

Introduction to CAS



Centera Overview

© 2005 EMC Corporation. All rights reserved.

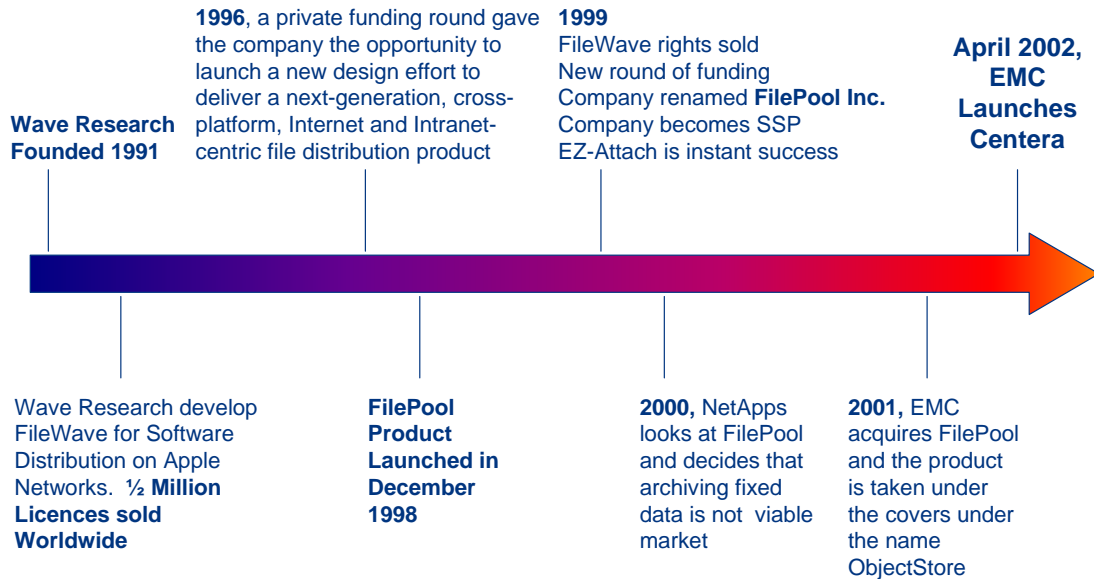


Introduction to Content Addressed Storage

Upon completion of this module, you will be able to:

- Explain the qualifications of fixed content
- Describe what information does not qualify as fixed content
- Differentiate between CAS, SAN and NAS
- List at least five advantages CAS has over traditional archival solutions
- List and describe five attributes of Centera functionality
- Describe the three different compliance licenses offered
- Describe the contents of a CDF
- Label and describe the physical components in a Centera rack

Centera's Timeline



© 2005 EMC Corporation. All rights reserved.

Module Title - 3

Mercer Road
Global Education & Productivity

Where Centera came from...

Quick Vocabulary

- BLOB: A unique data file, the atomic unit of work for a Centera
- CDF: Content Descriptor File, an XML based pointer to BLOBS
- Content Address: The globally unique, hashed address of a CDF or BLOB. Up to 64 alpha numeric characters with either a lowercase “e” in the center or a lowercase “x” or lowercase “r” for reflection (CDF or BLOB)
- C-CLIP: A package containing the users data and associated metadata. This term is often used interchangeably with Content Address
- XML Tag: An Element within a CDF
- Access Node: What talks to your LAN
- Storage Node: What holds all of your BLOBs and associated CDFs
- RAIN: Redundant Array of Independent Nodes
- Cluster: Up to 8 cubes of 16* nodes each. Cubes are interconnected by GB Ethernet
- Pool: A virtualized concept of a single large “pool” of storage. Typically bounded by a cluster
- S/AN: Storage on Access nodes, describes the ability to store data on access nodes

*Cubes made up of gen 4 hardware can have up to 16 node, older clusters can have up to 32 nodes.



What is Content Addressed Storage?

CAS: Is it a Revolutionary Paradigm Shift? (Or Just another TLA?)

TLA stands for “Three Letter Acronym”

What Exactly is Fixed Content?



**Generate
New Revenues**



**Improve
Service Levels**



**Leverage
Historical Value**

Digital assets retained for active reference and value

Electronic Documents

Contracts, claims, etc.
E-mail and attachments
Financial spread sheets
CAD / CAM designs
Presentations

Digital Records

Documents
— Checks, securities trades...
— Historical preservation
Photographs
— Personal / professional
Geophysical
— Seismic, astronomic, geographic

Rich Media

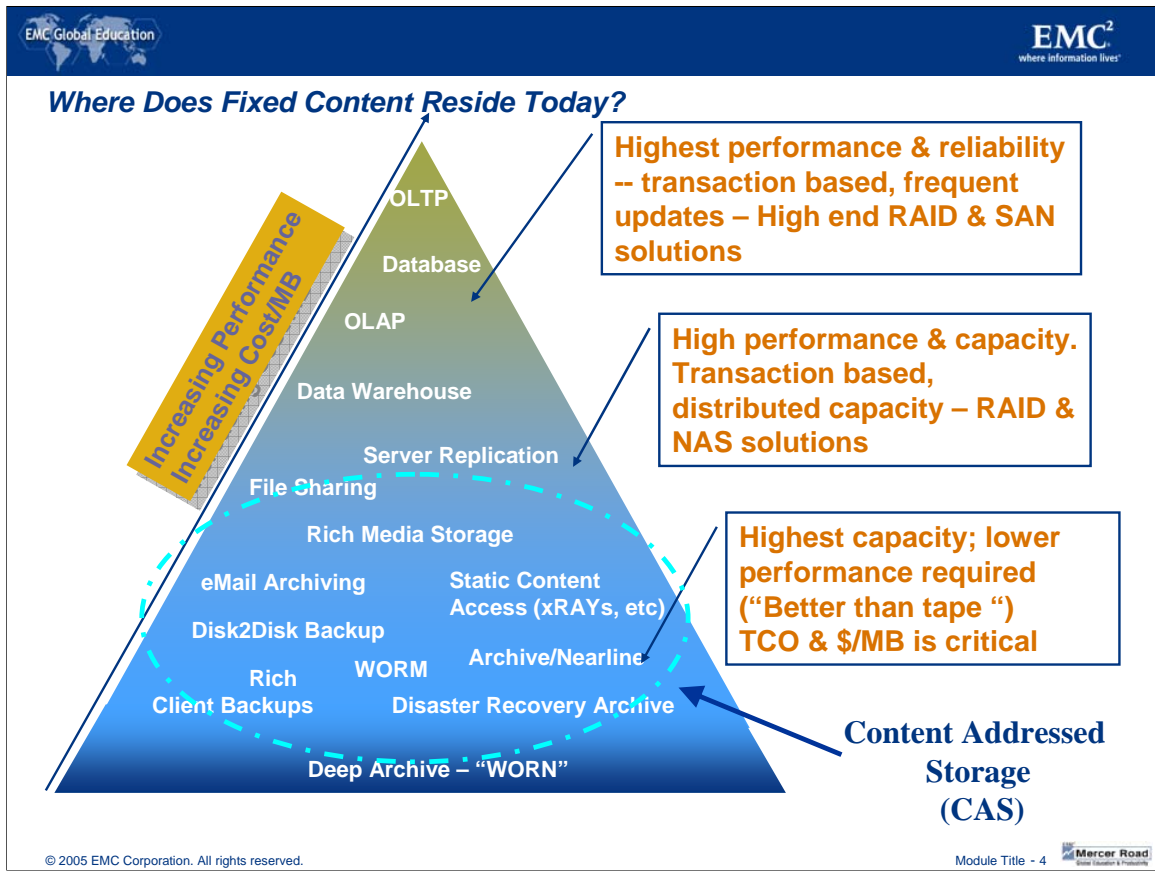
Medical
— X-rays, MRIs, CTI...
Video
— News / media, movies...
— Security surveillance
Audio
— Voicemail, radio

To understand CAS, it is helpful to understand where it is targeted and positioned. CAS is a unique concept that is specifically targeted to store “final form” (unchanging) data. Recent independent research shows that 75% of all information currently produced is fixed content. This is a market that was traditionally occupied by tape and optical disk and this IS the market that centra is trying to capture.



What isn't fixed content?

- Any information that has the propensity to change
- Documents that are “current”
- Documents that are not complete
- Documents that support On Line Transactional Processing (OLTP). An example of this would be a document that is used to keep track of your bank balance
- Data warehousing applications such as catalogs
- Any documents exposed to file sharing applications



Just to get an idea of the market for fixed content and where that data currently resides---

Where does the CAS market and Centera fit in the overall scheme of things??

Below SAN/DAS

Above WORM

And overlapping TAPE and Optical, and Low-end NAS and some NAS
(high capacity, lower performance)

Performance precludes it from web server type storage solutions



So, what's the problem?

A closer look at the issue reveals two things:

Storage is cheap...

COMPLEXITY & PEOPLE are expensive!

Consider the following points:

- A Terabyte of IDE storage costs ~\$1000 at maxtor.com. A P4, 3.0Ghz 512MB RAM workstation to mount it on the network is another \$499 at DELL. 20 machines like this give you 20,000,000,000,000 (2×10^{13} bytes) of NAS for under \$30k
- Human Resources, (people) are reoccurring costs that never go away

Commodity hardware alone is not expensive but if you couple commodity hardware with the inherent intelligence to replicate, regenerate, scale and manage itself, you add significantly more value because of what you can now do with the device.



Acquisition represents only 20% of TCO

You COULD buy cheap disk (JBOD storage) but that model has serious flaws:

- The JBOD model is inherently dangerous. It is often implemented as a RAID 0 solution as it is *cheaper* than RAID 5
- Not Scalable: There are now multiple, independent systems to integrate, maintain and ultimately fail
- Point-in-time Technology: Pain and cost to upgrade keeps legacy technology around long past it's usefulness
- It is much better to abstract storage from the systems used to access it

Source: IDG Reports

© 2005 EMC Corporation. All rights reserved.

Module Title - 6



JBOD “is what it is”, it doesn’t have imbedded intelligence to support any of the RAID standards and there are no provisions for long term admin and growth...They have to be managed as separate autonomous systems. This adds significant cost to the solution over the long term. Salary, benefits, office space etc. are all recurring management costs that never go away. It is extremely short sighted to disregard these costs and not account for them as part of the cost of the solution over time.

Additional challenges fixed content presents that preclude a NAS / cheap ATA solution

- Systems require guaranteed content authenticity and availability
- Long-Term Preservation Required (years-decades)
- Frequent and Fast Retrieval ("Internet" speeds)
- Minimal management and maintenance
- Shielded from technology obsolescence
- Tremendous Growth (pace is increasing)
- Simultaneous multi-user access
- Certain types of content present challenges in efficiency

All of these issues add to the complexity. These issues have to be managed to deliver low TCO... otherwise you are better off using tape, optical, or some other alternative archival solution.



© 2005 EMC Corporation. All rights reserved.

Module Title - 7

Mercer Road
Global Education & Productivity

Fixed content has attributes unique from any other type of content created today. These requirements are driven by a combination of business rules and regulatory standards that dictate the way these records are stored.

Typically fixed content is written once and then stored, unchanged over its lifecycle. During this retention the content must be guaranteed to remain authentic and accessible 100% of the time.

Regulatory requirements stipulate that in some industries this content be kept for 30+ years if not longer. Here are some examples of retention requirements:

- Seven years for check images, SEC information
- 21 years for Pediatrics
- Mortgage life plus 7 years
- Life of the patient for cancer care

Other requirements include:

Frequent and Fast Retrieval – although not accessed on a transactional basis fixed content is typically referenced in customer facing applications or time constrained situations (audit, medical emergency) for this reason online response times are becoming more and more critical.

Minimal management and maintenance – because of the low touch attributes of this content customers don't want to waste all their IT resources managing libraries, drives and media. For this reason fixed content, especially as it grows should require minimal storage management and maintenance. IT cycles should be spent on far more productive projects.

Module Title - 7

Mercer Road
Global Education & Productivity

Drawbacks of alternative archive solutions

- “How can I demonstrate that the records I’m turning over to the auditor are a precise copy of the original.”
 - **Current technologies do not guarantee record integrity**
- “It has taken more than 2 months to retrieve all my data back from optical.”
 - **Current media may be either on the shelf, not mounted or misplaced**
- “At these growing capacities managing all this content is a constant headache.”
 - **Current systems were not designed to handle today’s Web and Email volumes**
- “The vendor said the only place I can get a drive to read my platter is in a salvage yard”
 - **Media vendors are abandoning the optical storage market**
- “Even though my content has retention periods I can never find it to throw it away.”
 - **Media level management creates liability and inefficiency**



Current media alternatives, including tape and optical have many drawbacks. Speaking with customers we have listened to their biggest concerns with the current technology they are using to store fixed content. As you can see there is a clear demand for a better solution to storing fixed content.

EMC's Response: Centera

Purpose-Built Magnetic Disk Records Storage to Overcome Current Media Limitations and Facilitate Records Retention Compliance

- ✓ Guaranteed Content Authenticity
- ✓ Guaranteed Content Integrity
- ✓ Single Instance Storage
- ✓ Non-Rewriteable, Non-Erasable
- ✓ Record-level Retention, Protection and Disposition
- ✓ Shredded Deletes
- ✓ Seamless Content Migration: "Technology Proof"
- ✓ Record Level Auditability
- ✓ Faster Record Retrieval
- ✓ Self-Healing, Self Configuring
- ✓ Superior TCO



© 2005 EMC Corporation. All rights reserved.

Module Title - 1

Mercer Road
Global Education & Productivity

EMC's Centera is purpose built to overcome the the limitations of current media solutions while simultaneously exceeding the most stringent standards for records retention and preservation.

Here are some of the included benefits of Centera:

- Non-Rewriteable, Non-Erasable
- Record-level Retention, Protection and Disposition
- Assured Content Integrity and Availability
- Seamless Content Migration
- Record-level Auditability
- Faster Record Retrieval
- Superior TCO

Term #1: “Content Authenticity”

- A "hash output" (also called a "digest") is a kind of “fingerprint” for a variable length file of data; this output represents the contents of a file
- The digest can be used to verify if data is authentic or if it changed because of equipment failure or personal intent
- A good analogy we can think of is a “tamper proof seal for a software package”: if you open the file and change it, it's detected by the system
- Hashing algorithms are used in lots of places with the principal application being verification of authenticity
- Though this is a security feature in some systems, do not confuse this with “encryption”

A common use or application for hashing algorithms is in data networking. Network providers that are concerned with hackers getting into a network system often use optional hash codes with routing updates. Routing updates keep the network informed of changes in the topology due to congestion or equipment failure. Hackers often try to disrupt this normal operation by spoofing these updates and causing havoc in the routing system. To combat this many routing protocols support various hash functions that can be used as part of the update to verify that the incoming routing updates are properly generated by routers on the network and are not bogus updates created by hackers.

Example

- A file containing a single letter “a” generates an hashing algorithm of:
`0cc175b9c0f1b6a831c399e269772661`
- A file containing the letters “abc” generates:
`900150983cd24fb0d6963f7d28e17f72`
- In the examples above, it is possible to see how even a small change produces a radically different hashing algorithm
- Every unique bit sequence, in theory, will produce a *UNIVERSALLY UNIQUE* hashing algorithm signature
- It is this universal uniqueness that allows CAS to guarantee the authenticity and integrity of content.
- In fact, almost every feature of CAS is built on this foundation...

This slide shows the hashes generated for two files. The hashes are quite different. The fact that the hash in no way indicates the contents of the files contain adds a layer of security to Centera. In normal files systems, the file names often divulge what the file is about, however, a hash-based content address appears to be a random string of characters thereby obscuring the nature of the content from any observers.

Content Authenticity



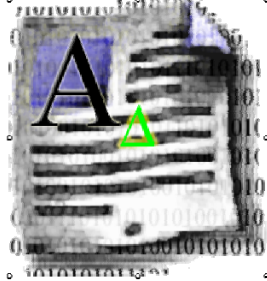
- Document is scanned using a hashing algorithm
- Universally unique document signature is created (hashing algorithm "Fingerprint")
- 128 bit binary address is hashed down to a variable length alphanumeric address. This "Content Address" is now the only identifier used for subsequent retrieval attempts

Content authenticity is achieved first with the generation of a CA, then the automated process of continuous checking and recalculation.

Content Authenticity (cont.)



ALGTLDNU3EDDS641VNISUS07DAK



- The only way to ask for an object from Centera is by its unique Content Address (or CA)
- If even 1 bit is changed, a new CA would result. **The unique signature of the content is also the “handle” it is accessed by.**
- It is impossible to “forge” or alter content as the only way to request the object is by its mathematically calculated identifier.

Though the slide is labeled as “content authenticity”, it also touches on the inherent security of the centra.

Term # 2: Content Integrity



- Every time an object is read, the Centera calculates the hashing algorithm of the object requested and compares the results.
- If an object fails to match its expected signature, it is rebuilt from its mirrored copy (which is also checked).
- This process also happens in the background, systematically testing every object in the cluster. Over time, every object is checked and rechecked continuously, guaranteeing content integrity even in the face of hardware failure, random error, or malicious intent.

Term #3: “Single Instance Storage”

- The universally unique hashing algorithm signature is also used to guarantee the storage of only a single instance of a blob object
- At write time, the Centera is polled to see if it already has an object with the same signature.
- If the object is already on the Centera, it is not transferred, and only a pointer to it is created (called a Content Descriptor File, C-Clip descriptor file, or CDF)
- Single instancing only refers to blobs, CDFs are never single instanced

Archived data is the fastest growing segment of the storage market. Whether it is external compliancy regulations or internal governance the challenge to store, maintain and manage this information becomes more and more challenging. With this in mind, any mechanism we can put in place to help us in any of these areas is a huge benefit, specifically, a mechanism that can help efficiency by ensuring we don't store multiple copies of the same document. Even if this only accounts for a small percentage of the users storage, it is that amount that will not have to be purchased, managed, powered, cooled etc. Over time this can account for significant savings!

It should be noted however that a variety of hashing mechanisms can be configured in the Centera, some of which do not offer single instance storage. The specific requirements need to be discussed between the appropriate members of the pre-sales team and the CIO or technology decision maker in the client environment. The various options will be reviewed later in this course.



Single Instance Storage Points

- Single instance storage is especially useful in environments with highly duplicated data (e-mail archiving for instance)
- Even 1 bit in difference will defeat single instance storage. If the time stamp at the top of 3 reports is the only difference, that will generate 3 content addresses
- In reality single-instance storage is best realized for items which were digital to begin with. Signal-to-noise ratios involved in scanning and sampling procedures mean that identical bitwise sequences are rare
- One large brokerage house realized almost a 20% savings in storage costs in their e-mail archiving environment



Term #4: Non Rewritable, Non Erasable

Term #5: Record-level Retention, Protection and Disposition

There are 3 versions of Centera:

- Basic Edition
- Governance Edition
- Compliance Plus Edition (aka CE+)
- Each version is purpose built to address specific, compliance and corporate governance driven needs
- Objects are protected by setting retention on an object by object basis or by associating the object to a “class” and assigning a retention to the class
- Additional metadata attributes provide file property information

Additional metadata is a function of the application and can also be added to the CDK as “custom metadata”. Metadata and metadata strategy are important as indexing tools use metadata to help with capacity planning and usage based billing. Metadata strategy is especially important with internally developed applications as including appropriate metadata entries from the beginning may add value to the archive at a future point in time by allowing more effective identification of specific useful content.



Basic Edition Centera

- Has no compliance edition enhancements
- Has support dial-in enabled
- Remote management is enabled
- Objects can be deleted anytime as retention is not enforced

It is important to mention that even with the non-compliant version of Centera, it is possible to protect objects from deletion. This can be accomplished by disabling the delete capability in the application profiles.

Governance Edition Centera

- Has Compliance Enhancements
 - Unless specifically set by the administrator or application, objects take on a default retention period of 0 (no specific retention period)
 - Objects which have a retention period which has not expired, cannot be deleted* by any means
- Phone in support is enabled
- Remote management is enabled
- Objects can be deleted as long as their retention period is 0*

*Privileged Delete is a feature supported from CentraStar Version 2.3 that allows the system administrator to enable an option that allows the API to delete objects before the retention period has expired

*Note that retention is expressed in seconds and is added to the write timestamp in the CDF and matched against the cluster time (UTC) to determine whether or not the clip can be deleted.

If privileged Delete is going to be enabled then the application should be set-up to make appropriate records showing why, when and by whom, the record was deleted. The Reflections feature of the Centera, which tracks deletes, could be used to record this information.



Compliance Edition Plus Centera

- Has Compliance Enhancements
 - *Unless specifically set by the application*, all objects have a default retention period of infinity (-1)
 - The default retention period can not be changed by the administrator
 - Objects which have a retention period which has not expired, cannot be deleted by any means
- Dial in support is DISABLED*
- Remote management is DISABLED
- Objects cannot be deleted until their retention period is 0

Even though the modem is enabled, it should be tested then physically disconnected in the case that EMC support needs to access the cluster remotely.

Some Retention Related Facts

- Retention period, as well as all Centera Meta Info (including the hashing algorithm of the binary object [or BLOB]) is kept in an XML file called a “Content Descriptor File”
- With single instance storage, multiple CDF’s can reference a single blob. Each CDF can have a DIFFERENT retention period as well
- When you remove an object from Centera, you are really only removing your CDF to it... The blob itself is removed only after all CDF’s which reference it have been deleted and no more references exist
- The actual removal of the blob is handled automatically by Centera’s garbage collection feature.
- **IMPORTANT:** CDF removal must be initiated *by the application*. A CDF will remain, even if it’s retention period has expired, until it is asked to be deleted by an API request. This is to ensure that “accidents” don’t happen with your data

The last bullet is very important. Blobs and CDFs do not go away if retention expires. This only means that the cdf/blob is a “candidate” for deletion. Thus a procedure needs to be in place within the application if you wish to delete data once the retention period has expired.



C-Clip Browser

- Write File
- Set Retention
- Read File
- Attempt Delete
- Write Duplicate
- Edit / Change Original
- Display CDF
- Delete Original

© 2005 EMC Corporation. All rights reserved.

Module Title - 14  Mercer Road
Global Education & Productivity

“C-Clip Browser” is a tool that demonstrates the concepts of single instance as well as retention. Displaying the CDF also gives you a first hand look at what Meta data is and what a real CDF might look like.

Term #6: Shredded Deletes (Department of Defense 5015)



© 2005 EMC Corporation. All rights reserved.

Module Title - 15  Mercer Road
Global Education & Productivity

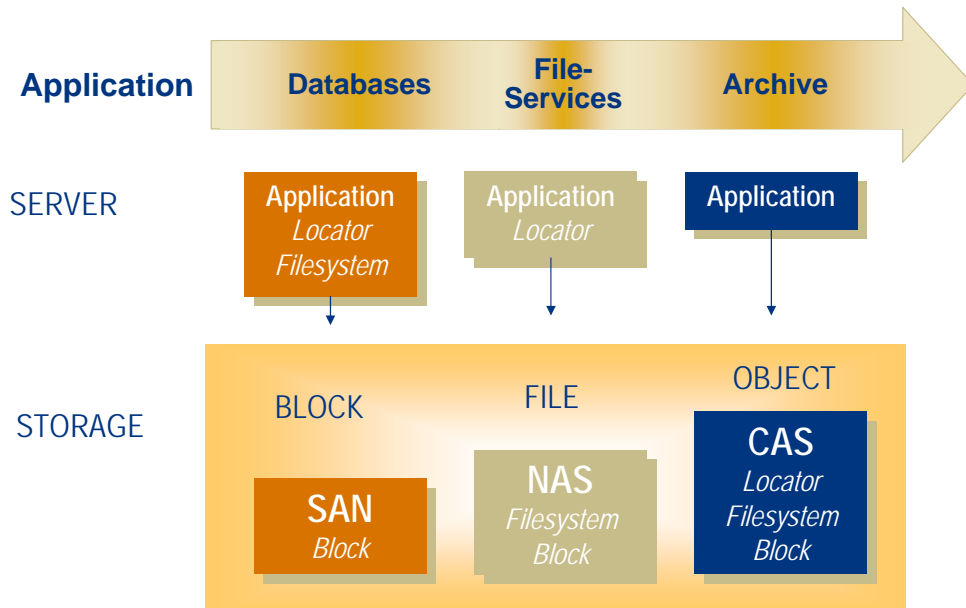
- DOD requirements for destruction of digital assets require deletion and 7 pass overwrites of file data with 0's and 1's.

Centera can be configured to execute a “shredded” delete in compliance with DOD 5015.

Deletion of blobs is handled by garbage collection so this activity occurs in the background.

Deletion of CDF's occurs upon request, but CDF's are typically very small, and therefore performance impact is trivial.

Data shredding is disabled by default on all Centera models. Only an EMC field engineer can enable data shredding on CE and CE+ models. On a Basic model, data shredding cannot be enabled.

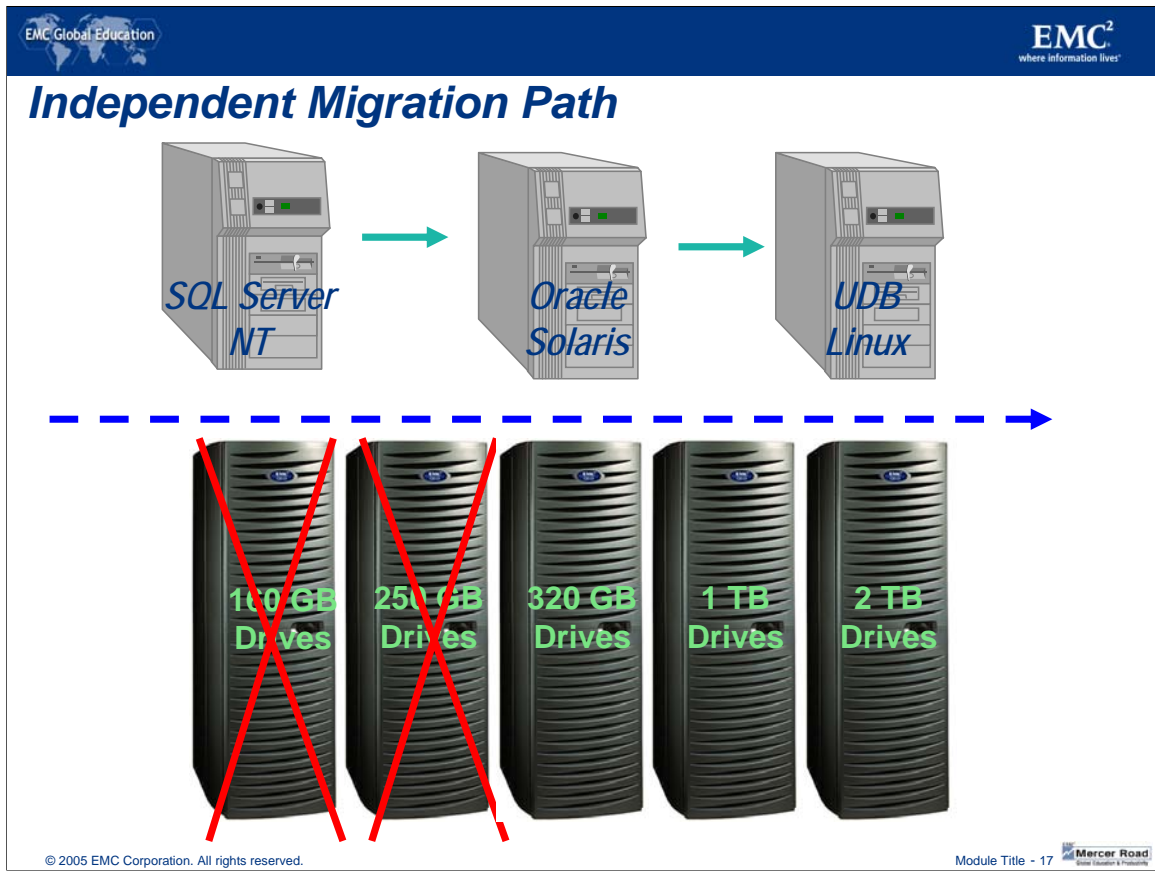
Term #7: Seamless Content Migration: “Technology Proof”

© 2005 EMC Corporation. All rights reserved.

Module Title - 16  Mercer Road
Global Education & Productivity

SANs, NAS and CAS are all purpose built solutions. SANs for OLTP and Data Warehousing, NAS for file sharing, and now CAS for fixedcontent applications.

Much the same way that NAS embedded the filesystem into the storage architecture and removed it from the application server, CAS removes the file locator from the application server and embeds it in the storage. This is the key to CAS. Information becomes portable without regard to platform or application. Future migration becomes simpler and information is vastly easier to access.



One of the most frustrating aspects of storing large archives is the inability to upgrade the storage medium without doing a wholesale data migration. This is because the storage medium and the logic controlling it are inextricably intertwined. If you decide to go to new tape drives or the manufacturer changes formats you must migrate everything from the old storage medium to the new. This requires a new filesystem with new locators and a new database to manage them.

With Centera, this is no longer the case. Because C-Clips are independent from the actual storage architecture and do not require a file-system or LUN architecture, it doesn't matter what kind of platform, database, or application they exist on.

As we look forward on the bottom, we have a pretty good idea that a year from now we will have faster processors, higher capacity drives, and more throughput on our networks. In five years we will have even more still. Yet these systems not only will be compatible with each other, they are totally independent from what type of database, platform, or application is used to manage the C-Clips. So, if six months or two years or five years from now a customer decides to go to a new database or platform or application it doesn't require any change on the Centera.

This is an independent migration path. Even though it is beyond our comprehension to imagine what types of database and storage technology will exist 20 or 50 years from now we know that



Term #8: Record-level Auditability

- Store record information in your database and in your CDF as well. CDF's can store application specific data
- Comparisons between the database and the guaranteed authentic CDF can highlight inconsistencies, either from bugs or malice
- CDF's can also be used to reconstruct data in the event of a disaster

This slide highlights the importance of metadata. EMC is of the opinion that more is better. When it comes time to organize or catalog the information stored on Centera. A sound metadata strategy ensures proper usage, bill-back or general organization requirements can be met if implemented BEFORE production data is written to Centera.

Inside a CDF

```

<?xml version='1.0' encoding='UTF-8' standalone='no'?>
<ecml version="3.0">
<eclipdescription>
<meta name="type" value="Standard"/>
<meta name="name" value="myClipName"/>
<meta name="creation.date" value="2004.06.18 15:03:31 GMT"/>
<meta name="modification.date" value="2004.06.18 15:04:05 GMT"/>
<meta name="creation.profile" value="testprof"/>
<meta name="modification.profile" value="testprof"/>
<meta name="totalsize" value="108"/>
<meta name="refid" value="2LO0S2FH10FR3FOJFVKD2GTOUO"/>
<meta name="prev.clip" value=""/>
<meta name="clip.naming.scheme" value="md5"/>
<meta name="numtags" value="1"/>
<meta name="retention.period" value="300"/>
<meta name="sdk.version" value="2.3.296"/>
<eclipcontents>
<myTagName>
<eclipblob hashing algorithm="4L0RJ0AVP6PCQx3QGOB5O7U5QK9"
size="108" offset="0"/>
</myTagName>
</eclipcontents>
</ecml>

```

This is an example of what a CDF *might* look like. Content of the CDF can vary depending on the application and whether custom metadata is being added in addition to the application metadata. For simplicity, it is easy to describe the CDF as being similar to the “properties” of any windows file.



Term #9: Faster Record Retrieval

- Spinning disk provides Sub-second “Time to First Byte” (200ms-400ms) in a single cluster
- No file system limitations on numbers of objects or sub-trees Content is accessed directly by address
- Random access is almost always faster than serial access
- All online spinning disk eliminates the hassle and delay of robotic jukebox type devices

Solid state memory access is the fastest... accessing DRAM for example is faster than the access time to any disk drive. Both are “random access” devices but the disk drive is mechanical hence slower than a purely electronic device.

Disk drives on the other hand are the fastest of the mechanical solutions. Tape is sequential access and CD ROMs have inherently slow transports.

Term # 10: Self Healing Self Configuring

- Self healing and self configuring features of Centera are implemented in the platform
- The platform refers to both hardware and software
- The hardware used in the platform is off the shelf commodity based hardware configured to EMC specifications
- The hardware is configured and implemented in a way that supports the overall objective of the platform which is to reliably secure customer archive data delivering superior performance at a competitive acquisition cost and much better long term ROI
- The product in many ways has the ability to manage itself

Some examples of self management:

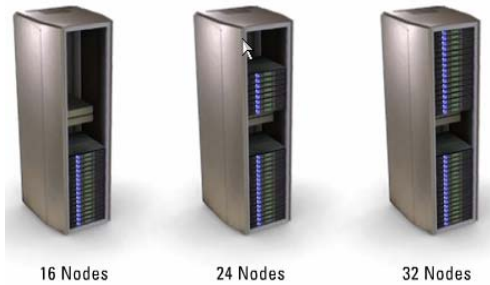
CentraStar can determine when a node has gone off line and compensate for the loss

CentraStar can determine the best place in the cluster to store incoming data at any given time

CentraStar can determine when it needs service and call home

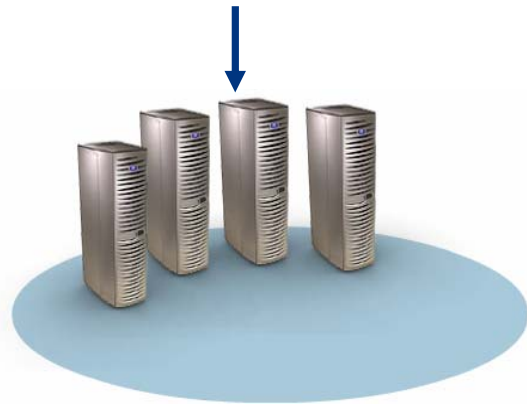
When adding new nodes, the system automatically will initiate an upgrade/downgrade to the right software level and automatically incorporate the new capacity in the cluster.

Rack / Cluster Configurations



Multiple Nodes in
a Cabinet

One or Multiple
Cabinets in a
Cluster



© 2005 EMC Corporation. All rights reserved.

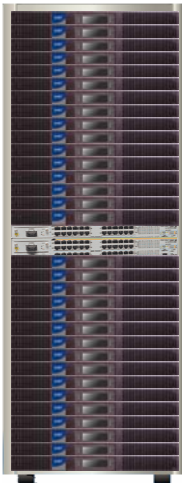
Module Title - 2

Mercer Road
Global Education & Productivity

A Centra hardware configuration can consist of a single rack/cabinet or multiple racks/cabinets. There are multiple nodes within each of the cabinets. When cabinets are connected, nodes become aware of each other. The nodes within a single cabinet, or connected cabinets, are referred to as a cluster. Clustered nodes are automatically aware of nodes that attach to, and detach from, the cluster.

Single Centera Rack / Cabinet / Cube Components

Redundant Array of Independent Nodes



- ***One rack contains 4, 8, 16, 24 or 32 nodes***
- ***Access nodes provide application server connectivity***
- ***Storage nodes store, protect, and validate information***
- ***Automatic node, drive & network configuration***
- ***Clustered operation with load balancing and self healing***
- ***Architected to EMC reliability, availability and serviceability standards***
- ***Two internal LAN switches***



Self Healing, Self Configuring Cont.

- ***Gigabit uplink modules for inter-cabinet connection (optional for Gen1,2&3)***
- ***Gigabit cluster access is available on Gen3,4 Hardware***
- ***Two external modems for dial backup***
- ***Two intelligent power distribution units are provided and should be connected to different sources at the site***
- ***Redundant component architecture in conjunction with the CentraStar redundant OS features ensure that Centera maintains continual data availability even when single components fail***

Node Characteristics at a Glance



- Customized Linux OS
 - Redhat prior to V3 CentraStar SuSE in V3
- Reiser FS
- Pentium CPU
 - >2 Ghz since V2 hardware
- 4 multi GB drives
- System Memory
 - 512 MB RAM for V3 hardware, 1 GB as of V4

© 2005 EMC Corporation. All rights reserved.

Module Title - 5

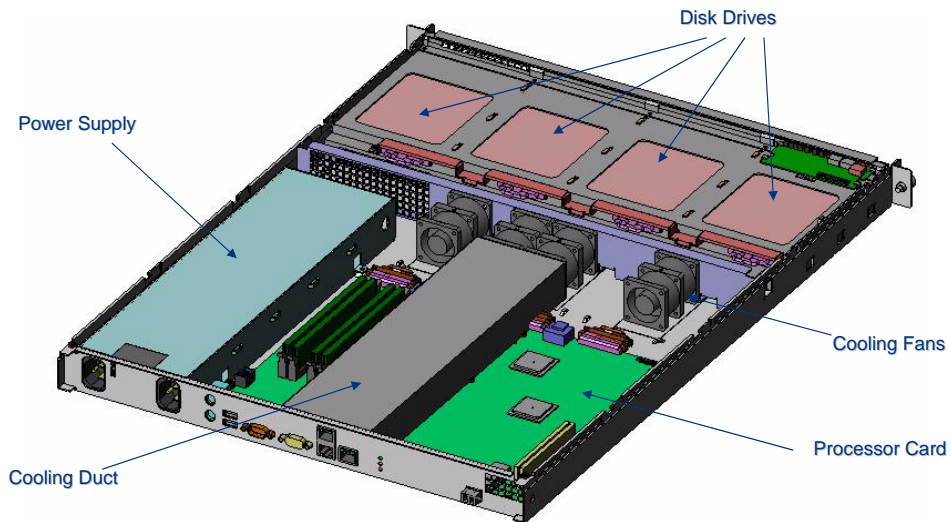


OS and FS are transparent to the outside world

Complete copy of Centera software on every node

Software installed on all nodes is identical – only the configured roles are different.

Internal Physical Layout



© 2005 EMC Corporation. All rights reserved.

Module Title - 6

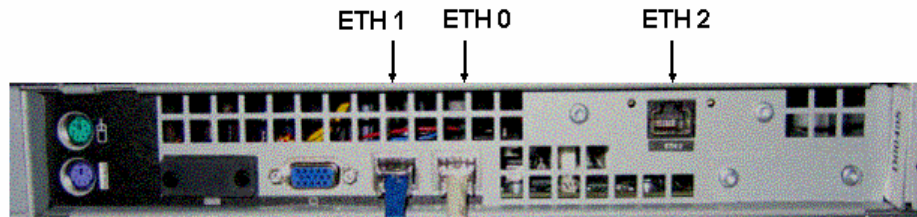
Mercer Road
Global Education & Productivity

This is the general layout of a node. There are minor differences from one version to the next but generally there are four disks mounted at the front of the unit, an off the shelf processor card and power supply assembly.

where information lives

EMC²
where information lives

Rear View of Node

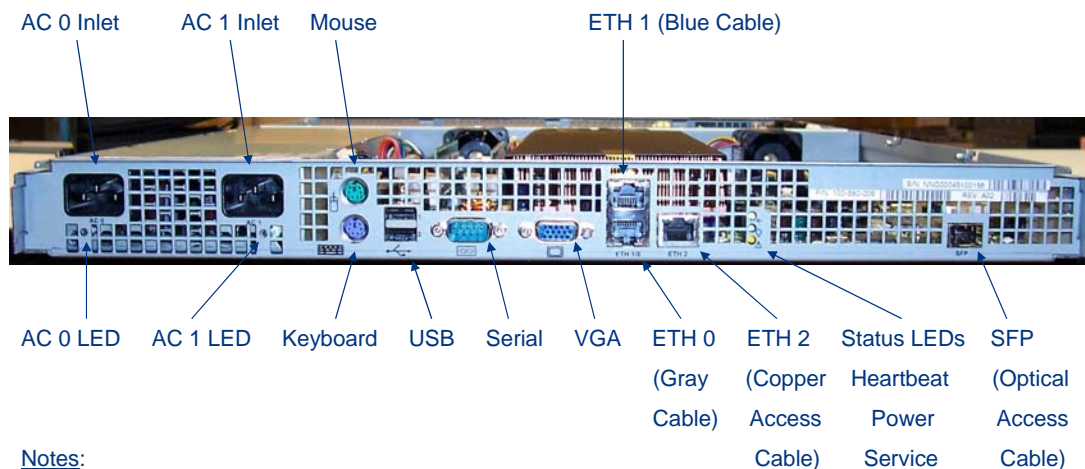


EMC Global Education
© 2004 EMC Corporation. All rights reserved.

EMC
Mercer Road
Global Education & Productivity

V4 Node – Physical Design

Mechanical Review Rear I/O and LEDs



Notes:

1. Notice that the sides of the nodes are notched in each corner so that the C-shaped mounting rails can be used.

SFP (Small Form-factor Pluggable) interface (Optical)

Version 2 vs. Version 3 - Node Comparison

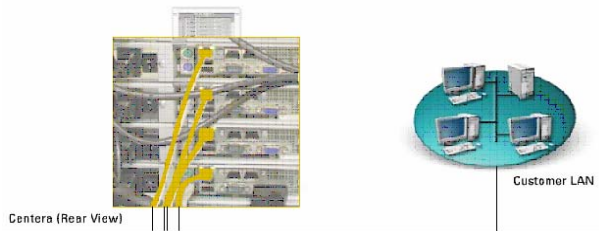
	<u>Version 2 Hardware</u>	<u>Version 3 Hardware</u>
Mainboard	Tyan S2420 (Pentium 3)	Tyan S2098 (Pentium 4)
Chipset	Intel 810e AGPset	Intel 845GL chipset
Front side bus	133mhz	400mhz
CPU	PIII Coppermine 1Ghz	P4 2.0A (2Ghz/.13u)
Disk drives	(4) Maxtor 250GB	(4) Maxtor 320GB
Memory	(2) SDRAM for 512mb total	(2) DDR for 512mb total
Power	ATX, 20-pin power connector	ATX 20-pin + 4-pin for fan
Bios	AMI BIOS 4Mbit flash	Award Bios 4Mbit Flash
Lan (onboard)	(1) 10/100 RJ45 port	(2) 10/100 RJ45 port
Lan (in riser)	(2) 10/100 RJ45 port	(1) Intel GigE port.
IDE Controller	Promise ATA133 Controller	Onboard ATA 100 Intel
Graphics	Intel Graphics	Intel Extreme Graphics
CPU Cooling	Heat Pipe	Redundant Fan Sink

V4 versus V3 Node Comparison

	CURRENT Configuration	NEW Configuration
	<u>Version 3 Hardware</u>	<u>Version 4 Hardware</u>
Mainboard	Tyan S2098 (Pentium 4)	Tyan S5158 (Pentium 4)
Chipset	Intel 845GL chipset	Intel E7221 Northbridge & ICH6-R Southbridge
Front side bus	400mhz	800mhz, parity protected
CPU	P4 2.0A (2Ghz/.13u)	2.8 GHz P4 w/1MB cache
Disk drives	(4) 5400 RPM Maxtor 320GB	(4) Maxtor 7200 RPM SATA-II 320GB
Memory	(2) DDR for 512mb total	(2) 512 MB DDR2-533 ECC DIMMs -1GB total
Power	ATX 20-pin + 4-pin for fan	ATX 300W Supply with integrated AC Transfer
Bios	Award Bios 4Mbit Flash	Award Bios 8Mbit Flash
Lan (onboard)	(2) 10/100 RJ45 port	2 Dual Intel 82546 GbE, 3 RJ45, 1 SFP shell
Lan (in riser)	(1) Intel GigE port.	No Risers
IDE Controller	Onboard ATA 100 Intel	4 Integrated SATA-II channels in ICH6-R
Server Mgmt	NEI custom HW & SW	Integrated Qlogic BMC & Standard IPMI SW

Cabling Nodes

Multiple
Access
Nodes



NETWORK
CABLES CONNECT
THE SWITCHES...



...TO THE
CORRESPONDING
NODES

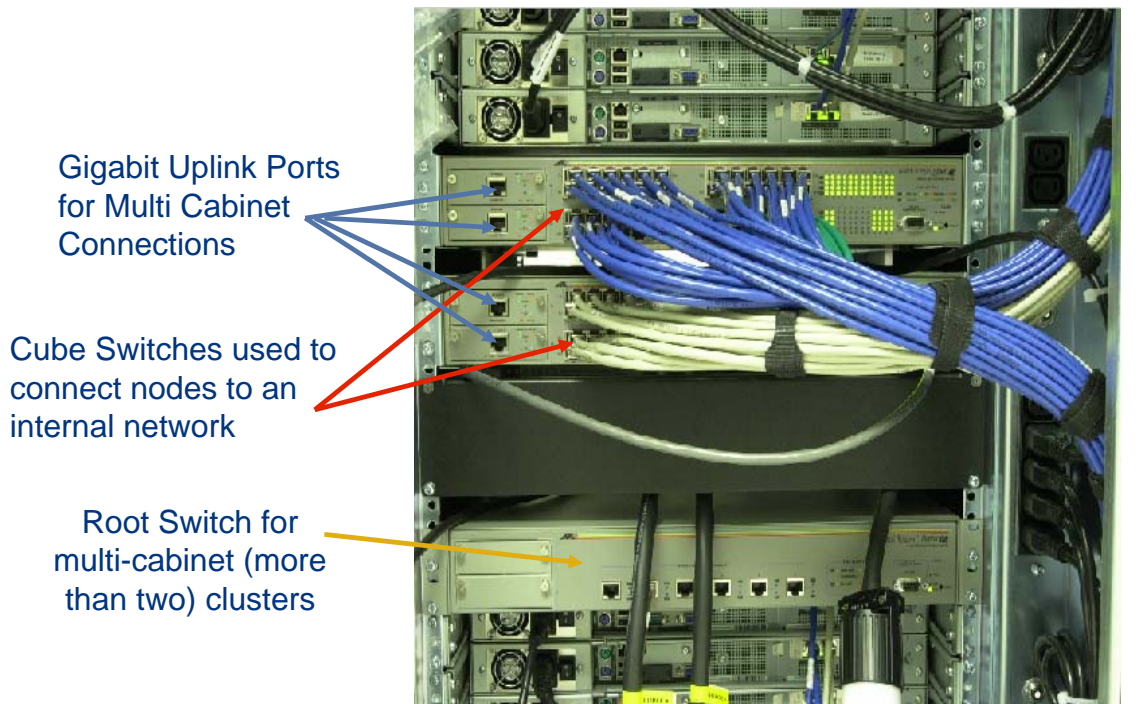


© 2005 EMC Corporation. All rights reserved.

Module Title - 11  Mercer Road
Global Education & Productivity

All nodes have three Ethernet ports referred to as ETH0, ETH1, and ETH2. Both Access nodes and Storage nodes use ETH0 and ETH1 for communication between the nodes over an internal network. Only the Access nodes use the ETH2 ports for communication with the customer's Application Servers. The throughput needs of the application will determine how many Access nodes are to be configured at the time of installation. Each Access node is connected to the application server infrastructure via 100 Megabit or GigE for V3 node connections.

Gen1,2,3 Switches for Internal Network



© 2005 EMC Corporation. All rights reserved.

Module Title - 12 Mercer Road
Global Education & Productivity

Gigabit Uplink Ports must be installed in the Cube Switches if a multi-cabinet cluster is planned. If the cluster is to go beyond two cabinets, then Root Switches must also be installed. These Root Switches would be installed in cabinets 1 and 3.

Note: If the installation was done at revision 1.x of the Centera hardware, both Root Switches would be installed in cabinet 1.

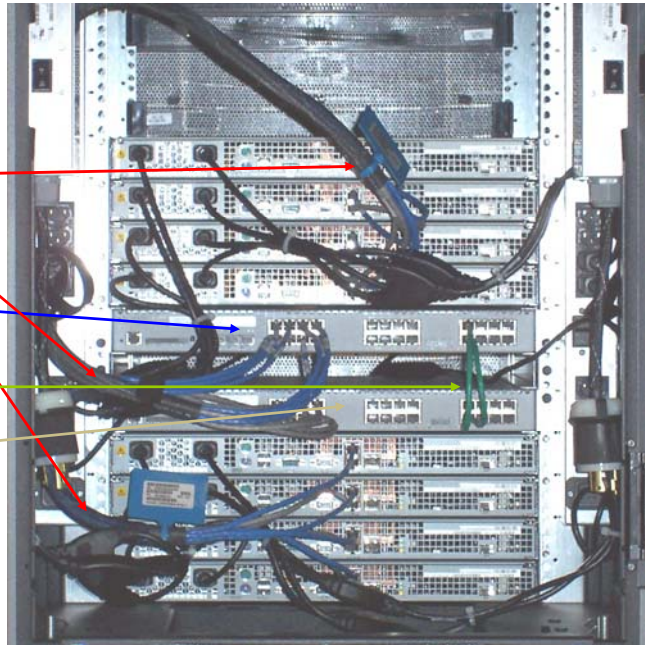
V4 NCS – for Internal / External Network

Wrapped cable (Red)
bundles routed with
“Dog Tag” attached

Cube Switch 1 (Blue)

Trunk Cables (Green)
(Cross-over cables)

Cube Switch 0 (Gray)

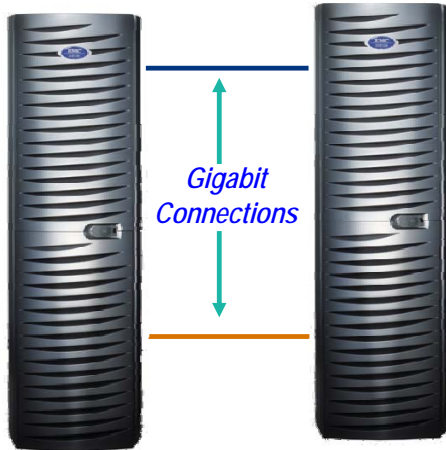


© 2005 EMC Corporation. All rights reserved.

Module Title - 13  Mercer Road
Global Education & Productivity

NCS (Network Cube Switch)

What the 2 racks look like...



- Limit of two racks "daisy chained together"
- >30 TB usable
- Any Access node can talk to any Storage node

The connection of two racks is relatively simple to perform. Gigabit uplink modules must be purchased first to be installed into the cube switches. These are then connected together to form a single cluster from two cabinets.

What more than 2 racks look like...



- >60 TB usable
- Any Access node can talk to any Storage node
- Connections are between root / NC switches

© 2005 EMC Corporation. All rights reserved.

Module Title - 15

Mercer Road
Global Education & Productivity

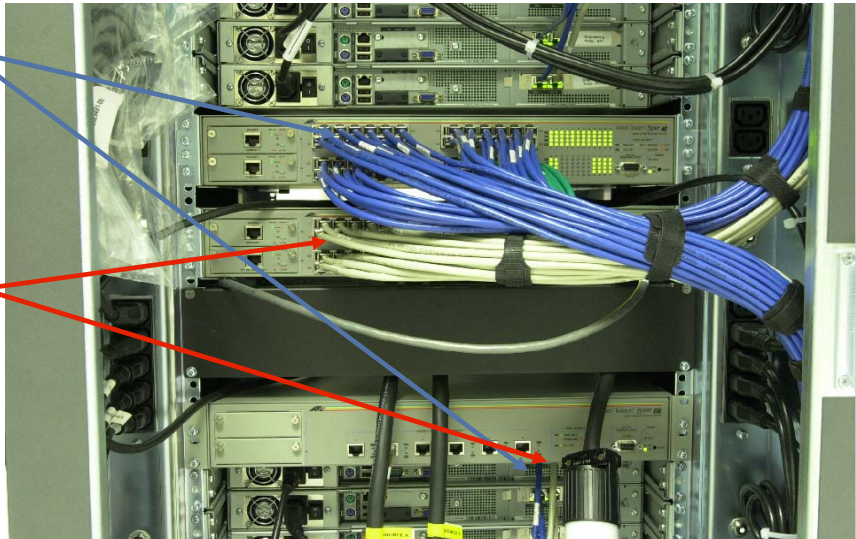
For a four or more rack cluster, there is a requirement to purchase root switches. These switches provide the connectivity to all of the cabinets in the cluster.

NCS (network cube switch)

Node Connections

Cube1 Cables
connected to
ETH1 Ports

Cube 0 Cables
connected to
ETH0 ports



Internal IP network assigned based on Cube Switch Port Connections

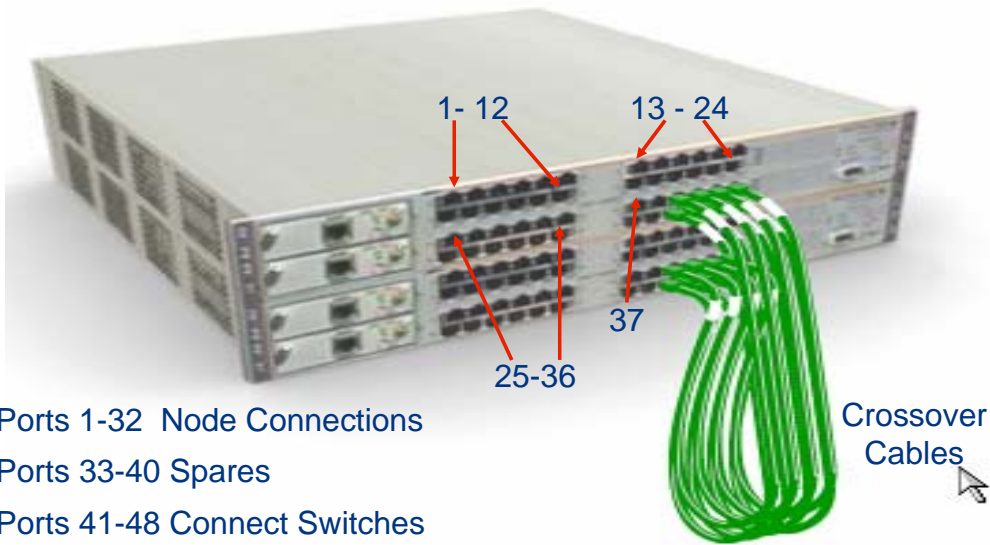
© 2005 EMC Corporation. All rights reserved.

Module Title - 16 Mercer Road
Global Education & Productivity

Node Connections

IP addresses are assigned to the nodes based on the Cube Switch port from which the cable is run. Nodes are cabled from bottom to top. Therefore, if the base address (chosen at Kickstart) for the Centera is 10.255.0.0, the bottom node in the first cabinet will be addressed 10.255.1.1. The third octet signifies cabinet 1. In another example, using a multi-cabinet cluster, the 13th node in cabinet 2 would be addressed 10.255.2.13.

Rapier Switch V1,2,3



- Ports 1-32 Node Connections
- Ports 33-40 Spares
- Ports 41-48 Connect Switches
- If cable moved to spare port, name/address of moved node changes

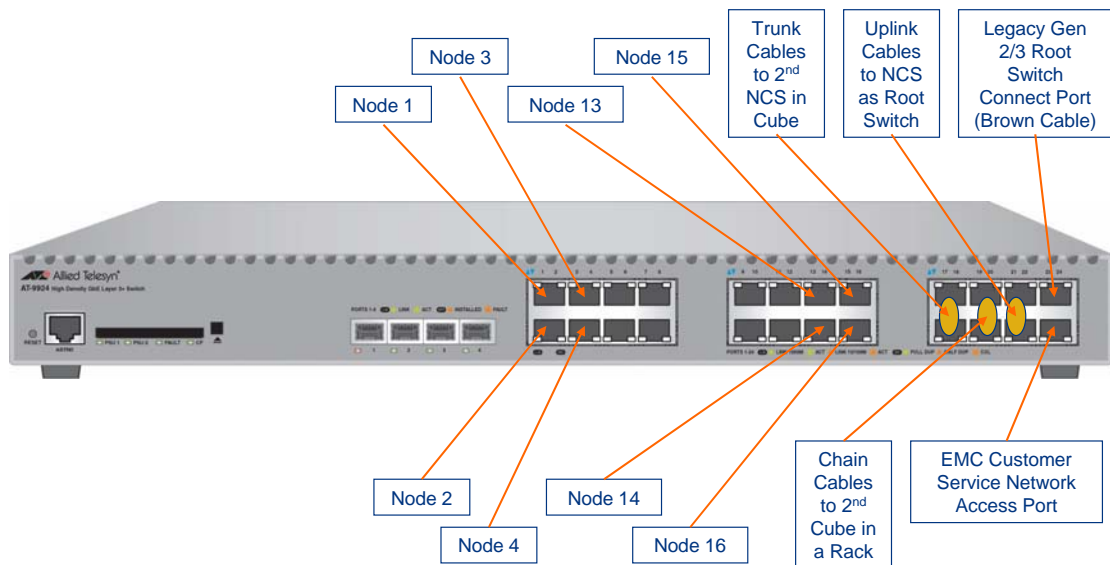
© 2005 EMC Corporation. All rights reserved.

Module Title - 17

Mercer Road
Global Education & Productivity

Ports 1-32 are used for node connections. Ports 33-40 are spares. Ports 41-48 use cross-over cables for communication between the two switches.

Network Cube Switch (NCS) V4




© 2005 EMC Corporation. All rights reserved.

Module Title - 18 

The new V4 24 port cube switch. Each port provides 10/100/1000 connectivity. The switch also uses redundant power inputs.

[illegible]

- Module Title - 19  Mercer Road
Global Education & Proficiency



EMC
Mercer Road
Global Education & Productivity

V1,2,3 Rapier 48(i) Cube Switch –vs- V4 NCS

Feature	Rapier 48(i) Cube Switch	9924T NCS Cube Switch
Port Count	48 10/100Mbit + 2 1000Mbit	24 10/100/1000 Mbit
Redundant PSU	No	Yes
Flash Ram	No	Yes
Serial Connection	Db9	RJ-45
Packaging	1.5 RU	1.0RU
OSPF	Yes	Yes
RIP	Yes	Yes
LACP	No	Yes
Jumbo Frames	No	Yes

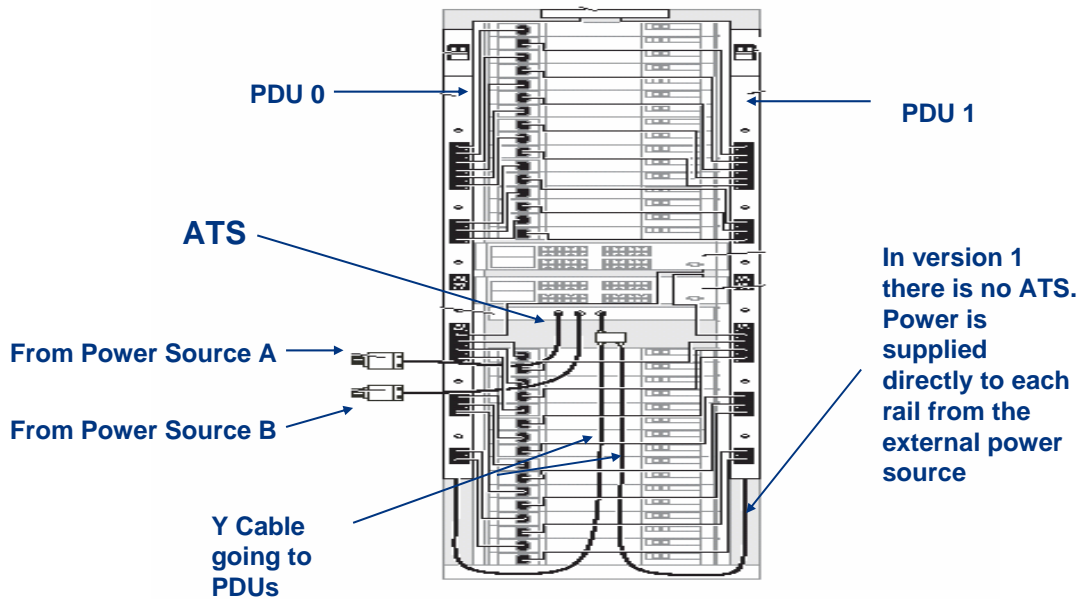
OSPF stands for: open shortest path first. It is a popular link state “interior” routing protocol and is supported by most router devices.

RIP stands for: routing information protocol. This is also a popular “interior” routing protocol but falls into the category of “distance vector”. Distance vector means that the distance or “hop count” is the only criteria used to determine the best path to the destination. By contrast a “link state” protocol such as OSPF considers other criteria such as delay, bandwidth and cost making it more sophisticated.

LACP stands for link aggregation control protocol (802.3ad). This protocol allows you to bundle several physical ports together to form a single logical channel.

Jumbo frame is the support of transferring frames above the standard Ethernet size of 1500 bytes, jumbo frame support can vary from router to router but typically tops out around 9000 bytes which allows support of an 8k payload and associated header.

V2 and V3 ATS



© 2005 EMC Corporation. All rights reserved.

Module Title - 21  Mercer Road
Global Education & Productivity

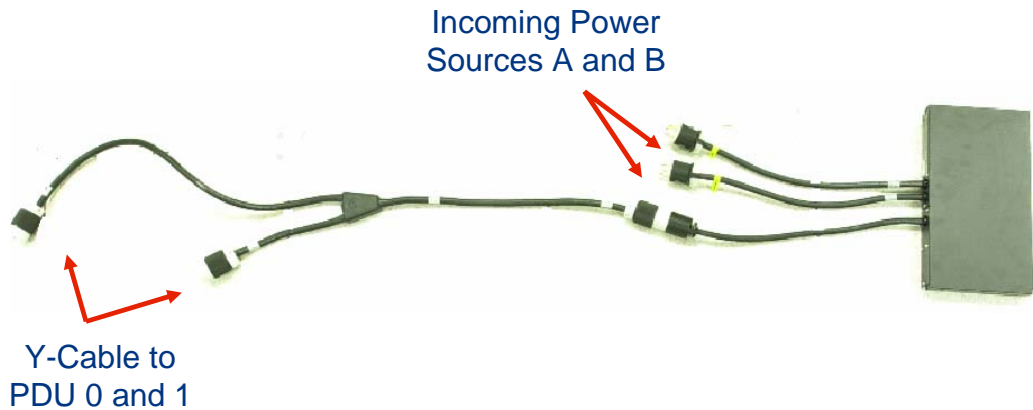
Two sources of power must be provided to the Centera. The ATS selects one of these sources to provide power to the entire rack. In the event of power loss from the first source, the ATS will switch over to the second source (20 milliseconds), thus preventing power loss to the Rack. The ATS thus provides power redundancy.

In version 1.x of Centera there was no ATS. There are two cables coming from the ATS switch which would be connected to power sources A and B. There is also a Y cable coming from the ATS switch and connected to PDU 0 and PDU 1.

Note: Currently, ATS switches cannot be retrofitted into version 1 cabinets.

Notice the location of the ATS in version 2. In version 1 this was the physical location of a root switch. Root switches are optional and used in a multi cabinet installation. If you have a multi cabinet installation that requires root switches, the one replaced by the ATS in cabinet 1 would now be placed in cabinet 3.

ATS Connections



© 2005 EMC Corporation. All rights reserved.

Module Title - 22  Mercer Road
Global Education & Productivity

ATS Connections

The two power sources made available to the Centera are referred to as power source A and power source B. Both power sources are connected to the ATS (A/C Transfer Switch). The output from the ATS is then connected to a Y cable. One output from the Y cable is then connected to PDU 0 while the other output from the Y cable is connected to PDU 1.

Indicators and Switches

DO NOT TOUCH
THESE SELECT
BUTTONS

A/C Source Indicators



Serial cable to Cube Switch 0.
Communicates its presence at
startup. Must be present for
ATS to communicate its
presence to the Cube Switch

© 2005 EMC Corporation. All rights reserved.

Module Title - 23 Mercer Road
Global Education & Productivity

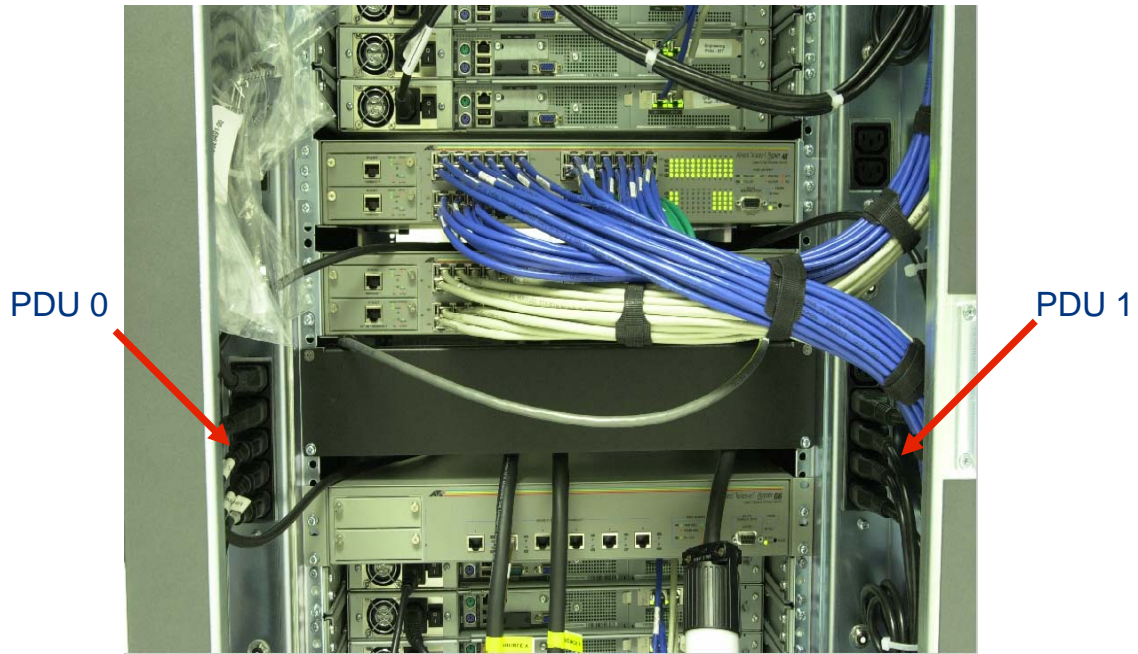
The A/C source indicator LEDs can be in 3 different states: Bright, Dim and Off.

If indicator A or B were bright, it would indicate that it was the power source being used. If it were Dim, it would indicate that the power is good but it is the backup power source. If it were off, it would indicate that there is no power (this is a problem).

Notice in this illustration that power source B LED is bright, while the power source A LED is off. This indicates that power source A has failed over, and power source B is being used. Notice that there is a serial cable from the ATS to the cube switch 0. This connection will not be utilized in version 1.2.SP1 of CentraStar. The configuration of the ATS switch is set when shipped from the factory, therefore the setting should not be changed.

DO NOT USE ANY OF THE BUTTONS WHICH COULD AFFECT THESE SETTINGS.

Power Rails



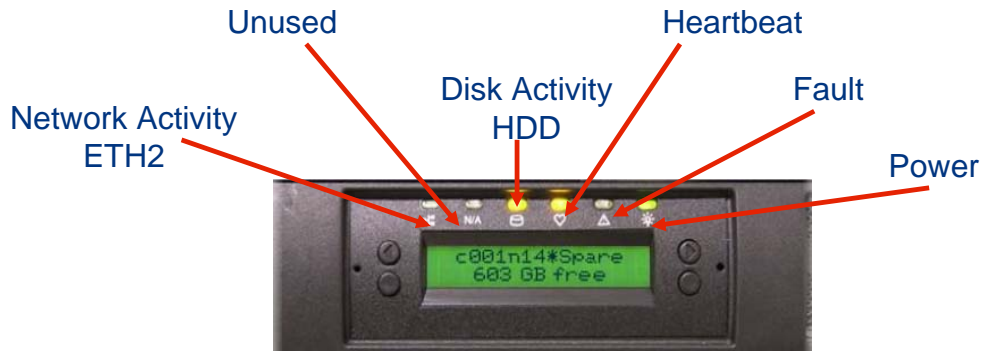
© 2005 EMC Corporation. All rights reserved.

Module Title - 24 Mercer Road
Global Education & Productivity

Power Rails

There are two power rails per cabinet which are used to power the nodes and switches. These power rails are also referred as PDU 0 and PDU 1. All even numbered nodes and Cube Switch 0 are plugged into PDU 0. All odd numbered nodes and Cube Switch 1 are plugged into PDU 1.

Front Panel & LEDs (Gen 1,2,3 nodes)



© 2005 EMC Corporation. All rights reserved.

Module Title - 25  Mercer Road
Global Education & Productivity

Front Panel LEDs

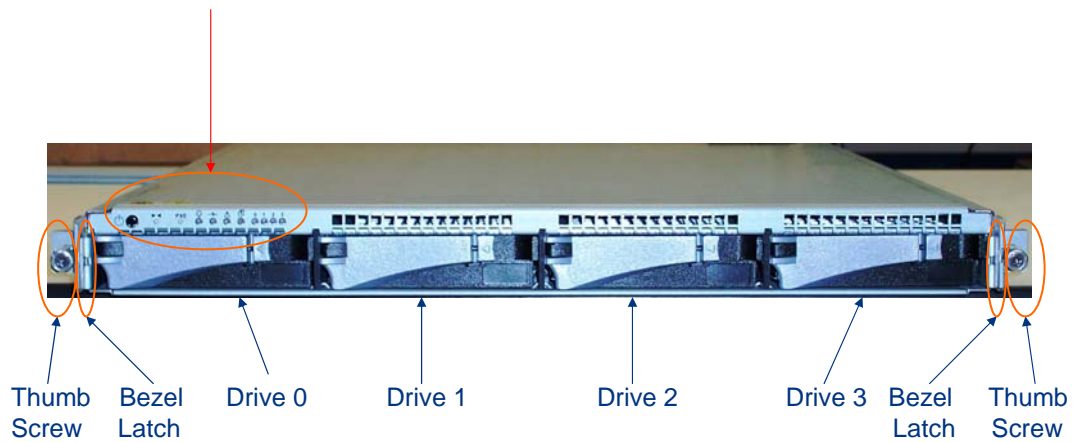
The front panel LEDs have the following functions:

- ETH2 - blinks when there is network activity, if the node is an Access Node.
- Unused
- HDD – blinks when there is disk activity on any disk in the node.
- Heartbeat – Should be blinking. LED staying on without blinking indicates a problem. If problem is detected, the node may reboot automatically in 2 minutes.
- Fault – If this LED is on, the software has detected a problem.
- Power – indicates that both AC and DC power are within specifications.

V4 Node – Physical Design

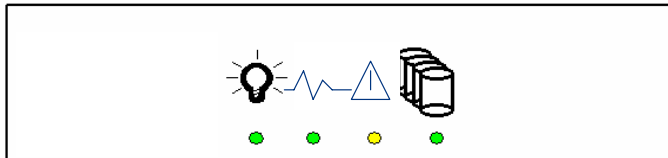
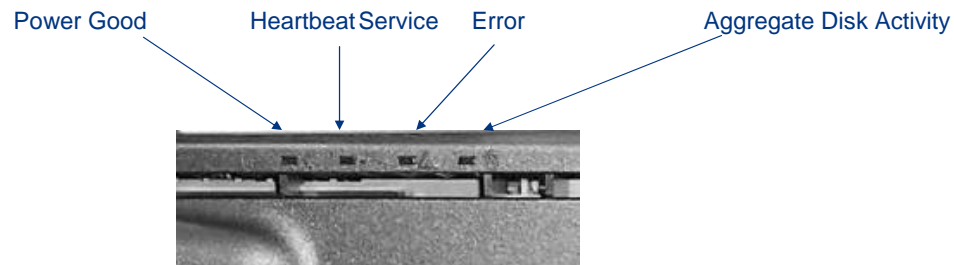
Front View with Bezel removed

Front Panel with additional buttons and LEDs visible

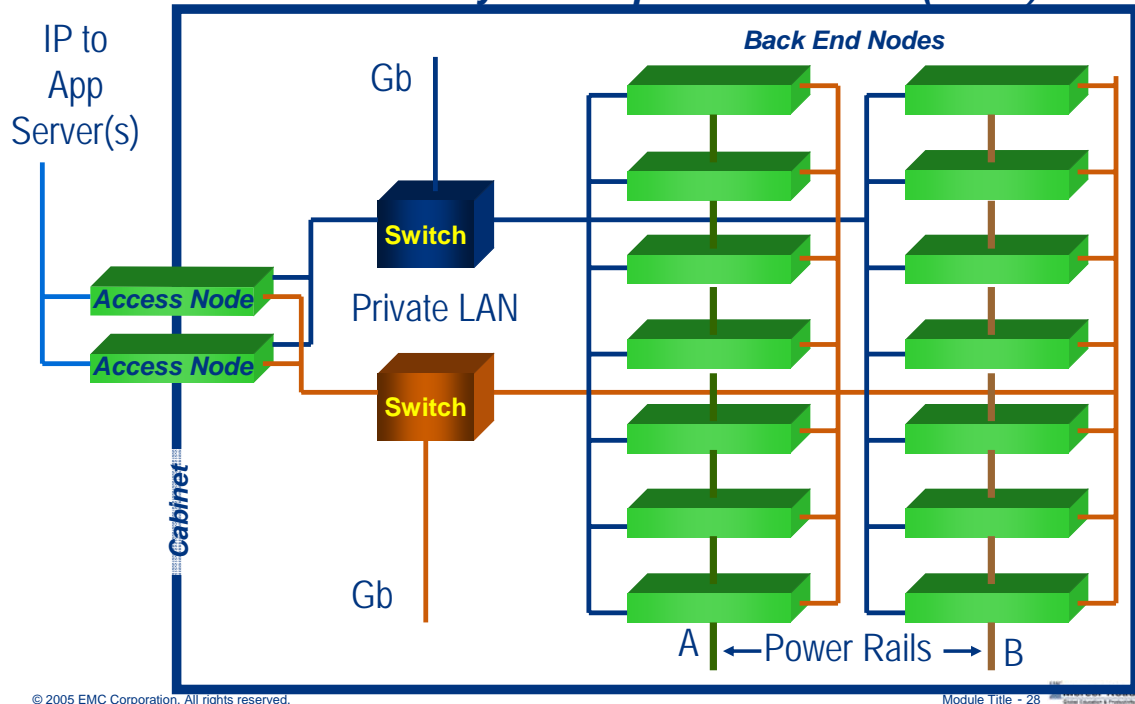


V4 Node – Physical Design

4 Node Status LEDs visible through the Bezel

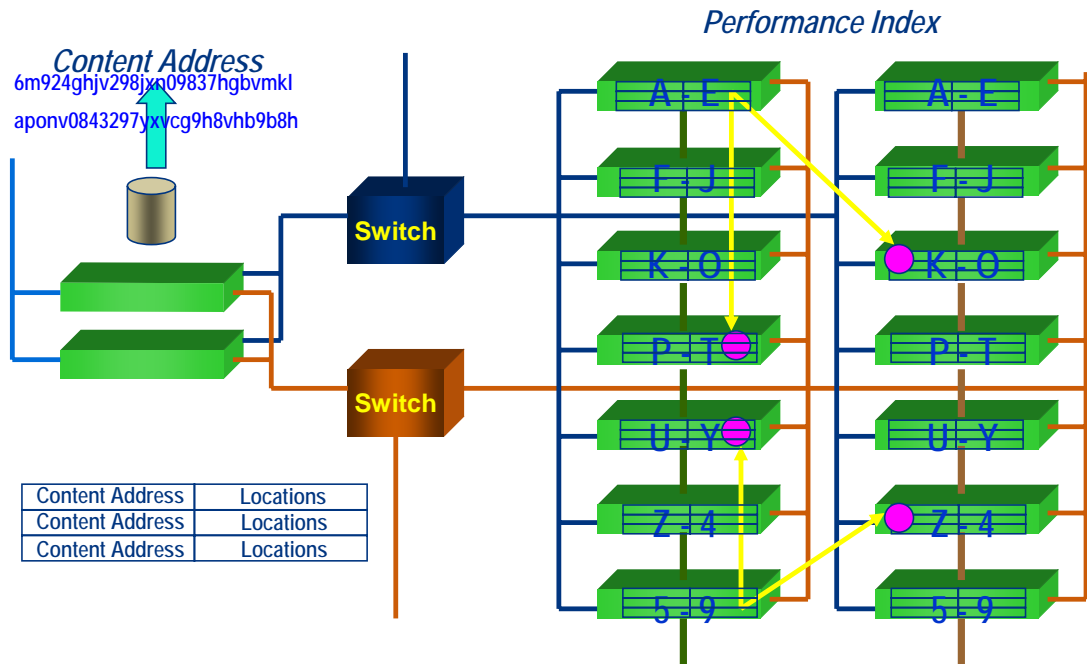


Centera's Architecture: Built on Redundant Array of Independent Nodes (RAIN)



This is what a Centera looks like internally. All the nodes are identical and upon power up configure themselves within a few minutes. The only thing the administrator does is determine how many nodes will be front-end nodes and how many will be back-end nodes. Front-end nodes give better performance and back-end nodes give more storage capacity.

Basic Sequence of Operation



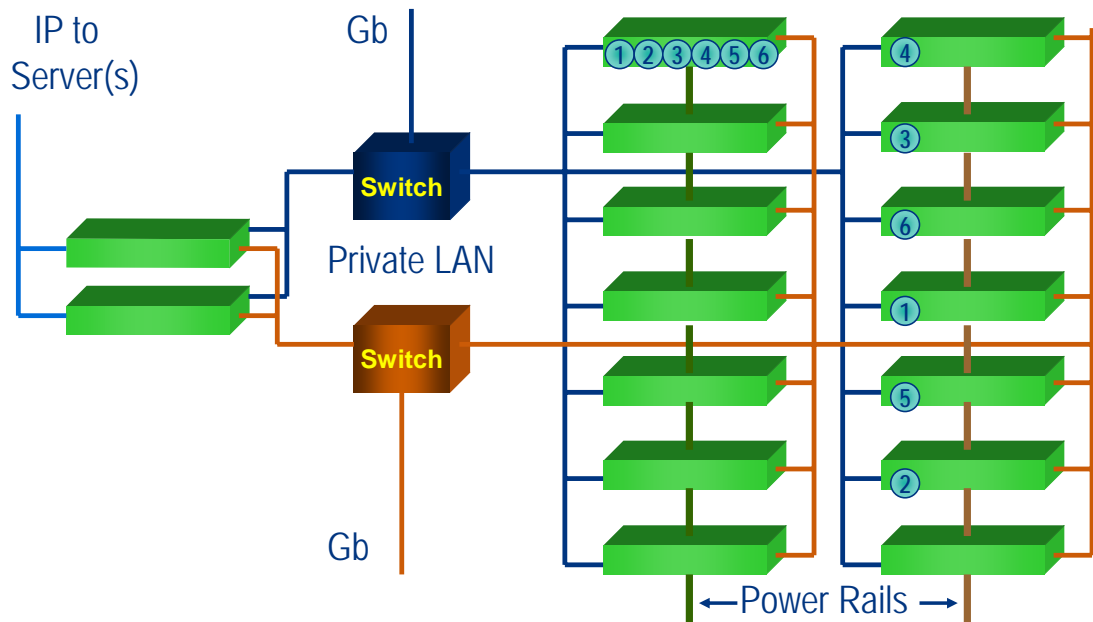
© 2005 EMC Corporation. All rights reserved.

Module Title - 29 Mercer Road
Global Education & Productivity

The performance index is a fast lookup table that is distributed over all of the backend nodes. There are two copies with one on each power rail. The performance index is broken up with each node taking a portion of the alphanumeric index which makes up the Content Addresses. The front-end nodes know which portions of the alphanumeric index are on each of the backend nodes. In the example above, if a read request comes in for a CA beginning with an A, the first node will look up the locations of the object. Of the two nodes designated in the performance index, the least busy will be selected to be read from and service the request.

If the performance index is corrupted, it will use the remaining copy. If both are corrupted, the system will work through broadcast queries to all nodes to find objects. This means it will still work, but more slowly. Typically response times will degrade from 200 –400 ms to about 1 second. Once the performance index is finished rebuilding itself, the system will renew normal operation.

Centera's Architecture: Intelligent Object Distribution

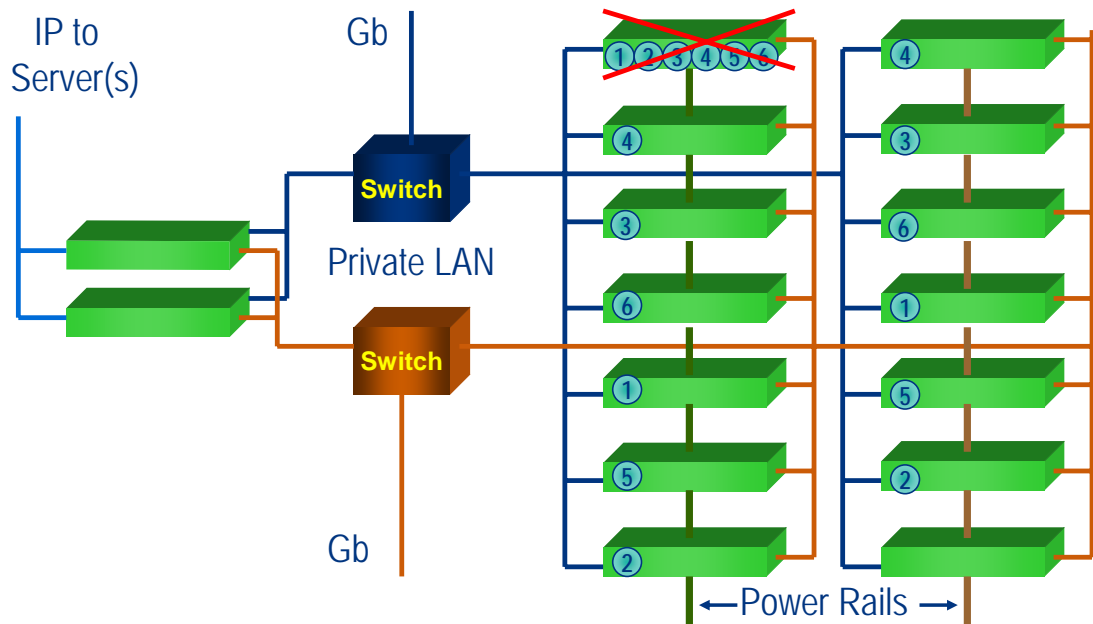


© 2005 EMC Corporation. All rights reserved.

Module Title - 30  Mercer Road
Global Education & Productivity

This slide gives an example of how objects might be written to Centera. The idea here is that disks and nodes are not mirrored, objects are. Also it is important to point out that when we write objects to centera they are not haphazardly written, there is a process that distributes them to the available media that tries to minimize data loss even if multiple devices fail.

Centera's Architecture: Self-Healing



© 2005 EMC Corporation. All rights reserved.

Module Title - 31  Mercer Road
Global Education & Productivity



Review of Centera Architecture

- **RAIN**
 - A full copy of Centera software on every node
 - Distributed work-load
 - No central or master node, all are identical
 - There is a principal node, but this is invisible to the user
- **Self-Managing**
 - Intelligent load balancing based on capacity and usage
 - LAN is private and requires no network administration
- **Self-Configuring**
 - New nodes are automatically configured
 - Must specify which role, node does the rest
 - Performance Index automatically scales as Centera grows
- **Self-Healing**
 - Reallocates and copies objects in case of disk or node failure



Term #11: Superior TCO Management

- Compared to other storage offerings Centera's self-managing, self-configuring and self-healing capabilities generate significant management cost savings
 - Managed TB's/FTE:
 - Centera = 250-350TB (CPM vs CPP)
 - Tape = 10-15TB
 - Optical Library = 2-10TB

It has been estimated that 1 person can manage >160TB's of Centera. This is compared to 1 person managing 10-15TB's of tape or 2-10TB's of optical.

The numbers for tape and optical here come from partner and customer feedback as well as numerous tco studies that have been done.

A typical FTE (full time employee) costs around \$100,000. Therefore 10-16x better management efficiencies for Centera will lead to an extreme cost advantage. This is the biggest TCO advantage that we have. It is useful here to talk about the self managing, configuring and healing capabilities that allow for hands-free management.

Lower Media Management Costs

Tape Drives & Cartridges

- **Media Mgmt.**
 - Multiple copies of cartridges are made for reliability
 1. Inside Silo
 2. Outside Silo
 3. Offsite On Shelf
- **Media Degradation**
 - Cartridges need to be refreshed every 2-4 years to ensure reliability
 - Media requires sampling to ensure integrity (tapes monitored by batch)
- **Migration**
 - when next generation of drives/media are released or upgrade is required:
 - HSM must be re-formatted to ensure compatibility
 - Total rebuilding of indexing database required



Optical Drives & Cartridges

- **Optical drives have changed formats 6 times in the last 7 years**
 - Not all drives/jukeboxes are backward compatible
 - Must reformat to ensure readability
 - Total rebuilding of indexing database required
- **Typically 2-3 copies of media are kept**
 1. Primary On-Site
 2. Secondary On-Site
 3. Offsite Copy

* Copies are rotated between 3 sites

© 2005 EMC Corporation. All rights reserved.

Module Title - 34 Mercer Road
Global Education & Productivity

Most people only pay attention to acquisition costs when considering tape and/or optical. However the story is much larger than that.

Optical and tape media require significant management beyond initial purchase.

Many copies of the data are kept b/c of the unreliability and deficiencies of the media and the libraries in which they are stored.

Often times up to 3 copies:

- Inside Silo
- Outside Silo
- Offsite On Shelf

Furthermore this media needs to be refreshed every 2-4 years from wear and tear and usage.

This is going to cost \$\$. Both for the new cartridges and the labor required to move data to the newer cartridges.

And then consider what happens if you move to a new generation of drives and or cartridges in the process.

Then you must ensure backwards compatibility which can be difficult. Typically a full data migration is required with a next generation of drives

And/or media. This means:

- HSM must be re-formatted to ensure compatibility
- Total rebuilding of indexing database required