Ext4 with Lustre From DDN Storage

DataDirect Network Storage , Inc

Shilong Wang

2017/10/19

About DDN Storage

- **High Performances Storage Vendor(HPC)**
  - Powers 2/3 of TOP 100 Supercomputers
- **Provides whole Storage solutions including hardware and software.**
  - SFA, IME
  - GPFS, Lustre, WOS
  - Sell most of Lustre and GPFS than any other vendors..
  - Refers to http://www.ddn.com for more!

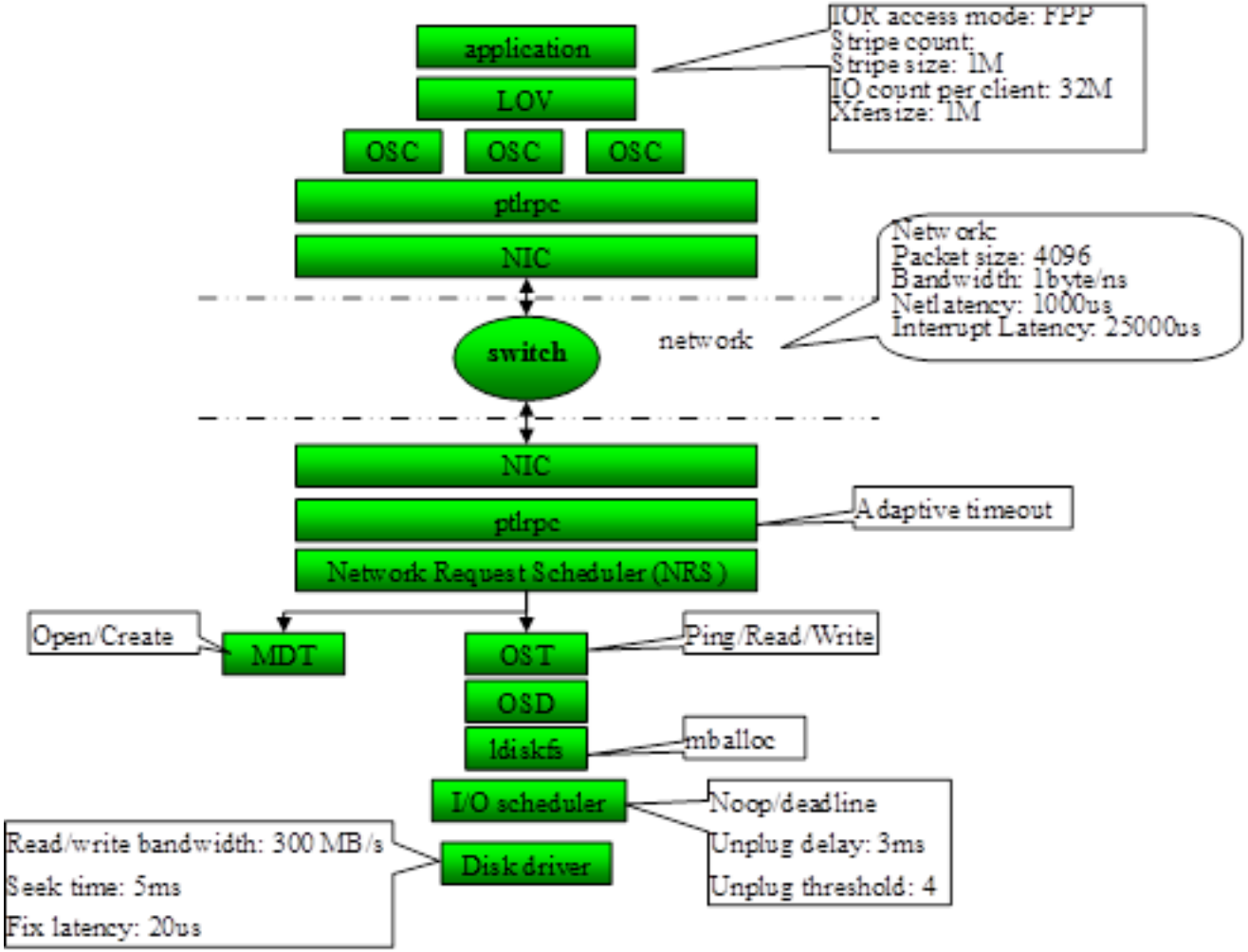
About our Developing Team

- We focus on Lustre Development and its extended software(monitoring, management tool, autotest system etc)
- Maintain DDN own Lustre
- Provides support to field engineer.
- 5 + 1 Developers works remotely.

Contributions to Linux ext4 and Lustre etc.

- **DDN grow up together with Lustre**
- **We contributed a lot to Lustre.**
  - 168 commits, 18K Lines change, many features like Ladvise, QoS, Lustre project quota. Large RPC support, Security etc…
- **Lustre contributed a lot to ext4**
  - Andreas Dilger(maintainer of ext4) is one of main architect of Lustre Filesystem
  - Lustre motivated many ext4 features: malloc, mmp, large EA, project quota etc.
- **DDN contributed to ext4 too.**
  - Third type of quota, similar idea of XFS named project quota
  - Metadata benchmark and improvements

How Lustre works with Ext4



Deployment of Ext4 with Lustre

- more than 300 systems are running DDN Lustre(Ext4 backend) in the world.
- Most of biggest DDN Lustre system running 8192+ clients mount a 25PB filesystem.

10 SFA14KE

4000 x 8 TB NLSAS

40 x raid6(8+2) x 10 system

Connected over 100GBps intel OPA network

Why Ext4?

- ExtN is supported in Lustre from the time when it is designed.
- ExtN is good compatibility, eg no extra efforts for upgrading from ext3 to ext4.
- Good Stability.(eg compared to Btrfs)
- Good performances compared to ZFS, especially Metadata performance.
- Better error isolation and mature fsck(very important!)
- XFS might be good replacement with Ext4, but adding another OSD backend needs many efforts.

Ext4 Problems

Filesystem Usage Limitation.

- File size limited to 16TB, very easy to hit limit with sparse file.
- Total number of inode still limited about 4 billion.
- Could not meet some customers requirements with one MDT, and it nearly reach people's expectation.

Solutions

- ZFS backend of Lustre
- Set Lustre stripe to split File to different OST.
- DNE feature which enabled cluster metadata server.
- It is time for ext5? ?

Ext4 problems

Ext4 Directory Limit

- Performance will drop a lot if one directory have more than 2 million files, XFS is not good at this too..
- 2 Level Hash Index Tree limit directory Size about 2GB
- Directory did not shrink with sub-dir/file removal which makes ENOSPC easier happen for large directory.
- We had several customers report reaching this limit.

Solutions

- 3 Level Hash index Tree Limitation have been merged by upstream recently(From Seagate motivated by Lustre)
- We walk around the problem by warning from the dmesg before reaching limit(50% and 75% of limit), we want to avoid large directory if possible for both performance and safety reasons.

Ext4 Problems

Bitmaps validations errors

- Some RHEL6 server happens very much
- Some OST will become read only and unavailable to be write.
- It is not easy to figure out reasons…

Solutions

- Corrupted block groups could be read only.
- Free counter reset to be zero, all allocations and de-allocation will be not allowable.
- Error state recorded in the superblock, e2fsck will be aware of it in the next run time.
- Just throw ext4 warning instead of error(Not RO in default)
- One of Android Team used same approach.

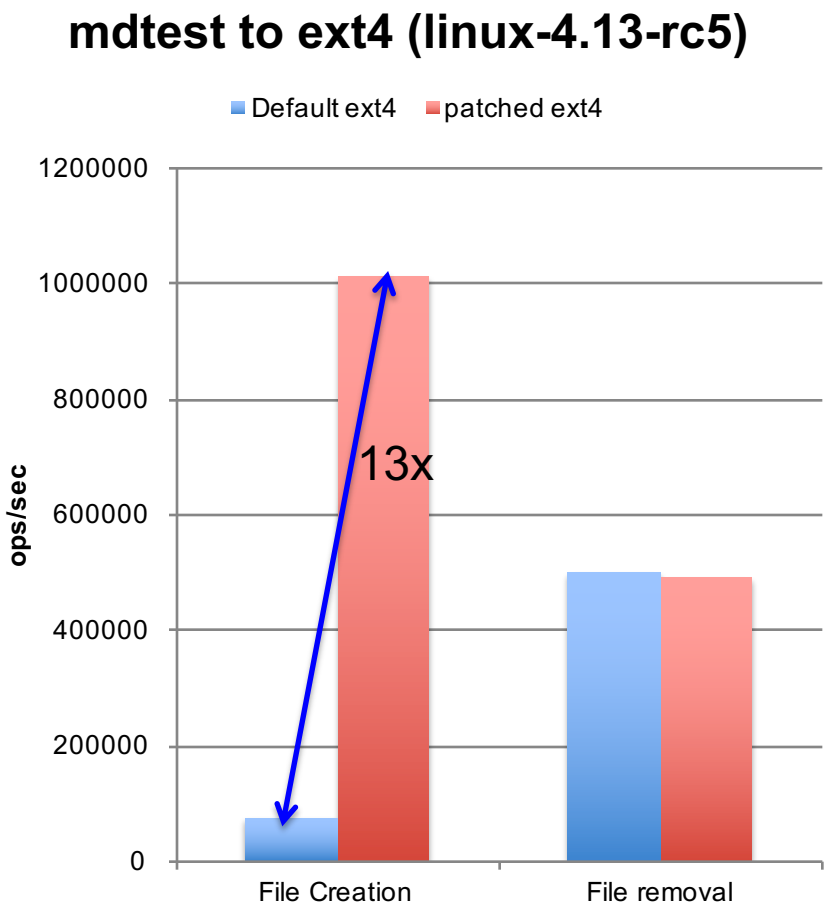
Ext4 performance

- **Fixed and merged upstream kernel (4.14)**

  Wang Shilong (2):
  ext4: cleanup goto next group
  ext4: reduce lock contention in __ext4_new_inode

- **13x performance improvement on file creation**
  - Run mdtest to ext4 directly
  - Unique directory/file operations
  - Quota disabled

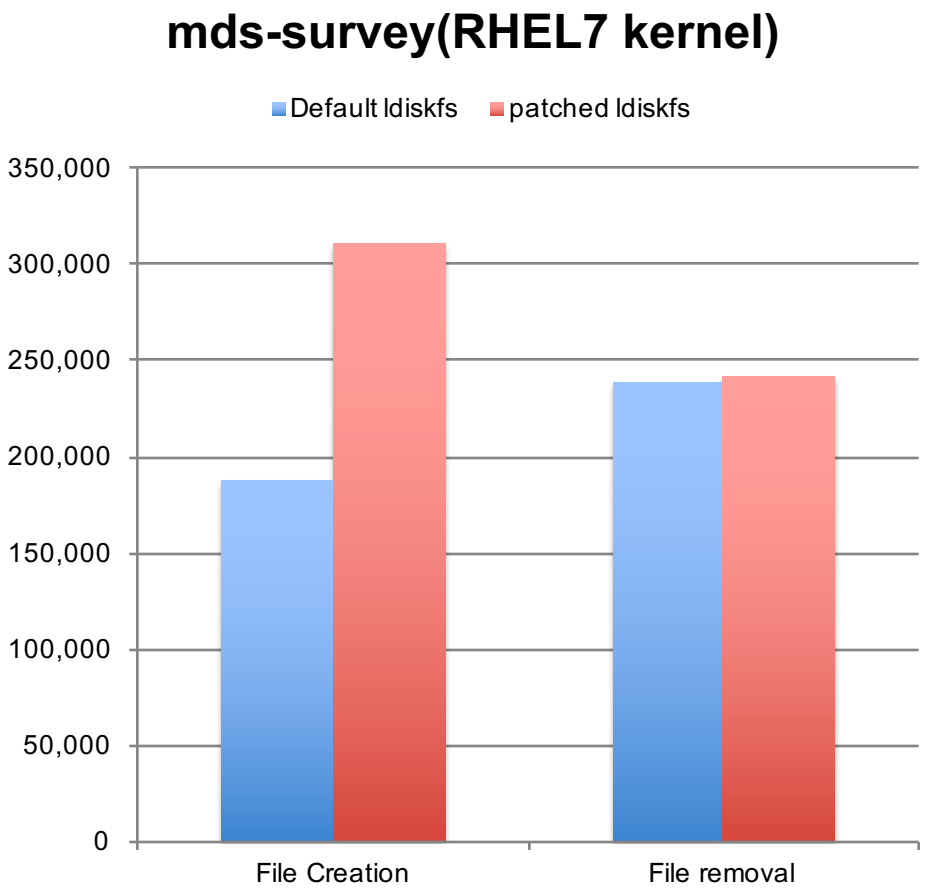
Test Configuration
1 x Xeon(R) Platinum 8160
128GB DDR4 memory
2 x RAID1 SSD with SFA7700/FC
1 x FDR Infiniband
Tested 1.28 Milition files with mdtest



Backport patch to RHEL7.3 kernel for Lustre-2.10

- **66% performance improvements on RHEL7**
  - Backported patch against ldiskfs in RHEL7 kernel
  - Run mds-survey to mount MDT
  - Lustre default quota setting (Enable user/group quota, but project quota is not enabled)

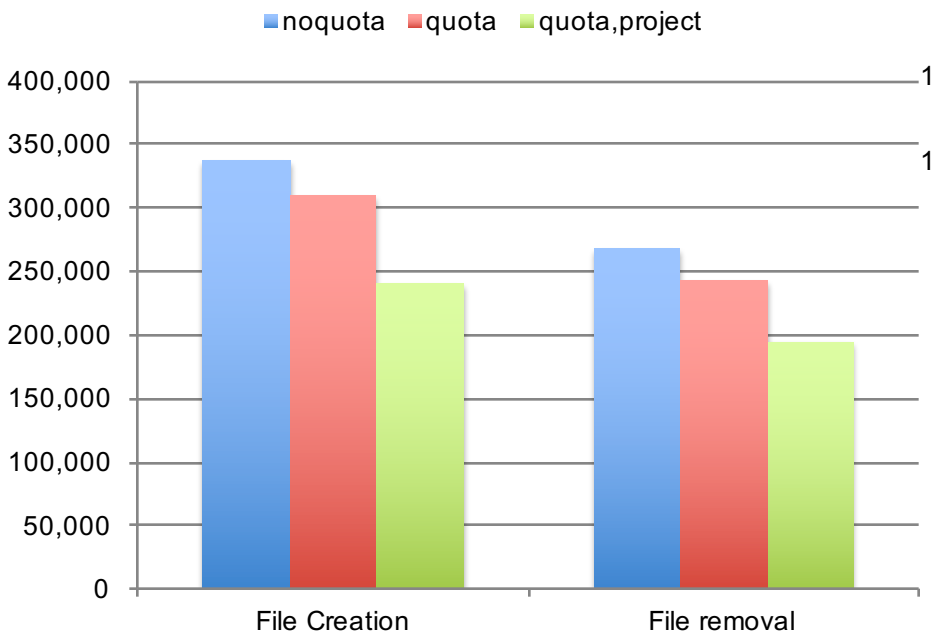
Test Configuration
1 x Xeon(R) Platinum 8160
128GB DDR4 memory
2 x RAID1 SSD with SFA7700/FC
1 x FDR Infiniband
Lustre-2.10.1RC1/RHEL7.3
Tested 1.28 Milition files with mdtest



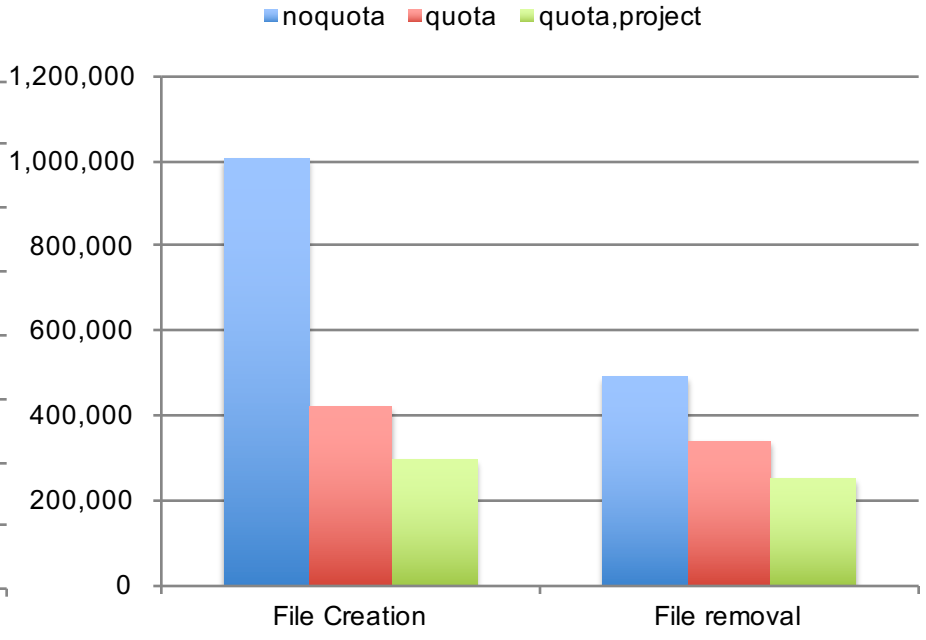
Quota scalability problem

- **File creation/unlink affects enabling quota**
  - Same behaviors on RHEL7 and upstream kernel
  - Project quota gives additional performance penalty

mds-survey(RHEL7 kernel)



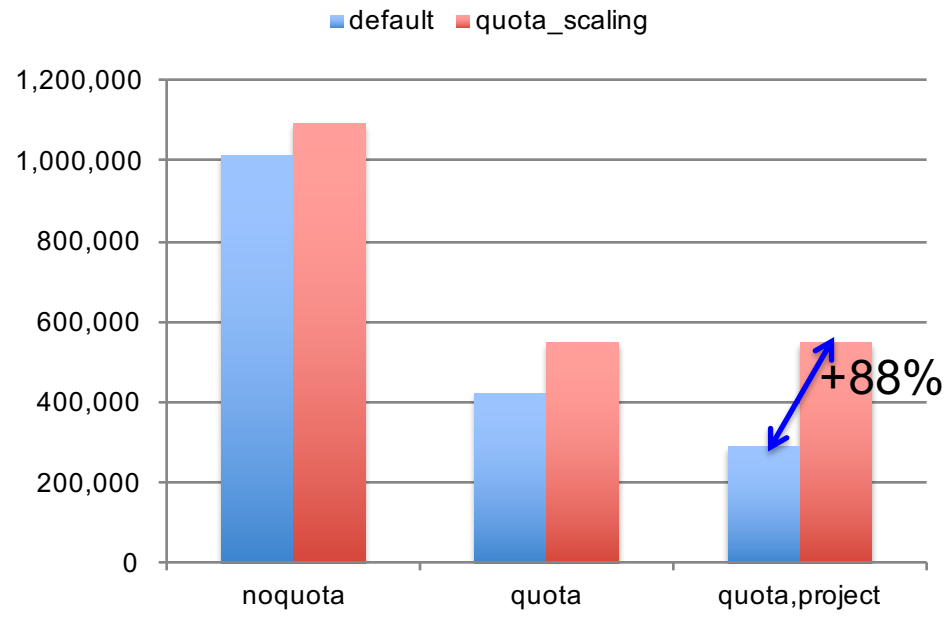
mdetst to ext4 (linux-4.13-rc5)



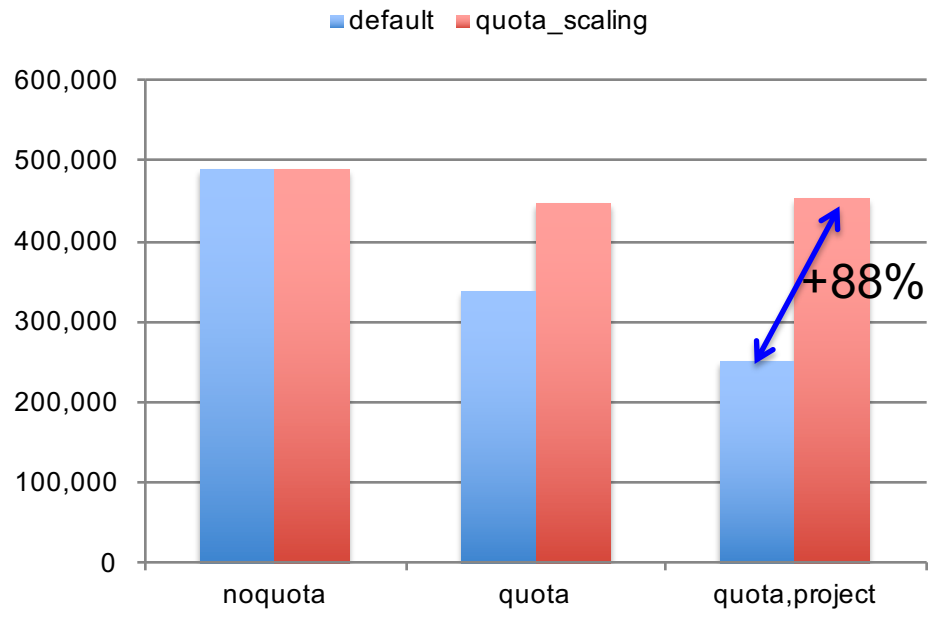
Quota scalability improvements in Ext4

- **New quota scaling patch introduced in upsteram kernel**
  - Tested new Jan Kara's quota scaling patches (merged in 4.14)
  - Huge performance gains when quota enabled

File Creation



File Removal



Thank you!