



Performance of Roadrunner at Scale Under a Realistic Workload

Adolfy Hoisie

**Leader, Computer Science for High-Performance Computing Group (CCS-1)
Performance and Architecture Lab (PAL)**

**Darren Kerbyson, Kevin Barker, Kei Davis, Mike Lang,
Scott Pakin, Jose Sancho
PAL, CCS-1**



LA-UR-07-7574



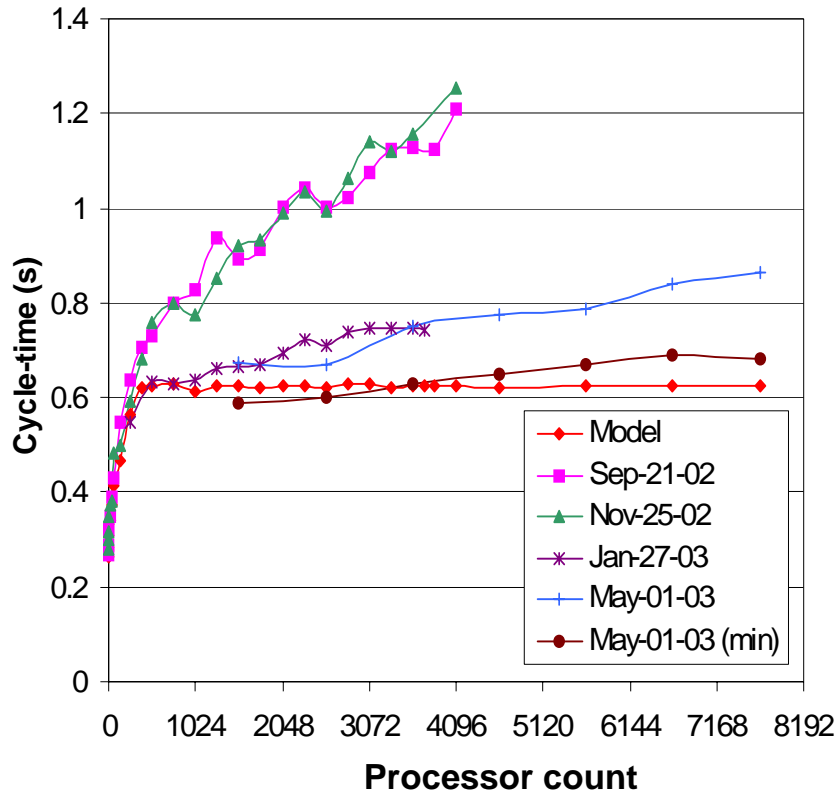
Outline

- **Performance modeling methodology**
- **Architecture and performance parameters review**
- **Application performance at scale**
 - VPIC
 - SPaSM
 - Sweep3D
 - Milagro
- **Performance prediction at scale**
- **Comparisons**

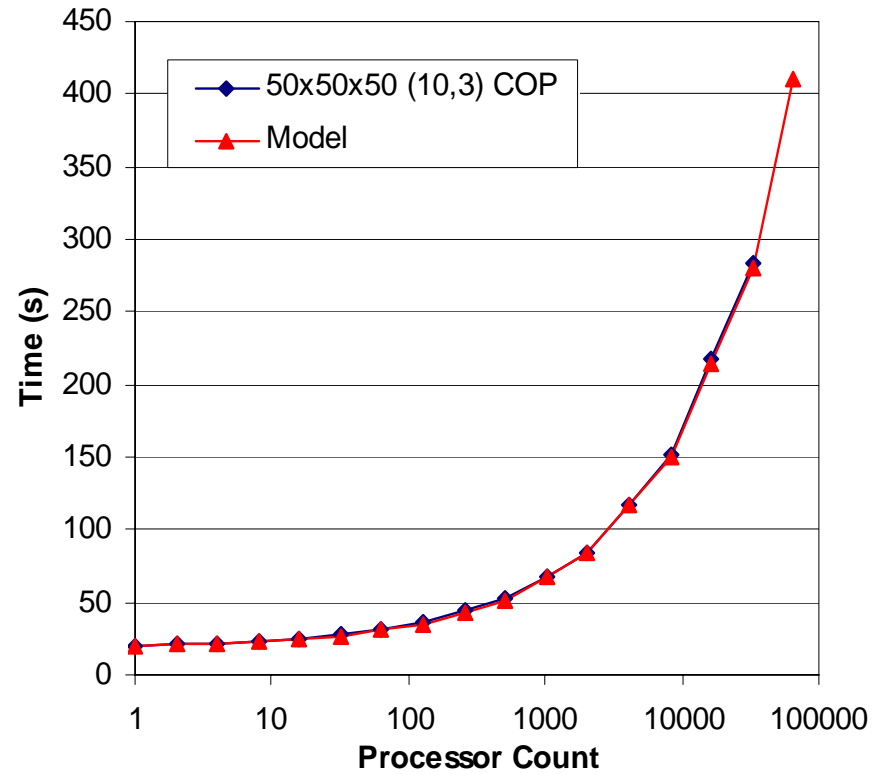
Performance Modeling in Action at LANL

- **Novel techniques developed by PAL at Los Alamos in the last decade**
- **Methods are quasi-analytical**
- **Models encapsulate performance of entire apps on full systems**
- **The workload considered already is large and diverse (ASC, SC, DARPA, NSF)**
- **Models were validated on most of the large supercomputers in the last decade**
- **Models are our tools for performance analysis.**
- **We apply modeling to : system design (PERCS, BG), system optimization (Purple, ASCI Q), application optimization, performance prediction for future architectures, performance calibration, etc**
- **Models are predictive, and highly accurate**

Accuracy of PAL's Performance Models

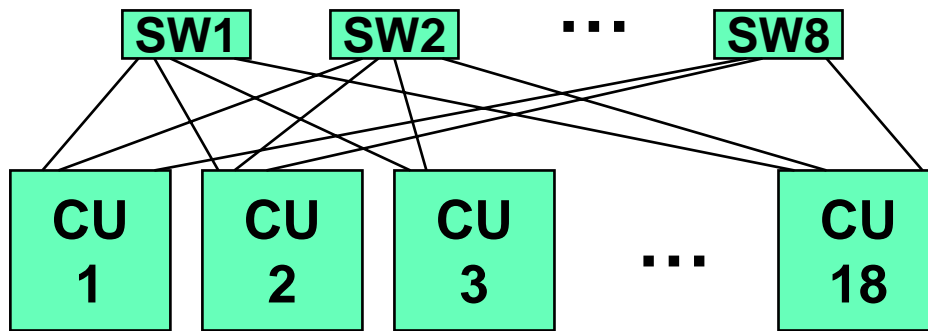


Sage on ASCI Q



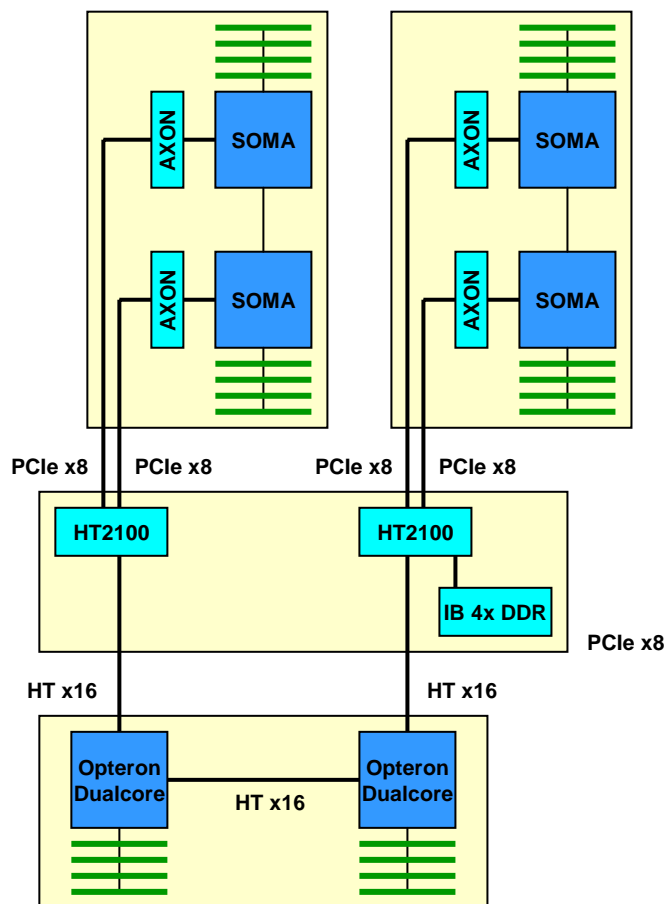
Sweep3D on BG/L

Essential System Peak Performance Parameters



- System = 18 CU = 3240 triblades
= 12960 (AMD cores + cell eDP)
- Peak DP flops = 1.33Pf/s
- Memory capacity=77 TBytes
- Peak memory bandwidth (cells) = 0.277PB/s

Essential Triblade Peak Performance Parameters



Tri-Blade 1

- **4 Cell eDP = 4x (PPU + 8 SPU)**
 - Cell eDP = 104 Gflop/s (DP)
 - = 208 Gflop/s (SP)
 - Memory bandwidth = 21.3 GB/s / cell
- **4 AMD cores 1.8GHz**
 - AMD = 3.6 Gflop/s (DP) / core
- **Cell <-> AMD**
 - Bandwidth = 2.0GB/s + 2.0GB/s
 - Latency ~1.5us
- **AMD <-> AMD (inter-node)**
 - Bandwidth = 2.0GB/s + 2.0GB/s
 - Latency ~ 1.5us

Data Movement Performance

Characteristics of RR: Input to Models

		Worst	Probable	Best
Single Cell -> Opteron (uni)	Latency	4.5us	3us	1.5us
	Bandwidth	1.2GB/s	1.4GB/s	1.6GB/s
All cells -> Opteron (uni)	Latency	5.5us	4us	2.5us
	Bandwidth	1.1GB/s	1.3GB/s	1.5GB/s
Single Cell -> Opteron (Bi)	Latency	5.5us	4us	3.5us
	Bandwidth	1GB/s	1.2GB/s	1.4GB/s
All cells -> Opteron (Bi)	Latency	6.5us	5us	3.5us
	Bandwidth	0.9GB/s	1.1GB/s	1.3GB/s
Infiniband (Uni)	Latency	2.2us	2.0us	1.8us
	Bandwidth	1.3GB/s	1.5GB/s	1.7GB/s
Infiniband (Bi)	Latency	2.7us	2.5us	2.3us
	Bandwidth	1.2GB/s	1.4GB/s	1.6GB/s

Applications

- 1) VPIC
- 2) SPaSM
- 3) PAL-Sweep3D
- 4) Milagro

- For each:
 - Input deck characteristics
 - Performance predictions
 - Performance advantage of RR with Cell
 - Bottlenecks: where the time is really spent
 - Comparison with Q
 - Comparison with a hypothetical multicore cluster

VPIC Model Input Parameters

- **Two species of particles: electrons and ions**
 - Electron particle transfer messages are the largest
 - Particle species have equal computational requirements
- **Problem size on all platforms is held fixed at 8M particles per species per node for both problems**
- **Iteration count:**
 - Electron performance sorting occurs every 50 iterations
 - Ion performance sorting occurs every 100 iterations
 - Model predictions are in terms of average iteration time

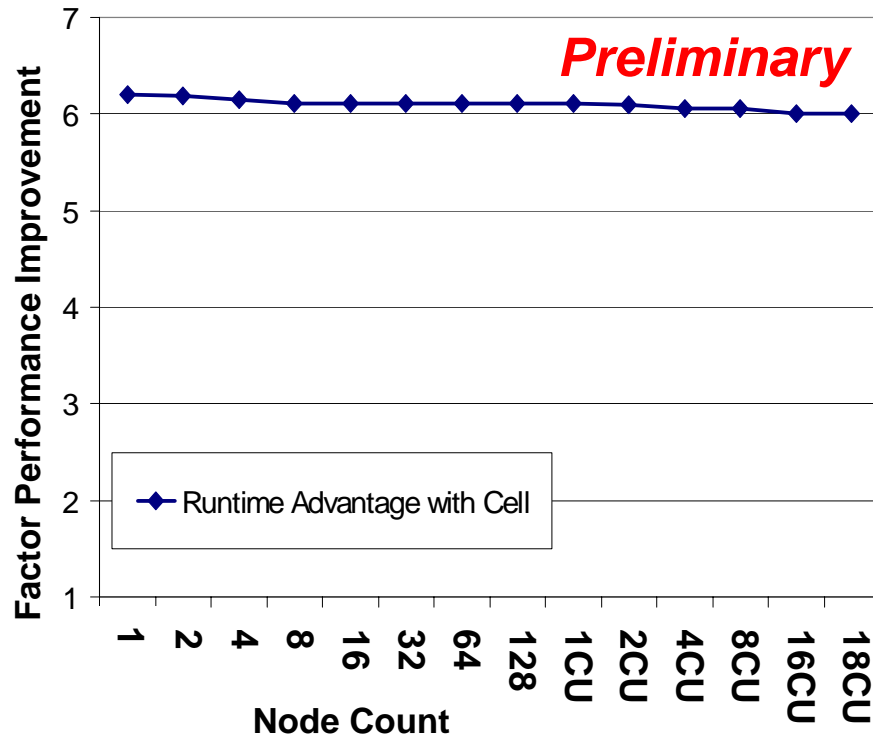
Input-deck	Hot 3D
Voxels / processor	16x16x16
Particles / Voxel	512
Particle species	2
Particles per node	16 M
Size of particle (Message Passing)	44B
Compute per particle (Opteron)	67 ns
Compute per atom (Cell-eDP)	12 ns

VPIC Workload Characteristics

- **Mapping of VPIC to the Triblade**
 - Processing:
 - » Cell - SPU: Particle push calculation
 - » Cell - PPU: Sorting, etc
 - » Opteron: MPI message relay
 - Message Passing: All messages originate on Cell and are relayed through Opteron
- **Message characteristics**
 - Particle transfer:
 - » One message per neighbor per iteration per species
 - » 44 Bytes per transferred particle (approx 50KB total)
 - Remaining messages are small; typically 4 bytes
 - Approx. 20-25 messages per neighbor per iteration
- **Overall performance is computationally bound on all platforms**
- **Model initially developed for non-accelerated VPIC**
 - Validated with high accuracy on 1024core AMD IB cluster
- **Refined to reflect hybrid implementation using message relay**
- **Model accuracy to within 5% on available hardware**

VPIC: RR Performance predictions

Runtime on Opterons / Runtime on accelerated RR

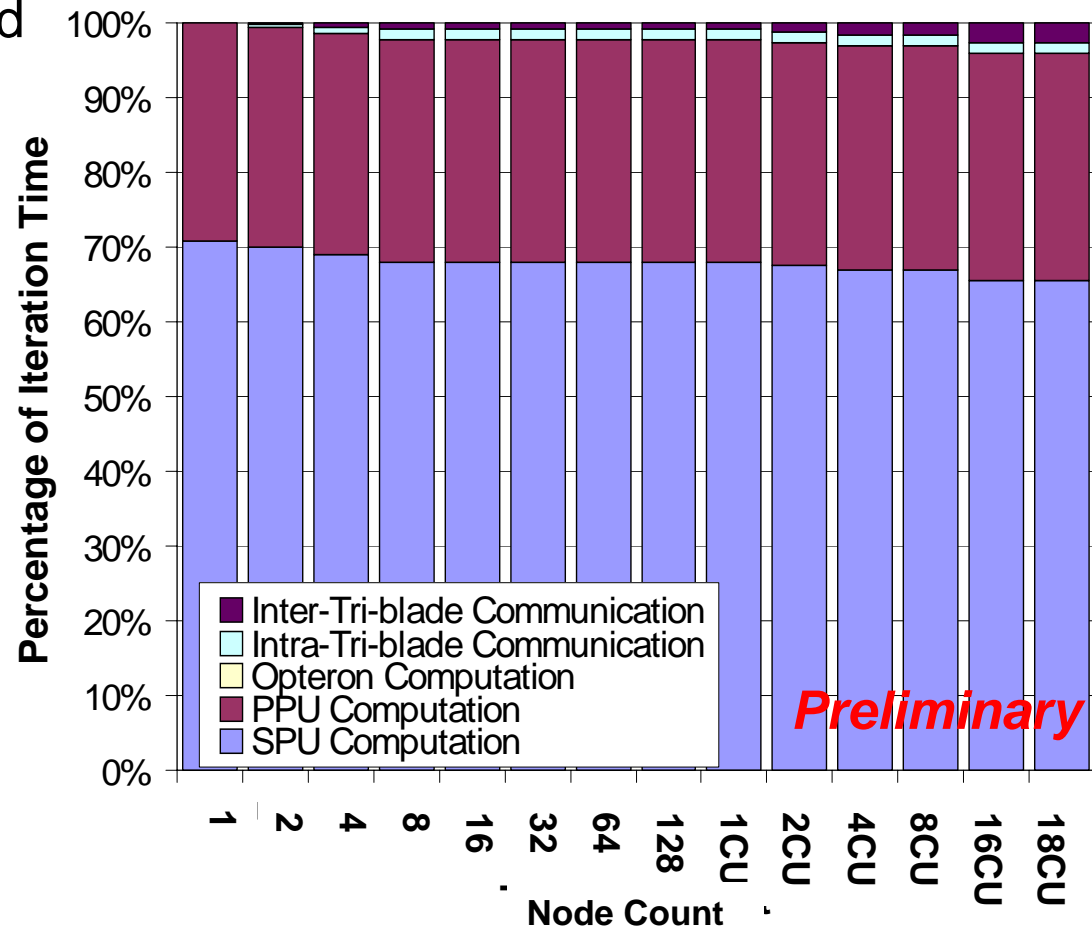


- Very Good scaling expected
- With current code, expect a factor of ~6x better performance using Cell

VPIC: Profiling

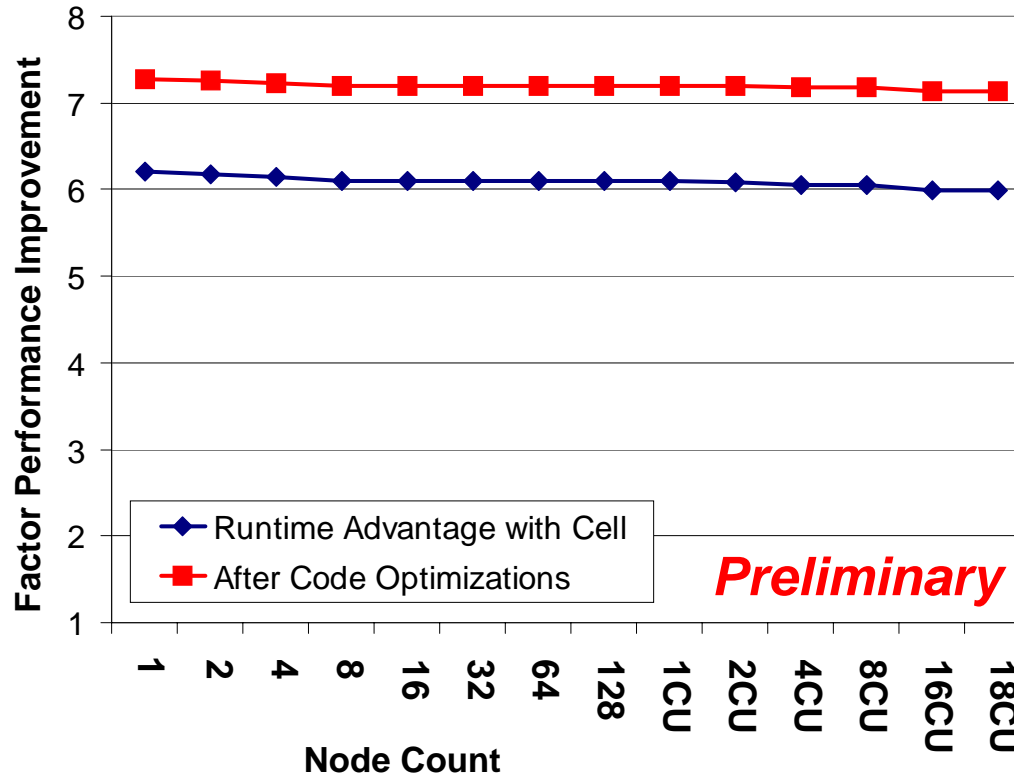
● Where is the time being spent ?

- Remains compute bound
- ~65% SPU
- ~31% PPU
- ~1 Cell \leftrightarrow Opteron
- ~3% Infiniband



VPIC: Possible Code Improvements

- **Between now and RR deployment expect:**
 - Migration of particle sort and graph operations from SPU to PPU (x0.5)



SPaSM Model Input Parameters

- **Single species of particles**
 - uniform spatial distribution (crystalline structure, possibly with small voids)
 - uniform, very short range interactions
- **Problem size on all platforms is held fixed at 1.5M particles (64*64*64 unit cells * 6 particles/cell)**
- **Iteration count:**
 - subsequent iteration times vary by a maximum of a few percent
 - model predictions are in terms of average time for a few typical iterations

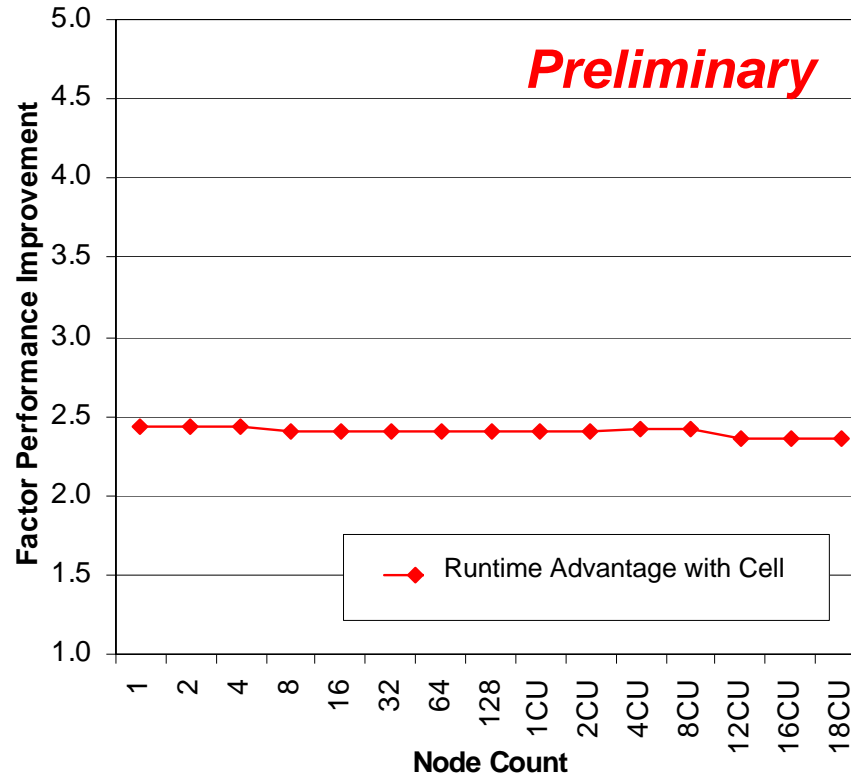
Input-deck	R2
Unit cells / processor	64x64x64
Computational cells / processor	46x46x69
Av. Atoms / c-cell	10.8
Skin Depth	2
Size of particle (Node<->Node)	590B
Size of particle (cell <-> Opteron)	132B
Compute per atom (Opteron)	1.23us
Compute per atom (Cell-eDP)	2.7us

SPaSM Workload Characteristics

- **Acceleration of major part of processing**
 - Accelerated 90% of original microprocessor cycles
- **Processing flow (cycle):**
 - Prepare data on AMD for Cell
 - Transfer data volume to Cell (~230MB)
 - Process data on Cell
 - Transfer data volume back to AMD (~230MB)
 - Postprocess on AMD
 - Update Particles on AMD
 - Exchange boundaries between AMDs (~250MB)
- **Model initially developed for non-accelerated SPaSM**
 - Validated with high accuracy on 1024core AMD IB cluster
- **Refined to reflect hybrid implementation (PNH)**

SPaSM: RR Performance predictions

Runtime on the base cluster / Runtime on accelerated RR

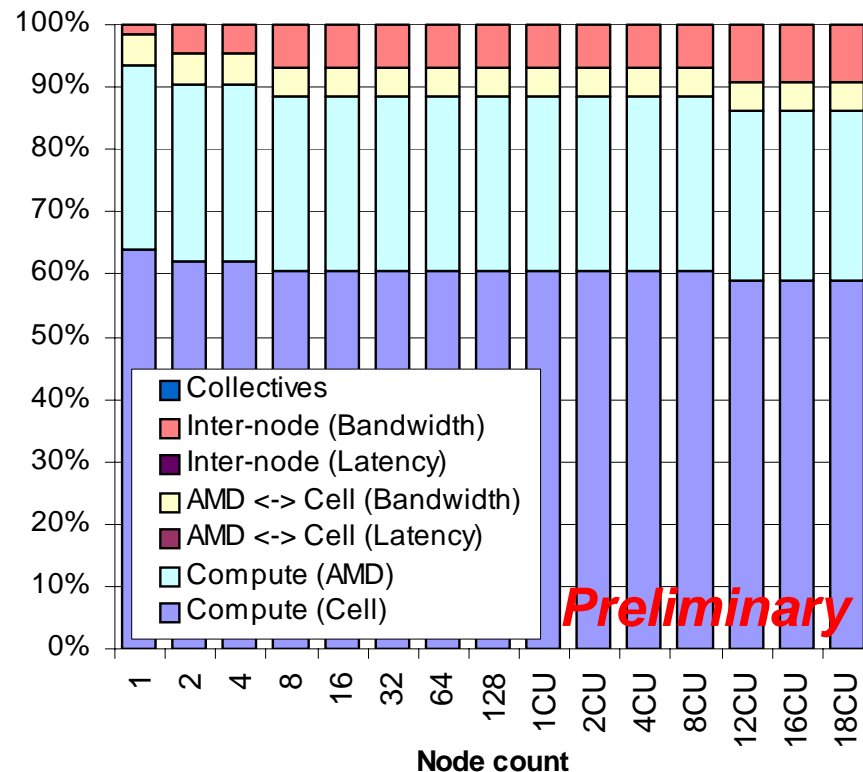


- Very Good scaling expected
- With current code, expect a factor of ~2.4x better performance using Cell

SPaSM: Profiling

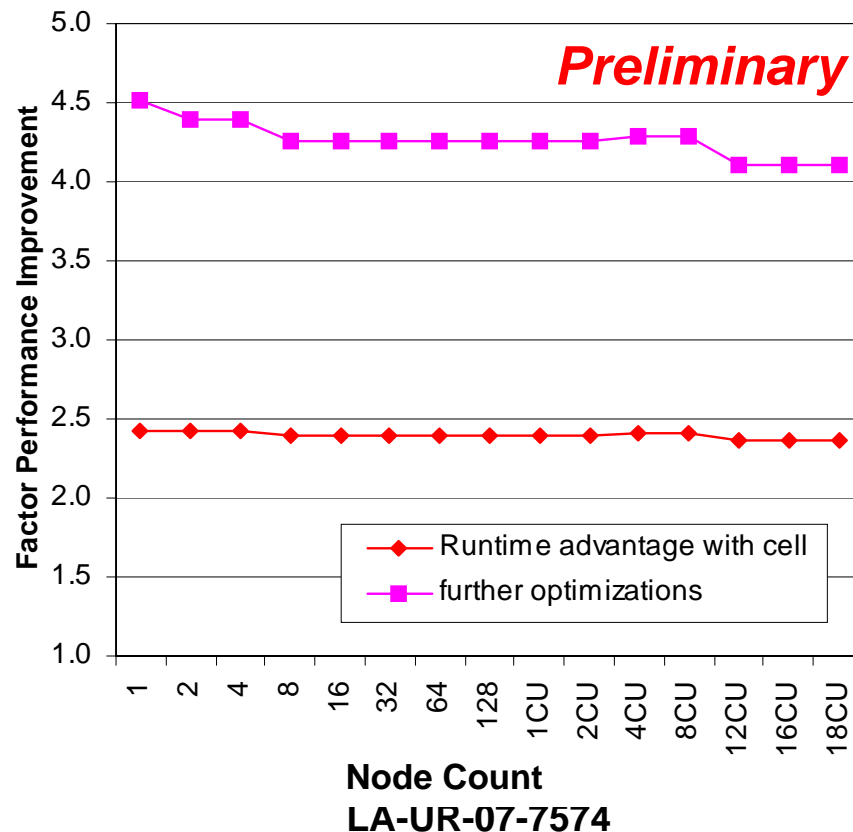
- Where is the time being spent ?

- Remains compute bound
- ~60% time on the Cell
- ~26% time on Opteron
- ~9% in Infiniband
- ~5% in Cell <-> Opteron



SPaSM: Possible code improvements

- **Between now and RR deployment expect:**
 - Improvement of cell computation (reduction of neighbors) (x0.6?)
 - Improvement on AMD side (x0.3?)



Sweep3D Model Input Parameters

- **PAL optimized version of Sweep3D for Cell**
- **Uses domain decomposition (in 2-D)**
 - Each SPE processes a defined subgrid
 - 32 subgrids per triblade
- **A key parameter is the computational block size**
 - Angles per block fixed at 6 (for high SPE compute efficiency)
 - K-planes per block is variable (decreases with scale for high parallel efficiency)

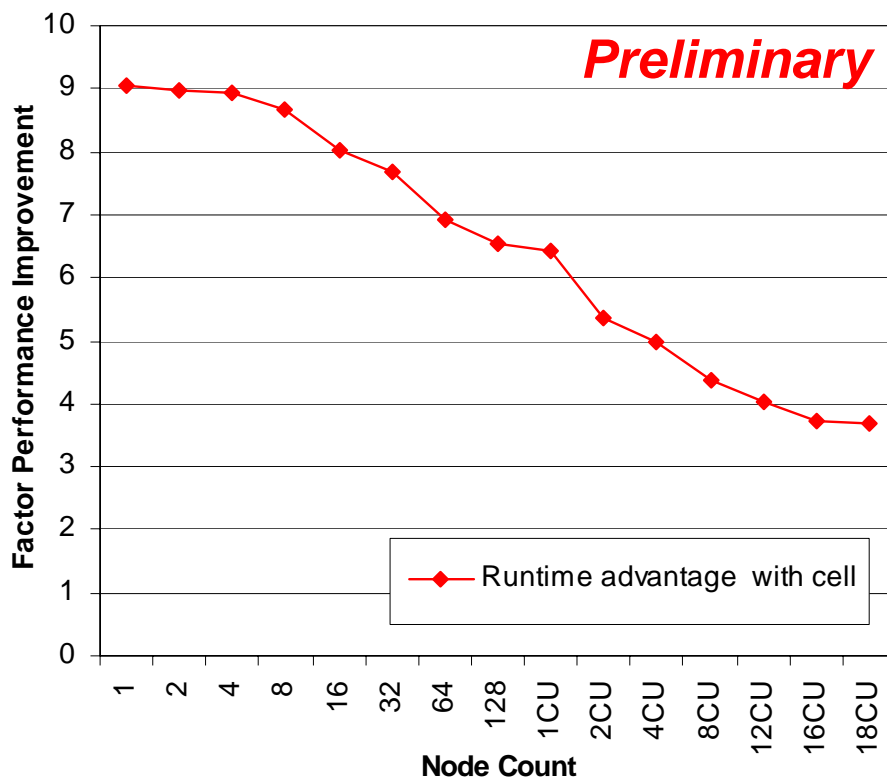
Sub-grid size per SPE (I x J x K)	5x5x400
K-planes per block	{1 .. 50}
Angles per block	6
Number of cycles	10
Grind time per grid-point per angle (eDP) NB variable depending on block-size	{29 .. 47} ns
Boundary surface (Bytes per grid-point per angle)	8

Sweep3D Workload Characteristics

- **Wavefront algorithm**
 - Pipeline characteristic whose length increases with scale
- **Mapping of Sweep3D to the Triblade**
 - Processing
 - » **Cell – SPU: main sweep processing**
 - » **Cell – PPU: DMA and inter-SPE communication management**
 - » **Opteron: No computation**
 - Message Passing: Originate on the Cell and relayed through Opterons
- **Message characteristics**
 - Fine-grained communications:
 - » **2 messages sent per SPE per block per cycle**
 - » **Sizes depend on block size, 240B -> 2,400B (typical)**
- **At small-scale performance is computational bound**
- **At large-scale performance is impacted by both message latency and increased pipeline length**
- **Performance Model validated on all large-scale systems**
- **Model adapted to reflect Cell->Opteron communications**

Sweep3D: RR Performance predictions

Runtime on the base cluster / Runtime on accelerated RR



- **Sweep3D sensitive to latency**
 - Increased due to Cell <-> Opteron
 - But expect some communication to be overlapped

Performance advantage of RR reduces with scale

Sweep3D: Profiling

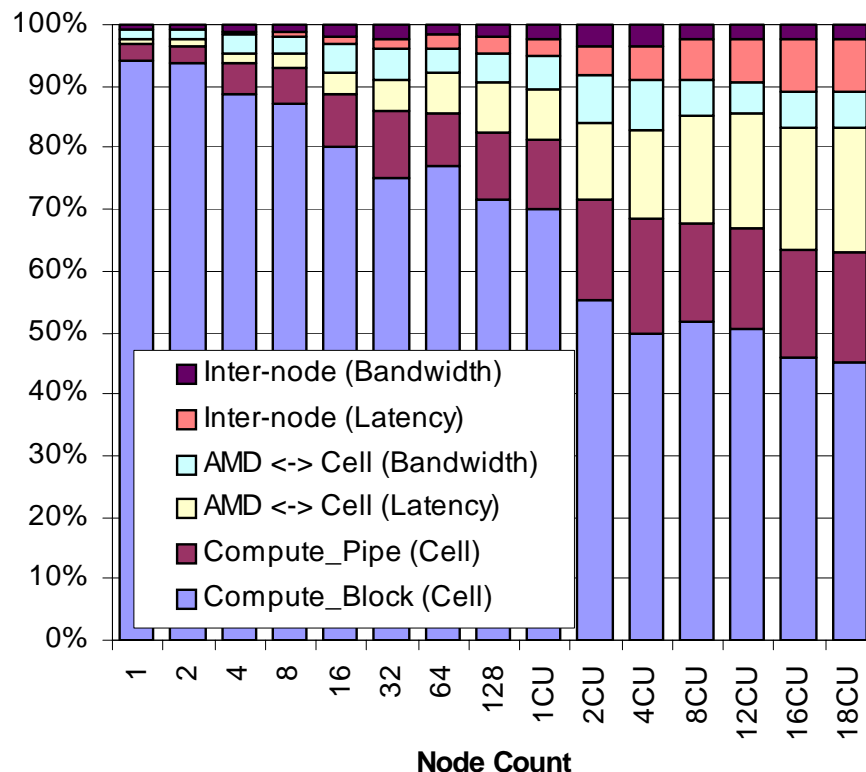
- **Where is the time being spent ?**

- ~63% Compute on Cell
- ~20% Latency (Cell <-> AMD)
- ~5% Bandwidth (Cell <-> AMD)
- ~8% Latency (Infiniband)
- ~3% Bandwidth (Infiniband)

- **Pipeline unavoidable**

- **Latency dominates communication (Cell <-> AMD is major component)**

- **Uses 'probable' HW parameters**



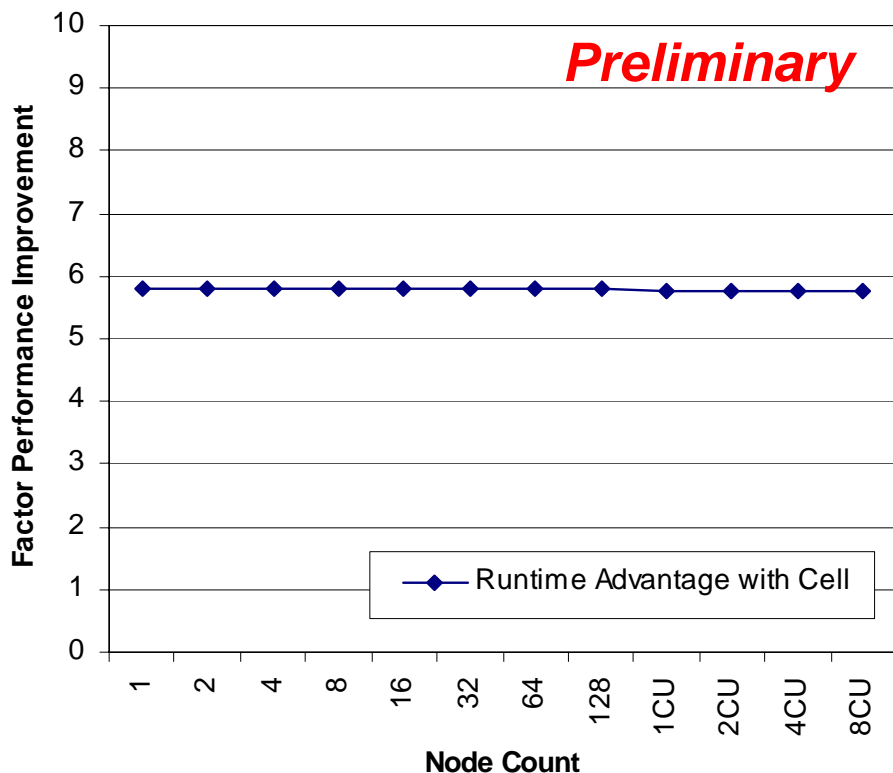
Milagro Model Parameters

- **Double-Bend input-deck**
 - RZwedge geometry
 - Replication: domain is replicated across all processors
- **Iterative**
 - Each processor transports n particles per cycle
 - » **PPE is master, and all work split across SPEs (slaves)**
 - Collectives at end of each iteration to merge tallies & update materials
- **Particles per processor per cycle is a key input**
- **Compute performance based on similar input-deck**

Input-deck	Double-bend
Geometry	RZ-wedge
Particles per cycle / processor	100K .. 500K
Size of collectives	4B, 8B, 21KB, 42KB

Milagro: RR Performance predictions

Runtime on Opteron / Runtime on accelerated RR



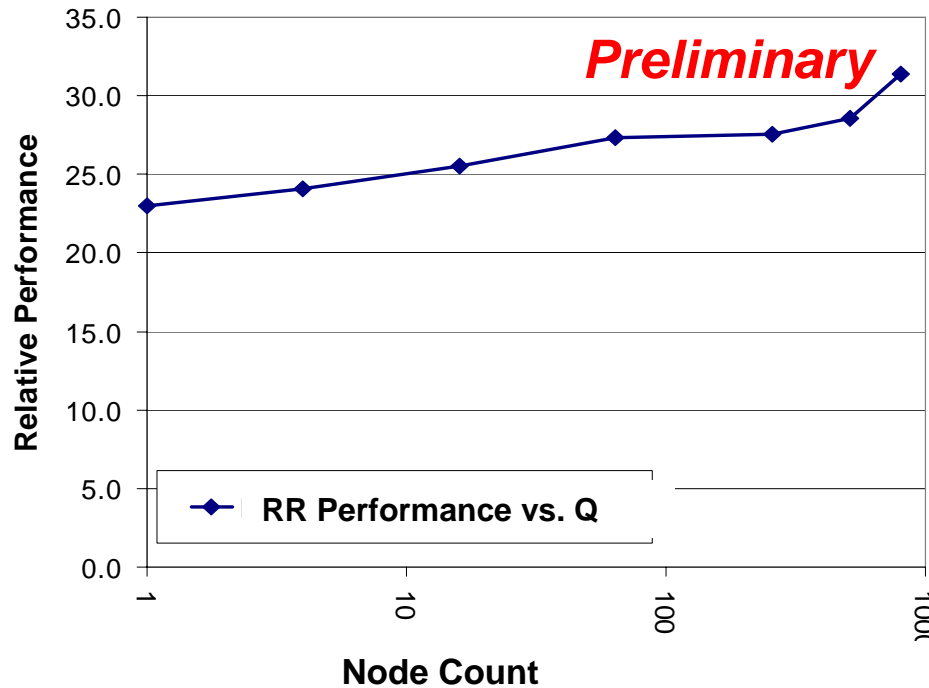
- Very Good scaling expected
- With current code, expect a factor of ~6x better performance using cell

Comparison to ASCI Q

- **Until relatively recently ASCI Q was the largest machine in use at Los Alamos**
- **4-processor (Alpha) EV68 nodes interconnected by Quadrics QNet-1.**
- **Peak speed of 20 Tflops**
- **Comparison made to insert a “historical” perspective in the analysis**

VPIC Performance vs. ASCI Q

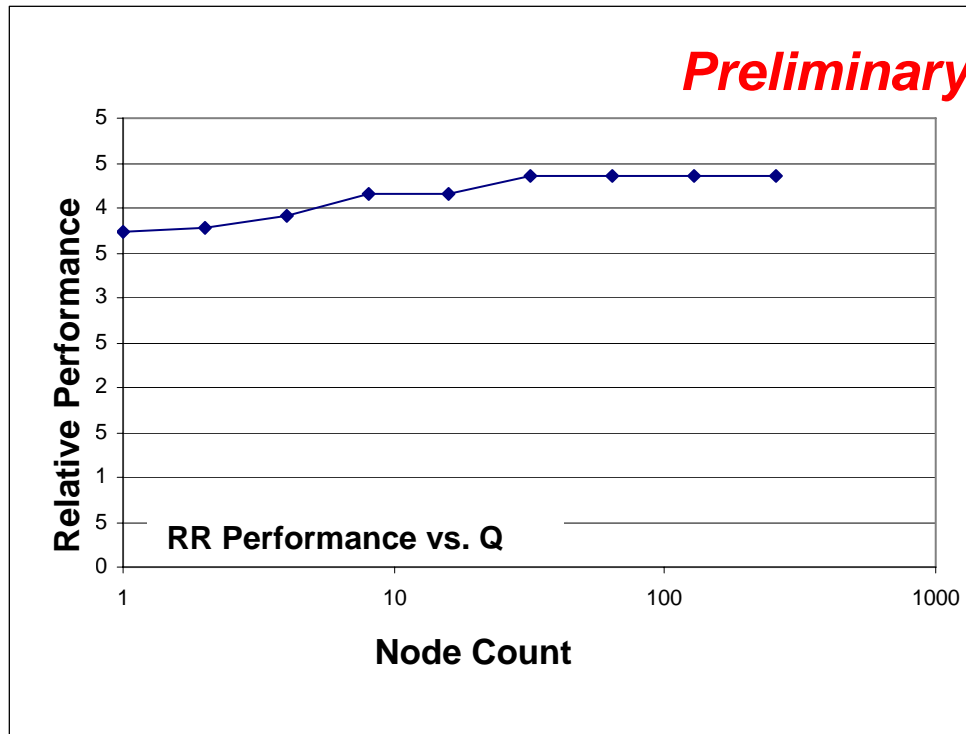
Runtime on Q / Runtime on RR



- Performance on Q is measured up to 3,200 processors
- Expect RR to achieve > 25x higher performance than Q

SPaSM: Performance vs. ASCI Q

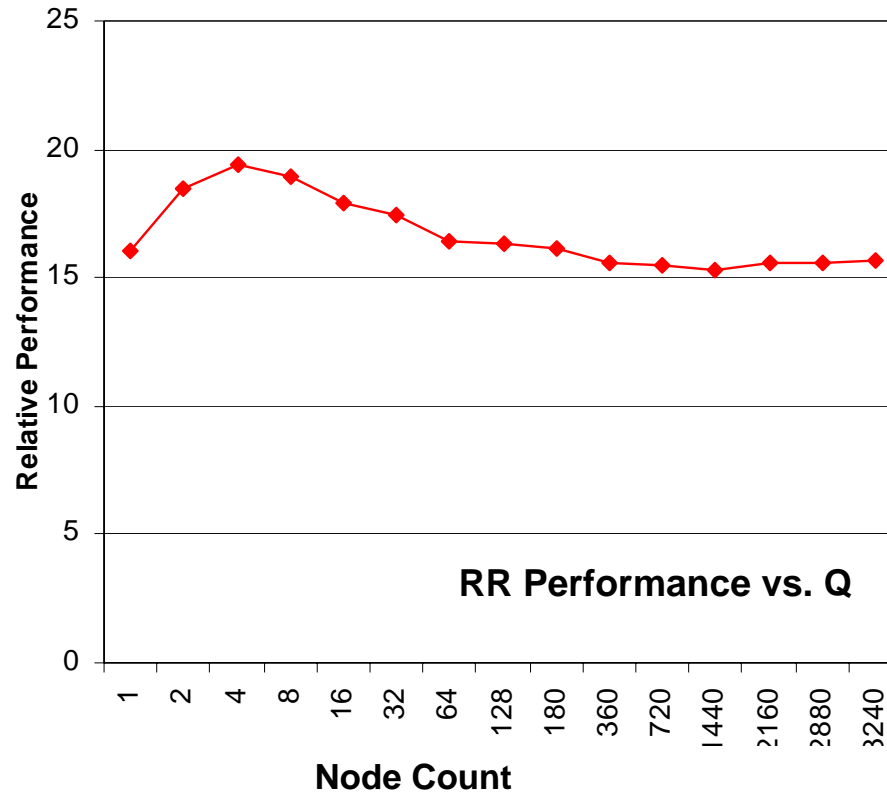
Runtime on Q / Runtime on RR



- Expect RR to achieve ~ 4.5x higher performance than Q

Sweep3D: Performance vs. ASCI Q

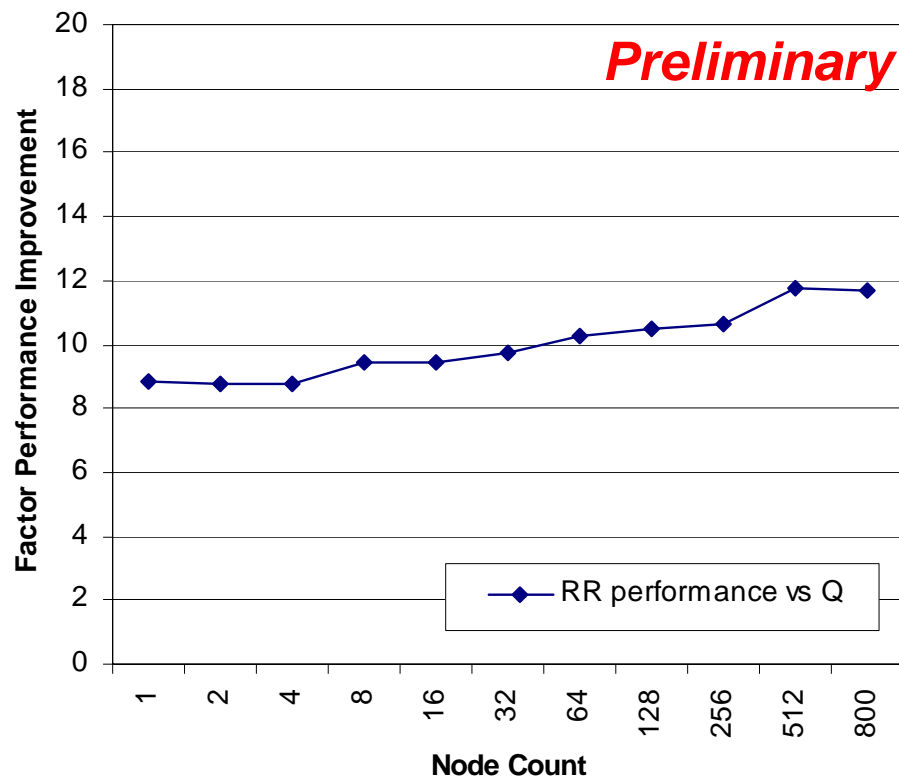
Runtime on Q / Runtime on RR



- Performance on Q is measured up to 800 nodes
- Expect RR to achieve ~15x higher performance than Q

Milagro: Performance vs. ASCI Q

Runtime on Q / Runtime on RR

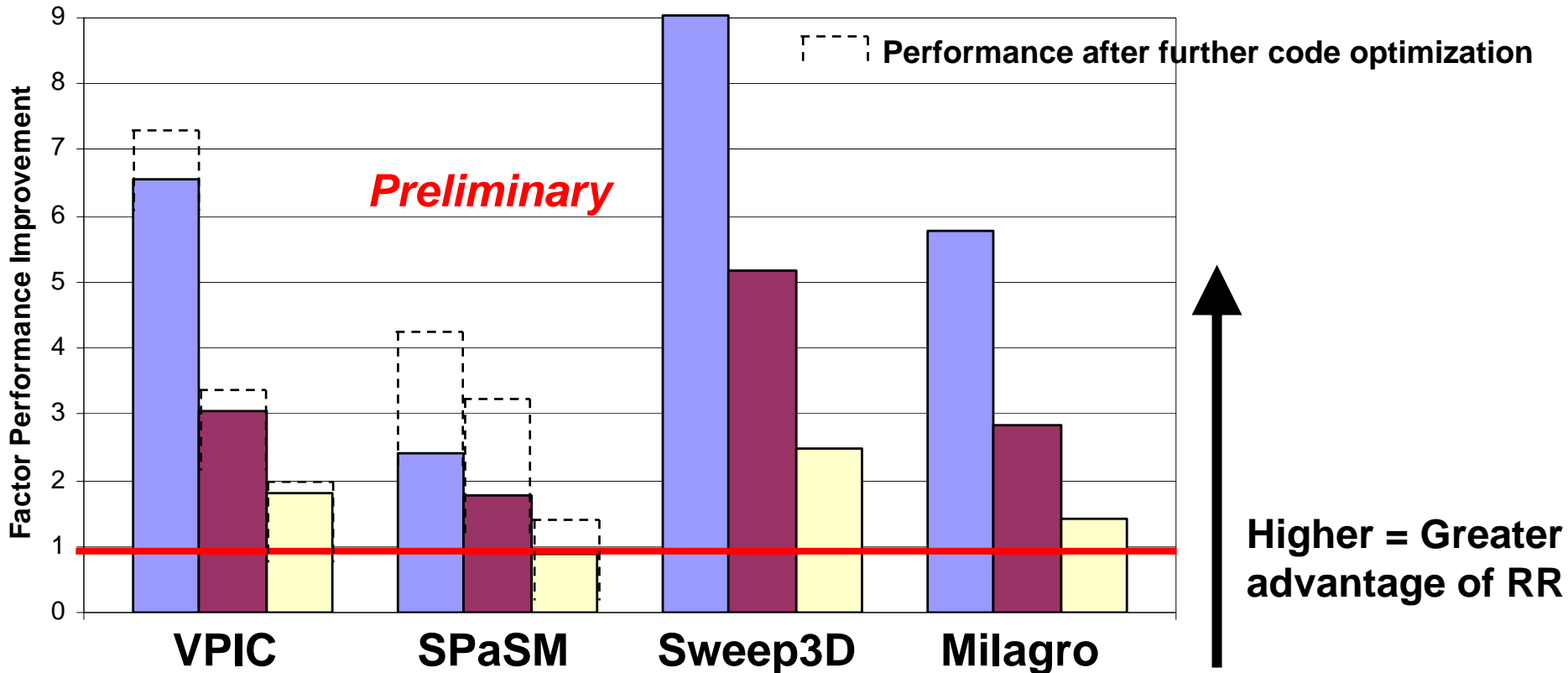


- Expect RR to be 9x-12x higher performance on Milagro than Q

Roadrunner Performance Relative to Hypothetical Systems

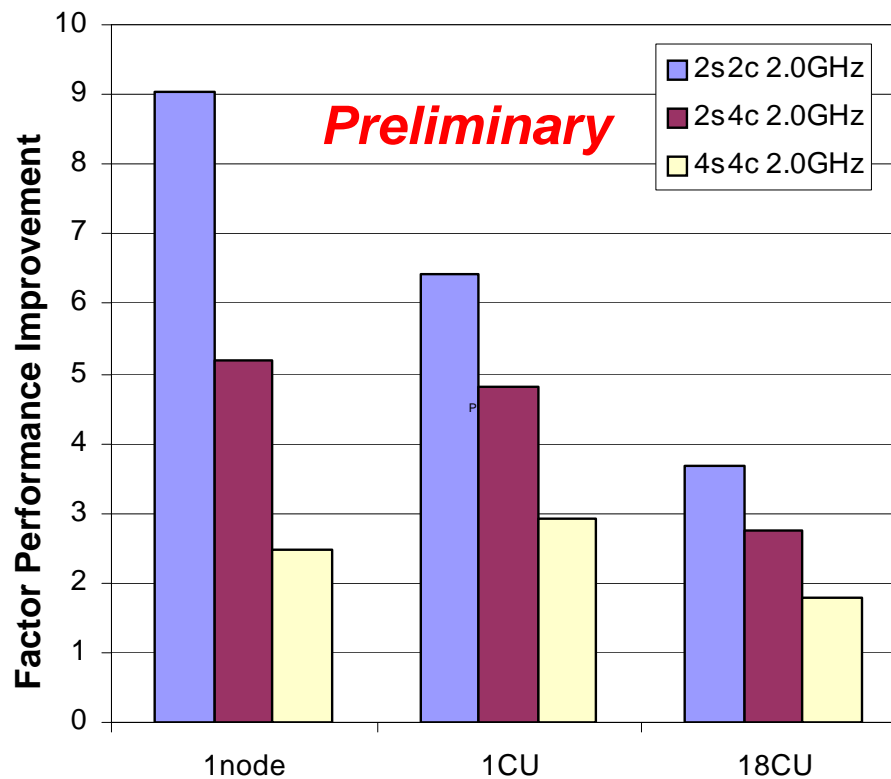
- **Nodes used for comparison:**
 - Triblade (4x cell-eDP, and AMD 2-socket x 2-core)
 - AMD Barcelona 2-socket x 4-core 2GHz)
 - AMD Barcelona 4-socket x 4-core 2GHz)
- **Fixed problem size per node**
 - when comparing node performance
- **Fixed application problem per socket**
 - When comparing core performance (for Barcelona)

Single Node 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

System Performance Comparison for Sweep3D



What's Next in RR Performance Work

- **Expansion of the modeling work to new applications of interest**
- **Application optimization steered by models**
- **Roadrunner system optimization**
- **Continuous monitoring of system performance during and after installation**
- **Performance “acceptance testing”**

Summary

- We have analyzed performance of RR under a realistic application workload of interest through predictive modeling
- VPIC, SPaSM and Sweep3D scale well on RR
- VPIC, SPaSM, Sweep3D exhibit high performance gains over the RR base cluster, in the range of 2.5-7
- Significant performance improvements over ASC Q were observed, between factors of 5-25
- Applications under consideration are faster on RR than on hypothetical systems using state-of-the-art multicore nodes