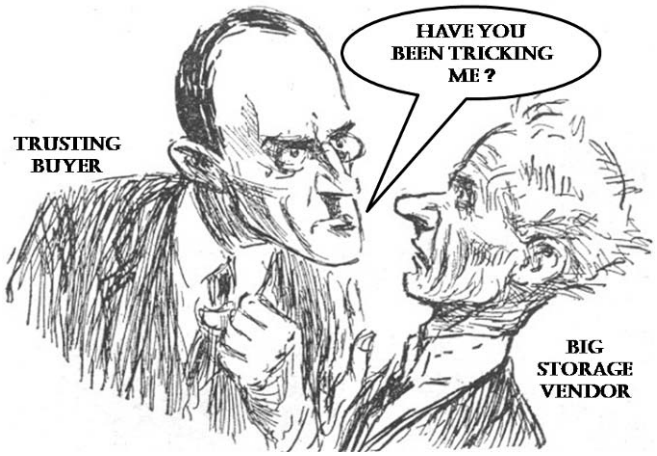
Always Useful News & Information from the Storage People at Coraid

Coraid News

Volume 1, Issue 1 Spring 2007

Important Disk Reliability Studies Published

Field Data Indicates Vendor MTBF Claims Overstated



Our Opinion
Coraid Inc.

Recently published studies about disk failure statistics indicate that the storage industry may be in for some big changes. Take a look at these field measured results from very large numbers of disks used in network storage systems.

5th USENIX Conference, Google Inc. – **"Failure Trends in a Large Disk Drive Population"**

Computer Science Department, Carnegie Mellon University - **"Disk failures in the real world: What does an MTTF of 1,000,000 hours mean to you?"**

continued on page 2

Who's Been Using AoE ?

The list of recognized users of EtherDrive AoE storage is growing by Petabytes each month. **ISPs, Hosting Companies, Universities, and Government Agencies** top the list, with smart **System Integrators** coming on strong. EtherDrive® storage is now deployed and in service in more than 40 countries around the world. That's not counting the ones that the **CIA, FBI, Army, Navy, and Air Force** have deployed somewhere out there.

AoE usage is growing exponentially with more than half of our EtherDrive customers each month coming back to buy more:-)

Popular EtherDrive Storage applications include:

- With NAS Gateway File Sharing Servers
- Disk-to-Disk Backup (now cheaper than tape)
- Email, Web, FTP Servers
- Video Surveillance Evidence Archives
- IPTV Stream Media Servers
- Shared Cluster Servers
- GRID Systems
- Database Servers
- XEN Virtual Servers

Bottom line? If an application works with a standard disk drive it should work great with EtherDrive storage.

AoE is Spreading

New AoE products for consumers are coming soon.

New AoE Product Introduced

We have it from reliable sources that a new AoE storage product for the consumer markets has been introduced at this springs CeBit show in Hanover Germany. The new consumer product is said to be a single disk toaster size unit with a PATA disk and a GigE connection.

continued on page 4

Inside This Issue

continued from page 1

http://216.239.37.132/papers/disk_failures.pdf

http://www.usenix.org/events/fast07/tech/schroeder/schroeder_html/index.html

Very interesting questions can be raised.

- Field data from a huge sample of hard disk drives operating under a variety of application loads and environments, shows that SATA, SCSI and Fibre Channel (FC) disks all have about the same failure rates. So it seems fair to ask, “Are SCSI and FC disks worth the 4x price premium over SATA disks?” Network storage users are discovering that most applications work just fine with SATA disks.
- The field data also shows that disk workload (duty cycle) doesn’t have much to do with failure rate. So running SATA disks at 100% duty cycle isn’t any less reliable than SCSI and FC disks? This also seems to imply that MAID (powering down idle disks) won’t improve disk reliability, but it may reduces power consumption.
- Seems like the field data shows that disk failure rates continue to increase with time, and do not have significant infant mortality, followed by a period of low failure rate (the bath tub curve of failure rates doesn’t apply to disks). If this is true, then why would you want your storage vendor to “burn in” the disks and then sell them to you?

Our position has been to let our customer buy their own SATA disks. This lets them save money and select just what they need for their specific application.

There are thousands of EtherDrive storage appliances in operation using just about every kind of SATA disk drive you can find. And with Coraid’s RAIDShield™ soft fail repair running on the SR1520 and SR1521 we don’t hear about excessive numbers of SATA disks being declared failed.

We welcome your opinions, feedback and field experience on these issues. We will share our findings with everyone in our next issue of Coraid News. Send us your comments to feedback@coraid.com

Thank you.

CTO’s Corner

Leading AoE Product Development

What’s Coming Next ?

By Brantley Coile

AoE and Coraid's SR products are a success. That doesn't mean that we at Coraid are resting on our laurels. Far from it - we're working as hard as ever on new products and improvements to our existing products.

The nice thing about having a product that is selling well is that we get to hear from a lot of people. We get to hear about their needs and wants. And we're working hard to provide for them. While I can't go into a lot of details until our products are officially released, I can tell you generally where we're heading.

We have faster versions in the lab of our existing products. We have new products that help people use those faster and existing products. A major effort is underway to help our customers manage their growing SR disk farms. Two things will be coming out that will help with this.

One new product will be an easy to configure appliance that will virtualize the storage into logical volumes made up of space on multiple SR units.

Also the little GUI elves are busy at work on producing an easy to use but still powerful graphical interface to help run an entire complex of SRs and virtualizers. In addition, there are products under development that will allow for offsite mirroring of ATA over Ethernet across internet links creating a virtual storage area network, complete with configurable encryption and authentication.

So, keep an eye on our web site for these and other exciting products from Coraid.

The EtherDrive Mechanic

Handy tips & tricks from a master EtherDrive Mechanic

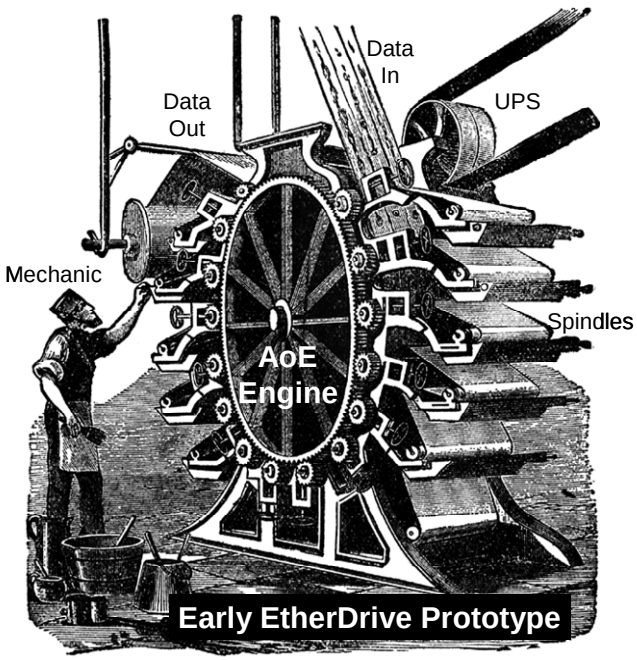
The mystery behind sharing data nicely

By Sam Hopkins

We are often asked, “Can one EtherDrive SR1521 provide storage for more than one host?”

Short answer – Yes, of course.

After studying the picture below read on for more details



See it still working in the Coraid Athens museum.

Sharing SR storage is a subject of much discussion for first time EtherDrive users. EtherDrive AoE storage is block access storage, it shares all the block sharing constraints of more traditional Storage Area Networks (SANs). Many users logically want to configure one large RAID and have all their servers just mount it. This is not possible using traditional filesystems.

Traditional local-access filesystems (ext3, XFS, JFS, etc) are designed with the assumption that the block device underlying the filesystem is only accessible from one server mounting the storage device. The filesystem implementation uses this assumption to simplify many tasks of interacting with the storage.

E.g., when a block is read from the storage device, it is stored in the server’s block cache. Traditional filesystems will use these cached blocks as long as necessary, assuming that (as the sole user of the storage device) the block contents must match the cached contents. With shared block storage this assumption is no longer valid. If a shared block device were to be mounted from two servers simultaneously, both servers would quickly fall victim to corruption as they each individually operated on potentially stale data.

It is often possible to mount a shared block device from multiple locations provided all users are mounting the device in readonly mode. In this case, no one can update the underlying device, and so no one should fall victim to the aforementioned block cache problem. This is entirely filesystem dependent; and as of this writing XFS will perform a few writes even when mounted as readonly.

To accomplish the shared use of a single block device, so called "cluster" filesystems have been invented. GFS, OCFS2, and Lustre are a few of the more popular choices. Because cluster filesystems are relatively new technology, they are often difficult to configure and maintain, requiring a thorough understanding of implementation specifics to be used effectively. It's a big investment, but for many users it's also a big win. For more information on using GFS, please see the Linux EtherDrive HOWTO section 5.10 "Q: Can you give me an overview of GFS and related software?"

Many users wanting to share the storage on the SR among several machines will partition the lblade and allocate one partition to each machine. As each machine will be responsible for a portion of the disk that does not overlap with any other machine, traditional filesystems can be used. Watch out for dos-style "fdisk" partition tables, though. The on disk partition format was not designed with today's large storage disks in mind. Each partition in a DOS-style partition can only be as large as 2TB. GUID Partition Table (GPT) is a good replacement that eliminates the limitations inherent in dos-style partition tables, permitting up to 128 partitions of (theoretically) 9.2 ZetaBytes (2^64 512B sectors) each.

Yet another way to share the SR storage is to front it by a Network Attached Storage (NAS) device that can perform AoE on its backside network (and be the sole user of the SR), exporting a mountable share on its frontside network. Clients can then share the storage using a file storage protocol like NFS or SMB. As these protocols work at a file level, they do not have the previously discussed shared storage constraints. The Coraid Linux NAS (CLN) Gateway Server provides this functionality in an easy to setup and install 64-bit Debian Linux system preconfigured for this task. For more information on the CLN Gateway Server, please see www.coraid.com.

- Cheers -

Dear Ed

Coraid's own Linux Guru Answers Your Most ~~Personal~~ Technical Questions

Can I break a CLN20-FT in half ?

(I think they meant to say "divide", not break. Our warranty doesn't cover intentional breakage.)

Q - Can I put one half of the CLN20-FT in one building and the other in a building across the street for better disaster protection?

Short answer – "Yes, of course."

(FYI - All of our short answers are Yes.)

A - Yes, by using an Ethernet-based heartbeat instead of a serial connection, you can. You will need one remote power switch for each building, though. There are two major ways to achieve such an arrangement.

Method One: Using a single Ethernet network.

If you have a high speed connection between the two buildings, you can put both server rooms in a single Ethernet broadcast domain. This configuration allows AoE to be performed from an AoE initiator in one building to an AoE target in the other.

For example, you could put CLN alpha in building 1 and CLN beta in building 2. With alpha acting as the primary host in the high availability cluster, alpha could perform Linux Software RAID 1 over two SR units, one in each building.

If beta had to take over, beta would power alpha down and attempt to start the RAID 1. If the SR in alpha's building was not available, beta would just start the RAID 1 in a degraded state, using only the closest SR.

continued from page 1

Tiny AoE Device

Why did they use AoE? Because AoE goes about 5 times faster than similarly sized NAS boxes. Check it out. Nice job Welland. We should see more boxes like this coming from other vendors soon.
http://www.welland.com.tw/html/network/network.html

Method Two: Using IP to connect multiple Ethernet networks.

If there is a routed IP network between the two buildings, AoE cannot be performed across buildings without tunneling it through a higher level network protocol. A more natural solution might be to use drbd, the distributed block device software.

http://www.drbd.org/

For example, you could put CLN alpha in building 1 and CLN beta in building 2. With alpha acting as the primary host in the high availability cluster, alpha would have a filesystem on a drbd. Whenever alpha writes to the drbd device, the data will be written to the local SR using AoE, and a note will be made in drbd's bitmap. This note tells drbd that it must sync the data over to the other building.

When alpha's drbd gets an opportunity, it sends the data via IP to the other CLN, beta. Then beta writes the data to its local SR.

If beta had to take over, beta would power alpha down and start its own drbd. Any written data would be committed first to the local SR and then synced to alpha via IP. If alpha wasn't available, the bitmap would serve as a record of all the data that still needed to be sent to alpha.

Dear Ed (cont'd)

I'm sure Ed's really happy to get another question.

Is intending to write worth writing about ?

(If you intend to write us with your question, please do.)

Q - Who needs the new Write Intent Bit Maps feature? And why?

"Yes, especially for RAID1 users"

(My boss thinks this is cool.)

A - The write-intent bitmap feature in the Linux Software RAID driver, md, is most interesting to anyone who has a RAID 1 over very large components. The bitmap will dramatically cut down on the time that is required for bringing the array from a "degraded" to a "normal" state.

For instance, with an md mirror over lblades* from two SR shelves, if one shelf suddenly powers down, then the md driver will mark that SR's AoE target as a "failed" or faulty Software RAID component.

The system administrator will then bring the SR back online, remove the failed component from the Software RAID 1, and add it back into the RAID 1. At this point, there is an up-to-date RAID 1 component, and an out-of-date component (the one that failed).

Now the md driver will do one of two things, depending on whether a write-intent bitmap is in use:

1) If a bitmap is in use, the blocks that have changed on the up-to-date component since they other went down are copied over to the restored (out-of-date) lblade.
2) If a bitmap is not in use, *all* of the blocks must be copied to the out-of-date lblade. For multi-terabyte lblades, that is a lot of copying.

While the restored lblade is being brought up to date, the array is in a degraded state, because its redundancy is not fully present. In a degraded state, the failure of a single RAID component will make the RAID itself fail. Minimizing the period of time in which the array is degraded increases the likelihood that the storage will remain available.

Coraid is currently testing the write-intent bitmap feature in conjunction with its Coraid Linux NAS (CLN) product, both in single NAS and High Availability (HA) NAS configurations. If you would like to test this beta feature, make sure you are using the latest version of the packages below. (The coraid-mdadm-pre package is not needed if you are not testing write-intent bitmaps.)

apt-get update # repeat if needed

apt-get install coraid-mdadm-pre coraid-kernel coraid-aoe

** lblade is Coraid's name for a disk on the network, accessed via the AoE protocol. An lblade can be a single disk or a RAID group of disks that looks like one large disk volume.*

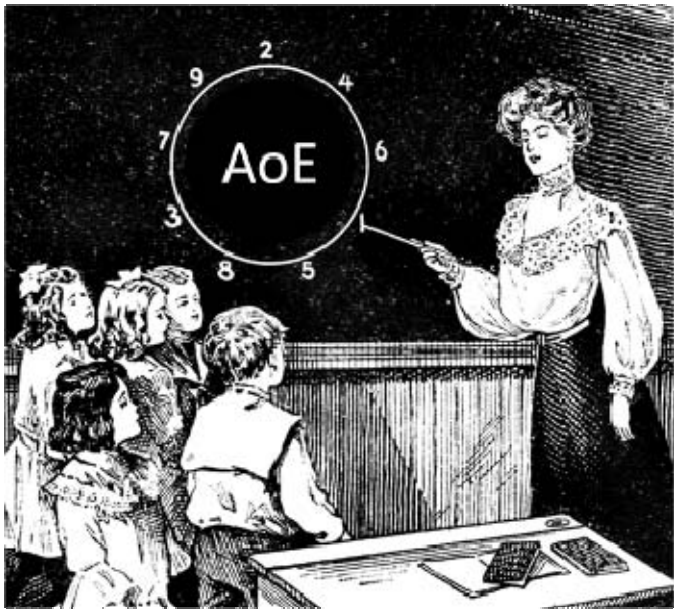
AoE and ZFS

(Solaris users are going to be happy.)

A free Solaris AoE driver (beta version) is available from Coraid. It is being tested with Sun's ZFS filesystem. Contact support@coraid.com for details.

ZFS is ideally suited for EtherDrive storage and offers an easily scalable filesystem and volume management features integrated into one open source package. With ZFS the storage pools support copy-on-write replication, unlimited snapshots, and dynamic striping for high performance. Just about everything your data might want. And since it's an open source project, more features and functions are sure to follow.

AoE IQ Test



How do you spell AoE ?.

1. **What is AoE?**
   a. A simple storage protocol.
   b. A fast storage protocol
   c. Block level disk access over Ethernet
   d. All of the above
2. **Does my host machine need a driver for AoE?**
   a. Yes, of course
   b. It's built into the Linux kernel
   c. Both of the above
3. **How does the AoE driver find a network disk?**
   a. Intuition
   b. Via AoE discovery packets
   c. Manual operator settings
4. **What is an Iblade?**
   a. Coraid's name for an AoE device.
   b. A single disk on the network
   c. A set of disks combined into a RAID
   d. Something to confuse EtherDrive users
   e. Almost all of the above
5. **How big can an Iblade be?**
   a. 750GB.
   b. Any Size
   c. 11.25TB
   d. It depends on how big the disks are
   e. Any size you can afford to buy

Recess

6. **Can an Iblade be shared?**
   a. No.
   b. Yes, see The EtherDrive Mechanic above
   c. Sometimes
   d. b and c
7. **Why doesn't AoE use TCP/IP?**
   a. Coraid doesn't understand TCP/IP
   b. TCP/IP slows things down
   c. AoE doesn't need it
   d. Coraid doesn't like TCP or IP
8. **Where can I use AoE storage?**
   a. Only in the kitchen
   b. Anywhere outside the kitchen
   c. Anywhere your mama says you can
   d. Anywhere a direct attached disk will work
9. **Is AoE a SAN or NAS protocol?**
   a. It's an Ethernet SAN protocol
   b. Can be used with a NAS server for sharing
   c. a and b
10. **Can AoE be used with a cluster filesystem?**
    a. No
    b. Yes
    c. Phone a friend

For more interesting information about Coraid and EtherDrive Storage Appliances, please see our web site www.coraid.com or give us a call at 706-548-7200.

Hold this up to a mirror to decode the answers to the AoE IQ test. Keep your score; it might come in handy some day.

1. d, 2. c, 3. b, 4. e, 5. d, 6. d, 7. c, 8. d, 9. c, 10. a