When Multicore Isn't Enough: Trends and the Future for Multi-Multicore Systems

Matt Reilly
Chief Engineer
SiCortex, Inc

SiCortex

The Computational Model

- For a large set of interesting problems (N is number of independent processes)

$$T_{sol} = T_{arith}/N + T_{mem}/N + T_{IO} + f(N)T_{comm}$$

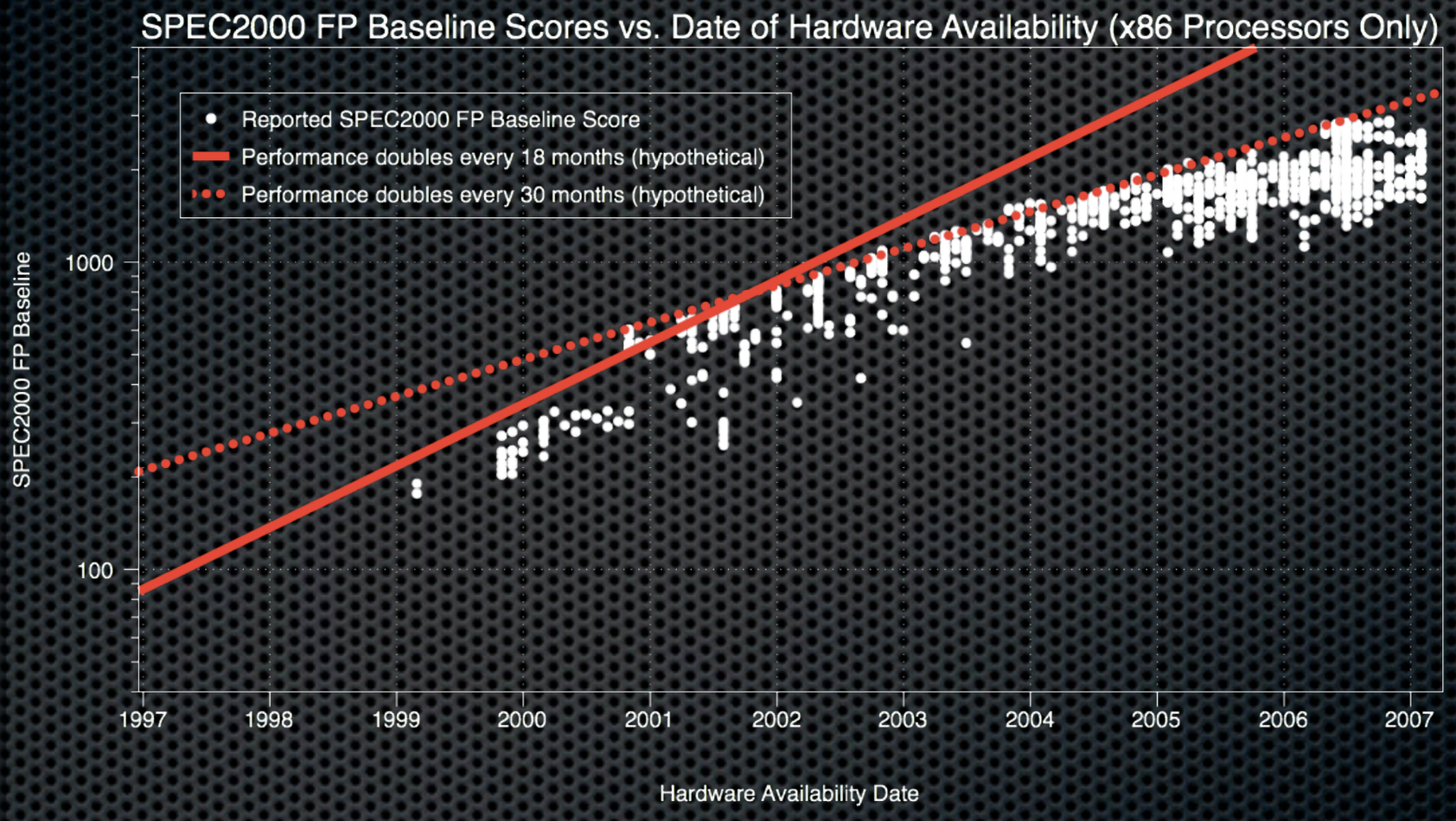
or

$$T_{sol} = MAX(T_{arith}/N,\ T_{mem}/N,\ T_{IO},\ f(N)T_{comm})$$

- For many interesting tasks, single CHIP performance is determined entirely by T_{mem} and memory bandwidth.

SiCortex

Why Multicore?



We don't get faster cores as often as we get more of them

Source: SPEC2000 FP Reports
http://www.spec.org/cpu/results/cpu2000.html

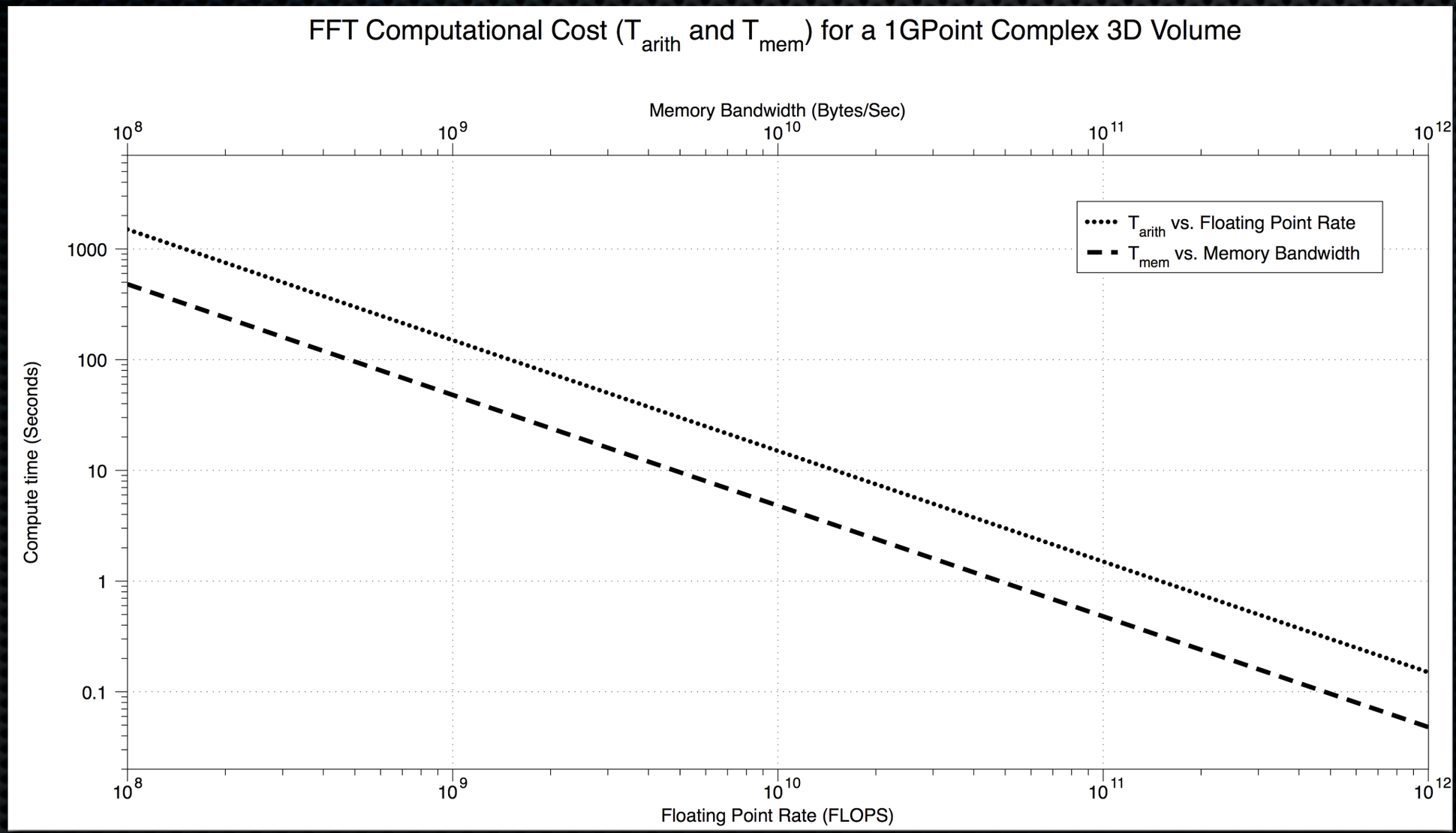
SiCortex

Compute Node Design: A Memory Game

- T_{arith} is becoming irrelevant. (Because N is getting large.)
- The design of the compute node is all about maximizing usable bandwidth between the compute elements and a large block of memory
  - Multicore
  - GPGPU
  - Hybrid/ScalarVector (*e.g.* Cell)
- The architecture choice drives the programming model, but all are otherwise interchangeable.

SiCortex

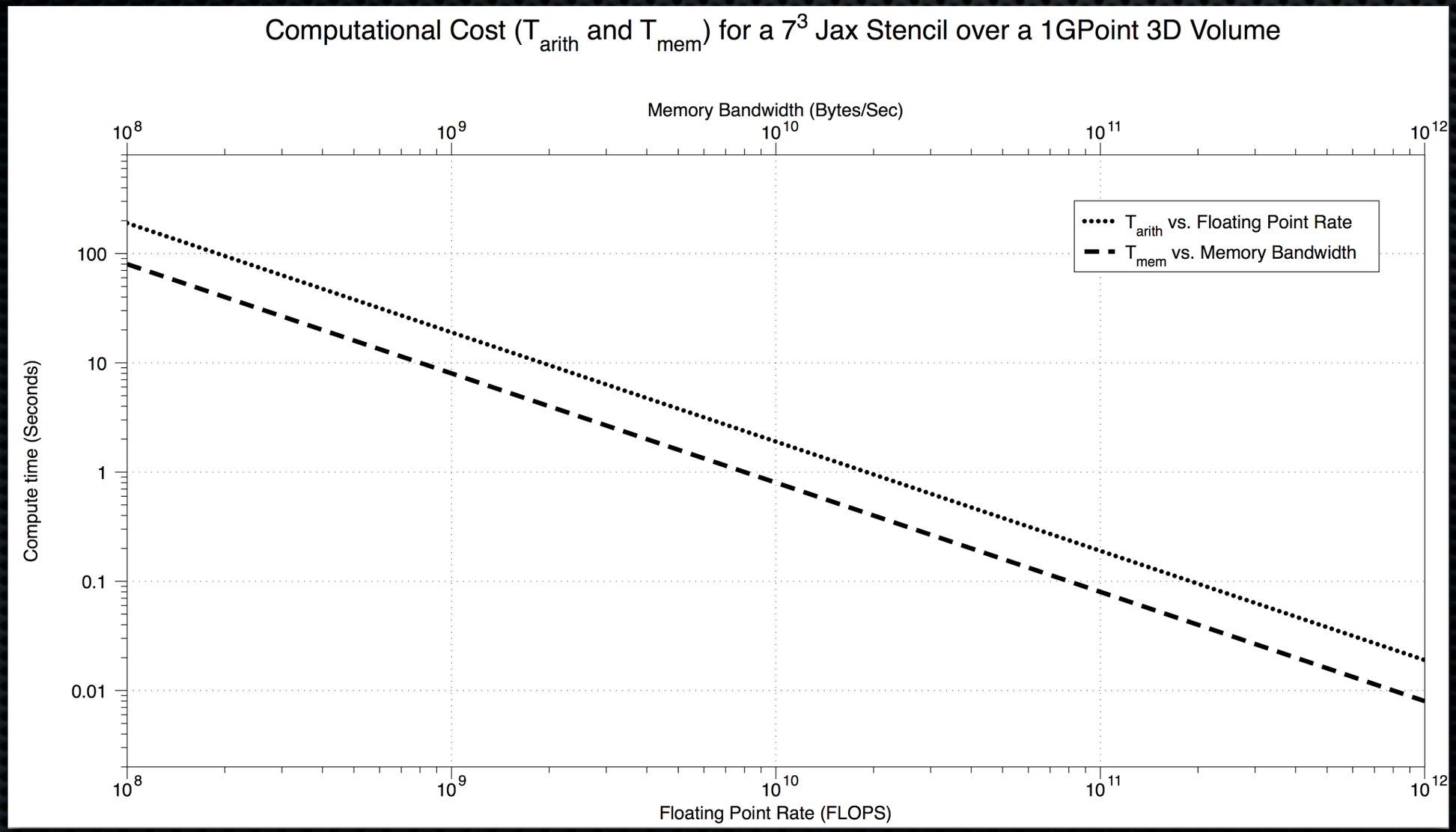
FFT Kernel



If Arithmetic is free, but pins are limited...

SiCortex

Stencil (Convolution) Kernel



0.4 Bytes/FLOP?

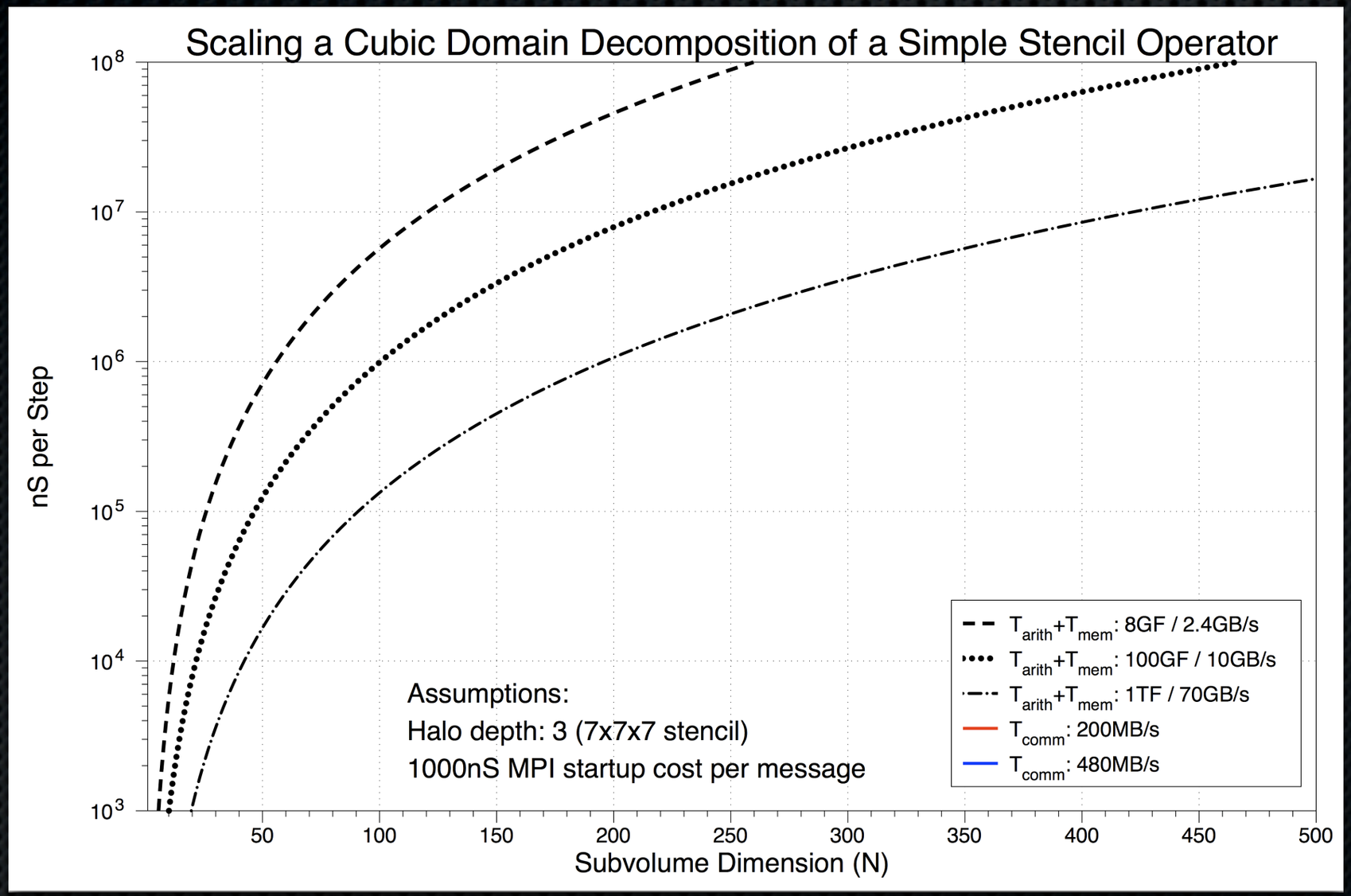
Then the processor spends 1/2 time waiting.

SiCortex

Alternatives

- Many fast cores on one die
  - Require commensurate memory ports: high pin count
  - High pin count and high processor count: large die
- A few fast cores on one die
  - Better balance T_{arith} : T_{mem}
  - Smaller die
- A few moderate cores on one die
  - Balance T_{arith} : T_{mem} : T_{comm}
  - Spend pins on other features

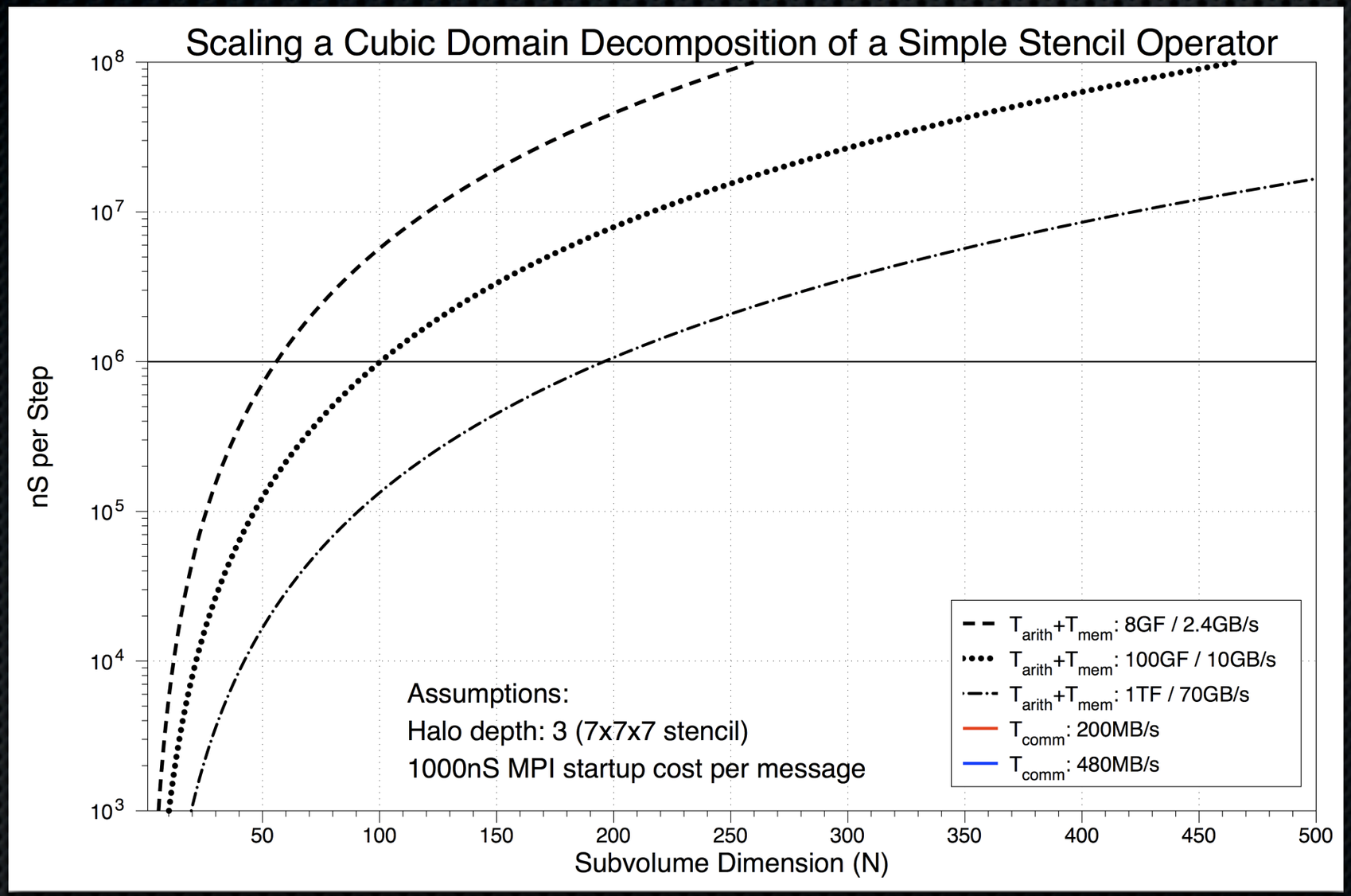
Cubic Domain Decomposition



Simple 7x7x7 “Jax” stencil operator over a large volume (1K^3 single precision): 19 flops per point.

SiCortex

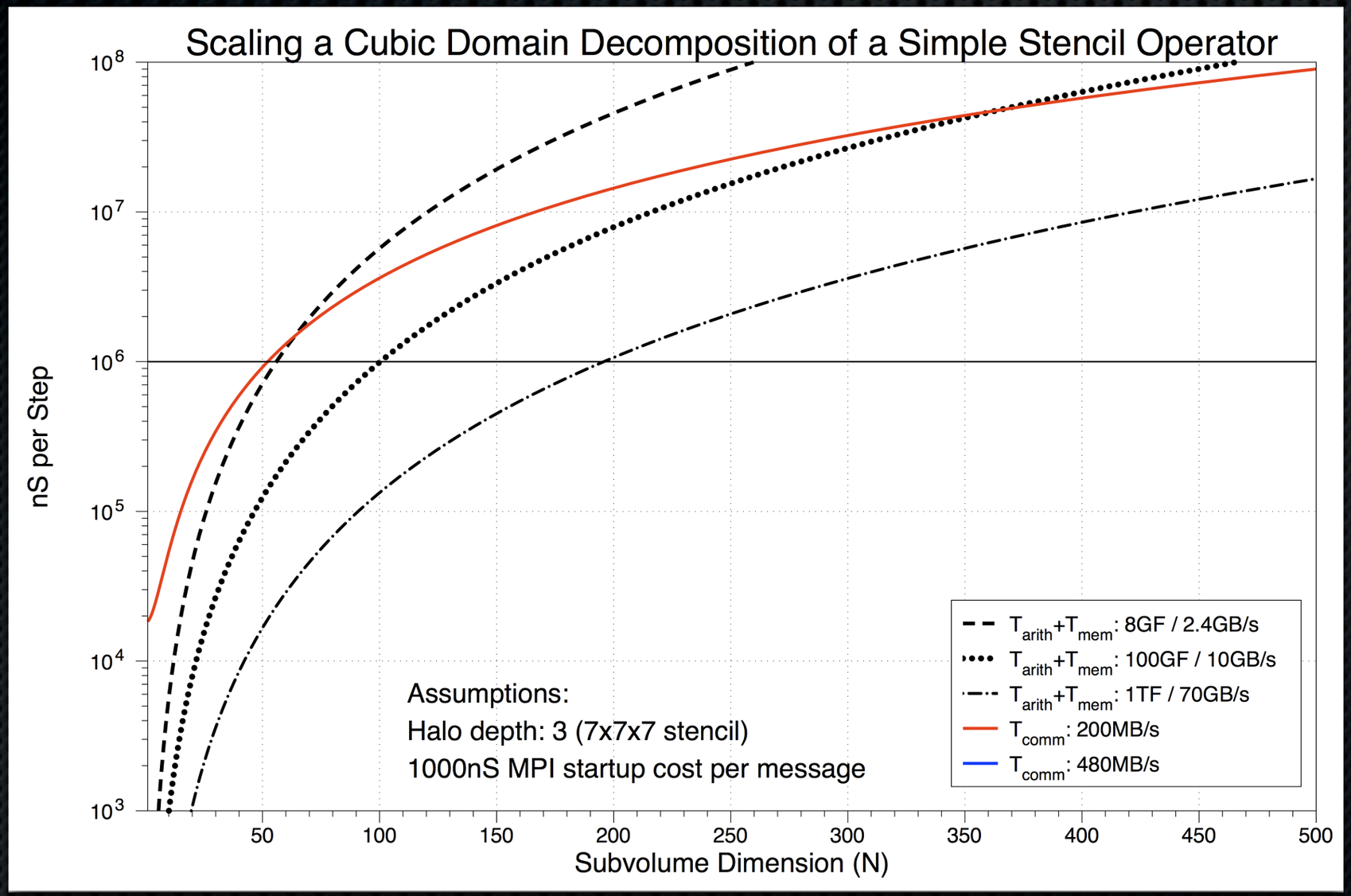
Cubic Domain Decomposition



Set a goal of completing a pass in 1mS.
Faster processors complete larger chunks of the total volume.

SiCortex

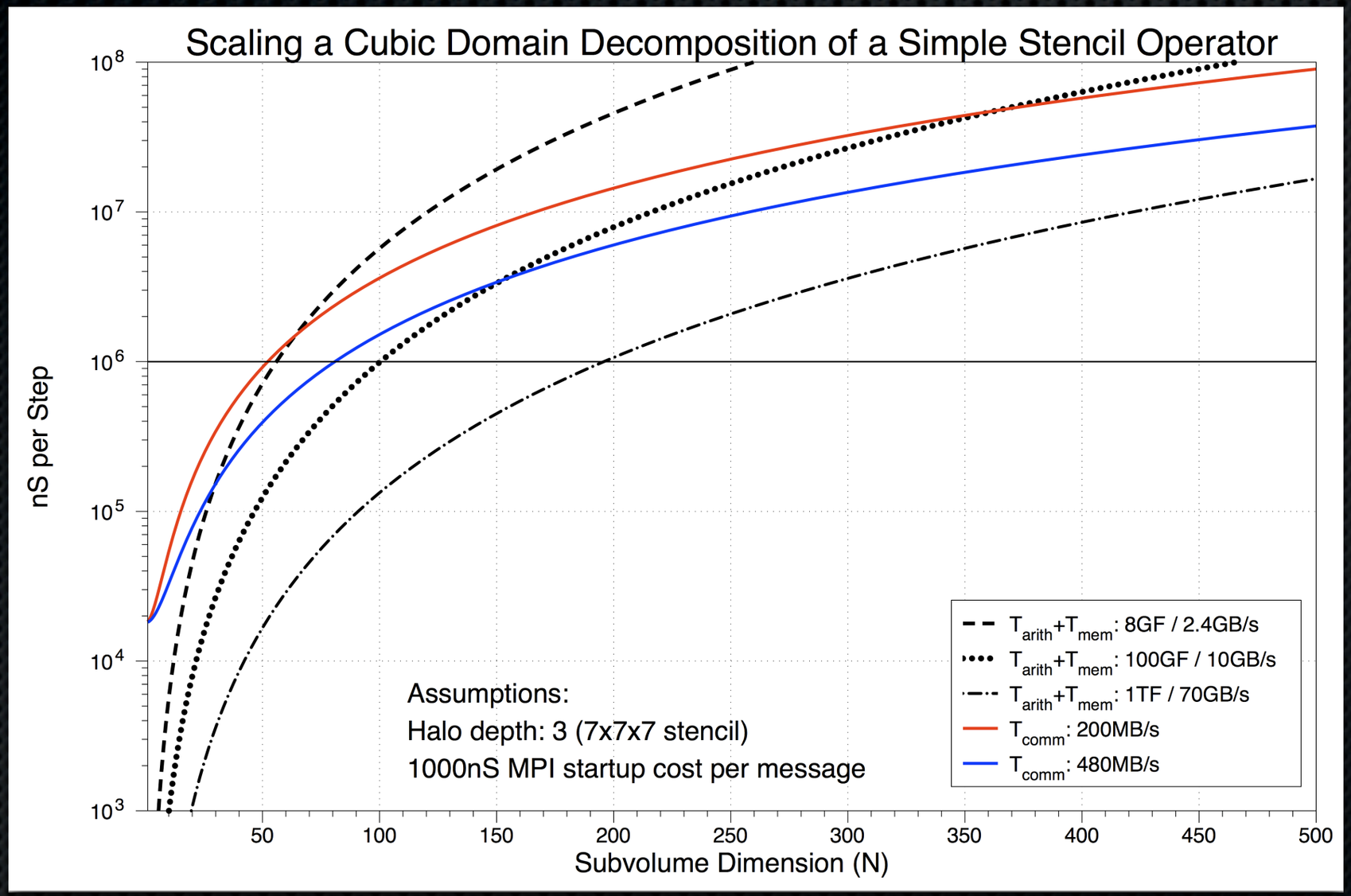
Cubic Domain Decomposition



Factor in T_{comm} and we find that a 200MB/s per-node link forces a chunk size of 50^3.

SiCortex

Cubic Domain Decomposition



If the goal is “time per step,” computation speed may not matter.

GPUs, FPGAs, Magic Dust, don’t help.

SiCortex

The Systems

A Family: From Production Systems to Personal Development Workstations

SiCortex

The SC5832

- 5832 Processors
- 7.7TB Memory
- > 200 FC I/O Channels
- Single Linux System
- 16KW
- Cool and Reliable

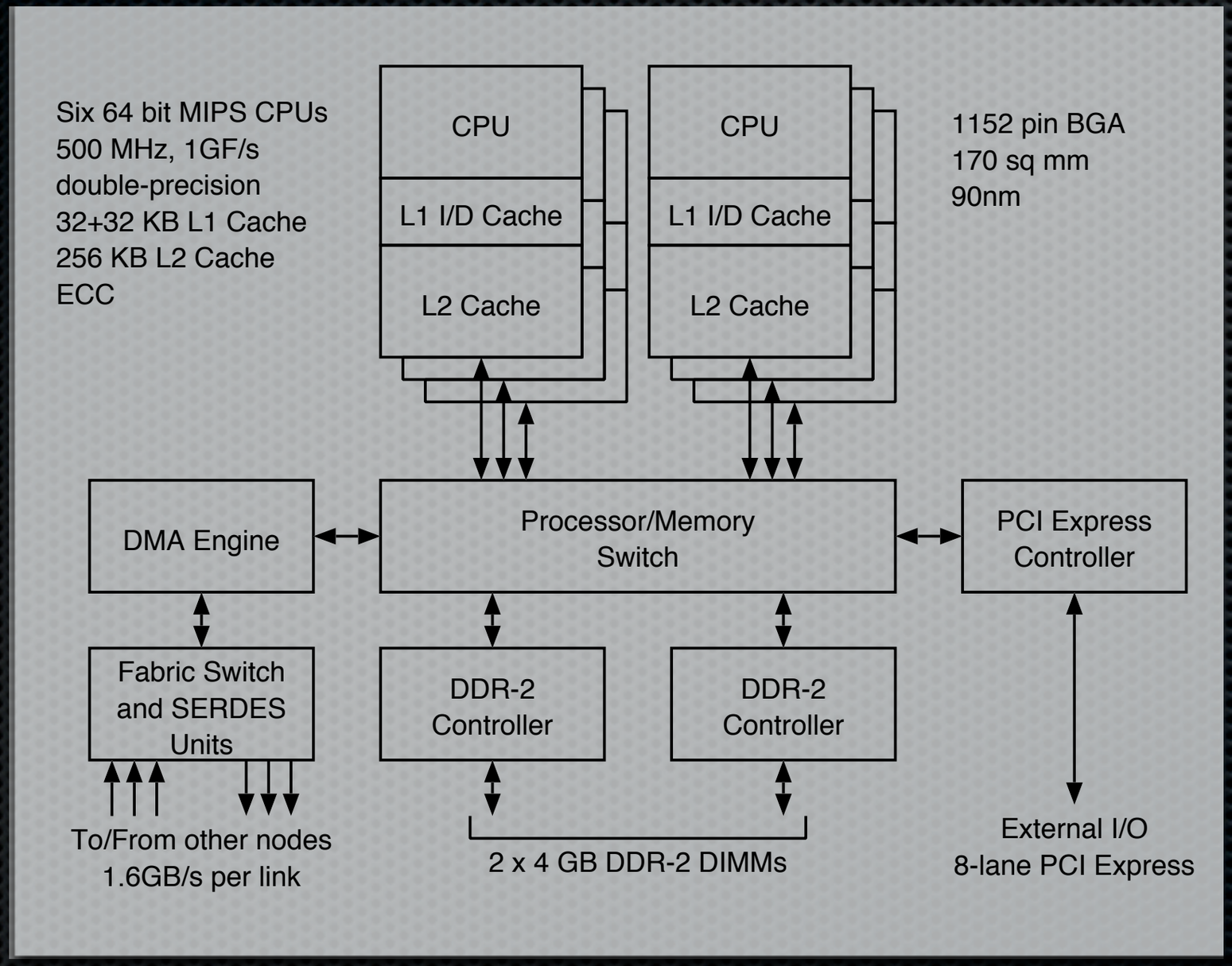
SiCortex

SiCortex in the Technical Computing Ecosystem

- Affordable, easy-to-install, easy-to-maintain
- Development platforms for high processor count applications
  - Rich cluster/MPI development environment
  - Systems from 72 to 5832 processors
- Production platforms in target application areas:
  - Multidimensional FFT
  - Large Matrix
  - Sorting/Searching

SiCortex

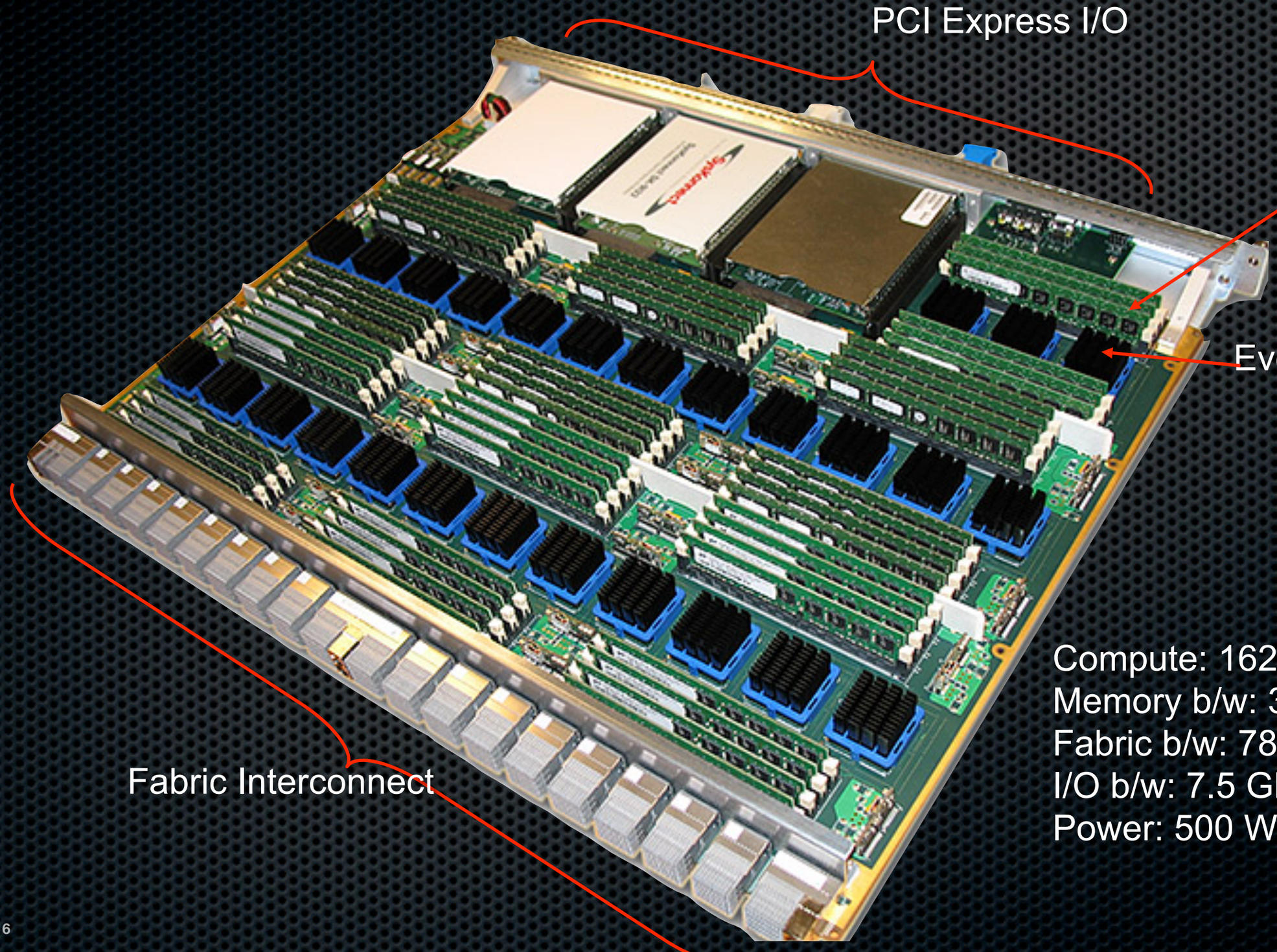
The SiCortex Node Chip



Six way Linux SMP with 2 DDR ports, PCI Express, Message controller, and fabric switch

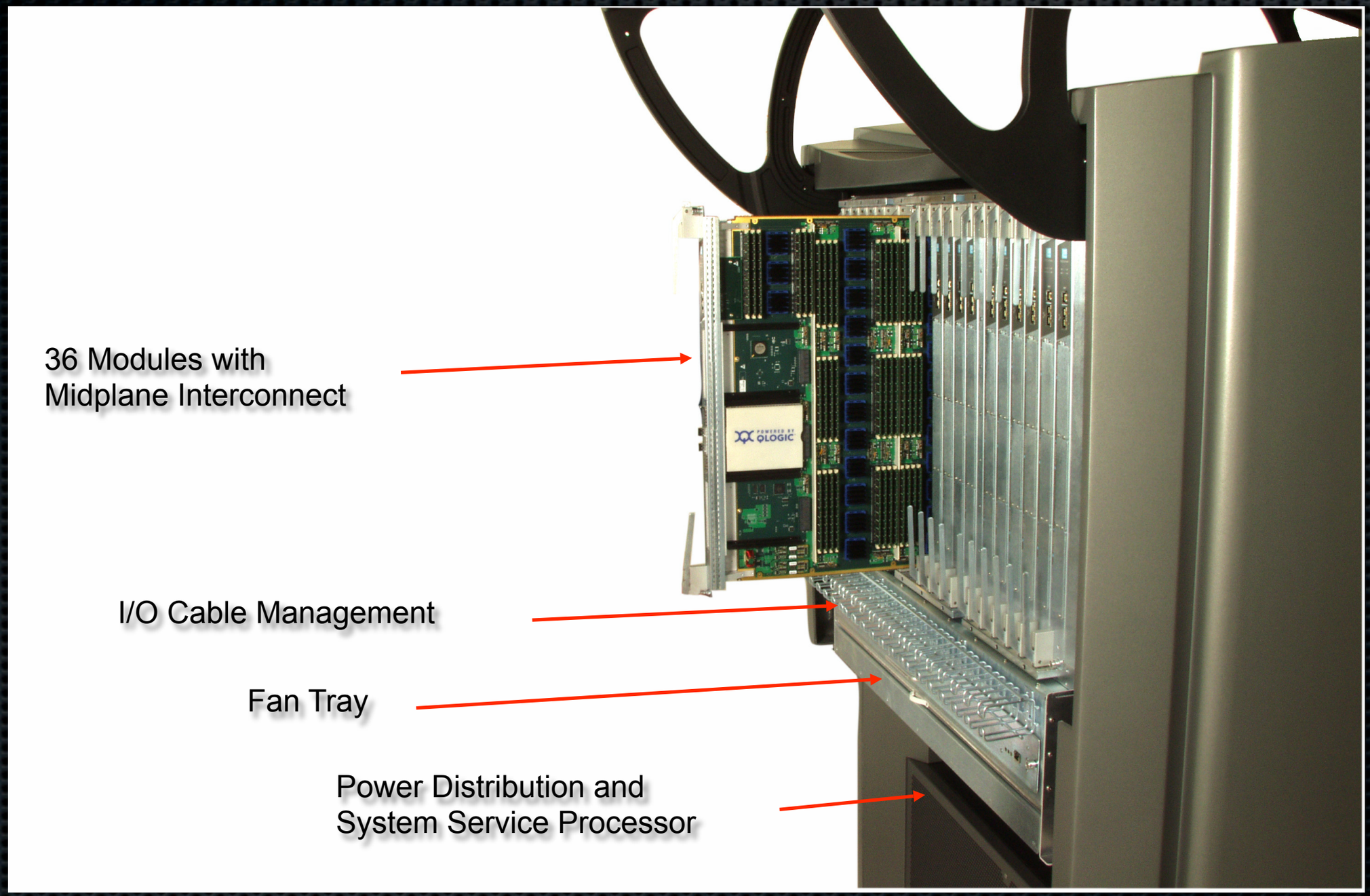
SiCortex

The SiCortex Module



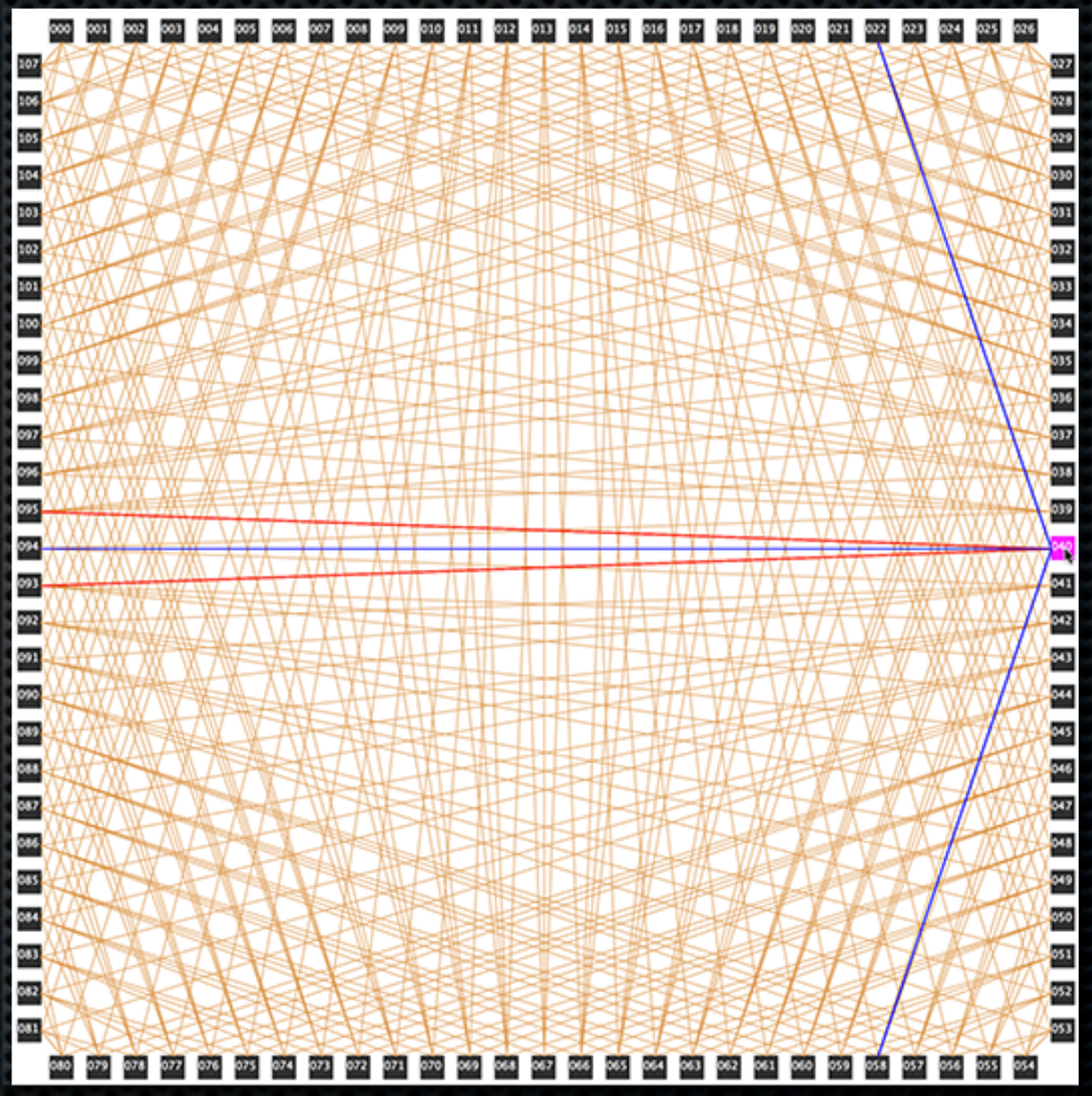
Compute: 162 GF/sec
Memory b/w: 345 GB/sec
Fabric b/w: 78 GB/sec
I/O b/w: 7.5 GB/sec
Power: 500 Watts

The SiCortex System



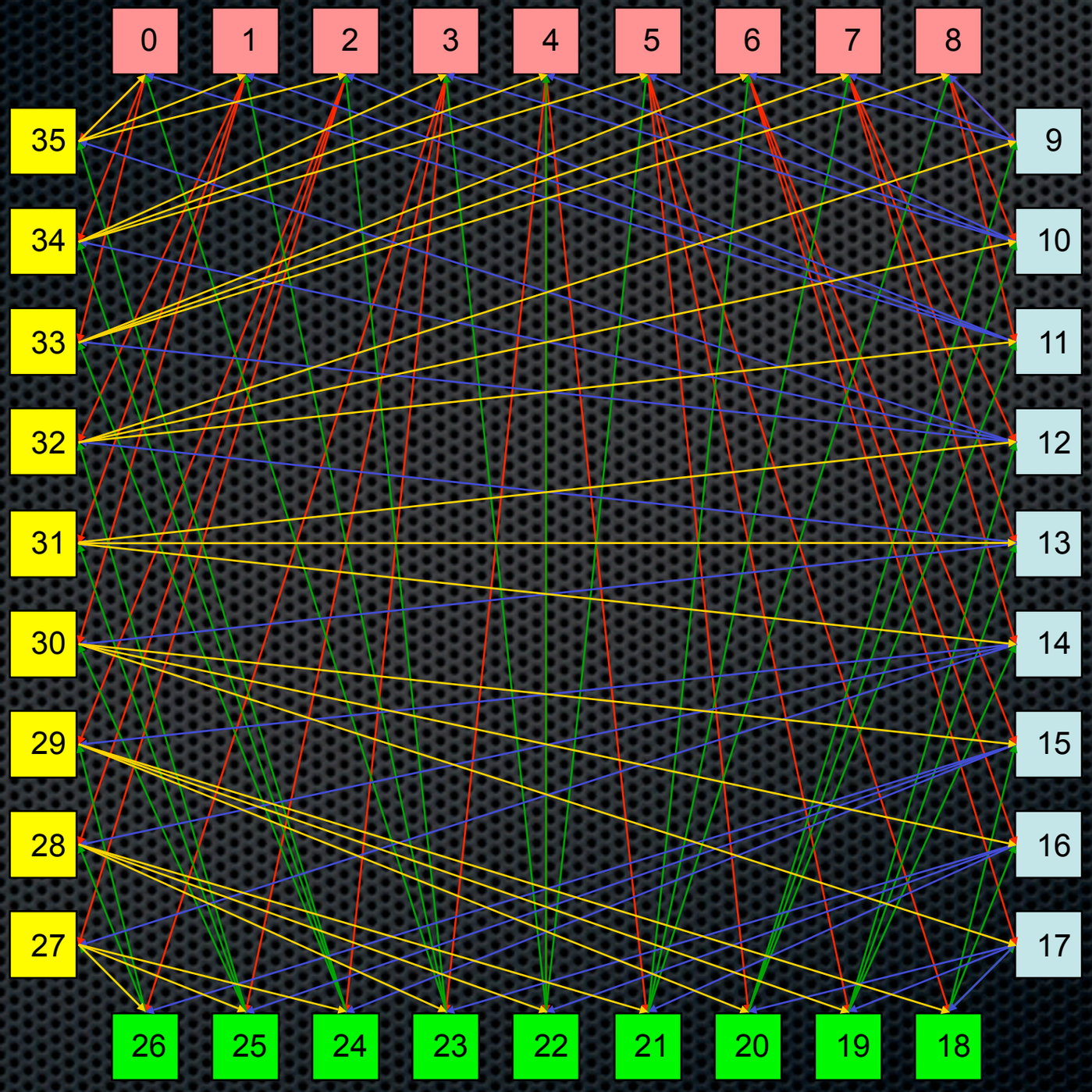
The Kautz Graph

- Logarithmic diameter
- Reconfigure around failures
- Low contention
- Very fast collectives



SiCortex

Thirty-six node Kautz graph

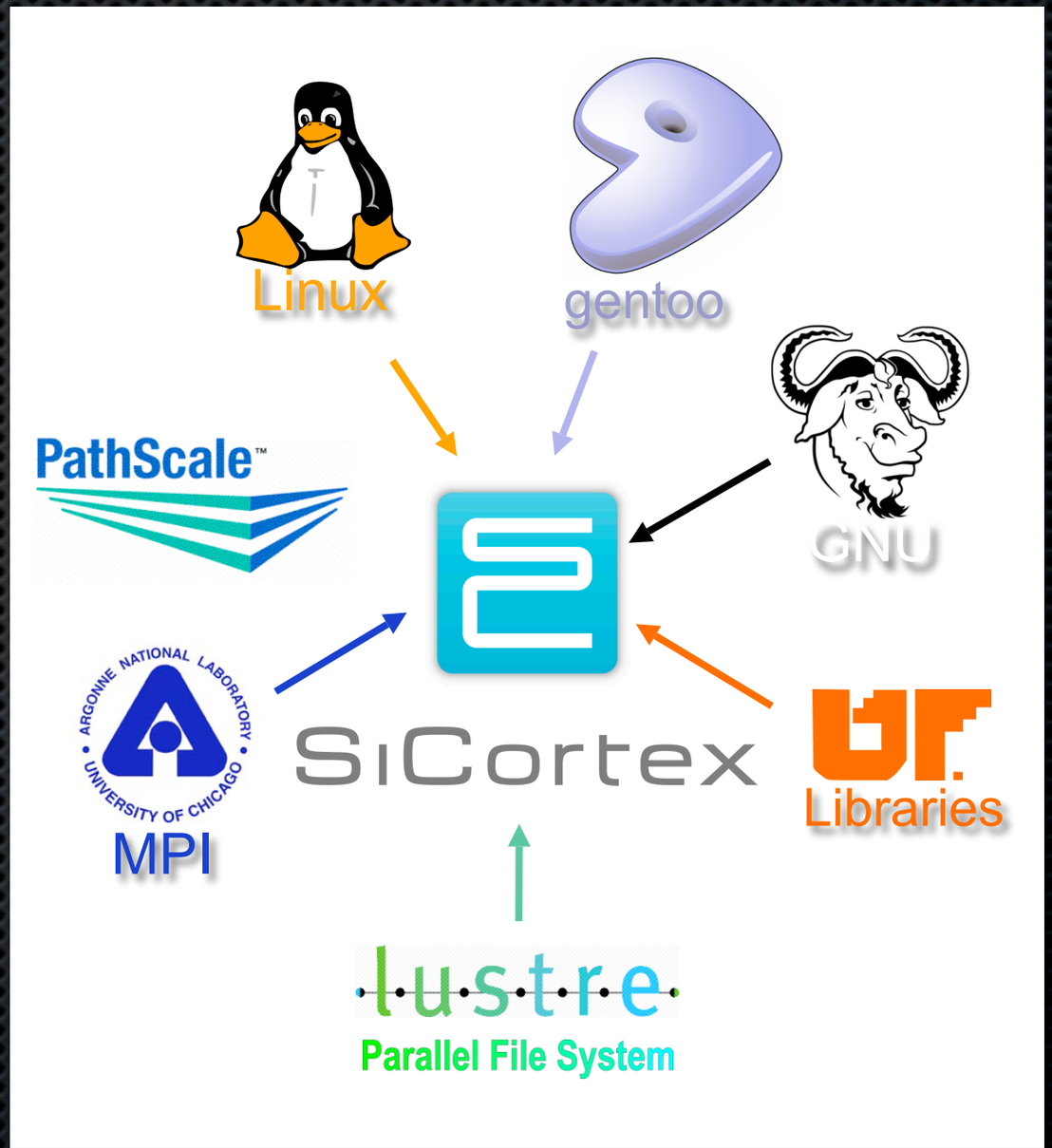


A pattern is developing

SiCortex

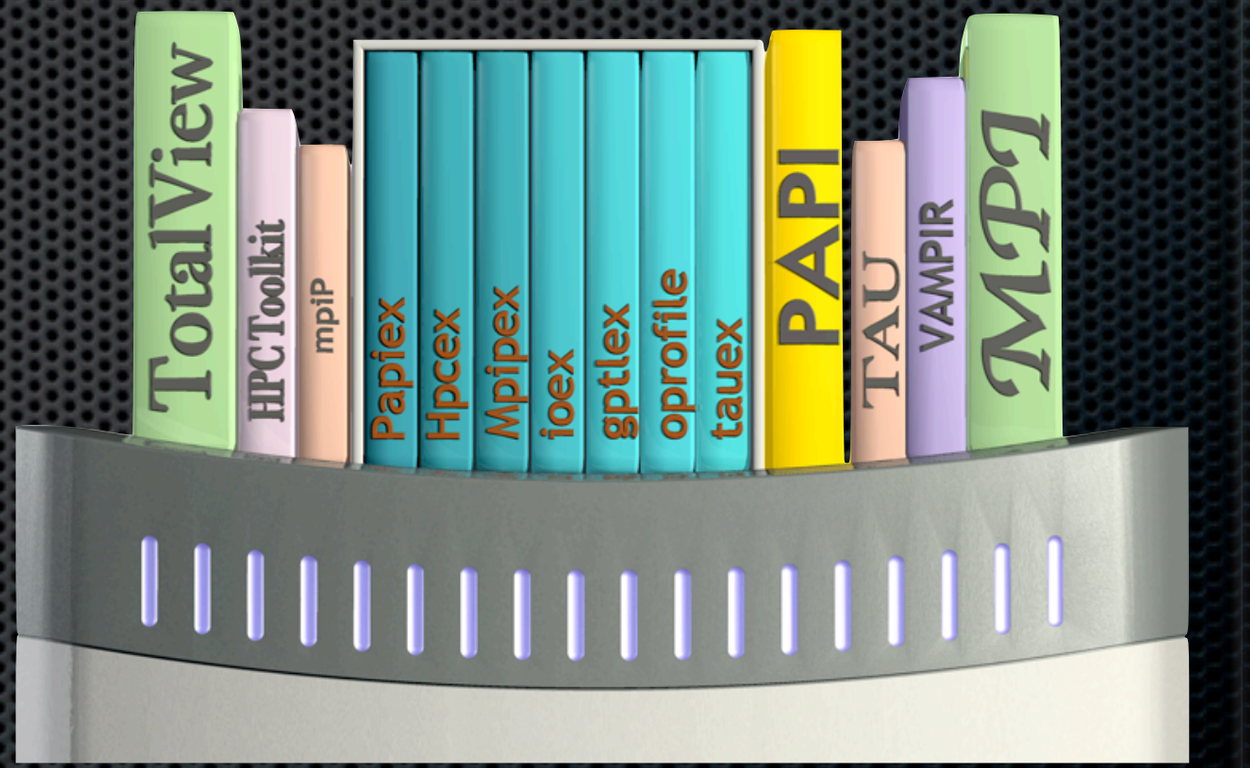
Integrated HPC Linux Environment

- Operating System
  - Linux kernel and utilities (2.6.18+)
  - Cluster file system (Lustre)
- Development Environment
  - GNU C, C++
  - Pathscale C, C++, Fortran
  - Math libraries
  - Performance tools
  - Debugger (TotalView)
  - MPI libraries (MPICH2)
- System Management
  - Scheduler (SLURM)
  - Partitioning
  - Monitoring
  - Console, boot, diagnostics
- Maintenance and Support
  - Factory-installed software
  - Regular updates
  - Open source build environment



Tuning Tools

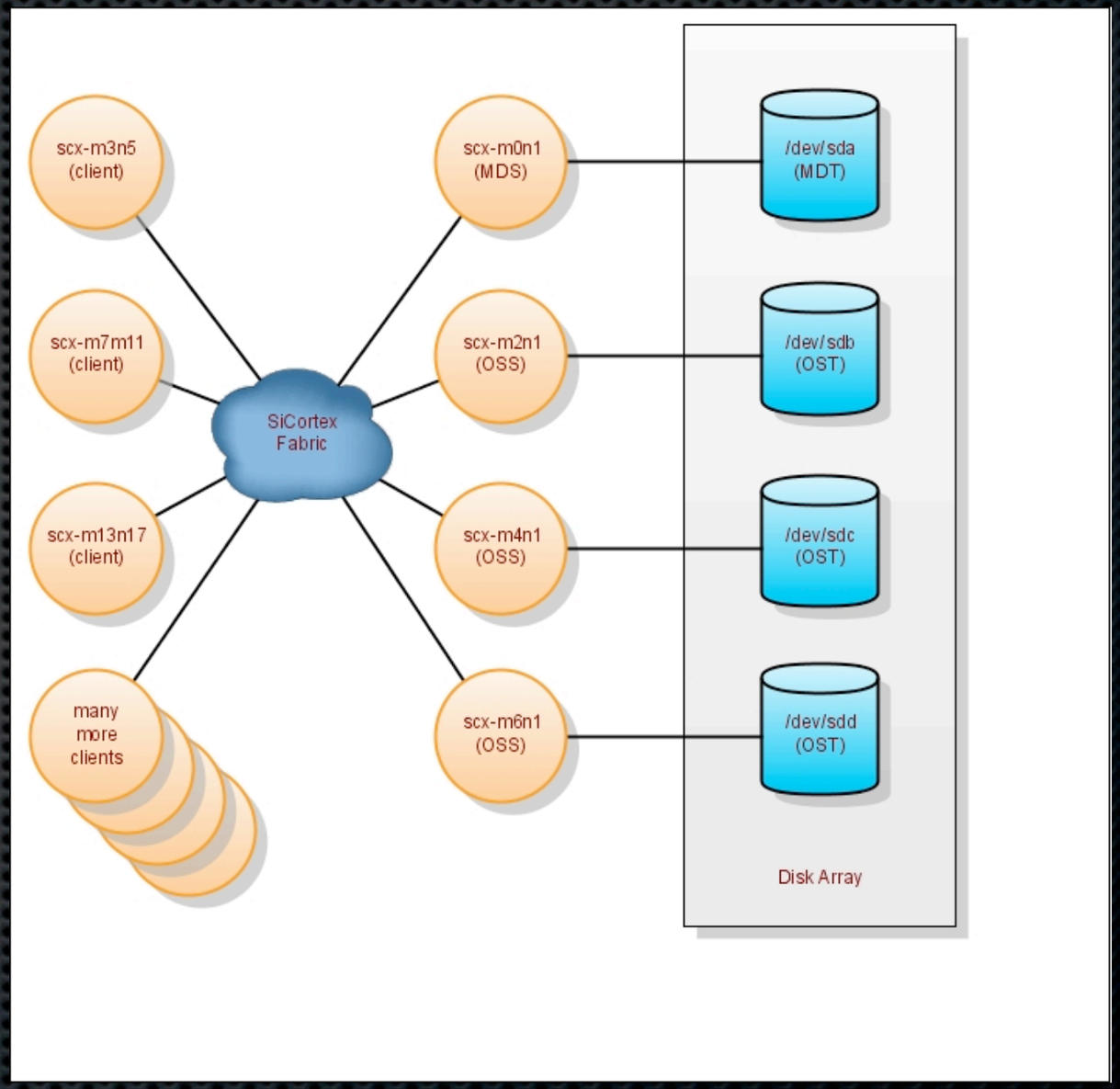
- Serial code (hpcex)
- Comm (mpiex)
- IO (ioex)
- System (oprofile)
- Hardware (papiex)
- Visualization (tau, vampir)



SiCortex

Parallel File System

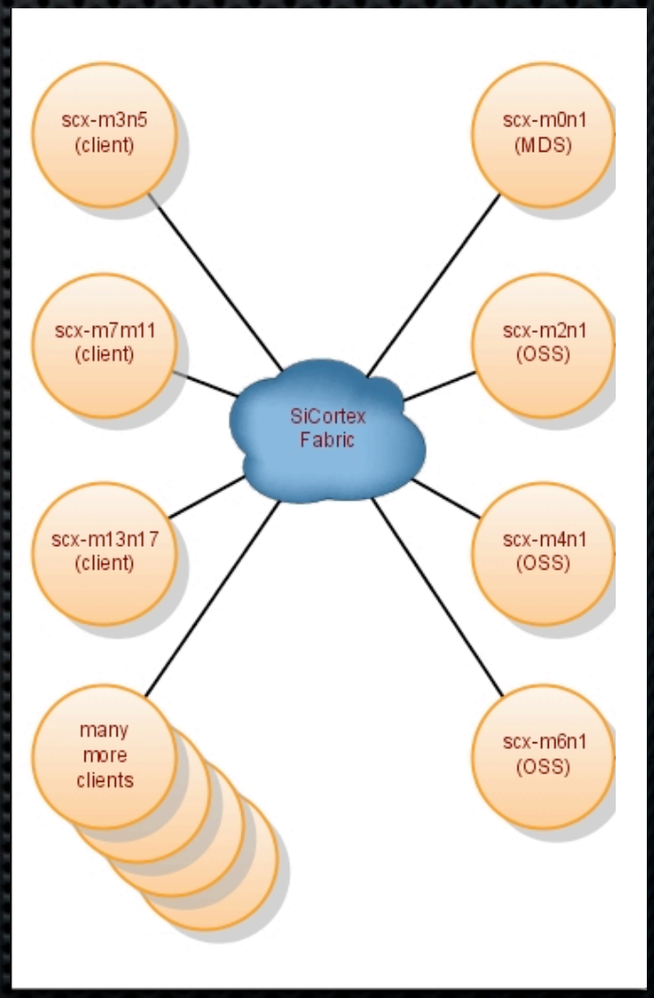
- Lustre Parallel File System
  - Open Source
  - Posix Compliant
- Native Implementation
  - Uses DMA Engine Primitives
- Scalable
  - Up to hundreds of I/O nodes



SiCortex

FabriCache

- RAM-backed file system
  - Based on Lustre file system
  - Store all data in Object Storage Server RAM
  - Present data as a file system
  - Scalable to 972 OSS nodes
- Similar to an SSD, but...
  - Higher bandwidth / lower latency
  - No external hardware required
  - Creating/removing volumes is easier
- Useful for...
  - Intermediate results
  - Shared pools of data
  - Staging data to/from Disk



SiCortex

MicroBenchmarks and Kernels

MPI Latency - 1.4 µsec

MPI BW - 1.5 GB/s

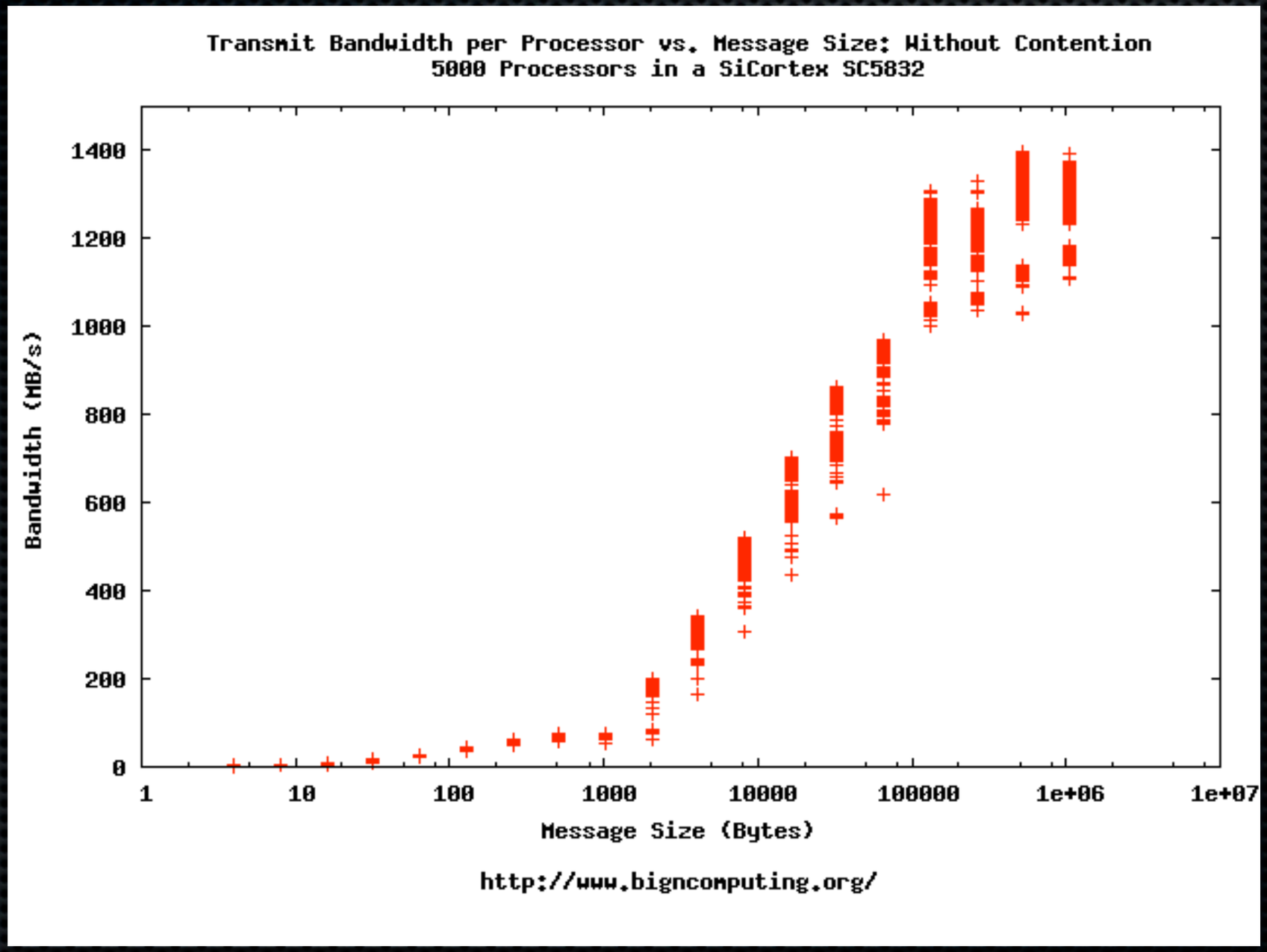
HPC Challenge work underway

SC5832, on 5772 cpus:

- DGEMM 72%
- HPL 3.6 TF (83% of DGEMM)
- PTRANS 210 GB/s
- STREAM 345 MB/s (1.9 TB/s aggregate)
- FFT 174 GF
- RandomRing 4 usec, 50 MB/s
- RandomAccess 0.74 GUPS (5.5 optimized)

SiCortex

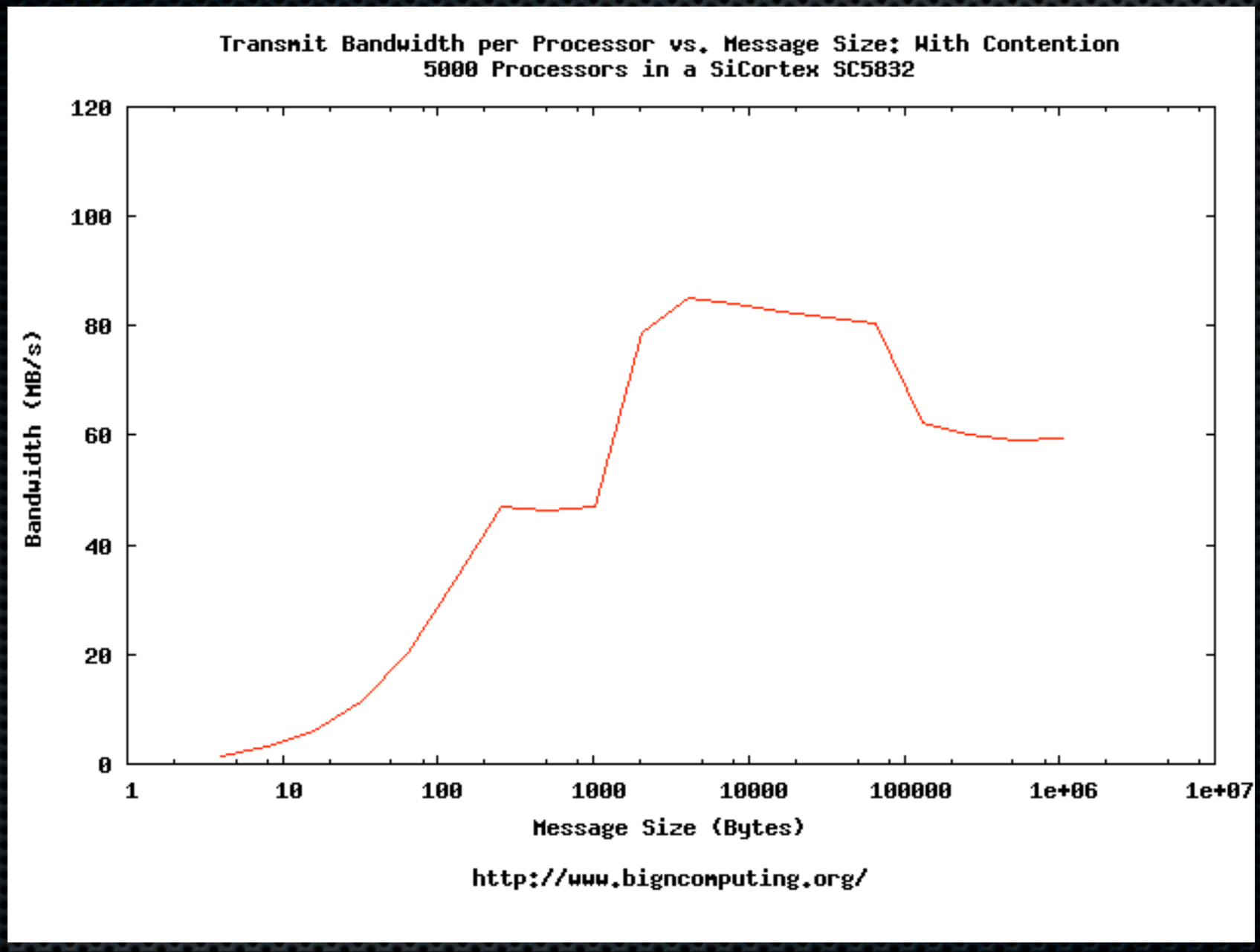
Zero contention message bandwidth?



Interesting relationship between message size and bandwidth

SiCortex

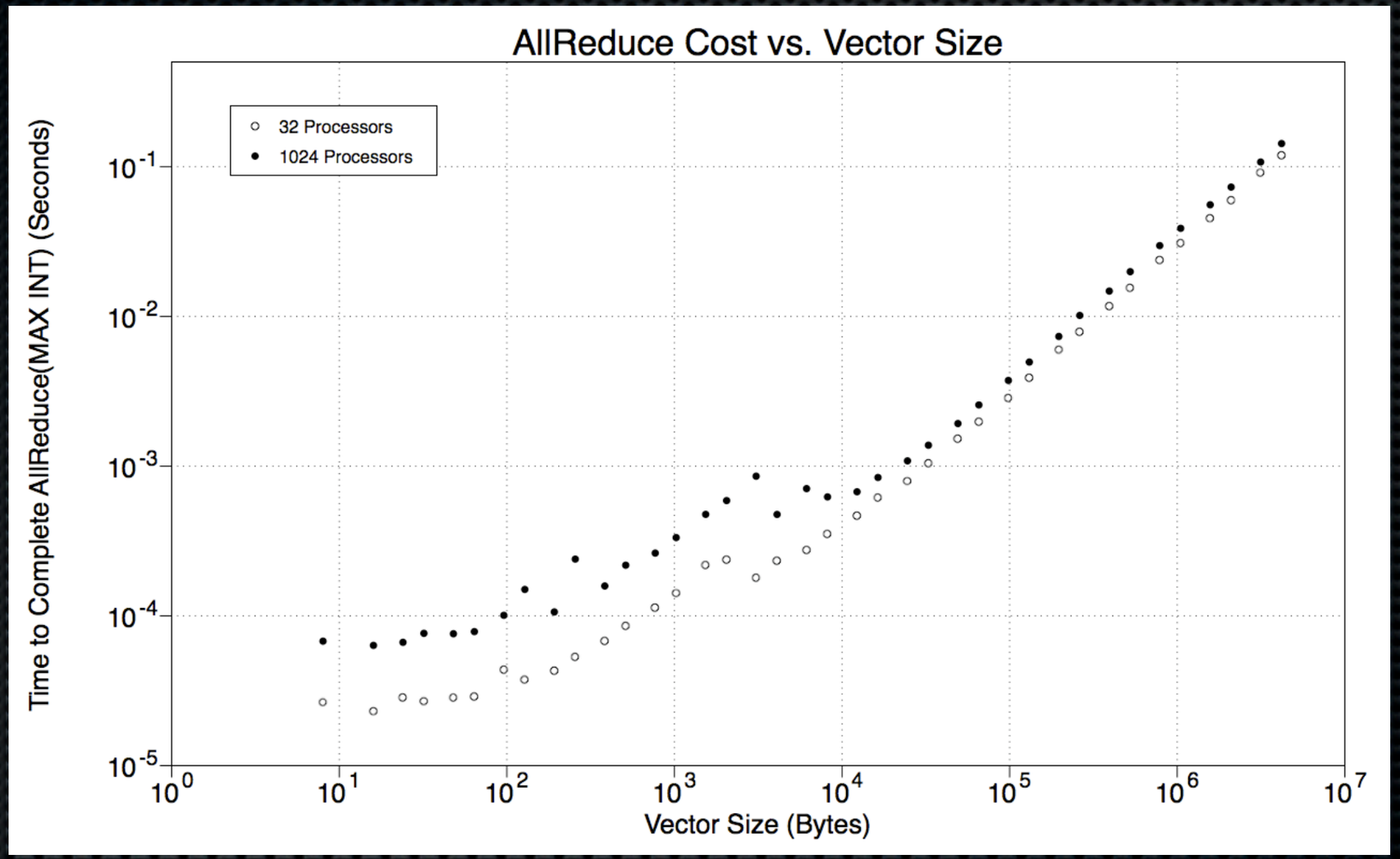
Communication in “real world” conditions



Contention matters. (For more, see Abhinav Bhatele’s work at http://charm.cs.uiuc.edu/ .)

SiCortex

What about Collectives?



Dependence on vector size is predictable.

SiCortex

What can it do?

- The machine shines on problems that require lots of communication between processes.

- TeraByte Sort
- Three-Dimensional FFT
- Huge systems of equations

SiCortex

TeraByte Sort

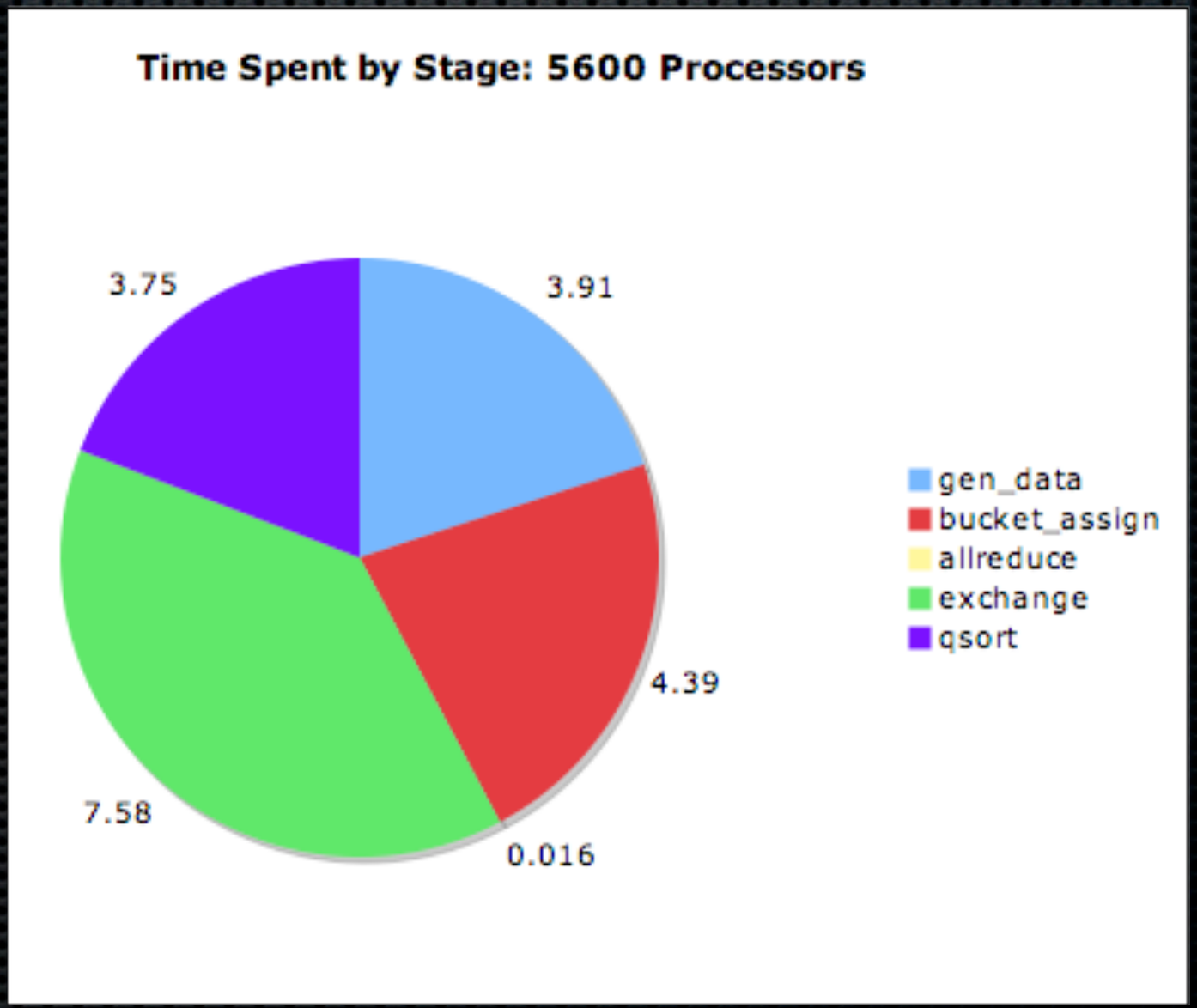
- Sort 10 billion 100 byte records (10 byte key).
- Leave out the IO (this isn't quite the Indy TeraSort benchmark)
- Use 5600 processors

- Key T_{comm} attributes:
  - Time to exchange all 1TB is about 4 sec +/-
  - Time to copy each processor's sublist is about 1 sec +/-
  - Global AllReduce for a 256KB vector is O(10mS)

SiCortex

Tuning....

- Improved QSort to the model target
- Bucket assignment is still very slow
- Exchange is still a little slow
- We can do better...



SiCortex

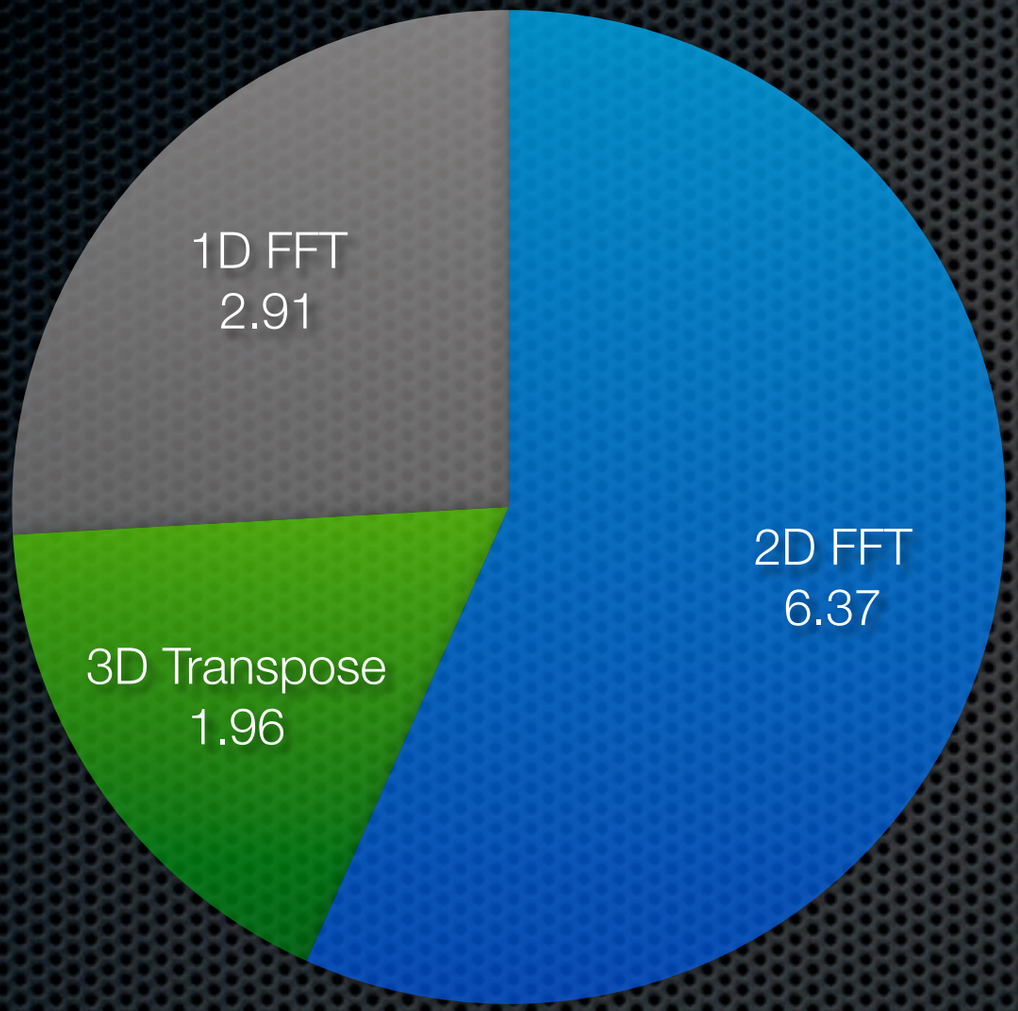
Three-Dimensional FFT

- 3D FFT of 1 billion point volume
- Use PFAFFT (prime factor analysis)
  - complex-complex single precision
    - 1040 x 1040 x 1040
- Two target platforms:
  - SC072 -- 72 processors
  - SC1458 -- 1458 processors

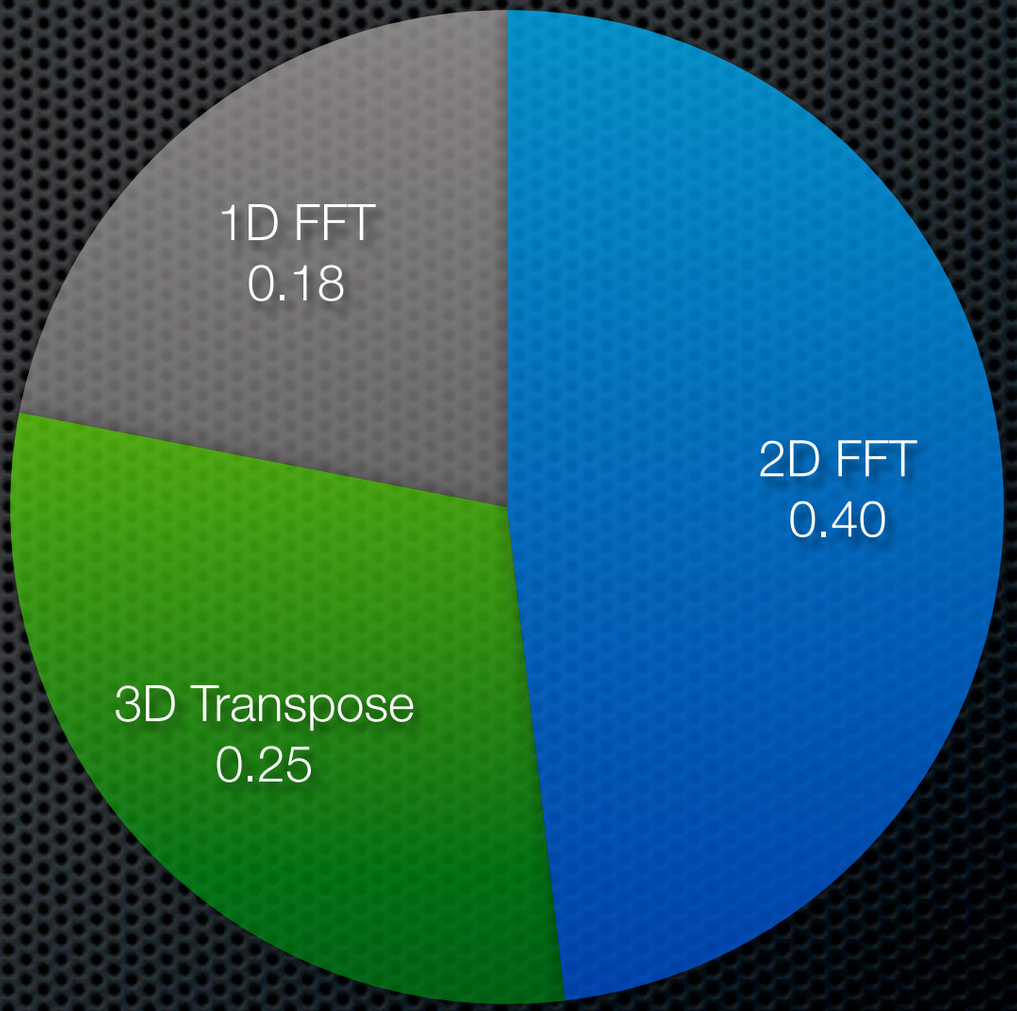
SiCortex

Results!

65 Processor 3D FFT



1040 Processor 3D FFT



FFTW3 is now producing comparable results.

SiCortex

Product Directions

- Revere the model: $T_{sol} = T_{arith}/N + T_{mem}/N + T_{IO} + f(N)T_{comm}$
- First generation emphasized T_{comm} and T_{IO}
- Second generation aimed at T_{mem} and T_{arith}
  - while taking advantage of technology improvements for T_{comm} and T_{IO}
- More performance per watt/cubic-foot/dollar
- Richer IO infrastructure
- “Special purpose” configurations

SiCortex

Take-away

- SiCortex builds Linux clusters
- With purposed built components
- Optimized for high-communication applications

High Processor Count Computing

SiCortex