Status update on QEMU PCI Express Support

Isaku Yamahata <yamahata@private.email.ne.jp>
<yamahata@valinux.co.jp>
VA Linux Systems Japan K.K.
LinuxConJapan 2011: June 3rd, 2011

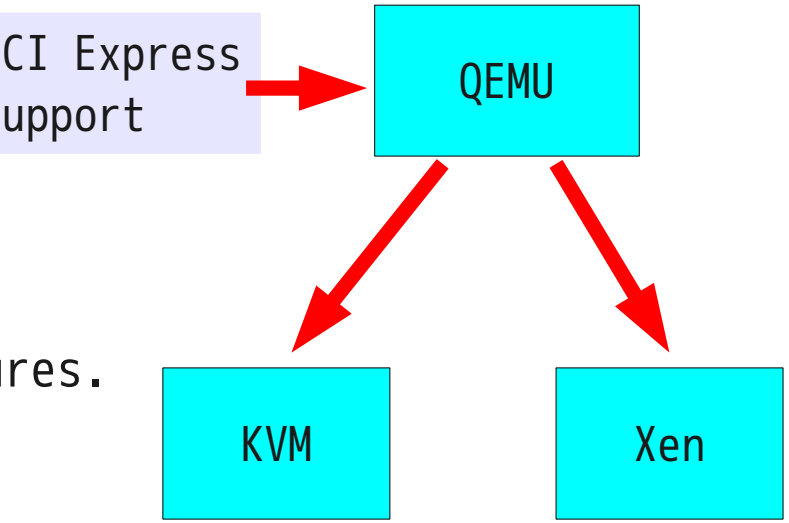
Agenda

- Introduction
- status update
- demo
- Future work
- Summary

Introduction

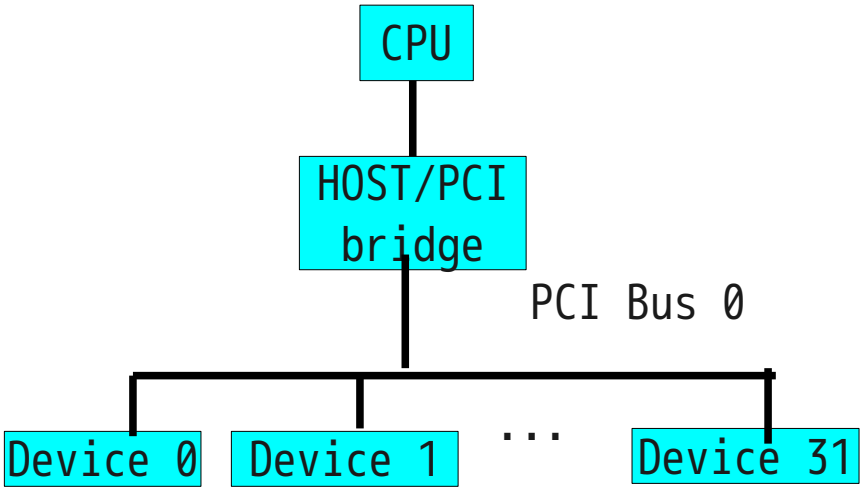
Background

- QEmu is used for device emulator for many virtualization technologies. KVM, Xen...
- QEmu supports PCI in a limited way, and doesn't support PCI Express.
- So do QEmu derivatives(KVM, Xen...).
- Enhance QEmu PCI layer and add PCI express
  - Fill those gaps
  - to enable KVM, Xen, ... to utilize those features.
  - Users always wants new hardware features...



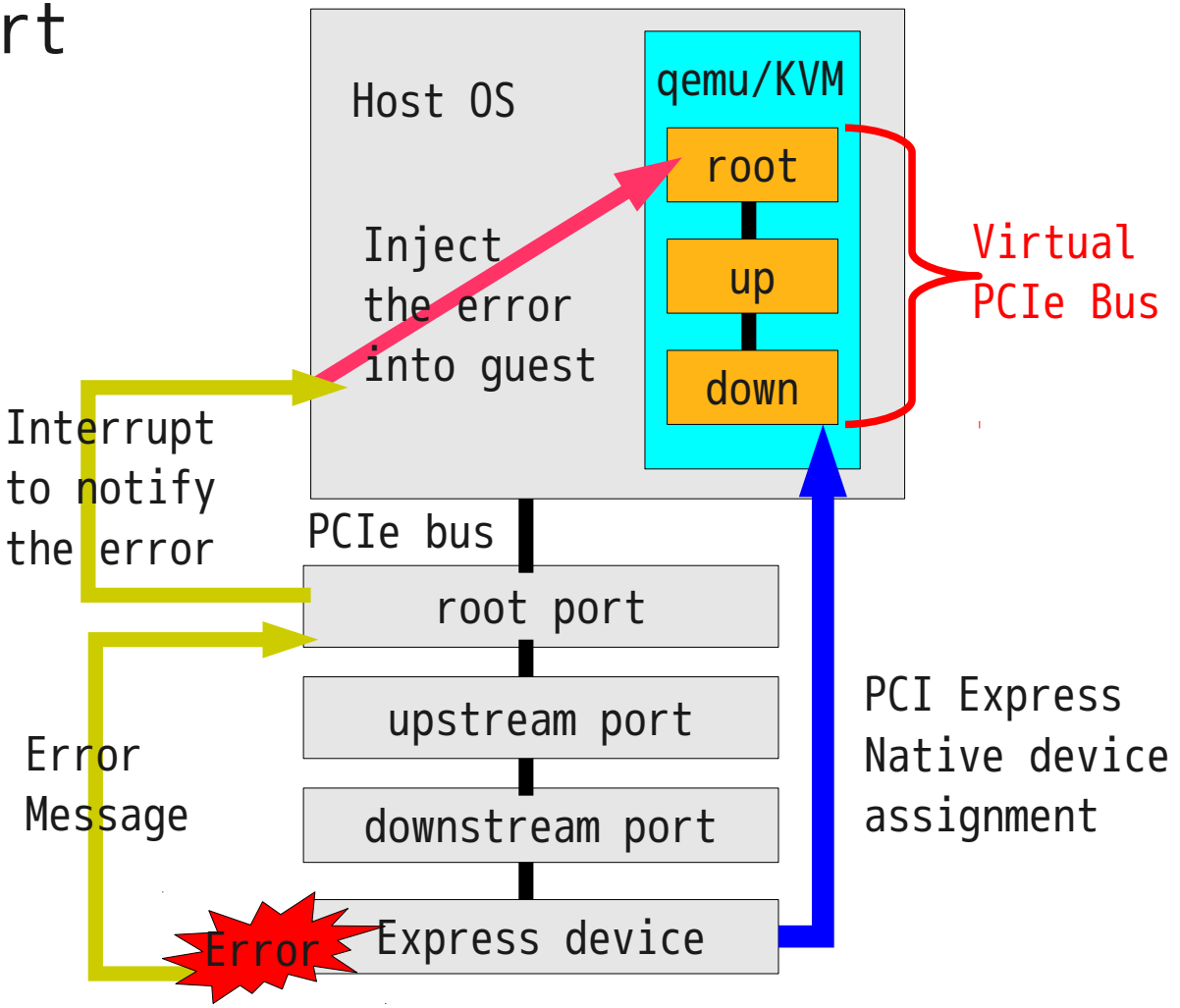
Goal in PCI area

- more features
  - 64bit BAR
  - Multifunction
  - Multi pci bus/segment
  - …
- clean up: remove many limitations
  - Only up to 32 slots
    - Some people wants several hundreds of hot-pluggable slots
  - ACPI-based hot plug is nasty/broken
    - Guest OS doesn't always corresponds in time

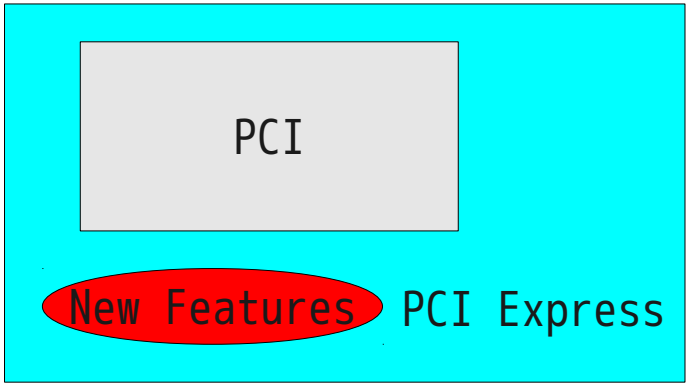


Goal in PCI Express area

- Enable QEmu to support PCI Express
- Enable PCI Express native device assignment with
  - Native hot plug
  - AER(error reporting)
- Then, bring Express support to qemu derivatives.



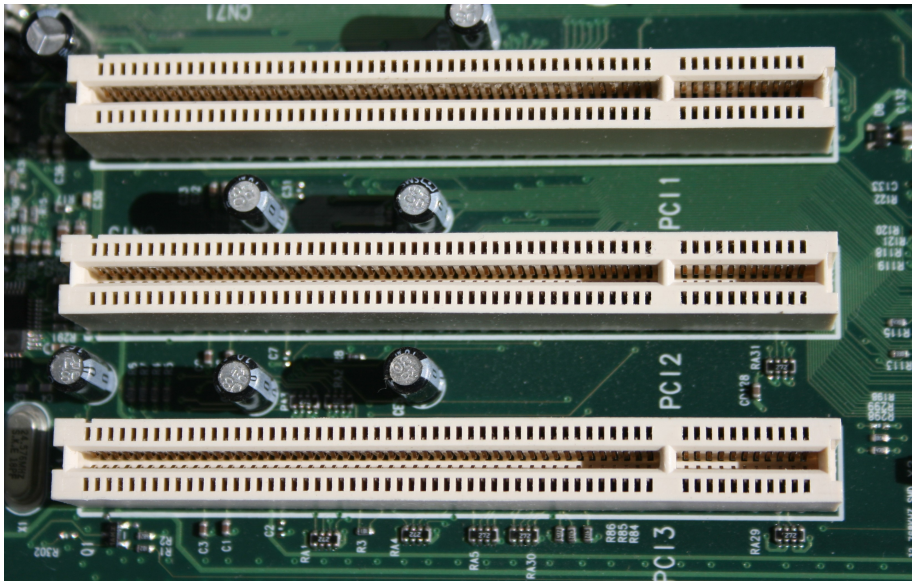
Why PCI Express?
Isn't it compatible with PCI?



- Yes it's upper compatible, but...
- many new native express features
  - They can be only used via express feature.
- Some devices(drivers) require native express
  - devices really use native express features
  - Its driver checks if the device is really express-enabled
  - Existing conventional PCI device assignment doesn't work
- Hardware certification requires express
- Developing new hardware: qemu is used for emulation when developing new hardware.

What's PCI?

- Peripheral Component Interconnect
- Year created:1992
- Parallel bus
- Has been widely adopted in the market



From Wikipedia

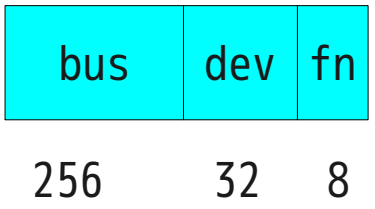
PCI features from software point of view

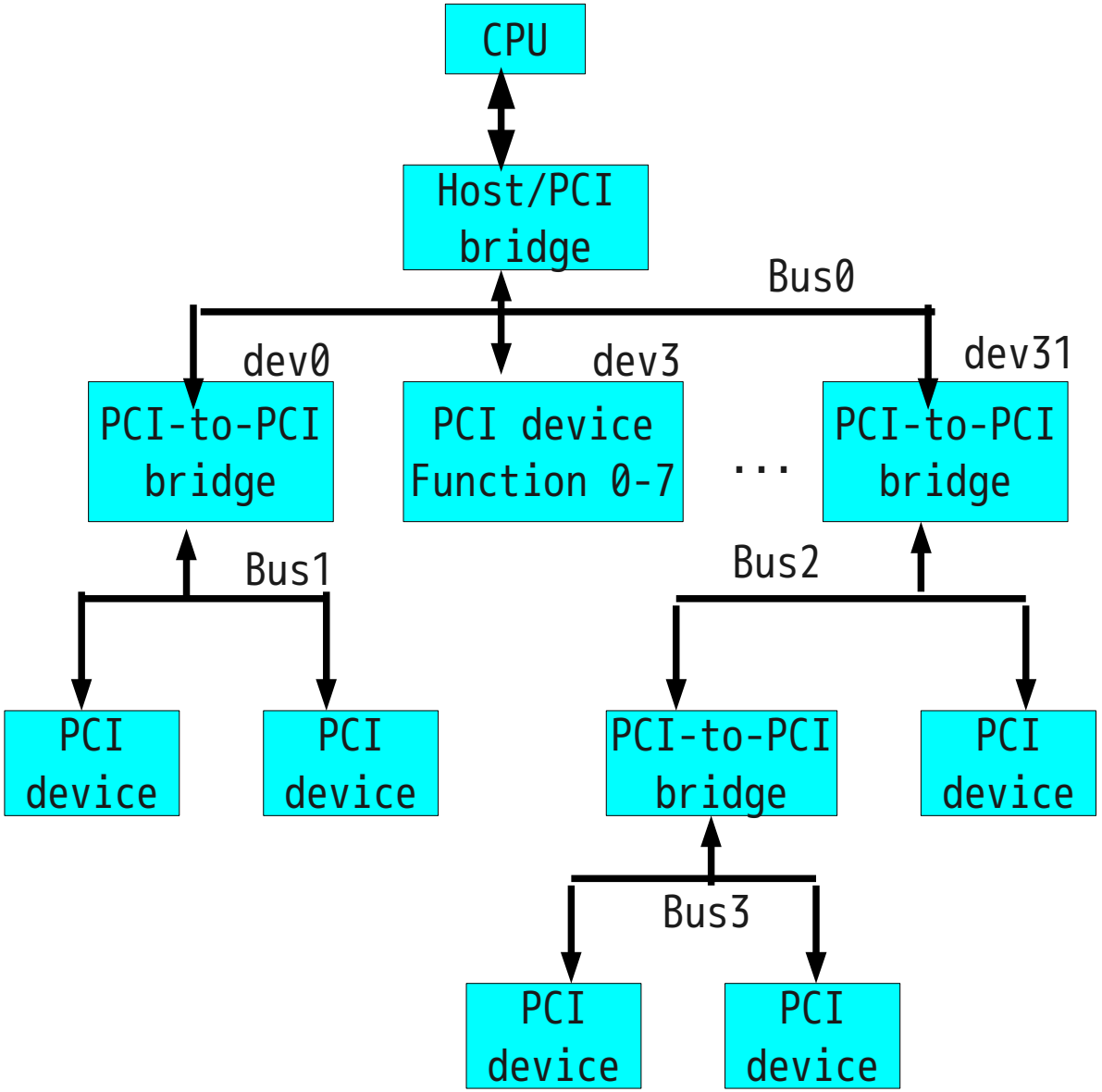
- Bus topology(bus addressing)
- 3 addressing spaces
- BAR(Base Address Register)
- interrupts

From wikipedia

PCI bus topology

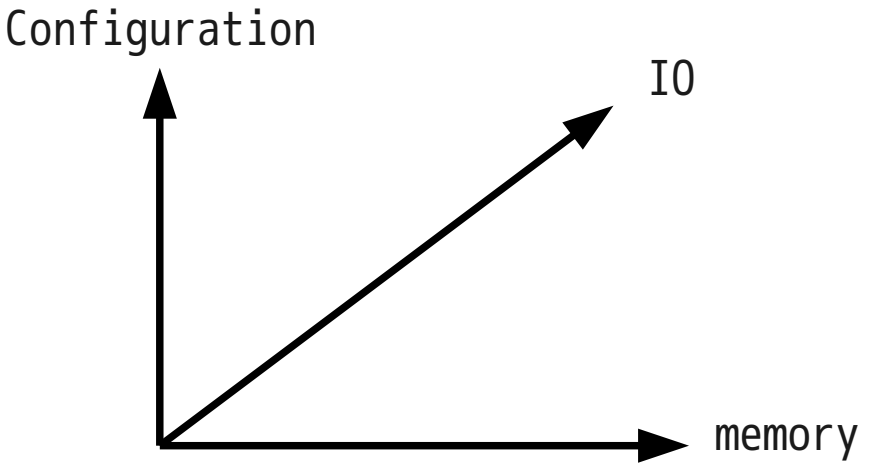
- Bus addressing
  - Bus numbering
- Bus, device, funciton





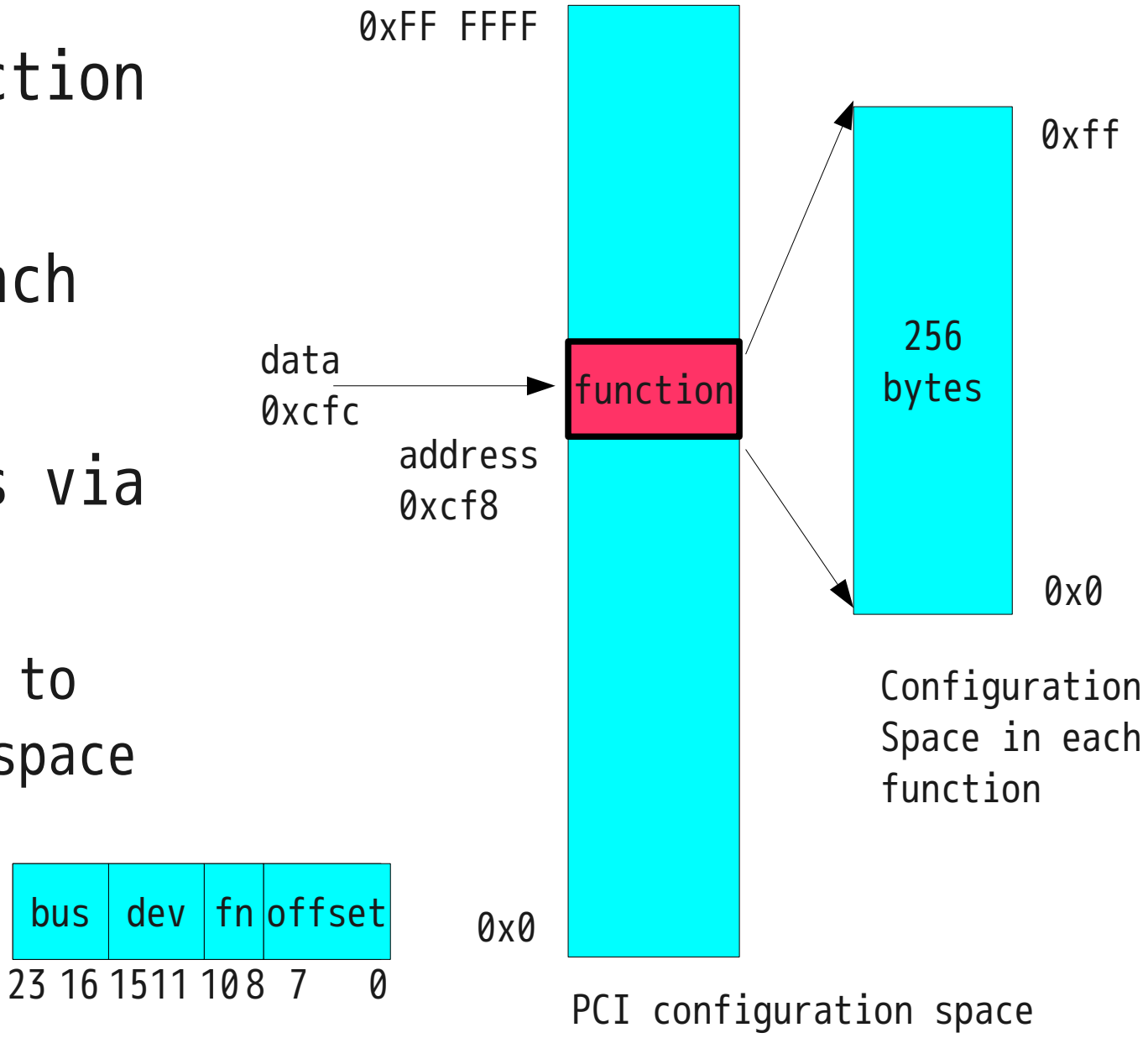
addressing spaces

- Memory: accessed via MMIO
  - Prefechable vs non-prefetchable
- IO: accessed via IOIO
- Configuration space



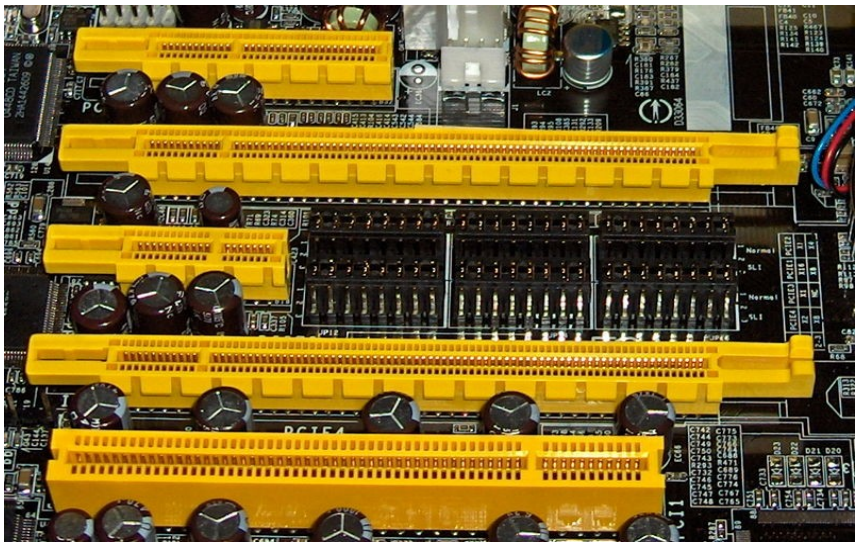
PCI configuration space

- Bus,device,function + offset
- 256 bytes on each function
- Indirect access via IO port
  - 0xcf8: address to configuration space
  - 0xcfc: data



What's PCI Express?

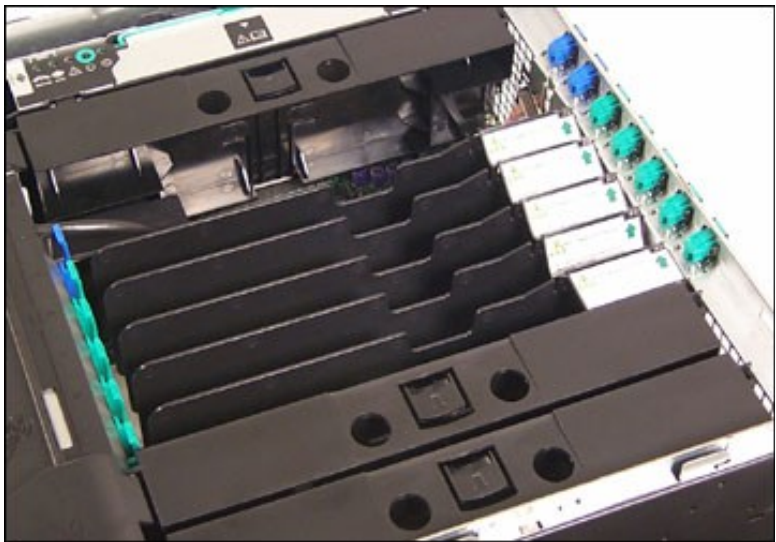
- Designed as a successor of PCI
  - Software compatible with PCI
  - Many improvements
  - Widely accepted in the market
  - Has been superseding PCI
- Year created: 2004
- Serial bus



From Wikipedia

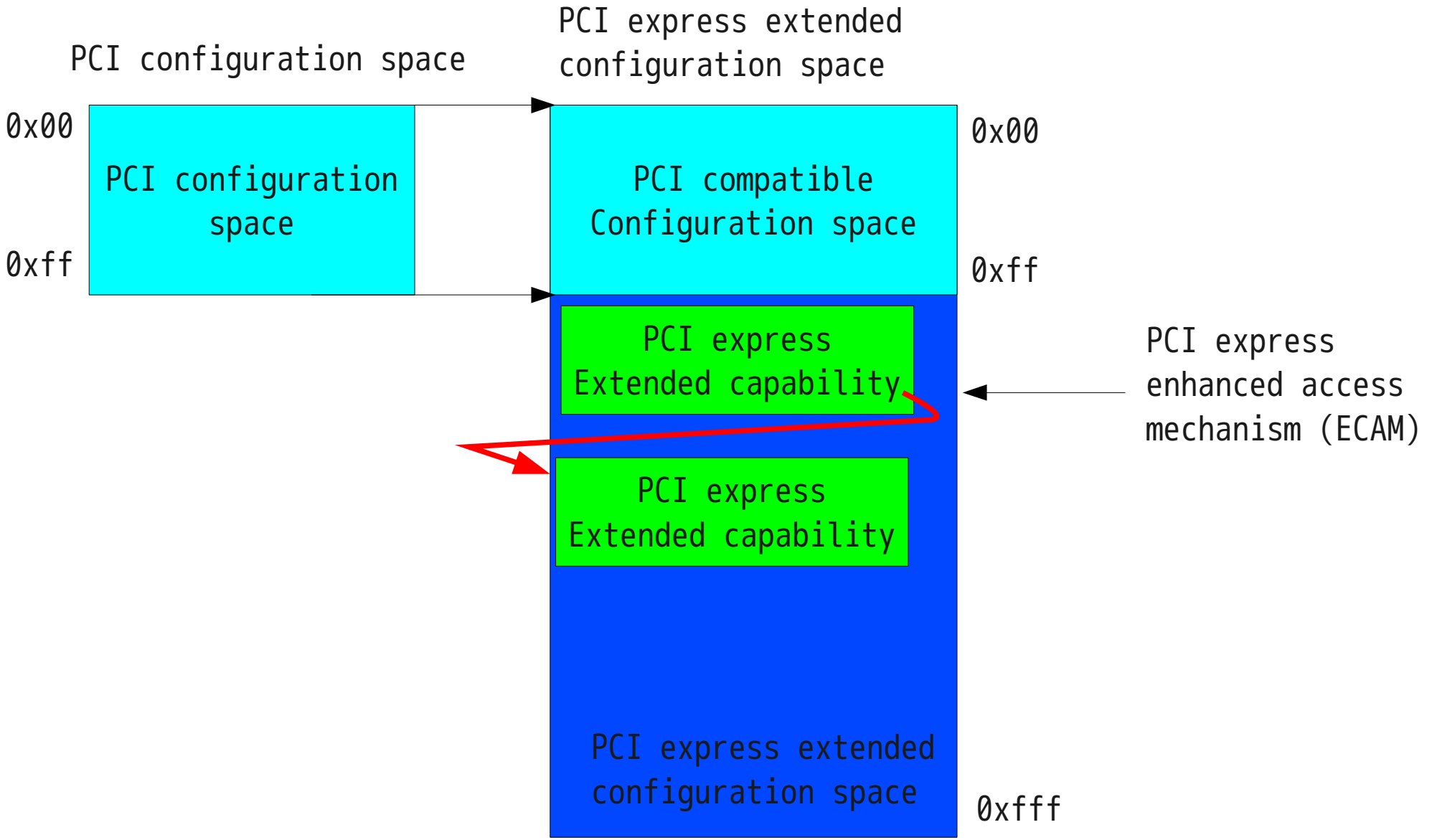
Express features from software point of view

- Many enhancements from PCI, for example
  - Extended configuration space
  - MMCONFIG: larger configuration space
  - Native hotplug:not ACPI based
  - Native power management
  - AER(Advanced Error Reporting)
  - ARI(Alternative Routing ID)
  - VC(Virtual Channel)
  - FLR(Function Level Reset)

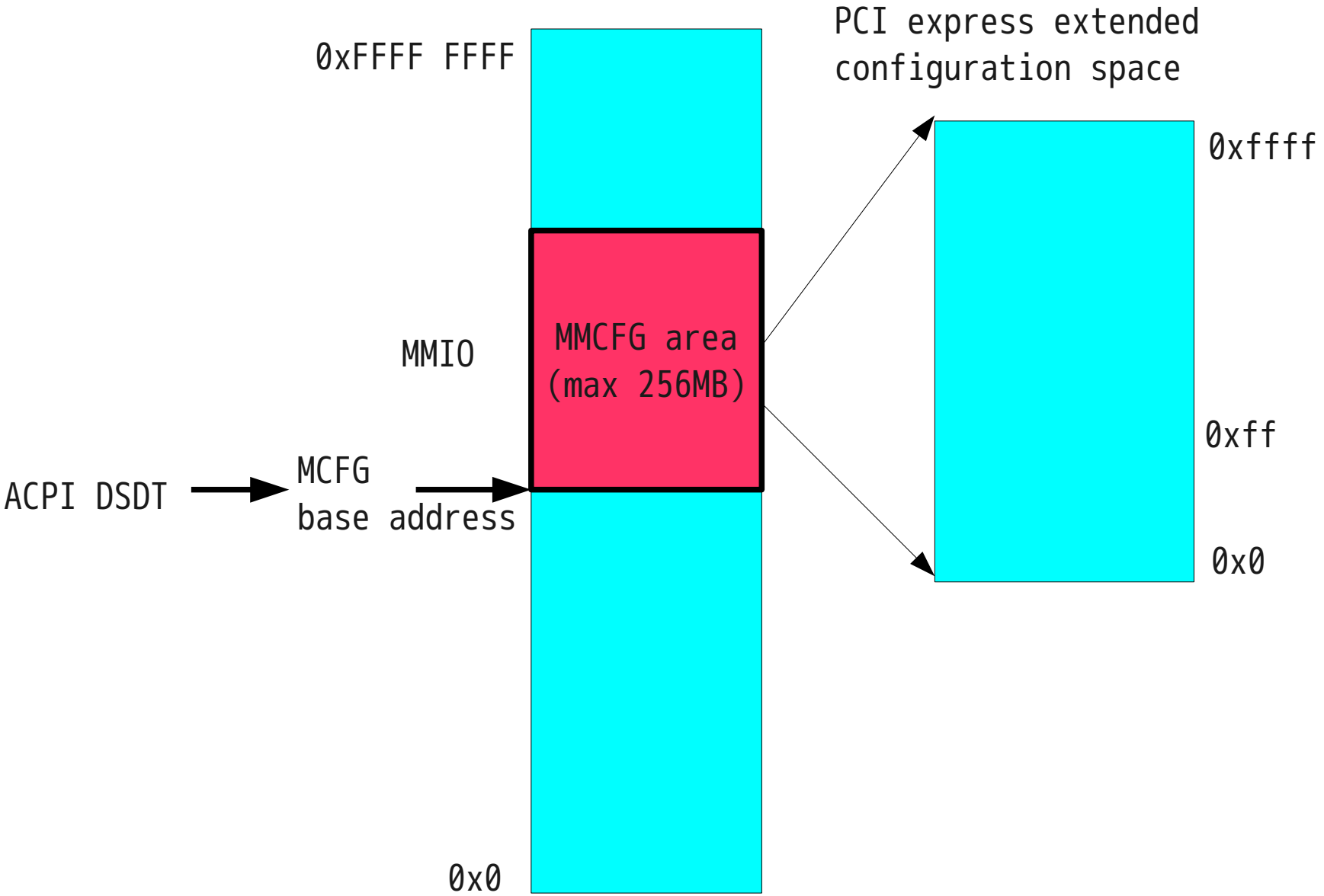


From http://cdnsupport.gateway.com/s/Servers/9715Server/54.jpg

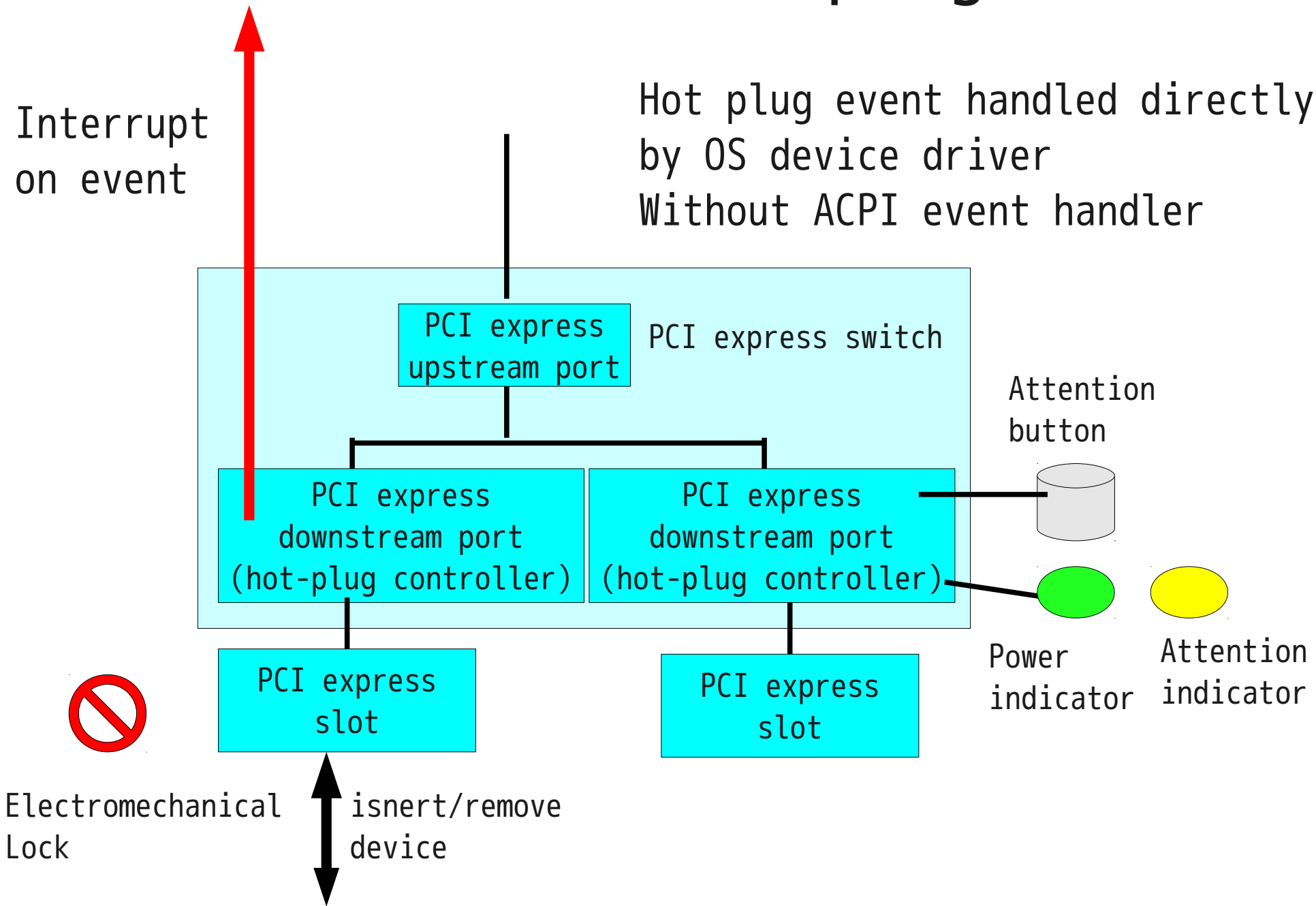
PCI express extended configuration space



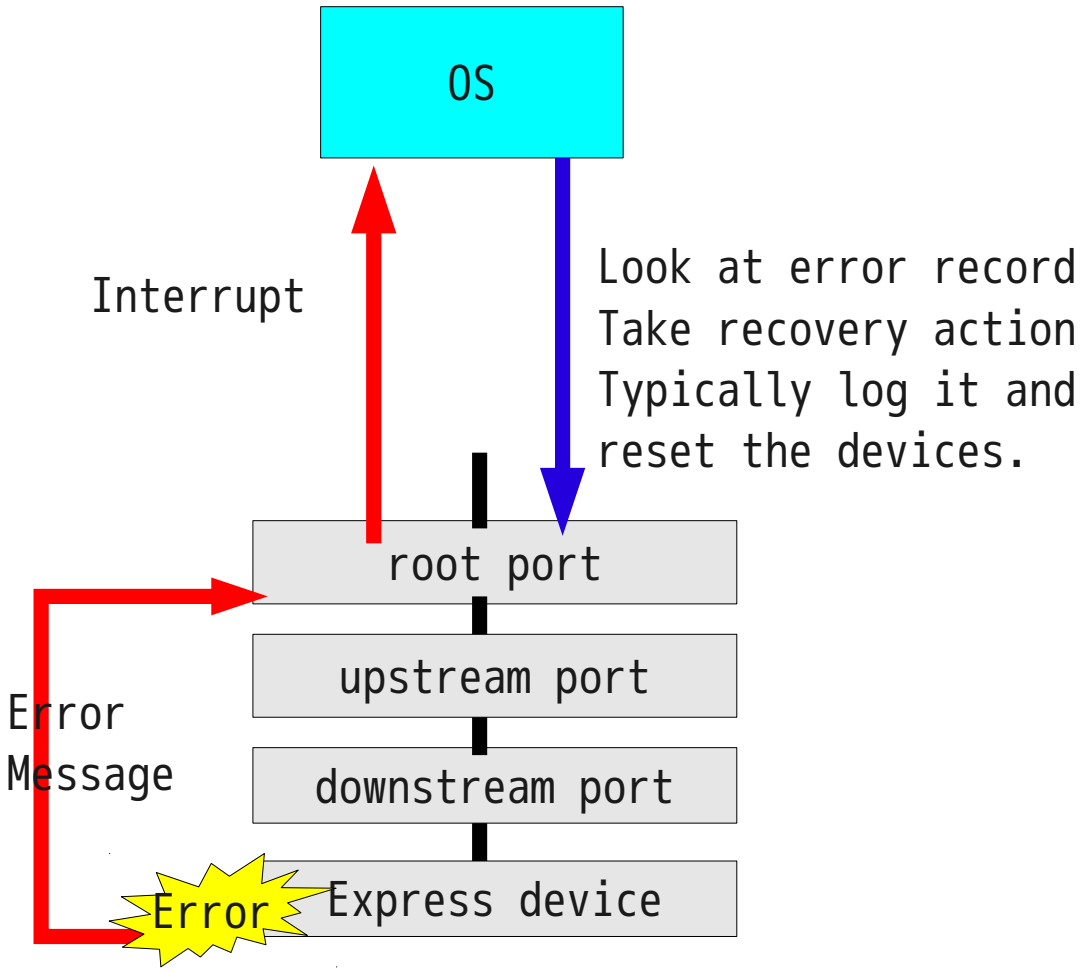
PCIe MMCONFIG



Native hot plug



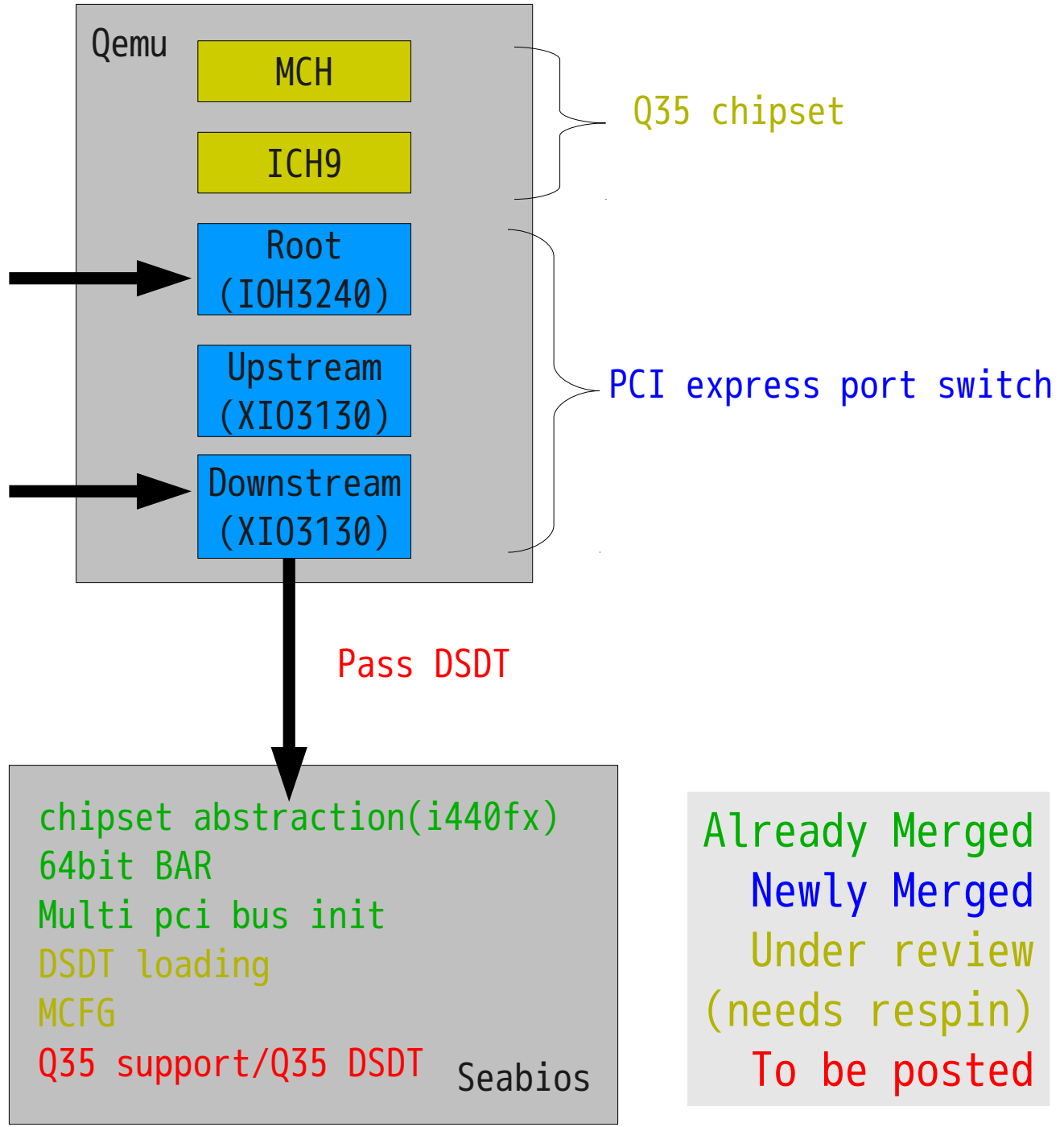
Advanced Error Reporting(AER)



- Standardized error reporting.
- Important for RAS

status update and implementation

I440fx chipset refactoring
64bit BAR
Extended config space
MMConfig
PCI-to-PCI bridge clean up
PCI bus reset



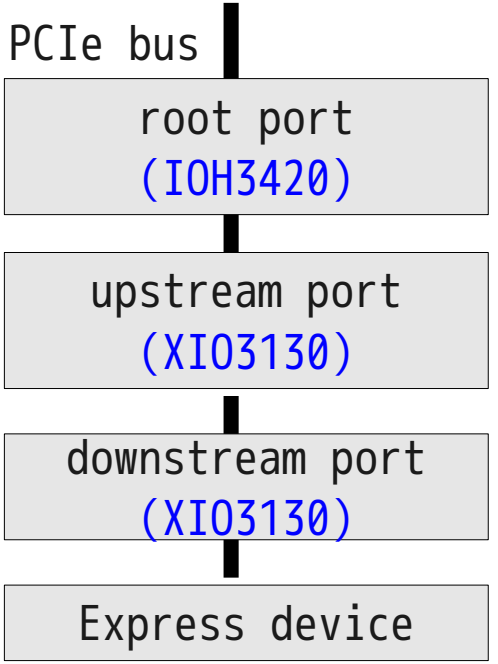
Hot plug function

Function	Supported?
Attention Button	yes
Power Controller	No
MRL Sensor	No
Attention Indicator	Yes
Power Indicator	Yes
Hot-Plug Surprise	Yes
EMI	Yes

Already Merged
Newly Merged
Under review
(needs respin)
To be posted

PCI Express port emulator

- Virtual PCI bridge
- Root/upstream/downstream port
  - All of three ports are needed.
  - Necessary for native hot plug, AER.
  - Native hotplug
  - AER
- IOH3420 and XIO3130 are chosen
  - Because there are few datasheet publicly available.



New chipset emulator

- Q35 chipset based
  - For Core2 Duo
  - North bridge: mch
  - South bridge: ich9
  - Release date: Sep 2007

From wikipedia

- In fact I have chosen Q35 because I have it available at hand.
  - Newer chipsets(ioh, ich10/pch) have mostly same feature from the point of view of emulation except graphics.

Q35 chipset emulator doesn't have

- IOMMU(VT-d) emulation
  - IOMMU support itself is a big topic
  - IOMMU emulation is coming by others
    - Only for emulated devices,
    - Not for direct assigned devices.
- Integrated graphic emulation
  - GPU itself is also a big topic and
  - many other people has been worked on

DEMO

Future work

Future work: PCI

- IRQ routing improvement
- Qdev id auto assignment
  - for pcie_aer_inject_error
- Hot plug
  - Improve pci hot plug framework
  - Multifunction hot plug
    - At this moment, qemu pci layer doesn't have the notion of pci lost
- PCI multi segment: For more slots
- Device-assignment: code consolidation, VFIO
- PCI BAR allocation
  - Currently new qemu RAM API is being discussed
- Listing supported pci device more user-friendly

Future Work: PCI Express

- PCI express native device assignment
  - Enhance VFIO for pcie
  - PCI express specific configuration registers should be virtualized
    - Device serial number cap, VSEC...
    - Native Power management
    - VC(Virtual channel)
  - AER(Advanced Error Report)
    - Catch the error in host.
      - Currently Linux AER port driver does only printk().
      - Poll errors from targeted devices.
    - inject errors from host to guest OS for RAS.
  - Assigning bus hierarchy tree
- QMP support
  - Hotplug LED indicator event(LED on/off/blink)

Future work: SeaBIOS

- MCFG
- Allowing Custom DSDT table
  - coreboot v.s. fw_cfg
  - Xen also wants this feature
- Smarter PCI BAR assignment
  - Gerd is tackling to this: RfC patch
    - Memory v.s. prefetchable memory
    - Reasonable error handling
    - 64bit BAR

Future work: qemu derivative

- KVM support
  - Jan Kiszka has reported some achievement
    - Boot with Kvm, in-kernel irq-controller(PIC, IOAPIC)
    - Device assignment
- Xen support
  - Xen has been trying to upstreaming their patches.
    - Switching to Seabios
    - We can reuse their device pass-thru code
      - Needs argument
  - If kvm supports pcie, xen will follow it.

Summary

- PCI Express is useful even in virtualized environment
- Q35 new chipset patch enables QEmu to support PCI Express
- The upstream merge is going on.
- qemu derivatives, KVM and Xen, will follow.

Thank you

Questions?