To Adapt, or Not to Adapt, That is THE Question

Matt Leininger (SNL/LLNL)
Mark Seager (LLNL)

OpenFabrics Developers Workshop
Sonoma, CA
April 30, 2007

Outline

- Short History of NNSA ASC and InfiniBand
- What is a Multi-Physics Application?
- Some Recent Examples of App Performance & Scalability
- Adaptive Routing
- Conclusions and Future Directions

NNSA ASC Tri-Lab Linux Capacity Cluster (TLCC07) Procurement

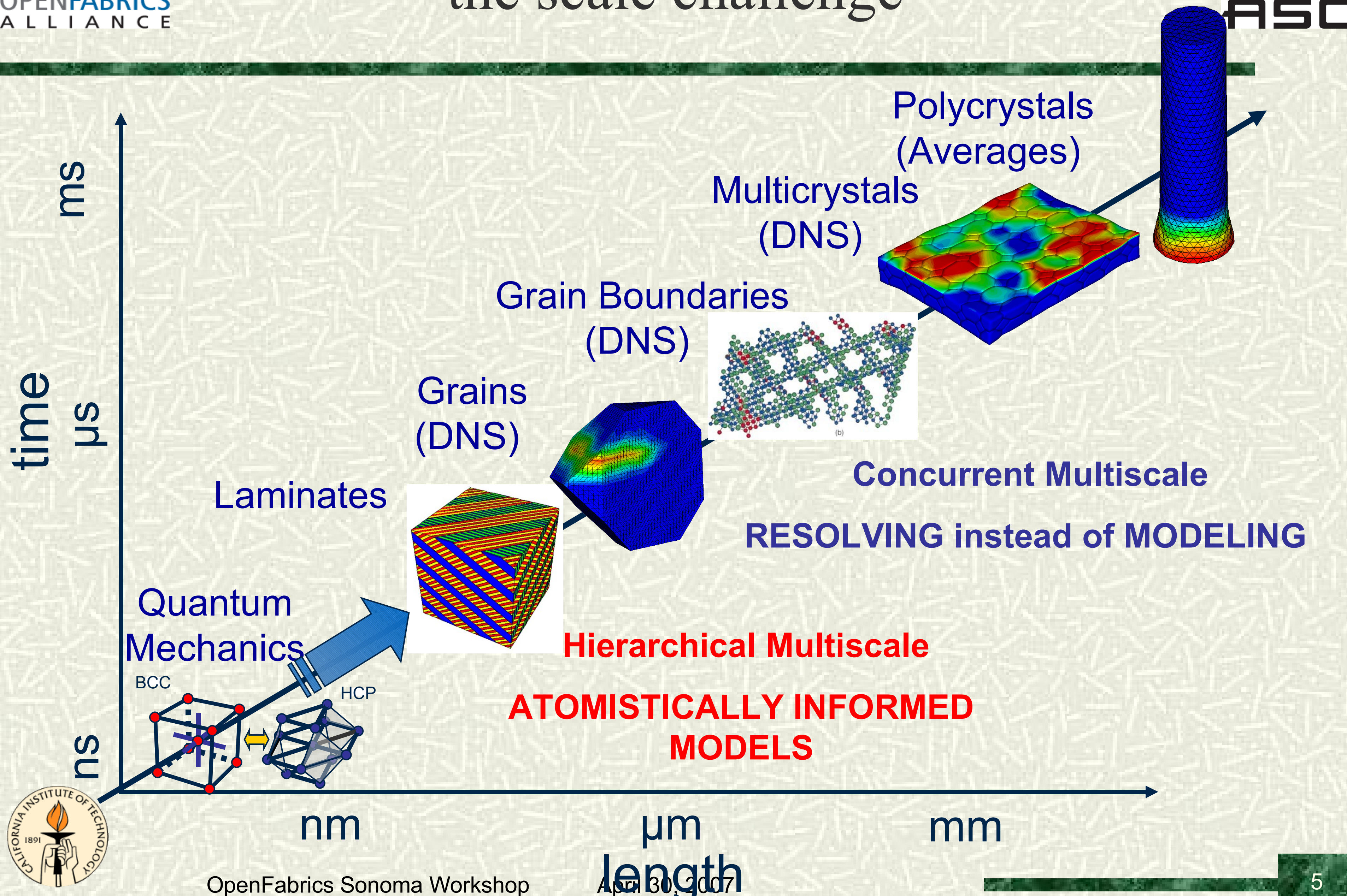
- Reduced Total Cost of Ownership
- Establish Foundation for Unified HW/SW environment among Tri-Labs
- ROBUST PRODUCTION CAPACITY
  - 100's of Tflops for capacity computing
- $30-45M over 4 Quarters
  - Clusters delivered to all three labs (LLNL, LANL, and SNL)
- Design based on Scalable Unit (SU) concept
  - SU's are cluster building blocks for 288, 576, 864, and 1,152 node systems

DRAFT RFP released
www.llnl.gov/asc/tlcc/rfp

NNSA ASC Has Long History with InfiniBand/OpenFabrics

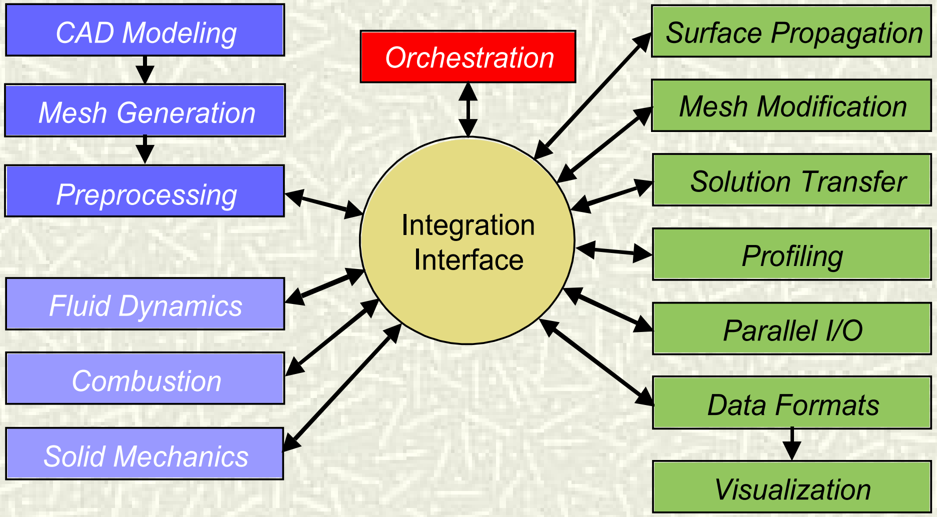
- ASC = Advanced Simulation Computing Program
- LLNL, LANL, and SNL
- Started Sonoma InfiniBand Workshop in 2003
- Built first 128, 256, and 4,500 node InfiniBand clusters
- Founding members of OpenIB → OpenFabrics
- Funding OpenFabrics Linux HPC development for ~2 years
- Matt's predictions:
  - Last Sonoma meeting: 9-10K IB nodes by 4QCY06
  - Actual: 12.5K nodes
  - Will be at 17K+ nodes by early 2008

Multi-scale Multi-physics modeling: the scale challenge



Fundamental Research in Integration Methodology for Rocket Simulation

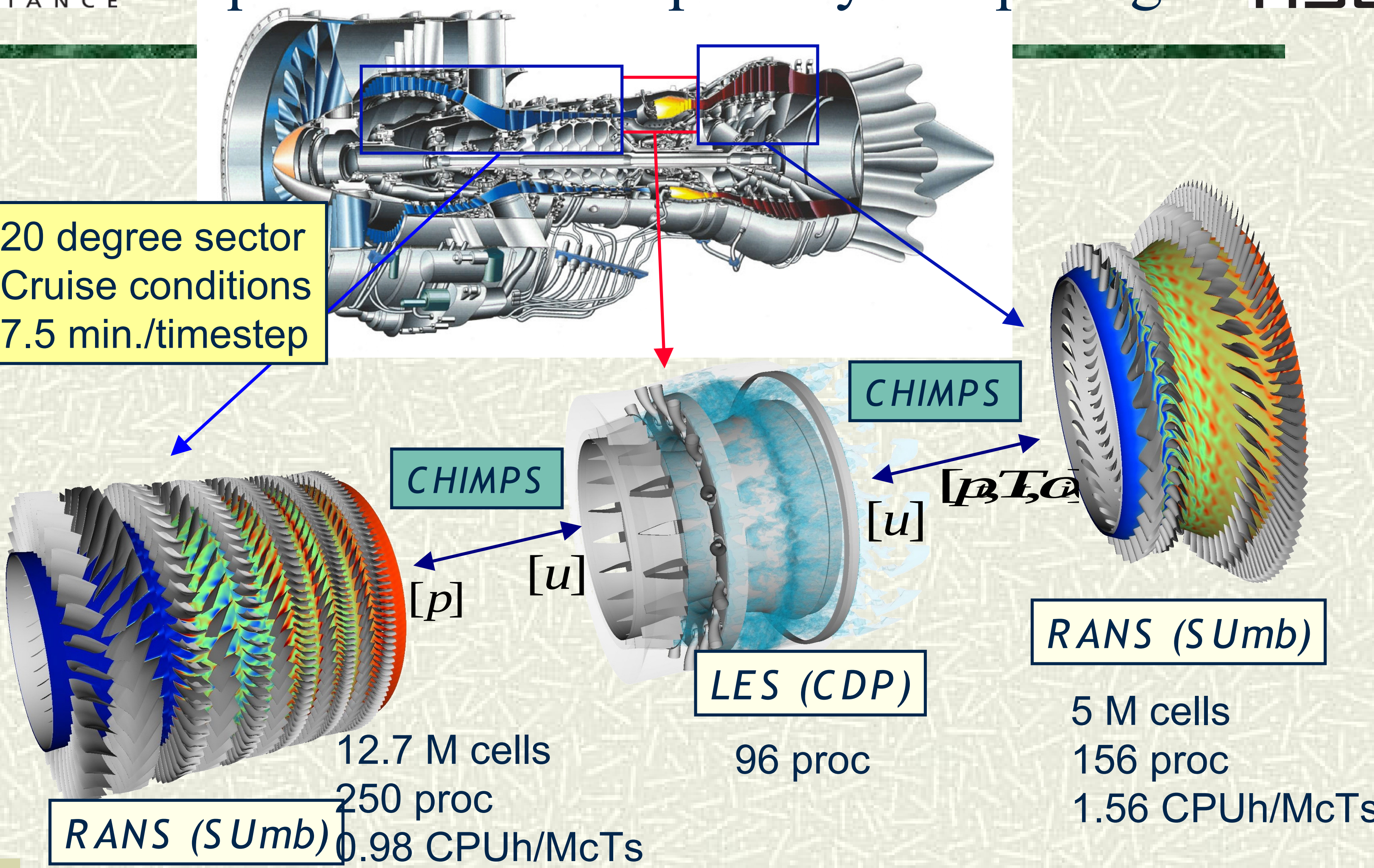
- Multiphase flow
  - Equilibrium Eulerian method for fine particles & Lagrangian super particles
- Turbulence modeling (Optimal LES)
- Time zooming
- Propellant morphology
  - Parallel packing code
- Propellant modeling
  - 3-D simulations with *Rocfire*



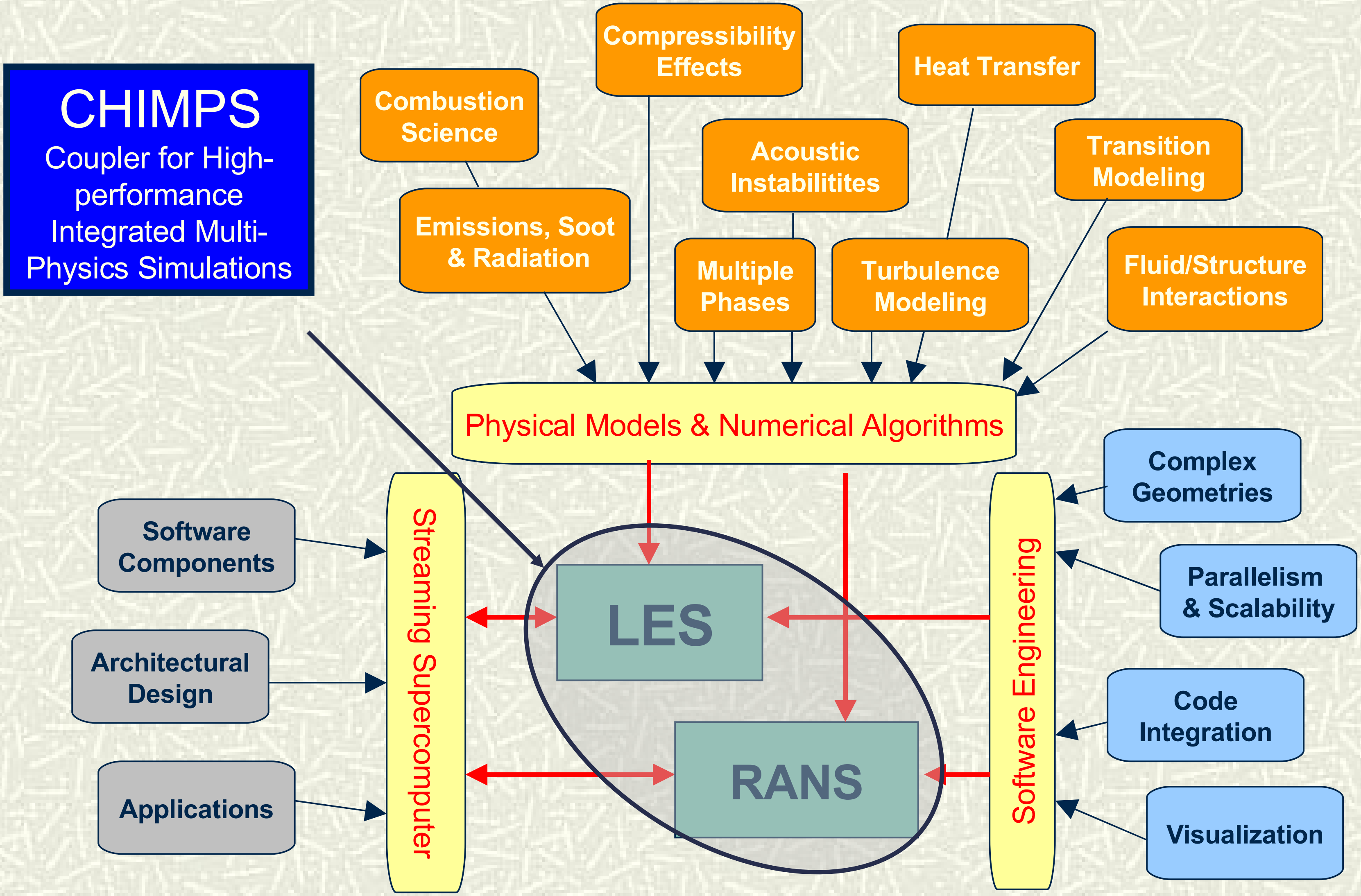
- Constitutive and damage modeling
  - Heterogeneous propellants and HE
  - Metallic components
- Crack propagation
  - Burning, pressure driven
- Multiscale materials modeling
- Molecular-level modeling of material interfaces
- Space-time discontinuous Galerkin methods

Full system jet engine configuration requires extreme capability computing

- 20 degree sector
- Cruise conditions
- 7.5 min./timestep



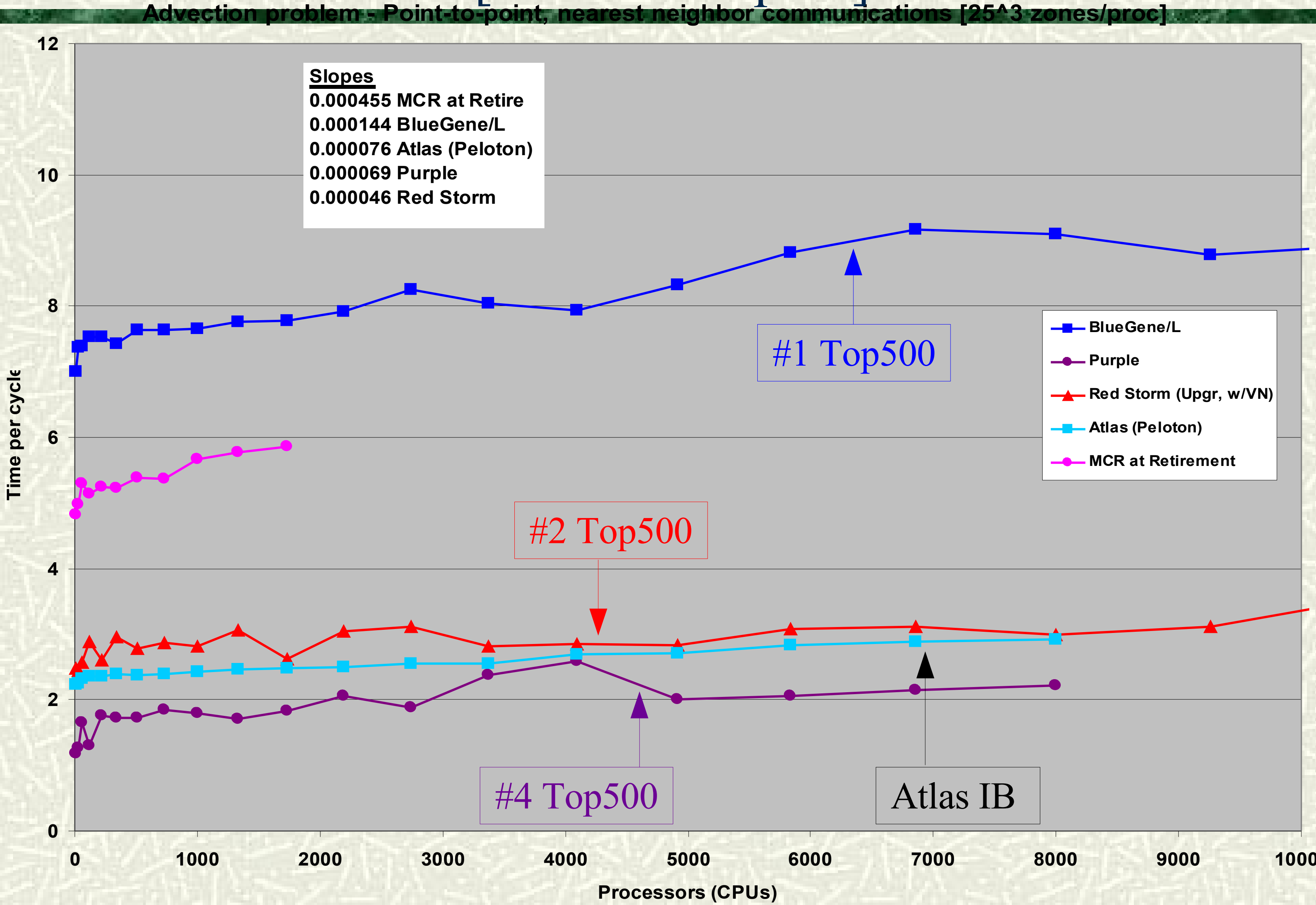
Fundamental Research Required in a Wide Range of Disciplines



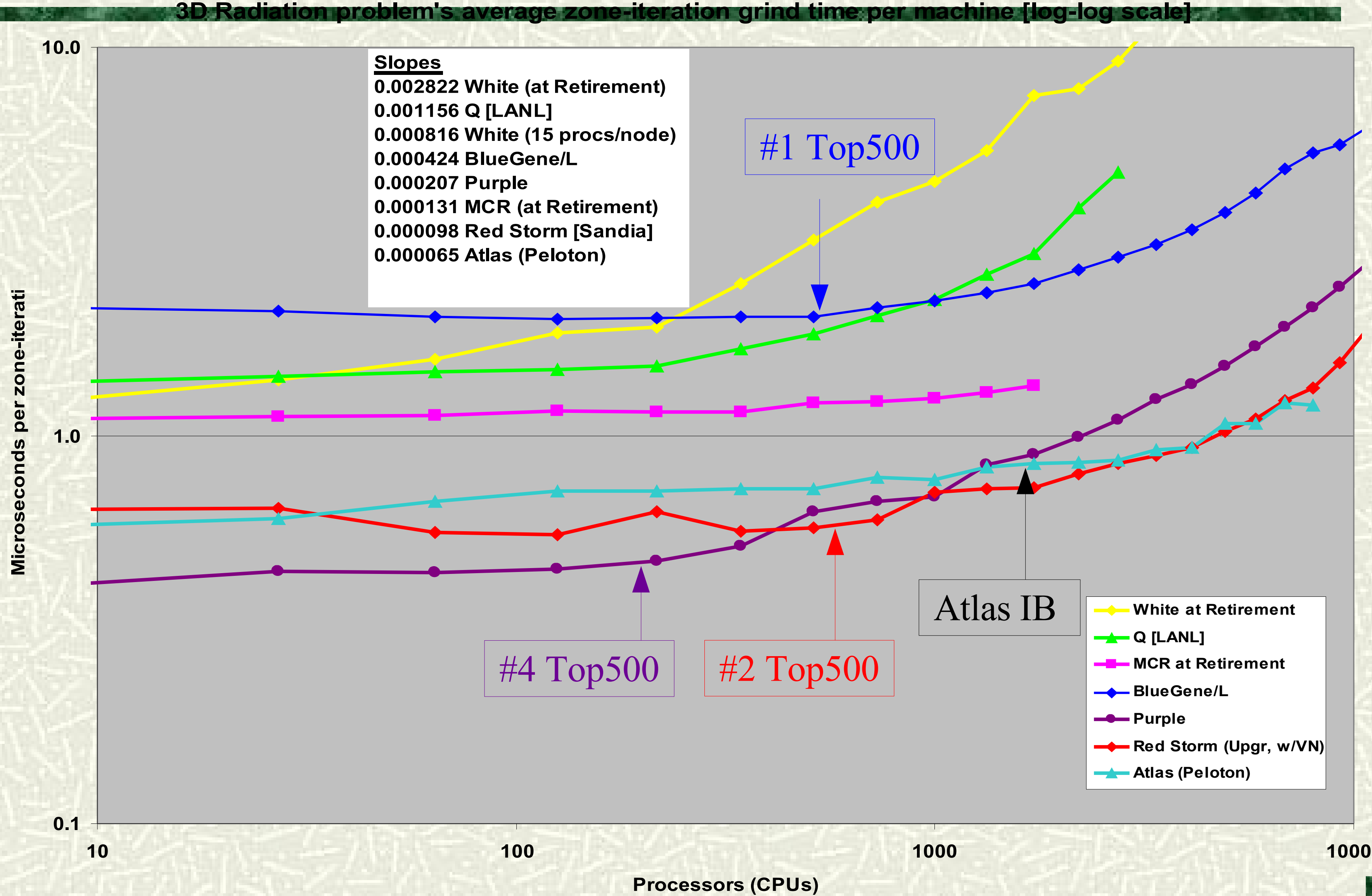
Some Examples of Application Performance and Scalability

- Hydrodynamic Advection
  - 3D Hydrodynamic advection
  - Same mesh size as for 3D radiation problem
  - Predominantly near-neighbor one-to-one communications pattern
    - Historically scales very well
- Radiation Diffusion
  - 3D Implicit thermal radiation diffusion
  - Stresses the solver (flux limited diffusion)
  - Has large number of iterations in matrix solve
    - Diagonally scaled conjugate gradient
  - Exhibits intensive collective communications
    - Each iteration has a reduction operation
    - Historically scales poorly, such as on ASC White

Hydrodynamic Advection Problem – Point-to-point, nearest neighbor communication [25^3 zones/proc]



3D Radiation Problem Average Zone-iteration Grind Time per Machine [log-log scale]



Current State InfiniBand

- Static Routing seems to work pretty well for some of our key scientific applications
- Other applications and benchmarks do not perform as well and spend 25-50% of runtime in network communication
- Our experience with fully adaptive routing, in non-IB networks, gives us confidence that we could be doing a lot better
- Technology Trends point to a growing need for adaptive routing

Computing Trends are Demanding Efficient Network Fabrics

- Moore's law will continue but only with multiple cores
  - Multi-core nodes will quickly reach over 1 Teraflop/s
  - More cores per node puts new stresses on the network fabric
  - Virtualization will lead to an overlay of traffic patterns and lead to unpredictable traffic patterns
- How do we deal with congestion?
  - Over-provision the network
    - Not scalable or cost effective in multi-core environment
  - Multi-path static routing
    - Small improvement over single path static routing
  - Back-off techniques to deal with head-of-line blocking (CCA)
- Future networks must incorporate effective congestion avoidance/management methods that do not assume *a priori* knowledge of traffic patterns

Application/Simulation Trends are Demanding Efficient Network Fabrics

- Concerned about not just today's applications but tomorrows multi-physics simulations
- BW to Flops ratios of 0.05 to 1.0 for capacity and capability systems
- ASC Applications today
  - On large SMPs will outpace the current IB 4X DDR/QDR technology
  - Multi-core chips increase network BW requirements
- ASC Applications future
  - Multi-core will outpace current IB roadmap leading to the need for more effcient fabric and likely multi-rail systems
- Simulation requirements are for 100's to 1000's of high resolution 2D and low res 3D runs and 10's of medium to high res 3D runs
- Algorithms incorporating more complex physics
  - non-local/global effects which lead to more stress on the network and MPI collective operations
  - Longer term: modern algorithms need to be latency tolerant

What is Adaptive Routing?

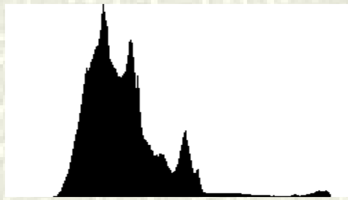
- Adaptive Routing is NOT
  - Multi-path static routing
    - Does not react to the state of the network
  - IBA Congestion Control Architecture
    - Network reacts to congestion/hotspots to reduce network load (“Back off”)
    - Only helps once congestion is happening
- Adaptive Routing is
  - A routing algorithm that makes decisions based on the network state, (queue occupancies/depth, least used channel, etc.) to select among alternative paths to deliver a packet
  - Spreads network traffic to reduce risk of congestion/hotspots

What is Adaptive Routing?

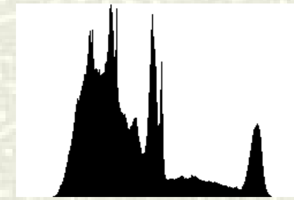
- Fully Adaptive Routing requires:
  - Switch/router chip logic for adaptive algorithm and state information
  - HCA to handle out-of-order delivery of packets
  - New InfiniBand silicon
  - Changes to the InfiniBand specification
- Adaptive routing becomes more important as
  - Radix of the switch/router chip increases
  - Size of the network fabric increases
  - Size of message/packet increases

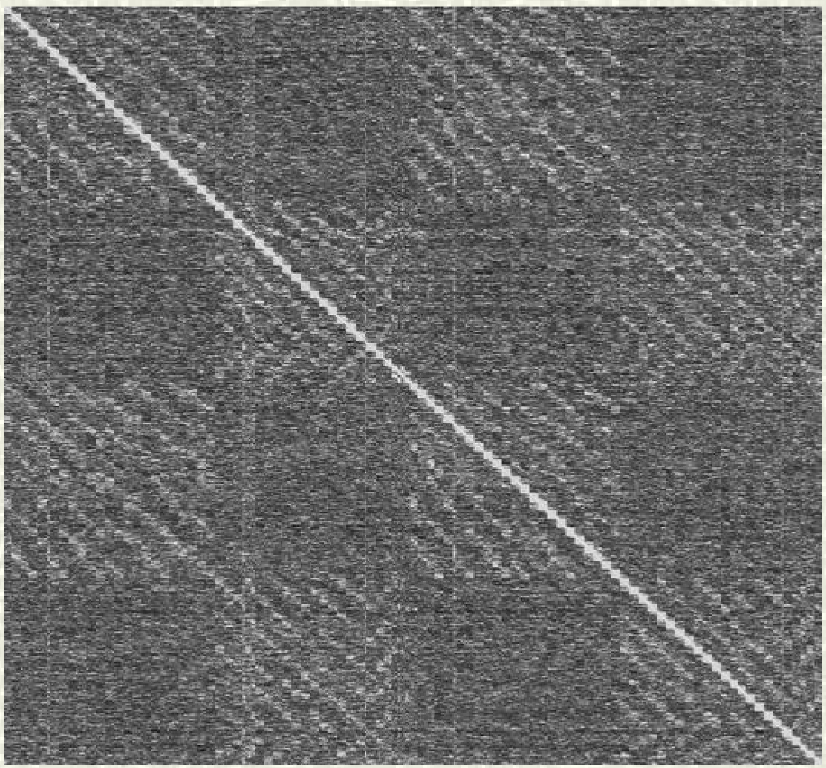
Static/Deterministic Routing on 1,152 Nodes of LLNL Atlas Cluster

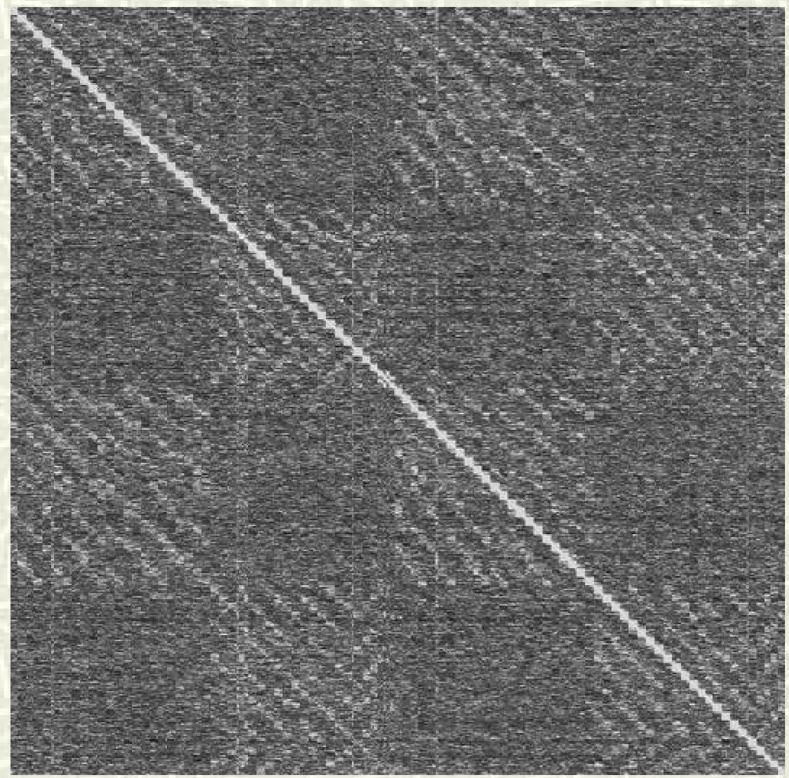
- MPI Send
- Max: 762 MB/s
- Average: 263 MB/s
- Min: 95 MB/s



- MPI Recv
- Max: 846 MB/s
- Average: 340 MB/s
- Min: 95 MB/s





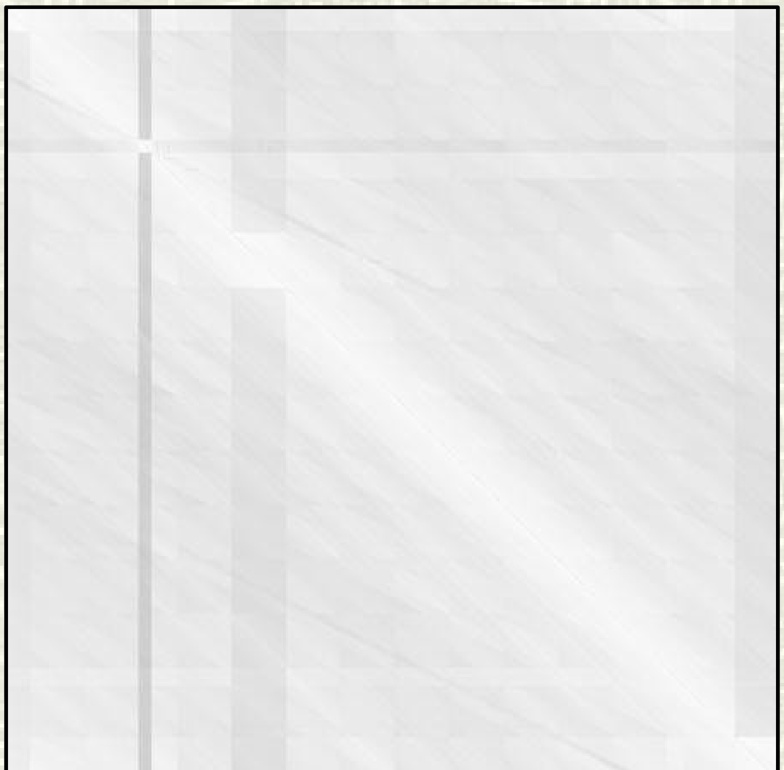


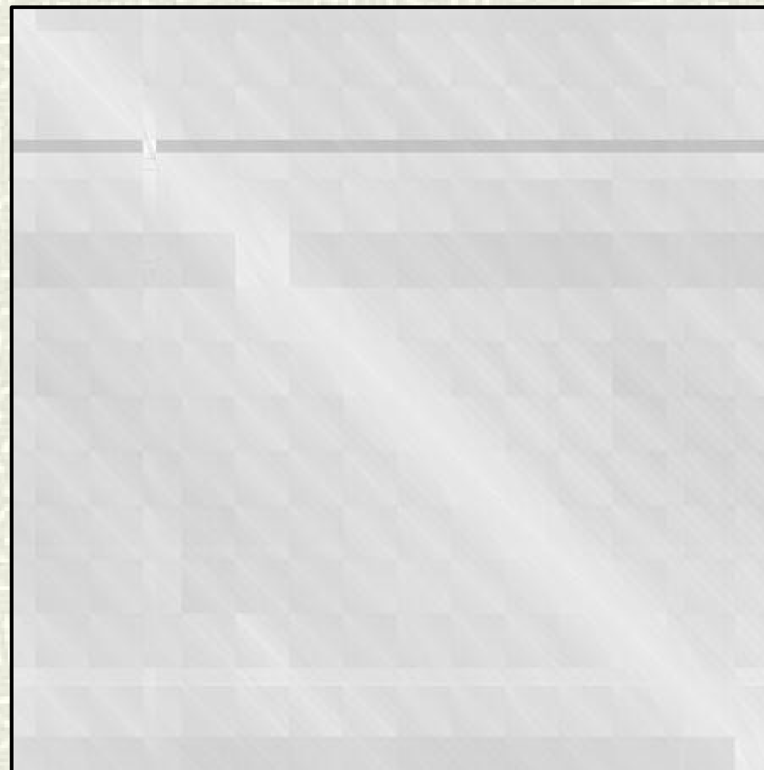
MVAPICH 0.9.7; 16 kB messages; 4X SDR InfiniBand

Adaptive Routing on 1,024 Nodes of LLNL Thunder Cluster

- MPI Send
- Max: 403 MB/s
- Average: 369 MB/s
- Min: 248 MB/s

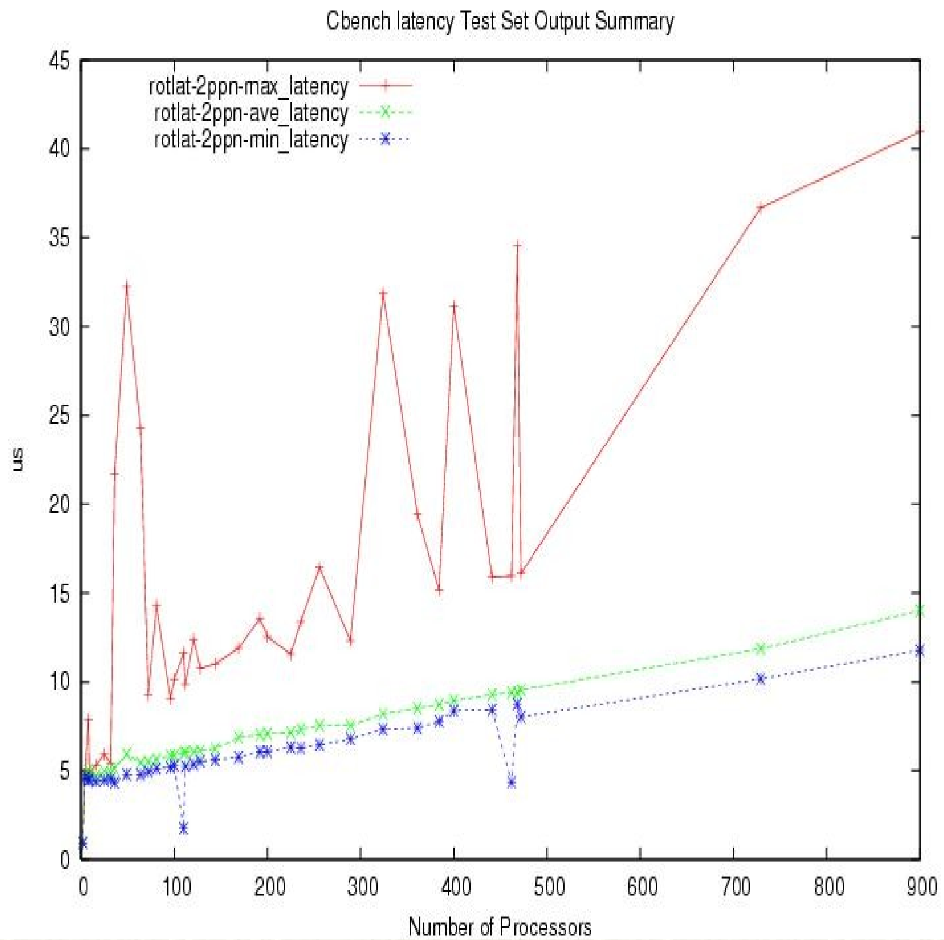
- MPI Recv
- Max: 427 MB/s
- Average: 370 MB/s
- Min: 246 MB/s

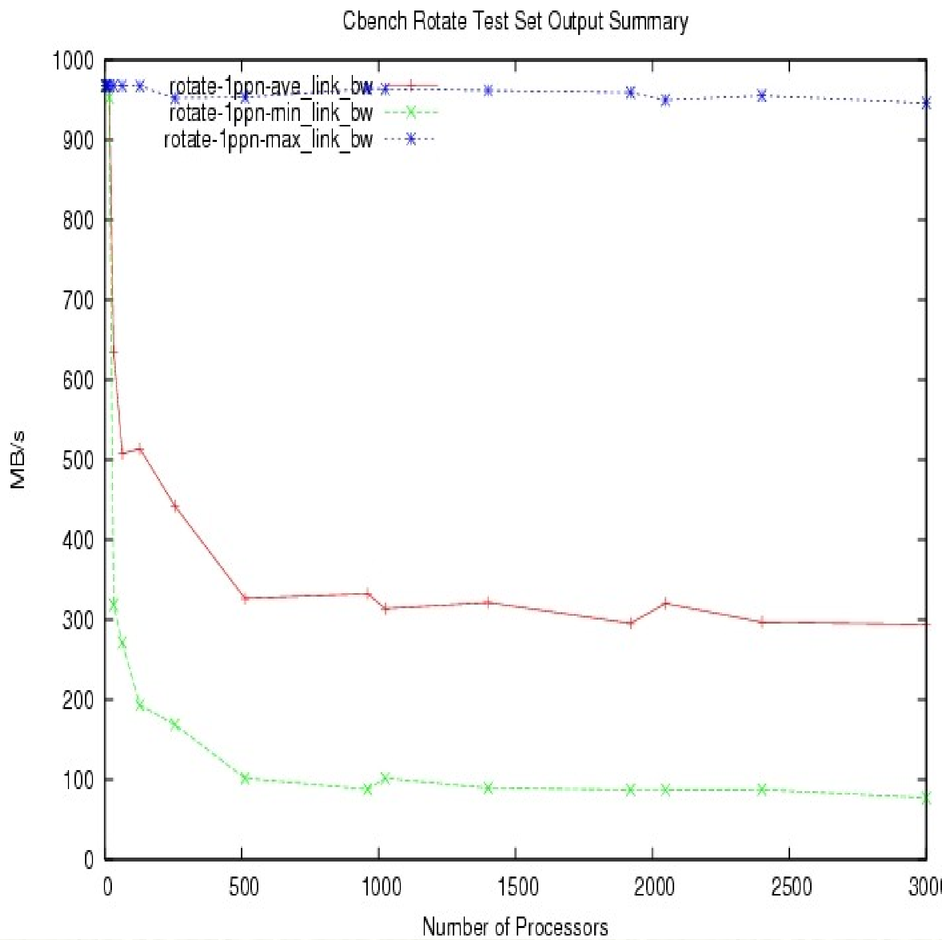




16 kB messages; Quadrics Elan4

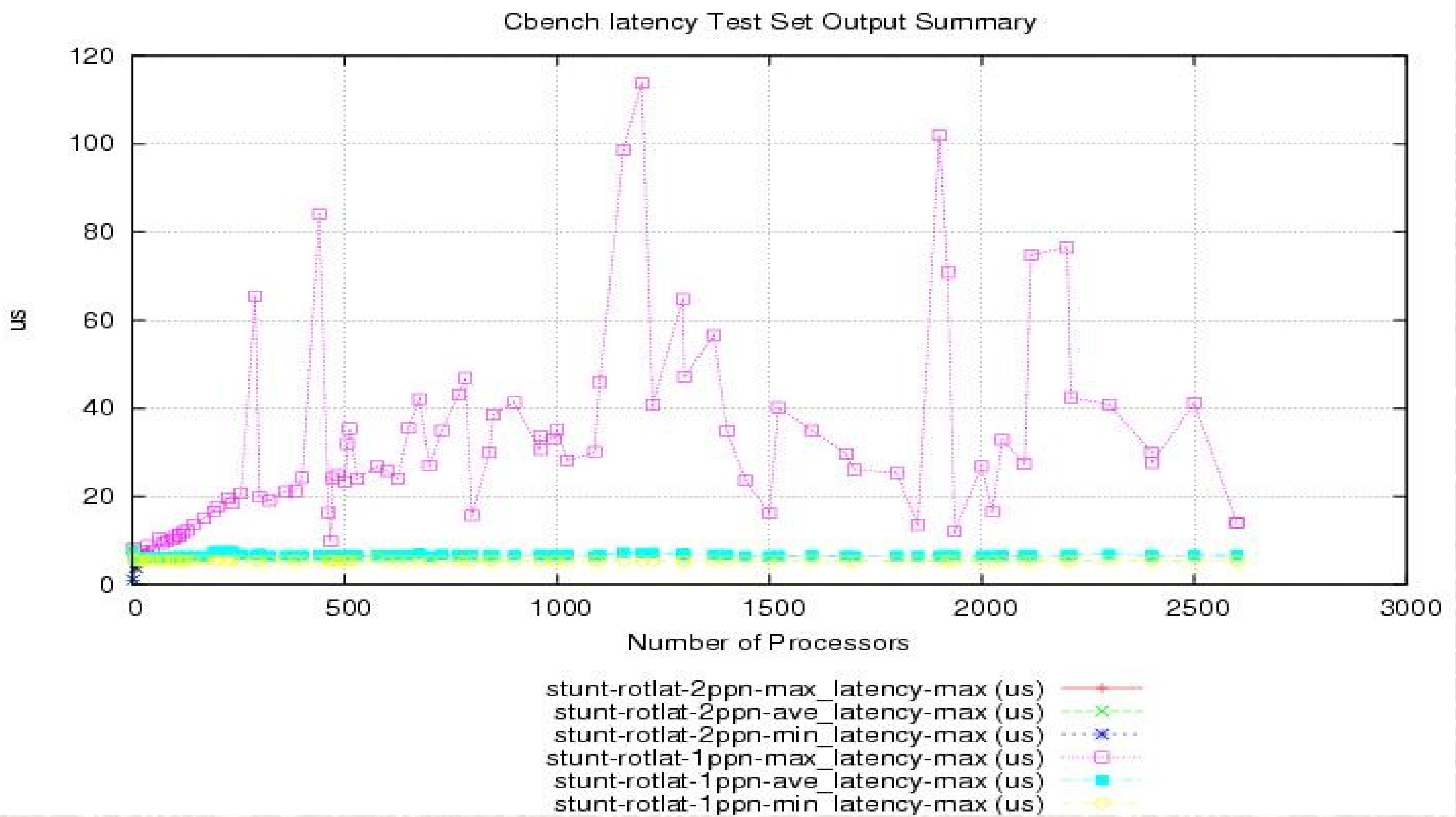
Static Routing on SNL Thunderbird Cluster





SDR InfiniBand

Static Routing on SNL Thunderbird Cluster



SDR InfiniBand

What About Optimized Static Routing? Part 1

- Recent work by LANL PAL group, Mellanox, and Valencia network group
- Static routing can out perform adaptive routing
  - If traffic pattern is know *a priori*
  - If traffic pattern is persistant
  - If traffic pattern is uniform (application is load balanced)
  - If the network has no faults (no bad/down links)

What About Optimized Static Routing? Part 2

- In a Multi-physics world application are not ideal or well-behaved
  - Traffic patterns are not always know *a priori*
  - Multi-physics applications will have many different communication phases
  - Can't change routes for each communication phase
  - Optimizing routes for each communication phase, for every application, is not a tractable solution
  - Network have faults – bad and/or downed ports and links
  - Applications tend to develop load imbalances since it is too expensive to correct for it each iteration
- Remember for Enterprise Computing virtualization will lead to similar issues
- Optimized Static/Deterministic routing requires too many components to work perfectly
- Any high performance routing method must perform well under non-ideal conditions and be robust under network link failures, application load imbalance, and a wide range of traffic patterns

Conclusions and Future Directions

- Trends in multi-core processor technologies are driving the need for more efficient utilization of network fabrics
- Problems with static routing today, with R&D showing that any future IB products need to incorporate adaptive routing
- Labs are starting R&D projects to use our real application traffic patterns in simulators of large scale network fabrics

Questions?

Matt Leininger (mlleini@sandia.gov and leininger2@llnl.gov)

Mark Seager (seager@llnl.gov)

Journal Articles on Adaptive Routing:

"Adaptive Routing in High-Radix Clos Network", J. Kim. W. J. Dally, D. Abts, SC2006
http://cva.stanford.edu/people/jjk12/sc06.pdf

"Adaptive Source Routing in Multicomputer Interconnection Networks"
Y. Aydogan, C. B. Stunkel, C. Aykanat, B. Abali