PCI Express support in qemu

Isaku Yamahata, VA Linux Systems Japan K.K.
<yamahata@private.email.ne.jp>
Akio Takebe, Fujitsu Limited
<takebe_akio@jp.fujitsu.com>

Xen Summit Asia Shanghai, China
November 20, 2009

Agenda

Introduction

New chipset emulator in qemu

Summary

Future work

Introduction

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

More PCI features are wanted

Currently Qemu only supports part of PCI specs. e.g.64bit BAR isn't supported.

More buses/slots

Qemu only supports single host bus (for PC emulation). Sub PCI bus isn't supported.

3+ pci bus(96+ slots)/96+ pcie slots are wanted.

Brige emulation: filtering aren't implemented.

Motivation(cont.)

- PCI express features
  - Hot plug, power management, ARI, AER
- Native pass through of PCI express device to guest OS
  - PCI express devices can be passed through as PCI device, though.
- Need to fill those gap between newer real hardware features and qemu emulation mainly in PCI related area.

New chipset emulator for new hardware features

Why new chipset?

Keep the currently supported chipset(I440FX, PIIX) for legacy compatibility.

Old OSes must run on it.

add new features for modern OSes without legacy compatibility.

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 view of emulation except graphics.

New chipset emulator(cont.)

Now the followings are working

64bit BAR

PCI express MMCONFIG

BIOS updates(MCFG, e820)

Linux boots happily using MMCONFIG

Windows XP installs and boots.

Currently those are under heavy review/rewrite process for qemu 0.12.0.

Q35 chipset emulator doesn't have

IOMMU(VT-d) emulation in qemu

It would make sense to support IOMMU for guest OS.

This requires full redesign of qemu DMA layer.

Graphic emulation

So it should be called P45 emulator, not Q35?

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 the 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

Summary: current status

PCI express

PCIe MMCONFIG	Merged.
Q35 chipset base	working. Pushing this now.
PCIe portemulator	WIP
pcie native hotplug	WIP
pcie passthrough	WIP
3+ pci bus	working(not merged yet)

pcbios

mcfg	working(not merged yet)
e820	working(not merged yet)
host bridge initiazatlin	working(not merged yet)
pci io/memory space initialization	working(not merged yet)
switching acpi table or passing acpi table outside qemu	WIP

sea bios: Not started yet

Future work

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 passthrough.
  - 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 have been created.

Qemu 0.12.0 release will use seabios instead of pcbios.

So patches for seabios are necessary for merging.

Future work:ACPI

BIOS 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: IOMMU?

IOMMU: Intel VT-d, AMD IOMMU

IOMMU emulator in qemu

The implementation will be interesting.

Requires full revise of qemu DMA layer

Shadowing IOMMU page tables for guest OS

For HVM guest

Future work: passthrough support in qemu?

Support passthrough in qemu.

Hopefully consolidate xen passthrough code into qemu.

By consolidating the passthrough code into qemu, the code base would get more tests and become more stable.

Thank you

Questions?

This work was partly funded by Ministry of Economy, Trade and Industry (METI) of Japan as the Secure Platform project of Association of Super-Advanced Electronics Technologies (ASET).