TWG Mission

- Work with the Lustre community to ensure that Lustre continues to support the stability, performance, and management requirements of the OpenSFS members as HPC compute platforms continue to scale
- Responsible for creating and managing the roadmap for the OpenSFS community
  - Gather requirements from the Lustre HPC community,
  - Prioritize and recommend development projects to the Board,
  - Initiate RFPs for important features, and
  - Work with contractors to meet these requirements
Who is the TWG?

-The following have attended TWG meetings and/or contributed content to our requirements:

- Bull/EOFS
- Cray
- DDN
- Fujitsu
- Indiana University
- LBL
- LLNL
- NRL
- ORNL
- RAID, Inc.
- Whamcloud
- Xyratex
Importance of Community

- Community contribution is crucial
  - Fujitsu
  - EOFS
- Broadening the scope of requirements
- No monopoly on good ideas
- Avoiding duplicate effort
Process History

- Developed feature-based proposals
  - Presented to Board in January 2011
  - Rejected in favor of requirement-based approach
- Gathered and prioritized requirements
  - Reached out to broader community members for requirements
  - Prioritized by consensus
Prioritized Requirements

- Near-term requirements
  - Metadata server performance
  - Metadata server scaling

- Long-term requirements
  - Support alternate storage backends
  - Scalable fault management
  - Start investigations of alternate storage backends

- Improve the code base
  - Reduce maintenance effort
  - Reduce cost of new features
Process History

‣ Presented requirements and recommendations to the Board March 2011
  ◦ http://goo.gl/cZSWG+
  ◦ Board accepted our recommendations

‣ Developed RFPs for top two priorities
  • RFPs open to the public April 7, 2011
    • Metadata Performance and Scalability
    • Space Quota Accounting and Enforcement
  • http://www.opensfsfs.org/?page_id=149
Roadmap Caveats

- OpenSFS doesn't have direct control
  - Development performed by contractors or members
  - Clearinghouse for requirements
  - Host community architecture discussions
- RFPs open
  - Expect some traditional ideas to be proposed
  - Encourage novel ideas
  - Tough to predict exact roadmap!
## Scaling Requirements

<table>
<thead>
<tr>
<th>Metric</th>
<th>Lustre 2.0/2.1</th>
<th>Q2 2012</th>
<th>Q1 2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>maximum number of files in file system</td>
<td>4 billion</td>
<td>100 billion</td>
<td>1 trillion</td>
</tr>
<tr>
<td>maximum number of files in directory</td>
<td>10 million</td>
<td>50 million</td>
<td>10 billion</td>
</tr>
<tr>
<td>maximum number of subdirectories</td>
<td>10 million</td>
<td>1 million</td>
<td>10 million</td>
</tr>
<tr>
<td>maximum number of clients</td>
<td>131072</td>
<td>64 thousand</td>
<td>128 thousand</td>
</tr>
<tr>
<td>maximum number of OSS nodes</td>
<td>-</td>
<td>1 thousand</td>
<td>4 thousand</td>
</tr>
<tr>
<td>maximum number of OSTs</td>
<td>8150</td>
<td>2 thousand</td>
<td>8 thousand</td>
</tr>
<tr>
<td>maximum OST size</td>
<td>16 TB</td>
<td>32 TB</td>
<td>128 TB</td>
</tr>
<tr>
<td>maximum file system size</td>
<td>64 PB</td>
<td>100 PB</td>
<td>256 PB</td>
</tr>
<tr>
<td>maximum file size</td>
<td>320 TB</td>
<td>1 PB</td>
<td>-</td>
</tr>
<tr>
<td>maximum object size</td>
<td>2 TB</td>
<td>16 TB</td>
<td>64 TB</td>
</tr>
<tr>
<td>peak aggregate file creates/s</td>
<td>-</td>
<td>200 thousand</td>
<td>400 thousand</td>
</tr>
<tr>
<td>peak directory listings/s (ls -l)</td>
<td>-</td>
<td>-</td>
<td>100 thousand</td>
</tr>
<tr>
<td>maximum single client open files</td>
<td>~3 thousand</td>
<td>100 thousand</td>
<td>-</td>
</tr>
<tr>
<td>peak single client file creates/s</td>
<td>-</td>
<td>30 thousand</td>
<td>-</td>
</tr>
</tbody>
</table>
Metadata Server Performance

- Vertical Scaling
  - LNET scaling
  - RPC/MDS operation scaling
  - Size-on-MDS
  - Other novel ideas proposed by respondents

- Horizontal Scaling
  - Phase 1 – distributed namespace
  - Phase 2 – striped directories

- Long-term
  - Rework service model
  - Network Request Scheduler
Alternate Storage Backends

- Ldiskfs is nearing the end of its useful life
  - Requires external assistance for redundancy
  - Increasing disk capacities require larger LUNs for efficiency
  - No checksum of data
  - No online filesystem consistency check
Alternate Storage Backends

- Refactor obdfilter to allow new backends
  - Object Storage Device interface work
  - Partially funded by LLNL

- Requires work on quota system
  - Currently intimate with details of ldiskfs quotas
  - Lustre quotas need to be independent of backend
  - RFP out for this work
Storage Backends

- Top contenders
  - Ldiskfs
  - ZFS
  - BTRFS
  - Another upstart?

- OSD work to support all of these
Other Requirements

‣ Reliability
  • Fault detection, recovery, reporting

‣ Layout improvements
  ◦ Allow layouts to adapt as the file grows and/or ages
  ◦ Dynamic storage balancing
  ◦ Snapshots/replication

‣ Environmental
  ◦ Patchless server/support for newer distros
  ◦ Support mixed endianness and page sizes
  ◦ Improved configuration
  ◦ Ipv6 support
Summary

‣ Full link to requirements document:

‣ TWG Archives
http://lists.opensfs.org/pipermail/twg-opensfs.org/

‣ Join us at discuss@lists.opensfs.org

dillowda@ornl.gov
carrier@cray.com