Whamcloud

# OST Pool Quotas

Sergey Cheremencev, Nathan Rutman, Cory Spitz, Hongchao Zhang

DDN® STORAGE

# ABSTRACT


- With the deployment of heterogenous clusters containing a mix of flash OSTs and disk OSTs, administrators may need to restrict use of higher-performance OSTs on flash. Lustre's OST Pools feature enables the grouping of similar OSTs into performance tiers and the assignment of file layouts into these tiers. However, this feature does not support limits on the usage of more desirable / more expensive / smaller capacity tiers.

- Quota controls are the practical solution to impose administrative limits on a cluster's resources. However, currently-available Lustre quotas are limited to filesystem-wide limits on a per-user, per-group or per-project basis.

- In this presentation, we describe a new design for pool quotas that extends Lustre's capability to limit allocations within pools. We outline the feature's design and introduce a strategy of using multiple quotas within a cluster.

# Why Tiers Within Lustre?

Whamcloud

- Tiered storage is rapidly becoming more prevalent
  - 100% flash is not economical… yet
  - still need a capacity tier, flash will augment disk
- Lustre performs on flash – therefore, flash should exist in the Lustre namespace
  - Bandwidth at server of NVMe flash + network exceeds disk + network
- Result: heterogeneous storage and tiers within Lustre clusters

# OST Pools for Tiers

Whamcloud

- OST Pools already exist to manage groupings of OSTs
- OST Pools do not provide administrative controls!
  - OST Pools are only a convenience to concisely describe file layouts on OSTs
- OST Pools != tiers, we want permissions and controlled access to tiers
  - However, we don't want or need to invent a new concept or construct
  - Let's try to use OST Pools to manage tiers

*mgs# lctl pool_new lustre.flash*
*mgs# lctl pool_add lustre.flash lustre-OST[0-10]*
*client$ lfs setstripe --pool flash /mnt/lustre/myflashdir*
*client$ dd if=/dev/zero of=/mnt/lustre/myflashdir/myfile bs=1g count=100*
*dd: error writing '/mnt/lustre/myflashdir/myfile': No space left on device*

# Why Pool Quotas

Whamcloud

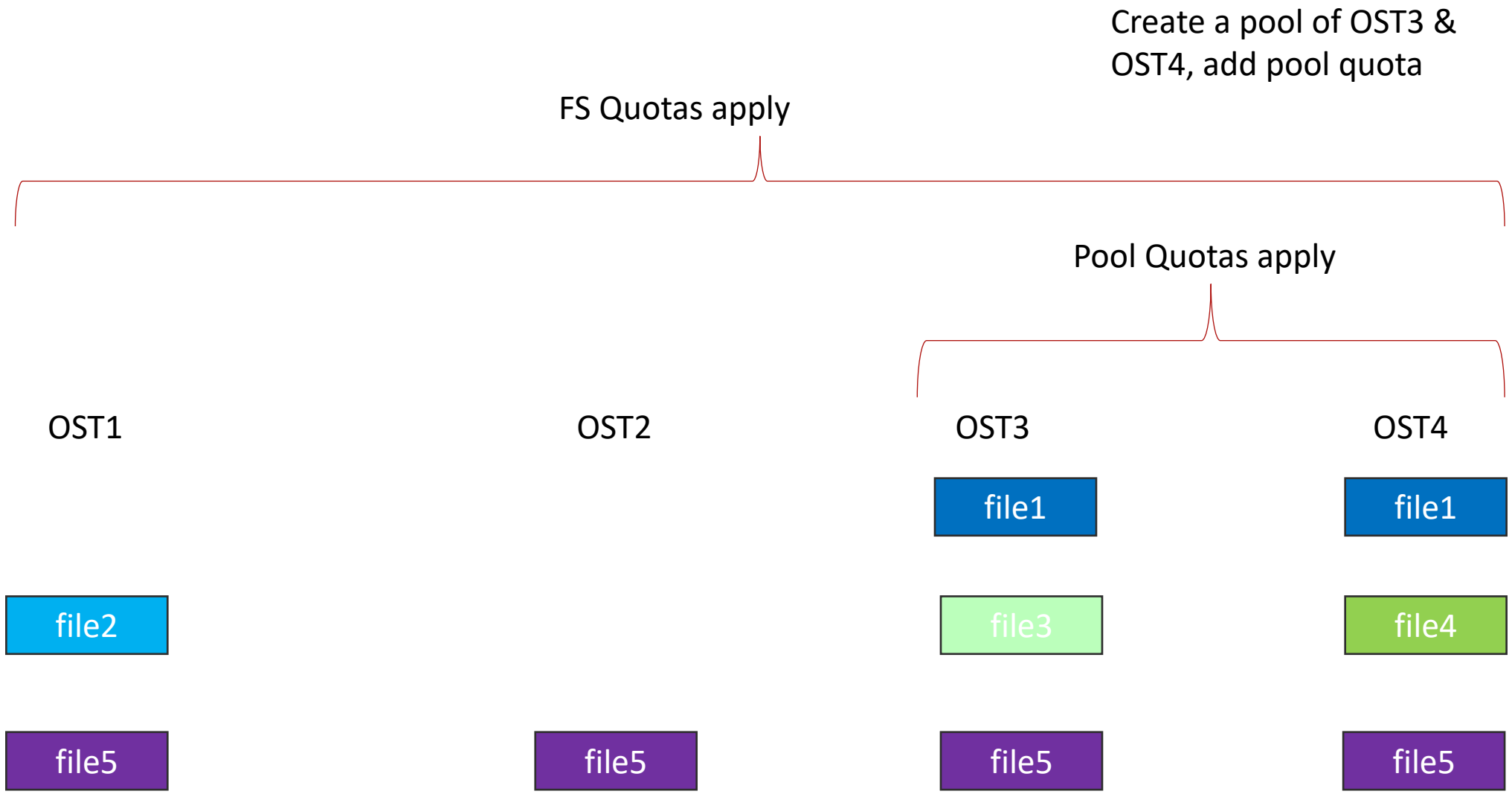
- It is too easy to fill a flash OST
  - Flash tiers are in high demand
  - Generally smaller capacity
  - Traditional quotas don't help
  - Sub-directory mounts don't help
  - PFL helps, but, it doesn't completely solve the problem
- Lustre has no method to limit the usage of desirable OSTs/tiers
- Quotas are the practical solution to impose administrative controls
- Problem: quotas are limited to filesystem-wide limits on a per-user/group/project basis
- Solution: Pool Quotas; that is, quotas for OST Pools

# Requirements for Pool Quotas

Whamcloud

- Basics: provide per-pool user/group/project capacity quotas
  - File quotas don't make sense for OST Pools
- Changes in pool definitions should dynamically affect the remaining pool quotas
  - OST may be added or removed from a pool at any time
- OST may be part of multiple pools, each with different pool quotas set
- Quota limit should be the minimum of all applicable limits
  - Of multiple OST Pools
  - Includes filesystem-wide quotas too
- Leverage existing quotas administrative interfaces, e.g.:
  - lfs setquota -u bob --pool flash --block-hardlimit=2T /mnt/lustre

# Quotas vs. Pool Quotas



# Implementation Approach

Whamcloud

- Start with the existing quotas implementation
- OSTs already request quota grants from quota master
- Current grant is determined on a filesystem-wide basis
- Aggregate resource used is tracked by each OST
- Teach the quota master(QMT) to consider OSTs within pool definitions!
  - grant requests from each OST must satisfy all limits
  - the min() of any fs-wide or per-pool limits

# Additional Details

Whamcloud

- Design approach means code changes only affect the MDS(QMT)
  - Files and objects do not "belong to" pools; OSTs belong to pools
  - OSTs don't need to understand which pool(s) they may be part of
  - Data written to an OST without a pool layout is still accounted for
- Easily cope with changes to pool definition
  - Pool definition changes can cause pool quota limit to be exceeded
  - In which case no new quota will be granted, but existing grants can continue to be used

# Administrator’s Example Usage

Whamcloud

- Create an OST Pool

  *mgs #lctl pool_new lustre.flash*

  *mgs #lctl pool_add lustre.flash lustre-OST[0-10]*

- Begin with a default Pool Quota

  # lfs setquota --pool flash -U --block-hardlimit 4M /mnt/lustre

- Then create or change Pool Quota as needed

  # lfs setquota -u quota_usr--pool flash --block-hardlimit 2g /mnt/lustre

  # lfs setquota -g quota_usr --pool flash --block-hardlimit 200g /mnt/lustre

  # lfs setquota -p proj --pool flash --block-hardlimit 20g /mnt/lustre

- Disable Pool Quota enforcement

  # lfs quotaoff --pool flash /mnt/lustre

# User's Example Usage

Whamcloud

- Review quota usage

```
# lfs quota -h -u quota_usr /mnt/lustre
Disk quotas for usr quota_usr (uid 60000):
     Filesystem  used  quota  limit  grace  files  quota  limit  grace
   /mnt/lustre   90M   200M   200M      -      1      0      0      -
uid 60000 is using default file quota setting

#lfs quota -h -u quota_usr --pool flash /mnt/lustre
Disk quotas for usr quota_usr (uid 60000):
     Filesystem  used  quota  limit  grace  files  quota  limit  grace
   /mnt/lustre   90M   100M   100M      -      1      0      0      -
```

- Upon EDQUOT user could have hit filesystem-wide or per-pool limit
- Remember, quota master grants min() of all limits
  - Which quota limit did you exceed?
  - As is today, but Pool Quotas add an extra dimension
  - All quota limits are available in report for inspection

# Things to Know


- OSTs may belong to multiple pools and pool membership can change
  - Objects count toward all pool limits and changes may push existing grant over limit
  - Approaching limit on one pool may slow performance for others (as qunit decreases)
- "No Pool Quota" not possible; Pool Quota of zero means "no limit"
- There are no MDT Pools, therefore Data on MDT space is not limited
- EDQUOT vs. ENOSPC
  - Even with Pool Quotas it is all too easy to get ENOSPC long before EDQUOT
  - Quotas are not space reservations

Whamcloud

# Thank You!

DDN® STORAGE