フロー処理やネットワーク機能の処理負荷を考慮したVSEリソース割当ての最適化

○村松 真† 川島 龍太† 齋藤 彰一† 松尾 啓志†
中山 裕貴†† 林 經正††

† 名古屋工業大学大学院
†† 株式会社ボスコテクノロジーズ

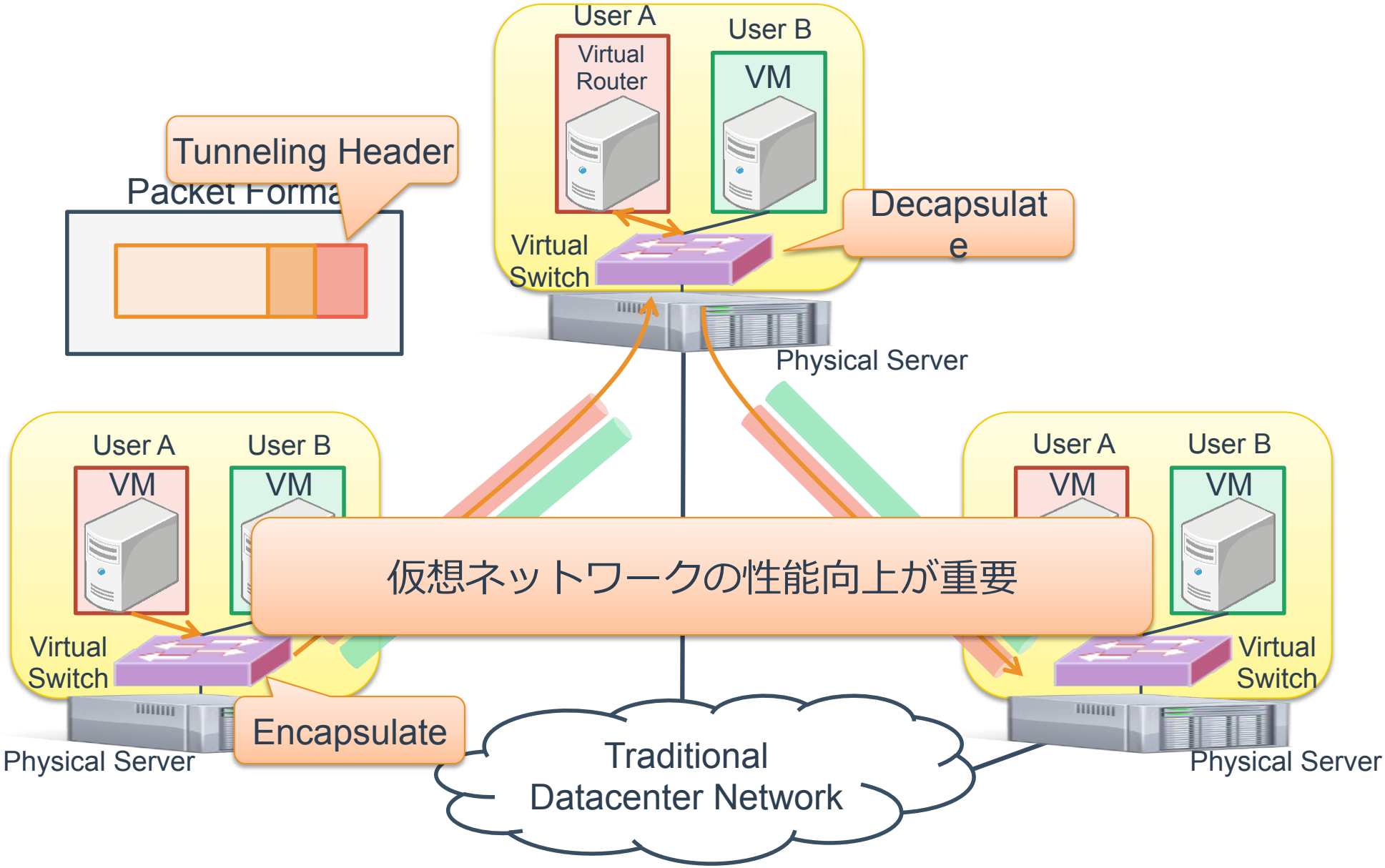
2015/04/22
CQ研究会

研究背景

- クラウドコンピューティングサービスの普及
  - データセンタの需要が増加
    - 膨大な数の物理サーバとストレージを管理

- マルチテナント方式を採用
  - 物理サーバ上に様々なテナントの仮想マシン(VM)を集約
    - 各テナントに独立した仮想環境を提供
    - 物理サーバの台数を削減
  - オーバレイプロトコルによる仮想ネットワークの構築
    - 既存のデータセンタネットワーク上で構築可能

マルチテナント方式の構成



仮想ネットワーク性能向上のアプローチ

- トンネリングプロトコルの改善

- ハードウェアスイッチ
  - CPU負荷の低減、通信の高速化

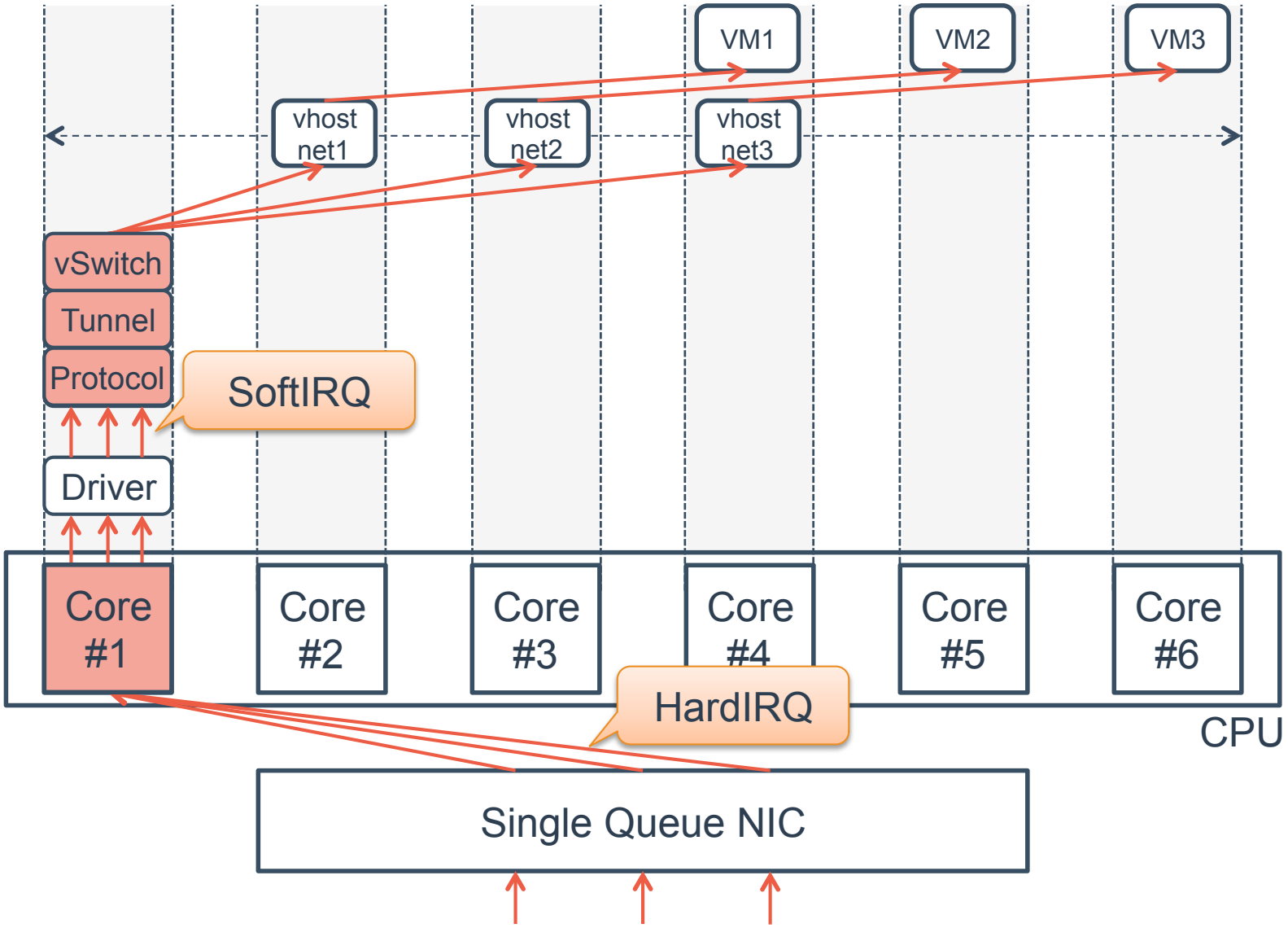
特定のハードウェア機能が必要となり
ベンダロックインに陥る可能性がある

- ホスト内処理の受信処理負荷に着目

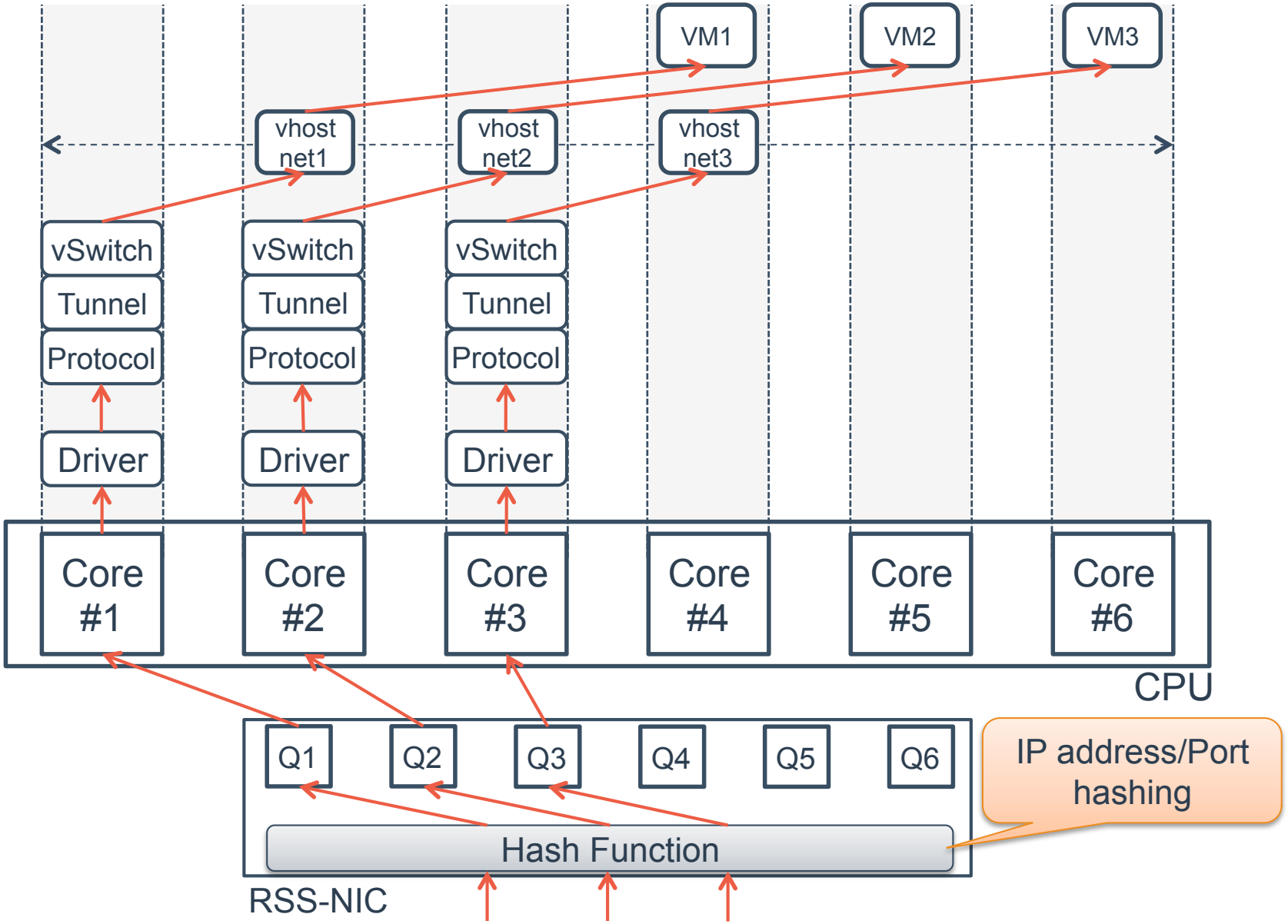
特定のハードウェア機能を利用せず
ホスト内受信処理のCPUリソース割当て方法の最適化

シングルキューNICにおける問題点

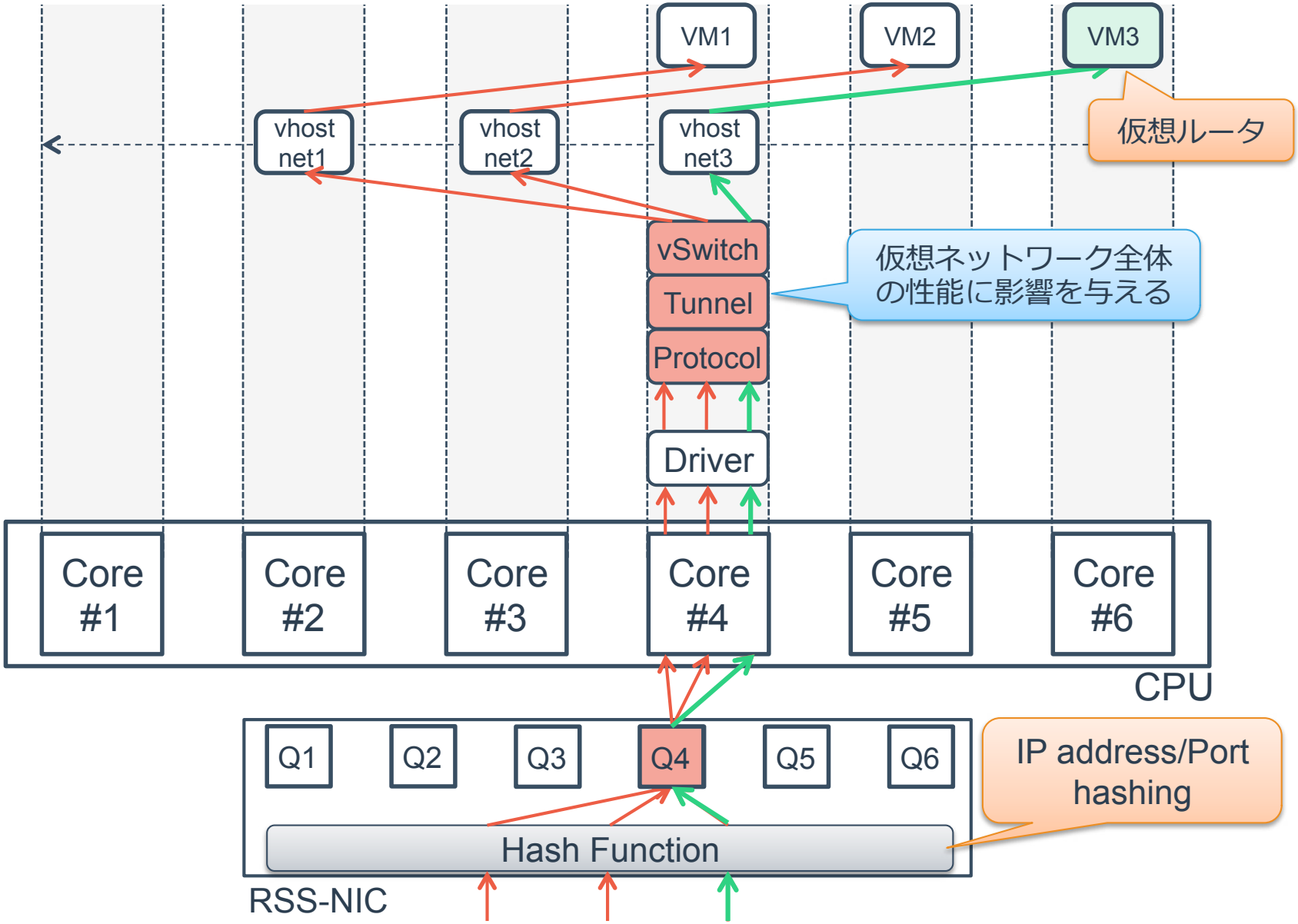
- 特定のCPUコアに受信処理負荷が集中



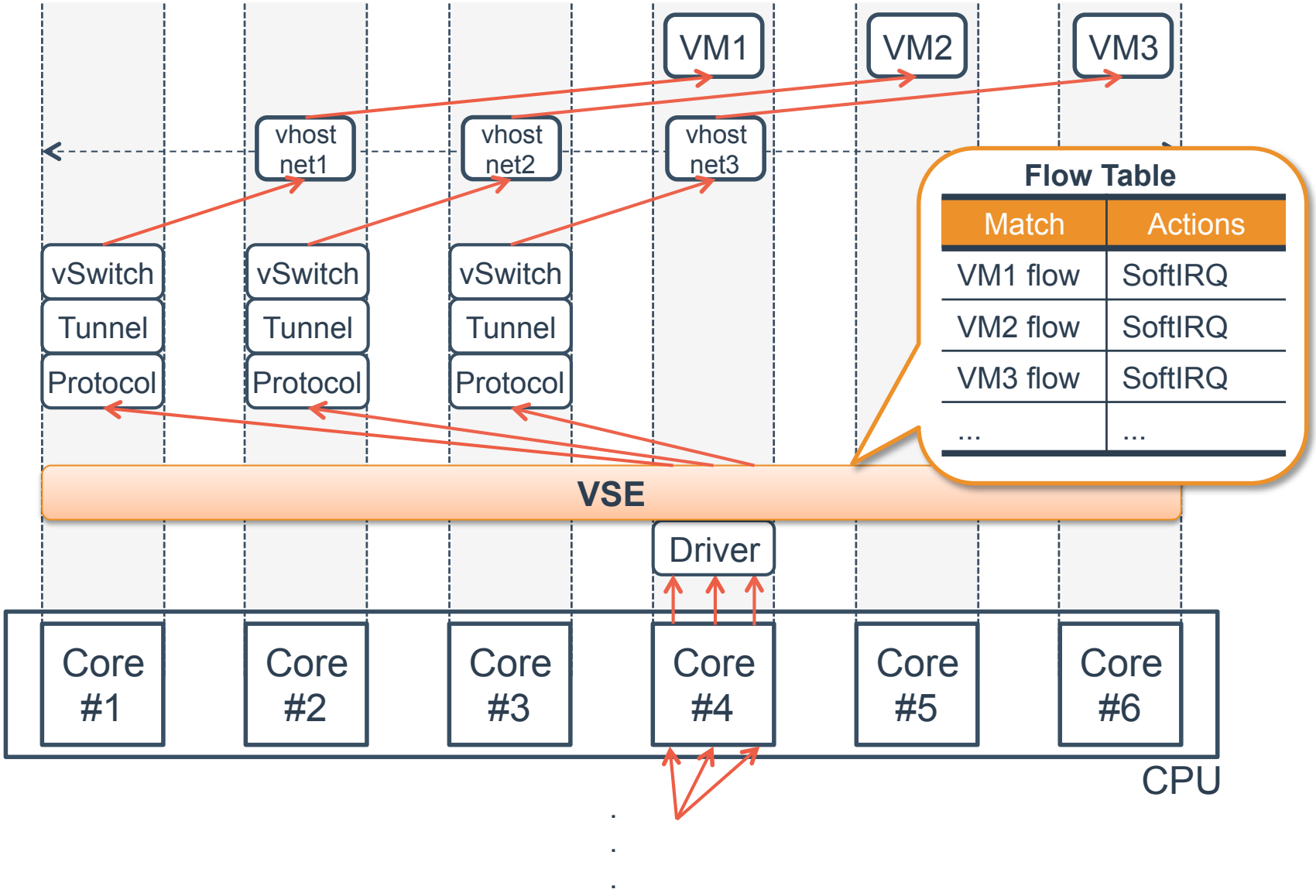
Receive Side Scaling (RSS)



Receive Side Scaling (RSS)

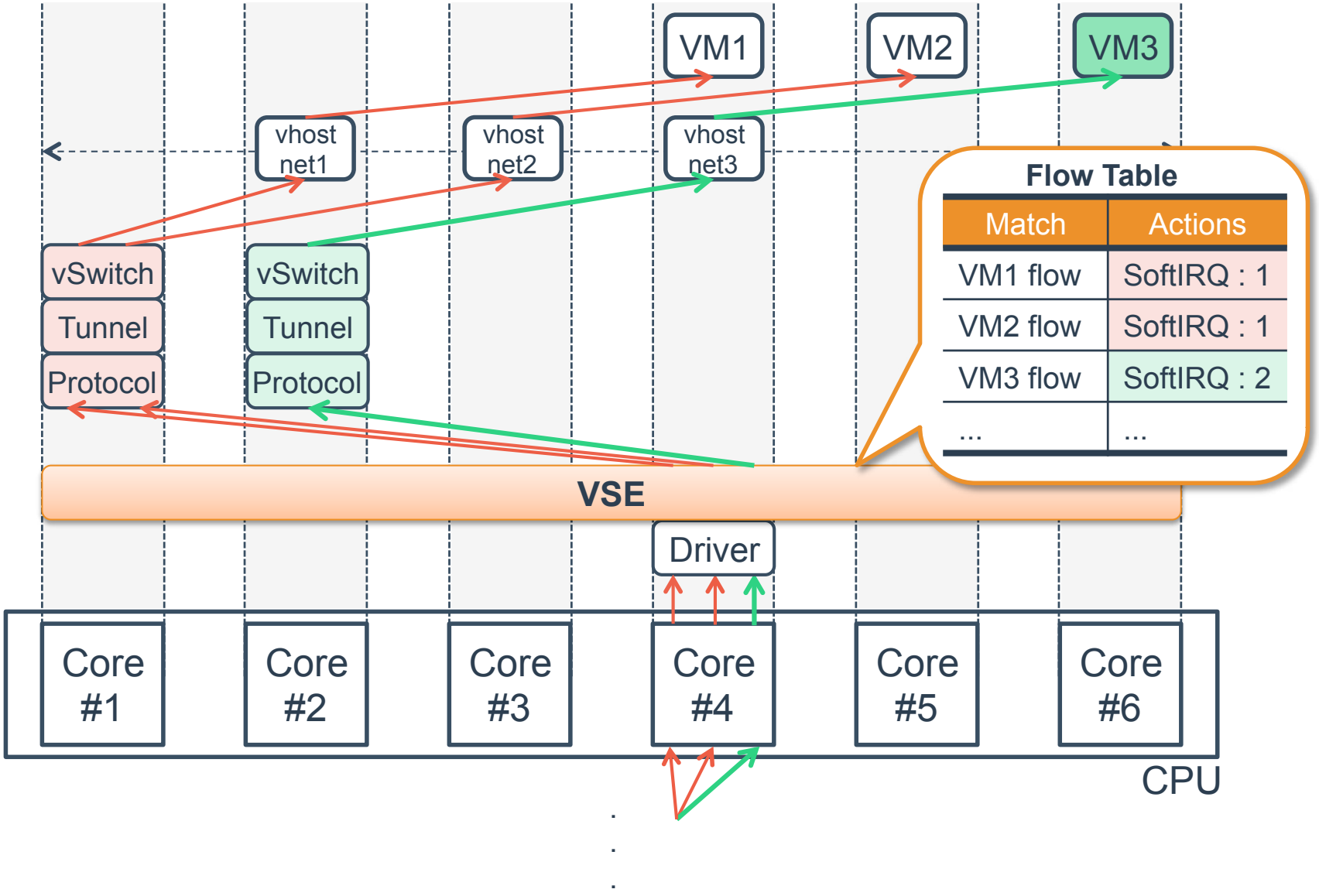


提案手法：Virtual Switch Extension (VSE)†



† “VSE: Virtual Switch Extension for Adaptive CPU Core Assignment in Softirq” 2014 IEEE 6th International Conference on Cloud Computing Technology and Science (CloudCom), Shin Muramatsu et al.

提案手法：Virtual Switch Extension (VSE)†



† “VSE: Virtual Switch Extension for Adaptive CPU Core Assignment in Softirq” 2014 IEEE 6th International Conference on Cloud Computing Technology and Science (CloudCom), Shin Muramatsu et al.

従来評価との相違点

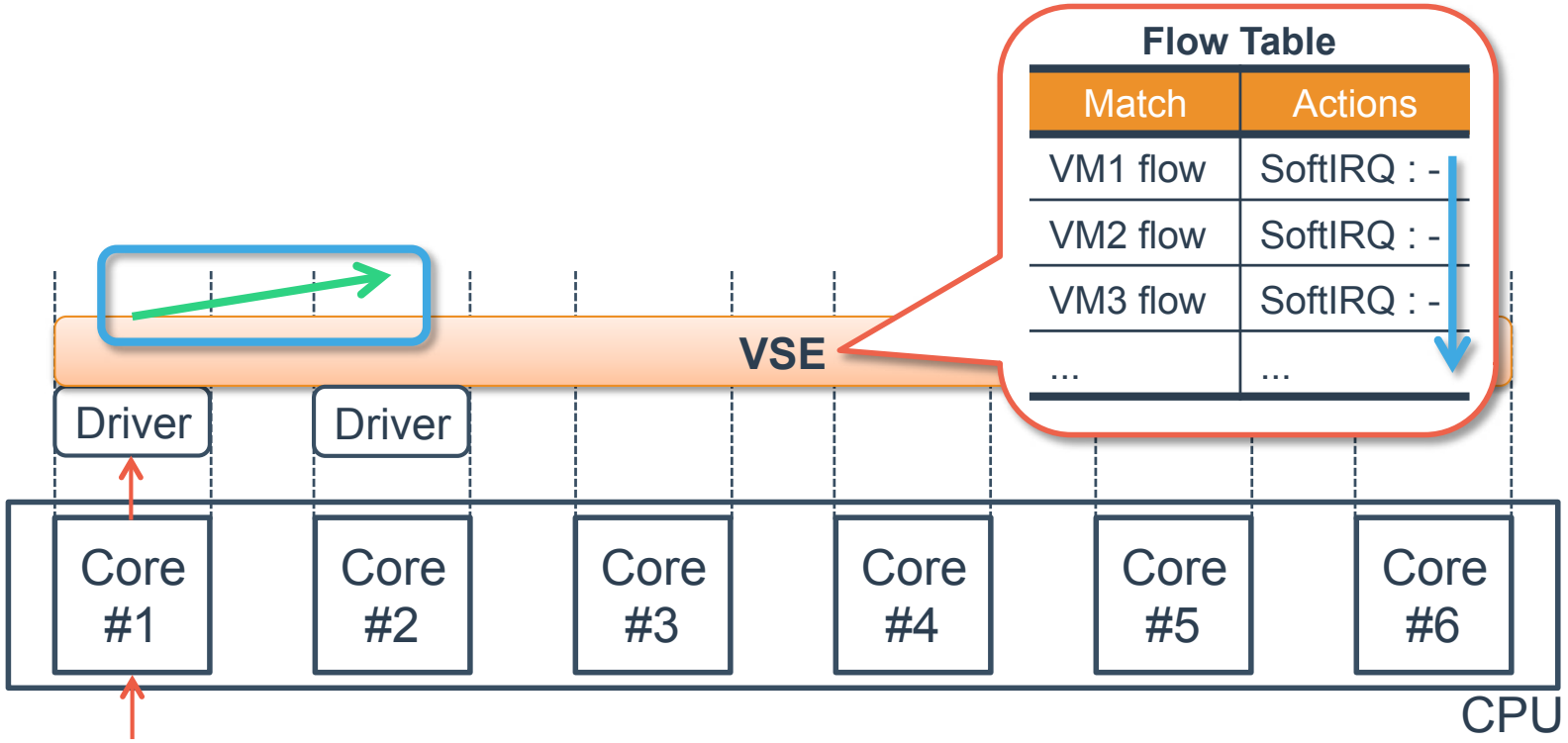
従来評価	本研究
• フロー処理負荷が非常に高い環境における評価のみ ➢ VXLAN over IPsec	• フロー処理負荷が比較的低負荷な環境における評価 ➢ VXLAN

- 低負荷なフロー処理における評価項目
  - ➢ VSE処理オーバヘッド
  - ➢ フロー/VMのCPUコア衝突

より一般的なケースでの通信性能への影響を評価

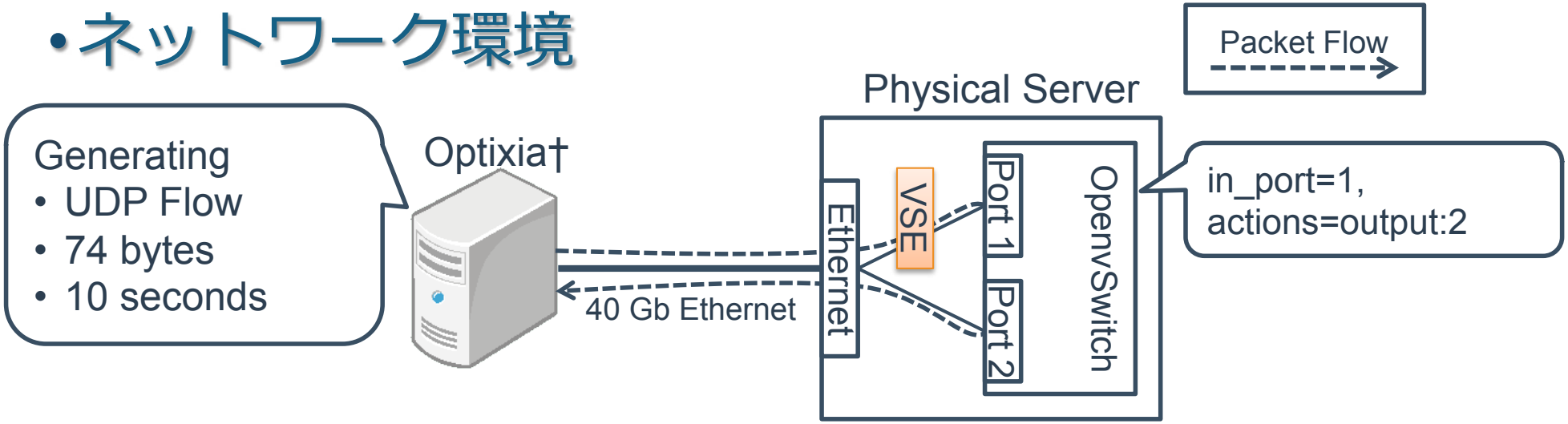
VSEにおける処理

ソフト割り込み先変更	フローテーブルのルックアップ処理
ハードウェア割込先CPUコアと 別のCPUコアへソフトウェア割込	線型検索によるエントリ検索



評価環境

- ネットワーク環境

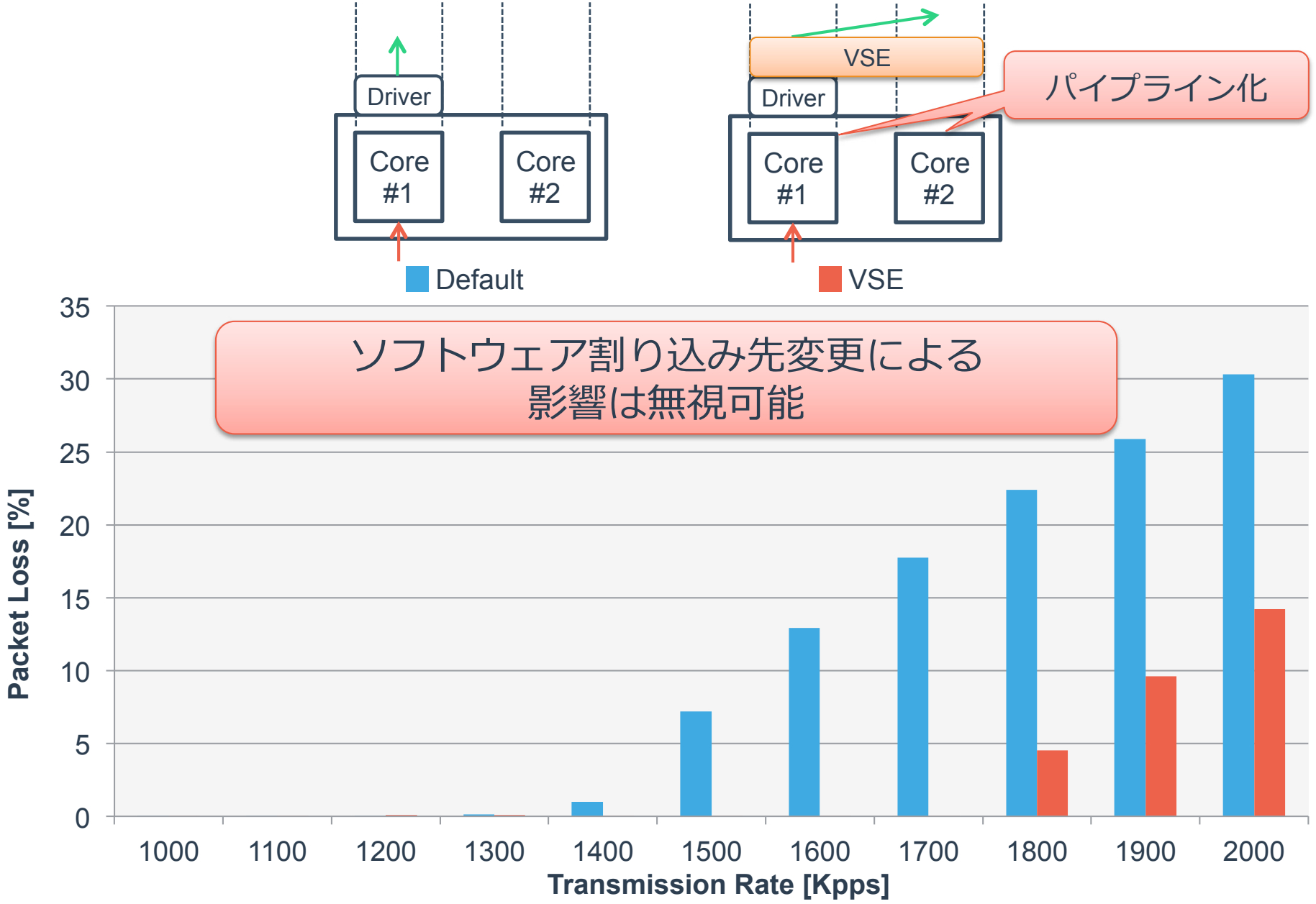


- マシン仕様

	Physical Server
OS	CentOS 6.6 (2.6.32)
CPU	Intel Core i7 (6 core)
Memory	16G bytes
Buffer	4M bytes
Network	40GBASE-SR4
Driver	MellanoxConnect-X(R)-3
vSwitch	OpenvSwitch 2.3.0

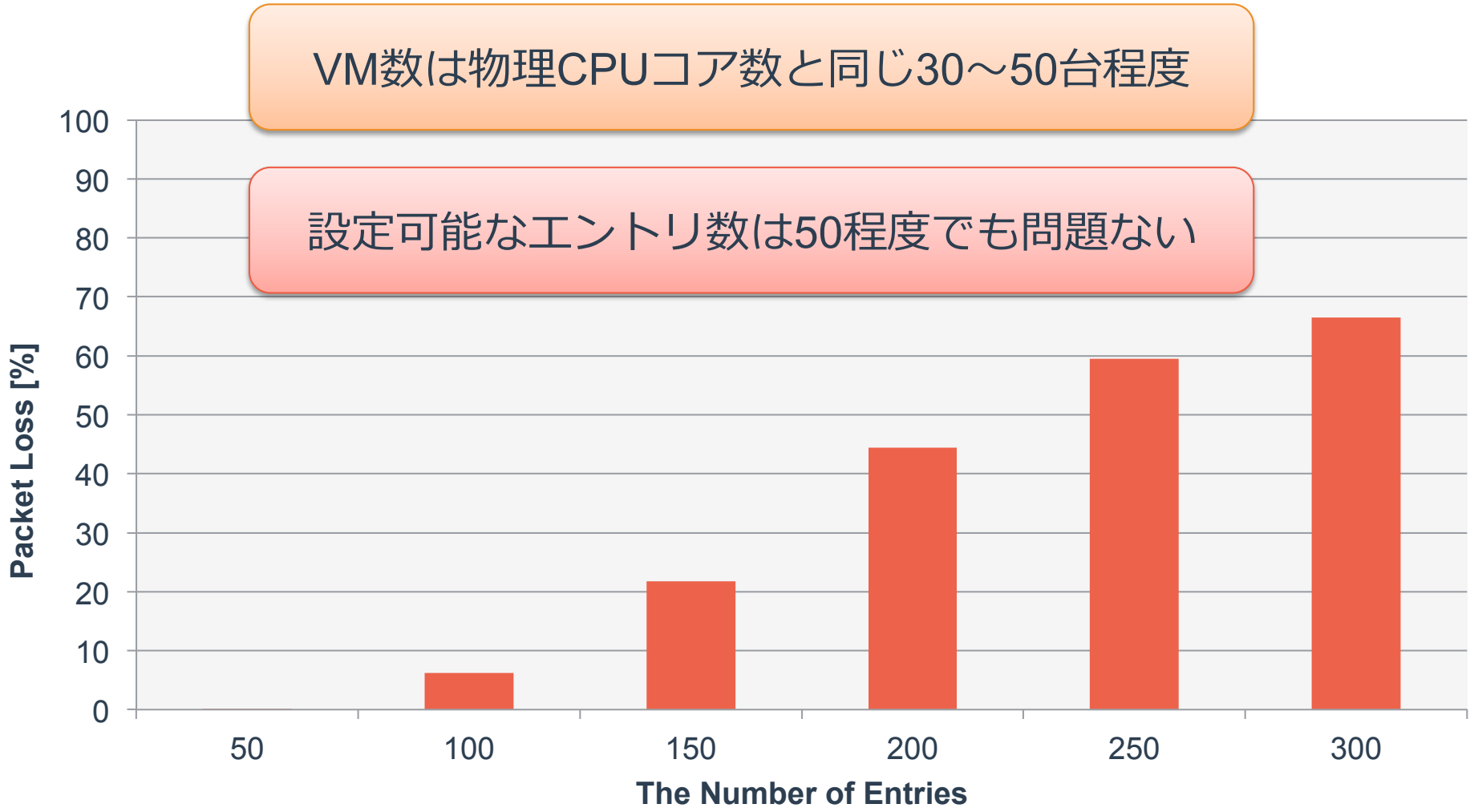
† Optixia XM2 “http://www.ixiacom.jp/sites/default/files/content/support/library/datasheets/ja_optixiaxm2.pdf”

評価結果：ソフトウェア割り込み先変更の影響



評価結果：エントリ数の影響

- 送信レートを1700Kppsに固定
  - マッチするエントリを末尾に設定



従来評価との相違点

従来評価	本研究
• フロー処理負荷が非常に高い環境における評価のみ ➢ VXLAN over IPsec	• フロー処理負荷が比較的低負荷な環境における評価 ➢ VXLAN

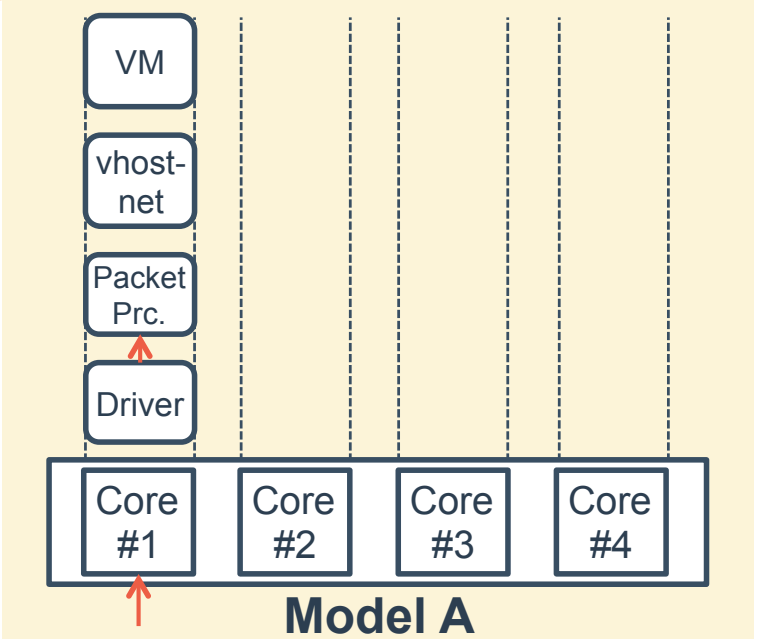
- 低負荷なフロー処理における評価項目
  - ➢ VSE処理オーバヘッド

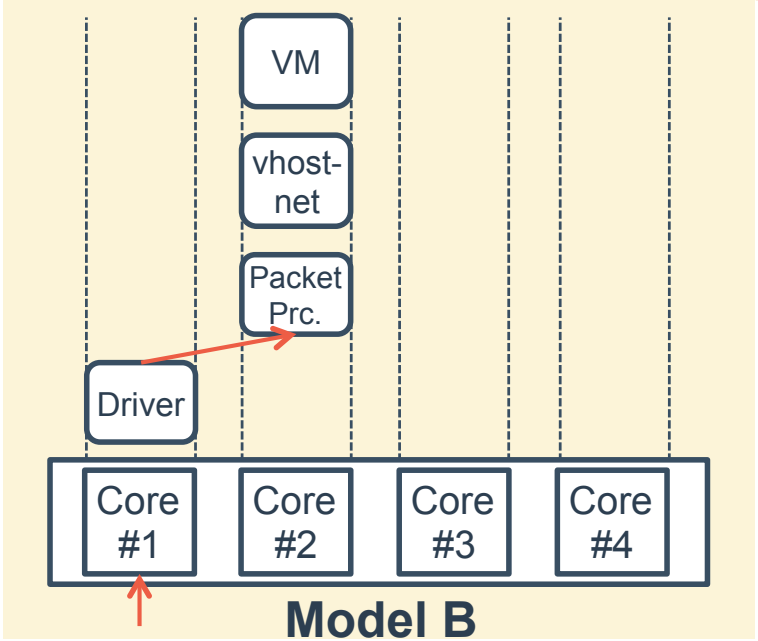
➢ フロー/VMのCPUコア衝突

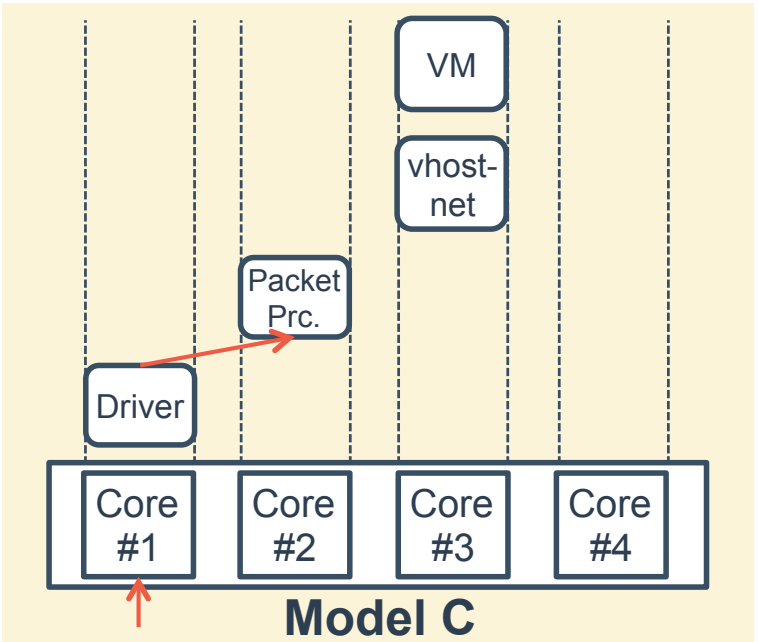
物理CPUコア数に対してVM数とフロー数の合計が大きい

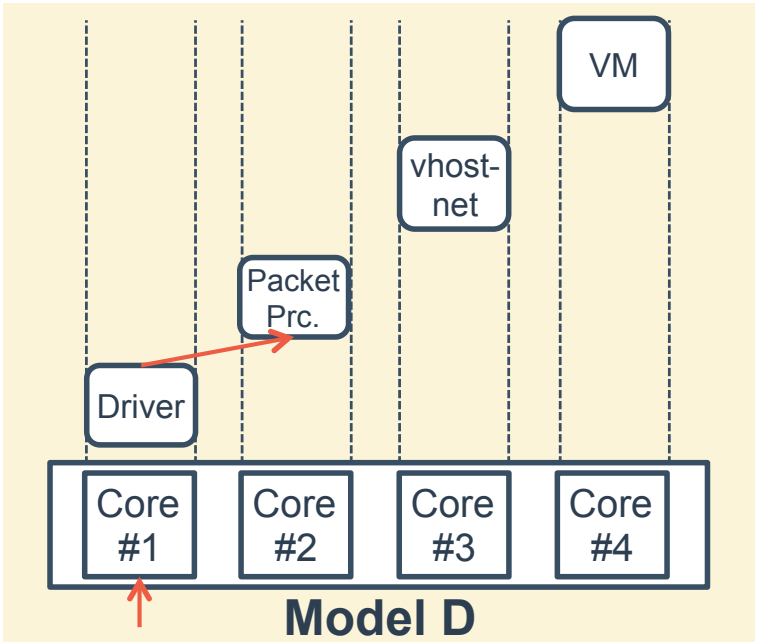
優先フローのためにどの処理を
別のCPUコア上で行うべきか検討

CPUリソース割当てモデル



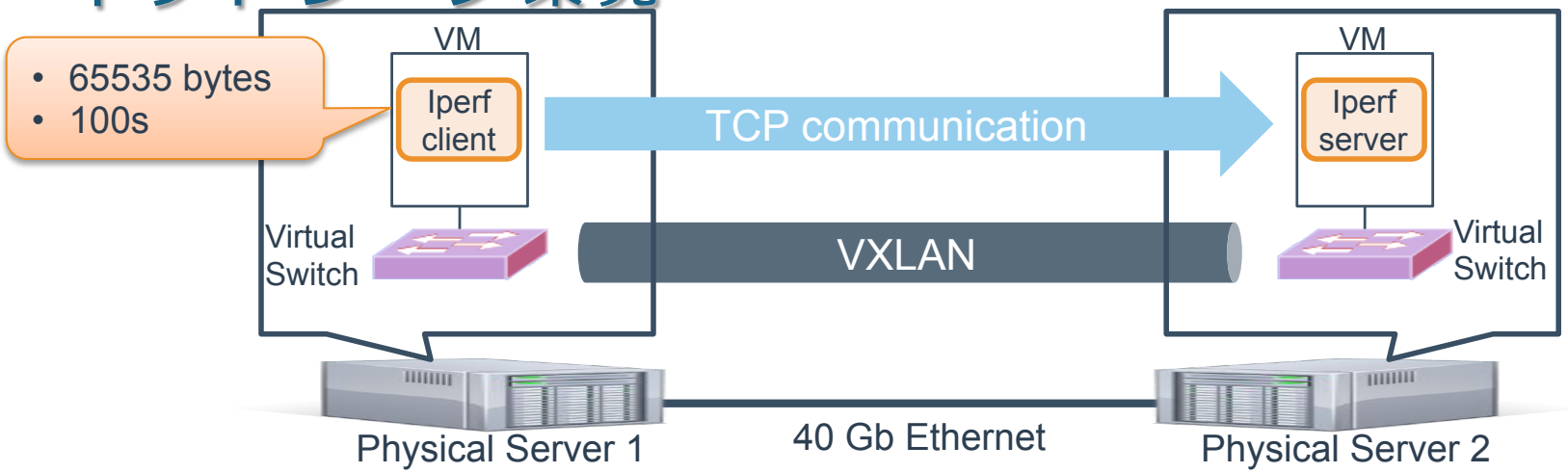






評価環境

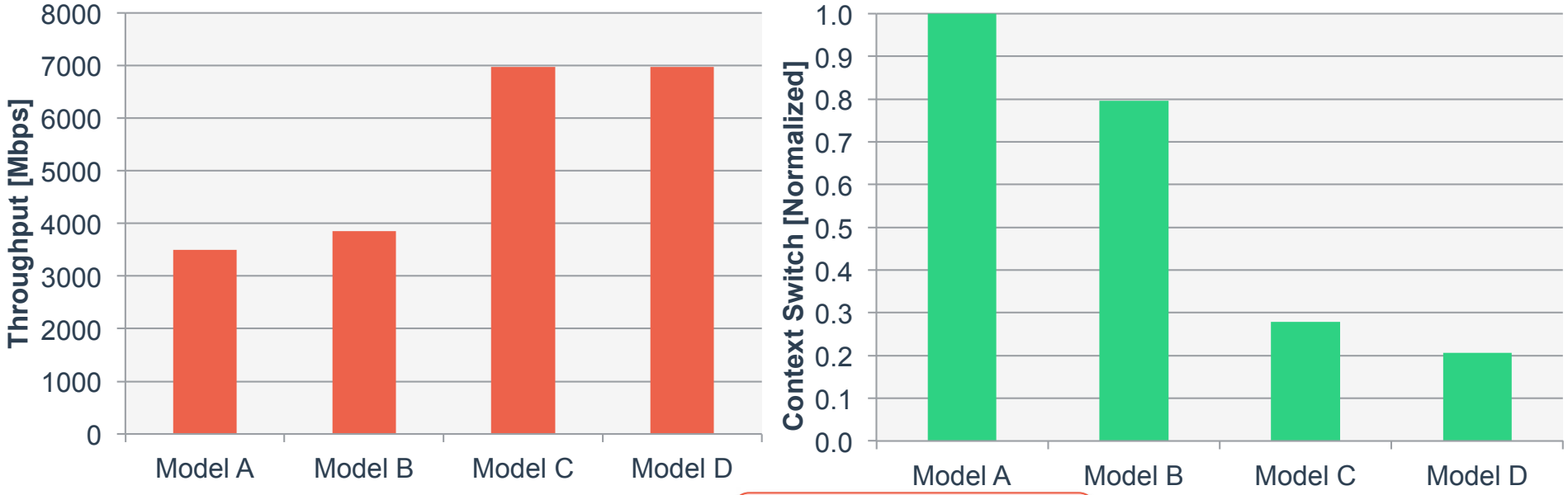
- ネットワーク環境

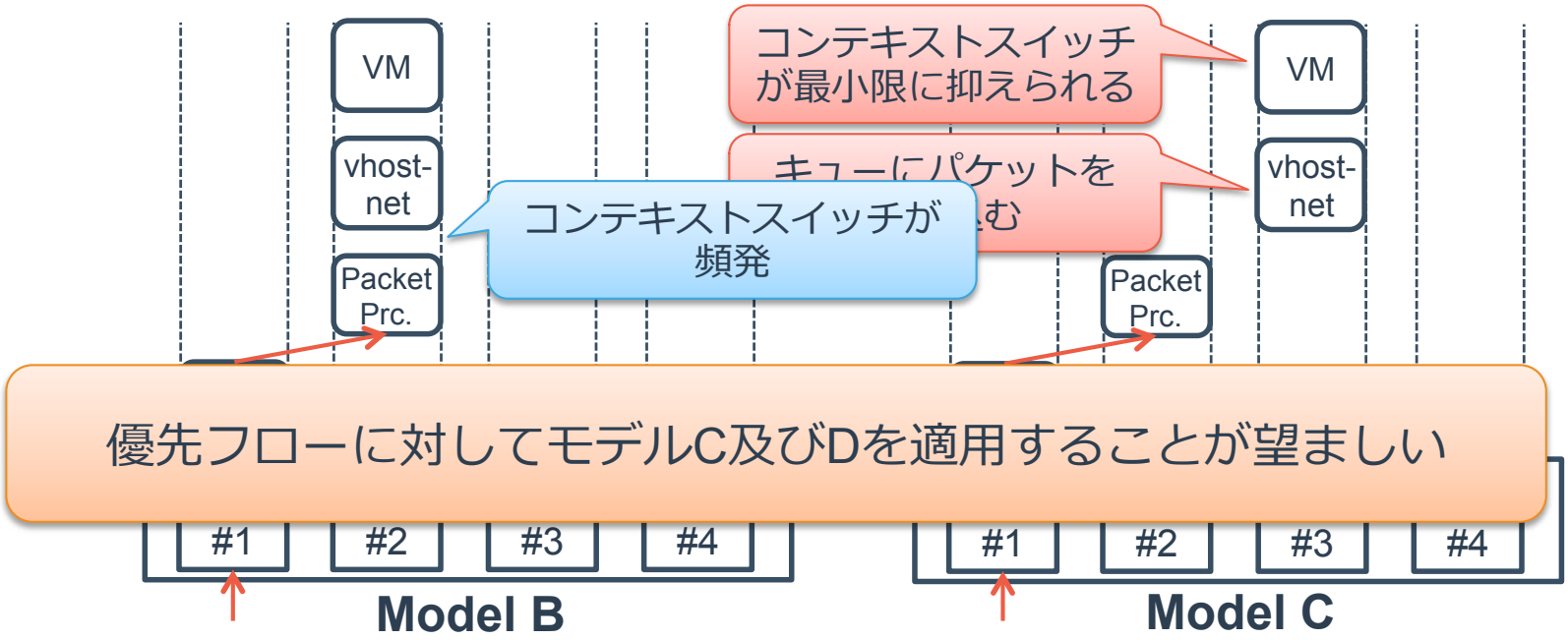


- マシン仕様

	Physical Server 1	Physical Server 2	VM
OS	CentOS 6.6 (2.6.32)		Ubuntu 14.04
CPU	Intel Core i7 (6 core)		1 core
Memory	16G bytes		2G bytes
Buffer	4M bytes		4M bytes
Network	40GBASE-SR4		-
Driver	MellanoxConnect-X(R)-3		virtio (MTU:1450)
vSwitch	OpenvSwitch 2.3.0		-

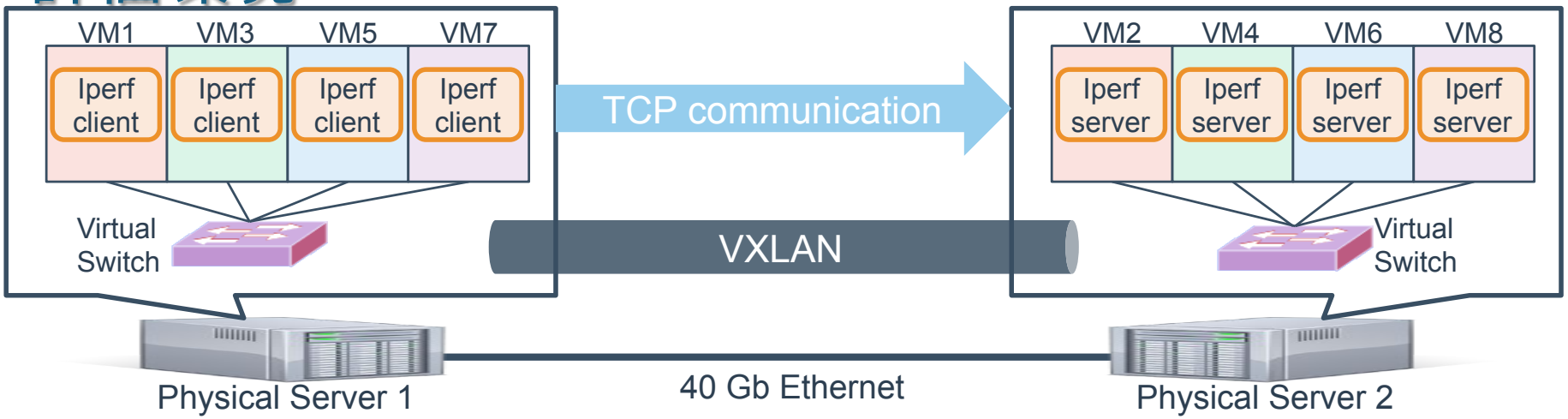
評価結果


Throughput [Mbps]
8000
7000
6000
5000
4000
3000
2000
1000
0
Model A
Model B
Model C
Model D
Context Switch [Normalized]
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Model A
Model B
Model C
Model D


VM
vhost-net
Packet Prc.
コンテキストスイッチが最小限に抑えられる
コンテキストスイッチが頻発
VM
vhost-net
Packet Prc.
優先フローに対してモデルC及びDを適用することが望ましい
#1
#2
#3
#4
Model B
#1
#2
#3
#4
Model C

優先フローの性能評価

- 評価環境

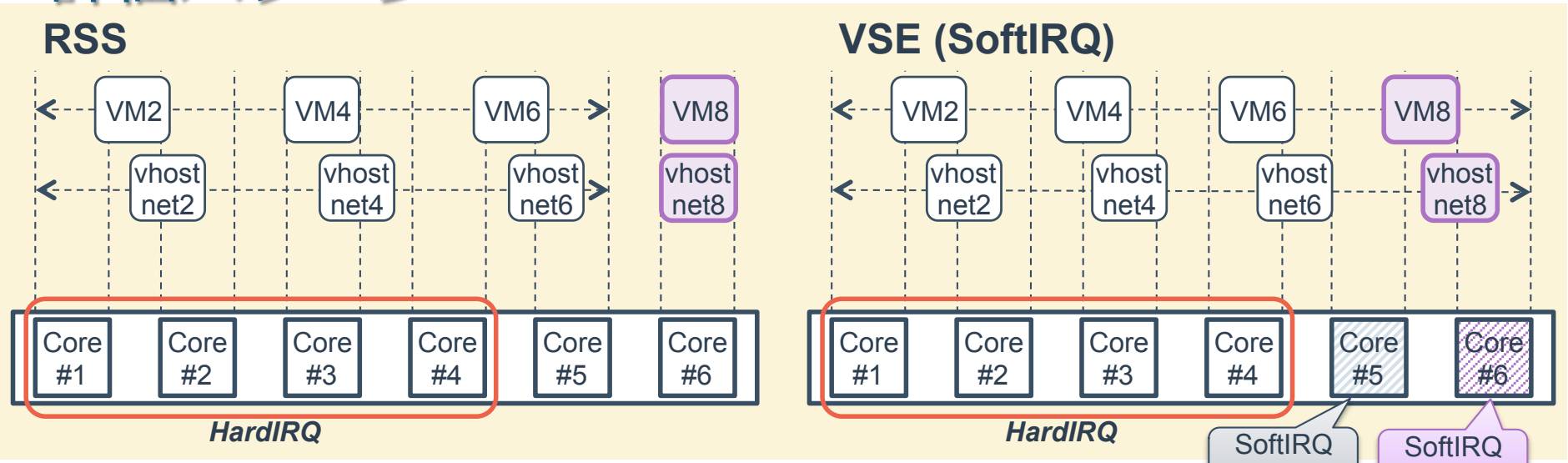


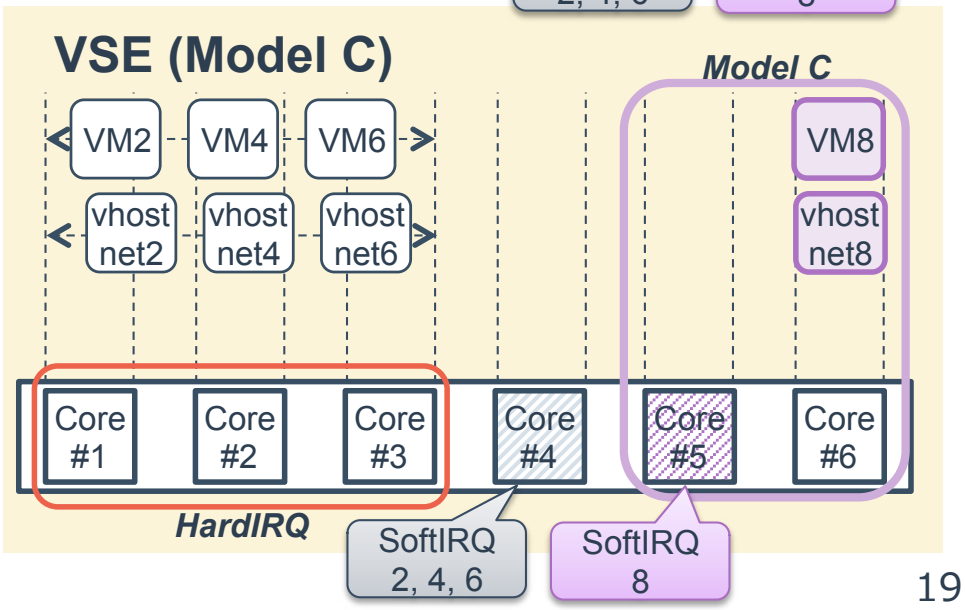
- ベンチマーク：Iperf

Protocol	TCP (Tunnel : VXLAN)
Packet Size	65535 bytes
Flow Duration Time	30 s
Flow Generation Time	15 times
Flow Generation Rules	Common for all patterns

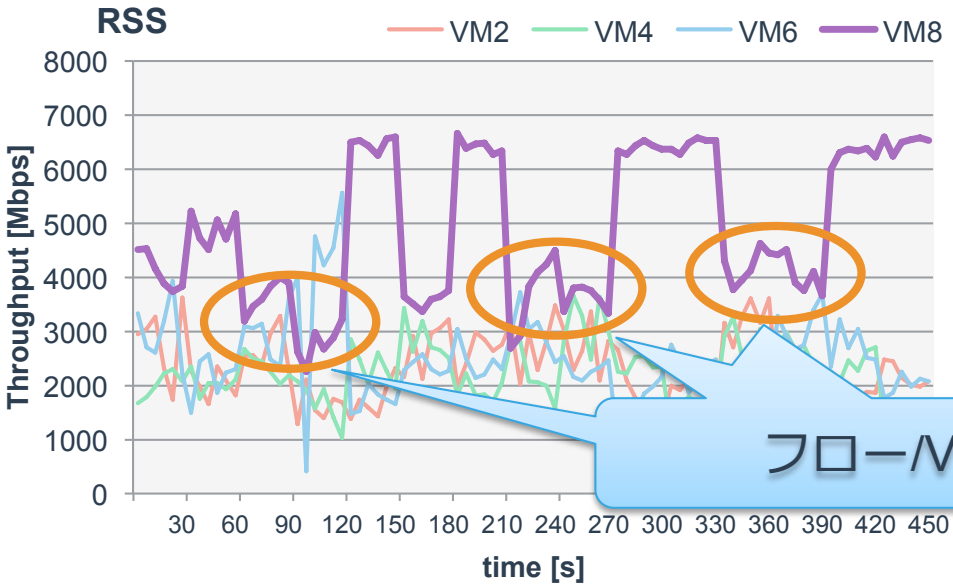
優先フローの性能評価

- 評価パターン

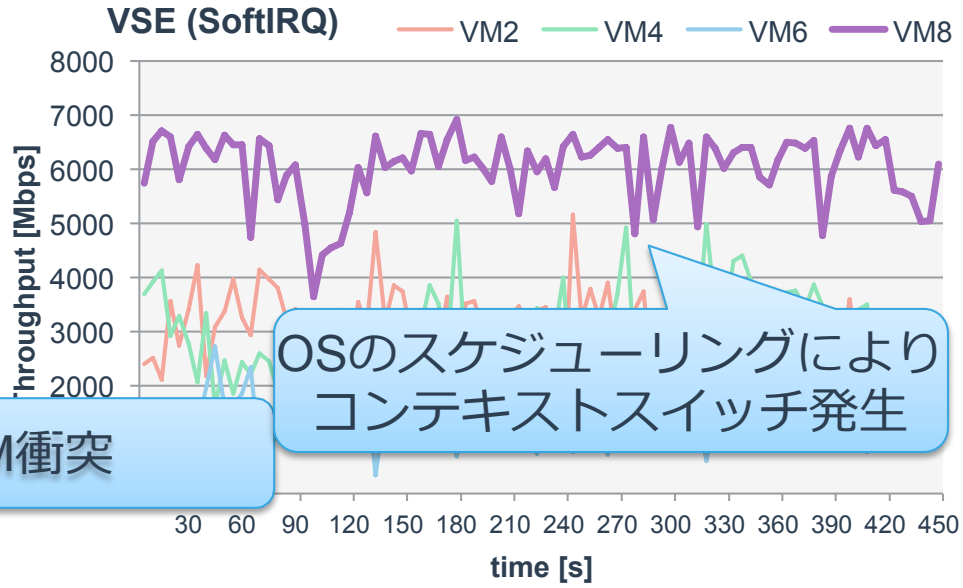




評価結果



フロー/VM衝突

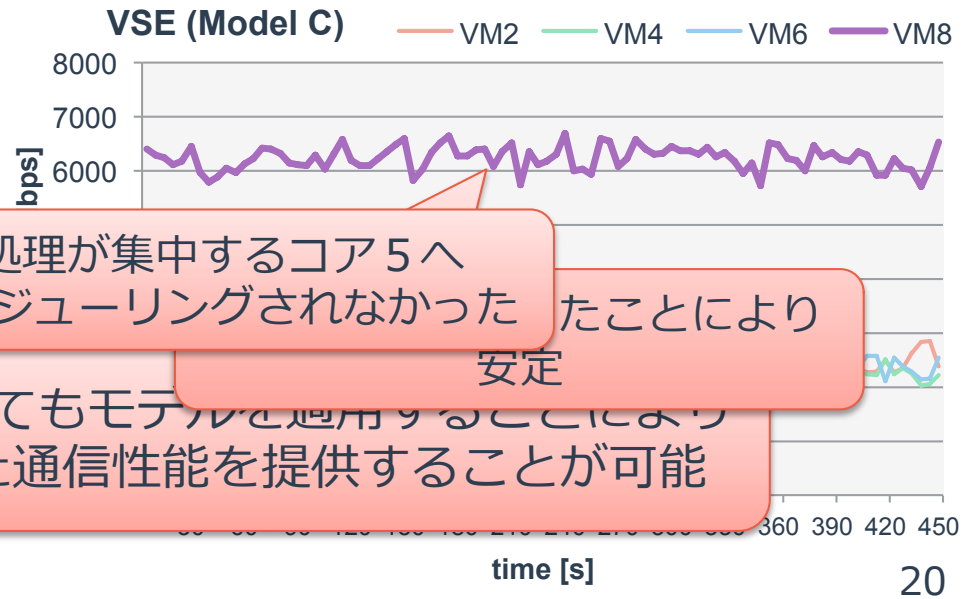


OSのスケジューリングにより
コンテキストスイッチ発生

VMとの衝突により低下

	Tota Ave.	VM8 Ave.
RSS	11934	
VSE (SoftIRQ)	13515	
V		

フロー処理が集中するコア５へ
VMがスケジューリングされなかった



たことにより
安定

フロー処理負荷が低い場合においてもモデルを適用することにより
常に高スループットかつ安定した通信性能を提供することが可能

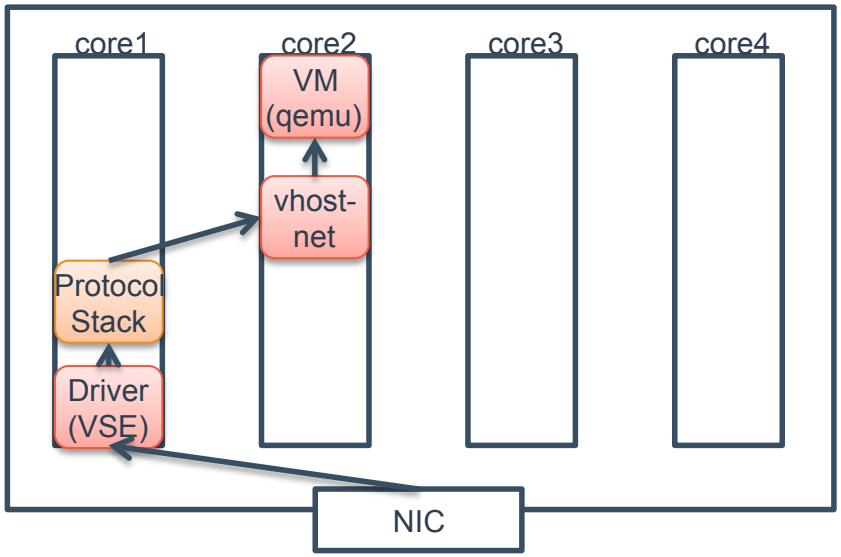
まとめと今後の課題

- まとめ
  - VSE処理オーバヘッドによる通信性能への影響を評価
    - VSE処理オーバヘッドは無視可能
    - VSEに設定可能なエントリ数は10～50
  - CPUリソース割当てモデルを提案
    - フロー処理負荷が低負荷である場合においてもVSEは優先フローに対しモデルを適用することで通信性能が向上かつ安定

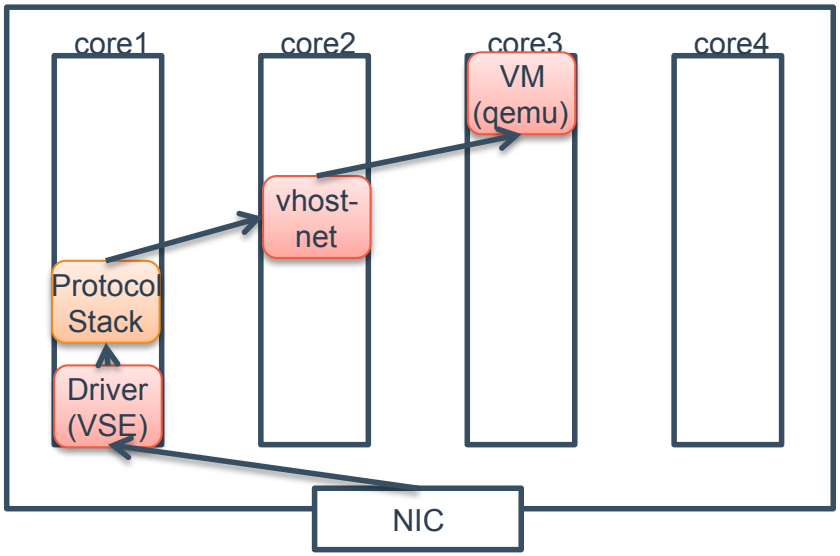
- 今後の課題
  - 各パケット処理における詳細なサービス時間の評価
  - 送信処理を考慮した場合のモデルの検討

パターン5~8

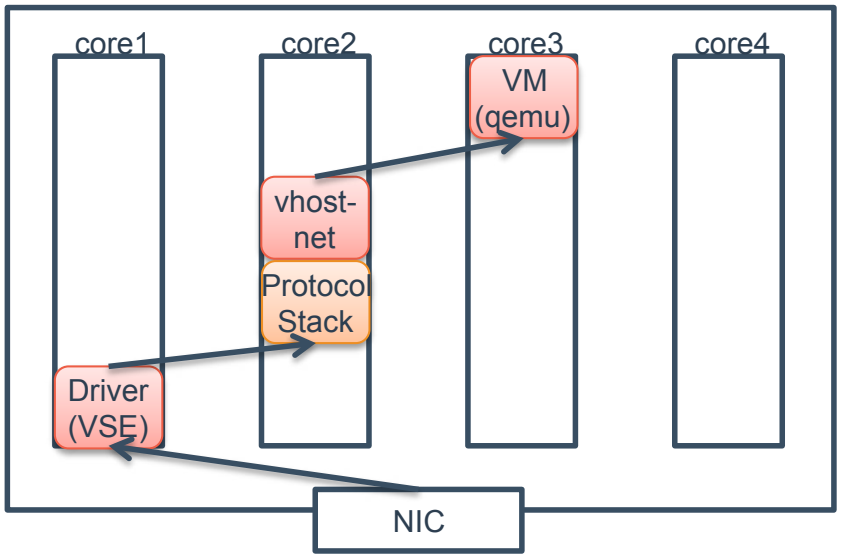
パターン5



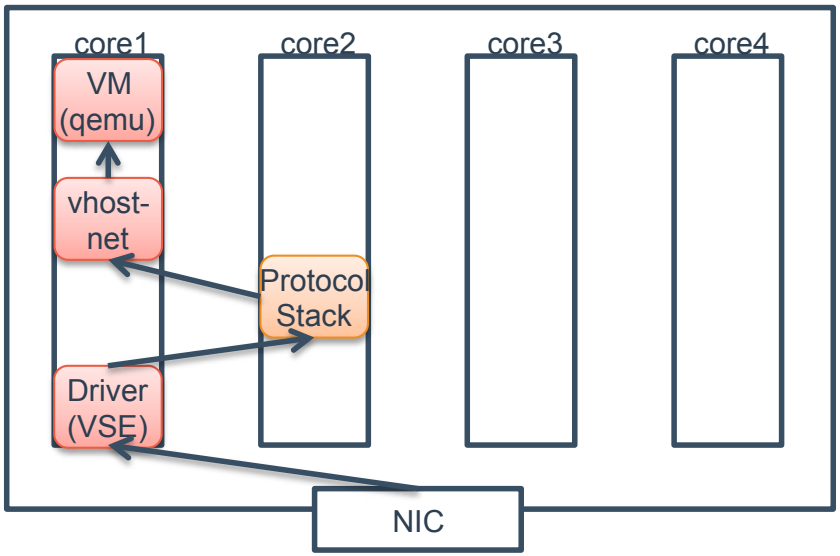
パターン6



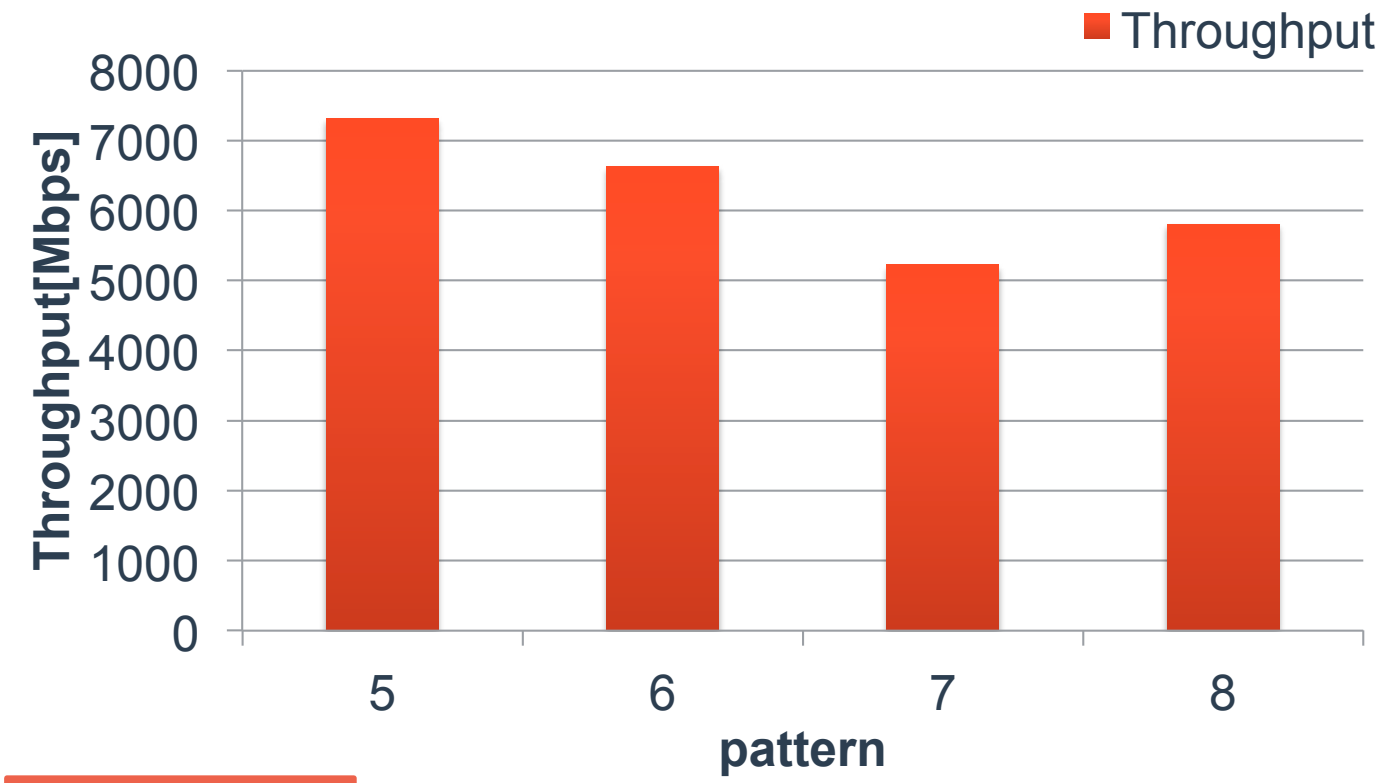
パターン7



パターン8



評価結果：パターン5~8



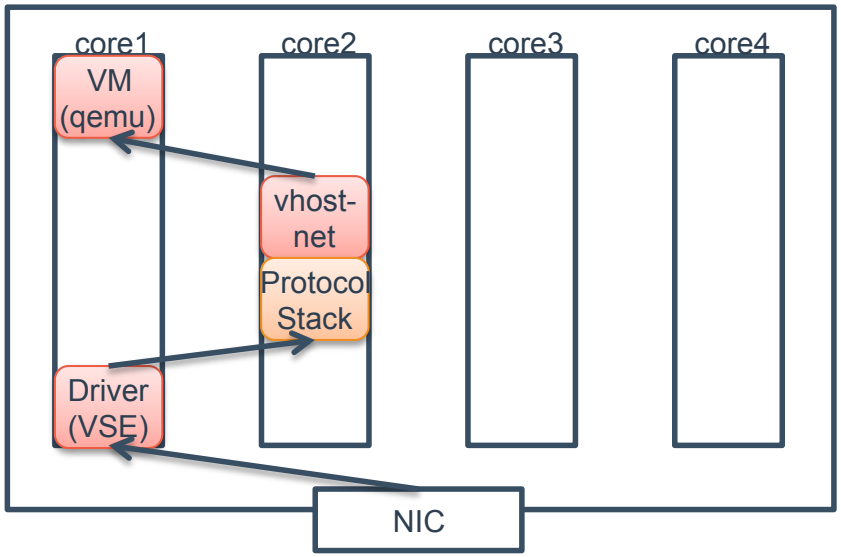
考察

- vhost-netと{HW-IRQ, SW-IRQ}が別々のコアかつvhost-netとVMが同一コアである場合が高速
- SW-IRQとvhost-netが同一コアである場合スループットが減少

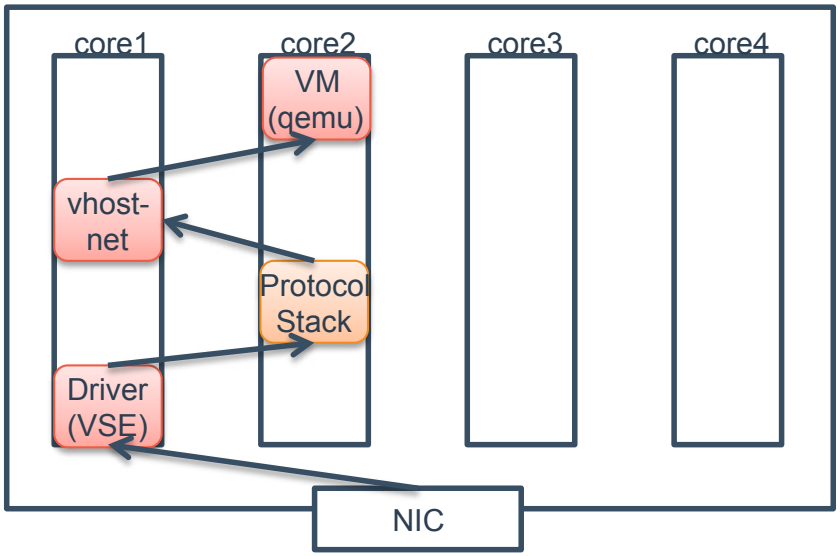
基本的にSW-IRQを行ったコアでvhost-netが動作

パターン9~12

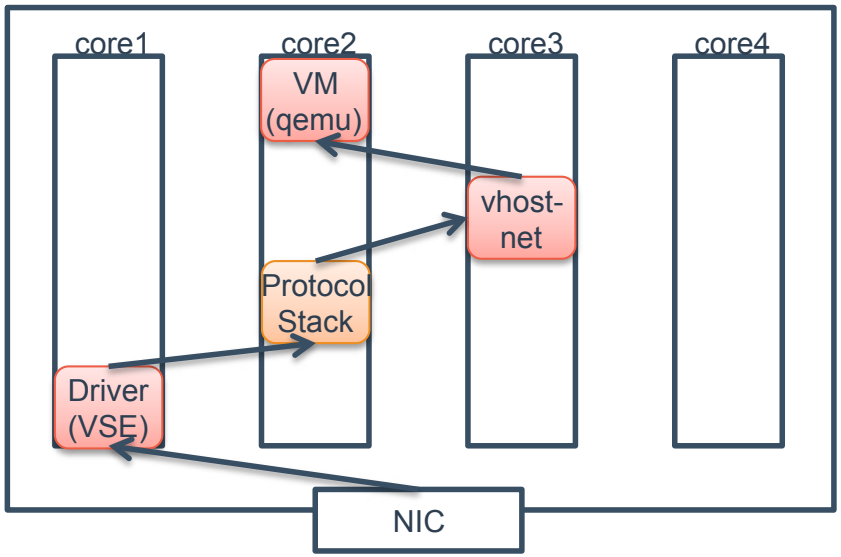
パターン9



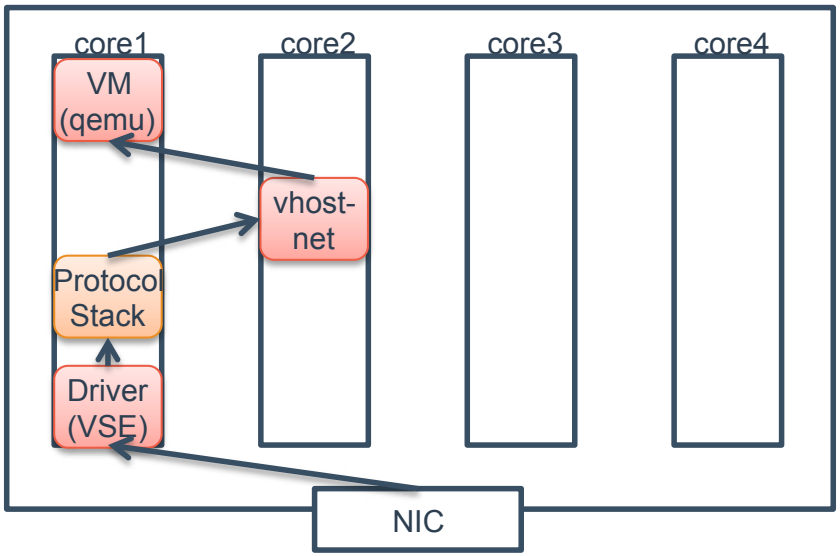
パターン10



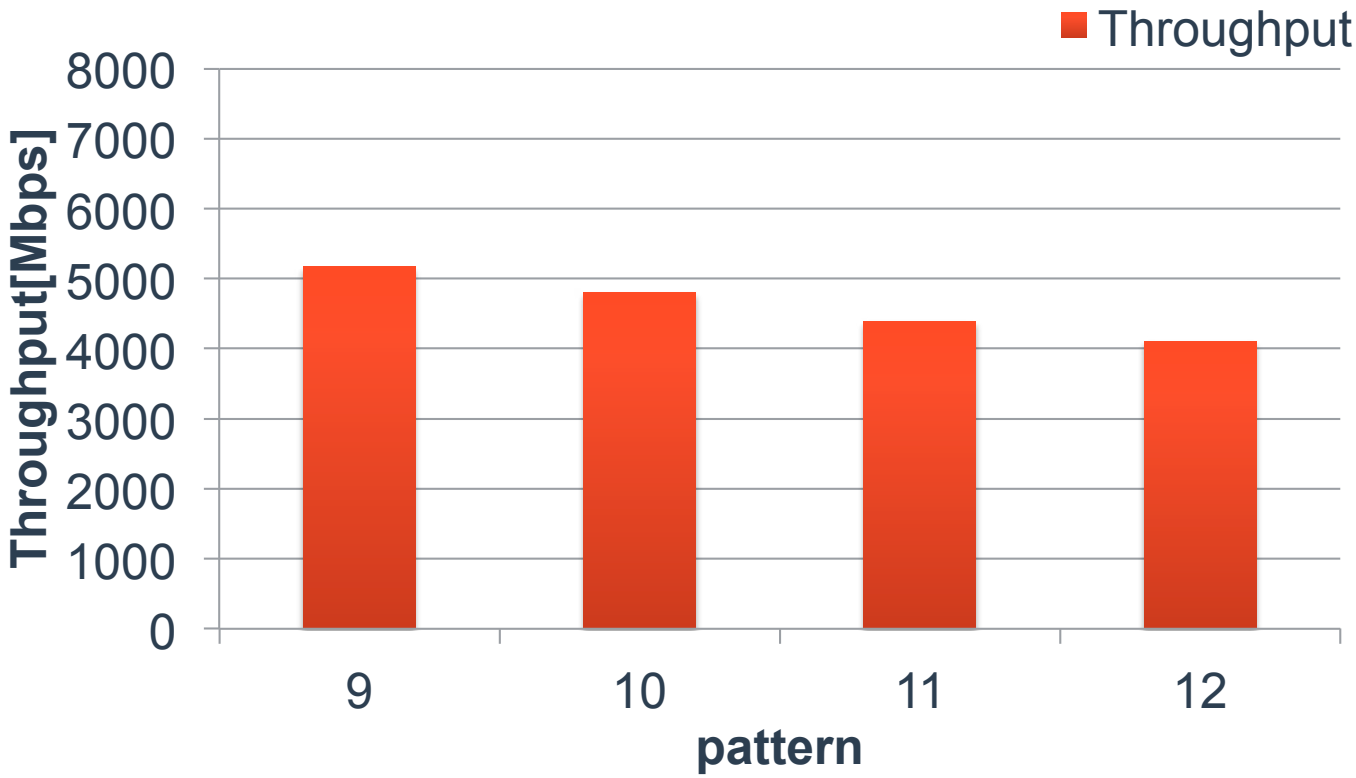
パターン11



パターン12



評価結果：パターン9~12

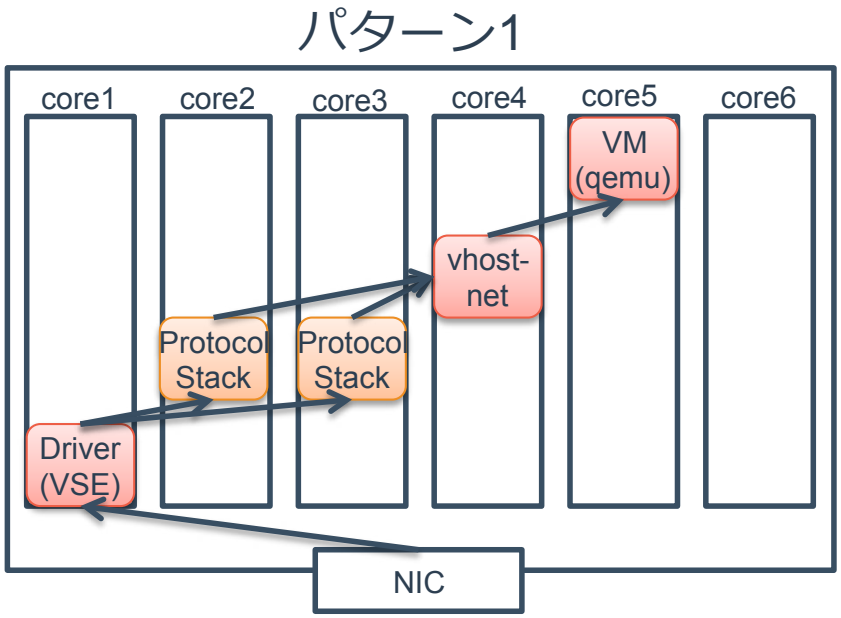


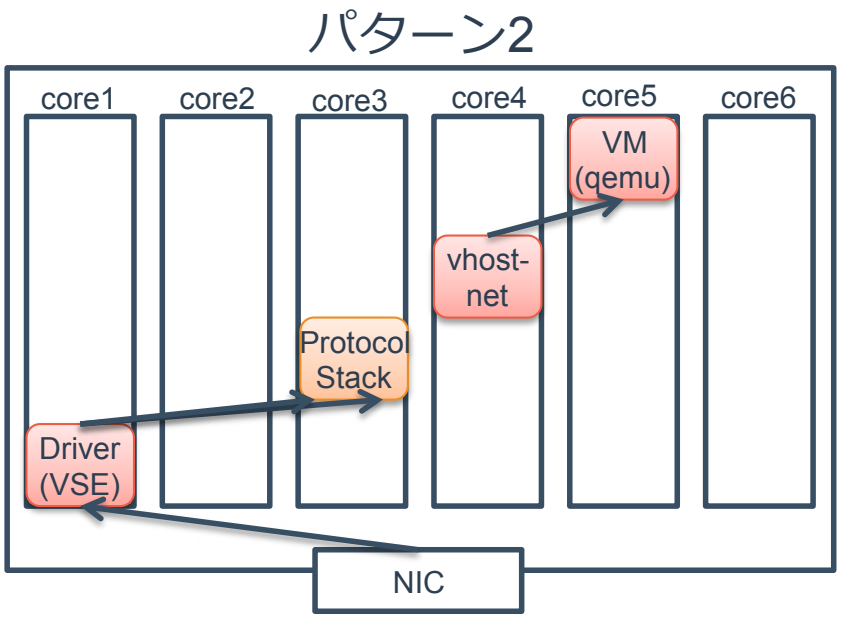
考察

- SW-IRQとVMが同一コアで動作する場合スループットが5Gbps以下になる
- コンテキストスイッチのオーバヘッド？

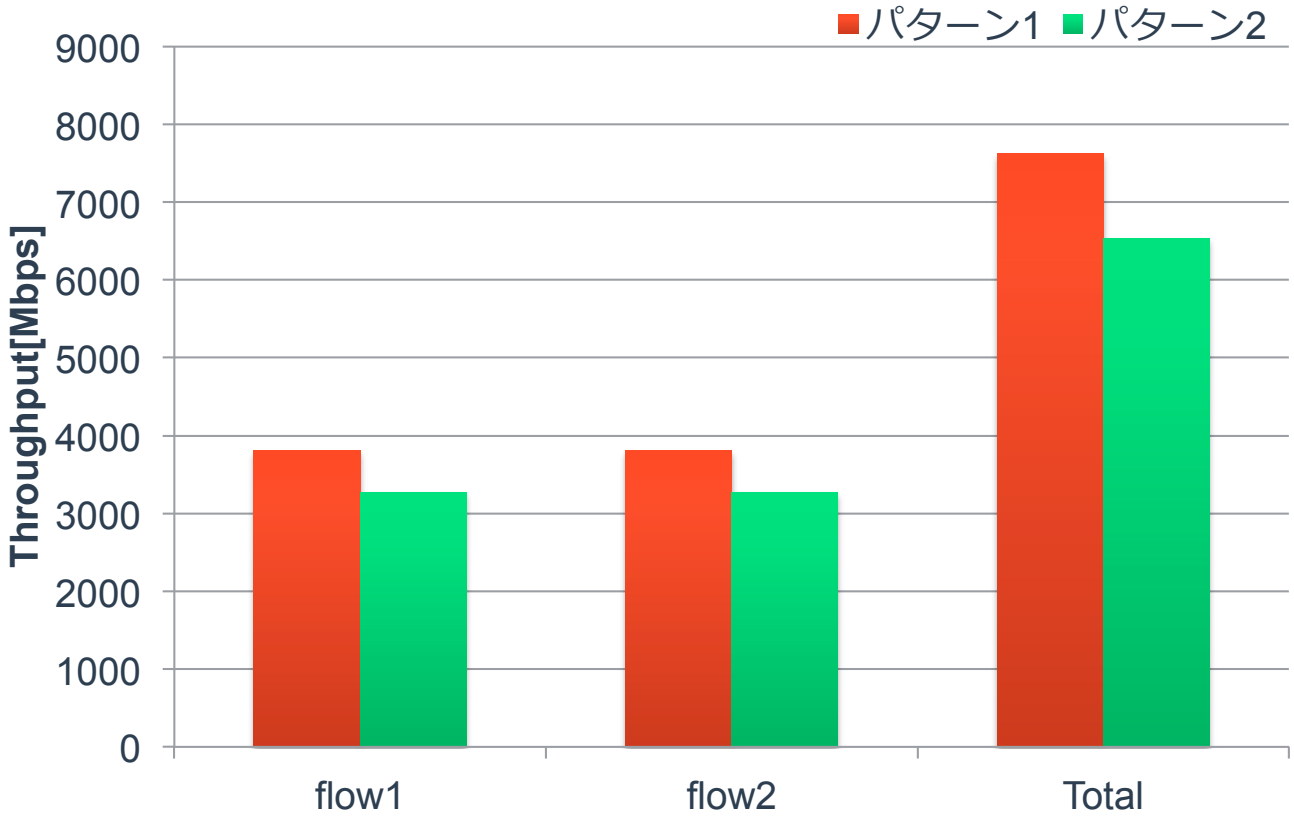
1VM-2Flows

- vhost-netがロックを使用？
  - 分散すると性能低下するか検証





評価結果

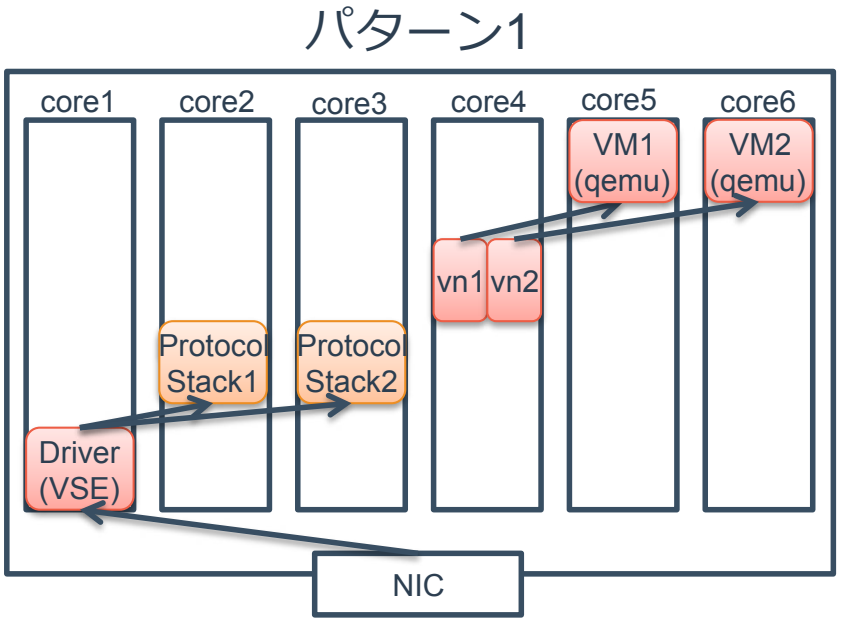


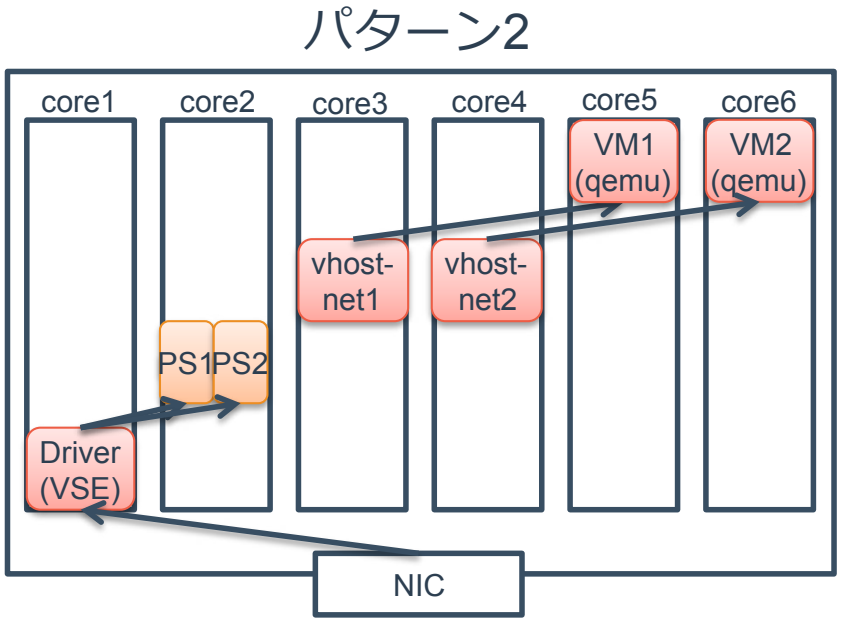
考察

- 2つのフローを別々のコアで処理しても遅くならなかった
  - むしろ若干速くなった
  - SW-IRQによるボトルネックが一番大きい？

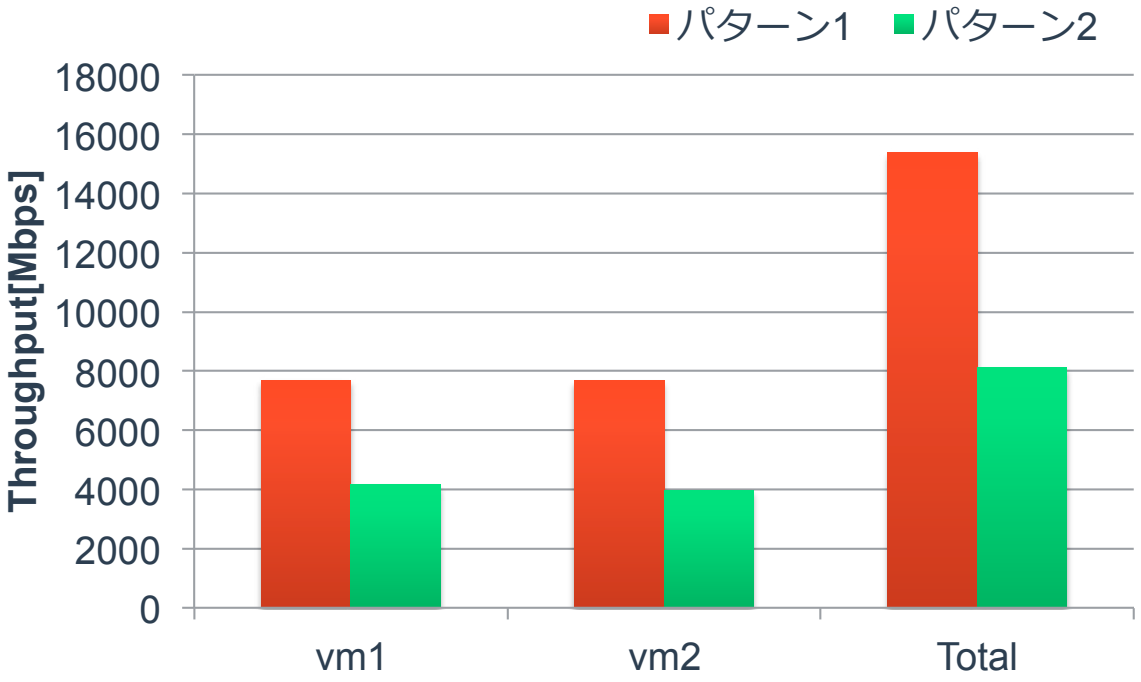
2VM-2Flows

- vhost-netが悪いのか，SW-IRQが悪いのか・・・





評価結果



考察

- 2つのvhost-netが衝突している場合よりもSW-IRQ先が衝突している時のスループットが大きく低下している
- SW-IRQ(Protocol Stack)の処理オーバヘッドが大きい