SR Redundancy and Throughput in Linux

S. Hopkins, M. Ennis
Coraid, Inc.

Summary. It is often requested that the SR be used in a configuration that permits simultaneous use of the network interfaces for aggregated throughput and redundancy. The standard solution for this is 802.3ad link aggregation, commonly known as bonding. Since AoE is its own Ethernet type and does not use IP, it has proved to be incompatible with some vendor implementations of 802.3ad bonding. To overcome this limitation, the Linux AoE driver implements a feature called multipathing to provide for round robin load balancing of multiple AoE targets on multiple client interfaces. This feature is available in 2.6 kernel Linux drivers starting with aoe6-33. Linux is currently the only system with an AoE driver implementing the multipathing feature.

There are considerations that must be made when deciding to use the second port of the SR in conjunction with multipathing. The remainder of this paper discusses issues specific to certain SR models and provides general guidelines for multipathing usage. When viewed from the rear, the first port on the SR is the left port and the second port is the right port.

SR1520 / SR420

Due to a hardware limitation in the SR1520 and SR420 series appliances, the second port provides slower throughput than the first when used for AoE. In general, Coraid recommends the second port of these appliances be used for CEC (Coraid Ethernet Console) management and not for AoE. There is no performance benefit to using both ports of these appliances with the aforementioned multipathing feature; there is in fact a performance penalty. The only benefit of using the second port for AoE is to have an additional path to the AoE storage in the event of network failure. This benefit cannot be obtained without paying the cost of reduced throughput.

The main cause for the limitation of the second interface lies in its location on the PCI bus; it must compete with the SATA cards for a limited bandwidth to memory. As a result, the second interface cannot sustain burst rates of packets as it cannot get them into memory fast enough to avoid overflowing its network receive FIFO. The result is that packets are frequently dropped. This problem is most evident when performing writes, as it is then that the receive FIFO is most stressed.

AoE drivers starting with aoe6-33 have a mechanism to reduce the number of outstanding packets a target can accept for cluster use of AoE targets. This feature helps with the receive problem of the second interface and reduces the number of retransmits that would otherwise occur.

If the SR is configured to export one lblade and there is only one host using the SR, then the second interface limitation will only cause a minor performance penalty when used with multipathing as shown in the following tables. If, however, there are multiple lblades being exported or there are multiple clients accessing the lblade, then there is a potential for the second interface to heavily penalize throughput for all lblades / users. Coraid strongly recommends not using the second interface for AoE in these configurations.

Figures A and B contain statistics achieved by averaging the results of three independent runs of `ddt` for the given SR configuration. `Ddt` is a simple tool that writes and reads to a file through a file system to determine the throughput capability of the file system and underlying storage. For a full description of `ddt`, please see the document titled SR Performance Analysis at the SR support page[1]. In the figures that follow, #IF is the number of interfaces used and all throughput values

1 See Appendix A for URL

are in KiB/s. Statistics are presented using jumbo[2] and standard Ethernet frame sizes.

The Linux client used for these tests was stuart, a dual CPU (dual core), system with 1GiB of RAM. The Linux kernel was 2.6.18, and the AoE driver was aoe6-44. The .config file for stuart's kernel is available for download1. The client network cards used the Intel 82546GB controller. For each configuration above, an XFS file system was placed on the resulting AoE device. The file system was mounted, and ddt was run against this mount point.

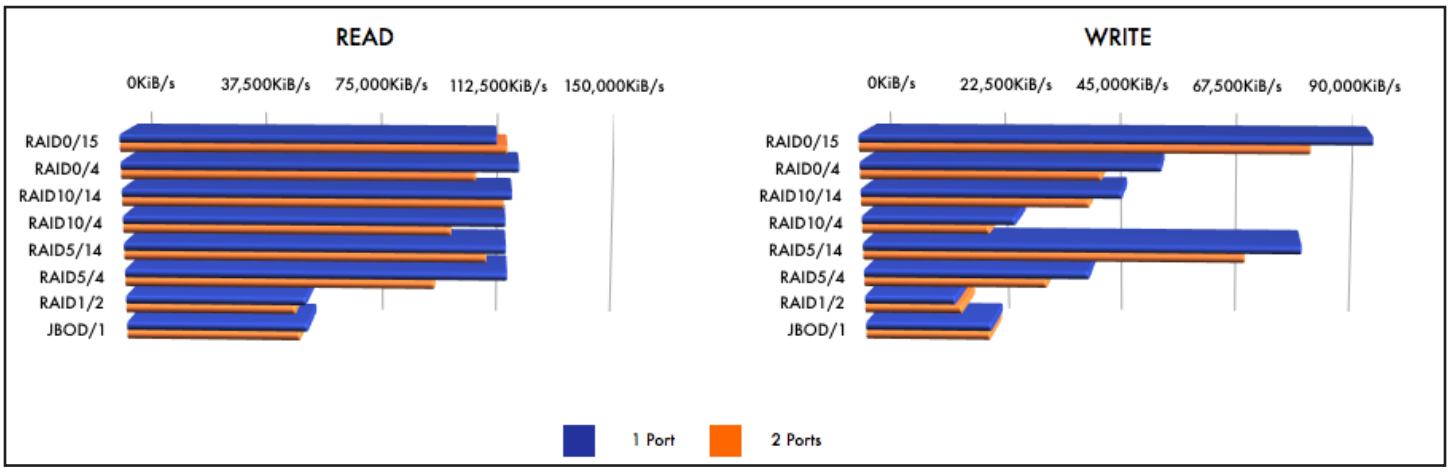
Both SR appliances were tested running the 20070111 release. For the dual port tests, two network ports on stuart were directly connected to both ports of the SR. For the single port tests, one network port on stuart was directly connected to the leftmost (first) port of the SR.

SR1521

The SR1521 appliance was introduced to overcome the limitations previously discussed with the SR1520. The SR1521 is capable of fully saturating both interfaces when used with the multipathing feature. The remaining throughput limitations are believed to be in software; Coraid is undergoing a full analysis of the SR RAID subsystem to identify bottlenecks.

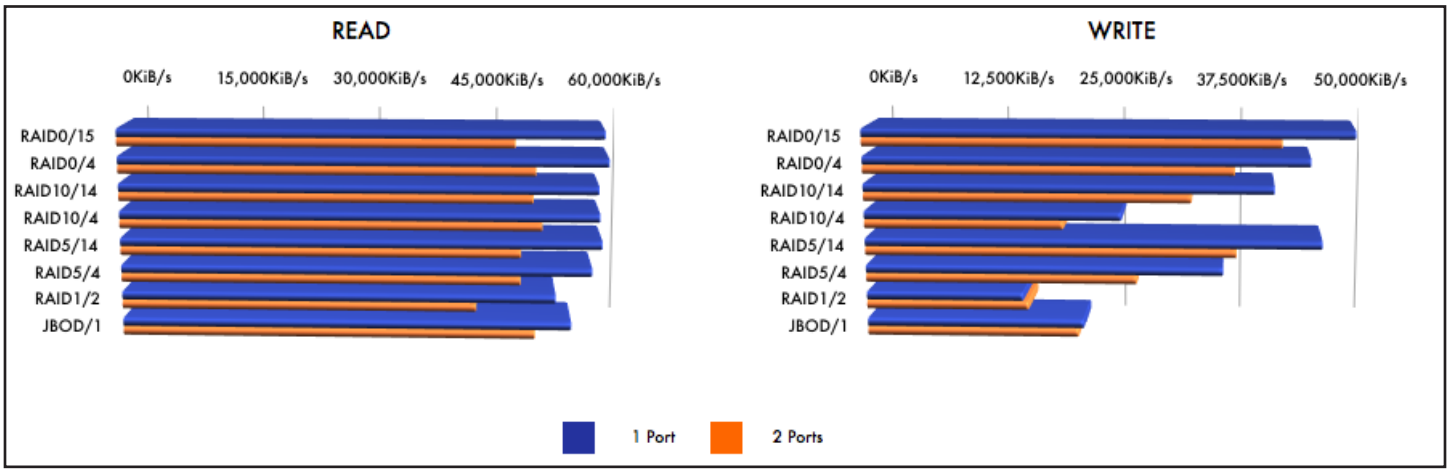
Figures C and D contain `ddt` results for the SR1521. All configuration parameters (Linux client, SR version) are identical for the SR1521 as for the previous SR1520 tests.

2 For an explanation of why we use an MTU of 4200, please see the Linux EtherDrive® FAQ entry 5.29, "Why does the Linux AoE driver only use 4132 byte Jumbo Frames?"



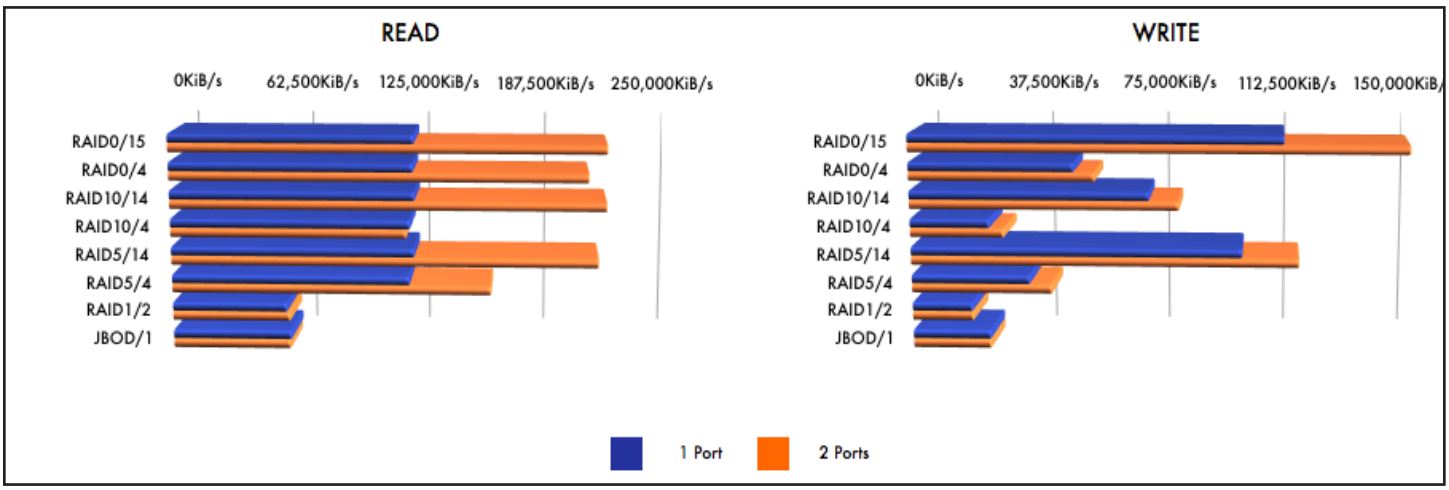
NUMBER OF PORTS USED	KiB / s	RAID0 15 DISK	RAID0 4 DISK	RAID10 14 DISK	RAID10 4 DISK	RAID5 14 DISK	RAID5 4 DISK	RAID1 2 DISK	JBOD 1 DISK
1	WRITE:	89,040.67	52,578.67	45,740.33	26,977.33	77,099.67	39,839.33	16,011.00	22,325.33
1	READ:	109,062.33	115,582.00	113,621.67	111,786.67	111,989.33	112,440.33	53,302.67	53,674.00
2	WRITE:	78,364.33	42,171.00	40,031.00	22,364.67	67,223.33	32,358.67	17,145.33	22,103.67
2	READ:	112,164.00	103,083.67	111,257.67	95,964.67	106,371.67	91,247.33	50,130.33	51,373.67

Figure A. Performance Data: SR1520, MTU of 4200 (Jumbo)



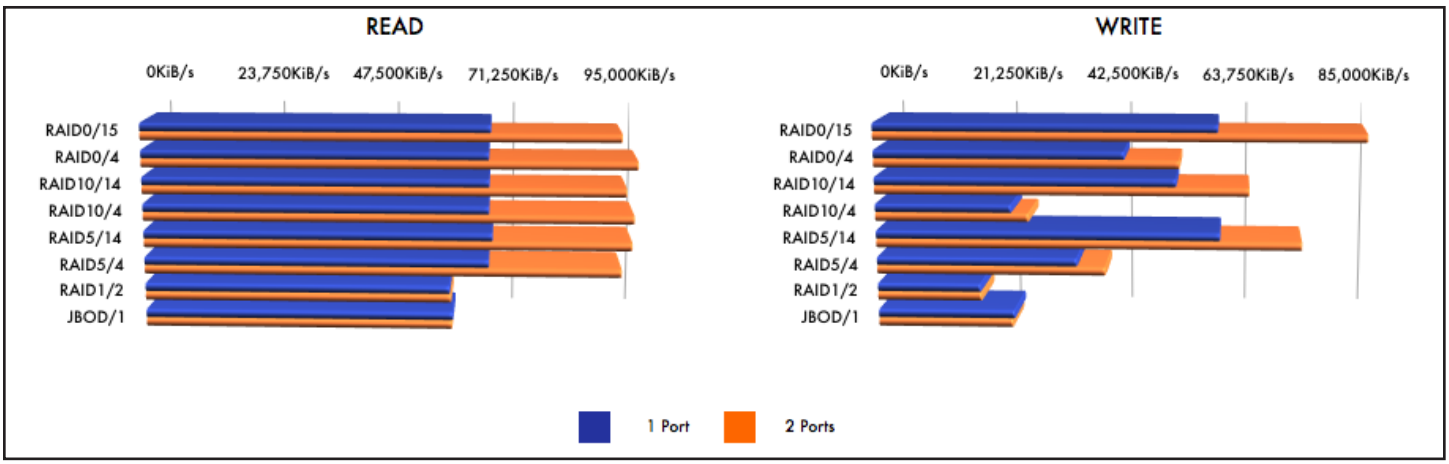
NUMBER OF PORTS USED	KiB / s	RAID0 15 DISK	RAID0 4 DISK	RAID10 14 DISK	RAID10 4 DISK	RAID5 14 DISK	RAID5 4 DISK	RAID1 2 DISK	JBOD 1 DISK
1	WRITE:	47,299.67	43,103.00	39,672.00	24,935.67	44,337.33	34,794.00	15,300.00	21,240.00
1	READ:	56,090.00	56,611.33	55,537.67	55,703.67	56,033.00	54,914.33	50,745.67	52,603.33
2	WRITE:	40,367.00	35,873.00	31,716.00	19,290.33	36,090.00	26,425.33	15,809.67	20,707.00
2	READ:	45,884.67	48,296.67	48,033.67	49,102.67	46,675.00	46,657.33	41,534.67	48,405.00

Figure B. Performance Data: SR1520, MTU 1500 (Standard)



NUMBER OF PORTS USED	KiB / s	RAID0 15 DISK	RAID0 4 DISK	RAID10 14 DISK	RAID10 4 DISK	RAID5 14 DISK	RAID5 4 DISK	RAID1 2 DISK	JBOD 1 DISK
1	WRITE:	109,228.67	48,151.00	70,009.67	23,058.00	97,212.00	34,741.67	17,472.67	22,854.00
1	READ:	118,961.00	118,417.00	119,198.00	116,934.33	118,969.33	116,551.33	56,034.33	58,133.00
2	WRITE:	145,482.67	54,254.00	78,704.33	27,320.67	113,814.00	41,361.00	17,772.00	22,914.33
2	READ:	212,323.00	203,633.67	212,401.33	115,040.33	208,849.67	156,546.00	57,612.00	57,706.33

Figure C. Performance Data: SR1521, MTU of 4200 (Jumbo)



NUMBER OF PORTS USED	KiB / s	RAID0 15 DISK	RAID0 4 DISK	RAID10 14 DISK	RAID10 4 DISK	RAID5 14 DISK	RAID5 4 DISK	RAID1 2 DISK	JBOD 1 DISK
1	WRITE:	57,324.67	41,826.00	50,428.67	22,796.33	57,760.00	33,752.67	17,362.67	23,125.33
1	READ:	65,042.00	64,727.00	64,842.00	64,799.67	65,484.33	64,807.67	57,472.00	58,127.33
2	WRITE:	81,922.00	50,971.67	62,564.00	25,701.33	71,408.67	38,548.67	17,551.67	22,823.67
2	READ:	89,180.00	92,215.33	90,227.00	91,739.00	91,441.67	89,580.33	57,750.33	57,944.67

Figure D. Performance Data: SR1521, MTU 1500 (Standard)

Appendix B - References

The `.config` file for stuart is available at:

`http://www.coraid.com/support/sr/stuart.config`

The SR support page includes the SR firmware, user manual, and related docs:

`http://www.coraid.com/support/sr/`

Please e-mail `support@coraid.com` with any questions or comments.