SOS 14

Challenges in Exascale Computing

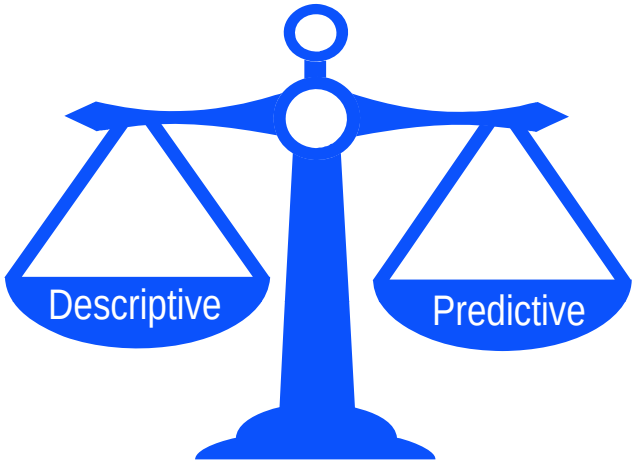
Acknowledgements

- **ANL** Rick Stevens, Pete Beckman, Ray Bair
- **BNL** Jim Davenport, Tom Shiagel
- **LBNL** Horst Simon, Kathy Yelick, John Shalf
- **LLNL** Mike McCoy, Mark Seager, Brent Gorda
- **LANL** Andy White, John Morrison, Cheryl Wampler
- **ORNL** Jeff Nichols, Al Geist, Jim Hack
- **PNNL** Steve Ashby, Moe Khaleel
- **SNL** Sudip Dosanjh, James Peery, Jim Ang

- **Ex officio** Fred Johnson, Paul Messina

The exascale draft plan has four high-level components

- **Science and mission applications**
- **Systems software and programming models**
- **Hardware technology R&D**
- **Systems acquisition, deployment and operations**



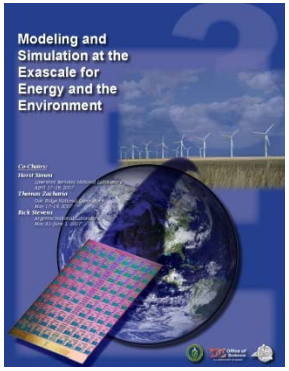
We are at the tipping point for predictive capability

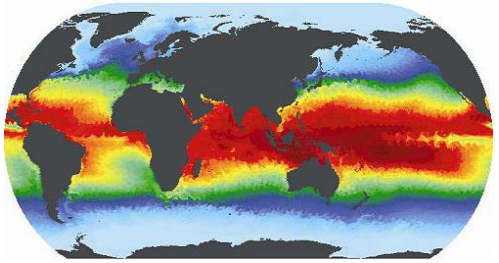
2018: The plan targets exascale platform deliveries in 2018 and a robust simulation environment and science and mission applications by 2020

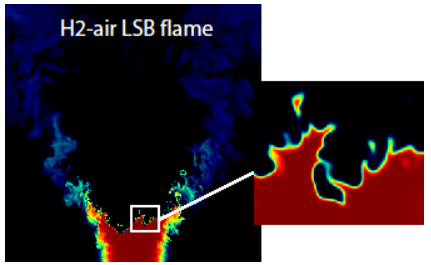
2015: Co-design and co-development of hardware, software, programming models and applications requires intermediate platforms in 2015

Process for identifying exascale applications and technology for DOE missions ensures broad community input

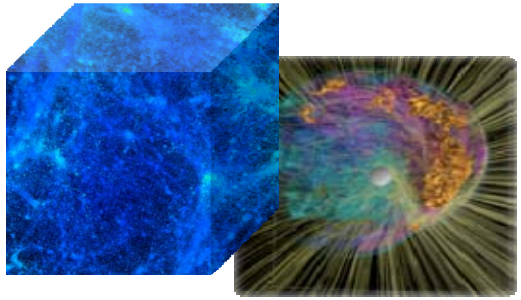
- **Town Hall Meetings April-June 2007**
- **Scientific Grand Challenges Workshops Nov, 2008 – Oct, 2009**
  - **Climate Science (11/08),**
  - **High Energy Physics (12/08),**
  - **Nuclear Physics (1/09),**
  - **Fusion Energy (3/09),**
  - **Nuclear Energy (5/09),**
  - **Biology (8/09),**
  - **Material Science and Chemistry (8/09),**
  - **National Security (10/09)**
  - **Cross-cutting technologies (2/10)**
- **Exascale Steering Committee**
  - **"Denver" vendor NDA visits 8/2009**
  - **SC09 vendor feedback meetings**
  - **Extreme Architecture and Technology Workshop 12/2009**
- **International Exascale Software Project**
  - **Santa Fe, NM 4/2009; Paris, France 6/2009; Tsukuba, Japan 10/2009**







MISSION IMPERATIVES



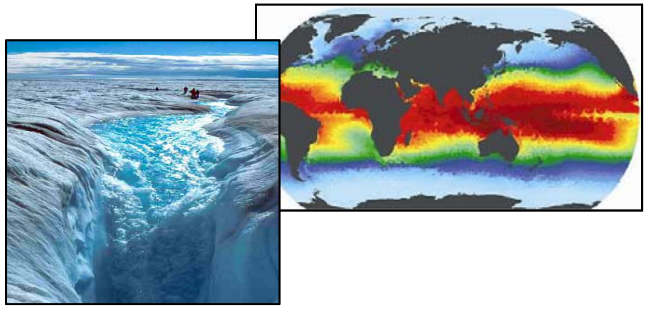
FUNDAMENTAL SCIENCE

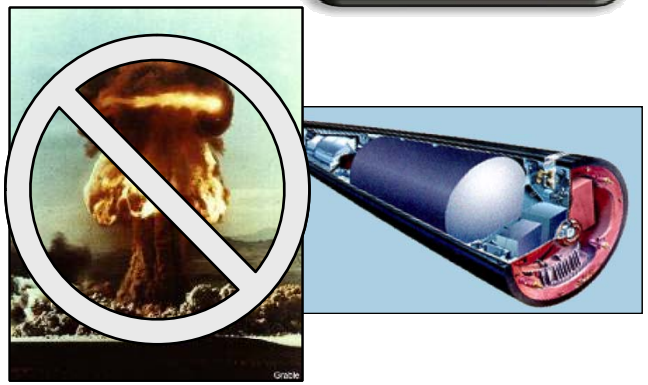
U.S. DEPARTMENT OF ENERGY

MISSION & SCIENCE NEEDS

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 systems**
- ***National Nuclear Security***: **Maintaining a safe, secure and reliable nuclear stockpile**
  - **Stockpile certification**
  - **Predictive scientific challenges**
  - **Real-time evaluation of urban nuclear detonation**





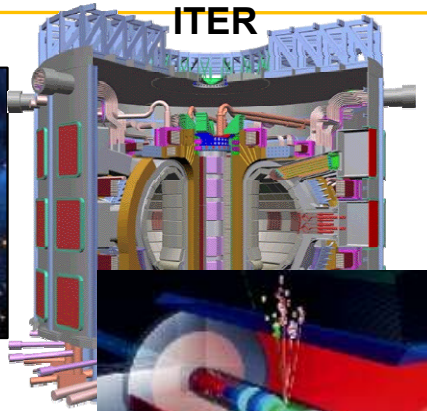
Accomplishing these missions requires exascale resources.

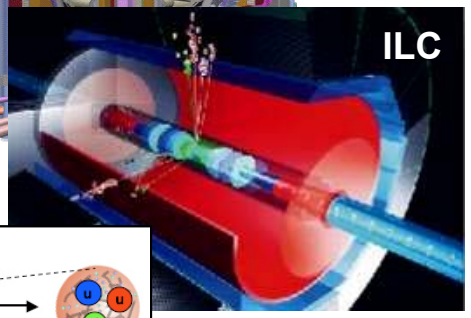
Exascale simulation will enable fundamental advances in basic science.

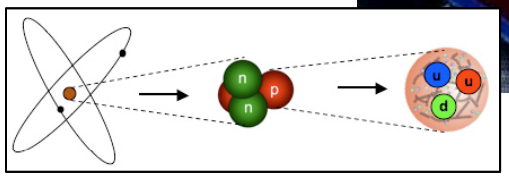
- **High Energy & Nuclear Physics**
  - Dark-energy and dark matter
  - Fundamentals of fission fusion reactions
- **Facility and experimental design**
  - Effective design of accelerators
  - Probes of dark energy and dark matter
  - ITER shot planning and device control
- **Materials / Chemistry**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

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

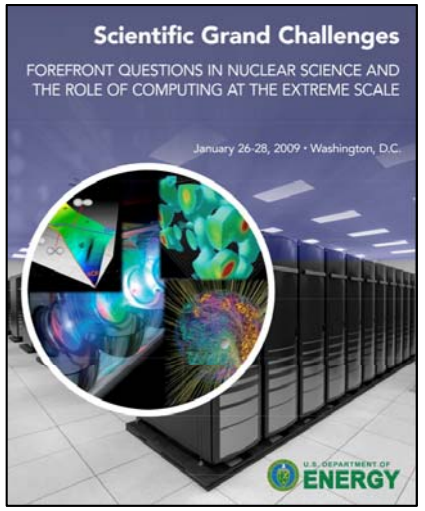
Hubble image of lensing

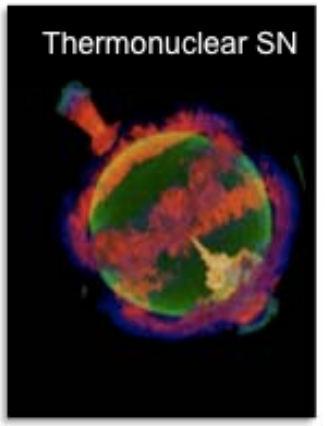






Structure of nucleons

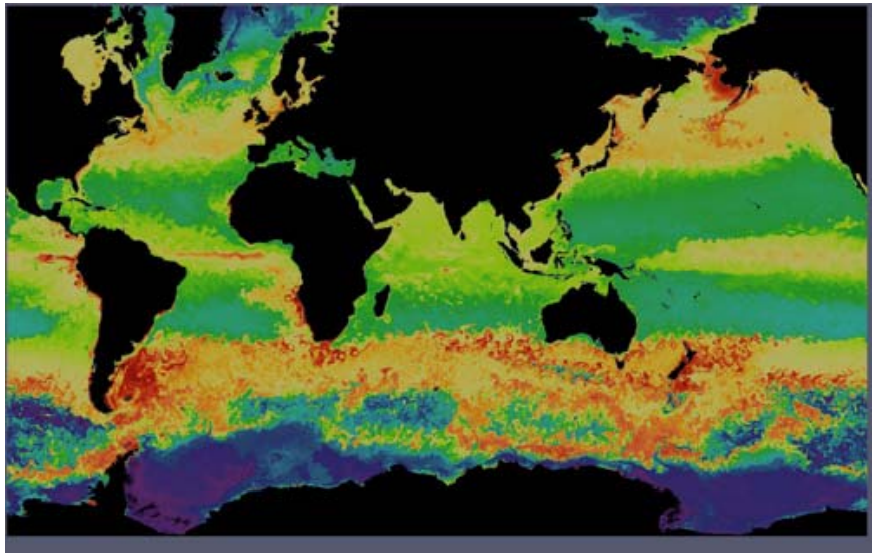




U.S. DEPARTMENT OF ENERGY

Exascale resources are required for predictive climate simulation.

- **Finer resolution**
  - **Provide regional details**
- **Higher realism, more complexity**
  - **Add "new" science**
    - **Biogeochemistry**
    - **Ice-sheets**
  - **Up-grade to "better" science**
    - **Better cloud processes**
    - **Dynamics land surface**
- **Scenario replication, ensembles**
  - **Range of model variability**
- **Time scale of simulation**
  - **Long-term implications**

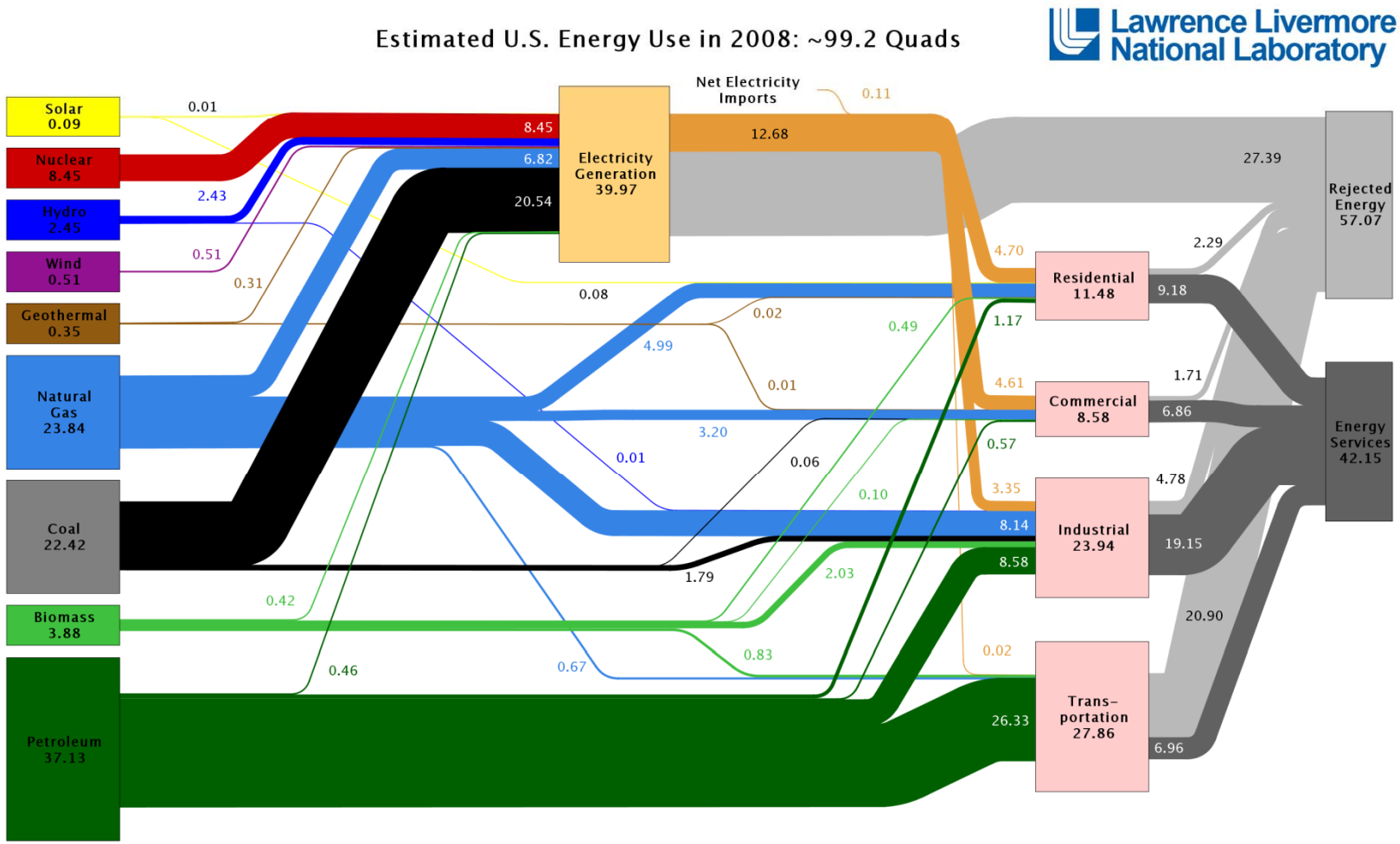


Ocean chlorophyll from an eddy-resolving simulation with ocean ecosystems included

It is essential that computing power be increased substantially (by a factor of 1000), and scientific and technical capacity be increased (by at least a factor of 10) to produce weather and climate information of sufficient skill to facilitate regional adaptations to climate variability and change.
World Modeling Summit for Climate Prediction, May, 2008

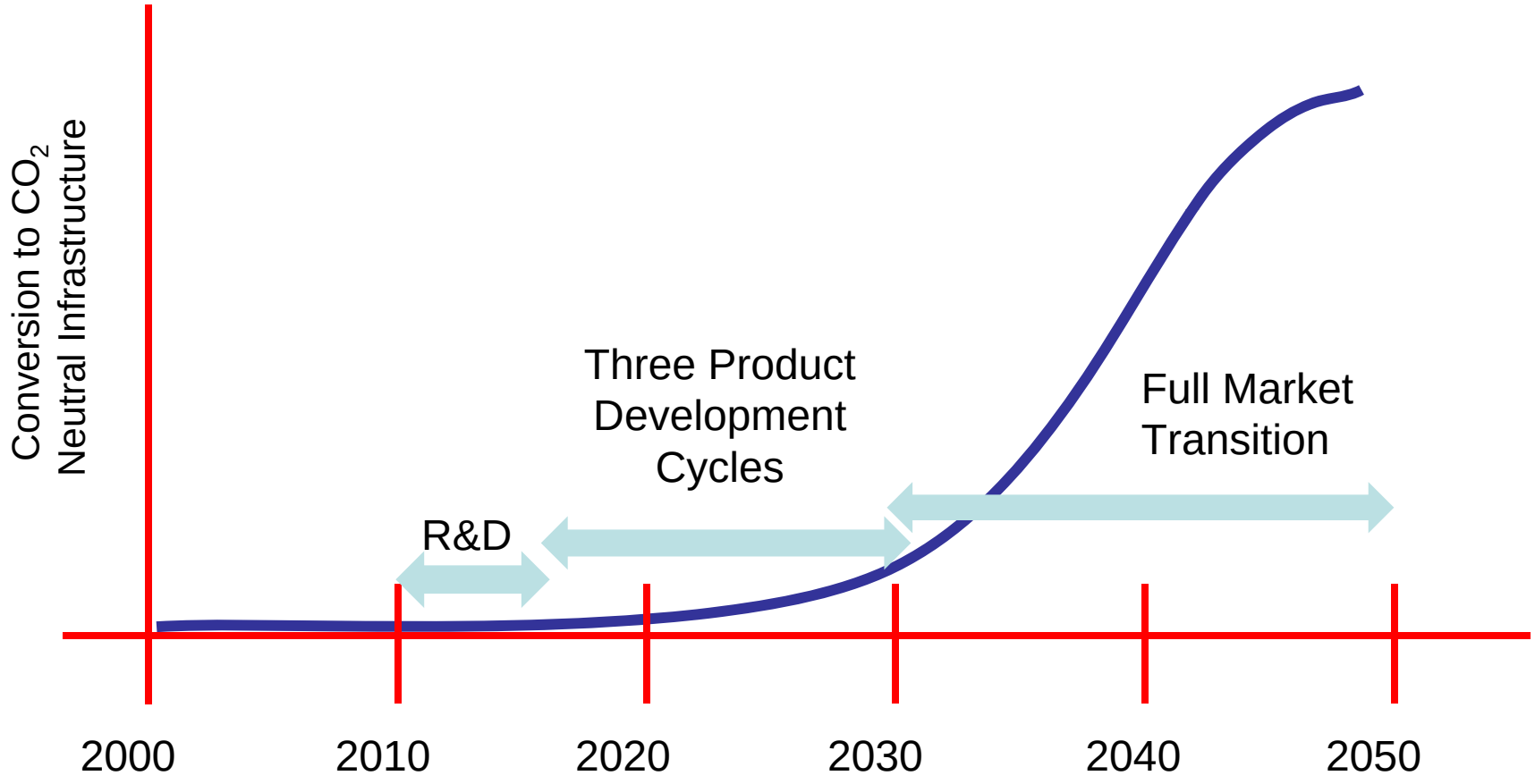
Adapted from *Climate Model Development Breakout Background*
Bill Collins and Dave Bader, Co-Chairs

US energy flows (2008, ≈ 104 Exajoules)

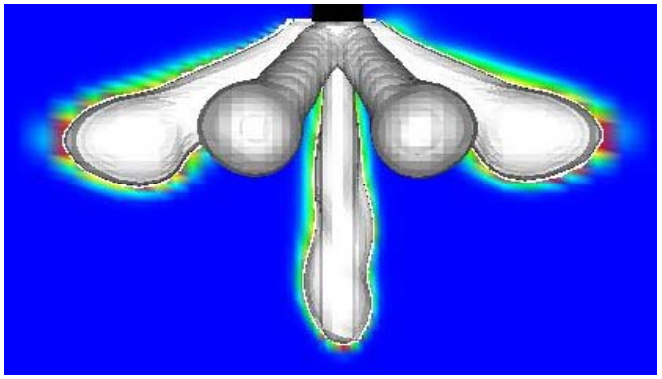


Source: LLNL 2009. Data is based on DOE/EIA-0384(2008), June 2009. If this information or a reproduction of it is used, credit must be given to the Lawrence Livermore National Laboratory and the Department of Energy, under whose auspices the work was performed. Distributed electricity represents only retail electricity sales and does not include self-generation. EIA reports flows for non-thermal resources (i.e., hydro, wind and solar) in BTU-equivalent values by assuming a typical fossil fuel plant "heat rate." The efficiency of electricity production is calculated as the total retail electricity delivered divided by the primary energy input into electricity generation. End use efficiency is estimated as 80% for the residential, commercial and industrial sectors, and as 25% for the transportation sector. Totals may not equal sum of components due to independent rounding. LLNL-MI-410527

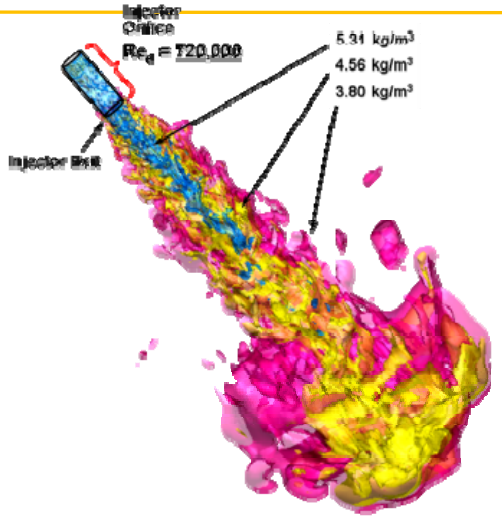
Product development times must be accelerated to meet energy goals



Simulation for product engineering will evolve from mean effects to predictive



RANS calculation for fuel injector captures mean behavior



LES calculation for fuel injector captures greater range of physical scales

Current CFD tools

- Reynolds-Averaged Navier-Stokes
- Calculate mean effects of turbulence
- Turbulent combustion submodels calibrated over narrow range
- DNS and LES for science calculations at standard pressures

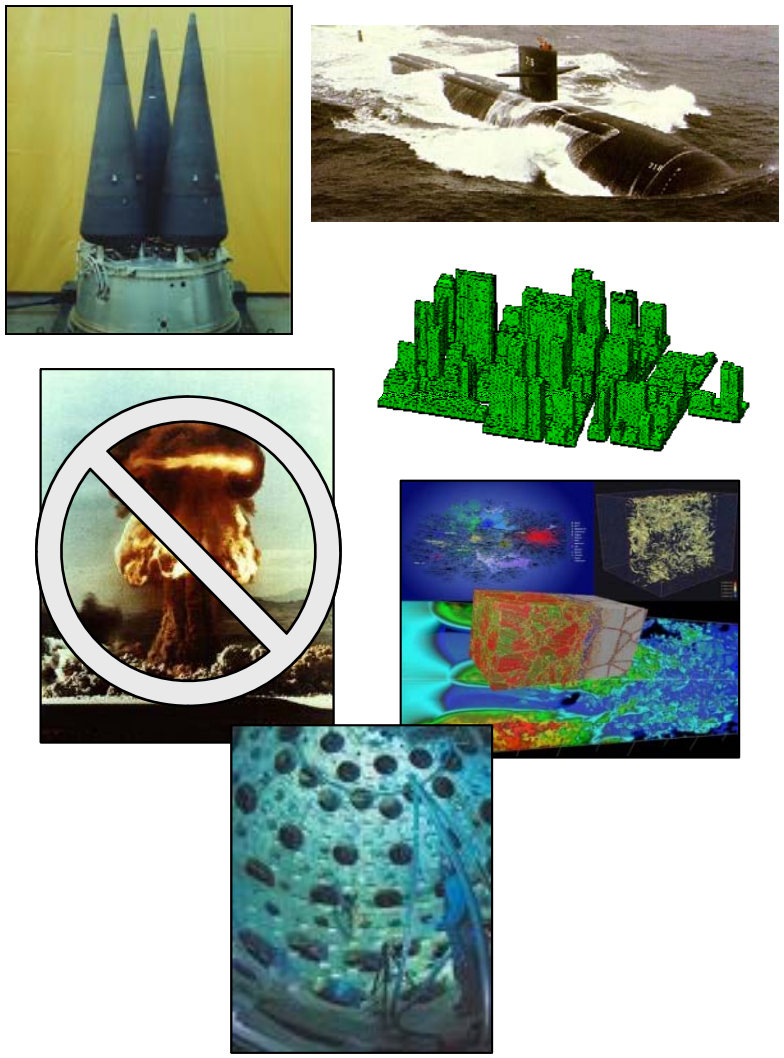
Future CFD tools

- Improved math models for more accurate RANS simulations
- LES with detailed chemistry, complex geometry, high pressures, and multiphase transport as we achieve exascale computing
- DNS for submodel development
- Alternative fuel combustion models

U.S. DEPARTMENT OF ENERGY

National Nuclear Security

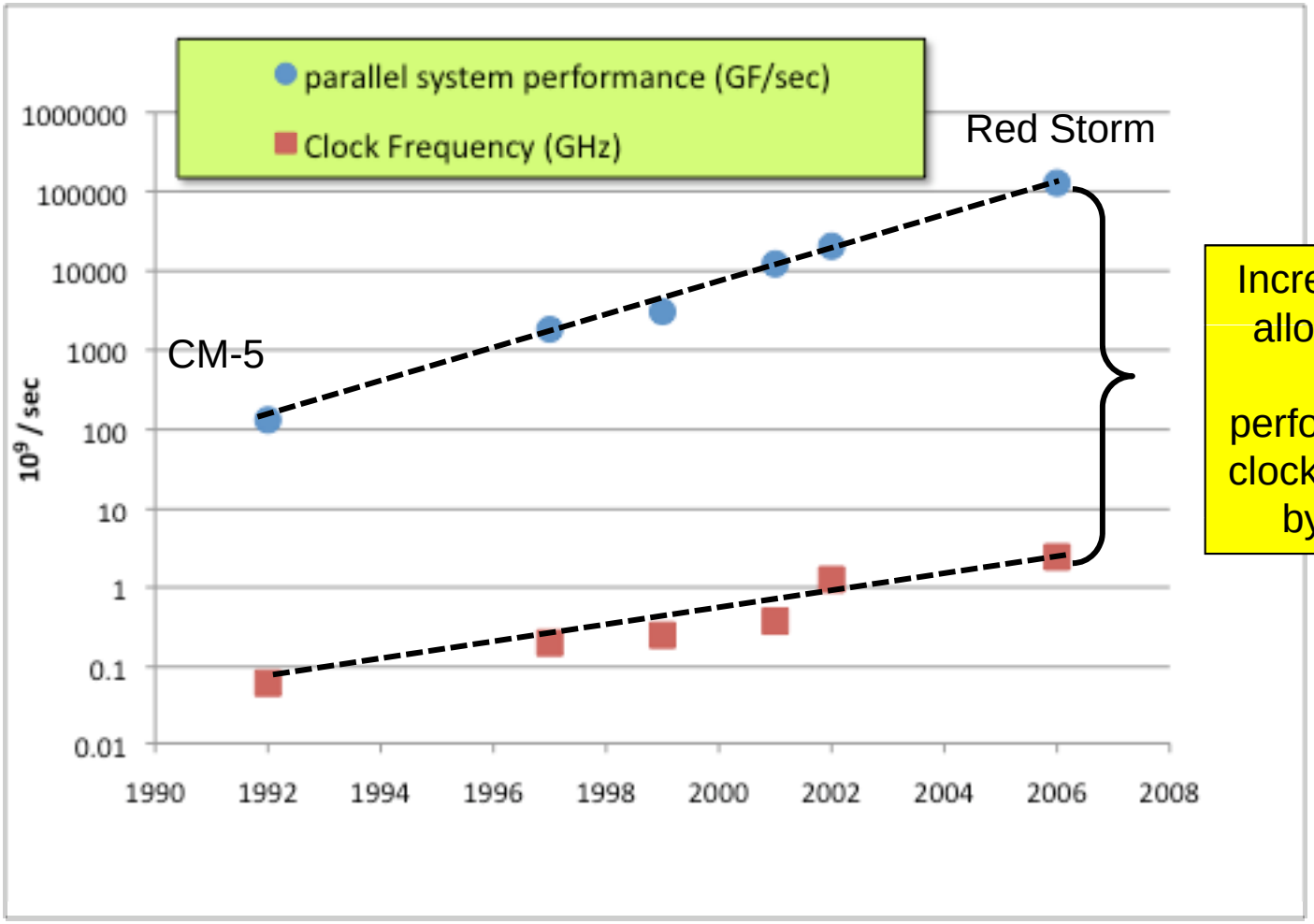
- **U.S. Stockpile must remain safe, secure and reliable without nuclear testing**
  - **Annual certification**
  - **Directed Stockpile Work**
  - **Life Extension Programs**
- **A predictive simulation capability is essential to achieving this mission**
  - **Integrated design capability**
  - **Resolution of remaining unknowns**
    - **Energy balance**
    - **Boost**
    - **Si radiation damage**
    - **Secondary performance**
  - **Uncertainty Quantification**
  - **Experimental campaigns provide critical data for V&V (NIF, DARHT, MaRIE)**
- **Effective exascale resources are necessary for prediction and quantification of uncertainty**



U.S. DEPARTMENT OF ENERGY

TECHNOLOGY NEEDS

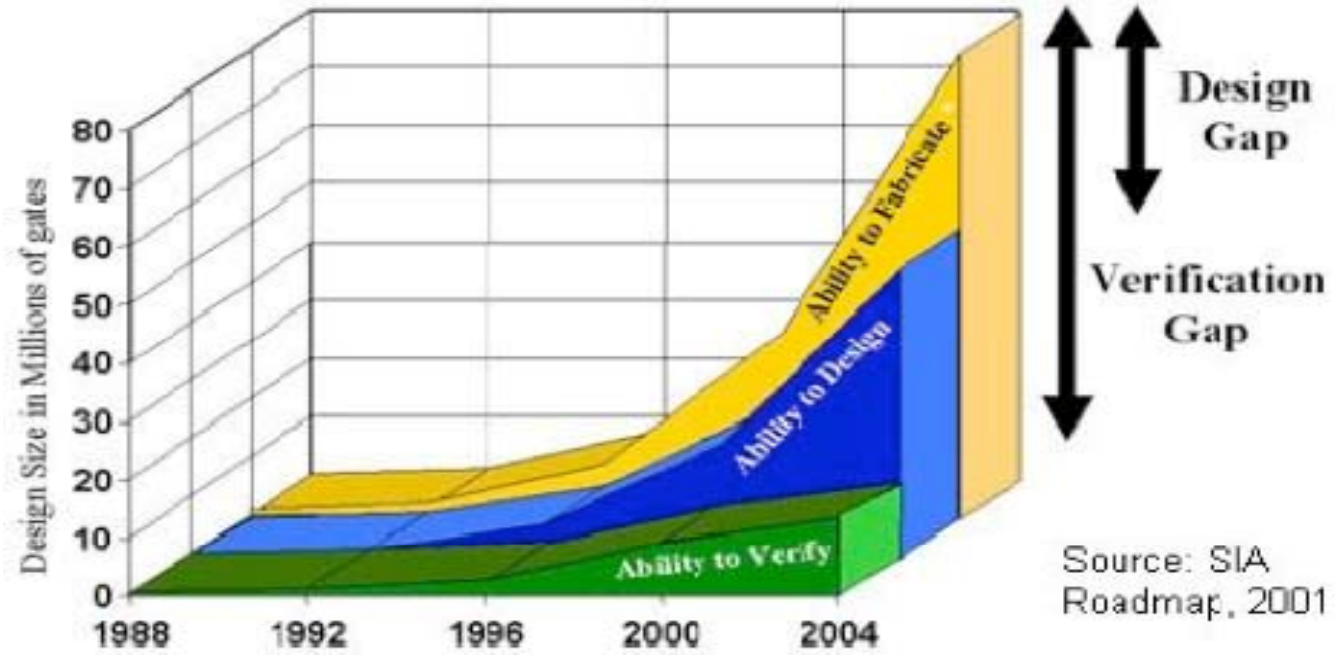
Concurrency is one key ingredient in getting to exaflop/sec



Increased parallelism allowed a 1000-fold increase in performance while the clock speed increased by a factor of 40

and power, resiliency, programming models, memory bandwidth, I/O, …

Many-core chip architectures are the future.



The shift toward increasing parallelism is not a triumphant stride forward based on breakthroughs in novel software and architectures for parallelism … instead it is actually a retreat from even greater challenges that thwart efficient silicon implementation of traditional uniprocessor architectures.
Kurt Keutzer

U.S. DEPARTMENT OF ENERGY

What are critical exascale technology investments?

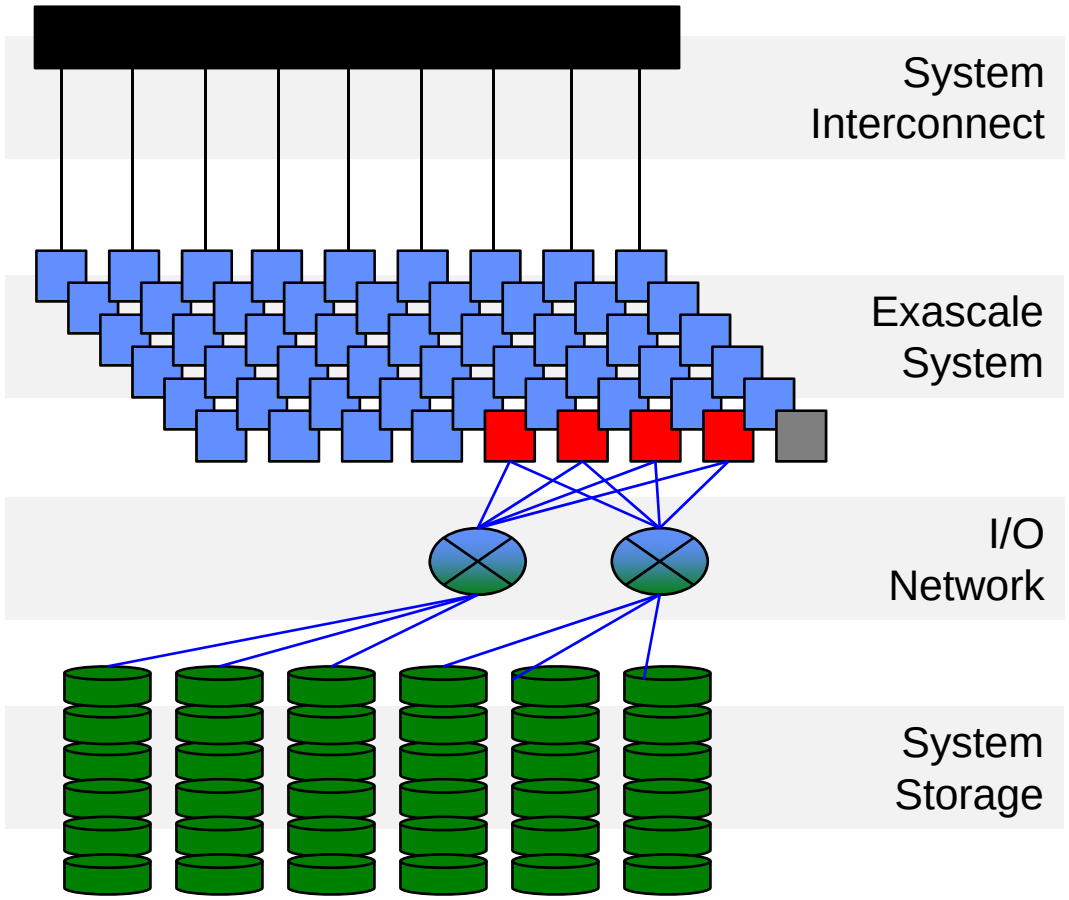
- **System power** is a first class constraint on exascale system performance and effectiveness.
- M**emory** is an important component of meeting exascale power and applications goals.
- **Programming model**. Early investment in several efforts to decide in 2013 on exascale programming model, allowing 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.*

U.S. DEPARTMENT OF ENERGY

Potential System Architecture Targets

System attributes	2010	"2015"		"2018"	
System peak	2 Peta	200 Petaflop/sec		1 Exaflop/sec	
Power	6 MW	15 MW		20 MW	
System memory	0.3 PB	5 PB		32-64 PB	
Node performance	125 GF	0.5 TF	7 TF	1 TF	10 TF
Node memory BW	25 GB/s	0.1 TB/sec	1 TB/sec	0.4 TB/sec	4 TB/sec
Node concurrency	12	O(100)	O(1,000)	O(1,000)	O(10,000)
System size (nodes)	18,700	50,000	5,000	1,000,000	100,000
Total Node Interconnect BW	1.5 GB/s	20 GB/sec		200 GB/sec	
MTTI	days	O(1day)		O(1 day)	

The high level system design may be similar to petascale systems

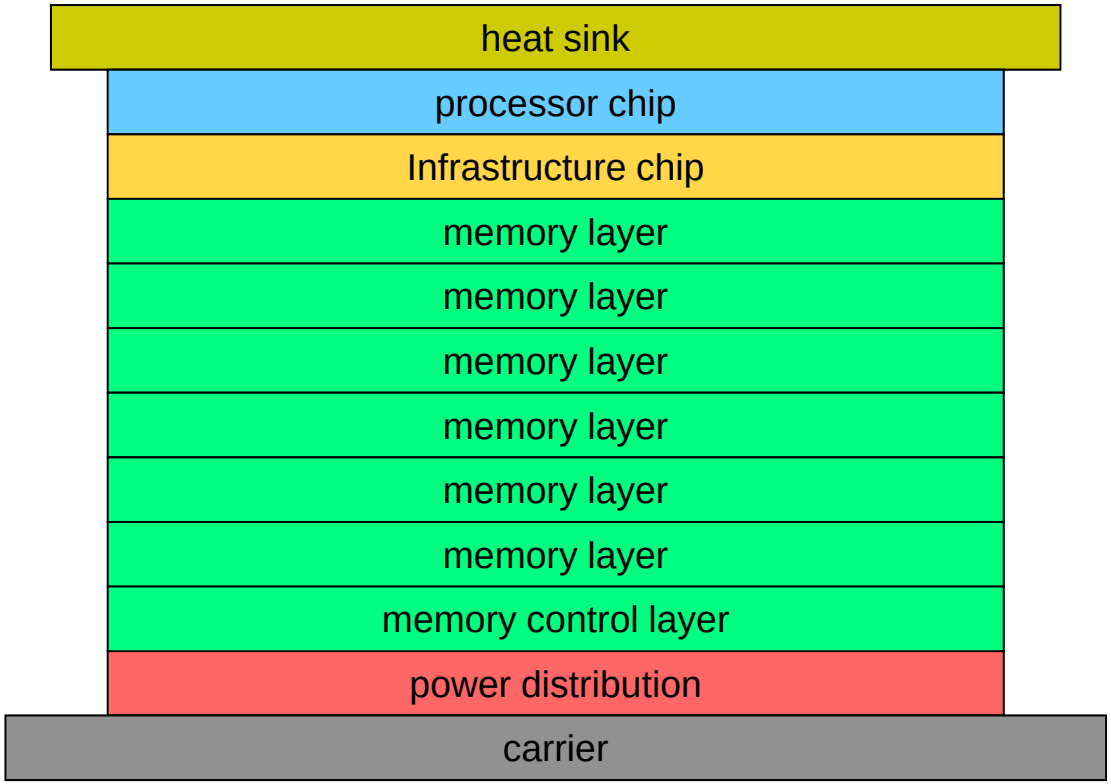


- New interconnect topologies
- Optical interconnect

- 10x – 100x more nodes
- MPI scaling & fault tolerance
- Different types of nodes

- Mass storage far removed from application data

The node is the key for exascale, as well as for ~ exascale.



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

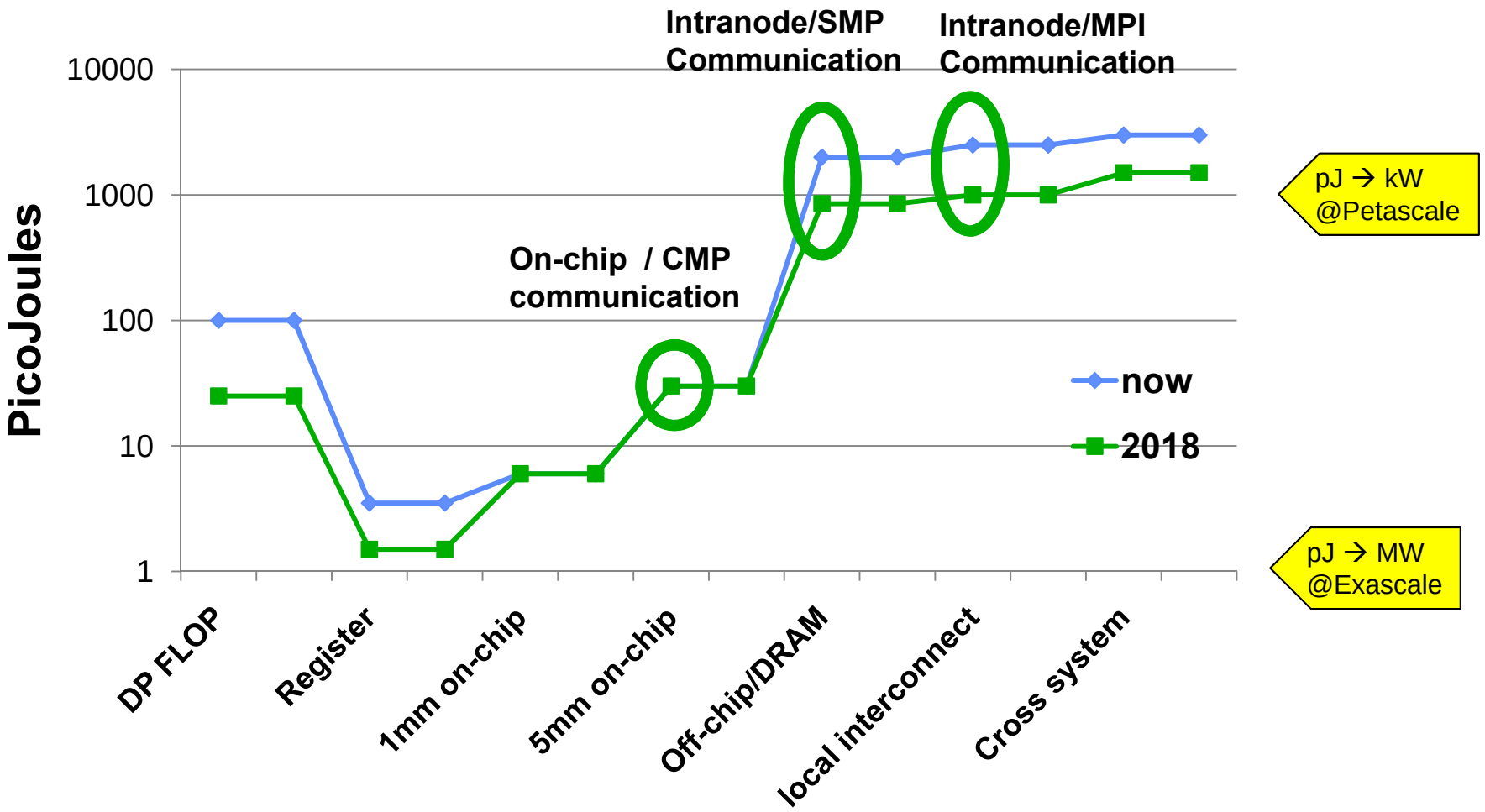
- 3d stacked memory

- Smart memory management

- Integration on package

U.S. DEPARTMENT OF ENERGY

Investments in architecture R&D and application locality are critical



"The Energy and Power Challenge is the most pervasive … and has its roots in the inability of the [study] group to project any combination of currently mature technologies that will deliver sufficiently powerful systems in any class at the desired levels."
DARPA IPTO exascale technology challenge report

Memory bandwidth and memory sizes will be >> less effective without R&D

- Primary needs are
  - Increase in bandwidth (concurrency can be used to mask latency, viz. Little's Law)
  - Lower power consumption
  - Lower cost (to enable affordable capacity)
- Stacking on die enable improved bandwidth and lower power consumption
- Modest improvements in latency
- Commodity memory interface standards are not pushing bandwidth enough

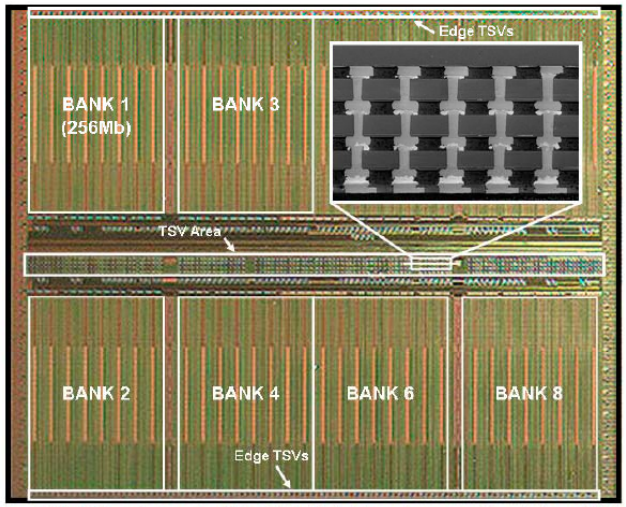


Figure 7.2.7: Die micrograph of the fabricated chip and cross-sectional view of TSVs. The chip size is 10.9×9.0mm².

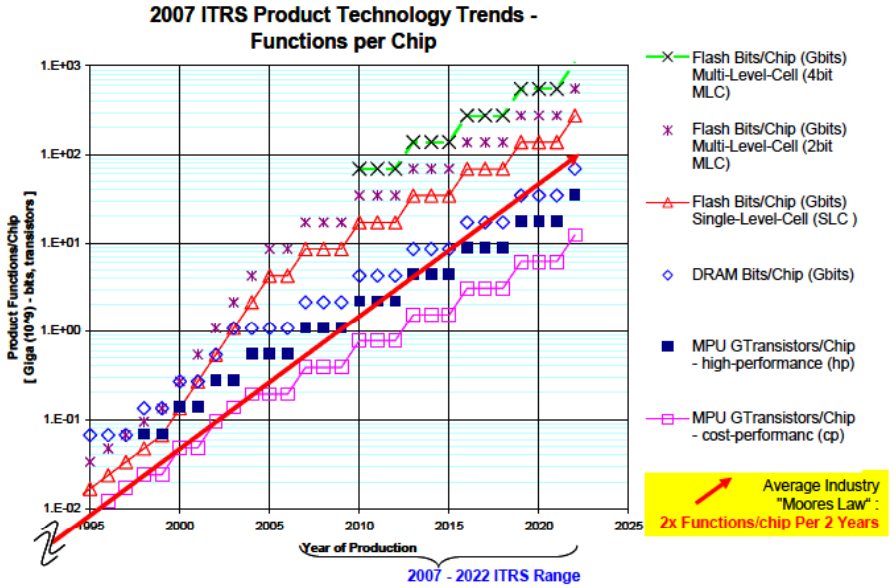
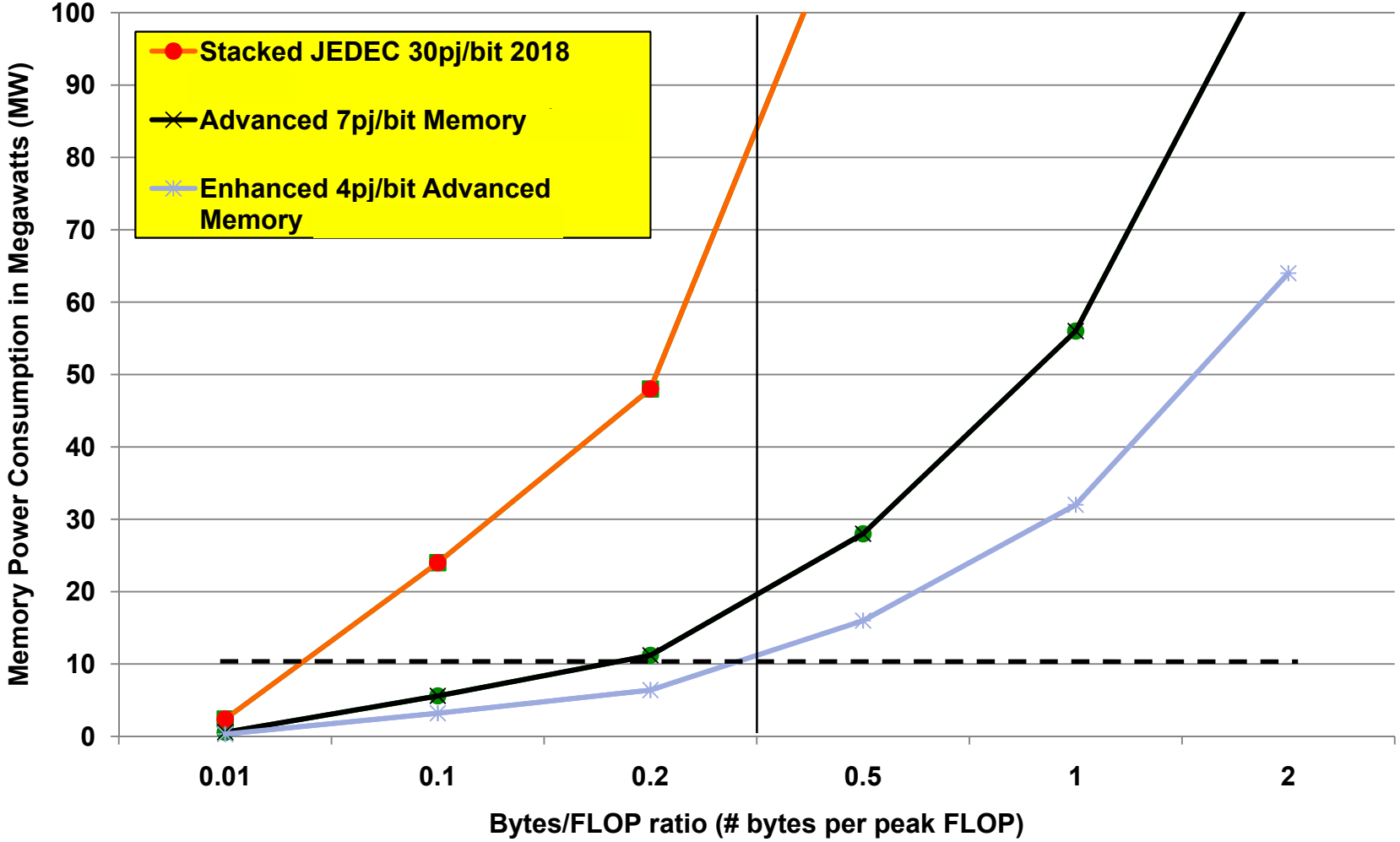


Figure ORTC2 ITRS Product Function Size Trends: MPU Logic Gate Size (4-transistor); Memory Cell Size [SRAM (6-transistor); Flash (SLC and MLC), and DRAM (transistor + capacitor)]--Updated

U.S. DEPARTMENT OF ENERGY

Investments in memory technology mitigate risk of narrowed application scope.

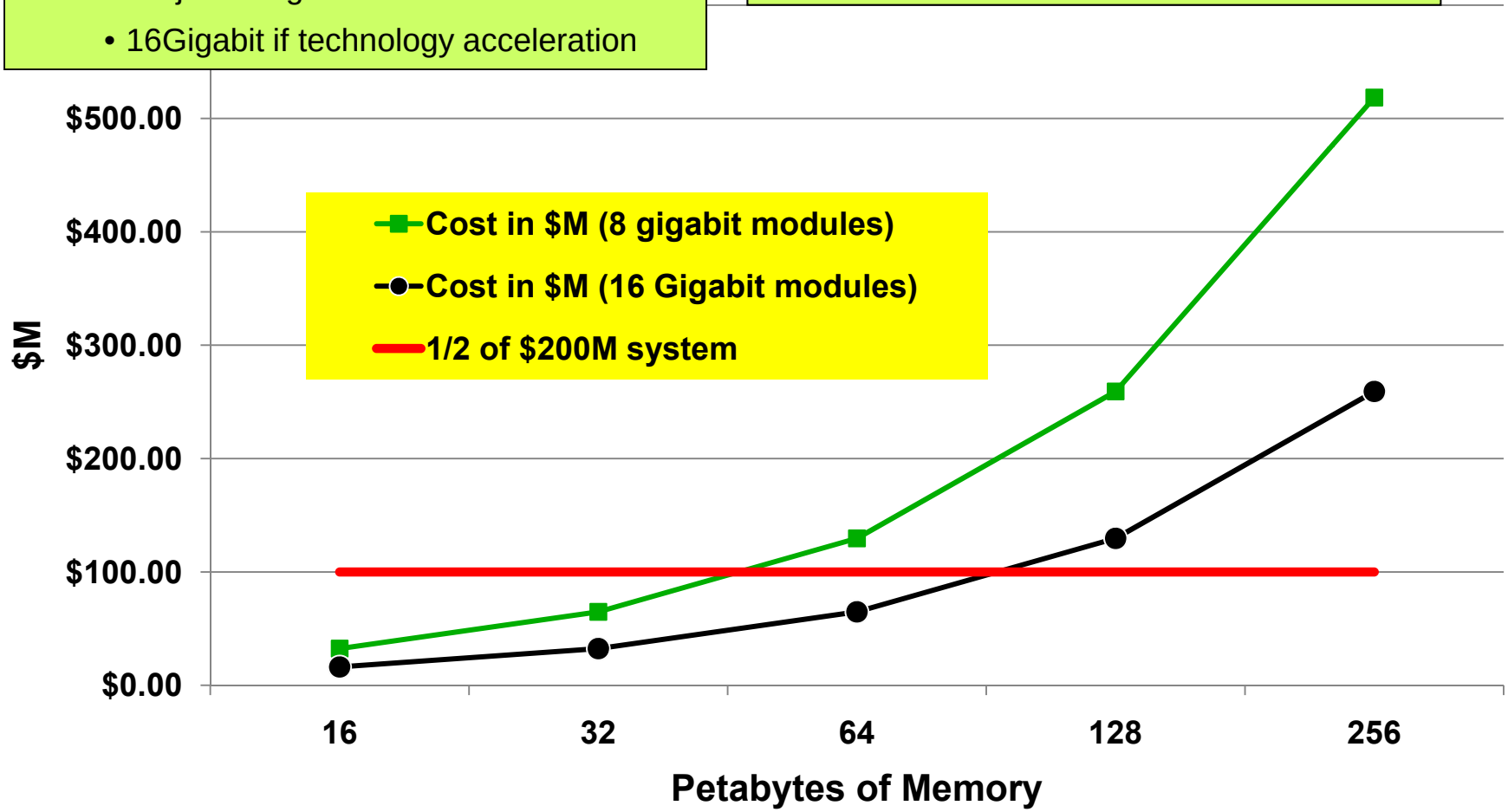


U.S. DEPARTMENT OF ENERGY

Cost of Memory Capacity for two different potential memory Densities

- Memory density is doubling every three years; processor logic, every two
  - Project 8Gigabit DIMMs in 2018
  - 16Gigabit if technology acceleration

- Storage costs are dropping gradually compared to logic costs
  - Industry assumption is $1.80/memory chip is median commodity cost



U.S. DEPARTMENT OF ENERGY

Factors Driving up the Fault Rate

It is more than just the increase in the number of components

Number of components both memory and processors will increase by an order of magnitude which will increase hard and soft errors.

Smaller circuit sizes, running at lower voltages to reduce power consumption, increases the probability of switches flipping spontaneously due to thermal and voltage variations as well as radiation, increasing soft errors

Power management cycling significantly decreases the components lifetimes due to thermal and mechanical stresses.

Resistance to add additional HW detection and recovery logic right on the chips to detect silent errors. Because it will increase power consumption by 15% and increase the chip costs.

Heterogeneous systems make error detection and recovery even harder, for example, detecting and recovering from an error in a GPU can involve hundreds of threads simultaneously on the GPU and hundreds of cycles in drain pipelines to begin recovery.

Increasing system and algorithm complexity makes improper interaction of separately designed and implemented components more likely.

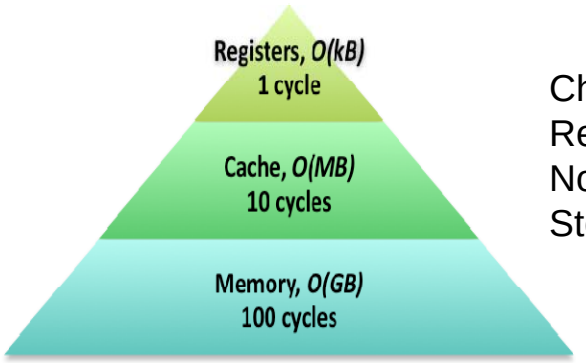
Number of operations (10^{23} in a week) ensure that system will traverse the tails of the operational probability distributions.

Need solutions for decreased reliability and a new model for resiliency

- **Barriers**
  - System components, complexity increasing
  - Silent error rates increasing
  - Reduced job progress due to fault recovery if we use existing checkpoint/restart
- **Technical Focus Areas**
  - Local recovery and migration
  - Development of a standard fault model and better understanding of types/rates of faults
  - Improved hardware and software reliability
    - Greater integration across entire stack
  - Fault resilient algorithms and applications
- **Technical Gap**
  - Maintaining today's MTTI given 10x - 100X increase in sockets will require:
    - 10X improvement in hardware reliability
    - 10X in system software reliability, and
    - 10X improvement due to local recovery and migration as well as research in fault resilient applications
  - .

Taxonomy of errors (h/w or s/w)

- **Hard errors**: permanent errors which cause system to hang or crash
- **Soft errors**: transient errors, either correctable or short term failure
- **Silent errors**: undetected errors either permanent or transient. *Concern is that simulation data or calculation have been corrupted and no error reported.*



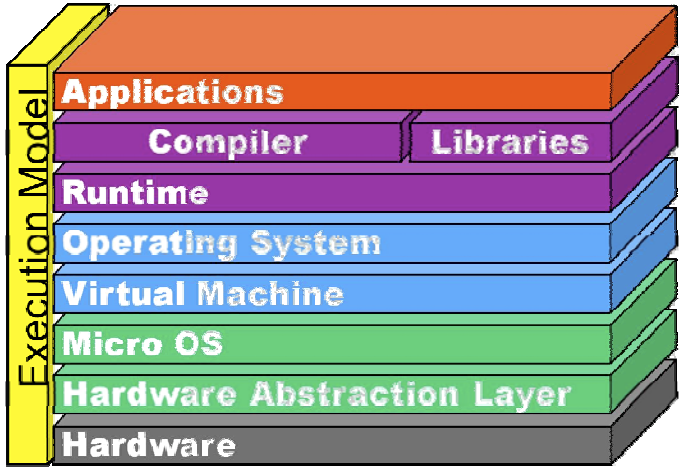
Checkpoint Restart to Node Local Storage

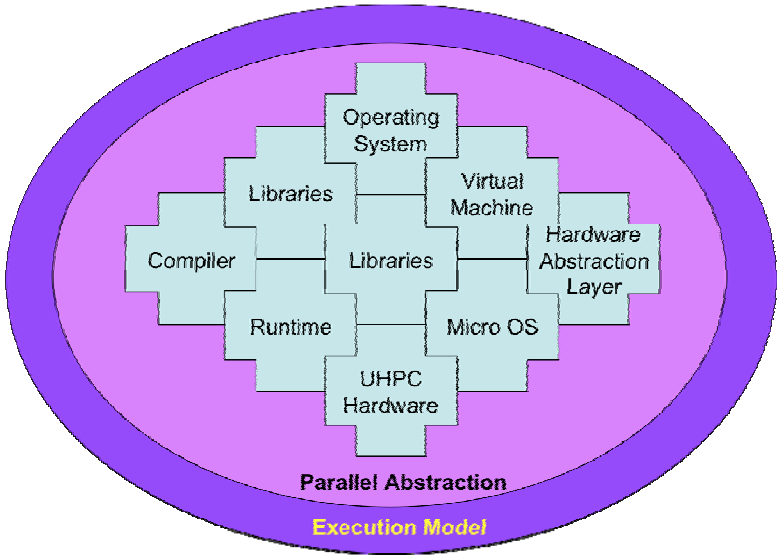
Need storage solution to fill this gap

U.S. DEPARTMENT OF ENERGY

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



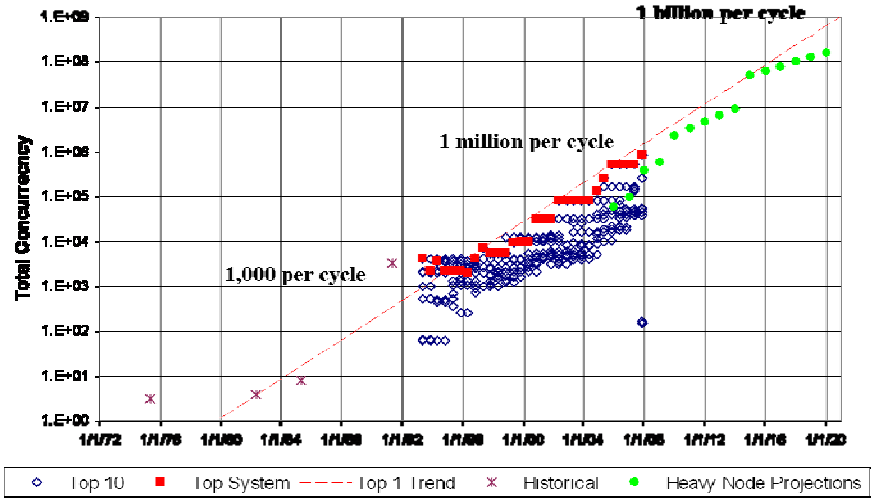


Software challenges in extreme scale systems, *Sarkar, 2010*

U.S. DEPARTMENT OF ENERGY

Programming models and environments require early investment.

- **Barriers:** Delivering a large-scale scientific instrument that is productive and fast.
  - O(1B) way parallelism in Exascale system
  - O(1K) way parallelism in a processor chip
    - Massive lightweight cores for low power
    - Some "full-feature" cores lead to heterogeneity
  - Data movement costs power and time
    - Software-managed memory (local store)
  - Programming for resilience
  - Science goals require complex codes

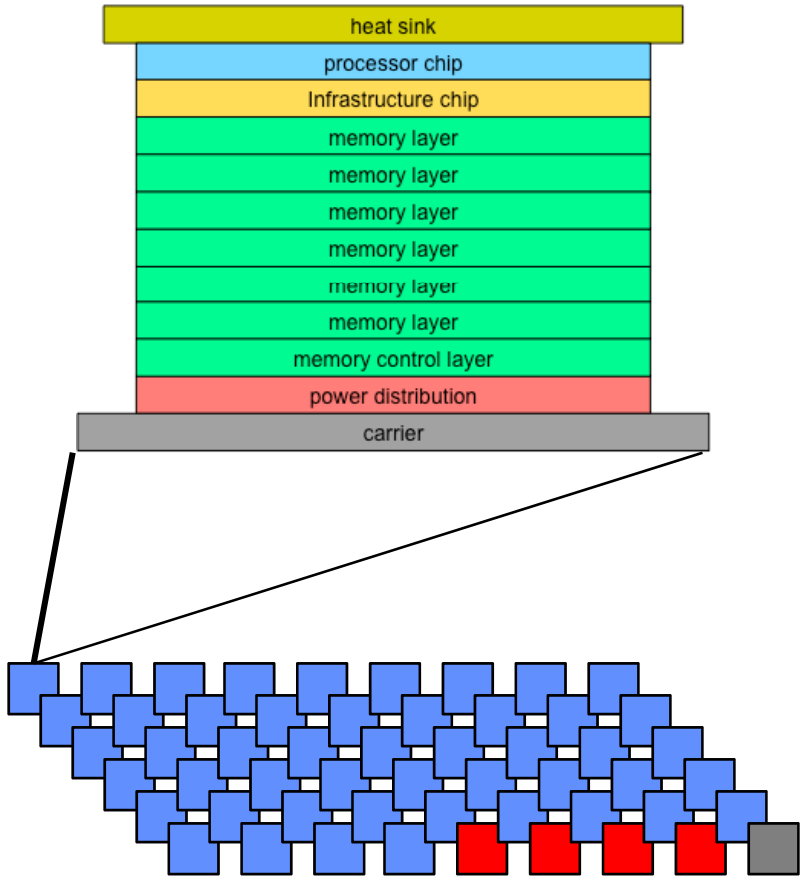


How much parallelism must be handled by the program?
From Peter Kogge (on behalf of Exascale Working Group), "Architectural *Challenges* at the Exascale Frontier", June 20, 2008

- **Technology Investments**
  - Extend inter-node models for scalability and resilience, e.g., MPI, PGAS (includes HPCS)
  - Develop intra-node models for concurrency, hierarchy, and heterogeneity by adapting current scientific ones (e.g., OpenMP) or leveraging from other domains (e.g., CUDA, OpenCL)
  - Develop common low level runtime for portability and to enable higher level models
- **Technical Gap:**
  - No portable model for variety of on-chip parallelism methods or new memory hierarchies
  - Goal: Hundreds of applications on the Exascale architecture; Tens running at scale

Programming Model Approaches

- **Hierarchical approach (intra-node + inter-node)**
  - **Part I: Inter-node model for communicating between nodes**
    - **MPI scaling to millions of nodes: Importance high; risk low**
    - **One-sided communication scaling: Importance medium; risk low**
  - **Part II: Intra-node model for on-chip concurrency**
    - **Overriding Risk: No single path for node architecture**
    - **OpenMP, Pthreads: High risk (may not be feasible with node architectures); high payoff (already in some applications)**
    - **New API, extended PGAS, or CUDA/OpenCL to handle hierarchies of memories and cores: Medium risk (reflects architecture directions); Medium payoff (reprogramming of node code)**
- **Unified approach: single high level model for entire system**
  - **High risk; high payoff for new codes, new application domains**

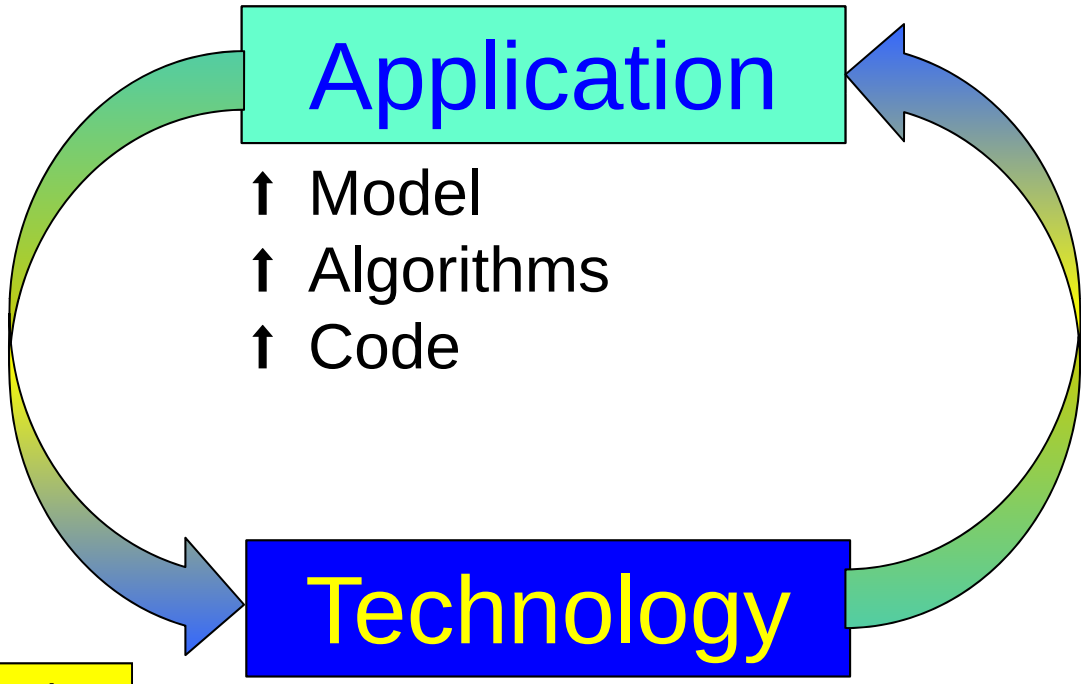


U.S. DEPARTMENT OF ENERGY

CO-DESIGN

Co-design expands the feasible solution space to allow better solutions.

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

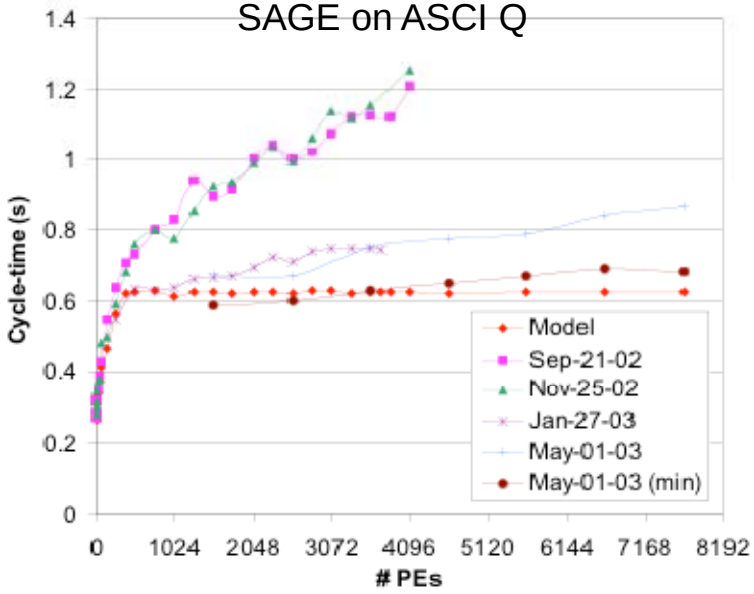
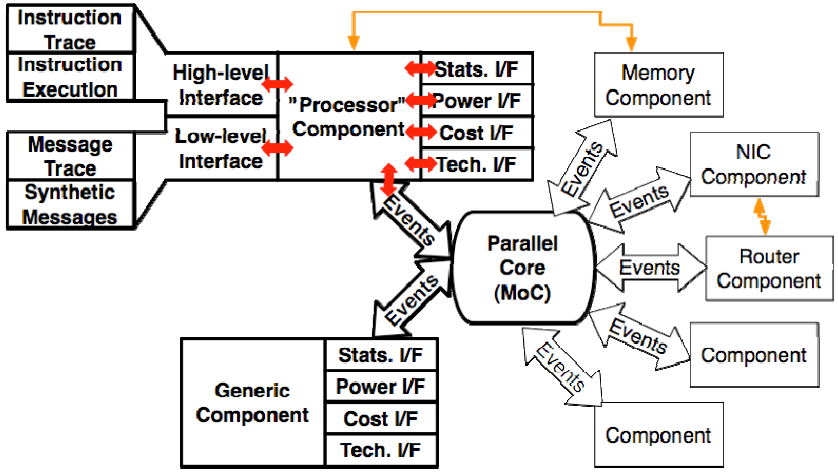


⊕architecture
⊕programming model
⊕resilience
⊕power

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.*

Exascale Initiative Steering Committee

Hierarchical {application, s/w, h/w} co-simulation a the key for co-design

- **Hierarchical co-simulation capability**
  - Discussions between architecture, software and application groups
  - System level simulation based on analytic models
  - Detailed (e.g. cycle accurate) co-simulation of hardware and applications
- **Opportunity to influence future architectures**
  - Cores/node, threads/core, ALUs/thread
  - Logic layer in stacked memory
  - Interconnect performance
  - Memory/core
  - Processor functionality
- **Current community efforts must work together to provide a complete co-design capability**