



Xen Extensions to Enable Modular/3rd Party Device Emulation for HVM Domains

John Zulauf
Staff Software Engineer
Simulation and Performance Team

Visual
Computing
Group

1

Problem Statement

- Solutions may require specific devices
 - Connectivity
 - Legacy requirements
 - New applications
- Adding a device for every HVM is difficult
 - Qemu-dm requires custom code per device
 - 1-1 Mapping supports only a single Domain
 - IOMMU requires a real device per Domain
- Modular Device Emulation
 - Must support a variety of configuration/resource allocations
 - Must support a variety of uses—emulation, sharing, ...
 - Connect easily to Qemu-dm
 - Must allow proprietary/value-add technologies

Visual
Computing
Group

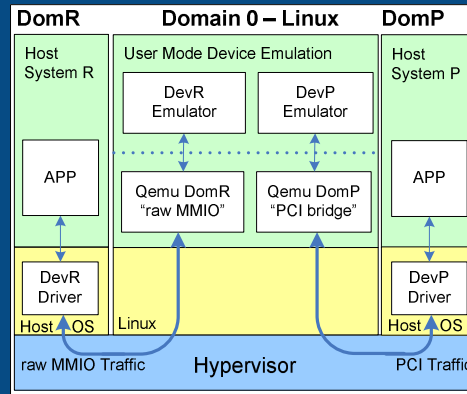
Intel is a trademark or registered trademark of Intel Corporation or its subsidiaries in the United States and other countries.
* Other names and brands may be claimed as the property of others.



2

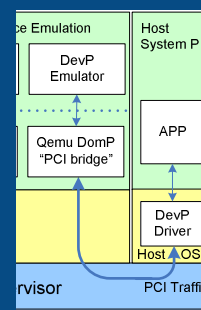
Approach/Xen Extensions

- Device Emulators
 - Standalone
 - Socket connected
- Connection Extensions
 - PCI-PCIe bridge
 - “raw” MMIO routing
- IO Handling Extensions
 - IO Merge
 - Asynchronous IO
- Interrupt Extensions



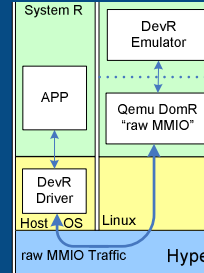
PCI-PCIe External Bridge

- Implements standard bridge
 - Intel ® 31154 PCI Bridge
 - Recognized by standard Host OS drivers
 - Uses PCIe TLP based wire protocol
 - Unix or TCP Domain Sockets
- Extensions support PCI-PCIe operations
 - INTx messaging, reordering, merging
- New CONFIG cycle support
 - Secondary bus routed through bridge
 - Bridged device discovery in hvmloder
- Enables proprietary device models & emulators
 - Socket connection isolates from GPL'd Qemu
- Supports IO merging & Asynchronous IO



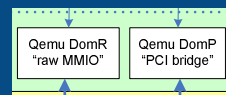
"raw MMIO" Socket

- Need for legacy/non-plug-and-play devices
 - Resource allocation fixed or non-standard
- Flexible range definitions
 - Declared through qemu-dm arguments
 - Read only, Write only, Prefetchable, Posted/Non-posted
- Reuses PCIe TLP Wire protocol & sockets
 - Non-standard uses support generic MMIO
- Supports IO merging & Asynchronous IO



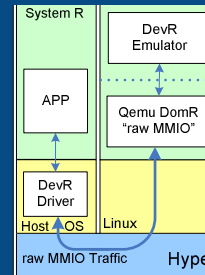
IO Handling Extensions

- Requires Support from Device
 - Device emulator registers additional callbacks
- IO Merging
 - Handles up to 4K rep;movs
 - Merged transactions, single iomem_index/mapcache
 - Prefetchable memory space **only**
 - 5x improvements vs. Dword over PCIe bridge
- Asynchronous IO
 - Required by both PCI-PCIe and "raw MMIO"
 - PCIe Deadlocks, MMIO "fence registers"
 - Only blocks VCPU making pending IO request
 - Qemu timer events free to operate
 - Other VCPU IO requests can be initiated/completed
 - Measurable performance increases for > 1 VCPU
 - Supports all memory space IOREQ_TYPE's



Interrupt Extension

- Supports Non-standard interrupt routing
 - “raw” device emulator to DomR
 - Examples:
 - DMA completion interrupt
 - External event interrupts DevR
- Backwards compatible
 - Won’t interfere with legacy or PCI
 - Additional set of VIOAPIC pins
 - Separate HVMOP to program

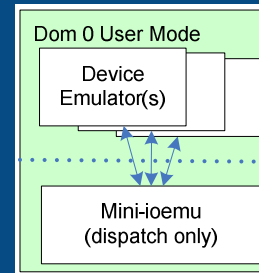


Results

- Interactive Performance
 - Able to run long term burn-in and stress tests
 - Transaction rate highest for 4B transfers
 - Bandwidth highest for 4K transfers
- Usable Development Platform
 - Simplified emulation development environment
 - No reconfiguration/installation of Xen
- Xen Aware Tuning – “para-optimizations”
 - MMIO VMEXIT frequently dominates performance
 - Recoding of driver and device runtime for “rep;mov”
 - Focus on DMA features when available in Dev*

Next Steps

- Enhanced PCIe support
 - Convert bridge to “modern” root port
 - Peer-to-peer root complex
 - MSI/MSI-x
 - 64 bit discovery
- “raw” MMIO
 - Allow dynamic resourcing (i.e. by DevR)
- Hypervisor MMIO tuning
- A “mini-ioemu”?
 - All devices modular
 - No devices built in



Acknowledgements

- VCG Management whose support made the development of this project possible
- The rest of the “Performance and Simulation” team that made it a reality

Questions?

**Visual
Computing
Group**

Intel is a trademark or registered trademark of Intel Corporation or its subsidiaries in the United States and other countries.
* Other names and brands may be claimed as the property of others.

