Linux Kernel 2.5

Not Your Grandmother's Kernel based on Dave Jones's post-halloween document

Muli Ben-Yehuda

mulix@mulix.org

IBM Haifa Research Labs

Regressions and Deprecations

- hptraid/promise RAID drivers are currently non functional and need updating.
- Some filesystems still need work (Intermezzo, UFS, HFS, HPFS..)
- A number of drivers don’t compile. They need work to convert them to the new APIs.
- khttpd (kernel web server) is gone - x15 can do it faster in user space.
- LVM1 has been removed, replaced with a better designed “device mapper”.

Modules

- The in-kernel module loader got re-implemented.
- You need replacement module utilities from http://www.kernel.org/pub/linux/kernel/people/rusty/
- Many modules are still broken, using the old API and semantics. They need converting to new way of doing things.
- New module semantics minimize races on module load and unload and simplify module handling code.

VM Changes

- The actual ’reverse mappings’ part of Rik van Riel’s rmap vm was merged. VM behaviour under certain loads should improve.
- rmap allows the VM to know, given a memory page, which processes are using it - more intelligent decisions on heavy load and swapout.
- VM tunables in /proc.
- The bdflush() syscall is now officially deprecated.
- Better behaviour under load, better behaviour (the kernel actually boots) on “enterprise” machines.

Kernel Preemption

- Users should notice much lower latencies especially in demanding multimedia applications.
- Code which is SMP safe should mostly be preempt safe. But, there are still cases where preemption must be temporarily disabled where we do not. These areas occur in places where per-CPU data is used.
- Several subsystems are not known to be preempt safe - be careful when enabling preempt.
- If you get “xxx exited with preempt count=n” messages in syslog, don’t panic, these are non fatal, but are somewhat unclean. (Something is taking a lock, and exiting without unlocking)

Scheduler Improvements

- Ingo Molnar reworked the process scheduler to use an O(1) algorithm.
- Users should notice no changes with low loads, and increased scalability with large numbers of processes, especially on large SMP systems.
- utilities for changing behaviour of the scheduler (binding processes to CPUs etc). http://tech9.net/rml/schedutils.
- sched_yield() and yield() can now make you sleep for a *long* time.
- 2.5 adds system calls for manipulating a task's processor affinity: sched_getaffinity() and sched_setaffinity()

Threading improvements

- Lots of work went into threading improvements. Some of the features of this work are:
  - Generic pid allocator (arbitrary number of PIDs with no slowdown, unified pidhash).
  - POSIX thread signals stuff (atomic signals, shared signals, etc.)
  - Threaded coredumping support
  - sys_exit() speedups (O(1) exit)
  - Generic, improved futexes, vcache
  - API changes for threading.
- Users should notice is a significant speedup in basic thread operations this is true even for old-threading userspace libraries such as LinuxThreads.

Kernel build system

- Quicker builds, less spontaneous rebuilds of files.
- make xconfig now requires the qt libraries.
- Make menuconfig/oldconfig has no user-visible changes other than speed, whilst numerous improvements have been made.
- make oldconfig *much* faster.
- ’help’ ’allyesconfig’ ’allnoconfig’ ’allmodconfig’.
- “make” does <arch-zimage> and modules.
- “make -jN” is now the preferred parallel-make execution.
- There is no need to run ’make dep’ at any stage

IO subsystem

- Considerable throughput improvements over 2.4 due to much reworking of the block and the memory management layers.
- Assorted changes throughout the block layer meant various block device drivers had a large scale cleanup whilst being updated to newer APIs.
- O_DIRECT improvements, size and alignment of data per device, not filesystem.
- Block devices can now access up to 16TB on 32-bit architectures, and up to 8EB on 64bit architectures.

POSIX ACLs and Extended attributes

- Finer grained access control lists to files, resources.
- Needs file system support, supported in 2.5 by EXT2, EXT3, others.
- Frequently requested feature from “enterprise” customers.
- Userspace tools available at http://acl.bestbits.at

ALSA

- Advanced Linux Sound Architecture. This offers considerably improved functionality over the older OSS drivers, but requires new userspace tools.
- Note that the OSS drivers are also still functional, and still present. Many features/fixes that went into 2.4 are still not applied to these drivers, and it’s still unclear if they will remain when 2.6/3.0 ships. The long term goal is to get everyone moved over to (the superior) ALSA.

System Calls

- Systems that support the SYSENTER extension (Basically Intel PPro and above, and AMD Athlons) now have a faster method of making the transition from userspace to kernelspace when a syscall is performed.
- Without an updated glibc, its unlikely that this will be noticable.
- sys_call_table is no longer exported to modules - modules can’t supply their own version of system calls.

IDE

- The IDE code rewrite was subject to much criticism in early 2.5.x, which put off a lot of people from testing. This work was then subsequently dropped, and reverted back to a 2.4.18 IDE status. Since then additional work has occured, but not to the extent of the first cleanup attempts.
- There are several Known problems with the current IDE code. IDE is biggest(?) show-stopper before 2.6.

SCSI

- Various SCSI drivers still need work, many don’t even compile currently.
- Various drivers currently lack error handling. These drivers will cause warnings during compilation due to missing abort: and reset: functions.
- Note, that some drivers have had these members removed, but still lack error handling.
- No major scsi infrastructure work done in 2.5.

Filesystems

- indexed directory support for EXT3.
- inode attributes (e.g. immutable) for reiserfs.
- Basic support has been added for NFSv4 (server and client)
- sysfs - a saner way for drivers to export their innards than /proc.
- IBM’s JFS, SGI’s XFS merged.
- HugeTLBfs - Files in this filesystem are backed by large pages if the CPU supports them.

OProfile

A system wide performance profiler has been included in 2.5. With this option compiled in, you'll get an oprofilefs filesystem which you can mount, that the userspace utilities talk to. The userspace utilities for this are very young, and still being developed. You can find out more at http://oprofile.sourceforge.net/oprofile-2.5.html

User Mode Linux

- UML is a port of the Linux kernel to a new architecture - Linux's own user space environment.
- or some kinds of kernel development (architecture independent, file systems, memory management), using UML is great.
- Also useful for web hosting, teaching students operating systems, etc.
- See http://user-mode-linux.sf.net for more information.

Kernel Hacking

- Linus now uses BitKeeper, a SCM system. The kernel is also available via CVS, subversion and other systems.
- rate of development has gone drastically up.
- kksymoops is merged - the kernel will now spit out automatically decoded oopses (no more feeding them to ksymoops). Always enable "Load all symbols for debugging/kksymoops".
- Various kernel sanity checks are now available. Look at your logs occasionally, and tell the kernel hackers if you find anything suspicious.

crypto

- A generic crypto API has been merged, offering support for various algorithms (HMAC,MD4,MD5,SHA-1,DES,Triple DES EDE, Blowfish)
- This functionality is currently only used by IPSec, but will later be extended to be used by other parts of the kernel. It’s possible that it will later also be available for use in userspace through a crypto device, possibly compatible with the OpenBSD crypto userspace.

New Ports

2.5 features support for several new architectures.

- x86-64 (AMD Hammer)
- ppc64
- UML (User mode Linux)
- uCLinux. 68k(w/o MMU) and v850.

Note that several in-tree ports are lagging behind their out-of-tree variants. Expected to improve as 2.6 draws nearer...

More Information

- Dave Jones's post-halloween document, aka 2.5 - what to expect. http://www.codemonkey.org.uk/post-halloween-2.5.txt
- Kernelnewbies Status page: http://www.kernelnewbies.org/status/
- The Linux-Kernel mailing list: http://www.tux.org/lkml/

See you at 2.7...