Linux kernel 2.4 internals

Peter De Schrijver
p2@mind.be

Overview

- MP scalability improvements
  - Softnet
  - Finer grained locking
  - Interrupt affinity
- Logical Volume Management
- > 1GB and > 64GB RAM (support for intel PAE)
- Network improvements
  - Netfilter/IP tables
  - IP QoS

Overview(2)

- Device filesystem
- New technologies
  - USB
  - AGP

Softnet

- "Multi-threaded" Network device layer
- Polling -> event driven
- Separate timeout handler
- all drivers adapted

Finer grained locking

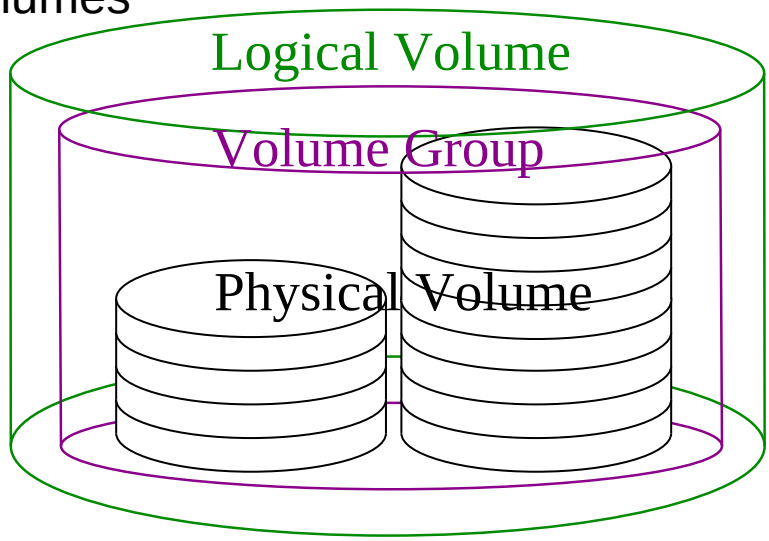
- New locking primitives
  - spinlocks
  - semaphores

Interrupt affinity

- Direct IRQ’s to specific CPU
- Improve locality
- used by TUX

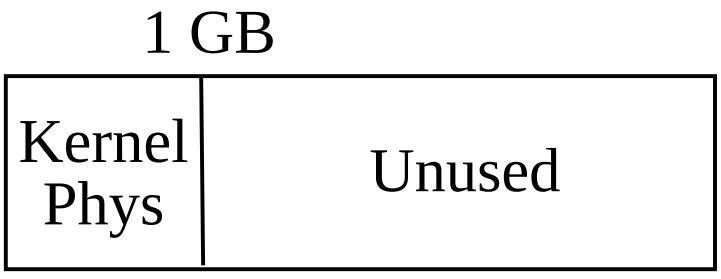
Logical Volume Management

- Similar to HP/UX LVM
- Physical volumes
- Volume Groups
- Logical Volumes

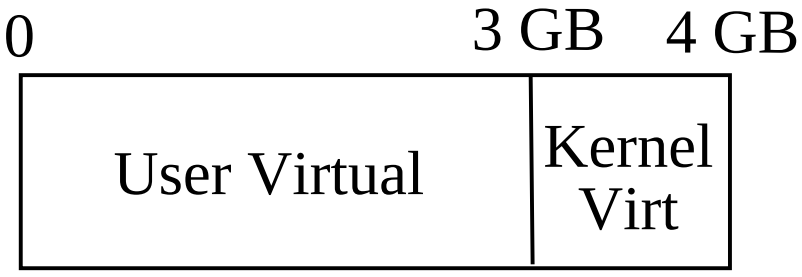


Highmem support

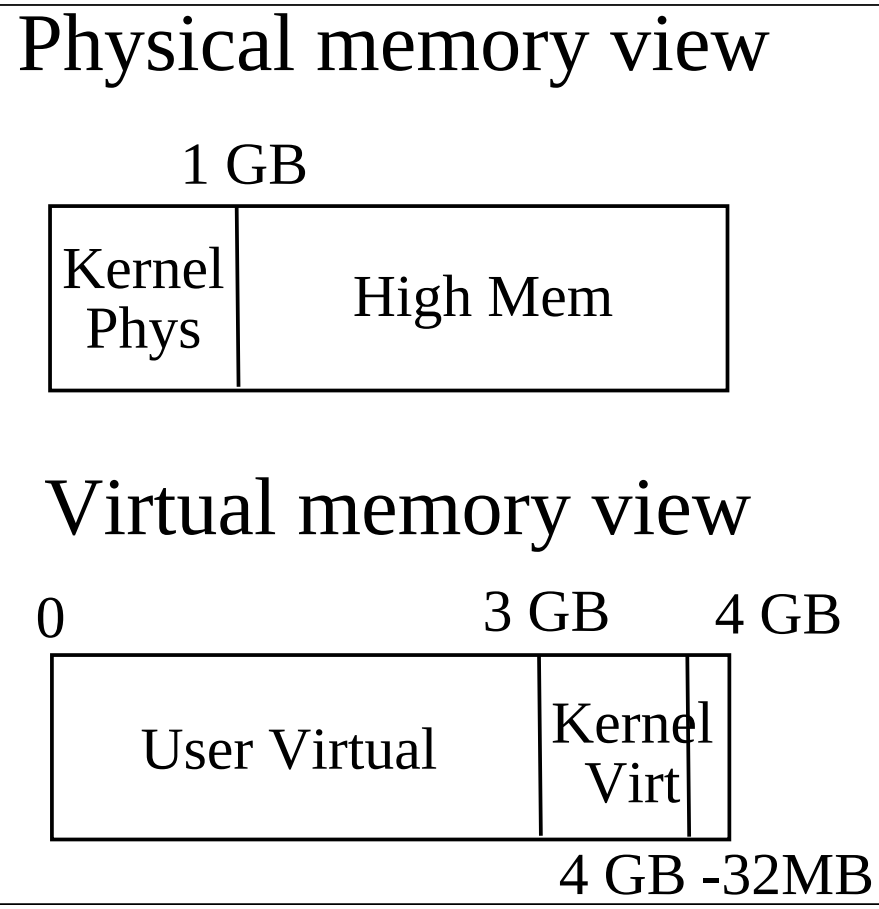
Physical memory view



Virtual memory view



Highmem support (2)



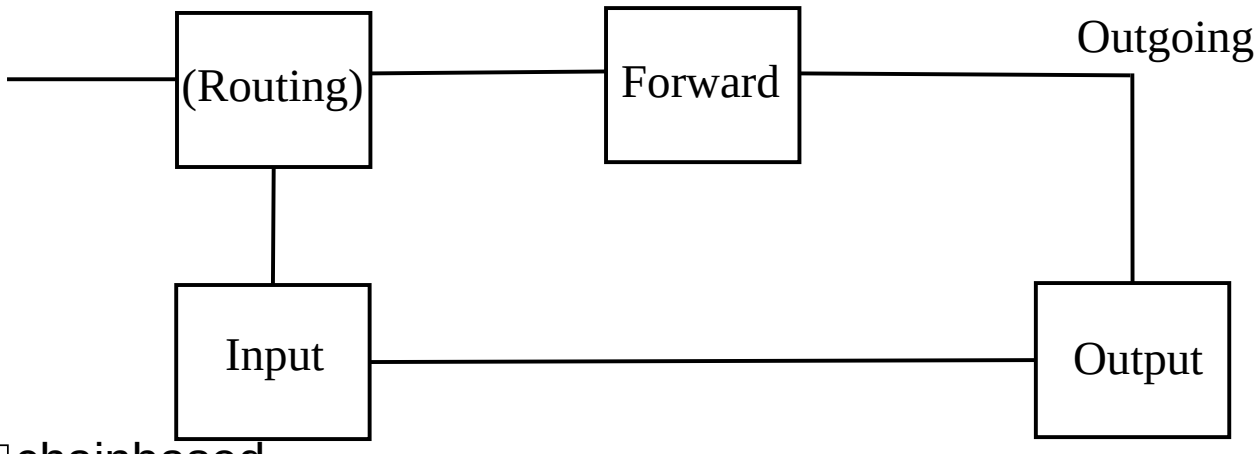
Netfilter

- kernel 2.0 : ipfwadm
- kernel 2.2 : ipchains
- firewall calls hardcoded
- hooks :
  - PRE_ROUTING
  - LOCAL_IN
  - FORWARD
  - LOCAL_OUT
  - POST_ROUTING
- actions :
  - NF_DROP
  - NF_ACCEPT
  - NF_QUEUE
  - NF_REPEAT
  - NF_STOLEN

Netfilter(2)

- packages using netfilter :
  - ipchains
  - ipfwadm+ipmasq
  - iptables
  - connection-tracking
  - NAT

IPtables

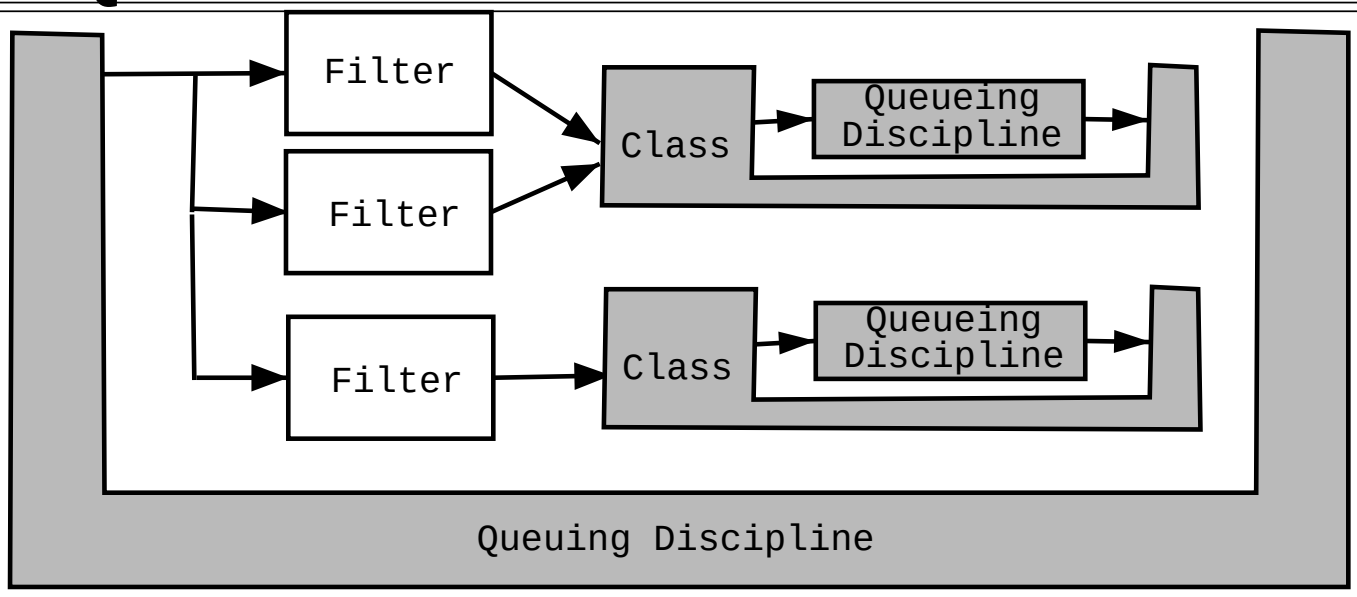


- chainbased
- new options :
  - rate limiting
  - queue for user space
  - return to end of chain

NAT

- SNAT (Source NAT) :
  - Always post routing
  - changes source IP & port
  - "internet connection sharing"
- DNAT (Destination NAT) :
  - pre routing
  - "Transparent proxying"

IP QoS



- Framework
  - Queueing disciplines
  - Filters
  - Classes

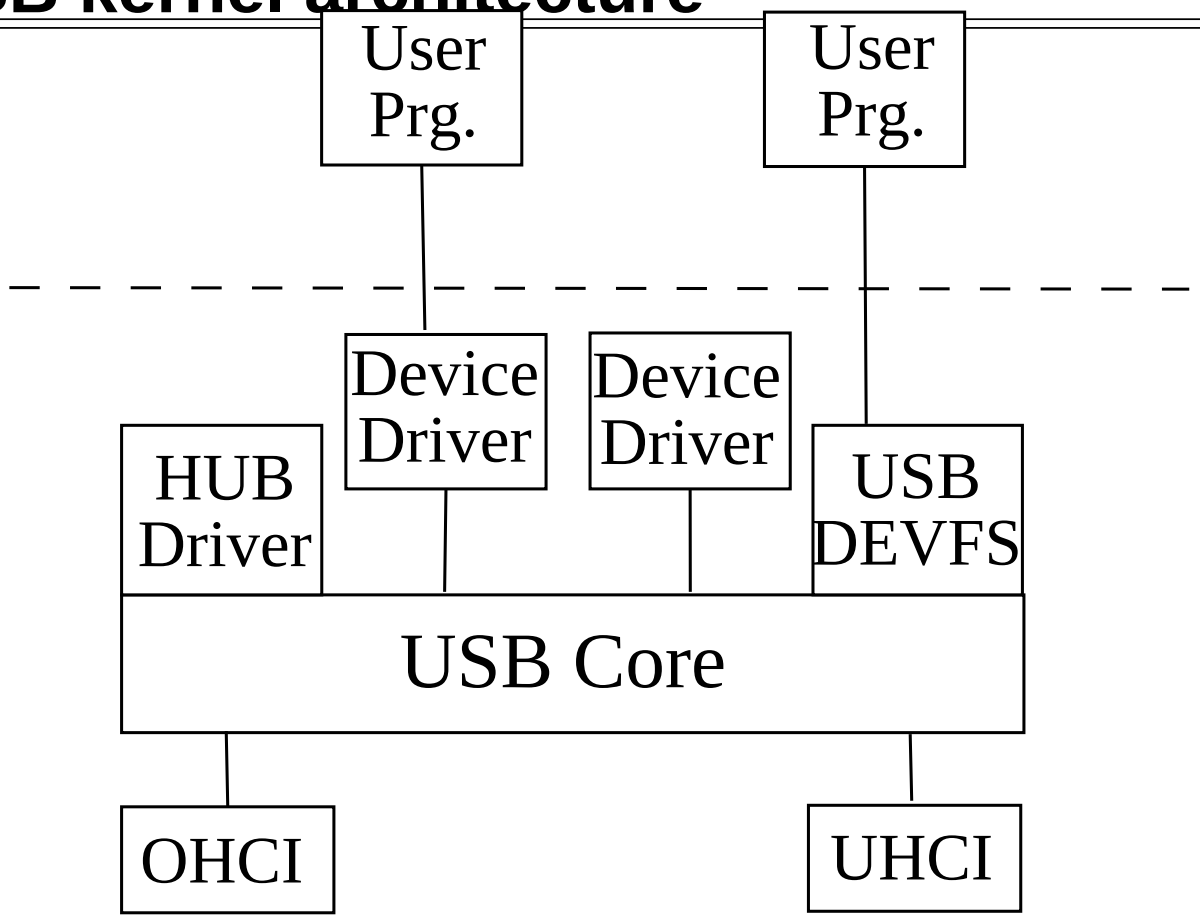
Device Filesystem

- Devices are special files
- Problems :
  - /dev management
  - /dev is too big
  - not enough major numbers
- Solution :
  - Pseudo filesystem similar to /proc
  - Drivers register by name
  - devfsd :
    - action at device create/delete time
    - set permissions
    - create symlinks

Device Filesystem(2)

- Advantages :
  - read-only rootfs
  - PTY security
  - speculative device scanning
  - intelligent device management

USB kernel architecture



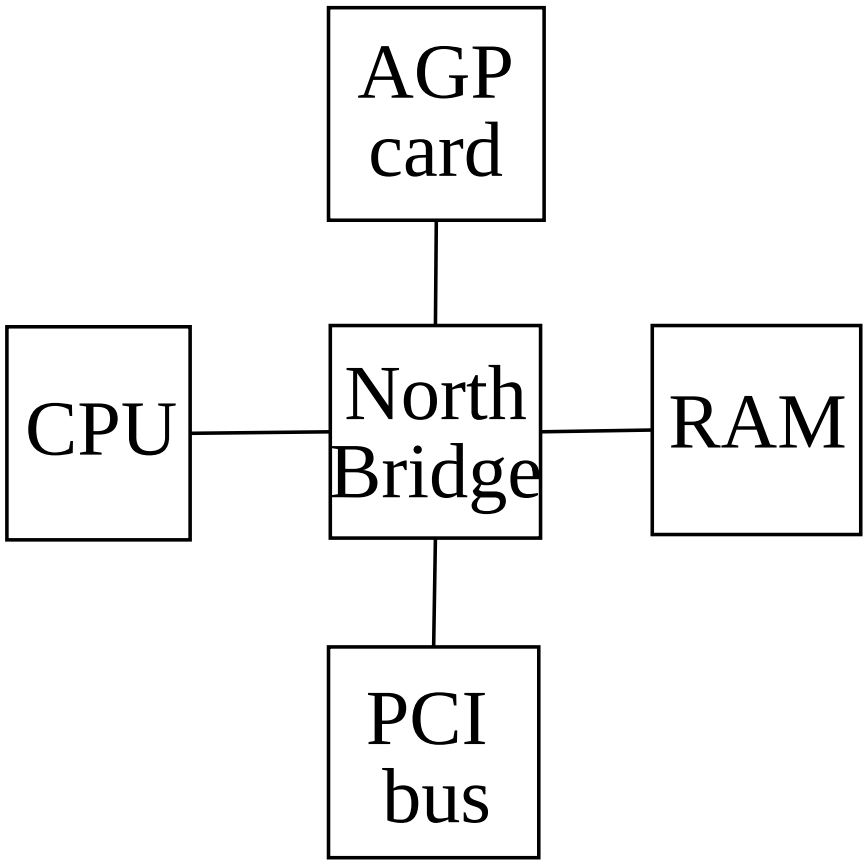
USB Daemon

- Triggers module load
- Starts script to configure devices
- Starts script to unconfigure devices

USB Device support

- standardized USB devices :
  - HID (keyboard, mouse, joystick, ...)
  - Audio
  - ACM Modems and ISDN TA’s
  - Printers
  - Bluetooth adapters
  - Storage devices
- not so standardized USB devices :
  - Scanners : sane
  - Webcams based on ov511, CPIA, IBM C-It
  - Diamond RIO 500
  - Prolific PL-2302 USB-to-USB
  - ADMtek Pegasus ethernet
  - DAB receiver from Institut fur Informatik TU-Munchen

AGP System architecture



AGP System architecture (2)

- High speed RAM access by AGP card
- Texture memory or framebuffer
- Large contiguous block
- => Problem in page-oriented systems
- Solution:
  - Graphics Address Remapping Table (GART)

AGP in linux

- /dev/agpgart
- AGP memory alloc/free
- mmap for fast access