Device Drivers

COMP9242
2011/S2 Week 5
Lecture outline

• Part 1: Introduction to device drivers
• Part 2: Overview of research on device driver reliability
• Part 3: Device drivers research at ERTOS
Some statistics

• 70% of OS code is in device drivers
  – 3,448,000 out of 4,997,000 loc in Linux 2.6.27
• A typical Linux laptop runs ~240,000 lines of kernel code, including ~72,000 loc in 36 different device drivers
• Drivers contain 3—7 times more bugs per loc than the rest of the kernel
• 70% of OS failures are caused by driver bugs
Part 1: Introduction to device drivers
OS archeology

The first (?) device drivers: I/O libraries for the IBM 709 batch processing system [1958]
OS archeology

The first (?) device drivers: I/O libraries for the IBM 709 batch processing system [1958]

Protection: prevent a user program from corrupting data belonging to the supervisor or to other programs
IBM 7090 [1959] introduced I/O channels, which allowed I/O and computation to overlap.
IBM 7090 [1959] introduced I/O channels, which allowed I/O and computation to overlap.

“the complex routines were required to allow even the simplest user program to take full advantage of the hardware, but writing them was beyond the capability of the majority of programmers.”

IBM 7094 [1962] supported a wide range of peripherals: tapes, disks, teletypes, flexowriters, etc.
IBM 7094 [1962] supported a wide range of peripherals: tapes, disks, teletypes, flexowriters, etc.
OS archeology

GE-635 [1963] introduced the master CPU mode. Only the hypervisor running in the master mode could execute I/O instructions.
Functions of a driver

- Encapsulation
  - Hides low-level device protocol details from the client
- Unification
  - Makes similar devices look the same
- Protection (in cooperation with the OS)
  - Only authorised applications can use the device
- Multiplexing (in cooperation with the OS)
  - Multiple applications can use the device concurrently
I/O device: a high-level view

- External medium
- I/O bus
- Internal logic
- Register file
- Bus interface
I/O devices in a typical desktop system
PCI bus overview

• PCI bus
  – Conventional PCI
    • Developed and standardised in early 90's
    • 32 or 64 bit shared parallel bus
    • Up to 66MHz (533MB/s)
  – PCI-X
    • Up to 133MHz (1066MB/s)
  – PCI Express
    • Consists of serial p2p links
    • Software-compatible with conventional PCI
    • Up to 16GB/s per device
PCI bus overview: memory space
PCI bus overview: DMA

CPU

Physical address space (FSB)

PCI controller

RAM

PCI memory space

Dev1  Dev2  Dev3
PCI bus overview: DMA

CPU → Dev1 → Dev2 → Dev3 → Physical address space (FSB)

PCI controller → IOMMU → RAM

PCI memory space

Dev1, Dev2, Dev3
DMA descriptors

Permanent DMA mappings
- Set up during driver initialisation
- Data must be copied to/from DMA buffers

Streaming mappings
- Created for each transfer
- Data is accessed in-place
PCI bus overview: interrupts

CPU

IRQ controller

PCI controller

RAM

Dev1

Dev2

Dev3

Dev1

Dev2

Dev3

PCI memory space

Physical address space (FSB)
PCI bus overview: config space

- PCI configuration space
  - Used for device enumeration and configuration
  - Contains standardised device descriptors

```
<table>
<thead>
<tr>
<th>Field</th>
<th>Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>Device ID</td>
<td>31-0</td>
</tr>
<tr>
<td>Vendor ID</td>
<td>16-15</td>
</tr>
<tr>
<td>Status</td>
<td>14-8</td>
</tr>
<tr>
<td>Command</td>
<td>12-9</td>
</tr>
<tr>
<td>Class Code</td>
<td>11-8</td>
</tr>
<tr>
<td>Revision ID</td>
<td>10-8</td>
</tr>
<tr>
<td>BIST</td>
<td>7-4</td>
</tr>
<tr>
<td>Header Type</td>
<td>3-0</td>
</tr>
<tr>
<td>Lat. Timer</td>
<td>16-15</td>
</tr>
<tr>
<td>Cache Line S.</td>
<td>0</td>
</tr>
<tr>
<td>Base Address Registers</td>
<td></td>
</tr>
</tbody>
</table>

Cardbus CIS Pointer

<table>
<thead>
<tr>
<th>Field</th>
<th>Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subsystem ID</td>
<td>28-24</td>
</tr>
<tr>
<td>Subsystem Vendor ID</td>
<td>24-20</td>
</tr>
</tbody>
</table>

Expansion ROM Base Address

Reserved

Cap. Pointer

Reserved

<table>
<thead>
<tr>
<th>Field</th>
<th>Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max Lat.</td>
<td>28-24</td>
</tr>
<tr>
<td>Min Gnt.</td>
<td>24-20</td>
</tr>
<tr>
<td>Interrupt Pin</td>
<td>20-16</td>
</tr>
<tr>
<td>Interrupt Line</td>
<td>16-12</td>
</tr>
</tbody>
</table>
```
PCI bus overview: I/O space

- I/O space
  - obsolete
Writing a driver for a PCI device

• Registration
  – Tell the OS which PCI device ID's the driver supports

• Instantiation
  – Done by the OS when it finds a driver with a matching ID

• Initialisation
  – Allocate PCI resources: memory regions, IRQ's
  – Enable bus mastering
  – Permanent DMA mappings: disable caching
Writing a driver for a PCI device

• Interrupt handler
  – Return ASAP to re-enable interrupts; perform heavy-weight processing in a separate thread

• DMA
  – Streaming mappings: may require bounce buffers

• Power management
  – Prepare the device for a transition into a low-power state
  – Restore device configuration during wake-up
USB bus overview

- **USB bus**
  - Host-centric
  - Distributed-system-style architecture (packet-based transfers)
  - Hot plug
  - Power management
    - Bus-powered and self-powered devices

- **USB 1.x**
  - Up to 12Mb/s

- **USB 2.0**
  - Up to 480Mb/s

- **USB 3.0**
  - Up to 4.8Gb/s
USB bus overview

Device 1

Device 2

Root hub

Device 3

Device 4

DMA

USB bus controller

DMA

Transfer descriptors

Completions
I/O devices in a typical desktop system
Driver stacking
Driver stacking

- CPU
- Memory controller
- PCI bus controller (ICH7)
- USB controller (EHCI)
- AX88772 Ethernet controller
- USB Ethernet controller (AX88772)
- Front-side bus
- TCP/IP stack
- AX88772 Ethernet driver
- hard_start_xmit(pkt)
Driver stacking

- CPU
- Memory controller
- PCI bus controller (ICH7)
- Front-side bus
- USB bus
- USB Ethernet controller (AX88772)
- USB controller (EHCI)
- USB EHCI controller driver
- AX88772 Ethernet driver
- TCP/IP stack
  - hard_start_xmit(pkt)
  - usb_submit_urb(urb)
Driver stacking

- CPU
- Memory controller
- PCI bus controller (ICH7)
- USB bus
- USB controller (EHCI)
- USB Ethernet controller (AX88772)
- PCI bus driver
- USB EHCI controller driver
- AX88772 Ethernet driver
- TCP/IP stack

Functions:
- hard_start_xmit(pkt)
- usb_submit_urb(urb)
- mem loads/stores
Driver stacking

- CPU
- Front-side bus
- Memory controller
- PCI bus controller (ICH7)
- USB bus
- USB controller (EHCI)
- USB Ethernet controller (AX88772)
- PCI bus
- AX88772 Ethernet driver
- USB framework
- USB EHCI controller driver
- PCI framework
- TCP/IP stack
Driver framework design patterns

The driver pattern

The bus pattern
Driver framework software architecture
Questions?
Part 2: Overview of research on device driver reliability
Some statistics

• 70% of OS code is in device drivers
  – 3,448,000 out of 4,997,000 loc in Linux 2.6.27

• A typical Linux laptop runs ~240,000 lines of kernel code, including ~72,000 loc in 36 different device drivers

• Drivers contain 3—7 times more bugs per loc than the rest of the kernel

• 70% of OS failures are caused by driver bugs
Understanding driver bugs

- Driver failures
Understanding driver bugs

- Driver failures
  - Memory access violations
  - OS protocol violations
    - Ordering violations
    - Data format violations
    - Excessive use of resources
    - Temporal failure
  - Device protocol violations
    - Incorrect use of the device state machine
    - Runaway DMA
    - Interrupt storms
  - Concurrency bugs
    - Race conditions
    - Deadlocks
User-level device drivers

- Each driver is encapsulated inside a separate hardware protection domain
- Communication between the driver and its client is based on IPC
- Device memory is mapped into the virtual address space of the driver
- Interrupts are delivered to the driver via IPC's
User-level drivers in $\mu$-kernel OSs

Diagram:
- User land
  - Driver
  - TCP/IP
  - Application
- Kernel
  - DMA
  - IPC
  - IPC
User-level drivers in µ-kernel OSs
User-level drivers in μ-kernel OSs

- Net filter
- Driver
- Application
- TCP/IP

User land

IPC

Kernel
User-level drivers in µ-kernel OSs
Driver performance characteristics
Driver performance characteristics

- **I/O throughput**
  - Can the driver saturate the device?

- **I/O latency**
  - How does the driver affect the latency of a single I/O request?

- **CPU utilisation**
  - How much CPU overhead does the driver introduce?
Improving the performance of ULD
Improving the performance of ULD

- Ways to improve user-level driver performance
  - Shared-memory communication
  - Request queueing
  - Interrupt coalescing
Implementing efficient shared-memory communication

**Issues:**
- Resource accounting
- Safety
- Asynchronous notifications
Rbufs

- Proposed in the Nemesis microkernel-based multimedia OS
Early implementations of ULD

- Michigan Terminal System [1970's]
  - OS for IBM System/360
  - Apparently, the first to support user-level drivers
- Mach [1985-1994]
  - Distributed multi-personality μ-kernel-based multi-server OS
  - High IPC overhead
  - Eventually, moved drivers back into the kernel
- L3 [1987-1993]
  - Persistent μ-kernel-based OS
  - High IPC overhead
  - Improved IPC design: 20-fold performance improvement
  - No data on driver performance available
More recent implementations

- **Sawmill [~2000]**
  - Multiserver OS based on automatic refactoring of the Linux kernel
  - Hampered by software engineering problems
  - No data on driver performance available

- **DROPS [1998]**
  - L4 Fiasco-based real-time OS
  - ~100% CPU overhead due to user-level drivers

- **Fluke [1996]**
  - ~100% CPU overhead

- **Mungi [1993—2006]**
  - Single-address-space distributed L4-based OS
  - Low-overhead user-level I/O demonstrated for a disk driver
Currently active systems

• Research
  - seL4
  - MINIX3
  - Nexus

• Commercial
  - OKL4
  - QNX
  - GreenHills INTEGRITY
User-level drivers in a monolithic OS

Ben Leslie et al. User-level device drivers: Achieved performance, 2005
User-level drivers in a monolithic OS

Ben Leslie et al. User-level device drivers: Achieved performance, 2005
User-level drivers in a monolithic OS

Ben Leslie et al. User-level device drivers: Achieved performance, 2005
User-level drivers in a monolithic OS

Ben Leslie et al. User-level device drivers: Achieved performance, 2005
User-level drivers in a monolithic OS

Ben Leslie et al. User-level device drivers: Achieved performance, 2005

• Performance
  - Up to 7% throughput degradation
  - Up to 17% CPU overhead
  - Aggressive use of interrupt rate limiting potentially affects latency (not measured).
Nooks

- A complete device-driver reliability solution for Linux:
  - Fault isolation
  - Fault detection
  - Recovery
Nooks

- A complete device-driver reliability solution for Linux:
  - Fault isolation
  - Fault detection
  - Recovery
Nooks

• A complete device-driver reliability solution for Linux:
  - Fault isolation
  - Fault detection
  - Recovery

- Driver
  - Heap
  - Stacks

- Shadow driver

- Isolation manager:
  - XPC
  - Copying/replication
  - Checking

- Linux kernel
  (read-only for the driver)
Nooks

• A complete device-driver reliability solution for Linux:
  – Fault isolation
  – Fault detection
  – Recovery

• Problems
  – The driver interface in Linux is not well defined. Nooks must simulate the behaviour of hundreds of kernel and driver entry points.

• Performance
  – 10% throughput degradation
  – 80% CPU overhead
Virtualisation and user-level drivers

- Direct I/O
Virtualisation and user-level drivers

- Paravirtualised I/O
Paravirtualised I/O in Xen

- Xen I/O channels are similar to rbufs, but use a single circular buffer for both requests and completions and rely on mapping rather than sharing.
Xen I/O channels

Response producer:
- rsp_prod_pvt
- req_cons
- nr_ents
- *sring

Request producer:
- req_prod_pvt
- rsp_cons
- nr_ents
- *sring

Shared:
- req_prod
- req_event
- rsp_prod
- rsp_event

PUSH_REQUESTS updates this pointer
req_prod<--req_prod_pvt

Unconsumed requests
0 1
255
254

Unconsumed responses

PUSH_RESPONSES updates this pointer
rsp_prod<--rsp_prod_pvt
Paravirtualised I/O in Xen

- Performance overhead of the original implementation: 300%
  - Long critical path (increased instructions per packet)
  - Higher TLB and cache miss rates (more cycles per instructions)
  - Overhead of mapping

- Optimisations
  - Avoid mapping on the send path (the driver does not need to “see” the packet content)
  - Replace mapping with copying on the receive path
  - Avoid unaligned copies
  - Optimised implementation of page mapping
  - CPU overhead down to 97% (worst-case receive path)
Other driver reliability techniques

- Implementing drivers using safe languages
  - **Java OSs: KaffeOS, JX**
    - Every process runs in a separate protection domain with a private heap. Process boundaries are enforced by the language runtime. Communication is based on shared heaps.
  - **House (Haskell OS)**
    - Bare-metal Haskell runtime. The kernel and drivers are in Haskell.
    - User programs can be written in any language.
  - **SafeDrive**
    - Extends C with pointer type annotations enforced via static and runtime checking
      ```c
      unsigned n;
      struct e1000 buffer * count(n) bufinfo;
      ```
Other driver reliability techniques

- Implementing drivers using safe languages
  - Singularity OS
    - The entire OS is implemented in Sing#
    - Every driver is encapsulated in a separate software-isolated process
    - Processes communicated via messages sent across channels
    - Sing# provides means to specify and statically enforce channel protocols
Other driver reliability techniques

- **Static analysis**
  - SLAM, Blast, Coverity
  - Generic programming faults
    - Release acquired locks; do not acquire a lock twice
    - Do not dereference user pointers
    - Check potentially NULL-pointers returned from routine
  - Driver-specific properties
    - “if a driver calls another driver that is lower in the stack, then the dispatch routine returns the same status that was returned by the lower driver”
    - “drivers mark I/O request packets as pending while queuing them”
  - Limitations
    - Many properties are beyond reach of current tools or are theoretically undecidable (e.g., memory safety)
Questions?