Linux 2.4

Changes and Enhancements

Acknowledgements

- Larry Woodman - Technical Director/Kernel Engineering, Mission Critical Linux
- Joe Pranevich
- Thomas Wolfgang Burger

Background

- Linux 0.95 in 1991
  - First release to public via Internet
- Linux 2.0
  - Starting to get public notice
- Linux 2.2
  - New file systems
  - Redesign of caching
  - Greater scalability
- Linux 2.4

Overview

- Linux kernel architecture features
- Linux kernel hardware support features
- File System Enhancements
- Networking Enhancements
- Device Support Enhancements

Architecture Features

- ELF and POSIX Foundation
  - More dependent on ELF
  - More POSIX compliant:
    - Clocks and Timers support

Architecture Features

- Memory Usage
  - About the same as 2.2
- Shared Memory
  - More compliant with Industry standards
  - Introduces a special “shared memory” filesystem

Architecture features

- 2.2 Page Replacement Problems
  - Page eviction
  - Simplistic NRU replacement
  - Clock algorithm can evict accessed pages
  - Sub-optimal reaction to variable load or load spikes after inactivity

Architecture features

- 2.4 Improvements:
  - Finer-grained SMP locking
  - Unification of buffer and page caches
  - Support for larger memory configurations
  - SYSV shared memory code replaced
  - Page aging reintroduced
  - Active & inactive page lists
  - Optimized page flushing
  - Controlled background page aging
  - Aggressive readahead

Architecture features

- SMP locking optimizations
  - Use of global “kernel_lock” was minimized.
  - More subsystem based spinlock are used.
  - More spinlocks embedded in data structures.
  - Semaphores used to serialize address space access.
  - More of a spinlock hierarchy established.
  - Spinlock granularity tradeoffs.

Architecture features

- Increased number CPUs supported
  - Static increase of maximum CPUs to 64.
  - Realistic scalability of up to 8 CPUs.
    - Bus saturation
    - SMP locking
  - Scheduler optimizations speed up selection of threads and context switching.

Architecture features

- Kernel multi-threading improvements
  - Multiple threads can access address space data structures simultaneously.
  - Single mem->msem semaphore was replaced with multiple reader/single writer semaphore.
  - Reader lock is now acquired for reading per address space data structures.
  - Exclusive write lock is acquired when altering per address space data structures.

Architecture features

- 32 bit UIDs and GIDs
  - Increase from 16 to 32 bit UIDs allow up to 4.2 billion users.
  - Increase from 16 to 32 bit GIDs allow up to 4.2 billion groups.

Architecture features

- 64 bit virtual address space
  - Architectural limit of the virtual address space was expanded to a full 64 bits.
  - IA64 currently implements 51 bits (16 disjoint 47 bit regions)
  - Alpha currently implements 43 bits (2 disjoint 42 bit regions)
  - S/390 currently implements 42 bits
  - Future Alpha is expanded to 48 bits (2 disjoint 47 bit regions)

Architecture features

- Unified file system cache
  - Single pagecache was unified from previous pagecache read/buffermem write functionality
  - Eliminates copying buffers from buffermem to pagecache on file read operations.
  - Reduces memory consumption by eliminating double buffered copies of file system data.
  - Eliminates overhead of searching two levels of data cache.

Architecture features

- Distributed Interrupts
  - Hardware interrupt service routines can be processed simultaneously on all CPUs.
  - Software interrupts (softIRQs) can be processed simultaneously on all CPUs.
  - SMP spin locks are maintained within device specific data structures.

Architecture features

- Increased number of threads and tasks
  - Default maximum number of tasks/address spaces was increased.
  - Default maximum number of threads per task was increased.
  - Configuration of both maximums was changed to be runtime tunable via /proc file system.
  - Scheduler optimizations minimize overhead of context switching between sibling threads.

Hardware Support Features

- IA64 Port and Architecture Optimizations
  - Support for IA64 processor features:
    - IA64 specific TLB optimizations.
    - Large rotating register file.
    - IA64 SMP specifics.
    - IA64 IO specifics.
  - 64 bit virtual address space.
    - Itanium is actually 51 bits; sixteen 47 bit regions.
  - NUMA support under development.

Hardware Support Features

- Alpha Architecture Optimizations
  - 64 bit virtual address space.
    - EV67 is 43 bits; half user, half kernel.
    - EV7 supports 48 bits; half user, half kernel.
  - 2TB(41 bit) physical address limit.
  - Highly accurate SMP compatible processor time optimizations.
  - NUMA support under development.

Hardware Support Features

- S/390 Architecture Optimizations
  - 64 bit virtual address space
    - 42 bits used - separate address spaces for users & kernel
  - 16EB physical address limit.
  - Highly accurate SMP compatible processor time optimizations.
  - NUMA support under development.

Hardware Support Features

- BIGMEM for IA32 (and other 32 bit systems)
  - 1GB physical memory limitation in the Linux kernel.
  - 4GB physical memory limitation for 32 bit systems.
  - 4GB physical memory optimizations in the Linux kernel.
  - 64GB physical memory using PAE on IA32.

Hardware Support Features

- Special instructions for some processors
  - Use of processor specific memory transfer instructions for:
    - Intel Pentium
    - AMD
    - Cyrix
    - WinChip

2.4 Kernel Hardware Support Features

- NUMA infrastructure
  - Machine independent Non-Uniform Memory Architecture (NUMA) infrastructure.
  - Support for:
    - multiple memory domains
    - processor subsets
    - binding of devices and interrupts to processors
  - Machine dependent NUMA portion under development for multiple architectures.

File System Enhancements

- File System size increase
  - File system data offset was increased from 31 bits to 44 bits in the VFS layer.
    - Increases file system size to 16TB.
    - Increases individual file size to 16TB.
    - Still need to consider file system overhead...
  - Several local file systems have been enhanced to take advantage of larger files.

File System Enhancements

- VFS layer redesign to use single cache
  - buffermem and pagecache functionality was unified in 2.4
  - VFS layer was changed to use pagecache for generic file read() and write() operations.
  - Eliminated coping between buffermem and pagecache.
  - Saves memory be eliminating multiple copies of buffered file system data.

File System Enhancements

- RawIO support to bypass file system cache
  - New RawIO interface was added to file systems.
  - This results in:
    - DMA directly to buffer wired in user address space.
    - Bypassing the pagecache.
    - Eliminates coping between pagecache pages and user buffer pages.
    - More efficient for databases.

File System Enhancements

- Several Journaling File Systems introduced
  - Pending file system updates are continually maintained in a single journal file.
  - The FSCK at reboot time is reduced to replaying the journal.
  - Speeds up reboot FSCK by several orders of magnitude.
  - ext3fs, reiserfs, xfs

File System Enhancements

- Inclusion of Logical Volume Manager into the Linux kernel
  - Allows file systems to span multiple disks.
  - Dynamic runtime resizing of file systems.
  - More flexible file system device management.
  - Standards compliant.
  - Familiar to users of commercial UNIX.

Networking Enhancements

- Network re-write for optimal performance
  - Redesigned to take advantage of improved multitasking and multithreading.
  - Improvement performance for simultaneous/multiple network interfaces.
  - Distributes networking load much more evenly on SMP systems.
  - Kernel uses wakeup_one to minimize wasted cycles

Kernel Networking Enhancements

- Firewall and IP functions placed in kernel
- Network subsystem split:
  - Packet filtering layer
  - Network Address Translation layer
- PPP code rewritten and modularized
- ISDN updated to support many new cards
- PLIP improved
- DECnet & ARCNet protocols supported
- Autodetection of Windows shares based on SMB
- Completely compatible to the letter of IPv4 spec

Networking Enhancements

- iptables/netfilter replacement for ipchains
  - Linux 2.2 replaced ipfwadm with ipchains.
  - Linux 2.4 replaced ipchains with iptables, also known as netfilter.
    - Includes capabilities to construct more sophisticated firewalls.
    - Can be used to implement NAT for supporting masqueraded private networks
    - Compatible with ipfwadm and ipchains command syntax.

Networking Enhancements

- Kernel based HTTP daemon
  - khttpd is a kernel daemon module which serves static web pages.
  - Can cooperate with Apache and other web servers to serve dynamic web pages.
  - Will result in significant web benchmarking improvement (SpecWeb, etc).

Networking Enhancements

- Fully compatible NFSv3 implementation
  - Fully compatible with version 3 of NFS distributed by Sun Microsystems.
  - Eases the burden of Linux sysadmins who maintain heterogeneous environments.
  - Also compatible with:
    - DECnet
    - ARCnet

Device Support

- 2.4 Supports:
  - Up to 10 IDE controllers
  - Up to 16 ethernet cards
  - Multiple AIPCs
  - SCSI TCQ (tagged command queuing)
  - RAID devices
  - ATM

Device Support

- Buses
  - Integrated into the new resource management subsystem
- Plug-N-Play
  - ISA & S/390 device configuration and detection
- USB
- I2O supported (PCI extension)
- PCMCIA support integrated

Device Support

- Framebuffers
  - New drivers and improvements to old
  - Support of many more “standard” VGA cards

Device Support

- Keyboards, Mice, Consoles, and Ports
  - USB support of keyboards and mice
  - Ability to redirect console output to parallel port
  - Serial support has same limitations as 2.2
  - Parallel port support has been overhauled
    - New generic driver
    - DMA support
  - IRDA support
  - Little work done on "WinModems"

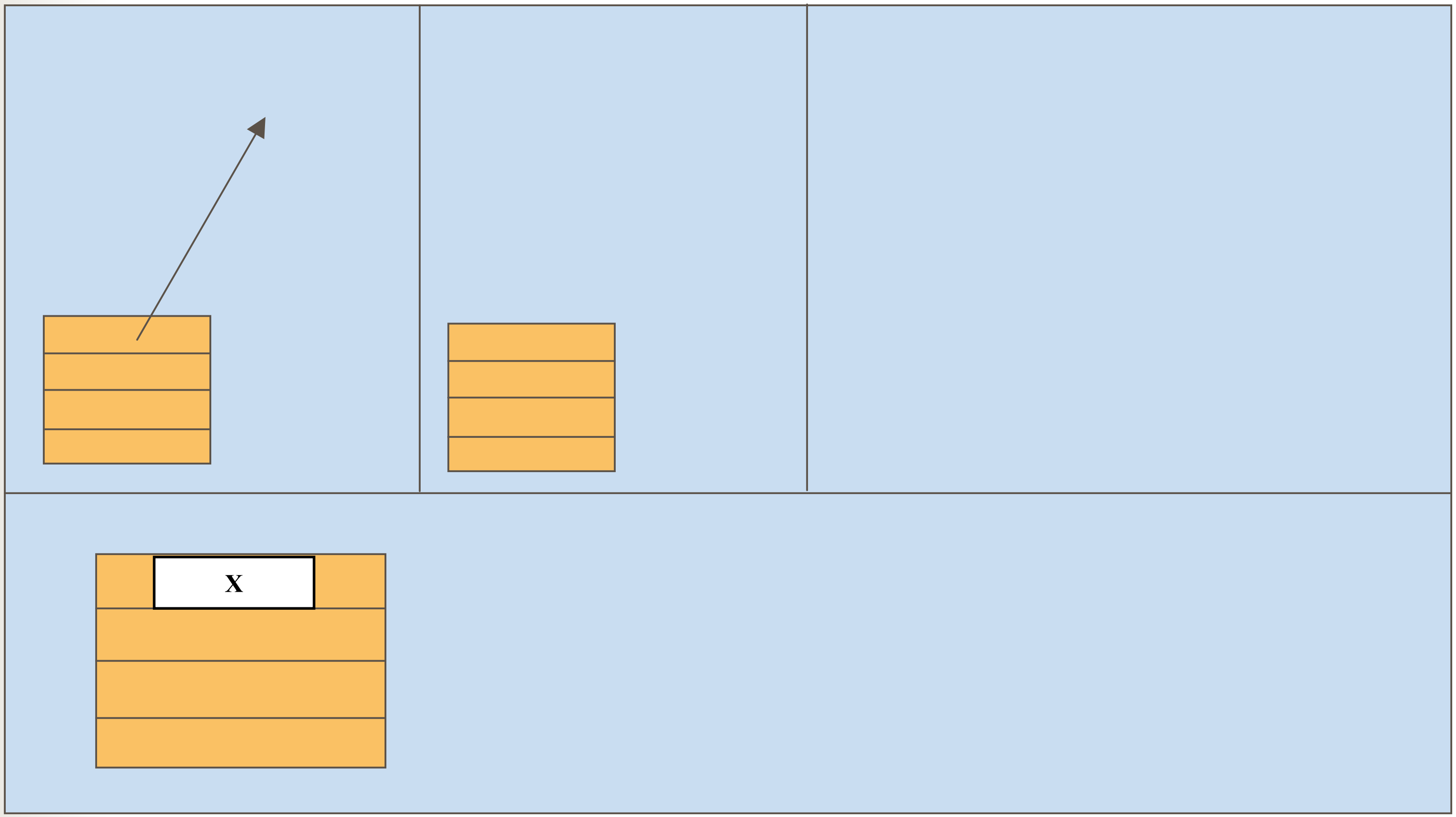
Device Support

- Accessibility
  - Support for speech synthesizer card
- Multimedia
  - No ground breaking changes
  - Updates and new drivers for variety of cards
    - Including full duplex support
  - Ease of configuration enhancements

Device Support

- S/390 Devices
  - 3270 as console and terminal
  - Tape support
  - Hipersockets
  - z/OS formatted disk (VTOC & DCBs)
  - PAGEX support (VM only)
  - Kernel in NSS (VM only)

PAGEX/PFAULT Support



PAGEX/PFAULT Support

- Eliminate overhead of double paging
- Page fault by Linux virtual machine usually puts it in wait state until VM gets page
- PAGEX/PFAULT handshaking allows VM to inform Linux of page request and have it dispatch another process
- When page operation is complete VM signals Linux again so it can mark task as dispatchable

PAGEX/PFAULT Support

- PAGEX
  - PROG 14 interrupt
  - 32-bit only
- PFAULT
  - External interrupt (x’2603’)
  - 32 & 64-bit systems
  - z/VM 4.2 required

What's still needed?

- Greater scalability above 8 processors
- NUMA
- Improved fiber-channel handling (requires an inappropriate amount of hand waving to work)
- >1TB per file system limit
- Poor I/O throughput on x86 class machines with very large amounts of memory
- Basic fail-over is there but not advanced clustering
- Logical volume manager needs more work