Exascale Computing and the Role of Codesign

Sudip Dosanjh

Co-director ACES and IAA
Served on DOE's Exascale Initative Steering Committee
Group Leader for Computer Science Research

Sandia National Labratories

DOE MISSION & SCIENCE NEEDS

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

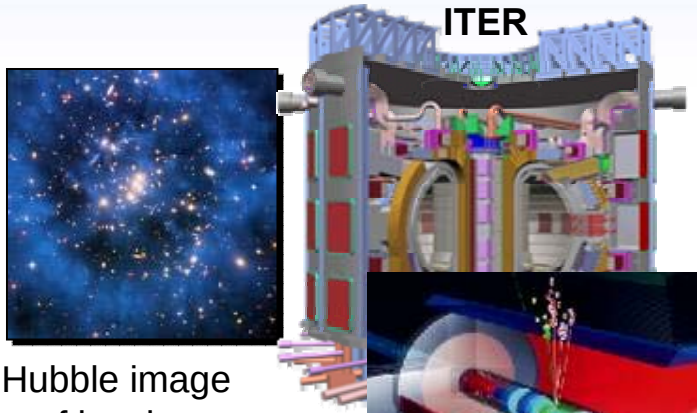
- ***Climate Change***: **Understanding, mitigating and adapting to the effects of global warming**
  - **Sea level rise**
  - **Severe weather**
  - **Regional climate change**
  - **Geologic carbon sequestration**
- ***Energy***: **Reducing U.S. reliance on foreign energy sources and reducing the carbon 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.

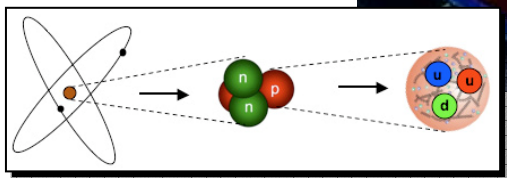
Exascale simulation will enable fundamental advances in basic science

- **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

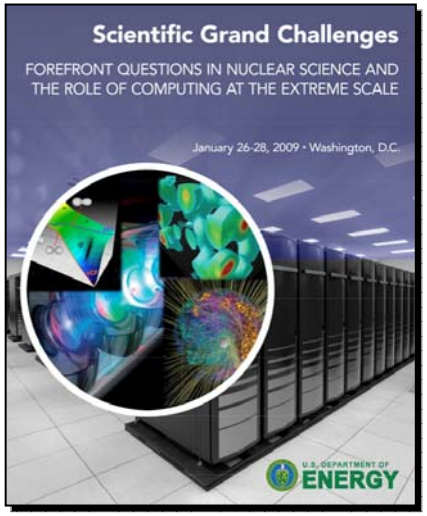
These breakthrough scientific discoveries and facilities require exascale applications and resources.

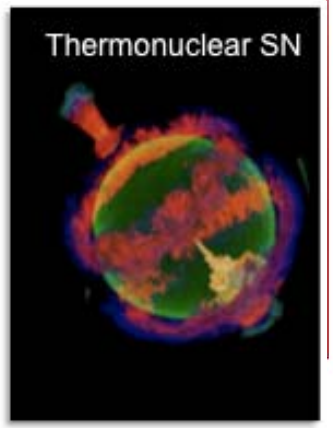


Hubble image of lensing



Structure of nucleons

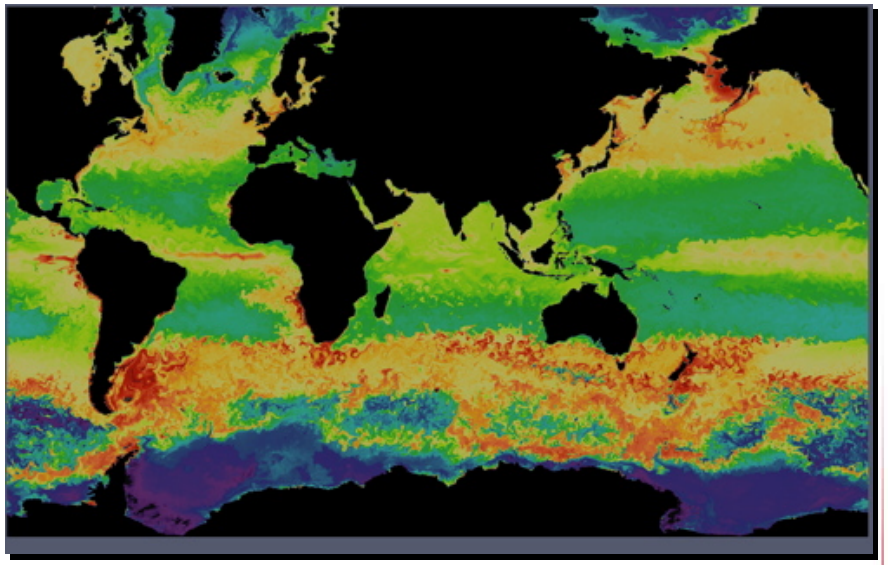




Laboratories

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

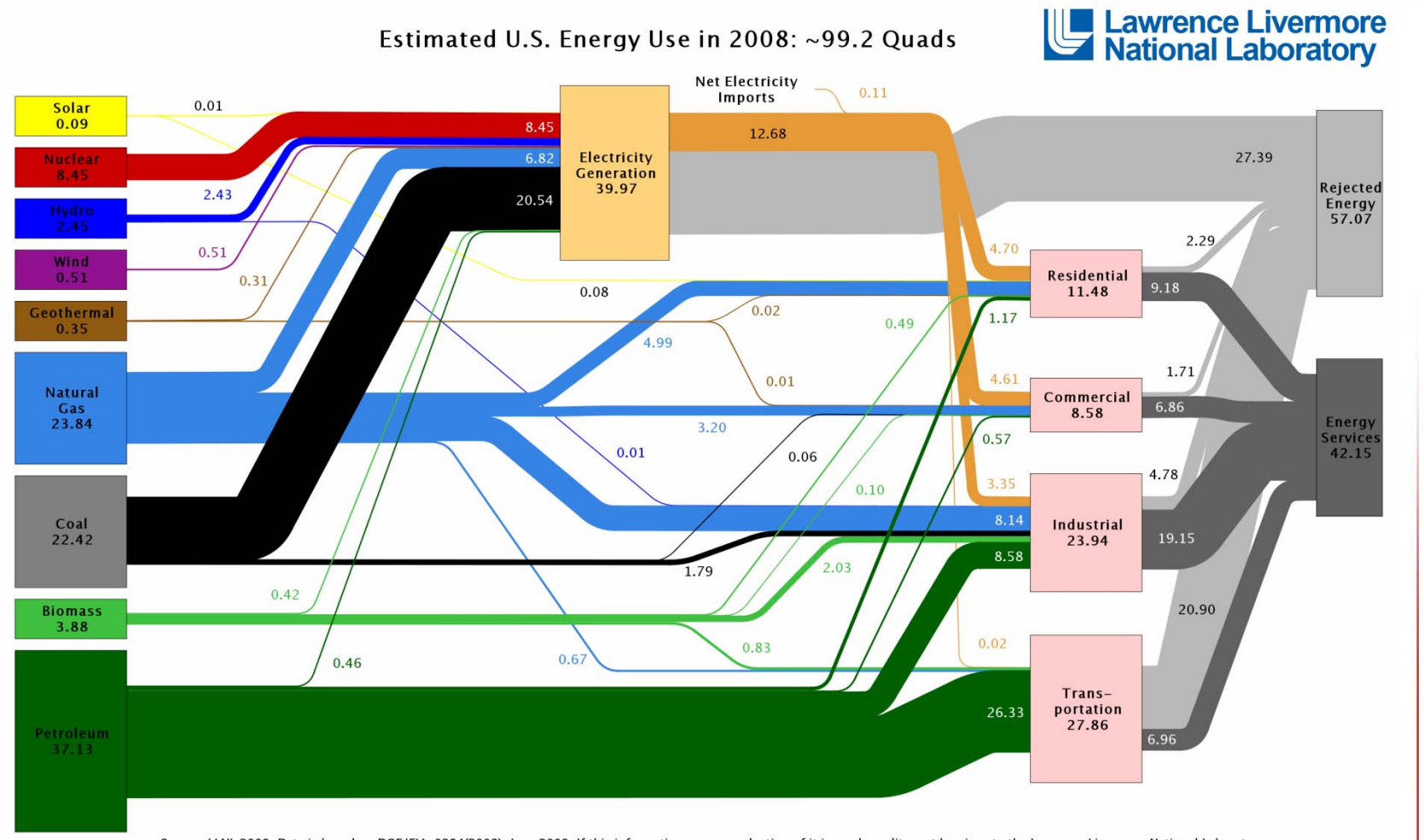


Ocean chlorophyll from an eddy-resolving simulation with ocean ecosystems included

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**

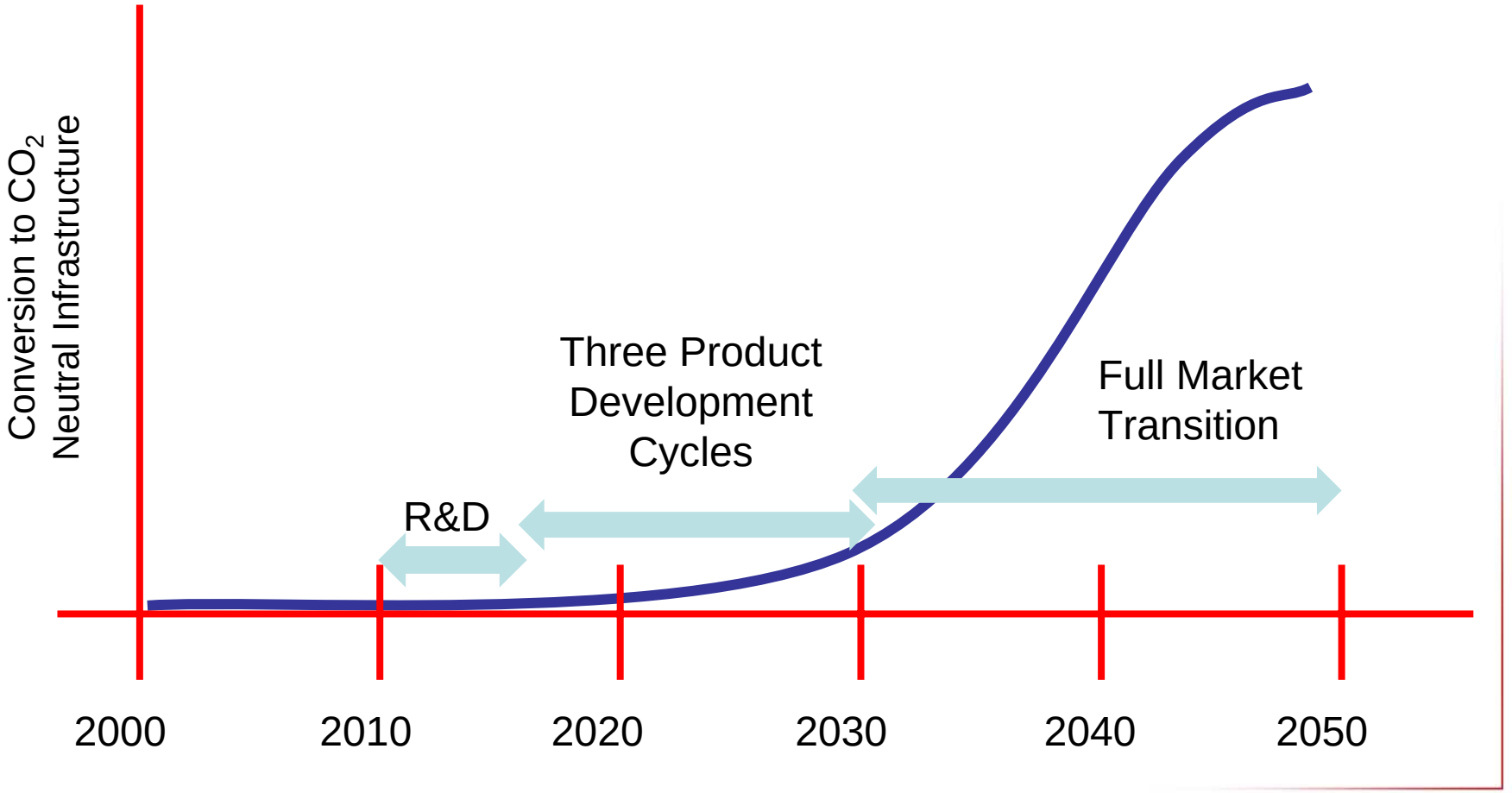
Adapted from*Climate Model Development Breakout Background*
Bill Collins and Dave Bader, Co-Chairs

US energy flows (2008, ≈ 104 Exajoules)



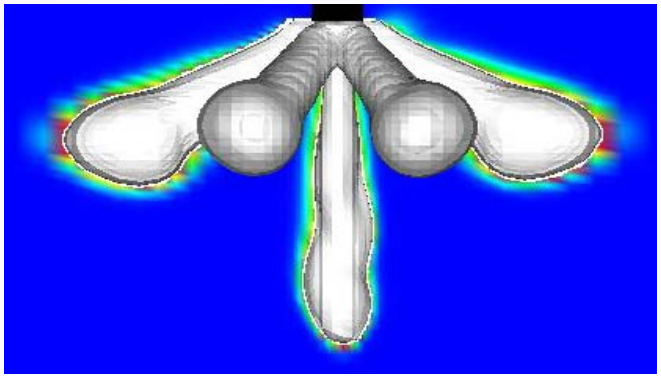
National Laboratories

Product development times must be accelerated to meet energy goals

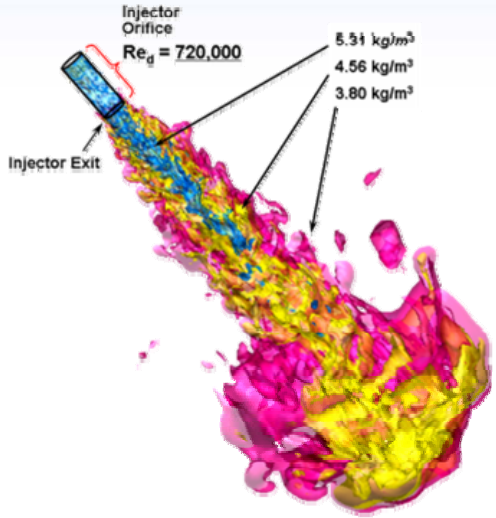


Sandia National Laboratories

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

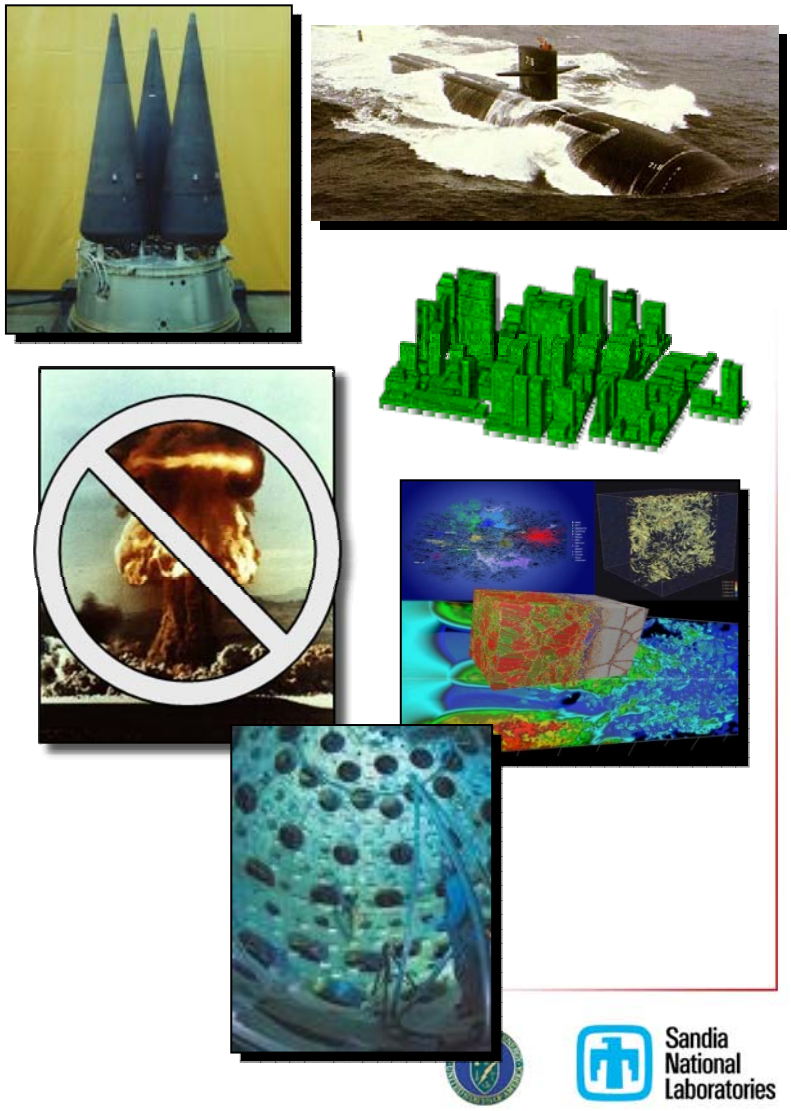
Future CFD tools

- 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

Sandia National Laboratories

National Nuclear Security

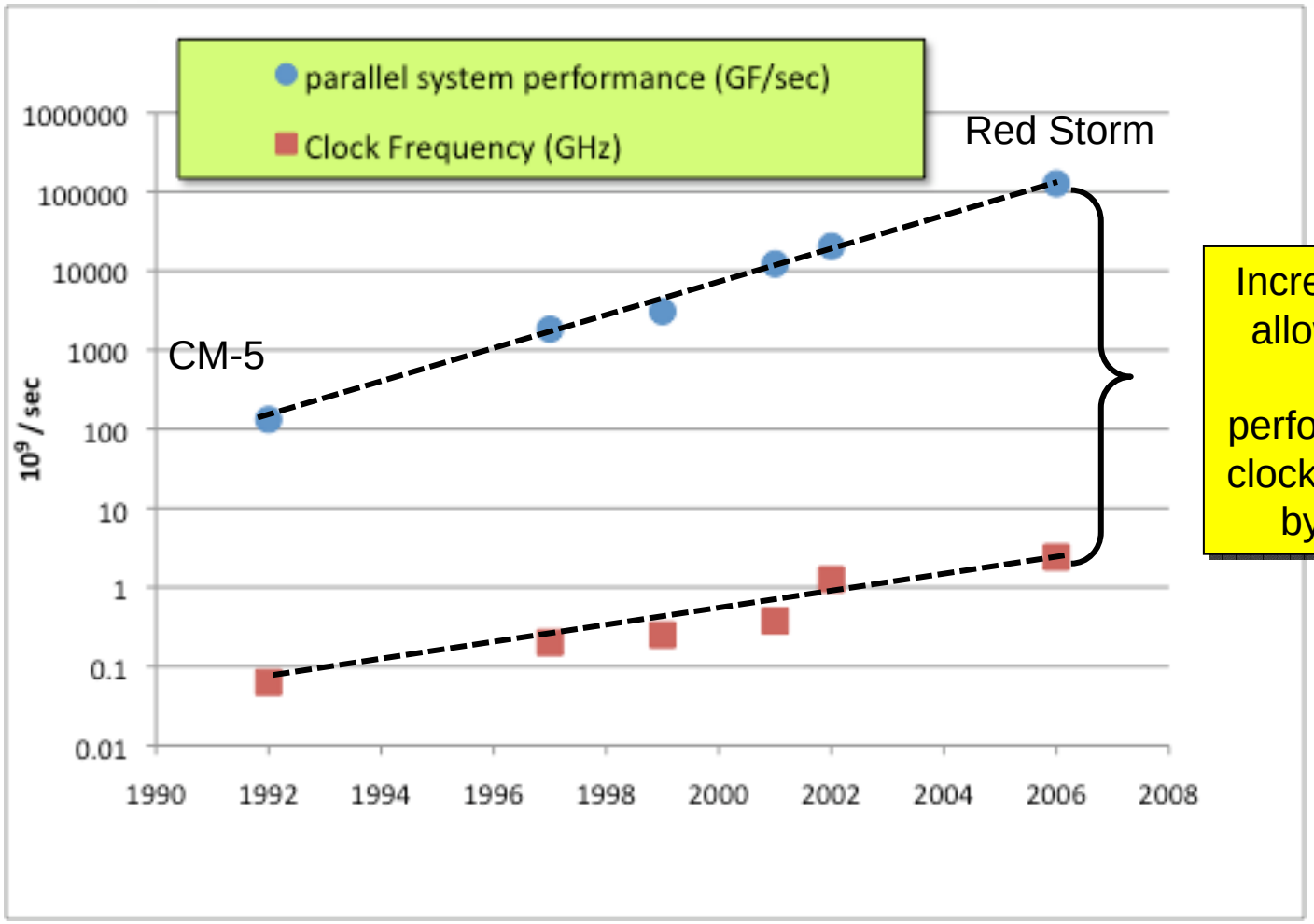
- U.S. Stockpile must remain safe, secure and reliable without nuclear testing
  - Annual certification
  - Directed Stockpile Work
  - Life Extension Programs
- A predictive simulation capability is essential to achieving this mission
  - Integrated design capability
  - Resolution of remaining unknowns
    - 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



Sandia National Laboratories

TECHNOLOGY NEEDS

Concurrency is one key ingredient in getting to exaflop/sec

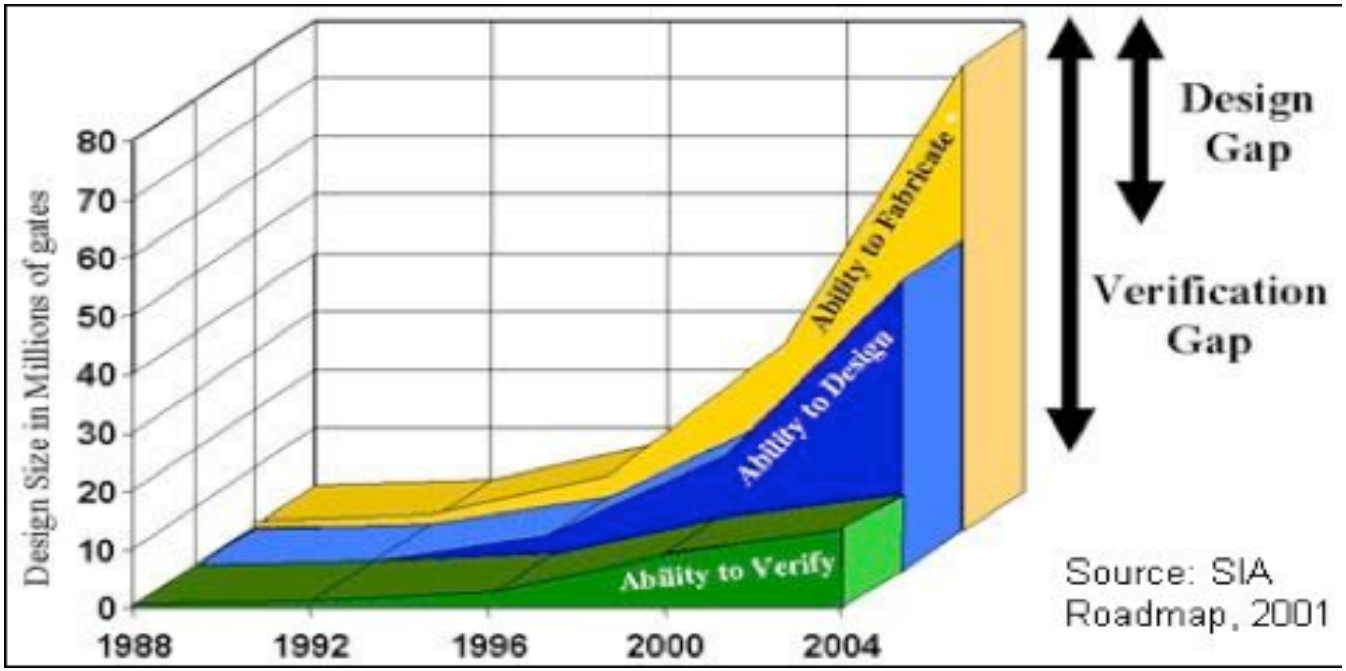


Increased parallelism allowed a 1000-fold increase in performance while the clock speed increased by a factor of 40

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

Sandia National Laboratories

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

Sandia National Laboratories

What are critical exascale technology investments?

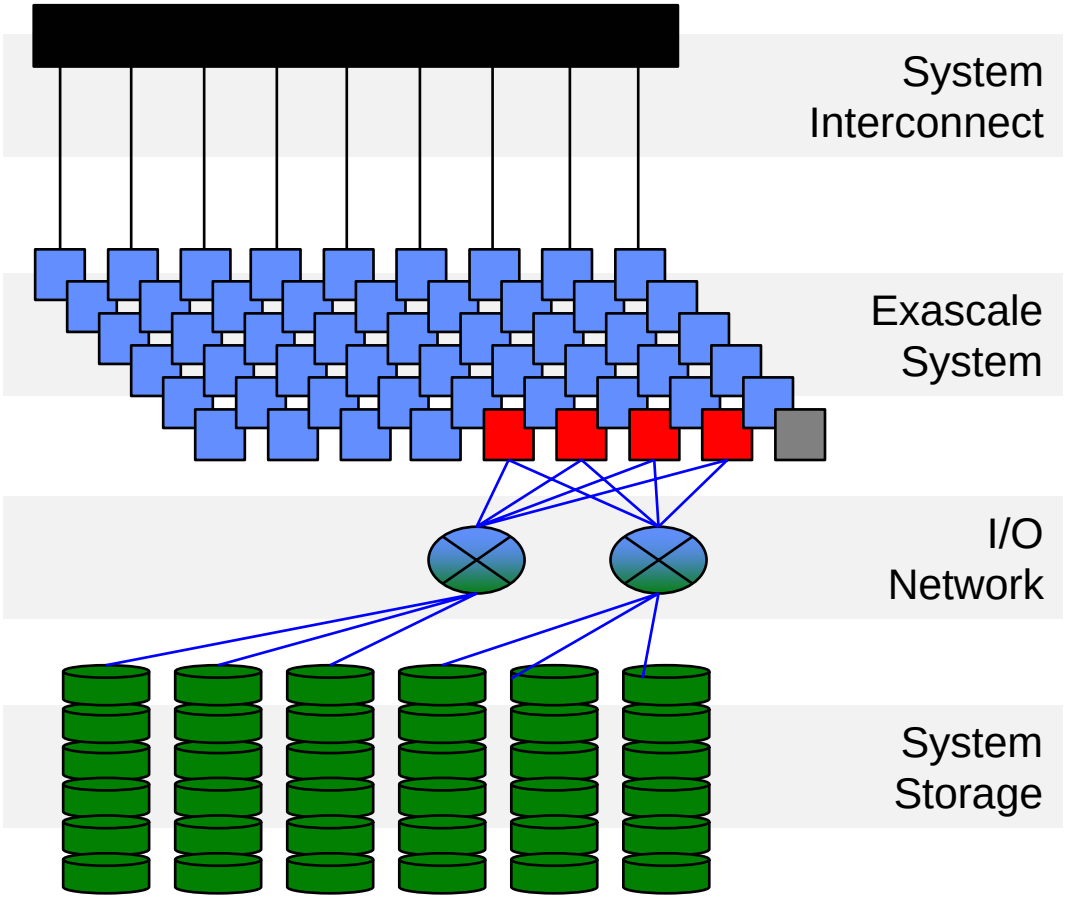
- **System power** is a first class constraint on exascale system performance and effectiveness.
- M**emory** is an important component of meeting exascale power and applications goals.
- **Programming model**. Early investment in several efforts to decide in 2013 on exascale programming model, allowing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

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	0.1TB/sec	1 TB/sec	0.4TB/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		200GB/sec	
MTTI	days	O(1day)		O(1 day)	

Sandia National Laboratories

The high level system design may be similar to petascale systems



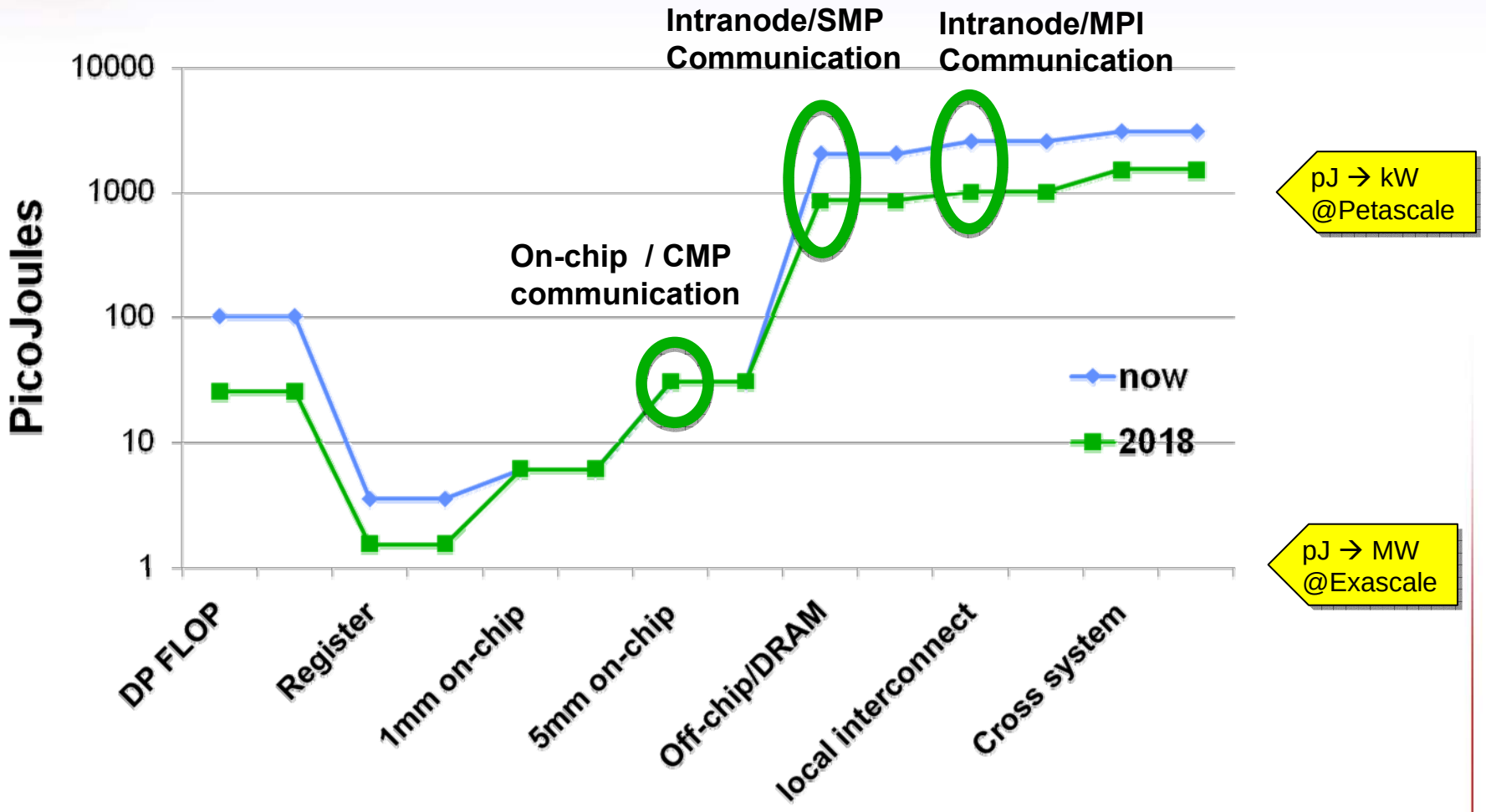
- New interconnect topologies
- Optical interconnect

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

- Mass storage far removed from application data

Sandia National Laboratories

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**

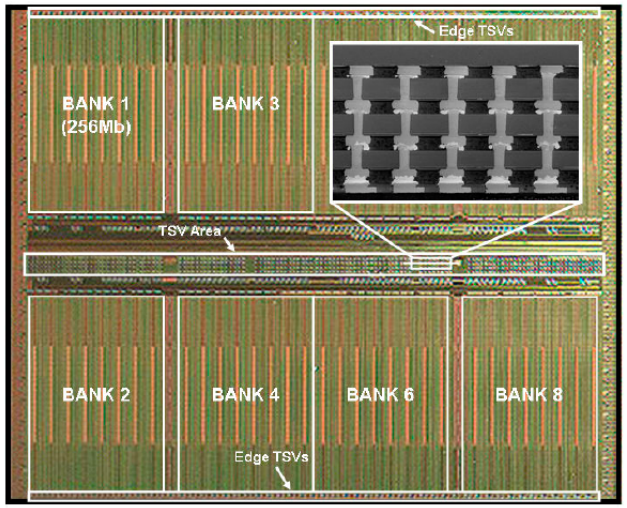


Figure 7.2.7: Die micrograph of the fabricated chip and cross-sectional view of TSVs. The chip size is 10.9×9.0mm².

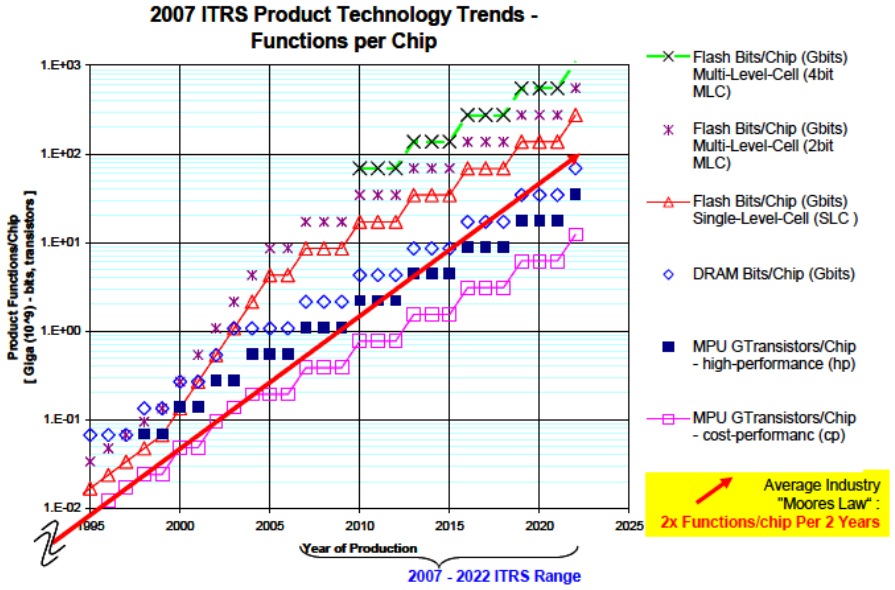
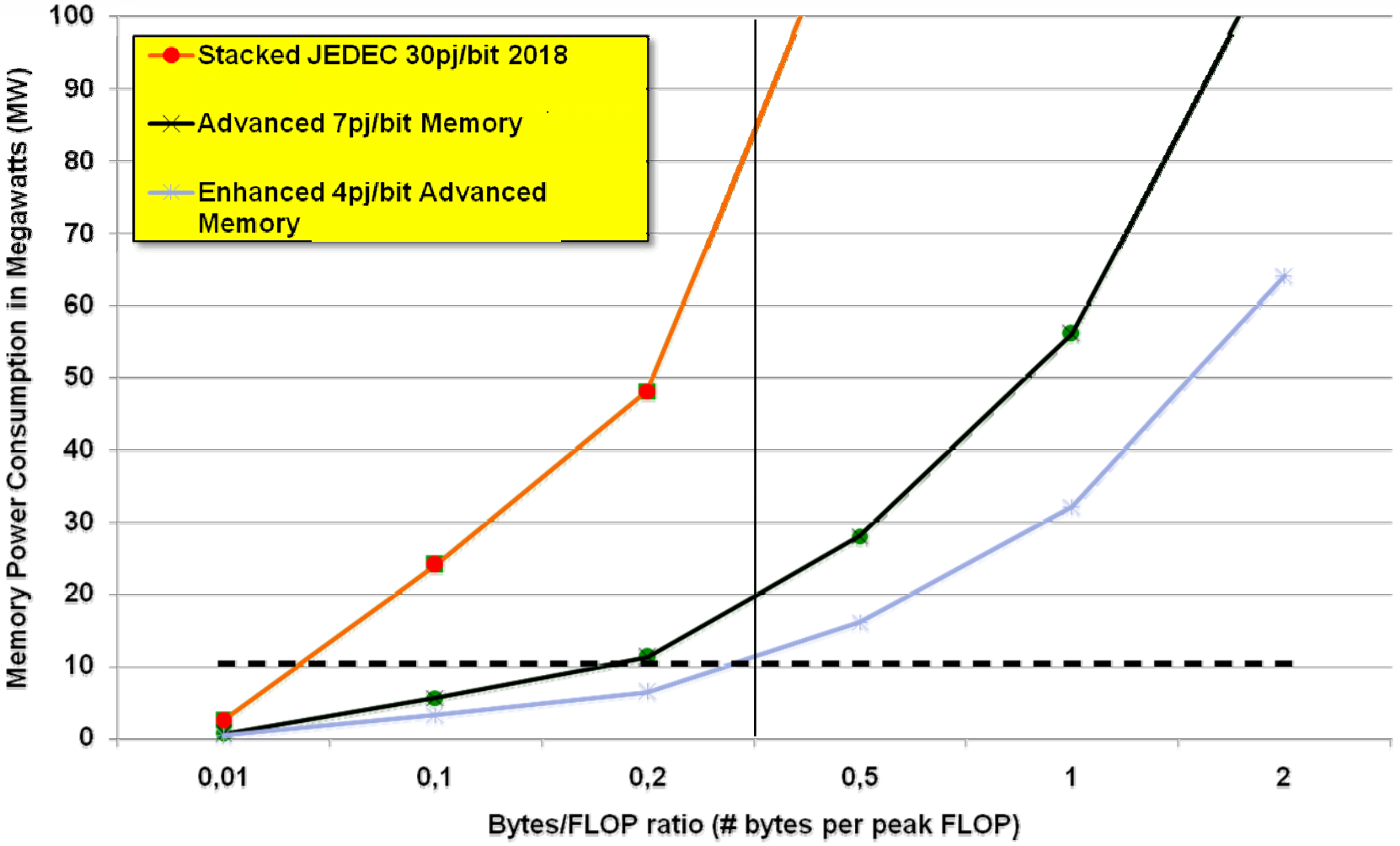


Figure ORTC2 ITRS Product Function Size Trends: MPU Logic Gate Size (4-transistor); Memory Cell Size [SRAM (6-transistor); Flash (SLC and MLC), and DRAM (transistor + capacitor)]--Updated

Investments in memory technology mitigate risk of narrowed application scope.

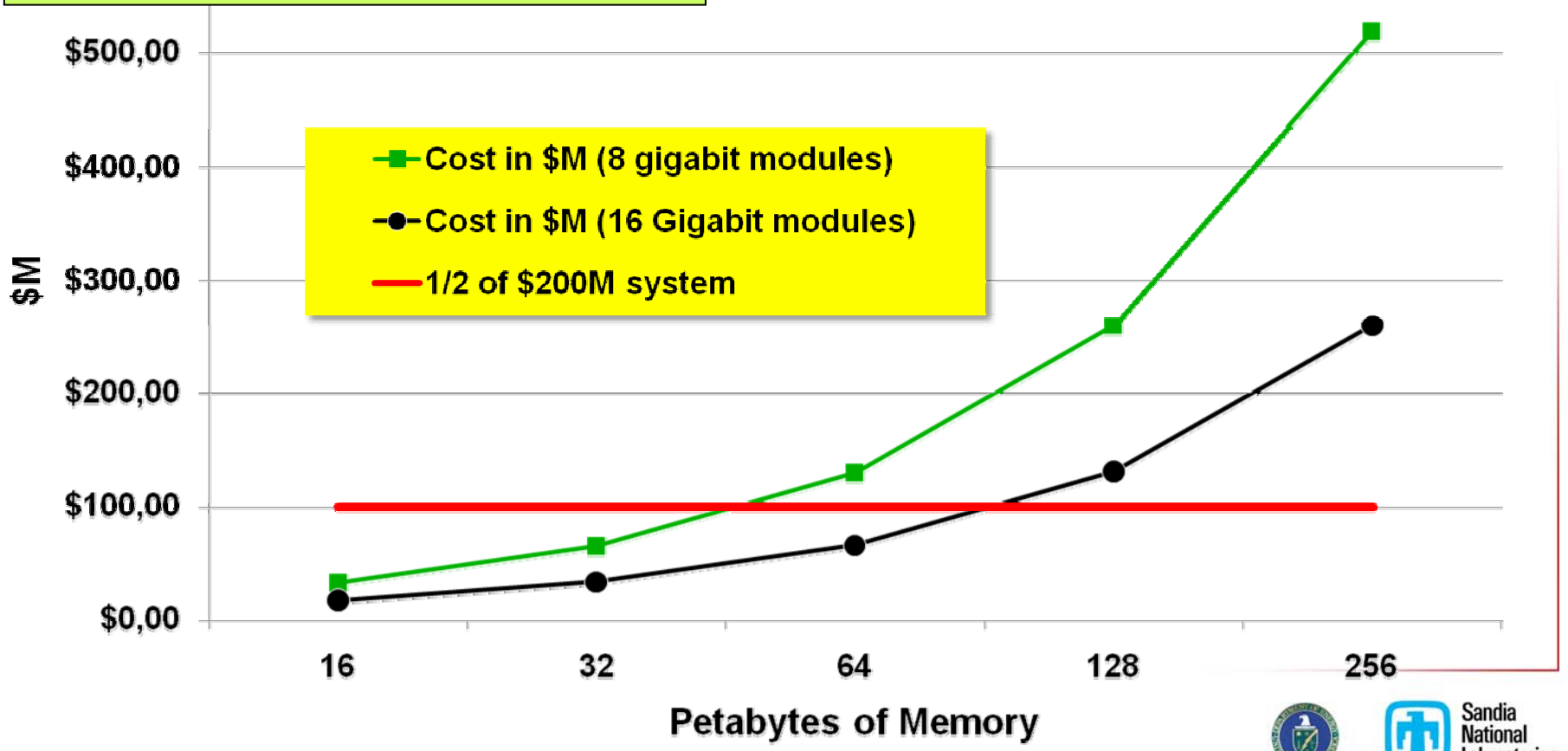


Sandia National Laboratories

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



Sandia National Laboratories

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:

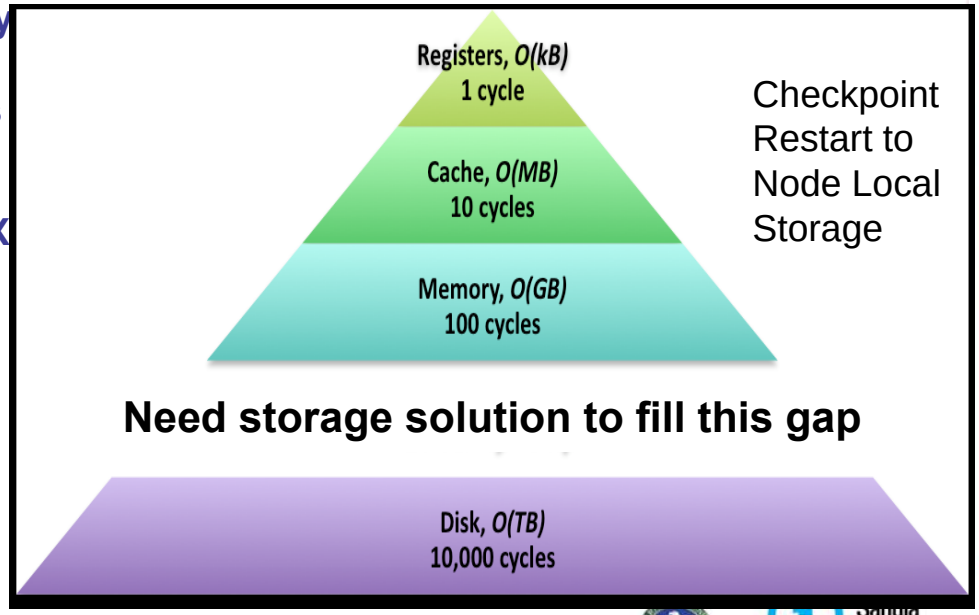
    10X improvement in hardware reliability

    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)

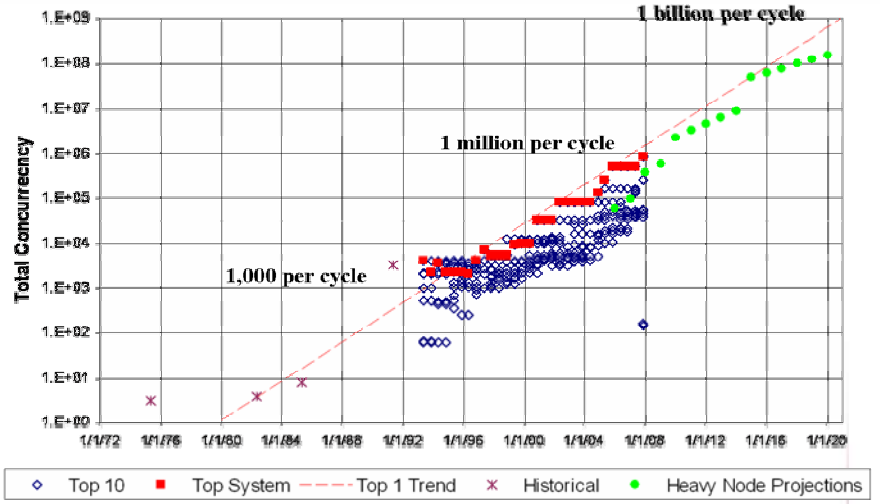
- **Hard errors**: permanent errors which cause system to hang or crash
- **Soft errors**: transient errors, either correctable or short term failure
- **Silent errors**: undetected errors either permanent or transient. *Concern is that simulation data or calculation have been corrupted and no error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



Sandia National Laboratories

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

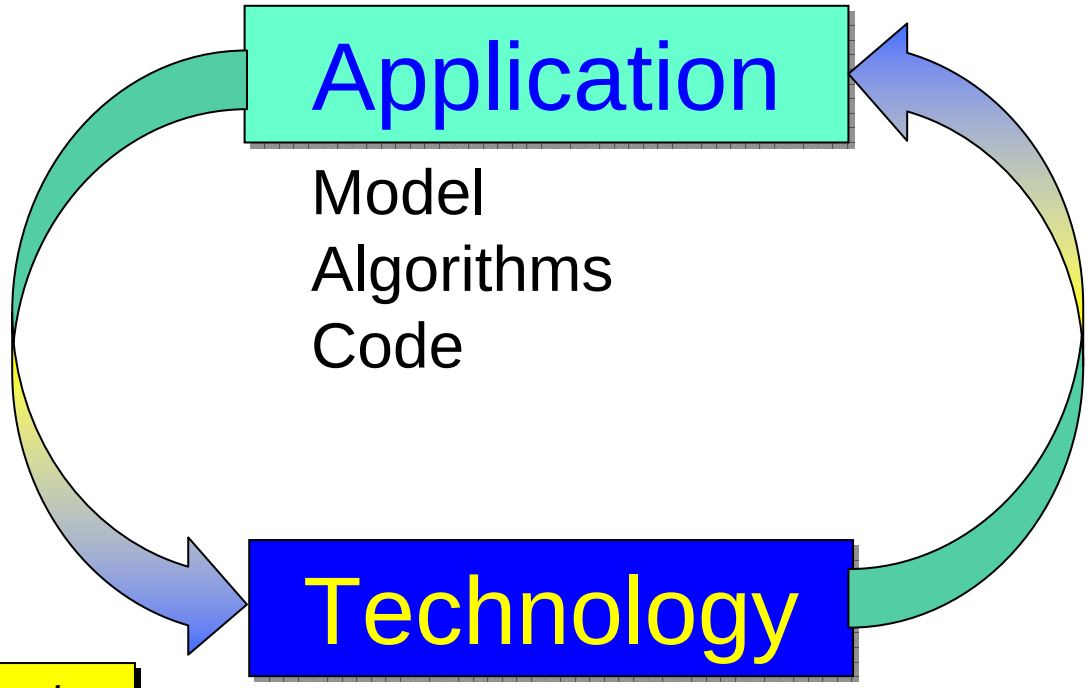
- **Technology Investments**
  - Extend inter-node models for scalability and resilience, e.g., MPI, PGAS (includes HPCS)
  - Develop intra-node models for concurrency, hierarchy, and heterogeneity by adapting current scientific ones (e.g., OpenMP) or leveraging from other domains (e.g., CUDA, OpenCL)
  - 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

Sandia National Laboratories

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

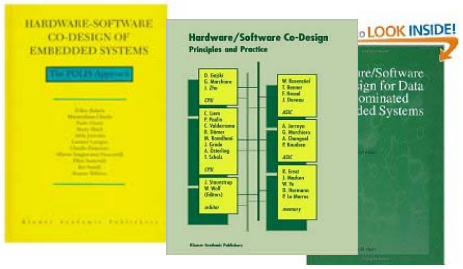
Now, we must expand the co-design space to find better solutions:
•new applications & algorithms,
•better technology and performance.

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

Sandia National Laboratories

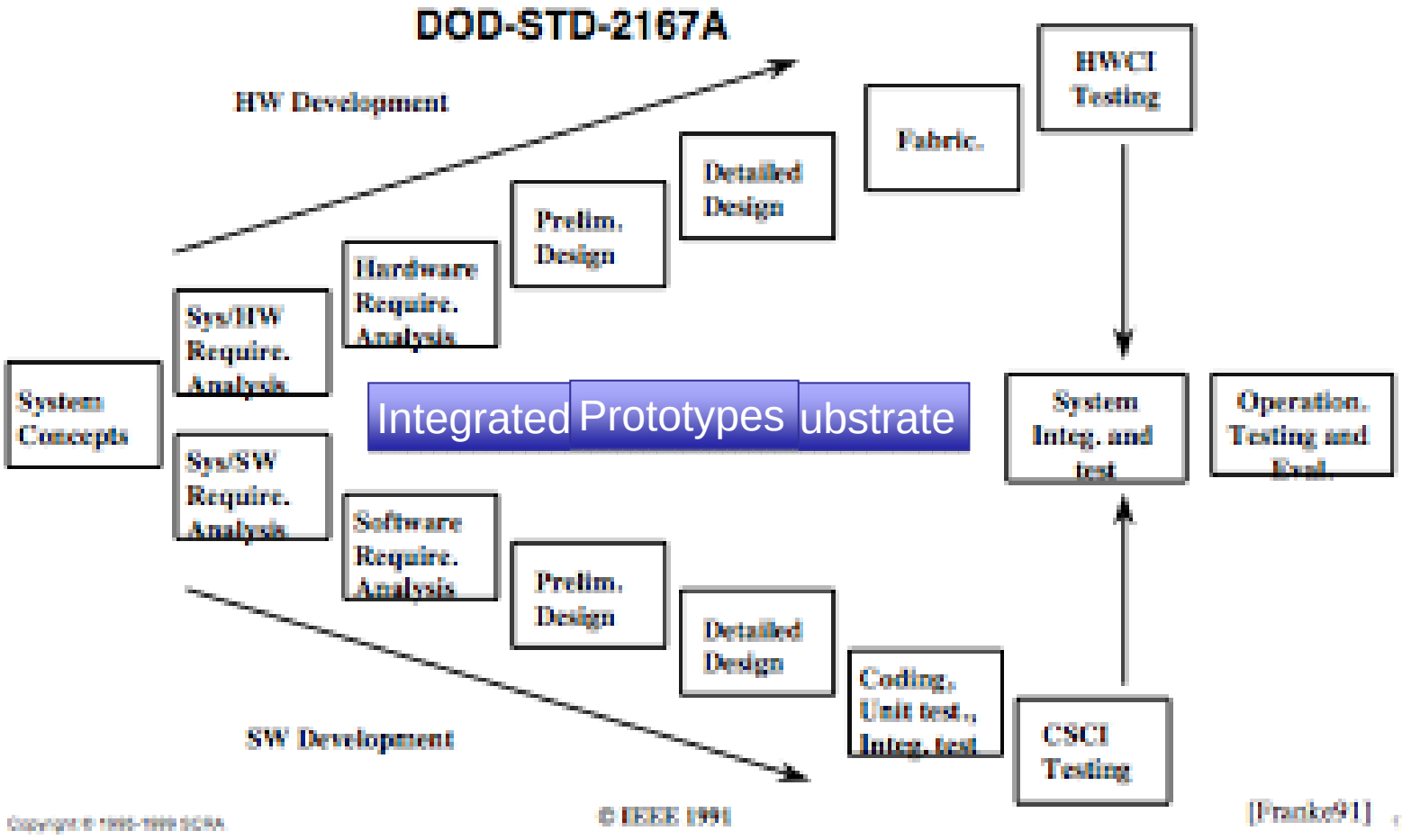
Hardware/Software co-design is a mature field in embedded computing

- Design of an integrated system that contains hardware and software
- Focus on embedded systems (cell phones, appliances, engines, controllers, etc.)
- Concurrent development of hardware and software
  - Interactions and tradeoffs
  - Partitioning is a focus
  - Must satisfy real-time and/or other performance/energy metrics/constraints



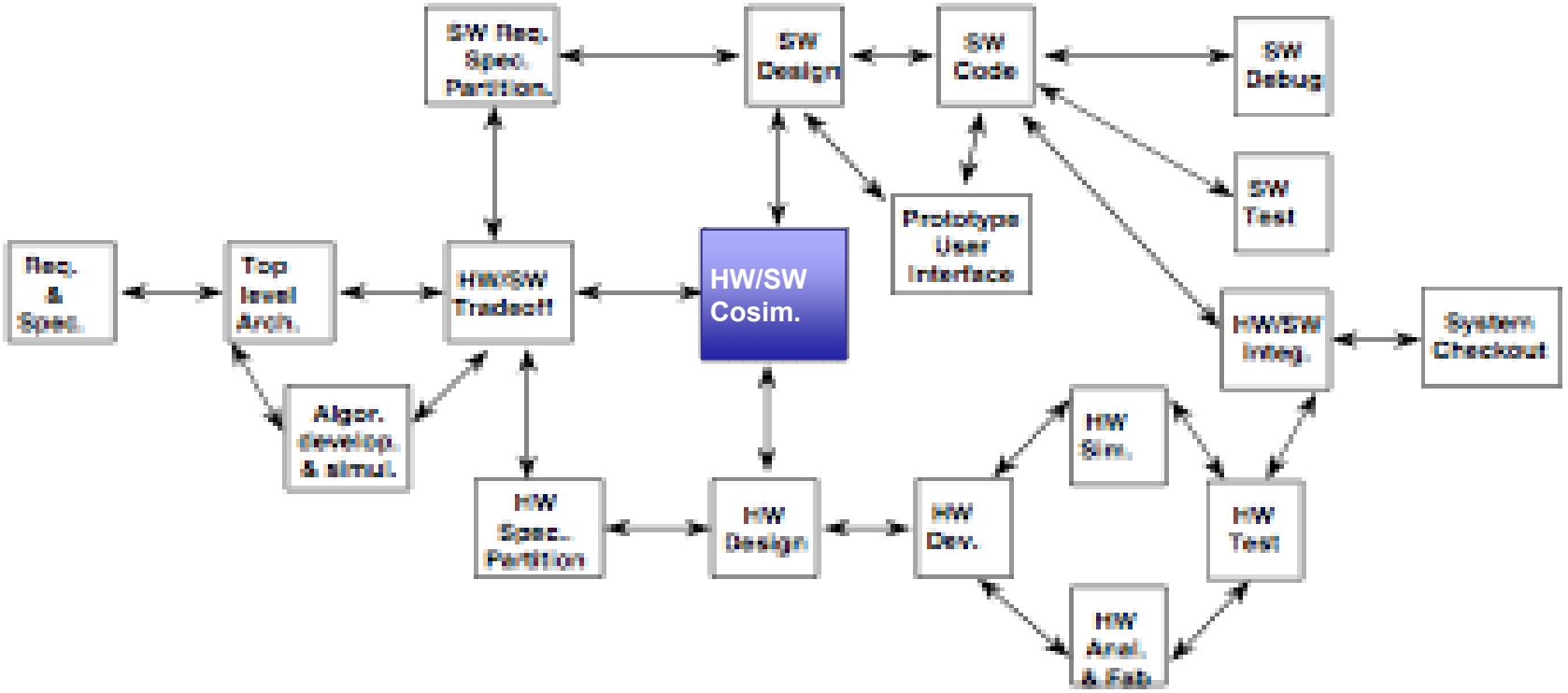
Sandia National Laboratories

Original DOD Standard for HW/SW co-development had shortcomings



Sandia National Laboratories

Lockheed Martin Co-design Methodology



Sandia National Laboratories

Why has co-design not been used more extensively in HPC?

- Leveraging of COTs technology
  - Almost all leadership systems have some custom components but HPC has benefited from the ability to leverage commercial technology
- HPC applications are very complex
  - May contain a million of lines of code
- ~15-20 years of architectural and programming model stability
  - Bulk synchronous processing + explicit message passing
- Lack of Adequate Simulation Tools
  - Often use Byte to Flop ratios and Excel spreadsheets
  - Industry simulation tools are proprietary

However, there are some HPC co-design examples and there are useful tools

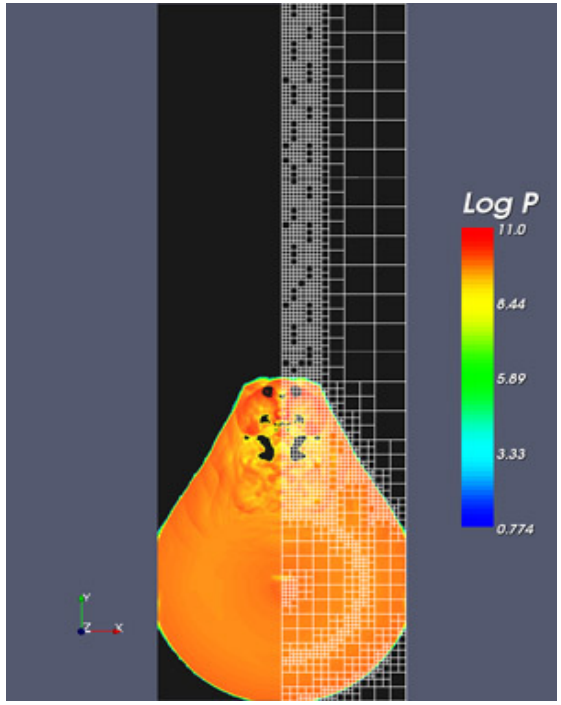
Basic performance modeling

CTH is DoD's most used code

Basic CTH Model

$$T = E(\kappa,\varphi)N^3 + C(\lambda + \tau k N^2) + S(\gamma log(P)) + L_{imbal}$$

- T is the execution time per time step
- N is size of an edge of a processor's subdomain
- C and S are number of exchanges and collectives
- P is the number of processors
- *k* is the number of variables in an exchange
- λ and τ are latency and transfer cost
- γ is the cost of one stage of collective
- $E(\kappa,\varphi)$ is the calculation time per cell
- L_{imbal} is a new term representing effects of load imbalance

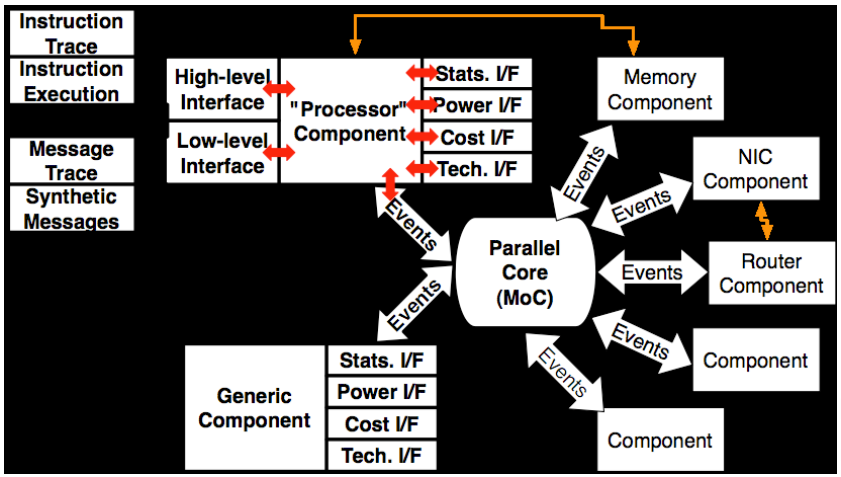


Limitations:

- Very simple architectural model
- Tuning parameters
- Need a new model when you change the application

Sandia National Laboratories

SST Simulation Project



- Parallel
- Parallel Discrete Event core with conservative optimization over MPI
- Holistic
- Integrated Tech. Models for power
- McPAT, Sim-Panalyzer
- Multiscale
- Detailed and simple models for processor, network, and memory

- Current Release (2.0) at

http://www.cs.sandia.gov/sst/

- Includes parallel simulation core, configuration, power models, basic network and processor models, and interface to detailed memory model

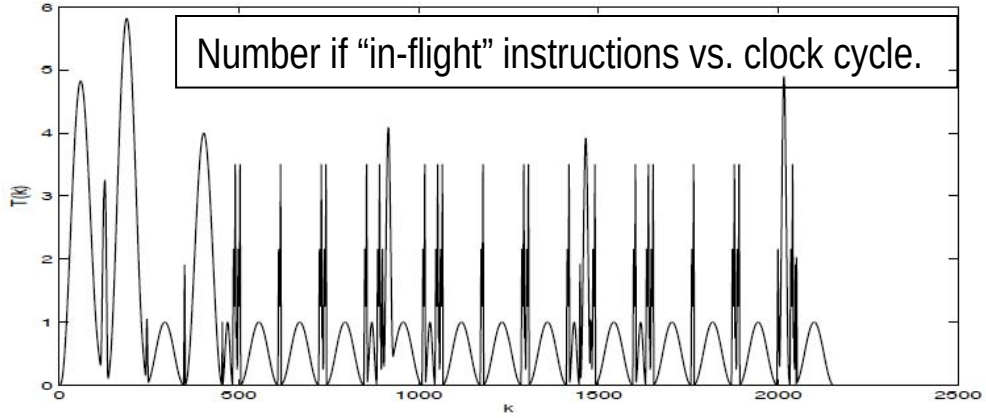
Sandia National Laboratories

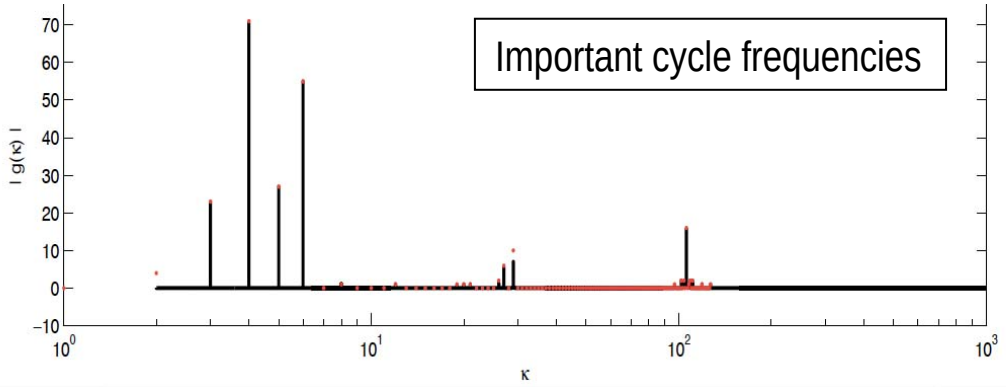
SST simulations have quantified the impact of the Memory Wall

- Most of DOE's Applications (e.g., climate, fusion, shock physics, …) spend most of their instructions accessing memory or doing integer computations, not floating point
- Additionally, most integer computations are computing memory Addresses
- Advanced development efforts are focused on accelerating memory subsystem performance for both scientific and informatics applications

SST is providing architectural insights to algorithms developers

- **Input: SST Trace for SpMV.**
- **Lots of instruction stream data.**
- **Model: Use restricted *sin*^{2}function to mark start/finish of each instruction.**
- **Use FFTs to analyze behavior.**





Trace fragment from SpMV inner loop

j	I_j	issue	complete	κ
59	bc	737	741	4
60	lwz	738	744	6
61	lfd	740	746	6
62	addi	742	746	4
63	addi	742	746	4
64	rlwinm	743	746	3
65	lfdx	744	850	106
66	fmadd	849	854	5
67	bc	850	854	4
68	lwz	851	857	6
69	lfd	853	859	6
70	addi	855	859	4
71	addi	855	859	4
72	rlwinm	856	859	3
73	lfdx	857	886	29
74	fmadd	885	890	5
75	bc	886	890	4
76	lwz	887	893	6
77	lfd	889	895	6
78	addi	891	895	4
79	addi	891	895	4
80	rlwinm	892	895	3
81	lfdx	893	899	6
82	fmadd	898	903	5
83	bc	899	903	4

Need to define HPC co-design methodology

- Could range from discussions between architecture, software and application groups to tight collaboration centered on the co-simulation of hardware and applications
- Opportunity to influence future architectures
  - Cores/node, threads/core, scheduling width/thread
  - Logic in memory subsystem
  - Interconnect performance
- HPC community must work together to define the next programming model