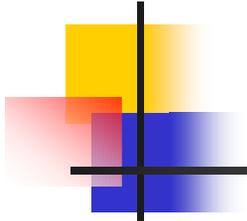


LQCD Project Status

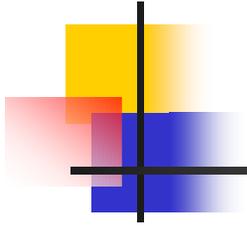
November 28, 2006

Don Holmgren/Jim Simone/Jim Kowalkowski

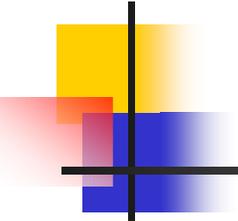


Outline

- Fermilab Lattice QCD Hardware Status
- National Lattice QCD Project
- Physics Production
- SciDAC-2 Lattice QCD Subprojects

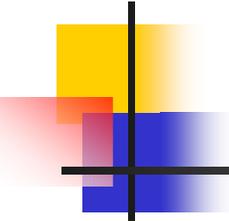


Hardware Status



"Kaon"

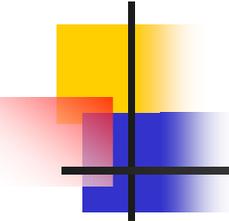
- First cluster at Fermilab purchased as part of the 4-year national project
- Awarded to Koi Computers, June 19, In Production Oct 3
- 600 dual processor Opteron 270 (dual core) nodes based on ASUS motherboards
- DDR Infiniband fabric (20 Gbps)
- #95 on the November 2006 Top500 list
 - Sustained 6.894 Tflop/sec, using 2400 processes
 - Very important to Mellanox
- Configuration generation started
 - Sustains 0.5 Tflop/sec on 256 nodes (1024 processes)
 - Produces a time step every 3.5 hours
 - Will run for next 9 months



Kaon – Issues

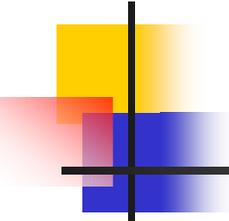
■ Stability

- During acceptance tests, ~ two dozen nodes crashed frequently under load with a kernel panic which indicated memory problems
- After memory replacement, these nodes functioned correctly
- In production in October and November, additional problems:
 - Nodes that crash – Koi replaces memory
 - Nodes that reset – isolated and investigated
 - ~ 75 nodes are misbehaving (eg, resets every few days)
 - Met with ASUS at SC'06 – they will work with Koi to fix



Kaon – cont'd

- Just discovered yesterday – “Machine Check Exceptions” in kernel logs
 - **MCELOG** interrupts errors – in a quick look, have seen:
 - Memory ECC corrections and non-corrections
 - CPU data cache and instruction cache (L1) corrections and non-corrections
 - CPU cache (L2) corrections and non-corrections
 - Will work to correlate the MCE's with reset frequency
- IPMI (2.0) over LAN – works flawlessly
 - Had to work with ASUS to get BIOS upgrades
 - Use *ipmitool*
- Serial over LAN – some problems
 - Spurious interrupts on COM1 on some nodes
 - High interrupt load slows those nodes



MCELOG

```
Nov  2 11:45:49 kaon0408 kernel: EXT3-fs: mounted filesystem with ordered data mode.  
Nov  2 11:45:49 kaon0408 kernel: Adding 4096532k swap on /dev/sda5.  Priority:-2 extents:1  
Nov  2 11:49:45 kaon0408 kernel: Machine check events logged  
Nov  2 11:54:45 kaon0408 kernel: Machine check events logged
```

```
/usr/sbin/mcelog
```

```
MCE 0
```

```
HARDWARE ERROR. This is *NOT* a software problem!
```

```
Please contact your hardware vendor
```

```
CPU 2 4 northbridge TSC c51e38b5b4fb1
```

```
ADDR dfef5f28
```

```
Northbridge ECC error
```

```
ECC syndrome = 57
```

```
bit46 = corrected ecc error
```

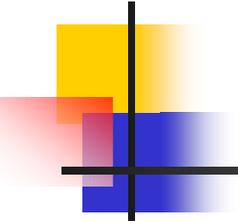
```
bit62 = error overflow (multiple errors)
```

```
bus error 'local node response, request didn't time out
```

```
generic read mem transaction
```

```
memory access, level generic'
```

```
STATUS d42bc00000000a13 MCGSTATUS 0
```



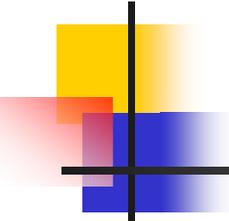
Pion/QCD

■ Pion

- 520 Pentium 640 (3.2 GHz) nodes
- SDR (10 Gbps) Infiniband
- 260 nodes purchased Spring 2005 from Koi
- 260 nodes purchased Fall 2005 from ACE

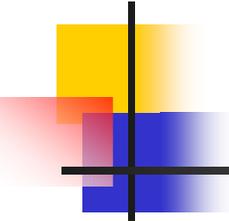
■ QCD

- 128 Pentium 4E (2.8 GHz) nodes
- Myrinet 2000 (2.5 Gbps)
- Nodes purchased Spring 2004 from Koi



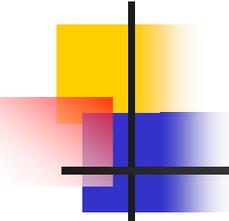
Pion Disk Errors

- Pion uses 40 GB SATA drives from Western Digital (WD400JD)
- Single bit errors were identified by Jim Simone in his analysis of physics output files produced during the previous months
 - By instrumenting physics code, we eliminated the Infiniband fabric as the source of the errors
 - Further analysis definitively showed that approximately 25 disks drives had single bit errors on reads at low rates ($< 1:10^{12}$)



Pion Disk Errors

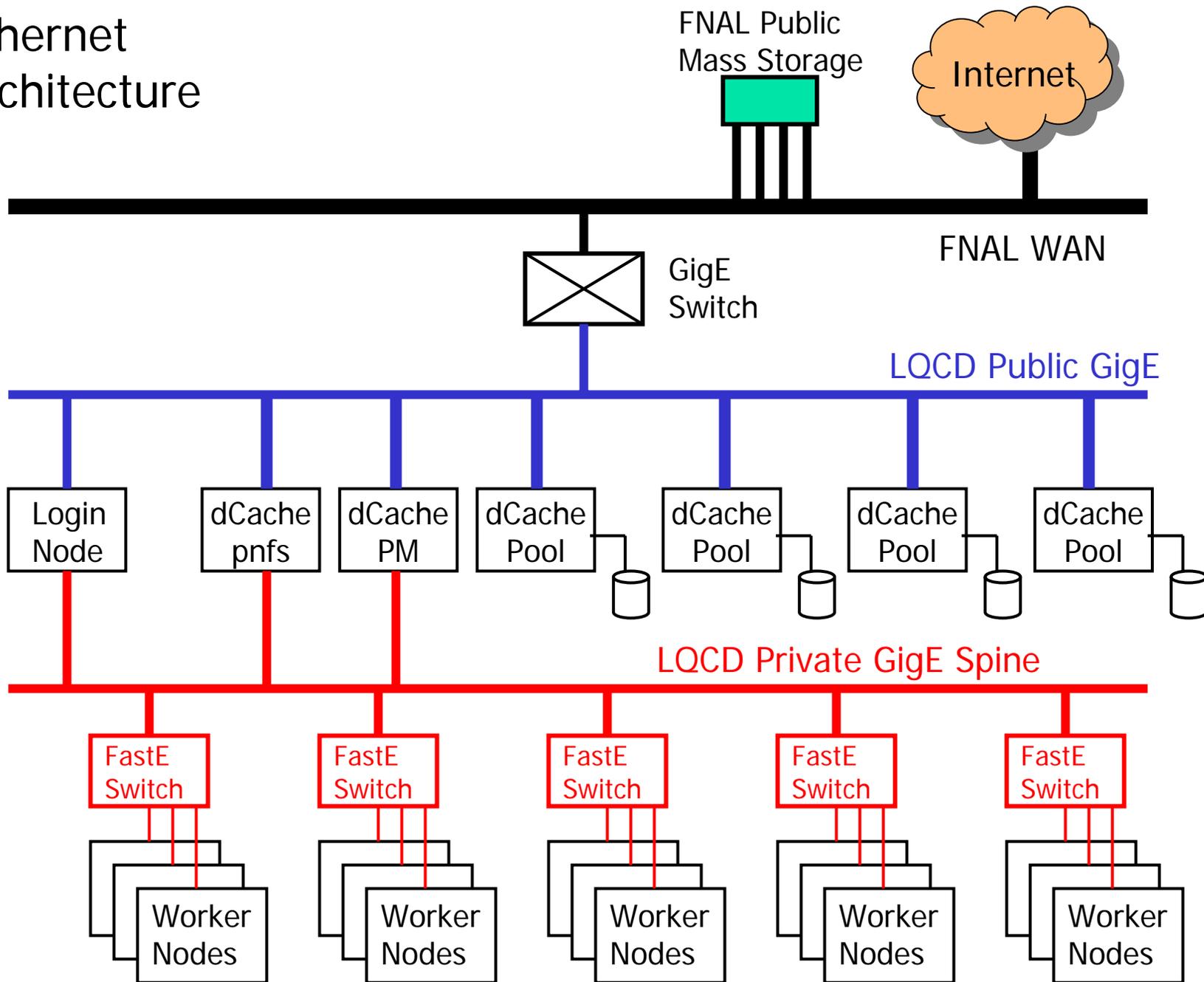
- 160 of the 260 drives included in the Spring 2005 half of Pion (purchased from KOI) were identified by Western Digital as having a potentially faulty DRAM chip used for the drives' read cache. None of the 260 drives of the same model included in the Fall 2005 half of Pion (purchased from ACE) had this problem.
- Western Digital supplied replacement drives
 - Koi worked with LQCD staff to do replacements in October
 - Further replacements will occur this month (27 disks with bad chips were mistakenly used as replacements)
- Physics codes all now verify CRC's of data files
- We regularly scan all disks on all clusters for similar errors



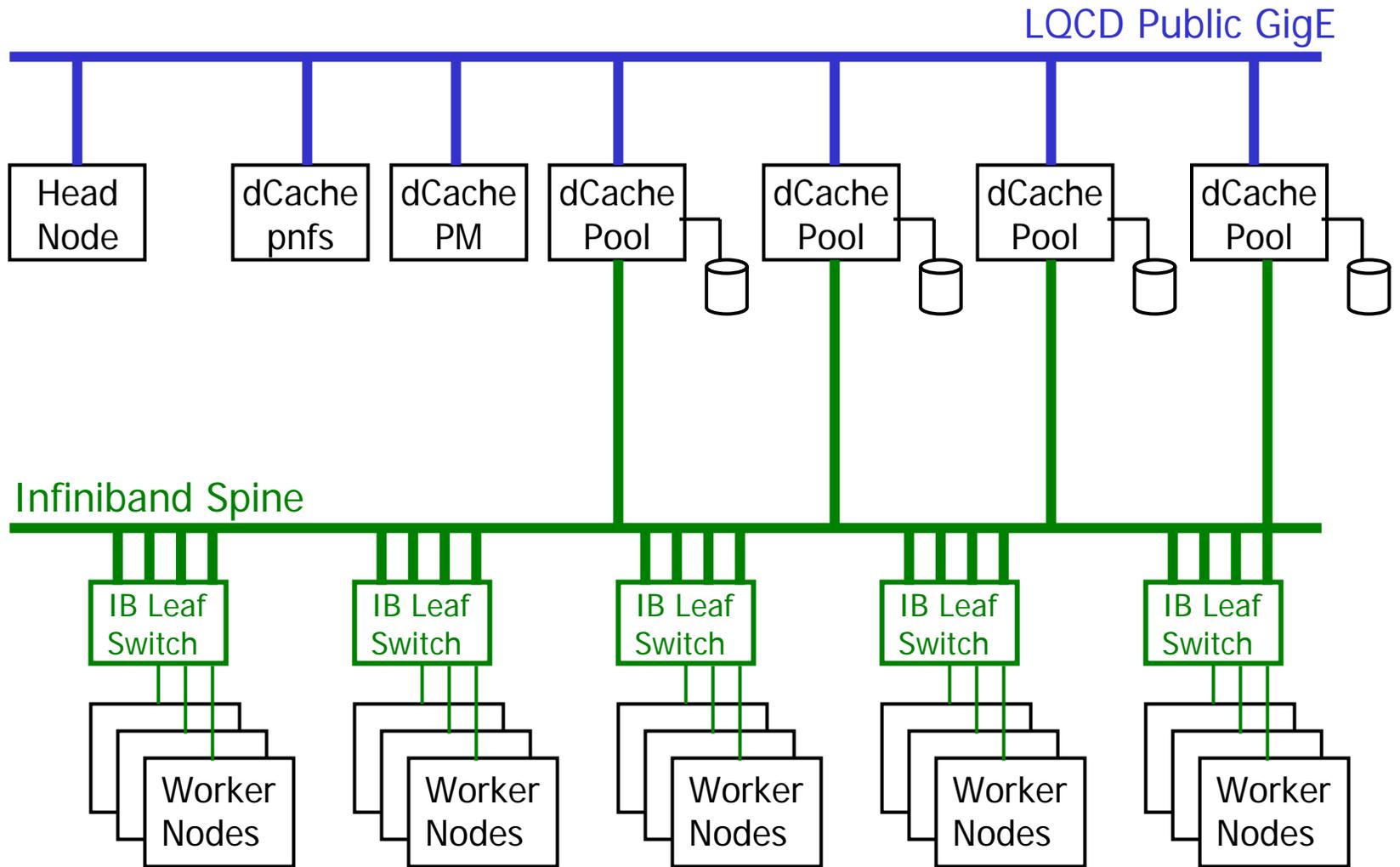
dCache

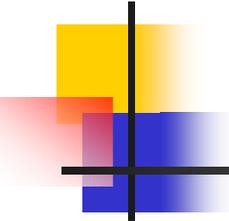
- Current configuration:
 - STKEN (public) dCache (/pnfs/lqcd) for access to Enstore
 - LQCD (private volatile) (/pnfs/volatile) dCache for high performance cluster-wide file system
 - 4 movers with 19 TBytes of attached disk
 - Movers are on both the Pion SDR Infiniband and Kaon DDR Infiniband fabrics (via dual port cards), and also send packets to the QCD Myrinet fabric via a gigE-to-Myrinet router
- Issues for users:
 - > 2 GB files appear differently in STKEN and volatile dCache /pnfs
 - All access has been via dccp; C and C++ access via libdcap is planned but we are still working on testing 64-bit port
 - Brittle w/respect to failed disks

Ethernet Architecture



Infiniband Architecture





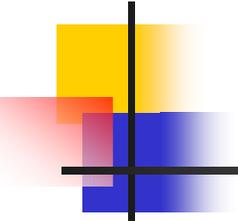
dCache cont'd

■ Immediate plans:

- Bring two additional movers online on volatile dCache (ready), each attached to a 8.4 TB disk array (in hand)
- Upgrade to newer version (1.7)
- Migrate data on older pool disks to new disk arrays, then reconfigure older disk arrays to RAID-5 (now at RAID-0)
- Encourage usage of libdcap (direct I/O from physics jobs) – critical as propagators increase to 8+ GB in size

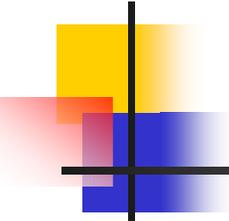
■ Longer term plans:

- Move tape-backed LQCD dCache from STKEN to LQCD-managed dCache
 - Currently LQCD users occasionally impact public STKEN users
 - Use 2 new movers, with 17 TB of storage (hardware in-hand)



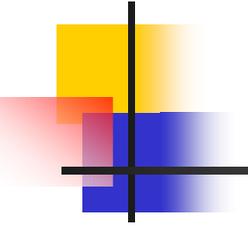
New Muon Computer Room

- Kaon is in the south half (Room 107)
 - Air conditioning is independent of rest of building (relies on external air cooled condensers)
 - Not counting AMR cluster, using 32 of 50 rack positions (60%), 25% of power/cooling (1 KVA)
 - So, any additional racks can be higher density
- QCD and Pion are in the north half (Room 108)
 - Rely on chillers which also serve the SMTF facility
 - During unscheduled early morning chiller outage in October we observed that Room 108 heated slowly and survived the 2-hour outage during production (temperatures did rise)
 - During upcoming outage cooling outage (so PPD can remove some ICW pipes) will use fans to improve circulation between 107 and 108



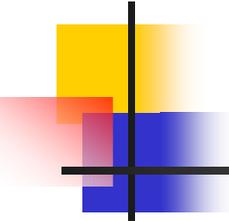
NML, cont'd

- Future needs:
 - National LQCD project plans call for new hardware at FNAL in FY08 and FY09
 - If possible, FY08 and FY09 purchases will be combined in FY08
 - By moving to higher density racks, NML would have sufficient space to house the FY08/FY09 cluster(s)
 - Infiniband and alternatives require proximity of racks, though remoting Infiniband to a different building will likely be an available (costly) option
 - Might have to move the AMR cluster, or decommission the existing AMR cluster and reuse QCD for AMR modeling



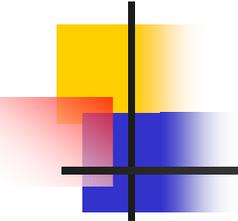
National LQCD Computing Project

- For this project, FNAL:
 - Operates QCD, Pion, Kaon, FY08/FY09 cluster(s)
 - Provides Project Manager and Associate Project Manager
 - Interacts extensively with JLab and BNL
 - JLab and FNAL operate similar Infiniband clusters
 - Reviewers urged FNAL and JLab to interact more closely on FY07 and future acquisitions – FNAL is helping to evaluate hardware for current JLab acquisition (MILC and other benchmarking on external Intel clusters)
 - Cooperate to ensure that users can transfer data files (FNAL: Kerberos, JLab/BNL: SSH firewalls), and that they see similar environments



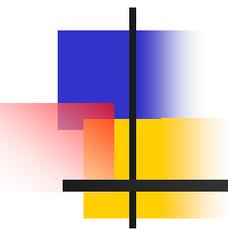
OMB-300 Issues

- National project is a “Major IT” project and so must submit and maintain OMB Exhibit 300 document for each future Budget Year
- Since September, we have had to respond to multiple data calls:
 - Operations Analysis – had to provide our Alternatives Analysis document (business case for why we acquire LQCD hardware) along with financial benefits, to the OCIO for their report to OMB
 - Risk Management – had to provide our Risk Management document to the OMB
 - Call from Senator Coburn – had to provide information to OCIO about any re-baselining of LQCD for report to Senator Coburn
- FY08 Exhibit 300 “passback” expected this week
 - Will likely require work on Privacy, Risk Management

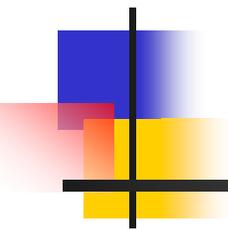


National Project Deliverables

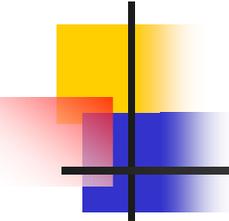
- Major deliverables each year are:
 - Capability of new hardware (Tflop/s deployed)
 - Availability of operated hardware (Tflop/s-yrs delivered)
- FY06
 - Promise: 1.8 Tflop/s (FNAL) and 6.2 Tflop/s-yrs
 - Result: 2.3 Tflop/s (FNAL) and 6.266 Tflop/s-yrs
- FY07: 2.9 Tflop/s (JLab) and 9.0 Tflop/s-yrs
- FY08: 4.2 Tflop/s (FNAL) and 12 Tflop/s-yrs
- FY09: 3.0 Tflop/s (FNAL) and 15 Tflop/s-yrs



Physics Production

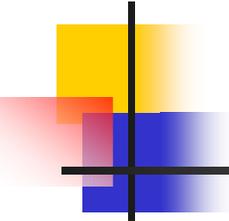


SciDAC-2 Subprojects



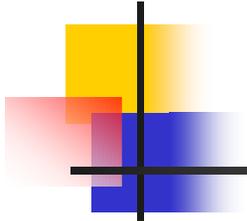
SciDAC-2: 64-bit Optimizations

- FNAL clusters:
 - QCD = 32-bit only
 - Pion = currently 32-bit, but 64-bit capable
 - Kaon = 64-bit, required for Opteron NUMA control
- C/C++ pieces of SciDAC libraries can be trivially built in 64-bit mode (validated)
- Assembler routines (SSE) can be optimized:
 - Different calling convention (registers, not stack) allows us to avoid gcc inline assembler macros
 - Code should be rewritten to use extra registers
- Schedule: finish by June 2007



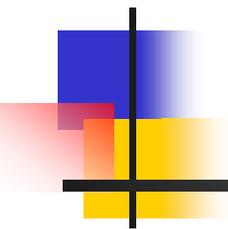
SciDAC-2: Opteron Optimizations

- SSE codes are in QLA, QDP libraries
 - Tuned using Intel processors; give significant speed-up
 - On “Kaon” (Opteron cluster), SSE routines are slightly slower than the “C” routines they replace
- Opteron optimizations are relevant to the Cray XT-3 machines at Pittsburgh and ORNL
- Planned work:
 - Understand Intel vs AMD SSE implementation differences
 - Follow AMD recommendations and tune SSE routines
 - Use AMD’s processor simulator to find and remove stalls
 - Investigate prefetching
- Schedule: finish by March 2007 (slip)

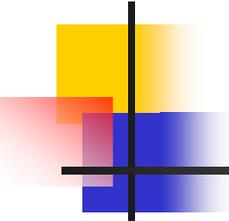


SciDAC-2: Multicore Optimizations

- All new commodity processors are dual or quad core
- On Kaon, we currently use a separate MPI process per core, each with a local lattice, and pass messages
 - Even via shared memory, this requires a memory copy
- Three approaches:
 - Use a separate MPI process per core, but use a single SYS V shared memory area for lattice
 - Use a single MPI process per node and OpenMP to thread loops
 - If OpenMP is not efficient, implement custom threaded library
- Schedule: at least 6 months to evaluate approaches

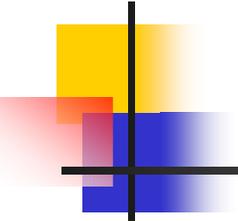


Extra Slides



Personnel

- Current:
 - Don Holmgren (95%) – including Project Mgmt
 - (Bakul Banerjee) (50%) – Project Mgmt
 - Amitoj Singh (100%)
 - Kurt Ruthmansdorfer (100%)
 - Nirmal Seenu (100%)
 - (Jim Simone) (50%)
 - (Jim Kowalkowski) (50%)
 - Bob Forster assisted with NML last year (100% after reorg?)
 - Also, this year Xian-He Sun is a guest scientist from IIT (workflow, reliability)
- Reorganization increases count by +1 (Rick Van Conant)



External Funding Sources

- Lattice QCD Computing (“OMB-300” Project)
 - Hardware: \$1548K in FY06, ~ \$50K in FY07 (storage), ~ \$2100K combined in FY08/FY09
 - 2.5 FTE (including 0.5 FTE for project management)
- SciDAC-2
 - ~ 3.0 FTE FY07
 - Slowly decreasing during subsequent 3.5 years