LA-UR- 11-06281

Approved for public release;
distribution is unlimited.

Title: Exascale Challenges and Opportunities

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

Intended for: Council on Competitiveness

Los Alamos
NATIONAL LABORATORY
EST.1943

Title: Exascale Challenges and Opportunities

Abstract: Up-date for the Council on Competitiveness.

E7

Exascale Challenges and Opportunities

for

Council on Competitiveness

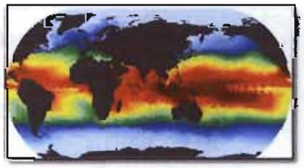
Andy White

Los Alamos National Laboratory

E7

DOE mission imperatives require simulation & analysis to inform 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 weapons science challenges
  - Directed Stockpile Work

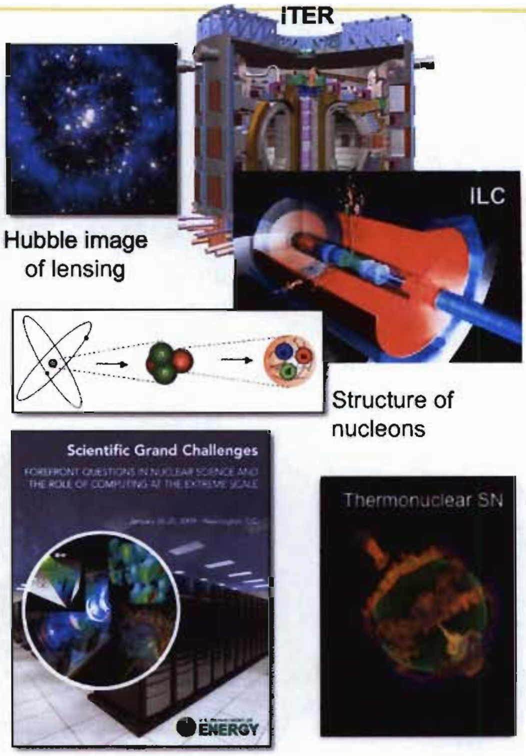


E7

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.


ITER
ILC
Scientific Grand Challenges
FOREFRONT QUESTIONS IN NUCLEAR SCIENCE AND THE ROLE OF COMPUTING AT THE EXTREME SCALE
ENERGY
Thermonuclear SN

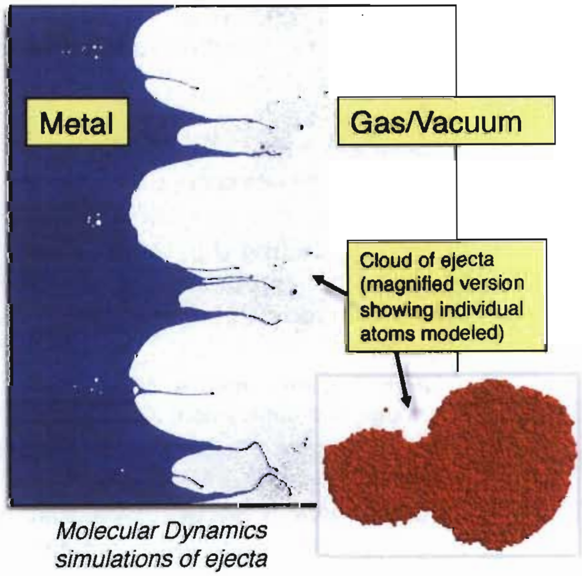
Hubble image of lensing

Structure of nucleons

E7

Increased computational power is driven by requirements to reduce uncertainty.

- Improved GEOMETRIC fidelity
  - Design features
  - UQ methodologies
  - Naturally 3D phenomena e.g. turbulence, material failure .....
- Improved NUMERICAL fidelity
  - Potentially important phenomena occur at 10x standard resolution
  - Bridging strongly-coupled multi-scale phenomena
  - Need to perform UQ studies over greater variable counts
  - Weapons science simulations displaying convergence at very large coverage
    - atoms or dislocations or ....
- Improved PHYSICS fidelity
  - Energy balance
  - Boost
  - Si radiation damage
  - Secondary performance

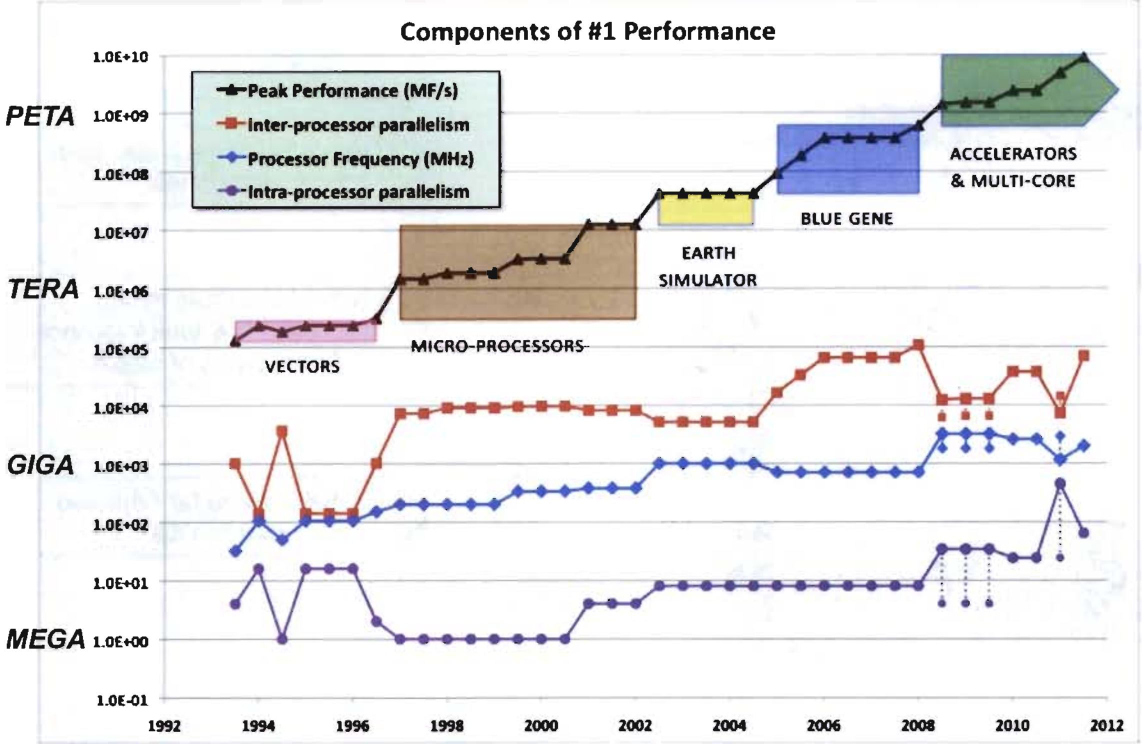

Metal
Gas/Vacuum
Cloud of ejecta (magnified version showing individual atoms modeled)

Molecular Dynamics simulations of ejecta

3D simulations of ductile spall failure with predictive potentials will require exascale resources

E7

History of high performance computing is multi-faceted.



E7

The exascale plan has three high-level components.

- Mission & science applications
  - Models
  - Applied math, libraries
- Software
  - Programming models
  - Tools, libraries, OS
- Hardware
  - Cross-cutting technology
  - Exascale Research & development
  - Acquisition & facilities

Co-design & Uncertainty Quantification

Los Alamos NATIONAL LABORATORY EST.1943

Argonne NATIONAL LABORATORY

Lawrence Livermore National Laboratory

OAK RIDGE National Laboratory

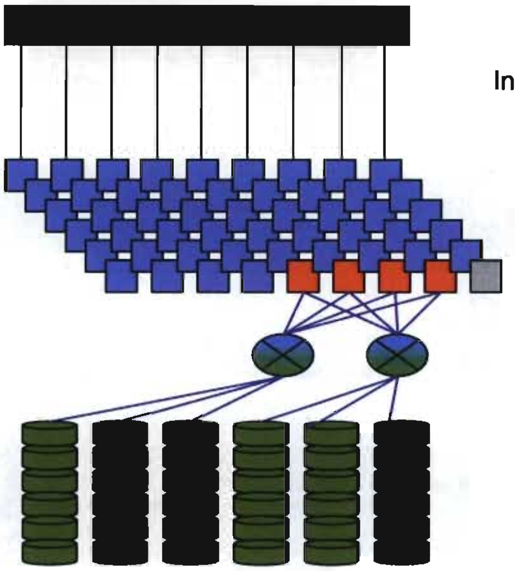
BERKELEY LAB

Sandia National Laboratories

Pacific Northwest NATIONAL LABORATORY

E7

The overall system design may be similar to today's systems



System Interconnect

Exascale System

I/O Network

System Storage

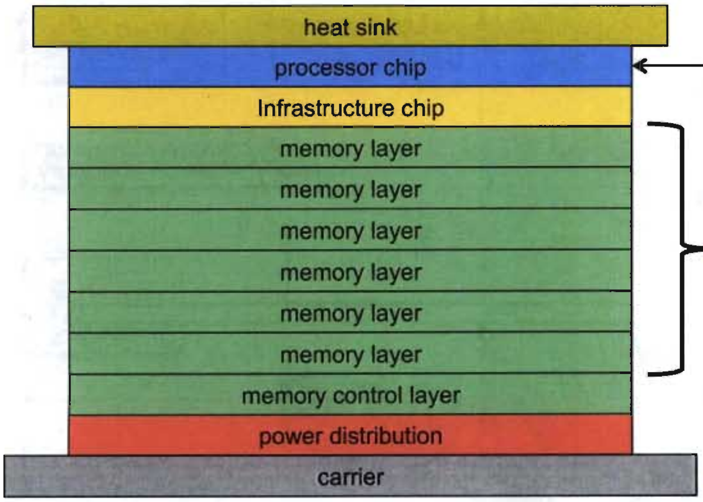
- New interconnect topologies
- Optical interconnect

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

- Global storage far removed from application data

E7

The processor is key for exascale, as well as petascale and terascale systems.



heat sink

processor chip

Infrastructure chip

memory layer

memory layer

memory layer

memory layer

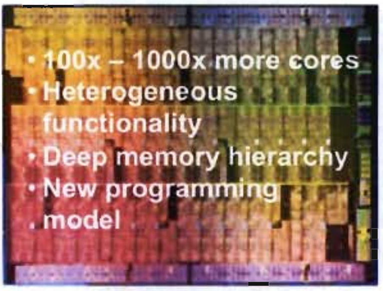
memory layer

memory layer

memory control layer

power distribution

carrier



- 100x – 1000x more cores
- Heterogeneous functionality
- Deep memory hierarchy
- New programming model

- 3d stacked memory

- Smart memory management

- Integration on package

E7

“You can run but you can’t hide.”

Joseph Louis Barrow

- **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.
- **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.
- **Power** will become the first class constraint on system performance and effectiveness at exa-scale, at peta-scale and at desktop-scale.
- **Memory** is not scaling with performance and memory hierarchies will be higher and deeper.
- **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.
- **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.*

E7

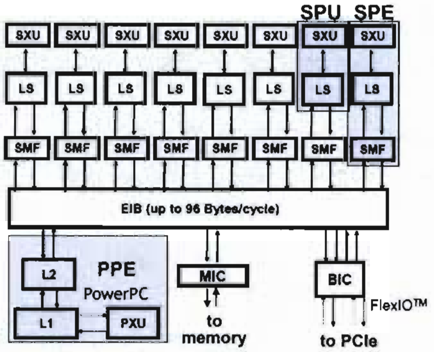
“The Future of Computing Performance: Game Over or Next Level?” NRC, 2011

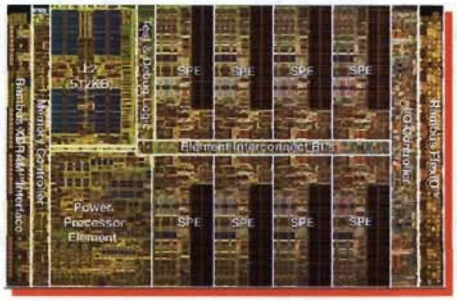
“The U.S. Computing Industry has been adept at taking advantage of increases in computing performance, allowing the United States to be a moving and therefore elusive target – innovating and improvising faster than anyone else.”

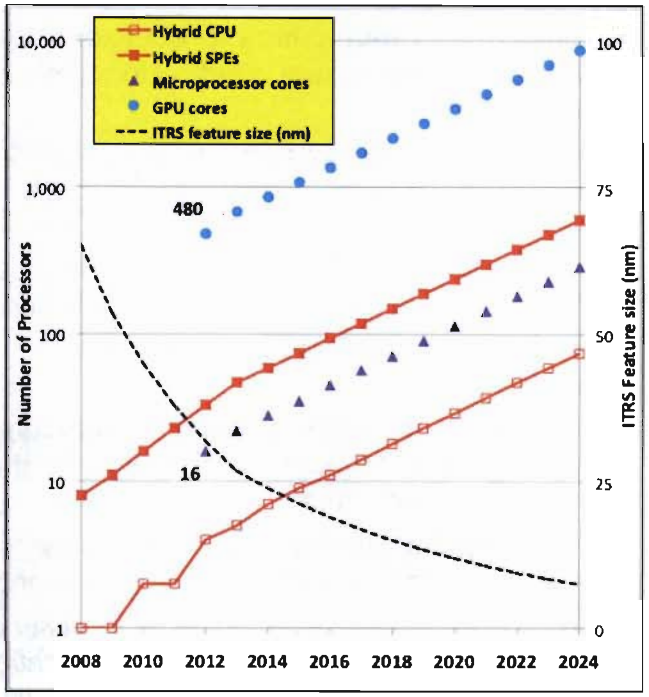
- Invest in research in and development of algorithms that can exploit parallel processing.
- Invest in research in and development of programming methods that will enable efficient use of parallel systems …
- Focus long-term efforts on rethinking of the canonical computing “stack” …
- Invest in research 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.”

E7 On-chip parallelism will continue to increase throughout the decade.







E7 Everyone is preparing for transformation.

AMD
The future is fusion

AMD:
Delivering heterogeneous computing

NVIDIA:
ARM CPU integrated with GPU

INTEL
Many Integrated Core architecture

Use parallelism to increase performance

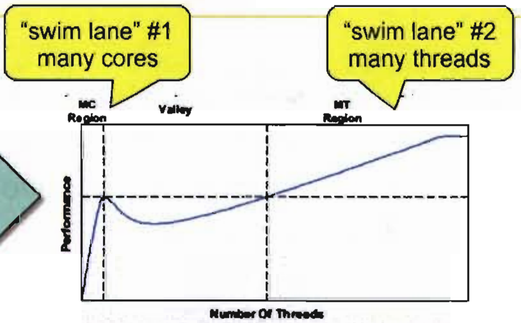
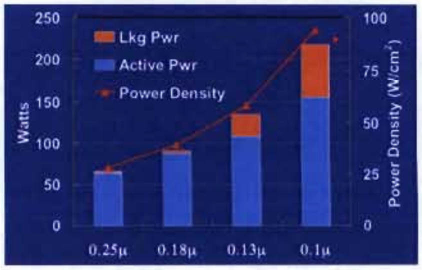


Fig. 1. Performance of a unified many-core (MC) many-thread (MT) machine exhibits three performance regions, depending on the number of threads in the workload.

Manage on-chip power consumption



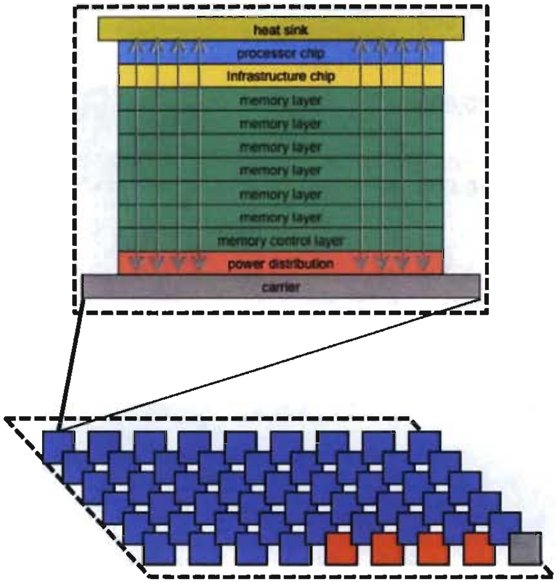
Chip power density = # gates * gate capacitance * frequency * voltage^2

E7

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

something new ...



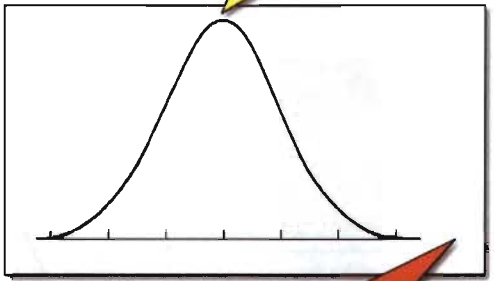
something old ...

Slide 13

E7

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

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

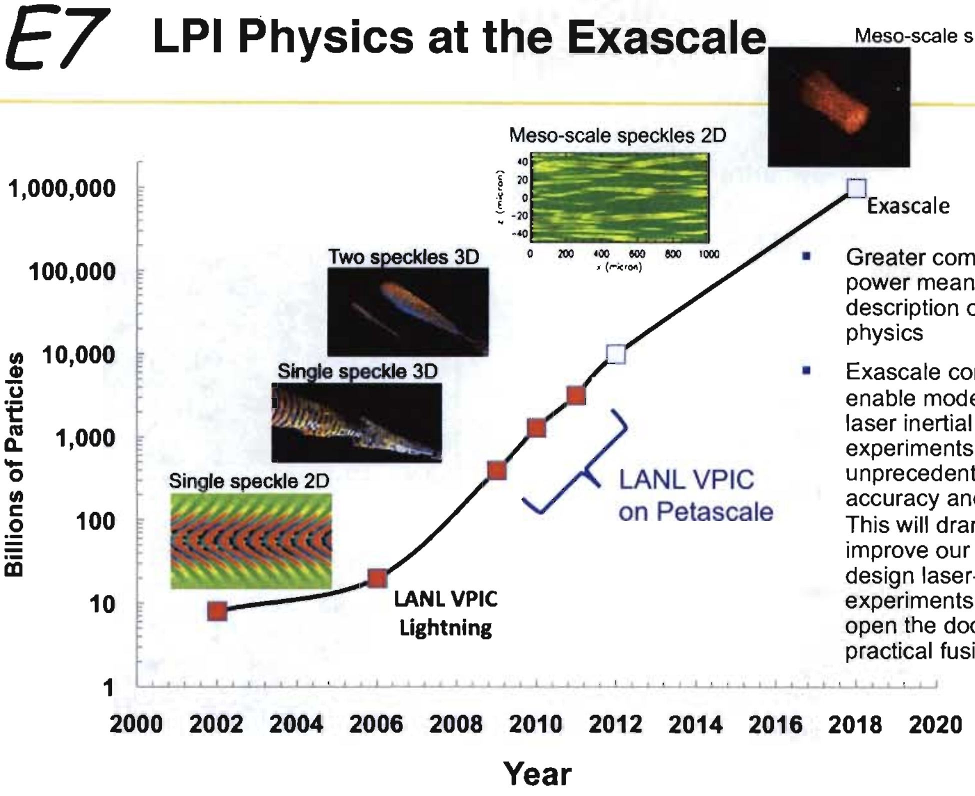


Number of operations ensure that system will sample the tails of the probability distributions

- 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	🟧	🟥

E7 LPI Physics at the Exascale

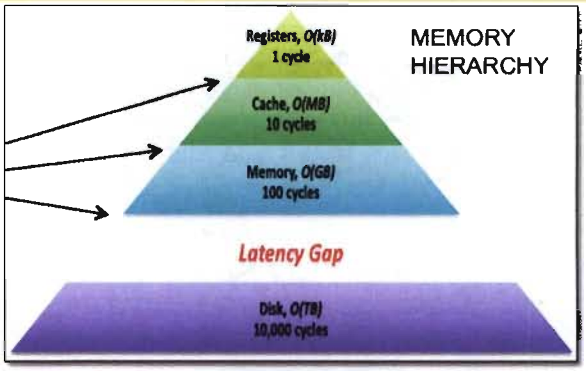


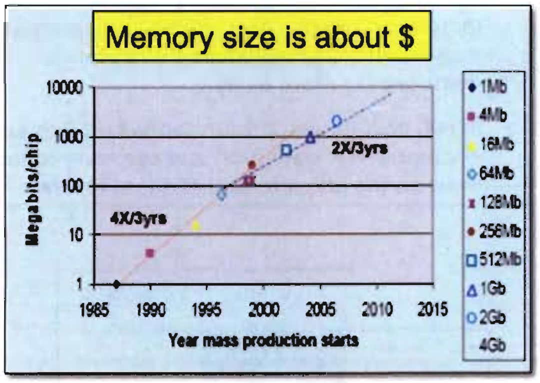
- Greater computer power means a better description of the physics
- Exascale computing will enable modeling of laser inertial fusion experiments at unprecedented levels of accuracy and fidelity. This will dramatically improve our ability to design laser-fusion experiments and could open the door for practical fusion energy

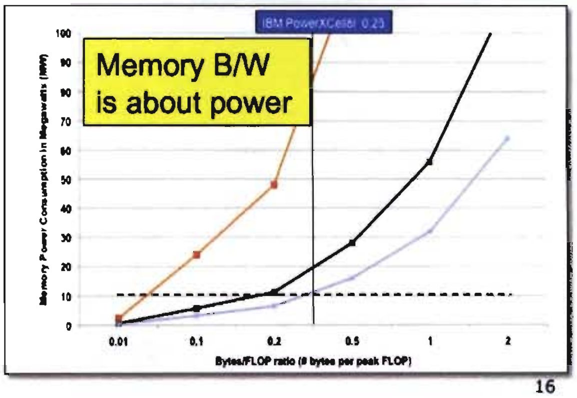
E7 Memory size, bandwidth and hierarchy will be challenges by 2018, if not sooner.

The memory hierarchy will be much richer at the end of the decade than it is now:

- *Software managed caches or scratch pad memory*
- *Very fast 3D stacked memory*
- *NVRAM for check-pointing and extended memory*

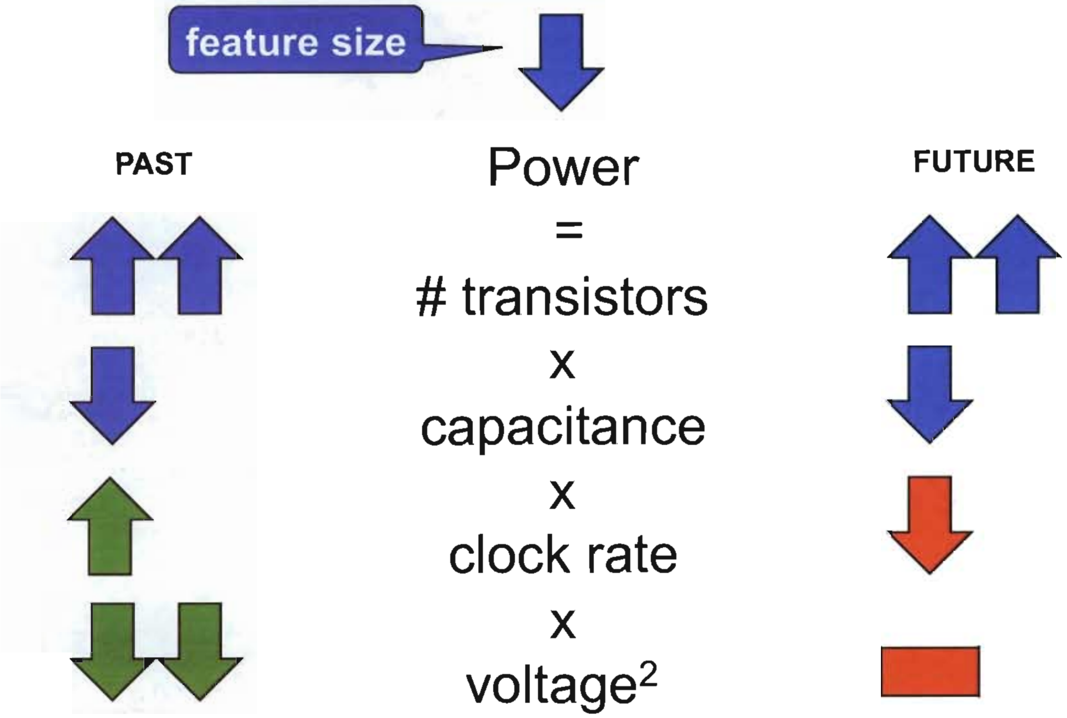






E7

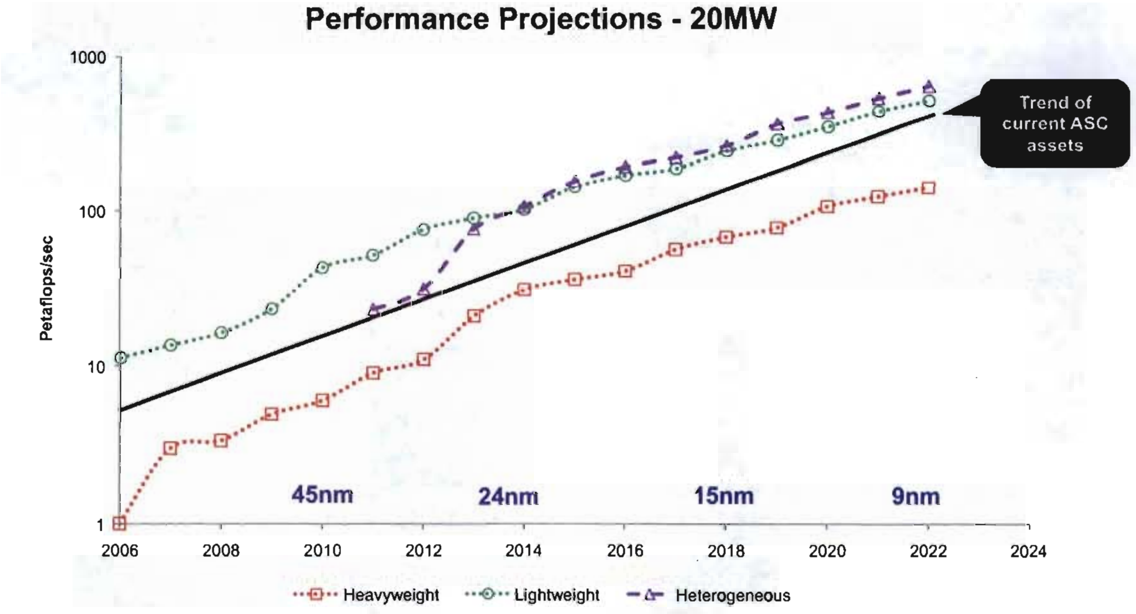
So what's happened to the good ol' days?



Power
=
# transistors
x
capacitance
x
clock rate
x
voltage^2

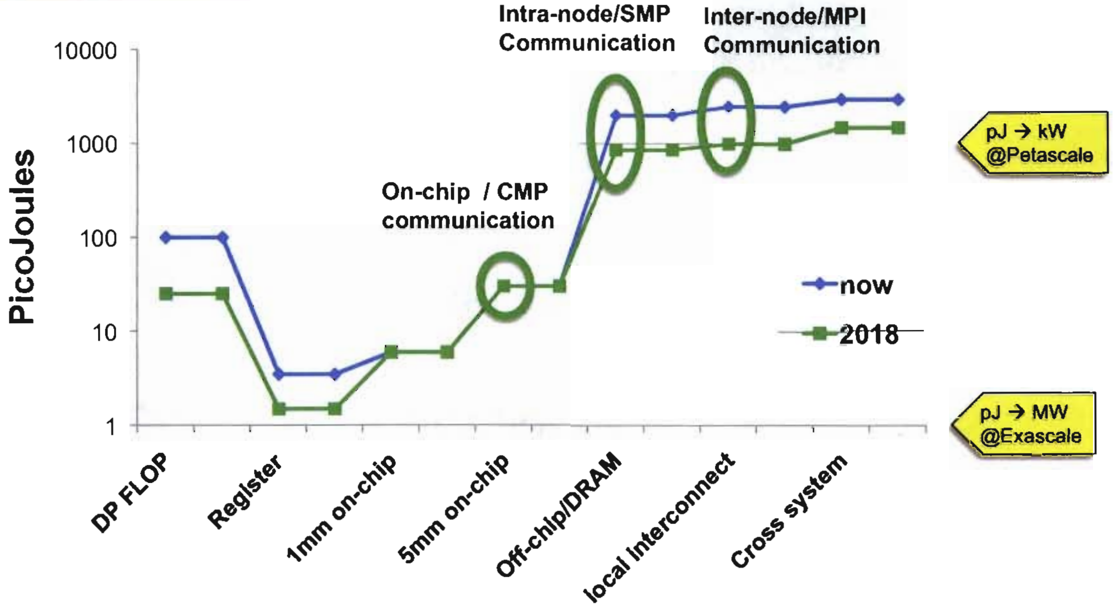
E7

Meeting the 20 MW power target will be a challenge



Trend of current ASC assets

… but moving data will be the real test of power for exascale.



> "The Energy and Power Challenge is the most pervasive."
> *DARPA IPTO exascale technology challenge report*

System architecture targets are aggressive in schedule and scope.

System attributes	2010	"2		2019-2020 targets	
Performance	2 PF/s	200 P		1 Exaflop/sec	
Power	6 MW	1		20 MW	
System memory	0.3 PB			128 PB	
Node performance	125 GF/s	500 G		1 TF/s	10 TF/s
Node memory BW (consistent with 0.4 B/F)	25 GB/s	200 G		400 GB/s	4 TB/s
Node concurrency	12	100	1,000	1,000	10,000
System size (nodes)	18,700	400,000	40,000	1,000,000	100,000
Node link BW (consistent with 0.1 B/F)	1.5 GB/s	50 GB		100 GB/s	1 TB/sec
Mean time before application failure	days			144 hours	
IO	0.2 TB/s			60 TB/s	

PEAK?
LINPACK?
AVERAGE?

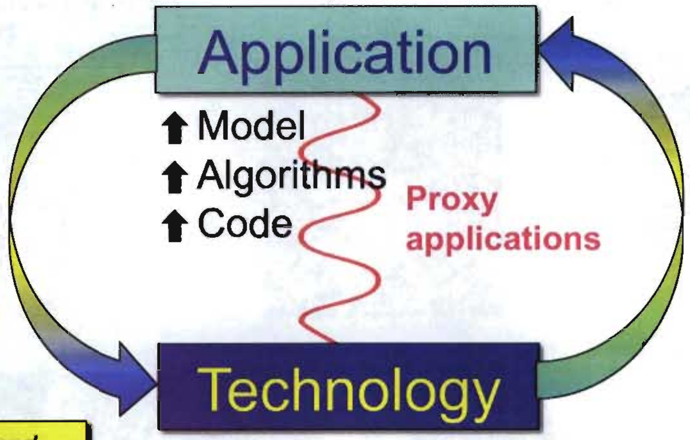
WHAT DO YOU MEAN BY "MEMORY"?

WHAT DO YOU MEAN BY "FAILURE"?

E7

Co-design is the essential opportunity in the exascale activities.

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



⬆ Model
⬆ Algorithms
⬆ Code

⊕ programming model
⊕ operating system
⊕ architecture

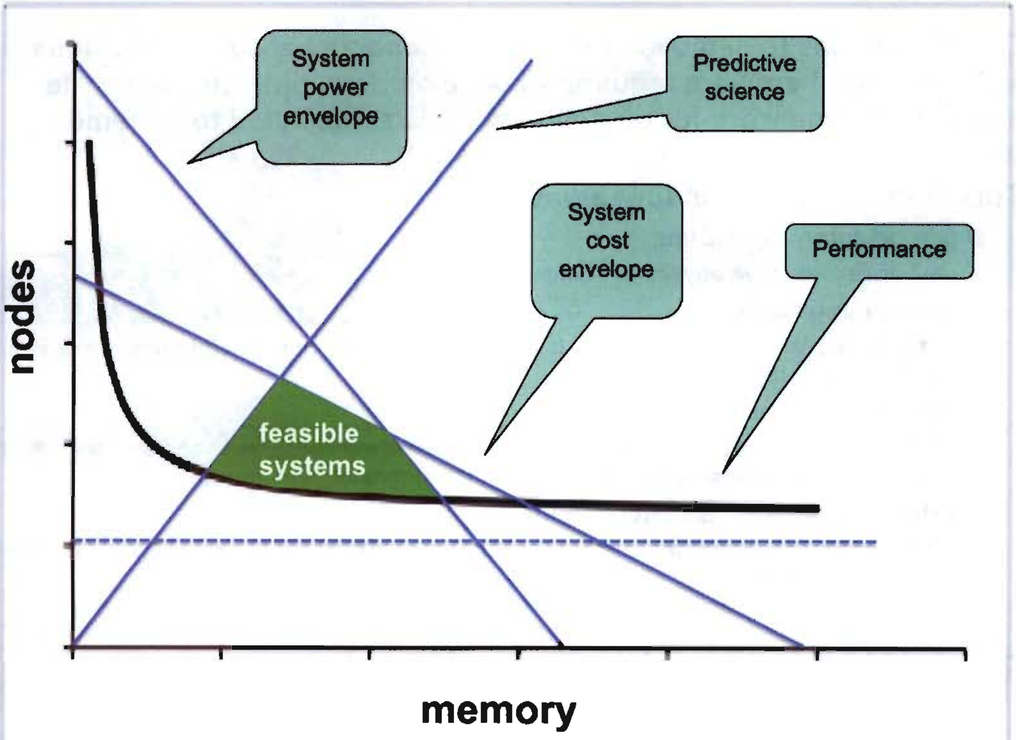
Key issues
Power?
Performance?
Price?
Parallelism?
Productivity?

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.

E7

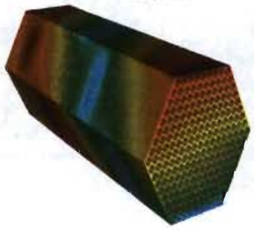
The trade space for exascale is very complex.



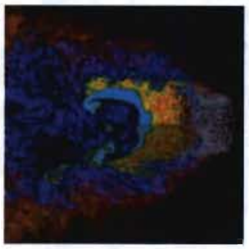
E7

First ASCR co-design centers are off and running

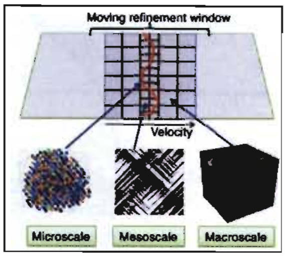
Advanced Reactors
ANL

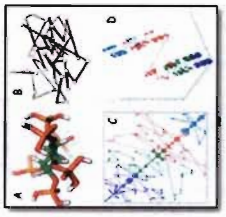


Combustion
SNL



Materials in Extreme Environments
LANL





Chemistry
ORNL

Earth systems
LANL & ORNL

Fusion
LBNL

High-energy Density
Physics
ANL

E7

Exascale Co-Design Center for Materials in Extreme Environments

- Our objective is to establish the interrelationship between algorithms, system software, and hardware required to develop a multiphysics exascale simulation framework for modeling materials subjected to extreme mechanical and radiation environments.
- This effort is focused in four areas:
  - Scale-bridging algorithms
    - UQ-driven adaptive physics refinement
  - Programming models
    - Task-based MPMD approaches to leverage concurrency and heterogeneity at exascale while enabling fault tolerance
  - Proxy applications
    - Communicate the application workload to the hardware architects and system software developers, and used in performance models/simulators/emulators
  - Co-design analysis and optimization
    - Optimization of algorithms and architectures for performance, memory and data movement, power, and resiliency

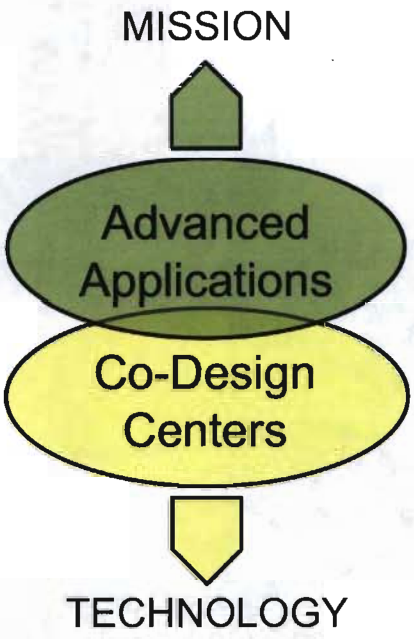
ExMatEx

Los Alamos National Laboratory Est. 1943 · Lawrence Livermore National Laboratory · Oak Ridge National Laboratory · Sandia National Laboratories · Stanford University · California Institute of Technology

E7

ASC has a broad application scope.

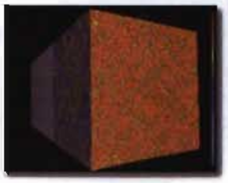
- ASC must meet the on-going needs of the weapons program
  - Provide increasingly better predictive physics and engineering capabilities
  - Provide increasingly more capable computational resources to support predictive science
  - Maintain the necessary core capabilities of the weapons program staff, 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



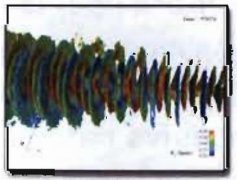
E7

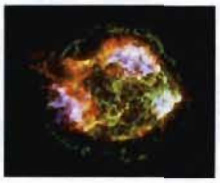
Roadrunner was a harbinger of this future.

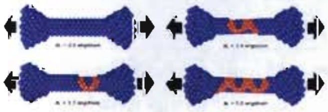
- Roadrunner was an exercise in co-design
  - Memory interface and DP-arithmetic
  - Software programming infrastructure (DaCS, ALF)
  - Redesign of system in flight
  - Focus on applications
- Roadrunner was a leap into the future
  - *First* computer to reach a petaflop
  - *First* heterogeneous supercomputer
  - *First* accelerated supercomputer
  - *First* supercomputer built from non-traditonal commodity processor
- Roadrunner defined advanced systems for ASC
  - Science at scale
  - Important physics modules
- Roadrunner open science provided resources for many important simulations --

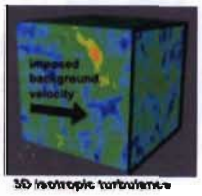

Shocks in metal

Magnetic reconnection


Laser plasma interaction


Core collapse supernova

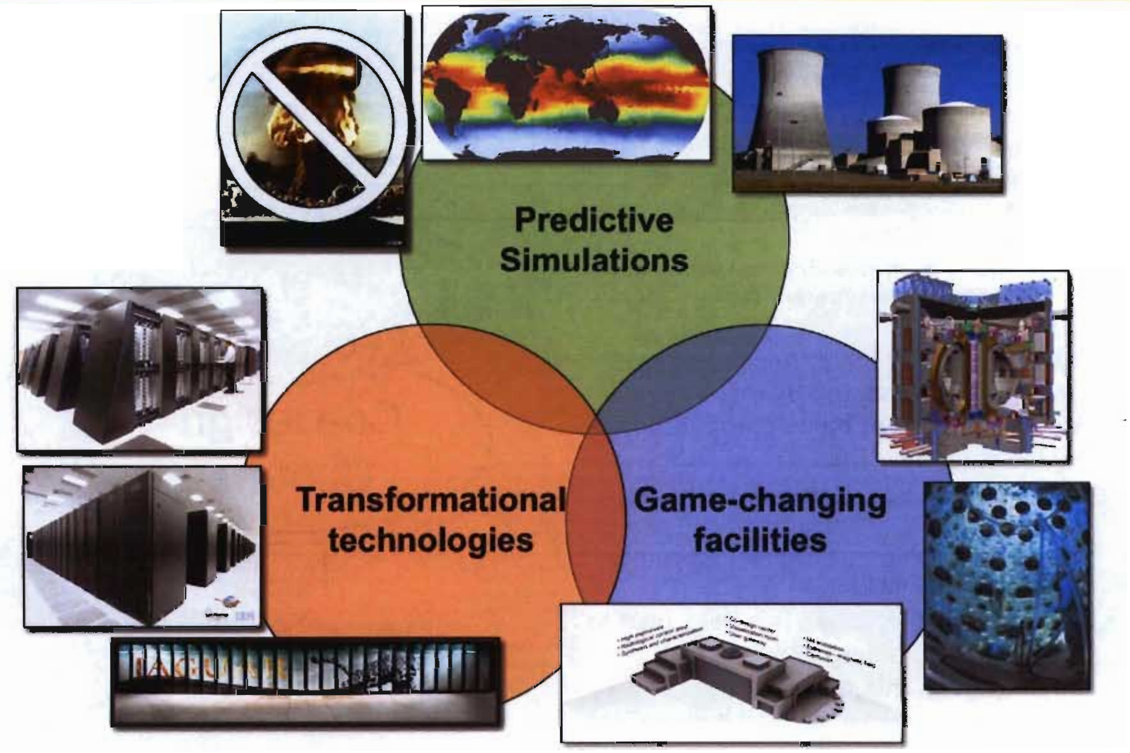

Accelerated MD



Turbulence with TN burn

E7

Co-design has an even broader context.



Predictive Simulations

Transformational technologies

Game-changing facilities

E7

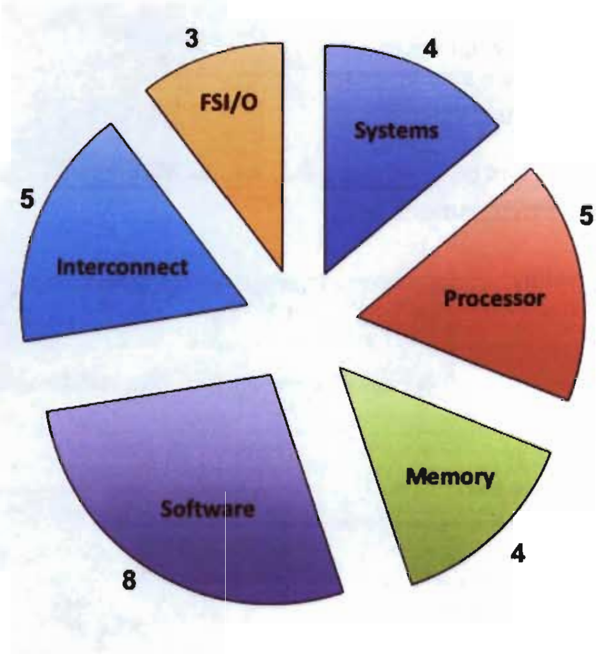
Applications and predictive science must transform with the technology.

- 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.

E7

There was a broad spectrum response to exascale R&D request for information.



- Variety of business opportunities
  - Create a competitive advantage in HPC
  - Leverage to create larger volume markets
- Variety of potential participants
  - Very large to very small
  - Broad portfolio to providing specific technologies
- Requirements
  - Coherence
  - Binding of R&D to mission

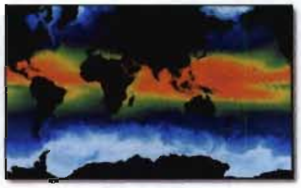
E7

The history of HPC is a sequence of key transformations.

Era	Mainframe	Vector	Single-level parallel	Multi-level parallel
Time frame	60s to mid-70s	mid-70s to early-90s	Early 90s to early 10s	Early 10s to mid-20s
Fundamental scale	System in a room	System in a chassis	System on a board	System on a chip
Free performance	---	16x	40x	1x
Dominant constraint	Floating point capability	Physical size of system	Interconnect scalability	Energy efficiency
Architectural challenges	---	Scatter-gather	Interconnect	Power Memory size Data motion Resiliency Heterogeneity
Programming model	Sequential processes	Vectorized sequential	Communicating sequential	Hierarchical parallel
Programming challenges	Expression of mathematical algorithms	Vectorized instructions & loops	Distributed applications & message passing	Data motion Multi-level parallelism Resiliency
Fundamental building block	Commercial CPUs	Custom CPUs	Commodity micro-processors	Heterogeneous cores
R&D initiatives	---	---	HPCC, ASCI	Exascale

E7

Success is clearly defined for the exascale initiative.



- Success of the initiative is:
  - Transformational capabilities in national nuclear security, climate, energy and science enabled by predictive exascale simulations
  - U.S. industry leadership in information technology lead by aggressive exascale technology development
  - Competitive advantage for U.S. energy-related and other industries
- Co-design of applications, computational environment and platforms is critical
  - Application teams must have dual responsibility
    - Mission/science
    - Exascale co-design
  - Simulation environment will
    - Have broad community participation
    - Be common across all applications and platforms
    - Leverage open source software and product support
    - Support both evolutionary and revolutionary approaches
  - Long term industry partnerships are essential to success of this 10 year initiative
    - Must leverage and influence the business plan of vendor partners
    - Joint R&D and leveraged community efforts reduce vendor risk
    - Having at least two tracks provide competitiveness, risk reduction and architectural diversity
    - As does deployment of at least two platform generations to get to Exascale