COMPUTER & COMPUTATIONAL SCIENCES

A Performance Analysis and Comparison of Roadrunner, Blue Gene/P, and an AMD Barcelona/InfiniBand Cluster

LACSS 2009

Kevin J. Barker, Kei Davis, Adolfy Hoisie, Darren J. Kerbyson, Mike 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

Los Alamos NATIONAL LABORATORY

Introduction and Motivation

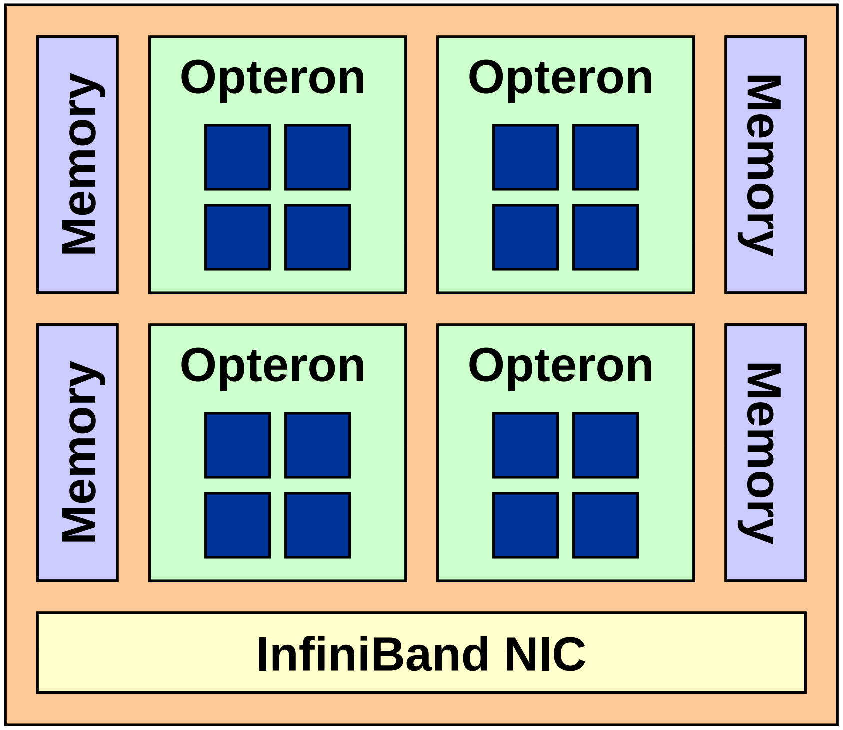
" 20% of a project's time is spent in trying to understand what to build, 80% is spent building it, and no time is spent trying to understand deeply, how well the design decisions were made in terms of performance delivered to users, and hence, how to proceed on the next system design."

- *David Kuck,*
  Kuck & Associates, Inc. and
  Univ. of Illinois, Emeritus
  "High-Performance Computing"
  Oxford U. Press, 1996

Systems Under Consideration

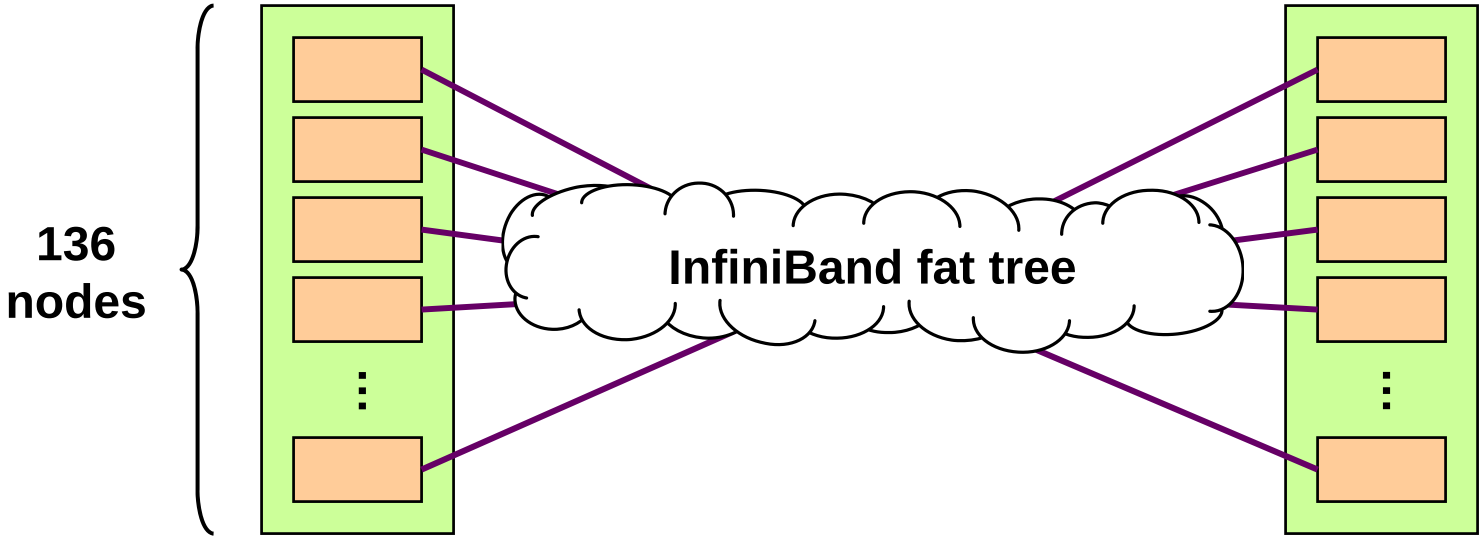
- **Lobo: Conventional cluster**
  - Commodity processors and network
- **Dawn: Traditional massively parallel processor**
  - Second-generation Blue Gene (Blue Gene/P)
  - Specially modified processors, custom networks
  - *Pros*: abundant parallelism, low-latency communication
  - *Cons*: weak processor cores, limited bandwidth
- **Roadrunner: Hybrid, accelerated cluster**
  - Commodity processors and network plus enhanced commodity processors as accelerators
  - *Pros*: immense peak performance per node, abundant parallelism
  - *Cons*: severely unbalanced communication/computation performance (few GB/s per flop/s) $\rightarrow$ significant NIC contention

Lobo Node Architecture



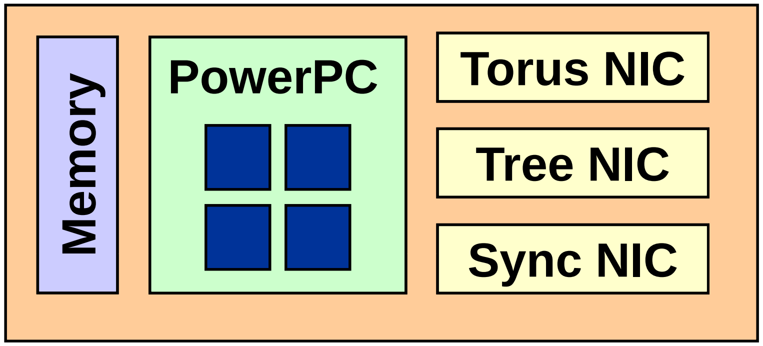
- **Quad-socket, quad-core CPUs**
  - AMD Barcelona 8354 @ 2.2 GHz
- **32 GB of memory per node**
  - 2 GB/core

Lobo System Architecture



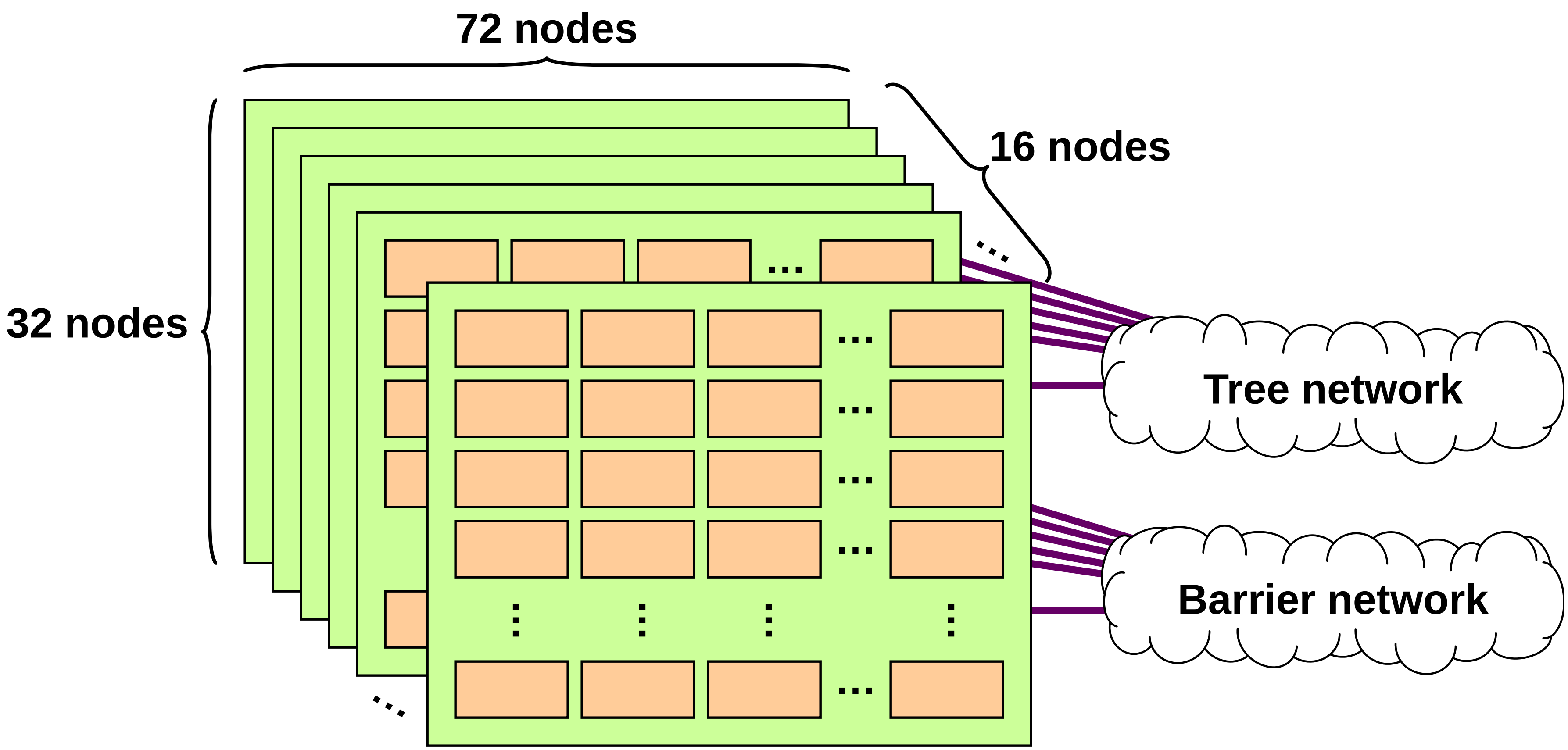
- 2 SUs × 136 nodes/SU × 4 sockets/node × 4 cores/socket = 4,352 cores (38.3 peak Tflop/s)
- 4x DDR InfiniBand (2 GB/s per link per direction)
- One 288-port InfiniBand switch

Dawn Node Architecture



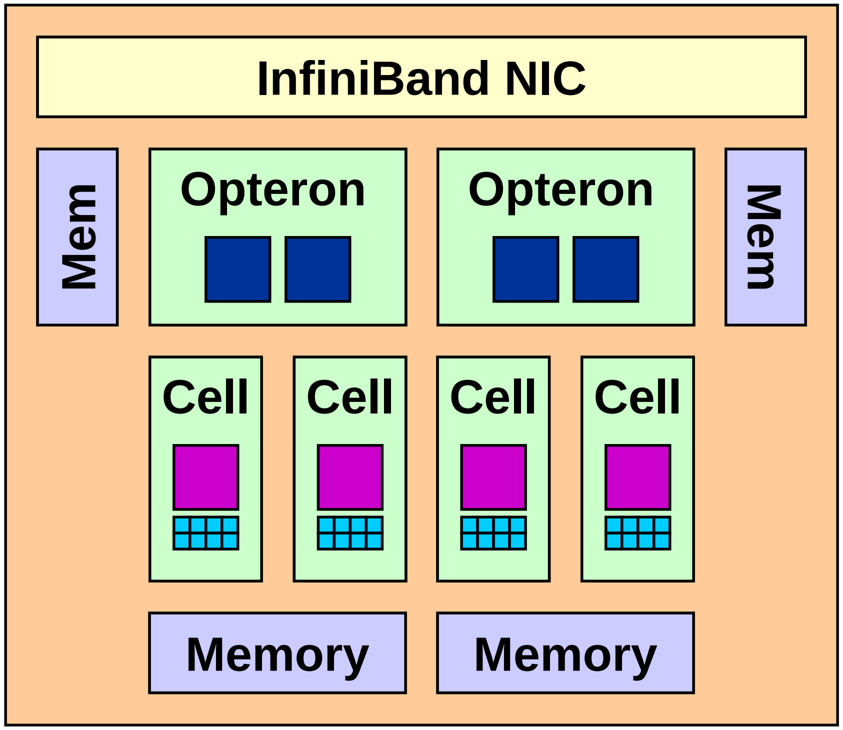
- **Single-socket, quad-core CPUs**
  - PowerPC 450d @ 850 MHz
- **4 GB of memory per node**
  - 1 GB/core

Dawn System Architecture



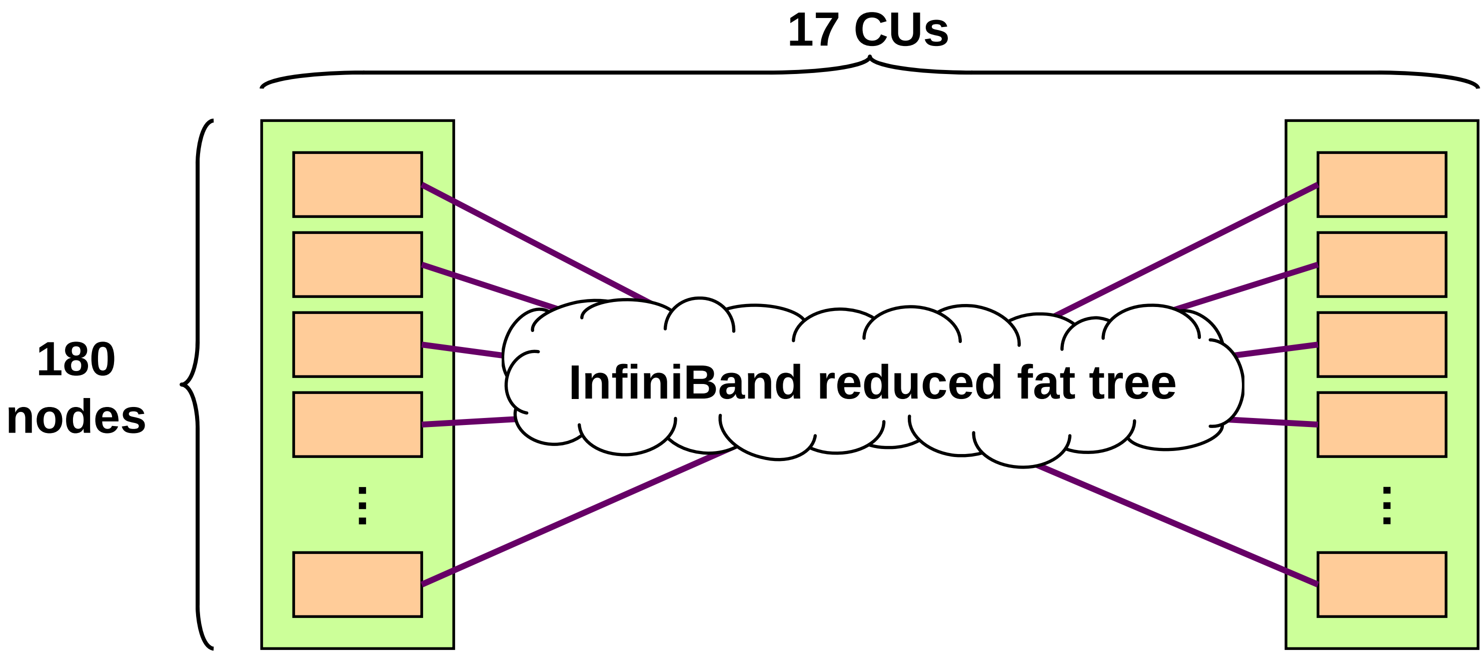
- 72 × 32 × 16 nodes × 4 cores/node = 147,456 cores (501.3 Tflop/s)
- 425 MB/s per torus link per direction × 6 links/node = 2.6 GB/s per direction per node

Roadrunner Node Architecture



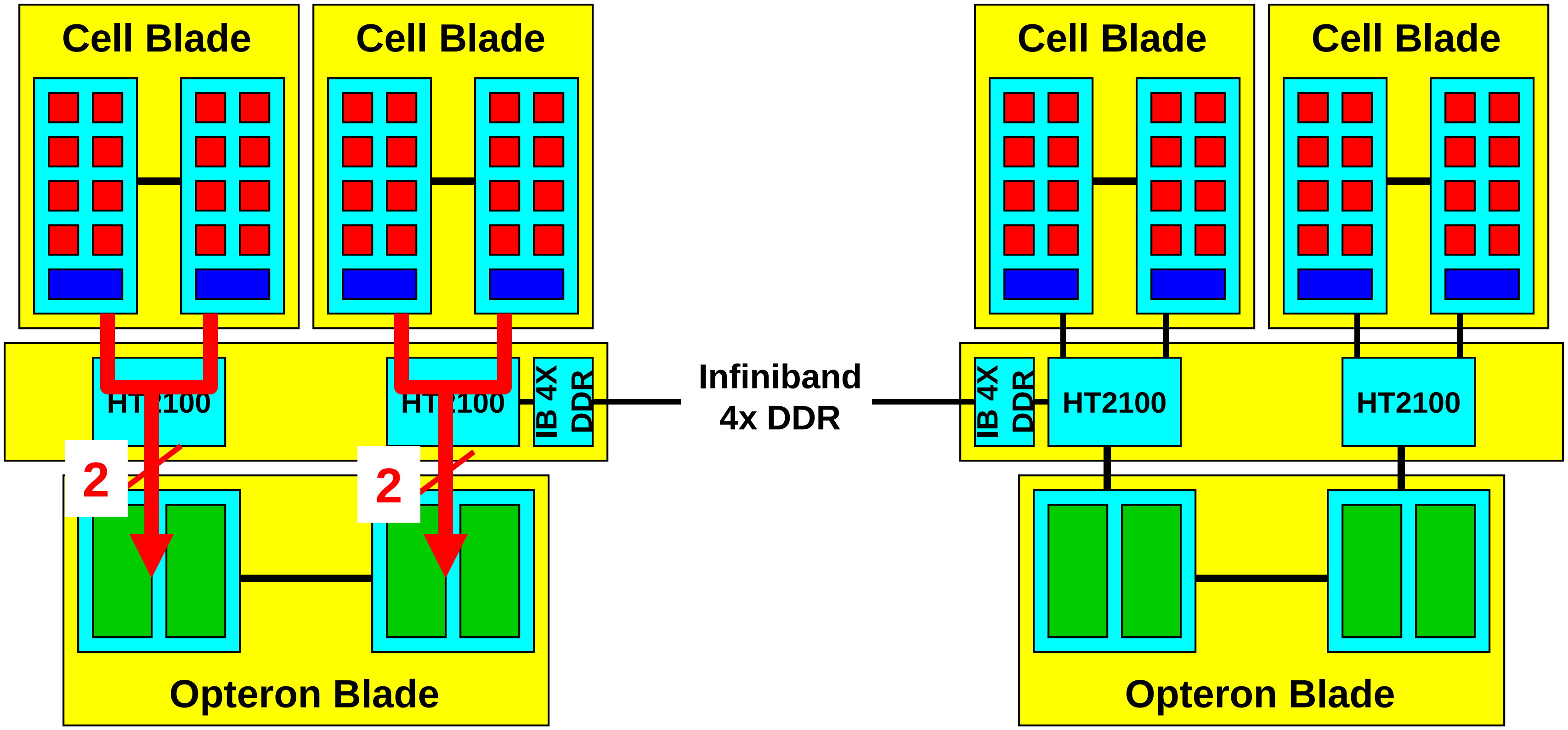
- **Dual-socket, dual-core CPUs**
  - AMD Opteron 2210 @ 1.8 GHz
- **4 Cell/B.E. accelerators (one per CPU core)**
  - PowerXCell 8i @ 3.2 GHz
- **32 GB of memory per node**
  - 4 GB/Opteron core + 4 GB/Cell socket

Roadrunner System Architecture



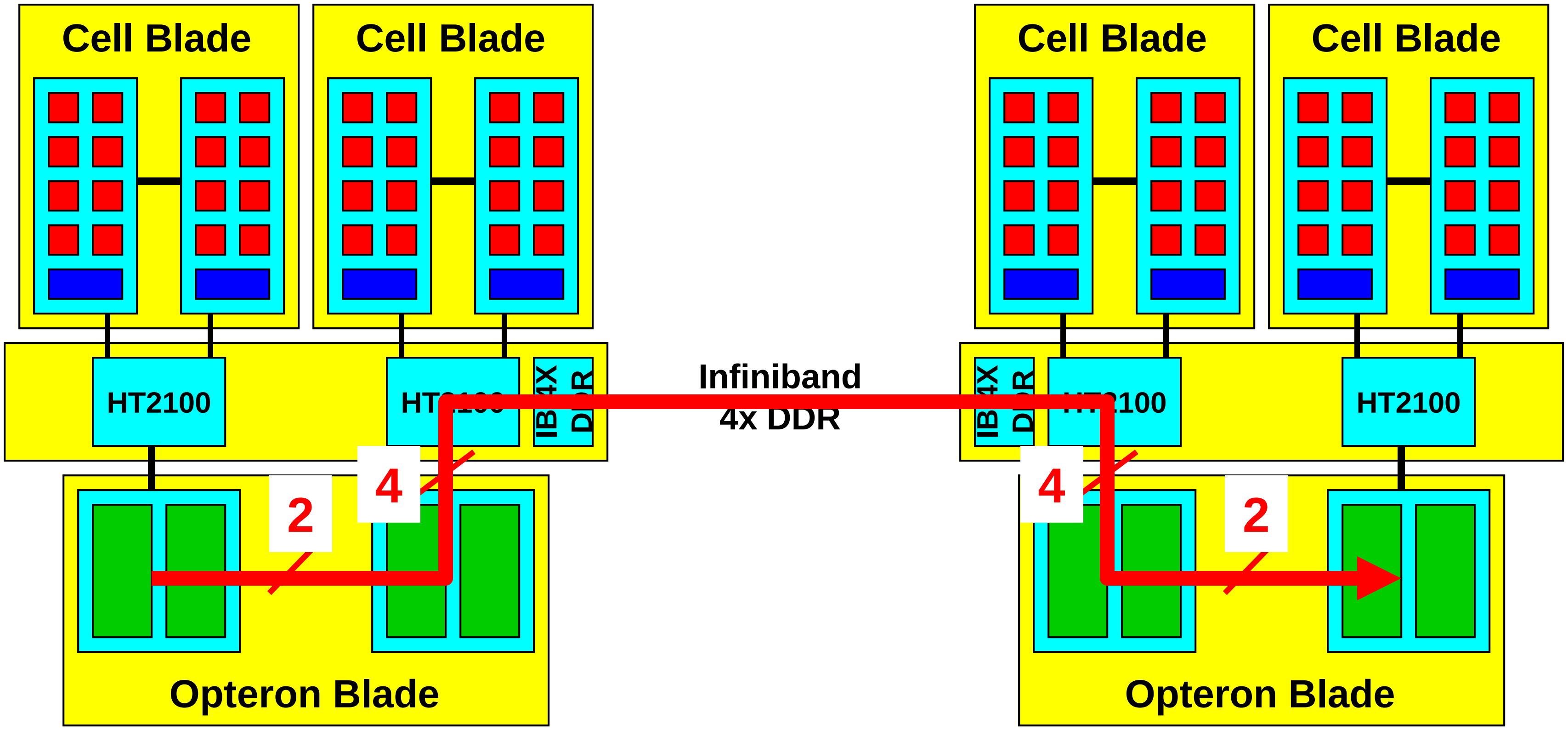
- 17 CUs × 180 nodes/CU × {2,4} sockets/node × {2,9} cores/socket = 122,400 cores (1,393 peak Tflop/s)
- 4x DDR InfiniBand (2 GB/s per link per direction)
- 2 levels of InfiniBand (intra- and inter-CU)

Example: Communication steps



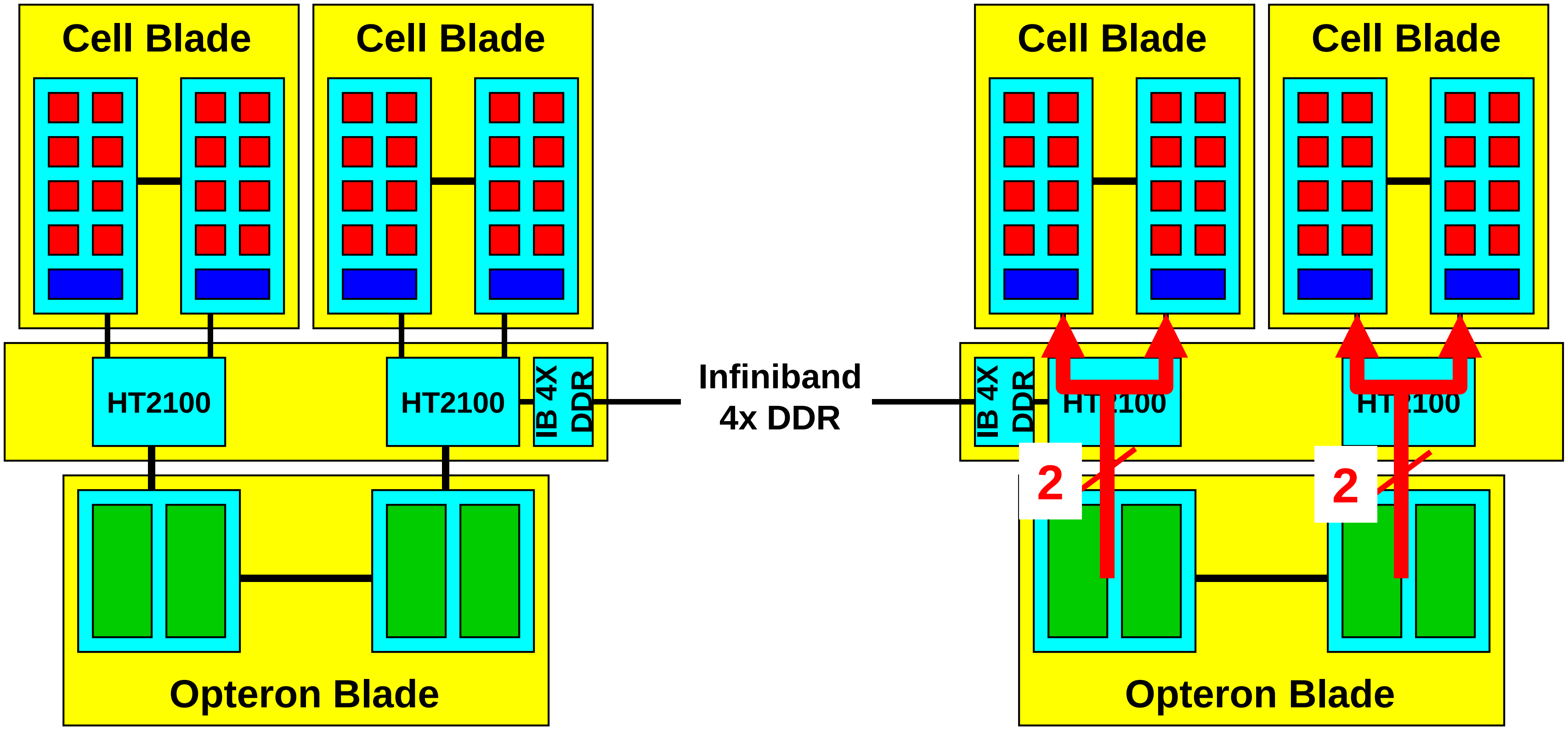
1) Cells (Node 1) -> Opterons (Node 1)
2) Opterons (Node 1) -> Opterons (Node 2)
3) Opterons (Node 2) -> Cells (Node 2)

Example: Communication steps



1) Cells (Node 1) -> Opterons (Node 1)
2) Opterons (Node 1) -> Opterons (Node 2)
3) Opterons (Node 2) -> Cells (Node 2)

Example: Communication steps



1) Cells (Node 1) -> Opterons (Node 1)
2) Opterons (Node 1) -> Opterons (Node 2)
3) Opterons (Node 2) -> Cells (Node 2)

Data Movement Performance Characteristics of Roadrunner: 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**

Summary of Architectural Characteristics

Feature	Lobo	Dawn	RR
Cores/node	16	4	40
Nodes/system	272	36,864	3,060
Cores/system	4,352	147,456	122,400
Memory/node (GB)	32	4	32
Streams mem. BW/socket (GB/s)	7.4	10.0	22.2
Streams mem. BW/node (GB/s)	18.8	10.0	88.9
Network BW/node/dir. (GB/s)	2	2.5 (÷6)	2
Peak performance (Tflop/s)	38	501	1,393 (44 Base)

No one system is clearly superior → use performance models to compare

What is a Performance Model?

- **Analytical expression of performance in terms of application and system characteristics**
  - May be embodied as mathematical formulas, Excel spreadsheets, Perl scripts, etc. (It doesn't matter.)
- **Precise description of an application in terms of system resources**
  - Which resources substantially determine execution time?

    CPU speed/core count, network latency/bandwidth/topology, memory hierarchy sizes/speeds, …
  - When is each resource used?

    during an iteration, between iterations, every nth iteration, …
  - What determines how much each resource is used?

    processor count, memory capacity, physics modules included, …

Large-Scale System Comparison

- **Performance models can be used to compare the performance of large systems**
  - Measurement is not always possible
    - Access may be limited
    - Systems may not yet be available (e.g., in the procurement of a future system)
  - Predict and compare performance of a workload on a set of systems
  - Determine the system characteristics that most limit performance
- **We compared performance of three supercomputers on a realistic workload combining benchmarking and modeling**
  - For this short talk, only Sweep3D (hybrid on RR) and SAGE are presented.

Case Studies

Two case studies chosen from many applications that have been modeled

1) Sweep3D

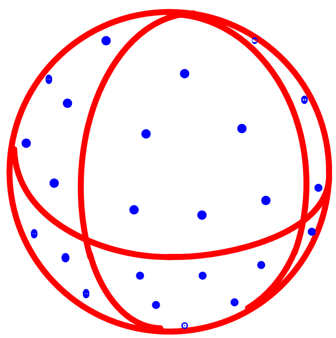
- Deterministic S_N Transport
  - Structured mesh
  - 2-D data decomposition
  - Pipelined wavefront processing

2) SAGE

- Hydrodynamics code
  - Structured Adaptive mesh
  - 1-D data decomposition

Case Study I: S_N Transport

- **Solve the particle transport equation, where the density distribution of particles N(<u>x</u>, E, Ω, t) is the unknown**
- **Use discrete directions Ω**
  - S_N has N*(N+2) total directions spread out in 3-dimensions
  - e.g., S_6 has 48 total directions, or 6 directions per octant



- **SWEEP3D code: 1-group, Cartesian-grid kernel**
  (http://www.c3.lanl.gov/par_arch/Software.html)

"Performance and Scalability Analysis of Teraflop-Scale Parallel Architectures Using Multidimensional Wavefront Applications", A. Hoisie, O. Lubeck, H. Wasserman, The Int. J. of High Performance Computing Applications, Sage Science Press, 14(4), Winter 2000

Sweep3D Workload Characteristics

- **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 → 4,800B (typical)
- **At small scale, performance is compute-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 additional Cell → AMD communications**

Case Study II: Hydrodynamics

- **SAGE – SAIC's Adaptive Grid Eulerian hydrocode**
- **Hydrodynamics code with Adaptive Mesh Refinement (AMR)**
- **Applied to: water shock, energy coupling, hydro instability problems, etc.**
- **Represents a large class of production ASC cycles at LANL**
- **Routinely run on 1,000s of processors**
- **Scaling characteristic: Weak**
- **Data Decomposition (Default): 1-D (of a 3-D AMR spatial grid)**

"Predictive Performance and Scalability Modeling of a Large-Scale Application", D.J. Kerbyson, H.J. Alme, A. Hoisie, F. Petrini, H.J. Wasserman, M. Gittings, in Proc. SC, Denver, 2001

Model Accuracy

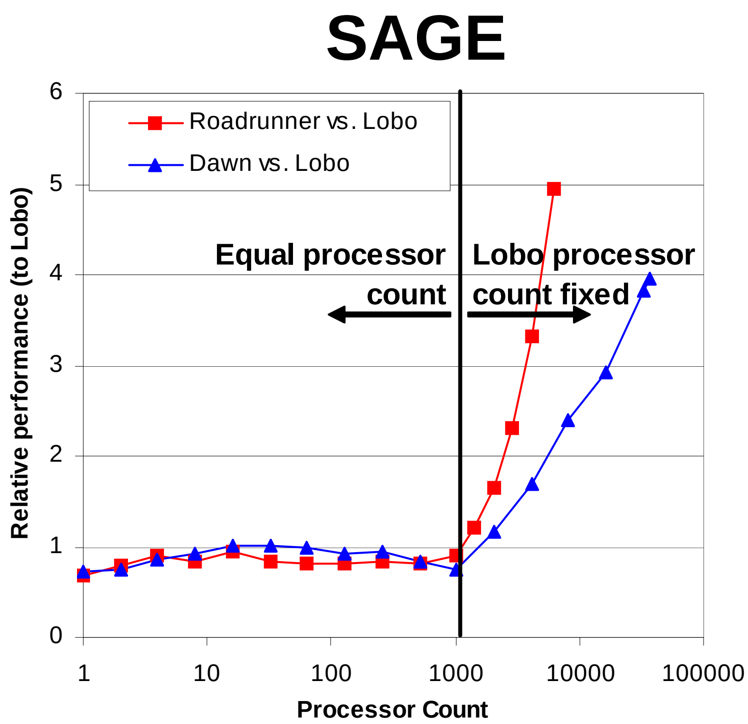
- **Maximum modeled error excluding outlying "rogue" points**

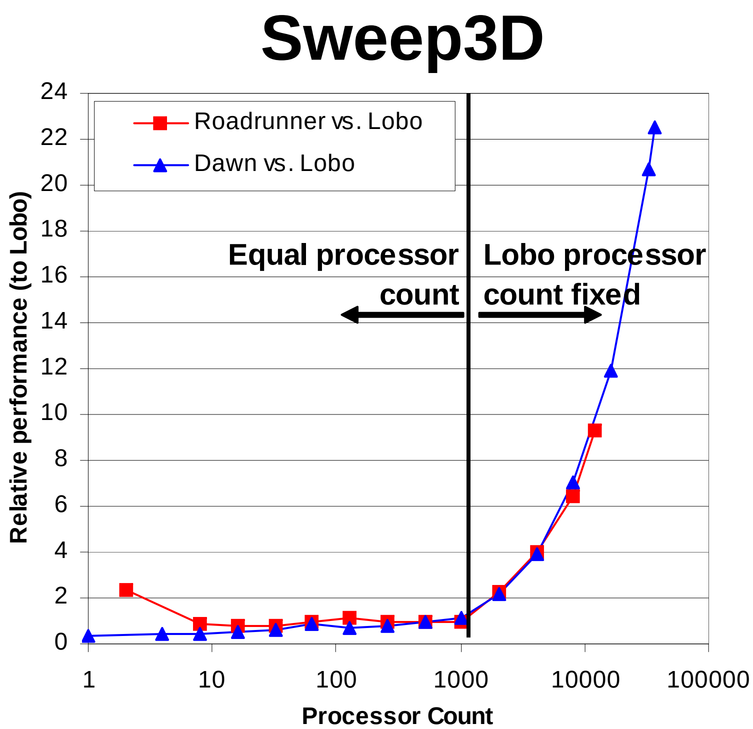
	Lobo	**Dawn**	**Roadrunner**	
SAGE	< 7%	< 10%	< 4%	
Sweep3D	< 14%	< 4%	< 8%	**Non-Hybrid**
			< 11%	**Hybrid**

FYI, two other applications we also looked at:

VPIC	< 6%	< 1%	< 4%	**Non-Hybrid**
			< 8%	**Hybrid**
Partisn	< 6%	< 12%	< 4%	

Measuring Application Performance





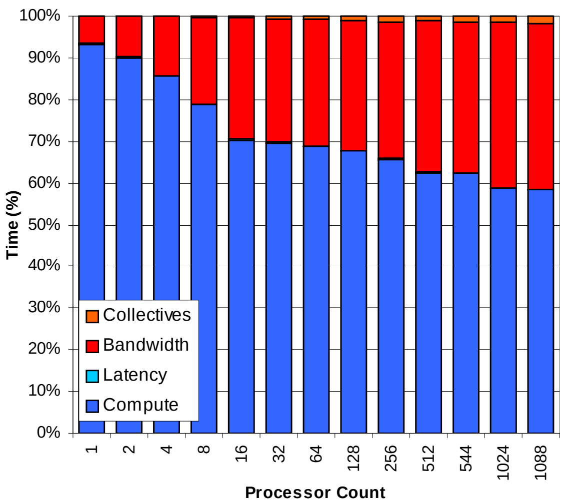
- **Roadrunner Base > Dawn on SAGE**
- **Dawn > Roadrunner Hybrid on Sweep3D**
- **Can we use modeling to explain this discrepancy?**

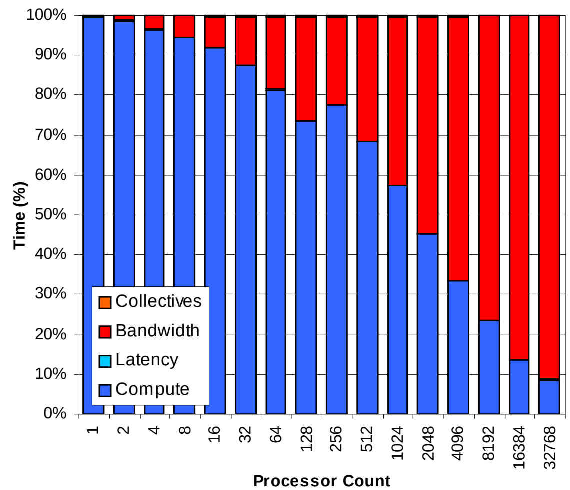
Los Alamos
NATIONAL LABORATORY
EST. 1943

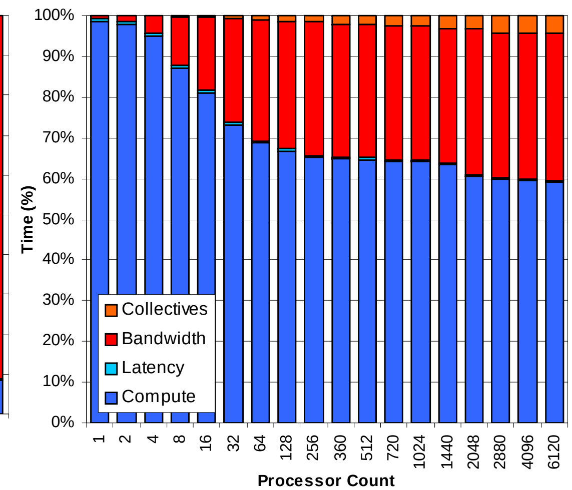
Using Modeling to Identify Performance Bottlenecks

Lobo | **Dawn** | **Roadrunner**

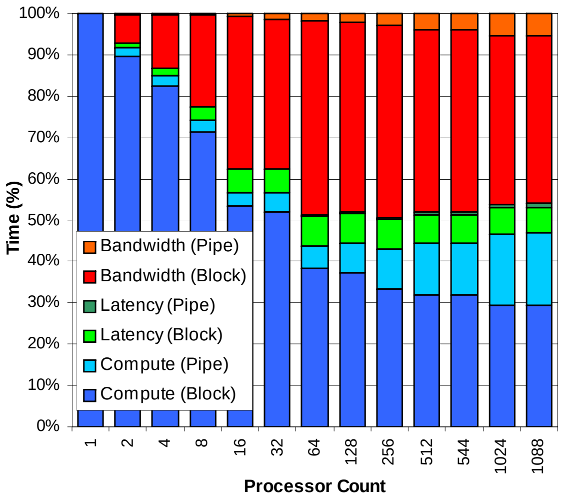
SAGE

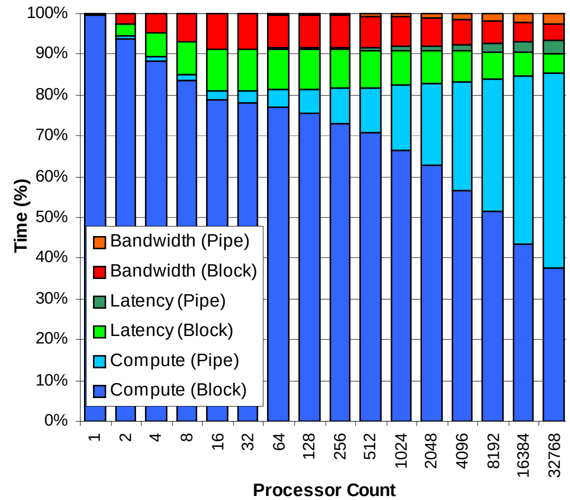


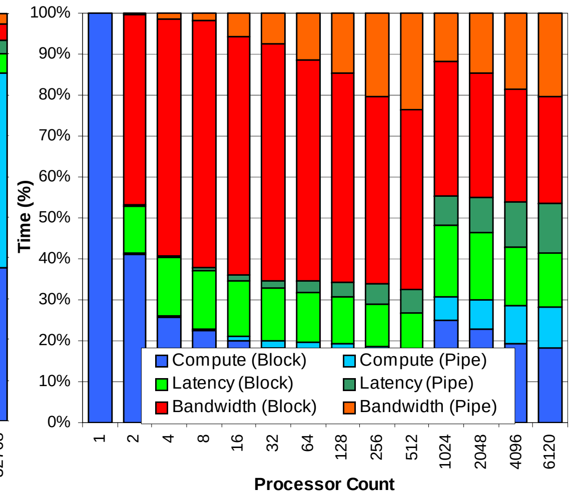




Sweep3D





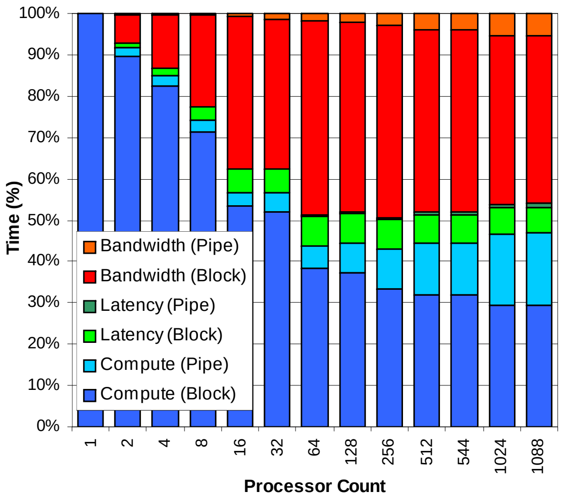


Using Modeling to Identify Performance Bottlenecks

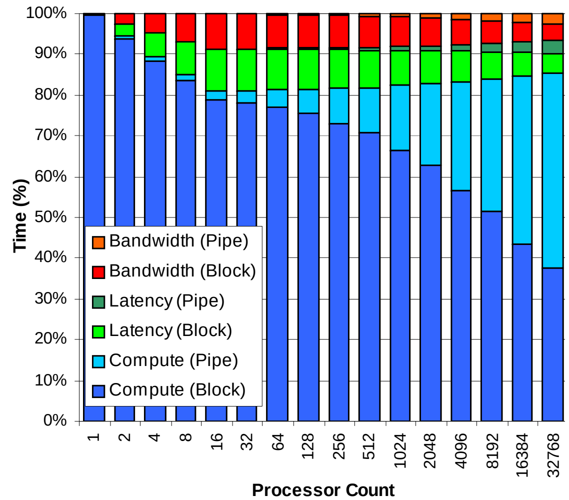
- **Sweep3D transmits a large number of small/medium-sized messages; also, pipeline effects limit parallel efficiency**
- **Would expect latency to dominate; in fact,**
  - Few networks are bandwidth-optimized for Sweep3D's message sizes
  - Lobo is 50-50 compute/bandwidth due to NIC contention (16 procs)
  - Dawn spends 50% of its time stalled waiting for data (pipeline effects)
  - Roadrunner required different blocking at 2K procs; data aggregation helped with pipelining effects, but deep comm. hierarchy hurts perf.

Sweep3D

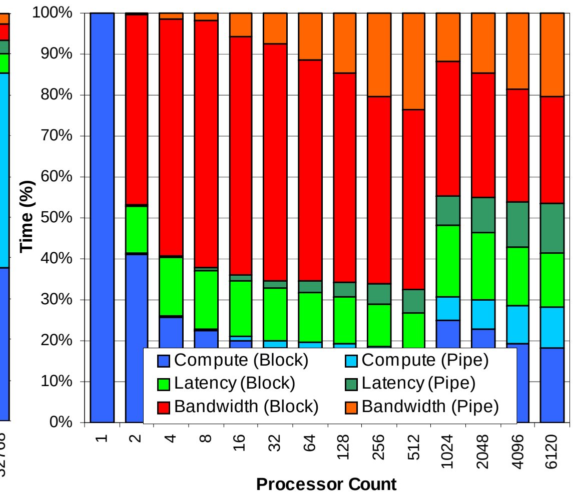
Lobo



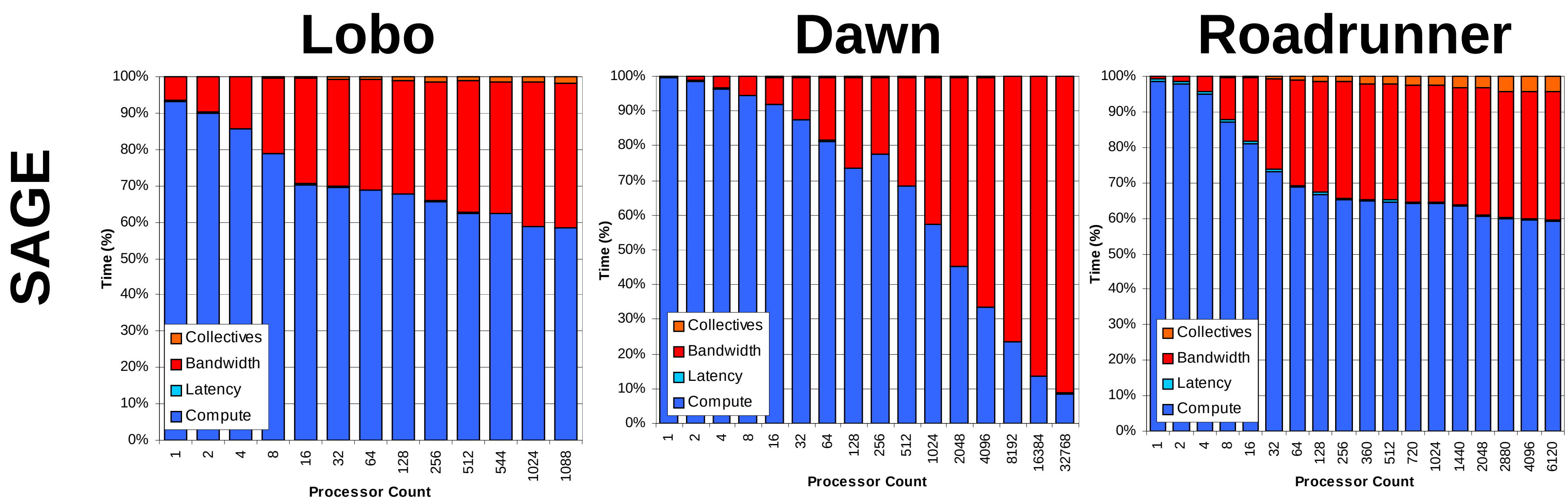
Dawn



Roadrunner



Using Modeling to Identify Performance Bottlenecks

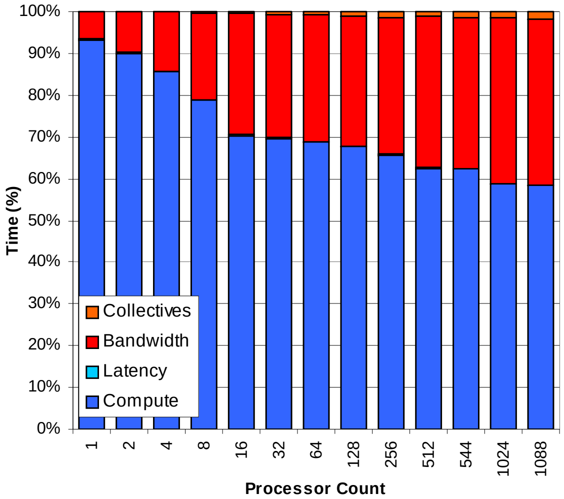


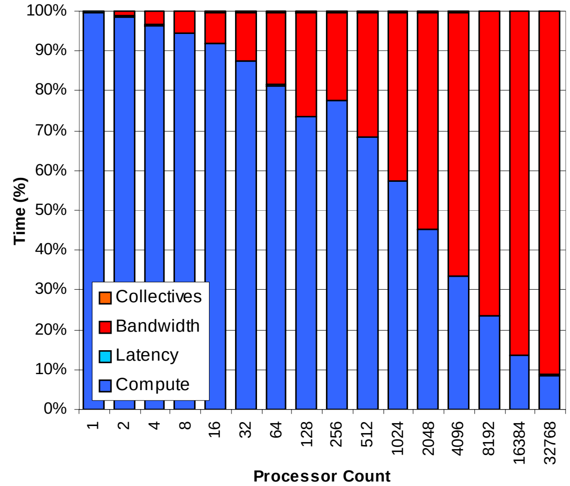
- **SAGE transmits a large volume of large messages**
- **Lobo and Roadrunner Base (same IB fat-tree network) gradually lose performance to bandwidth**
- **Dawn's limited link bandwidth and susceptibility to network contention in the torus rapidly let bandwidth dominate performance**

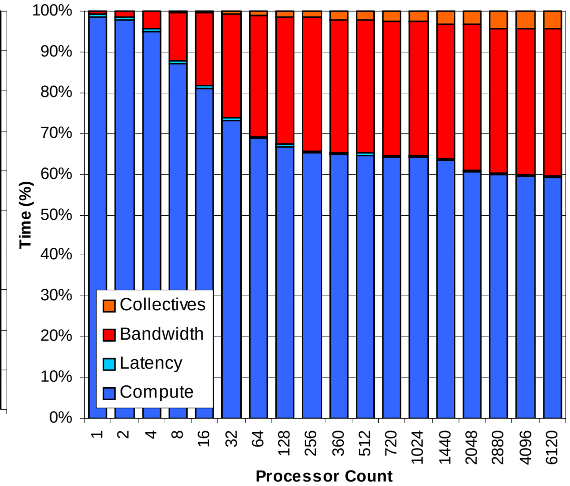
Using Modeling to Identify Performance Bottlenecks

Lobo | Dawn | Roadrunner

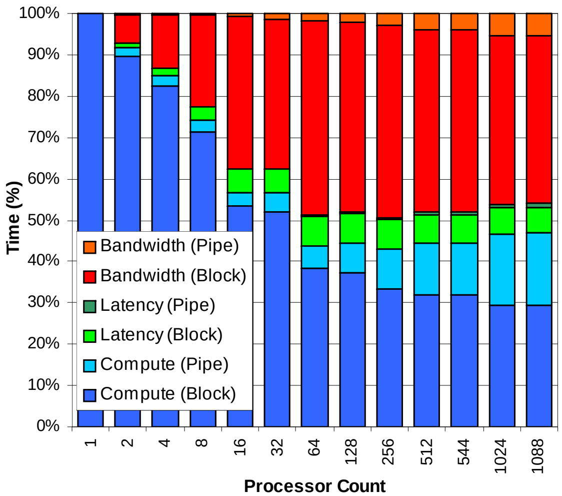
SAGE

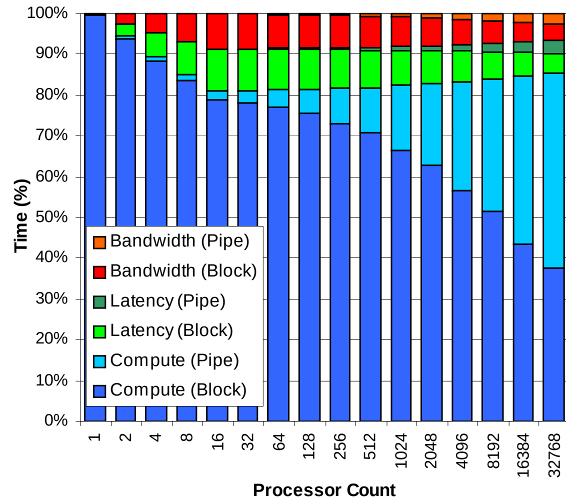


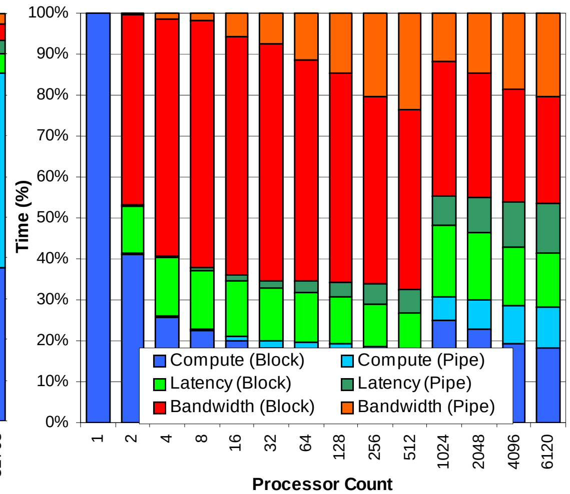




Sweep3D







Summary

- **Performance is workload-dependent**
- **Different systems → different bottlenecks**
  - SAGE is compute-bound on Lobo and Roadrunner Base but bandwidth-bound on Dawn
  - Sweep3D is compute-bound on Dawn and Roadunner Base but communication bound on Roadrunner Hybrid and 50-50 compute/communicate on Lobo
- **Different applications → different bottlenecks**
  - Dawn is bandwidth-bound on SAGE but compute-bound on Sweep3D
- **Modeling can help explain performance measurements**
  - Dawn has more processors than Roadrunner Base, but Roadrunner Base is faster on SAGE
    - » Model shows Dawn's relatively poor bandwidth limits its performance
  - Roadrunner Hybrid has higher per-node peak than Dawn, but Dawn is faster on Sweep3D
    - » Model shows Roadrunner Hybrid is bottlenecked by communication

Summary

- **Performance models are useful tools for exploring system performance at all stages of development**
  - Predicting performance during procurement/assessment
    - Comparing performance of hypothetical machines (impossible to do empirically!)
  - Validating performance during installation
  - Monitoring performance during system upgrades
- **Performance models are equally useful for understanding how software changes will impact performance**