Specializing General-Purpose Computing

A New Approach to Designing Clusters for High-Performance Technical Computing

Win Treese

SiCortex, Inc.

What the heck does that mean?

- High-performance computing often uses specialized hardware
  - Supercomputers
  - experiments with graphics processors
- General-purpose computing doesn't optimize for technical computing

With some problems...

Supercomputers

- Expensive
- Not on the same technology curve
- Different programming environment

General-purpose computing

- Amazing technology curve
- Optimized for desktop and enterprise applications

A Challenge: The Best of Both

- Use general-purpose hardware components
- With a standard programming environment
- And SYSTEM DESIGN for technical computing

The Roadmap

- A bit of history
- A bit about high-performance technical computing (aka "HPTC")
- Linux clusters for HPTC
- Designing a new system for HPTC
- What we are building

A Bit of History

The SUPERCOMPUTER

But all is not well in supercomputer land...

- You have to pay a lot for them
- You have write your program differently
- You have to find some high priests to take care of them
- Supercomputer companies don't make money

...so let's use lots of little computers

- PCs are cheap
- Linux is free
- Commodity interconnect (Ethernet) is cheap

The (Beowulf) Cluster is born

A Small Visualization Cluster

Some characteristics of high-performance technical computing

Some typical applications

- Climate and weather models
- Geophysics
- Complex financial modeling
- Mechanical design
- Finite element analysis
- Fluid dynamics
- Life sciences analysis and simulation
- Top-secret stuff
- ...and many others

What are they like?

- Can run for weeks
- Consume all the cycles you can afford
- Not very cache-friendly
- Parallelism often demands good communications
- Large data sets (input and output)
- Many are in Fortran!
- ...but also in C, C++, Java, Perl, Python, etc.

The Market for HPTC

- HPTC is now mainstream computing!
- Over $6 billion in Linux cluster hardware sales in 2006
- Petascale computing is hot for research, but there is a real market now for teraflops

Linux Clusters
and
High-Performance Technical Computing

So clusters are great, right?

- Cheap, because they use cheap PCs
- Expandable
- Easy to get started
- Software is free
- They ride the desktop/ server technology curve
- Interconnect (Ethernet) is cheap
- Emerging de facto standards
  - Linux
  - Message Passing Interface (MPI)
  - C, Fortran, etc.

...but not perfect

- Computational efficiency is often low
- Use lots of power
- Generate lots of heat
- Many parts to fail
- ...with a desktop MTBF design
- Interconnect is slow: XXX microseconds for MPI on Ethernet
- ...or expensive: using Infiniband can increase the price of a node by 50%

And software rules!

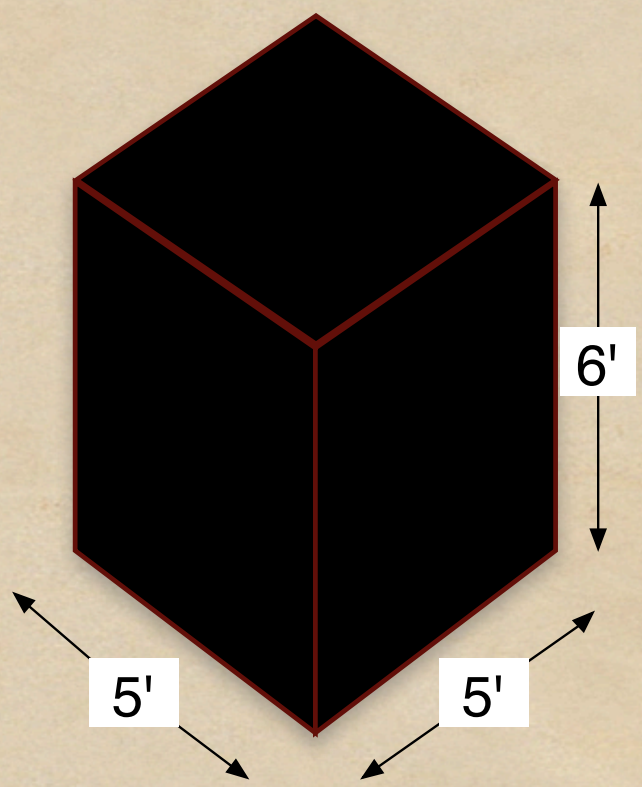
- Software investment is the significant cost
- Replace the cluster, but keep the software

What if we redesign the system with the same programming interface?

Designing a New System for High-Performance Technical Computing

A Design Challenge

- 1000 nodes in this box
- ...all running Linux
- Near-microsecond MPI latency
- Air-cooled

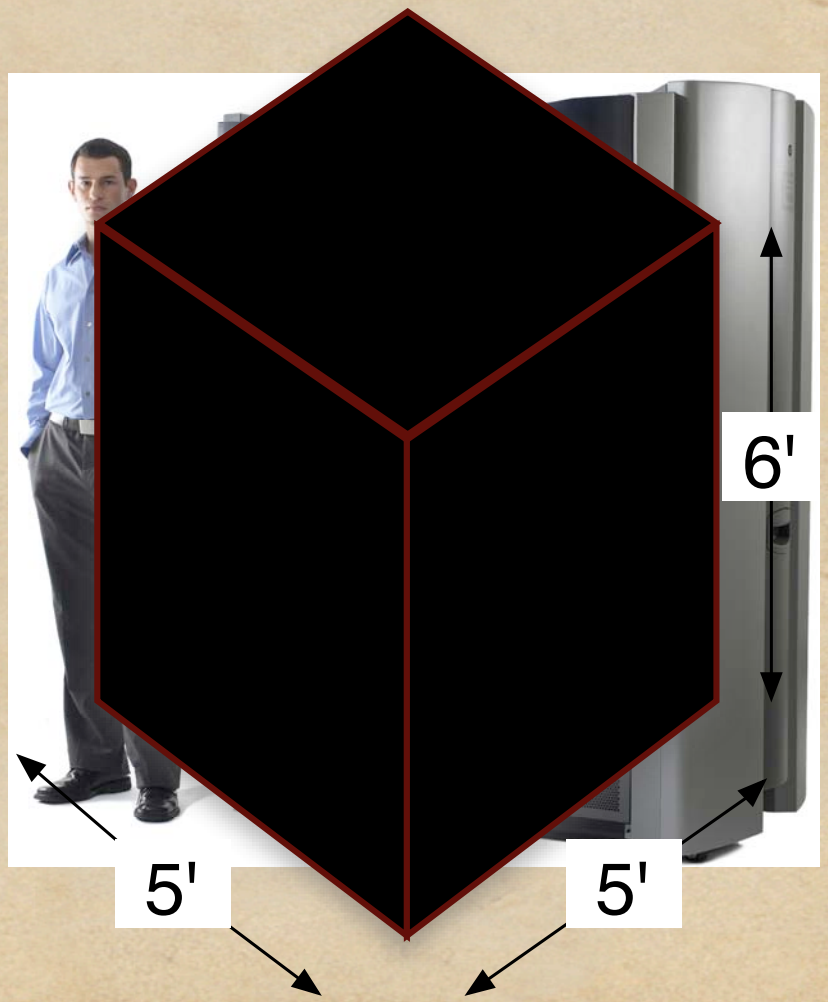


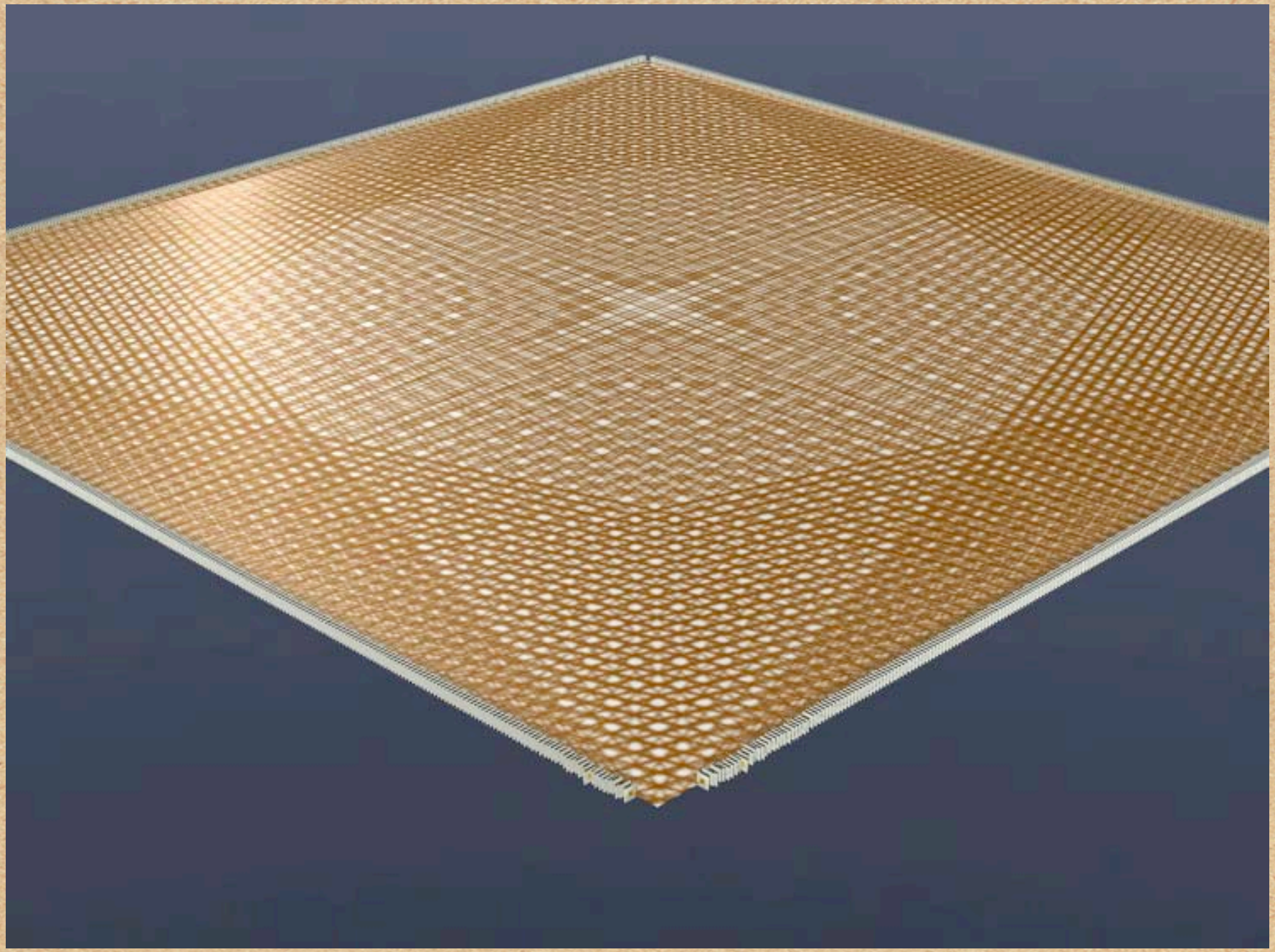
The logic of low power

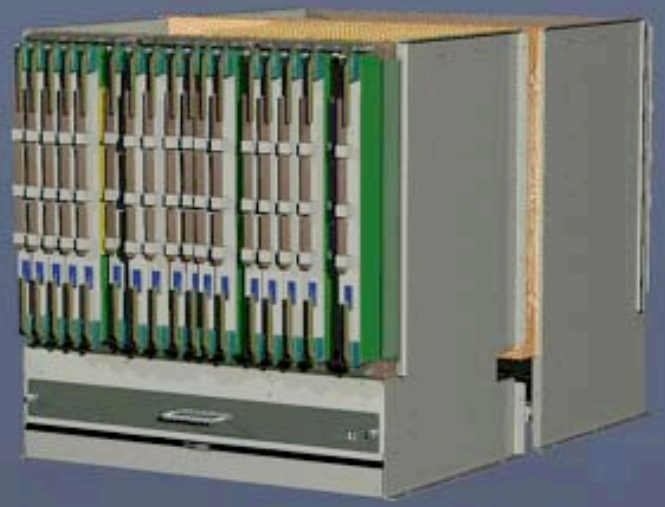
- Low power ⇒ less heat
- Less heat ⇒ parts closer together
- Parts closer together ⇒ shorter wires ⇒ easier high-performance interconnect
- Less heat ⇒ greater reliability
- Burn less power waiting for memory

The SC5832

- 5832 Gigaflops
- 7776 Gigabytes ECC memory
- 972 6-core 64-bit nodes
- 2916 2 GByte/s fabric links
- about 1 microsecond MPI latency
- 108 8-lane PCI-Express
- 18 KW
- 1 Cabinet







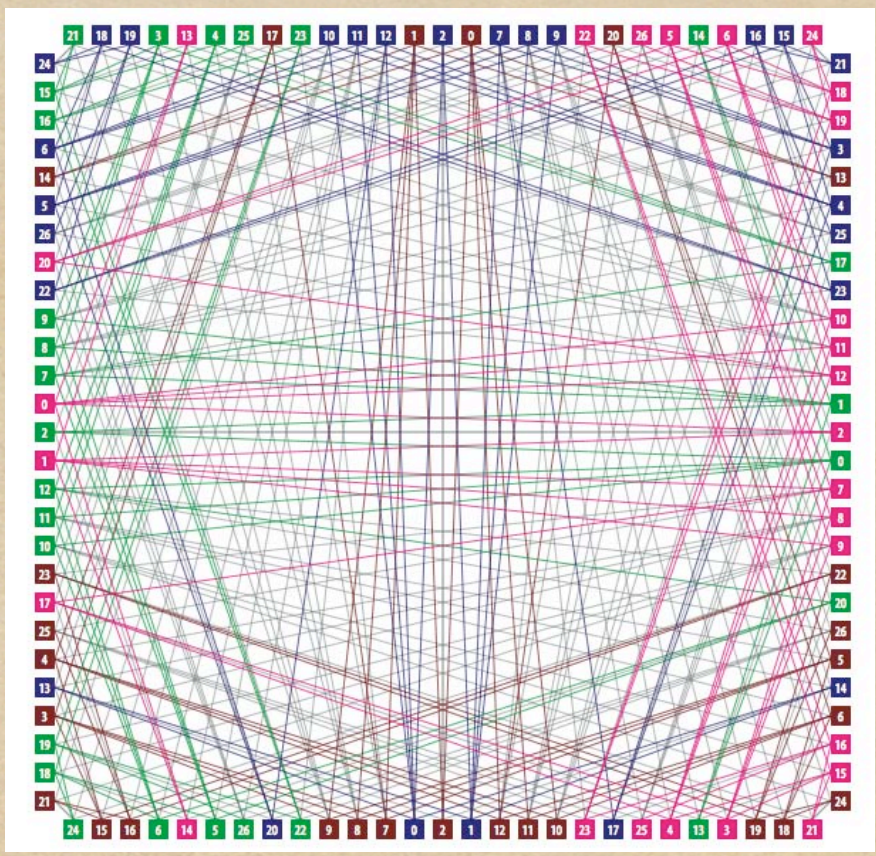
The SC648

- 648 Gigaflops
- 864 Gigabytes ECC RAM
- 108 6-core 64-bit nodes
- 324 2 GB/s fabric links
- about 1 microsecond MPI latency
- 12 8-lane PCI-Express
- 2 KW
- 1/2 standard 19" rack

Software

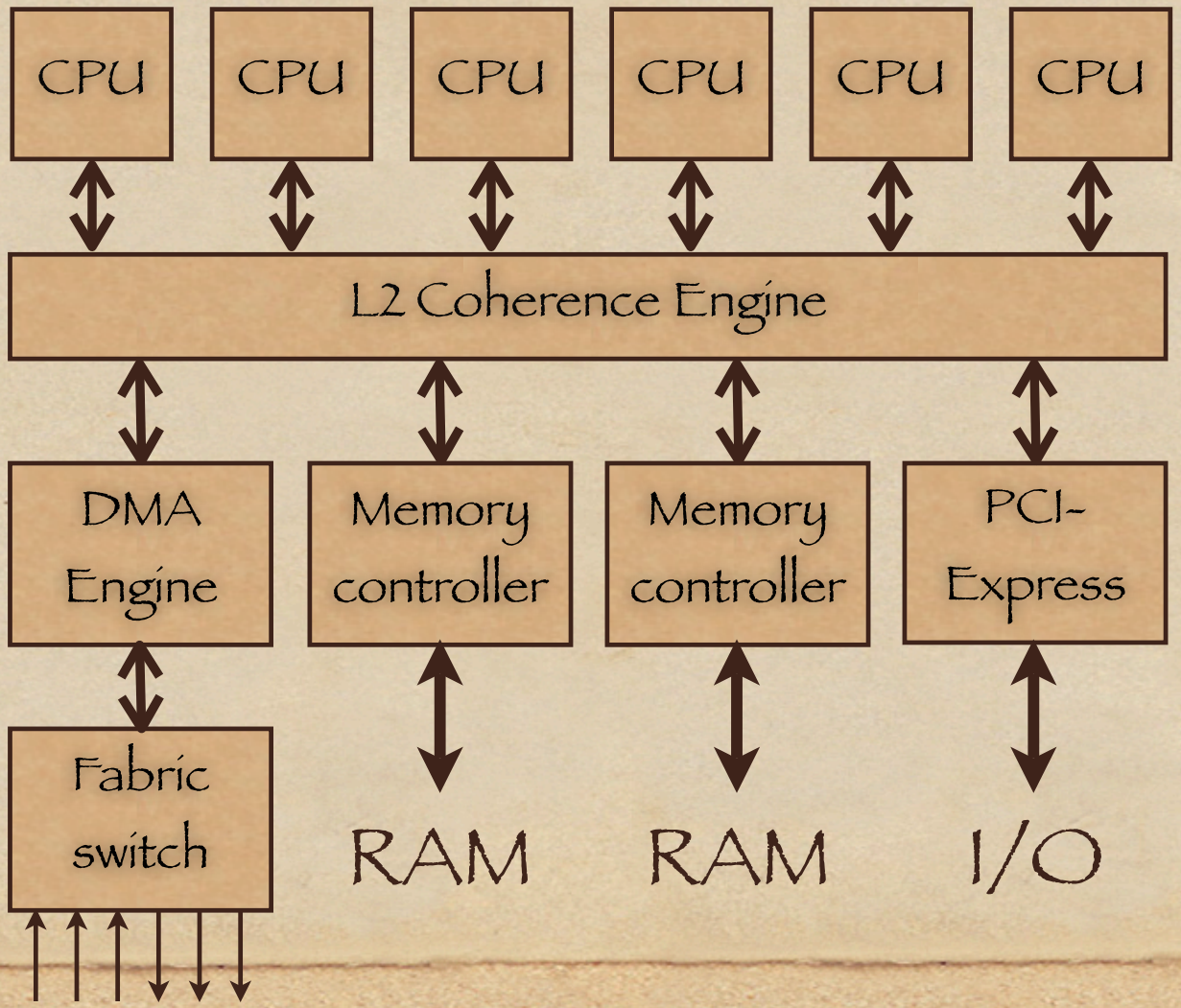
- It's just Linux
- gcc
- MPI
- etc.
- ...even Emacs!
- All open source

Interconnect fabric

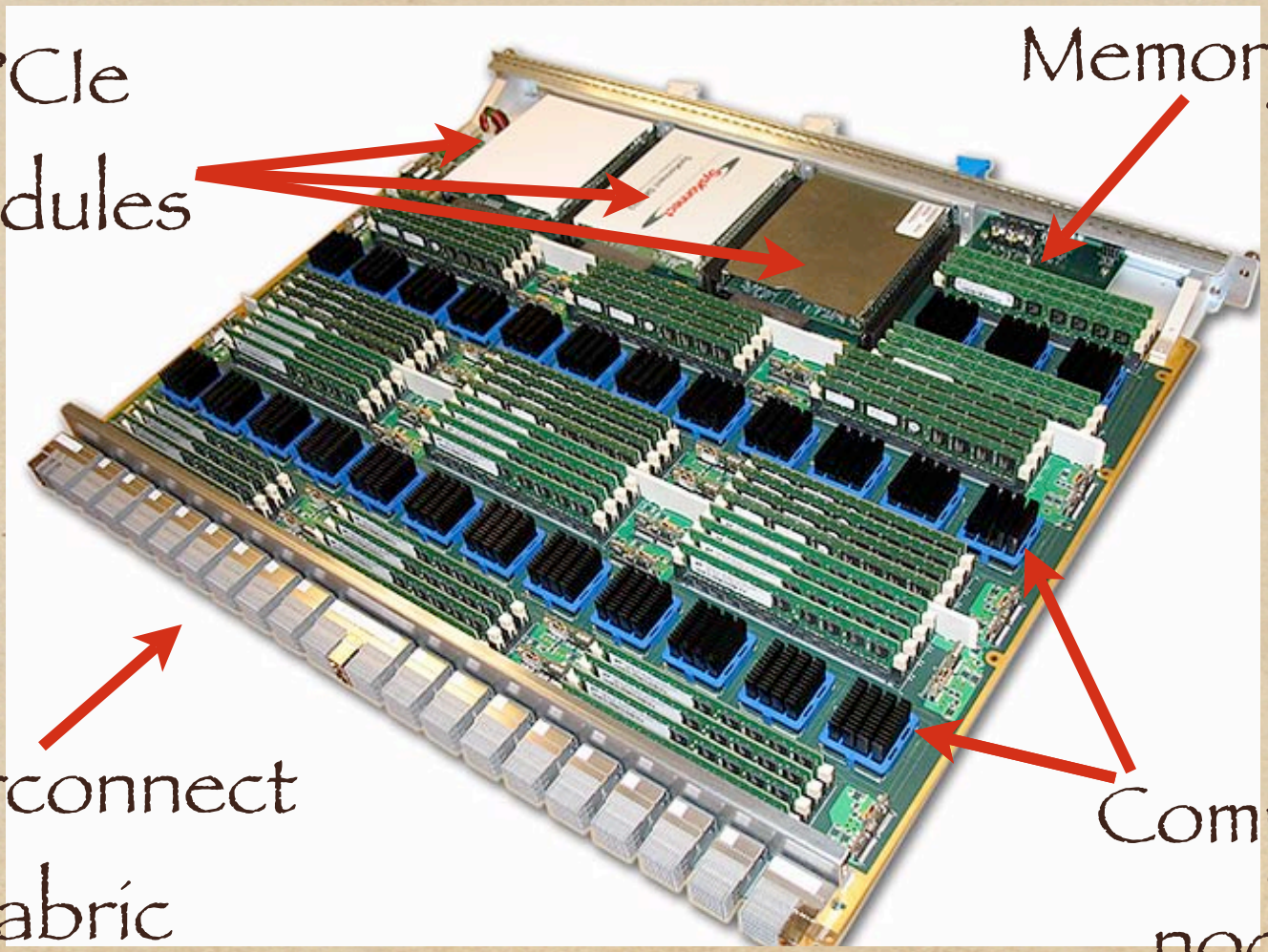


- Log diameter
- Multiple paths
- Cost-effective

A Cluster Node Chip



27-Node Module



Design for reliability

- Lower parts count
- Lower power = less heat = less stress
- All RAMs have ECC
- Redundancy in interconnect

Parallel I/O

- Integrated Lustre cluster filesystem
  - Open source
  - POSIX-compliant
- Multiple uses
  - Direct-connect storage
  - External Lustre servers
  - RAM-based filesystem

What have we learned?

- Take general computing techniques
- ...with some knowledge about the applications
- Mix well

Powerful and usable computing

Specializing General-Purpose Computing

Win Treese

SiCortex, Inc.

win.treese@sicortex.com

or

treese@acm.org