LA-UR- 11-02135

Approved for public release; distribution is unlimited.

Title: Exascale: Applications, Architecture and Co-Design

Author(s): Andrew B. White, Jr
ADTSC

Intended for: Predictive Science Panel
April 12, 2011
Los Alamos National Laboratory

Los Alamos
NATIONAL LABORATORY
EST. 1943

Exascale Applications, Architecture and Co-Design

Abstract. Computing technology will change in fundamental ways over the next decade. Co-design of applications, software and hardware is essential to managing this complex change. This talk describes the DOE's exascale initiative to address these challenges.

Exascale Applications, Architectures ~~y~~ and Co-design

Predictive Science Panel

April 12, 2011

Andy White

Technology challenges, and opportunities, will be significant over the next decade.

- ASC must meet the on-going needs of the weapons program
  - Provide increasingly better predictive physics and engineering capabilities
  - Provide increasingly more capable computational resources to support predictive science
  - Maintain the necessary core capabilities of the weapons program staff, i.e. right sizing
- ASC's ability to meet these needs will be severely impacted by the transformation of basic computational technology over the next decade
  - Performance of existing codes will stall, at best
  - Both capability and capacity resources will be difficult to use
- The exascale initiative is a DOE plan to meet these challenges and take advantage of this opportunity

“You can run but you can’t hide.”

Joseph Louis Barrow

- **Power** will become the first class constraint on system performance and effectiveness at exa-scale, at peta-scale and at desktop-scale.
- **Tomorrow’s on-chip multi-processor** will have an 100 – 1000x increase in parallelism; architecture is critical to meet power, performance, price, productivity & predictive goals.
- **Reliability and resiliency are very difficult at this scale** and require new error handling model for applications and better understanding of effects and management of errors.
- **Memory** is not scaling with performance and memory hierarchies will be higher and deeper.
- **Tomorrow’s programming model** will be different on tomorrow’s chip multi-processors, whether exascale or not. Early investment is critical to provide applications effective access to 2015 system.
- **Operating and run-time systems will be redesigned** to effectively management massive on-chip parallelism, system resiliency and power.
- ***Co-design*** *requires a set of hierarchical set of performance models, simulators and emulators as well as agile compact applications to mediate interactions among applications, software and architecture communities.*

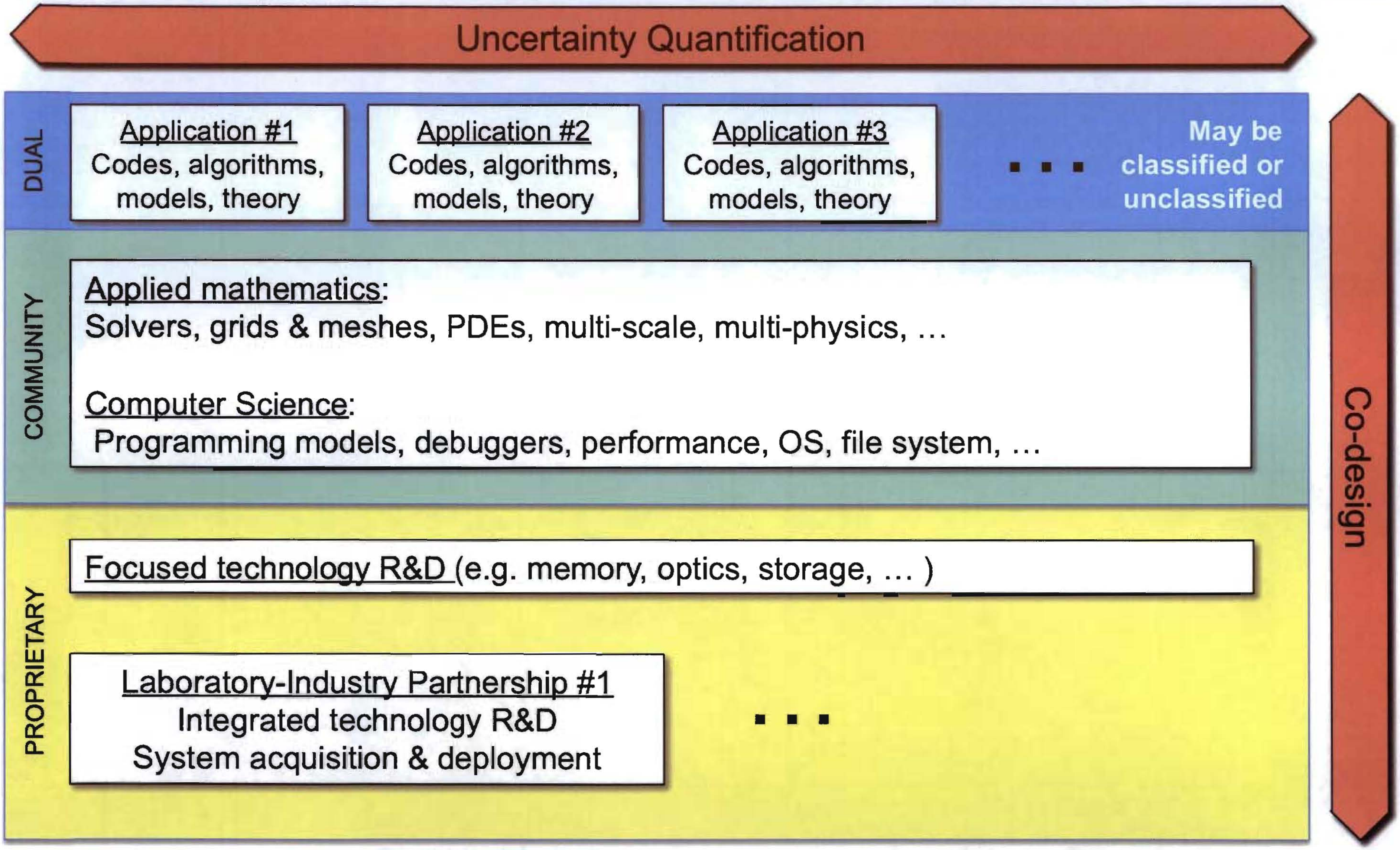
"The Future of Computing Performance: Game Over or Next Level?" NRC, 2011

> *"The U.S. Computing Industry has been adept at taking advantage of increases in computing performance, allowing the United States to be a moving and therefore elusive target – innovating and improvising faster than anyone else."*

- Invest in research in and development of algorithms that can exploit parallel processing.
- Invest in research in and development of programming methods that will enable efficient use of parallel systems …
- Focus long-term efforts on rethinking of the canonical computing "stack" …
- Invest in research on and development of parallel architectures driven by applications, …
- Invest in research and development to make computer systems more power efficient at all levels of the system …

> **"There is no known alternative to parallel systems for sustaining growth in computing performance; however, no compelling programming paradigms for general parallel systems have yet emerged."**

Collaboration and co-design are the key ingredients of the exascale initiative.



Uncertainty Quantification

DUAL

Application #1
Codes, algorithms, models, theory

Application #2
Codes, algorithms, models, theory

Application #3
Codes, algorithms, models, theory

▪ ▪ ▪

May be classified or unclassified

COMMUNITY

Applied mathematics:
Solvers, grids & meshes, PDEs, multi-scale, multi-physics, …

Computer Science:
Programming models, debuggers, performance, OS, file system, …

PROPRIETARY

Focused technology R&D (e.g. memory, optics, storage, …)

Laboratory-Industry Partnership #1
Integrated technology R&D
System acquisition & deployment

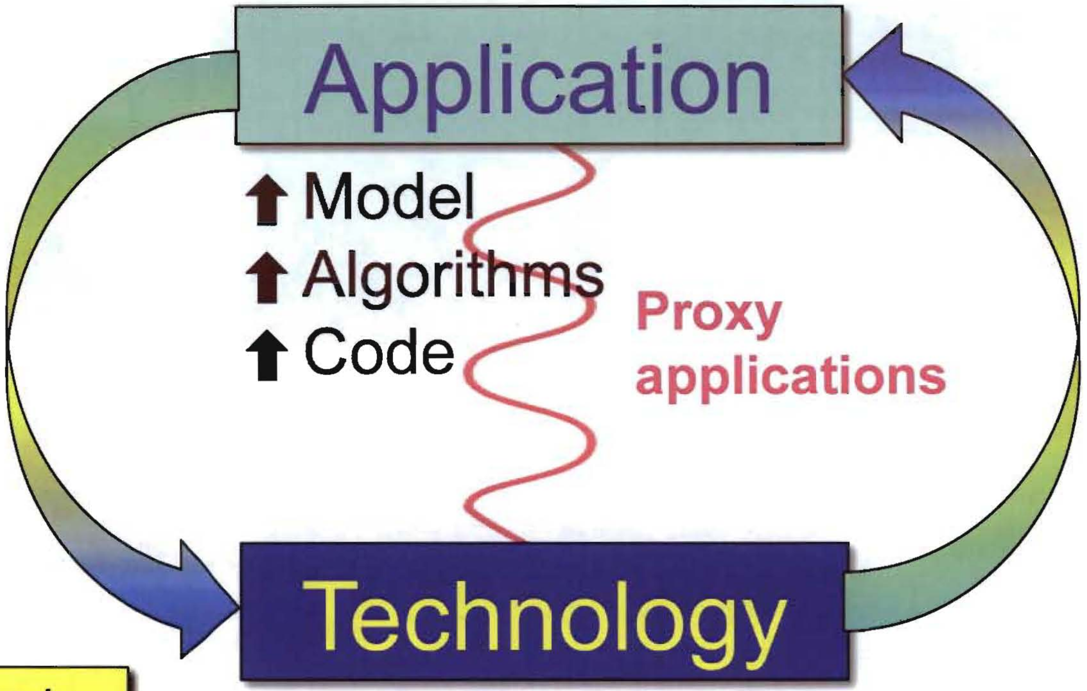
▪ ▪ ▪

Co-design

Co-design is essential to manage complexity and optimize the outcome.

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

Key issues
Power?
Performance?
Price?
Parallelism?
Productivity?



Application

↑ Model
↑ Algorithms
↑ Code

Proxy applications

Technology

⊕ programming model
⊕ operating system
⊕ architecture

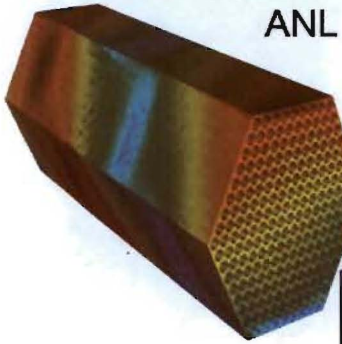
Now, we must expand the co-design space to find better solutions:

- *new applications & algorithms,*
- *better technology and performance.*

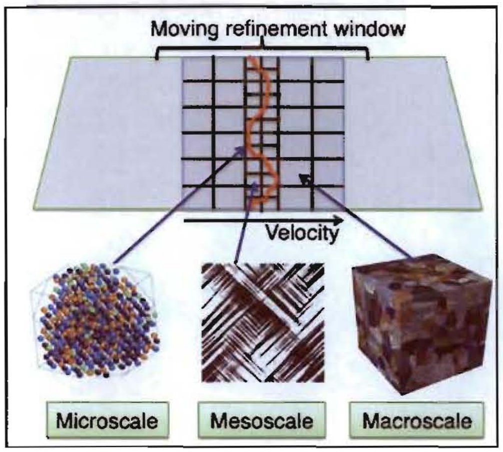
Technology driven:
Fit your application to this technology.
Sub-optimal.

Seven ASCR application-driven co-design centers are under consideration.

Advanced Reactors
ANL

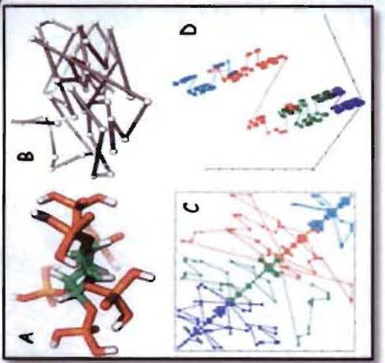


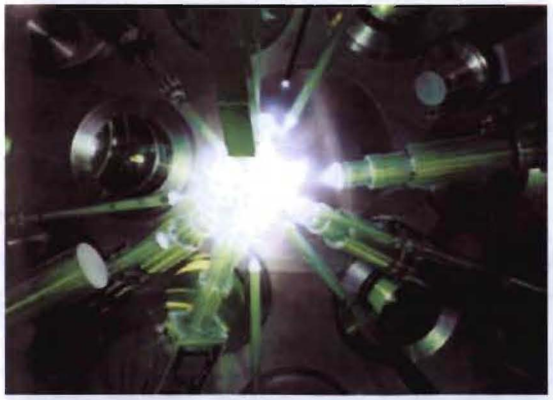
Earth systems
LANL & ORNL



Materials in Extreme Environments
LANL

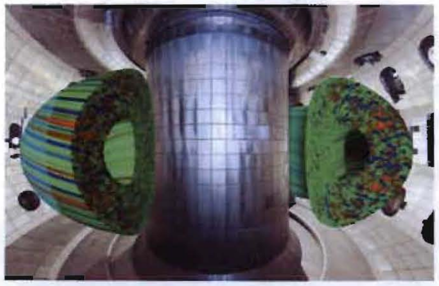
Chemistry
ORNL

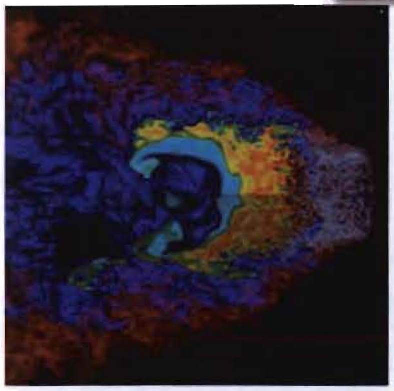




High-energy Density Physics
ANL

Fusion
LBNL

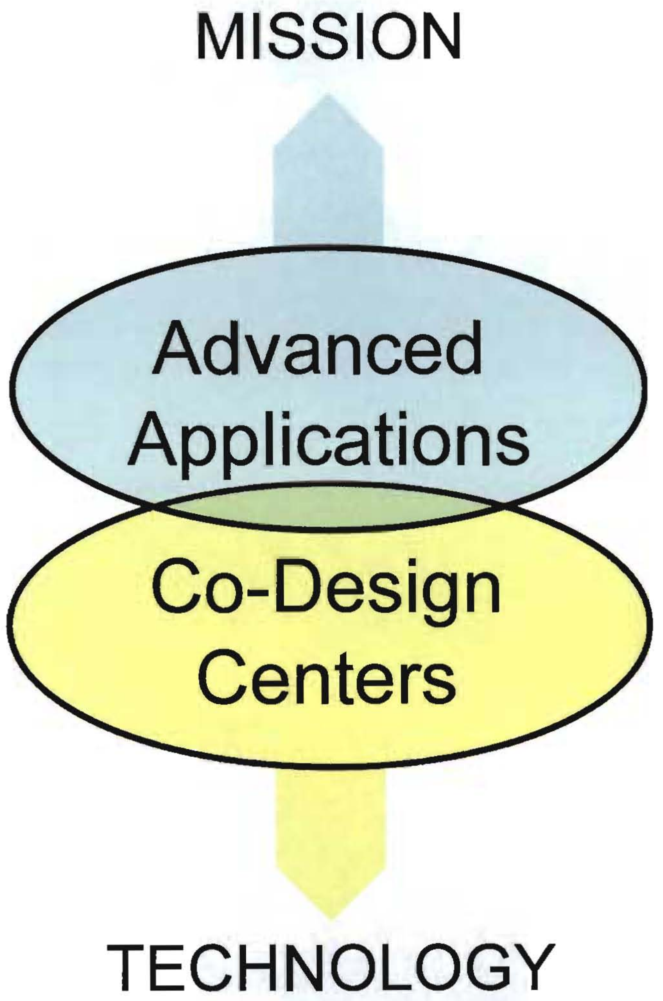




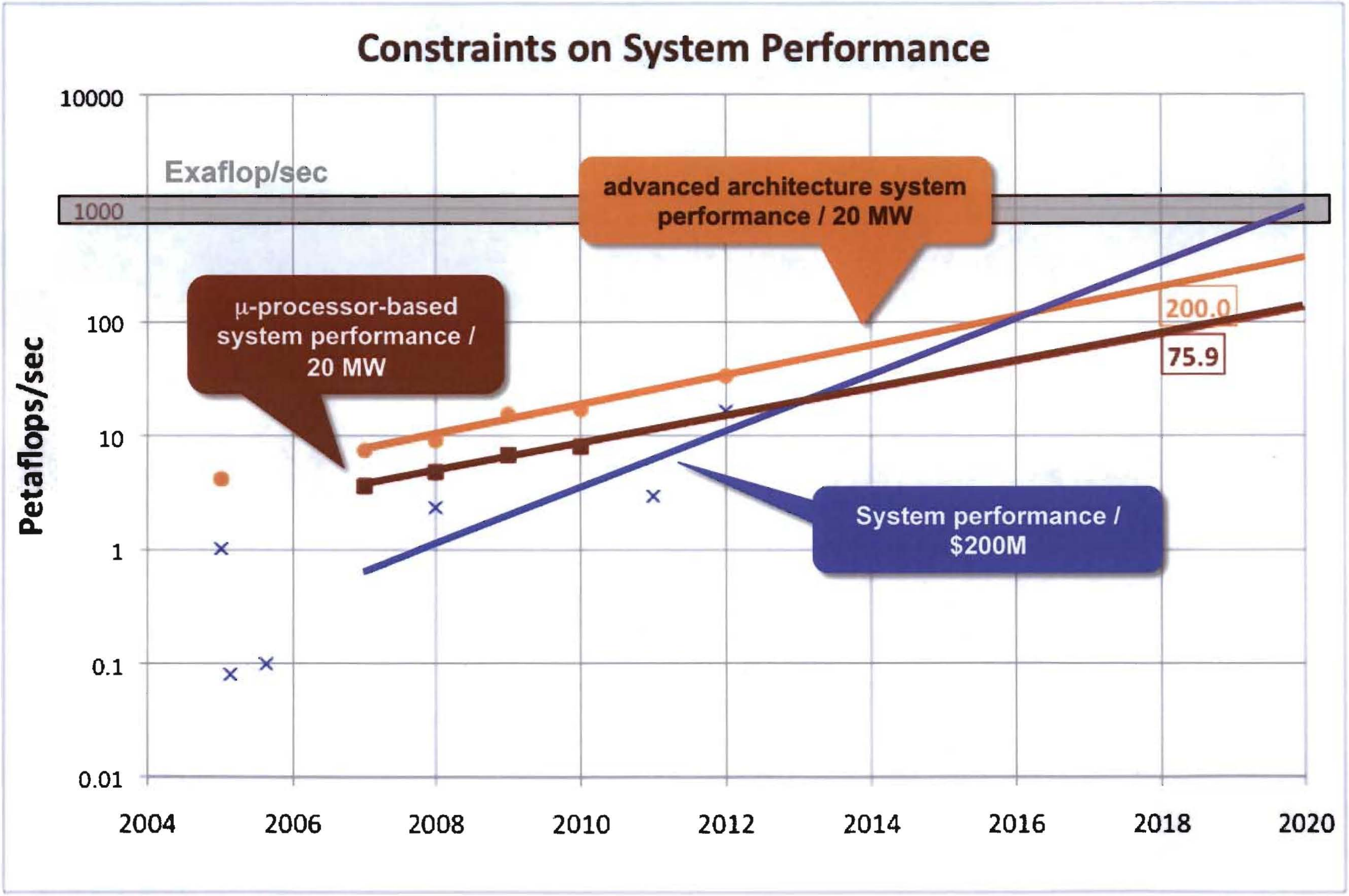
Combustion
SNL

ASC is considering additional co-design activities and advanced applications.

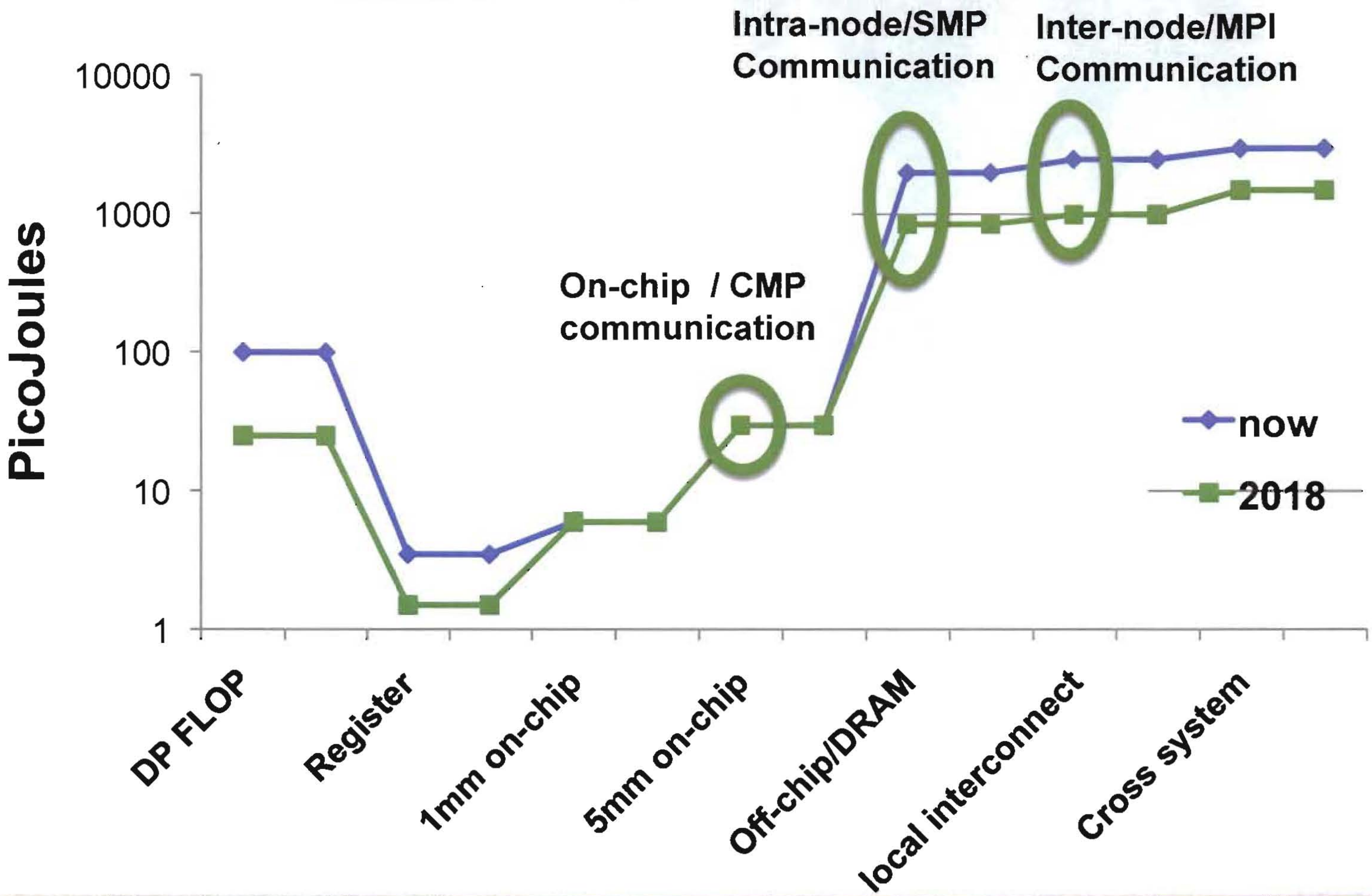
- **Co-design:**
  - **Focus on mini, proxy, or skeleton applications to facilitate co-design among apps, software and hardware technology efforts**
- **Advanced Applications:**
  - **Focus on advanced applications linking the co-design effort with ASC mission space.**
  - **This has no ASCR counter-part.**



Power will be a significant constraint on system performance.



Data movement and data locality are critical factors for managing power.

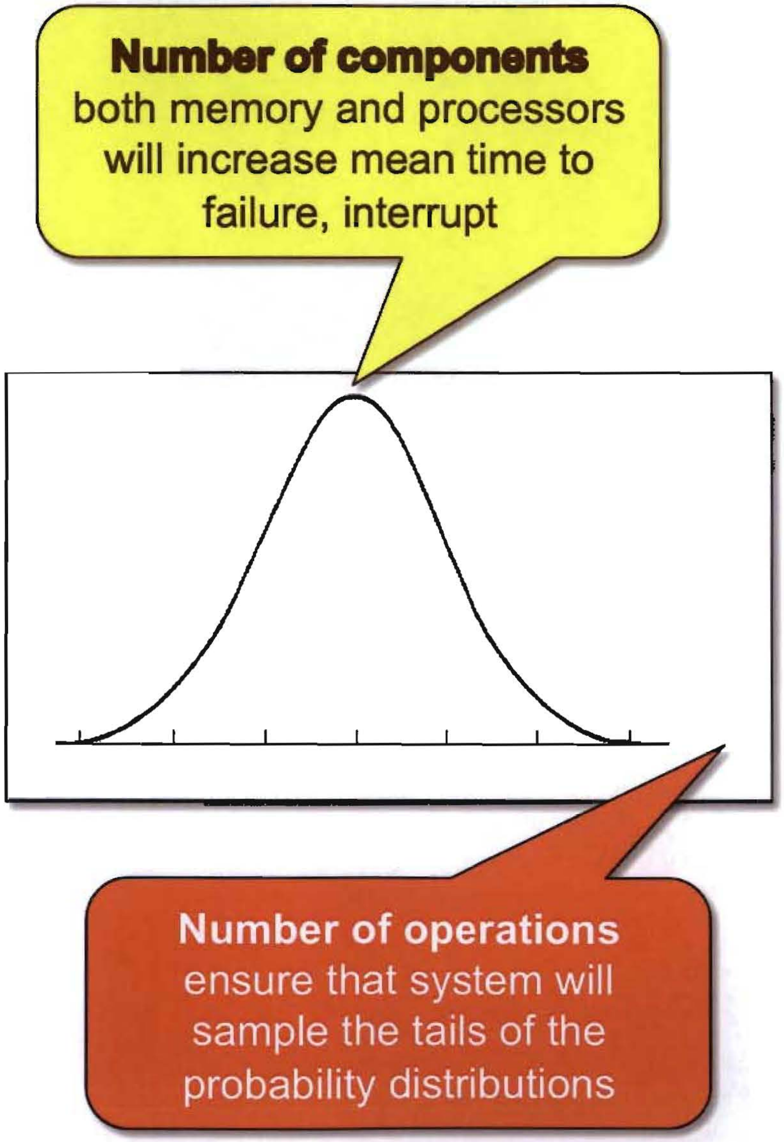


pJ → kW @Petascale

pJ → MW @Exascale

> "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*

Resiliency issues will affect hardware, software and perhaps even applications.



Number of components both memory and processors will increase mean time to failure, interrupt

Number of operations ensure that system will sample the tails of the probability distributions

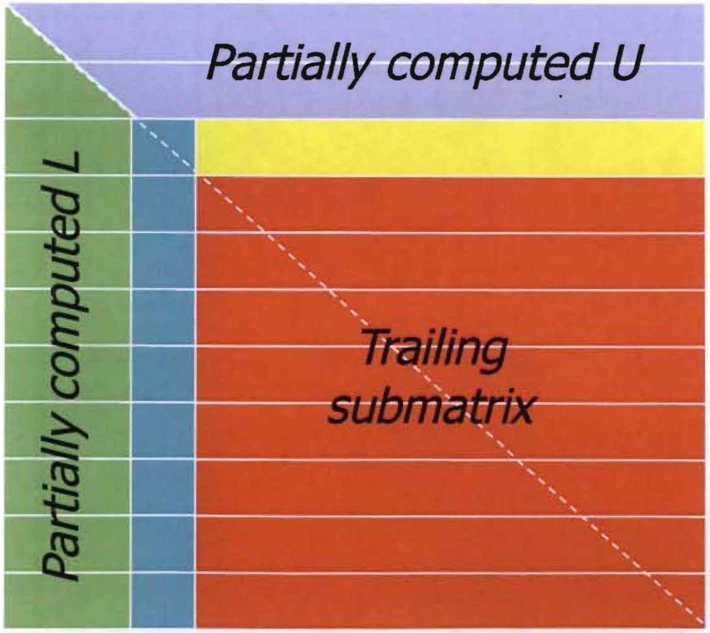
- Smaller circuit sizes, running at lower voltages to reduce power consumption, increases the probability of errors
- Heterogeneous systems make error detection and recovery even harder, for example, error recovery on GPU system will require managing up to 100 threads
- Increasing system and algorithm complexity makes improper interaction of separate components more likely.
- In will cost power, performance and $ to add additional HW detection and recovery logic right on the chips to detect silent errors.

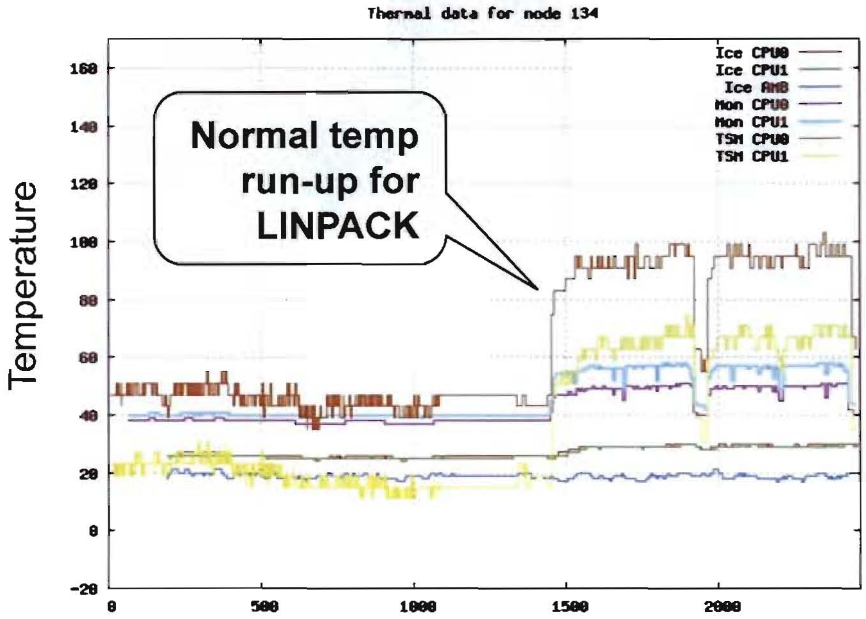
	Transient	Persistent
Detected	▨	▨
Undetected	▨	▨

Science Partnership for Extreme-scale Computing

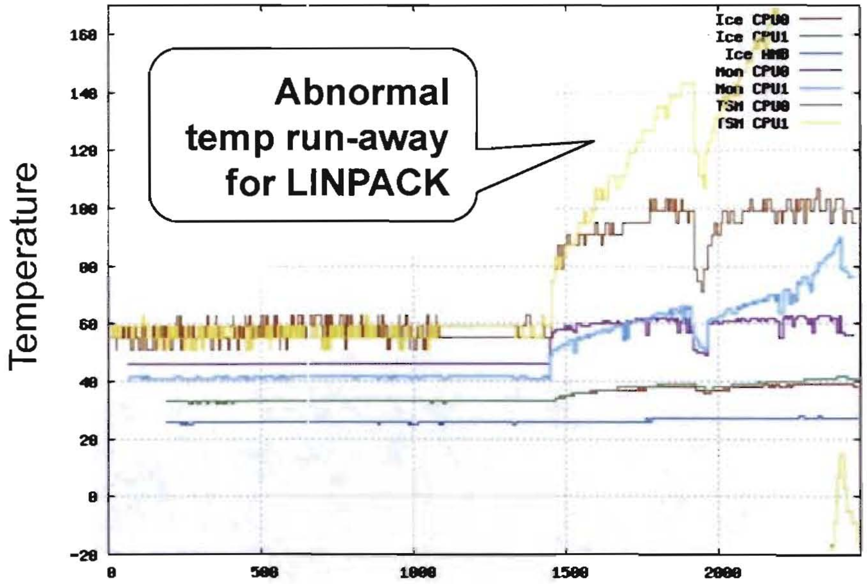
The rate and effect of undetected (aka silent) errors must be better understood.

- High Performance Linpack is an excellent tool for stressing systems



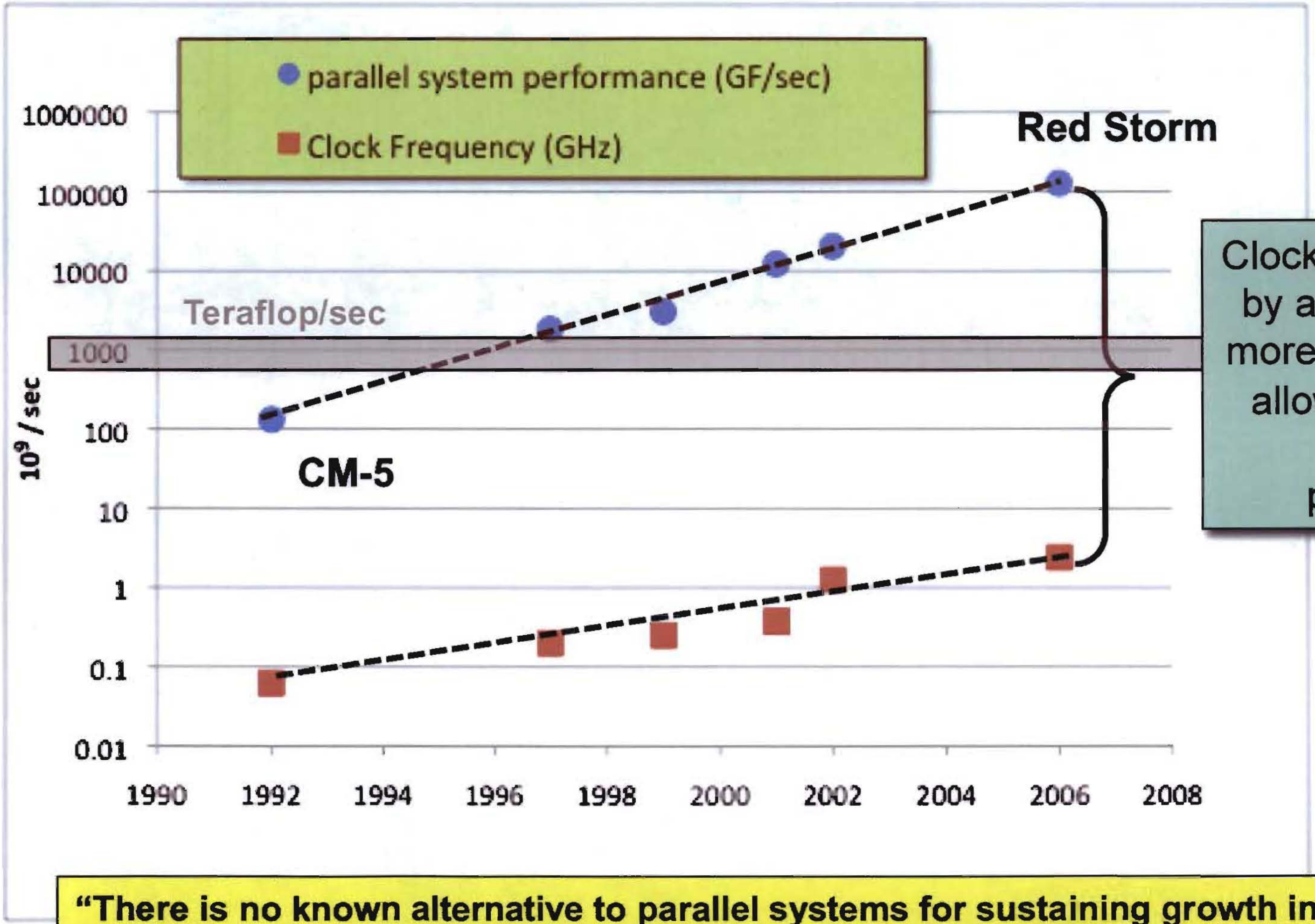


Normal temp run-up for LINPACK



Abnormal temp run-away for LINPACK

The last factor of 1000 was made possible by HPCC, ASCI and the marketplace.



Clock speed increased by a factor of 40 and more parallelism (25x) allowed a 1000-fold increase in performance.

"There is no known alternative to parallel systems for sustaining growth in computing performance; however, no compelling programming paradigms for general parallel systems have yet emerged." NRC, 2011

The next decade will see significant architectural innovation.

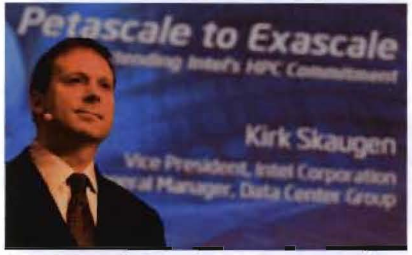
AMD

The future is fusion

AMD:
Delivering heterogeneous computing

Project Denver
NVIDIA-Designed
High Performance ARM Core

NVIDIA:
ARM CPU integrated with GPU



Petascale to Exascale
Extending Intel's HPC Commitment

Kirk Skaugen
Vice President, Intel Corporation
General Manager, Data Center Group

INTEL
Many Integrated Core architecture

Use parallelism to increase performance

"swim lane" #1
many cores

"swim lane" #2
many threads

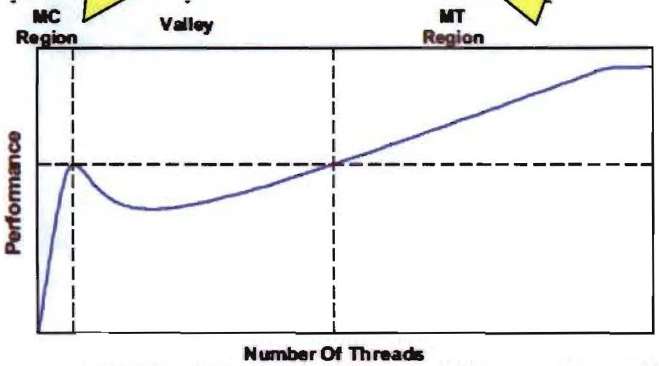
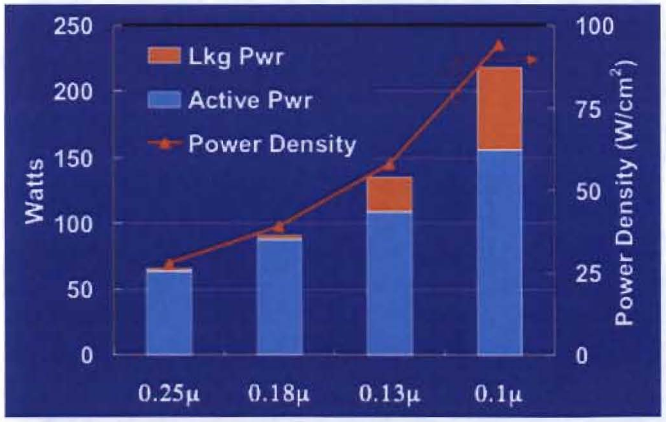


Fig. 1. Performance of a unified many-core (MC) many-thread (MT) machine exhibits three performance regions, depending on the number of threads in the workload.

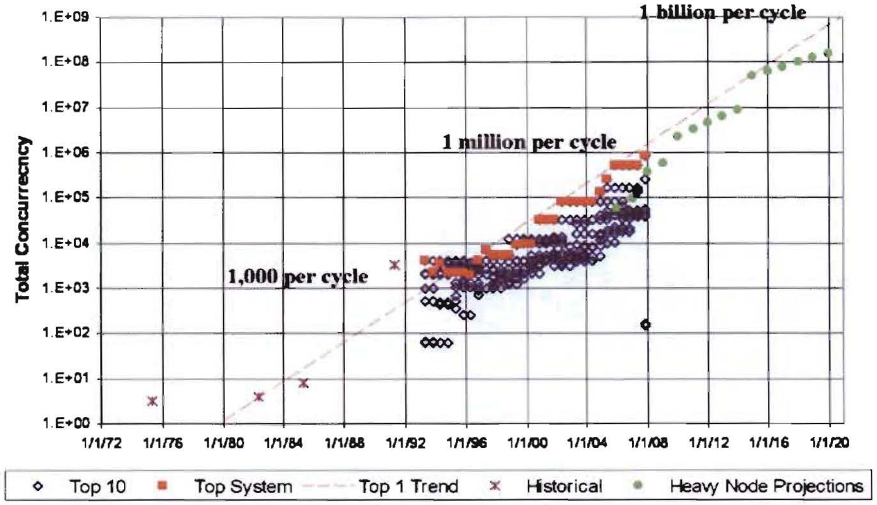
Mange on-chip power consumption



Chip power density = # gates * gate capacitance * frequency * voltage^2

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
    - Maybe 100B threads!
  - 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)
    - Additional levels of memory hierarchy (NVRAM)
  - Programming for resilience
  - Science goals require complex codes



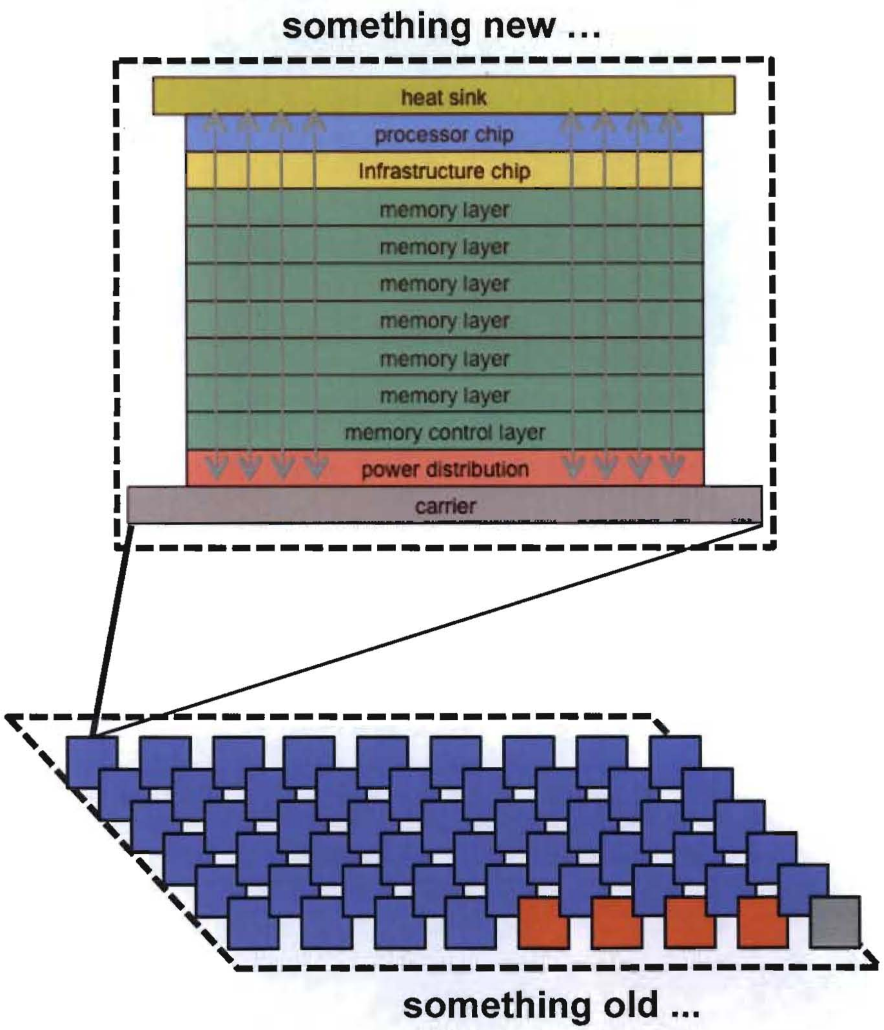
How much parallelism must be handled by the program?

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 models requires a dual approach.

- 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; provides path for incremental progress
    - 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

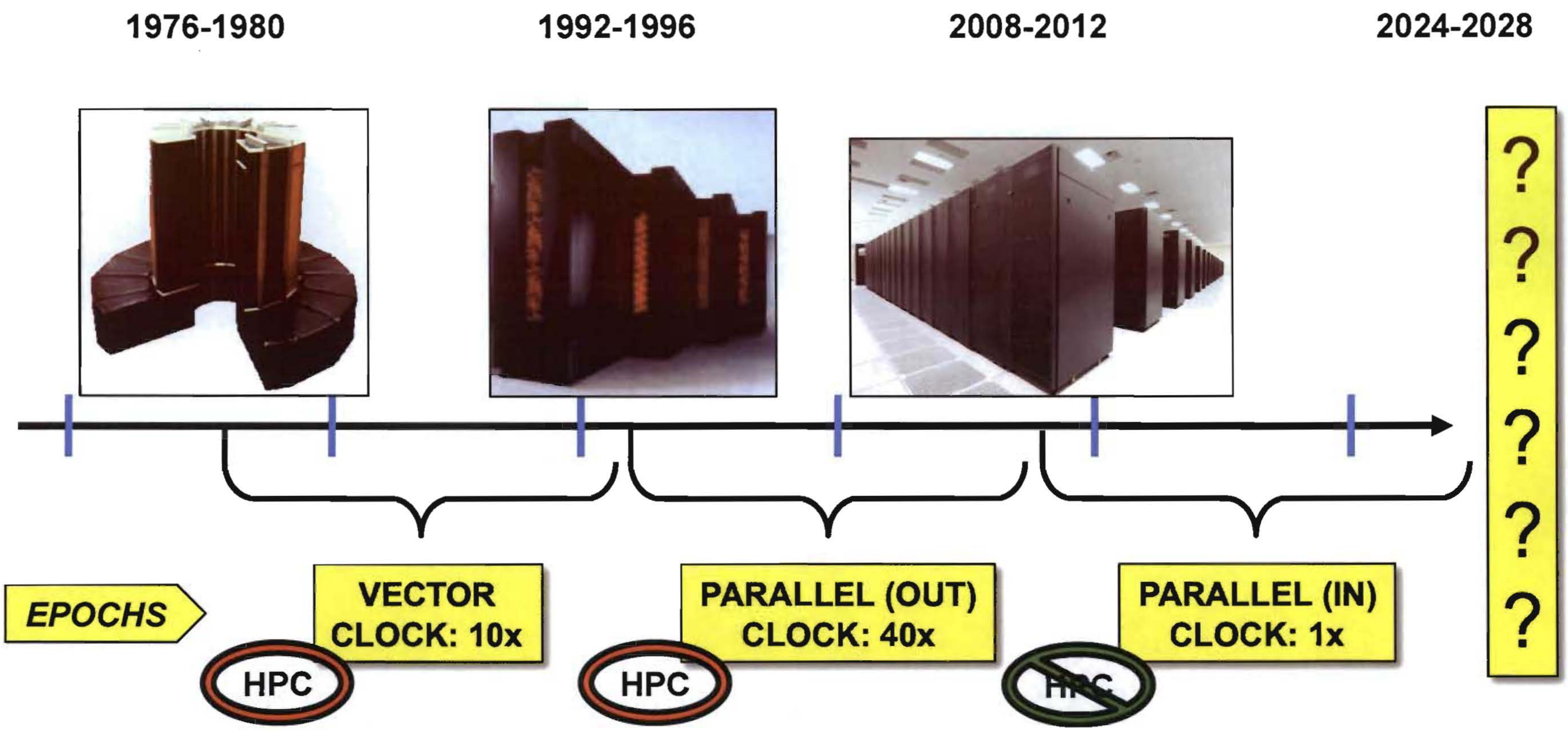


Applications and predictive science must transform with the technology.

- Power will be the number one architectural constraint
  - Applications will be effected by power efficient architectures
  - Applications may be directly involved in managing system power
  - Load balancing will have a new dimension
- On-chip: ten thousand way parallelism, deeper/higher memory hierarchies, 100x more upsets/sec
  - New programming models, languages and run-time systems
  - Fault-aware applications and fault-tolerant algorithms
- Cheap flops, expensive data motion, very expensive I/O
  - Remap multi-physics and algorithms to maximize data reuse and locality
  - Data analysis on-the-fly and embedded UQ
  - Reformulate algorithms to trade flops for memory use

… but this is not just a challenge; it is also an opportunity to transform our capability to do predictive science and engineering.

We also need to prepare for the next transformation.



1976-1980

1992-1996

2008-2012

2024-2028

EPOCHS

VECTOR
CLOCK: 10x

HPC

PARALLEL (OUT)
CLOCK: 40x

HPC

PARALLEL (IN)
CLOCK: 1x

~~HPC~~

?
?
?
?
?
?

The 2009 ITRS predicts the continued growth of transistors/chip for the next decade, but it will probably not be possible to continue to increase the transistor density (number of transistors per unit area) of CMOS chips at the current pace beyond the next 10 to 15 years.