

# パススルーI/Oの実装と今後

2008年11月21日

島田 雄二

NECシステムテクノロジー(株)

# 目次

1. パススルーI/Oの実装

2. 今後の予定

3. パススルーI/Oの課題

4. まとめ

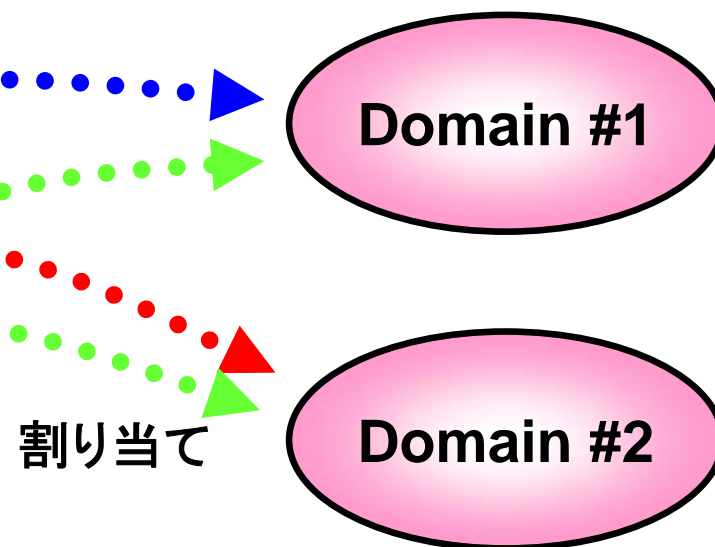
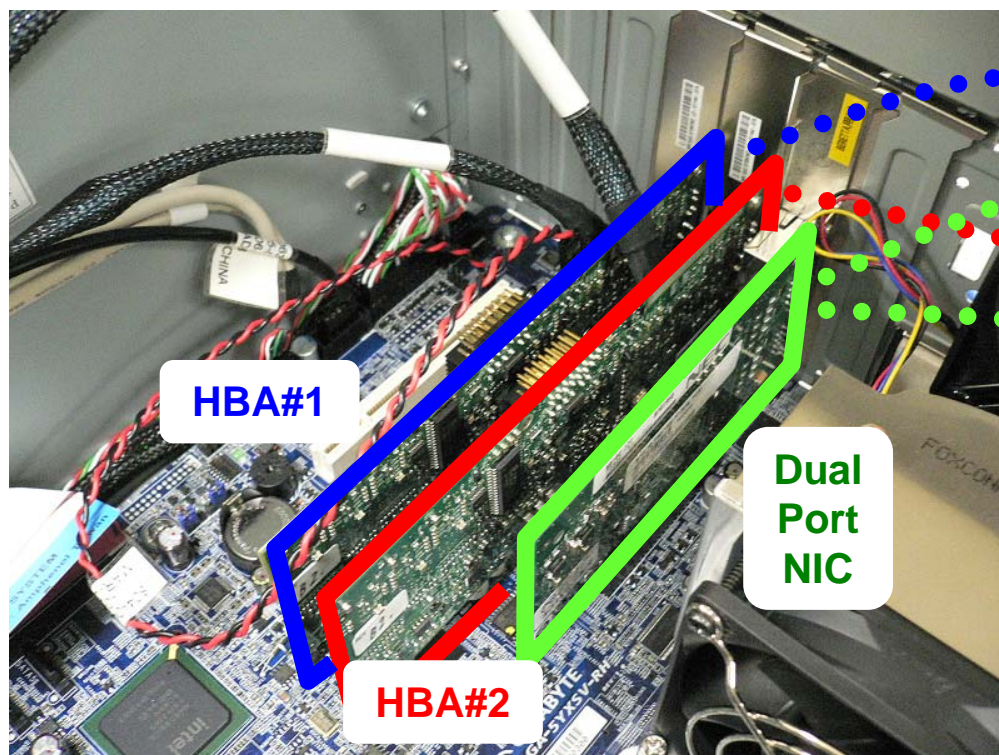
# 1. パススルーI/Oの実装

# パススルーI/Oとは

- DomainへI/Oデバイスを割り当て、Guestソフトウェアから直接制御する

HBAを割り当て → ストレージへ高速アクセス

NICを割り当て → ネットワークへ高速アクセス



本発表ではHVM DomainのパススルーI/Oについて説明する

Xen Summit Tokyo 2008

# パススルーI/O概要(1)

- **MMIO仮想化**

- Guest Physical AddressからMachine Addressへページ単位で対応付け
- I/Oデバイスのリソース再割り当て
  - BIOSが割り当てたMMIOリソースが、ページ・サイズにアラインされていないならば、Domain0 Linuxがページ・サイズにアラインされているMMIOリソースをI/Oデバイスに再割り当て

改造点1

- **コンフィグレーション・レジスタ・アクセス仮想化**

- ioemuはGuestソフトウェアからのコンフィグレーション・レジスタ・アクセスをトラップし、ビットのタイプに応じてそのままパススルーまたはエミュレート

改造点2

Xen Summit Tokyo 2008

# パススルーI/O概要(2)

- **DMA仮想化**
  - デバイス・ドライバはI/OデバイスにGuest Physical Addressを書き込む
  - I/OデバイスはGuest Physical Addressを使用してDMAを実行
  - チップセット上のIOMMUがGuest Physical AddressをMachine Addressに変換
- **ポートI/O仮想化**
  - HypervisorはIN/OUT命令をトラップし、Guestソフトウェアの代わりにI/Oデバイスにアクセス
- **割り込み仮想化**
  - HypervisorはI/Oデバイスから割り込みを受け、割り込みコントローラをエミュレートし、割り込みをHVM Domainへインジェクト

# 改造点1 - I/Oデバイスのリソース再割り当て

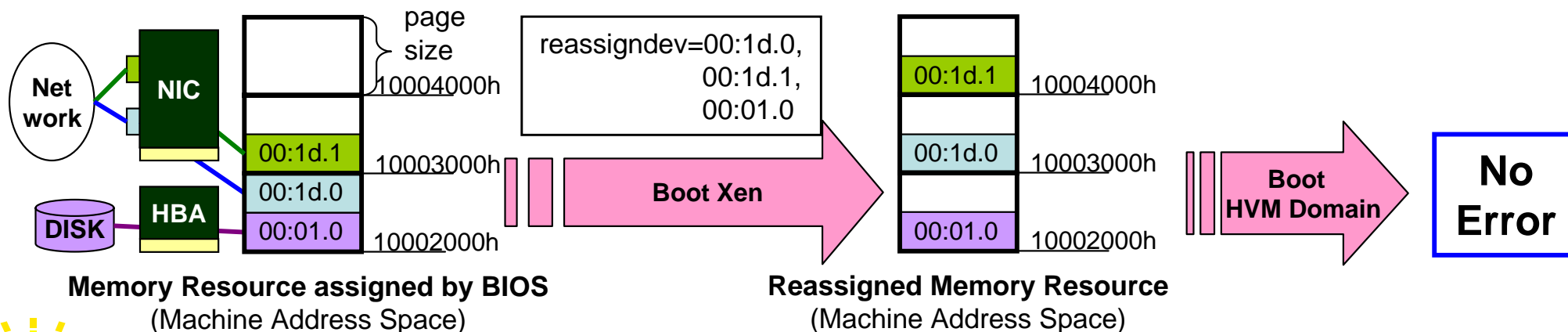
## 問題

- BIOSにより割り当てられたリソースがページ・サイズにアラインされていなければ、HVM DomainへのI/Oデバイス割り当てに失敗

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

## 改造内容

- BIOSが割り当てたリソースを開放し、I/Oデバイスにページ・サイズにアラインされたリソースを再割り当て
- リソース再割り当て対象のI/Oデバイスを指定するブート・パラメータを追加



BIOSによるリソースの割り当て状況に関わらず、HVM DomainにI/Oデバイスを割り当てることが可能に。

Xen Summit Tokyo 2008

## 改造点2 - コンフィグレーション・レジスタ・アクセス仮想化(1)

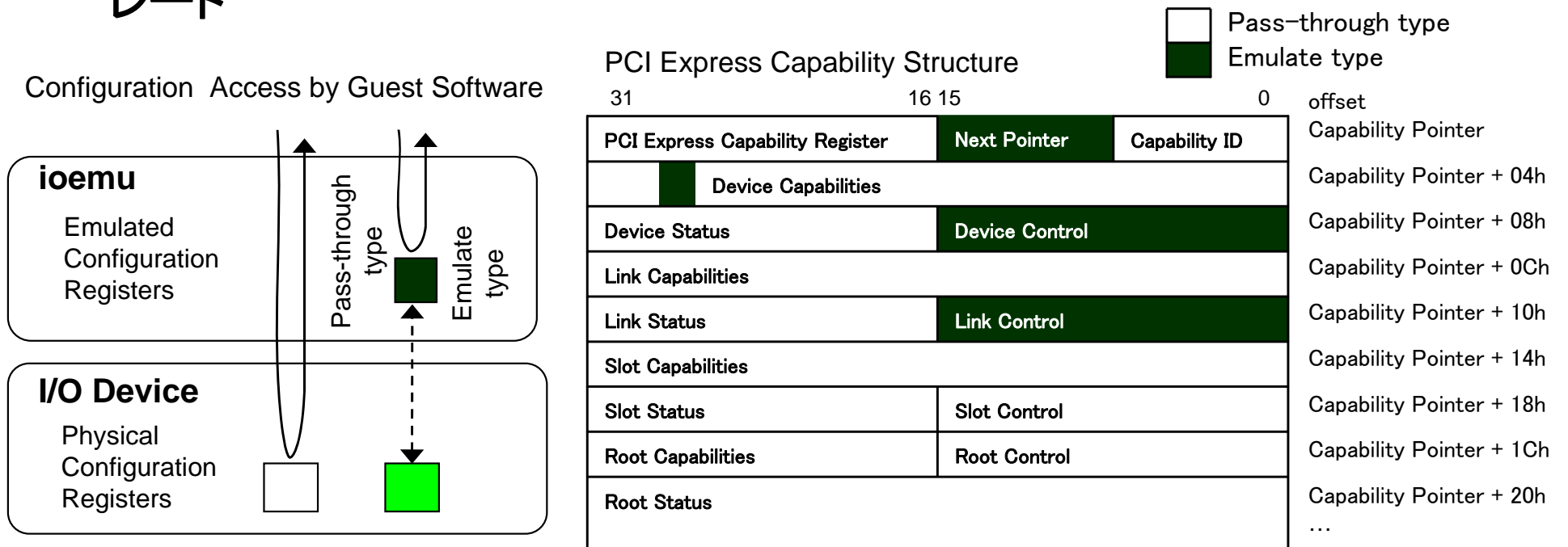
- 問題
  - Configuration Header Type 0以外のレジスタはリード・オンリーであった為、HVM Domainに割り当てることのできるI/Oデバイスが限定
- 改造内容
  - Guestソフトウェアが次のレジスタにリード/ライト・アクセス可能とし、他のレジスタは隠蔽
    - Configuration Header Type 0
    - MSI Capability Structure
    - MSI-X Capability Structure
    - PCI Express Capability Structure
    - PCI Power Management Capability Structure
    - Vital Product Data Capability Structure
    - Vendor Specific Capability Structure
    - Device Specific Register (HeaderとCapability Structureを除く)

Xen Summit Tokyo 2008

# 改造点2 - コンフィグレーション・レジスタ・アクセス仮想化 (2)

## ● 改造内容(続き)

- Guestソフトウェアが直接制御可能なビットはパススルー
- BIOS設定値を保護すべきビット、特殊な処理が必要なビットはエミュレート



HVM DomainにUSBやSAS HBAなど様々なデバイスを割り当て可能に。  
不正なGuestソフトウェアからシステム全体を保護。

Xen Summit Tokyo 2008

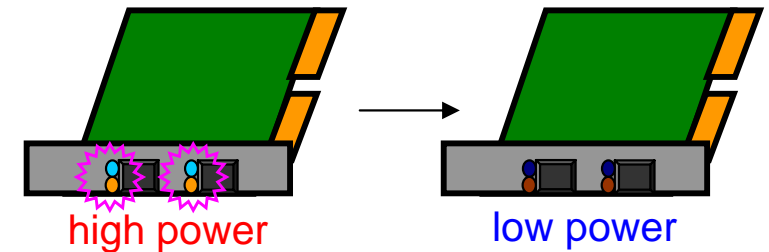
## 2. 今後の予定

# 今後の予定 (1)

- **GuestソフトウェアからI/Oデバイスの電源制御**
  - デバイスの電源制御／状態レジスタ(PMCSR)による直接制御
  - D0～D3hotをサポート



省電力化。



- **割り込みのリダイレクトを削減**

- Xen 3.3.0において、割り込みは一つのプロセッサに集中(仮想プロセッサ0番)
  - そのプロセッサが、Guestソフトウェアが指定したプロセッサでなければ割り込みをリダイレクト
- 割り込みのリダイレクトを削減するために、Guestソフトウェアが設定したプロセッサへの割り込み配送をサポート



割り込み配送のパフォーマンス向上。

Xen Summit Tokyo 2008

# 今後の予定 (2)

- ioemuのログ改善

- ioemuのログに日付情報、プロセスIDを付加

example

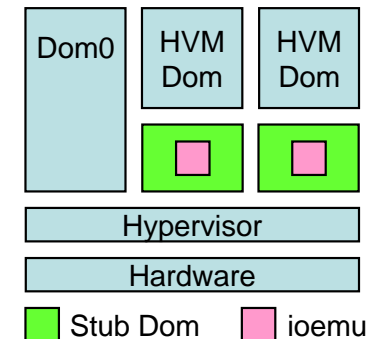
```
[2008-11-21 16:10:05 4567] can't store dev vc name for domid 1 in /parallel/0 from a stub domain  
[2008-11-21 16:10:06 4567] qemu_map_cache_init nr_buckets = 10000 size 3145728  
[2008-11-21 16:10:06 4567] shared page at pfn 1ffe
```



障害発生時の解析負担軽減。

- Stub DomainによるパススルーI/Oをサポート

- Stub DomainはDomain0の代わりにI/Oをエミュレーション
- Stub Domainを使用したパススルーI/Oを可能に



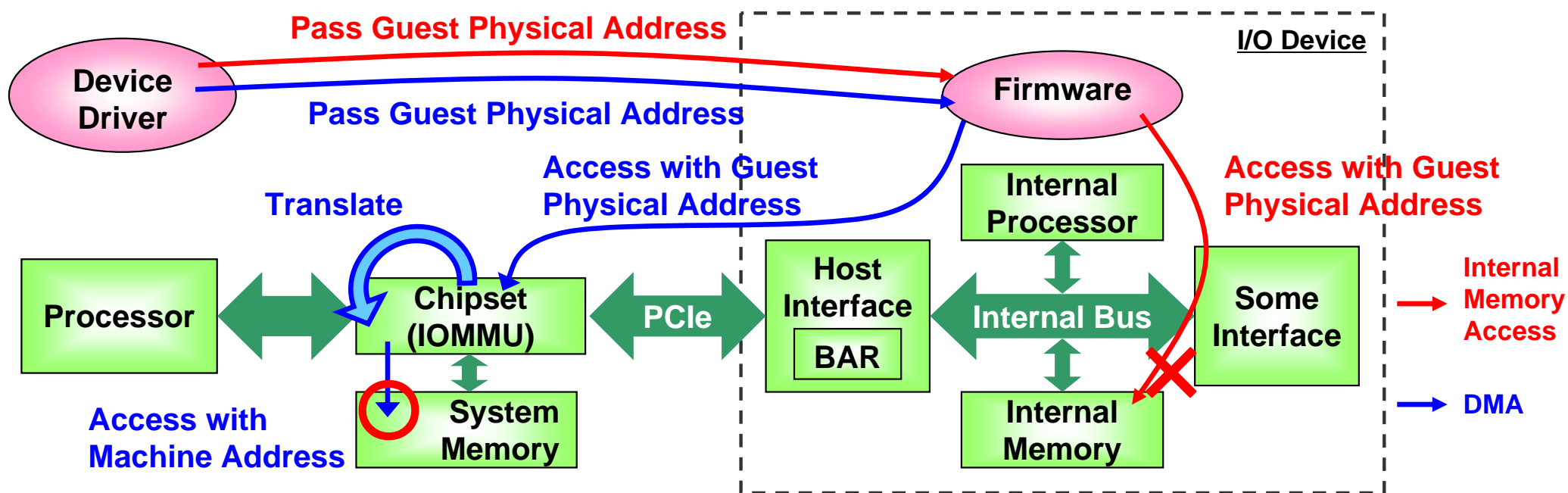
スケーラビリティ向上、Domain0縮小。

Xen Summit Tokyo 2008

### 3. パススルーI/Oの課題

# 課題1 - パススルーI/Oに適さないデバイス

- デバイス・ドライバから取得したPhysical Addressを内部メモリへのアクセスに使用するI/Oデバイスは動かない



- パススルーI/Oを使用する際に、デバイス・ドライバが使用するPhysical Addressは”Guest Physical Address”であり、Machine Addressとは異なる。
- DMA時は、IOMMUがGuest Physical AddressをMachine Addressへ変換する。
- 内部メモリへのアクセス時は、IOMMUはAddressを変換しない。その結果、アクセスに失敗する。

Xen Summit Tokyo 2008

# 課題1 – この問題を引き起こす「仮想化の穴」

- I/Oデバイスがデバイス・ドライバから取得したGuest Physical AddressをDMA以外の目的で使用する場合は、Guest Physical AddressをMachine Addressへ変換することができない
- これはIOMMUベースのHypervisorで共通する問題



Adapter VendorがI/OデバイスをパススルーI/Oに適する設計にしてほしい。

## 課題2 – ioemuがPCIeをサポートしていない

- ioemuは以下の機能が無い

- MMCFGメカニズム

- メモリ空間にアクセスすることでコンフィグレーション・レジスタにアクセス
- オフセットが100hからFFFhのレジスタへのアクセスに必要

- 100hからFFFhにあるCapability Structure

- 例: Device Serial Number Capability Structure
- PCI Express VSEC Structure

- Root Port

- PCIe固有の機能を制御するレジスタを持つデバイス

- Advanced Error Reporting (AER)

- PCIeエラーを通知

- PCIeホット・プラグ

- ACPIホット・プラグより新しいホット・プラグの仕組み

### その結果...

Guestソフトウェアが、必要なレジスタ、Capability Structureへアクセスできない

Guestソフトウェアからみたトポロジが、PCIeマシンではない

GuestソフトウェアはPCIeエラーをリカバリできない

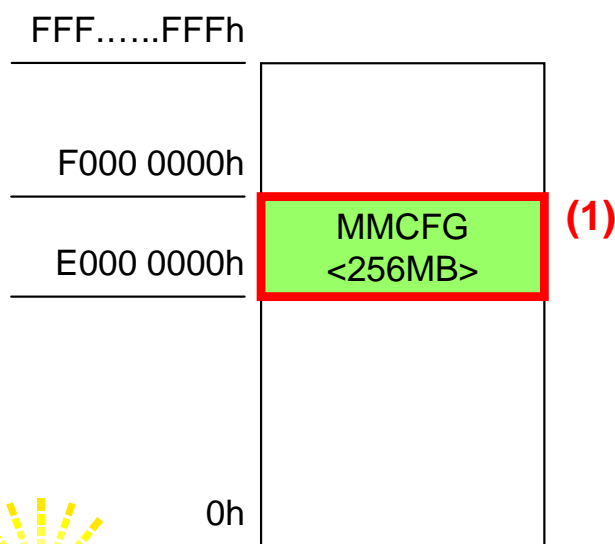
ACPIホット・プラグを使用しなければならない

Xen Summit Tokyo 2008

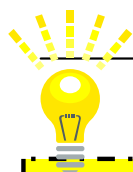
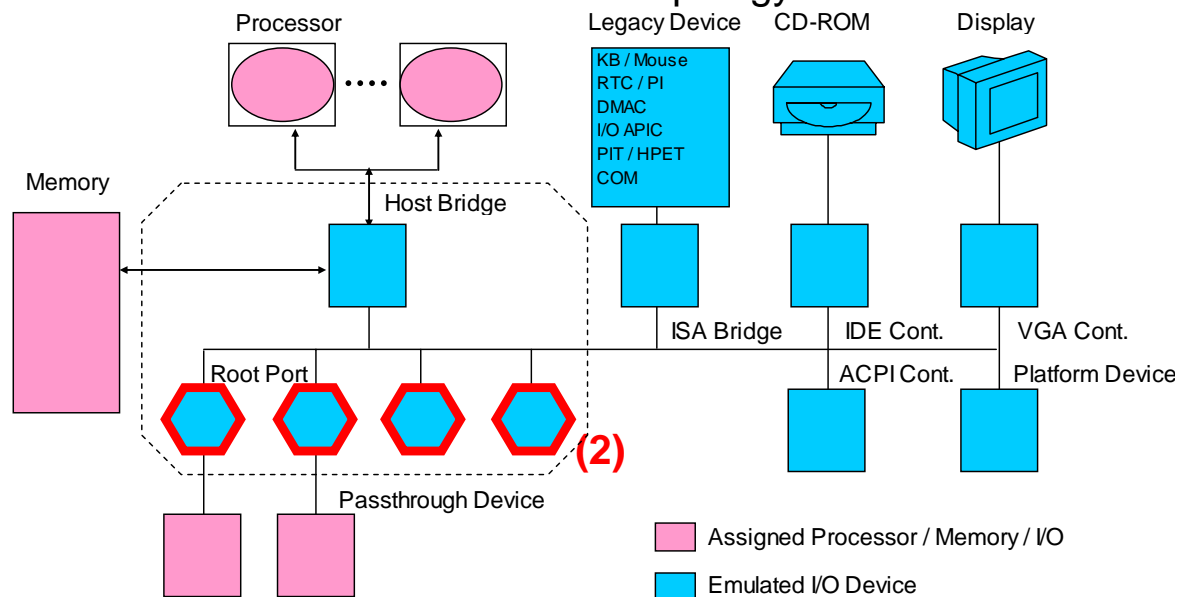
## 課題2 - PCIeをサポートするための開発

- まずはMMCFGメカニズム(1)とRoot Portエミュレータ(2)をioemuに追加
- 次に、他のものを実装

Guest Physical Address Space



Virtual Machine Topology



一人の開発者/一つのチームが全てを開発することは難しい。みなさんの協力が必要。

Xen Summit Tokyo 2008

## 4. まとめ

# まとめ

- **パススルーI/Oは実用的になってきた**
- **我々は今後もパススルーI/O関連の機能を開発していく**
- **パススルーI/Oには解決すべき課題が残っている**
- **様々な技術者の力が必要**

# 謝辞

- 皆様に深く感謝致します。
  - Xen Summit Tokyo 2008関係者の方々
  - Xenコミュニティ技術者の方々

Empowered by Innovation

**NEC**