COLUMBIA RBRC UKQCD QCDOC

QCDOC Project
Supercomputing 2004

Peter Boyle
for the QCDOC collaboration

COLUMBIA RBRC UKQCD QCDOC

QCDOC Collaboration

Columbia:	N. Christ, S. Cohen, C. Cristian, C. Kim, L. Levkova, X. Liao, G. Liu, R. Mawhinney, A. Yamaguchi
UKQCD:	P. Boyle, B. Joo
Riken-Brookhaven:	S. Ohta (KEK), T. Wettig (Regensburg)
SciDac:	C. Jung, K. Petrov
IBM Research:	A. Gara, D. Chen

Introduction

Quantum Chromodynamics On a Chip

Scientific instrument designed to simulate the strong nuclear force that binds quarks in protons, QCD.

- Quarks and gluons analogous to electrons and photons
- Non-linear - must be solved numerically
- Perform Feynman path integral for QCD in $(2fm)^4$ 4-torus importance sampling, Markov, Metropolis etc...
- 32^4 box $\Rightarrow 10^8$ dimensional integral
- Each sample costs around $\simeq 10^{16} - 10^{20}$ flops.
- Hundreds of Teraflop-years (or more)
- Dirac equation must be solved millions of times along Markov chain

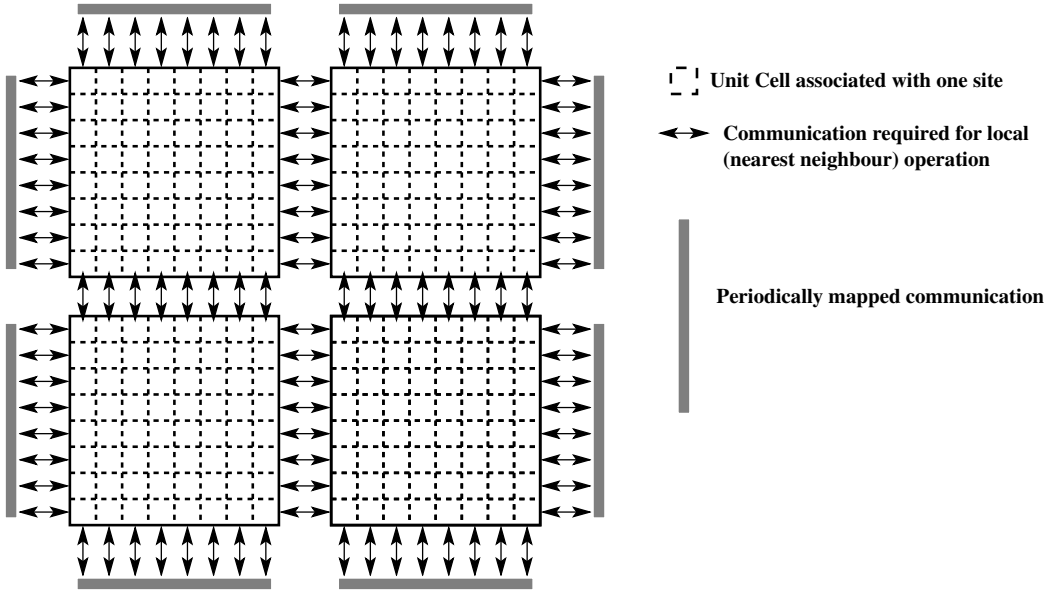
COLUMBIA RBRC UKQCD QCDOC

Dirac Equation

$$\sum_{\mu=x,y,z,t} \gamma^{\mu}(\partial_{\mu} + igA_{\mu}(x))\psi(x) = m\psi(x)$$

The QCD Gluon field A_μ is a 3×3 complex analog of the electromagnetic vector potential. $\Rightarrow$ 2 GB/s memory BW per Gflop/s

ψ is a 3 (color) $\times$ 4 (spin) complex field



COLUMBIA RBRC UKQCD QCDOC

Problem characteristics

Krylov solvers dominate: Dirac matrix multiply + global summation

- Simple nearest neighbour stencil.
- Communications are nearest neighbour on Toriodal Network
- Very large machines ⇒ small local sub-lattice
- Challenge is to provide the data fast enough
- Interconnect as agressive as a local DRAM system
- Communications repetitive
- Communicate in multiple directions simultaneously
- Predictable memory access patterns ⇒ prefetch

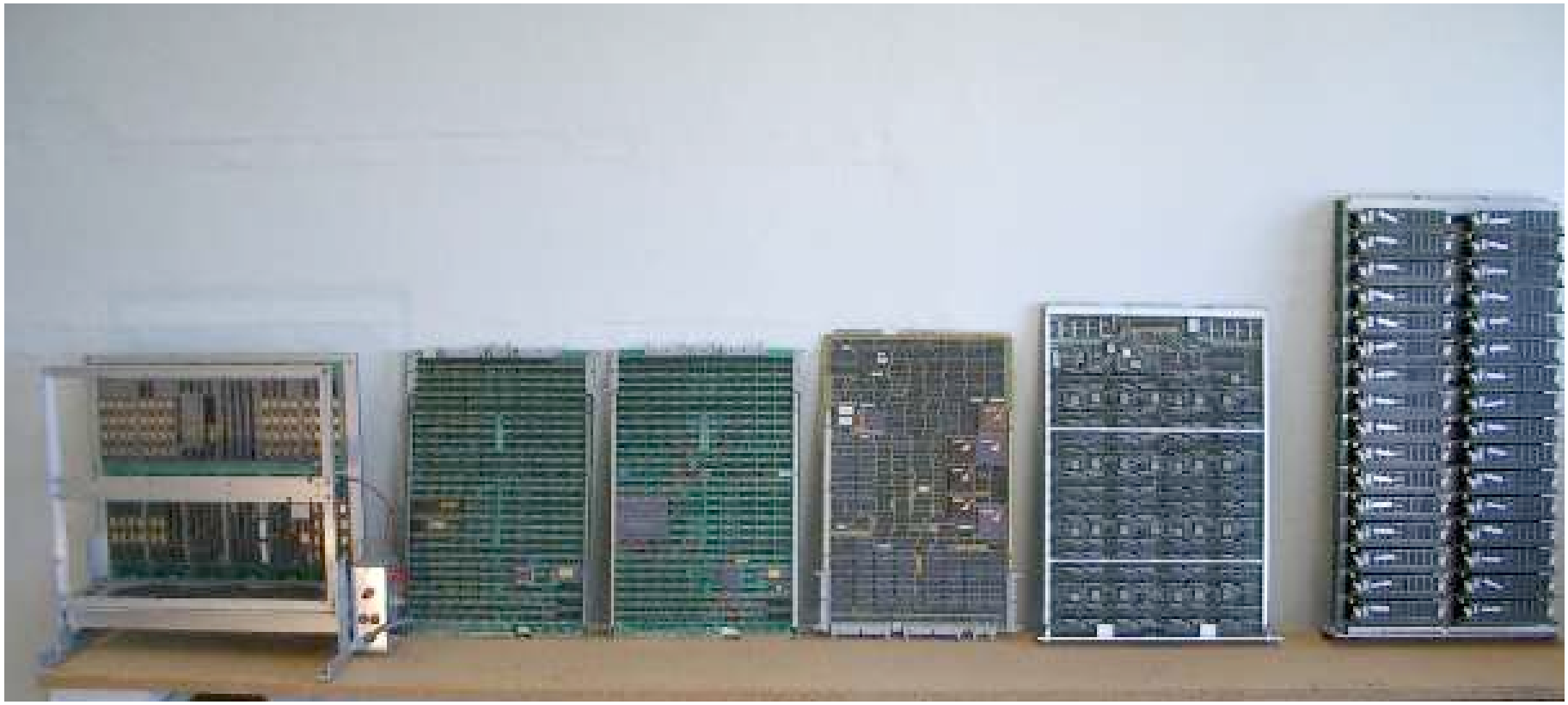
Exploit simplifications: special purpose machines increase scalability

COLUMBIA RBRC UKQCD QCDOC

Machine	Date	Processor (FPU precision)	Nodes	Speed (Gflops)	Memory (GBytes)
Previous:					
3x3 multiplier	1983	CU (16) + PDP11	1	0.001	192 Bytes
16-node	1985	286/TRW (22)	16	0.25	0.016
64-node	1987	286/Weitek (32)	64	1.0	0.128
256-node	1989	286/Weitek (64)	256	16.0	0.5
CU QCDSP	1998	TI DSP (32)	8,192	400	16
RBRC QCDSP	1998	TI DSP (32)	12,288	600	24
Current:					
RBRC QCDOC	2004	440 PPC (64)	12,288	10,000	1,570
UKQCD QCDOC	2004	440 PPC (64)	12,288	10,000	1,570
US LGT QCDOC	2005	440 PPC (64)	12,288	10,000	1,570

Also: http://www.netlib.org/utk/lsi/pcwLSI/text/node38.html

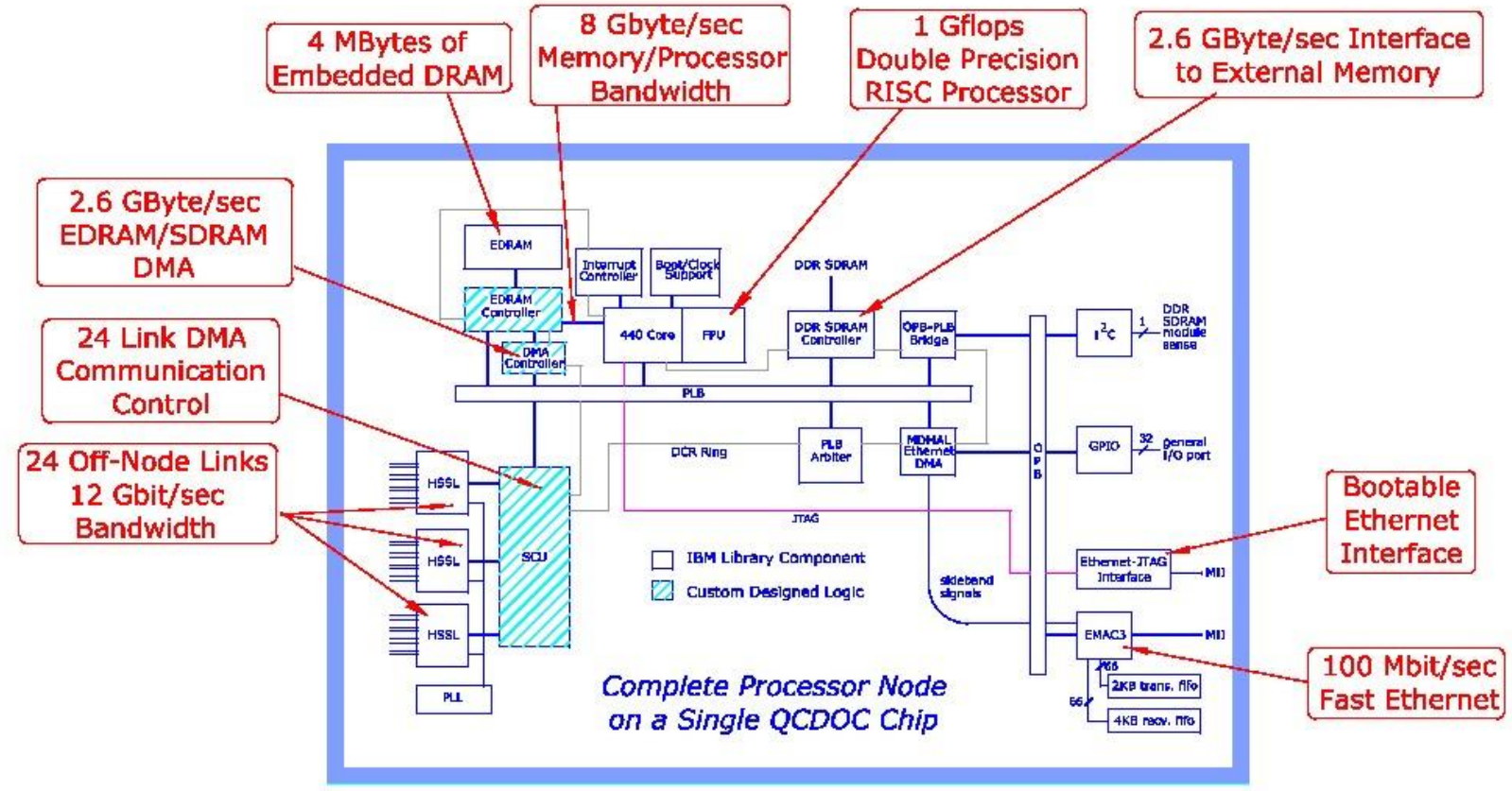
Caltech, APE, Fermilab, GF11, QCDPAX, CP-PACs, GigE mesh.



COLUMBIA RBRC UKQCD QCDOC

QCDOC Architecture

- 6 dimensional torus
- QCD can be maximally spread out in four/five dimensions
- Extra dimensions allow partitioning
- Over 10k nodes $\Rightarrow$ small subvolume per node
- fast on-chip memory
- high performance nearest neighbour communication
- 24 DMA FIFO's to nearest neighours.
- Frequent global summation : Hardware assist
- Communication performance $\simeq$ memory performance


4 MBytes of Embedded DRAM
8 Gbyte/sec Memory/Processor Bandwidth
1 Gflops Double Precision RISC Processor
2.6 GByte/sec Interface to External Memory
2.6 GByte/sec EDRAM/SDRAM DMA
24 Link DMA Communication Control
24 Off-Node Links 12 Gbit/sec Bandwidth
Bootable Ethernet Interface
100 Mbit/sec Fast Ethernet
EDRAM
EDRAM Controller
DMA Controller
Interrupt Controller
Boot/Clock Support
440 Core
FPU
DDR SDRAM
DDR SDRAM Controller
OPB-PLB Bridge
PLB
DCR Ring
PLB Arbiter
MIDHAL Ethernet DMA
OPB
I2C
DDR SDRAM module sense
GPIO
general I/O port
HSSL
HSSL
HSSL
SCU
PLL
JTAG
IBM Library Component
Custom Designed Logic
sideband signals
Ethernet-JTAG Interface
MII
EMAC3
2KB trans. fifo
4KB recv. fifo
Complete Processor Node on a Single QCDOC Chip

QCDOC Asic Technology

- IBM SA27E 0.18μ mixed logic + Embedded DRAM process
- 420 MHz IBM PowerPC SoC (440 CPU + FPU)
- 5 Watts allowing high density
- Custom 4MB on-chip embedded memory controller (IBM Research, Yorktown Heights)
- Custom communication network and DMA (Columbia)
- Custom Ethernet-JTAG boot/diagnostic protocol (IBM Research, Yorktown Heights)
- Hardware and Software *Co-Designed*

Serial Communications Unit

- 12 $\times$ 420Mbit/s LVDS serial links using IBM HSSL transceivers
- Single bit detect, ACK/NACK, hardware retransmit, checksums
- Block-strided descriptor based DMA
- Concurrently runs all links efficiently
- Hardware assist for global summation

Prefetching Edram Controller

- Wide 1152 bit backend to Embedded DRAM macros with ECC.
- Multiple ports: efficient concurrent access for Comms and Compute
- Each PEC port linearly prefetches two streams
- Double buffers gather writes

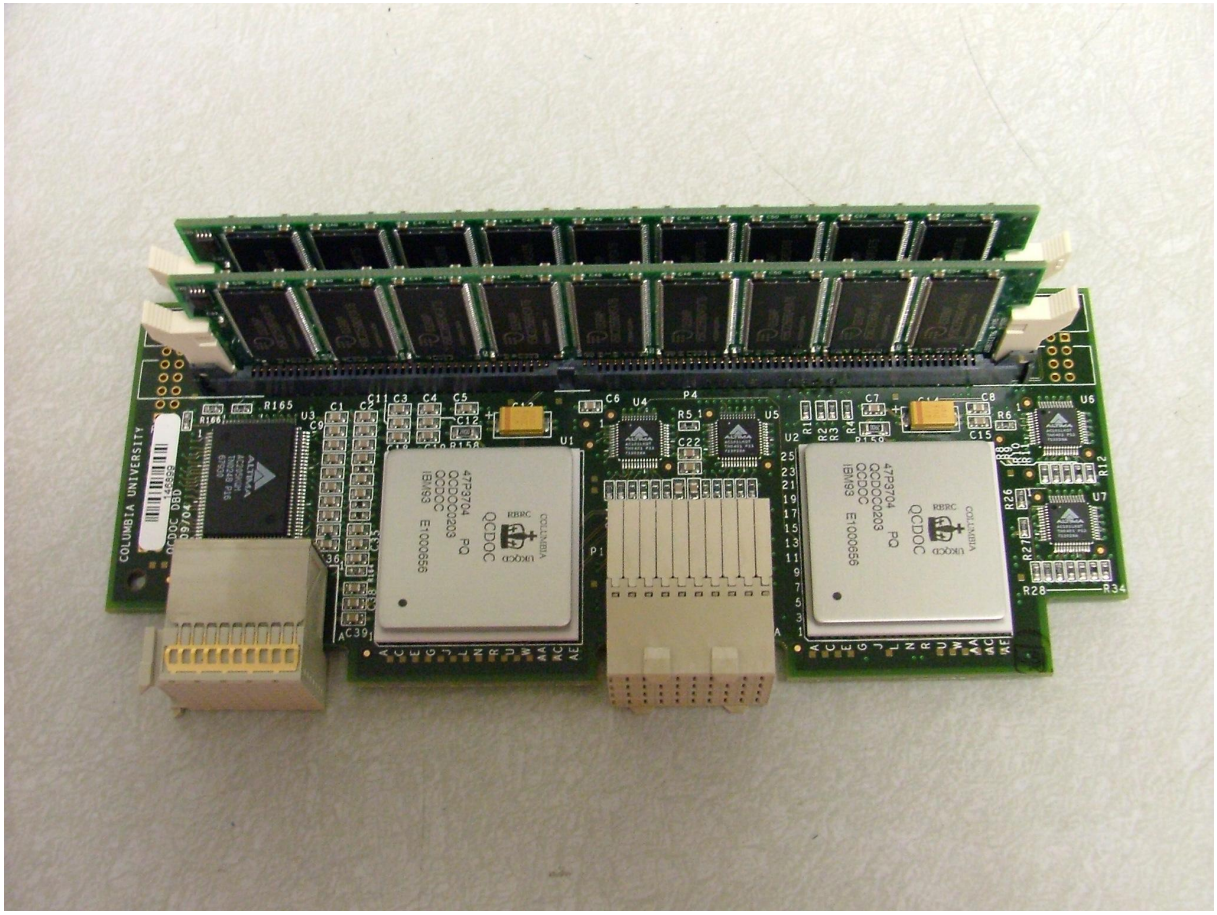
COLUMBIA RBRC UKQCD QCDOC

QCDOC Asic, 1 Gflop/s

COLUMBIA
RBRC UKQCD
QCDOC

QCDOC Daughterboard 2 Gflop/s, $450

Two independent compute nodes + Ethernet system

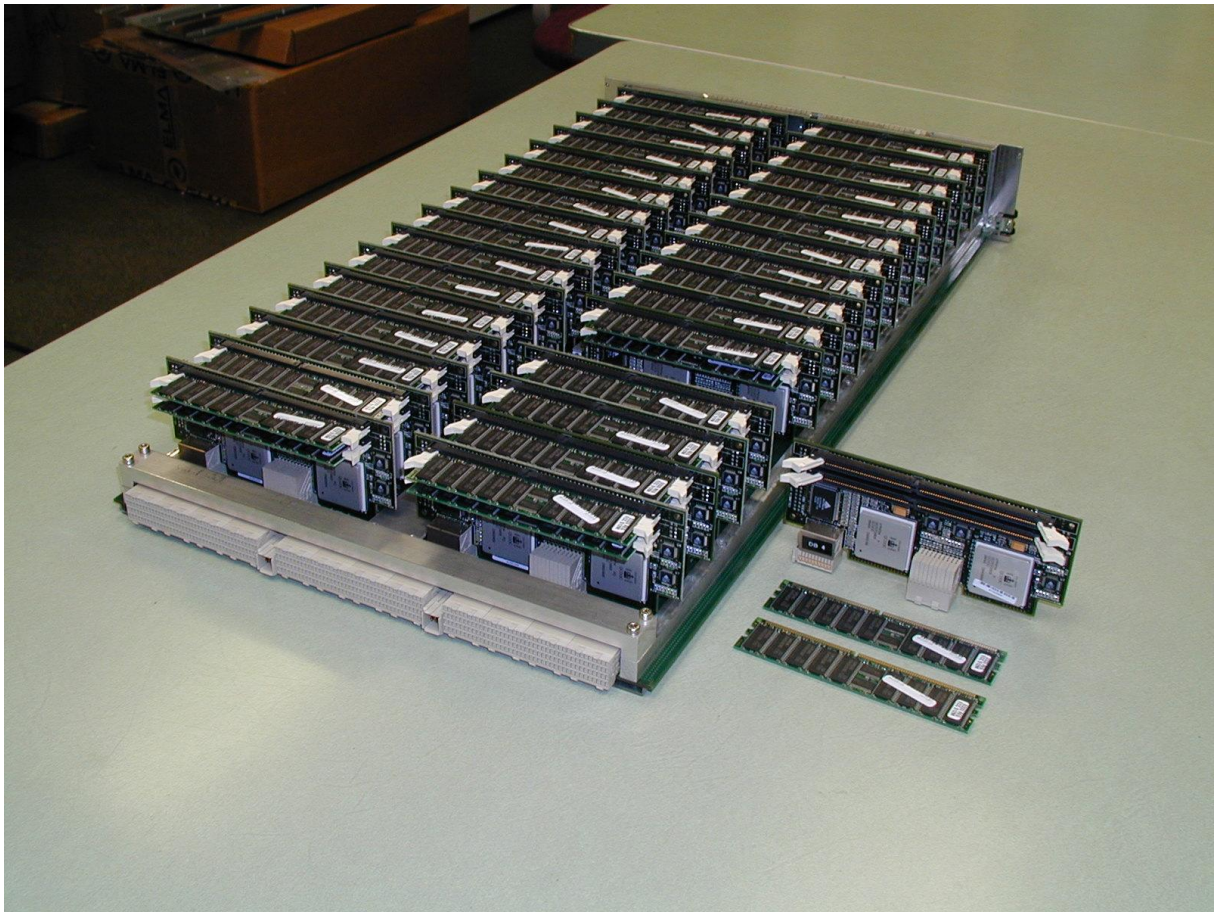


COLUMBIA RBRC UKQCD QCDOC

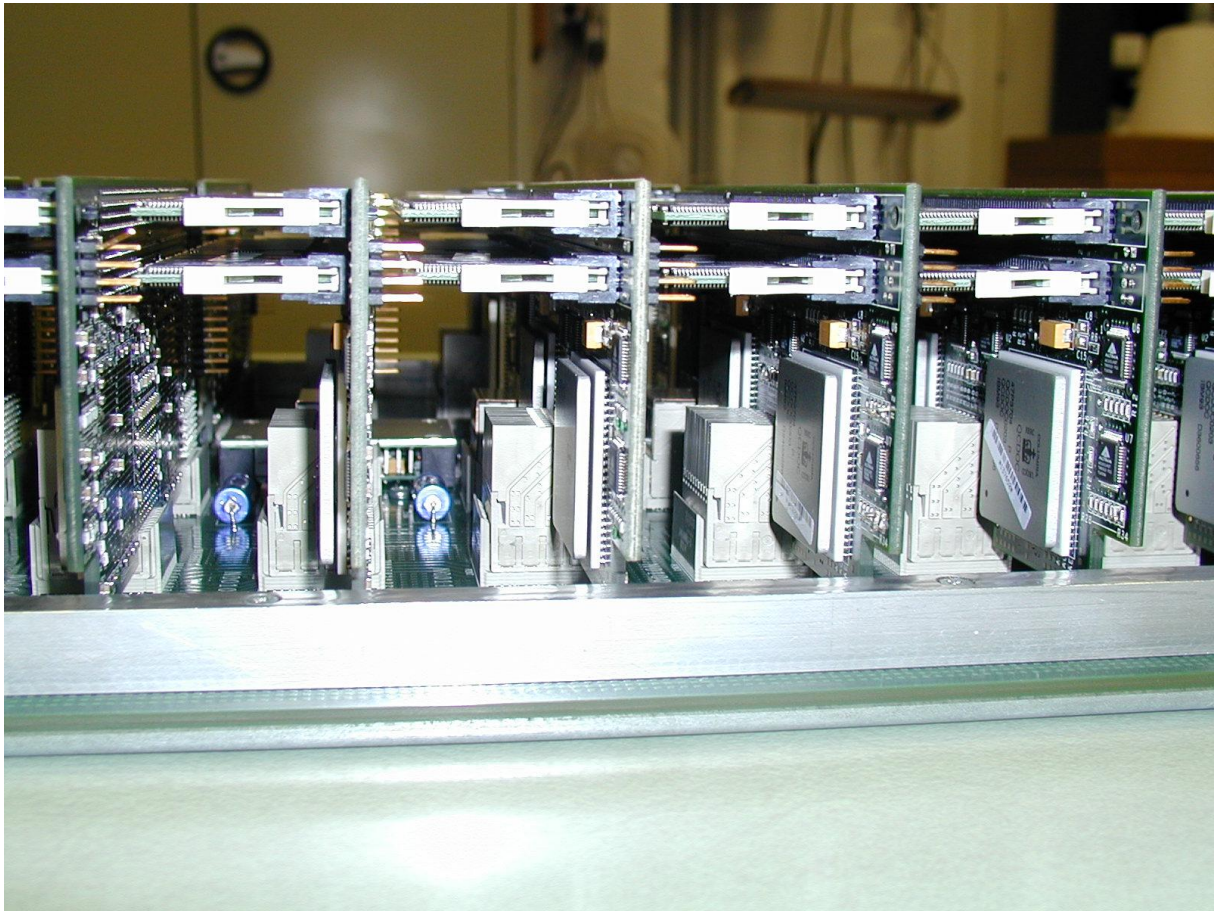
QCDOC Motherboard, $17k

64 Gflop/s, 32 Daughterboards, 64 nodes, very dense

768 Gigabit/s internode, 800 Mbit/s Ethernet.



QCDOC Motherboard



COLUMBIA RBRC UKQCD QCDOC

4096 node prototype, 4Tflop/s, $1.6 million

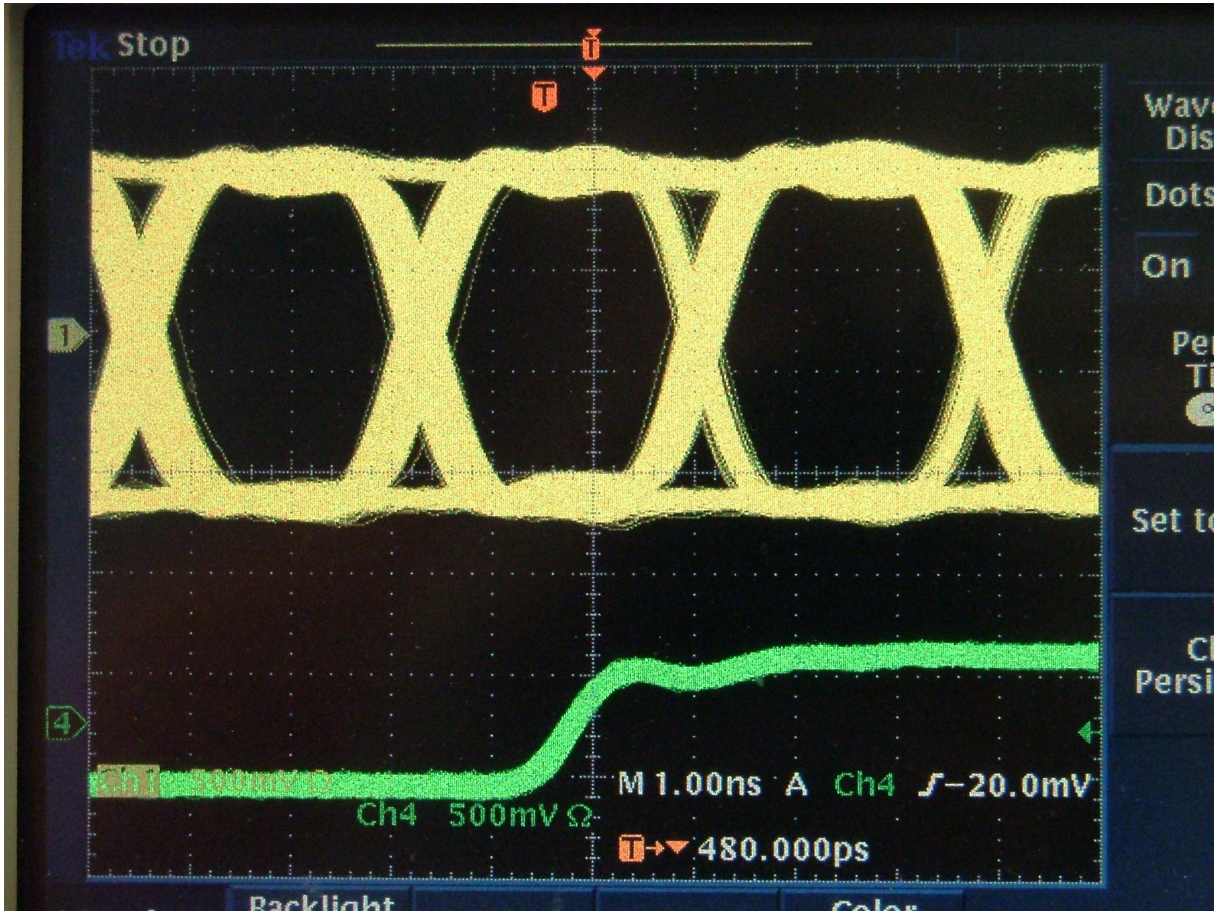
(Air cooled with chilled water heat exchanger)

COLUMBIA RBRC UKQCD QCDOC

12288 node UKQCD machine on the BNL machine floor.
10.3 Tflop/s peak (current clock speed), $5 Million
UKQCD machine disassembled and in shipping,
12288 node RBRC machine nearly complete,
12288 node DOE SciDAC machine by March

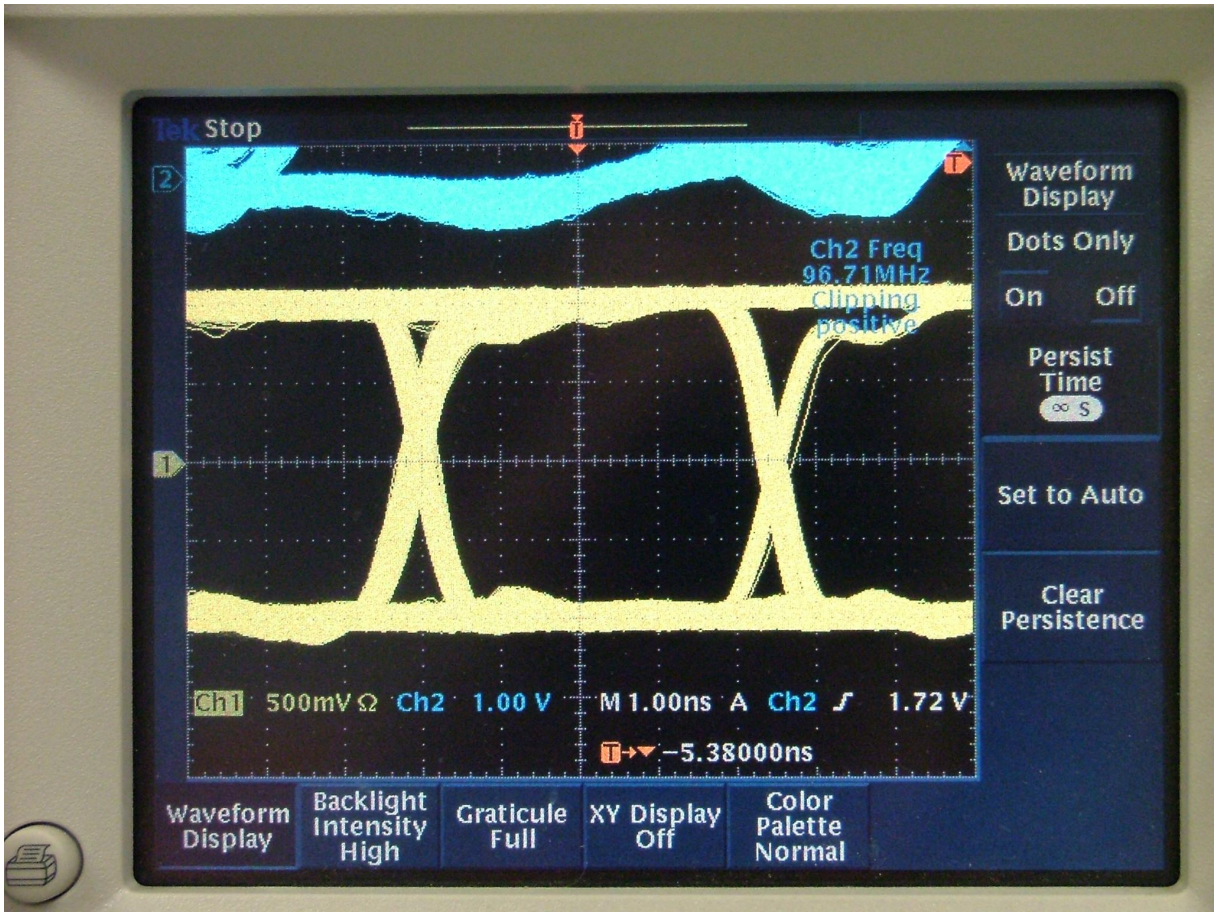
COLUMBIA RBRC UKQCD QCDOC

Serial Communications Data Eye



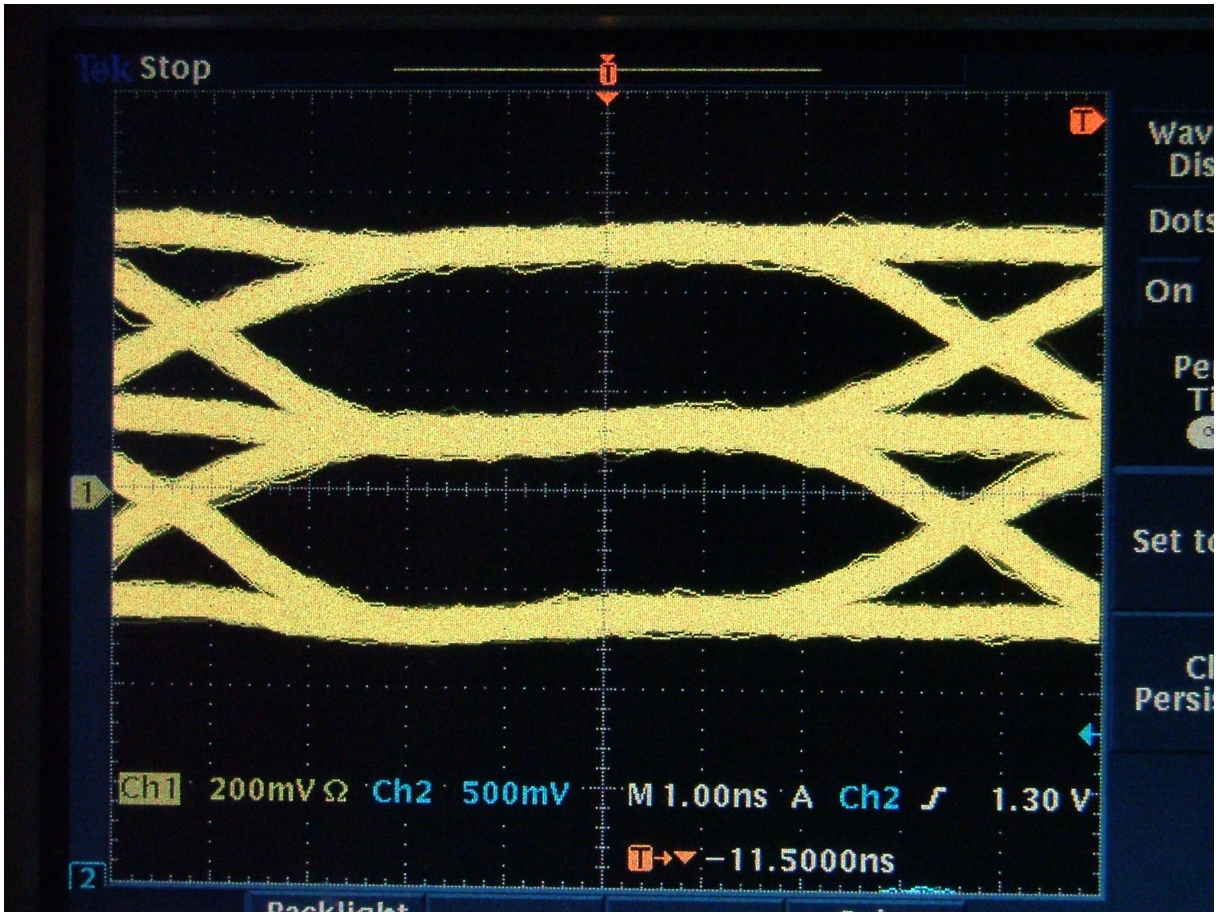
COLUMBIA RBRC UKQCD QCDOC

DDR memory data eye (write path)



COLUMBIA RBRC UKQCD QCDOC

Ethernet Data Eye (three level signal)



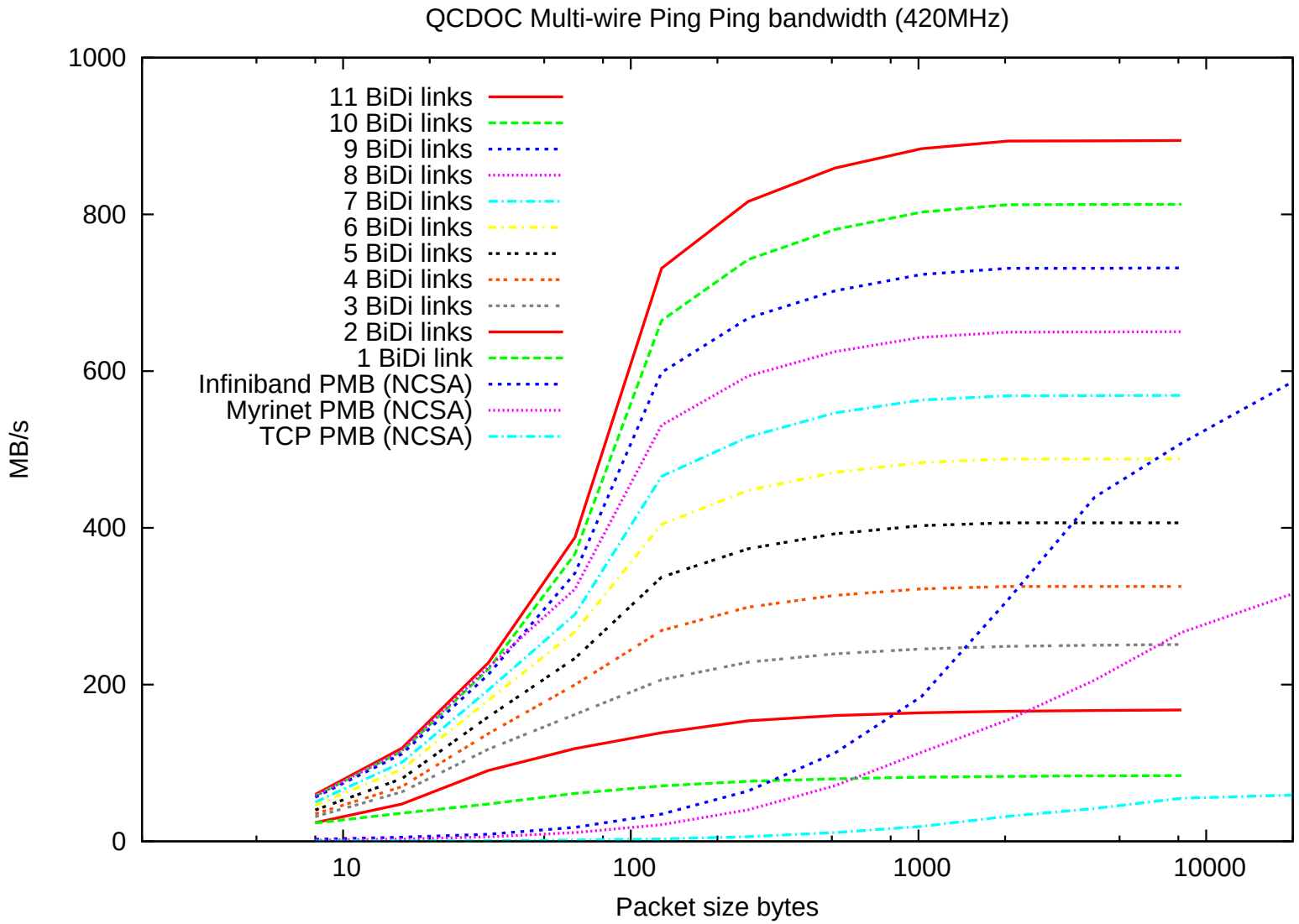
COLUMBIA RBRC UKQCD QCDOC

Performance

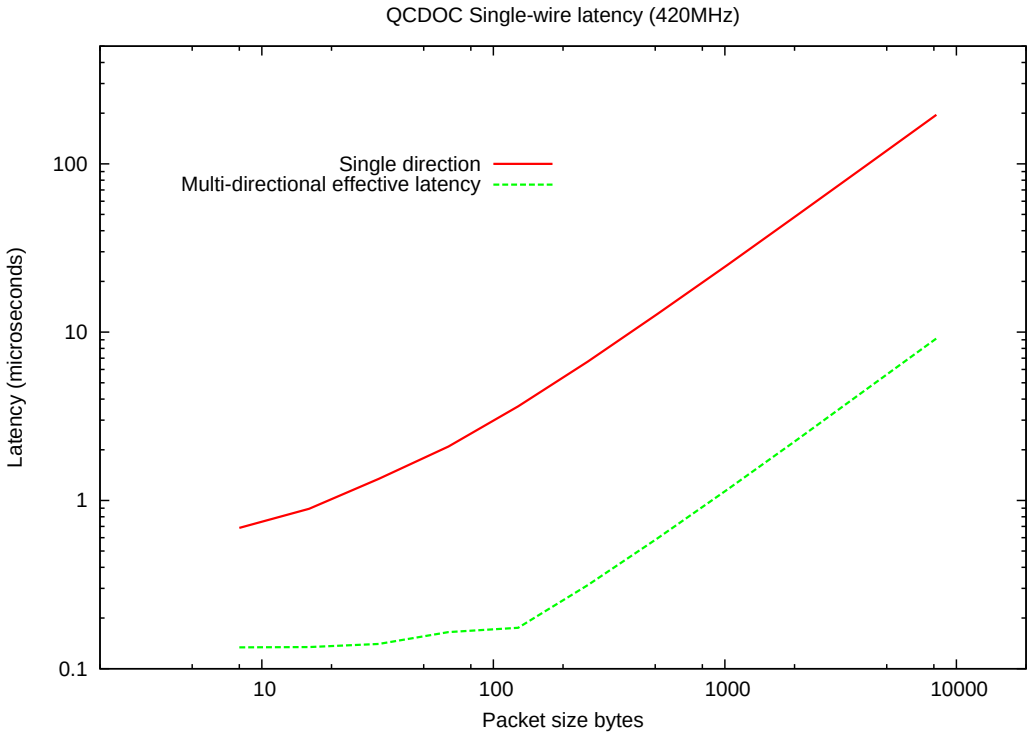
- Memory system performance
- Network performance
- Application performance

Streams performance

Compiler/Options/Code	Comment	Memory	MB/s
xlc -O5 -qarch=440	vanilla source	Edram	747
gcc-3.4.1 -funroll-all-loops -fprefetch-loop-arrays -O6	vanilla source	Edram	747
gcc-3.4.1 -funroll-all-loops -fprefetch-loop-arrays -O6	__builtin_prefetch	Edram	1024
Assembly	Auto-generated asm	Edram	1670
xlc -O5 -qarch=440	vanilla source	DDR	260
gcc-3.4.1 -funroll-all-loops -fprefetch-loop-arrays -O6	vanilla source	DDR	280
gcc-3.4.1 -funroll-all-loops -fprefetch-loop-arrays -O6	__builtin_prefetch	DDR	447
Assembly	Auto-generated asm	DDR	606


QCDOC Multi-wire Ping Ping bandwidth (420MHz)
MB/s
Packet size bytes
0
200
400
600
800
1000
10
100
1000
10000
11 BiDi links
10 BiDi links
9 BiDi links
8 BiDi links
7 BiDi links
6 BiDi links
5 BiDi links
4 BiDi links
3 BiDi links
2 BiDi links
1 BiDi link
Infiniband PMB (NCSA)
Myrinet PMB (NCSA)
TCP PMB (NCSA)

- Multi link bandwidth as good as CPU memory bandwidth!
- Up to 8,000,000 ping ping packets per second using all 12 links
- Single link half RTT Latency: 690 ns.
- Single link ping pong obtains 50% max bandwidth on 32 byte packets.

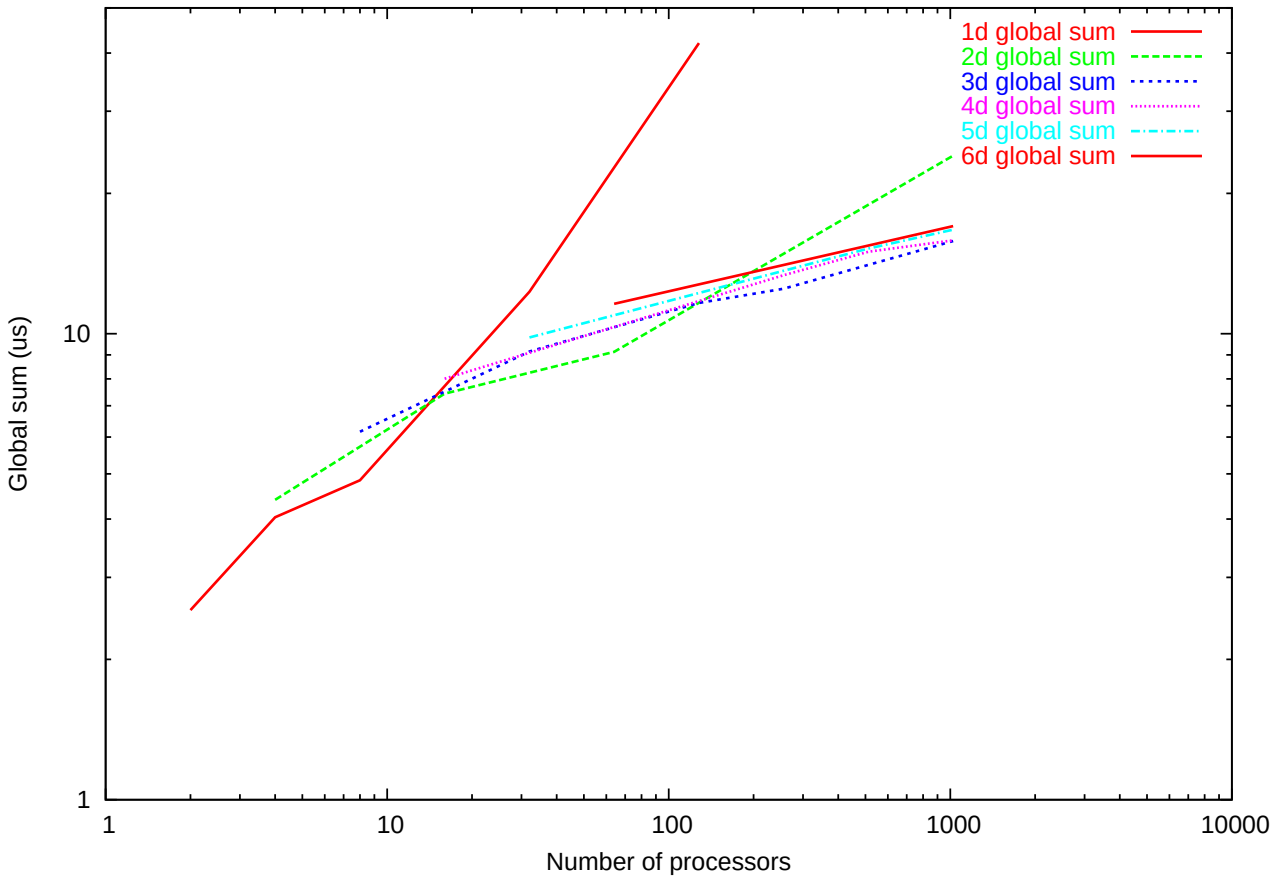


Global reduction

- Hdw acceleration for "All-to-All" along an axis.
- CPU performs arithmetic.

$$\text{GlobalSum} \rightarrow 300\text{ns} \times \frac{1}{2} D (N_{\text{proc}})^{\frac{1}{D}}$$

- Slope of log-log plot asymptotically $\frac{1}{D}$
- 1024 node global sum in *under* 16 μs.
- 20-30 μs (estimated) for $D \geq 3$ on 12k nodes


1d global sum
2d global sum
3d global sum
4d global sum
5d global sum
6d global sum
Global sum (us)
10
1
1
10
100
1000
10000
Number of processors

Application code performance

Various discretisations, 4^4 local volume

Action	Nodes	Sparse Matrix	CG Performance
Wilson	512	44%	39%
AsqTad	128	42%	40%
DWF	512	46%	42%
Clover	512	54%	47%

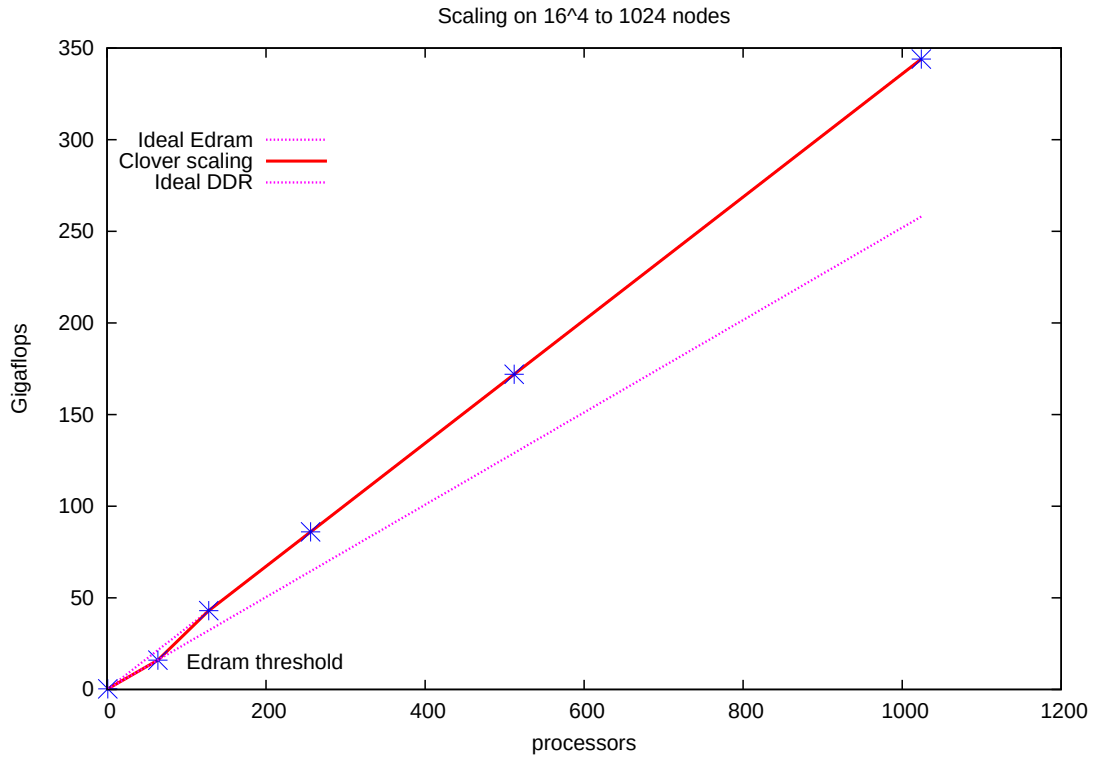
On 2^4 surface to volume ratio is 4

QCDOC nodes can apply Wilson sparse matrix in around 20 μs for 44% of peak with 16 distinct communications.

COLUMBIA RBRC UKQCD QCDOC

Scaling

16^4 on 1k nodes (equivalent to 32^4 on 16k nodes)

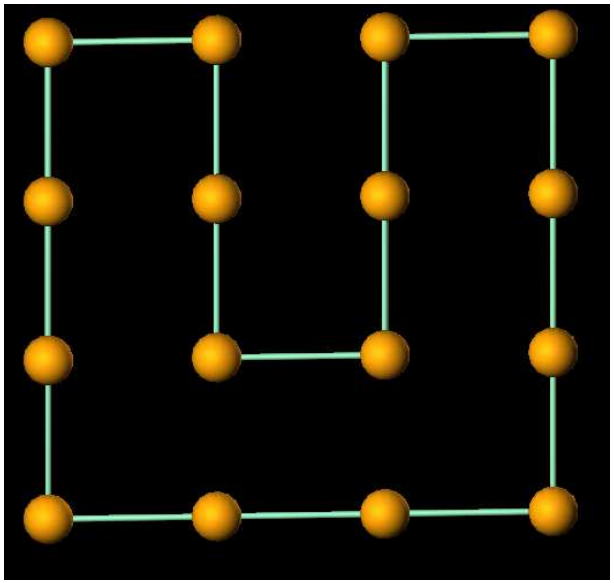


Expect 4 to 5 Tflop/s sustained on large machines

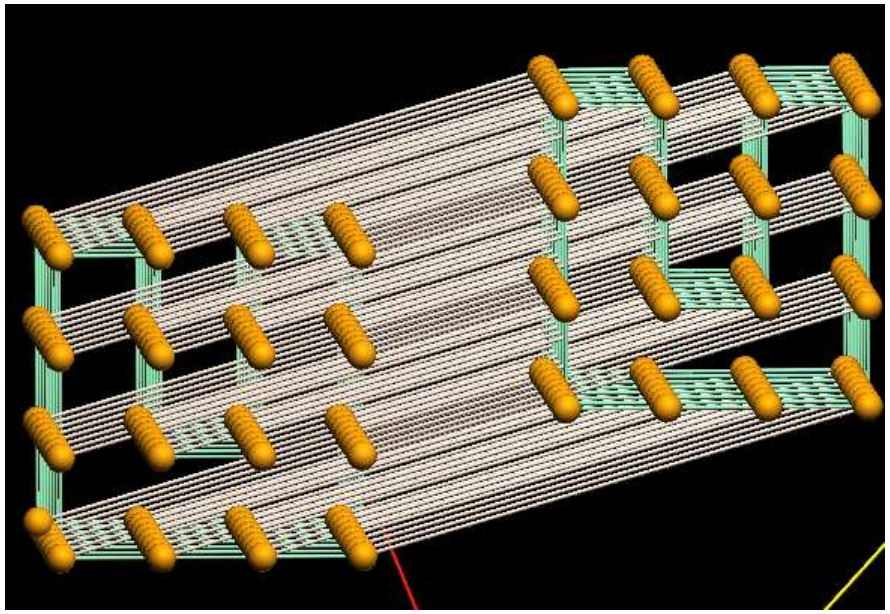
Partitioning

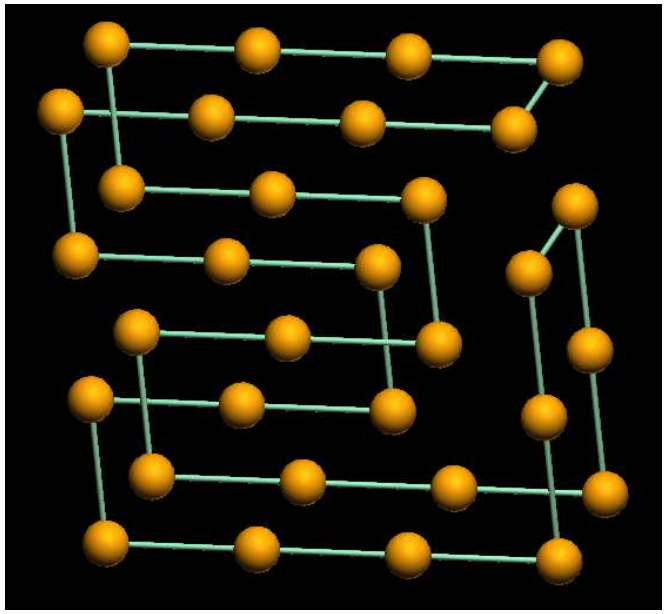
- Machine is a $2k \times 2l \times 2m \times 2 \times 2 \times 2$ hyper-torus.
- Partitions are 6d hyper-rectangular slices of machine
- Locally redefine Application to Machine axis map
- Present 1,2,3,4,5 or 6 dim torus to applications
- Logical nearest neighbor is physically the nearest neighbor

Remapped slice of machine: lower dimensional torus:



- Fold any two dimensions independent of others
- Can iterate to fold multiple dimensions together





Software environment

Custom node kernel

- Integrated diagnostics
- One application process
- No timer driven scheduling
  → avoid scheduler induced slowdown
- Map physical memory in TLB
  → avoid TLB misses
  → Zero-copy DMA
- User space access to communications hardware reduces latency
- POSIX compliant single process subset (Cygwin/newlib libc)
- Custom Ethernet driver
- Custom NFS client provides I/O to host and trivially parallel file system

Multi-threaded *qdaemon* on host boots machine in under 2 minutes
Boot via JTAG download to I/D cache in parallel
Qdaemon provides diagnostics, program load, user interface.

Reliability

- Careful and controlled boot process
- Log and test each memory, device, bus, bridge before use
- Discovers and checks machine topology
- Count and report DDR,Edram ECC errors
- Count and report SCU link errors
- Report checksum mismatches on any link in machine
- Optionally terminate jobs on any correctable errors
- Application level: run 10% reproducibility testing long term
- 100% reproducibility testing during shakeout phase

Compile Chain & Programming environment

- C/C++ with message passing extensions
- Two standard "C" and "C++" compile tools: GCC, XLC
- Library for internode message passing (QMP, SCU)
- QMP is SciDAC QCD cross platform library for message passing
- Lean libraries reduce latency
- Library supports pre-registered channels and multidirectional optimisation
- Software routing for multi-hop packets
- Memory allocation extensions for Edram
  `void *qalloc(int flags,size_t bytes)`
- Standard compliant: benefits everyone
  documentation is *already* written & code is portable

Summary

- Machine is very scalable on Cartesian nearest neighbor problems
- Exceptional internode latency/bandwidth characteristics
- Sub-microsecond ping-pong
  8 million small messages per second
  half maximum on 32 byte packets
- Exceptional global reduction performance
- Good memory performance: 2 Bytes/Flop
- Ideal scaling obtained for a very tightly coupled problem
- Novel dimension folding scheme allows 1d to 6d problems
- Low power and low cost allows very large systems to be built
- \$1 USD per sustained Mflop/s at 5 Tflop/s *on a non-trivial problem*
- 30 peak Tflop/s aggregate on three machines
- Machines will be run 24×7 for 5 years in a dedicated fashion