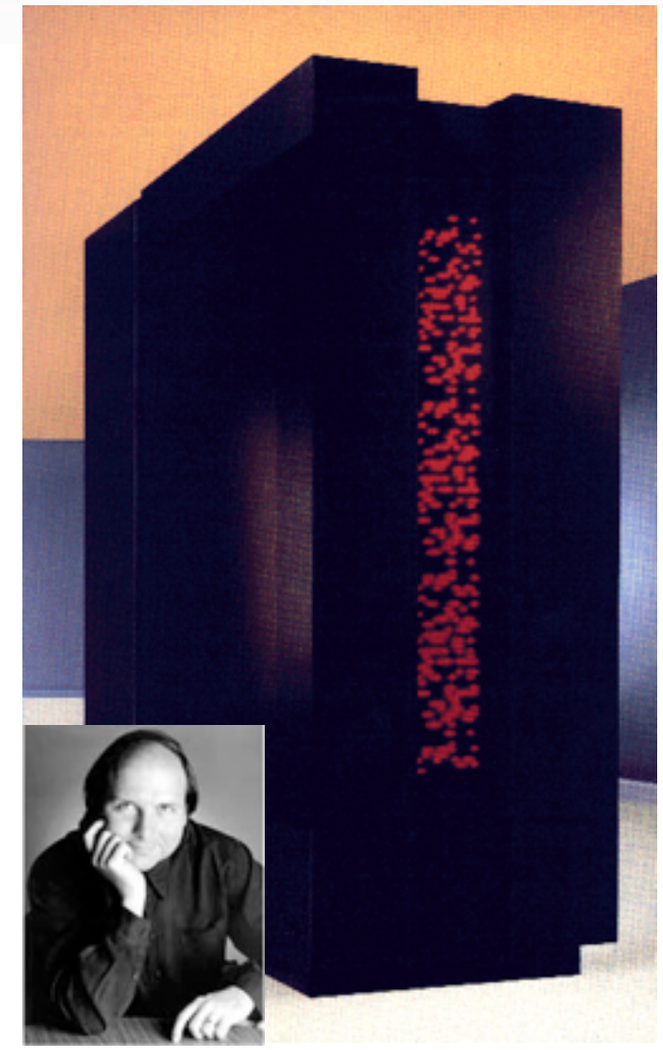
X-caliber and Exascale Grand Challenge Research Summary

Richard C. Murphy
Scalable Computer Architectures Department
Sandia National Laboratories
Affiliated Faculty, New Mexico State University

September 16, 2011

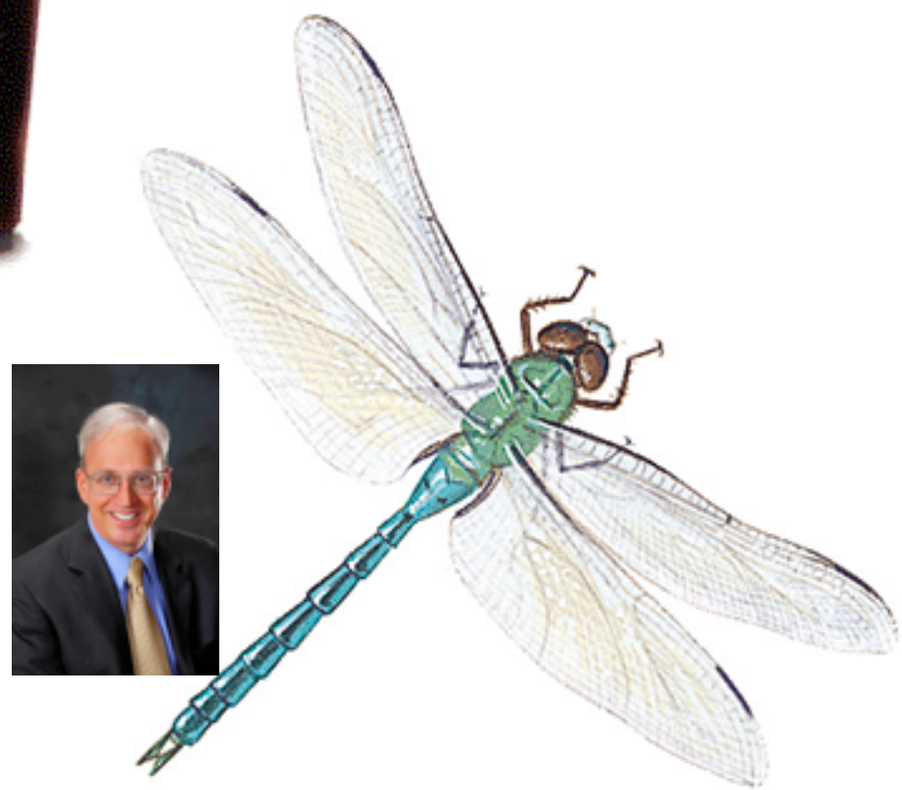
Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company,
for the United States Department of Energy's National Nuclear Security Administration
under contract DE-AC04-94AL85000.

An Architect's Worry about the Exascale Era: As Told by MIT Alumni



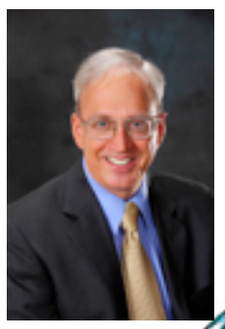
Step 1: Choose your favorite MIT Alumni's Computer from the early 1990s as your compute node... errr swim lane

An Architect's Worry about the Exascale Era: As Told by MIT Alumni (continued)



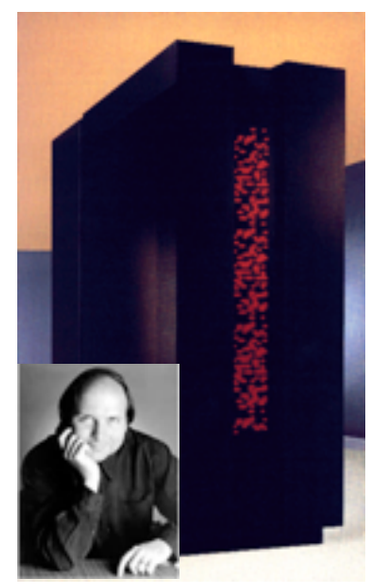
Step 2: Pick your favorite MIT Alumni's network topology and wire up whatever bandwidth you think you can afford

An Architect's Worry about the Exascale Era: As Told by MIT Alumni (continued)



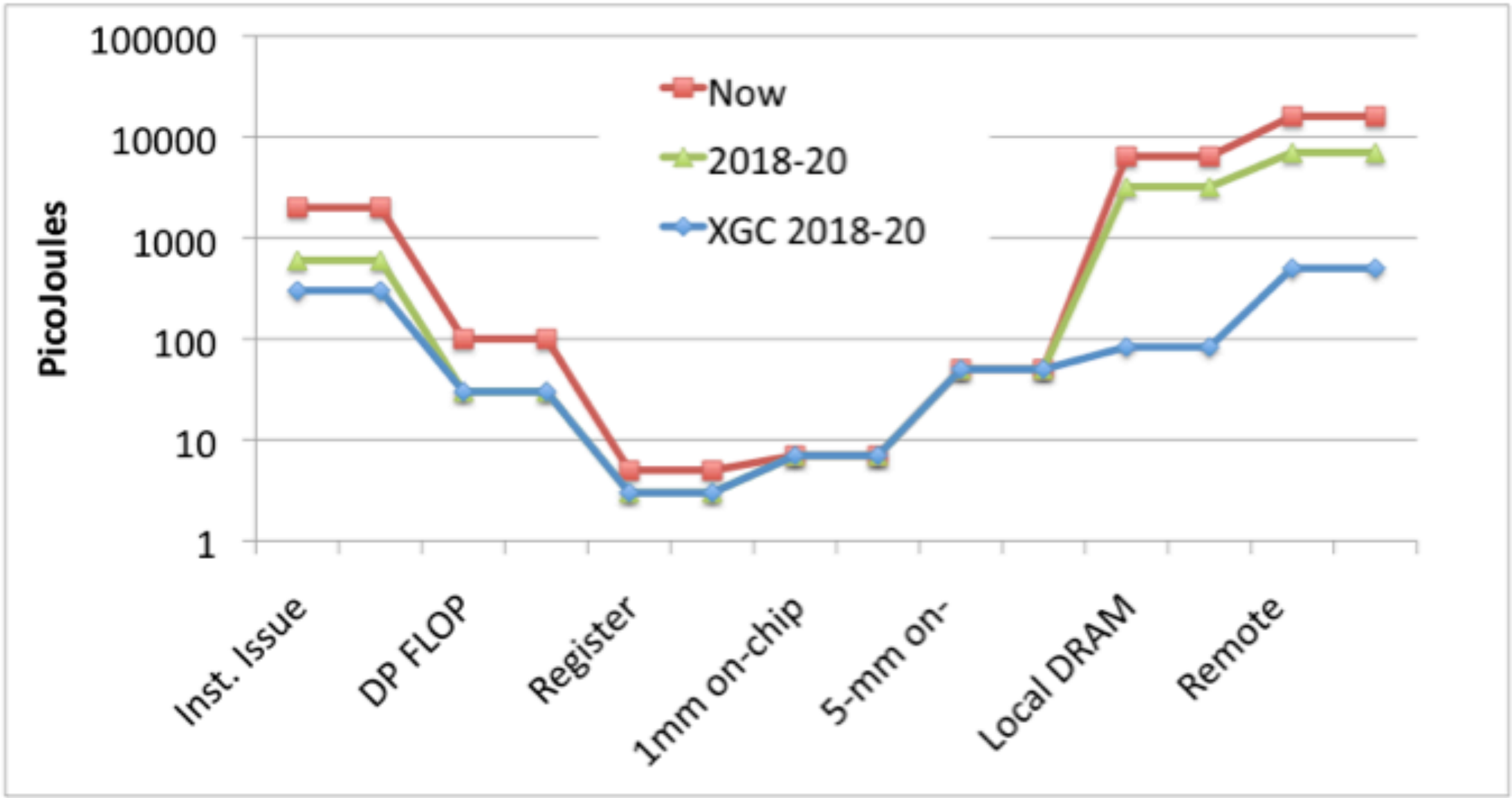
Step 3: Mix in MPI... and a touch of your favorite MIT Alumni's Alternative Programming Model

An Architect's Worry about the Exascale Era: As Told by MIT Alumni (continued)



We all know this approach is subject to criticism...

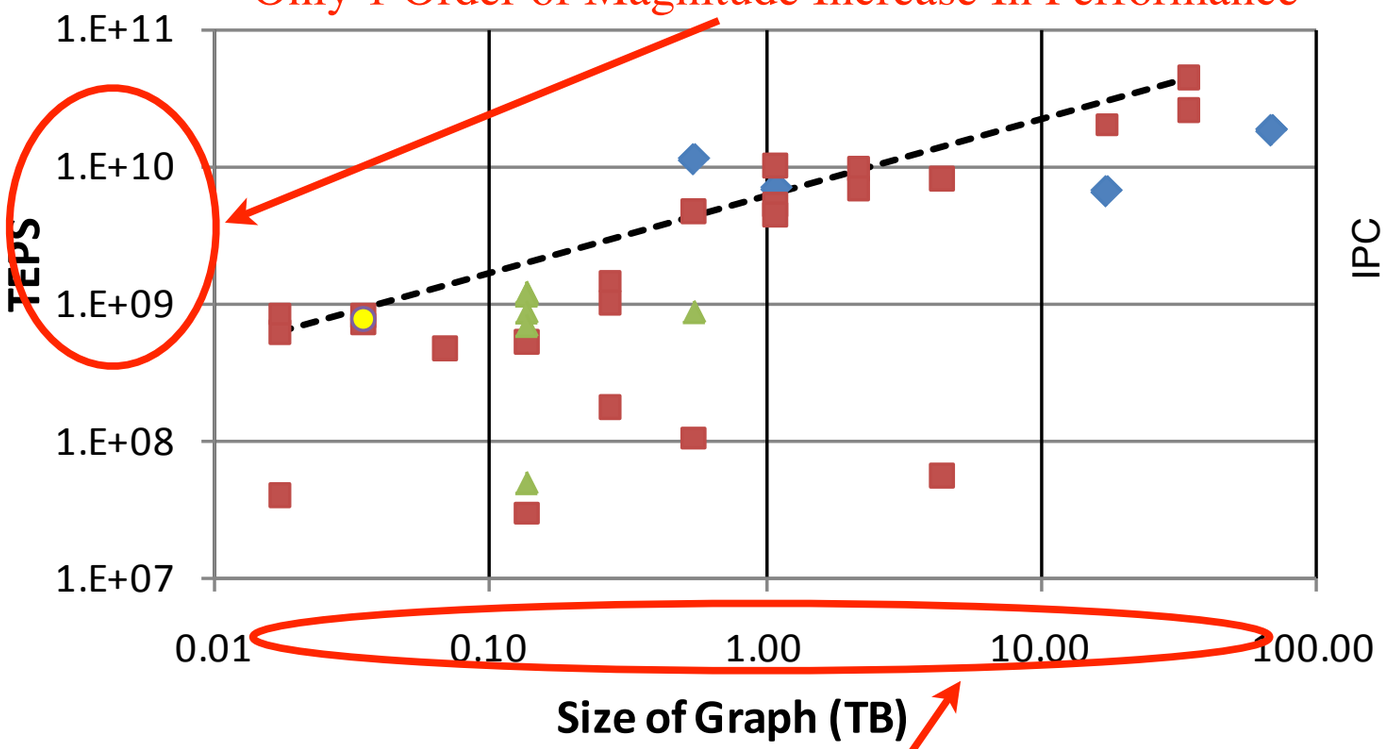
Motivation


100000
10000
1000
100
10
1
PicoJoules
Now
2018-20
XGC 2018-20
Inst. Issue
DP FLOP
Register
1mm on-chip
5-mm on-
Local DRAM
Remote

Memory Trends are Terrible... Memory Determines Performance

TEPS vs Problem Size

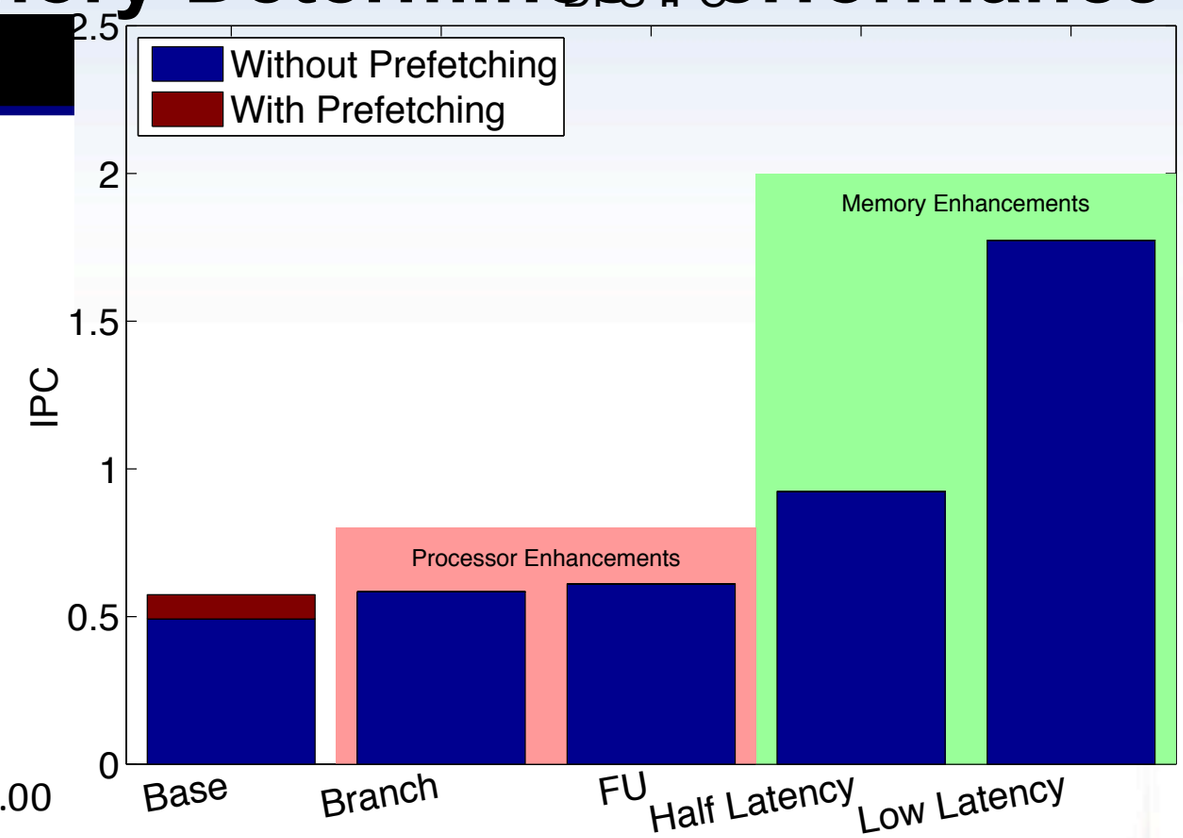
Only 1 Order of Magnitude Increase In Performance

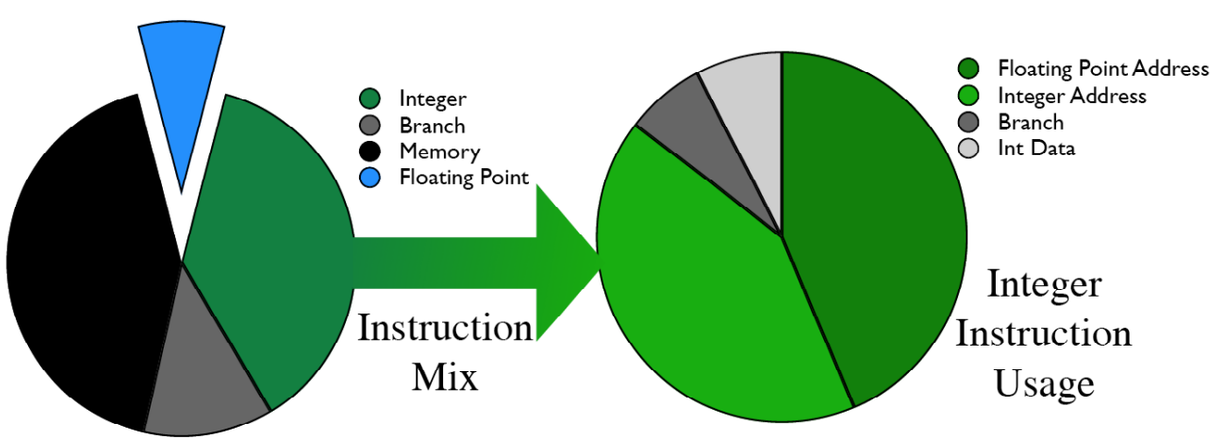


BlueGene x86 XMT Other 1.2B TEPS/TB

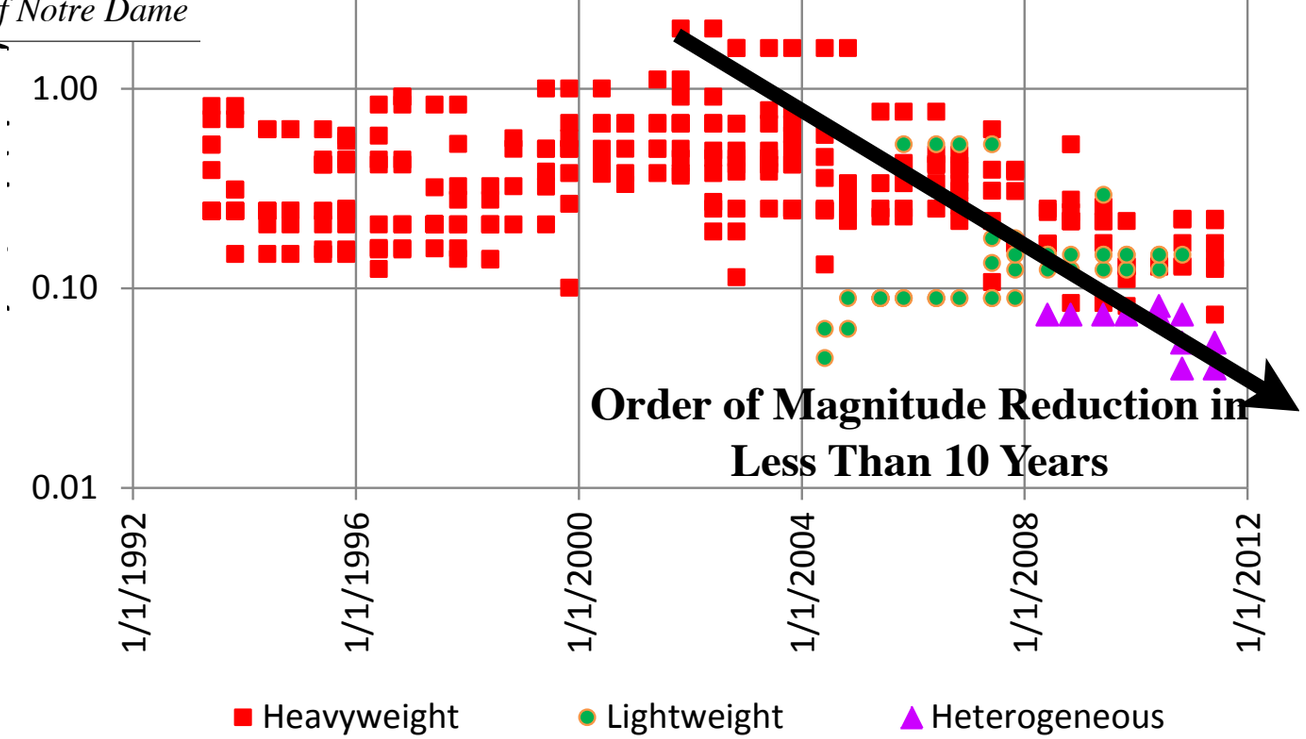
3 Order of Magnitude Increase in Problem Size

Data Courtesy of Peter Kogge, University of Notre Dame





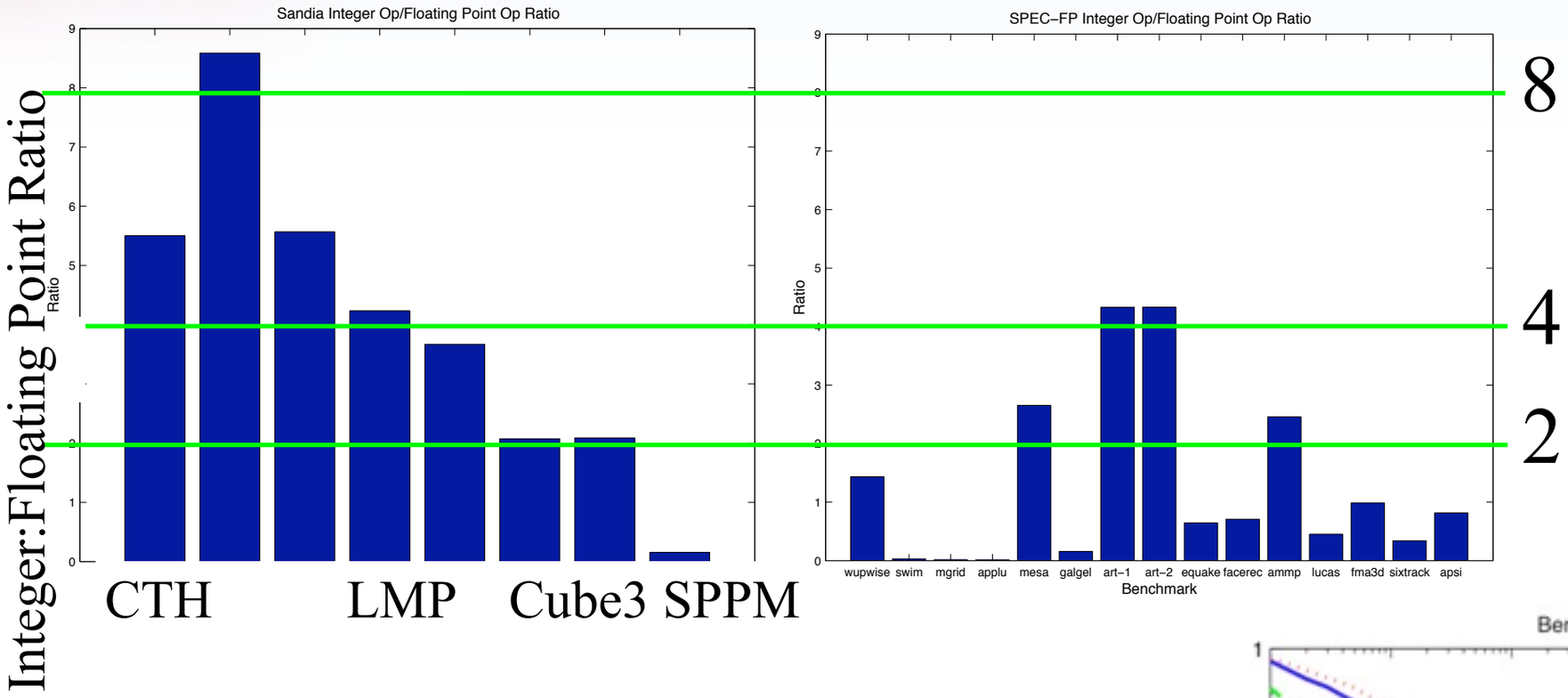
Memory Capacity Trend



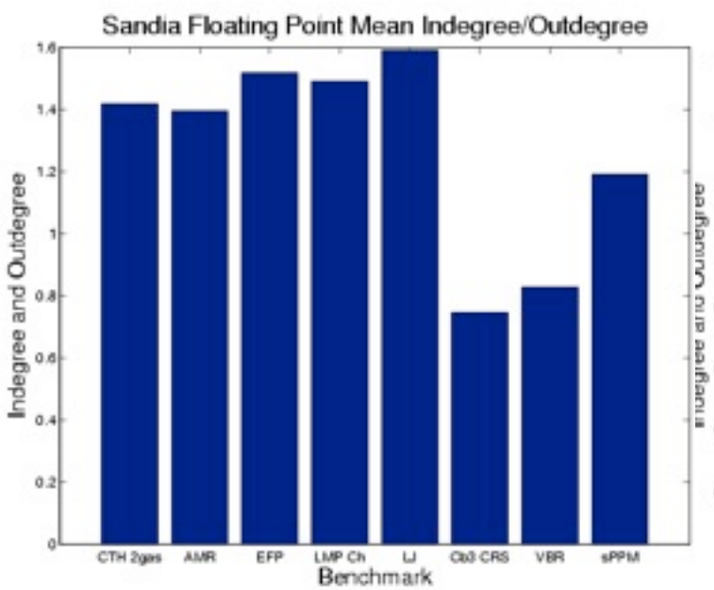
UNIVERSITY OF NOTRE DAME

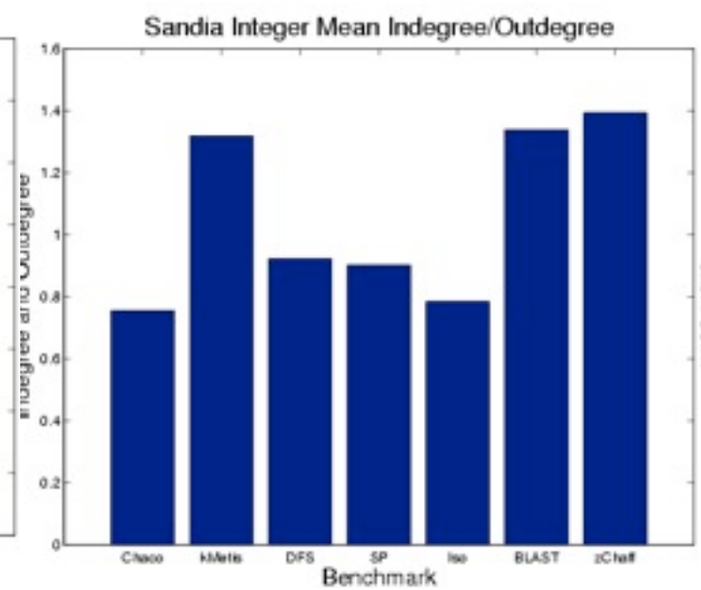
Application Properties: FLOPS and Locality

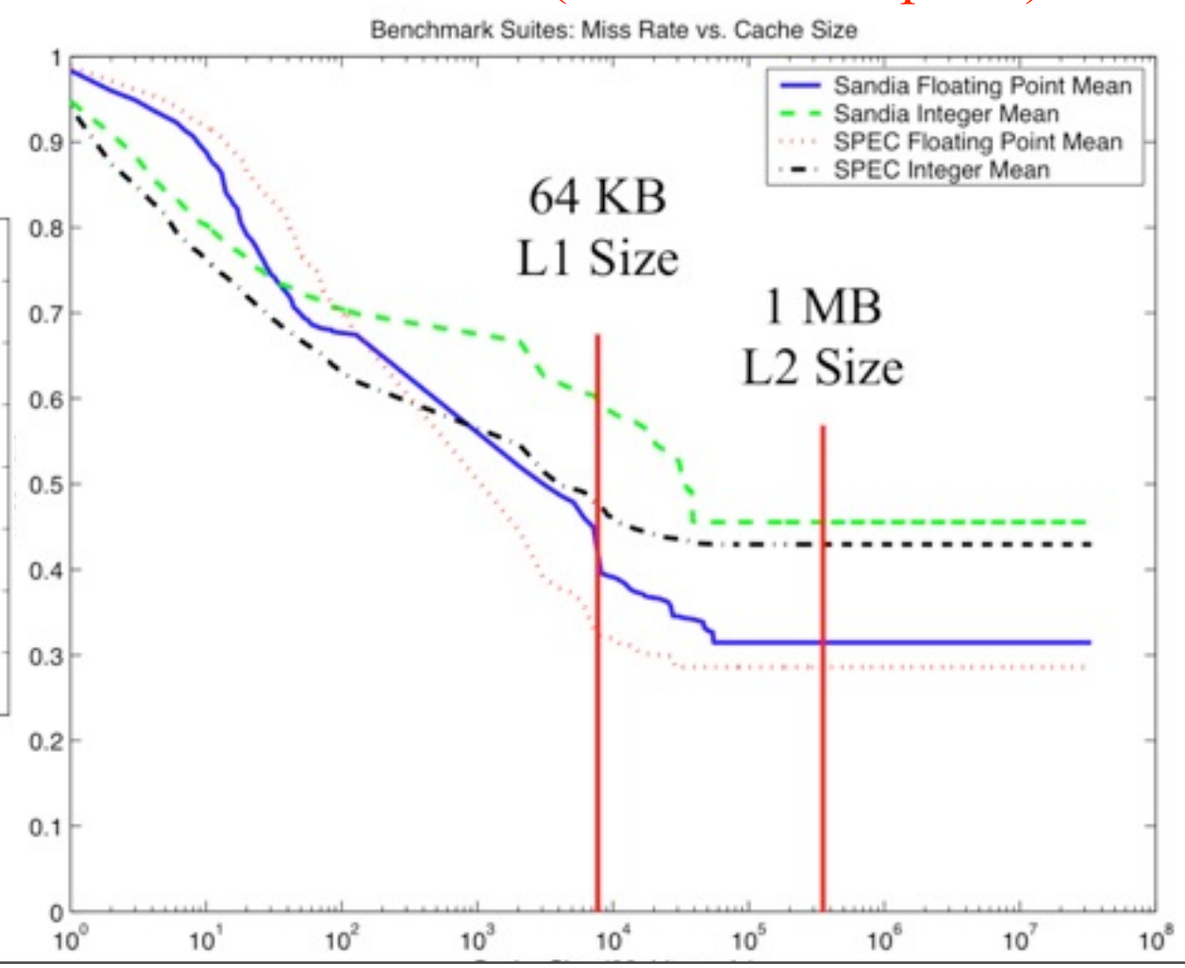
Real Apps Average 5.5X the number of Integer Operations/FLOP



Very little Temporal Locality
(64k - 1MB capture)

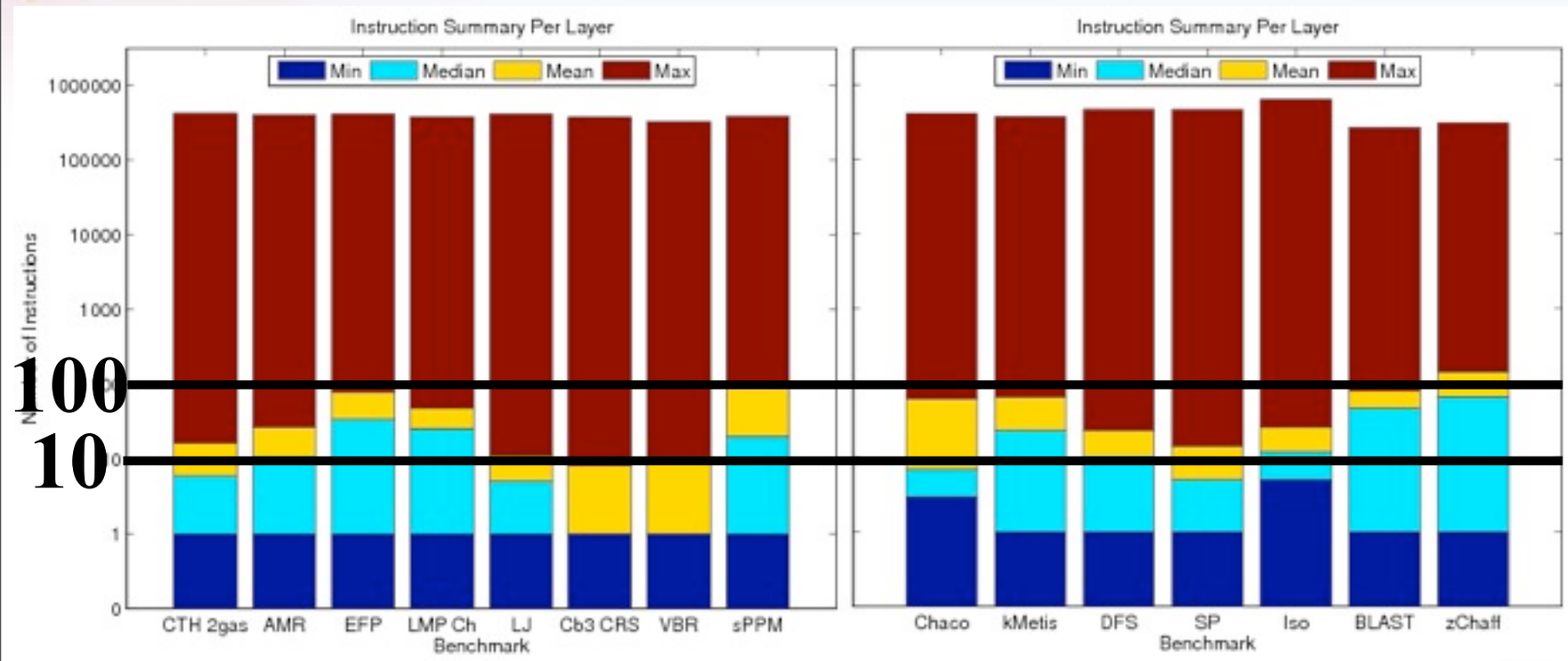






VERY FEW HIGHLY SHARED VALUES
MEDIAN "USE DISTANCE" IS 1.0

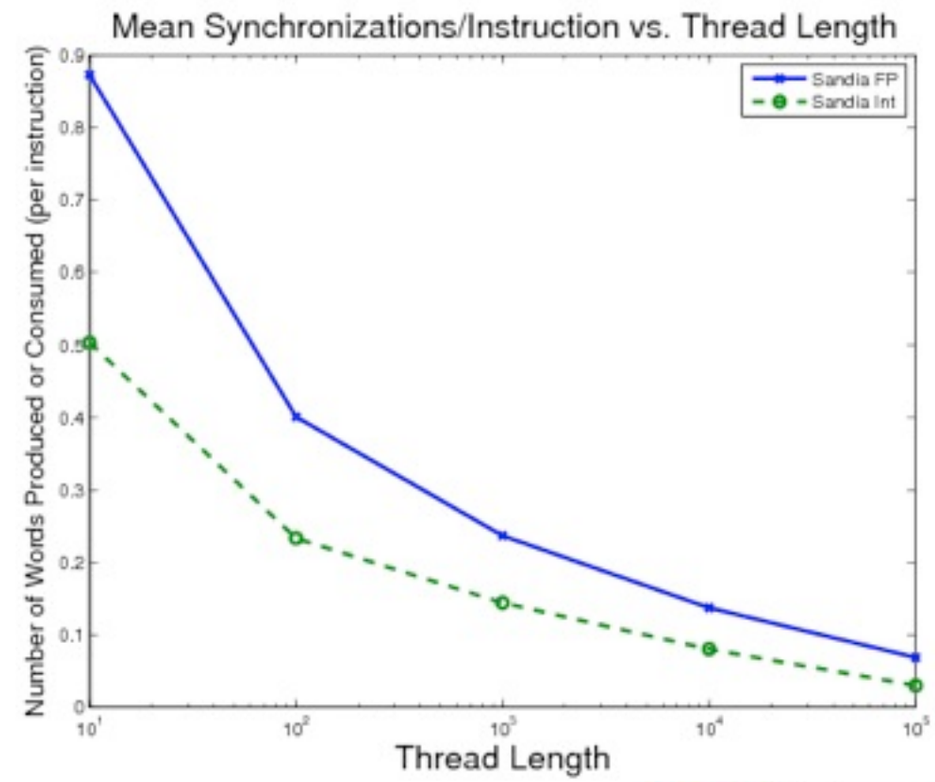
Application Properties: Concurrency



Strong Scale: 2 orders of magnitude to be more power-efficient

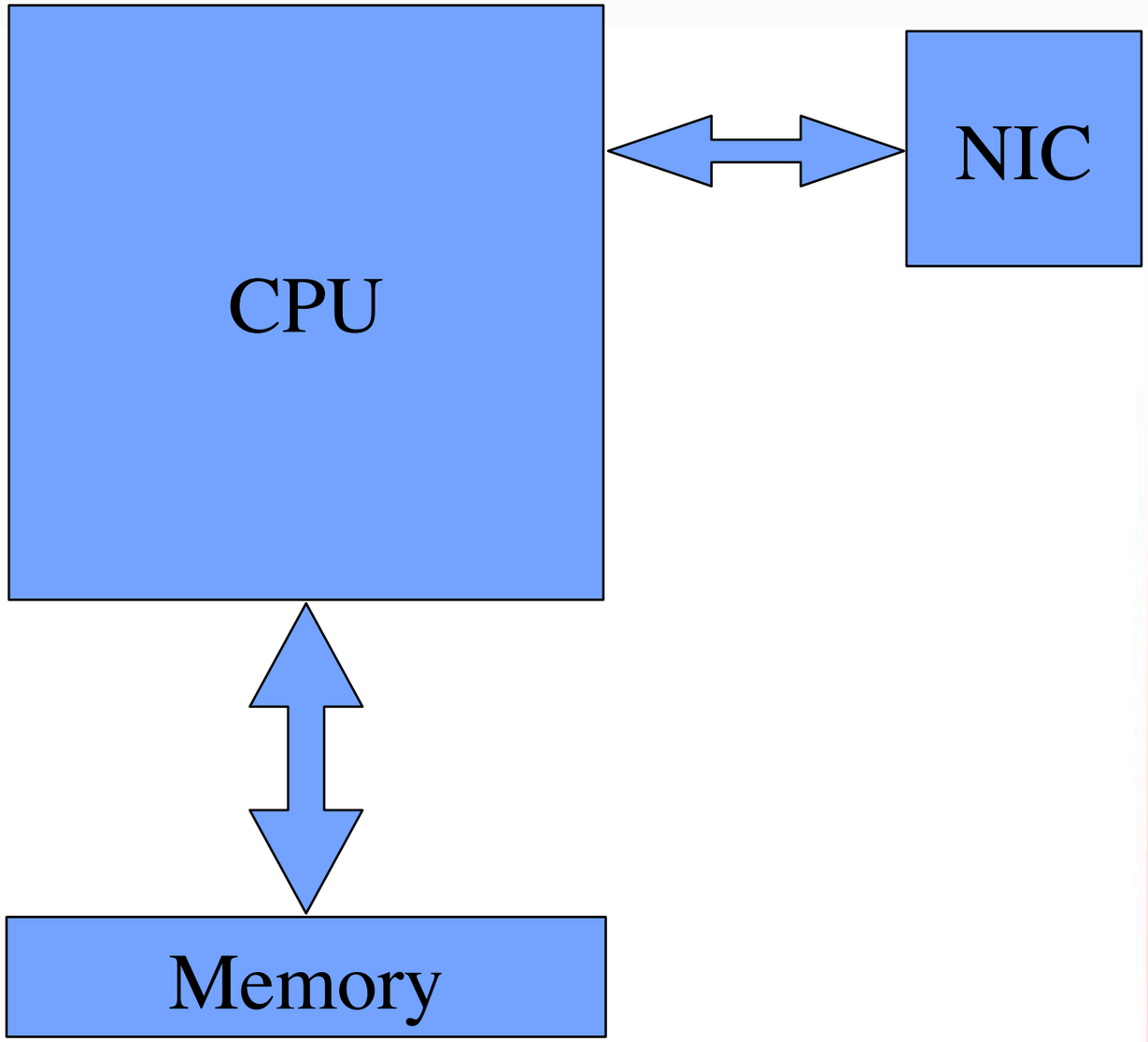
Weak Scale: the rest (3+)

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



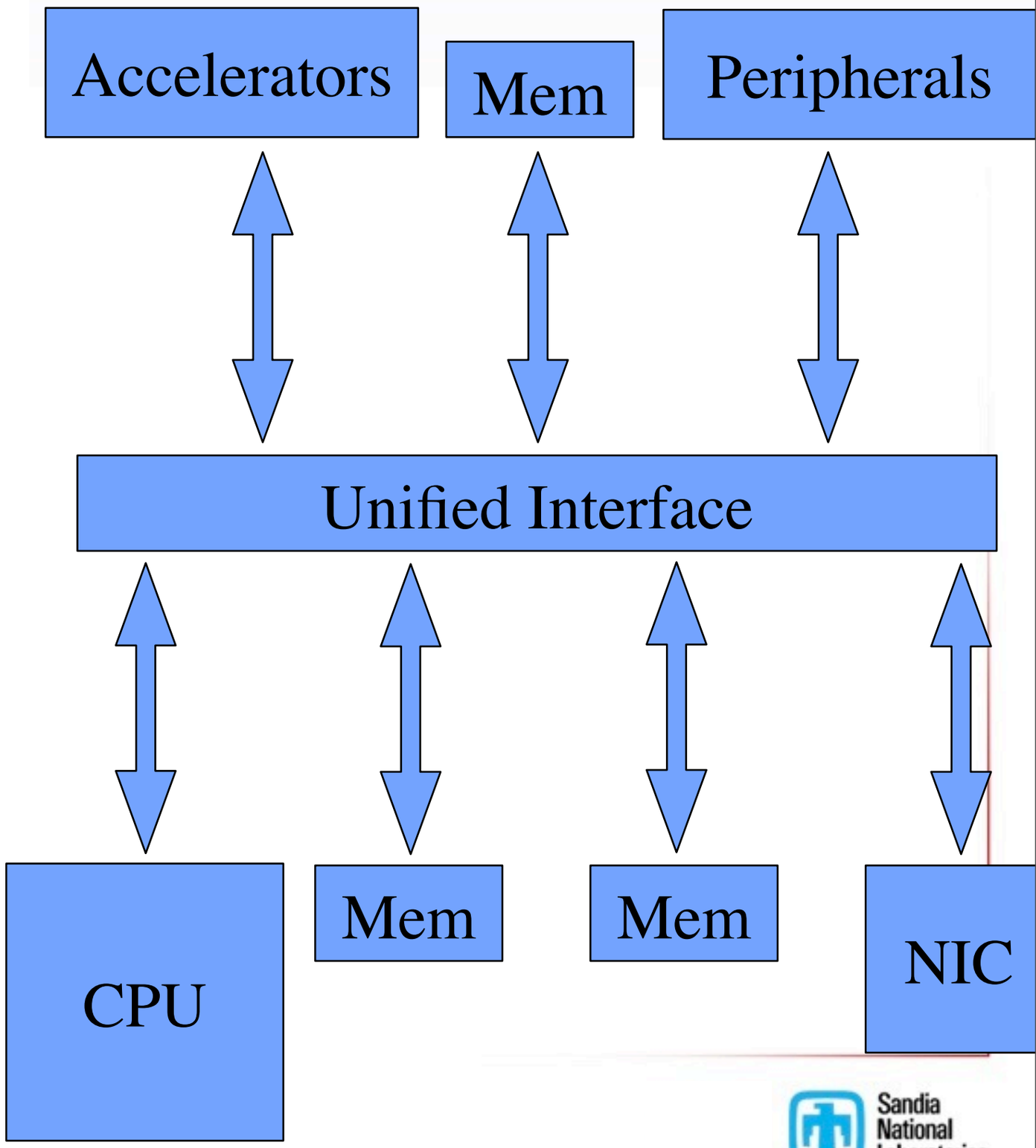
Today's Main Supercomputer Architecture Challenge

- Cannot source and sync the network and CPU simultaneously
  - Because the CPU goes to memory every:
    - 100 instructions (on a good day)
    - < every 10 instructions (for a "hard" problem)
- Commercial accelerators suffer the same problem
- "Data" devices are equally starved (HDDs, SSDs, etc.)
  - Real industry (e.g., "cloud") problems are entirely I/O bound with < 2% processor utilization



We Need a Radical System Reorganization... and memory abstraction is the key

- We need to be able to source/sync lots of data to/ from the memory system independently
- Memory abstraction is required
  - Can be distributed among memory devices
  - Other alternative is centralized through the processor
  - Memory's the only space with enough "state" to hold the current state of all memory words
  - Creates a more composable system

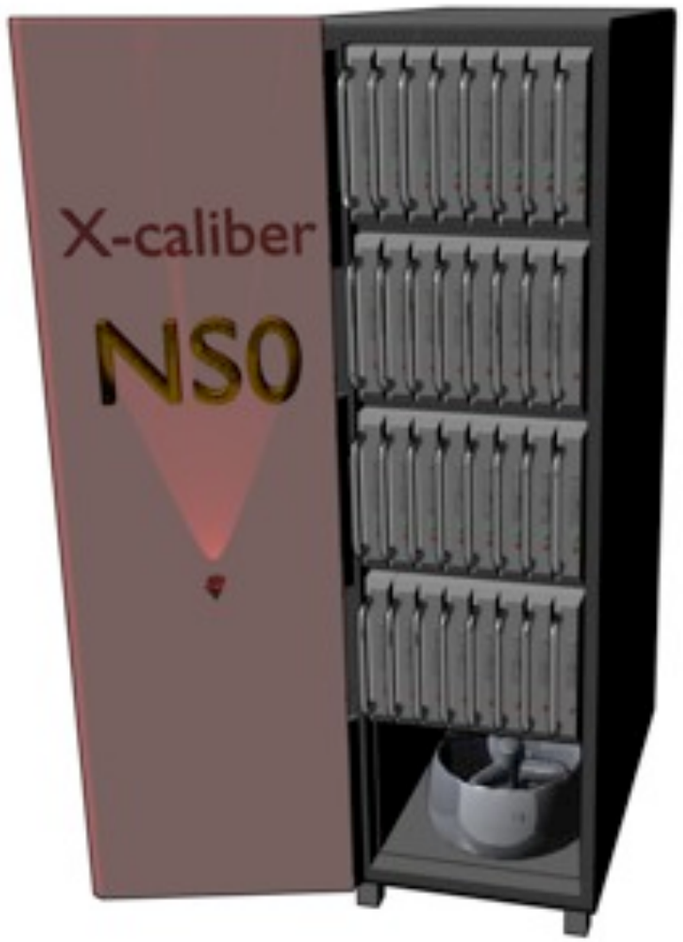


Sandia National Laboratories

But you're a supercomputer guy... there's no faster path to bankruptcy than trusting a supercomputer guy

- Commercially (and in the government) compute capability doubles every 14 months
  - But data doubles every 9
  - Is this the "data wall"?
- Five Business Areas You Should Not Ignore
  - Cybersecurity
  - Medical Informatics
  - Data Enrichment
  - Social Networks
  - Symbolic Networks
- New benchmark for "big-data" problems: graph500
  - www.graph500.org

X-caliber Rack Scale Prototype



- **Rack Scale**
  - **Processing:128 Nodes, 1 (+) PF/s**
  - **Memory: 128 TB DRAM**
    - **0.4 PB/s Aggregate Bandwidth**
  - **Network**
    - **0.13 PB/sec Injection, 0.06 PB/s Bisection**
- **Two outputs of the project:**
  - **Advanced Memory System**
    - **Embedded Memory Processor**
  - **ParalleX Execution Model**
    - **Sandia ParalleX Runtime (SPR)**

Deployment	Nodes	Topology	Compute	Mem BW	Injection BW	Bisection BW
Module	1	N/A	8 TF/s	3 TB/s	1 TB/s	N/A
Deployable Cage	22	All-to-All	176 TF/s	67.5 TB/s	22.5 TB/s	31 TB/s
Rack	128	Flat. Butterfly	1 PF/s	.4 PB/s	0.13 PB/s	0.066 PB/s
Group Cluster	512	Flat. Butterfly	4.1 PF/s	1.6 PB/s	0.52 PB/s	0.26 PB/s
National Resource	128k	Hier. All-to-All	1 EF/s	0.4 EB/s	0.13 EB/s	16.8 PB/s
Max Configuration	2048k	Hier. All-to-All	16 EF/s	6.4 EB/s	2.1 EB/s	0.26 EB/s

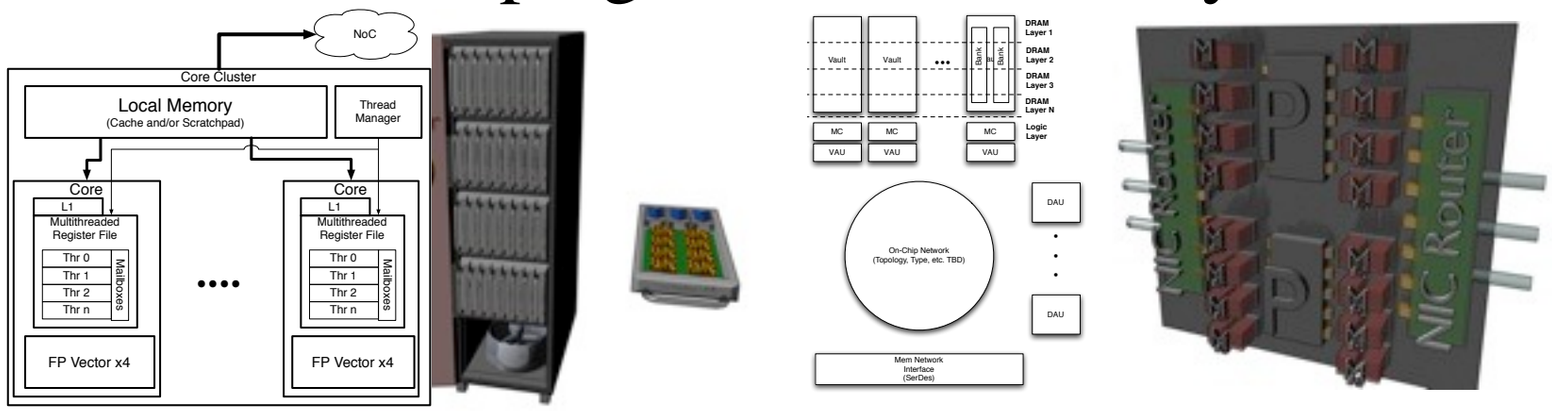
XGC: Approach via Five Thrust Areas

Safety and Security | Reentry | Circuity | Graph | Stream

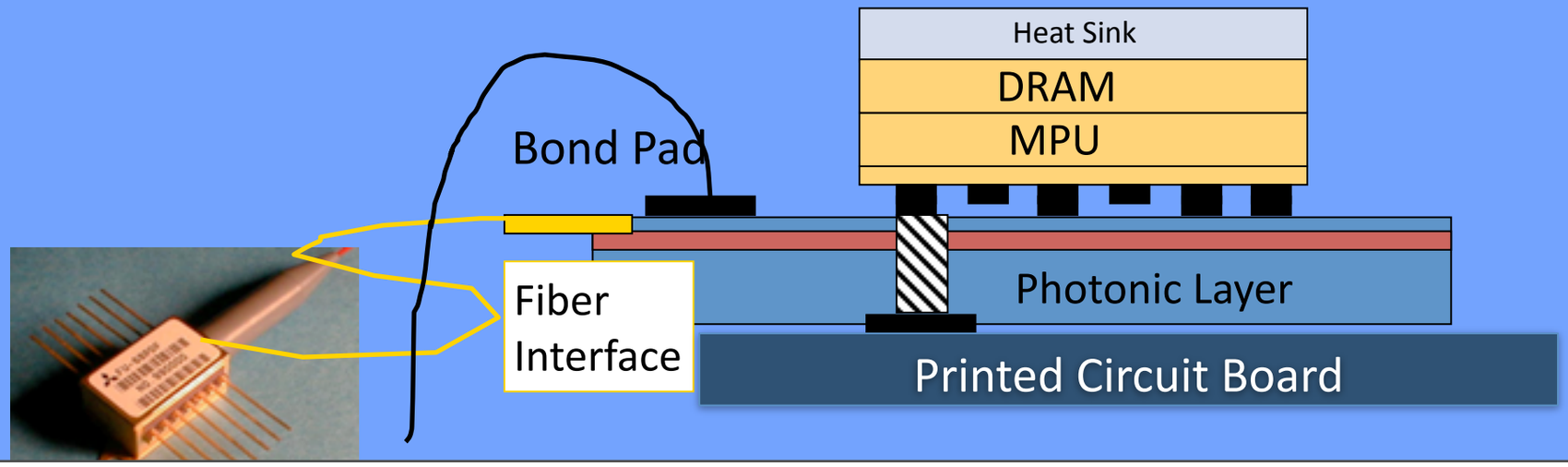
(5) Application Drivers

(4) System Software - enabling a new model of computation

(3) Architecture - Coping with Concurrency and Data Movement



(2) Microsystems - Key Data Movement Enabling Technologies



(1) Baseline - what happens if we do nothing?

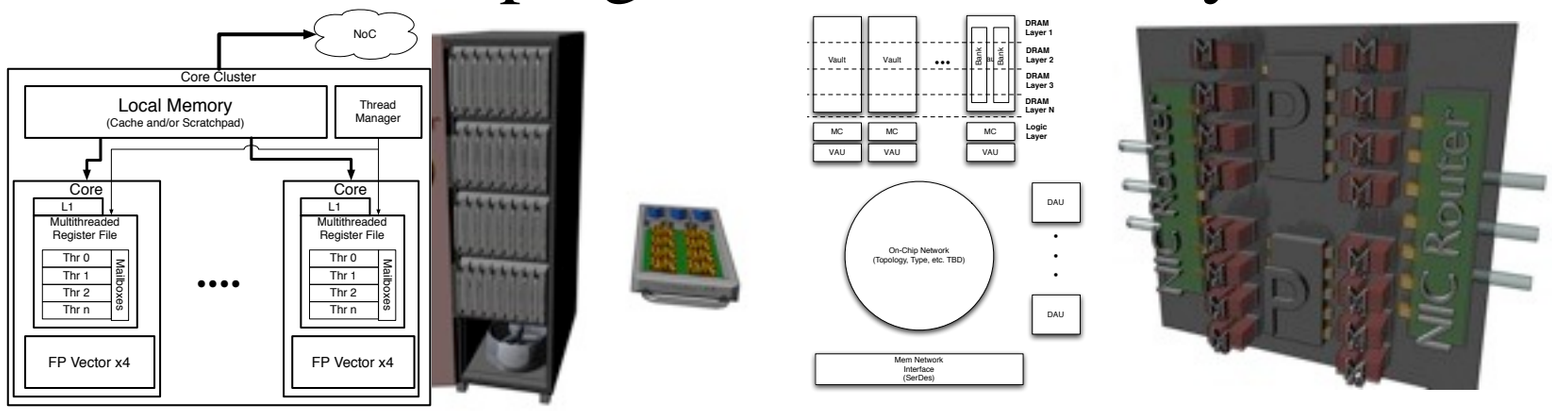
XGC: Approach via Five Thrust Areas

Safety and Security | Reentry | Circuity | Graph | Stream

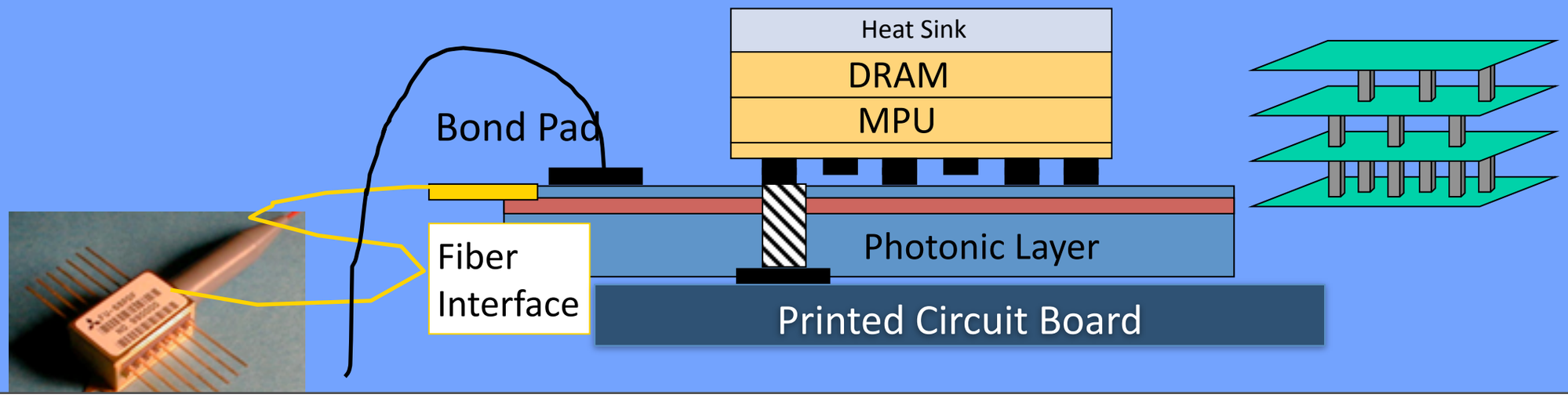
(5) Application Drivers

(4) System Software - enabling a new model of computation

(3) Architecture - Coping with Concurrency and Data Movement

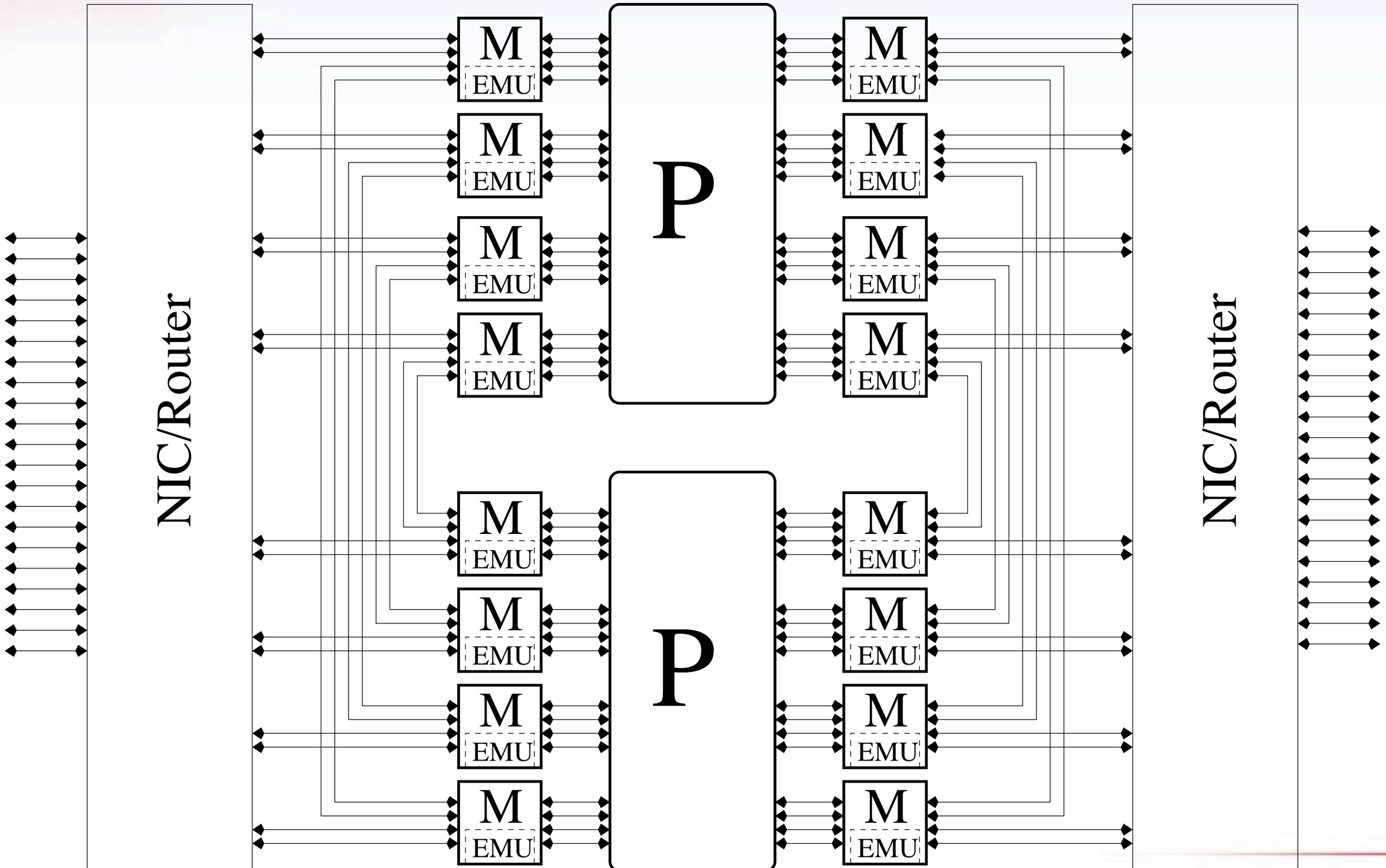


(2) Microsystems - Key Data Movement Enabling Technologies



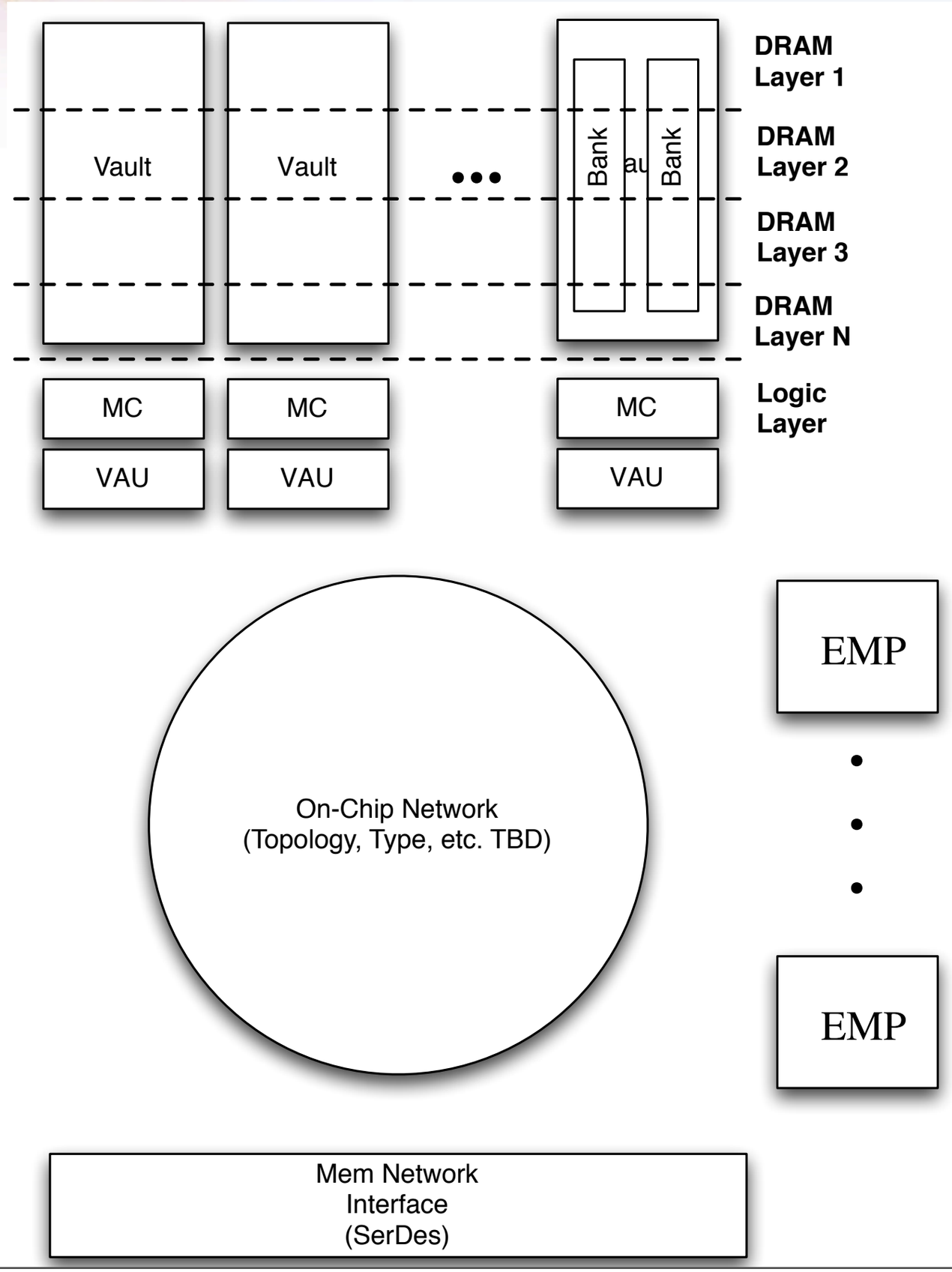
(1) Baseline - what happens if we do nothing?

X-caliber Node Architecture (Enabling Memory Abstraction)



Friday, September 16, 2011

Memory System (M) and Embedded Memory Processor (EMP)

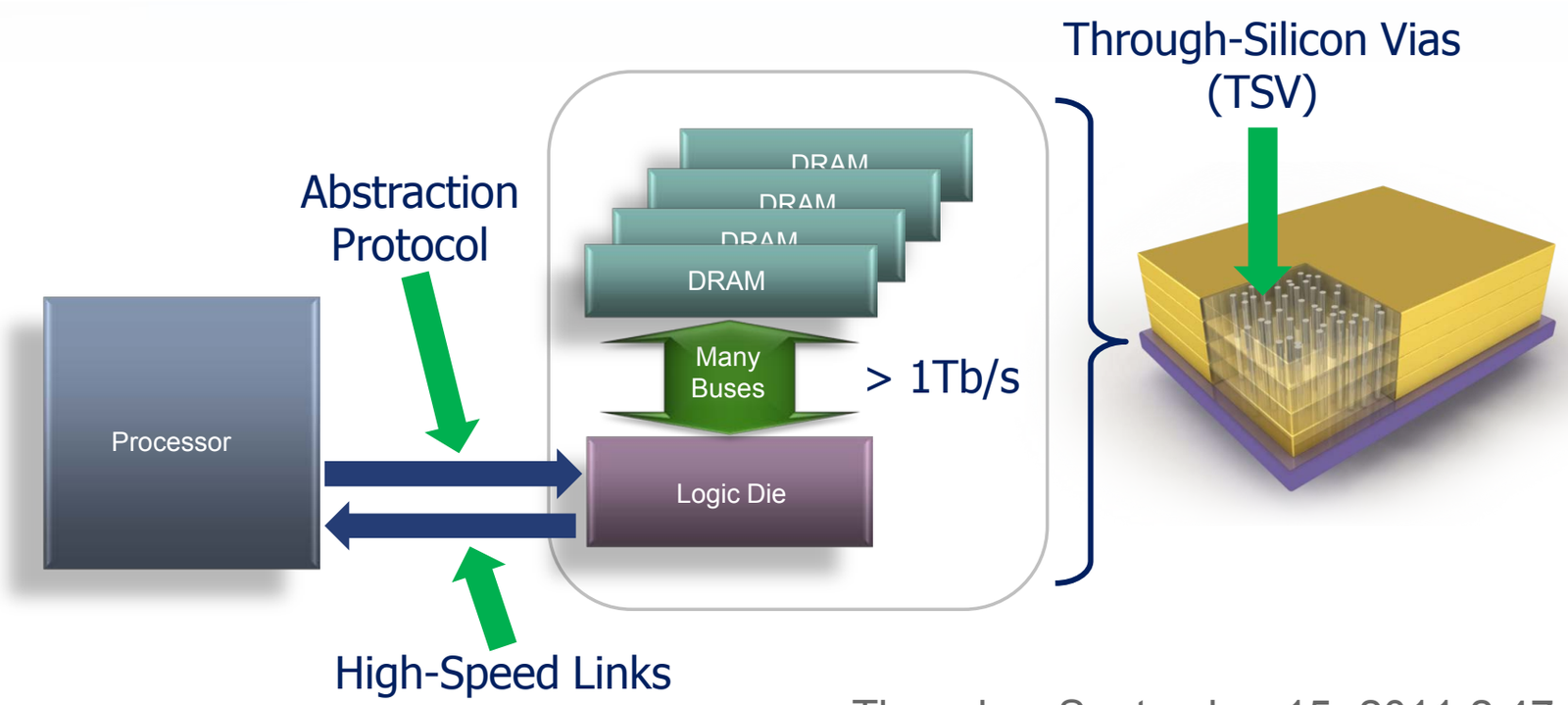


- Two computation Units
  - Right next to the DRAM vault memory controller (VAU)
  - To aggregate between DRAM vaults (EMP)
- "Memory Network" Centric
- Homenode for all addresses
  - Owns the address, data, and its state, "coherency"
- Three Control-Flow Options
  - In the Processor ("Memory is the Accelerator"), conventional
  - In the Memory System ("Processor is the Accelerator"), our approach
  - Both, probably un-programmable
- At 1-2 GHz, 4 EMPs per vault
- 64 vaults
- 2-4K threads per node in the memory system!

Sandia National Laboratories

But is it real?

Hybrid Memory Cube (HMC)



Notes: Tb/s = Terabits / second
HMC height is exaggerated

Micron

Thursday, September 15, 2011 2:47 PM EDT

Intel Debuts More Powerful Chips for Cloud Computing

By David Zielenziger

Intel, the world's biggest chipmaker, introduced new low-power and multi-core chips specifically designed for cloud computing applications. Some could have many as 50 core processors.

NNSA National Nuclear Security Administration

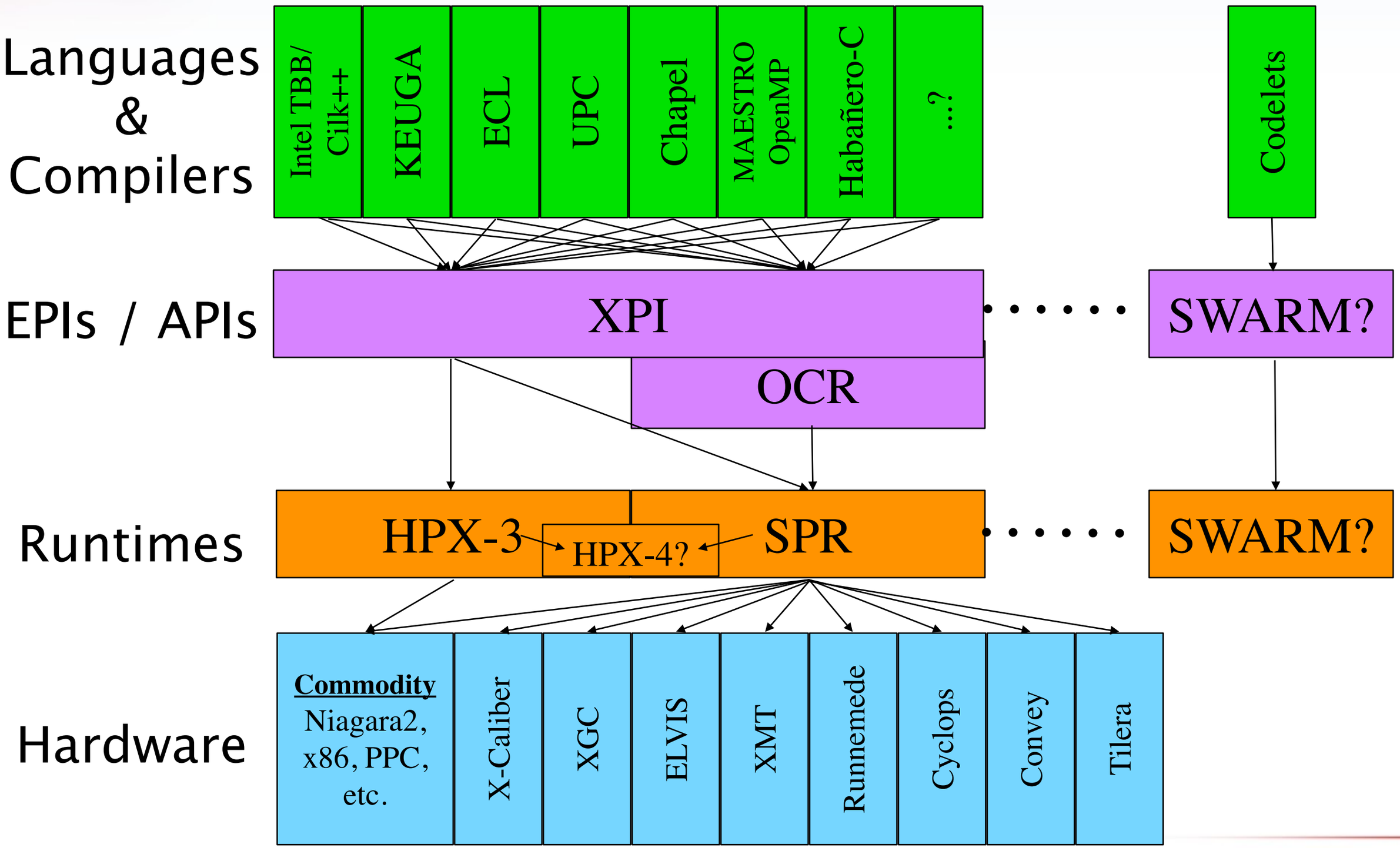
Sandia National Laboratories

System Balance

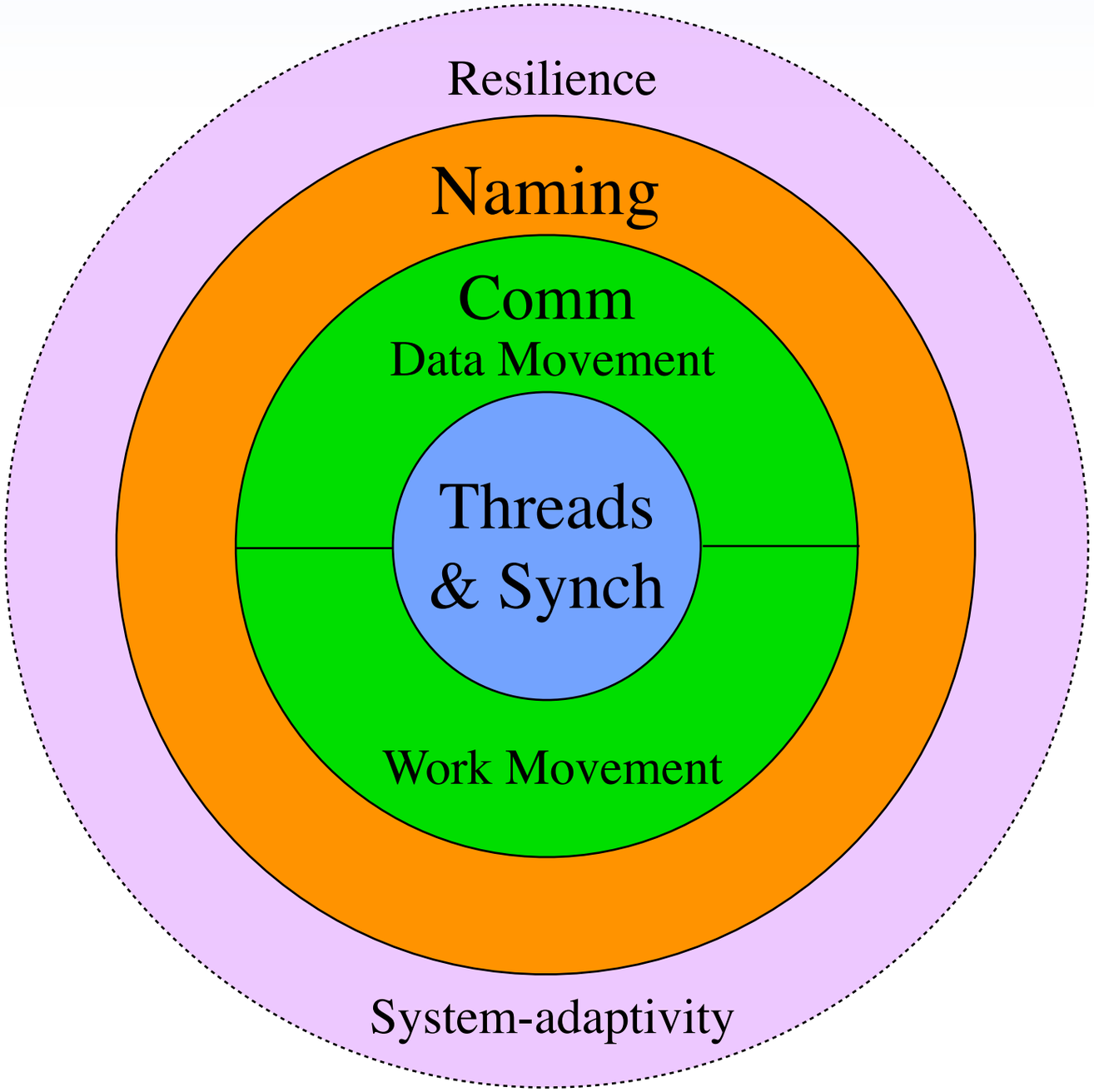
- System Balance
  - Because we're memory centric, we're focused on bandwidth, capacity, and scalability of the memory system (near and far)
  - X-caliber compared to the state of the art (scaled to 2018):
    - 5X the FLOPs of Red Storm
    - 2X the memory capacity
    - Similar network bandwidth ratio
  - 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

Execution Model Instantiation as a Runtime



Execution Model Onion



Where are we departing from the roadmap?

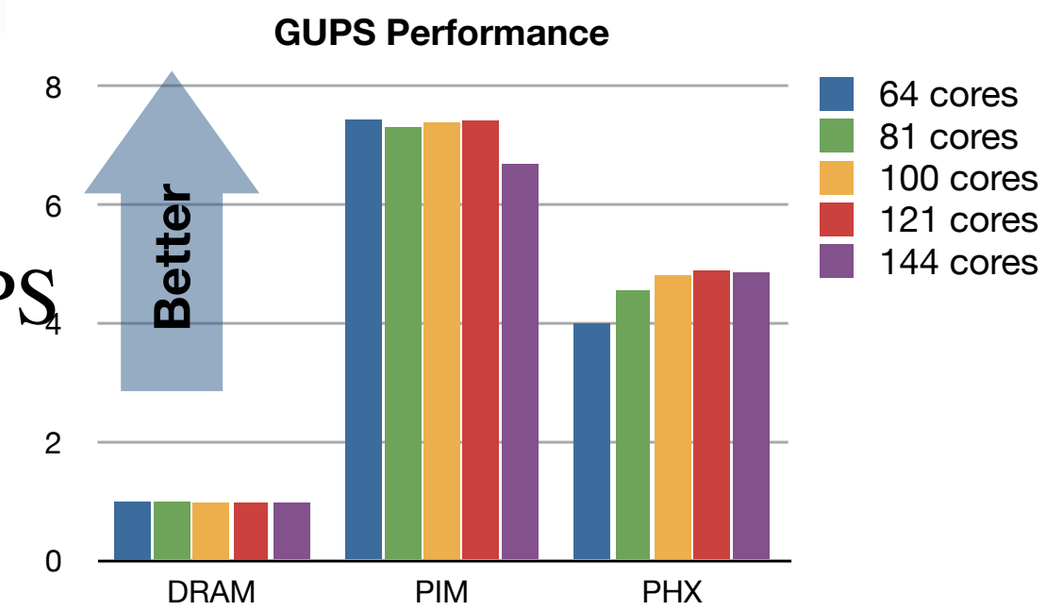
Element	ParalleX	GPUs	Stylized CSP	PGAS
Concurrency	Threads/Codelets	SIMD/lock-step threads	Ranks/ Processes	Processes
Coordination	Lightweight Control Objects (fine-grained)	Local Memory/ Explicit	BSP	BSP
Movement	of Work: Parcels of Data: PGAS + Bulk	Bulk Data Transfer (weak memory system)	Bulk Data Transfer	Data Only (load + store)
Naming	Global Name Space Global Address Space	Global Address Space	Explicit by Rank	Global Address Space
Introspection and Adaptivity	System Knowledge Graph/Dynamic	None/Static	None/Static	None/Static

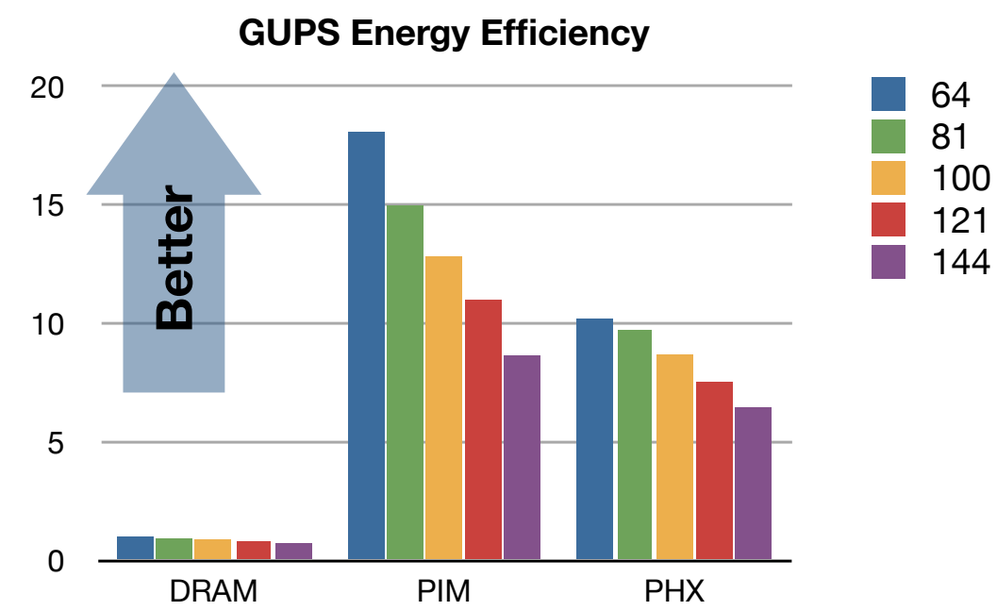
Early Results

Performance

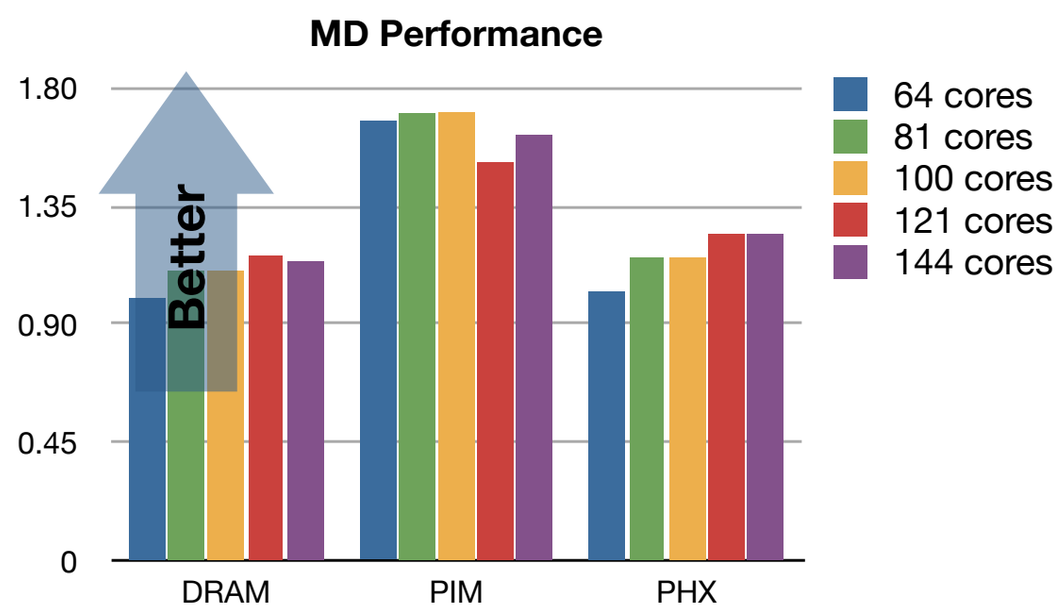
Energy Efficiency

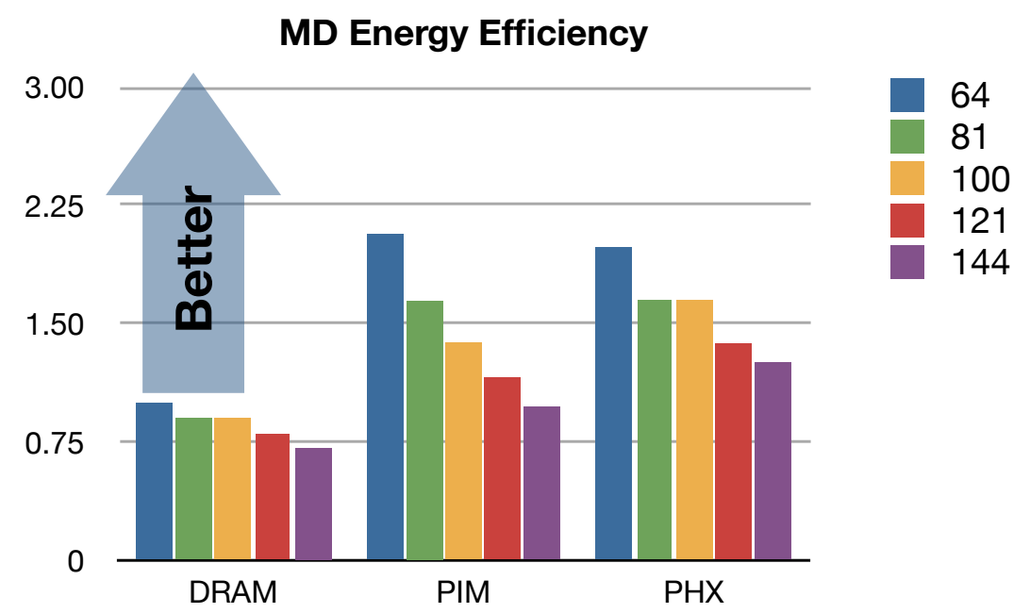
GUPS





MD



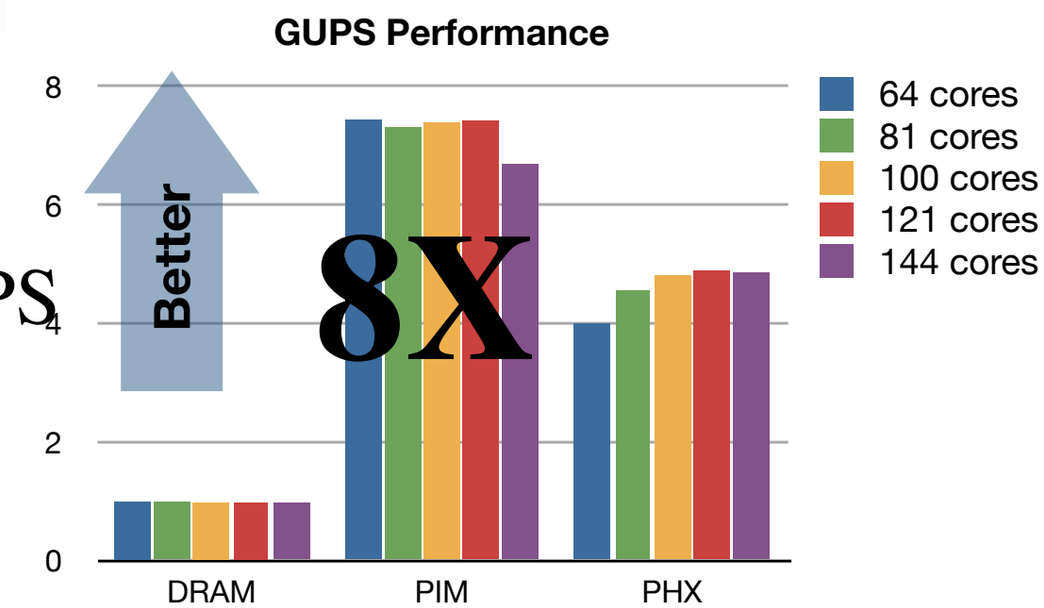


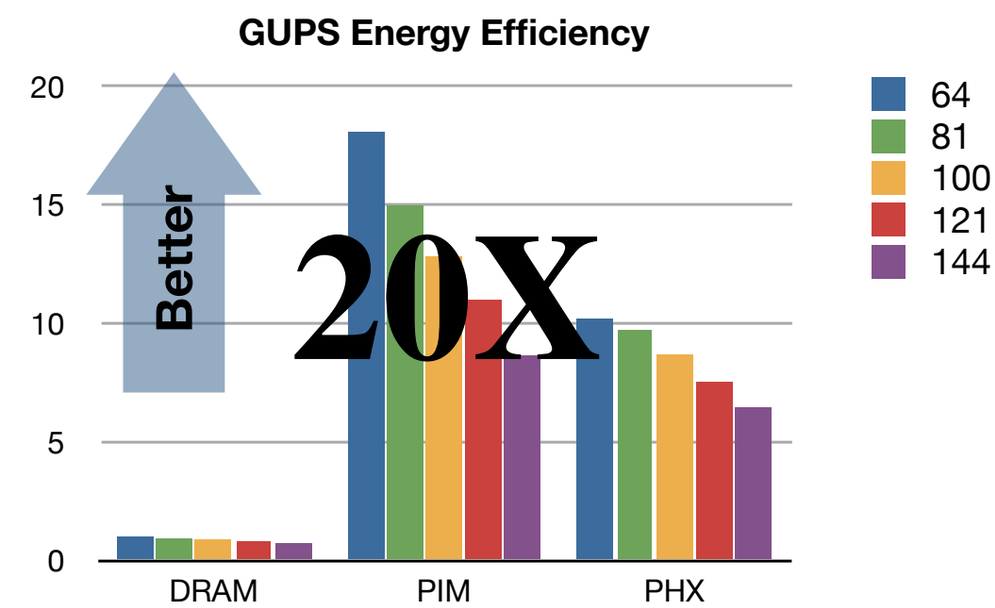
Early Results

Performance

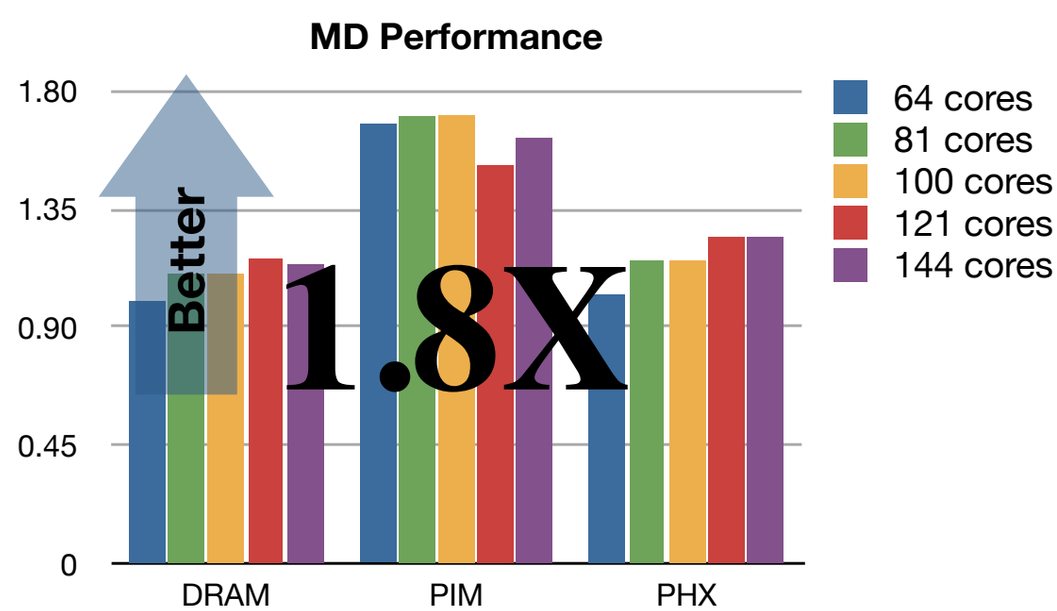
Energy Efficiency

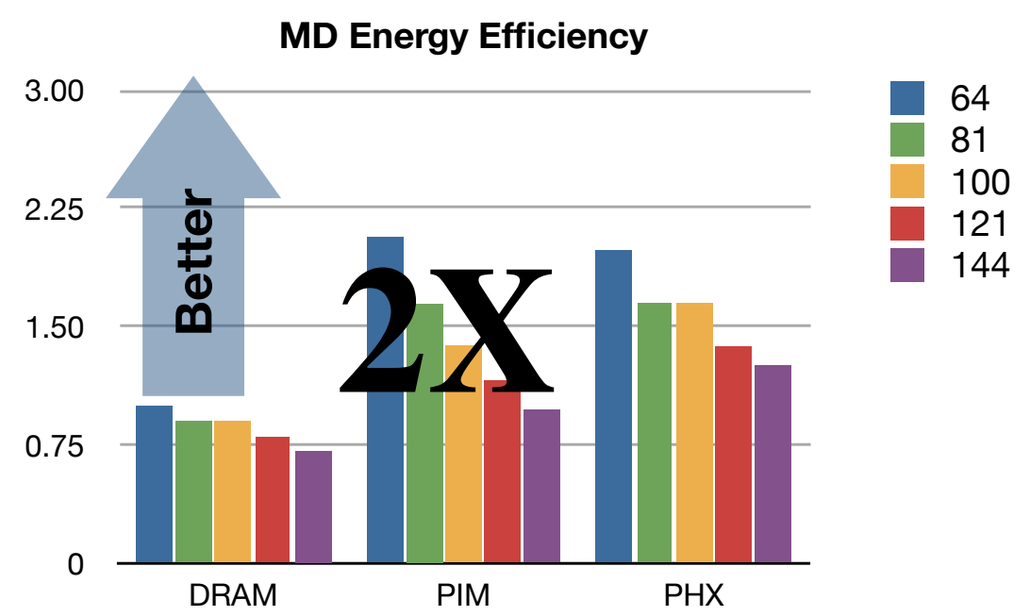
GUPS





MD

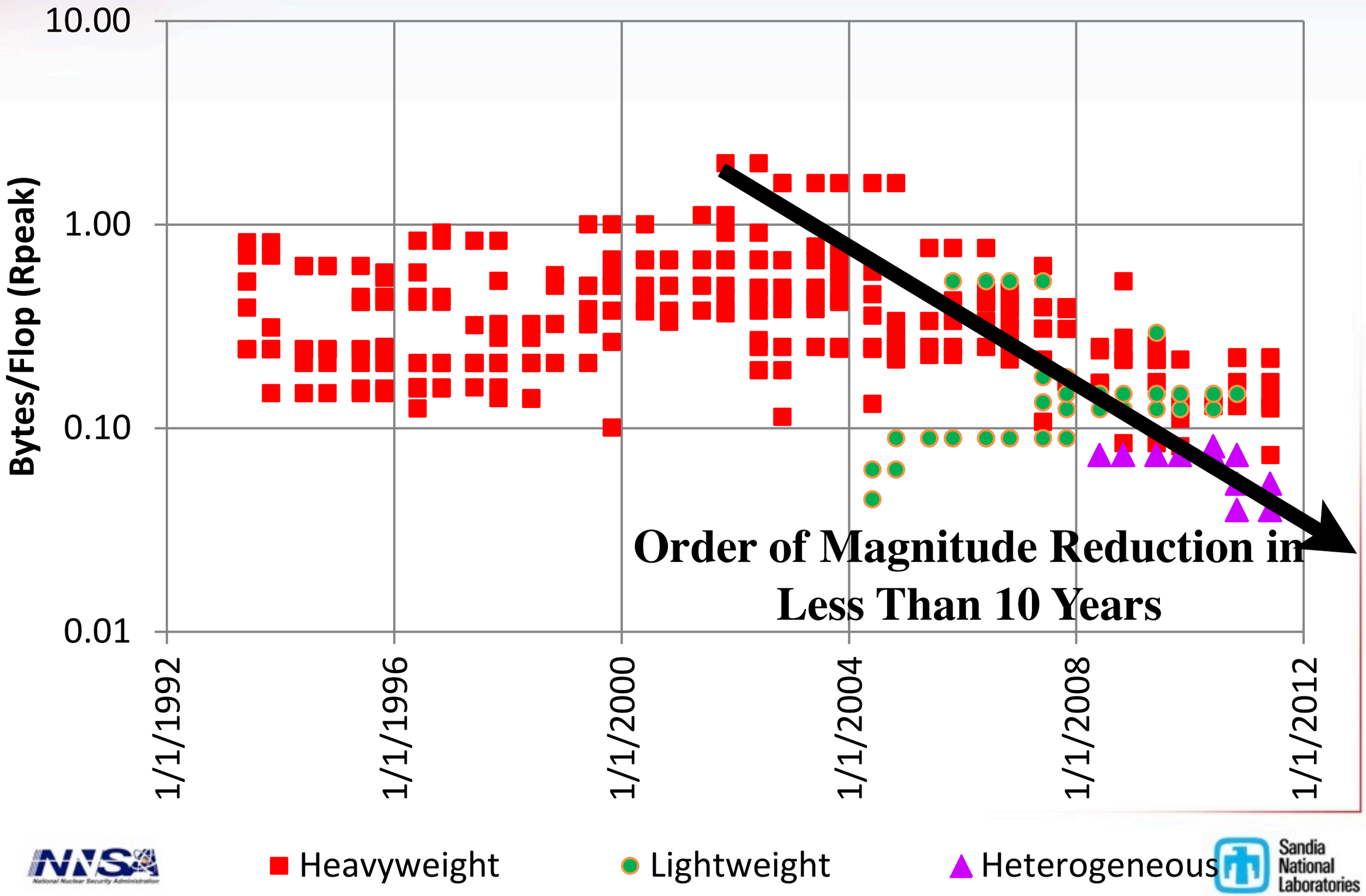




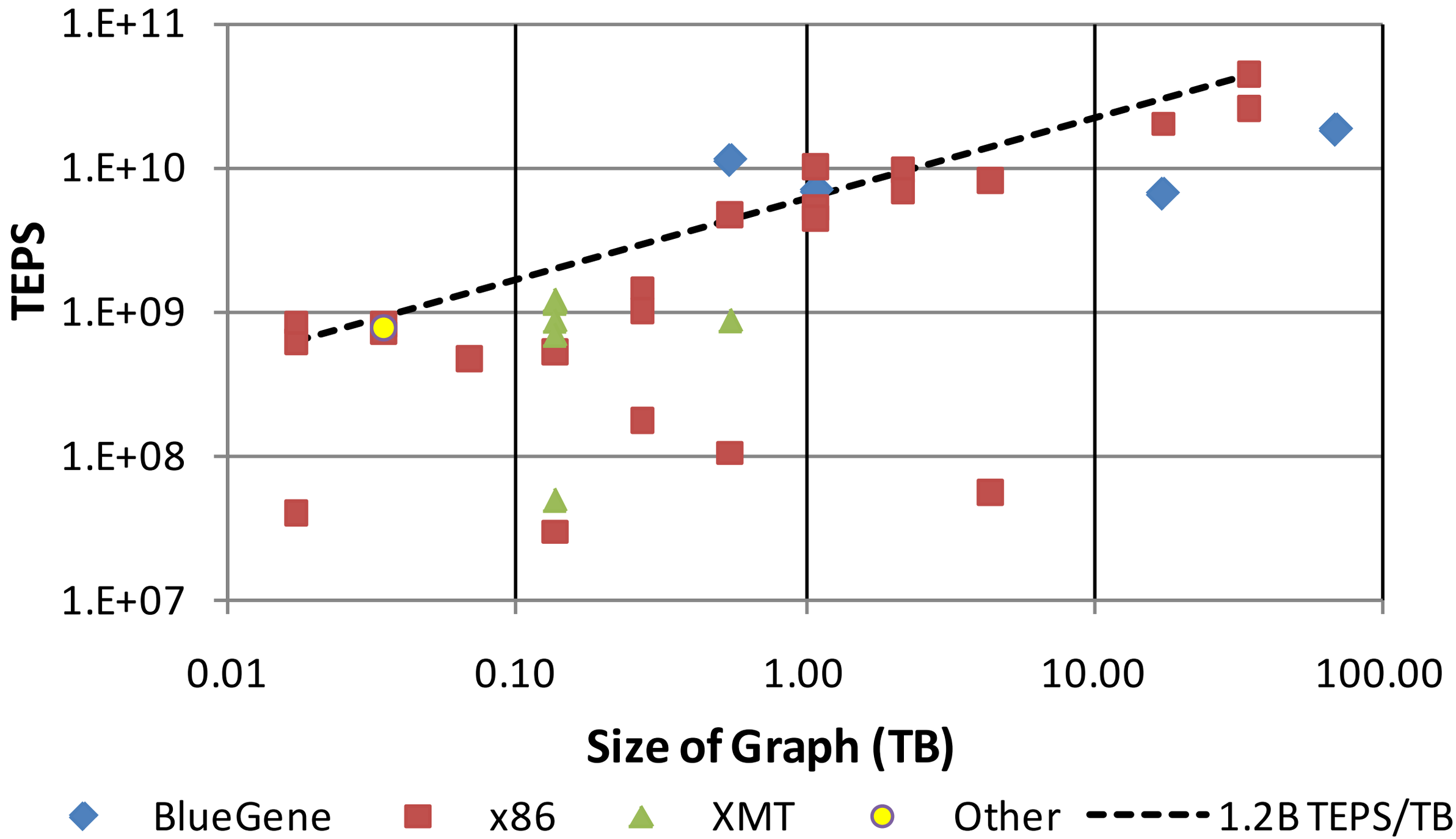
WE CAN BUILD THIS TODAY!

Bonus Material

Memory Capacity Trend

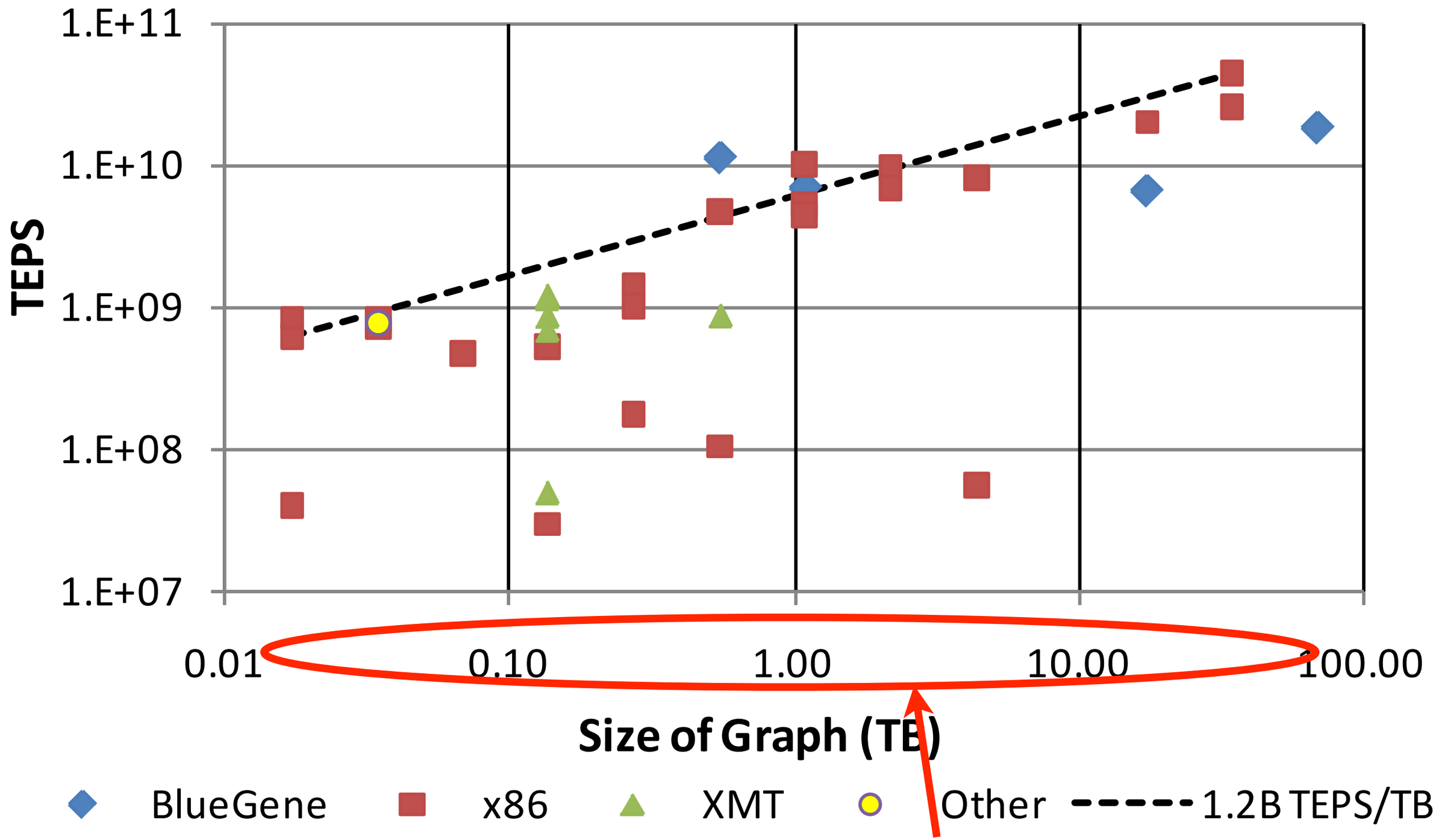


TEPS vs Problem Size



Data Courtesy of Peter Kogge, University of Notre Dame

TEPS vs Problem Size

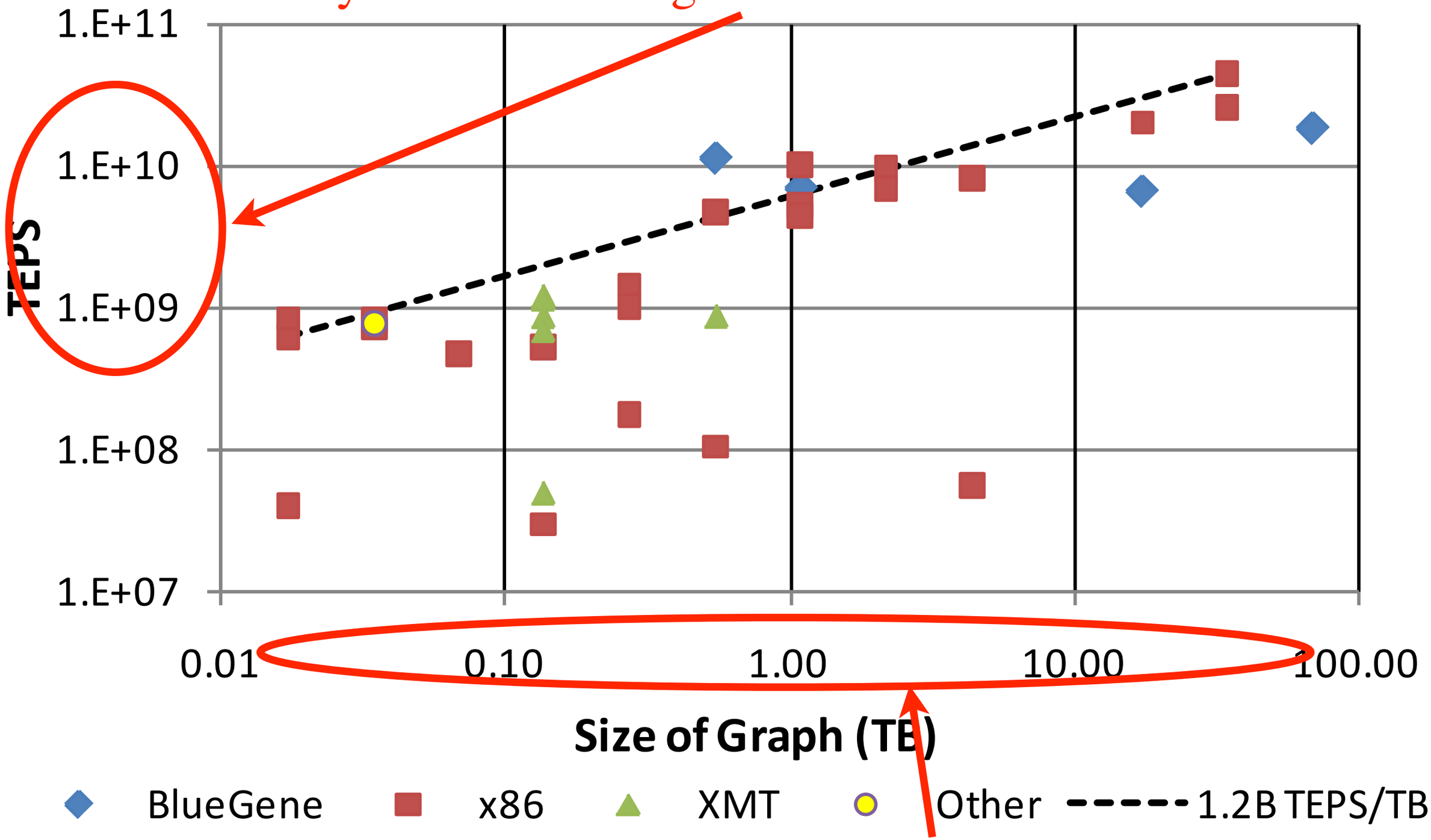


3 Order of Magnitude Increase in Problem Size

Data Courtesy of Peter Kogge, University of Notre Dame

TEPS vs Problem Size

Only 1 Order of Magnitude Increase In Performance



3 Order of Magnitude Increase in Problem Size

Data Courtesy of Peter Kogge, University of Notre Dame