SOS10
Complexity in Scalable Parallel File Systems

Harriet Coverston
Distinguished Engineer
Sun Microsystems

March 9, 2006

Harriet Coverston Bio

- Distinguished Engineer at Sun Microsystems
- VP of Technology and founder of LSC, Inc., which was acquired by Sun in 2001
- SAM-QFS architect
  - QFS is a high performance shared file system
  - SAM is an automated data management system (HSM)
- 35 years of Kernel design and development in HPC operating systems
  - LLNL, Control Data Corporation, LSC, Sun Microsystems
- Currently the DARPA HPCS file system team lead

Bandwidth (1 of 2)

- In the next decade is the bandwidth transferred into or out of one “high end computing file system” (a) going down 10X or more, (b) staying about the same, (c) going up 10X or more, or (d) “your answer here”, as a result of the expected increase in computation speed in its client clusters/MPPs, and why?
- Bandwidth will increase to support the increased processing engines
- Bandwidth will go up 10X to 20X or more
  - 2 GB/sec per TF = 2TB/sec peak for a petaflop machine
  - 1TB/sec sustained I/O bandwidth for a petaflop machine

Bandwidth (2 of 2)

- Metadata ops (IOPS) will not require more bandwidth; however, there will be more concurrent metadata ops
  - Increasing concurrent metadata operations are required to support the increasing data retrieval requirements
    - Finding any one file among trillions of files
    - Finding anything in the petabytes of data (data mining)
- Need to distinguish between high-bandwidth data access and low-bandwidth metadata ops because we cannot average the two types of I/O
- Conclusion: Bandwidth will increase 10X or more to feed the ever increasing processors

Spindle Count (1 of 2)

- In the next decade is the number of magnetic disks in one "high end computing file system" (a) going down 10X or more, (b) staying about the same, (c) going up 10X or more, or (d) "your answer here", as a result of the expected increase in computational speed in its client clusters/MPPs, and why?
- Disk drive capacity doubles every 18 months
  - Doubling of capacity equates to doubling of density
  - Doubling of capacity only increases bandwidth by 40%
    - Dumping the data off the disk or rebuilding the disk only increases by 40%
  - Each doubling of capacity does not effect seek performance and access to the data

Spindle Count (2 of 2)

- Disk drive latency is the about the same as 1990 drives and this is the bottleneck
  - Many transaction applications do not use the entire capacity of the disk drive
- Disk drive count must increase to solve the metadata ops problem
  - Solid state disks can help, but the cost is high and capacity is low
- Conclusion: spindle count will increase to support the ever increasing requirement to shorten “time to first byte”

Concurrency (1 of 2)

- In the next decade is the number of concurrent streams of requests applied to one “high end computing file system” (a) going down 10X or more, (b) staying about the same, (c) going up 10X or more, or (d) “your answer here”, as a result of the expected increase in concurrency in client clusters/MPPs, and why?
- HPC compute scaling means more processors which means more concurrent threads which means more concurrent I/O requests
- Concurrency is going up at least 10X or more
  - DARPA HPCS requires 32K file creates/sec.
  - DARPA HPCS creates 10K files in the same directory for application checkpoint

Concurrency (2 of 2)

- Scaling metadata
  - Distributed metadata for fault tolerance and performance
  - Concurrency
    - One thread opening m files
    - n threads opening 1 file
    - n threads opening m files
  - Trillions of files
- Conclusion: concurrency will increase to support the ever increasing processors appetite for data

Seek Efficiency (1 of 2)

- In the next decade is the number of bytes moved per magnetic disk seek in one “high end computing file system” (a) going down 10X or more, (b) staying about the same, (c) going up 10X or more, or (d) “your answer here”, as a result of the expected increase in computational speed in its client clusters/MPPs, and why?
- Data capture (satellite, seismo, signal, ...) will require larger request sizes
  - 16MB request size expected to grow to hide the latency
  - Requirement for minimum of 90% of the channel speeds, scaling up linearly
  - DARPA HPCS requires 30GB/sec per node

Seek Efficiency (2 of 2)

- Many HPC applications need to locate the file and mine the data in the file
  - Data is becoming more like a data base
    - Indexing required
  - The amount of data per seek for these applications is shrinking, probably approaching 1K
- SLAC's Petacache project is abandoning disk (6ms) latency, for DRAM and SSD, Richard Mount says
  - “Unfettered sparse access by 100,000s of processes to kilobyte objects in petabytes of immutable data”
- Conclusion: request sizes are both increasing and shrinking for different applications

Failure Domains (1 of 2)

- In the next decade is the number of independent failure domains in one “high end computing file system” (a) going down 10X or more, (b) staying about the same, (c) going up 10X or more, or (d) “your answer here”, and why?
- Number of disk drives expected to increase to 10,000 (max 60,000) for a petascale machine
- The “disk drive” fault domain will increase; however, the exposure can be mitigated
  - RAID 5 or 6
  - N-levels of mirroring
  - Combination of RAID and mirroring

Failure Domans (2 of 2)

- With 10,000 drives, always expect a broken drive
  - Rebuild times are increasing with increasing capacity
- Replicating data, self-repairing storage systems, fail-in-place models, and end-to-end check data, are needed
  - Mirroring is now economically viable because of the capacity increases
- Object-based storage is aware of its data and can independently act on the failing data
- Conclusion: the number of failure domain may stay the same, but the “size” of each independent failure domains will probably increase significantly

Coping with Complexity (1 of 2)

- Coping with Complexity: If you have answered (c) for one or more times, explain why these large increases are not going to increase the complexity of storage software significantly? Are you relying on the development of any currently insufficient technologies and if so, which?
- Intelligent, secure storage (T10/1355 object-based storage is the protocol) will help solve storage complexities
  - Move higher-level storage functions out to the devices
  - Execute these functions in parallel in order to scale
  - Support secure client access because credentials are checked on every access

Coping with Complexity (2 of 2)

- Object-based storage enables interchange because it is a standard
  - Standards are good because users not locked into one vendor
  - Open source does not mean it is a standard (remember Unitree)
- Intelligent storage devices will be self-repairing, self-managing, secure, redundant, and self-protecting of the data that they hold

Development Time Trends

- If complexity is increasing in high end computing file systems, is the time and effort required to achieve acceptable 9s of availability at speed (a) going down 10X or more, (b) staying about the same, (c) going up 10X or more, or (d) "your answer here", and why? Are you relying on the development of any currently insufficient technologies, and if so, which?
- Yes, it is taking more time to achieve acceptable 9s of availability
- Operations cost keep exponentially increasing
  - > $1 in storage costs $8 dollars to manage (Gartner, 2003)
- How do we stop this trend?

Automate Management

- Management
  - Self-managing infrastructure
  - Ease of repair
  - Performance management
- Access Preservation
  - Nodes enter/exit without disruption
  - Fault tolerant operation
  - Continuous access
- Data Preservation
  - Disaster recovery
  - Versioning for checkpoint support

Intelligent Storage

- Move data-centric operations to storage devices
- Explore implications of object-based storage
  - Quality of service
  - Security
  - Data integrity
- Automate storage system management
  - Storage nodes become peers of compute nodes
  - Standards – use SNIA SMI initiative
- Enable data sharing with object storage pNFS
  - Contributing to the IETF object storage pNFS draft
    - Plan to demonstrate interoperability

SOS10
Complexity in Scalable Parallel File Systems

harriet.coverston@sun.com