INTERNATIONAL
SUPERCOMPUTING CONFERENCE

Exascale Computing Panel

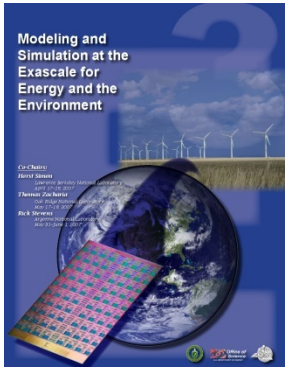
Jack Dongarra
University of Tennessee & Oak Ridge National Laboratory, USA

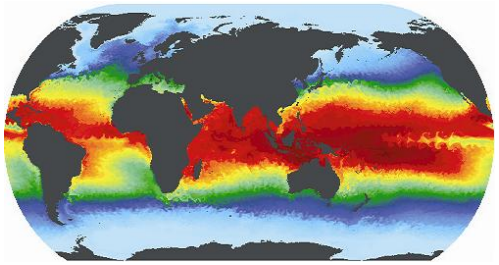
Questions for the panelists

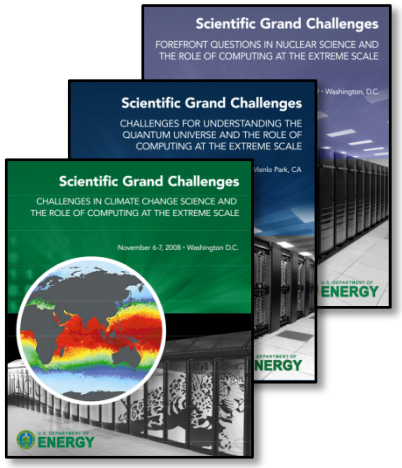
- What application of Exascale computing could justify such a huge investment?

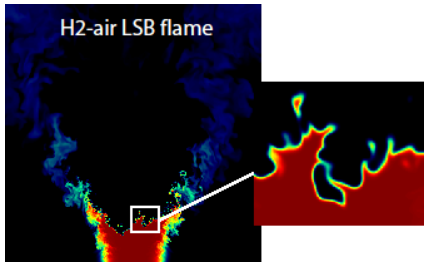
Broad Community Input

- Town Hall Meetings April-June 2007
- Scientific Grand Challenges Workshops Nov, 2008 – Oct, 2009
  - Climate Science (11/08),
  - High Energy Physics (12/08),
  - Nuclear Physics (1/09),
  - Fusion Energy (3/09),
  - Nuclear Energy (5/09),
  - Biology (8/09),
  - Material Science and Chemistry (8/09),
  - National Security (10/09)
- Exascale Steering Committee
  - "Denver" vendor NDA visits 8/2009
  - Extreme Architecture and Technology Workshop 12/2009
  - Cross-cutting workshop 2/2010
- International Exascale Software Project
  - Santa Fe, NM 4/2009
  - Paris, France 6/2009
  - Tsukuba, Japan 10/2009
  - Oxford, UK, 4/2010

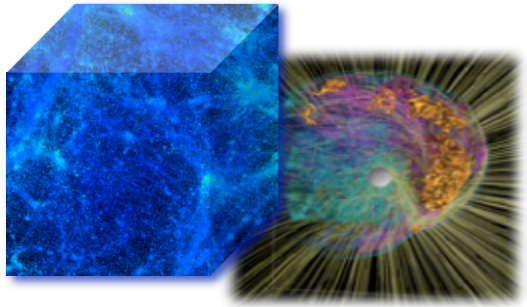








MISSION IMPERATIVES



FUNDAMENTAL SCIENCE

Science and Engineering Drivers

- Climate
- Nuclear Energy
- Combustion
- Advanced Materials
- CO_2 Sequestration
- Basic Science

- Common Needs
  - Multiscale
  - Uncertainty Quantificc
  - Rare Event Statistics

The Exascale Draft plan has Four High-Level Components

- **Science and engineering mission applications**
- **Systems software, tools and programming models**
- **Computer hardware and technology development**
- **Systems acquisition, deployment and operations**

The plan is currently under consideration for a national initiative to begin in 2012

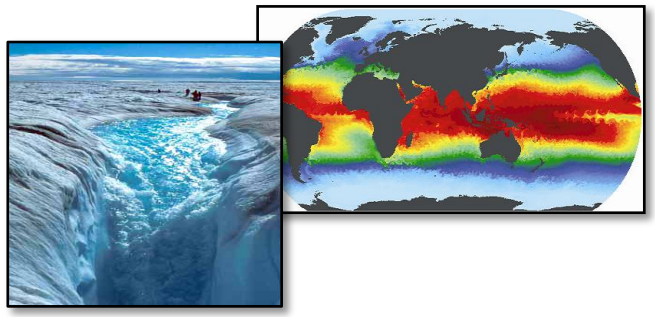
Three early funding opportunities have been release by DOE this spring to support preliminary research

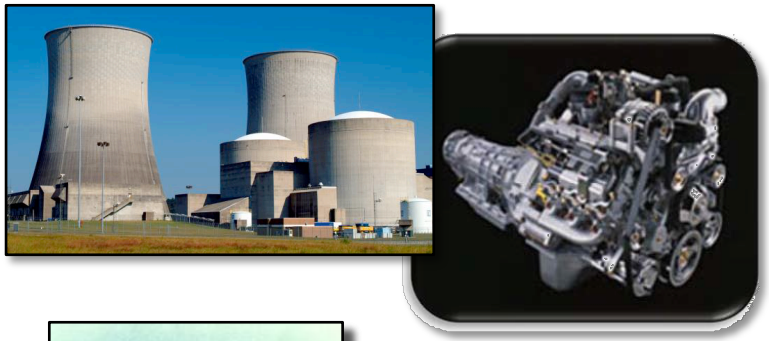
The plan targets exascale platform deliveries in 2018 and a robust simulation environment and science and mission applications by 2020

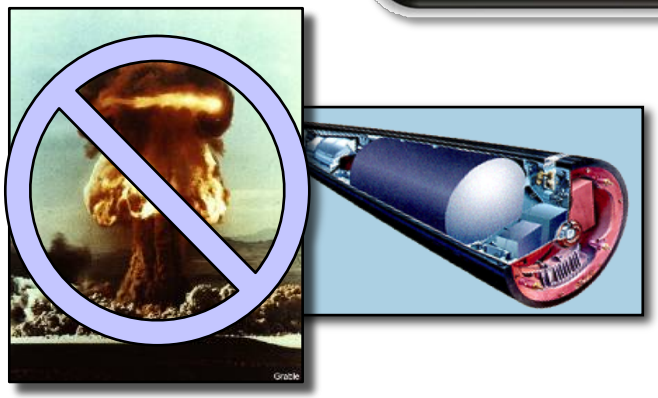
Co-design and co-development of hardware, system software, programming model and applications requires intermediate (~200 PF/s) platforms in 2015

DOE mission imperatives require simulation and analysis for policy and decision making

- *Climate Change*: Understanding, mitigating and adapting to the effects of global warming
  - Sea level rise
  - Severe weather
  - Regional climate change
  - Geologic carbon sequestration
- *Energy*: Reducing U.S. reliance on foreign energy sources and reducing the carbon footprint of energy production
  - Reducing time and cost of reactor design and deployment
  - Improving the efficiency of combustion energy sources
- *National Nuclear Security*: Maintaining a safe, secure and reliable nuclear stockpile
  - Stockpile certification
  - Predictive scientific challenges
  - Real-time evaluation of urban nuclear detonation





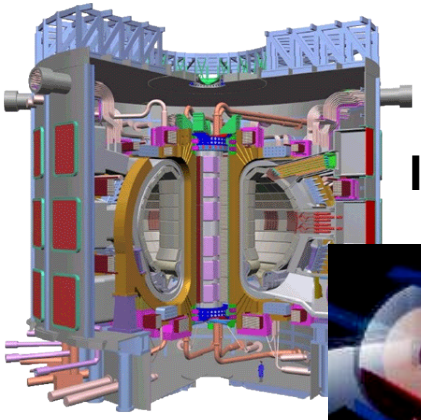


Accomplishing these missions requires exascale resources.

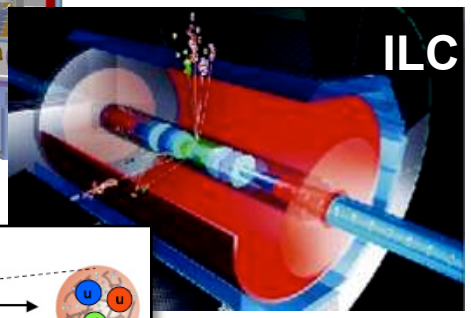
Simulation enables fundamental advances in basic science.

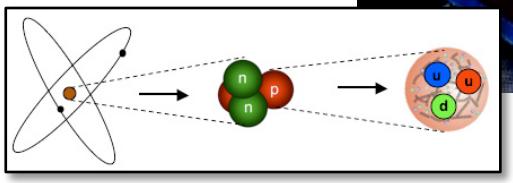
- **Nuclear Physics**
  - Quark-gluon plasma & nucleon structure
  - Fundamentals of fission and fusion reactions
- **Facility and experimental design**
  - Effective design of accelerators
  - Probes of dark energy and dark matter
  - ITER shot planning and device control
- **Materials / Chemistry**
  - Predictive multi-scale materials modeling: observation to control
  - Effective, commercial, renewable energy technologies, catalysts and batteries
- **Life Sciences**
  - Better biofuels
  - Sequence *to* structure *to* function

These breakthrough scientific discoveries and facilities require exascale applications and resources.

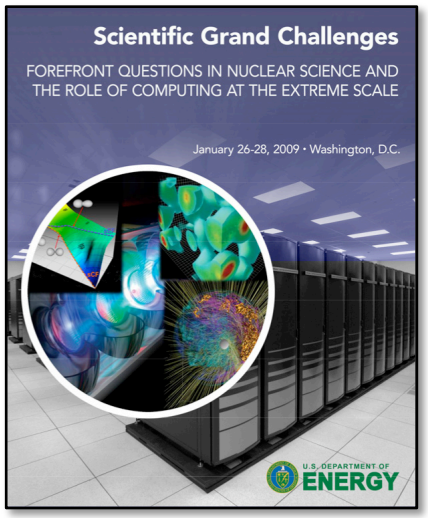


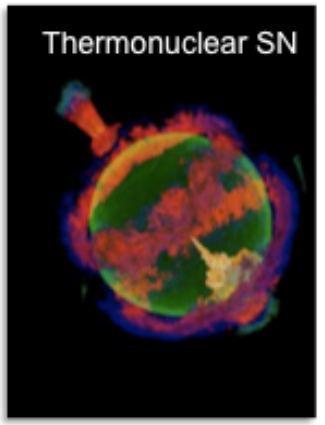
ITER





Structure of nucleons





2. Extrapolating the TOP500 predicts an exascale system in 2018 time frame. Can we simply wait for an exascale system to appear in 2018 without doing anything out of the ordinary?

biggest challenges

- Increasing imbalance among processor speed, interconnect bandwidth, and system memory
- Memory management will be a significant challenge for exascale science applications due to their deeper, complex hierarchies and relatively smaller capacities, and dynamic, latency tolerant approaches must be developed
- Software will need to manage resilience issues more actively at the exascale
- Automated, dynamic control of system resources will be required
- exascale programming paradigms to support 'billion-way' concurrency

What are critical exascale technology investments?

- **System power** is a first class constraint on exascale system performance and effectiveness.
- **Memory** is an important component of meeting exascale power and applications goals.
- **Programming model**. Early investment in several efforts to decide in 2013 on exascale programming model, allowing exemplar applications effective access to 2015 system for both mission and science.
- **Investment in exascale processor design** to achieve an exascale -like system in 2015.
- **Operating System strategy for exascale** is critical for node performance at scale and for efficient support of new programming models and run time systems.
- **Reliability and resiliency are critical at this** scale and require applications neutral movement of the file system (for check pointing, in particular) closer to the running apps.
- ***HPC co-design strategy and implementation*** *requires a set of a hierarchical performance models and simulators as well as commitment from apps, software and architecture communities.*

Major Changes to Software

- **Must rethink the design of our software**
  - Another disruptive technology
    - Similar to what happened with cluster computing and message passing
  - Rethink and rewrite the applications, algorithms, and software
- **Numerical libraries for example will change**
  - For example, both LAPACK and ScaLAPACK will undergo major changes to accommodate this

Five Important Features to Consider When Computing at Scale

1. Effective Use of Many-Core and Hybrid architectures
   - Break fork-join parallelism
   - Dynamic Data Driven Execution
   - Block Data Layout
2. Exploiting Mixed Precision in the Algorithms
   - Single Precision is 2X faster than Double Precision
   - With GP-GPUs 10x
   - Power saving issues
3. Self Adapting / Auto Tuning of Software
   - Too hard to do by hand
4. Fault Tolerant Algorithms
   - With 1,000,000's of cores things will fail
5. Communication Reducing Algorithms
   - For dense computations from O(n *log* p) to O(*log* p) communications
   - Asynchronous iterations
   - GMRES k-step compute (x, Ax, A^2x, … A^kx)

A Call to Action

- Hardware has changed dramatically while software ecosystem has remained stagnant
- Need to exploit new hardware trends (e.g., manycore, heterogeneity) that cannot be handled by existing software stack, memory per socket trends
- Emerging software technologies exist, but have not been fully integrated with system software, e.g., UPC, Cilk, CUDA, HPCS
- Community codes unprepared for sea change in architectures
- No global evaluation of key missing components

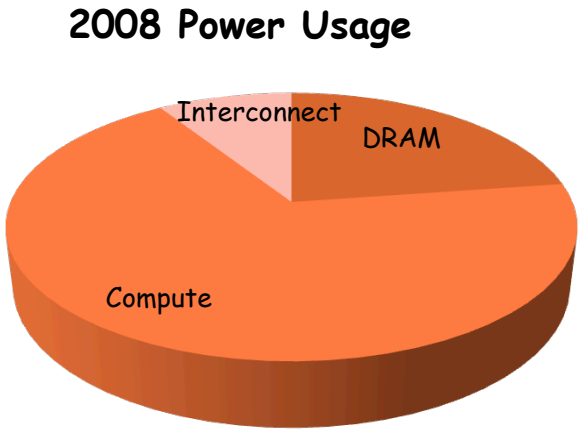
3. What are the principal hardware and software challenges in getting to a useable, 20MW exascale system in 2018?

Exascale Systems Targets

Systems	2010	2018	Difference Today & 2018
System peak	**2 Pflop/s**	**1 Eflop/s**	**O(1000)**
Power	**6 MW**	**~20 MW (goal)**	
System memory	0.3 PB	32 - 64 PB	O(100)
Node performance	125 GF	1.2 or 15TF	O(10) – O(100)
Node memory BW	25 GB/s	2 - 4TB/s	O(100)
Node concurrency	12	O(1k) or O(10k)	O(100) – O(1000)
Total Node Interconnect BW	3.5 GB/s	200-400GB/s (1:4 or 1:8 from memory BW)	O(100)
System size (nodes)	18,700	O(100,000) or O(1M)	O(10) – O(100)
Total concurrency	225,000	O(billion) + [O(10) to O(100) for latency hiding]	O(10,000)
Storage Capacity	15 PB	500-1000 PB (>10x system memory is min)	O(10) – O(100)
IO Rates	0.2 TB	60 TB/s	O(100)
MTTI	days	O(1 day)	- O(10)

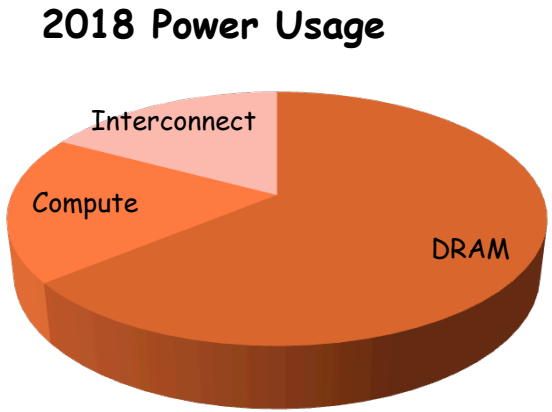
Memory Power Consumption

- **Power Consumption with standard Technology Roadmap**



70 Megawatts total

- **Power Consumption with Investment in Advanced Memory Technology**



20 Megawatts total

Reducing power is fundamentally about architecture choices and process technology

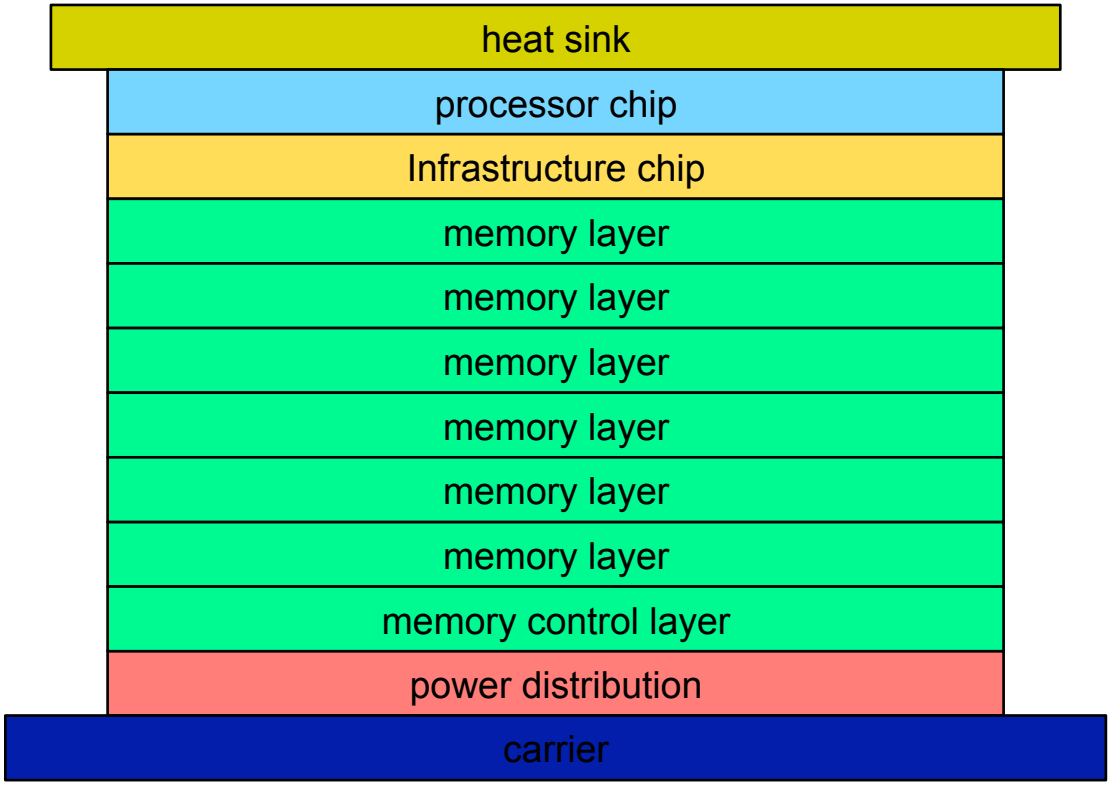
- **Memory (2x-5x)**
  - New memory interfaces (chip stacking and vias)
  - Replace DRAM with zero power non-volatile memory
- **Processor (10x-20x)**
  - Reducing data movement (functional reorganization, > 20x)
  - Domain/Core power gating and aggressive voltage scaling
- **Interconnect (2x-5x)**
  - More interconnect on package
  - Replace long haul copper with integrated optics
- **Data Center Energy Efficiencies (10%-20%)**
  - Higher operating temperature tolerance
  - Power supply and cooling efficiencies

The Path Forward

¨ **Research Needed to Achieve Exascale Performance**

- Extreme voltage scaling to reduce core power
- More parallelism 10x – 100x to achieve target speed
- Re-architecting DRAM to reduce memory power
- New interconnect for lower power at distance
- NVM to reduce disk power and accesses
- Resilient design to manage unreliable transistors
- New programming models for extreme parallelism
- Applications built for extreme (billion way) parallelism

A key for the next decade – exascale & otherwise - is the node



- 100x – 1000x more cores
- Heterogeneous cores
- New programming model

- 3d stacked memory

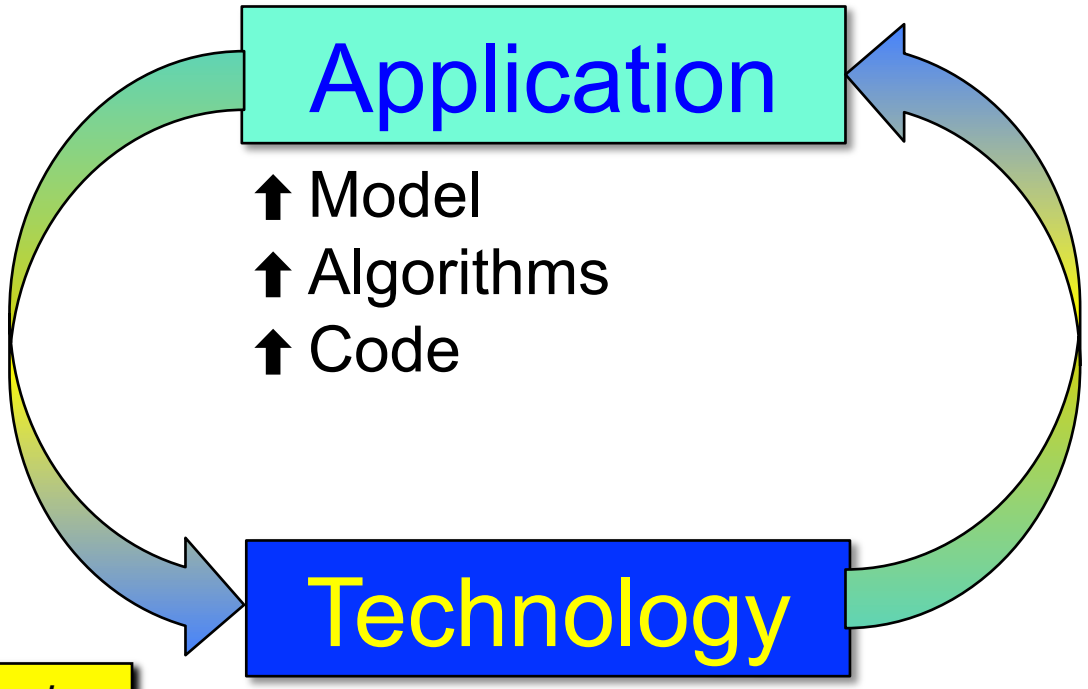
- Smart memory management

- Integration on package

4. What applications will be ready to run on an exascale system in 2018? What needs to be done over the next decade to develop these applications?

Co-design is a fundamental tenet of the initiative

Application driven:
Find the best technology to run this code.
Sub-optimal



⊕ architecture
⊕ programming model
⊕ resilience
⊕ power

Technology driven:
Fit your application to this technology.
Sub-optimal.

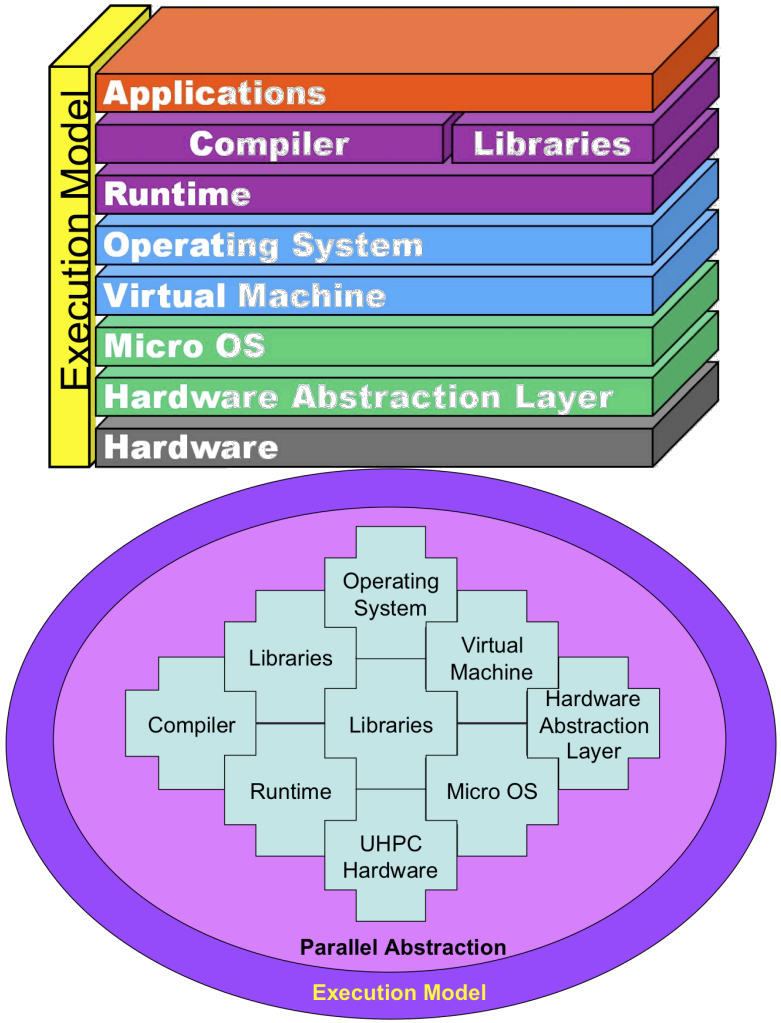
Now, we must expand the co-design space to find better solutions:
- *new applications & algorithms,*
- *better technology and performance.*

Co-Design

- Designing computer architectures and system configurations that will be both affordable and an appropriate match for current and future high-end science applications with reasonable implementation effort;
- Devising mathematical models, numerical software, programming models, and system software that enable implementation of complex simulations and achieve good performance on the new architectures.

System software as currently implemented is not suitable for exascale system

- **Barriers**
  - System management SW not parallel
  - Current OS stack designed to manage only O(10) cores on node
  - Unprepared for industry shift to NVRAM
  - OS management of I/O has hit a wall
  - Not prepared for massive concurrency
- **Technical Focus Areas**
  - Design HPC OS to partition and manage node resources to support massively concurrency
  - I/O system to support on-chip NVRAM
  - Co-design messaging system with new hardware to achieve required message rates
- **Technical gaps**
  - 10X: in affordable I/O rates
  - 10X: in on-node message injection rates
  - 100X: in concurrency of on-chip messaging hardware/software
  - 10X: in OS resource management



Software challenges in extreme scale systems, *Sarkar, 2010*

International Exascale Software Program

Improve the world's simulation and modeling capability by improving the coordination and development of the HPC software environment

Workshops:

Build an international plan for coordinating research for the next generation open source software for scientific high-performance computing

www.exascale.org

International Community Effort

- **We believe this needs to be an international collaboration for various reasons including:**
  - The scale of investment
  - The need for international input on requirements
  - US, Europeans, Asians, and others are working on their own software that should be part of a larger vision for HPC.
  - **No global evaluation of key missing components**
  - **Hardware features are uncoordinated with software development**

Four Goals for IESP

- Strategy for determining requirements
  - clarity in scope is the issue
- Comprehensive software roadmap
  - goals, challenges, barriers and options
- Resource estimate and schedule
  - scale and risk relative to hardware and applications
- A governance and project coordination model
  - Is the community ready for a project of this scale, complexity and importance?
  - Can we be trusted to pull this off?

Roadmap Components

www.exascale.org

INTERNATIONAL EXASCALE SOFTWARE PROJECT

ROADMAP

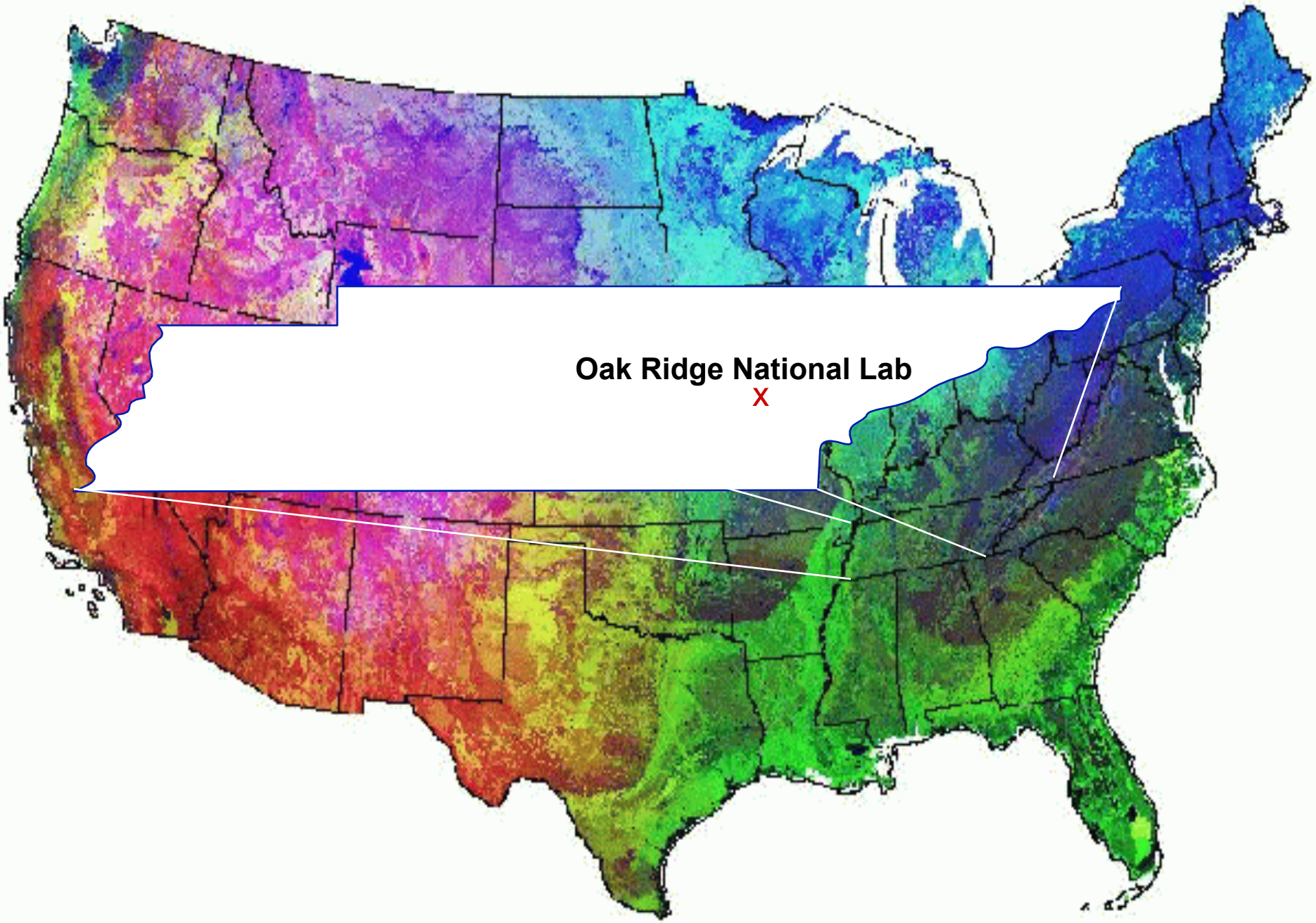
- www.exascale.org

Jack Dongarra, Pete Beckman, Terry Moore, Jean-Claude Andre, Jean-Yves Berthou, Taisuke Boku, Franck Cappello, Barbara Chapman, Xuebin Chi, Alok Choudhary, Sudip Dosanjh, Al Geist, Bill Gropp, Robert Harrison, Mark Hereld, Michael Heroux, Adolfy Hoisie, Koh Hotta, Yutaka Ishikawa, Fred Johnson, Sanjay Kale, Richard Kenway, David Keyes, Bill Kramer, Jesus Labarta, Bob Lucas, Barney Maccabe, Satoshi Matsuoka, Paul Messina, Bernd Mohr, Matthias Mueller, Wolfgang Nagel, Hiroshi Nakashima, Michael E. Papka, Dan Reed, Mitsuhisa Sato, Ed Seidel, John Shalf, David Skinner, Thomas Sterling, Rick Stevens, William Tang, John Taylor, Rajeev Thakur, Anne Trefethen, Marc Snir, Aad van der Steen, Fred Streitz, Bob Sugar, Shinji Sumimoto, Jeffrey Vetter, Robert Wisniewski, Kathy Yelick

www.exascale.org

SPONSORS

- 5. When will the first sustained exaflop/sec be achieved, on what code and where?


Oak Ridge National Lab
X