Of FLITS and FLOPS: Balancing Energy and Interconnect Performance

Scott Hemmert

Scalable Computer Architectures
Sandia National Laboratories

SAND2009-2588C

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.

NNSA National Nuclear Security Administration

Sandia National Laboratories

Contributors

- Sandia
  - Jim Ang
  - Brian Barrett
  - Ron Brightwell
  - Kurt Ferreira
  - Sue Kelly
  - Jim Laros
  - Kevin Pedretti
  - Courtenay Vaughan
- Indiana University
  - Torsten Hoefler

System-level Interconnect and Energy

- Interconnect performance is the key factor in determining how well many applications scale
- With increasing bandwidths, interconnect power is becoming a real concern
  - Serdes don't turn off well (OK, they turn off fine, they just don't turn back on quickly, due to channel initialization times)
    - Uses power whether valid data is moving through the network or not
- A lot of discussion lately on minimizing picojoules/bit
- However, interconnects are not used in isolation and a system view is vital to maximizing energy efficiency
  - NIC and router architectures, topologies and MPI implementations all play an important role

Application Case Study: CTH

- **CTH is a multi-material, large deformation, strong shock wave, solid mechanics code developed at Sandia National Laboratories. CTH has models for multi-phase, elastic viscoplastic, porous and explosive materials.**

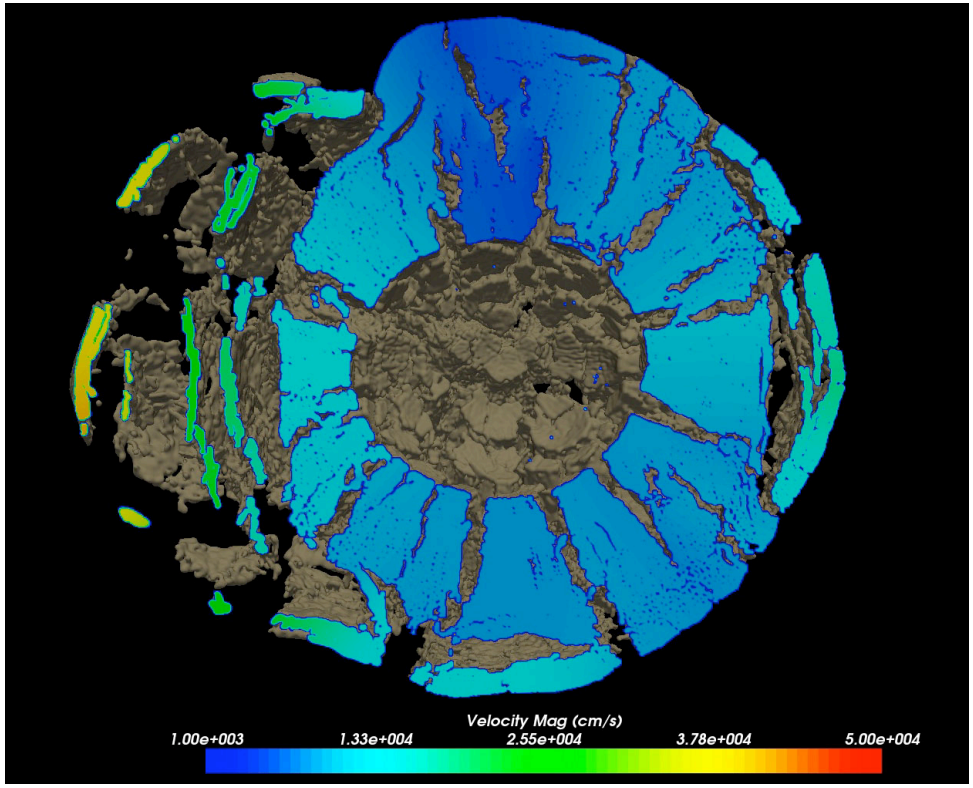
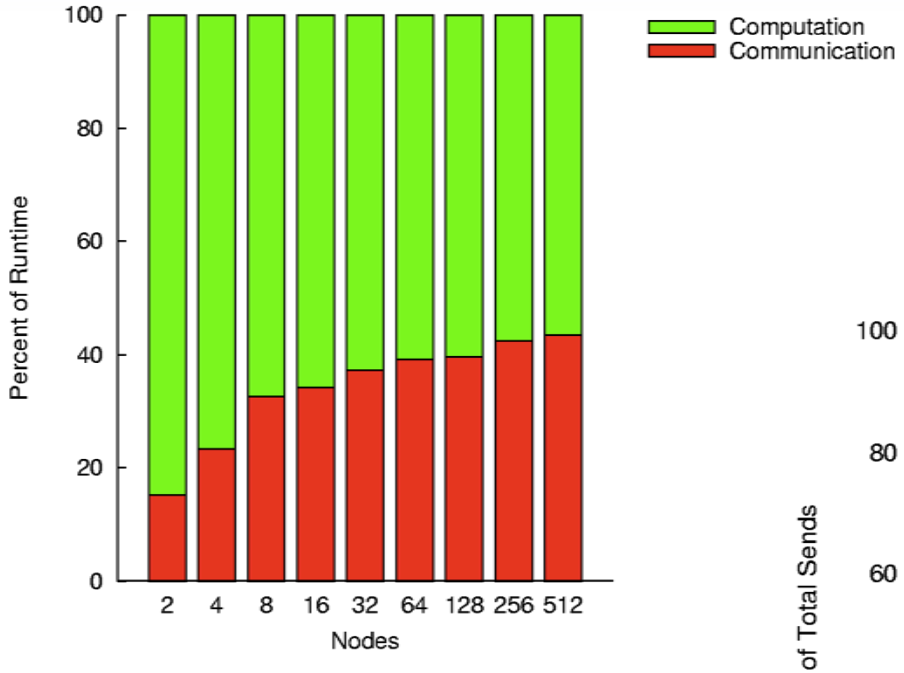


Image courtesy of ASC

Asteroid Golevka measures about 500 x 600 x 700 meters. In this CTH shock physics simulation, a 10 Megaton explosion was initiated at the center of mass. The simulation ran for about 15 hours on 7200 nodes of Red Storm and provided approximately 0.65 second of simulated time.

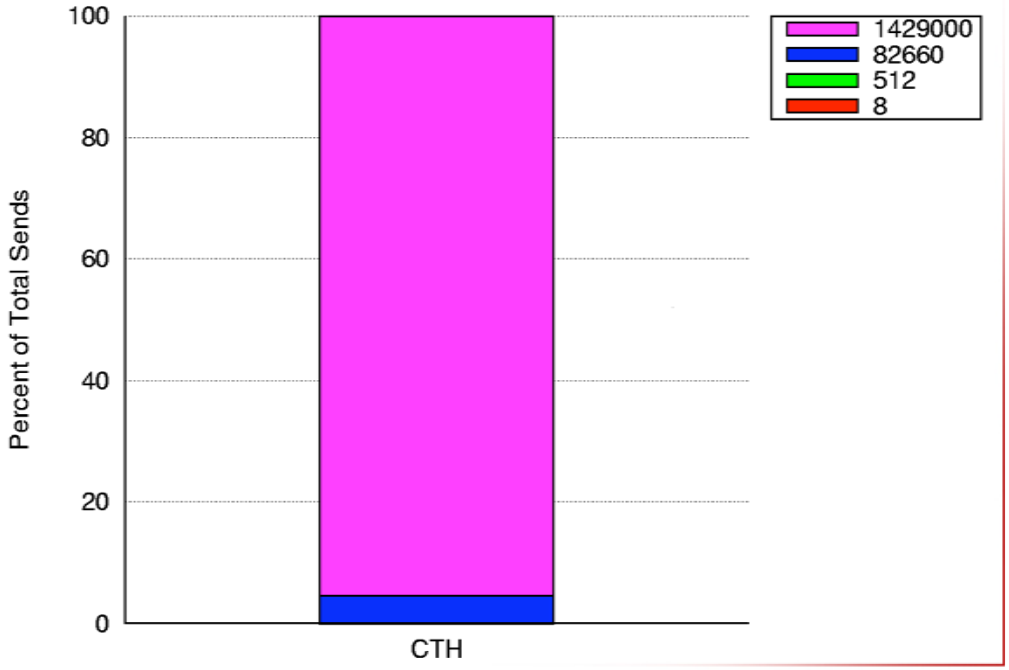
Application Case Study: CTH

Shaped Charge Problem (weak scaling)



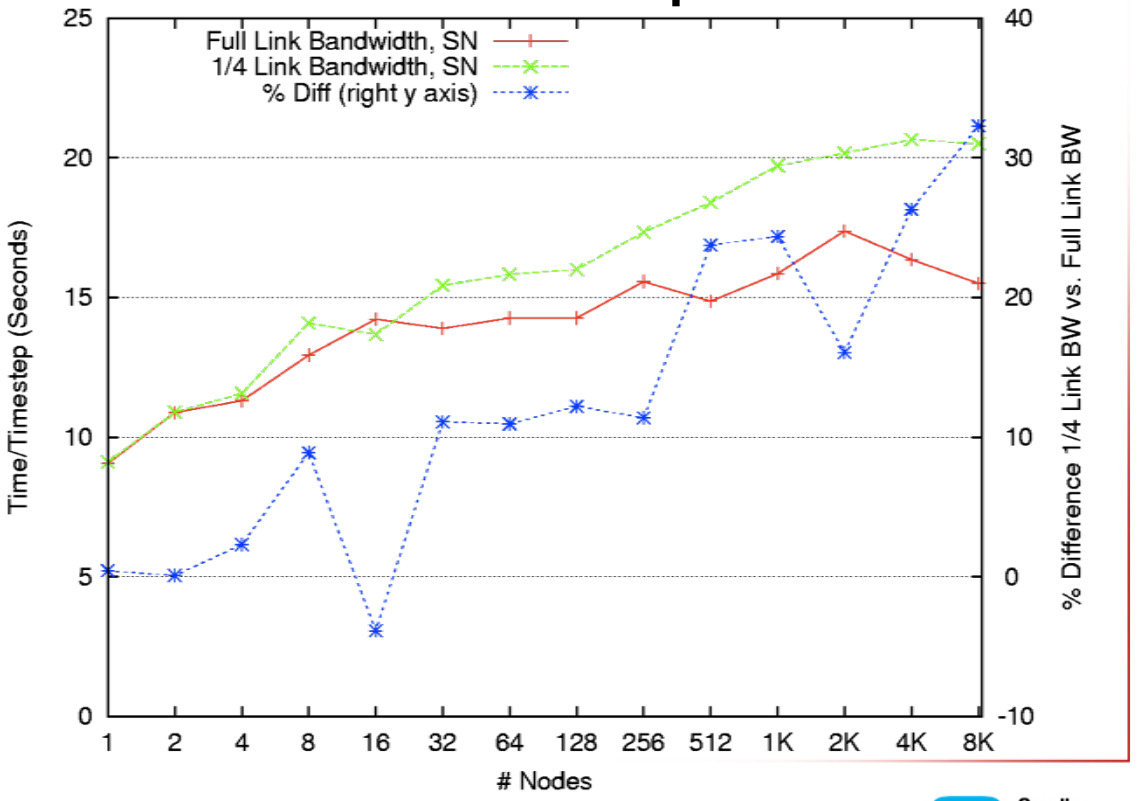
As job size increases, communication time can grow to consume around 40-50% of the runtime.

CTH communication is dominated by long messages.



CTH Bandwidth Degradation Study

- Uses capabilities built into the Red Storm SeaStar interconnect to turn off interconnect router lanes at boot time
  - Links are made up of 4 3-bit subchannels that can be independently enabled
- Measure application performance at full and one-quarter link bandwidth
- At largest measured job size, quartering bandwidth leads to 32% longer runtime



NNSA National Nuclear Security Administration

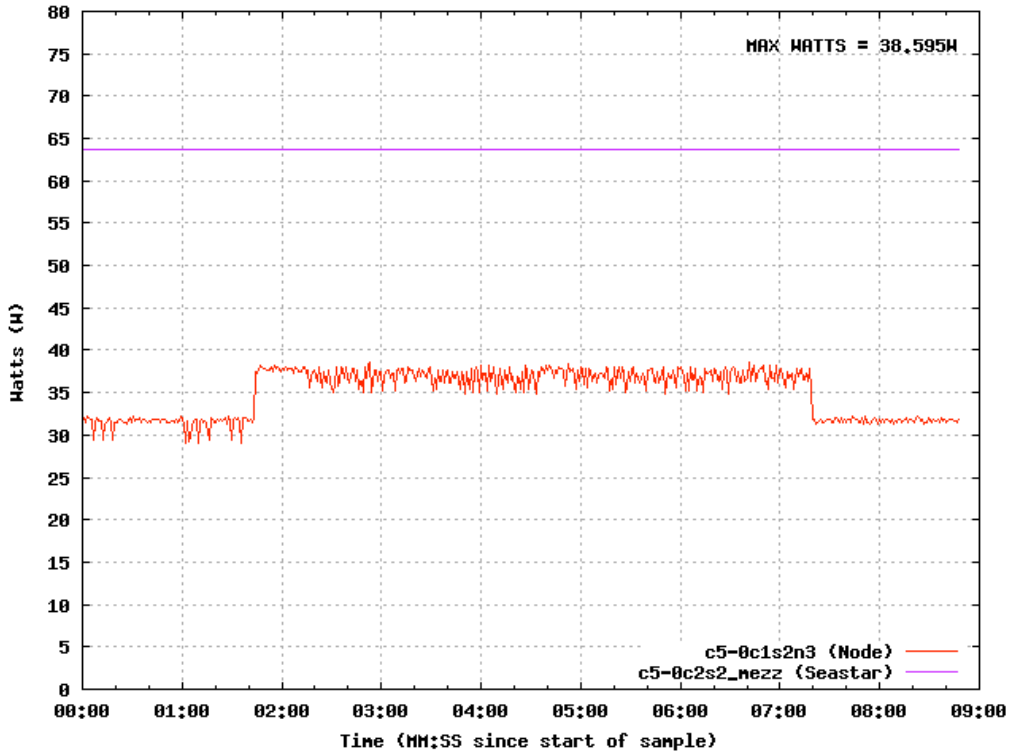
Sandia National Laboratories

CTH Power Signature Study

- **Power measured using Red Storm's built-in current monitors**

Total Node Power:

CPU: 37 (red)
SeaStar: 16 (blue/4)
Memory: 20 (estimated)
73 Watts



National Nuclear Security Administration

Sandia National Laboratories

Putting it all Together

- **Assume interconnect power drops linearly with bandwidth**
  - **68% of the performance for 25% of the interconnect power**
- **Total power for ¼ bandwidth = 61 Watts (down from 73 watts)**
  - **68% of the performance for 83.6% of the system power**

- **Total Energy for two cases assuming full bandwidth runtime of X**

$$\text{Energy}_{\text{full}} = 73X \qquad \text{Energy}_{1/4} = 1.32X * 61 = 80.5X$$

$$\frac{\text{Energy}_{1/4}}{\text{Energy}_{\text{full}}} = \frac{80.5X}{73X} = 1.10$$

- **Net energy increase of 10% for ¼ bandwidth case**
  - **Keep in mind this doesn't count the energy used for the file system attached to the machine or other machine room costs**

Application Case Study: POP

- POP is an ocean circulation model derived from earlier models of Bryan, Cox, Semtner and Chervin in which depth is used as the vertical coordinate. The model solves the three-dimensional primitive equations for fluid motions on the sphere under hydrostatic and Boussinesq approximations.

- POP sends small-ish messages (one run showed 16KB average message size) and spends large portions of it's MPI time in MPI_Allreduce (at large node counts)

- POP is generally believed to be a latency and/or message rate bound application

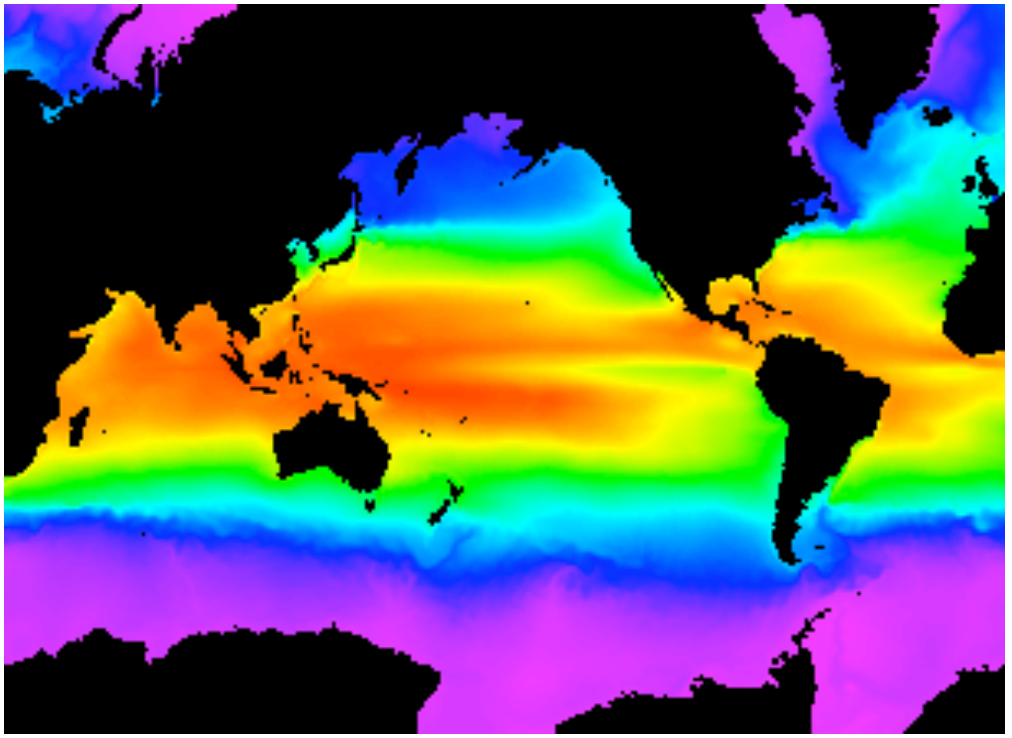
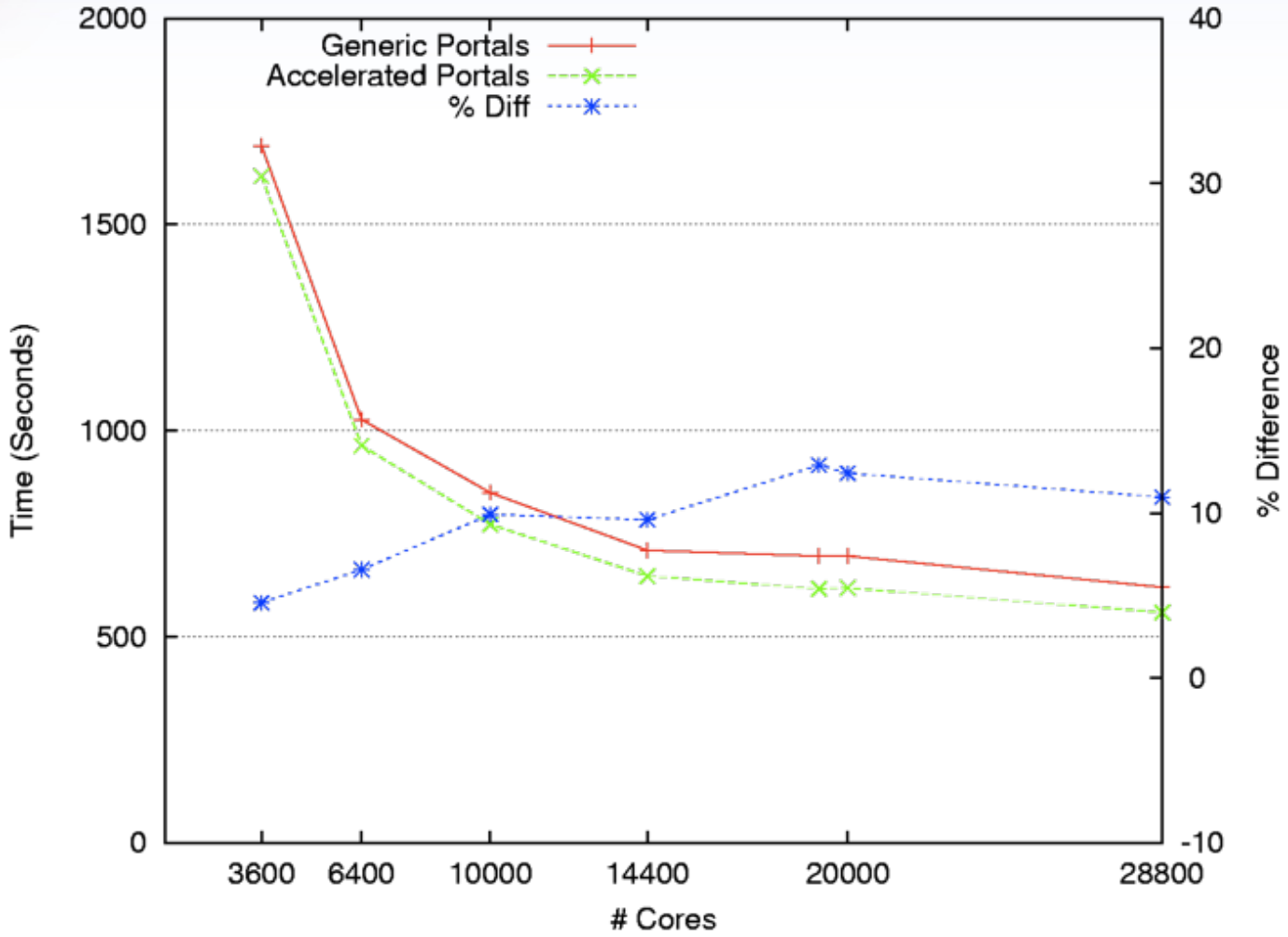
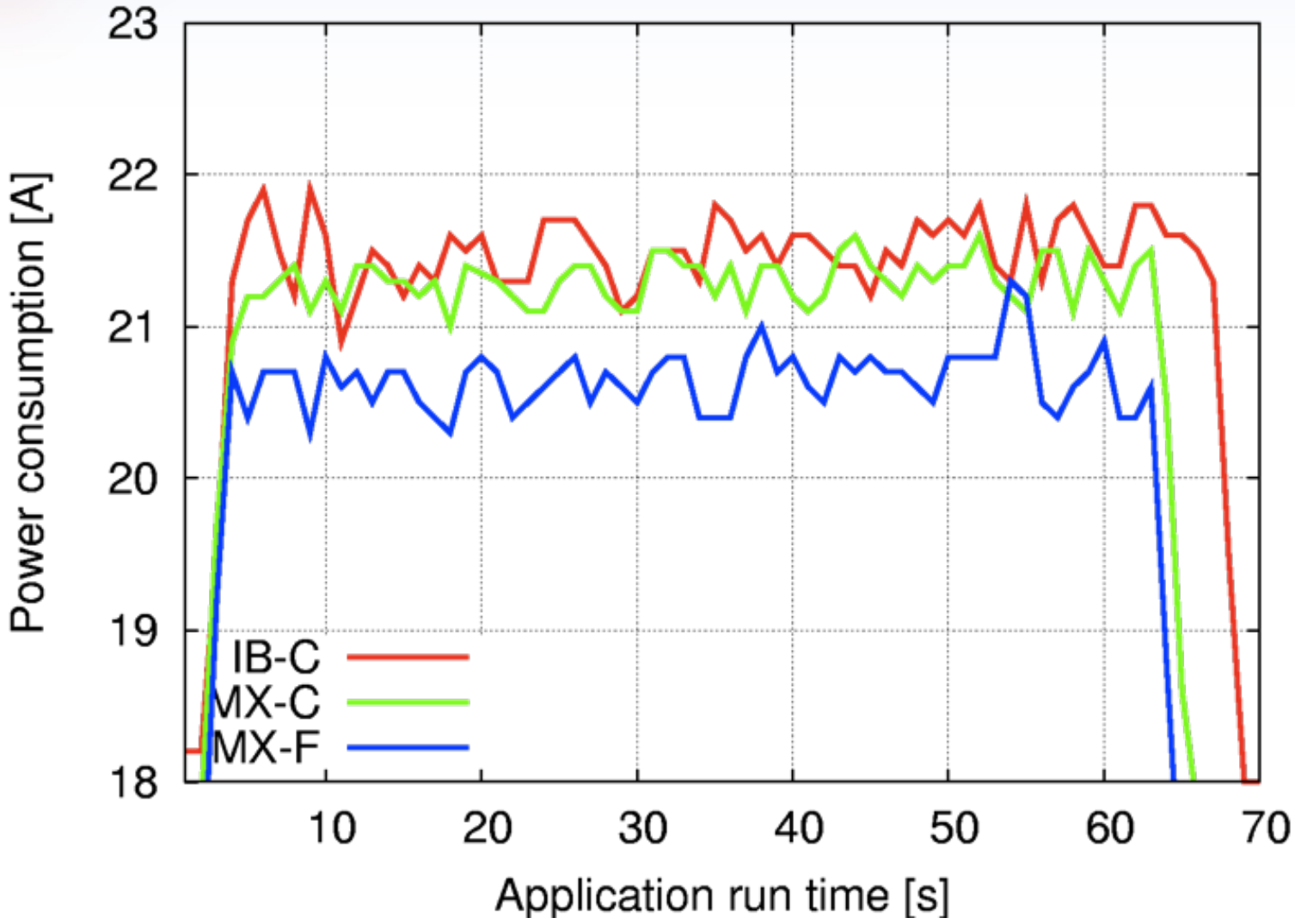


Image and description from http://www.lanl.gov/orgs/t/t3/codes/pop.shtml

Red Storm GP vs AP



Power Study from Indiana University



Torsten Hoefler, Timo Schneider and Andrew Lumsdaine, "A Power-Aware, Application-Based Performance Study of Modern Commodity Cluster Interconnection Networks." To appear in IPDPS/CAC09, May 2009.

National Nuclear Security Administration

Sandia National Laboratories

System View is Vital

- First example showed a case where higher interconnect power leads to lower energy to solution
- Second example illustrates how advanced features which add very little to system power can improve performance, thus improving energy to solution

- The system view is critical
  - Interconnect is not an isolated system and only accounts for a portion of the total system power
  - Thus, higher interconnect power can actually lead to lower energy
  - Understanding the true impact of the interconnect trade-offs can lead to more energy efficient systems

Energy Fallacies

- **Areas were wrong assumptions can lead to less energy efficient solutions**
  - **Microbenchmarks**
  - **High Radix Routers**
  - **Tight Integration**

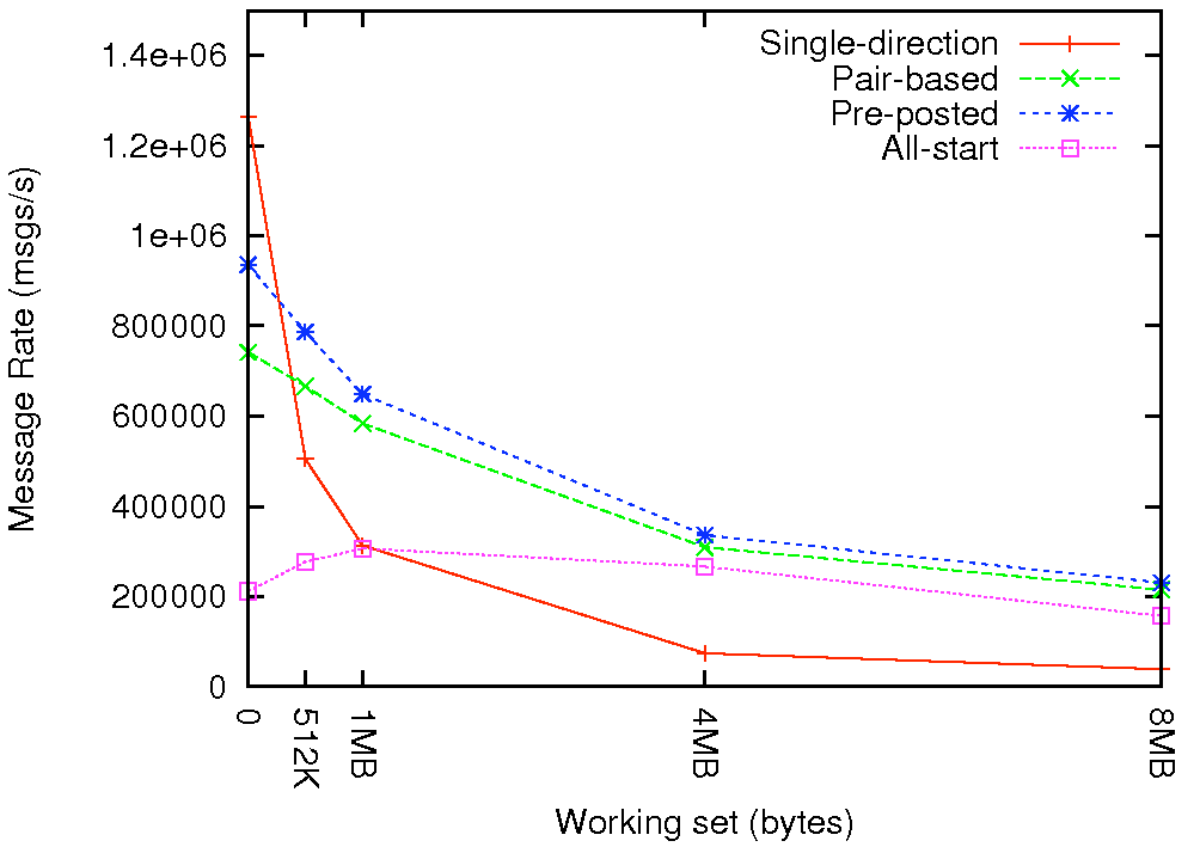
Microbenchmarks

- Fallacy: Optimizing interconnects and MPI implementations to microbenchmarks will necessarily improve application performance (or at least won't hurt it).
- Any optimization that reduces performance without reducing power will lead to less energy efficient system
  - Conversely, any optimization that increases performance without increasing power will lead to more energy efficient systems

- Removing useful advanced features to improve NetPipe latency and bandwidth will not generally translate to improved application performance (and may actually make it worse)
- Coalescing identical zero-byte messages will not help any application of which I am aware
- Measuring message rate under ideal conditions does not provide useful information about message rate achievable by an application

Sandia Message Throughput Benchmark

- **Measures message rate using communication patterns mimicking those of scientific applications**
  - **Simulation of computation/communication phase with variable working set sizes (compute stage modeled by touching data to invalidate some portion of cache)**
  - **Each MPI rank both sends and receives**
  - **Variable number of peers**



High Radix Routers

- **Fallacy: Small, globally random traffic is the only important pattern for HPC (Corollary: MPI doesn't matter anymore)**
- **What's driving high radix routers***

$$k \ln^2 k = \frac{B t_r \ln N}{L}$$

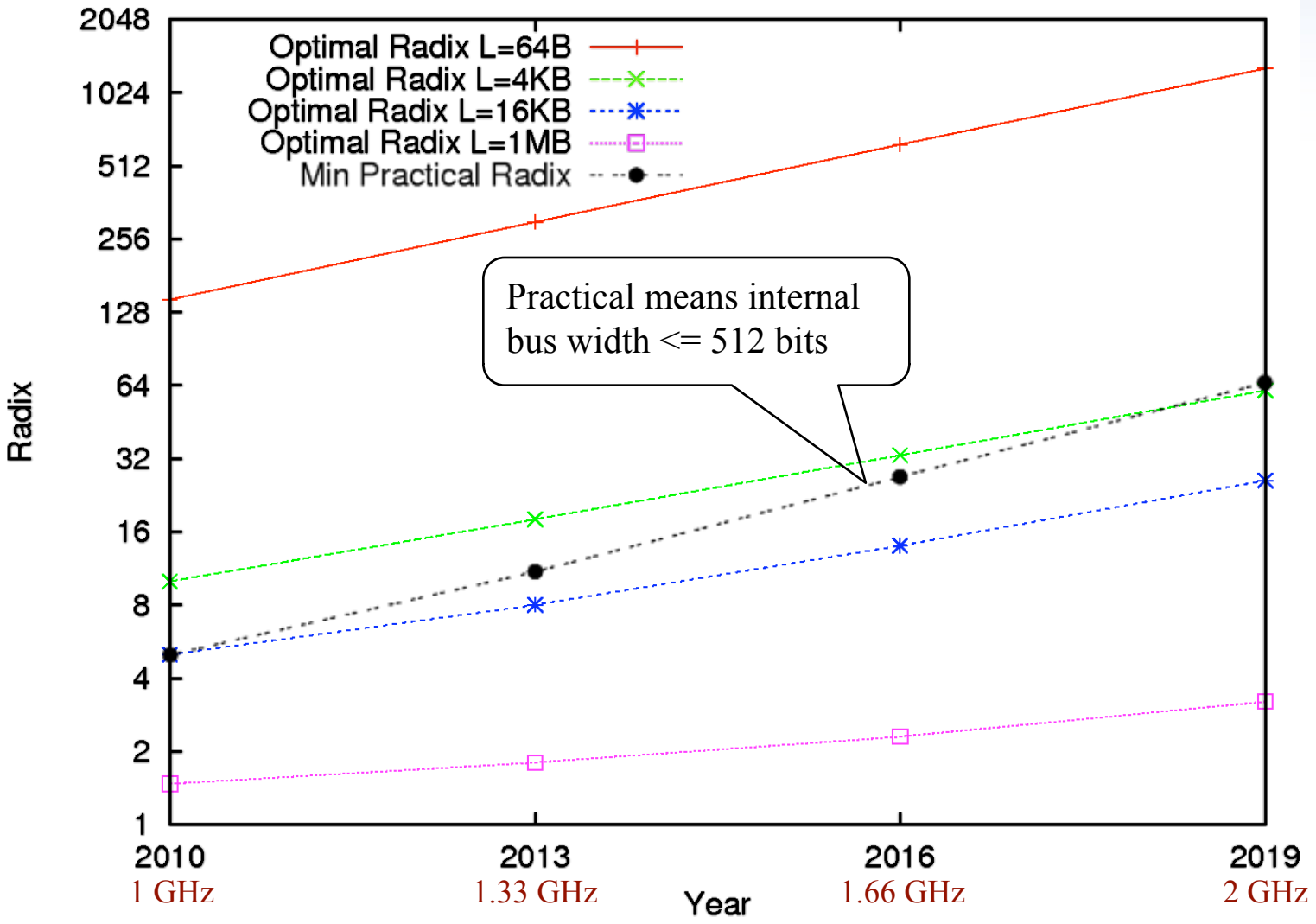
k = optimal radix
B = total router bandwidth
t_r = latency of a single router
L = length of packet (message)
N = number of nodes

- **Consider for 2010: B = 5Tbit, t_r = 35ns & N = 32k**

Cache Line Size Transfers:	Memory Page Size Transfers:	POP Size Transfers:	CTH Size Transfers:
L = 64B	L = 4KB	L = 16KB	L = 1MB
Radix = 144	Radix = 10	Radix = 5	Radix = 1.5

*From: John Kim, William J. Dally, Brian Towles, Amit K. Gupta. Microarchitecure of a High-Radix Router. In *Proceedings of the 32nd International Symposium on Computer Architecture (ISCA '05).*

Future Radix Trends



Assumptions:

2x machine size every 3 years

5 ns reduction in latency every 3 years

3x switch bandwidth every 3 years

National Nuclear Security Administration

Sandia National Laboratories

Comparison of Theoretical Modern Networks

	Dragonfly*	3D Torus	Fat Tree
Router Radix	64	7	32
Notes	2 links per node, group size of 256, 128 groups	32x32x32	Full bisection, based on 512 port switches
Number of Switches	4096	32768	7168
Bisection BW (Bi-directional)	80 TB/s	91 TB/s	160 TB/s
Node BW (Bi-directional)	9.8 GB/s	44.6 GB/s	9.8 GB/s
Max Hops	5	48	7

Assumptions: 2.5Tbit/s total switch bandwidth, 32k nodes

*John Kim, William J. Dally, Steve Scott, Dennis Abts, "Technology-Driven, Highly-Scalable Dragonfly Topology." In Proceedings of the 35th International Symposium on Computer Architecture (ISCA '08).

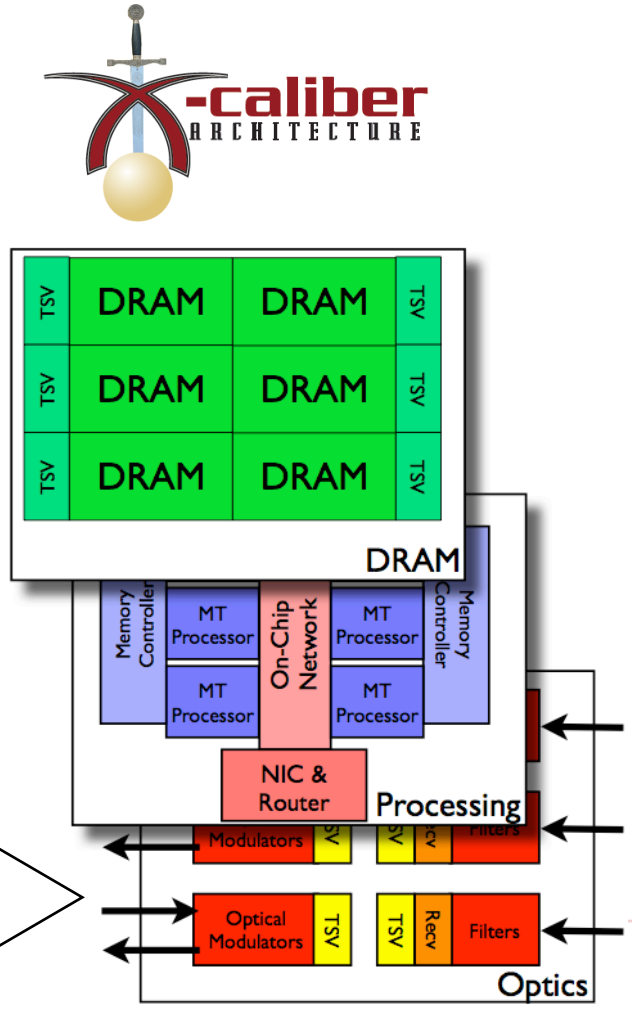
High Radix: The Wave of the Future?

- High radix routers are definitely in our future
  - Available bandwidth on a chip continues to increase
  - Core clock speeds are not
  - But, we have a few more years where low radix networks are feasible, and for certain workloads possibly favorable
    - Still need to study the energy/performance trade-offs of newer networks
- New work into hybrid network topologies is a good start in how to best use high radix routers
  - However, networks seem to be optimized for global random traffic
  - Need a better feel for how these types of topologies will impact performance of traditional scientific applications
  - What is the right radix??

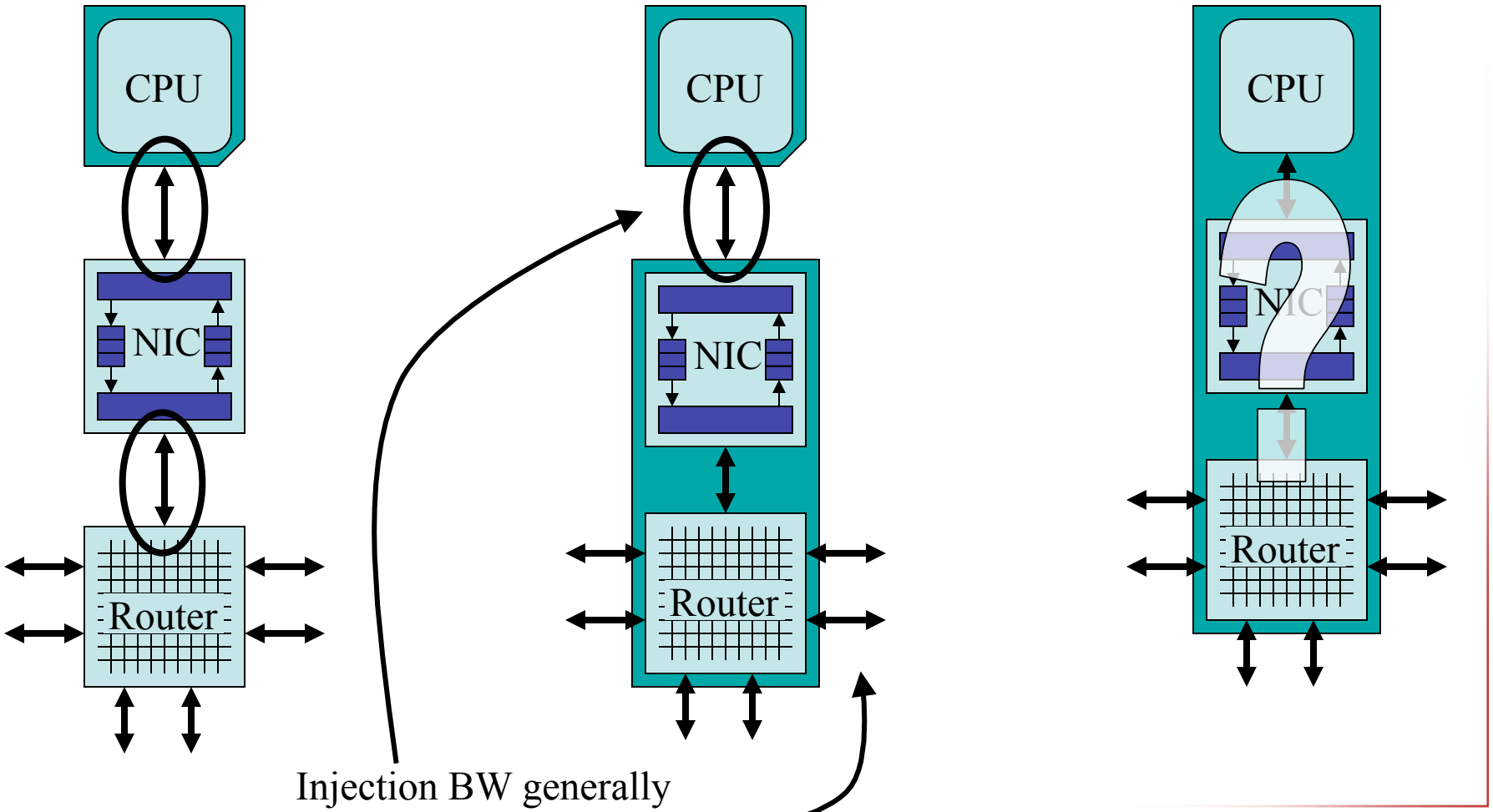
Tight Integration

- **Fallacy: Integrating router, NIC and processor onto same package magically provides access to more usable interconnect bandwidth.**
- **The problem is not with tight integration, it's understanding how that integration affects the balance of the system**

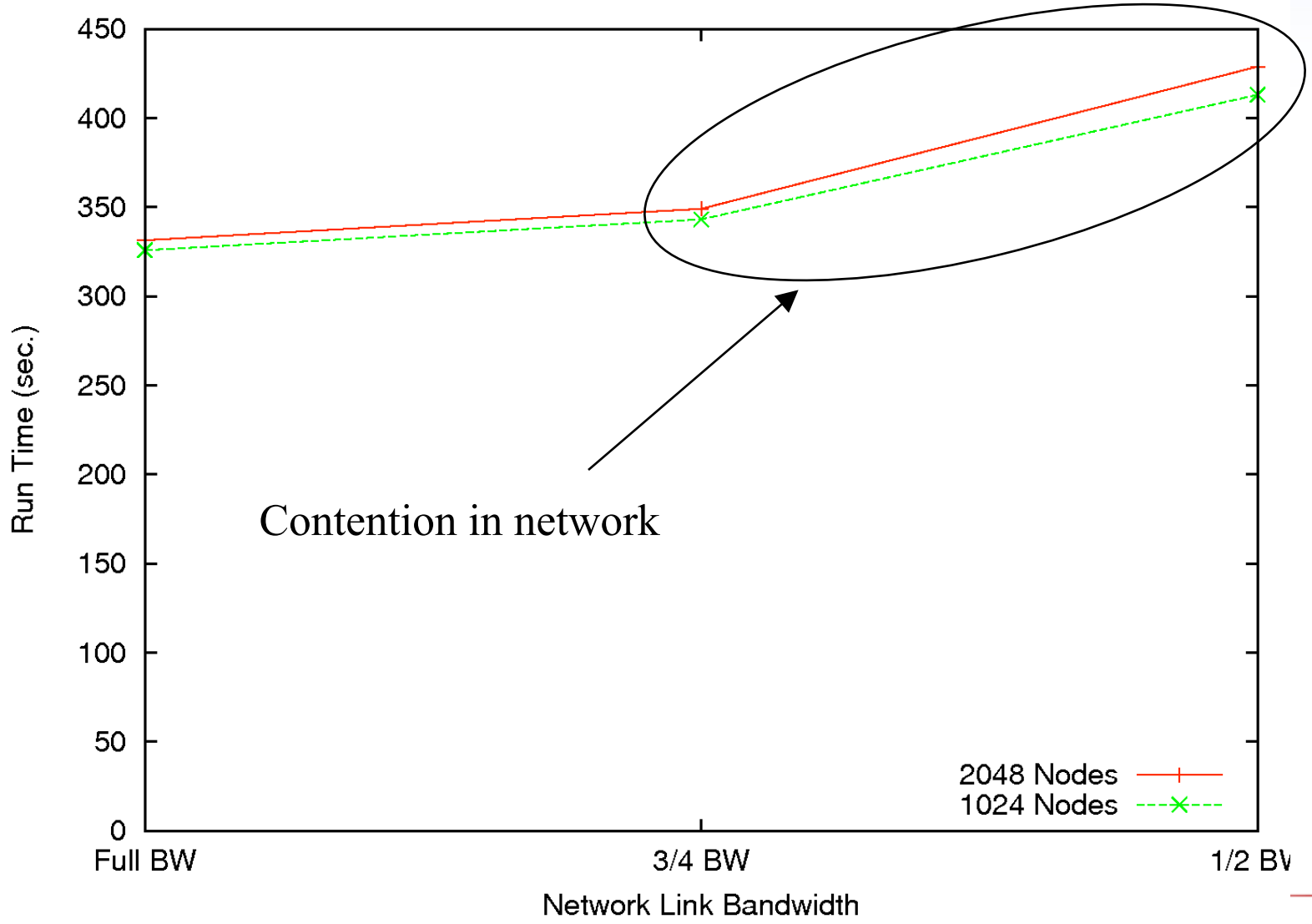
Million dollar question:
How much of this bandwidth is genuinely available to the processors?



Tight Integration and Injection Bandwidth



CTH Example: Network Contention



Being Honest with Ourselves

- What we realized: Most of the bandwidth into the stack is not typically usable by the cores in the stack
  - Most of the bits flowing in are not destined for that node
  - Most of the bandwidth going out is already being used by other traffic
- Expect to get the same utilization as when the router is off-chip
- Two approaches in the end:
  - Marketing approach: Count all the bandwidth
    - Detrimental impact on application performance/energy due to poor balance
  - Technical approach: Properly balance the system based on usable bandwidth

Conclusions

- It's not necessarily about power, it's about energy to solution
  - Higher power systems can actually lead to lower energy to solution
  - When peak power is a limiter, likely better off with a "smaller", more balanced system, than a larger, unbalanced system
- It's not about peak FLOPS/Watt, it's about the percent of peak that can be sustained
  - We pay an energy penalty for unused operations
  - With rising awareness of energy-efficient computing, FLOPS/Watt threatens to become the new HPL. Let's not let this happen!
- This talk focused on interconnects, but other areas are equally important
  - What's the application impact of slower, less complex cores
    - Can in-order cores use wide floating-point units?
    - Can applications scale to the dramatically increased number of cores?
- Components should be designed with a system view and understanding of the application needs

Questions?