# High Availability for the Lustre File System


OAK RIDGE NATIONAL LABORATORY

MANAGED BY UT-BATTELLE FOR THE DEPARTMENT OF ENERGY

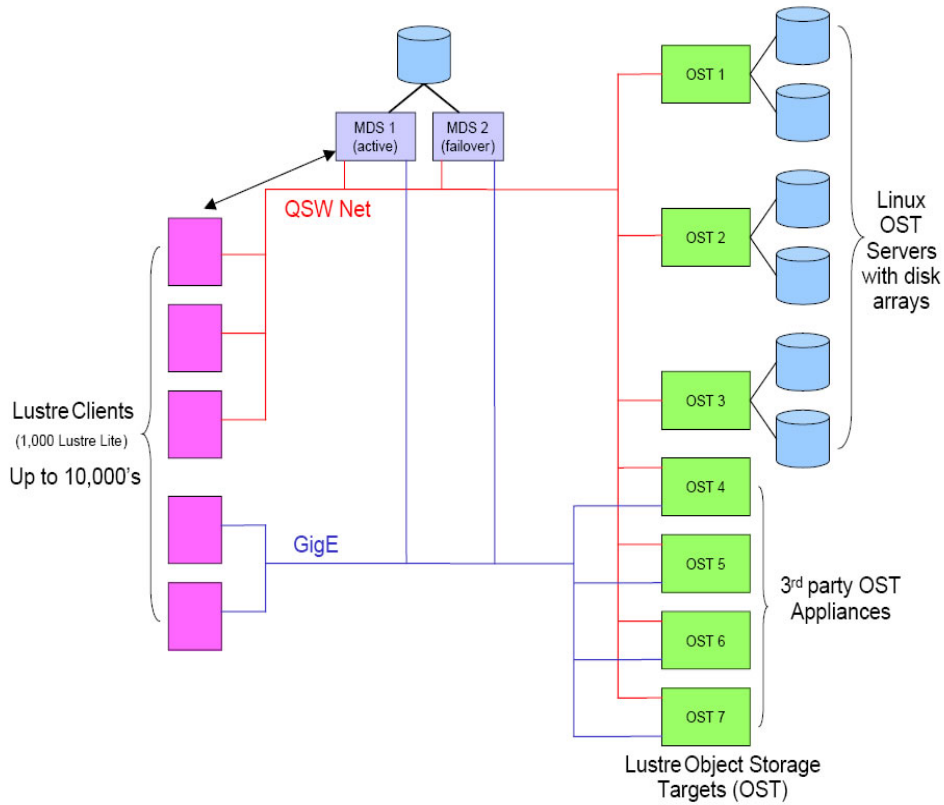
# Overview

- Introduction
  - Lustre
  - Symmetric Active/Active Replication
  - Preceding Projects
- Development
  - Replication Method
  - Preliminary System Design
  - Lustre Networking
  - Final Prototype Design
  - Prototype Implementation Limitations
- System Tests
  - Functionality
  - Performance
- Summary

# Introduction
## *Lustre*

*"Lustre is a scalable, secure, robust, highly-available cluster file system."*

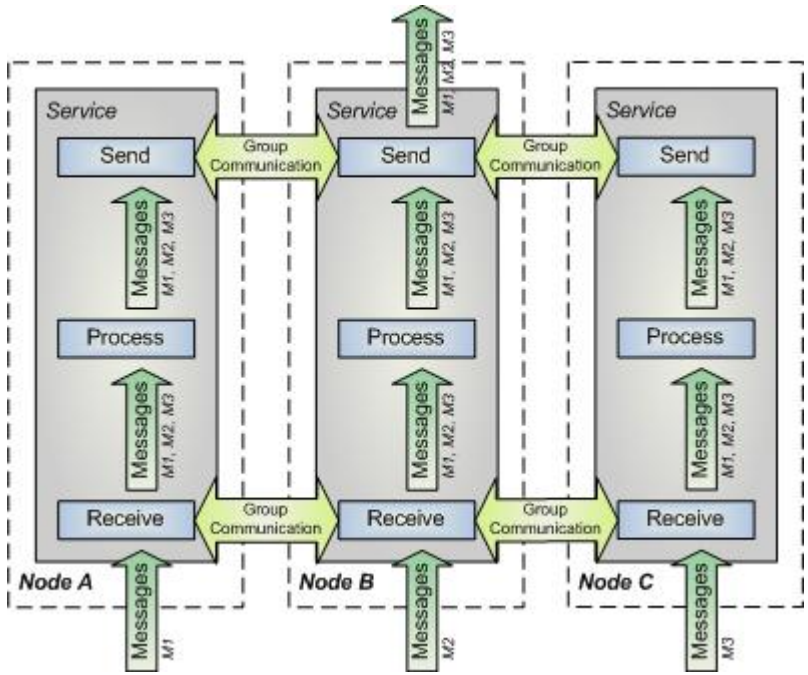
- file data is physically stored on the OSTs using a RAID pattern
  - already redundancy √
- file metadata is stored on a single MDS disk
  - Single Point of Failure!



# Introduction
## *Symmetric Active/Active Replication*

- replication of the MDS on several nodes
- group communication system Transis ensures total message order
- state is never lost as long as one node is up

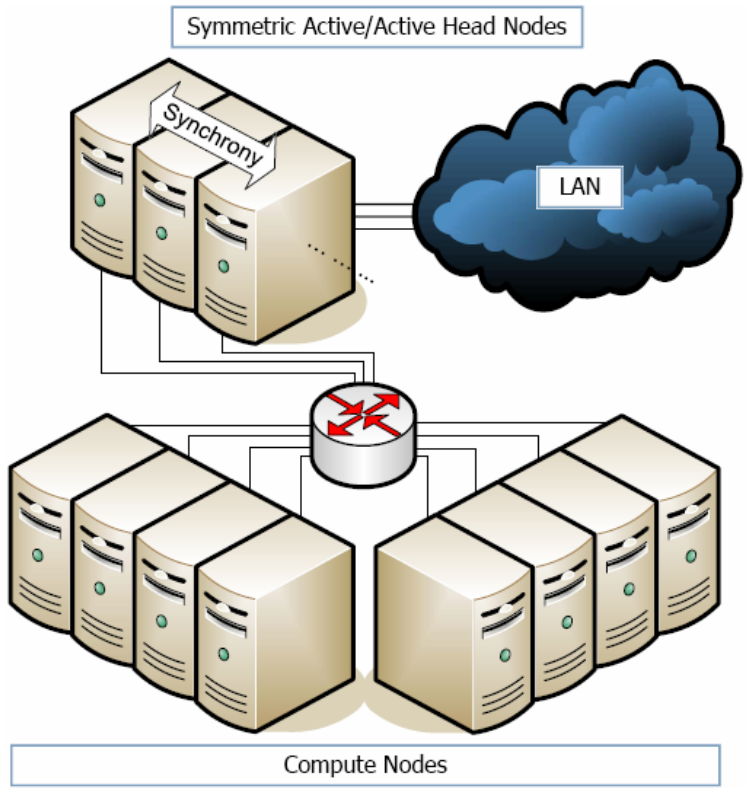
➢ Elimination of Single Point of Failure




*symmetric active/active architecture*

# Introduction
## *Preceding Projects*

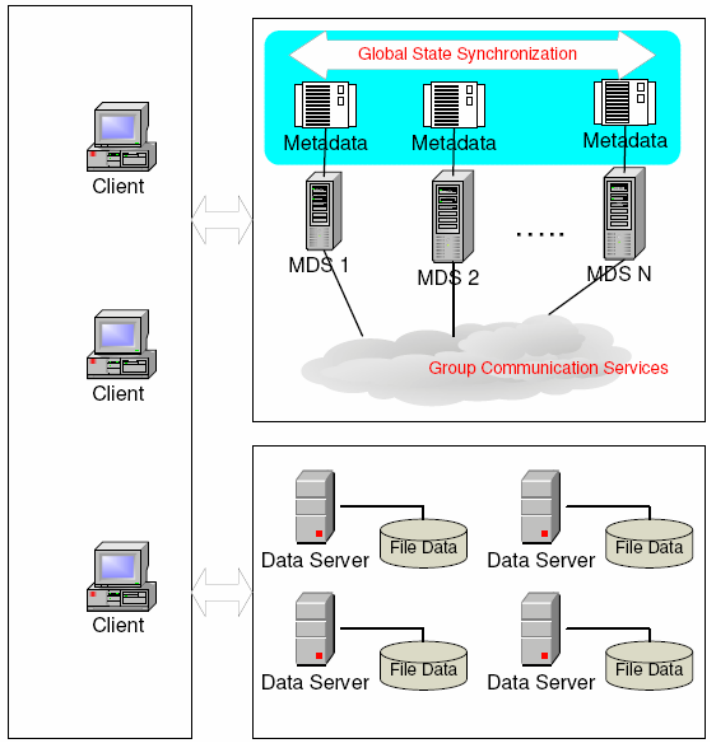
The **JOSHUA** project offers symmetric active/active HA for HPC job and resource management services. It represents a virtually synchronous environment using external replication providing HA without any interruption of service and without any loss of state.




Advanced Beowulf Cluster Architecture with Symmetric Active/Active High Availability for Head Node System Services

# Introduction
## *Preceding Projects*

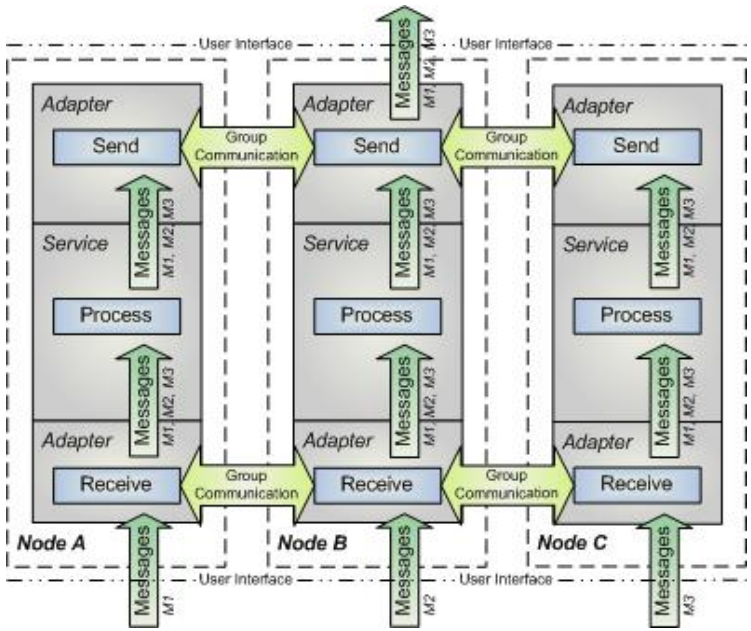
The **Metadata Service for Highly Available Cluster Storage Systems** project targets the symmetric active/active replication model using multiple redundant service nodes running in virtual synchrony.




Active/Active Metadata Servers in a Distributed Storage System

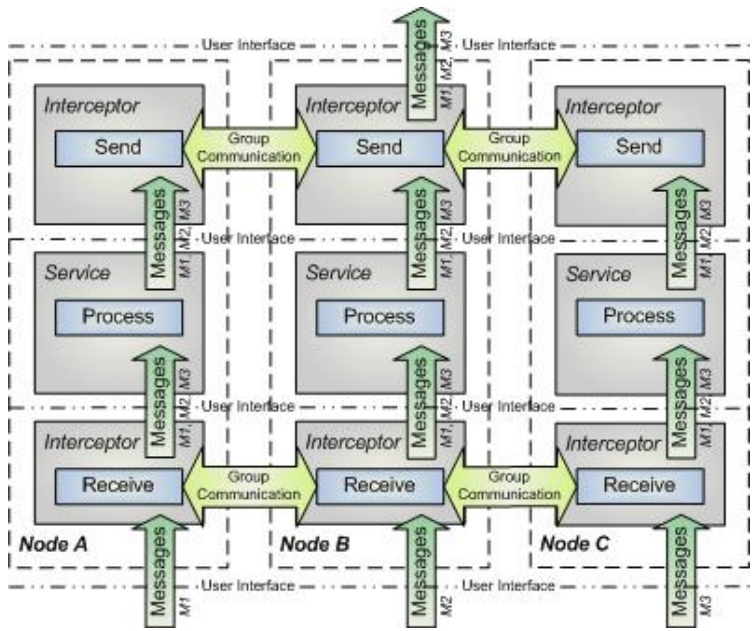
# Development
## *Replication Method*

internal




- faster, due to no inter-process communication
- kernel development

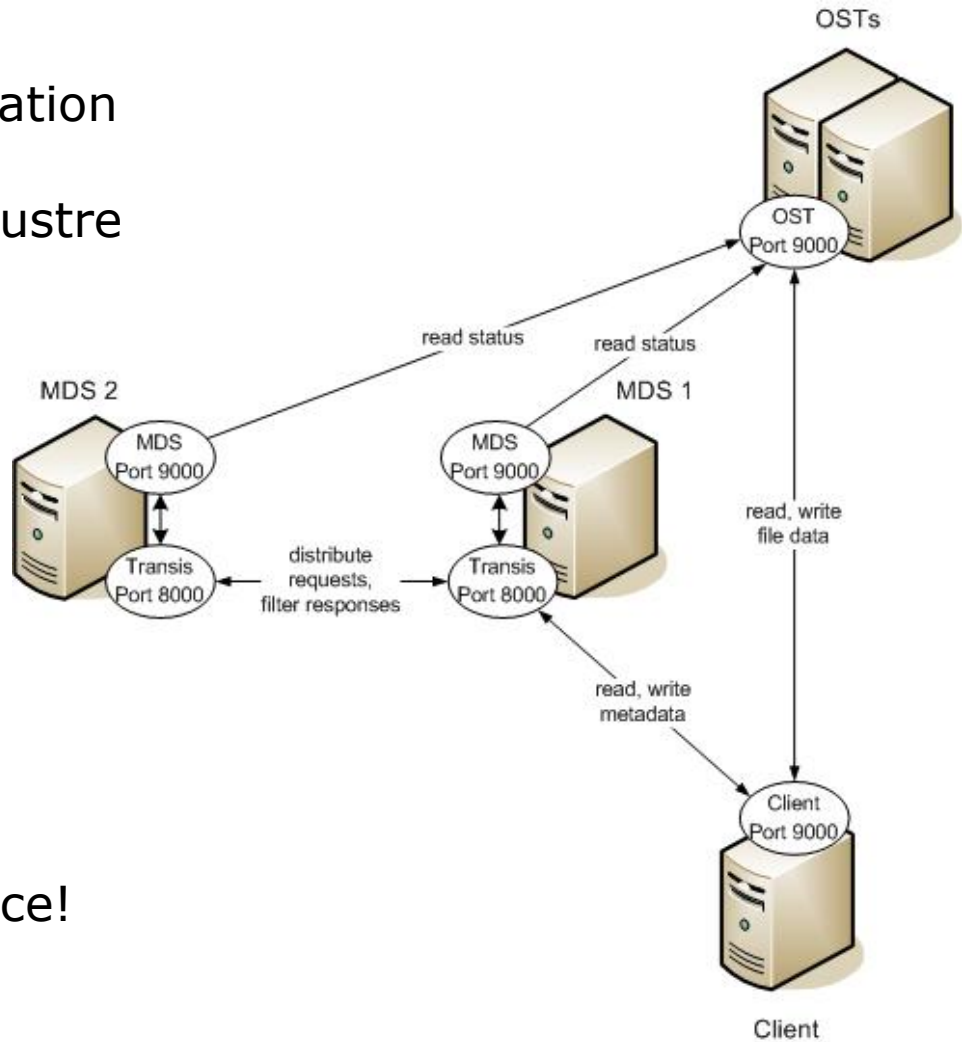
external




- no need to touch Lustre code
- modular design

# Development
## *Preliminary System Design*

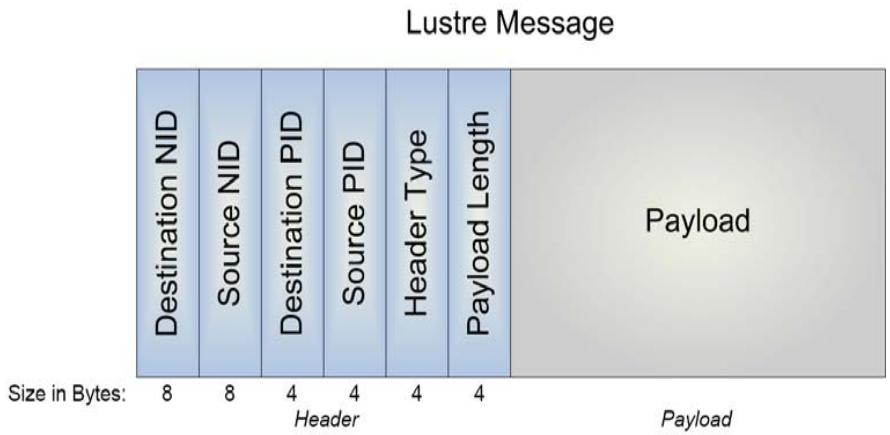
- uses external replication method
- Transis intercepts Lustre messages




Doesn‘t work in practice!

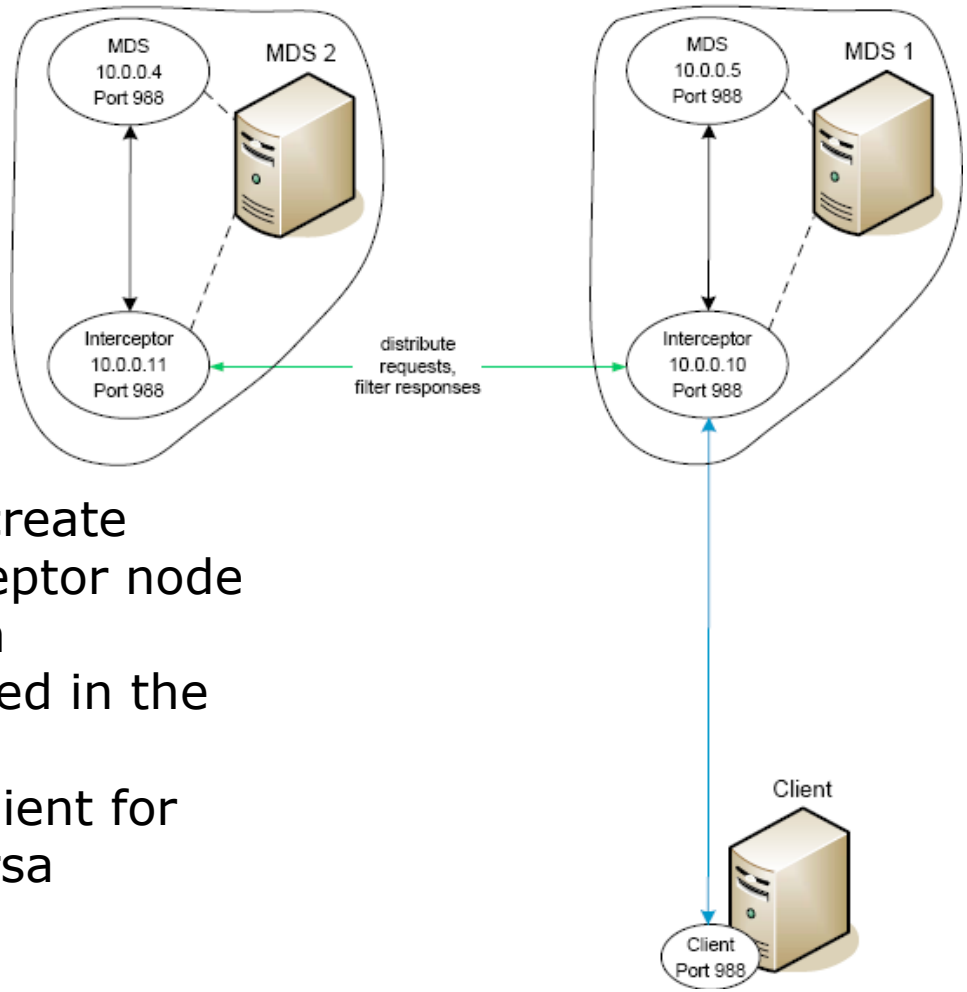
# Development
## *Lustre Networking*

- Lustre needs all components running on port 988
- message source and destination are checked
- Lustre doesn't allow rerouting
  - only direct sent messages are accepted



# Development
## *Final Prototype Design*

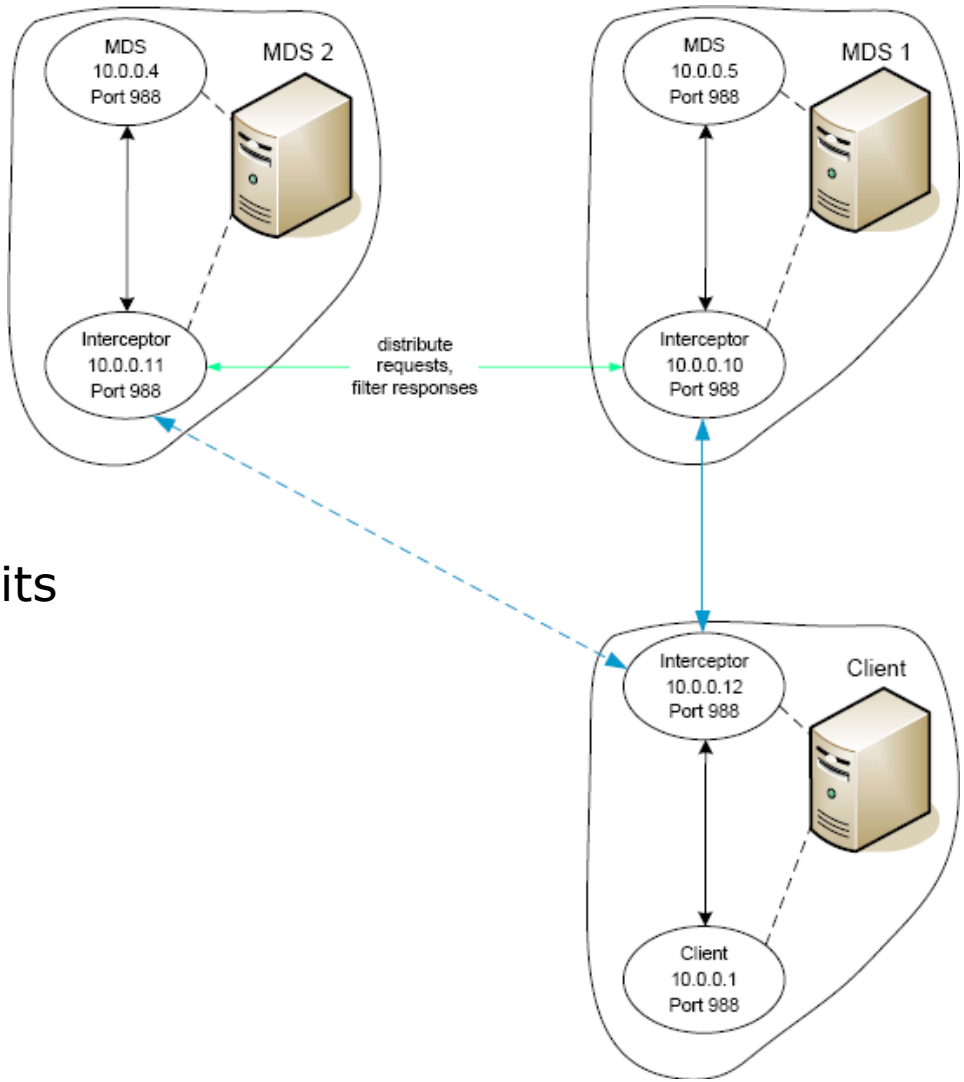
*Prototype 1*




- use of IP aliasing to create second virtual Interceptor node
- group communication functionality is included in the Interceptor
- Interceptor acts as Client for the MDS and vice versa

# Development
## *Final Prototype Design*

*Prototype 2*




- each Client node has its own Interceptor
- Client-Interceptor is capable of failover to another MDS node in case of error

# Development
## *Prototype Implementation Limitations*

The design of Lustre doesn't allow implementation of all prototype features.

***Distributed locking mechanisms within Lustre***

- Each MDS tries to get the same lock

***Existing active/standby failover behaviour of the MDS***

- only one running MDS allowed at a given time
- only two MDS can be configured

***Only three connections per node allowed***

- one client uses three connections
- all clients are routed through one interceptor

# System Tests
## *Functionality*

Due to restrictions caused by the Lustre design, functionality tests could only performed part wise.

***Working functionality of the prototypes:***

- Message Routing
- Group Communication System
- File Operations: read, write, create, delete

***Missing functionality of the prototypes:***

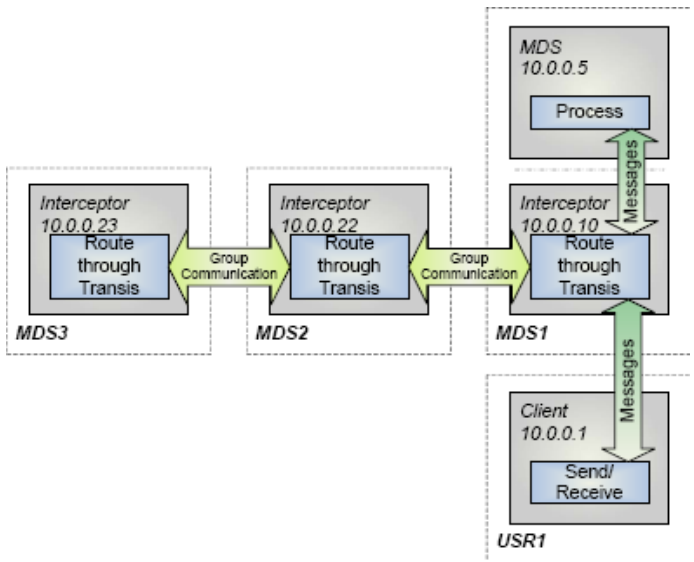
- multiple running MDS at the same time
- connection failover

*The results give proof of working components, but an active/active HA solution of Lustre could not be tested.*

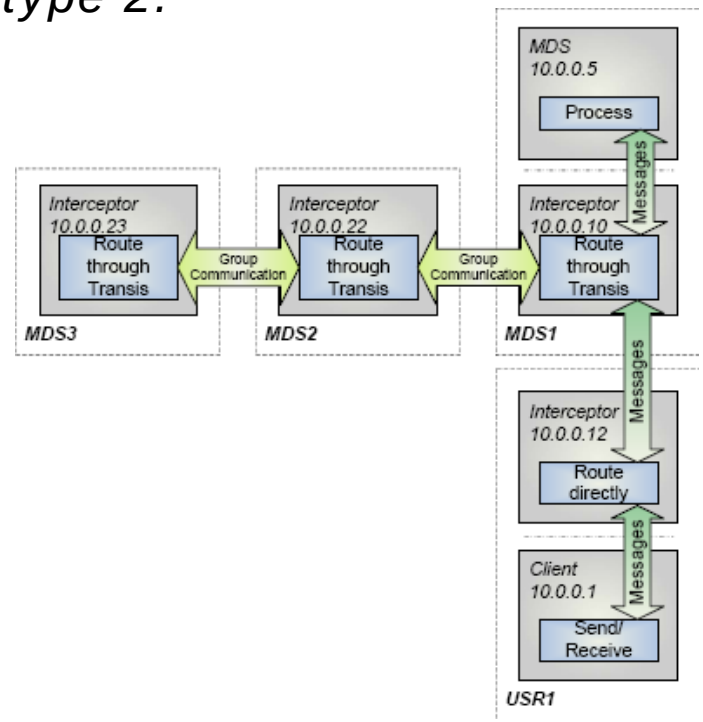
# System Tests
## *Performance*

Implemented prototypes consist of all components a working HA solution needs.

*A full working HA prototype would have almost the same impact on performance like Prototype 2.*




Test Setup Prototype 1




Test Setup Prototype 2

# System Tests *Performance*

Performance tested with own benchmark program.

All performance tests have been done with 100MBit and 1GBit network.

*The standard Lustre performs up to 89 times faster than the tested prototypes.*

| Operations per second | 100 Mbit | | | | | |
|---|---|---|---|---|---|---|
| | 1 file | | | 100 files | | |
| | create | read | delete | create | read | delete |
| Standard Lustre | 538.199 | 462.389 | 1,129.114 | 551.799 | 452.459 | 1,721.659 |
| Prototype 1, 1 GM | 6.103 | 11.840 | 12.165 | 8.030 | 8.052 | 24.064 |
| Prototype 1, 2 GM | 6.104 | 11.846 | 12.170 | 8.026 | 8.060 | 23.775 |
| Prototype 1, 3 GM | 6.108 | 11.844 | 12.165 | 8.025 | 8.062 | 23.964 |
| Prototype 2, 1 GM | 6.056 | 11.758 | 12.094 | 7.966 | 8.051 | 23.895 |
| Prototype 2, 2 GM | 6.051 | 11.732 | 12.047 | 7.964 | 8.045 | 23.889 |
| Prototype 2, 3 GM | 6.037 | 11.782 | 12.092 | 7.918 | 8.046 | 23.894 |

| Time taken for one operation (msec) | 100 Mbit | | | | | |
|---|---|---|---|---|---|---|
| | 1 file | | | 100 files | | |
| | create | read | delete | create | read | delete |
| Standard Lustre | 1.858 | 2.163 | 0.886 | 1.812 | 2.210 | 0.581 |
| Prototype 1, 1 GM | 163.859 | 84.463 | 82.202 | 124.538 | 124.186 | 41.557 |
| Prototype 1, 2 GM | 163.827 | 84.418 | 82.172 | 124.593 | 124.074 | 42.061 |
| Prototype 1, 3 GM | 163.707 | 84.433 | 82.202 | 124.607 | 124.041 | 41.729 |
| Prototype 2, 1 GM | 165.125 | 85.050 | 82.686 | 125.529 | 124.211 | 41.849 |
| Prototype 2, 2 GM | 165.248 | 85.240 | 83.009 | 125.558 | 124.299 | 41.860 |
| Prototype 2, 3 GM | 165.647 | 84.874 | 82.698 | 126.298 | 124.290 | 41.852 |

| Operations per second | 1 Gbit | | | | | |
|---|---|---|---|---|---|---|
| | 1 file | | | 100 files | | |
| | create | read | delete | create | read | delete |
| Standard Lustre | 622.247 | 550.658 | 1,330.973 | 636.749 | 520.497 | 1,951.101 |
| Prototype 1, 1 GM | 6.181 | 12.710 | 12.314 | 8.157 | 8.252 | 24.359 |
| Prototype 1, 2 GM | 6.140 | 12.038 | 12.221 | 8.082 | 8.179 | 24.238 |
| Prototype 1, 3 GM | 6.128 | 11.939 | 12.207 | 8.067 | 8.138 | 24.209 |
| Prototype 2, 1 GM | 6.138 | 12.144 | 12.248 | 8.106 | 8.217 | 24.224 |
| Prototype 2, 2 GM | 6.091 | 11.926 | 12.156 | 8.023 | 8.134 | 24.037 |
| Prototype 2, 3 GM | 6.086 | 11.900 | 12.142 | 8.010 | 8.125 | 24.021 |

| Time taken for one operation (msec) | 1 Gbit | | | | | |
|---|---|---|---|---|---|---|
| | 1 file | | | 100 files | | |
| | create | read | delete | create | read | delete |
| Standard Lustre | 1.607 | 1.816 | 0.751 | 1.570 | 1.921 | 0.513 |
| Prototype 1, 1 GM | 161.786 | 78.680 | 81.211 | 122.598 | 121.184 | 41.052 |
| Prototype 1, 2 GM | 162.871 | 83.071 | 81.825 | 123.734 | 122.269 | 41.257 |
| Prototype 1, 3 GM | 163.193 | 83.762 | 81.919 | 123.964 | 122.882 | 41.308 |
| Prototype 2, 1 GM | 162.920 | 82.348 | 81.649 | 123.364 | 121.696 | 41.282 |
| Prototype 2, 2 GM | 164.165 | 83.850 | 82.263 | 124.646 | 122.937 | 41.602 |
| Prototype 2, 3 GM | 164.310 | 84.033 | 82.359 | 124.840 | 123.078 | 41.630 |

# System Tests
## *Performance*

***Overhead to the system:***

*JOSHUA*
256% with four group members

*Prototype 2*
8815%! with three group members

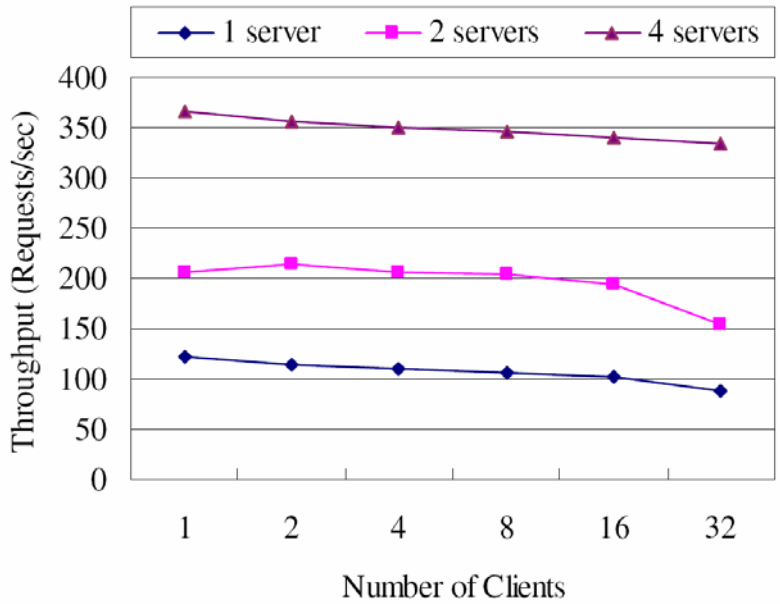
***Read request throughput:***

*Metadata Service project*
125 with one server
360 with four servers

*Prototype 2*
520 with standard Lustre
8 with three group members

| System | # | Latency | Overhead |
|---|---|---|---|
| TORQUE | 1 | 98 ms | |
| JOSHUA/TORQUE | 1 | 134 ms | 36 ms / 37% |
| JOSHUA/TORQUE | 2 | 265 ms | 158 ms / 161% |
| JOSHUA/TORQUE | 3 | 304 ms | 206 ms / 210% |
| JOSHUA/TORQUE | 4 | 349 ms | 251 ms / 256% |

Job Submission Latency Comparison of Single vs. Multiple Head Node HPC Job and Resource Management




Read Request Throughput Comparison of Single vs. Multiple Metadata Servers

# Summary

- A working symmetric active/active HA solution for the MDS of Lustre cannot be provided within this project
- Significant performance impact of proposed HA solution
- Further analyses of Lustre needed
- Full working production type HA prototype requires changes in the entire Lustre design

# High Availability for the Lustre File System

Thank you for your attention.


OAK RIDGE NATIONAL LABORATORY

MANAGED BY UT-BATTELLE FOR THE DEPARTMENT OF ENERGY