New Features in Linux Kernel v2.6 for high-end computing

4th Annual Workshop on Linux Clusters
for Super Computing
Linköping
October 22-24, 2003

Jes Sorensen
Wild Open Source Inc.
jes@wildopensource.com
http://www.wildopensource.com/

Acknowledgements

- I wouldn't be able to keep track of all these changes in the upcoming Linux kernel on my own
- I would like to thank the entire Linux community, and in particular:
  - William Lee Irwin III, IBM
  - Andrew Morton, OSDL
  - Dave Jones, SuSE

Agenda

- What is the 2.6 fuss all about?
- What applications will benefit the most from 2.6?
- Kernel improvements for scalability
- Q&A

2.6+ and all the fuzz

- Linux 2.5.x is a development kernel, not intended for production use
    - 2.6.0-testX is named to encourage users to test the kernel before the final release
- Linux 2.6 will be the first major stable kernel release since 2.4.0 - (January 7th, 2001)
- Current distributions all ship 2.4.x
- Wide adoption will happen when the major distributions start shipping 2.6

What applications will gain from 2.6?

- 2.6 will not be amazingly faster than 2.4 for normal workloads on normal machines
- It should be more pleasant to use on the desktop due to improved fairness and lower latency
- Major improvements in the areas where 2.4 was sub-optimal:
  - large number of processors
  - large amounts of memory
  - large number of threads and/or processes
  - large disks and number of disks
  - high performance networking
  - improved error handling

New O(1) scheduler

- Scheduling performance almost independent of number of processes and threads in the system
- Scales better on SMP systems - per CPU runqueues
- Interactive tasks receive special treatment, resulting in better interactive performance under load
- CPU affinity support
- Hyperthreading aware
- New improved thread implementation, NPRT (Native POSIX Threading Library)
- Fast userspace mutexes (Futex)

SMP locking

- All global locks removed from VM layer (pagemap_lru_lock and pagecache_lock)
- Block layer lock removed
- cli()/sti() (global interrupt disable) has been eliminated
- Developers are running Linux on 128 CPU machines (and bigger!)

Virtual Memory layer improvements

- Virtual Memory layer has been improved substantially
- RMAP (reverse mapping) provides virtual→physical and physical→virtual mappings
- Allows for more intelligent decisions on what memory to swap out/in to/from disk
- Large page support through hugetlbfs

Storage/disk handling

- New logical volume manager: Device Mapper (LVM2)
  - LVM1 has been removed completely
  - Compatible with LVM1 disk format
  - Requires new tools
- Large block device support:
  - 16TB on 32 bit architectures
  - 8EB on 64 bit architectures
- Large dev_t allows for large numbers of disks

Improved block I/O layer

- Rewritten and much improved error handling
- io_request_lock is gone - global lock which was used by all block device drivers and block layer
- “biobufs” (block I/O buffer), I/O requests larger than PAGE_SIZE
- 64 bit DMA directly to HIGHMEM
- SCSI is undergoing major improvements
    - Support for thousands of disks
- Multiple I/O schedulers available + scheduler tuning via sysfs

More block layer features

- Large block device limits:
  - 16TB (32 bit architectures)
  - 8EB (64 bit architectures)
- Asynchronuous I/O support (AIO) for filesystem I/O (requires O_DIRECT), no socket support (yet)

File Systems

- New high performance file systems: XFS, JFS
- Indexed directory support for ext3
- Large File System support
  - 2.4: 1TB or 2TB (1TB due to driver bugs)
  - 2.6: 10^18 bytes for XFS
- Larger files:
  - 16TB with a 4KB PAGE_SIZE (ia32)
  - 64TB with a 16KB PAGE_SIZE (ia64)
- devfs scaled down, eventually replaced by udev:
  - Uses /sbin/hotplug – device names policy in userland
- Basic NFSv4 support

Networking

- Networking was one of 2.4's stronger sides, expect fewer major performance improvements
- Improved NAPI support (reduces overhead under load by switching to polling mode)
- More driver performance updates and 10-Gigabit Ethernet support
- Native IPSec support
- Bridging firewall support

Other updates worth noticing (non performance etc.)

- New module loader, __init sections in modules are now freed
- PCI domain support (large # of PCI busses)
- New Input Layer (multiple keyboards etc)
- New system wide profiler (Oprofile)
- ACPI updates, better power management and software suspend-to-diskn
- x86-64, PPC64 and UML (user-mode-linux)

Questions?