The Past and Future of Lustre Configuration

nathan_rutman@xyratex.com

xyratex

A Deep Dive into the Evolution of a Lustre Feature

xyratex

The Bad Old Days: Lustre 1.4

- lmc...

```
lmc -m local.xml --add node --node localhost
lmc -m local.xml --add net --node localhost –nid localhost --nettype tcp
lmc -m local.xml --format --add mds --node localhost -- mds mds-test --fstype ext3 --dev \
/tmp/mds-test  –size 50000
```

- ...to create an XML file...

```
<profile name="PROFILE_host" uuid="PROFILE_host_UUID">
 <ldlm_ref uuidref="ldlm_UUID"/>
 <network_ref uuidref="NET_host_tcp_UUID"/>
```

- ...distribute out-of-band (NFS)...
- ...lconf to parse the XML (python, libxml)...

```
lconf --node client /lustre/config.xml
```

- … called lctl to set up the obd’s

```
+ /usr/sbin/lctl
  cfg_device MDC_head4_mds-test_MNT_localhost
  setup mds-test_UUID localhost_UUID
  quit
```

Loads Better: Lustre 1.6

- Introduced MGS and MGC
- MGC sends info to MGS the first time a server is started
  - nids, failover partner, params (from mkfs)
  - hence “mountconf”
- MGS creates per-OST config llogs, adds to MDT and client llogs
  - behind the scenes, these kept the lctl format
- Servers and clients get logs from MGS every time they start
- Use ldlm “config” lock to notify of updates
  - OST addition
  - param changes

Param updates

- 1.4: proc

```
/proc/fs/lustre/obdfilter/testfs-OST0013/sync_journal=0
```

No way to change params globally

- 1.6: lctl set_param

```
lctl set_param obdfilter.testfs-OST0013.sync_journal=0
```

Just called local proc.

- Add global setting via MGS: lctl conf_param

```
lctl conf_param testfs-OST00013.ost.sync_journal=0
```

Permanent Parameters

What idiot changed the syntax?

```
 * lustre/mgs/mgs_handler.c
 * Author: Nathan Rutman <nathan@clusterfs.com>
```

Explain thyself:

- Solaris in the wind, no proc
- Hierarchy
  - Per-server config files
  - Service is primary instead of module

    ```
    obdfilter.testfs-OST0013.param
    testfs-OST00013.ost.param
    ```

  - Opportunity to fix names

But that never happened :(

Problems Today

- Want to harmonize the naming
- No globbing: *.ost.param
- Each conf_param needs to be sent to the right config log
    - e.g. lov changes to both MDS and client
- Each param (group) handler needs special plumbing
- Some services don’t have config files (ldlm)
- So many params *can’t even be set*

Let’s fix it

- Recognize that proc isn’t going away
- Create a new universal “param” config log
- Distribute it to everyone
- Parse it using an upcall to lctl set_param
    - Use globbing to control matches
    - Exact same processing path
- No need for MGS to figure targets, no need to add handlers
- Nodes will ignore params that don’t match

Landed in Lustre 2.5

- LU-3155
  - Temporary:

```
set_param obdfilter.*.client_cache_seconds=15
```

  - Permanent:

```
set_param -P obdfilter.*.client_cache_seconds=15
```

The Future

But

- Configuration llogs are opaque
- Writeconf’s are still required in some corner cases

Vision

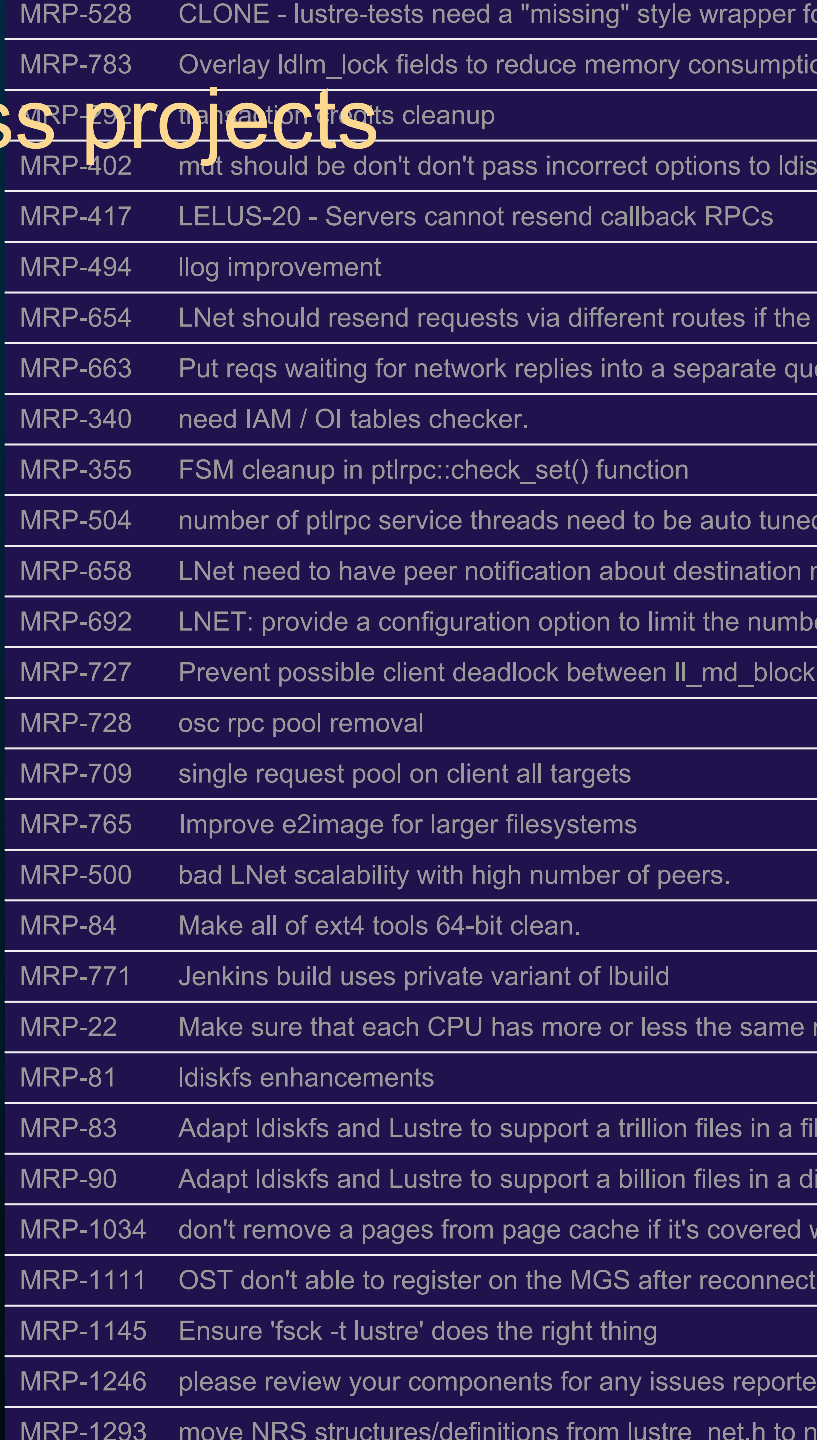
- Get rid of config llogs
- With imperative recovery we don’t need to predefine failover locations
- Servers just report their NID as they start services
- MGS just needs to track who started where
- And a simple text file on the MGS for the params: /etc/lustre/testfs.txt

And now for something completely different

xyratex

Xyratex in-progress projects

- PDRAID
- T10 end-to-end (client to drive)
- Performance improvements
  - SSF
  - small file
  - NRS policies
- WNC
- Xperior
- Lustre internal improvements



MRP-528	CLONE - lustre-tests need a "missing" style wrapper fo
MRP-783	Overlay ldlm_lock fields to reduce memory consumptio
MRP-292	transaction credits cleanup
MRP-402	mdt should be don't don't pass incorrect options to ldis
MRP-417	LELUS-20 - Servers cannot resend callback RPCs
MRP-494	llog improvement
MRP-654	LNet should resend requests via different routes if the
MRP-663	Put reqs waiting for network replies into a separate qu
MRP-340	need IAM / OI tables checker.
MRP-355	FSM cleanup in ptlrpc::check_set() function
MRP-504	number of ptlrpc service threads need to be auto tune
MRP-658	LNet need to have peer notification about destination r
MRP-692	LNET: provide a configuration option to limit the numb
MRP-727	Prevent possible client deadlock between ll_md_block
MRP-728	osc rpc pool removal
MRP-709	single request pool on client all targets
MRP-765	Improve e2image for larger filesystems
MRP-500	bad LNet scalability with high number of peers.
MRP-84	Make all of ext4 tools 64-bit clean.
MRP-771	Jenkins build uses private variant of lbuild
MRP-22	Make sure that each CPU has more or less the same r
MRP-81	ldiskfs enhancements
MRP-83	Adapt ldiskfs and Lustre to support a trillion files in a fi
MRP-90	Adapt ldiskfs and Lustre to support a billion files in a di
MRP-1034	don't remove a pages from page cache if it's covered w
MRP-1111	OST don't able to register on the MGS after reconnect
MRP-1145	Ensure 'fsck -t lustre' does the right thing
MRP-1246	please review your components for any issues reporte
MRP-1293	move NRS structures/definitions from lustre_net.h to n

Thank You

nathan_rutman@xyratex.com

xyratex