Everyday* Lustre

A short survey on Lustre tools

Sven Trautmann
Engineer, Lustre Giraffe Team, Sun Microsystems

Wednesday, September 9, 2009
Outline

• Motivation
• Setup Tools
• Management Tools
• Monitoring
• Conclusion
Who am I, What am I doing here?

- Engineer in Lustre Giraffe team since October last year
- current product: SunHPC Stack Linux Edition 2.0.1
- We try to provide a complete Linux HPC solution including
  - provisioning tools with BoIB (cobbler, oneSIS)
  - monitoring tools (ganglia, nagios)
  - managing tools (powerman, conman, freeipmi, cfengine, ...)
  - scheduling (slurm, SGE)
  - development tools (Clustertools, Sun Studio, ...)
  - and of course: Lustre server and client packages + OFED stack
- currently at Aachen University, Sandia NL, ANU/BoM, CLUMEQ ...
More giraffe information

• http://www.sun.com/software/products/hpcsoftware/

• http://blogs.sun.com/giraffe/

• linux_hpc_swstack@lists.lustre.org

• irc.freenode.net, Port 6667, #sunhpc

• http://www.randomgiraffefacts.com (not related)
Motivation

• Lustre is a complex system, learning curve is pretty steep

• starts with setting up Lustre and continues with managing and monitoring

• most tasks can be handled by scripts, most sides have their own solution

• there are a few tools out there to make working with Lustre easier

• this talk will discuss some of these tools
Lustre setup tools
The Lustre setup experience

- install Lustre packages (client & server)
- create software RAIDs (optional)
- setup Lustre network(s)
- format all MDT and OST file systems
- configure HA solution of your choice (optional)
- mount all Lustre targets
- try to mount Lustre file system from clients
- fix some problems and try again
Lustre setup tools

- three solutions (sorry if I missed any)
  - lustre_admin (Bull) last release in 2007
  - shine (CEA)
    - http://sourceforge.net/projects/lustre-shine/
  - lustre_config
- comes with Lustre
- a little outdate, still worked for 1.8.0.1
shine overview

• CLI tools designed to setup and manage the Lustre file system
• implemented in Python
• clustershell Python module is used to communicate with other nodes
• in heavy development, open source
• main developers work for CEA
shine workflow

• create configuration file for every Lustre file system
• shine install -m example.lmf
• shine format -f <fsname>
• shine start -f <fsname>
• shine mount -f <fsname>
shine model file example

```bash
[root@sven00 shine]# cat /etc/shine/models/example.lmf
fs_name: sven
nid_map: nodes=sven0[0-1] nids=sven0[0-1]@tcp0
mount_path: /sven
mgt: node=sven01 dev=/dev/sdb1
mdt: node=sven01 dev=/dev/sdb2
ost: node=sven01 dev=/dev/sdc
client: node=sven00
failover: no
description: my Lustre Filesystem
stripe_size: 1048576
stripe_count: 1
mdt_mkfs_options: -b 4096 -i 4096
mdt_mount_options: acl,user_xattr
ost_mkfs_options: -b 4096
ost_mount_options: extents,mballoc
mount_options: acl,user_xattr
quota: no
mgt_mount_path: /mnt/$fs_name/mgt
mdt_mount_path: /mnt/$fs_name/mdt/$index
ost_mount_path: /mnt/$fs_name/ost/$index
```
shine installation

root@sven00 shine]# shine install \
   -m /etc/shine/models/example.lmf
Using Lustre model file /etc/shine/models/example.lmf
Configuration files for file system sven have been installed successfully.

Lustre targets summary:
  1 MGT on sven01
  1 MDT on sven01
  1 OST on sven01

Use `shine format -f sven' to initialize the file system.
[root@sven00 shine]#
[root@sven00 shine]# shine format -f sven
Format sven on sven01: are you sure? (y)es/(N)o: y
Starting format of 3 targets on sven01
[18:54] In progress for 3 target(s) on sven01 ...
[18:54] In progress for 2 target(s) on sven01 ...
Format successful.

FILESYSTEM COMPONENTS STATUS (sven)
+-----+--+-------+------------+
|type |# |nodes  |   status   |
+-----+--+-------+------------+
|MGT  |1 |sven01 |offline (1) |
|MDT  |1 |sven01 |offline (1) |
|OST  |1 |sven01 |offline (1) |
+-----+--+-------+------------+

[root@sven00 shine]#
shine start & mount

[root@sven00 ~]# shine start -f sven
Starting 3 targets on sven01
[21:27] In progress for 1 target(s) on sven01 ...
Start successful.

FILESYSTEM COMPONENTS STATUS (sven)

<table>
<thead>
<tr>
<th>type</th>
<th>#</th>
<th>nodes</th>
<th>status</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>sven01</td>
<td>online (1)</td>
</tr>
</tbody>
</table>

[root@sven00 ~]# shine mount -f sven
Starting sven clients on sven00...
Mount successful on sven00
[root@sven00 ~]#
shine comments

• shine and clustershell need to be installed on all nodes, MDS, OSS and clients
• no fail-over (as of shine 0.903)
• no entry in modprobe.conf for other networks than tcp
• only one file system per MGS (?)
• almost no documentation (?)
• no software RAID setup
• shine looks very promising though
lustre_config

• create a csv file describing the Lustre file system
• enable password-less ssh login to all Lustre nodes
• **call:** lustre_config <csv_file>
Lustre csv file

f0011, options lnet networks=o2ib0,/root/mds.fs,/mnt/mds,mgs|mdt,,,,,--device-size=16777216 --param="mdt.group_upcall=/usr/sbin/l_getgroups" --stripe-count-hint=4,-i 4095,,

f0012, options lnet networks=o2ib0,/root/oss.fs,/mnt/ost,ost,,5.6.9.11@o2ib,0,--device-size=33554432,-i 16357,,

f0013, options lnet networks=o2ib0,/root/oss.fs,/mnt/ost,ost,,5.6.9.11@o2ib,1,--device-size=33554432,-i 16357,,

f0014, options lnet networks=o2ib0,/root/oss.fs,/mnt/ost,ost,,5.6.9.11@o2ib,2,--device-size=33554432,-i 16357,,

f0015, options lnet networks=o2ib0,/root/oss.fs,/mnt/ost,ost,,5.6.9.11@o2ib,3,--device-size=33554432,-i 16357,
lustre_config comments

• comes with Lustre package
• bash script, uses ssh
• currently no new development
• lots of features (try lustre_config -h)
  • fail-over, lustre networks ...
Management Tools
What can be managed?

- Lustre parameters can be tuned/changed
  - some parameters need to be changed on all nodes
- tools like pdsh can help
- other management tasks: deactivating OSTs
- mounting/un-mounting clients
- shine can mount clients and tune Lustre (in theory, not tested yet)
Lustre Monitoring Tools
Monitoring tools

• general monitoring tools
  • ganglia, nagios, cacti ...

• filesystem tools
  • collectl, ...

• Lustre specific tools
  • LMT2
Ganglia, Nagios ...

Overview of unspecified

CPUs Total: 3
Hosts up: 2
Hosts down: 1

Avg Load (15, 5, 1m):
2%, 6%, 13%

Localtime: 2009-05-06 02:08

Cluster Load Percentages
- 25-50 (33.33%)
- 0-25 (33.33%)
- down (33.33%)

Show Hosts: yes  no  | unspecified load_one last hour sorted descending | Columns 4 5  Size small

192.168.93.201
doederal
Last heartbeat 51s ago

head

192.168.93.202

load_one last hour
(now 0.37)

load_one last hour
(now 0.01)
### Current Network Status
- Last Updated: Thu May 7 15:01:25 CEST 2009
- Updated every 90 seconds
- Nagios® 3.1.0 - www.nagios.org
- Logged in as nagiosadmin

View History For all Hosts
View Notifications For all Hosts
View Host Status Detail For all Hosts

---

### Service Status Details For All Hosts

<table>
<thead>
<tr>
<th>Host</th>
<th>Service</th>
<th>Status</th>
<th>Last Check</th>
<th>Duration</th>
<th>Attempt</th>
<th>Status Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>localhost</td>
<td>Current Load</td>
<td>OK</td>
<td>05-07-2009 14:58:42</td>
<td>3h 20h 43m 1s 3s</td>
<td>1/4</td>
<td>OK - load average: 0.98, 0.51, 0.35</td>
</tr>
<tr>
<td>localhost</td>
<td>Current Users</td>
<td>OK</td>
<td>05-07-2009 14:59:20</td>
<td>3h 20h 41m 53s</td>
<td>1/4</td>
<td>USERS OK - 2 users currently logged in</td>
</tr>
<tr>
<td></td>
<td>HTTP</td>
<td>CRITICAL</td>
<td>05-07-2009 14:57:57</td>
<td>0h 0h 6m 28s</td>
<td>4/4</td>
<td>Connection refused</td>
</tr>
<tr>
<td></td>
<td>PING</td>
<td>OK</td>
<td>05-07-2009 15:00:35</td>
<td>3h 20h 46m 20s</td>
<td>1/4</td>
<td>PING OK - Packet loss = 0%, RTA = 0.07 ms</td>
</tr>
<tr>
<td></td>
<td>Root Partition</td>
<td>WARNING</td>
<td>05-07-2009 14:58:34</td>
<td>1h 17h 4m 2s</td>
<td>4/4</td>
<td>DISK WARNING - free space: / 2548 MB (1.7% inode=90%)</td>
</tr>
<tr>
<td></td>
<td>SSH</td>
<td>OK</td>
<td>05-07-2009 14:59:12</td>
<td>3h 21h 14m 43s</td>
<td>1/4</td>
<td>SSH OK - OpenSSH_4.3 (protocol 2.0)</td>
</tr>
<tr>
<td></td>
<td>Suse Users</td>
<td>OK</td>
<td>05-07-2009 14:59:40</td>
<td>3h 20h 44m 28s</td>
<td>1/4</td>
<td>SWAP OK - 100% free (2015 MB out of 2015 MB)</td>
</tr>
<tr>
<td></td>
<td>Total Processes</td>
<td>OK</td>
<td>05-07-2009 15:01:12</td>
<td>3h 20h 43m 4s</td>
<td>1/4</td>
<td>PROCS OK, 38 processes with STATE = R/SZDT</td>
</tr>
</tbody>
</table>

8 Matching Service Entries Displayed
Ganglia, Nagios ...

- these tools were not developed for Lustre
- some parameters can be monitored using these tools
- e.g. network load, software RAID state
- most Lustre specific things can be monitored, but not out of the box
collectl

• developed by HP
• shows file system statistics
• latest version supports Lustre
• was in SunHPC Linux stack 1.x, not in 2.0
• will probably be in 2.1 again
# CPU output example

<table>
<thead>
<tr>
<th>cpu</th>
<th>sys</th>
<th>inter</th>
<th>ctxsw</th>
<th>KBRead</th>
<th>Reads</th>
<th>KBWrit</th>
<th>Writes</th>
<th>KBIn</th>
<th>PktIn</th>
<th>KBOut</th>
<th>PktOut</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1000</td>
<td>125</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1010</td>
<td>175</td>
<td>48</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1037</td>
<td>129</td>
<td>0</td>
<td>0</td>
<td>717</td>
<td>95</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1014</td>
<td>150</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1002</td>
<td>140</td>
<td>0</td>
<td>0</td>
<td>44</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1013</td>
<td>155</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1000</td>
<td>134</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1012</td>
<td>164</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>998</td>
<td>133</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1012</td>
<td>141</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1003</td>
<td>151</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1011</td>
<td>155</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>996</td>
<td>141</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
LMT 2 overview

- development started at LLNL
- monitors Lustre File System servers
- collects data using the Cerebro monitoring system and stores it in a MySQL database
- GUI and CLI clients display historical and real time data
LMT dependencies

- LMT uses Cerebro to collect data
- Cerebro needs genders and LLNL specific genders extension gendersllnl
- genders and gendersllnl need pdsh and some Perl modules
- LMT Client’s CLI commands and GUI is implemented in Java
- LMT uses MySQL database to store data
Cerebro

• Cerebro is a monitoring tool, sharing roots with Ganglia

• https://computing.llnl.gov/linux/cerebro.html

• Cerebro comes with pre-defined metrics e.g. for loadavg and memory usage

• configuration similar to Ganglia’s gmond

• uses multi-cast or uni-cast, just like Ganglia
LMT server agent

- provides new Cerebro metrics
- lmt-server-agent package provides Lustre metrics for Cerebro

```bash
root@sven01 ~]# cerebro-stat -l | grep lmt
lmt_mds
lmt_oss
lmt_ost
[root@sven01 ~]# cerebro-stat -m lmt_oss
sven01: 1.0;sven01;0.400000;18.965605
[root@sven01 ~]# cerebro-stat -m lmt_ost
sven01: 1.0;sven01;sven-OST0000;524232;524288;7906228;8256952;0;0
[root@sven01 ~]#
```
LMT Server Setup I

• LMT server provides binding between cerebro and mysql database

• LMT server can run on any node configured as Cerebro server

• needs to run a MySQL service

• one table for every Lustre file system

• MySQL database may be used by other applications as well
LMT Server Setup II

[root@sven00 ]# yum install mysql-server
...
[root@sven00 ]# /etc/init.d/mysqld start
...
[root@sven00 ]# /usr/bin/mysqladmin -u root password 'changeme'
[root@sven00 ]# yum install lmt-server
...
[root@sven00 ]# cat /usr/share/lmt/cron/lmtrc
filesys.1.name=sven
filesys.1.mountname=/sven
filesys.1.dbhost=localhost
filesys.1.dbport=3306
filesys.1.dbuser=lwatchadmin
filesys.1.dbauth=
filesys.1.dbname=filesystem_sven
[root@sven00 ]#
create LMT tables

filesystem
{
    name        sven
    mountpoint /sven
}

mds
{
    name        sven01
    uuid        sven-MDT0000
    nid         192.168.93.200@tcp0
    device      /dev/sdb2
}

ost
{
    name sven01
    uuid sven-OST0000
    nid {NODENAME}@tcp0
    device /dev/sdc
    skip 1
}
LMT database

```bash
# create_lmtconfig -a -d -t -f lmt-config.txt > lmtdb.in
# mysql -p mysql <lmtdb.in
```

```
[root@sven00 cerebro]# mysql -p
Enter password:
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 2
Server version: 5.0.45 Source distribution

Type 'help;' or '\h' for help. Type '\c' to clear the buffer.

mysql> use filesystem_sven
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
mysql> select * from OST_DATA;

+--------+-------+------------+-------------+---------+-------------+-------------+-------------+-------------+
| OST_ID | TS_ID | READ_BYTES | WRITE_BYTES | PCT_CPU | KBYTES_FREE | KBYTES_USED | INODES_FREE | INODES_USED |
+--------+-------+------------+-------------+---------+-------------+-------------+-------------+-------------+
|      1 |     1 |          0 |           0 |    NULL |     7906228 |      350724 |      524232 |          56 |
|      1 |     2 |          0 |           0 |    NULL |     7906228 |      350724 |      524232 |          56 |
| ...    |      1 |    276     |  1935671296 |    NULL |     7064496 |     1192456 |      524200 |          88 |
|      1 |     277 |          0 |  1935671296 |    NULL |     7064496 |     1192456 |      524200 |          88 |
| ...    |      1 |    278     |  1935671296 |    NULL |     7064496 |     1192456 |      524200 |          88 |
+--------+-------+------------+-------------+---------+-------------+-------------+-------------+-------------+
278 rows in set (0.03 sec)

mysql> quit
Bye
[root@sven00 cerebro]#
```
LMT Clients

• All clients are implemented in Java
• just MySQL clients, which could be easily implemented in many languages
• two CLI clients ltop and lstat
• one GUI client lwatch
• extra config file in /usr/share/lmt/etc.
Itop and Istat

[root@sven00 ~]# ltop
2009-07-26 00:05:40 --- sven: OST Report

OST Name        Read (MB/s)  Write (MB/s)
sven-OST0000    0.00          0.00

[root@sven00 ~]# lstat

-------- 2009-07-26 00:06:25 --------

[OST Name]    [Read MB/s]    [Write MB/s]
sven-OST0000  0.00            0.00

-------- 2009-07-26 00:06:30 --------

[OST Name]    [Read MB/s]    [Write MB/s]
sven-OST0000  0.00            0.00

-------- 2009-07-26 00:06:35 --------

[OST Name]    [Read MB/s]    [Write MB/s]
sven-OST0000  0.00            0.00
lwatch
Iwatch
What’s missing?

• there are a few tools out there which make working with Lustre easier

• problem: no complete solution, combining all the tasks

• shine looks promising, but: not all features are implemented yet

• LMT 2 looks good, but: setup is hard and relies on a lot of LLNL packages
Wish list

• Lustre management/monitoring solution
• setting up Lustre with fail over
• plug-ins for different HA solutions
• monitor errors on all nodes
• Lustre self-healing and/or provide hints on how to fix problems
• integrated monitoring solution
• ...

Wednesday, September 9, 2009
Thank you! Questions? Comments?
svtr@sun.com

Wednesday, September 9, 2009