



Towards Invisible Storage

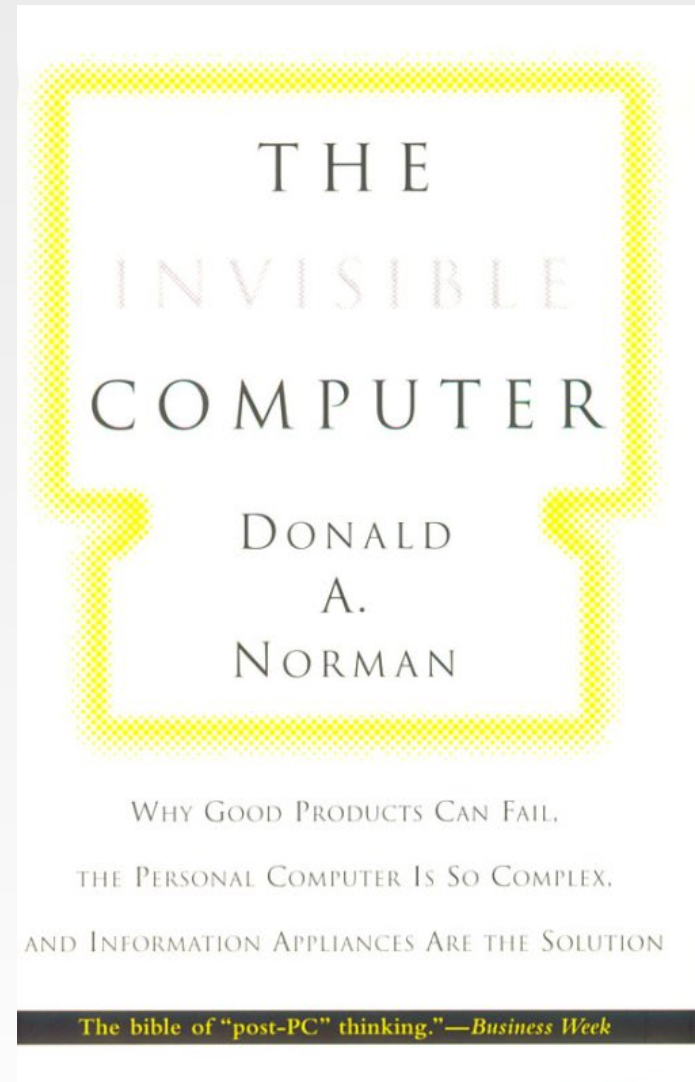
Alain Azagury
IBM XIV Business Executive



Acknowledgement

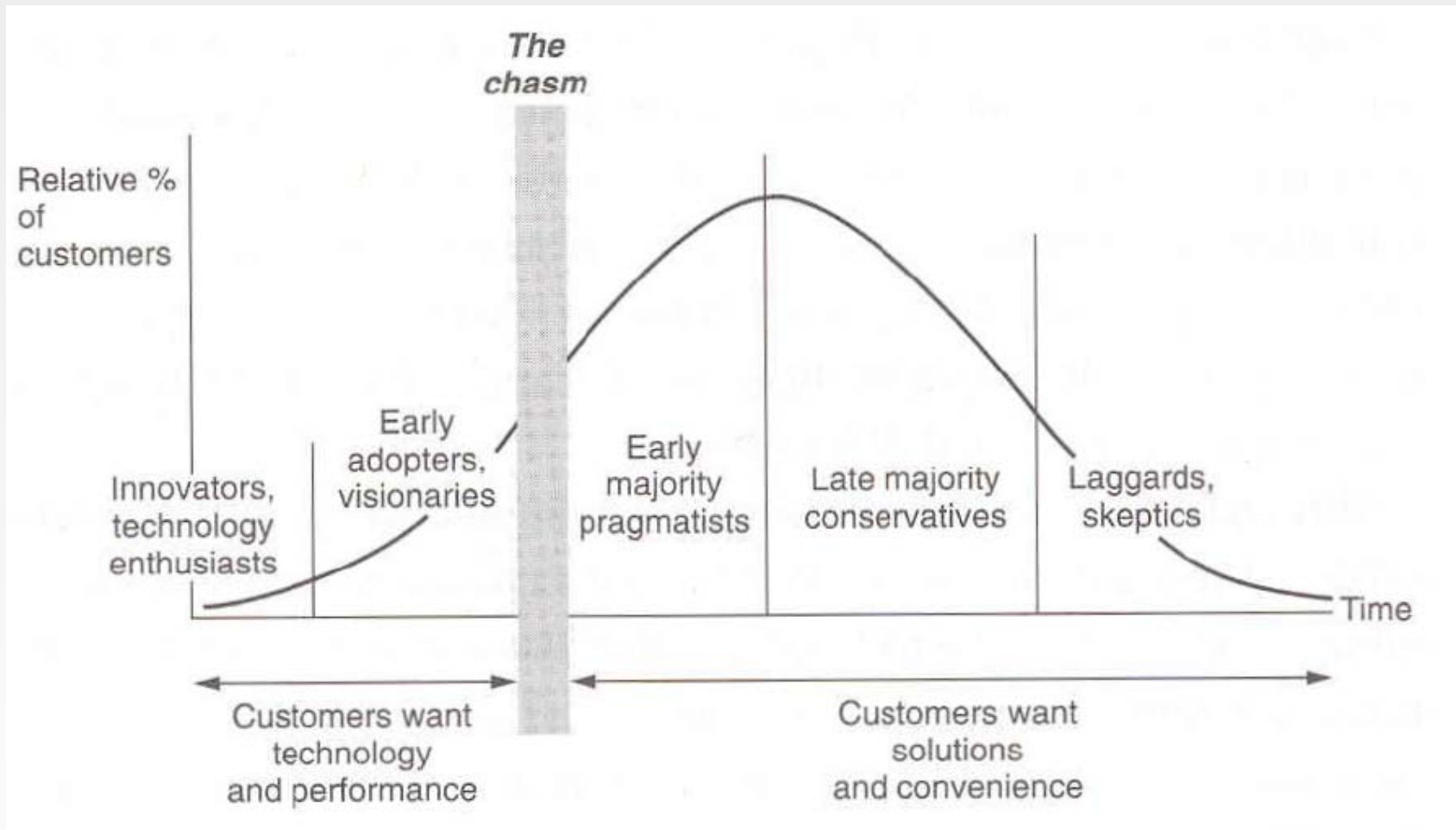
Influenced by Donald A. Norman's book
"The Invisible Computer"

Why Good Products Can Fail,
The Personal Computer is so Complex
And Information Appliances are the Solution



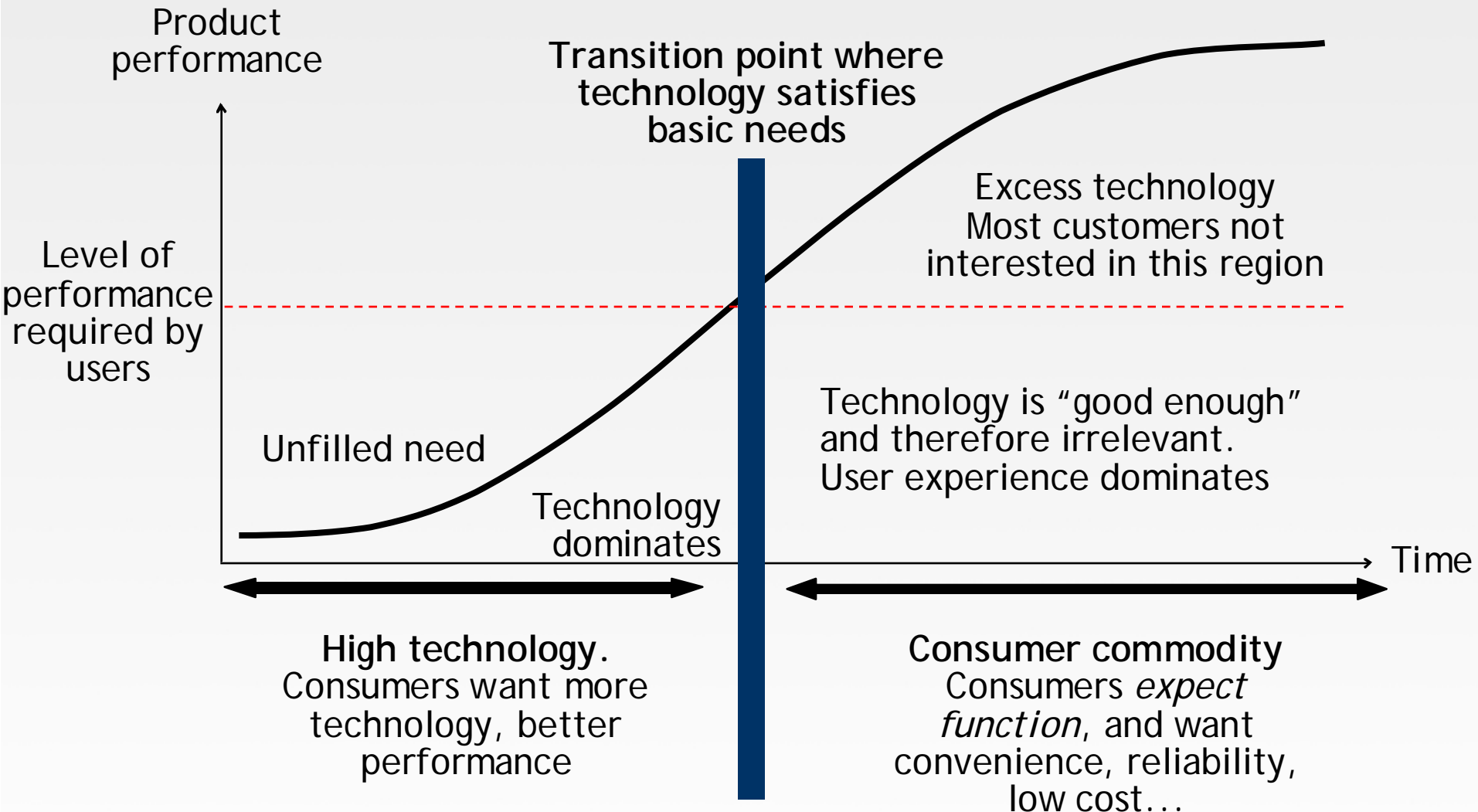
The Change in Customers as a Technology Matures

From Norman, 1998 (modified from Moore, 1995)

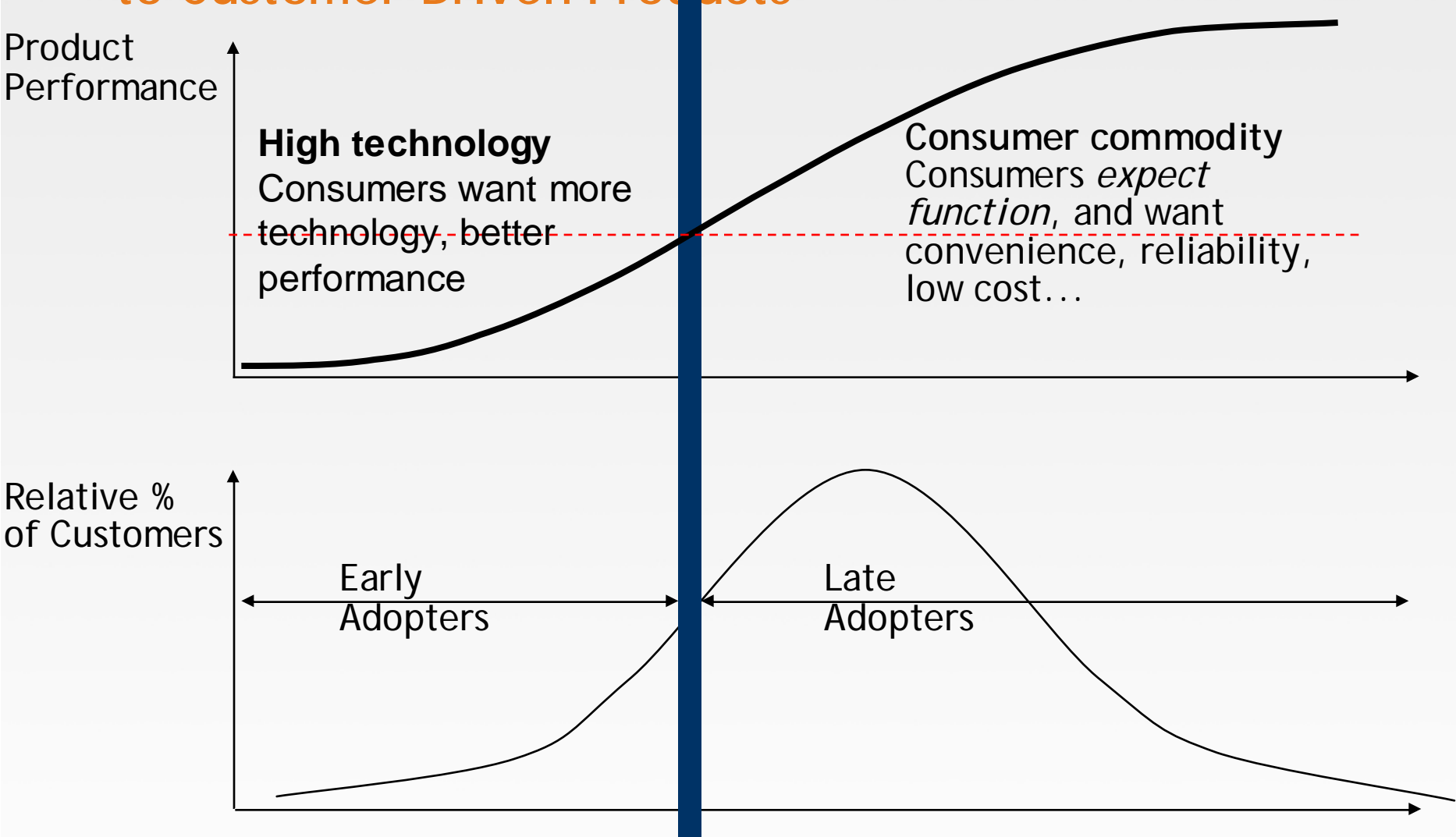


The Needs-Satisfaction Curve of a Technology

From Norman, 1998 (modified from Christensen, 1997)

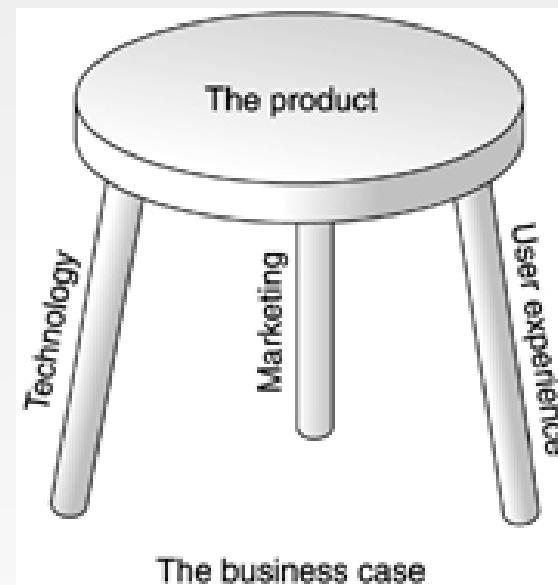


The Change from Technology-Driven Products to Customer-Driven Products



From Personal Computers to Enterprise Storage

- Enterprise Storage is crossing the chasm... and Storage consumers *expect function, and want convenience, reliability, low cost...*
- Enterprise storage users are more sophisticated than today's average Personal Computer User
 - Therefore, they were able to trade-off “lack of convenience” for technology features
- However, maturity of IT (and the current financial crisis) are forcing the Enterprise Storage market to cross the chasm



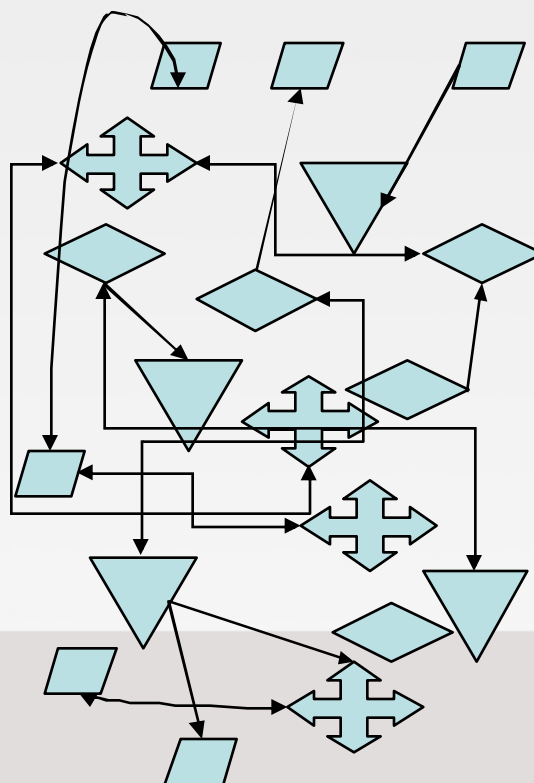
The balanced product stool.
Source: Norman 1998.

What's, why and where to Reinvent?

Data

0101001011101010111
0110101010111010101
0110101100110101011
0101110101010101011
0101110110101010101
0101010101010111101
0101010111010101010
1110101101011101010
1010111010101011110
1010101011011101010
1010111111101010111
0101101010011101010
0010101010100010101

Architecture



Disks



The Need - DATA

Information Explosion Creates Storage Challenges

How much data does mankind store?

- IDC says about 161 exabytes in 2006
- By 2010, we'll reach 988 exabytes
- That's 600% growth in 4 years

161,000 PB

**We must provide a simple solution
for the storage needs of the
modern enterprise**

988,000 PB

The Media - DISKS

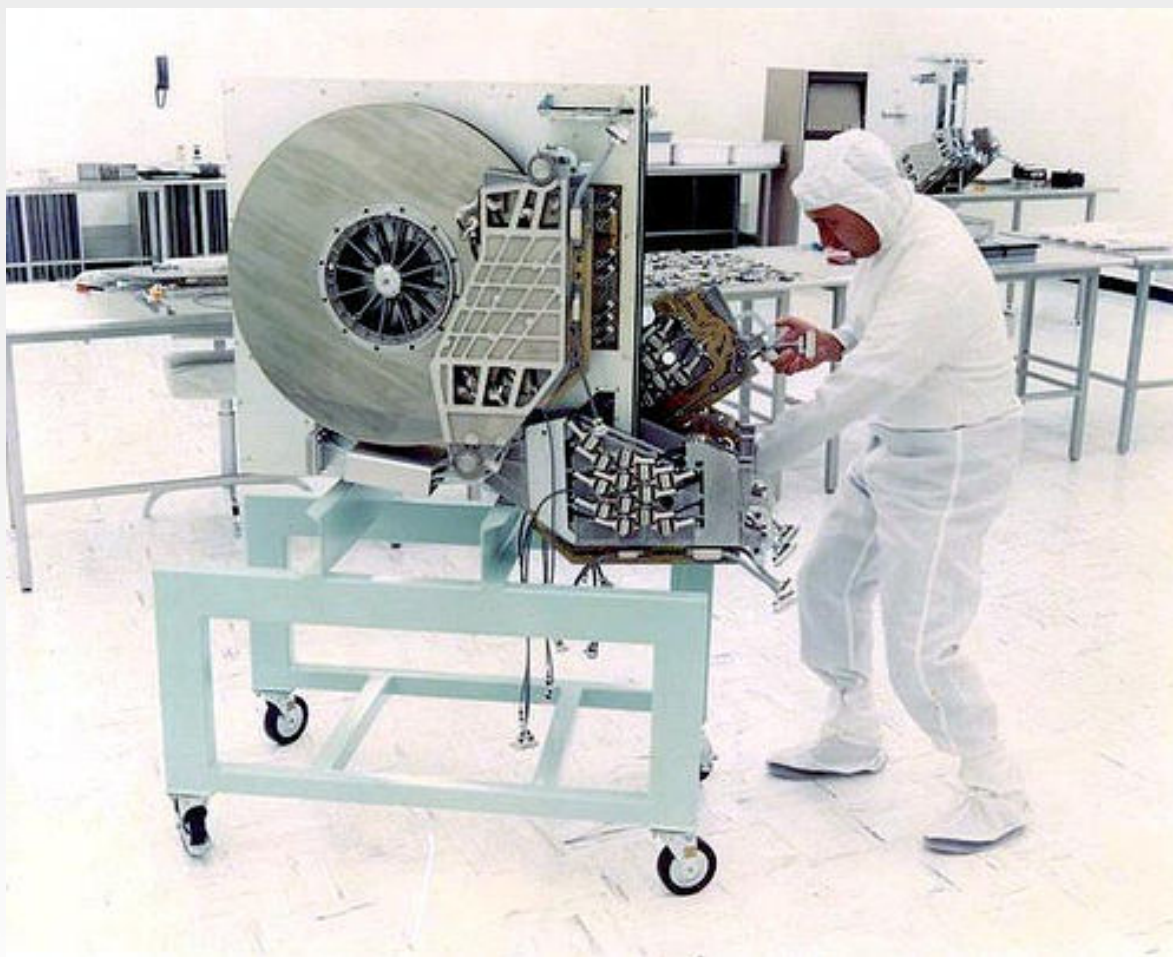
Modern Data Storage

- **Magnetic Tape** - IBM pioneered the magnetic tape in 1952, realizing that both punch cards and ticker tape were far too slow
- **Magnetic Disk** - In 1956 a small team of IBM engineers in San Jose introduced the first computer disk storage system. The 305 RAMAC could store five megabytes of data on 50 disks, each 24 inches in diameter.

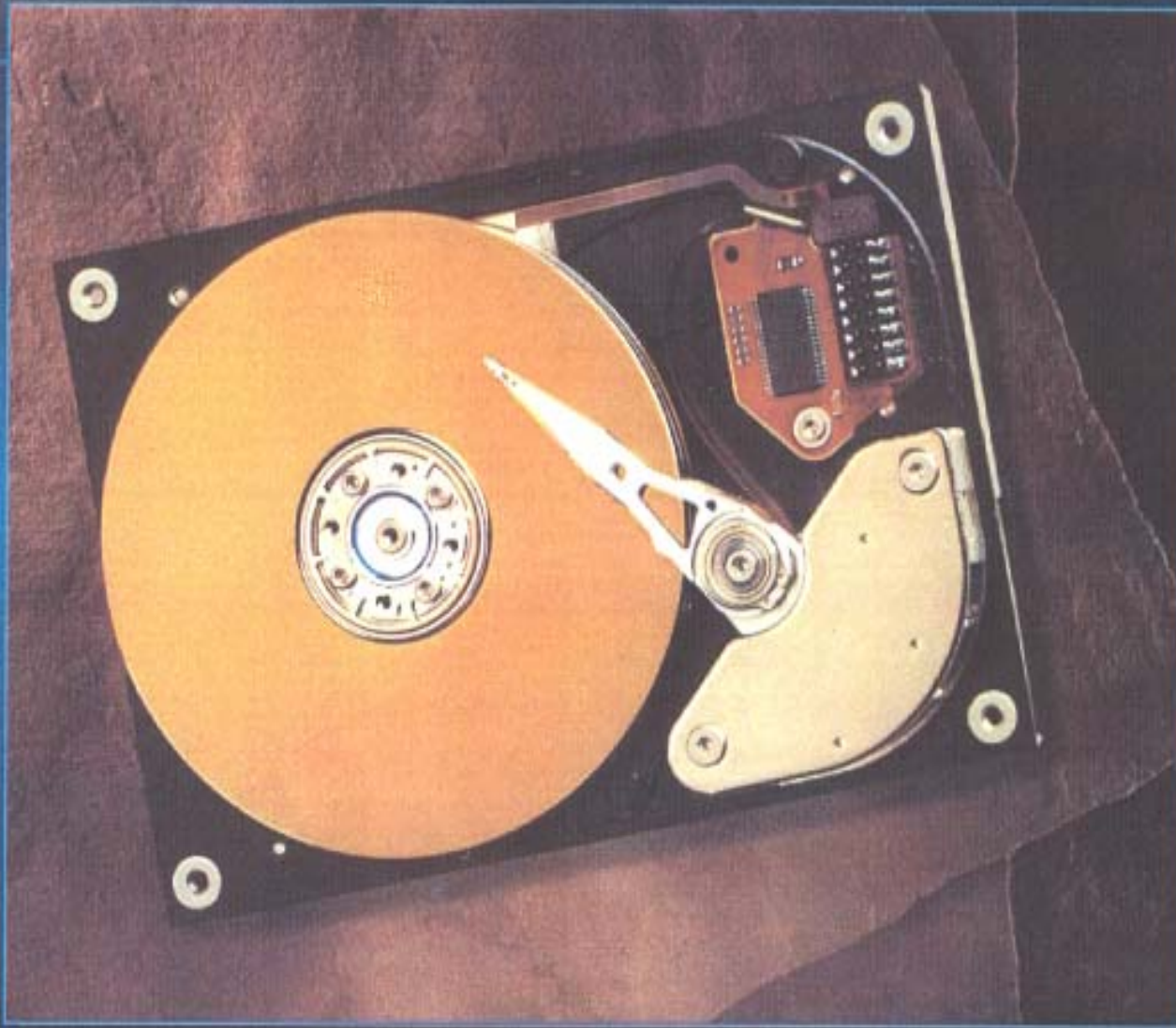
Evolution of Storage Media



Hard Drives in 1975



The 1996 Ewing Lecture



IBM

Memor - 12/1/96

Seagate

The means - Architecture

Key Attributes for Enterprise Storage Solutions

(Remember the key Needs-Satisfaction attributes of Consumer Commodity)

- **Reliability** - Business data more critical than ever, with no tolerance for downtime for most applications - requirements now greater than 5 nines
- **Convenience**
 - **Performance** - Consistent performance under all conditions, eliminating hot spots and staying consistent during rebuilds after hardware failures
 - **Manageability** - Total system virtualization with emphasis on ease of use
- **Cost** - Reasonable cost so business can concentrate its efforts on its core business and not on IT
- **Functionality** - Tier 1 functions (e.g. replication, thin provisioning) that scale with no performance penalty and are inherently built-in to the architecture

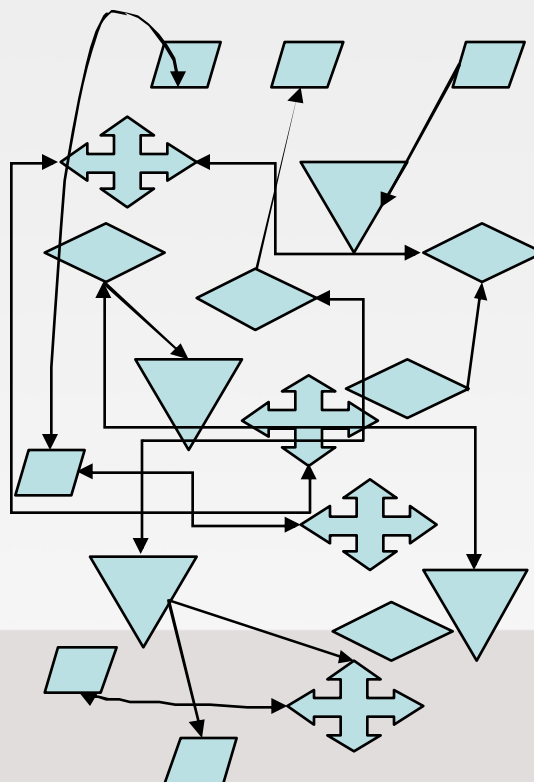
All of these key attributes -- with unlimited scalability

What's, why and where to Reinvent?

Data

0101001011101010111
0110101010111010101
0110101100110101011
0101110101010101011
0101110110101010101
0101010101010111101
0101010111010101010
1110101101011101010
1010111010101011110
1010101011011101010
1010111111101010111
0101101010011101010
0010101010100010101

Architecture



Disks

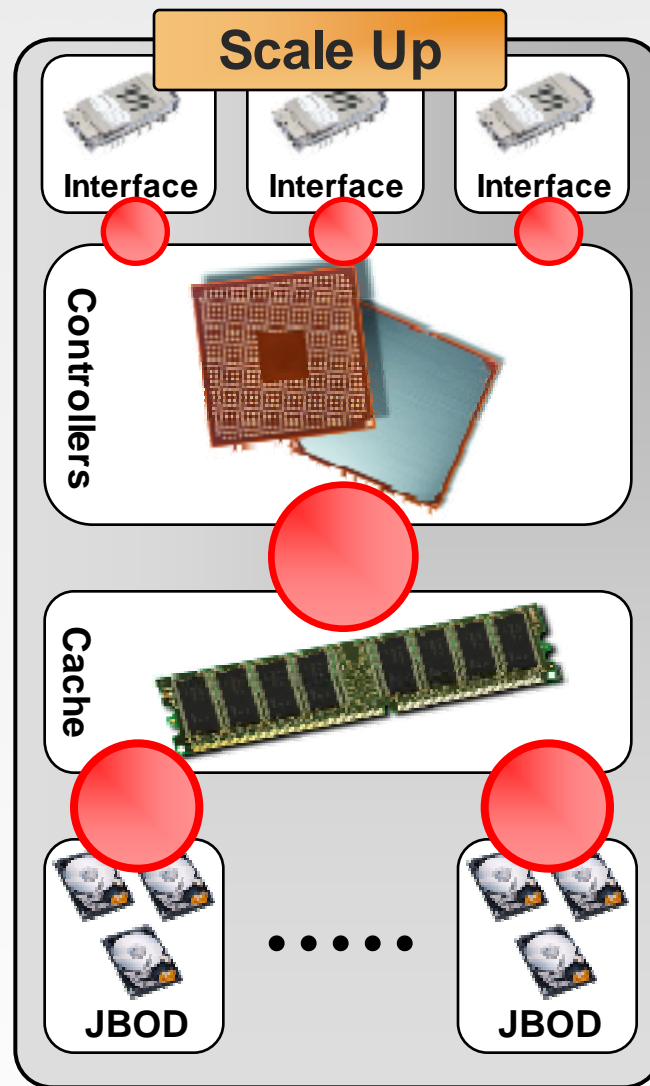


Current Enterprise Storage Solutions

Building blocks:

- Disks
- Cache
- Controllers
- Interfaces
- Interconnects

With the current architecture, scalability is achieved by using more powerful (and more expensive) components



Current Enterprise Storage Solutions

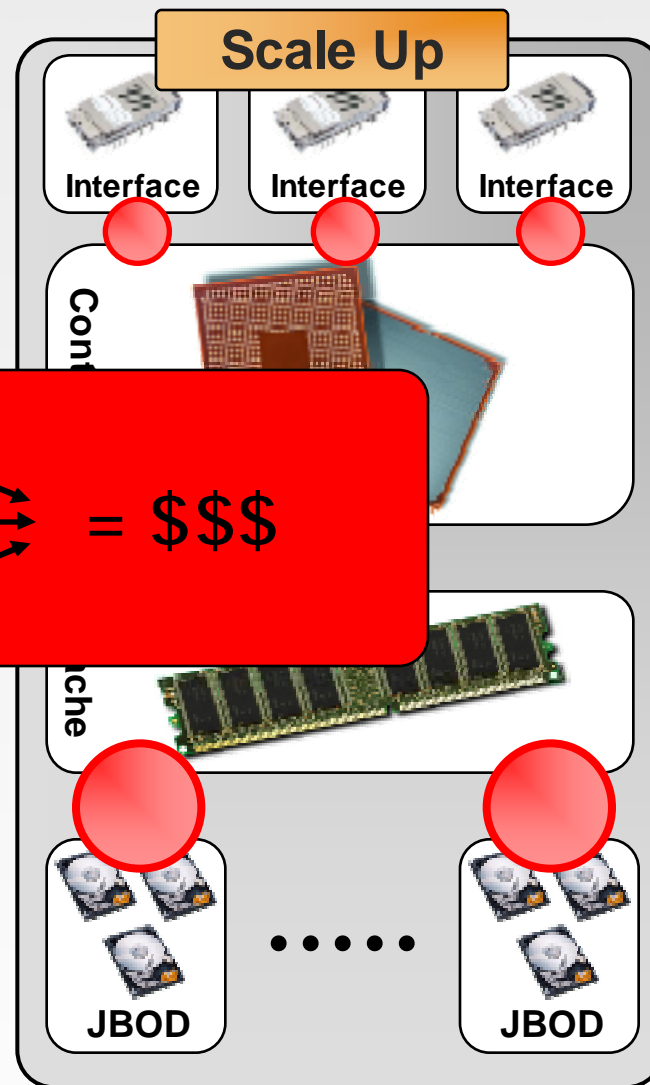
Building blocks:

- Disks
- Cache
- Controllers
- Interface
- Interconnect

PERFORMANCE
RELIABILITY
SCALABILITY

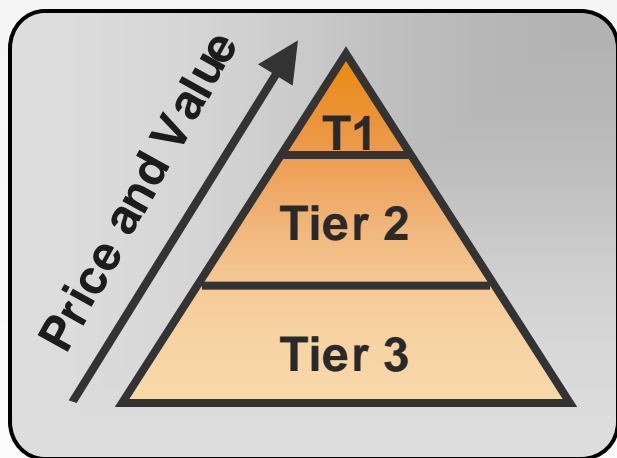
= \$\$\$

With the current architecture, scalability is achieved by using more powerful (and more expensive) components



Available Solutions Add Cost and Complexity: Creating the Need for Information Lifecycle Management

- ILM attempts to cope with storage pains via multi-tiered storage
 - Tiered storage management and data classification are costly and complex
 - Excessive data movements create reliability and performance issues
 - Utilization rates remain low (50% or less), with limited ability to execute thin provisioning



Imagine prioritizing electricity at home...



Laundry Power?



Lamp Power?



TV Power?

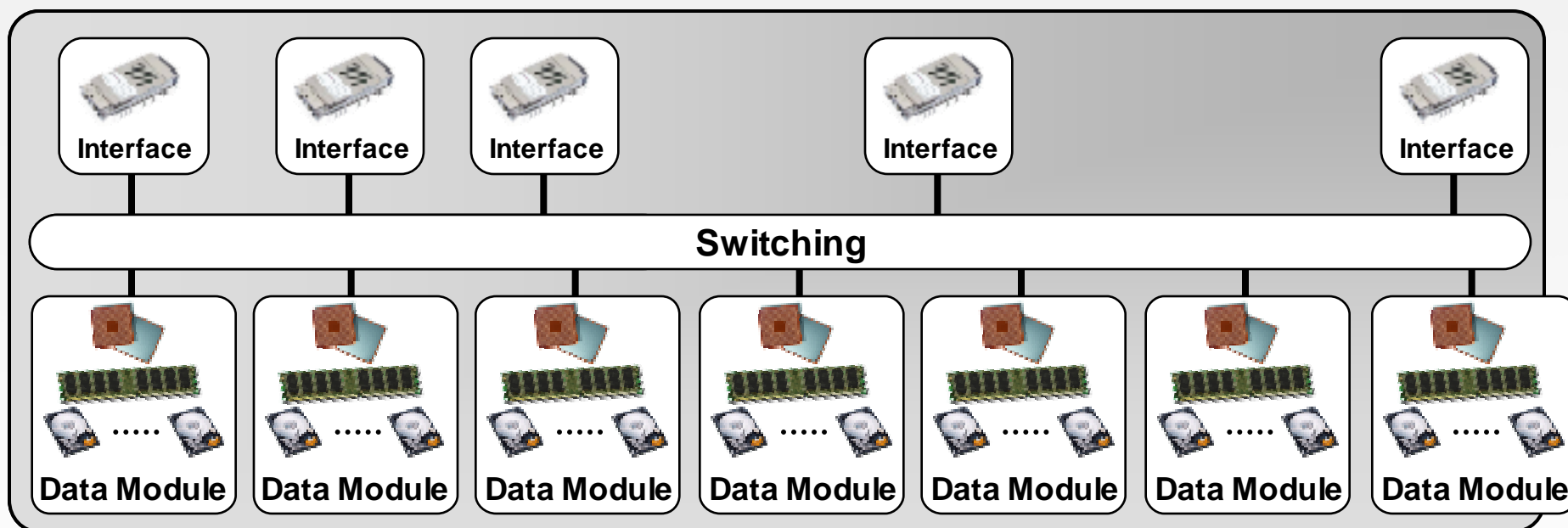
The Next Generation Architecture

The Next Generation Architecture

Design principles:

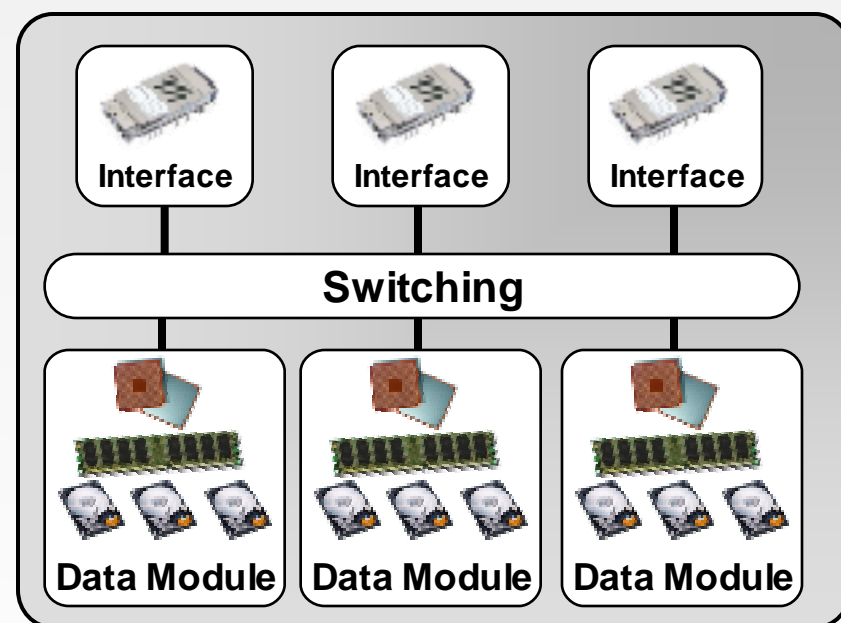
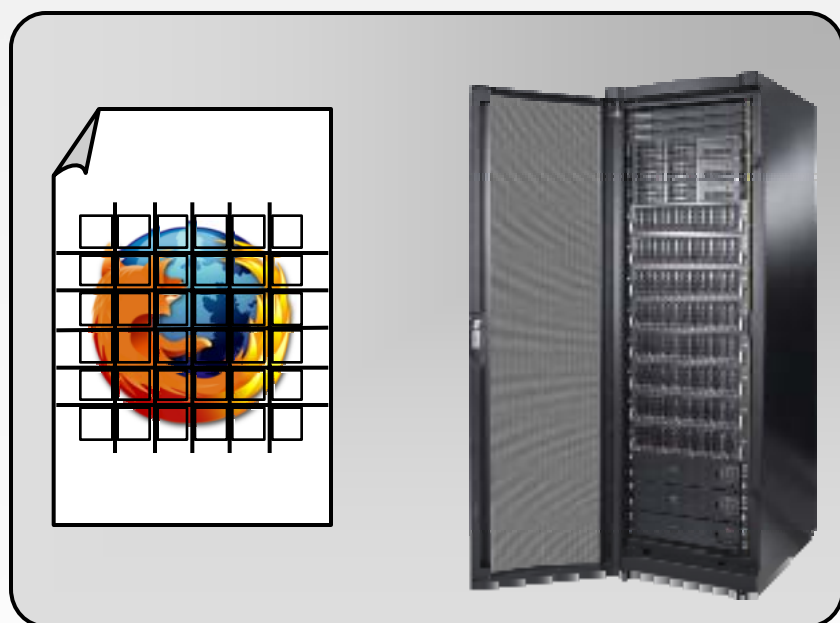
- Massive parallelism
- Granular distribution
- Off-the-shelf components
- Coupled disk, RAM and CPU
- User simplicity

Scale Out



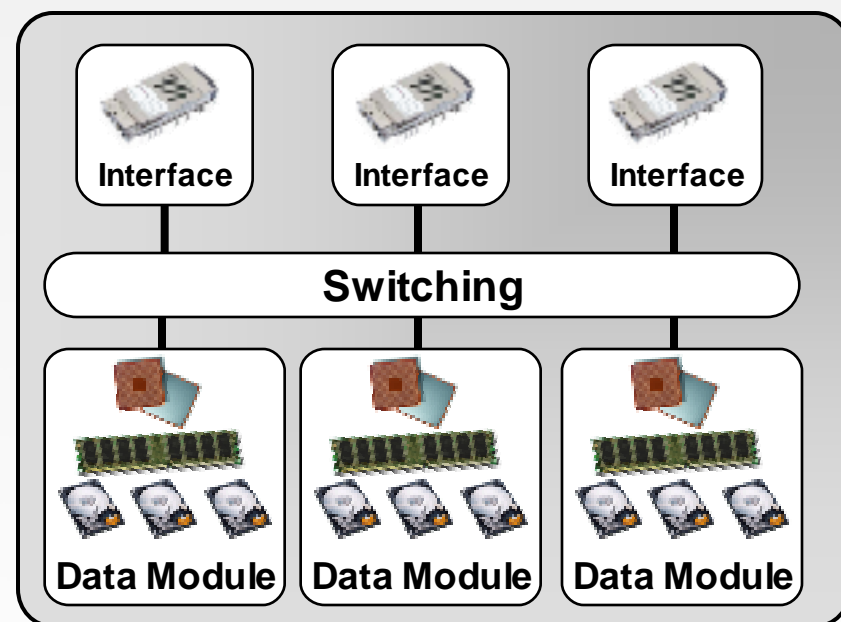
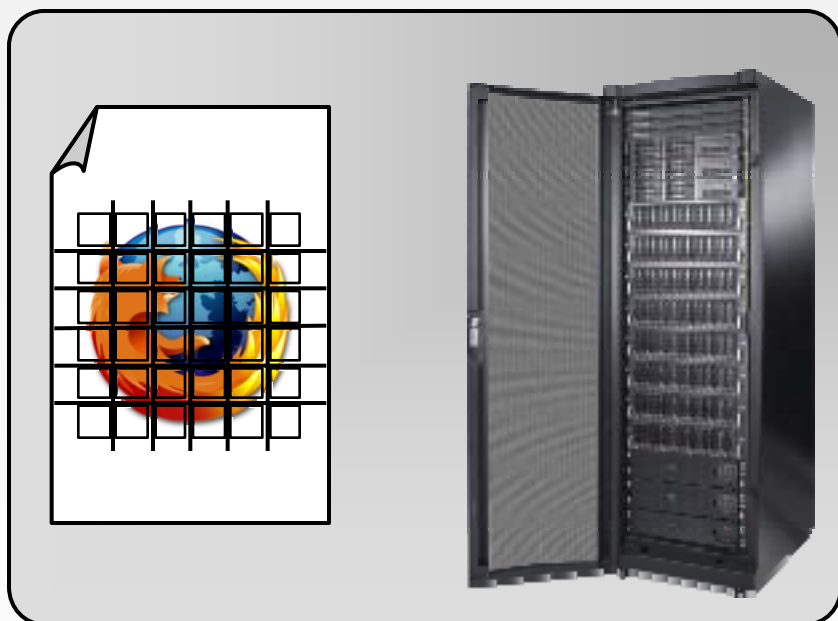
System Distribution Algorithm

- Each volume is spread across all drives
- Data is “cut” into 1MB “partitions” and stored on the disks
- XIV’s distribution algorithm automatically distributes partitions across all disks in the system pseudo-randomly



System Distribution Algorithm

- Each volume is spread across all drives
- Data is “cut” into 1MB “partitions” and stored on the disks
- XIV’s distribution algorithm automatically distributes partitions across all disks in the system pseudo-randomly

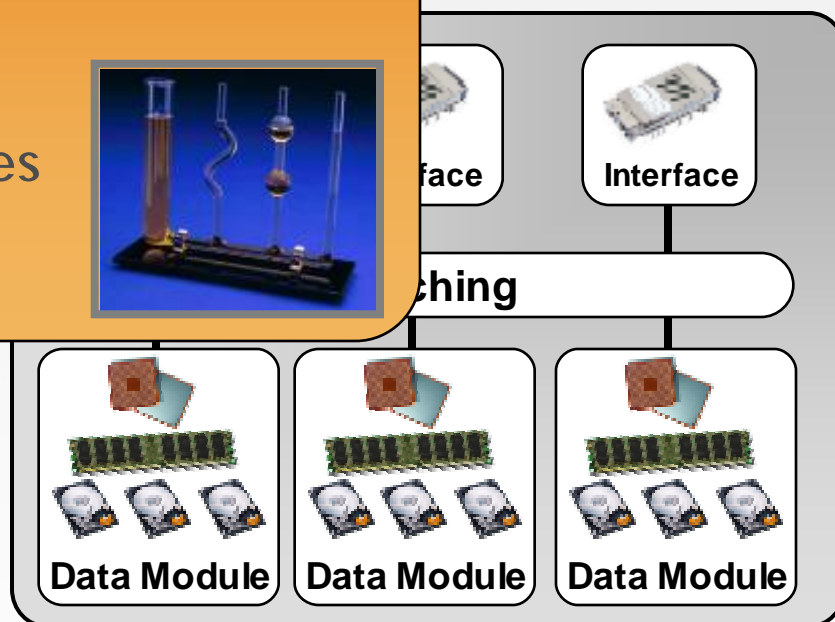


System Distribution Algorithm

- Each volume is spread across all drives
- Data is "c" on the disks
- XIV's distribution algorithm aims for constant disk equilibrium randomly

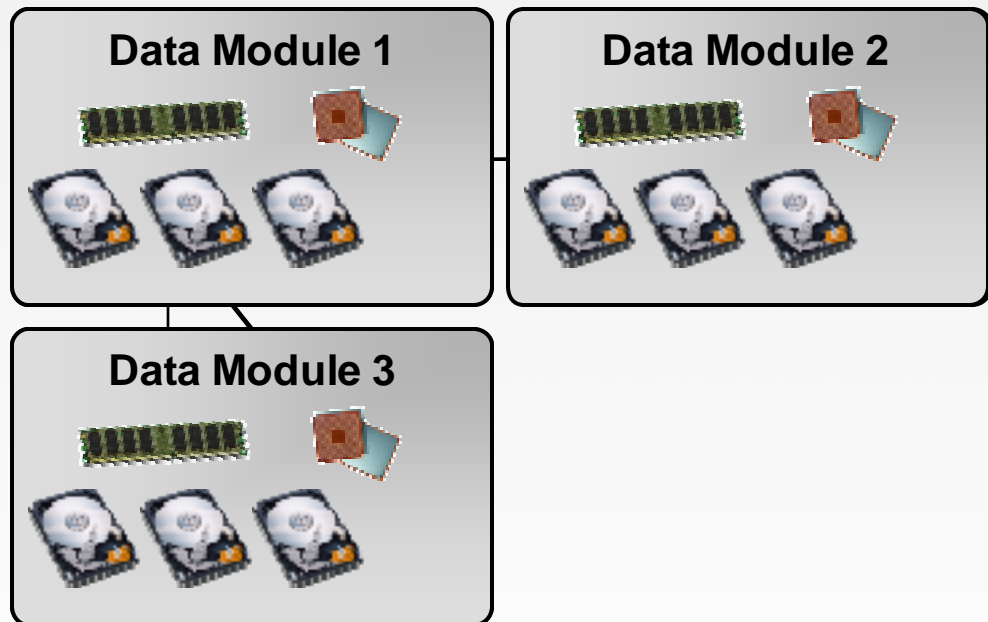
XIV disks behave like connected vessels, as the distribution algorithm aims for constant disk equilibrium.

Thus, XIV's overall disk usage approaches 100% in all usage scenarios.



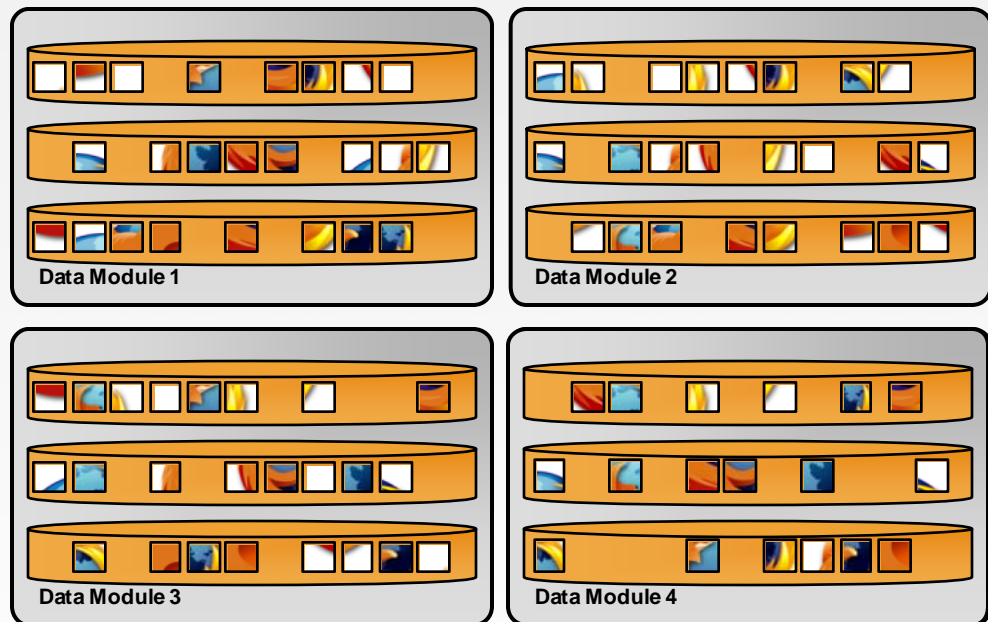
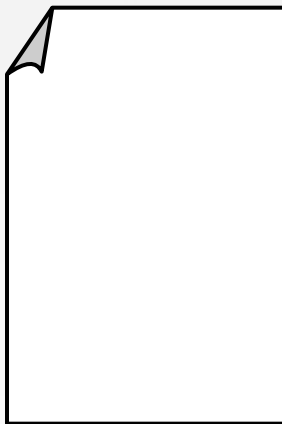
Distribution Algorithm on System Changes

- Data distribution only changes when the system changes
 - Equilibrium is kept when new hardware is added
 - Equilibrium is kept when old hardware is removed
 - Equilibrium is kept after a hardware failure



Distribution Algorithm on System Changes

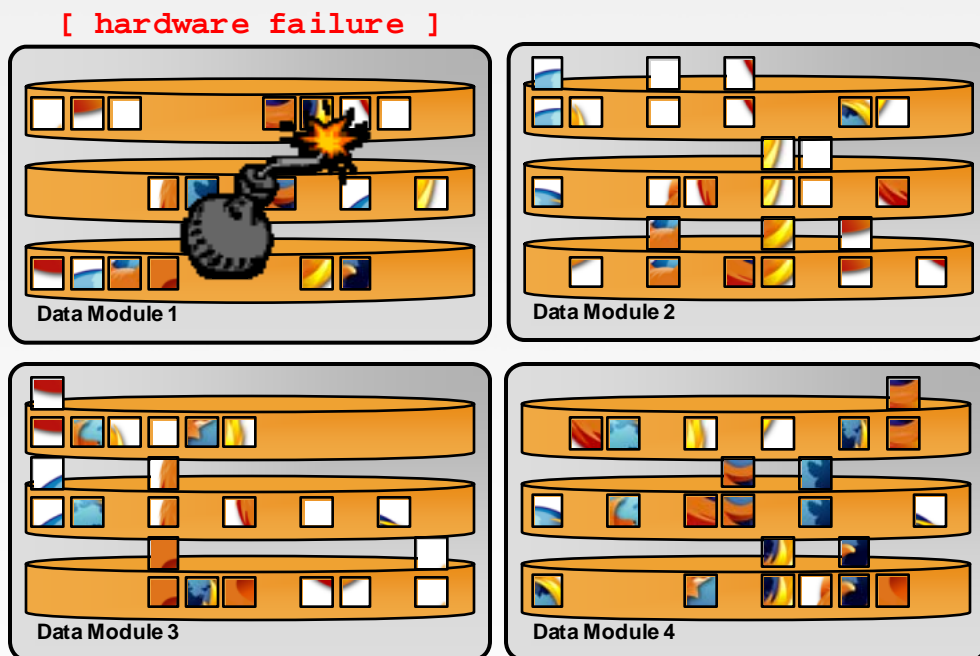
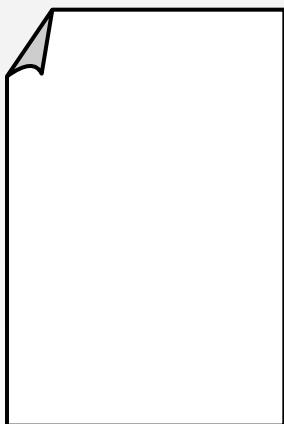
- Data distribution only changes when the system changes
 - Equilibrium is kept when new hardware is added
 - Equilibrium is kept when old hardware is removed
 - Equilibrium is kept after a hardware failure



[hardware upgrade]

XIV Distribution Algorithm on System Changes

- Data distribution only changes when the system changes
 - Equilibrium is kept when new hardware is added
 - Equilibrium is kept when old hardware is removed
 - Equilibrium is kept after a hardware failure

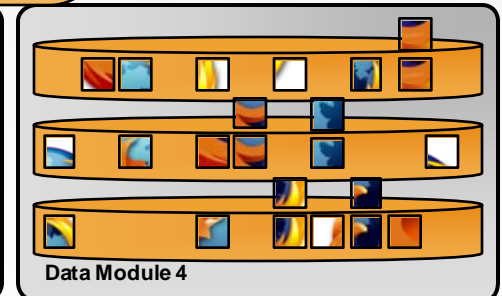
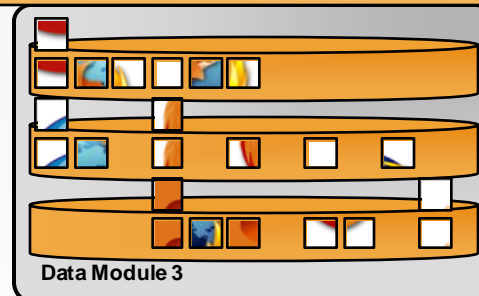
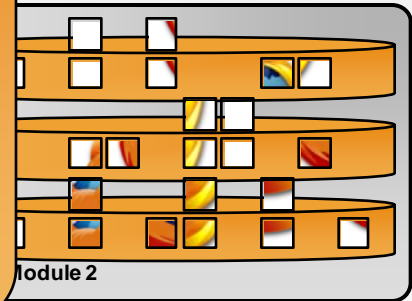


XIV Distribution Algorithm on System Changes

- Data distribution only changes when the system changes
 - Equilibrium is kept when new hardware is added
 - Equilibrium is maintained when hardware is removed
 - Equilibrium is maintained when hardware is reconfigured

The fact that distribution is full and automatic makes sure all spindles join the effort of data re-distribution after configuration change.

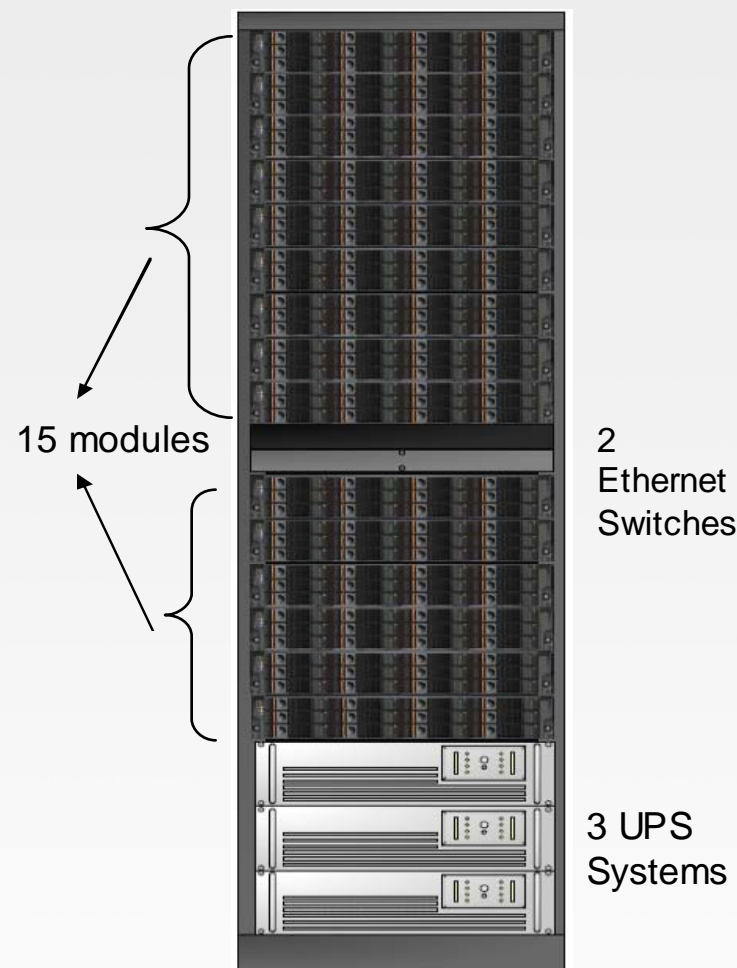
Tremendous performance gains are seen in recovery/optimization times thanks to this fact.



IBM XIV Storage System Hardware Platform

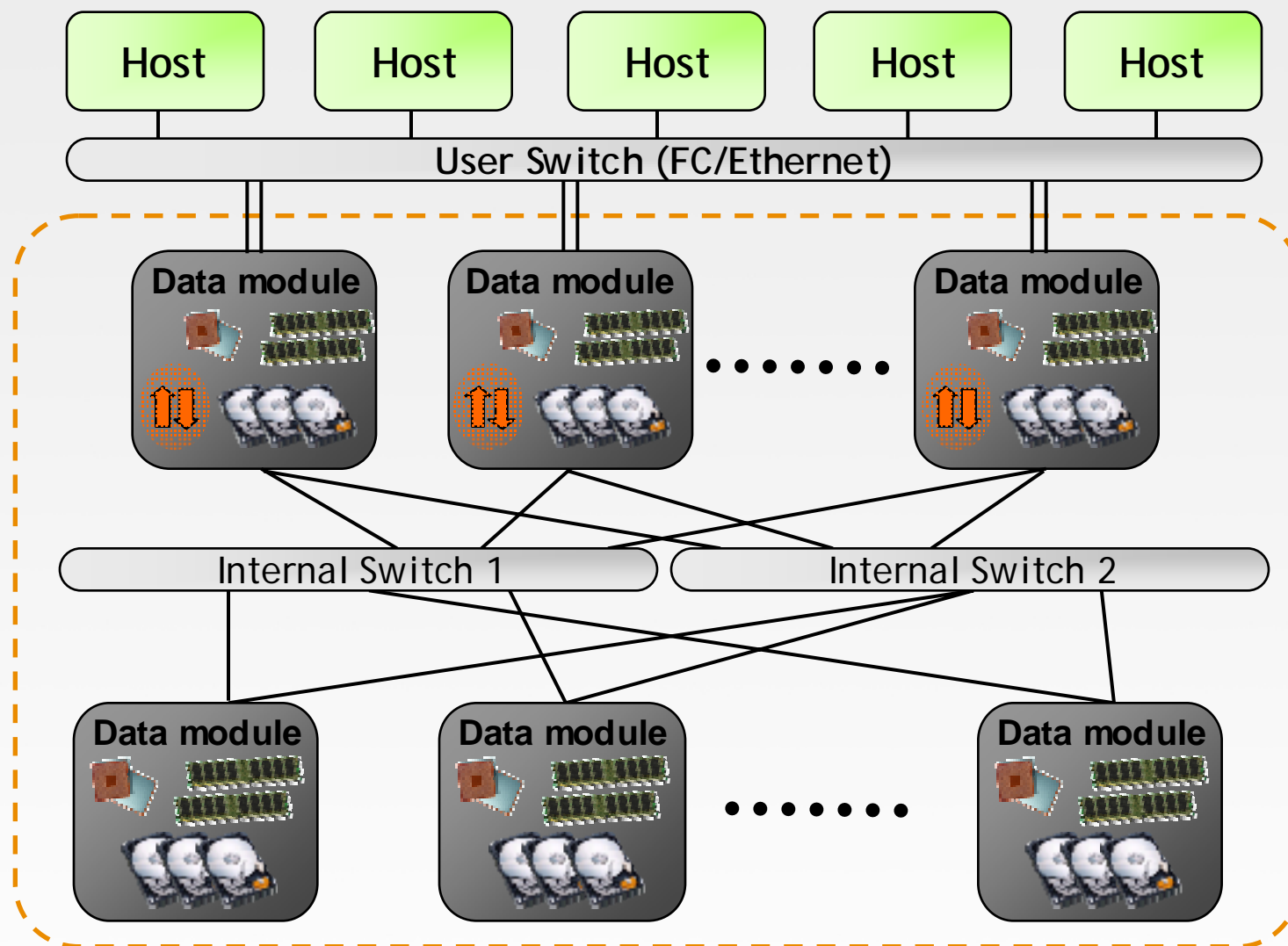
Machine Type: 2810-A14

- 180 disks per rack
 - 15 modules per rack
 - 12 disks per 2U module
 - 1TB 7200RPM SATA disk drives
- 80TB usable capacity for a single rack
- 120GB of system cache per rack (8GB per module)
- Up to 24 4GB FC host ports
- 6 1Gb iSCSI host ports
- 3 UPS systems



IBM XIV Storage

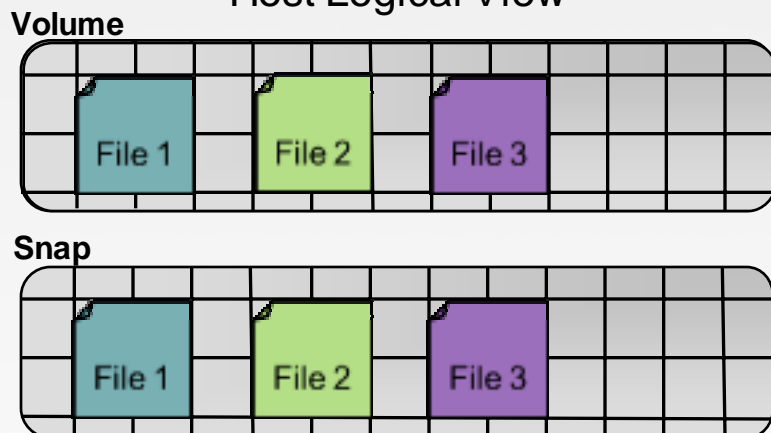
IBM XIV Storage System's Grid Architecture



SNAPs with No Limitations

- SNAPs creation/deletion is instantaneous
- High Performance WITH SNAPs
- Unlimited number of SNAPs

Host Logical View

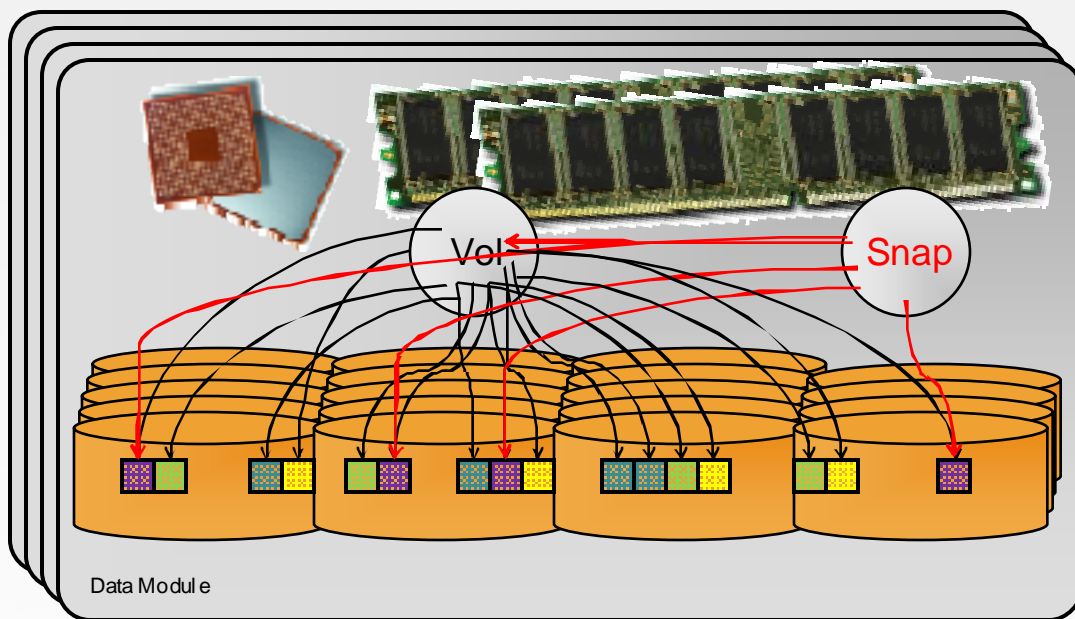


As Host Writes data, it is placed randomly across system in 1MB chunks

Each Server has pointers in memory to the disks that hold the data locally

On a SNAP, each Server simply points to original volume. Memory only Operation

XIV Physical View



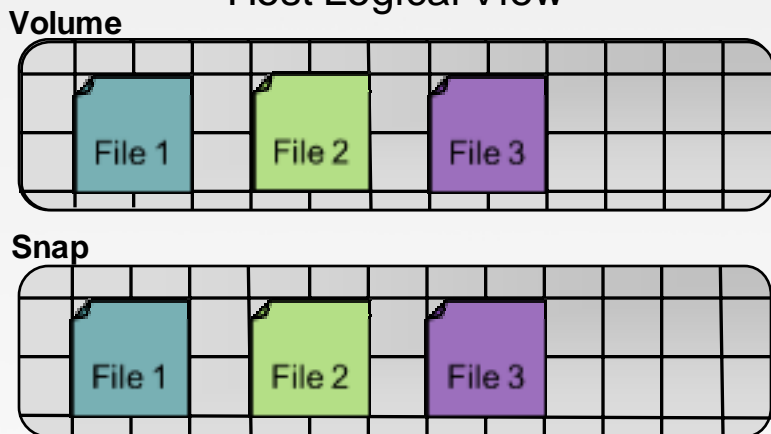
SNAPs with No Limitations

- SNAPs creation/deletion is instantaneous
- High Performance WITH SNAPs
- Unlimited number of SNAPs

Distributed SNAP on each Server.
Extremely fast memory operations

Accessing SNAPs is as fast as
accessing production volumes

Host Logical View

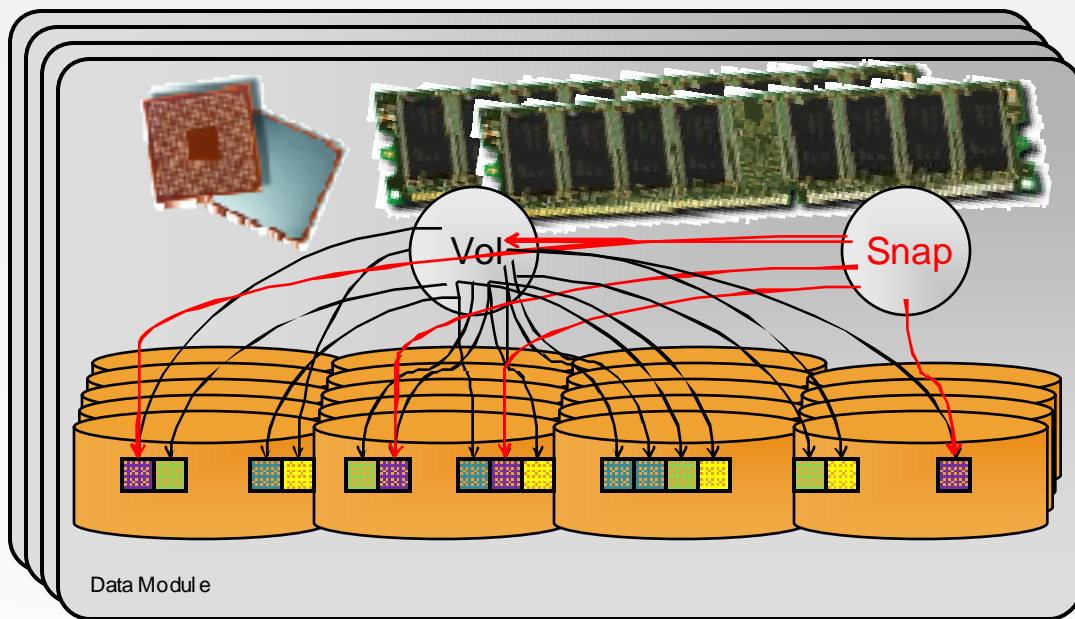


As Host Writes data, it is placed
randomly across system in 1MB chunks

Each Server has pointers in memory to
the disks that hold the data locally

On a SNAP, each Server simply points to
original volume. Memory only Operation

XIV Physical View



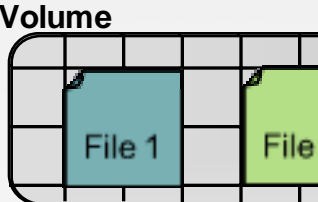
SNAPs with No Limitations

- SNAPs creation/deletion is instantaneous
- High Performance
- Unlimited

High Performance, Unlimited SNAPs provide:

- Easier Physical Backup to Tape
- Instant recovery from Logical Backup
- Easy creation of Test Environment
- Boot-from-SAN with easy rollback
- Easy Data-Mining on Production data

Volume



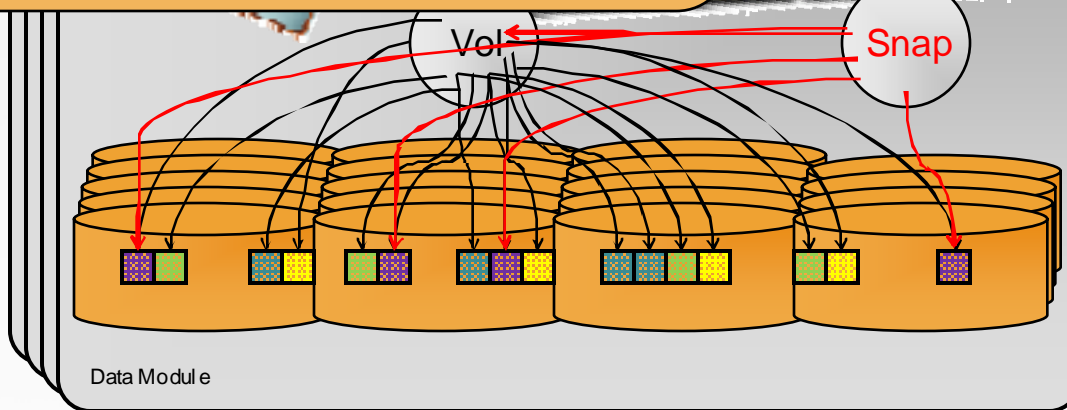
Snap



As Host Writes data, it is placed randomly across system in 1MB chunks

Each Server has pointers in memory to the disks that hold the data locally

On a SNAP, each Server simply points to original volume. Memory only Operation



... and More Tier-1 Functionality Built in to the Architecture

- **Thin Provisioning**
 - Installing physical capacity only if and when needed
- **Automatic Data Migration**
 - Online data migration from other Storage arrays with no down time, no host configuration and no administration effort
- **Remote Mirroring for Disaster Recovery**
 - Low granularity - any to any volume replication, with automatic Snap to keep copies self-consistent even during re-sync after link failure
- **And more...**

IBM XIV Storage Simple and Intuitive Management

- Intuitive GUI (Java based) with Script Generator
- No dedicated management station
- Command Line Interface (CLI)
- XML over SSL
- Event management (SNMP)
- Complete Event Logging
- Events notification via email, SNMP and SMS
- Role based management:
 - Storage Admin
 - Application Admin
 - Operator

IBM XIV Storage Simple Intuitive Management

example: Creating a Volume

The screenshot shows a 'Create Volumes' dialog box. At the top, there is a 'Select Pool' dropdown menu set to 'PriorityApps_0'. Below this, it states 'Total Capacity: 13400 GB of Pool: PriorityApps_0'. A progress bar shows three segments: a green segment labeled '3075', a yellow segment labeled '3367', and a grey segment labeled '6957'. Below the progress bar are three status indicators: a green dot labeled 'Allocated', a yellow dot labeled 'Total Volume(s) Size', and a grey dot labeled 'Free'. In the center, there is a form with three fields: 'Number of Volumes:' with a value of '1', 'Volume Size:' with a value of '3367' and a 'GB' unit dropdown, and 'Volume Name:' with a red asterisk and the value 'Email_Vol_1'. At the bottom are 'Create' and 'Cancel' buttons.

Category	Value
Allocated	3075
Total Volume(s) Size	3367
Free	6957

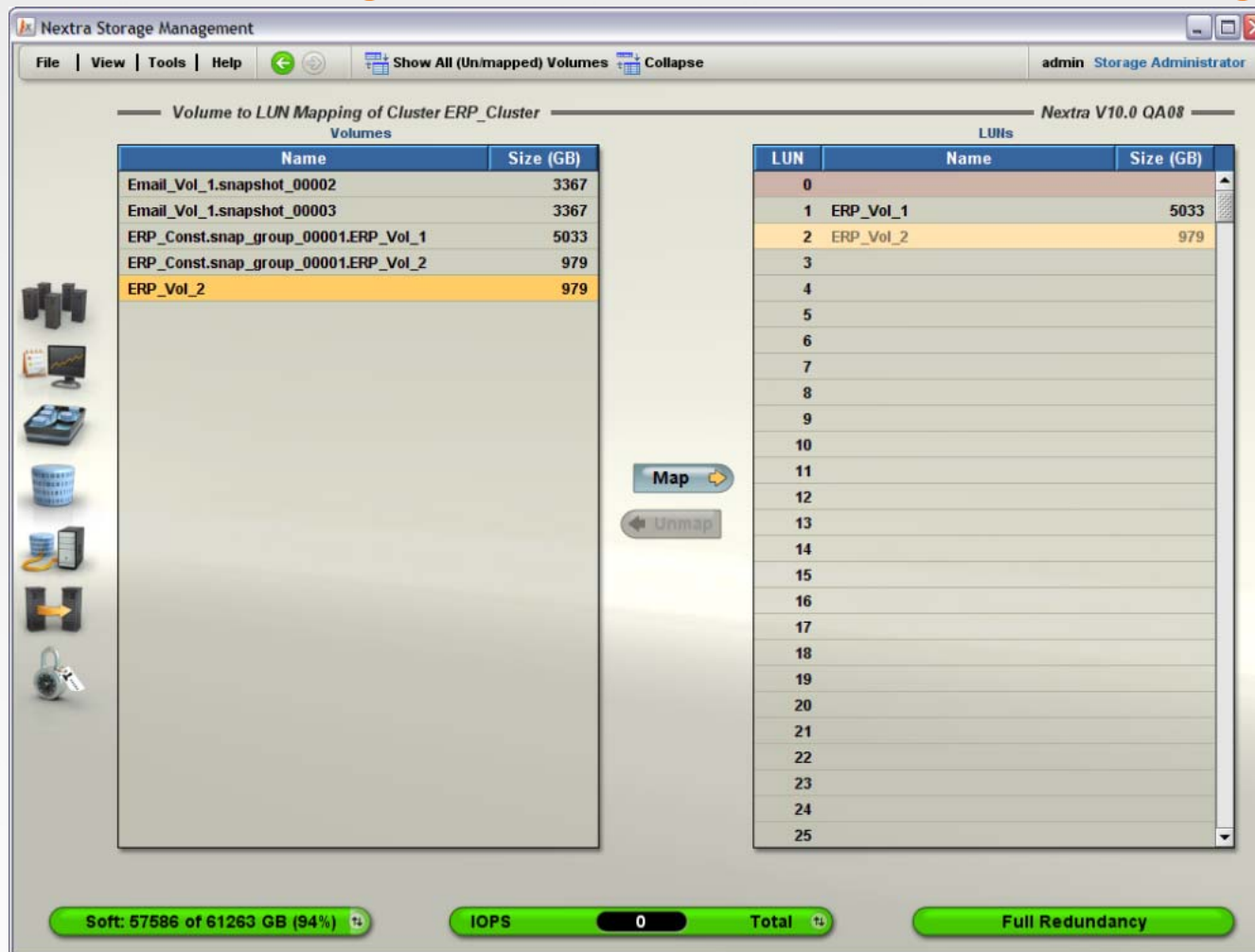
Number of Volumes: 1

Volume Size: 3367 GB

Volume Name: *Email_Vol_1

- Used capacity is always known !

IBM XIV Storage: Volume to LUN Mapping



Nextra Storage Management

File | View | Tools | Help | Show All (Unmapped) Volumes | Collapse | admin Storage Administrator

Volume to LUN Mapping of Cluster ERP_Cluster

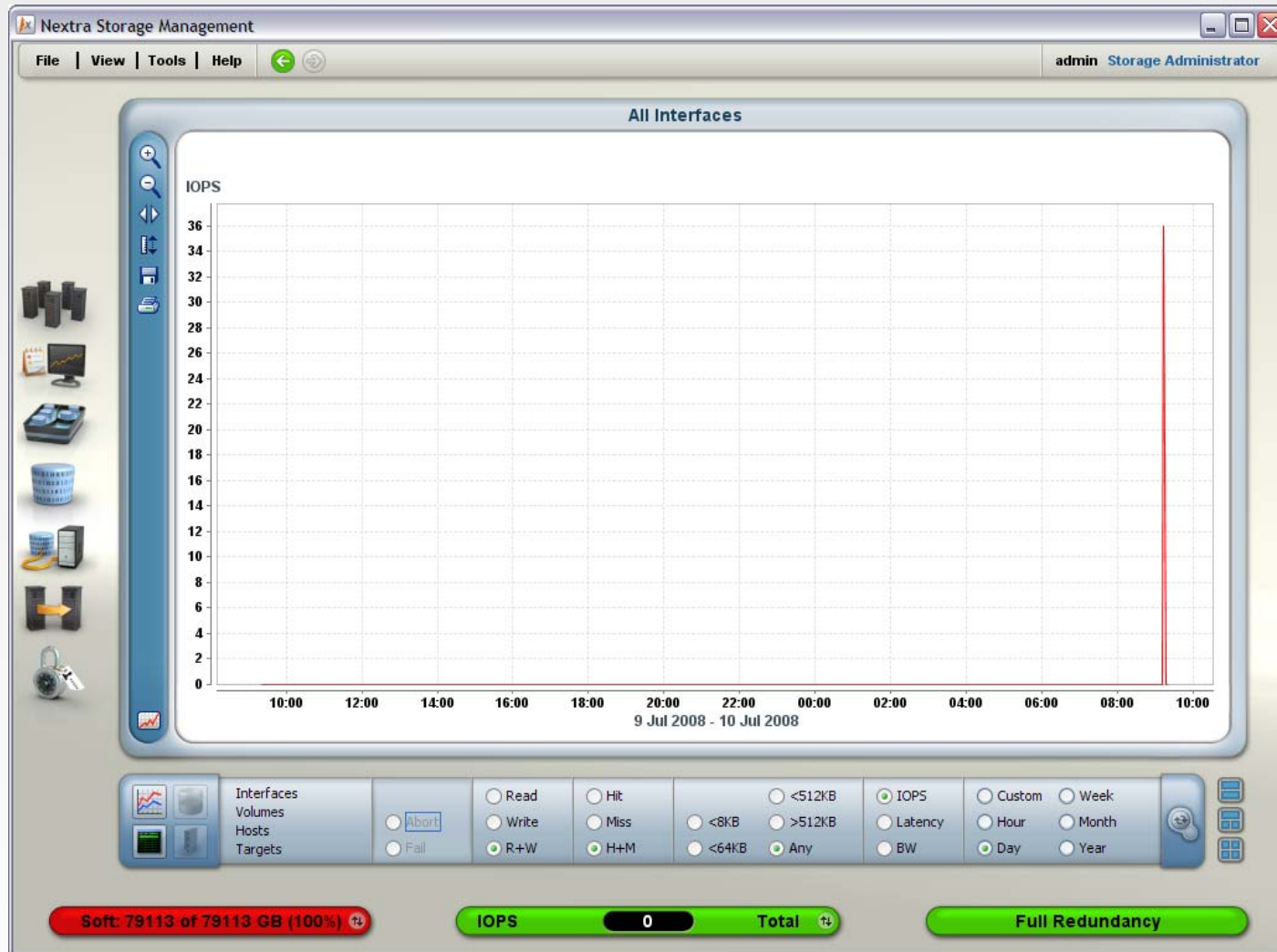
Volumes	
Name	Size (GB)
Email_Vol_1.snapshot_00002	3367
Email_Vol_1.snapshot_00003	3367
ERP_Const.snap_group_00001.ERP_Vol_1	5033
ERP_Const.snap_group_00001.ERP_Vol_2	979
ERP_Vol_2	979

LUNs		
LUN	Name	Size (GB)
0		
1	ERP_Vol_1	5033
2	ERP_Vol_2	979
3		
4		
5		
6		
7		
8		
9		
10		
11		
12		
13		
14		
15		
16		
17		
18		
19		
20		
21		
22		
23		
24		
25		

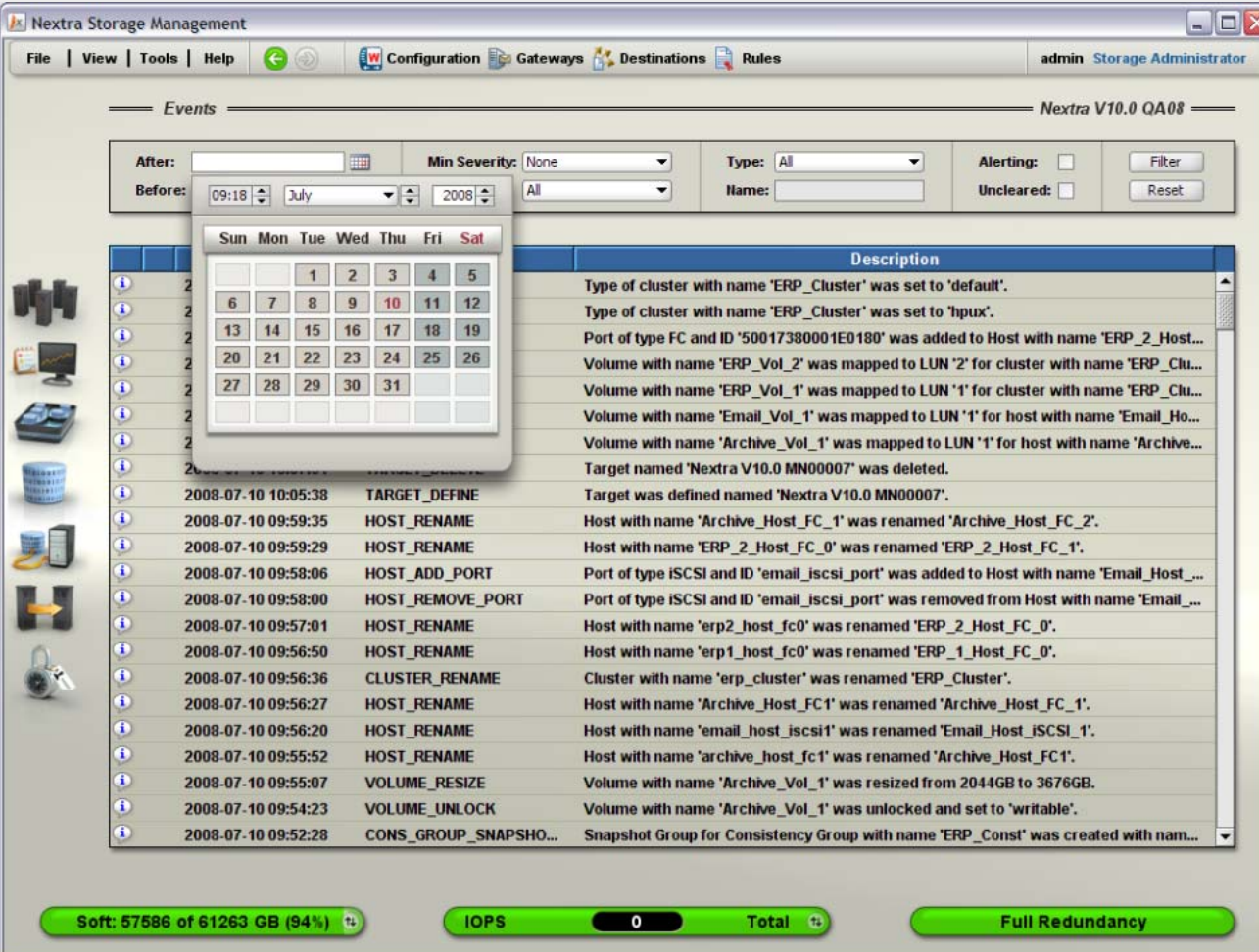
Map Unmap

Soft: 57586 of 61263 GB (94%) IOPS 0 Total Full Redundancy

IBM XIV Storage: Monitoring



IBM XIV Storage: Events Log



Nextra Storage Management

File | View | Tools | Help | Configuration | Gateways | Destinations | Rules | admin Storage Administrator

Events Nextra V10.0 QA08

After: [Date Picker] Min Severity: None Type: All Alerting: ☐ Filter
 Before: 09:18 July 2008 All Name: [Text Box] Uncleared: ☐ Reset

		Sun Mon Tue Wed Thu Fri Sat						
			1	2	3	4	5	
6	7	8	9	10	11	12		
13	14	15	16	17	18	19		
20	21	22	23	24	25	26		
27	28	29	30	31				

		Description
		Type of cluster with name 'ERP_Cluster' was set to 'default'.
		Type of cluster with name 'ERP_Cluster' was set to 'hpux'.
		Port of type FC and ID '50017380001E0180' was added to Host with name 'ERP_2_Host...'
		Volume with name 'ERP_Vol_2' was mapped to LUN '2' for cluster with name 'ERP_Clu...
		Volume with name 'ERP_Vol_1' was mapped to LUN '1' for cluster with name 'ERP_Clu...
		Volume with name 'Email_Vol_1' was mapped to LUN '1' for host with name 'Email_Ho...
		Volume with name 'Archive_Vol_1' was mapped to LUN '1' for host with name 'Archive...
		Target named 'Nextra V10.0 MN00007' was deleted.
		Target was defined named 'Nextra V10.0 MN00007'.
2008-07-10 10:05:38	TARGET_DEFINE	Target was defined named 'Nextra V10.0 MN00007'.
2008-07-10 09:59:35	HOST_RENAME	Host with name 'Archive_Host_FC_1' was renamed 'Archive_Host_FC_2'.
2008-07-10 09:59:29	HOST_RENAME	Host with name 'ERP_2_Host_FC_0' was renamed 'ERP_2_Host_FC_1'.
2008-07-10 09:58:06	HOST_ADD_PORT	Port of type iSCSI and ID 'email_iscsi_port' was added to Host with name 'Email_Host_...
2008-07-10 09:58:00	HOST_REMOVE_PORT	Port of type iSCSI and ID 'email_iscsi_port' was removed from Host with name 'Email_...
2008-07-10 09:57:01	HOST_RENAME	Host with name 'erp2_host_fc0' was renamed 'ERP_2_Host_FC_0'.
2008-07-10 09:56:50	HOST_RENAME	Host with name 'erp1_host_fc0' was renamed 'ERP_1_Host_FC_0'.
2008-07-10 09:56:36	CLUSTER_RENAME	Cluster with name 'erp_cluster' was renamed 'ERP_Cluster'.
2008-07-10 09:56:27	HOST_RENAME	Host with name 'Archive_Host_FC1' was renamed 'Archive_Host_FC_1'.
2008-07-10 09:56:20	HOST_RENAME	Host with name 'email_host_iscsi1' was renamed 'Email_Host_ISCSI_1'.
2008-07-10 09:55:52	HOST_RENAME	Host with name 'archive_host_fc1' was renamed 'Archive_Host_FC1'.
2008-07-10 09:55:07	VOLUME_RESIZE	Volume with name 'Archive_Vol_1' was resized from 2044GB to 3676GB.
2008-07-10 09:54:23	VOLUME_UNLOCK	Volume with name 'Archive_Vol_1' was unlocked and set to 'writable'.
2008-07-10 09:52:28	CONS_GROUP_SNAPSHO...	Snapshot Group for Consistency Group with name 'ERP_Const' was created with nam...


Soft: 57586 of 61263 GB (94%) IOPS 0 Total Full Redundancy

System Power Usage

- Power consumption of a system comparable to XIV is 180-380W per raw TB
 - Typically using 146GB 15K rpm disks (380W per TB)
- Power consumption of an XIV rack is 7.7KW
 - 180TB raw capacity, 79TB net capacity
 - 42W per raw TB today
- Rack power consumption will not change much with 2TB disks
 - But capacity will double
 - Consumption per raw TB expected to drop to 21W

System Power Usage

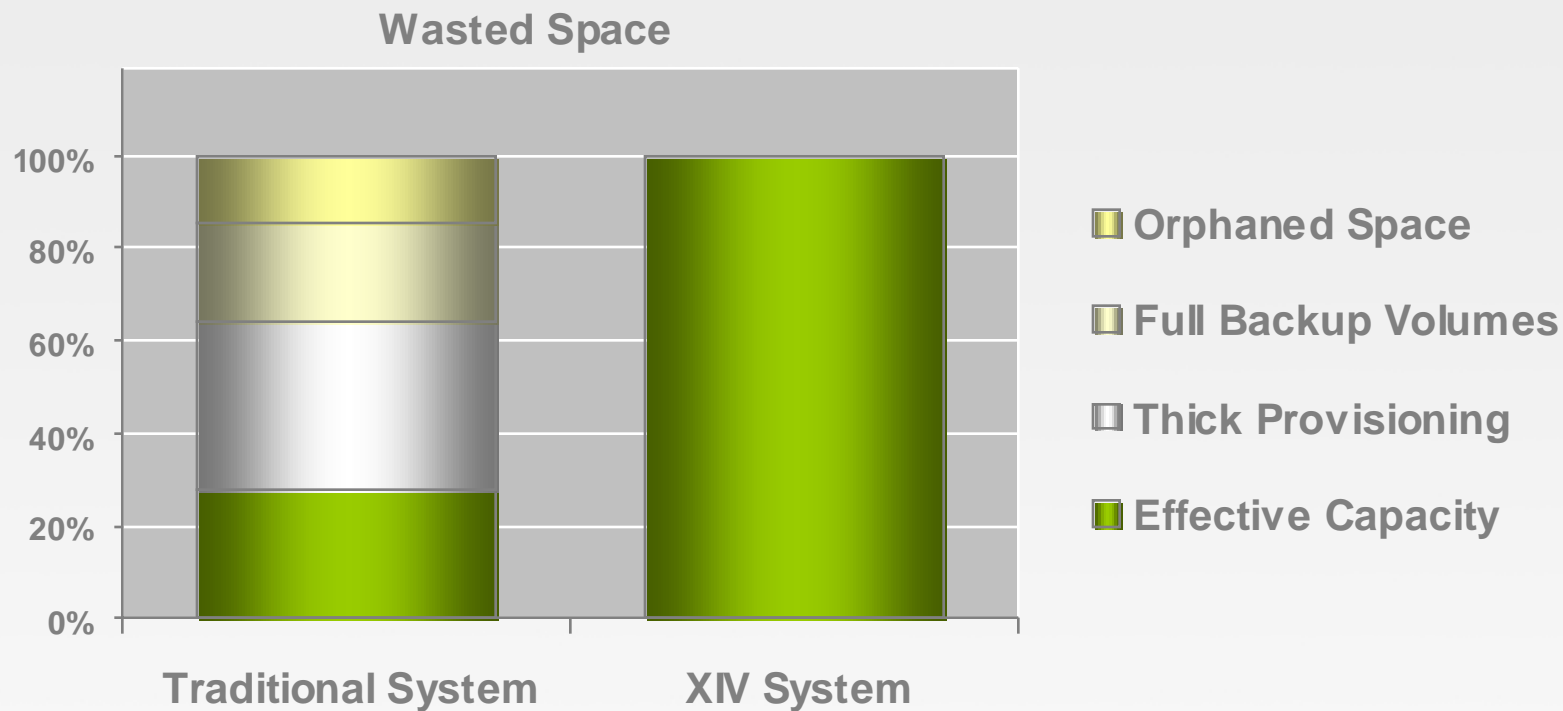
- Power consumption of a system comparable to XIV is 180-380W per raw TB
 - Typically using 146GB 15K rpm disks (380W per TB)
- Power consumption of an XIV rack is 7.7KW
 - 180TB raw capacity, 79TB net capacity
 - 42W per raw TB today
- Rack power consumption will not change much with 3TB disks



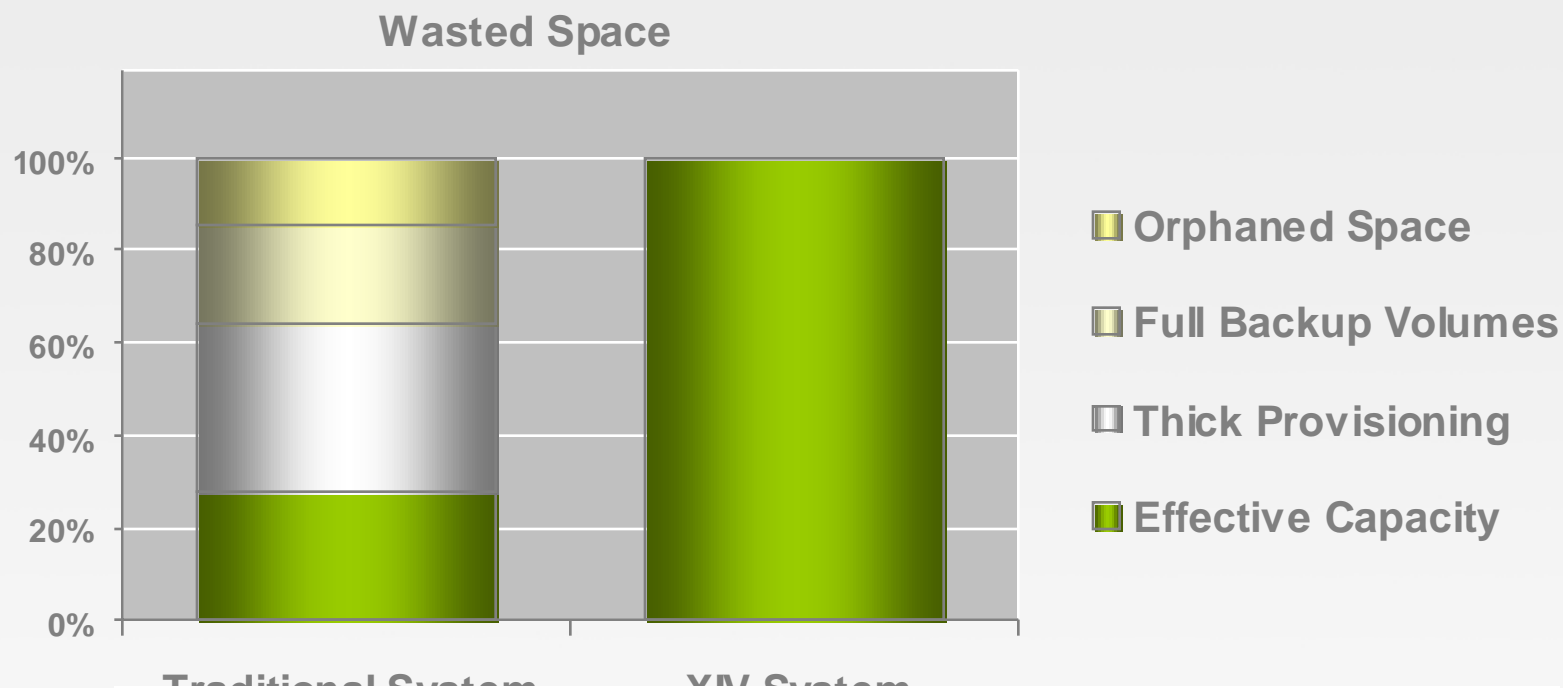
The new solution uses 4 to 9 times less power for the same (or better) performance and reliability levels

21W

Stretching a TB to the Max



Stretching a TB to the Max



Real-life capacity gain with XIV

- Meet the same functional needs with much less net TBs

Customer Success Story

Customer Problem

Bank has 7TB Oracle Database for logging activities (compliance).
 Extreme performance requirements.
 Tried Hi End tier 1 systems without success.
 Hot backup was not possible with current storage.

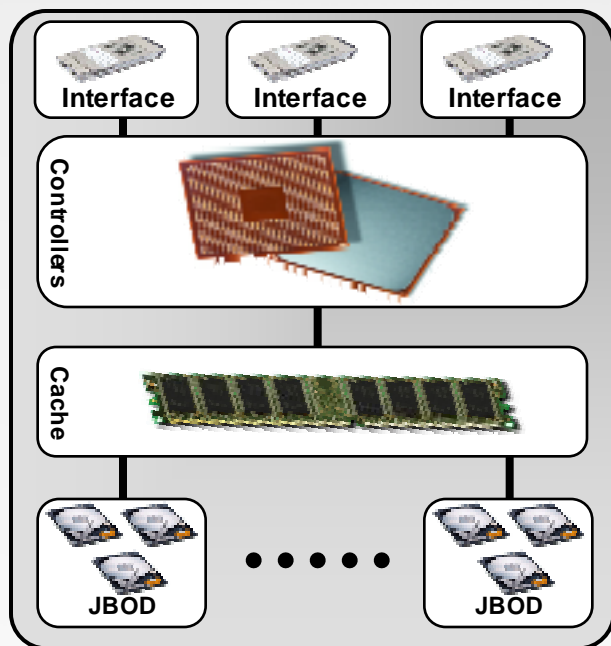
With XIV

XIV Technology

Achieve higher TPM than other high-end systems	High Spindle Utilization
Able to do hot backups with no performance impact	Distributed snapshot algorithm
Now taking 4 daily snapshots for backup Snapshots are saved for a week Can instantly return to any of the 28 snapshots	Efficient, differential snapshots

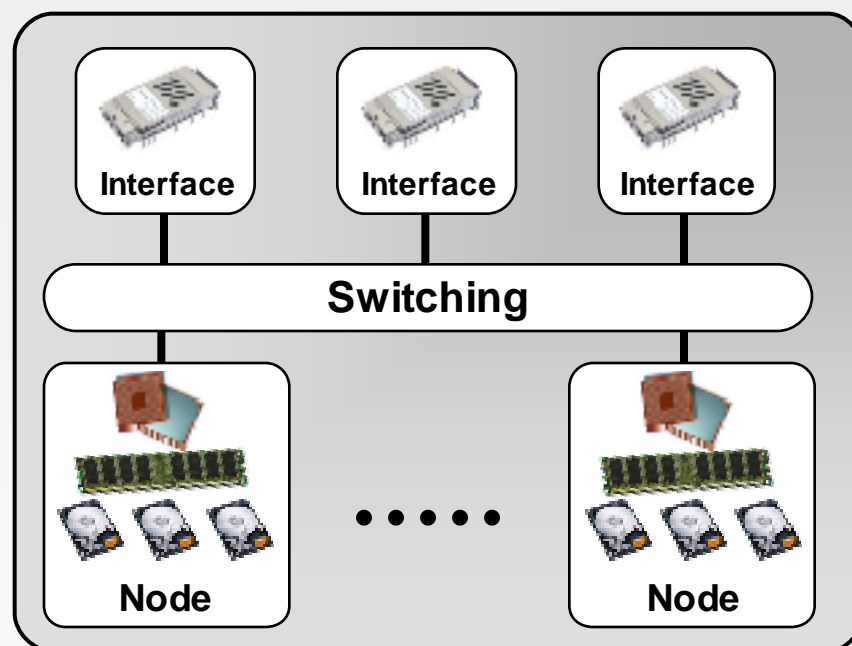
A Novel Storage Architecture

- Dual node Clusters
- Tightly coupled
- Custom HW design
- Expensive components



- Long, complex development cycles
- System exposed on failures
- Complex reactive service
- Requires tuning for optimal performance

- Fast, efficient development cycles
- Self Healing
- Scheduled, convenient service
- Autonomic tuning



- Scalable Grid nodes
- Open: node independent
- Commodity HW building blocks
- Low cost components

The Bottom Line: Real-World Benefits

- **Reliability**
 - Revolutionary self-healing that takes minutes, not hours
 - Grid “WEB” resiliency
- **Convenience**
 - **Performance**
 - Massive parallelism, spindle utilization, and cache effectiveness boost performance dramatically
 - No need to “optimize disk layout” or manage “data tiers”
 - **Manageability**
 - A logical volume has only two parameters: name and size
- **Cost**
 - Off-the-shelf components, SATA large drives
 - Self-healing allows scheduled visits for maintenance
 - Practically eliminates time spent for array management
 - Power saving
- **Functionality** - Tier 1 functions (e.g. replication, thin provisioning) that scale with no performance penalty and are inherently built-in to the architecture

Practically unlimited scalability of capacity and performance

Add capacity together with CPU, cache and bandwidth

Summary

[...] the primary motivation behind the information appliance is clear: simplicity.

Design the tool to fit the task so well that the tool becomes part of the task, feeling a natural extension of the work [...]

The Invisible Computer
Donald A. Norman



Thank You

Alain Azagury
azagury@il.ibm.com

