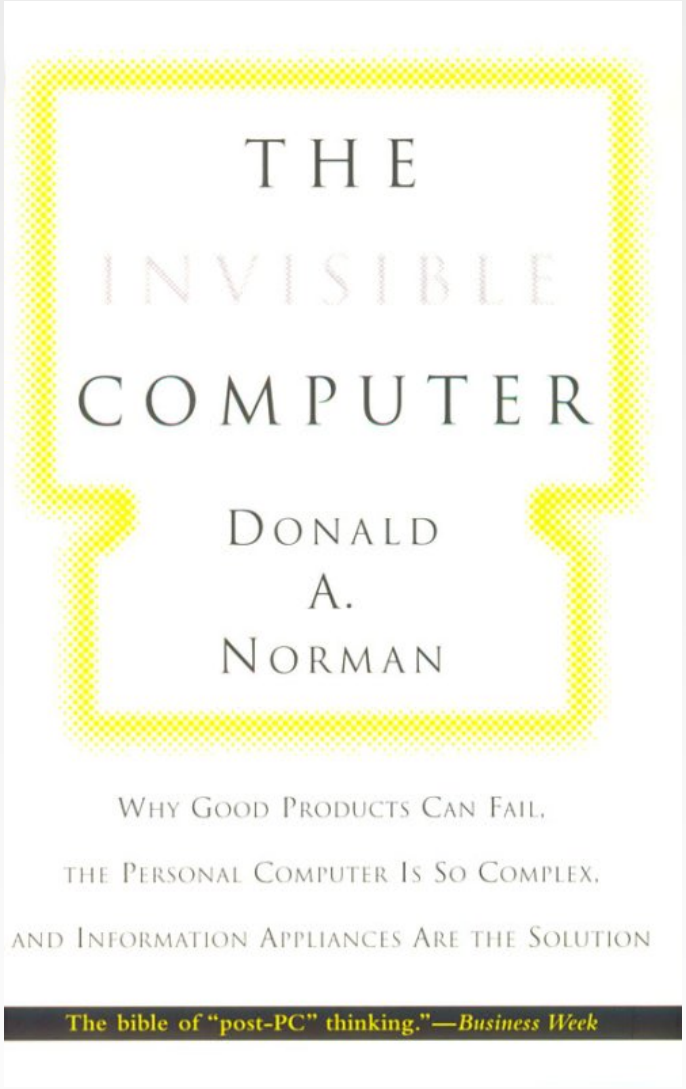
# Towards Invisible Storage

Alain Azagury
IBM XIV Business Executive

# Acknowledgement

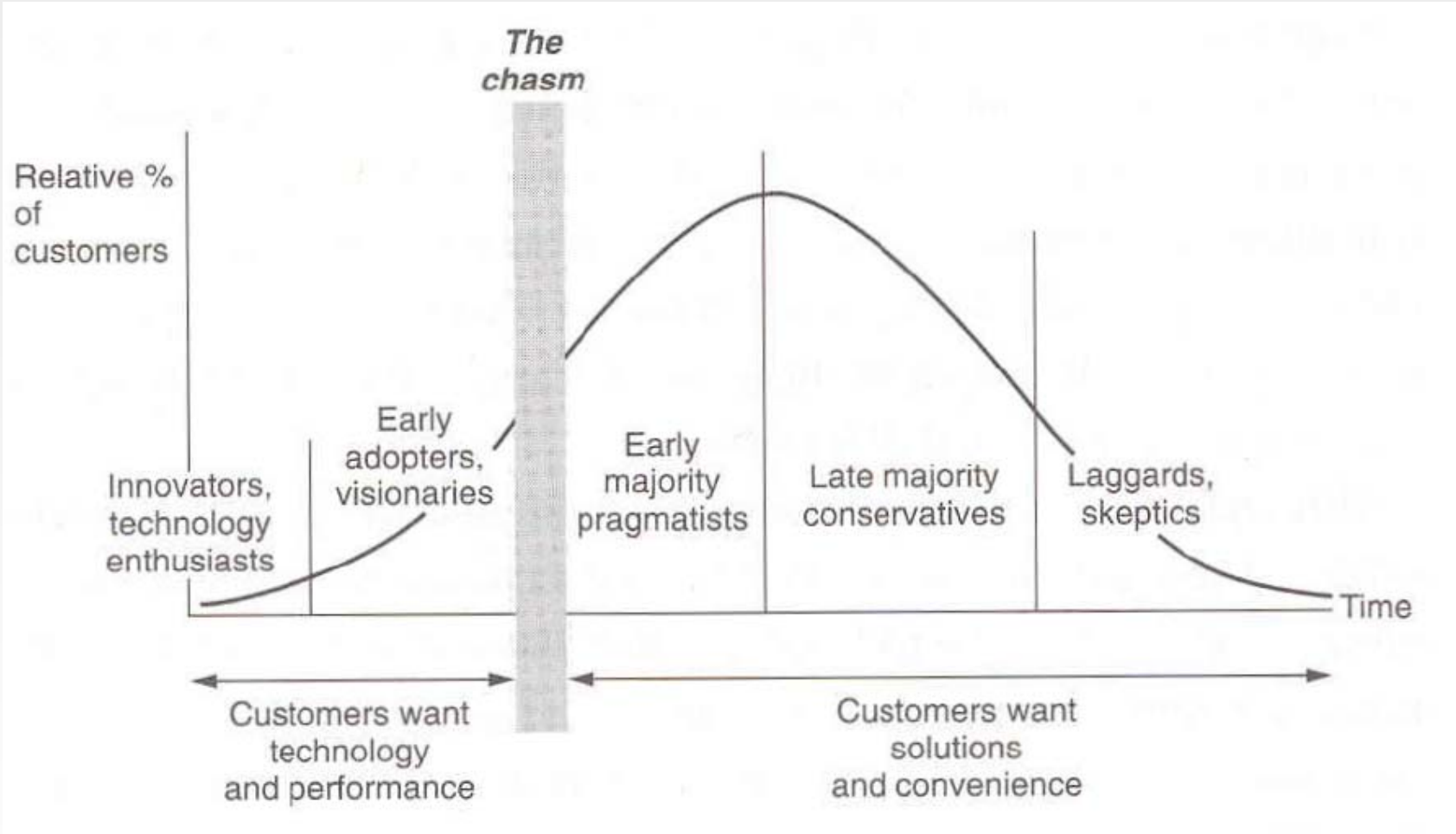
Influenced by Donald A. Norman's book
"The Invisible Computer"

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



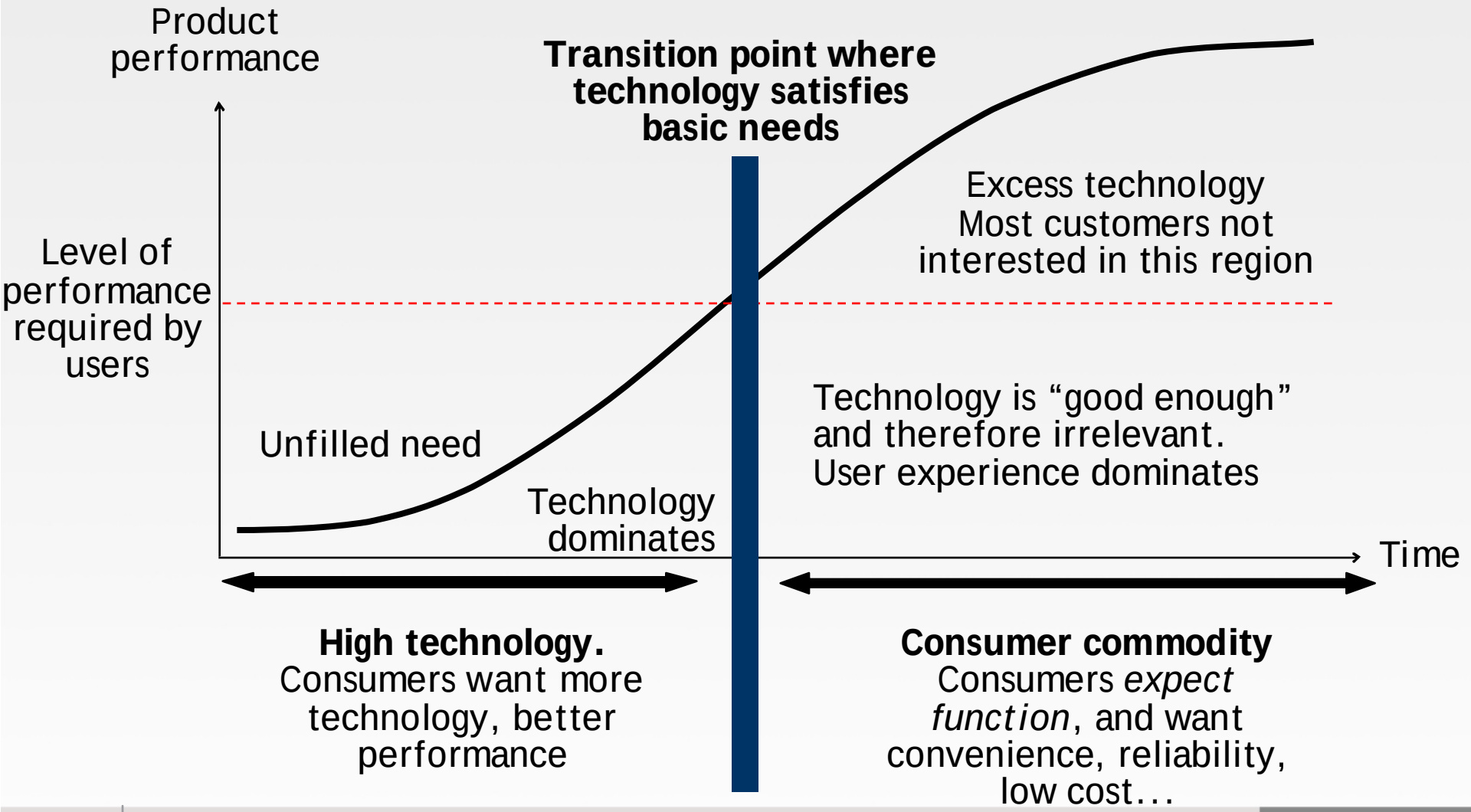
# The Change in Customers as a Technology Matures

## From Norman, 1998 (modified from Moore, 1995)

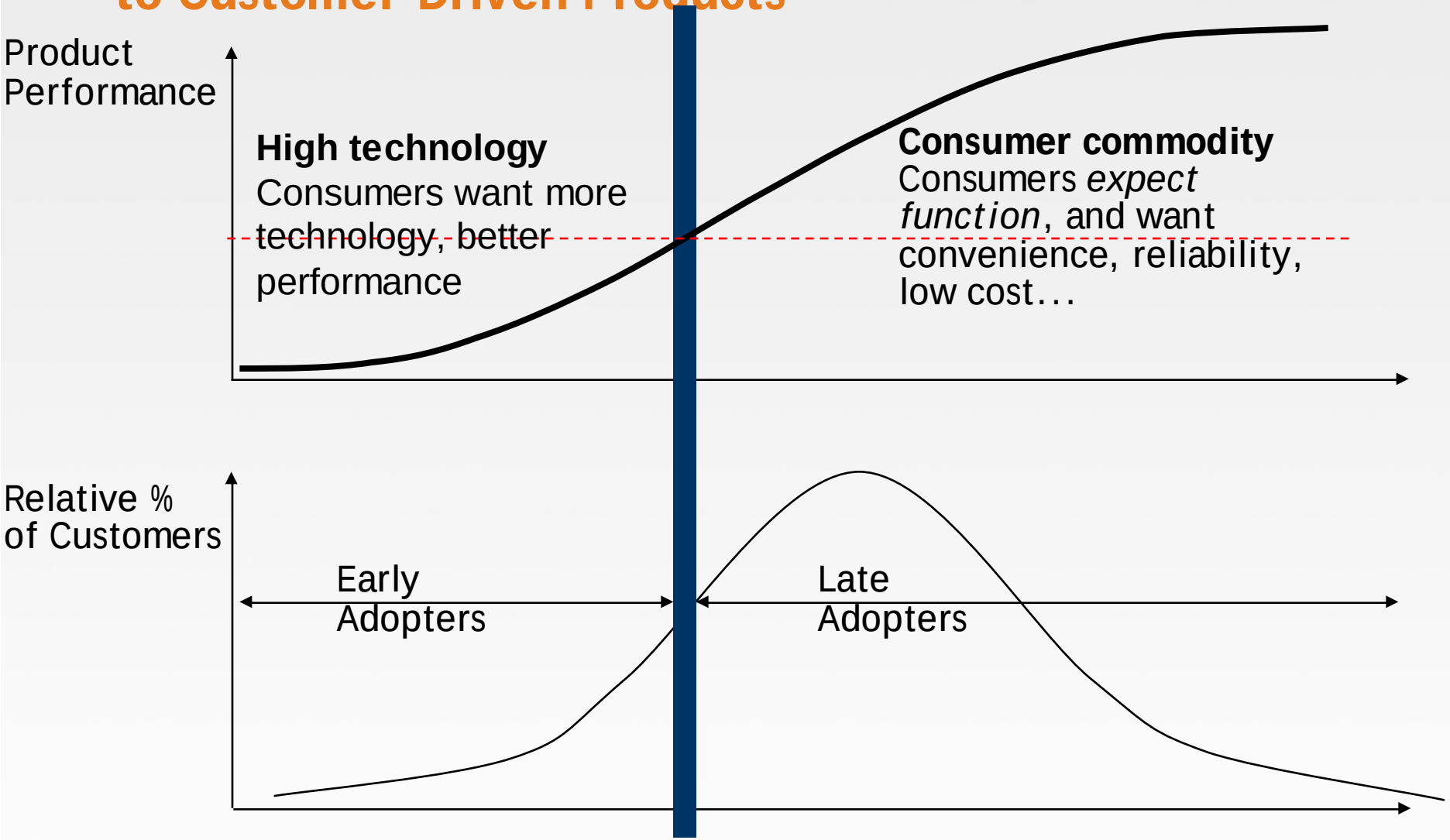


# The Needs-Satisfaction Curve of a Technology

## From Norman, 1998 (modified from Christensen, 1997)

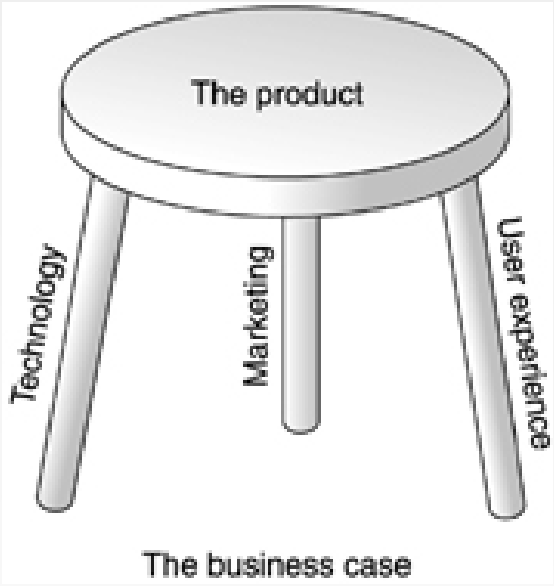


# The Change from Technology-Driven Products to Customer-Driven Products



# 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



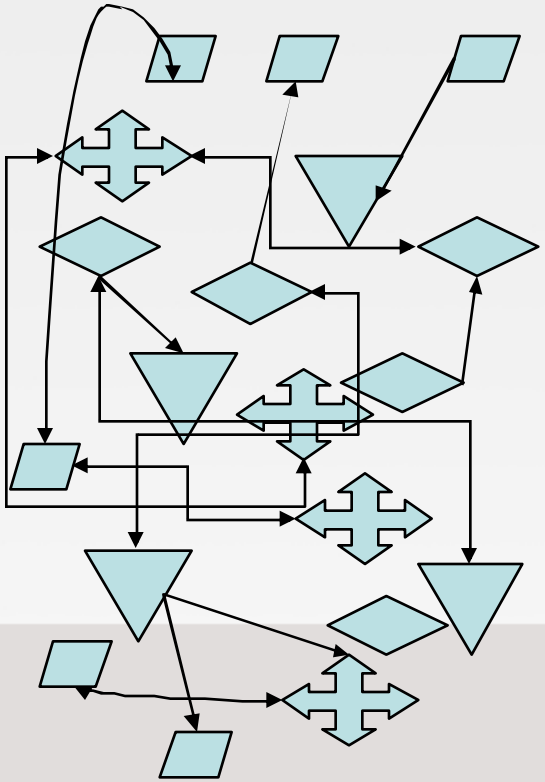

The balanced product stool.
Source: Norman 1998.

# What's, why and where to Reinvent?

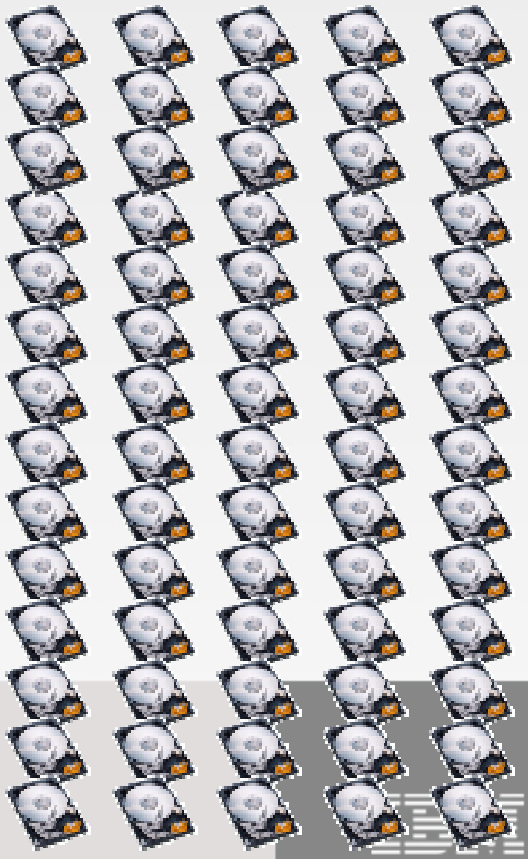
## Data

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

## Architecture



## Disks

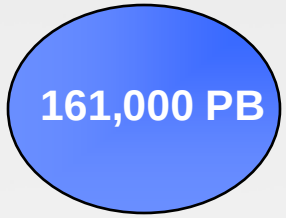


# The Need - DATA

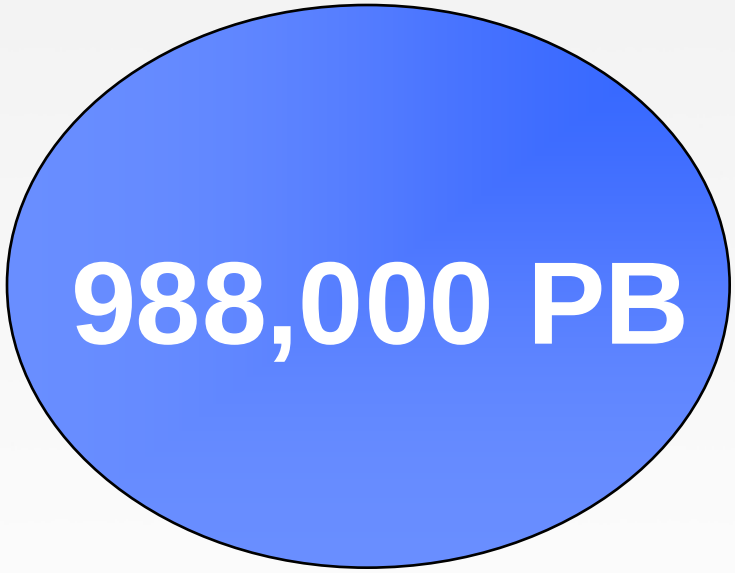
# Information Explosion Creates Storage Challenges

**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**




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



# The Media - DISKS

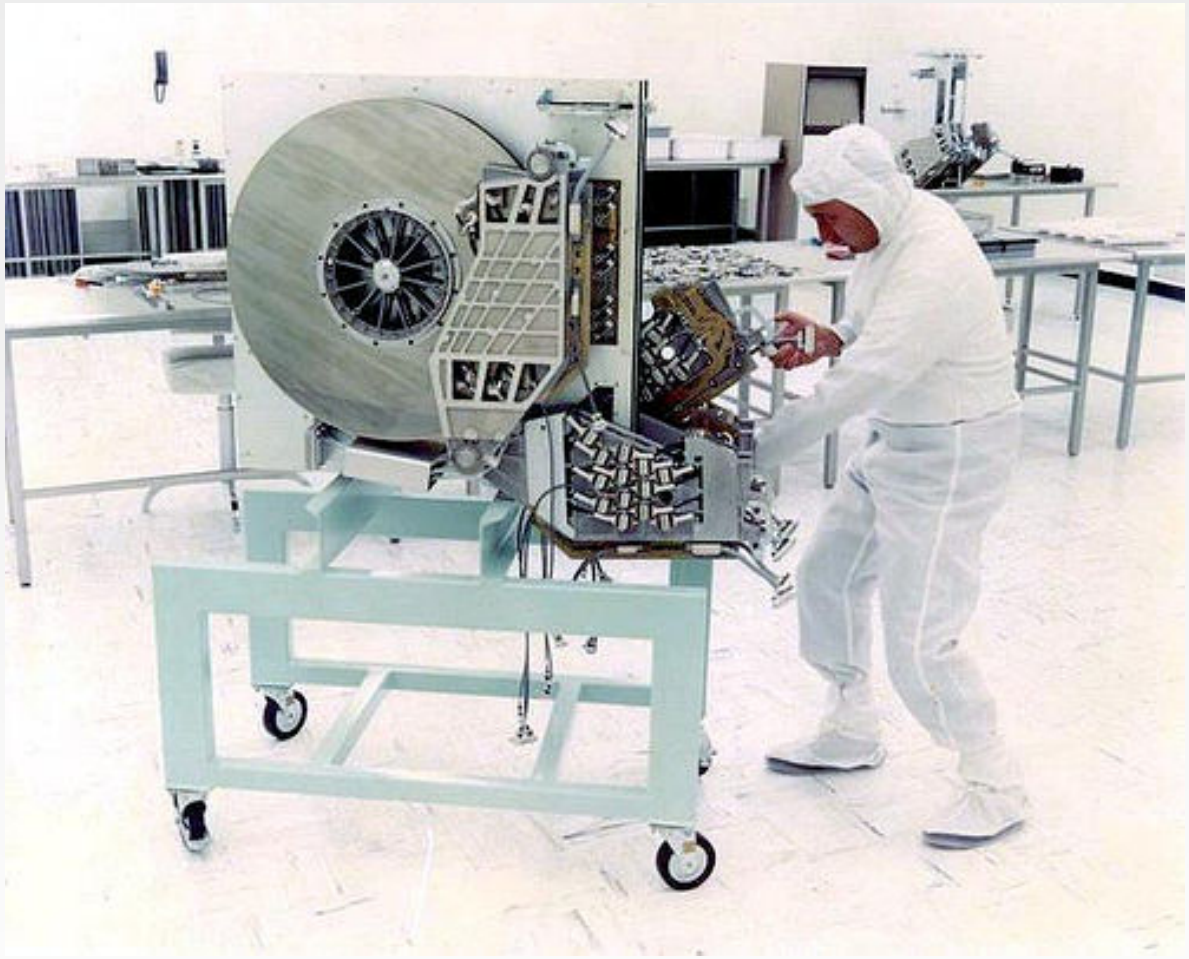
# Modern Data Storage

- Magnetic Tape - IBM pioneered the magnetic tape in 1952, realizing that both punch cards and ticker tape were far too slow

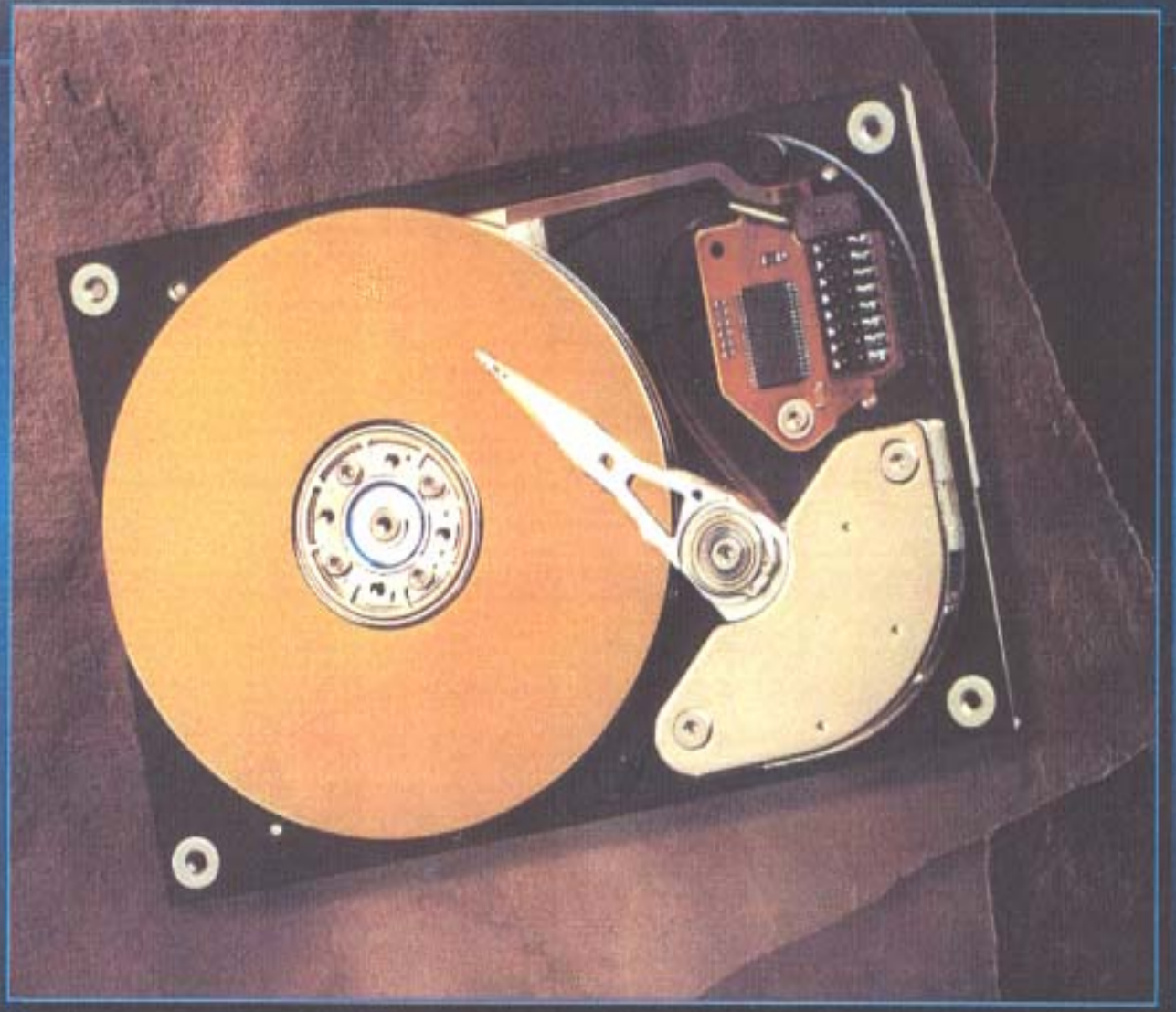
- Magnetic Disk - In 1956 a small team of IBM engineers in San Jose introduced the first computer disk storage system. The 305 RAMAC could store five megabytes of data on 50 disks, each 24 inches in diameter.

# Evolution of Storage Media

# Hard Drives in 1975



# The 1996 Ewing Lecture



IBM

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

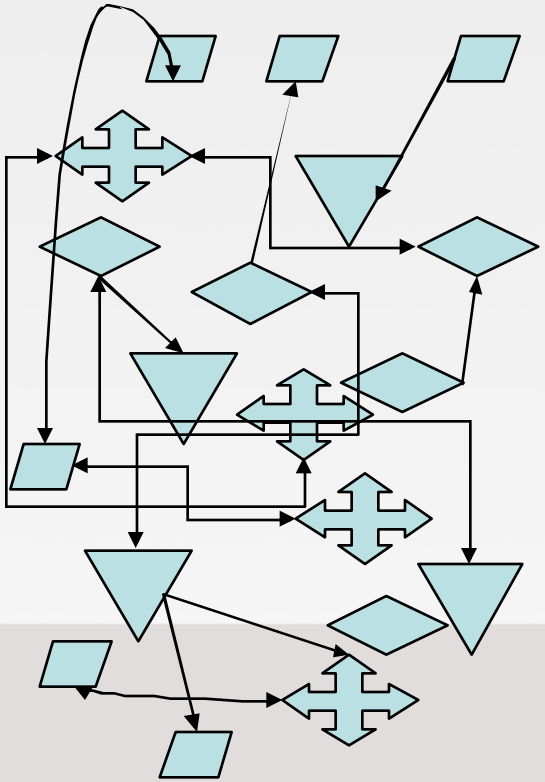
**All of these key attributes -- with unlimited scalability**

# What's, why and where to Reinvent?

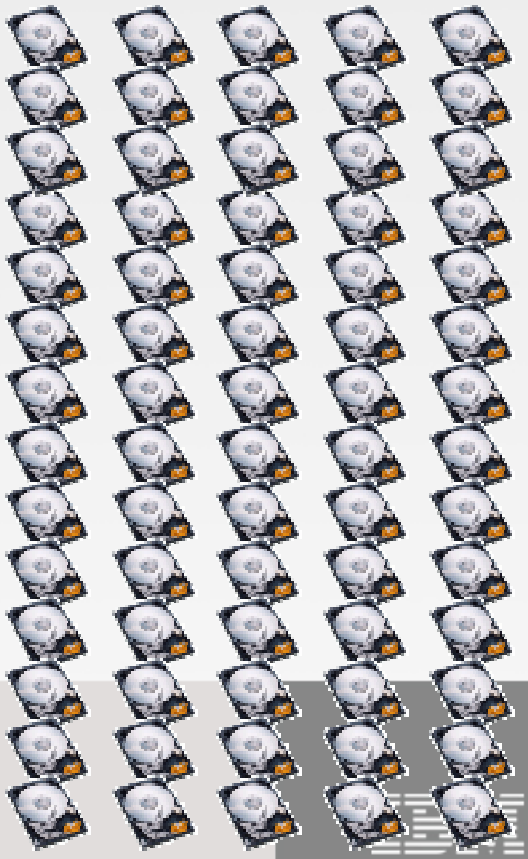
## Data

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

## Architecture



## Disks

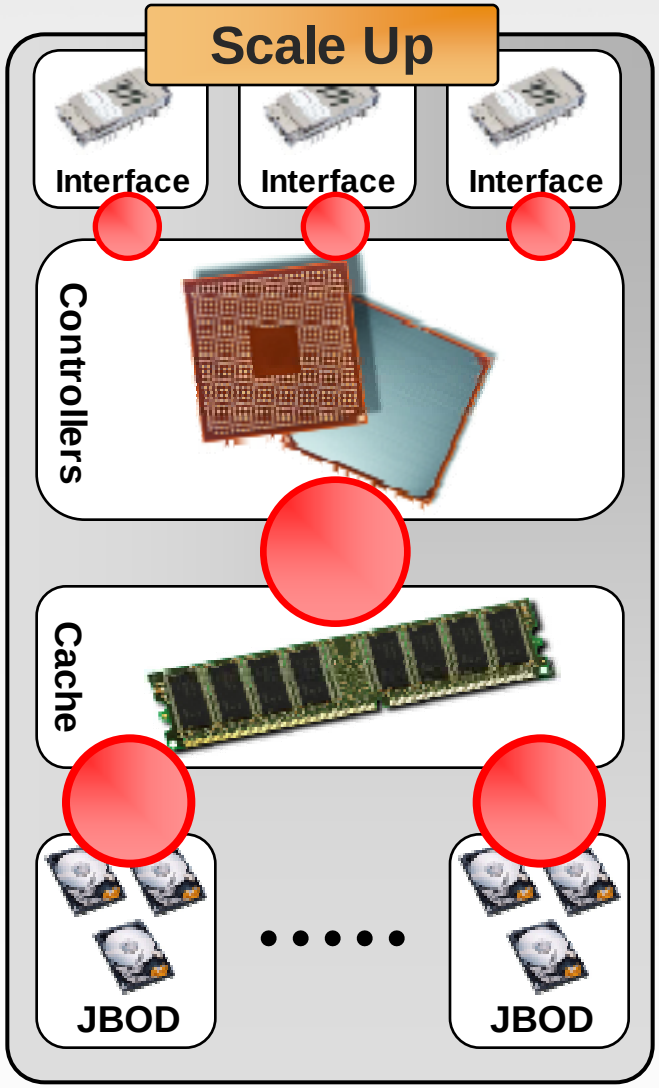


# Current Enterprise Storage Solutions

**Building blocks:**

- Disks
- Cache
- Controllers
- Interfaces
- Interconnects

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



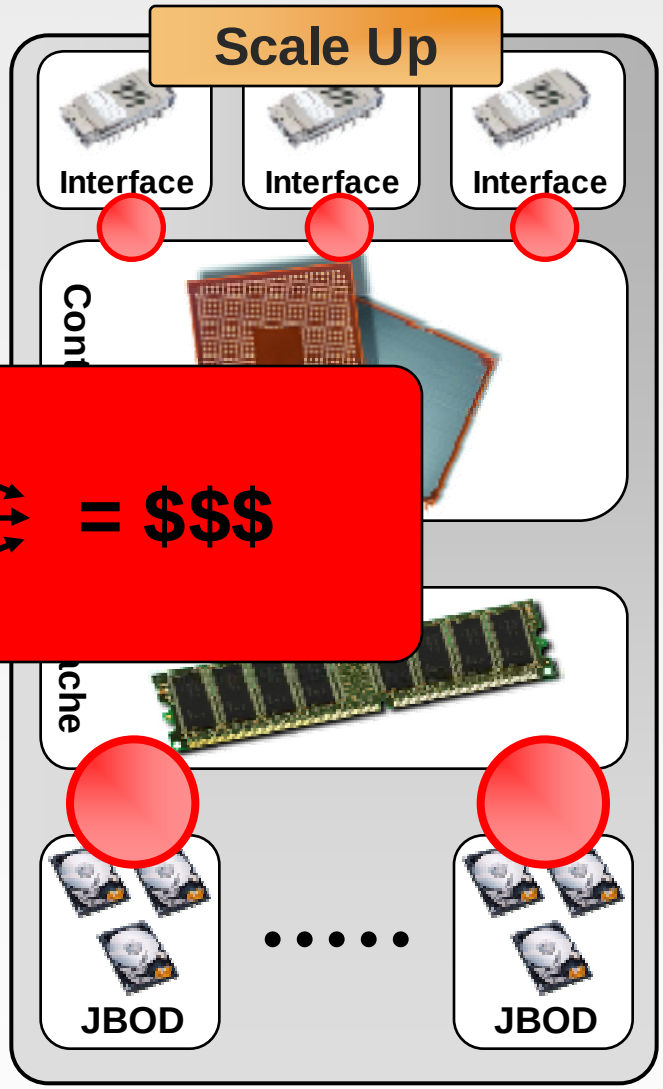
# Current Enterprise Storage Solutions

**Building blocks:**

- Disks
- Cache
- Controllers
- Interf
- Interc

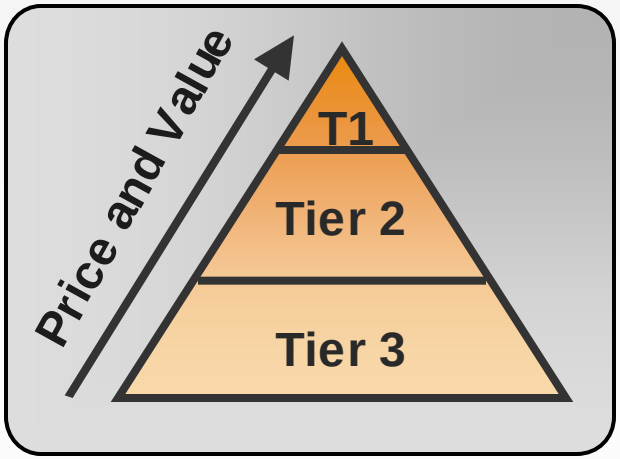
**PERFORMANCE**
**RELIABILITY** → = $$$
**SCALABILITY**

**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?**

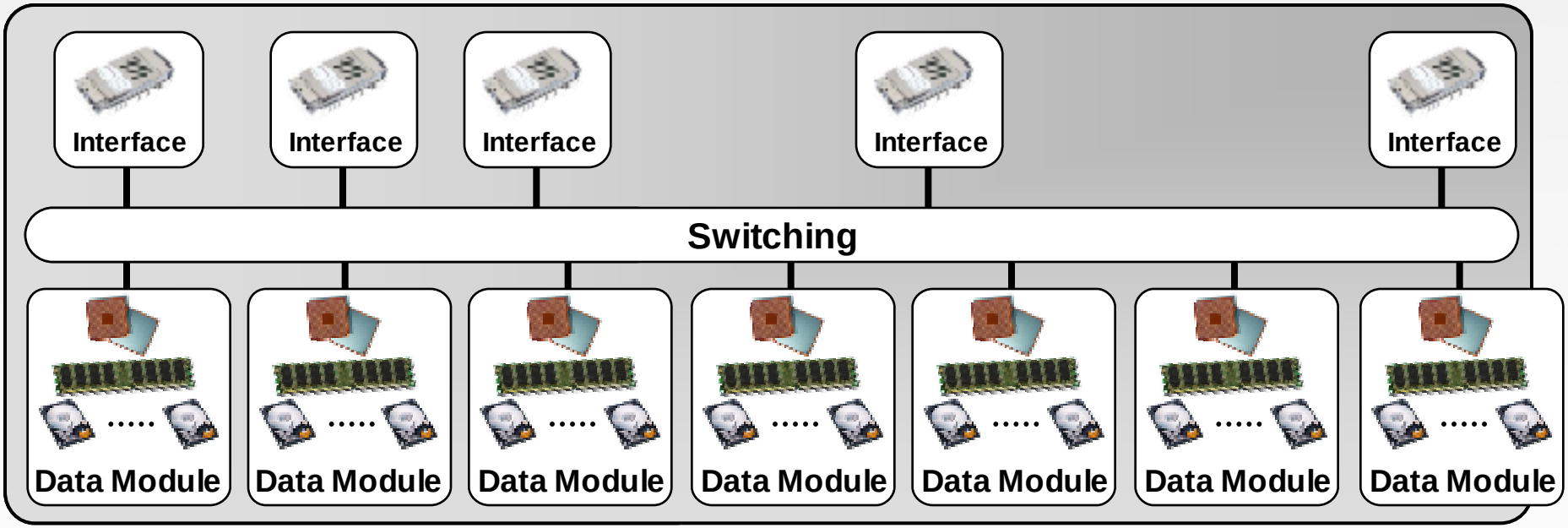
# The Next Generation Architecture

# The Next Generation Architecture

## Design principles:

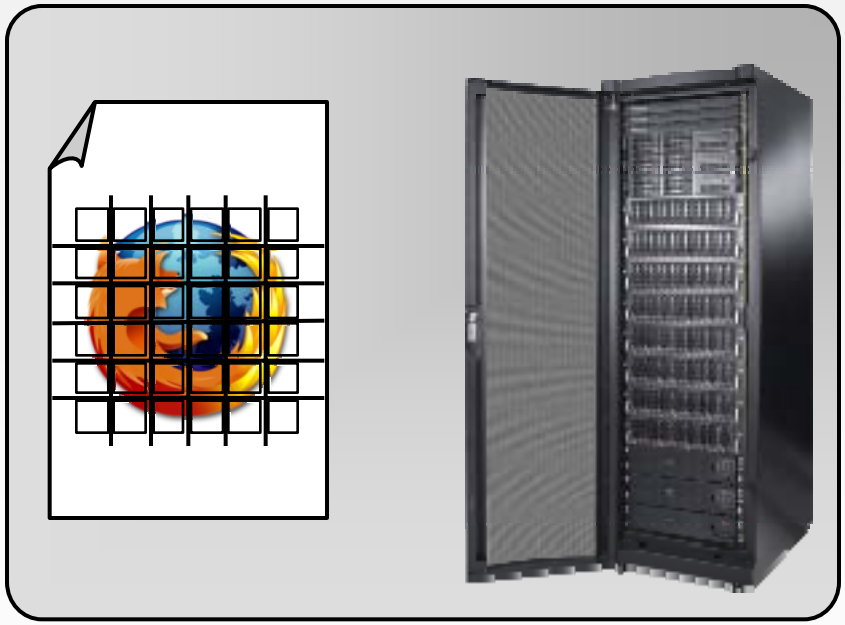
- Massive parallelism
- Granular distribution
- Off-the-shelf components
- Coupled disk, RAM and CPU
- User simplicity

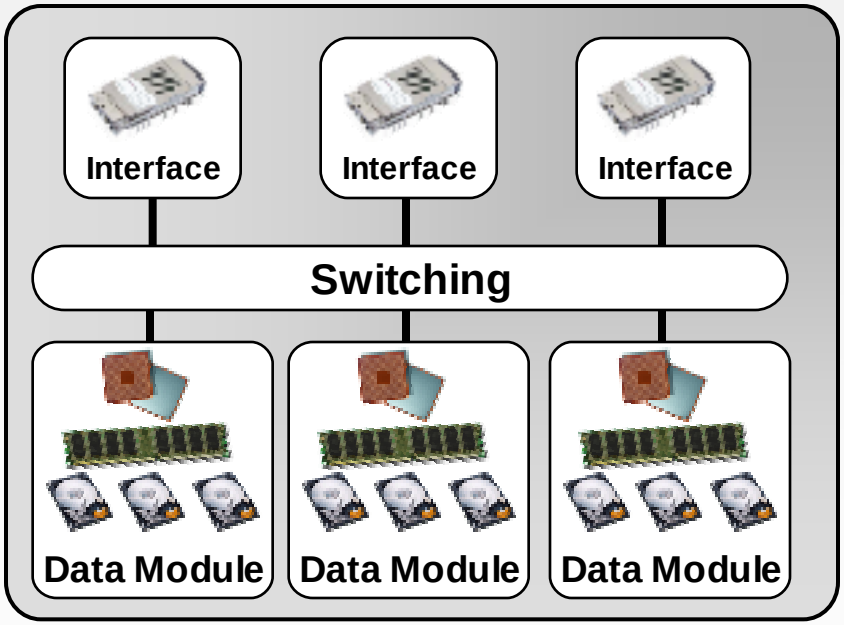
**Scale Out**



# System Distribution Algorithm

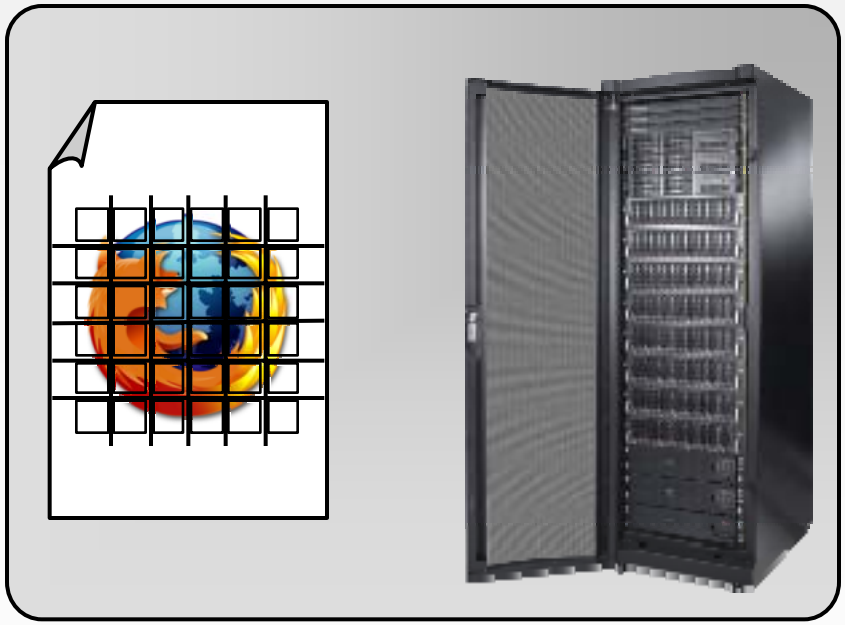
- Each volume is spread across all drives
- Data is “cut” into 1MB “partitions” and stored on the disks
- XIV’s distribution algorithm **automatically** distributes partitions across **all** disks in the system pseudo-randomly

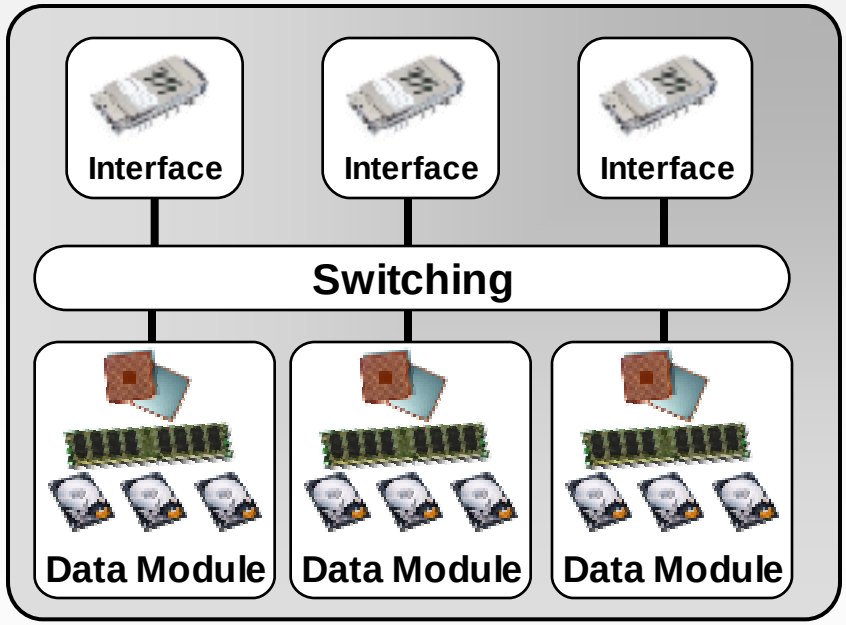




# System Distribution Algorithm

- 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



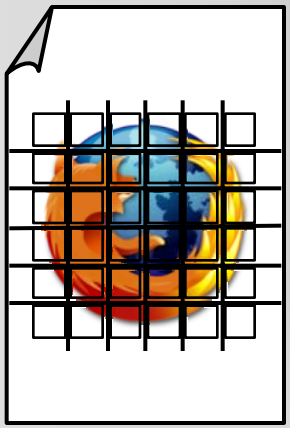


# System Distribution Algorithm

- Each volume is spread across all drives
- Data is "c on the disks
- XIV's distr ributes partitions randomly

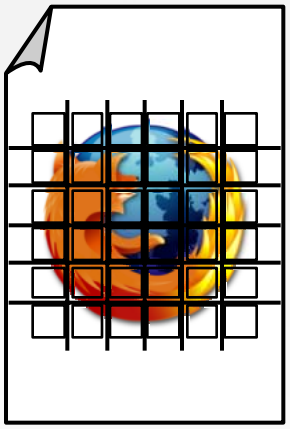
**XIV disks behave like connected vessels, as the distribution algorithm aims for constant disk equilibrium.**

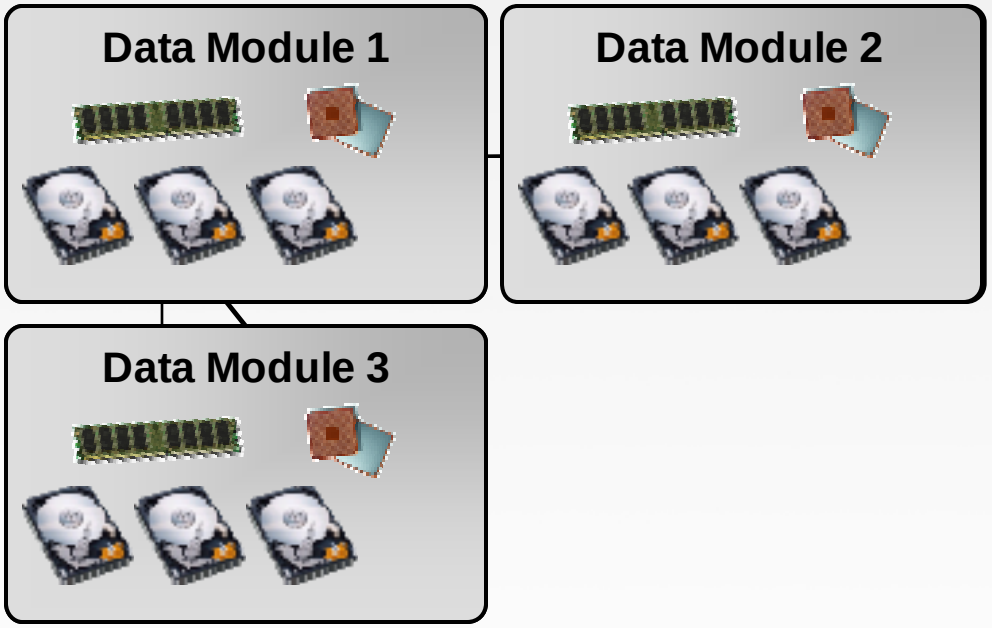
**Thus, XIV's overall disk usage approaches 100% in all usage scenarios.**



# Distribution Algorithm on System Changes

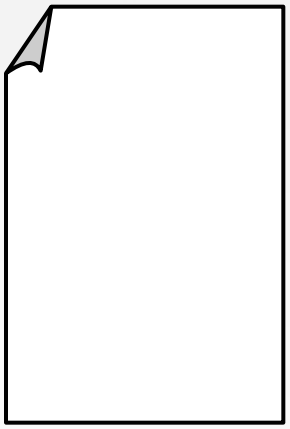
- Data distribution only changes when the system changes
  - Equilibrium is kept when new hardware is added
  - Equilibrium is kept when old hardware is removed
  - Equilibrium is kept after a hardware failure

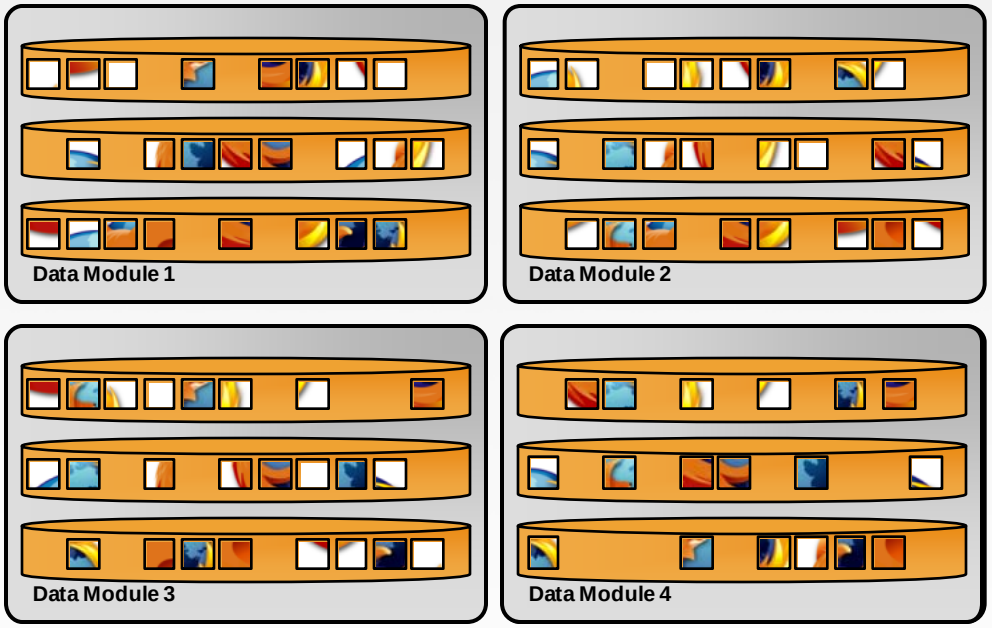




# Distribution Algorithm on System Changes

- 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

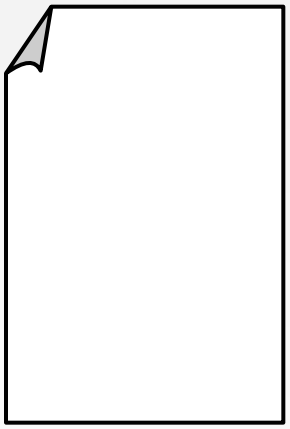


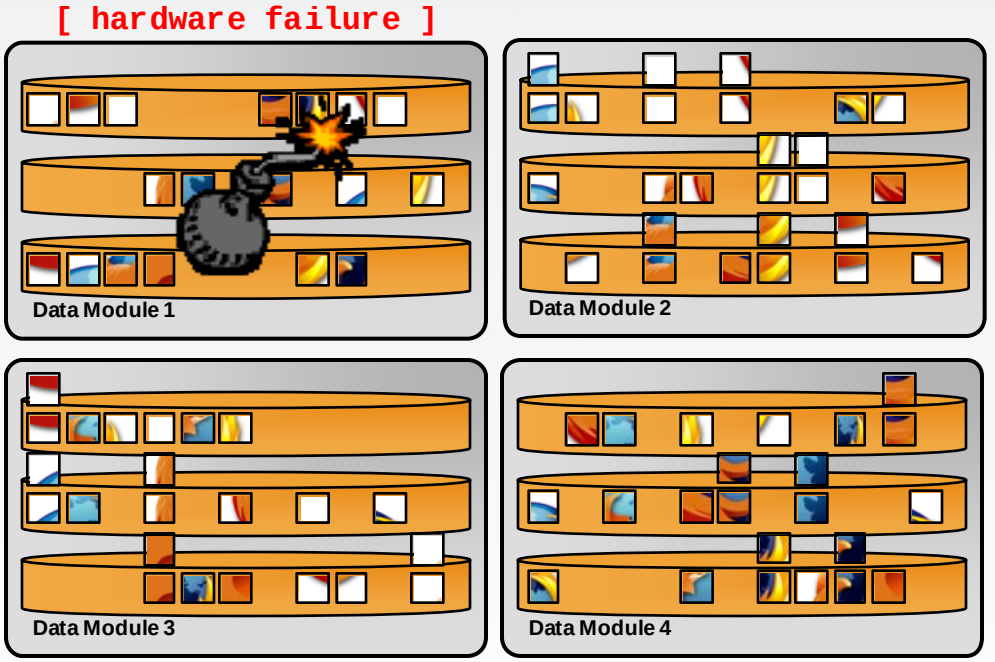



[ hardware upgrade ]

# 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



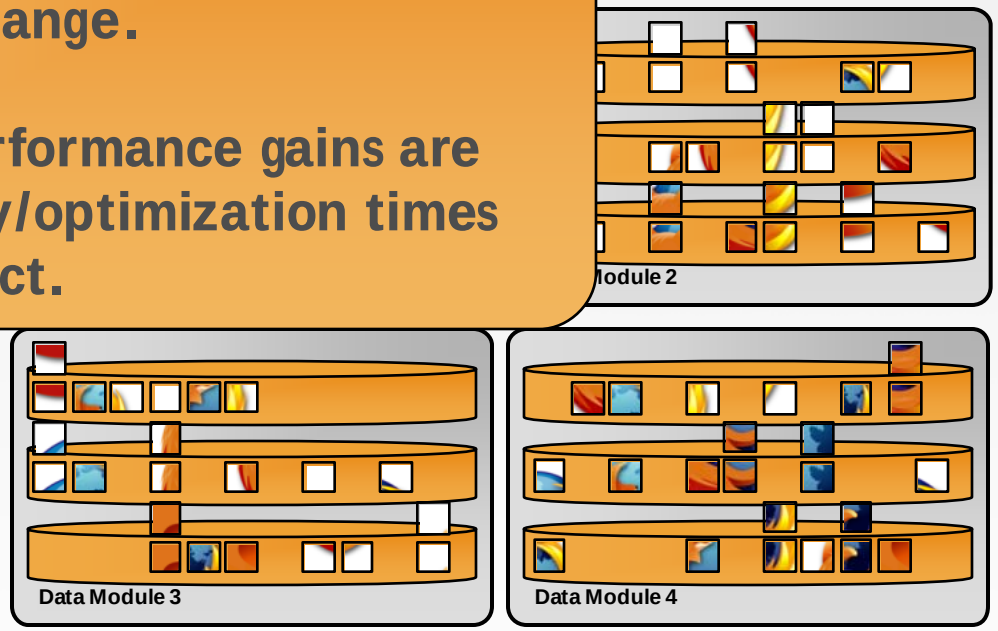


# XIV Distribution Algorithm on System Changes

- Data distribution only changes when the system changes
  - Equilibrium is kept when new hardware is added
  - Equilib…oved
  - Equilib…

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

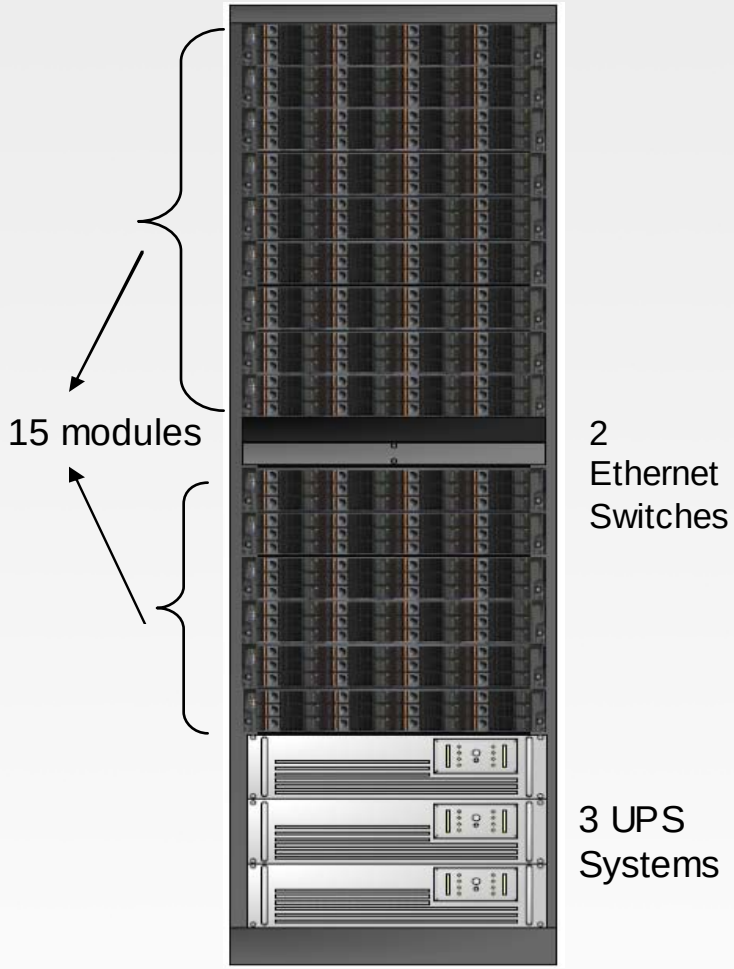
**Tremendous performance gains are seen in recovery/optimization times thanks to this fact.**



# IBM XIV Storage System Hardware Platform

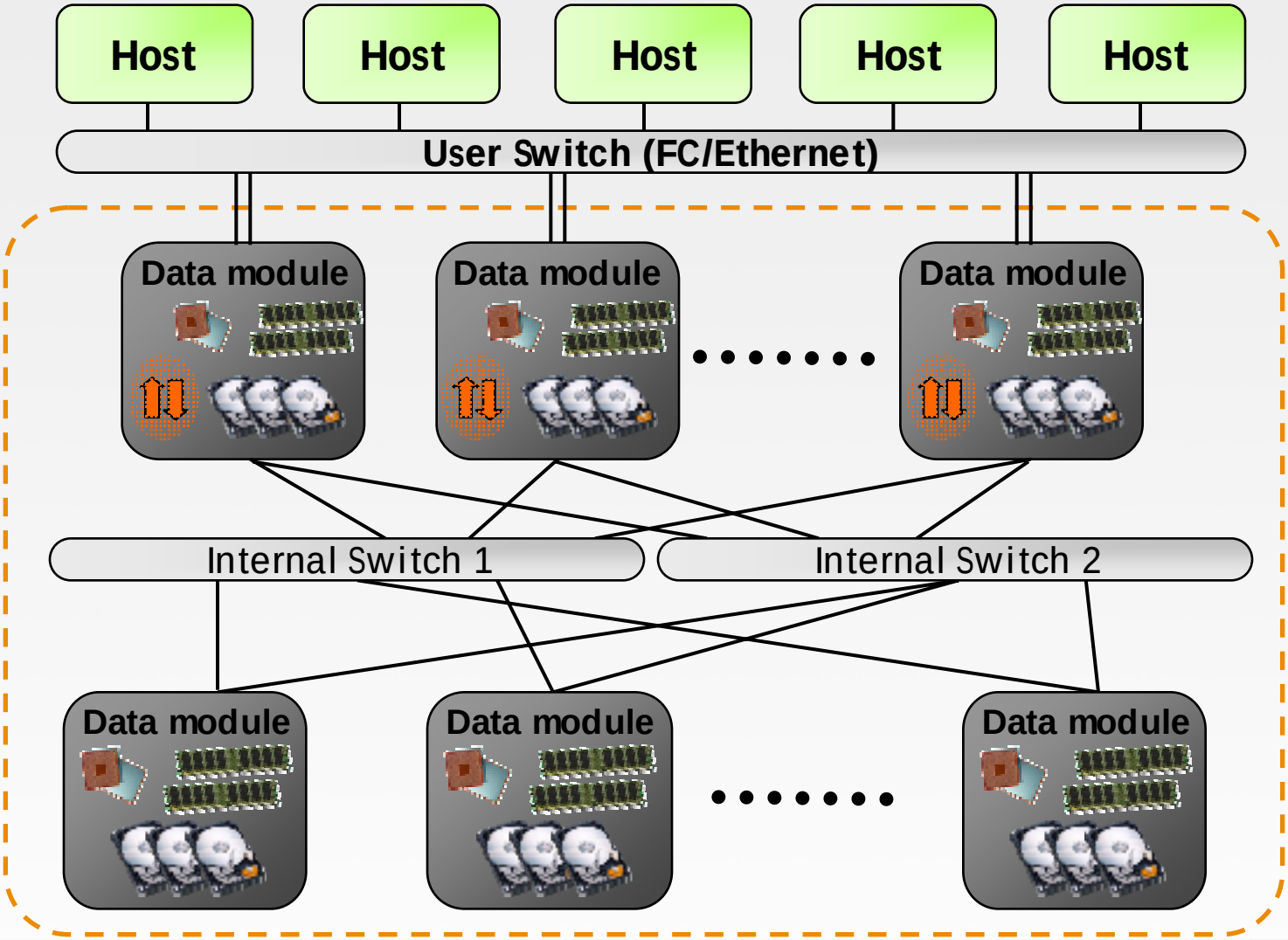
Machine Type: 2810-A14

- 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

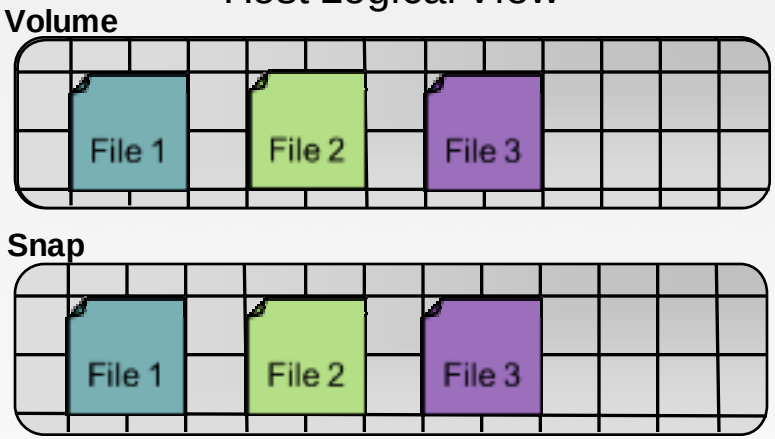
# IBM XIV Storage System's Grid Architecture



# SNAPs with No Limitations

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

Host Logical View

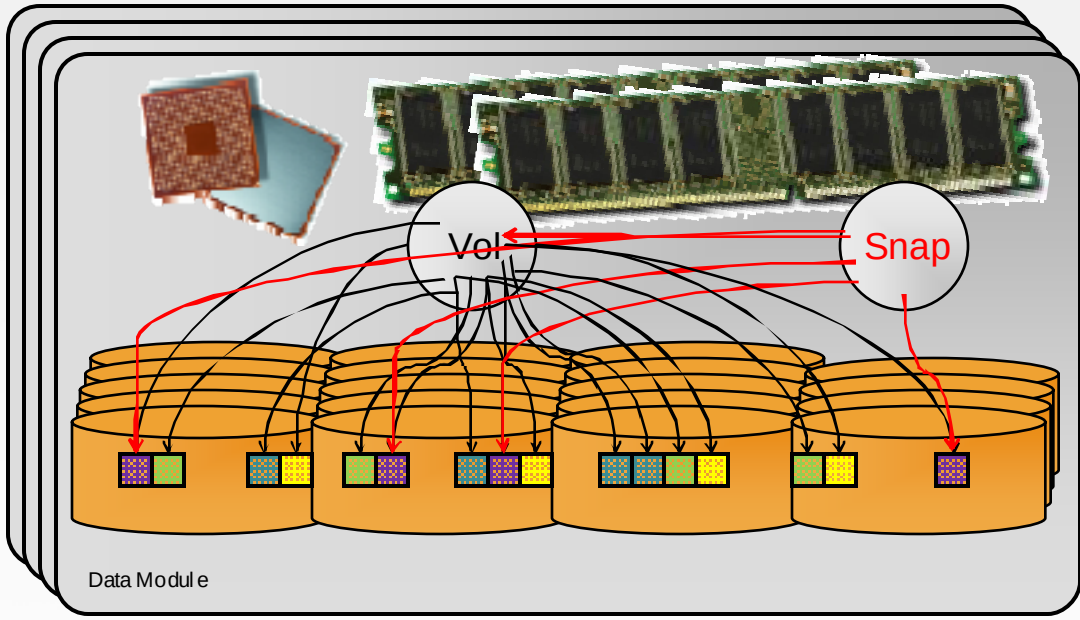



As Host Writes data, it is placed randomly across system in 1MB chunks

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



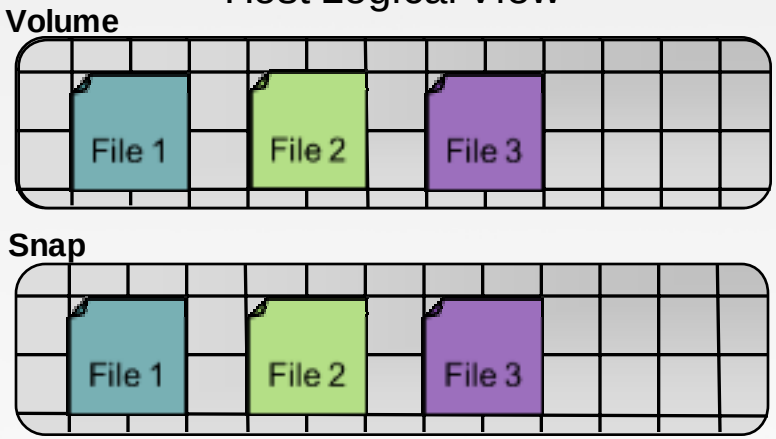
# SNAPs with No Limitations

- SNAPs creation/deletion is instanta
- 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

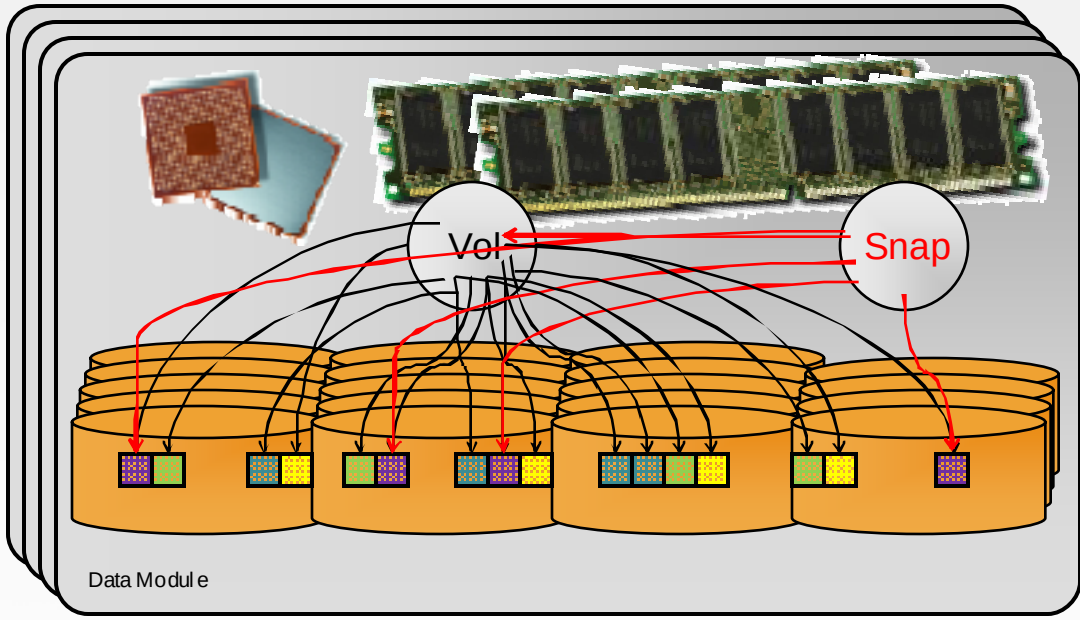



As Host Writes data, it is placed randomly across system in 1MB chunks

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 instant
- High
- Unlim

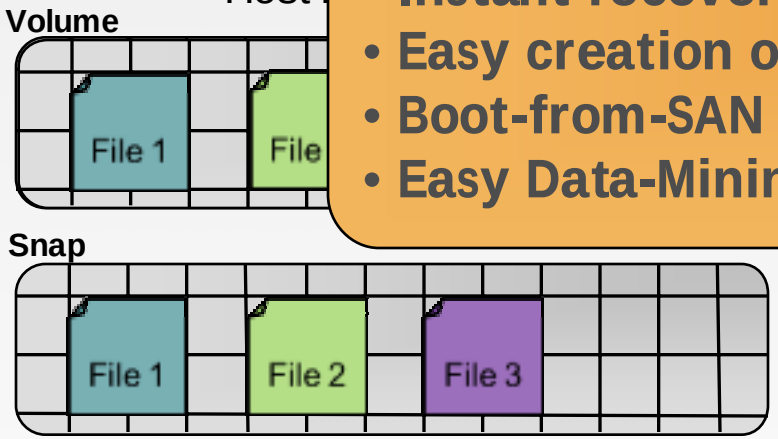
Host

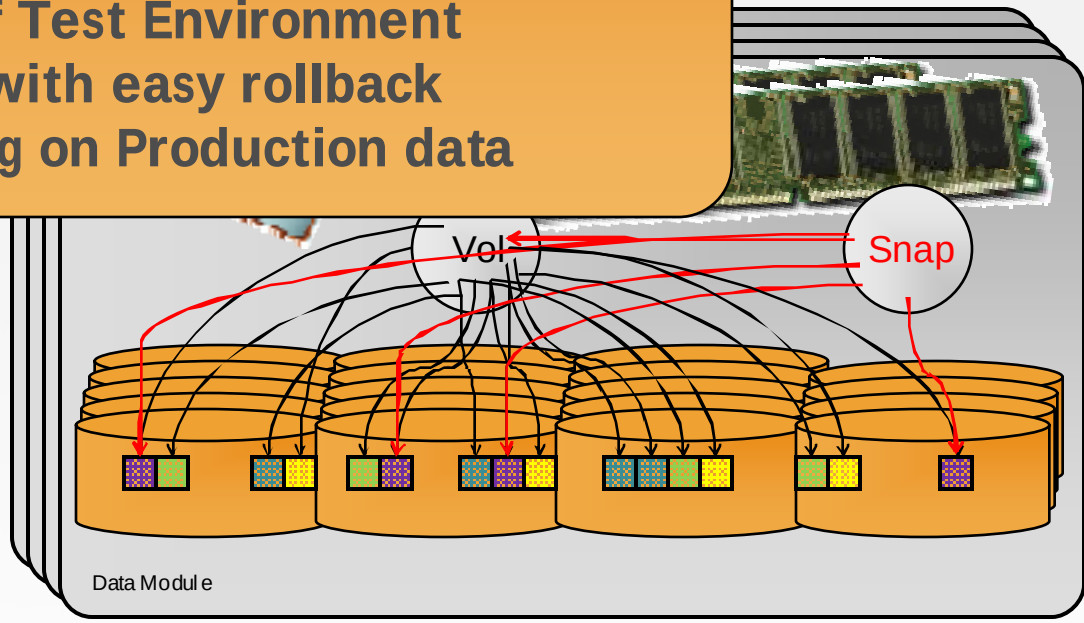
**High Performance, Unlimited SNAPs provide:**

- **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**

Distributed SNAP on each Server.
mory operations

as fast as
on volumes




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

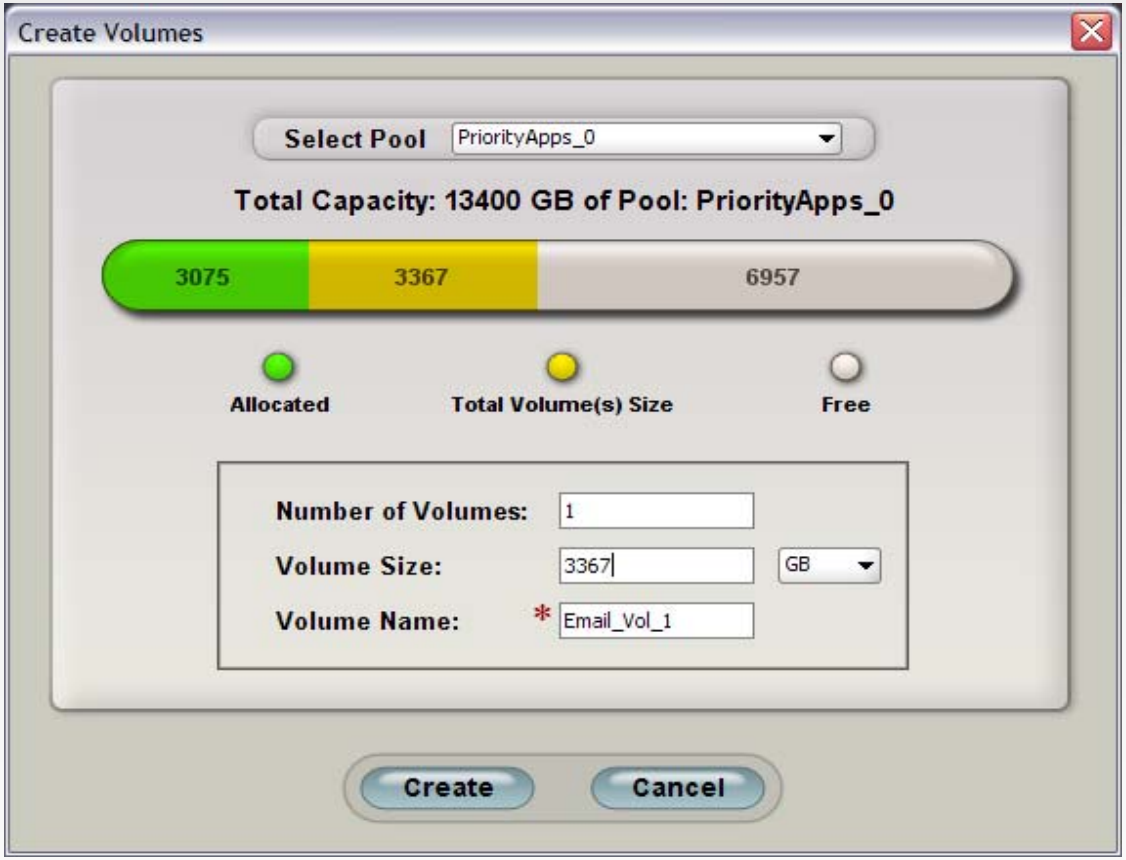
# … and More Tier-1 Functionality Built in to the Architecture

- **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

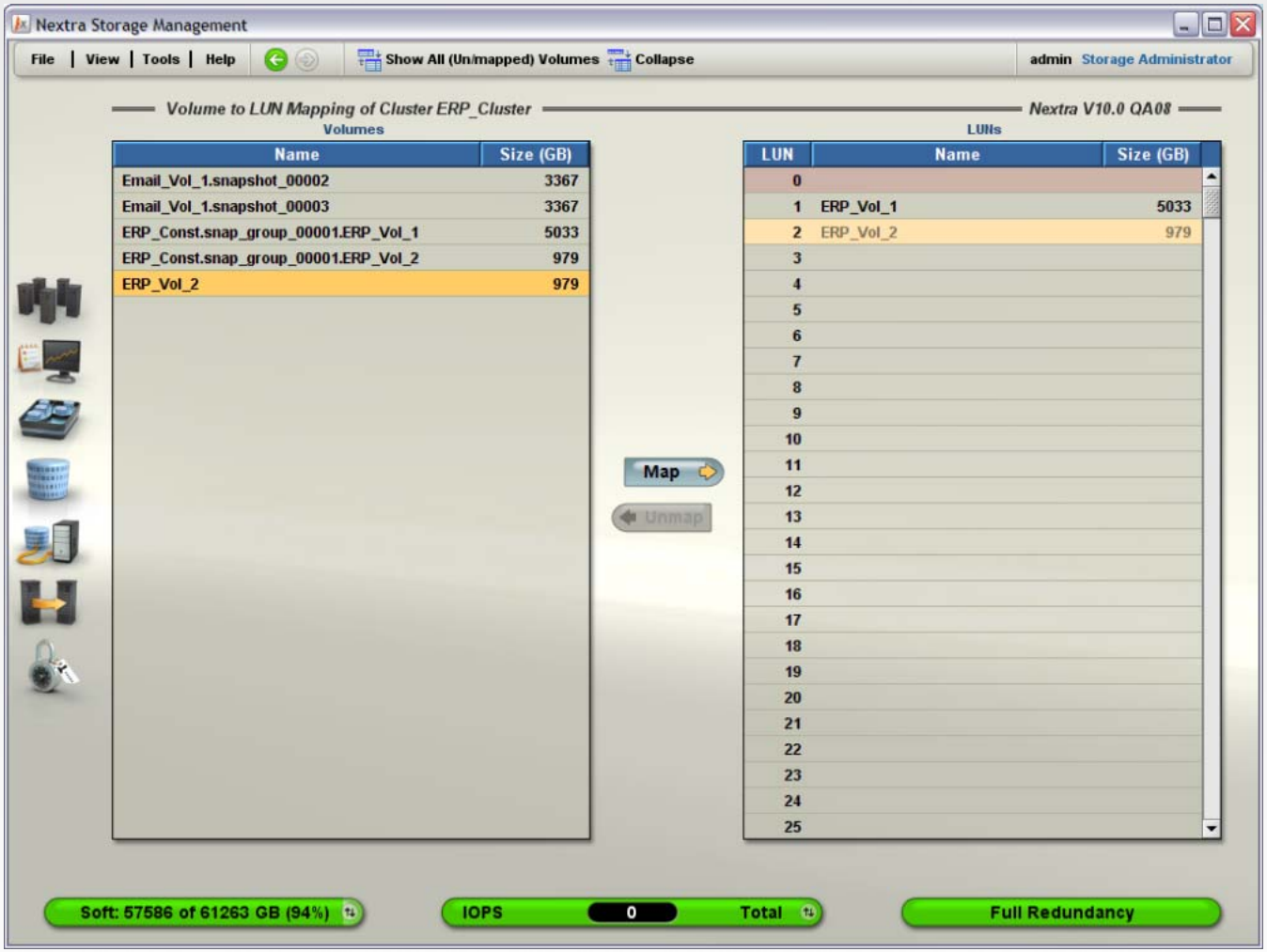
- Intuitive GUI (Java based) with Script Generator
- No dedicated management station
- Command Line Interface (CLI)
- XML over SSL
- Event management (SNMP)
- Complete Event Logging
- Events notification via email, SNMP and SMS
- Role based management:
  - Storage Admin
  - Application Admin
  - Operator

# IBM XIV Storage Simple Intuitive Management example: Creating a Volume

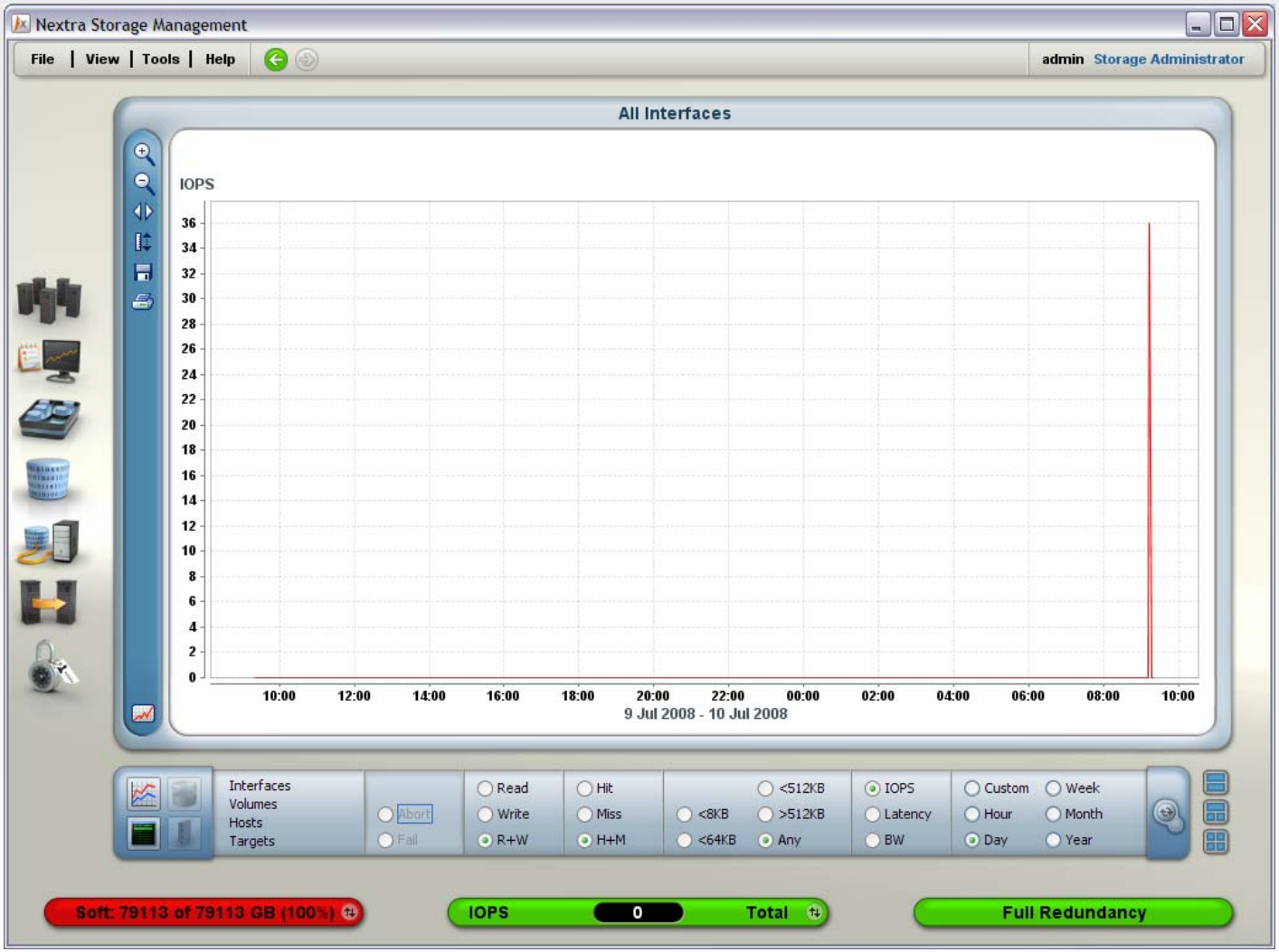



- Used capacity is always known !

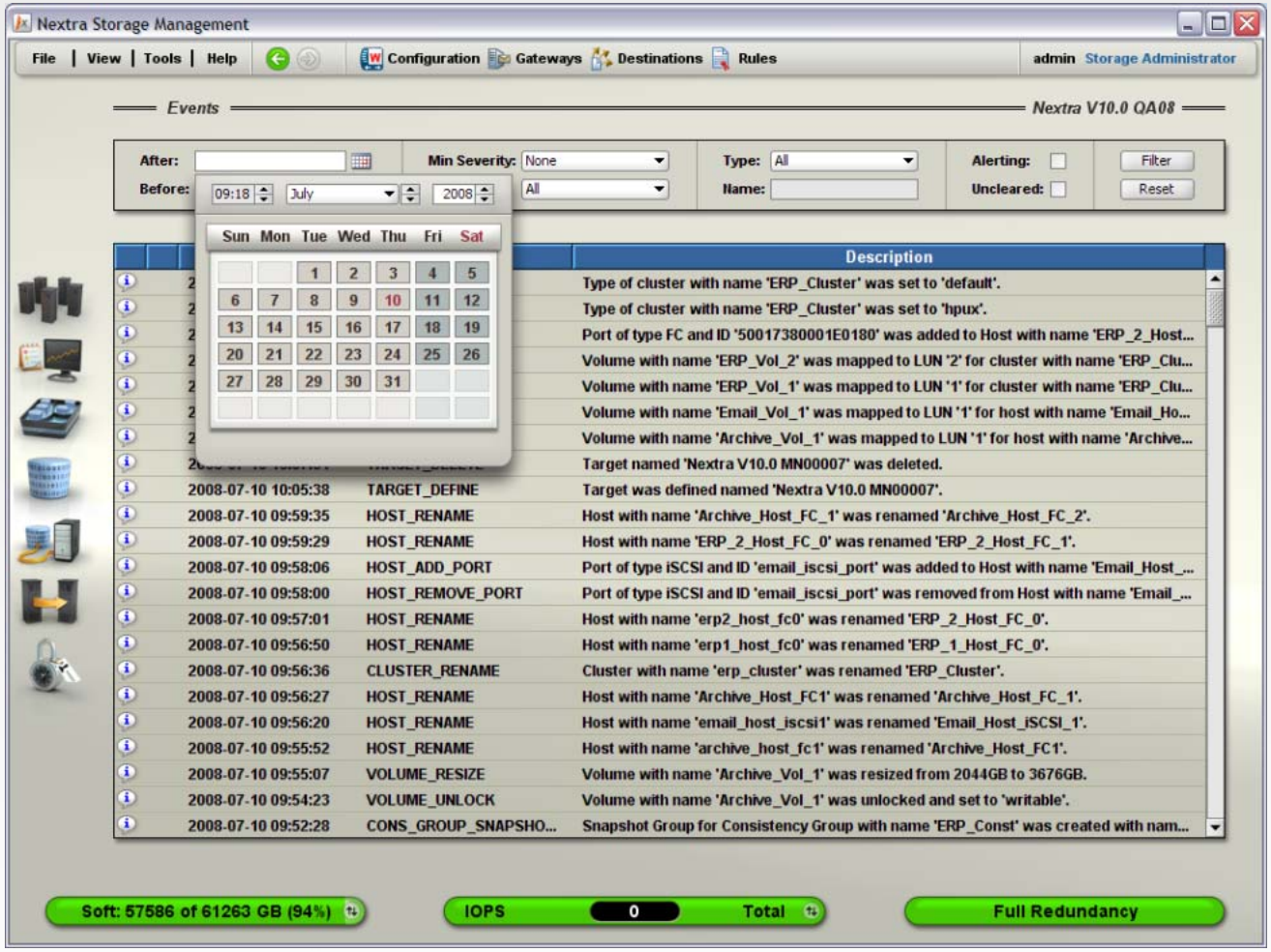
# IBM XIV Storage: Volume to LUN Mapping



# IBM XIV Storage: Monitoring



# IBM XIV Storage: Events Log

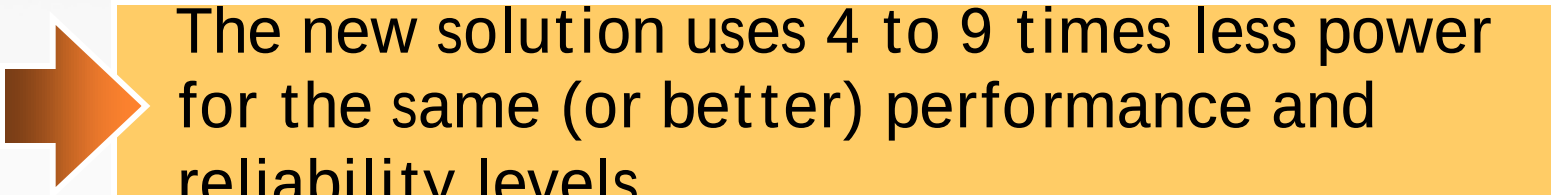


# System Power Usage

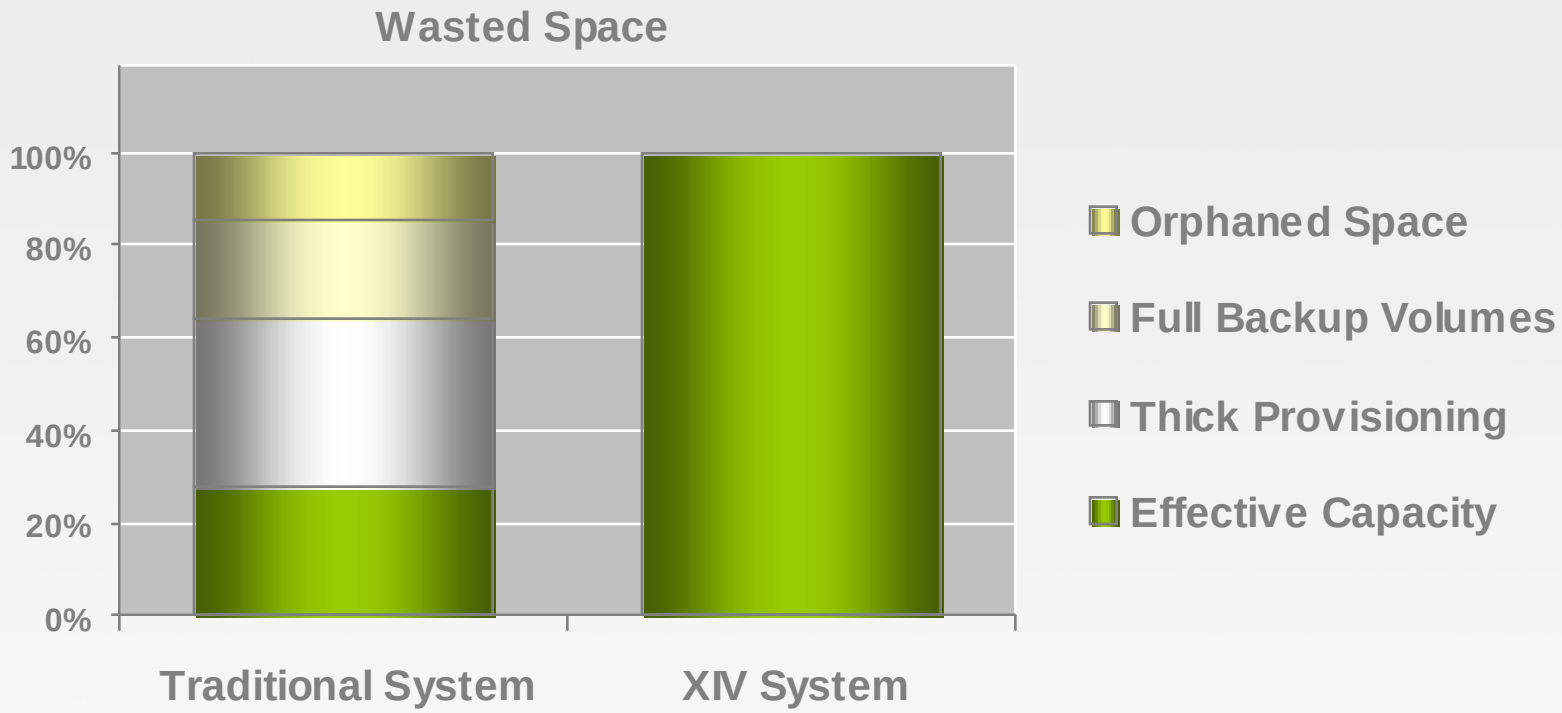
- Power consumption of a system comparable to XIV is 180–380W per raw TB
  - Typically using 146GB 15K rpm disks (380W per TB)
- Power consumption of an XIV rack is 7.7KW
  - 180TB raw capacity, 79TB net capacity
  - 42W per raw TB today
- Rack power consumption will not change much with 2TB disks
 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

# System Power Usage

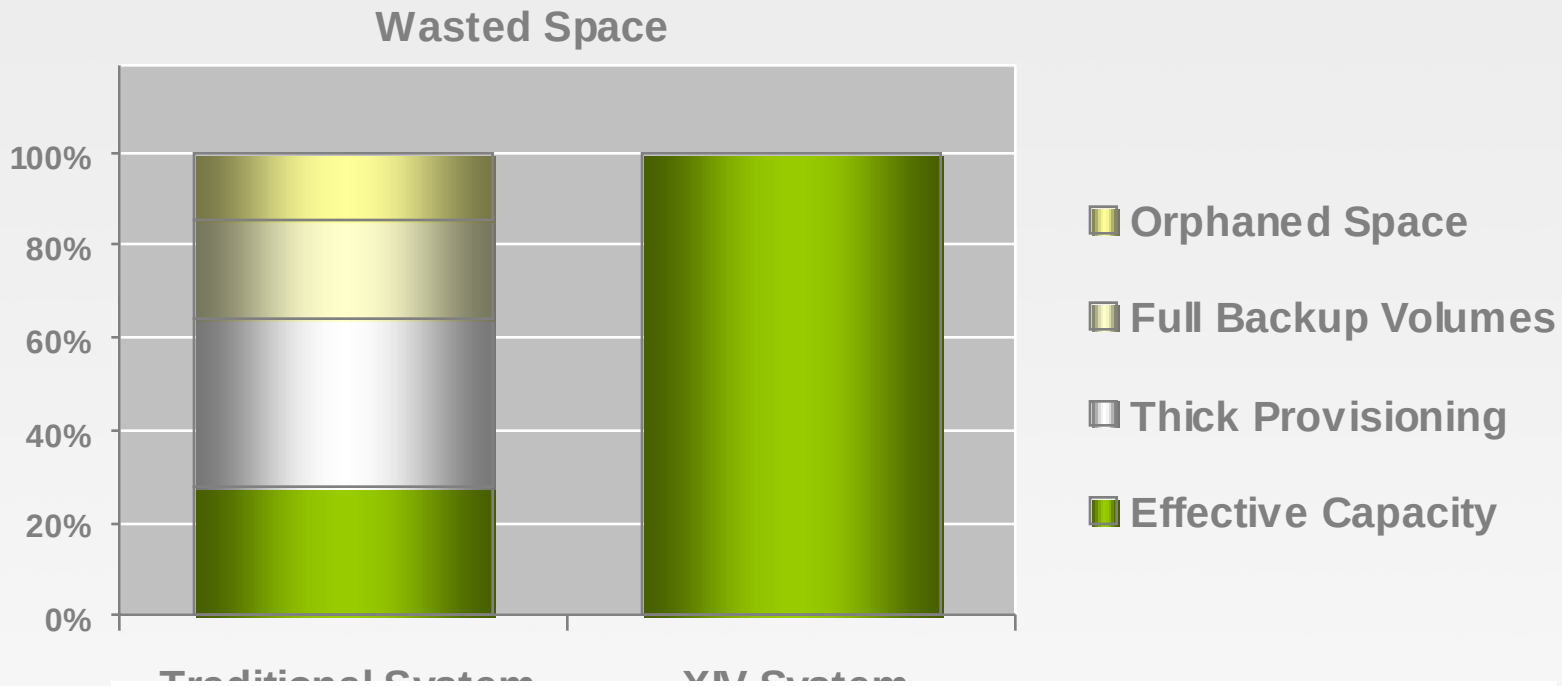
- Power consumption of a system comparable to XIV is 180–380W per raw TB
  - Typically using 146GB 15K rpm disks (380W per TB)
- Power consumption of an XIV rack is 7.7KW
  - 180TB raw capacity, 79TB net capacity
  - 42W per raw TB today
- Rack power consumption will not change much with 2TB disks



# 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

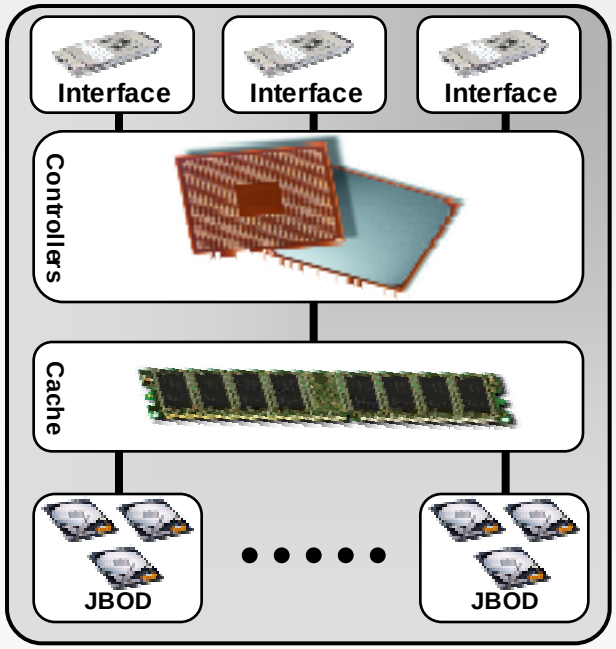
## Customer Problem

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<br>Snapshots are saved for a week<br>Can instantly return to any of the 28 snapshots | Efficient, differential snapshots |

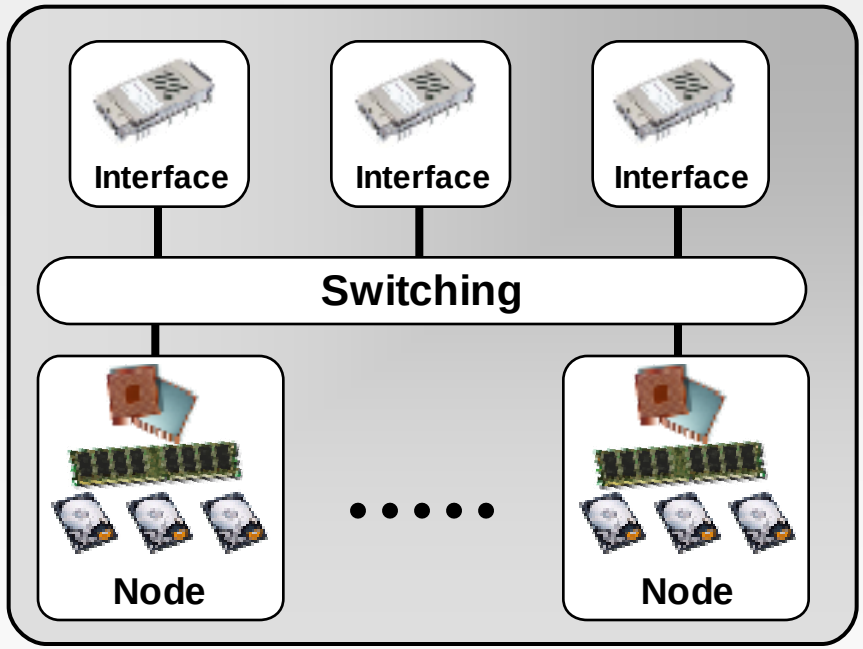
# A Novel Storage Architecture

- 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