

An Overview of Coraid Technology and ATA-over-Ethernet (AoE)

Michael A. Covington

for
Coraid, Inc.

2008

1. Introduction

All Coraid products revolve around one goal: making disk storage available through networks. This white paper briefly summarizes the key concepts and underlying technology.

2. Sharing files or sharing disks?

There are two ways to store your data remotely on a network. You can deal with files or disks.

If your main goal is to share files among multiple computers, you probably want to do the networking at the file level. That is, put the files on a server, and use a protocol such as NFS or CIFS (the basis of Windows File Sharing) to access the files from more than one computer. This is usually called *network attached storage* (NAS), and its server may be called a *NAS gateway*, but, for clarity, we will call it **file sharing** and we will call the server a **file server** or **filer**. The file server "owns" the file — maintains its file name, keeps track of its creation and modification time, administers permissions, etc. — and shares the file with other computers.

If what you want to do is access disk drives — not files — over the network, so that your computers can use the drives any way they want, you'll want what others call *storage area networking* (SAN); we call it **disk sharing** (even though you normally only "share" a disk with one CPU), and we call the server a **disk server**. More strictly speaking, a storage area network (SAN) is a separate network set up for disk access, to avoid competing with traffic on the main network.

Disk sharing is simpler than file sharing because operations aren't duplicated. With file sharing, when a client computer writes a file, it has to assemble the data into a file and write it onto the server, which (again) constructs it into a file and writes it onto the disk. With disk sharing, the client computer does only what it would have had to do anyhow, with a locally attached disk drive, except that the disk drive isn't locally attached — it's on the network.

ATA over Ethernet (AoE) is a protocol for disk sharing, i.e., for placing a disk drive outside a computer and joining the drive and the computer by Ethernet.

Coraid's product lines include both disk servers (storage appliances) and Linux-based file servers.

3. Why use disk sharing?

There are several reasons you might want to use disk sharing:

- To allow several computers to access the same disk drive (using different disk partitions, or using a cluster file system to share a disk between CPUs).
- To allow "main" and "standby" computers to use the same disk, so the standby computer can be put into action if the main computer fails. (This is not possible if the disk is inside the failed computer.)
- To locate your disk drives separately from the computer (by a distance of feet or miles, for physical security or convenience).
- To attach a high-performance disk system to a computer without opening the box (just use the Ethernet port that is already there).
- To add more disk drives to a computer than will physically fit in the box.

You can of course combine file sharing with disk sharing. File servers can contain their own disks and/or use disks located elsewhere via disk sharing. A file server need not actually have a disk drive of its own - it can get to its disks through AoE.*

4. What is ATA?

ATA is the usual way personal computers communicate with their disk drives. It originated in the late 1980s when PC manufacturers decided to move the drive controller into the drive itself, giving it a very simple and quick communication interface to the motherboard. These are known as IDE (integrated drive electronics) drives, and ATA is the protocol.

*See Erik Quanstrom, "The diskless fileserver," presented at the Second International Workshop on Plan 9, December 2007, <http://cm.bell-labs.com/iwp9/papers/23.disklessfs.pdf>.

ATA replaced an earlier system where the CPU connected to a disk controller card which connected to the disk. On personal computers, ATA has largely replaced other disk protocols such as SCSI and ESDI.

There are two physical ways to connect ATA disk drives to PC motherboards, PATA (parallel ATA) and serial ATA (SATA, newer and faster). ATA over Ethernet can be viewed as a third way, where the disk drive is actually outside the computer box, connected to it via Ethernet.

5. What is Ethernet?

Ethernet is a system that allows many computers to communicate over a single system of cables using *data packets*, brief bursts of data. It is called *Ethernet* because it borrows many concepts from radio transmission, and "ether" ("aether") is the invisible material that was once thought to carry radio waves.

Ethernet is used with several kinds of cable; twisted-pair wiring (Cat 5, Cat 5e, and Cat 6) is now the most familiar kind within offices, but coaxial cables were used originally, and fiber optics are used for links that carry heavy traffic.

On top of Ethernet are several layers of *protocols* (ways of organizing information): IP (for assigning addresses to machines and routing packets), TCP (for delivering data consisting of multiple packets), and, higher up, SMTP, FTP, HTTP, and all the familiar protocols of the Internet.

6. What is ATA over Ethernet (AoE)?

ATA over Ethernet is the simplest possible way of sharing a disk drive through a network. The communication that would take place between motherboard and IDE disk drive is arranged into data packets and sent through the Ethernet.

AoE is not built on IP, TCP, or other protocols. It is simple and direct. Packets are addressed to devices using their low-level Ethernet addresses (MAC addresses), not IP numbers.

7. Is Ethernet fast enough for disk drives?

Obviously, yes, because people are using all sorts of networked and external file systems all the time, many of them considerably slower than AoE. For comparison, here are some communication speeds:

Typical 7200-rpm PATA disk drive: 72 MB/s (sustained throughput)

Typical 7200-rpm SATA disk drive: 105 MB/s (sustained throughput)

Wireless-N: 30 MB/s

USB 2.0 "hi-speed": 60 MB/s

IEEE-1394 (FireWire): 200 MB/s

Gigabit Ethernet (1000baseT): 125 MB/s

10-Gigabit Ethernet: 1,250 MB/s

All these speeds are in megabytes, not megabits, per second; the latter would be a factor of 8 to 10 higher depending on the data encoding. Note that Gigabit Ethernet is faster than USB and comparable to FireWire and to the speed of the disk itself.

8. AoE packets are not routable — and that's a good thing

Because they do not use IP addressing, AoE packets are not routable — that is, a router cannot send them to another network. Lacking IP addresses, from a router's viewpoint they simply have nowhere to go. (They *do* make it through the switches that are used within local-area networks.)

Non-routability is a good thing because it means AoE packets are inherently secure. You don't have to worry about them escaping onto the Internet, or about hackers on the Internet getting into your disk drives.

Of course, remote access to an AoE device through the Internet can be achieved through *tunneling* — you can use software to convert local packets into routable packets at both ends of a link. Many of us are familiar with tunneling as the mechanism behind virtual private networking (VPN), which gives us secure access to our files from far away. The point is, AoE starts out impenetrable, and if you want to access it over the Internet, you have to make special arrangements. It will not be remotely accessible by accident.

9. How do you configure an AoE server through the network?

AoE servers don't have IP addresses, so you can't Telnet to them. Instead, Coraid provides CEC (Coraid Ethernet Console), which provides a terminal interface to Coraid AoE devices. You can get it from Coraid or from <http://aoetools.sourceforge.net>.

10. How AoE compares to iSCSI and Fibre Channel

Fibre Channel and iSCSI are two alternatives to AoE for communicating with a disk drive through a network. Both of them are based on SCSI rather than ATA, which means they use a more complicated protocol designed for a variety of devices (scanners, printers, etc.), not just disk drives. Accordingly, they include more overhead in each data packet and perform more processing.

As its name suggests, Fibre Channel was developed for fiber optics, but it is also usable with wired Ethernet, either by itself (Fibre Channel over Ethernet, FCoE) or with IP routing (FCIP).

Note that FCoE postdates AoE, and unlike AoE, does not pass through ordinary network switches. FCoE was adapted from a non-Ethernet network with guaranteed delivery of packets, so it requires special switches when carried over Ethernet. By contrast, AoE complies with Ethernet's normal practice of "best-effort delivery" with automatic retries when a packet is lost.

iSCSI is SCSI over TCP/IP over Ethernet. That is, the iSCSI protocol uses Ethernet to transport SCSI commands and data using the routable TCP/IP protocol. Compared to ATA, there is additional overhead and processing because of SCSI, TCP, and IP, all three of which ATA does not use.

11. What's in the data packets

To further understand the difference between AoE, Fibre Channel, and iSCSI, consider what is in the data packets. A typical AoE data packet looks like this:

```
8-byte Ethernet frame preamble  
14-byte Ethernet header (tagged 0x88A2)  
10-byte AoE header  
12-byte ATA command  
0 to 1024 bytes of data (more if your network supports jumbo frames)  
4-byte Ethernet frame CRC
```

That's all — just 48 bytes plus the data. By contrast, iSCSI and Fibre Channel wrap the disk drive in the SCSI command set, then wrap SCSI in a transport protocol. AoE simply delivers ATA communication through the network.

12. Not just one disk, but many

To be shared via AoE, a disk drive doesn't have to *be* a plain ATA disk, it just has to *act* like one. Thus, AoE is a great way to share RAID systems and present them to the computer as if they were simple disk drives, off-loading all the RAID work onto the AoE disk server.

13. What is RAID?

A RAID system (Redundant Array of Inexpensive Disks) uses multiple disk drives to emulate one disk drive with higher reliability and/or greater speed.

RAID systems improve speed by *striping*. That is, consecutive blocks of data are spread across different drives so that they can be read in immediate succession, without waiting for the read-write heads to move from each block to the next on a single drive.

The simplest way for a RAID system to improve reliability is to maintain two copies of all the data, so that if one drive fails, the other one will still work. More sophisticated systems use error-correcting codes to guarantee that any single disk in an array can be reconstructed if it fails.

Either way, the computer interacts with the RAID system as if it were one disk.

RAID is not a substitute for backups. An error-correcting RAID system only protects against single disk drive failures, not against anything that destroys all the disks at once (such as a lightning strike). Further, RAID does not preserve *past* versions of your data; if you delete something, or an intruder makes changes, the RAID system cannot reconstruct it. That's why you need backups.

14. More than one Ethernet, too

You can send your AoE disk traffic over the same network cables as your TCP/IP network traffic — but you don't have to, and if performance or security is important, you probably shouldn't.

Instead, you can set up a separate subnetwork for AoE disk traffic, or even just run another network cable from your CPU server to your disk server, using a second Ethernet port on the CPU server. Coraid disk servers can have 2, 4, 6, or more Ethernet ports, enabling very fast communication.

A wide variety of combinations is possible and useful. Bear in mind that AoE packets do not pass through routers, and TCP/IP packets do.

15. AoE is an open protocol

The ATA over Ethernet protocol is simple and fully documented. Third parties can and do make AoE drivers for various operating systems. Many of these drivers are freeware.

16. Coraid disks can be read without Coraid hardware

Because a Coraid appliance (disk server) shares the disk rather than the filesystem, the disk is formatted just the way the computer would ordinarily format it — with one small exception. The first 3 sectors (LBA 0 to 2) are taken up with

Coraid-specific data. LBA 3 is presented to the computer as LBA 0, and so on.

Those 3 sectors enable Coraid RAID systems to reconstruct themselves when you move the disks into a new Coraid chassis. You do not have to set up — or even know — the configuration.

If you move a single (non-RAID) disk drive from a Coraid disk server to an ordinary computer, all you have to do is adjust the partition table to skip over those three sectors, and you can use it without further modification. (Non-RAID setups are often called JBOD, "just a bunch of disks.")

Even Coraid's RAID algorithms are open. You can take the disks out of a Coraid RAID array and reconstruct the data without using Coraid hardware, using software from <http://aoetools.sourceforge.net>.

17. It's easy to find the disk in your server room

Unlike the SCSI protocol, which allows only 16 devices per SCSI bus, the AoE protocol provides comprehensive, simple ways to identify your disks.

In the AoE protocol itself, each disk has a unique Ethernet address (MAC address).

In Coraid storage appliances, each disk is identified by shelf number and disk number. You have to number the shelves yourself, since nobody knows where you are going to place them, but the disks within each shelf are numbered for you.

18. AoE disk servers work well with virtualized CPU servers

It is now common practice to use hypervisors such as Xen, VMWARE, Microsoft Virtual PC, etc., to *virtualize* computers that are used as servers — that is, the server runs as a simulation inside another computer. Modern computer architectures can do this efficiently. Advantages include:

- The ability to run more than one virtual CPU on the same real CPU;
- The ability to experiment with the configuration of a copy of a server without disrupting the original one;
- The ability to move a virtual server to another real server at any time, to improve performance or to correct a hardware failure;
- The ability to back up an entire server very easily, since from the outer (real) computer's point of view, it is simply one large disk file.

AoE can be used on the virtual computer or the real one (or both). The usual arrangement is to use AoE only on the real computer, so that only the real

computer needs AoE drivers; the virtual computer thinks it has a locally attached disk.

19. The disk storage can be virtualized, too

The Coraid VS disk storage virtualizer does the opposite of CPU virtualization; it virtualizes the disk. That is, the disk storage, whatever form it may actually take, can be presented to the computer(s) as something else.

The real disk (typically a Coraid RAID system) is divided into units of perhaps 4 MB, which you can arrange into virtual disks any way you like. For example, you can combine multiple 22-terabyte RAID systems into a virtual disk that is even larger. Or (more commonly) you can divide up a large disk system into smaller virtual disks any way you want, and you can change partition sizes at any time without regard to the physical arrangement of the data. You can even use two RAID systems to maintain simultaneous copies of a single virtual disk, so that complete failure of a RAID system would not bring down your server.