FSTRIM

Marc Skinner
Principal Solutions Architect
mskinner@redhat.com

The SSD problem?

- Solid State Drives are NAND backed storage
- Divided into blocks then pages
- A block is the minimum unit that can be erased
- A page is the minimum unit that can be written
- Blocks can not be erased if they contain pages that are not empty
- Device will need to move pages of data to free up blocks to be erased
- Overtime, device becomes more and more fragmented
- Device performance decreases as page re-allocation increases

TRIM – the solution

- No TRIM
  - When OS deletes file – it marks index as file deleted, but doesn't notify SSD
  - Device will continue to degrade, getting worse and worse over time
- With TRIM
  - OS notifies SSD to blocks/pages that are empty
  - Internally defrags free/used blocks/pages for optimization

What the FSTRIM?

- A tool for “defragging” solid state drives
- Provided by util-linux rpm
- Is my disk trimable?
  - cat /sys/block/sda/queue/discard_max_bytes
    - If non-zero = YES
- Real time vs batch
  - Real time
    - Add “discard” option in /etc/fstab
    - Adds overhead during each file operation
  - Batch
    - Add a weekly or daily fstrim task in cron
    - Batches the operation to off hours for a given mount point
    - fstrim /home

TRIM and layers

- For TRIM to work
  - Must be enabled on each layer
    - File system
    - Encryption Layer
    - Software Raid – mdadm
    - LVM
    - Device

File System Support

- GFS2, ext4, ext3, XFS

LUKS – Linux Encryption

```
$ cat /etc/crypttab
$ luks-a2af767b-c62f-4123-acef-20de4e9f3bab UUID=a2af767b-c62f-4123-acef-20de4e9f3bab
none luks,discard

$ dracut -f

$ reboot

$ dmsetup table /dev/dm-11 --showkeys
0 104853504 crypt aes-xts-plain64
432140ae96e29cbc4cd1c4f56d686da706d467dc322b1fce2c7b12558ddb46a5443f266977c50dee92eca96
701767f10b844bf7a6baeb1161f45514cb59d450e 0 253:6 4096 1 allow_discards
```

LVM Support

```
$ vi /etc/lvm/lvm.conf
issue_discards = 1

$ dracut -f

$ reboot
```

Software Raid – mdadm/dm-raid

- RHEL 6.5
  - RAID 0/1/10 supported
- RHEL 6.6 and RHEL 7
  - Will add RAID 4/5/6

Real mode setup

```
$ cat /etc/fstab
/dev/mapper/vg_t530-lv_iso   /ISO        ext4    defaults,discard        1  1
```

Batch mode setup

```
$ vi /etc/cron.weekly/dofstrim
$ cat /etc/cron.weekly/dofstrim
#! /bin/sh
for mount in / /boot /home; do
        fstrim $mount
done
```

Real Time vs Batch

```
$ mount -o discard /dev/vg0_hal/lv_discard /DISCARD/

$ mount /dev/vg0_hal/lv_no_discard /NO_DISCARD/

$ mount | grep DISCARD

/dev/mapper/vg0_hal-lv_discard on /DISCARD type ext4
(rw,relatime,seclabel,discard,data=ordered)

/dev/mapper/vg0_hal-lv_no_discard on /NO_DISCARD type ext4
(rw,relatime,seclabel,data=ordered)
```

```
$ cat /home/mskinner/discard_test.sh

#!/bin/bash

echo "Testing $1"

cp -r /usr/lib $1

cp -r /usr/src $1

cp -r /usr/share $1

dd if=/dev/zero of=/$1/BIG bs=4k count=1M

rm -rf /$1/BIG

rm -rf /$1/usr/share

rm -rf /$1/usr/src

rm -rf /$1/usr/lib

echo "Done"
```

Real Time Test

```
$ time /home/mskinner/discard_test.sh /DISCARD
Testing /DISCARD
dd: error writing ‘//DISCARD/BIG’: No space left on device
118181+0 records in
118180+0 records out
484065280 bytes (484 MB) copied, 3.05687 s, 158 MB/s
Done

real    1m46.878s
user    0m1.171s
sys     0m21.443s
```

Batch Test

```
$ time /home/mskinner/discard_test.sh /NO_DISCARD
Testing /NO_DISCARD
dd: error writing ‘//NO_DISCARD/BIG’: No space left on device
118128+0 records in
118127+0 records out
483848192 bytes (484 MB) copied, 2.5086 s, 193 MB/s
Done

real    0m55.602s
user    0m0.511s
sys     0m9.592s

$ time fstrim /NO_DISCARD

real    0m0.135s
user    0m0.000s
sys     0m0.002s
```

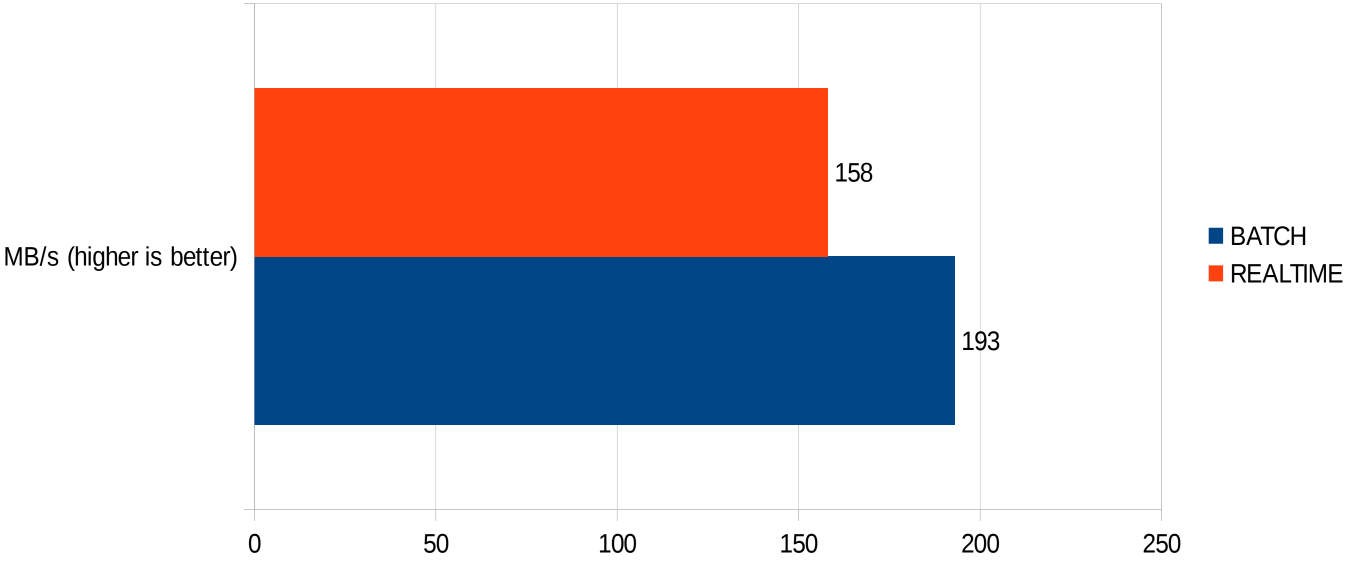
Real Time vs Batch Results

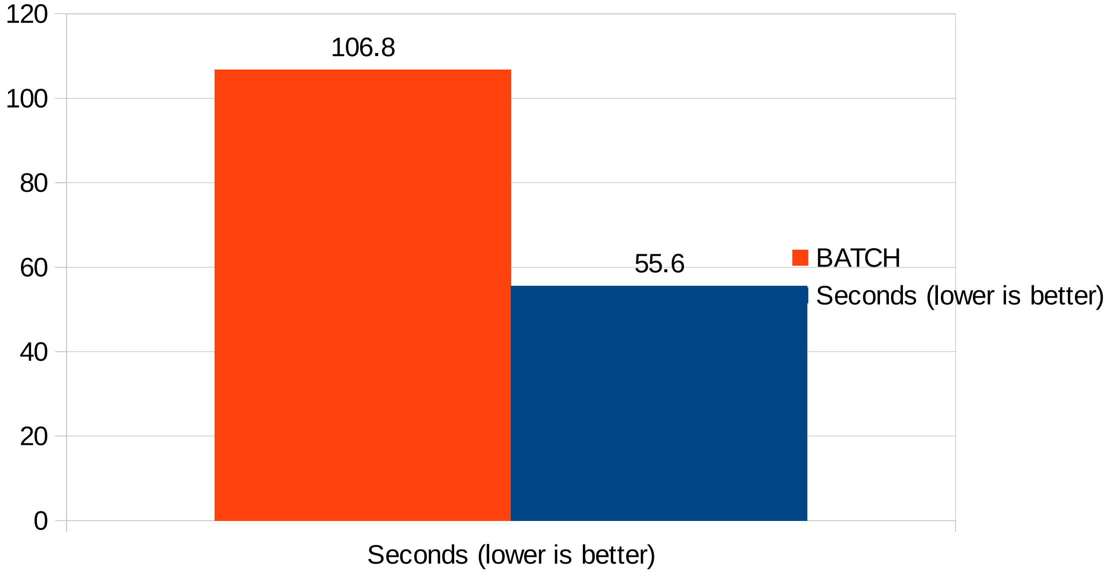
Real Time @ 158 MB/s

```
real    1m46.878s
user    0m1.171s
sys     0m21.443s
```

Batch @ 193 MB/s

```
real    0m55.602s
user    0m0.511s
sys     0m9.592s
```





Questions?