Lustre Monitoring Tool Version 3

Jim Garlick

garlick@llnl.gov

Livermore Computing

Lawrence Livermore National Laboratory

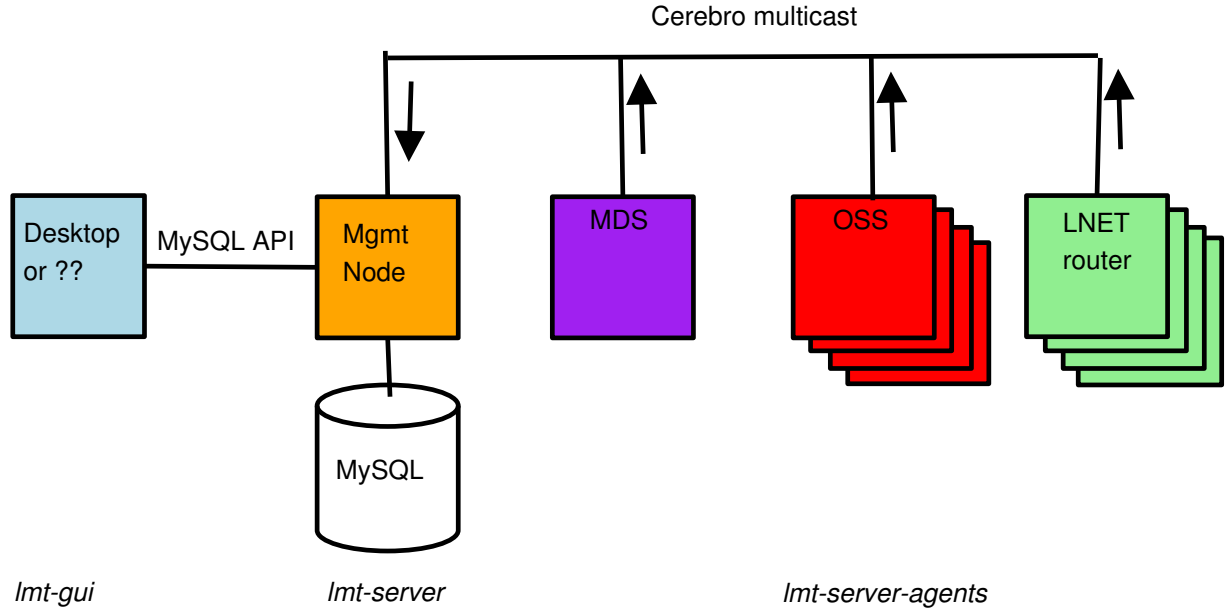
LLNL-PRES-459655-DRAFT

Background

- Lustre provides some very nice metrics in `/proc`, but no mechanism to aggregate metrics across a file system.
- LMT version 1 was a python application for visualizing Lustre metrics, mainly used in-house at LLNL.
- LMT version 2 rewrite in C/Java added MySQL for capturing historical data, plotting capability in the GUI client, and `ltop` text utility.
- MySQL history is cool! [*Uselton 2009 CUG*]
- `ltop` in particular has become a useful sys admin tool.

LMT Overview

The Lustre Monitoring tool uses cerebro and MySQL for data collection and storage. Data can be mined directly from MySQL or visualized with LMT clients (`ltop, lwatch`).



LMT Version 3 uses the same architecture as Version 2.

LMT Version 2 Problems

Based on LLNL experience and lmt-discuss mailing list:

- Lustre config must be expressed in an odd language, then pre-loaded into MySQL.
- Nothing functions until both MySQL and cerebro are up.
- Poor error handling and logging make debug difficult.
- There are two overlapping config files in odd locations.
- The cerebro module code is prototype quality and brittle.

frustration for new users and maintainer!

Improved in Version 3

- Lustre config is automatically determined on the fly.
- `ltop` now functions as soon as cerebro is up.
  MySQL is actually optional now.
- Error handling and logging are rewritten/improved.
- Cerebro module code has been refactored/rewritten.
- There is a single new config file: `/etc/lmt/lmt.conf`
- More data is collected/shown in `ltop` [*demo later*].

Unchanged in Version 3

- The architecture is the same (except `ltop`).
- The database schema is unchanged.
- The `lwatch/lstat` java clients are unchanged (moved to separate lmt-gui package).
- Cron aggregation scripts that convert high $\rightarrow$ low-res MySQL sample data still exist (kludge!).

LMT Version 3 Setup

1. Install packages.
2. Configure cerebro and restart cerebrod on LMT and Lustre servers. (Test with `ltop`.)
3. Run `mysql_secure_installation` or equiv, then `/usr/share/lmt/mkusers.sql`
4. Set up `/etc/lmt/lmt.conf`.
5. Create databases for each file system to be monitored: `lmtinit -a fsname`. (Test with `lwatch`.)
6. Add cron job for aggregation scripts.

LUA-based lmt.conf

```
lmt_cbr_debug = 0
lmt_proto_debug = 0
lmt_db_debug = 0

lmt_db_host = nil
lmt_db_port = 0
lmt_db_rouser = "lwatchclient"
lmt_db_ropasswd = nil
lmt_db_rwuser = "lwatchadmin"

f = io.open("/etc/lmt/rwpasswd")
if (f) then
  lmt_db_rwpasswd = f:read("*all")
  f:close()
else
  lmt_db_rwpasswd = nil
end
```

Version 3 Metric Protocol Changes

- mds.v2 $\rightarrow$ mdt.v1:
  - MDS + multiple MDT data in one metric.
  - Dropped 60 seldom-used mdops ($81 - 60 = 21$).
- oss.v1 + ost.v1 $\rightarrow$ ost.v2:
  - OSS + multiple OST data in one metric
  - Added IOPS, lock count, lock grant/cancel rate, (re-)connects, recovery state.
- osc.v1: OST state from MDS pov, definitive OST list.
- router.v1: No change.

New data is only displayed in ltop, not stored in MySQL.

Future Work

- LMT schema needs a revision to accommodate new data.
- `lwatch` should display new data.
- Add support for Lustre 2.x ioctl interface.
- Add support for Lustre 2.x ZFS servers.
- New `ltop` screens for routers, MDS's, etc.
- `ltop` should also support direct `/proc`, MySQL.
- Use lnet/ptlrpc to gather data, tighter integration with Lustre.

What other Lustre metrics to be monitored/visualized?

LMT Support and Downloads

- Google code project:

  `http://code.google.com/p/lmt`.

- Email support list:

  `http://groups.google.com/lmt-discuss`.