High Throughput Computing
Linux Clusters, Storage, Grids

Jamshed H. Mirza
Server Group
mirza@us.ibm.com

Notice

All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only. Contact your IBM local Branch Office or IBM Authorized Reseller for the full text of a specific Statement of General Direction.

IBM may have patents or pending patent applications covering subject matter in this presentation. The furnishing of this presentation does not give you any license to these patents.

The information contained in this presentation has not been submitted to any formal IBM review and is distributed AS IS.

The following terms are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both: IBM, IBM logo, AIX, AIX/L, PowerPC, RS/6000, SP, Netfinity, pSeries, xSeries, Chipkill, ServeRAID.

Intel and Pentium are trademarks or registered trademarks of Intel Corporation in the United States, other countries, or both.

Myrinet is a trade name of Myricom, Inc.

LINUX is a registered trademark of Linus Torvalds

Other company, product, and service names may be trademarks or service marks of others.

Linux is important to IBM

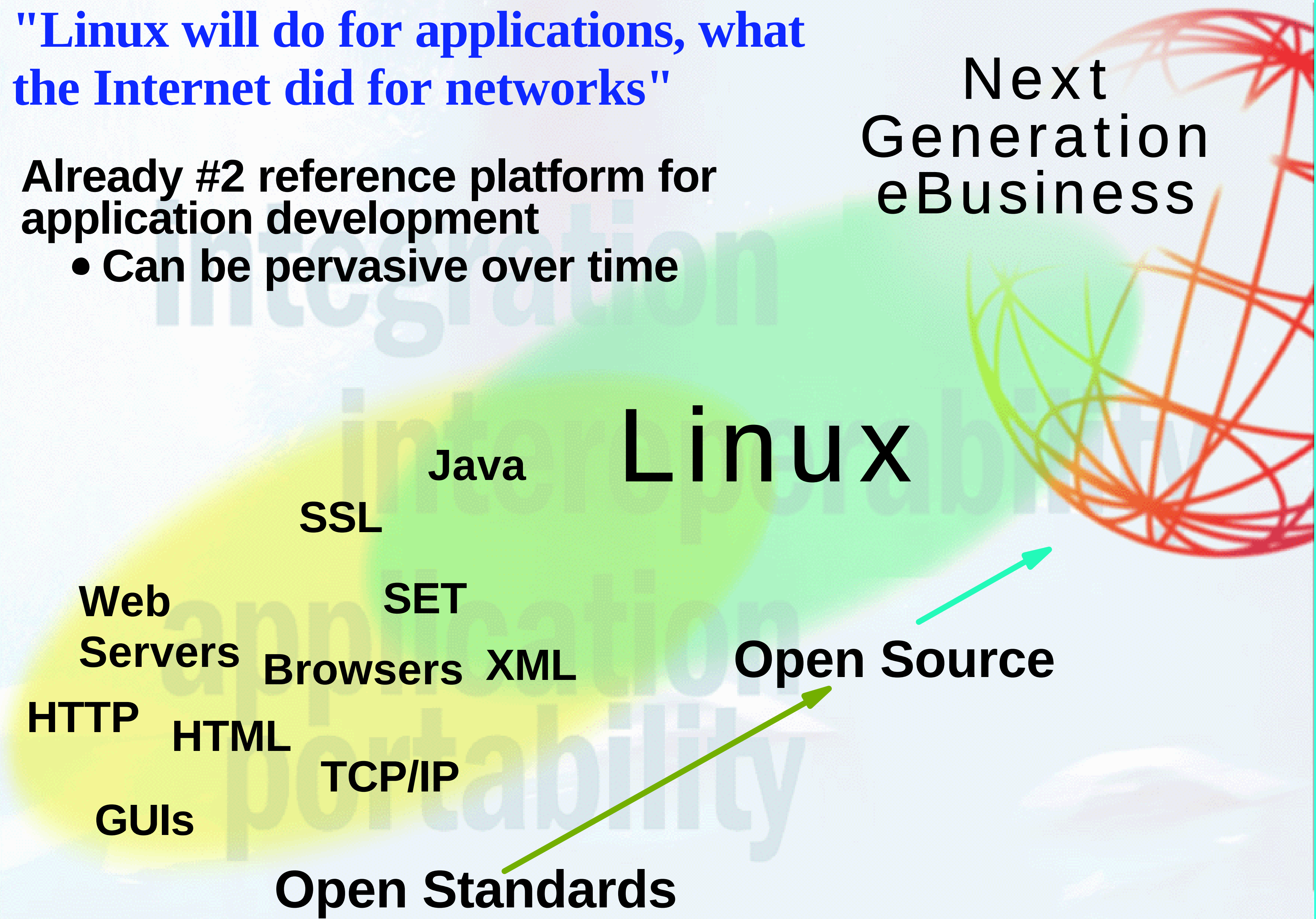
- Entrenched Internet technology
- Increasingly used in the HPC market
- Can become the volume Application Development and Deployment environment
- Potential to be a key technology for the next generation eBusiness

The Potential of Linux

"Linux will do for applications, what the Internet did for networks"

Already #2 reference platform for application development

- **Can be pervasive over time**



Linux

Linux has real and perceived limitations today for pervasive, enterprise-wide use

IBM sees Linux as a strategic technology

- We are investing considerable resources and $$, and contributing key IBM technology to making it enterprise-ready
- We will work with the community to do that

Unix and other proprietary operating systems will continue to exist for foreseeable future

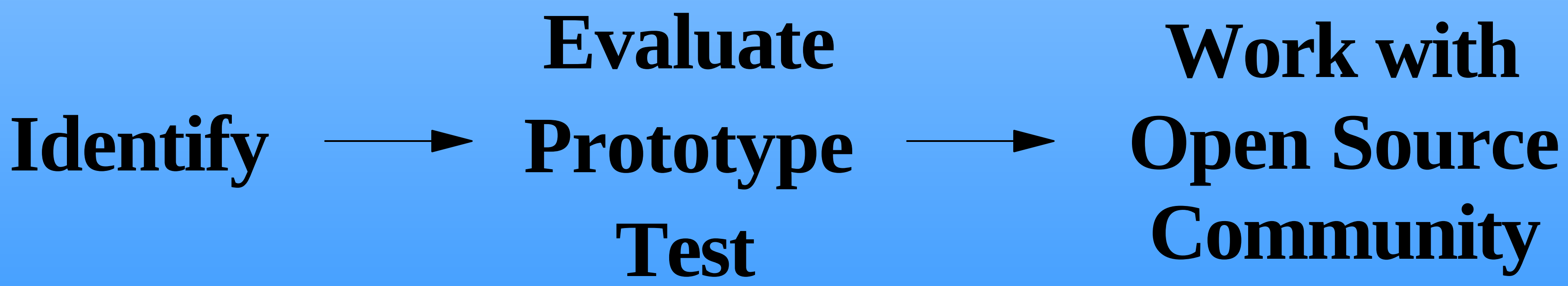
- Large investment in software, applications, data
- IBM continues to invest in AIX and pSeries systems

Linux Technical Strategy

- Support Linux on all IBM platforms
- Strong affinity between IBM operating Systems and Linux
  - Example: AIX/L
- Work with the Linux community to infuse technology into the Linux kernel
- Deliver robust Linux Cluster solutions based on Open Source and IBM technologies
- Encourage adoption of Linux

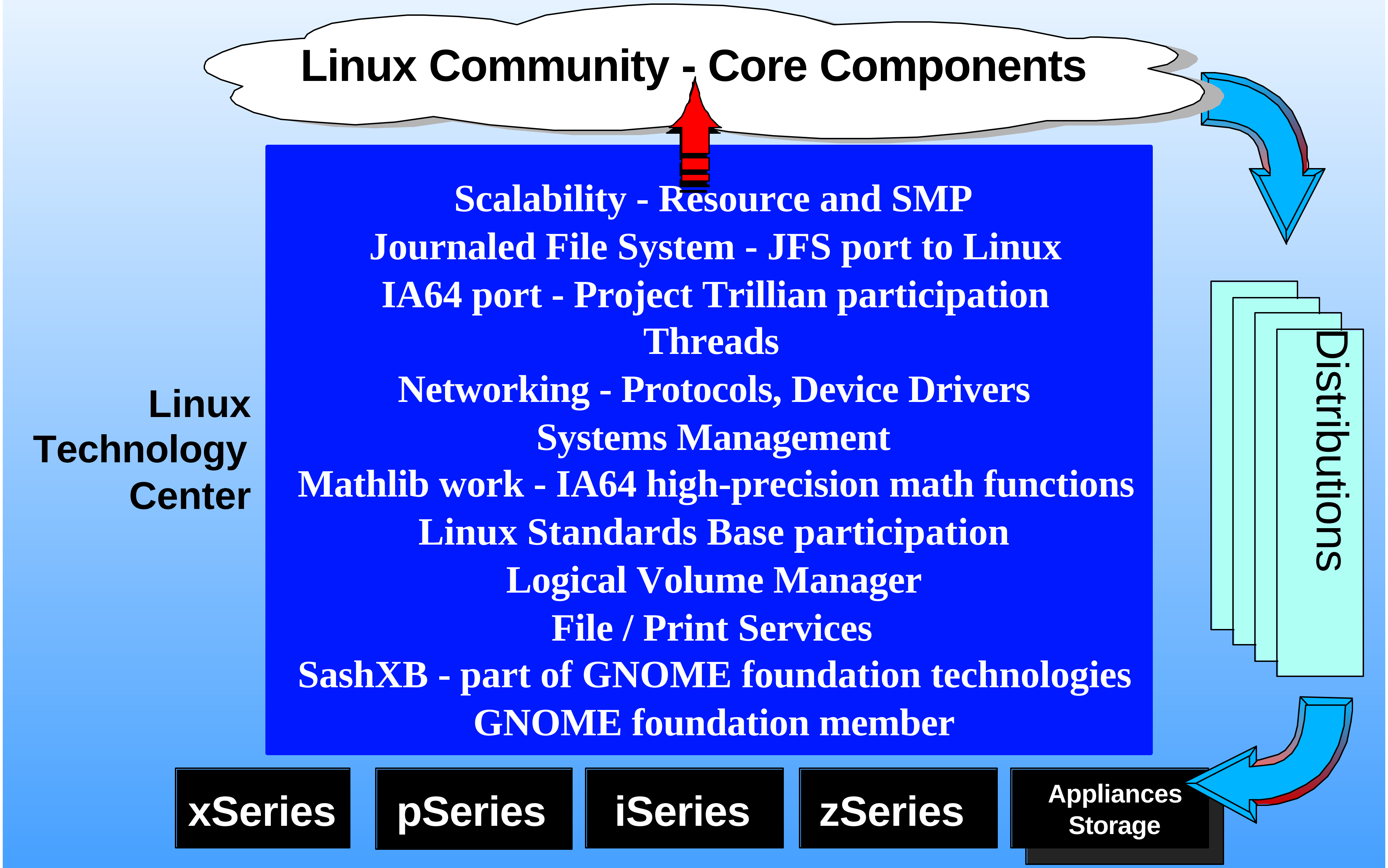
Goal: Accelerate maturation of Linux into Enterprise

- Distributed - worldwide organization of ~200 developers
- IBM's primary interface to open source Linux community
- Identify and work on enhancements for enterprise-class capability



Identify → Evaluate Prototype Test → Work with Open Source Community

Evolving Linux



Linux Community - Core Components

Linux Technology Center

Scalability - Resource and SMP
Journaled File System - JFS port to Linux
IA64 port - Project Trillian participation
Threads
Networking - Protocols, Device Drivers
Systems Management
Mathlib work - IA64 high-precision math functions
Linux Standards Base participation
Logical Volume Manager
File / Print Services
SashXB - part of GNOME foundation technologies
GNOME foundation member

Distributions

xSeries

pSeries

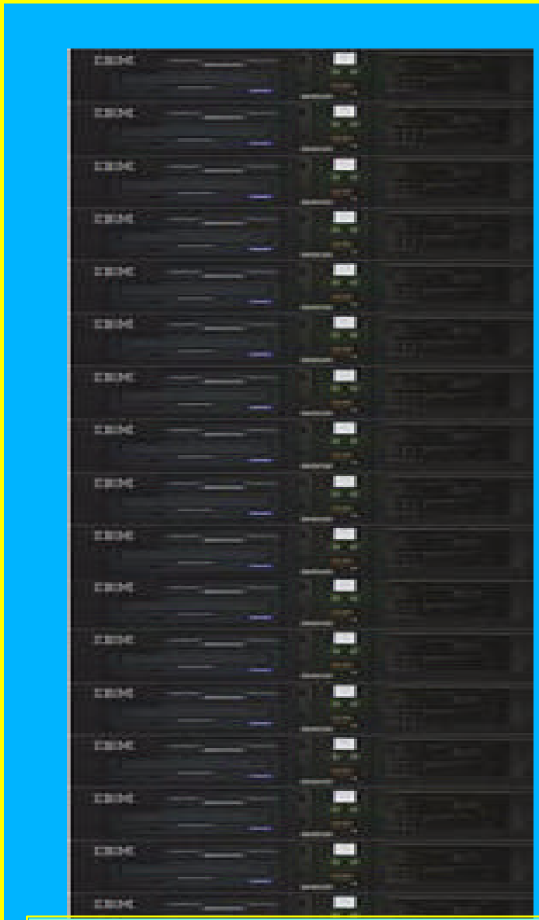
iSeries

zSeries

Appliances
Storage

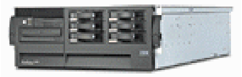
IBM Linux Cluster solution for S&TC

- Prepackaged, prevalidated Linux Cluster
- 1U and 3U 2-way IA-32 servers
  - PowerPC and IA-64 in '02
- Cluster and management networks, remote control
- Fully integrated, high availability storage solution
- Comprehensive Systems Management (CSM)
- Cluster File System (GPFS)
- S&TC optimized solution includes
  - High-performance Myrinet 2000 network
  - High-performance compilers (Fortran, C, C++, OpenMP)
  - Parallel Debugger (TotalView)
  - Job Management software (PBS)
- Optional enterprise service/support for IGS



2-way 1.26 GHz
4 GB RAM
146 GB disk

Rack-optimized IA-32 Systems

	xSeries 330 	xSeries 340 	xSeries 350 	xSeries 370
Processor	**2-way SMP**	**2-way SMP**	**4-way SMP**	**8-way SMP**
Package	**1U**	**3U**	**4U**	**8U**
Max Memory	**4GB**	**4GB**	**16GB**	**32GB**
Internal HDD	**2**	**3**	**6**	**2**
PCI slots	**2**	**5**	**6**	**12**

Integrated Service Processor

Netfinity Director

Software Rejuvenation

Processor Deallocation

Cable Chaining Technology

Predictive Failure Analysis

Hot Plug disk, adapters, fans, power

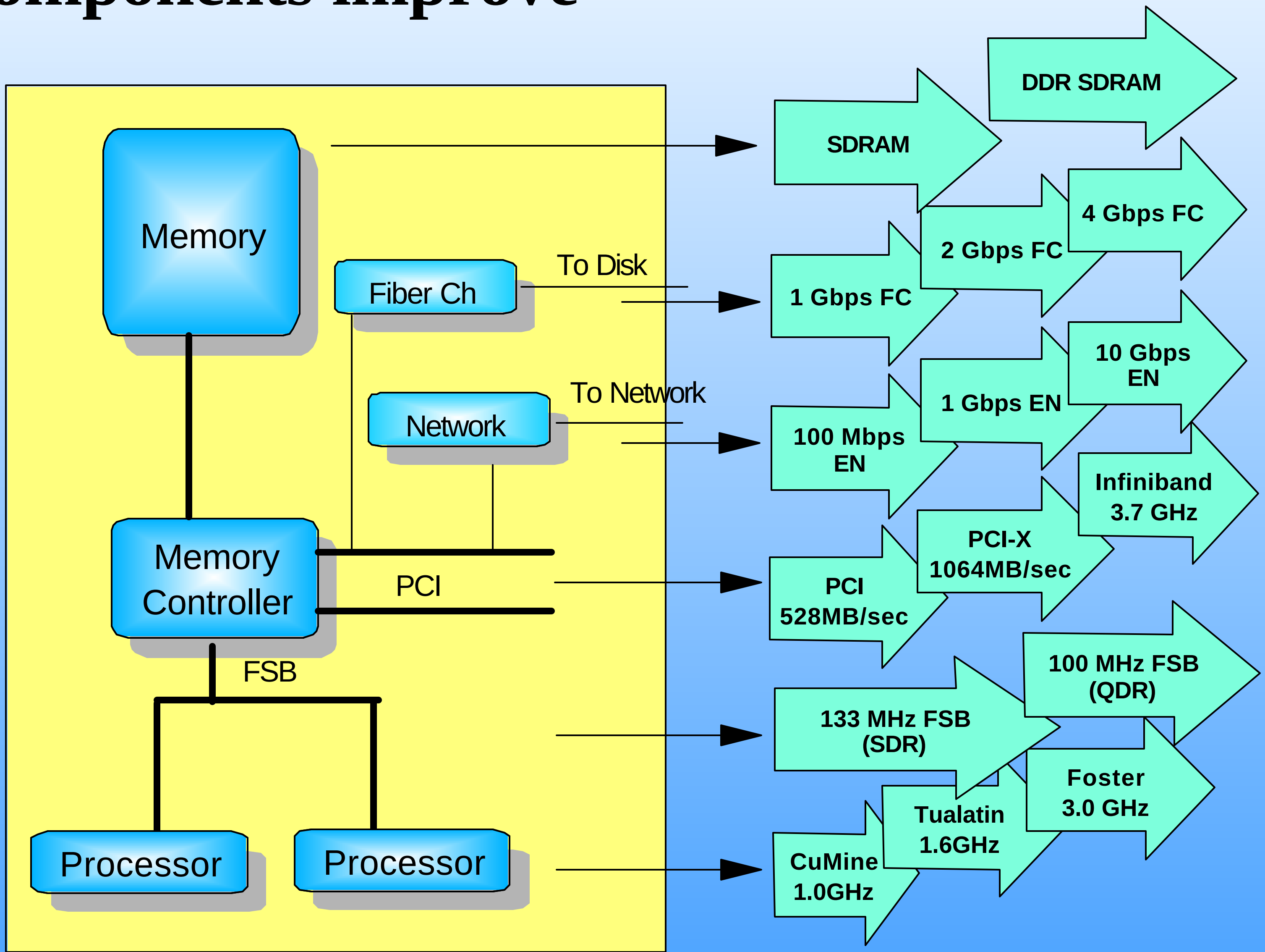
LightPath Diagnostics

ChipKill Memory

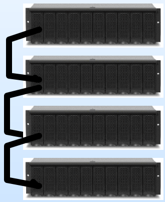
Varying capability in various models

Intel-Based Server Roadmap

All components improve



Rack-optimized Systems



IA-64 4-w Itanium 4-w McKinley 4-16w McKinley

IA-32

Follow Intel Curve

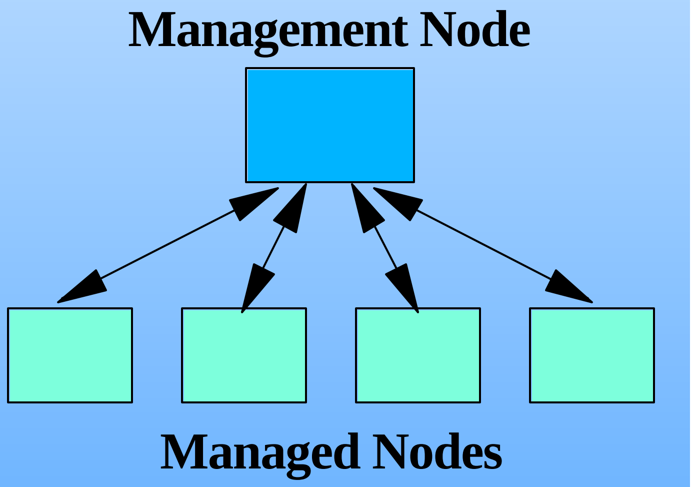
Ultra320 SCSI

Blade-based Dense Servers

Cluster Systems Management for Linux

CSM allows a cluster to be managed as a single entity from a single point of control

- **Remote hardware control and monitoring**
  - Power on/off/reset
  - Monitor environmental conditions
- **Remote console function**
  - Access to cluster servers prior to OS installation or when network access is unavailable
- **Software installation**
  - Cluster-wide parallel install
- **Distributed Shell, Node Groups**
  - Execution of arbitrary commands or scripts on all or some of the servers in the cluster



Cluster Systems Management for Linux

Allows a cluster to be managed as a single entity from a single point of control

- Configuration File Manager
  - Enables administrator to set up configuration files in a central place
  - An agent that pulls any changes down to each server in the cluster
- Distributed Management Server
  - Coordination for various management functions
  - Persistent repository of cluster configuration
  - Heartbeat function
  - Liveness state that can be assessed by other applications

Cluster Systems Management for Linux

Allows a cluster to be managed as a single entity from a single point of control

- Event Response Resource Manager
  - Mechanism for automatic response to specific events
  - Set of predefined events and actions that are commonly used in managing a cluster will be provided

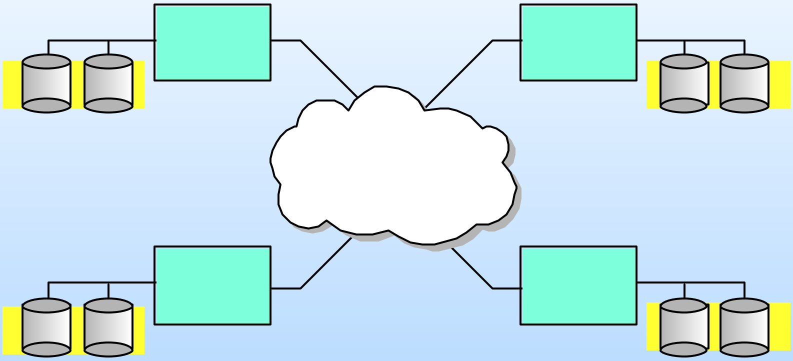
- Probe manager
  - Set of probes to check consistency of cluster configuration information and diagnose configuration errors

Much of CSM is based on mature SP technology

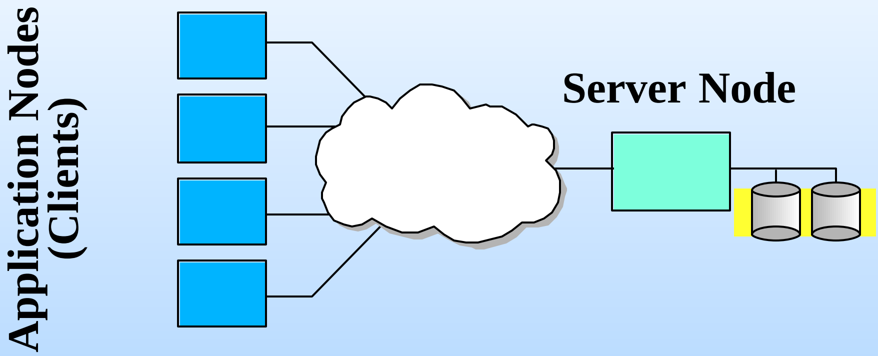
Used in over 10,000 SP systems today

Gone through multiple releases over past 9 years

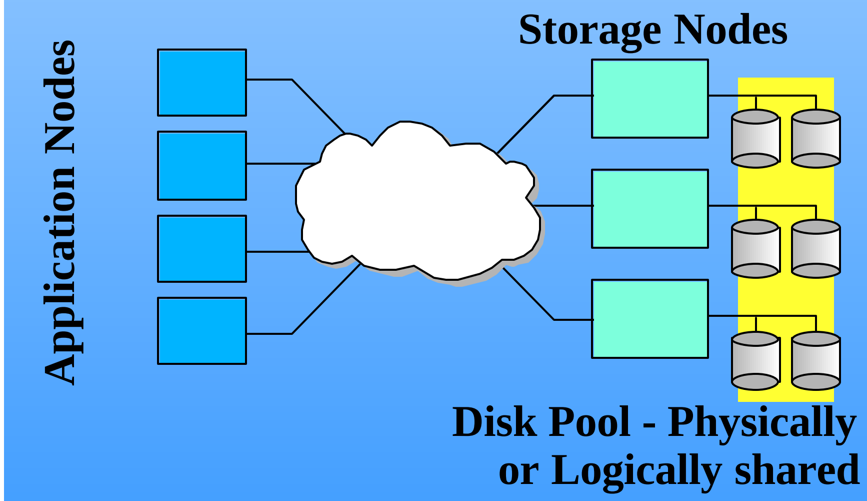
Scalable Cluster File System



- **Native File System**
  - No file sharing - application can only access files on its own node
  - Applications must do their own data partitioning or replication



- **DCE Distributed File System**
  - Application nodes share files on server node
  - Coarse-grained (file or segment level) parallelism
  - Server node is performance and capacity bottleneck



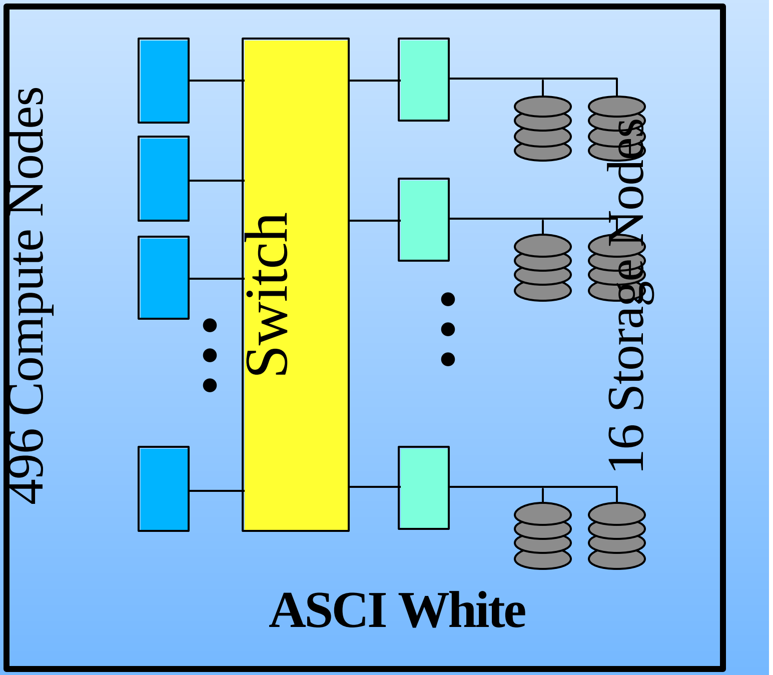
- **GPFS Parallel File System**
  - Striped across multiple disks on multiple storage nodes
  - Independent GPFS instances run on each application node
  - Storage nodes used as "block servers"
    - all instances can access all disks

GPFS

- Posix standards-compliant
- Uniform access via (logically or physically) shared disks
- High capacity - tens of TB per file system, ~TB per file
- High throughput
  - Wide striping and large data blocks
  - Client caching via distributed locking
  - Parallel access via fine-grained (byte range) locking
  - High sequential throughput via aggressive prefetch
- Reliability and fault-tolerance - node and disk failures
  - Journaling, data replication, RAID support
  - High-availability infrastructure
- Export via NFS and DFS
- MPI-IO optimizations (Not fully exploited by MPICH for Linux)
- Data Management API for Hierarchical Storage Management (Not supported in initial Linux release)

GPFS

- GPFS is in it's fifth release and is used in widely used by our RS/6000 SP customers (AIX)
- ASCI White
  - 512 nodes, 8K CPUs
  - 150+ TB
  - 12 GB/s to/from single file (or multiple files)
- Can be used as scalable NFS or DFS
- Now available on IA32/Linux
- Used in future NAS solutions from IBM



IBM Global Services

Education & Training

- Classroom or via web
- Available in 20 countries, multiple languages
- How-to (Redbooks) for Linux

Service & Support

- 24 X 7 enterprise level support
- All major distributions

Professional Services

- Comprehensive enterprise services for Linux
- Infrastructure consulting and planning
- Installation
- Configuration
- Application enablement

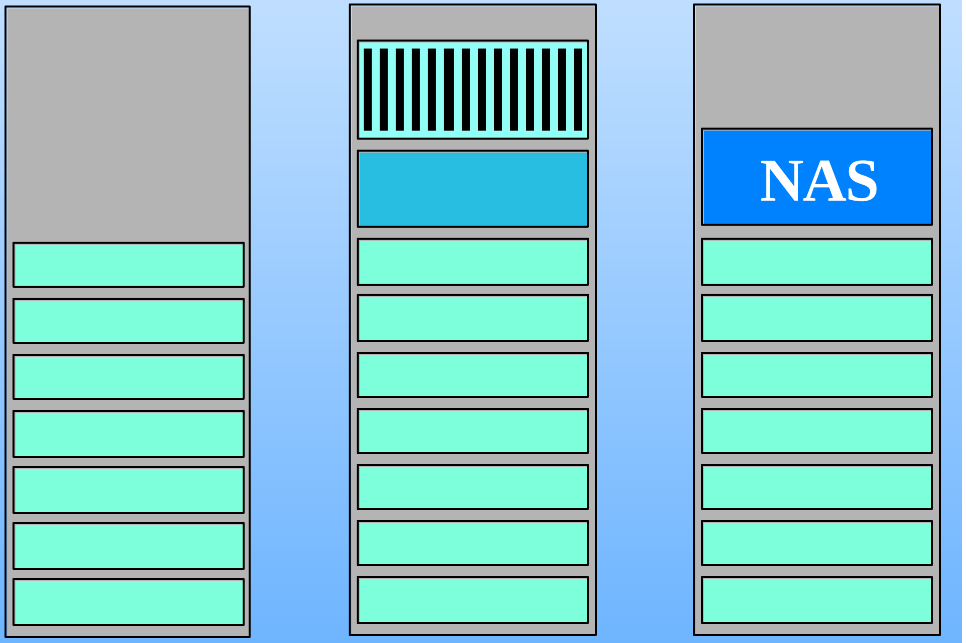
Linux Clusters - Future

IA-64 and PowerPC Linux

Scalability optimizations

Functional Enhancements

- Install
- Automated operations
- Security
- Usability
- ...



Support other nodes

Integration of NAS

Integrated AIX & Linux clusters

Linux Cluster Successes

U of New Mexico

- 256 x330s
- 80th on the 12/00 Top500 Supercomputers list

Maui High Performance Computing Center

- 288 x330s
- One of the larger SP sites

NCSA

- Support next generation Grid
- xSeries servers: 512 x330s and 100+ IA-64 nodes (1 TF each)
- IBM SW to support scaling, management and application in a tera-scale Linux cluster environment

Royal Dutch Shell

- Tera-scale seismic processing
- 1024 x330s (1+Tflop)
- IBM Global Services to design, build, and implement

MDS Proteomics

- Two 100-node x330 clusters
- 80th on the Top500 Supercomputers list

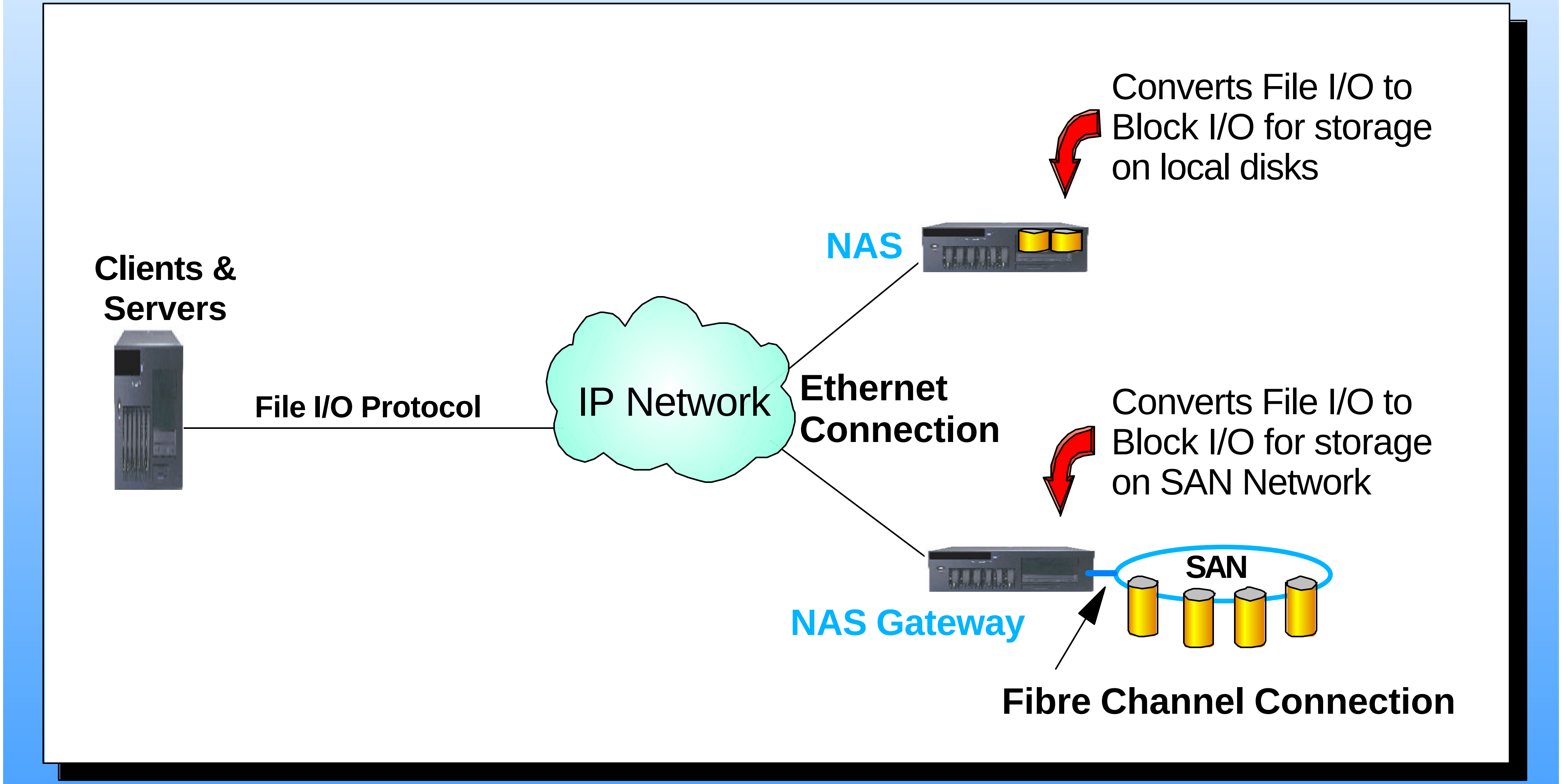
weather.com

- One of "top 25" web sites
- xSeries servers with Linux, Websphere Commerce Suite, IBM Global Services design approach
- Cost, availability, scalability requirements

Whole range of Storage Solutions

- **Direct and SAN-attached storage**
- **NAS**
- **iSCSI**

NAS and NAS Gateways

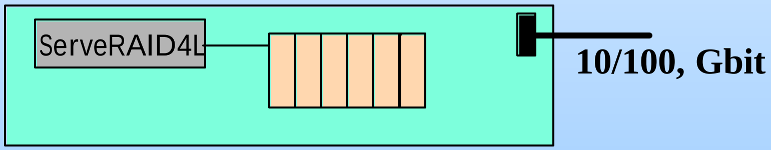


Converts File I/O to Block I/O for storage on local disks

Converts File I/O to Block I/O for storage on SAN Network

NAS 200 Low-End Models

Workgroup Model



Uni or dual IA32 processors server

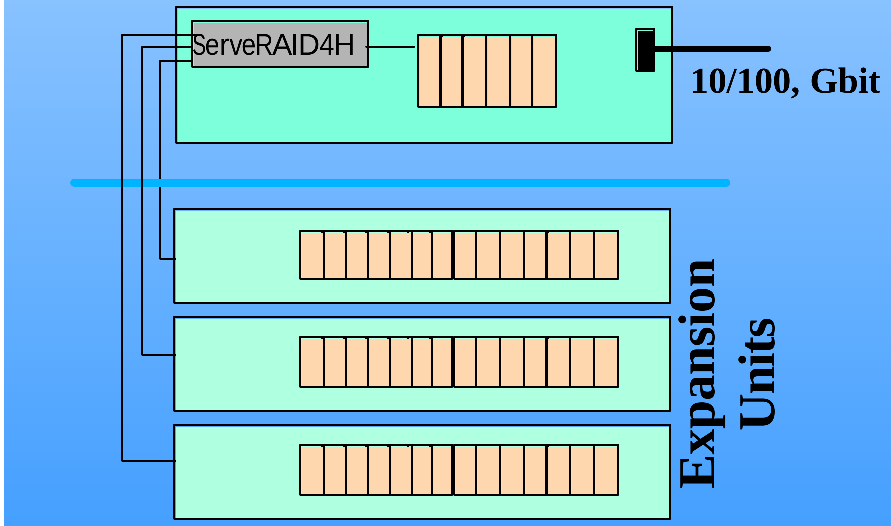
1-channel RAID controller

108GB to 216 GB

Support for multiple File System protocols

CIFS, NFS, HTTP, Novelle, ...

Departmental Model

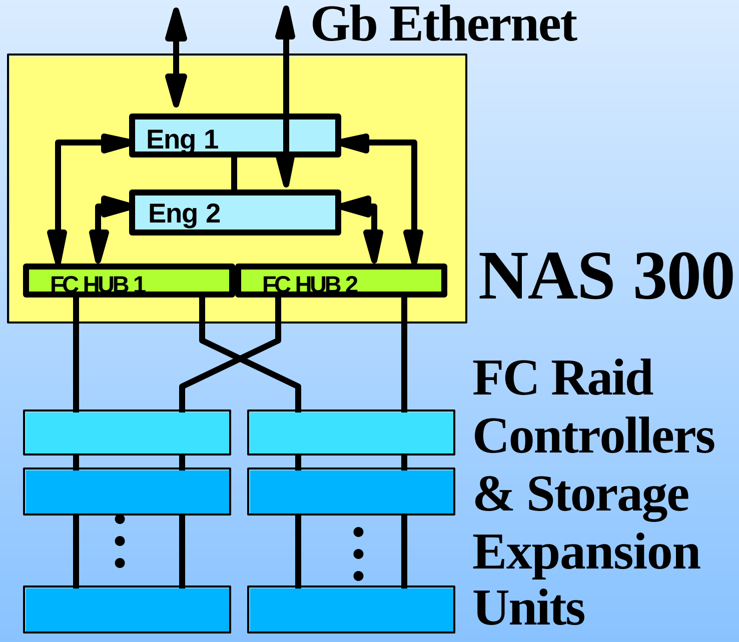


Dual IA32 processors server

4-channel RAID controller

216 GB to 1.7 TB

NAS 300 and 300G High-End Enterprise Models

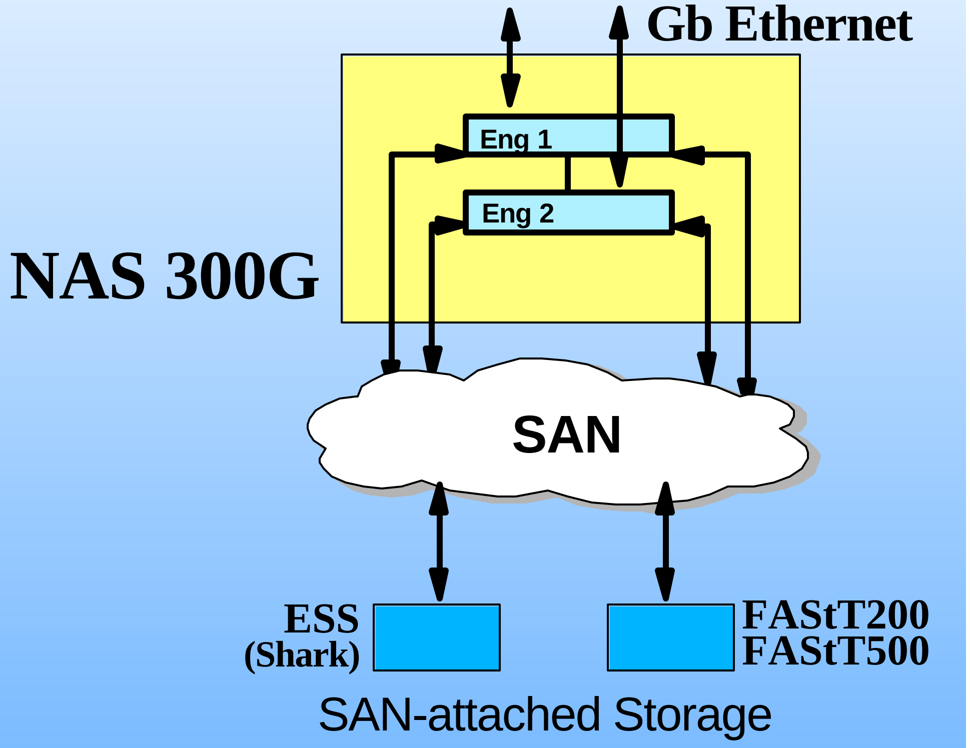


One or two dual-processor engines

Dual engines and redundant components for fault-tolerance

Cluster or Failover mode

360 GB to 3.4 TB



Allows IP Clients and Servers access to SAN storage without fiber channel connection

Dual-engine option for Fault-tolerance

Access to up to 11 TB storage

On The Roadmap

Scalable NAS

Multiple dual-engine NAS engines

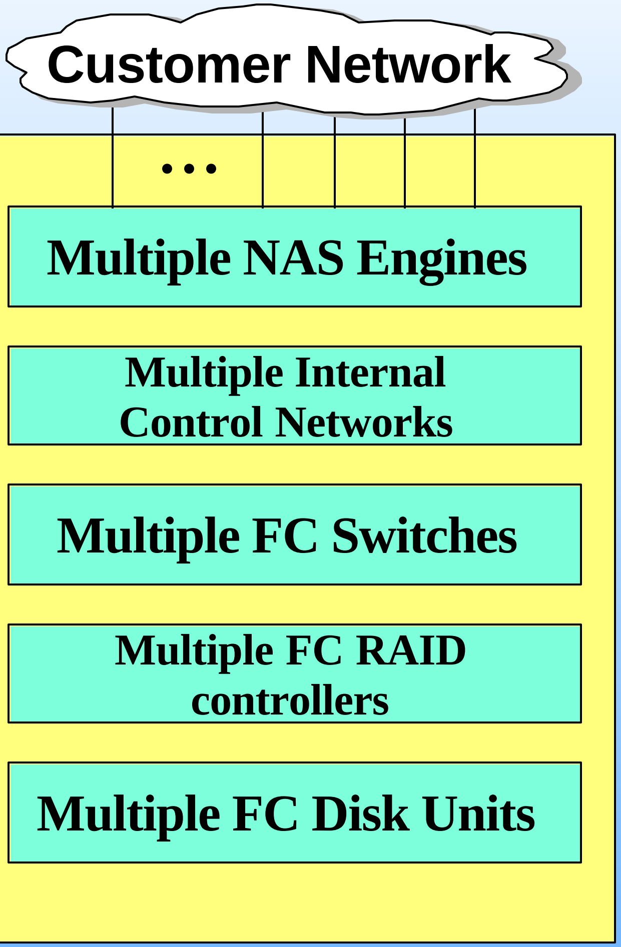
Redundant components

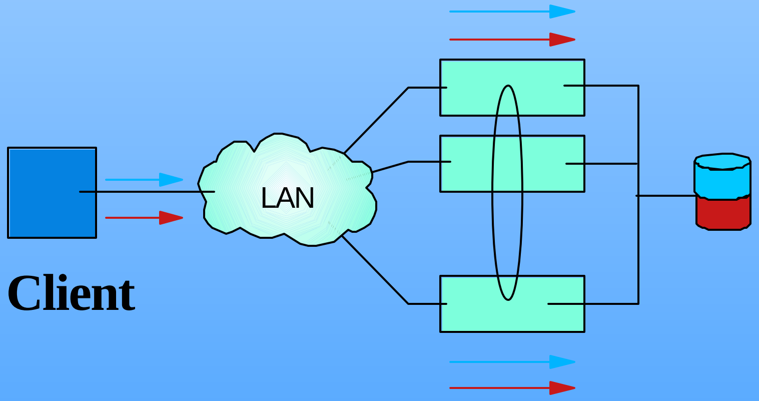
Cluster Shared File System

- Single scalable File System
- HA, load-balancing

Gateway model for existing SAN infrastructure

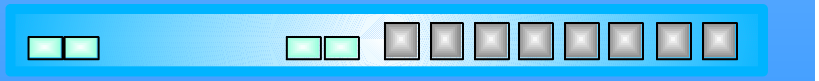
Scalable capacity and performance





Any node can read or write to any piece of data...
CONCURRENTLY

Low-cost IDE Disk based NAS



iSCSI

- SCSI storage protocol encapsulated in a TCP/IP message and transported over an IP network
- Enables IP Storage Area Network (IP SAN)

Why TCP/IP?

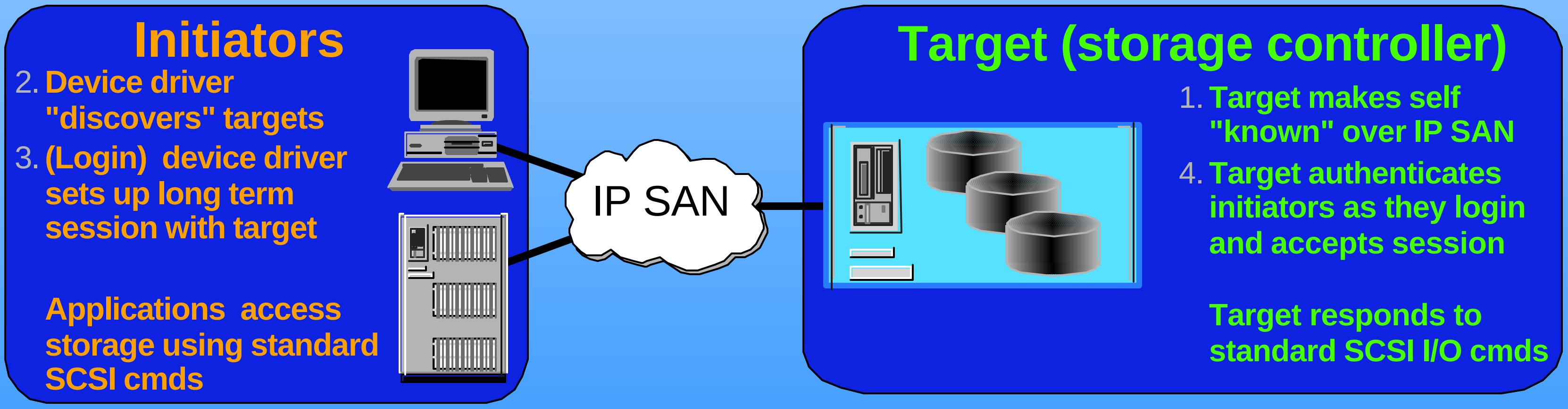
- Guaranteed in order delivery
- Ubiquitous (20+ yrs old)
- Supports long distances

Why SCSI?

- Standard storage protocol in use today

Why IP SAN?

- Pooling, Scaling, etc.
- Availability
- Interoperability
- Single skill base
- Economies of scale
- Advanced capabilities



Initiators

2. Device driver "discovers" targets
3. (Login) device driver sets up long term session with target

Applications access storage using standard SCSI cmds

Target (storage controller)

1. Target makes self "known" over IP SAN
4. Target authenticates initiators as they login and accepts session

Target responds to standard SCSI I/O cmds

IBM TotalStorage IP Storage 200i

- 108 GB to 1.7 TB
- RAID for performance and availability
- Remote/centralized storage management through Web-based GUI
- Addition of storage and administration while online
- Pre-installed software
- Linux, Windows NT, Windows 2000 clients

Model 100

- Entry model
- 800 MHz PIII
- One-channel RAID controller
- 10/100, Gbit or Gbit Fiber
- Hot swap drives, fans, power
- 108 to 216 GB capacity

Model 200

- Two 800 MHz PIIIs
- 4-channel RAID controller
- Up to 6 internal 36 GB disk
- 10/100, Gbit, or Gbit Fiber
- Hot swap disk, fans, power
- Up to 3 external enclosures (3U each)
  - Each up to 14 36 GB hot-swap disks
- 108 GB to 1.74 TB capacity

On The Roadmap ...

- Hardware offload
- Low-cost IDE option

IBM and Grids

Grid intersect several key IBM strategies and initiatives

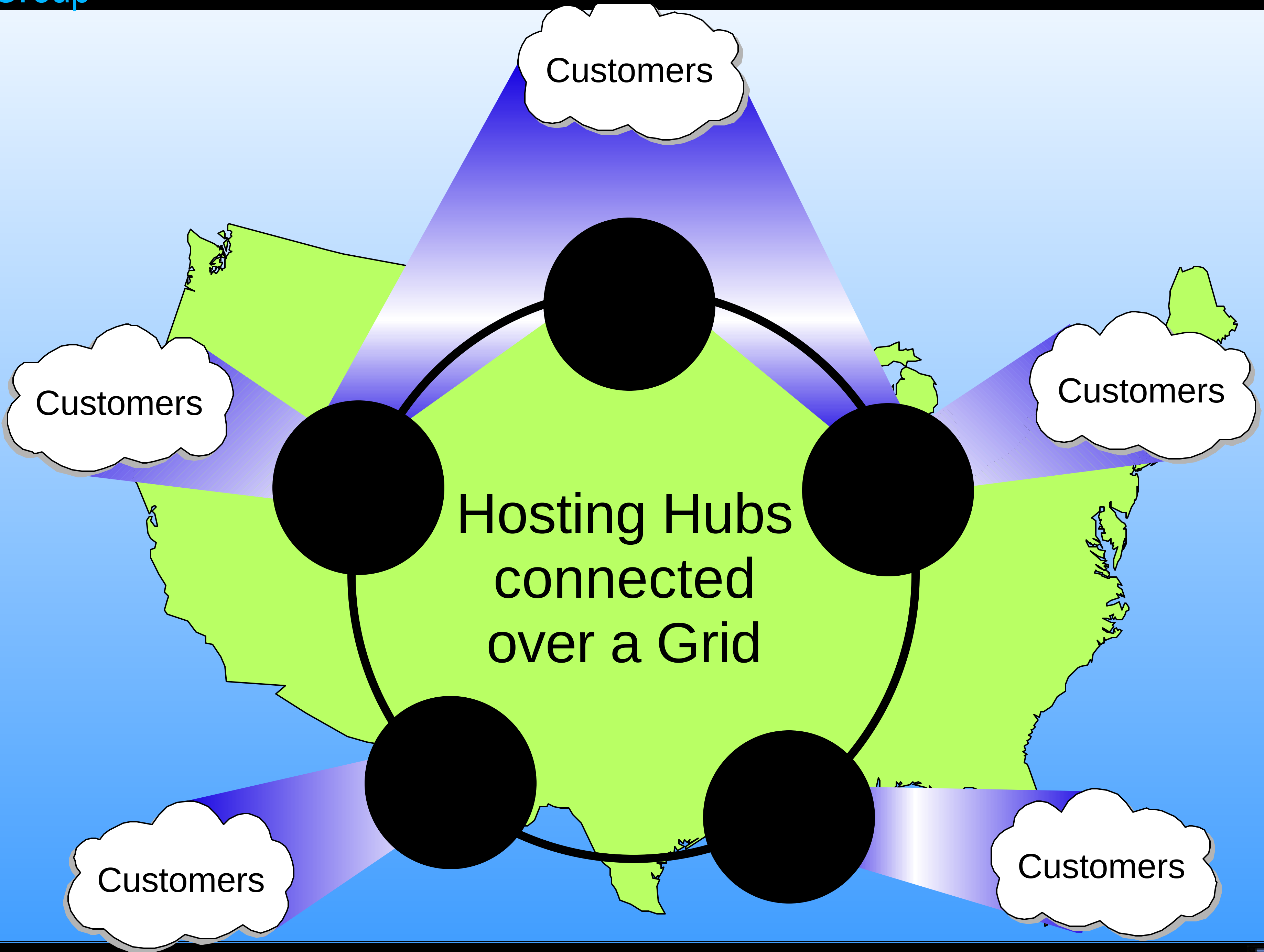
- Deep Computing
- eLiza
- eSourcing
- eServices

IBM's eSourcing Initiative

- Delivery of standardized processes, applications, and infrastructure over the network as a service on a pay-as-you-go basis
  - Business Functions: CRM, eCommerce, Supply Chain, ...
  - IT Functions: Security, Web Hosting, Storage Services, Systems Management, ...
- $4B to add 50 hosting centers worldwide to serve as eSourcing hubs
- First step in the Utility model

Connect hubs into a Grid over time

eSourcing


Customers
Customers
Customers
Customers
Customers
Hosting Hubs
connected
over a Grid

IBM and Grids

- We believe Grids will emerge much as eSourcing and eServices will
- We recognize Grid Computing as a key strategic area
- Irving Wladawsky-Berger anointed to lead the Grid Computing Initiative
  - Forming a cross-unit design council
  - Align with eSourcing strategy
  - Engage with Grid development community
  - Encourage joint University research in appropriate areas

IBM Grid Initiative

"Although other companies have expressed interest [in Grids], Foster and Hey said IBM has shown the most significant support so far. 'IBM is distinguished by farsightedness and enthusiasm,' Hey said. 'This stuff, to be significant in the long term, has to move into the commercial space, and IBM has stepped up,' Foster said."

New York Journal News, Aug 2.

Game Plan

Work with the community

IBM technology where relevant

Grid-enable IBM products

Promote use in a wider segment

Very similar to how we approached the Linux Initiative

Assets or Building Blocks

- Systems
  - Servers, Storage, Linux, Clusters
- Number of technologies available or under development which can be applied to or extended for grid computing
  - Data access, data management
  - Resource management/Workload Management
  - Security
  - Resource publication and discovery
  - Performance monitoring
  - QoS
  - Enterprise Linux
  - Web Services
  - ...

Concluding Remarks

- Linux is a key component of IBM's strategy. We are following a multi-pronged approach to accelerate adoption of Linux into the Enterprise
  - Support Linux on all IBM platforms
  - Build strong affinity between IBM operating Systems and Linux
  - Work with the Linux community to infuse technology into the Linux kernel
  - Facilities to assist migration to Linux
  - Deliver robust Linux-based solutions using Open Source and IBM technologies
- We are focusing on developing and deploying technology that will make the configuration, management, and efficient use of Linux systems and Linux Clusters easier in the Enterprise
- Grid computing intersect several key IBM initiatives and strategies. We will work with the community to define and deploy a robust infrastructure and accelerate its adoption across a wider segment

Linux chez IBM

IBM Linux Marketing

Linux at IBM

www.ibm.com/linux

www.ibm.com/developerworks

oss.software.ibm.com/developer/opensource/linux/

Merci!

Questions?