Cluster Deployment at Fermilab

Don Holmgren

http://lqcd.fnal.gov/talks/

All Hands Meeting – Feb. 21, 2003

Outline

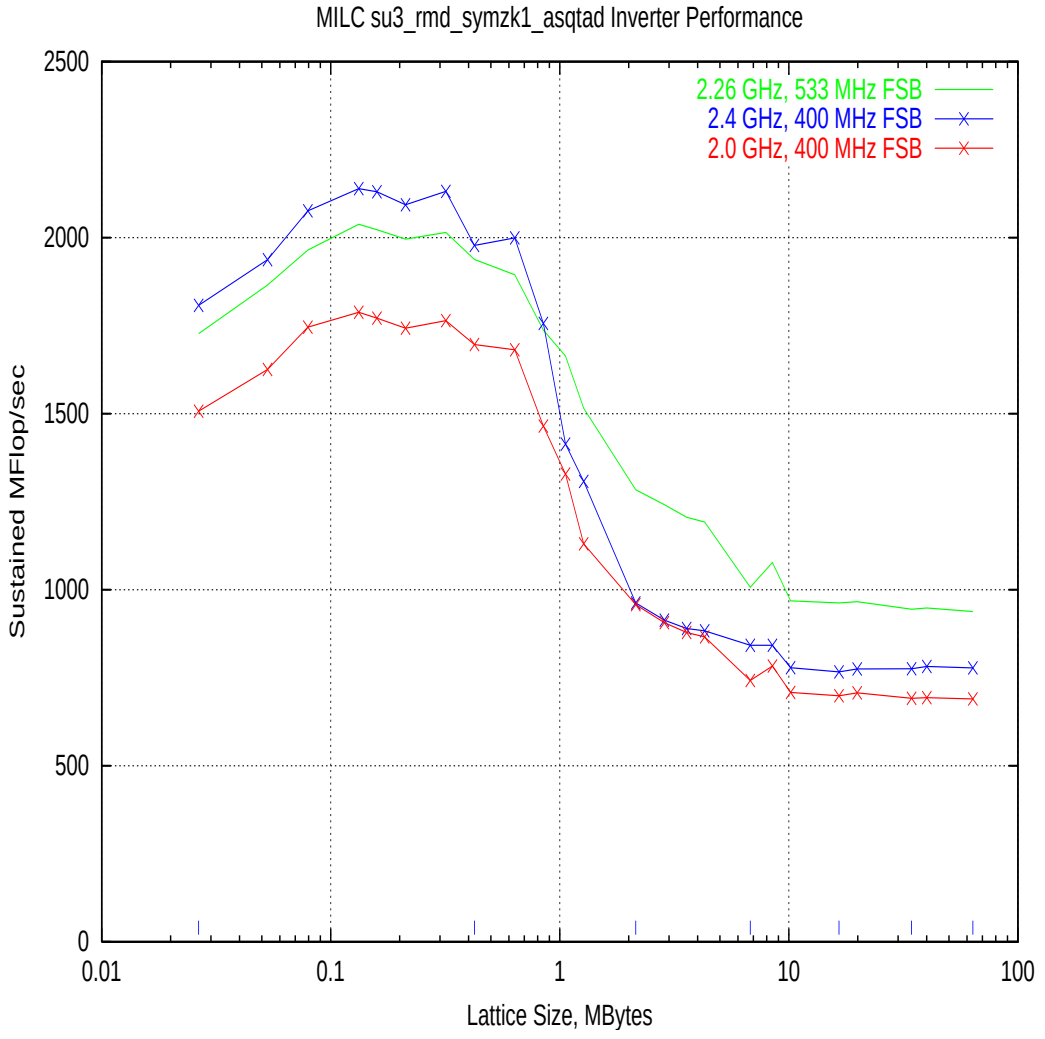
- Facility details
- Using the FNAL clusters
- Other SciDAC work at FNAL
- Future plans

Cluster Details

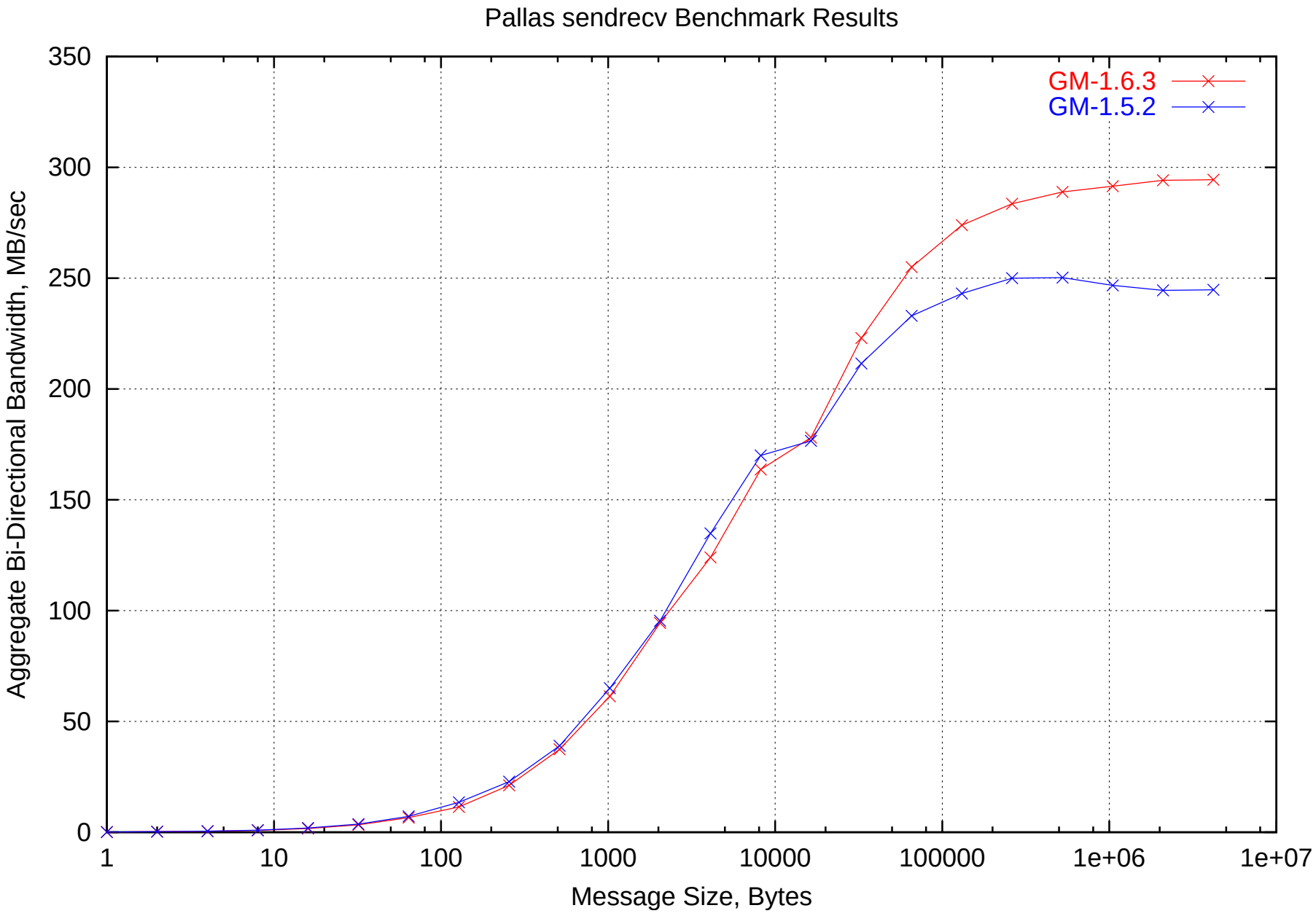
- Available hardware:
  - 48 dual 2.0 GHz Xeon systems (Steel Cloud, Reston, VA)
    - 1 GB DDR memory (E7500 chipset - 400 MHz FSB)
    - 17 GB scratch disk on each node
    - all machines with Myrinet 2000 (2MB, 133 MHz LANai-9)
    - 16 of the machines are also on a 2-D GigE mesh, 4 NIC's per node
      - mesh can be reconfigured via patch panel
  - 128 dual 2.4 GHz Xeon systems (CSI Inc, Alpharetta, GA)
    - same memory and disk configuration
    - all machines with Myrinet 2000 (112 with 133 MHz LANai-9, 16 with 200 MHz)
  - Head node (lqcd.fnal.gov)
    - 4 x 1.5 GHz Xeon processors
    - 4 GB memory
    - 0.9 TByte RAID-5 disk array (will purchase 2 x 1.8 TByte arrays)
    - 3 x 300 MByte network-attached storage units

Single Node Performance

- Improved Staggered Performance
  - Each site is 1656 bytes large
  - Blue ticks mark $(2,4,6,8,10,12,14)^4$
  - L2 cache (512K) near 4^4
  - FPU dominates for lattices smaller than 4^4
  - Memory bandwidth dominates for lattices larger than 4^4
  - 533 MHz FSB systems just now available with PCI-X



Performance - Networking - Pallas sendrecv



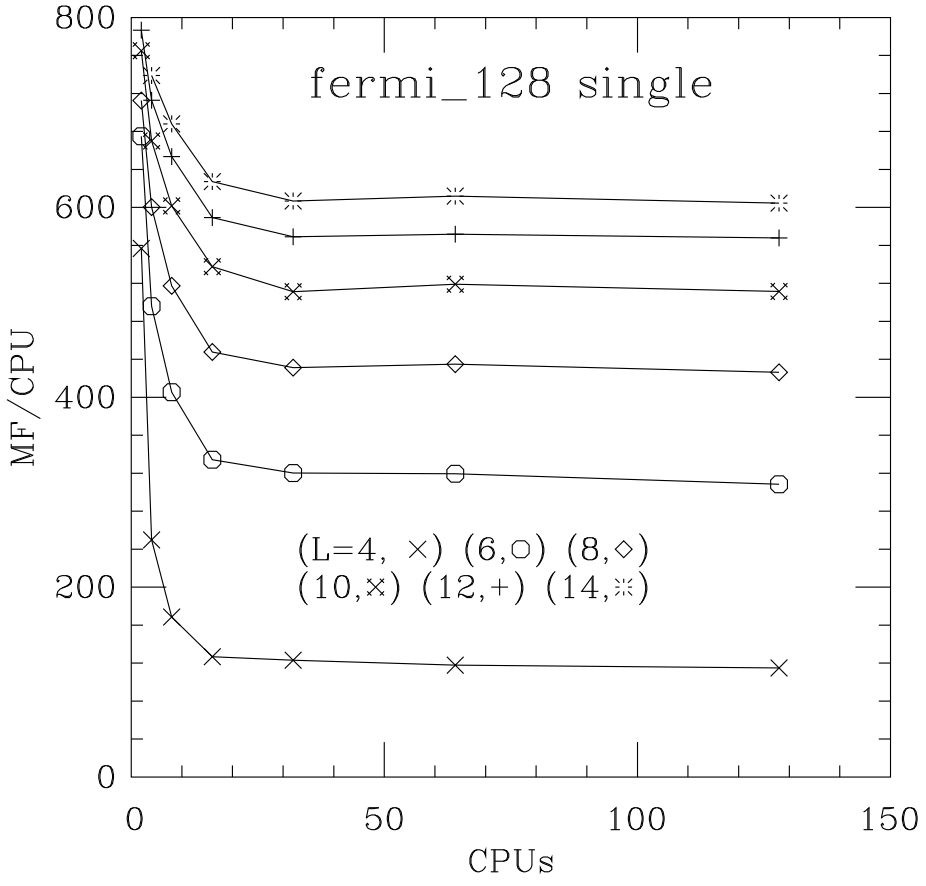
Performance - Multinode

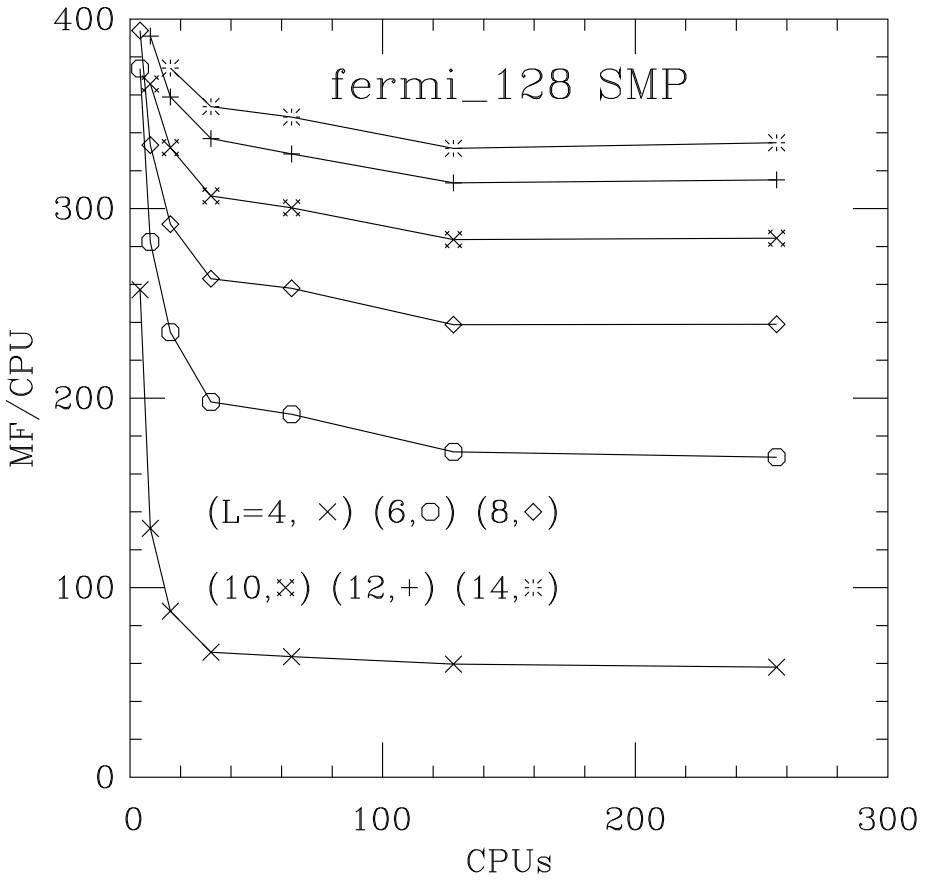
- HPL (top500.org): R_{max} = 550 GFlop/sec for 128-node cluster

  - #97 ranking on latest list

  - 10% slower than 2.4GHz dual Xeon RDRAM clusters at Utah, Cornell

- MILC scaling (L^4) - Improved Staggered Inverter





Other Relevent FNAL Computing Assets

- Connection to the Internet
  - OC3 connection to MREN
  - OC12 connection to ESNET
  - likely future connection to “Starlight” interchange
    - ESNET backbone connection at 10 Gbps
    - also Abilene, Internet2
- GRID-Enabled Mass Storage
  - 5 STK silos, 5800 cartridges each (9940A and 9940B drives)
  - 7 AML2 Quadra Towers, 5000 cartridges each (LTO and LTO2 drives)
  - currently 650 TBytes stored, more than 14 TBytes moved per day
  - designed for 1 Petabyte/year

Software for Mass Storage

- Direct access to tape via Enstore software
- Also, can use tapes via disk cache (FNAL/DESY dcache) using:
  - vanilla FTP (only for reading, not writing)
  - Kerberized FTP
  - X509 certificates
  - GridFTP
- Storage Resource Manager (SRM) project:
  - collaboration with JLAB
  - version 1 API in production
  - version 2 API being coded, will give the following high level services:
    - reliable copy between storage repositories (*eg* NCSA to FNAL)
    - primitives to assist schedulers

Using the FNAL Clusters

- Building code
  - Available compilers:
    - gcc 2.95.3 (/usr/bin)
    - gcc 3.2 (/opt/bin)
    - pgi 4.0.2 (via *setup pgi*)
    - will add Intel compiler
  - MPI:
    - mpich-gm (/usr/local/mpich)
    - mpich-gm-pgi (/usr/local/mpich-pgi)
  - QMP:
    - over mpich-gm (/usr/local/qmp)
    - over gm (/usr/local/qmp-gm, “single port”)
  - VMI (/usr/local/vmi)

Using the FNAL Clusters

- Running your code
  - Batch system is OpenPBS with Maui scheduler
    - scheduler restricts jobs to be contiguous within each clusters
    - 4 nodes on 48-node cluster reserved for 5 minute jobs
  - MPI jobs:
    - use *miprun*
    - node file will automatically be generated from information from PBS
  - VMI jobs:
    - use vmi-launch, node information automatically taken from PBS
  - QMP jobs:
    - use QMPrun
    - user script currently must generate node list from `$PBS_NODEFILE`

Allocations and Quotas

- Allocation restrictions to begin in April:
  - PBS will only accept jobs with project identifiers
    - example: `qsub -A` *myProject* `-l nodes=16 run.script`
    - users can belong to multiple projects
    - accounting will run nightly
- Storage quotas:
  - 4 GB per user backed-up home area
  - project-based data disk quotas
  - project-based tape quotas
  - scratch areas will be available

Web Resources

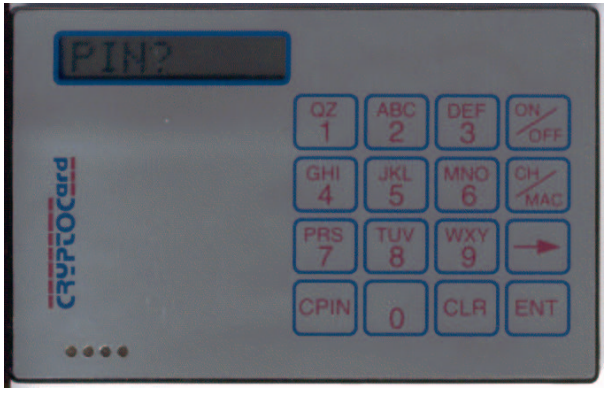
- See http://lqcd.fnal.gov/. Available information:
  - cluster status at a glance
  - node mapping of all jobs
  - documentation
  - user guides
  - benchmark results

Getting Accounts

- See http://lqcd.fnal.gov/user_accounts/
- Summary of procedure:
  - submit a successful proposal to the Scientific Program Committee
  - request Fermilab Visitor ID
  - request Kerberos principal and cryptocard
  - request account on lqcd

Kerberos at Fermilab

- According to lab policy, all logins require Kerberos
  - prevents clear-text passwords, and allows detailed logging of all computer access
  - web pages, other read-only resources do not require Kerberos
- How Kerberos works:
  - to login to a machine, you need a `ticket`
  - a Kerberos principal allows you to obtain a `ticket-granting ticket (TGT)`
    - principals are of the form `username@FNAL.GOV`
    - the Kerberos `kinit` client can be used to get a TGT
    - or, you can use a cryptocard



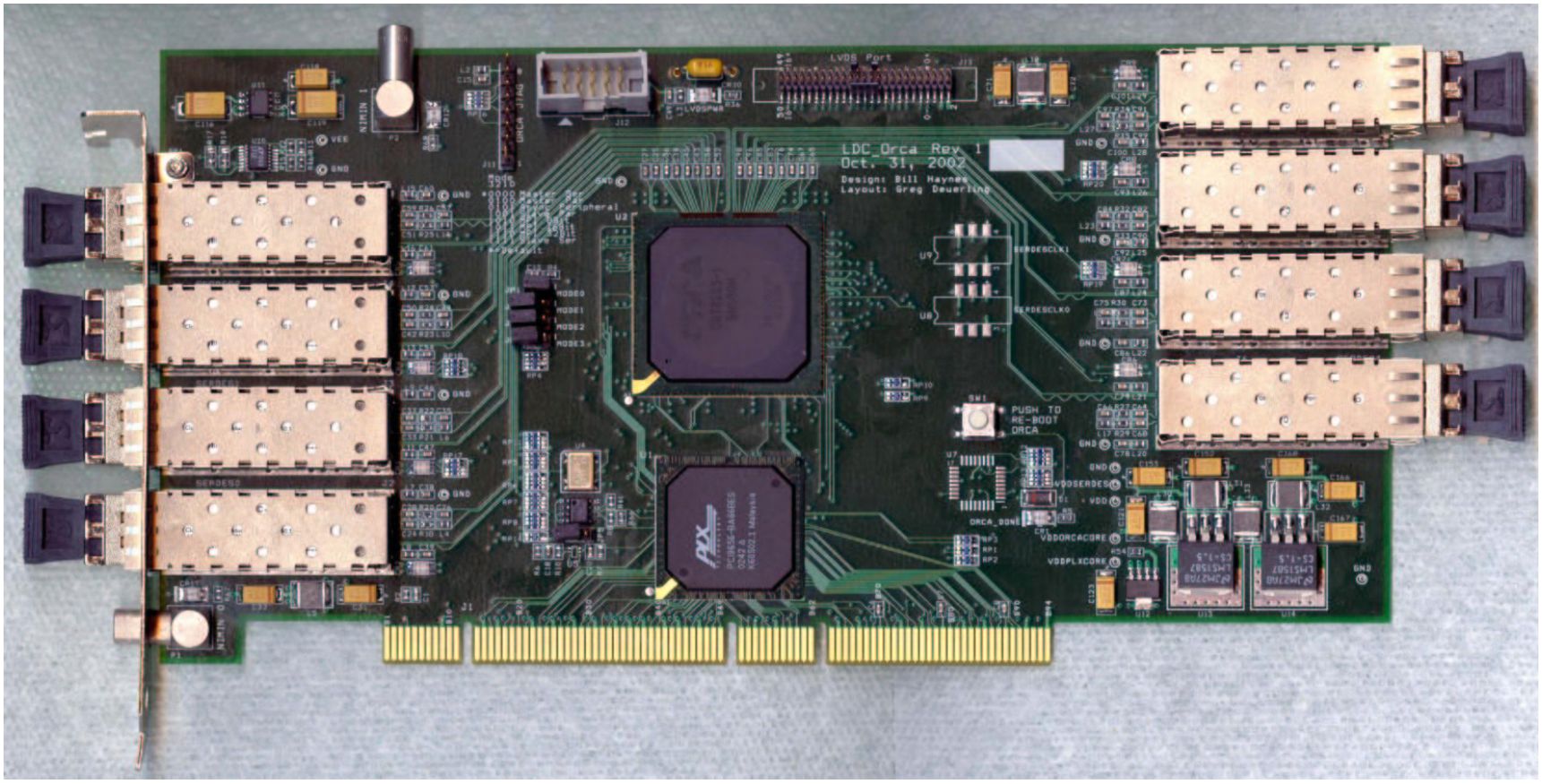
Kerberos Continued

- Standard KRB5 clients can be used (`/usr/kerberos/bin` in many Linux distributions)
  - see web pages for `krb5.conf` file
  - kerberized clients are telnet, rsh, rlogin, ftp, ssh (all encrypted by default)
  - non-kerberized clients will result in a cryptocard challenge
  - for ssh with cryptocard, be sure to hit “return” to `password` prompt
- Obtaining kerberos
  - Standard MIT KRB5 may be installed on your system already, or
  - use Fermi versions - see `http://www.fnal.gov/docs/strongauth/`
  - use statically-linked versions from `http://lqcd.fnal.gov/kerberos/`
- You may already have Kerberos principals from other sites (eg NCSA)
  - if so, you’ll need to modify /etc/krb5.conf to add FNAL.GOV realm and KDC’s
  - holding TGT’s from two realms simultaneously is tricky but possible
    - this is useful, for example, for moving data between sites
  - don’t use `kinit` over unencrypted connections! (FNAL will catch you)

Prototype Routed Mesh Network

- Motives
  - High performance, switched networks (Myrinet, Quadrics, SCI) have good bandwidth, low latencies, mature software, and high prices
  - Switched gigabit ethernet suffers from lower bandwidth, higher latencies, limited switch sizes, immature software, but low prices
  - Gigabit ethernet meshes are very cheap and have good latencies, but immature software, poor non-nearest-neighbor performance, and rigid configurations
- Weapons
  - FPGA's with multiple high speed serial links are now available
  - Some FPGA's will also have multiple PowerPC CPU's aboard
- Opportunities
  - Fermilab is already building prototype PCI cards with these FPGA's for data acquisition
  - Simple nearest-neighbor mesh appears straightforward - with more complex firmware, routing in the network is possible

FPGA-based NIC



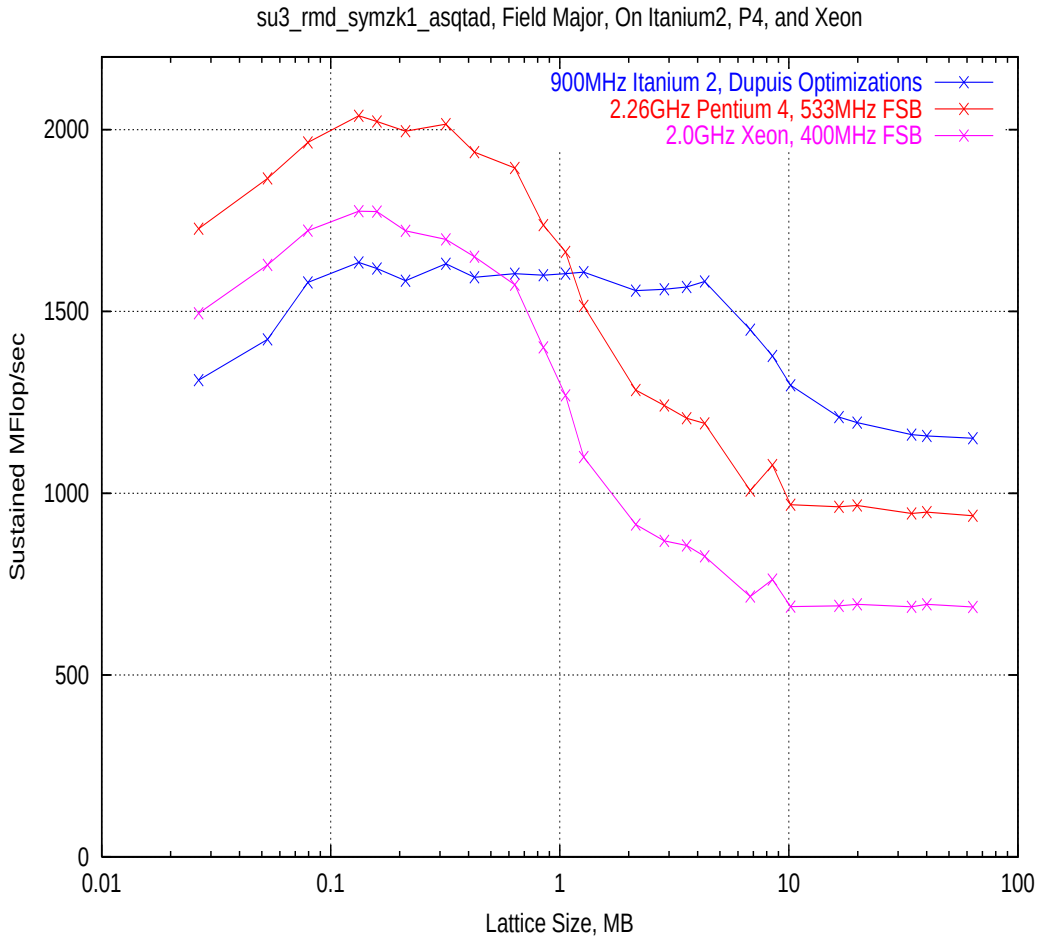
- 8 bidirectional fiber or copper 2 Gbps links (reconfigurable)
- fast/wide PCI interface (PCI-X in next generation)
- long-term goal is to build 4-D or higher mesh with routing
- FPGAs with CPUs could allow sums, reductions to be done by the network

Prototype Mesh Network Schedule

- First boards arrived: February 2003
- First firmware (PIO from computer, DMA to computer) and initial testing: March/April 2003
- Protocol design and implementation: thru Summer
- QMP implementation: Fall 2003
- Strict nearest-neighbor first, routing later

Prototype Itanium2 Cluster

- Encouraging Itanium2 results obtained with help from HP, Summer 2002
- Extensive software development necessary to get good performance
- Prototype cluster:
  - 2 dual 900 MHz machines, one Linux, the other dual boot HPUX/Linux
  - 6 additional single 900 MHz machines
  - Myrinet (LANai 7)
  - SCI (Wulfkit)
  - Online next month



Cluster Strategy

- Clusters show promise as long term, renewable compute resources:
  - first, establish significant facilities (eventually order 1000 machines?)
  - then, continuously track the latest commodity components
  - refresh a fraction of the facilities periodically (replace 33% annually?)
- This can work, as long as:
  - the software is appropriately structured (QMP/QLA allow machine abstractions)
  - commodity hardware stays balanced for our problem (sufficient I/O and memory bandwidth for floating point capability)
  - the non-commodity pieces (networking) don’t break the budget

The Next Fermilab Cluster

- Next FNAL cluster purchase will be late Summer 2003
- Possible architectures:
  - 533 MHz FSB dual Xeon unless something better
  - 800 MHz FSB single P4, only if PCI-X
  - Itanium2 if software development is tenable
  - AMD Hammer/Clawhammer/Sledgehammer if AMD delivers
  - PPC970 is a potential wildcard
- Possible networks:
  - GM over Myrinet
  - GigE over Myrinet
  - GigE mesh