Storage Design for High Capacity and Long Term Storage

Balancing Cost, Complexity, and Fault Tolerance

DLF Spring Forum, Raleigh, NC
May 6, 2009

Lecturer:

Jacob Farmer, CTO
Cambridge Computer

My Background

CAMBRIDGE
Computer
ARTISTS IN DATA STORAGE

- 20+ years experience working with data protection, archiving, and storage technologies
- CTO of Cambridge Computer since 1991
  - Cambridge Computer provides consulting and a variety of services with a focus on data storage.
  - 25% of my time is spent circulating in the industry, meeting with vendors, and researching trends
  - 75% of my time is end-user facing
- My work is across all industries.
  - About 50% of my business is higher education
    - Often get involved in projects with the libraries
  - Plenty of other industries struggle to store and manage digital objects
- Lecturer for Usenix

My Contact Info

CAMBRIDGE Computer
ARTISTS IN DATA STORAGE

- Jacob Farmer, CTO
  - Cambridge Computer
  - 271 Waverley Oaks Road
  - Waltham, MA 02452
  - Jfarmer@CambrigdeComputer.com
  - Twitter: JacobAFarmer

A Little Plug for Usenix

CAMBRIDGE
Computer
ARTISTS IN DATA STORAGE

USENIX

- I volunteer my time to lecture for the Usenix Association
  - Usenix is a nonprofit whose mission is the advancement of information technology.
- "Usenix on the Road" – A free program that brings Usenix storage technology tutorials to college campuses.
  - Topics include:
    - Tiered Storage and Archiving: Best Practices for Life Cycle Management and Digital Preservation
    - Next Generation Storage Networking: Beyond Conventional SAN & NAS
    - Eliminating Backup System Bottlenecks

Pain Points for Long Term Storage

CAMBRIDGE
Computer
ARTISTS IN DATA STORAGE

- **Conventional file storage technology does not scale up gracefully.**
  - Disk-based file systems are limited in capacity and as they grow they become more prone to failure and corruption.
    - The bigger they are the harder they fall.
  - Corruption is hard to detect and harder to fix once you detect it.
- **Conventional backups do not scale.**
    - When your disk capacity doubles, you have to double both the capacity AND the performance of the backup system.
- **Unconventional storage technology tends to be expensive.**
  - Not just acquisition cost, but total cost of ownership

Saving Graces

CAMBRIDGE
Computer
ARTISTS IN DATA STORAGE

- **Enterprise SAN and NAS systems are not necessarily the solution.**
  - These products are designed to handle the complexity and diversity of a large data center.
  - They might have some useful properties, but you can find their essential value elsewhere.
  - Save your money!
- **Fixed content is easier to manage than dynamic content.**
  - Files that do not change are easier to back up and replicate.

Cost = Not Just Acquisition Cost

CAMBRIDGE
Computer
ARTISTS IN DATA STORAGE

- Maintenance
  - Is it affordable. What if you opt not to pay it?
- Upgrades
  - How much do they cost relative to acquisition cost?
- Opportunity costs
  - As storage gets cheaper and cheaper, does your system allow you to take advantage of the declining costs?
- Product life cycle
  - Are you forced to replace your equipment at an inopportune time?
  - What costs are associated with replacing it?
    - Data Migration?
- Environmental – Power, cooling, physical space
- Operational costs

What Do We Care About Other than Costs?

CAMBRIDGE
Computer
ARTISTS IN DATA STORAGE

- Fault tolerance
  - Uptime
- Data Protection / Redundancy
- Data Integrity Assurance
- Performance
  - But maybe we have a high performance version of the file that is separate from the preservation format
- Operational Efficiency / Ease of Use
- Environmental
  - Rack space, power, and cooling

CAMBRIDGE
Computer
ARTISTS IN DATA STORAGE

Storage 101

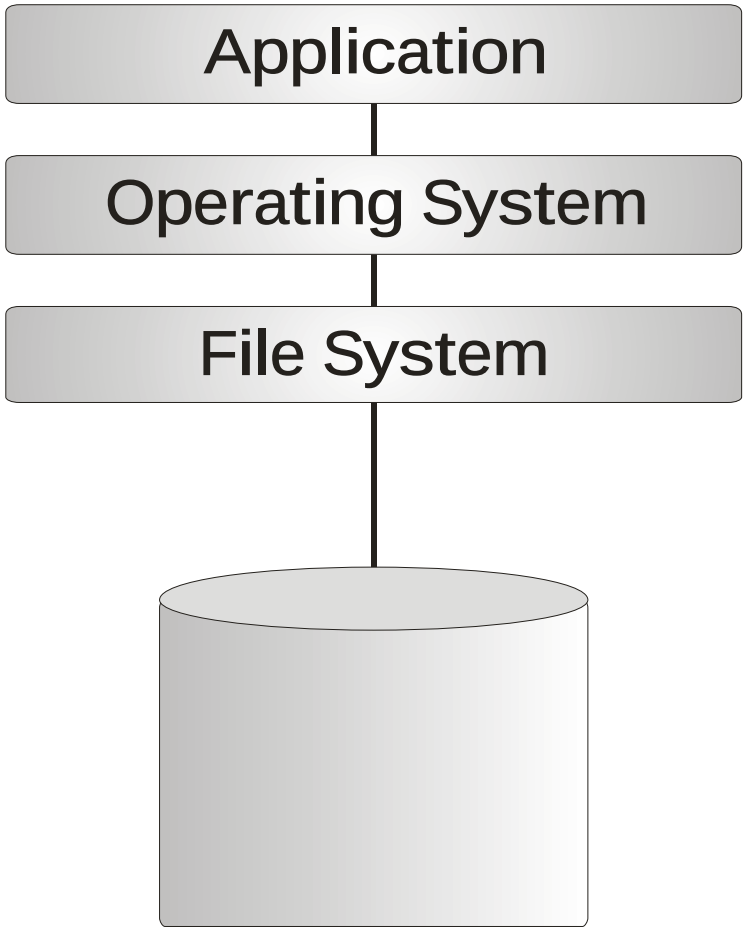
Files v. Blocks

CAMBRIDGE
Computer
ARTISTS IN DATA STORAGE

- **Blocks**
  - Least common denominator in conventional storage technologies.
  - A block is a unit of data storage.
  - Hard drives and RAID arrays serve requests for blocks.
- **Files**
  - Objects consisting of multiple blocks.
  - Blocks are organized into files by file systems, which are like databases of all of the files, their attributes and records of the blocks that make up the files.

The Storage I/O Path

CAMBRIDGE Computer
ARTISTS IN DATA STORAGE



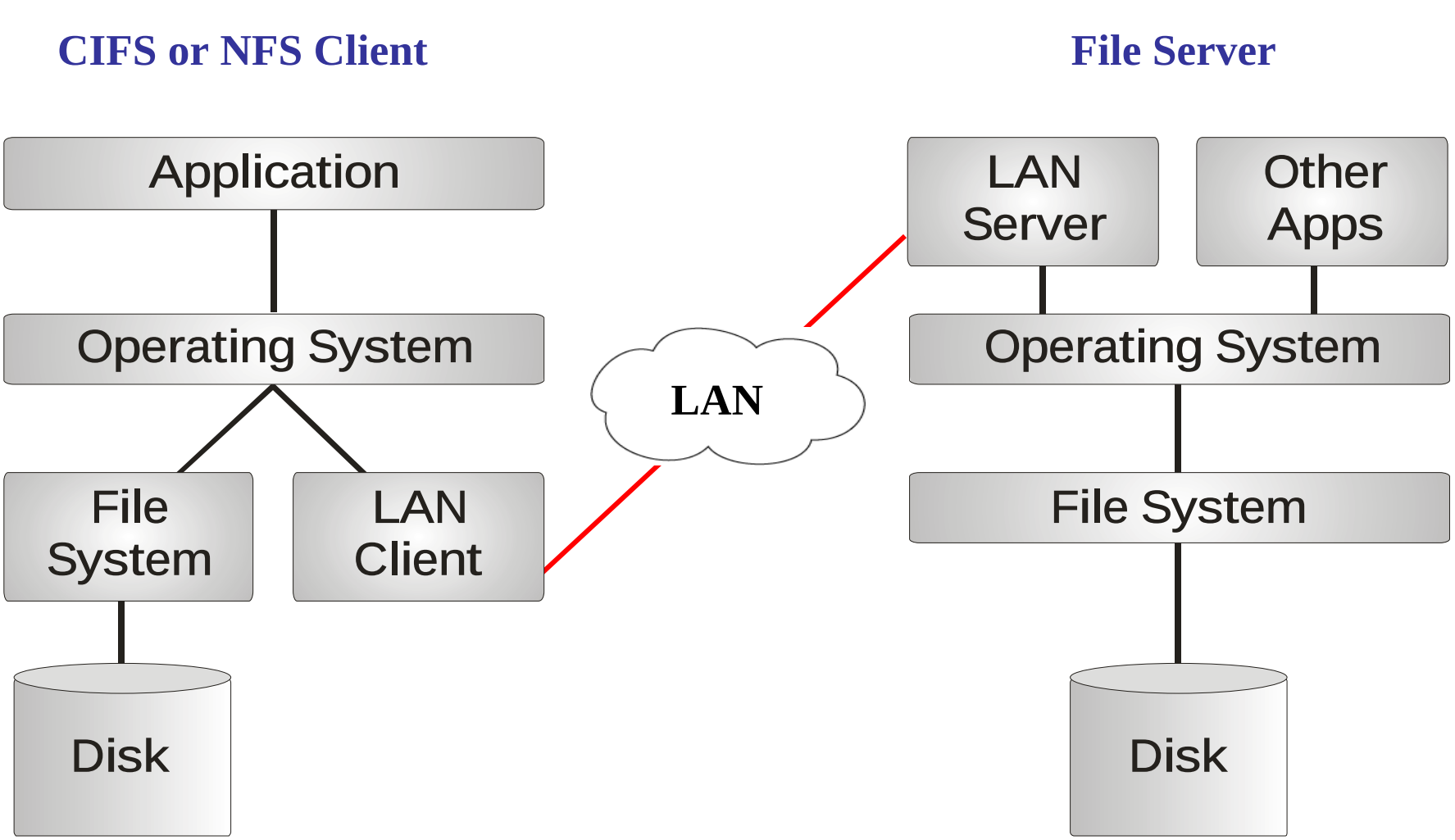
Storage has a layered architecture, very much like a network stack.

Disk drives store data in blocks. Each block has a unique numerical address.

Disk devices (hard drives, RAID systems, etc.) are like “block servers”, meaning you ask them to perform operations on specific blocks.

Network File Services

CAMBRIDGE Computer
ARTISTS IN DATA STORAGE



NAS = Fancy Word for File Server

CAMBRIDGE
Computer
ARTISTS IN DATA STORAGE

- NAS = Network Attached Storage
  - Marketing term to refer to a file server appliance.
- Appliance
  - Marketing term to describe a turnkey product based on hardware and software.
  - Typically, the software and the hardware are inextricable.
    - When you refresh the hardware, you replace the software.
    - When you want to replace the software, you dump the hardware.

File Systems and Their Limitations

CAMBRIDGE
Computer
ARTISTS IN DATA STORAGE

- Conventional file systems do not scale very well.
- Microsoft NTFS can address up to 256TB of capacity in a single file system.
  - But no one in their right mind would ever do this.
- Leading NAS products typically do not offer more than a few TB in a single volume.
- Large file systems can become corrupt and it is very difficult to identify and root out corruption.
- Backing up large file systems takes forever.
  - Restore takes even longer!

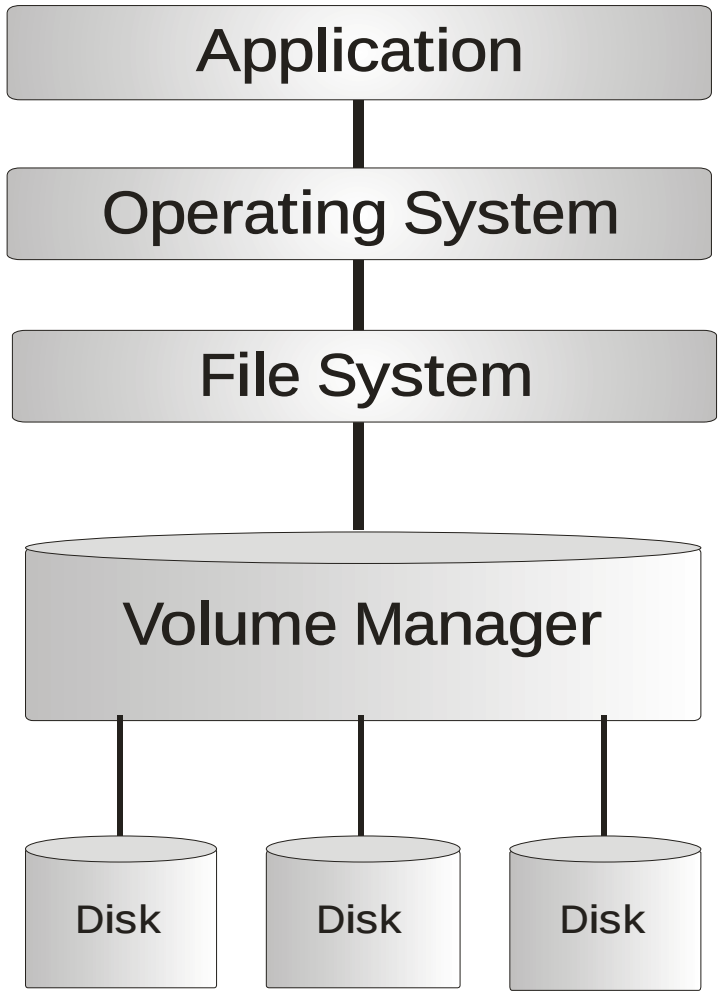
Block-Level Abstraction

CAMBRIDGE Computer
ARTISTS IN DATA STORAGE

Abstraction of the physical disk.

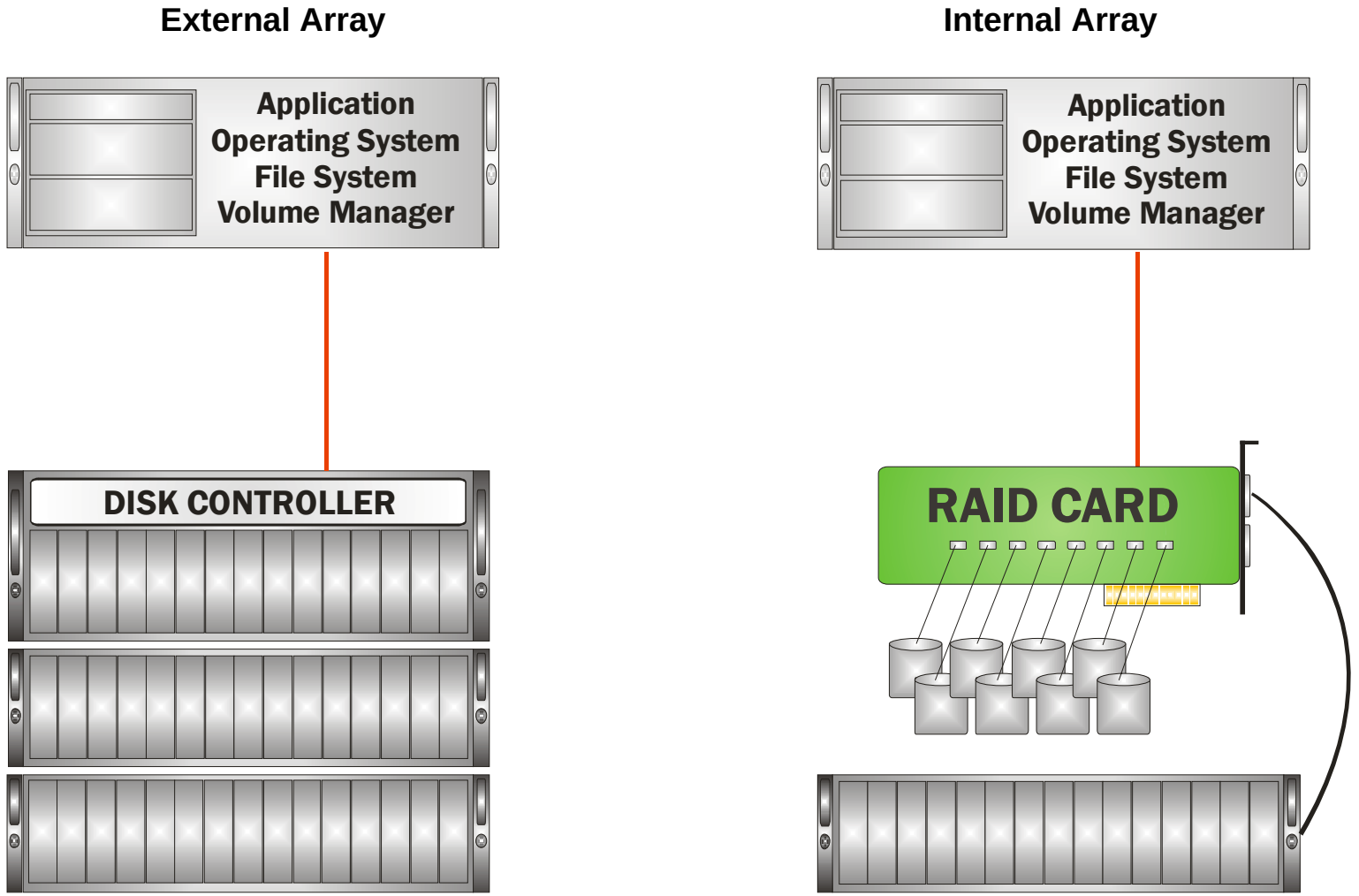
File system asks for specific block addresses and volume manager fulfills request from the disks.

- Software RAID
  - Hard Drive Fault Tolerance
  - Spindle Aggregation
- Host-based mirroring
- Volume-level snapshots
- Volume-level replication



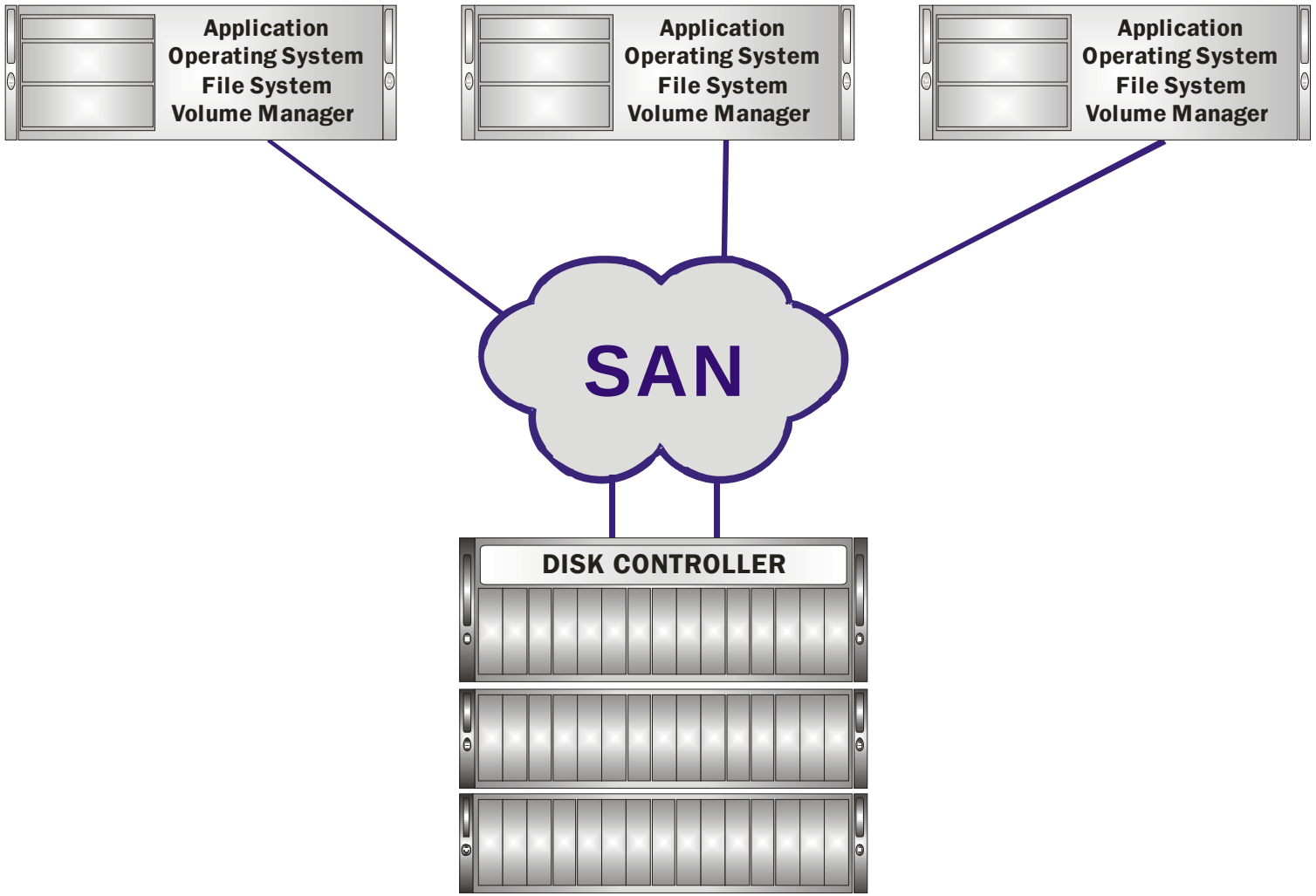
RAID Arrays = Hardware-Enabled Block Abstraction

CAMBRIDGE Computer
ARTISTS IN DATA STORAGE



SAN Array = Centralized, External Abstraction

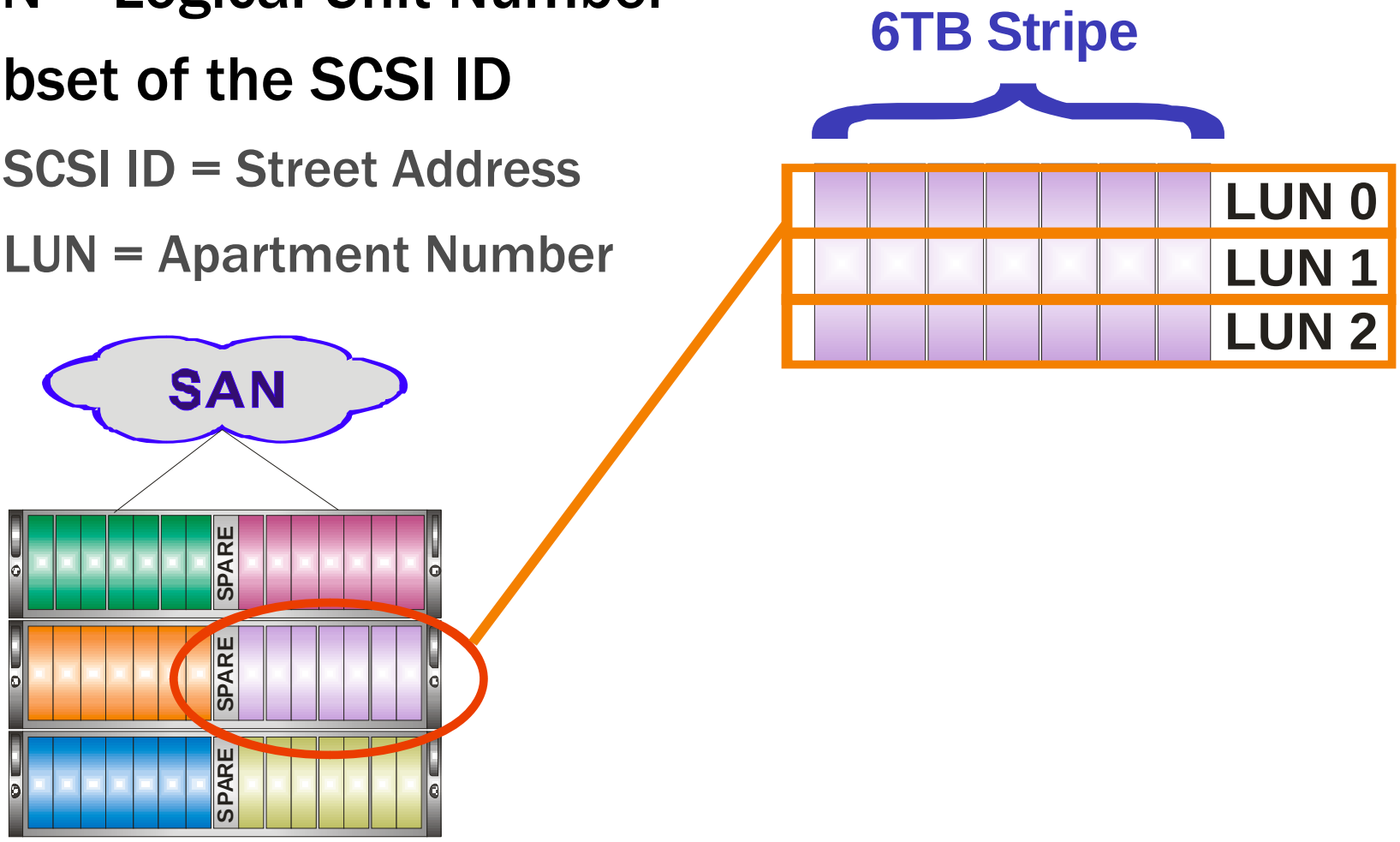
CAMBRIDGE Computer
ARTISTS IN DATA STORAGE



LUNS – Logical Hard Drives Carved Out of a Disk Array

CAMBRIDGE Computer
ARTISTS IN DATA STORAGE

- LUN = Logical Unit Number
- Subset of the SCSI ID
  - SCSI ID = Street Address
  - LUN = Apartment Number



The Layers

CAMBRIDGE Computer
ARTISTS IN DATA STORAGE

Web Server | Web Server

Asset Management

File System | File System

LUN | LUN | LUN | LUN | LUN | LUN | LUN | LUN

Another 3-Letter Acronym - CAS

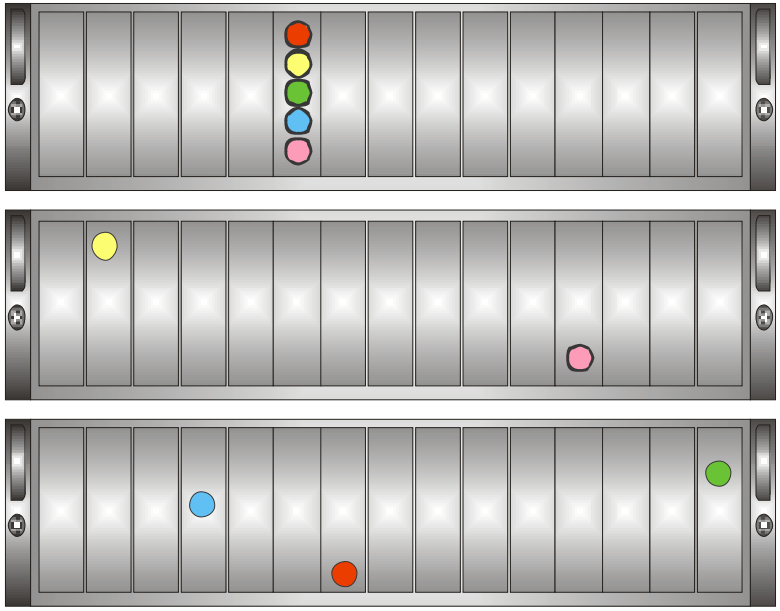
CAMBRIDGE
Computer
ARTISTS IN DATA STORAGE

- CAS – Originally coined by EMC to describe their Centera Product
  - CAS = Content Addressable Storage
  - The file is addressed by a unique number that is derived by hashing the file.
- New meaning – "Content Aware Storage"
  - All this means is that the storage sub-system knows that it is holding a bunch of objects and the management logic is tailored to object management.
    - Other than that, there is no official industry definition.

Reducing the Size of the Redundancy Unit – Clustered "Object" Storage

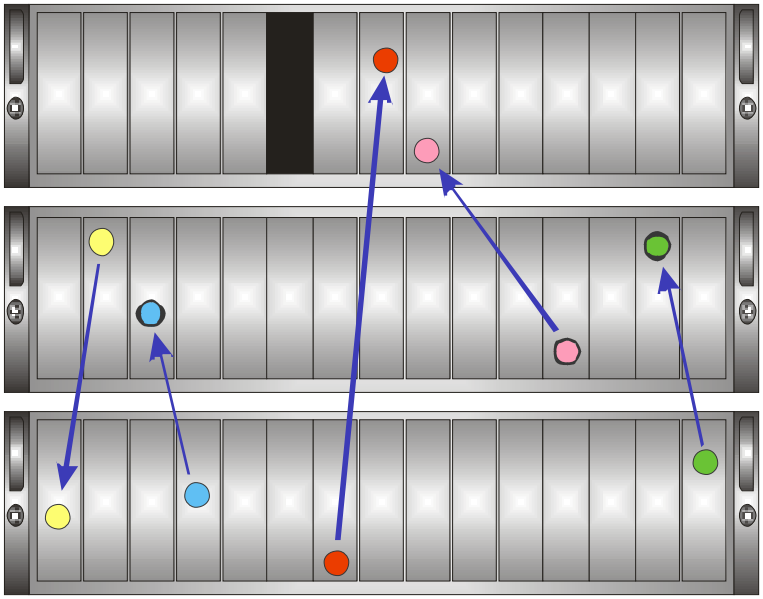
CAMBRIDGE Computer
ARTISTS IN DATA STORAGE

Normal Operation



Small units are stored redundantly across the pool of drives and cabinets.

Drive Failture



Recovery from a failed drive is highly parallel.

Properties Associated with Content-Aware Storage

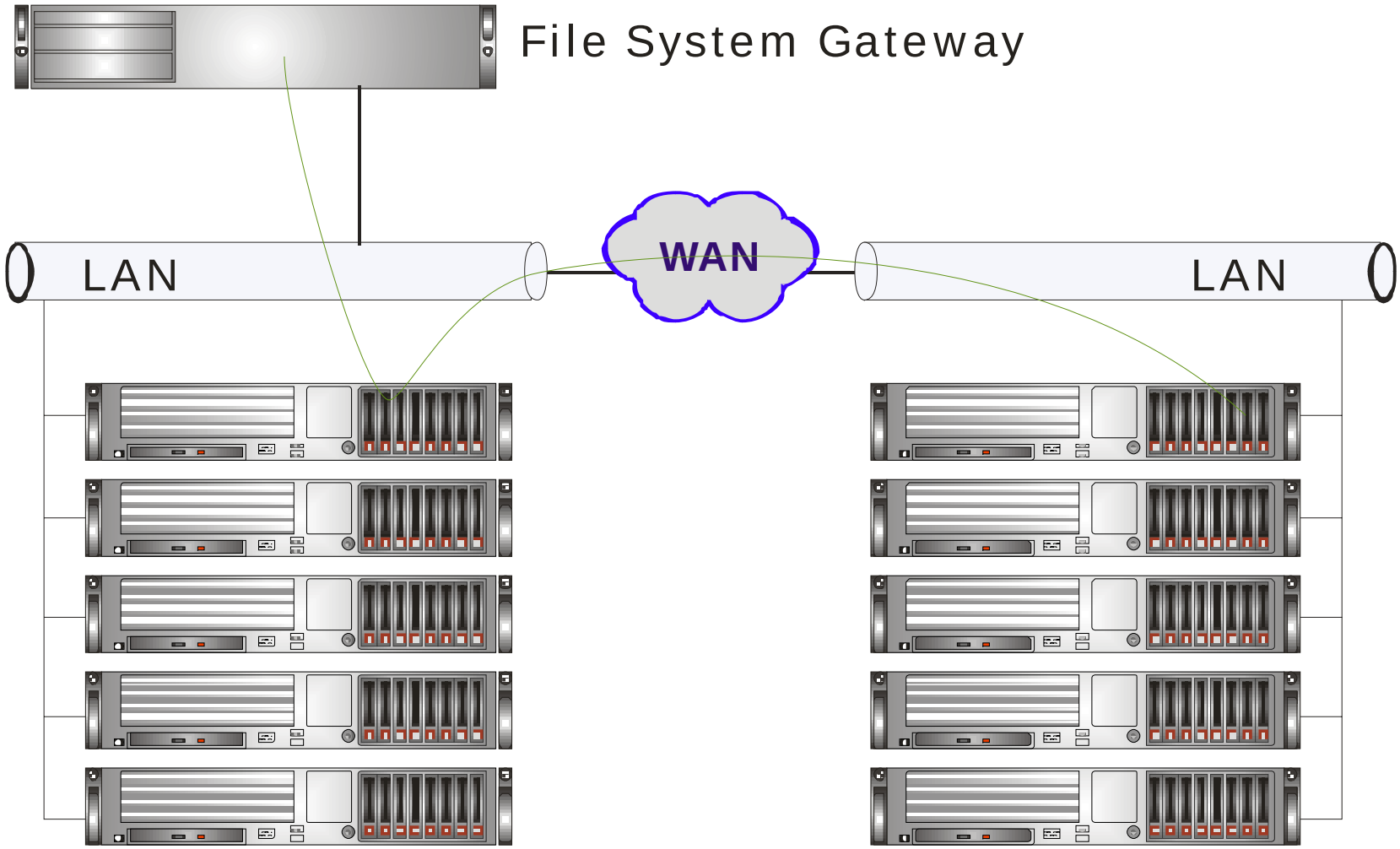
CAMBRIDGE
Computer
ARTISTS IN DATA STORAGE

- Self healing
- Immutability (write once)
- Single instance storage
  - a/k/a deduplication
- File integrity checking
  - If redundant copies are stored, corrupt files might be automatically replaced with a non-corrupt file

- Many CAS systems offer a gateway that emulates a conventional file system.

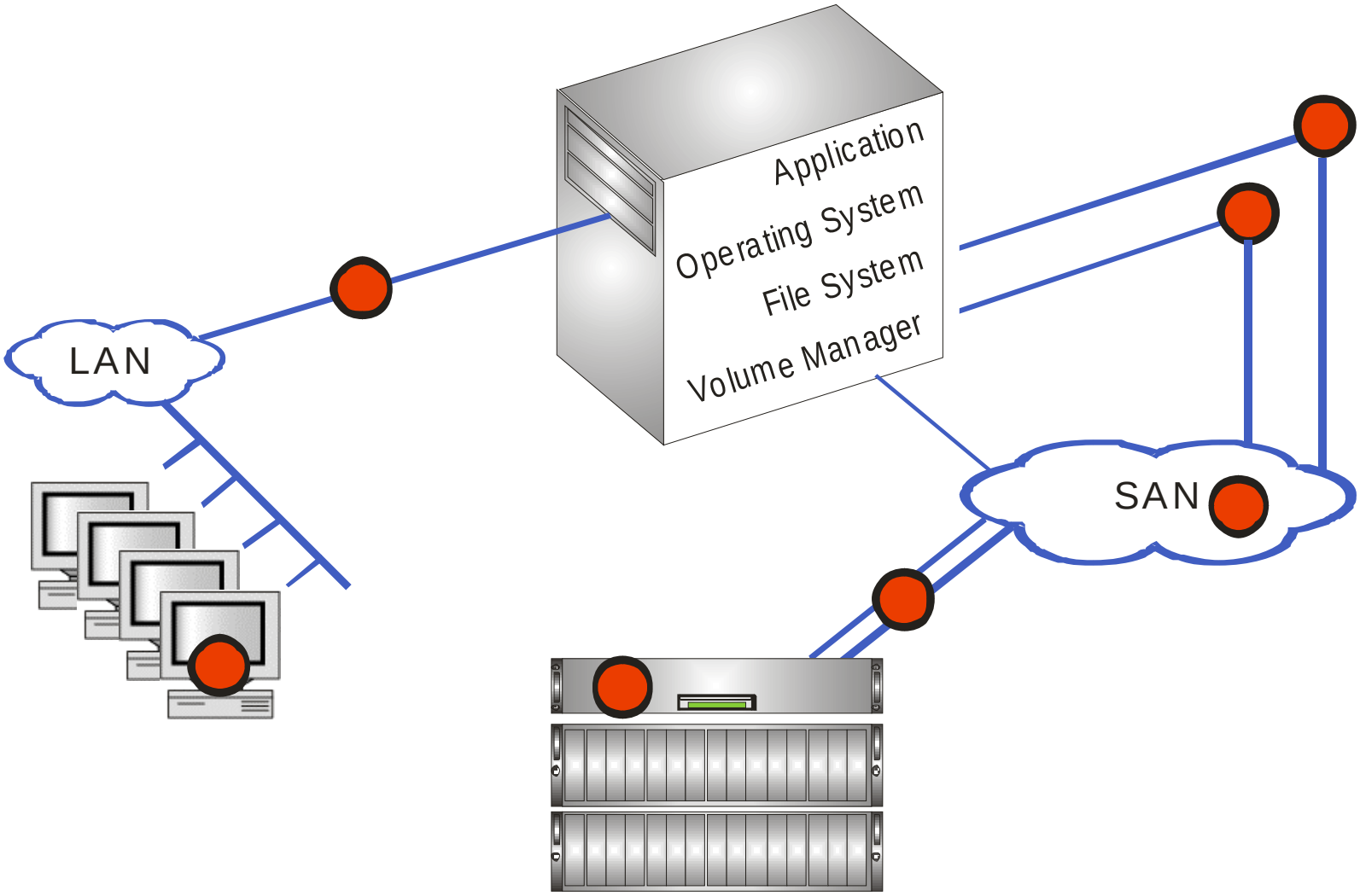
File System Gateway in Front of Replicated CAS

CAMBRIDGE Computer
ARTISTS IN DATA STORAGE



Storage Logic Can Live Pretty Much Anywhere

CAMBRIDGE Computer
ARTISTS IN DATA STORAGE



CAMBRIDGE
Computer
ARTISTS IN DATA STORAGE

Three Ways to Grow Big

Monoliths, Aggregates, and Federations

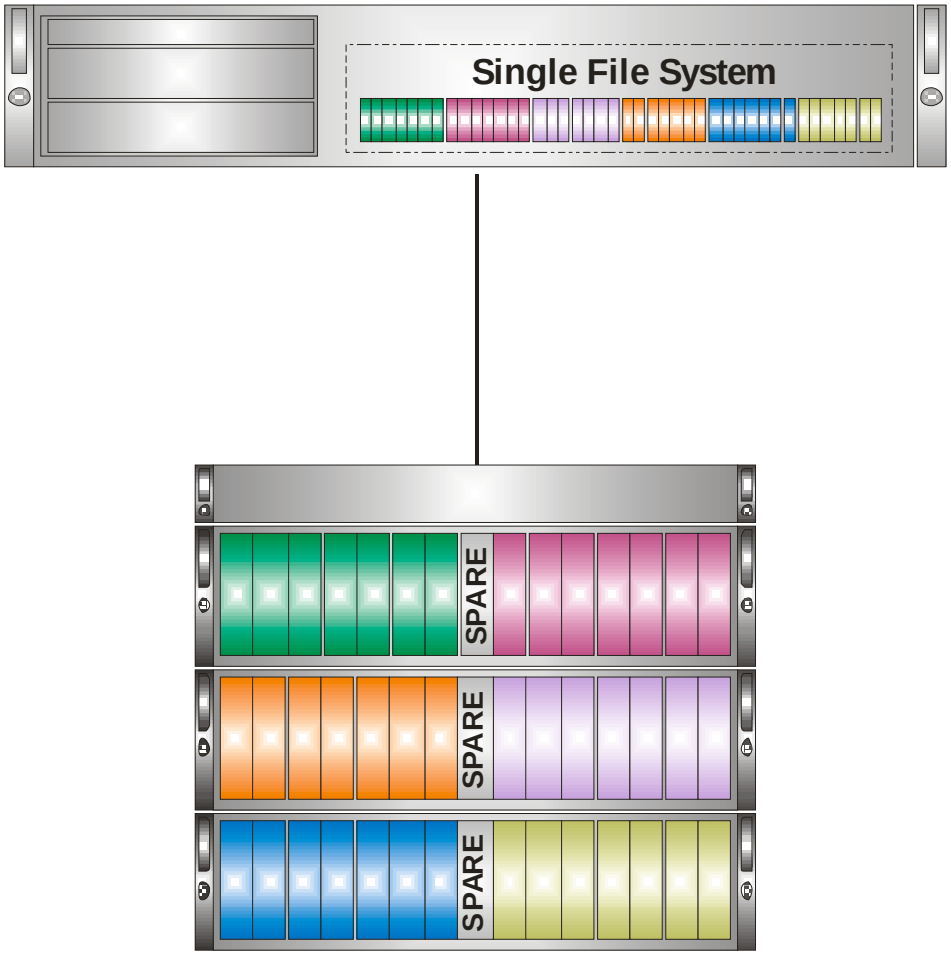
Monoliths

CAMBRIDGE
Computer
ARTISTS IN DATA STORAGE

- Monoliths are single systems that meet all of your foreseeable needs.
- Monoliths are relatively easy, very reliable, and some scale pretty big (1-2PB)
- As monoliths get really big, you encounter some problems.
  - Might be expensive
  - Replacement at end of life cycle can be painful
    - Much harder to extract a wisdom tooth than an incisor
  - If they break, it can be really messy
- The moment you outgrow your monolith, you lose the benefits of having a monolith.

A Simple Monolithic File Server

CAMBRIDGE Computer
ARTISTS IN DATA STORAGE



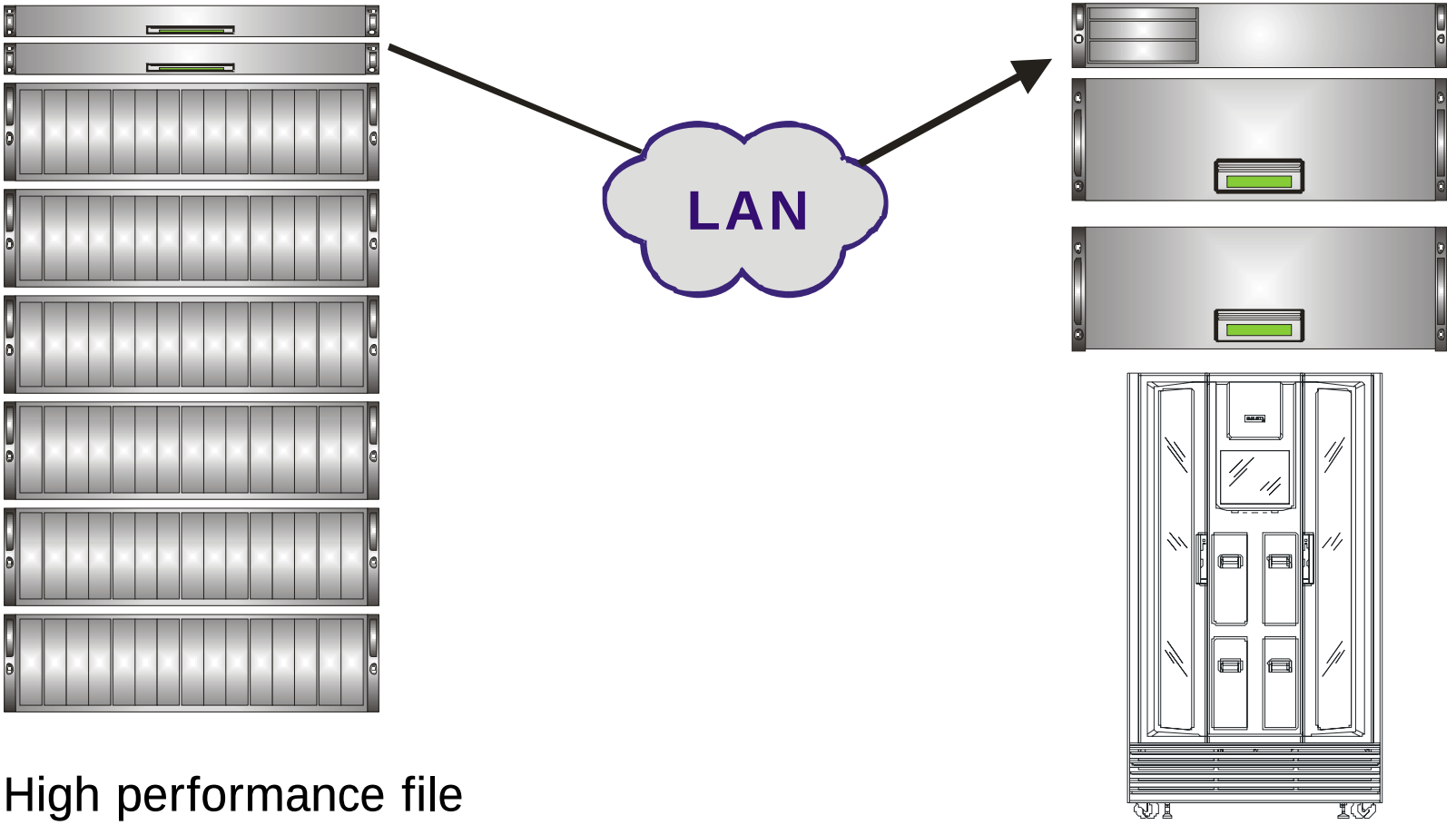
Replication and Backup of a Monolithic File Server

CAMBRIDGE
Computer
ARTISTS IN DATA STORAGE

- **Replication**
  - Replication can be as simple as a file copy.
    - Once a day, copy all files that were added since yesterday.
  - Alternatively, there are directory synchronization solutions on the market that are very inexpensive.
- **Backups**
  - Since the files do not change, there is no need to do full backups once a week.
  - Only back up unique files.

Using Archival File System (or CAS) as Backup Target for a Large File Server

CAMBRIDGE Computer
ARTISTS IN DATA STORAGE



High performance file system or NAS

Tiered Archival File System

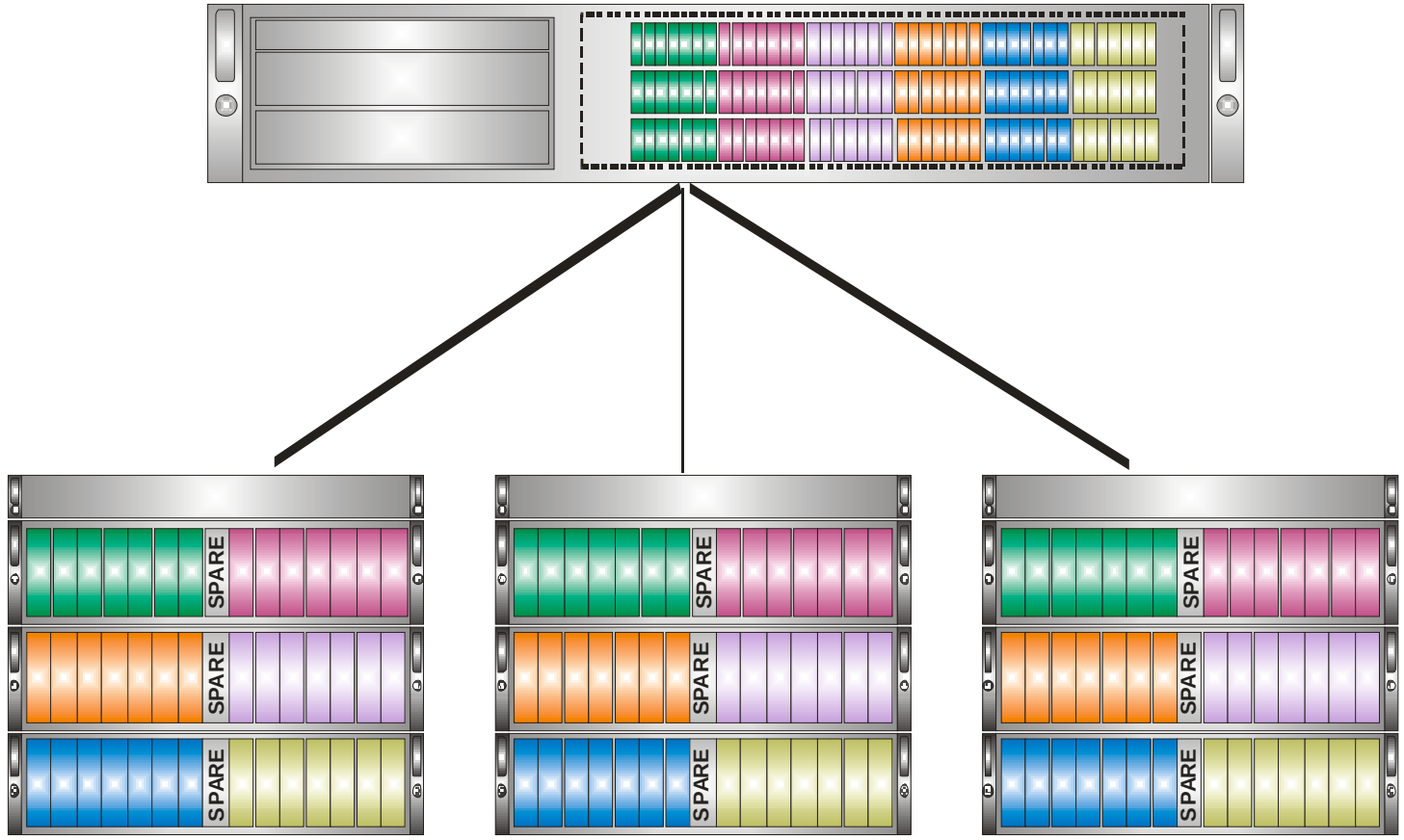
Aggregates

CAMBRIDGE
Computer
ARTISTS IN DATA STORAGE

- Aggregates are systems built of individual components that together behave like a monolithic system.
  - No one piece of the system can work by itself.
- The challenge with aggregates is that if any one component fails, the whole system could be compromised
  - The more components you have, the higher the probability of failure
  - Aggregates require some kind of redundancy model

What Happens When You Outgrow the Monolith

CAMBRIDGE Computer
ARTISTS IN DATA STORAGE

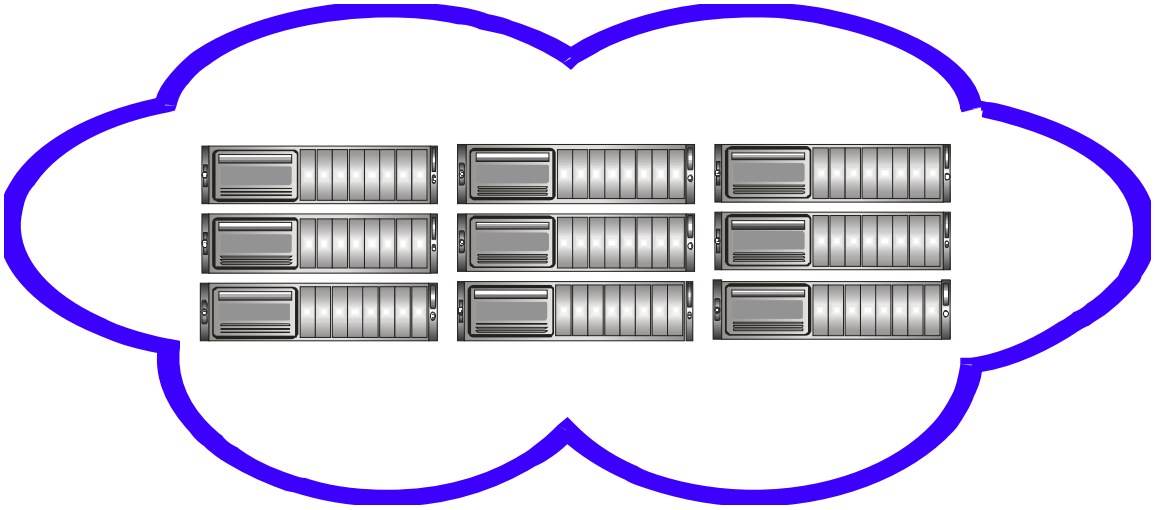


The more pieces you have, the more likely you are to have a piece fail.

Aggregated File System with Parity Protection

CAMBRIDGE Computer
ARTISTS IN DATA STORAGE

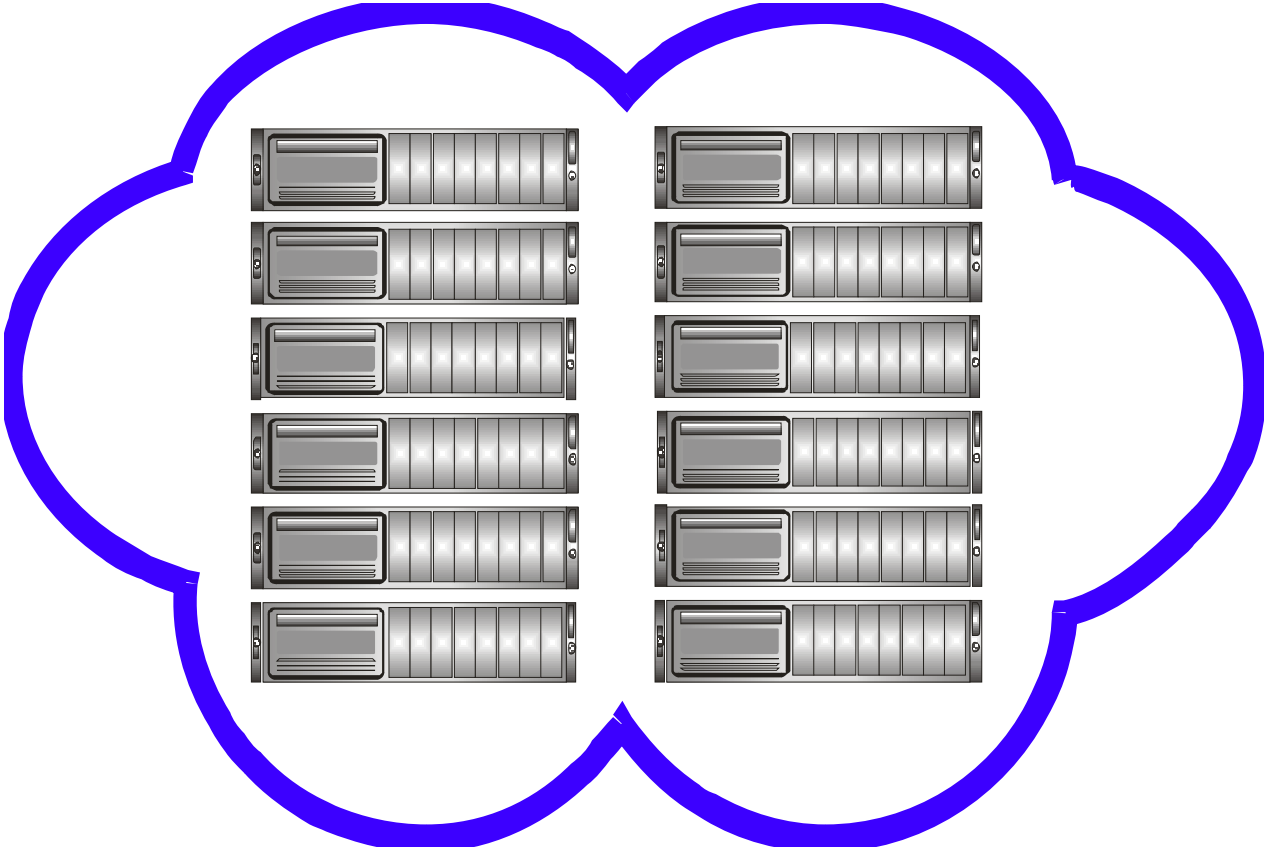
- Data is striped across members of the aggregate with a parity scheme that allows one or more boxes to fail.



Usable capacity is N-1. Some systems allows N-2, N-3, etc.

Aggregated System with 1:1 Redundancy

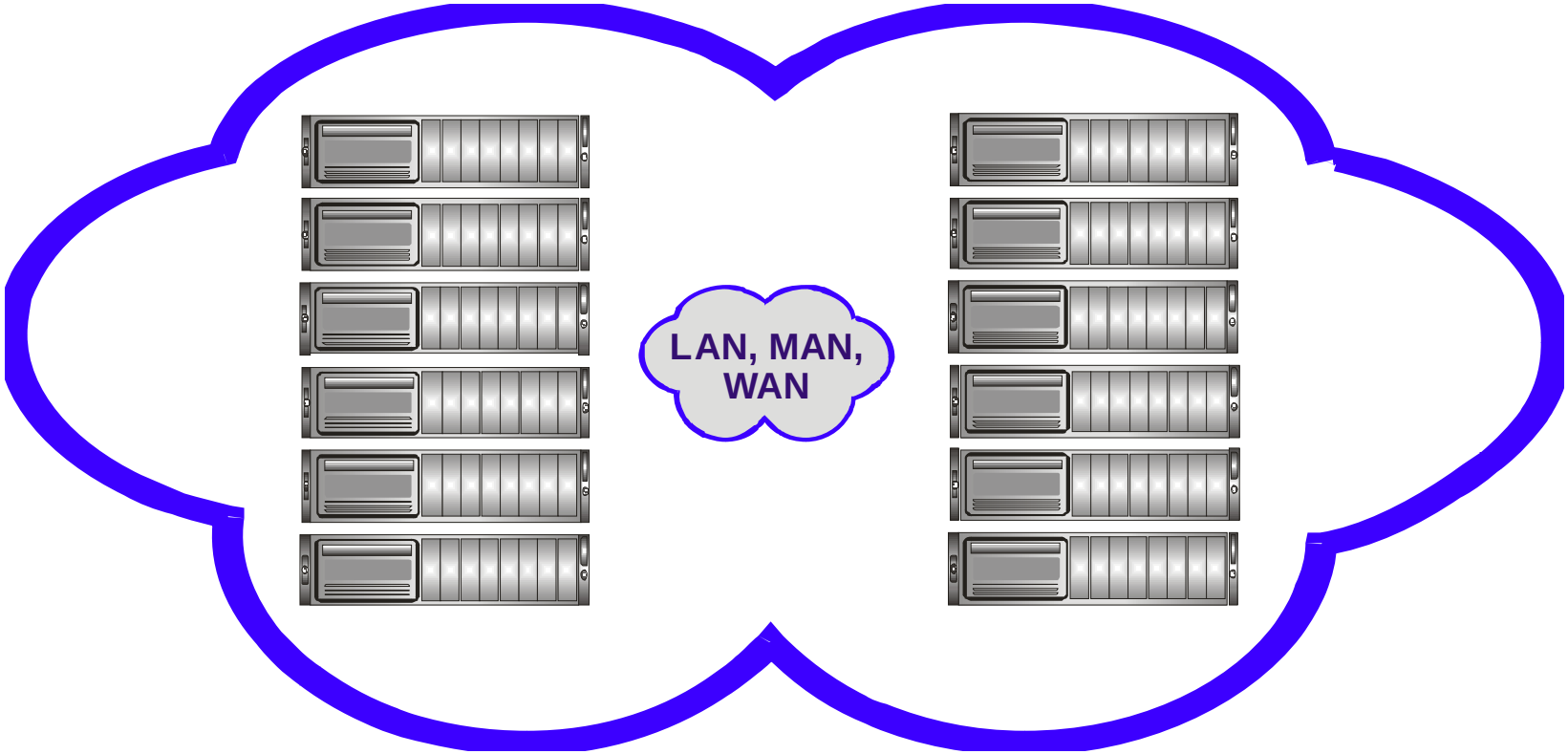
CAMBRIDGE Computer
ARTISTS IN DATA STORAGE



You lose 50% of your capacity. Maybe you go 3-way redundancy and lose 66%

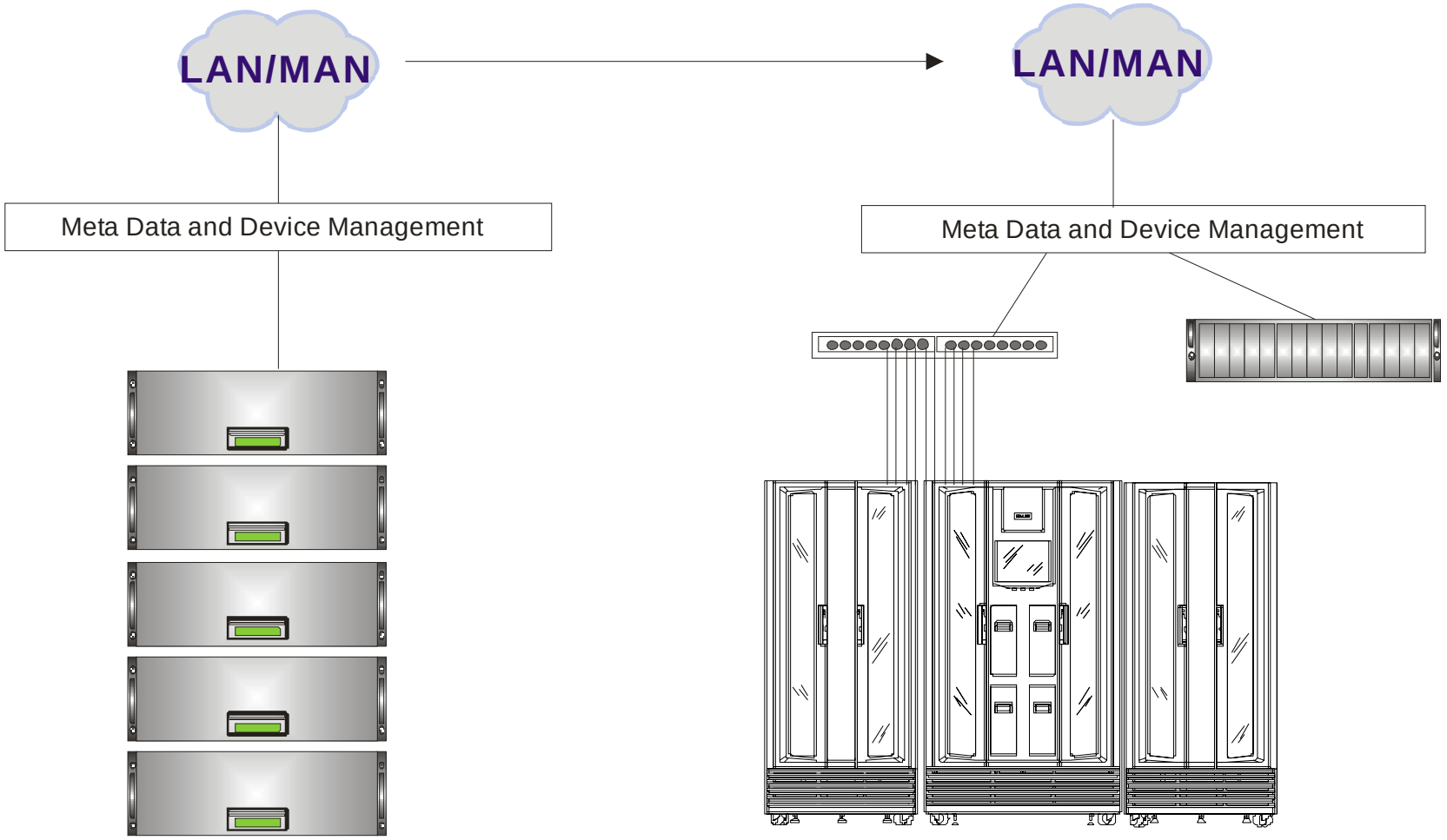
1:1 Redundancy Between Sites
HA, Fault Tolerance, and DR in One

CAMBRIDGE Computer
ARTISTS IN DATA STORAGE



N-Way Redundancy – Mirroring Between Disk and Tape

CAMBRIDGE Computer
ARTISTS IN DATA STORAGE



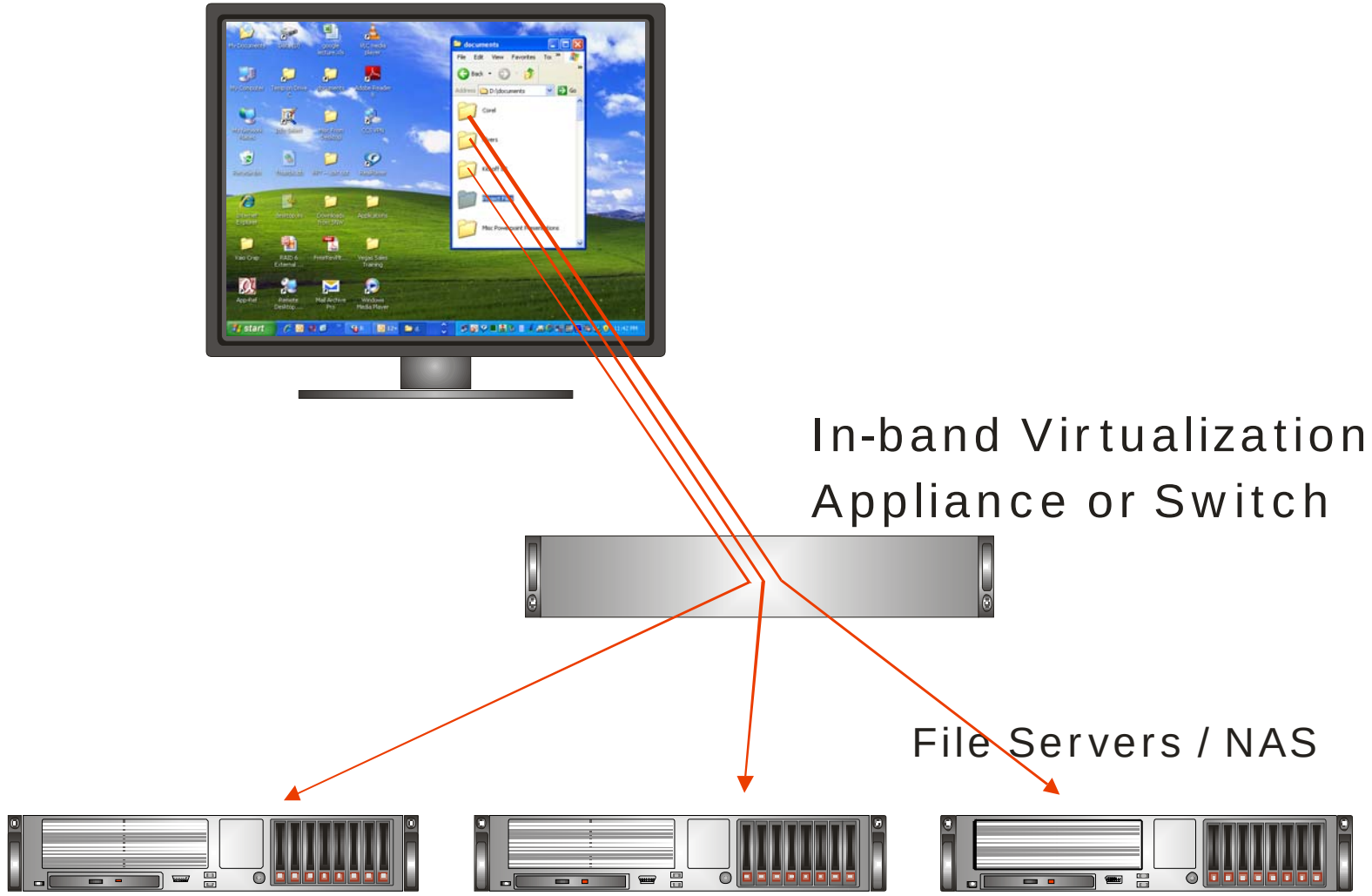
Federations

CAMBRIDGE
Computer
ARTISTS IN DATA STORAGE

- A logical grouping of independent systems that behave as one global file system.
- Different from aggregates in that each member of the federation is a free-standing element.
  - If the federations were to dissolve or melt down, data is still intact and accessible from the individual members.
  - Some federated solutions allow you to interact directly with the members.
    - This is particularly useful when federating with high performance file systems.
- Members of a federation can have very different properties.

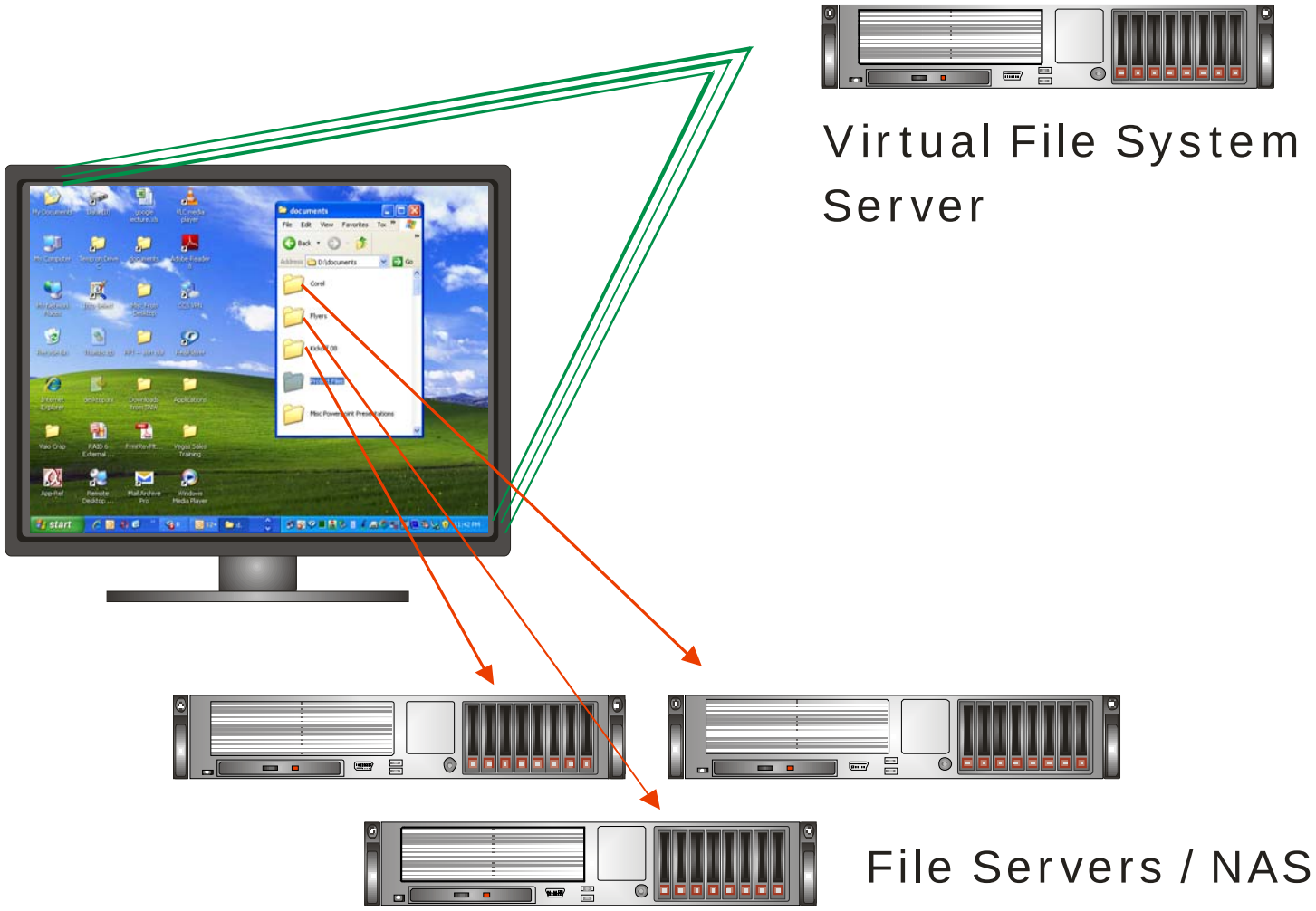
In-Band File System Federation

CAMBRIDGE Computer
ARTISTS IN DATA STORAGE



Out-of-Band File System Federation

CAMBRIDGE Computer
ARTISTS IN DATA STORAGE



Life Cycle Management with Federated Virtual File Systems

CAMBRIDGE
Computer
ARTISTS IN DATA STORAGE

- Most virtual file systems offer some kind of integrated tiering logic
  - Might work on the directory level or the file level.
  - Policy engines will vary in capabilities
- Sophisticated features:
  - Write staging – new data is written into high performance pool.
  - Transparent download to local disk for faster processing
  - Management of files across a wide area
    - Web portal
    - Caching and management of local copies

Content Classification with Federated, Virtual File Systems

CAMBRIDGE Computer
ARTISTS IN DATA STORAGE

- Extensible metadata
  - The central database can typically be customized to serve as a catalog of all files in the file system.
- Users can annotate their files and directories.
- Metadata can be associated with files as they are created and used.
- Tags from embedded file metadata can be ingested into the file system catalog
  - TIFF images, DICOM, etc.
- Virtual directories can be created with a SQL query.

Other Useful Features of Federated File Systems

CAMBRIDGE Computer
ARTISTS IN DATA STORAGE

- Federated user directories
  - Share files between different institutions with a single sign-on.
- File access auditing
- Backup
  - Maintain backup copies locally or in other sites
- Self service delegation of file access privileges
  - Users can easily grant access to their files to other users and groups.