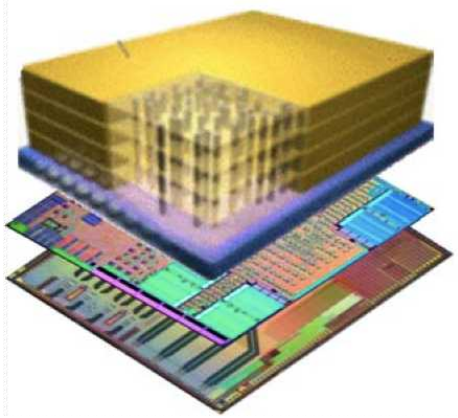
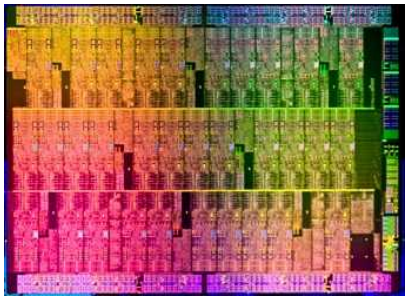
SAND-2012-####P

SAND2012-5741P

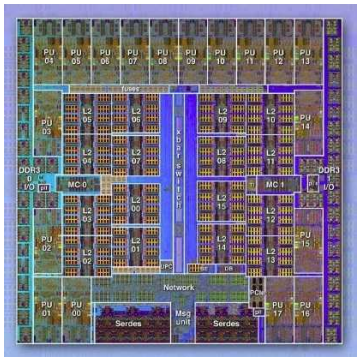




Hardware Issues:

Industry Hardware Trends – what will happen if we don't do anything?

JASON Exascale Study
June 27-29, 2012
The MITRE Corporation
La Jolla, CA



James Ang, Ph.D.
Scalable Computer Architectures
Sandia National Laboratories, NM
jaang@sandia.gov

Sandia National Laboratories

Exceptional service in the national interest

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Discussion Topics

Sandia National Laboratories

- **What is different about Exascale?**
- **Moore's Law**
- **Top500 List and Linpack**
- **Hardware Components**
  - Processor Trends
  - Memory Trends
  - Resilience Trends
  - Interconnection Network Trends
- **An Exaflop Linpack system ≠ a DOE mission Exascale system**
- **Co-Design to design better hardware & architectures**
- **Hardware Recommendations**

What is Different about Exascale?

Sandia National Laboratories

- **Time – through the lens of a decade:**
  - Opportunity for several generations of Hardware
  - At most, one generation of Application Software
- **Computer Architecture/Hardware Changes:**
  - Not just an opportunity for change
  - Essential to meet DOE mission application needs
- **International Competition**

China, Japan, Europe, India and Russia have significant investments in Exascale

Sandia National Laboratories

Japan To Invest $1.3 Billion In New Supercomputer

SERKAN TOTO | posted on Friday, August 12th, 2011

There is a list of the world's 500 supercomputers, and the last t updated, back in June this year, (pictured) came out on top, takin from Tianhe-1A (a supercompute

It was the first time since 2004 fo claim those bragging rights, and largest business newspaper The that the government is already thinking at will happen in 2020: by then, the plan is t a computer that handles exascale computing or, in other words, one million trillion operations second (that computer would be 100 times more powerful than K).

Enter China: A prototype four-core Loongson 3 will be produced at commercial scale by STMicro starting this year. Credit: Institute of Computing Technology, Chinese Academy of Sciences

COMPUTING

China Details Homemade Supercomputer Plans

The machine will use an unfashionable chip design.

TUESDAY, JA

Audio

It's officia Dawning home-gro of the Loc CPUs at t

Plan panel capabilitie

G. SRINIVASAN

Dr Ashwani Kumar

NEW DELHI, MAY 25:

India is all set to step up its supercomputing capabilities and capacity with the Planning Commission seeking to take this as "a principal initiative in the Twelfth Five-Year Plan (2012-17) in consultation with other line Ministries and research and scientific institutions".

Disclosing this to *Business Line* here in an exclusive talk, the Minister of State for Planning and Science & Technology and Earth Science, Dr Ashwani Kumar, said that the Plan panel has got a specific report commissioned on the subject and recently flagged off the first round of discussion.

CHINA'S CAPABILITIES

Elaborating upon the "overarching" priority to underpin supercomputing capabilities of the country, Mr Kumar contended that in 2007-08, the country's supercomputing capabilities were more or less equal to that of China but since then

HPC wire

Since 1986 - Covering the Fastest Computers in the World and the People Who Run Them

Select Language | Translation Disclaimer

Welcome Guest
Subscribe | Sign In

Since 1986 - Covering the Fastest Comp

Visit additional Tabor Communication Publications

Home | News | Features | Blogs | HPC Markets | Whitepapers

Home | News | Photo Gallery

Digital Manufacturing report

HPC in the Cloud

2011

March 27, 2012

Russia Sets Its Sights on Exascale

Michael Feldman

While much attention has focused on China's rising supercomputing pro lately, a number of other countries, including Russia, are also quickly m their HPC resources. With their eyes set on exascale, Russia is planning over a billion dollars this decade to field at least one such system by 20

January 06, 2011

European Exascale Project Drives Toward Next Supercomputing Milestone

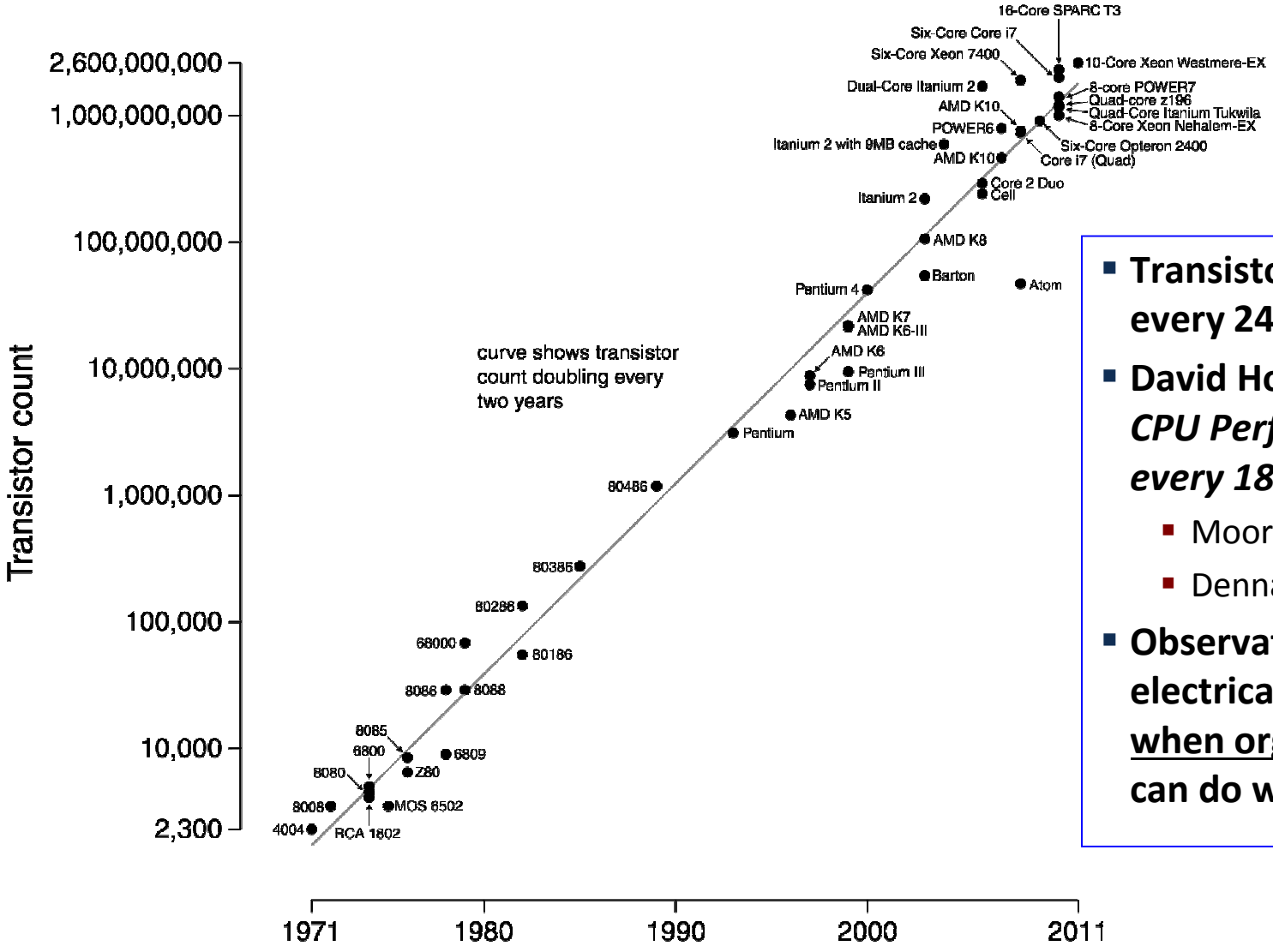
With petascale systems now deployed on three continents, the HPC industry is already looking toward the next milestone in supercomputing: exascale computing.

The European Union Exascale Project has language calling for *Geographically Restricted* IP

Moore's Law 1971-2011:

Growth in Transistor Count

Sandia National Laboratories

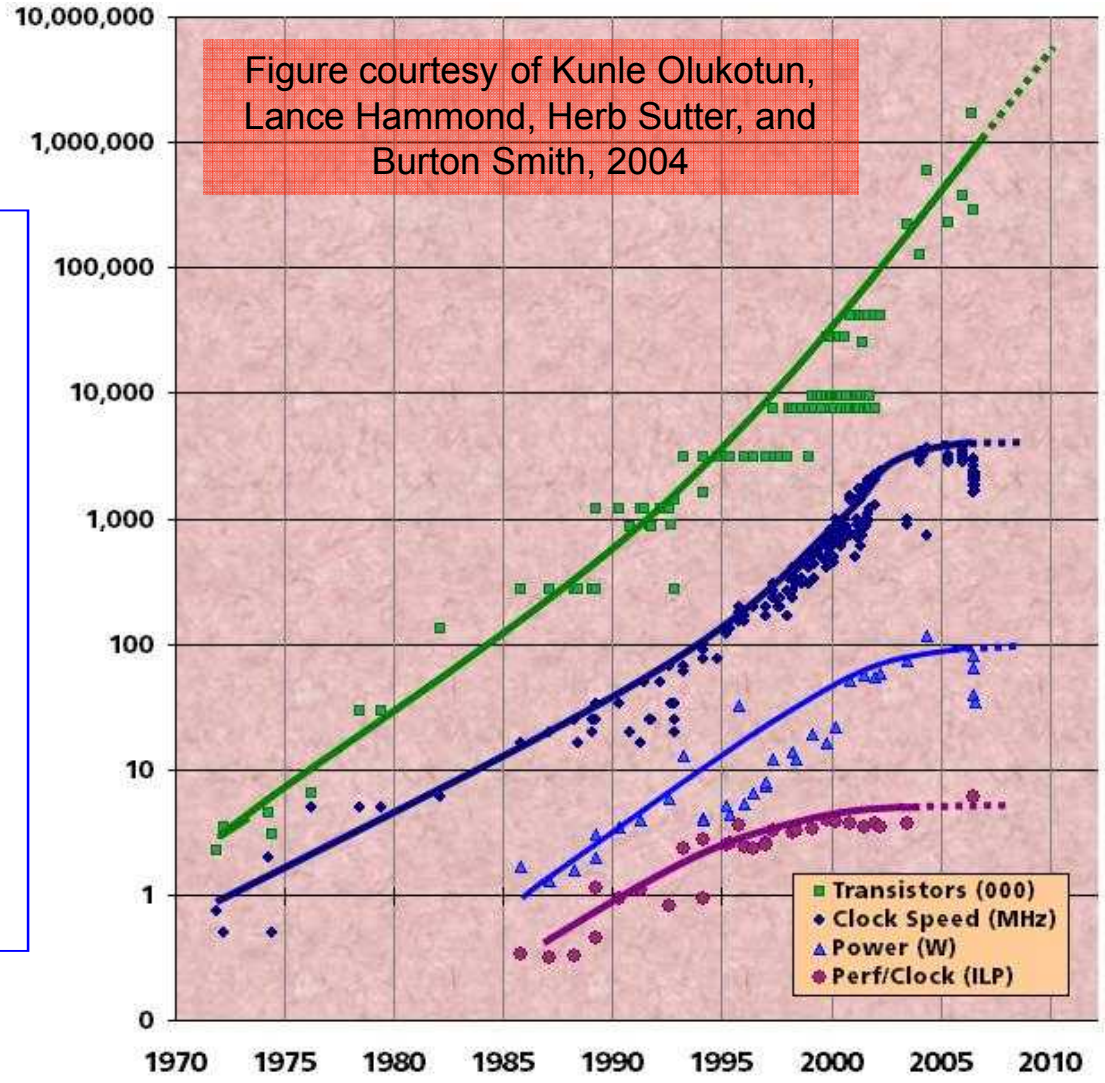


- **Transistor count doubles every 24 months**
- **David House, Intel: *CPU Performance doubles every 18 months due to*:**
  - Moore's Law
  - Dennard Scaling
- **Observation of what electrical engineers, <u>when organized properly</u>, can do with silicon**

Olukotun's 2004 Projections on Moore's Law and other trends

Sandia National Laboratories

- **Moore's Law continues**
  - Transistor count still doubles every 24 months
- **Dennard scaling stalls – key parameters flatline:**
  - Voltage
  - Clock Speed
  - Power
  - Performance/clock



Moore's Law (2010)

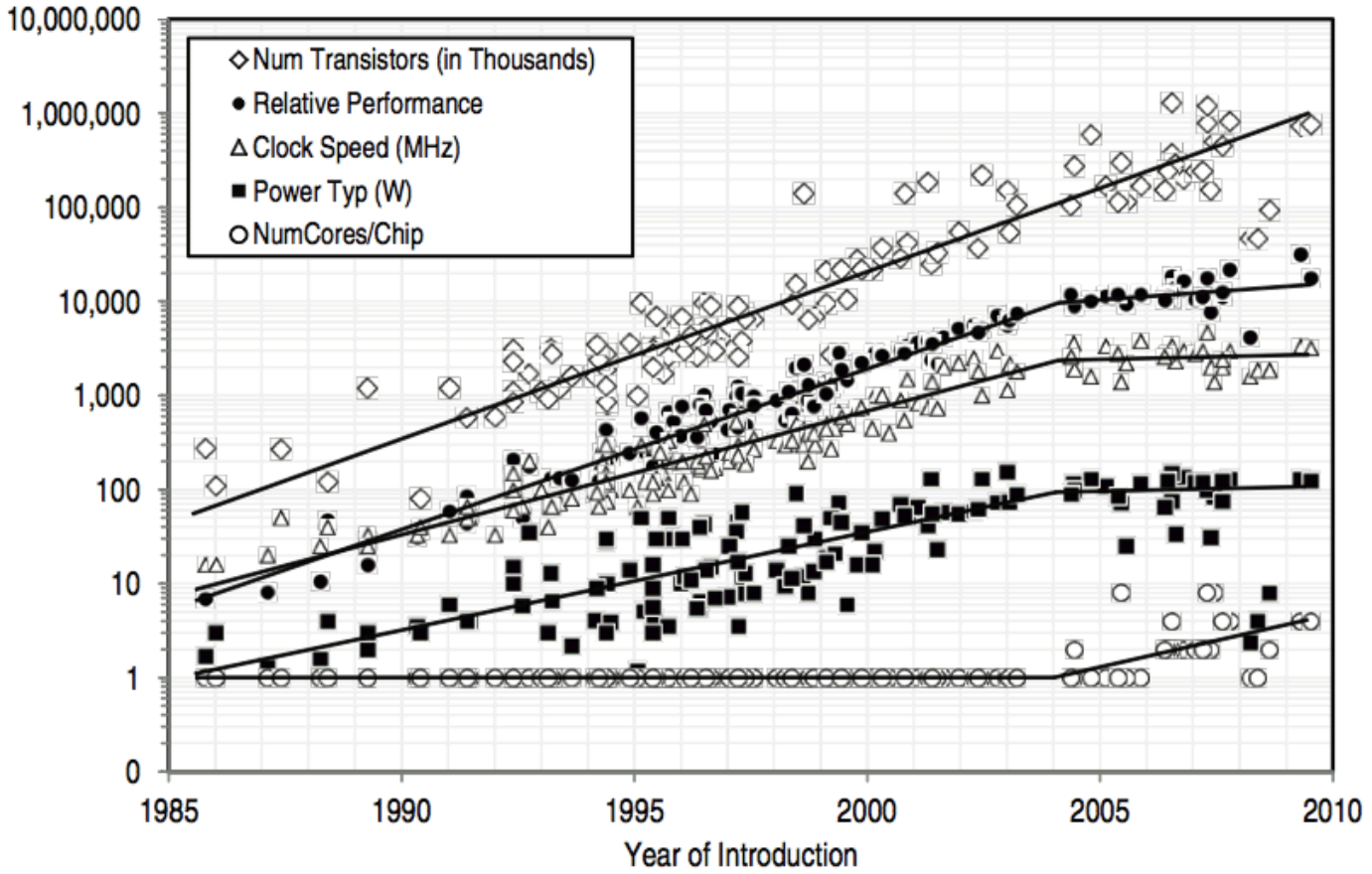


FIGURE 2.1 Transistors, frequency, power, performance, and cores over time (1985-2010). The vertical scale is logarithmic. Data curated by Mark Horowitz with input from Kunle Olukotun, Lance Hammond, Herb Sutter, Burton Smith, Chris Batten, and Krste Asanoviç.

The Future of Computing Performance: Game Over or Next Level, Samuel Fuller and Lynette Millet, Eds., National Academy Press, 2011

Laws of Physics will Halt Moore's Law

Sandia National Laboratories

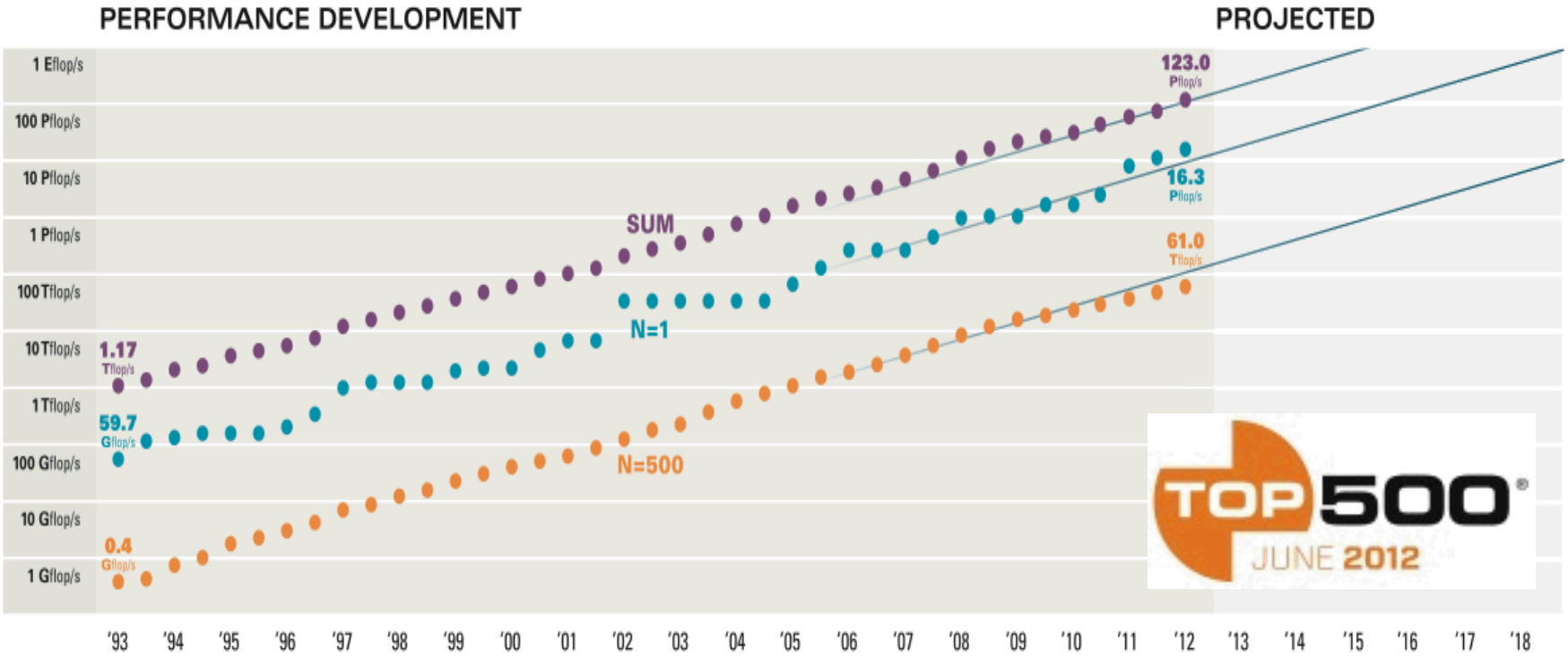
High-performance Logic Technology Requirements (ITRS 2011)

Year	2012	2013	2014	2015	2016	2017	2018	2019	2020
Gate Length (nm)	22	20	18	17	15.3	14	12.8	11.7	10.6
Equivalent Oxide Thickness	●	●	●	●	●	●	●	●	●
Source-Drain Leakage	●	●	●	●	●	●	●	●	●
Threshold Voltage	●	●	●	●	●	●	●	●	●
CV/I Intrinsic Delay	●	●	●	●	●	●	●	●	●
Total Gate Capacitance	●	●	●	●	●	●	●	●	●
Drive Current	●	●	●	●	●	●	●	●	●

- **Time line shown for best performing multi-gate transistor technology.**
- **Similar timelines exist for other functional components; *e.g.*, memory, RF logic.**

● technology available
● solutions known
● no known solutions

Top500: Linpack Performance

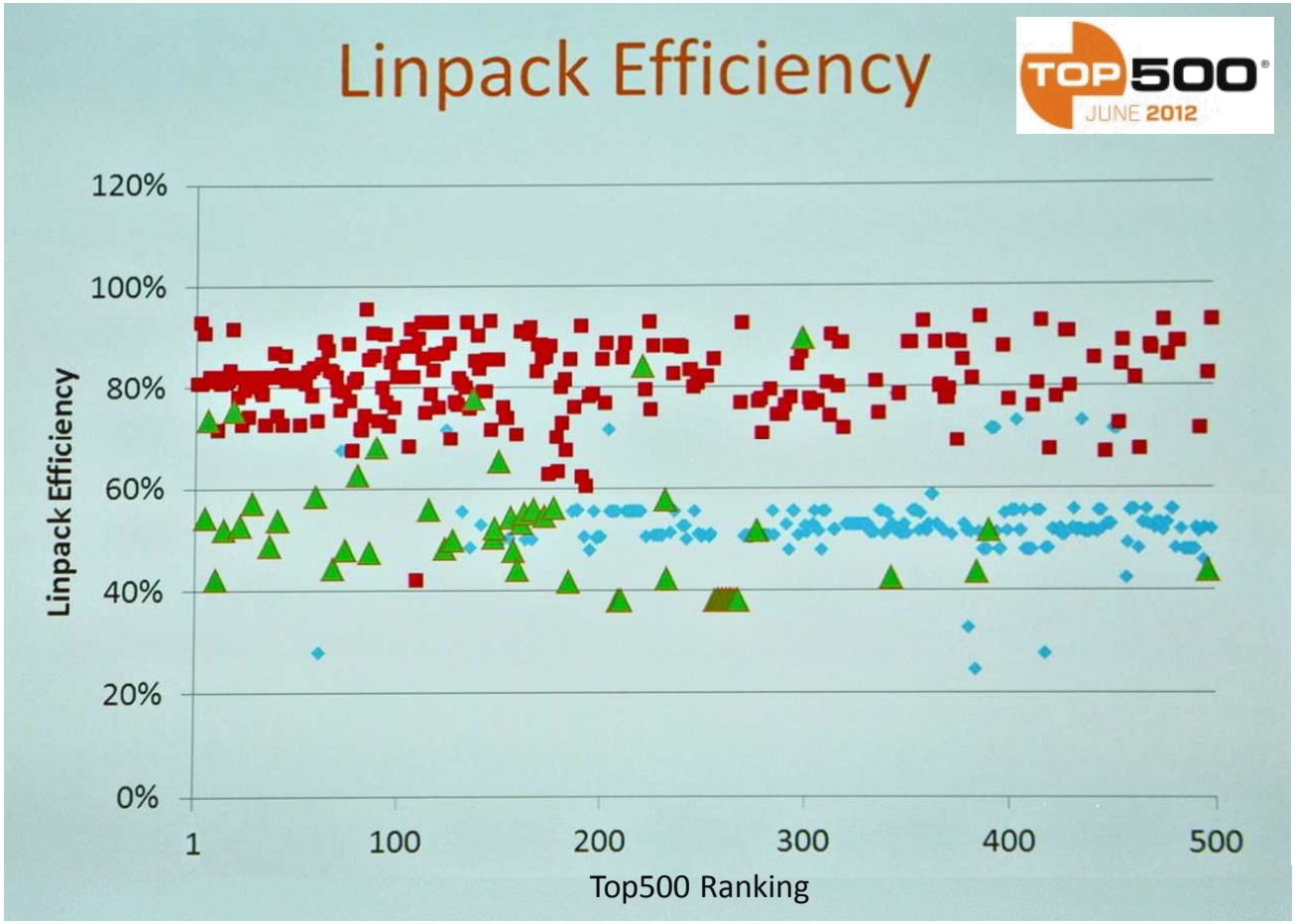


- **Industry will likely achieve ExaFlops in 2018-2020**
- **By analogy to " Moore's Law" the Top500 is an indication of what computer architects and system integrators, when organized properly, can do with integrated circuits**
- **The projection can become a self-fulfilling prophesy as resources and budgets are allocated to meet expectations**

Hardware Issues for 2020 ExaFlop System

Sandia National Laboratories

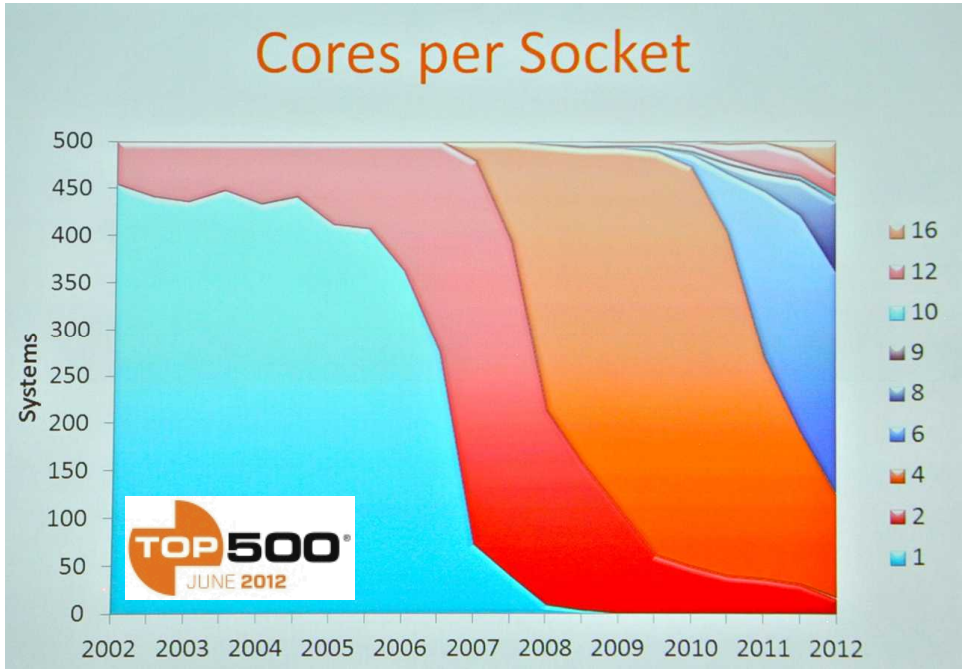
- **While Exaflop Linpack efficiency may be over 80% of peak, DOE mission applications often require orders of magnitude more memory bandwidth and capacity, and interconnection network performance**



Processor Trends

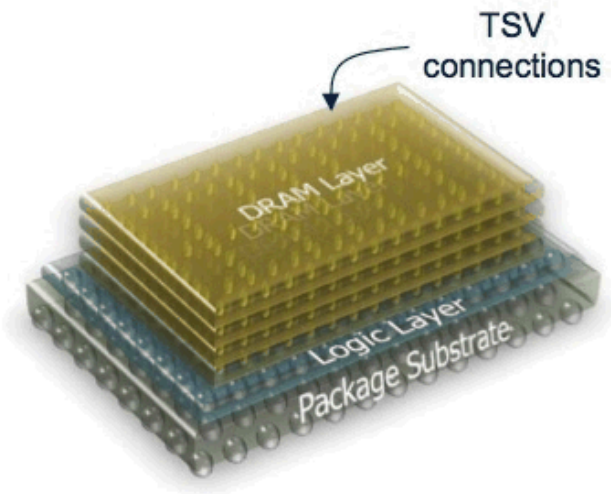
Sandia National Laboratories

- **Increases in Concurrency**
  - Scale of Multi-core
  - Scale of Many-core – SIMD wide-vector instructions
  - Number of threads per core
- **SoC Integration**
  - NIC or NIC/Router
  - Memory Controllers
  - Optical networking: Si-Photonics
- **Opportunities for fine-grained power monitoring and control**
  - Multiple power planes – Lessons from smartphones
  - Moore's Law may provide more transistors than we can afford to simultaneously power up – dark silicon
  - 3DI enables adaptive power management beyond processor



Memory Trends

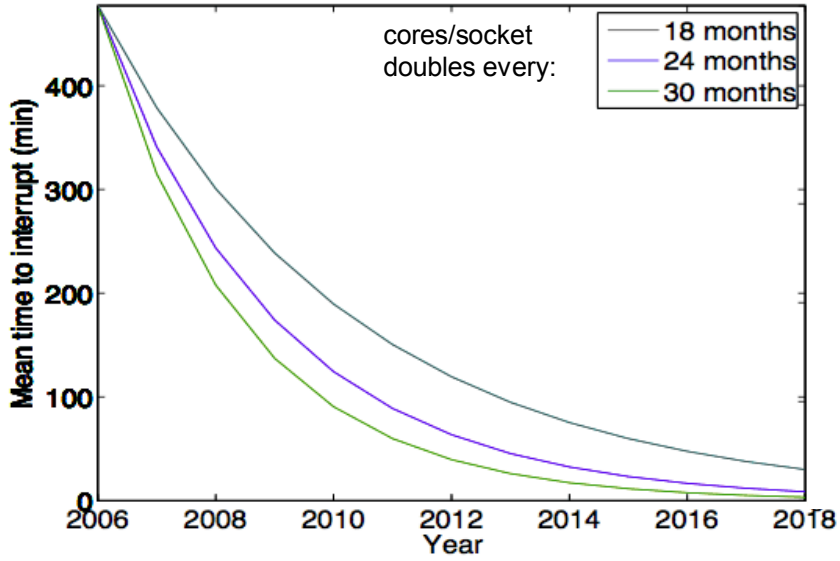
- **Memory Interface Technology (power efficiency & capacity)**
  - Power consumption of commodity JEDEC DDRn roadmap is infeasible for HPC
  - JEDEC Wide I/O Mobile with TSV for stacking on SoC (mobile) processors
  - Micron Technology - Hybrid Memory Cube (HMC) Consortium

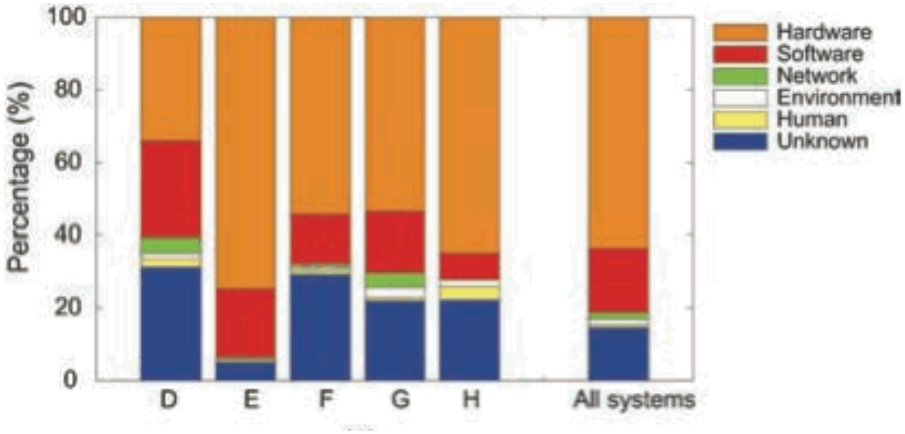


- **Industry R&D in NVRAM**
  - NAND/NOR FLASH, PCRAM, STT-RAM, ReRAM (Memristor), etc.
  - Primary focus is for storage applications
  - Research efforts in Industry to develop “Universal Memory” – performance of SRAM, cost of DRAM, non-volatility/lower cost of FLASH

Resilience Trends

- Most interrupts are attributed to hardware
- Top500 performance trends are expected to be sustained by increasing core count
- Resilience requirements may lead to more integrated SoC solutions, e.g. IBM's BG architecture

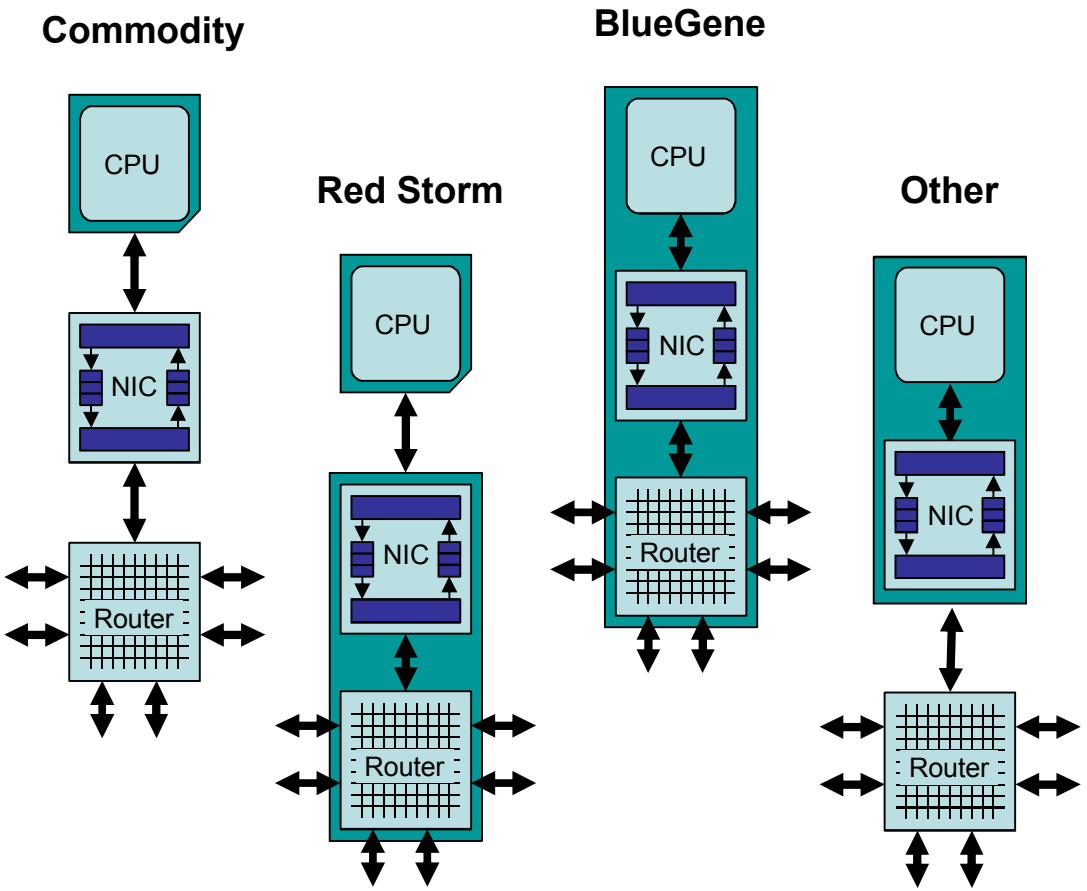




Ref: "Failure Tolerance in Petascale Computers", Schroeder and Gibson, 2007 CTWatch.

Interconnection Network Trends

- Bandwidth and latency performance
- Message injection rate
- Integration addresses Resilience and Energy efficient performance
- Optical Network Technology: Si Photonics



Industry co-designs for Linpack

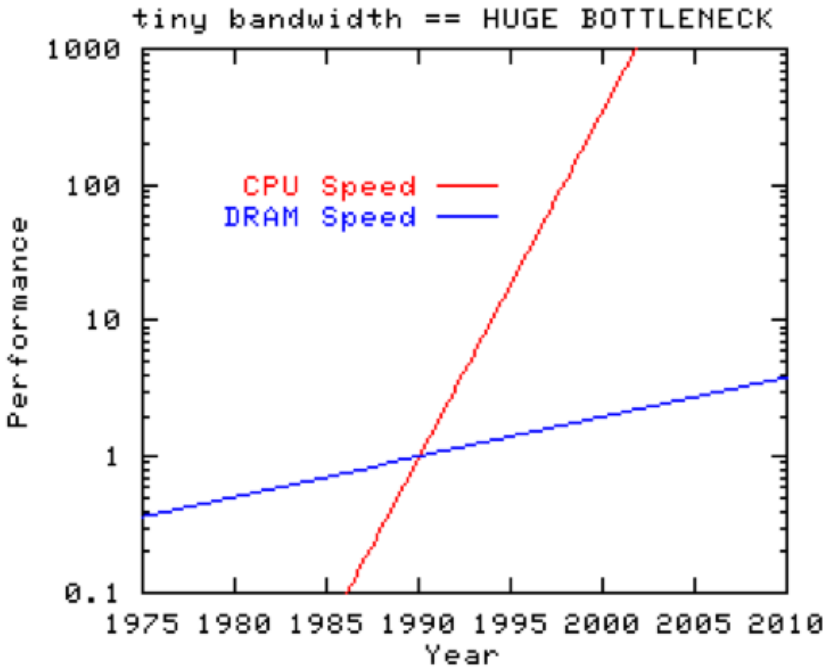
Sandia National Laboratories

- **A 2020 Linpack supercomputer will not be useful for DOE Mission Applications**
  - Limited memory capacity
  - Very poor data movement performance – both up the memory hierarchy, and across the interconnection network
  - Power requirement – well over 20 MW
- **We need to provide industry with better targets for our mission applications**

HPC is Primed for a Paradigm Change

- **Multi-core exacerbates the memory wall and data movement problem**
- **Co-design is an implicit statement that commodity processors need redesign for HPC**

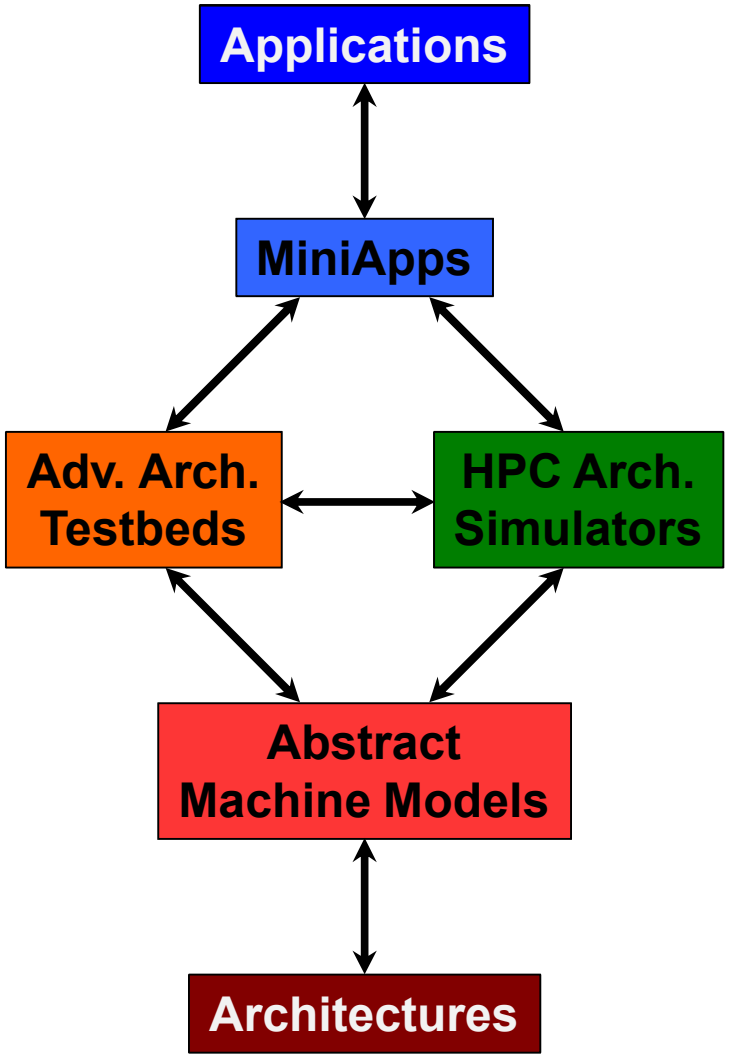
The Memory Wall



John McCalpin, STREAM Sustainable Memory Bandwidth in HPC
http://www.cs.virginia.edu/stream/

Our View of Co-design

- **Key Co-design Capabilities**
  - HPC Architectural Simulators – flexible to accommodate fidelity/speed tradeoffs
  - Development and evolution of Proxy Applications: miniApps
  - Advanced architecture testbeds
  - Definition, development and evolution of Abstract Machine Models (Proxy Architectures)
- **More details in talks by:**
  - Barrett – Proxy Applications & Advanced Architecture Testbeds
  - Rodrigues – HPC Architectural Simulators



A Critical Issue / Challenge:

Commodity adoption of capabilities for HPC

- **Business case for integration of HPC co-design innovations into commodity hardware designs**
- **DOE has an opportunity because Industry may welcome good ideas for how to use extra transistors**
  - Co-design can help define new HW/SW capabilities
  - PathForward Program can help DOE buy influence

Co-Design is an Optimization Problem

Sandia National Laboratories

- **The traditional goal –**
  ***Minimize* Time to Solution**
- **The new goal for Exascale HPC –**
  ***Balance* minimizing both Time and Energy to Solution**
- **Objective functions versus Constraints**
- **The Goal of Co-design is general purpose HPC:**
  Support the DOE portfolio of Mission applications
  - Not one-off special purpose computers

Hardware Recommendations

Sandia National Laboratories

- **Invest in Data movement capabilities to address the Memory Wall and improve Interconnection Network performance**
- **Buffer our application code base from the disruptive changes in computer architectures through R&D in advanced system software capabilities, e.g., adaptive runtime system software:**
  - Support for concurrent computation and communication
  - Fine-grained energy/power measurement and control
  - Real-time response to component failures
- **Develop a co-design constraint/objective function that factors preserving an evolutionary path for our application portfolio into the Co-design process**

Sandia National Laboratories

Questions?

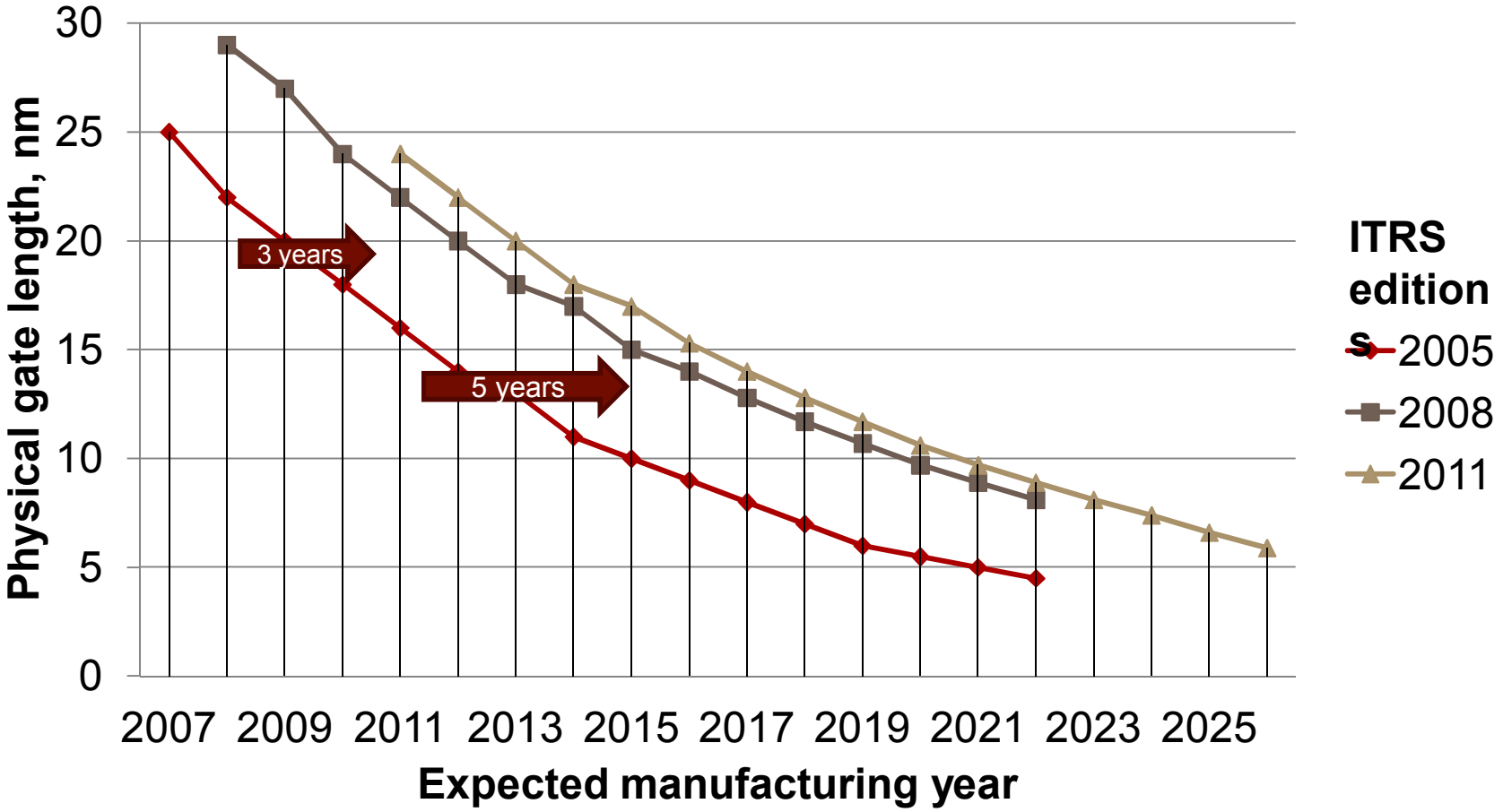
ITRS projections for physical gate lengths (nm) of high performance logic technology

Sandia National Laboratories

Report year / Projected year	2001	2003	2005	2007	2008	2009	2011
2007	25	25	25	25	x	x	x
2008		22	22	22	29	x	x
2009		20	20	20	27	29	x
2010	18	18	18	18	24	27	x
2011			16	16	22	24	24
2012		14	14	14	20	22	22
2013	13	13	13	13	18	20	20
2014			11	11	17	18	18
2015		10	10	10	15	17	17
2016	9	9	9	9	14	15.3	15.3
2017			8	8	12.8	14	14
2018		7	7	7	11.7	12.8	12.8
2019			6	6	10.7	11.7	11.7
2020			5	5.5	9.7	10.7	10.6

ITRS projections for gate lengths (nm) for 2005, 2008 and 2011 editions

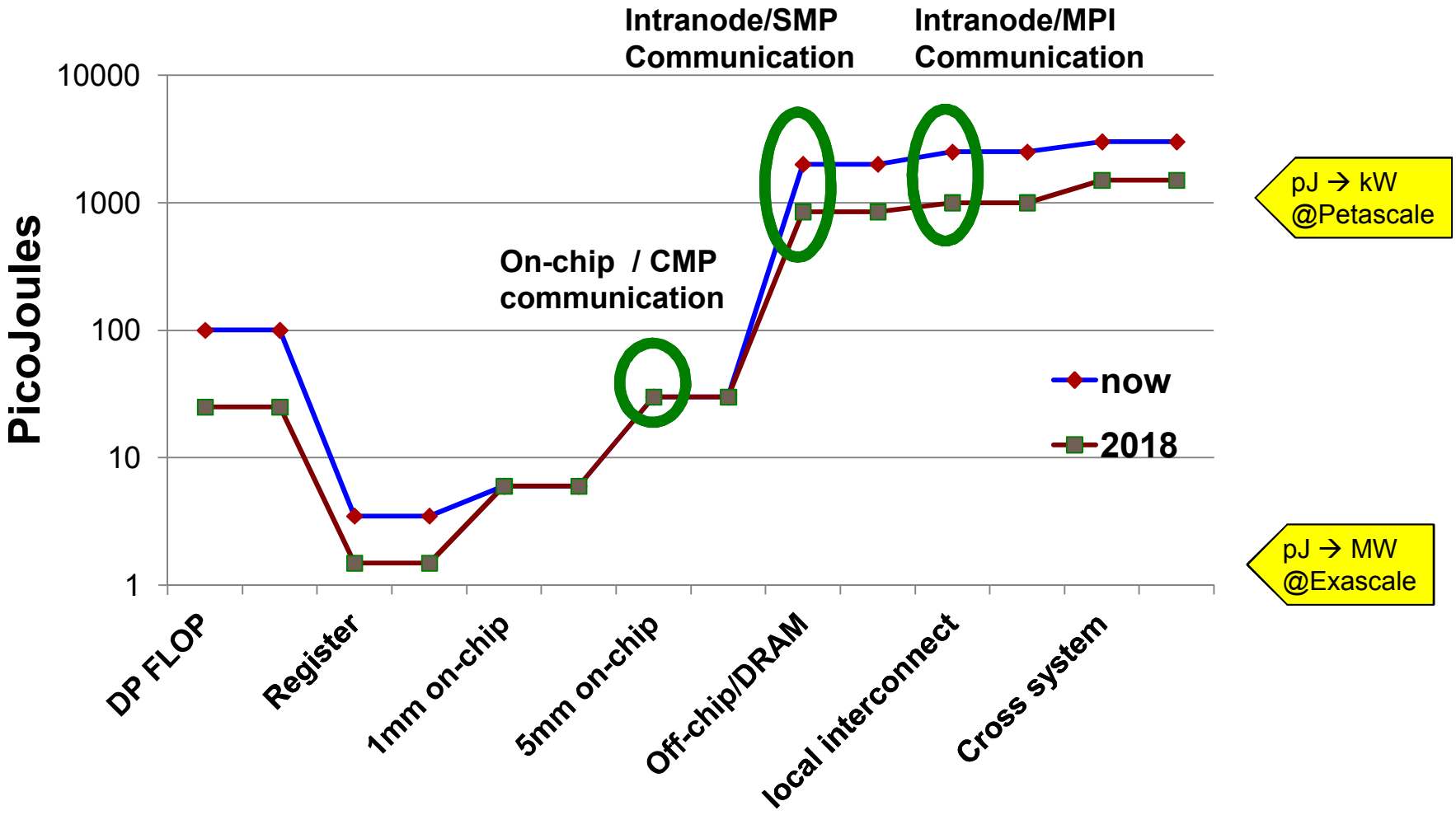
Sandia National Laboratories



Note the rapid 3- and then 5-year shifts in ITRS projections for physical gate lengths.

Memory Bandwidth & Data Movement Performance will be Energy limited

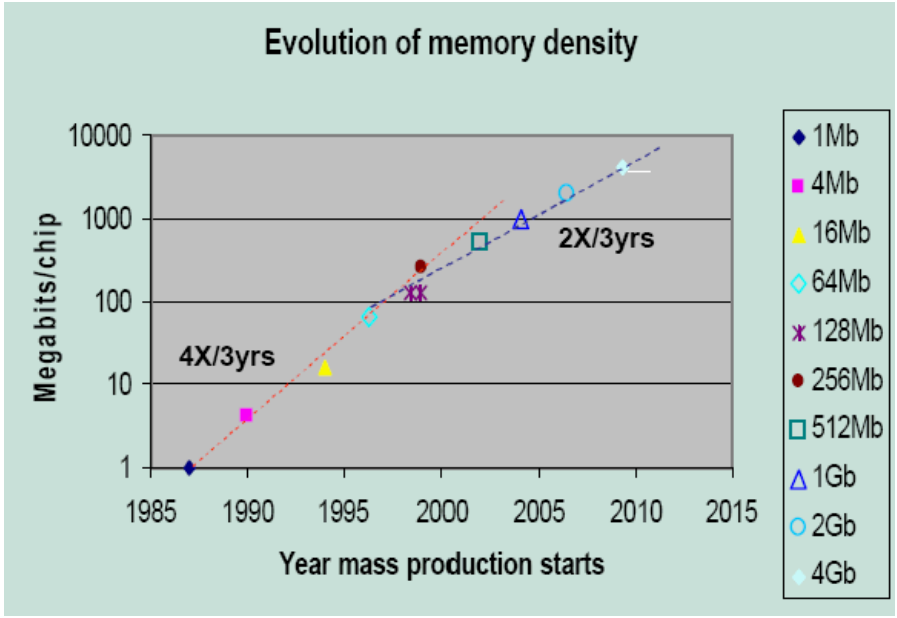
Sandia National Laboratories

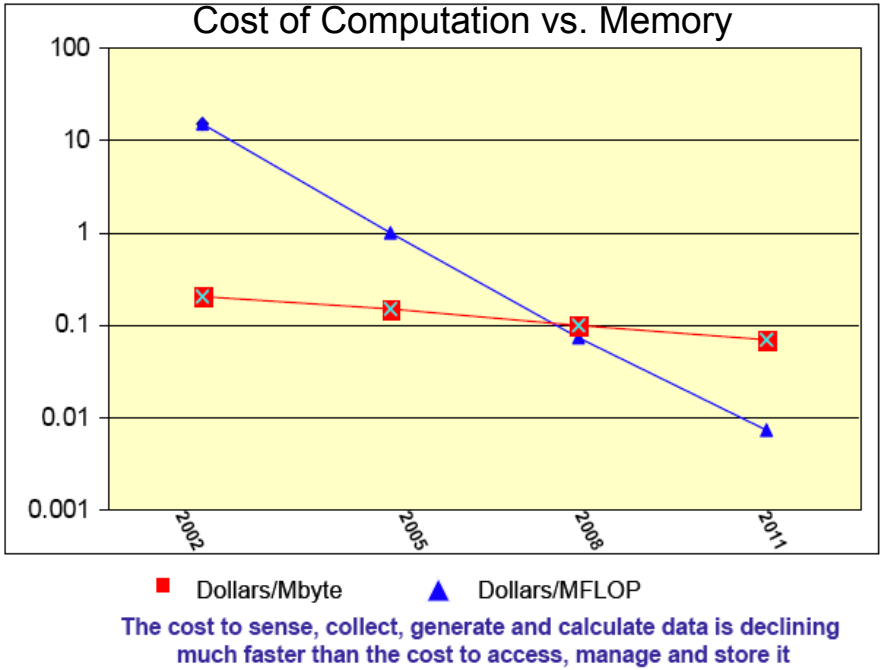


ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems, Peter Kogge, Editor and Study Lead, DARPA-TR-2008-13, 2008.

Projections of Memory Density Improvements

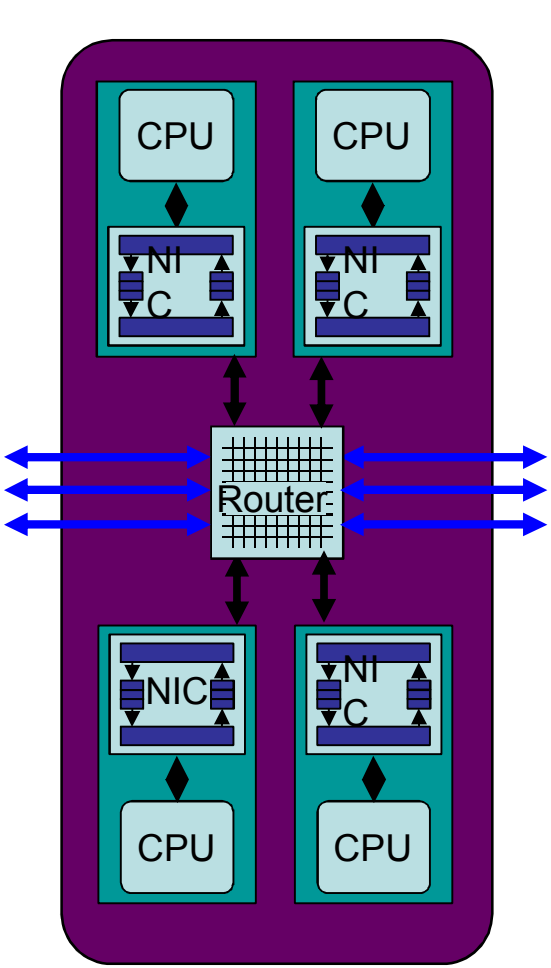
- Memory density is doubling every three years; processor logic is every two
  - Project 8Gigabit DIMMs in 2018
  - 16Gigabit if technology acceleration (or higher cost for early release)
- Storage costs (dollars/Mbyte) are dropping gradually compared to logic costs
  - Industry assumption: $1.80/memory chip is median commodity cost

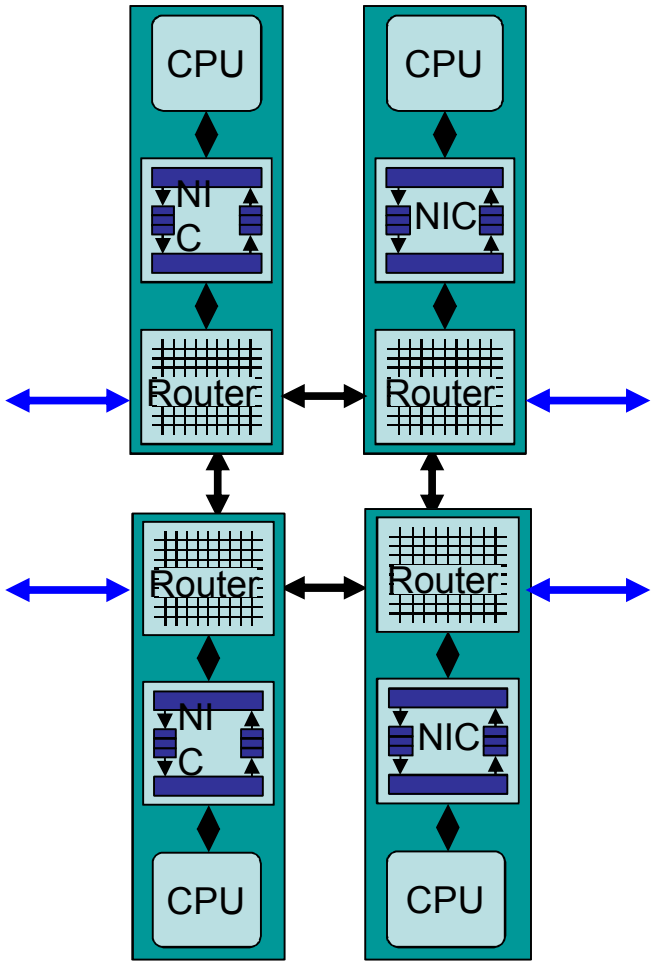




Source: David Turek, IBM

Other Interconnection Network Integration Possibilities

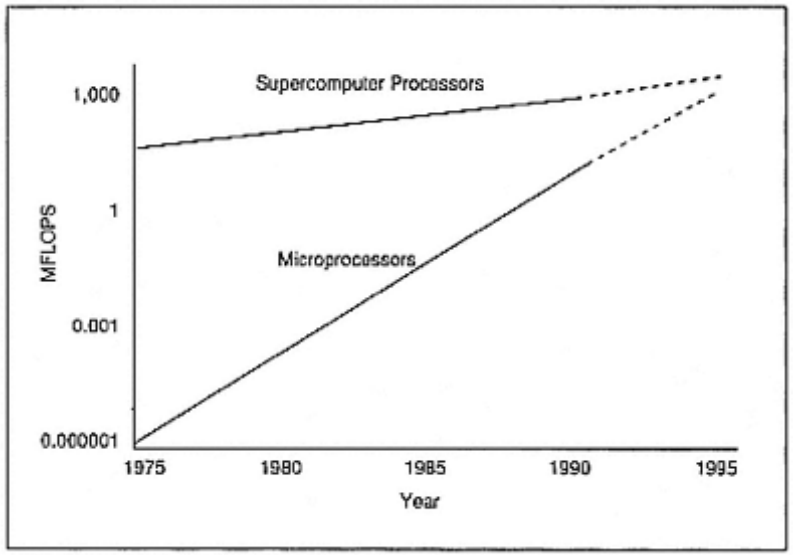




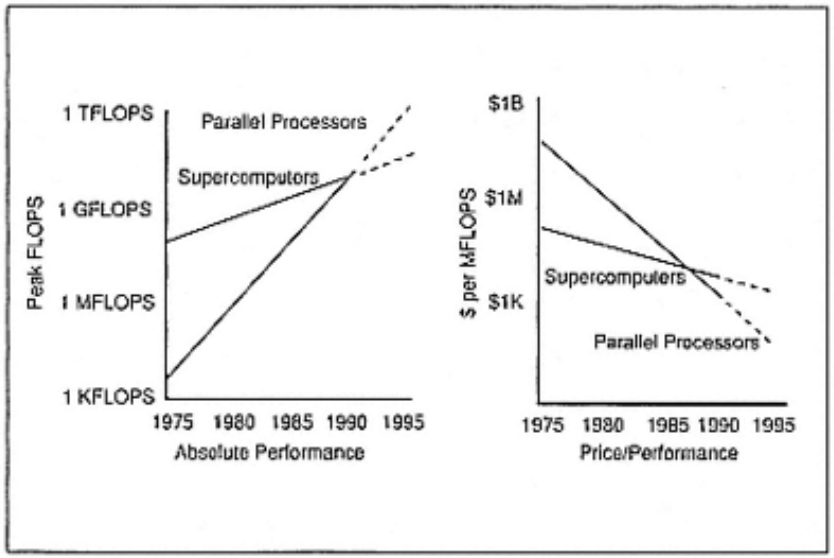
HPC Paradigm: Custom vs Commodity

Sandia National Laboratories

- **The last paradigm shift in HPC was the move from Vector to MPP supercomputers**
- ***The Attack of the Killer Micros* - Eugene Brooks, LLNL**



Absolute performance gains of microprocessors versus supercomputers

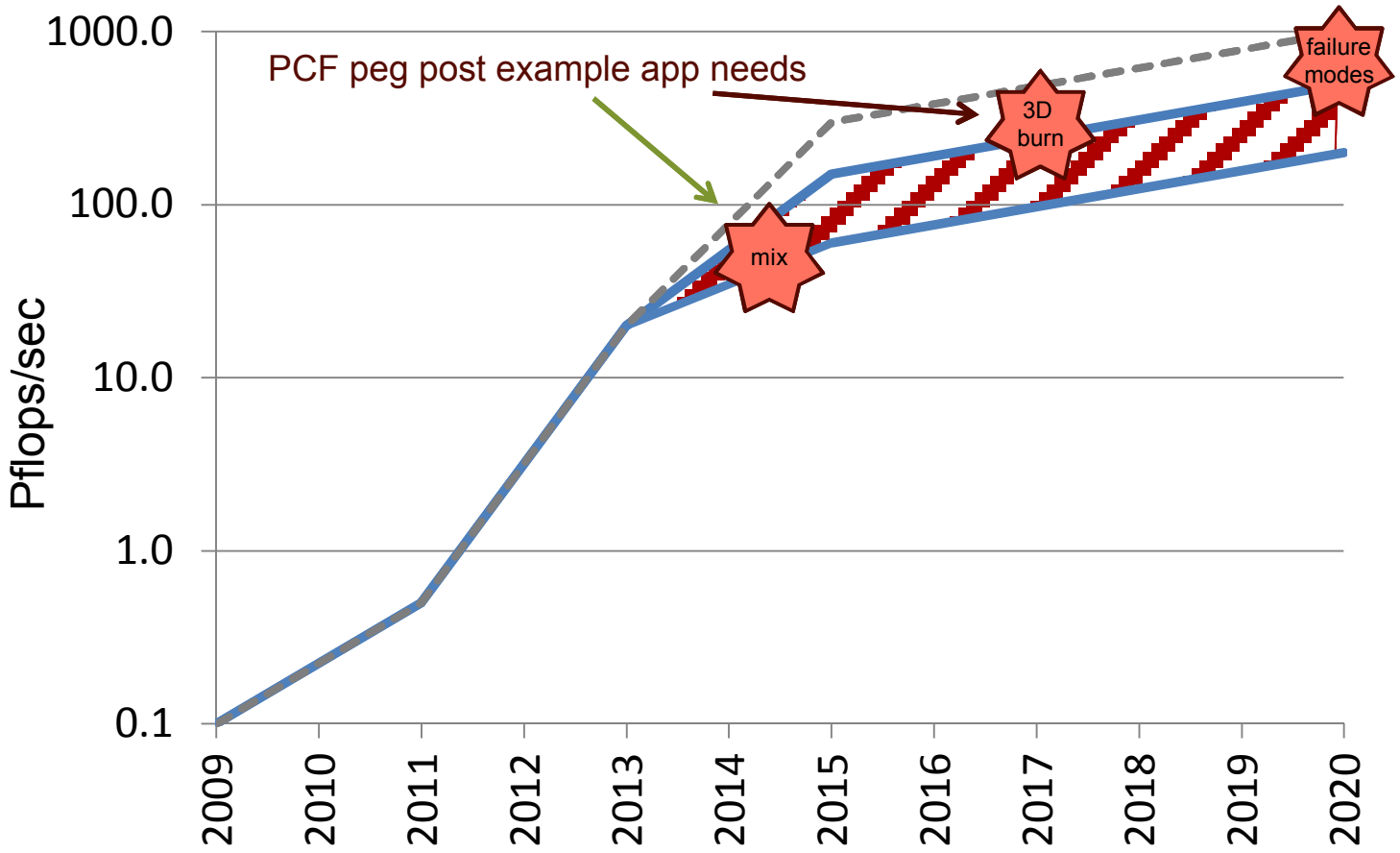


Improvements in parallel processors versus supercomputers

Attack of the Killer Micros, UC Press E-Books Collection, 1982-2004, http://publishing.cdlib.org/ucpressebooks/view docId=ft0f59n73z&chunk.id=d0e14780&toc.depth=1&toc.id=d0e14710&brand=eschol

Industry Responses to DOE Exascale RFI indicate the magnitude of the Data Movement Gap

Sandia National Laboratories



PCF Target performance (capability runs)

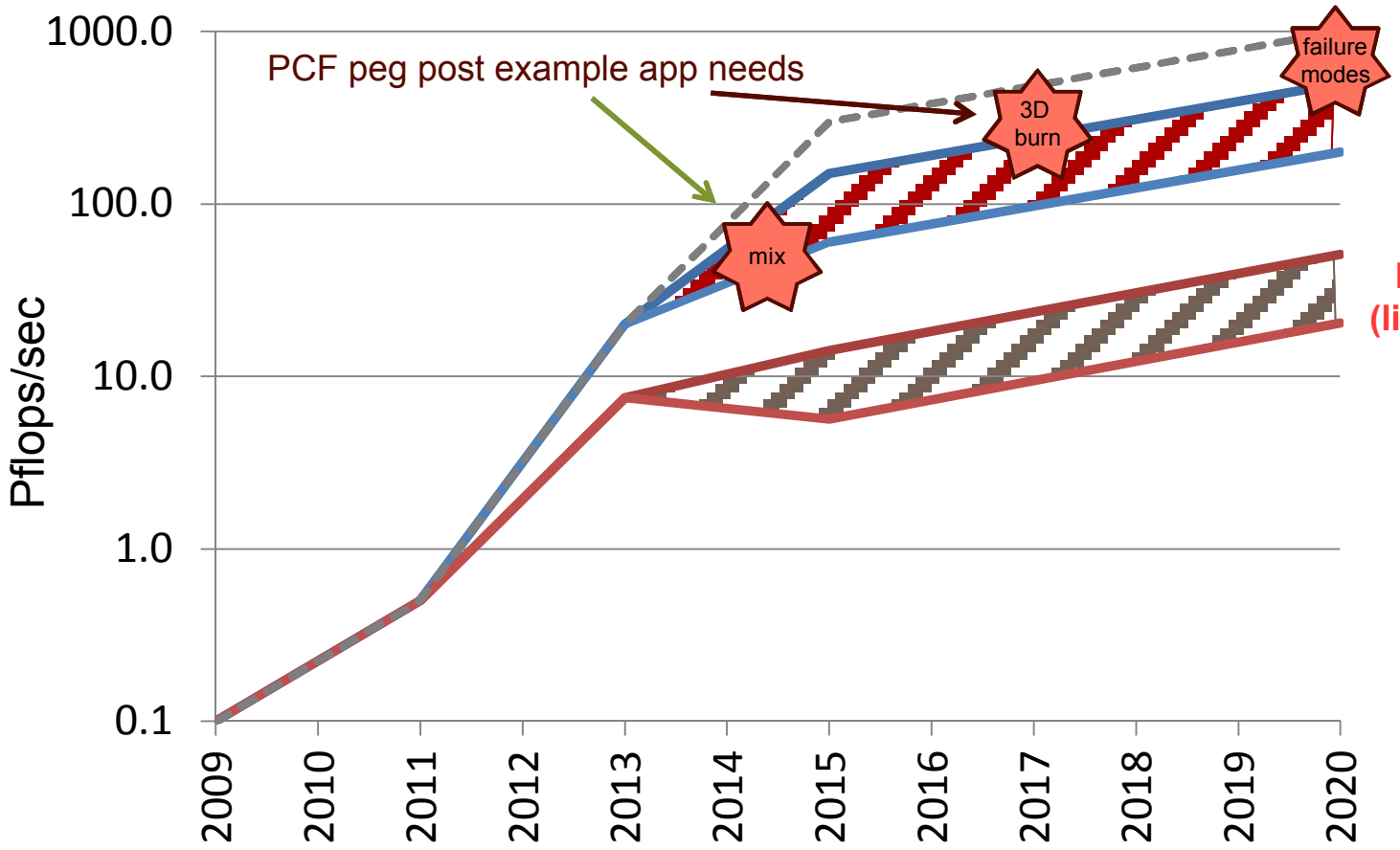
Vendor peak performance (no DOE investment)

Courtesy of Bert Still, LLNL

ASC apps require major work to avail fine-grained parallelism. Vendor roadmaps currently incur memory bandwidth and capacity limitations. Thus, effective utilization of machines remains largely flat (bottom curve), even with > 100x peak performance. ASC programmatic demands continue rising (Predictive Capability Framework (PCF) peg posts above)

Issue #1: Peak Performance requires ~250MB memory capacity per thread, can only use ~10% of threads

Sandia National Laboratories



PCF Target performance (capability runs)

Vendor peak performance (no DOE investment)

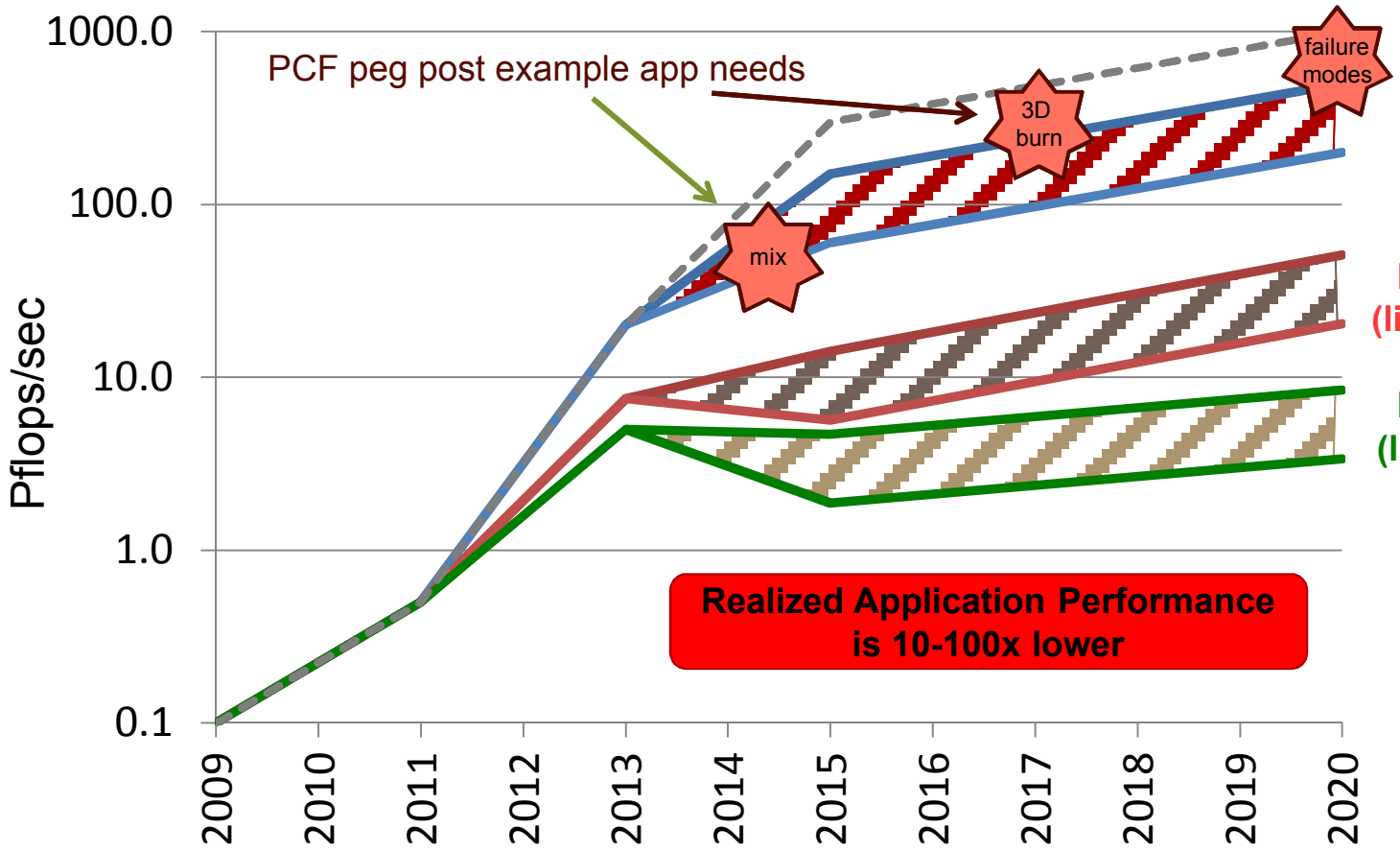
Peak performance (limited by mem size)

Courtesy of Bert Still, LLNL

ASC apps require major work to avail fine-grained parallelism. Vendor roadmaps currently incur memory bandwidth and capacity limitations. Thus, effective utilization of machines remains largely flat (bottom curve), even with > 100x peak performance. ASC programmatic demands continue rising (Predictive Capability Framework (PCF) peg posts above)

Issue #2: Peak Performance requires ~500MB/s memory BW per thread, can only use ~2-4% of threads

Sandia National Laboratories



PCF Target performance (capability runs)

Vendor peak performance (no DOE investment)

Peak performance (limited by mem size)

Peak performance (limited by mem size and bandwidth)

Courtesy of Bert Still, LLNL

ASC apps require major work to avail fine-grained parallelism. Vendor roadmaps currently incur memory bandwidth and capacity limitations. Thus, effective utilization of machines remains largely flat (bottom curve), even with > 100x peak performance. ASC programmatic demands continue rising (Predictive Capability Framework (PCF) peg posts above)