U.S. DEPARTMENT OF ENERGY

Exascale Applications and Technologies for DOE Mission Needs

Mark Seager

Lawrence Livermore National Laboratory

28 September 2010

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344 LLNL-PRES-459917

U.S. DEPARTMENT OF ENERGY

A decadal DOE plan for providing exascale applications and technologies for DOE mission needs

Rick Stevens and Andy White, co-chairs

Pete Beckman, Ray Bair-ANL; Jim Hack, Jeff Nichols, Al Geist-ORNL; Horst Simon, Kathy Yelick, John Shalf-LBNL; Steve Ashby, Moe Khaleel-PNNL; Michel McCoy, Mark Seager, Brent Gorda-LLNL; John Morrison, Cheryl Wampler-LANL; James Peery, Sudip Dosanjh, Jim Ang-SNL; Jim Davenport, Tom Schlagel, BNL; Fred Johnson, Paul Messina, ex officio

U.S. DEPARTMENT OF ENERGY

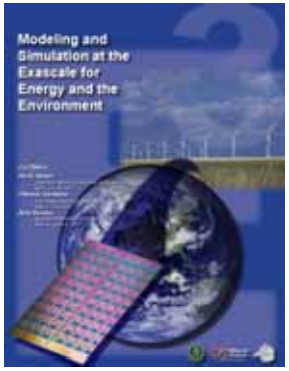
Outline of talk

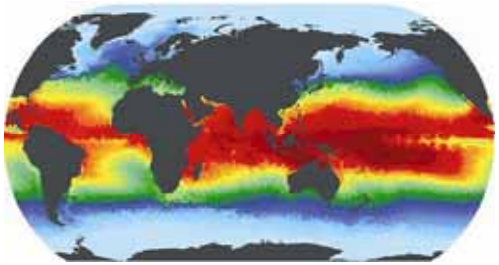
- **Exascale Initiative Background**
- **Technology needs**
- **Co-design**

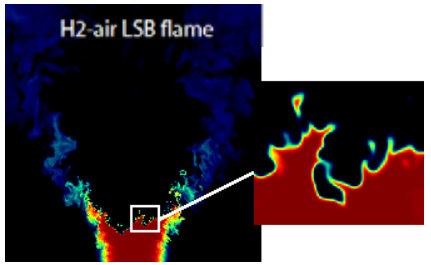
Innovation on power and cooling essential to enable exascale and larger IT market growth ➔ co-design of systems, applications and data centers

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

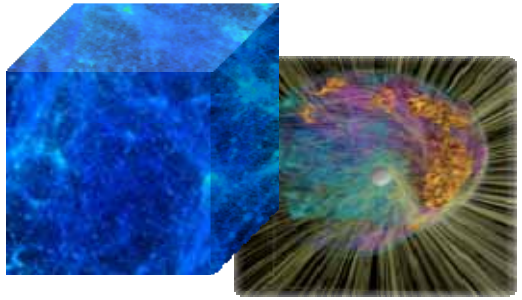
- **Town Hall Meetings April-June 2007**
- **Scientific Grand Challenges Workshops Nov, 2008 – Oct, 2009**
  - **Climate Science (11/08),**
  - **High Energy Physics (12/08),**
  - **Nuclear Physics (1/09),**
  - **Fusion Energy (3/09),**
  - **Nuclear Energy (5/09),**
  - **Biology (8/09),**
  - **Material 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**







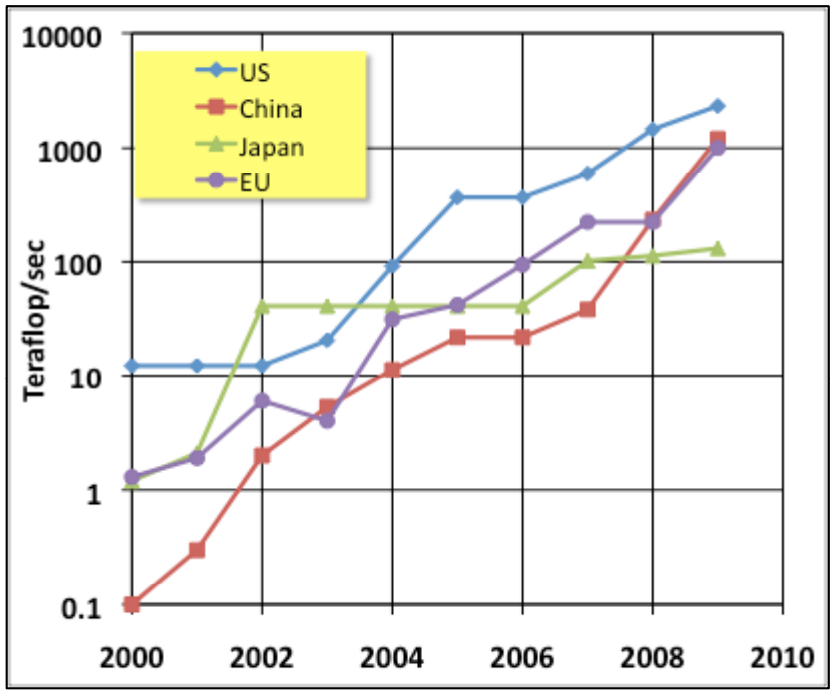
MISSION IMPERATIVES



FUNDAMENTAL SCIENCE

Computational science, exascale computing & leadership in science and technology

- **The future will require certification of complex engineered systems and analysis of climate mitigation alternatives with quantified levels of uncertainty**
  - **New fuels and reactors**
  - **Stewardship without nuclear tests**
  - **Carbon sequestration alternatives**
  - **Regional climate impacts**
- **Broader application of exascale computing can provide tremendous advantages for fundamental science and industrial competitiveness**
  - **Renewable energy and energy storage**
  - **Prediction and control of materials in extreme environments**
  - **Understanding dark energy and dark matter**
  - **Clean and efficient combustion in advanced engines**



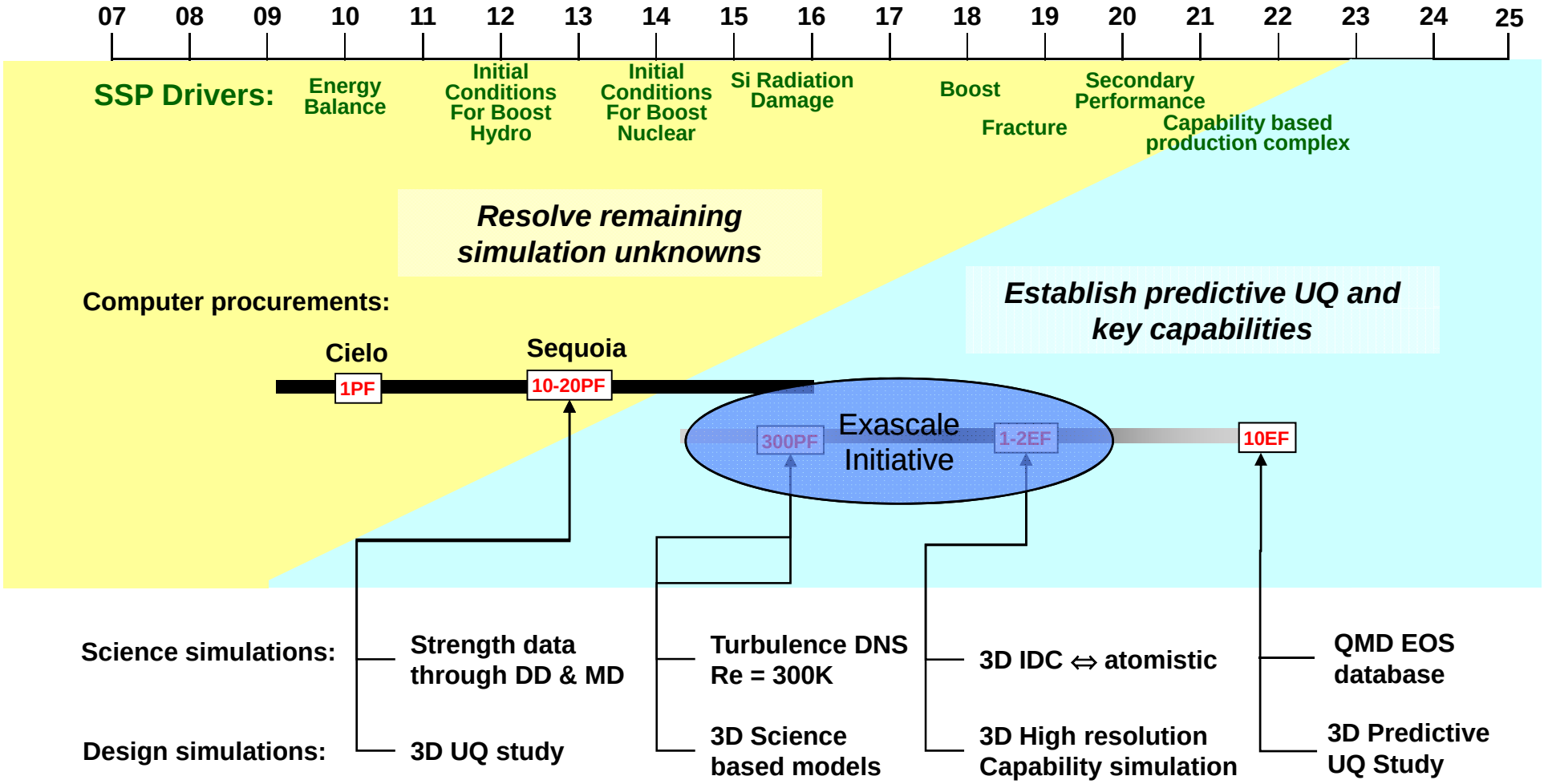
International Competition in HPC

Chart shows most capable system for each year in TOP500

"The United States led the world's economies in the 20th century because we led the world in innovation. Today, the competition is keener; the challenge is tougher; and that is why innovation is more important than ever. It is the key to good, new jobs for the 21st century."
President Barack Obama, August 5, 2009

U.S. DEPARTMENT OF ENERGY

A sequence of powerful systems must support the PCF & exascale is required to support predictive 3D UQ



Predictive capability is required to perform stockpile assessments as we approach a smaller stockpile, a leaner production complex and to support the NPR LEP modalities

A successful exascale initiative will require careful coordination across five efforts

A model much like that managed by ASC over the past decade is essential – indeed, vendors will be depending critically on deliverables from other components!

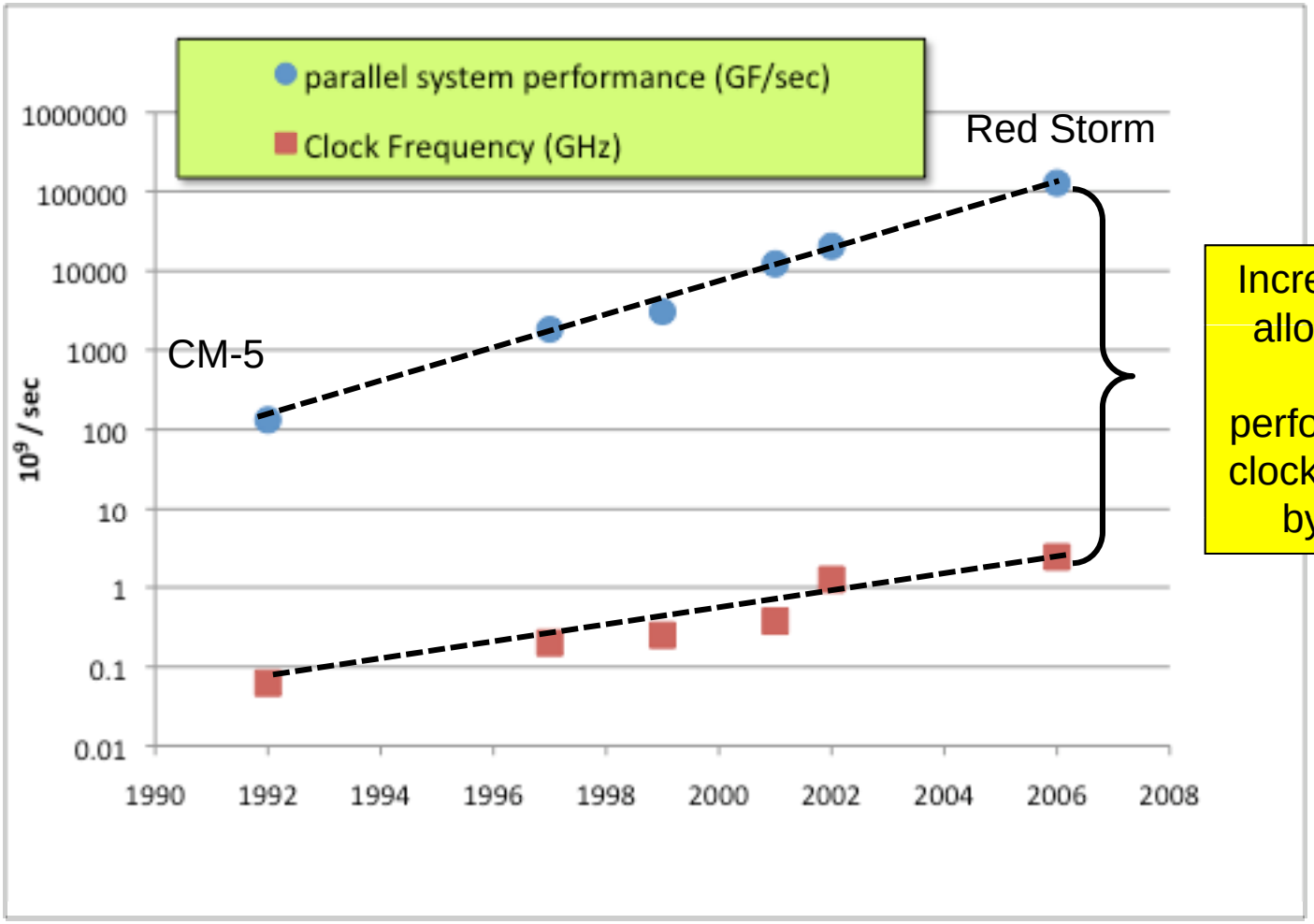
Potential System Architecture Targets

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)	

U.S. DEPARTMENT OF ENERGY

TECHNOLOGY NEEDS

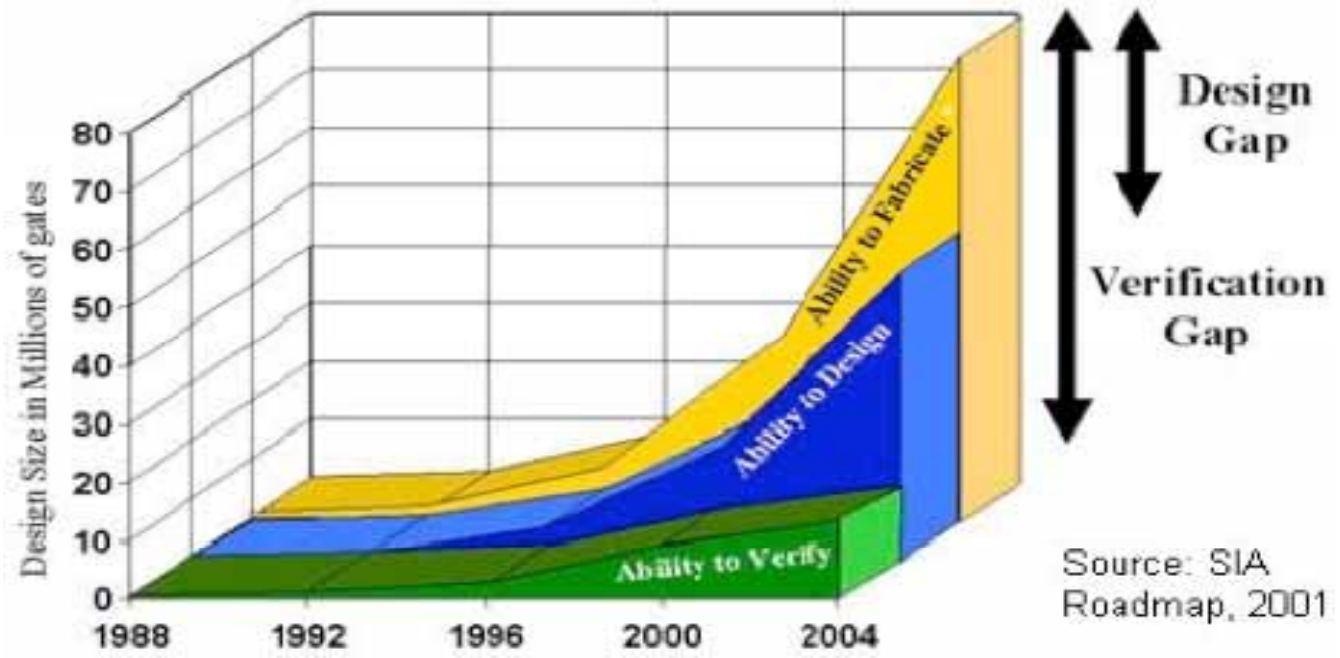
Concurrency is one key ingredient in getting to exaflop/sec



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.*

Swim lanes affect the number of threads that the system needs to support.

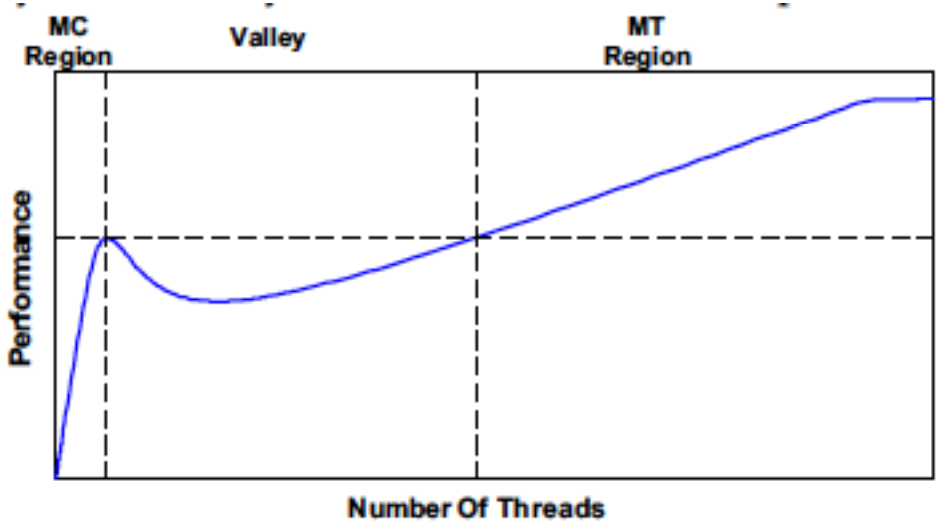
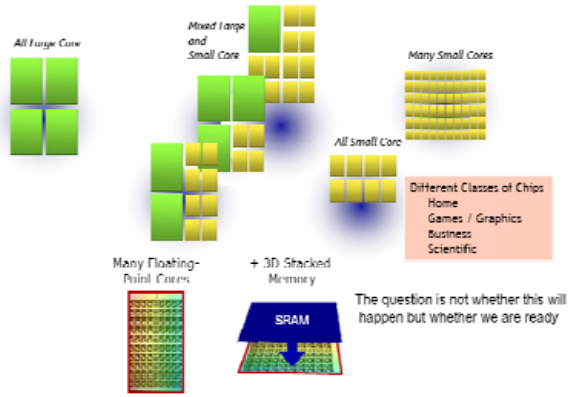
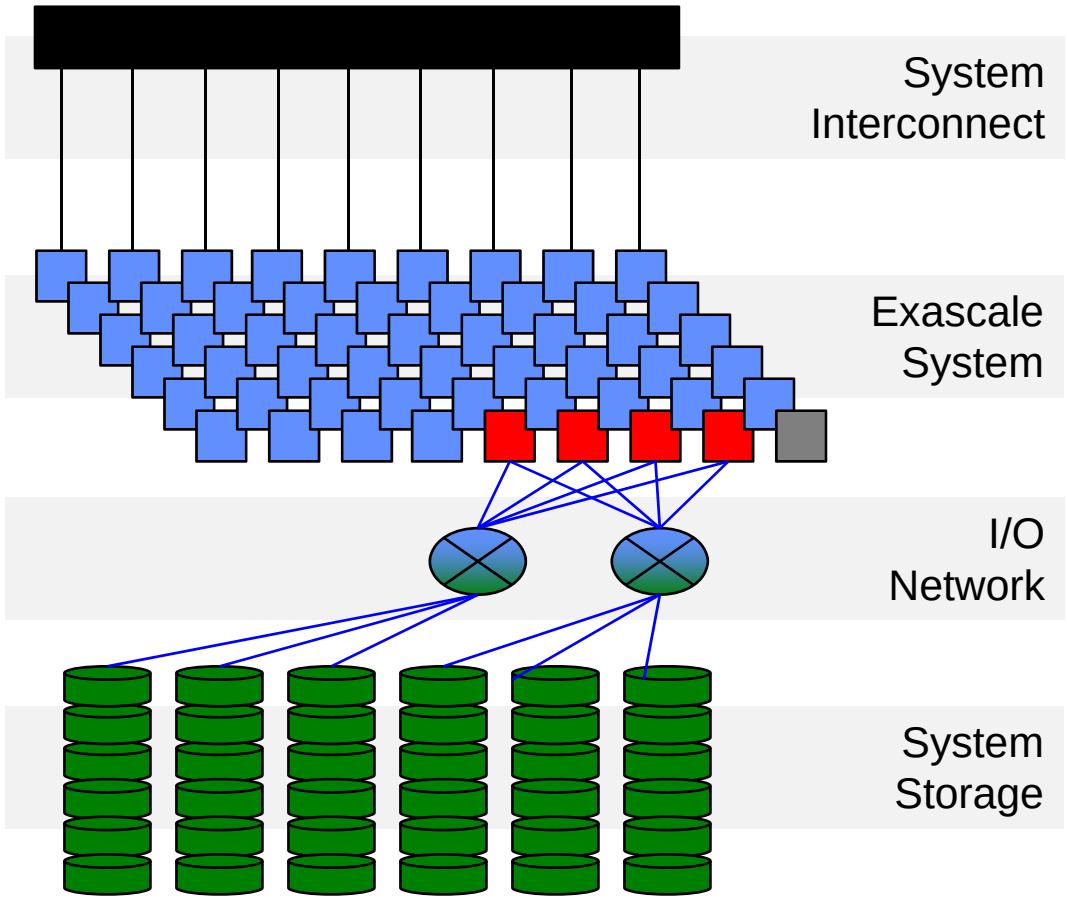


Fig. 1. Performance of a unified many-core (MC) many-thread (MT) machine exhibits three performance regions, depending on the number of threads in the workload.



There are currently two basic design points for achieving high performance in technical applications. In the future it is expected that these design points will become more Integrated.

The high level system design may be similar to petascale systems

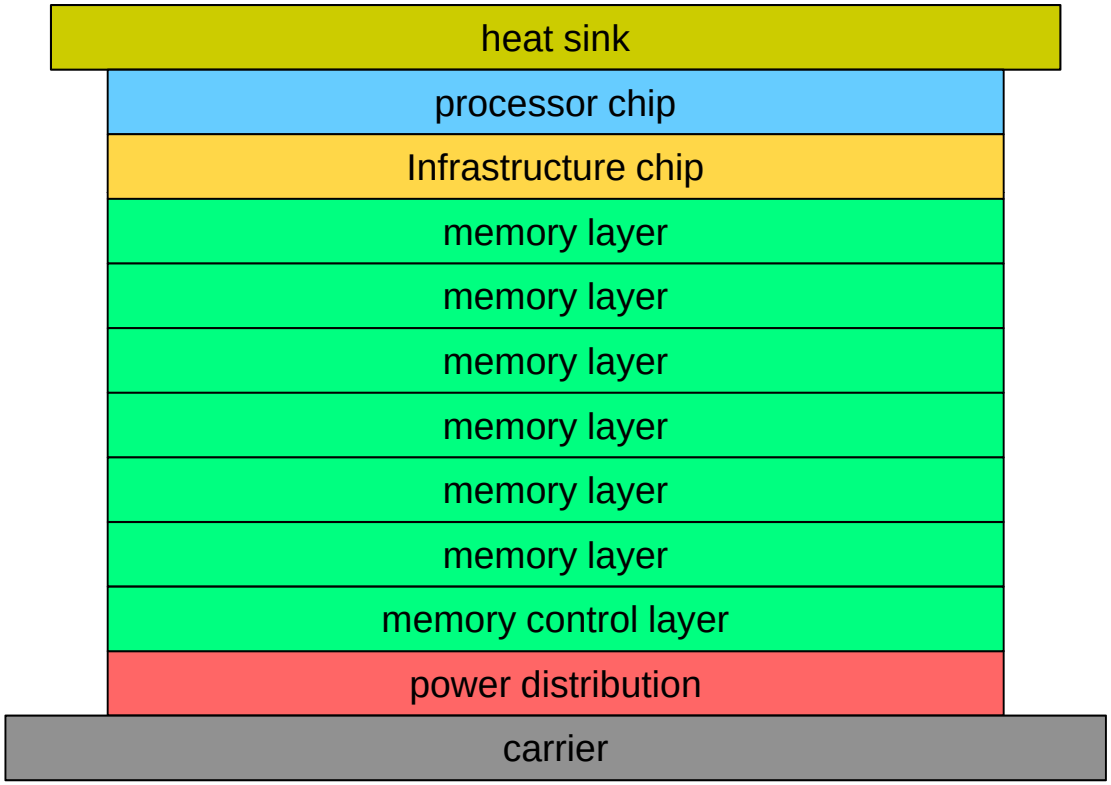


- New interconnect topologies
- Optical interconnect

- 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

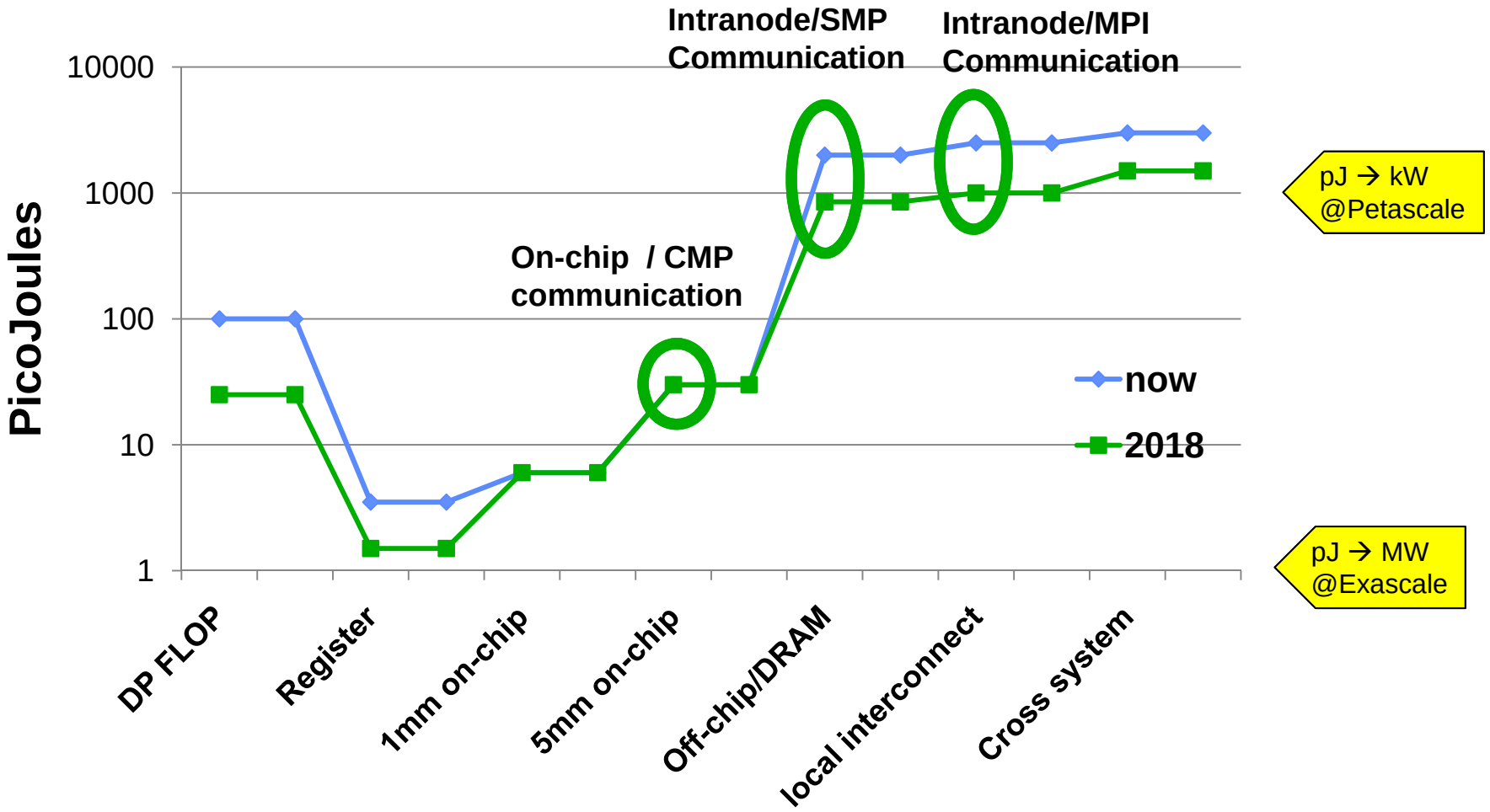
- 3d stacked memory

- 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

Example of architectural reduction of processor power consumption.

TABLE I

PROCESSOR CONFIGURATIONS AND DETAILS

Ensemble Processor			
Technology	TSMC CL013G (V_{DD}=1.2V)		
Clock Frequency	200 MHz		
Average Power	28 mW		
Multipliers	16-bit + 40-bit acc.	16.5 pJ/op	
IRFs	64 128-bit registers	16 pJ/read	18 pJ/write
XRFs	32 32-bit registers	14 pJ/read	8.7 pJ/write
ORFs	8 32-bit registers	1.3 pJ/read	1.8 pJ/write
ARF	8 16-bit registers	1.1 pJ/read	1.6 pJ/write
Ensemble Memory	8KB	33 pJ/read	29 pJ/write
RISC Processor			
Technology	TSMC CL013G (V_{DD}=1.2V)		
Clock Frequency	200 MHz		
Average Power	72 mW		
Multiplier	16-bit + 40-bit acc.	16.5 pJ/op	
Register File	40 32-bit registers	17 pJ/read	22 pJ/write
Instruction Cache	8KB (2-way)	107 pJ/read	121 pJ/write
Data Cache	8KB (2-way)	131 pJ/read	121 pJ/write

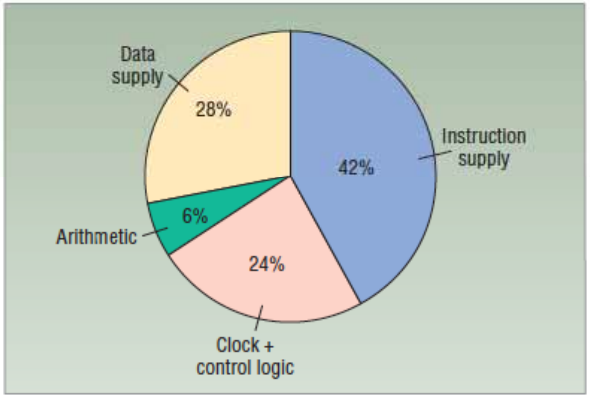


Figure 1. Embedded processor efficiency. Supplying data and instructions consumes 70 percent of the processor's energy; performing arithmetic consumes only 6 percent.

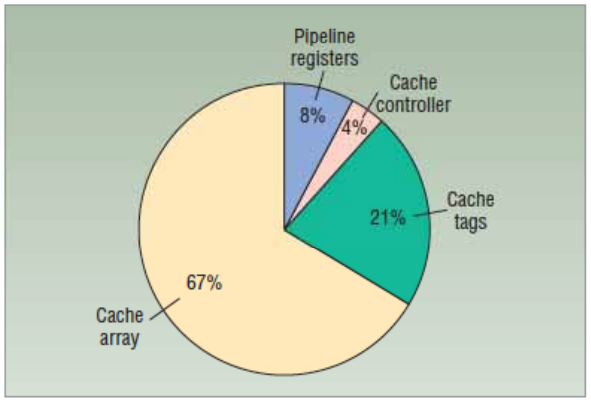


Figure 2. Instruction-supply energy breakdown. The 8-Kbyte instruction cache consumes the bulk of the energy, while fetching each instruction requires accessing both directions of the two-way set-associative cache and reading two tags.

An Energy-Efficient Processor Architecture for Embedded Systems, IEEE Computer Arch Letters, 2007

Reducing power is fundamentally about architectural choices & process technology

- **Memory (2x-5x)**
  - New memory interfaces (optimized memory control and xfer)
  - Extend DRAM with 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

Tera➔Peta-Scale trends are not sustainable

System	date	peak	nodes	cores	power	Facilities Impact
BluePacific ID	1996	0.136	512	512	0.125	B113 Air Handler
BluePacfic TR	1997	0.75	512	2048	0.25	
BluePacific SST	1998	1.3	1452	5808	0.433	B113 Doubling
White	2000	12.3	512	8192	1.0	B453 doubling
BlueGene/L	2004	367	65536	131072	1.8	
Purple	2005	100	1536	12288	4.8	New Building
Dawn	2008	1000	73728	294912	2.3	
Sequoia	2011	20000	98304	1572864	8.0	B453 doubling

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

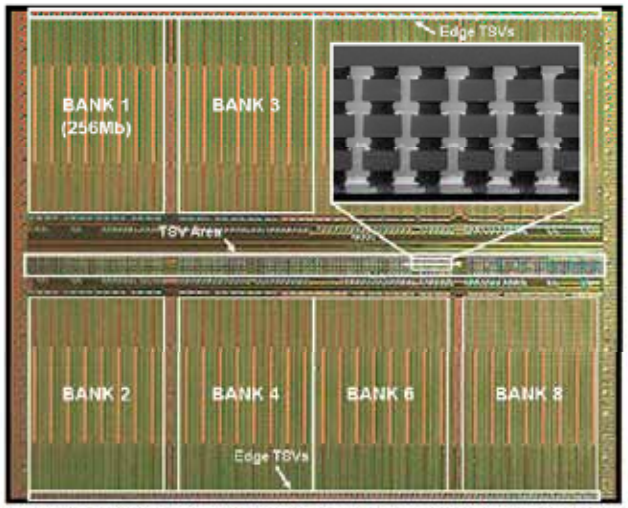


Figure 7.2.7: Die micrograph of the fabricated chip and cross-sectional view of TSVs. The chip size is 10.9×9.0mm².

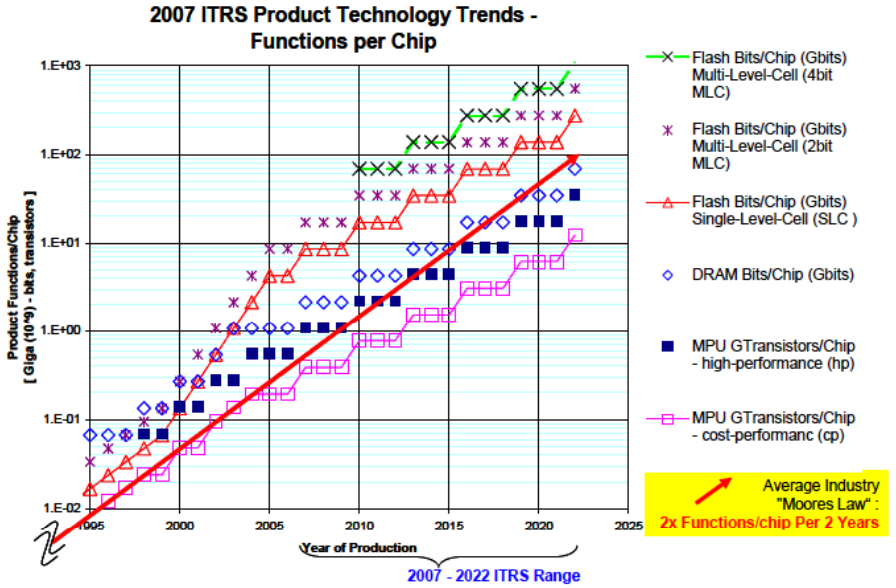
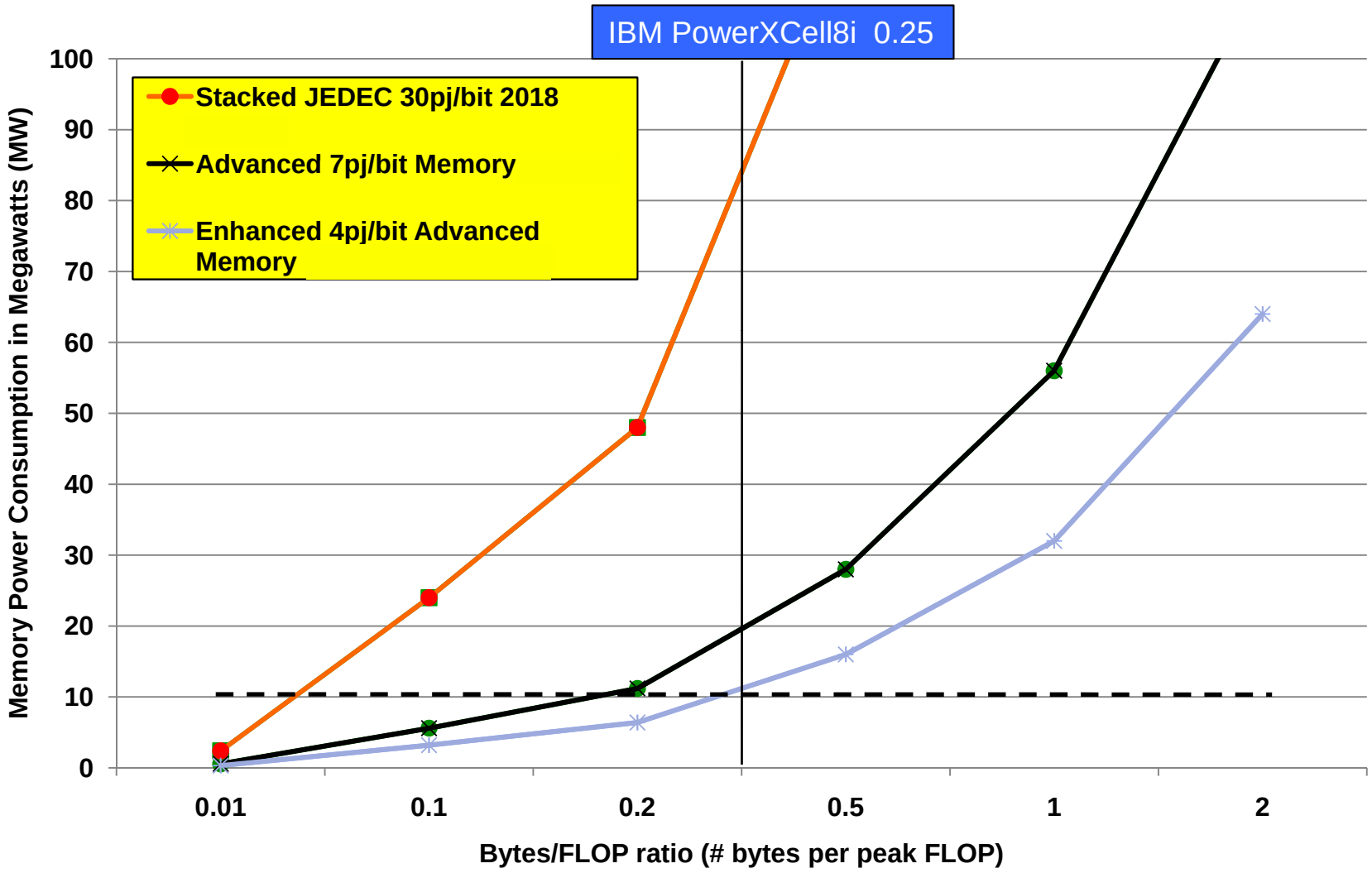


Figure ORTC2 ITRS Product Function Size Trends: MPU Logic Gate Size (4-transistor); Memory Cell Size [SRAM (6-transistor); Flash (SLC and MLC), and DRAM (transistor + capacitor)]--Updated

U.S. DEPARTMENT OF ENERGY

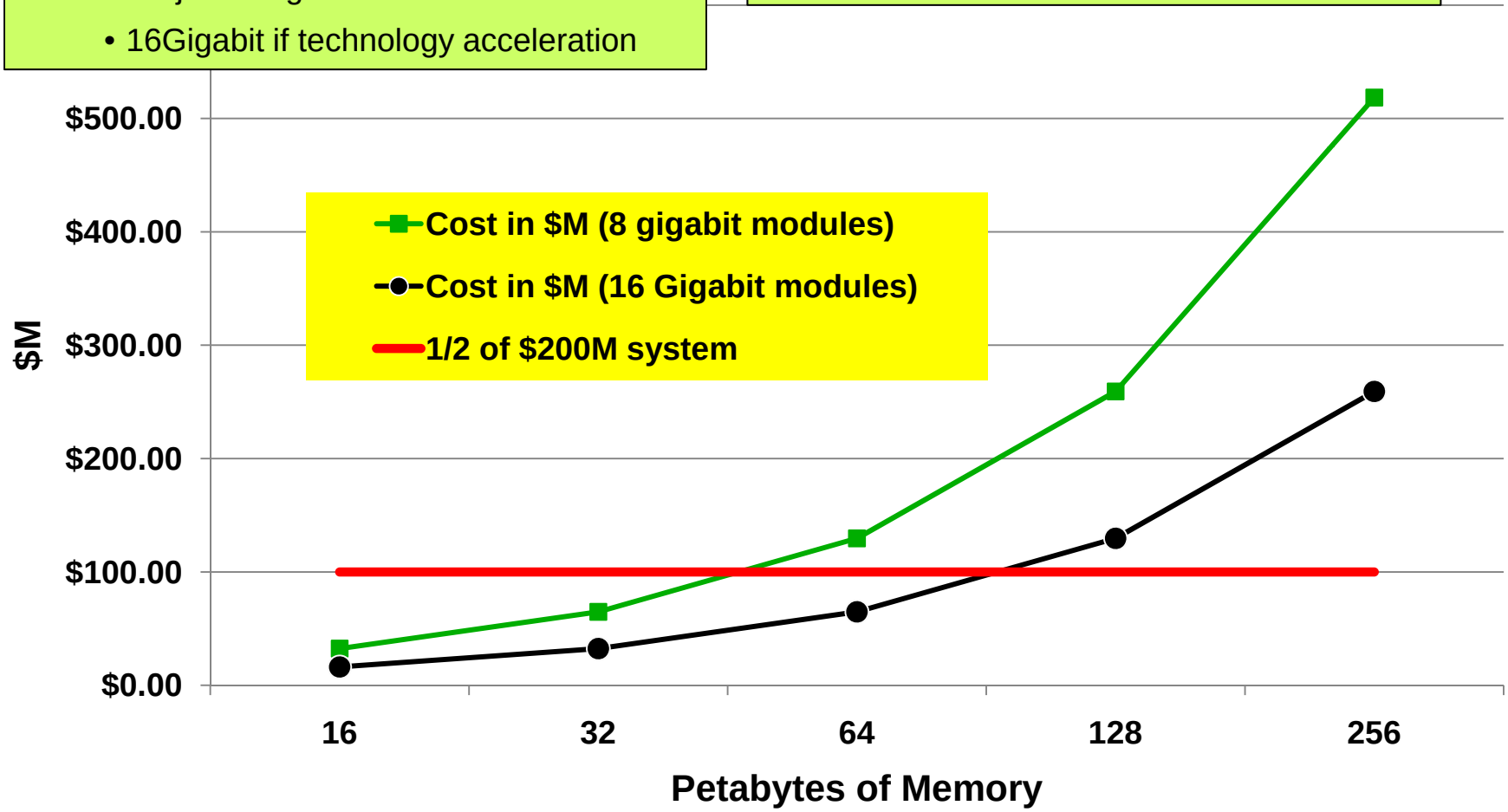
Investments in memory technology mitigate risk of narrowed application scope.



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

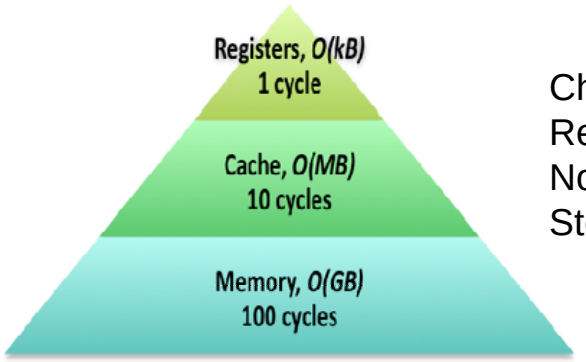


Need solutions for decreased reliability and a new model for resiliency

- **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

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.

Resilience gap analysis

The following analysis (with items in priority order) take into account the need for near-term mitigations and the longer-term R&D needed for resilience at the Exascale.

1. Existing fault tolerance techniques (global checkpoint/global restart) and will be unpractical at Exascale. Local checkpoint techniques for saving and restoring state need to be developed into practical solutions before 2015
2. There is no standard fault model, nor standard fault test suite or metrics to stress resilience solutions and compare them fairly.
3. Errors, fault root causes, propagation, and rate of silent errors are not well understood

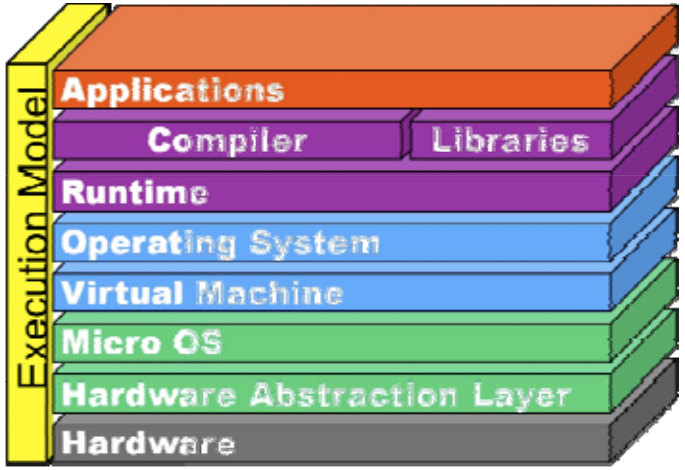
Primary risks

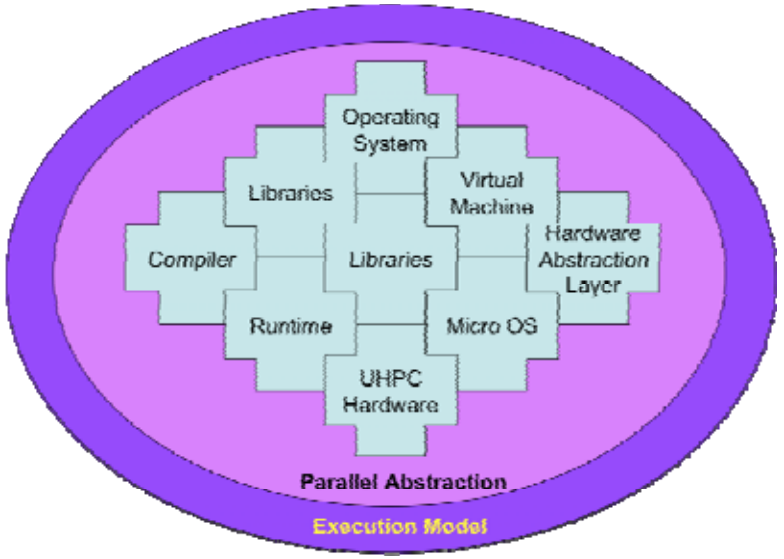
2015 system (moderate)
The amount of data needing to be check pointed and the expected rate of faults for petascale and larger systems are already exposing the inadequacies traditional checkpoint/restart techniques.
Mitigation: Local checkpoint schemes, either on board or in network, required to for resilience on 200 PF system

2018 System (high)
If the relevant components of the HW/SW stack are not fault tolerant, then even relatively short-lived applications are unlikely to finish or worse, may terminate with an incorrect result.
Mitigation: community R&D effort as described above, co-designed with apps and vendors

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

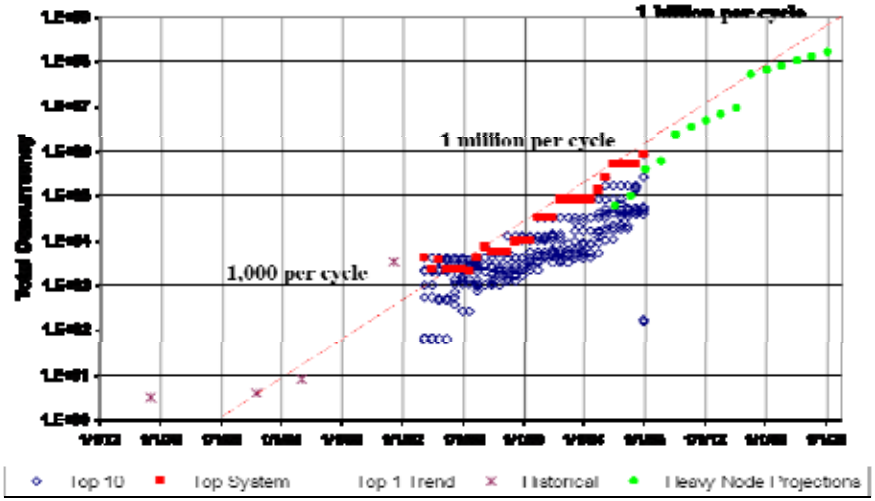
Factors Leading to Gap

It is not only the massive increase in concurrency, but also the change of architecture

- **OS**
  - **Current OS designs focus on homogeneous cores, memory structures, and tasks**
  - **Designs to manage 256 "full" cores or efficiently coordinate thousands of stream processors for HPC applications do not exist**
  - **HW makers are moving to on-chip page-mapped memory, but no OS features have been developed to leverage this for HPC applications**
- **I/O**
  - **Current designs are primarily file based, and cannot efficiently optimize HPC workloads for aggregation, ordering, and patterns**
  - **Chip makers are putting NV RAM on or close to die, but file-based I/O paradigms are not suited to leverage this development**
  - **Currently, I/O is "far", through many hardware layers (torus, I/O forwarding, Infiniband, RAID controller, SCSI, etc). I/O balance quickly falling behind – new integrated design approach required**
- **Messaging & Run-time Systems**
  - **HW put/get message queues for interconnect must rapidly evolve to support massive concurrency – however SW community has not explored how to manage millions of msg endpoints, dynamic mapping of memory buffers, or fault resilience at that scale**

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
- **Technology Investments**



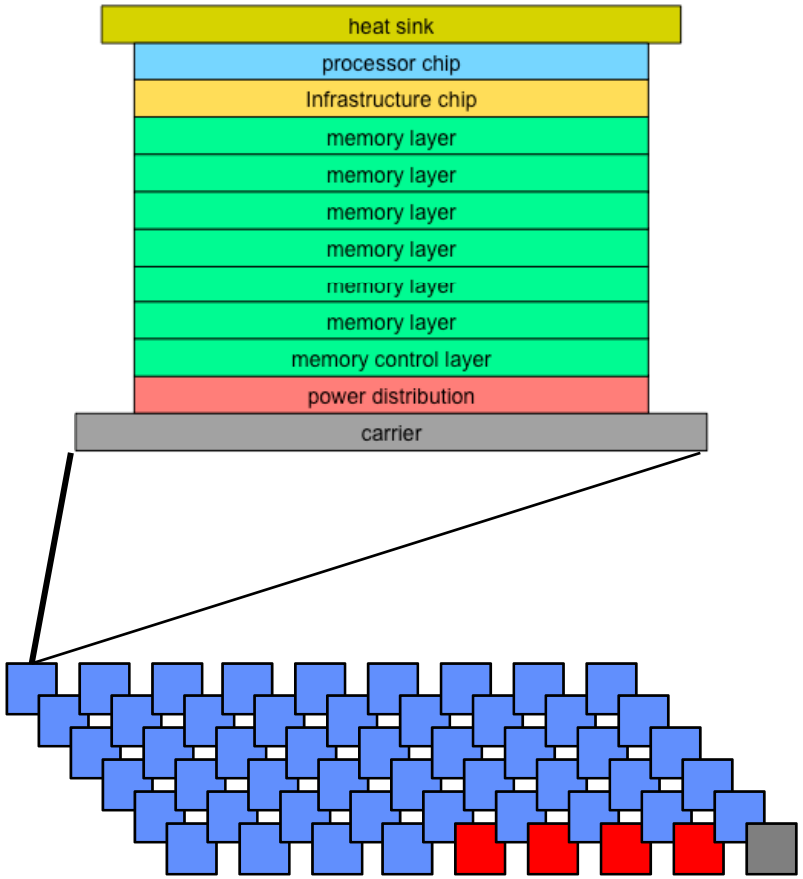
How much parallelism must be handled by the program?

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

  - 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

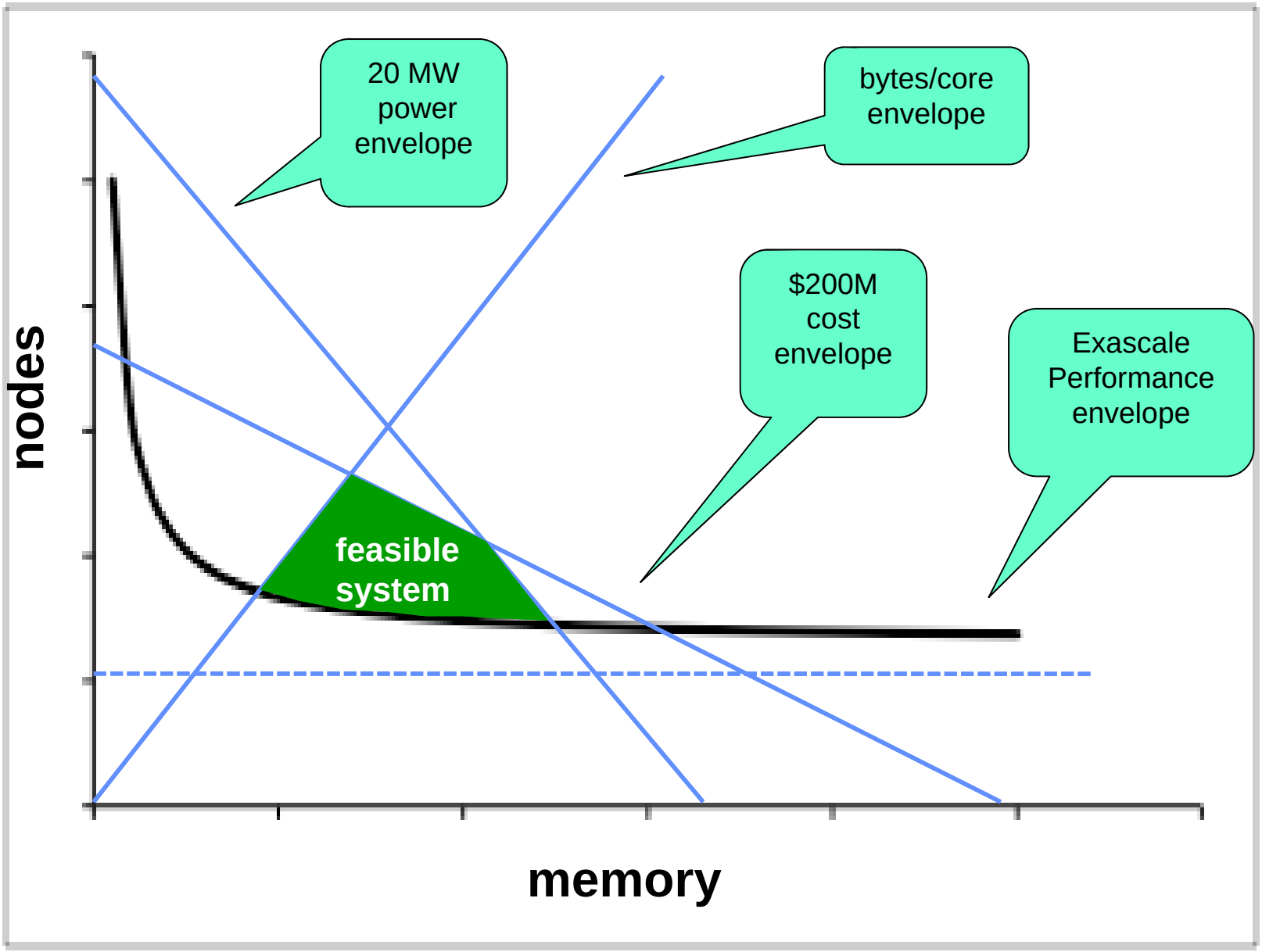
- **Hierarchical approach (intra-node + inter-node)**
  - **Part I: Inter-node model for communicating between nodes**
    - **MPI scaling to millions of nodes: Importance high; risk low**
    - **One-sided communication scaling: Importance medium; risk low**
  - **Part II: Intra-node model for on-chip 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

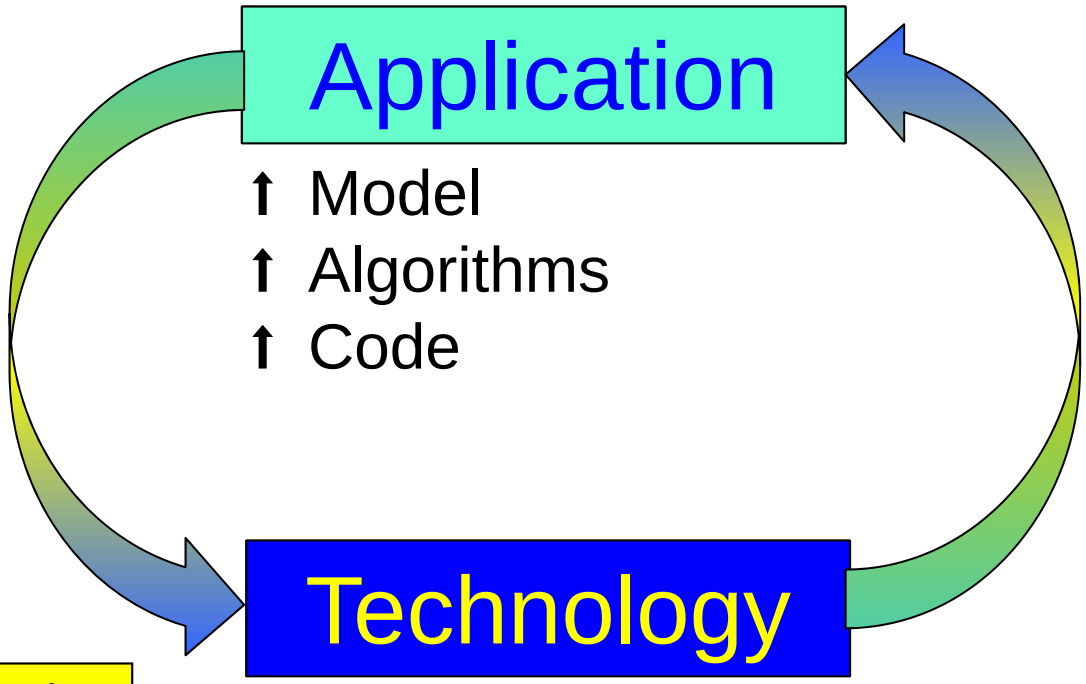
CO-DESIGN

The trade space for exascale is very complex.



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