

ATA Over Ethernet

Making SAN Simple

Brantley Coile
CORaid, Inc



What is SAN?

- Block storage on a wire
- IBM System/360 channel/controller/disk
- HIPPI (High Performance Parallel Interface)
 - 100Mbps 50-wire twisted pair
 - 200Mbps over fiber optics
- Replaced by Fibre Channel

What is SAN?

- Fibre Channel (FC)
 - replace bulky HIPPI with fiber optics
 - Compete with IBM's SSA
 - Work started in 1985
 - First standard ratified in 1997
 - 200MBps, 1997 (full duplex)
 - 800MBps, 2005 (full duplex)
 - Mainframe requirements very complex



What is SAN?

- iSCSI
 - Access Fibre Channel like services over the Internet
 - Fibre Channel over TCP/IP
 - Enable tier one data centers use Ethernet for storage

What is SAN?

- FCoE, Fibre Channel over Ethernet
 - Fibre Channel over local Ethernet
 - Replaces iSCSI
 - (Not really Ethernet; loss-less)

Why New SAN?

- Current SAN is mainframe centric
- Current SAN has many features not needed by the Unix culture.
- eg logging into switches, passwords for disks, universal unique names for disks, etc
- Connection based limits performance
- New kind of SAN was needed



ATA-over-Ethernet

- AoE is to FC culture what Unix is to Mainframe
- AoE is easy to understand, use and deploy
- AoE has high performance
- AoE products are very affordable
- Been in kernel.org since 2005



Features of AoE

- Simple additions to Ethernet and ATA
- Carries ATA commands and responses
- Multiple outstanding requests
- Discovery
- Fencing
- Reservation



How Simple is AoE?

How Simple is AoE?

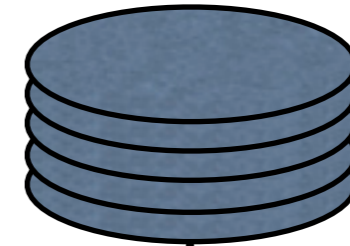
You are about to become an expert!



Linux



AoE SAN



- AoE targets appear as block devices in `/dev/etherd`
- AoE storage arrays are assigned a “shelf” number
- Individual logical units in a shelf have LUN numbers
- This 24 bit number forms a unique target address.

Setup AoE Target

```
SR shelf unset> shelf 42
SR shelf 42> make 9 raid5 42.0-11
SR shelf 42> make 10 raid5 42.12-22
SR shelf 42> spare 42.23
SR shelf 42> online 9 10
```

Complete configuration of an SR242I



AoE on Linux

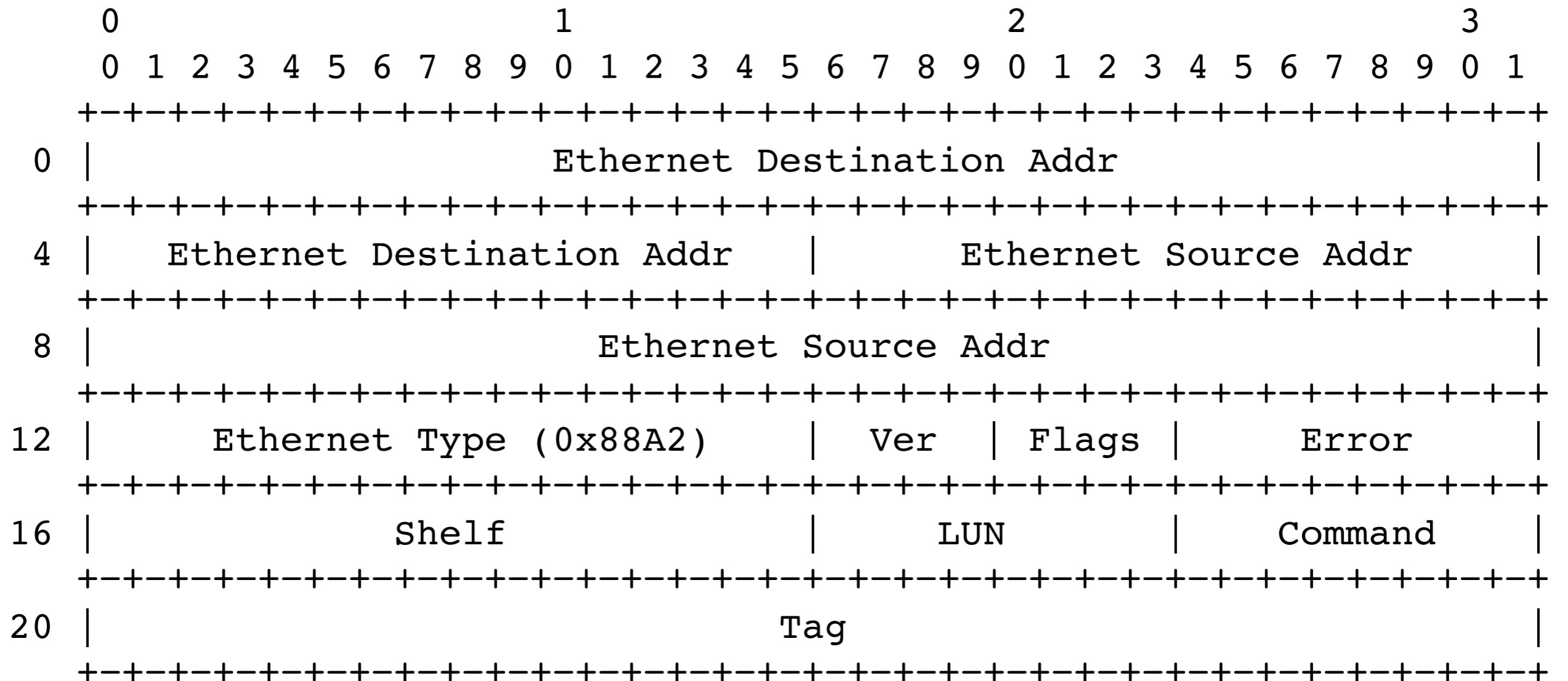
```
ellijay:/home/ecashin# ls -l /dev/etherd
total 0
brw-rw---- 1 root disk 152, 1792 Jul 20 13:15 e42.9
brw-rw---- 1 root disk 152, 1904 Jul 28 10:44 e42.10
```



AoE is RPC Based

- Remote procedure calls
- Initiator sends request
- Target responses
- Many outstanding requests
- All ports can be used
- Request and responses have same formats

Common Header



Flags Field

The Flags field contains bitwise flags defined as follows:

```
+--+--+--+--+
|R|E|0|0|
+--+--+--+--+
```

The R bit is set if the message is a response. The E bit is set in a response message if the associated command message generated an AoE protocol error. The Z bits are reserved and must be set to zero.



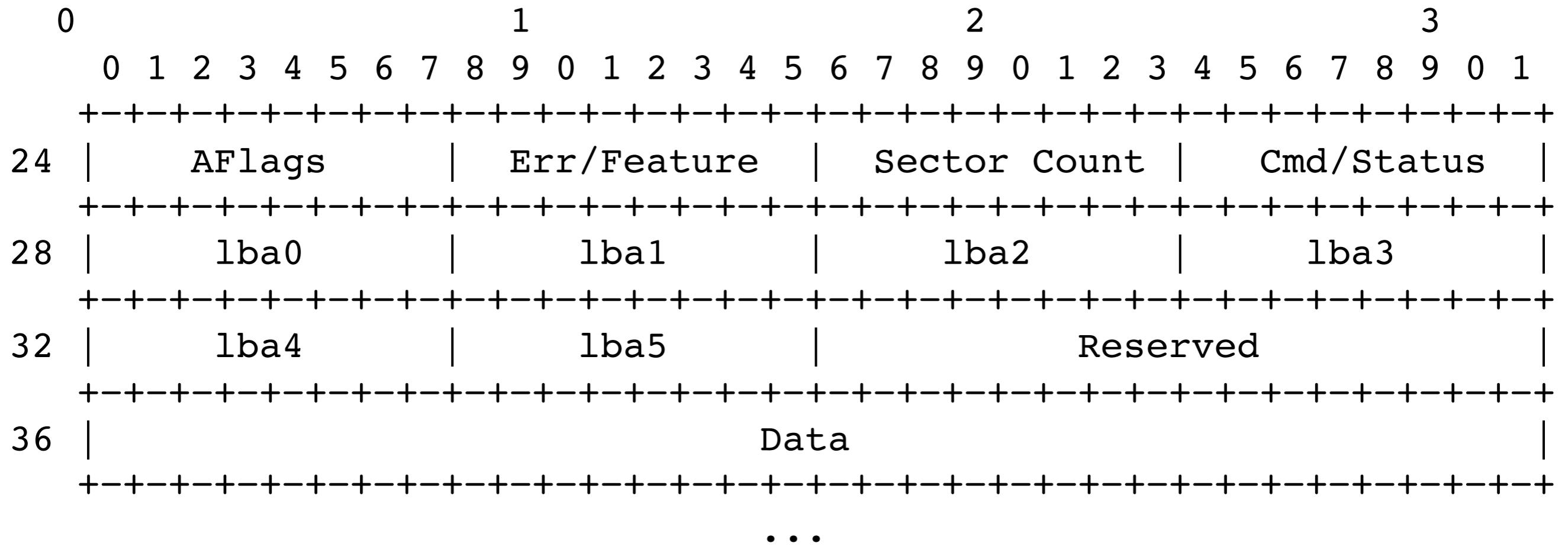
Error Values

- 1 unrecognized command
- 2 bad argument parameter
- 3 device unavailable
- 4 config string present
- 5 unsupported version
- 6 target is reserved

AoE Commands

- 0 issue ATA operation
- 1 query config information
- 2 mask list fencing
- 3 reserve target

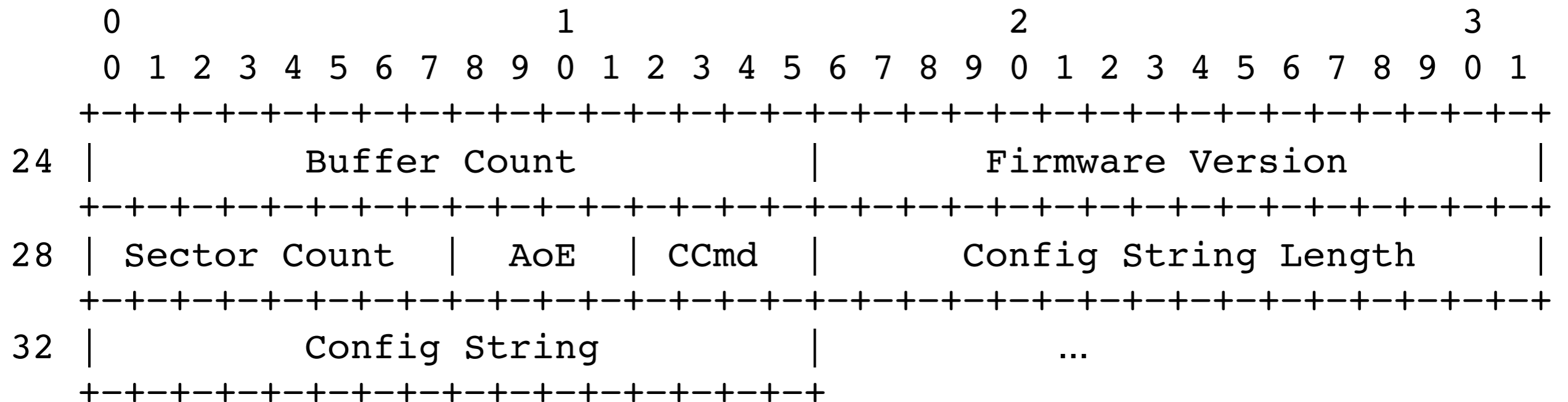
Disk Command Hdr



Query Config

- Broadcast queries to discover targets
- Claim targets
- Targets broadcast notices to network
- Includes information for driver
 - max frame size
 - hint at number of outstanding requests

Query Config Hdr



Query String

- 16 bit length
- 0-1024 bytes of binary data
- Convention is sequence of UTF-8 strings
- First word is the form
 - <domain name>.<application>
 - eg com.coraid.vs ...

Query Commands

- 0 read config string
- 1 test config string
- 2 test config string prefix
- 3 set config string
- 4 force set config string

Fencing

- By default, all MACs are allowed
- Create list of MACs that are allowed exclusively
- MAC can be added and removed
- Fencing persistent

Reserve Target

- Specify a set of MAC address that are only allowed to do IO
- Does not live beyond reboots
- Can be forced open

And That's it!

- 12 page specification
 - <http://support.coraid.com/documents/AoEr11.pdf>
- Easier to configure and deploy
- Faster than FC in ESX tests
- More affordable than iSCSI devices



ATA-over-Ethernet lacks galactic unique target names, application level CRC checks, expensive switches, and switch port logins.

-- Brantley Coile



ATA-over-Ethernet lack galactic unique target names, application level CRC checks, expensive switches, and switch port logins.

-- Brantley Coile

These features fill a much needed gap.

-- Ken Thompson



Questions?

