Linux_IO_Tuning

Jump to bottom

Kyle Anderson edited this page on Mar 12, 2016 · 1 revision

Table of Contents

Disks

Read Ahead

Check out the read-ahead of your disks:

```
root@archive:~# blockdev  --report
RO    RA   SSZ   BSZ   StartSec            Size   Device
rw   256   512  4096          0   2000398934016   /dev/sda
```

You can set it with:

```
 blockdev --setra N /dev/sda
```

Useful if you know that a larger read-ahead will help your application. (It will not if you are doing mostly random reads/writes.

You can only know if it helps if you test.

Same as:

```
/sys/block/sda/queue/read_ahead_kb
```

Notice the different units.

Queues / NCQ

Make sure NCQ is supported:

```
hdparm -I /dev/sda | grep NCQ
```

The ammount of commands in flight can make a big difference, especially on iSCSI links. Make sure your values are close to these:

```
4096 > /sys/block/sda/queue/max_sectors_kb
64 > /sys/block/sda/queue/nr_requests
31 > /sys/block/sda/device/queue_depth
```

Software Raid

Read Ahead

Depending on what raid level you have and stripe settings, Linux may be fetching a lot of data for a small random read. It *may* be helpful to set a larger read-ahead:

```
blockdev --setra 65536 /dev/md0
```

Stripe Cache

The stripe cache can consume a lot more ram for a lot more performance.

```
cat /sys/block/md0/md/stripe_cache_size
256
echo 32768 > /sys/block/md0/md/stripe_cache_size
```

Try this script for determining your best cache size if memory is no problem:

```
#!/bin/bash
for cache_size in 256 512 768 1024 2048 4096 8192 16834 32768; do
 for i in {1..3}; do
```

```
 echo ${cache_size} > /sys/block/md0/md/stripe_cache_size
 sync
 echo 3 > /proc/sys/vm/drop_caches
 echo "stripe_cache_size: ${cache_size} (${i}/3)"
 # for write
 dd if=/dev/zero of=/dev/md0 bs=3145728 count=5460
 # for read
 dd if=/dev/md0 of=/dev/null bs=3145728 count=5460
 done
done
```

See: http://peterkieser.com/2009/11/29/raid-mdraid-stripe_cache_size-vs-write-transfer/

More

- Lots of stuff: https://raid.wiki.kernel.org/index.php/Performance

File Systems

Chunk / Stripe / Stride

Ext performance can be increased if you make it aware of your raid layout. Definitions:

- Stride (How many consecutive filesystem blocks on a raid chunk on one disk): chunk size / filesystem block size (`mdadm --detail /dev/md0 | grep Chunk` divided `tune2fs -l | grep Block`)
  - Example: 512k/4k = 128 blocks
- Stripe Width (How many filesystem blocks do you need to fill one raid "unit"): stride * (num raid disks - num parity disks)
  - Example: 128 * (4 - 1) = 384 (raid5 with 4 disks)
- See: https://raid.wiki.kernel.org/index.php/RAID_setup#ext2.2C_ext3.2C_and_ext4

Ext

```
# tune2fs -E stride=STRIDE,stripe-width=WIDTH
tune2fs -E stride=128,stripe-width=384 /dev/md0
```

XFS

```
mkfs -t xfs -d sunit=128 -d swidth=384 /dev/md0
```

You must either set that at mkfs time or put it in fstab?

Category:RaidCategory:Disks

Clone this wiki locally