

Ten tips for better AoE performance and reliability in Linux

Objective

Congratulations! You have chosen to use our Coraid storage products. Here are some tips on how to test our products and how to get the best performance out of our Coraid SR and LD product lines. These tips are designed to help you to increase the overall reliability and performance of Coraid's AoE storage products.

Solutions

1. Disk performance

Only enterprise- or raid-class disks that are designed for constant 24/7 use are supported - please visit our [disk compatibility table](#) for more information. Also, we suggest that you use similar if not the same disks within the chassis. This is to avoid performance problems found by using a comparatively slower disk within a LUN. If you use a slow disk with a group of faster disks, the overall performance of the LUN will be degraded as the limiting factor will be the slower disk.

2. Use previous performance metrics as a reference

To give the performance of the Coraid SR/LD proper context, please reference our [performance analysis documentation](#) for points of reference. We have also written a performance program called [ddt](#) which will test throughput. We recommend that you use `ddt` for getting accurate performance measurements. To obtain the most accurate results we suggest that you set the total I/O for any `ddt` test to 1.5 times total RAM installed in the machine being used as the AoE initiator. Your mileage may vary depending on your internal RAID configuration as well as your network configuration; for more information on `ddt` please see its man page.

```
[root@CentOSServer2 ddt-8]# ddt -t 3G /mnt
Writing to /mnt/ddt.19533 ... syncing ... done.
sleeping 10 seconds ... done.
Reading from /mnt/ddt.19533 ... done.
3072MiB   KiB/s CPU%
Write    378915  20
Read     512891  30
```

3. Use a 9K MTU size

For best performance, all devices on the network that connects to the Coraid SR/LD should be using a 9K MTU size. AoE equipment works best with this MTU size because it lowers the amount of CPU cycles needed for processing each individual packet on both the target and the host. By lowering the total amount of packets across the network, it can reduce congestion and processing overhead. This is a general suggestion and may not be suitable or viable for some particular networking equipment. Only performance testing on your specific networking hardware can determine if this is a viable MTU size.

We also suggest that you use “server-class” NICs and switches that can support more than a 9K MTU size, as many times the advertised maximum MTU size is not truly practical. If a 9000 MTU is not possible, the next best situation is an MTU of 8192 or 4200, depending on which switch equipment you have. Sometimes this is set manually on the switch. In Linux you can set then verify the MTU size for your network interfaces with the *ifconfig* command.

4. Format your devices with XFS

We recommend that you format the LUNs that you create with the XFS file system. It has proven to have the best performance compared to other Linux filesystems (ext3, reiserfs, jfs) and does very well with large files. It will however use a large amount of virtual memory. We recommend that you use the most current version of the XFS filesystem as there were some particularly heinous bugs in previous versions. Here is an example of a mounted LUN formatted with XFS within the context of the other mounted filesystems.

```
[root@CentOSServer2 ~]# mount | grep -e xfs -e ext3
/dev/sdb1 on / type ext3 (rw)
/dev/etherd/e2.7 on /mnt type xfs (rw)
```

5. Dedicate a separate LAN to AoE traffic

AoE traffic can live on any Ethernet network. But as traffic on the Ethernet network increases, so does the latency of the data transfer. This is why, under most circumstances, we suggest that you dedicate a separate LAN to AoE traffic. This means to dedicate separate switches or VLANs to the devices and their host servers. This should not increase the need for expensive additional hardware, as AoE traffic does well on relatively inexpensive switches such as the Dell Powerconnect 6224.

6. Wait for LUN rebuild

When a RAID-based LUN is first built, or rebuilt, parity needs to be written to the disks. In such cases, the performance of the LUN is degraded even though it fully accessible from the network. If you are doing performance testing, under most tests a large RAID 5 has the best performance, but we suggest that you only do performance testing after parity has been written to the device. In the following example a RAID 5 LUN being initialized. The when command will tell you when the RAID parity write will finish. Notice how internal RAID rebuilds tend to be very fast (in this example, around 510MB/s).

```
LD shelf 2> make 0 raid5 2.4-15
beginning building parity: 0.0
LD shelf 2> list -l
0 3300.759GB offline
0.0 3300.759GB raid5 initing 0.12%
0.0.0 normal 300.069GB 2.4
0.0.1 normal 300.069GB 2.5
0.0.2 normal 300.069GB 2.6
0.0.3 normal 300.069GB 2.7
0.0.4 normal 300.069GB 2.8
0.0.5 normal 300.069GB 2.9
0.0.6 normal 300.069GB 2.10
0.0.7 normal 300.069GB 2.11
0.0.8 normal 300.069GB 2.12
0.0.9 normal 300.069GB 2.13
0.0.10 normal 300.069GB 2.14
0.0.11 normal 300.069GB 2.15
LD shelf 2> when
0.0 510158 KBps 1:52:21 left
```

7. Connect your storage appliance to a UPS

By connecting the Coraid SR/LD appliance as well as the host server and networking hardware to an uninterpretable power supply (UPS), you are increasing the reliability of the devices. As there is currently no battery-backed cache on the Coraid box, power loss equals possible data loss. UPSs are a way to insure the data is not lost and is a way to withstand interruptions in power delivery caused by natural events or unintended disconnections.

8. Use LUN masking

If you are not using a clustered filesystem like GFS or CXFS, we recommend that you use LUN masking. This is a way to prevent other hosts to write to a non-clustered filesystem which is intended for only one server. By default, the LUN's mask list is empty permitting anyone access it on the network. The MAC address masking is set for each individual LUN. Here is an example to limit access to the LUN number 7 to only one MAC address, and block all others. This tool is similar to the access control model of /etc/hosts.allow, where the hosts specified are permitted and the rest are denied.

```
LD shelf 2> mask
7
LD shelf 2> mask 7 +00:30:58:61:C5:05
LD shelf 2> mask
7 00305861c505
```

9. Use the latest driver

As bugfixes and performance enhancements are part of our driver development, we recommend that you use the latest version of our driver and tools. You can find out the version of your AoE using the *aoe-version* command. The newest drivers will be available [from our website](#). Although upgrading the driver is not always the first step at increasing performance, it may be a way to upgrade performance and reliability if other steps have not helped.

```
[root@CentOSServer2 ~]# aoe-version
aoetools:    64
installed aoe driver: 64
running aoe driver: 64
```

10. Reloading the aoe module

If you have made any changes, and still see issues, it is recommended that you reload the aoe module. You can use an aoe-init script, or use `#rmmod aoe` and then `#modprobe aoe`. If you need installation tips for specific distributions please visit our [Linux support page](#).

Additional Information

For more information please reference the [EtherDrive HOWTO](#) and our [Linux support page](#). If you have questions about the procedures outlined here or comments/suggestions regarding the document, please contact us at support@coraid.com.