



---

## An Overview of Red Hat Advanced Server V2.1 Reliability, Availability, Scalability and Manageability (RASM) Features

### **Abstract**

This white paper provides a technical overview of Red Hat Linux Advanced Server V2.1 RASM features. Intended for systems administrators, DBAs and IT management, the paper outlines key distinguishing features in Advanced Server which deliver on RASM requirements for enterprise-class IT operations. It is suitable for people who have a technical understanding of operating system software platforms.

## Table of Contents

<b>Introduction</b> .....	<b>3</b>
<b>Reliability</b> .....	<b>4</b>
<b>Availability</b> .....	<b>5</b>
Cluster Manager Operation Overview.....	6
Cluster Manager Application Support.....	6
<b>Scalability</b> .....	<b>8</b>
Asynchronous I/O Support.....	8
I/O Spinlock Contention Reduction.....	9
Improved Process Scheduler.....	10
Bounce Buffer Elimination.....	12
<b>Manageability</b> .....	<b>14</b>
Network Console.....	14
Netcrashdump.....	14
Red Hat Network.....	14
<b>Summary</b> .....	<b>16</b>

## Introduction

Red Hat Linux Advanced Server V2.1, released in May 2002, provides enterprise class features that enable Linux-based solutions to be deployed across the widest range of enterprise IT environments. Working in collaboration with leading industry software vendors including Oracle and VERITAS, Red Hat engineers enhanced the capabilities of the Linux operating system in many areas. This paper describes Red Hat Linux Advanced Server's features that are critical for the mission critical enterprise application and database deployments on which companies rely for their operations and business continuity - Linux system RASM.

The terms **Reliability, Availability, Scalability** and **Manageability** are widely used when describing the requirements for deploying an enterprise system. These systems must meet RASM targets that are more stringent than those required by off-the-shelf consumer systems. The initial investments in hardware, storage, management and applications are considerable. The design and deployment phases are more complex, the life of the system is longer, and maintenance and upgrade costs are significant.

Linux has a proven track record in general purpose and mid-range server workloads. Customers are now looking to deploy Linux in more complex, mission critical environments. To successfully move Linux into more critical IT environments it is necessary to extend the capabilities of the operating system in all four RASM dimensions, to meet or exceed RASM capabilities found in proprietary enterprise platforms. Red Hat Linux Advanced Server is the Linux platform which delivers on the RASM requirements of the enterprise market

While this paper has been organized to consider RASM features individually, it is important to recognize that, in production IT environments, there are no hard and fast boundaries between them. They are heavily interdependent on each other - a system that is extremely reliable but cannot be scaled to meet a customer's growth requirements would not, ultimately, meet the needs of the enterprise. Similarly, RASM capabilities are not purely technical in nature. The reliability of a system can be greatly impacted by the soundness of the vendor's qualification and testing processes, and the speed of its response to problems. Likewise, the manageability of a system can be impacted by the quality of training received by the system administrators from system software vendors or their partners.

The following sections of this paper will detail Red Hat Linux Advanced Server RASM features. In combination with Red Hat's global support infrastructure, unmatched technical capabilities, and expanding Independent Software Vendors (ISVs) support and certifications, Red Hat Linux Advanced Server is poised to meet the needs of the most demanding enterprise environments.

## Reliability

Creating a reliable system goes far beyond classic software troubleshooting.- it requires an understanding of all aspects of the customer's expectations for an IT solution over its full life-cycle. Understanding this, Red Hat has taken a holistic approach to ensuring that Advanced Server delivered the highest levels of overall reliability. All aspects of system reliability were considered:

- To ensure that **Advanced Server** software components are of the highest quality, Red Hat Engineering subjected it to the most stringent qualification and testing over an extended period of time. Also, extensive stress testing and quality assurance was performed using major enterprise applications and large system configurations. Through extensive, longer testing schedules and focus on those operating system features that are most critical to enterprise applications, **Advanced Server** has become the most reliable Red Hat platform for commercial environments. This work set the foundation for the reliability of the overall system.
- Red Hat has worked with major application vendors, including Oracle, IBM, SAP, VERITAS, Computer Associates and others, to ensure that their products work seamlessly with **Advanced Server**. It is expected that major application vendors will take advantage of **Advanced Server** stability by selecting it as the only Red Hat Linux platform on which they will qualify their products. This single certification platform for application vendors will simplify customer deployments and system administration. It will also provide a level of consistency for customers that the more frequent release cycles of consumer Linux operating systems cannot provide.
- Acting on feedback from both enterprise customers and key enterprise system vendors and ISV partners, Red Hat has extended the life cycle of Advanced Server releases, distinguishing it from traditional consumer “release early, release often” Red Hat Linux platforms. New consumer versions are typically released approximately every 6 months. This imposes an excessive testing, qualification and migration burden on enterprise customers, who value stability and planning time to execute version upgrades. Consequently, new versions of **Advanced Server** will be released on a 12-18 month schedule. This release cycle will ensure the timely integration of key enterprise-class technology developed by Red Hat and other members of the open source community, while mitigating the risk to enterprise customers of adopting bleeding-edge innovations that characterize consumer releases.
- Red Hat has made a commitment to ISV partners that the **Advanced Server** product family will provide upward compatibility of system APIs. This means that applications certified on one release will typically not require any modifications or testing cycles before being certified on new releases (unless the application is enhanced to take advantage of new APIs). This will greatly increase application stability over the long term.

- Savvy enterprise IT managers recognize that deploying a solution without access to top quality professional services and technical support inherently reduces a system's reliability. To address the necessity of highly qualified support and maintenance, **Advanced Server** is only available for deployment as a support and maintenance subscription. In contrast, consumer-focused Red Hat Linux products can be purchased without on-going support and maintenance. **Advanced Server** is provided by Red Hat on an annual subscription basis with remedial and errata (patch) services included. There are three variants of **Advanced Server**, each providing a different level of service, with coverage ranging from basic installation/configuration support, through to enterprise-class service with 24x7 coverage and 1 hour response. All variants include access to Red Hat Network for system maintenance and management services, Red Hats Internet-based systems management facility. With Red Hat Network, systems are maintained with the latest tested, certified patches, ensuring maximum reliability and security.

Multi-disciplinary Red Hat teams have worked together with application partners throughout the **Advanced Server** development process to ensure that the final customer solution achieves the highest levels of reliability

## Availability

From a user's viewpoint, the term *availability* is easily defined the users applications and data should be continuously available. To achieve this goal requires a mixture of hardware and software. Sufficient hardware resources are required to ensure that the failure of any single hardware component does not compromise the entire system (such systems are often referred to as having *no single point of failure*). The software must be able to control the hardware components so that the users and their applications are minimally impacted by dynamic changes in the configuration.

**Advanced Server** provides two integrated high availability technologies. The first, IP Load Balancing (Piranha), provides network load balancing for environments such as web server farms. IP Load Balancing will distribute incoming network requests across a group of servers, which then service the request. Load balancing yields improved performance and, if a single server fails, incoming requests will continue to be redistributed across the remaining servers. From the remote users viewpoint, the network requests can be serviced continuously, even in the face of a server failure. The IP Load Balancing technology in **Advanced Server** has been available for more than two years in previous Red Hat OS releases.

The second high availability technology, Red Hat Cluster Manager, is being

introduced with **Advanced Server**. Cluster Manager provides high availability by using a technology widely used by other operating systems - application *failover*. Application failover is used in products such as Microsoft Windows 2000 Advanced Server, Sun Cluster and Compaq TruCluster. Cluster Manager has been specifically developed for use with today's commodity hardware products; it does not require expensive, special-purpose hardware components.

## Cluster Manager Operation Overview

The simplest Cluster Manager configuration comprises a pair of servers and an external SCSI or Fibre Channel storage array. Both servers are connected to the external storage array and access its disks directly. The Cluster Manager software is used to control access to storage partitions, so that only one server can access a particular partition at a time. This is required because standard applications do not support concurrent access to their data files from multiple systems. Each server will then operate in the same manner as if it were a single standalone system, running applications and accessing data on its allocated storage partitions.

In addition to their connections to the shared storage array, the two servers (or *nodes*) are also connected to each other using a network or serial interface so that they can communicate with each other. In the event that one of the servers shuts down or fails, the other server will detect the event and will automatically start, or failover, the applications that were previously running on the failed server. Because both servers are connected to the external shared storage, the operational server can access the failed server's disk partitions and its applications can continue to operate normally. If necessary, the remaining server will also take over the IP address of the failed server, so that network operations can continue without interruption.

## Cluster Manager Application Support

Cluster Manager is suitable for a wide range of applications, including:

- **Generic, unmodified applications.** Most custom, in-house applications can be used in Cluster Manager environments. This applies to any application that can tolerate a few seconds of downtime.
- **Databases.** Cluster Manager is the ideal way to deliver highly available databases, including Oracle 8i/9i, DB2, MySQL and PostgreSQL.
- **File Serving.** Cluster Manager brings high availability to file serving environments such as NFS V2 and V3, and SMB/CIFS (using Samba).
- **Enterprise Commercial Applications.** Cluster Manager can be used with applications such as SAP, Oracle Application Server and Tuxedo.
- **Internet, and Open Source applications.** Cluster Manager fully supports the

most popular Internet and Open Source applications (e.g. Apache).

- **Messaging.** Cluster Manager can be used with leading messaging applications such as Sendmail and Lotus Domino.

Figure 1 shows a typical Cluster Manager configuration. Note that the crucial technical feature of a Cluster Manager cluster is that the storage is shared, allowing any server to host any application and directly access its data. For additional information on Cluster Manager please refer to other white papers at <http://www.redhat.com>.

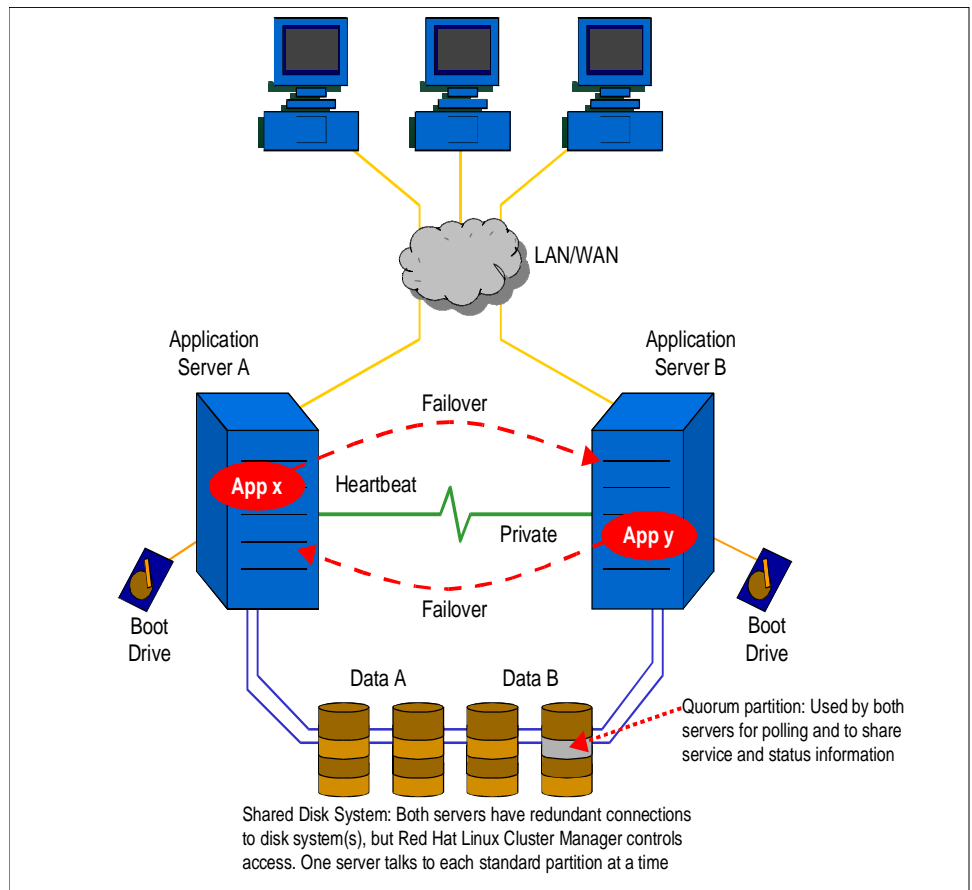


Figure 1 – General Layout of a Cluster Manager Configuration

## Scalability

The scalability of a system refers to its ability to provide linear, predictable support for ever-larger systems. While the Red Hat Linux 7.x series operating system provides excellent support for entry-level, general purpose servers and appliances, the potential performance of larger systems cannot be fully realized with the standard Linux kernel. **Advanced Server** is optimized for Intel SMP systems with more than a single processor (2 to 8-way) systems. Red Hat has provided several significant kernel enhancements in **Advanced Server** which ensure that applications can scale linearly to maximize CPU throughput and processing power. These kernel enhancements are described in the following sections.

## Asynchronous I/O Support

In a standard Linux environment, all I/O read operations issued by a single process are *synchronous*; that is, once an application has issued a read I/O it will stall until the I/O is complete. During the stall, other processes may run, and issue I/Os, but they too will be synchronous. For smaller, less complex applications, the I/O stall is often acceptable because no further processing is feasible until the results of the I/O are complete (for example, further computation is dependent on the I/O data). But for large applications that service many clients, such as a database or messaging server, the ability to issue multiple I/Os and continue processing while awaiting completion is critical to providing rapid response times and high throughput. Bear in mind that a typical disk I/O will take on the order of a few milliseconds, enough time for a modern CPU to perform several hundred thousand instructions.

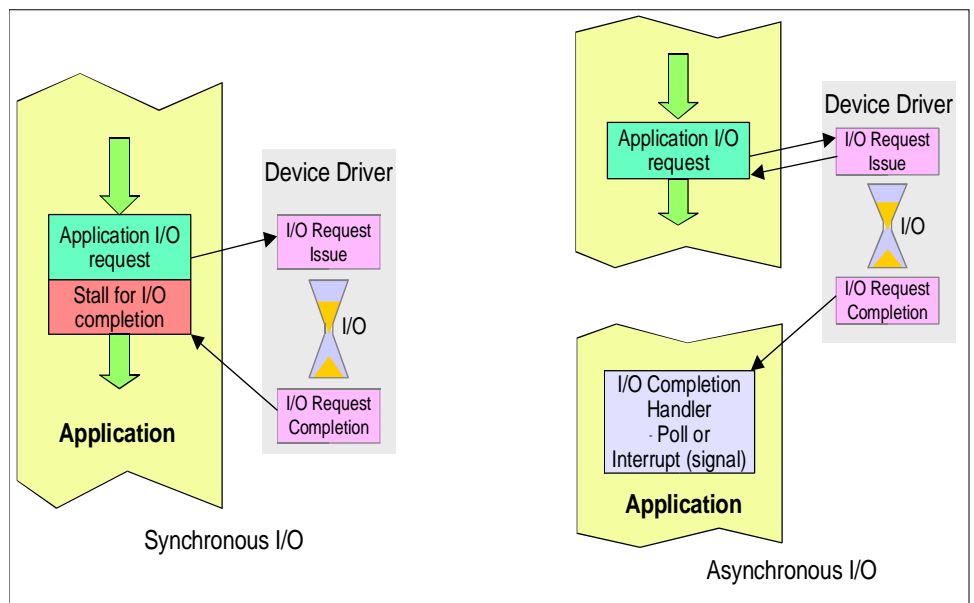
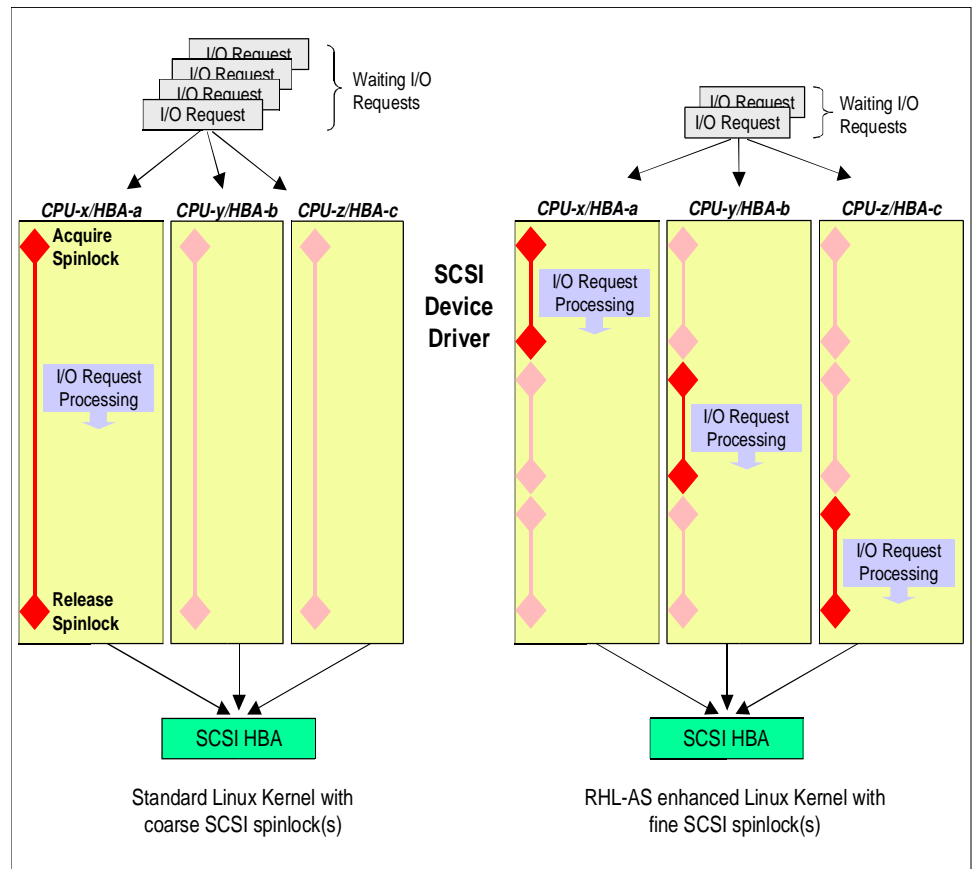


Figure 2 - Synchronous and Asynchronous I/O code flow

Asynchronous I/O support in **Advanced Server** allows a process to issue a read I/O and immediately continue processing. The application is notified of I/O completion by (1) a process-level software interrupt (a *signal*) and/or (2) by regular polling of an *event flag* (the kernel will set the event flag when the I/O is complete, which the application will subsequently detect). Enhancements to the I/O API allow an application to specify signal and event flag parameters. Note that an application must be modified before it can take advantage of this feature. The diagrams in Figure 2 trace an application program flow with synchronous and asynchronous I/O support.

## I/O Spinlock Contention Reduction

The Linux kernel uses *spinlocks* to coordinate access to critical kernel code in SMP systems. A spinlock is used, for example, to ensure that only one CPU at a time can execute the memory allocation routine. The Linux 2.4 kernel greatly improved the granularity of kernel spinlocks to ensure that access to different sections of the kernel was controlled by different spinlocks. The result is that different CPUs can simultaneously handle different kernel functions, increasing the kernel performance of the system. In **Advanced Server**, Red Hat engineers have extended this capability to the SCSI I/O subsystem to improve the I/O performance of multiprocessor systems. In the standard Linux kernel, the SCSI I/O spinlock granularity is *coarse* - each lock controls access to a large section of the driver and its data structures - resulting in limited opportunity for parallel SMP execution within the driver. With **Advanced Server** enhancements, the spinlock granularity is *fine*, allowing multiple I/Os to be processed by the device driver in parallel in SMP systems with multiple SCSI host bus adapters (HBAs). Specifically, the SCSI mid-layer spinlock was replaced by a per controller spinlock for a range of common SCSI adapter drivers (as of May 2002, these included the Adaptec AIC-7xx, Adaptec AAC RAID, QLogic QLA2200 & QLA2300). This results in increased I/O performance, especially for I/O intensive applications.



**Figure 3 - Coarse and Fine Granularity SCSI spinlocks**

he diagrams in Figure 3 provide a simplistic view of a device driver's spinlock usage with and without the **Advanced Server** enhancements. Note that the actual number of spinlocks required will vary depending on the I/O function, and that the overall level of parallelism achievable will depend on the number of SMP CPUs and SCSI HBAs available. Also note that no application modifications are necessary for a system to benefit from this enhancement.

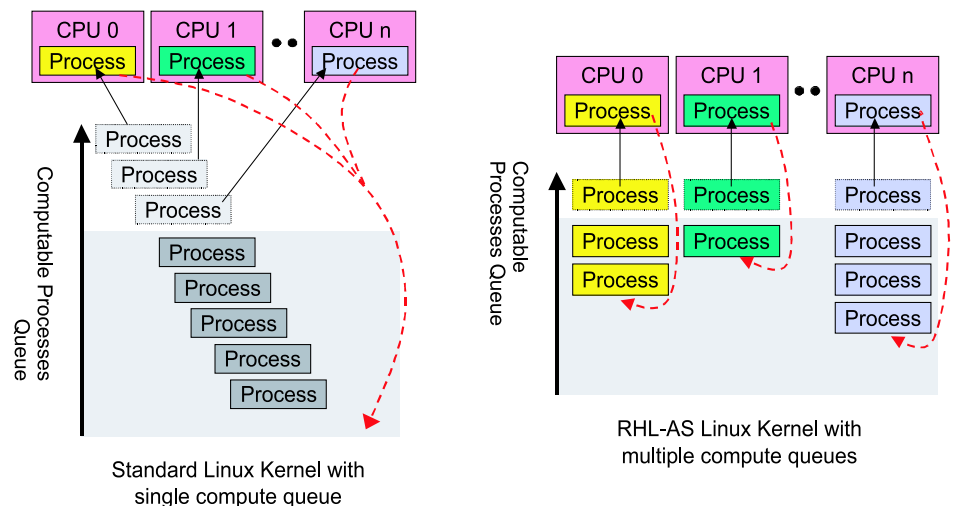
## Improved Process Scheduler

**Advanced Server** also provides an enhancement to the Linux process scheduler that has been explicitly designed for large-scale enterprise systems with multiple CPUs. In the standard Linux kernel there is a single scheduler compute queue that is, all processes that are computable are linked onto the same queue, waiting for the scheduler to place them into execution on a CPU. The scheduler will take processes from the head of the queue and place them on the first available CPU. A single process will often execute on a different CPU each time it is scheduled. With today's modern CPUs which generally contain very large Level 1 and Level 2 caches, this proves to be inefficient - data cached in one CPU is of no value if the process subsequently gets scheduled on another CPU. Also, scheduler spinlock contention - gating parallel access to the

scheduler code and the single compute queue - can become excessive in heavily loaded systems.

**Advanced Server** scheduler enhancements work to eliminate these inefficiencies by scheduling a process on the same CPU whenever possible. This means that data and instructions held in the CPU cache will continue to be available for the process each time it executes - greatly improving performance. The enhanced scheduler maintains a compute queue for each CPU in the SMP system. Once a process has been scheduled on a CPU it has an *affinity* for that CPU, and it will be rescheduled there until forced onto another CPU. An additional benefit of performing scheduling activities on a per-CPU basis is that the scheduler spinlock contention that occurs in the single compute queue implementation is eliminated, further improving performance. Process affinity is ignored only when appropriate - for example, when the process' CPU is executing a different process and another CPU is idle, or when there is a large disparity in the length of the compute queues. As an indication of the impact of this enhancement, Red Hat test labs have observed approximately one million context switches per second (CX/sec) on 2-way SMP system (over 400% faster than previously) and over 6M CX/sec on an 8-way system (6000% faster than before).

Process affinity is valuable in large systems with multiple CPUs and many active processes. In systems where CPUs are regularly idle, process affinity is not explicitly maintained, so the performance advantage is lost. The diagrams in Figure 4 show the two methods of scheduler operation. Note that no application modifications are necessary for applications to benefit from process affinity

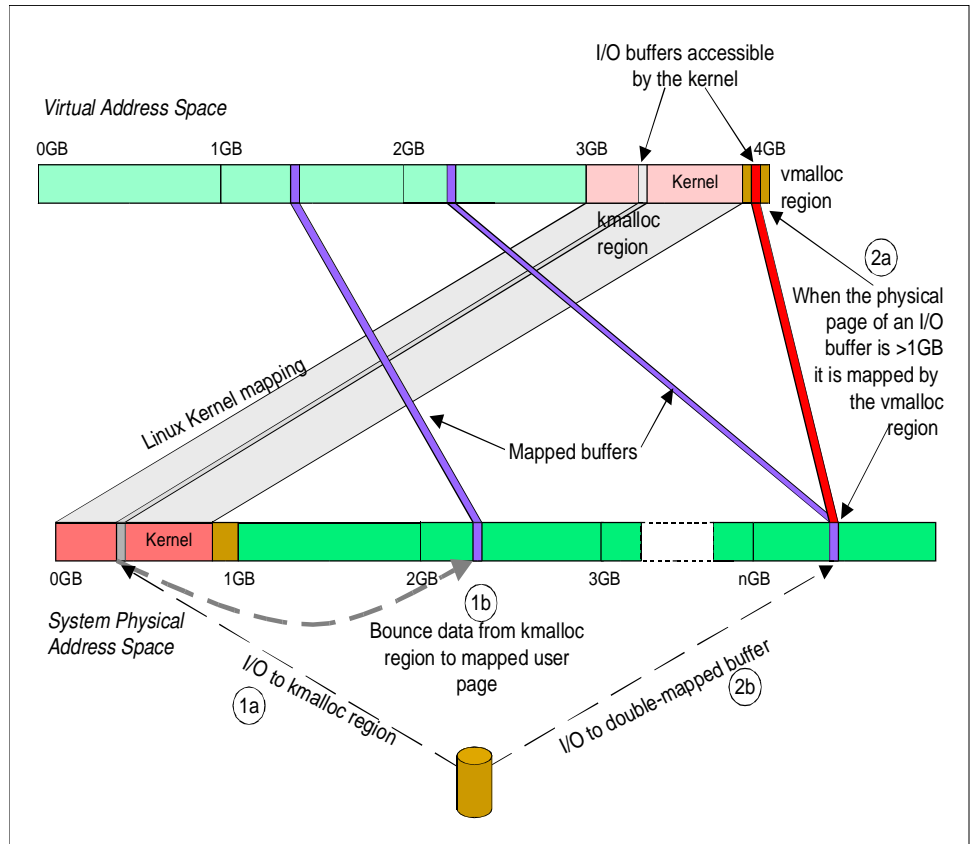


**Figure 4 - Single and Multiple Scheduler Compute Queues**

## Bounce Buffer Elimination

Bounce buffer elimination is another area where **Advanced Server** has been specifically enhanced for large systems (over 1GB main memory). Depending on the specific I/O operation, the enhancement will generally drastically reduce or eliminate buffer copying while allowing device drivers to access the contents of an I/O buffer regardless of its location in memory (a common device driver requirement). The requirement is easily met for I/Os that occur in the 0-1GB region of main memory, because the kernel virtual address space, itself 1GB, permanently maps it. However, the kernel is not able to map the entire memory of systems with over a gigabyte of memory, so if an I/O transfer is performed to physical addresses over 1GB, the device driver will be unable to access the buffer. To overcome this problem, the standard Linux kernel will perform all I/Os into interim buffers that are located in the first gigabyte of memory (15/16GB to be exact), and later *bounce*, or copy, the data to its final buffer above 1GB. Copying data buffers is inefficient, and impacts system performance.

To eliminate buffer bouncing, the **Advanced Server** kernel uses a section of the kernel virtual address space (called the *vmalloc* region, 1/16 GB in size) whose virtual pages can be used to map any physical address. Enough *vmalloc* pages to contain the I/O buffer are mapped directly to the buffer in high memory. The physical buffer, which is already mapped by the I/O requestor in user address space, is now double mapped. Using *vmalloc* addresses the device driver can now directly access the I/O buffer, regardless of where it is located in physical memory, and no bounce operation is necessary. This double mapping feature is extremely efficient and provides considerable I/O performance improvement. Applications do not have to be modified to benefit from non-bounced I/Os.



**Figure 5 - Bounce I/O vs. non-bounce I/O**

Figure 5 shows two I/Os in a large memory system to two user buffers mapped into high memory. I/O #1 is to a kmalloc buffer and then bounced to the user buffer. I/O #2 is directly to the I/O requestor's buffer above 1GB, and the buffer has been double mapped in the kernel's vmalloc space, eliminating the bounce operation. Note that for clarity the picture is somewhat simplified.

## Manageability

The manageability of a system can have a large effect on its cost effectiveness and long term TCO. Complex, error-prone or time-consuming administration tools can lead to extended downtimes and dissatisfied users. Consequently, all Red Hat Linux products provide a comprehensive suite of management features, leveraging the very strong Linux command line environment and the many GUI-based administration utilities for use with GNOME and KDE.

For the initial release of **Advanced Server**, attention has been paid to enhancing the ability to monitor systems and to resolve operating system kernel-level problems. These features are essential for delivering enterprise-level system support services and are not well developed in standard consumer Linux products.

## Network Console

The Network Console feature allows all kernel messages to be logged by another system on the network. The remote system runs a *syslog* server, which is able to collect the messages - including kernel crash signature messages - from a group of systems. This feature can be used to centralize and simplify the handling of message logs from multiple servers, even those located an extended distance from the syslog server. Importantly, it can be used to ensure that information critical to the identification and resolution of system crashes is not lost. In many cases the messages logged by the Network Console feature will contain enough information to resolve a problem after its first occurrence.

## Netcrashdump

The Netcrashdump feature allows an **Advanced Server** system to transmit a complete crash dump across the network to a sink node. This differs from traditional crash dumping techniques, which usually write dumps to swap partitions or special files on the failing system. In a Linux environment, where the range of potential disk controllers is very large, providing a reliable disk-based crash dumping facility is impractical. Dumping across the network, on the other hand, is straightforward, and has the added benefit that crash analysis is performed on a different system. This means that the failed system can be quickly recovered to its usual operation. A single sink node can be used to accept crash dumps from many servers, simplifying the management and archiving of crash dumps.

## Red Hat Network

Red Hat Network is Red Hat's internet-based systems management infrastructure. It provides automated notification and errata (patch) updates for **Advanced Server** systems. While onsite, phone and email support provide the timely resolution that **Advanced Server** customers require for their specific

issues, Red Hat Network is used for routine system maintenance activities. Red Hat Networks automated services greatly increase the productivity of system administration staff, by offloading tedious repetitive maintenance activities and ensuring that they are aware of the latest security and Severity 1 bugs. Red Hat Network has been fully integrated into the **Advanced Server** product.

## Summary

**Red Hat Linux Advanced Server** has been specifically designed for deployment in mission critical, enterprise environments. Providing support for large systems - with multi-gigabyte memories and multiple CPUs - **Red Hat Linux Advanced Server** enables Linux solutions to be successfully deployed further up the corporate IT infrastructure than ever before. And the high availability clustering capability enables customers to achieve seamless business continuity for their IT operations.

To complement **Advanced Server's** technical features, the product includes an annual Red Hat service subscription. Three service variants are available, each delivering a different level of support - from a basic service that provides support for Installation and Configuration to a fully-featured service level agreement (SLA) with 24x7 coverage and 1 hour response.

Throughout the product's development, careful attention has been paid to the complete **Advanced Server** environment beyond the merely technical - such as its support services, its certification processes, and its commitment to providing stable application interfaces.

With **Red Hat Linux Advanced Server**, customers will benefit from the most advanced **Reliability, Availability, Scalability** and **Manageability** capabilities ever provided in a Linux solution.

For additional information please refer to <http://www.redhat.com> or call 1-888-2REDHAT