June 26, 2012

Data Storage

Architectures and Networking

XSEDE

Extreme Science and Engineering Discovery Environment

Agenda

- Overview of Storage Architectures
  - Parallel and WAN File Systems
  - Tape and other Archive Resources
  - “Cloud” and other service-based storage
- Networks and Data Transfer Considerations
  - Research and Education Networks and XSEDEnet
  - Tools and Approaches for managing data in the WAN
- Strategies for Data Management in XSEDE
- Q&A

Background and Goals

- Attendees should complete the session with an improved understanding of storage systems, and the networks that connect them, on campus and in XSEDE

- Part of a series of training sessions on Data Management Infrastructure, Planning, and Execution
- Other courses have focused on planning and writing
- This session focuses on technical foundations
- Suggestions for future topics are encouraged

Storage Architectures - Hardware

- “Data Servers” & I/O Nodes
- Storage media
  - Online and Offline/Removable media
  - Floppies, Disks, Tapes and Chips
  - Wide range of capacity and performance characteristics
  - Trend is capacity increase > performance increase
- Redundant Array of Inexpensive Disk (RAID)

Storage Architectures - Software

- File System: ubiquitous organizing paradigm for storage media
  - Manages structure of directories and files across physical layer
  - USB Drives, DVDs, laptop hard drives, etc all have file systems
- Most file systems act on single, online storage elements
- Many technologies utilize multiple and/or heterogeneous elements, including hardware and software
- Parallel and Distributed File Systems
- Hierarchical Storage Managers
- Distributed Data Management tools

Storage Architectures - Networking

- Storage-Area Network (SAN) – connects storage to systems
  - Fiber Channel, Infiniband, Ethernet
- Local-Area Network – connects systems
  - Ethernet, TCP/IP, NFS, custom file system protocols
- Wide-Area Network – connects sites
  - Similar to LAN in technology but very different performance characteristics

Parallelism and Redundancy in Data

- Parallelism - similar to Cluster Computing:
  - Node X/Disk X cannot do/store my work
  - 1000 of Node/Disk X can do/store my work
- Redundancy/Failure protection also important
- Redundant Array of Inexpensive Disk (RAID)
  - RAID 0 – Multiple drives acting as one
  - RAID 1 – Multiple drives mirroring each other'
  - RAID 5, 6, etc – Data protection + parallelism

Parallelism/Redundancy in File Systems

- Many disks in RAID volumes
- Connect via SAN or directly to many servers
- Servers provide coordinated access to data
- Can distribute data based on files or blocks
- Can replicate data based on files or blocks
- In XSEDE, typically used for high performance cluster file systems

Common Parallel File Systems

- Lustre –
  - Open Source, community-supported
  - Many Linux nodes and EXT2/3/4 file systems
  - Operates over TCP/IP or HPC interconnects
- General Parallel File System (GPFS) –
  - Commercial product from IBM
  - AIX/Linux/Windows, Proprietary file system format, many nodes and volumes
  - TCP/IP + HPC Interconnects

WAN File Systems

- Use existing parallel file system technology
- Can be used for “global” site file systems
- Also used over wide-area networks:
  - TeraGrid, XSEDE, Europe’s DEISA
- Advantage of easy access to data
- Challenges in identity management, performance, and stability

Hierarchical Storage Managers

- Originally for dealing with offline media
  - Tapes, DVDs, etc – high capacity, low performance
- Disk-based cache to store data while active
- "Stage" data to tape or other offline media when inactive
- Retrieve data from tape to disk on request
- Policies used to determine when and where to move data, can be of varying complexity

HSM Hardware Configurations/Issues

- Modern tape libraries hold up to 10K tapes
- Tape capacities 5TB or more
- Tapes subject to mechanical, other faults
  - Multiple copies are common
- Archive capacities now 10s of petabytes
- Data movement still limited to GBs/sec
- Size and performance of cache is critical

“Cloud” Storage and Service-based storage

- Object storage/Cloud Storage
  - Store data on many servers in many data centers
  - Lots of copies, minimal data protection
  - No concept of “folders”, use Unique IDs
- Different paradigm from file systems
- Goal is to be inexpensive, easy to maintain
- Open-source software for implementation
- Commercial storage providers

XSEDE Ecosystem Examples

- HSM: TACC Ranch, NICS HPSS
- Lustre parallel FS : TACC Lonestar $SCRATCH and $WORK
- Lustre WAN-FS : Albedo, IU DC-WAN
- Outside of XSEDE:
  - GPFS parallel/WAN-FS: TACC Corral
  - Storage Services: Amazon, Google Drive, DropBox, etc

Some Usage Considerations

- Awareness of where data is stored and where it needs to be for each workflow step is crucial
- Lustre and GPFS
  - Control of parallelism especially important for Lustre (stripe width)
  - Understanding metadata performance
- HSM: Pre-stage data before access, stage as little data as possible

Data Transfer Protocols and Paradigms

- Simple protocols are single-threaded:
  - HTTP, FTP, SSH/SCP
  - Widely supported, easy to use
  - Very little opportunity for performance tuning
- Specialized protocols allow parallelism:
  - GridFTP, iRODS, bbcp
- Control mechanisms for many transfers/files:
  - Globus Online – Data, iRODS, GFFS

Some Data Transfer & Management Tools

- GridFTP – Extension to FTP protocol to support parallel transfer and large transfers
- GO-Data – Tool/Interface to manage GridFTP transfers
- Global Federated File System – XSEDE service for managing distributed data/resources
- iRODS – Open source client/server tools for data replication and distributed management

Issues in Data Transfer and Networking

- The Speed of Light is still a problem
  - Long-distance networks are relatively slow
  - Bandwidth is increasing but latency remains
- Parallelism to the rescue again:
  - Multi-threaded transfers
  - Multi-node transfers (striping)
- Other optimizations:
  - Reusing network connections
  - Bundling small files

Research and Education Networks

- Many states and regions have RENs:
  - Texas LEARN, California CENIC
- RENs typically are connected to or coincident with national nets like NLR, Internet 2
- Most educational institutions, and XSEDE, are connected to one or more of these networks
- Important to ensure your traffic uses non-commercial networks where possible

Understanding Network Performance

- The slowest link sets the upper limit to bandwidth
- XSEDE network = 10Gb/sec
- Campus networks = 1 or 10Gb/sec
- Campus connections = 10Mb-1Gb/sec
- Very important to understand the full path traveled by your data.

Data Transfer Performance

- Each element is a potential bottleneck
- Need to understand performance of source and target storage as well as the network
- In many cases, simple options are close to the performance of complex options
- Can use memory-to-memory test transfers to generate baseline network numbers
- Talk to site staff about storage performance

Common Network "Gotchas"

- Campus-level firewalls restricting access to ports used by GridFTP/GO-Data/GFFS/etc
- MTU mismatches (9000 vs 1500)
- Bandwidth throttling/maximum transfer limits
- 100Mb networking equipment on campus
- Incorrect/suboptimal routing of traffic

Data Management Strategies - 1

- Goal should be to move as little data as possible as few times as possible
- Where does data start?
- Where is data utilized?
- Understand resources in terms of performance/capacity relative to your data
- XSEDE archives are not good long-term homes for data!

Data Management Strategies – 2

- Find a long-term location for your data
  - Campus-level shared resources
  - Discipline-specific shared resources
  - Commercial or self-run resources
- Store only what you need in XSEDE systems
  - Working data sets
  - Simulation output
- Always copy data to the long-term resources

Contact Information

Chris Jordan

data@tacc.utexas.edu


Our reach will forever
exceed our grasp, but,
in stretching our horizon,
we forever improve our world.
XSEDE
Extreme Science and Engineering
Discovery Environment