Symmetric Active/Active Metadata Service for Highly Available Cluster Storage Systems

Li Ou¹, Xubin (Ben) He¹, Xin Chen¹, Christian Engelmann²,³, Stephen L. Scott²

¹ Tennessee Tech University, Cookeville, USA
² Oak Ridge National Laboratory, Oak Ridge, USA
³ The University of Reading, Reading, UK
Background and Motivation

- Large-scale HPC systems with 10,000-210,000 processors
- 100-480 TFlop/s LINPACK performance and 10-100TB RAM.
- Current systems: IBM Blue Gene/P and Cray XT4 - Top500.org
- Next-generation: Peta-scale IBM Blue Gene/P and Cray ??
- Applications: Climate change, nuclear astrophysics, fusion energy, materials sciences, biology, nanotechnology, …
- Single application runs for days, weeks, and even months

- Reliability, availability, and serviceability is crucial for success
- Head and service nodes are single points of failure and control
- Efficient redundancy strategies are needed for high availability
Active/Standby with Shared Storage

- Commonly employed technique
- Single active head node
- Backup to shared storage
- Simple checkpoint/restart
- Fail-over to standby node
- Possible corruption of backup state when failing during backup
- Introduction of a new single point of failure
- Correctness and availability are NOT ALWAYS guaranteed

 Metadata servers (MDS) of Parallel Virtual File System (PVFS) and Lustre (not part of standard install)
Symmetric Active/Active Redundancy

- Many active service nodes
- Work load distribution
- Symmetric replication between service nodes
- Continuous service
- Always up-to-date
- No fail-over, no restore-over
- State-machine replication
- Virtual synchrony model
- Complex algorithms (process group communication)
- JOSHUA prototype for Torque batch job queue and resource manager
- PVFS MDS (this paper)
Internal Replication Method

Output Unification

Virtually Synchronous Processing

Input Replication
Symmetric Active/Active Metadata Service Architecture

[Diagram showing the architecture of a symmetric active/active metadata service with client nodes, metadata servers (MDS 1, MDS 2, ..., MDS N), and data servers with file data.]
Symmetric Active/Active Metadata Service Design for the Parallel Virtual File System

**Metadata Server**
- Locking Service
- File Handle Management

**Membership Management**
- Transaction Control
- Basic Metadata Services
- Scheduler
- Request Interface

**Transis Group Communications**

**Client APIs:**
- create file, delete file, truncate file, create directory, delete directory, read directory, rename, set attributes
- read attributes, lookup path

<table>
<thead>
<tr>
<th></th>
<th>read</th>
<th>update</th>
<th>write</th>
</tr>
</thead>
<tbody>
<tr>
<td>read</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>update</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>write</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

MDS Record Locking Conflicts Resolved by Transaction Control
Symmetric Active/Active PVFS Metadata Service Request Handling

MDS Read Request Handling

1. Client requests metadata services.
2. Metadata Server processes request.

MDS Write Request Handling

1. Client requests file handle management.
2. Metadata Server processes request.
3. File Handle Management handles request.

Transaction Control

Request Interface

Transis Group Communications

MDS B

MDS A

Client

create file
delete file
create directory
delete directory
rename set attributes
Experimental Setup

• Transis v1.03 group communication system with fast delivery protocol (see ICCCN 2007 paper)
• Parallel Virtual File System (PVFS) v2
• ORNL XTORC cluster
  – Dual Intel Pentium 2GHz nodes
  – 768MB memory and 40 GB hard disk space per node
  – 100MBit/s Fast Ethernet (full duplex)
  – Linux Fedora Core 5 operating system
• MPI-based benchmark using multiple clients to send concurrent MDS read and write requests
Symmetric Active/Active PVFS Metadata Service Performance and Overhead

Throughput

100 Mbps Ethernet LAN Environment

PVFS
A/A 1
A/A 2
A/A 4

Number of Clients

Throughput (requests/sec)

0 20 40 60 80 100 120

1 2 4 8 16 32

Writing Throughput
Symmetric Active/Active PVFS Metadata Service Performance and Overhead

Reading Throughput

Throughput (requests/sec)

Number of Clients

100 Mbps Ethernet LAN Environment
Symmetric Active/Active PVFS Metadata Service Availability Improvement

- \( A_{\text{component}} = \frac{\text{MTTF}}{(\text{MTTF} + \text{MTTR})} \)
- \( A_{\text{system}} = 1 - (1 - A_{\text{component}})^n \)
- \( T_{\text{down}} = 8760 \text{ hours} \times (1 - A) \)
- Single node MTTF: 5000 hours
- Single node MTTR: 72 hours

<table>
<thead>
<tr>
<th>Nodes</th>
<th>Availability</th>
<th>Est. Annual Downtime</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>98.58%</td>
<td>5d 4h 21m</td>
</tr>
<tr>
<td>2</td>
<td>99.97%</td>
<td>1h 45m</td>
</tr>
<tr>
<td>3</td>
<td>99.9997%</td>
<td>1m 30s</td>
</tr>
<tr>
<td>4</td>
<td>99.999995%</td>
<td>1s</td>
</tr>
</tbody>
</table>

Single-site redundancy for 7 nines does not mask catastrophic events.
Summary and Future Work

- Developed a symmetric active/active metadata service (MDS) for the Parallel Virtual File System (PVFS)
- Solution ensures no loss of service and no loss of state
- Minimal performance impact for MDS write requests
- Significant performance gain for MDS read requests
- Significant availability improvement for MDS
- Employed concepts are applicable to any networked file system that utilizes a MDS

- Ongoing work focuses on more complex scenarios in service-oriented architectures (SOAs), such as dependencies between multiple services
Symmetric Active/Active Metadata Service for Highly Available Cluster Storage Systems

Li Ou¹, Xubin (Ben) He¹, Xin Chen¹, Christian Engelmann²,³, Stephen L. Scott²

¹ Tennessee Tech University, Cookeville, USA
² Oak Ridge National Laboratory, Oak Ridge, USA
³ The University of Reading, Reading, UK