

Lattice QCD Project Status Report

1/29/2004

- Don Holmgren
- Amitoj Singh
- Eric Neilsen
- Jim Simone
- Valery Sergeev



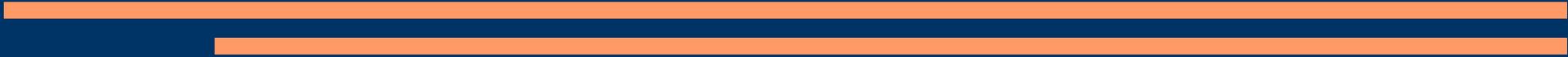
Outline

- Operations
- Near and Far Term Expansions
- SciDAC Software Work
- Plans



Operations

- Overview of installed hardware
- Maintenance, administration, monitoring, alarms
- Documentation
- Utilization
- Physics results

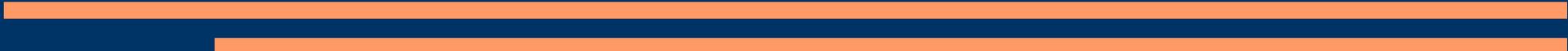


Installed Hardware *“qcd80”*

- Dual 700 Mhz Pentium III cluster
 - Installed November 2000, \$1490/node
 - Purchased from SGI
 - 256 MB memory/node (upgraded from 128 MB)
 - Myrinet 2000 (copper cabling), \$1500/node
 - Uses 80 ports of our first 128-port Myrinet 2000 switch
 - Remote hardware management via IPMI
 - Integrated BMC (baseboard management controller)
 - Out-of-band communications via serial port
 - Purchased via a supplemental DOE grant and CD funds
 - Out of warranty as of November 2003
 - Very lightly used
 - Several nodes are no longer usable
 - Plan to remove from Myrinet fabric (more later)
-
-

Installed Hardware *1st SciDAC Cluster - "nqcd"*

- Dual 2.0 Ghz Xeon cluster
 - 50 nodes, installed July 2002, \$1594/node
 - Purchased from SteelCloud, Reston, VA
 - 1 GB memory per node
 - Myrinet 2000 (fiber cabling), \$1400/node
 - Uses 48 ports of first 128-port Myrinet 2000 switch
 - IPMI using add-in BMC from SuperMicro
 - Out-of-band management over ethernet
 - Not nearly as reliable as we'd hoped



Installed Hardware

2nd SciDAC Cluster - "w"

- Dual 2.4 GHz Xeon cluster
 - Installed January 2003, \$1704/node
 - Purchased from CSI, Alpharetta, GA
 - 1 GB memory per node
 - Myrinet 2000 (fiber cabling), \$1300/node
 - Uses all 128 ports of our 2nd Myrinet 128-port switch
 - IPMI via same SuperMicro BMC option card

Maintenance/Administration

- Common to each of the clusters:
 - PBS batch system
 - Maui scheduler
 - Automated network-based installs
 - Using PXE, implemented late 2000
 - Also, netboot into DOS for firmware, BIOS
 - FNAL IPMI software
 - Read sensors, motherboard logs from Linux
 - Reset, power on/off/cycle from serial line (Holmgren, 1999) or ethernet (Singh, 2003)

Monitoring

- “Nannies” (Amitoj Singh)
 - Clients on workers, server on head node
 - Clients “write only”, send data to server via SYSLOG (UDP)
 - Clients monitor:
 - Health (temperatures, fans)
 - Uptime
 - PBS client status
 - CPU status (frequency)
 - Disk space
 - Myrinet health
-
-

Monitoring

- Server Nanny:
 - Parses Syslog every 2 minutes
 - Generates MRTG health and FLOP plots
 - Monitors average temperatures and declares temperature alarms
 - Interacts with NGOP to send temperature alarms
 - Monitors RAID boxes
 - Over serial lines (ugh)
 - Agent written by Daiya Miazato
 - Monitors PBS server, Maui scheduler
 - Monitors NFS server
 - Issues mail alerts
-
-

Web Interfaces

File Edit View Go Communicator Help

Back Forward Reload Home Search Netscape Print Security Shop Stop

Location: <http://lqcd/>

Internet Google Dejanews Lookup New&Cool

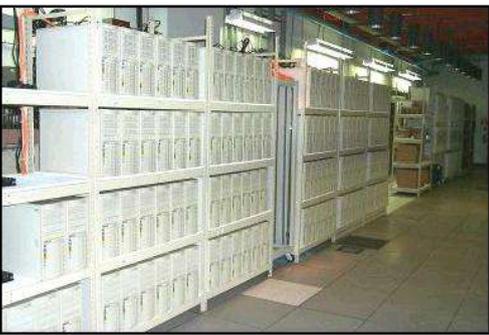


Fermilab Fermilab at Work Theoretical Physics Dept. Distributed Systems Projects Group, Integrated Systems Development Dept., Fermilab O

Tue Jan 27 11:57:06 CST 2004

				1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	2	2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
2	2	2	2	2	2	3	3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
3	3	3	3	3	3	4	4	4	4	4	4	4	4	4
4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
4	4	4	4	4	4	5	5	5	5	6	6	6	6	6
7	7	7	7	7	7	7	7	7	7	7	7	7	7	7

LQCD System Status



The QCD clusters

webmaster@lqcd.fnal.gov

File Edit View Go Communicator Help

Back Forward Reload Home Search Netscape Print Security Shop Stop

Location: <http://lqcd.fnal.gov/cgi-bin/stat>

Internet Google Dejanews Lookup New&Cool

lqcd.fnal.gov Status

Tue Jan 27 11:57:06 CST 2004

w101	w102	w103	w104	w105	w106	w107	w108	w109	w110	w111
w112	w113	w114	w115	w116	w117	w118	w119	w120	w121	w122
w123	w124	w201	w202	w203	w204	w205	w206	w207	w208	w209
w210	w211	w212	w213	w214	w215	w216	w217	w218	w219	w220
w221	w222	w223	w224	w301	w302	w303	w304	w305	w306	w307
w308	w309	w310	w311	w312	w313	w314	w315	w316	w317	w318
w319	w320	w321	w322	w323	w324	w401	w402	w403	w404	w405
w406	w407	w408	w409	w410	w411	w412	w413	w414	w415	w416
w417	w418	w419	w420	w421	w422	w423	w424	w501	w502	w503
w504	w505	w506	w507	w508	w509	w510	w511	w512	w513	w514
w515	w516	w517	w518	w519	w520	w521	w522	w523	w524	w601
w602	w603	w604	w605	w606	w607	w608	nqcd0101	nqcd0102	nqcd0103	nqcd0104
nqcd0105	nqcd0106	nqcd0201	nqcd0202	nqcd0203			nqcd0206	nqcd0301	nqcd0302	nqcd0303
nqcd0304	nqcd0305	nqcd0306	nqcd0401	nqcd0402	nqcd0403	nqcd0404	nqcd0405	nqcd0406	nqcd0501	
	nqcd0504	nqcd0505	nqcd0506	nqcd0601	nqcd0602	nqcd0603	nqcd0604	nqcd0605	nqcd0606	nqcd0701
nqcd0702	nqcd0703	nqcd0704	nqcd0705	nqcd0706	nqcd0801	nqcd0802	nqcd0803	nqcd0804	nqcd0805	nqcd0806

free : 65 down : 0 offline : 7 reserve : 0 job-exclusive : 104 job-sharing : 0 Usage 61%

Job List

Ref Id	Job Id	Job Name	User	Time Use	S	Queue	Nodes
	926075	launch.nos	Don Holmgren	--	R	workq	64
2	926070	gluon_stuf	Matthew Nobes	00:00	R	workq	32
3	926064	pQuen	Jim Simone	01:08	R	workq	4
4	926063	pQuen	Jim Simone	01:10	R	workq	4

Temperature **COOL**

[latest node health statistics](#) [MRTG health plots](#) [uptime](#) [flop](#) [qcdhome](#)

Refresh Rate : 60 secs

last modified 11/20/2003 lqcd-admin@fnal.gov

Web Interfaces

File Edit View Go Communicator Help

Back Forward Reload Home Search Netscape Print Security Shop Stop

Bookmarks Location: <http://lqcd.fnal.gov/cgi-bin/stat?health=uptime> What's Related

Internet Google Dejanews Lookup New&Cool

uptime Status

Tue Jan 27 12:01:44 CST 2004

Hostname	Uptime Statistics	Load Average
nqcd0102	11:55am up 73 days, 18:09	0.00, 0.00, 0.00
nqcd0103	11:55am up 73 days, 18:20	0.00, 0.00, 0.00
nqcd0104	11:56am up 73 days, 18:23	0.00, 0.00, 0.00
nqcd0105	11:56am up 34 days, 20:55	0.00, 0.00, 0.00
nqcd0106	11:56am up 73 days, 18:23	0.00, 0.00, 0.00
nqcd0201	11:56am up 63 days, 19:40	0.00, 0.00, 0.00
nqcd0202	11:57am up 73 days, 18:22	0.00, 0.00, 0.00
nqcd0203	11:57am up 63 days, 22:22	0.00, 0.00, 0.00
nqcd0204	11:57am up 73 days, 1:36	0.00, 0.00, 0.00
nqcd0205	11:57am up 73 days, 18:20	0.00, 0.00, 0.00
nqcd0206	11:58am up 63 days, 19:43	0.00, 0.00, 0.00
nqcd0301	11:56am up 39 days, 36 min	0.00, 0.00, 0.00
nqcd0302	11:59am up 63 days, 19:42	0.00, 0.00, 0.00
nqcd0303	11:57am up 62 days, 23:41	0.00, 0.00, 0.00
nqcd0304	11:58am up 63 days, 19:42	0.08, 0.02, 0.01
nqcd0305	11:58am up 63 days, 19:42	0.00, 0.00, 0.00
nqcd0306	11:58am up 64 days, 5:39	0.00, 0.00, 0.00
nqcd0401	11:57am up 73 days, 18:02	0.00, 0.00, 0.00
nqcd0402	11:59am up 46 days, 19:21	0.00, 0.00, 0.00
nqcd0403	11:59am up 73 days, 18:23	0.00, 0.00, 0.00
nqcd0404	11:59am up 73 days, 18:09	0.00, 0.00, 0.00
nqcd0405	11:57am up 73 days, 18:20	0.00, 0.00, 0.00
nqcd0406	11:59am up 73 days, 18:23	0.00, 0.00, 0.00
nqcd0501	10:06am up 73 days, 16:23	0.00, 0.00, 0.00
nqcd0502	11:55am up 1:20	0.00, 0.00, 0.00
nqcd0503	11:55am up 1:20	0.00, 0.00, 0.00
nqcd0504	11:54am up 1:20	0.00, 0.00, 0.00
nqcd0505	10:06am up 73 days, 16:12	0.00, 0.00, 0.00
nqcd0601	11:59am up 64 days, 5:36	0.00, 0.00, 0.00
nqcd0602	11:57am up 46 days, 21:47	0.00, 0.00, 0.00
nqcd0603	11:59am up 46 days, 21:50	0.20, 0.17, 0.07
nqcd0604	11:56am up 46 days, 21:46	0.00, 0.00, 0.00
nqcd0605	11:58am up 63 days, 19:46	0.24, 0.14, 0.05
nqcd0606	11:57am up 63 days, 19:40	0.73, 0.17, 0.06
nqcd0701	11:56am up 73 days, 18:18	0.00, 0.00, 0.00
nqcd0702	11:59am up 63 days, 19:42	0.13, 0.12, 0.05
nqcd0703	11:59am up 63 days, 19:42	0.16, 0.13, 0.05
nqcd0704	11:57am up 63 days, 19:40	0.23, 0.05, 0.02

Web Interfaces

File Edit View Go Communicator Help

Back Forward Reload Home Search Netscape Print Security Shop Stop

Location: <http://lqcd.fnal.gov/cgi-bin/stat?health=MRTG=w603> What's Related

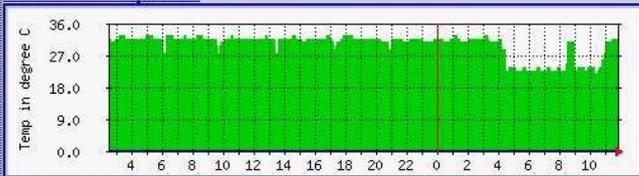
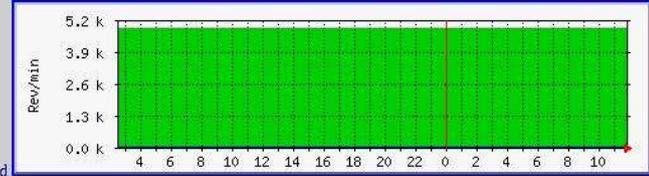
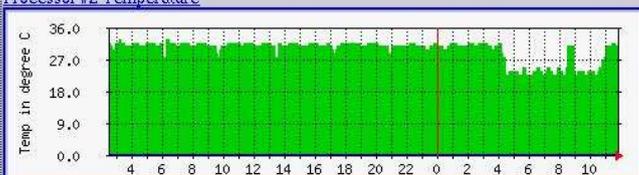
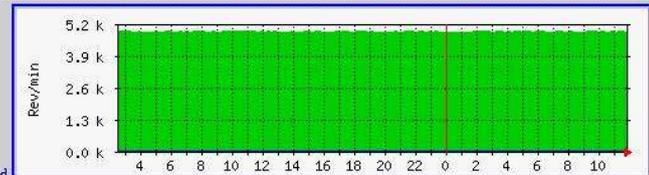
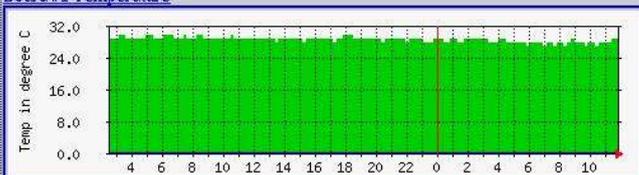
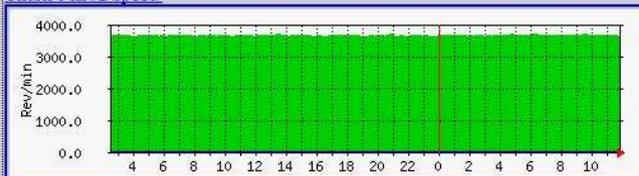
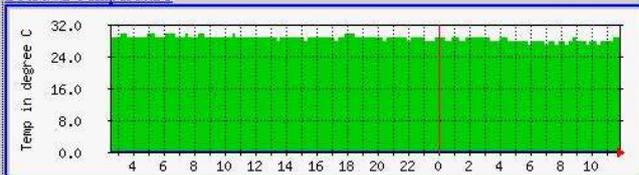
Internet Google Dejanews Lookup New&Cool

MRTG Status

Tue Jan27 11:59:17 CST 2004

MRTG Plots for w603



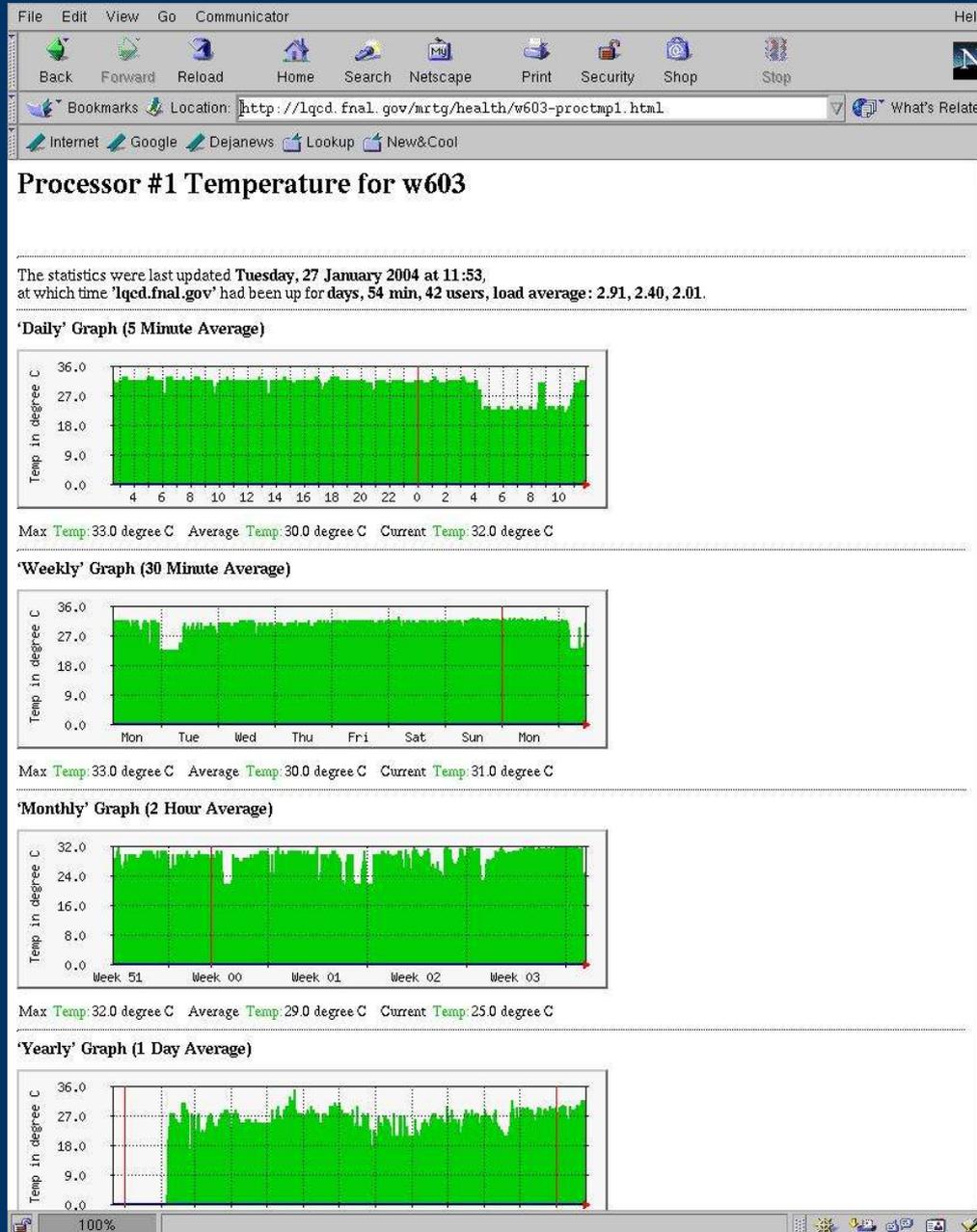
<p><u>Processor #1 Temperature</u></p> 	<p><u>Fan #1 Speed</u></p> 
<p><u>Processor #2 Temperature</u></p> 	<p><u>Fan #2 Speed</u></p> 
<p><u>Board #1 Temperature</u></p> 	<p><u>Chasis Fan #2 Speed</u></p> 
	<p><u>Board #2 Temperature</u></p> 

Hostname:

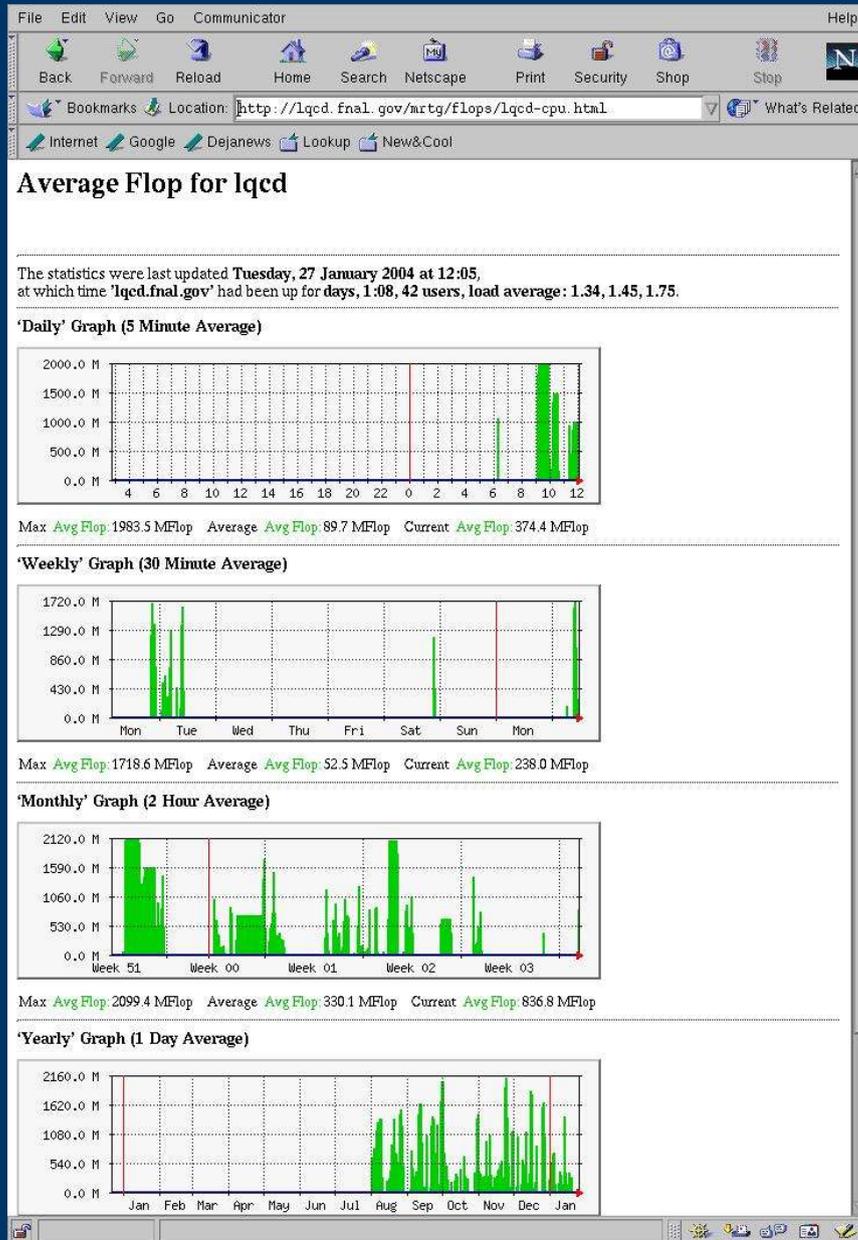
[latest node health statistics](#) [MRTG health plots](#) [uptime](#) [flop](#) [gcdhome](#)

100%

Web Interfaces



Web Interfaces



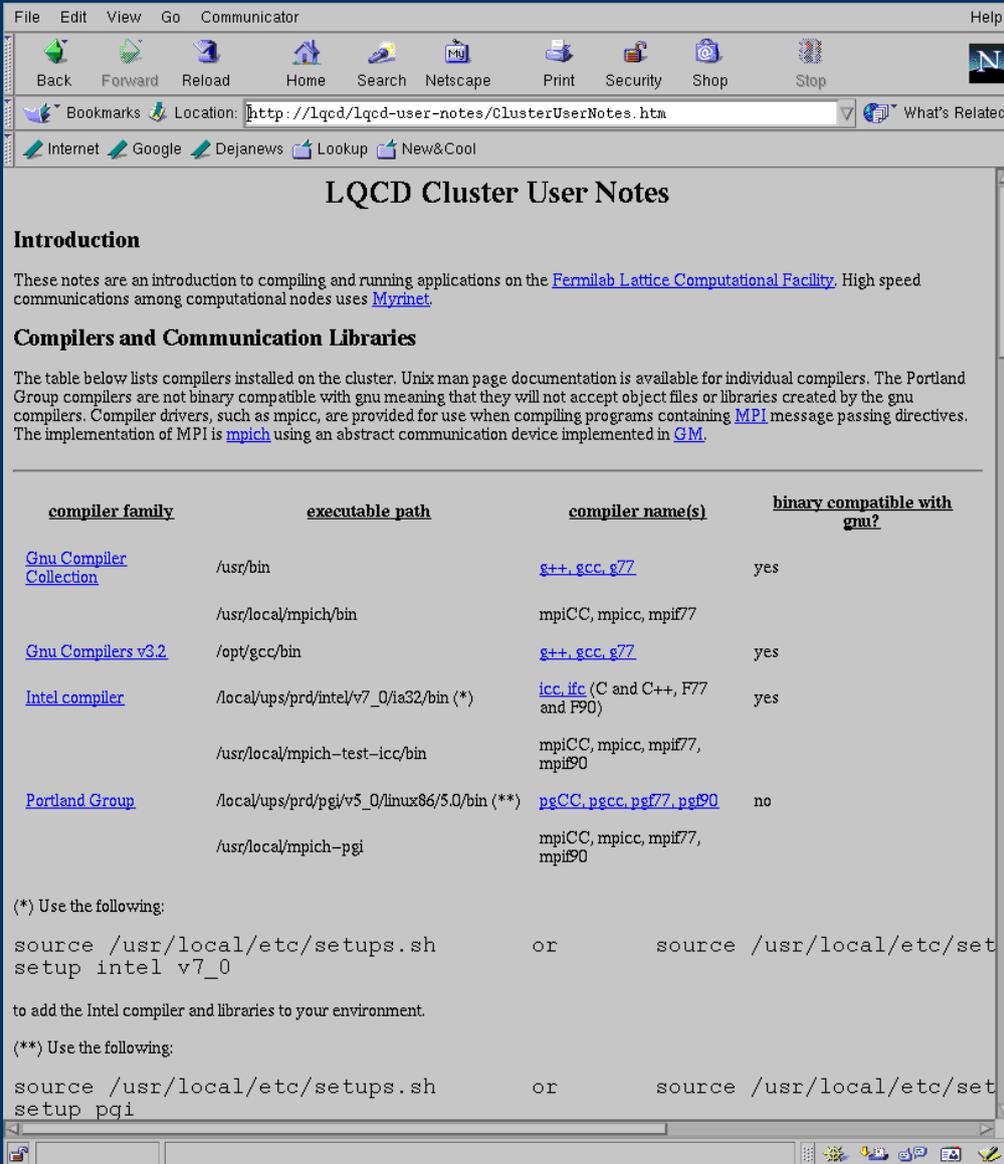
- FLOP counter
 - Raw data from processor performance counters via FNAL TRACE software
 - Graphs show average FLOPS/sec/node for the 176 SciDAC nodes
 - Not quite working:
 - Busy processors report infrequently
 - Total FLOPS are correct, the rates are not
 - Further Kernel mods necessary to fix

Temperature Alarms

- New Muon chillers and Lieberts are not monitored by the site system
 - A high temperature alarm is raised if either:
 - A large number of worker nodes report high temperatures. Automated shutdown will occur after a set period (2 hours) of temperature over threshold.
 - An automated temperature sensor trips and phones the operators and helpdesk
 - Responses:
 - During normal working hours, handled by LQCD project team
 - Outside NWH, operators will call Jack MacNerland and the duty mechanic. ISA group primary will monitor and if necessary manually power down systems at New Muon
 - Incidents this year:
 - Have had 2 chiller failures during NWH, fixed by reset
 - Have had 1 failure during off-hours, reset by duty mechanic
-
-

User Documentation

- User notes
 - Compilers
 - MPI
 - Using the batch system
 - Mass storage
 - Kerberos
 - Getting accounts



The screenshot shows a Netscape browser window with the address bar containing `http://lqcd.lqcd-user-notes/ClusterUserNotes.htm`. The page title is "LQCD Cluster User Notes".

Introduction

These notes are an introduction to compiling and running applications on the [Fermilab Lattice Computational Facility](#). High speed communications among computational nodes uses [Myrinet](#).

Compilers and Communication Libraries

The table below lists compilers installed on the cluster. Unix man page documentation is available for individual compilers. The Portland Group compilers are not binary compatible with gnu meaning that they will not accept object files or libraries created by the gnu compilers. Compiler drivers, such as mpicc, are provided for use when compiling programs containing MPI message passing directives. The implementation of MPI is [mpich](#) using an abstract communication device implemented in [GM](#).

compiler family	executable path	compiler name(s)	binary compatible with gnu?
Gnu Compiler Collection	/usr/bin	g++ , gcc , g77	yes
	/usr/local/mpich/bin	mpiCC, mpicc, mpif77	
Gnu Compilers v3.2	/opt/gcc/bin	g++ , gcc , g77	yes
Intel compiler	/local/ups/prd/intel/v7_0/ia32/bin (*)	icc , ifc (C and C++, F77 and F90)	yes
	/usr/local/mpich-test-icc/bin	mpiCC, mpicc, mpif77, mpif90	
Portland Group	/local/ups/prd/pgi/v5_0/linux86/5.0/bin (**)	pgCC , pgcc , pgf77 , pgf90	no
	/usr/local/mpich-pgi	mpiCC, mpicc, mpif77, mpif90	

(*) Use the following:

```
source /usr/local/etc/setup.sh or source /usr/local/etc/set
setup intel v7_0
```

to add the Intel compiler and libraries to your environment.

(**) Use the following:

```
source /usr/local/etc/setup.sh or source /usr/local/etc/set
setup pgi
```

System Documentation

- Lattice QCD Twiki
 - Very useful!
 - Where we store:
 - Startup, shutdown instructions
 - System maintenance notes
 - Design notes
 - Internal HOWTOs

File Edit View Go Communicator Help

Back Forward Reload Home Search Netscape Print Security Shop Stop

Bookmarks Location: <http://lqcd.twiki/bin/view/Lqcd/WebHome> What's Related

Internet Google Dejanews Lookup New&Cool

WebHome

TWiki . Lqcd . WebHome

Logged in as TWikiGuest

People

TWiki System

Lqcd Web

WebChanges

WebIndex

WebNotify

WebStatistics

Knowledge Web

Sandbox

Welcome to the home of TWiki.Lqcd. This is a repository for informal documentation and folklore relating to the SciDAC LatticeQCD cluster at NewMuon at Fermilab. See the [InstructionsForNewUsers](#) for more information on starting to use this wiki.

Information about the SciDAC LatticeQCD Cluster at NewMuon

More formal documentation can be found at <http://lqcd.fnal.gov>. Other informal information can be found in the archives of the [lqcd-users](#) and [lqcd-admin](#) mailing lists.

- [UsefulClusterSoftwareExternalDocumentation](#)
- [ClusterHardwareHistory](#)
- [NewMuonClusterOverview](#)
- [HowTo](#) documents
- [ShellHistories](#) and scripts showing operations of interest
- [ClusterAdminNotes](#)
- [LqcdBackupPlan](#)
- [LatticeQCDAundryList](#)
- [AcronymList](#)
- [OutsideArchives](#) for lattice QCD
- [ClusterDCacheNotes](#)
- [InternationalLatticeDataGrid \(ILDG\)](#), including SRM plans

Site Tools of the Lqcd Web

Search (More options in [WebSearch](#))

- [WebChanges](#): Display recent changes to the Lqcd web
- [WebIndex](#): List all Lqcd topics in alphabetical order. See also the faster [WebTopicList](#)
- [WebNotify](#): Subscribe to an e-mail alert sent when something changes in the Lqcd web
- [WebStatistics](#): View access statistics of the Lqcd web
- [WebPreferences](#): Preferences of the Lqcd web ([TWikiPreferences](#) has site-wide preferences)

Notes:

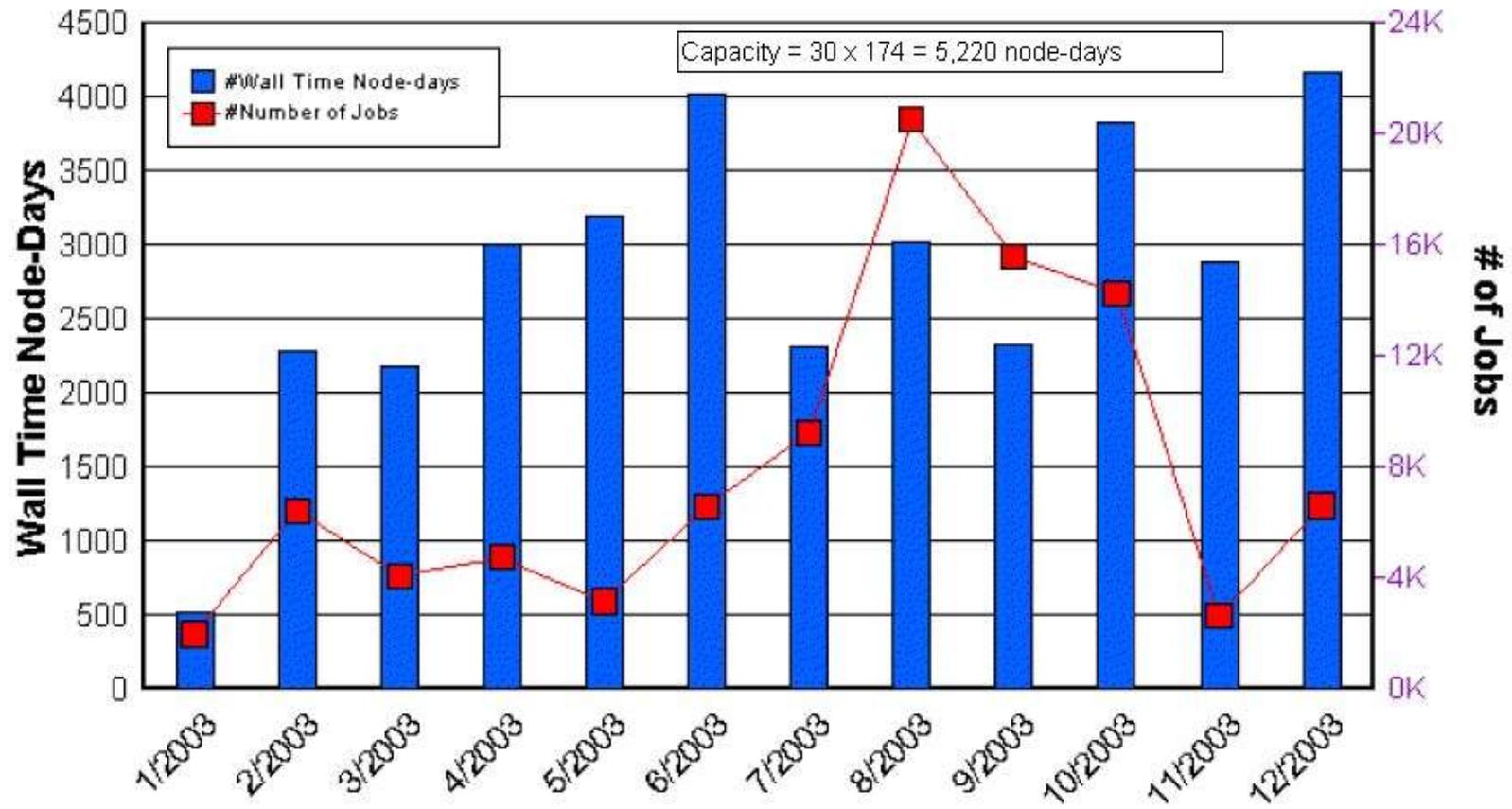
100%

Utilization

- Time on the SciDAC clusters (“nqcd”, “w”) is allocated by a scientific program committee
 - Current allocations: Cornell (incl. Automated perturbation theory by SFU), Charmonium, Heavy-Lite
 - Accounting controls are in place, but have not been necessary to date because peak demand has not exceeded capacity
 - Time on the pre-SciDAC cluster was available to interested parties. Usage this year by:
 - MILC (Carleton Detar - thermodynamics)
 - SDSS (“co-adds”)
 - SciDAC Accelerator Simulation
 - CDMS (in the last week)
-
-

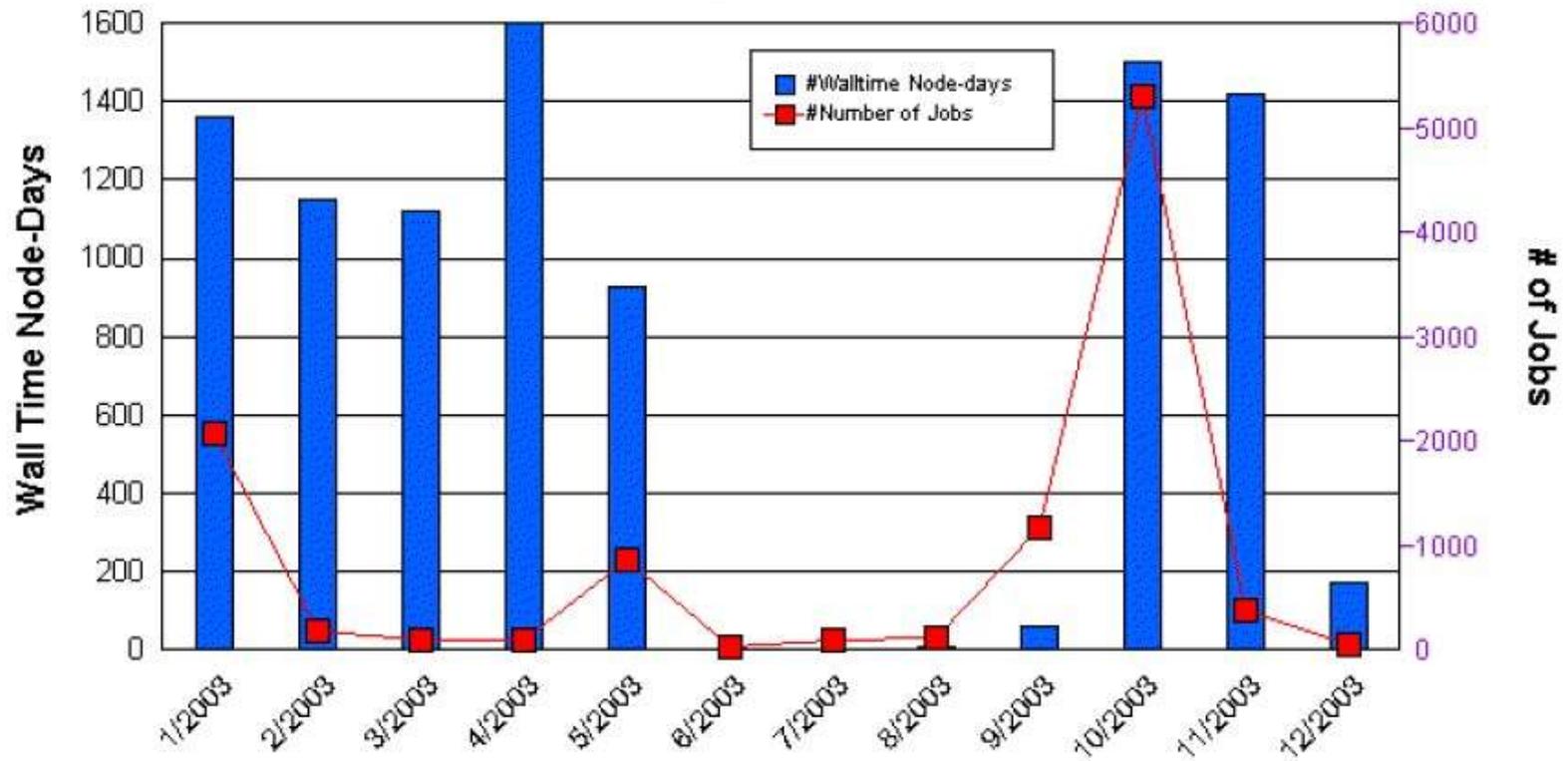
Utilization

Lattice QCD Cluster - SCIDAC



Utilization

Lattice QCD Cluster - General



Lattice QCD Cluster - General

Physics Results

VOLUME 92, NUMBER 2

PHYSICAL REVIEW LETTERS

week ending
16 JANUARY 2004

High-Precision Lattice QCD Confronts Experiment

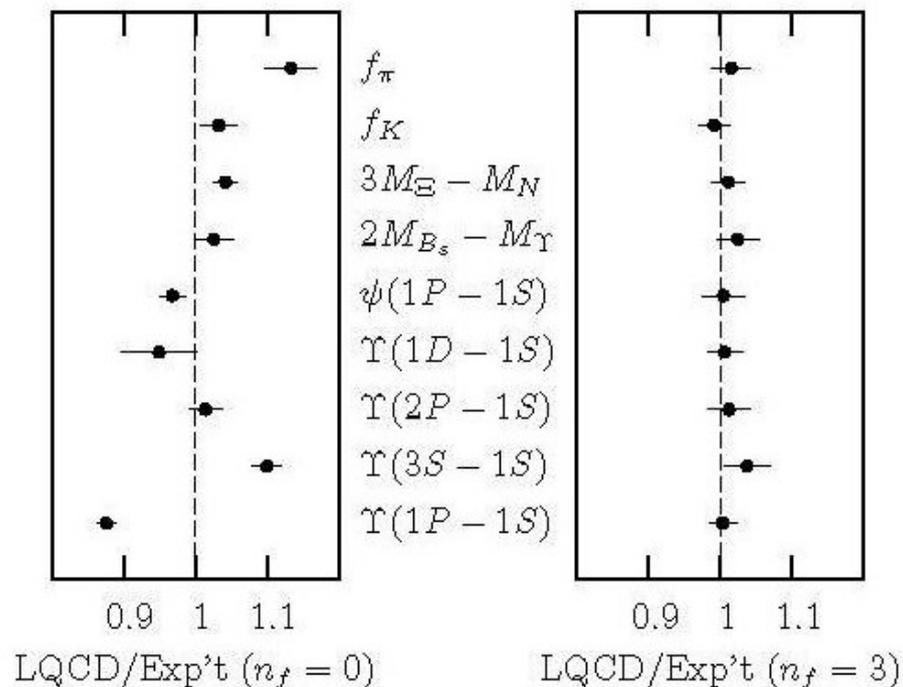
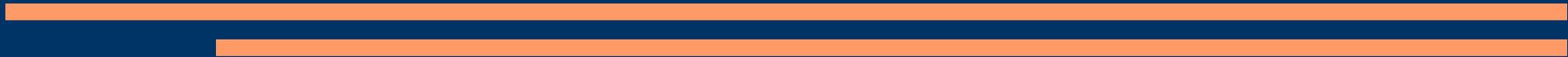


FIG. 1: LQCD results divided by experimental results for nine different quantities, without and with quark vacuum polarization (left and right panels, respectively). The top three results are from our $a = 1/11$ and $1/8$ fm simulations; all others are from $a = 1/8$ fm simulations.

- Papers with data generated partially on FNAL clusters:
 - PRL Article: “High-Precision Lattice QCD Confronts Experiment”
 - 31 Citations according to SPIRES
 - Anticipate *Nature* article on this paper in early February
 - Discussed in *Science*, “Calculating the Incalculable” 16 May 2003, 300, 1076.
 - Physics Today, Feb 2004, “Lattice Chromodynamics Comes of Age”

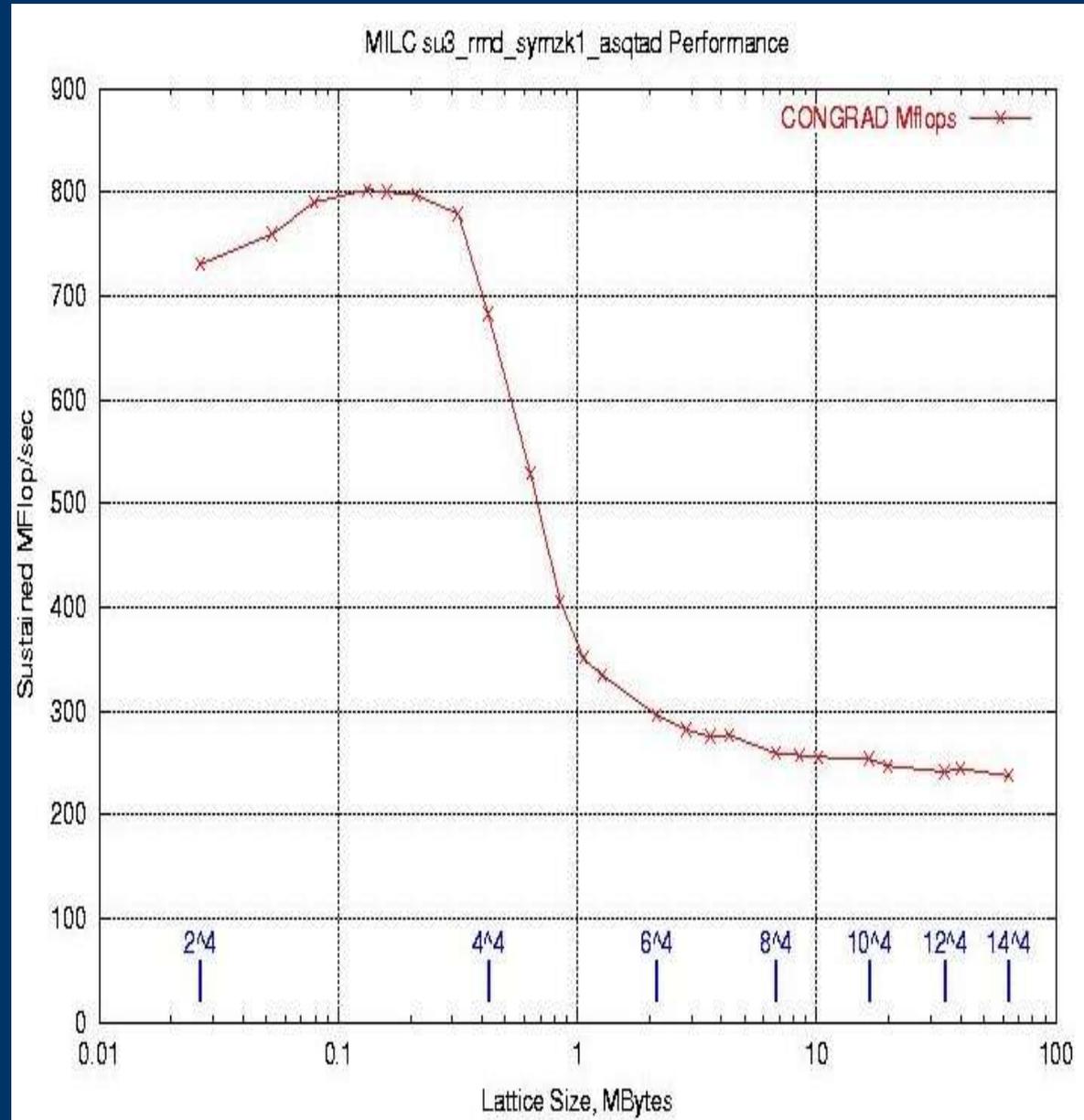
Near and Long Term Expansions

- Designing Clusters:
 - Lattice QCD Constraints
 - Processor Choices
 - Network Choices
- Winter 2004 Upgrade
- Fall 2004 Upgrade



Lattice QCD Constraints

- Lattice code is:
 - Floating point intensive
 - Memory BW intensive
 - Communications intensive
- Typical single node “fingerprint” is shown:
 - Graph shows sustained floating point performance as a function of lattice size
 - Best floating point performance in cache
 - L2 size is 512KB
 - On a cluster, typically run at lattice sizes of 10 MBytes or more per node



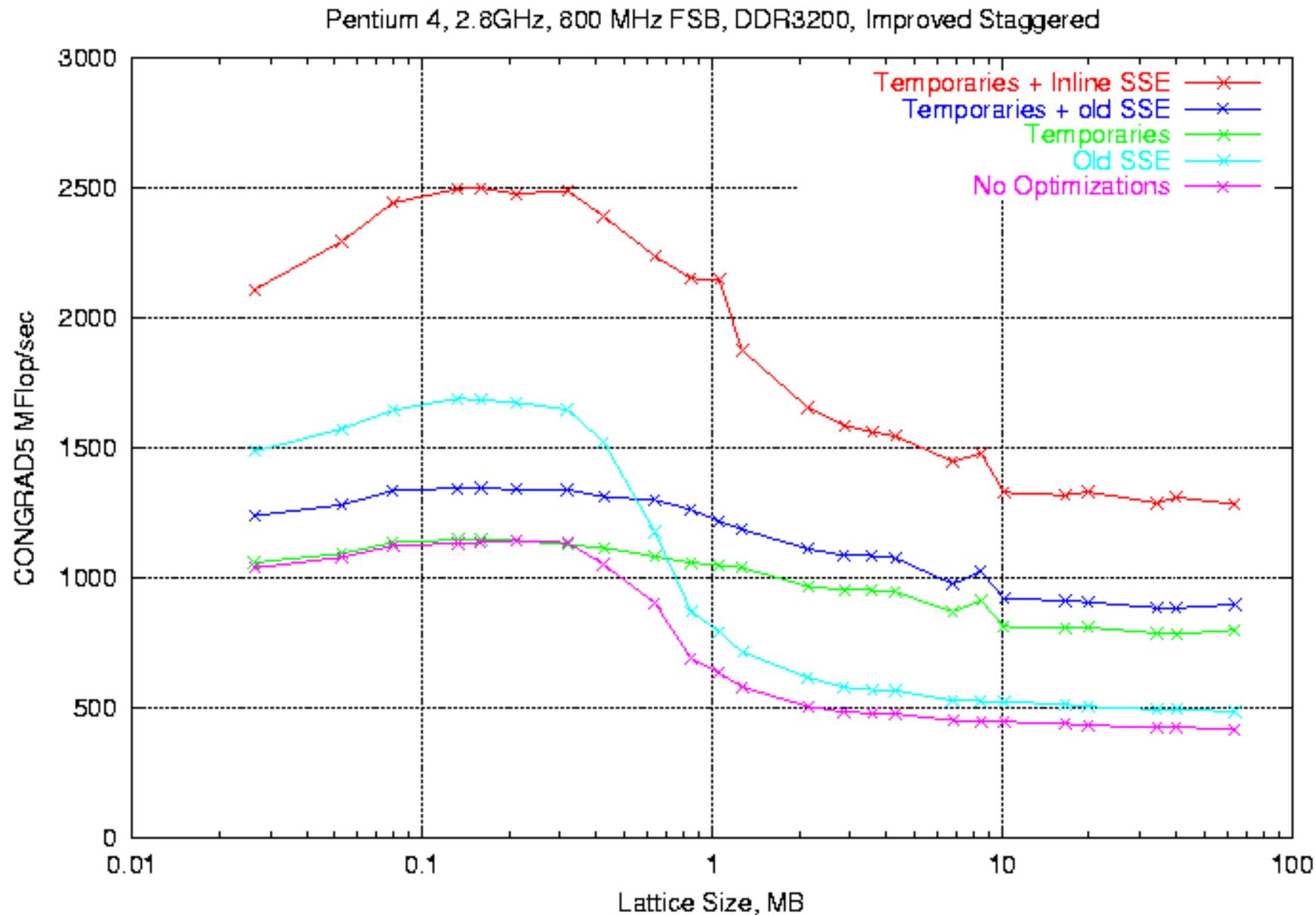
CPU Performance Regimes

- In cache:
 - Dominated by floating point throughput
 - Nearly all FLOPS are SU3 matrix-vector multiplies (complex 3X3 matrices, complex 3x1 vectors)
 - Strong improvements result from assembly language techniques, including SSE (x86), 3DNow (AMD), and AltiVec (IBM/Motorola PPC)
 - Cluster performance is constrained by network performance, especially latency
 - In main memory:
 - Dominated by memory bandwidth
 - Strong improvements result from careful data layout, cache management, prefetching
 - Cluster performance is constrained by memory and I/O bandwidth
-
-

CPU Choices

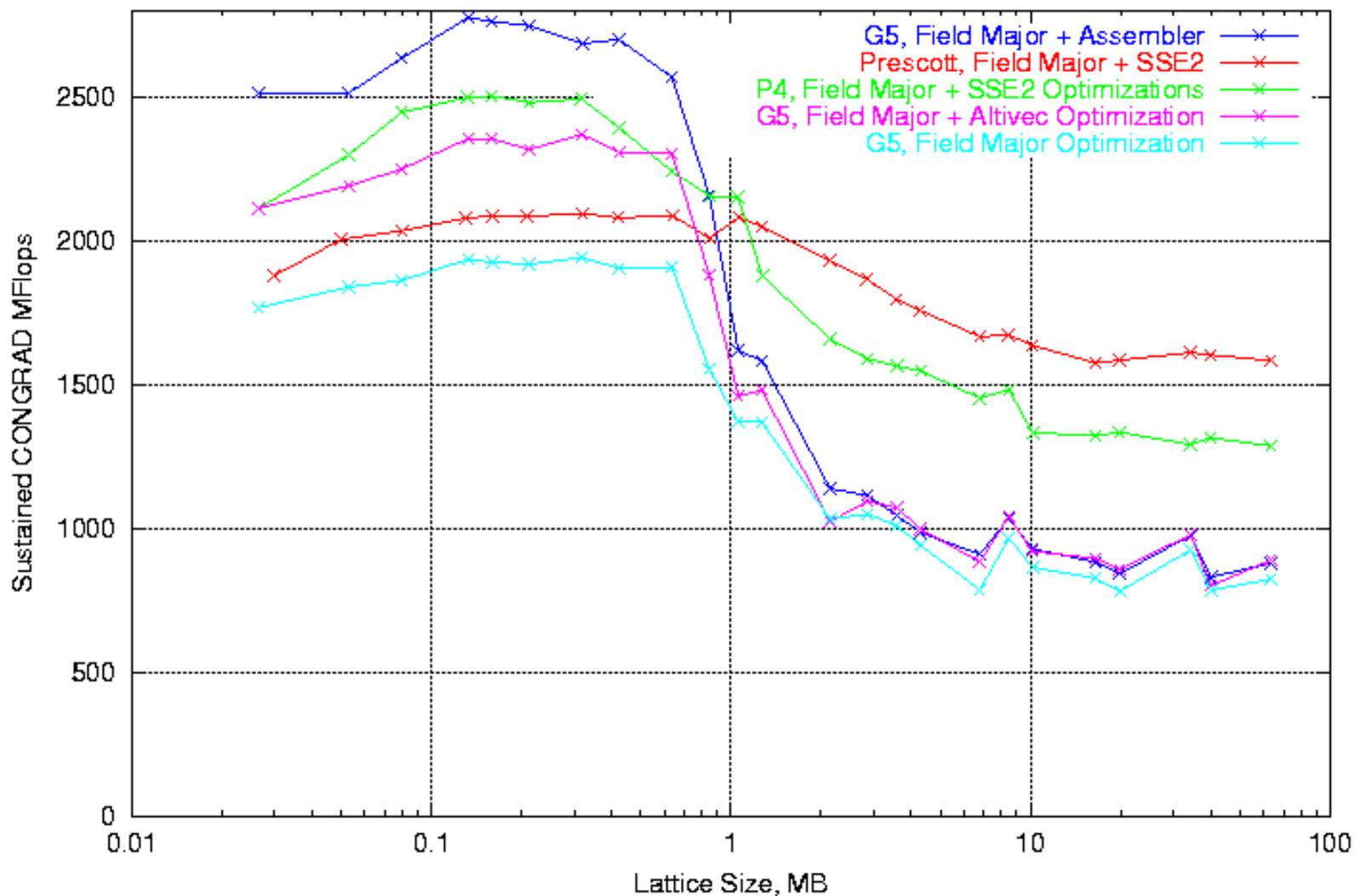
- Intel
 - Xeon
 - SMP systems share slower memory bus
 - Best PCI buses
 - Pentium 4
 - Highest available memory bandwidth
 - Until this quarter, poor PCI buses
 - Pentium 4E (“Prescott”)
 - Intriguingly good main memory performance
 - Itanium2
 - Expensive, best performance but most difficult to code
 - AMD
 - Opteron
 - SMP systems have better memory BW than dual Xeon
 - Athlon64
 - Not tested yet
 - IBM
 - PowerPC 970 (aka Apple G5)
 - Best in-cache performance
 - Disappointing memory bandwidth
-
-

CPU Optimizations



CPU Comparisons

MILC Improved Staggered Performance, 2.8 GHz P4 vs 2.8 GHz Prescott vs 2.0 GHz G5

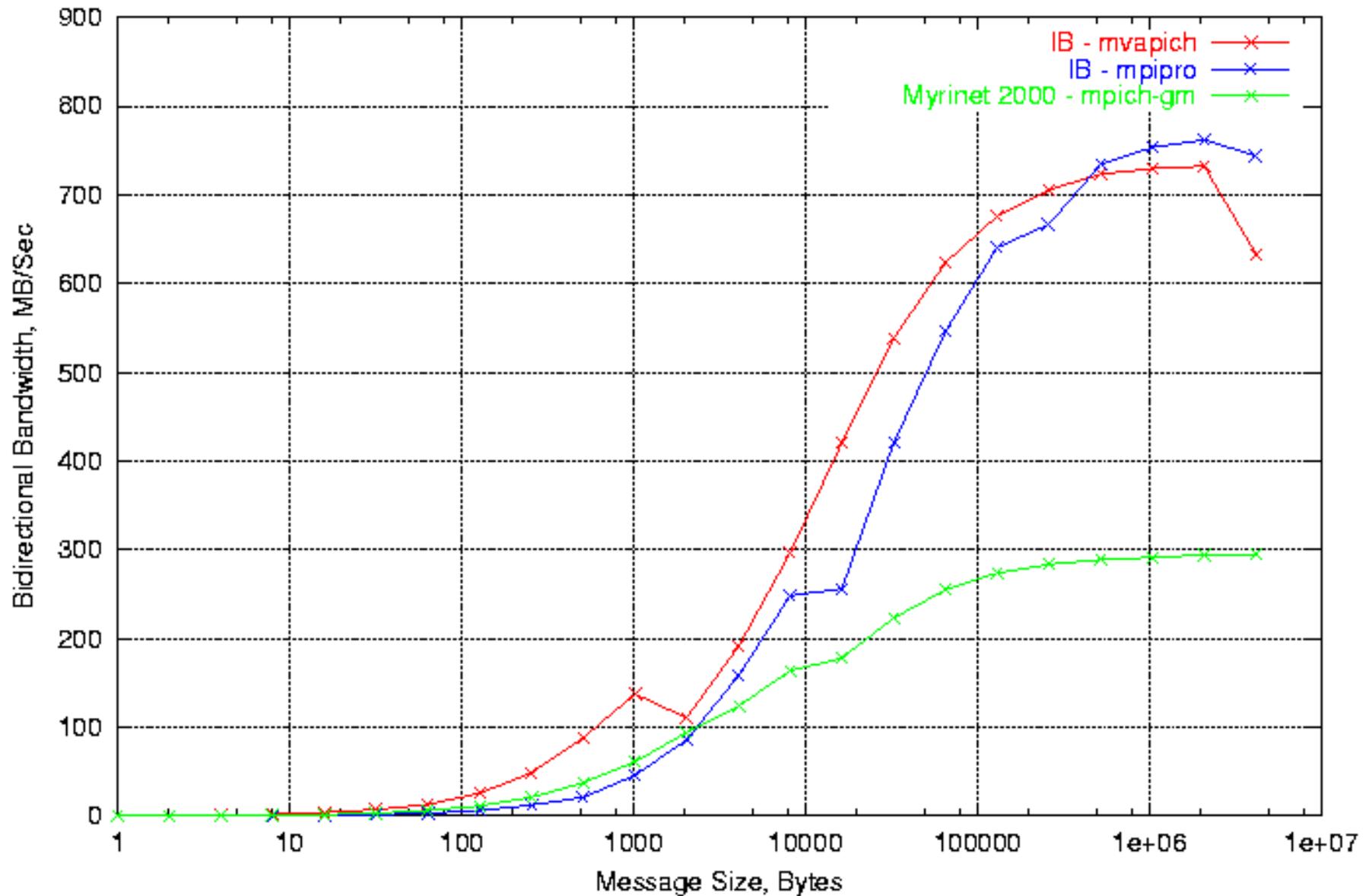


Network Choices

- Ethernet
 - Fast ethernet was already too slow for Pentium II
 - Multiple GigE NICs per node fast enough
 - Switches are expensive and have high latency
 - Meshes are under investigation (JLAB)
 - Myrinet
 - Most common fabric for mid-size HPC clusters
 - Costs as much as the node
 - Reaching bandwidth limits (2.5 Gbps)
 - Infiniband
 - 4X bandwidth now (10 Gbps), 12X next year
 - Multi-vendor
 - Multi-application (HPC is only part of market)
 - Others
 - Quadrics: \$\$\$, ASCI favorite
 - SCI: 3D torus, Myrinet-like cost, infrequently used
-
-

Network Performance

Pallas Sendrecv Benchmark, Infiniband vs Myrinet



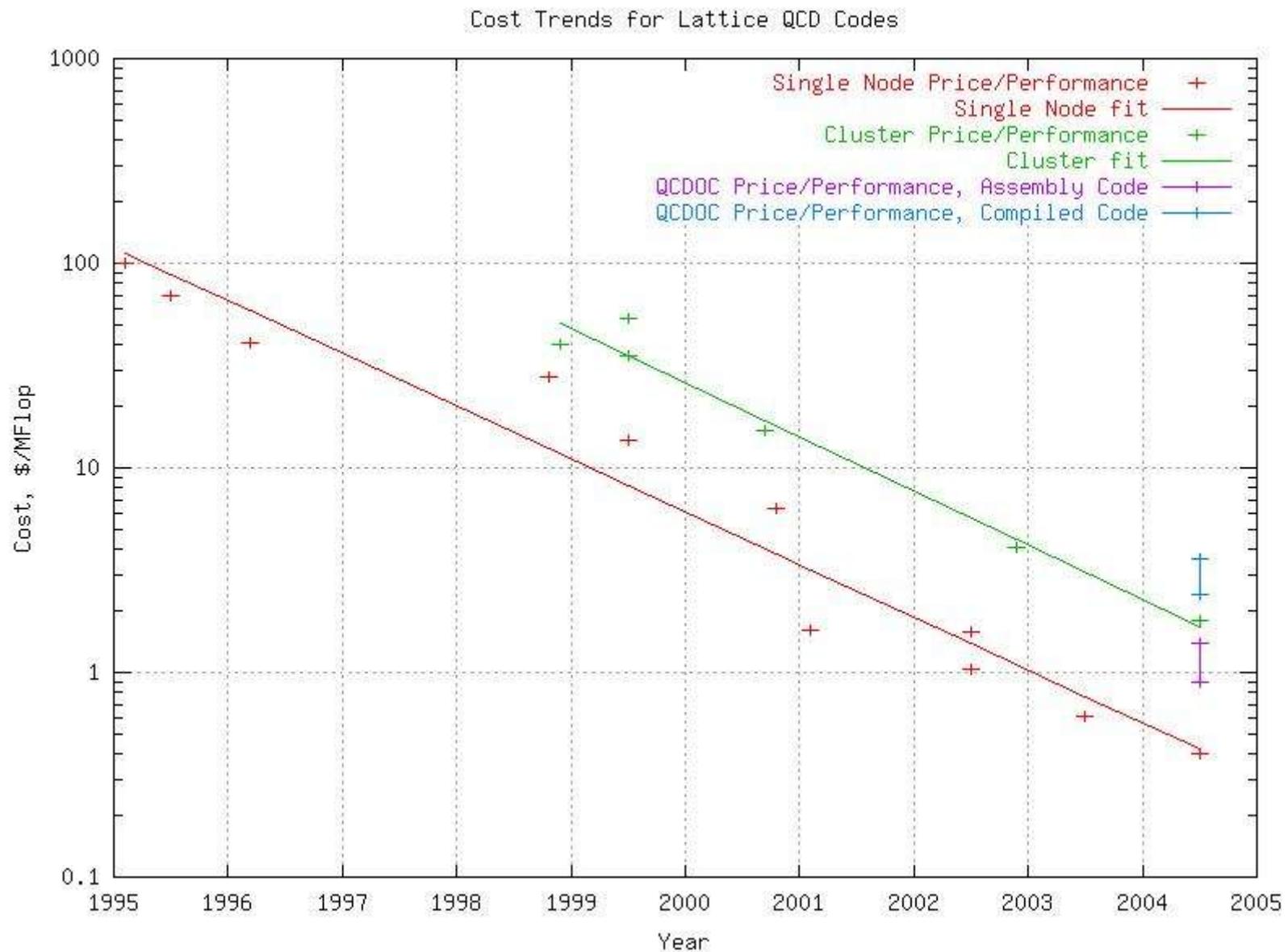
Winter Upgrade

- Move “qcd80” off of Myrinet fabric
 - Use switched fast ethernet fabric
 - Likely useful for automated perturbation theory
 - Replace Pentium III nodes with Pentium 4E nodes
 - About \$900/node (1 GB memory, 3.0 Ghz)
 - Evaluating beta Intel motherboards this week
 - First re-use of network fabric (a practice posited long ago)
 - Incremental advantage: gain 1 Gflop/sec, lose 60 Mflop/sec - \$0.96/MFlop
 - Infiniband investigation
 - Buy small switch (24 ports)
 - Start SciDAC software work (port *QMP*)
 - Switched GigE investigation
 - Buy 2 Myrinet GigE blades
 - Use low latency drivers from SciDAC collaborators
-
-

Fall Upgrade

- We are ~~cursed~~ blessed with much industry churn
 - G5 vs Itanium2 vs P4E vs AMD
 - New chipsets
 - Infiniband
 - PCI Express
 - Evaluate and pick best performance/price
 - Best guess:
 - 250 single processor nodes (\$800) or 125 dual processor nodes (\$1500)
 - Infiniband fabric
 - PCI Express NICs or embedded I.B. Interfaces
 - Will easily meet or drop below \$2/MFlop on most demanding codes
-
-

Price/Performance Trends



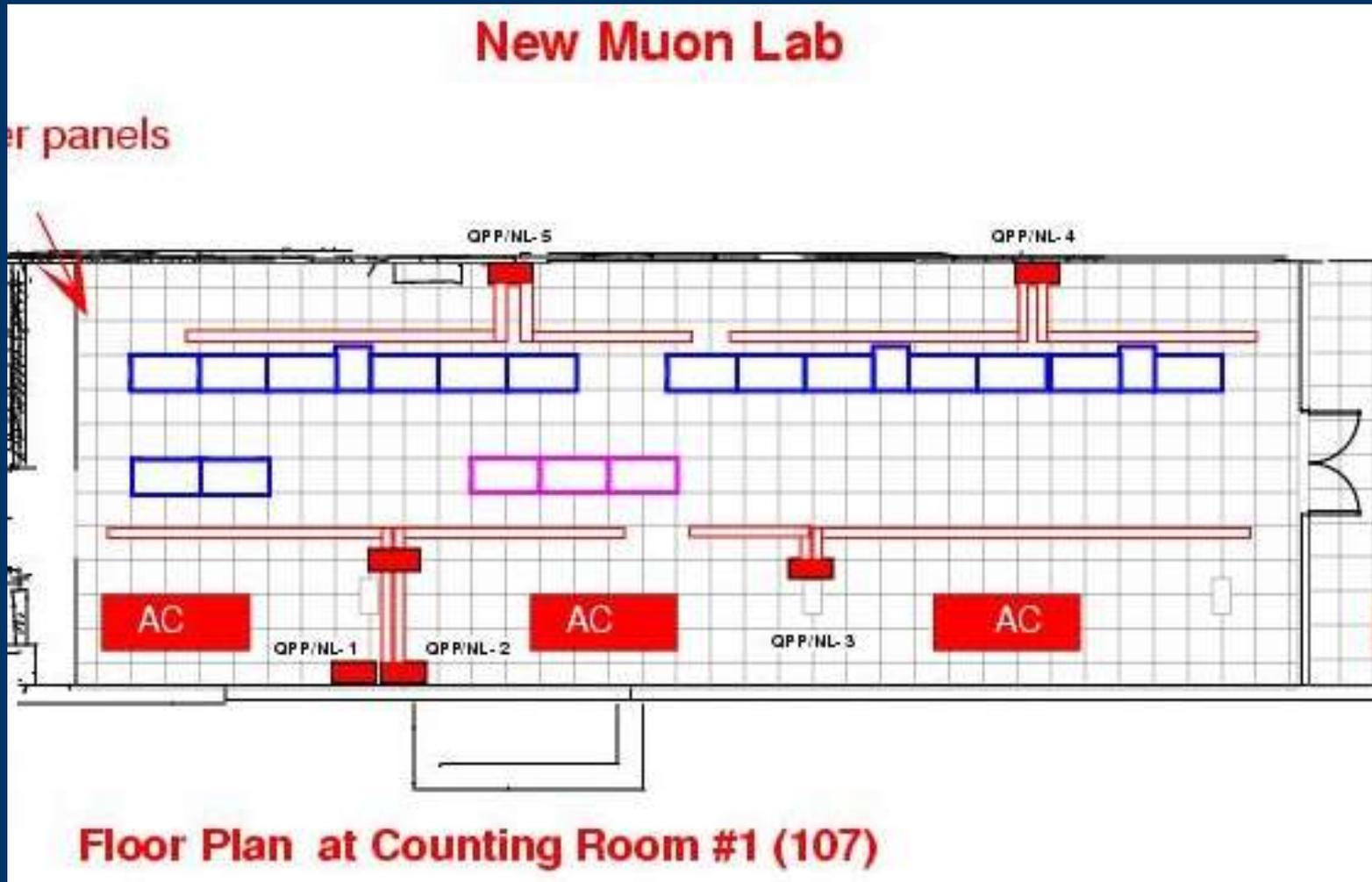
Why Clusters and QCDOC this year?

- The analysis codes running now, which have not been implemented in QCDOC assembler, will run faster and more cost effectively on clusters.
 - There exists a large backlog of MILC-generated lattices which need to be analyzed; clusters are very effective for this.
 - Analysis jobs, as opposed to unquenched configuration generation, run best on small numbers of nodes. It is entirely unknown how well QCDOC will perform on jobs like this.
 - Neither QCDOC nor clusters have proven the ability to sustain 1 TFlop on a single job; both approaches have merit.
-
-

New Muon Facility Issues

- Security
 - Building, computer room are unlocked during the day, with lots of foot traffic
 - Per Jed Brown, computer room is being re-keyed
 - Cooling
 - With chiller repair done (Summer 2003), we now have backup cooling
 - 3rd Liebert is partially installed
 - Space
 - Room for CDF, LQCD Winter expansion, BTeV test stand
 - Fall expansion will require CDF to move
-
-

New Muon Layout



SciDAC Software Work

- Processor optimizations
- User environment
- The SciDAC software stack
 - QIO
- Data handling
 - Metadata
 - Data movement (SRM et al)
 - Middleware
 - Interactions with ILDG



Processor Optimizations

- Microbenchmarks
 - SU3 kernels
 - Memory bandwidth
 - Optimizations for specific processors:
 - X86:
 - SSE2, SSE3
 - Prefetching
 - G5:
 - Altivec, assembly routines
 - Itanium2:
 - Intel optimization course (Singh, Simone, Neilsen, Holmgren, November 2003)
 - Work needed this Spring
-
-

Microbenchmarks – SU3 Kernels

- QCDSTREAM

- <http://lqcd/qcdstream/>
- Measures memory bandwidth, performance of SU3 matrix-vector and matrix-matrix operations
- Revealed bug in Intel C compiler, fixed in Release 8

QCDSTREAM Benchmark

Introduction

QCDSTREAM was inspired by John McCalpin's [STREAM](#) sustainable memory bandwidth benchmark. Lattice QCD physics codes are typically memory bandwidth or floating point limited. Data access patterns and the interactions of structure layout with cache line lengths and memory controller dynamics can significantly impact performance of these codes. QCDSTREAM is designed to investigate these effects, as well as to measure the performance boost obtained with SSE versions of complex matrix-vector and matrix-matrix operations (see also these [SSE discussions](#)).

Details

Memory Bandwidth

In the STREAM "Copy" measurement, the rate of movement of double precision values between arrays is measured. QCDSTREAM adds measurements of the rate of movement using floats and long doubles. The latter measurement uses inline SSE instructions to take advantage of the 128-bit wide SSE registers and cache-bypass write operations. Because of SSE usage, this section of QCDSTREAM will only run on Pentium III, Pentium 4, and Athlon processors which support SSE.

Here is a sample extracted from a 1.7 GHz Pentium 4 Xeon (Foster) run:

Function	Rate (MB/s)	Mean time	Min
Float Copy:	1234.0 +/- 0.9	0.0260 +/- 0.00002	0.02
Double Copy:	1253.9 +/- 1.0	0.0256 +/- 0.00002	0.02
SSE Copy:	2121.2 +/- 4.9	0.0151 +/- 0.00004	0.01

Unlike STREAM, which computes rates using the minimum time values, QCDSTREAM uses the mean time and estimates the uncertainty using the standard deviation.

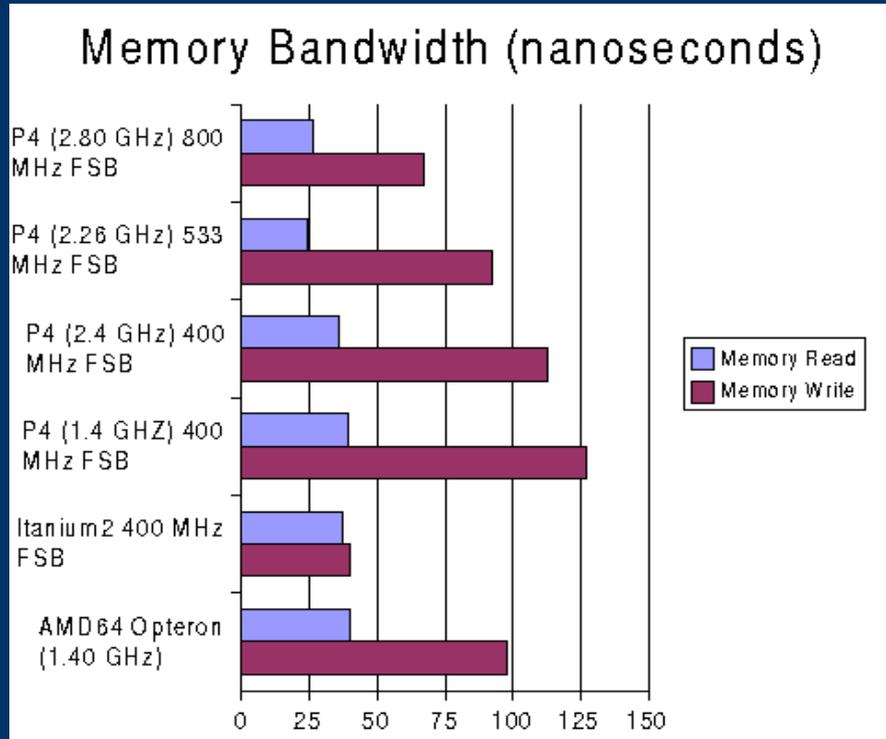
Matrix-Vector

In the "MatVec" section, QCDSTREAM calculates sustained MFlop/sec during matrix-vector multiplies. The matrices are 3x3 complex, and the vectors are 3x1 complex. There are two stanzas, labelled [MILC MatVec](#) and [SSE MatVec](#). The MILC version uses C-language code, and the SSE version uses inline GCC assembler.

Here is a sample extracted from the same Xeon run:

Function	Rate (MFlop/s)	Mean time	Min
MILC MatVec In Cache:	780.4 +/- 0.2	0.0188 +/- 0.00001	0.01

Microbenchmarks – Memory Performance



- Memory access in LQCD code:
 - Many reads
 - Few writes
 - Benchmarks like STREAMS measure read/write pairs
 - For predicting performance, we need separate read and write performance measurements

Optimizations – SSE

File Edit View Go Communicator Help

Back Forward Reload Home Search Netscape Print Security Shop Stop

Bookmarks Location: <http://lqcd.fnal.gov/sse/inline.html> What's Related

Internet Google Dejanews Lookup New&Cool

Inline SSE MILC Math Routines

- [NASM to Inline GCC Translator](#), [nasm2c.pl](#)
- [Performance](#) of the inline routines
- [Obtaining and using](#) the inline routines
- [Optimizing MILC Math Routines with SSE](#)
- [Catalog of MILC Math Routines](#)

Code written in NASM assembler tends to be easy to read, and the programmer can readily add documentation. Object files generated using nasm may be linked with C object modules generated by many compilers, including gcc, pcc (Portland Compiler Group), and icc (Intel C++ compiler). However, all routines must be implemented as subroutines, with a corresponding overhead penalty.

NASM to Inline GCC Translator

We've implemented a nasm-to-inline-gcc translator, [nasm2c.pl](#), which generates gcc assembler macros from the the NASM source codes. For this translator to succeed, the following conventions must be used in NASM source codes:

- All `push`, `mov`, `pop`, `add`, and `ret` operations must be associated solely with the stack handling operations used to reference arguments. Lines with these codes are deleted from the inline macros.
- Any references to memory (*i.e.*, using `[. . .]` NASM constructs) *must* include in the comment field on the same line a construct, offset by `< . . . >`, which gives a C-language reference to the address, referenced by the macro argument. For example, if the inline macro definition is

```
#define _inline_sse_mult_su3_nn(aa,bb,cc)
```

where `(aa, bb, cc)` are of type `(su3_matrix *)`, a possible source line containing a reference construct would be:

```
movss xmm3,[eax] ; <(aa)->e[0][0].real>
```

Here, in the NASM version `[eax]` would dereference the first argument, passed on the stack, to `mult_su3_nn(a,b,c)`. In the translated inline version, where no stack is used to transfer arguments, the code directly references `a->e[0][0].real`.

- In the current version, all labels are ignored, and branches will not work.
- By default, three macro arguments `aa, bb, cc` are assumed. Override this in the invocation of `nasm2c.pl`, *eg*:

```
./nasm2c.pl sse_routine.nas "aa,bb0,bb1,bb2,bb3,cc"
```

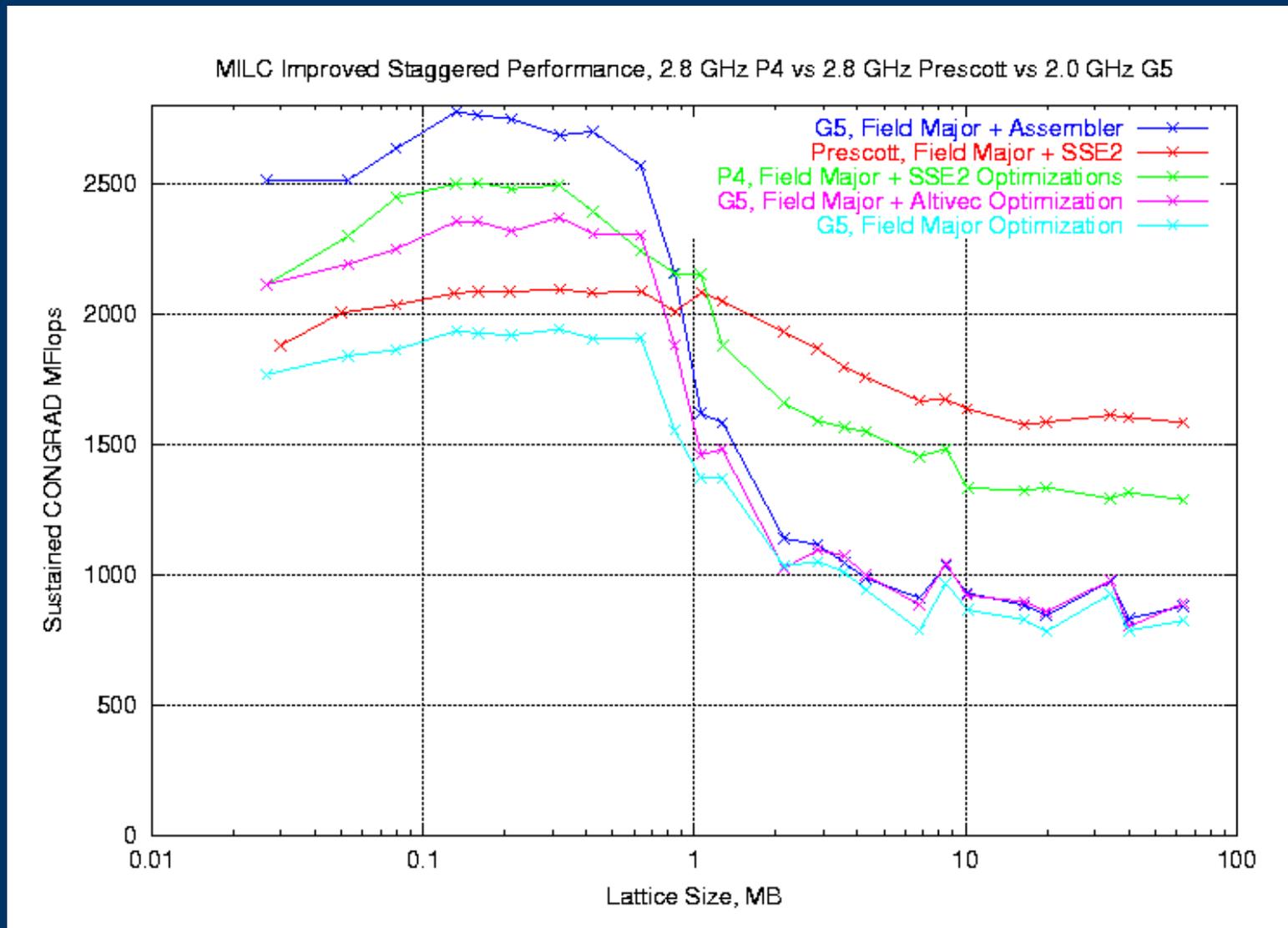
Note that there should be corresponding `< . . . >` reference constructs in your NASM code.

As an example, here is the NASM version of `add_su3_vector`:

```
;; sse_add_su3_vector( su3_vector *a, su3_vector *b, su3_vector *c)
```

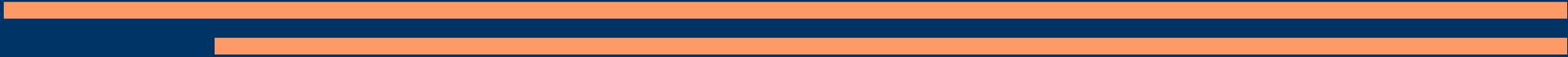
- <http://lqcd/sse/inline.html>
- Optimizations for x86, AMD using SIMD facilities
 - Similar work done for PPC970/G5 using Altivec
 - Also, assembly coding for PPC970/G5

P4, P4E, G5 Optimizations



User Environment

- JLAB, BNL, FNAL are working to establish a common user environment
 - Batch usage
 - Filesystem layout
 - Accounting
 - File movement
- See <http://lqcd/runTimeEnv.html>
 - Surprisingly short
 - A work in progress – much awaits QCDOC progress



SciDAC software

- SciDAC software steering committee
 - Holds weekly phone conferences.
 - Members representing Columbia, Jlab, FNAL, MILC and MIT.
 - FNAL: D. Holmgren, J. Simone, *et al.*
 - Charges:
 - Common user environment ✓
 - Portable software libraries
 - Data management
-
-

SciDAC software libraries

- Portable, scalable framework for writing lattice QCD applications
 - API support in C and C++
 - Basis for new parallel codes
 - Leverage existing application codes (e.g. MILC)
 - Optimizations for clusters and QCDOC
 - Support parallel I/O
 - Promote standard data formats
-
-

SciDAC library API's

Applications can call any of three levels:

- Level 3:
 - optimized solvers for Dirac operator
- Level 2:
 - QDP (QCD Data Parallel) lattice-wide operations
 - QIO (I/O) binary data and metadata
- Level 1:
 - QLA (QCD Linear Algebra) site-by-site
 - QMP (Message Passing) communication operations more suited to lattice than MPI



Fermilab efforts

- Processor-specific optimizations (QLA) ✓
- QIO/metadata(XML) library design
 - APIs for C++ and C
 - lib using xerces(Apache) or libxml2(Gnome)
 - Data binding C structs \Leftrightarrow XML
- Data formats
 - binary data format
 - metadata XML markup
 - file packaging: wrapper

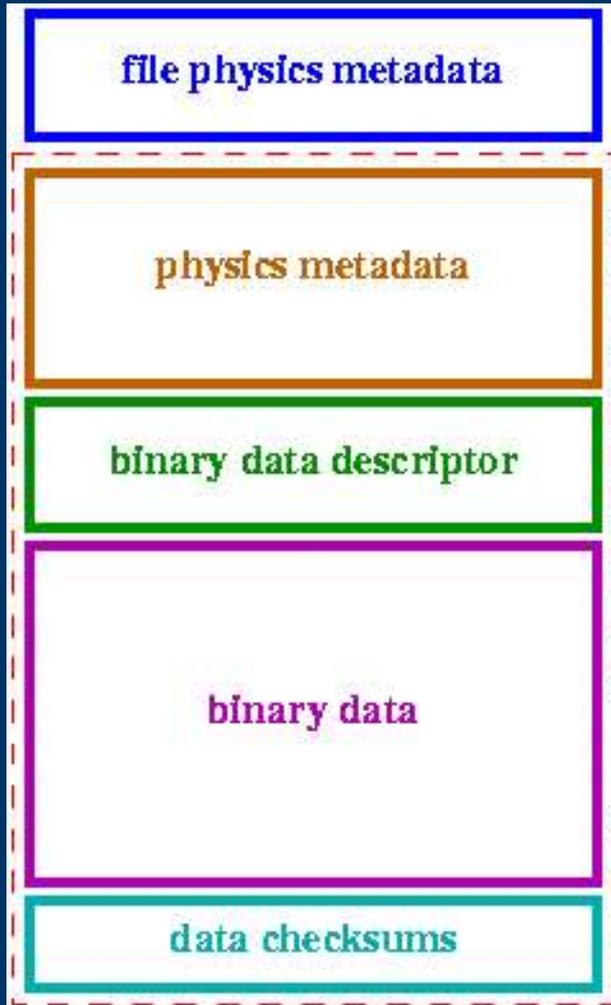
QIO/metadata XML

- API features
 - Export data structures as XML.
 - Bind XML document to data structures.
 - Support conversions to/from intrinsic C types and arrays of thereof.
 - Recursively handles composite types built out of simpler types.
 - Navigate metadata via subset of Xpath expressions
 - /asqtadGaugeAction/coupling/beta
 - Hides the XML parser implementation.
-
-

QIO/XML status

- Draft C API designed.
 - Partial/prototype implementation in C over libxml2.
 - An equivalent C++ library based on an earlier draft API implemented over libxml2 by B. Joo (UKQCD).
 - C implementation needs to be brought beyond the prototype stage and extended to the complete set of supported data types.
-
-

SciDAC draft file format



- XML markup of metadata
- Schema: standard XML
- Multiple records/file
- Markup for array layout (BinX?)
- Binary data: QLA types, site-major lexicographic order
- Wrapper: DIME, cpio or CERN wrapper?

SciDAC file format issues

- BinX?
 - proposed DataGrid standard for markup of binary array dimensionality, sizes, data type, endianness...
 - Complicated XML schema
 - Future: allow automatic translations.
 - Not necessary to characterize SciDAC data.
 - DIME?
 - Direct Internet Message Encapsulation
 - Proposal for IETF standardization withdrawn.
 - Alternatives: tar, cpio (enstore) or CERN wrapper (enstore, huge files)
-
-

Data management

- Local management issues
 - Manage data flowing between endpoints: cluster workers, large raid disk arrays, enstore tapes, external sources/sinks.
 - Flat filesystems highly desirable.
 - Auto archiving/replication of valuable data.
 - Auto recovery of space used by temp. files.
 - Imported/exported data issues
 - Large volumes (15TB@FNAL) scattered around many institutions.
 - Existence/location poorly advertised.
 - Metadata difficult to obtain.
-
-

Data management status

- Begin testing private dCache for Fermilab LQCD facility.
- Prototyped SQL database design for storage of gauge configuration metadata.
- Testing SRM interface to enstore.
- But, grid authentication mechanisms not always straightforward.



lattice data grid

- 
 - www.iqcd.org/ildg
 - www.nesc.ac.uk Edinburgh
 - Virtual workshops: 12/02, 5/03, 12/03
 - Chair: A. Ukawa '04, R. Kenway '03
 - U.S. Board member: R. Brower '04
 - Global membership: Australia, Europe (Italy, Germany, UK), Japan and U.S. (SciDAC)
-
-

ILDG goals

- Data grid
 - Link data repositories in member countries
 - Common webservices middleware
 - User interface via web browsers and APIs for scripting languages
 - Data movement/replication (SRM)
 - Share database technology
 - Discovery through metadata searches
 - Share valuable gauge configurations
 - Standardize binary data representation
 - Standardize metadata (XML Schema)
-
-

ILDG working groups

- Grid Architecture and middleware
 - FNAL: E. Neilsen, D. Holmgren, J. Simone
 - Grid **middleware** for data management
- Metadata
 - FNAL: J. Simone
 - Focus on XML schema design for lattice QCD gauge configurations
 - Standard binary data format for configurations
 - Extend (meta)data standards to other data?



ILDG Implementation

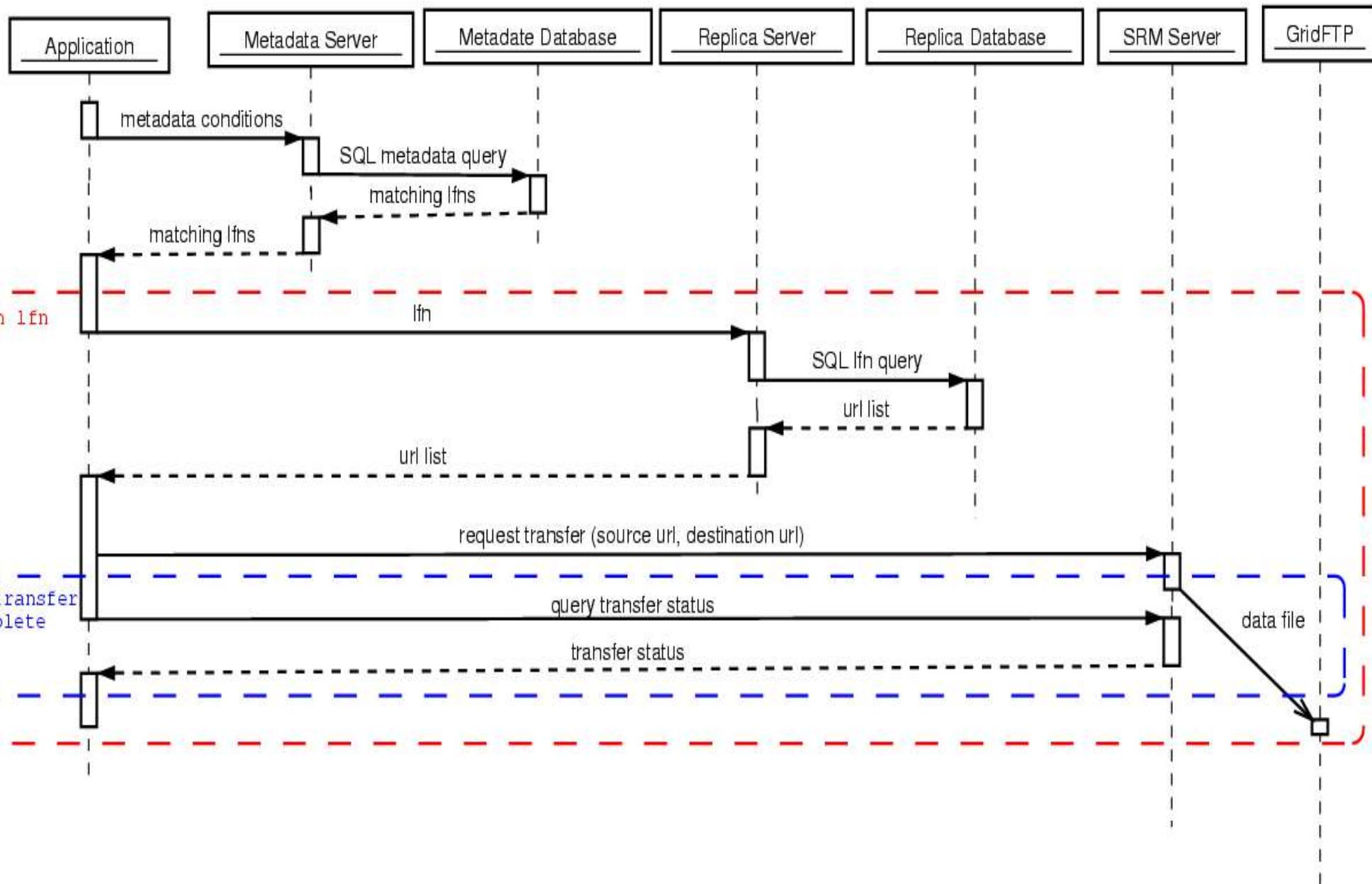
- Security will be based on Globus GSI tools.
- Web services (SOAP, WSDL) interfaces are being established by ILDG collaborators.
- SRM will provide the interface to mass storage systems.
- It will be left up to the collaborating institutions to implement the agreed upon interfaces.



Anticipated architecture

- Leverage grid data and replica management architecture of other projects.
 - Chervenak *et al.* 2002
 - Allcock *et al.* 2002
- Application uses four services
 - Webservice registry
 - Metadata catalog
 - Replica catalog
 - Storage Resource Manager (SRM)

Interaction for Data Transfer



ILDG current status

- Several institutions have produced prototypes of expected elements, but nothing interoperable.
 - Attempted “live demos” at LATTICE'03.
 - Fermilab (Eric) is driving the documentation of requirements and use cases.
 - Significant progress, but more feedback welcome from rest of the working group.
-
-

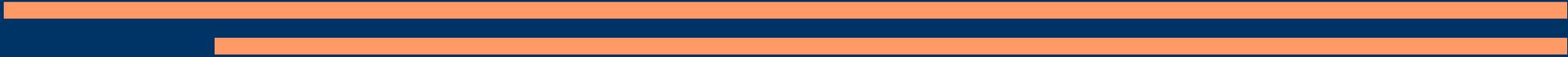
ILDG metadata

- QCDML an XML Schema for metadata markup of gauge configurations.
- Flexible design to structure metadata corresponding to diverse LQCD actions in use.
- Extensible to encompass future actions.
- XML structured to facilitate browsing and searching for gauge configurations by metadata.



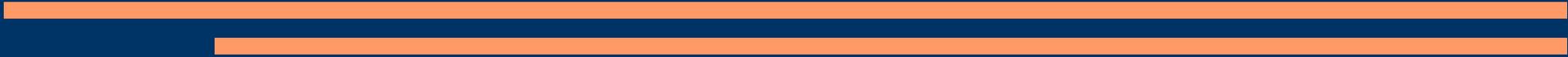
Top level QCDML objects

- management
 - collaboration, references, history, file version, ...
- implementation
 - machine name, hardware type, software versions, ...
- markovChain
 - action and physics couplings, algorithm, precision, ...
- MarkovStep
 - Individual configuration series name, sweeps, global file name, ...



Metadata wg status

- Draft v4 of QCDML markup for gauge configurations.
- Latest version iterates on comments from SciDAC and others.
- Main changes:
 - XML to include standard references for action and algorithms.
 - Structural changes to simplify searches based on parameters of action.
- Revisit file format standardization
 - BinX for array markup?
 - DIME wrapping?



Plans

- Local storage
- Hardware acquisitions
- Software



Plans – Local Storage

- Local data requirements:
 - O(10 Tbytes)
 - Flat directory space
 - Load balanced and throttled
 - Accessible from all workers
 - High rate connectivity to Enstore
- Proposed solution (work in progress):
 - “Resiliant” dCache for non-migrated data
 - Multiple spindles, buses, dCache servers
 - Hope to avoid thrashing at startup of large job sets
 - Local (to New Muon) dCache for data which migrates
 - GigE connectivity to FCC
 - Currently on our head node
 - Will move to dedicated node

Plans - Acquisitions

- Winter expansion
 - 80 single P4E's
 - Use existing Myrinet
 - Assess whether P-III Fast Ethernet cluster is of use
- Fall expansion
 - O(250) single P4E's
 - Infiniband



Plans - Software

- Communications software (QMP) for Infiniband
 - Additional microbenchmarks
 - Processor optimizations
 - Itanium2
 - Opteron
 - Assist with Level 3 inverter for MILC on x86
 - Middleware
 - Archive wrapper for > 8GB files
 - Metadata database investigations
 - Parallel I/O
 - PVFS?
 - Lustre?
-
-