

# CluB: A CLuED0 Backend Cluster

Dave Evans, Roger Moore, Dugan O'Neil

Analysis Computing Meeting  
December 13, 2001

## Outline

- What is CluB?
- Design of CluB
- Institute Contributions
- Integration with CLuED0
- Hardware and Technical Considerations
- Where do we put it?
- Network Considerations and Tests
- System Administration
- Plans/Costs
- Conclusions

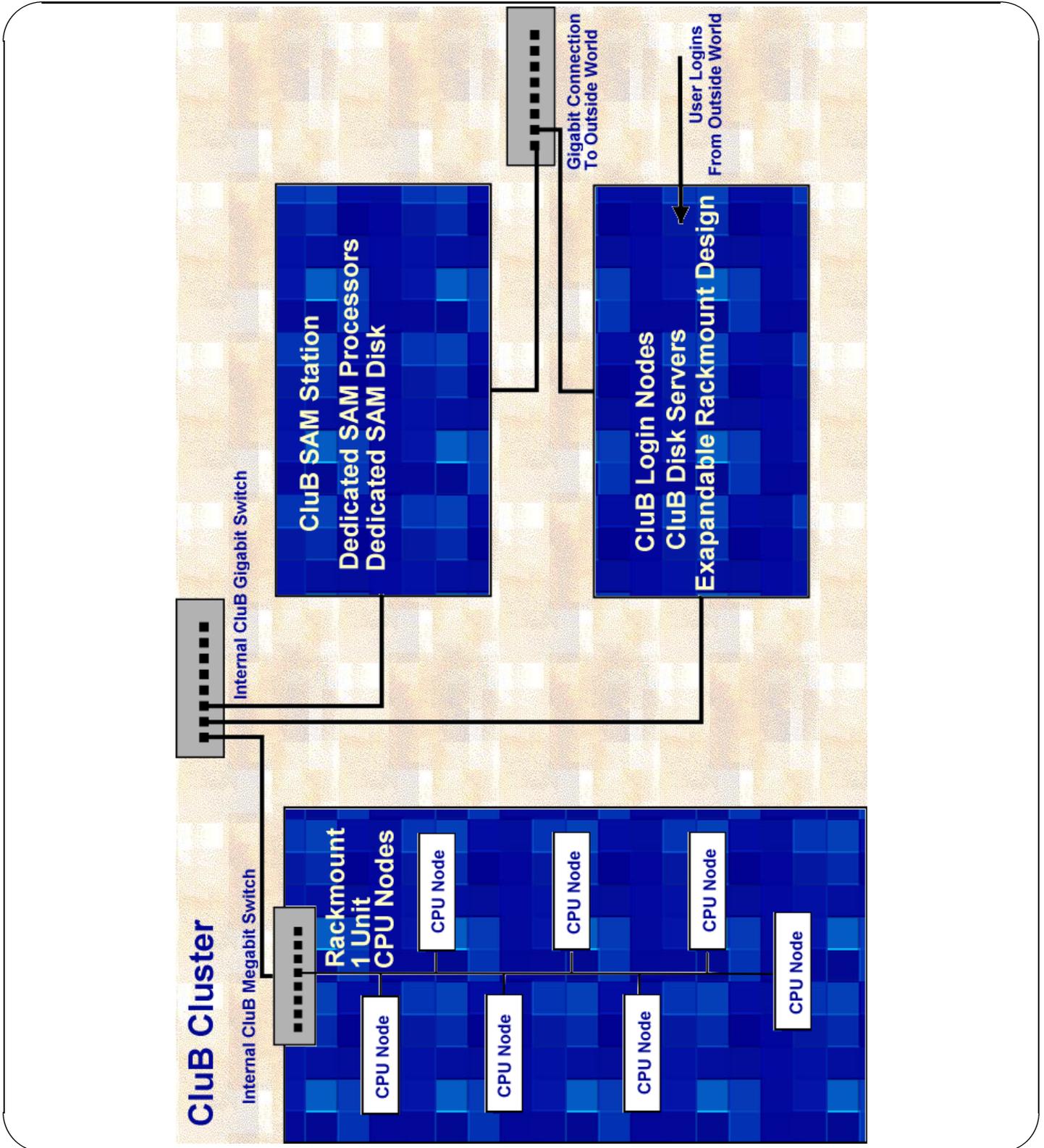
## What is CluB

- CLuED0 is a Linux cluster at D0 providing centralized services and management for desktop machines. Currently 148 nodes, 375 users, 45 institutes.
- CluB is a proposed backend computing cluster for CLuED0.
- D0 computing model calls for a level of computing (IO) between the 10Tb level (d0mino) and the 100Gb level (CLuED0). We want something to crunch through 1Tb of data.
- Should have good integration with desktop cluster. Ease of use, access of output, etc.
- Institutes should be able to contribute directly to the building of the cluster and have some say over its use.
- This is a central resource. The offline resources board has some say as well.

## Design of CluB

- Based on a linux farm model
- Two types of machines:
  - disk servers
  - cpu servers
- Private network for cpu servers. Disk servers accessible from outside.
- SAM station with large central cache

# Design of CluB



## SAM's CluB

- Large ( $\geq 1\text{Tb}$ ) central SAM cache, small caches on each processing node
- Central machine can be a group of dual-processor machines. Easily expanded.
- Only central cache machines access ENSTORE. Processing nodes rcp files from central cache.
- Transfers into the central machine can be throttled by SAM. Max number of simultaneous transfers.
- Number of simultaneous rcp connections from processing nodes can be limited as well (xinetd).

## Integration with CLuED0

- In order to make most efficient use of CluB it needs to be well integrated with desktop environment. Users should be able to
  - submit batch jobs to CluB
  - retrieve the output from CluB to view on CLuED0
- CluB must run PBS for batch jobs
- We need to be able to mount CluB disks from CLuED0 to look at our ROOT files
- It would be most convenient/efficient to use the SAM disk cache on CluB to serve CLuED0. CLuED0 needs some dedicated disk servers for SAM anyway.

## Hardware and Technical Considerations

- All machines are single or dual processor.
  - more processors per box are more expensive
  - linux is best-tested on “desktop” machines.
  - easier for institutes to contribute
  - more processors per box often means slower processors
  - more easily expandable...just add more cheap boxes
  - easy to share within CluB
  - easy to share with d0mino backend, farms, L3, etc.
- AMD processor preferred over Intel
  - significantly faster than P4
  - significantly cheaper than P4
  - easy to dual
- Disk arrays are ide using 3-ware raid controller
  - BIG price difference with fibre-channel. ~\$5000 buys more than 800Gb of space and two fast processors.
  - 4 such systems in clued0 for  $\geq 1$  year without problems
  - opensource driver. worked in RH6.1, works in RH7.1
  - hardware raid 0, 1, 5, 10

## Hardware and Technical Considerations

- Machines are uniform hardware from selected vendor.
- PBS for batch
  - same as CLuED0, d0mino backend (even d0mino if you want)
  - LSF costs money
  - easy to allocate resources according to contribution
  - MAUI scheduler seems robust and very felxible

## Institute Contributions

- Important to find a design made of affordable building blocks. Any institute in D0 should be able to contribute.
- Institute can buy a dual cpu box for  $\sim$ \$2000. Disk server for  $\sim$ \$5000.
- Disk and cpu resources would be shared
  - in practice machine X does not belong to a particular institute. They get a share equivalent to  $X/\text{total}$  of the system resources.
  - disk is shared by traditional D0 model. 50% to institute and 50% to physics group of their choice
  - cpu can be shared in the same way. Contribute a dual processor machine and the batch system allows half credit to be given to institute and half to physics group.
  - very dynamic resource allocation. If nobody is using the cpus today they can be soaked-up by someone who needs them. People cannot come back and ask for the disk they bought 5 years ago....they effectively buy space, not disk.

## Where Do We Put It?

- There are basically two places to choose from: FCC and D0
- Within CluB it doesn't matter which you choose. All traffic is on private network. Only disk servers have access to outside network. What matters is getting data to CluB and getting results out to CLuED0.
- CluB needs to be SAM-friendly and needs to allow transfers straight from tape or from d0mino cache.
- Users need easy access to batch output from their CluED0 desktops. Wherever CluB is put we should allow CluED0 machines to mount large disk areas to access output.
- In future we may wish to use such a facility to do things like PROOF (parallel ROOT). Other examples?
- CLuED0 needs SAM infrastructure at D0 regardless of what is done with CluB.
- We favour D0 as the best location for CluB. Easier to integrate with desktops, more control over network use (see following slides), shared SAM infrastructure with CLuED0.

## Network Considerations and Tests

- What are the implications for the D0 network for CluB in either FCC or DAB?
- If CluB is at FCC it has no problems transferring data files into the SAM cache. However, how do we get the results to our desktops? Mount disk areas? A large number of small access over our link to FCC, no controls.
- If CluB is at D0 we use SAM to manage our bandwidth to Feynman. Control number of simultaneous transfers from tape/d0mino. Transfers are fewer, but larger. SAM gives us control.
- If central cache is shared with CLuED0 we save multiple copying across the Feynman link.
- What about internal network at D0? This is impacted most directly by how people use CLuED0, not where we put CluB.

## Network Considerations and Tests

- How well does this work? We did a test of a SAM configuration on CLuED0 which is similar to the design for CluB (smaller scale)
  - one central cache - 100Gb
  - 5 desktop clients - 10Gb each
  - 5 simultaneous transfers from FCC
  - no limit on simultaneous rcps on server, 2 rcps per client
- Bottleneck: not network. Load/bandwidth on central machine. 45% cpu/encp, 20% cpu/rcp from d0mino, 10% per client.
- Feynman link spike reached approximately 250Mbits/s. Almost all dur to our tests. No effect on desktop use.
- Conclusion: if we limit ourselves to 5 simultaneous transfers from FCC we are fine. We could increase this with networking improvements.

## System Administration

- Should be easier than CLuED0:
  - no interactive users (batch only)
  - uniform hardware
  - we start with CLuED0 experience and tools
- Who does it? A cooperative effort of institutes and computing division would be best.
- Once it is installed it could be managed by skilled shifters. Pool includes CLuED0 admins and D0-CD people.
- Need more “hands-on” support from D0-CD people if it becomes a 24-hour resource.

## Plans/Costs

- Institutes will be responsible for growth of CluB, but D0 must plant the seed.
- Example (rough estimates):
  - 1 SAM server with 8 disks - \$5000
  - 1 Disk server with 8 disks - \$5000
  - 2 cpu nodes - \$4000
  - networking - ?? - depends on existing equipment
- Once networking and disk servers are in place institutes can begin to contribute cpu and/or disk

## Conclusions

- CluB is a processing farm designed as a backend to CLuED0. It should handle datasets between the small (CLuED0) and the large (d0mino).
- Designed to make institute contributions easy. Sharing between institutes and physics groups.
- It must be close friends with SAM. Central large cache, many small caches.
- Preferred location is at D0. Better integration with desktops, more control over traffic on Feynman link.
- D0 needs to provide network infrastructure and a small CluB (a stick?) to get started. Institutes provide the rest.