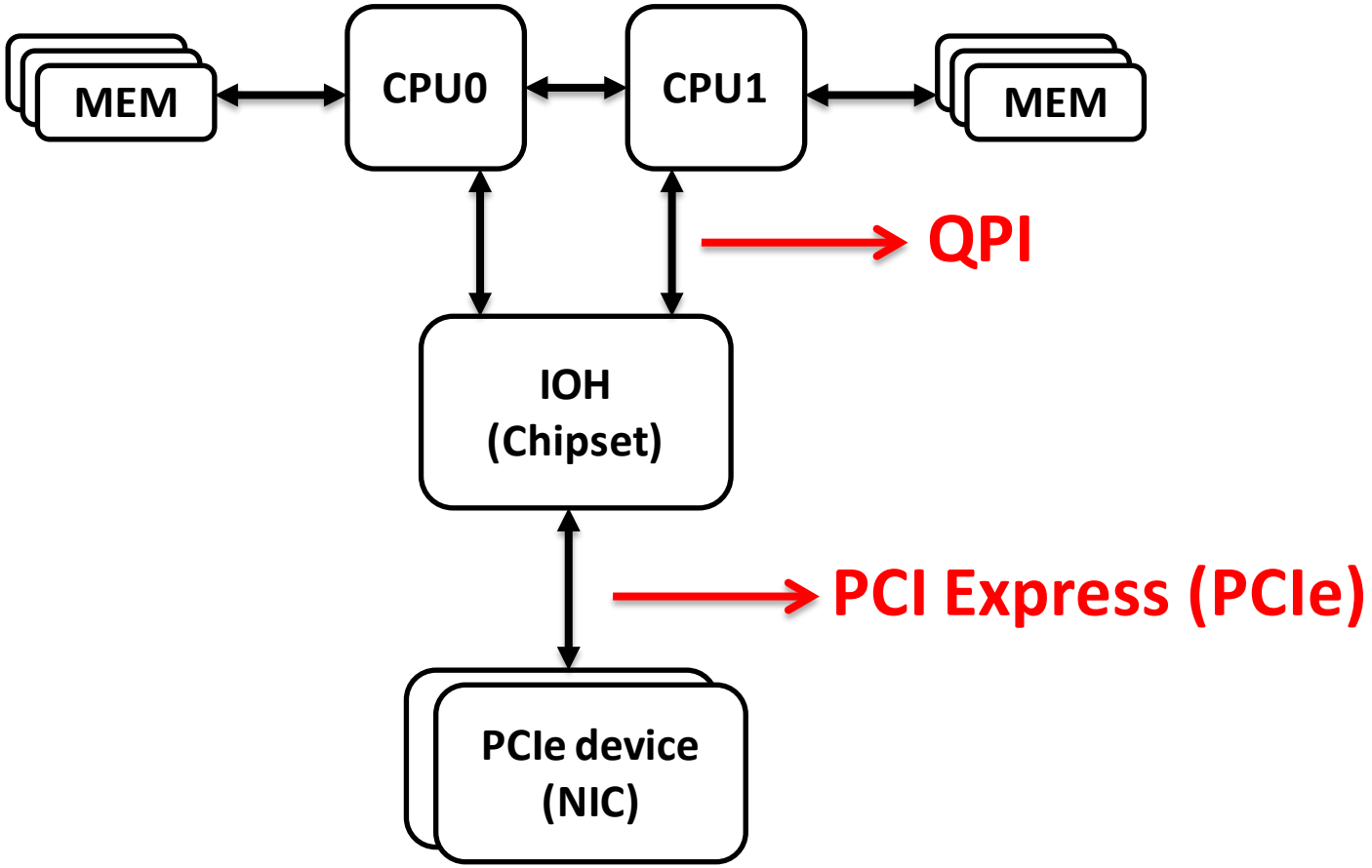
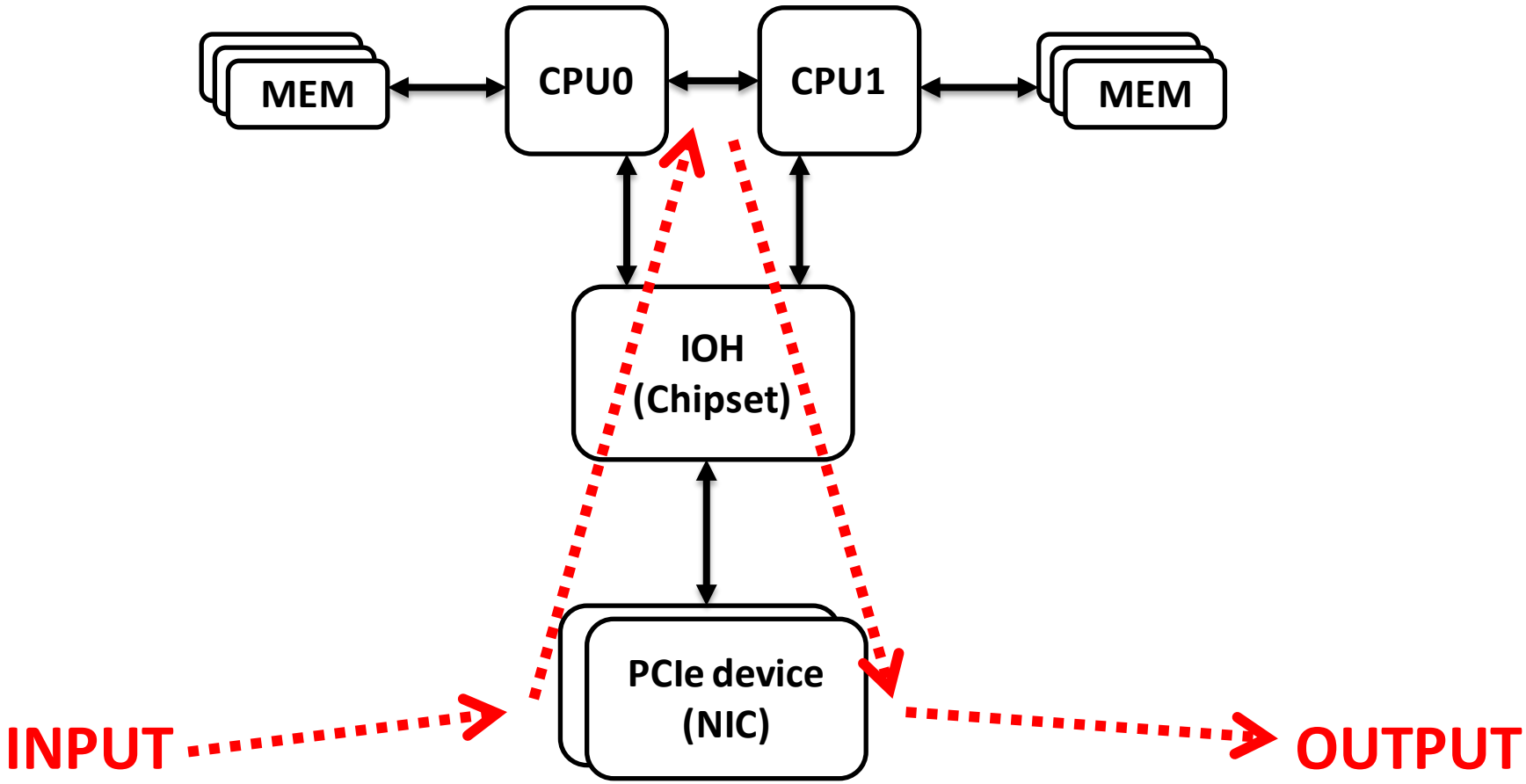
パケット転送を支える技術

海老澤 健太郎 @ パラレルス株式会社
Twitter: @ebiken

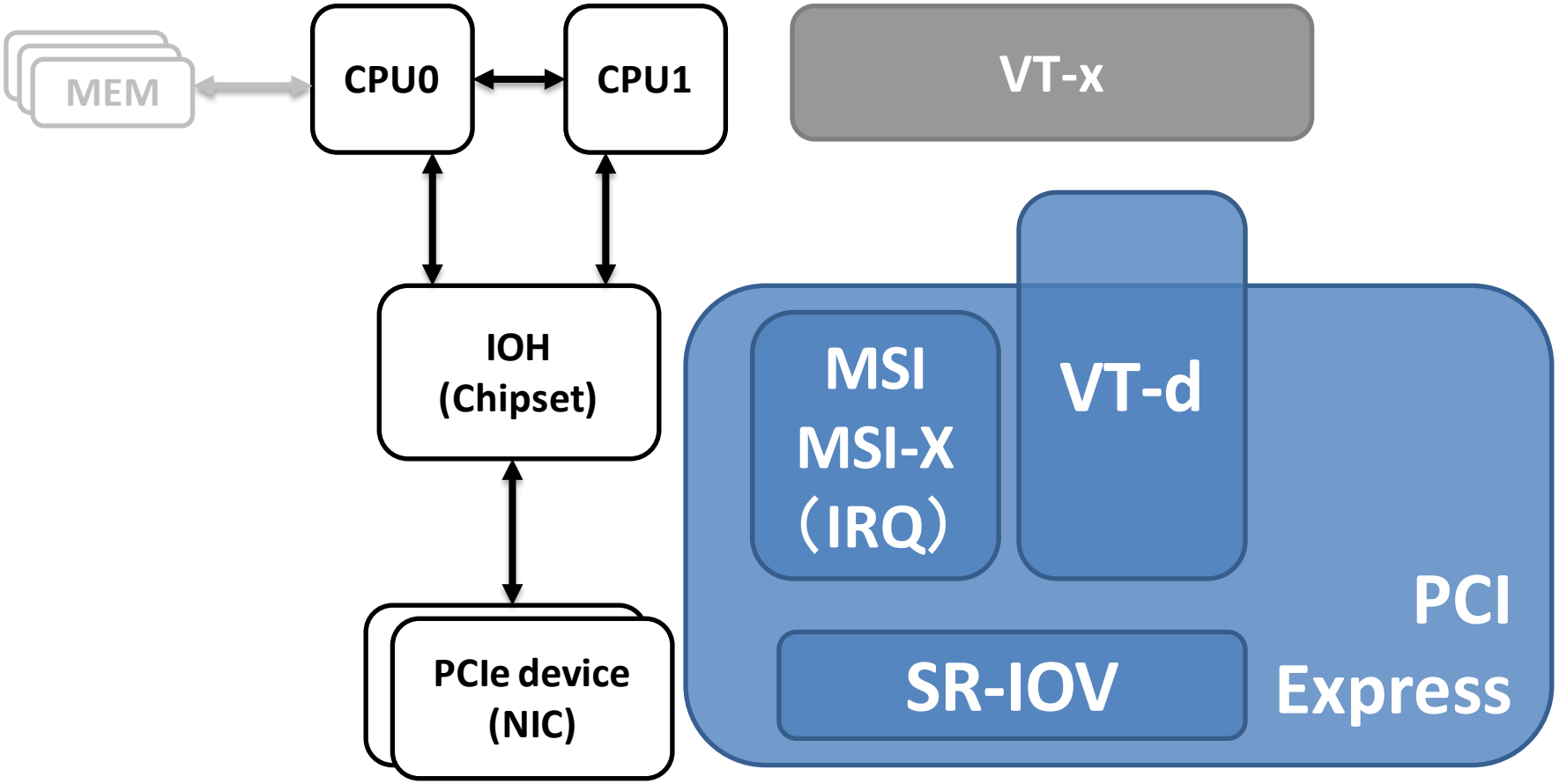
サーバーの物理構造


MEM
CPU0
CPU1
MEM
QPI
IOH
(Chipset)
PCI Express (PCIe)
PCIe device
(NIC)

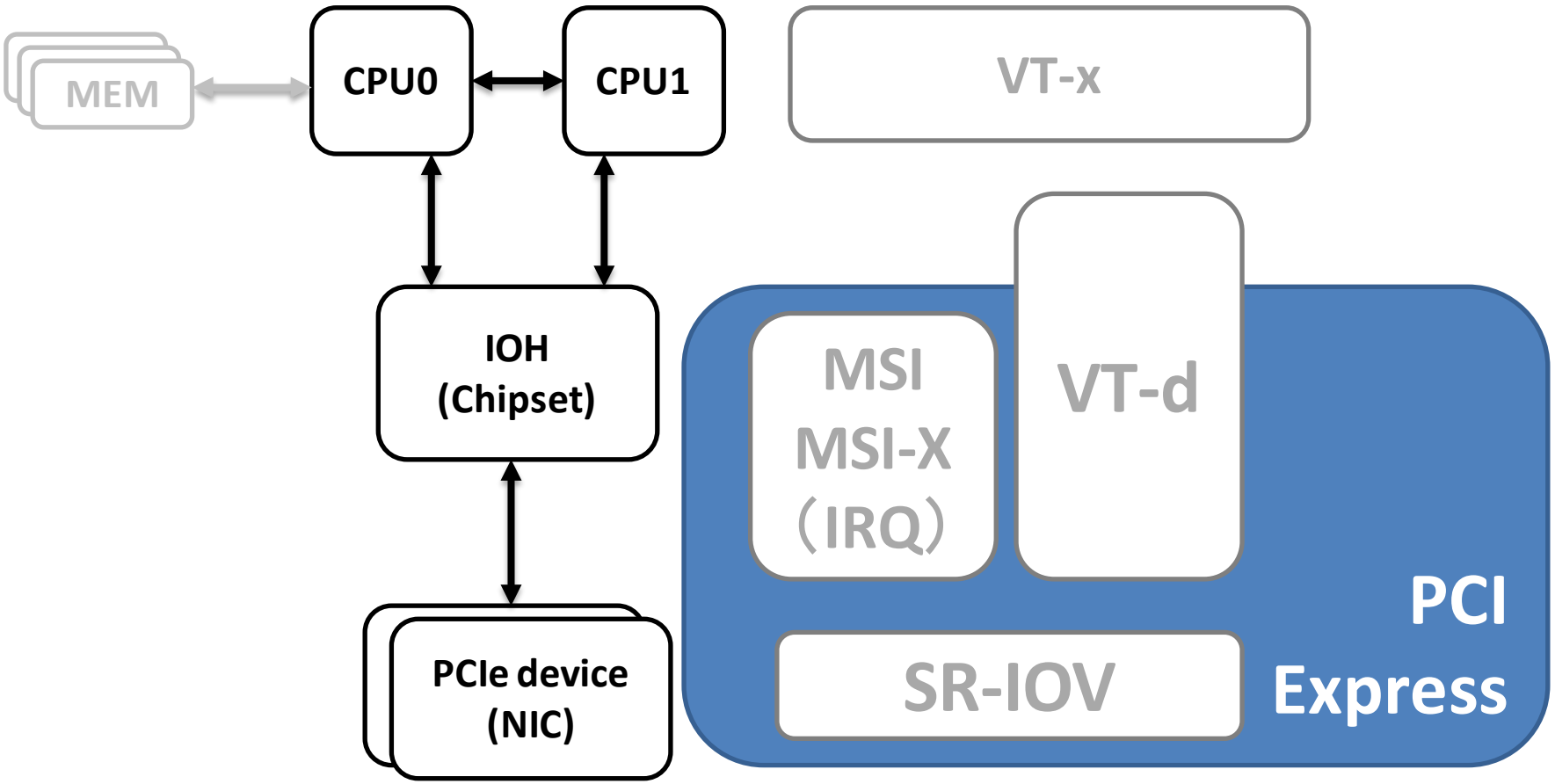
パケットの流れ



仮想化＆パケット転送 技術マップ



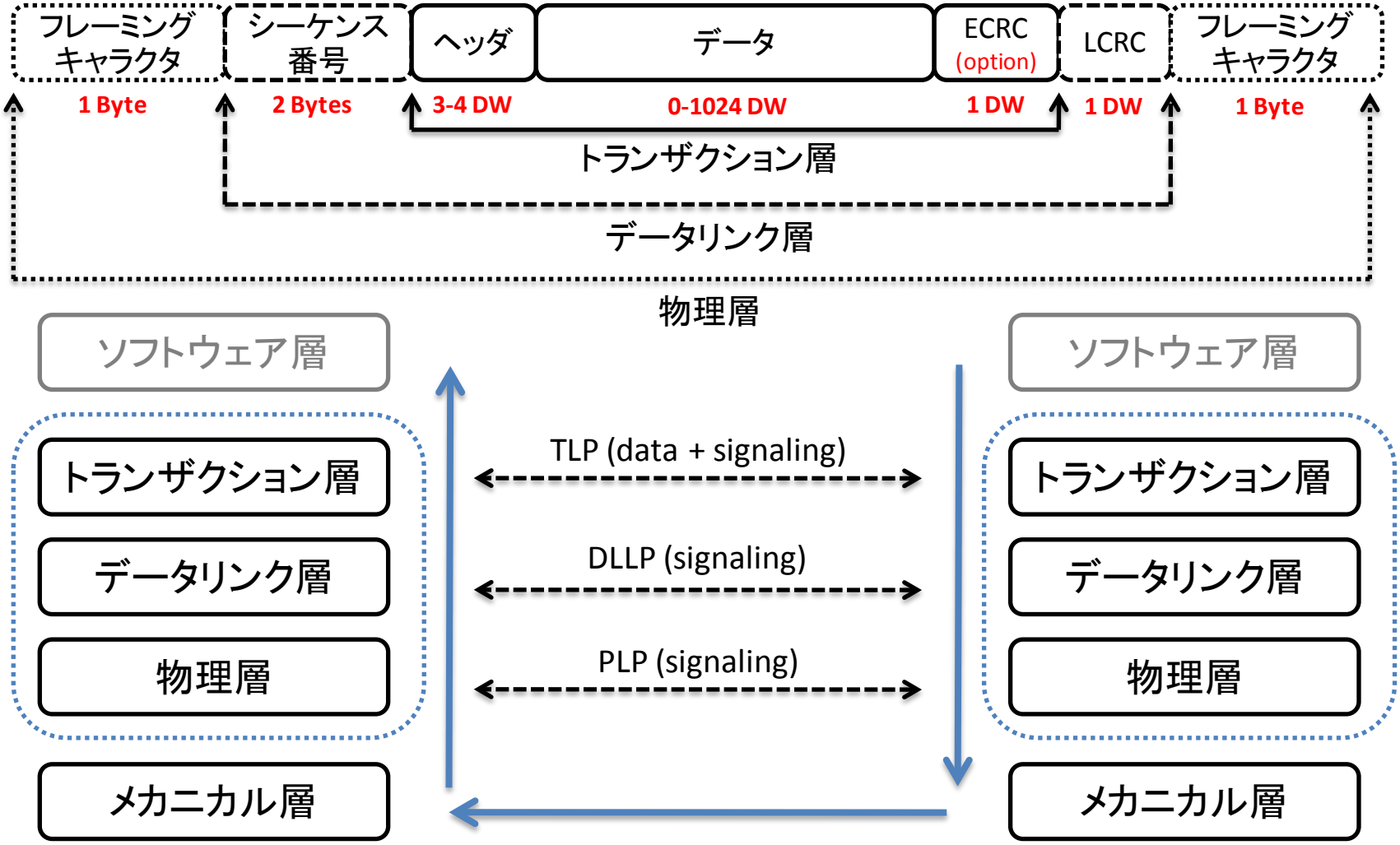
PCI Express - PCIe



PCI Express - PCIe とは？

- 「パケットベース」のプロトコル
- CRCやデータコーディングによるエラー検出
  - 8b10bエンコーディングオーバーヘッドに考慮（20%）
- point-to-point接続のシリアル通信
  - リンク（バス）をスイッチで繋ぐツリー構造
  - ポート速度を処理するのに十分なバス幅を確保

パケット構造


フレーミング キャラクタ
シーケンス 番号
ヘッダ
データ
ECRC (option)
LCRC
フレーミング キャラクタ
1 Byte
2 Bytes
3-4 DW
0-1024 DW
1 DW
1 DW
1 Byte
トランザクション層
データリンク層
物理層
ソフトウェア層
ソフトウェア層
トランザクション層
データリンク層
物理層
メカニカル層
TLP (data + signaling)
DLLP (signaling)
PLP (signaling)
トランザクション層
データリンク層
物理層
メカニカル層

PCIe の帯域

	伝送速度	実効速度 （8b10b考慮）
PCIe 1.1 x 4	10Gbps x 2	8Gbps x 2
PCIe 1.1 x 8	20Gbps x 2	16Gbps x 2
PCIe 2.0 x 4	20Gbps x 2	16Gbps x 2
PCIe 2.0 x 8	40Gbps x 2	32Gbps x 2

10Gbps x 2 port x Full Duplex には
PCIe 2.0 x 8 Lane 必要

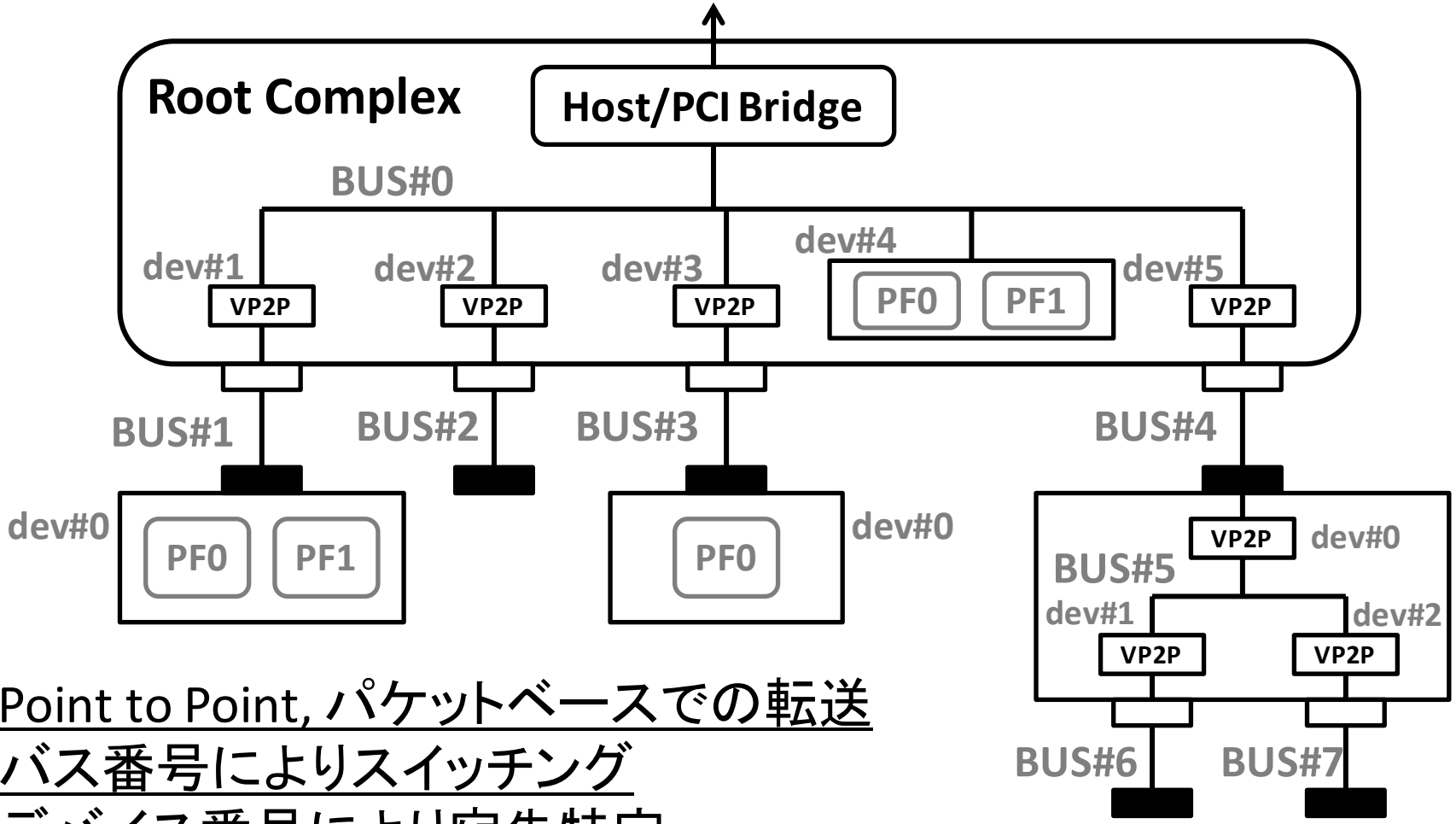
PCIe デバイスをLinuxから見てみる

```
[ebiken@iwpf01 ~]$ lspci
... snip ...
01:00.0 Ethernet controller: Broadcom Corporation NetXtreme II
01:00.1 Ethernet controller: Broadcom Corporation NetXtreme II
03:00.0 Ethernet controller: Intel Corporation 82599EB 10-Gigab
03:00.1 Ethernet controller: Intel Corporation 82599EB 10-Gigab
```

BUS:Device.Functionで認識

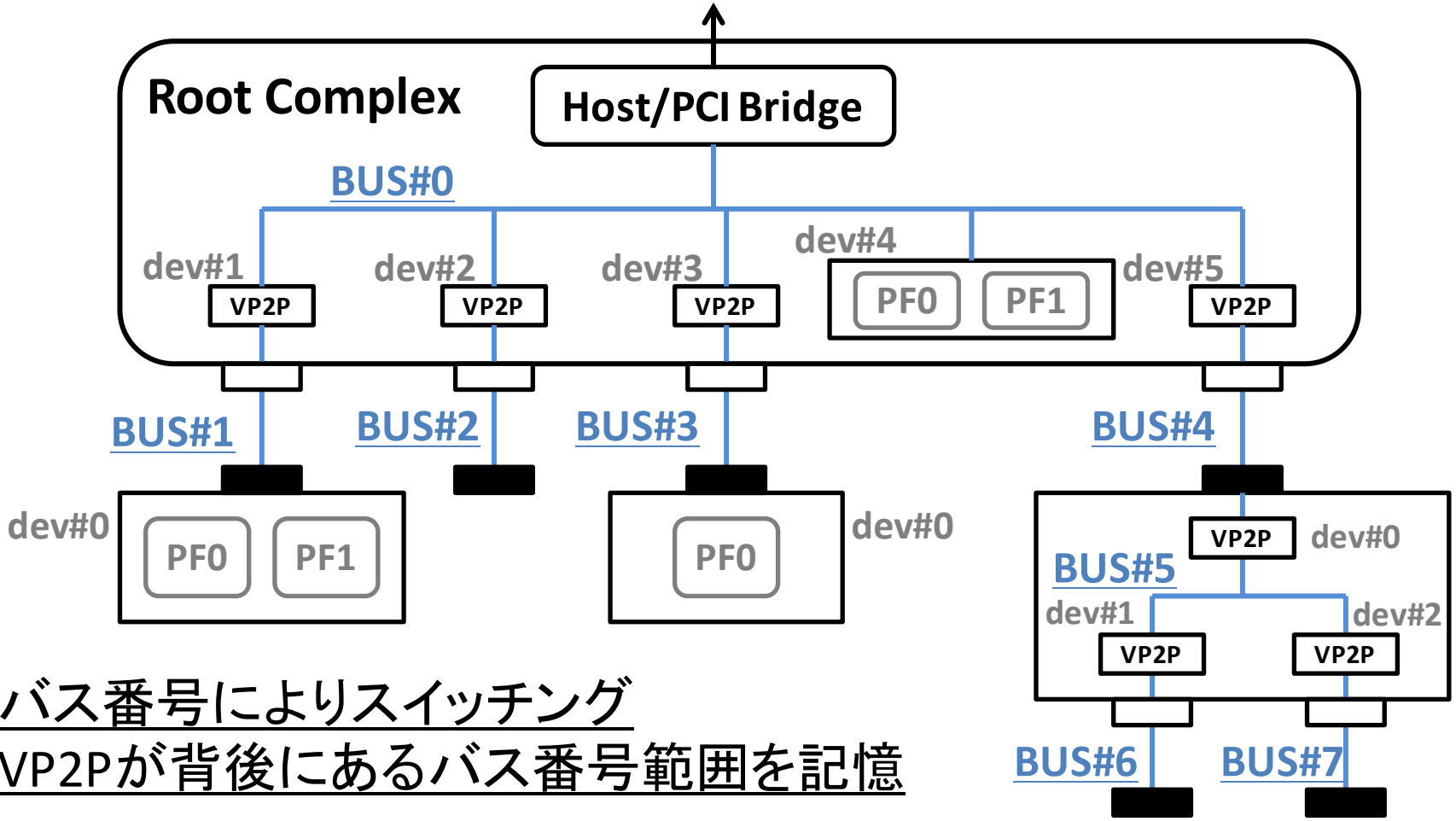
```
[ebiken@iwpf01 ~]$ lspci
... snip ...
01:00.0 Ethernet controller: Broadcom Corporation NetXtreme II BCM5716 Gigabit Ethernet (rev 20)
01:00.1 Ethernet controller: Broadcom Corporation NetXtreme II BCM5716 Gigabit Ethernet (rev 20)
03:00.0 Ethernet controller: Intel Corporation 82599EB 10-Gigabit SFI/SFP+ Network Connection (rev 01)
03:00.1 Ethernet controller: Intel Corporation 82599EB 10-Gigabit SFI/SFP+ Network Connection (rev 01)
```

PCIe Topology – 物理配線



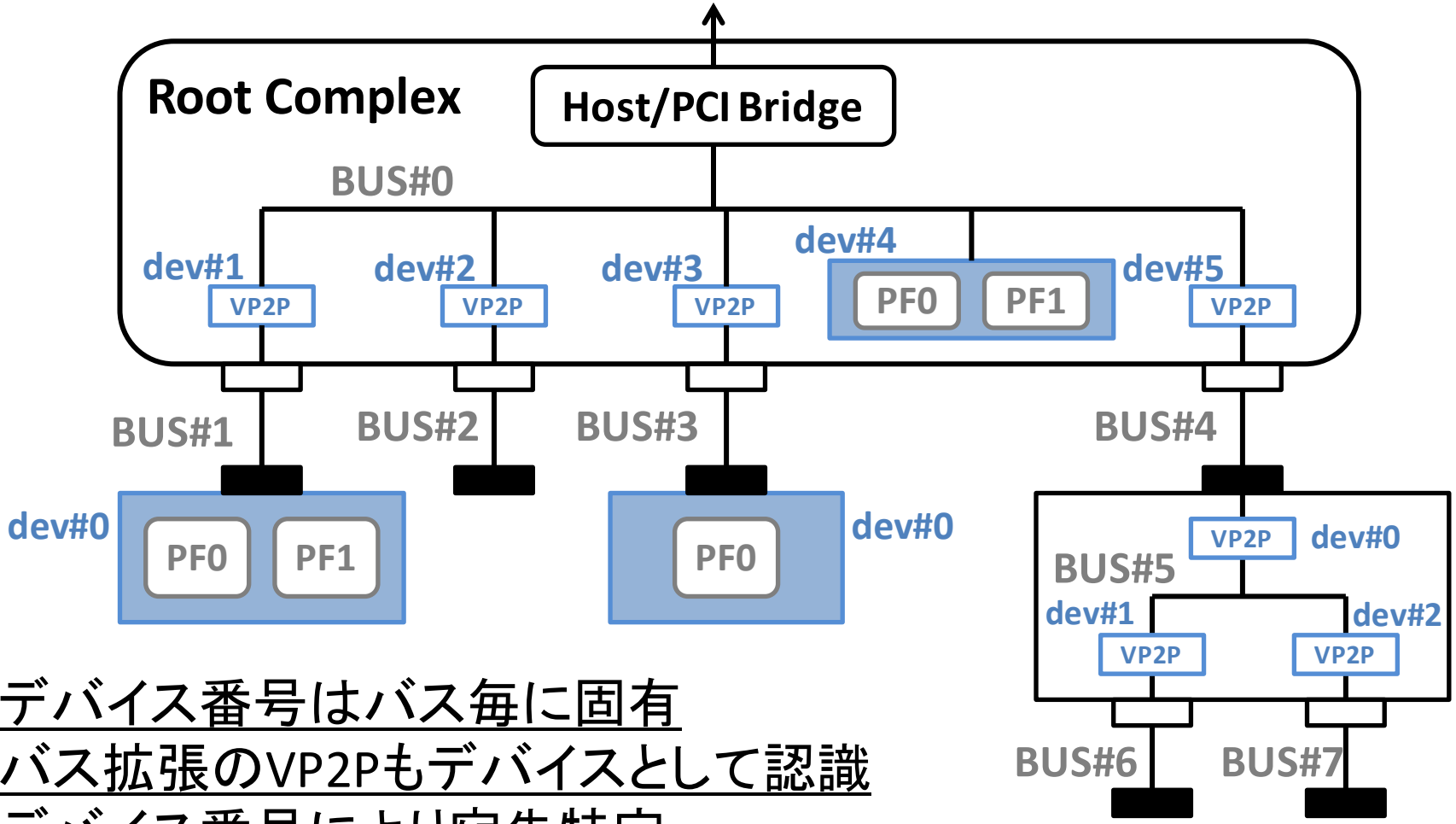
- Point to Point, パケットベースでの転送
- バス番号によりスイッチング
- デバイス番号により宛先特定

PCIe Topology - BUS:Device:Function



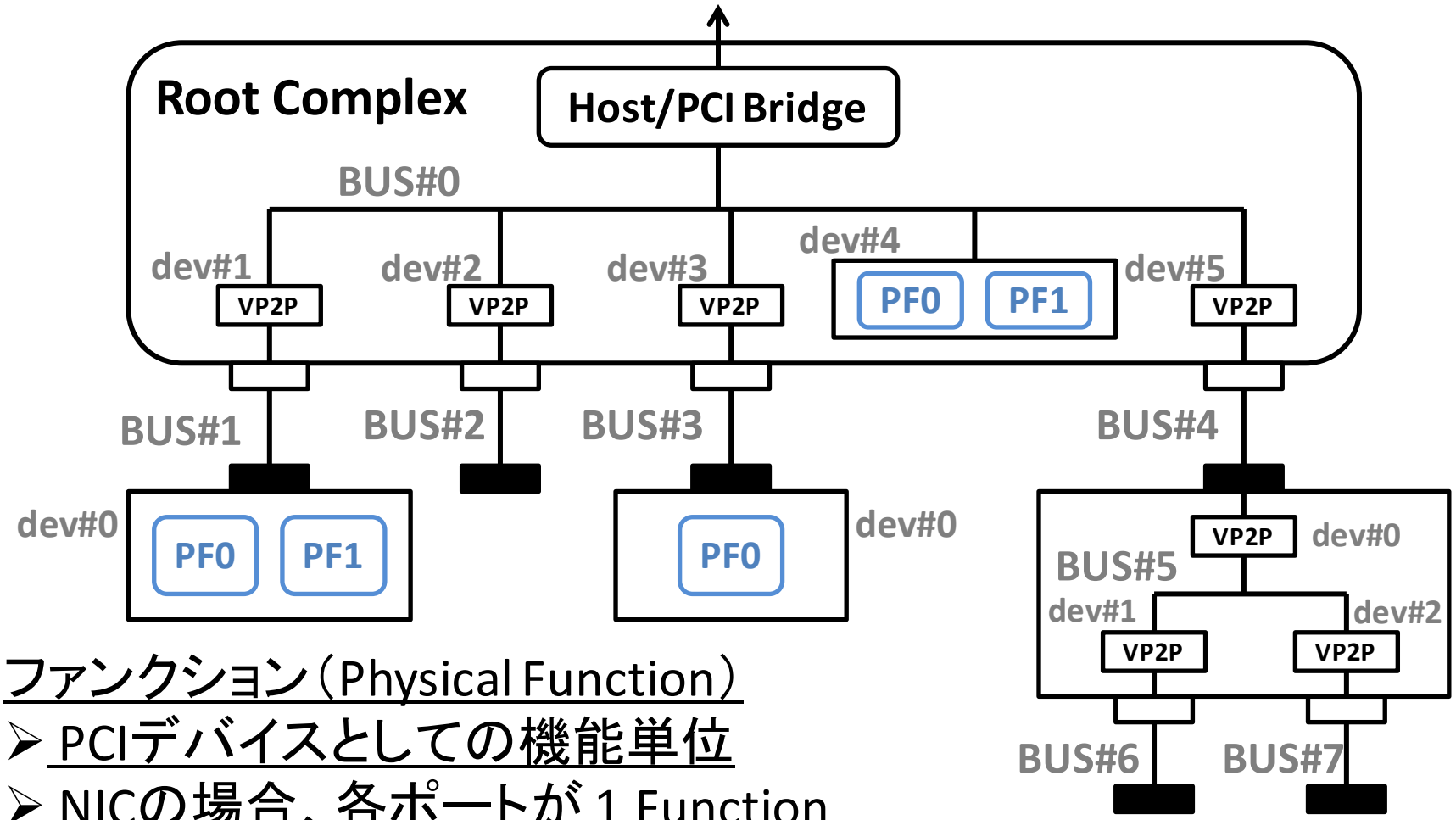
- バス番号によりスイッチング
- VP2Pが背後にあるバス番号範囲を記憶

PCIe Topology - BUS:Device:Function



- デバイス番号はバス毎に固有
- バス拡張のVP2Pもデバイスとして認識
- デバイス番号により宛先特定

PCIe Topology - BUS:Device:Function

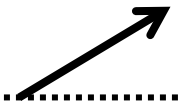


ファンクション（Physical Function）

- PCIデバイスとしての機能単位
- NICの場合、各ポートが 1 Function
- Sound+Videoカードならそれぞれが Function

PCIe トポロジーをOSから見てみる

```
[ebiken@iwpf01 ~]$ lspci -tv
... snip ...
 ¥-[0000:00]-+-00.0  Intel Corporation 5500 I/O
             +-01.0-[01]--+-00.0  Broadcom Corp
             |            ¥-00.1  Broadcom Corp
             +-03.0-[02]--
             +-07.0-[03]--+-00.0  Intel Corpora
             |            ¥-00.1  Intel Corpora
```



```
[ebiken@iwpf01 ~]$ lspci -tv
... snip ...
 ¥-[0000:00]-+-00.0  Intel Corporation 5500 I/O Hub to ESI Port
             +-01.0-[01]--+-00.0  Broadcom Corporation NetXtreme II BCM5716 Gigabit Ethernet
             |            ¥-00.1  Broadcom Corporation NetXtreme II BCM5716 Gigabit Ethernet
             +-03.0-[02]--
             +-07.0-[03]--+-00.0  Intel Corporation 82599EB 10-Gigabit SFI/SFP+ Network Connection
             |            ¥-00.1  Intel Corporation 82599EB 10-Gigabit SFI/SFP+ Network Connection
             +-14.0  Intel Corporation 5520/5500/X58 I/O Hub System Management Registers
             +-14.1  Intel Corporation 5520/5500/X58 I/O Hub GPIO and Scratch Pad Registers
```

PCIe トポロジーをOSから見てみる

```
[ebiken@iwpf01 ~]$ lspci -tv
... snip ...
 ¥-[0000:00]-+-00.0  Intel Corporation 5500 I/O
             +-01.0-[01]--+-00.0  Broadcom Corp
             |            ¥-00.1  Broadcom Corp
             +-03.0-[02]--
             +-07.0-[03]--+-00.0  Intel Corpora
             |            ¥-00.1  Intel Corpora
```

バス番号 : デバイス番号 : ファンクション番号

※ファンクション×２＝２ポート

デバイスの機能を確認

[ebiken@iwpf01 ~]$ ***lspci -v -s 03:00.0***
03:00.0 Ethernet controller: Intel Corporation ***82599EB*** 10-Gigabit SFI/SFP+ Network Connection (rev 01)
Subsystem: Intel Corporation Ethernet Server Adapter ***X520-2***
Flags: bus master, fast devsel, latency 0, IRQ 38
Memory at d9900000 (64-bit, prefetchable) [size=512K]
I/O ports at fcc0 [size=32]
Memory at d98f8000 (64-bit, prefetchable) [size=16K]
Capabilities: <access denied>
Kernel driver in use: ixgbe
Kernel modules: ixgbe

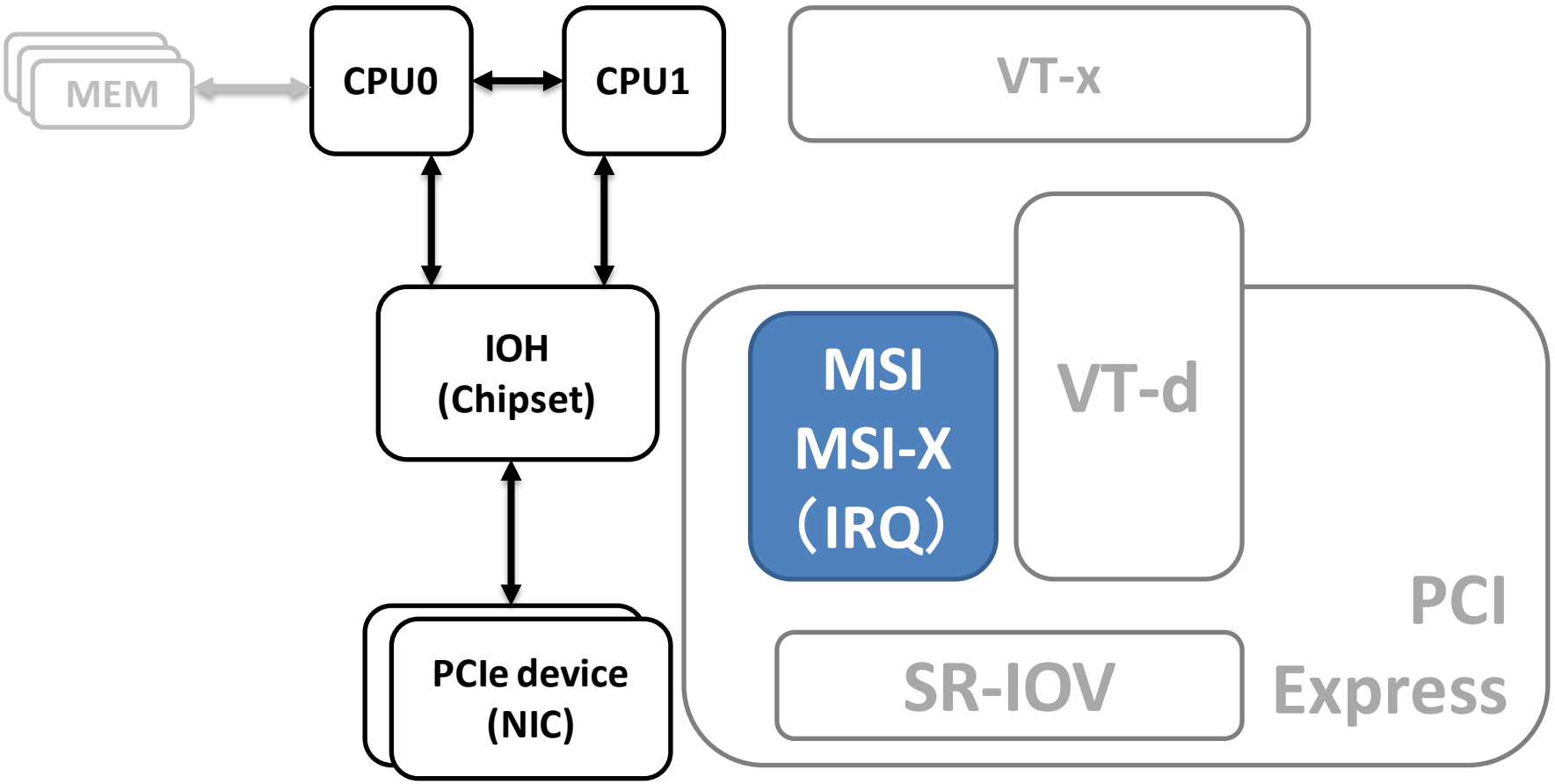
root権限必要

[ebiken@iwpf01 ~]$ ***sudo lspci -v -s 03:00.0***

03:00.0 Ethernet controller: Intel Corporation 82599EB 10-Gigabit SFI/SFP+ Network Connection (rev 01)

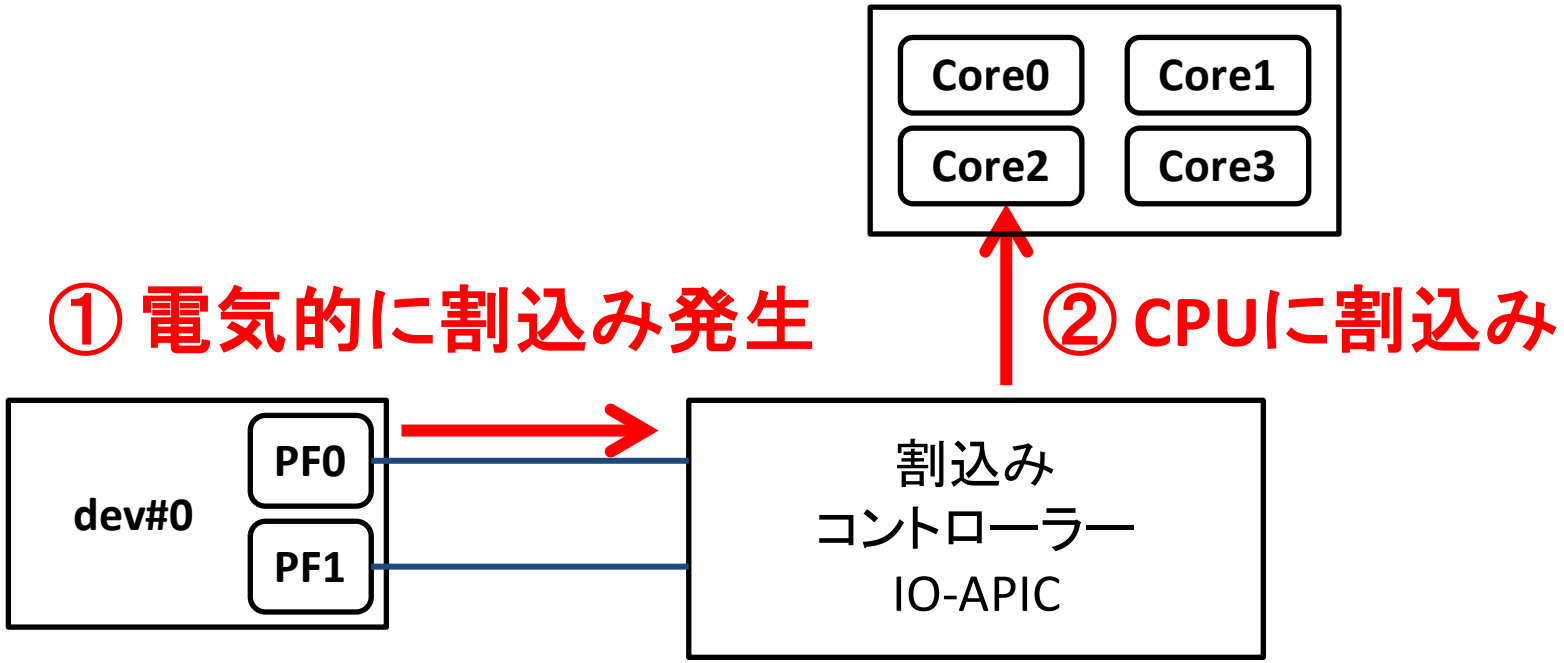
Subsystem: Intel Corporation Ethernet Server Adapter X520-2
Flags: bus master, fast devsel, latency 0, IRQ 38
Memory at d9900000 (64-bit, prefetchable) [size=512K]
I/O ports at fcc0 [size=32]
Memory at d98f8000 (64-bit, prefetchable) [size=16K]
Capabilities: [40] Power Management version 3
Capabilities: [50] MSI: Enable- Count=1/1 Maskable+ 64bit+
Capabilities: [70] MSI-X: Enable+ Count=64 Masked-
Capabilities: [a0] Express Endpoint, MSI 00
Capabilities: [100] Advanced Error Reporting
Capabilities: [140] Device Serial Number 00-1b-21-ff-ff-74-5a-f8
Capabilities: [150] Alternative Routing-ID Interpretation (ARI)
Capabilities: [160] Single Root I/O Virtualization (SR-IOV)
Kernel driver in use: ixgbe
Kernel modules: ixgbe

割込み：MSI / MSI-X


MEM
CPU0
CPU1
VT-x
IOH
(Chipset)
MSI
MSI-X
（IRQ）
VT-d
PCI
Express
SR-IOV
PCIe device
(NIC)

割込み：Pin Based Interrupt(Legacy)

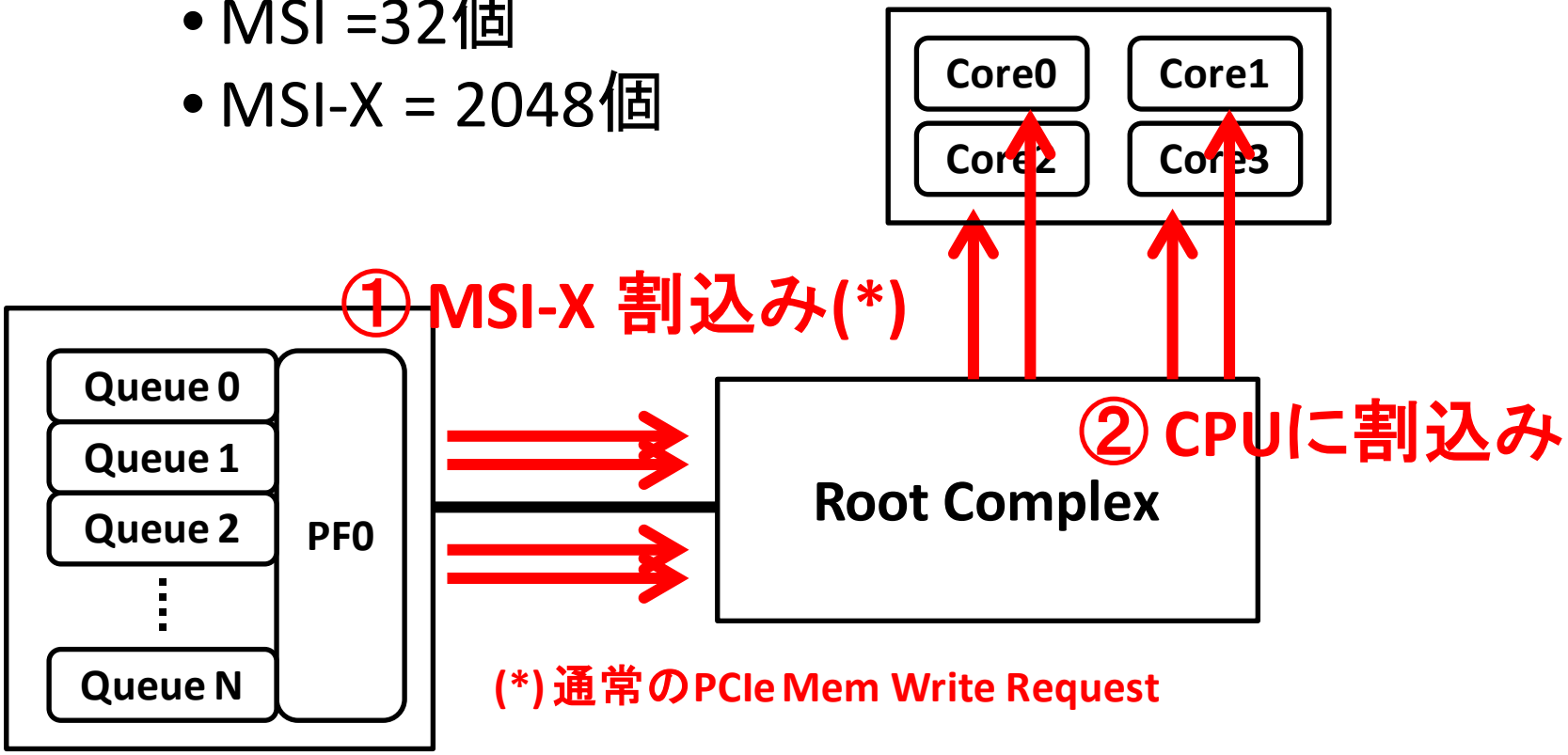
割込み線＝1ファンクション1本
Portで送受信した割り込みは同じCPU(Core)に割り込み



割込み : MSI / MSI-X

MSI = Message Signaled Interrupt
Queue毎に割込むCPU(Core)を変更可能

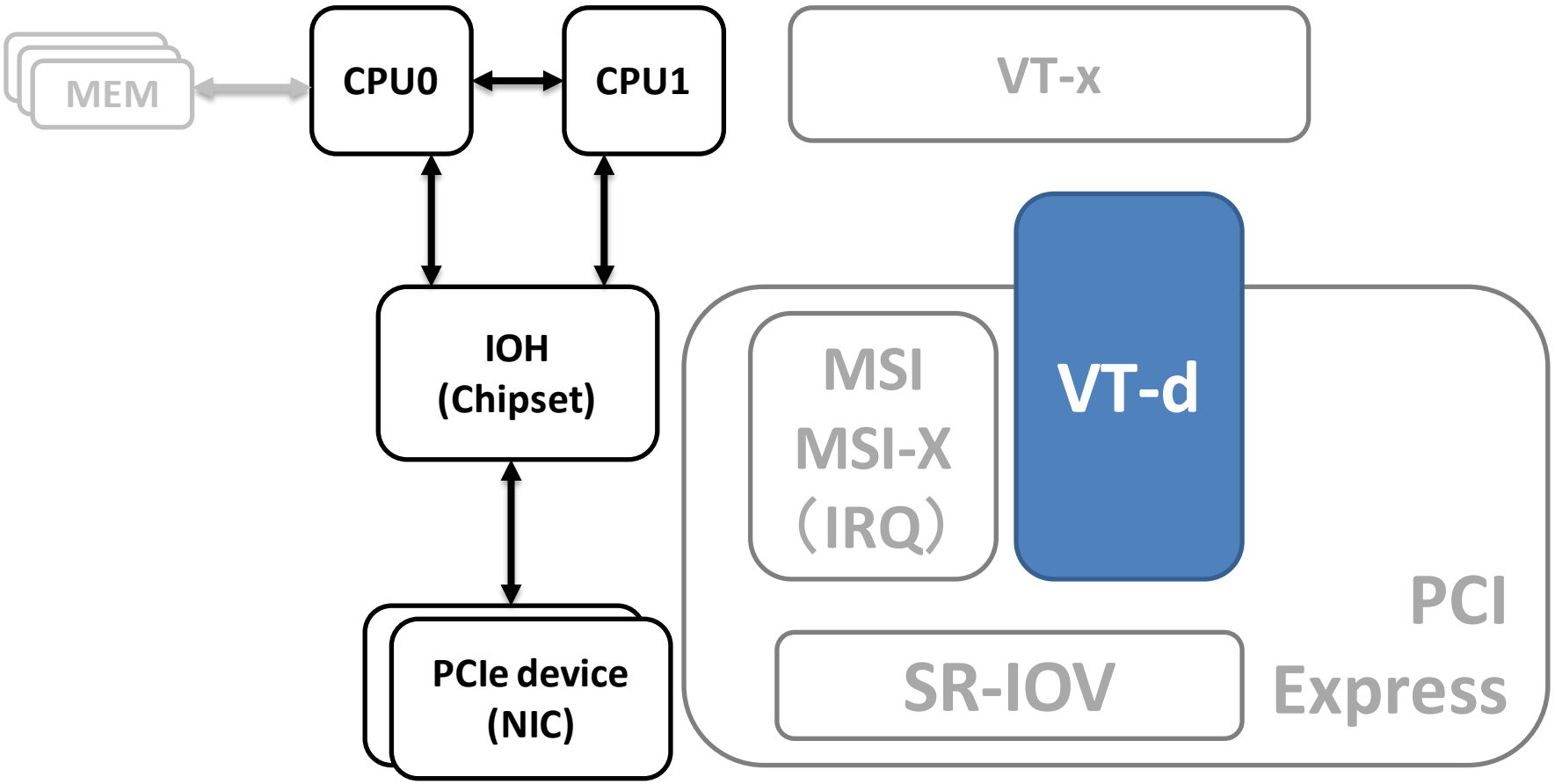
- MSI =32個
- MSI-X = 2048個



割込み：IRQ, MSI/MSI-X

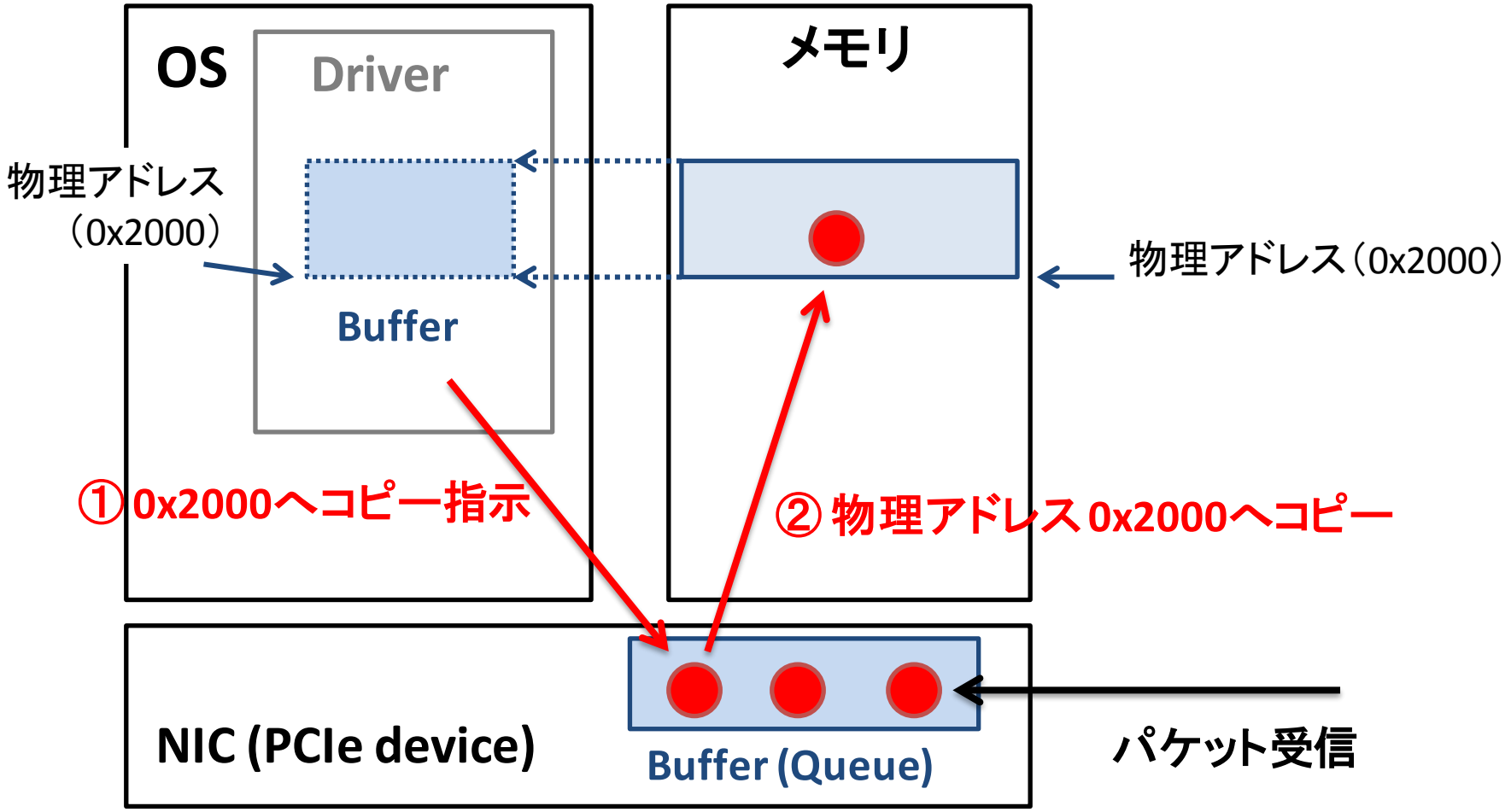
```
[ebiken@iwpf01 ~]$ cat /proc/interrupts
 CPU0   CPU1   CPU2   CPU3   CPU4   CPU5   CPU6   CPU7
... snip ...
  83: 153048      0      0      0      0      0      0      0   PCI-MSI-edge  p1p1-TxRx-0
  84: 158095     34      0      0      0      0      0      0   PCI-MSI-edge  p1p1-TxRx-1
  85:  88240      0     34      0      0      0      0      0   PCI-MSI-edge  p1p1-TxRx-2
  86: 156169      0      0     44      0      0      0      0   PCI-MSI-edge  p1p1-TxRx-3
  87: 155602      0      0      0     34      0      0      0   PCI-MSI-edge  p1p1-TxRx-4
  88: 168235      0      0      0      0     34      0      0   PCI-MSI-edge  p1p1-TxRx-5
  89: 195289      0      0      0      0      0     34      0   PCI-MSI-edge  p1p1-TxRx-6
  90: 163879      0      0      0      0      0      0     34   PCI-MSI-edge  p1p1-TxRx-7
  91:      4      0      0      0      0      0      0      0   PCI-MSI-edge  p1p1:lsc
  92: 171226      0      0      0      0      0      0      0   PCI-MSI-edge  p1p2-TxRx-0
  93: 196258     34      0      0      0      0      0      0   PCI-MSI-edge  p1p2-TxRx-1
  94: 175363      0     34      0      0      0      0      0   PCI-MSI-edge  p1p2-TxRx-2
  95:     24 878678      0     44      0      0      0      0   PCI-MSI-edge  p1p2-TxRx-3
  96: 172140      0      0      0     34      0      0      0   PCI-MSI-edge  p1p2-TxRx-4
  97: 152054      0      0      0      0     34      0      0   PCI-MSI-edge  p1p2-TxRx-5
  98: 163481      0      0      0      0      0     34      0   PCI-MSI-edge  p1p2-TxRx-6
  99:  88243      0      0      0      0      0      0     34   PCI-MSI-edge  p1p2-TxRx-7
 100:      4      0      0      0      0      0      0      0   PCI-MSI-edge  p1p2:lsc
```

Virtualization Technology for directed I/O

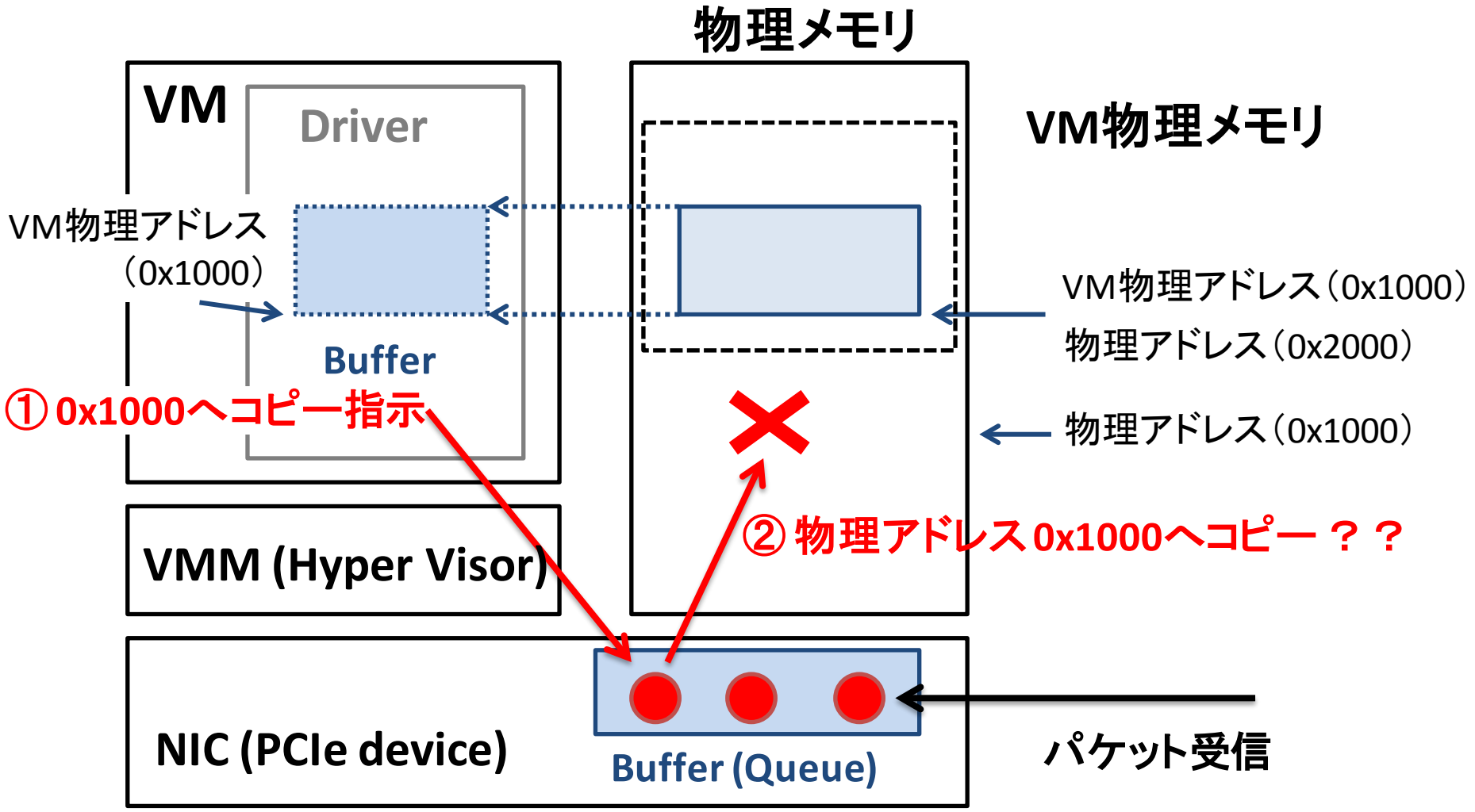


DMA : Direct Memory Access

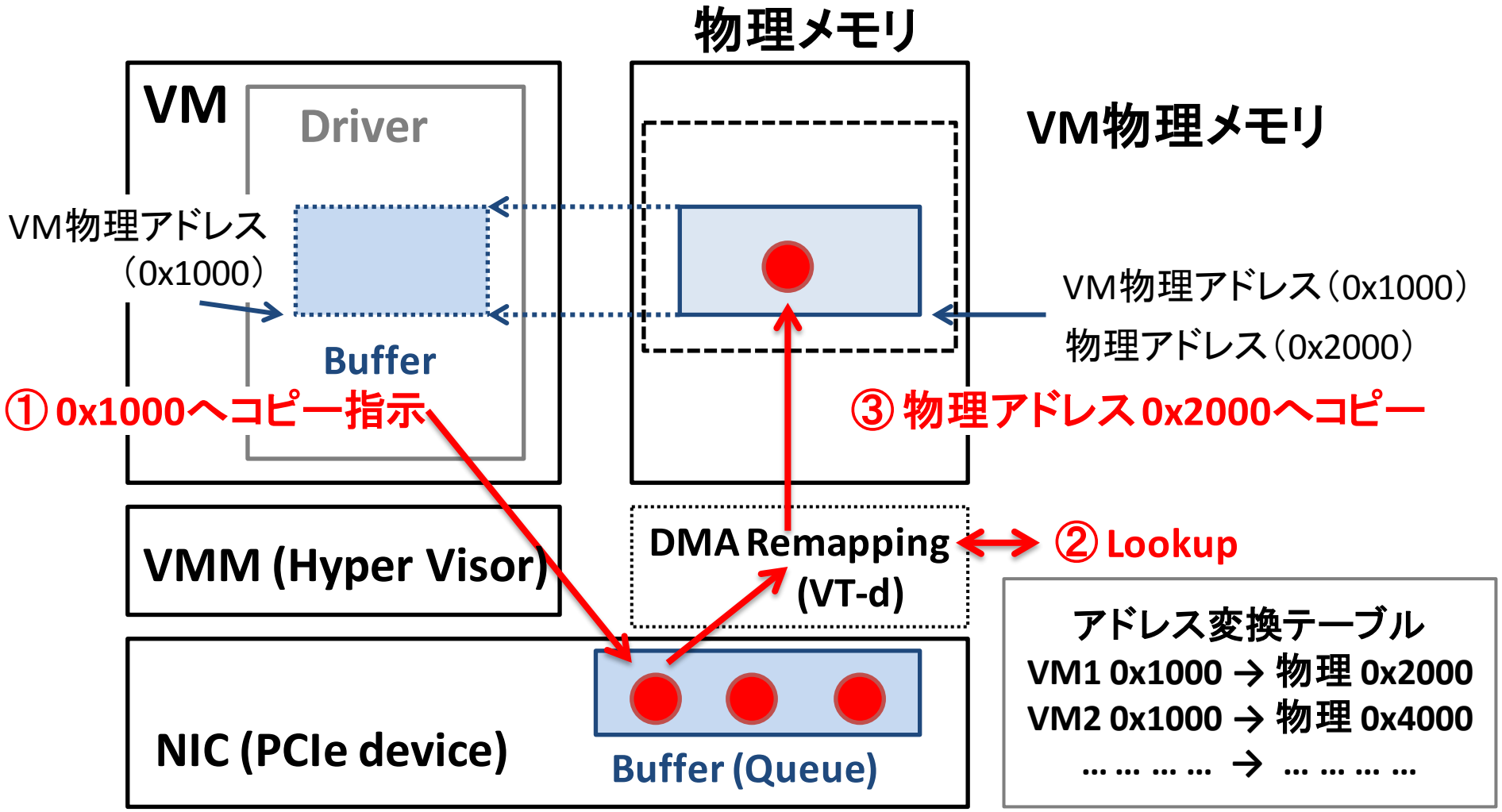
CPUを介さずパケットをコピー



仮想環境（VM）の場合


物理メモリ
VM
Driver
VM物理メモリ
VM物理アドレス
（0x1000）
Buffer
① 0x1000へコピー指示
VM物理アドレス（0x1000）
物理アドレス（0x2000）
物理アドレス（0x1000）
② 物理アドレス0x1000へコピー？？
VMM (Hyper Visor)
NIC (PCIe device)
Buffer (Queue)
パケット受信

仮想環境（VM）＋ VT-d サポート

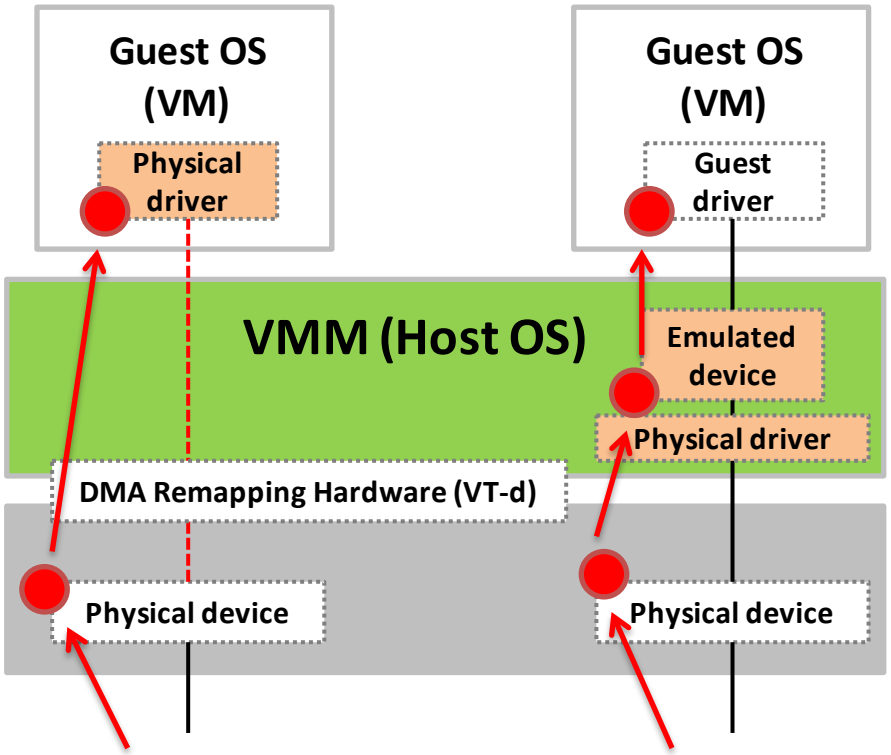


VT-d によるI/O デバイス割り当て

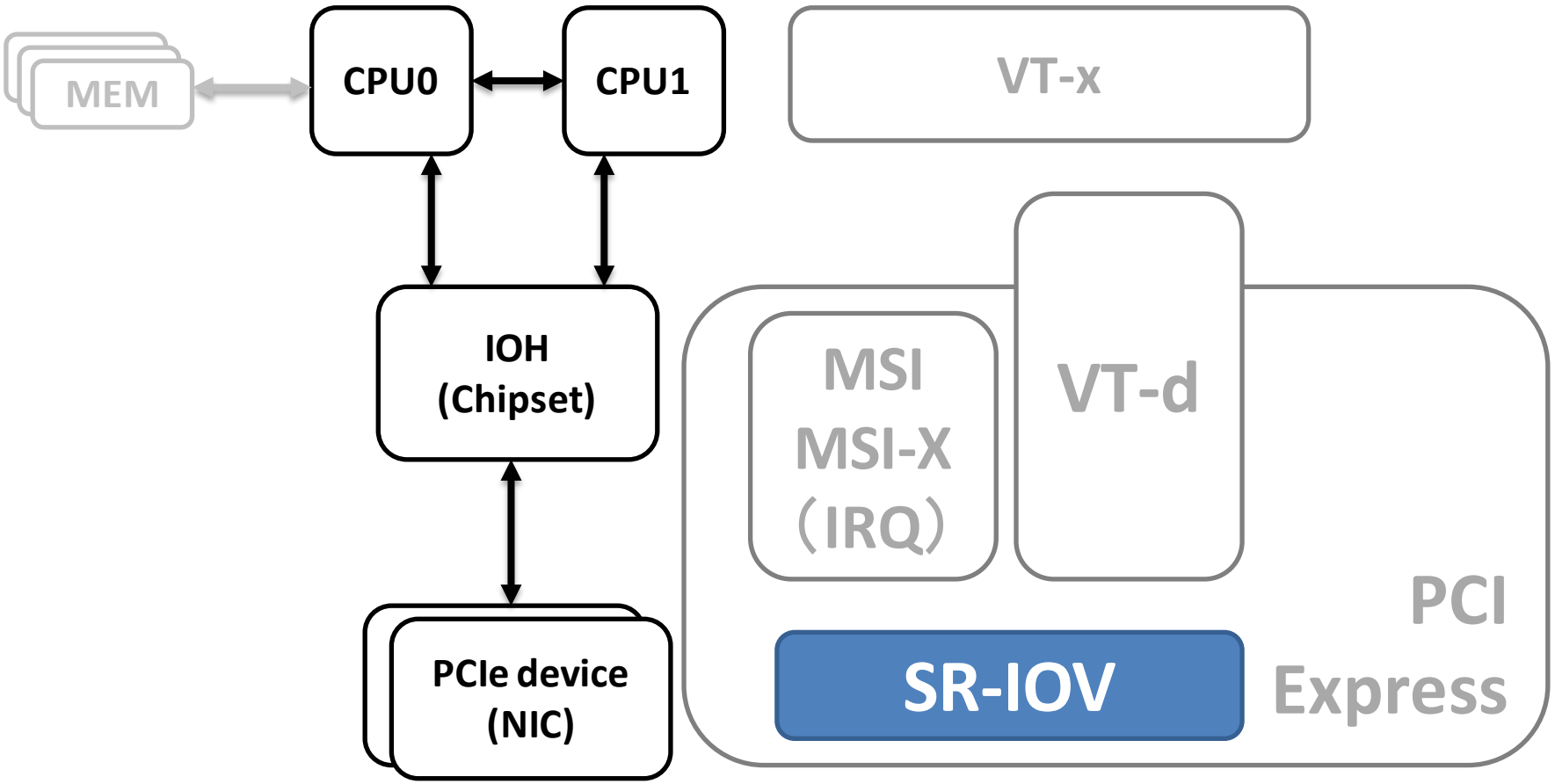
- 物理デバイス（NIC）を VMに割り当て
- VMから直接見える（割込みのみVMM経由）
- パケットをDMA転送
- オーバーヘッド＝極小

1ポート（※）に1VMのみ

（※）1 PCIe Function



SR-IOV : Single Root - IO Virtualization



SR-IOV & Virtual Function

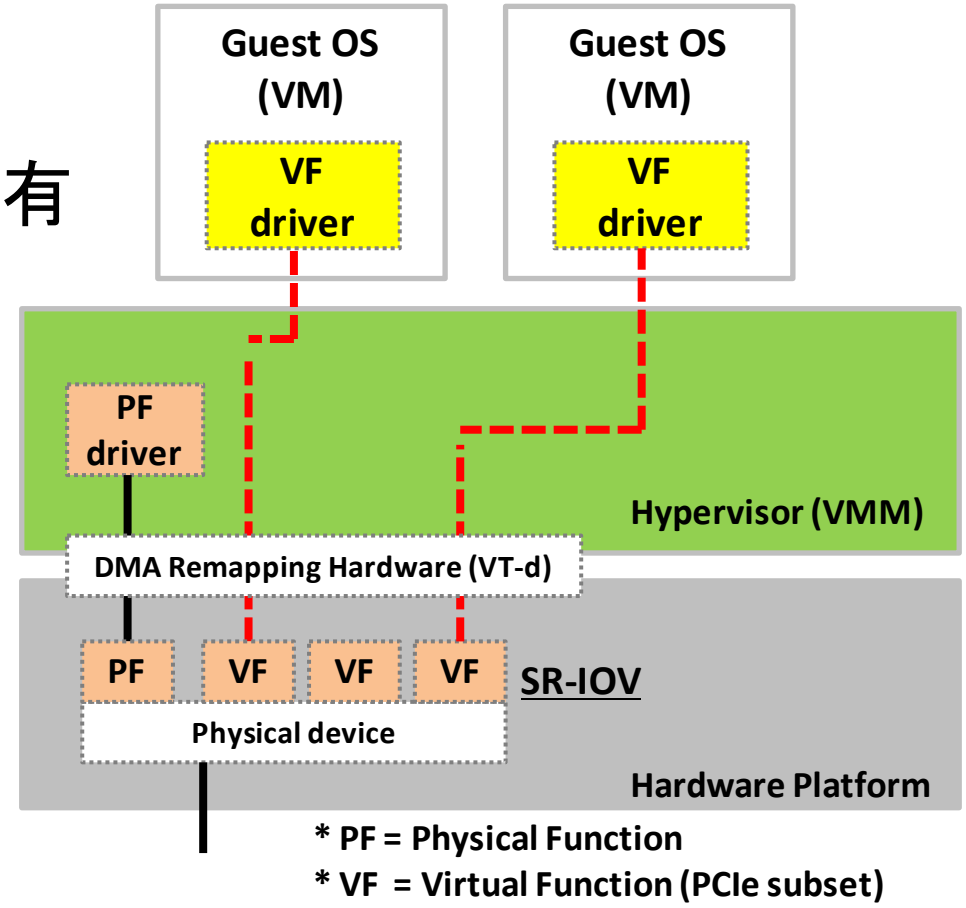
- SR-IOV = Single Root - IO Virtualization
- PCIe Device (NIC) 機能
- Virtual Function = Physical Functionのサブセット
  - 設定はPhysical Function経由
  - データ送受信は Virtual Function <-> Driver 直接

I/O デバイス割り当て＋共有（SR-IOV）

物理デバイス（NIC）を VMに割り当て＋共有
オーバーヘッド＝極小

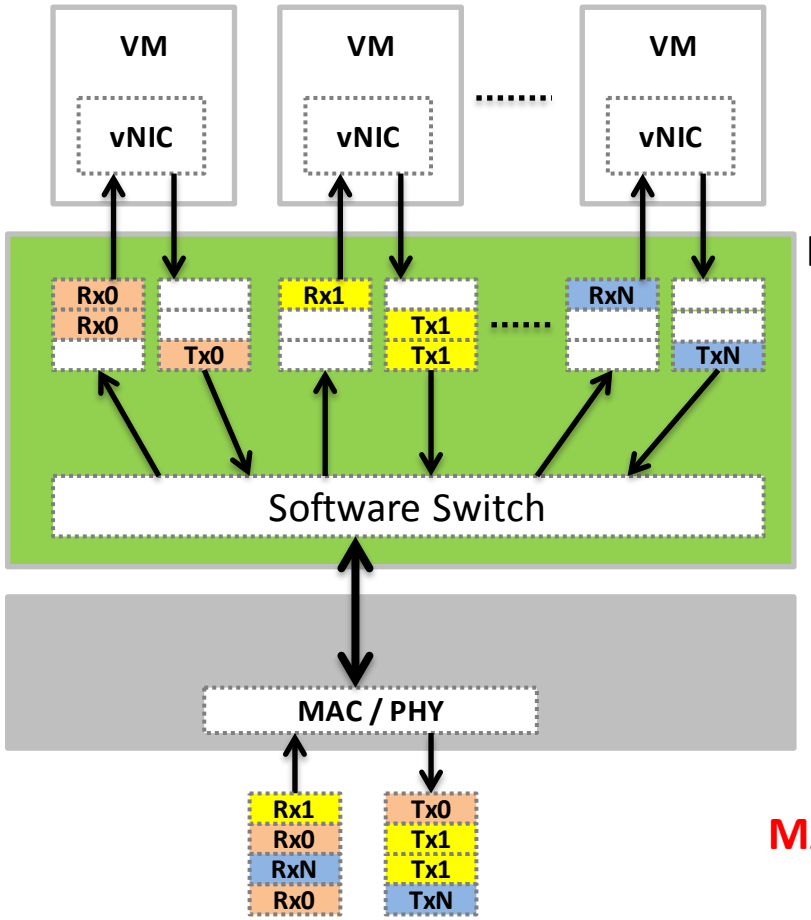
複数VMでの物理ポート共有

VMからは VF=NIC(Port)
に見える

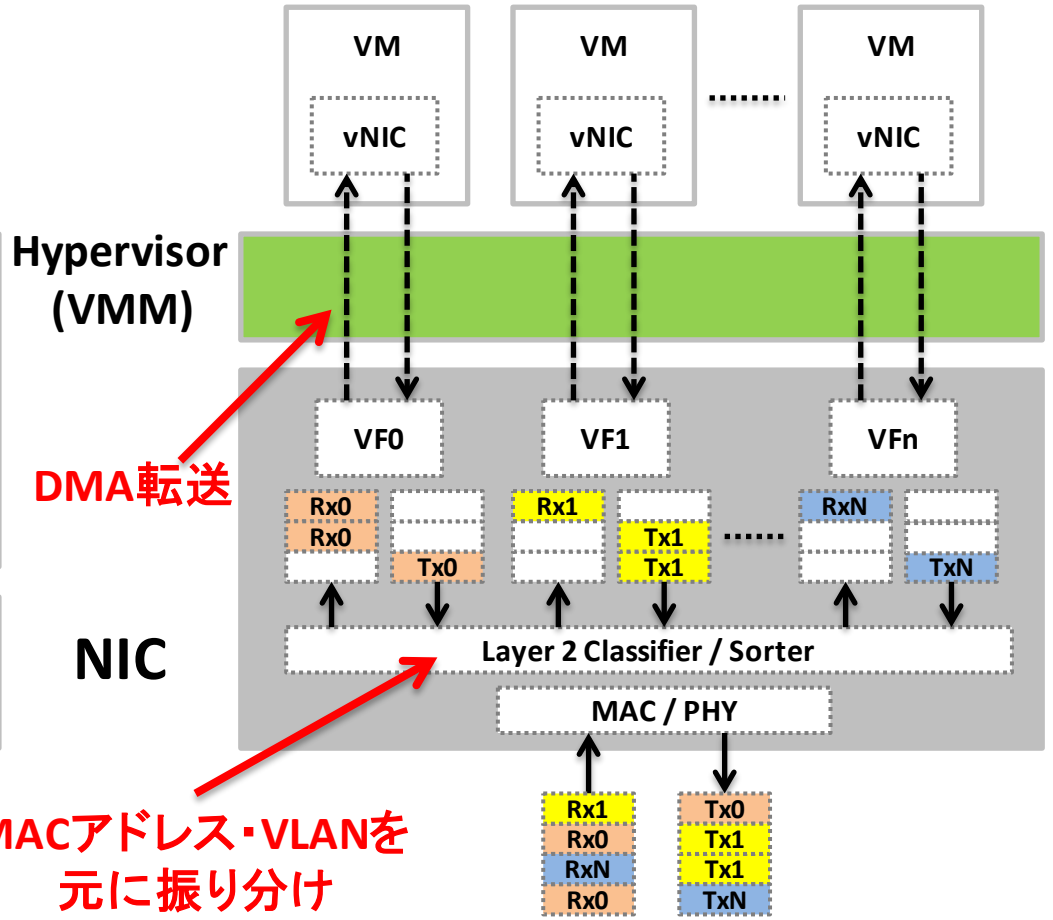


SR-IOV パケットフロー

SR-IOV 無し : VMMがパケット振り分け



SR-IOV 有り : NICがパケット振り分け
VMMはパス・スルー



SR-IOV : lspci -tv @ VMM (Host)

```
[ebiken@iwpf01 ~]$ lspci -tv
 ¥-[0000:00]-+-00.0  Intel Corporation 5500 I/O Hub to ESI Port
             +-01.0-[01]--+-00.0  Broadcom Corporation NetXtreme II BCM5716 Gigabit Ethernet
             |            ¥-00.1  Broadcom Corporation NetXtreme II BCM5716 Gigabit Ethernet
             +-03.0-[02]--
             +-07.0-[03]--+-00.0  Intel Corporation 82599EB 10-Gigabit SFI/SFP+ Network Connection
             |            ¥-00.1  Intel Corporation 82599EB 10-Gigabit SFI/SFP+ Network Connection
             +-14.0  Intel Corporation 5520/5500/X58 I/O Hub System Management Registers

[ebiken@iwpf01 ~]$ sudo modprobe -r ixgbe
[ebiken@iwpf01 ~]$ sudo modprobe ixgbe max_vfs=2,2

[ebiken@iwpf01 ~]$ lspci -tv
 ¥-[0000:00]-+-00.0  Intel Corporation 5500 I/O Hub to ESI Port
             +-01.0-[01]--+-00.0  Broadcom Corporation NetXtreme II BCM5716 Gigabit Ethernet
             |            ¥-00.1  Broadcom Corporation NetXtreme II BCM5716 Gigabit Ethernet
             +-03.0-[02]--
             +-07.0-[03]--+-00.0  Intel Corporation 82599EB 10-Gigabit SFI/SFP+ Network Connection
             |            +-00.1  Intel Corporation 82599EB 10-Gigabit SFI/SFP+ Network Connection
             |            +-10.0  Intel Corporation 82599 Ethernet Controller Virtual Function
             |            +-10.1  Intel Corporation 82599 Ethernet Controller Virtual Function
             |            +-10.2  Intel Corporation 82599 Ethernet Controller Virtual Function
             |            ¥-10.3  Intel Corporation 82599 Ethernet Controller Virtual Function
```

SR-IOV設定

BUS [03] に Device 10 (Virtual Function) 生成

SR-IOV : lspci -tv @ VM (Guest)

```
[root@dutvm-sriov ~]# lspci -tv
-[0000:00]-+-00.0  Intel Corporation 440FX - 82441FX PMC [Natoma]
           +-01.0  Intel Corporation 82371SB PIIX3 ISA [Natoma/Triton II]
           +-01.1  Intel Corporation 82371SB PIIX3 IDE [Natoma/Triton II]
           +-01.2  Intel Corporation 82371SB PIIX3 USB [Natoma/Triton II]
           +-01.3  Intel Corporation 82371AB/EB/MB PIIX4 ACPI
           +-02.0  Cirrus Logic GD 5446
           +-03.0  Intel Corporation 82559 Ethernet Controller Virtual Function
           +-04.0  Red Hat, Inc Virtio block device
           +-05.0  Red Hat, Inc Virtio memory balloon
           \-06.0  Intel Corporation 82559 Ethernet Controller Virtual Function
```

BUS [00] : device 06 . function 0

Intel Corporation 82559 Ethernet Controller Virtual Function

Virtual Functionが直接見えている

VF Capabilities @ VMM (Host)

PF依存情報は[virtual]

```
[ebiken@iwpf01 ~]$ sudo lspci -v -s 3:10.0
03:10.0 Ethernet controller: Intel Corporation 82599 Ethernet Controller Virtual Function (rev 01)
        Subsystem: Intel Corporation Device 0003
        Flags: bus master, fast devsel, latency 0
        [virtual] Memory at c0000000 (64-bit, non-prefetchable) [size=16K]
        [virtual] Memory at c0100000 (64-bit, non-prefetchable) [size=16K]
        Capabilities: [70] MSI-X: Enable+ Count=3 Masked-
        Capabilities: [a0] Express Endpoint, MSI 00
        Capabilities: [100] Advanced Error Reporting
        Capabilities: [150] Alternative Routing-ID Interpretation (ARI)
        Kernel driver in use: ixgbevf
        Kernel modules: ixgbevf
```

Capabilitiesに SR-IOV無し

driver, modules は ixgbevf (not ixgbe)

VF Capabilities ＠ VM（Guest）

PCI Physical Function として見えている

([virtual] 無し)

```
[root@dutvm-sriov ~]# lspci -v -s 00:06.0
00:06.0 Ethernet controller: Intel Corporation 82559 Ethernet Controller Virtual
 Function (rev 01)
        Subsystem: Intel Corporation Device 0003
        Physical Slot: 6
        Flags: bus master, fast devsel, latency 0
        Memory at f202c000 (32-bit, non-prefetchable) [size=16K]
        Memory at f2030000 (32-bit, non-prefetchable) [size=16K]
        Capabilities: [a0] Express Endpoint, MSI 00
        Capabilities: [70] MSI-X: Enable+ Count=3 Masked-
        Kernel driver in use: ixgbevf
        Kernel modules: ixgbevf
```

VM側でもixgbevf 使用

CapabilitiesにPF機能無し

[100] Advanced Error Reporting

[150] Alternative Routing-ID Interpretation (ARI)

/proc/interrupts @ VMM (Host)

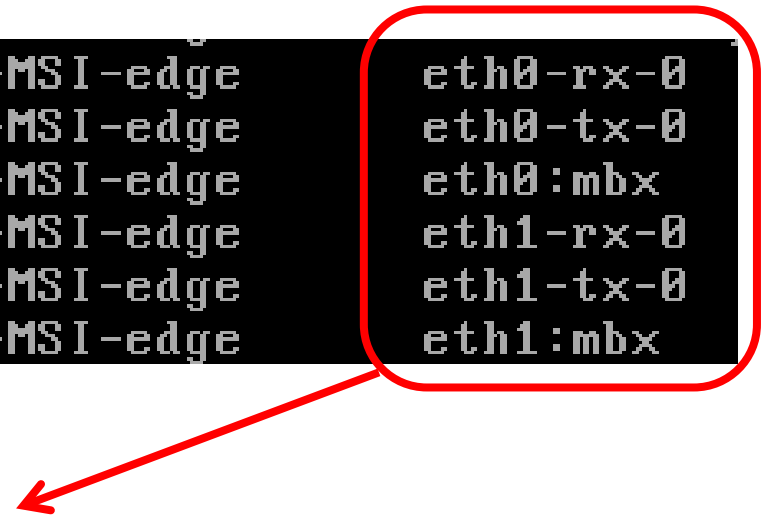
```
[ebiken@iwpf01 ~]$ cat /proc/interrupts
           CPU0       CPU1       CPU2       CPU3       CPU4       CPU5       CPU6       CPU7
... snip ...
  84:          0          0          0          0          0          0          0          0   PCI-MSI-edge p1p1-TxRx-0
  85:          0          0          0          0          0          0          0          0   PCI-MSI-edge p1p1
  86:          0          0          0          0          0          0          0          0   PCI-MSI-edge p1p2-TxRx-0
  87:          0          0          0          0          0          0          0          0   PCI-MSI-edge p1p2

  88:          0          0          0          0          0          0          0          0   PCI-MSI-edge kvm:0000:03:10.0
  89:          0          0          0          0          0          0          0          0   PCI-MSI-edge kvm:0000:03:10.0
  90:          0          0          0          0          0          0          0          0   PCI-MSI-edge kvm:0000:03:10.0
  94:          0          0          0          0          0          0          0          0   PCI-MSI-edge kvm:0000:03:10.1
  95:          0          0          0          0          0          0          0          0   PCI-MSI-edge kvm:0000:03:10.1
  96:          0          0          0          0          0          0          0          0   PCI-MSI-edge kvm:0000:03:10.1
```

PFの割込み

VF割込み

/proc/interrupts @ VM（Guest）

```
 42:          4          7   PCI-MSI-edge      eth0-rx-0
 43:          5         12   PCI-MSI-edge      eth0-tx-0
 44:          4          4   PCI-MSI-edge      eth0:mbx
 45:          8          3   PCI-MSI-edge      eth1-rx-0
 46:         11          6   PCI-MSI-edge      eth1-tx-0
 47:          4          3   PCI-MSI-edge      eth1:mbx
```

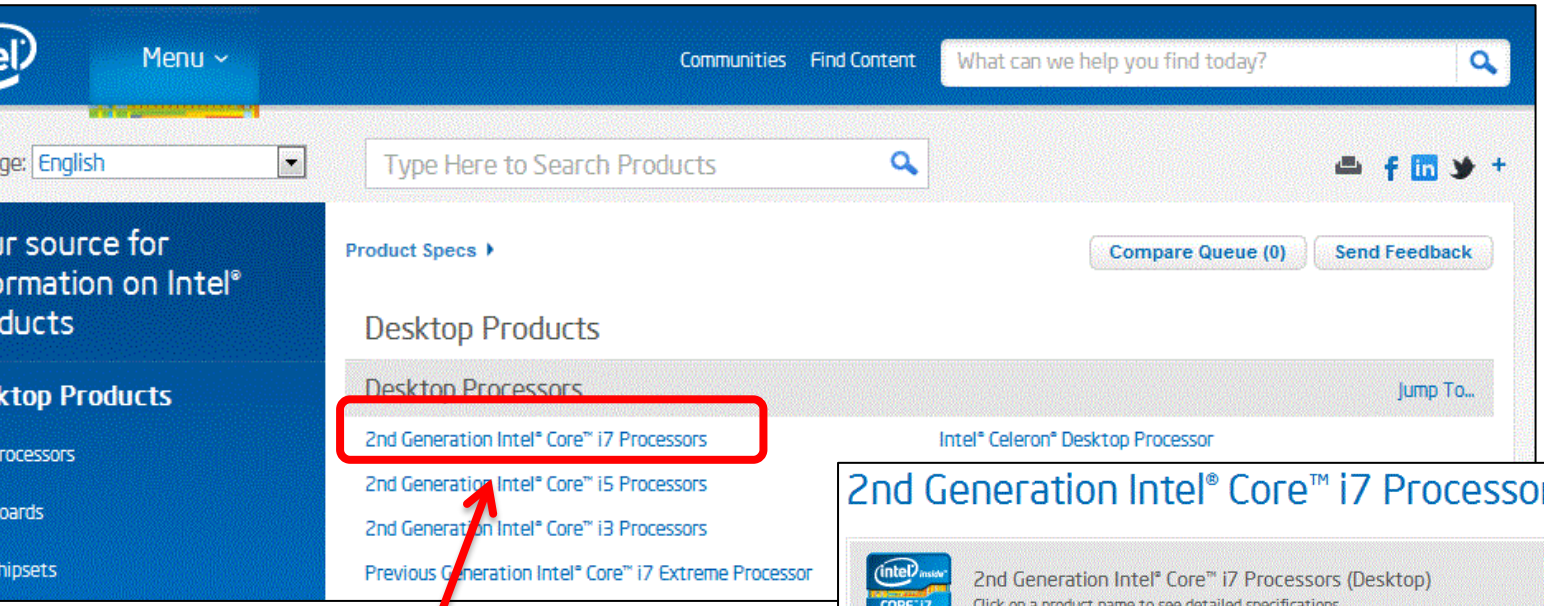


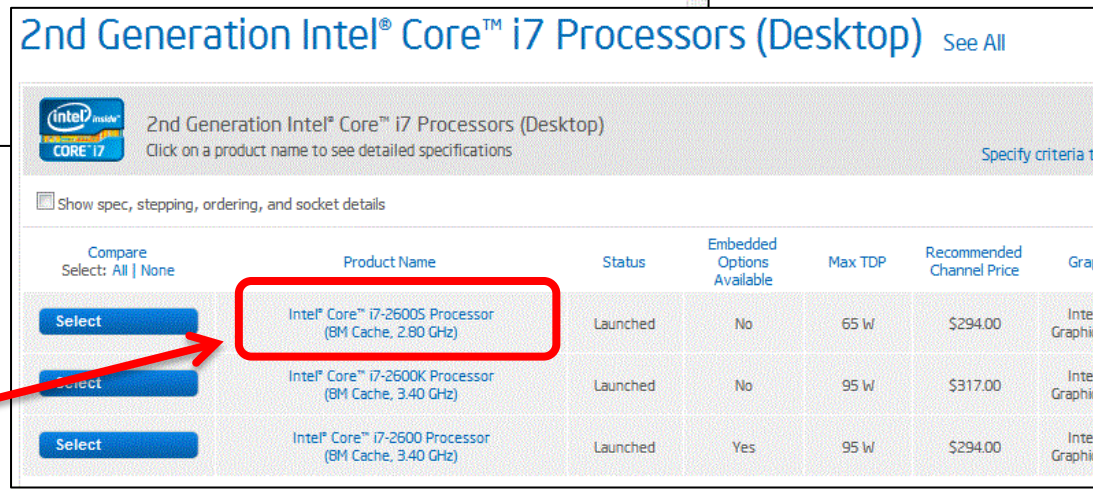
- eth0, eth1 として認識
- eth0:mbx, eth1:mbx = PF/VF間通信に使用

VT-x, VT-d, SR-IOV
サポート確認方法

Intel CPU VT-x, VT-d サポート確認方法

http://ark.intel.com/

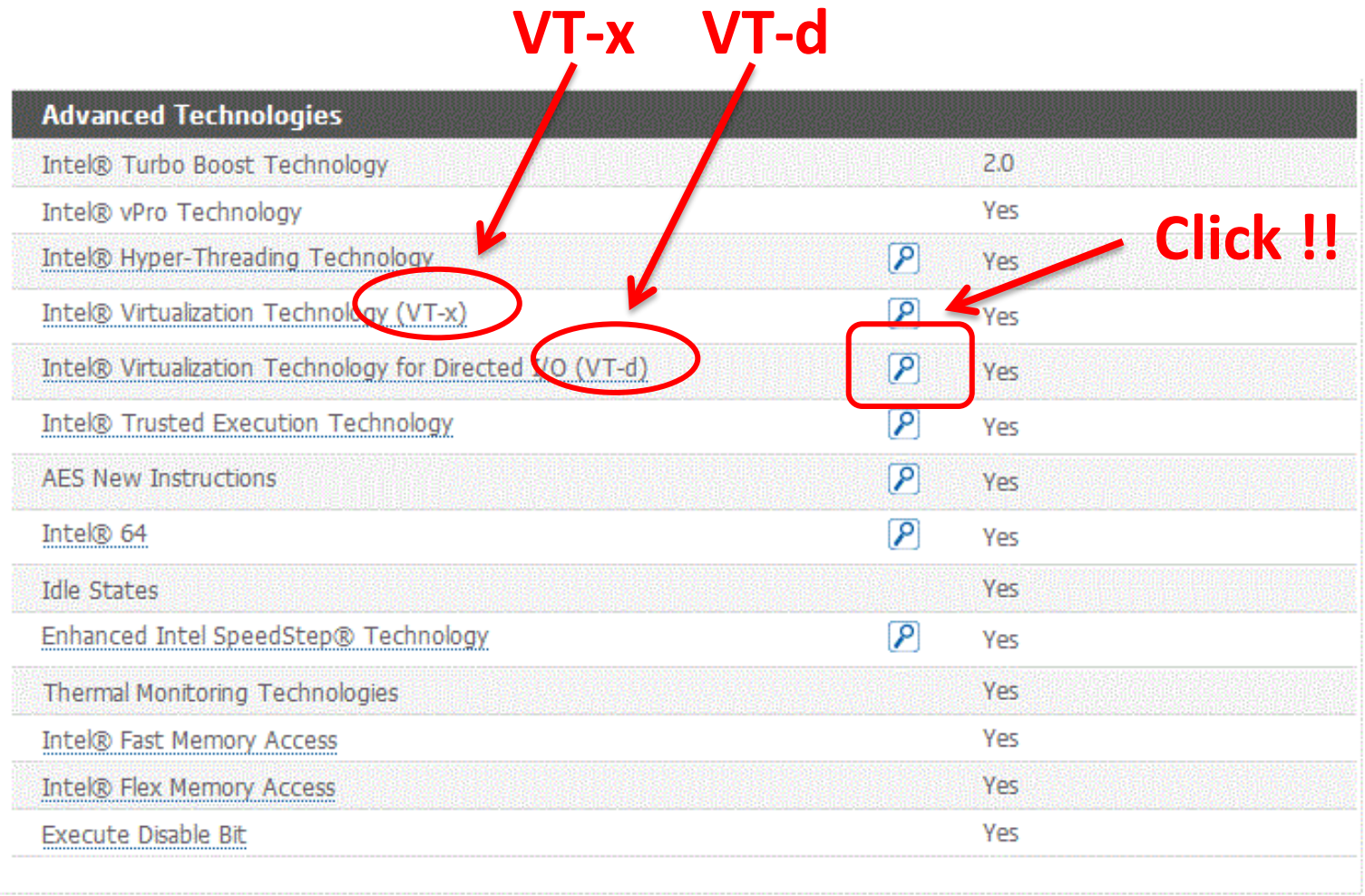




1) Click !!

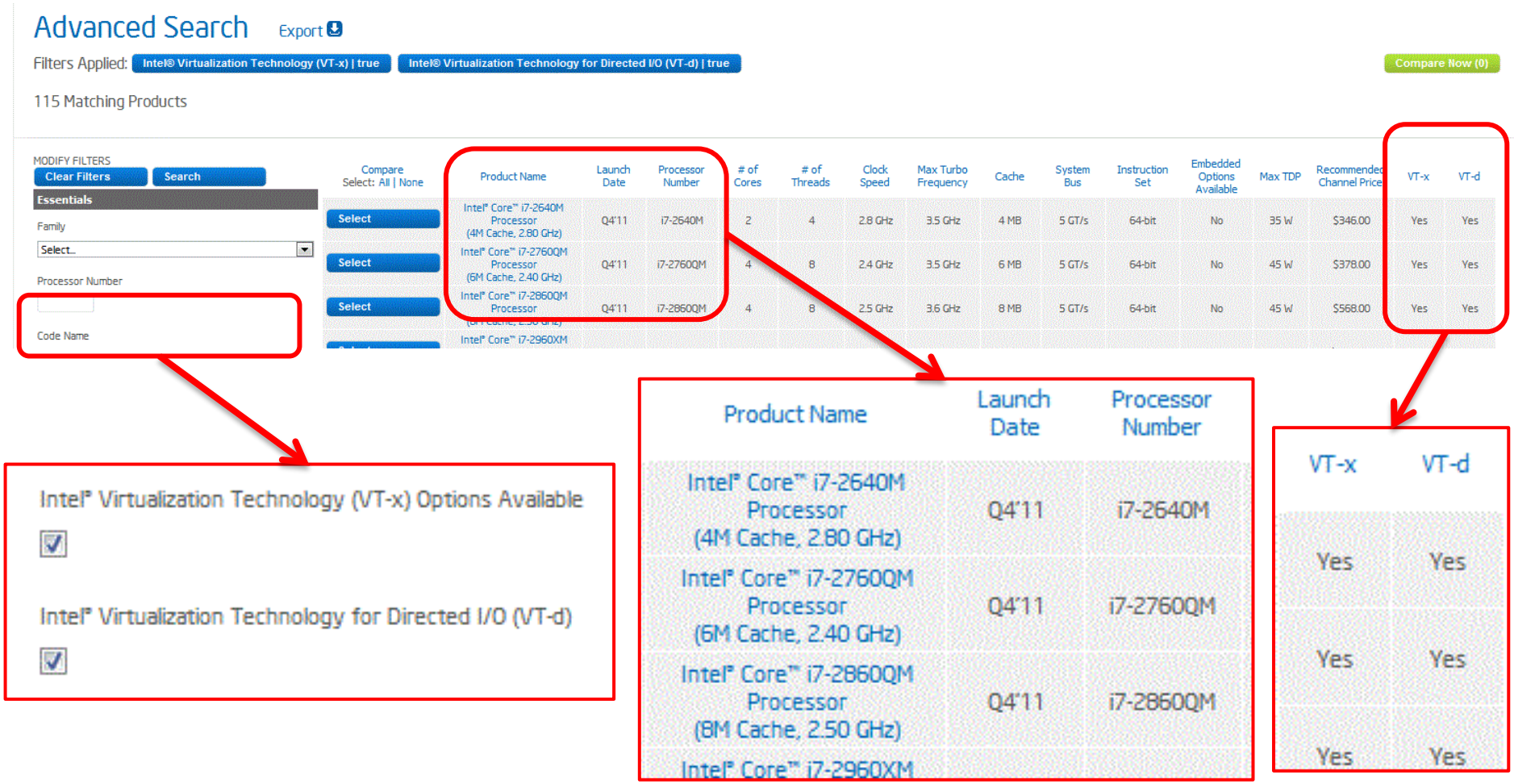
2) Click !!

Intel CPU VT-x, VT-d サポート確認方法



Intel CPU VT-x, VT-d サポート確認方法

http://ark.intel.com/search/advanced/?s=t&VTX=true&VTD=true



SR-IOVサポートする Intel NIC一覧

http://www.intel.com/support/network/adapter/pro100/sb/CS-031492.htm

Intel® Server Adapters
FAQs: Using SR-IOV with Intel® Ethernet Server Adapters

Which Intel® Ethernet Adapters support SR-IOV?

- Intel® Ethernet Server Adapter X520-DA2
- Intel® Ethernet Server Adapter X520-SR1
- Intel® Ethernet Server Adapter X520-SR2
- Intel® Ethernet Server Adapter X520-LR1
- Intel® Ethernet Server Adapter X520-T2
- Intel® Gigabit ET Dual Port Server Adapter
- Intel® Gigabit EF Dual Port Server Adapter
- Intel® Gigabit ET2 Quad Port Server Adapter

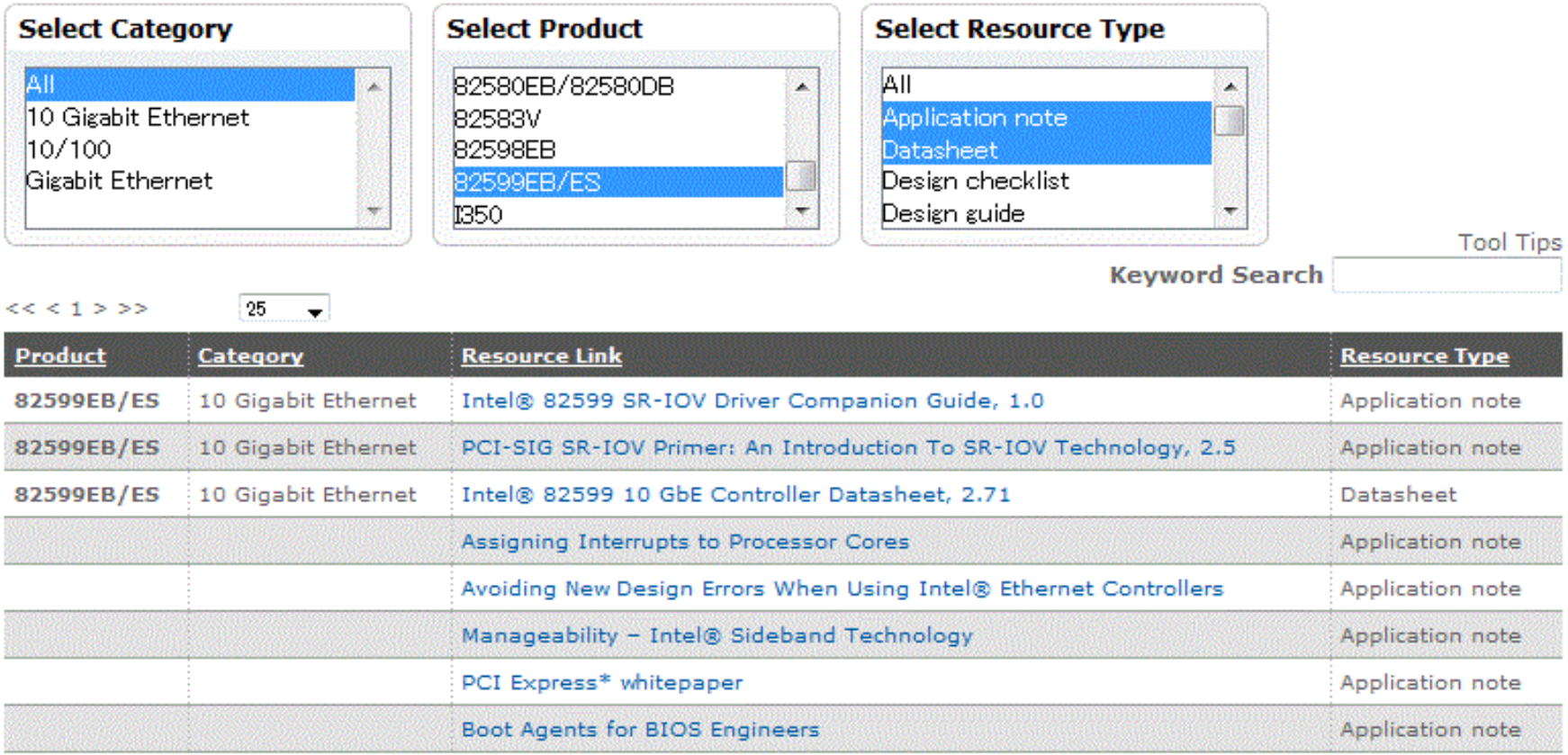
Which hypervisors support SR-IOV on Intel® Ethernet Adapters?
Xen Hypervisor*
KVM* (Kernel Based Virtual Machine)

Which operating systems and distributions include virtual function driver support for guests?
Virtual Function (VF) drivers for guest operating systems are available for:

- Windows Server 2008 R2*
- Windows Server 2008*, 32-bit and 64-bit
- Linux* 2.6.30 kernel or later
- Red Hat Enterprise Linux* 5.4
- Red Hat Enterprise Linux 5.5
- SUSE Linux Enterprise Server* 11 SP1

Intel NIC VT-d, SR-IOV サポート確認方法

http://www.intel.com/products/ethernet/resource.htm
Controllerの型番からデータシートや参考資料が入手可能



Product	Category	Resource Link	Resource Type
82599EB/ES	10 Gigabit Ethernet	Intel® 82599 SR-IOV Driver Companion Guide, 1.0	Application note
82599EB/ES	10 Gigabit Ethernet	PCI-SIG SR-IOV Primer: An Introduction To SR-IOV Technology, 2.5	Application note
82599EB/ES	10 Gigabit Ethernet	Intel® 82599 10 GbE Controller Datasheet, 2.71	Datasheet
		Assigning Interrupts to Processor Cores	Application note
		Avoiding New Design Errors When Using Intel® Ethernet Controllers	Application note
		Manageability – Intel® Sideband Technology	Application note
		PCI Express* whitepaper	Application note
		Boot Agents for BIOS Engineers	Application note

参考文献

- Books
  - PCI Express 設計の基礎と応用（CQ出版社）
    - ISBN 978-4-7898-4641-7
  - PCI Express System Architecture (MINDSHARE)
    - ISBN 978-0-321-15630-3
  - Linux デバイスドライバ プログラミング（ソフトバンククリエイティブ株式会社）
    - ISBN 978-4-7973-4642-8
  - Understanding The Linux Kernel (O'REILLY)
    - ISBN 978-0-596-00565-8
- PCI-SIG
  - PCI Express Base Specification, Revision 2.1, March 4, 2009
  - Single Root I/O Virtualization and Sharing Specification, Revision 1.0, September 11, 2007
- Intel
  - Intel Virtualization Technology for Directed I/O, Architecture Specification, Revision: 1.3
  - Intel 82599 10 GbE Controller Datasheet, Revision Number: 2.7
  - Intel 82599 SR-IOV Driver Companion Guide, Revision 1.00, May 2010
  - PCI-SIG SR-IOV Primer, An Introduction to SR-IOV Technology, Revision 2.5, January 2011
  - Product Specs from http://ark.intel.com/

予備スライド