

# Server Architectures: Data Storage

January 2005

René J. Chevance

## Foreword

- This presentation is an introduction to a set of presentations about server architectures. They are based on the following book:

**Serveurs Architectures: Multiprocessors, Clusters, Parallel Systems, Web Servers, Storage Solutions**  
René J. Chevance  
Digital Press December 2004 ISBN 1-55558-333-4  
<http://books.elsevier.com/>

This book has been derived from the following one:

**Serveurs multiprocesseurs, clusters et architectures parallèles**  
René J. Chevance  
Eyrolles Avril 2000 ISBN 2-212-09114-1  
<http://www.eyrolles.com/>

The English version integrates a lot of updates as well as a new chapter on Storage Solutions.

Contact: [www.chevance.com](http://www.chevance.com)

[rjc@chevance.com](mailto:rjc@chevance.com)

## Organization of the Presentations

- Introduction
- Processors and Memories
- Input/Output
- Evolution of Software Technologies
- Symmetric Multi-Processors
- Cluster and Massively Parallel Machines
- ➔ Data Storage (this document)
  - See contents on next slide
- System Performance and Estimation Techniques
- DBMS and Server Architectures
- High Availability Systems
- Selection Criteria and Total Cost of Possession
- Conclusion and Prospects

Page 3

© R.J Cheavance

## Data Storage - Contents

- Data Storage
  - Storage Issues
  - Magnetic Disks
  - File Systems
  - Remote File Access
    - NFS, CIFS, DAFS
  - Disk organizations
    - JBOD, SBOD
    - RAID
  - Storage Virtualization
  - Scatter/Gather
  - Comparing the various RAID levels
    - RAID Performance
    - RAID Implementation
  - Architectural options for storage virtualization
  - Storage Architectures: DAS, SAN, NAS and iSCSI
    - Integration of Fibre Channel and Internet
    - Integration of SAN and NAS
  - Summary of storage architecture options
  - SNIA Architecture Model
  - Storage Management
  - Data Compression
  - Commercial Storage Subsystems: BlueArc, EMC DMX, IBM ESS, Network Appliance
  - Data Backup and Restore
  - Real-time Data Copy
  - Resource Optimization in Backup and Restore
  - Data Life Cycle
  - Technologies Supporting Backed-up Data
  - RAIT and Tape Virtualization
  - Data Archiving
  - Storage of Reference Information

Page 4

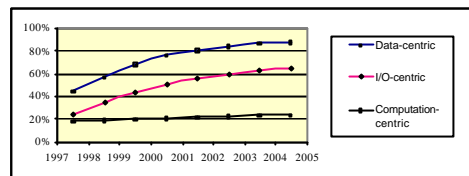
© R.J Cheavance

## Storage Issues

- Units being used in storage

Name	Abbreviation	Value
Gigabyte	GB	10 <sup>9</sup> bytes
Terabyte	TB	10 <sup>12</sup> bytes
Petabyte	PB	10 <sup>15</sup> bytes
Exabyte	EB	10 <sup>18</sup> bytes

- Allocation of Expenditures for Servers (Source Gartner)



- Data Centric: Most of Company's data is made accessible (e.g. Data Warehouse, Decision Support)
- I/O Centric: Applications stressing I/O performance (e.g. OLTP)
- Computation Centric: Applications having not such extreme requirements on I/O performance or storage size

Page 5

© R.J Chevance

## Storage Issues(2)

- World-wide production of original content in 2002 (Source [LYM03])

Medium	Content Type	TB/yr Upper estimate	TB/yr Lower estimate
Paper	Books	39	8
	Newspapers	138.4	27.7
	Office Documents	1,397.5	279.5
	Mass market periodicals	52	10
	Journals	6	1.3
	Newsletter	0.9	0.2
	<b>Subtotal</b>	<b>1,633.8</b>	<b>326.7</b>
Film	Photographs	375	37.5
	Cinema	6,078	12
	Made for TV films	2,531	2.53
	TV series	14,155	14,155
	Direct to video	2,49	2,49
	X-Rays	20	20
	<b>Sub-total</b>	<b>420,254</b>	<b>76,687</b>
Optical	Audio CD	58	6
	CD ROM	1.1	1.1
	DVD	43.8	43.8
	<b>Subtotal</b>	<b>102.9</b>	<b>50.9</b>
Magnetic	Videotape	1,340,000	1,340,000
	Audiotape	128.8	128.8
	Digital Tape	250	250
	Mini Digital Videocassettes	1,265,000	1,265,000
	Floppy disc	80	80
	Zip	350	350
	Audio MiniDiscs	17	17
	Flash	12	12
	Hard Disk	1,986,000	403
	<b>Sub-total</b>	<b>4,999,230</b>	<b>3,416,230</b>
<b>Total</b>		<b>5,421,220.7</b>	<b>3,493,294.6</b>

Page 6

© R.J Chevance

## Storage Issues(3)

### ■ Size of the Internet in Terabytes in 2002 (Source [LYM03])

Media	2002 (TB)
Surface Web	167
Deep Web	91.85
Email (originals)	440,606
Instant messaging	274
<b>Total</b>	<b>532,897</b>

### ■ Data Storage server objectives:

- Availability and reliability of data
- Performance (response time and bandwidth)
- Scalability
- Ease of operation and administration
- TCO (Total Cost of Ownership)
- Backup and Restore

Page 7

© R.J Chevance

## Storage Issues(4)

### ■ Storage Facts Figures, Estimates and Rules of Thumb (Source [www.horison.com](http://www.horison.com))

Average annual storage demand rate (primary occurrence of data, all platforms)	50-60%
Amount of disk data stored on Unix, Windows 2000 and Linux systems (estimate)	85%
Average disk allocation levels for z/OS (eSeries mainframes using DFSMS suite)	60-80%
Average disk allocation levels for iSeries (AS/400 servers)	60-80%
Average disk allocation levels for Unix/Linux	30-50%
Average disk allocation levels for Windows 2000/NT	20-40%
Ratio of block data to file data	1.5/1
Average annual disk drive capacity increase	60%
Average annual disk drive performance improvement (seek; latency, and data rate)	<10%
Increase in disk drive capacity per actuator since the disk drive in 1956	39260x
Increase in native tape cartridge capacity since the first tape cartridge in 1984	1250x
Average multi-user server utilization (% busy)	25-40%
Average tape cartridge utilization levels for virtual tape systems	60-85%
Estimated range of disk data managed per administrator (distributed systems - Windows 2000, Unix, Linux)	500 GB - 1 TB

Page 8

© R.J Chevance

## Storage Issues(5)

- **Storage Facts Figures, Estimates and Rules of Thumb (Source [www.horison.com](http://www.horison.com)) continued**

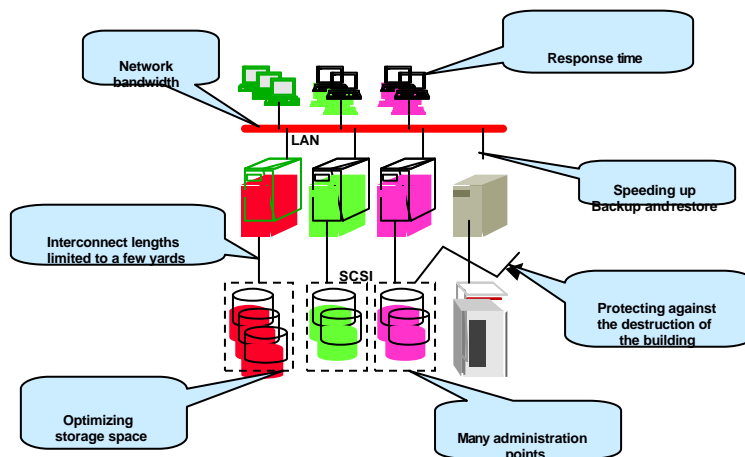
Estimated range of disk data managed per administrator (z/OS, mainframe)	> 30 TB
Estimated range of automated tape data managed per administrator (all platforms)	40 TB + External Backup (varies widely based on library size)
Average CAGR of e-mail message size	90%
Average growth rate of e-mail spam	~350%
Estimated percentage of SANs that are homogeneous (the same OS)	75% (Unix and Windows 2000 systems)
Average size of e-mail message and attachments in 2002	50 kb
Average size of e-mail in 2007 (estimate)	650 kb
Number of e-mails sent daily in 2001	12 000 000
Number of e-mails sent daily in 2005 (estimate)	>35 000 000
Percentage of all e-mail traffic that is spam (also known as bandwidth burning)	62% (and growing)
Annual growth in all Internet traffic	80%
Percentage of digital data stored on single user systems	56%
Percentage of digital data stored on removable media (tape and optical)	>80%

Page 9

© R.J Cheavance

## Storage Issues(6)

- **Issues for Information System architects (inspired by Bull)**

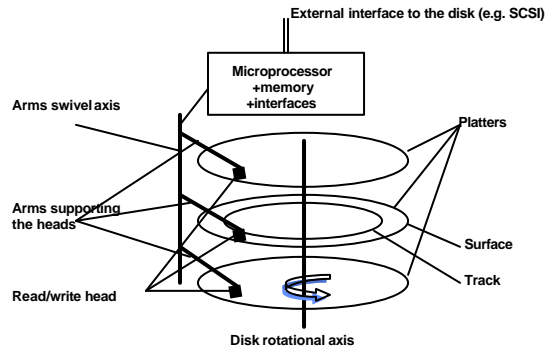


Page 10

© R.J Cheavance

## Magnetic Disks

### ■ Structure of a disk (simplified)



### ■ Components of disk access time (for a read operation):

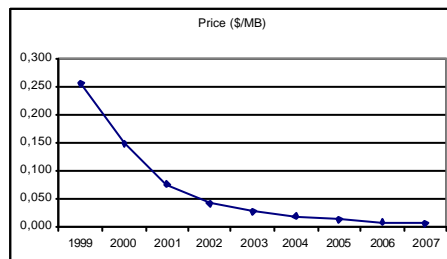
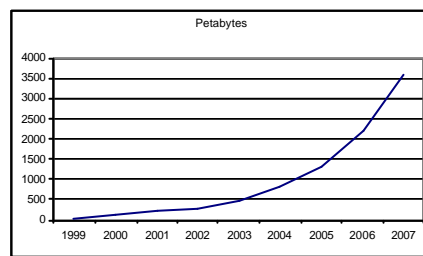
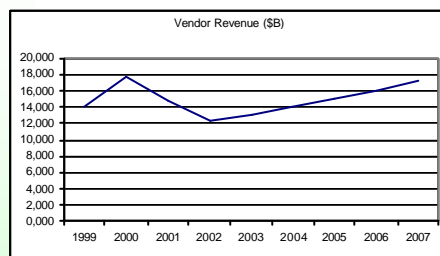
- Seek time
- Rotational latency
- Internal transfert time (from disk into disk electronic)
- External transfert time (e.g. onto the SCSI bus)

Page 11

© R.J Chevence

## Magnetic Disks(2)

### ■ Worldwide Market for External Controller-based Disk Storage (Source Gartner)

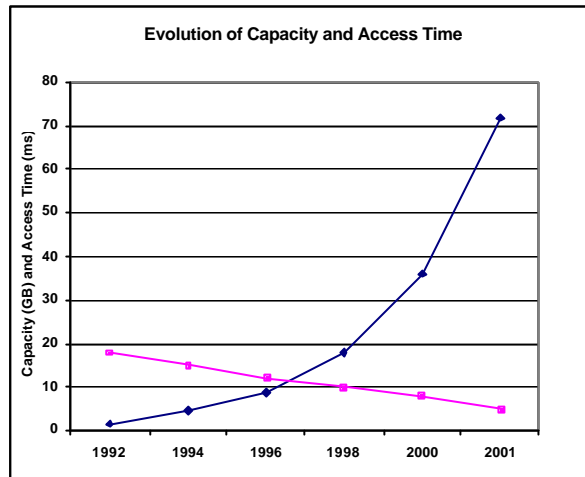


Page 12

© R.J Chevence

## Magnetic Disks(3)

### ■ Evolution of disk capacity and access time

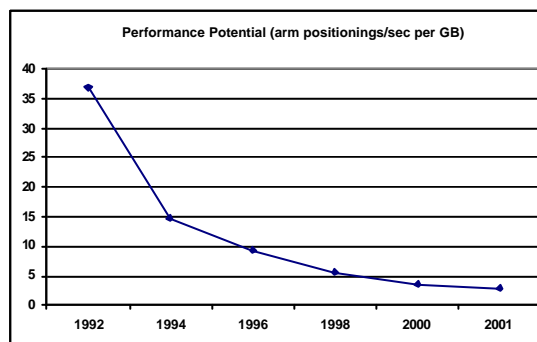


Page 13

© R.J Cheavance

## Magnetic Disks(4)

### ■ Evolution of access density



- In addition to the reduction of the access density, the increase in disk capacity tends to reduce the number of disks necessary to hold a database, thus stressing the number of accesses to disks

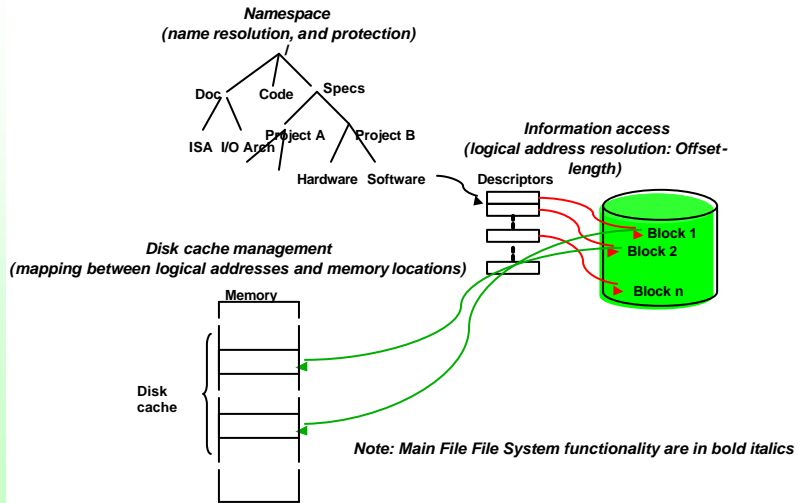
- Possible ways to solve the problem:

- Spread over several disks in parallel (RAID)
- Caching data in DRAM (issue: writes are not securized)
- Storage subsystems with securized memory

Page 14

© R.J Cheavance

## ■ File system functionality

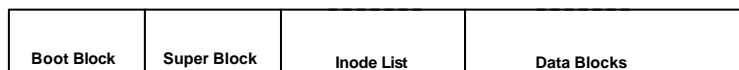


Page 15

© R.J Cheavance

Note: Through an unfortunate choice of language, the term "File System" refers both to the collection of files in a system and to the software which manages those files

## ■ General structure of a Unix File System



### ■ Super Block contents:

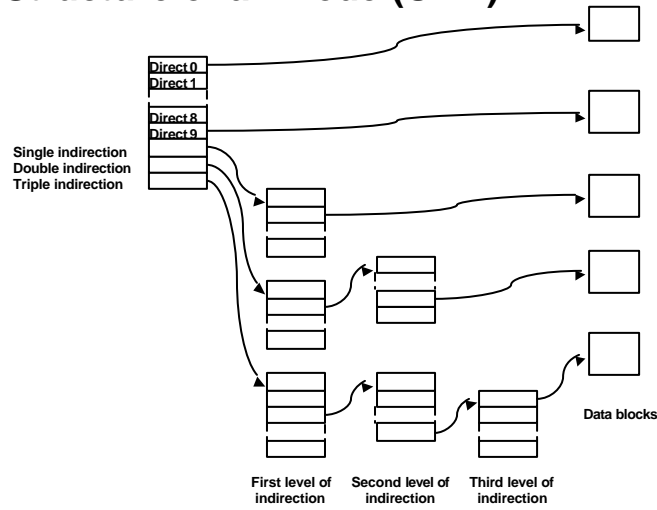
- the size of the file system
- the number of unused (free) blocks in the file system
- a list of the free blocks
- the index of the first free block
- the size of the list of file descriptors (inodes)
- the number of free inodes
- the list of free inodes
- the index of the first free inode
- locks for the lists of free blocks and inodes
- a flag indicating that the Super Block has been modified.

Page 16

© R.J Cheavance

## File Systems(3)

### ■ Structure of an inode (Unix)

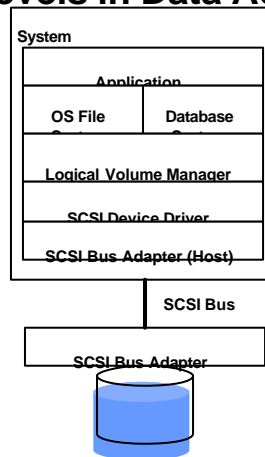


Page 17

© R.J Chevance

## File Systems(4)

### ■ Functional Levels in Data Access



A file system can be layered on top of a Logical Volume Manager. The role of an LVM is to provide the file system with an abstraction of the concept of a storage volume.

Page 18

© R.J Chevance

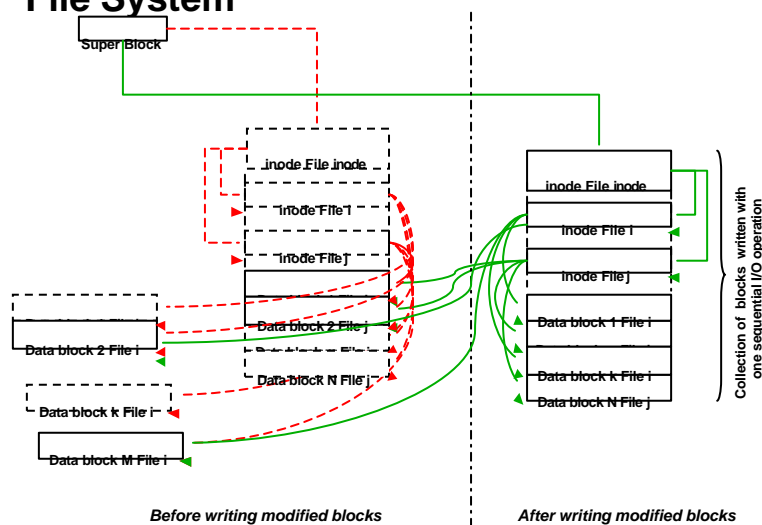
### ■ Journaled File Systems

- With a classical File System, an incident causes the File System, at system reboot, to perform a sequential search over the set of disks in order to check the coherency of the meta data (i.e. data describing the structure of the File System (*fsck* on Unix))
- In a journaled file system, all actions which change the meta data of the file system (for example, the meta data in a Unix file system which is represented by the i-nodes; or the internal data structures of NTFS, the native file system for Windows Server 2003) are managed as transactions, and are therefore able to be cancelled or re-done without side effects if they do not complete. This approach makes it possible to reduce start-up times for the systems by cancelling or replaying actions which were not completed at the time of the failure incident

Page 19

© R.J Chevance

### ■ Optimized file systems - Log Structured File System



Page 20

© R.J Chevance

## File Systems(7)

### ■ Log Structured File System

- Optimization of the write operations based upon the following principles:
  - no in-place modification of data
  - replacement of several random operations by one sequential I/O operation, collecting up all the blocks which need writing
- Example: Network Appliance's WAFL (Write Anywhere File Layout)
- Comparison of "Classic" and Log-structured File Systems

	Log Structured File System	"Classic" File System
Advantages	Improved write performance due to minimizing head movement Simplification of backup operations (snapshot, incremental, differential,...)	Widespread adoption Well-proven technology
Disadvantages	Requires custom, complex implementation Parameterizing the free-space reclamation operation	Write performance Backup is more complex

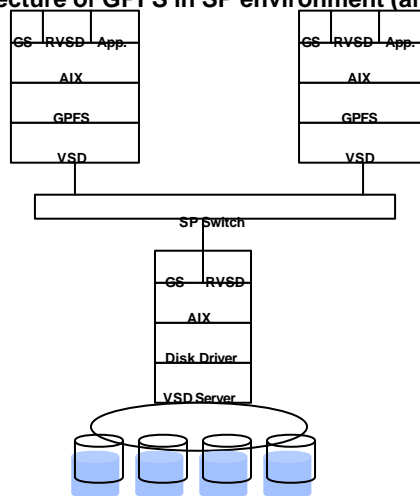
Page 21

© R.J Cheavance

## File Systems(8)

### ■ Parallel File Systems (exploiting parallelism in a cluster or MPP environment)

- Architecture of GPFS in SP environment (after IBM)



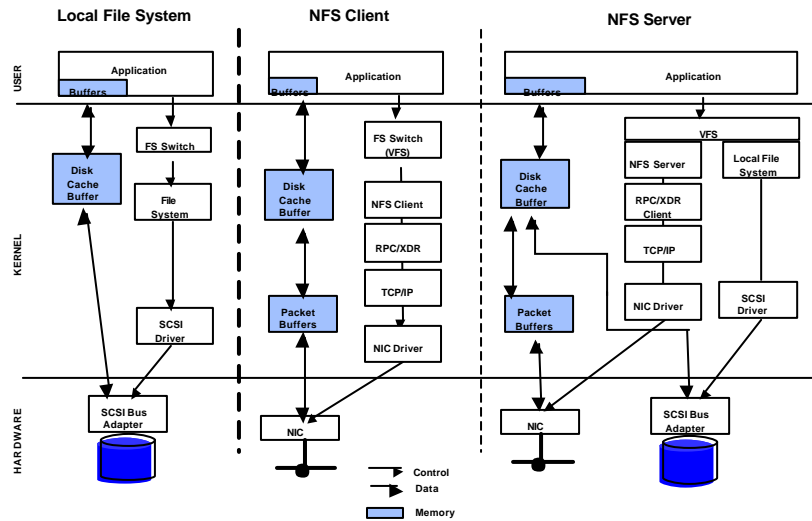
Page 22

© R.J Cheavance

## Remote File Access

### ■ NFS (Network File System)

#### □ Software and hardware layers in NFS



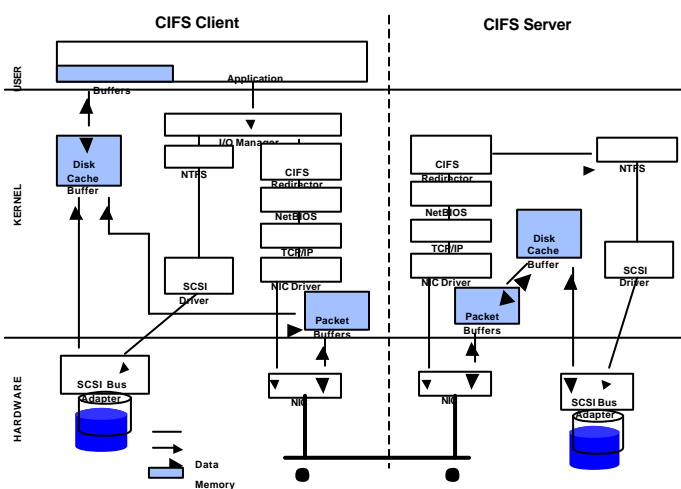
Page 23

© R.J Chevance

## Remote File Access(2)

### ■ Common Internet File System (CIFS)

#### □ Architecture of CIFS - Windows



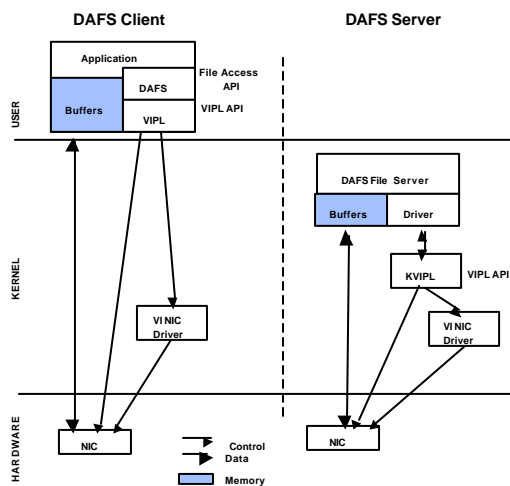
Page 24

© R.J Chevance

## Remote File Access(3)

### ■ DAFS Direct Access File System

#### □ Software layers involved in DAFS file access



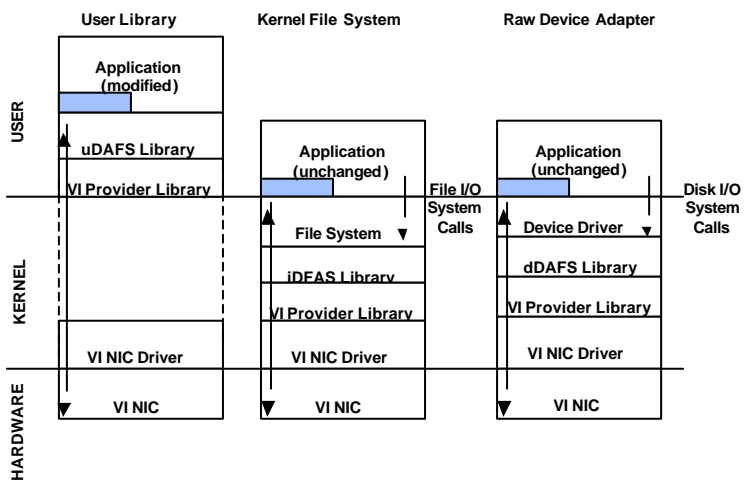
Page 25

© R.J Cheavance

## Remote File Access(4)

### ■ DAFS Direct Access File System(2)

#### □ DAFS implementation options



Page 26

© R.J Cheavance

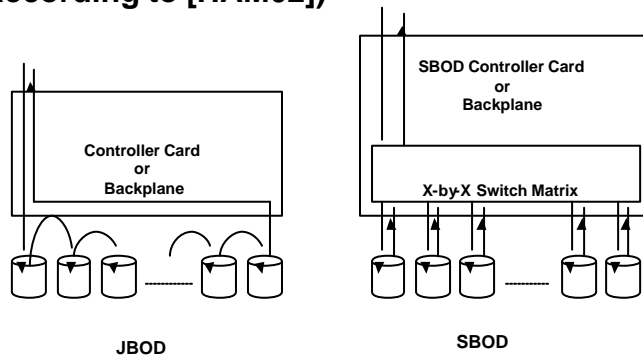
# Disk Organizations

Page 27

© R.J Cheavance

## JBOD and SBOD

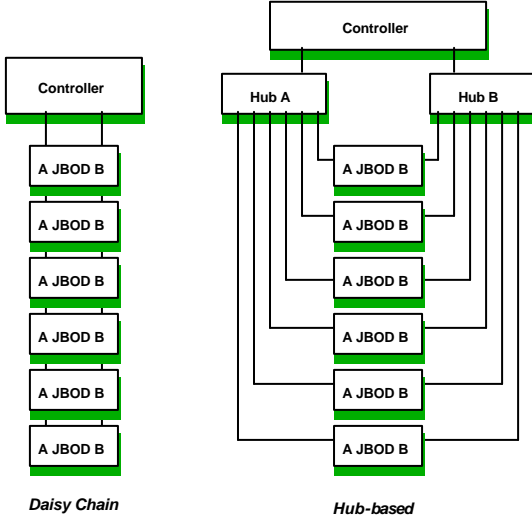
- **JBOD = Just a Bunch Of Disks**
  - Objectives: capacity and reduced cost
- **SBOD = Switched Bunch Of Disks**
  - Objectives: capacity, availability and performance
- **Comparison of JBOD and SBOD architectures (according to [HAM02])**



Page 28

© R.J Cheavance

■ JBOD Topologies

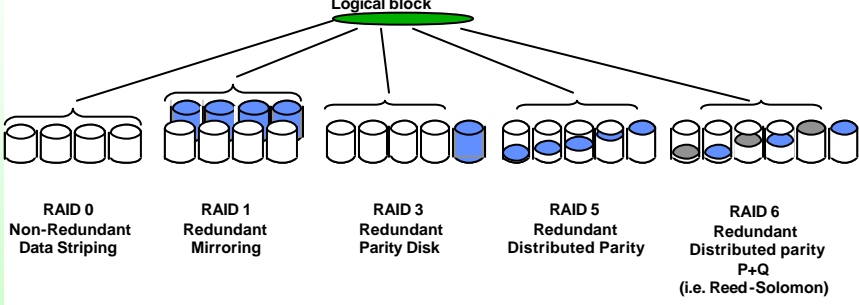


Page 29  
© R.J Chevance

■ RAID Organization

- RAID : Redundant Array of Independent (Inexpensive) Disks
- Concept formalized by a research team at the University of Berkeley
- Principles:
  - Distributing the data across several disks (data striping)
  - Redundancy (except for RAID 0) based upon XOR operation

■ RAID Architectures

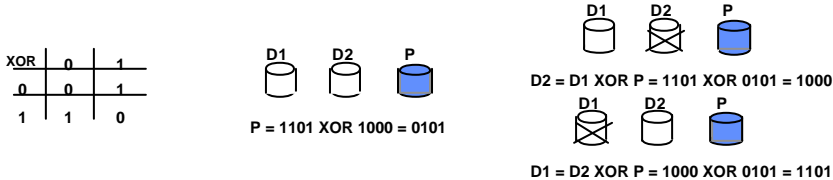


- MAID (Massive Array of Inactive Disks): collection of SATA or PATA disks only active when accessed. Low acquisition cost and reduced electrical consumption. Can be used wherever access time is not a key factor (e.g. cache to a tape library)

Page 30  
© R.J Chevance

# RAID(2)

## XOR-based redundancy scheme

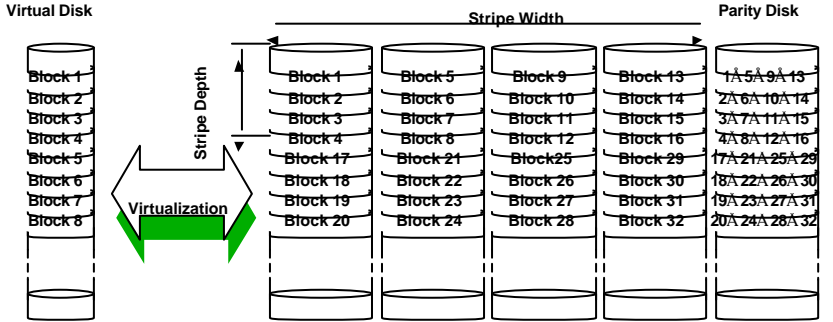


## Use cases for RAID:

- RAID 0: performance without redundancy
- RAID 1: performance and expensive redundancy ( 2 x disks)
- RAID 3: cost effective redundancy (1 parity disk for N data disks) and high performance for large data transfers
- RAID 5: cost effective redundancy (1 parity disk for N data disks) and high performance data transfers. Compared with RAID 3, the distributed parity removes contention on parity disk and performance on data updates is improved
- RAID 6: same basic characteristics as RAID 5, but with the ability to survive the concurrent failure of two disks rather than just one

# Storage Virtualization

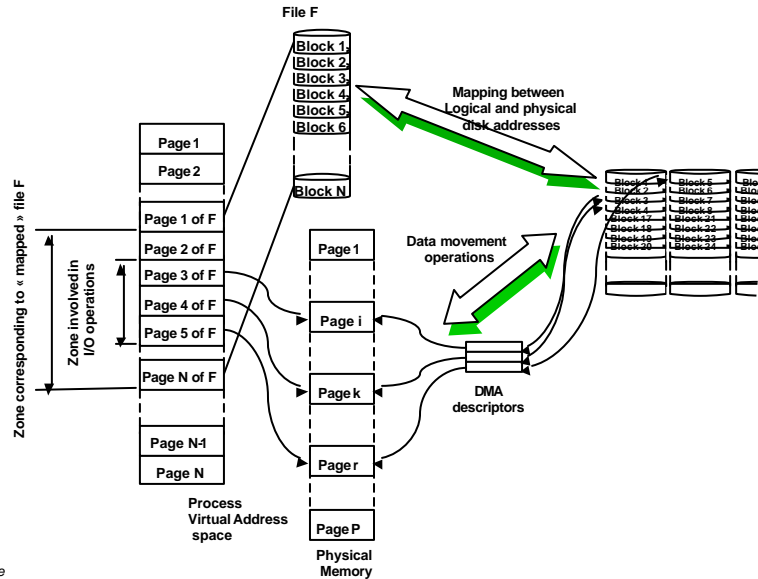
- Naive implementation of RAID leads to a very large (but reliable) disks
- Storage virtualization brings the vision of small virtual disks
- Advantages:
  - Offers several virtual disks of capacity and RAID levels according to their usage
  - Hide physical differences in disk sizes and technologies
  - Optimize the use of installed configurations
- RAID Implementation with Virtual Disks



Note: stripe depth (4 in this example) and stripe width (again, 4 in this example) are independent

## Scatter/Gather

### Scatter Reads and Gather Writes



Page 33

© R.J Cheavance

## Comparing the various RAID levels

### Comparison of the most popular RAID organizations (inspired by Veritas)

RAID Type	Name	Storage cost	Relative data availability	Large sequential read speed	Large sequential write speed	Random read bandwidth
0	Data striping	> 1	Lower than that of a conventional organization	Higher - Depends on the number of parallel disks	Higher - Depends on the number of parallel disks	Higher
1	Mirroring (of order M; M=2)	x M (see note[1])	> RAID 3 & RAID 5 < RAID 6	Up to M times a single disc	Lower than a single disc	Up to M times a single disc
0+1	Striped mirror (of order M; M=2 usually)	M x N	> RAID 3 & RAID 5 < RAID 6	Up to M times a RAID 0 equivalent	Can be higher than that of the single disc as a function of N	Up to M times a RAID 0 equivalent
3	Parity disk	N + 1	>> conventional disc	Higher Depends on the number of parallel disks and the need to compute parity (< RAID 0)	Higher Depends on the number of parallel disks and the need to compute parity (< RAID 0)	Higher
5	"spiral" parity	N + 1	>> conventional disc - RAID 3	< RAID 0 because of the parity check	< RAID 0	Higher > RAID 3
6	Double "spiral" parity	N + 2	Higher than all the other types	Slightly > RAID 5	< RAID 5 (2 parity 'blocks')	Slightly > RAID 5

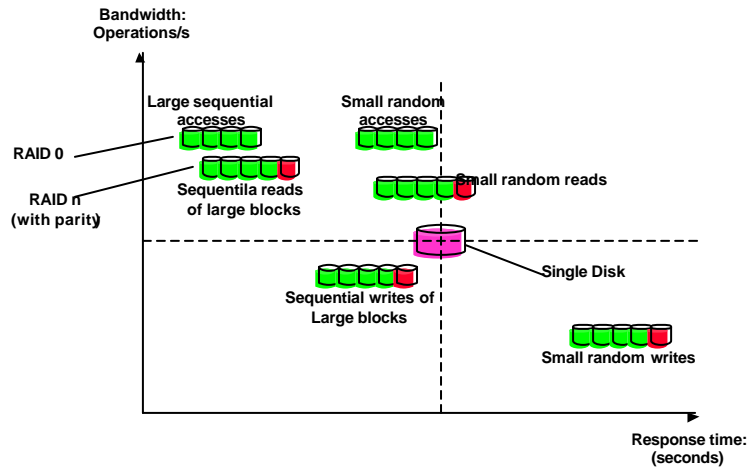
Page 34

© R.J Cheavance

Note [1]: If redistributing the data across a number of disks is unnecessary, use of the simple mirror results in a doubling of the number of disks required for data storage. It is also possible to have more than 2 mirrors (the number M is used in the table to specify the number of mirrors)

## RAID Performance

### ■ RAID Performance Positioning (derived from Veritas)



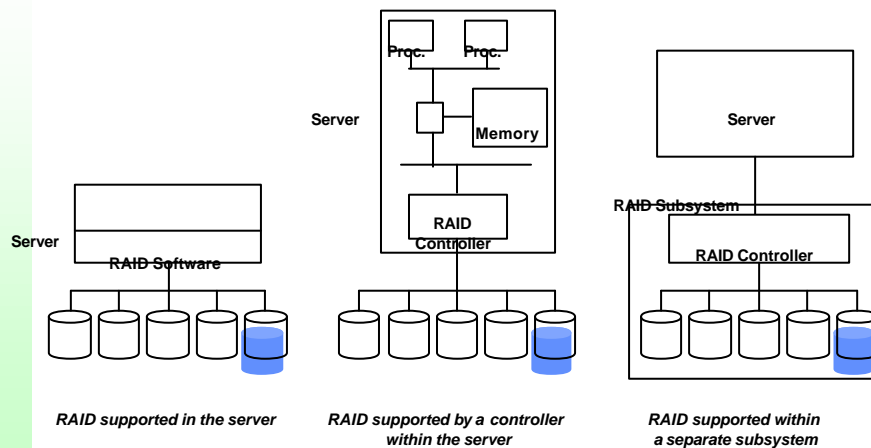
Page 35

© R.J Cheavance

## RAID Implementation

### ■ Architectural options for RAID Implementations

#### □ System architecture options



Page 36

© R.J Cheavance

## RAID Implementation(2)

### ■ Comparing System Architecture Options for RAID

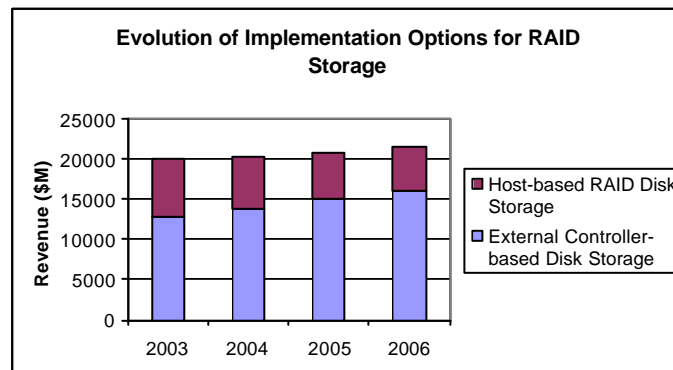
	RAID supported by the server	RAID supported by controller in server	RAID supported within a specialized subsystem
<b>Advantages</b>	<ul style="list-style-type: none"> <li>• Low cost</li> <li>• High connectivity (i.e., the server's innate connectivity)</li> <li>• Scalability (increasing server performance increases RAID performance)</li> <li>• High availability (no extra hardware elements involved)</li> </ul>	<ul style="list-style-type: none"> <li>• Moderate cost</li> <li>• Good execution times and good bandwidth (specialized hardware)</li> </ul>	<ul style="list-style-type: none"> <li>• Connectivity usually high (constrained by the subsystem) along with the possibility of connecting multiple subsystems</li> <li>• High bandwidth (specialized hardware)</li> <li>• Good write performance, if a write cache is available</li> <li>• High availability (doubling internal controllers and multiple access paths)</li> <li>• Independence between host interconnect (e.g. FC-AL) and disk (e.g. SCSI).</li> </ul>
<b>Disadvantages</b>	<ul style="list-style-type: none"> <li>• Server performance is impacted by the extra load of implementing RAID functionality</li> <li>• Data availability demands mean that the disks must have dual access interfaces to allow recovery after failure of the server</li> </ul>	<ul style="list-style-type: none"> <li>• Number of disks supported constrained by the connectivity capabilities of the controller</li> <li>• Data availability demands mean that the disks must have dual access interfaces to allow recovery after failure of the server or the controller</li> </ul>	<ul style="list-style-type: none"> <li>• Specialist hardware (redundant secure cache)</li> <li>• Higher cost</li> <li>• better response time than a pure server-based solution thanks to the server/subsystem interconnect</li> </ul>

Page 37

© R.J Chevanee

## RAID Implementation(3)

### ■ Evolution of the market for RAID support options (Source data Gartner 2004)

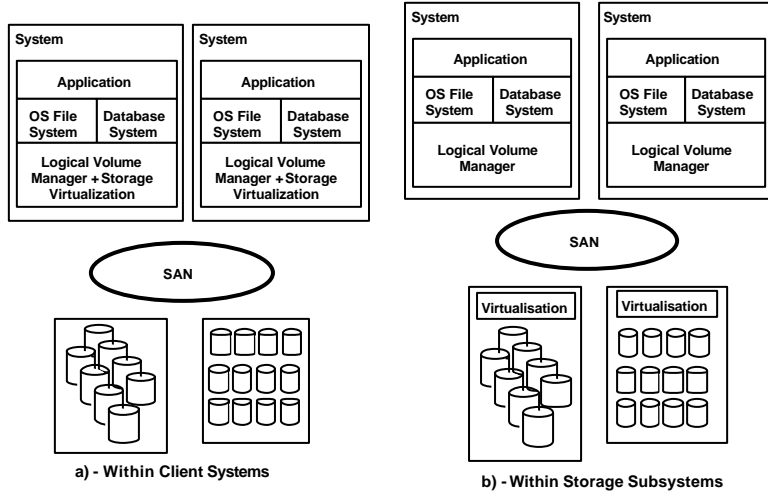


Page 38

© R.J Chevanee

## Architectural options for storage virtualization

### ■ System Architecture Options for Storage Virtualization

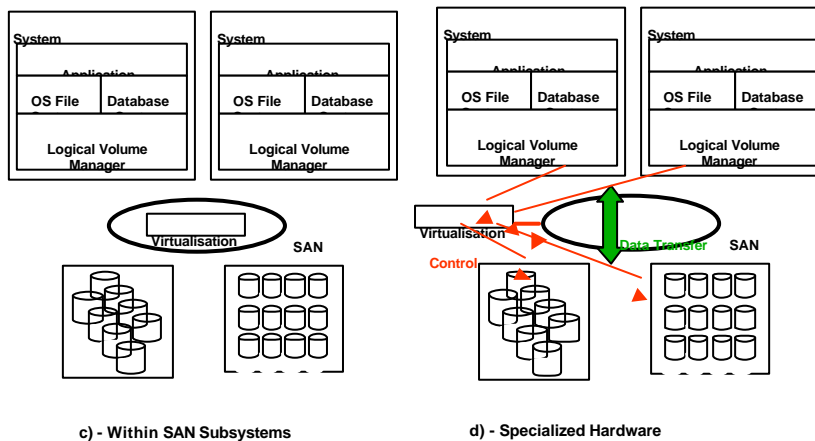


Page 39

© R.J Chevanca

## Architectural options for storage virtualization (2)

### ■ System Architecture Options for Storage Virtualization (continued)



Page 40

© R.J Chevanca

## Architectural options for storage virtualization(3)

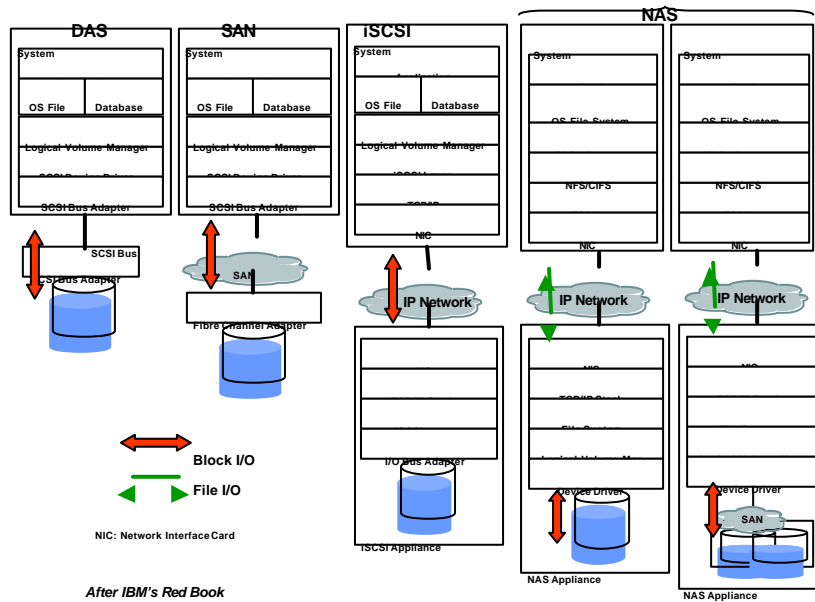
### ■ Comparison of System Architecture Options for Storage Virtualization

	Client systems	Storage Subsystem	SAN Subsystem	Specialist hardware
<b>ADVANTAGES</b>	<ul style="list-style-type: none"> <li>Virtualization based on proven principles</li> <li>Narrow integration with File Systems and DBMS</li> </ul>	<ul style="list-style-type: none"> <li>Allows the support of heterogeneous storage (technology and vendor independence)</li> </ul>	<ul style="list-style-type: none"> <li>Ability to connect diverse clients</li> </ul>	<ul style="list-style-type: none"> <li>Centralized control</li> <li>High performance due to separation of control and data transfer</li> <li>Supports heterogeneous clients</li> </ul>
<b>DISADVANTAGES</b>	<ul style="list-style-type: none"> <li>Global visibility of storage means that clusterization techniques must be used</li> <li>Administrative complexity</li> </ul>	<ul style="list-style-type: none"> <li>Multiple points of administration</li> <li>Solution proprietary to each vendor</li> <li>Global visibility of storage means that clusterization techniques must be used</li> <li>Qualification of the solution</li> <li>Cost of the various subsystems</li> </ul>	<ul style="list-style-type: none"> <li>Global visibility of storage means that clusterization techniques must be used</li> <li>Need to use clusterization techniques for availability</li> <li>Limits choice to hardware that can support virtualization</li> <li>Interoperability concerns between different vendors</li> </ul>	<ul style="list-style-type: none"> <li>Requires special drivers in the clients</li> <li>Difficulty of qualifying the solution in a heterogeneous environment</li> <li>High availability requires redundant equipment</li> <li>Complexity of connection</li> </ul>

Page 41

© R.J Cheavance

## Storage Architectures: DAS, SAN, NAS and iSCSI



Page 42

© R.J Cheavance

## Storage Architectures: DAS, SAN, NAS and iSCSI(2)

### ■ Comparison of DAS, NAS, SAN and iSCSI

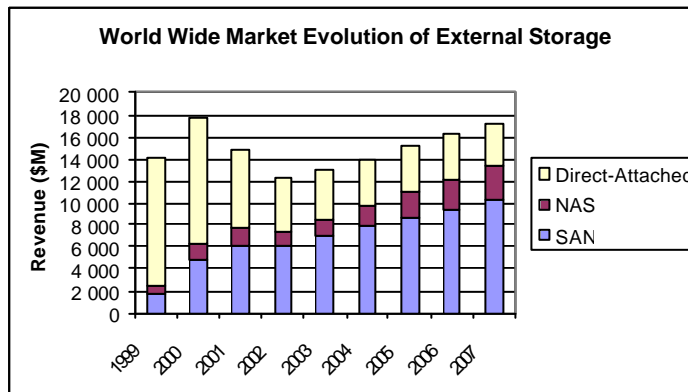
	DAS	NAS	SAN	iSCSI
Type of connection	<ul style="list-style-type: none"> <li>• SCSI</li> <li>• FC-AL</li> </ul>	<ul style="list-style-type: none"> <li>• Fast Ethernet</li> <li>• Fibre Channel</li> </ul>	<ul style="list-style-type: none"> <li>• Fibre Channel</li> </ul>	<ul style="list-style-type: none"> <li>• Internet</li> </ul>
Remote connection	<ul style="list-style-type: none"> <li>• Typically no</li> </ul>	<ul style="list-style-type: none"> <li>• Yes</li> </ul>	<ul style="list-style-type: none"> <li>• Possible</li> </ul>	<ul style="list-style-type: none"> <li>• Yes</li> </ul>
Type of I/O	<ul style="list-style-type: none"> <li>• Block</li> </ul>	<ul style="list-style-type: none"> <li>• File</li> </ul>	<ul style="list-style-type: none"> <li>• Block</li> </ul>	<ul style="list-style-type: none"> <li>• Block</li> </ul>
Performance	<ul style="list-style-type: none"> <li>• High</li> </ul>	<ul style="list-style-type: none"> <li>• Limited by the network</li> </ul>	<ul style="list-style-type: none"> <li>• Higher</li> </ul>	<ul style="list-style-type: none"> <li>• Limited by the network</li> </ul>
Data sharing	<ul style="list-style-type: none"> <li>• Implies NFS or CIFS</li> </ul>	<ul style="list-style-type: none"> <li>• Native</li> </ul>	<ul style="list-style-type: none"> <li>• Difficult (in 2002)/</li> </ul>	<ul style="list-style-type: none"> <li>• Difficult (in 2002)</li> </ul>
Cost reduction	<ul style="list-style-type: none"> <li>• No</li> </ul>	<ul style="list-style-type: none"> <li>• Yes</li> </ul>	<ul style="list-style-type: none"> <li>• Yes</li> </ul>	<ul style="list-style-type: none"> <li>• Yes</li> </ul>
Investment separation	<ul style="list-style-type: none"> <li>• No</li> </ul>	<ul style="list-style-type: none"> <li>• Yes</li> </ul>	<ul style="list-style-type: none"> <li>• Yes</li> </ul>	<ul style="list-style-type: none"> <li>• Yes</li> </ul>
Scalability	<ul style="list-style-type: none"> <li>• No</li> </ul>	<ul style="list-style-type: none"> <li>• Yes</li> </ul>	<ul style="list-style-type: none"> <li>• Yes</li> </ul>	<ul style="list-style-type: none"> <li>• Depends on network</li> </ul>
data Availability	<ul style="list-style-type: none"> <li>• Limited</li> </ul>	<ul style="list-style-type: none"> <li>• Yes if redundant</li> </ul>	<ul style="list-style-type: none"> <li>• Yes if redundant</li> </ul>	<ul style="list-style-type: none"> <li>• Yes if redundant</li> </ul>
Centralization of management and support	<ul style="list-style-type: none"> <li>• Typically no</li> </ul>	<ul style="list-style-type: none"> <li>• Yes</li> </ul>	<ul style="list-style-type: none"> <li>• Yes</li> </ul>	<ul style="list-style-type: none"> <li>• Yes</li> </ul>
Management	<ul style="list-style-type: none"> <li>• Traditional</li> </ul>	<ul style="list-style-type: none"> <li>• SNMP</li> </ul>	<ul style="list-style-type: none"> <li>• Difficult (in 2002)</li> </ul>	<ul style="list-style-type: none"> <li>• Difficult (in 2002)</li> </ul>
LAN-Free Backup	<ul style="list-style-type: none"> <li>• No</li> </ul>	<ul style="list-style-type: none"> <li>• Depends on NAS</li> </ul>	<ul style="list-style-type: none"> <li>• Yes</li> </ul>	<ul style="list-style-type: none"> <li>• Depends on iSCSI</li> </ul>
Server-Free Backup	<ul style="list-style-type: none"> <li>• No</li> </ul>	<ul style="list-style-type: none"> <li>• Depends on NAS</li> </ul>	<ul style="list-style-type: none"> <li>• Yes</li> </ul>	<ul style="list-style-type: none"> <li>• Depends on iSCSI</li> </ul>
Security	<ul style="list-style-type: none"> <li>• By the server</li> </ul>	<ul style="list-style-type: none"> <li>• By the servers and the network</li> </ul>	<ul style="list-style-type: none"> <li>• By the servers and storage network</li> </ul>	<ul style="list-style-type: none"> <li>• By the servers and network</li> </ul>
Installation	<ul style="list-style-type: none"> <li>• Specific to the server</li> </ul>	<ul style="list-style-type: none"> <li>• Simple</li> </ul>	<ul style="list-style-type: none"> <li>• Difficult (in 2002)</li> </ul>	<ul style="list-style-type: none"> <li>• Difficult (in 2002)</li> </ul>

Page 43

© R.J Cheavance

## Storage Architectures: DAS, SAN, NAS and iSCSI(3)

### ■ Worldwide Market for External Storage (Source data: Gartner 2003)



Market size: \$14B in 2001, \$17.3B in 2007

Page 44

© R.J Cheavance



## Storage Architecture Options

### ■ Summary of Storage Architecture Options

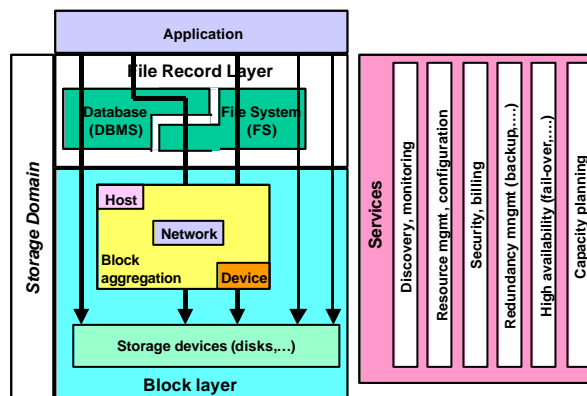
		Media	
		Direct (direct transfer To memory)	Network (TCP/IP)
Protocols	Blocks	DAS SAN	iSCSI
	File	DAFS	NAS

Page 47

© R.J Chevanee

## SNIA Architecture Model

- Model proposed by the Storage Networking Industry Association
  - Provide a common framework for the description of storage architectures
  - Does not make assumptions about implementation
- SNIA shared storage architecture (after SNIA)

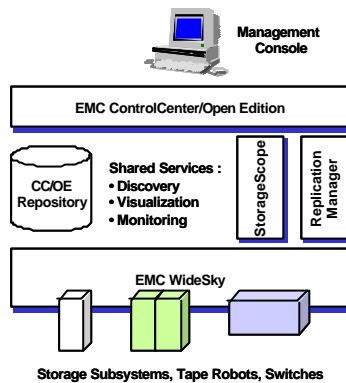


Page 48

© R.J Chevanee

## Storage Management

- Due to the growing importance of storage expenditures, the management of storage resources is a key factor in the search for TCO reduction
- Example of an approach: EMC's WideSky



- Components:
- WideSky which provides a homogeneous vision of the various storage equipment;
  - ControlCenter/Open Repository Edition: a database containing information on the collection of elements
  - a number of services shared by the offer's components - configuration discovery, visualization, operational monitoring...
  - StorageScope: tool for reporting of the use of storage resources;
  - Replication Manager;
  - ControlCenter/Open Edition: centralized management tool for storage resources.

Page 49

© R.J Cheavance

## Data Compression

- Data compression is widely used to save storage space as well as bandwidth
- Two classes of data compression techniques:
  - lossless compression: data is compressed without losing any information
  - lossy compression: data is compressed, but information is lost (used for pictures, video and audio)
- Description of these techniques is out of the scope of this presentation. Techniques used for lossless compression and described in the book:
  - Dictionary-based Compression
  - Fixed Bit Length Packing
  - Run Length Encoding (RLE)
  - Huffman Encoding (static or dynamic)
  - LZ77 Encoding
  - LZ77 Encoding
  - Arithmetic Coding

Page 50

© R.J Cheavance

## Data Compression(2)

- Data compression can be done on the host systems or within storage subsystems

- Comparison of Host-based Compression and Storage subsystem-based compression (derived from [BMC 01])

	Host-based Compression	Storage Subsystem-based Compression
<b>Advantages</b>	<ul style="list-style-type: none"> <li>• selective compression (e.g., one does not compress system data, file catalog etc.)</li> <li>• ability to choose the compression technique to fit the characteristics of the data</li> <li>• reduction in I/O traffic between host and disk subsystem</li> <li>• does not require any special capabilities in the storage subsystem</li> </ul>	<ul style="list-style-type: none"> <li>• best use of the storage capacity provided by the storage elements</li> <li>• improvement in host system performance (due to off-loading the compression task)</li> </ul>
<b>Disadvantages</b>	<ul style="list-style-type: none"> <li>• performance impact unless either specialized coprocessors are used or the number of processors is increased, either of which increases host system cost</li> </ul>	<ul style="list-style-type: none"> <li>• inability to be selective in what gets compressed and how it is compressed</li> <li>• no reduction in host-subsystem I/O traffic</li> <li>• disk occupation not known until data is compressed by the subsystem</li> </ul>

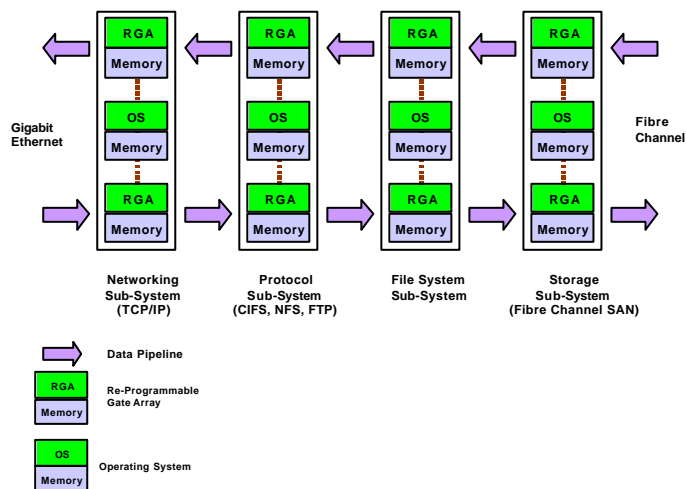
Page 51

© R.J Cheavance

## Commercial Storage Subsystems

- BlueArc

- Gate array-based implementation of a NAS

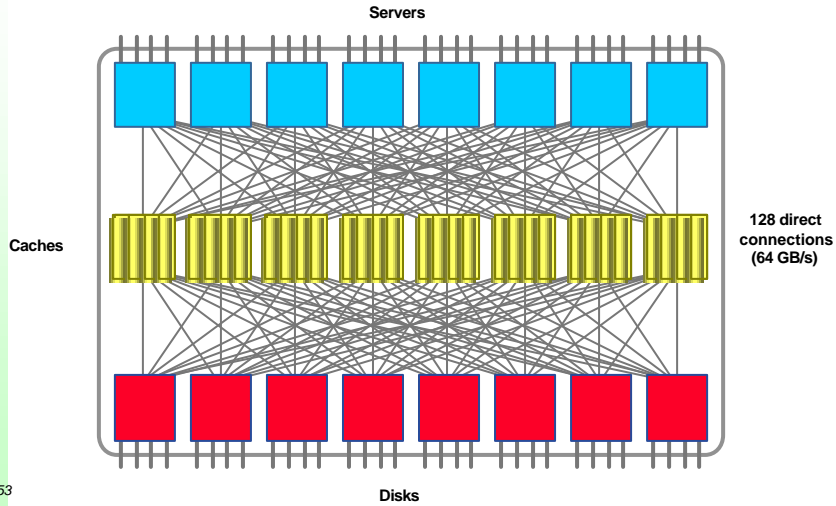


Page 52

© R.J Cheavance

## Commercial Storage Subsystems(2)

### ■ EMC Symmetrix DMX Series

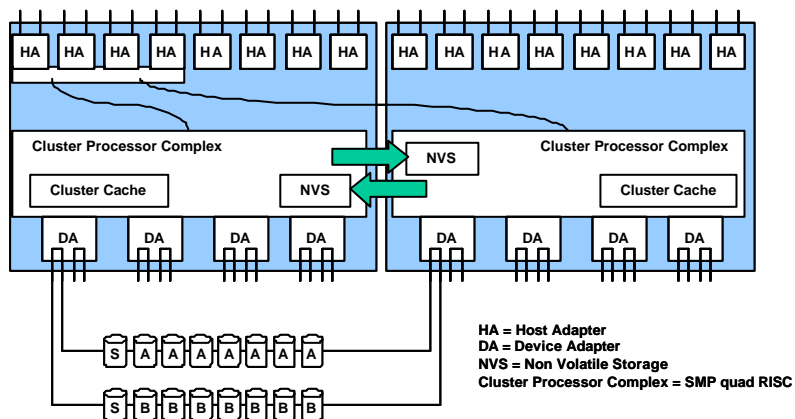


Page 53

© R.J Cheavance

## Commercial Storage Subsystems(3)

### ■ IBM ESS (Enterprise Storage Server) after IBM

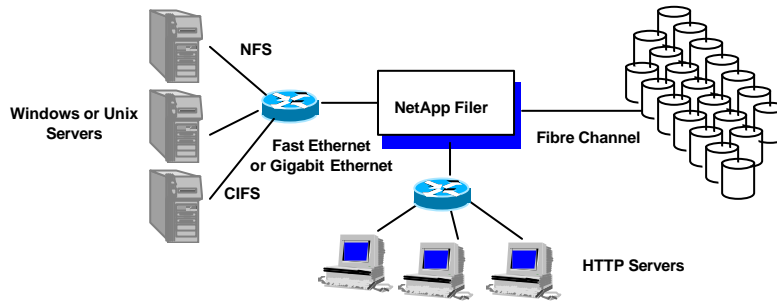


Page 54

© R.J Cheavance

## Commercial Storage Subsystems(4)

### ■ Network Appliance

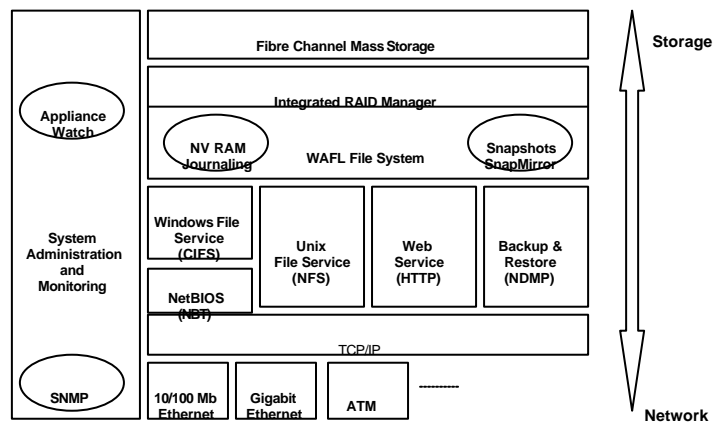


Page 55

© R.J Cheavance

## Commercial Storage Subsystems(5)

### ■ Network Appliance – Software Architecture



WAFL = Write Anywhere File Layout (log structured File system supporting the snapshot functionality)

Page 56

© R.J Cheavance

## Data Backup and Restore

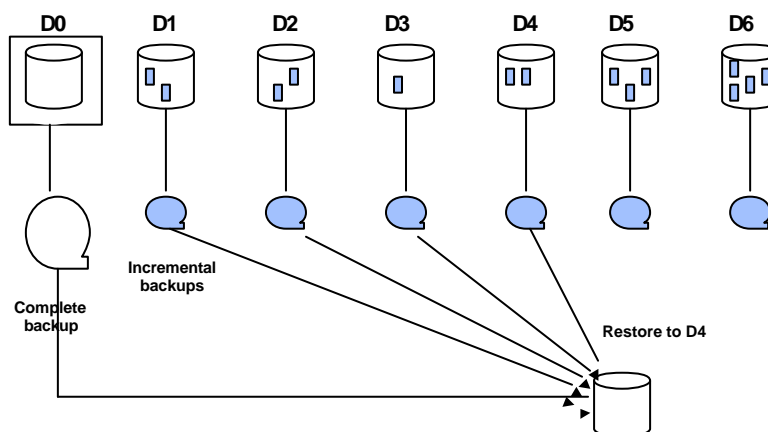
- Due to the ever growing importance of data, backup and restore are key to protect from the consequences of:
  - system failures
  - data destruction, for example by a fire
  - incorrect software operation (OS, DBMS or application)
  - human error
- Two key objectives have a strong influence on the choice of a backup/restore solution i.e. choice of a technology (support and robot) and policy. These objectives are:
  - Recovery Point Objective (RPO) which the point at which the system state must be restored;
  - Recovery Time Objective (RTO) which is the time allowed for restoring system operation (full or degraded service since all the applications are not critical from business continuity point of view)

Page 57

© R.J Cheavance

## Data Backup and Restore(2)

### ■ Incremental Backup



Restoring at Dx:

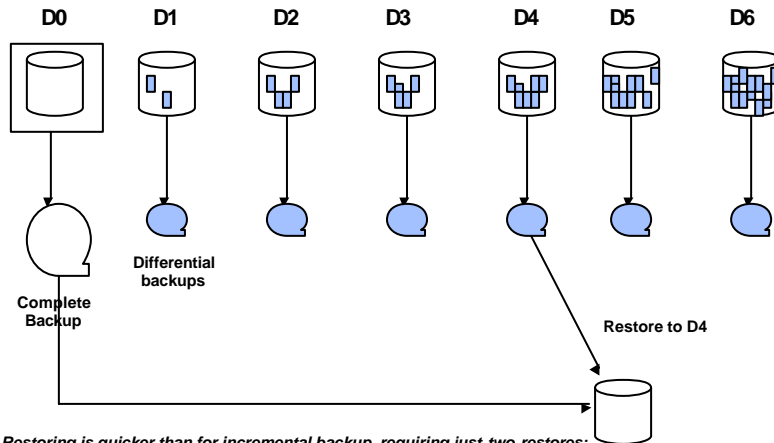
- restoring data from the most recent complete backup
  - successively restoring data from each incremental backup, up to the desired day
- Issue: time to restore

Page 58

© R.J Cheavance

## Data Backup and Restore(3)

### ■ Differential backup



Restoring is quicker than for incremental backup, requiring just two restores:

- restoration of the last complete backup
- restoration of the appropriate differential backup

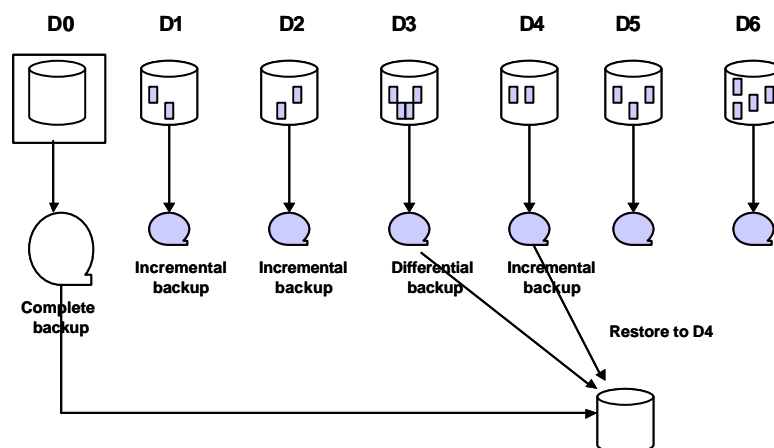
Page 59

Issue: growing volume of differential backups

© R.J Chevance

## Data Backup and Restore(4)

### ■ Incremental and differential Backup



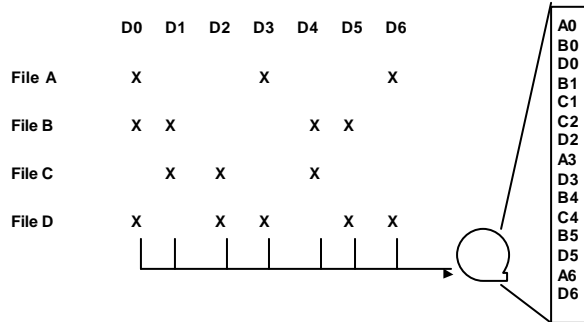
Page 60

© R.J Chevance

## Data Backup and Restore(5)

### ■ Progressive backup

#### □ Backup operations



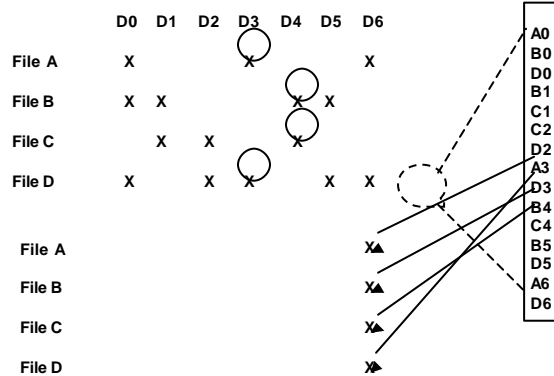
Page 61

© R.J Chevance

## Data Backup and Restore(6)

### ■ Progressive backup

#### □ Restore operation (e.g. at D4)



*In the restore operation, the database is consulted to identify exactly those backups needed to restore to the specified point, thus (in general) reducing the amount of data transferred and thus the time necessary.*

Page 62

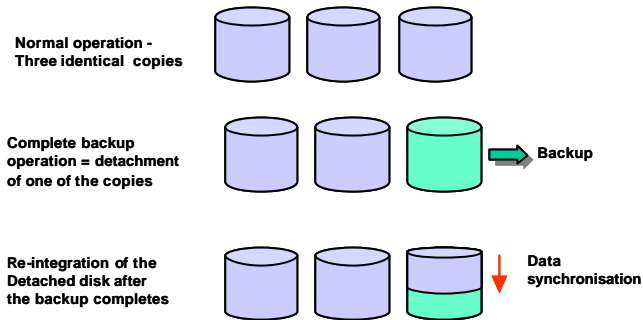
© R.J Chevance

## Data Backup and Restore(7)

- **Fast complete backup:**

- **Creating a complete backup is an expensive operation which can perturb normal system operation. To alleviate this nuisance, it is possible to implement a complete backup by starting a backup whose destination is not the “active” disk space as in a normal restore, but a new archive (e.g. Tivoli’s Instant Archive).**

- **Complete Backup in a RAID 1 Environment**

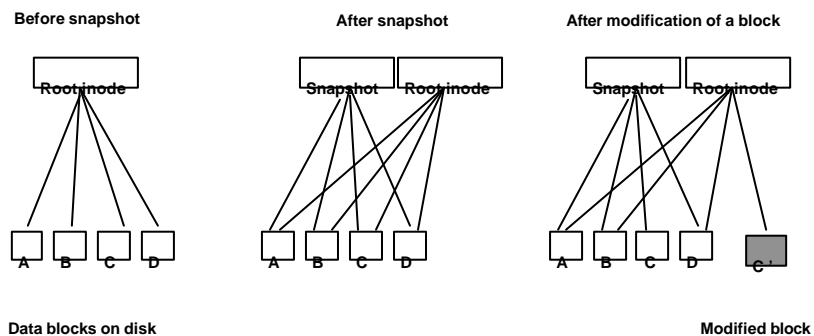


Page 63

© R.J Chevence

## Data Backup and Restore(8)

- **File System Snapshot (fast save operation)**

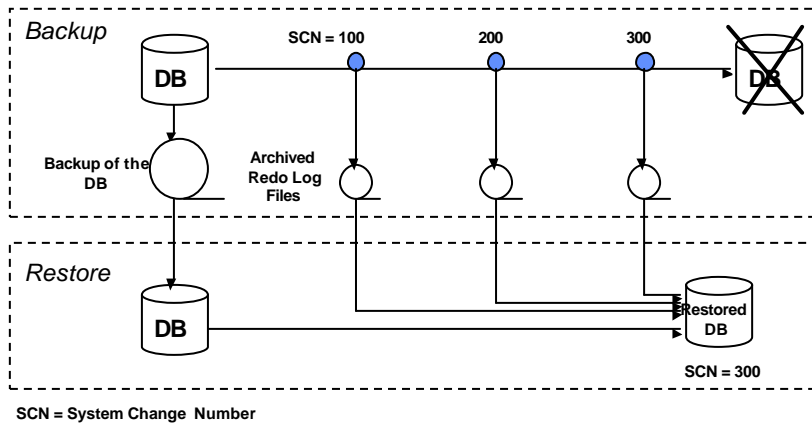


Page 64

© R.J Chevence

## Data Backup and Restore(9)

### Backup and Restore in a DBMS Context (Open Database backup with Oracle)

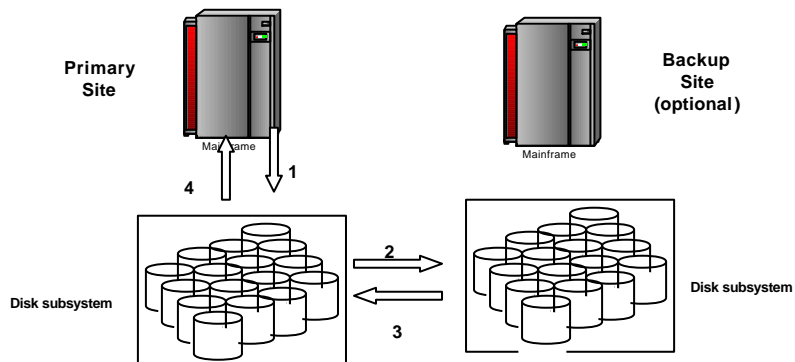


Page 65

© R.J Chevance

## Real-time Data Copy

- Related to disaster recovery rather than backup.  
Principle: maintaining a synchronous copy of data to a remote site
- PPRC/HRC/SRDF Synchronous Copy

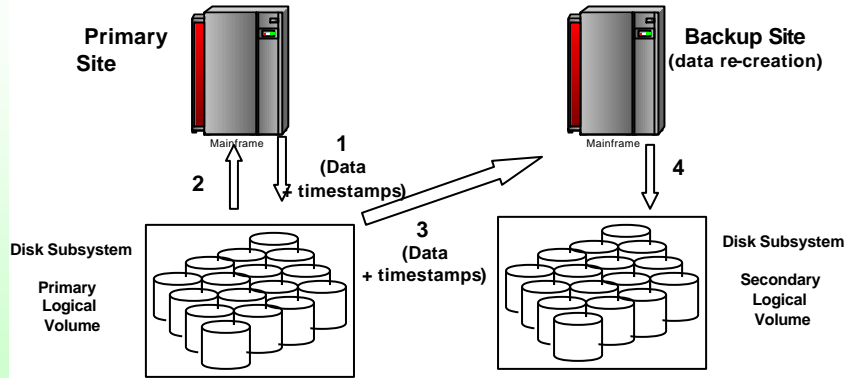


Page 66

© R.J Chevance

## Real-time Data Copy(2)

### ■ Differential update (XRC and HXRC)



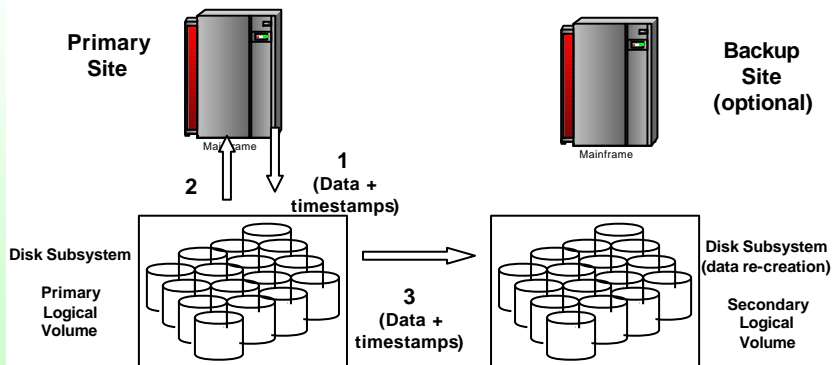
Page 67

© R.J Cheavance

## Real-time Data Copy(3)

### ■ Differential Update within the HARC Disk Subsystem

- Removes the need for a remote site
- Requires total compatibility between the disk subsystems

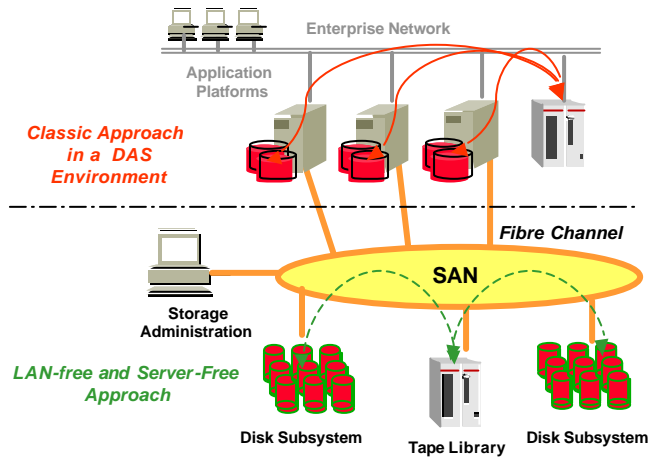


Page 68

© R.J Cheavance

## Resource Optimization in Backup and Restore

### ■ Server-free and Lan-free Backup/Restore

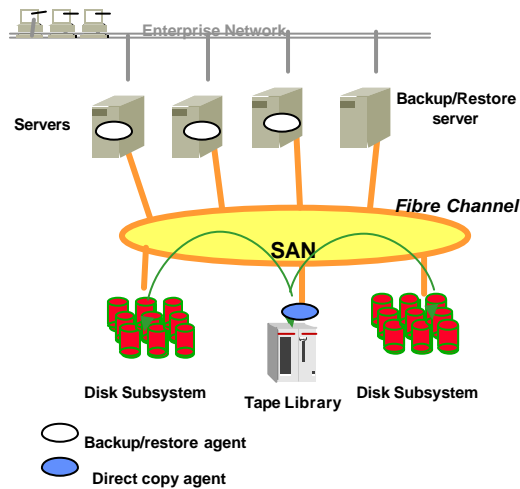


Page 69

© R.J Cheavance

## Resource Optimization in Backup and Restore(2)

### ■ NDMP (Networked Data Management Protocol) Operation



Page 70

© R.J Cheavance

## Data Life Cycle

- The data life cycle used to be based upon frequency of access. After a certain period of time, frequency of access to data decreased and data was candidate for migration towards lower cost media or even deletion. There are now other factors to consider such that government regulations for retention of data or the value of data.
- Classical life cycle was composed of three stages:
  - Online: retrieval time in ms, life time about 7 days, typically kept on disks
  - Nearline: retrieval time in seconds, life time of about 60 days, typically kept on low cost disks such as MAID or tape library
  - Archival and/or deletion (CDs,....)
- As retrieval activity decrease , the volume of stored data increase.

Page 71

© R.J Cheavance

## Technologies Supporting Backed-up Data

- There are two major backup media technologies:
  - magnetic media, used to hold backup data for reasonably short periods - perhaps a few days to a few weeks
  - optical or magneto-optical media, used to hold backup data for extended periods - a few years
- Choice criterias (on a TCO basis):
  - cost per GB of storage (\$/GB)
  - cost per unit of bandwidth (\$/MB/s)
  - amortizing the acquisition price of the hardware (robots and any computer system(s) necessary for running the robot)
  - hardware maintenance costs
  - cartridge price and cartridge replacement costs
  - the cost of backup management software for the servers along with the price of any agents needing to be installed on various client systems
  - software maintenance cost
  - the cost of personnel necessary to implement the solution
  - the cost of keeping the system operational
  - the cost of using the system - for example, the cost of repetitive verification tasks, such as checking that the overnight backup worked properly - and the cost of export/import of cartridges between the home site and a backup site, or a cartridge storage site.
  - ...

Page 72

© R.J Cheavance

## Technologies Supporting Backed-up Data(2)

### ■ Comparison of various Magnetic Cartridge technologies (Source [QUA03])

	SAIT-1	SDLT320	Ultrium LTO (Generation 2)
Capacity (native)	500 GB	160 GB	200 GB
Capacity (compressed)	1.3 TB	320 GB	400 GB
Transfer Rate (native)	30 MB/s	16 MB/s	30-35 MB/s
Transfer Rate (compressed)	72 MB/s	32 MB/s	60-70 MB/s
MTBF at 100% duty cycle	>300000 hours	250000 hours	250000 hours
Media Formulation	AME	MP	MP
Media Form Factor	Single reel, half-inch cartridge	Single reel, half-inch cartridge	Single reel, half-inch cartridge
Media length	600 m	600 m	600 m
File Access Time	70 s	70 s	52 s
Media Load Time	23 s	40 s	15 s
Memory in Cassette	Yes	No	Yes
WORM Capable	Yes	No	No
Published Roadmap # of Generations		4	4
Roadmap Maximum Capacity (Native)	10 TB	1.2 TB	1.6 TB
Roadmap Maximum Performance (Native)	240 MB/s	100 MB/s	160 MB/s
Multiple Manufacturing Sources for Drives and Media	Yes	No	No

Page 73

© R.J Cheavance

## RAIT and Tape Virtualization

### ■ RAIT (Redundant Array of Independent Tapes)

- Organization similar to RAID

### ■ Tape Virtualization

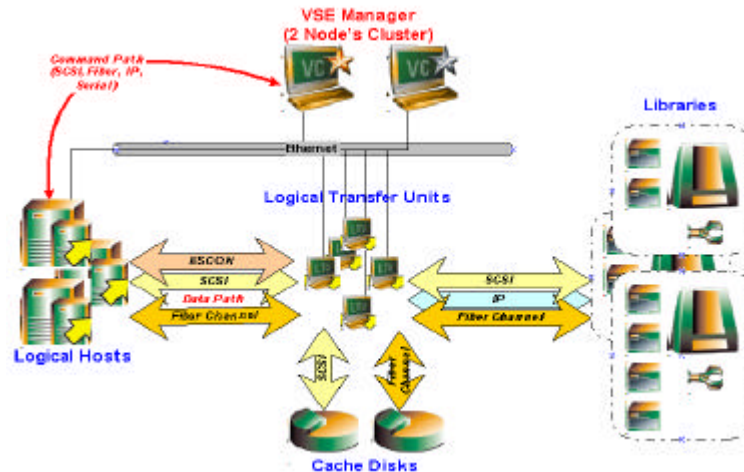
- Example Neartek's VSengine
- Product functionality:
  - independent of any particular backup peripheral
  - supports simultaneous, sharable backups for heterogeneous hosts
  - centralized management of resources and of their sharing
  - fast data backup and restore thanks to a RAID cache
  - the ability to provide multiple volumes on a single cartridge (Dynamic Volume Stacking)
  - automatic volume duplication
  - compatible with the most widely-used backup software, such as Netbackup, ArcserveIT, TSM, Networker...

Page 74

© R.J Cheavance

## RAIT and Tape Virtualization(2)

### ■ VSEngine Architecture (source Neartek)



Page 75

© R.J Cheavance

## Data Archiving

- Life time to backed up data is usually limited to few weeks or months
- There are needs to keep data for longer periods (e.g. government regulations,...). This is the purpose of Data Archiving.
- Since archived data is kept unchanged, the term Reference Information is being used to refer to such data.
- Summary of the Characteristics of Several Archiving Technologies

Technology	Capacity (GB)	Longevity (years)	Use
12" WORM	30	50 to 100	Very long-term archiving
5.25" Magneto-optic	9	30	Fast access to archived material
CD and DVD 120 mm	3	30	Low cost archiving
Tape	200+	10	High capacity, short lifespan

Page 76

© R.J Cheavance

## Storage of Reference Information

- Several systems have been proposed for storage and management of Reference Information
- Comparison of the Characteristics of SAN, NAS and CAS (Content Addressable Storage, EMC term for Reference Information) (Source EMC)

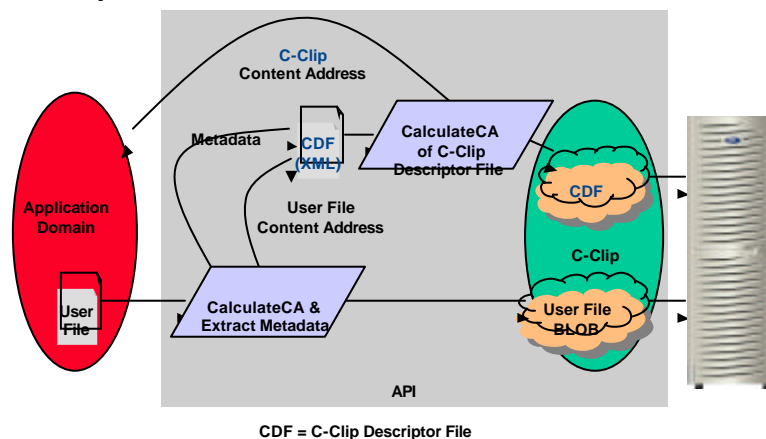
	SAN	NAS	CAS
Typical application	OLTP Decision support	CAD/CAM Collaborative work	Management of contents
Nature of information	Modifiable		Fixed
Technological Difficulty	Performance	Modifiable file sharing	Scalability Longevity
Access method	By address		By contents
Type of data stored	Volume	File	Object with metadata

Page 77

© R.J Chevance

## Storage of Reference Information(2)

- Overview of Centera Functionality (after EMC)

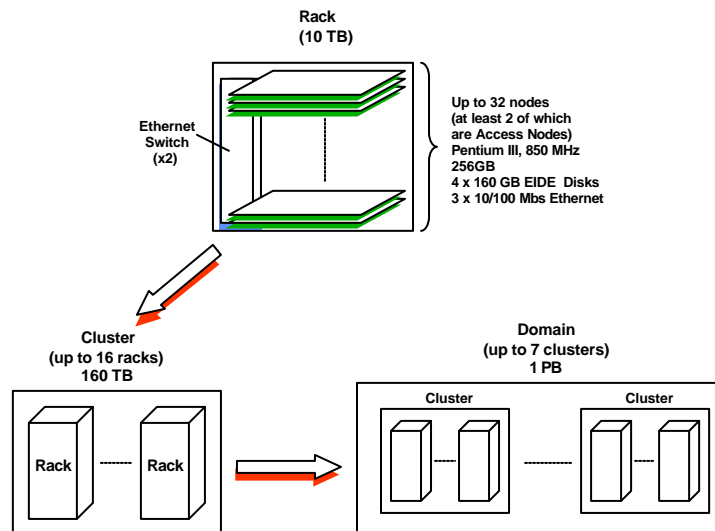


Page 78

© R.J Chevance

## Storage of Reference Information(3)

### ■ Centera Architecture (after EMC)



Page 79

© R.J Cheavance

## References

- [BMC01] BMC Software « The Case for Host Processor Compression » White Paper 02/2001 [www.bmc.com](http://www.bmc.com).
- [HAM02] Tom Hammond-Doel "Changing from shared bandwidth to SBOD" Storage Networking World Online April 29th, 2002 <http://www.snwonline.com/>
- [LYM03] Peter Lyman, and Varian Hal R. "How Much Information", 2003. <http://www.sims.berkeley.edu/how-much-info-2003> (6/2004)
- [QUA03] Qualstar 2003 White Paper "New Tape Technologies Offer More Data on Less Tape" available on [www.qualstar.com](http://www.qualstar.com)

Page 80

© R.J Cheavance