Red Hat Enterprise Linux: Creating a Scalable Open Source Storage Infrastructure

By Alan Radding and Nick Carr

Abstract

This paper discusses the issues related to storage design and management when an IT infrastructure is migrated from a vertically scaled, large SMP architecture to a horizontally scaled server infrastructure--with large numbers of low-cost commodity servers running Linux to deliver industrial-strength processing. A discussion of the storage capabilities of Red Hat Enterprise Linux , including kernel-level, file systems, logical volume management and other features is included, along with an overview of the storage capabilities of Red Hat's ISV partners. The paper concludes with a discussion of the challenges of storage design and management in a horizontally scaled infrastructure, and offers guidance on the best storage management options available to customers planning large-scale Linux deployments.

June 2004

Introduction: Storage for Horizontal Scalability

Many companies are discovering that horizontal scalability is a better, more cost-effective way to meet their growing requirements for server processing capability. Horizontal scalability is the process of deploying large numbers of low-cost commodity servers running an open source operating system like Red Hat Enterprise Linux to deliver large-scale, industrial-strength processing. Previously, organizations requiring such large-scale processing followed a vertical approach to scalability by deploying increasingly larger, expensive, proprietary SMP servers.

The horizontal approach to scalability offers a number of advantages. By taking advantage of low-cost commodity processors, the servers themselves, even deployed in large volumes, are less expensive than the proprietary vertically scaled servers, yet deliver as much or more processing power. In addition to low cost, the horizontal approach offers more flexibility and availability options.

However, despite the clear advantages of such horizontal scalability, some companies have been hesitant to implement adoption programs due to concerns about how to integrate rapidly proliferating storage subsystems to their existing enterprise storage infrastructure, particularly their storage area network (SAN). While they are successfully managing their SAN and network storage infrastructure with a small number of large servers, they are concerned that migrating to an environment with a large number of small servers will considerably complicate storage management. Quite understandably, these companies want to continue to experience the benefits of shared access provided by networked enterprise storage and the low total cost of ownership it delivers.

Fortunately, with Red Hat Enterprise Linux, organizations can pursue a horizontal scalability strategy and fully leverage their existing investment in shared enterprise storage. Additionally, technologies that enable local (non-networked) storage to be consolidated and managed as part of the overall storage infrastructure are developing rapidly. Servers running Red Hat Enterprise Linux can connect with the organization's NAS and SAN storage via IP, or Fibre Channel. In addition, companies will find an extensive range of software to support most storage devices and hardware components, as well as interoperability with leading third-party storage management tools, such as those from VERITAS.

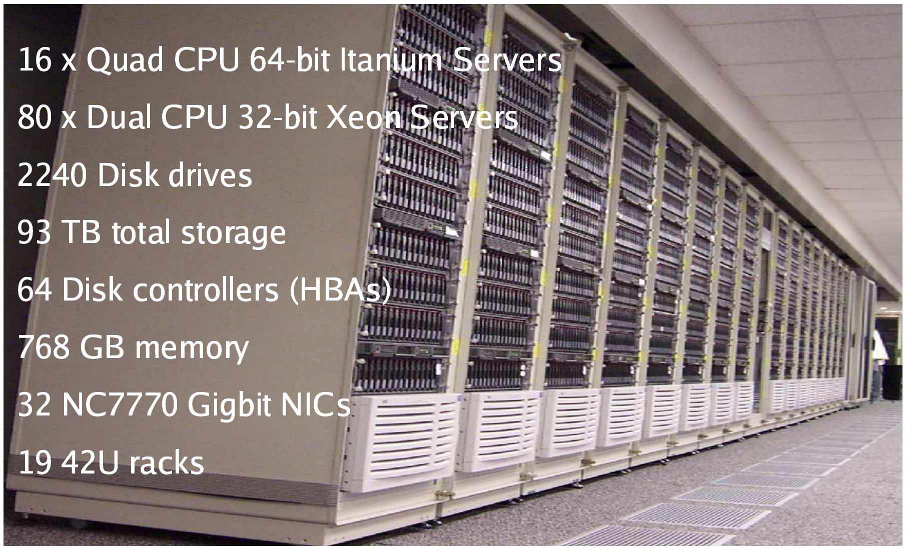
This paper should put managers at ease concerning any storage issues arising from an open source-based horizontal scalability strategy built around Red Hat Enterprise Linux. It will describe Red Hat's current storage capabilities, identify challenges of enterprise storage, and introduce Red Hat storage enhancements to meet these challenges.

Red Hat Enterprise Linux core storage capabilities

Out of the box, Red Hat Enterprise Linux comes with the kind of robust capabilities needed for enterprise storage. Linux in general brings an extensive set of drivers for myriad storage hardware—disk arrays, host bus adapters, network cards. Drivers are readily available and regularly updated by hardware vendors or Linux distribution maintainers. When it comes to storage in general, Linux is on par with any flavor of UNIX and other operating systems.

Red Hat Enterprise Linux kernel

In technical terms, the current Red Hat Enterprise Linux v.3 kernel provides support for volumes up to 1TB in size. Recent file and I/O system development work has enabled much larger file systems to be supported and the next major release of Red Hat Enterprise Linux, which will incorporate the Linux 2.6 kernel, extending the block device limit to 2^64 bytes--virtually limitless. In a 64-bit system this essentially translates into unlimited volume size. In general all the usual limitations—file system size, file size, buffer cache size, I/O transfer size, etc. —are being steadily worked on by Red Hat and the open source community as demanded by customers. In addition, upcoming versions of Red Hat Enterprise Linux will have support for tens of thousands of devices, overcoming the current upper limit of only a few hundred. This is an important capability, as storage area networks with thousands of storage devices are no longer uncommon. New systems will be able to accommodate increasingly large physical storage systems, comprising ever more host bus adapters and devices. The recently published world record TPC/C benchmark, achieved by a partnership of Oracle, Red Hat, and HP, included a storage system with over 2000 physical disks providing a total of 93TB of data, hosting a 90TB database. So the ultimate scalability of Linux is not in doubt. The photograph below shows the actual benchmark configuration. Details of the benchmark can be found at http://www.tpc.org.


16 x Quad CPU 64-bit Itanium Servers
80 x Dual CPU 32-bit Xeon Servers
2240 Disk drives
93 TB total storage
64 Disk controllers (HBAs)
768 GB memory
32 NC7770 Gigbit NICs
19 42U racks

Of greater interest from an enterprise storage standpoint is Red Hat Enterprise Linux' s support for journaled, networked, and clustered file systems as well as logical volume managers, and its interoperability with third-party management storage software.

File systems

Red Hat Enterprise Linux allows organizations to run a choice of file systems. Particularly interesting to enterprise storage users are the default Linux file system, Ext3 (Third Extended file-system), NFS (Network File System), and Red Hat' s GFS cluster file system. Specifically:

- Ext3 is a journaling file system, which uses log files to preserve the integrity of the file system in the event of a sudden failure. It is the standard file system used by all Red Hat Enterprise Linux systems.
- NFS is the de facto standard approach to accessing files across the network
- GFS (Global File System) allows multiple servers to share access to the same files on a SAN while managing that access to avoid conflicts. Sistina Software, the original developer of GFS, was acquired by Red Hat at the end of 2003. Subsequently, Red Hat contributed the software to the open source community under the GPL license, in keeping with Red Hat' s development methodology. GFS is provided as a fully supported, optional layered product for Red Hat Enterprise Linux systems.

Using an NFS-served Ext3 file system, or GFS, enterprises are able to deploy large pools of shared storage that can be accessed by multiple servers.

Logical Volume Manager

Red Hat Enterprise Linux includes the Logical Volume Manager (LVM), which provides kernel-level storage virtualization capabilities. The LVM allows system administrators to combine physical storage elements into a collective storage pool, which can then be allocated and managed according to application requirements, without regard for the specifics of the underlying physical disk systems. The LVM is the key to efficient, responsive enterprise storage management.

Initially developed by Sistina and now part of the standard the Linux kernel, LVM has been adopted by the open source development community for online disk storage management with all major Linux distributions. LVM provides robust, enterprise-level volume management capabilities that are consistent with the leading, proprietary enterprise operating systems. Development of LVM capabilities is progressing rapidly, and as of mid-2004 these include:

- Storage performance and availability management by allowing for the easy addition and removal of physical devices and through dynamic disk volume resizing. Logical volumes can be resized dynamically online, while the Ext3 supports offline file system resizing (requiring unmount, resize, and mount operations).
- Disk system management that enables system administrators to easily upgrade disks, remove failing disks, reorganize workloads, and adapt storage capacity to changing system needs

Additional and future features

Included in Red Hat Enterprise Linux is a capability called Devlabel, which allows administrators to give storage devices persistent names. This eliminates the problem of device names changing when a system is reconfigured. Devlabel can also be used to create consistent names on systems with shared storage and has a provision for automatically detecting and naming multi-path devices.

Multi-path access to storage is essential to continued availability in the event of path failure (such as failure of a Fibre Channel adapter). When providing multi-path support the operating system must correctly identify the same single device at the end of multiple paths, as opposed to configuring multiple devices. Red Hat Enterprise Linux's multi-path device driver (MD driver), recognizes multiple paths to the same device, eliminating the problem of the system assuming each path leads to a different disk. MD driver combines the paths to a single disk, enabling failover to an alternate path if one path is disrupted.

Additional features that are currently available and/or under development for delivery between now and the next 12-18 months include:

- Snapshot capability to enable fast, consistent backups or point-in-time recovery of damaged or mistakenly deleted data.
- Host-based RAID-1 (mirroring), which enable disks to be combined into high availability configurations regardless of their physical connectivity (with controller based RAID-1 all devices in the mirror set are usually connected to the same controller).
- Delivery of all storage management capabilities for use in clustered environments. This ensures that all systems in the cluster maintain a consistent view of the underlying storage configuration, both from a hardware and software standpoint.

Storage capabilities from Red Hat partners

In addition to providing storage management capabilities in the base Red Hat Enterprise Linux product set or as optional layered products, Red Hat has worked closely with leading storage software providers to ensure their products are fully supported. This flexible approach results from the realization that while some customers wish to use open source storage management software, others have standardized or proprietary third-party solutions. Providing a choice of solutions is a critical feature of Red Hat' s product strategy. An example of this approach is that storage management products from VERITAS, a recognized leader in the field, are supported on Red Hat Enterprise Linux. Red Hat also partners with other leading storage vendors, including Network Appliance (NetApp) and EMC/Legato.

Storage Options

Enterprises today embrace a wide range of storage options. Although SAN and NAS have emerged as the preferred enterprise storage approach, direct attached storage remains widespread throughout the enterprise. Red Hat Enterprise Linux supports the full set of enterprise storage options:

- Direct attached storage
    - SCSI
    - ATA
    - Serial ATA
    - SAS (Serial Attached SCSI)

- Networked storage
    - SAN (access to block-level data over Fibre Channel or IP networks)
    - NAS (access to data at the file level over IP networks)

- Storage interconnects
    - Fibre Channel (FC)
    - iSCSI
    - GNBD (global network block device)
    - NFS

Challenges of enterprise storage

Traditionally storage has been directly associated with a specific server through direct attached storage. In the open systems environment, this typically took the form of a SCSI connection to a storage array. More recently, those storage arrays were configured in any of a number of RAID (redundant array of inexpensive disk) formats for purposes of high availability and performance.

Today, enterprises still deploy large amounts of direct attached storage, although the proportion is steadily declining as the deployment of networked storage, particularly SAN, increases. According to industry researcher IDC: In the US market for external disk storage systems, revenue for storage attached network (SAN) systems will exceed the combined value of network attached storage (NAS) and direct attached storage (DAS) shipments in 18 of 20 vertical markets by 2007.

This shift to networked storage is being driven by cost and productivity. Although simple to deploy, direct attached storage proves costly and cumbersome as the number of servers and their corresponding storage proliferates throughout the enterprise. Management of direct attached storage requires intensive amounts of labor, and it is impossible, short of physically pulling disk and cables, to shift a server’s unused storage capacity to another server that needs more storage. Similarly, it is complicated for users to navigate to different data accessed through different servers. These factors drive up the cost of operating the storage environment while driving down worker productivity.

Networked storage, on the other hand, enables the enterprise to create logical pools of storage that can be accessed by any authorized user or application over the network. Although the initial implementation is more complicated and costly, networked storage can be managed centrally and, with the right management tools, allocated and reallocated as needed without having to physically move disks and cables. In addition, data can be efficiently backed up or replicated for availability. As such, networked storage lowers the cost of operating the storage environment, enables greater storage utilization, and expedites access to storage, thus increasing worker productivity.

Specifically, enterprise system administrators want to implement their storage in such a way as to:

- Share data among multiple servers, enabling any server to access any data (for which it is authorized) anywhere on the network
- Scale (up or down) as needed by adding and removing, allocating and re-allocating storage on the fly
- Converge and consolidate data by combining block-level and file-level data within the same storage infrastructure
- Ensure high availability through the use of mirroring, replication, snapshots
- Protect data through backup and recovery
- Control costs and reduce total cost of ownership through centralization, virtualization, and automation

Red Hat Enterprise Linux provides the capabilities that enable enterprise storage managers to meet these challenges.

Advanced networked storage options for Red Hat Enterprise Linux

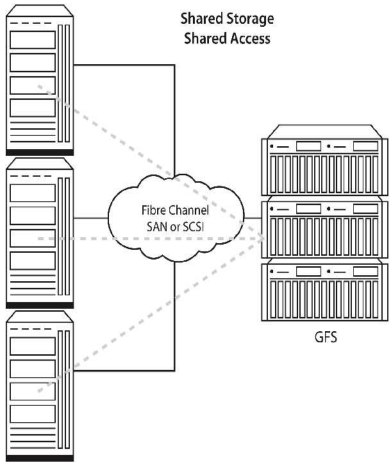
Using the core storage capabilities described above – LVM, GFS, devlabel, etc. Red Hat Enterprise Linux provides the features necessary to make it effective for enterprise storage. Red Hat GFS is the single most important of all these features. It allows multiple servers to attach to a SAN and share a common file system mapped onto a shared storage device, an architecture known as a data-sharing cluster. A cluster file system such as GFS eliminates conflicts and problems that arise when different servers want to access the same file at the same time. Using the GFS file system, system administrators can configure data sharing clusters using a group of Red Hat Enterprise Linux servers. Authorized servers and users can access and share data on the existing enterprise SAN via FC or iSCSI (or GNBD, describe below) connections.

The combination of Red Hat GFS and LVM establishes the foundation for an advanced enterprise storage infrastructure that provides the following:

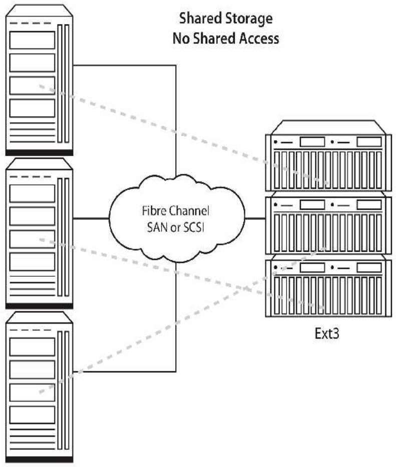
- Enterprise-wide data sharing among multiple servers
- Virtualization of multiple disk storage systems
- Easy scalability, up or down
- Virtual single-system image for ease of management and ease of use

Enterprise storage topologies

Storage deployments for the enterprise use several different topologies, selected based on performance and cost. At the high end (in terms of cost and performance) is the SAN. In a typical SAN configuration each server is configured with a Fibre Channel host adapter that is connected to an external Fibre Channel storage array. In a large configuration a Fibre Channel hub or switch will be used to increase the connectivity, and improve performance and availability. SANs provide direct block-level access to storage. When deploying a SAN with the Ext3 file system, each server mounts and accesses disk partitions individually. Concurrent access is not possible. When a server shuts down or fails, the clustering software will “failover” its disk partitions so that a remaining server can mount them and pick up its tasks. Deploying GFS on SAN-connected servers allows full sharing of all file system data, concurrently. GFS technology provides unprecedented flexibility and scalability. These two configuration topologies are shown in the diagrams below.

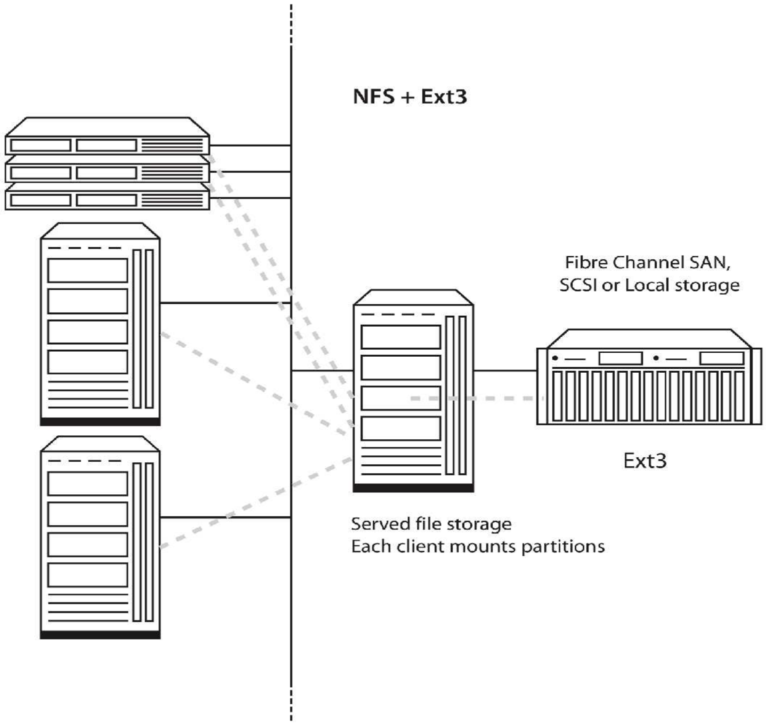


Shared physical storage
Servers mount GFS disk partitions concurrently
Server failure does not perturb other servers



Shared physical storage
Each server mounts Ext3 disk partitions privately

NFS environments have been available for many years and their topologies are well known. In general, an NFS file server, usually configured with local storage, will serve file-level data across a network to remote NFS clients. This topology is best suited for non-shared data files (individual users'   directories, for example) and is widely used in general-purpose computing environments. NFS configurations generally offer lower performance than block-based SAN environments, but they are configured using standard IP networking hardware so offer excellent scalability. They are also considerably cheaper. An NFS topology is shown in the following diagram:

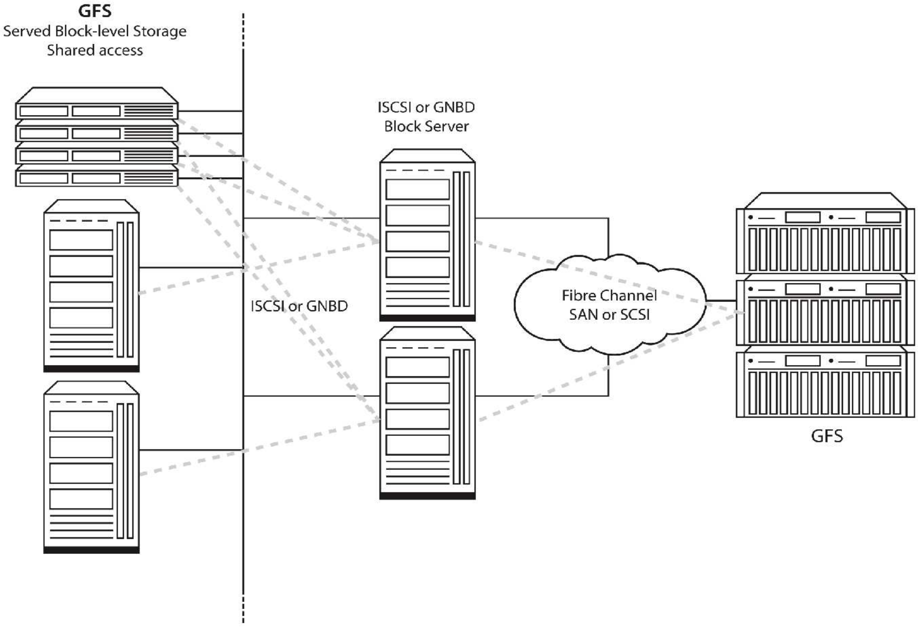


Combining the performance and sharing capabilities of a SAN environment with the scalability and cost effectiveness of a NAS environment is a highly desirable goal for customers and storage hardware and software vendors. A topology that achieves this uses SAN technology to provide the core (“back end”) physical disk infrastructure, and then uses block-level IP technology to distribute served data to its eventual consumer across the network. The emerging technology for delivering block-level data across a network is iSCSI. This has been developing slowly for a number of years, but as the necessary standards have stabilized, adoption by industry vendors has started to accelerate considerably. Linux support for iSCSI exists today and is maturing rapidly.

As an alternative to iSCSI, Red Hat Enterprise Linux provides support for Red Hat’s Global Network Block Device (GNBD) protocol, which allows block-level data to be accessed over TCP/IP networks. The combination of GNBD and GFS gives system adminstrators additional configuration flexibility for sharing data on the SAN and throughout the enterprise. This topology allows a GFS cluster to scale to hundreds of servers, which can concurrently mount a shared file system without the expense of including a Fibre Channel HBA and associated Fibre Channel switch port with every machine. In effect, system administrators can make SAN data available to many other systems on the network without the expense of a Fibre Channel SAN connection. Today, GNBD and iSCSI offer similar capabilities, however GNBD is a mature technology while iSCSI is still relatively new. Red Hat provides GNBD as part of Red Hat Enterprise Linux so that customers can deploy IP network-based SANs today. As iSCSI matures it is expected to supplant GNBD, offering better performance and a wider range of configuration options. An example configuration is shown in the diagram below.

Using iSCSI or GNBD, storage architects decouple the storage network port connection from the servers, effectively putting multiple servers through the same storage network port connection. This approach allows the system architect to achieve specific price, performance and storage capacity objectives more easily than with systems designed using only Fibre Channel for the SAN. With the ability to independently scale storage and capacity, a wide variety of file and database applications can be addressed in a cost-effective manner. In just the same way, a SAN switch with both Fibre Channel and iSCSI ports can amortize expensive Fibre Channel connectivity across a less expensive, more scalable Ethernet-based IP network.

For example, an organization can use GFS to cluster 128 servers connected via Gigabit Ethernet to 16 GNBD servers connected, in turn, to a SAN network with 16 shared storage devices. Instead of paying about $500,000 to directly attach each GFS server to the SAN (128 servers at $3000 per director port cost plus $1000 per FC HBA) the cost with GNBD would be about $100,000 (16 Linux GNBD servers, each with a FC HBA, at $2500 each and one 32-port FC switch at $60,000). The Red Hat GNBD/GFS approach saves $400,000 over the pure FC SAN approach. Add the extra savings of an open source Linux operating system and commodity servers and the financial advantages of the Red Hat approach are overwhelming.



Red Hat Enterprise Linux storage implementation scenarios

With Red Hat Enterprise Linux, organizations have several options. As upcoming enhancements are rolled out, those options will continue to increase. The following table presents four common enterprise storage scenarios using Red Hat Enterprise Linux.

Scenerio Description	Components	Advantages	Drawbacks
NFS to NAS Connect multiple servers to a NAS array to access files	• Red Hat Enterprise Linux running NFS • NAS array • Ethernet/IP link • NIC	• Industry-standard NAS configuration	• Not POSIX compliant • Slow • Limited file sharing
ISCSI to SAN Connect multiple servers via IP to the FC SAN to access block-level data	• Red Hat Enterprise Linux running GFS • iSCSI Linux driver (available with Red Hat Enterprise Linux in late 2004 or early 2005) • NIC • Fibre Channel (FC) SAN • Switch with FC and iSCSI ports	• Low cost (no server HBA) • POSIX-compliant	• Medium performance
Servers to FC SAN Connect multiple Linux servers to a SAN via FC	• Red Hat Enterprise Linux • FC HBA for each server • 1 multi-port FC switch	• Industry-standard SAN configuration • Easy to deploy	• Higher performance for FC-attached servers • No Data sharing between servers unless used with GFS • Costly (need FC HBA on each server)
Servers to converged SAN/NAS Multiple servers share file and block data on NAS and SAN	• Red Hat Enterprise Linux running GFS • iSCS/GNBD • FC HBA and GNBD for a subset of servers • Switch with FC and iSCSI ports	• Low cost (fewer FC HBAs) • POSIX-compliant	• More complex configuration

Note: All the above scenarios can take advantage of the other storage management capabilities provided by Red Hat Enterprise Linux, such as LVM, Devlabel, etc.

Conclusion: Linux-Based Utility Computing

For large enterprises, scalability is the issue. Previously forced into deploying costly, proprietary, vertically scaled servers, enterprises can now deploy large sets of open source, commodity servers in a horizontal scalability strategy and achieve the same levels of processing power for far less cost. And, they gain flexibility that is not possible with a vertical approach.

Such horizontal scalability can lead an organization toward utility computing, where server and storage resources are added as needed, on the fly. With Red Hat Enterprise Linux, managers can achieve substantial server and storage flexibility—the ability to add and remove servers and storage and to redirect and reallocate storage resources dynamically. This brings them to the brink of on-demand utility computing but without the considerable price associated with today' s utility computing offerings.

Even before organizations go that far, an enterprise storage strategy involving Red Hat Enterprise Linux will fit into their existing NAS and SAN storage environments. There, it will deliver immediate benefits in terms of storage consolidation, efficient management, low cost, and flexibility.