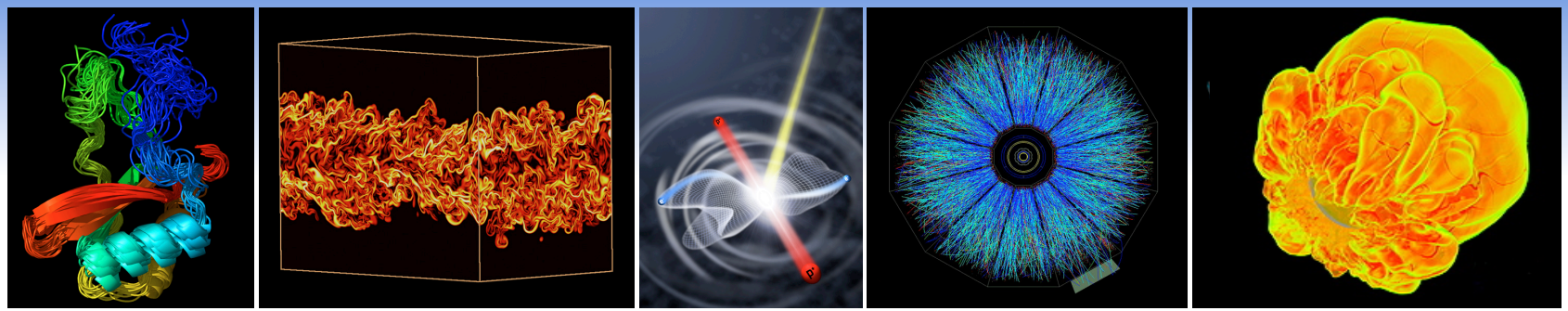


Exascale Computing Technology Challenges

John Shalf

National Energy Research Supercomputing Center

Lawrence Berkeley National Laboratory

Petascale Data Storage Workshop at SC2010

New Orleans, November 15, 2010

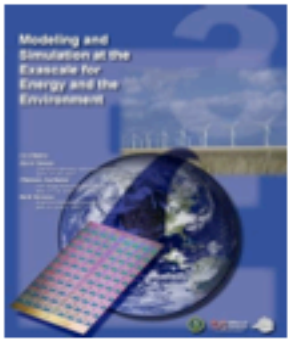
Office of Science
U.S. DEPARTMENT OF ENERGY

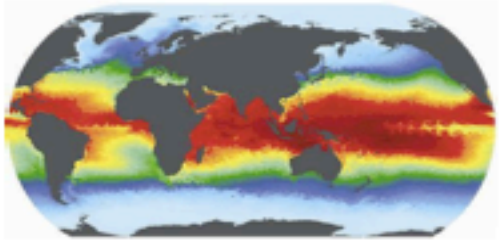
NERSC

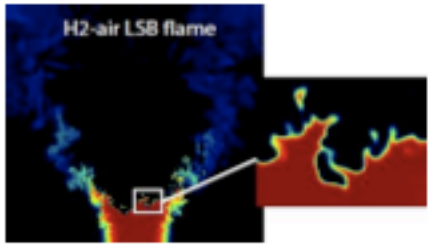
Berkeley Lab

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

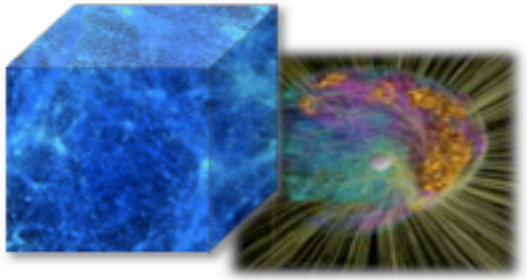
- Town Hall Meetings April-June 2007
- Scientific Grand Challenges Workshops Nov, 2008 – Oct, 2009
  - Climate Science (11/08),
  - High Energy Physics (12/08),
  - Nuclear Physics (1/09),
  - Fusion Energy (3/09),
  - Nuclear Energy (5/09),
  - Biology (8/09),
  - Material Science and Chemistry (8/09),
  - National Security (10/09)
- 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



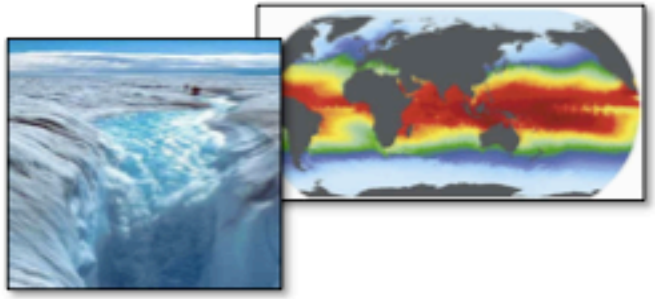
FUNDAMENTAL SCIENCE

Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

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 sources
- ***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.

Technology Disruptions on the Path to Exascale

- **Gigaflops to Teraflops was highly disruptive**
  - Moved from vector machines to MPPs with message passing
  - Required new algorithms and software
- **Teraflops to Petaflops was *not* very disruptive**
  - Continued with MPI+Fortran/C/C++ with incremental advances
- **Petaflops to Exaflops will be highly disruptive**
  - No clock increases → hundreds of simple "cores" per chip
  - Less memory and bandwidth → cores are not MPI engines
  - x86 too energy intensive → more technology diversity (GPUs/ accel.)
  - Programmer controlled memory hierarchies likely
- **Computing at every scale will be *transformed (not just exascale)***

Exascale Architecture Constraints

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)	

Exascale Initiative Steering Committee
(circa December 9, 2009)

Systems	2009	2015 +1/-0	2018 +1/-0
System peak	**2 Peta**	**100-300 Peta**	**1 Exa**
Power	**6 MW**	**~15 MW**	**~20 MW**
System memory	0.3 PB	5 PB	64 PB (+)
Node performance	125 GF	0.5 TF or 7 TF	1-2 or 10TF
Node memory BW	25 GB/s	1-2TB/s	2-4TB/s
Node concur	12	O(100)	O(1k) or 10k
Total Node Interco	3.5		00GB/s 4 or 1:8 from memory
System size (no			00) or O(1M)
Total concurre			O 0) for latency hiding
Storage	15 P	150 P	00 PB (>10x system memory is min)
IO	0.2 TB	10 TB/s	60 TB/s (how long to drain the machine)
MTTI	days	O(1day)	O(1 day)

60 MW over budget

OOOPs!

Systems	2009	2015 +1/-0	2018 +1/-0
System peak	**2 Peta**	**100-300 Peta**	**1 Exa**
Power	**6 MW**	**~15 MW**	**~20 MW**
System memory	0.3 PB	5 PB	64 PB (+)
Node performance	125 GF	0.5 TF or 7 TF	2 TF or 10TF
Node memory BW	25 GB/s	0.2TB/s or 0.5TB/s	0.4TB/s or 1TB/s
Node concurrency	12	O(100)	O(1k) or 10k
Total Node Interconnect BW	3.5 GB/s	100-200 GB/s 10:1 vs memory bandwidth 2:1 alternative	200-400GB/s (1:4 or 1:8 from memory BW)
System size (nodes)	18,700	50,000 or 500,000	O(100,000) or O(1M)
Total concurrency	225,000	O(100,000,000) *O(10)-O(50) to hide latency	O(billion) * O(10) to O (100) for latency hiding
Storage	15 PB	150 PB	500-1000 PB (>10x system memory is min)
IO	0.2 TB	10 TB/s	60 TB/s (how long to drain the machine)
MTTI	days	O(1day)	O(1 day)

The REAL Exascale Constraints

First Generation

- 300PF
- 15MW
- $200M
- Deliver by 2015

Second Generation

- 1 Exaflop
- 20MW
- $200M
- Deliver by 2018

Do not get caught up in the tyranny of the spreadsheet!

A Revolution is Underway

- **Rapidly Changing Technology Landscape**
  - **Evolutionary change between nodes** *(10x more explicit parallelism)*
  - **Revolutionary change within node** *(100x more parallelism, with diminished memory capacity and bandwidth)*
  - **Multiple Technology Paths** *(GPU, manycore/embedded, x86/PowerX)*

- **The technology disruption will be pervasive *(not just exascale)***
  - ***Assumptions that our current software infrastructure is built upon are no longer valid***
  - **Applications, Algorithms, System Software *will all break***
  - **As significant as migration from vector to MPP (early 90's)**

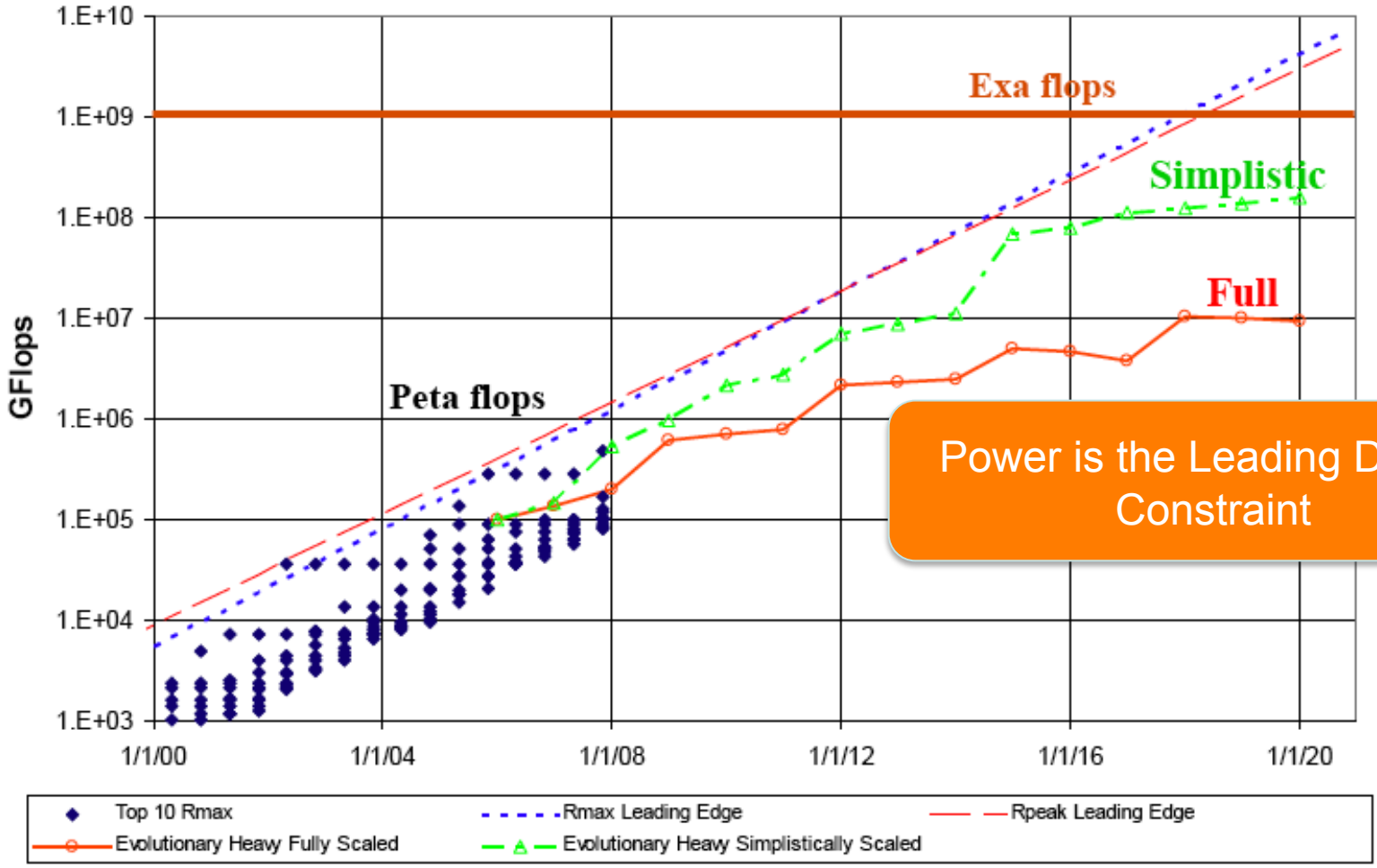
- **Need a new approach to ensuring continued application performance improvements**
  - **This isn't just about Exaflops – this is for all system scales**

NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

Part I

Power Crisis in HPC

Current Technology Roadmaps will Depart from Historical Gains



Power is the Leading Design Constraint

From Peter Kogge, DARPA Exascale Study

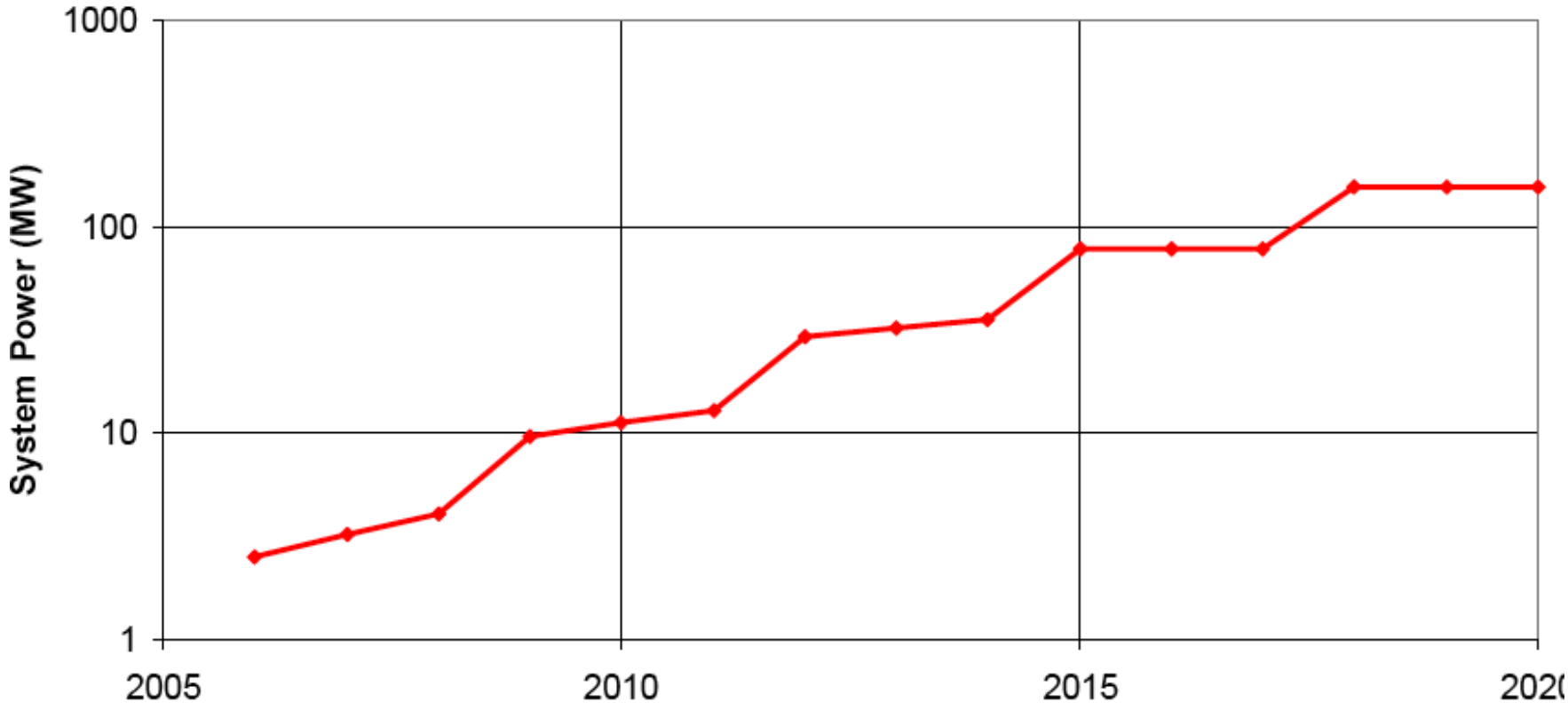
NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

Office of Science U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

… and the power costs will still be staggering



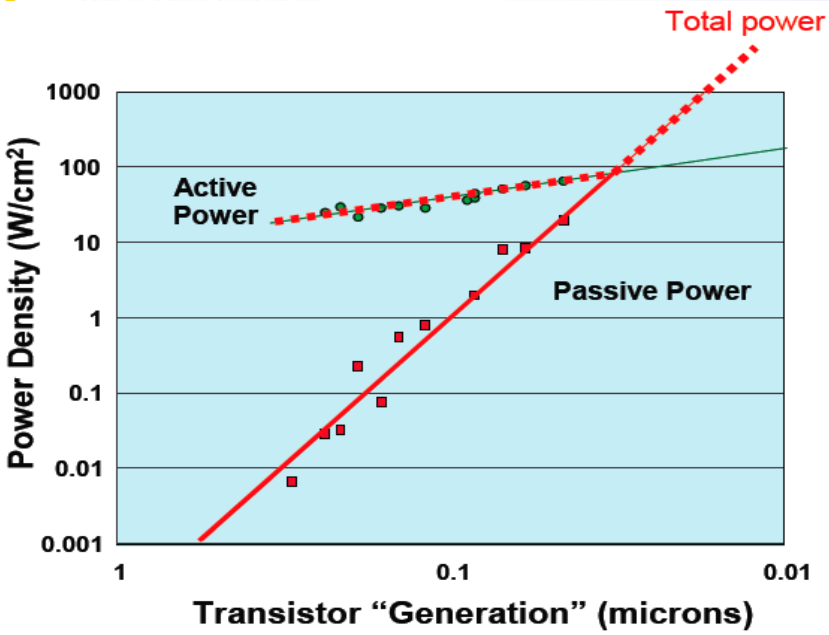
From Peter Kogge,
DARPA Exascale Study

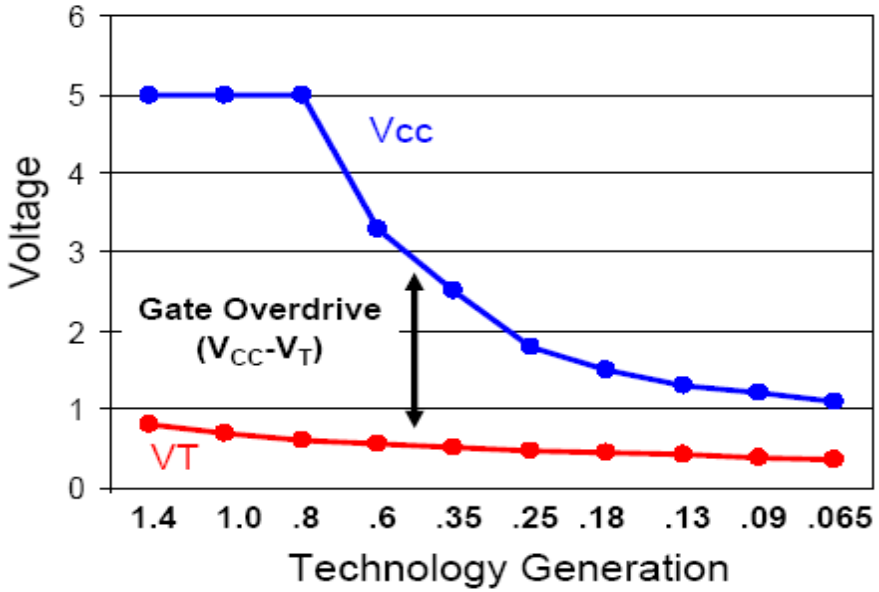
$1M per megawatt per year! (with CHEAP power)

Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

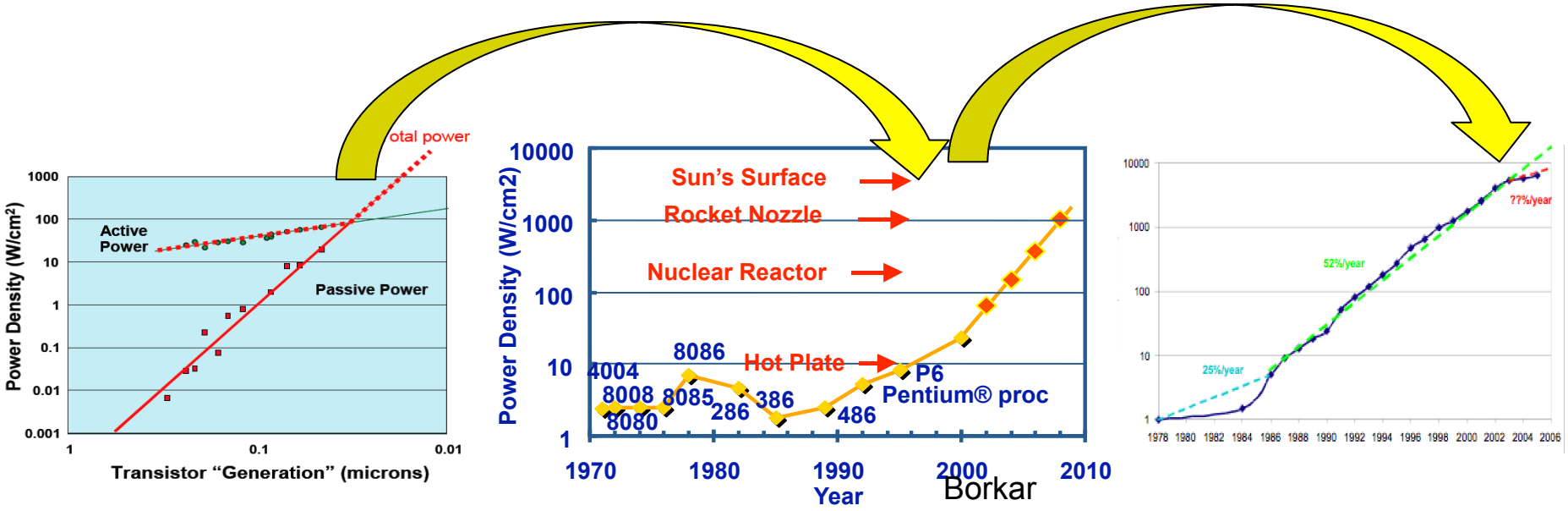
Primary Design Constraint: *POWER*





- **Total Energy = Active Power + Leakage Power**
- **Active Power = C * V^2 * F**
  - **This is energy required to charge & discharge capacitance of transistor**
  - **Dennard recognized capacitance is reduced proportional to die shrink**
  - **Power neutral if you drop supply voltage and increase clock frequency**
- **Leakage Power = V * $I_{leakage}$**
  - **Voltage is so low that cannot turn transistor entirely on or off**
  - **So transistors must either "leak" or run much slower**

Primary Design Constraint: *POWER*



- **No room for Dennard scaling (leakage power caught up to us)**
- **Without changes, we will get exponential growth in power**
- ***So, clock frequencies stalled in 2002 (Patterson Graph)***

NERSC
NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

The Challenge

Where do we get a 1000x improvement in performance with only a 10x increase in power?

How do you achieve this in 10 years with a finite development budget?

Where do we get 1000x performance improvement for 10x power?

1. Processors
2. On-chip data movement
3. System-wide data movement
4. Memory Technology
5. Resilience Mechanisms
6. Exascale Data Storage (EDSW?)

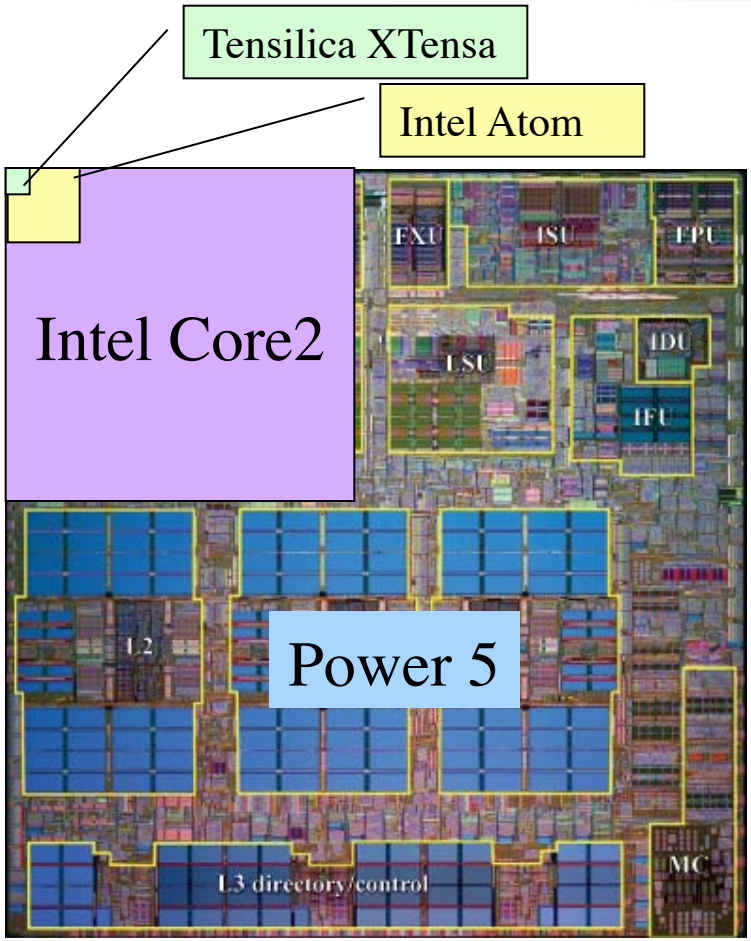
Processors: What are the problems?

(Lessons from the Berkeley View)

- **Current Hardware/Lithography Constraints**
  - **Power limits leading edge chip designs**
    - Intel Tejas Pentium 4 cancelled due to power issues
  - **Yield on leading edge processes dropping dramatically**
    - IBM quotes yields of 10 – 20% on 8-processor Cell
  - **Design/validation leading edge chip is becoming unmanageable**
    - Verification teams > design teams on leading edge processors

- **Solution: Small Is Beautiful**
  - **Simpler (5- to 9-stage pipelined) CPU cores**
    - Small cores not much slower than large cores
  - **Parallel is energy efficient path to performance:CV^2F**
    - Lower threshold and supply voltages lowers energy per op
  - **Redundant processors can improve chip yield**
    - Cisco Metro 188 CPUs + 4 spares; Sun Niagara sells 6 or 8 CPUs
  - **Small, regular processing elements easier to verify**

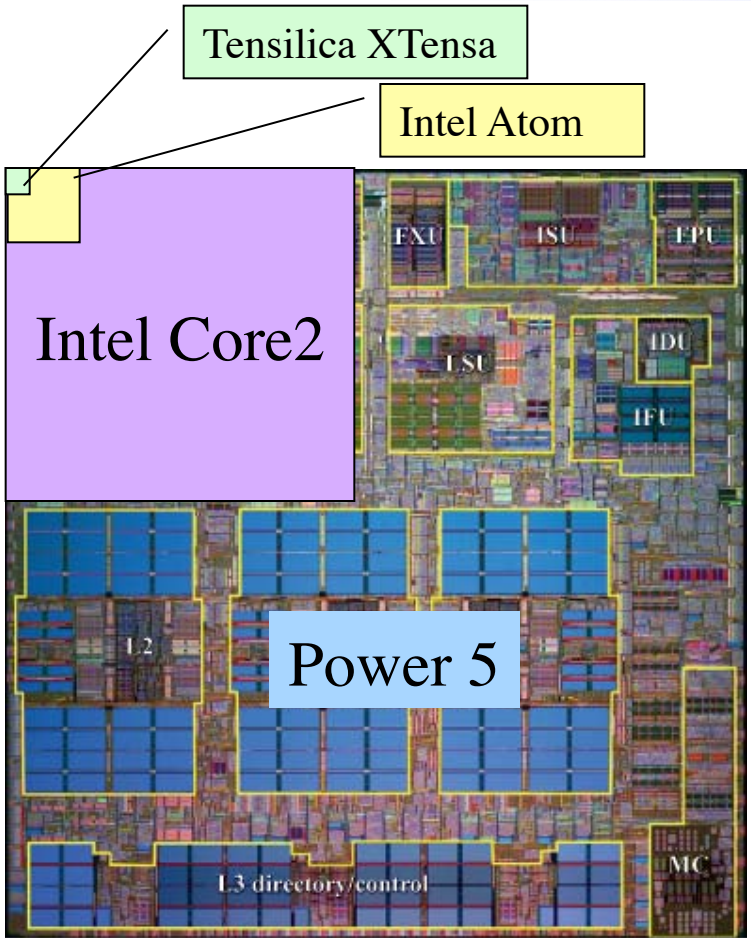
Low-Power Design Principles



- **Cubic power improvement with lower clock rate due to V^2F**
- **Slower clock rates enable use of simpler cores**
- **Simpler cores use less area (lower leakage) and reduce cost**
- **Tailor design to application to REDUCE WASTE**

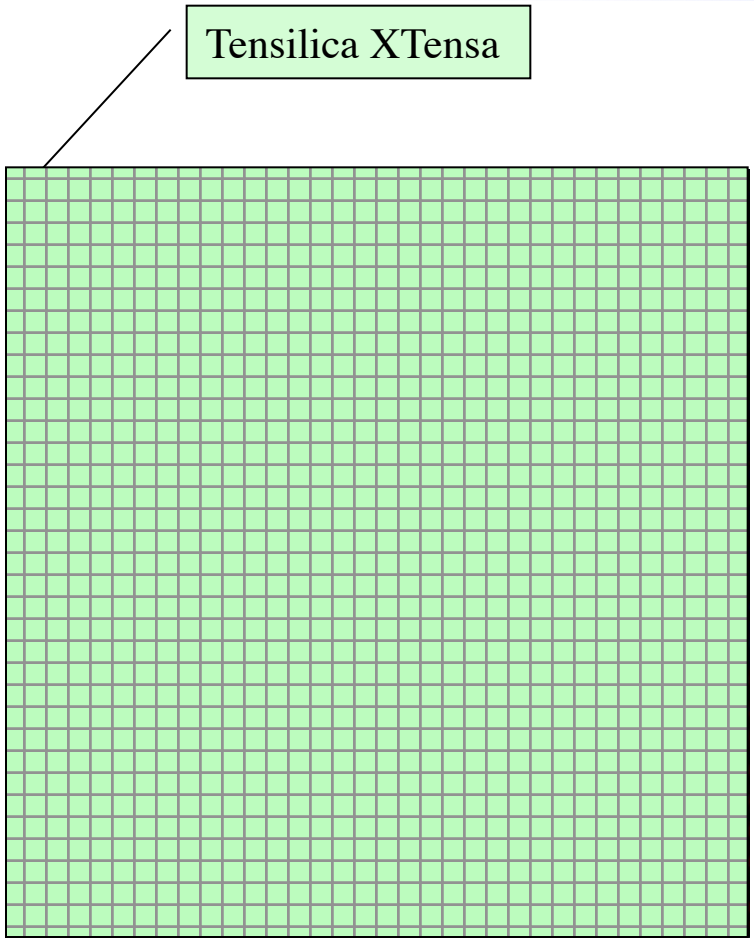
This is how iPhones and MP3 players are designed to maximize battery life and minimize cost

Low-Power Design Principles



- Power5 (server)
  - 120W@1900MHz
  - Baseline
- Intel Core2 sc (laptop) :
  - 15W@1000MHz
  - *4x more FLOPs/watt than baseline*
- Intel Atom (handhelds)
  - 0.625W@800MHz
  - 80x more
- Tensilica XTensa DP (Moto Razor) :
  - 0.09W@600MHz
  - 400x more *(80x-120x sustained)*

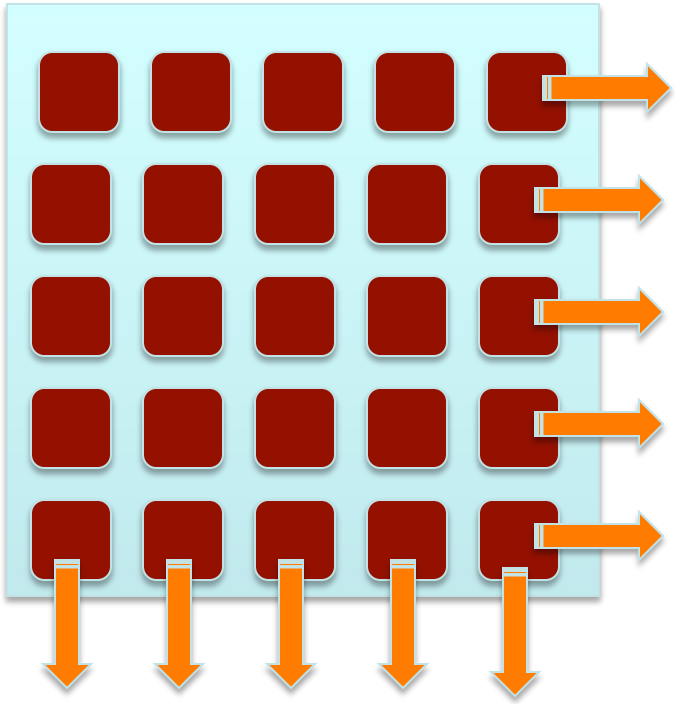
Low Power Design Principles



- Power5 (server)
  - 120W@1900MHz
  - Baseline
- Intel Core2 sc (laptop) :
  - 15W@1000MHz
  - *4x more FLOPs/watt than baseline*
- Intel Atom (handhelds)
  - 0.625W@800MHz
  - 80x more
- Tensilica XTensa DP (Moto Razor) :
  - 0.09W@600MHz
  - 400x more (80x-100x sustained)

Even if each simple core is 1/4th as computationally efficient as complex core, you can fit hundreds of them on a single chip and still be 100x more power efficient.

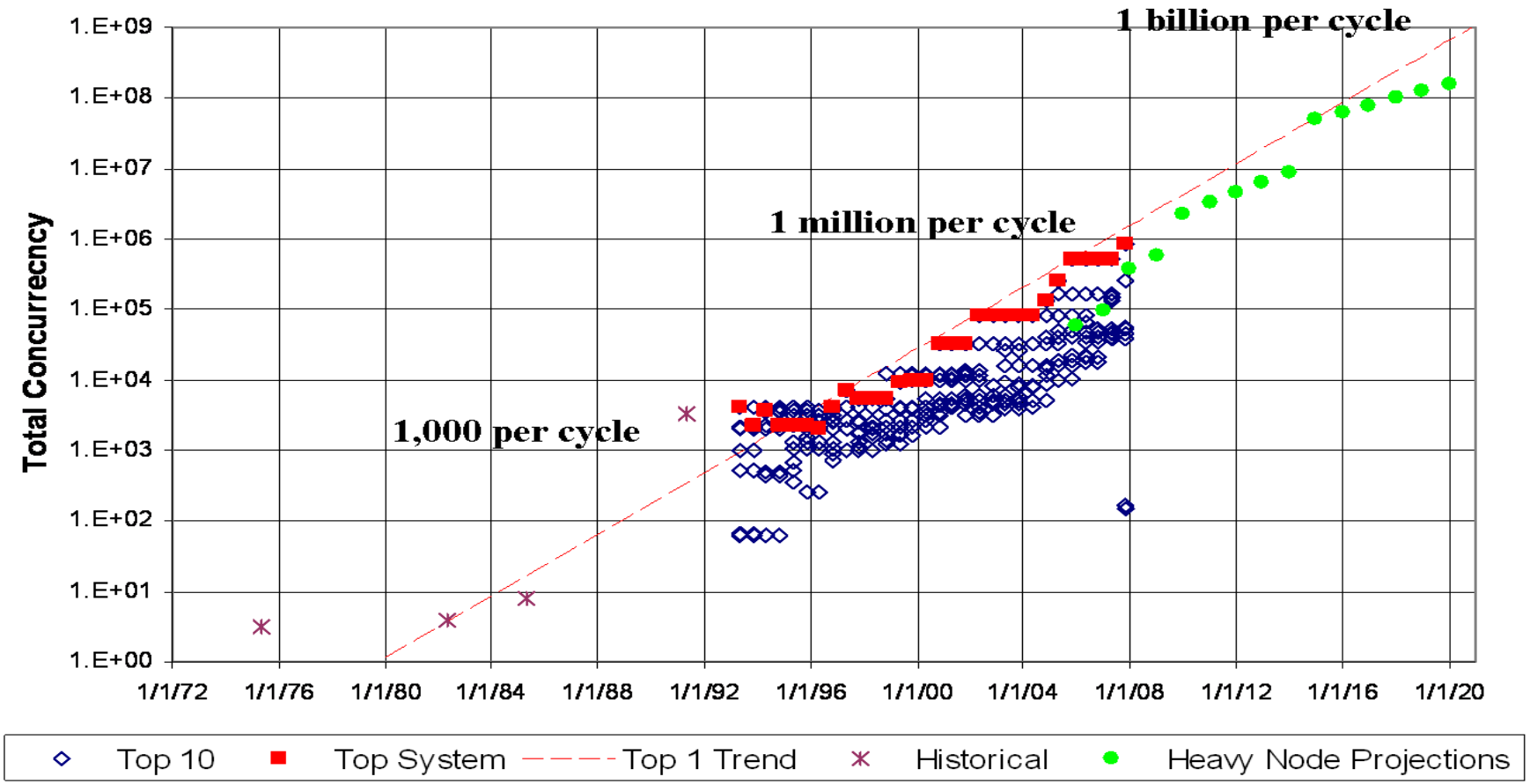
Future of On-Chip Architecture
(San Diego Meeting)



Scale-out for Planar geometry

- **~1000-10k simple cores /Chip**
  - 4-8 wide SIMD or VLIW bundles
  - Either 4 or 50+ HW threads
- **On-chip communication Fabric**
  - Low-degree topology for on-chip communication (torus or mesh)
  - *Scale cache coherence?*
  - Global (nonCC memory)
  - Shared register file (clusters)
- **Off-chip communication fabric**
  - Integrated directly on an SoC
  - Reduced component counts
  - Coherent with TLB (no pinning)

Projected Parallelism for Exascale



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

Need 1Million-way parallelism to reach an Exaflop . . .

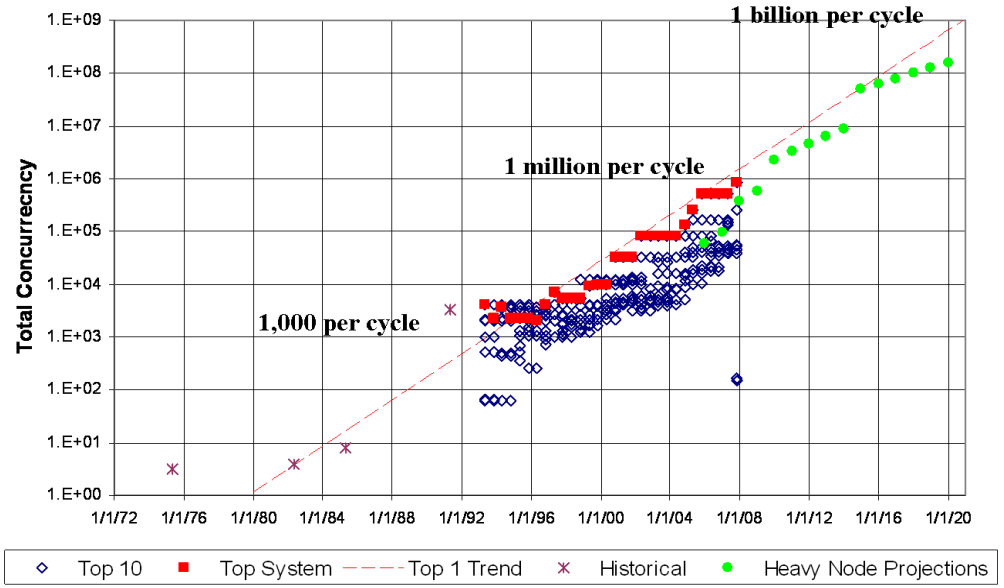
. . . And possibly another 100x just to hide latency

Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

Conclusion: Solving Logic Power Drives Move to Massive Parallelism

- **Future HPC must move to simpler power-efficient core designs**
  - Embedded/consumer electronics technology is central to the future of HPC
  - Convergence inevitable because it optimizes both cost and power efficiency



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

- **Consequence is massive on-chip parallelism**
  - A thousand cores on a chip by 2018
  - 1 Million to 1 Billion-way System Level Parallelism
  - *Must express massive parallelism in algorithms and pmodels*
  - *Must manage massive parallelism in system software*

NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

The Cost of Data Movement

How do those cores talk to each other?

The problem with Wires:

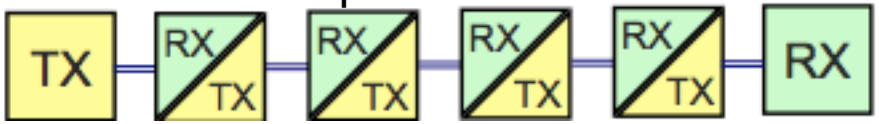
Energy to move data proportional to distance

- **Cost to move a bit on copper wire:**
  - **Power = bitrate * Length² / cross-section area**

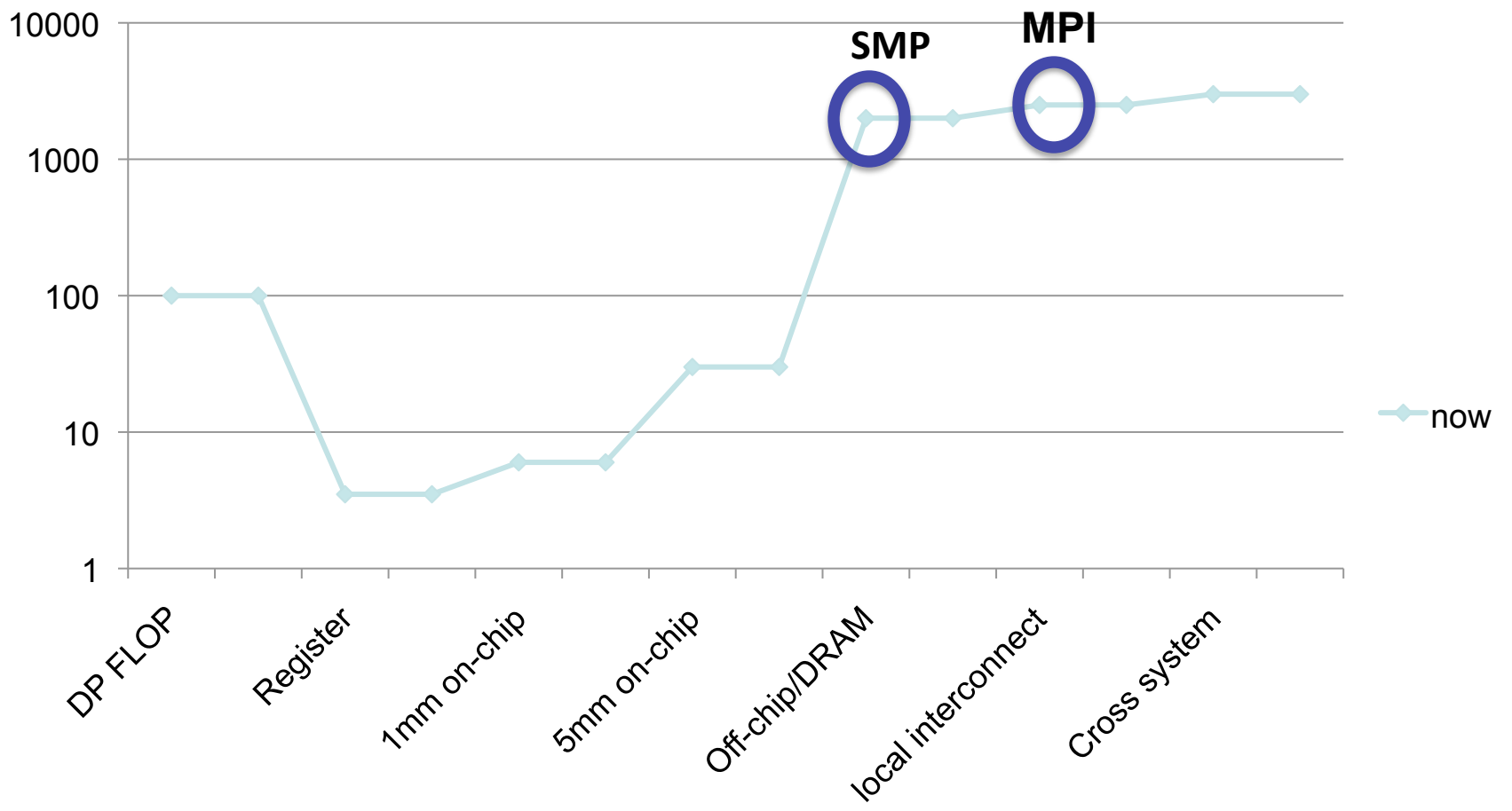
- **Wire data capacity constant as feature size shrinks**
- ***Cost to move bit proportional to distance***
- ***~1TByte/sec max feasible off-chip BW (10GHz/pin)***
- ***Photonics reduces distance-dependence of bandwidth***

Photonics requires no redrive and passive switch little power

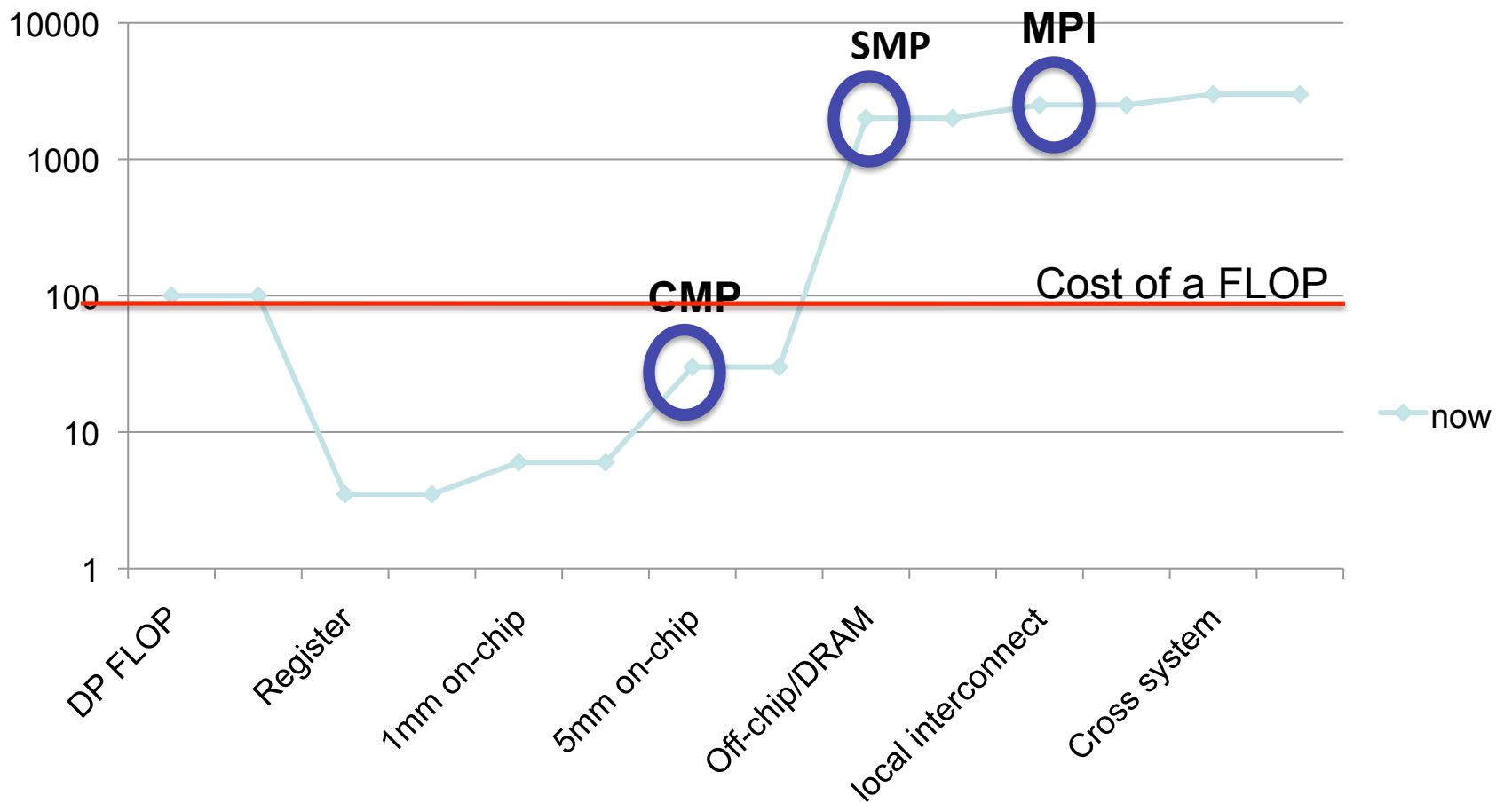
Copper requires to signal amplification even for on-chip connections



The Cost of Data Movement



The Cost of Data Movement



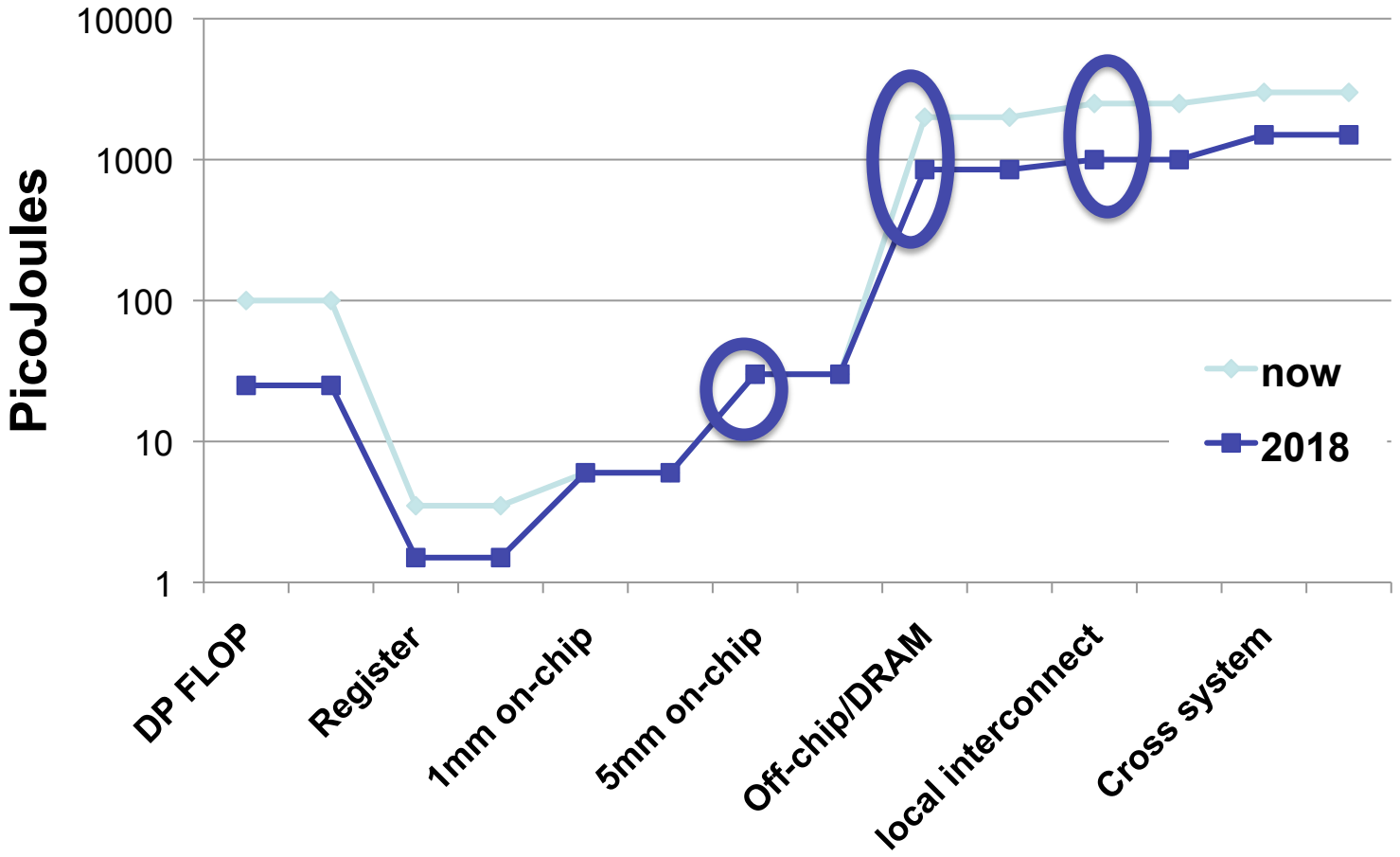
NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

Office of Science U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

The situation will not improve in 2018

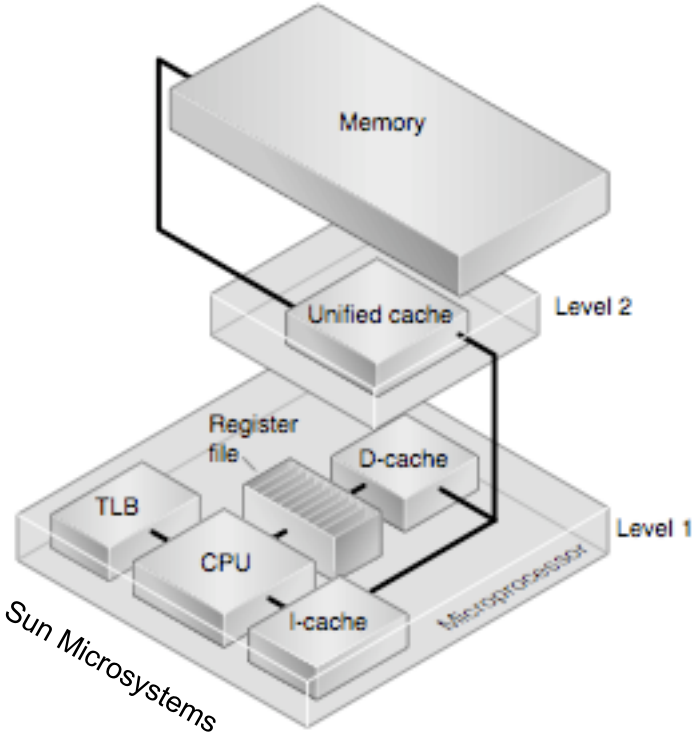
Energy Efficiency will require careful management of data locality



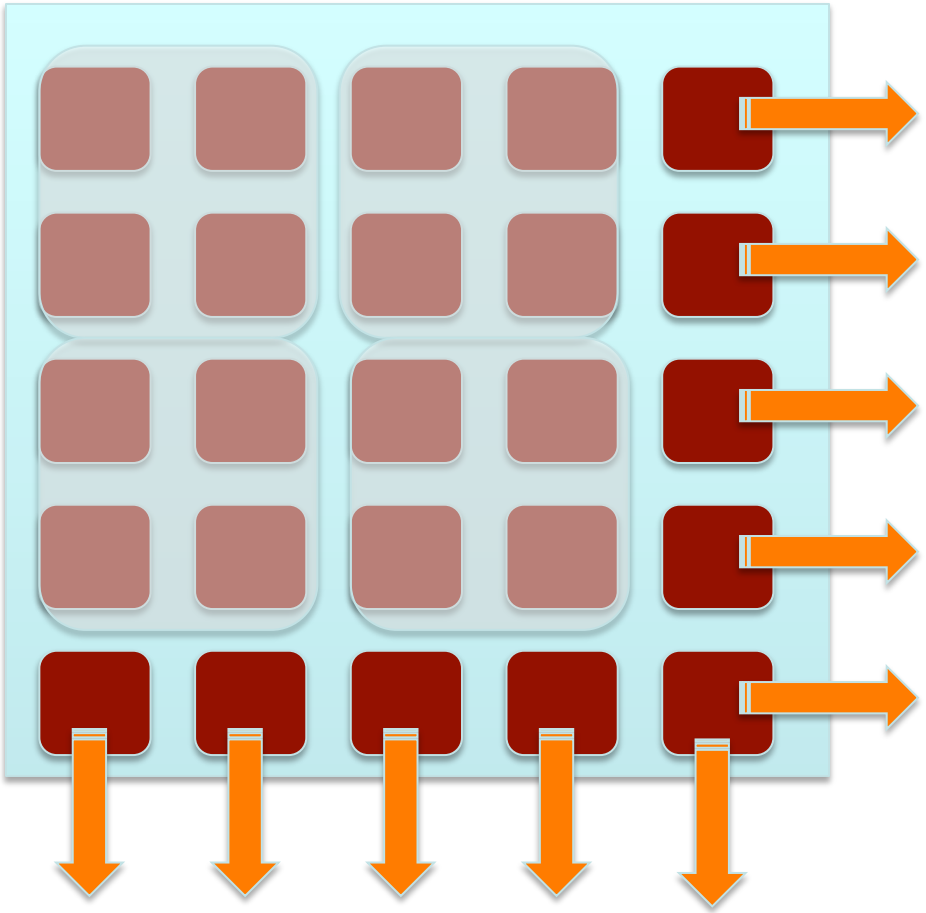
Important to know when you are on-chip and when data is off-chip!

Locality Management is Key

Vertical Locality Management



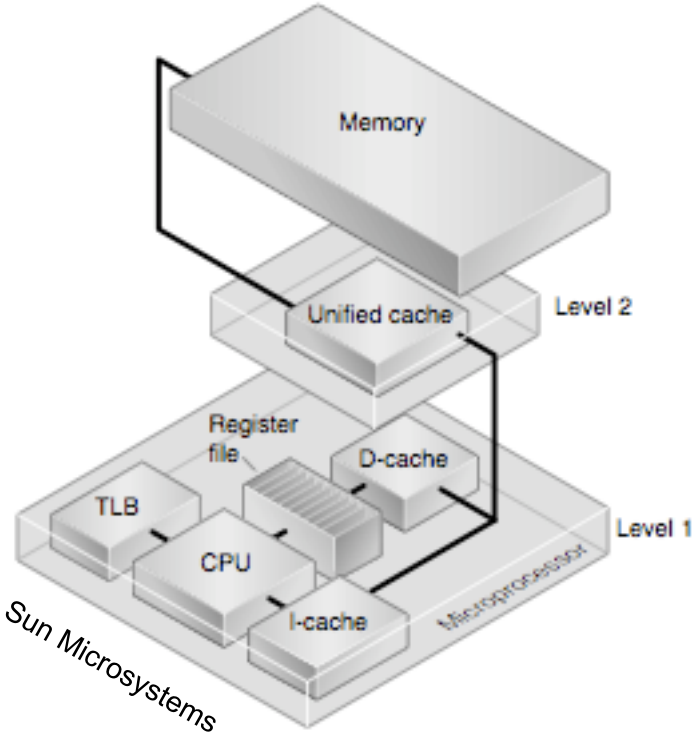
Horizontal Locality Management



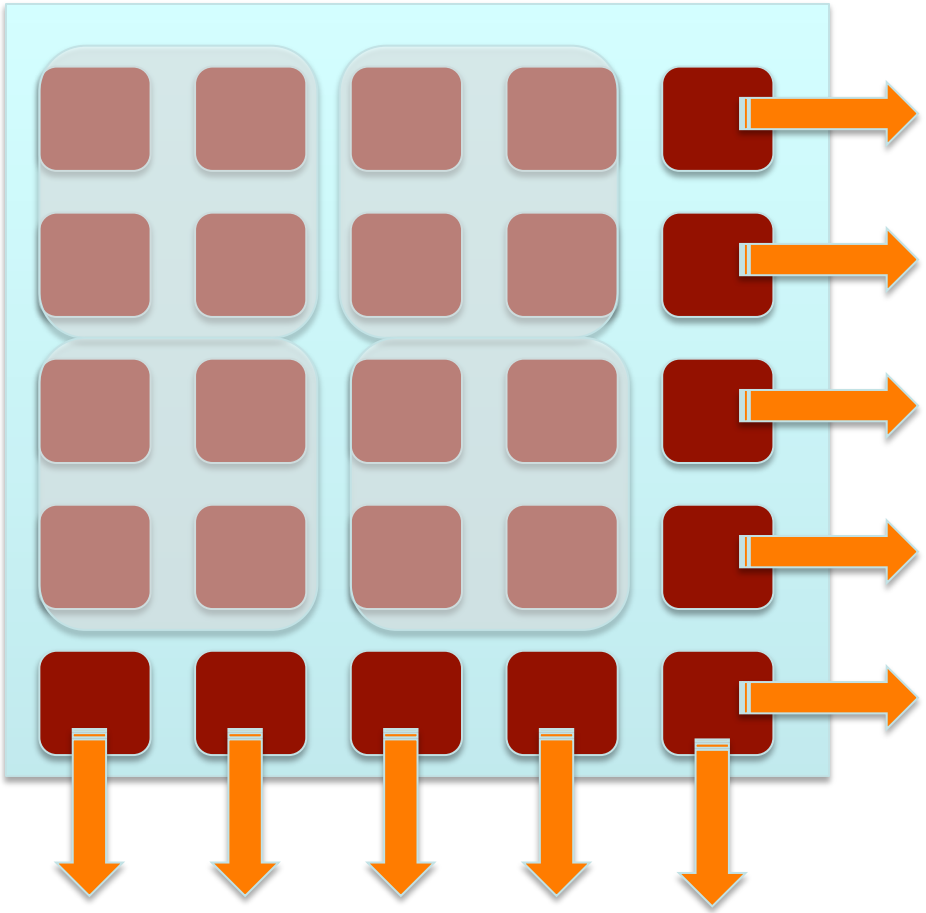
Can no longer assume everything is "equidistant"

Locality Management is Key

Vertical Locality Management



Horizontal Locality Management



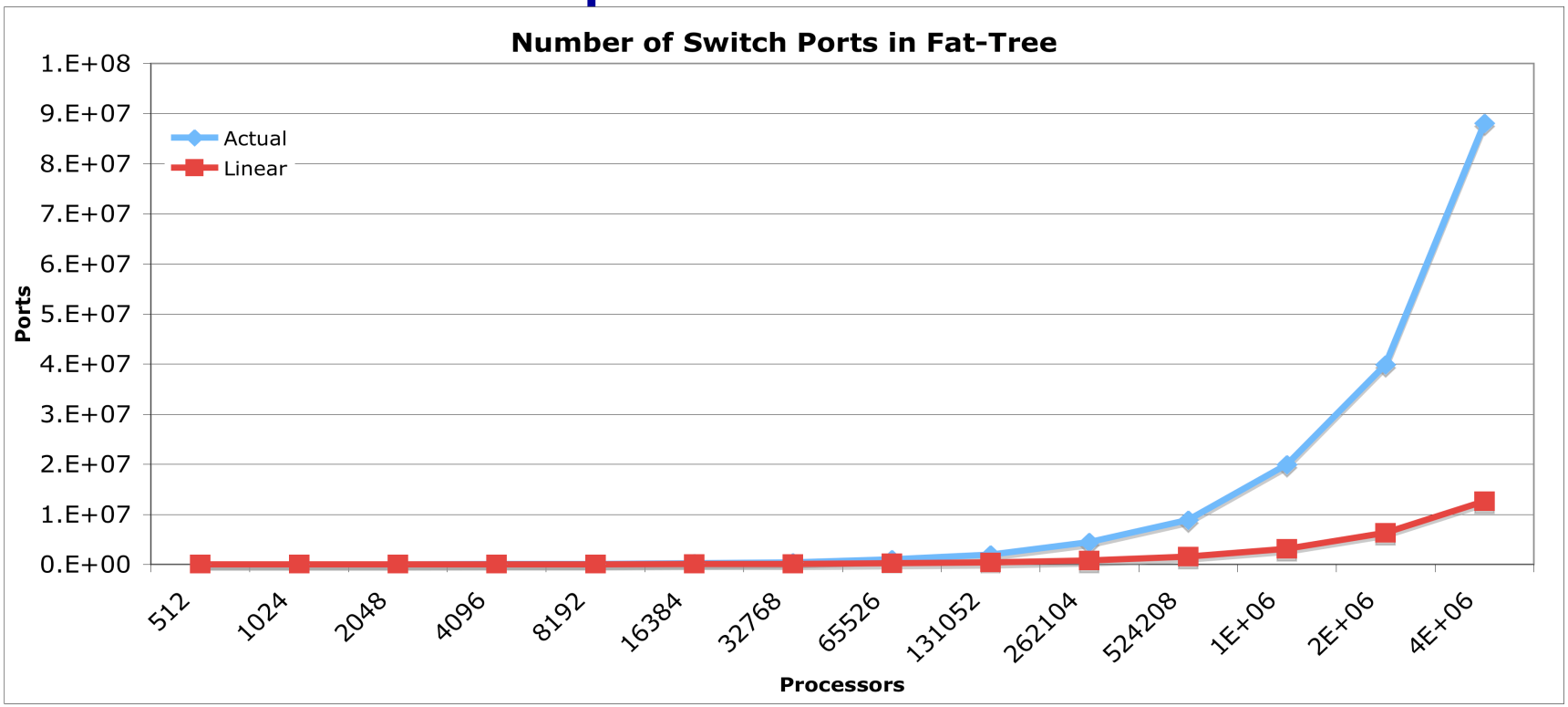
Consider how this will affect shared I/O systems?

NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

Interconnects

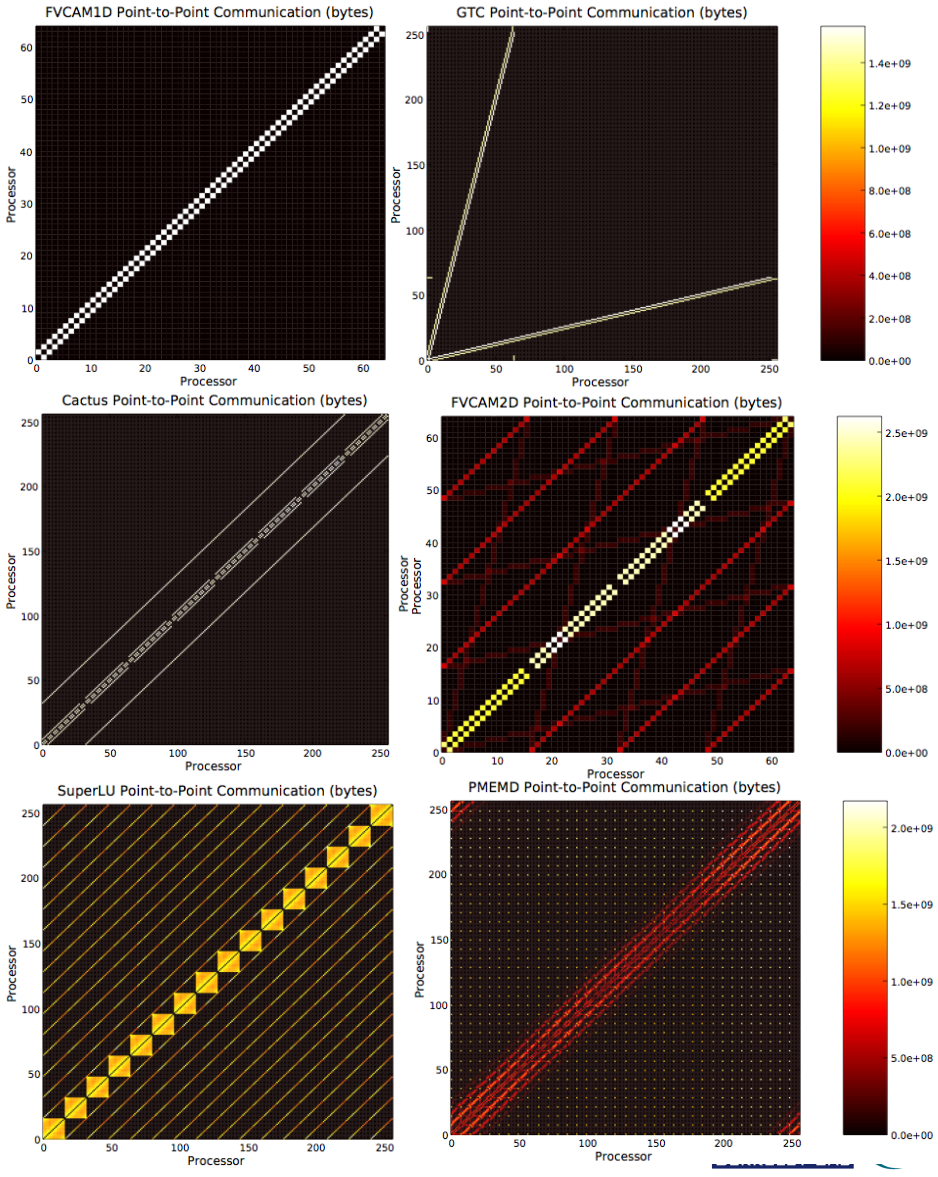
Interconnect Cost
(Scalable Topologies)

- **Fully-connected networks scale superlinearly in cost, but perform the best**
- **Limited-connectivity networks scale linearly in cost, but introduce new problems**



Interconnect Design Considerations for Message Passing Applications

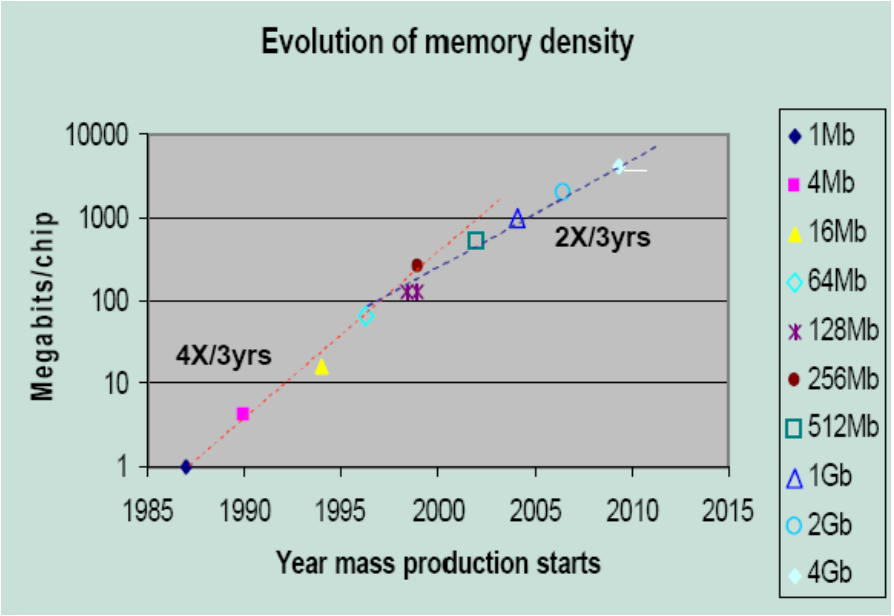
- **Application studies provide insight to requirements for Interconnects (both on-chip and off-chip)**
  - On-chip interconnect is 2D planar (crossbar won't scale!)
  - Sparse connectivity for most apps.; crossbar is overkill
  - No single best topology
  - Most point-to-point message exhibit sparse topology + often bandwidth bound
  - Collectives tiny and primarily latency bound
- **Ultimately, need to be aware of the on-chip interconnect topology in addition to the off-chip topology**
  - Adaptive topology interconnects (HFAST)
  - Intelligent task migration?

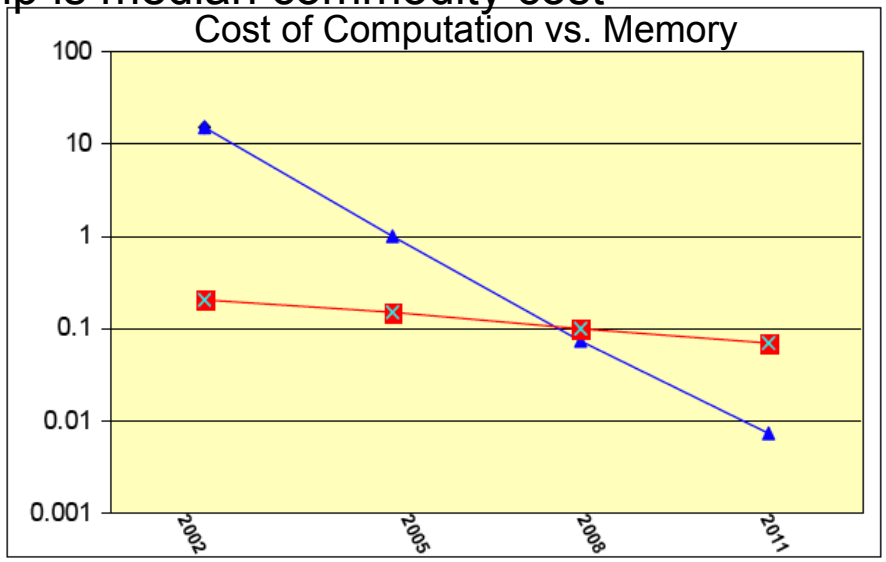


Memory

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

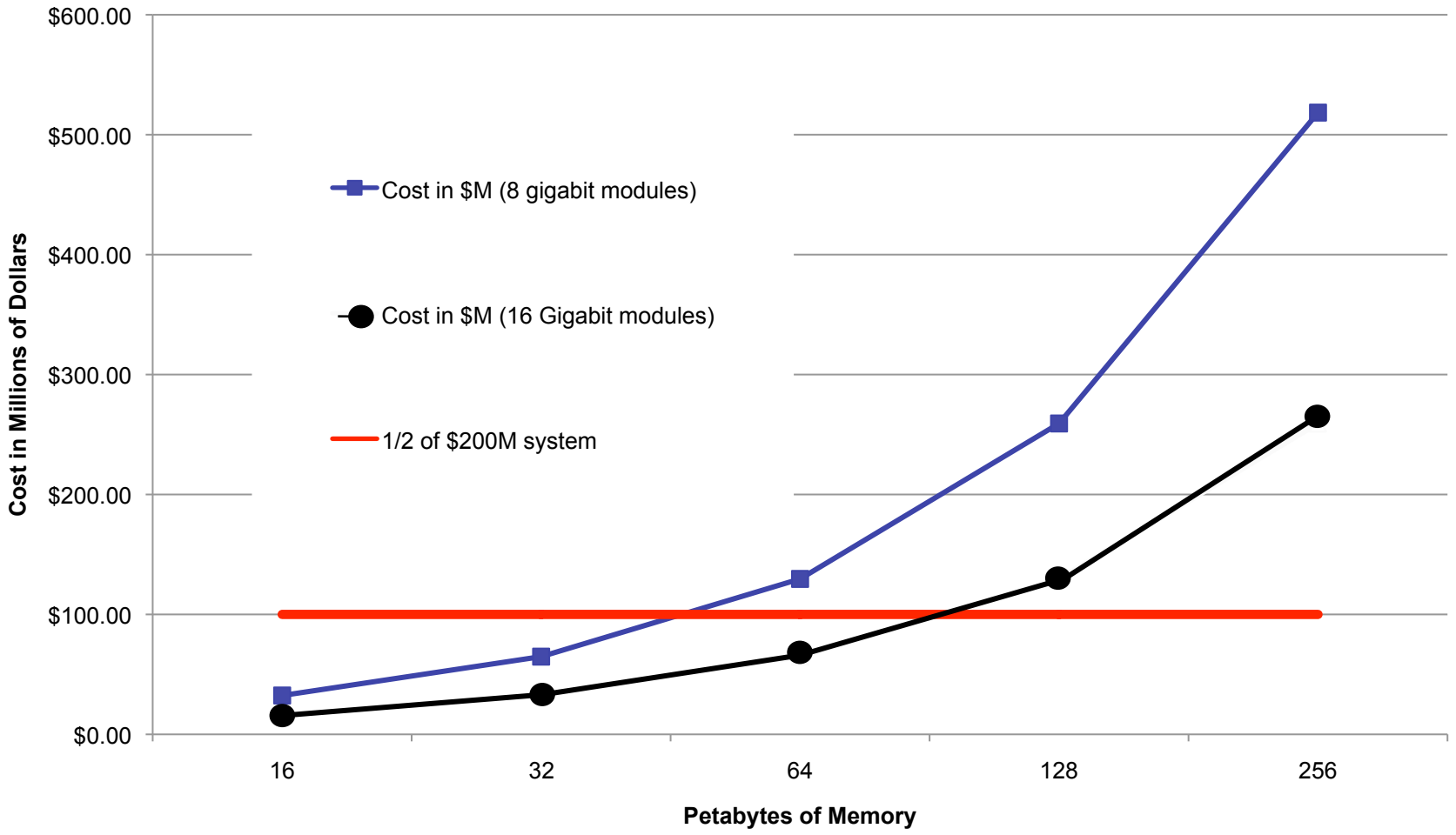




The cost to sense, collect, generate and calculate data is declining much faster than the cost to access, manage and store it

Source: David Turek, IBM

Cost of Memory Capacity
2 different potential Memory Densities

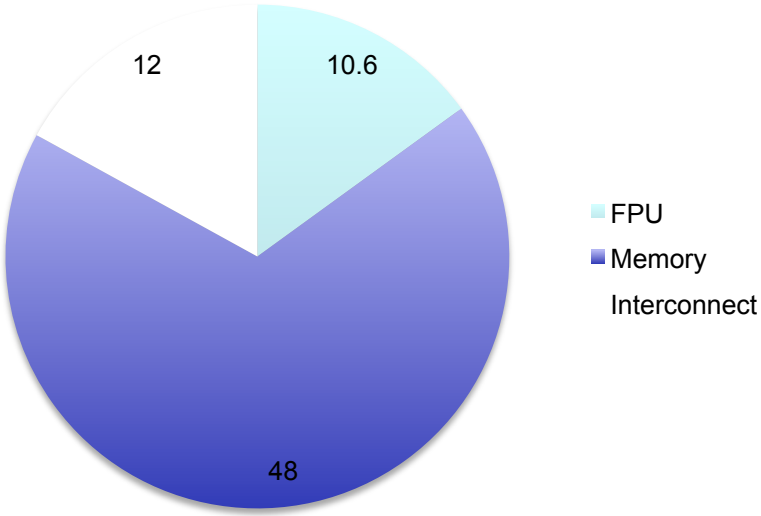


Forces us to strong scaling

Forces us to memory conservative communication (GAS)

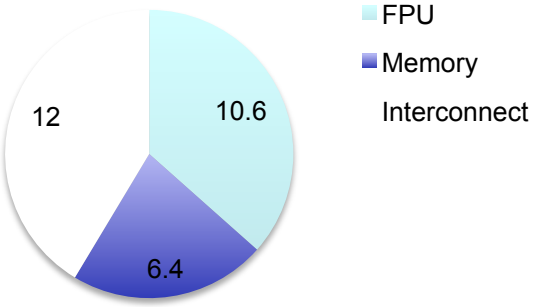
Exascale Memory Power Consumption (San Diego Meeting)

- **Power Consumption with standard Technology Roadmap**



70 Megawatts total

- **Power Consumption with Investment in Advanced Memory Technology**

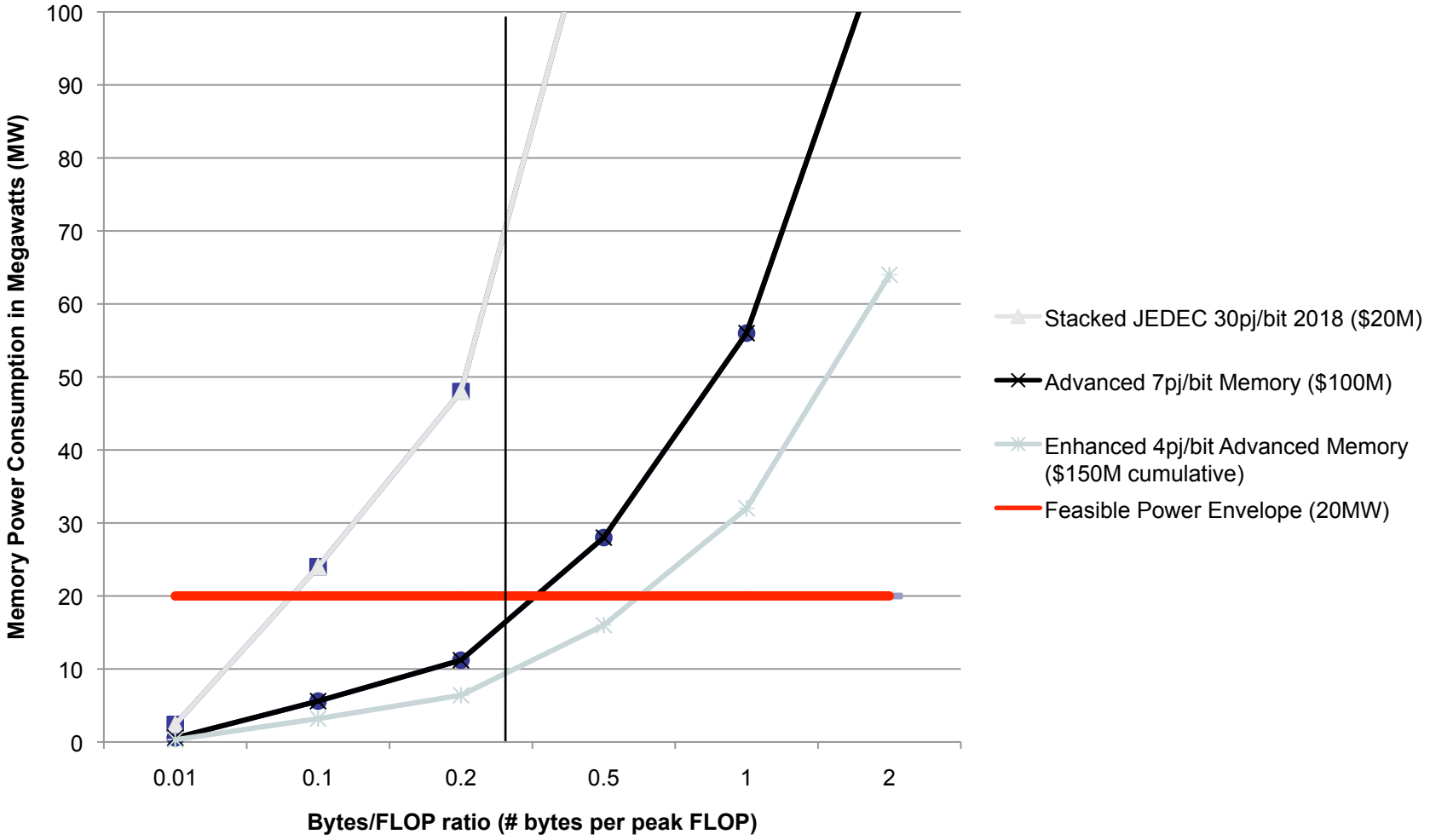


20 Megawatts total

	Phase I	Phase II	Phase III
Capacity/Cube:	512MB	2GB	2GB
Bandwidth/Cube:	128 GB/s	128 GB/s	256GB/s
Energy/bit:	10.0 pJ	7.0 pJ	5.0 pJ

NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

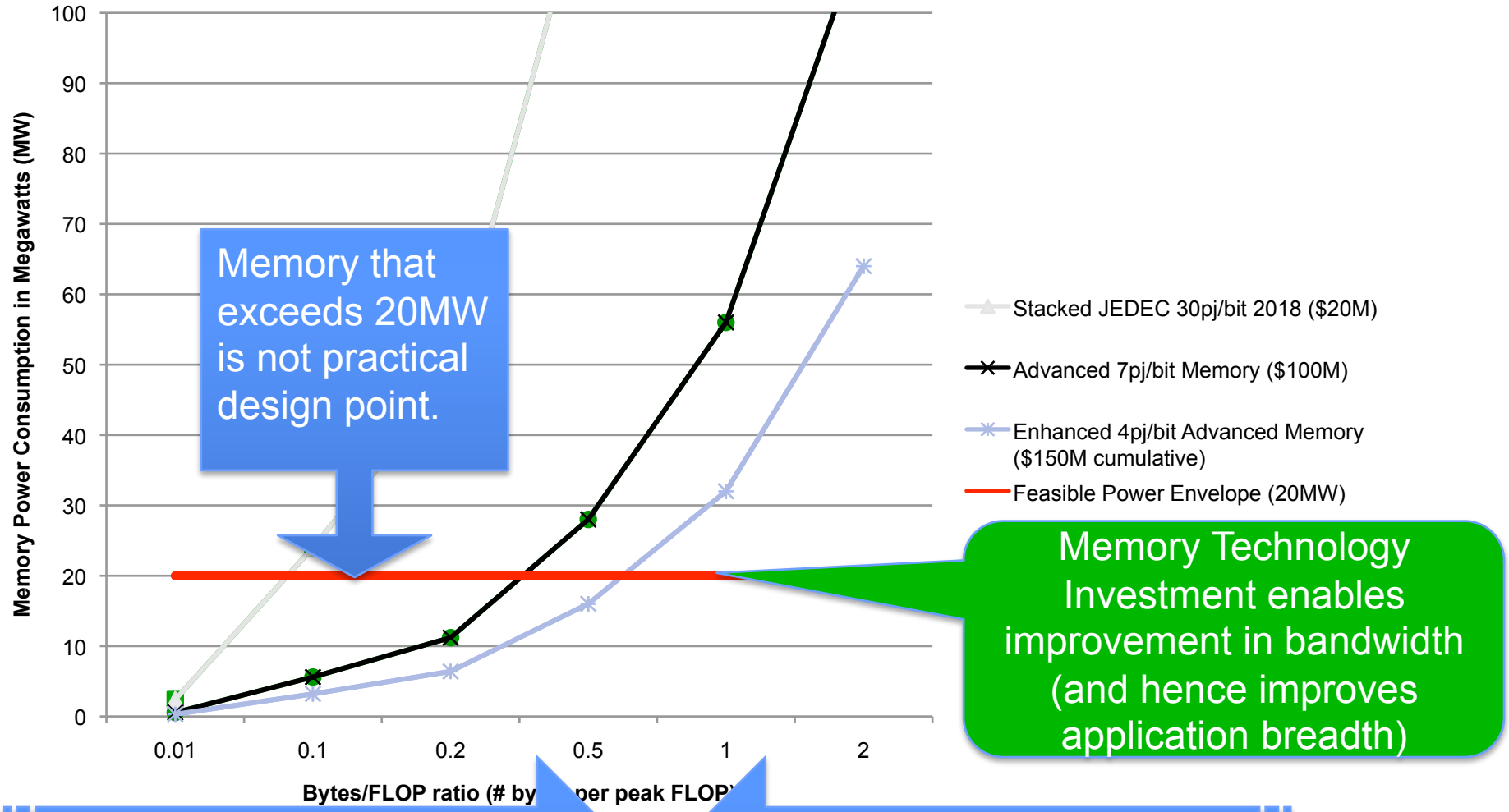
Memory Technology Bandwidth costs power



Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

Limiting Memory Bandwidth Limits System Scope



Application performance and breadth pushes us to higher

Power pushes us to lower bandwidth

Conclusions

- **Memory technology requires major reorganization** *(if domestic industry stays alive)*
  - **More ranks/banks, Less over-fetch, new drivers**
  - **Chip stacking or optical memory interfaces**
  - **New nonvolatile memory technologies**

- **Failure to invest in memory technology means**
  - **We will have to live with less memory (more emphasis on strong scaling)**
  - **We will have lower memory bandwidth/ computational performance (< 0.01 bytes/flop)**

NERSC
NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

Fault Resilience

Chip with FIT rate 1000 fails once every 16 years

A room full of them will fail every few minutes

Fault Tolerance/Resilience

- **Hard Errors: *proportional to component count***
  - Spare cores in design (Cisco Metro)
  - SoC design (fewer components and fewer sockets)
  - Use solder (not sockets)
  - Fewer sockets (pushes us to 10TF chip to keep # sockets const.)
- **Soft Errors: *cosmic rays randomly flip bits***
  - Simpler low-power cores expose less surface area
  - ECC for memory and caches
  - On-board NVRAM controller for localized checkpoint
  - Checkpoint to neighbor for rollback (LLNL SCR)
- **Silent errors: Sometimes RAID & ECC are not enough**
  - End-to-End protection schemes (ZFS)
  - Byzantine Fault Tolerance (BFT)

Industry Trends in Fault Resilience

- Industry must maintain constant FIT rate per node
  - ~1000 failures in time
- Moore's law gets us 100x improvement
  - But still have to increase node count by 10x
- So we will own 10x worse FIT rate
  - MTTI 1week to 1 day
  - MTTI 1 day to 1 hour

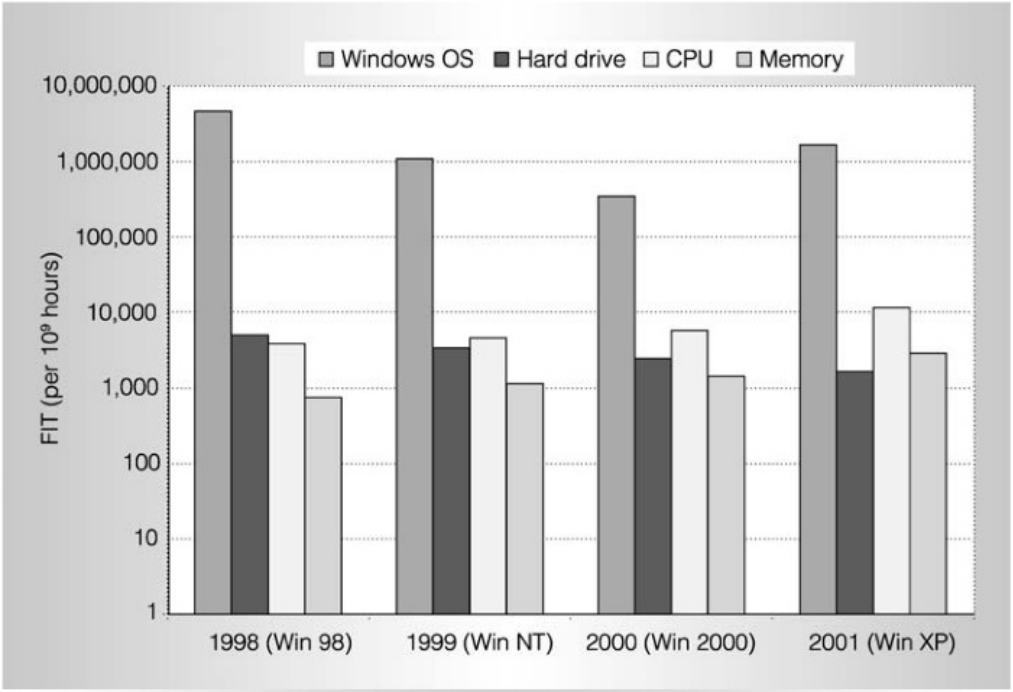


Figure 2. Failures in billions of hours of operation.[2–5]

- Localized checkpointing
  - LLNL SCR to node-local NVRAM
  - More user-assistance in identifying what data to checkpoint

NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

Co-Design

Involve Applications Developers in Navigating Complex Trade-offs

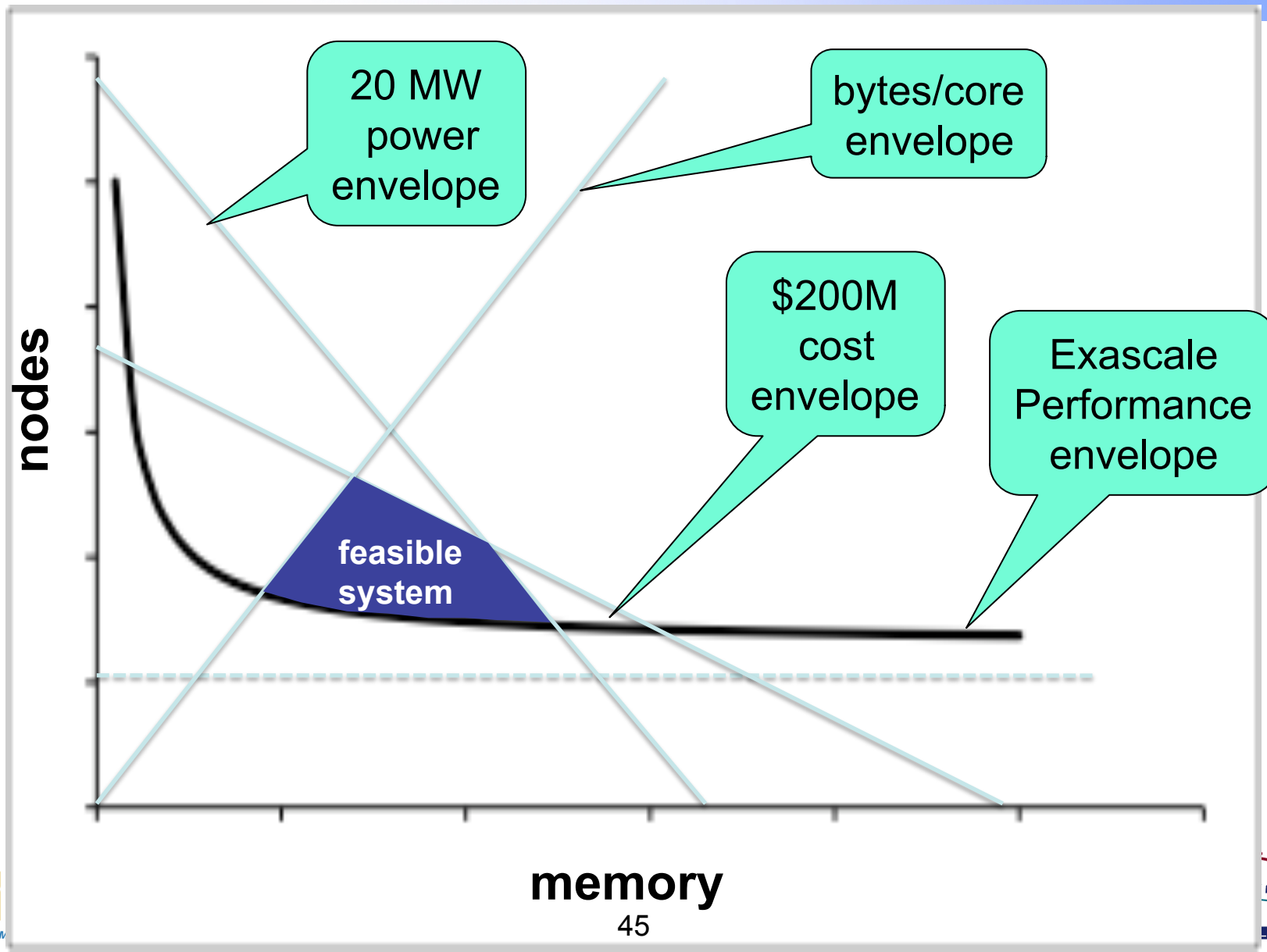
Office of Science U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

Changing Notion of "System Balance"

- If you pay 5% more to double the FPUs and get 10% improvement, it's a win (despite lowering your % of peak performance)
- If you pay 2x more on memory BW (power or cost) and get 35% more performance, then it's a net loss (even though % peak looks better)
- *Real example: we can give up ALL of the flops to improve memory bandwidth by 20% on the 2018 system*
- We have a fixed budget
  - Sustained to peak FLOP rate is *wrong* metric if FLOPs are cheap
  - Balance involves balancing your checkbook & balancing your power budget
  - Requires a application co-design make the right trade-offs

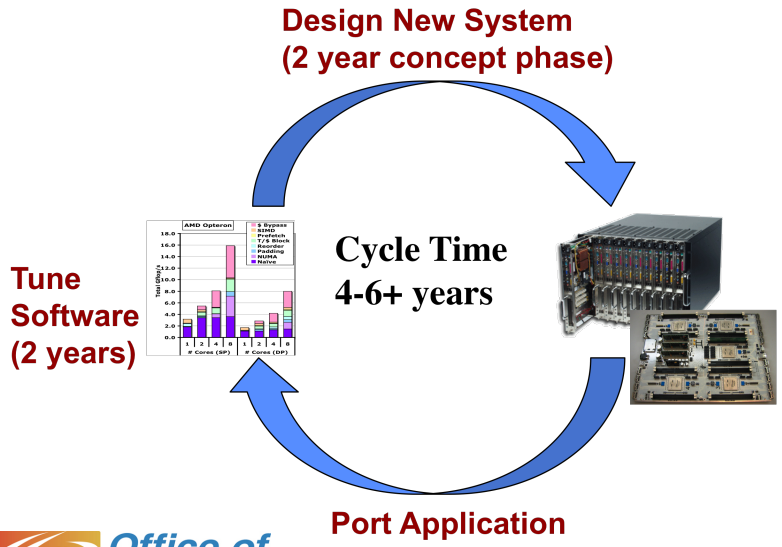
DOE Roadmap: The Trade Space for Exascale is Very Complex.

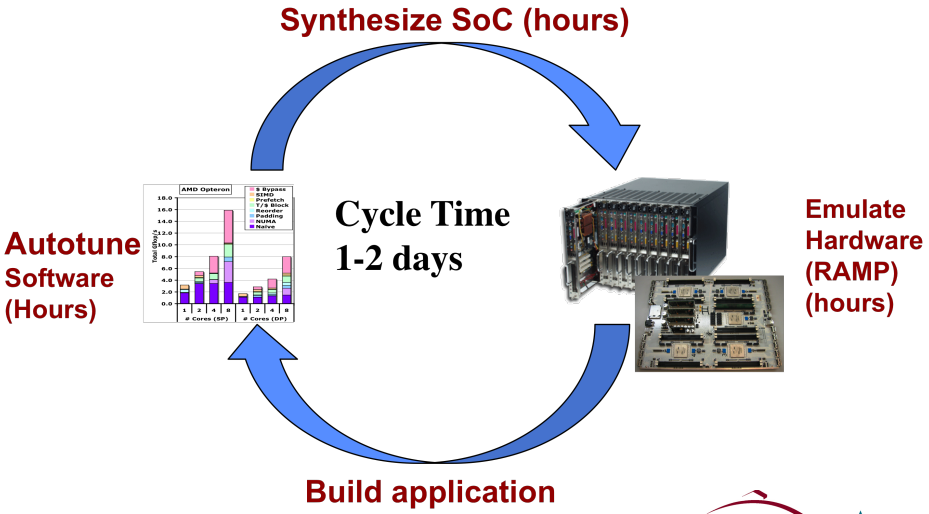


NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

Inserting Scientific Apps into the Hardware Development Process

- **Hardware Architectural Simulation**
  - **Simulate hardware *before* it is built!**
  - **Break slow feedback loop for system designs**
  - **Tightly coupled hardware/software/science co-design** *(not possible using conventional approach)*





Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

Exascale I/O

Office of Science U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

Conclusions from Exascale I/O Meetings Series

Exascale I/O Strategy

I/O Technology

- **Mechanical Disk storage: spindle limited**
  - Requires exponentially more devices (more subject to failure)
  - Need to purchase more capacity than we want to get bandwidth
- **NVRAM/FLASH: way faster than disk, but expensive**
  - Can easily purchase sufficient bandwidth
  - But cannot afford the capacity that we need

- ***Gary's "Reese's Peanut Butter Cup" solution: Hybrid I/O with NVRAM for defensive I/O that bleeds off to disk***

- **Shared Filesystems vs. Distributed Filesystems**
  - Difficult to scale POSIX consistency model to exascale
  - Consider how to integrate node-localized storage into hierarchy
  - How does one manage a distributed filesystem?

Other I/O Issues

- **Defensive I/O (for ~10x higher MTTI)**
  - **Localized Checkpointing:** SCR to local NVRAM could supply required bandwidth
  - *How does one manage node-distributed persistent storage?*
- **Analysis I/O**
  - **In-situ (locality aware) data analysis**: e.g. MapReduce: Layout data across cluster and ship computation to the storage (functional semantics)
  - **Object database storage** (HDF, NetCDF) pushed into the storage infrastructure (interoperate with locality-aware storage)

All requires a lot more discussion

(which should happen here)

Overall Conclusions

- **Supercomputers are power limited**
  - Limited by end of Dennard scaling for logic
  - Limited by energy cost of moving bits
- **Primary growth in explicit parallelism is on-chip**
  - *100x growth in parallelism on-chip*
  - *10x growth in parallelism off-chip*
- **Need a new abstract machine model that reflects hierarchical power costs**
  - *Current abstract machine model has flat or 2-level costs, which do not match technology trends*
  - *Will require fundamental advances in technology and system architecture*
  - *Will result in disruptive changes to programming model*

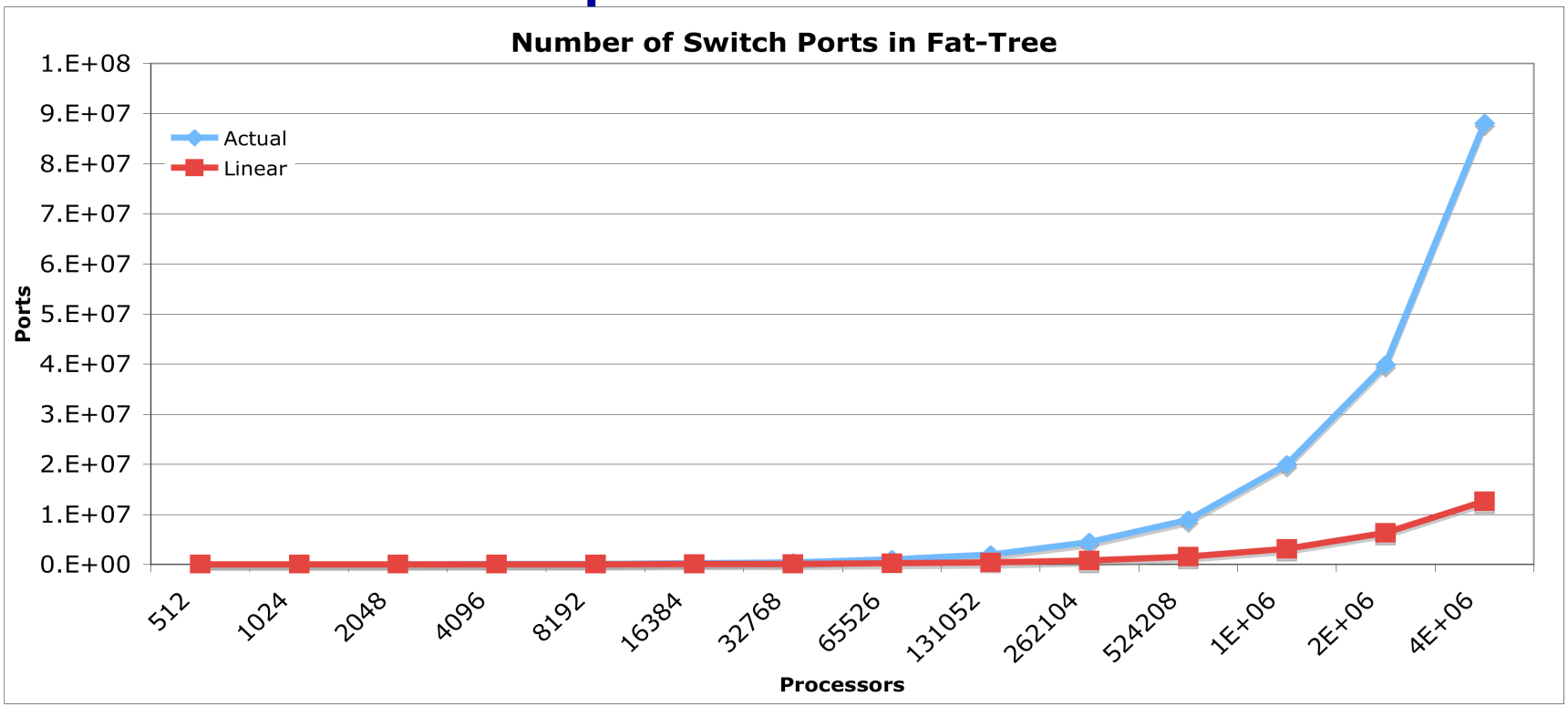
More Info

- **DOE Exascale**
  - http://extremecomputing.labworks.org/
  - http://www.exascale.org/
- **NERSC Advanced Technology Group**
  - http://www.nersc.gov/projects/SDSA

Bonus Material

Interconnect Cost
(Scalable Topologies)

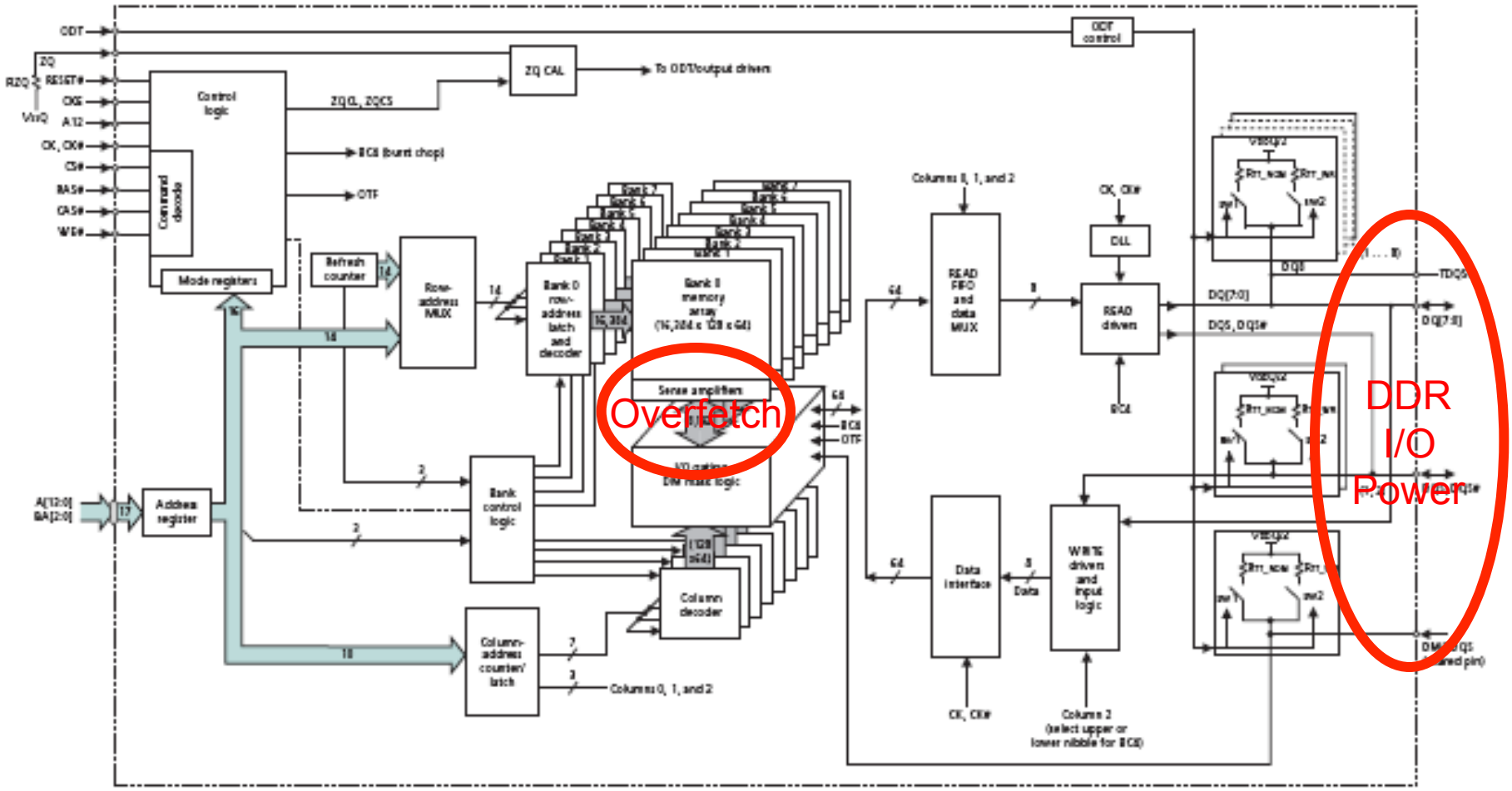
- **Fully-connected networks scale superlinearly in cost, but perform the best**
- **Limited-connectivity networks scale linearly in cost, but introduce new problems**



Systems	2009	2015 +1/-0	2018 +1/-0
System peak	**2 Peta**	**100-300 Peta**	**1 Exa**
Power	**6 MW**	**~15 MW**	**~20 MW**
System memory	0.3 PB	5 PB	64 PB (+)
Node performance	125 GF	0.5 TF or 7 TF	2 TF or 10TF
Node memory BW	25 GB/s	0.2TB/s or 0.5TB/s	0.4TB/s or 1TB/s
Node concurrency	12	O(100)	O(1k) or 10k
Total Node Interconnect BW	3.5 GB/s	100-200 GB/s 10:1 vs memory bandwidth 2:1 alternative	200-400GB/s (1:4 or 1:8 from memory BW)
System size (nodes)	18,700	50,000 or 500,000	O(100,000) or O(1M)
Total concurrency	225,000	O(100,000,000) *O(10)-O(50) to hide latency	O(billion) * O(10) to O (100) for latency hiding
Storage	15 PB	150 PB	500-1000 PB (>10x system memory is min)
IO	0.2 TB	10 TB/s	60 TB/s (how long to drain the machine)
MTTI	days	O(1day)	O(1 day)

Systems	2009	2015 +1/-0	2018 +1/-0
System peak	**2 Peta**	**100-300 Peta**	**1 Exa**
Power	**6 MW**	**~15 MW**	**~20 MW**
System memory	0.3 PB	5 PB	64 PB (+)
Node performance	125 GF	0.5 TF or 7 TF	2 TF or 10TF
Node memory BW	25 GB/s	0.2TB/s or 0.5TB/s	0.4TB/s or 1TB/s
Node concurrency	12	O(100)	O(1k) or 10k
Total Node Interconnect BW	3.5 GB/s	100-200 GB/s 10:1 vs memory bandwidth 2:1 alternative	200-400GB/s (1:4 or 1:8 from memory BW)
System size (nodes)	18,700	50,000 or 500,000	O(100,000) or O(1M)
Total concurrency	225,000	O(100,000,000) *O(10)-O(50) to hide latency	O(billion) * O(10) to O(100) for latency hiding
Storage	15 PB	150 PB	500-1000 PB (>10x system memory is min)
IO	0.2 TB	10 TB/s	60 TB/s (how long to drain the machine)
MTTI	days	O(1day)	O(1 day)

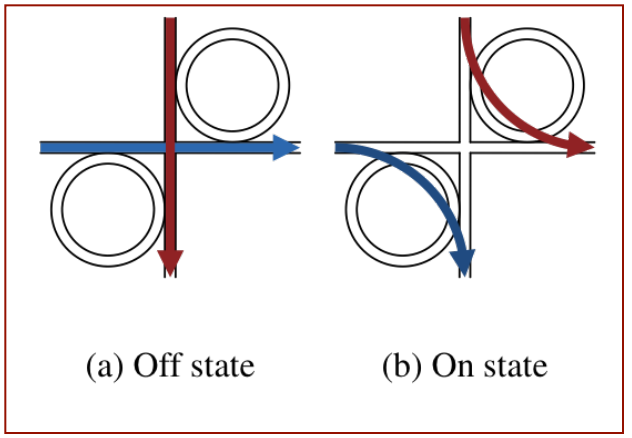
1Gbit DDR3 Architecture

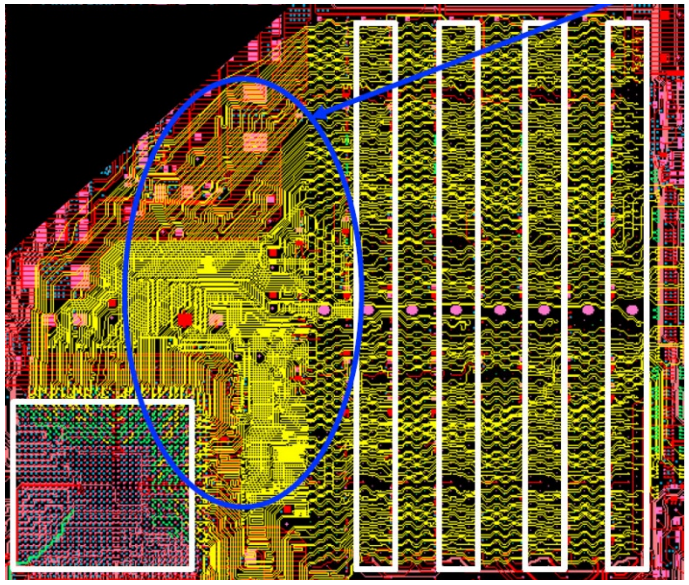


Slide from Dean Klein (Micron Technology)

Optical Memory Interfaces

- **On chip:**
  - Optical interconnect enabled with Si photonic ring resonators
  - Integrates with conventional CMOS
  - Up to 27x power improvement
- **Off Chip:**
  - DDR interface power hungry
    - Cu line capacitance
    - Large voltage swing
  - Optical link much more efficient
    - Very small voltage modulation required
    - 50x reduction in interface power
- **Unified optical fabric to reduce optical / electrical conversion**
- **Stacking to improve density**

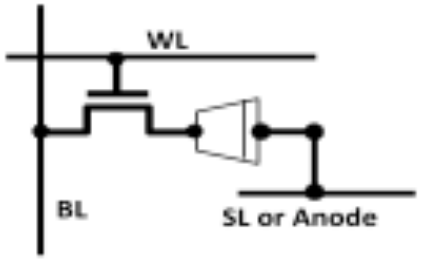


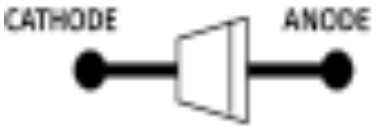


Wiring of a single channel DDR to the Memory controller (Intel)

Looking Beyond DRAM

- **Resistive Change RAM (ReRAM)**
  - **Nonvolatile - no refresh required!**
  - **No high-voltage requirement**
  - **Less energy / write (compared to FLASH)**
  - **More robust than FLASH**
    - **More cycles to cell wear out**
  - **Lower read energy than DRAM**
    - **< 1V read-out voltage**
  - **Similar density to flash**
    - **MLC capable**
    - **2-4x DRAM**
  - **Read / write speeds comparable (or better!) than DRAM**
  - **Integrates very well with existing CMOS processes**





Overall 10x reduction in power with a 4x increase in density