Integrating External Storage Servers with the XT3

Pittsburgh Supercomputing Center

Jason Sommerfield, Paul Nowoczinski,

J. Ray Scott, Nathan Stone

Project Goals

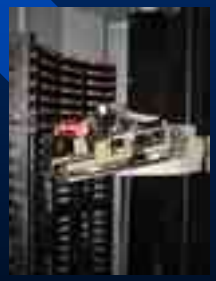
- Expose WAN bandwidth to the XT3
- Facilitate efficient transfers to/from the PSC archiver
- Get the most out of each SIO node
  - More fully utilize the PCI-X on each SIO node
  - Potentially reassign or share SIO nodes between multiple purposes
- Allow for further expansions (e.g. Vis) an incremental performance improvements

Outline

- Background
- Observations & Motivation
- Project Goals
- Plan of record
- Status
- Continued/future efforts

Selected PSC Systems

- XT3 : ~2066p
- TCS: ("LeMieux") 3000p AlphaServer SC cluster
- Rachel/Jonas: 4x64P AlphaServers
- Archiver: 3-level HSM w/PBs of tape capacity (<1PB in use to date)



Selected External Considerations

- Vast majority of users are remote
- PSC is a member of the NSF Teragrid project
  - PSC has a capability platform focus
  - Increased need for frequent bulk data transfers between compute & storage resources of other sites
  - 3x10Gb/s connection to Teragrid (“ETF”) backbone

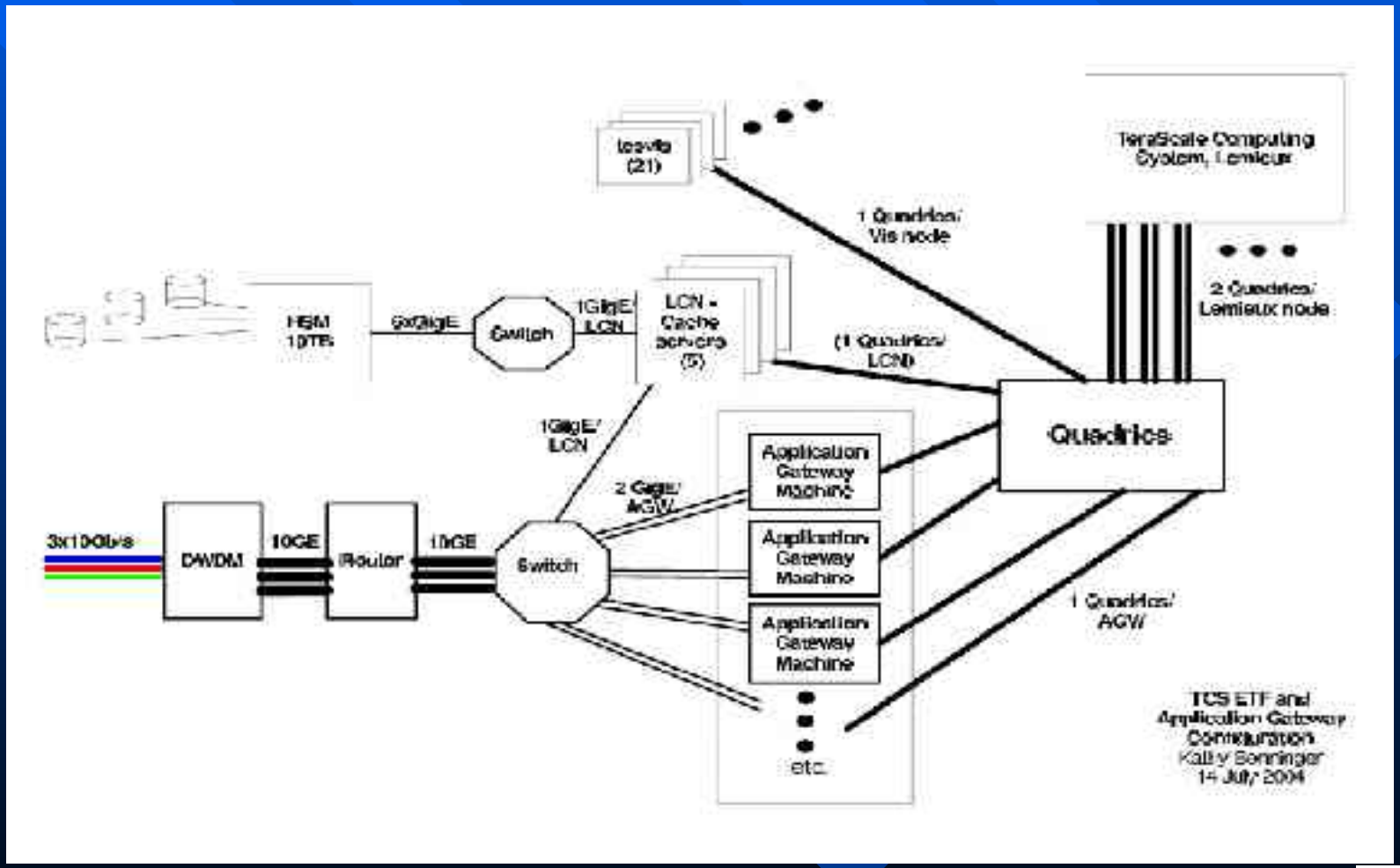
Storage at PSC

- Per system resources
  - slow & steady home directories
  - fast & temporary scratch space
- Access to central archiver
  - High speed access from local compute systems
  - Good, parallel paths to Teragrid
- Lustre testbeds
- Emerging Lustre deployment for XT3
  - 200+TB, 10-15GByte/s performance target

Related PSC Efforts of Note

- TCSIO
  - High speed parallel transport suite supporting advanced features including parallel & 3rd party transfers with rcp-like (simple) syntax
- Application Gateways (AGWs)
  - Novel project to extend the Teragrid network bandwidth into the TCS compute system using commodity hardware.
- Scalable Lightweight Archival Storage Hierarchy (SLASH)
  - Distributed file caching

Application Gateway Nodes



XT3 IO Overview

- A small subset of nodes (~46 for us) run Linux & have PCI-X slots
- These Service & IO (SIO) nodes have a few roles
  a) A few perform system tasks (boot, database)
  b) Some attach to disks via Fibre Channel (for Lustre)
  c) Some operate 10GigE interfaces for external communications
  d) Others are typically assigned as login nodes

Key Observations

- The main use of the 10GigE bandwidth is likely for file transfer.
  - file transfers also involve the Disk IO nodes in order to actually read or write the data
  - other (e.g. interactive sessions) connections could be supported with less aggregate connectivity
- Neither current 10GigE or 2Gb FC adapters typically fill the PCI-X bus
- SIO nodes are running Linux on Opterons, so adapters supported in that general environment stand a chance here

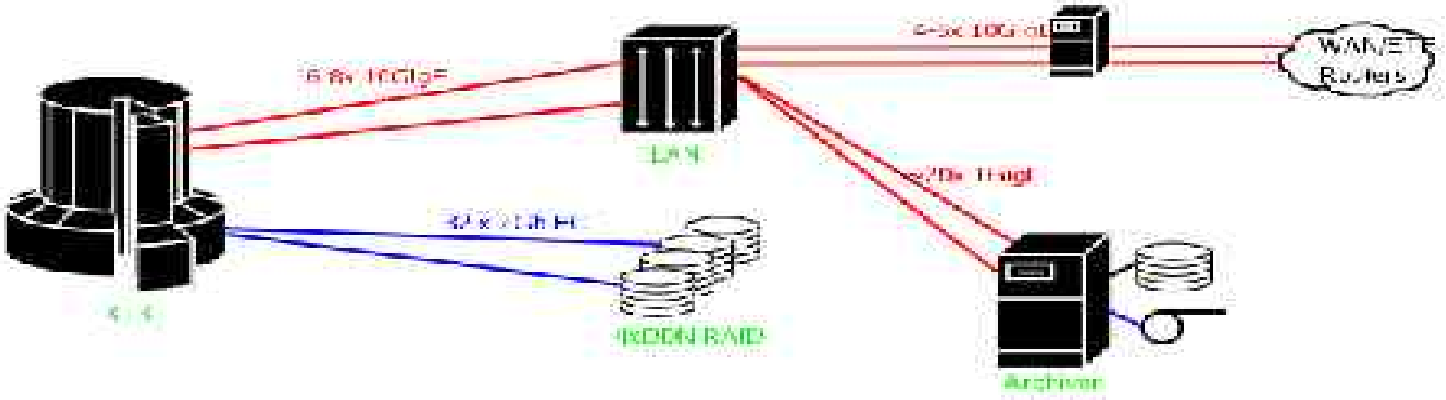
Project Goals (revisited)

- Expose Teragrid bandwidth to the XT3
- Facilitate efficient transfers to/from the archiver
- Get the most out of each SIO node
  - More fully utilize the PCI-X on each SIO node
  - Potentially reassign or share SIO nodes between multiple purposes
- Allow for further expansions (e.g. Vis) an incremental performance improvements

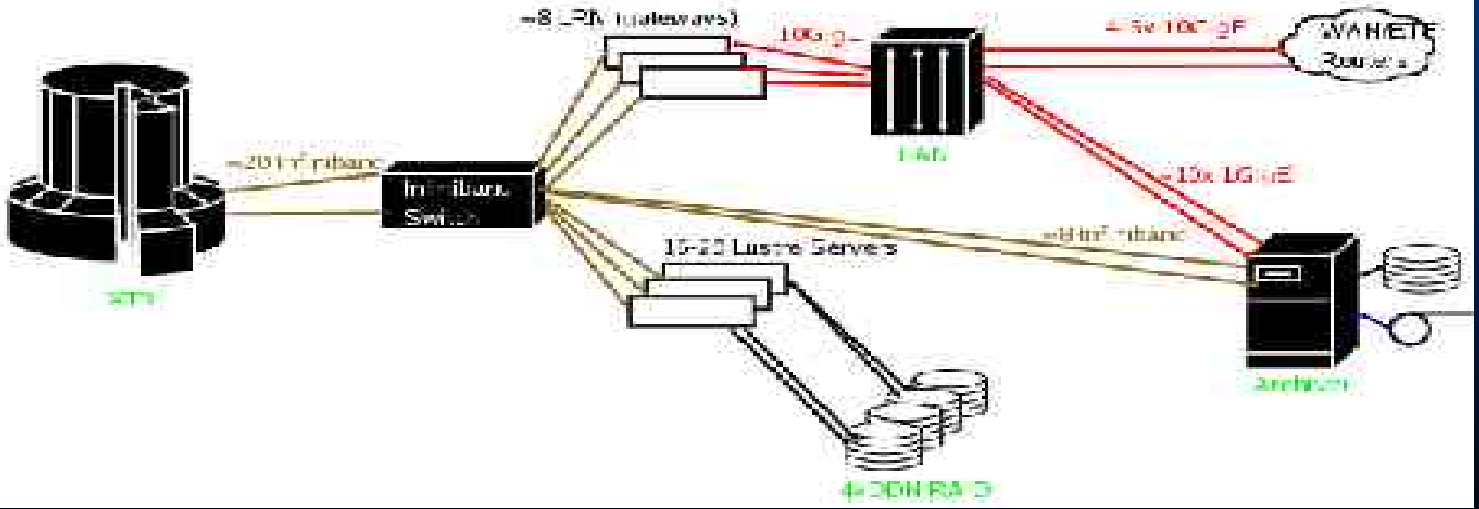
The Plan (, Man)

- Populate the SIO nodes with high speed, multi-protocol (Infiniband) adapters
- Relocate the disk storage to external servers to be accessed over Infiniband
- Develop a path from the Infiniband network to the archiver, the Teragrid, and other systems of interest

Base Configuration

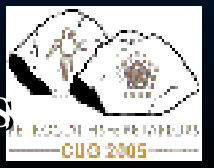


Modified Configuration



Status

- Infiniband in XT3
  - Building kernel modules requires presence of something very near the source for the SIO kernel running
  - Using vendor packaged snapshot of OpenIB (e.g. Mellanox's IB Gold) was the most successful & least painful path for 2.4 kernel support
    - » Documentation also better
  - User level protocols functioning, but some features (e.g. adapter firmware upgrade) still a little challenging via SIO node
  - OpenIB effort (boosted by ASC support) evolving quickly and will likely continue to produce the best software stack, but current revisions require 2.6 kernels



Status (2)

- Lustre over Infiniband
  - Running in small PSC testbeds (commodity Opteron & Xeon systems)
  - Infiniband NAL functioned roughly “out of the box”
  - Reasonable resiliency & performance
    - » Single thread write speeds ~400MByte/s (1 thread) to 2 OSTs (~400MB/s local disk bandwidth per OST)
    - » Multiple thread (6 over 2 clients) write speeds up to ~900MB/s to 3 OSTs (~400MB/s local disk bandwidth per OST)

Status(3)

- Lustre on XT3
  - Currently using current default IO mechanism from compute nodes
    - » ~200MByte/s through yod
  - Aggregate writes from SIO nodes approaches current disk channel rates
    - » ~380MB/s to 2 OSTs with one 2Gb FC to DDN RAID each
  - Expanding Disk subsystem (>200TB, 20-32 FC links)

Status(4)

- Infiniband to 10GigE routing nodes (LRNs)
  - Infiniband vendors developing appliance type version of gateways
    - » GigE available, 10GigE under development
  - AGW-like alternative prototypes systems undergoing testing & tuning
    - » Commodity PC with with PCI-Express Infiniband adapter & 10GigE card of choice
    - » Basic [IP] routing between interfaces works
    - » Node-to-node IP performance improving
      - Currently in the 3-3.5Gb/s range between commodity nodes

Next Steps

- Test SCSI over Infiniband (SRP) as interim means to permit disk & network traffic from the same SIO nodes
  - Investigate concurrent & time-shared multi-purposing of select SIO nodes
- Implement a functional Portals "router" to pass Lustre traffic from compute node **through** SIO node to external Lustre servers
  - Implement an optimized Portals router for Lustre

Next Steps (2)

- Investigate load balancing issues
  - Lustre access through multiple Portals routers
    - » Possibly a non-fixed number of routers
  - Network load balancing through multiple LRNs
- Revisit LRN issues
  - Load balancing of data streams—primarily to Lustre
  - Should XT3 use SDP to LRNs?
  - Should LRNs run data services (e.g. gridftp)?
- Pursue SLASH integration with Lustre

References

- Ongoing work related to this talk
  - http://www.psc.edu/~jasons/xt3
- OpenIB Open Source Infiniband Software effort
  - http://www.openib.org
- TCSIO (parallel and high perfomance file IO suite)
  - http://www.psc.edu/research/presentations/2003/TerascaleIOSolutions.pdf
- Application Gateway Nodes & Qsockets software
  - http://www.psc.edu/~jheffner/talks/agw.pdf
- Scalable Lightweight Archival Storage Hierarchy (SLASH)
  - http://www.storageconference.org/2005/papers/25_nowoczynskip_slash.pdf

Questions or Comments ?

jasons@psc.edu