Storage Industry Trends and Storage Clustering Keynote Address IEEE/NASA Mass Storage Conference April 2002

Matthew T. O'Keefe, Ph.D.
Founder and Chief Technical Officer
Sistina Software, Inc.
Minneapolis, Minnesota
okeefe@sistina.com

SISTINA
Infinitely Smarter Storage

Agenda

- Merrill Lynch/McKinsey Enterprise Storage Study
- Storage Clustering — Definition and Advantages
- Storage Clustering Applications
  - Compute Clusters, Edge Serving, Parallel Databases
- Questions?

The Storage Report – Customer Perspectives & Industry Evolution

Merrill Lynch

- Thomas Kraemer, James Berlino

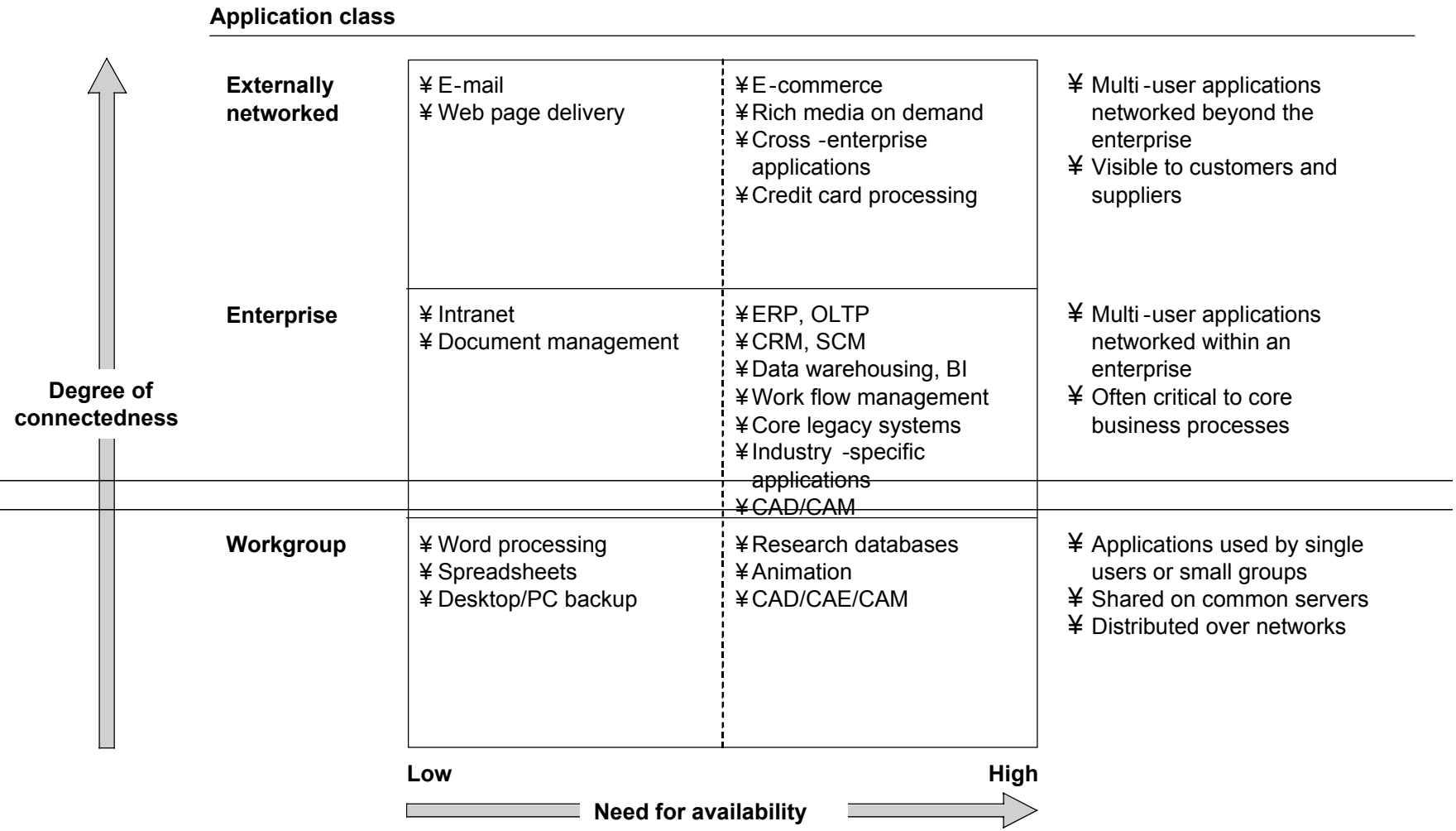
McKinsey & Company

- John Griffin, Doug Haynes, Thomas Herbig, Peter Stern, Alberto Torres

- Market surveys of several hundred storage industry customers on their storage needs
- In-depth interviews with storage customers across multiple industry sectors to understand capacity and performance needs, customer economics, purchasing processes, and decision criteria.

SISTINA
Infinitely Smarter Storage

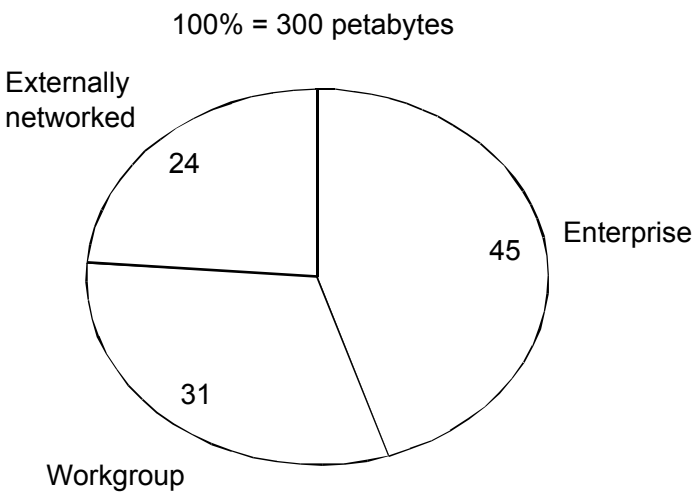
Applications Classes



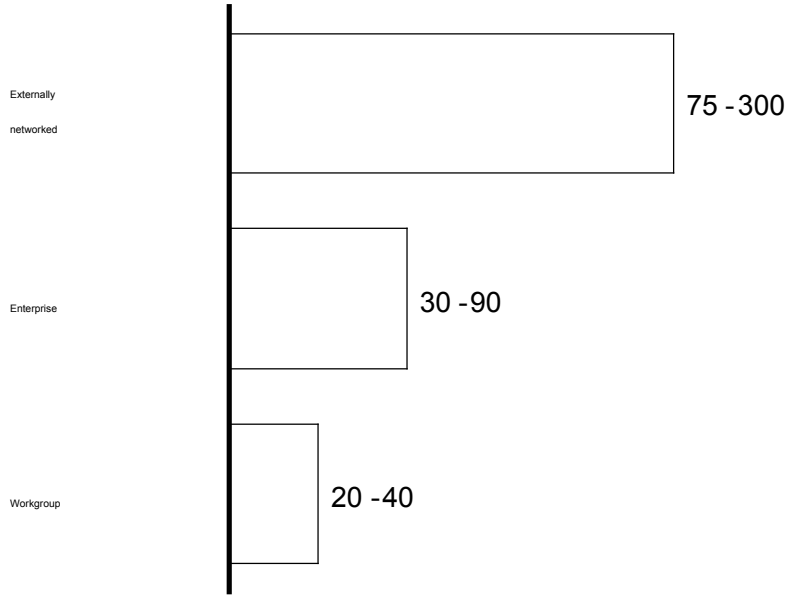
Source: Customer interviews; McKinsey and Merrill Lynch

Where is the Fastest Growth?

Breakdown of 2000 capacity shipped by application class
Percent



Annual capacity growth through 2003
Percent

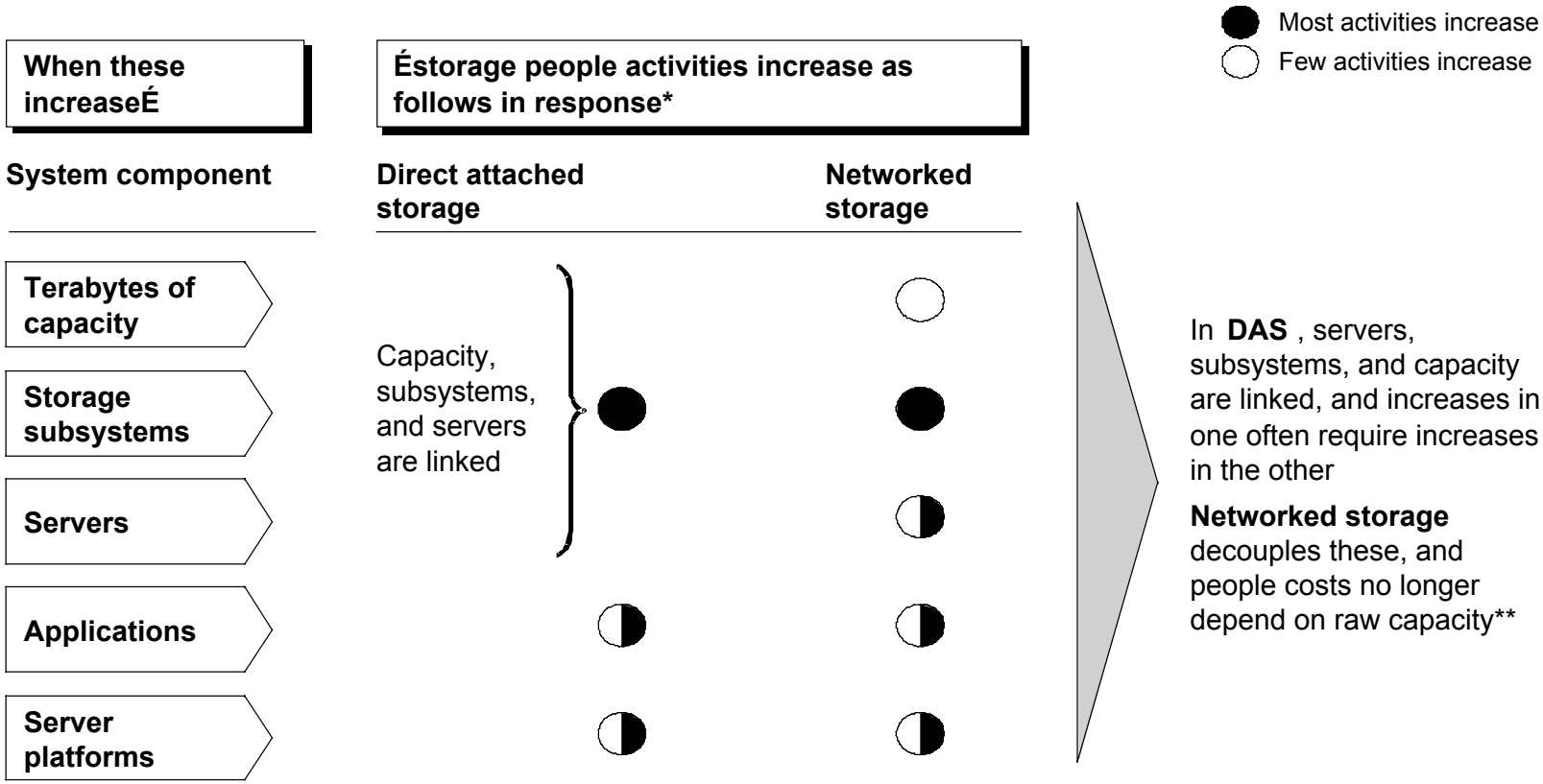


Source: Customer interviews; Merrill Lynch Reality Check survey of 110 C IOs; IDC; Forrester; McKinsey and Merrill Lynch

SISTINA
Infinitely Smarter Storage

People costs, capacity, and NAS/SAN

A NEW DIRECTION IN DATA STORAGE & MANAGEMENT



* Main activities are scaling, backing up, archiving, recovering, managing access, device and capacity monitoring, training, performance tuning. People costs often scale less ste eply than system components due to economies of scale

** When capacity growth exceeds 100% a year, they link terabytes to subsystems and again increase people costs

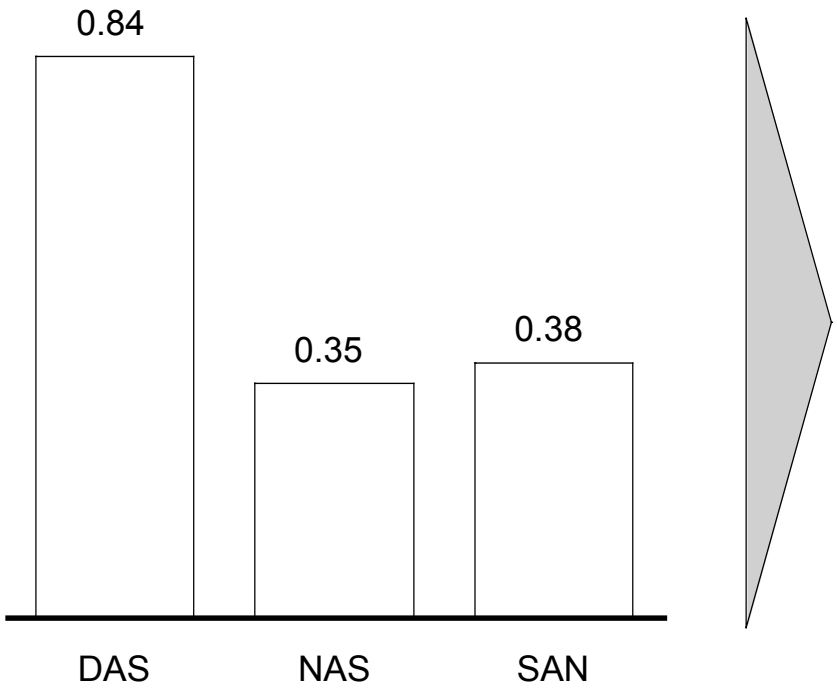
Source: Gartner; customer interviews; McKinsey and Merrill Lynch

SISTINA

Infinitely Smarter Storage

Network Storage has Better TCO than DAS

3-year TCO by storage architecture*
$ per megabyte of user data



- Cost saving of SAN and NAS driven by:
  - Improved disk utilization
  - Centralized management
  - Tape drive consolidation
- NAS does not have SAN network and installation charges

* Based on 2TB of user data; as of March 2001

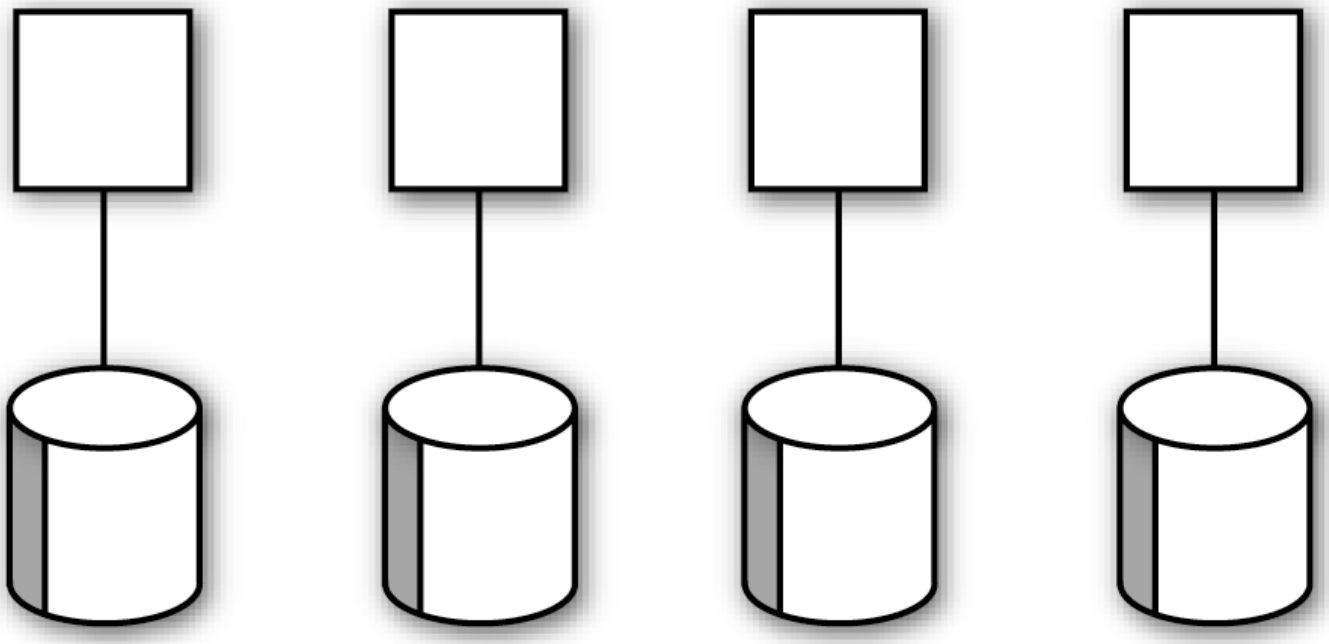
Source: Customer interviews; expert interviews; McKinsey and Merrill Lyn ch

SISTINA
Infinitely Smarter Storage

Local Disk Approach is Inefficient

UNDERUTILIZATION

Inefficient space allocation resulting from local disk approach

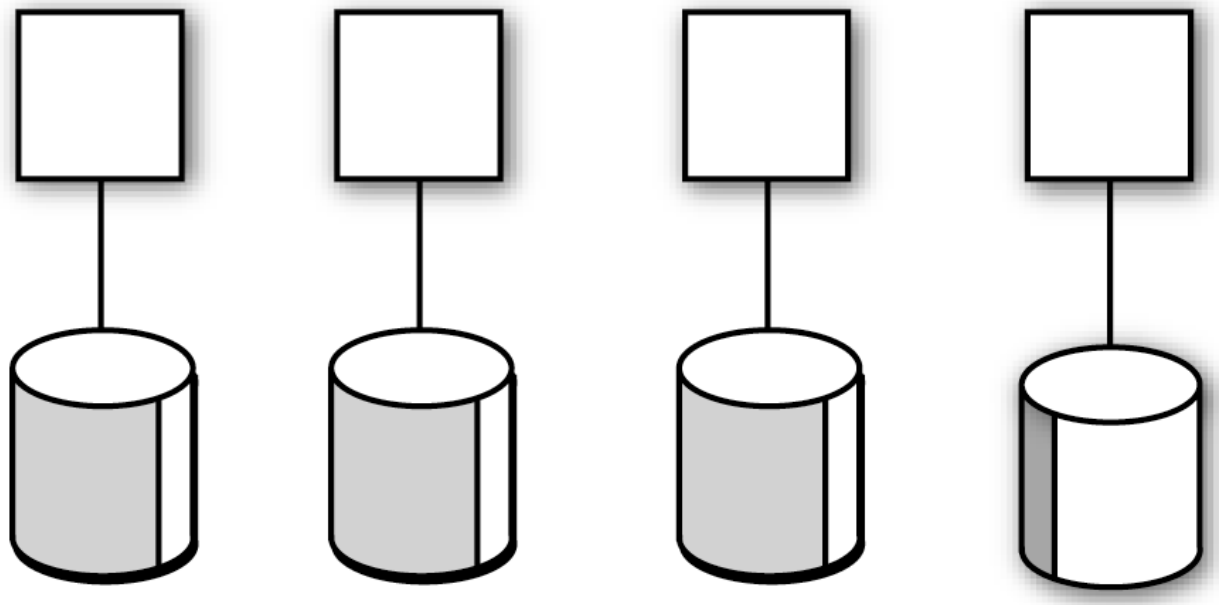


Each local disk only 10% utilized

SISTINA
Infinitely Smarter Storage

Lack of Balanced Usage Across Local Disks

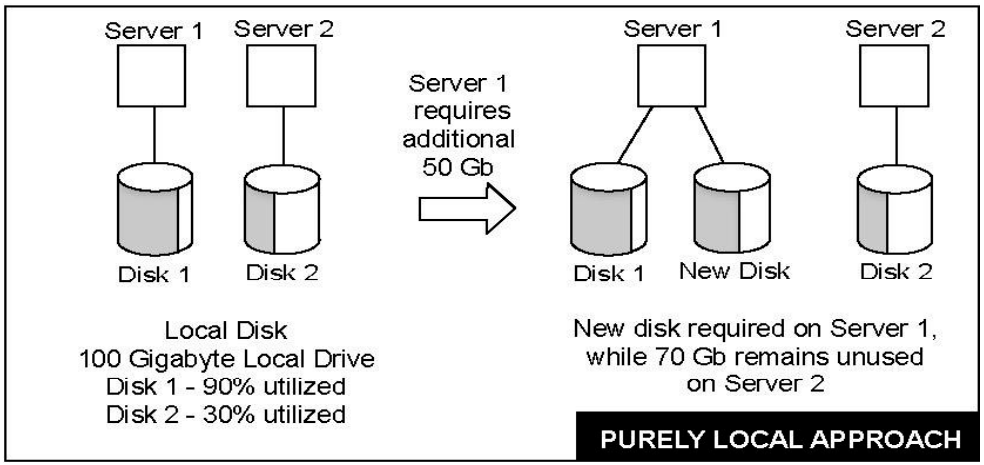
MIXED UTILIZATION RATIO

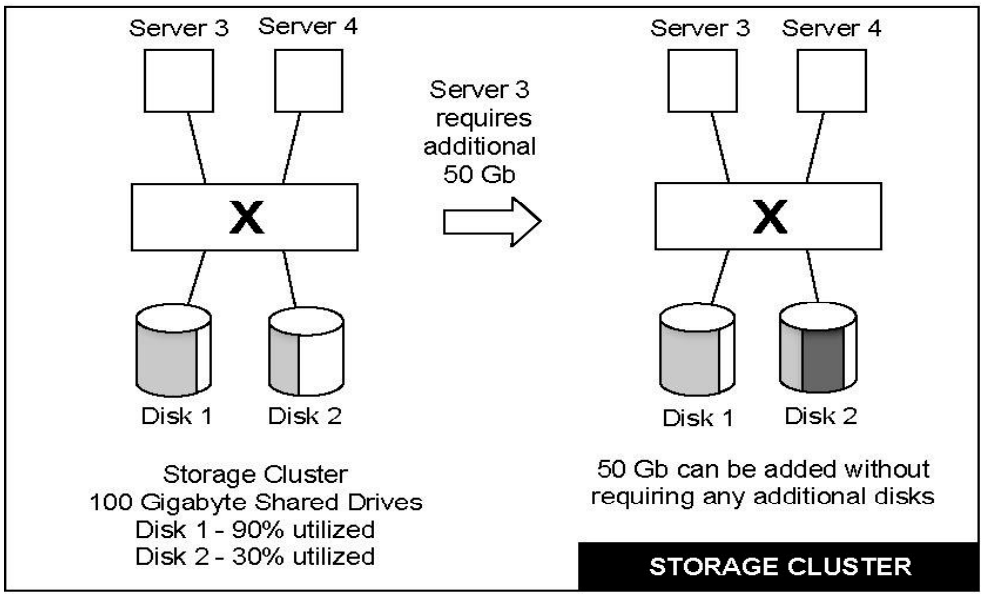


Three local disks 90% utilized, one local disk 10% utilized

SISTINA
Infinitely Smarter Storage

DYNAMIC SPACE UTILIZATION WITH STORAGE CLUSTERING VS. A PURELY LOCAL DISK APPROACH

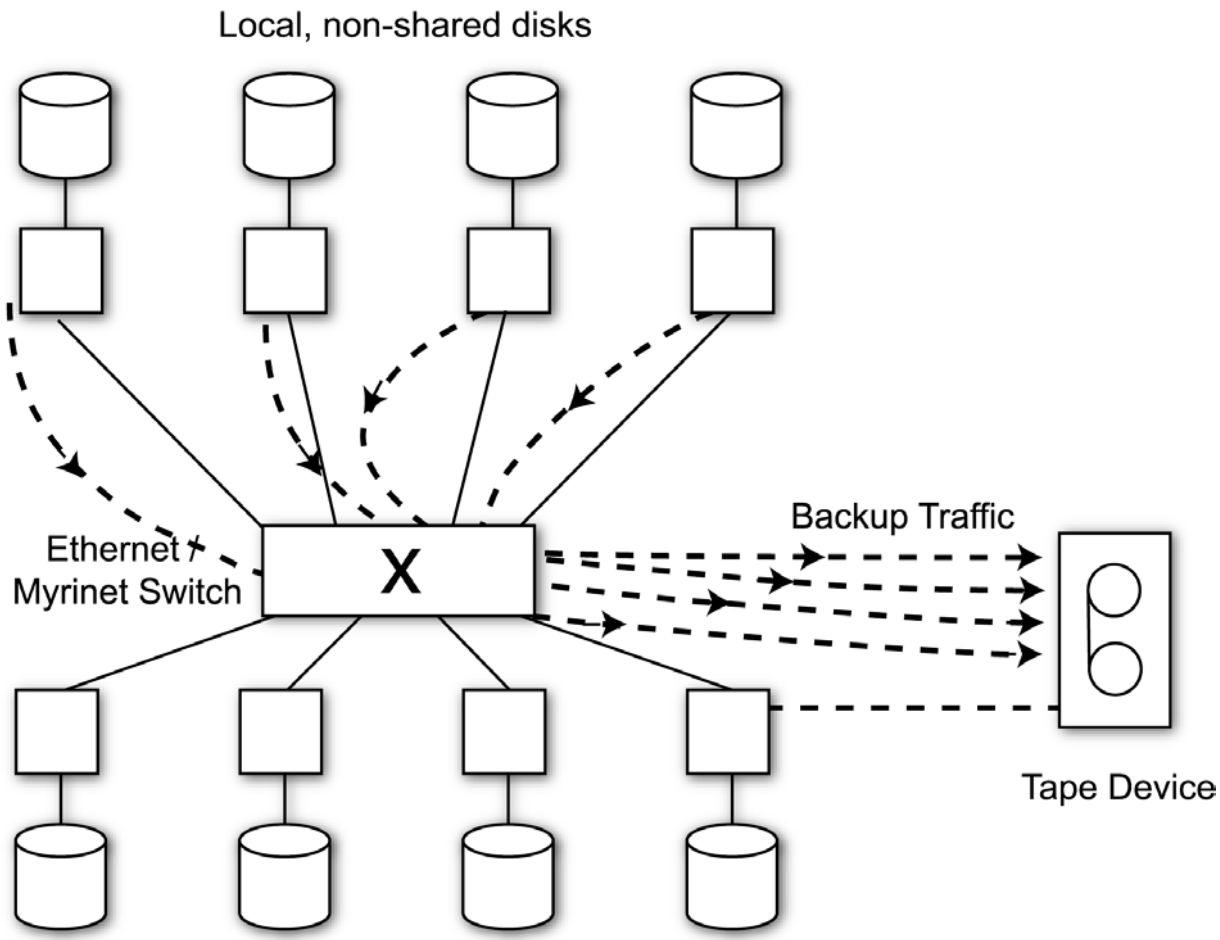




- Without shared storage, in this example a new server-attached disk is required
- With shared storage, no new disk is required, can efficiently utilize existing storage

SISTINA
Infinitely Smarter Storage

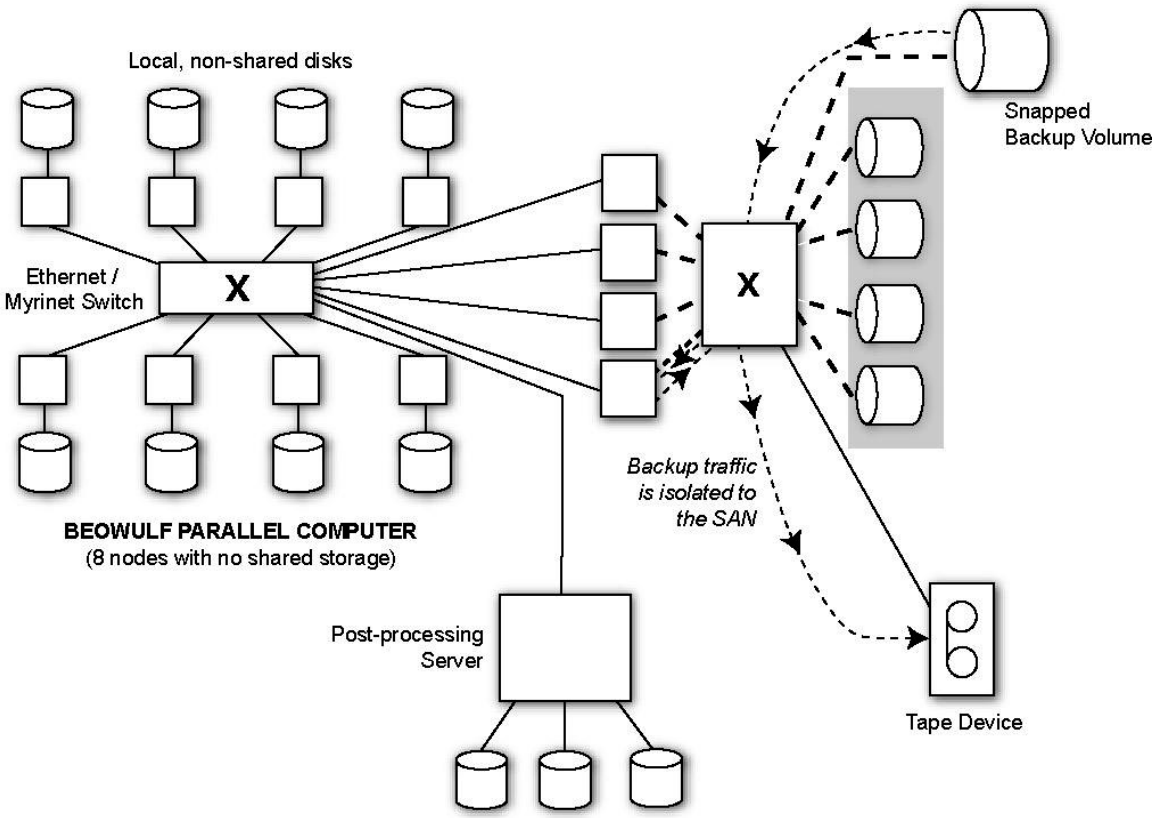
Backup Traffic Can Saturate a Compute Network



BEOWULF PARALLEL COMPUTER
(8 nodes with no shared storage)

SISTINA
Infinitely Smarter Storage

Backup Traffic Off-Loaded to Storage Cluster



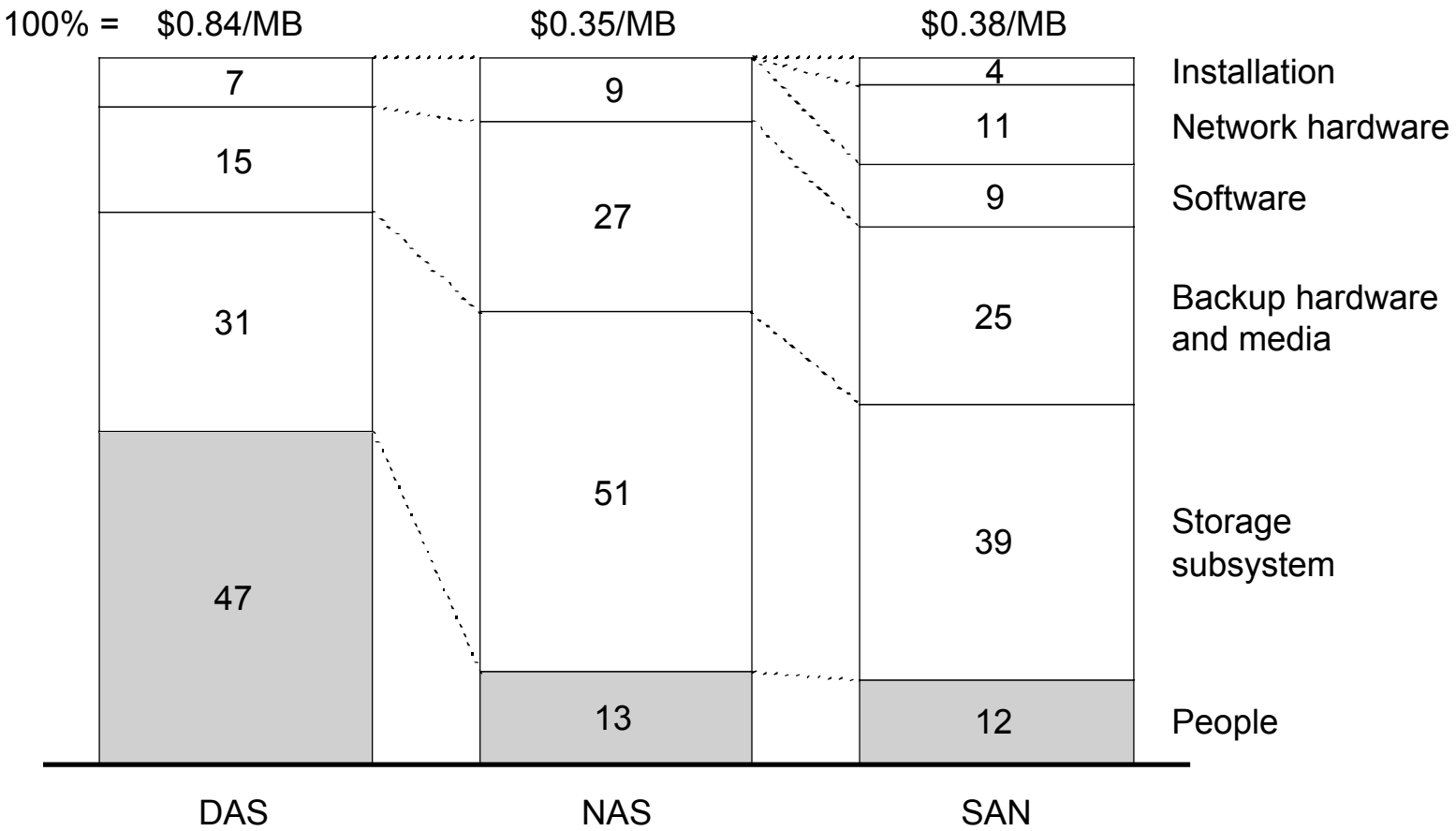
- Volume snapshot of GFS file system made first
- Read-only GFS mount then backed up across SAN
- Can be done on-line without disrupting computations
- Off-line storage consolidation

SISTINA
Infinitely Smarter Storage

Relative Cost Components

Cost breakdown of storage architectures*

Percent



* Based on 2TB of user data and 10 servers; as of March 2001; see Appendix E for assumptions

Source: Customer interviews; expert interviews; McKinsey and Merrill Lynch

Storage Decision Drivers

- Three performance characteristics are the primary drivers of storage decisions:
  - ease of management
  - availability
  - throughput
- Reliability and scalability are not the primary differentiators.
  - reliability is important, but today's storage systems are sufficiently reliable that this is no longer a differentiator

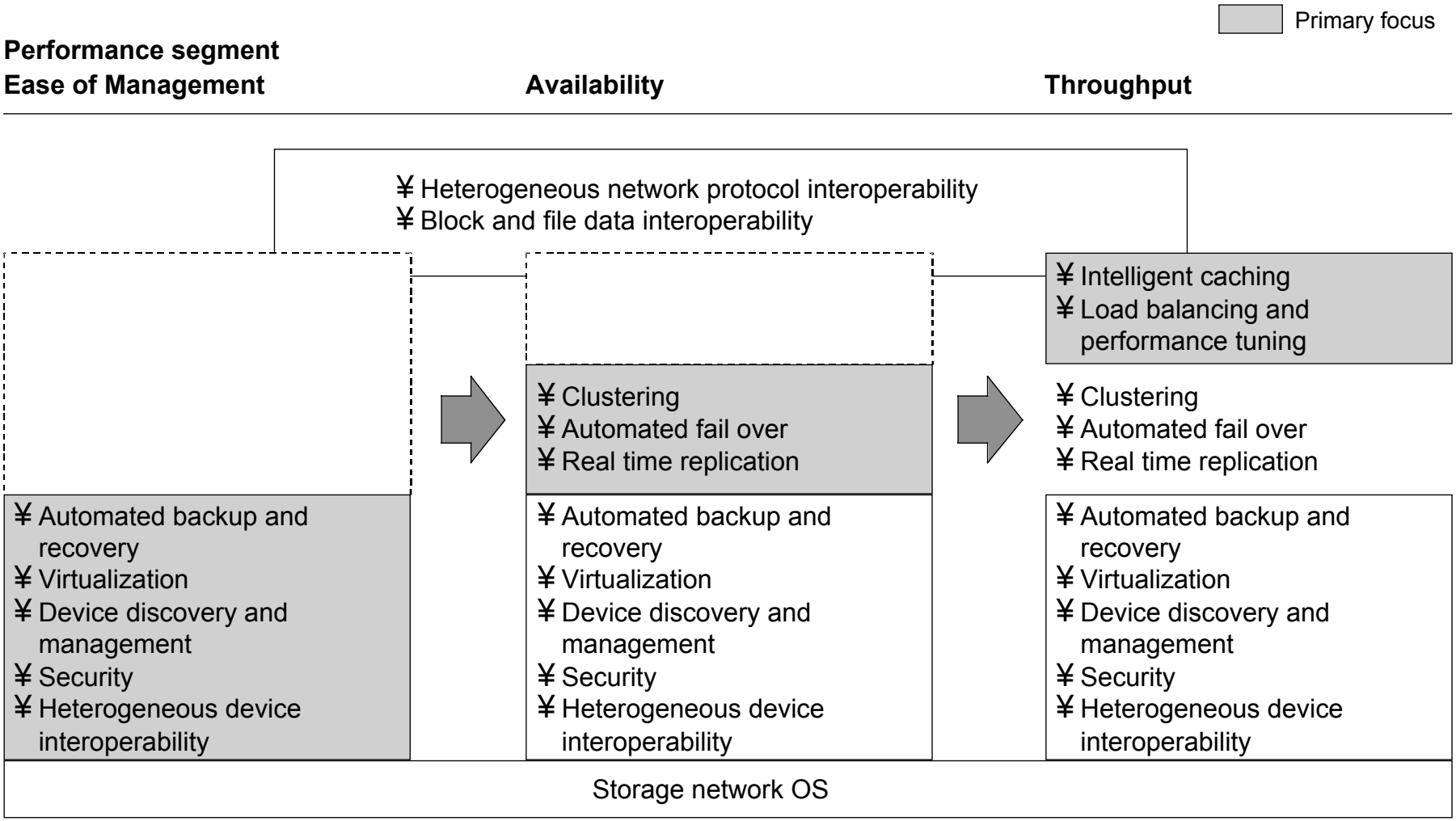
Storage Decision Drives

- Scalability is no longer a major concern for most customers for two reasons:
  - improvements in storage density are accelerating more rapidly than the growth in demand for storage
  - networked storage greatly simplifies scaling

- Vendors should focus on the next set of customer challenges:
  - ease of management: NAS
  - availability: SAN
  - throughput: storage clustering (SAN + CFS + NAS)

SISTINA
Infinitely Smarter Storage

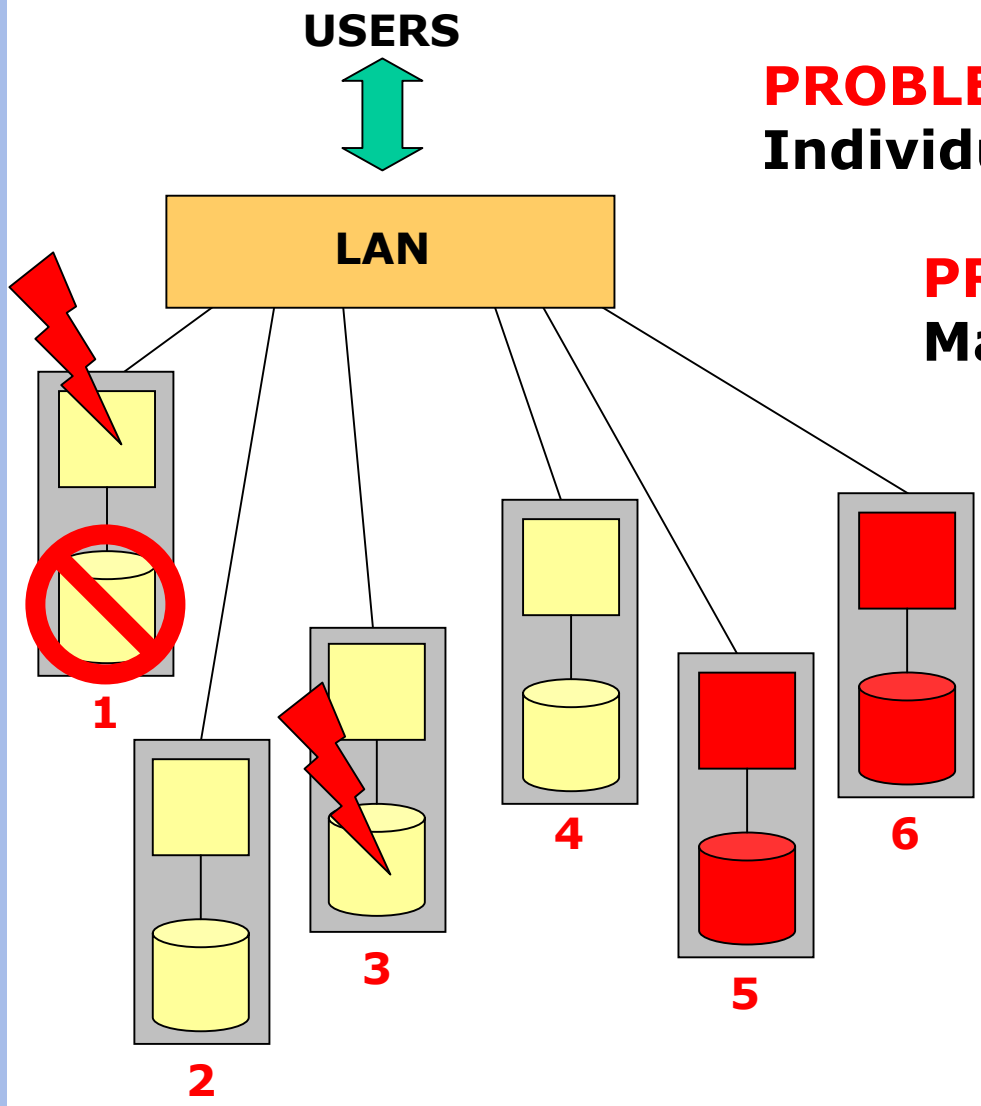
Requirements by Performance Segment



Source: Customer interviews; McKinsey and Merrill Lynch

SISTINA
Infinitely Smarter Storage

Today's Infrastructure



PROBLEM: Servers Attach Individually to Disparate Disks

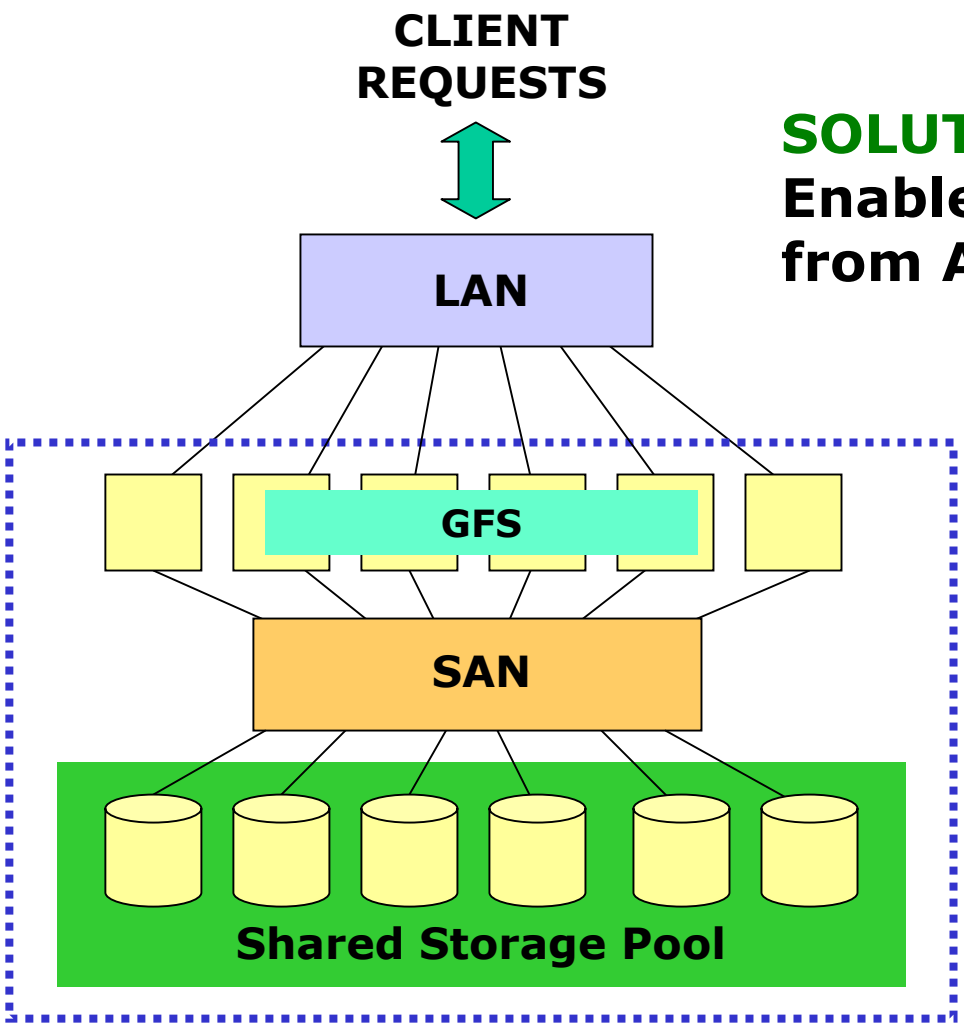
PROBLEM: Many Separate Management Domains

PROBLEM: Inefficient Scalability

PROBLEM: Data Accessibility

SISTINA
Infinitely Smarter Storage

Storage Clustering



SOLUTION: Storage Clustering Enables a SAN to Create a Path from All Servers to All Disks…

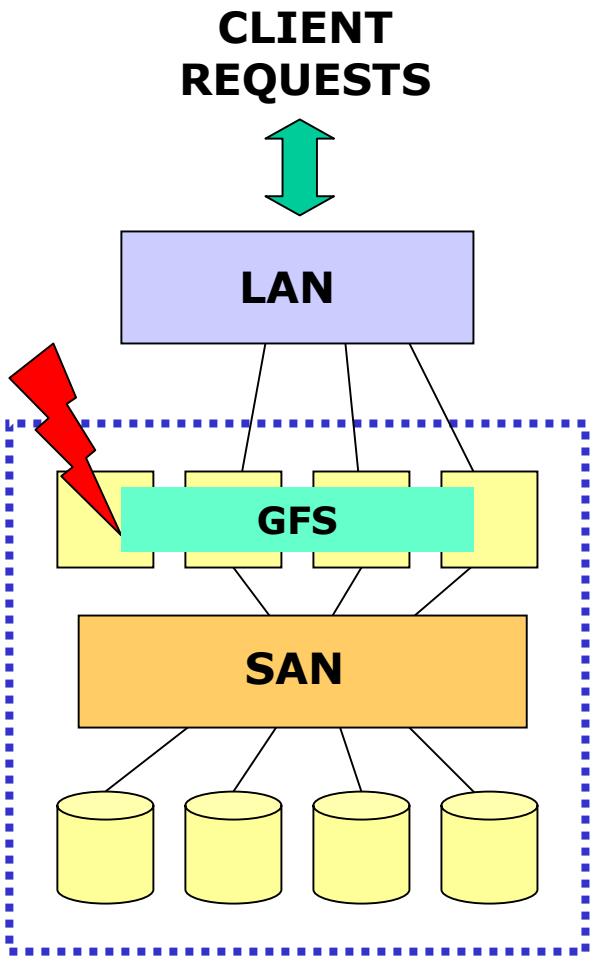
SOLUTION: One Management Domain with Shared Storage

SOLUTION: Efficient Server Scalability

SOLUTION: Efficient Storage Scalability

SISTINA
Infinitely Smarter Storage

Storage Clustering



Major Advantage:

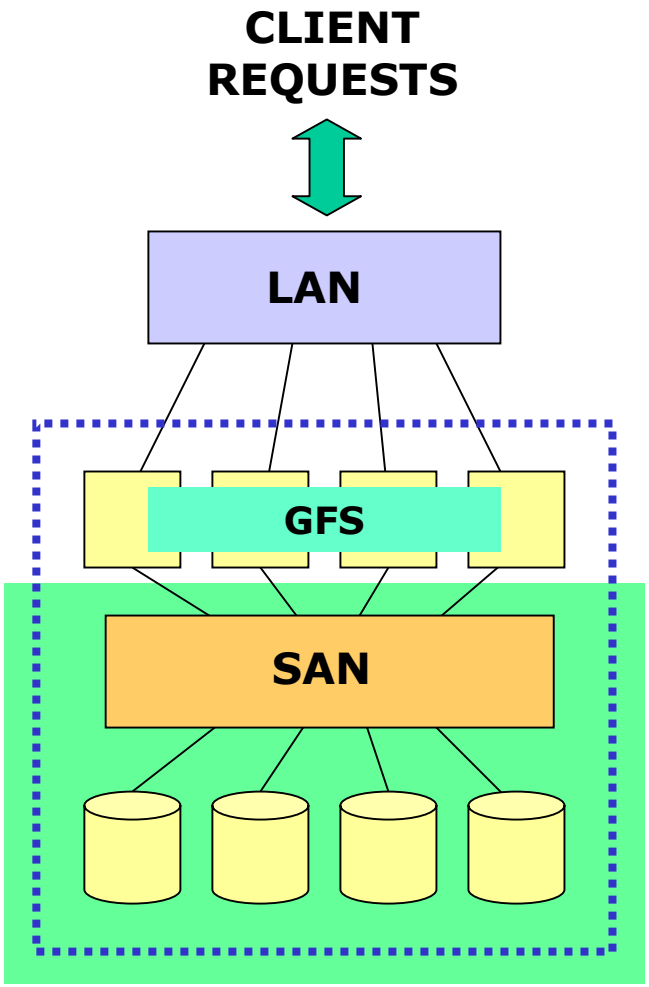
Constant Accessibility

In the case of a server failure the load is balanced across the remaining servers without access interruption.

The failed server can be brought back on-line while the system remains running.

SISTINA
Infinitely Smarter Storage

Storage Clustering



Major Advantage:

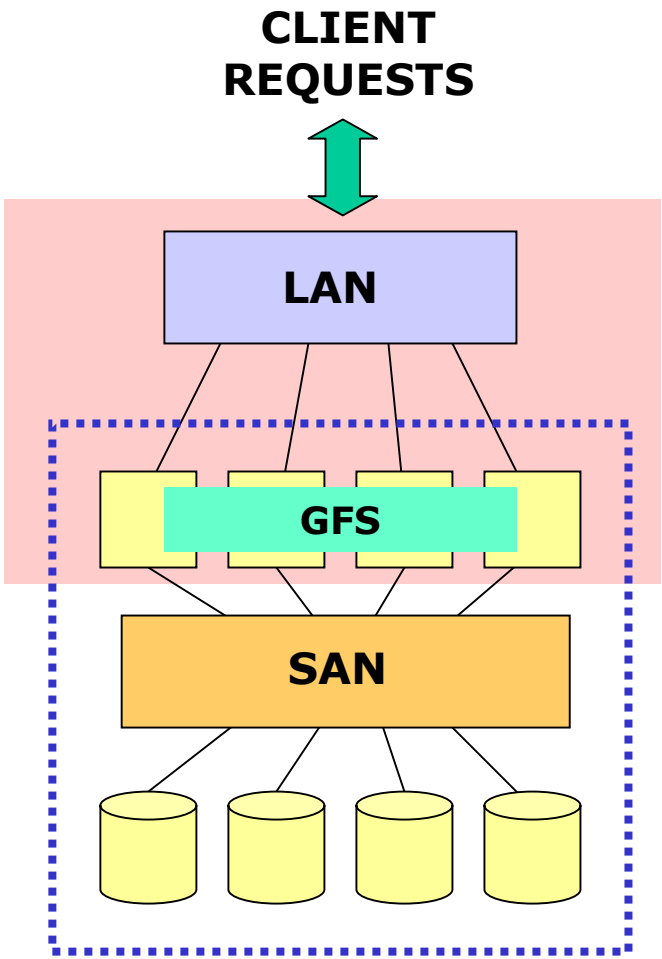
Supports Storage Virtualization

Virtualization allows the aggregation and management of blocks.

A cluster file system completes the virtualization concept by enabling data sharing.

SISTINA
Infinitely Smarter Storage

Storage Clustering



Major Advantage:

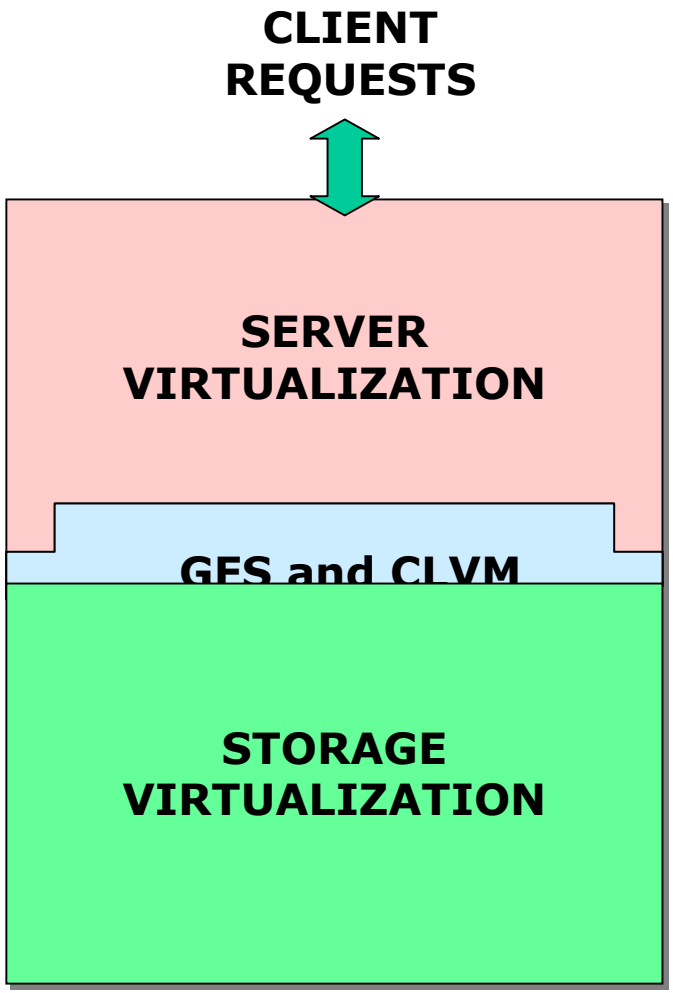
Supports Server Virtualization

A cluster file system and SAN enables the evolution of clustered applications beyond pair-based, failover software.

Storage clustering technology reduces the expense and complication of HA software.

SISTINA
Infinitely Smarter Storage

A Cluster File System as Middleware

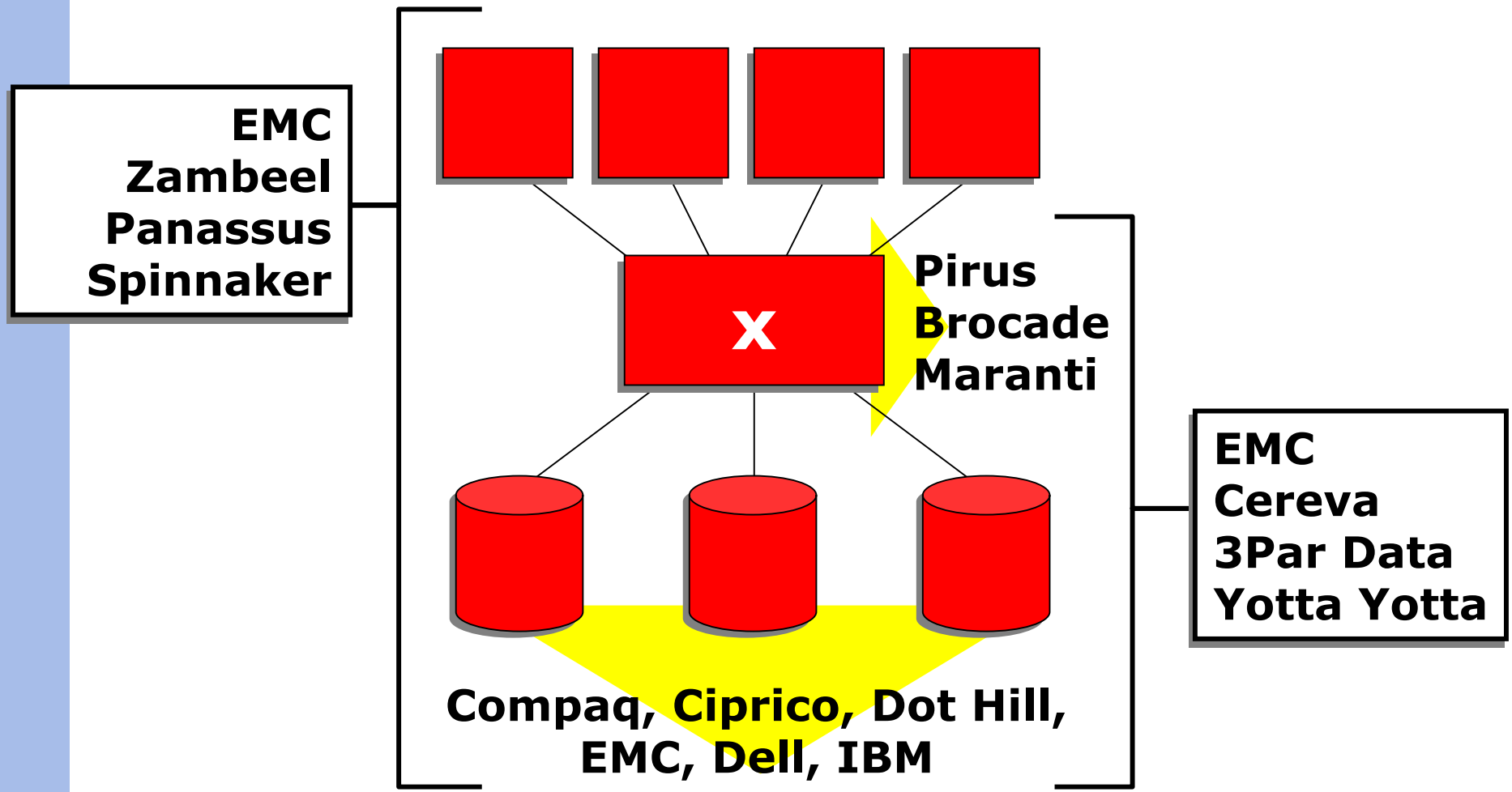


Major Advantage:

A Cluster File System and Cluster LVM Enables Server and Storage Virtualization to Operate Together

SISTINA
Infinitely Smarter Storage

Lots of Innovation

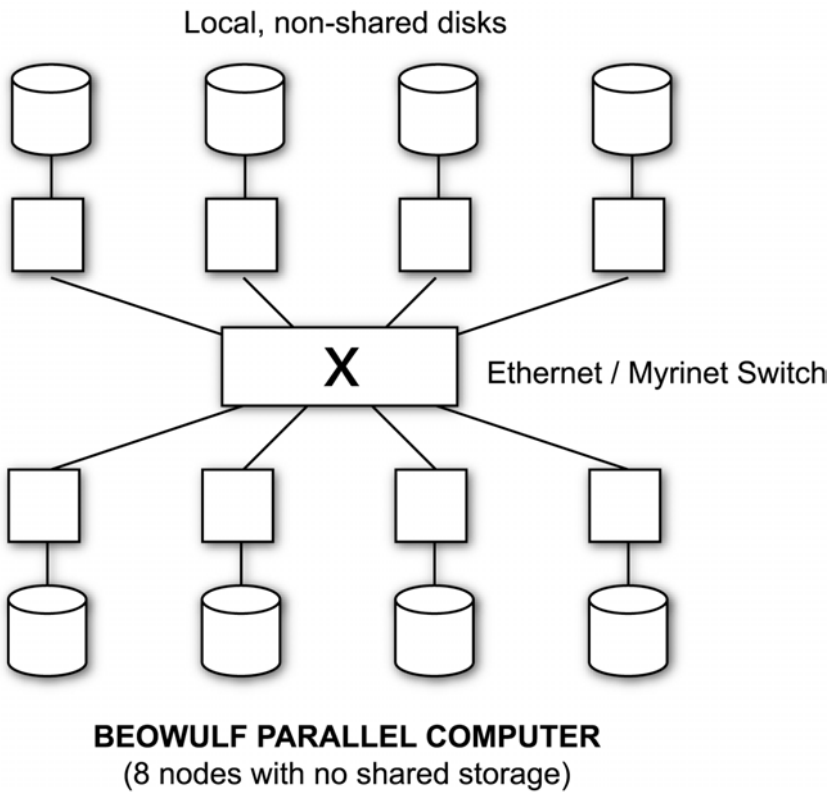


SISTINA
Infinitely Smarter Storage

Storage Clustering Applications

- Compute Clusters
- Edge Serving
- Parallel Database Serving

Compute Clusters

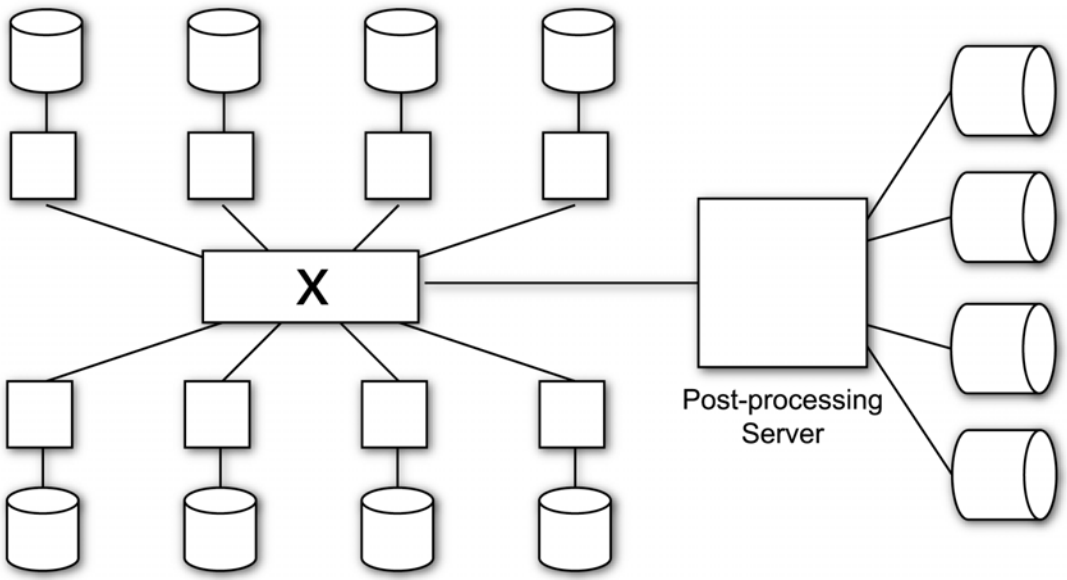


BEOWULF PARALLEL COMPUTER
(8 nodes with no shared storage)

- Compute clusters becoming increasingly popular for HPC
- Basic concept: commodity processors, network, and disks make HPC affordable and scalable
- Hard to program, no fault-tolerance, but cheap

SISTINA
Infinitely Smarter Storage

Pre- and Post-Processing

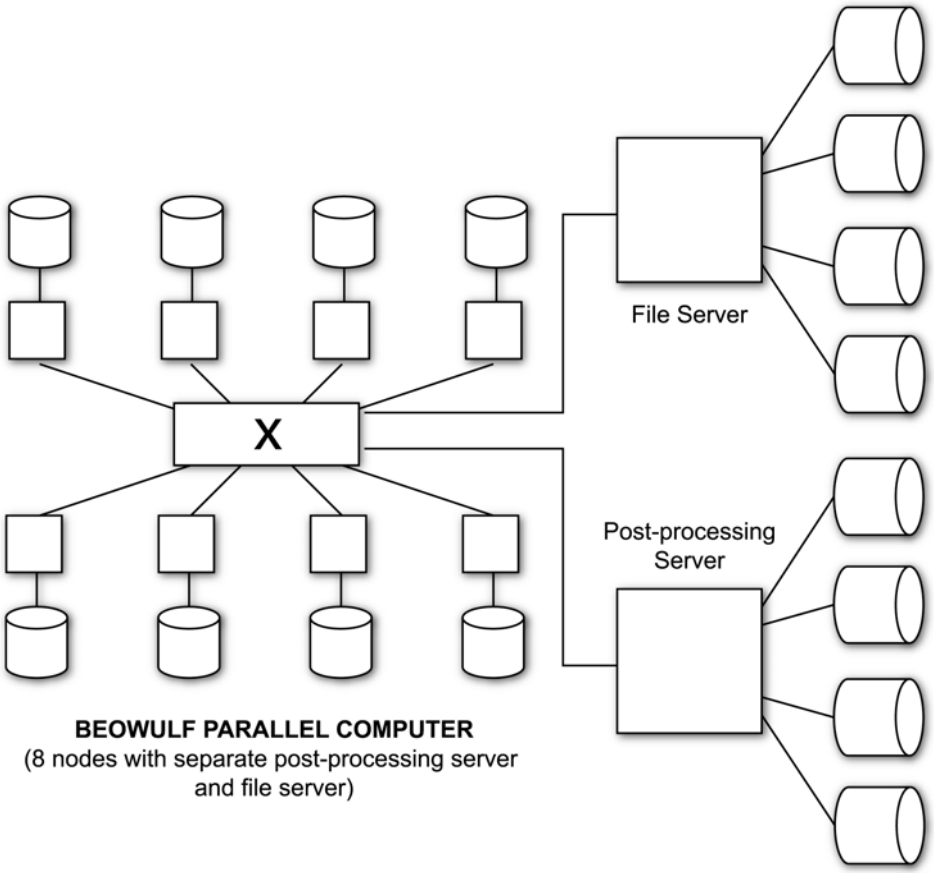


BEOWULF PARALLEL COMPUTER
(8 nodes with separate post-processing server)

- Pre- and Post-processing servers often used to process computed data
- ftp or NFS used to get data from each compute node
- Or the data is just ignored

SISTINA
Infinitely Smarter Storage

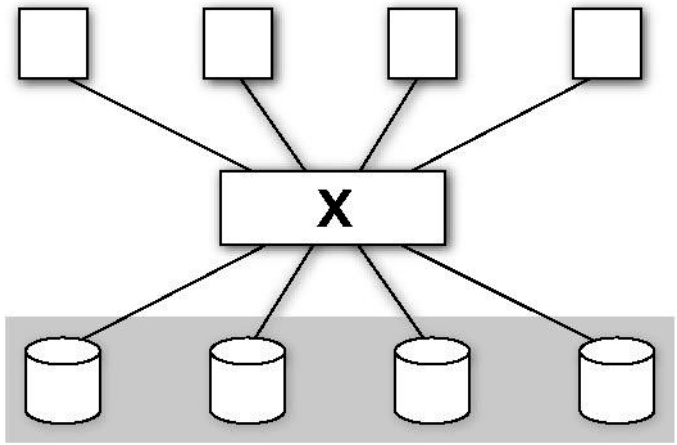
Sharing with One File Server



- NFS file server can be used to share data between compute cluster and outside processing nodes
- Results in replicated data: on both cluster and the file server
- Slow, does not scale

SISTINA
Infinitely Smarter Storage

GFS-Based Storage Clusters



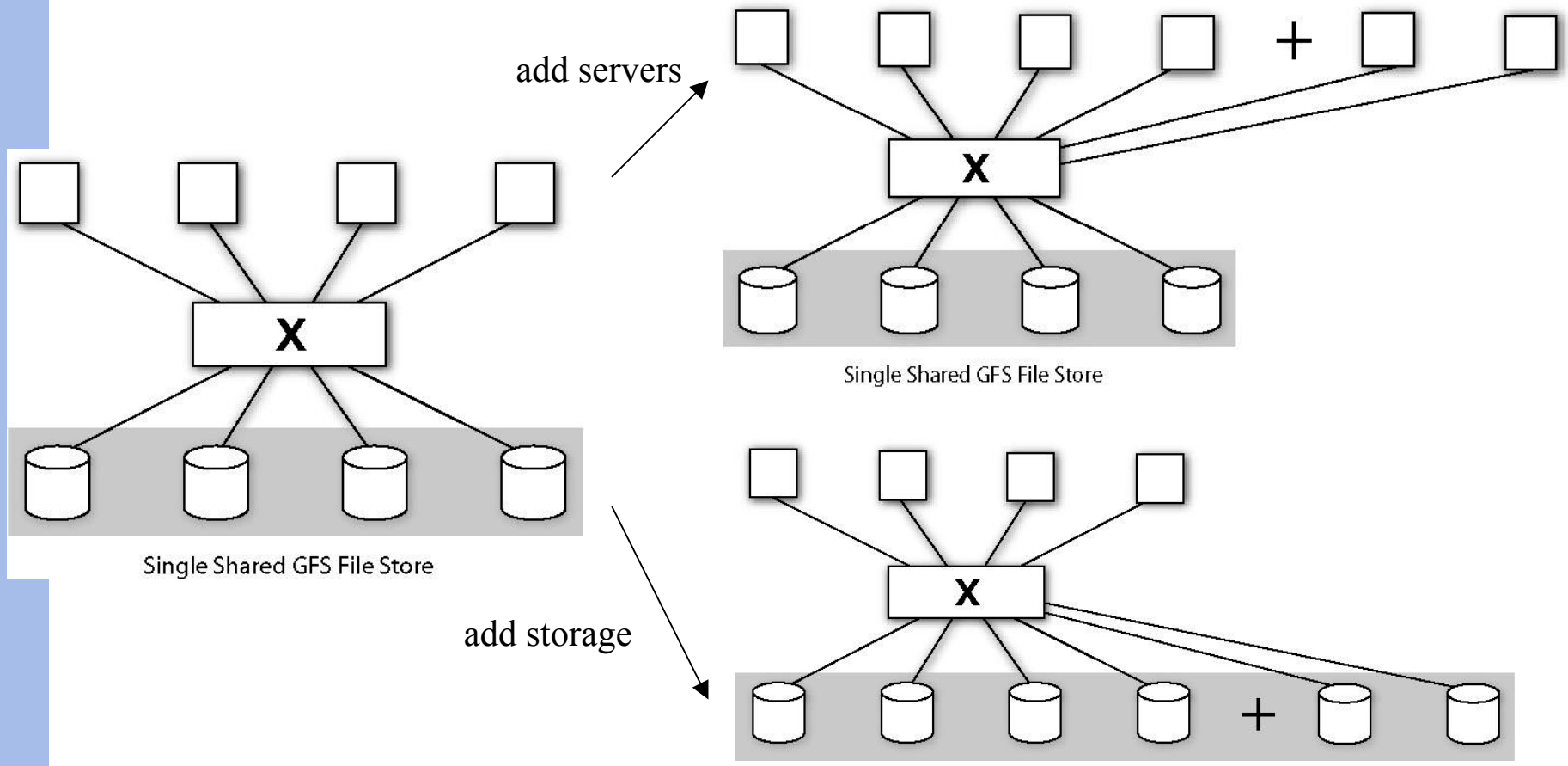
Single Shared GFS File Store

- Storage cluster consists of
  - •storage area network
  - •shared storage devices,
  - •cluster file system/volume manager running on the servers
- Sistina's GFS and CLVM: cluster file system for Linux: can be used to build a Linux-based storage cluster

SISTINA
Infinitely Smarter Storage

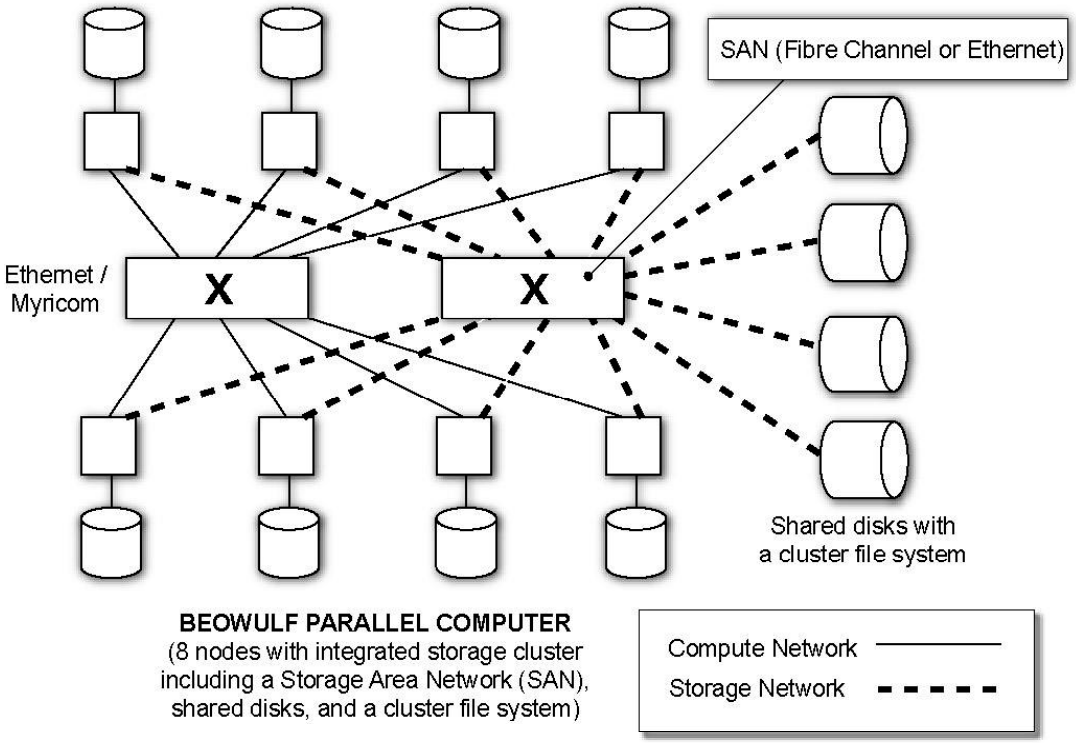
Storage Cluster Scalability

- Scale either servers or storage



SISTINA
Infinitely Smarter Storage

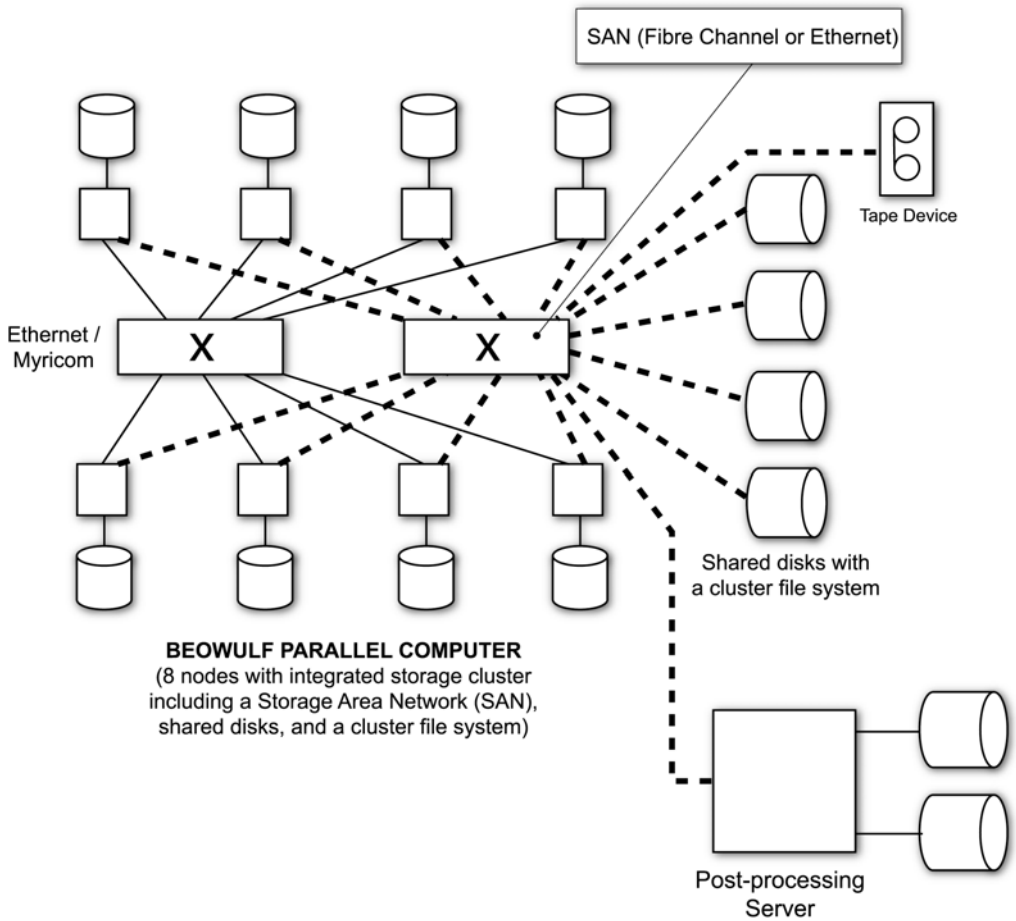
Compute Cluster with Integrated Storage Cluster



- Storage cluster: servers sharing storage with a cluster file system over a SAN
- Implies set of shared storage devices
- SAN with FC or IP

SISTINA
Infinitely Smarter Storage

Post-Processing Server Can Use Shared Storage



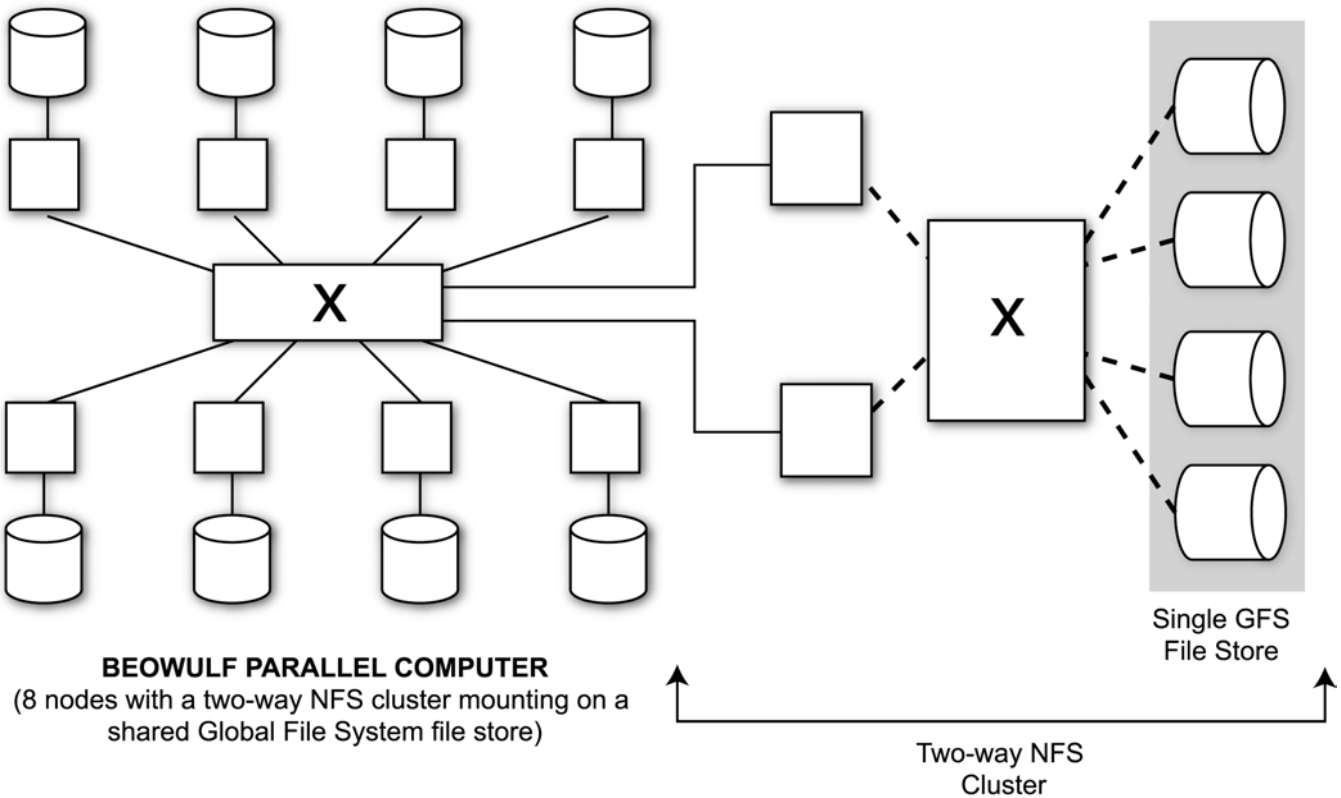
- Pre- or Post-processing hardware can be attached to the storage cluster
- One image of data between the compute cluster and other servers
- Consolidates storage into single, easily managed pool — avoids replication

SISTINA
Infinitely Smarter Storage

GFS-Based Shared Storage

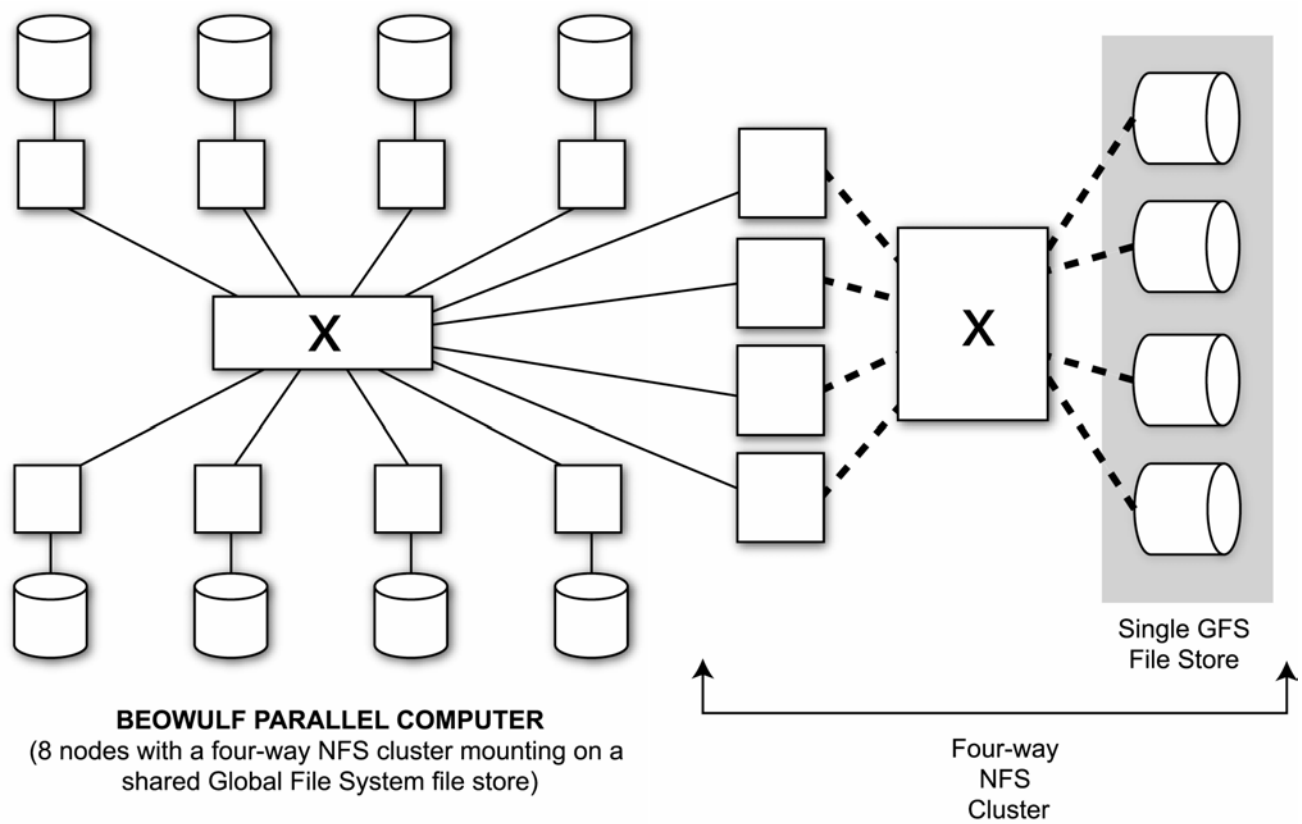
- Shared storage is great, but FC-based SANs are still expensive relative to compute node cost
- Instead of a single large NFS server, system architects can use GFS to enable an NFS cluster to provide scalable storage service for a large compute cluster

NFS Clustering for Shared Storage: 2-way



SISTINA
Infinitely Smarter Storage

NFS-Clustering for Shared Storage: 4-way



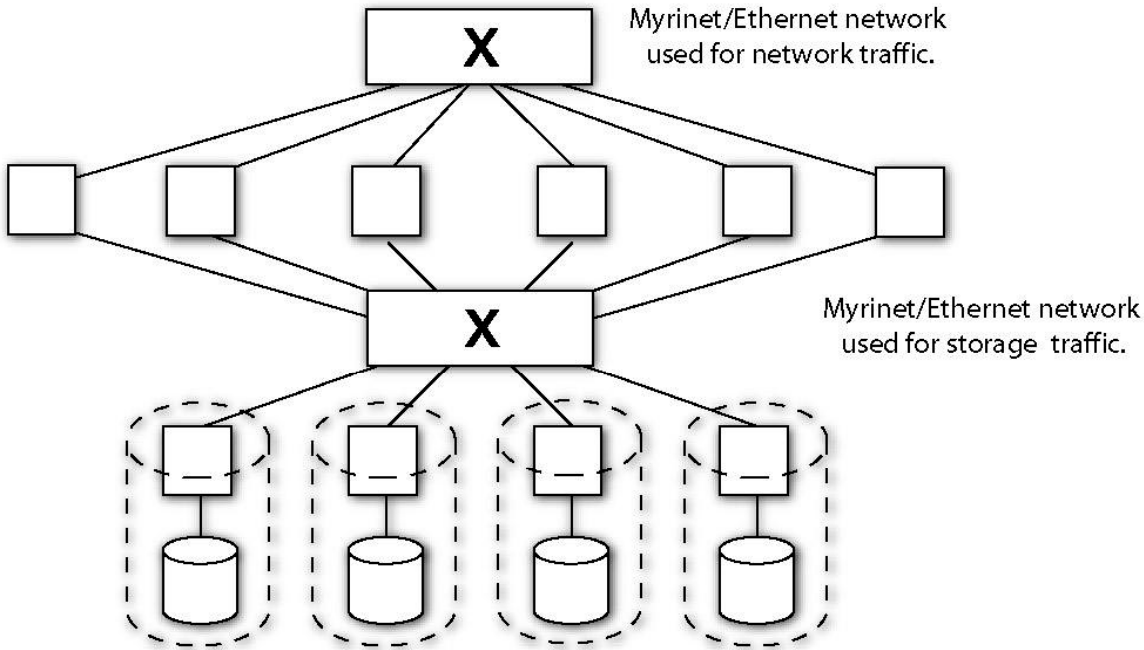
- Advantage: expensive SAN hardware not required for every node
- Disadvantage: not quite as scalable

SISTINA
Infinitely Smarter Storage

Table 1
Storage Clustering Alternatives Comparison Matrix

	File Protocol	Hardware Cost	Management Cost	Performance IO	Scalability (Bandwidth & Capacity)	Availability (Downtime & Failover)
Fully-integrated Storage Cluster	Local	Medium (IP) High (FC)	Low	High	High	High
N-way NFS Cluster	NFS	Medium	Medium	Medium - High	Medium	High
Single Node Shared NFS Server	NFS	Medium	Medium	Low	Low	Low
No Shared Storage	None	Low	Very High	Very Low	Very Low	Very Low

Storage Clustering without Fibre Channel



SHARED VIRTUAL DISKS USING GNBD
6 compute nodes connected to 4 virtual disks
over Ethernet or Myrinet (without Fibre Channel).

SISTINA
Infinitely Smarter Storage

Edge Serving — Definition

- At boundary between external Internet and internal data center operation, organizations typically deploy servers to implement a variety of client-server protocols
- These servers provide external client access to internal server data via web, email, file, or database protocols

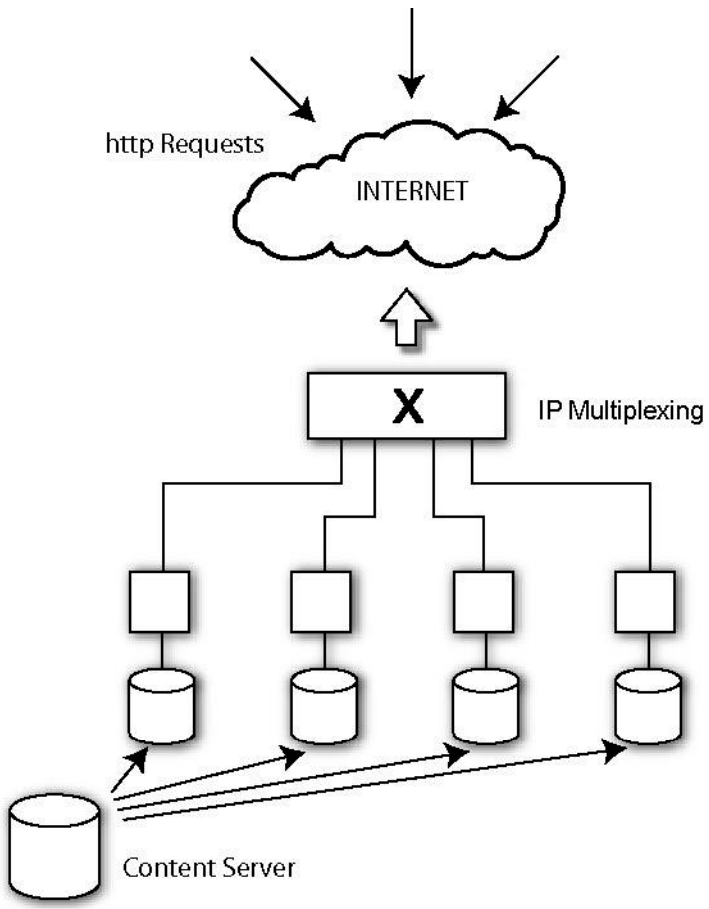
Edge Server Clustering Protocols

- Web Services
  - http and http proxy (e.g., squid cache)
- Mail Services
  - pop, imap, smtp, mta
- File Services
  - ftp, NFS, CIFS, Appletalk, Oracle IFS
- Client Login Services
  - ssh, telnet, etc.

Edge Server Farms

- A collection of edge servers will be referred to as an *edge server farm*
  - *multiple servers required to meet load requirements and to increase availability*
- Edge server farms are *homogeneous* if all servers in the farm are providing the same service or set of services

Edge Server Farms



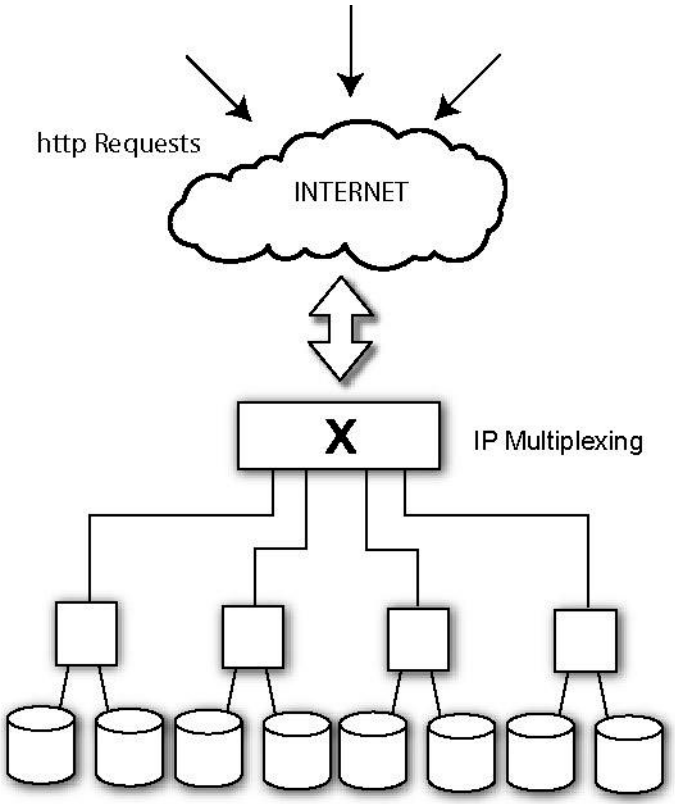
- Multiple servers provide content through IP load balancing switch to clients over Internet

- Content must be duplicated, no data sharing between servers, complex to manage and scale

- Scalable from hardware cost perspective, but complexity increases with each server added

SISTINA
Infinitely Smarter Storage

Edge Server Farms



- If content and storage requirements double, storage on each node must be doubled
- Replicated content complex to manage and update
- Replicated content works only for simple, read-only protocols (http) not file serving (NFS,CIFS)

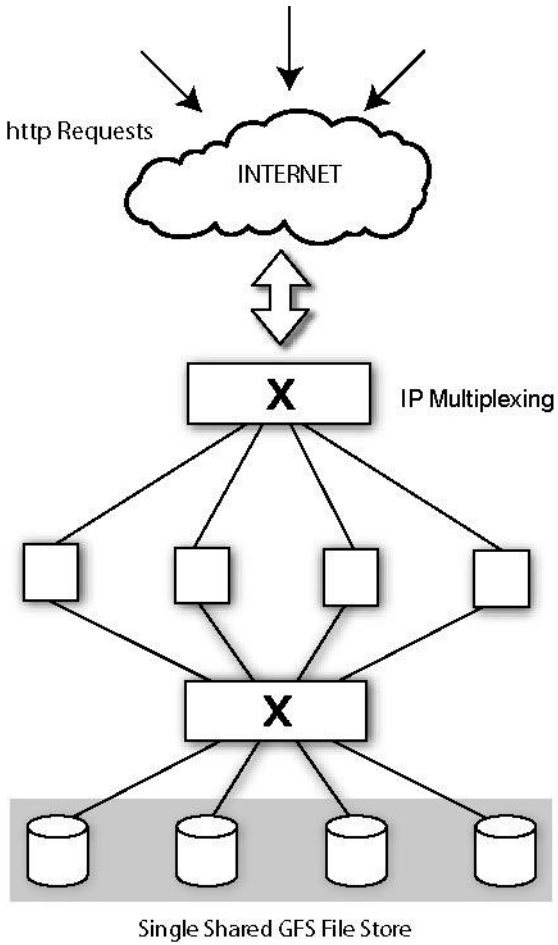
SISTINA
Infinitely Smarter Storage

- Farms don’t scale well from a management or efficiency standpoint (see Sun’s latest ads about consolidation)
- Clustered, shared storage solves the problems of edge server farm management, scaling, performance, and availability
- GFS and CLVM provide clustered storage for Linux

Storage Clustering

- Storage cluster: servers sharing storage with a cluster file system and volume manager over a storage area network
- Integrating an edge server farm with a storage cluster creates an *edge server cluster*

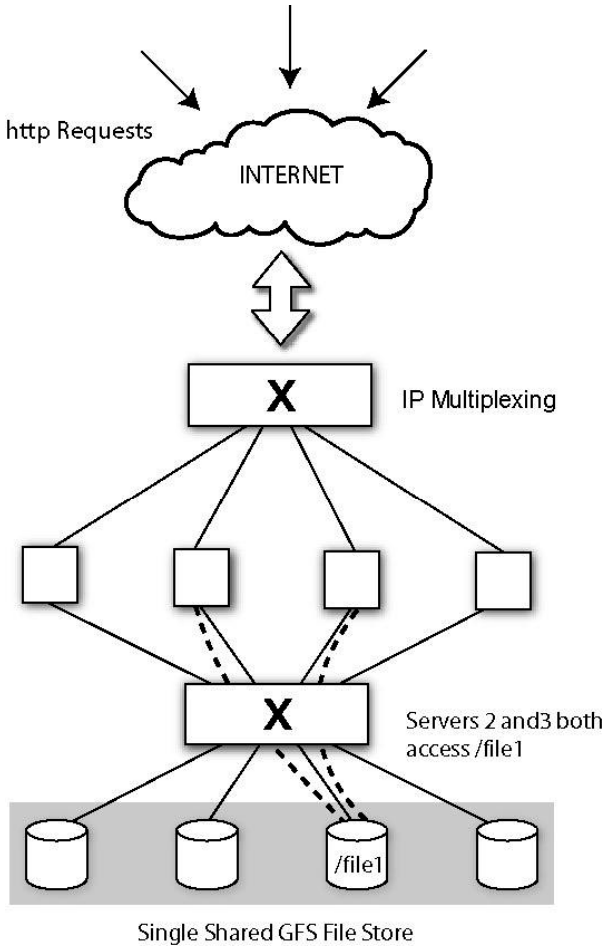
Edge Server Cluster



- Servers share storage: either storage or servers can be scaled independently, integrates fail-over and load balancing
- Edge server clusters exploit commodity technologies to achieve enterprise-class scalability, performance, and availability
- Approach works well for most client-server protocols

SISTINA
Infinitely Smarter Storage

Shared File Access in an Edge Server Cluster



- Storage area networks are not good enough: cluster file system provides shared read/write access to applications across the edge server cluster
- IP Load Balancer is an important part of the architecture: helps virtualize the edge servers from the standpoint of outside clients

SISTINA
Infinitely Smarter Storage

Edge Serving

- About 40% of our users are building edge servers
  - applications include web serving, ftp, and email
  - availability is a big factor, especially for email
  - performance is also a factor
  - users are learning the storage management advantages of clustering as they deploy Sistina's storage cluster technology

Edge Serving Applications

- We are collaborating with the Samba team to insure CIFS clustering support
- We are developing an NFS cluster server stack to integrate with GFS and CLVM
  - current Linux NFS stack works fine, but additional performance and capabilities can be provided
- We are beginning to work with load balancing switch vendors

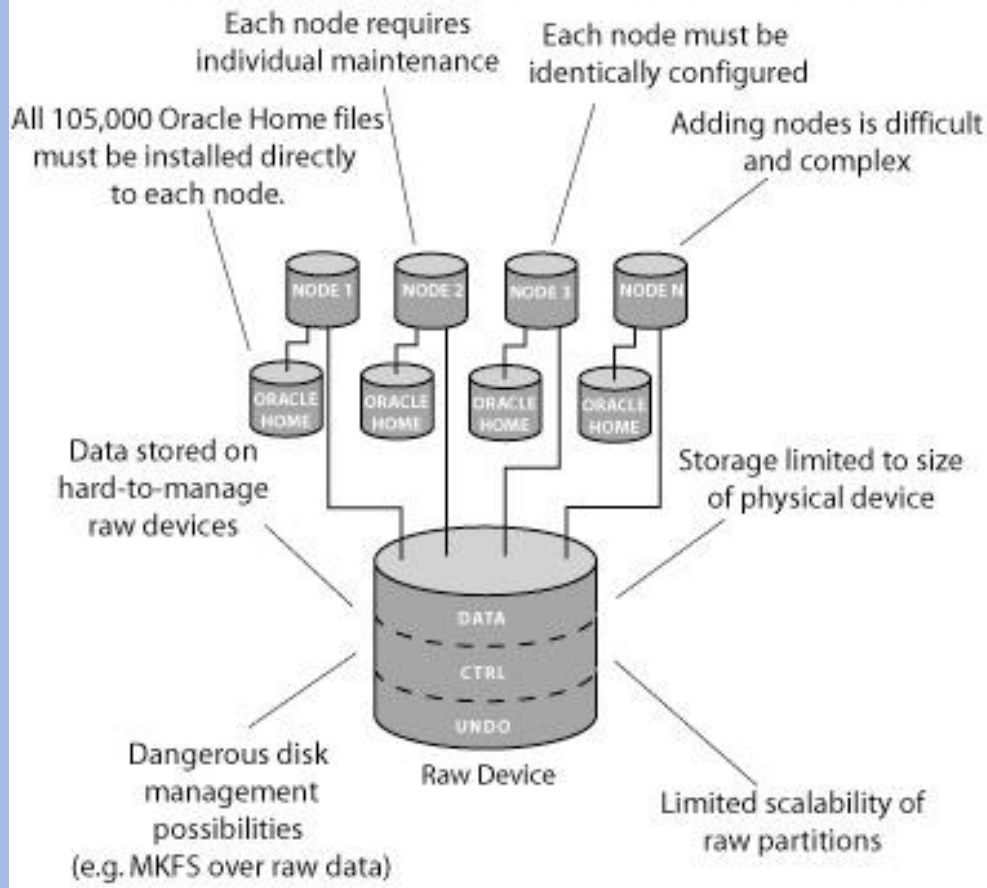
Summary on Edge Serving

- Edge server clustering can greatly simplify service and storage management complexity
- It can also increase ease of management, throughput, and availability
- GFS cluster file system and CLVM logical volume management technologies enable edge server clusters in Linux

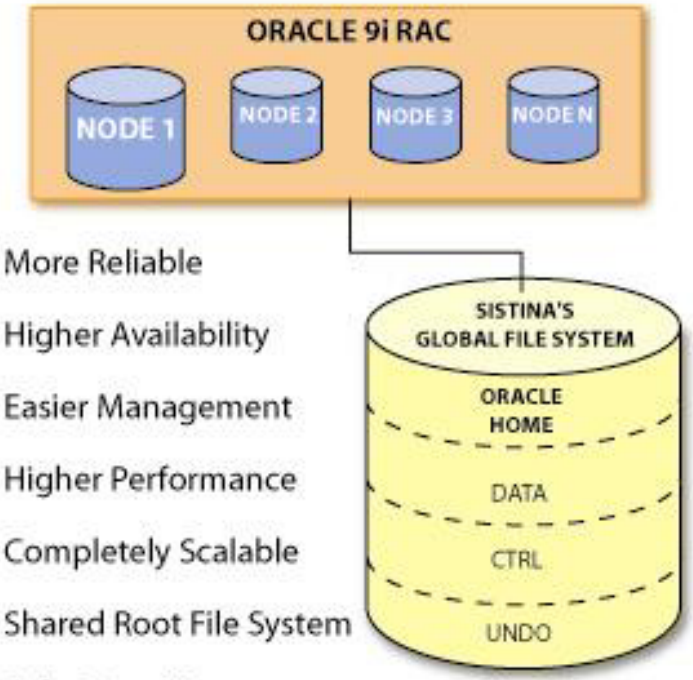
Oracle Night and Day

A NEW DIRECTION IN DATA STORAGE & MANAGEMENT

TRADITIONAL ORACLE OPS ARCHITECTURE



ORACLE 9i RAC WITH SISTINA'S GFS



More Reliable

Higher Availability

Easier Management

Higher Performance

Completely Scalable

Shared Root File System

Safer Data Management

Data Files in File Systems

Lowest TCO via Commodity Hardware

SISTINA

Infinitely Smarter Storage

Cost Savings Example (Legacy)

Qty	Vendor	Model	Description
2	Sun	Enterprise 10000	64 466MHz Sparc II processors
2	Sun	Solaris 8	Included w/Sun System cost
40	Sun	Sbus HBA	Dual-Loop FC-AL Sbus HBA
6	Brocade	Silkworm 2040	8 port Fibre Switch
4	EMC	Clariion FC4500	Rack Mountable Raid 0+1 Disk Array
64	EMC	Clariion 18GB Drives	18GB drive for Clariion
2	Cisco	Catalyst 2924	24-port 10/100 Switch (Enterprise Edition)
2	Sun	Sun Clusters	HA Agent for Failover
2	Sun	SC Agent for Oracle	HA Agent for Oracle Database
Total			**$3,414,194.00**

SISTINA
Infinitely Smarter Storage

Cost Savings Example (w/GFS)

Qty	Vendor	Model	Description
64	Dell	1U Rack Server	Dual Pentium III @ 933MHz
64	SuSE	SuSE 7.3	Linux Operating System
64	Qlogic	QLA2200f	64bit, 64MHz, Copper, HBA
5	Brocade	Silkworm 2040	8 port fibre switch
4	EMC	Clariion FC4500	Rack Mountable Raid 0+1 Disk Array
64	EMC	Clariion 18GB Drives	18GB drive for Clariion
4	Cisco	Catalyst 2924	24-port 10/100 Switch (Enterprise Edition)
64	Sistina	GFS 5.0	Cluster File System for Linux
Total			**$225,654.00**

Summary

- Enterprise applications require ease of management, availability, throughput
- Storage clustering technology can provide all three — SAN and NAS alone do not
- Faster, cheaper SANs and virtualization, coupled with cluster file systems provide storage clustering foundations
- Thanks for your attention
- Questions?