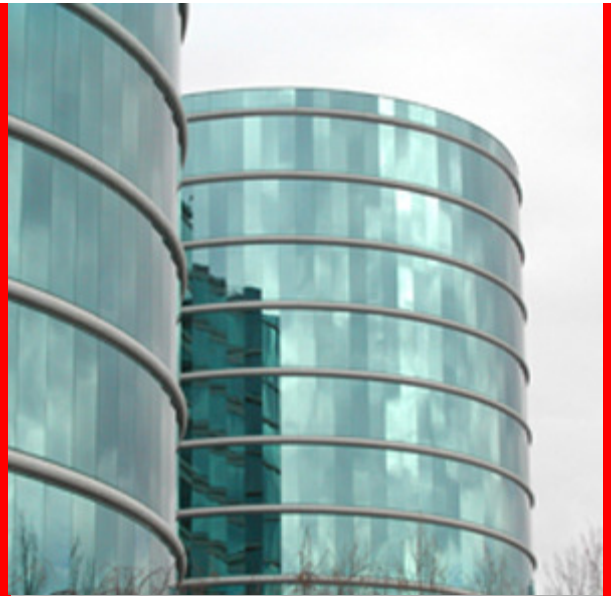

ORACLE®



ORACLE

Port quota to osd

Landen
October 27, 2010

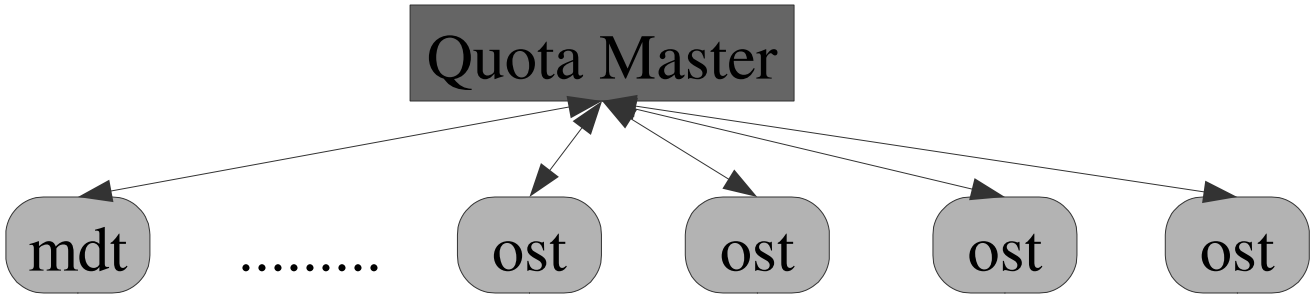
Agenda

- Architecture Overview
- New Quota Interface
- Main structures
- Main issues

Architecture Overview

Architecture Primer

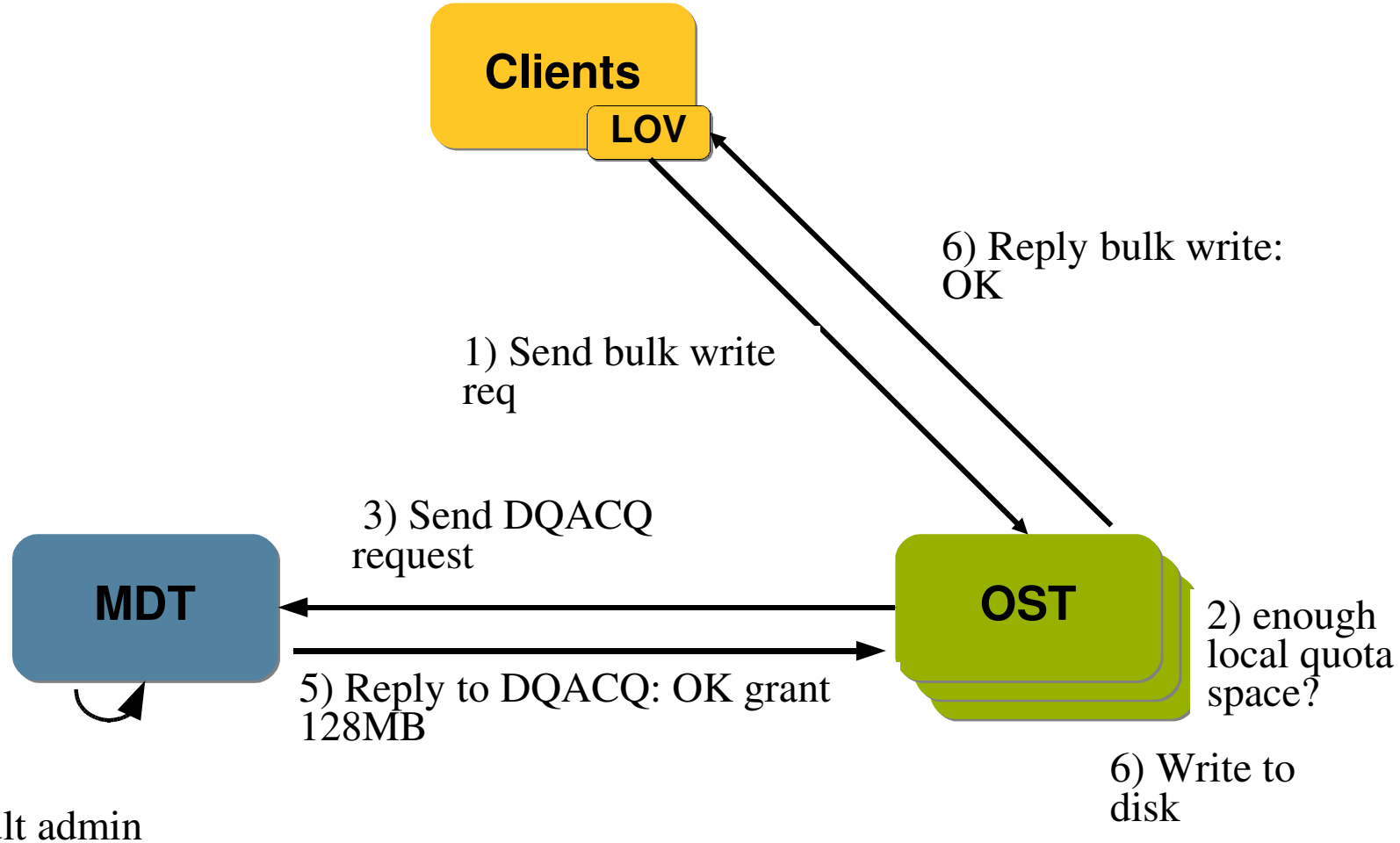
- A centralized server hold the cluster wide limits: the quota master(s)
  - guarantees that global quota limits are not exceeded
  - track quota usage on slaves
- Quota slaves
  - all the OSTs and MDT(s)
  - manage local quota usage/hardlimit
  - acquire/release quota space from the master



Acquire/Release Protocol

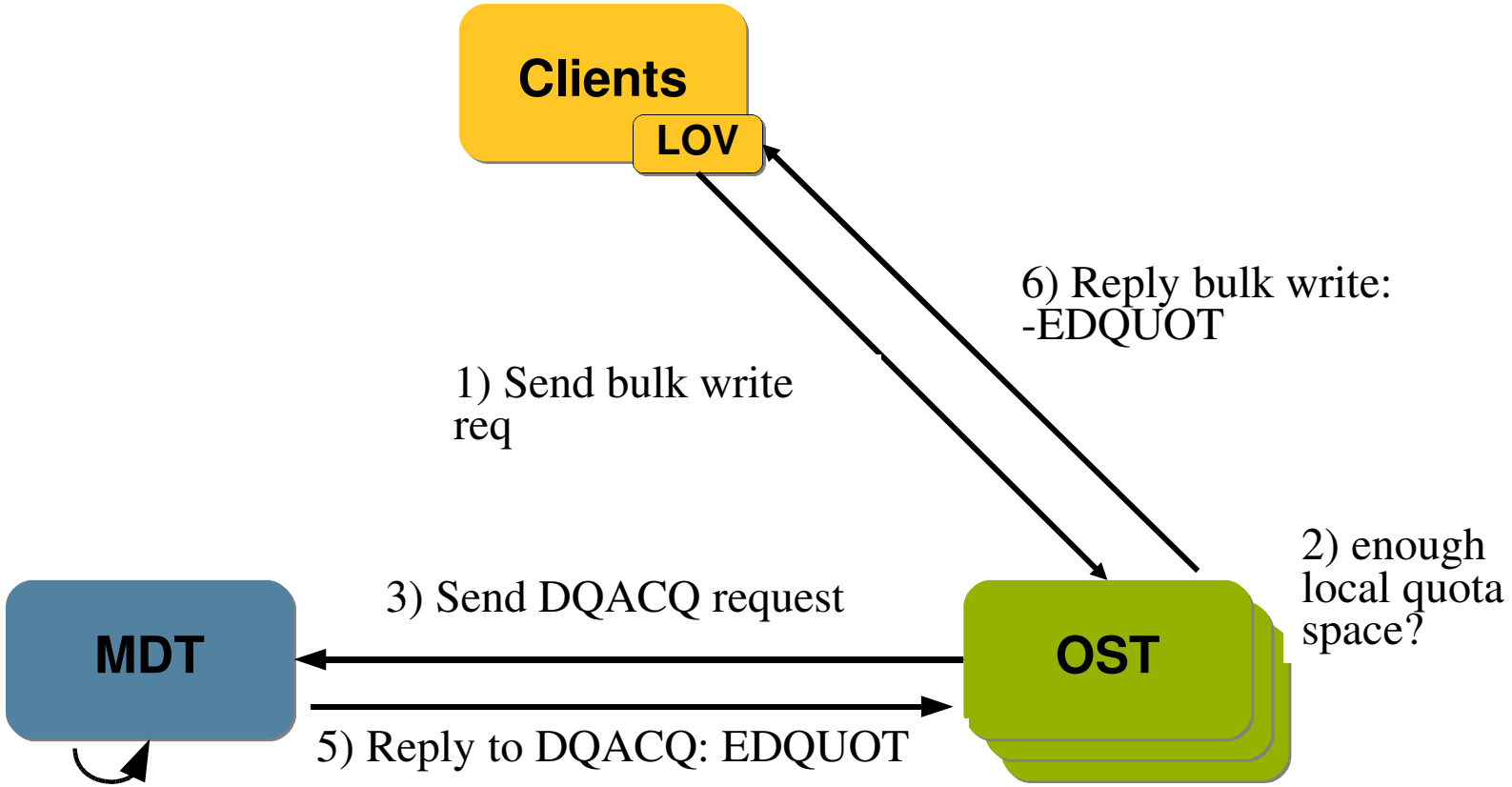
- Two different RPC types
  - DQACQ = Disk Quota ACQuire
  - DQREL = Disk Quota RELease
- DQACQ/DQREL RPCs are
  - initiated by slaves
  - processed by master(s)
- increase/lower the local **hardlimit** on slaves
- increase/decrease administrative **usage** on the master

Quota protocol overview:
Enough quota



4) Consult admin quota files. Enough space.

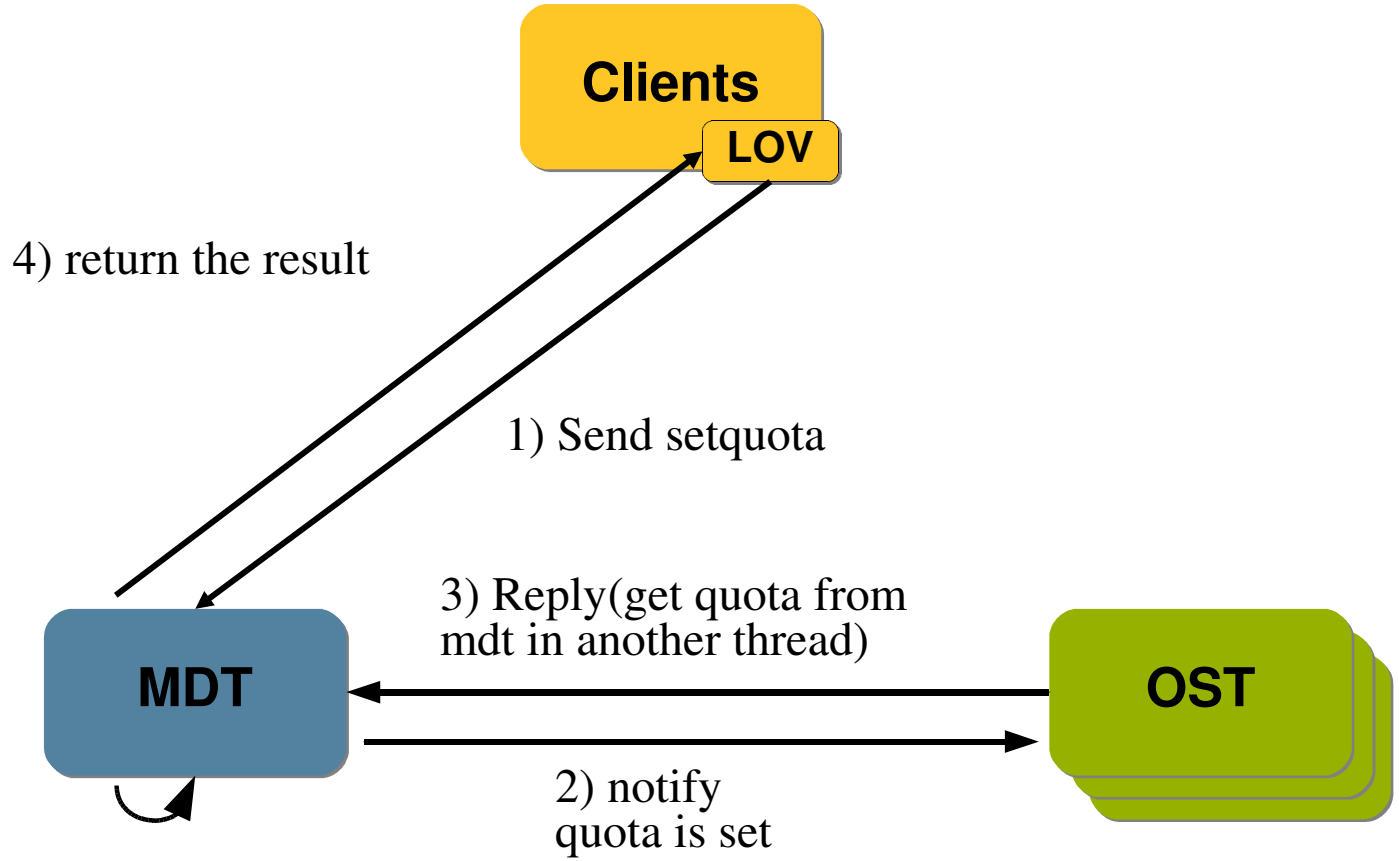
Quota protocol overview:
Quota exceeded - EDQUOT



4) Consult admin

quota files. Quota exceeded

Quota protocol overview:
Setquota



New Quota Interface

Virtual Quota Interface

- Quota Master
  - QUOTA_MASTER_UID_OID
  - QUOTA_MASTER_GID_OID
  - REMOTE_QUOTA_LIMIT_UID_OID
  - REMOTE_QUOTA_LIMIT_GID_OID
  - REMOTE_QUOTA_USAGE_UID_OID
  - REMOTE_QUOTA_USAGE_GID_OID
- Quota Slave
  - QUOTA_SLAVE_LIMIT_UID_OID
  - QUOTA_SLAVE_LIMIT_GID_OID
  - QUOTA_SLAVE_USAGE_UID_OID
  - QUOTA_SLAVE_USAGE_GID_OID

Quota objects relationships

Quotacheck, change qunit, setquota and getquota use lod/osp

Quota Objects on ldiskfs-osd

Quota OID	on-disk representation
QUOTA_MASTER_UID_OID	/OBJECTS/admin_quotafile_v2.usr
QUOTA_MASTER_GID_OID	/OBJECTS/admin_quotafile_v2.grp
QUOTA_SLAVE_LIMIT_UID_OID	iam(located by oi)
QUOTA_SLAVE_LIMIT_GID_OID	iam(located by oi)
QUOTA_SLAVE_USAGE_UID_OID	/lquota_v2.user
QUOTA_SLAVE_USAGE_GID_OID	/lquota_v2.group

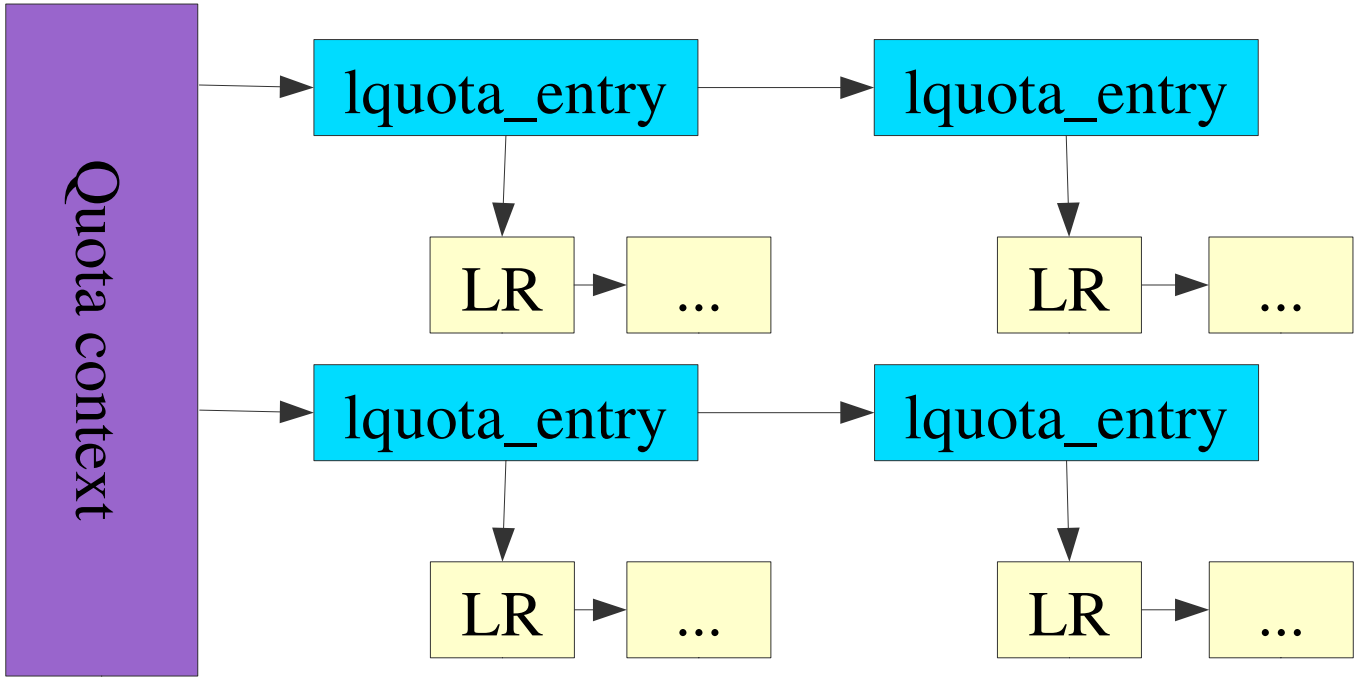
Quota Objects on kdmu-osd

Quota OID	on-disk representation	purpose
QUOTA_MASTER_UID_OID	LL_USERQUOTA_M ZAP (new)	global user settings
QUOTA_MASTER_GID_OID	LL_GROUPQUOTA_M ZAP (new)	global group settings
QUOTA_SLAVE_USAGE_UID_OID	DMU_USERUSED_OBJECT ZAP	local user space accounting
QUOTA_SLAVE_LIMIT_UID_OID	LL_USERQUOTA_S ZAP (new)	local user settings
QUOTA_SLAVE_USAGE_GID_OID	DMU_GROUPUSED_OBJECT ZAP	local group space accounting
QUOTA_SLAVE_LIMIT_UID_OID	LL_GROUPQUOTA_S ZAP (new)	local group settings

Main structures

Quota main structures

- Quota context
- lquota_entry
- Lquota_requests



Main Issues

Quota hook

Quota hook will be added in os:

- lquota_start_quota() in osd_declare_write_commit()
  - it will update pending quota info
  - Send sync quota requests if necessary
  - Trigger -EDQUOT

```
osd_declare_write_commit() {
    ...
     osd_estimate(..., &env->lc_ses->quota_info);
     rc = lquota_start_quota(env,   ...);
     If (rc)
          return -EDQUOT;
     ....
}
```

- lquota_end_quota() in osd_dtrans_stop()
  - it will release pending quota
  - send async acq/rel quota req if needed

Quota enforcement

- Trigger -EDQUOT in lquota_start_quota() in osd_declare_write_commit()
- Trigger -EDQUOT when:
  - Blocks + metadata + pending_data + usage > limit
  - Quota requests returned from mds return nothing
  - No other requests for this uid/gid is inflight
  - No adjusting qunit is on the way

thread 1	thread 2
Add pending	
Update on disk	
	Add pending
	Update on disk
Release pending	
	Release pending

Quotaon ldiskfs-osd always

- We can do this in two ways:
  - Move quota limits to a new iam object
  - cfs_cap_raise(CFS_CAP_SYS_RESOURCE);
- Benefits:
  - Can reduce quotacheck(scan inode by inode);
  - Get usage for uid/gid even if quota is off

Quotacheck

- Create quota files/zap if necessary
- Make quota usage on qm = quota limits on qs
- Every target scans inode by inode(22741)

Where is quota context?

- Have two kinds of quota context:

Master quota context and slave quota context

- Slave quota hook is at osd, so slave quota context should be places in osd_device. (import issue)

- Master quota will be at mdt_device(maybe mgs later)

Estimate meta blocks

- For ldiskfs-osd, use fsfilt_ldiskfs_get_mblk()
- For kdmu-osd, we just return 2x new_blocks. It is impossible to predict how much data will be created.

How to use lod/osp if lov/osc is gone

- When setquota, getquota, adjust qunit, quota master uses lov/osc to do broadcast to quota slaves. When lod/osp is landed, we should use it instead.

- We need create virtual objects for this goal too and assign new index operations for them.


ORACLE®