Lustre User Group 2009

Workshop

Andreas Dilger
Sun Microsystems

Overview

- Ext4 features
- Multi Mount Protection
- RAID tuning
- OST Pools
- File size and Glimpse
- Timeouts and Eviction

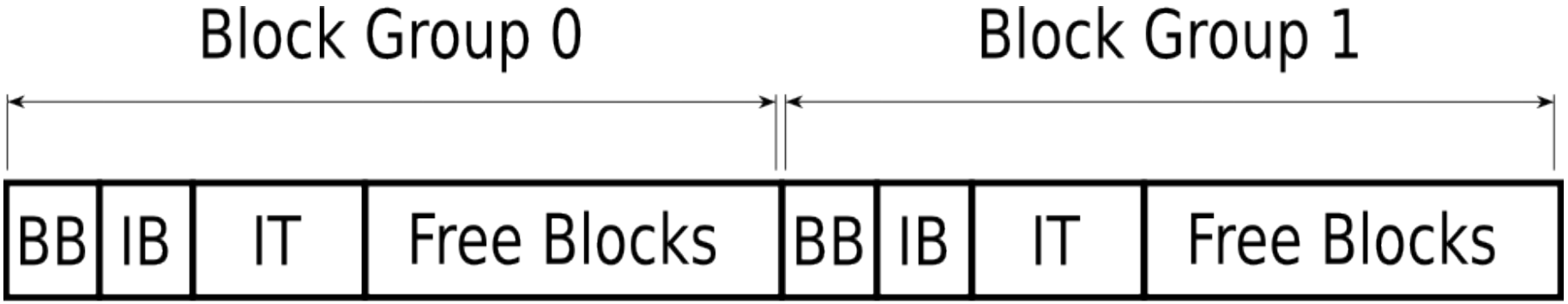
Ext4 Features

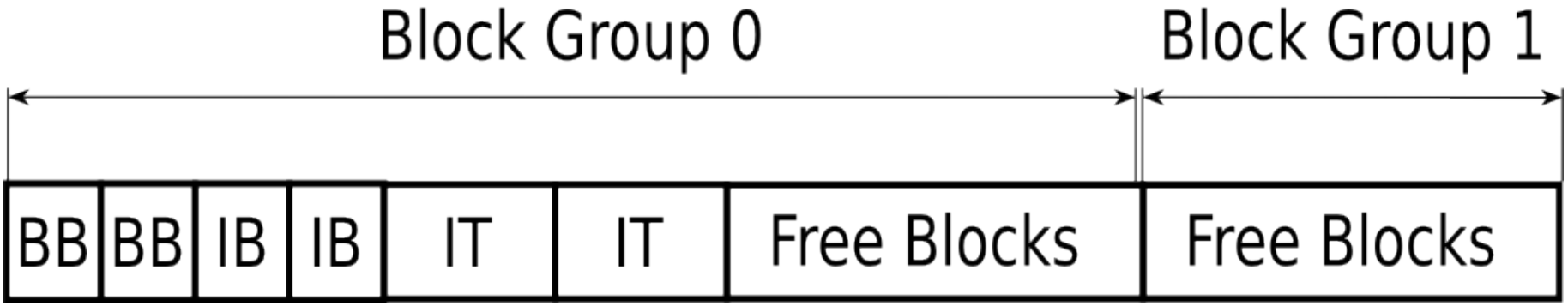
- Fast extended attributes (1.4)
- Extent support (1.4)
- Multiple block allocator (1.4)
- Inode versioning (VBR,1.8)
- Uninitialized groups for faster fsck (1.6.5)
- Delayed block allocation (all)
- File extent map (FIEMAP, 1.8)
- Nanosec timestamp (1.6.5)
- Flexible Inode Placement
- Larger Files (> 2TB)
- Persistent file preallocation (sys_fallocate)
- Larger file system (>16TB)

Ext4 - Flexible block groups

- Remove restrictions on the location of bitmaps and inode table.
- Bitmaps and inode table can group to simulate a block group larger than what a single bitmap can address.
- Allow for new allocation strategies that exploits the new meta-data allocation.
- Take advantage of new disk data densities to avoid costly seeks.
- Help keep keep as much uninitialized data as possible when using uninit_bg to keep fsck times low.

Ext4 – Flexible Block Group Layout





Ext4 - 16 Threads FFSB results

Op	Ext4	Ext4(flex_bg)	% change
read	96778	119135	18.76%
write	143744	174409	17.58%
create	1584997	1937469	18.19%
append	46735	56409	17.14%
delete	93333	113598	17.83%
Total	6514.51	7968.24	18.24%

Ext4 – e2fsck improvements

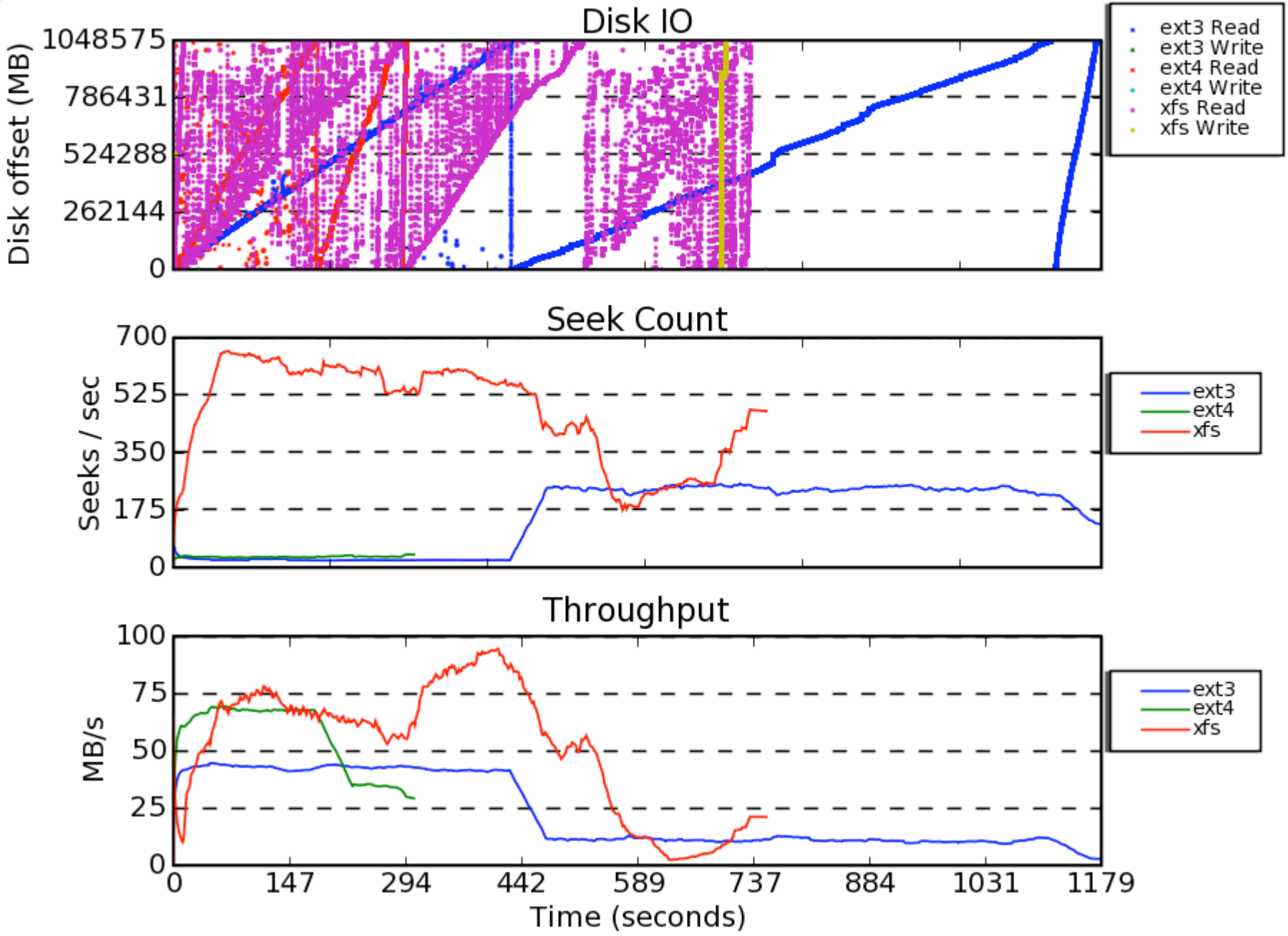
- Skip uninitialized bitmaps.
- Skip unused inodes
- Checksum group descriptors
- TODO: support for > 16TB filesystems
- TODO: fine-grained journal checksum

mke2fs -O uninit_bg /dev/sda

tune2fs -O uninit_bg /dev/sda

e2fsck -fy /dev/sda

fsck comparison: 50M inodes, 1T fs



Ext4 – current state of affairs

- Available in RHEL5.3 update
- Available in SLES11
- Will be available in RHEL6
- Lustre ldiskfs ported to ext4 baseline
- Lustre will use if available in 1.8.x

- Ldiskfs has MMP outstanding

Ext4 – Multi Mount Protection

Important for failover nodes

Delays mount/e2fsck by 10s of seconds

mke2fs -O mmp /dev/sda

tune2fs -O mmp /dev/sda

tune2fs -F -O ^mmp /dev/sda

LDISKFS-fs warning (device sda): Device is already active on another node.

LDISKFS-fs warning (device sda): MMP failure info: last update time: 1239822128, last update node: lin-cli1, last update device: sdb

RAID tuning

Inform filesystem of RAID layout
Delays mount/e2fsck by 10s of seconds

e.g. 64kB RAID6 6+2 (4kB blocks, 4kB page)

64kB/stride / 4kB/block = 16 block/stride
64kB * 6 stripes / 4kB/block = 96 blocks

mke2fs -E stride-size=16 -E stripe-width=96
tune2fs -E stripe-width=96

lctl conf_param lustre.osc.max_pages_per_rpc=192

OST Pools

Named groups of OSTs

Useful for heterogeneous storage

- Advisory OST selection only

```
mgs# lctl pool_add lustre.slow lustre-OST[0-13]
mgs# lctl pool_add lustre.fast lustre-OST[14-22]

lfs setstripe -c 2 -p fast /mnt/lustre/mydir
lfs setstripe -c 2 -p slow /mnt/lustre/yourdir

lctl get_param -N lov.lustre*.pools.*
lctl get_param -n lov.lustre*.pools.fast
```

THANK YOU

<adilger@sun.com>