SAND2010-3666P

Advanced Architectures for Exascale Computing

Ron Brightwell

Scalable System Software

Predictive Engineering Science Panel (PESP)

Sandia National Laboratories

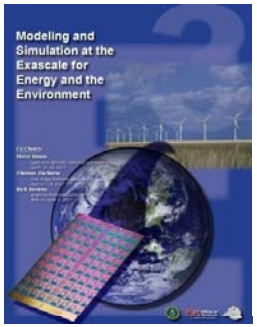
Albuquerque NM

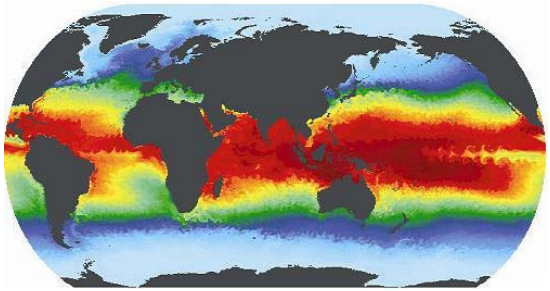
June 2-4, 2010

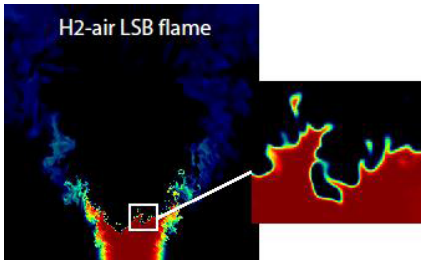
Sandia is a multi program laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Process for Identifying Exascale Applications and Technology Ensures Broad Community Support

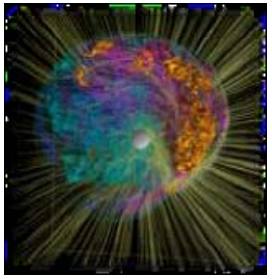
- **Town Hall Meetings: April-June 2007**
- **Scientific Grand Challenges**
- **Workshops: Nov'08 – Oct'09**
  - Climate Science (11/08)
  - High Energy Physics (12/08)
  - Nuclear Physics (1/09)
  - Fusion Energy (3/09)
  - Nuclear Energy (5/09)
  - Biology (8/09)
  - Material Science and Chemistry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







Mission Imperatives

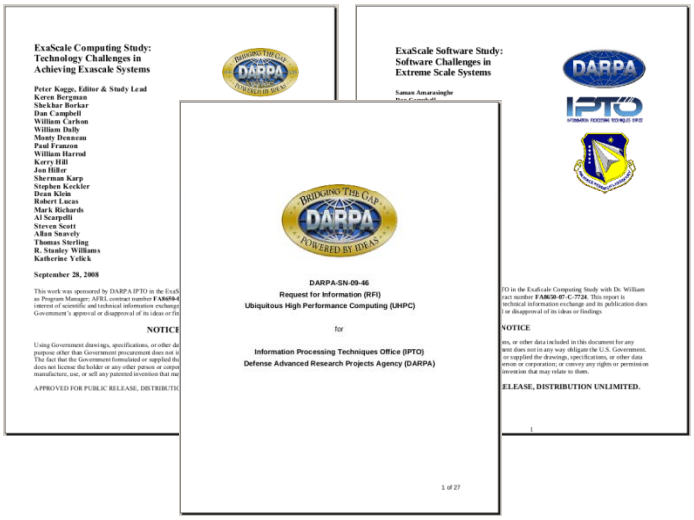


Fundamental Science

Sandia National Laboratories

Other Agencies Also Pursuing Exascale

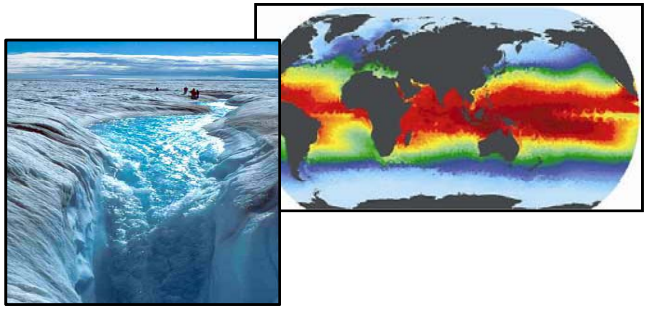
- **DARPA**
  - Exascale Hardware Report
  - Exascale Software Report
  - Ubiquitous High-Performance Computing BAA

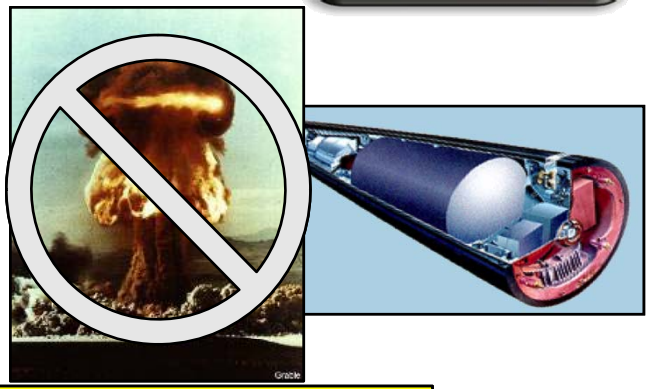


- **NSF**
  - G8 Research Councils Initiative on Multilateral Research, Interdisciplinary Program on Application Software towards Exascale Computing for Global Scale Issues.

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.

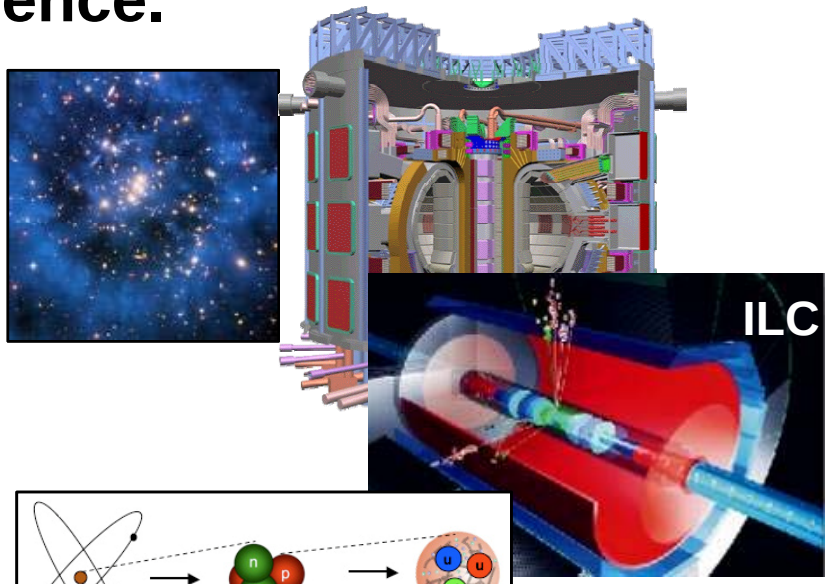
ASC

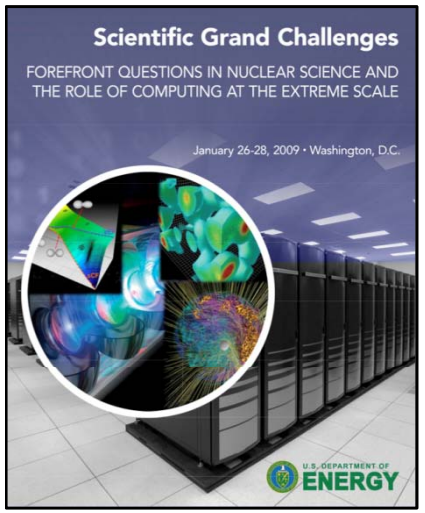
Sandia National Laboratories

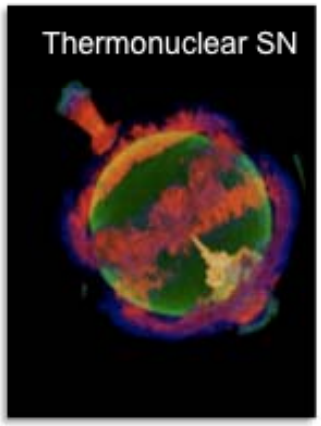
Exascale simulation will enable fundamental advances in basic science.

- **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.

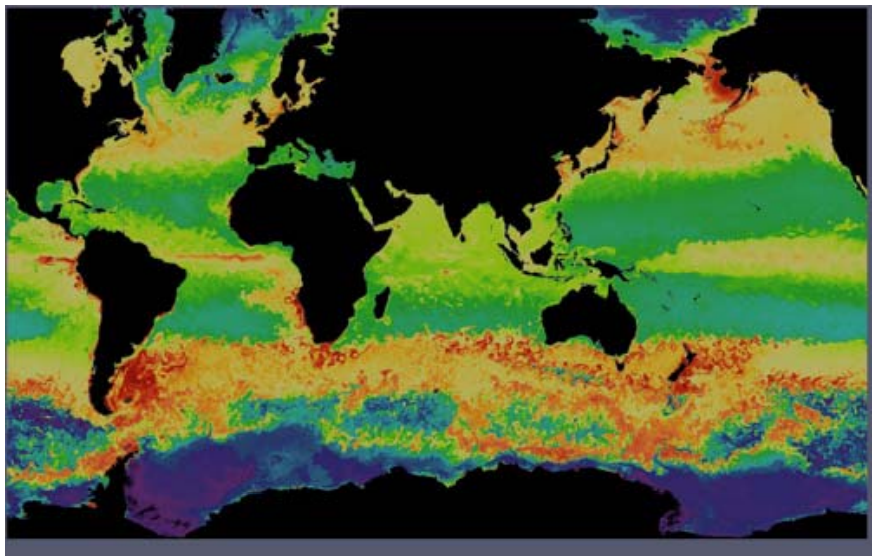






Exascale resources are required for predictive climate simulation.

- **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

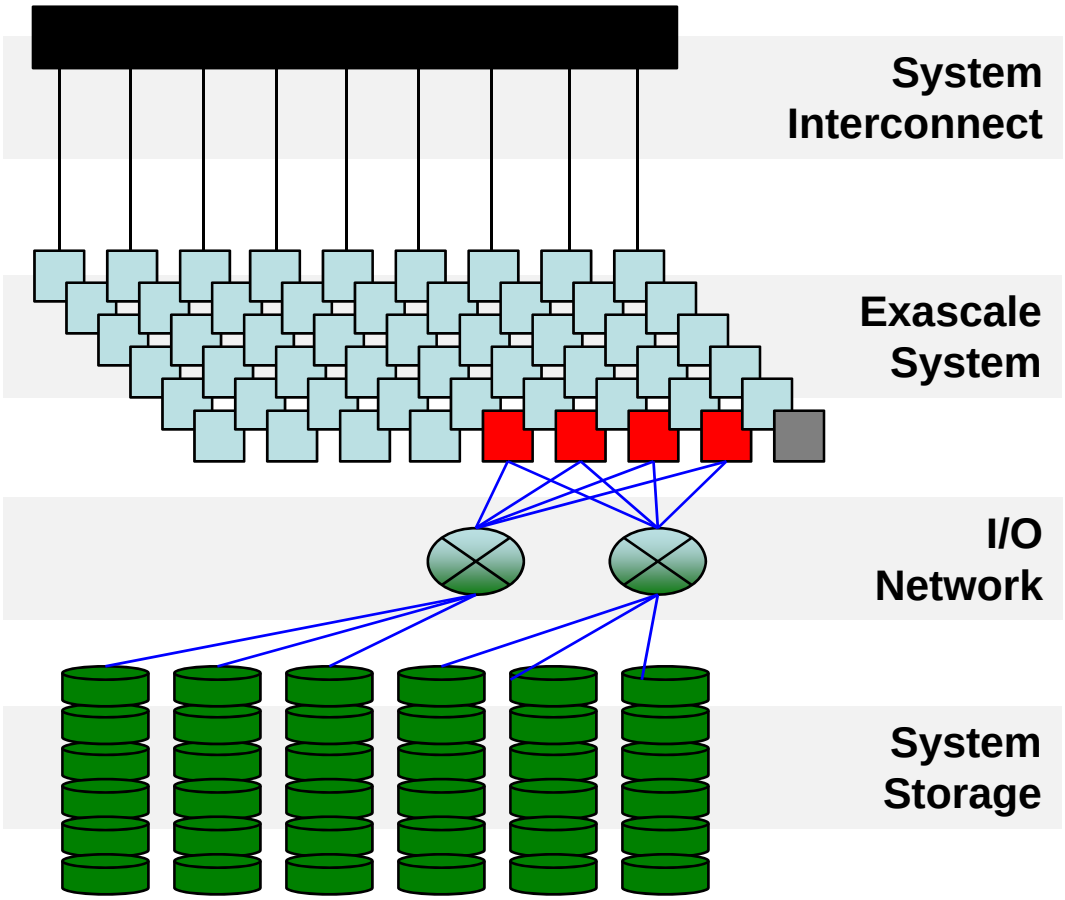
What are critical exascale technology investments?

- 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.

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

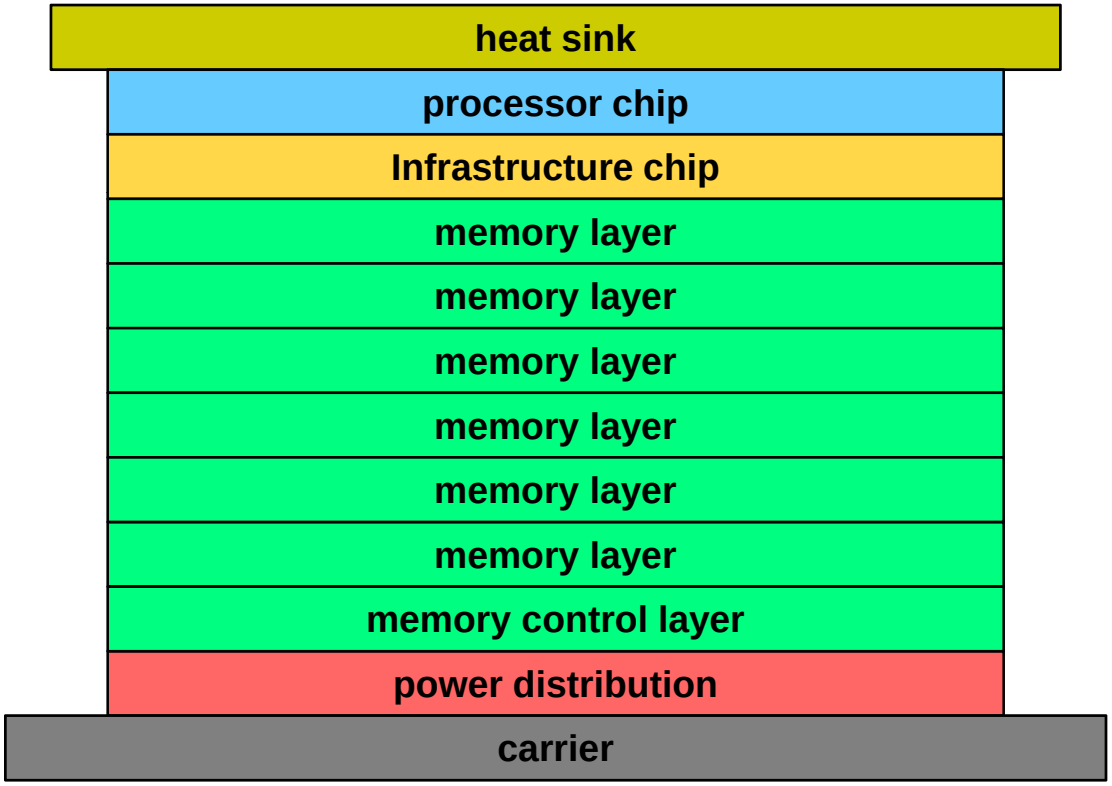


- New interconnect topologies
- Optical interconnect

- 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 beyond exascale



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

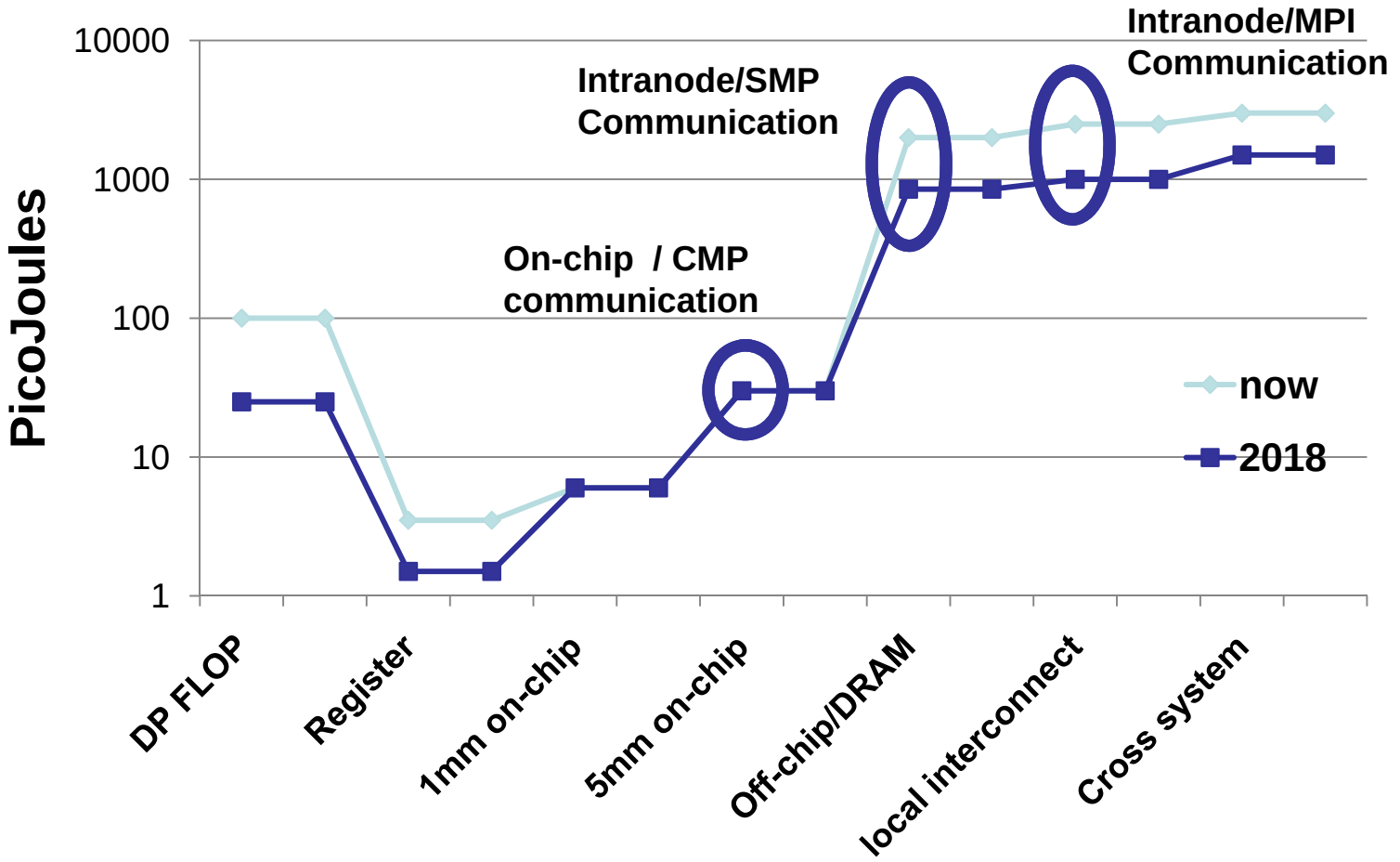
- 3d stacked memory

- Smart memory management

- Integration on package

Sandia National Laboratories

Investments in architecture R&D and application locality are critical

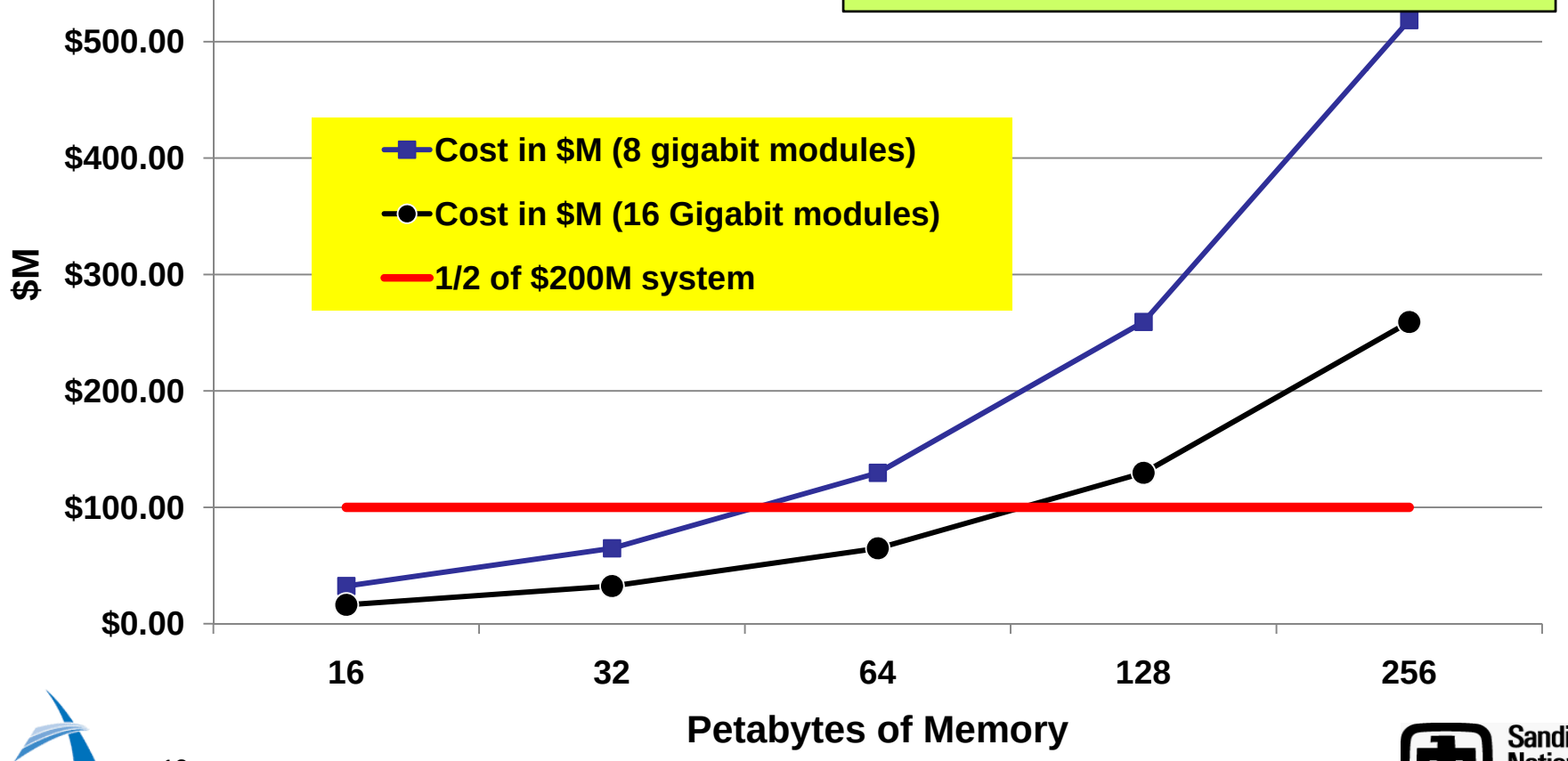


"The Energy and Power Challenge is the most pervasive … and has its roots in the inability of the [study] group to project any combination of currently mature technologies that will deliver sufficiently powerful systems in any class at the desired levels."
DARPA IPTO exascale technology challenge report

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

Factors Driving Up 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 (1023 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:

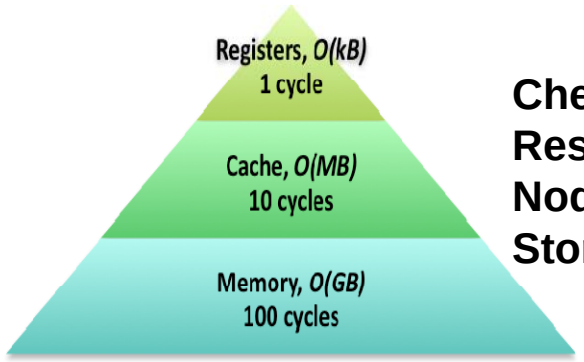
    10X improvement in hardware reliability

    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.***



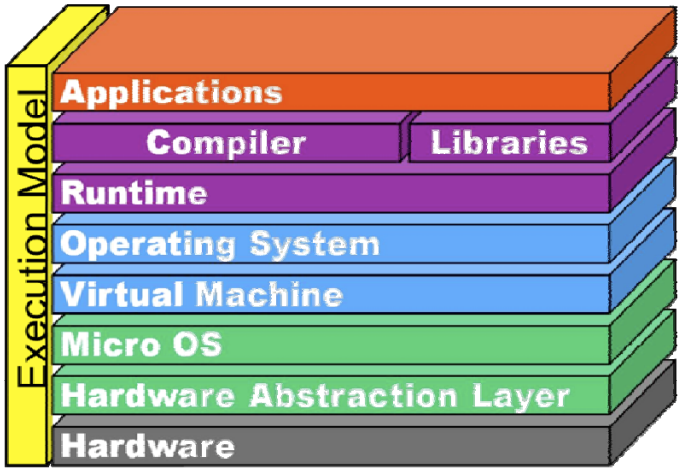
Checkpoint Restart to Node Local Storage

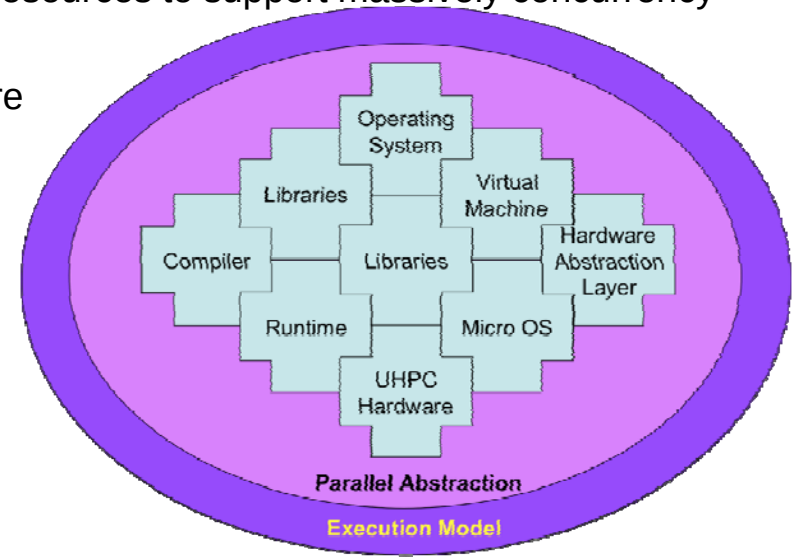
Need storage solution to fill this gap

Sandia National Laboratories

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





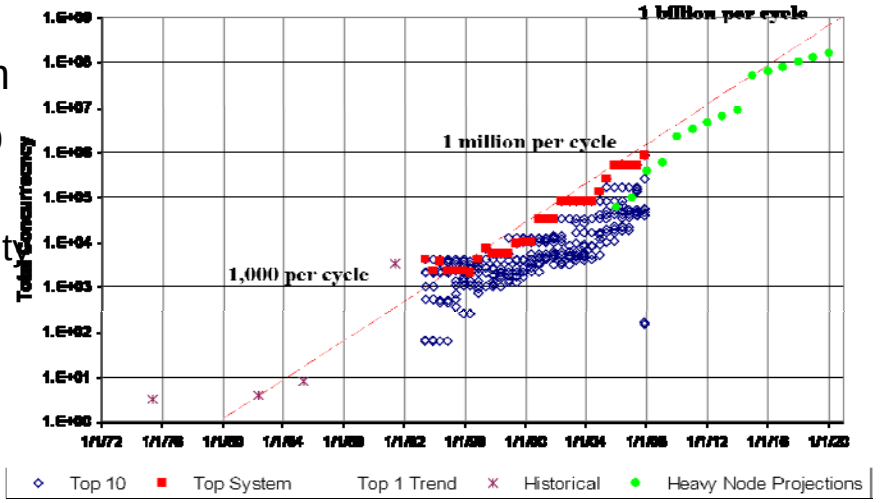
Software challenges in extreme scale systems, ***Sarkar, 2010***

15

Programming models and environments require early investment.

- 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
- **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
- **Technology Investments**

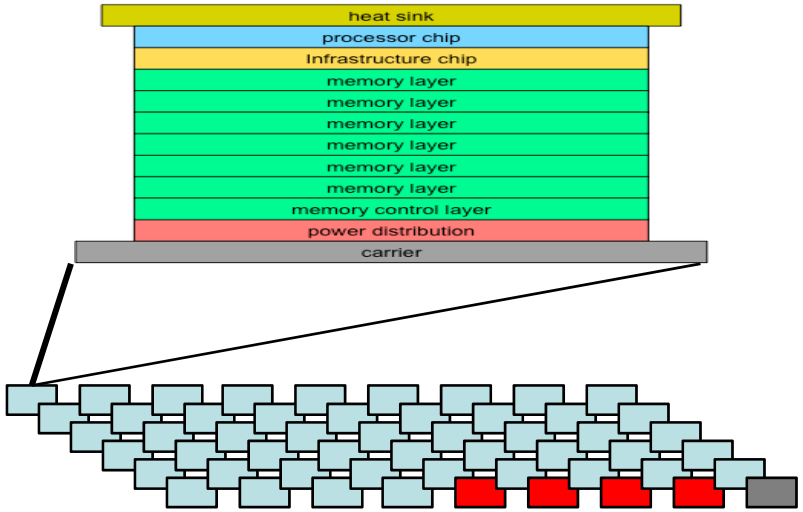


How much parallelism must be handled by the program?
From Peter Kogge (on behalf of Exascale Working Group), "Architectural *Challenges* at the Exascale Frontier", June 20, 2008

Sandia National Laboratories

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



Co-design is a key element of the Exascale strategy

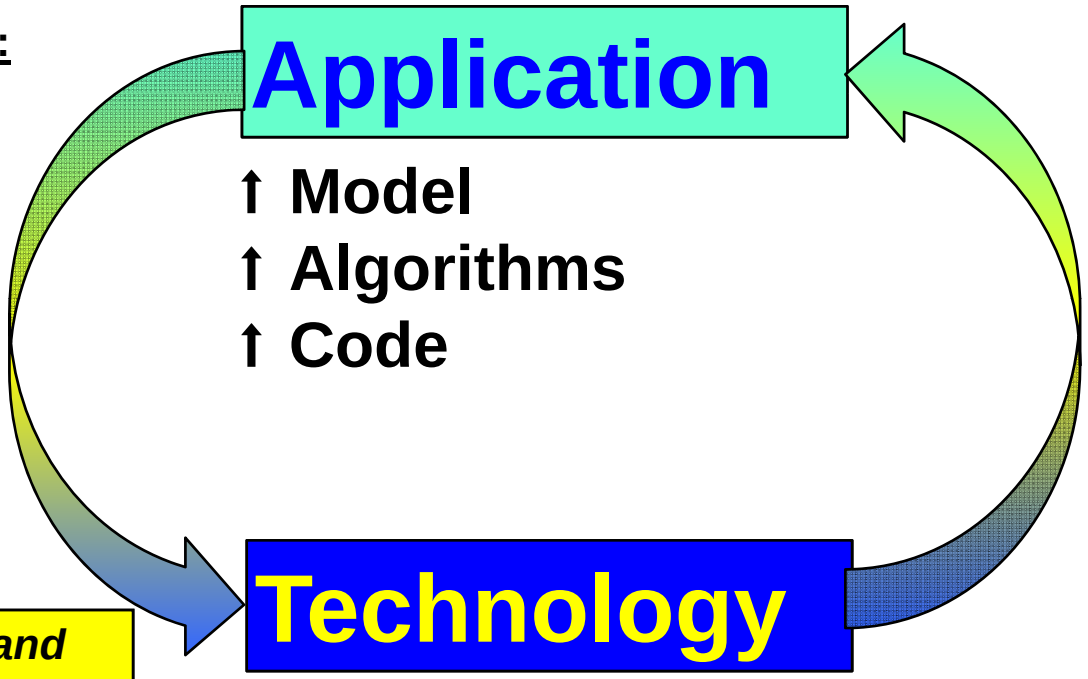
- **Architectures are undergoing a major change**
  - Single thread performance is remaining relatively constant and on chip parallelism is increasing rapidly
  - Hierarchical parallelism, heterogeneity
  - Massive multithreading
  - NVRAM for caching IO
- **Applications will need to change in response to architectural changes**
  - Manage locality and extreme scalability (billion-way parallelism)
  - Potentially tolerate latency
  - Resilience?
- **Unprecedented opportunity for applications/algorithms to influence architectures, system software and the next programming model**
  - Hardware R&D is needed to reach exascale
- **We will not be able to solve all of the exascale problems through architectures work only**

Co-design space is subject to other constraints

- **Power, system cost, R&D costs**
- **Physical limitations**
- **Multiple applications**
- **Goal is to build a sustainable infrastructure with broad market support**
  - ***Extend beyond natural evolution of commodity hardware to create new markets***
  - ***Create system building blocks that offer superior price/performance/programmability all scales (exascale, departmental and embedded)***

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

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



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

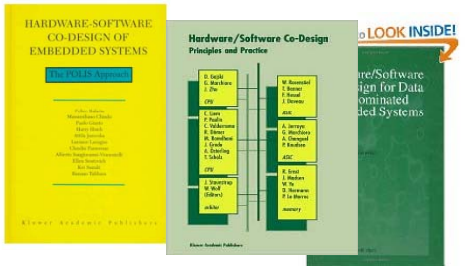
Technology driven:
Fit your application to this technology.
Sub-optimal.

ASC

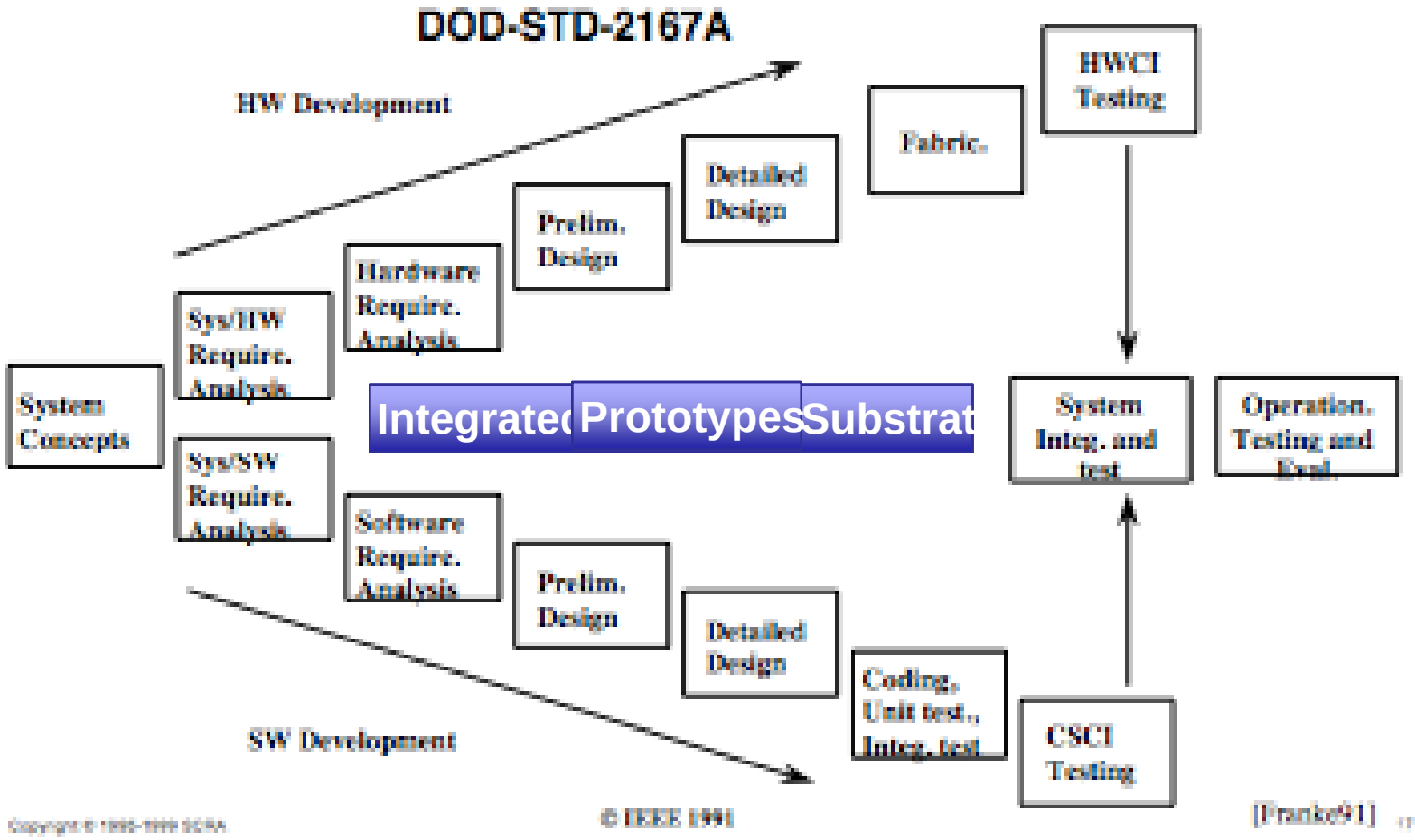
Sandia National Laboratories

Hardware/Software co-design is a mature field

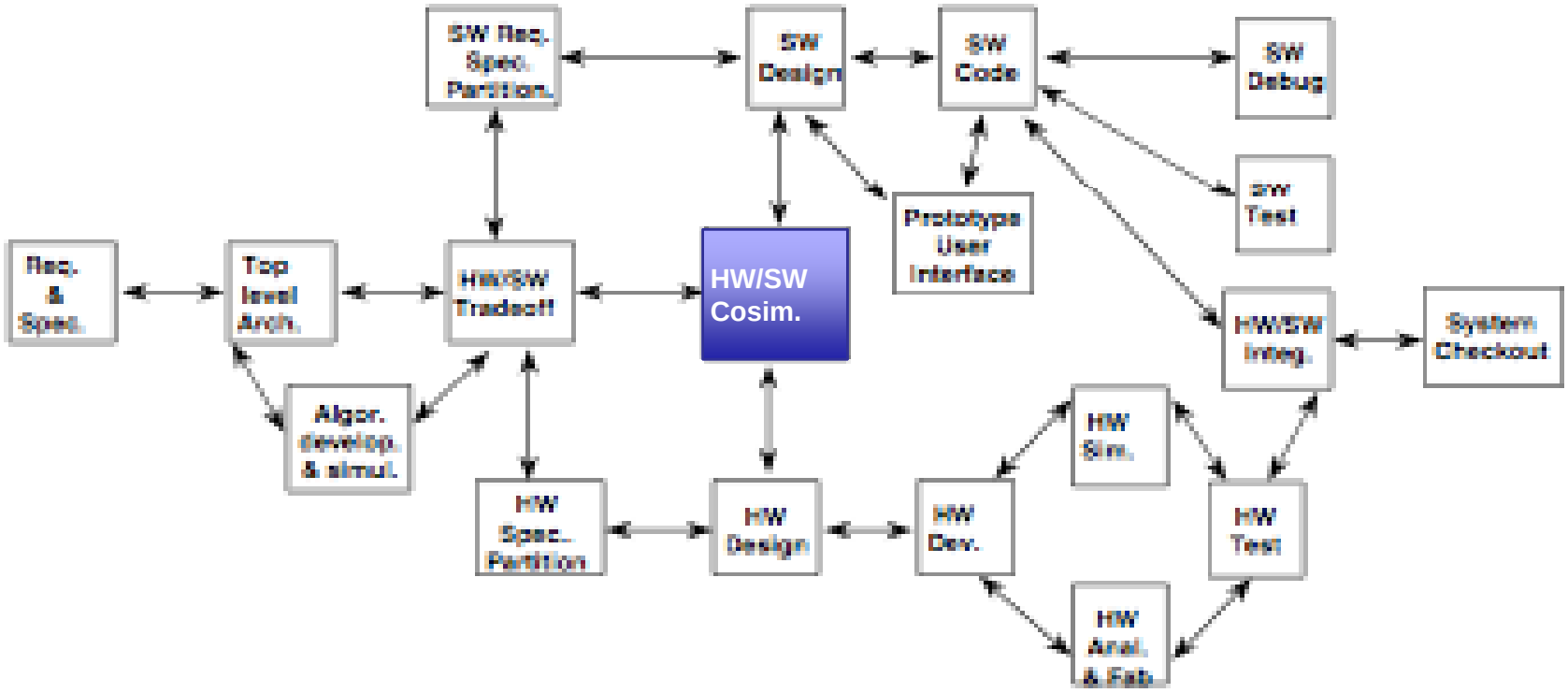
- **Design of an integrated system that contains hardware and software**
- **Focus on embedded systems (cell phones, appliances, engines, controllers, etc.)**
- **Concurrent development of hardware and software**
  - Interactions and tradeoffs
  - Partitioning is a focus
  - Must satisfy real-time and/or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



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



Lockheed Martin Co-design Methodology



Sandia National Laboratories

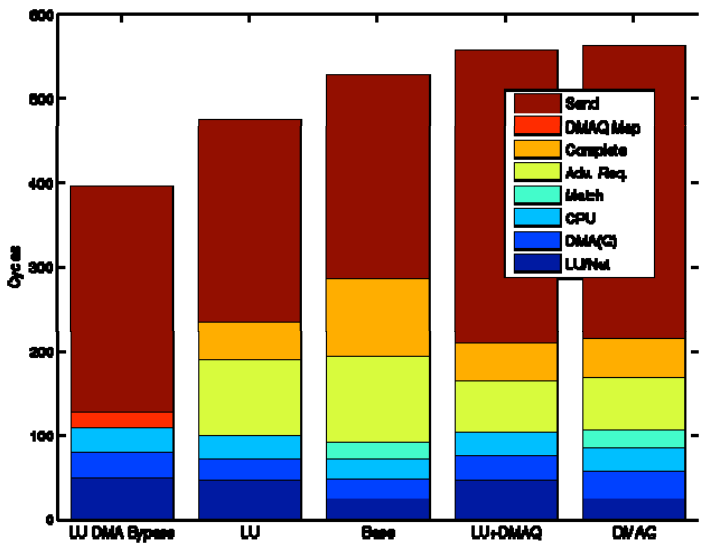
Why has co-design not been used more extensively in HPC?

- **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

NIC Architecture Co-design

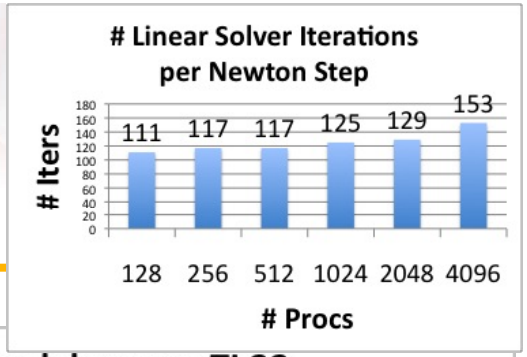
- Prevailing architectural constraints have driven many applications to highly bursty communication patterns
- In a power constrained world this trend will be unsustainable due to inefficient use of the system interconnect
- Design Goal: Produce a NIC architecture that enables overlap through high message rates and independent progress

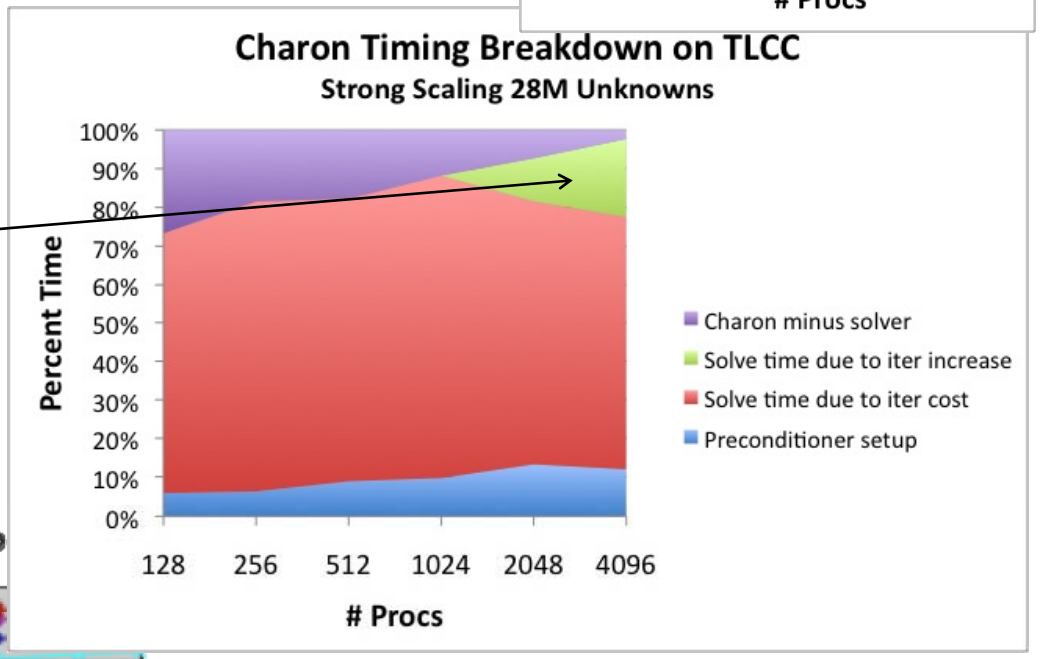


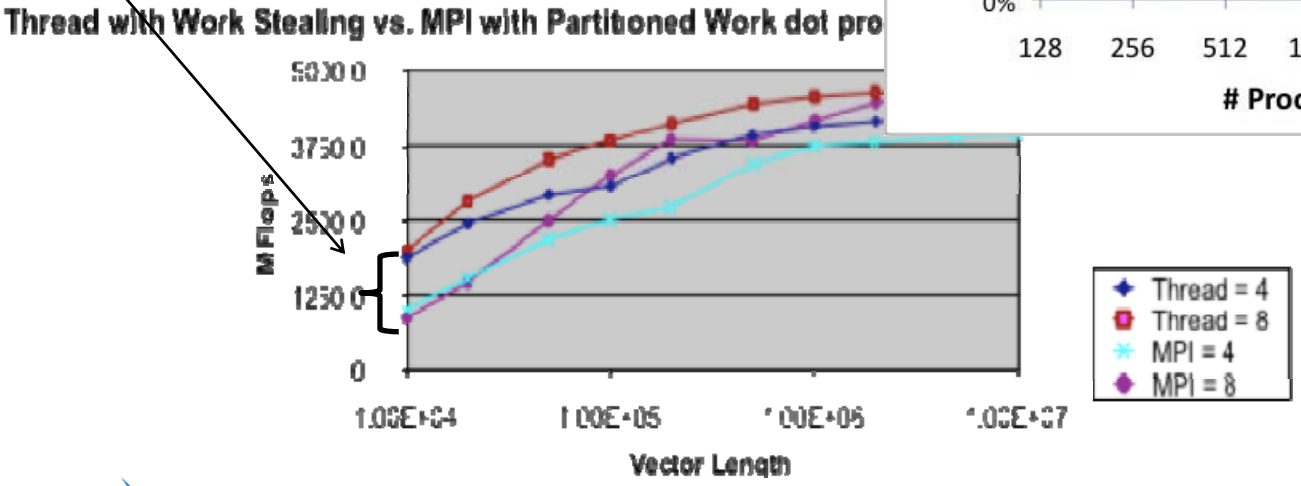
- Using simulation, NIC hardware & software and host driver software were simultaneously profiled for various architecture choices
- Trade-offs:
  - Which architectural features provide performance advantages
  - What software bottlenecks need to be moved to hardware
  - Which functions can be left to run on NIC CPU or in the host driver
- Next step: rework applications (or portions) to take advantage of the new features and provide feedback for more architectural improvements

Co-design of algorithms and runtime libraries (IAA)

- **Selective Shared Memory Use:**
  - **App: 4096 MPI tasks**
  - **Solver: 256 MPI tasks, 16-way threading**
- **Robustness:**
  - **117 iterations (not 153)**
  - **Eliminates green region**
- **Speed: Threading (carefully used):**
  - **Same asymptotic speed as MPI**
  - **Faster ramp up: 2X or more**
  - **Better load imbalance tolerance**







Bottom line: Hybrid parallelism promises better:

-Robustness,

-Strong scaling and

-Load balancing.

ASC

Sandia National Laboratories

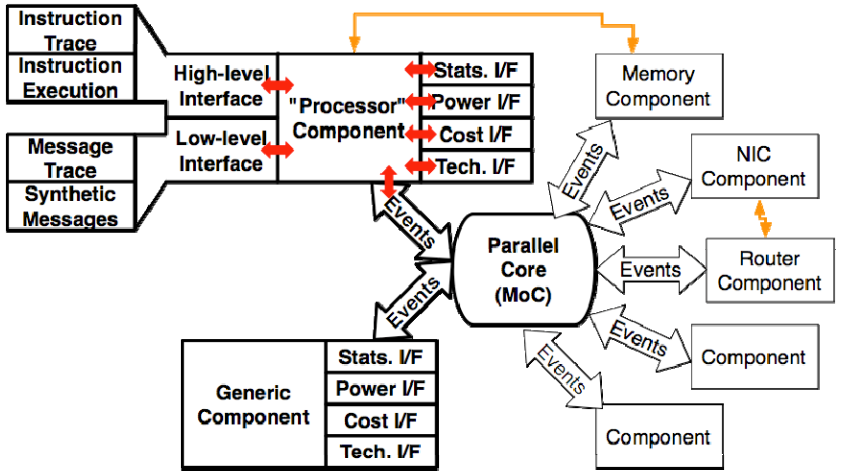
System Simulation Workshop was co-sponsored by IAA and LBNL

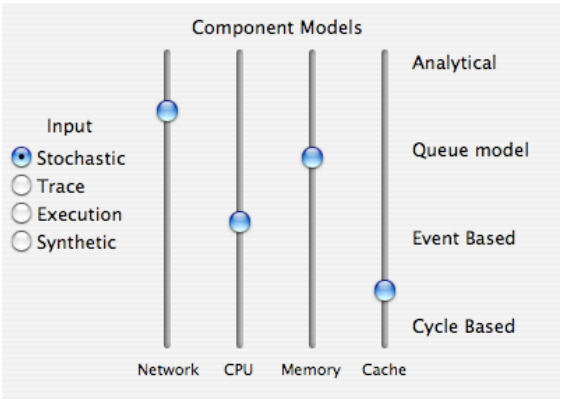
September 2009 in Boulder, CO

- **Goals**
  - **Define requirements for a common community product for HPC architectural simulation.**
  - **Addresses the needs of multiple HPC communities**
  - **Form the foundation of an ecosystem for HPC simulation**
- **Audience: ~50 participants from labs, universities, industry**
  - **Brought together simulation "providers" and "customers"**
- **Findings**
  - **HPC is faced by fundamentally new challenges (Hardware, Software, Scale, Power) and needs new simulation capabilities to confront them**
  - **Many simulation models share common implementation themes**
  - **Most participants believe a common simulation platform is desirable, beneficial and technically feasible**
  - **Verification and Validation are major concerns**

System simulation should be a key enabling technology

- **Co-simulation of hardware and software**
  - Assess architectural choices and their impact on applications
  - Identify bottlenecks and enable the development of algorithms for future architectures
- **Key features**
  - Open source with the ability to interface to proprietary software
  - Holistic: performance, power, area, cost, reliability analysis
  - Modular and multiscale (cycle accurate to analytical)
  - Input traces as well as joint execution
  - Parallel
  - FPGA acceleration





Sandia National Laboratories

SST Simulation Project

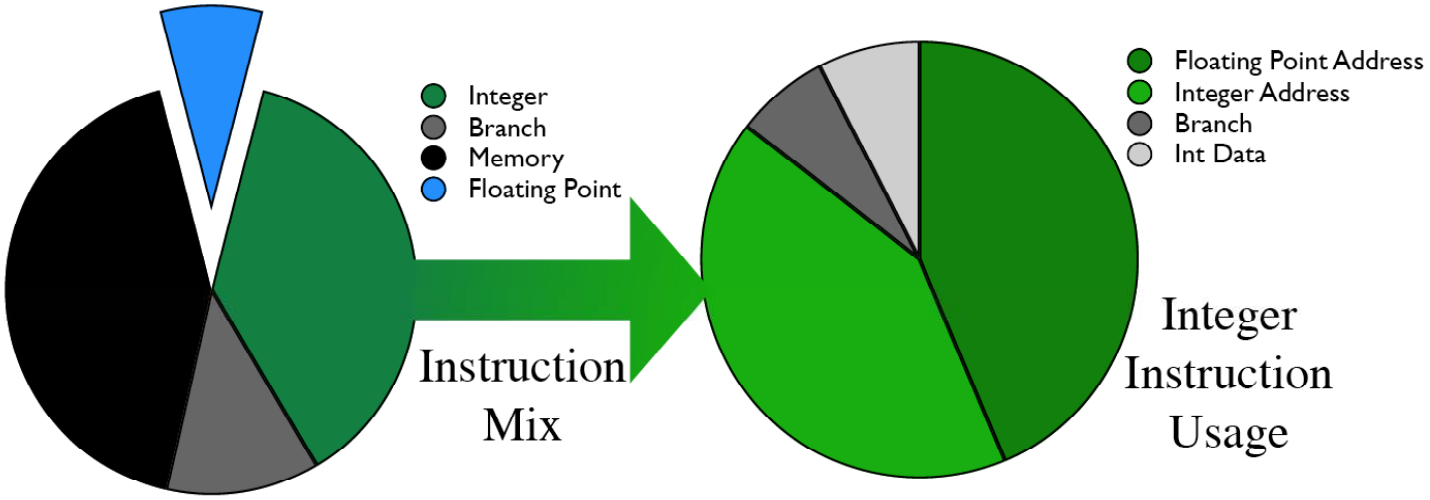
- Parallel
- Parallel Discrete Event core with conservative optimization over MPI
- Holistic
- Integrated Tech. Models for power
- McPAT, Sim-Panalyzer
- Multiscale
- Detailed and simple models for processor, network, and memory

- 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

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**

Conclusion:
Need to define HPC co-design methodology

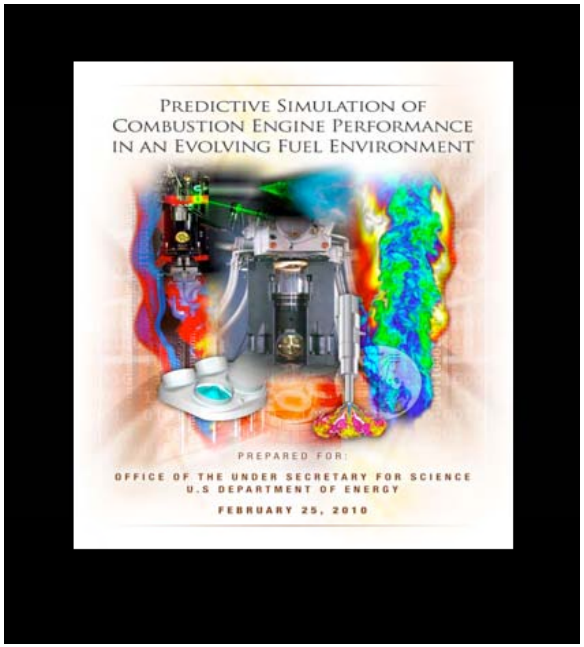
- **Could range from discussions between architecture, software and application groups to tight collaboration centered on the co-simulation of hardware and applications**
- **Opportunity to influence future architectures**
  - Cores/node, threads/core, scheduling width/thread
  - Logic in memory subsystem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

How do we consider multiple applications?

Sandia is influencing Exascale discussions at the national level

- **Helped lead development of DOE's Exascale technology roadmap**
  - Basis for Strayer/Meisner presentation to Brinkman, Koonin, D'Agostino, Chu
- **Influenced key elements of strategy, especially co-design**
  - 2 plenary presentations on co-design
  - Widely acknowledged by community as critical to success
- **Worked with Rick Stulen and SNL/CA CRF to establish combustion as a key DOE Exascale application**
- **One journal article and one invited paper in 2010 on exacale computing**
- **Sandia is leading a DARPA/UHPC proposal team**

Sandia was heavily involved in Exascale cross-cutting workshop (1/10 in DC)



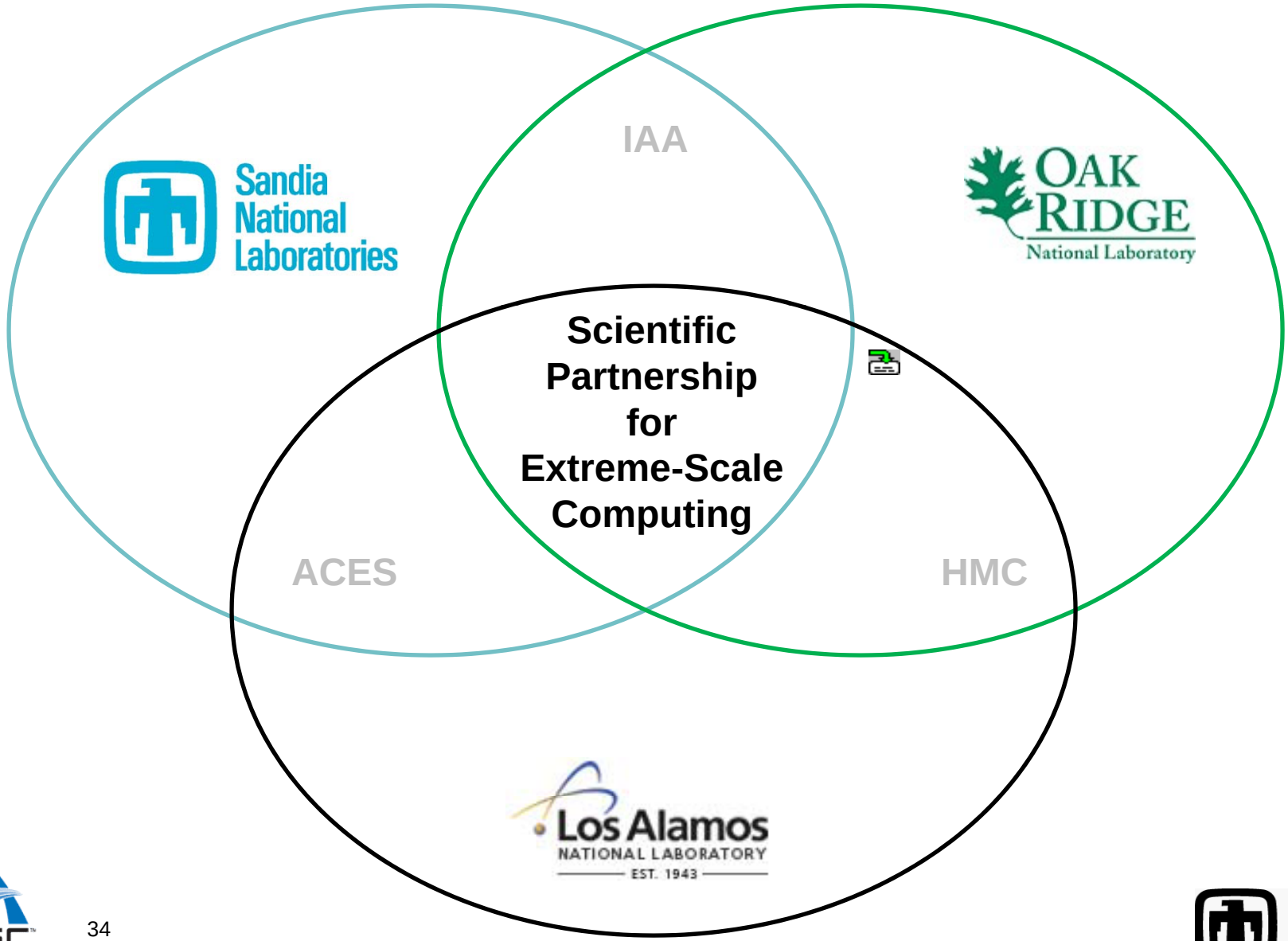
Sandia National Laboratories

DOE asked Sandia to establish strategic Micron collaboration to address memory wall and energy/power challenge

Activity	Results
2-way NDA (Nov'07)	Foundation for discussions of Micron's proprietary 3DI stacked memory design
Advanced Memory for DOE Architectures	PIM LDRD Effort, ASC/CSSE L2 Milestone: Evaluate Advanced Memory Subsystems – Due 9/10
IAA Activities	Dean Klein is a member of the IAA Advisory Board IAA Workshop on Memory Opportunities for HPC (1/08) IAA Workshop on HPC Architectural Simulation (9/09)
Alignment with other agencies	Alignment of ASC/CSSE, DoD/ACS, & IAA support to Integrate UMd's Memory Simulator (DRAMsim with SST)
Proposal partnerships	IAA Memory Proposal, DOE/ASCR FOA, DARPA/UHPC
CRADA	Micron-Sandia collaboration to analyze advanced concepts for error correction in advanced memory designs (5/10 - pending)

- Technical exchanges from 7/06 – present
  - Approximately two dozen face-to-face technical meetings
  - Catalyst for collaboration with other DOE/NNSA labs, DoD and universities

A New Alliance for Exascale



Sandia National Laboratories

LANL, ORNL and Sandia have formed a strategic alliance to reach Exascale

- **Mission:**
  - Development of extreme-scale applications and technologies for DOE mission needs
- **Goals:**
  - Advance DOE's mission through extreme-scale computing
  - Provide national leadership for Exascale computing and beyond
- **Charter:**
  - Exclusively collaborate on platform development/management proposals in response to DOE ASCR/ASC Exascale calls for proposals
  - Coordinate efforts on Office of Science and NNSA platforms as appropriate
  - Develop a joint technology roadmap and coordinated strategy for reaching Exascale applications and computing in this decade
  - Develop long-term collaborations with vendors to advance partnership strategy of sustainable extreme-scale computing
  - Perform and coordinate needed architectures, computer science and mathematics research
  - Initiate and coordinate Extreme-scale applications efforts especially in materials, catalysis, climate and combustion

35