

# The CLuEDØ Linux Cluster

Roger Moore  
Michigan State University

# CLuEDØ

- CLuEDØ is a cluster of Linux desktop machines
  - "Cluedo" is the board game that was renamed "Clue" when it was first introduced to the US
  - Original name a play on the British game "ludo"
- Original cluster developed at MSU with the aim to:
  - Harness desktop's CPU
  - Maintain a uniform user environment
  - Reduce system administration
  - Keep things secure!
- Ported cluster to FNAL when I moved out here

# Current Statistics

- Total of 164 machines with 180 CPUs ranging from 200-2000MHz
- 370+ users from 57+ institutes
  - Majority of DO now has a CLuEDØ account
- 5 geographic sub-clusters
  - Plum, White, Green, Scarlett, Mustard
- 12 system administrators
- 5 large disk servers all using IDE RAID arrays
  - 4x640GB + 1x800GB = 3.28 TB
  - Top group buying a new one with 8x160GB disks (1.28TB), cost ~\$5-6k.

# RedHat Linux 7.1

- **Currently running RedHat Linux 7.1**
  - Updates regularly applied by autoRPM
- **Several extra packages added**
  - Open PBS batch queue system
  - Kerberized OpenSSH, XEmacs, Acroread etc. installed as standard
  - XFS journaling file system for RAID arrays
- **Support for most of latest hardware**
  - Up to date Kernels (2.4.16)
  - WebCAM video conferencing via VRVS and via GNOME-Meeting (H323)

# Disk Configuration

- AutoFS used to NFS mount disks
  - Improves NFS reliability since mounts timeout and are re-made as needed
- Several AutoFS volumes
  - /home: user & services home areas
  - /rooms: big IDE raid servers
  - /work: work (scratch) area, one per node
  - /clued0: local cluster utilities, programs etc
    - Includes OpenOffice, AutoRPM directories etc.
- No insurmountable NFS problems to date
  - Biggest pain is Ripon or dØ2ka going down causing "Stale NFS handle"
  - Fixed to some extent by cron scripts

# LDAP

- Cluster configuration stored and exported via LDAP
  - Modern replacement for NIS/YP
- Improved security
  - Kerberos authentication
  - Encrypted data transport
- Slave servers and automatic failover
  - Automatic updates, no remake database
- ...but close to "bleeding edge"
  - Problems getting things to work initially but now stable
- LDAP is becoming the new standard

# CLuMP

- **Clustered LinUx Management Program:**
  - Home written package of python scripts
- **Entire cluster configuration stored in LDAP database**
  - CLuMP extracts this information and generates configuration files for on each node
  - Adding new node or user will automatically update all cluster nodes
- **Allows a structured configuration**
  - Cluster → Netgroup → Node
- **Restoring/replacing nodes very simple**
  - Configuration saved in LDAP database

# PBS: Portable Batch System

- Two compatible versions
  - OpenPBS: Open source and free
  - PBSPro: Commercially supported version
- CLuEDØ uses OpenPBS
  - Lot cheaper than LSF (it's free!)
  - Scalable to >500 nodes (more that we need)
  - Maui scheduler used for maximum flexibility
    - Also being updated to have support for GRID, see <http://www.supercluster.org/>
- Supports interactive jobs so debugging runs don't interfere with batch system
  - "cluesow -I" gives new shell prompt

# PBS Configuration

- Four queues available
  - SHORT: up to 2 hours CPU/4 hours wall
  - MEDIUM: up to 12 hours CPU/48 hours wall
  - LONG: up to 72 hours CPU/96 hours wall
  - GENERAL: sorts and pipes to other queues
- Users specify limits at submission time
  - Scheduler benefits correct estimates
- CPU shared according to institute's contribution to CLuEDØ
  - Implemented by scheduling priorities
- Jobs executed on fastest available node
  - CPU time allowed scaled to 1GHz P-III
    - i.e. 1 hour limit scales to 2 hours on 500MHz node

# DØ Environment

- Full DØ development environment available on every workstation
  - Mounted from dØ2ka, always up to date thanks to Alan and Paul
  - Local packages mounted from CLuEDØ server for historical reasons (could probably now switch dØ2ka as well)
    - KAI used to need individual licenses
    - Some local customization needed for dØcvs
- Home directories also mounted from dØ2ka
  - Dave Fagan does our backups
  - New scheme for combined dØmino and CLuEDØ home directories

# SAM on CLuEDØ

- SAM installed but still not yet working...
- SAM still undergoing rapid development
  - Things change and break the CLuEDØ setup regularly
  - CLuEDØ is first SAM cluster implementation
- PBS adapter hard coded with executable and queue names (doesn't use PBS API)
  - Work around for queues found
  - Special scripts needed to adapt CLI interface
- Still possible conflicts with PBS scheduling vs. SAM scheduling
  - Need SAM running first to determine if this will be a problem

# Offsite CLuEDØ?

- Could use CLuEDØ Linux distribution offsite as well
  - Installer customized for FNAL but "expert" mode available, could add "offsite" mode
- PBS batch system supports remote submission of batch jobs
  - Could submit jobs directly to FNAL from anywhere
    - But need to ensure FNAL computer security requirements met...PBS doesn't do Kerberos directly
- Documentation needed for server install
  - Workstation install easy once server setup
- UPS/UPD RPMs included for DØ code

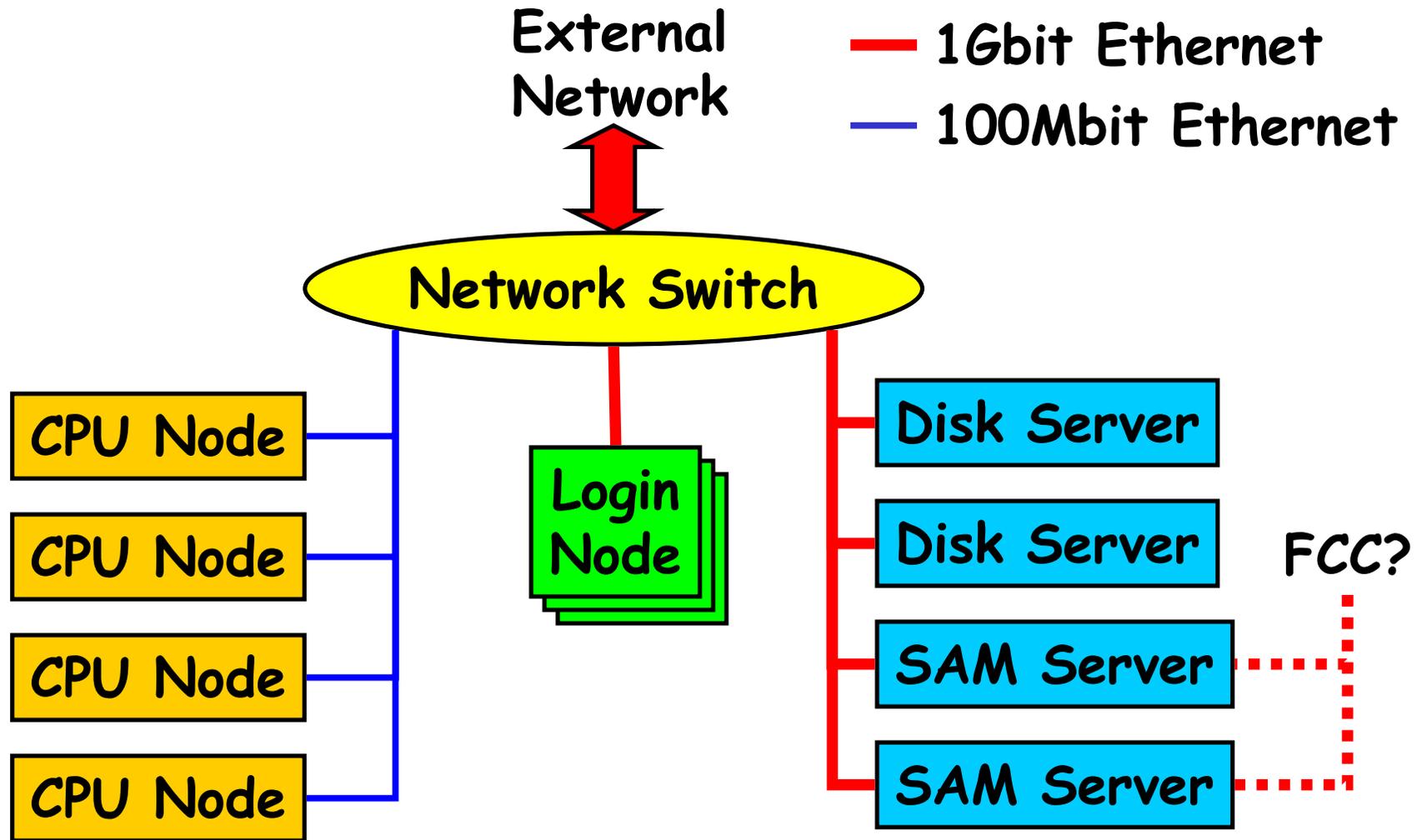
# The Future

- Just over a year ago...
  - 30 machines, 7 institutes and 5 sysadmins
  - We won't keep growing at the same rate!
- CLuEDØ now has an abundance of CPU
  - now need to concentrate on I/O
  - Local network limits analysis jobs to ~200GB data
  - Need to organize SAM and disk servers
- Two analysis cluster proposals accepted by Analysis Computing board:
  - CLuB: A CLuEDØ Back-end cluster
  - DØmino back-end Linux cluster

# CLuB

- The design goals of CLuB are:
  - Execute mid-sized analysis jobs ~2TB
  - Allow institutes to contribute resources
    - Modular design
  - Integrate with the CLuEDØ cluster
    - Accept batch jobs from CLuEDØ nodes
    - Provide CLuB/CLuEDØ shared disk and SAM servers
  - Provide for easy remote access
- Design based closely on CLuEDØ but some differences
  - Single location to allow fast, local networking
  - More uniform hardware (don't choose your own)

# CLuB Design



# Design Details

- Jobs submitted via PBS
  - Read/write data via SAM or local disk servers
- External access to job output
  - RCP/SCP/FTP for remote sites
    - Automate? Still need a cache in case network down when job completes
    - Long term: OpenAFS or replacement
  - NFS for CLuEDØ (RCP etc. still allowed)
- Login nodes provide remote access
  - Also remote access for CLuEDØ

# SAM

- All data input handled by SAM
  - Read from Enstore or DØmino cache
  - Output can be written into SAM but not required
- Only SAM servers have access to external data sources
- Each CPU node has local cache
  - Use RCP to distribute data between caches
  - SAM takes into account local cache contents when assigning files from a project to a batch job

# Hardware

- CPU nodes
  - Fastest, reasonable dual processors available
    - Currently dual 1.6GHz Athlon MP
  - 2GB memory (DØ code needs a lot!)
  - 80GB disk: almost all SAM cache
  - Cost: ~\$2.5k
- Disk/SAM servers
  - Same base as CPU nodes
  - 3ware Escalade IDE RAID controller
    - 8x160GB IDE disks = 1.28TB disk
  - Cost: ~\$5.5k
- Racks with remotely switchable power
  - Power cycle nodes remotely

# Procurement

- Initial cluster seed bought by DØ
  - Network switch (if needed, FNAL specified)
  - Racks, power, terminal server (for serial console)
  - One disk, one SAM server and 10 CPU nodes
- Maintain detailed machine specifications with a single vendor
  - One specification each for disk and CPU servers
  - Update specification (and possibly vendor) on 0.5-1 year timescale
  - Institutes buy direct from vendor

# System Maintenance

- Administer CLuB as a CLuEDØ subcluster
  - Approx. same setup which existing CLuEDØ administrators already know
  - More machines but similar hardware and most without interactive users
  - No 24 hour support but could add remote sysadmins in, for example, Europe
- We would welcome any new sysadmins
  - Need working knowledge of Linux/UNIX
    - But you don't need to be an expert!
  - Fringe benefit: learn how to change job priorities on the batch system! 😊

# Location

- **Feynman Computing Centre**
  - Power, A/C, racks no problem, space?
  - No physical access to machines
  - CLuEDØ access to disk/SAM servers over FCC-DØ link
    - Small files but frequent access
    - No control (bandwidth for NFS/rcp/scp)
- **DAB 2<sup>nd</sup> floor (Pete Simon)**
  - Power, A/C, racks no problem
  - Need to move some offices though
  - Access to Enstore/DØmino over FCC-DØ link
    - Big files but less frequent access
    - SAM can control bandwidth

# Resource Allocation

- CLuB will be a central DØ resource
  - Administered by ORB
- Follow standard DØ practice
  - 50% goes to the institute
  - 50% goes to a physics group of their choice
- Apply this to both disk and CPU
- Share CPU by batch system scheduler
  - Means unused CPU can be used by other groups
  - Over several days allocation will work out (if the system is fully loaded)

# Conclusion

- CLuB complementary to DØmino back-end
  - Better integration to desktop but smaller data volumes
  - Possibly move hardware between them
- Putting together detailed proposal
  - requested by Analysis Computing Committee
- Help very welcome
  - Only Dugan O'Neil, Dave Evans and myself currently involved in CLuB