SAND2008-4701P

High Performance Computing at Sandia

Douglas Doerfler
Organization 1422
Sandia National Laboratories

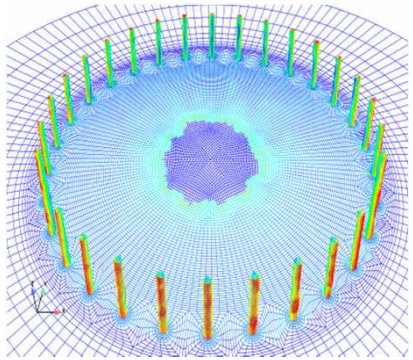
Kansas State University Visit

July 16th, 2008

SAND2008-xxxxP
Unlimited Release
Printed July, 2008

Center 1400: Multidisciplinary Research Has Been our Key to Success

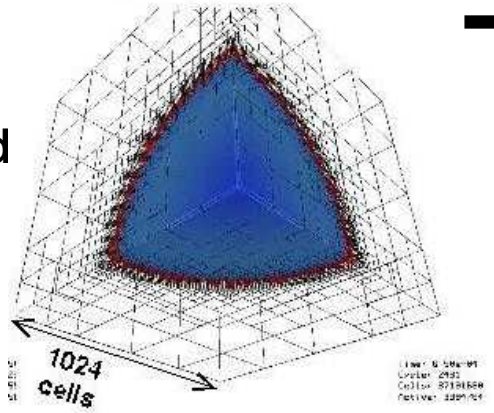
Computing Architectures And Systems



Computational Science Applications

+

Algorithms and Enabling Technologies



Hardware

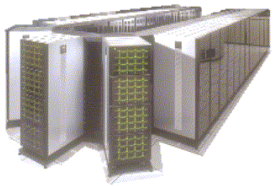
Sandia National Laboratories

Sandia has made key contributions to the development of MP computing technology


CM-2


nCUBE-2


iPSC-860


Paragon


ASCI Red


Cplant


Red Storm

1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006

Gordon Bell Prize

R&D 100
Parallel Software

Patent
Meshing

R&D 100
Dense Solvers

R&D 100
Storage

Gordon Bell Prize

World Record
Teraflops

Gordon Bell Prize

R&D 100
Allocator

R&D 100
Trillinos

Gordon Bell Prize

SC96 Gold Medal
Networking

R&D 100
Signal Processing

Mannheim
SuParCup

Karp Challenge

World Record
281 GFlops

R&D 100
Aztec

Patent
Data Mining

R&D 100
Salvo

R&D 100
Meshing

World Record
143 GFlops

Fernbach
Award

Patent
Parallel Software

Patent
Paving

Patent
Decomposition

Sandia National Laboratories

Red Storm Machine Specifications

- Massively Parallel Processor*
  - 284 TFLOPs
  - 75 TB RAM
  - 12,960 nodes; 38,400 processor cores
  - 21 TB/sec bisection bandwidth
  - Nearly 2PB of disk space
    - 60 GB/s BW per side

*compute section statistics only

Sandia National Laboratories

SNL/NNSA Re-establish Cray in the Supercomputing Marketplace

- Cray's product roadmap was strictly vector architecture based with the X-1 line
  - **This provided solutions to a limited market segment**
  - This was Cray's initial proposal for the Red Storm contract
  - X-1 line essentially saturated the market within 1 year of it's availability
- SNL/NNSA recognized the potential for Cray to market a commercial processor based MPP architecture
  - **High risk venture with the potential for international impact**
  - SNL was the architect and provided SW expertise (Cray's SW light-weight OS expertise was held at bay due to terms of the acquisition from SGI by Tera)
  - Cray provided HW and systems expertise
- Office of Science & Oak Ridge National Lab Peta-scale Initiative
  - **Have capitalized on this investment by NNSA**, with the XT3 based Cray road map providing the basis for their Petascale Initiative
- Testimonials
  - "Together with **Sandia National Laboratories, who partnered with Cray in designing the 'Red Storm' architecture**, we are very excited that PSC has selected 'Red Storm' for their very diverse and demanding scientific supercomputing workload."
    - Peter Ungaro, Cray President and CEO [http://www.hpcwire.com/hpcwire/hpcwireWWW/04/0507/107608.html]
  - "Many HPC systems today are designed to excel on peak performance and Linpack numbers that are poor predictors of actual problem-solving performance on many end user applications. Like the Cray T3E before it, the new **Cray XT3 is designed for high performance on large-scale customer HPC applications and workloads**."
    - Earl Joseph, IDC Program Vice President [http://www.cray.com/products/xt3/index.html]

Sandia National Laboratories

XT3 is an International, Commercial Success

- Major XT3 installations worldwide (19 sites)
  - SNL: 13,500 nodes, 125 TF
  - Pittsburgh Supercomputing Center: 2,200 nodes, 10 TF
  - Oak Ridge: originally 5,200 nodes, 54 TF; now 12,000 nodes, 122 TF
  - Swiss National Supercomputing Center: 1,600 nodes, 8.5 TF
  - Atomic Weapons Establishment, UK: 3,900 nodes, 40 TF
  - Army Corp of Engineers, Engineering Research Development Center: 4,000 nodes, 40 TF
  - LBNL-NERSC: 9,500 nodes, 100 TF
- Recent Announcements:
  - ORNL: 25,000 nodes, 1 PF plus 130 TF upgrade to current system
  - Army Corp of Engineers, Engineering Research Development Center: 70 TF
  - Engineering and Physical Sciences Research Council, UK: 50 TF
- Plus multiple small (2 to 4 cabinet) installations in US, Japan, Europe, Canada and Australia

Sandia National Laboratories

New Challenges: The NNSA Complex Transformation

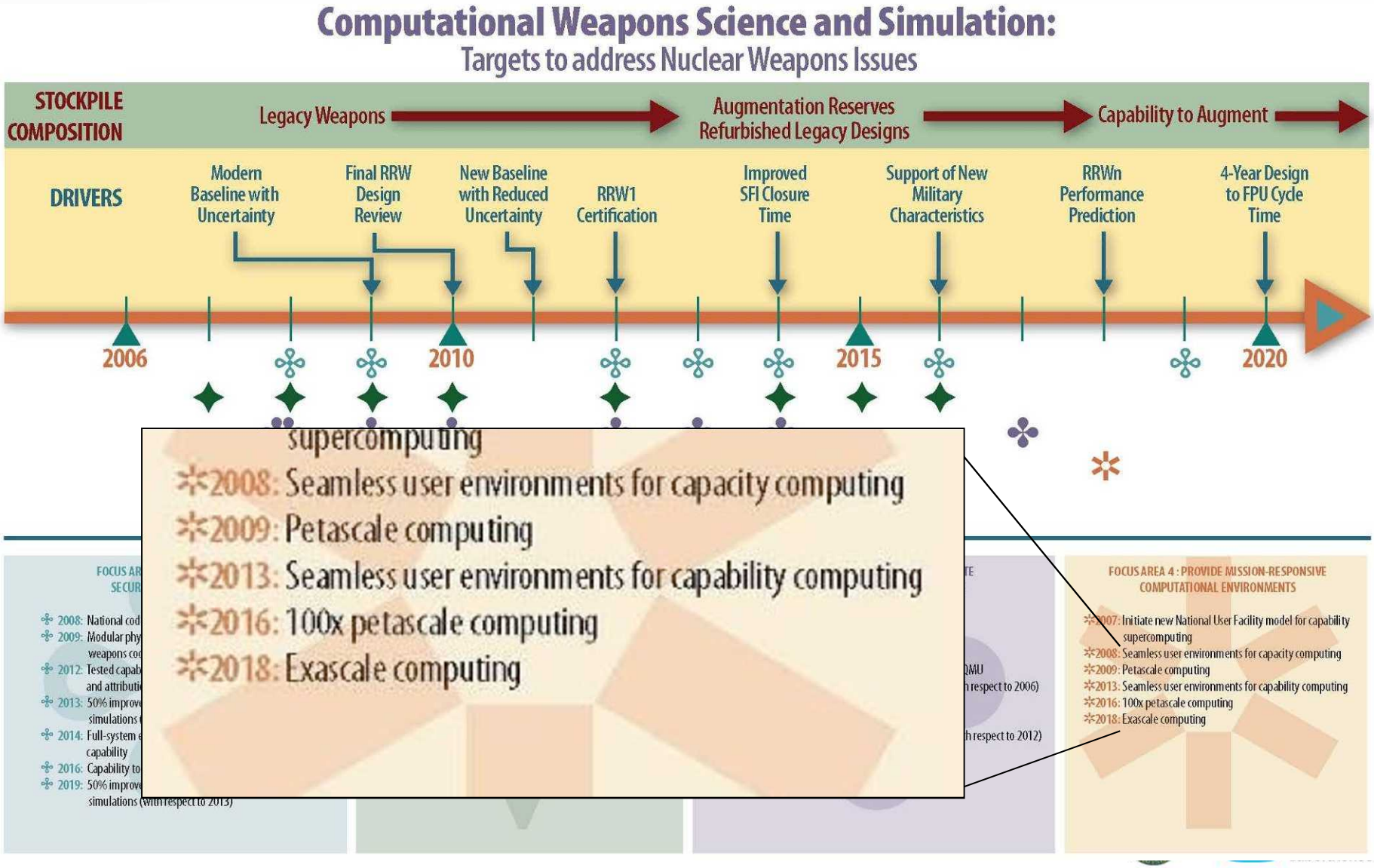
- Rightsizing our nuclear complex
- Transform the SSP in partnership with the DOD
- Modernized, cost effective NW complex
- Integrated, interdependent enterprise, best business practices
- Drive the essential science and technology base
- Urgency of beginning transformation now

Sandia National Laboratories

Vision of the Future Complex

- A smaller, safer, more secure and less expensive enterprise that leverages the scientific and technical capabilities of our workforce, and meets national security requirements
- Eliminate redundancies and improve efficiencies by consolidating missions and capabilities
- Many changes at all NNSA sites
- NNSA preferred alternative: Two capability supercomputing sites:
  - ACES
  - LLNL
- SNL continues to site NNSA capacity computers, other platforms

ASC Roadmap


Computational Weapons Science and Simulation:
Targets to address Nuclear Weapons Issues
STOCKPILE COMPOSITION
Legacy Weapons
Augmentation Reserves Refurbished Legacy Designs
Capability to Augment
DRIVERS
Modern Baseline with Uncertainty
Final RRW Design Review
New Baseline with Reduced Uncertainty
RRW1 Certification
Improved SFI Closure Time
Support of New Military Characteristics
RRWn Performance Prediction
4-Year Design to FPU Cycle Time
2006
2010
2015
2020
supercomputing
2008: Seamless user environments for capacity computing
2009: Petascale computing
2013: Seamless user environments for capability computing
2016: 100x petascale computing
2018: Exascale computing
FOCUS AREA 4 : PROVIDE MISSION-RESPONSIVE COMPUTATIONAL ENVIRONMENTS
2007: Initiate new National User Facility model for capability supercomputing
2008: Seamless user environments for capacity computing
2009: Petascale computing
2013: Seamless user environments for capability computing
2016: 100x petascale computing
2018: Exascale computing

ACES: The NNSA New Mexico Alliance for Computing at Extreme Scale

- 3/2008: LANL & SNL Memorandum of Understanding
- Joint design, architecture, development, deployment and operation of production capability systems for NNSA
- Driven by mission needs
- Commitment to the development and use of world class computing
- Continued leadership in high performance computing
- Sharing intellectual capabilities of both laboratories
- Controlled release of information

Sandia National Laboratories

ACES Strategy

- Align with and influence industry roadmaps
- Co-architect platforms, applications and algorithms recognizing that new architectures are likely to influence applications and algorithms
- Ensure pragmatic migration of ASC codes to new platforms with significant performance gains
- Encourage and foster credible competition in the supercomputing industry (procurements will be open and competitive)
- Actively promote public standards
- Focus on a broad range of applications
- Impact the supercomputing industry through market acceptance of designs and component technologies
- Be driven by cost, risk and benefit analyses
- Partner with the DOE's Office of Advanced Scientific Computing Research and other government agencies (notably DARPA and other DoD agencies)

Zia: The next ASC capability platform

- ASC Purple
  - Deployed at LLNL in 2005
  - 92.781 TF/s peak, Linpack Rmax 75.760 TF/s
  - Will be 5 years old in 2010
- ASC Red Storm
  - Deployed at SNL in 2005, upgraded in 2007
  - 127.531 TF/s peak, Linpack Rmax 102.200 TF/s
  - Will be 5 years old in 2010
- Need for Purple replacement & capability increase in 2010
  - ACES Design Team developing RFP for Zia
  - Capability production in 2010

Sandia National Laboratories

Zia Goals

- Production capability
  - Capable of running a single application across the entire machine
- Petascale
- RFP will specify minimum peak, aggregate memory bandwidth and interconnect bandwidth
- Large memory per core
- Easy migration of existing codes with reasonable increase in performance
- Key challenges: Power, reliability, scalability, usability

Institute for Advance Architectures and Algorithms (IAA)

A collaboration between Sandia National Laboratories (NNSA) and Oak Ridge National Laboratory (DOE Office of Science) too enable …

- Focused R&D on key impediments to high performance in partnership with industry and academia
- Foster the integrated co-design of architectures and algorithms to enable more efficient and timely solutions to mission critical problems
- Partner with other agencies (e.g., DARPA, NSA …) to leverage our R&D and broaden our impact
- Impact vendor roadmaps by committing National Lab staff and funding the Non-Recurring Engineering (NRE) costs of promising technology development and thus lower risks associated with its adoption
- Train future generations of computer engineers, computer scientists, and computational scientists, thus enhancing American competitiveness
- Deploy prototypes to prove the technologies that allow application developers to explore these architectures and to foster greater algorithmic richness

Industry Trends

Existing industry trends not going to meet DOE HPC application needs

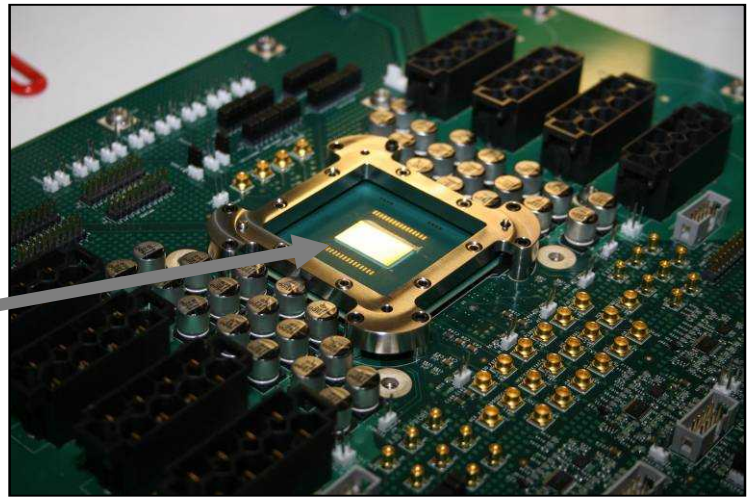
- Semi-conductor industry trends
  - Moore's Law still holds, but clock speed now constrained by power and cooling limits
  - Processors are shifting to multi/many core with attendant parallelism
  - Compute nodes with added hardware accelerators are introducing additional complexity of heterogeneous architectures
  - Processor cost is increasingly driven by pins and packaging, which means the memory wall is growing in proportion to the number of cores on a processor socket
- Development of large-scale Leadership-class supercomputers from commodity computer components requires collaboration
  - Supercomputer architectures must be designed with an understanding of the applications they are intended to run
  - Harder to integrate commodity components into a large scale massively parallel supercomputer architecture that performs well on full scale real applications
  - Leadership-class supercomputers cannot be built from only commodity components

Sandia National Laboratories

Moore's Law Continues

But in a different Way:
More functionality, Clock rate increases flatten out

1997: ASCI Red
- **1 TeraFLOP in a room**
- **2,500 ft^2 & 500 KW**



Moore's Law + Multicore → Rapid Growth in Computing Power

2010-2012: TeraFLOP on a chip
275 mm^2 (size of a dime) & < 150 W

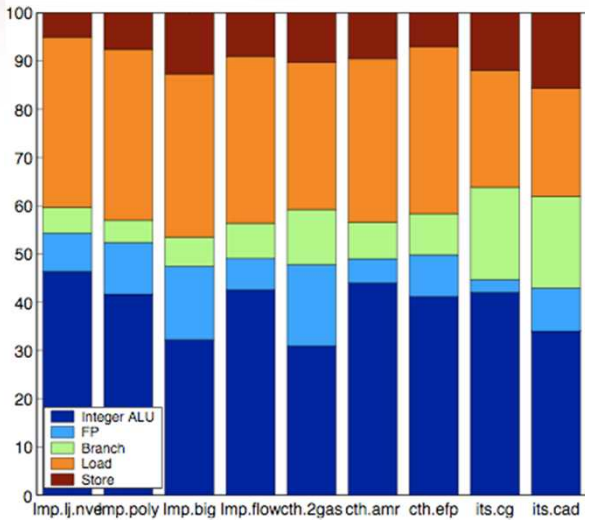
Sandia National Laboratories

Impediments to Useful Exascale Computing

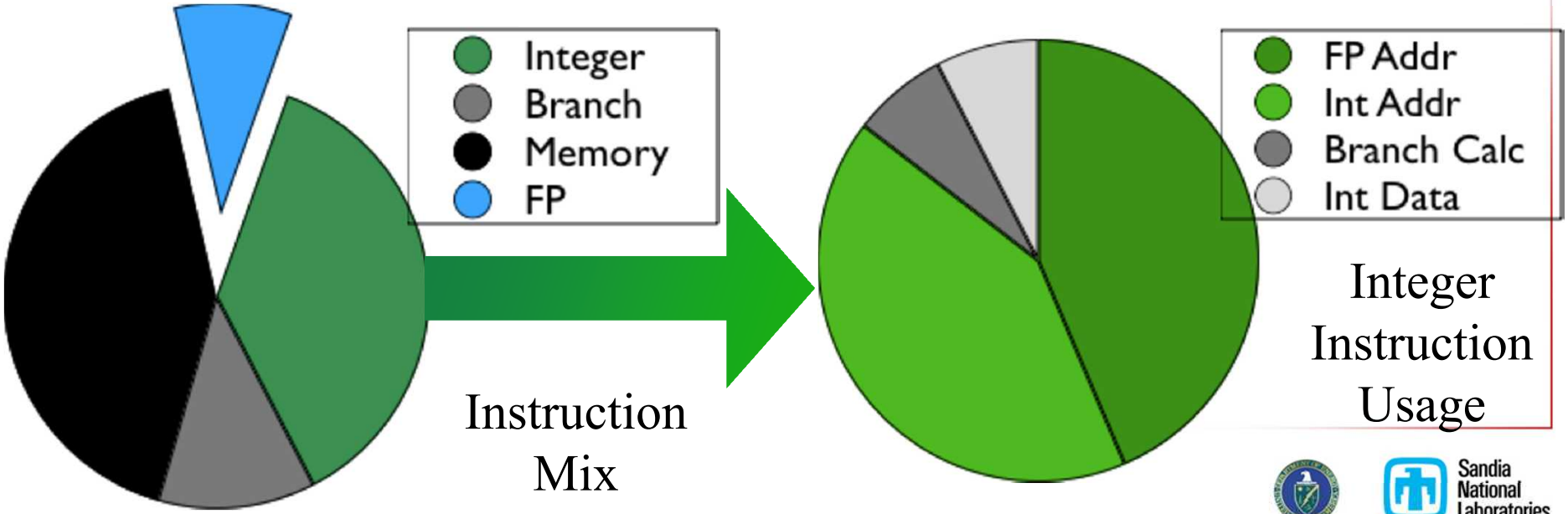
- Data Movement
  - Local
    - cache architectures
    - main memory architectures
  - Remote
    - Topology
    - Link BW
    - Injection MW
    - Messaging Rate
  - File I/O
    - Network Architectures
    - Parallel File Systems
    - Disk BW
    - Disk latency
    - Meta-data services
- Power Consumption
  - Do Nothing: 100 to 140 MW
- Scalability
  - 10,000,000 nodes
  - 1,000,000,000 cores
  - 10,000,000,000 threads
- Resilience
  - Perhaps a harder problem than all the others
  - Do Nothing: an MTBI of 10's of minutes
- Programming Environment
  - Data movement will drive new paradigms

Sandia National Laboratories

Memory Operations Dominate



- FP ops ("Real work") < 10% of Sandia codes
- Several Integer calculations, loads for each FP load
- Memory and Integer Ops dominate
  - ...and most integer ops are computing memory addresses

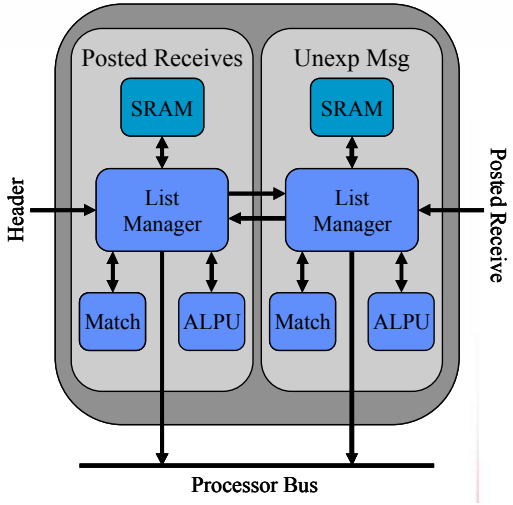


Sandia National Laboratories

High Speed Interconnects are Essential to Ensure Scalability

Vision: Ensure next generation interconnects satisfy HPC needs

Approach: Provide understanding of application needs, explore designs with simulation, prototype features with vendors

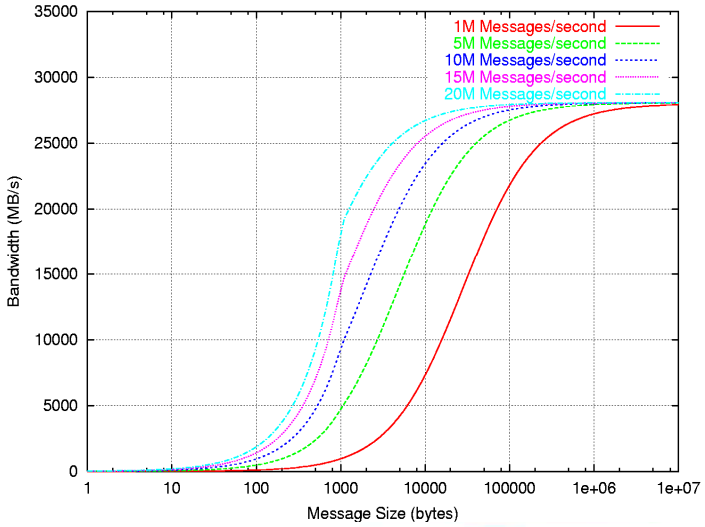


Long Term Goals:

- *Scalability*: >100,000 ports (including power, cabling, cost, failures, etc.)
- *High Bandwidth*: 1TF sockets will require >100GBps
- *High Message Throughput*: >100M for MPI; >1000M for load/store
- *Low Latency*: Maintain ~1us latency across system
- *High Reliability*: $<10^{-23}$ unrecovered bit error rate

Near Term Goals:

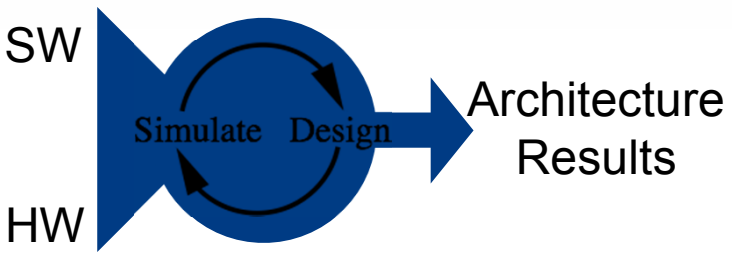
- Identify interconnect simulation strategy
- Characterize interconnect requirements on mission apps
- Develop MPI models & tracing methods
- Pursue small collaboration project with industry partner

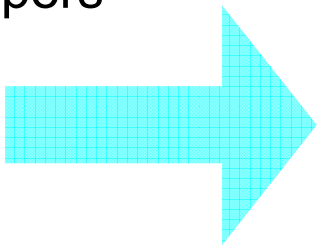


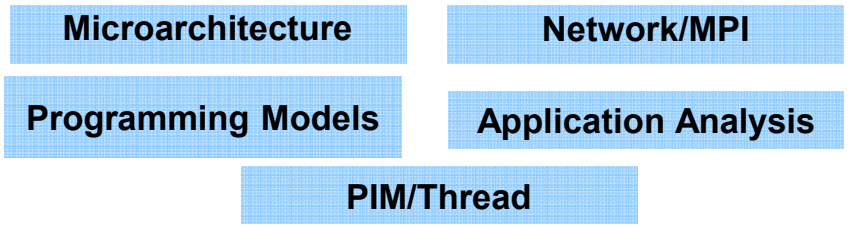
Sandia National Laboratories

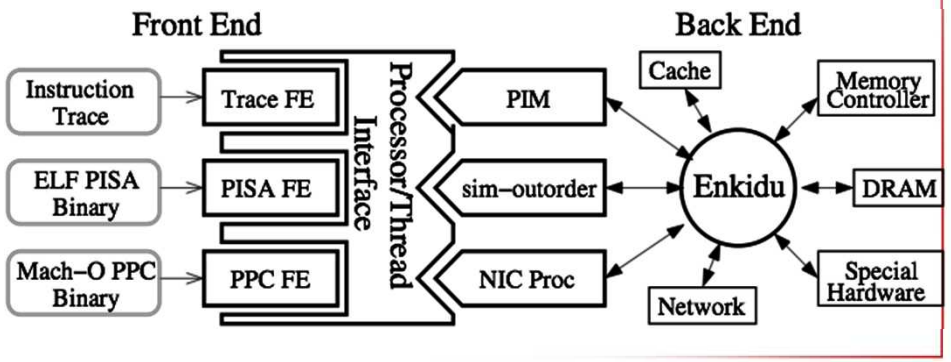
Structural Simulation Toolkit

- Novel Architectures require novel hardware and software
- Customers
  - Application Developers
  - System Architects
  - Microarchitects
- Requirements
  - Multiple Programming Models (Front-ends)
  - Multiple Architectures (Back-ends)
  - Fast turnaround
  - Low Overhead
  - Modular / Reusable









Sandia National Laboratories

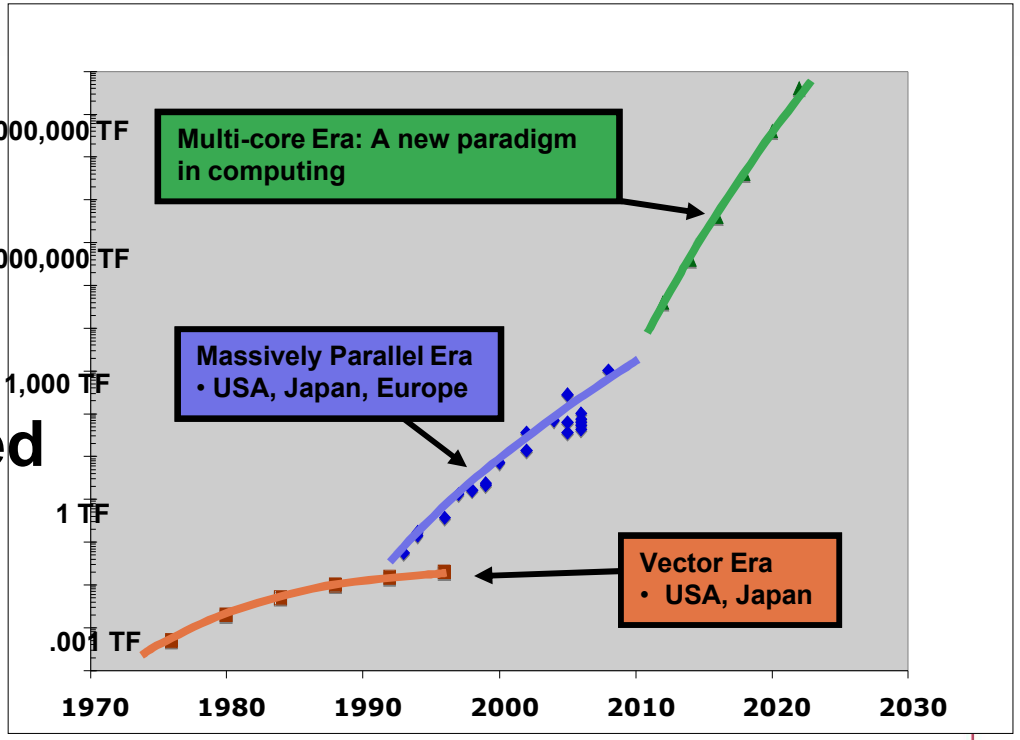
Sandia & the DOE are Looking to Enable a New Paradigm Shift in Computing

New Initiatives:

- **ACES: Alliance for Computing at Extreme Scale**
  - **SNL/LANL partnership**
- **IAA: Institute for Advanced Architectures**
  - **SNL/ORNL partnership**

Ongoing Initiatives:

- **Capability Computing**
  - **Red Storm**
- **Capacity Computing**
  - **Thunderbird, NWCC, ICC, …**

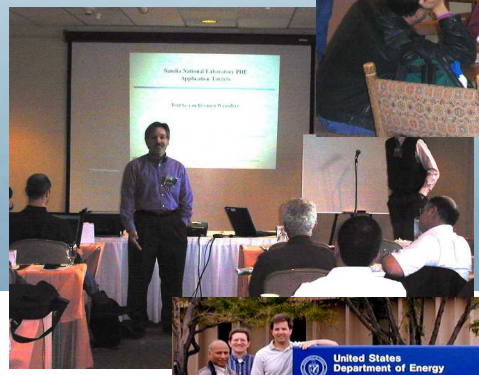


Sandia National Laboratories

CSRI
Computer Science Research Institute

Goal: engage university faculty and industry experts in on-site collaborative research in computer science, computational science and mathematics in support of Sandia's modeling and simulation capabilities.

- New building "done"
- 42 short-term visitors
- ~50 summer students and faculty
- 1 long-term faculty visitor
- 3 workshops with ~50 external attendees
- 7 fellows (HPCS, NPSC, CSGF)



Sandia National Laboratories

Backup Slides

Solvers (Trilinos)

Recent Work

- Trilinos 5.0 released; >1000 downloads
- Focus on continued package development and software engineering

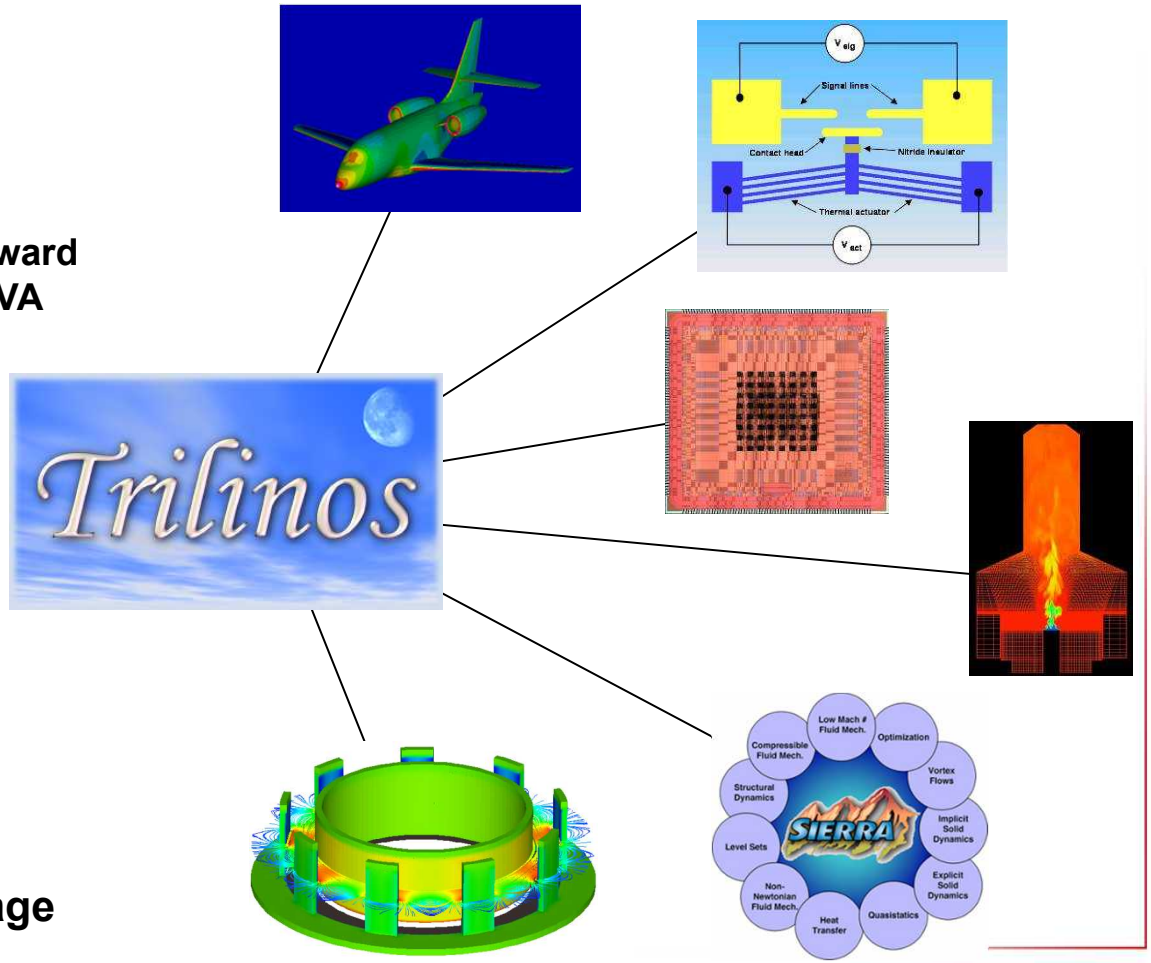
Awards

- R&D 100 (in 2004)
- IEEE HPC Software Challenge Award
- Sandia ERA team award and NOVA nomination

First software architecture to allow community development

- Two-level design:
  - Self-contained packages
  - Leveraged common tools.
- Allows rapid algorithmic development and delivery

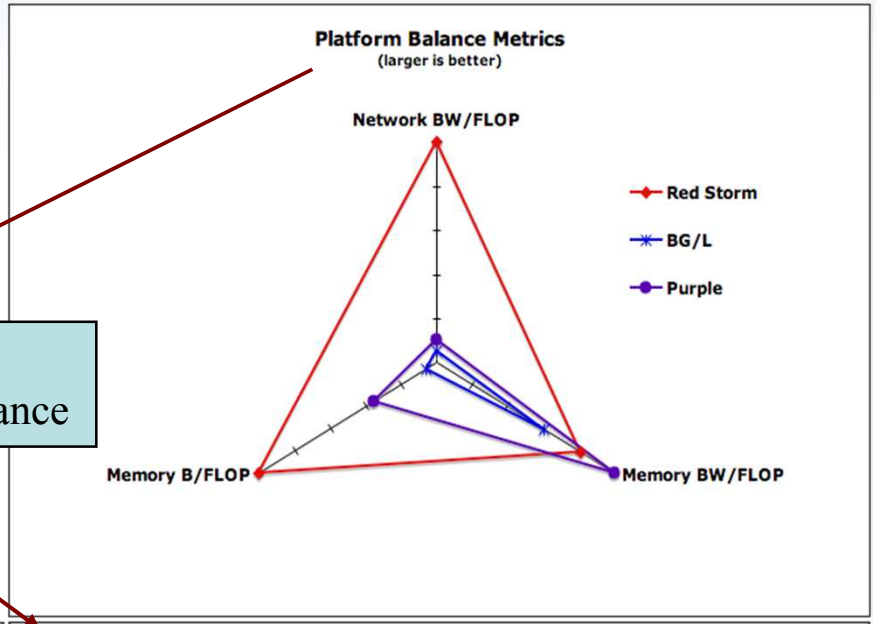
Leveraging investments in software infrastructure without compromising individual package autonomy



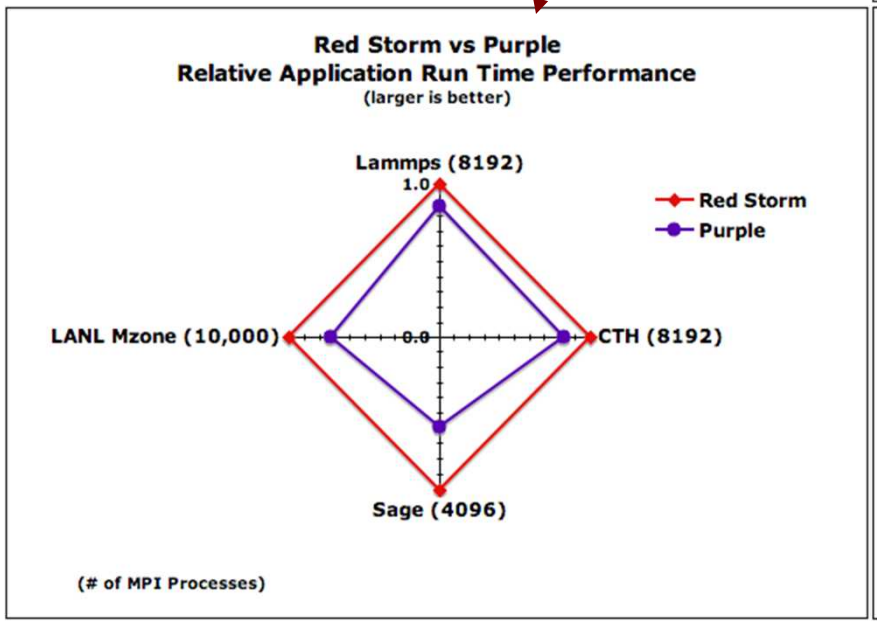
Sandia National Laboratories

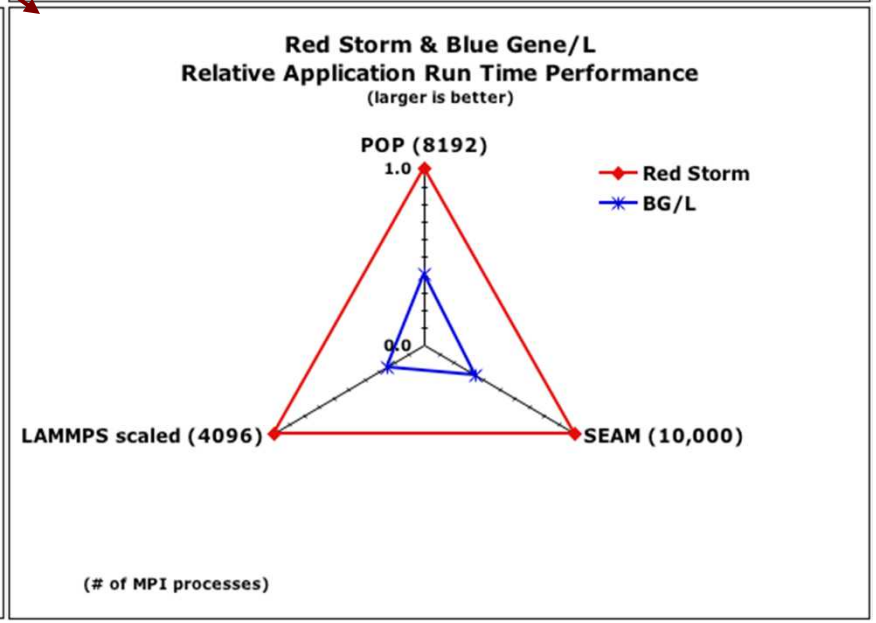
The Impact of a Balanced Architecture

- Architectural balance with low system noise is the key to a scalable platform



Well Balanced Traits Translate to
High Real World Application Performance





Challenges in Building Scalable HPC Networks

Scott Hemmert
Scalable Computer Architectures
Computation, Computers, and Mathematics Center
Sandia National Laboratories
Albuquerque, NM

March 11, 2008

Sandia is a Multiprogram Laboratory Operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy Under Contract DE-ACO4-94AL85000.

Scale of Future Systems

- Capability Class Computing
  - Ability to run jobs across entire machine
    - Whole machine jobs are the standard, not the exception
  - Parallel efficiency likely determines application efficiency
  - Interconnect performance is the most critical component in determining parallel efficiency
    - High reliability
    - Balanced system
      - Past experience shows a good network balance at 1 Byte/FLOP of aggregate bandwidth, both into a node and within each link

- Well provisioned networks can allow for future upgradability
  - Multiple generations of processors in same socket
  - Even rev the board with no wiring change
- 2009-2010 Timeframe
  - 20k-40k Sockets @ 40-100 Gflops/socket
- 2013-2014 Timeframe
  - 20k-100k Sockets @ 400-2000 Gflops/socket

Red Storm

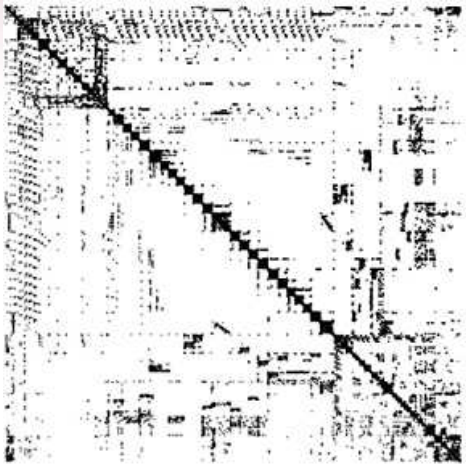
(2004) 40 TF: 10360 sockets @ 4.8 Gflops/socket

(2006) 124 TF: 12960 sockets @ 9.6 Gflops/socket

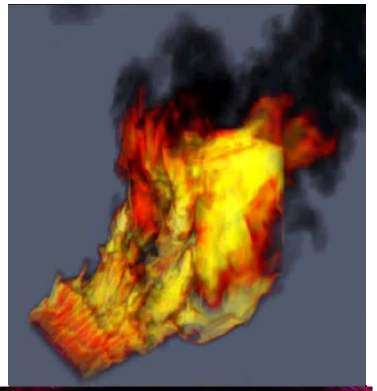
(2008) 284 TF: See press release…

Sandia National Laboratories

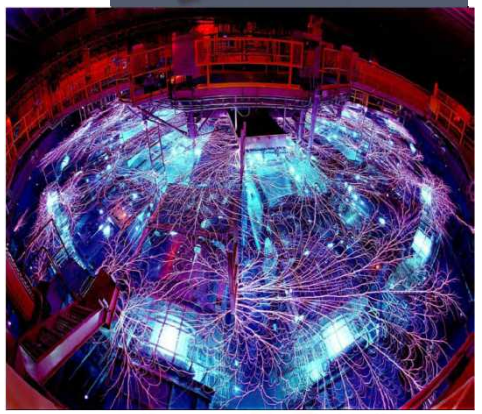
Application Characteristics: Traditional Physical Simulations



- Large-scale physics and engineering applications
  - Able to utilize the entire Red Storm machine
- Basis in physical world leads to natural 3-D data distribution
  - Communication to nearest neighbors, in 3 dimensions
  - Peers limited even for adaptive mesh refinement



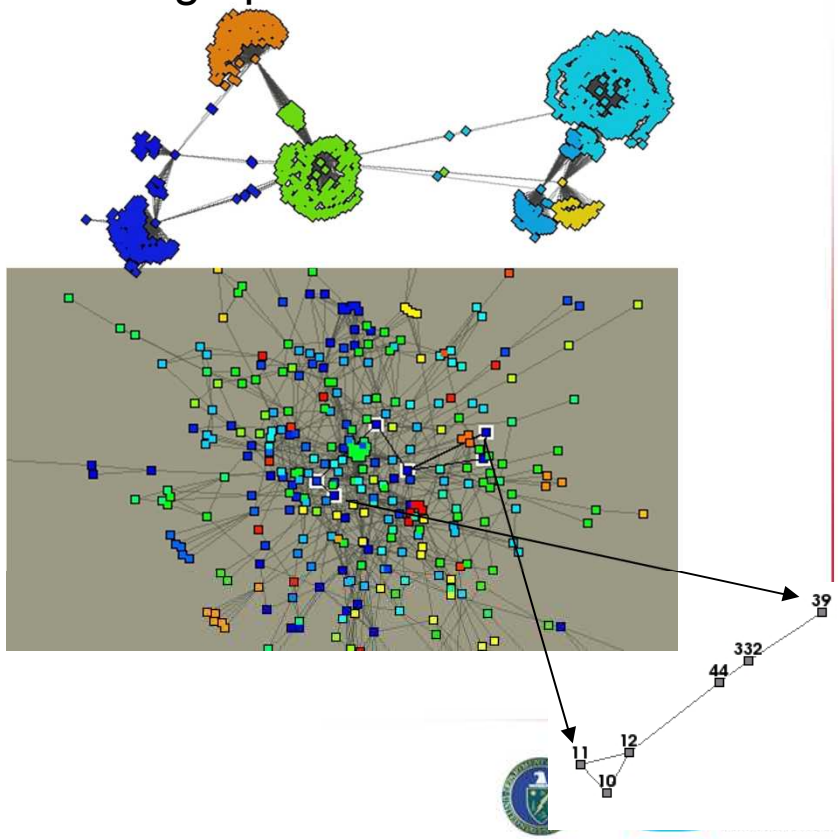
- Ghost cell update messages sent from packed buffers
  - MPI historically bad at sending derived datatypes
  - Poor message rates led to buffering non-contiguous slices of the 3-D data space
- Point-to-point communication largely ghost cell updates, range in size from word-length to megabytes
- Collective communication
  - Double precision floating point all-reduce
  - Varied sized broadcasts, particularly during startup
- Ghost cell updates implicitly synchronize time-steps



Application Characteristics: Emerging Informatics Applications

- Graph-based informatics applications emerging as an important application space
- No natural data partitioning
  - Random communication patterns
  - Fully-connected point-to-point communication graph

- Very small (word size) messages
- Higher injection rate requirements than physics codes
  - More outstanding requests
- Communication paradigm still being developed
  - Non-MPI matching requirements could help with injection rate
  - Remote addressing, ordering issues still open to exploration



Challenge Areas for HPC Networks

- The traditional “big three”
  - Bandwidth
  - Latency
  - Message Rate (Throughput)
- Other important areas for “real applications” versus benchmarks
  - Allowable Outstanding Messages
  - Host memory bandwidth usage
  - Noise (threading, cache effects)
  - RDMA effects
  - Topology
  - Reliability

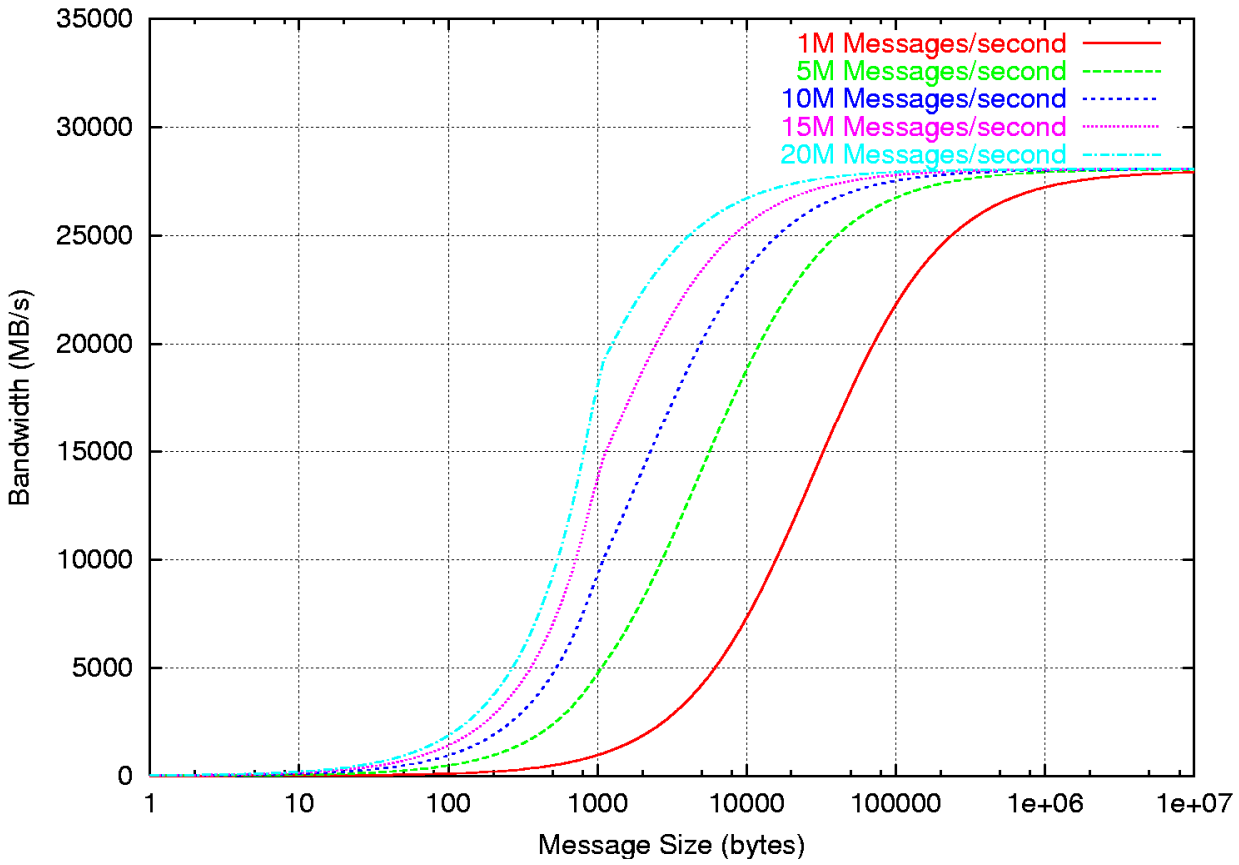
Requirements for 2009-2010 Timeframe

- For 40-100 GFlop nodes we would like to see unidirectional network bandwidth of 20-50 GB/sec

Network	Bandwidth (GB/sec/direction)	Latency (MPI) (nanoseconds)	Throughput (MPI) (messages/sec/dir)
Red Storm (2004)	2	4000	400,000
Next Gen (2009)	30	800	20,000,000
Improvement Needed	15X	5X	50X

- Off-load needed to meet these requirements for operating conditions of real applications
  - Systems tend to be analyzed using microbenchmarks which simulate ideal (i.e., unreal) operating conditions

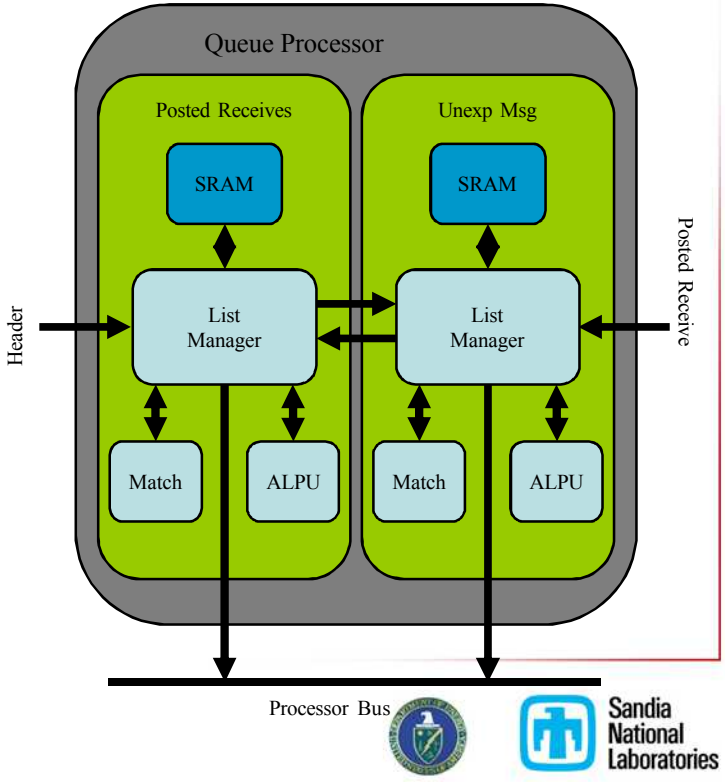
High Message Throughput is Vital



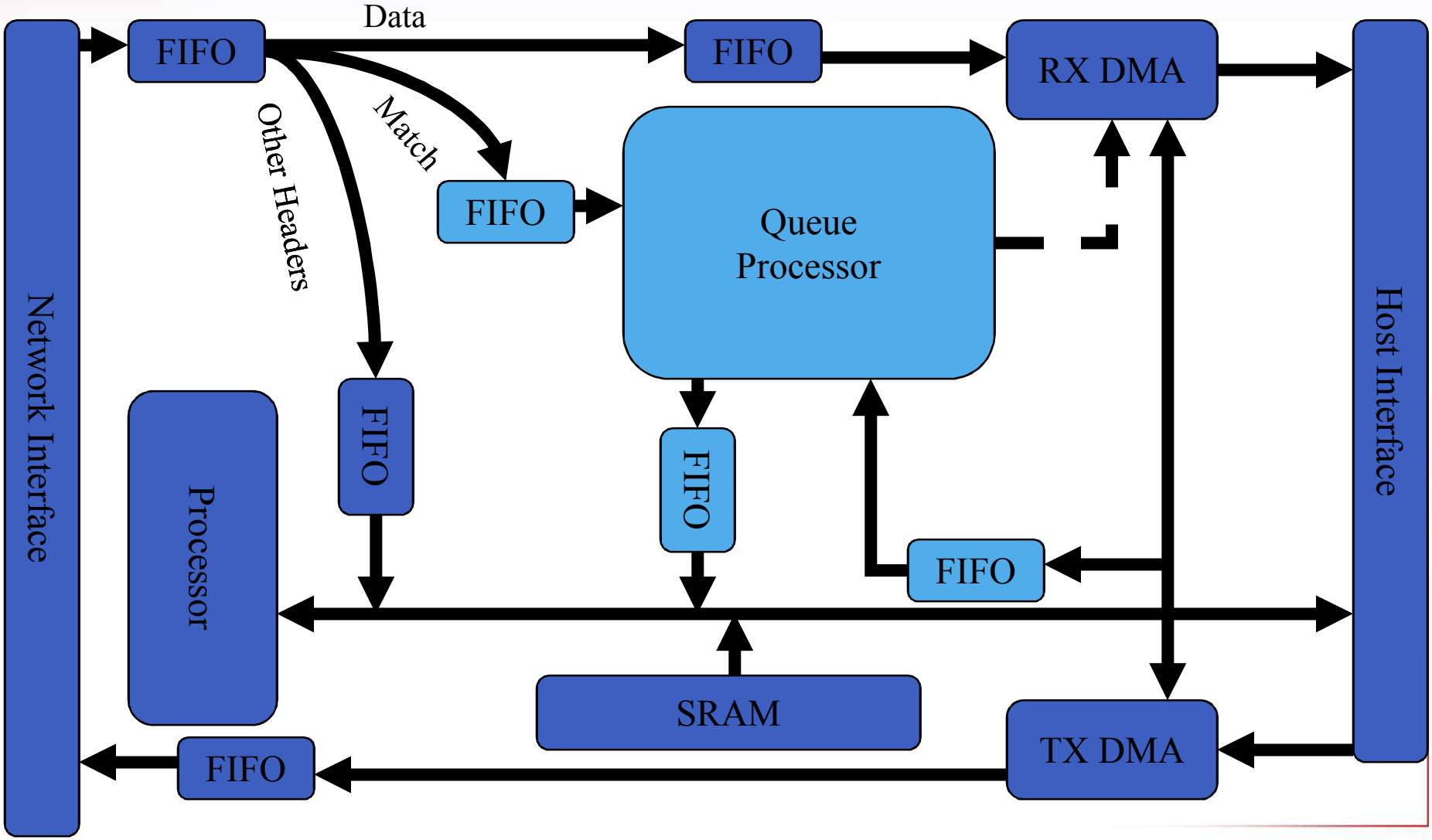
Message rate determines the minimum message size needed to saturate the available network bandwidth

High Message Throughput Challenges

- 20M messages per second implies a new message every 50ns
- Significant constraints created by MPI semantics
- On-load approach
  - Roadblocks
    - Host general purpose processors are inefficient for list management
    - Caching (a cache miss is 70-120ns latency)
      - Microbenchmarks are cache friendly, real life is not
  - Benefits
    - Easier & cheaper
- Off-load approach
  - Roadblocks
    - Storage requirements on NIC
    - NIC embedded processor is worse at list management (than the host processor)
  - Benefits
    - Opportunity to create dedicated hardware
    - Macroscale pipelining

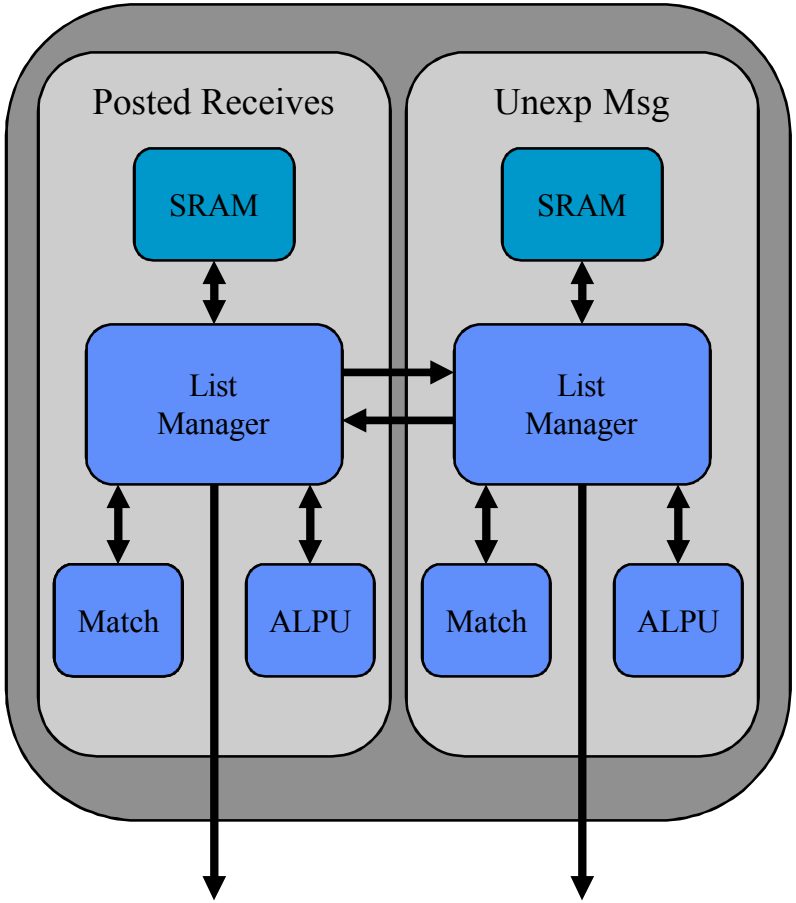


Proposed NIC Architecture



Sandia National Laboratories

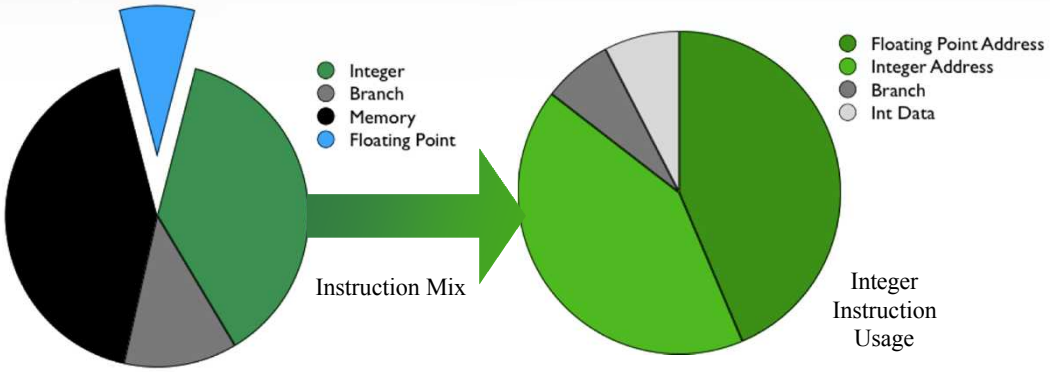
Queue Processor



Microcoded Match Unit

- **Accomplishes the matching which compares headers and posted receives**
- **Provides flexibility**
  - **Change meaning/location of match bits**
  - **Add additional functionality**

Minimizing Memory Bandwidth Usage in Network Stack



Memory Usage in Sandia Applications

- Memory bandwidth is most often the limiting factor for on node performance
- We must minimize the use of host memory bandwidth in the network stack

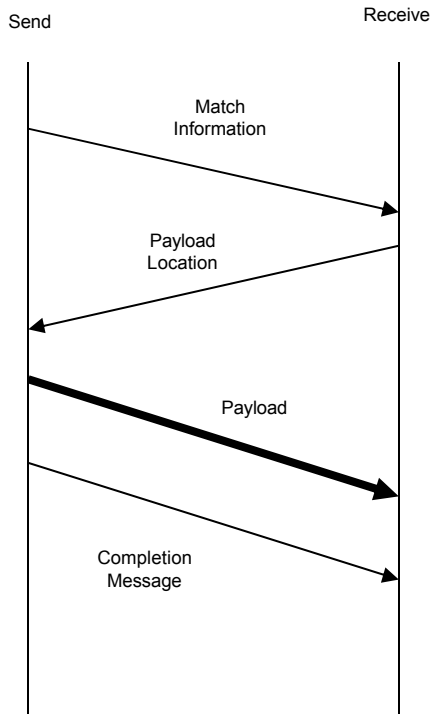
- Bounce buffers (or any other copying) incur a 2x memory bandwidth penalty
- A fast off-load approach can minimize host memory bandwidth utilization
  - Allows the NIC to determine where received messages need to be put in host memory and DMA the data directly there, eliminating the need for bounce buffers
  - High message rate can reduce the need for buffering of non-contiguous data

Network Induced Noise

- Traditional physics codes exhibit well balanced computation phases
  - Late-comers to communication phases slow down entire job
  - Important to minimize random interruptions to computation phase
- Causes of system noise:
  - Operating system timers
  - Interrupt based message processing
  - Threaded communication libraries (for asynchronous progress)
- Use of dedicated cores for operating system and network
  - Solves noise problems due to application being swapped off the processor
  - Shared cache structures mean operating system and network can still negatively impact application performance

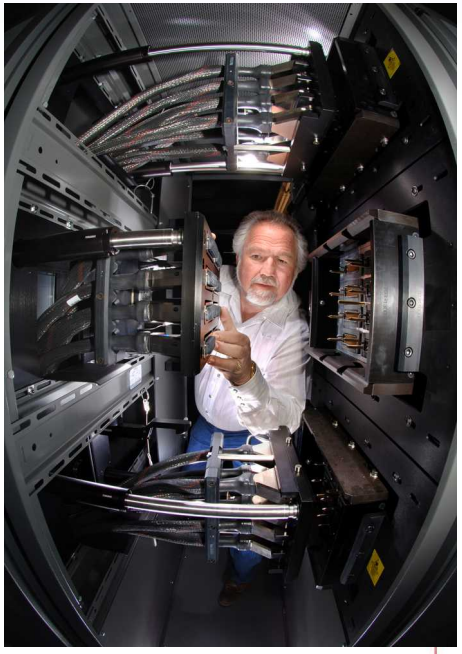
RDMA Effects

- OS Bypass RDMA provides high bandwidth and low CPU overhead
- Message destination of endpoint / virtual address means software message matching
- Control messages needed to determine remote data location
  - Destination cannot be determined receive-side
- Some networks do not provide remote completion, requiring another control message
- Asynchronous progress without threads difficult, particularly for expected receives



Topology

- No single network topology is best for all applications
- Meshes
  - Advantages: high local bandwidth, low wiring complexity, ability to easily add nodes (no distinct steps in expandability curve)
  - Disadvantages: high maximum latency, low global bandwidth
  - Works well for physical simulations which tend to talk nearest neighbor
- Trees
  - Advantages: high global bandwidth, low global latency
  - Disadvantages: high wiring complexity, lower local bandwidth, discrete steps in network topology (limiting expandability)
  - Works well for random accesses patterns and apps with lots of global communication



Red Storm
Red/Black Switching

- Hierarchical/Hybrid networks
  - Tend to inherent the strengths and weaknesses of the building blocks
- Take home message: Network routers must be designed to allow different topologies with the same silicon

Summary

- Biggest obstacles for future interconnects
  - Message throughput
  - Efficient host memory bandwidth usage by network stack
  - Microbenchmarks
    - Or more accurately, designing to microbenchmarks with no thought to real application operating conditions

Sandia National Laboratories

Bonus Slides

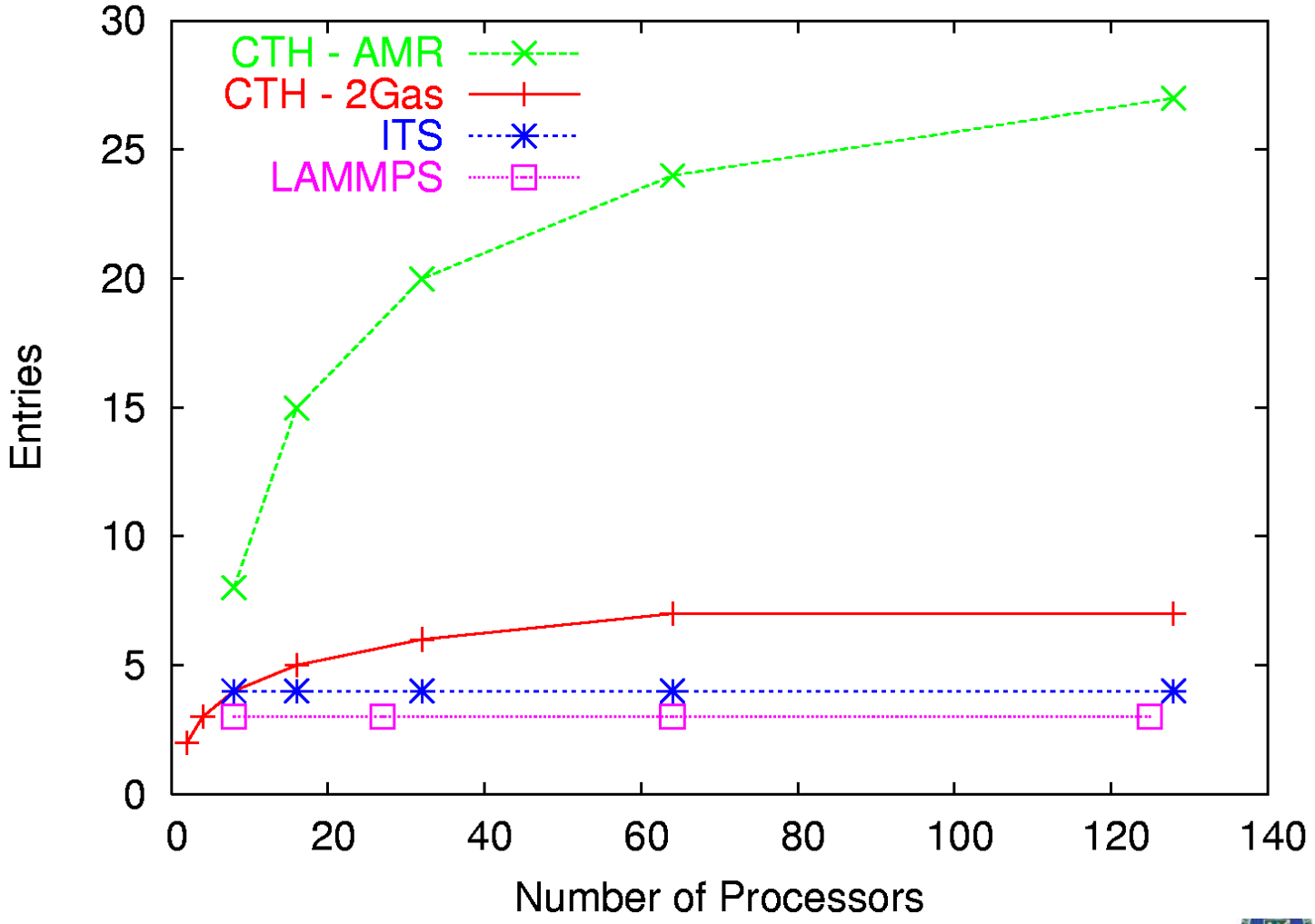
If we get here you didn't ask enough questions

Sandia National Laboratories

MPI Queue Processing

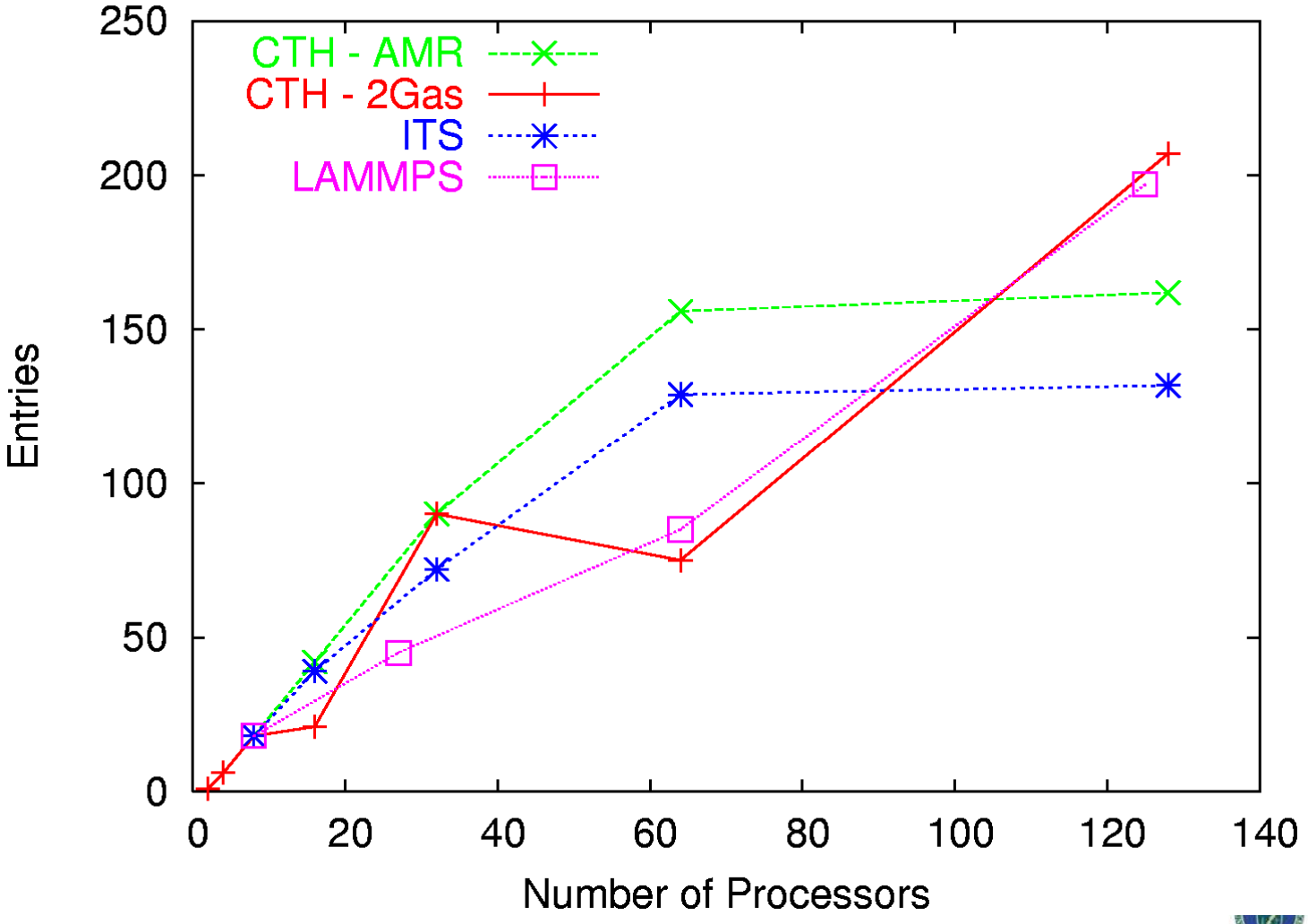
- Posted Receives Queue
  - Queue of receives that have been posted, but have not yet been received
  - Only occurs with non-blocking communications
  - Traversed each time a message is received
- Unexpected Message Queue
  - Queue of messages that have arrived, but do not have a matching receive call
  - Occurs for both blocking and non-blocking communications
  - Traversed each time a receive is posted

Maximum Posted Queue Search



Sandia National Laboratories

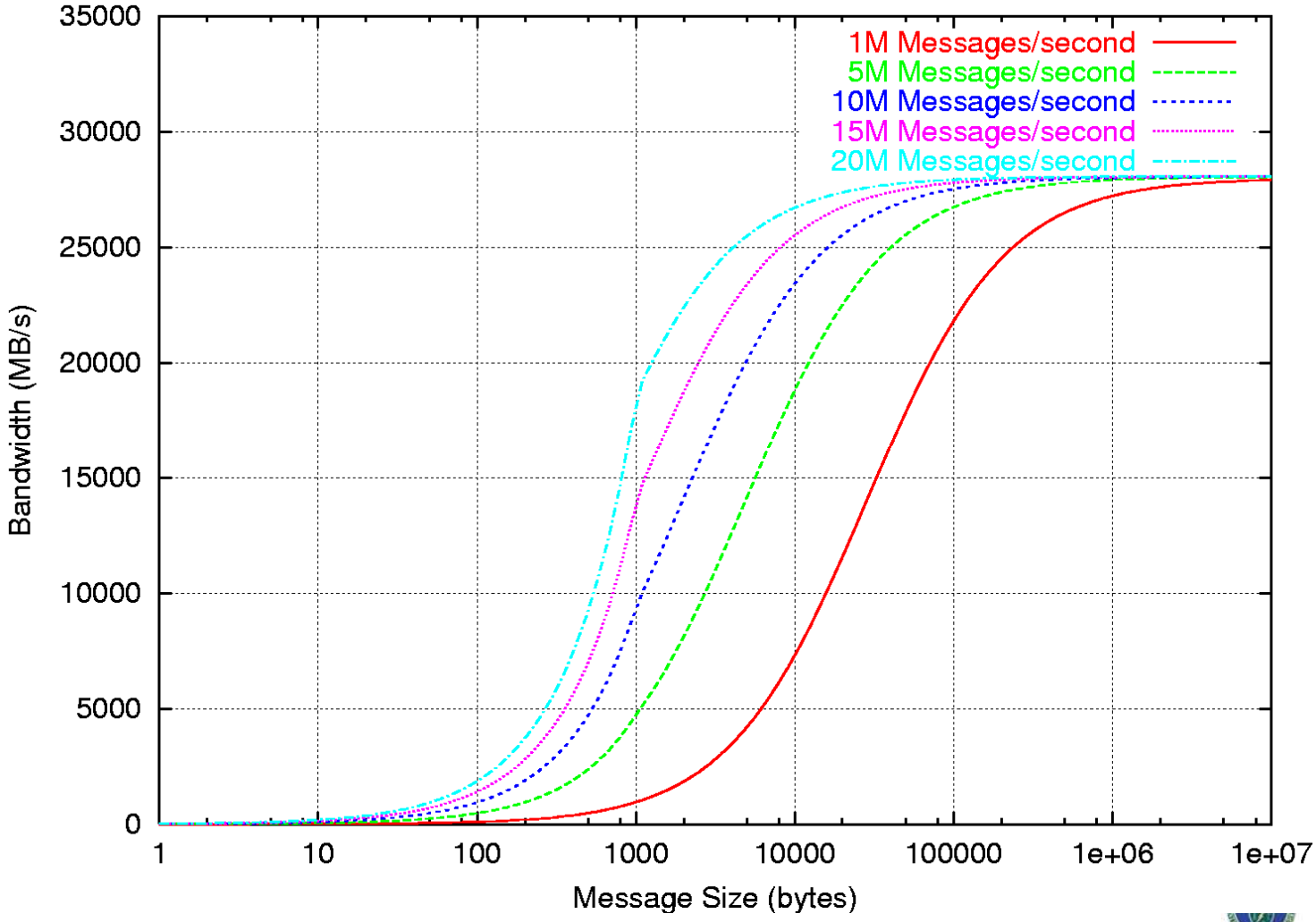
Maximum Unexpected Queue Search



Motivation

- Applications have long queues
- Good NICs offload MPI semantics
- Result: long queues are traversed by an embedded processor on the NIC
  - This decreases throughput for long queues
  - Throughput will become key as bandwidths increase
- Question: how can we increase the performance of NIC based offload of MPI processing?

High Message Throughput is Vital

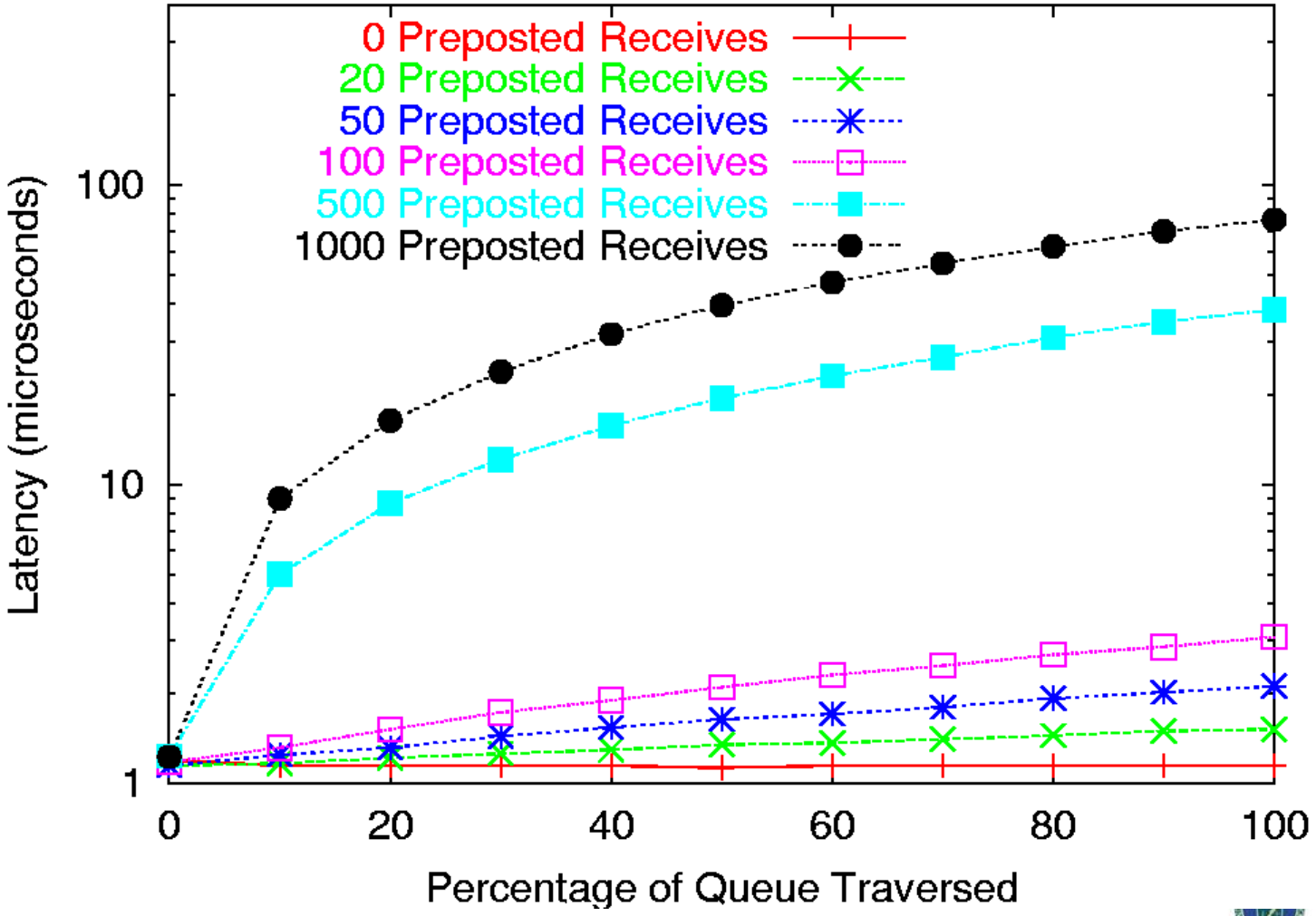


Sandia National Laboratories

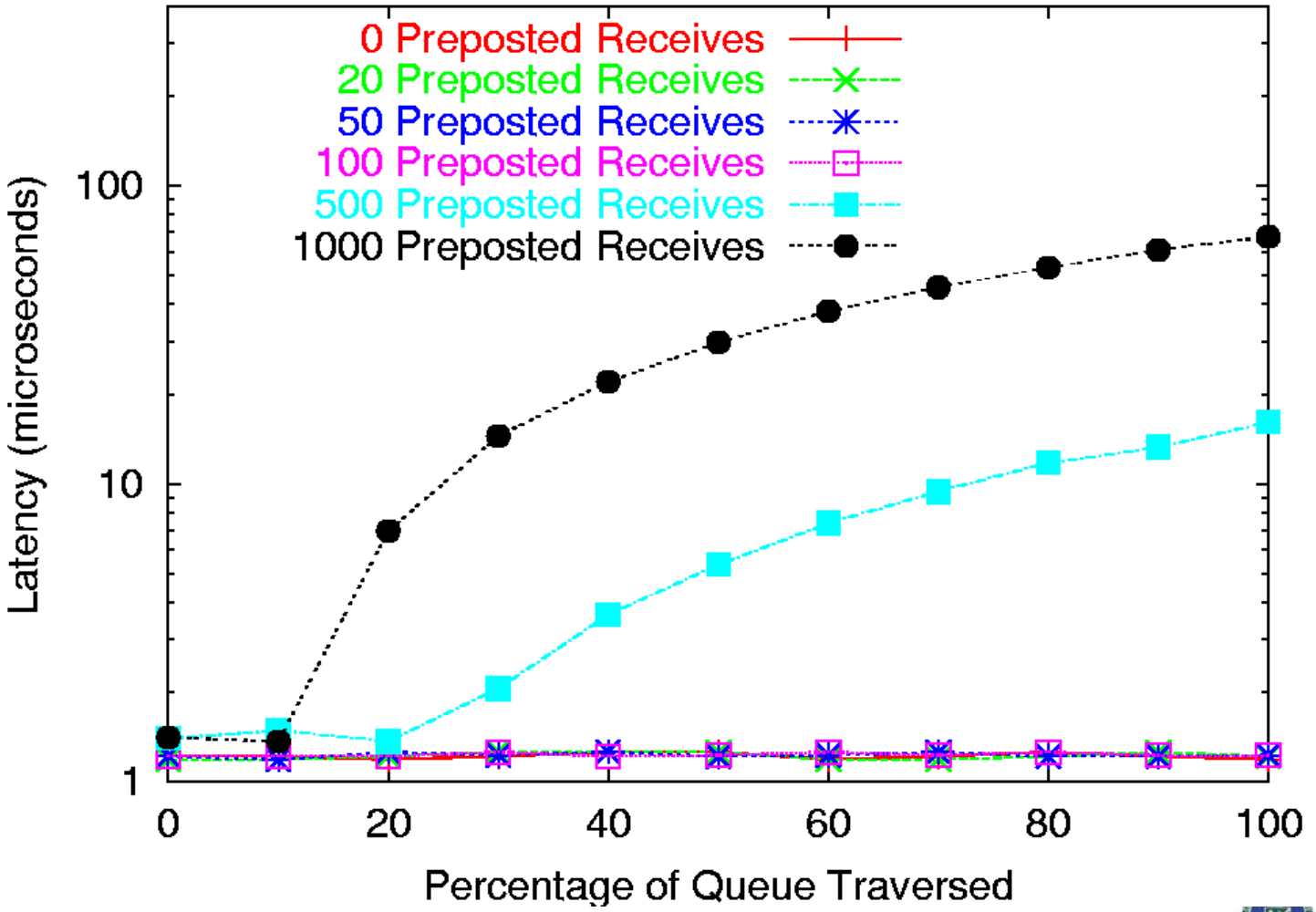
Difficult Challenges

- Significant constraints created by MPI semantics
  - Match criteria can contain wildcards
  - Matching must maintain strict ordering
  - High turn-over rate for match entries
- Even list management can be a large overhead on the NIC processor
  - Recognize new posted receive
  - Insert item into NIC list
  - Move to completion queue after match
  - Delete after receive completed

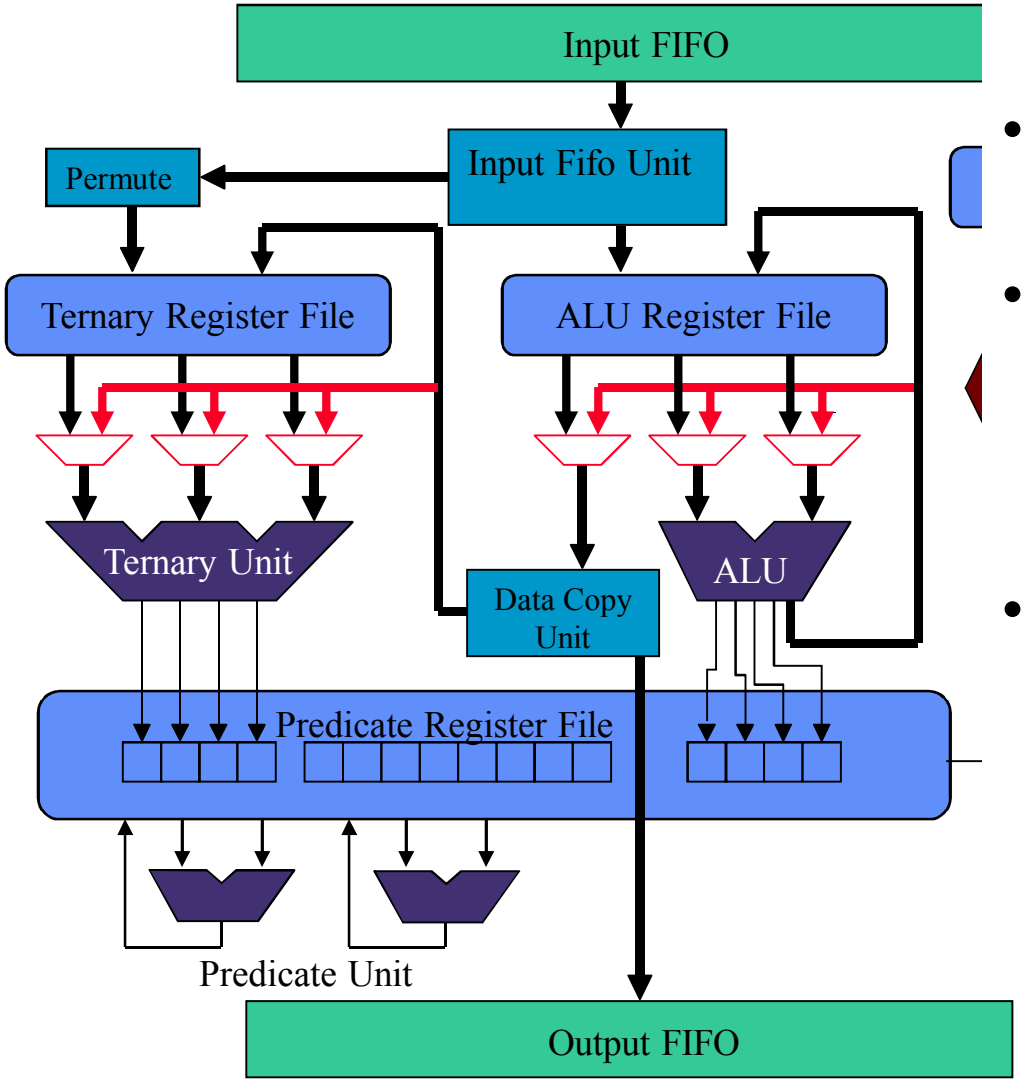
Posted Queue Results – Baseline



Posted Queue Results – 128 Entry ALPU



Match Unit Architecture



Architecture Drivers

- **High throughput**
  - **3 stage pipeline**
- **Irregular data alignment**
  - **SIMD operation**
  - **Permute units**
- **Program Consistency**
  - **Forwarding in datapath**
  - **Read before write in register file**

Sandia National Laboratories

SIMD Operation

- Similar to typical SIMD operations, but primary motivator is the ability to operate on data with irregular alignment
- A single 64-bit operation is actually 4 16-bit operations
- 16-bit operations can be “bridged” to form larger operations
- 3 SIMD bits specify which boundaries are bridged
  - Examples:
    - 000: 16-16-16-16; 100: 32-16-16; 011: 16-48; 111: 64

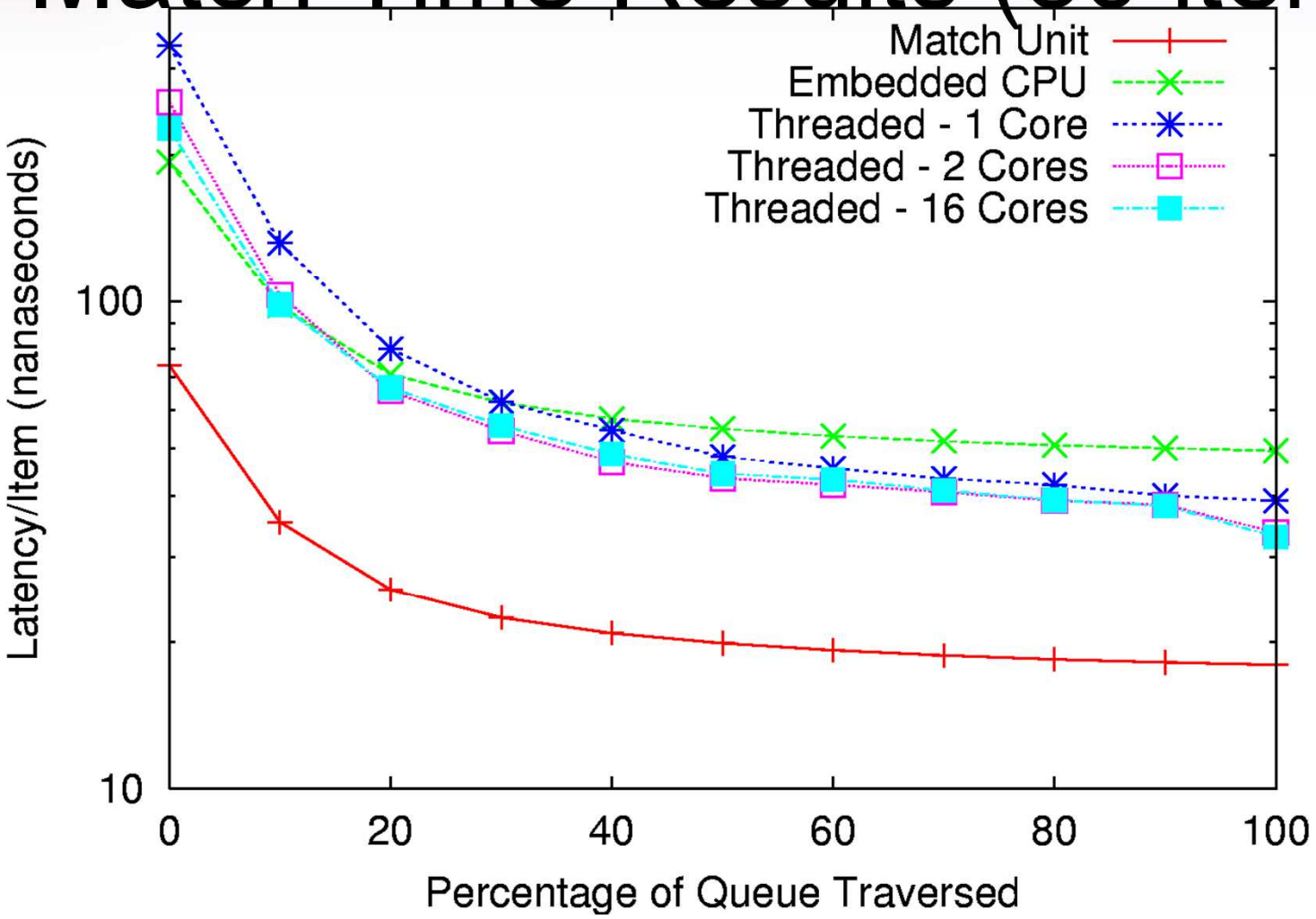
Evaluation Methodology

- Simulated 5 configurations
  - Microcoded match unit
  - Embedded processor similar to PowerPC 440
  - 3 multithreaded units
    - 1, 2 and 16 cores
- Execution-based simulator for processors (SST)
- Cycle accurate simulator for match unit (JHDL)
- All configurations run at 500MHz
- Varying memory bandwidths
  - Match unit: 8 bytes/cycle
  - Embedded proc/1 and 2 MT cores: 16 bytes/cycle
  - 16 MT cores: 32 bytes/cycle

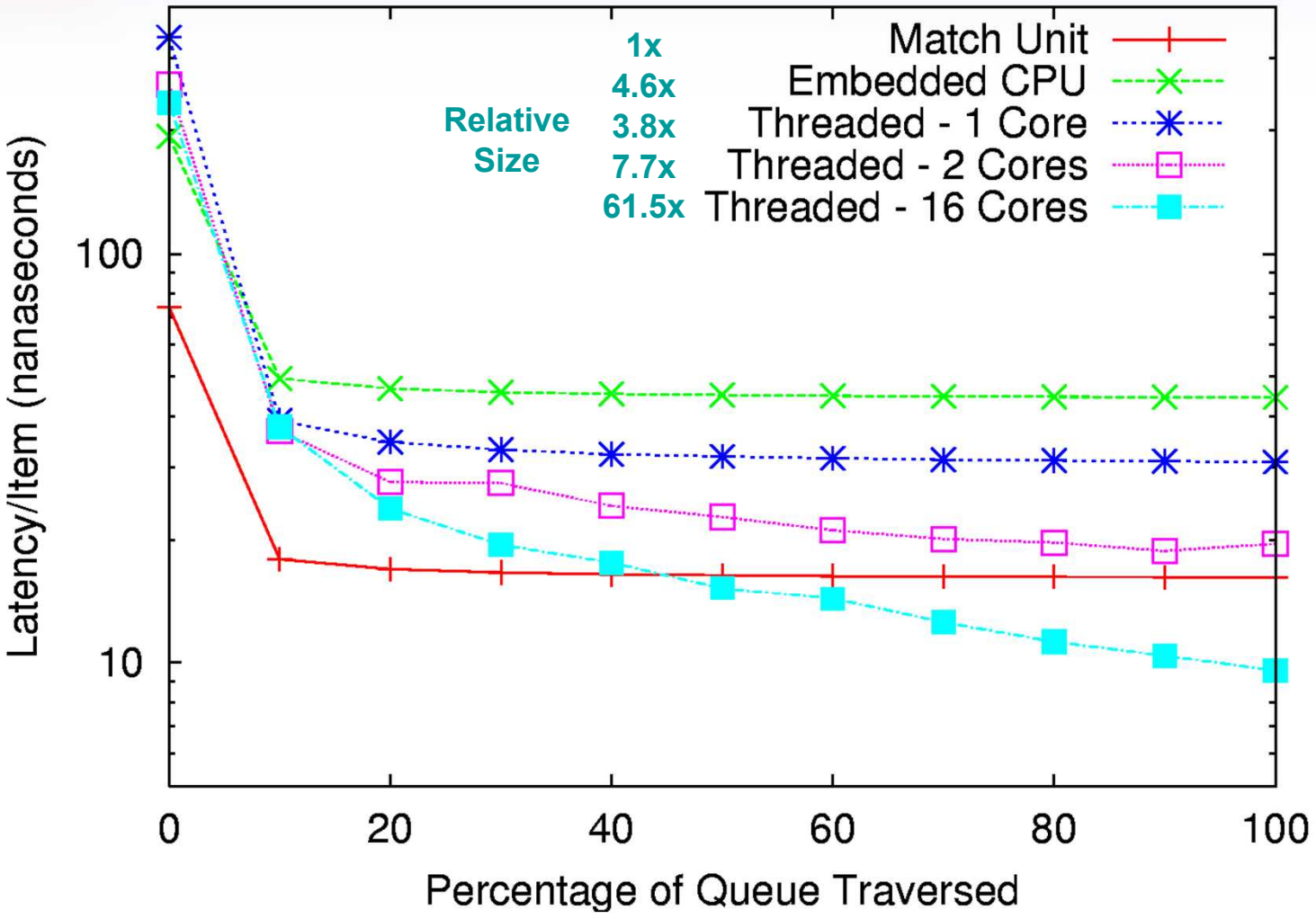
Benchmark Characteristics

- Measures time to match a header given the length of the posted receives queue and the percentage of the queue traversed
- To facilitate timing for the multithreaded units, 64 identical headers are sent
  - Multithreaded units provide no advantage for processing a single header
- To measure only match time, the queue item is not deleted on a match
- Instruction cache is primed before timing begins

Match Time Results (30 items)

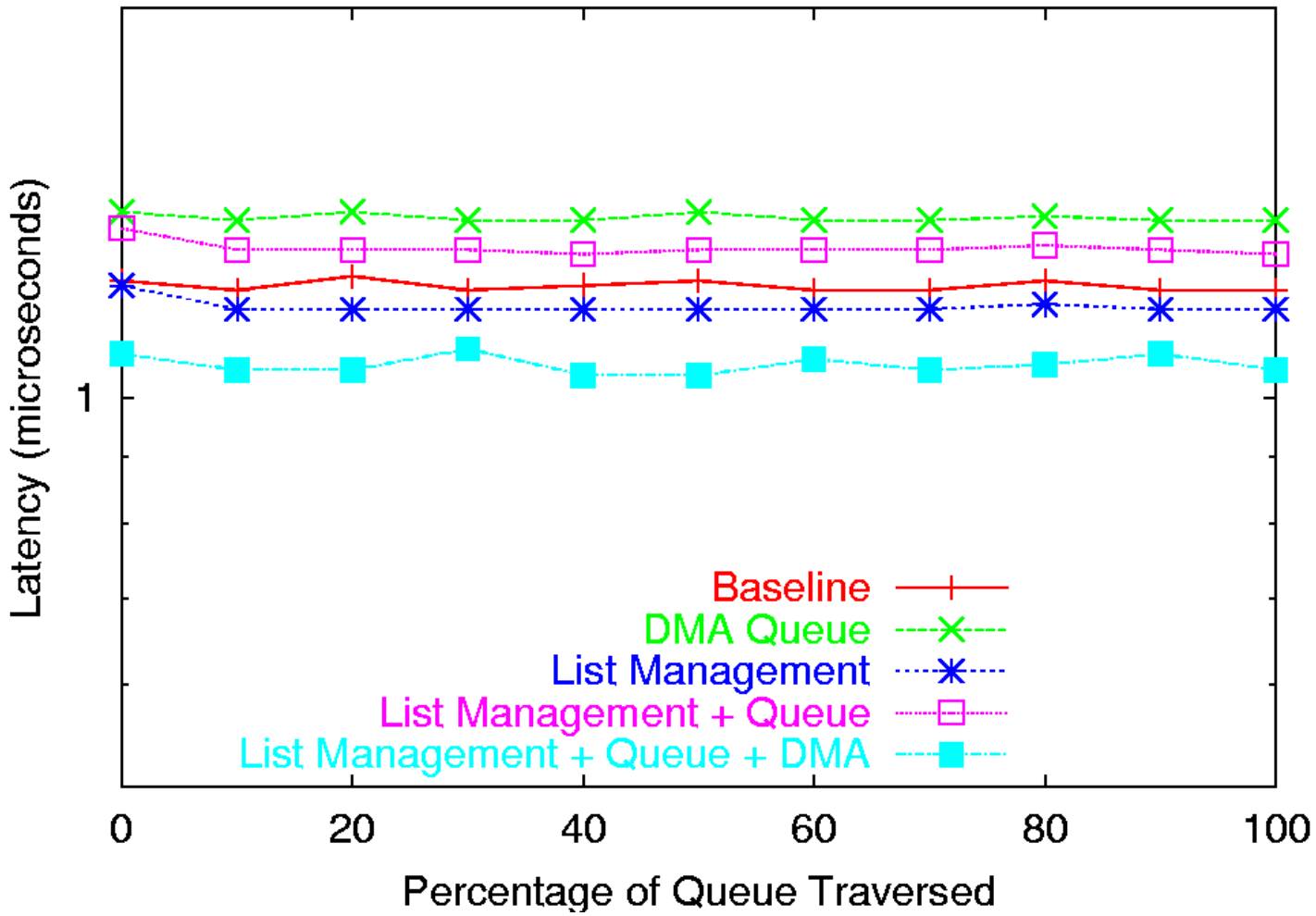


Match Time Results (300 items)

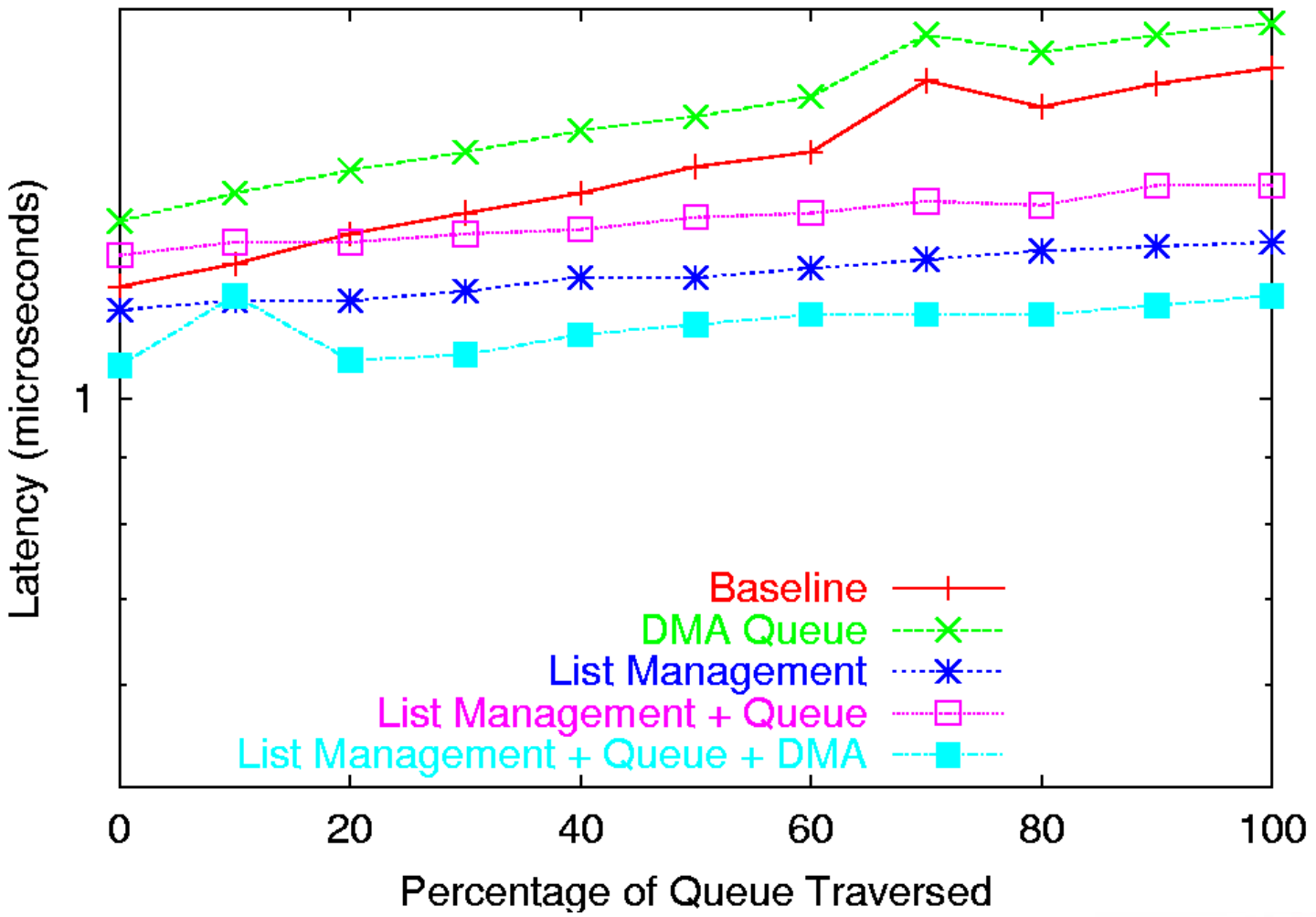


Sandia National Laboratories

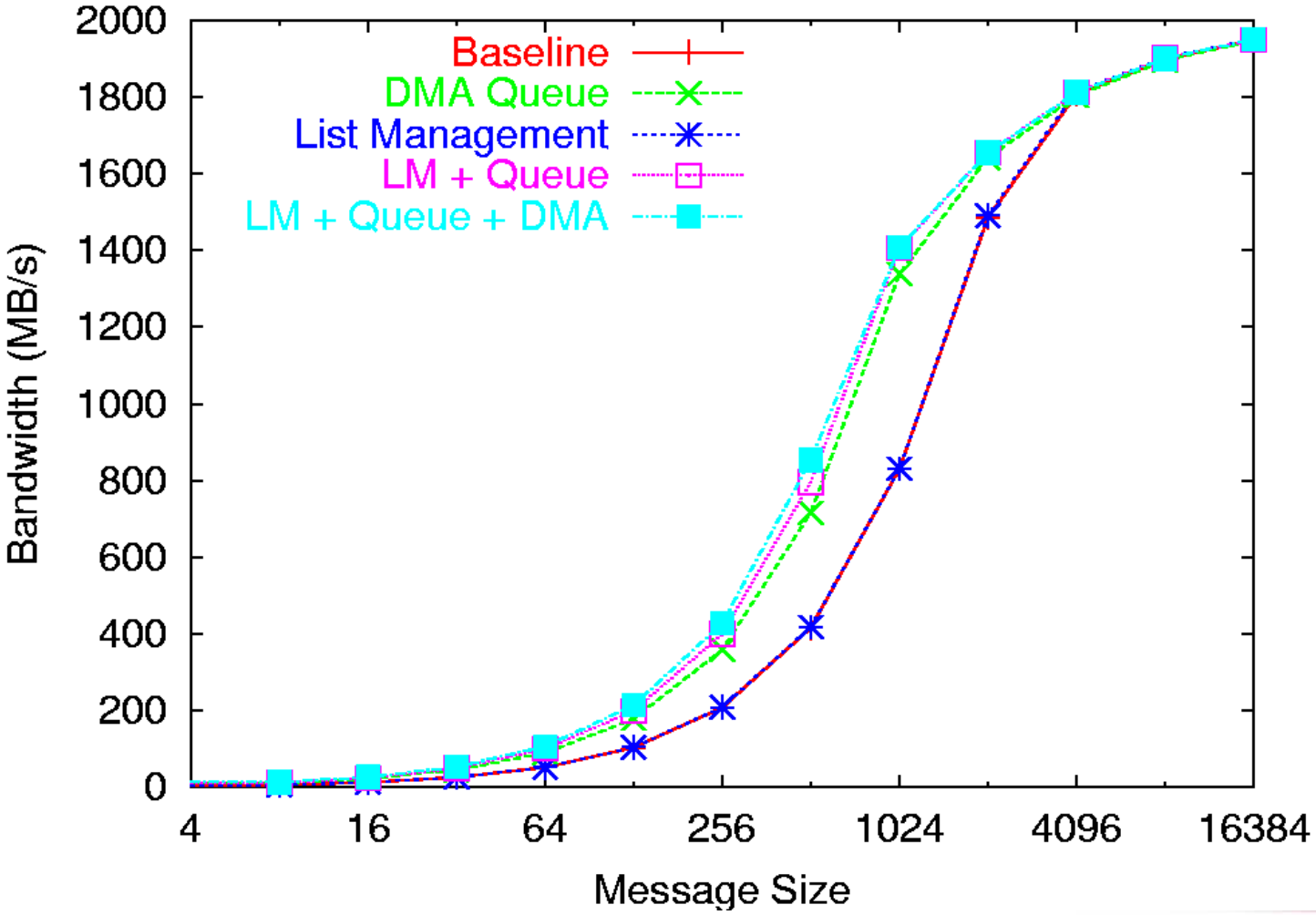
Zero Queue, Short Message



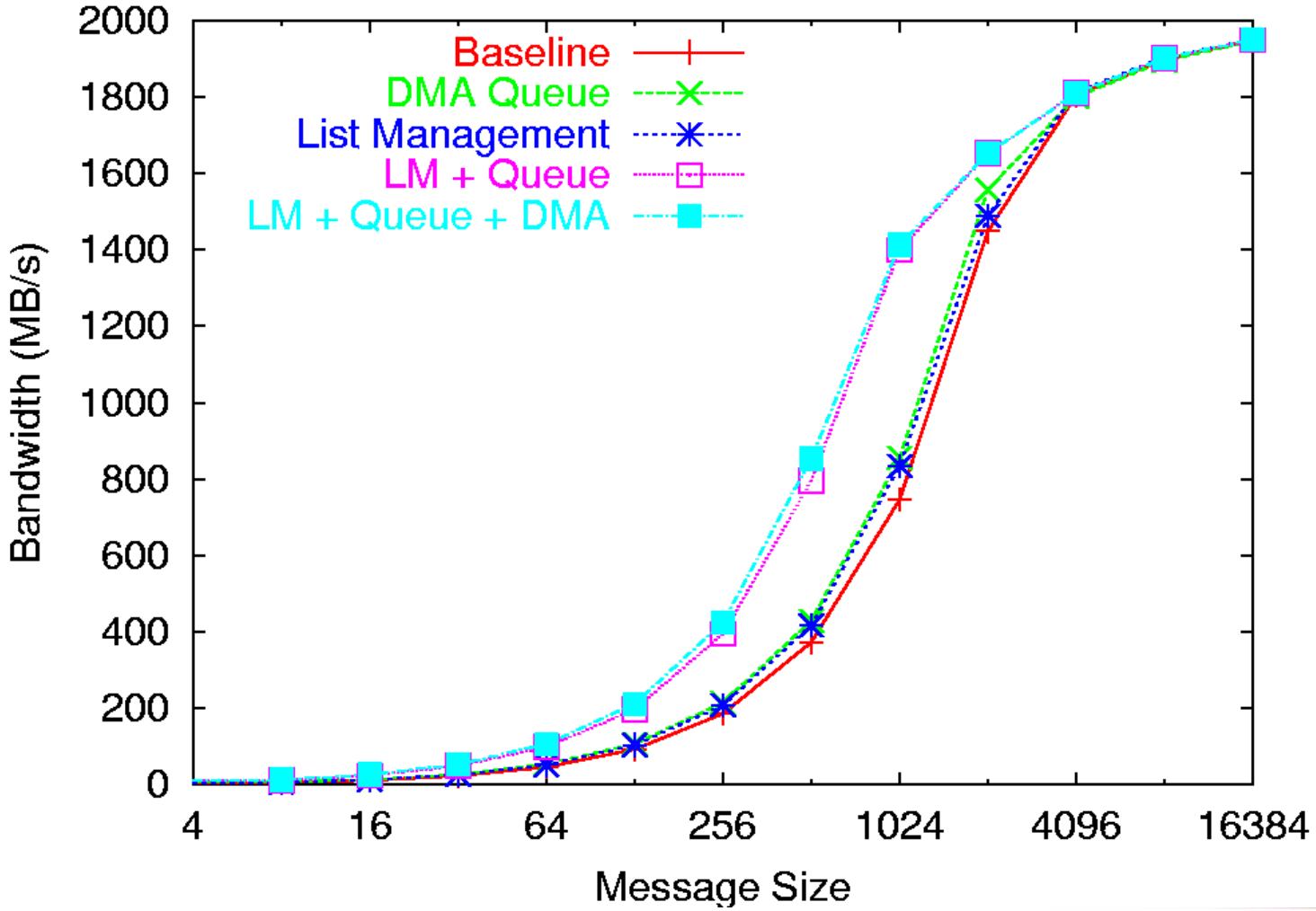
Length 30 Queue, Short Message



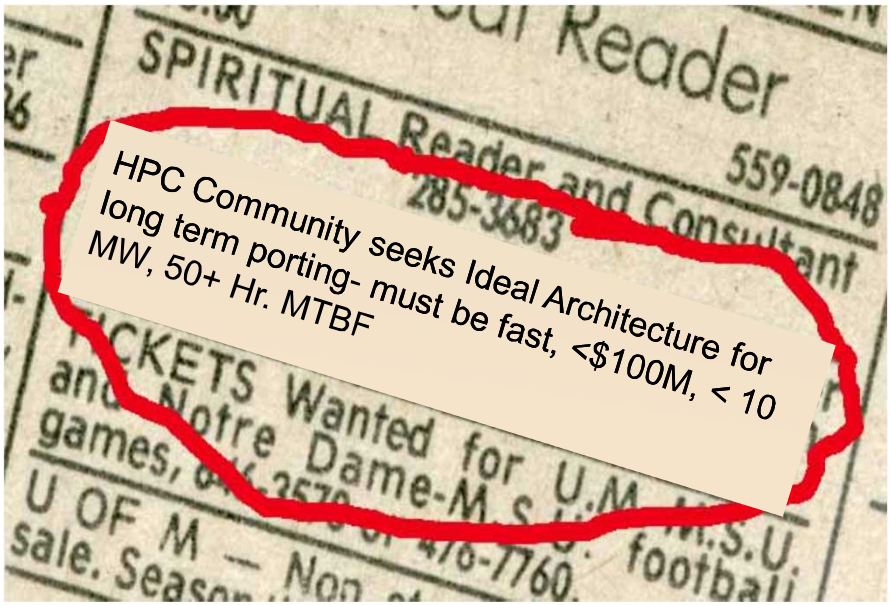
Zero Queue, Streaming



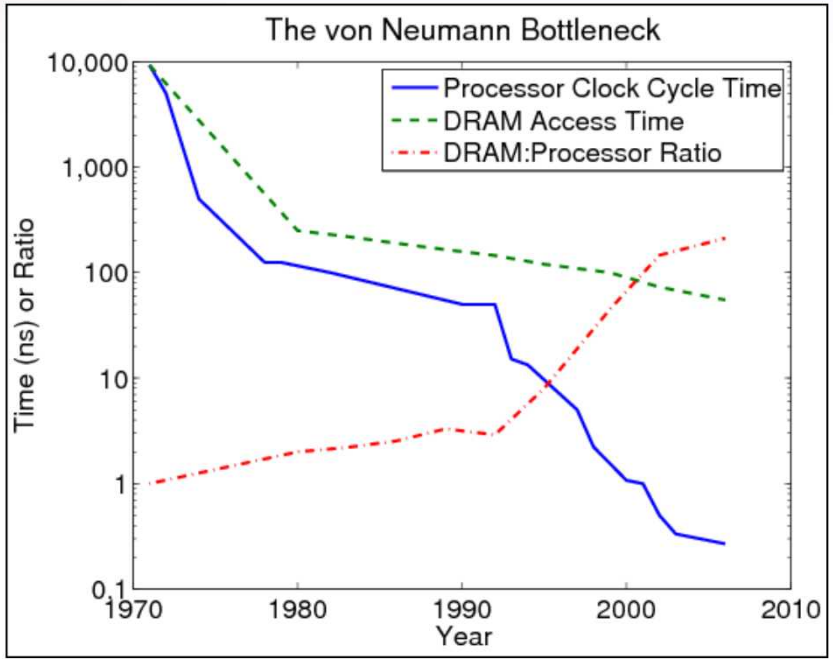
Length 30 Queue, Streaming



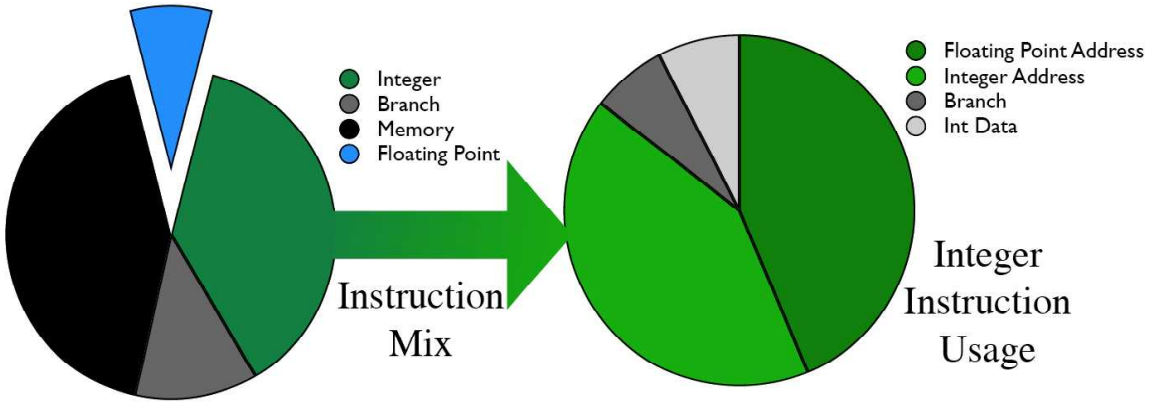
Memory Research


SPIRITUAL Reader
559-0848
285-3683
HPC Community seeks Ideal Architecture for long term porting- must be fast, <$100M, < 10 MW, 50+ Hr. MTBF
TICKETS Wanted for U.M.
Notre Dame-M
games,
U OF M — Non
sale. Season
476-7760.
football

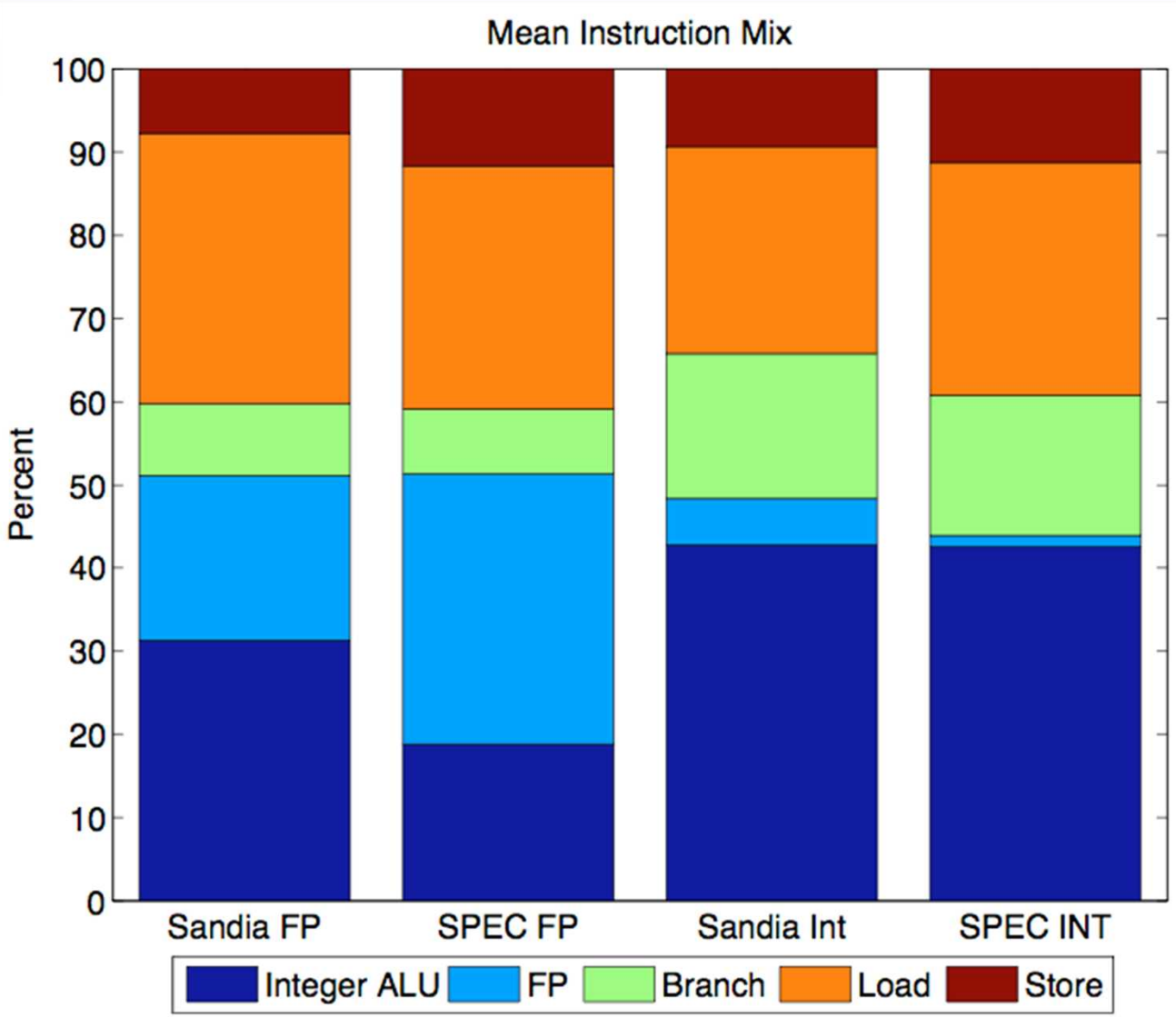
The Problem: Hardware



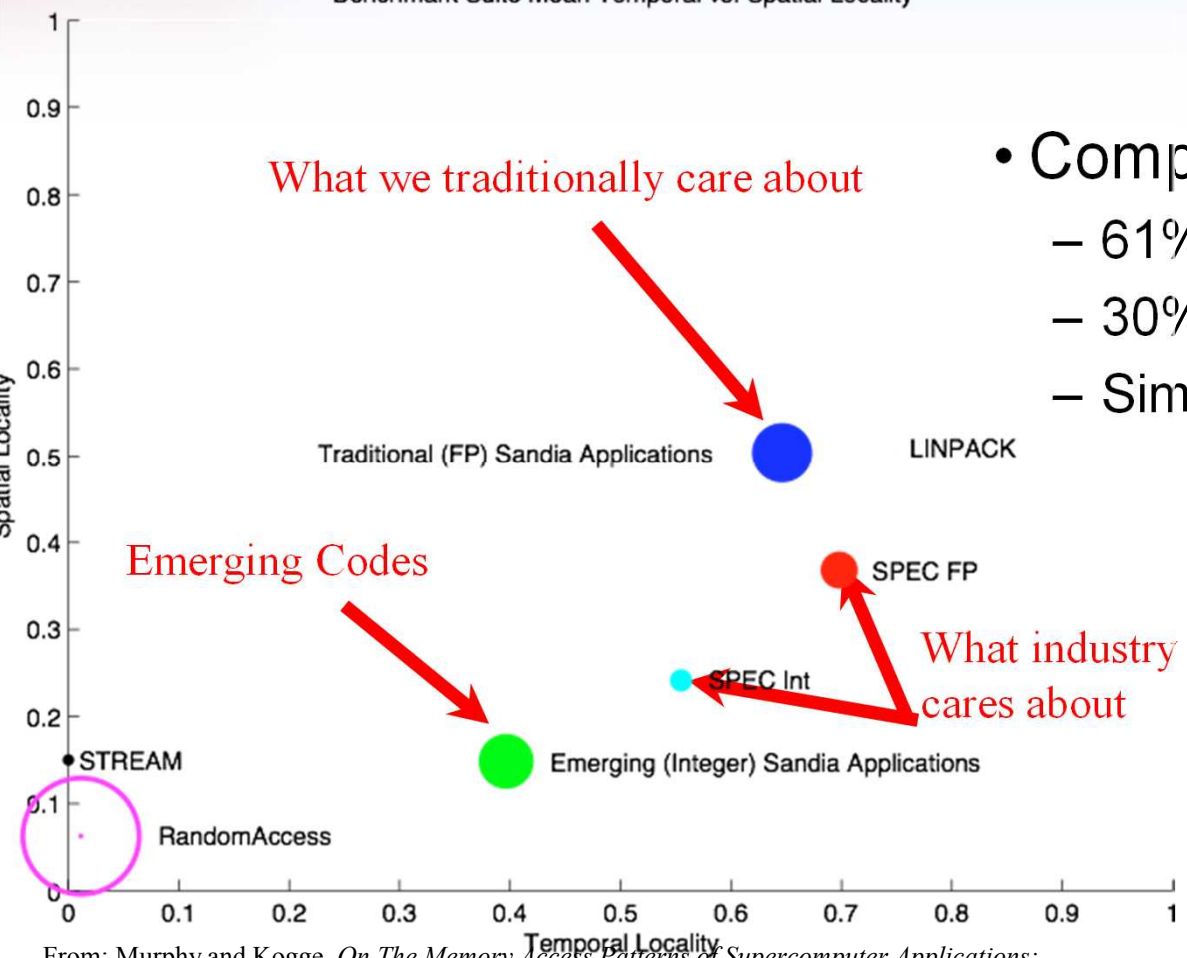
- The gap between processor and memory performance continues to grow
- Future increases in performance will require increased concurrency
- Most *compute* instructions are computing memory addresses, not data value



Real FP Applications Don’t Do Much FP



The Problem: Software



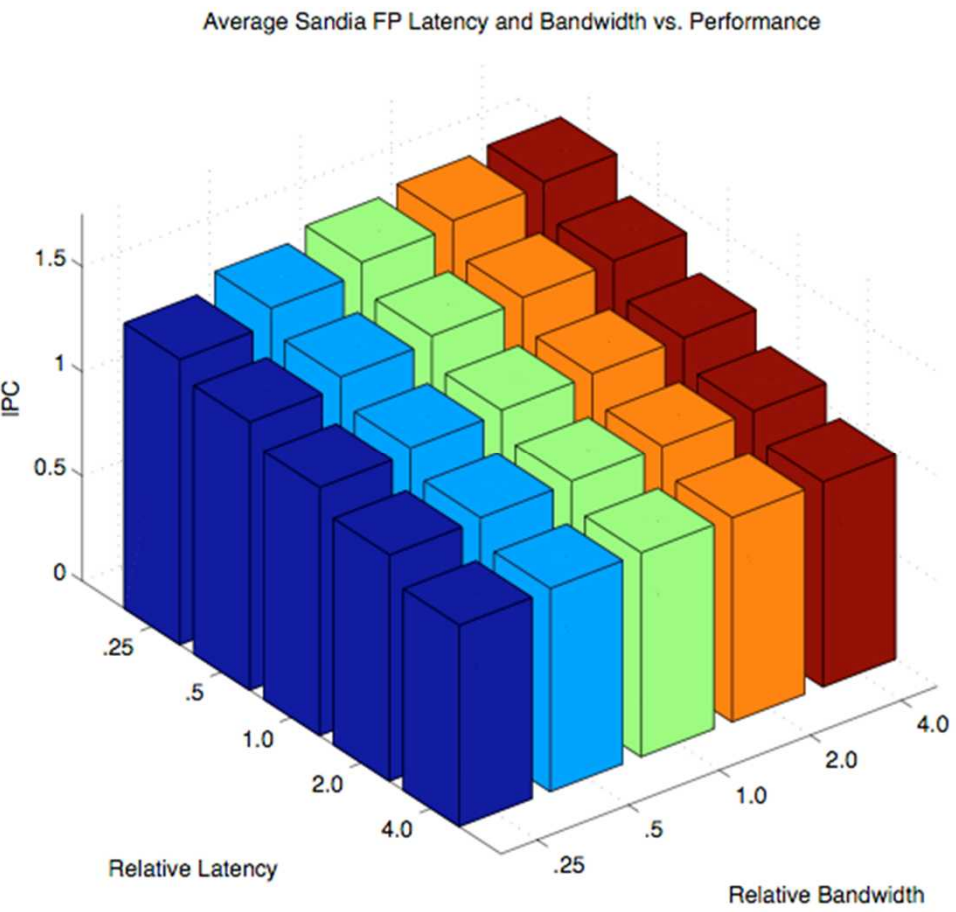
- Compared to traditional apps:
  - 61% the temporal locality
  - 30% the spatial locality
  - Similar data set size

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

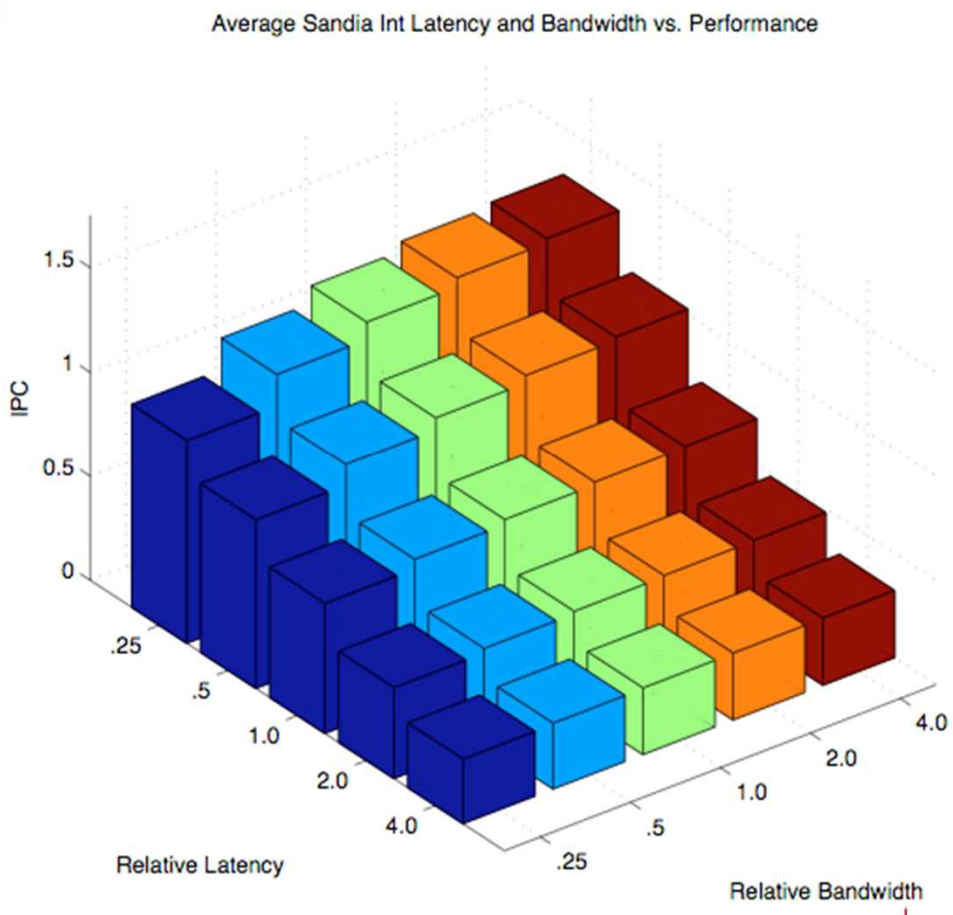
VENDORS DO NOT DESIGN MACHINES TO ADDRESS OUR PROBLEMS!

Latency/Bandwidth Tradeoffs

Floating Point



Integer



Catamount: The Light Weight Kernel (LWK) on Red Storm

July 16, 2008

Sue Kelly

Sandia National Laboratories, Dept 1423

smkelly@sandia.gov, 505-845-9770

Systems and Software History

CM-2
1989

nCUBE-2
1990

iPSC-860
1992

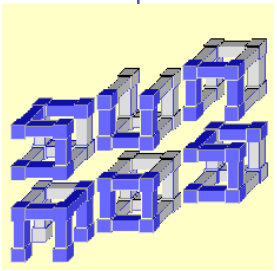
Paragon
1993

ASCI Red
1996

Cplant
1998

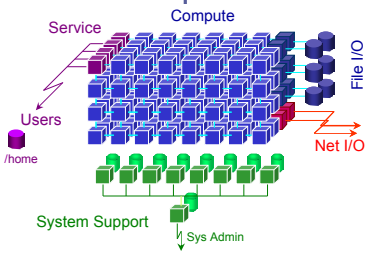
Red Storm
2005

1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007

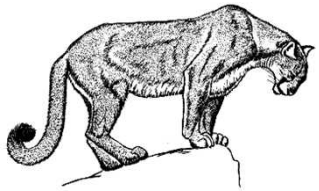


SUNMOS
1991 - 1997

Portals
1992 –



Partition Model
1993 –



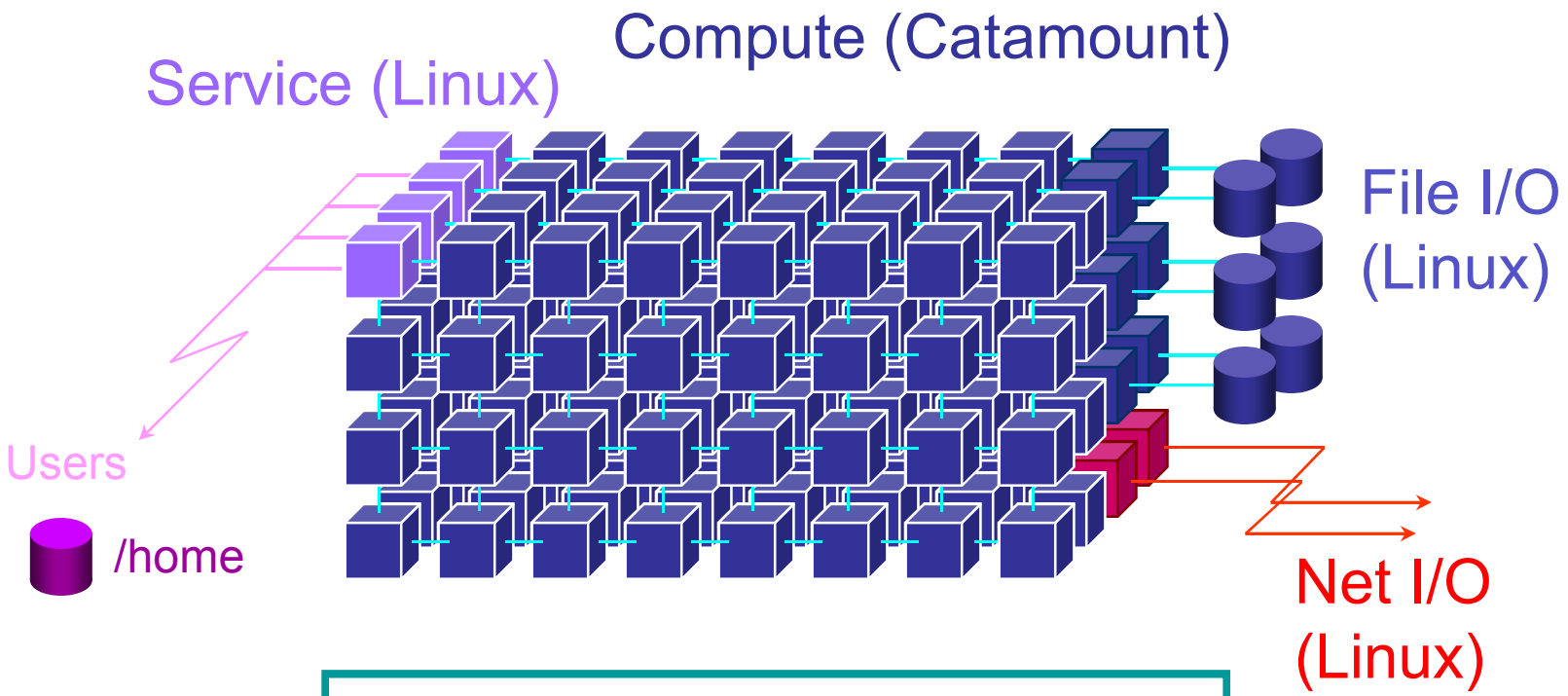
Puma, Cougar,
Catamount
1993 –

Cplant
1997 – 2005

Kitten
2007 –

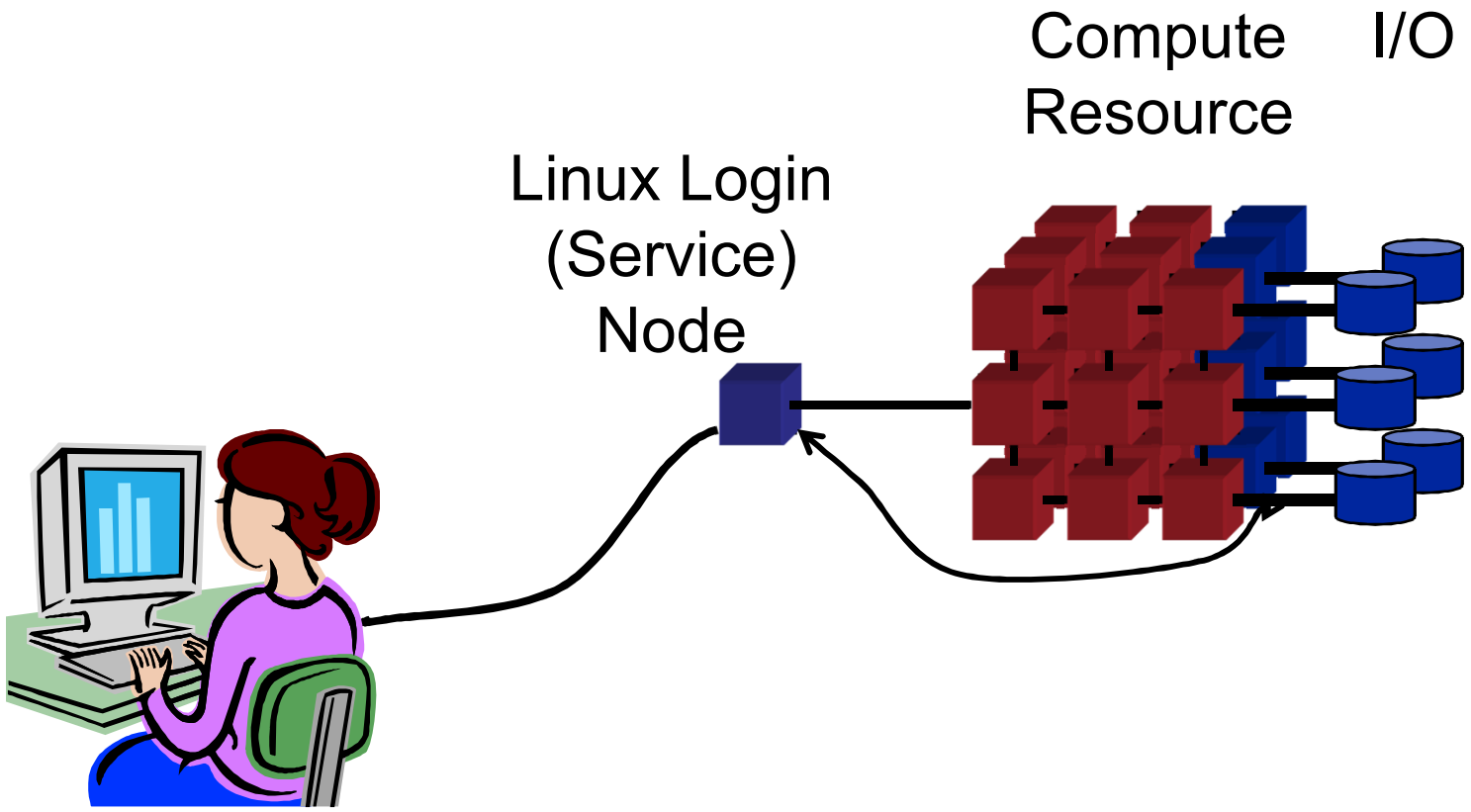
Sandia National Laboratories

Conceptual Partition Model



Partitioning applies to both the hardware and the software

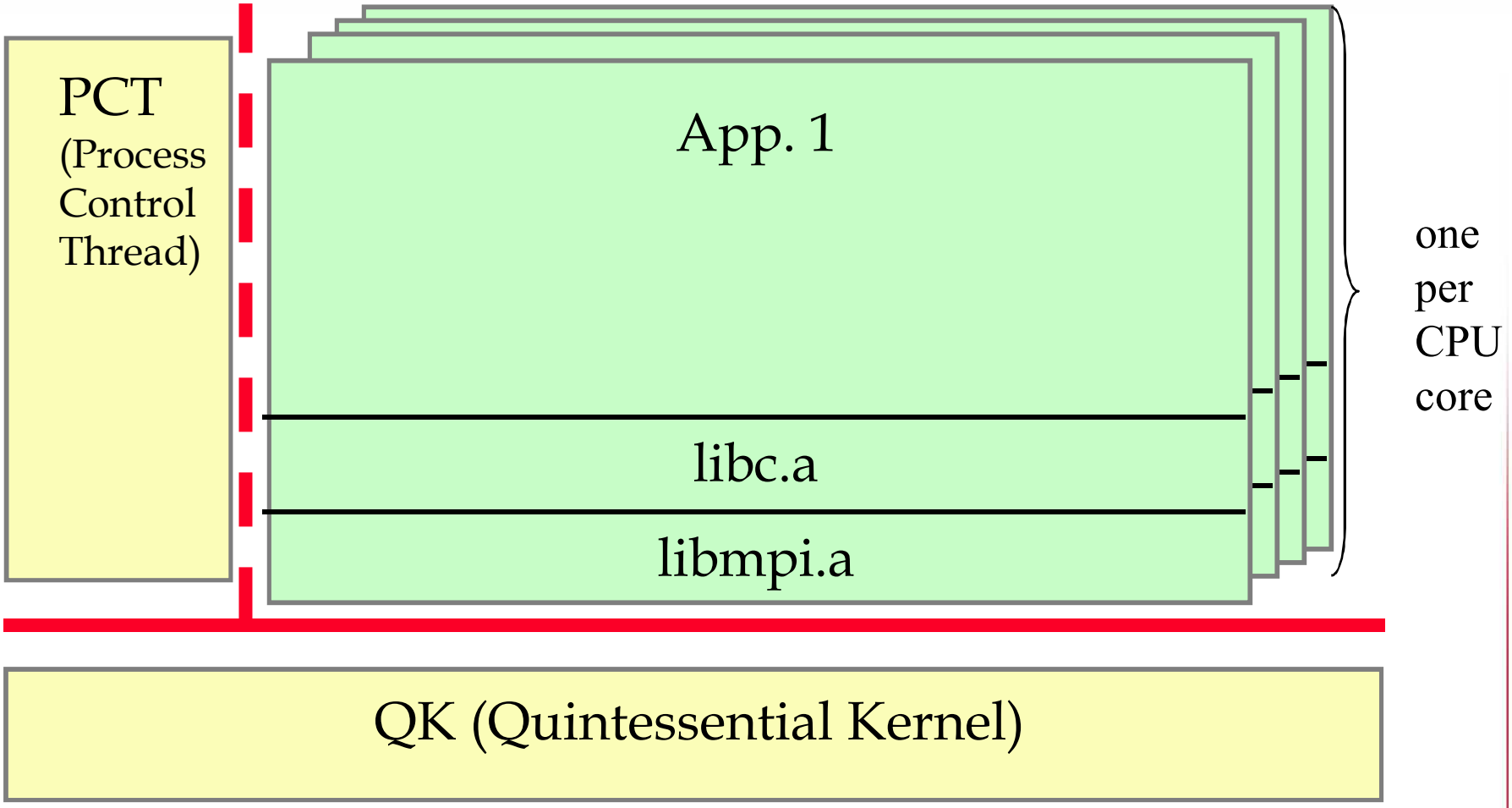
Usage Model



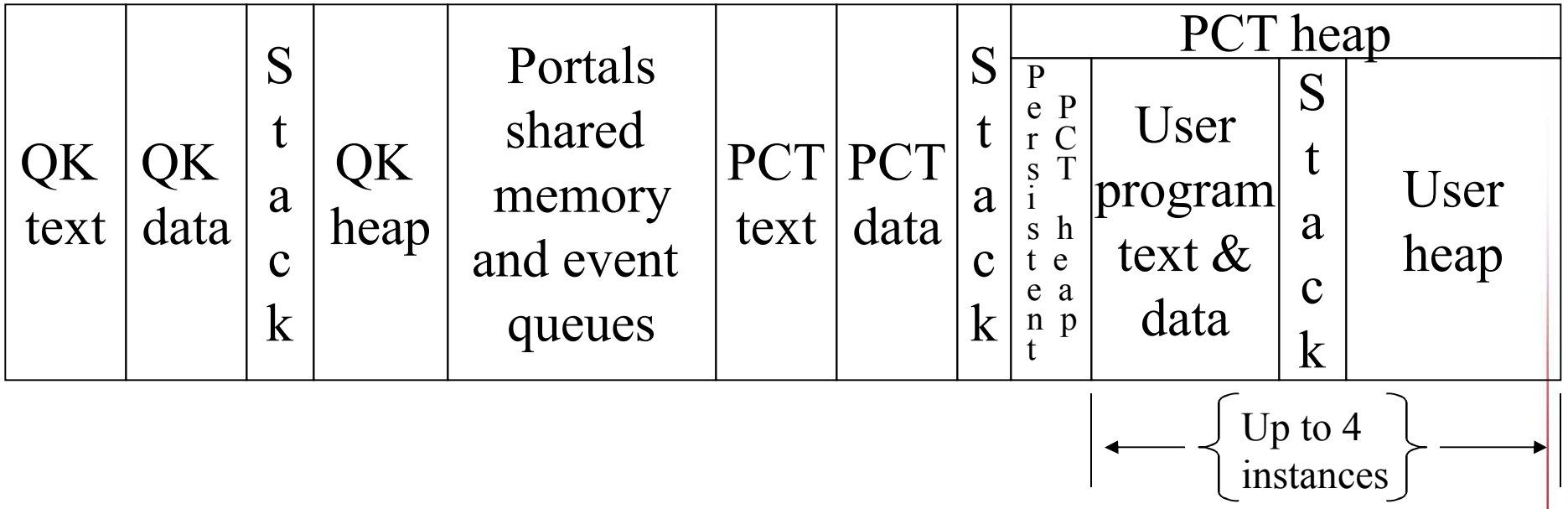
A Lightweight Compute Node Operating System is a Fundamental Part of the Sandia Architecture

- It is essential for
  - Maximizing CPU resources
    - Reduce OS and runtime system overhead
  - Maximizing memory resources
    - Small memory footprint, large page support
  - Maximizing network resource
    - No virtual memory, physically contiguous address mapping
  - Increasing reliability
    - Small code base, reduced complexity
  - Deterministic performance
    - Repeatability
  - Scalability
    - OS resources must be independent of job size
- Others have realized these benefits
  - nCUBE (Vertex), Cray T3 (UNICOS/mk), IBM BG/L (HPK), IBM BG/P (CNK)

Compute Node Software Components



Catamount LWK Physical Memory Layout



Note: not to scale

Portals Was Designed for MPI at Scale

- Connectionless
- Supports MPI matching semantics
  - Allows for offloading MPI matching to NIC
  - Very low CPU overhead for both small and large messages
- Maximizes overlap of computation and communication
- Provides scalable and efficient support for buffering unexpected messages
  - Does not require extra flow control in MPI

Sandia National Laboratories

Summary

- Sandia remains committed to system software partitioning, based on function
- Sandia remains committed to Light Weight Kernel technology and custom network protocols for scientific computation
- New programming models dictate careful consideration of new features, while preserving performance of existing message-based parallel programming

Sandia National Laboratories

Application Performance On Multicores

Douglas Doerfler

CSRI Seminar

Feb. 4th, 2008

SAND2008-1084P
Unlimited Release
Printed February, 2008

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.

Executive Summary

It’s the memory subsystem!
(not quite, but pretty darn close)

Sandia
National
Laboratories

What my talk is NOT about

- Massively parallel performance (I.e. performance off the node)
- Software programming models and Operating System Impact
  - A topic of a future System Engineering Seminar Series (SESS)
- Architecture research

Multicore is here, are we ready?

- Dual-core is mainstream
  - Not a big deal, process level parallelism
- Quad-core is almost mainstream
  - In general, still no strategy for SW
  - It's an SMP?
  - MPI everywhere?
- Eight core is in the near future
- 10's to 100's of cores in the next 3 to 5 years?
- Platform Roadmap
  - Red Storm - dual-core AMD
  - TLCC - quad-core AMD
  - ASC/NMHPC next generation capability system - N-core ?
  - ASC/Sequoia UQ platform in 2010 - ?

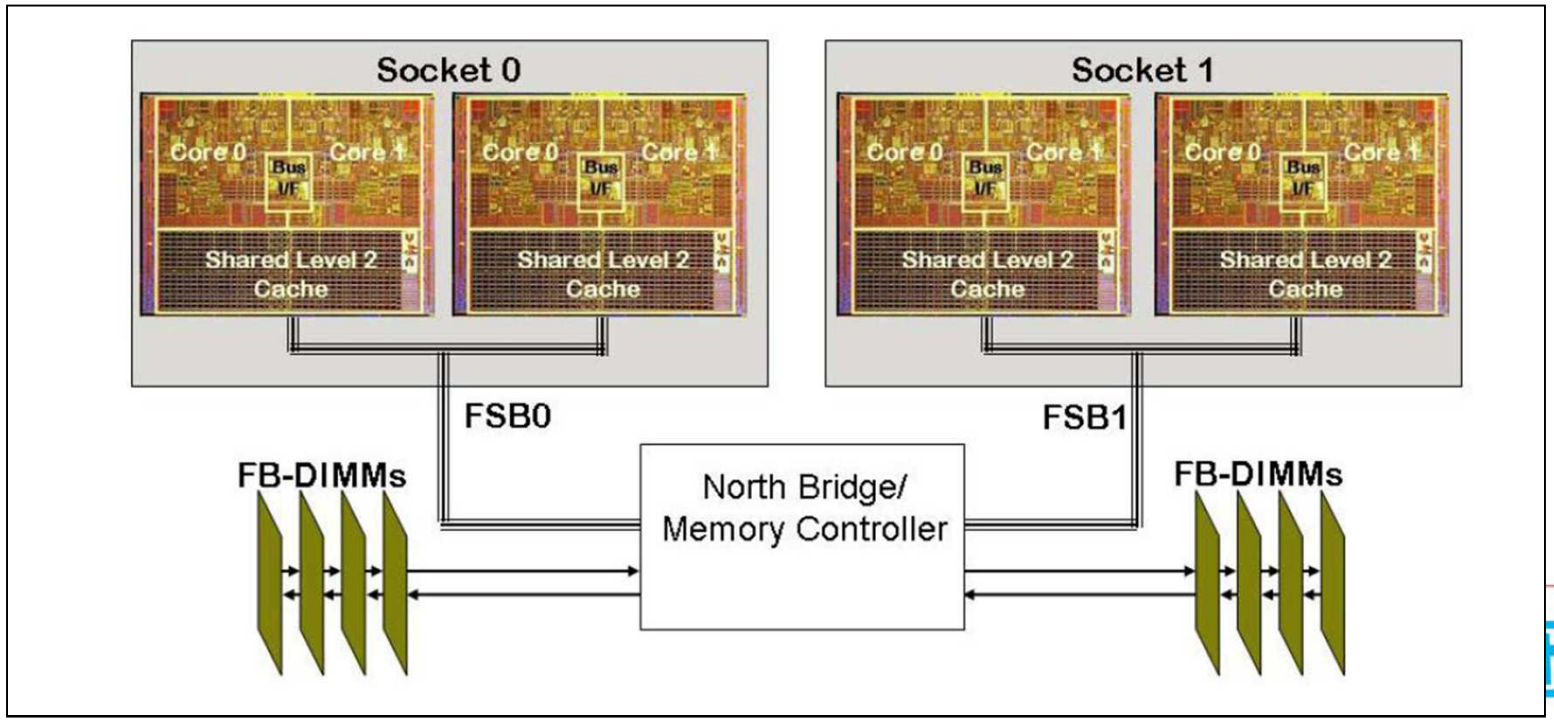
Sandia National Laboratories

What is Multicore?

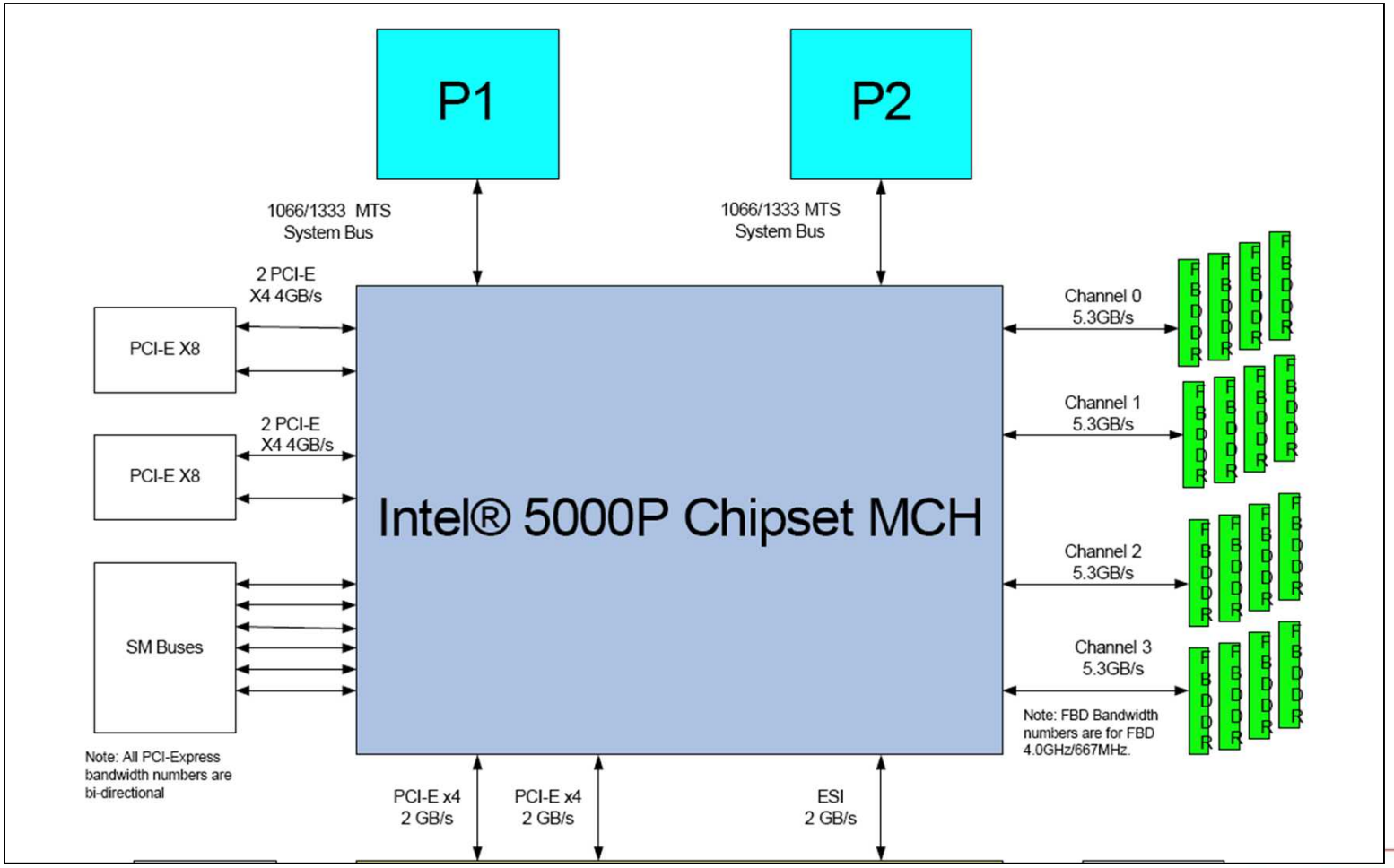
- For this talk, general purpose processors only
- Multicore comes in many flavors
  - How does a processor company differentiate?
  - Distinguishing architecture features
    - Cache hierarchy
    - Bus
    - Memory controller
    - Core architecture, including # of cores

Intel: Clovertown/Harpertown (don't ask me to explain Intel nomenclature!)

- Pseudo quad-core
  - Two dual-core die MCM
- Clovertown
  - 65 nm process
  - Core 2 vs Core 2 Duo?
- Harpertown
  - 45 nm process
  - New microarchitecture Core 2 Duo Extreme?
- 4 FLOPs/cycle/core

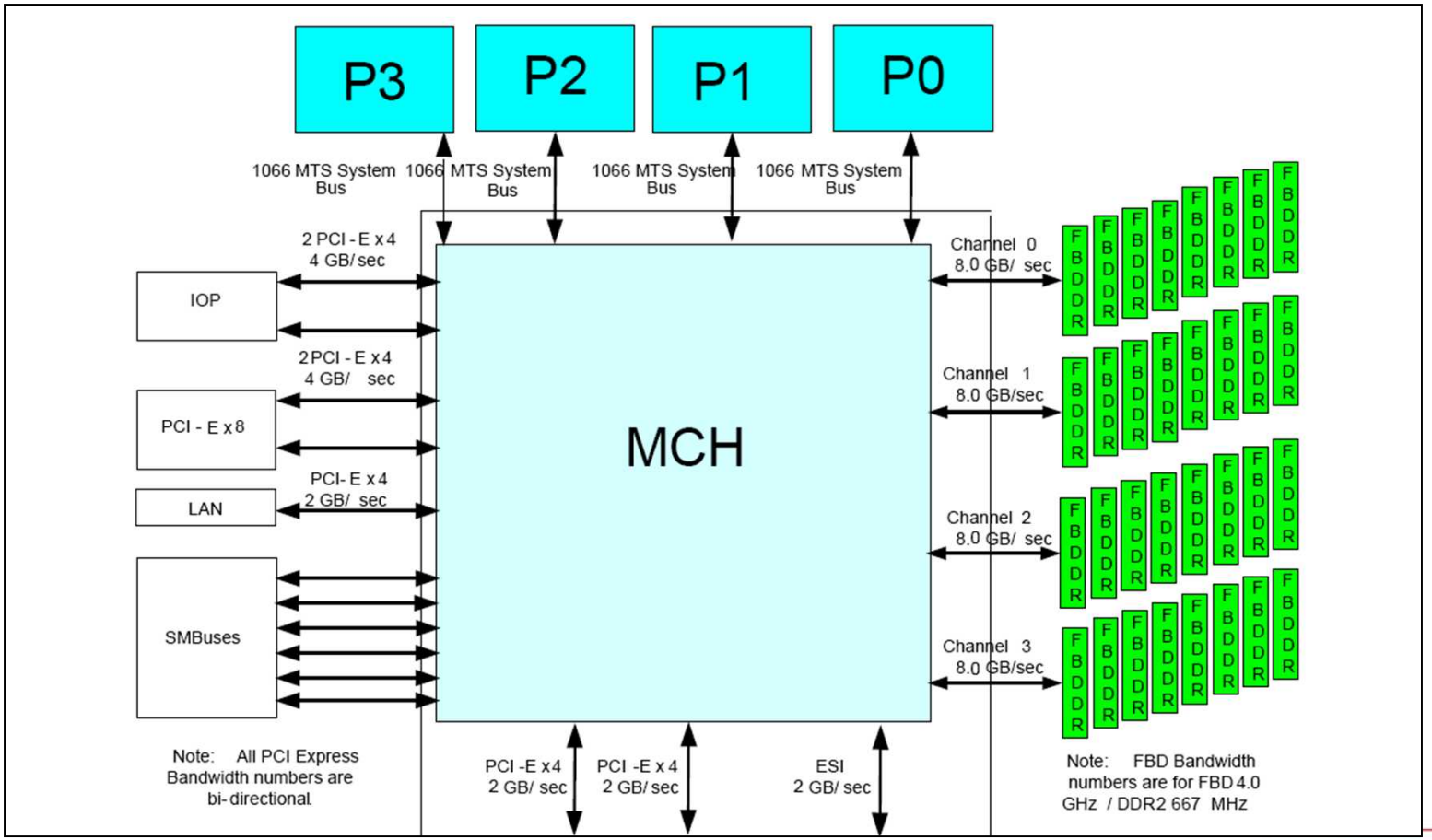


Intel 5000P Northbridge/Memory Controller



Sandia National Laboratories

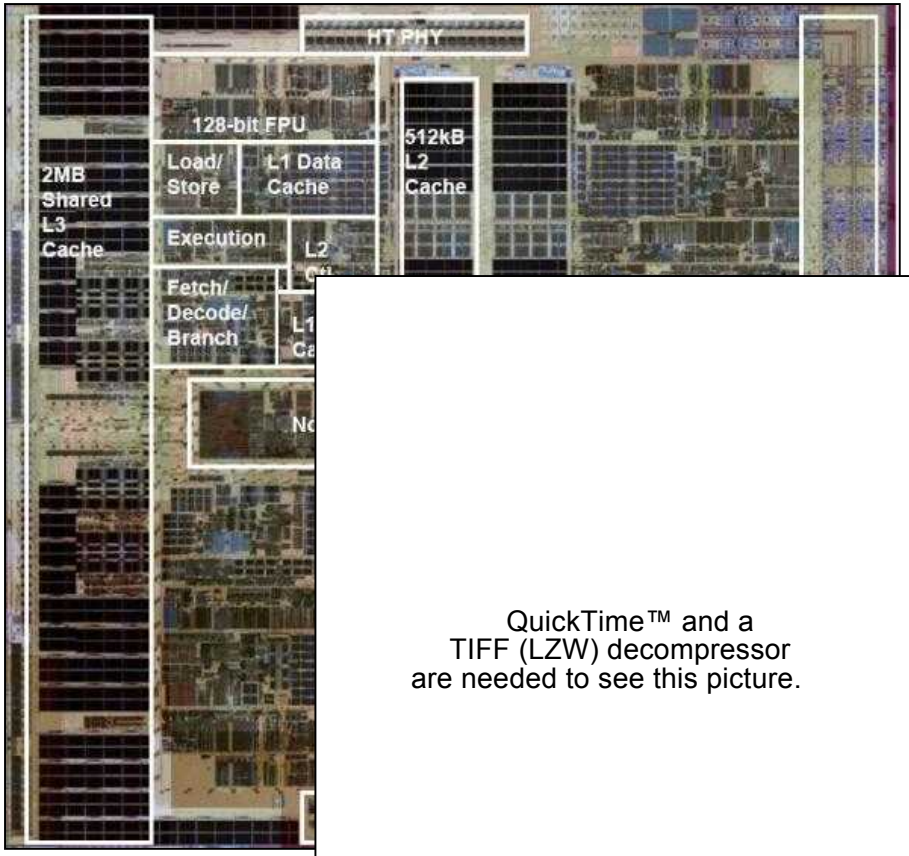
Intel 7300 Northbridge: Four Sockets



Sandia National Laboratories

AMD: Barcelona

- "True" quad-core
- Integrated memory controller
- Uses DDR2/3
  - I.e. no FBDIMM
- Hypertransport for multi-socket nodes
  - NUMA issues
- Dual-channel DDR memory controller
- 2 MB shared LLC
- 4 FLOPs/cycle/core



QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

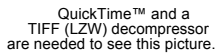
Sun UltraSPARC Niagra/T2

- “True” eight core die
- Each core supports 8 threads - 64 total threads
- Simple core microarchitecture
- Four dual-channel FBDIMM memory controllers!
- 4 MB shared LLC (L2)
  - 1 MB/MCU
- 1 FLOP/cycle/core!

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

IBM Power6

- Dual-core!
- High clock rate!
  - Target 4+ GHz
- Dual-channel DDR2/3
- 32 MB LLC (L3)
  - But external to die
- 4 FLOPs/cycle/core

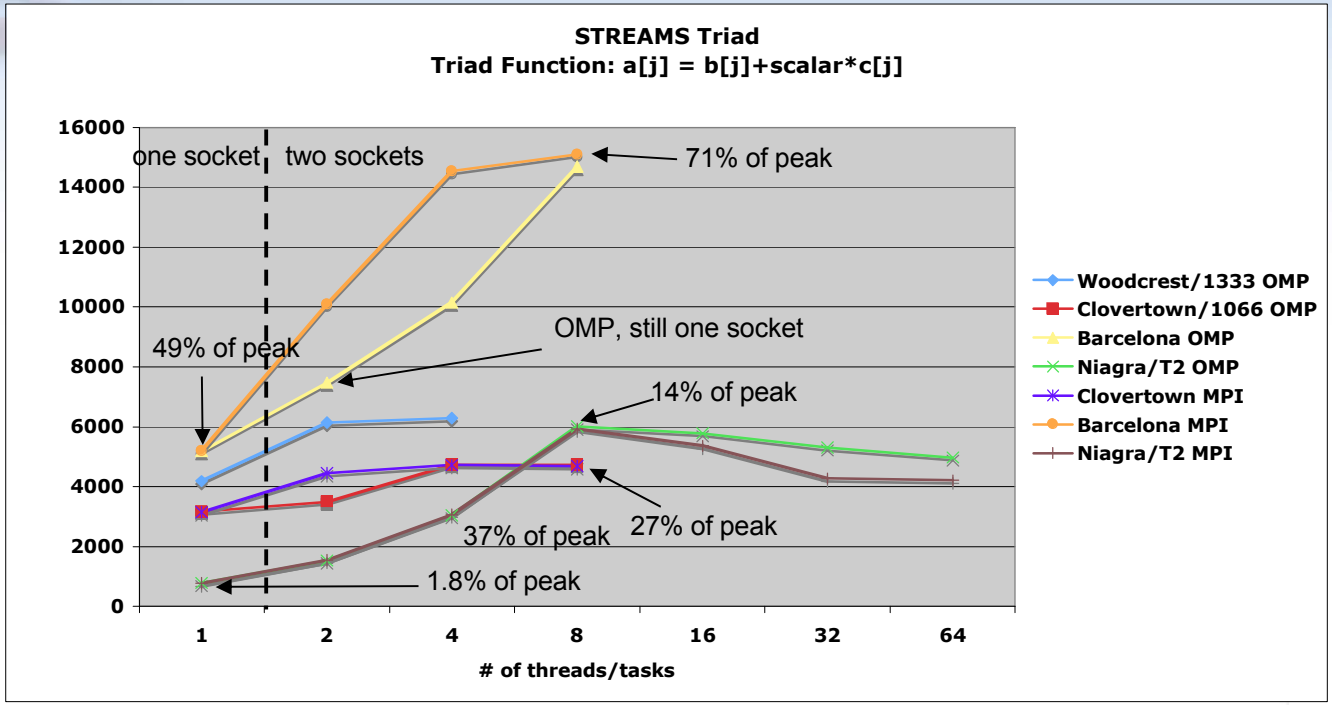


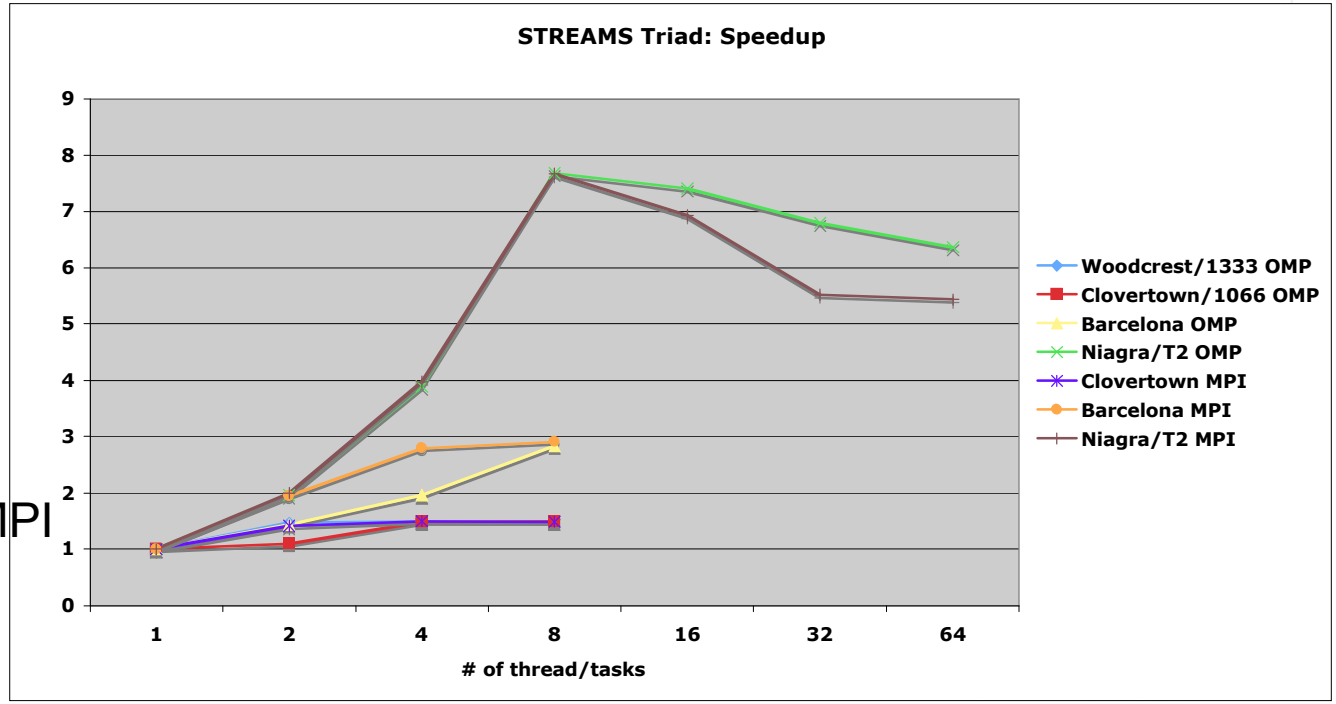
Test Systems by the Numbers

	Clovertown	Barcelona	Niagra/T2
Core frequency	1.86	2.2	1.4
# cores/socket	4	4	8
Execution arch.	out of order	out of order	in order
Total # cores	8	8	8
# sockets	2	2	1
Total Threads	8	8	64
L1 Dcache	32KB/core	64KB/core	8KB/core
L2 cache	2 x 4MB	512KB/core	4 x 1MB
L3 cache	-	2MB	-
DRAM	667MHz FB	667MHz DDR	667MHz FB
# mem. channels	4	2 x 2	8
Peak read BW	21.3	2 x 10.7	42.7
Peak write BW	10.7	same bus	21.3
FLOPs/cycle	4	4	1
Peak FLOPs	59.5	70.4	11.2

STREAMS

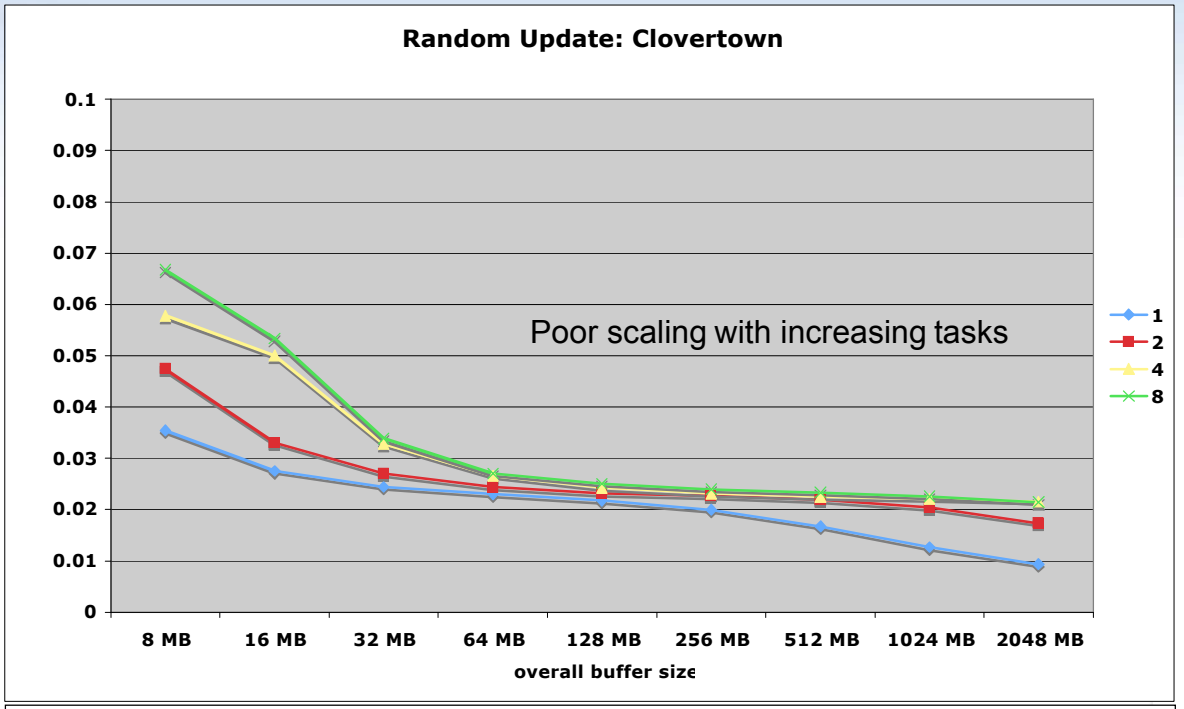
- MPI and OMP
- Two socket nodes
  - 1: 1 core
  - 2: 1 core/socket
  - 4: 2 cores/socket
  - 8: 4 cores/socket
- Clovertown and Barcelona have same peak BW
- Clovertown effectively saturates at 1 core/socket
- Barcelona effectively saturates at 2 cores/socket
- Niagra/T2 becomes FLOPs bound
  - One socket node
- Note Task scheduling differences between MPI and OMP

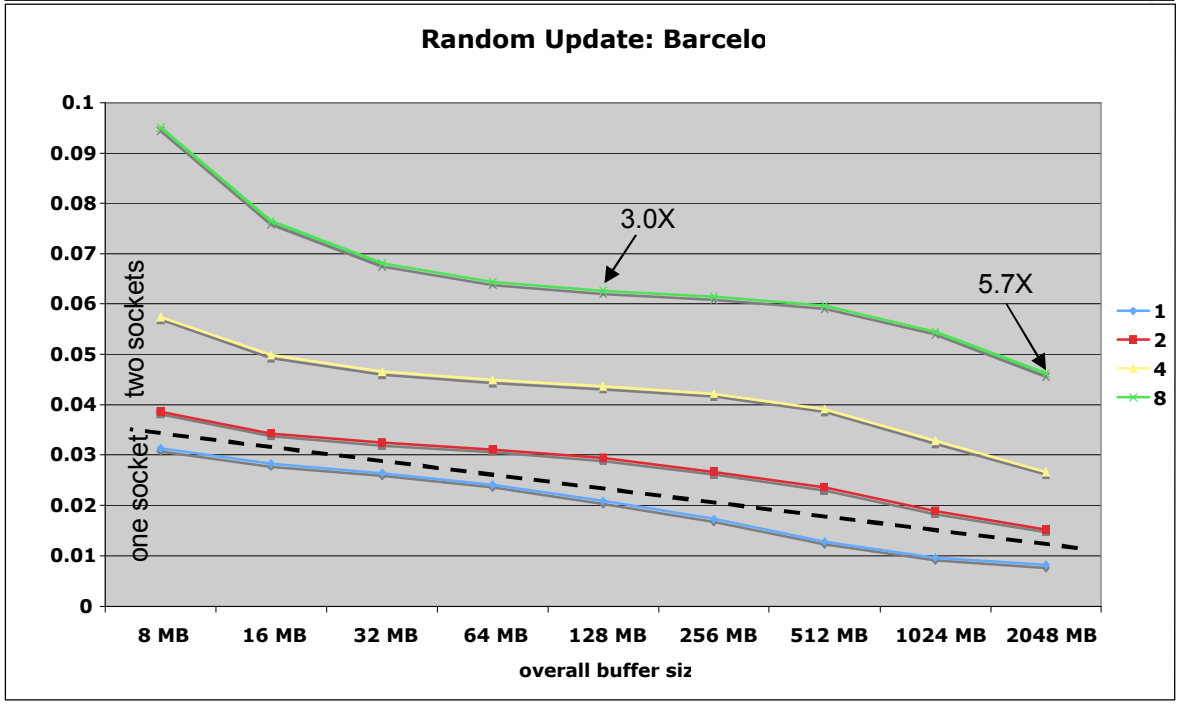




Random Update

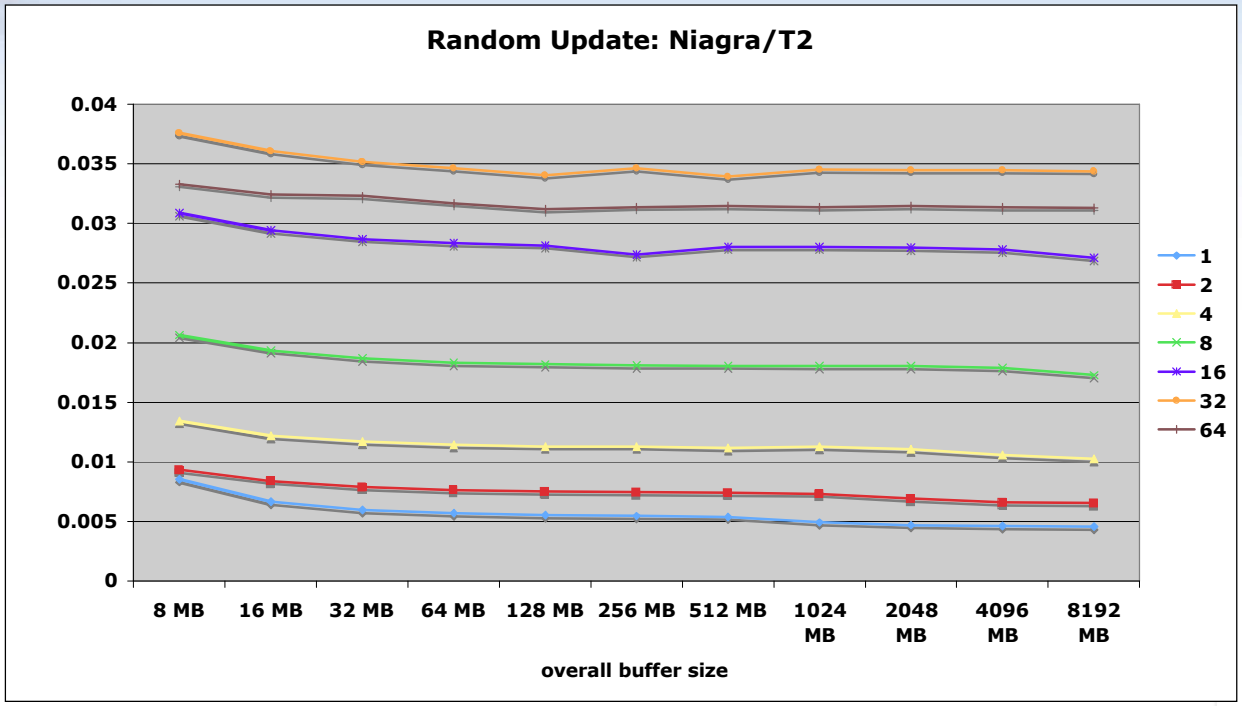
- Two socket nodes
  - 1: 1 core
  - 2: 1 core/socket
  - 4: 2 cores/socket
  - 8: 4 cores/socket
- Varying buffer size
  - Cache effects
  - TLB effects
- Higher is better
- Clovertown: extra cores don't add to performance
- Barcelona: memory subsystem effective at handling extra requests

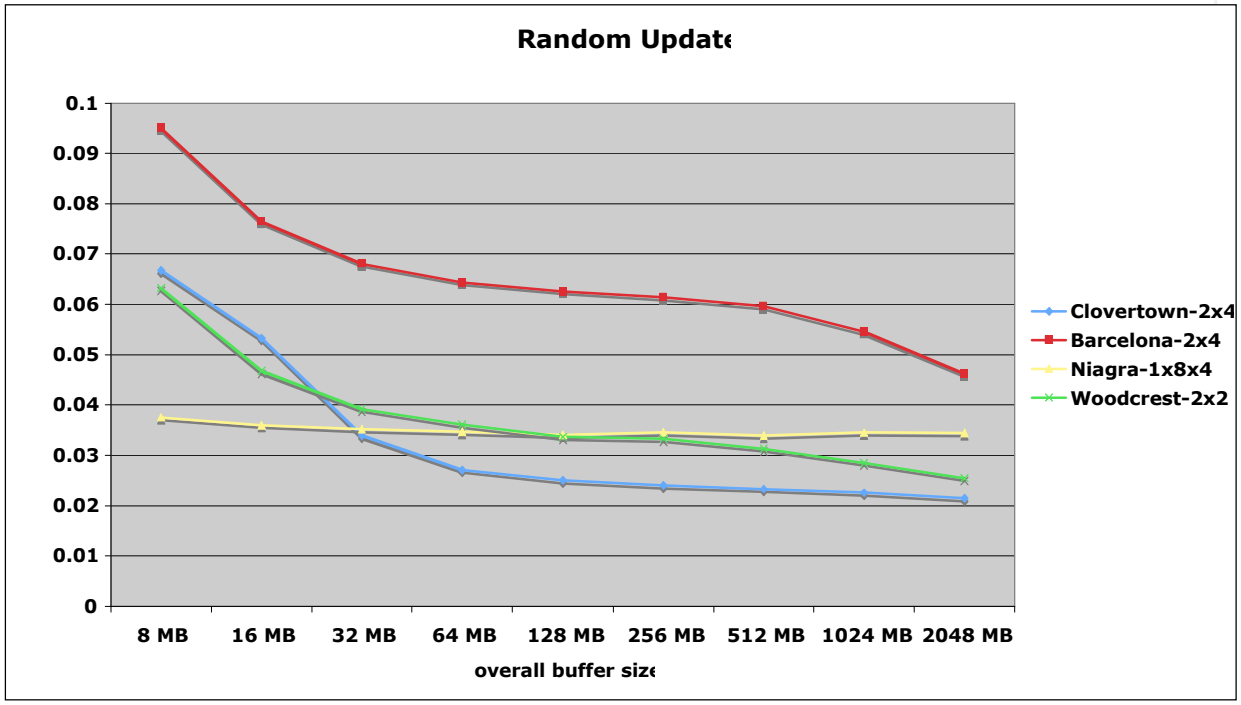




Random Update Cont'd

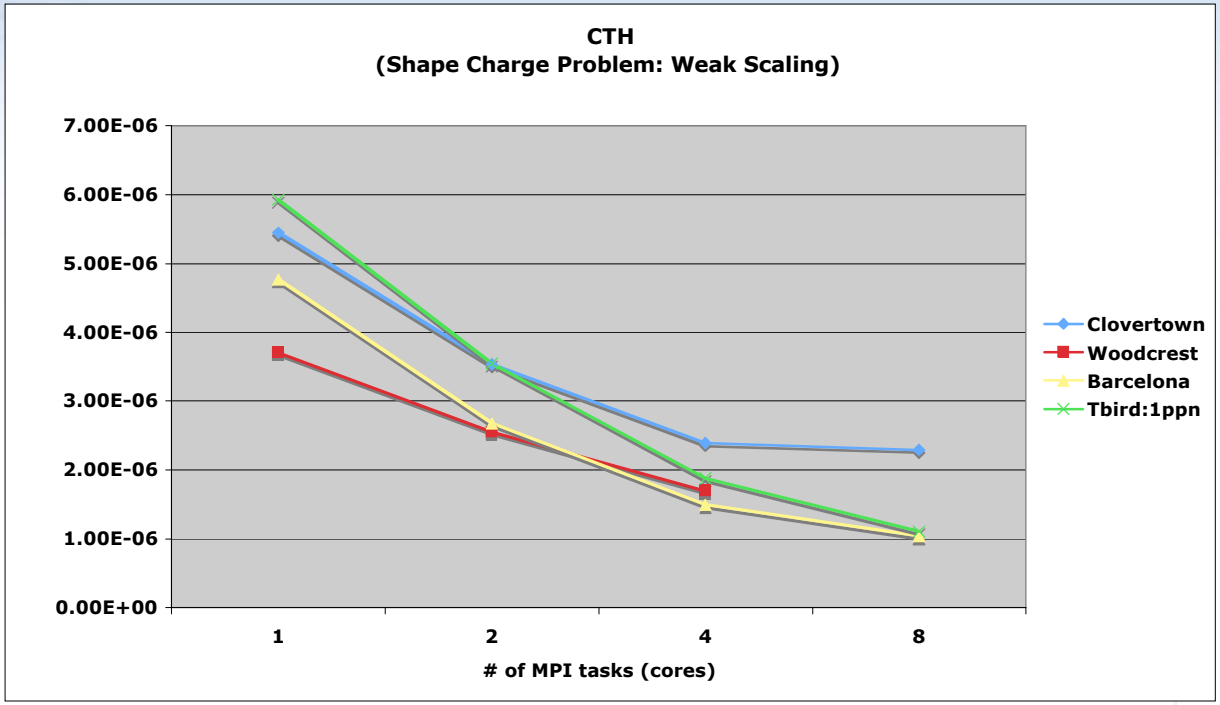
- Niagra/T2 performance
  - Excellent scaling to 16 threads
  - Excellent scaling with buffer size
  - 64 threads < 32 threads
- Best case performance across architectures
  - Barcelona best up to 2048 MB buffer, after that?
  - Intel architecture exhibits lowest performance

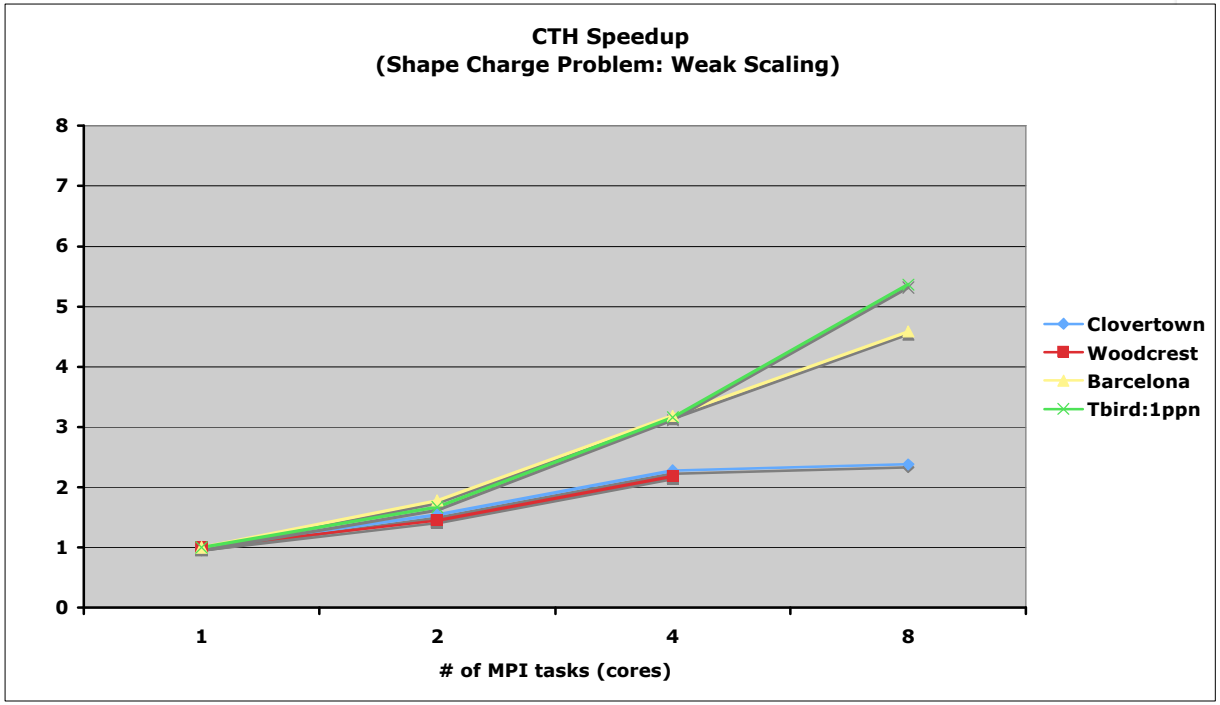




CTH

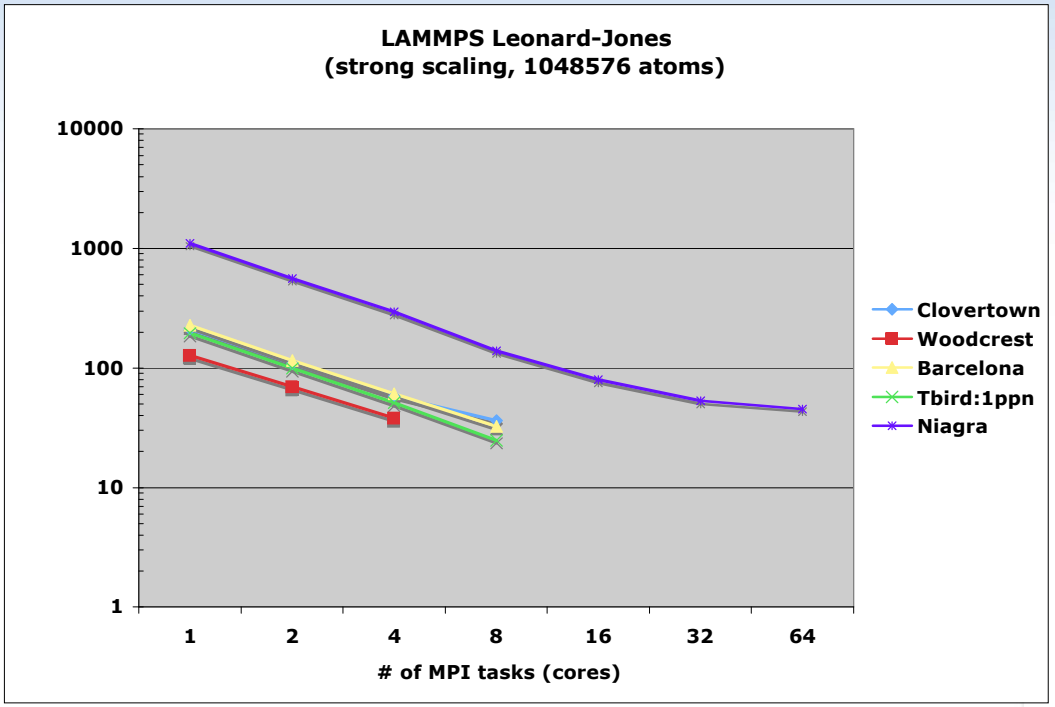
- Shape Charge Problem
- Weak Scaling
  - This is how we plan to use multicore, I.e. spec'ing a minimum GB/core
  - Intracore MPI
- Tbird 1ppn is used as a measure of "ideal" scaling
- Lower is better
- At 8 cores, Barcelona is as fast as 8 Pentium's with Infiniband!
- Clovertown/Woodcrest northbridge issues are evident

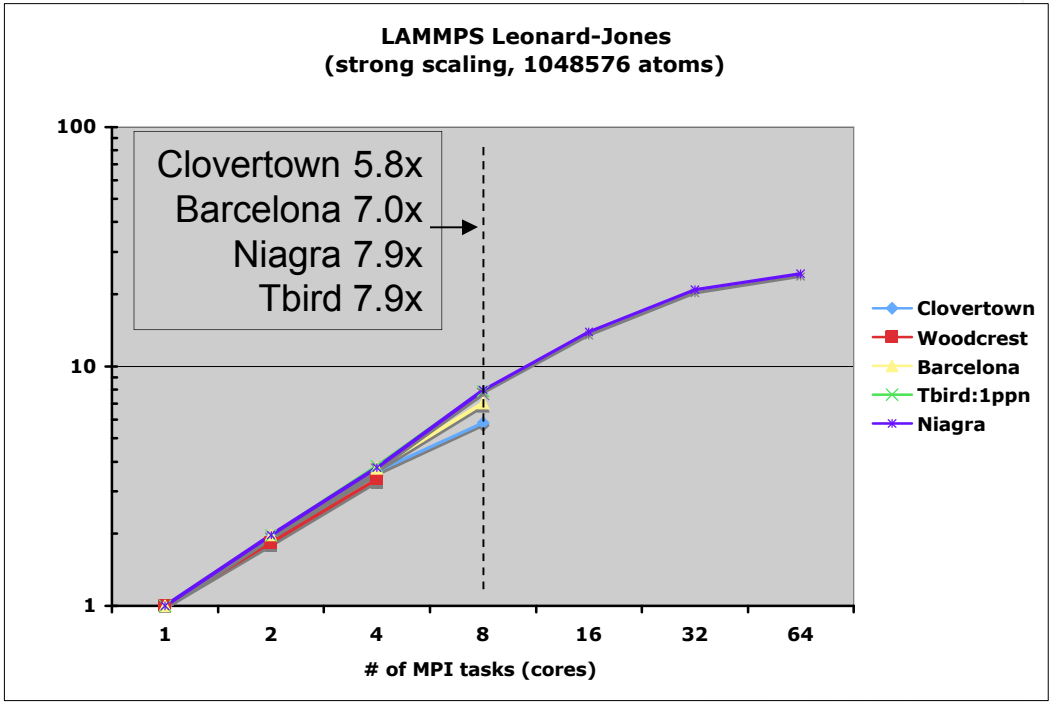




LAMMPS LJ

- Strong Scaling
  - Weak scaling produces similar results
- Lower is better
- All architectures scale very well up to the number of cores/socket
- Niagra never achieves performance of x86-64 architectures, despite excellent scaling





?s & Discussion

Sandia
National
Laboratories