Overview

The Past
- The Evolution of High Performance Computing

The Present
- High Throughput (Technical) Computing
  - Applications
  - Architectural Characteristics
- Cluster Computing
  - Trends and Challenges
- Scalable Storage Architectures
- Object Storage: Scalable Storage for High Throughput Computing
- Panasas ActiveScale™ Storage Cluster

The Future …
The Evolution of HPC: Clusters

Recent Top500.org highlights

- It's all (mostly) about clusters
  - #3 Virginia Tech X (“SuperMac”)
    - Apple G5 workstations; Mellanox Infiniband interconnect
  - #4 NCSA Tungsten
    - Dell PowerEdge nodes; Myrinet interconnect
  - #5 DOE PNNL
    - HP/It2 nodes; Quadrics interconnect
  - #6 Los Alamos
    - AMD Opteron; Myrinet interconnect

- A total of 189 systems are now using Intel processors. Six months ago there were 119 Intel-based systems on the list and one year ago only 56.
- 208 systems are now labeled as clusters, up from 149. This makes clustered systems the most common architecture in the TOP500.

Cluster Computing Architecture
Technology Trends Driving Clusters

Commoditization of processing
- Phase 1: Vector -> RISC
- Phase 2: RISC -> CISC/EPIC/CMP (workstation / desktop PC processors)
  - Vector instruction sets; FPGA accelerators

Commoditization of networking (ongoing)
- Significant per-port cost reductions for high speed interconnects
- Increased usage of commodity switching (GbE)

Improved management tools
- Provisioning, monitoring, and reporting

Maturation of the software stack
- Compilers, libraries, job schedulers

High Throughput Computing

Key Application Drivers
- High performance computing (clusters)
  - Tens to thousands of processors
- High data throughput requirements
  - Multi-GB/s aggregate
- Large datasets
  - GBs to TBs working store
- Production-oriented
  - A day’s worth of processing in a day

Life Science
- Bioinformatics
- Sequence Analysis
- Expression Analysis

Gov’t Science
- Simulation
- Visualization
- Homeland Defense

Oil and Gas
- Seismic processing
- Earth science

Media
- Rendering
- Post-production
- Streaming

Commercial HPC
- Financial analysis
- Aeron/Auto Design
Cluster Computing Architecture

Applications
Serial and Parallel

Cluster Middleware
(Management, DRM, Parallel programming support)

Compute Nodes
(CPU, Memory, Local disk)
Compute Nodes
(CPU, Memory, Local disk)
Compute Nodes
(CPU, Memory, Local disk)

Node Interconnect

Shared Storage

Shared Storage: The Promise

Shared storage cluster computing

- Compute anywhere model
  - Partitions available globally; no replicates required (shared datasets)
  - No data staging required
  - No distributed data consistency issues
  - Reliable checkpoints; application reconfiguration
  - Results gateway

- Enhanced reliability via RAID
- Enhanced manageability
  - Policy-based management (QoS)
A Common Production Computing Model

Example: Shared Storage Sequence Analysis

Sequence Extraction

Sequence Acquisition

Compute Cluster

(Sequence Analysis)

Shared Storage System

(server, NAS)

Data Preparation

/panasas/ (RAID)

Storage System

(Analysis Working Store and Results Gateway)

Analysis and Visualization

Cluster Computing Operational Challenges

<table>
<thead>
<tr>
<th>Operational Challenges</th>
<th>Scalable Shared Storage System Features and Benefits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Application Development</td>
<td></td>
</tr>
<tr>
<td>Multiple filesystems and mountpoints</td>
<td>Unified filesystem and single global namespace simplify development</td>
</tr>
<tr>
<td>Ensuring I/O efficiency and scalability</td>
<td>Scalable shared storage simplifies application development</td>
</tr>
<tr>
<td>High concurrency; high aggregate I/O</td>
<td>Highly scalable storage system</td>
</tr>
<tr>
<td>Job scheduling driven by data availability</td>
<td>Simplified job scheduling; compute anywhere</td>
</tr>
<tr>
<td>Rapid checkpointing to non-volatile storage</td>
<td>High throughput speeds checkpointing</td>
</tr>
<tr>
<td>Application Data Sharing (pipelines; analysis/viz apps)</td>
<td>Shared checkpoints enable dynamic repartitioning</td>
</tr>
<tr>
<td>Archiving, Staging, De-staging</td>
<td>High performance storage minimizes delays</td>
</tr>
<tr>
<td>Load Balancing (capacity and performance)</td>
<td>Automatic capacity and load balancing</td>
</tr>
<tr>
<td>Disaster Recovery and Business Continuity</td>
<td>Fast online recovery; hardware redundancy</td>
</tr>
<tr>
<td>Backup, Restore, Nearline</td>
<td>Scalable backup support (per volume); very high performance disk-to-disk backup and recovery</td>
</tr>
<tr>
<td>Consistent point-in-time backups</td>
<td>Snapshots</td>
</tr>
<tr>
<td>Regulatory (Sarbane-Oxley)</td>
<td>TBD</td>
</tr>
<tr>
<td>System Management</td>
<td></td>
</tr>
<tr>
<td>Capacity increments</td>
<td>Single system image and rapid provisioning; capacity balancing</td>
</tr>
<tr>
<td>Failure / Recovery</td>
<td>High performance backup and recovery with multiple clients</td>
</tr>
<tr>
<td>Volume configuration and management</td>
<td>Virtual volumes simplify data management tasks</td>
</tr>
<tr>
<td>User file recovery (snapshots)</td>
<td>Snapshots support point-in-time recovery of user files</td>
</tr>
</tbody>
</table>
Shared Storage Challenges

Performance, scalability and manageability issues
- Single file system performance limited
- Multiple volumes and mount points
- Manual capacity and load balancing
- Large quantum upgrade costs

Overcoming the Bandwidth Bottleneck
Los Alamos National Laboratory recently received the Storage Networking World “Best Practices in Storage” Award for Innovation and Promise, based on their leadership in promoting object-based storage for high throughput computing and ushering in the age of the commodity supercomputer.
Cluster Computing Storage Approaches

**Local Disk Solutions**
- Excellent scalability
- Poor manageability
- Data staging
- Brittle filesystem abstractions (PVFS)

**SAN Filesystem Solutions**
- Hardware provisioning and sharing architecture
- Limited concurrency
- Cluster partitioning
- Expensive
- Complex management
  - LUN, volume, filesystem
  - Manual load balancing

**NAS Solutions**
- File sharing architecture
- Simple management
- Performance bottlenecks
- Manual load balancing

**Clustered Filesystem Solutions**
- The promise of scalable shared storage
- Integration and hardware optimization currently required
- Complex management
  - Hardware and software

Motivation for a New Storage Architecture

- **A highly scalable, interoperable, shared storage system**
  - Improved storage management
    - Self-management, policy-driven storage (i.e. backup and recovery)
  - Improved device and data sharing
    - Shared devices and data across OS platforms
  - Improved storage performance
    - Quality of service, differentiated services
  - Improved scalability
    - Of performance and metadata (i.e. free block allocation)
Object Storage Fundamentals

- An object is a logical unit of storage
  - Like simple files with uninteresting names
  - Lives in flat name space with ID
  - Identified by a <server, group, object #> number (many bits)

- Contains application data and attributes
  - Metadata: block allocation, length (similar to inode)
  - QoS requirements, capacity quota, etc.
  - Some well known, others opaque to storage server

- Has file-like methods
  - create, delete, read, write, getattrib, setattrib

- Three types of objects:
  - Root Object - one per device
  - Group Object - a “directory” of objects
  - User Object - for user data

- Objects enforce access rights
  - Strong capability-based access control

Object Storage System Concept

How it works:
Object-based Storage is an emerging networked storage standard that delivers unprecedented performance and scalability within a single management domain.

1. Client processes make independent file requests (e.g., stat, open) of the clustered metadata manager.
2. Authorized clients receive object maps and capabilities.
3. Clients use their credentials to perform parallel direct-to-disk requests across the GbE switching network.
OSD File System Architecture

Traditional Model
- APPLICATIONS
  - SYSTEM CALL INTERFACE
  - File System
    - User Component
    - Storage Component
  - Sector/LBA INTERFACE
  - Block I/O Manager

Object-Based Model
- APPLICATIONS
  - SYSTEM CALL INTERFACE
  - File System
    - User Component
    - OSD System
    - Storage Component
  - OSD INTERFACE
  - Block I/O Manager

Object abstraction
- Re-divides responsibility for managing the access to data
- Assigns to the storage device additional activities (space management)

Scale-out (Cluster) Computing

Traditional Parallel Computing
- Large SMPs (Scale Up)
  - Issues
    - Complex Scaling
    - Limited Bandwidth
    - I/O Bottleneck
    - Inflexible
    - Expensive
  - Single data path
- Monolithic Storage

Distributed Computing
- Linux Compute Cluster (Scale Out)
  - Benefits
    - Linear Scaling
    - Extreme Bandwidth
    - High Random I/O
    - Ease of Mgmt
    - Lower Cost
  - Parallel data paths
- Panasas Storage Cluster

High concurrency
High aggregate bandwidth
Unprecedented Performance and Scalability

**Scalability Axes**

- **Capacity**
- **Performance**
  - Sequential throughput; IOPS
  - Single stream (session); Aggregate
- **Concurrency (# users)**
- **Operational**
  - RAS – reliability, availability, serviceability
    - Backup & recovery; disaster recovery; business continuation
  - Staffing and expertise
  - Support
- **Architectural**
  - Technology platform; commoditization; standards
Object Architecture Momentum

- Panasas helping lead industry adoption
  - OSD Working group
    - 26 members including: Intel (LEAD) with Panasas & others
  - NSIC NASD group chaired by Gibson
  - Evangelizing through industry events
- Building on industry-wide acceptance
  - EMC, IBM, Hitachi & Seagate endorsements
- Converging with I/O
  - A new landscape for Objects
    - Storage and IPC
    - IPC and Networking
    - Storage and Networking

Converging with I/O

Panasas Represents …

- … a Novel Networked Storage Architecture Delivering Unprecedented Performance and Scalability
- … an Emerging Storage Standard that Leverages Commodity Components
- … a Highly Scalable Parallel and Distributed File System
- … an easy-to-deploy and easy-to-manage Shared Storage System well suited for Scale-out architectures
- … a Financially Sound Company Delivering a Supported Object Storage File System
  - Strong customer success: largest Linux storage sale ever
  - World-class Strategic Partners
Panasas Storage Cluster Hardware

- Stores 5 TBs per shelf
- Orchestrates system activity
- Balances objects across StorageBlades
- Stores data objects
- Smart disks using Serial ATA
- 240, 500 GB
- Dual power supplies
- Redundant fans
- Battery back up
- 16-Port GE Switch Blade

Single Shelf Performance & Scaling

![Graph showing single shelf performance and scaling]

- Blue line: Read (64K)
- Pink line: Write (64K)
Unified Shared Data Access Architecture

Tailored implementation and performance

Unified Namespace

Multiprotocol Support

NFS/CIFS
Up to 7x Random I/O Scalability

DirectFLOW
Up to 30x Data Throughput

StorageCluster

Component Architecture Convergence

Microsoft
- COM (VB)
- DCOM
- MTS
- .NET

Java
- Java/Swing
- RMI
- Beans
- J2EE

OMG CORBA
- CORBA
- IDL
- IIOP
- OMA Services
- MDA

WWW
- HTTP (GET/PUT)
- URLs
- CGI
- ASP / JSP

Web Services
(XML / SOAP)
Grid Computing

- Distributed Resource Sharing in a multiple authority, WAN environment
  - Processing
  - Data / storage
  - Applications
- Dynamic membership
- Heterogeneous (HW / OS)
- Fine-grained security architecture

WAN Data Sharing (DataGrids)
Diskless Cluster Computing

- Elimination of non-essential components
  - Increased reliability
  - Reduced heat/footprint improve density
- Elimination of non-essential drivers
  - Increases node stability
  - Reduces node/cluster boot time
- Single System Image models
  - Scyld, OpenMOSIX
- Completing the "Commodity Supercomputer"