PCI Express in QEmu

Isaku Yamahata <yamahata@private.email.ne.jp>
<yamahata@valinux.co.jp>
VA Linux Systems Japan K.K.
KVM-forum 2010: August 10, 2010
Agenda

- Introduction
- Current status and future work
- Summary
Introduction
Why PCI Express?

- New features: enhancements as a successor
  - Used as express is widely accepted in the market.
- Some device drivers require express
  - They check if the device is really express
  - Existing PCI device assignment isn't enough
- Hardware certification requires express
- Current PCI support is also limited
PCI Express features from software point of view

- Many enhancements from PCI, for example
  - 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)
Native hot plug

Interrupt on event

Hot plug event handled directly by OS device driver
Without ACPI event handler

Insert/remove device
Advanced Error Reporting (AER)

- Standardized error reporting.
- Important for RAS

OS

Interrupt

Look at error record
Take recovery action
Typically log it and reset the devices.

root port

upstream port

downstream port

Express device

Error Message

Error

Error
PCI express extended configuration space

PCI configuration space

PCI express extended configuration space

PCI compatible Configuration space

PCI express Extended capability

PCI express Extended capability

PCI express extended configuration space

PCI express enhanced access mechanism (ECAM)
PCIe MMCONFIG

0xFFFF FFFF  PCI express extended configuration space

MMIO

MCFG base address

0x0

MMC\ F area (max 256MB)

0xffff

0xff

0x0
Goal

- Enable QEmu to support PCI Express
- Enable PCI Express native device assignment with
  - Native hot plug
  - RAS
- Then, bring Express support to qemu derivative, ie KVM.
Current implementation and future work
I440fx chipset refactoring
64bit BAR
Extended config space
MMConfig
PCI-to-PCI bridge clean up
PCI bus reset

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

Pass DSDT (avoid rom size limit)
PV pci bus numbering
Pass hint for pci bus number
chipset abstraction(i440fx)
64bit BAR
Multi pci bus init
DSDT loading
MCFG
Q35 support

Qemu
MCH
ICH9
Root
upstream
downstream
Q35 chipset
New DSDT
PCI express port switch
AER error injection
pcie_aer_inject_inject
Native hotplug
pcie_abp

Merged
Under review
To be posted
New chipset emulator (Q35 based)

- Why new chipset?
  - Current QEmu chipset is
    - I440FX/PIIX: 10+ years since its release
    - Flat PCI bus (single PCI bus)
  - Discard legacy compatibility
    - It's very difficult to test various legacy OSes
    - Only for modern OSes
- Introduce new functionality with new chipset
  - Q35 (MCH/ICH9) chipset based emulator
  - PCI Express
  - Multi pci bus (PCI-to-PCI bridge, pci express port switch)
  - It lacks iommu/graphics emulation so it should be called P45?
### Current status

#### QEemu

<table>
<thead>
<tr>
<th>Items</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>64bit BAR</td>
<td>Merged</td>
</tr>
<tr>
<td>PCI Bridge lib</td>
<td>Merged to PCI branch</td>
</tr>
<tr>
<td>PCI Bus reset</td>
<td>Under review</td>
</tr>
<tr>
<td>MMCONFIG (PCI layer)</td>
<td>Merged</td>
</tr>
<tr>
<td>PCIe port switch</td>
<td>To be posted</td>
</tr>
<tr>
<td>Including</td>
<td></td>
</tr>
<tr>
<td>native hotplug</td>
<td></td>
</tr>
<tr>
<td>AER error injection</td>
<td></td>
</tr>
<tr>
<td>DSDT overriding</td>
<td>posted (to be resend)</td>
</tr>
<tr>
<td>Q35 Chipset</td>
<td>To be posted</td>
</tr>
<tr>
<td>PV PCI bus numbering</td>
<td>To be posted</td>
</tr>
</tbody>
</table>

#### Seabios

<table>
<thead>
<tr>
<th>Items</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>64bit BAR</td>
<td>Merged</td>
</tr>
<tr>
<td>Multi pci bus</td>
<td>Merged</td>
</tr>
<tr>
<td>Chipset abstraction</td>
<td>Merged</td>
</tr>
<tr>
<td>DSDT overriding</td>
<td>Under review</td>
</tr>
<tr>
<td>MCFG</td>
<td>Under review</td>
</tr>
<tr>
<td>Q35</td>
<td>To be posted</td>
</tr>
<tr>
<td>Q35 DSDT</td>
<td>To be posted</td>
</tr>
<tr>
<td>PV pci bus numbering</td>
<td>To be posted</td>
</tr>
</tbody>
</table>

#### VGABios

<table>
<thead>
<tr>
<th>Items</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>VBE</td>
<td>Waiting Gerd's patch</td>
</tr>
</tbody>
</table>
Future Work

- Upstream merge
- PCI express native device assignment
  - PCI express specific configuration registers should be virtualized
    - Device serial number cap, VSEC...
  - 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.
- Native Power management
- VC(Virtual channel)
- Assigning bus hierarchy tree
Future Work(cont.)

- PCI Express device emulator?
  - Currently pcie port switch only
  - Interesting express native device?
    - IGB? IGBVF? IXGB? IXGBVF?
- Possibly, VT-d/IOMMU shadow paging for nested virtualization ...
  - Qemu iommu emulation is coming. So device assignment version would be wanted.
Summary

- PCI Express is useful even in virtualized environment
- Q35 new chipset patch enables QEmu to support PCI Express
- It benefits all qemu derivatives, including KVM.
Thank you

Questions?