Empowered by Innovation NEC

第3回ネットワーク仮想化時限研究会

進化型ネットワーク仮想化基盤を実現するネットワーク処理計算ノードの提案

2012年3月2日

○神谷　聡史、狩野　秀一、孫　雷、百目木　智康(*)、河邉　岳彦(*)

日本電気株式会社、システムプラットフォーム研究所

(*) 日本電気通信システム

本研究の一部は，独立行政法人情報通信研究機構(NICT) の委託研究「新世代ネットワークを支えるネットワーク仮想化基盤技術の研究開発」(課題ア：統合管理型ネットワーク仮想化基盤技術の研究開発) によるものです。

内容

- 背景／進化型ネットワーク仮想化基盤
- Programmer：プログラマ -ネットワーク処理計算ノード-
- プログラマの課題
- 提案プログラマアーキテクチャ
- 要素技術検証
  - KVM環境でのvSwitch offloadの実現と評価
- まとめ

背景

高度ICT社会の実現と多様なネットワークサービスの実現

↓

ネットワークへの多種多様な要望に対応する
新世代通信基盤の実現に対する期待

↓

ネットワークへの要望は今後変遷
並行検証、随時実現が必要に

↓

進化型ネットワーク仮想化基盤

- 共通のネットワーク基盤を論理的に分割し、提供
- 各論理ネットワーク上で様々な通信処理を実施・検証

ネットワーク仮想化基盤

- ネットワーク内の資源(計算資源、リンク資源等)を仮想化
- 仮想化基盤利用者毎に、資源集合体:スライスを用意
- 利用者は各自スライス内で自ら考案のネットワーク方式を構築・検証

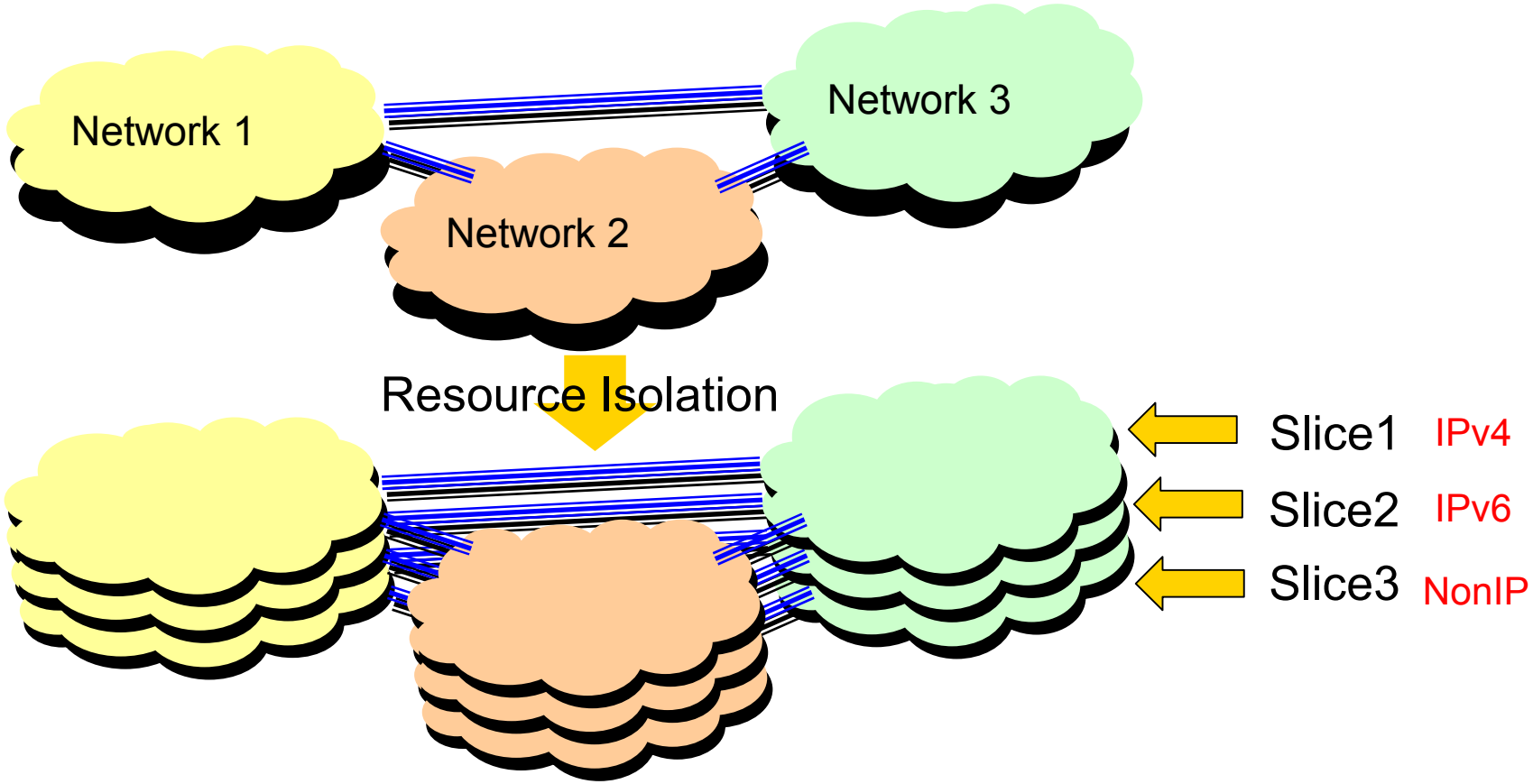
ネットワーク仮想化基盤の進展

1. ネットワーク仮想化基盤 (旧ネットワーク仮想化基盤)
   - NICT仮想化ノードプロジェクトにて研究開発
     (NICT, 東大, NTT, 日立, 富士通, NEC: 2008-2010).
   - NICT JGN-Xに展開済

2. 進化型ネットワーク仮想化基盤 (新ネットワーク仮想化基盤)
   - NICT委託研究「新世代ネットワークを支えるネットワーク仮想化基盤技術の研究開発」:課題ア にて現在研究開発中
     (NTT, 東大, 日立, 富士通, NEC: 2011-2014)

ネットワーク仮想化の概念：Slice [Nakao10]

Slice (スライス)
「ネットワーク全体で予約可能なコンピュータ・ネットワーク資源の集合体」



[Nakao10] A. Nakao, “Network Virtualization as Foundation for Enabling New Network Architecture and Application,” IEICE Trans. Commun., Vol E93-B, No. 3, pp.454-457, March 2010.

（旧）ネットワーク仮想化基盤

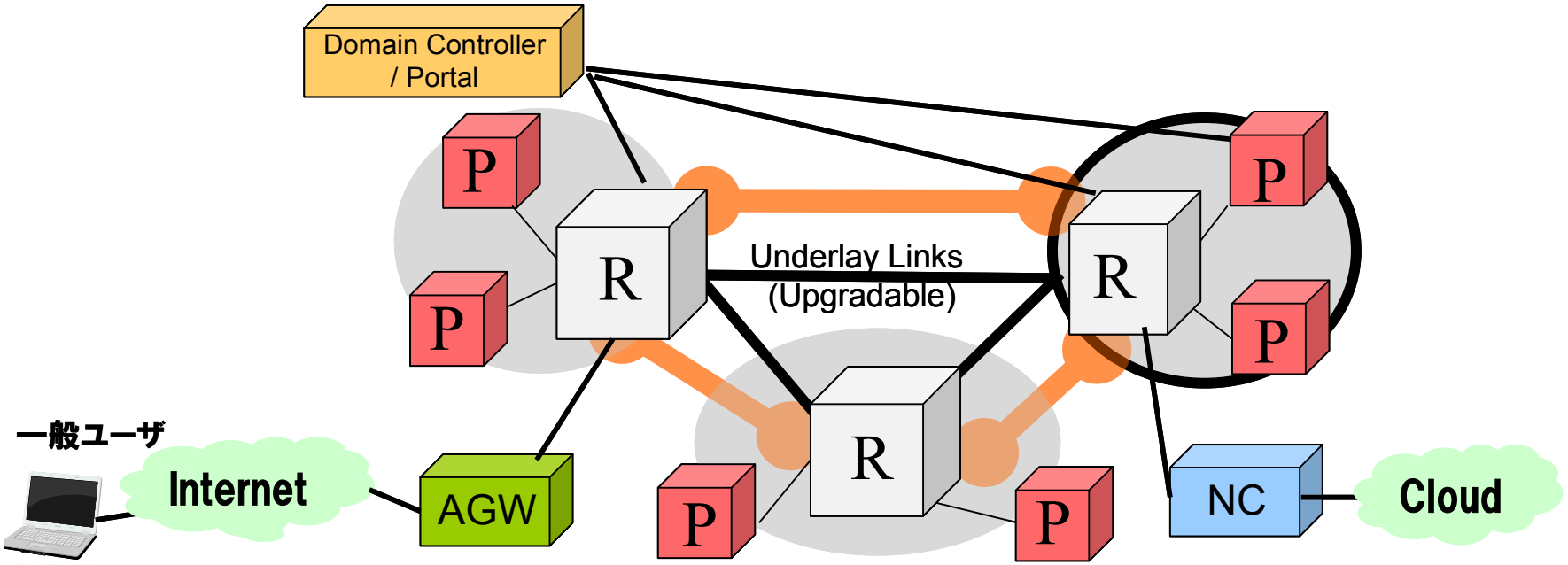
資源管理システム、資源提供機器、利用者アクセス機器より構成

管理プレーン

- Domain Controller (DC), Portal, **各データプレーン機器管理**

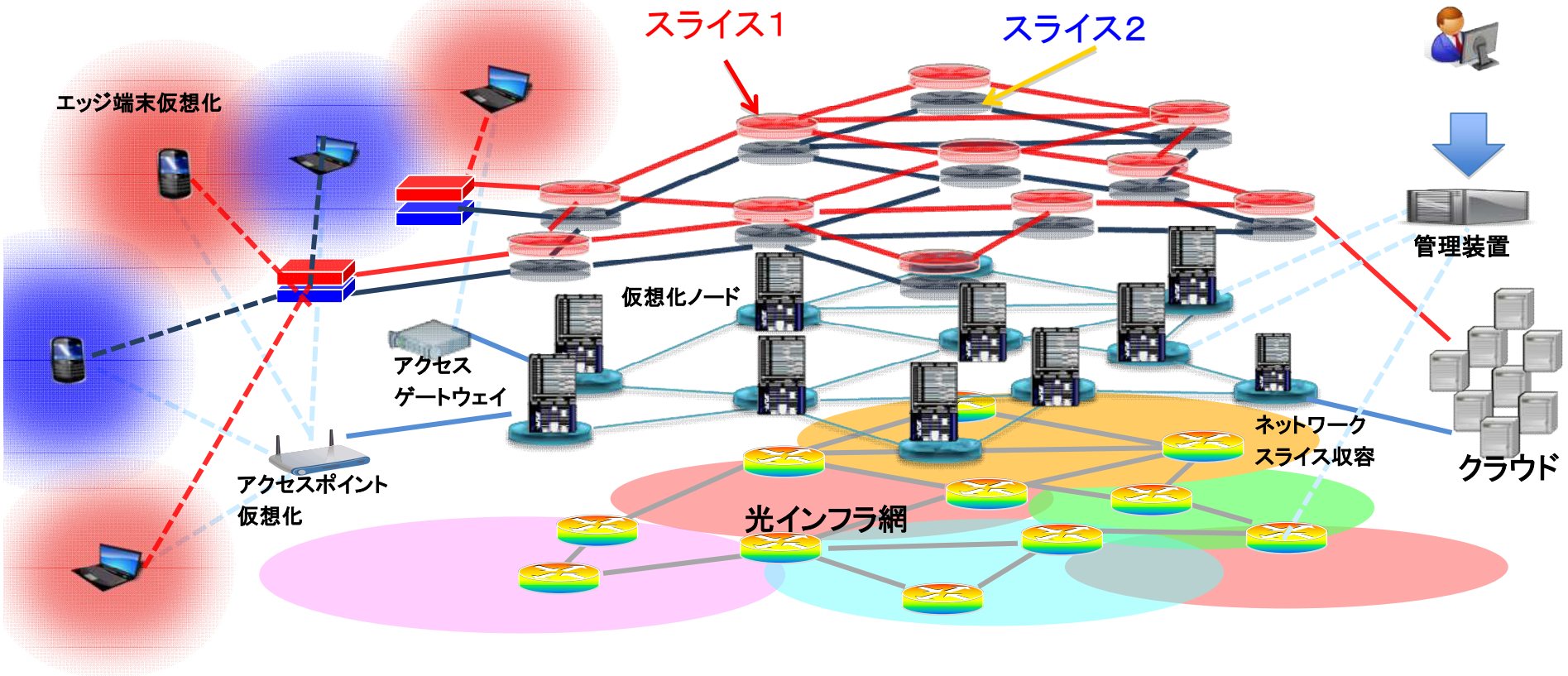
データプレーン

- Vnode(**仮想化ノード**) = Programmer + Redirector：**仮想化基盤資源提供**
- Access Gateway：**スライス構築者・利用者収容**
- Network Connector：**外部ネットワーク／クラウド収容**



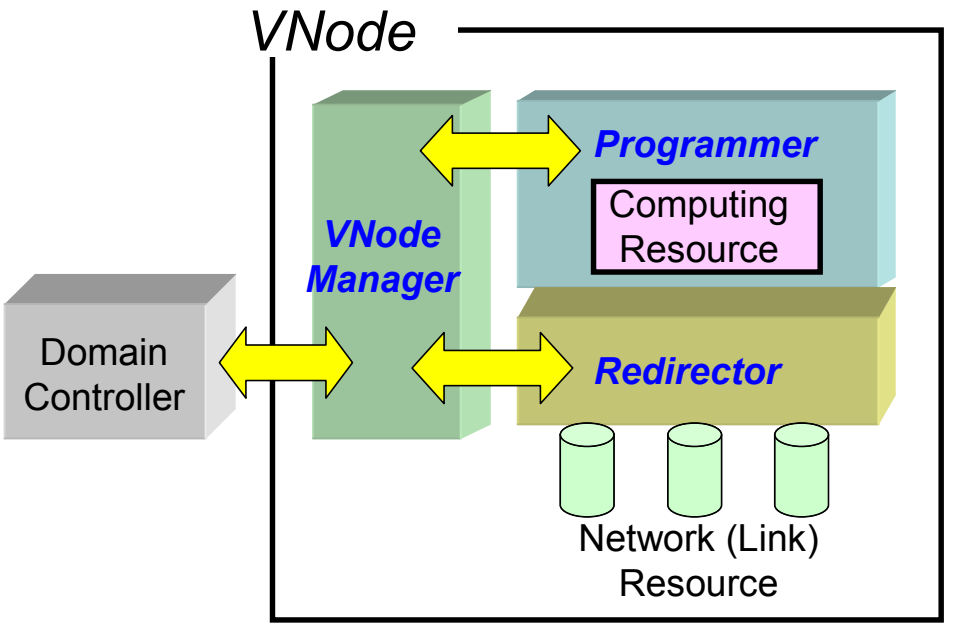
進化型ネットワーク仮想化基盤

（旧）ネットワーク仮想化基盤から、機能強化、収容機器拡大、容量拡大



仮想化ノード(VNode)

- （新・旧）仮想化基盤上で、物理資源を仮想化してスライスへ資源提供をするノード
- 構成
  - ネットワーク計算資源提供機能部分：
    Programmer（プログラマ）
  - リンク資源提供機能部分：
    Redirector（リダイレクタ）
  - ノード管理機能部分：
    VNode Manager (VNM)



“Programmer”（プログラマ）-ネットワーク処理計算ノード-

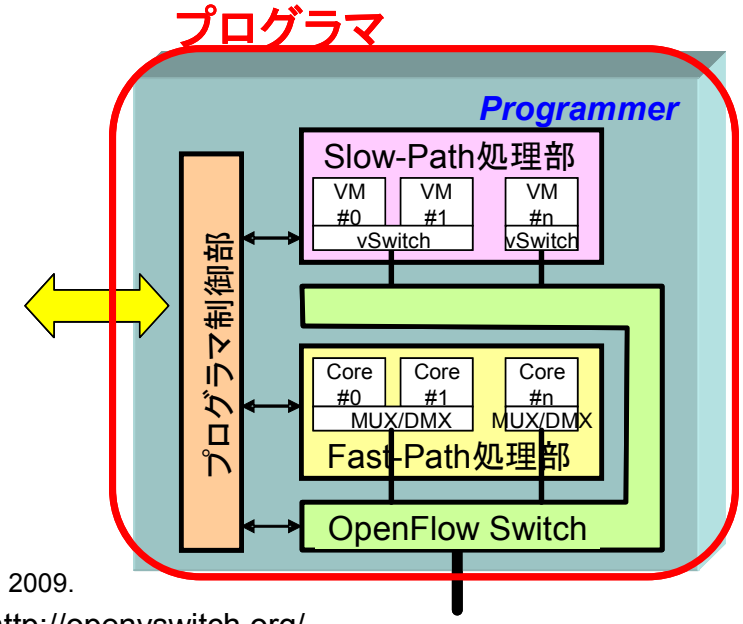
ネットワーク処理におけるプログラマビリティ（プログラム可能性）を提供

多様なネットワーク機能実現への対応

- Slow Path :仮想マシン（VM）による柔軟なプログラミング環境を提供
  - Intel Architectureサーバ上にKVM環境で実現[KVM]
  - VM-外部ネットワーク間通信制御にOpen vSwitchを使用[OpenvSwitch]
- Fast Path : 高いパケット転送性能を有するプログラミング環境を提供
  - Network Processorにて実現

拡張性を有したアーキテクチャ

- 装置内の計算資源接続にOpenFlowスイッチ技術を利用[OpenFlow]
  - 物理OpenFlowスイッチ（物理計算資源間）＋Open vSwitch（Slow Path内VM）
  - 各種計算資源を自由に組み合わせ可能なネットワークノードを構築



[KVM] “Kernel based virtual machine,” http://www.linux-kvm.org/page/Main_Page.

[OpenFlow] OpenFlow Switch Consortium, OpenFlow Switch Specification, version 1.0.0, Dec, 2009.

[OpenvSwitch] Open vSwitch project, Open vSwitch – an open virtual switch, http://openvswitch.org/.

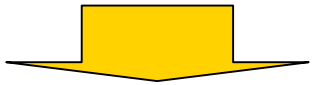
旧仮想化基盤・プログラマの課題

- プログラム性とパフォーマンスの両立
  - Slow-Pathのパフォーマンス向上
- 簡素で柔軟な管理システムの実現
- スライス全体におけるリソースアイソレーション
  - プログラマ内の資源確保／割り当て
- 利便性向上

Slow-Pathのパフォーマンス向上：課題と解決手段

課題

- Slow-Pathを構成するIAサーバ+KVMは、高い計算性能を保有するも、ネットワークI/O性能が不足
  - VM-外部ネットワーク間の通信ネック：
    vSwitch（ソフトウェアスイッチ）の性能向上が必要
  - ネットワークI/Oの物理帯域自体の増強



解決手段

- vSwitchの性能改善→ vSwitch offload
- ネットワークI/O自体の性能向上
  →広帯域化(GbE→10GbE)、I/O仮想化(SR-IOV)

提案プログラマアーキテクチャ

ノード内IAサーバのI/Oボトルネックとソフトスイッチ性能ネックを改善

特徴

① Slow-Pathのネットワーク I/O性能向上：

a) vSwitch性能向上 (vSwitch offload)

➔目的：ソフト処理のボトルネック解消、動的切り換え実現

b) I/O仮想化(SR-IOV)

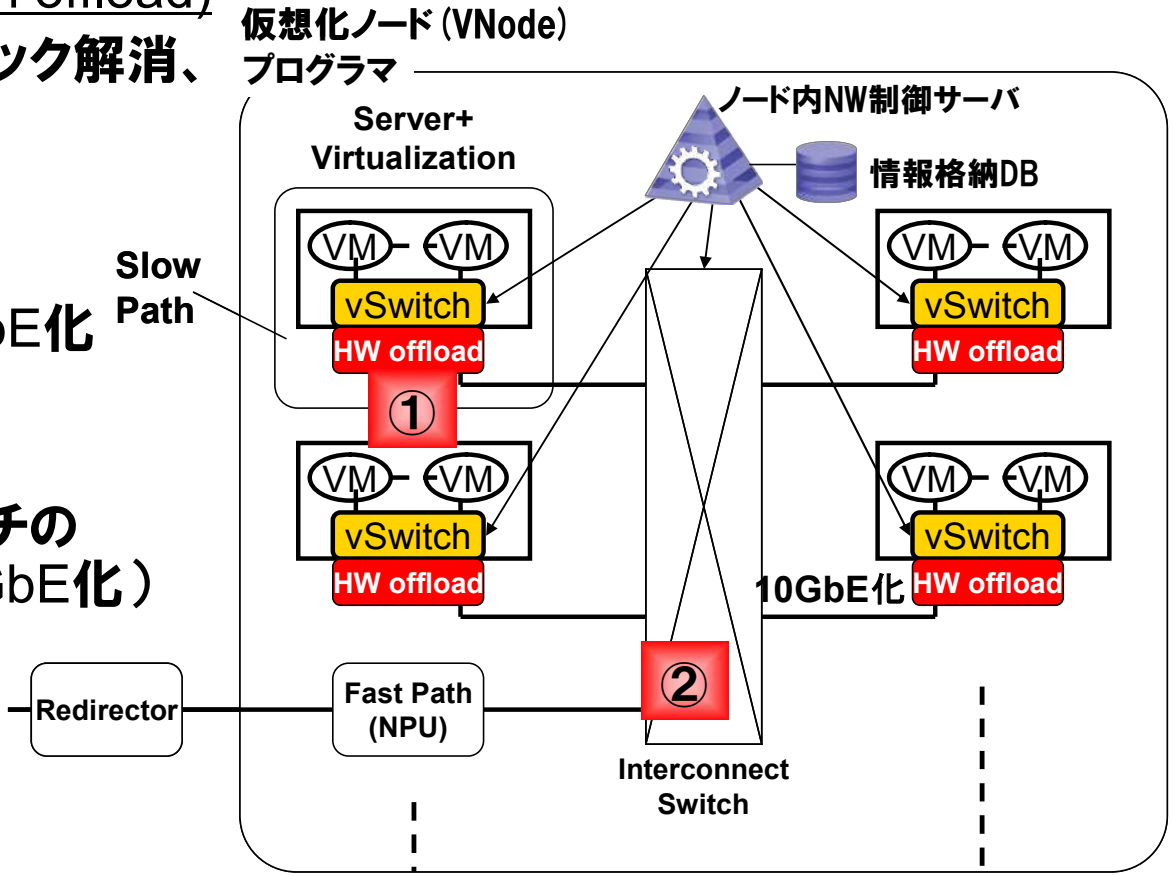
➔目的：仮想化I/O高性能化

c) プログラマ内部NICの10GbE化

➔目的：広帯域化

② 内部物理OpenFlowスイッチの大容量化（GbE接続→10GbE化）

➔目的：広帯域化



要素技術検証

プログラマへの適用を想定して要素技術の事前評価を実施

Slow-Pathのパフォーマンス向上

- KVM環境でのvSwitch offloadの実現と評価

Empowered by Innovation NEC

Slow-Pathのパフォーマンス向上
- KVM環境でのvSwitch offloadの実現と評価

目的・実現機能

目的

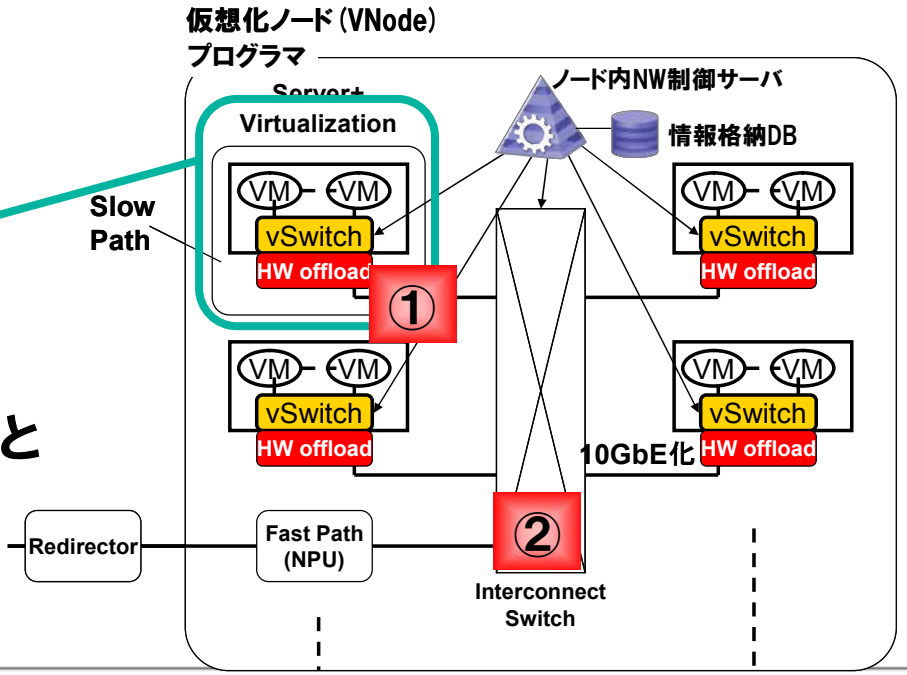
- 進化型ネットワーク仮想化基盤のネットワーク処理計算ノード(プログラマ)における、ネットワークI/O性能の向上、そのための要素技術の事前評価

実現機能

- ① Slow Path部分のI/O性能向上:
  a) vSwitch性能向上 (vSwitch offload) 、
  b) 標準I/O仮想化(SR-IOV) 、
  c) 10GbE化
- ② 内部OpenFlowスイッチの大容量化(GbE接続→10GbE)

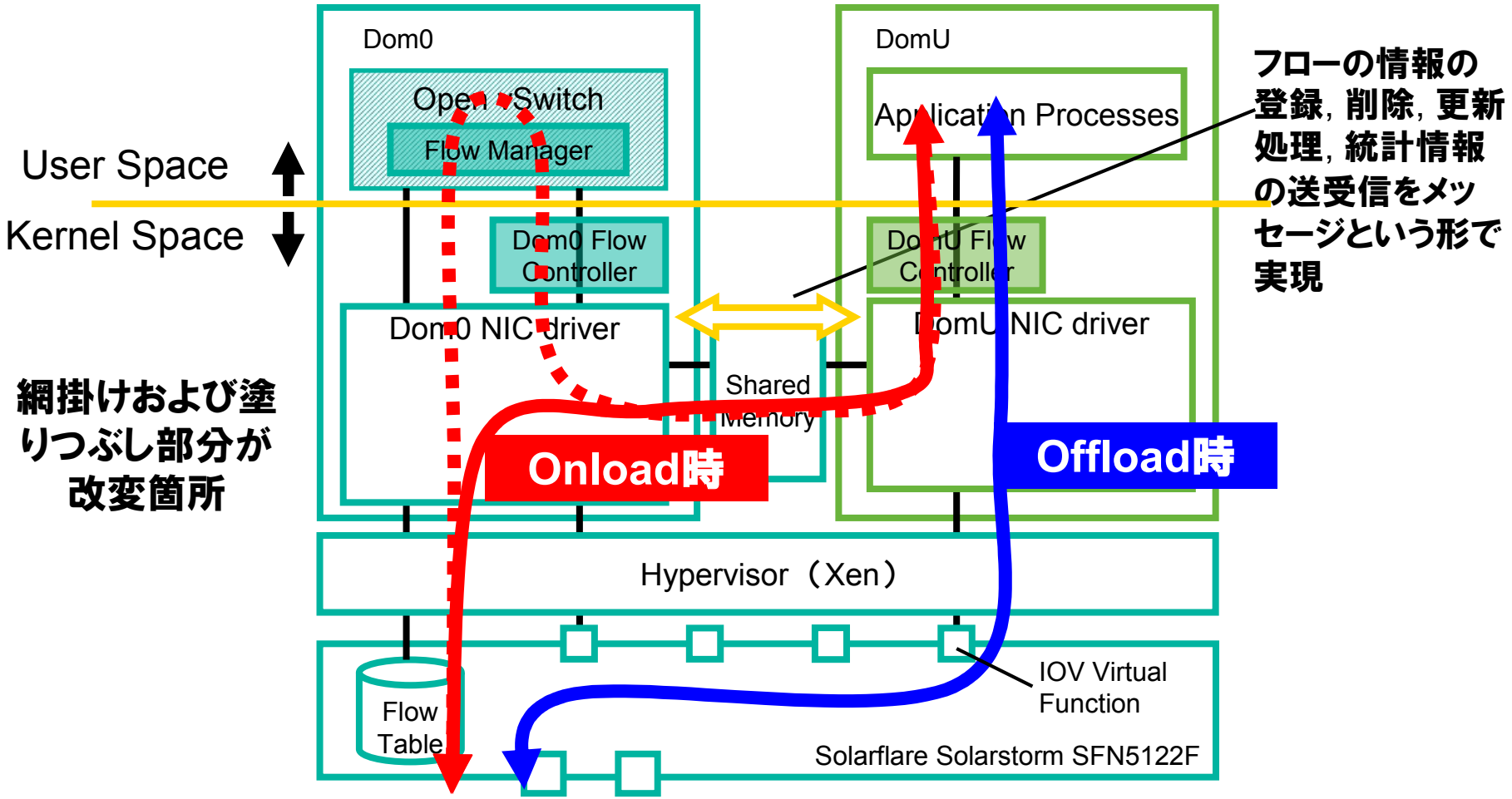
①-a) の実現手段

- KVMによるvSwitch offload実現
- b) SR-IOV 、c) 10GbE化の効果と合わせて検証・評価



従来技術：Xenserver環境のvSwitchオフロード [Tsuji11]

Xenserver上で、Solafrare社NIC (SFN5122F) の独自IOMMUを使用して実現



[Tsuji11] 辻聡 他, “vswitch 処理の動的オフロード方式の実装と評価,”信学技報 NS2011-29, pp69-74, May 20, 2011

Slow Path部 vSwitchオフロード方針

① サーバ仮想化機構としてKVMを使用（Xenserverからの変更）

理由：プログラマのサーバ仮想化環境を継承

② I/O仮想化機構としてSR-IOVを使用（独自IOMMUから）

理由：GuestVMに対するポータビリティの改善

③ 動的オフロード切り換えの実現

目的：ノード内資源の有効活用

課題：Host/Guest間通信実装, NIC Flow Table操作

Solafrare社NIC (SFN5122F) 利用[Solarflare]

[Solarflare] Solarflare Communications, Inc., “Solarflare SFN5122F Dual-Port 10G Ethernet Enterprise Server Adapter”. Available at http://www.solarflare.com/products/Solarflare_10GbE_NIC_SFN5122F_v011711.pdf.

KVM化上の課題

Host/Guest間通信

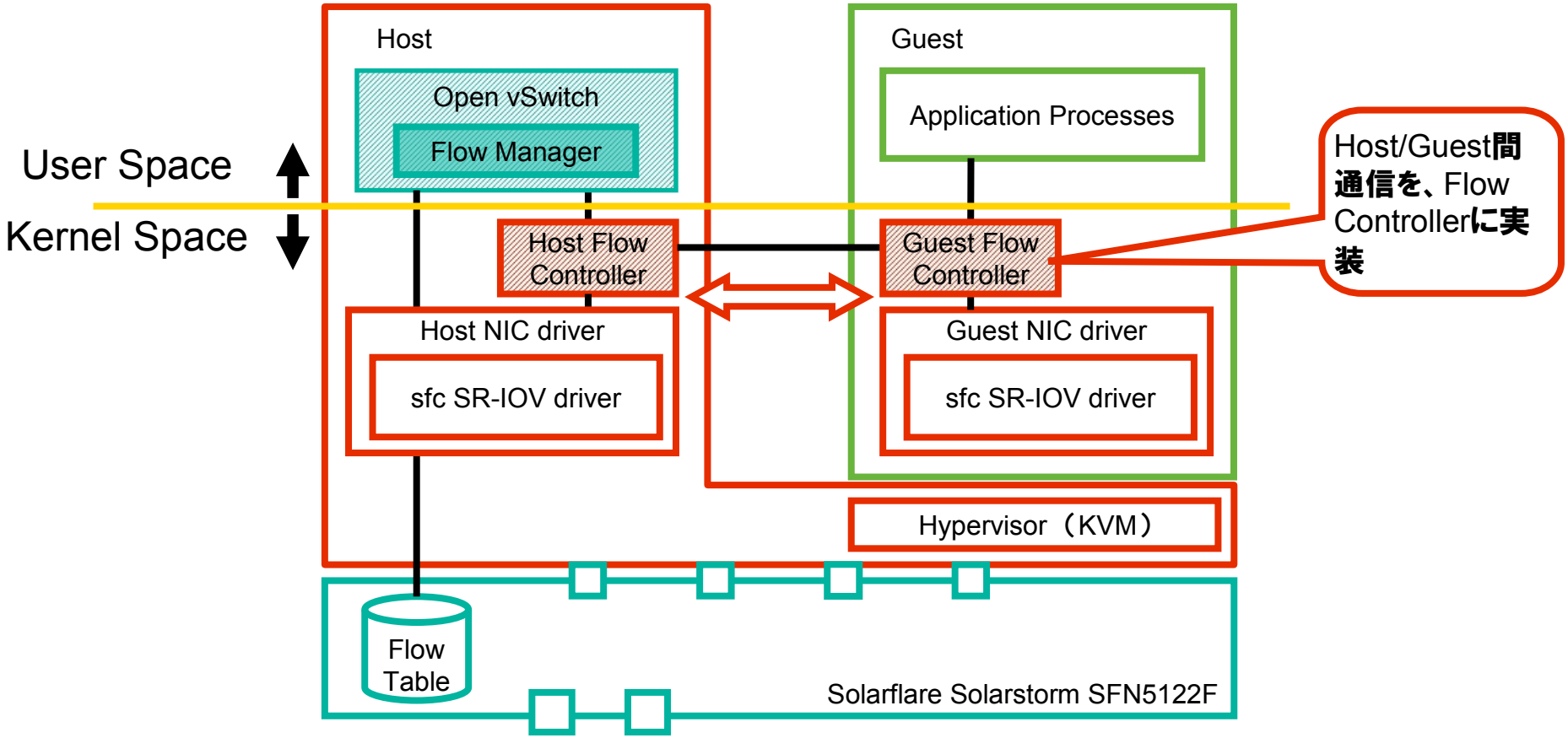
- KVMのHost/Guest間には、Xenでは用意されていたDom0/DomU間通信相当を提供する機能がない。

考案方式

- Offload処理実現のために追加搭載したFlow Controller 部にHost/Guest間通信機能を実装

考案方式：KVM+SR-IOVでのvSwitchオフロード

Host/Guest間通信機能をFlow Controller に搭載



KVM+SR-IOV+vSwitch Offload 処理性能評価：項目

評価項目

送受信スループット（bps）、CPU使用率

注）CPU使用率：HOST CPU使用率を測定：
算出：100% - %idle-%guest

評価パラメータ

VM数	1，2，4，8
Port数	1，2
MTU	9000
vSwitch offload	offload, onload
SR-IOV	on
コア割付	**なし**

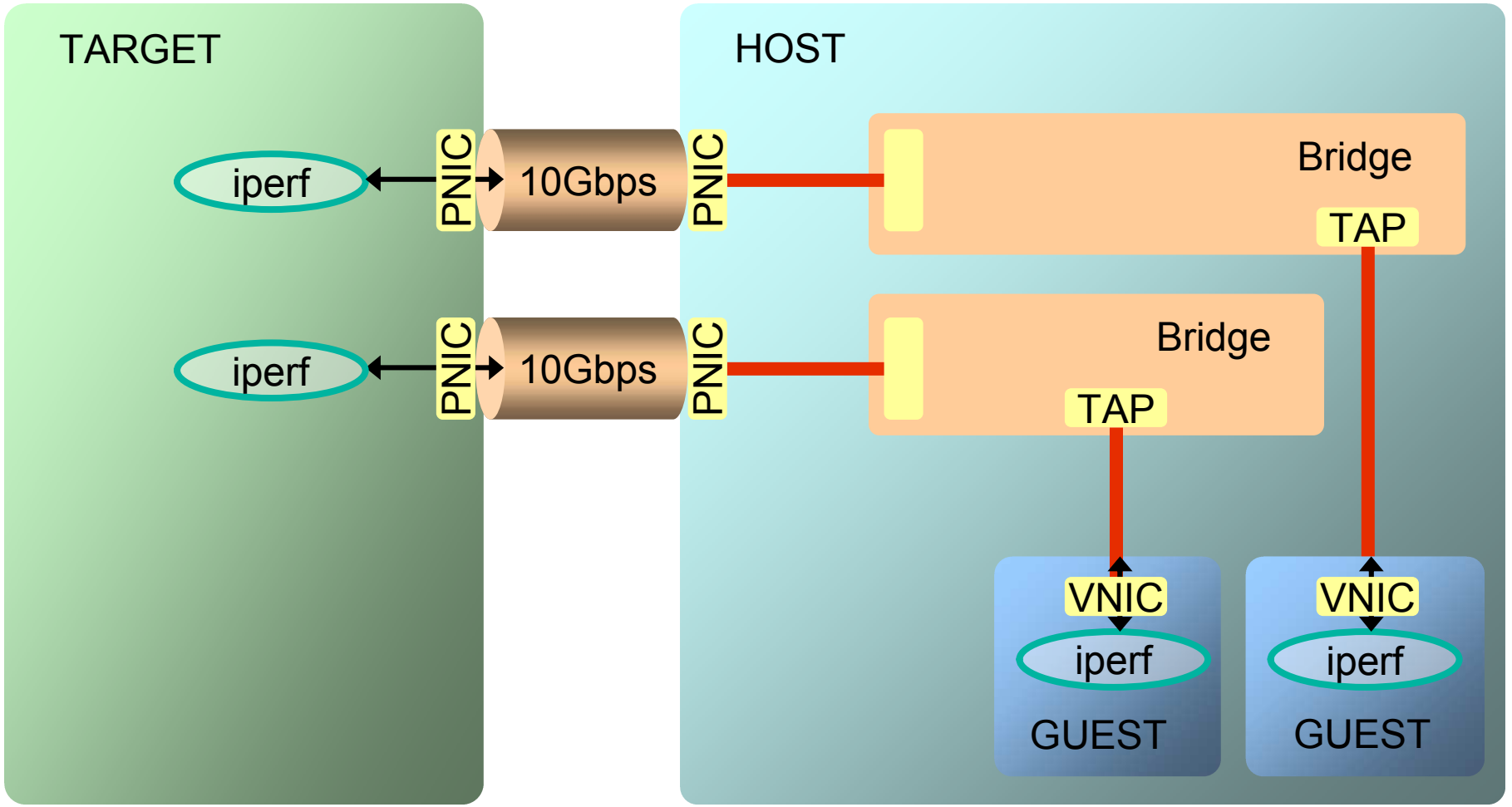
KVM+SR-IOV+vSwitch Offload 処理性能評価:条件

評価環境諸条件

	TARGET	HOST
OS	Fedora 15 (kernel-2.6.38.6-26)	
CPU	Intel(R) Xeon(R) X5680 3.33GHz（**コア数:**24）, HT:2 /core	
Memory	80GB	
HDD	500GB	
NIC	BCM57711	SFN5122F
Driver Version(HOST)	1.70.00-0	3.2.0.6040 (SF-103848-LS)
Driver Version(GUEST)	ー	sfc-xnap-virtio-v1_0_0_0012 (SF-105740-LS)
Software(Traffic)	iperf version 2.0.5 (08 Jul 2010) pthreads	iperf version 2.0.5 (08 Jul 2010) pthreads
Software(CPU)	mpstat (sysstat-9.0.6.1-14)	mpstat (sysstat-9.0.6.1-14)

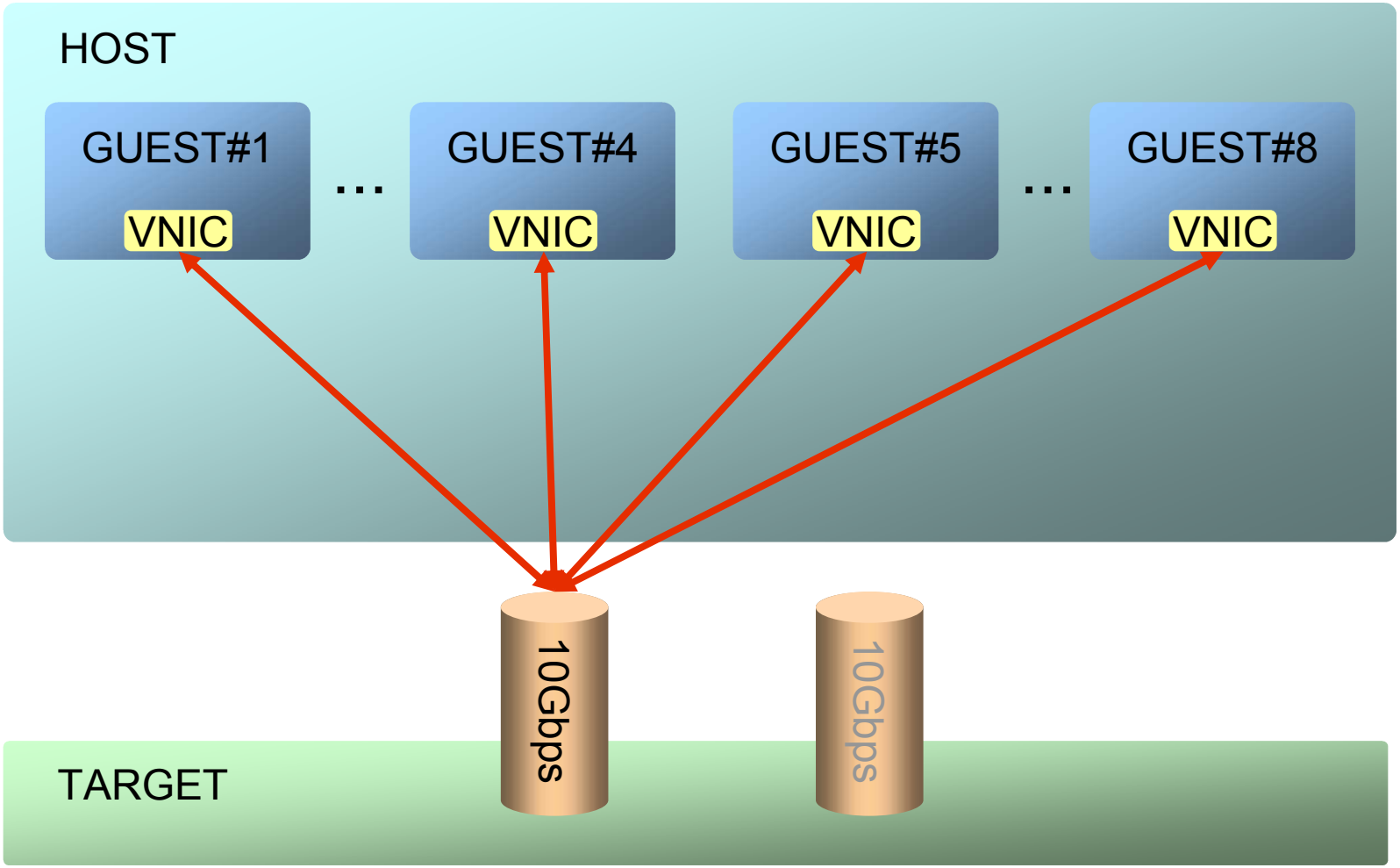
KVM+SR-IOV+vSwitch Offload 処理性能評価：評価系

iperf稼働：TARGETは物理サーバ上、HOST/GUESTは仮想マシン上



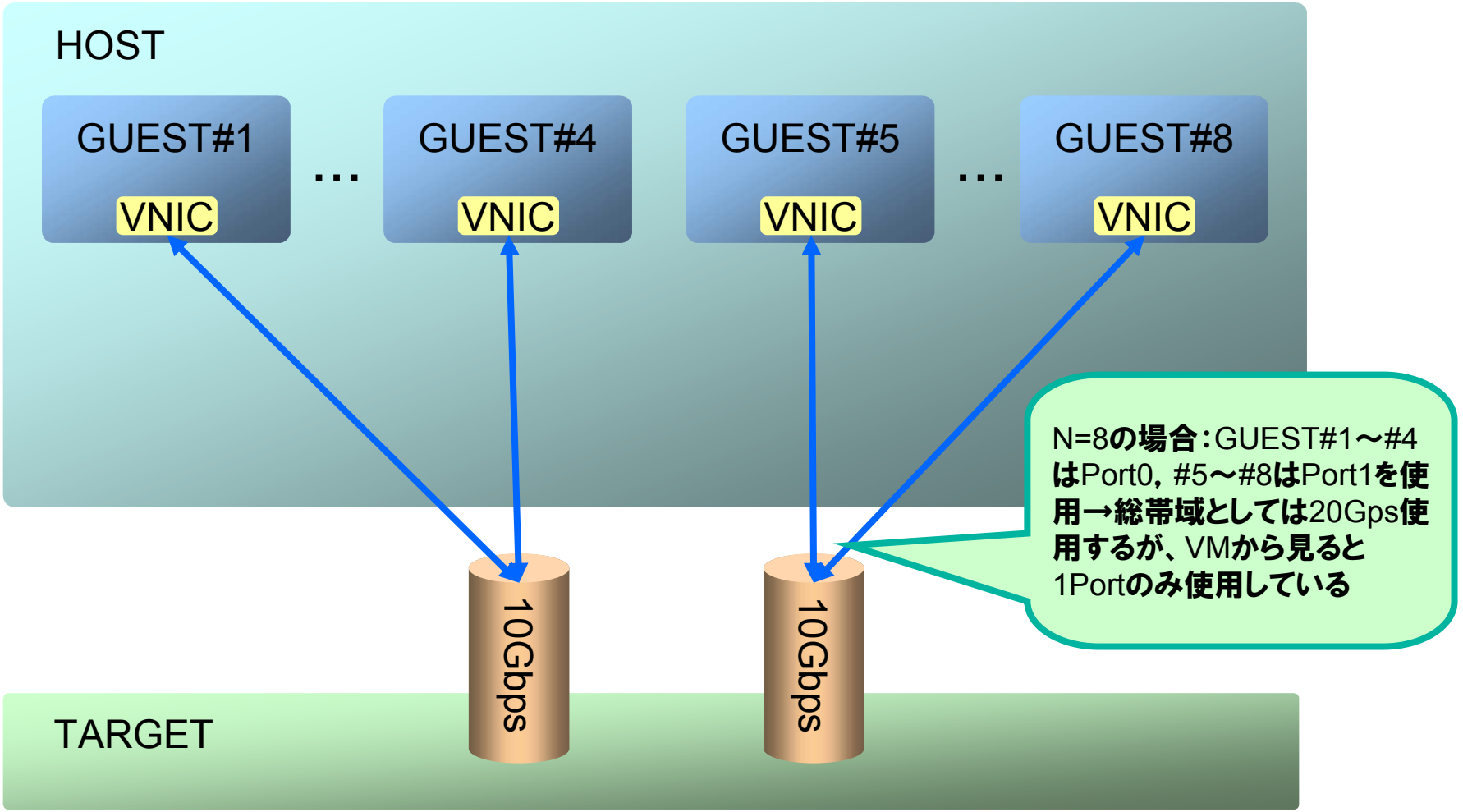
1Port使用時のデータフロー割り当て

1VMにつき1VNIC使用。全VMのデータフローを単一物理ポートに割り当て



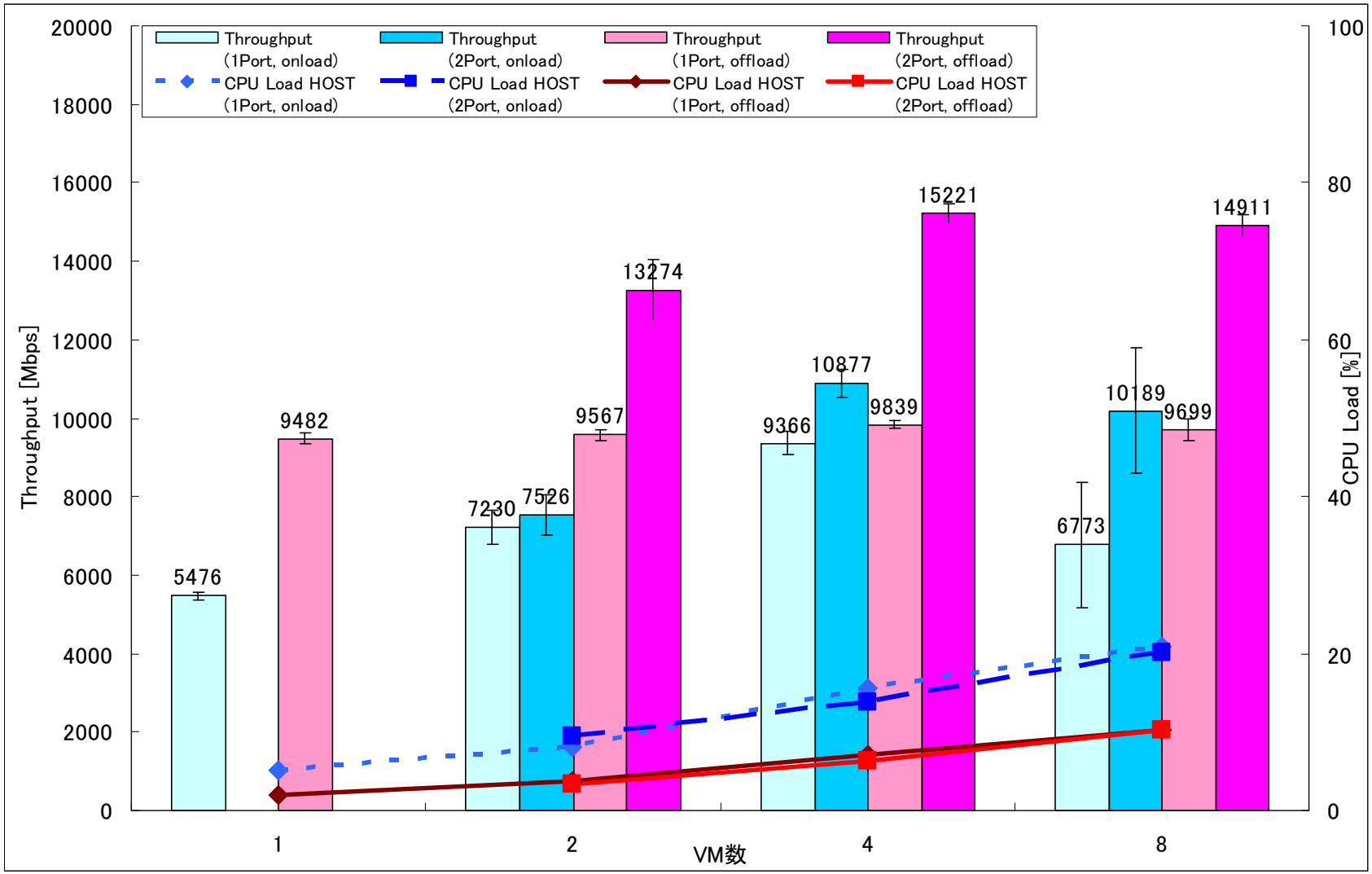
2Port使用時のデータフロー割り当て

1VMにつき1VNIC使用。VM毎に使用する物理ポートを一つに固定



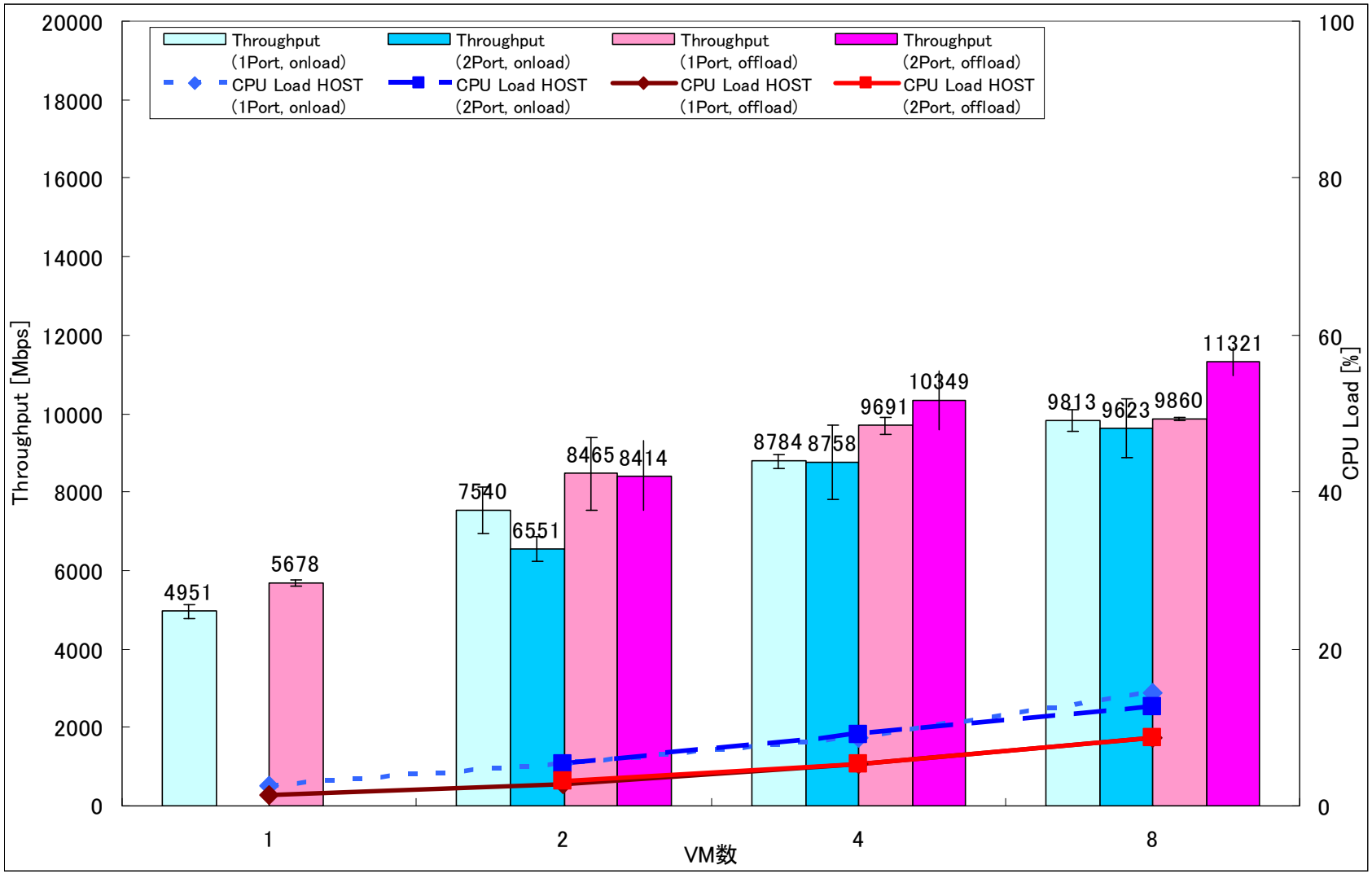
KVM+SR-IOV+vSwitch Offload 処理性能評価結果(１)

受信



KVM+SR-IOV+vSwitch Offload処理性能評価結果(2)

送信



KVM+SR-IOV+vSwitch Offload処理性能評価考察

受信側は多ポート化+複数VM時での性能改善を確認
送信側は改善度小：今後詳細分析を実施

■受信

- ✓Throughput: vSwitch Offload 時 ＞ vSwitch Onload時
- ✓Offload時のThroughputは、15.2Gbpsで頭打ち
  - ➢1Port利用では、1VMでほぼWire-rate (9.9Gbps) に到達。
    2Port利用では、1Port利用時の性能×2倍に達せず。
- ✓Host CPU使用率：Offload時 ＜ Onload時
- ✓Host CPU使用率は、VM数にほぼ比例して増大

■送信

- ✓Throughput: vSwitch Offload 時 ＞ vSwitch Onload時
  だがOffload機能On時の改善度合いは小。
- ✓Offload時のThroughputは、11.3Gbpsで頭打ち。
- ✓Host CPU使用率：Offload時 ＜ Onload時
  - ➢Host CPU使用率はまだかなりの余裕あり（On時 10%未満）。

まとめ

- 進化型ネットワーク仮想化基盤実現に向けた「プログラマ」（ネットワーク処理計算ノード）を提案した。
- 実現に必要となる要素技術として、KVM環境でのvSwitch offloadの実現と評価を行った。

今後の活動

- プログラマの試作と動作検証・性能検証
- 他課題検討
  - 簡素で柔軟な管理システムの実現
  - スライス全体におけるリソースアイソレーション
    - プログラマ内の資源確保／割り当て
  - 利便性向上

参考文献

- [Nakao10] A. Nakao, “Network Virtualization as Foundation for Enabling New Network Architecture and Application,” IEICE Trans. Commun., Vol E93-B, No. 3, pp.454-457, March 2010.
- [Nakao12]**中尾 彰宏**, “VNode: A Deeply Programmable Network Testbed Through Network Virtualization,” **第3回NV研究会**, Mar. 2, 2012
- [Tsuji11] **辻聡 他**, **“**vswitch **処理の動的オフロード方式の実装と評価,”信学技報** NS2011-29, pp69-74, May 20, 2011
- [KVM] “Kernel based virtual machine,” http://www.linux-kvm.org/page/Main_Page.
- [Xen] http://wiki.xen.org/wiki/Network_Throughput_Guide #Recommended_TCP_settings_for_a_VM
- [Solarflare] Solarflare Communications, Inc., “Solarflare SFN5122F Dual-Port 10G Ethernet Enterprise Server Adapter”. Available at http://www.solarflare.com/products/Solarflare_10GbE_NIC_SFN5122F_v011711.pdf.
- [SRIOV] PCI-SIG I/O Virtualization, “http://www.pcisig.com/specifications/iov/”.
- [VMDc] http://www.intel.com/network/connectivity/solutions/vmdc.htm
- [OpenFlow] OpenFlow Switch Consortium, OpenFlow Switch Specification, version 1.0.0, Dec, 2009.
- [OpenvSwitch] Open vSwitch project, Open vSwitch – an open virtual switch, http://openvswitch.org/.

Empowered by Innovation