Getting Ready for Hybrid Multicore Computing
or
On Data Movement, Pico Joules, and Codesign (oh my!)

Richard C. Murphy, Ph.D.

X-caliber Project PI

Scalable Computer Architectures Department

Sandia National Laboratories

Affiliated Faculty, New Mexico State University

March 22, 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.

Where are we? What does the future look like?

- The Technology Facts
  - Exascale in 2018 will require 125-500MW without intervention
  - Scaling will come from parallelism, not improved clock cycle time
  - Data movement dominates energy and performance
- Application Trends
  - Science codes will be increasingly unstructured
  - New data analytics applications show even less structure (and may eclipse traditional HPC by the end of the decade)
- Results for the architect
  - Today's design targets (particularly at the high end) represent the past, not the future
  - CPUs and GPUs are likely to converge over the next decade
  - We have a limited opportunity to affect programming and execution models to the benefit of our applications

What is Hybrid Multicore?

- Conventional, Processor-centric definition:
  - Hybrid: CPUs and GPUs in some combination
  - Multicore: lots of these on a chip
- A more data-centric definition:
  - Hybrid: put the functionality where it can achieve the desired computation with the minimum number of pico Joules
  - Multicore: achieve performance through parallelism not energy
- What's required:
  - Lots of support for data movement
  - Simple, likely heterogeneous compute elements
  - Advances in programming models, dynamic runtime systems, resource management, and underlying implementation technology
- Note: this is UBIQUITOUS (not just exascale)
  - As Dan Hitchcock said this morning, thinking about Exascale impacts ALL scales (terascale desktops, petascale racks, etc.)

What kind of research path forward do we need?

- The EI requires a DOE/SC pathfinding research component
  - See Kogge's IEEE Spectrum Article: http://spectrum.ieee.org/computing/hardware/exascale-computing-by-2015
  - Exascale report projections were likely optimistic by about 10X
  - Our application base may change
    - Informatics applications are important to DOE, especially ASCR
- Data movement dominates FLOPS
  - FLOPS have to be supported by memory access
  - Some proposed designs are 5-10X the FLOPS of Red Storm but...
    - 1/2 the relative network bandwidth
    - 2X the memory capacity
    - NOT a true "exascale" design
- Applications are the goal, and the power budget a constraint
  - Not the reverse!

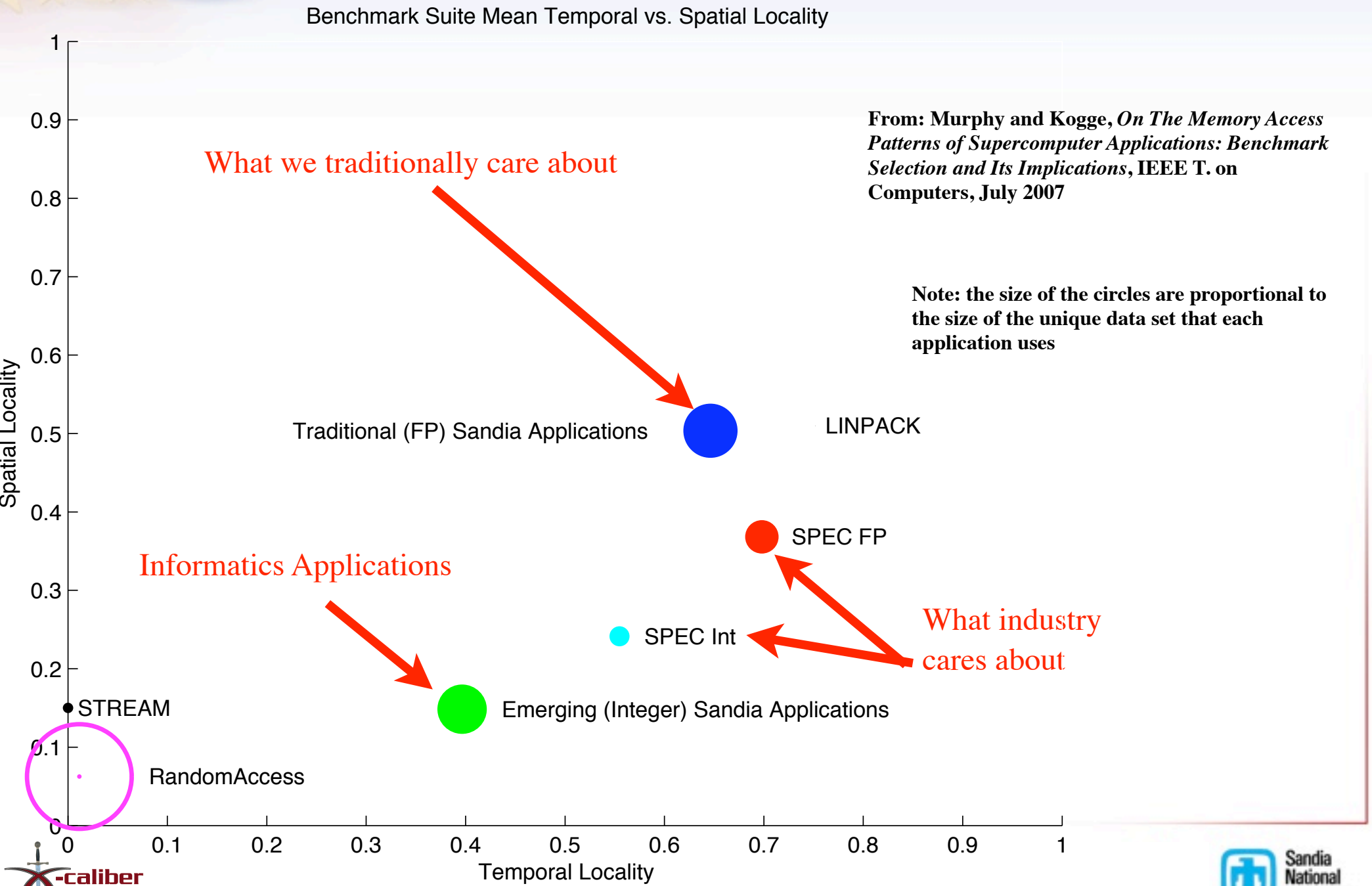
HTMT: A Historical Perspective

- Early Petaflops Effort (1996-1999)
  - NSF, DARPA, NASA, NSA
- One of 8 NSF-sponsored petaflops design points in a 6 month study
  - Would it be useful for DOE to have our own exascale design points?
  - Or, should they be exclusively generated by industry?
- 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-targeted 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

- Required Today (as discussed in the EI workshops):
  - Multithreading
    - We're stuck with this no matter what
  - Message-Driven Computation
    - Lightweight Active Messages/Parcels
  - Distributed global shared memory
- Lacking in today's machines (but likely necessary for Exascale):
  - Dynamic adaptive resource management and load balancing
  - Smart memory operations, percolation for prestaging computation
  - Data vortex for high bandwidth low latency for short messages
- Most of our memory work with Micron comes out of this PIM heritage
- X-caliber has all these things in one form or another

How are applications changing?



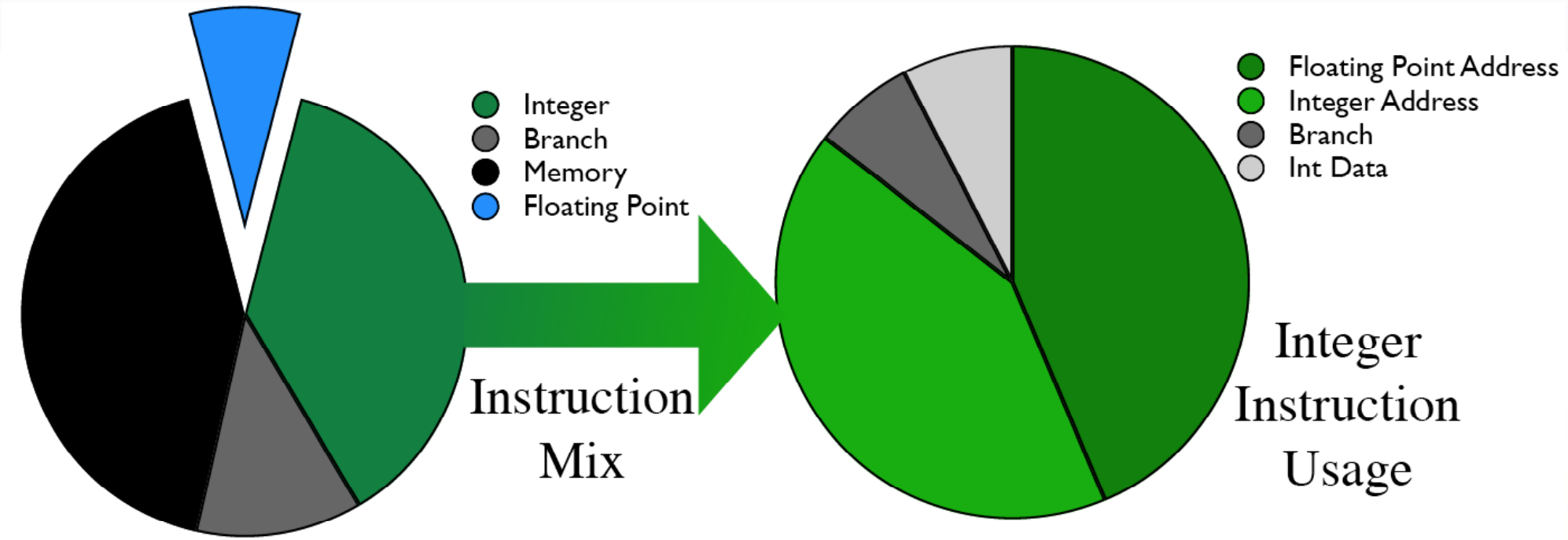
From: Murphy and Kogge, *On The Memory Access Patterns of Supercomputer Applications: Benchmark Selection and Its Implications*, IEEE T. on Computers, July 2007

Note: the size of the circles are proportional to the size of the unique data set that each application uses

X-caliber ARCHITECTURE

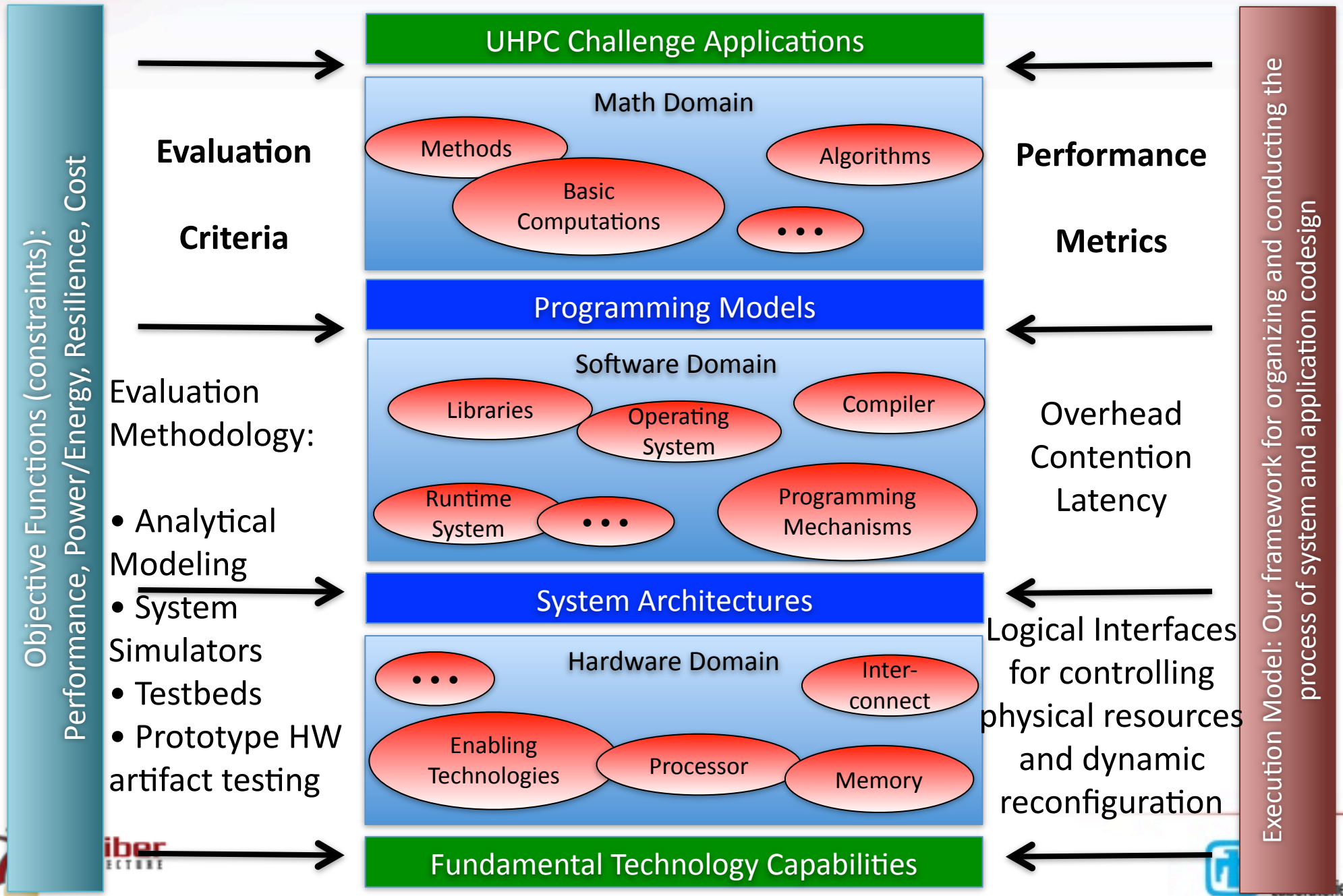
Sandia National Laboratories

What about DOE Physics Applications?



Most Physics Applications Primarily Do SLOW Memory References

What is codesign?



How does the X-caliber team think about codesign?

- **Model of Computation (AKA: Execution Model)**
  - Enables discussion of the semantics of a machine separate from the implementation... why?
    - How else do people at different layers communicate new ideas?
    - How else do you optimize between layers?
  - Not the traditional approach of a hardware implementation being thrown (at application developers) over the fence
- **My five elements of an execution model...**
  - Concurrency
  - Coordination
  - Movement
    - of work
    - of data
  - Naming
  - Introspection

ParalleX

Element	Parallex Mechanism
Concurrency	Lightweight Threads/Codelets (lightweight, h/w scheduled, for latency tolerance not throughput!)
Coordination	Lightweight Control Objects (LCOs) for construction of mutexes, futures, producer/ consumer interactions, etc.
Movement	Of Work: Parcels (lightweight active messages) Of Data: PGAS and Bulk Transfer
Naming	Global Name Space and Global Address Space
Introspection	Unified publication at all levels via System Knowledge Graph (SKG)

ParalleX vs. Today's Dominant Model

Element	Parallex Mechanism	Stylized Communicating Sequential Processes
Concurrency	Lightweight Threads/ Codelets	MPI Ranks/Processes
Coordination	Lightweight Control Objects (LCOs) (fine-grained)	Bulk Synchronous (or maybe by teams and messages) (coarse-grained)
Movement	of Work: Parcels of Data: PGAS and Bulk	of Work: None of Data: Bulk
Naming	Global Name Space Global Address Space	Coarse, rank/node names
Introspection	System Knowledge Graph (enables dynamic/ adaptive)	Not specified by the model, in practice out-of-bands RAS network

System Balance for Petascale Racks

- 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 (in a rack)
    - Half the memory capacity (or less)

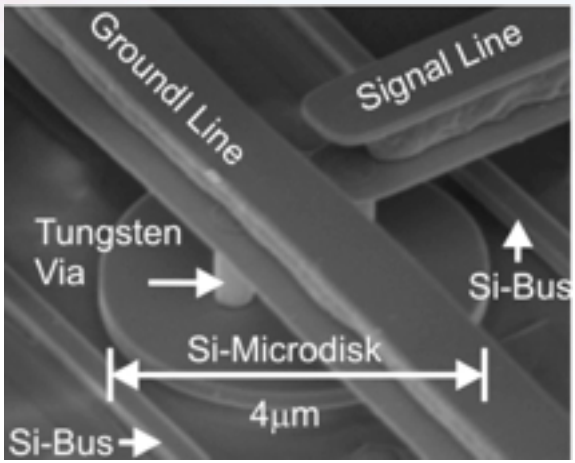
System	Injection BW	FLOPS	B/F	Comment
X-caliber	133 TB/s - 266 TB/s	1.0 - 1.4 PF/s	0.095 - 0.266	Adaptive
Other Proposals	205 TB/s	2.6 PF/s	0.0788	Static

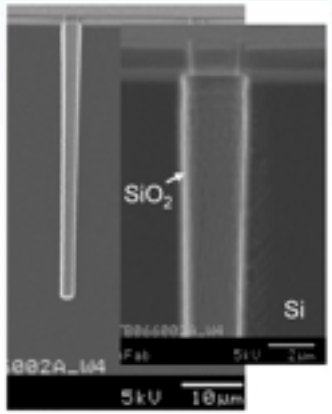
DARPA Challenge Problems

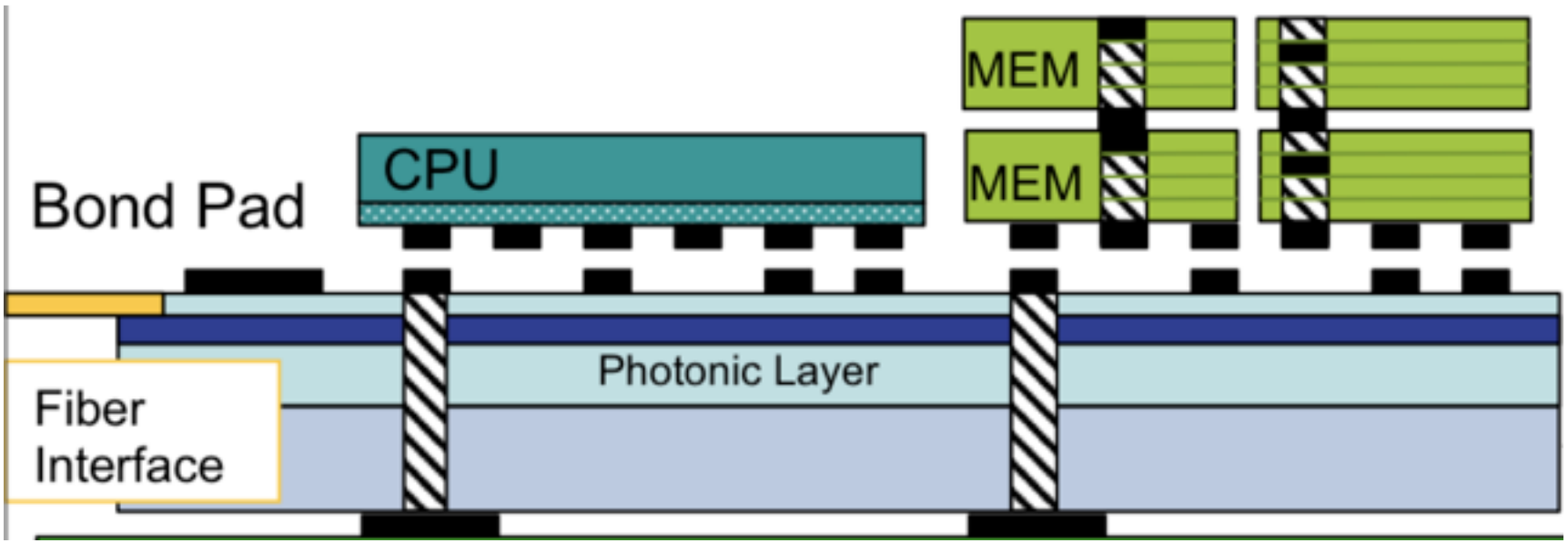
Problem Area	Standin Problem	Executes	Researcher Responsible	Quality
Graph	Graph500 Concurrent Search	EMP	Brian Barrett and Bruce Hendrickson	Integer Pointer Dereference
Stream	GUPS	EMP	Steve Plimpton	Input + Integer Pointer Dereference (latter harder)
Decision Support	Chess	EMP	Thomas Sterling	Integer Pointer Dereference
Shock Physics	MiniFE	EMP + P	Mike Heroux	Integer Pointer Dereference + 12% FP
Molecular Dynamics	MiniMD	P	Marc Snir and Steve Plimpton	15% FP with lots of local references

Our initial DOE challenge problems will also be informed by the ASCR codesign centers

Our Enabling Technologies: Advanced Packaging, 3D Integration, Optics





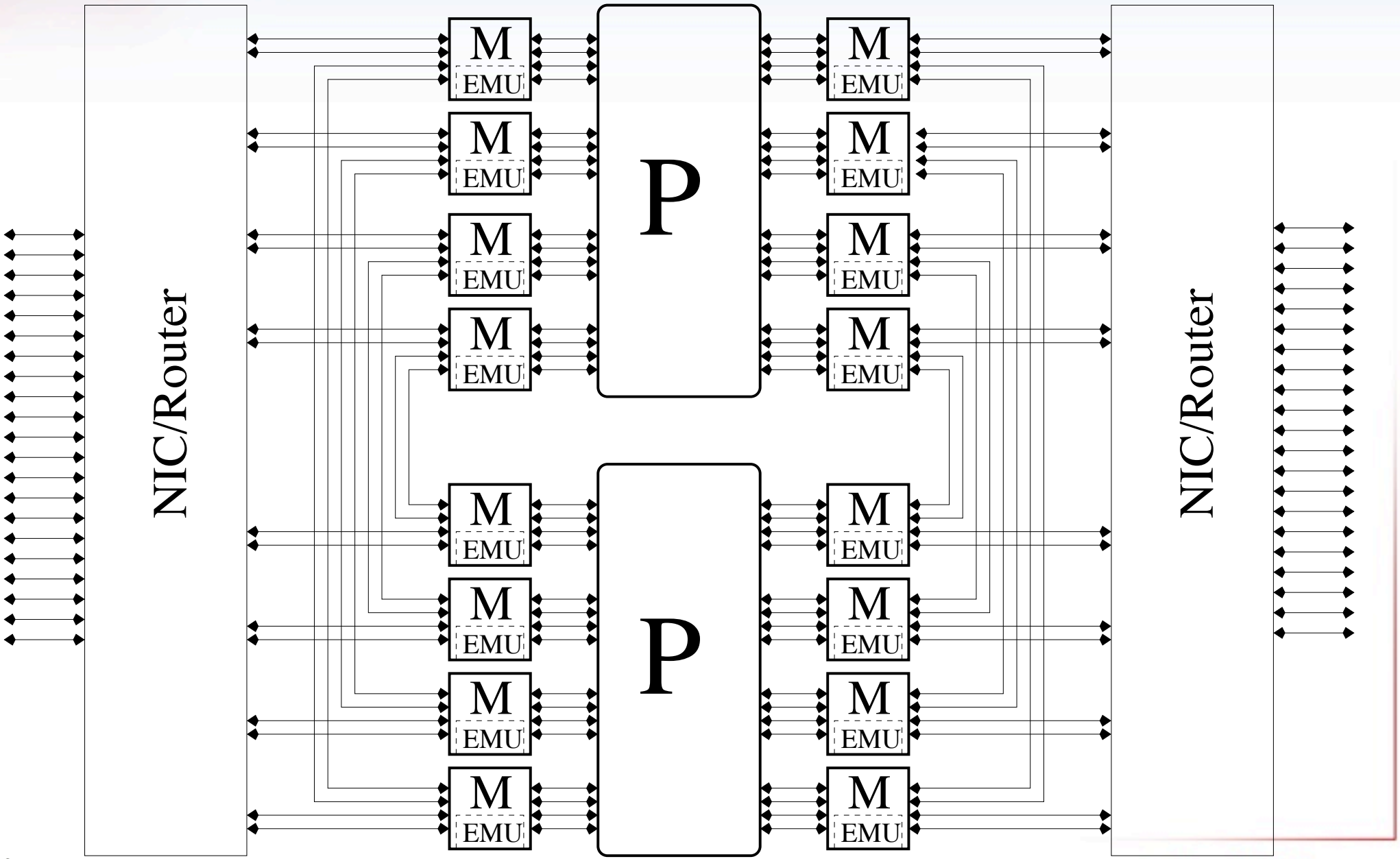


We are leveraging Intel's investment in low-power circuits

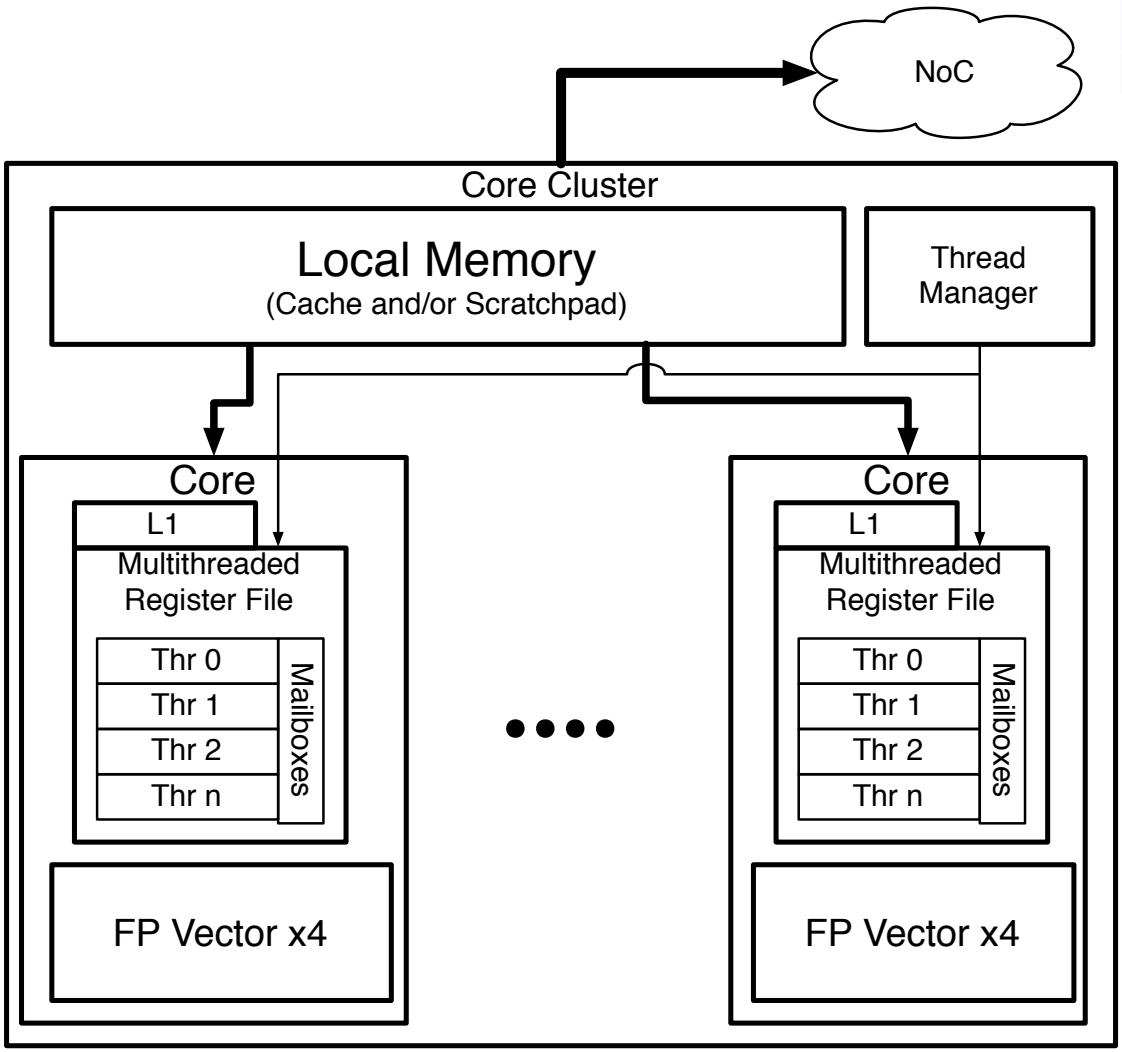
X-caliber ARCHITECTURE

Sandia National Laboratories

Node Architecture (Continued)

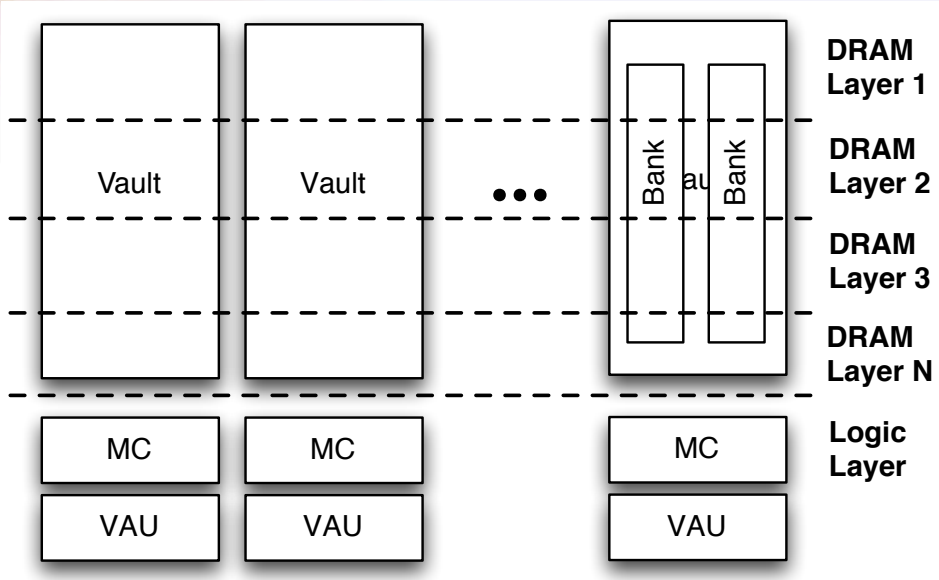


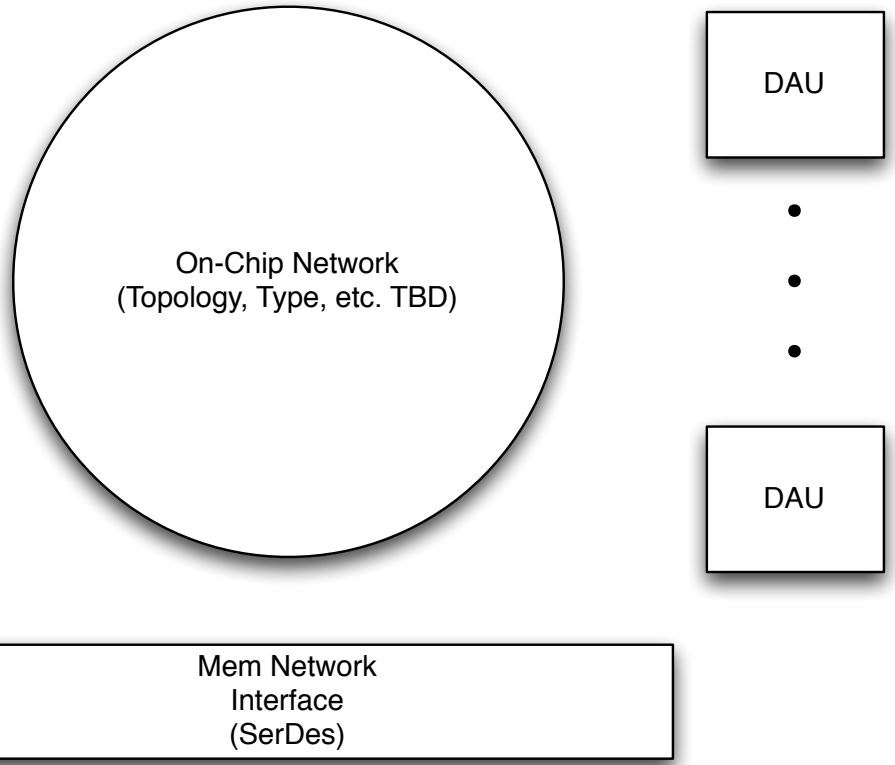
Processor (P)



- X-caliber more concerned with data movement
- Hybrid CPU-like and GPU-like architecture
- Heavily Threaded and Vectored
- Client of the Memory Network
- Owns only cache/ scratchpad memory

Memory System (M)





- Two computation Units
  - Right next to the DRAM vault memory controller (VAU)
  - To aggregate between DRAM vaults (DAU)
- "Memory Network" Centric
- Homenode for all addresses
  - Owns the "address"
  - Owns the "data"
  - Owns the "state" of the data
  - Can build "coherency"-like protocols via local operations
  - Can support PGAS-like operations
  - Can manage thread state locally

Sandia National Laboratories

Thread Coalescing Observation

- System design as given is oriented in two modes of operation: high temporal locality/low temporal locality
- The "right" view of this may actually be thread-rich and thread-starved (some anecdotal XMT-evidence for this)
- If so ... lightweight threads may be:
  - very small in state (8-registers-ish)
  - XMT-like scheduling with lightweight synchronization (including synchronization on a register!)
- In "thread-starved" mode we may want hardware to use the same resources to create "heavyweight" threads automatically
  - coalescing 4 under-used cores in this mode could create 1 32-register thread
  - Tomasulo's would allow hardware to expand the register set cheaply and DOES NOT have to be coupled with speculation
  - < 10 cycles to memory (at least locally) looks more like an IBM-360 than a super-scalar, speculative, out-of-order system

Sprinting

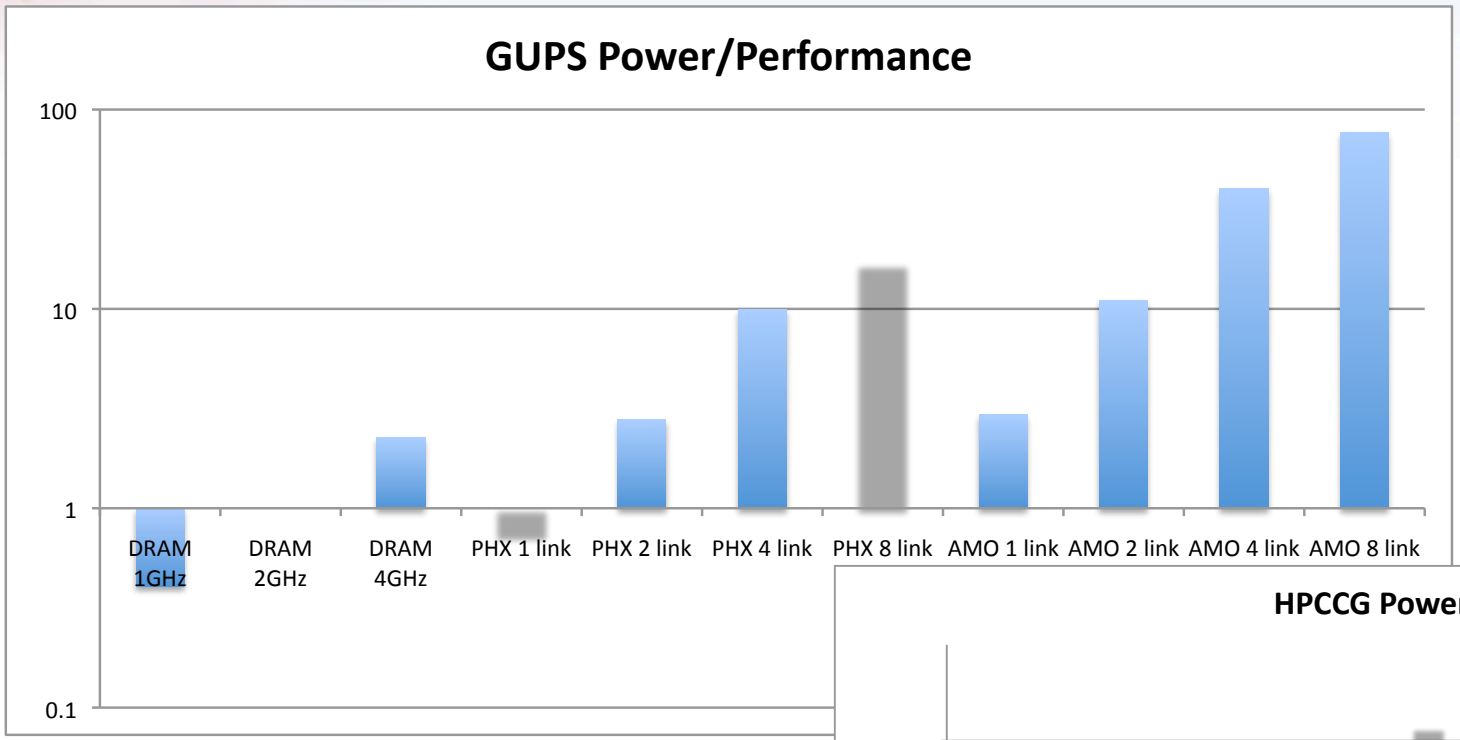
- Every major component of the system has the capability to "sprint" by operating outside it's nominal power envelope
  - Processor: Increases the clock rate from 1.5 GHz to 2.5 GHz
    - Can be applied to half the cores and allow "ping-ponging"
  - Memory: Additional memory links (increasing concurrency and bandwidth) can be powered up in sprint mode
  - Network: Sprint on injection bandwidth from 512 GB/sec on the NIC to 1 TB/s
- Decisions about when to sprint are made dynamically by the runtime and OS

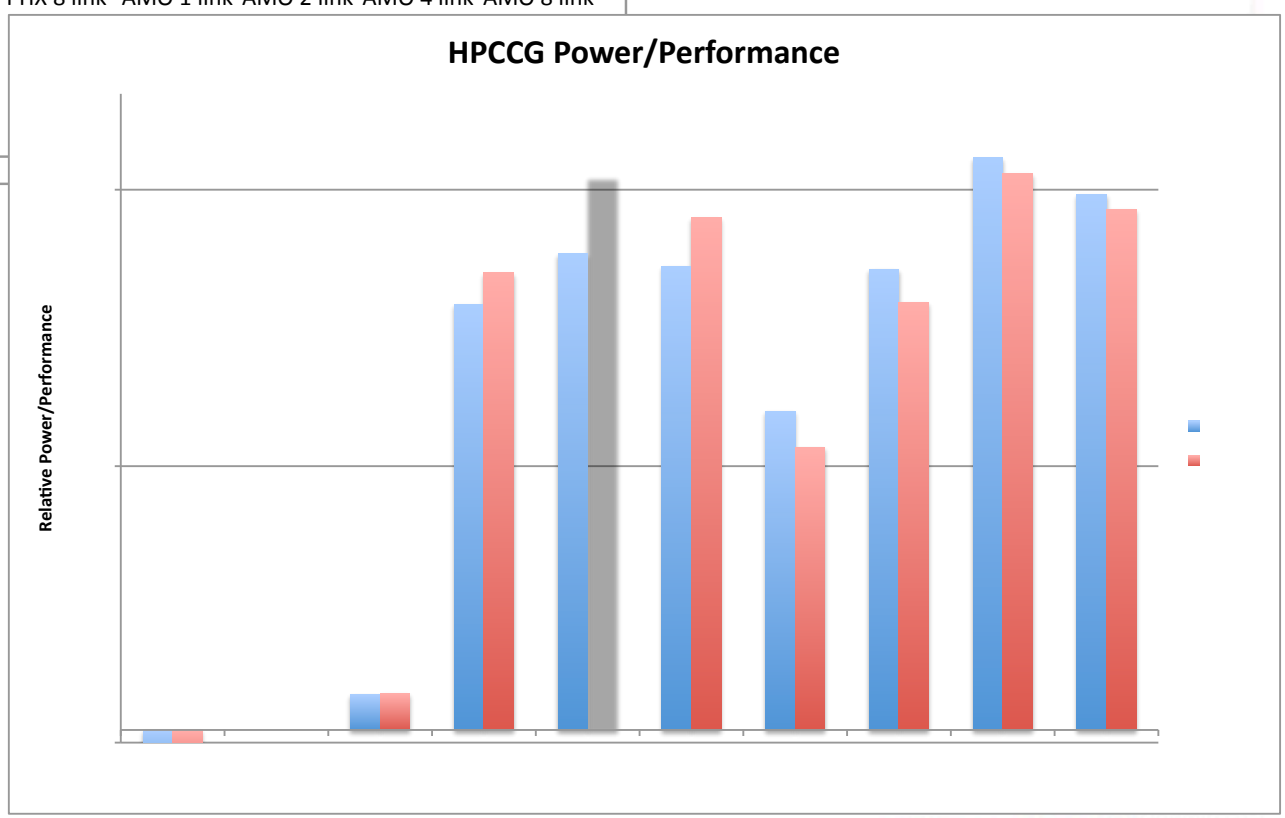
Target Scales

- **Rack Scale**
  - **Processing:128 Nodes, 1 (+) PF/s**
  - **Memory:**
    - **128 TB DRAM**
    - **0.4 PB/s Aggregate Bandwidth**
  - **NV Memory**
    - **1 PB Phase Change Memory (addressable)**
    - **Additional 128 for Redundancy/RAID**
  - **Network**
    - **0.13 PB/sec Injection, 0.06 PB/s Bisection**

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

Nearly 100X Improvement In Power/Performance over Conventional Memory Roadmaps in GUPS





A Joint Intel/Sandia Roadmap for Exascale

- Foster closer engagement with Intel Labs
- Combined activity arising from UHPC to identify the research required for 2016 proof of concept platform
  - Low-power circuits and SerDes
  - Memory Architecture
  - Acceleration architecture for target applications
  - Communications and IO networks
    - MPI, lightweight active messages, and codelets
  - Execution Model
    - Draft at April 4th meeting
  - Programming Model
    - Draft at April 4th meeting
  - Application Understanding
- Preliminary draft roadmap April/May 2011

Initial Intel/Sandia Findings from UHPC

- **Execution and Programming Model Change is required**
  - MPI+Threads is not viable for systems with millions of cores
  - Need to manage and utilize resources in a new way
  - Need better ways of capturing dependencies and data structure
- **Execution Model Acceptance/Adoption will be a challenge**
  - Need a broad government/Industry/Academia commitment.
  - Need implementation on cluster systems today
  - Need support for tools developers to target new model
  - Need motivation for application experts to program to new model

Initial Intel/Sandia Findings from UHPC (continued)

- Codesign is key to efficiency
  - Need to define application target that represents DOE/DoD and Industry missions for Exascale
  - We need to decide if one system fits all of if some application customization of hardware is viable
- Fundamental Technology Advances have to occur to support Exascale
  - Circuits, data locality management, photonics, Architecture, Model of Computation
- Now is the time to start the disruptive research
  - Production processor roadmaps take 5+ years to impact
  - Results have to be demonstrated by 2013-2015 to impact the exascale roadmap
  - Clear technology transitions (off-ramps) for disruptive technology

Murphy's Recommendations... we need to:

- Continue to quantify "what if we do nothing"?
  - Peter Kogge should extend and refine the Exascale Report's targets
    - For DOE/SC: lightweight, heavyweight, and add hybrid
- Create an application-driven DOE execution model(s)
  - MPI + Threads is the starting point, but NOT the end-point
- Proffer DOE Exascale Design Points to complement industry design points
  - Highly custom, like X-caliber (targeted for DOE/SC apps)
  - Semicustom SoC based on something highly commodity like ARM
    - Powerful network and memory, homogeneous low-power compute
  - Semicustom application-specific targets like Green Flash
  - Homogeneous SoC (Cray XMT ramped up + MPI)
  - Heterogeneous and highly reconfigurable
- Recommendation: Give small, focused groups 6 months and a charter to produce "notional systems/execution models" for discussion

Thank You!