(A Bowl full of) Lustre Tools

LAD 2013 – Paris

Zellescher Weg 12
WIL A 208
Tel. +49 351 - 463 – 34217

Michael Kluge (michael.kluge@tu-dresden.de)
Content

- About this talk
- Systematic overview
- Tools, tools, tools, …
- Wrap up
Problem Statement

- Lots of presentations about tools at LUG
- Lots of sites use tools
- Lots of homebrews
- No tool page (community task)
- Some very useful stuff already vanished …
Systematic Tool Overview

Options for categorizing tools:

- Target audience/area
- Ease of use
- Availability
- Maintenance status
- License
- Owner
Target Audience/Area

- Administrators
  - Setup/Management
  - Monitoring
  - Maintenance tasks

- Users
  - Daily work

- System Architects
  - Benchmarking
  - Performance analysis
Content

- About this talk
- Systematic overview
- **Tools, tools, tools, ...**
- Wrap up
Tools for Administrators

- Setup/Management
- Monitoring
- Maintenance tasks
Tools for Administrators (Setup/Management)

- Complete management solutions
- Commercially available from many vendors: Intel

- Hadoop Adapter
  Lustre storage for MapReduce applications

- Intel® Manager for Lustre* Software
  Configure, Monitor, Troubleshoot, Manage

- REST API
  Extensibility

- Management and Monitoring Services

- Lustre File System
  Full distribution of open source Lustre software

- Storage Plug-in
  Integration

Intel value-added software  Open source software

http://lustre.intel.com
Tools for Administrators (Setup/Management)

- Intel® Manager for Lustre

http://lustre.intel.com
Tools for Administrators (Setup/Management)

- Complete management solutions
- Commercially available from many vendors: Terascale

http://www.terascala.com
Tools for Administrators (Setup/Management)

- Complete management solutions
- Commercially available from many vendors: Xyratex

http://www.xyratex.com
Tools for Administrators (Setup/Management)

ClusterStor Manager

http://www.xyratex.com
Tools for Administrators (Setup/Management)

- Shine: open source
  - Uses model file that describes the setup
  - LAD’12 presentation ([www.eofs.org](http://www.eofs.org))
  - Remote administration via ssh
    - shine install/format/start/mount ...
  - status view
  - management of file system tunings
  - parallel execution of commands


Tools for Administrators (Monitoring)

- Monitoring is a big deal
  - **Usage/Quota, Health**
  - Lustre Log Files
  - Performance

- ne2scan
- Nagios plugins (Lustre health, multipath, controllers, …)
- Shine
- Robinhood
Tools for Administrators (Monitoring)

- Robinhood: open source
- Accounting and monitoring
  - fast "du" and "find"
- Policy engine
- Alerts
- up to date (consumes changelogs)
- Web interface

http://robinhood.sf.net/
Tools for Administrators (Monitoring)

**Space used per group**

<table>
<thead>
<tr>
<th>Group</th>
<th>Space used</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>544.09 TB</td>
<td>12,708,782</td>
</tr>
<tr>
<td></td>
<td>302.19 TB</td>
<td>2,743,617</td>
</tr>
<tr>
<td></td>
<td>166.26 TB</td>
<td>3,370,537</td>
</tr>
<tr>
<td></td>
<td>150.26 TB</td>
<td>1,995,581</td>
</tr>
<tr>
<td></td>
<td>82.22 TB</td>
<td>1,393,682</td>
</tr>
<tr>
<td></td>
<td>79.94 TB</td>
<td>7,465,108</td>
</tr>
<tr>
<td></td>
<td>73.46 TB</td>
<td>522,402</td>
</tr>
</tbody>
</table>

View all groups
Tools for Administrators (Monitoring)

- Monitoring is a big deal
  - Usage/Quota, Health
  - **Lustre Logs**
  - Performance

- Syslog ++
- Event correlation (ORNL, SEC, Splunk)
Tools for Administrators (Monitoring)

- Monitoring is a big deal
  - Usage/Quota
  - Lustre Log Files
  - **Performance**

- xltop, lmt, Tacc-stat
- DDNtool
Tools for Administrators (Monitoring)

- xltop: daemons on all servers and compute nodes
- master combines data into different views

```
<table>
<thead>
<tr>
<th>FILESYSTEM</th>
<th>MDS/T</th>
<th>LOAD1</th>
<th>LOAD5</th>
<th>LOAD15</th>
<th>TASKS</th>
<th>OSS/T</th>
<th>LOAD1</th>
<th>LOAD5</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>taurus-scratch</td>
<td>1/1</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
<td>900</td>
<td>4/48</td>
<td>133.22</td>
<td>70.07</td>
<td></td>
</tr>
</tbody>
</table>
```

Michael Kluge
Tools for Administrators (Monitoring)

- lmt: daemons on all servers
- MySQL database with historic data

https://github.com/chaos/lmt/wiki
https://computing.llnl.gov/linux/cerebro.html
Tools for Administrators (Maintenance Tasks)

- Common things to do:
  - Archive and backup
  - move the whole file system around
  - Purge old files
  - Export file systems
  - fsck?

- Robinhood

- Quit a few home grown parallel copy tools

- NFS Ganesha
Tools for Users

- Administrators
  - Setup
  - Monitoring
  - Maintenance tasks

- Users
  - Daily work

- System Architects
  - Benchmarking
  - Performance analysis

http://www.glasbergen.com
Main areas of interest:
- What is my code doing
- What is the fastest way to get my stuff done
- Programming support?

No tool to report Lustre specific job characteristics
- Request sizes, time spent doing I/O
- IOTA @ ORNL?

Quite a few tools to handle common tasks
Tools for Users

- Common tasks
  - Copying, tar/untar, moving files …
  - Most important: parallelism, stripes and extended attributes

- Parallel unix file system tools
  - lustre_rsync
  - pcp
  - mutil, retools
  - spdcp?, pltar?
  - some generic tools: dcp, bbcp, mtcp
Tools for System Architects

- Administrators
  - Setup
  - Monitoring
  - Maintenance tasks

- Users
  - Daily work

- System Architects
  - Benchmarking
  - Performance analysis
Benchmarks working group at ORNL
- Covers everything from block I/O to metadata
- Application kernels

Some tools shipped with the Lustre I/O kit
Tools for System Architects (Performance Analysis)

- Lustre MDS trace (ORNL)
  - Collect sample of MDS RPCs and calculates properties
- RPC visualization (ZIH)
  - see all RPCs floating around between the different machines
- Wireshark patches (Intel)
  - debug Lustre at the wire level
- SystemTap scripts
  - Low overhead, very flexible
  - Probes can be attached to every function
Content

- About this talk
- Systematic overview
- Tools, tools, tools, ...
- Wrap up
# Admin Tools Overview (1)

<table>
<thead>
<tr>
<th>Tool</th>
<th>Owner/Author</th>
<th>Last Update/Status</th>
<th>Status</th>
<th>Description</th>
<th>Lizenz</th>
</tr>
</thead>
<tbody>
<tr>
<td>home grown parallel copy tool</td>
<td>many (at least 3)</td>
<td>LUG 2013</td>
<td>unavailable</td>
<td>presentation from Marc Stearman</td>
<td>not described</td>
</tr>
<tr>
<td>Shine (uses ClusterShell)</td>
<td>CEA</td>
<td>up to date</td>
<td>alive</td>
<td>setup and management, including HA</td>
<td>GPL</td>
</tr>
<tr>
<td>NFS Ganesha</td>
<td>CEA</td>
<td>up to date</td>
<td>alive</td>
<td>Lustre aware user space NFS server</td>
<td>GPL</td>
</tr>
<tr>
<td>DDNTool</td>
<td>ORNL</td>
<td>2010</td>
<td>unavailable</td>
<td>continuous monitoring of controller data</td>
<td>not described</td>
</tr>
<tr>
<td>lltop/xltop</td>
<td>TACC</td>
<td>2011</td>
<td>probably alive</td>
<td>continuous monitoring of /proc data from servers and clients</td>
<td>GPL</td>
</tr>
<tr>
<td>lmt</td>
<td>LLNL</td>
<td>2011</td>
<td>probably alive</td>
<td>continuous monitoring of /proc data from servers (and history)</td>
<td>GPL</td>
</tr>
<tr>
<td>Lustre RPC trace</td>
<td>Intel HPDD</td>
<td>up to date</td>
<td>alive</td>
<td>records all RPCs</td>
<td>GPL</td>
</tr>
<tr>
<td>MDS RPC Trace</td>
<td>ORNL</td>
<td>LUG 2010</td>
<td>unavailable</td>
<td>short term monitoring of all RPCs on the server</td>
<td>not described</td>
</tr>
<tr>
<td>Nagios extensions</td>
<td>Many</td>
<td>up to date</td>
<td>alive</td>
<td>Admin Support, Management</td>
<td>not described</td>
</tr>
<tr>
<td>routerstat</td>
<td>built in</td>
<td>up to date</td>
<td>alive</td>
<td>prints LNET router statistics</td>
<td>GPL</td>
</tr>
<tr>
<td>Robinhood</td>
<td>CEA</td>
<td>up to date</td>
<td>alive</td>
<td>usage monitor, purging, HSM</td>
<td>GPL</td>
</tr>
</tbody>
</table>
## Admin Tools Overview (2)

<table>
<thead>
<tr>
<th>Tool</th>
<th>Owner/Author</th>
<th>Last Update/Status</th>
<th>Status</th>
<th>Description</th>
<th>Lizenz</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intel Manager for Lustre</td>
<td>Intel HPDD</td>
<td>up to date</td>
<td>alive</td>
<td>complete stack, setup and management, including HA and monitoring, has plugins for external metrics from vendors</td>
<td>Proprietary</td>
</tr>
<tr>
<td>Terascala LustreStack</td>
<td>TeraScala</td>
<td>up to date</td>
<td>alive</td>
<td>complete vendor stack</td>
<td>Proprietary</td>
</tr>
<tr>
<td>Xyratex ClusterStore</td>
<td>Xyratex</td>
<td>up to date</td>
<td>alive</td>
<td>complete vendor stack</td>
<td>Proprietary</td>
</tr>
</tbody>
</table>
## User Tools Overview

<table>
<thead>
<tr>
<th>Tool</th>
<th>Owner/Author</th>
<th>Last Update/Status</th>
<th>Status</th>
<th>Description</th>
<th>Lizenz</th>
</tr>
</thead>
<tbody>
<tr>
<td>home grown parallel copy tools</td>
<td>many (at least 3)</td>
<td>LUG 2013</td>
<td>unavailable</td>
<td>presentation from Marc Stearman</td>
<td>not described</td>
</tr>
<tr>
<td>lustre_rsync</td>
<td>built in</td>
<td>up to date</td>
<td>alive</td>
<td>copies a whole file system to another place</td>
<td>GPL</td>
</tr>
<tr>
<td>Intel HPDD Hadoop</td>
<td>Intel HPDD</td>
<td>up to date</td>
<td>alive</td>
<td>Hadoop with Lustre backend</td>
<td>Proprietary</td>
</tr>
<tr>
<td>OLCF pltar</td>
<td>ORNL</td>
<td>2010</td>
<td>unavailable</td>
<td>parallel tar tool</td>
<td>?</td>
</tr>
<tr>
<td>OLCF spdcp</td>
<td>ORNL</td>
<td>2008</td>
<td>unavailable</td>
<td>parallel copy tool</td>
<td>not described</td>
</tr>
<tr>
<td>parallel copy tools (mutil) mcp, msum</td>
<td>NASA</td>
<td>2012</td>
<td>alive</td>
<td>parallel copy tool</td>
<td>GPL</td>
</tr>
<tr>
<td>parallel copy tools (mutil) mtar, m<em>zip</em>, mrsync</td>
<td>NASA</td>
<td>2013</td>
<td>waiting for release</td>
<td>parallel copy tools</td>
<td>not described</td>
</tr>
<tr>
<td>pcp</td>
<td>Guy Coates</td>
<td>up to date</td>
<td>alive</td>
<td>parallel copy tool</td>
<td>GPL</td>
</tr>
<tr>
<td>Robinhood</td>
<td>CEA</td>
<td>up to date</td>
<td>alive</td>
<td>usage monitor, purging, HSM</td>
<td>GPL</td>
</tr>
</tbody>
</table>
## Performance Tools Overview

<table>
<thead>
<tr>
<th>Tool</th>
<th>Owner/Author</th>
<th>Last Update/Status</th>
<th>Status</th>
<th>Description</th>
<th>Licence</th>
</tr>
</thead>
<tbody>
<tr>
<td>lustre-iokit</td>
<td>Intel HPDD</td>
<td>up to date</td>
<td>alive</td>
<td>tools for benchmarking Lustre systems</td>
<td>GPL</td>
</tr>
<tr>
<td>Lustre RPC trace</td>
<td>Intel HPDD</td>
<td>up to date</td>
<td>alive</td>
<td>records all RPCs</td>
<td>GPL</td>
</tr>
<tr>
<td>MDS RPC Trace</td>
<td>ORNL</td>
<td>LUG 2010</td>
<td>unavailable</td>
<td>Generate report from short term monitoring of all RPCs on the server</td>
<td>not described</td>
</tr>
<tr>
<td>routerstat</td>
<td>built in</td>
<td>up to date</td>
<td>alive</td>
<td>prints LNET router statistics</td>
<td>GPL</td>
</tr>
<tr>
<td>System Tap scripts for Lustre</td>
<td>Jason Rappleye</td>
<td>2012</td>
<td>unavailable</td>
<td>create your own metric, on/off</td>
<td>not described</td>
</tr>
<tr>
<td>Wireshark patches for LNET</td>
<td>Intel HPDD</td>
<td>2013</td>
<td>alive</td>
<td>packet decoding for Wireshark, flow graphs</td>
<td>GPL</td>
</tr>
<tr>
<td>Lustre RPC visualization</td>
<td>ZIH</td>
<td>2011</td>
<td>unavailable</td>
<td>Analysis of RPC traces</td>
<td>not described</td>
</tr>
</tbody>
</table>
Similar Information Sources

- https://wiki.hpdd.intel.com/display/PUB/Lustre+Tools
- https://wiki.hpdd.intel.com/display/PUB/Third+Party+Tools
Where to go from here

- What did I miss?
- Any updates?

- Need to put this information somewhere
  - http://goo.gl/gWpxjP