Facility Evolution

Shigeki Misawa
RHIC Computing Facility
Brookhaven National Laboratory

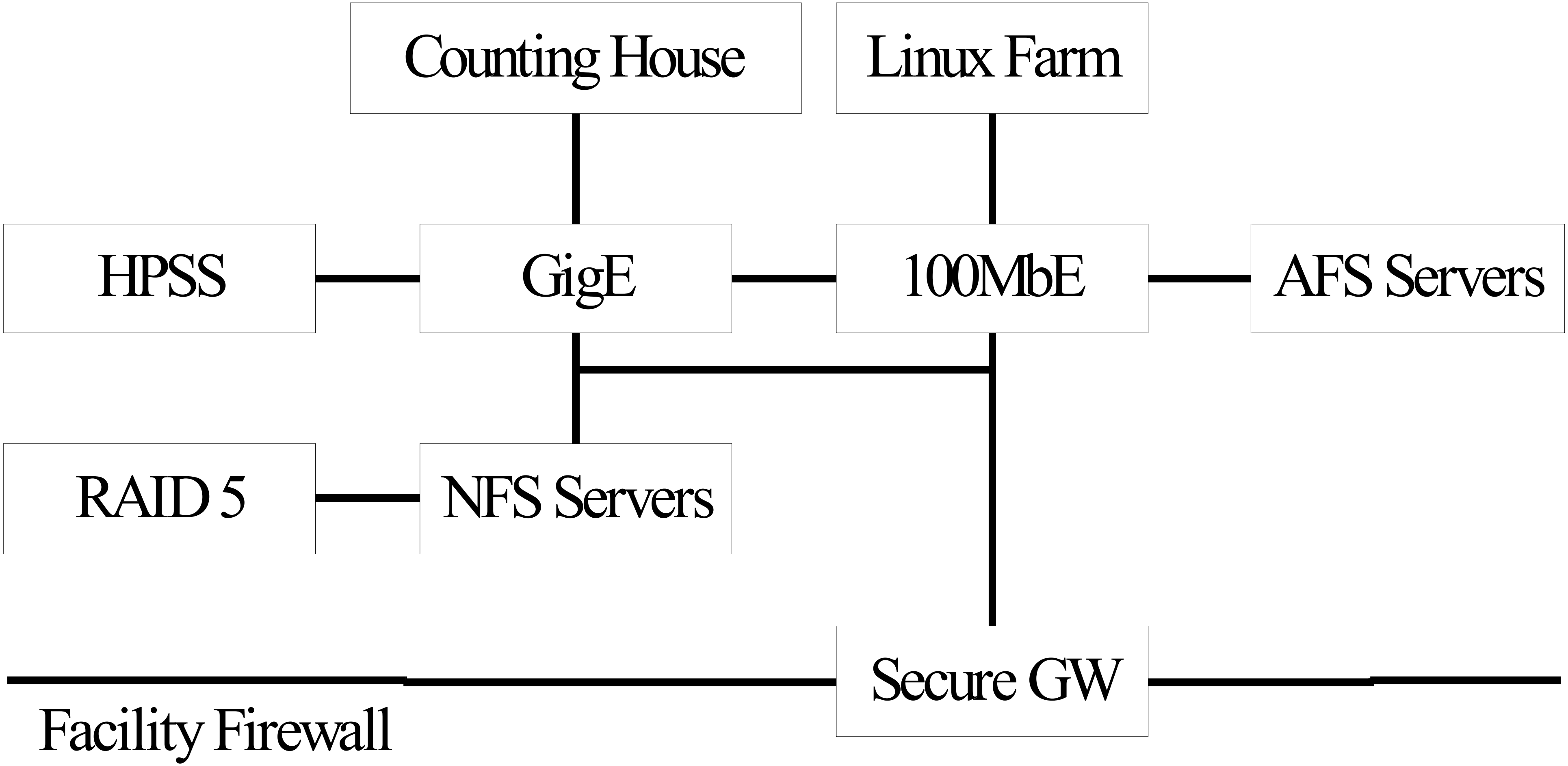
Change Drivers at the Facility

- Users
- Component Costs
- Technology
- Security
- Operational Experience

User Desires

- Unlimited on line storage space
- Uniform, global file namespace
- Infinite file system bandwidth
- Unlimited I/O transaction rates
- Infinite processing power
- No learning curve
- High availability
- $0 cost

Current Configuration



Current Configuration (cont'd)

- HPSS managed tape archive
  - 10 IBM Server
  - 9840/9940 tape drives
- NFS Servers
  - 14 E450
  - 5 V480
  - 100TB RAID5
    - MTI and Zyzzx(CMD)
    - Brocade SAN
- GigE Backbone
  - Alcatel (PacketEngine)
  - Cisco
  - Alteon (phased out)
  - SysKonnect NIC
- ~1000 Dual CPU Linux nodes w/local scratch
- 2 AFS Cells
  - IBM AFS 3.6.x

Current Configuration (cont'd)

- Secure Gateways
  - Ssh/bbftp
- Facility Firewall
- Software for optimized file retrieval from HPSS
- Custom and LFS batch control
- Limited management of NFS/local scratch space
- Other experiment specific middle-ware

Matching User Desires

- HPSS provides near infinite storage space, although not on line
- NFS provides large amounts of on line storage, uniform, global namespace
- Linux Farm provides significant amount of processing power.
- Relatively low cost

- Issues
  - High bandwidth ?
  - Transaction rate ?
  - High availability ?
  - Learning curve ?

User Processing Models

- Reconstruction
  - Processing models among users are similar
  - Processing is well defined
- Analysis
  - Wide range of processing styles in use
  - Transitioning to finite set of processing models difficult to institute.
    - Requires high level experiment acceptance
    - Requires adoption of processing models by individual users

Security Perspective

- Facility Firewall
- On and Off site considered hostile
- Gateways (ssh, bbftp) bypass FW, provide access with no clear text passwords
- Exposure is getting smaller
- Management and monitoring getting better
- Primary vulnerablities
  - User lifecycle management
  - Reusable passwords
  - Distributed file system

Changes affecting security

- Rigorous DOE mandated user life cycle management
- Kerberos 5 and LDAP authorization/authentication
- Tighter network security (inbound/outbound access)
- Grid integration
  - Web services
  - Grid daemons
  - Deployment vs maintenance

HPSS

- Provides virtually unlimited storage ability
- Pftp access inconvenient and can be problematic
- General user access to HSI, a good thing ?
- Work arounds
  - Limit direct access
  - Identify and eliminate problematic access
  - Utilize optimization tools
- Prognosis:
  - Good enough, too much invested to switch

Gigabit/100Mb Ethernet

- GigE not a performance issue
  - Driving it is another question
- GigE port/switch costs an issue
- Vendor shake out somewhat of an issue
- Mix of copper/optical technology a nuisance
- 100Mb compute node connectivity sufficient for the forseeable future

Linux Farm

- Adequately provides needed processing power
- Split between pure batch and batch/interactive nodes
- Nodes originally disk light, now disk heavy
  - Cheaper ($/MB) than RAID 5 NFS disk
  - Better performance
  - Robust with respect to NFS/HPSS hiccups
  - Changing processing model

Linux Farm Limitation

- Individual nodes are now mission critical (since they are stateful)
- Maintenance an issue (HD must be easily swappable)
- Nodes no longer identical
- Node lifecycle problematic
- Local disk space management issues
- Non global namespace and access to data
- EOL of CPU vs EOL of Disk

NFS Servers

- Providing relatively large amounts of storage (~100TB)
- Availability and reliability getting better
  - Servers not a problem
  - Problems with all disk components, RAID controllers, hubs, cables, disk drives, GBICs, configuration tools, monitoring tools, switch, ...
- Overloaded servers are now the primary problem

NFS Servers

- 4x450MHz E-450, 2x900MHz V480 (coming on line soon)
- SysKonnect GigE NIC (Jumbo/non-jumbo)
- MTI and Zyzzx RAID 5 storage
- Brocade Fabric
- Veritas VxFS/VxVM

Performance of NFS Servers

- Maximum observed BW 55 MB/sec (non jumbo)
- NFS Logging recently enabled (and then disabled)
  - Variable access patterns
  - 1.4TB/day max observed BW out of a server
  - Max MB transferred usually, not the most highly accessed
  - Data files accessed, but shared libraries and log files are also accessed.
  - Limited statistics makes further conclusions difficult to make
- NFS Servers and disks are poorly utilized

NFS Logging (Solaris)

- Potentially a useful tool
- Webalizer used to analyze resulting log files
  - Daily stats on 'hits' and KB transferred
  - Time, client, and file distributions
- Poor implementation
  - Generation of binary/text log files problematic
    - Busy FS -> observed 10MB/minute, 1-2 GB/hour log write rate
    - Under high load, nfslogd cannot keep up with binary file generation
    - nfslogd unable to analyze binary log files > 1.5 GB
    - nfslogd cannot be run offline

NFS Optimization

- Without usage statistics cannot tune
  - File access statistics
  - Space usage
  - I/O transactions/FS
  - BW / FS
- Loosely organized user community makes tracking and controlling of user behavior difficult

Future Directions for NFS Servers

- Continued expansion of current architecture
  - Current plan
- Replace with fully distributed disks
  - Needs middleware (from grid?) to manage. Can users be taught to use the system ?
- Fewer but more capable servers (i.e., bigger SMP servers)
  - Not likely ($cost)
  - Will performance increase ?

Future Directions for NFS Servers

- More, cheaper NFS appliances
  - Currently not likely (administrative issues)
- IDE RAID systems ?
  - Technological maturity/administrative issues
- Specialized CIF/NFS servers ?
  - Cost/technological maturity/administrative issues
- Other file system technologies ?
  - Cost/technological maturity/administrative issue

AFS Servers

- Current State
  - 2 AFS Cells
  - IBM AFS 3.6.x
  - ~12 Servers
- System working adequately

- Future direction
  - Transition to OpenAFS
  - Transition to Kerberos 5
  - AFS home directories ?
  - LSF/Grid integration ?
  - Security issues ?
  - Linux file servers ?
    - Stability
    - Backups

Grid Integration

- Satisfying DOE cyber security requirements
- Integration of Grid authentication and authorization with site authentication and authorization
- Stateful grid computing, difficult issues
- Stateless grid computing, not palatable to users
- Transition from site cluster to grid enabled cluster can be achieved in may ways with different tradeoffs.

Future direction issues

- Current facility implements mainstream technologies, although on a large scale.
- Current infrastructure is showing the limitations of using these technologies at a large scale.
- Future direction involves deploying non-mainstream (though relatively mature) technologies or immature technologies in a production environment.
- As a complication, some of these new technologies replace existing mature systems that are currently deployed.