

Development of I/O Pass-through: Current Status & the Future

Nov 21, 2008

Yuji Shimada

NEC System Technologies, Ltd.

Agenda

1.Implementation of I/O Pass-through

2.Future Enhancement Plan

3.Challenges for I/O Pass-through

4.Conclusion

Xen Summit Tokyo 2008

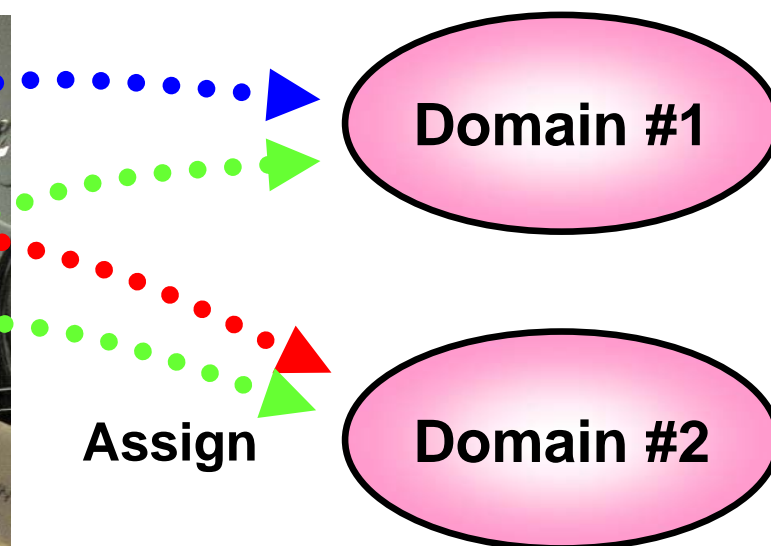
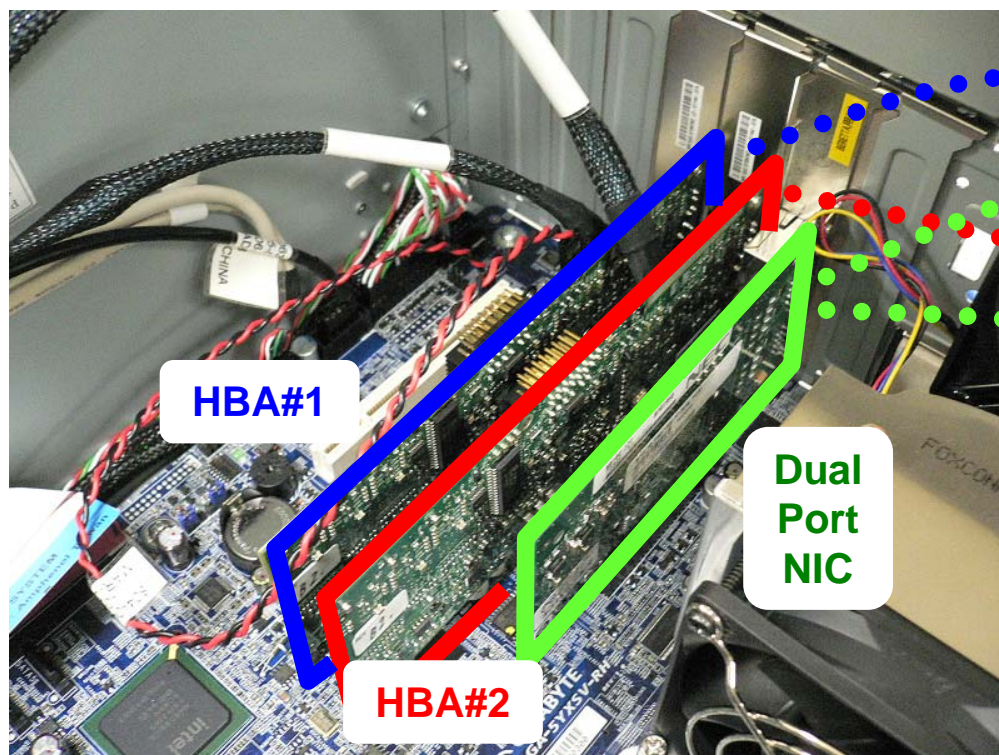
1. Implementation of I/O Pass-through

What is I/O Pass-through

- **Assigning I/O device to domain and directly use from guest software**

Assign HBA → High performance access to storage

Assign NIC → High performance access to network



In this presentation I will talk about I/O pass-through for HVM domain

Xen Summit Tokyo 2008

Overview of I/O Pass-through (1)

- **MMIO Virtualization**

- The mapping from guest physical address to machine address is done per-page
- Reassign Memory Resources to I/O Device
 - If BIOS assigned non-page-aligned MMIO resource, then domain0 linux reassign page-aligned MMIO resource to I/O device

Enhancement 1

- **Configuration Register Access Virtualization**

- ioemu traps configuration register access from guest software, and emulates or pass-through as to bit type

Enhancement 2

Xen Summit Tokyo 2008

Summary of I/O Pass-through (2)

- **DMA Virtualization**
 - Device driver writes guest physical address to the I/O device
 - I/O device executes DMA by using guest physical address
 - IOMMU on chipset translates guest physical address to machine address
- **Port I/O Virtualization**
 - Hypervisor traps IN/OUT instructions and accesses I/O device instead of guest software
- **Interrupt Virtualization**
 - Hypervisor receives interrupt from I/O device, emulates interrupt controller, and injects interrupt to HVM domain

Enhancement 1 – Reassign Memory Resources to I/O Device

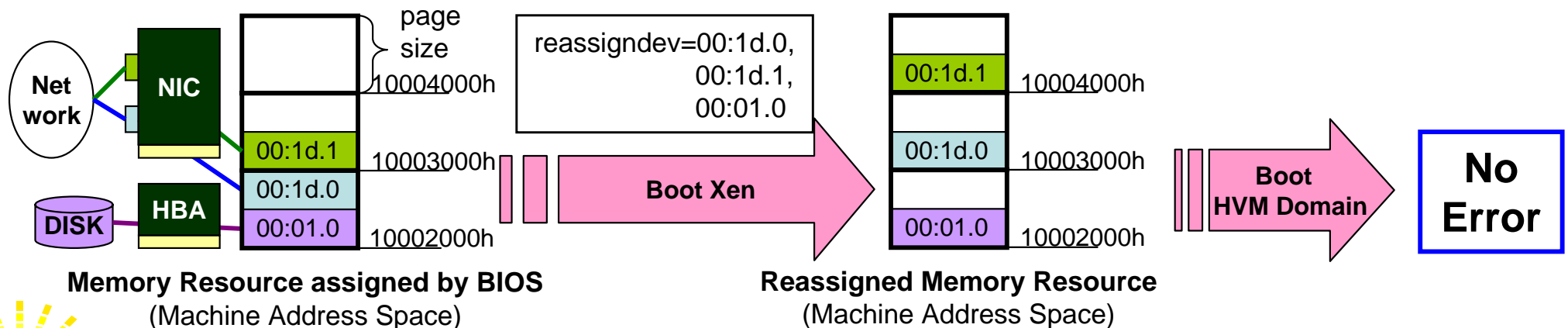
- **Problem**

- If the resources assigned by BIOS is not page-aligned, assigning I/O device to HVM domain fails

Error: pci: 0000:00:1d.0: non-page-aligned MMIO BAR found.

- **Enhancement**

- Release resources set by BIOS, and reassign page-aligned resources to I/O device
- Add boot parameter to specify I/O device to reassign resources



We can assign I/O device to HVM domain regardless of the initial resource assignment by BIOS.

Xen Summit Tokyo 2008

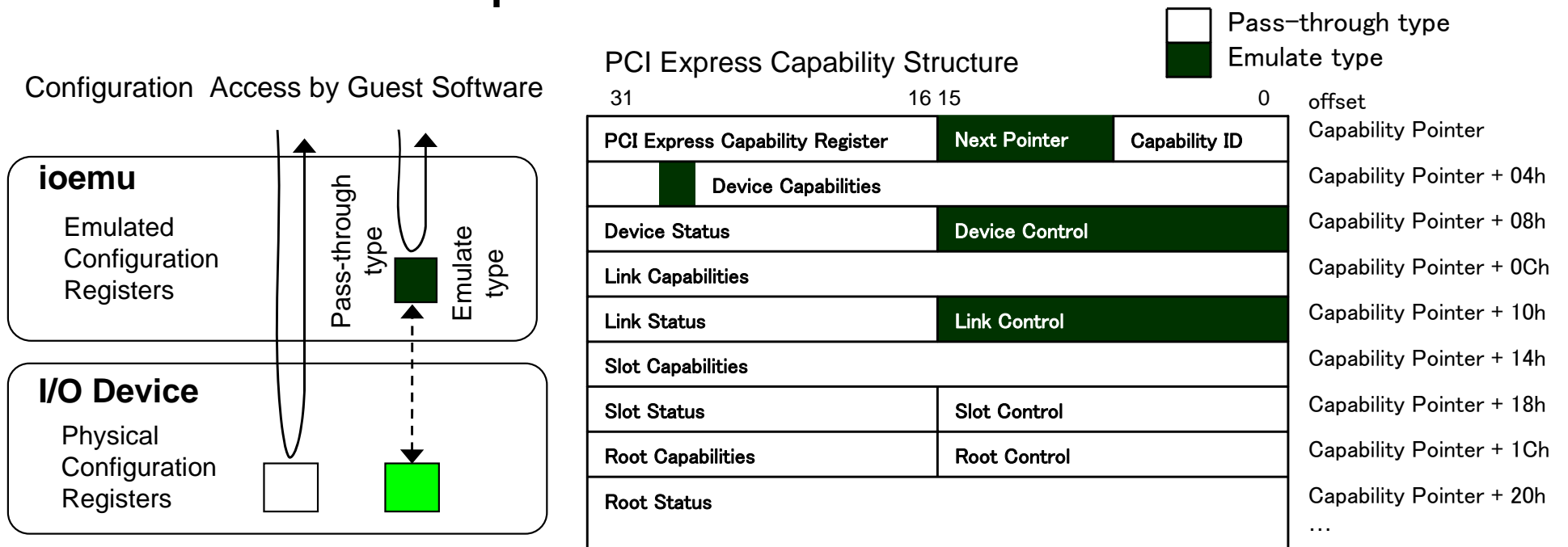
Enhancement 2 – Configuration Register Access Virtualization (1)

- **Problem**
 - I/O device which can be assigned to HVM domain was limited because the registers other than Configuration Header Type 0 were read only
- **Enhancement**
 - **Allow guest software to read/write access to the following registers, and hide other registers**
 - Configuration Header Type 0
 - MSI Capability Structure
 - MSI-X Capability Structure
 - PCI Express Capability Structure
 - PCI Power Management Capability Structure
 - Vital Product Data Capability Structure
 - Vendor Specific Capability Structure
 - Device Specific Register (excluding Header & Capability Structure)

Xen Summit Tokyo 2008

Enhancement 2 – Configuration Register Access Virtualization (2)

- Enhancement (cont.)
 - Pass-through the directly controllable bit by guest software
 - Emulate the bit that BIOS initial settings should be protected and the bit which has special reasons



We can assign several types of I/O device, USB and SAS HBA etc. Protect the whole system from invalid guest software behavior.

2. Future Enhancement Plan

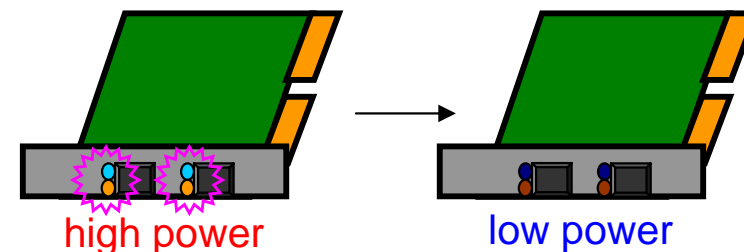
Future Plan (1)

- **Manage I/O device's power from Guest software**

- Direct management with PMCSR (Power Management Control/Status Register) of Device
- Support D0, D1, D2, and D3hot



Save power of I/O device.



- **Reduce interrupt redirection**

- On Xen 3.3.0, an interrupt is delivered to one processor (vCPU#0)
 - If that processor is not the one specified by guest software, then interrupt is redirected
- Support interrupt delivery to processor specified by guest software in order to reduce redirection



Improve performance of interrupt delivery.

Xen Summit Tokyo 2008

Future Plan (2)

- **Improve ioemu log**

- Attach date and pid information to the log of ioemu

example

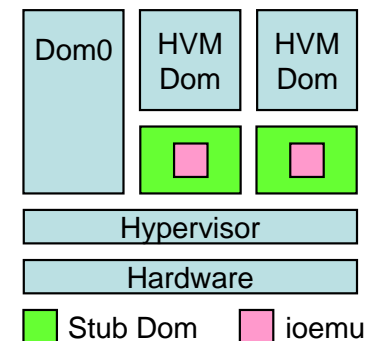
```
[2008-11-21 16:10:05 4567] can't store dev vc name for domid 1 in /parallel/0 from a stub domain  
[2008-11-21 16:10:06 4567] qemu_map_cache_init nr_buckets = 10000 size 3145728  
[2008-11-21 16:10:06 4567] shared page at pfn 1ffe
```



Make it easy to analyze the log when any trouble happens in a system.

- **Support I/O pass-through using stub domain**

- Stub domain emulates I/O instead of domain0
- Improve stub domain to enable I/O pass-through



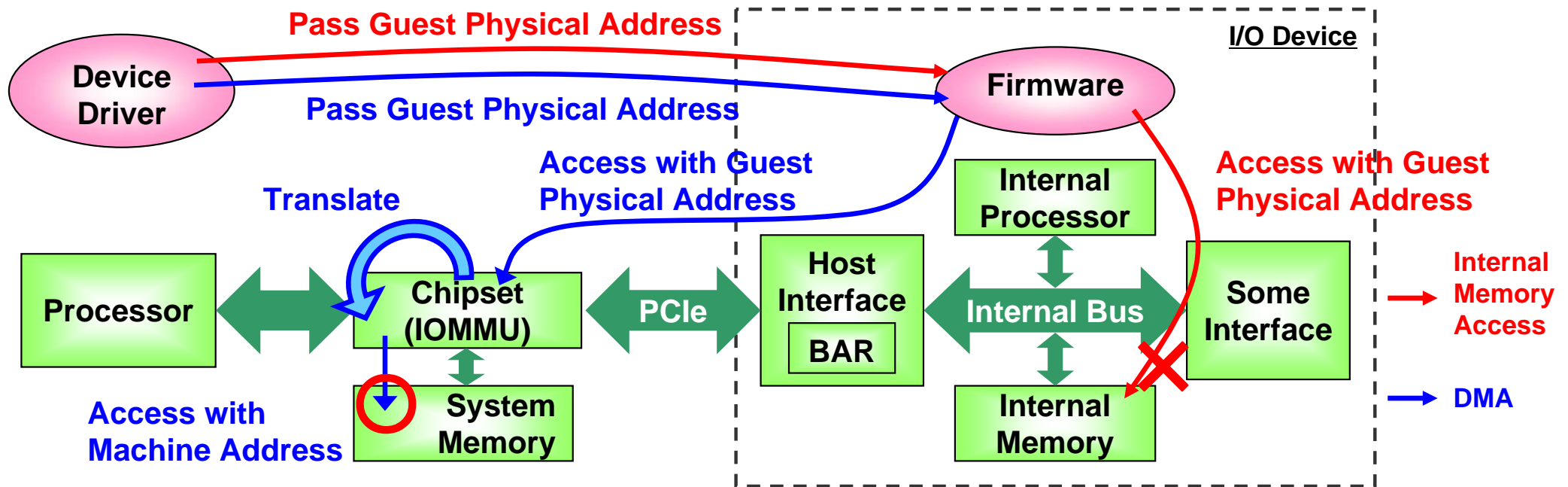
Improve the scalability and make domain0 smaller.

Xen Summit Tokyo 2008

3. Challenges for I/O Pass-through

Issue 1 – Unsuitable Device for I/O Pass-through

- I/O device which accesses internal memory with physical address obtained from device driver does not work



- When we use I/O pass-through, physical address used by device driver is “guest physical address”. It is distinguished from machine address.
- On DMA, IOMMU translates guest physical address to machine address.
- On internal memory access, IOMMU does not translate addresses. It causes access failure.

Xen Summit Tokyo 2008

Issue 1 – Virtualization Hole Causing the Issue.

- **No one translates guest physical address to machine address if I/O device uses guest physical address obtained from device driver for the purpose other than DMA**
- **This is common problem among IOMMU based Hypervisors**



We expect the adapter vendor to design I/O device suitable for I/O pass-through.

Issue 2 – ioemu does not Support PCIe

- **ioemu does not have following functions**

- **MMCFG mechanism**

- Accessing configuration registers by accessing memory space
- Needed to access configuration register at offset from 100h to FFFh

- **Capability Structure from 100h to FFFh**

- e.g. Device Serial Number Capability Structure
- PCI Express VSEC Structure

- **Root Port**

- Device with registers to control PCIe-specific function

- **Advanced Error Reporting (AER)**

- Reporting PCIe error to software

- **PCIe Hot-plug**

- It is newer mechanism than ACPI hot-plug

As a result...

Guest software can't access register or capability structure it needs

Topology at guest software's view is not PCIe machine

Guest software can't recover PCIe error

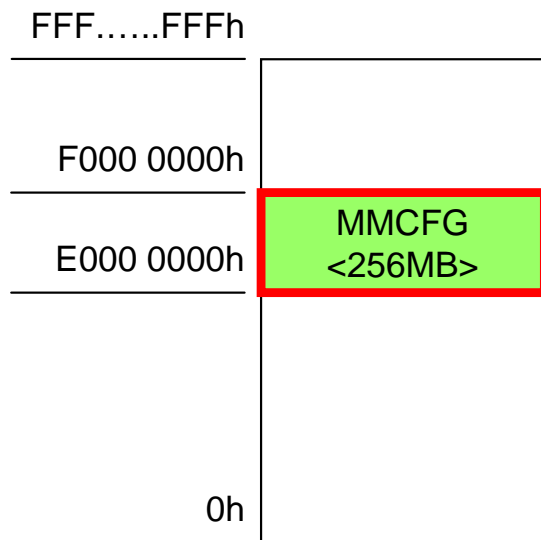
Guest software has to use ACPI hot-plug

Xen Summit Tokyo 2008

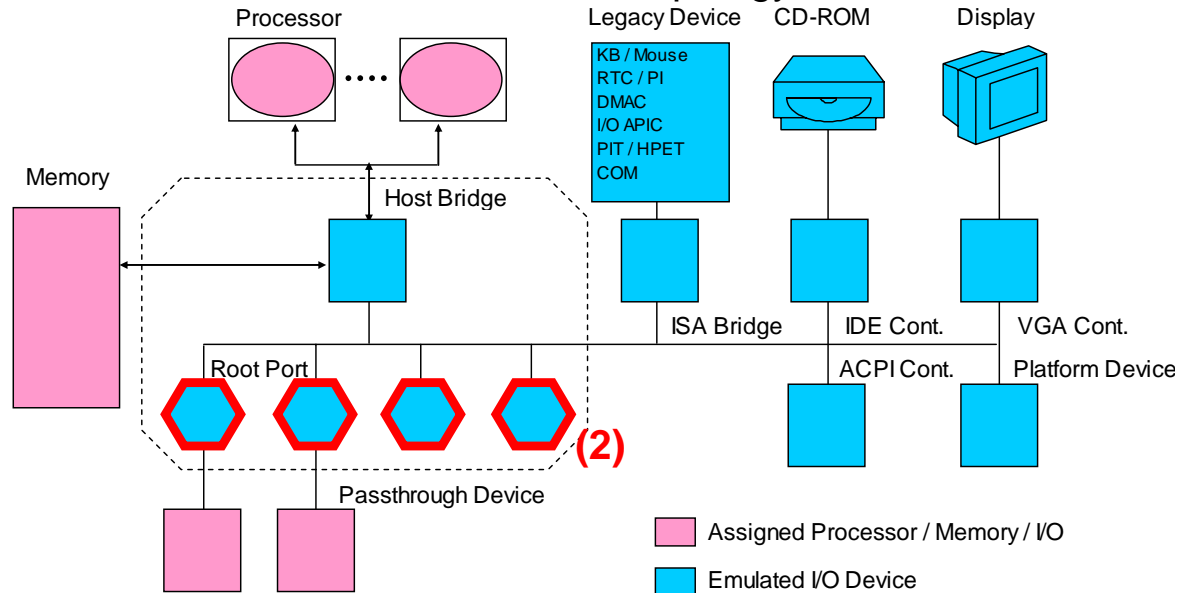
Issue 2 – Development to Support PCIe

- First, add MMCFG mechanism(1) and Root Port emulator(2) into ioemu
- Then, implement others

Guest Physical Address Space



Virtual Machine Topology



It is difficult for one developer or a single team to do all the implementations. We need cooperation from all of you!

Xen Summit Tokyo 2008

4. Conclusion

Conclusion

- **I/O pass-through becomes practicable now**
- **We will continue developing the functions related to I/O pass-through**
- **There are still some issues for I/O pass-through**
- **We need many developers cooperation for progress**

Acknowledgements

- **Thank you very much to all of you:**
 - **All the people concerned with Xen Summit Tokyo 2008**
 - **All the developers in Xen Community**

Empowered by Innovation

NEC