SAND2011-5475C

Through the Exascale Looking-Glass and What Alice Found There

Richard C. Murphy

Scalable Computer Architectures Department

Sandia National Laboratories

Affiliated Faculty, New Mexico State University

August 3, 2011

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company,
for the United States Department of Energy's National Nuclear Security Administration
under contract DE-AC04-94AL85000.

What have we learned from UHPC?

- DOE and DoD have very different computing agendas, but the same fundamental set of needs
  - –Lots of address generating tasks (I like threads, but...)
  - –Lightweight synchronization
  - –Global naming
  - –Mix of message passing, PGAS, and work moving models
- Execution Models are the most important tool of codesign
  - –This needed to be solidified last year, so we're behind
  - –Sandia and Intel are in the process of transitioning from a joint "UHPC" model to a community model
- Performance is about the data movement system not the compute engine
- 3D Integration and Silicon Photonics are the most important technology investments

Example from the EI RFI

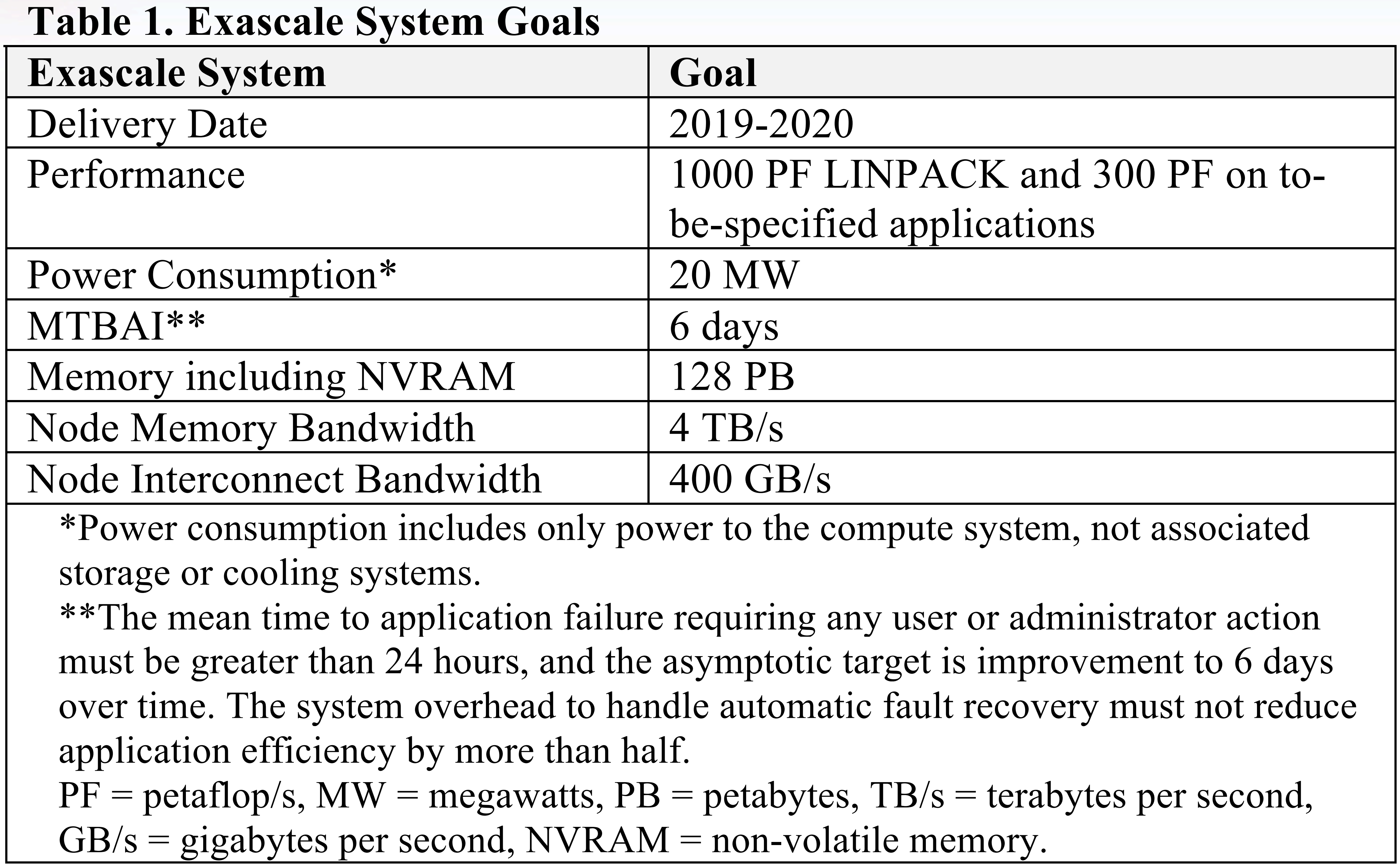


Table 1. Exascale System Goals

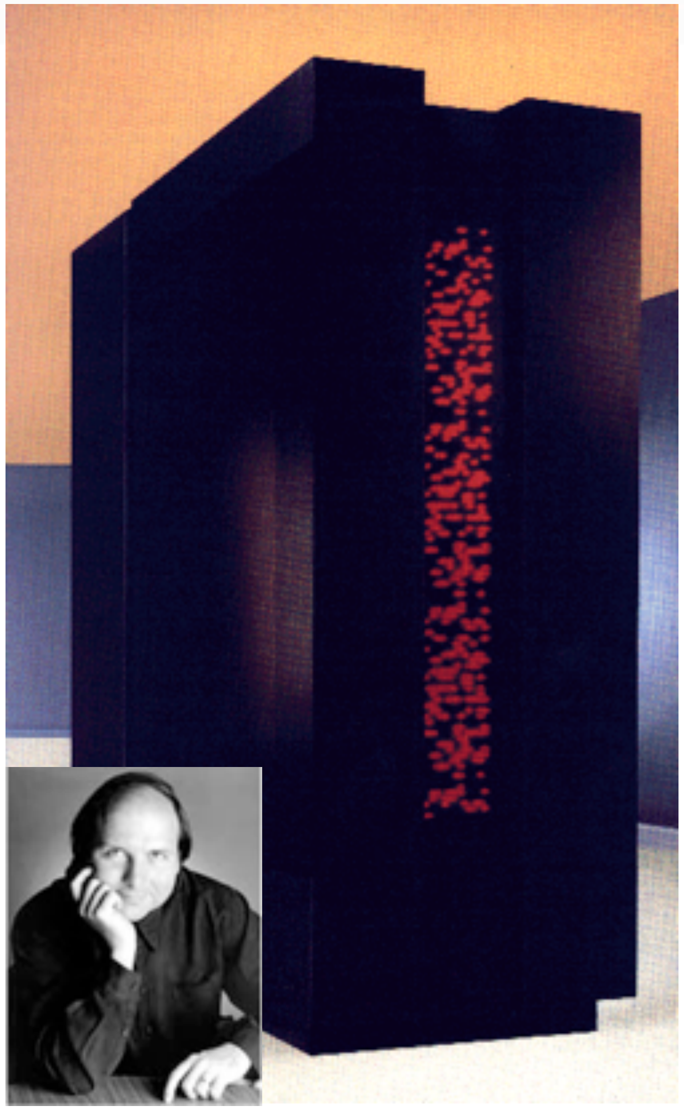
Exascale System	Goal
Delivery Date	2019-2020
Performance	1000 PF LINPACK and 300 PF on to-be-specified applications
Power Consumption*	20 MW
MTBAI**	6 days
Memory including NVRAM	128 PB
Node Memory Bandwidth	4 TB/s
Node Interconnect Bandwidth	400 GB/s

*Power consumption includes only power to the compute system, not associated storage or cooling systems.

**The mean time to application failure requiring any user or administrator action must be greater than 24 hours, and the asymptotic target is improvement to 6 days over time. The system overhead to handle automatic fault recovery must not reduce application efficiency by more than half.

PF = petaflop/s, MW = megawatts, PB = petabytes, TB/s = terabytes per second, GB/s = gigabytes per second, NVRAM = non-volatile memory.

DOE Has Over-constrained the Exascale Problem



Step 1: Choose your favorite MIT Alumni's Computer from the early 1990s as your compute node... errr swim lane

DOE Has Over-constrained the Exascale Problem

Step 2: Pick your favorite MIT Alumni’s network topology and wire up whatever bandwidth you think you can afford

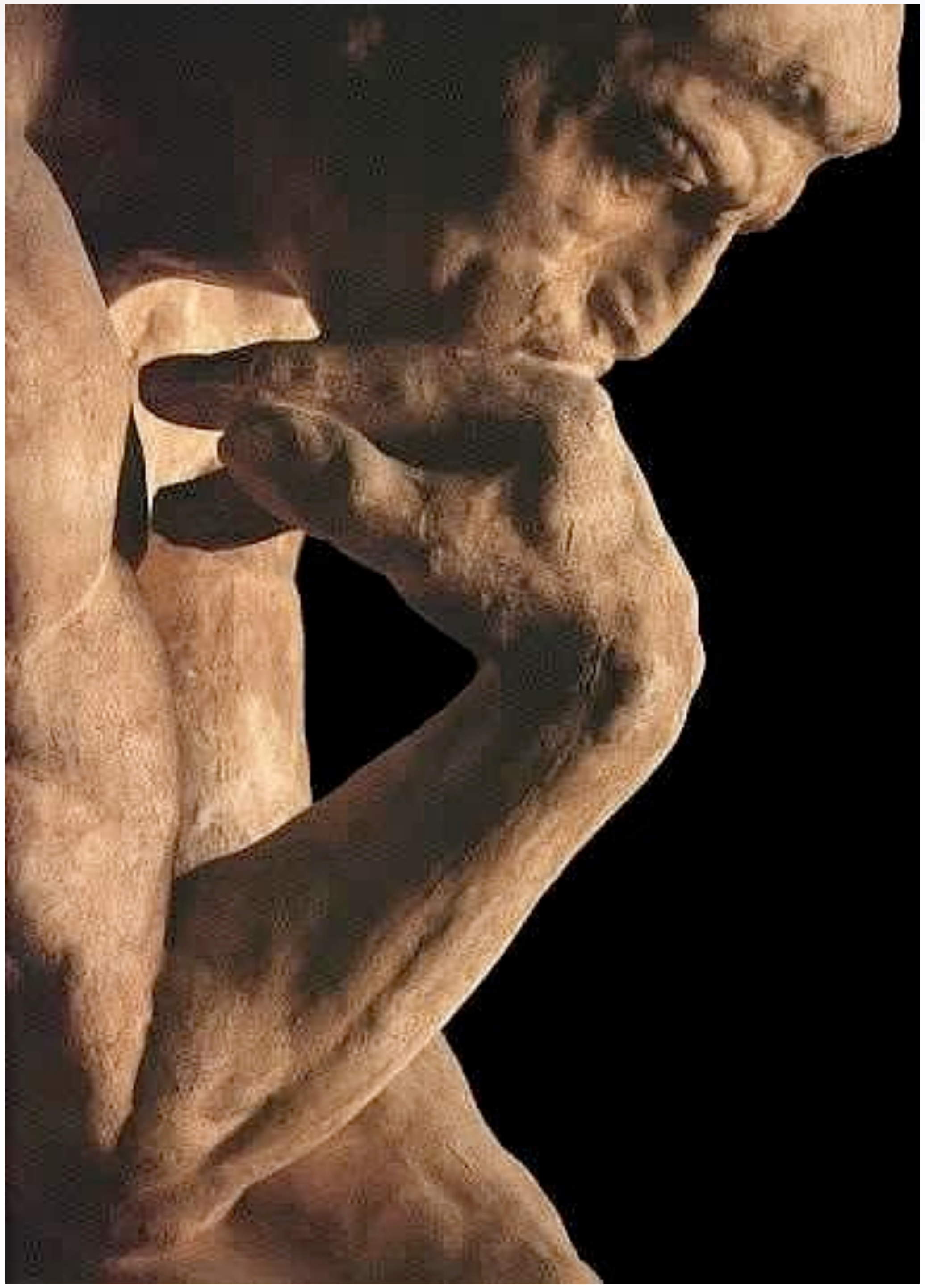
DOE Has Over-constrained the Exascale Problem

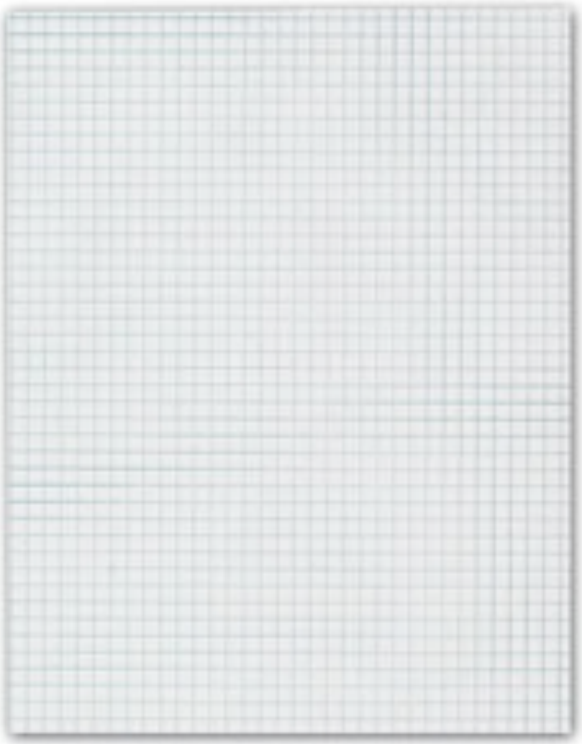
Step 3: Mix in MPI... and a touch of your favorite MIT Alumni's Alternative Programming Model

DOE Has Over-constrained the Exascale Problem

We all know this approach is subject to criticism...
(I personally think it’s a disaster)

DoD Approach





Differences in Viewpoint

DOE	DoD
•Applications that are older than I am	•Kleenex code in a rapidly changing environment
•Risk lowered by preserving validated applications (at least in the NW complex)	•Risk lowered by covering more space and timely generation of results
•3-Dimensional Physics	•N-dimensional space
•Simple (3D) naming scheme	•Complex, machine-wide naming scheme
•Amenable to halo exchange	•GUPS-oriented
•FLOPS are important but not dominant	•FLOPS are much less important

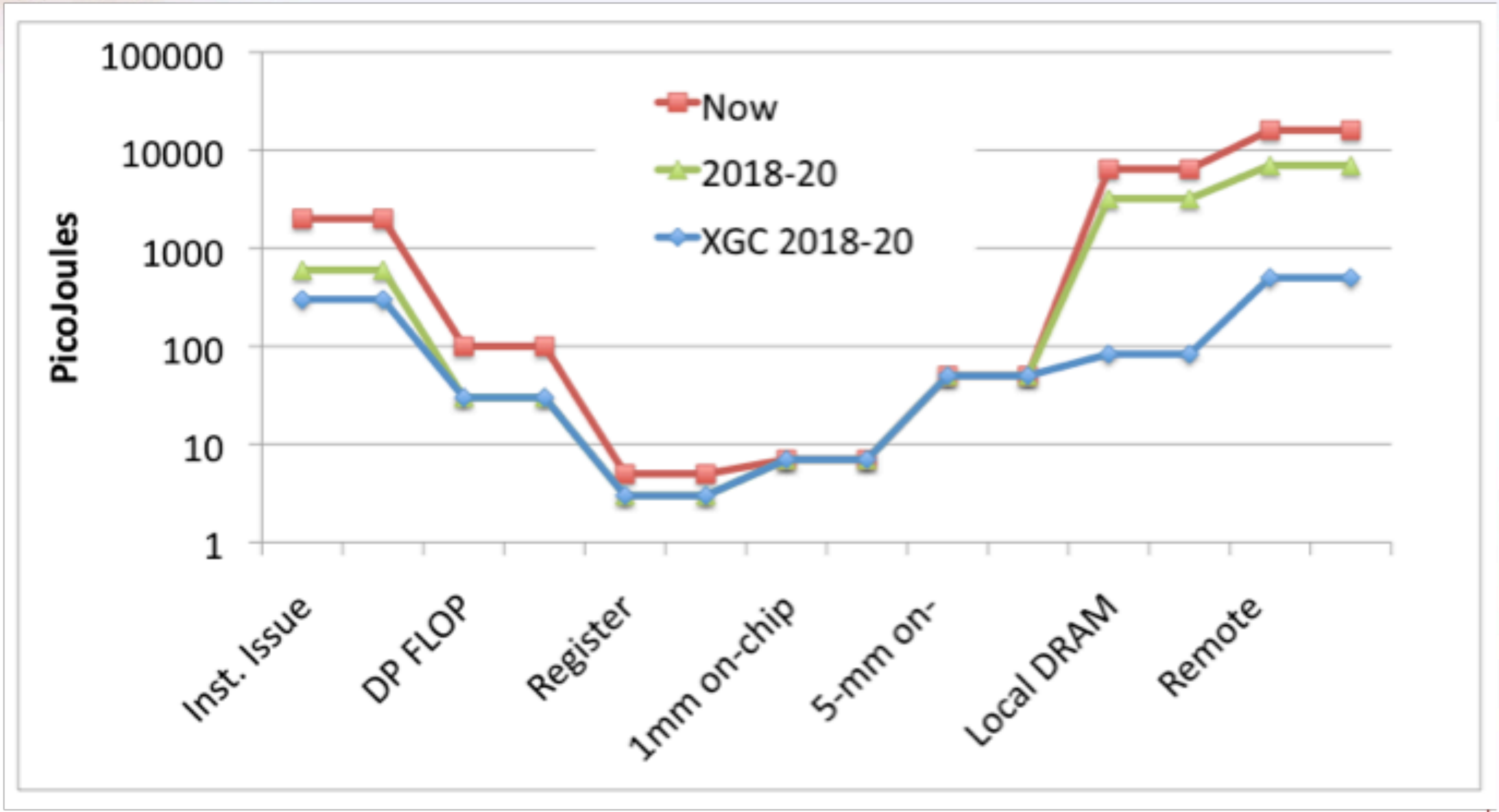
Thoughts on Vintage DoD Computing

- Early Petaflops Effort (1996-1999)
  - NSF, DARPA, NASA, NSA
  - DOE stayed out because the mission need could be met with commodity (but we're paying the price now)
- One of 8 NSF-sponsored petaflops design points in a 6 month study
- We were able to get to petascale a decade later
  - Without addressing the fundamental energy issues
  - Without programming model innovation, which we know we need

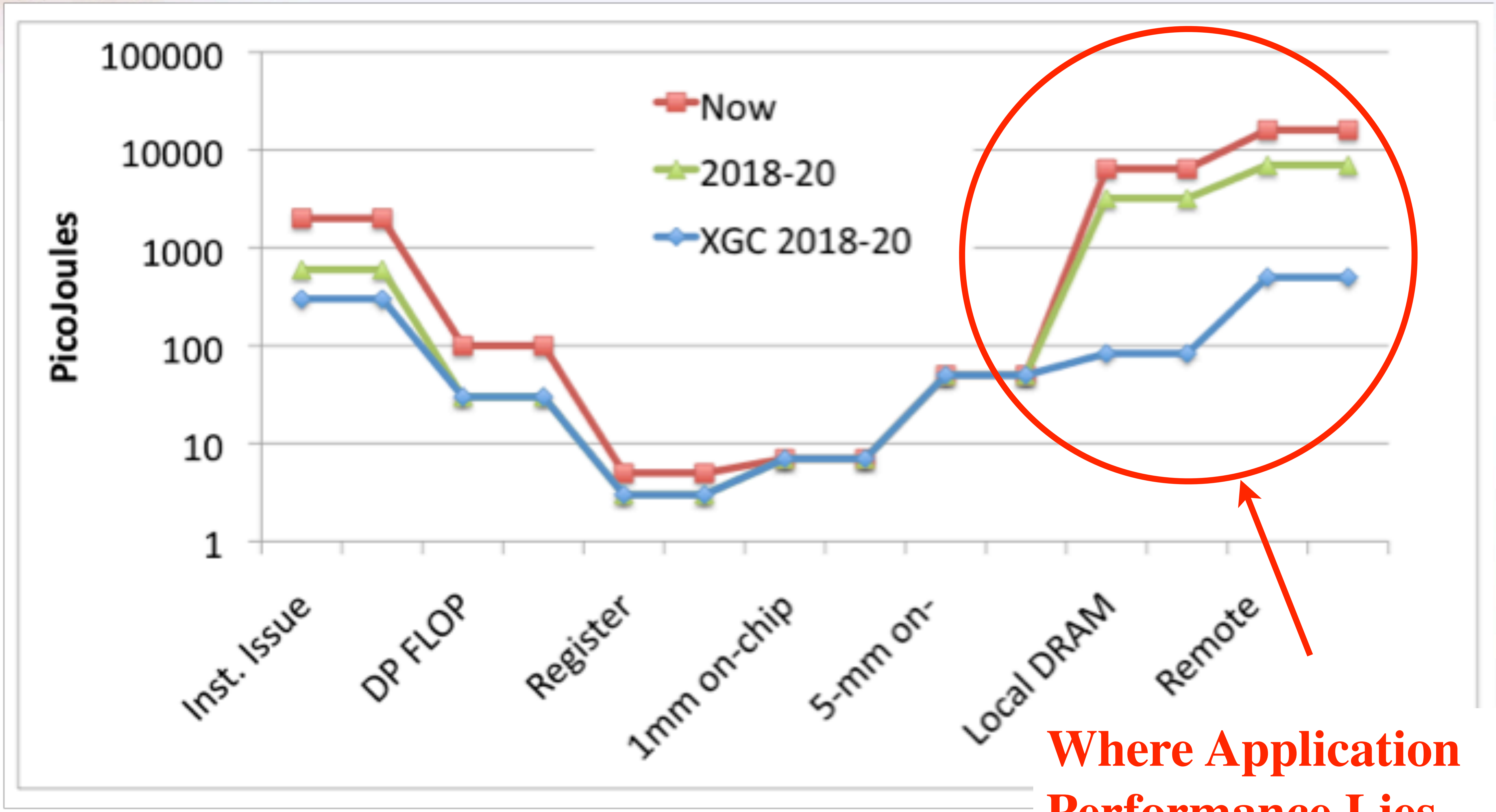
  - Without broad agreement between government agencies
- Consider the power envelopes:
  - 2007 HTMT Design Point: 2.4 MW
    - Scaled (unfairly) by Moore's Law: < 1.2MW
  - 2008 Road Runner PF/s: 2.4 MW
  - 2008 Jaguar PF/s: 7 MW

Key concepts from HTMT drive today's Exascale research agenda (threads, message-driven computation, global shared memory)

Where do we focus our attention?



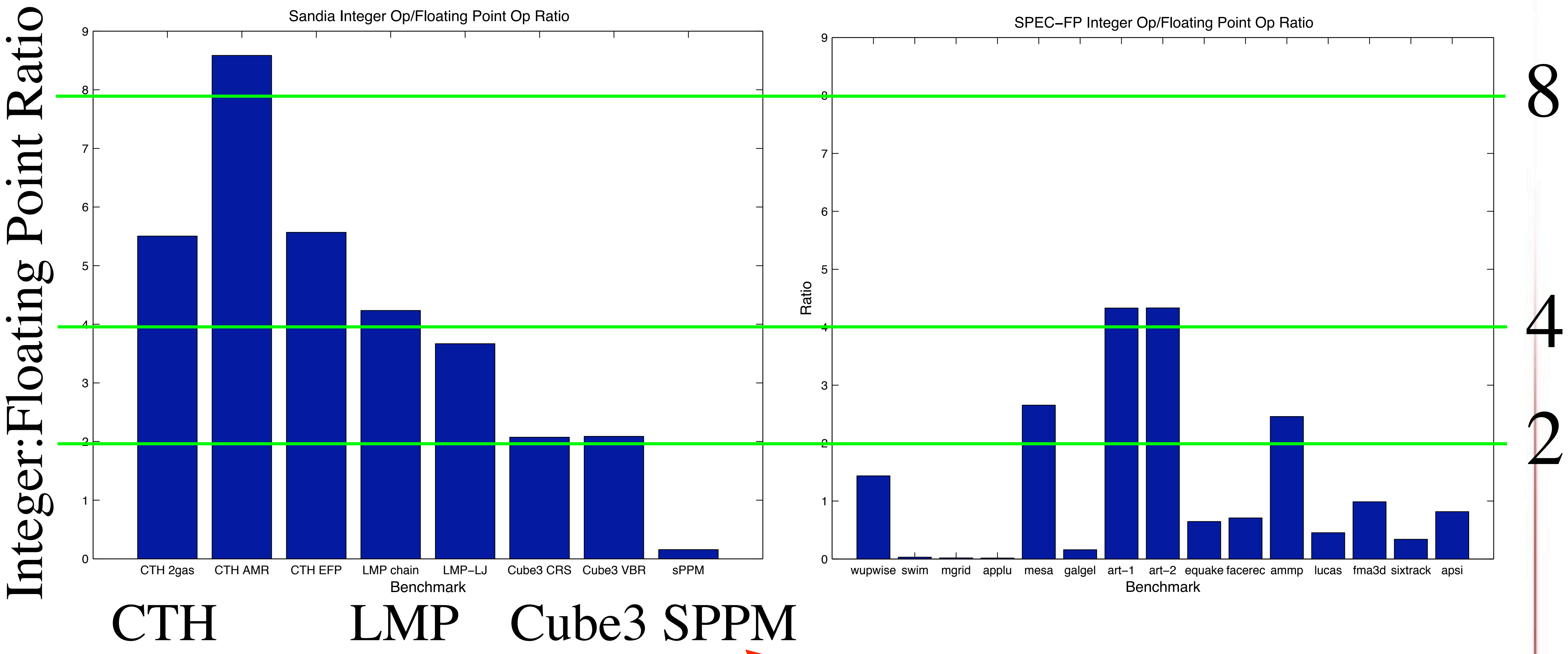
Where do we focus our attention?



Where Application Performance Lies

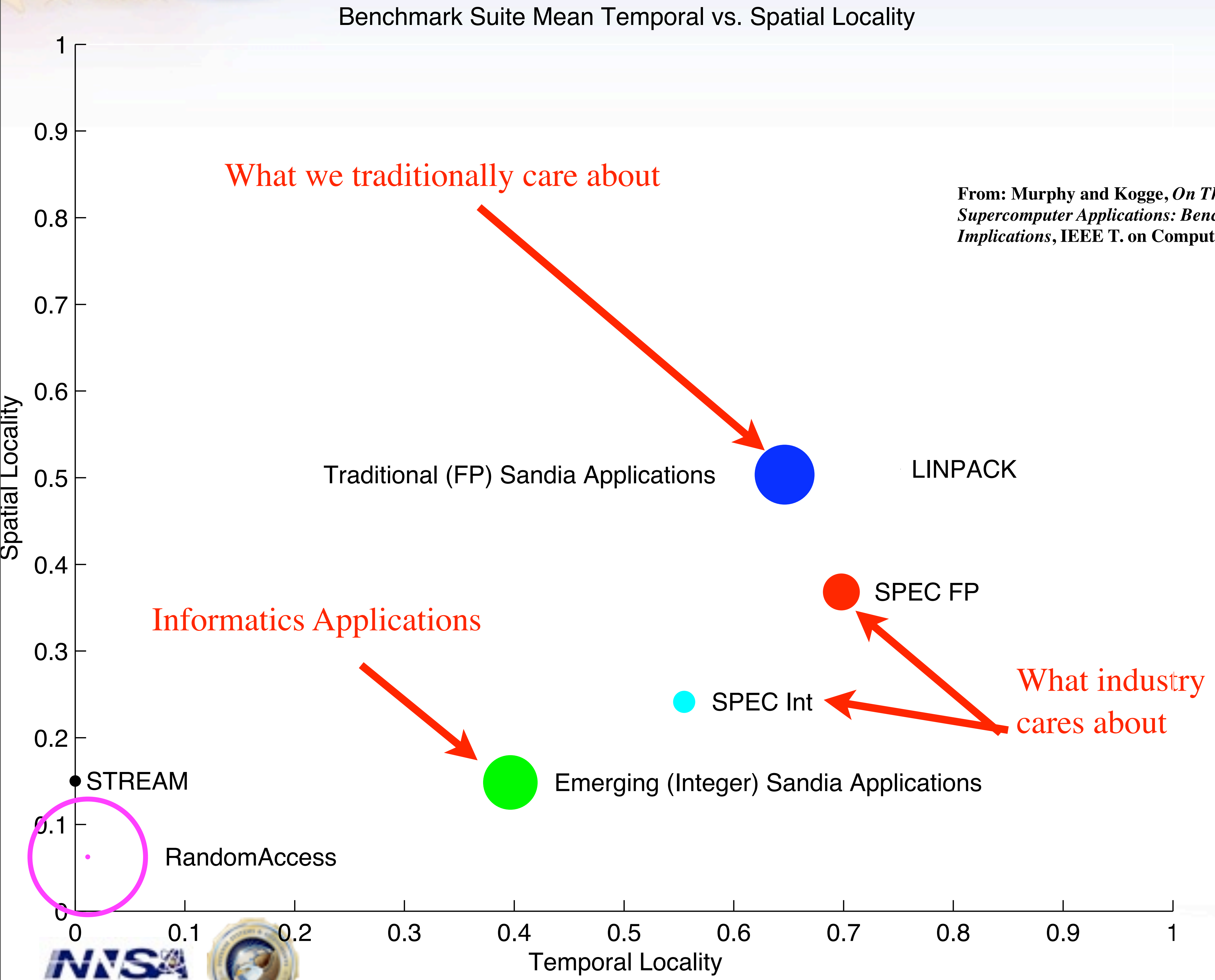
Integer/Floating Point Ratio

Sanida FP Averages 5.5X the number of Integer Operations/FLOP



Only Artificial Benchmark

There ISN'T Much Inherent Locality in Real Codes



What we traditionally care about

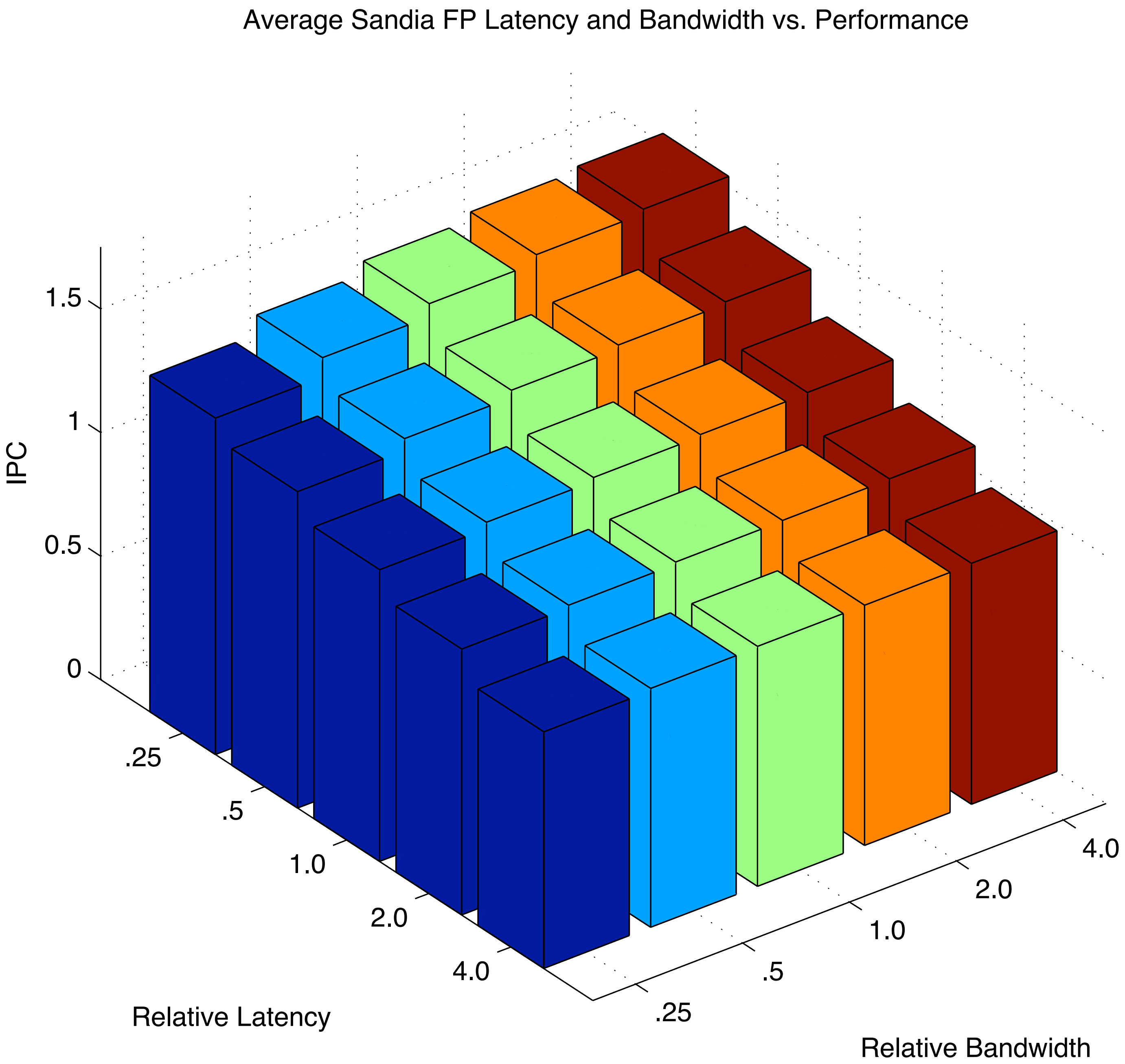
Informatics Applications

What industry cares about

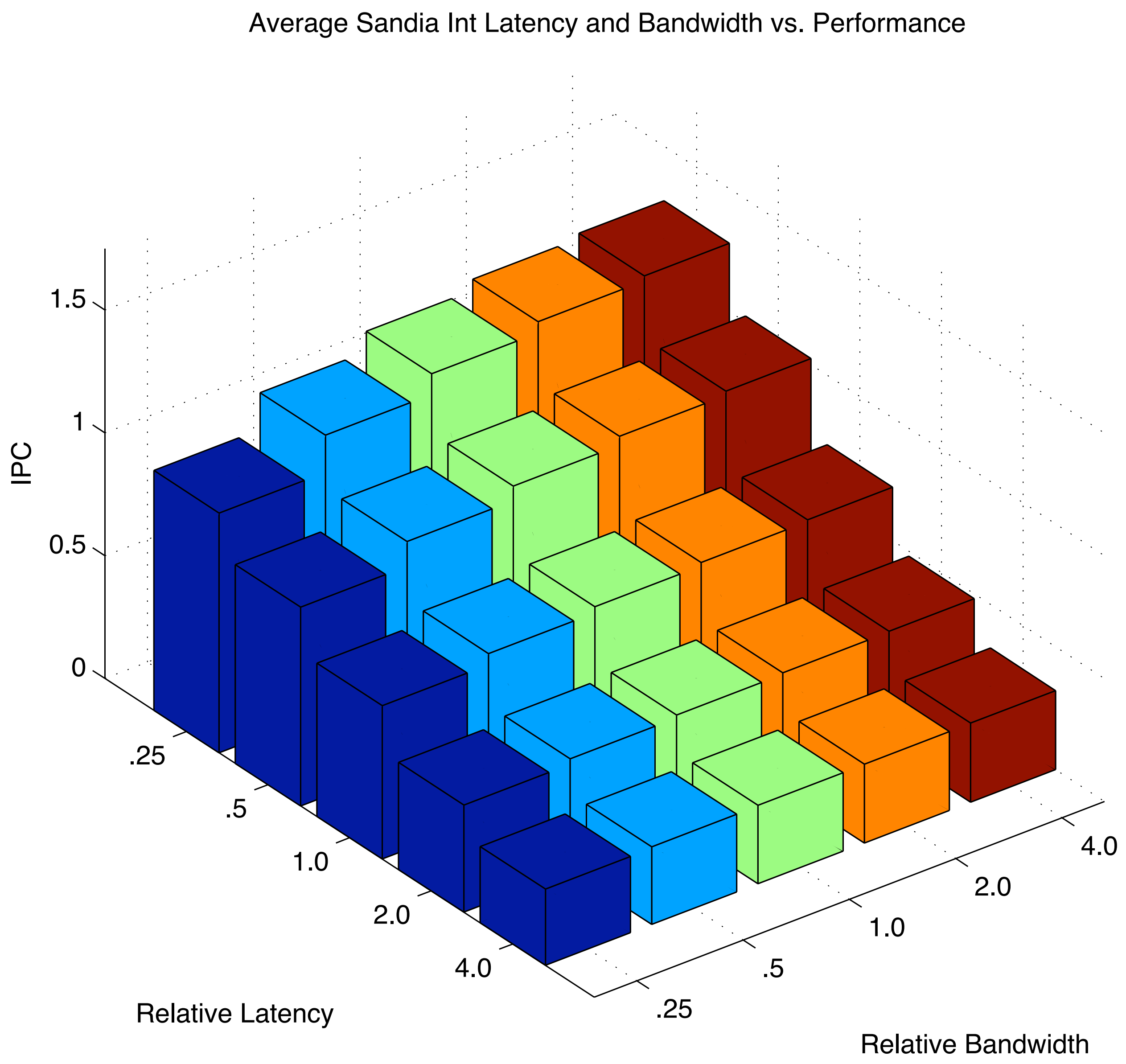
From: Murphy and Kogge, *On The Memory Access Patterns of Supercomputer Applications: Benchmark Selection and Its Implications*, IEEE T. on Computers, July 2007

Latency Dominates Bandwidth (Concurrency Decreases Effective Latency)

Physics



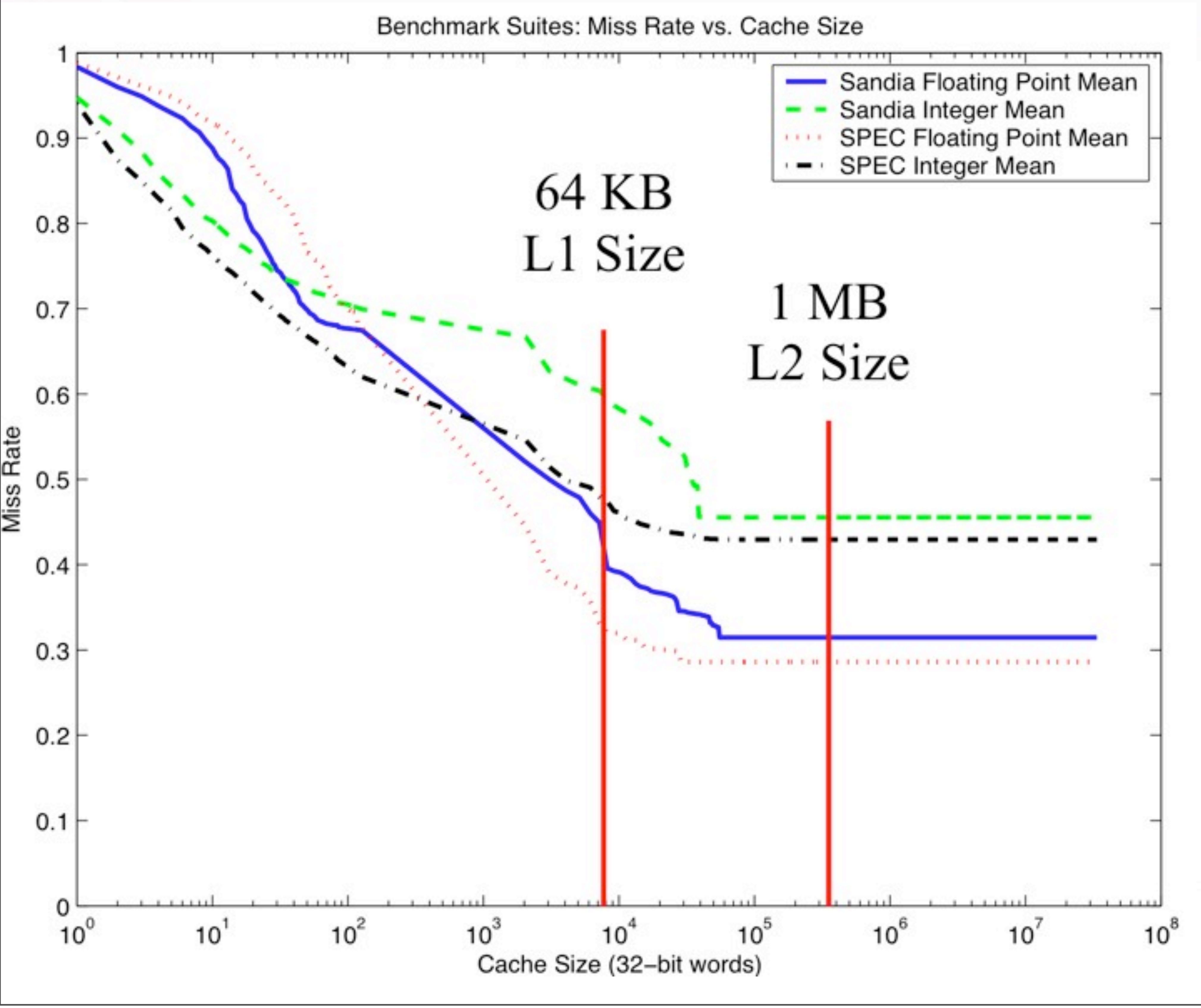
Informatics



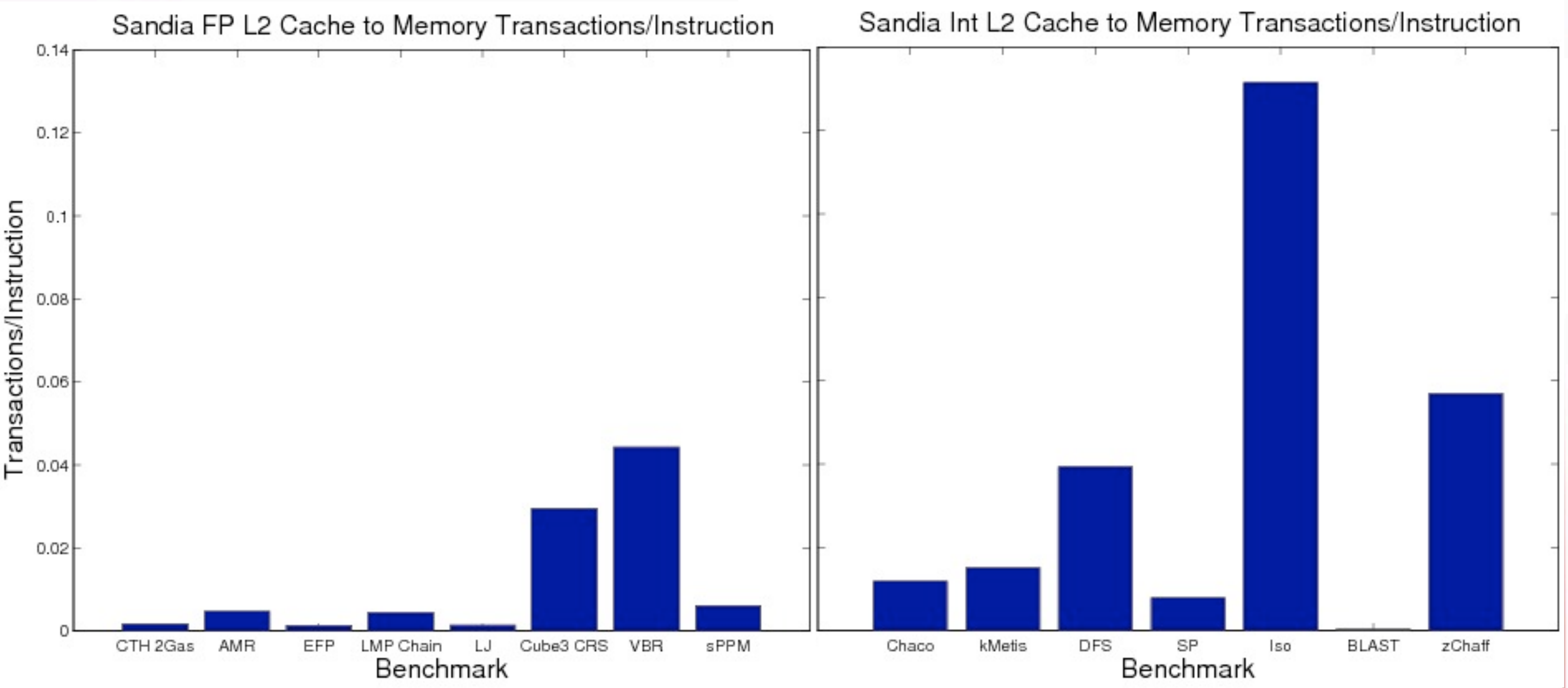
"Message Rate" more important than bandwidth

From: Murphy, *On the Effects of Latency and Bandwidth on Supercomputer Application Performance*, in the Proceedings of the IEEE International Symposium on Workload Characterizattion 2007 (IISWC07), Boston, MA September 27-29, 2007.

Temporal Miss Rate Results



Comparison to a Conventional Processor



Conventional L2 Cache to Memory Miss Rate

- L1: 64 KB 8-way Set Associative 64-byte Block with Write-Back

- L2 1 MB 4-way Set Associative 128-byte Block with Write-Back

Temporal reuse curve “flat” at 1 MB cache size!

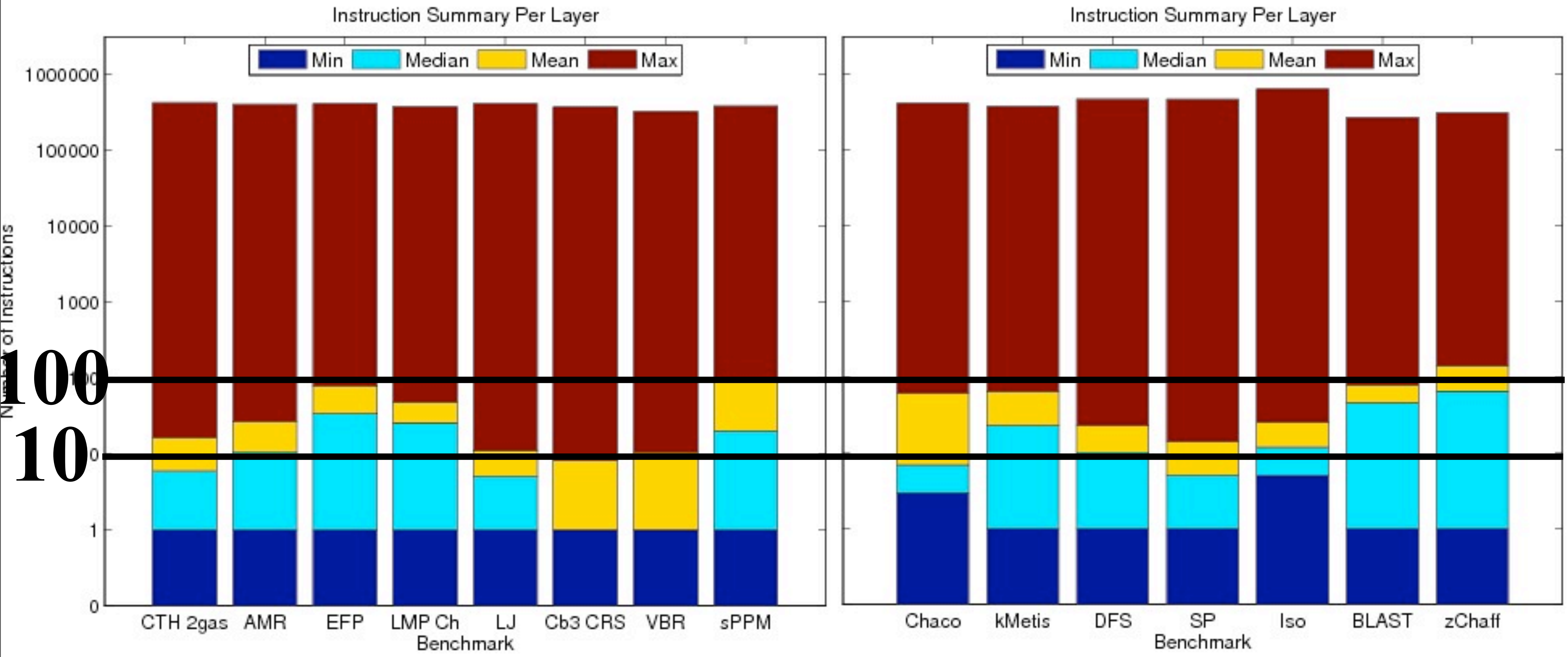
Conclusion: Managing Locality Will NOT Solve Our Problem... there isn’t much!

Concurrency

- Transition from a 3 or 4 GHz clock-rate to a 1-1.5 GHz clockrate
  - 2-4x more concurrency (units), similar performance at similar power
- Amdahl Fraction for strong scaling?
- Power budget transition from 125 MW to 25 MW
  - Another factor of 5
- Easily 10-100X more concurrency required to achieve the same performance at the required efficiency!
- PLUS three orders of magnitude more concurrency to get from peta to exa

Applications see a concurrency problem, not an energy problem!

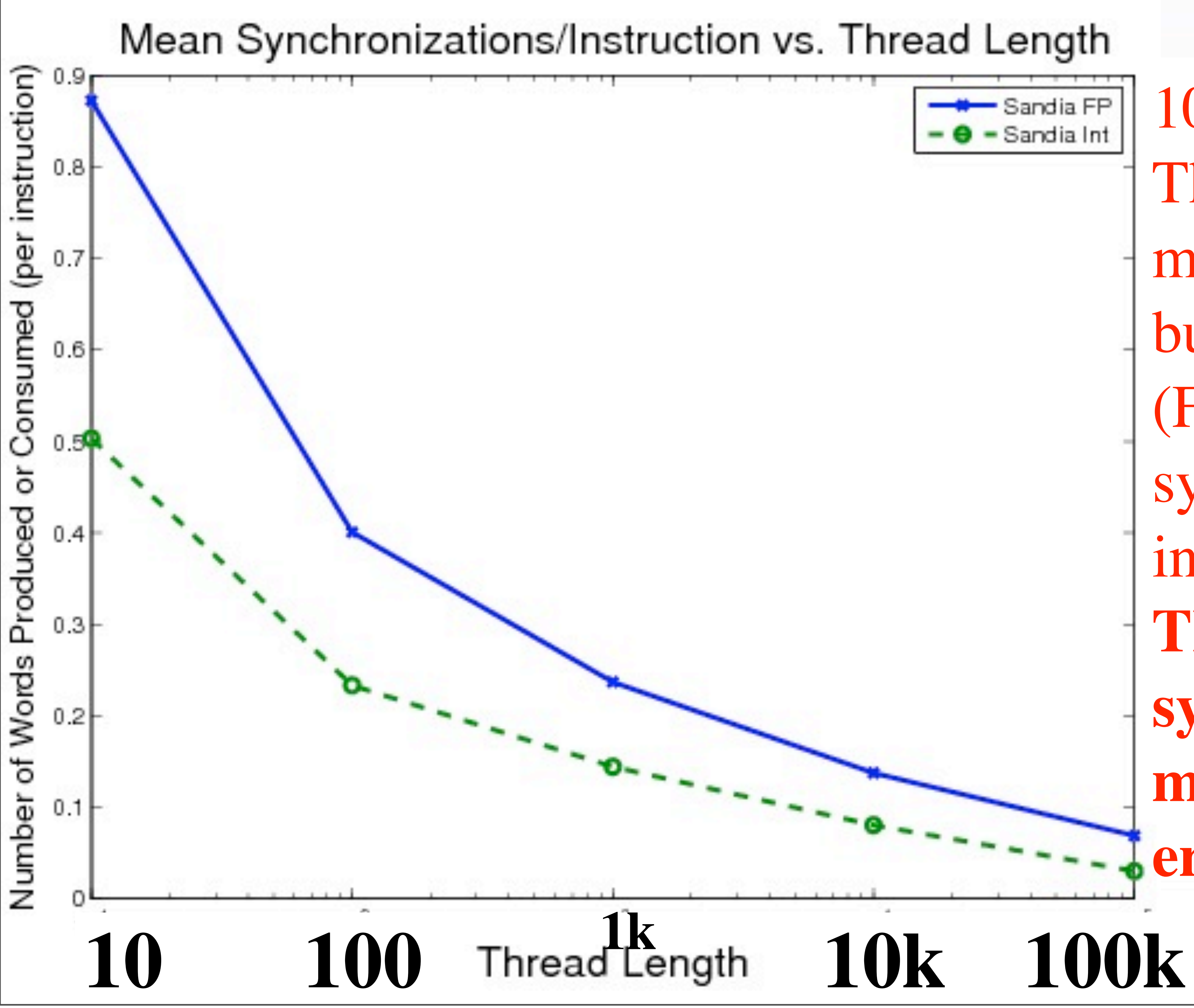
What concurrency is available?



We have the 2-order of magnitude "strong" factor, but that's optimistic!

What about the 3-order of magnitude weak factor?

At what synchronization cost?

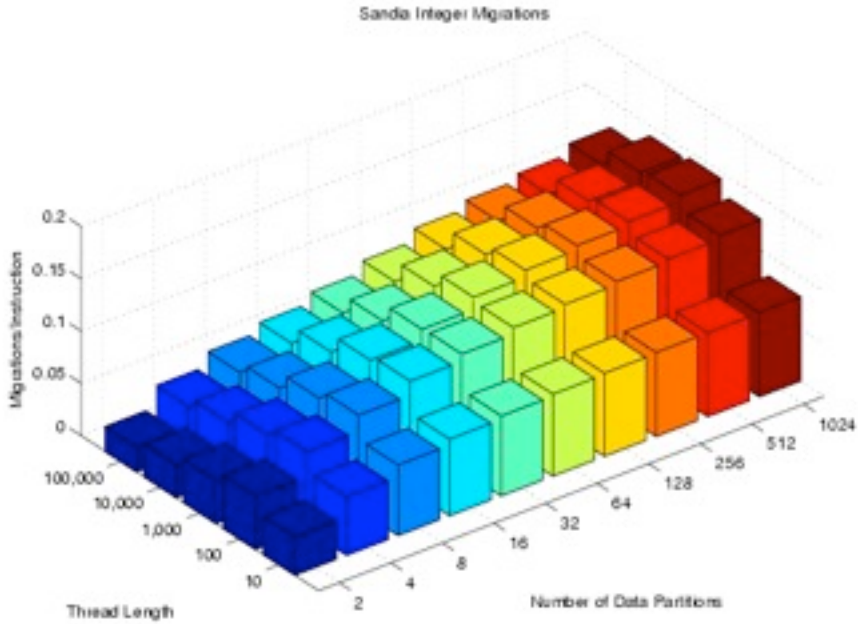


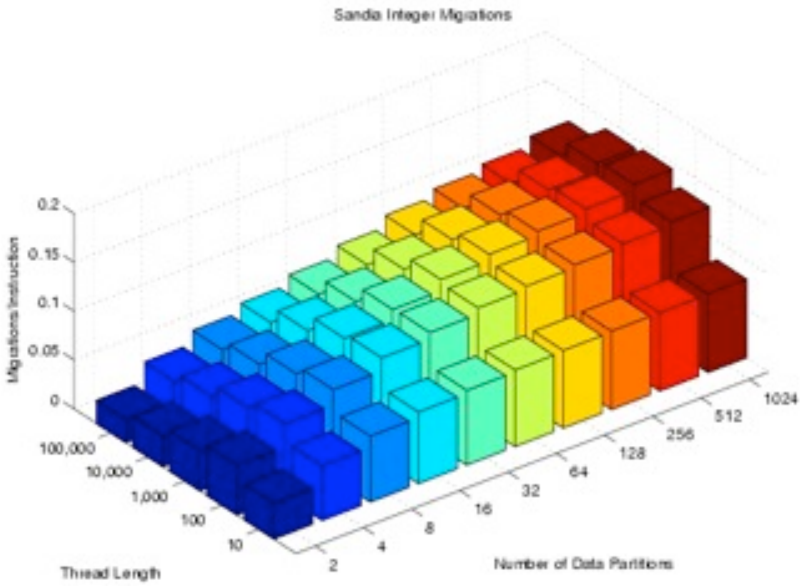
10 Instruction Threads offer more parallelism but require 0.9 (FP)and 0.5 (Int) synchronizations/ instruction

Therefore synchronizations must be cheap... err free!!!

Consequences

- Every core in the system has to know about every other core in the system
  - –We can’t afford the energy for today’s loose coupling
  - –We can’t manage the concurrency (even locally) with today’s model and today’s coupling -- the synchronization problem is too hard
- New models are required (work moving!)
  - –5-400x improvement in data movement over the application suite
    - THIS IS WHERE YOUR ENERGY GOES!
  - –Reduced thread state size (15% of a modern register file)





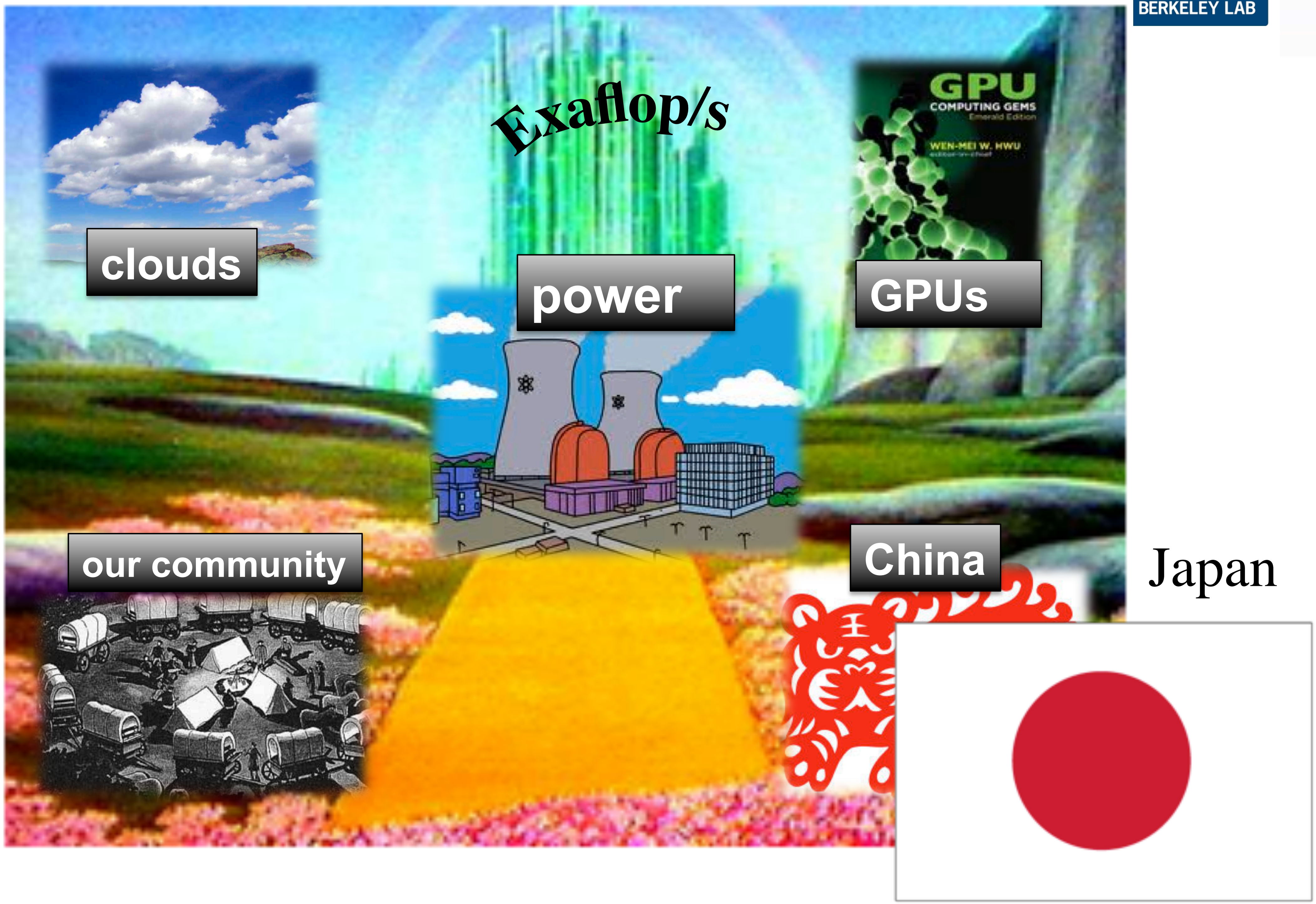
Weak Scaling and Balance

- System Balance
  - Because we're memory centric, we're focused on bandwidth, capacity, and scalability of the memory system (near and far)
  - X-caliber compared to the state of the art (scaled to 2018):
    - 5X the FLOPs of Red Storm (in the petascale rack)
    - 2X the memory capacity
    - Similar network bandwidth ratio
  - Other approaches (aggregate from what I've seen):
    - 10X the FLOPs of Red Storm, Half or less the memory capacity

System	Injection BW	FLOPS	B/F	Ratio	Comment
X-caliber	133 TB/s - 266 TB/s	1.0 - 1.4 PF/s	0.095 - 0.266	1.21 - 3.38	Adaptive
Typical Exascale Thinking	205 TB/s	2.6 PF/s	0.0788	0.82 - 0.30	Static

Horst Simon’s Distractions

The Road to Exaflop/s – four distractions and a road block



“Exascale’s too important to suck through a nuclear weapons straw”

- DOE should do it on behalf of the government
- Naming is the fundamental difference between DOE and DoD applications
  - 3D to ND transition requires efficient naming
  - I don’t care how you implement it (provided it’s energy efficient!)
  - Some of the energy budget has to go here
- This is where I believe the commercial opportunities are, e.g., Graph500 Business Areas
  - Cybersecurity, Data Enrichment, Medical Informatics, Social Networks, Symbolic Networks
- DoD will also push the address generating tasks requirement
- We need benchmarks that reflect this - GUPS not FLOPS

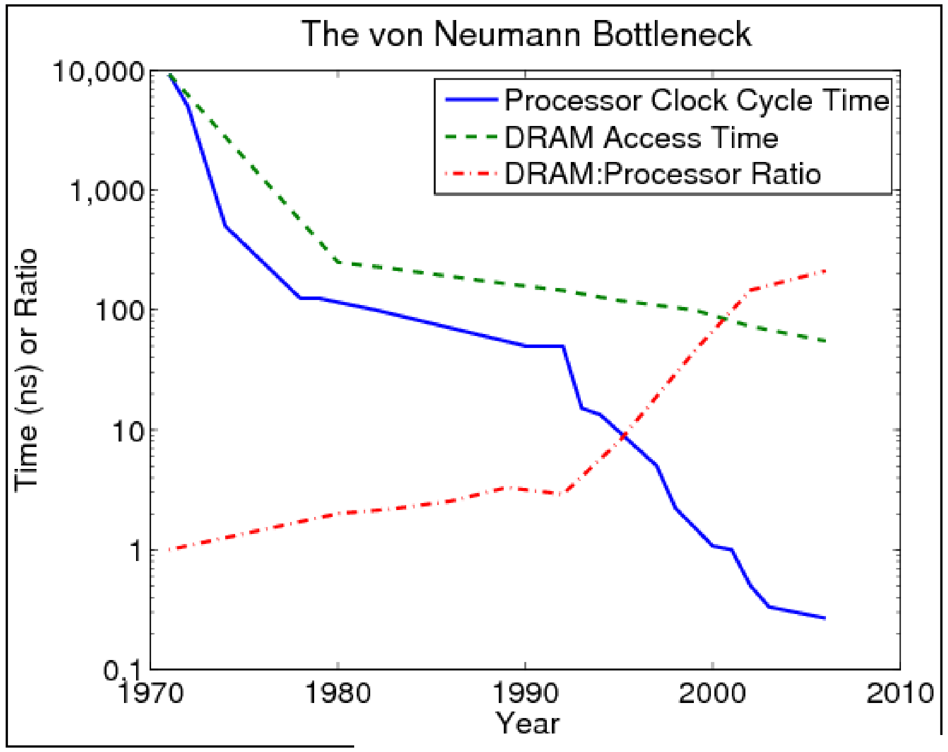
An Exascale Initiative in Five Thrusts

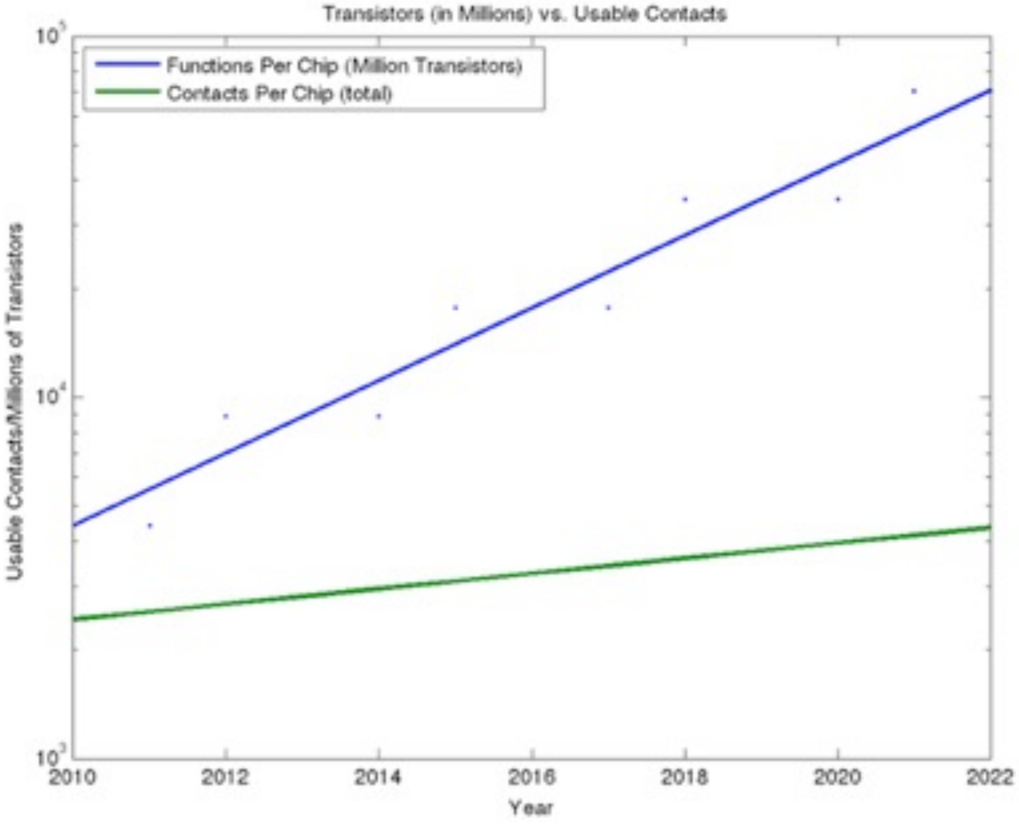
- Establish a Baseline: What happens if we do nothing?
- Devices and Integration
  - Enabling technologies for low-energy data movement: 3DI/Photonics
- Architecture
  - Optimize the architecture for energy and simplicity
  - Expose data movement and it's costs
- System Software
  - Create minimal interfaces to expose and manage data/work movement
- Applications
  - Rethink applications to be data-movement-centric instead of processor-centric
  - Define the ultimate metrics for evaluation
- Cross-cutting Theme: Codesign

3D Integration and Memory

- 3D Heterogeneous Integration
  - MPU, DRAM, Si Photonics
- Memory Challenges
  - Rich Atomics:
    - Fetch and Add (MTA)
    - Fetch and Other...
    - In-Memory Copy/DMA
    - BIST, refresh, scrubbing
    - Page Ops: Zero, Fill, etc.
    - CAS
    - Cray-like Gather/Scatter
    - Dereferencing Gather/Scatter (graph)
    - Synchronization (B+B, random)
  - PIM (increase address generators, and do it where power-efficient!)

$$\text{Throughput} = \frac{\text{Concurrency}}{\text{Latency}}$$





Technology Investments: 3D Integration and Silicon Photonics

$$\boxed{\frac{8MW}{8*10^{18}b/s} \approx 1pJ/bit}$$

Modulation (E to O):

- **3 fJ/bit** modulation has been demonstrated[3]

Modulator thermal control and trimming:

- 106 GHz within die frequency variation has been measured and a thermal resonance shift of 4.4 μW/GHz has been demonstrated[1] → **<23 fJ/bit>** thermal trimming power
- 4.4 fJ/bit-°C thermal control has been demonstrated[2], so for +/- 10 °C swing in operating temperature → **<44 fJ/bit>** thermal control power

Optical Demux:

- 2 ring-filter with thermal control and trimming → **<134 fJ/bit>**

Receiver (O to E):

- A -18.9 dBm sensitivity integrated Ge receiver has been demonstrated[4] at 5 Gbps, with a total power consumption of **690 fJ/bit**

Optical Source:

- -18 dBm power required at receiver. Assuming 2 dB/facet coupling loss (demonstrated) and another 2 dB of on-chip loss then for a wall plug efficiency of 10 % (includes TEC power consumption) then 1 mW is required for the optical source → **100 fJ/bit**

Total power consumption: $\boxed{\text{~1 pJ/bit}}$

[1] W. Zortman, *et al*. to be published

[2] C.T. DeRose, *et al*. CLEO (2010)

[3] W. Zortman, *et al*. CLEO (2010)

[4] Zheng, *et al*. Opt. Exp. **18** pp 204-211 (2009)

Final Thoughts: Memory Abstraction

- Lots of simple memory operations should occur locally
  - Memory Controller Functions
  - Error Correction and Management
  - Hierarchy Abstraction
    - NVRAM to boost bytes/core
    - DRAM to boost performance
- The critical system-visible change
- Requires tight coordination with the NIC

Thank You