TECHNISCHE UNIVERSITÄT DRESDEN

Center for Information Services and High Performance Computing

# (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)

ZIH

Zentrum für Informationsdienste und Hochleistungsrechnen

# 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

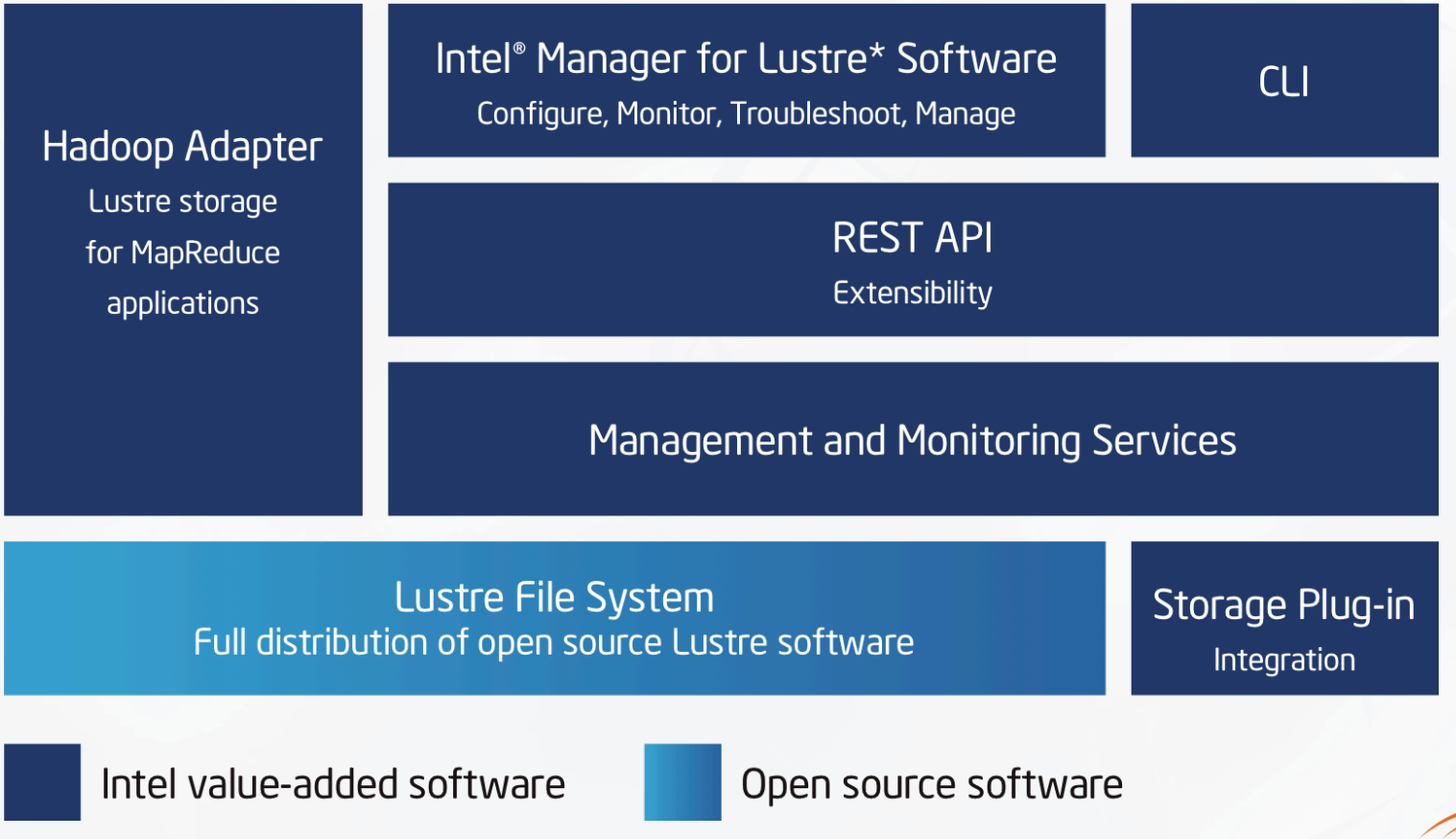
- About this talk
- Systematic overview
- **Tools, tools, tools, …**
- Wrap up

# Tools for Administrators

- **Administrators**
  - Setup/Management
  - Monitoring
  - Maintenance tasks

# Tools for Administrators (Setup/Management)

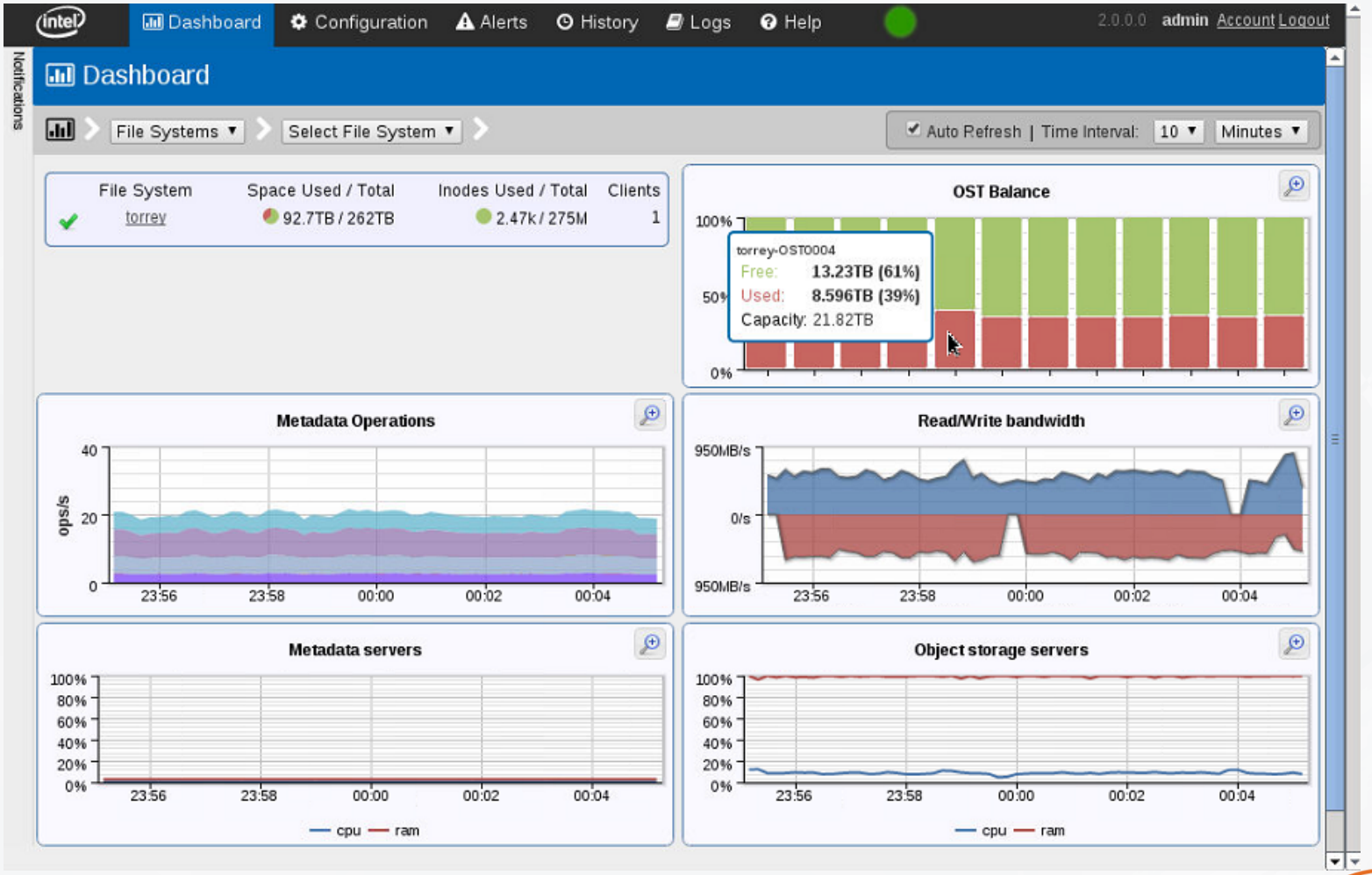
- Complete management solutions
- Commercially available from many vendors: Intel




http://lustre.intel.com

# Tools for Administrators (Setup/Management)

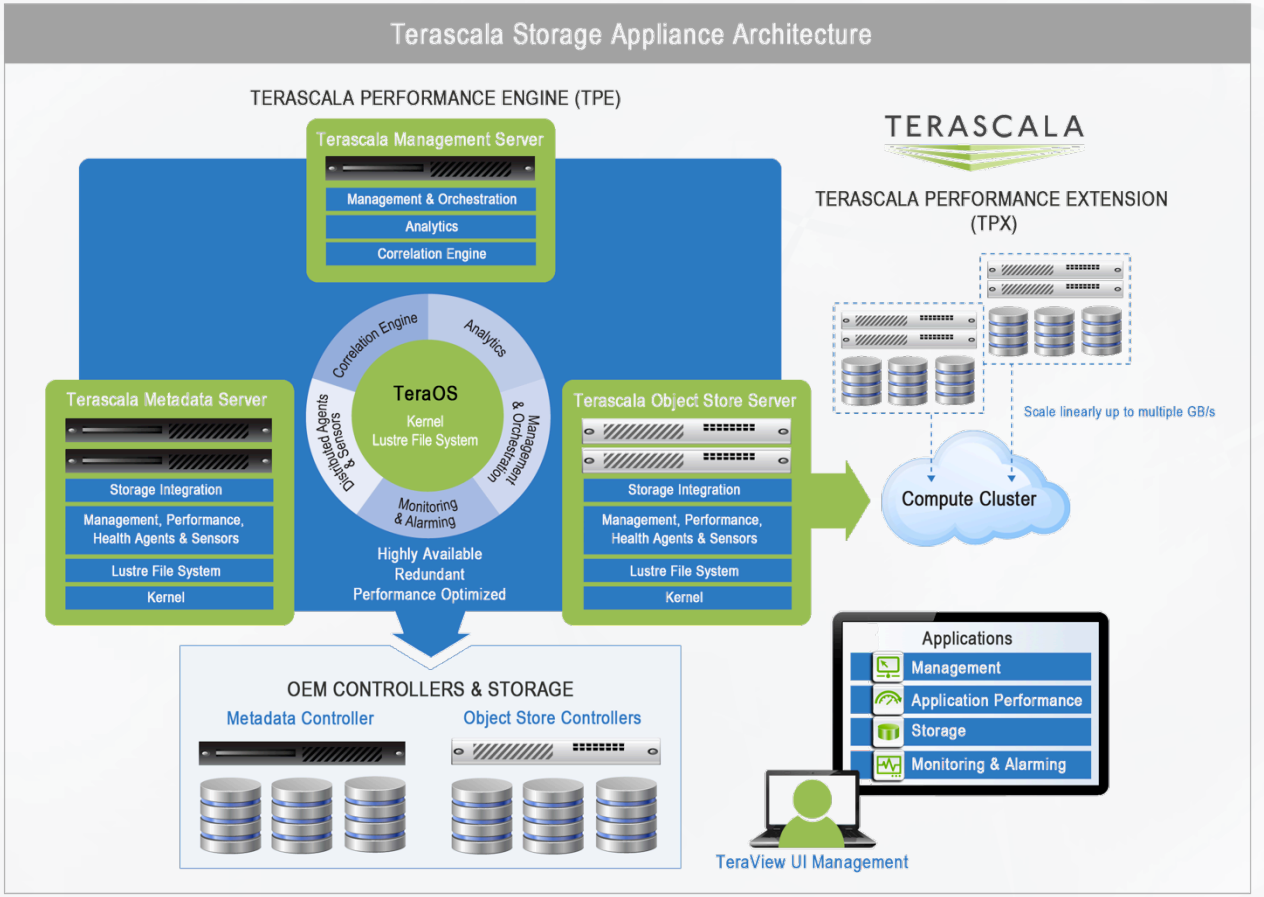
- Intel® Manager for Lustre




http://lustre.intel.com

# Tools for Administrators (Setup/Management)

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




http://www.terascala.com

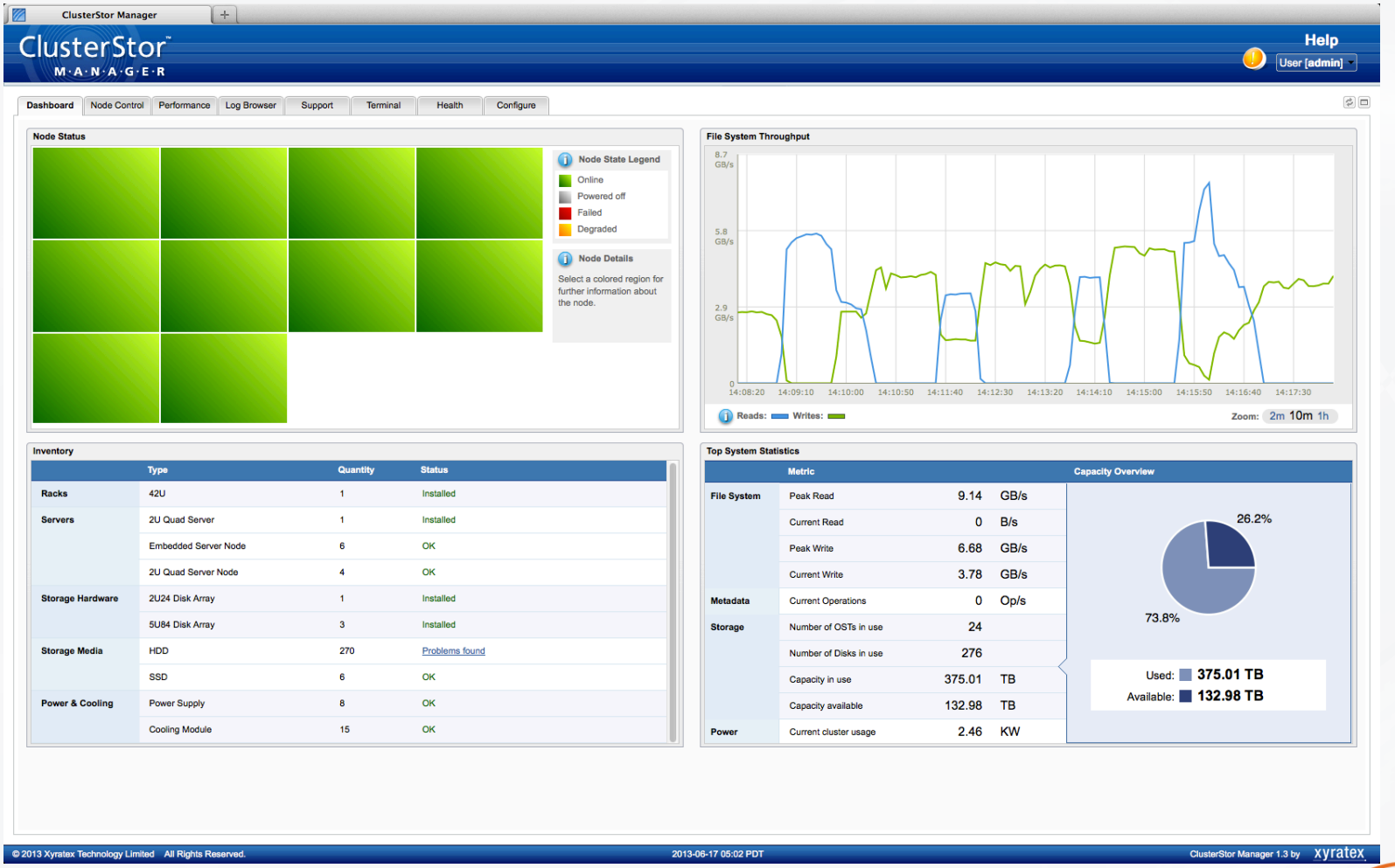
# Tools for Administrators (Setup/Management)

- 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)
  - Remote administration via ssh
    - shine install/format/start/mount …
  - status view
  - management of file system tunings
  - parallel execution of commands

- http://lustre-shine.sourceforge.net

# 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)

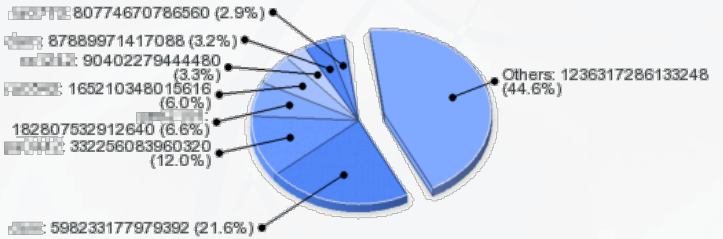
Robinhood Policy Engine

/scratch

- Users
- Groups
- Sizes
- Search

Volume | Count | List

Space used per group




| Group | Space used | Count |
|---|---|---|
| [illegible] | 544.09 TB | 12 708 782 |
| [illegible] | 302.19 TB | 2 743 617 |
| [illegible] | 166.26 TB | 3 370 537 |
| [illegible] | 150.26 TB | 1 995 581 |
| [illegible] | 82.22 TB | 1 393 682 |
| [illegible] | 79.94 TB | 7 465 108 |
| [illegible] | 73.46 TB | 522 402 |

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

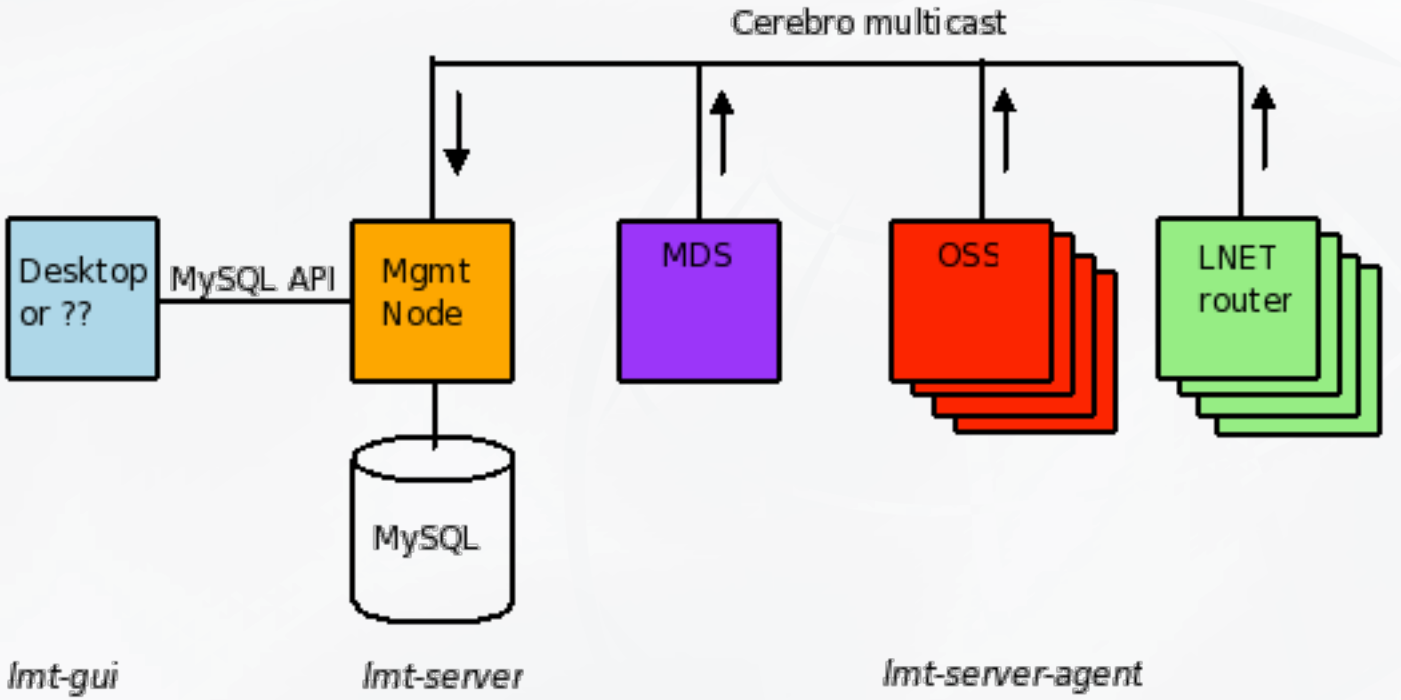
# Tools for Administrators (Monitoring)

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

```
FILESYSTEM          MDS/T   LOAD1   LOAD5 LOAD15   TASKS      OSS/T   LOAD1   LOAD5 L
taurus-scratch        1/1    0.00    0.01    0.00     900       4/48  133.22   70.07
HOST                SERV              WR_MB/S      RD_MB/S      REQS/S  JOBID
taurusi1016         taurusoss2          661.027        0.000    1366.529  IDLE
taurusi1059         taurusoss4          403.514        0.000     806.573  IDLE
taurusi1179         taurusoss4          402.491        0.000     804.996  IDLE
taurusi1046         taurusoss4          388.884        0.000     777.859  IDLE
taurusi1166         taurusoss4          381.185        0.000     762.410  IDLE
taurusi1069         taurusoss3          359.913        8.066     736.237  IDLE
taurusi1003         taurusoss3          357.138       16.121     747.431  IDLE
taurusi1051         taurusoss3          347.576        0.000     695.303  IDLE
taurusi1171         taurusoss3          344.440        0.000     689.061  IDLE
taurusi1056         taurusoss2          341.106        0.000     682.302  IDLE
taurusi1176         taurusoss2          339.234        0.000     679.487  IDLE
taurusi1112         taurusoss1          333.233        0.001     666.238  IDLE
taurusi1070         taurusoss3          332.911        0.000     667.700  IDLE
taurusi1049         taurusoss4          331.206        0.000     664.657  IDLE
taurusi1004         taurusoss3          330.407        0.000     661.556  IDLE
taurusi1010         taurusoss1          330.103        0.002     660.564  IDLE
taurusi1169         taurusoss4          329.139        0.000     658.413  IDLE
taurusi1048         taurusoss3          328.273        0.000     656.548  IDLE
taurusi1168         taurusoss3          322.521        0.000     645.048  IDLE
taurusi1072         taurusoss3          322.139        0.000     644.360  IDLE
taurusi1006         taurusoss3          317.360        0.000     634.719  IDLE
1-21 out of 1194                                xltop - Wed Sep 11 12:49:39 2013
```

# 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

# Tools for Users

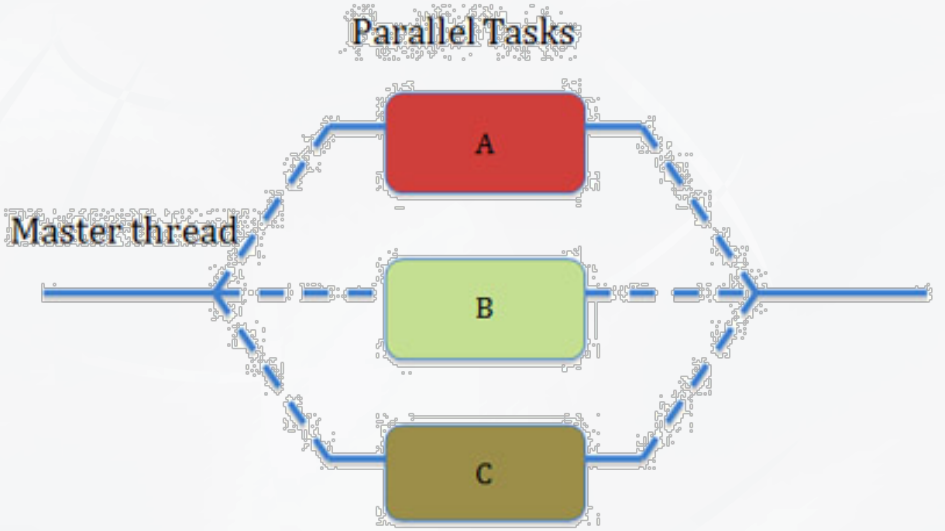
- Main areas of interest:
  - What is my code doing
  - What is the fastest way to get my stuff done
  - Programming support?

- 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

# Tools for System Architects (Benchmarking)

- Benchmarking working group at ORNL
  - Covers everything from block I/O to metadata
  - Application kernels
  - http://wiki.opensfs.org/Benchmarking_Working_Group
- 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)

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

# Admin Tools Overview (2)

| Tool | Owner/Author | Last Update/ Status | Status | Description | Lizenz |
|---|---|---|---|---|---|
| Intel Manager for Lustre | Intel HPDD | up to date | alive | complete stack, setup and management, including HA and monitoring, has plugins for external metrics from vendors | Proprietary |
| Terascala LustreStack | TeraScala | up to date | alive | complete vendor stack | Proprietary |
| Xyratex ClusterStore | Xyratex | up to date | alive | complete vendor stack | Proprietary |

# User Tools Overview

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

# Performance Tools Overview

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

# Similar Information Sources

- http://wiki.lustre.org/index.php/Diagnostic_and_Debugging_Tools
- 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