Scaling Beyond Commodity

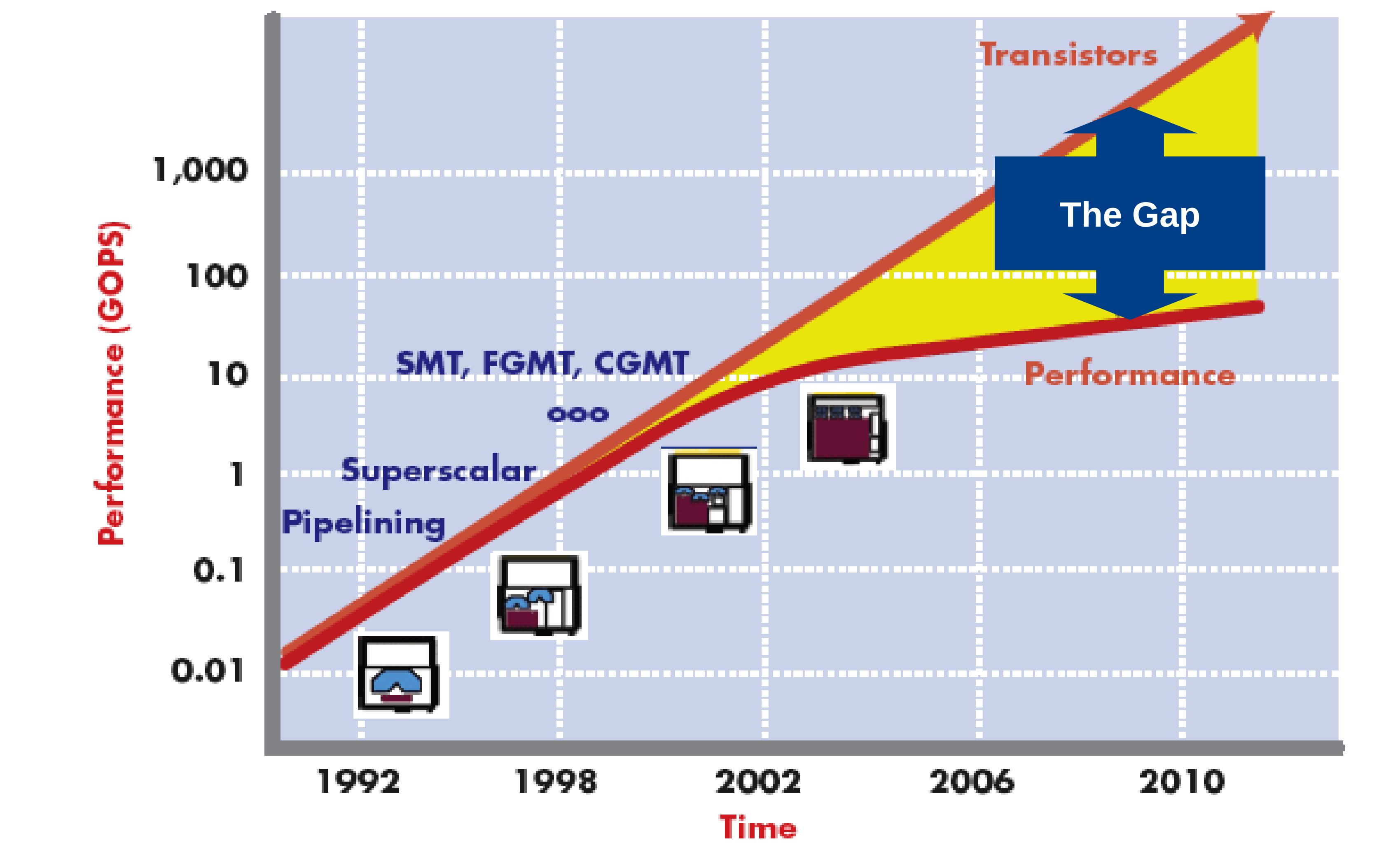
Key Challenges in moving towards Exaflop computing

John Levesque

Director, Cray Supercomputing Center of Excellence, Cray Inc.

September 2007

Single Processor Performance …
No Longer Tracking Moore's Law



Increasing Importance of Scaling

- Rate of increase has increased with advent of multi-core chips
- Sold systems with more than 100,000 processing cores today
- Million processor systems expected within the next five years
  - Equivalent to the entire Top 500 list today

Average Number of Processors Per Supercomputer (Top 20 of Top 500)

Source: www.top500.org

History of some “Unix-based” Cray systems

(about $20M each)

Cray 2
4 CPUs

Cray Y-MP
8 CPUs

Cray T90
16 CPUs

Cray T3E
1024 CPUs

Cray X1E
256 CPUs

Cray XT4
16384 CPU cores

1986 1990 1994 1996 2004 2007

Processors	4	8	16	1024	256	16384 **4096 X**
Memory	2GB	256 MB	4 GB	512 GB	1 TB	16TB **8192 X**
Frequency	240 Mhz	166 Mhz	440 Mhz	600 Mhz	1.1 Ghz	2.6 Ghz **11 X**
Peak	1.9 Gflops	2.6 Gflops	28 Gflops	1.2 Tflops	4.6 Tflops	150 Tflops **78,000 X**
Boot Time	~20 minutes	~20minutes	~20 minutes	~20 minutes	~20 minutes	~20 minutes **1X**

Realities

Supercomputing with commodity processors will become almost solely focused on ***scalability***

+

The flattening of the per-core performance trends has renewed interest in ***novel processing architectures*** and ***accelerator*** technologies

Some Customer Pain Points

Clusters are still hard to use and manage

- Power, cooling and floor space are major issues
- Third party software costs
- Weak interconnect performance at all levels
- Applications & programming - hard to scale beyond a node
- RAS is a growing issue
- Storage and data management
- Multi-processor type support and accelerator support

Where Should We Invest?

- Five areas to invest that yield big payoffs in scalability:

1) Reliability & Manageability
2) Interconnect Technology
3) Packaging for Performance
4) Scalable Software
5) Application Support

Reliability at Scale

$(\text{Probably})^{1,000,000}$ = ?

Probably Not

Reliability Features Needed At Scale

- Simple, microkernel-based software design
- Redundant Power Supplies and Voltage Regulator Modules (VRMs)
- Small number of moving parts
- Limited surface-mount components
- All RAID devices connected with dual paths to survive controller failure
- Interconnects with link-level reliable transport

One vs. Many

Beowulf

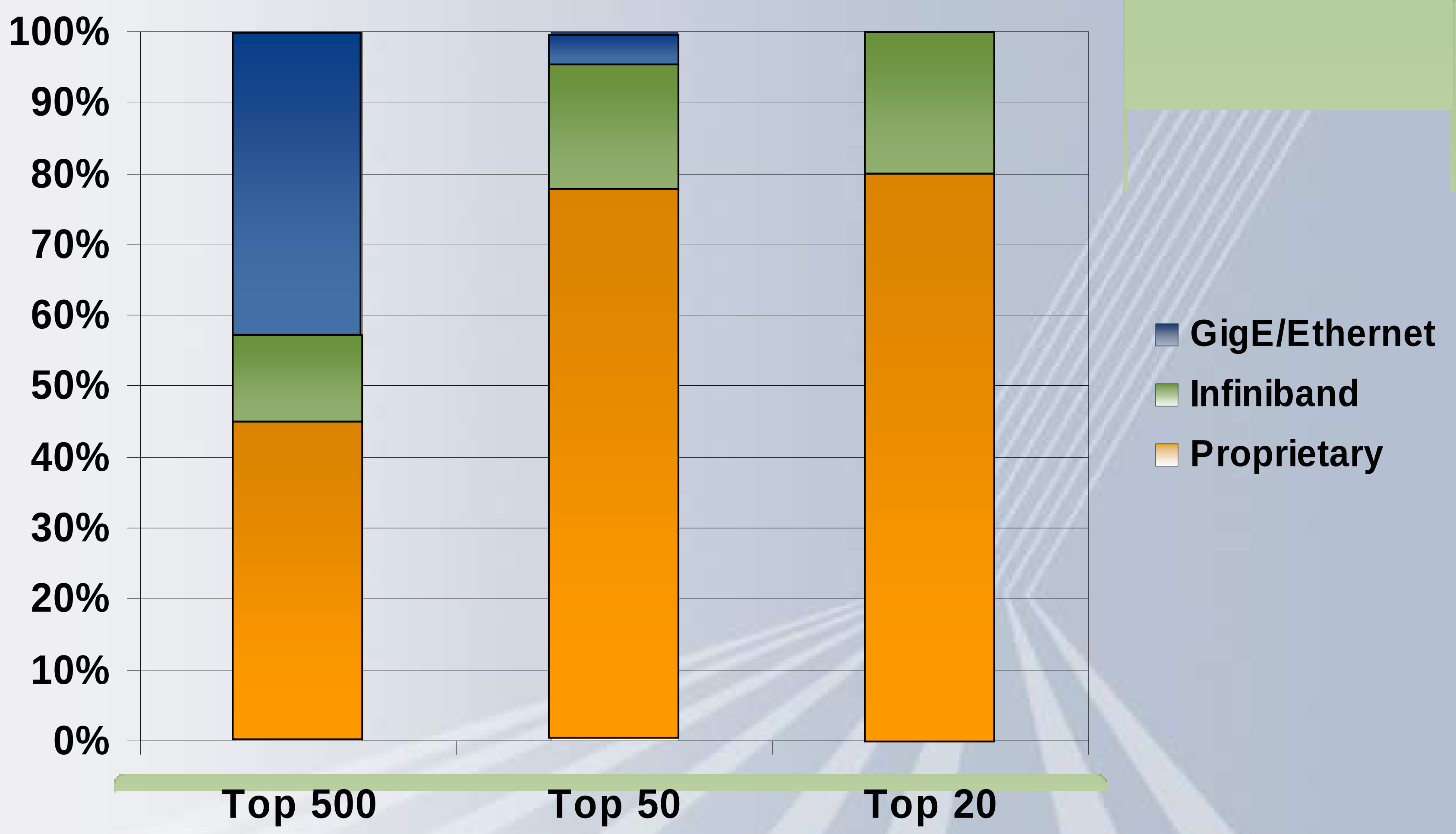
vs.

Cray-o-Wulf

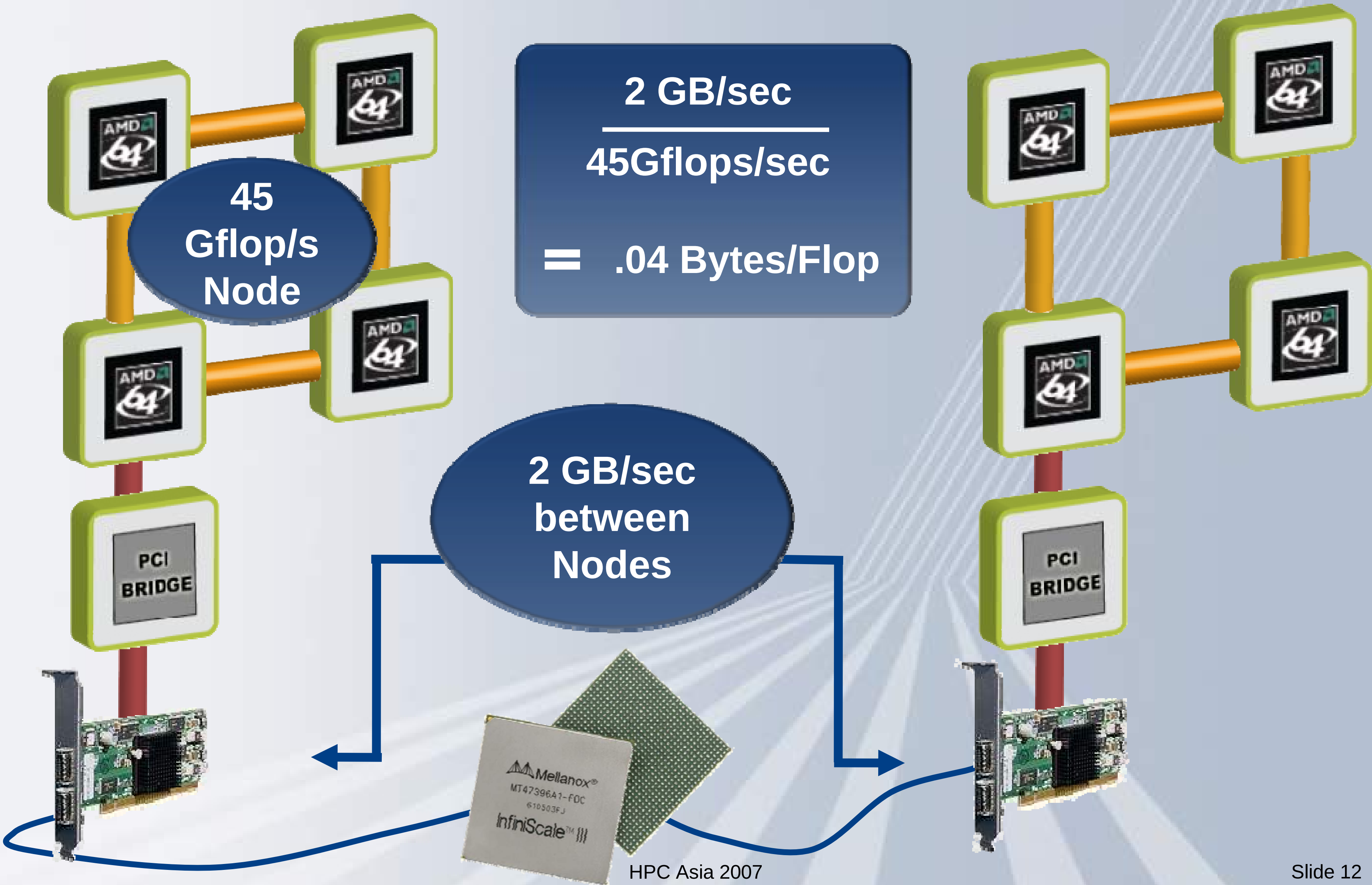
- Current commodity clusters have roughly 250 fans per cabinet
- MTBF for fans alone in a 10-cabinet system is 26 hours

Do We Still Need Custom Interconnects?

Interconnects in the Top 500



Balance Points – 2-node Beowulf

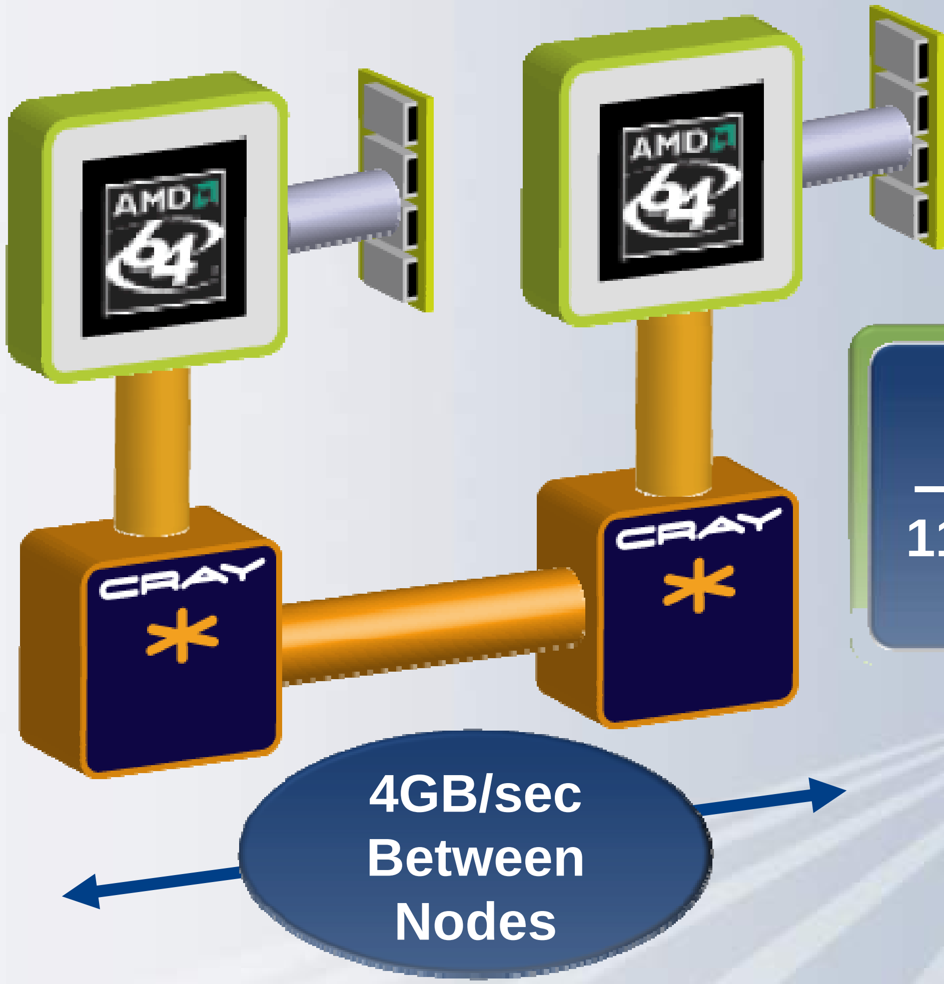


Slide 12

---

m2 miaenno, 5/9/2007

Balance Points – 2-node Cray XT4 Architecture



11Gflops/s per node

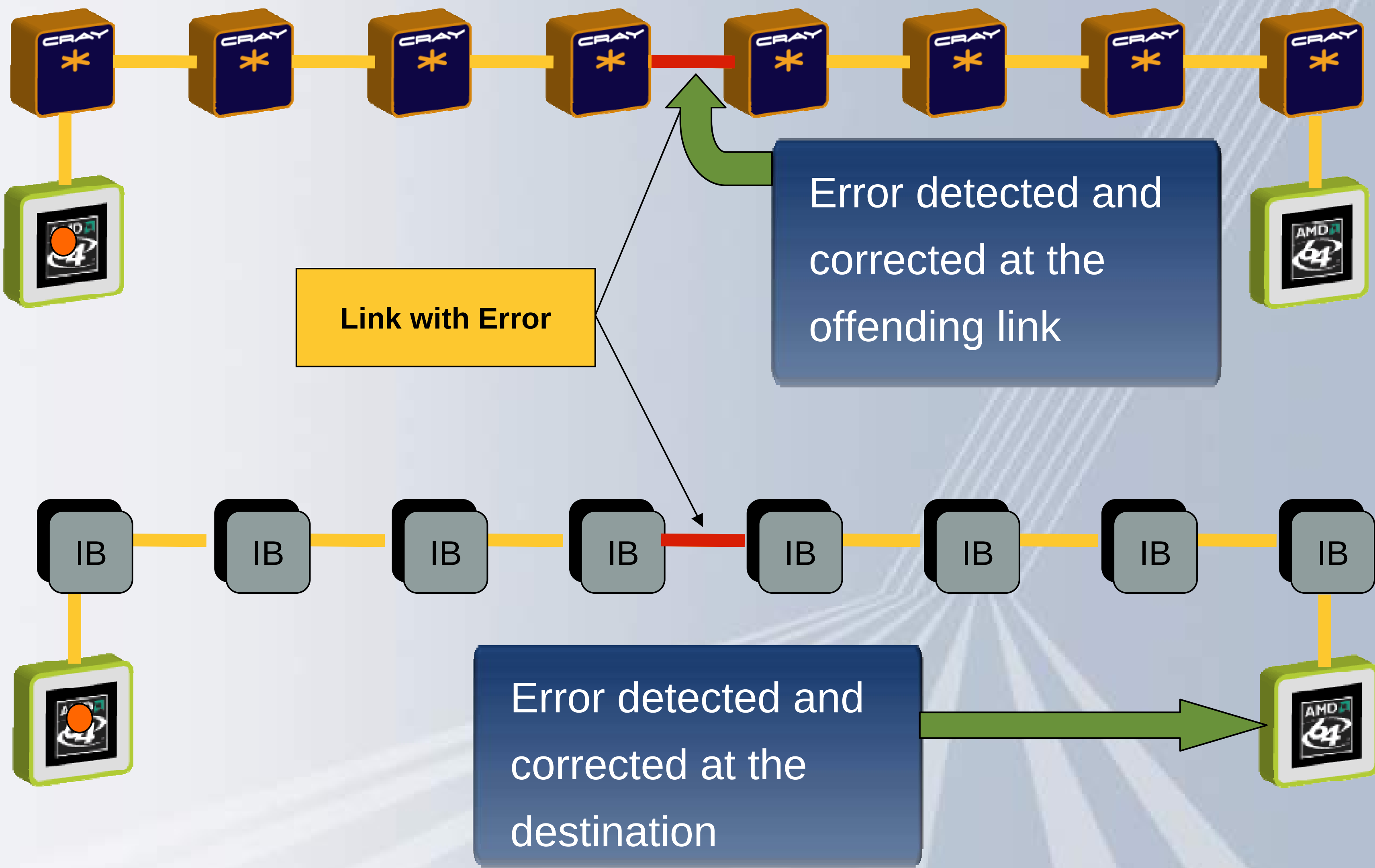
$$\frac{4\ \text{GB/sec}}{11\text{Gflops/sec}} = .36\ \text{Bytes/Flop}$$

10X

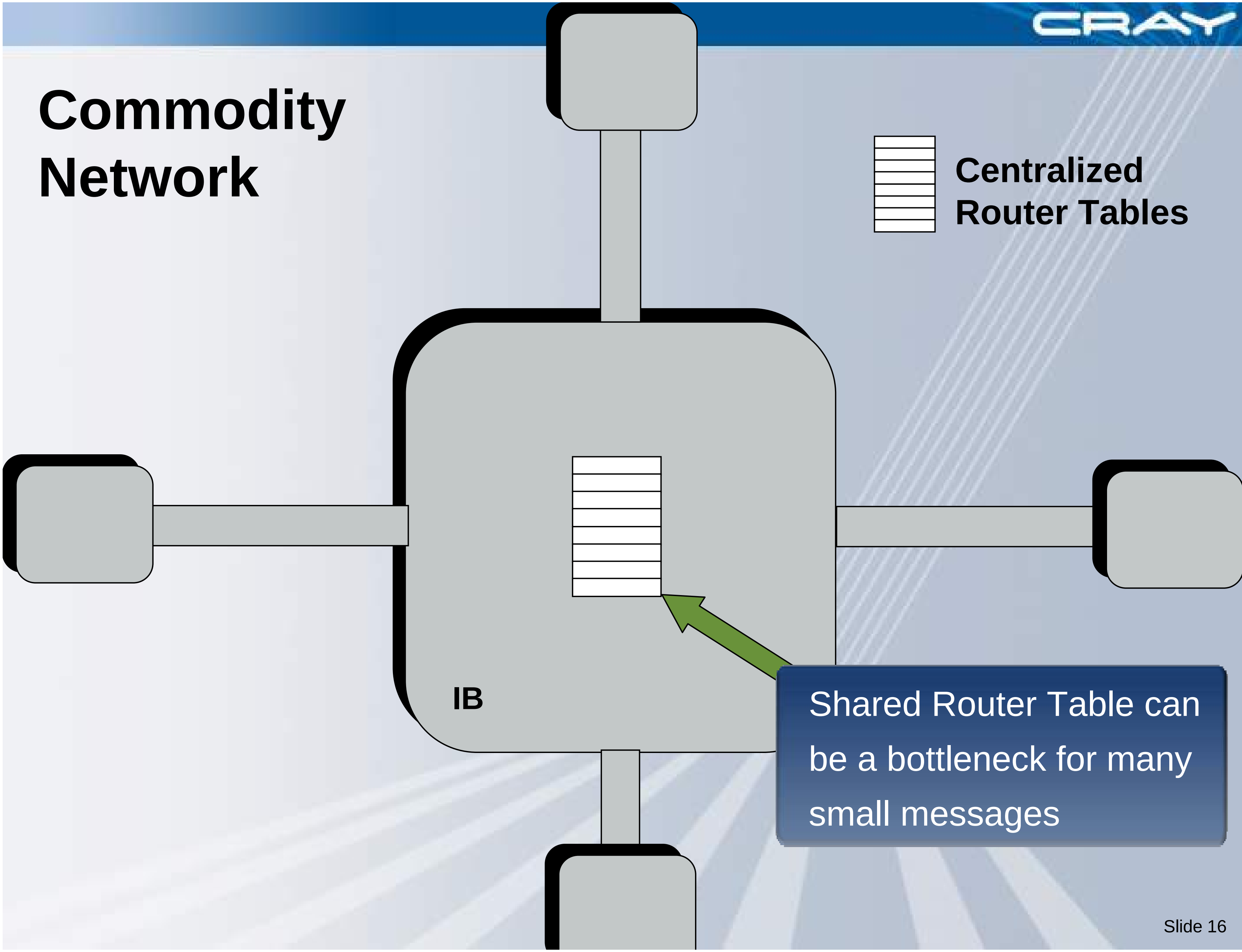
Everything Is Interrelated

- Providing high bandwidth requires many high-speed cables
- Air simply cannot be pushed through the cabinets from front to back
- Packing systems more densely is needed due to cable reach declining
  - ~ 5m at 20 Gb/s
- Bottom to top cooling is necessary
- Liquid cooling could become a requirement

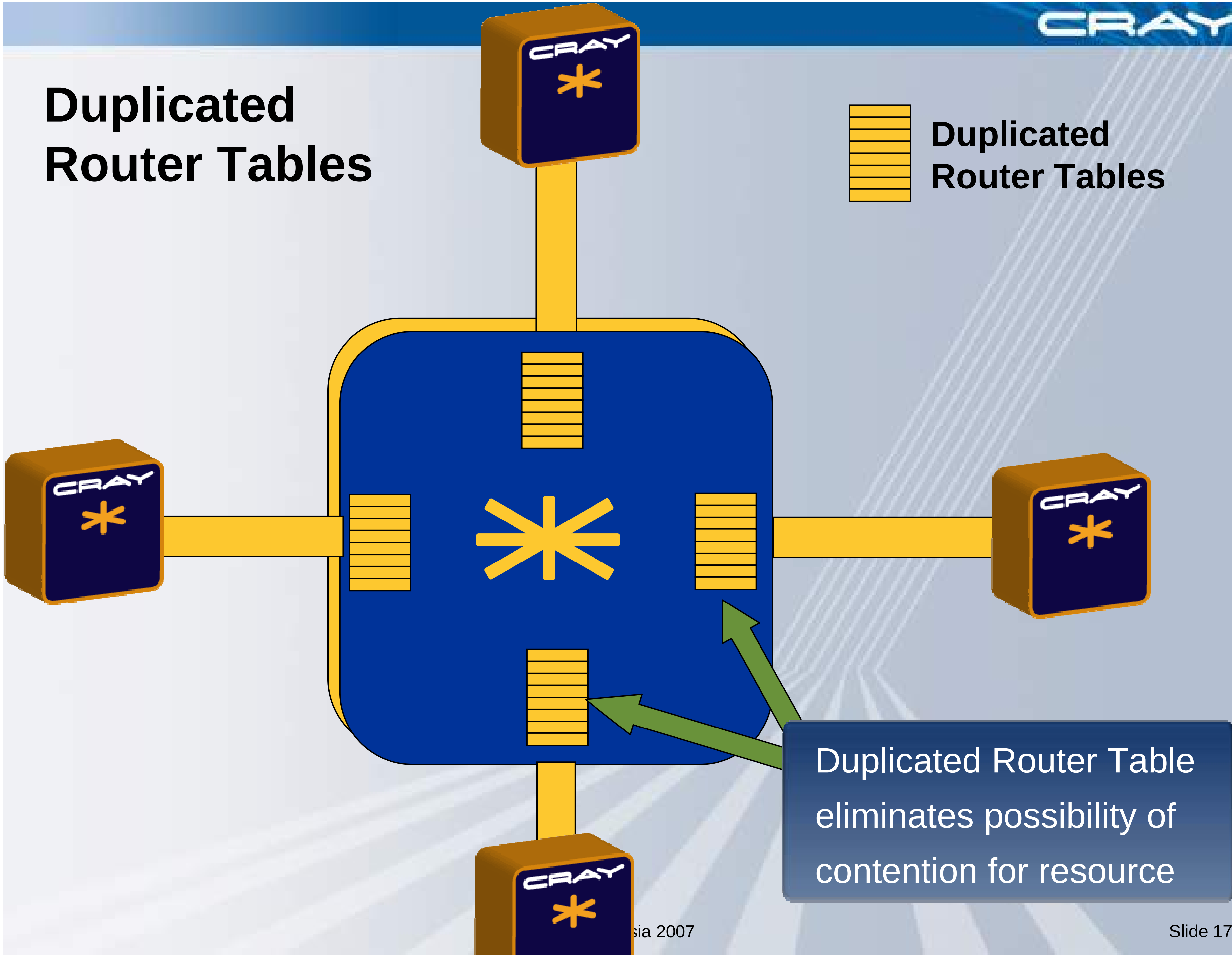
The Importance of Link Level Reliability



Commodity Network

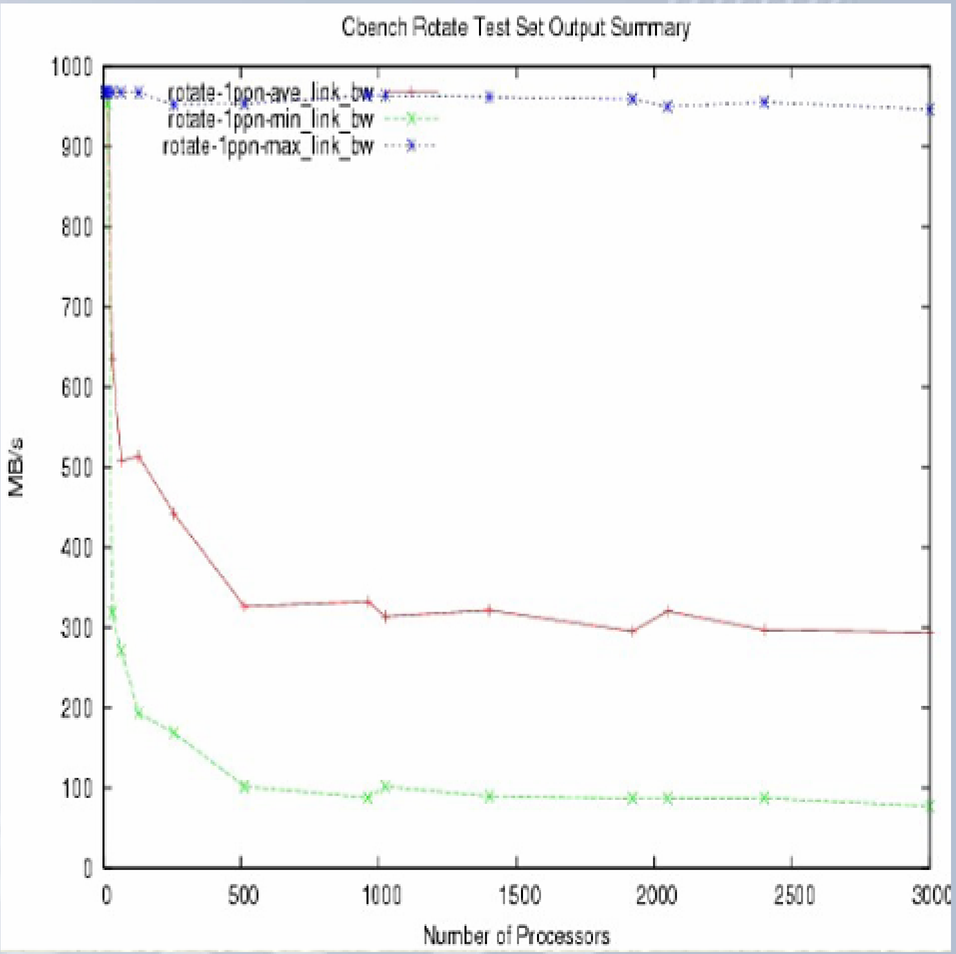


Duplicated Router Tables



Transpose Performance on Large IB Cluster

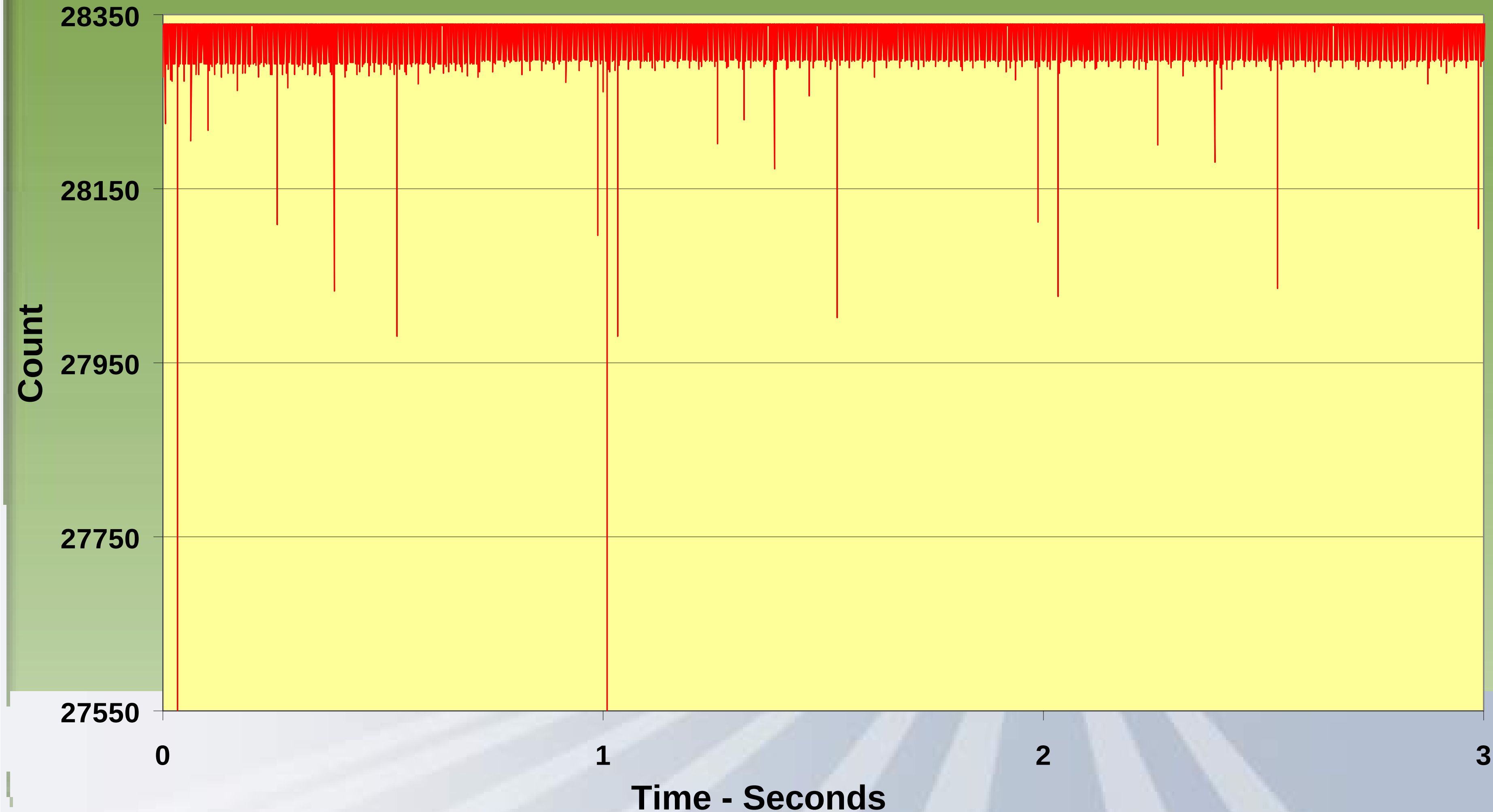
- IB shows a large spread between maximum and minimum performance (almost 10X)
- In MPP computing, we always wait for the slowest processor, so the *minimum* values are more important than the maximums
- Solutions include over-provisioning the interconnect and adaptive routing



Source: Presentation by Matt Leininger & Mark Seager, OpenFabrics Developers Workshop, Sonoma, CA, April 30th, 2007

FTQ Plot of Catamount Microkernel

FTQ Plot of Stock SuSE (most daemons removed)



FTQ plot of CNL

Application Support

- Best in Class MPI
- Best in Class Scientific Library Routines
- Best in Class Performance Tools
- Cray Supercomputing Centers of Excellence to assist researchers in porting/optimizing applications

Realities

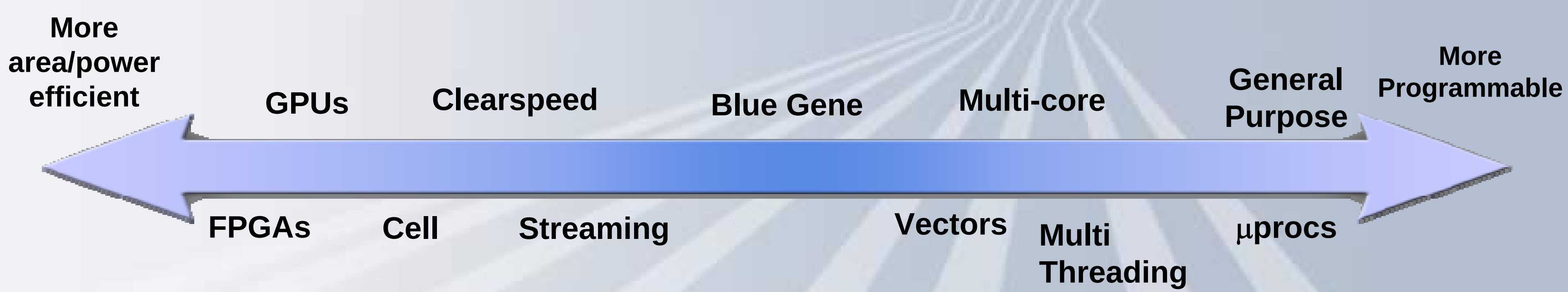
Supercomputing with commodity processors will become almost solely focused on ***scalability***

+

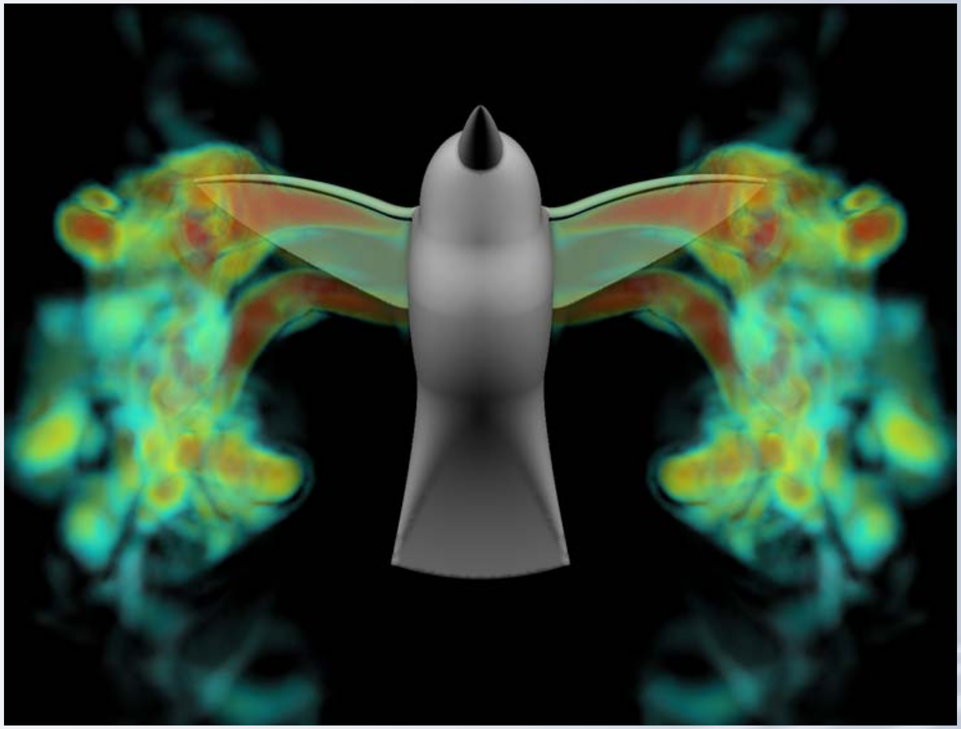
The flattening of the per-core performance trends has renewed interest in ***novel processing architectures*** and ***accelerator*** technologies

Should We Accelerate?

- Slowing single-thread performance may make special-purpose designs more attractive
- Commodity Multi-core processors have issues with memory bandwidth balance and latency tolerance
- There is a trade-off between power efficiency and programmability
- There is no such thing (today), as a general purpose accelerator



Why Vectors? Accelerating challenging memory addressing patterns through global addressing



The Simulation Challenge

- Simulate a flapping wing for development of Unmanned Aerial Vehicle
- Around 5.5 million tetrahedral elements

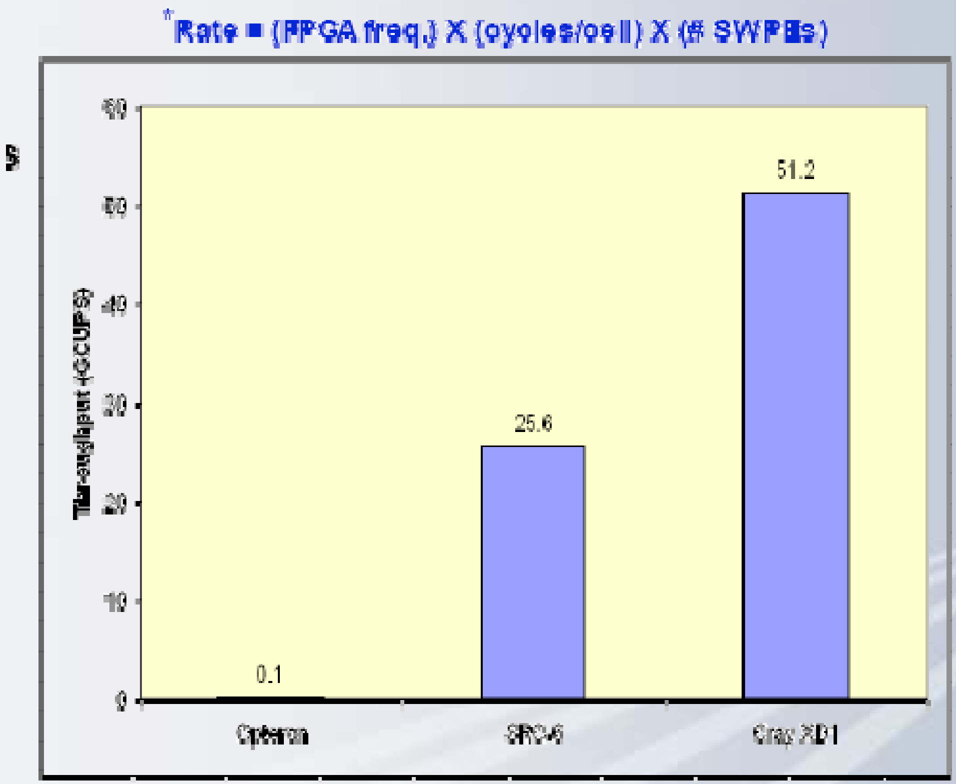
Solution

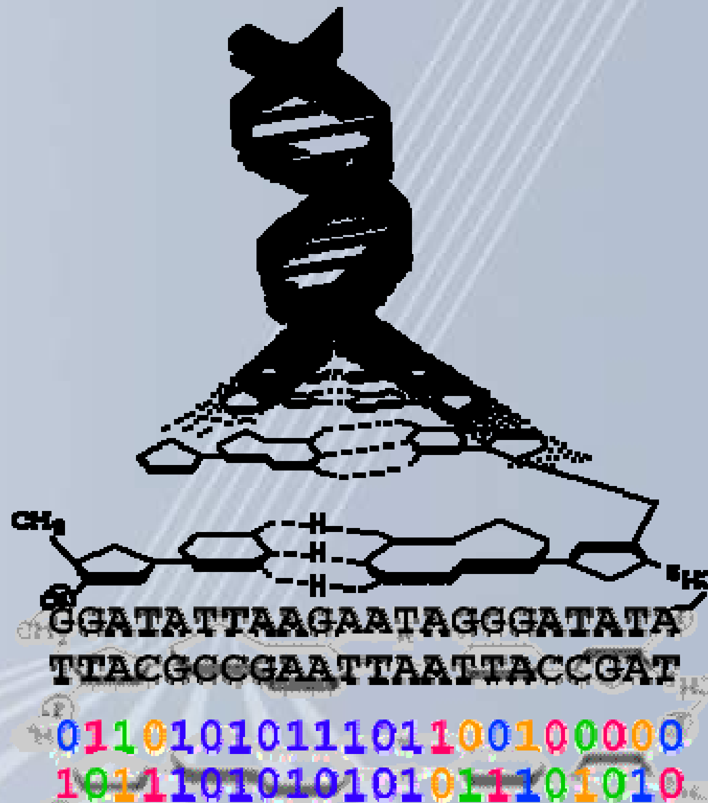
- Need new CFD solution written in modern programming language
  - “XFlow” Dynamic Mesh CFD code
  - Language - Unified Parallel C
- Customer reports largest ever adaptive mesh simulation
- “Could not be programmed in MPI”

New application opens door to modeling whole new realm of simulation and modeling

Why FPGAs? Example: Smith Waterman Search

- Previous FPGA product over 500X faster than Opteron processor





Source: George Washington University

Why Massive Multithreading?

- Driving applications are Informatics Graph-Based algorithms
- Problems of interest are large and require Terabytes of memory to hold
- Problems have no locality and are not partitionable
- Most of these types of problems cannot be coded with the MPI programming model

Case Study: MTA-2 vs. BlueGene/L

- With LLNL, implemented s-t shortest paths in MPI
- Ran on IBM/LLNL BlueGene/L, world's fastest computer

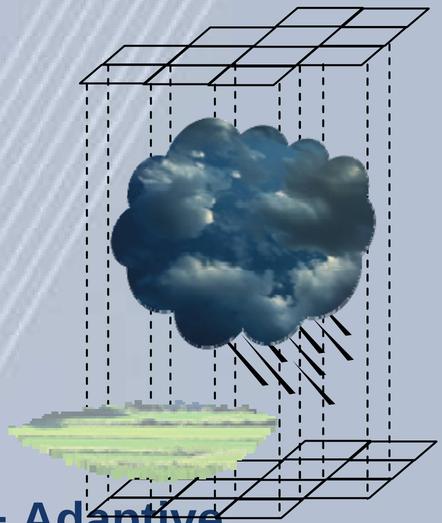
- Finalist for 2005 Gordon Bell Prize
  - 4B vertex, 20B edge, Erdős-Renyi random graph
  - Analysis: touches about 200K vertices
  - Time: 1.5 seconds on 32K processors
- Ran similar problem on MTA-2
  - 32 million vertices, 128 million edges
  - Measured: touches about 23K vertices
  - Time: 0.7 seconds on one processor, 0.09 seconds on 10 procs
- Conclusion: 4 MTA-2 procs = 32K BlueGene/L procs

Hybrid Supercomputing to Adaptive Supercomputing

- Today – Hybrid Supercomputing
  - Multiple processor types in the same system
  - Software to allow them to be easily used and administered
  - Heterogeneous Workflows
- Tomorrow – Adaptive Supercomputing
  - Flexible processor that takes on different “personalities” while operating within a *single code*
    - Scalar
    - Vector
    - Multi-threading
  - Software (compilers & languages) to take advantage of these features

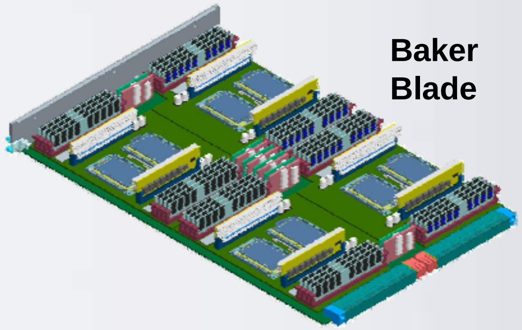
Example Application:
Weather Research & Forecasting (WRF) Model

- **Operational forecasting, environmental modeling, & atmospheric research**
  - Key application for Cray (both vector & scalar MPP systems)
- **Code characteristics:**
  - Most of the code vectorizes really well
    - Dynamics and radiation physics
  - Part of the code is serial
  - Cloud physics is parallel, but doesn't vectorize
    - Little FP, lots of branching and conditionals
    - Vertical columns are all independent
      - $\Rightarrow$ very amenable to multithreading

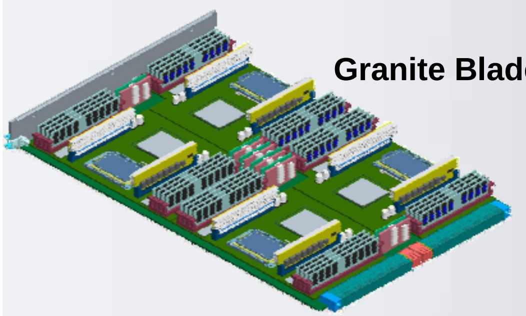


- **Accelerating on Cascade Adaptive Processor (Opteron + Adaptive Vector/Multithreading Accelerator)**
  - Serial code runs on Opteron
  - Vector code runs on accelerator in vector mode
  - Cloud physics runs on accelerator in multithreaded mode
  - Optimal performance on each code segment

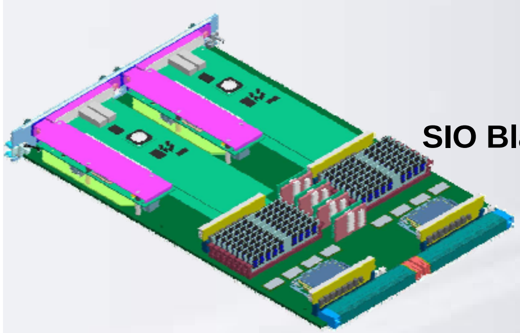
Cascade Packaging



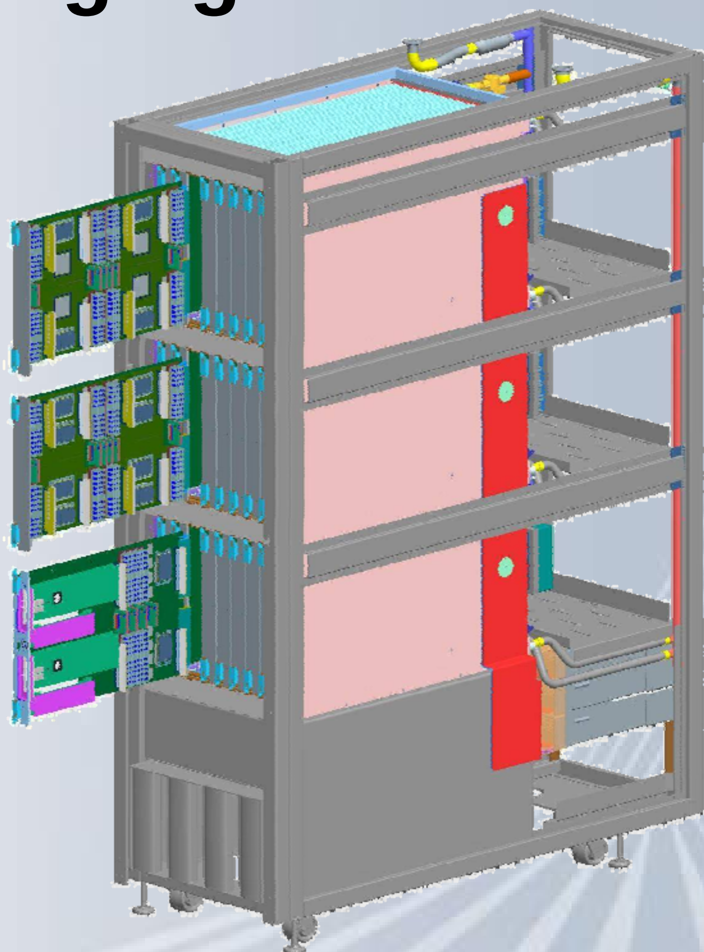
Baker Blade

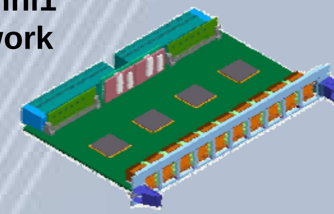


Granite Blade

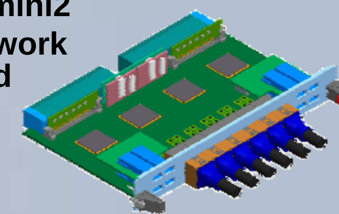


SIO Blade

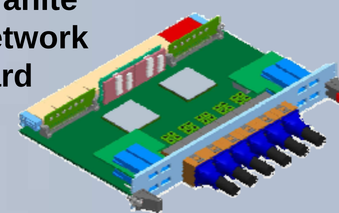




Gemini1 network card



Gemini2 network card



Granite network card

Cascade Cabinet

- Improved TCO due to density, cooling and upgradeability enhancements
- Extensible over multiple years

Summary

- Supercomputing using commodity processors is becoming more and more about *scalability*
- "Beyond Commodity" investment is required in:
  - Reliability & Manageability
  - Interconnect
  - Packaging
  - Software
  - Application Support
- Accelerator technologies are gaining interest
- Today – Hybrid Supercomputing
  - Hetrogenous workflows
- Tomorrow – Adaptive Supercomputing
  - Broad application acceleration

Thank You!