Red Hat Global File System

HP User Society / DECUS
18. Mai 2006

Joachim Schröder
Red Hat GmbH

redhat

Two Key Industry Trends

- Clustering (scale-out) is happening
  - 20% of all servers shipped will be clustered by 2006. - Gartner
  - Linux clusters are growing at a cagr of 44% per year. - IDC
  - 30%+ of Red Hat inquiries are about implementing clusters.
  - Clustering software is a $500m+ market today.
- Storage capacity and complexity are growing
  - Storage (capacity) will continue to grow at 40-60% annually. - Aberdeen.
  - SAN management software use grew by 36% year over year... ... driven "primarily by customer need to support larger and more complex storage networks". - IDC
  - Storage management software is a $5B+ market today.

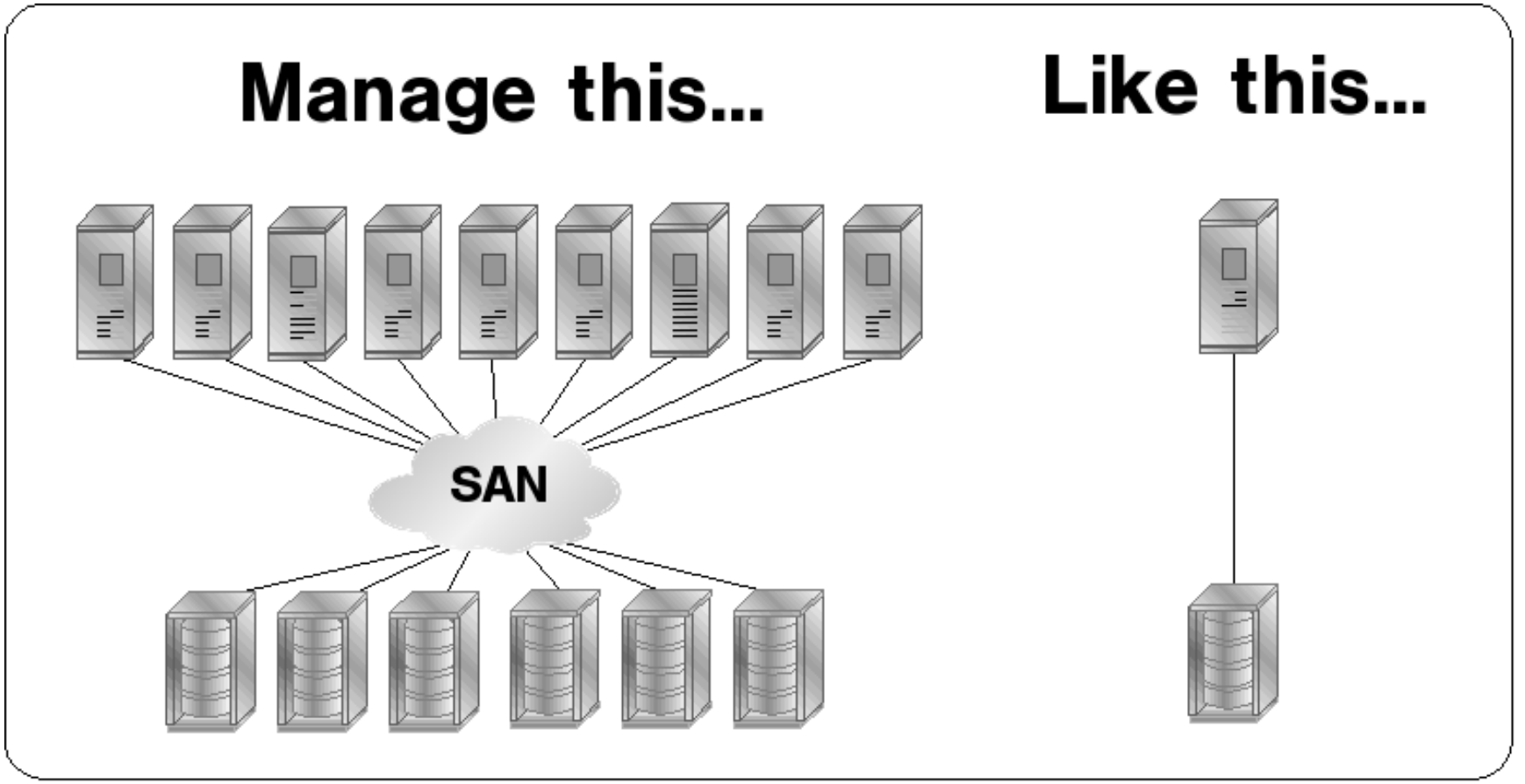
redhat

Clustering - motivation

- Reduce TCO by:
  - Consolidate services to fewer systems
  - Simplify management tasks
  - Increase service uptime
- Approaches:
  - Scale-up: few very powerful systems running multiple services
  - Scale-out: many small systems collaborate on several services
- Requirements:
  - Data sharing
  - Monitoring infrastructure
  - Service management layer

redhat

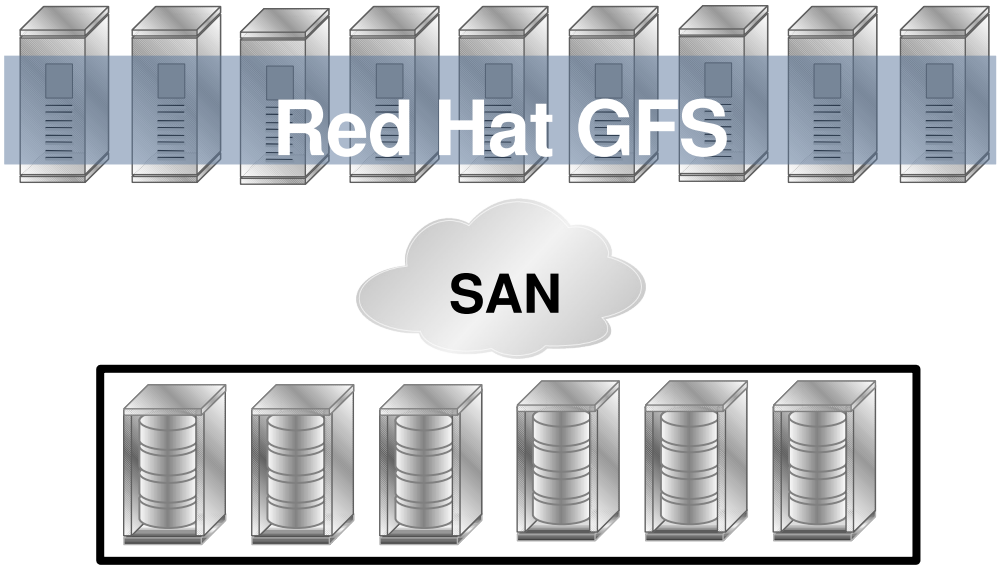
What if you could...



redhat

Red Hat Global File System (GFS)

Allows a **cluster of Linux servers** to **share data files** in a **common pool of storage**



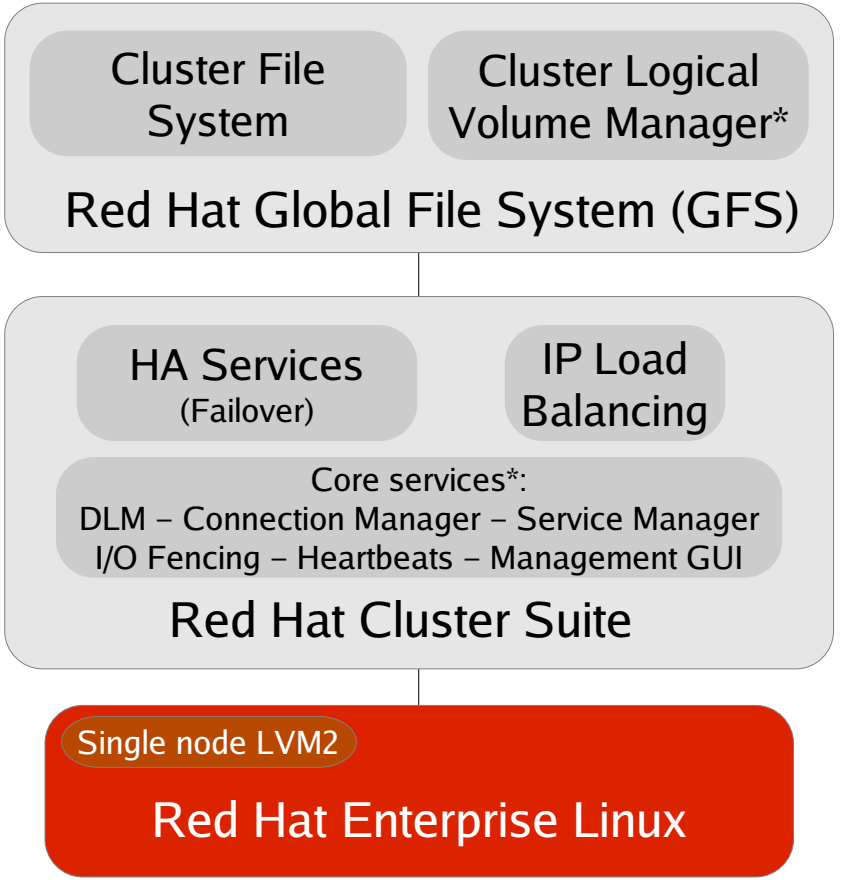
redhat

Storage Management Buzzwords

- Direct-attached storage - *Declining*
  - Storage connected directly to a server (like the hard-drive on your computer)... handles blocks or files
- SAN (Storage Area Network) - *Growing*
  - A specialized network of storage devices connected to servers, usually through a central storage switch using FibreChannel... handles blocks
- NAS (Network Attached Storage) - *Growing*
  - A storage device that is connected to a standard IP network, and has a network name... typically handles files
- IP SAN (IP-based Storage Area Network) - *Growing*
  - A network of storage devices connected to servers using a standard IP network... handles blocks

Red Hat Clustering Architecture

- Red Hat Cluster Suite provides
  - Application failover
  - Improves application availability
  - Included with GFS
  - Core services for enterprise cluster configurations*
- Red Hat Global File System (GFS)
  - Cluster-wide concurrent read-write file system
  - Improves cluster availability, scalability and performance
  - Includes Cluster Logical Volume Manager (CLVM)*



** These features will be available with launch of Cluster Suite v4 and GFS 6.1 – availability concurrent with RHEL4 Update 1*

redhat

Red Hat Cluster Suite: Core Cluster Services

- Core functionality for both Clustering and GFS is delivered in Red Hat Cluster Suite
  - Membership management
  - I/O fencing
  - Lock management
  - Heartbeats
  - Management GUI
- Support for up to 300 nodes
- Two selectable lock management models
  - Client-server with SLM/RLM (single/redundant lock manager)
    - Was in prior version
  - Distributed Lock Manager (DLM)
    - New with Red Hat Cluster Suite v.4
    - Open, stable API – consistent with VMS DLM

redhat

Red Hat GFS: The best cluster file system for Linux

- The **only native 64-bit** general purpose Cluster File System for Linux – support for x86, IA64, AMD64, and EM64T
- **Most scalable Cluster File System on Linux** – supports up to 256 nodes (8x nearest competitor)
- **Tightly integrated with Red Hat Enterprise Linux** (no patching, no lag from RHEL updates), includes up to **7 years of support**
- **Only 3rd party cluster file system validated by Oracle for Oracle RAC**
- **Only open source (GPL) general purpose cluster file system**, headed for upstream adoption:

"Significant features will be added [to the kernel] in the next year, including NFS 4 and clustering file support, possibly Red Hat's Global File System technology."
- Andrew Morton, 2.6 Kernel Maintainer, at Feb. 2005 OSDL Summit

redhat

Red Hat Global File System v6.1

- New version for Red Hat Enterprise Linux v.4
  - Uses new common cluster infrastructure in Red Hat Cluster Suite (included)
- Provides two major technologies
  - GFS cluster file system – concurrent file system access for database, web serving, NFS file serving, HPC, etc. environments
  - CLVM cluster logical volume manager
- Much faster fsck
  - Ported to GFS 6.0 on RHEL 3 Update 5
- Data and meta-data journaling (per-node journals, clusterwide recovery)
- Maximum filesize & file system size: 16TB with 32-bit systems, 8EB with 64-bit systems

redhat

GFS Features (continued)

- Supports file system expansion
- Supports F/C SAN, iSCSI, GNBD
- Uses a Distributed Lock Manager (DLM)
- Supports up to 300 nodes
- Dynamic Inodes allocation
  - No need to guess at file system creation time
- Support x86, x86-64 and ia64 in same cluster
  - Platform Independent metadata
- Extendible Hashing Directories for fast access
- Fuller utilizes Linux buffer and page cache

An open source, **POSIX-compliant, cluster file system.**

redhat

GFS Features (continued)

- Data and meta-data journaling
  - per-node journals, clusterwide recovery
- Context Dependent Path Names (CDPN)
  - Based on hostname, OS, uid, gid, sys, mach
- ACL support
- Quota support
- Freeze/Unfreeze
  - Create consistent backup views.

An open source, **POSIX-compliant, cluster file system.**

redhat

Distributed Lock Manager

- Red Hat Cluster Suite v.4 includes a Distributed Lock Manager (DLM)
  - Primarily used by Red Hat Global File System
  - Available for general purpose use by any application
- A DLM is a highly functional, distributed (cluster-wide), application synchronization subsystem
  - Processes use the DLM to synchronize access to a shared resource (e.g. a file, program, or device) by establishing locks on named resources
  - Provides a collection of services
    - Multiple lock spaces and concurrency (lock) modes
    - Lock hierarchies/domains (resources & subresources)
    - Range locking
    - Lock conversions & value blocks

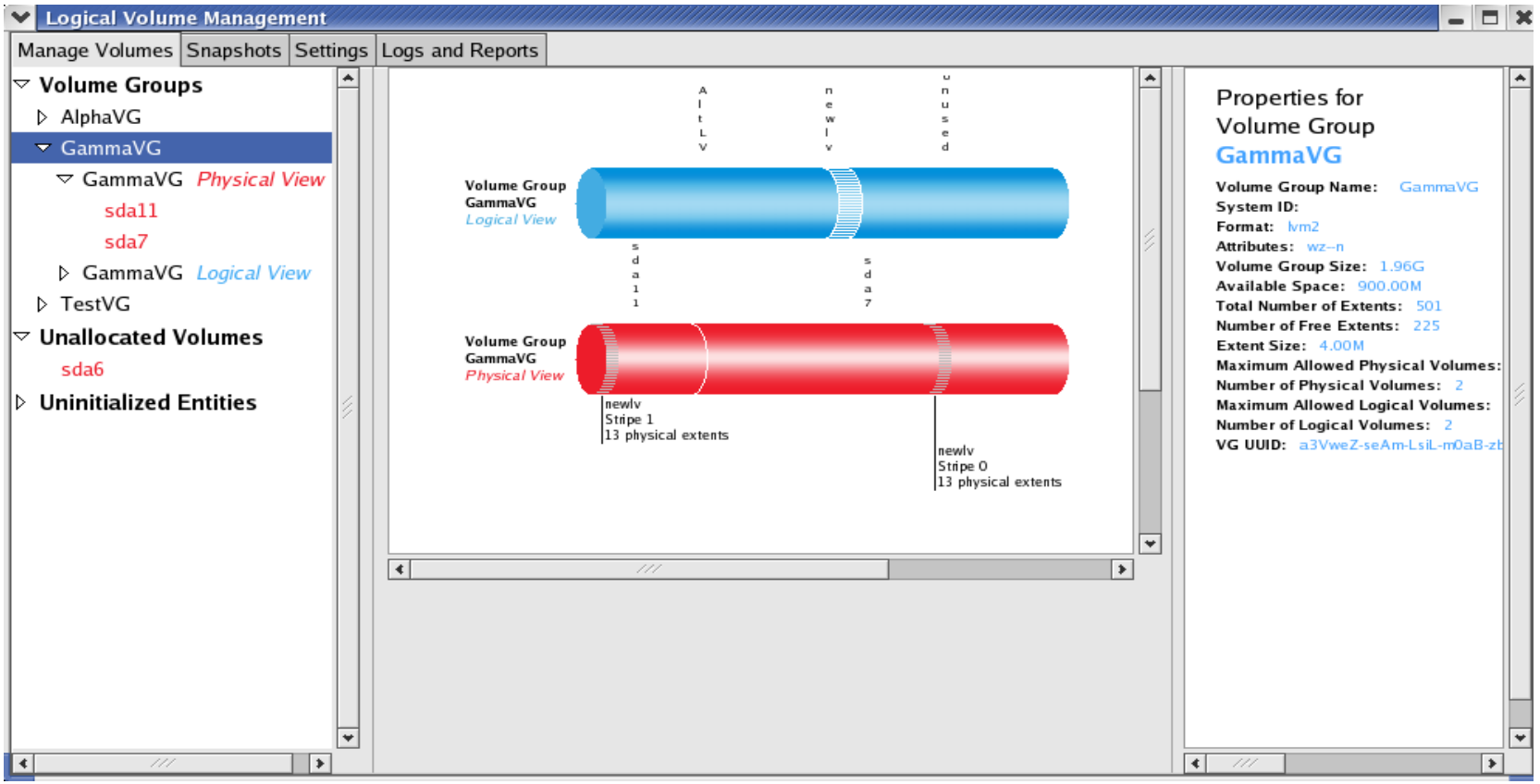
redhat

Cluster Logical Volume Manager (CLVM)

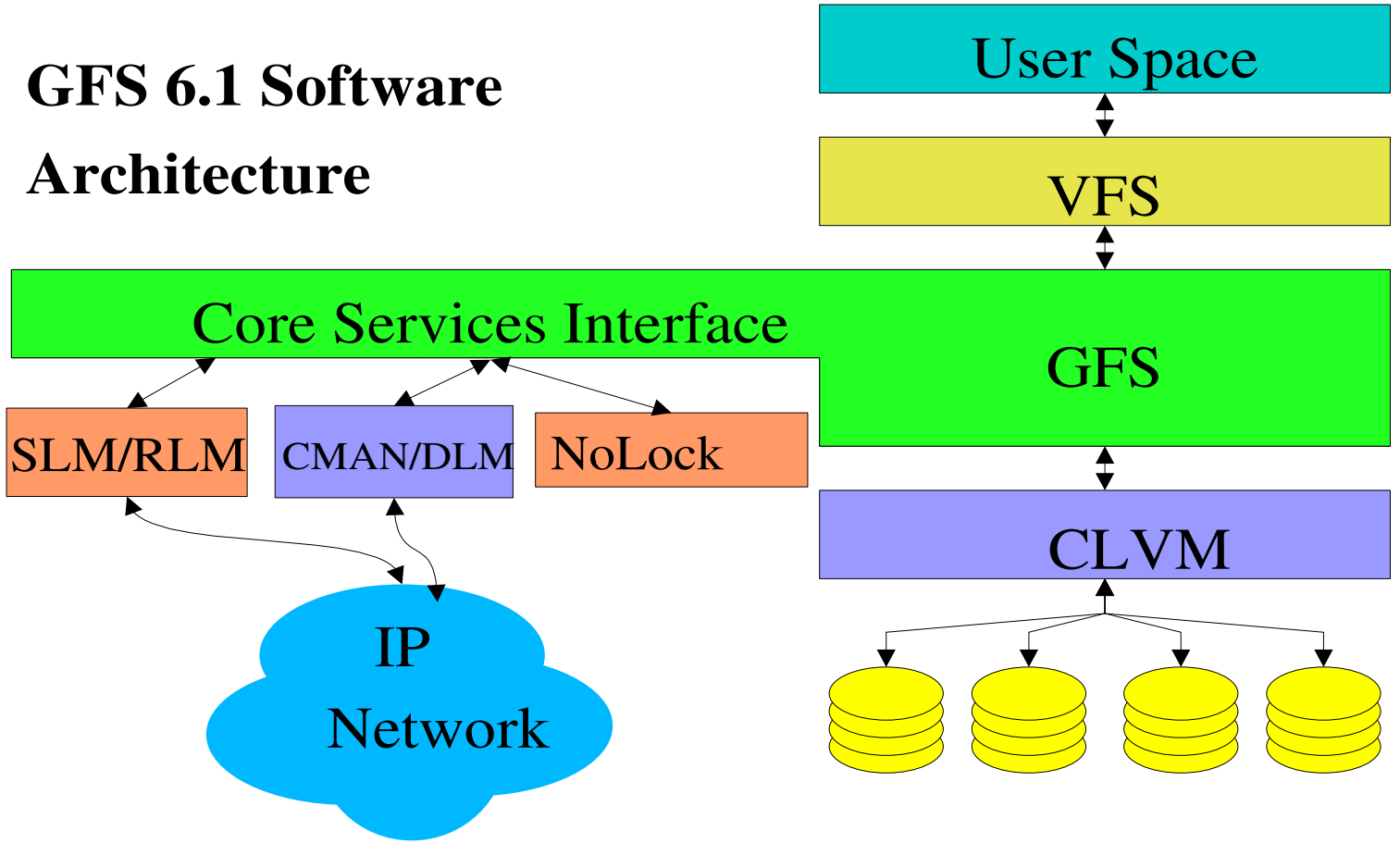
- CLVM builds upon LVM 2.0 and the kernel device mapper component included in 2.5 and 2.6 Linux distributions
- Essentially a cluster-aware version of LVM 2.x
- Commands, features, functions all work just fine in a cluster, any Linux server may mount any volume
- Provides
  - Cluster safe volume operations
  - Cluster-wide concatenation and stripping of volumes
  - Dynamic Volume resizing
  - Cluster-wide snapshots (mid '06)
  - Cluster-wide mirroring (mid '06)
  - Other RAID levels (mid '06)

Storage: Logical Volume Management (LVM2)

- LVM2 provides significantly improved GUI-based storage management capabilities
  - Goal to provide consistent, easy to understand, administrator interface

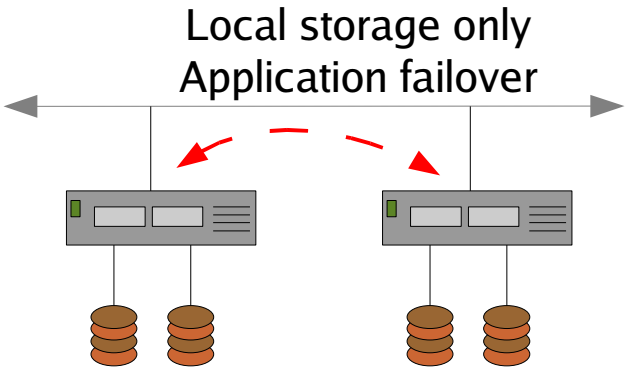


Red Hat Global File System v.6.1



Use Case: High availability with local data

- Red Hat Enterprise Linux (ext3) + Red Hat Cluster Suite v4
- Requires NO ADDITIONAL HARDWARE
  - No physically shared storage
    - Prior to RHEL4, physically shared storage was required
- Ensures that an application stays running
  - Monitors all cluster nodes
  - Fails-over applications/services from stopped nodes
  - Simple management GUI
  - Service definitions are automatically propagated and synchronized, cluster-wide



H/A Cluster

Topology:
- 2 or mode nodes
- NIC card
- NO shared storage

Applications:
- Read-only data
- Small web serving
- NFS/FTP serving
- Edge of network

Data model:
- Unshared, static data
- Replication w/ rsync
- ext3 file system

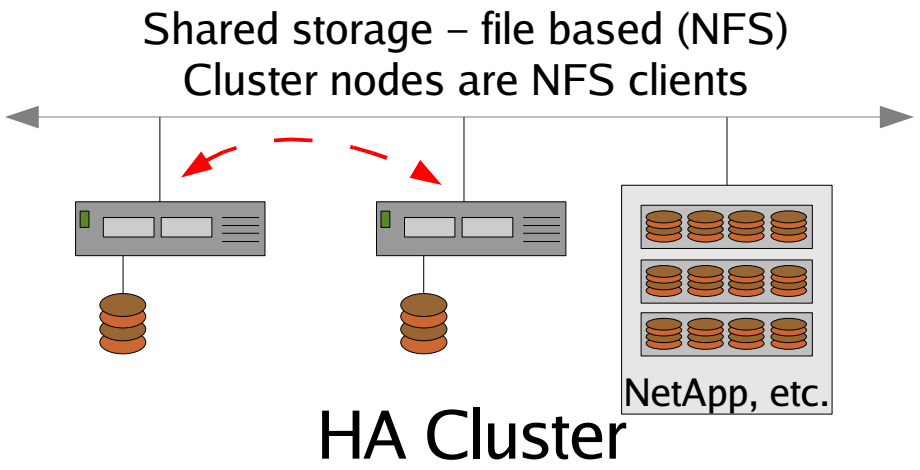
Use Case: Shared file datastore, NAS

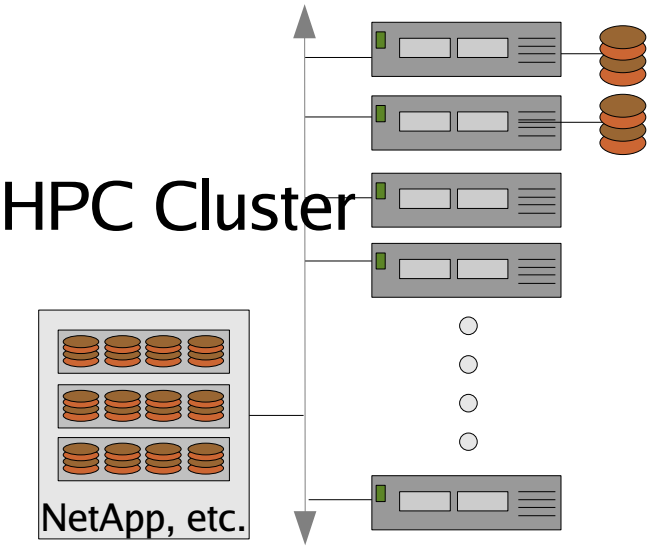
- Red Hat Enterprise Linux (NFS)
- Uses an external NFS file server

Addnl. Cluster features:
- Increased storage capacity
- Applications can failover and access the same data as before
- NFS sharing model

Applications:
- Small/medium web serving
- NFS/FTP serving
- Read/write data
- File level access

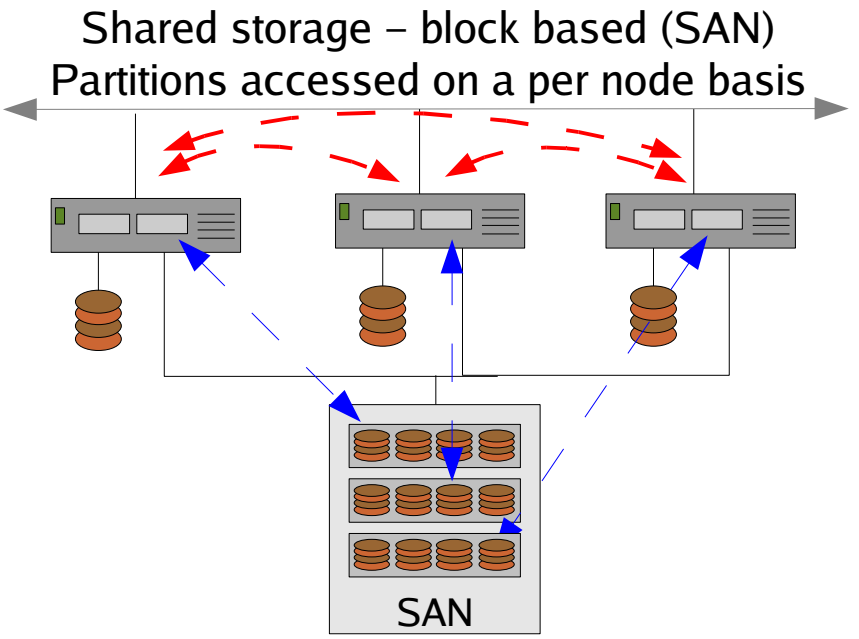
Data model:
- Physically shared data
- File based (NFS)





Use Case: Shared block datastore, SAN

- Red Hat Enterprise Linux (ext3, integrated HBA drivers)
- Offers higher, SAN-based, performance



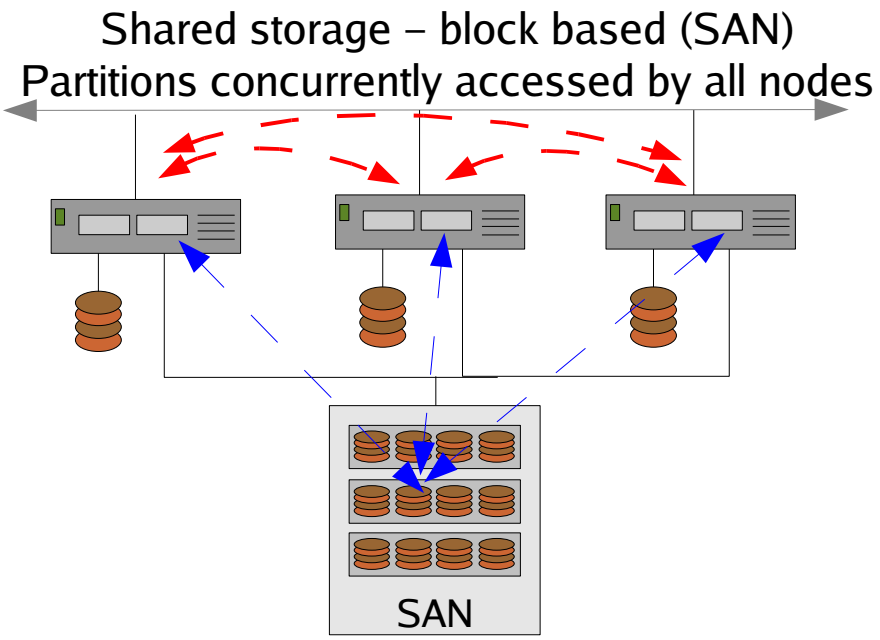
Addnl. Cluster features:
- SAN access provides:
-- improved performance
-- heterogeneous access by other systems

Applications:
- Medium/large web serving
- Medium database
- Read/write data
- Block level access

Data models:
- Storage Area Network
- Physically shared storage
- Block based
- ext3 file system

Use Case: Concurrent file system access

- Red Hat Enterprise Linux + Red Hat Global File System
  - Includes Red Hat Cluster Suite
- Shared file system access
- Scalable performance
- Same hardware as previous configurations
- Common software infrastructure with previous configuration



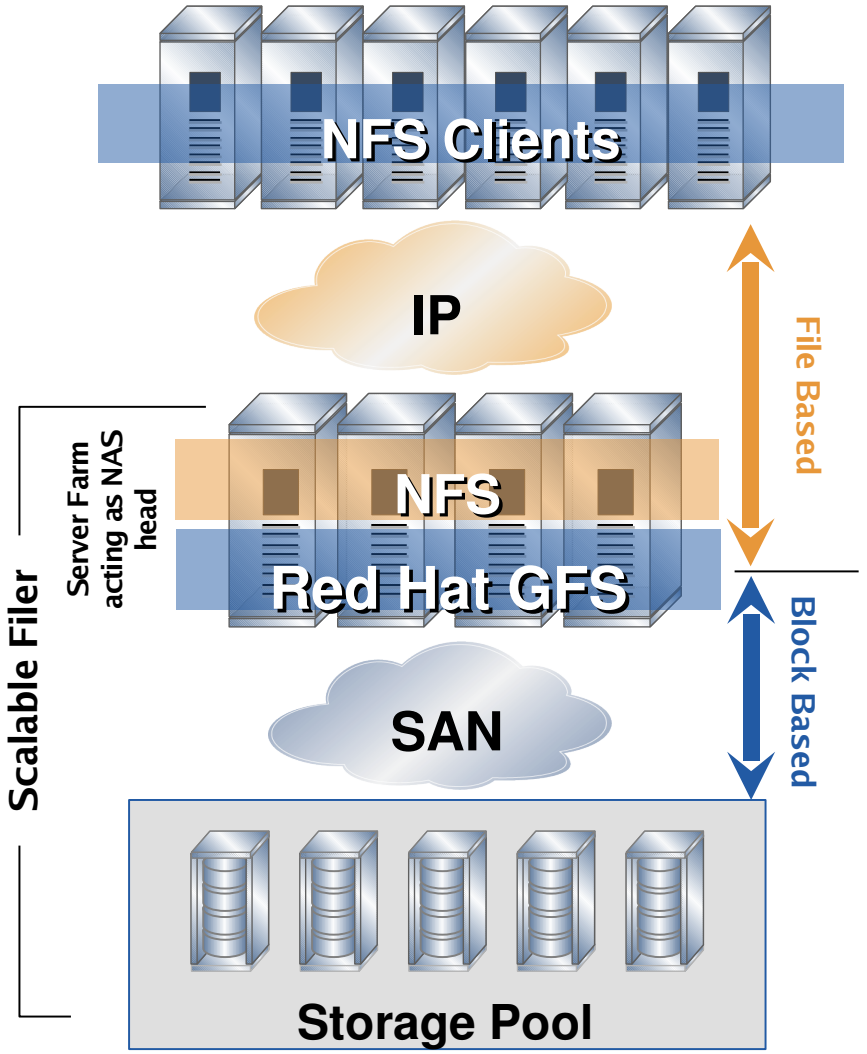
Applications:
- Large web serving
- Large database
- Read/write data
- Block level access
- Concurrent access

Data models:
- Storage Area Network
- Shared file system access
- Block based
- GFS file system

Addnl. Cluster features:
- SAN access provides:
-- improved performance
-- heterogeneous access by other systems

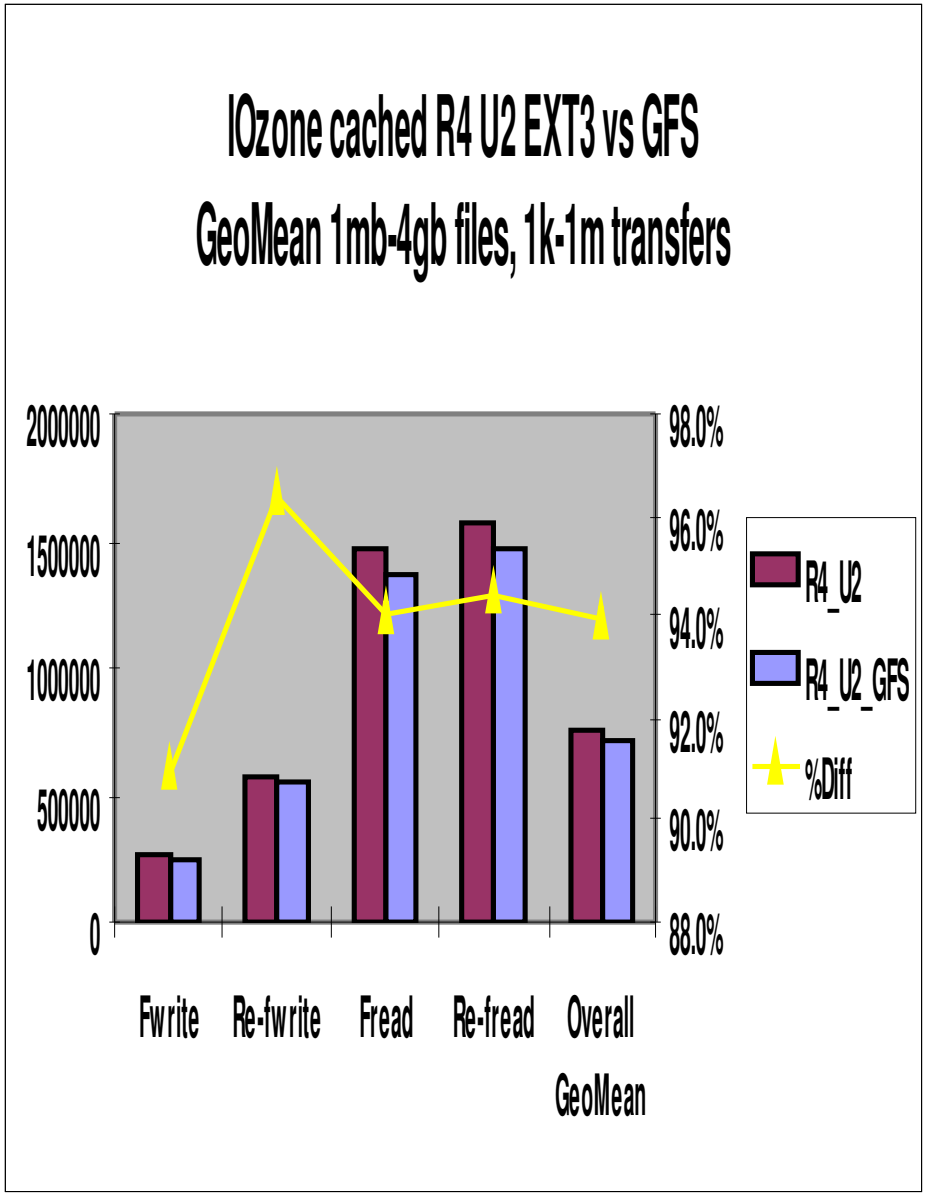
redhat

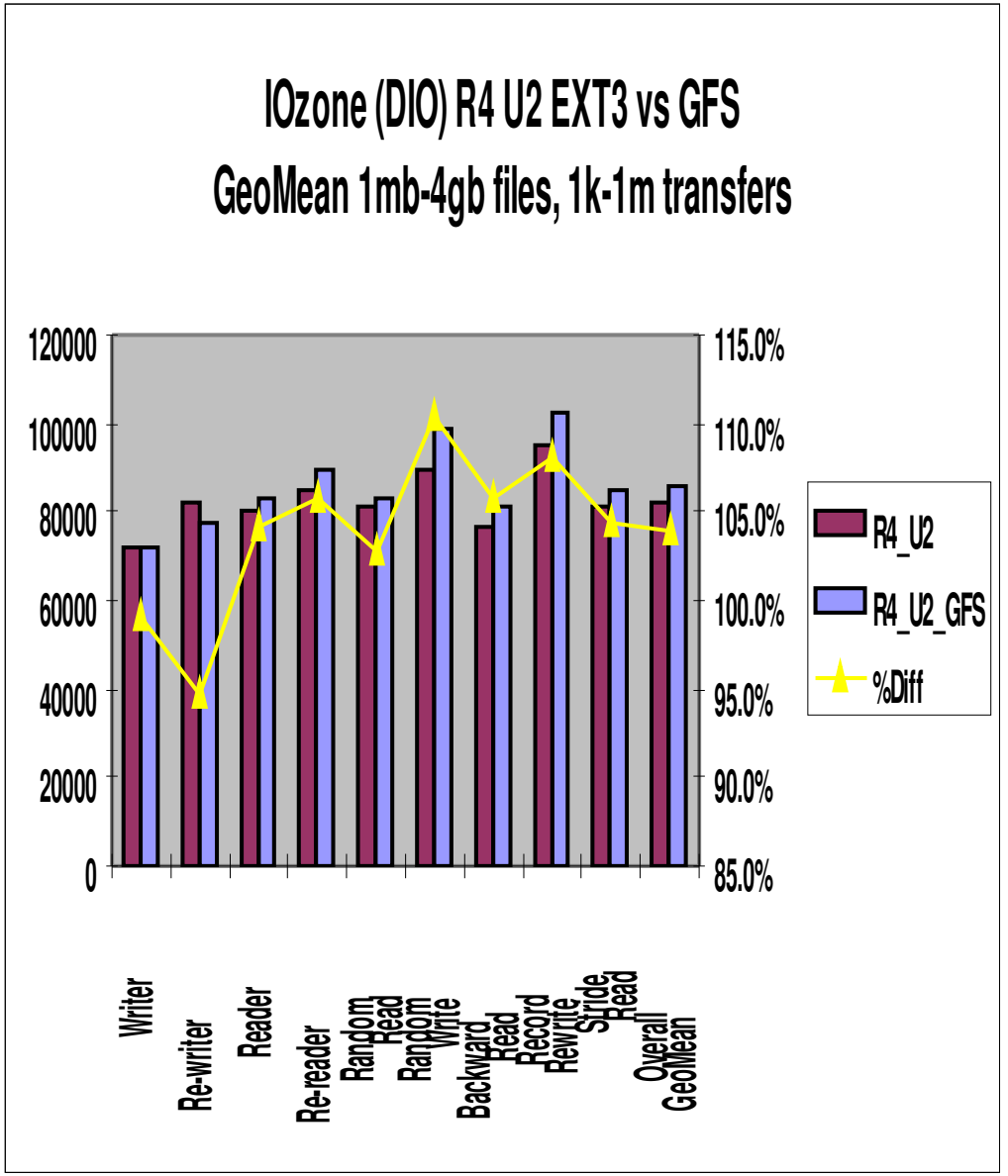
Red Hat GFS can make NFS scale & perform



- Eliminate NFS performance and scalability limitations
- Create a scalable NAS-like cluster with no single point of failure
- Dynamically add compute and I/O resources
- Streamline development environments and accelerate build times
- Eliminate need for data duplication
- Real case: 256 nodes, 10 GFS servers serving NFS

GFS vs EXT3 Iozone Comparison





redhat

Oracle RAC without GFS

- ***Management Tasks for an 8-node Oracle RAC cluster:***
  - An Oracle home directory for each node (and node-specific information for each node) must be created. An eight-node configuration will have:
    - Eight local disks, one for each node (or similarly, eight physical or virtual storage devices on a storage network)
    - Eight Oracle home directories, and node-specific information on each of the eight local disks—requiring each local disk to be managed and protected
  - A raw storage device must be created and used for each table space
    - A minimum of 10 raw devices to manage, including devices for Oracle tables (system, undo, redo, temp, users, tools, index) and user space data.
  - Downtime and administrative effort
    - Raw devices cannot be dynamically expanded. Therefore, the table spaces mapped to these devices cannot be expanded dynamically while the Oracle cluster remains online. In addition, updates to the Oracle home directories on each node are required whenever the Oracle program files are updated.

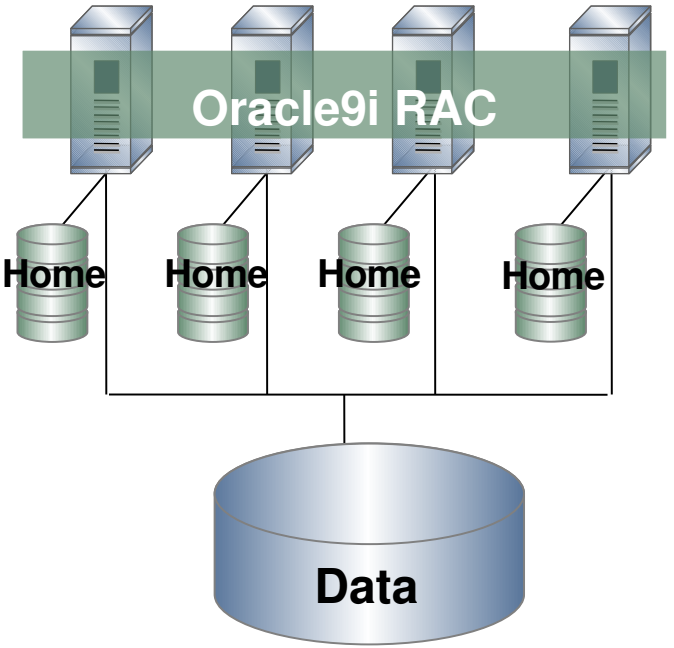
redhat

Oracle RAC with Red Hat GFS

- ***Management Tasks for an 8-node Oracle RAC cluster:***
  - ONE SINGLE Oracle home directory for the entire cluster is created. An eight-node configuration will have:
    - Shared Oracle Home directory, with centrally stored node-specific information for each of the 8 nodes.
  - Storage can be provisioned and grown dynamically
    - GFS has raw-like performance
    - No longer necessary to take Oracle offline to grow storage
  - Reduced Downtime and administrative effort
    - Oracle updates can be applied once, for the entire cluster
  - Any other POSIX compliant applications can be run on the RAC nodes

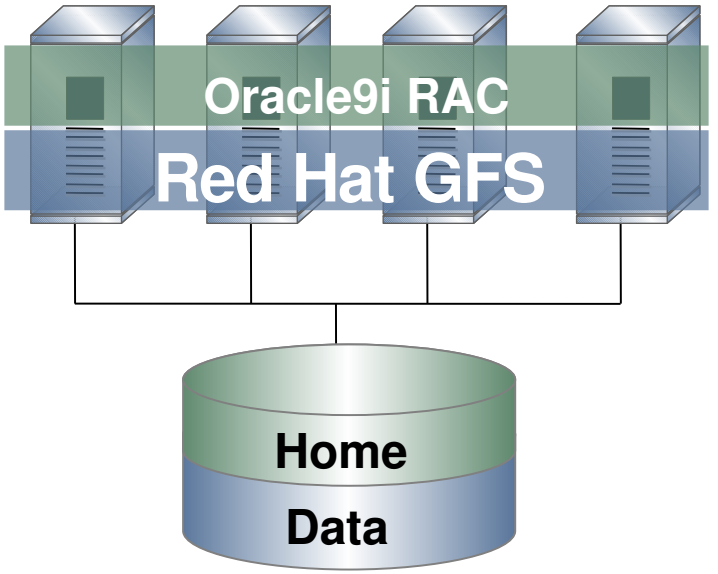
Red Hat GFS & Oracle 9i RAC

Without GFS



- 105,000 Oracle Home files replicated on private direct-attached storage
- Additional storage and management overhead

With GFS



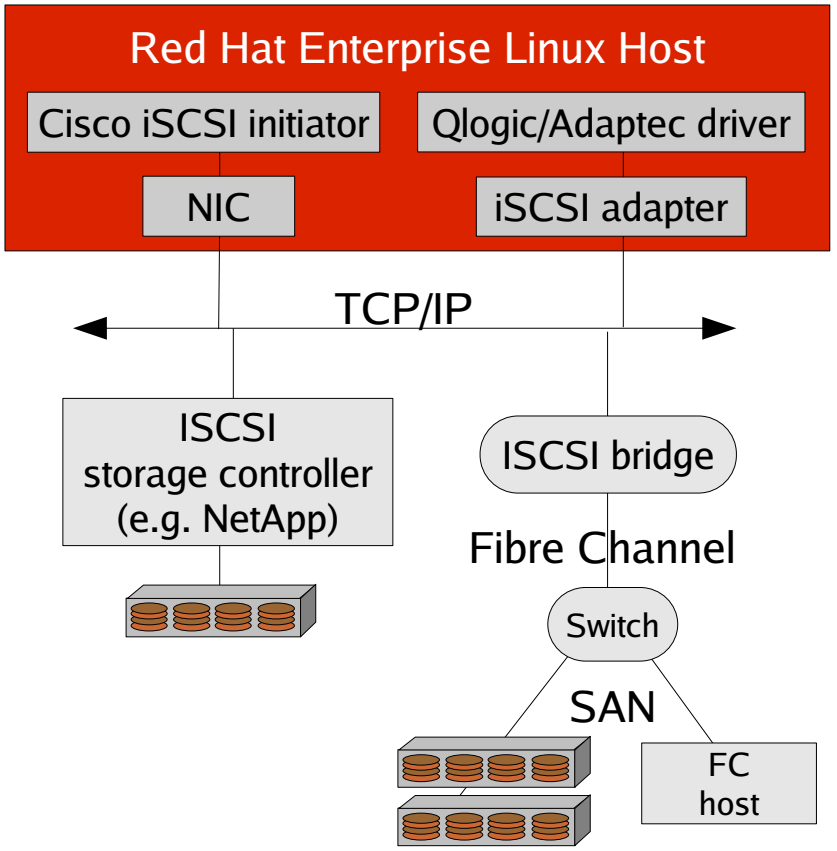
- Oracle Home files shared
- Data sharing for DB & other apps
- Embedded Volume Management
- Database management and capacity planning simplified

Fibre Channel Device Drivers

- RHEL3 U4
  - Emulex Driver Update - 7.1.14
    - EMC certification on version included in U4
  - Qlogic Driver Update - 7.01.01
  - I/O Performance
    - Some improvement due to virtual memory scheduling changes
- RHEL4
  - Greatly increased support with over 4,000 SCSI devices/paths (was 256 in RHEL 3)
  - Driver versions tracking upstream submissions very closely
  - Goal is to keep them current as much as possible (e.g. F/C 4GB, SATA 2)

iSCSI

- Low-cost Enterprise SAN connectivity
- Driver in RHEL 3 U4
  - Open source Cisco implementation
  - Initiator only
- RHEL4 Update 2
  - Rewrite for 2.6 kernel (based on 2.4 driver)
  - Undergoing rapid change and development
  - Driving upstream adoption
- Boot support planned for a later update
- No iSCSI Target yet
  - Currently no viable open source option
  - Possible feature in later
- Qualified with major storage vendors
  - NetApp
  - EMC
  - EqualLogic



redhat

Red Hat and GFS Joint Linux Clustering Solution

- HP Serviceguard for Linux creates a cluster of Linux servers that make application services available despite hardware or software failures or planned downtime.
- Red Hat Global File System (GFS) is a cluster file system that enables multiple servers to simultaneously read and write to a single shared file system on a Storage Area Network.
- Recently tested. Working together, now provide best of breed clustering technologies for Linux. The combination offers greater availability of applications and data using a cluster of low-cost Linux servers.
  - White Paper
  - Solution Brief
  - Joint marketing

Questions?

Vielen Dank!

Joachim Schröder, Solution Architect

joachim.schroeder@redhat.com