Virtualisation
Peripheral Buses

COMP311 2008

Jamie Curtis

Virtualisation

- Allows a “Host” OS to execute a “Guest” OS as a task
  - No need for “Guest” OS’s cooperation
  - Host is often called a Hypervisor or VMM
- VMM controls devices
  - Often presents virtual devices to guest OS’s

Virtualisation cont.

- PC architecture makes this tricky
  - Protected mode introduced privilege rings
    - 4 Levels – 0,1,2,3 with decreasing privilege
  - Operating systems assume full control (ring 0)
    - and userspace at ring 3
    - There is no control on reading this state
  - Some instructions silently fail if not run at the correct privilege level instead of causing a fault

Virtualisation cont.

- 4 approaches to fix this
  - Emulation
  - Paravirtualisation
  - Binary Translation
  - Hardware Assisted
- Emulation
  - Fake the entire machine
  - Very slow but doesn’t require the same host architecture as the guest

Paravirtualisation

- Alter the guest so that it doesn't use the "bad" instructions.
  - Guest instead calls out to the VMM
- Very fast with minimal overhead
- Requires support from all guest OS's
  - Effectively limited to open source OS's
- Championed by Xen

Binary Translation

- The VMM now dynamically translates the byte stream before it's executed, replacing "bad" instructions as it goes
  - Lots of optimisations make this not nearly as bad as it sounds.
- Allows running un-modified OS's
- Performance hit can be anything from 5 – 60% depending on workload.
- Championed by VMware

Hardware Virtualisation

- Adds another privilege level for the VMM
- Hardware maintains state for each guest
- VMM gets very flexible control of what causes faults
- Intel and AMD again have similar specs (but incompatible !)
  - AMD – "*Pacifica*" – "*AMD-V*"
  - Intel – "*Vanderpool*" – "*Intel VT*"

Hardware Virtualisation cont.

- Initial solutions don't deal with enough in hardware, causing them often to be slower than BT
  - Transition into and out of guests is very slow
- VMware and Xen both have added support
- Hardware support is getting more complete in newer versions

Virtual I/O Devices

- Typically VMMs provide virtual hardware drivers to the guests
  - These may then map onto real hardware inside the VMM
- Allowing secure and separated direct access to hardware is difficult

Extensions

- AMD and Intel’s latest architectures add a number of important virtualisation enhancements
  - Primarily focused around making fewer calls into the VMM
- Three main enhancements
  - Nested Page Tables
  - Tagged TLB
  - Device Exclusion Vector (DEV)

Nested Page Tables

- Current designs use shadowed page tables
  - The MMU is setup for the current guest / process
  - Changes to the MMU are trapped by the VMM
- Nested page tables replicate the page table per VM.
  - Looks similar to the virtual address space provided by the OS to processes, just one more level up
  - Makes lookups through the page table slower

Tagged TLB

- TLB’s cache lookups through the page tables
  - Typically they are flushed each time you switch process or VM
- Tagged TLB’s add a tag to reference which VM this tag relates to
  - No longer have to flush TLB’s when switching VM
  - Makes VM – VMM – VM transitions cheaper
  - Reduces the performance hit of nested page tables

Device Exclusion Vector

- Contains a mapping of what pages a device can access
- Allows the VMM to give DMA access to specific pages
  - Allows a device to do DMA directly into a specific VM's memory space

Virtualisation Future

- Linux now has 4 full virtualisation packages
  - VMWare , Xen, KVM, Lguest
- Intel and AMD working on next-gen hardware support
- Dense, multi-core solutions making virtualisation very attractive
  - Power and space efficient
- Keeps getting more and more important

PC Buses

- ISA is the first generation bus
  - 8 bit on IBM XT
  - 16 bit on 286 or above (16MB/s)
- Extended through VESA system
  - Tied into the 486 Memory bus
- IBM tried to make a Licensed bus
  - MicroChannel Architecture (MCA)

PCI

- Designed started by Intel around 1990
  - 1.0 Specification released in 1992
  - 2.0 Specification released in 1993
    - 2.0 specification was the first to define physical slots
- PCI now controlled by the PCI SIG
- Introduced as a 32bit 33MHz bus
  - Allowing for a total bus bandwidth of 133MB/s
- For electrical reasons the bus is limited to 5 physical slots

PCI

- PCI 2.0 used 5v signalling
  - PCI slot keying determined voltage
- Later versions of PCI introduced 3.3v signalling
  - Cards can be 5v only, 3.3v only or universal
- PCI was first true PnP bus
  - Cards contain “Configuration Space” detailing their requirements.
  - Host OS allocates resources according to requirements

PCI

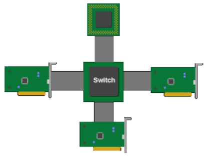
- Problem is PCI is too slow for many new applications
- To make PCI faster there are two options
  - Wider bus
  - Faster clock
- Wider bus makes motherboard layout harder and more expensive
- Increasing clock also makes layout harder, but also reduces number of devices
  - Each device on the bus creates more noise

PCI Express

- Formally called 3GIO
- Standardised by PCI SIG
- Designed to give much higher performance than PCI while maintaining software compatibility
- Completely redesigned physical and electrical layer. Transparent to software

PCI Express

- Full duplex serial connection
  - Differential 8B/10B serial signalling
  - 2.5Gbps per direction
    - 250MB/s per direction
- Point to point
  - PCI Express switch in the center

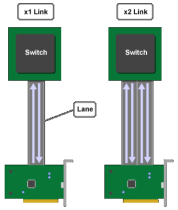


PCI Express

- Packet switched system
- Central switch allows Quality of Service
  - Real time (streaming) packets can take priority over other types of data
- 250MB/s is still not enough for 3D cards !

PCI Express Lanes

- A single card can use multiple PCI Express lanes
  - Each byte in turn is striped across a different lane
  - 1x, 2x, 4x, 8x, 16x and 32x are standardised



PCI Express Lanes

- Three different issues
  - Card size
  - Connector size
  - Link size
- The connector must be the same size or larger than the card
- The link may be the same size or smaller then either the card or connector

PCI Express 2.0

- Standardised in Jan 07
- First chipsets (Intel X38) arriving now
- Clock has doubled from 2.5GHz to 5GHz
  - Still 8b/10b
  - 500MB/s per direction per lane
- Fully backwards compatible with 1.1 cards
- New power connector
  - Changed from 75w 6 pin to 150w 8 pin for increasing GPU power demands

PCI Express 3.0

- In development
- Standards expected in 2009
- Drops 8b/10b, 8GHz clock
  - Bandwidth up to 1GB/s per lane per direction

USB

- Universal Serial Bus
- Controlled by the USB Implementers Forum (USB-IF)
  - Formed in 1995
- USB 1.0 specification released in 1996
- Designed to be the universal connector
  - Replacing PS/2, serial, parallel, game ports etc
  - The ideal aim of a “Legacy Free” PC is still to be realised.

USB

- USB 1.1 specification released in 1998
- Under USB 1.1 devices can operate in one of two speeds
  - Low-Speed (1.5 Mbps)
  - Full-Speed (12 Mbps)
- USB 2.0 specification released in 2000
  - Introduced Hi-Speed (480 Mbps) mode

USB

- USB is designed as a bus technology
  - By using hubs you can connect up to 127 devices
- USB is half duplex
  - USB cables contain a differential data pair and a power pair
- USB is a Master – Slave arrangement.
  - Uses different ports to distinguish this
    - A ports (upstream)
    - B ports (downstream)

USB

- USB is a smart bus
  - Devices can run in three different modes
    - Interrupt, Bulk, Isochronous
  - Isochronous and Interrupt devices request a bandwidth
  - Host will allocate up to 90% of the bus to them
  - Bulk transfers get whatever is left
  - Isochronous data is not error corrected
  - Devices detail power usage and host can stop a device powering up

USB

- Master – Slave mode causes problems for many devices
  - For example, a PDA syncs data to host via USB but also wants to be able to have a USB keyboard plugged into it.
  - A digital camera wants to be plugged into a printer to print photos.
    - Both a Master and a Slave arrangement
  - USB On-The-Go introduced in 2001

USB

- Why was USB so successful ?
  - Big industry support
  - Simple
  - Standardised host controller interface
  - Standardised protocols for many devices
    - Printer
    - Keyboards + Mice
    - Storage

USB 3.0

- Specification to be released in 2008
- Adds fibre-optic connection into the same cable
  - Existing copper links for backwards compatibility and power
- “More than 10x the bandwidth”
  - Therefore at least 4.8Gbps
- Increased power efficiency