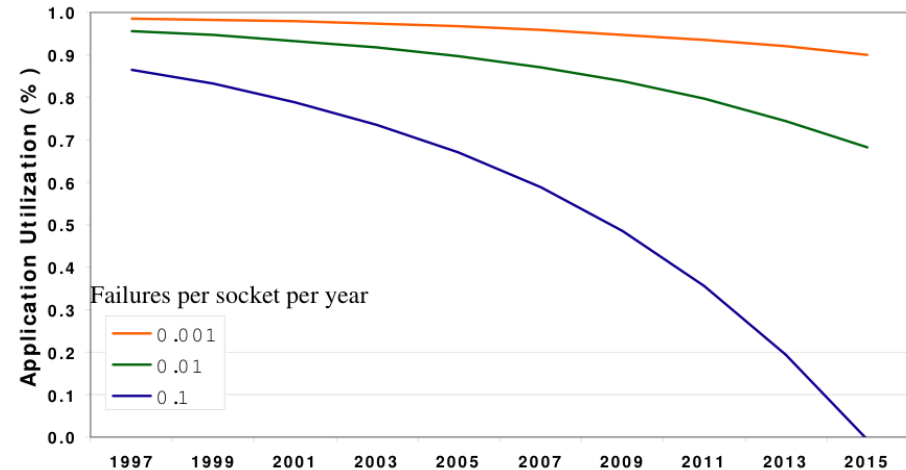
- Number of system components increasing faster than component reliability
- Mean time between failures of minutes or seconds for exascale
- Silent error rates increasing
- No job progress due to fault recovery if we use existing checkpoint/restart

Technical Focus Areas

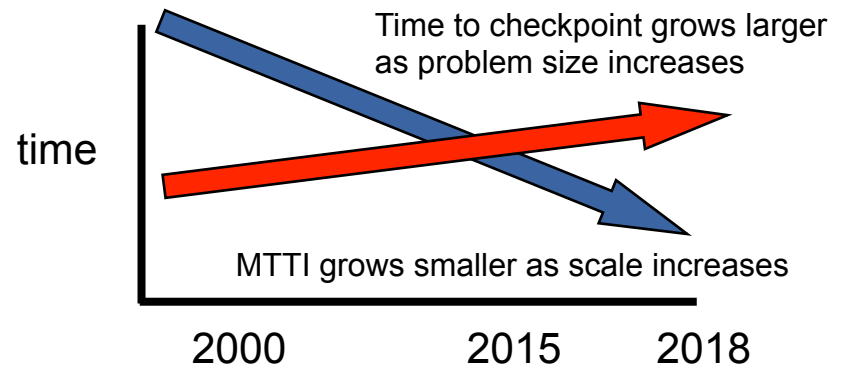
- Improved hardware and software reliability
 - Better Reliability, Availability and Serviceability (RAS) collection and analysis (root cause)
 - Greater integration
- Fault resilient algorithms and applications
- Local recovery and migration

Technical Gap

- Need 100X improvement in MTTI so that applications can run for many hours. Goal is **10X** improvement in hardware reliability. Local recovery and migration may yield another 10X. However, for exascale, applications will need to be fault resilient



Effective application utilization (including checkpoint overhead) at 3 rates of hardware failure



By Exascale, checkpoint/restart no longer viable

System Software Scalability

■ Barriers

- Fundamental assumptions of system software architecture did not anticipate exponential growth in parallelism
- Requirements for resilience at scale
- IO wall reducing effectiveness of simulation environment

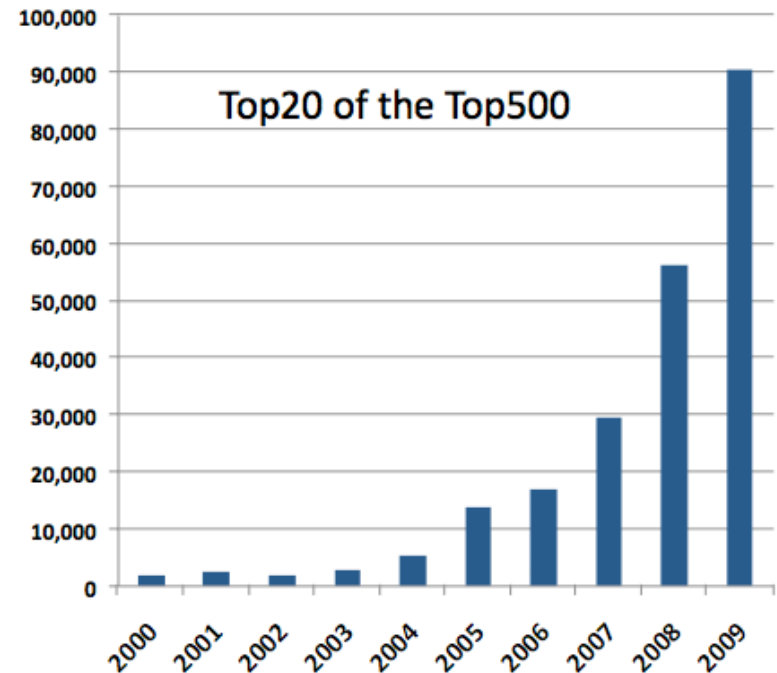
■ Technical Focus Areas

- System Hardware Manageability
- System Software Scalability
- Applications Scalability
- Supporting investments in infrastructure to support systems
- Initial deliveries to validate software and operations path

■ Technical Gap

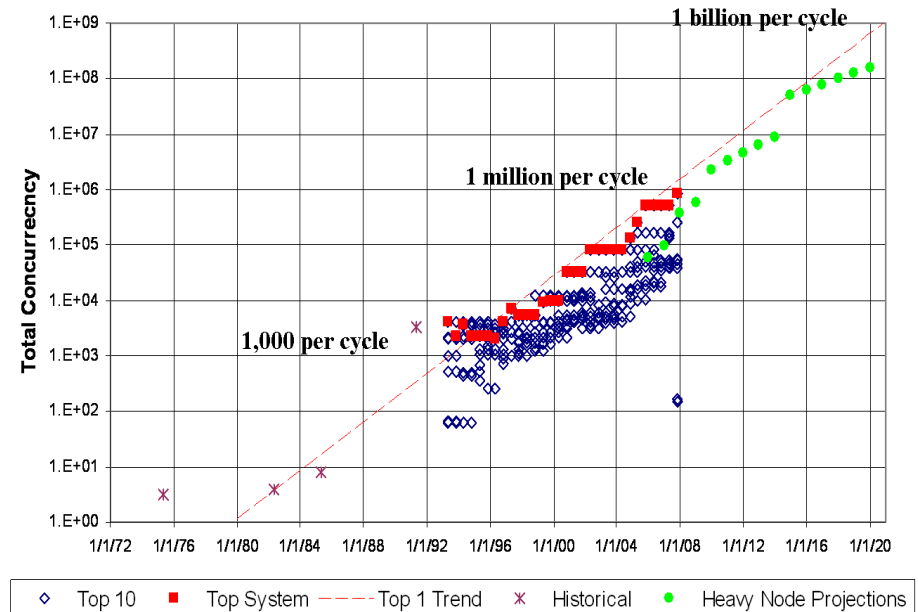
- **1000x** improvement in system software scaling
- **100x** improvement in system software reliability
- Need application hooks into Reliability, Availability and Serviceability (RAS) system

Average Number of Cores Per Supercomputer



Programming Models and Environments

- **Barriers-** Delivering a complex large-scale scientific instrument that is productive and fast.
 - O(1B) way parallelism in Exascale system
 - O(1K) way parallelism in a processor
 - Effective management of locality
 - Complexity of scientific applications
 - Programming for resilience
 - Science goals require complex codes



How much parallelism must be handled by the program?

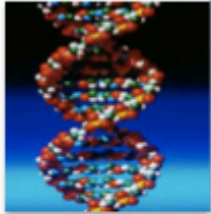
From Peter Kogge (on behalf of Exascale Working Group), "Architectural Challenges at the Exascale Frontier", June 20, 2008

■ Technical Focus

- Evolutionary: extend existing models used in science for scalability and to hide system complexity, e.g., heterogeneity and failures
- Moderate: leverage emerging models in scientific computing
- Revolutionary: develop a new paradigm for high usability at extreme scales

- **Technical Gap:** Productivity, Performance and Correctness for **1000x** more parallelism while increasing programming productivity of scientists by **10x**

Extreme Scale Science is Causing a Data Explosion

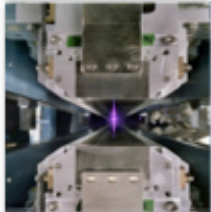


Genomics

Data Volume increases to 10 PB in FY21

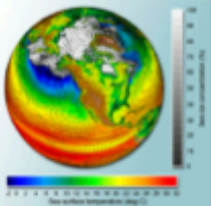


High Energy Physics (Large Hadron Collider)
15 PB of data/year



Light Sources

Approximately 300 TB/day



Climate

Data expected to be hundreds of 100 EB

Driven by exponential technology advances

Data sources

- Scientific Instruments
- Sensors/Treaty monitoring
- Scientific Computing Facilities
- Simulation Results

Big Data and Big Compute

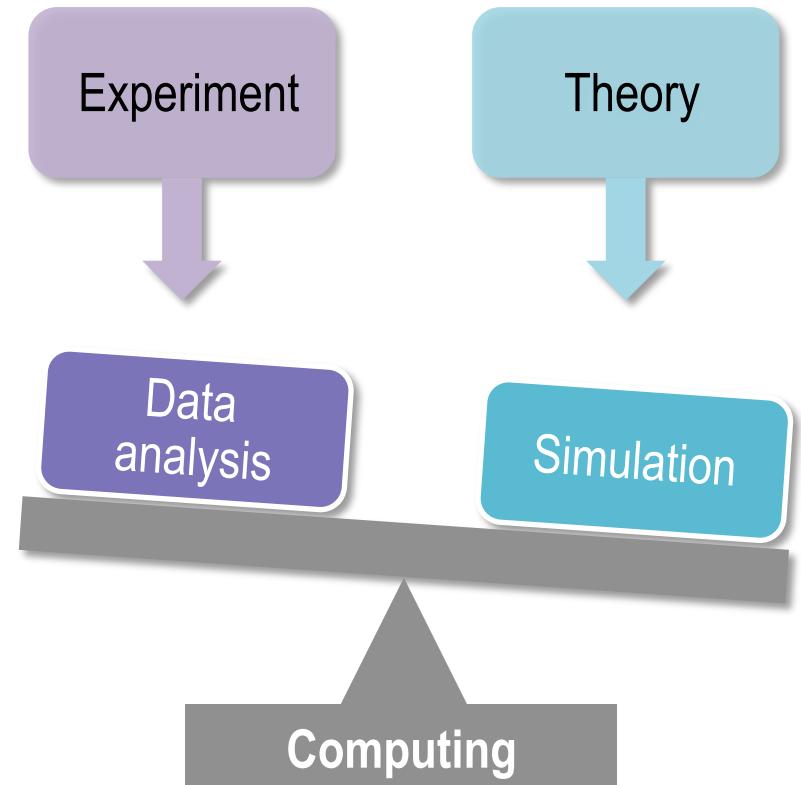
- Analyzing Big Data requires processing (e.g., search transform, analyze,...)
- Extreme scale computing will enable timely and more complex processing of increasingly large Big Data sets

“Very few large scale applications of practical importance are NOT data intensive.” – Alok Choudhary, IESP, Kobe, Japan, April 2012

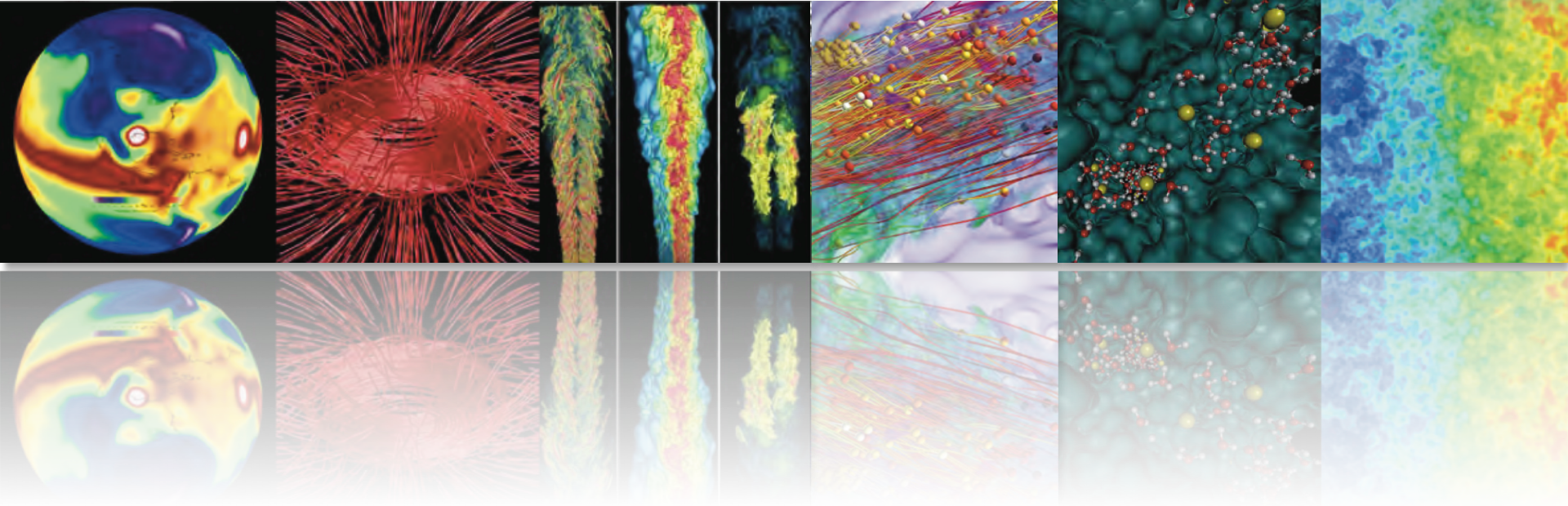
Data analysis and simulation both require Exascale

- Simulation and data are critical to DOE
- Both need more computing capability
- Both have similar HW technology requirements
 - High bandwidth to memory
 - Efficient processing
 - Very fast I/O
- Different machine balance may be required

- Big data: Analyzing and managing large complex data sets from experiments, observation, or simulation and sharing them with a community
- Simulation: Used to implement theory; helps with understanding and prediction



The Workload Manager's Role in Advanced Computing



Resource management requirements at extreme scale

- Need to provide extreme scalability
- Need to address memory locality
- Need to provide fault tolerance
- Need to manage heterogeneity of resources
- And meet of all these needs while under a strict power budget

Managing power and energy

- Schedule power
 - Schedule jobs to collectively remain within a power envelope
 - Schedule based on time-varying power costs
 - Reduce power consumption for less urgent jobs
- Schedule based on power sources with awareness of resources sharing the same PDUs
 - Help balance a power load
 - Make it easier to maintain a consistent machine room temperature

Managing memory and storage concurrency and locality

- Memory is now the expensive component versus processors
- Memory affinity becomes crucial
- Schedule I/O use
 - Schedule the staging of data before allocating computing resources
 - Schedule data storage after releasing allocated resources

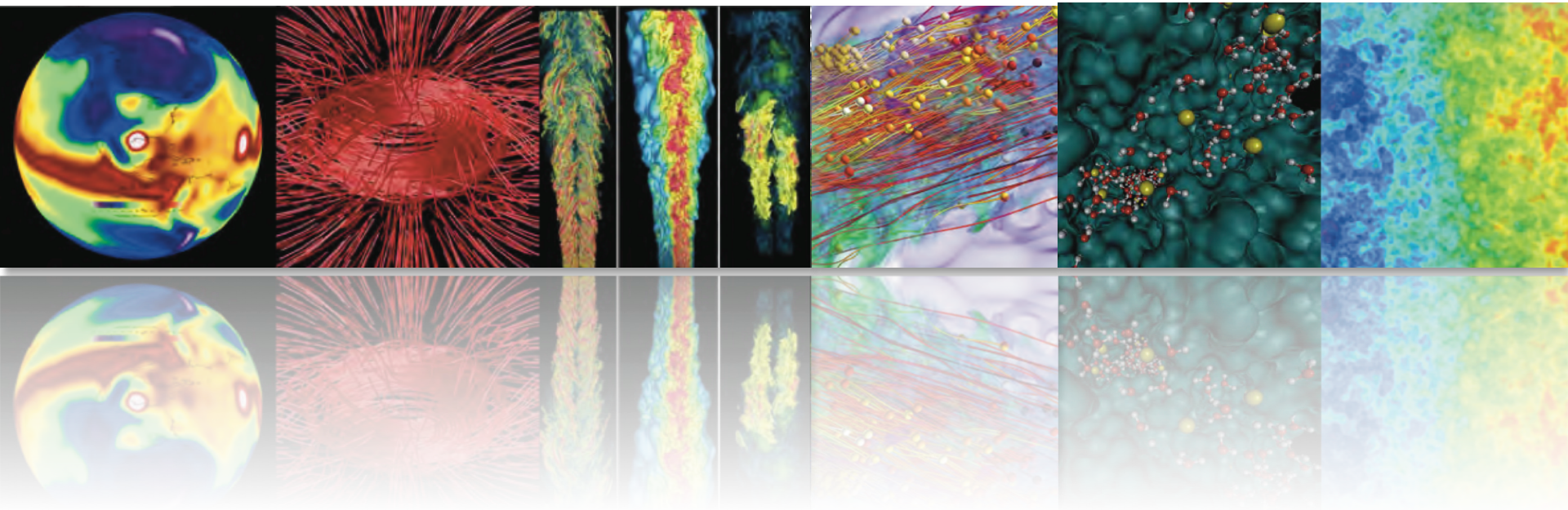
Enabling reliability and resiliency

- Schedule replacement resources when failures are imminent or detected
- Enable dynamic resource allocation to allow allocation of replacement to currently running job
- Incorporate ways to measure a job's progress so hung or erroneous computations are flagged or terminated
- Aggregate a rich set of resource and job statistics to help correlate factors responsible for job failures
 - Ideally include system monitoring info
 - Could provide the necessary evidence to recreate or replay job anomalies
- Minimize down time - allow software updates or reconfiguration to happen across allocated resources

Ease of Use

- Make the process of running a job as simple for the user as possible.
- Reduce the need for the user to ask any of these questions:
 - Why was my job submission rejected?
 - Why is my job not running?
 - Why did my job terminate sooner than I expected?
- Provide a set of graphical tools that allows the user to navigate the plethora of job specification options and run their jobs following simple and intuitive patterns.
- Provide a way to auto-tune the scheduler configuration to achieve specific goals: availability, throughput, efficient use of resources, etc.

Summary



Summary

- The need for advanced computing is increasing
- The technology required to meet that need is changing dramatically
- This presents challenges and opportunities

Thank You

