Archival Storage

Lecture 11

November 21, 2006

Plan for today

- Design and Architecture for Fixed Content
- Paper discussion

Case study: EMC Centera

Designing a content-addressable government compliant-object store cluster

Digital and Mostly Fixed Content

Music & Video

YAHOO! music

MP3s & More (Audio & Video Downloads)

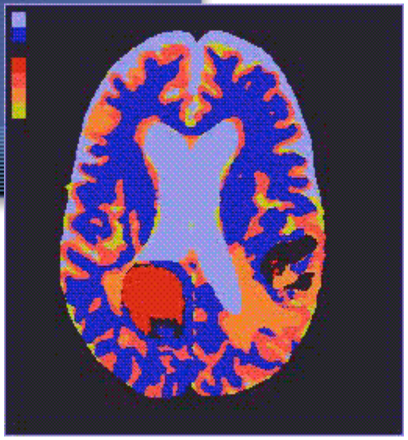
Downloadable Music & Videos

Search

WARNER BROS. PICTURES

Medical Imaging

Centricity™ PACS

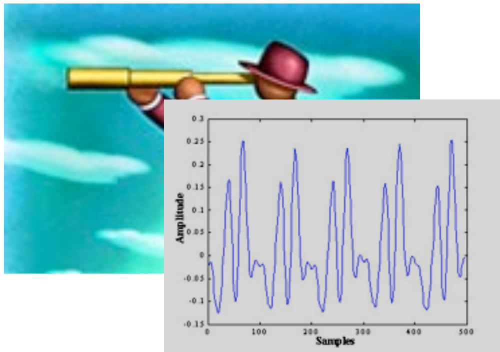


Microsoft® Exchange Server

Lotus.

E-mail

Video Surveillance & Voice Recording



RFID

MasterCard

VISA

Virtually all financial transactions

5 5th Mar 1974 01-06-35

National Westminster Bank Limited
Old Bond Street Branch
47 Old Bond Street, London W1X 4HS

Pay Tony Hill or order
Eleven pounds 96p £11—96.

PAID

APPLE PUBLISHING LTD

⑈632070 ⑈01⑉0636⑆ 01043315⑈ ⑇0000001196⑇

Photo Services

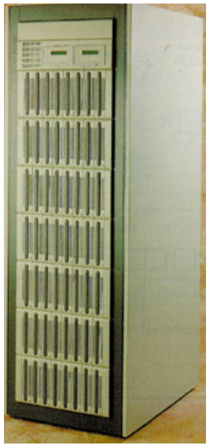
Kodak EasyShare Gallery

Get the government involved…

- Regulations and requirements for data retention
  - To prevent ENRON, catch Martha Steward, spy on people, etc.
    - Sarbanes-Oxley, SEC 17.4a…
  - Throw in contradicting regulatory policies
    - US laws vs. EU privacy protection

- Storage used to be easier (maybe)
  - Shred paper documents
  - Use WORM media
    - Write to tape, destroy it
    - Laser disk JukeBoxes
  - But…
    - Management challenges
    - Speed of access to data

Enter On-line Archives



1993 – $6/MB
Cost to store a 30MB object: $180

Can store only 400 objects

400:1 pricing difference

2003 – 1.5¢/MB
Cost to store a 30MB Object: 45¢

Can store 800,000 objects

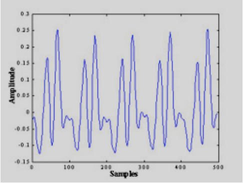
Now, it is economically viable to use on-line disk-based storage instead of WORM technologies

- Somewhat larger purchase price and higher operational costs offset savings in data and content management including fast access

First Generation of a New Access Method

- Recognize a legal record as a unit of transfer
  - Store an E-mail, an X-ray, a digital voice recording
- Handle basic legal record requirements
  - Retention, immutability, etc.
- Audit actions
  - Deletion is an obvious one
  - Auditing reads is important as well
- Handle Trillions of objects
- Why?

Document Mngmt.
File Backup



5

5th Mar 1974

01-06-36

National Westminster Bank Limited
Old Bond Street Branch
47 Old Bond Street, London W1X 4HS

Pay Tony Hill or order

Eleven pounds 96P

£11-96.

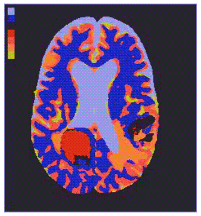
APPLE PUBLISHING LTD

PAID 18 MAR 1974

UPTON PARK

20-99-04

⑈632070 ⑈01⑈0636⑆ 01043315⑈ ⑇0000001196⑇



How do we Access Trillions of Items

- Use Content Addressing
  - Define a “GUID” address based on the content bit pattern
    - MD5, SHA-1, HAVAL, … hashes
- Have a flat large address space
  - No external explicitly maintained hierarchy
    - Internally, there must naturally be some hierarchy or structure
- Decouple the address/name from the structure of storage

Nothing new (so far)

Why is Content Addressing Important?

- **Content authenticity**
  - Unique “fingerprint” is generated from the content itself
  - Content is validated on delivery
  - Content integrity is continuously validated in background
- **Content Address is location independent**
  - Address is globally unique
  - Not a place in a hierarchy (file system)
  - Not a place in a disk array (logical volume)
- **Identical objects are only stored once**

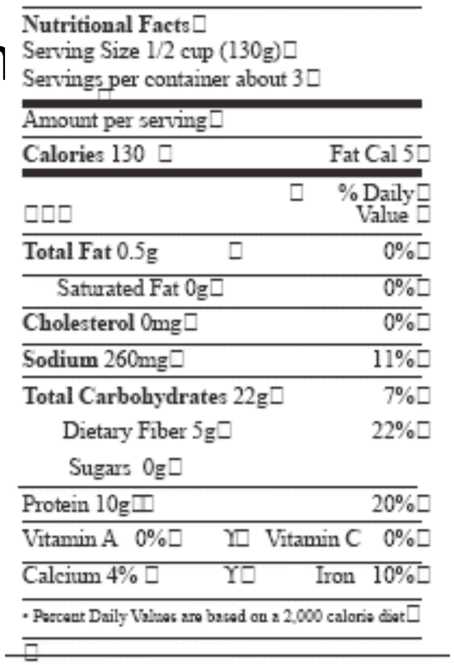
MetaData: The Second Key Component

Imagine a can shelf without labels

- can’t determ
- can’t figure
- don’t know
  has been on

Adding labels fixes many of these problems

Standardized labeling allows multiple vendors to consistently represent information to consumers



Nutritional Facts	
Serving Size 1/2 cup (130g)	
Servings per container about 3	
Amount per serving	
Calories 130	Fat Cal 5
	% Daily Value
Total Fat 0.5g	0%
Saturated Fat 0g	0%
Cholesterol 0mg	0%
Sodium 260mg	11%
Total Carbohydrates 22g	7%
Dietary Fiber 5g	22%
Sugars 0g	
Protein 10g	20%
Vitamin A 0% Y	Vitamin C 0%
Calcium 4% Y	Iron 10%

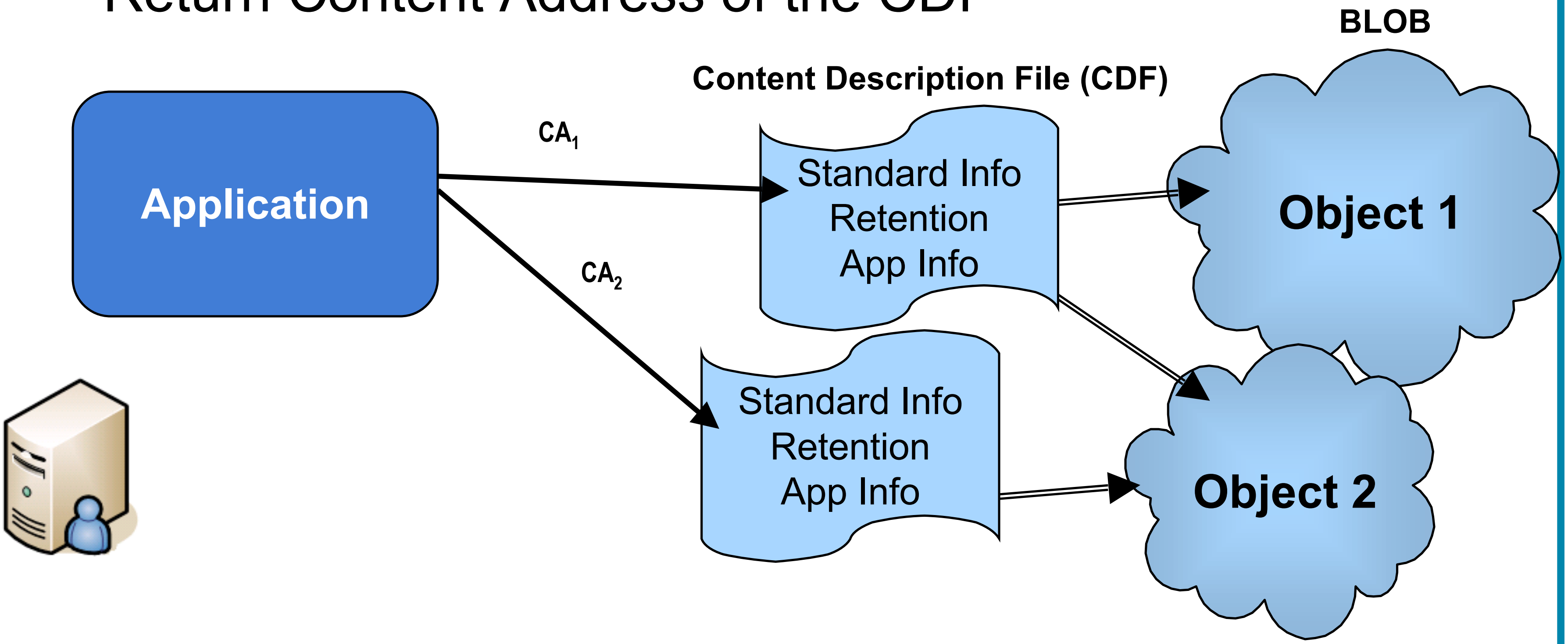
• Percent Daily Values are based on a 2,000 calorie diet

Centera: A (New) Kind of Object-based Storage

- **Stores Any Kind of *Fixed* Content**
  - Satisfies Non-erasable/Non-rewriteable regulatory requirements
- **Content Addressed Storage**
  - Content authentication
- **Extensible metadata stored with each object**
- **Scaling computational power with capacity**
  - **Computation close to data**

Centera Object Model for Fixed Content

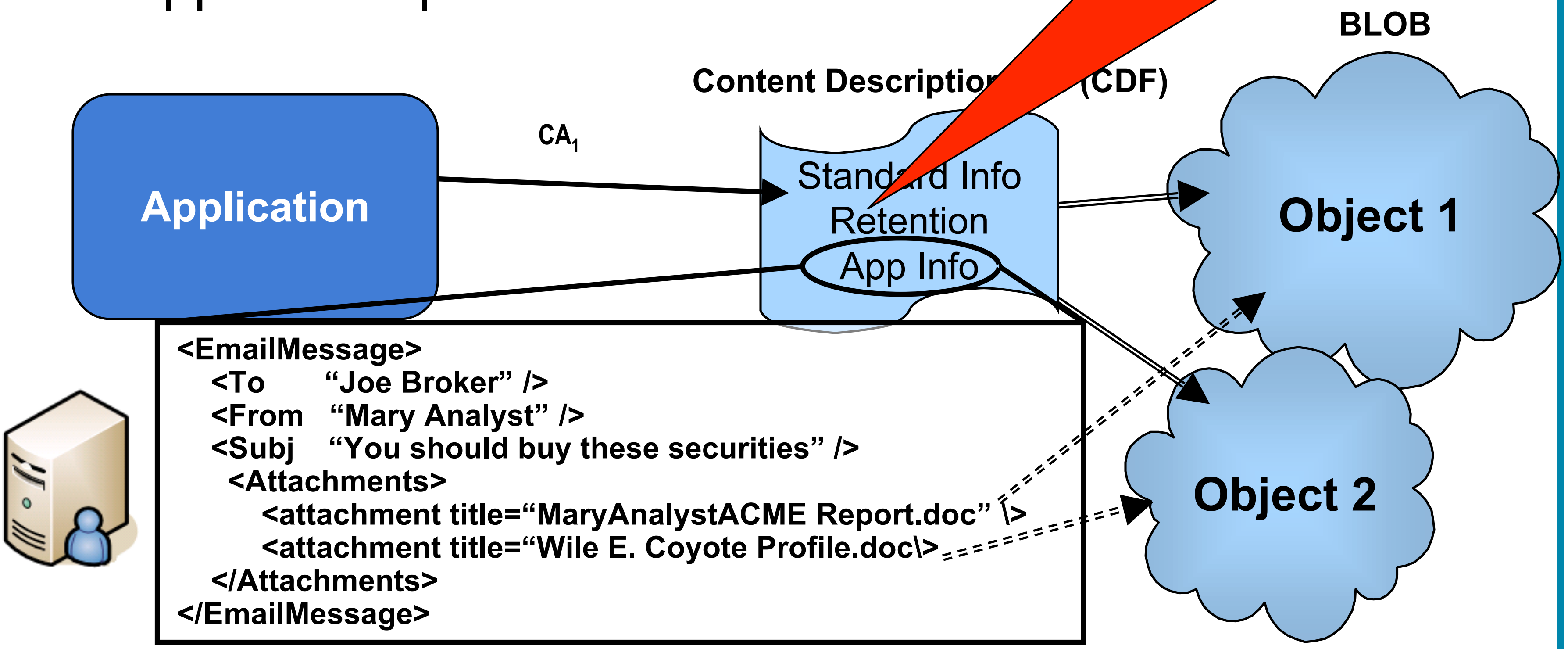
- Annotate fixed content with arbitrary metadata
- Store separately data and metadata (object attributes)
  - two Centera objects (CDF & Blob)
- Return Content Address of the CDF



CDF: Extensible Metadata

- Standard attributes
  - creation date, etc.
- Special added attributes
  - retention period, etc.
- Application-provided information

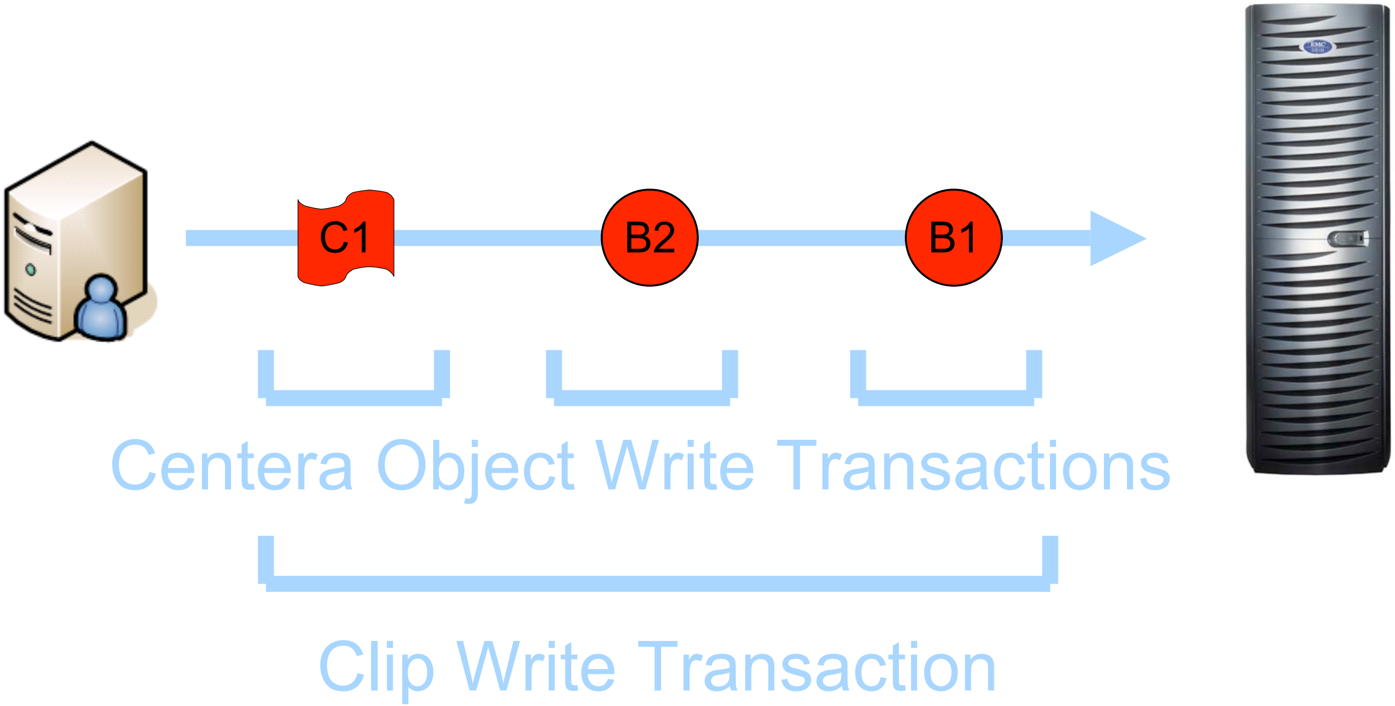
XML Object containing metadata and references/pointers to Blobs. The Content Address of the CDF is returned to the applications.



```
<EmailMessage>
   <To      "Joe Broker" />
   <From   "Mary Analyst" />
   <Subj    "You should buy these securities" />
     <Attachments>
        <attachment title="MaryAnalystACME Report.doc" \>
        <attachment title="Wile E. Coyote Profile.doc\>
   </Attachments>
</EmailMessage>
```

Write Transaction: CDFs and Blobs

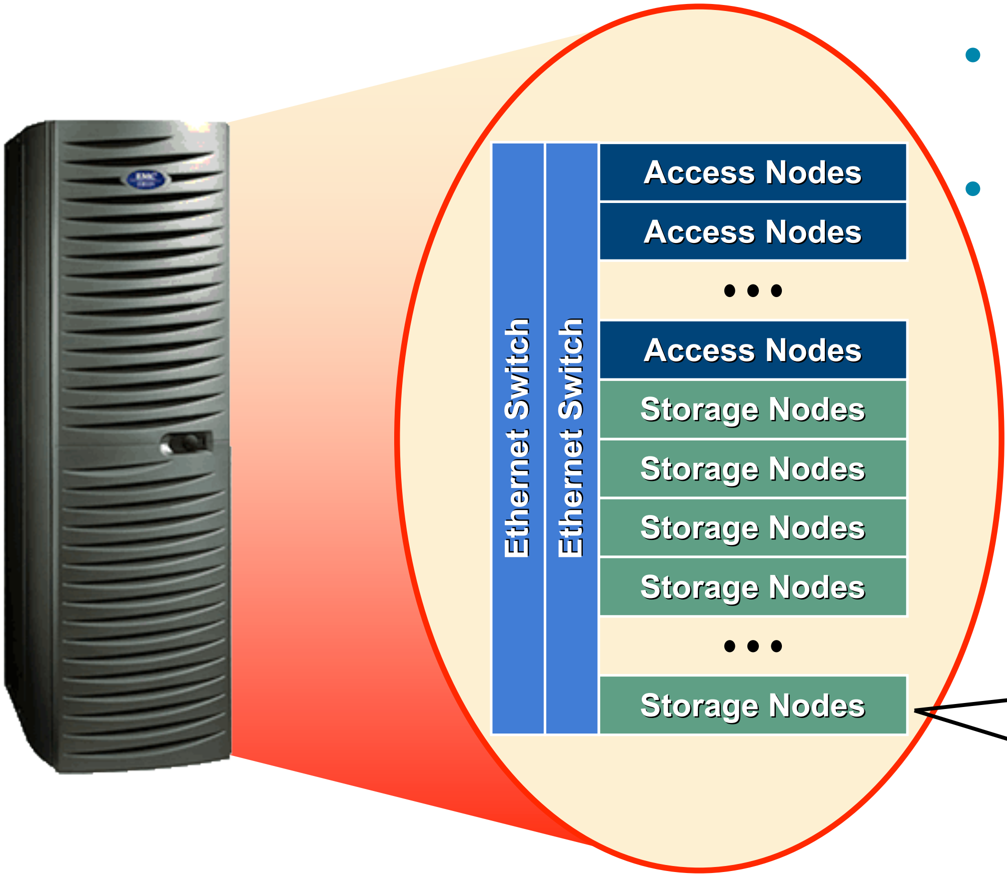
- Clip Write Transaction
  - N+1 Centera Object Write Transactions
  - CDF follows BLOBs



- Return CDF’s CA to the user
  - Content of CDFs defined by attributes and BLOB’s CA

Centera HW Architecture

- Node-based cluster
  - “homogenous”
- No single point of failure
  - Content protection
- Leverage Commodity HW
  - Cost-effective solution
  - Can do frequent HW refresh



Ethernet Switch

Ethernet Switch

Access Nodes

Access Nodes

• • •

Access Nodes

Storage Nodes

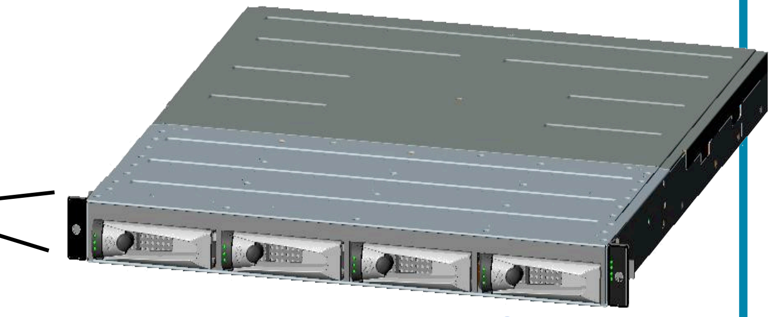
Storage Nodes

Storage Nodes

Storage Nodes

• • •

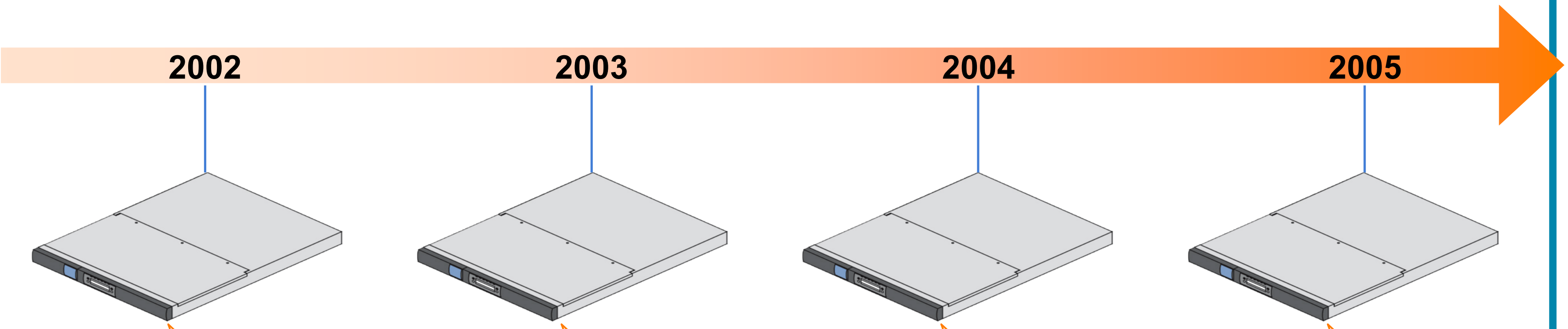
Storage Nodes



Intel Pentium 4 with 1 GB RAM
4 x 500 GB SATA
3 x Gb copper NIC

Access Nodes provide external API access
Storage Nodes store and protect information
A single physical node can have *both* personalities

Leveraging Commodity at Right Price Point



2002

Gen1
866MHz P3
256MB RAM
160GB Drives
100Mbit Network

2003

Gen2
1GHz P3
512MB RAM
250GB Drives
100Mbit Network

2004

Gen3
2GHz P4
512MB RAM
320GB Drives
1x GigE Network

2005

Gen4
2.8GHz P4
1024 MB RAM
320/500GB Drives
All GigE Network

5400 RPM P-ATA Disks
100 Mbit Internal Net

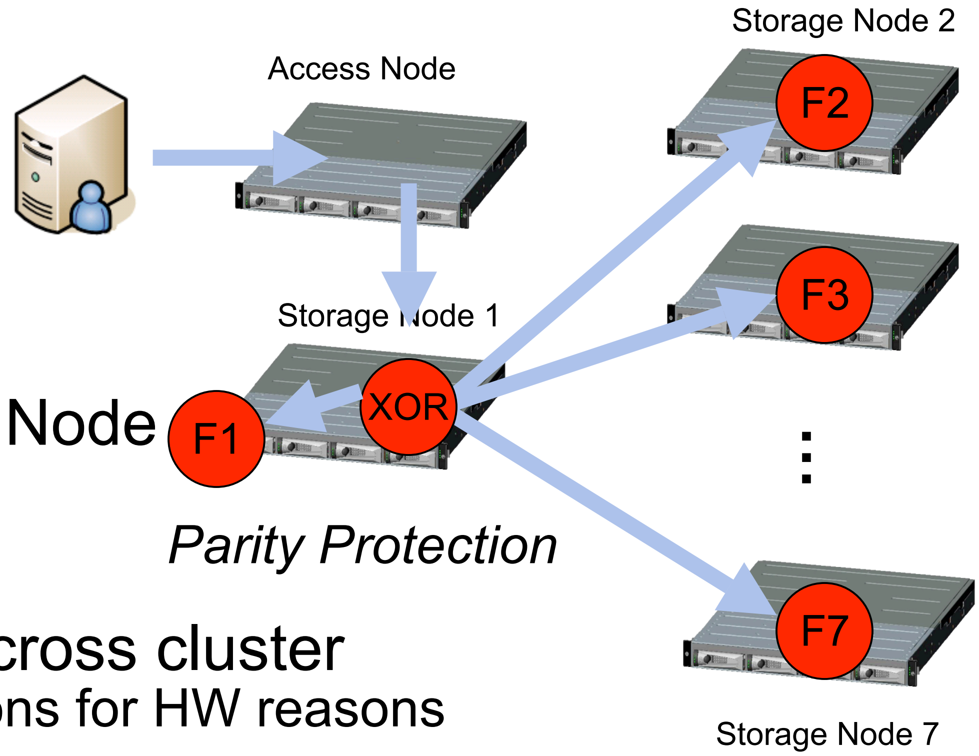
7200 RPM SATA Disks
GigE Internal Net

100 Mbit external AN connections

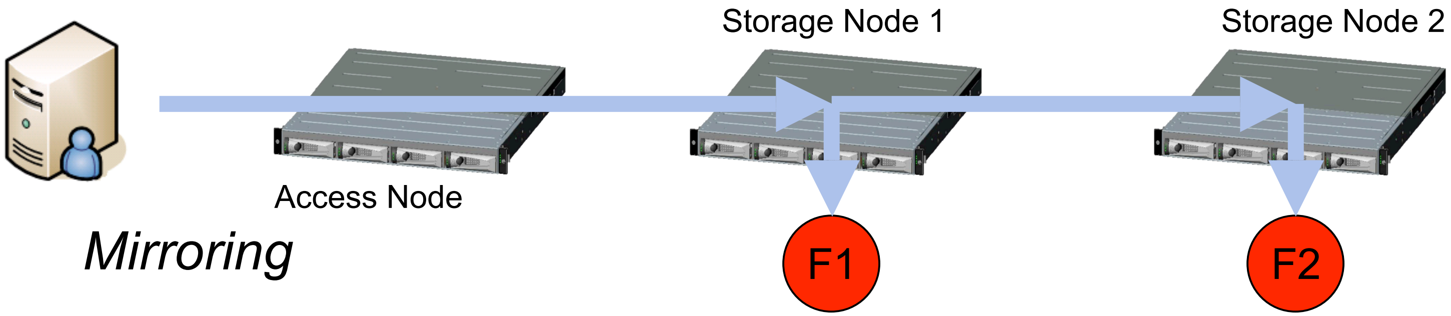
GigE AN external connections

Protecting Fixed Content Stored on a Cluster

- Protection Schemes
  - Mirroring
  - Parity Protection
- Access Node = Storage Node
  - Logical Node Roles
- Fragments distributed across cluster
  - some placement restrictions for HW reasons

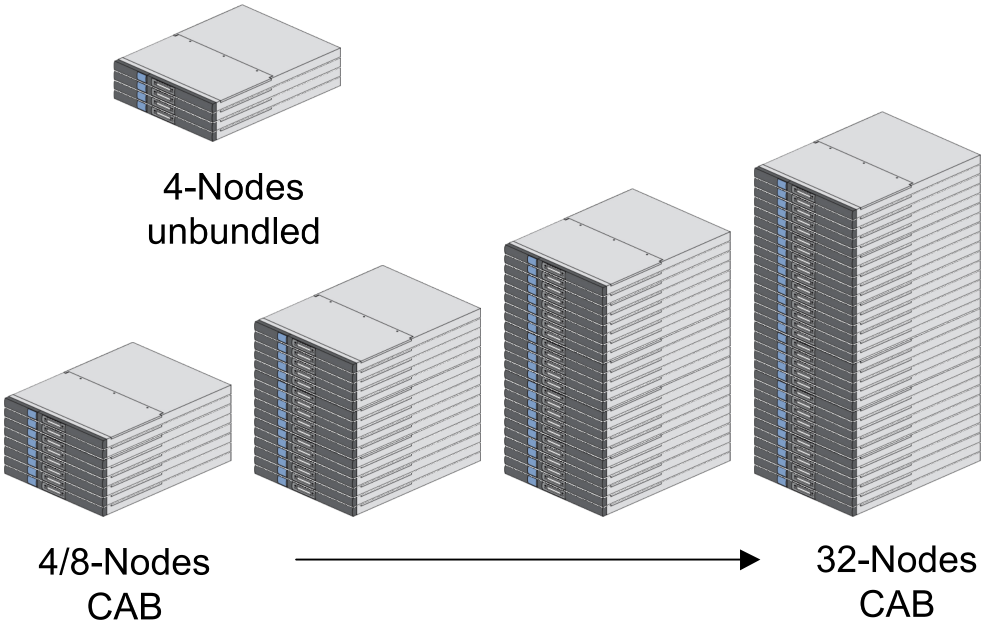


Parity Protection

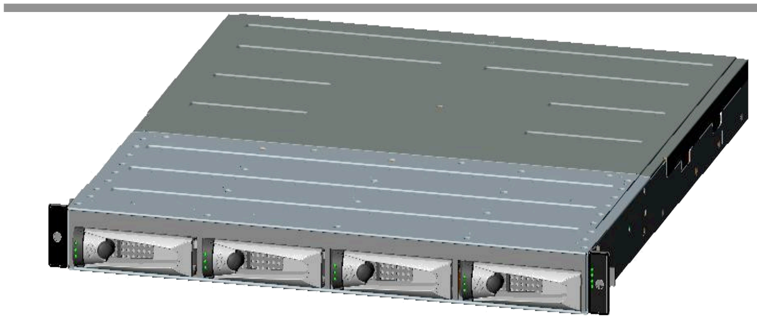


Mirroring

HW Packaging and Scalability



Multiple cabinets in a single cluster

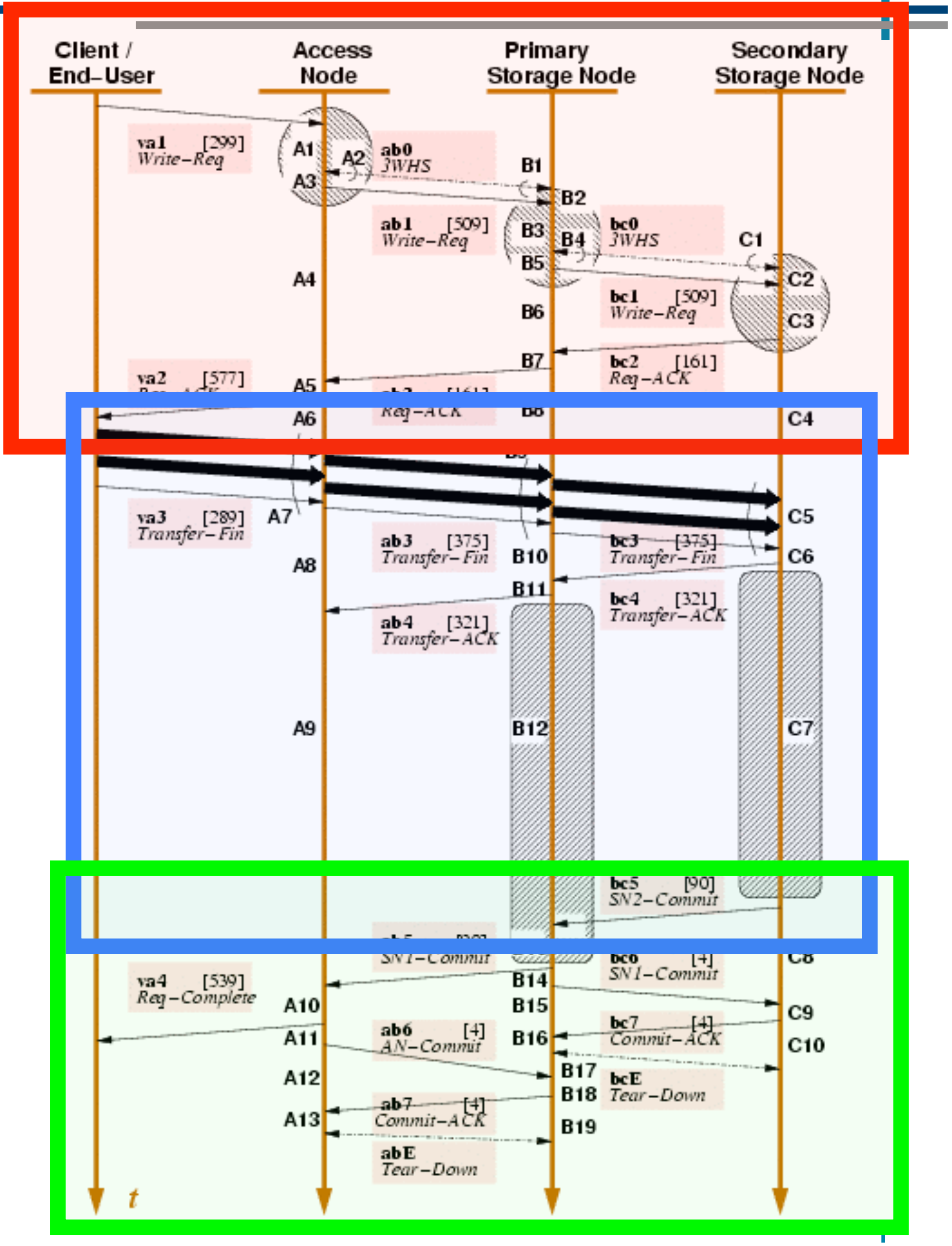


Intel Pentium 4 with 1 GB RAM
4 x 500 GB SATA
3 x Gb copper NIC

- Centera Cube
  - two switches for redundancy
  - 2x GbE to facilitate additional racks.
  - 32 or 16 nodes
- Centera Cabinet
  - No single point of failure
  - Dual AC power
  - Half nodes/switches are on one AC rail
  - One or two cubes
- Scalability (4-128/256 nodes)
  - Take advantage of Parallel Processing
  - Link cubes through uplinks and “root” switches
  - Scale CPU with Storage Capacity
    - Processing Power
    - Bandwidth

Object Write Protocol

- Initial
  - Misc. checks
  - Load balancing
- Data Xfer
  - Write to nodes
- Commit
  - Respond to client
  - Update internal metadata



Key:
Data reliably stored before ACK

Write Protocol Features

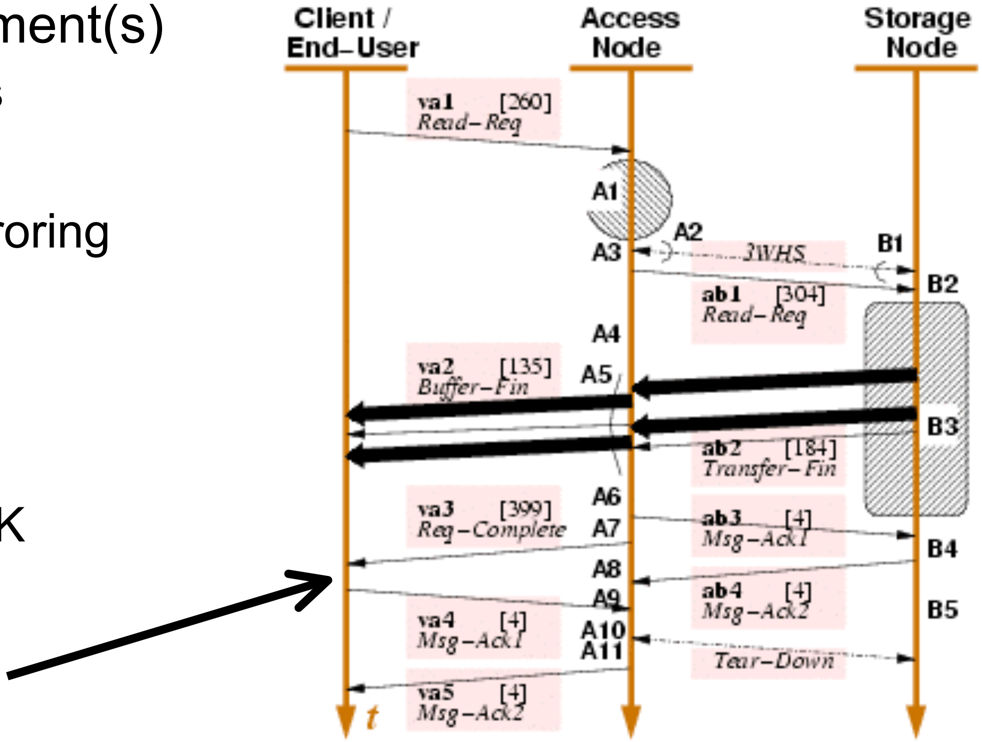
- Workload changes over time affect protocol changes
  - large objects, single-threaded access
  - Small objects, multithreaded access

- Current version
  - Persistent TCP connections throughout
  - Messages not sent over UDP anymore

- Leverage & harden commodity components
  - FS journaling allows grouping of multiple fsync()s
  - Write-barrier patch to Linux
    - can use drives with Write Cache ON
      - FLUSH Disk Cache
      - Write
      - FLUSH Disk Cache

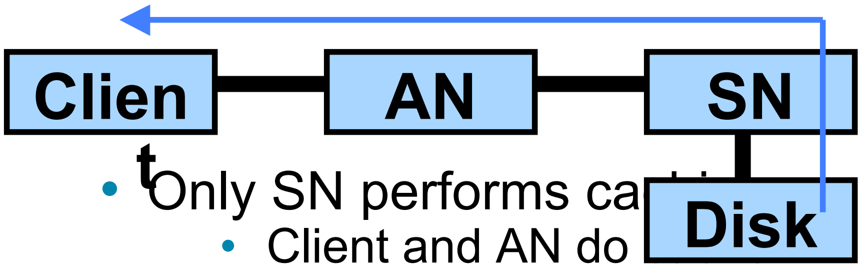
Object Read Protocol

- A1: Check access credentials
- A1: Locate object’s fragment(s)
  - DHT across all nodes
- Select Suitable SN(s)
  - Load balancing if Mirroring
- Transfer Data to Client
- Check integrity/ACK
- Exchange update info
  - Piggyback on the ACK

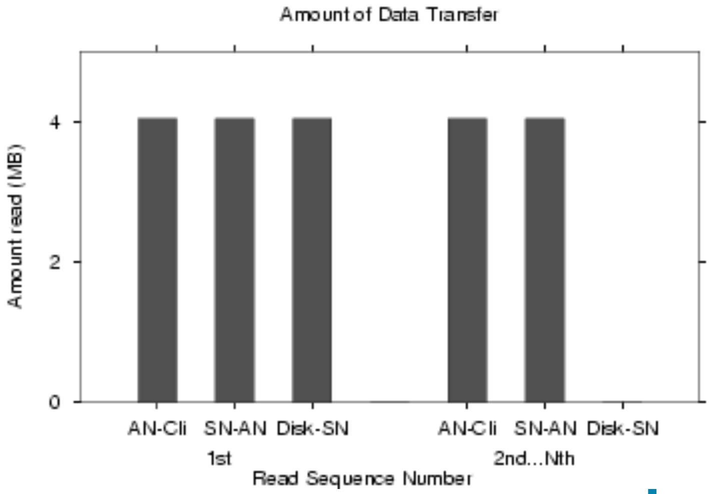


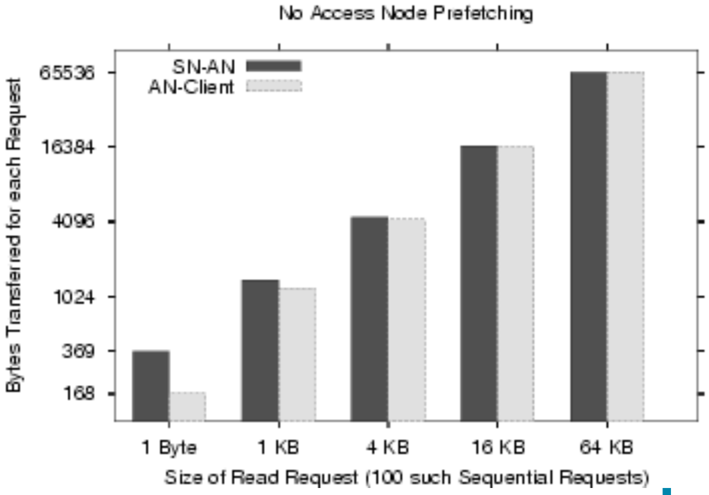
Key:
Data integrity is checked end-to-end for every read

Read Caching and Prefetching



- Only SN performs ca
  - Client and AN do
- Only SN performs prefectching
  - No prefetching across network





- Leverage commodity components
  - Node-local FS prefetching
- Workload focus
  - Reads are not (yet) primary workload
  - Writes and queries are

Architectural features at a glance

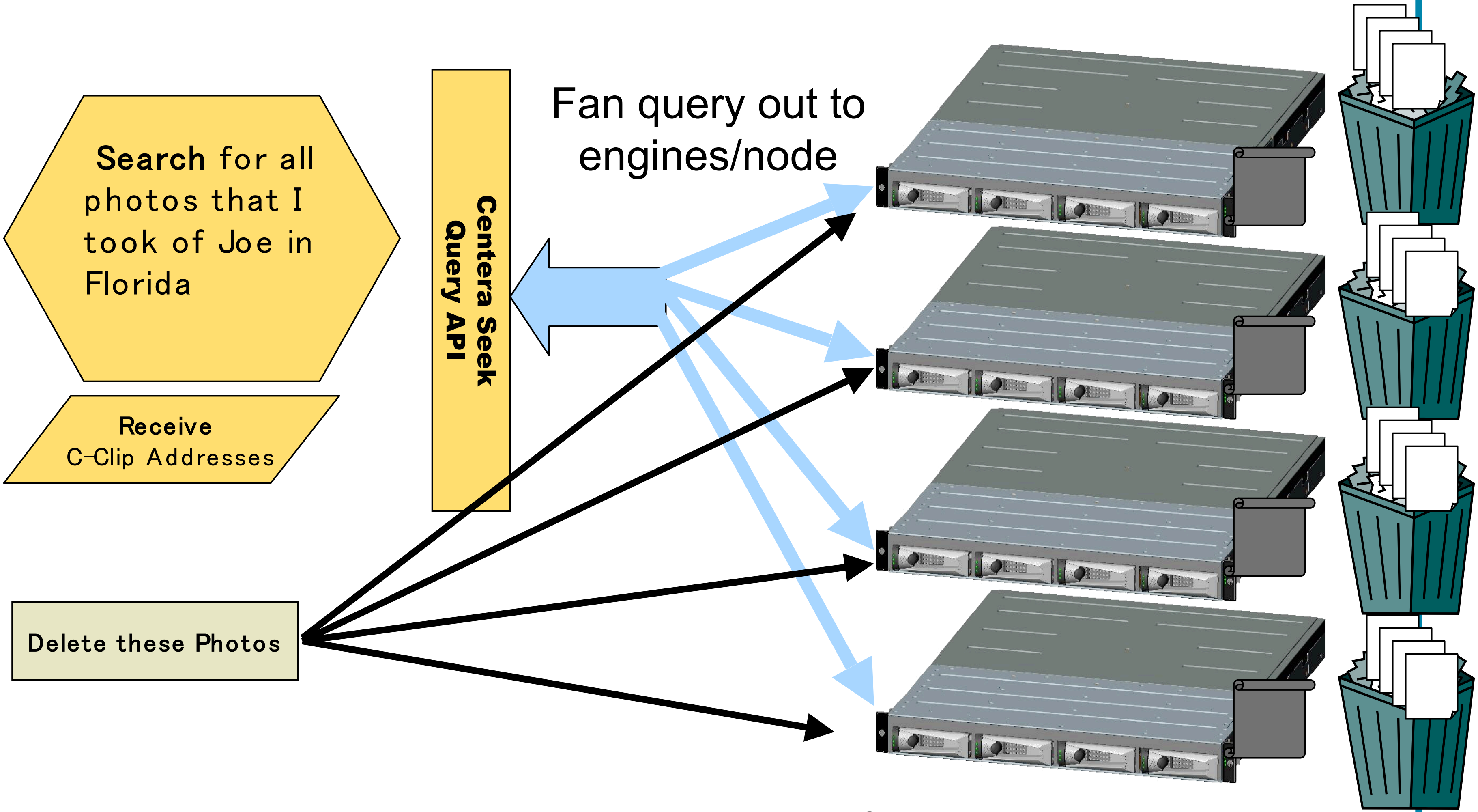
- Centera is a form of an Object Store
  - each node is a storage “brick” with object interface
    - “brick” stores a fragment of an object
- Self-healing
  - Regenerate missing fragments due to disk/node failure
  - Per-object recovery can take advantage of parallelism
- Reclamation of capacity
  - garbage collect unreferenced/expired objects
- Content integrity checking
  - Proactively scrubbing data against checksum (Content Addr.)
- Single instancing of information at object level

Architecturally not unlike other “brick” projects

“Plain ol’ Google” isn’t good enough

- Finding information is about crawling and indexing
  - Ad hoc organization
  - You cannot plan on organization for long-term retention
- Sometimes you need very precise query results
  - Google gives you best efforts
- Temporal views are very important
  - Display my most recent work
  - Display the work I did before I made the mistake
- Context can be crucial
  - Show me the E-mail received by Joe Broker from any person who is an Analyst
  - Find all the legal documents
- Sometimes you want to index “non-text-based” information

Beyond Data Mining



Search for all photos that I took of Joe in Florida

Receive C-Clip Addresses

Centera Seek Query API

Fan query out to engines/node

Delete these Photos

Query results may stay resident on nodes

Centera API: Making it all possible

- Functions implemented by an application-linked library
  - Authentication & access control
  - Load balancing across cluster's Access Nodes
  - Data xfer (R/W)
  - Querying for content (with sufficient permissions)
- Applications can take advantage of Centera-unique features

- Alternative access methods
  - Centera Universal Access server (NFS, CIFS, http, ftp)
    - Applications can use some Centera features w/o rewriting/recompiling
    - Not meant for performance, or as a "pure NAS" device
  - Centera AP is not standardized: XAM standard proposal
    - effort initiated by EMC, IBM, Sun, HP, Archivias
    - SNIA "Content-Aware Storage" TWG

Centera Security and Auditing Features

- Access privileges based on application and user profiles
  - Audit logging on access/read/write/delete

Feature	Basic	Governance Edition	Compliance Edition Plus
Retention Enforcement			
Default Retention			
Purge Blob			
Delete CDF			
Audited Deletes			
Privileged Delete			
Content Shredding			
Remote Management			

Summary: What works well

- Scaling R/W performance with cluster size

- Object Model
  - Content Authenticity
  - Governance Edition & Compliance Edition+
  - Applications routinely extend attributes via XML
    - overcome the performance overhead of XML through indexing

- Taking advantage of commodity components
  - Build on top of high-level constructs
    - OS features: Linux (Q&A) with FS
    - Put everything into user-level processes
  - Fast refresh of both HW and SW

- Alternative access methods
    - Not all applications can/are willing to change
    - Over time can migrate to Centera API/XAM

Fixed Content-Related Research

- Common block redundancy elimination
  - How to do storage savings, single instancing of 2-4KB blocks
  - Works better at larger scale, but hard to make it work distributed
  - compression works just as well
- Single-instance storage
  - “object-level” is sufficient
  - not for free even though it is “trivial” with Content Address
- Content-based intrinsic data placement (DHT-like)
  - Want the flexibility of placing data where I want it
  - Migration of content is important (HW refresh, etc) without any churn
    - Physical transfer and simple update of a reference works great

Alternatives: NetApp R200

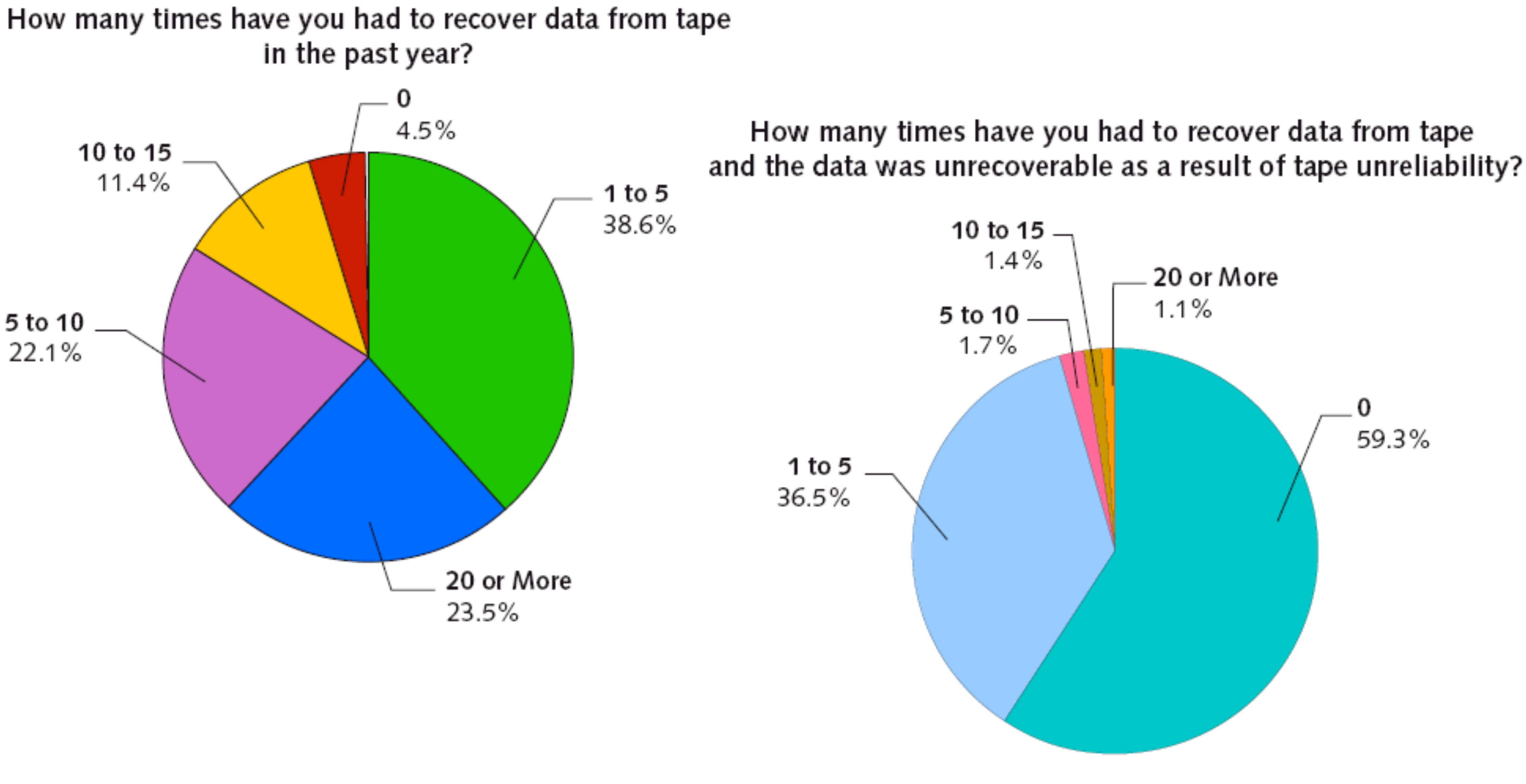
- FAS 960 Filer head
  - 2 uP's
  - 14 320GB drives per DS-14 shelves
  - 2 to 24 shelves in two racks (96 TB)
  - PATA drives with SATA-FC adapters
- $/MB declines as capacity grows
  - But limits scalability
  - Performance starts at higher
- Access Method based on NFS/CIFS
  - Retention Data
    - Volumes with special semantics
      - set atime to retention period
      - Commit by setting bits to R/O

Let’s do Some Numbers

Exhibit 1

Customer Recovery of Data from Tape

Source: Sunbelt Software and the Yankee Group, 2004



How many times have you had to recover data from tape in the past year?

How many times have you had to recover data from tape and the data was unrecoverable as a result of tape unreliability?

Note: Totals may not equal 100% due to rounding.

What's next …

- Next lecture: 11/28
- Readings will be posted tomorrow