Virtualization BOF

Isaku Yamahata <yamahata@valinux.co.jp>

Japan Linux Symposium October 23, 2009

Agenda

- Introduction
- New chipset emulator in qemu
- Other desired features
- QEMU
- (Any other virtuliazation topics?)

Introduction

- The active development in virtualization has shifted from cpu itself to surrouding area.
  - Eg, IO, guest firmware, FT/HT
- Virtualization target has diverged from server to Desktop/Client/Embedded
  - So do Required features.

Introduction(cont.)

- QEMU is the key component of IO emulation
  - QEMU is heavily utilized by virtulalization projects as device emulator.
  - Topics are mainly on IO device emulation/guest firmware.
- Other virtualization related topics are also welcome.

New chipset emulator for new hardware feature

Background

- Current Qemu emulates
  - For Pentium Pro/II/III
  - North bridge: I440FX
  - South bridge: PIIX3 (and PIIX4 for acpi power management and pci hot plug)
  - Hardware release date: May 1996
- Too old compared to new real hardware features

Motivation

- PCI
  - Qemu only support part of PCI specs. e.g.64bit BAR
  - More buses/slots
    - Qemu only support host buses (for PC emulation)
    - 3+ pci bus(96+ slots)/96+ pcie slots
    - Brige emulation: filtering
- PCI express isn't supported
  - PCI express has more advanced features

Motivation(cont.)

- Native direct attachment of PCI express device to guest OS
  - Currently it can be attached as PCI device
    - Xen calls it pci passthrough
    - Kvm calls it device assignment
  - PCI express has more features
- Need to fill the gap between newer real hardware feature and qemu emulation mainly in PCI related area.

Challenges

- Chiken and egg
  - qemu emulated devices doesn't use those new features
  - The new features haven't been provided, qemu emulated devices won't use them
- Testing
  - Testing those features without real user
  - Variety of target. PCI is used by many targets
  - PCI express direct attach is a way for test, but their qemu aren't based on very unstable version

Why PCI express?

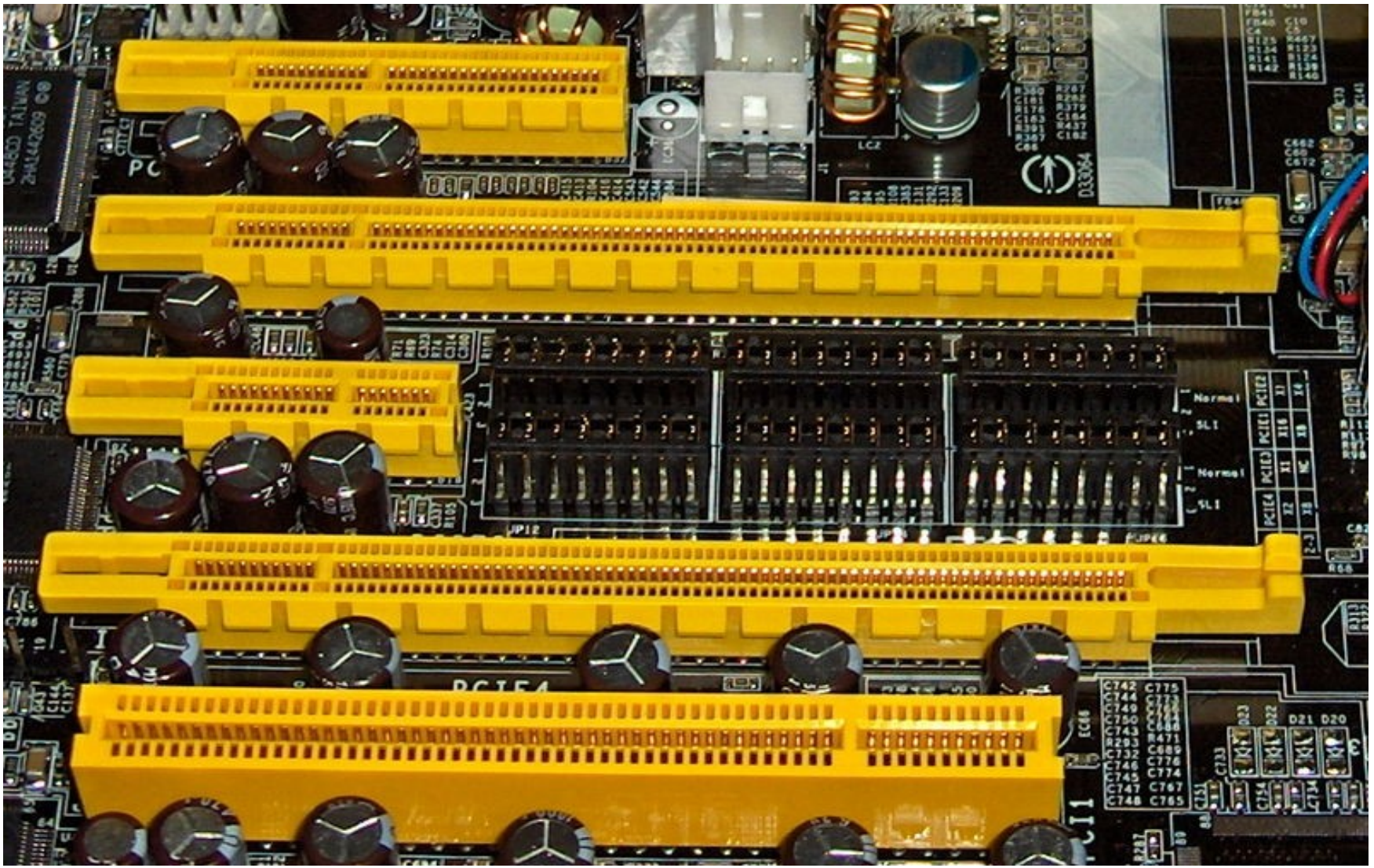
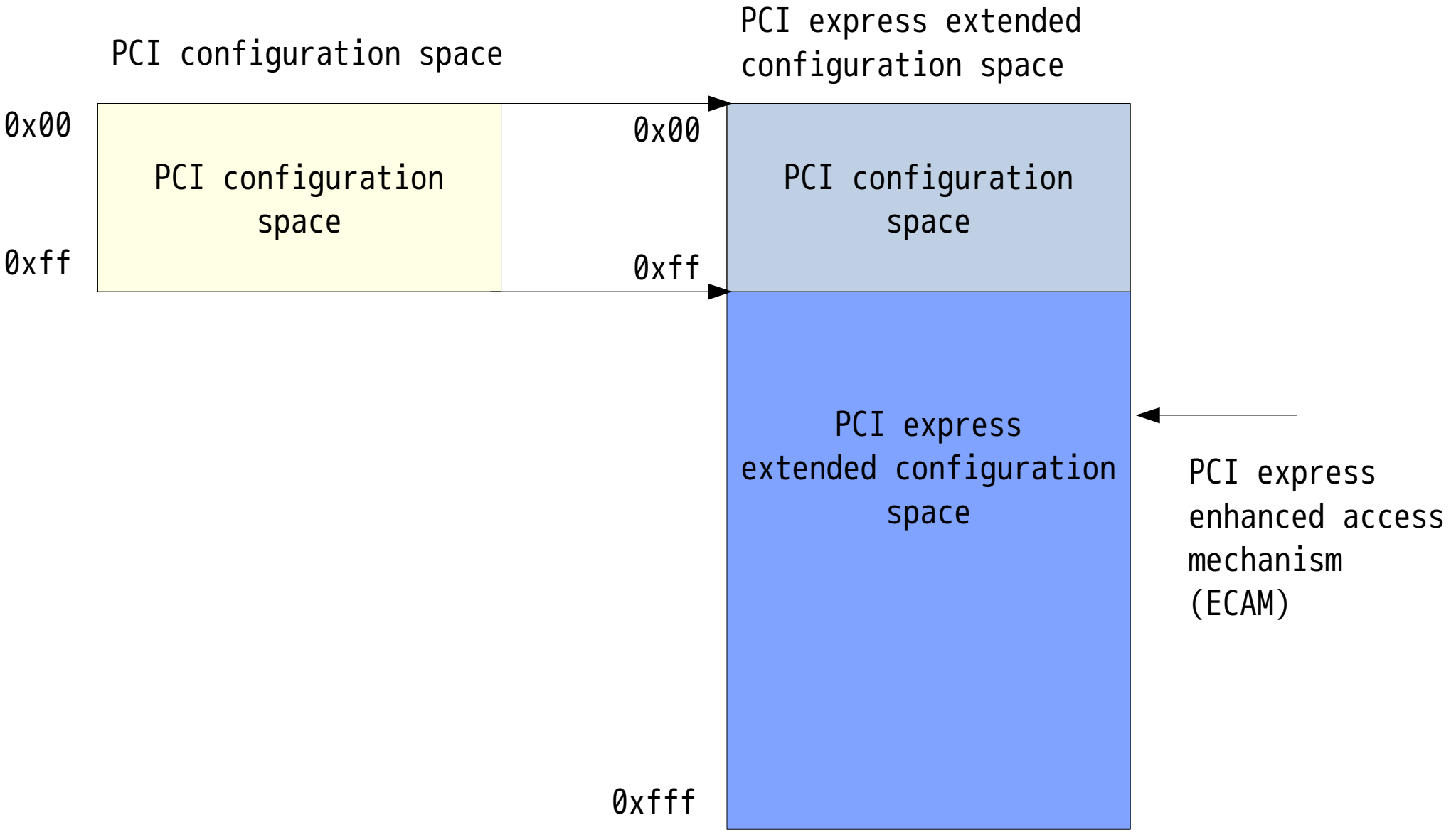


Photo:Cited from wikipedia

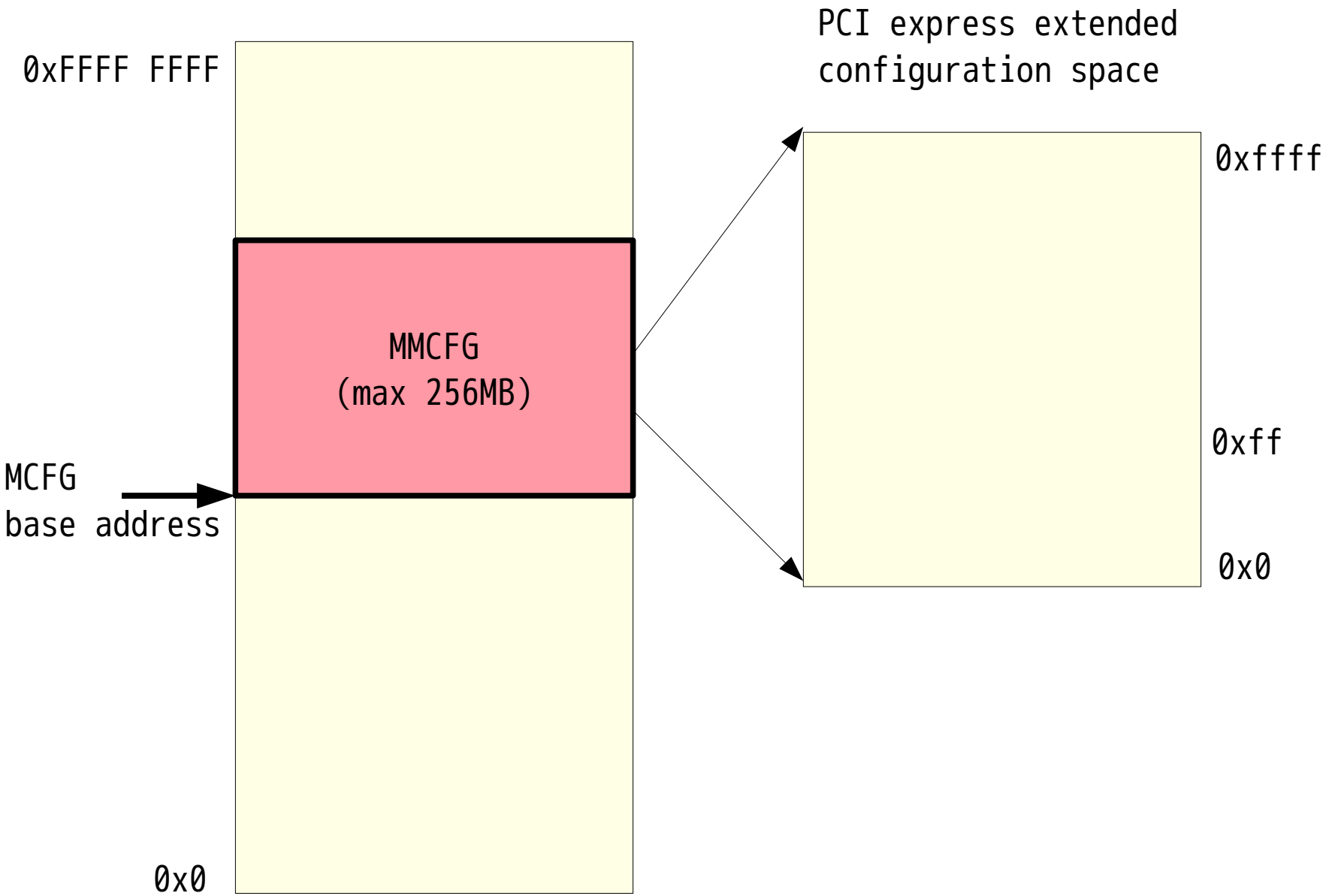
Features from software point of view

- MMCONFIG (>0xff configuration space)
- PCI express native hot plug (not ACPI based)
- AER(Advanced Error Reporting)
- ARI(Alternative Routing ID)
- PCI express native power management

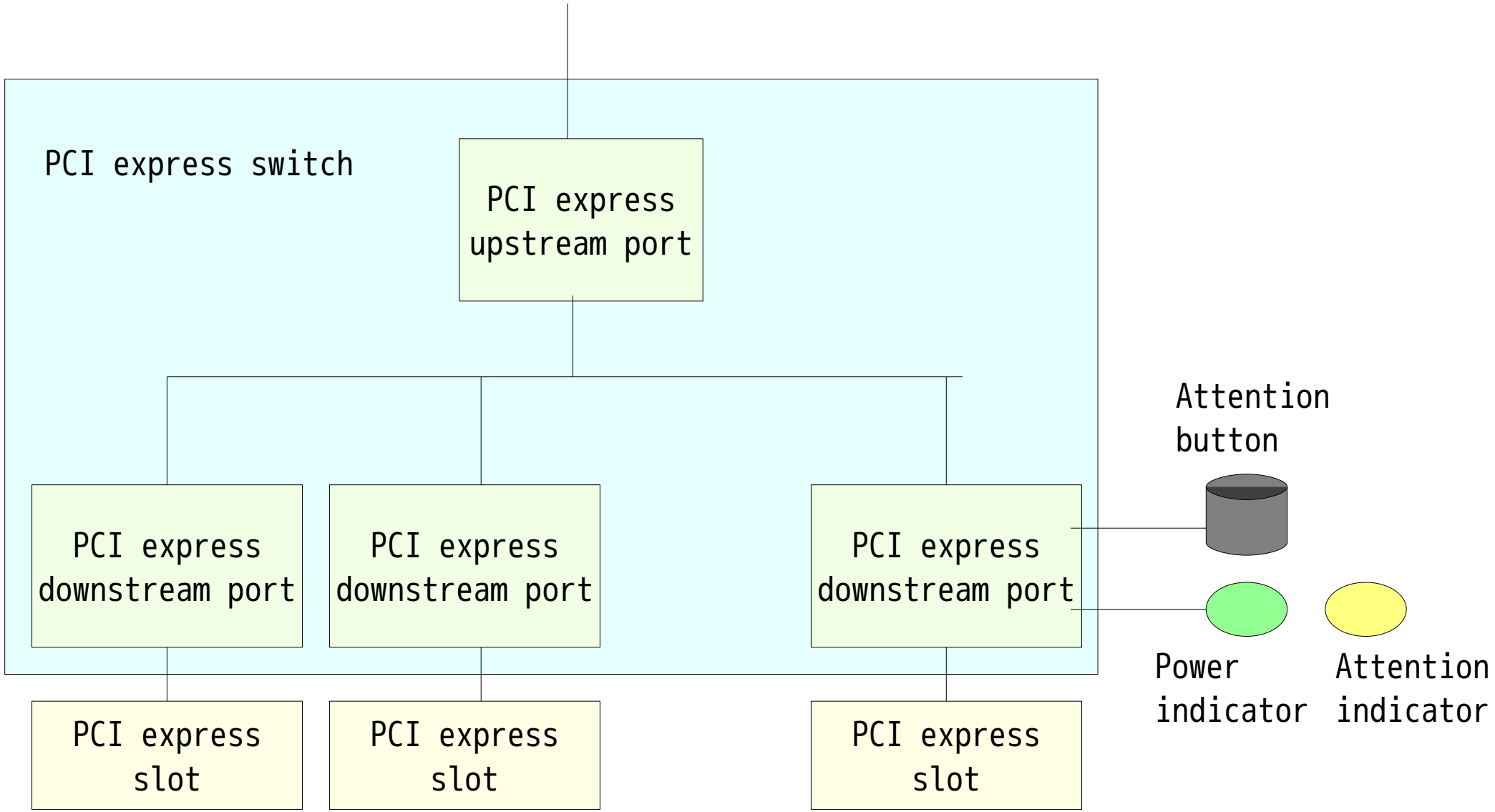
PCI express extended configuration space



PCIe MMCONFIG



PCI express hot plug



New chipset emulator

- Q35 chipset based
  - For Core2 Duo
  - North bridge: mch
  - South bridge: ich9
  - Release date: Sep 2007
  - In fact I have chosen Q35 because I have it available at hand.
    - Newer chipsets(gmch/ioh, ich10) have mostly same feature from the point of emulation except graphics.
  - Now it lacks iommu/graphics emulation so it should be called P45?

New chipset emulator(cont.)

- Now the followings are working
  - 64bit BAR
  - PCI express MMCONFIG
  - BIOS updates(MCFG, e820)
  - Linux boots happily using MMCONFIG
    - Haven't tested other OSes.

BIOS

- ACPI MCFG to specify MMCONFIG area
- E820 update
  - Make e820 code 64bit aware.
    - So far it filled higher bits with zero.
  - Linux requires MCFG area is covered by e820 reserved area
  - Otherwise Linux thinks that it's bios bug and avoids to use MMCONFIG.

BIOS(cont.)

- PCI initialization
  - Teach bios new chipset
  - PCI IO/memory area assignment for multi pci bus.

ACPI

- ACPI tables update
  - FADT
  - MCFG
  - DSDT
    - PCI express(PNP0A08)
    - PCI routing table

Future work:PCI express

- PCI express hot plug will be provided as pcie switch emulator (not integrated into chipset)
  - Many (96+) port wanted
- ARI(alternative routing ID)

Future work:PCI express(cont.)

- PCI express native direct attach.
  - PCI express specific configuration registers should be virtualized
    - Device serial number cap, VSEC...
  - AER(Advanced Error Report): passing errors to guest OS
  - Power management
- Multi PCI domain?
  - More slots

Future work:BIOS

- pcbios(bochs bios) vs seabios
  - Pcbios is from bochs.
  - Seabios is more clean and featured.
- Qemu switches from pcbios to seabios
  - Now qemu uses pcbios so that patches for pcbios has been created.
  - Qemu 0.12.0 release will use seabios instead of pcbios.
  - So patches for seabios is necessary for merging.

Future work:ACPI

- Code change is small, however acpi table change would be large.
  - Have two tables (more in future?), and switch it dynamically?
  - pass tables outside qemu, say, by command line option.
    - Requires interface between qemu and bios
    - fw_cfg
  - Dynamically generating acpi code?
    - COREBOOT does.

Future work: Direct attach in qemu?

- Does qemu want the feature?
  - Hopefully consolidate xen and kvm code.

Summary: current status

PCI express

PCIe MMCONFIG	Working Under heavy review for merge.
Q35 chipset	working. Waiting for pcie MMCONFI
PCIe portemulator	WIP
pcie native hotplug	WIP
pcie passthrough	WIP
3+ pci bus	working

pcbios

mcfg	working
e820	working
host bridge initiazatlin	working
pci io/memory space initialization	working
switching acpi table or passing acpi table outside qemu	WIP

Other desired features?

Other hot plug

- SATA/eSATA hotplug?
  - AHCI

Other feature: IOMMU

- IOMMU: Intel VT-d, AMD IOMMU
- Usage model?
  - Guest OS wants IOMMU?
- IOMMU emulator in qemu
  - Implementation will be interesting.
- Shadow paging of IOMMU for guest OS
  - At the moment DMA fault and restart isn't possible due to PCI specification.

Other feature: graphics

- Integrated Graphics of gmch
  - Anyway GPU support is highly wanted.
- GPU passthrough

Other feature: APICs

- IOAPIC
  - IOAPIC is performance critical, so IOAPIC emulation is done in kernel/hypervisor.
  - Does it make sense to address IOAPIC in qemu?
  - More than 24 pins
  - Multi IOAPIC

Other feature: firmware

- gPXE
  - More device support
    - igb, igbvf
- Guest UEFI
  - Tristan Gingold created guest UEFI using edk2.tianocore.org

HT/FT

- Kemari

PCI DMA fault/restart?

- DMA fault/restart?

QEMU

Long term(?) qemu feature

- Nested VT-x
  - AMD SVM is there
- Threading QEMU
  - For guest SMP
  - For scalability
- Machine config file
  - Allow more flexible machine
    - For more complex pci bus topology
    - Device tree

Thank you

backup

Qemu next release

- Qemu 0.11.0 released
- Now planning for 0.12.0
  - Anthony Liguori thinks aiming for early to mid December
  - three month cycle

Planned features

- Qdev
- Vmstate
- seabios switch
- gPXE switch
- KVM
  - In-kernel APIC
  - guset SMP
- Multiport virtio-console

Planned features

- Machine monitor protocol
  - Robust UI for human and machine
  - QObject
- NEC PC-9821