Linux Kernel 2.6 Features in Red Hat Enterprise Linux

Author
Nick Carr
Enterprise OS Marketing Manager

Abstract
This technical brief demonstrates that the Linux 2.6 kernel features that are most important for commercial and enterprise deployments have been backported and included in the Red Hat Enterprise Linux version 3 product family.

Table of Contents

Overview

The Red Hat Enterprise Linux 3 family of products has been developed for customers who are deploying Linux solutions in commercial and enterprise environments. It includes features and capabilities chosen to meet their requirements, including:

- An enhanced level of maturity, stability, reliability and integration of software features, beyond that provided by typical open source community projects.
- Extensive support for commercial applications by leading Independent Software Vendors (ISVs).
- Extensive support for hardware systems--servers, clients, storage & network systems--by leading Original Equipment Manufacturers (OEMs).
- Comprehensive training, maintenance and support services that cover the complete system, and will be available throughout the multi-year life of the deployment.
- Proven performance and scalability through the availability of published benchmark results from accepted industry auditing groups.
- Access to latest technology features that are appropriate for commercial customer applications.

With Red Hat Enterprise Linux, Red Hat provides a complete product solution that meets these requirements through a combination of leading open source engineering and test capabilities, close partnerships with major ISVs & OEMS, and comprehensive services delivered by Red Hat and its partners.

Kernel Technology Selection

Choosing the most appropriate kernel technology for Red Hat Enterprise Linux required us to include the latest features while maintaining stability and quality. The Linux kernel evolved rapidly, with exciting new features still appearing continuously.

To provide the best mix of technology and stability, Red Hat created the Red Hat Enterprise Linux 3 products using a hybrid kernel approach: Features from the 2.6 kernel that would be of most value to commercial/enterprise customers and ISVs were backported for use with the latest and most stable 2.4 kernel. As a result of this approach, Red Hat Enterprise Linux products are provided with kernels that include the latest 2.6 features and are also exceptionally stable. Note that many of the new 2.6 kernel features were developed by Red Hat engineers, so Red Hat is unique in having the technical capability to achieve this best-of-both-worlds combination.

Kernel Features in Red Hat Enterprise Linux 3

The table below outlines the most important new kernel features and notes which ones are included in Linux 2.6 and the Red Hat Enterprise Linux 3 product family. The following sections of this paper provide additional technical details on the most important of these features.

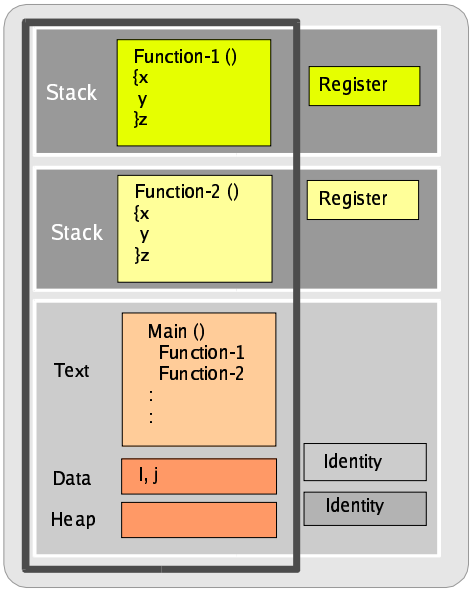
	In Linux 2.6 Kernel	In Enterprise Linux 3	Provides:
Native Posix Thread Library (NPTL)	Yes	Yes	High performance POSIX compliant multi-threading
Kernel IPSec	Yes	Yes	IPSec layer available for use by kernel modules
Asynchronous I/O (AIO)	Yes	Yes	Improved application performance
O(1) Scheduler	Yes	Yes	Highly scalable SMP scheduler
OProfile	Yes	Yes	CPU-hardware-based performance monitoring
kksymoops	Yes	Yes	Improved kernel bug reporting
Reverse Map Virtual Memory (rmap VM)	Yes	Yes	Performance improvement in memory constrained systems
HugeTLBFS	Yes	Yes	Performance improvement for large virtual memory applications (e.g. Databases)
Remap_file_pages	Yes	Yes	Kernel memory optimization for shared memory applications
2.6 Network stack features (IGMPv3, Ipv6, etc.)	Yes	Yes	Improved network performance & messaging
IPvs	Yes	Yes	Network load balancing
Access Control Lists (ACLs)	Yes	Yes	Improved file system security management
4GB-4GB memory split	No	Yes	Greatly increased x86 physical memory support and larger application address space
Scheduler support for hyperthreaded CPUs	No	Yes	Improved hyperthreaded CPU performance. (2.6 implementation not yet comparable)
Block I/O (BIO) block layer	Yes	No	Major rewrite of the I/O subsystem (stabilization and driver support in progress)
Support for > 2TB file system	Yes	No	Support for very large volumes. Red Hat Enterprise Linux 3 supports up to 1 TB.
New I/O elevators	Yes	No	Fine tuning for I/O subsystem performance (stabilization in progress)
XFS filesystem	Yes	No	High performance file system
Interactive scheduler response tuning	Yes	No	Scheduler improvements for interactive tasks (stabilization in progress)

Technical Capabilities in Linux 2.6 and Red Hat Enterprise Linux 3

This section describes some of the more important capabilities that have been newly included in the Linux 2.6 kernel and also backported for inclusion in the Red Hat Enterprise Linux product family.

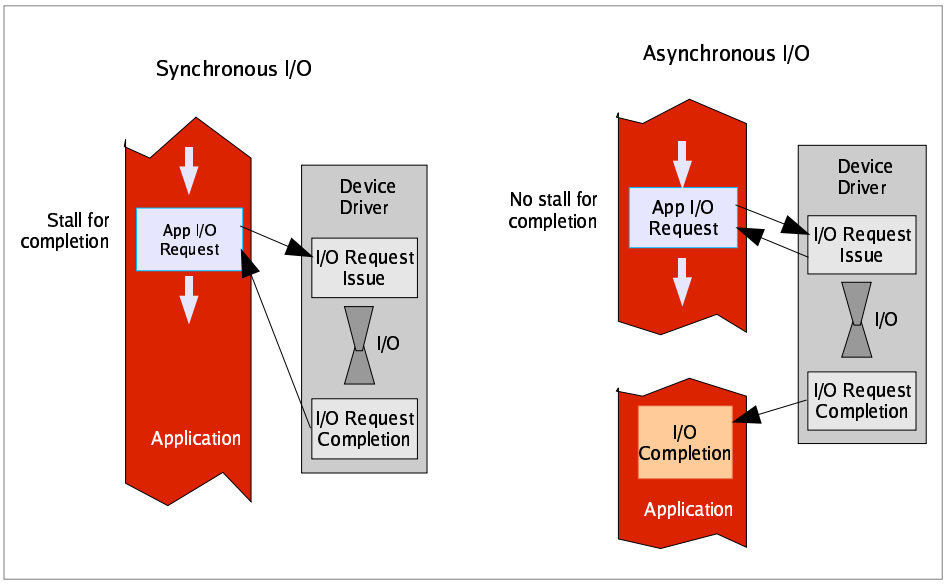
Native Posix Threading Library

This Linux 2.6 kernel feature, originally designed and implemented by Red Hat, provides excellent performance for multi-threaded applications (for example, Java applications). It enables multi-threaded applications that previously required the performance offered by proprietary Unix systems to be successfully deployed on Red Hat Enterprise Linux. The implementation provides full POSIX compliance, support for Thread Local Storage and Futex-based synchronization. Informal benchmarks have shown very high thread creation/deletion performance (in excess of 50,000/second).



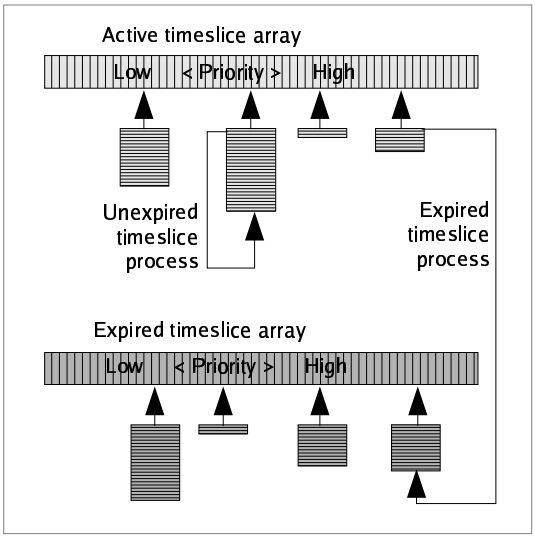
Asynchronous I/O

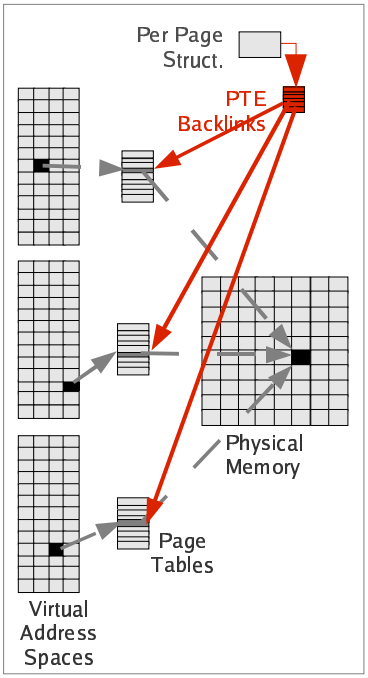
Asynchronous I/O support was originally included in Red Hat Enterprise Linux 2.1, almost two years before the Linux 2.6 kernel was released. This feature allows processes to continue running after issuing a disk read/write I/O. Previously, processes were required to wait for their disk I/O requests to complete before they could continue processing. The feature is particularly useful for processes that issue multiple writes in rapid succession, such as database writer processes. However it can be very useful for any multi-user application.



O(1) Scheduler

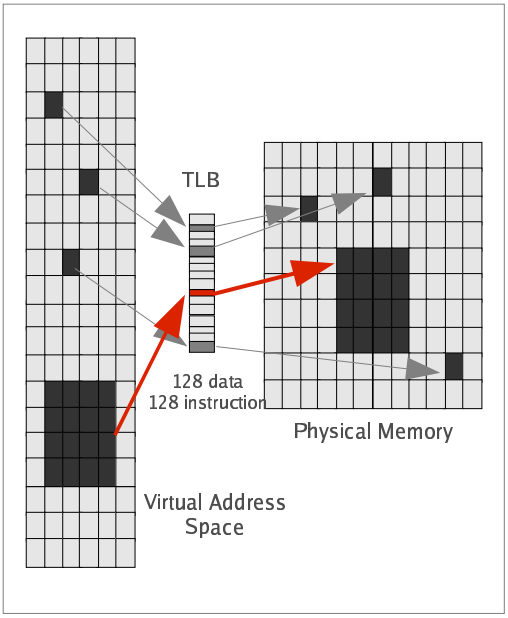
The O(1) scheduler provides greatly increased scheduling scalability. This increase has been achieved by a full redesign of the scheduler algorithms so that the time taken to choose a process for placing into execution is constant, regardless of the number of processes. The new scheduler scales very well, regardless of process count or processor count, and imposes a low overhead on the system. The algorithm uses two process priority arrays--active and expired. As processes are allocated a timeslice by the scheduler, based on their priority and prior blocking rate, they are placed in a list of processes for their priority in the active array. When they expire their timeslice they are allocated a new timeslice and placed on the expired array. When all processes in the active array have expired their timeslice, the two arrays are switched, restarting the algorithm. For general interactive processes (as opposed to real-time processes) this results in high-priority processes (which typically have long timeslices) getting more compute time than low-priority processes, but not to the point where they can starve the low-priority processes completely.





Rmap VM (Reverse Map Virtual Memory)

This feature, also developed by Red Hat engineers, improves the operation of the kernel's memory management subsystem, resulting in improved performance for memory constrained systems, NUMA systems, and systems with large aggregate virtual address spaces. Previously, the Linux memory management system could efficiently perform virtual to physical address translation using page tables. However, translating a physical address to its corresponding virtual address(es) was inefficient and time consuming, because it required the memory management software to traverse every page table in the system. In today's large systems this operation is unacceptably slow. With Rmap VM additional memory management structures have been created that enable a physical address to be back-translated to its associated virtual address(es) quickly and easily.

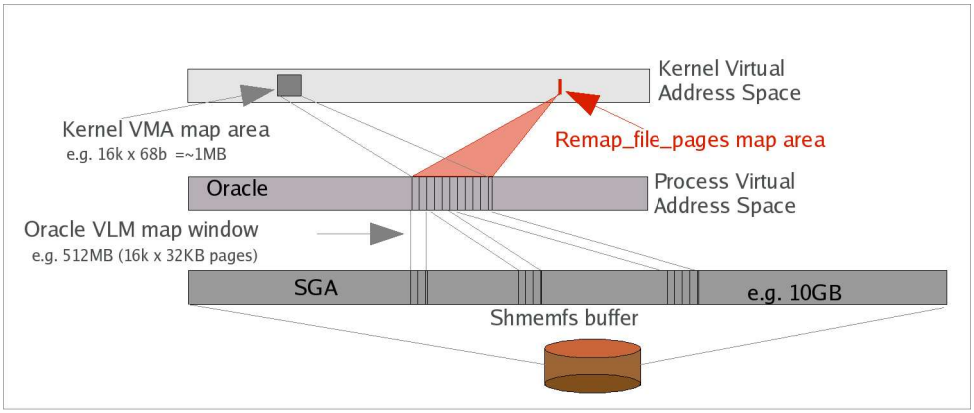


Huge TLBfs

This memory management feature is valuable for applications that use a large virtual address space. It is especially useful for database applications. The CPU's Translation Lookaside Buffer is a small cache used for storing virtual to physical mapping information. By using the TLB a translation can be performed without referencing the in-memory page table entry that maps the virtual address. However, to keep translations as fast as possible the TLB is typically quite small, so it is not uncommon for large memory applications to exceed the mapping capacity of the TLB. The HugeTLBfs feature permits an application to use a much larger page size than normal, so that a single TLB entry can map a correspondingly larger address space. A HugeTLB entry can vary in size, but, as an example, in an Itanium 2 system a huge page might be 1000 times larger than a normal page. This allows the TLB to map 1000 times the virtual address space of a normal process without incurring a TLB cache miss. For simplicity, this feature is exposed to applications by means of a file system interface.

Remap_file_pages

This is another memory management feature that is suitable for large memory and database applications. It is primarily useful for x86 systems that use the shmemfs (shared memory file system). A shmemfs memory segment requires kernel structures for control and mapping functions, and these structures can grow unacceptably large given a large enough segment and multiple sharers. For example, a 512MB segment requires approximately 1MB of kernel mapping structures per accessor. Large database applications that create hundreds of concurrent threads (one for each SQL query, for example) can quickly consume all available free kernel address space. The Remap_file_pages feature modifies the shmemfs management structures so that they are significantly smaller (less than 100 bytes). This permits much larger shmemfs segments and thousands of concurrent users to be supported.



Oprofile

Oprofile is an open source project that provides comprehensive code profiling capabilities. Oprofile uses hardware counters inside the CPU to accumulate its performance measurements. It can be used to profile kernel code, device drivers and modules, and application code, with minimal system overhead. Implemented as a profiler daemon and provided with a number of post-profiling tools, Oprofile can be used by programmers to help resolve issues such as loop unrolling, branch mispredictions, and poor cache utilization.

Networking

Red Hat has implemented most of the Linux 2.6 kernel network stack in its Enterprise Linux family of products. This includes features such as kernel level IPsec support, improved IPv6 support and kernel support for IGMP V2 and V3. The net effect of this work is that Red Hat Enterprise Linux 3 network performance and functionality is essentially the same as the Linux 2.6 kernel.

Access Control Lists

The addition of Access Control List (ACL) support to the Ext2/3 file systems greatly increases flexibility and security of file and file system management. Standard Linux file systems provide the classic r,w,x security triplets for all file system objects, which are then gathered under the UID/GID grouping mechanism. ACLs are ideal for use in situations where this standard security mechanism is inadequate. By using ACLs, file system objects and accessors can be defined and grouped with much greater flexibility--as examples, accessors can be members of multiple groups, and objects can provide different levels of access for different users and groups on an individual basis.

Kksymoops

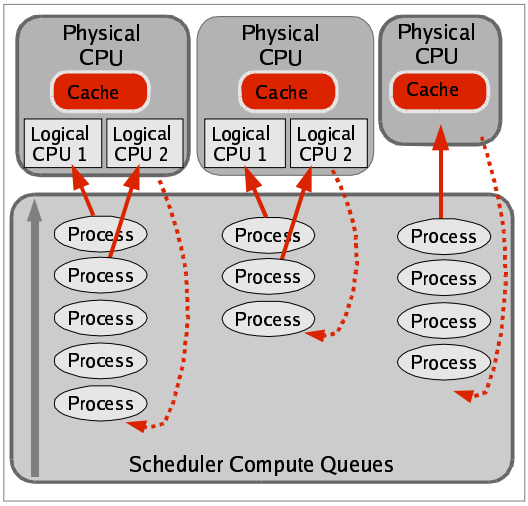
This feature will hopefully never be needed by a customer--it is useful to aid debugging and resolution of system crashes (oops). Without this feature a system crash is accompanied by a console display that is not annotated, making it extremely hard to decipher. Kksymoops provides automated annotation of console crash reports, greatly assisting problem resolution. For Red Hat Enterprise Linux products, which are fully supported by Red Hat and its partners, features such as Kksymoops are critical to delivering the level of service that enterprise/commercial customers require.

Technical Features Unique to Red Hat Enterprise Linux 3

The following section outlines a few of the features that are currently unique to the Red Hat Enterprise Linux product family. They were included due to their clear suitability for the enterprise/commercial market, and are expected to be delivered in upcoming Linux 2.6 kernel releases.

Hyperthreading Scheduler

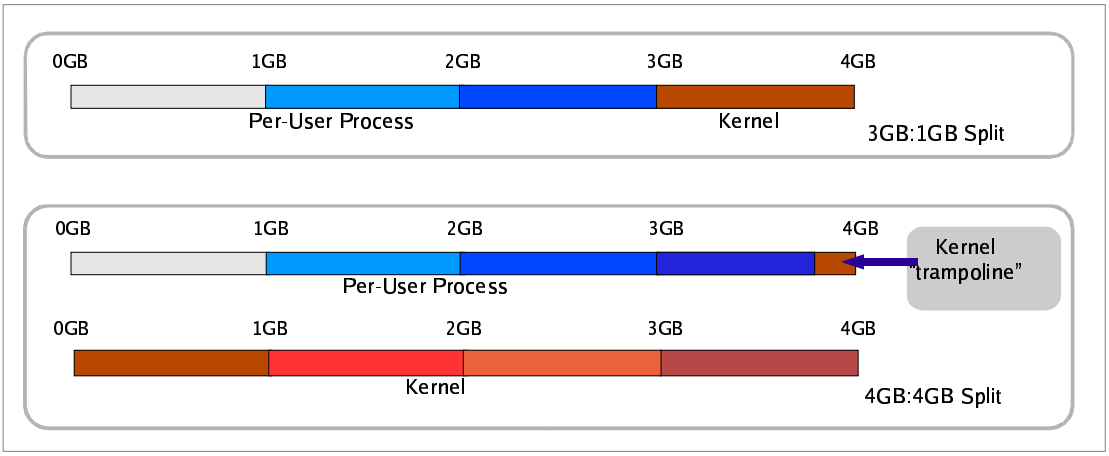
Scheduler support for hyperthreaded CPUs is another feature that was developed by Red Hat. Hyperthreading support ensures that the scheduler can distinguish between physical CPUs and logical (hyperthreaded) CPUs. Scheduler compute queues are implemented for each physical CPU, rather than each logical CPU (as was the case previously). This results in processes being evenly spread across physical CPUs, thereby maximizing utilization of resources such as CPU caches and instruction buffers.



4GB-4GB Split

Of the many new technologies included in Red Hat Enterprise Linux perhaps the most exciting is the enhanced handling of user process and kernel virtual address space in Intel x86-compatible systems. Traditionally, 32-bit x86 systems had a fixed 4GB virtual address space, which was allocated such that the kernel had 1GB and each user process 3GB (referred to as the 3-1 split). This allocation has been in use for many years, but has become restrictive as physical memory sizes have grown. With Red Hat Enterprise Linux version 3 it is possible to configure a 4-4 split, where each user process and the kernel are allocated 4GB of virtual address space. There are two important benefits to this new feature:

- The larger kernel virtual address space allows the system to manage more physical memory – up to 64GB of main memory is supported by Enterprise Linux 3 on x86-compatible systems.
- The larger user virtual address space allows applications to use approximately 30% more memory (3.7-3.8GB), improving performance for applications that take advantage of the feature. This feature is unique to Red Hat Enterprise Linux and means that x86-compatible systems can be expected to have a longer life-span and better performance.



Crash Dump

Red Hat Enterprise Linux has provided a network-based crash dump capability since it first shipped. While this feature is not included in the community-based Linux kernels, Red Hat believes that it is an essential capability for any supported operating system environment that customers rely on for their business-critical deployments. The feature enables systems to dump their memory contents to a remote “netdump” server in the event of a kernel crash. The crash can then be analyzed using Red Hat's enhanced analysis tools by Red Hat support services personnel.

Summary

This technical brief demonstrates that the Linux 2.6 kernel features that are most important for commercial and enterprise deployments have been backported and included in the Red Hat Enterprise Linux 3 product family. The overall value of these features is perhaps best reflected in the recent world record TPC/C benchmark result achieved using Red Hat Enterprise Linux. This benchmark demonstrated outstanding performance and price performance, audited at over 1,000,000tpmC and $5.52/tpmC. Also, the benchmark configuration, a 16 node cluster with over 90TB of online storage, clearly showed the scalability of Red Hat Enterprise Linux systems.

Significantly, the new 4GB-4GB memory split feature for large x86-based systems has been included in Red Hat Enterprise Linux 3 prior to its availability

in the Linux 2.6 kernel. As a result, Red Hat Enterprise Linux 3 supports more physical memory and larger applications than standard Linux 2.6 kernels.

Red Hat Enterprise Linux 3 provides the best technology combination for enterprise deployments – leading edge features with stability and maturity.