June, 2014

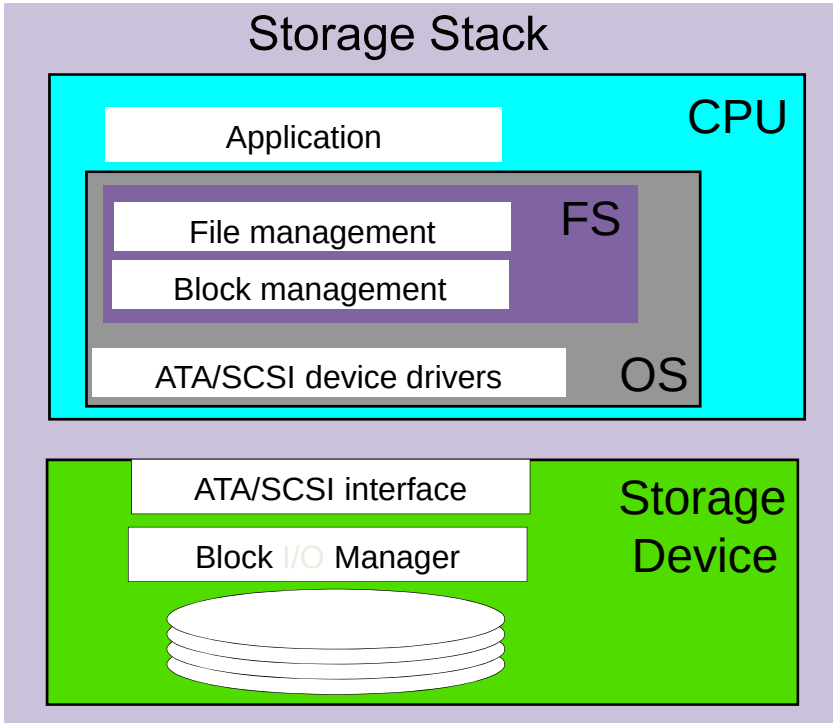
# Introduction to Kinetic Key Value Storage


*Dave Anderson*

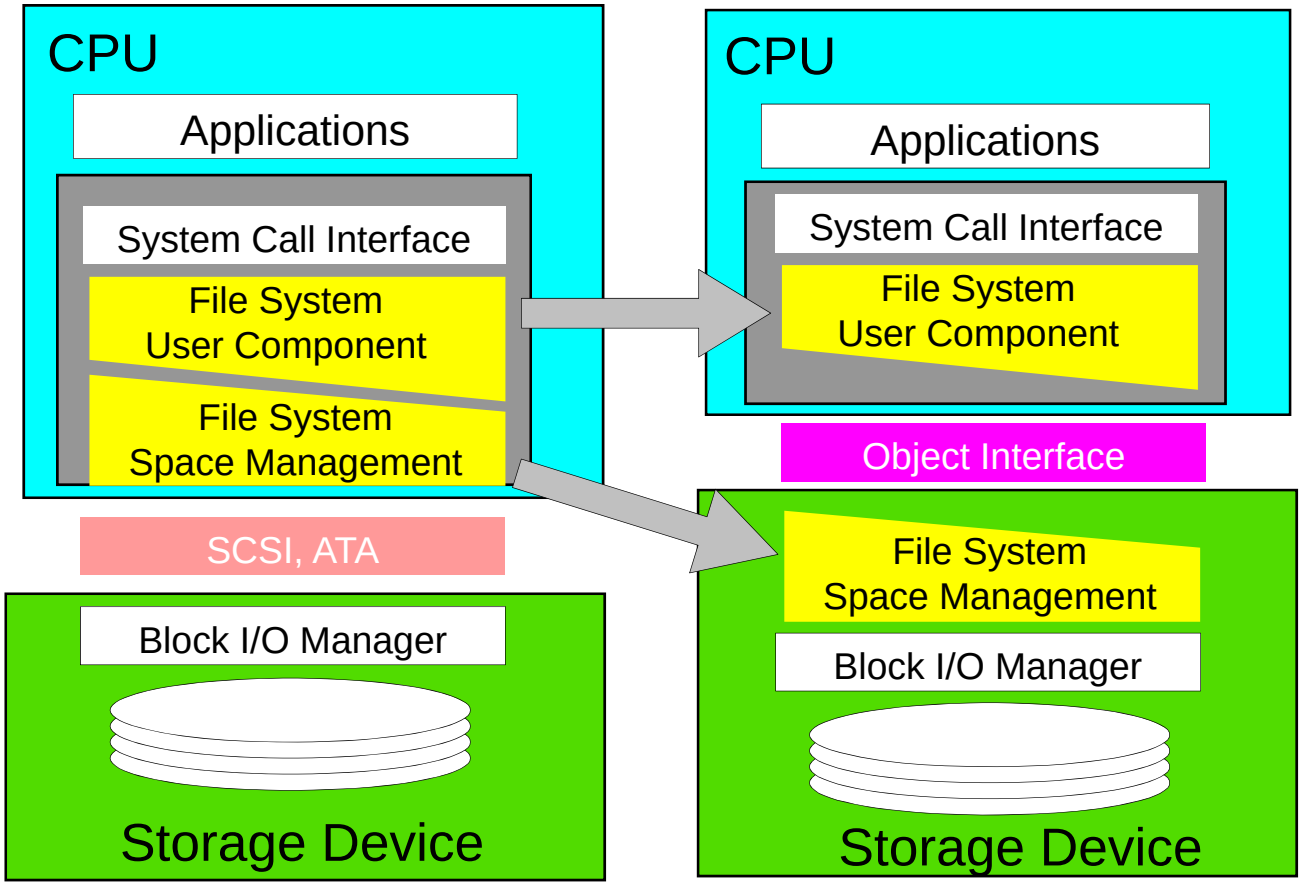
# File Systems: 2 main Components

- **Parts of a file system**
  - **Application supporting piece – the part you build and see:**
    - **Data files – naming, attributes**
    - **Directory structure – subdirectory hierarchy**
  - **Storage supporting piece – what affects the storage:**
    - **Space management structures**

- **Part of the Storage Stack**
  - **Isolation at each layer**
  - **Allow changes without affecting other layers**



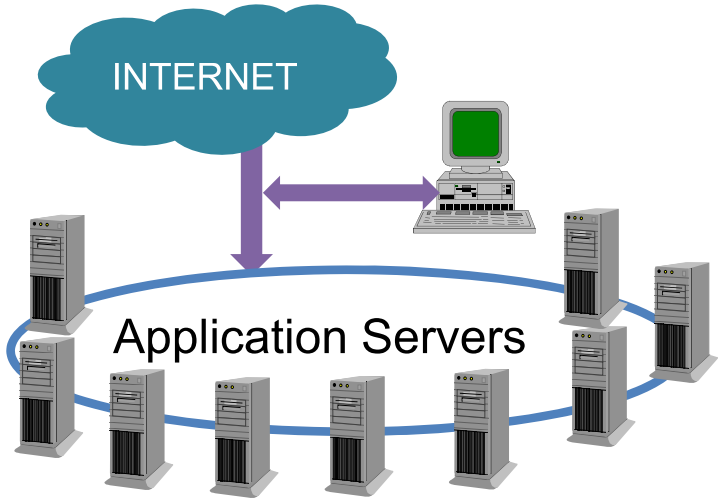
# QUIZ: A New Standard Interface



# CSPs & the File System Problem

**Internet traffic arrives & gets scheduled**

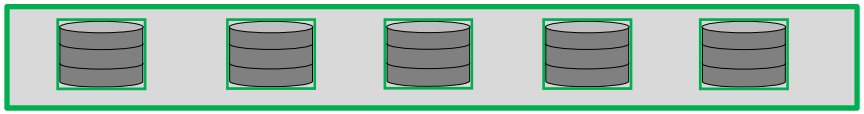
- **Goal: balance workload across all systems**
- **Requires: Any server talks to any storage**
- **Desire: global file system**




**File system gets in the way**

- **File system tied to storage device**
- **All servers update common metadata?**
  - **How to handle contention?**
  - **How to handle integrity**

??????????????????



**No array controller: creates single point of failure, expensive, doesn't scale**

- **Replication is essential to performance and reliability goals**
- **So, how to attach them?????**

# The Answer: Objects!

**App servers have global file system**

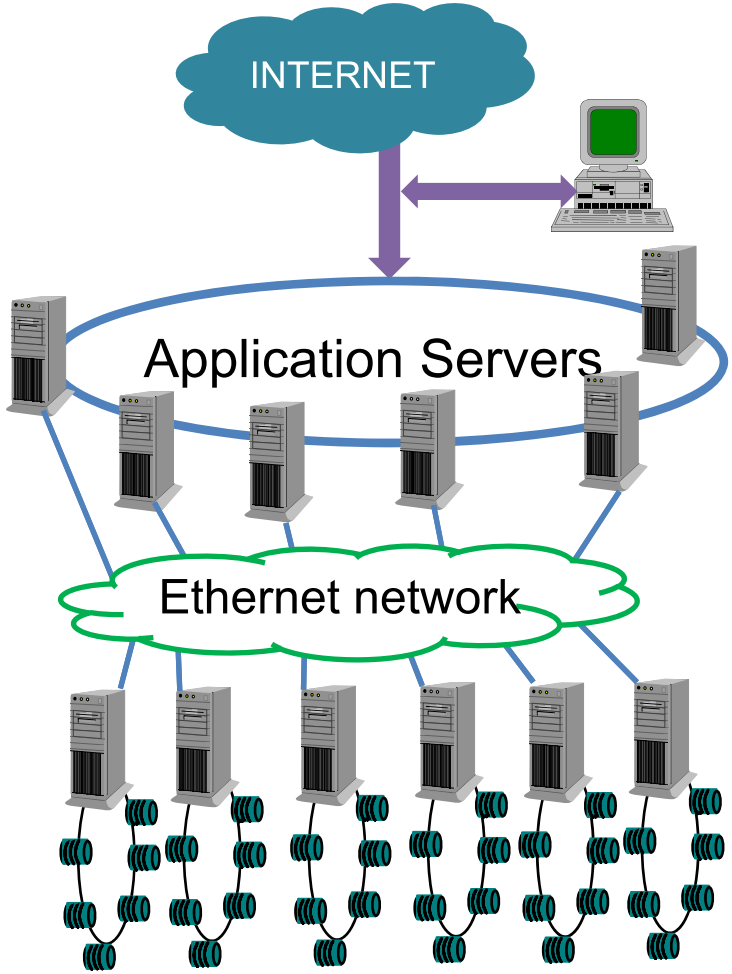
- **Talk objects (key-value) with storage**
- **metadata processing in storage servers**
  - **Single point of metadata update**
  - **At point of convergence for I/O**
  - **metadata engines = storage servers**

**Now, no object storage, soooo…**

**New tier of servers running object protocol**

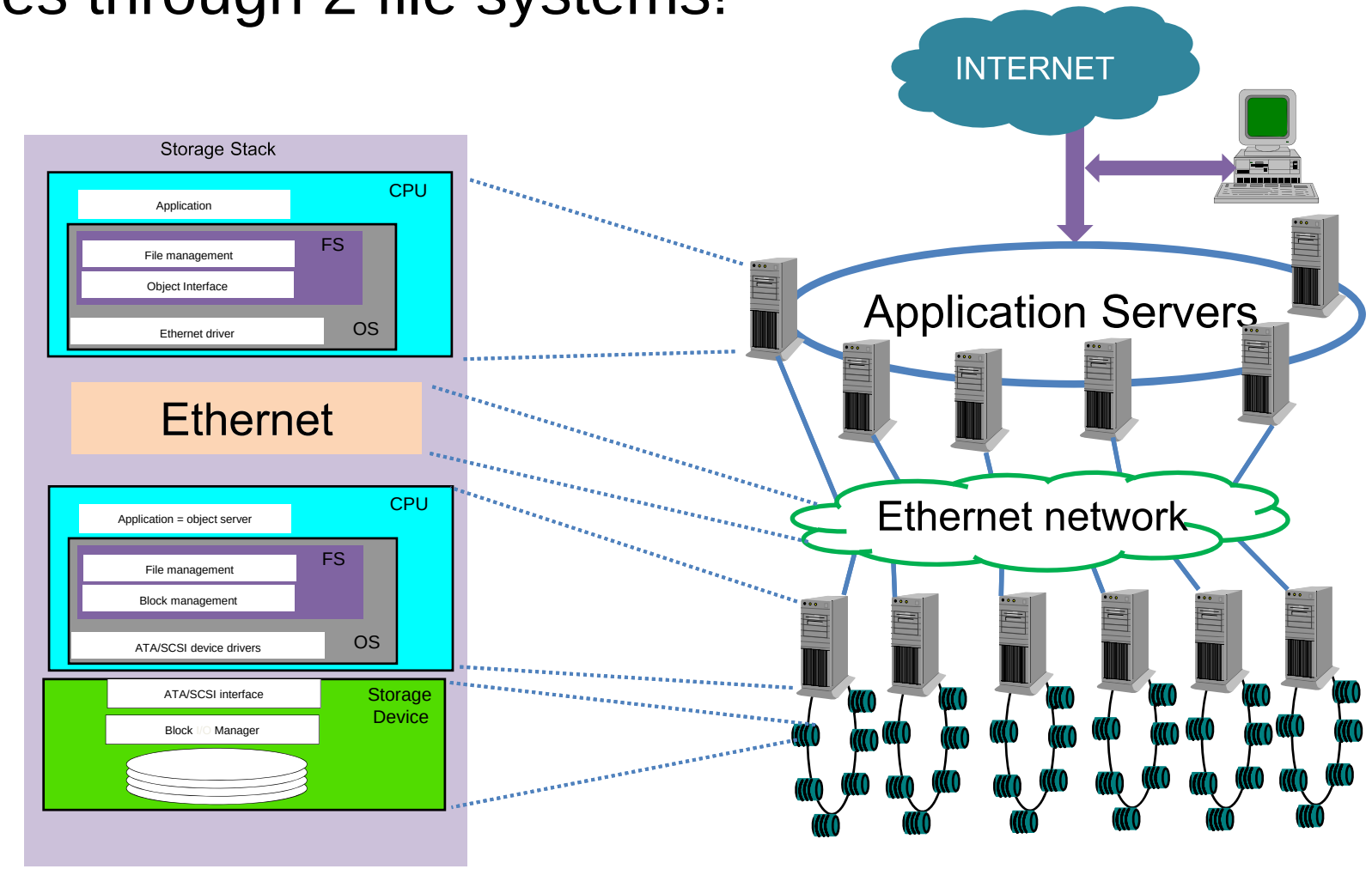
- **Storage servers run LINUX file system**
- **Storage servers have SATA HDDs**

**But, what has happened to the storage stack?**



# The Storage Stack: More Complex

I/O goes through 2 file systems!



# Kinetic: Object interface in the Drive!

**Eliminates Storage servers! (disaggregation)**

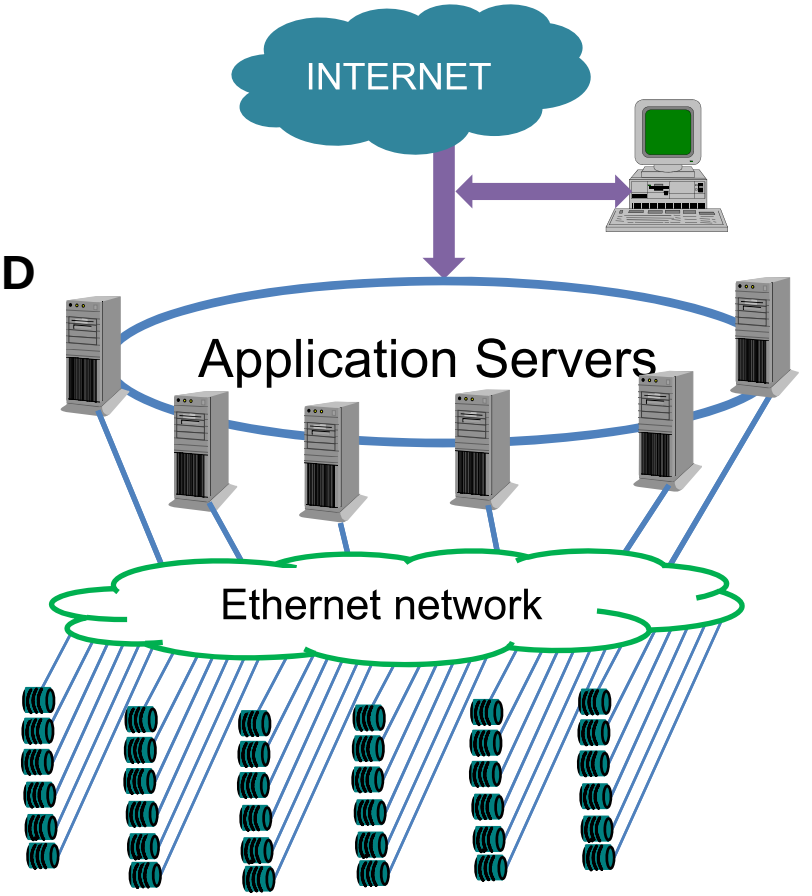
**Multiplies metadata engines! (1 per drive)**

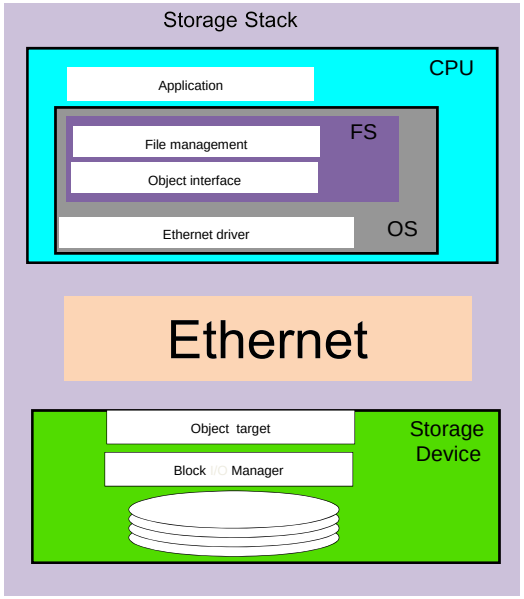
**Eliminates extra file system layer!**

**Confines many metadata operations in HDD**

**Enables 3rd party copies**

**Supports more secure computing**



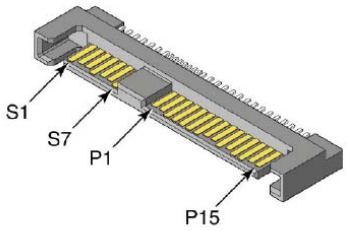



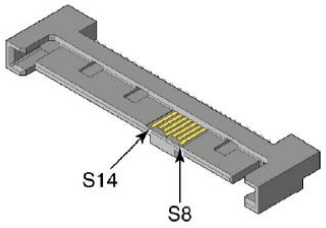
# SAS & Kinetic Open Storage

## Drive technical overview

- Standard form factor
- “SAS connector”
- 2 SAS ports

- Standard form factor
- “SAS connector”
- 2 Ethernet ports
  - Repurposed SAS piece parts
    - 1st Ethernet port (S2/S3, S5/S6)
    - 2nd Ethernet port (S9/S10, S12/S13)
  - 1-Gps Ethernet (SGMII) ports
  - Concurrent access
  - Network failure tolerance
  - Separate access types
  - TCP/IP: IPv4, IPv6



# Kinetic Drive Chassis

PRODUCT SHEET

**Newisys® EDA-4605**
4U-60 3.5" Ultra Dense
Ethernet Drive Storage Array

NEWISYS®


Ultra-Efficient Ethernet-Direct HDD Storage System delivering Object storage to Cloud and Big Data Deployments

The Newisys EDA-4605 paves the way for the latest developments in Storage technology targeting Ethernet-direct Hard Disk Drive technology. Optimized for Object storage, the EDA-4605 implements an ultra-efficient storage platform for Cloud and Big Data deployments. It is ideal for scale-out and distributed storage solutions.

Unlike traditional storage boxes, the EDA-4605 provides redundant Ethernet fabrics that connect directly to the disk drives, eliminating many layers of overhead and enabling new levels of storage scalability. Whether deployed in Cloud installations, for Big Data or the traditional data center, the EDA-4605 delivers object storage at unprecedented efficiency.

The first disk product to leverage the EDA-4605 is the new Seagate Kinetic Open Storage drive.

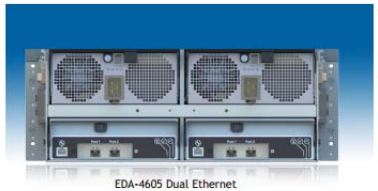
With up to 60 x 3.5" Seagate Kinetic Open Storage drives per 4U enclosure, the industry-unique Newisys EDA-4605 is an ultra-dense, space and power saving, complexity-reducing storage solution. The Newisys EDA-4605 fits nicely into a standard 19" wide, 1m deep, rack that easily creates a 15 drives/U object storage building block. With 4TB drives, this can deliver 2.4 Petabyte per standard 42U rack, and can easily scale out beyond that.

### Product Highlights

- Optimized for Ethernet-Direct HDDs, such as the novel Seagate Kinetic Open Storage Drives
- Ideal building block for Object Storage deployments
- Reduces complexity and improves efficiency for Cloud and Big Data Object Storage installations
- Full-featured, highly available, high performance Object Storage

**Product Features**

- Four 10GbE connections to the datacenter
- Redundant, 1GbE connections to each of the 60 HDDs
- Dual, redundant, hot-pluggable Ethernet Switch and Management (ESM) modules
- Dual, redundant, hot-pluggable, high efficiency power supply and fan units (PSU)
- Redundant hot-pluggable, system blowers implemented in the PSUs
- Modular design increases product configuration flexibility
- Standard chassis customization and branding available



EDA-4605 Dual Ethernet Switch and Management

### Enterprise Class Features

- Operation up to 35°C inlet temperature
- Individual HDD power control and service action LED
- Integrated Enclosure Management providing:
  - Enclosure and HDD temperature monitoring; Fan and PSU monitoring and control
  - Event reporting (IPMI) and logging Secure management access to the datacenter
  - Automatic enclosure management failover between redundant Service Processors
- Redundancy for key components:
  - Dual Ethernet paths to each HDD
  - Dual PSUs and redundant fans along with power inlets
  - Dual Ethernet Switch and
  - Management modules (redundant Ethernet fabrics, data ports, management ports, and Service Processors)
- Hot-swap for key components:
  - HDD and interposer in the carrier PSUs and fans
  - Ethernet Switch and Management modules
- Integrated cable management
- Tool-less service

### Ethernet Switch and Management modules

- Two Newisys Ethernet Switch and Management modules per system, with each module supporting 60 x 1GbE HDD ports and 2 x 10GbE datacenter network ports.
- Each HDD connects to both of ESMs in the EDA-4605 enclosure via both a data-plane (SGMII) and a low-speed management plane (I2C).
- Individual drive power control, and individual drive temperature monitoring
- ESMs configurable as both active/passive for redundancy or active/active for bandwidth
- Service Processor for management on each Ethernet Switch and Management module
- Auto failover to other Service Processor if faults occur
- 1GbE datacenter data management connection - RJ45
- Option for secure in-band or out-of-band (independent of data plane) management connections
- Secure Socket Layer (SSL) and web-based (HTTPS) host interfaces

### Capacity

- Up to 60 x Seagate Kinetic Open Storage drives or 240TB capacity per 4U enclosure (using 4TB HDDs).
- Up to 2.4 Petabyte per standard 42RU 19" rack.

### Safety Standards

- IEC/EN/UL/CSA 60950-1, 2nd Edition

### Environmental Protection

- RoHS and WEEE compliant

### 4U Rackmount Enclosure

- Compliant with EIA-310-D 19" rack specification
- System dimensions without cable management arms:
  - 6.90 in. H x 16.56 in. W x 35.1 in. D (175.3 mm H x 420.6 mm W x 891.5 mm D)
- System dimensions with cable management arms:
  - 6.90 in. H x 16.56 in. W x 39.17 in. D (175.3 mm H x 420.6 mm W x 994.9 mm D)
- Weight for Newisys EDA-4605 without HDDs Single Shipping Pack: 125 lbs (56.8 kg) max
- Weight for Newisys EDA-4605 w/ 60 HDDs Single Shipping Pack: 235 lbs (106.8 kg) typical

### Disk Drives

- 60 independent Seagate Kinetic Open Storage drives with dual Ethernet interfaces
- Failover by each Ethernet Switch and Management module to each drive
- Form factor: 3.5" Seagate Kinetic Open Storage drives
- Drive interface: 2 x 1GbE, 1 x I2C
- Individual drive power control
- Individual drive service action LED

### AC Power

- Input voltage: Auto ranging, 90-264V AC
- Input frequency: 47-63Hz
- Maximum system continuous DC output power rating: 1465W (20W HDDs)
- Typical system continuous DC output power rating: 865W (10W HDDs)

### Monitoring and Reporting

- Monitoring for temperature, advanced power and cooling modules including blower speed control, disk drives and Ethernet Switch Module(s)
- In-band reporting of all serial number, part number and revisions of each FRU and chassis

### Operating Environment

- Temperature: 5° to 35°C
- Altitude: -200 to 10,000 ft.

### About Newisys

Newisys, a product division of Sanmina Corporation, provides advanced data center products including solid-state memory and disk-based storage server appliances, JBOD storage systems and other products for a variety of data center and storage applications. Visit www.newisys.com for more information.

**Contact Newisys:**
30 East Plumeria Drive
San Jose, CA 95134
Toll-free: 855 639 7838
International: +1 408 964 3730
www.newisys.com

# Kinetic Key-Value objects detail

What makes up an object?

- Each object is a tuple of the following
  - Key
  - Value
  - Access Tag (aka version)
  - Value checksum (aka tag)
  - Value checksum algorithm specifier

K/V object = {key, value, version, EDC}

# Kinetic Open Storage API

## Quick Tour

- Key ordering
  - Keys have lexographic order

  1, 10, 11, 12, 2, 21, 22, 230, 231,

  A, Ax, Ay, Az, B, B8, B9, Ba, Bb
- Key schemas
  - Opaque to device
  - Key schema design pattern
    - » High order key space bytes first
    - » Low order key space bytes last

| Partition | Collection | Name | Chunk id |
|---|---|---|---|

# Kinetic Open Storage API

## General k/v access

k/v object = {key, value, version, EDC}

key specifiers = [this key, next key, previous key]

key range =

[start, end] or (start, end) or [start, end) or (start, end]

peer-to-peer data copying

# Kinetic Open Storage API

## Security

- At-rest encryption and Instant Secure Erase
- Message authentication
- Client authorization
  - Restricted operations
  - Restricted key space


- Transport Layer Security (TLS, SSL)
  - Advised for security management

# Kinetic: Object interface in the Drive!

**Eliminates Storage servers! (disaggregation)**

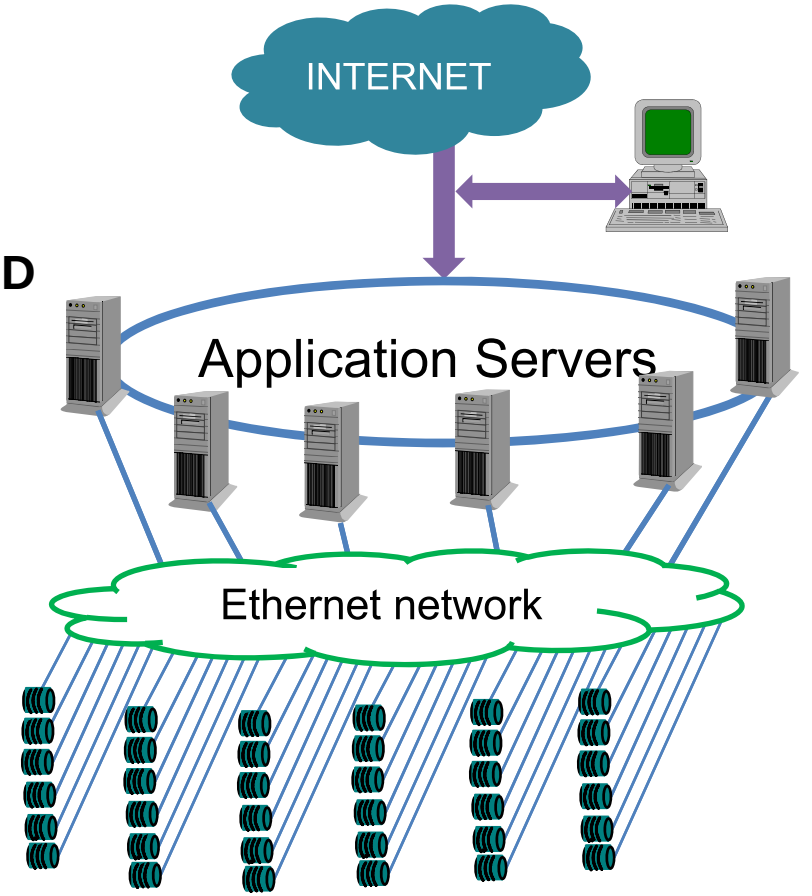
**Multiplies metadata engines! (1 per drive)**

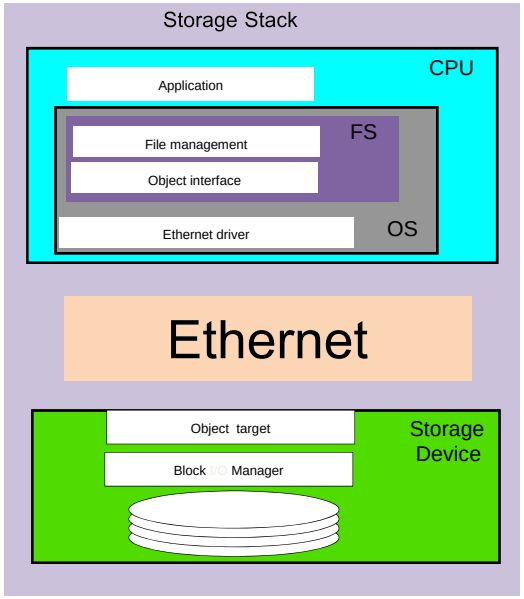
**Eliminates extra file system layer!**

**Confines many metadata operations in HDD**

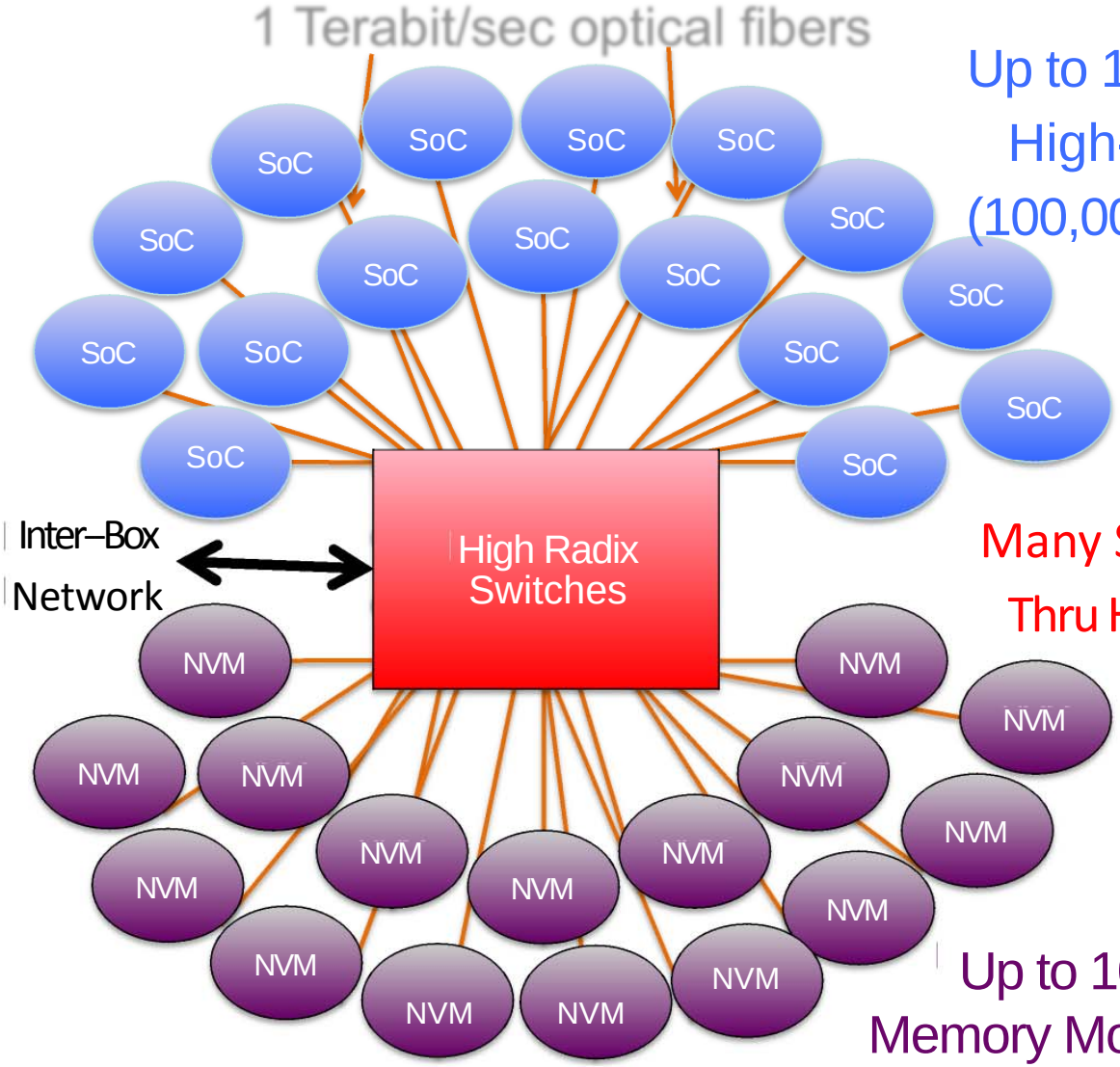
**Enables 3rd party copies**

**Supports more secure computing**



# FireBox Overview




Up to 1000 SoCs +
High-BW Mem
(100,000 core total)

Many Short Paths
Thru High--Radix Switches

Up to 1000 NonVolatile
Memory Modules (100PB total)

All slides with the “Aspire” logo at the top are from a Presentation by University of California - Berkeley at FAST 2014

# What About Future Enhancements?

Think of this Kinetic Drive as just the starting point

Ethernet & protocol buffers make enhancements really easy

Security

- Adding functionality based on SED technology opens big possibilities
- Already a lot of thinking on this

More information:

- http://seagate.com/www/kinetic
- https://developers.seagate.com
- http://github.com/Seagate