COMPUTER, COMPUTATIONAL & STATISTICAL SCIENCES

Office of Science
U.S. DEPARTMENT OF ENERGY

ASC

LA-UR-08-2847

Moore, More Cores, and More Application Performance

Darren J. Kerbyson

with

Kevin J. Barker, Kei Davis, Adolfy Hoisie, Michael Lang, Scott Pakin, Jose Carlos Sancho

Performance and Architecture Laboratory (PAL)

http://www.c3.lanl.gov/pal

Computer, Computational & Statistical Sciences Division

Los Alamos National Laboratory

NNSA

Los Alamos NATIONAL LABORATORY

Cores

Complexity

Constraints

C

"The future will be like the present only more so"

Groucho Marx

Giga-flops 10^9

Tera-flops 10^{12}

Peta-flops 10^{15}

Exa-flops 10^{18}

Don't matter a-flops 10^{21}

Matt Reilly, SiCortex

Outline of talk

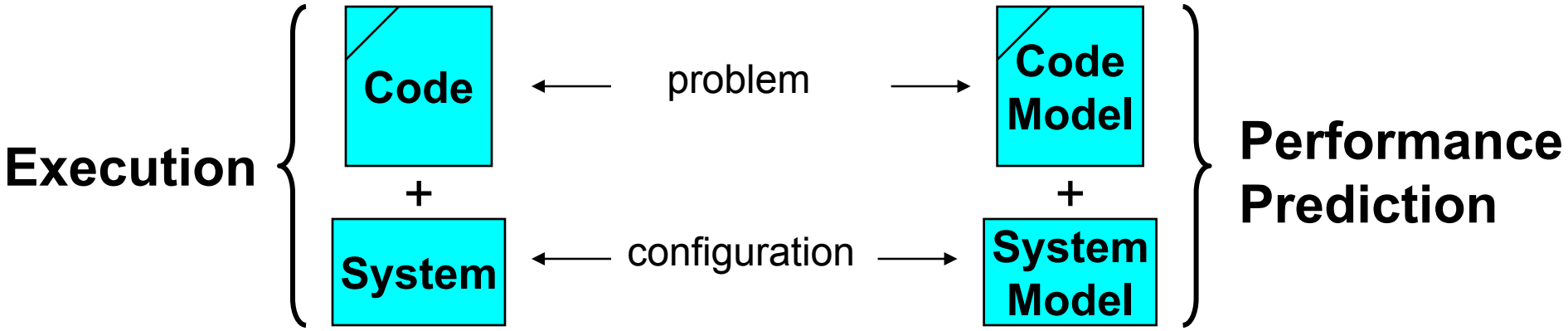
- **Performance at the *µSystem* scale**
  - Quad-core node level performance
- **Performance at the *mSystem* scale**
  - Some different networks
- **Performance at the *System* scale**
  - Dual-core to Quad-core upgrade
  - Accelerated system

Examples drawn from

Roadrunner, PERCS, AMD, Intel, SiCortex, Cray,

Many applications

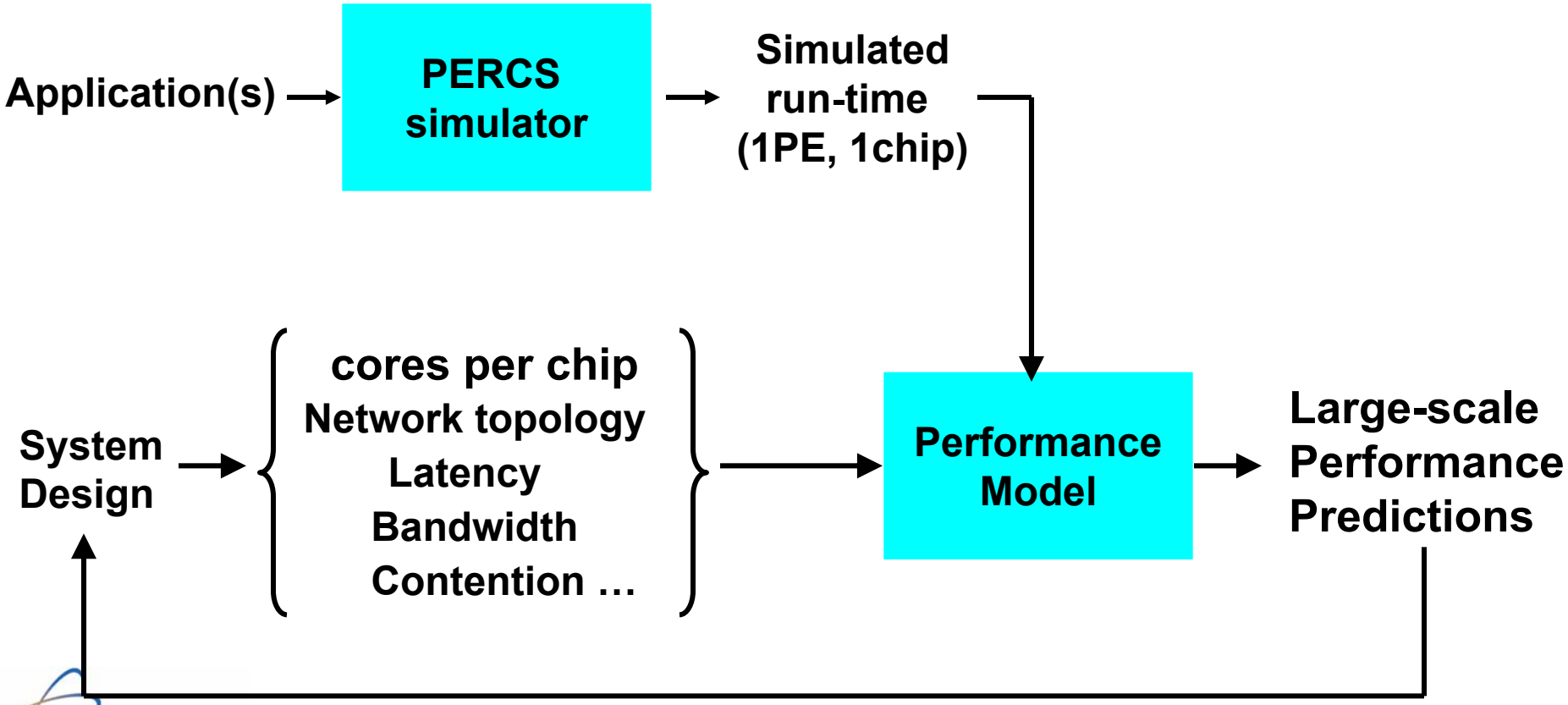
Question: How can I analyze the performance of a non-existent Machine?



- **Answer: Need a model.**
- **A model should encapsulate the understanding of:**
  - What resources an application uses during execution
  - How often it does it
  - How its usage changes when scaling
  - How long the system takes in order to satisfy the resource requirements
- **Application centric view – what the application doing**

Design Space Exploration: Performance Modeling for IBM PERCS (HPCS and BlueWaters)

- Input to model: single-PE / single-chip performance from Mambo

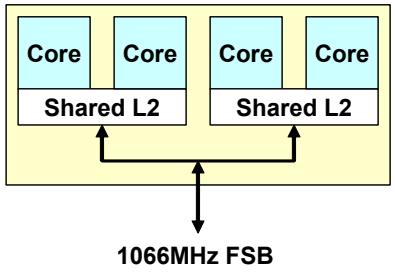


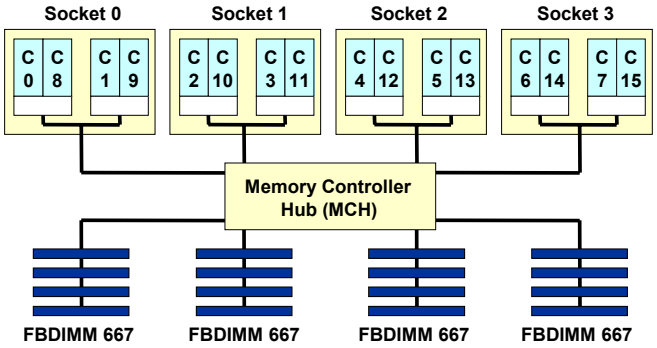
Performance at the *µSystem* scale: Quad-cores

- **Two quad-core architectures:**
  - Intel Tigerton, 4-socket, 2 dies per socket, 2 cores per die
  - AMD Barcelona, 4-socket, 1 die per socket, 4 cores per die

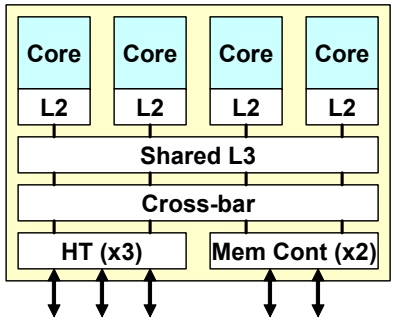
Socket | **Node (4-sockets)**

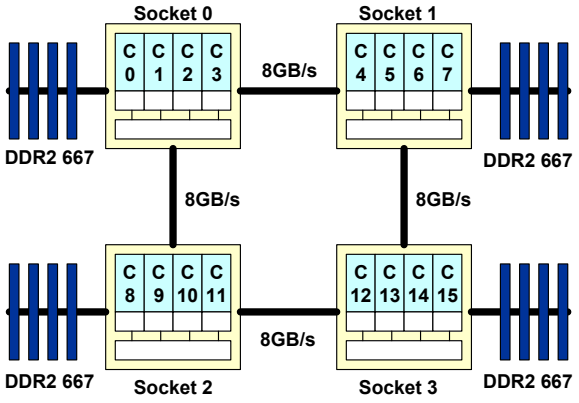
Intel



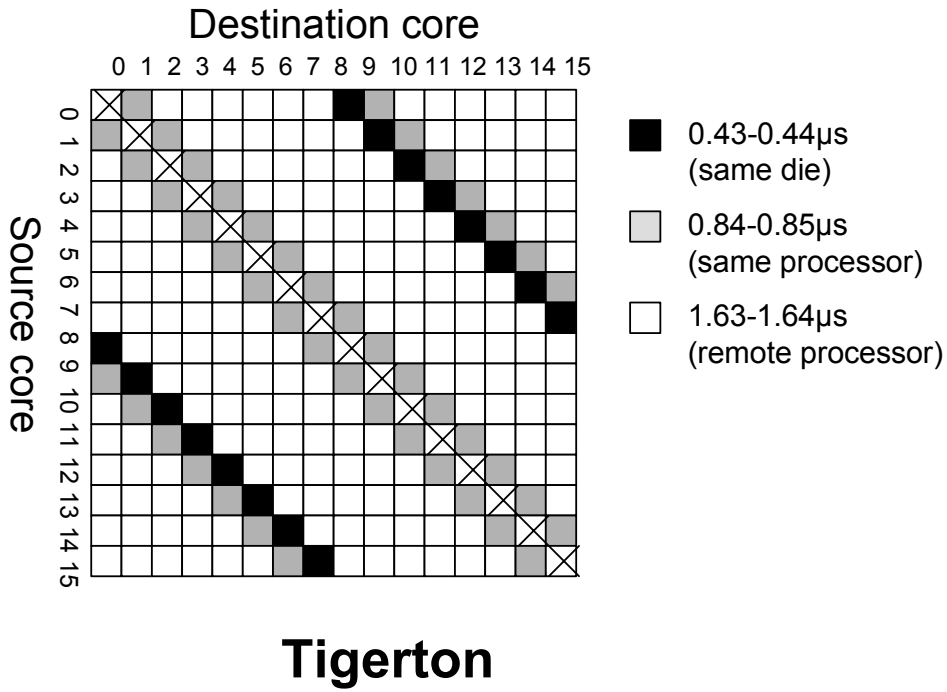


AMD

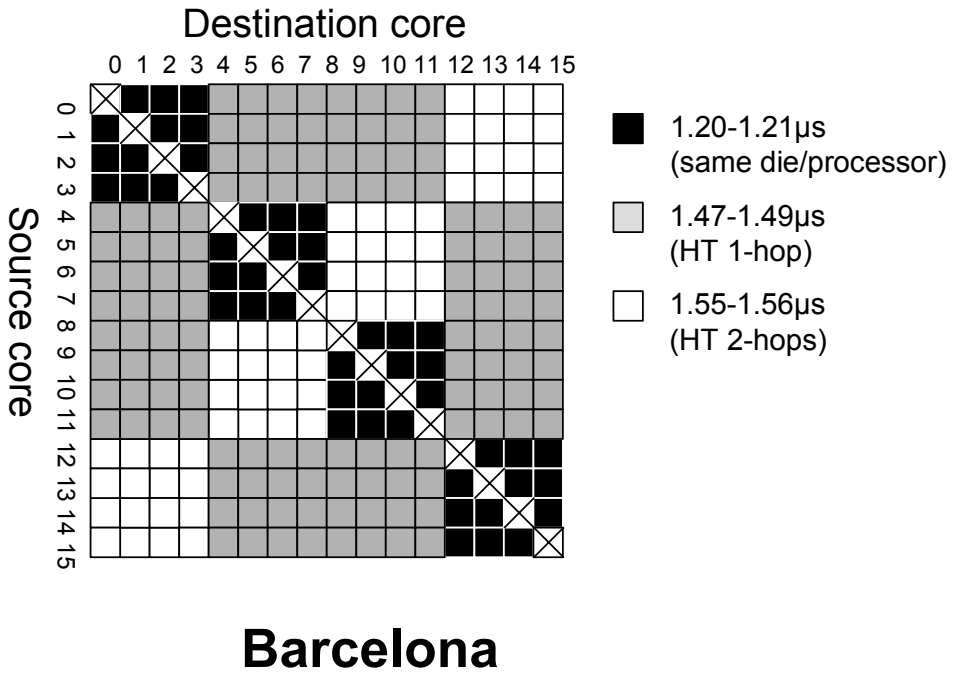




Processor locality



Tigerton

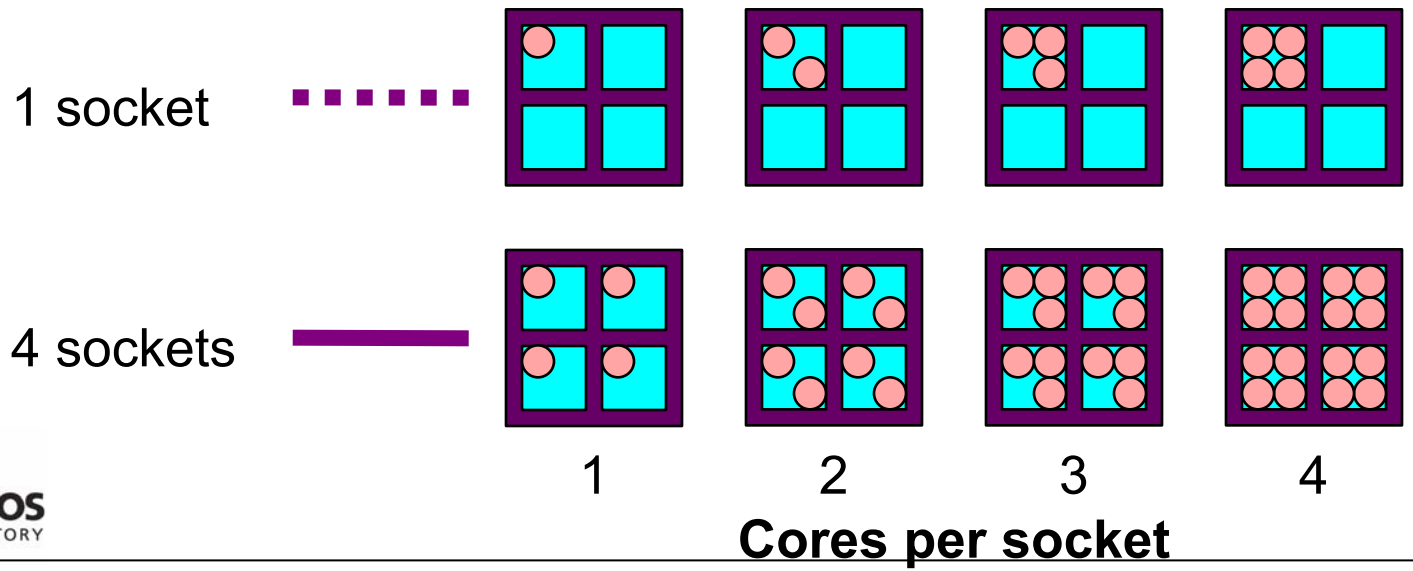


Barcelona

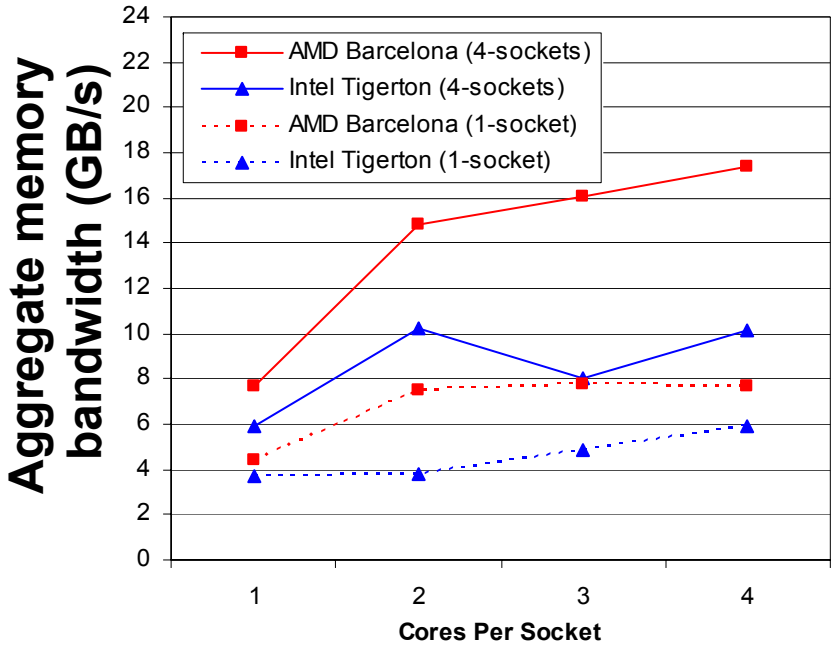
- **For performance experiments, need to know core ordering**
- **MPI ping-pong test from every core to every other core**
  - *Xeon X7350*: same die (DCM), same socket, different socket
  - *Barcelona*: same die/socket, one HT hop, two HT hops

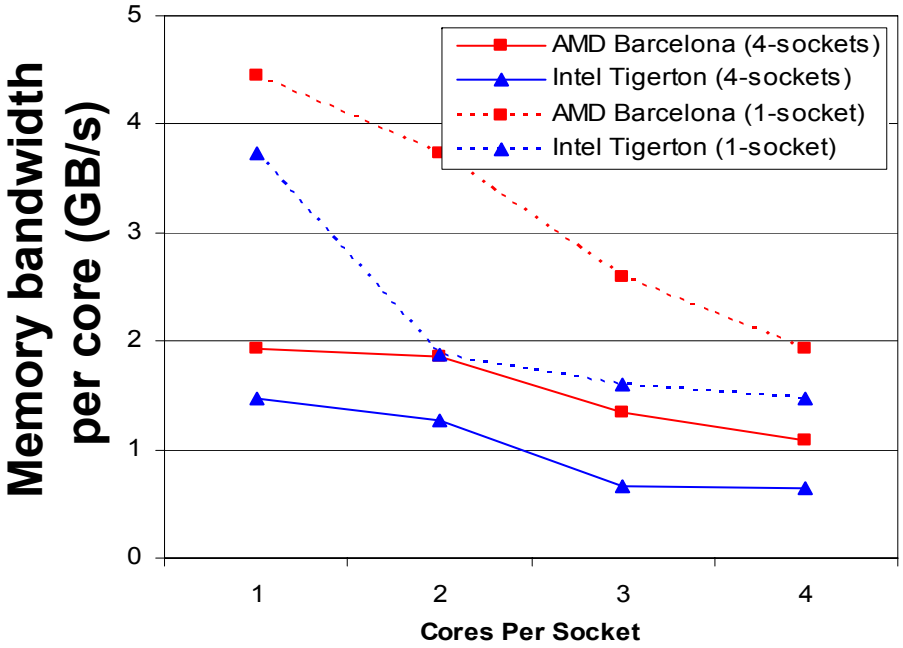
Measurement Methodology

- **Constant problem size per socket**
  - Strong scaling within a socket
  - Weak scaling across sockets
- **Mimics typical usage**
  - Weak scaling
  - Use all of the available memory in a node
- **Experiments:**



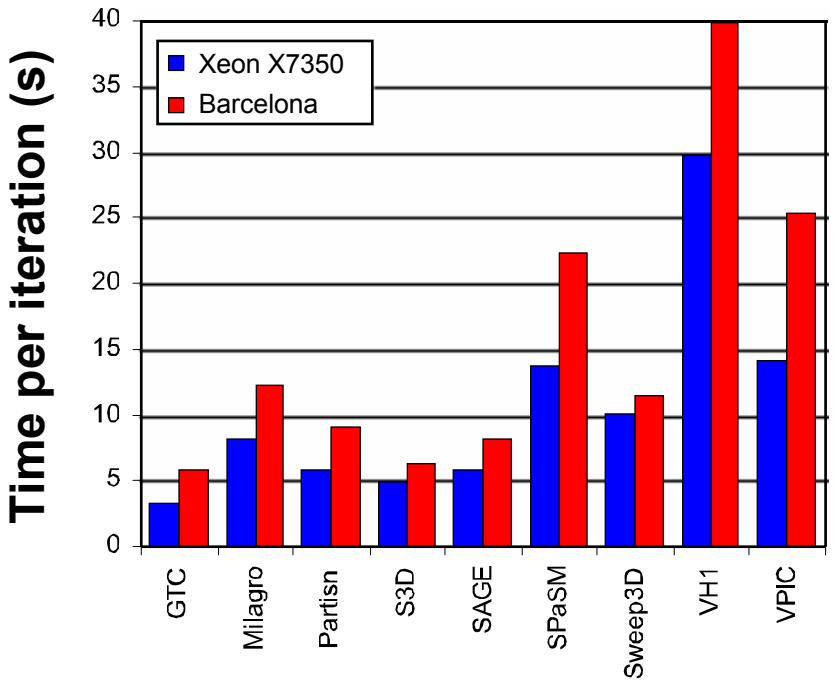
Microbenchmark: Memory bandwidth

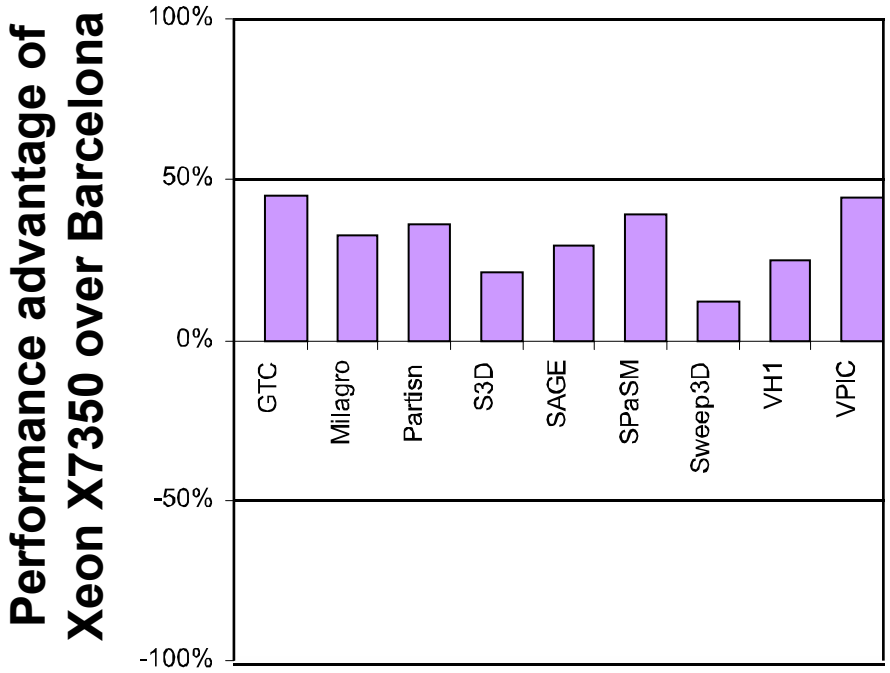




- **Streams triad**
- **Barcelona observes superior memory bandwidth to Xeon X7350 both per core and aggregate**

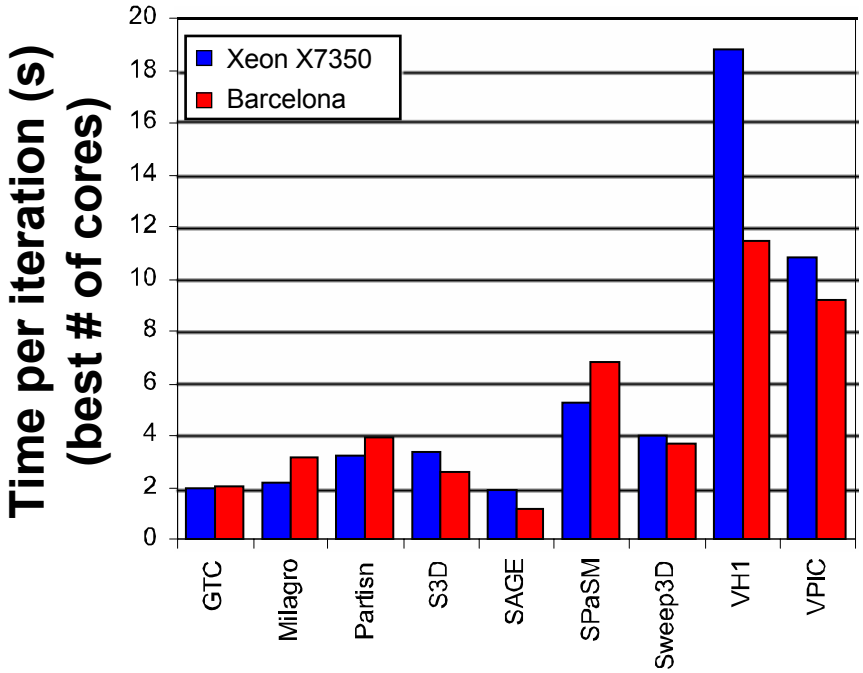
Single-Core Performance

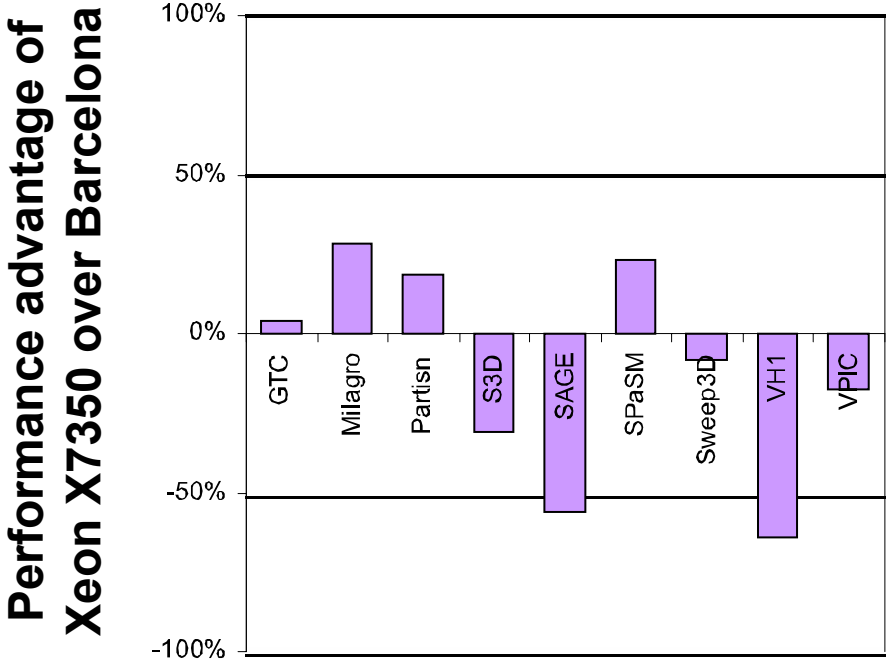




- **Xeon X7350 faster than Barcelona on all single-core tests**
  - 50% higher clock speed
  - Double the cache per core
  - Only 20% less memory bandwidth

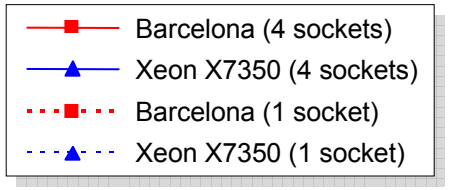
Full-Node Performance

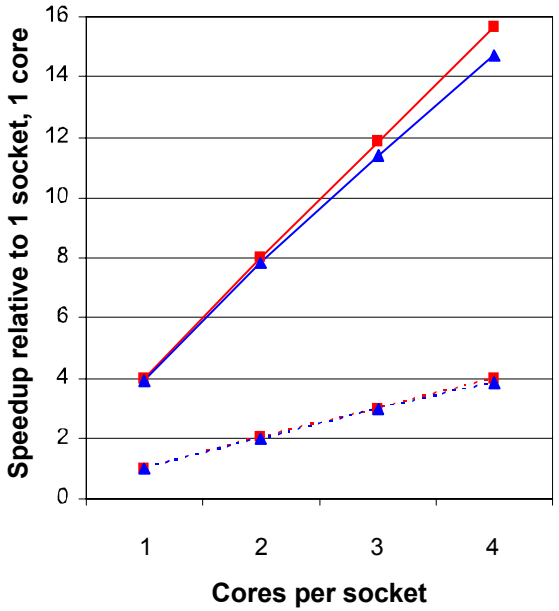




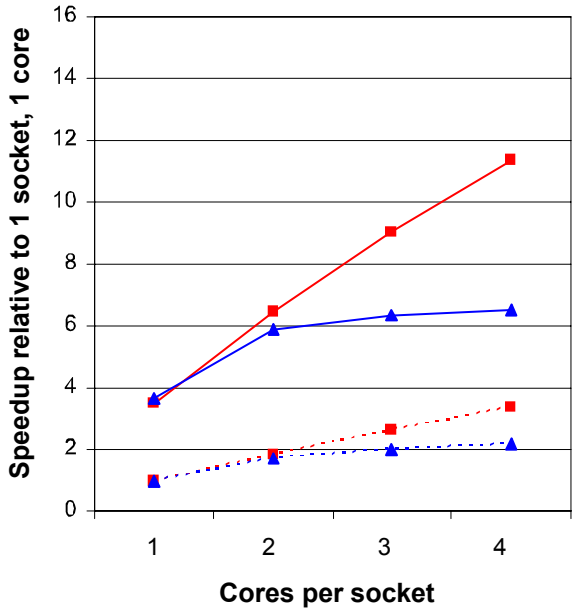
- **Barcelona outperforms Xeon X7350 on over half the applications studied**
  - 1.75X more per-core bandwidth at 16 cores (1.1 vs. 0.63 GB/s)

Application Scalability

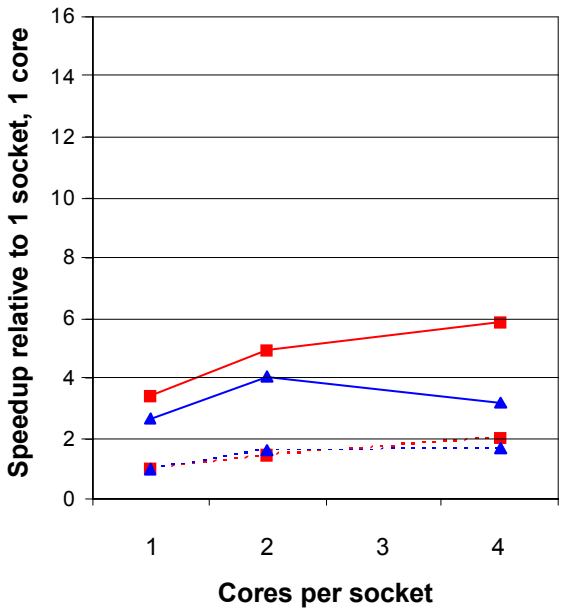




***Best*: Milagro**



***Average*: GTC**



***Worst*: Partisn**

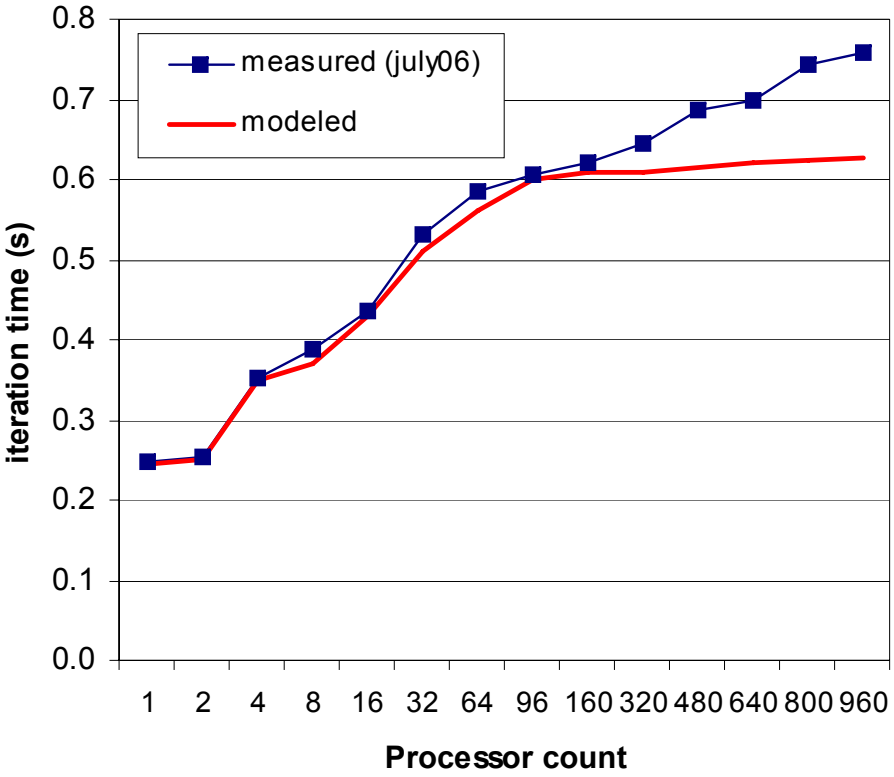
- **Milagro, SPaSM, and Sweep3D (compute-bound)**
  - Good speedup on both Xeon X7350 and Barcelona
- **VH1, GTC, VPIC, and S3D (neither compute- nor memory-bound)**
  - Good speedup on Barcelona, poor speedup on Xeon X7350
- **SAGE and Partisn (memory-bound)**
  - Poor speedup on both Xeon X7350 and Barcelona

Los Alamos NATIONAL LABORATORY EST.1943

A look at *mSystems*

- **Connectivity is an important issue**
  - Topologies
  - Routing
- **Hierarchical communication structures**
  - Traditionally: intra- & inter-node
  - Additionally: NoCs (Network on Chips)
    - » **Already see this on embedded devices: e.g. PicoChip, Cswitch, Tilera, and Cell-BE**
- **Take a look here at some existing, and possible, networks**
  - Infiniband
  - Meshes: Cray XT
  - Kautz: SiCortex
  - ‘All-to-all’: OCS

Infiniband: an example of *Model Driven Optimization*



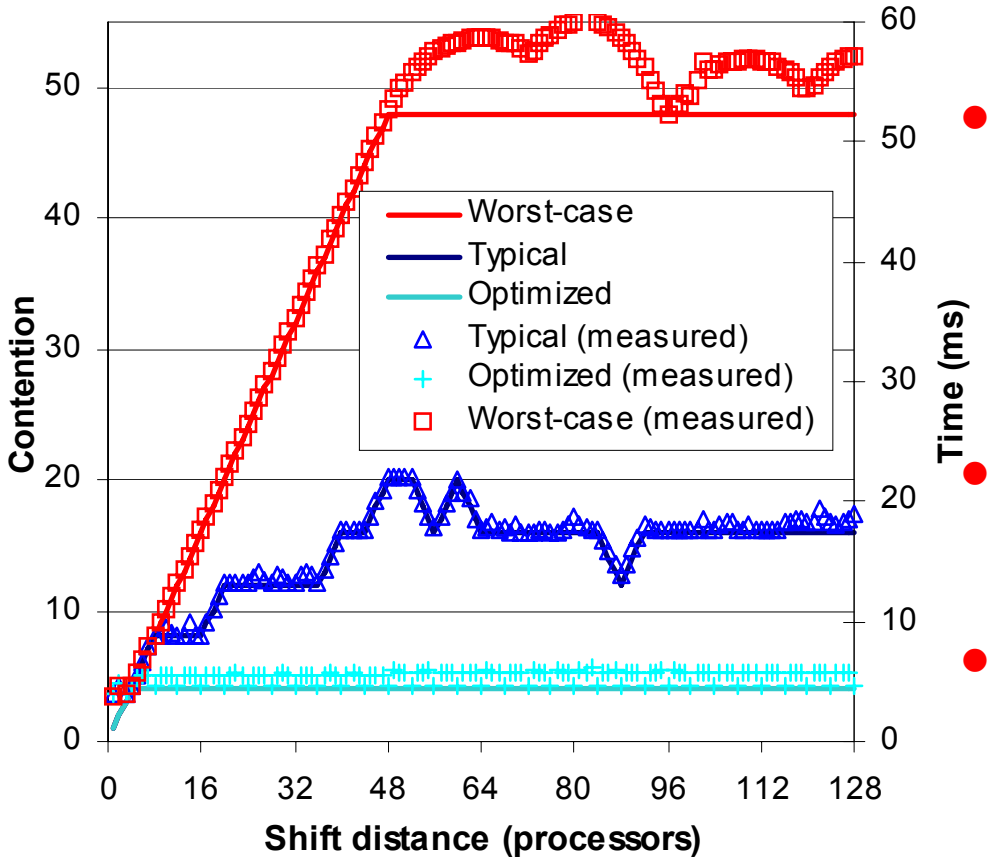
- **Example:**
  SAGE, 256 node,
  288-port IB 4x SDR

- **Model**
  - Developed several years ago
  - Good prediction accuracy
  - Include node -> network conter
  - Includes contention in mesh networks (e.g. BG/L) NOT fat-trees

- **No significant network contention observed on other Fat-tree networks (Quadrics)**

Contention can be an issue

- **Use logical-shift communication pattern**
  - P_i -> P_{i+d} where d = 1..128
- **Maximum modeled contention plotted (1024 PE job)**



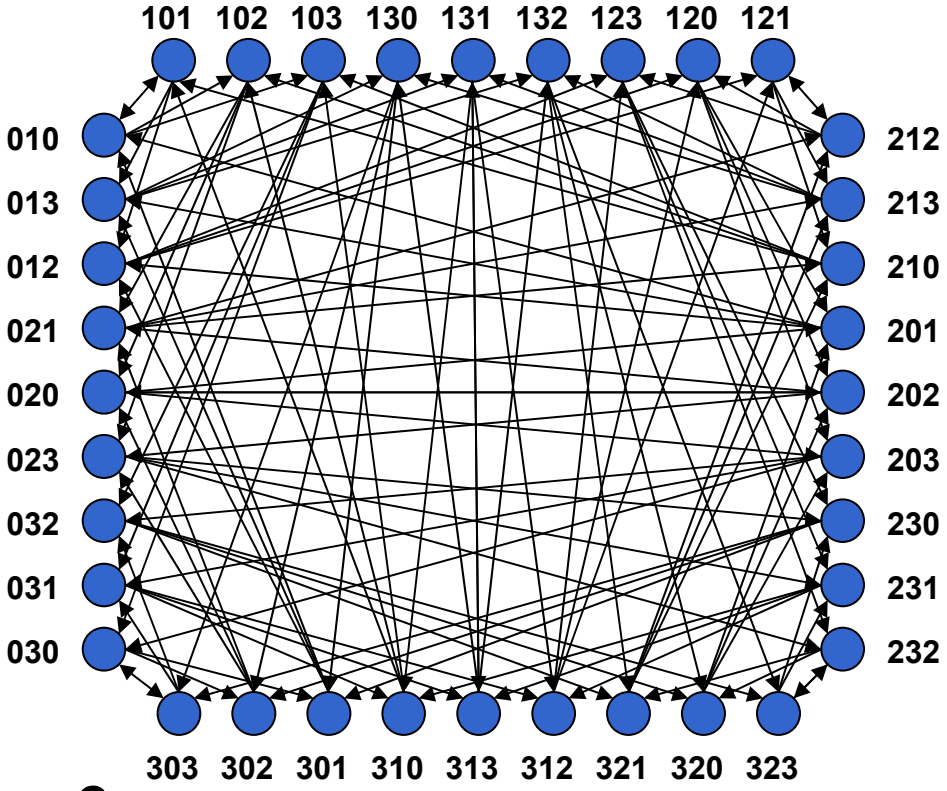
- **Worst-case: max of 48 (# PEs attached to 1 switch)**
- **Typical: contention generally increases with shift distance**
- **Optimized: max of 4 (bottleneck is node-size, PEs)**

Optimization of Infiniband for Scientific Applications. LSPP, IPDPS 2008, Miami, FL, April 2008

Kautz Graph topology

- **Kautz Graph:**
  - Largest node count for a given degree and diameter
- **SiCortex: Degree 3**
  - 3 input and 3 output links

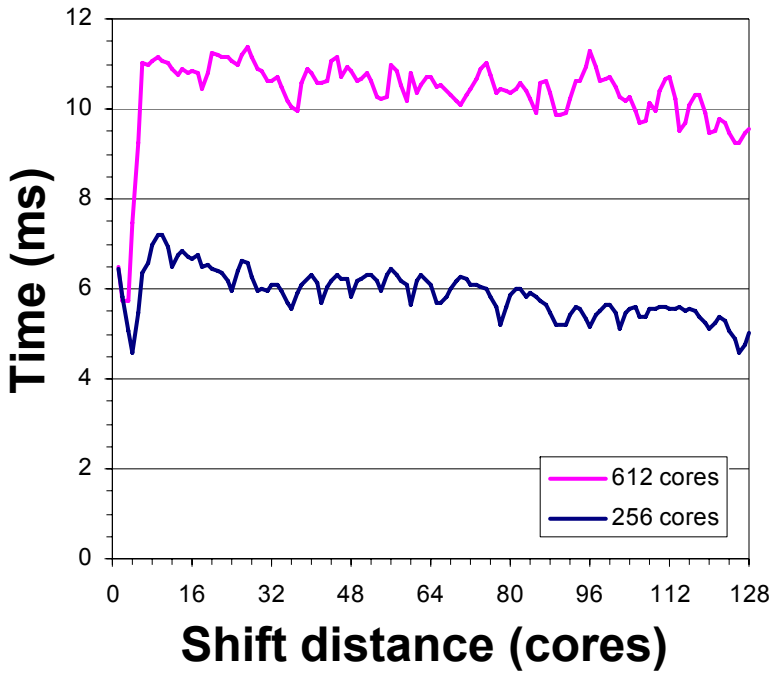
Diameter	*Node Count*	*SiCortex System*
2	*12*	*SC072*
3	*36*	
4	*108*	*SC648*
5	*324*	
6	*972*	*SC5832*

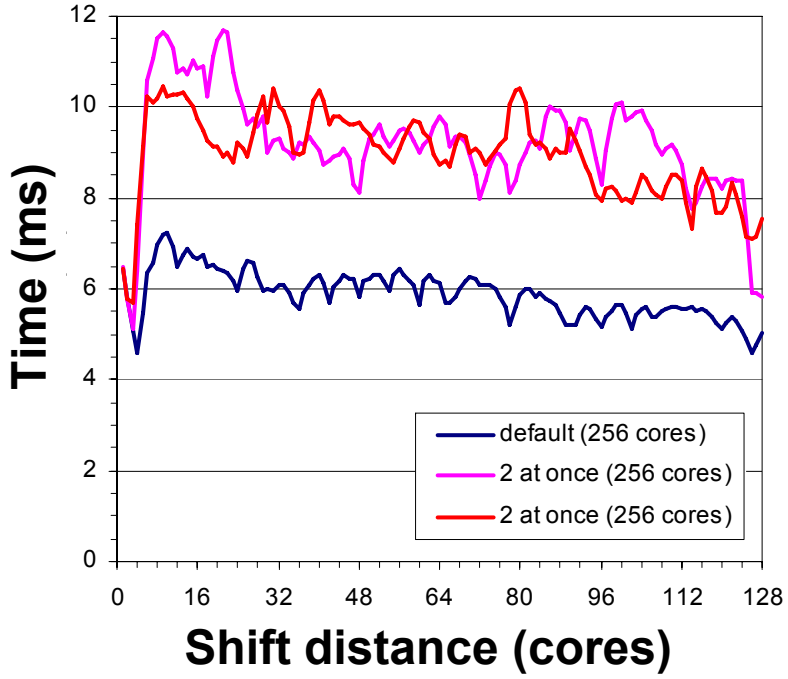


- **Example: Degree 3, diameter 3**
  - Node name: 3 symbols of a 4-character alphabet, no two adjacent symbols the same
  - Rule for node connections: XYZ -> YZ[W|X|Y]

Contention in the SiCortex Kautz Network

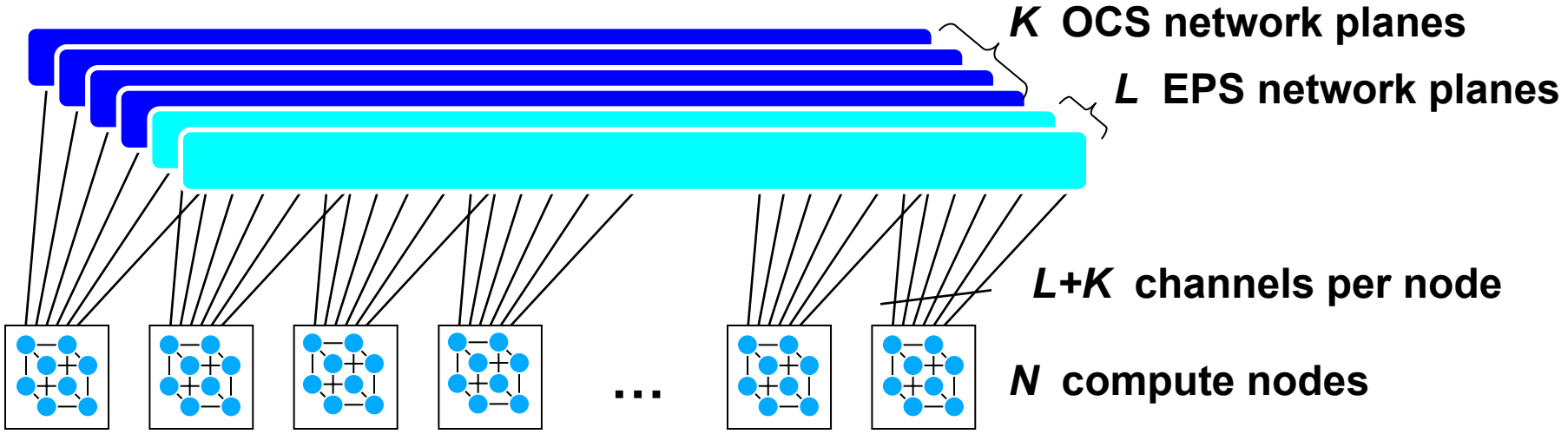
- **logical-shift communication pattern**
  - P_i -> P_{i+d} where d = 1..128





- **Available bandwidth can be used by smaller jobs**
  - Larger jobs can suffer increased contention (greater time)

What about a fully connected network? OCS - System Concept (HPCS, IBM)

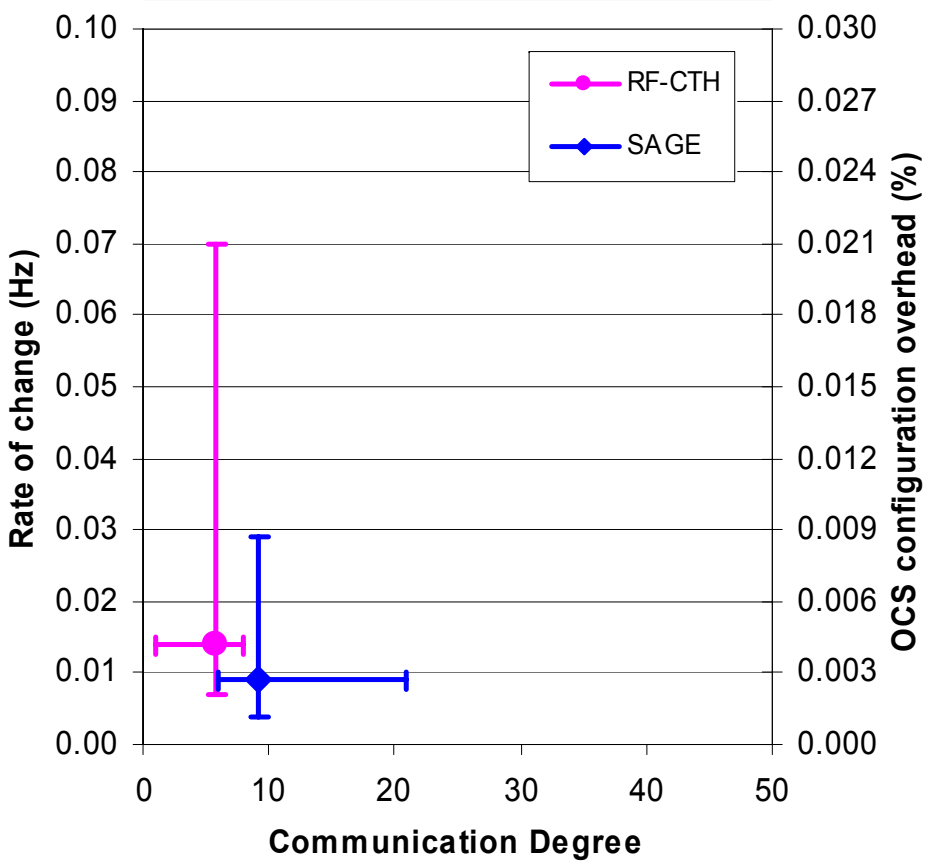


- **Bandwidth where it is needed (nodes actually communicating)**

- **Nodes: *m* PEs, (*L*+*K*) > m communication links**
- **Optical Circuit Switching (OCS) network planes**
- **Electronic Packet Switched (EPS) network planes**
  - low bandwidth links (~10% of OCS)
  - collectives

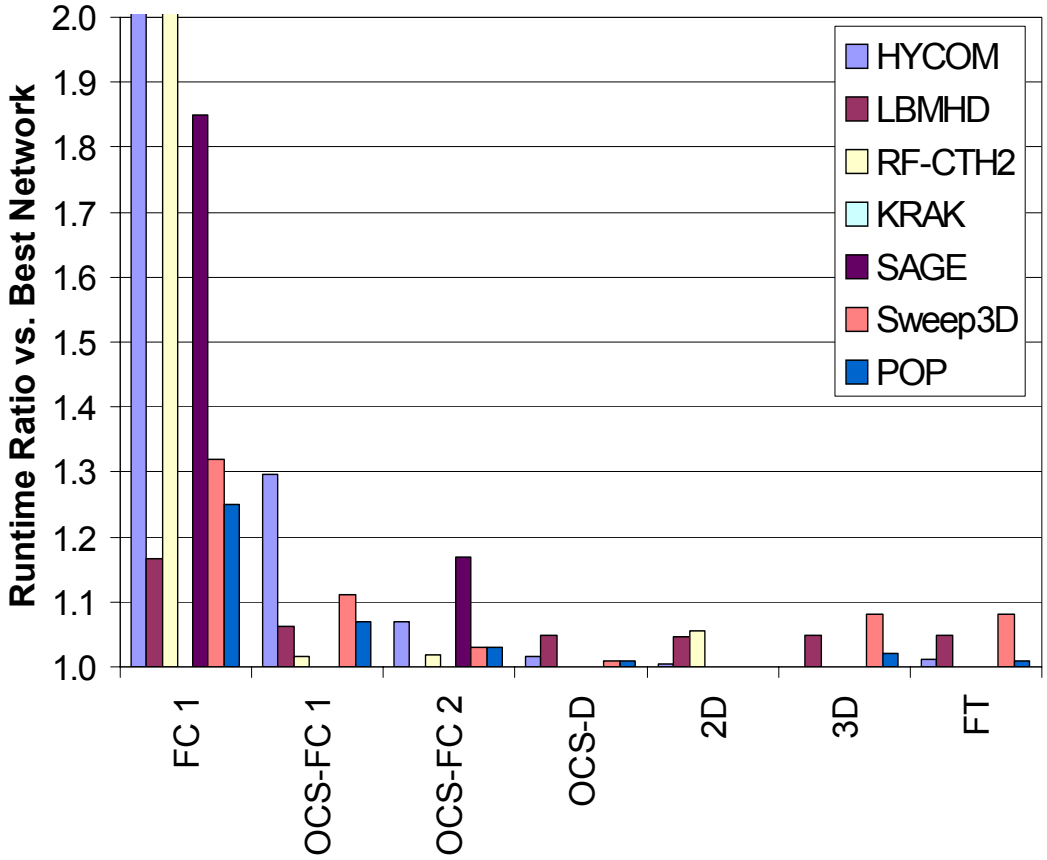
Communication degree: temporal analysis

- **Degree vs. rate-of-change (Hz)**
  - Higher rate-of-change means higher OCS set-up costs
- **e.g. 3ms OCS set-up:**
  - OCS overhead between 0% and 0.021%.
- **Using both OCS and EPS:**
  - Degree reduced
  - Rate-of-change unaltered



OCS performance: comparable to best

- **Analyzed performance of OCS in various system configurations**
- **Example: 2,048 PE job (256-node system, 64-way)**
  - FC Fully-connected 1-hop
  - OCS 1-hop or 2-hop
  - 2D, 3D meshes
  - FT Fat-tree
  - OCS-D OCS-Dynamic
- **Best hardware latency of 50ns, 4GB/s links**



- **Graph shows relative performance of each network relative to the best performing network**

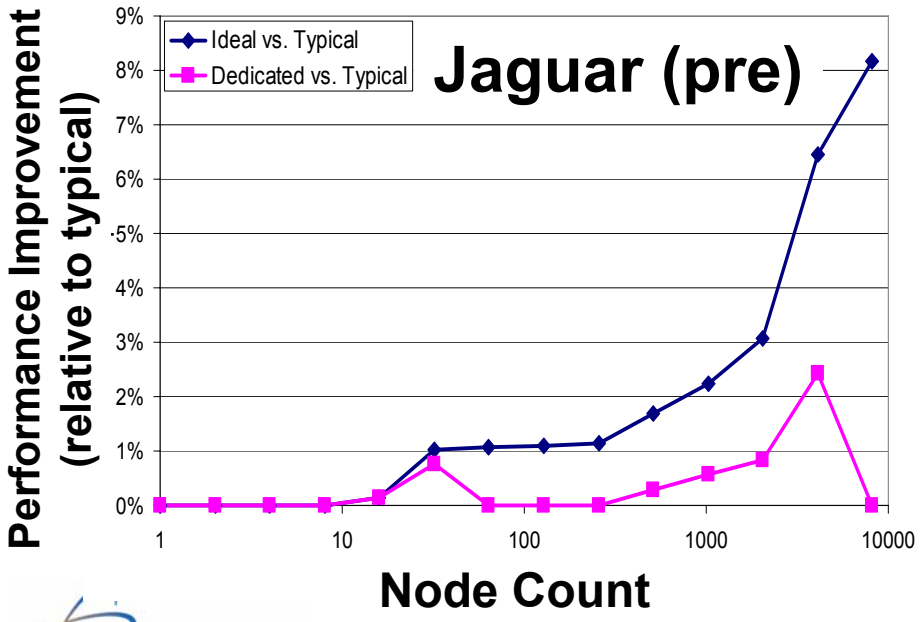
Performance Analysis of an Optical Circuit Switched Network for Peta-scale systems. EuroPar, August 2007

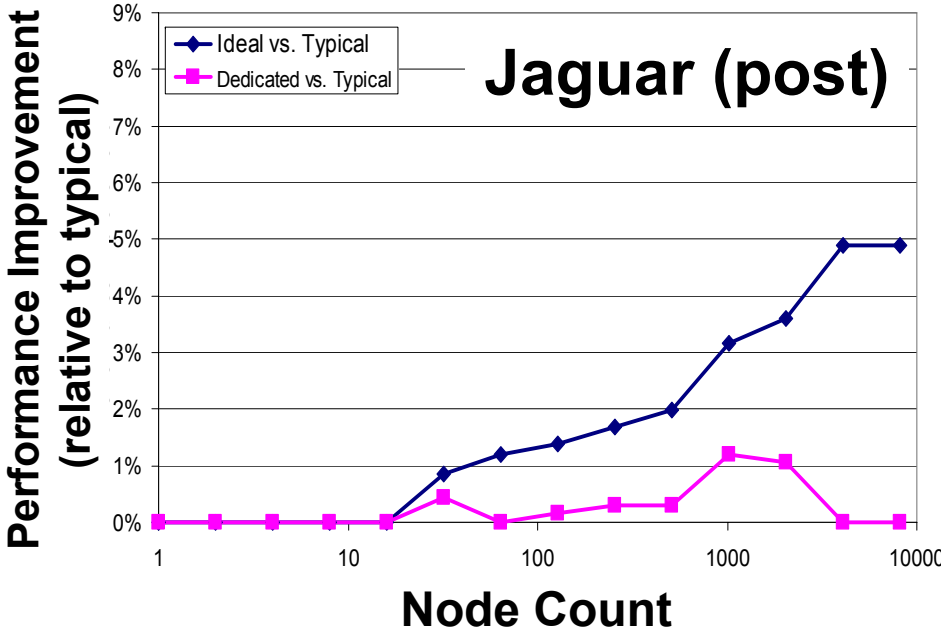
Jaguar *System* upgrade @ ORNL

- **Main aspects of Jaguar upgrade:**
  - Dual-core -> Quad-core
  - SeaStar 2 -> SeaStar 2+
- **Developed application performance models**
  - GTC and S3D
- **Models Validated on existing hardware**
  - Jaguar (pre-upgrade) & AMD/Infiniband system
- **Models used to predict performance**
  - Jaguar (post-upgrade)
- **Models used to explore network contention issues**

Contention in the XT4

- Jaguar pre- and post-upgrade
- Different allocations considered:
  - *Typical* – assigned by the scheduler
  - *Dedicated* – using the first n nodes of the system
  - *Ideal* – layout of nodes matches application



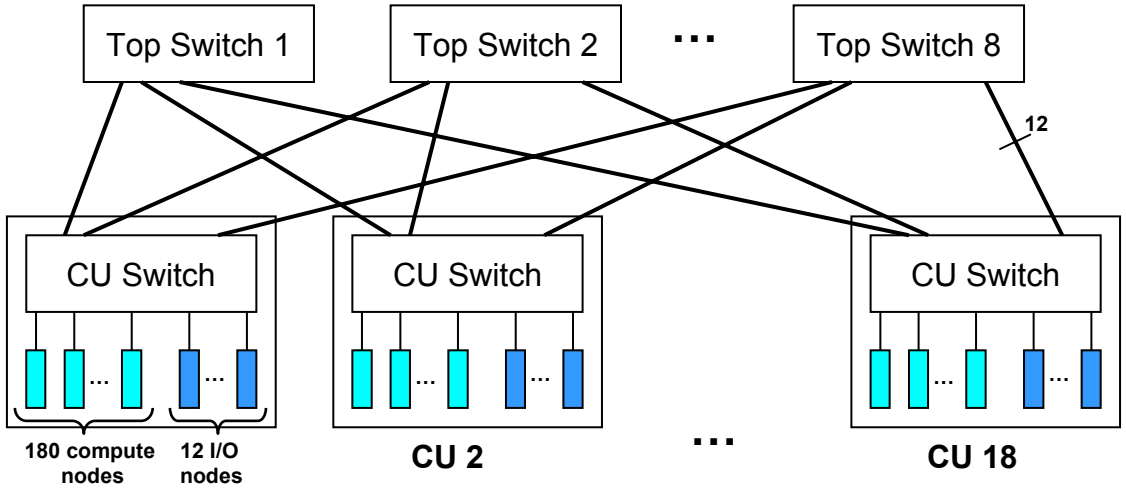


Example: GTC

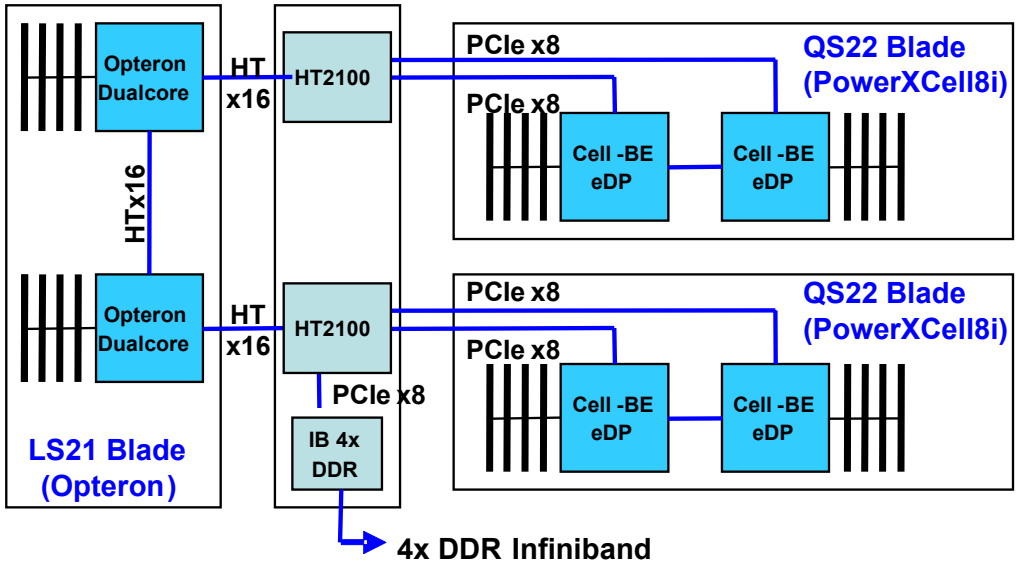
Roadrunner *System*

- **18 Connected Units**
  - 180 compute-nodes ea.
- **Infiniband DDR 4x**
  - Full fat-tree within CU
  - Half fat-tree between CUs

System	
CU count	18
Node count	3,240
Peak Performance (DP)	1.46 Pflops/s
Connected Unit (CU)	
Node count	180
Peak performance per CU	80.9 Tflops

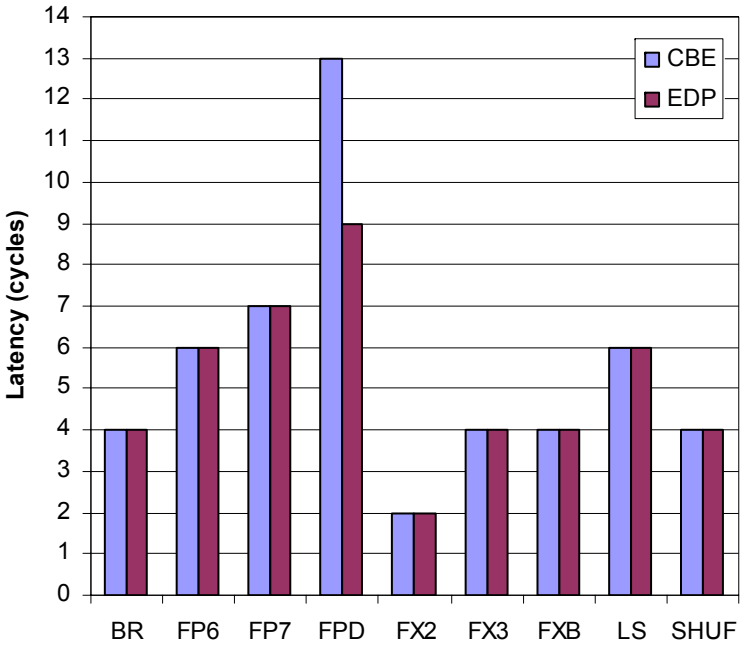


Roadrunner node – a 'triblade'

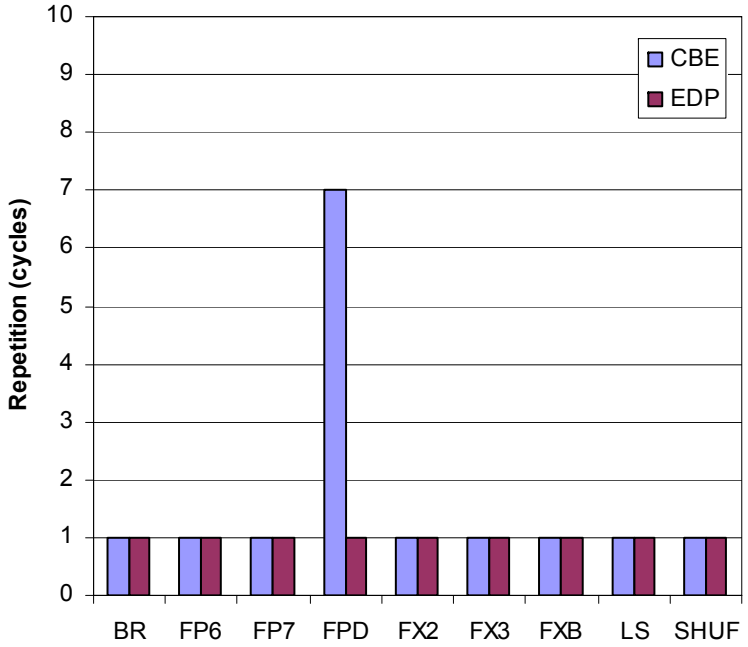


Node (triblade)	1 Opteron blade	2 Cell blades
Processor count	2	4
Processor-core count	4	4 PPEs, 32 SPEs
Clock Speed	1.8 GHz	3.2 GHz
Peak-performance per node (DP)	14.4 Gflops/s	435.2 Gflops/s
Memory per processor	4 GB (800MHz DDR2)	4 GB (800MHz DDR2)

PowerXCell8i : Instruction characteristics



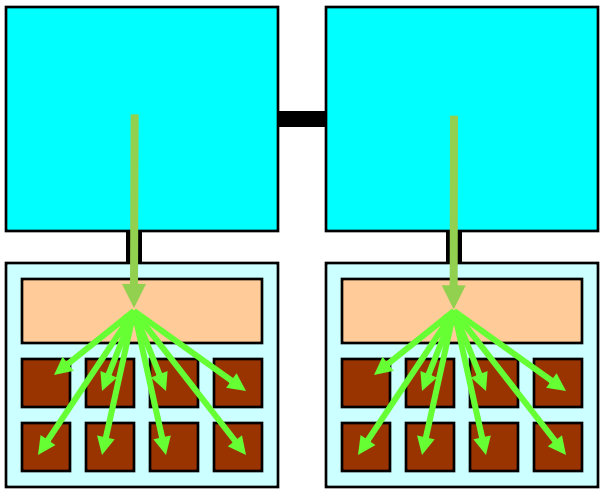
Instruction Latency

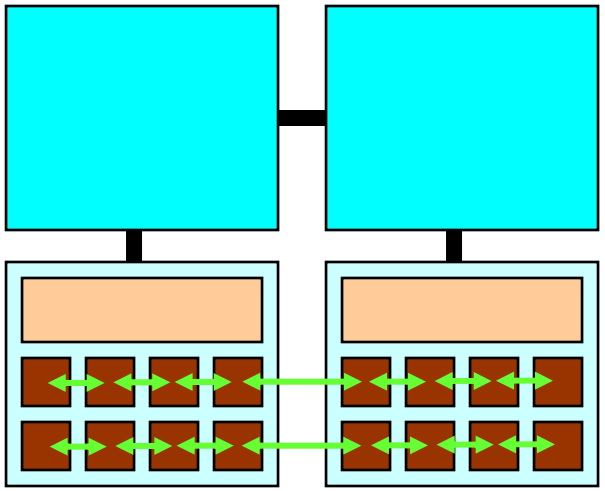


Repetition Delay

- **Two different implementations of the Cell-Broadband Engine**
  - PowerXCell 8i version has 7x improved FPD repetition delay, and
  - Slightly lower pipeline latency

Using of accelerators



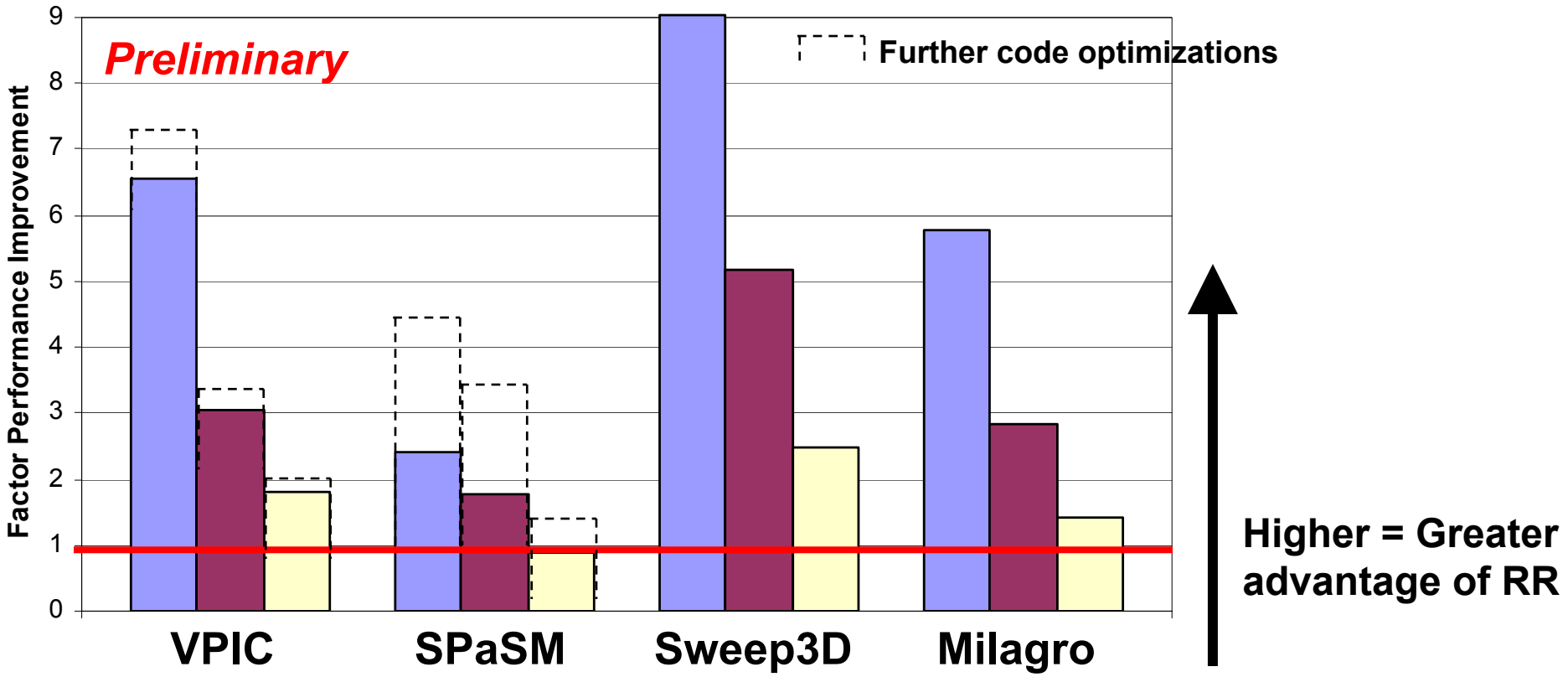


- **General accelerator approach**
  - One MPI rank per Opteron
  - SPE = accelerator
  - Opterons see each other and their local SPEs
  - Opteron pushes work (data) to SPEs and receives results

- **Cell-Messaging-Layer**
  - One MPI rank per SPE
  - Opteron = NIC & extra storage
  - SPEs see each other and their local Opteron
  - SPEs communicate directly with other SPEs
  - PPE provides support
  - “Cluster of 100,000 SPEs”

Los Alamos NATIONAL LABORATORY EST.1943

Roadrunner performance comparison



RR without Cell (2-socket x 2-core) vs. RR with Cells

Barcelona (2s x 4c) vs. RR with Cells

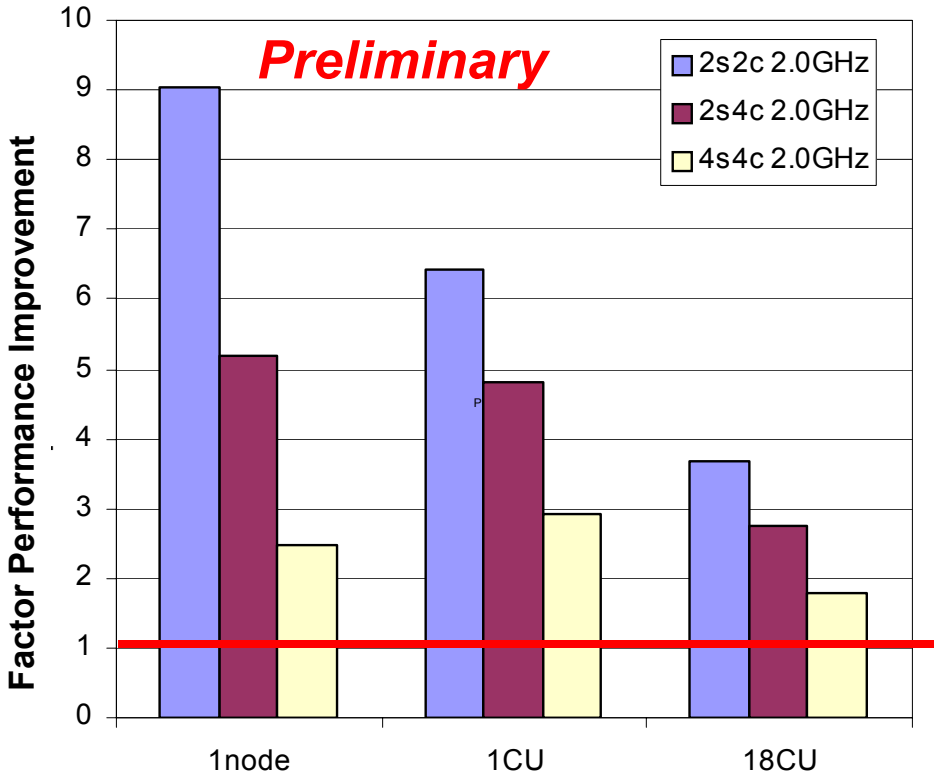
Barcelona (4s x 4c) vs. RR with Cells

CCS-2 & CCS-1 code teams

Roadrunner Performance Comparison: for Sweep3D

Performance of

Roadrunner
vs.
equivalent
Quad-core System



Advancing Architectures

- **Technology:**
  - Heterogeneity, accelerators , GPUs
  - Clusters on a chip (cores++, networks)
    - » **Network hierarchy (cf memory hierarchy)**
  - Integrating processors on top of memory, or
  - Integrating memory on top of processors
  - Silicon Photonics
  - Hierarchical Connectivity (many levels of networks)
- **Workload:**
  - Programming models
  - Code optimizations
    - » **Overlap: communicate and compute**
    - » **Overlap: memory and compute (SW prefetching)**
- **All of the above ?**

Performance modeling can help in this process

Summary

Core performance + application performance model =

Performance Exploration

Predictions at scale
Predictions on new systems
Predictions in the design space

μSystem **: quad-core nodes**

mSystems **: networks increasingly important**

Infiniband, Kautz, OCS

Systems **: Modeling used to examine:**

Jaguar - performance during system upgrade

Roadrunner – performance in advance of deployment & compare against other state-of-the-art systems