NFV網路加速技術介紹

工研院資通所人工智慧運算平台組
李育緯 技術經理
Email: rayinlee@itri.org.tw

講者介紹

李育緯 Ray

現任：

工研院資通所人工智慧運算平台組技術經理

(前資料中心架構與雲端應用組)

專案經歷：

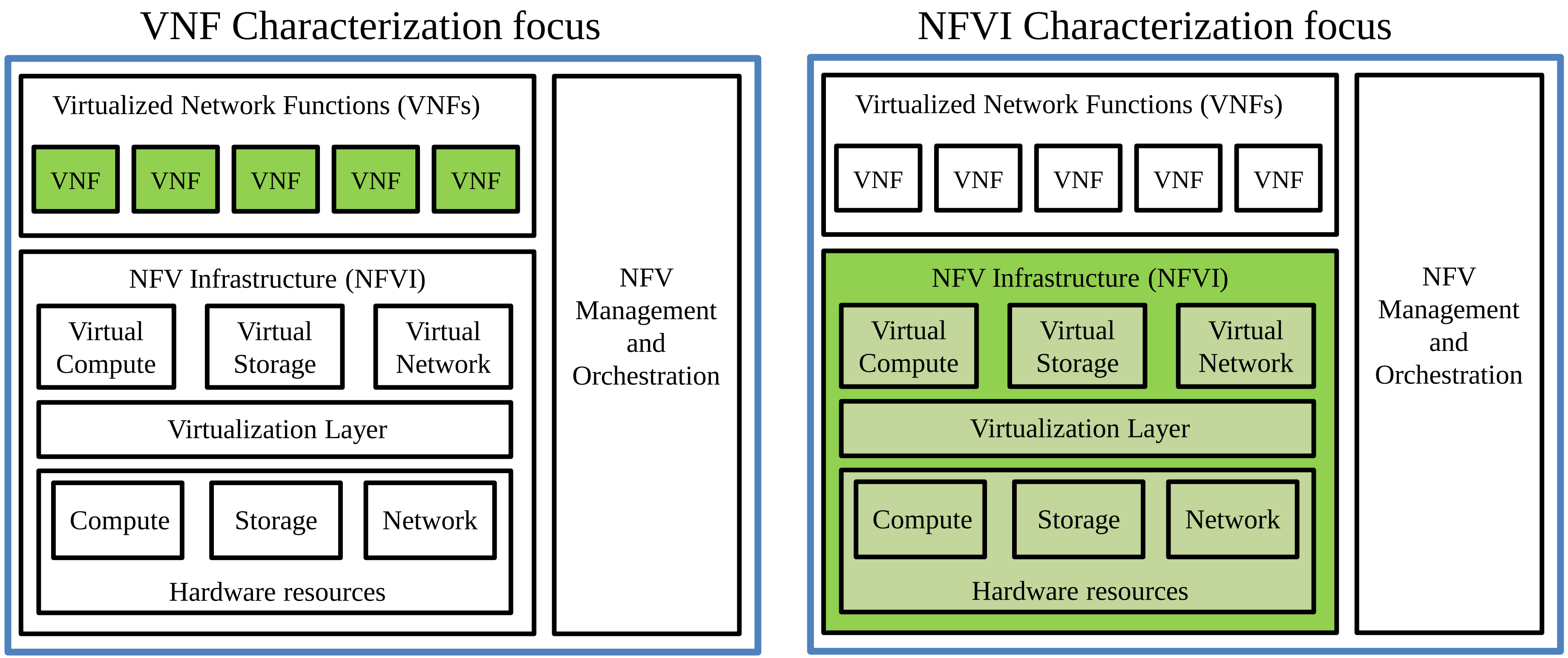
- **NFV效能實驗室技術推廣**
- **5G NFVI平台研發與5G專網系統整合**
- **混合式軟體定義網路系統研發**
- **國產Cloud OS資料中心網路系統研發**

Outline

- NFV Performance Lab介紹
  - NFVI效能測試與優化技術
  - SDWAN效能測試與優化技術
  - Service Function Chain效能測試與優化技術

NFV Performance Lab

- Intel and ITRI build the NFV performance lab cooperatively
- Goal: NFV performance characterization
- NFVI characterization requires some “VNF”
- VNF characterization requires some NFV infrastructure

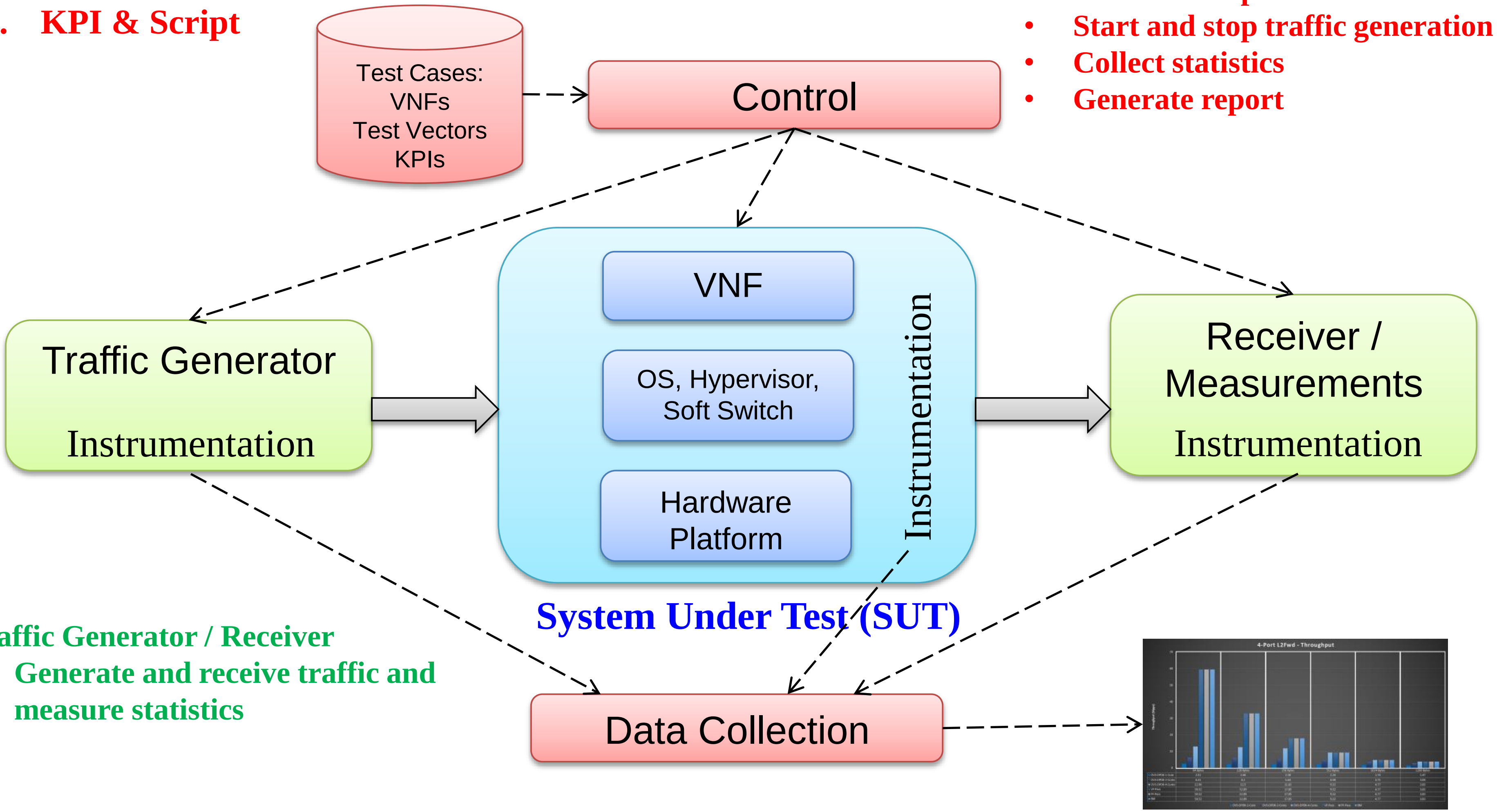


VNF and NFV Infrastructure Characterization

1. System Under Test (SUT)
2. Traffic Generator
3. KPI & Script

Control and Data Collection

- Test cases measurement
- Start and stop SUT workload
- Start and stop traffic generation
- Collect statistics
- Generate report



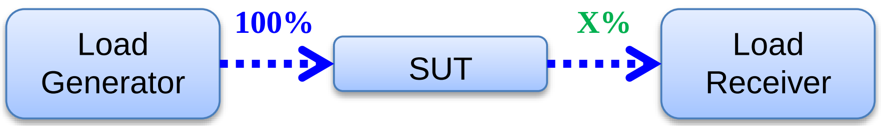
Traffic Generator / Receiver

- Generate and receive traffic and measure statistics

Throughput

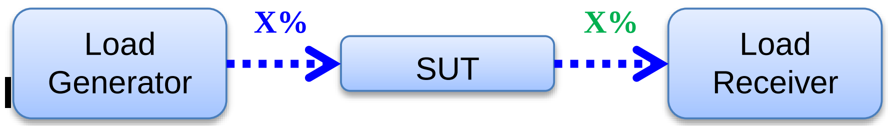
Forwarding Rate at Maximum Offered Load (FROML, RFC 2889)

- Generate at line rate, and measure forwarded packets
- Easy to run, fast
- Measure high load behavior (overload?)
- Might represent a completely different number than maximum forwarding rate



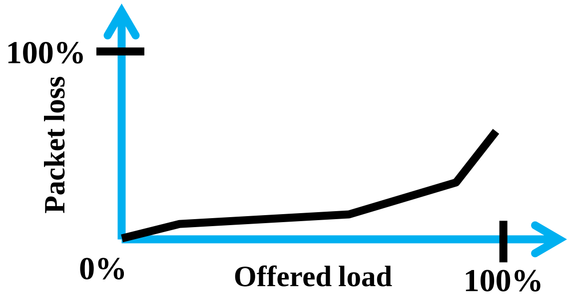
Maximum Forwarding Rate (MFR, RFC 2544)

- Start generating at 100% line rate and binary search for higher rate without packet loss
- Quite fast
- Might be very sensitive to spurious packet l



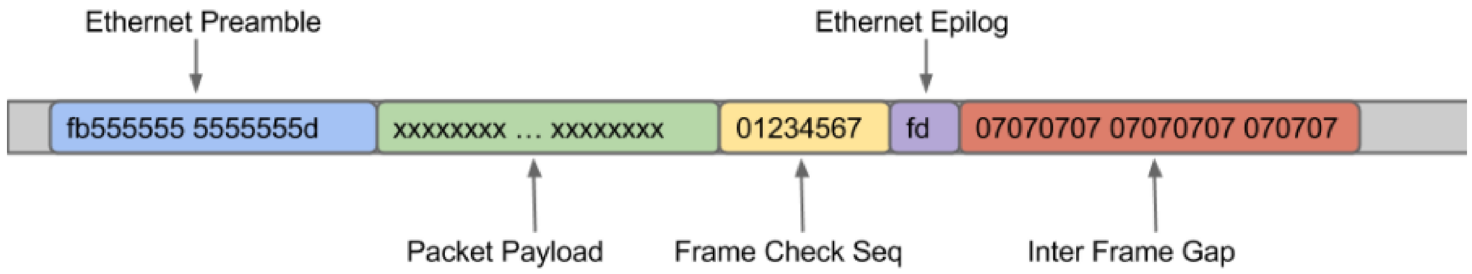
Measure packet loss for any rate, starting from 0.1% to 100% of line rate

- Slow
- Better picture of the performance
- Will highlight spurious packet loss



何謂10G Bit Line Rate

- **在NFV的領域，因為使用CPU來執行虛擬化網路功能(Router、Switch、Firewall等)，在討論NFV的效能時，我們要了解所謂的10G Bit究竟有多少封包要處理。**
- **通常我們以64 Bytes的封包大小來估算應該要處理的封包數量**
  - **64 Bytes封包包含60 Bytes的Payload與4 Bytes的Frame Check Sequence，再加上L1的20 Bytes的Overhead，封包大小總共是84 Bytes。**
    - **10×10^9 / 84 x 8 = 14.88 x 10^6 (14.88百萬個封包)，每秒要處理14.88百萬個封包，相當於1個封包只有67ns的時間可以處理**
    - **DDR4的Memory Access Latency約15ns**
- **若封包大小是1500 Bytes的話，又會如何呢?**
  - **1500 Bytes封包，加上L1 20Bytes的Overhead，封包大小總共是1520 Bytes**
    - **10×10^9 / 1520 x 8 = 822,368 (82萬個封包)，每秒要處理82萬個封包，相當於1個封包有1216ns的時間可以處理。**



Preamble	blue	8 bytes
Payload	green	variable
Frame Check Sequence	yellow	4 bytes
Epilogue	purple	1 bytes
Inter Frame Gap (+Epilogue its 12B)	Red	11 bytes

Source: https://fmad.io/blog-what-is-10g-line-rate.html

Intel Xeon Processor E5-2695 v4
NFVI Performance Report

Produced by ITRI Performance Lab

NFVI Characterization

VNF:
L2-Fwd
L3-Fwd
ACL
MPLS

Packet Size:
64 Bytes
128 Bytes
256 Bytes
512 Bytes
1024 Bytes
1280 Bytes

Test Cases:
VNFs
Test Vectors
KPIs

Control

Control and Data Collection
- **Test cases measurement**
- **Start and stop SUT workload**
- **Start and stop traffic generation**
- **Collect statistics**
- **Generate report**

Traffic Generator

Instrumentation

VNF

OS, Hypervisor, Soft Switch

Hardware Platform

Instrumentation

Receiver / Measurements

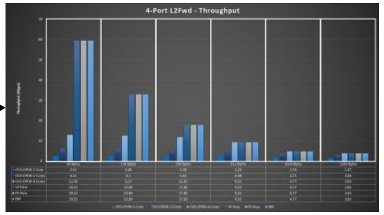
Instrumentation

System Under Test (SUT)

Traffic Generator:
PROX

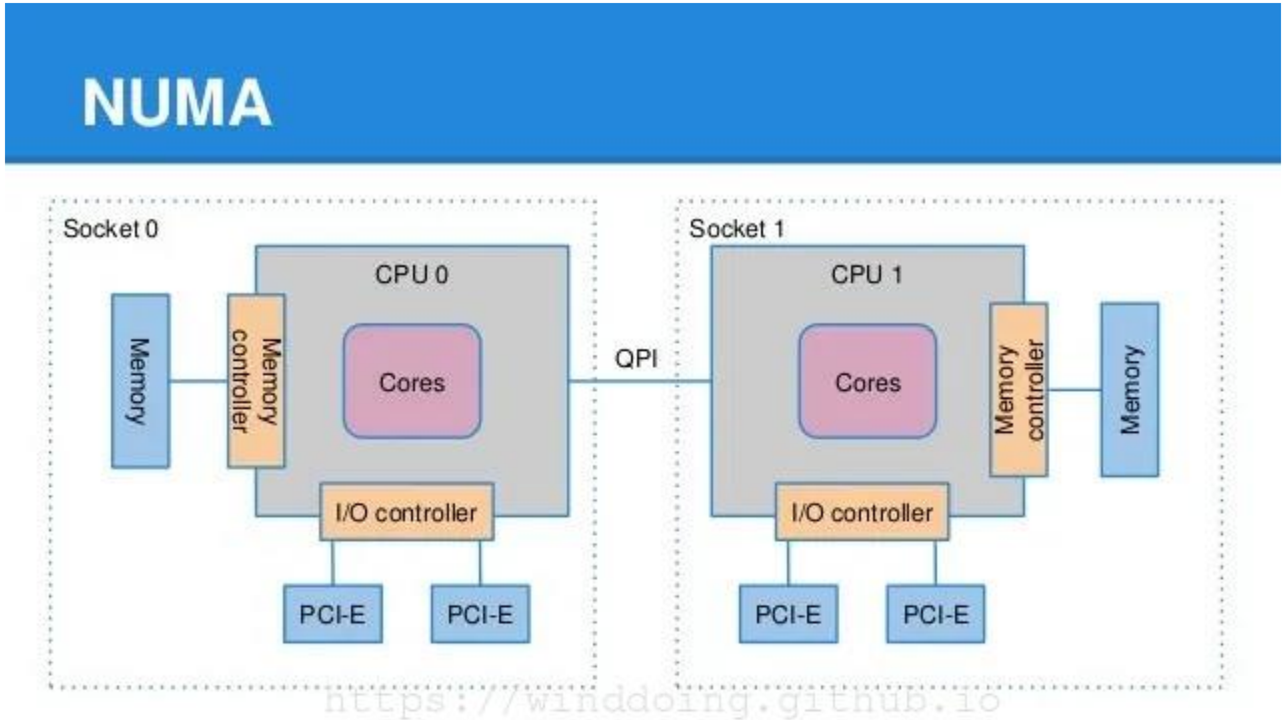
Traffic Generator / Receiver
- **Generate and receive traffic and measure statistics**

Data Collection



NUMA概念解說

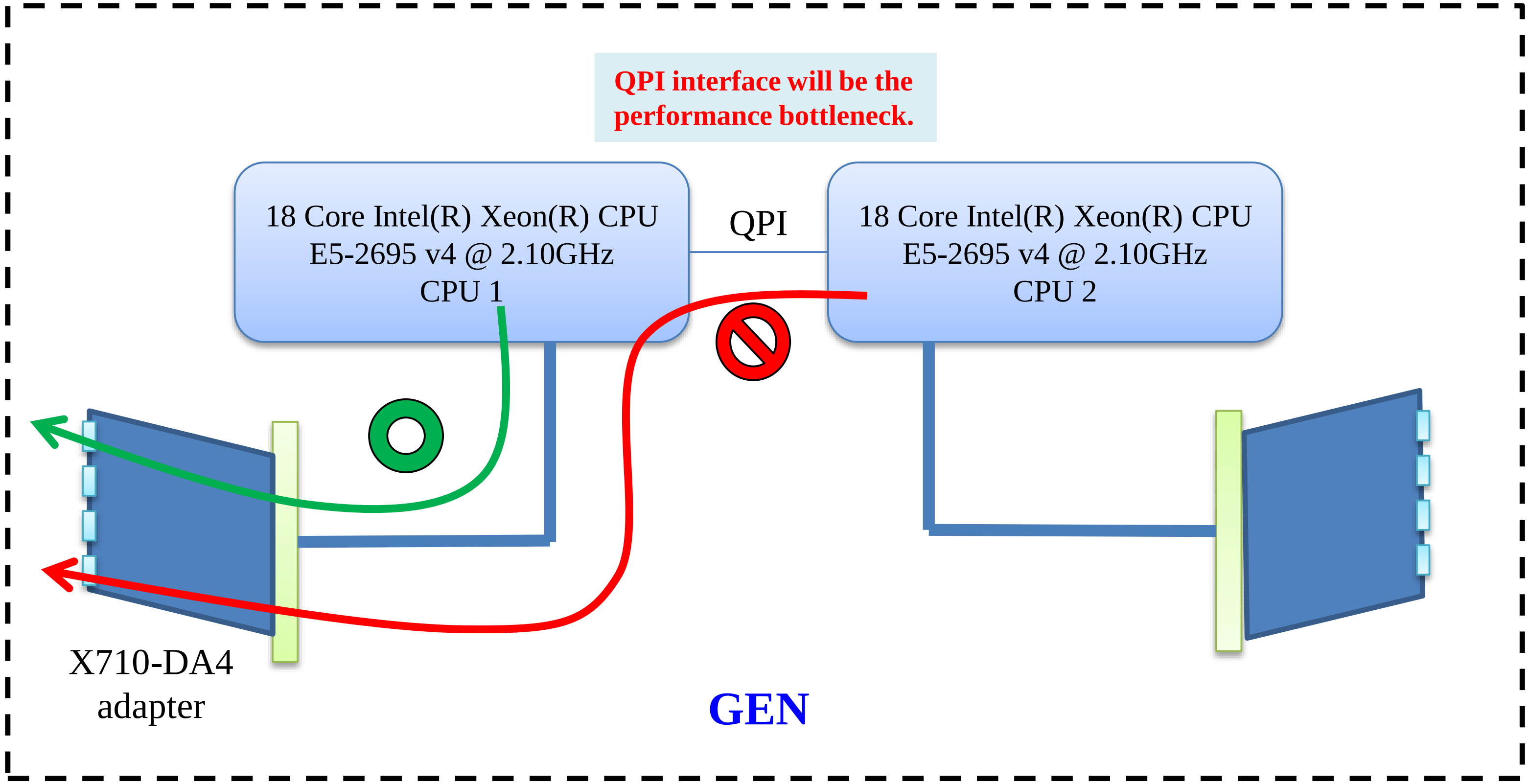
- **Non-uniform memory access (NUMA)系統是指一個主機板上有多個CPU，各CPU有各自的記憶體與IO控制器，CPU之間透過QPI介面可存取遠端的記憶體與IO設備，但CPU存取本地記憶體與IO設備的延遲與頻寬會優於存取遠端的記憶體與IO設備。**



Source: https://winddoing.github.io/post/13d4e2a6.html

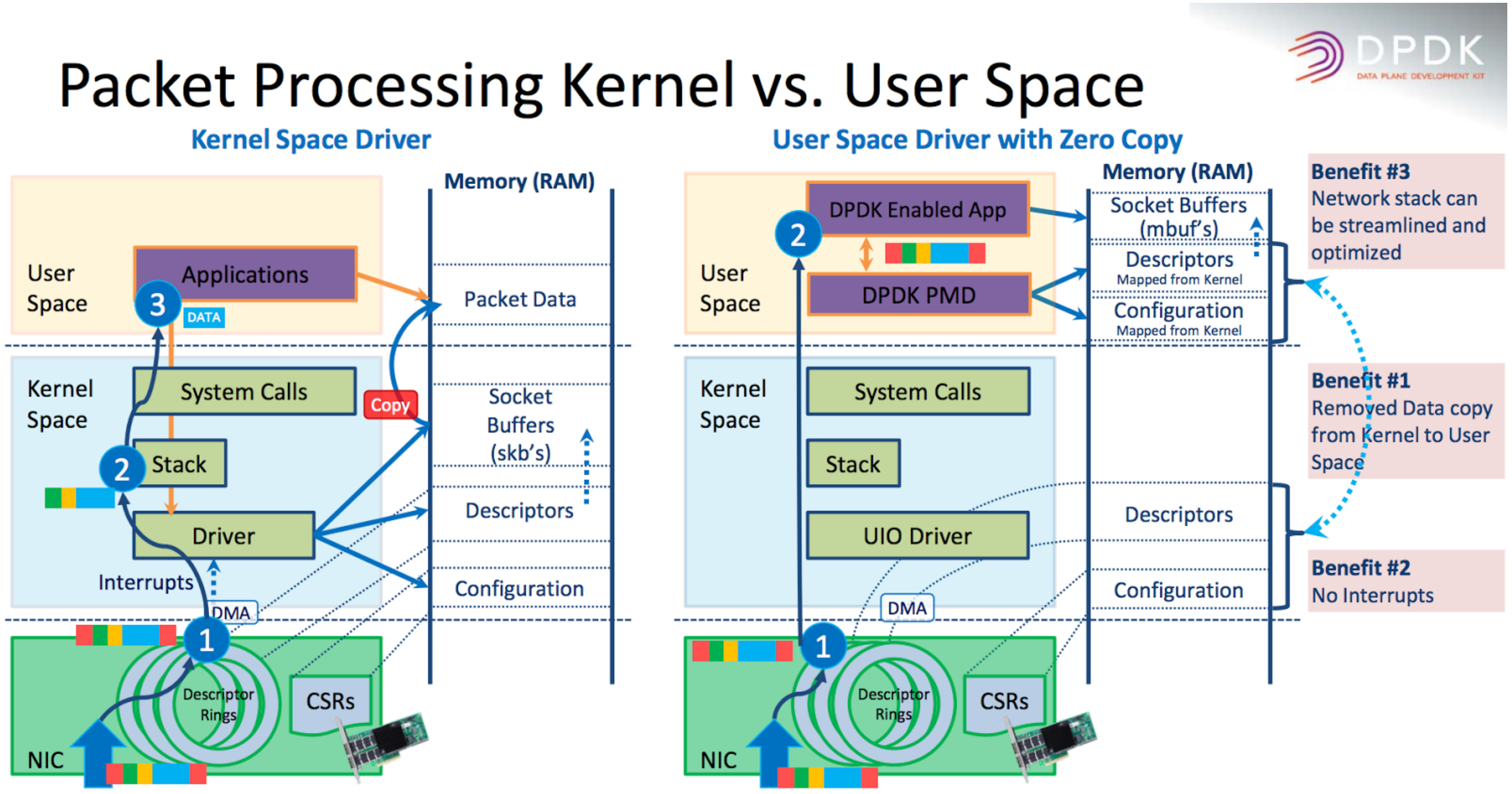
The Mapping Between CPU and NIC

- **The CPU core and NIC for packet generator should be in the same CPU socket, or the QPI interface will be the performance bottleneck.**
  - **10GbE line rate (64bytes packet) = 10.00e9 bits / (8 bits * (64 + 20)bytes) = 14.88e6 packets.**
  - **CPU and NIC in same socket, 10G port could generate 14.88 Mpps.**
  - **CPU and NIC in different socket, 10G port only could generate 10 Mpps.**



DPDK介紹

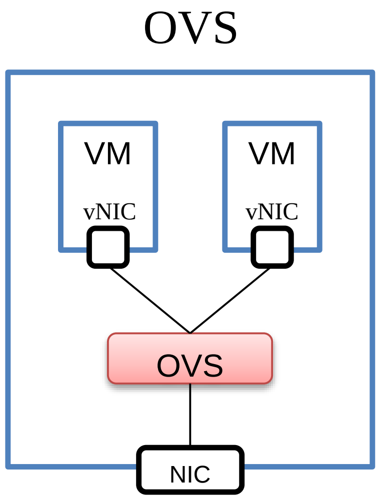
- **DPDK (Data Plane Development Kit)是由Intel提供的開發工具集，不同於Linux系統以通用性設計為目的，而是專注於網路應用中數據封包的高性能處理。**

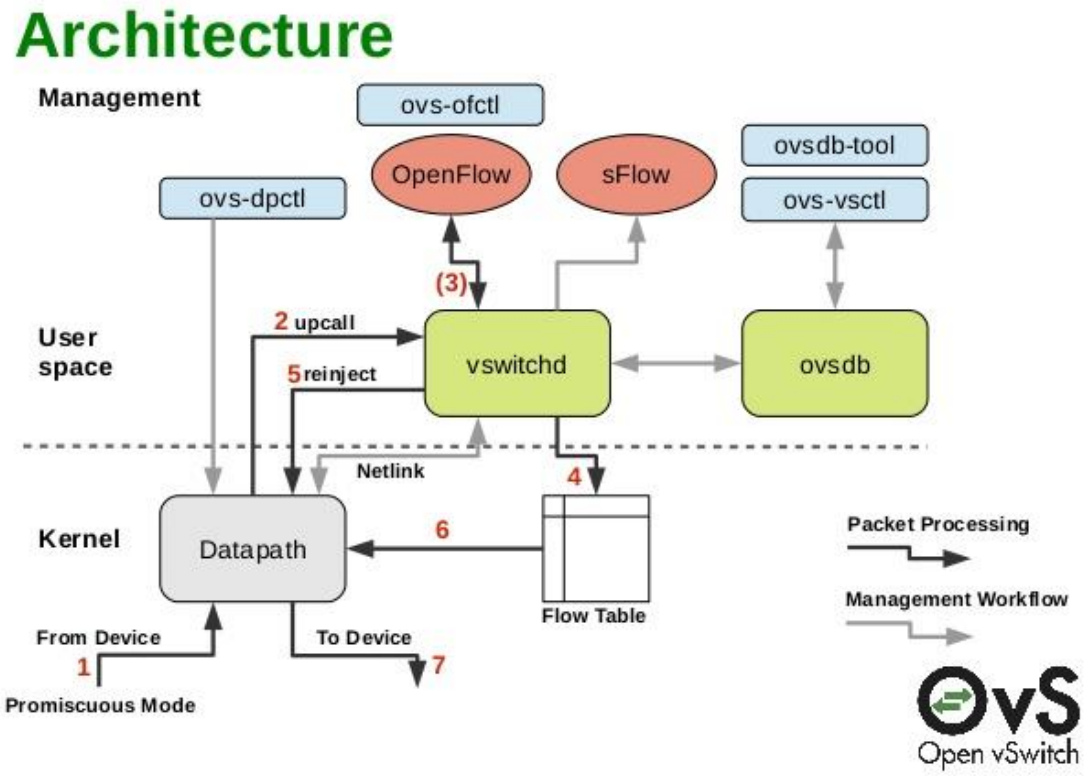


Source: https://www.slideshare.net/MichelleHolley1/dpdk-1805-inflection-point

OVS介紹

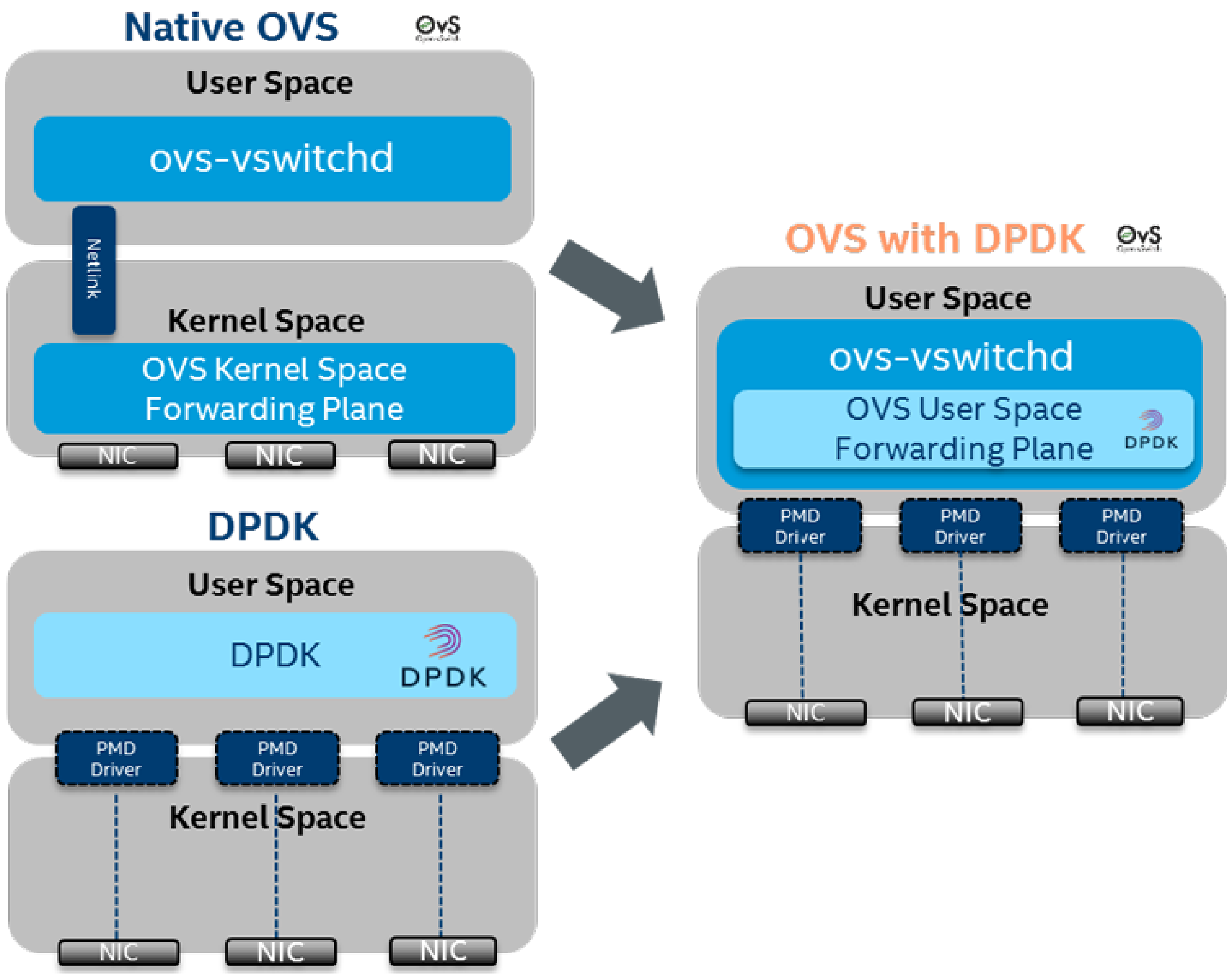
- **Open vSwitch(OVS)是開源的虛擬交換器，支援VLAN/VxLAN/NVGRE等網路隔離功能，也支援QoS、sFLOW與OpenFlow協定。**





Source: https://hustcat.github.io/an-introduction-to-ovs-architecture/

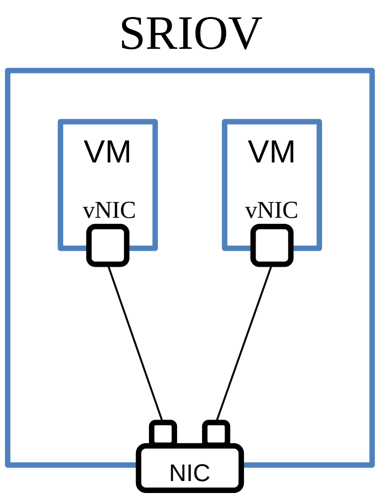
OVS + DPDK

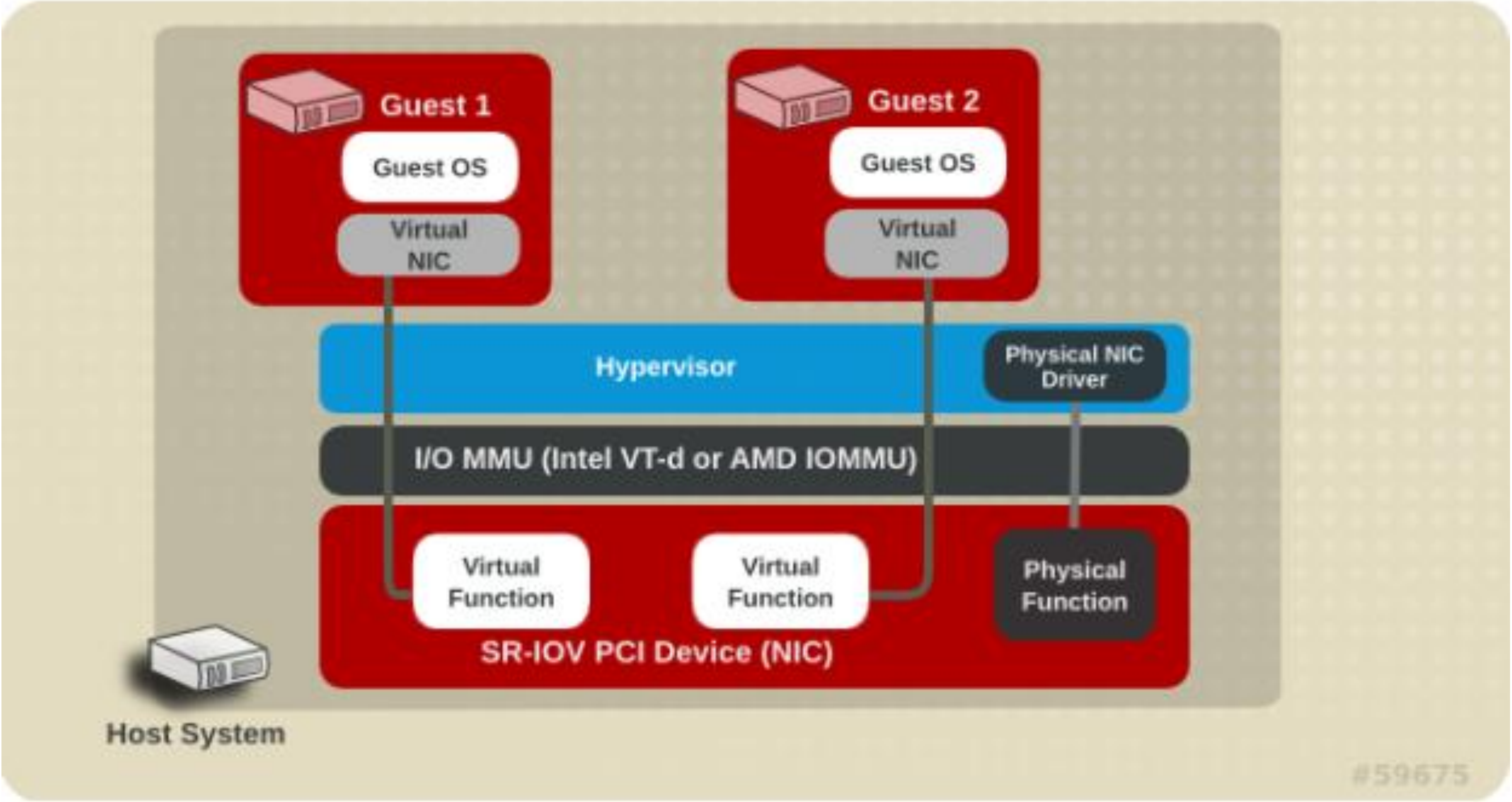


Source: https://sites.google.com/a/cnsrl.cycu.edu.tw/da-shu-bi-ji/openvswitch/dpdk-ovs

SRIOV網卡

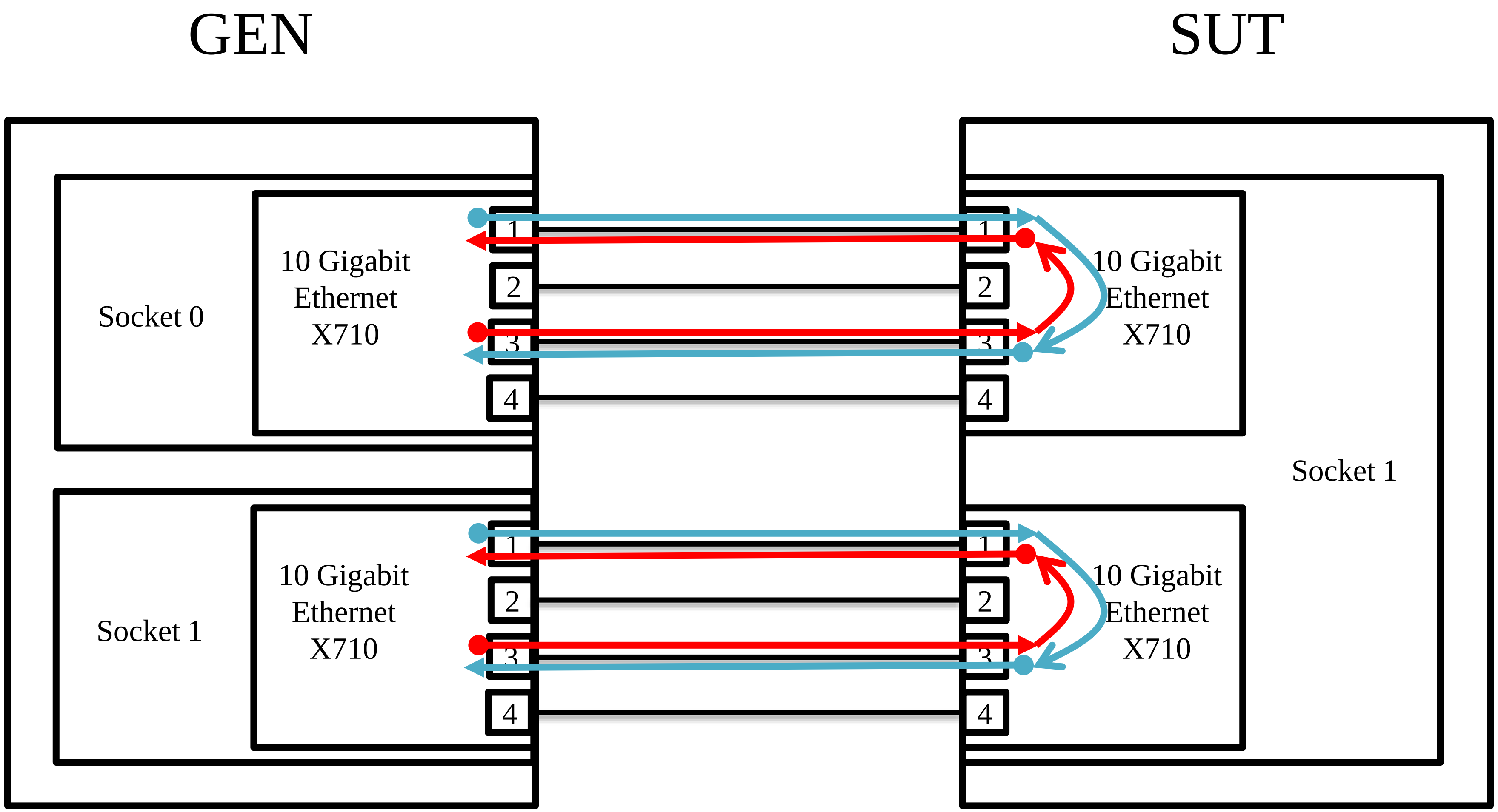
- Single Root I/O Virtualization (SR-IOV)。SR-IOV為PCI-SIG標準，允許PCIe的I/O裝置以多個實體與虛擬裝置呈現。



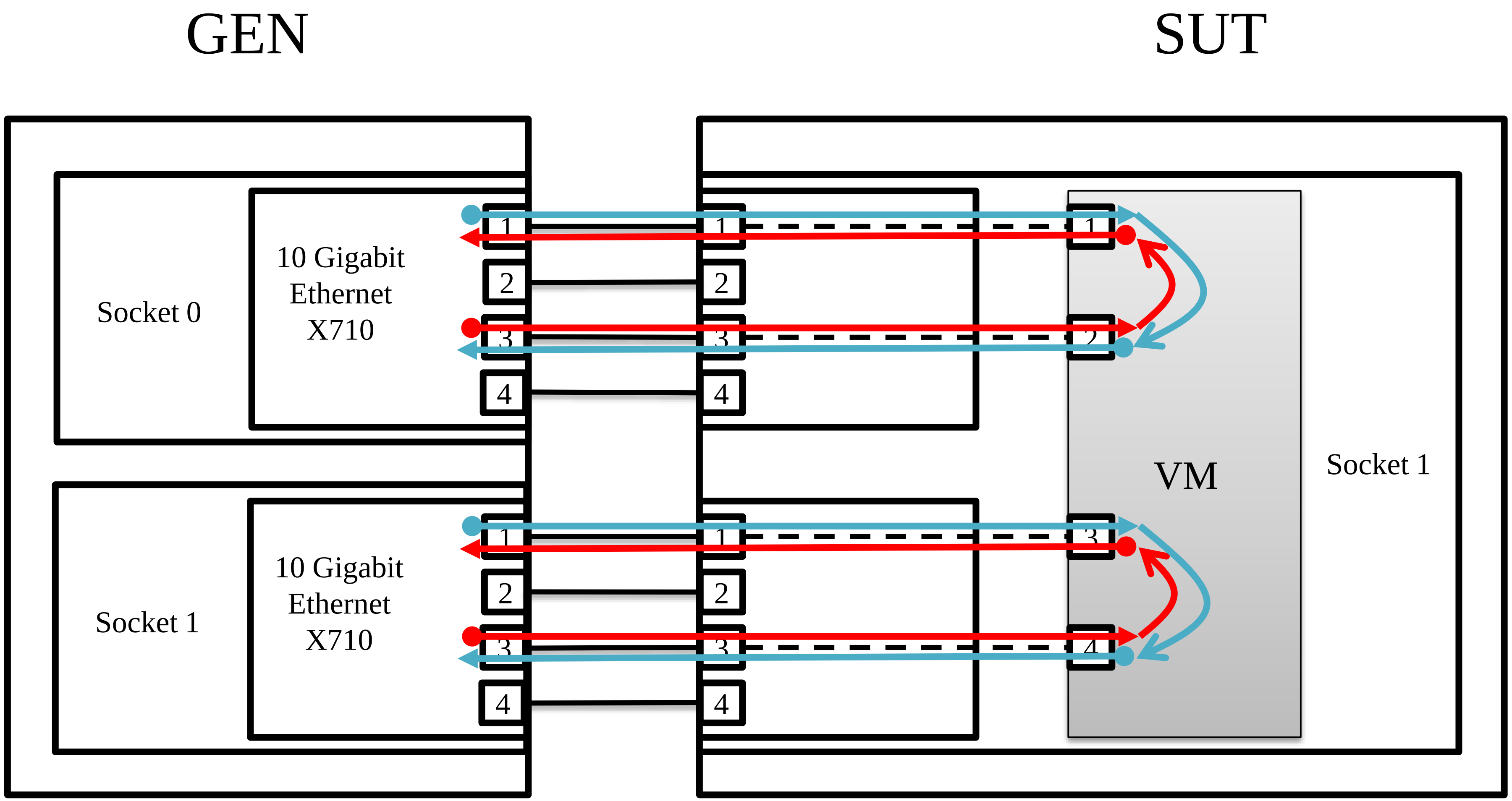


Source: https://access.redhat.com/documentation/en-us/red_hat_enterprise_linux/6/html/virtualization_host_configuration_and_guest_installation_guide/chap-virtualization_host_configuration_and_guest_installation_guide-sr_iov

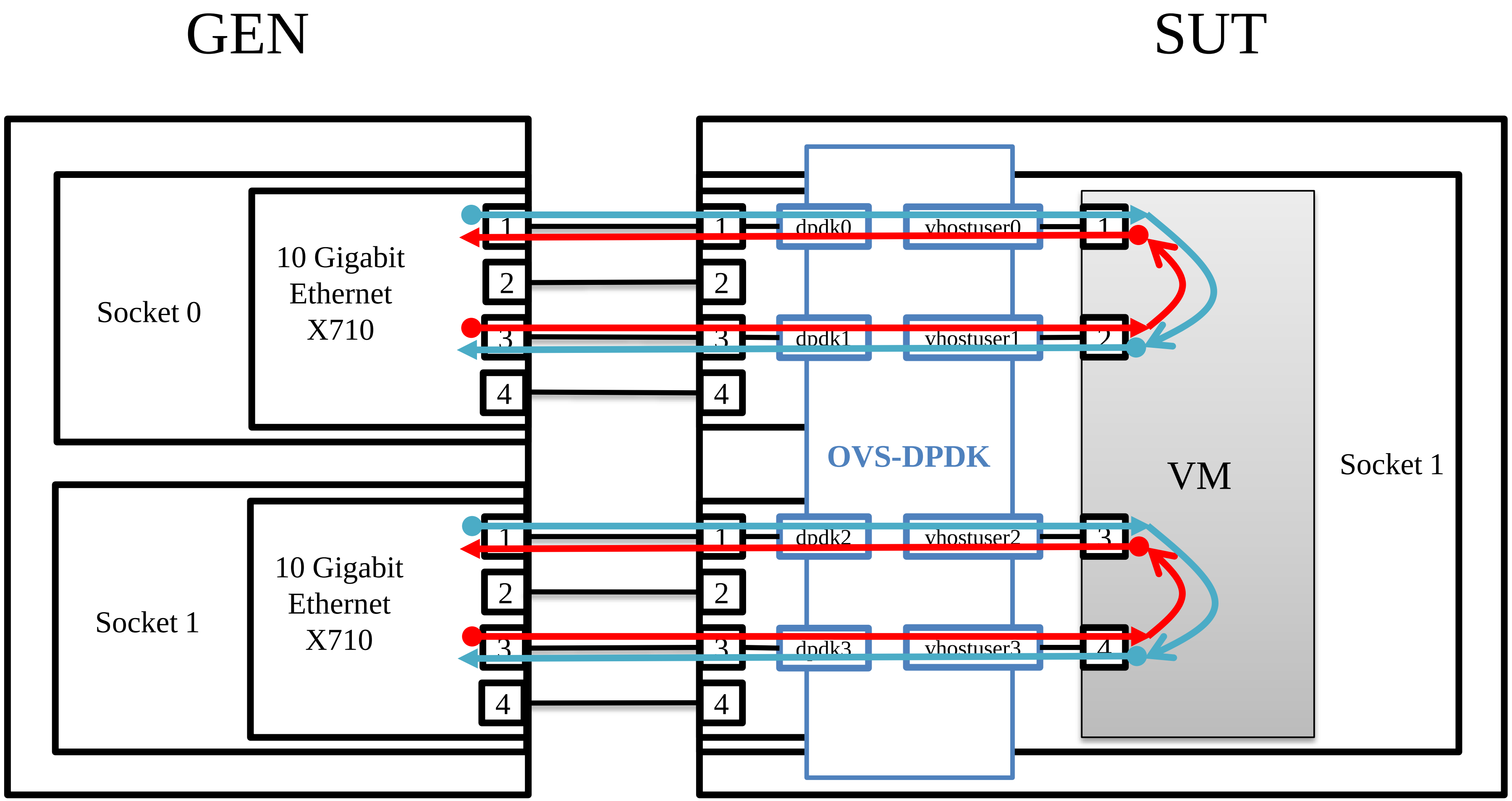
Baremetal Architecture



SRIOV Passthrough Architecture



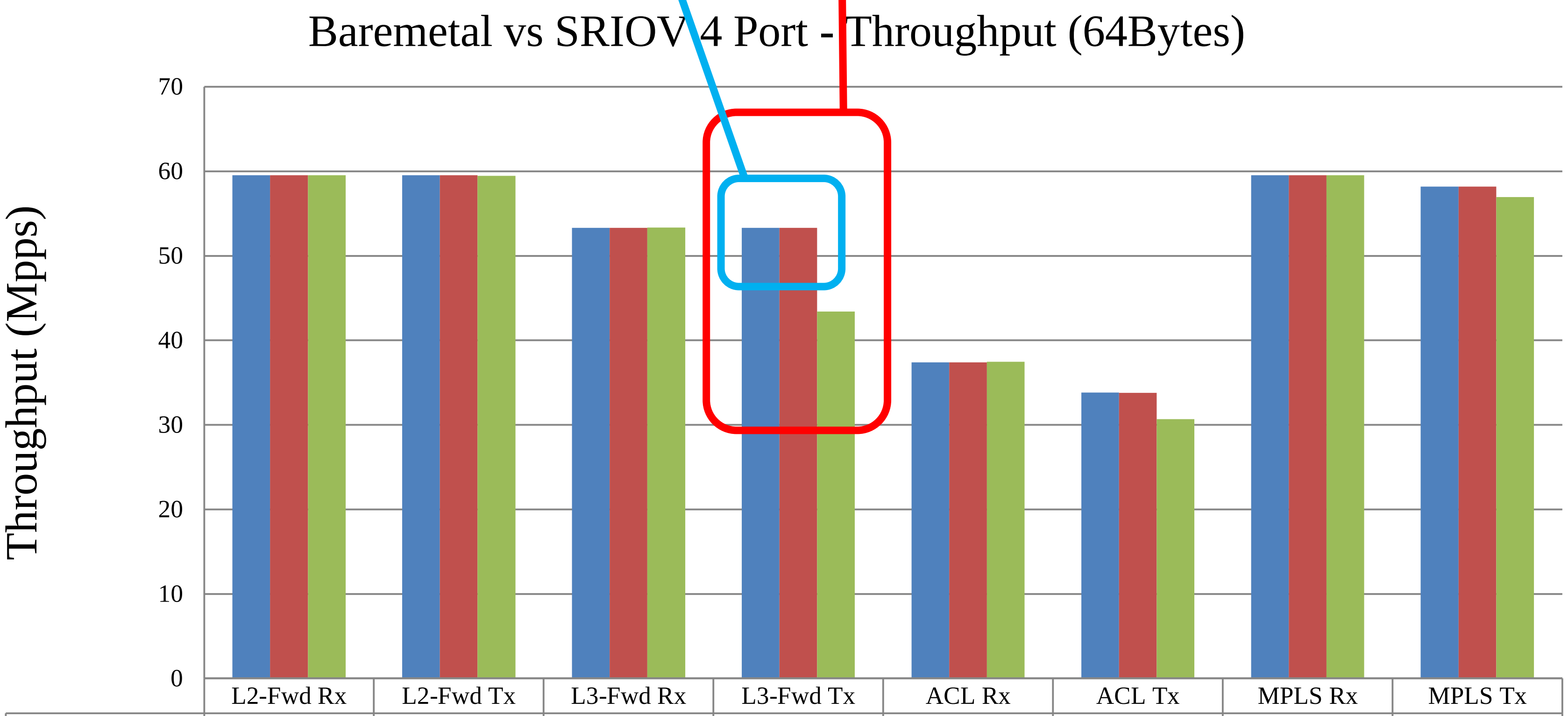
OVS-DPDK Architecture



Baremetal vs SRIOV Performance Result

VM performance with NUMA-aware configuration is similar to Baremetal performance.

VM without NUMA-aware configuration may decrease 10 – 20% performance.

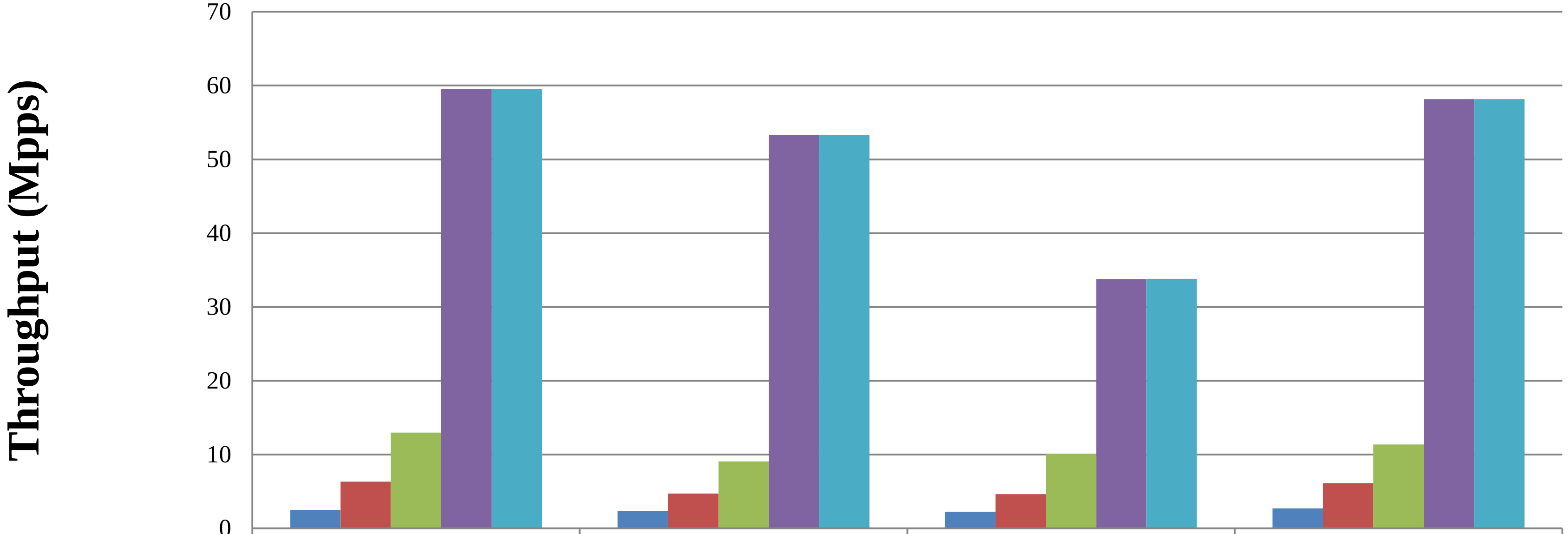


	L2-Fwd Rx	L2-Fwd Tx	L3-Fwd Rx	L3-Fwd Tx	ACL Rx	ACL Tx	MPLS Rx	MPLS Tx
BM	59.52	59.52	53.29	53.29	37.38	33.82	59.52	58.17
VM w/ NUMA	59.52	59.52	53.29	53.29	37.41	33.78	59.52	58.17
VM w/o NUMA	59.52	59.46	53.36	43.4	37.46	30.66	59.52	56.96

Baremetal vs SRIOV vs OVS-DPDK Performance Result

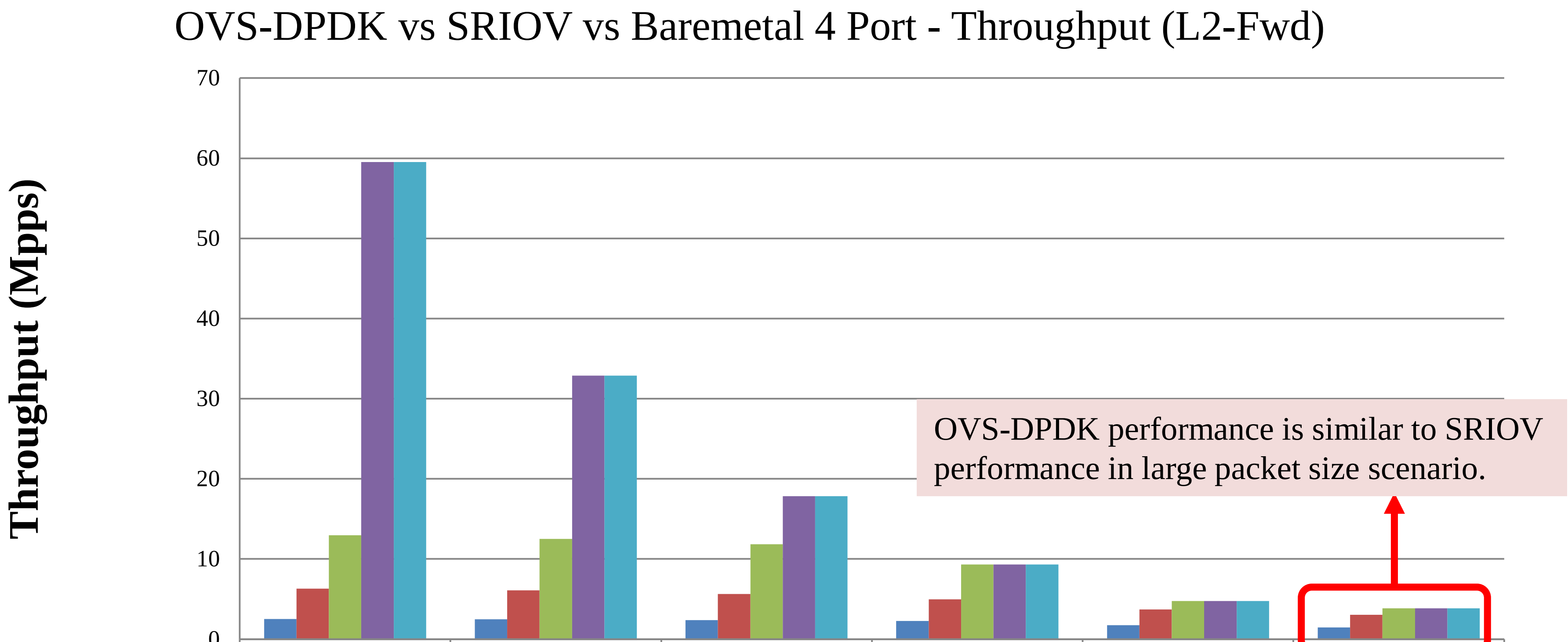
SRIOV could get better performance than OVS-DPDK in small packet size scenario.

OVS-DPDK vs SRIOV vs Baremetal 4 Port - Throughput (64Bytes)



	L2-Fwd Tx	L3-Fwd Tx	ACL Tx	MPLS Tx
OVS-DPDK-1-Core	2.51	2.33	2.24	2.71
OVS-DPDK-2-Cores	6.31	4.71	4.64	6.11
OVS-DPDK-4-Cores	12.96	9.06	10.1	11.36
SRIOV	59.52	53.29	33.78	58.17
BM	59.52	53.29	33.82	58.17

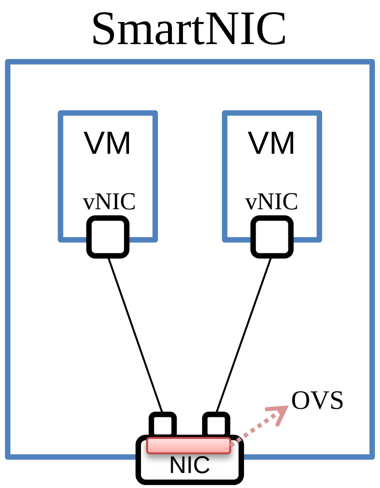
Baremetal vs SRIOV vs OVS-DPDK Performance Result

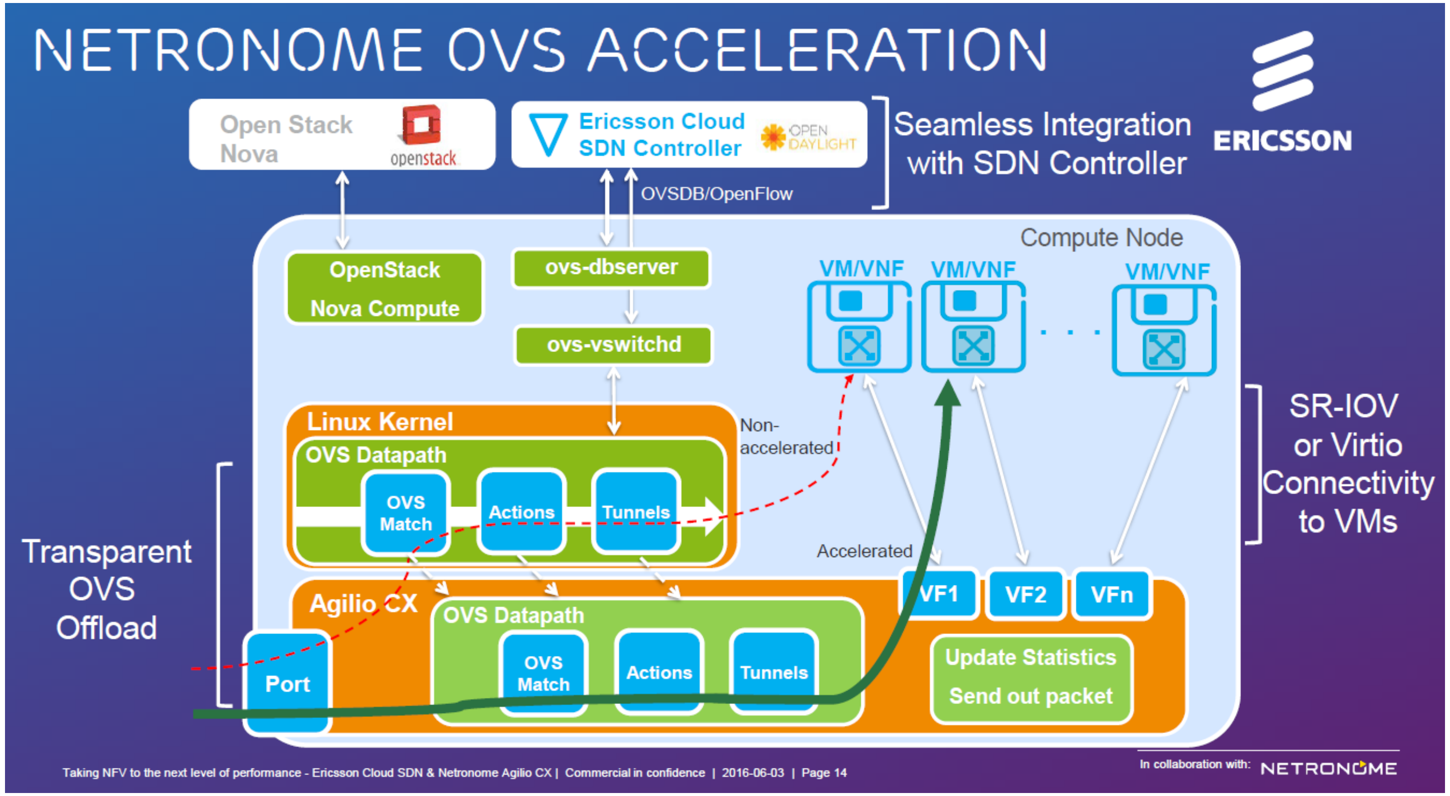


	64 Bytes	128 Bytes	256 Bytes	512 Bytes	1024 Bytes	1280 Bytes
OVS-DPDK-1-Core	2.51	2.46	2.38	2.26	1.74	1.47
OVS-DPDK-2-Cores	6.31	6.1	5.64	4.98	3.71	3.04
OVS-DPDK-4-Cores	12.96	12.5	11.82	9.32	4.77	3.83
SRIOV	59.52	32.89	17.85	9.32	4.77	3.83
BM	59.52	32.89	17.85	9.32	4.77	3.83

SmartNIC

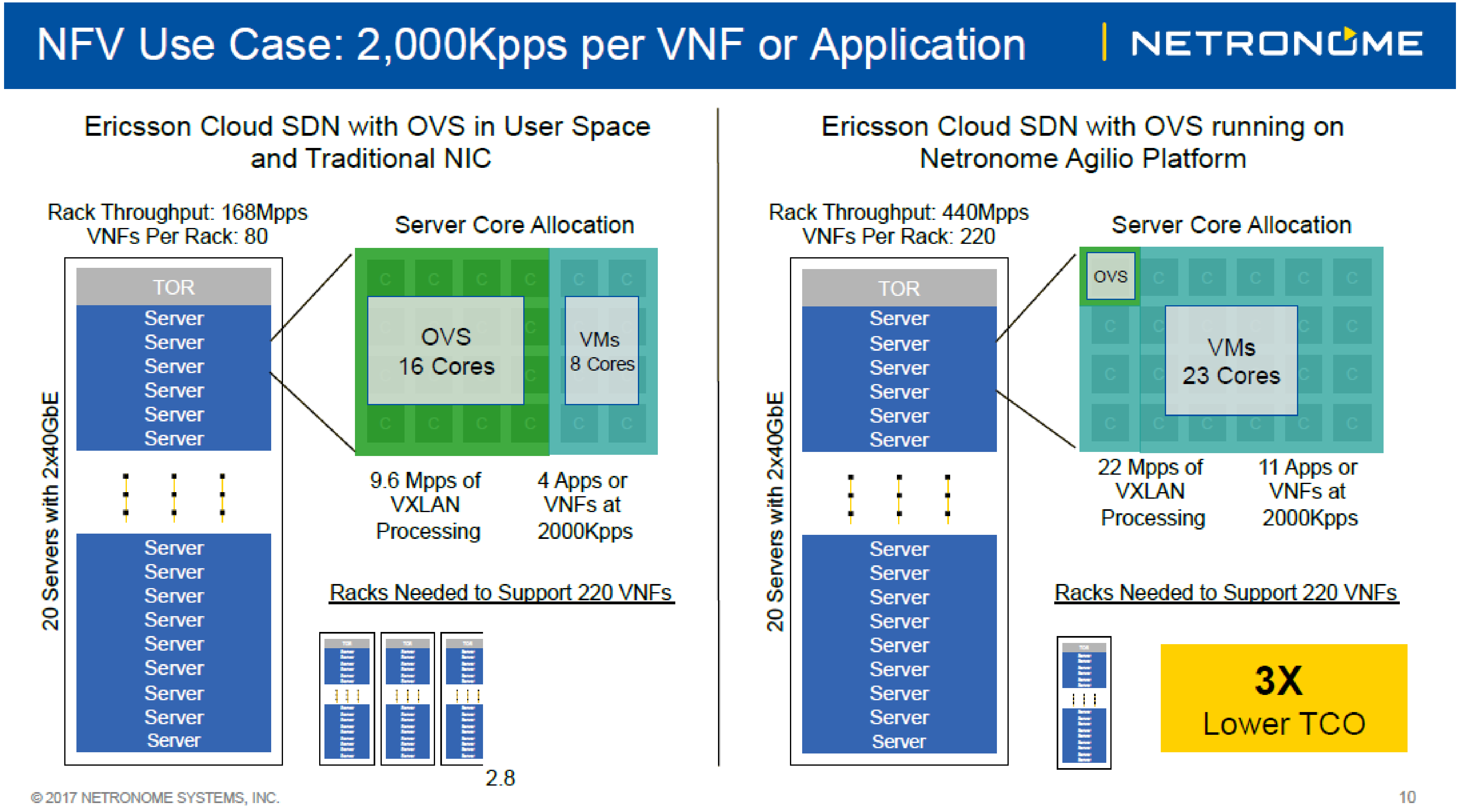
- **Netronome SmartNIC將OVS的功能做到硬體網卡中，可增加網路傳輸效能，並減少CPU資源的損耗。**





Source: https://www.slideshare.net/Netronome/ericsson-cloud-sdn-netronome-agilio-cx-taking-nfv-to-the-next-level-of-performance

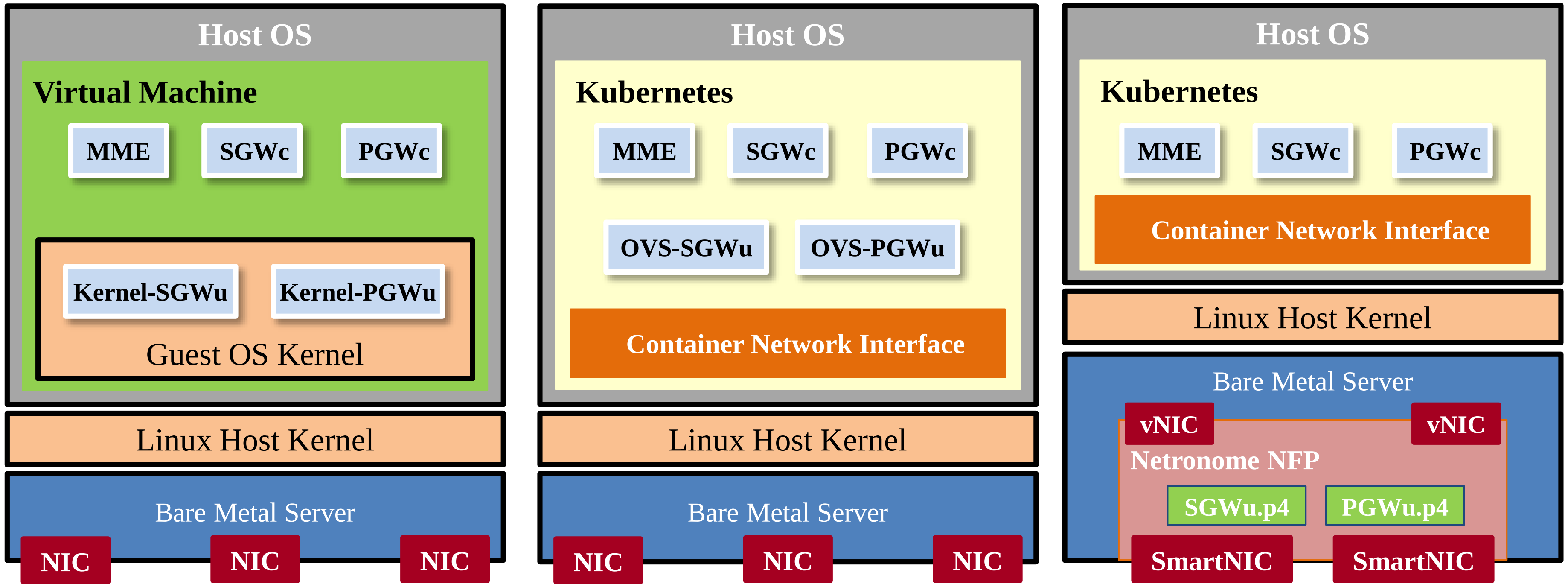
Ericsson Cloud SDN & Netronome Agilio CX



Source: https://www.slideshare.net/Netronome/ericsson-cloud-sdn-netronome-agilio-cx-taking-nfv-to-the-next-level-of-performance

III vEPC Data Plane Enhancement

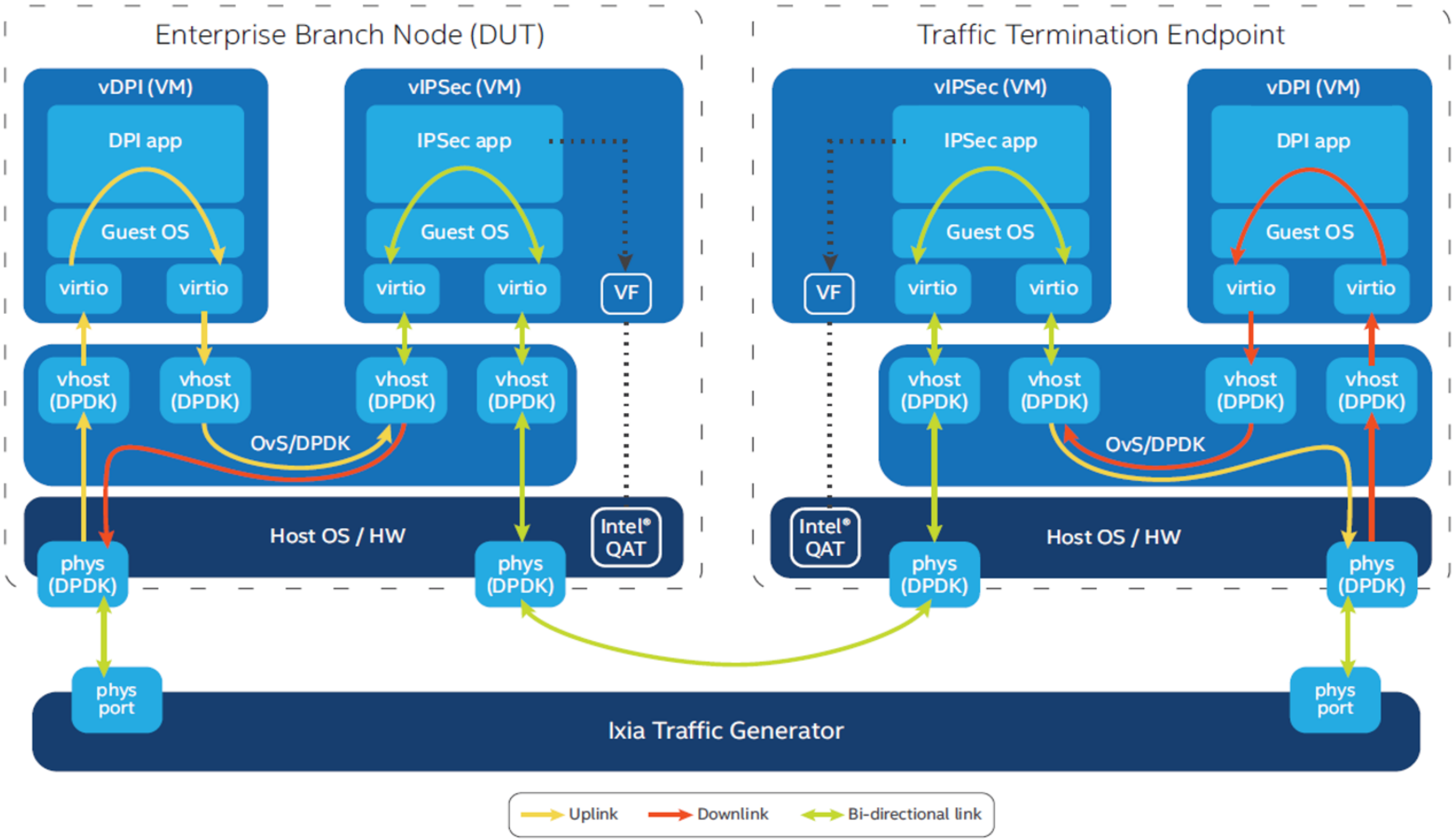
- 工研院協助資策會團隊優化vEPC的網路效能，並整合到NFVI平台。
- 三階段Data Plane效能優化:
  - Kernel-based GTPU data plane, throughput is about 800Mbps.
  - DPDK-based GTPU data plane, throughput is about 4Gbps.
  - SmartNIC-based GTPU data plane, throughput is about 9Gbps.
- 現今III 5GC的網路效能可達 25Gbps – 40Gbps



Intel Atom Processor C3758
SDWAN Performance Report

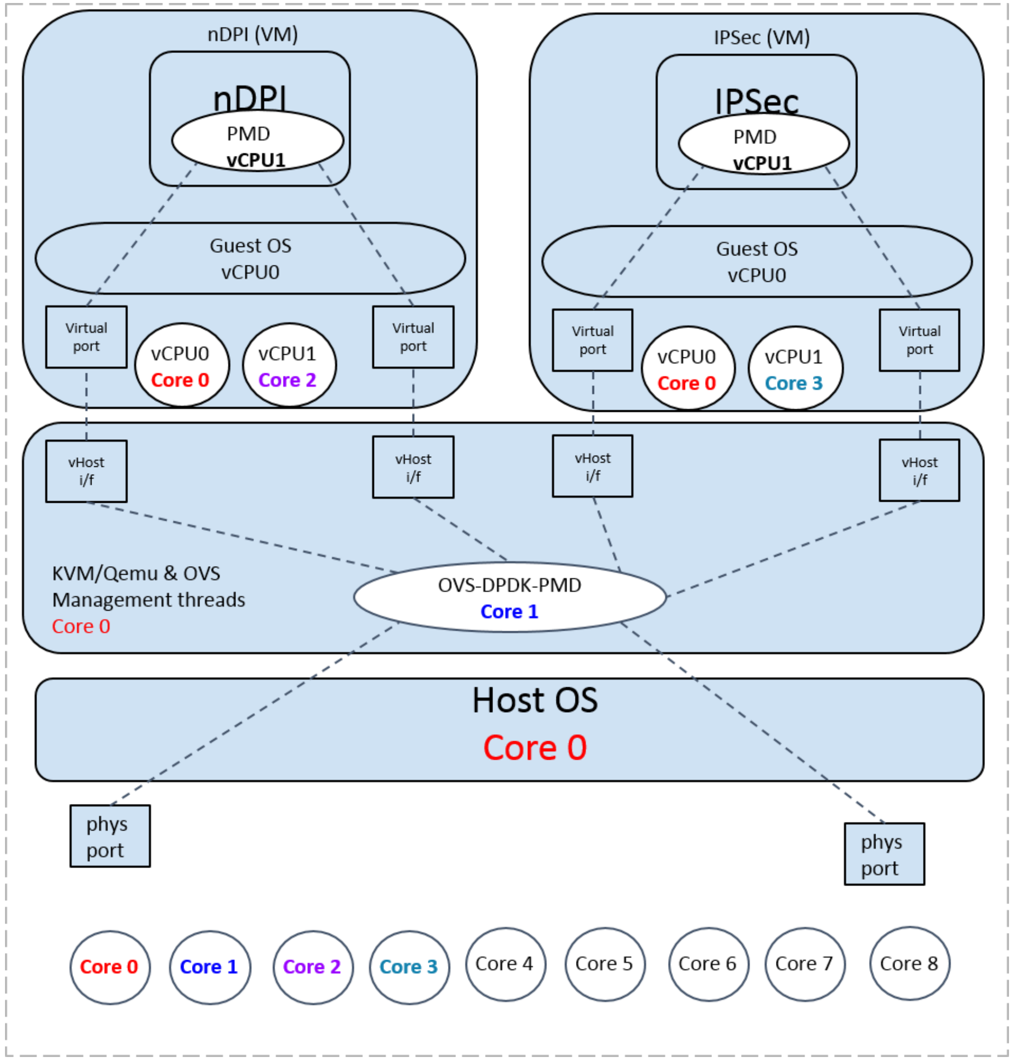
Produced by ITRI Performance Lab

SDWAN Topology



CPU Core Assignment – 4 Cores, 1 PMD Thread

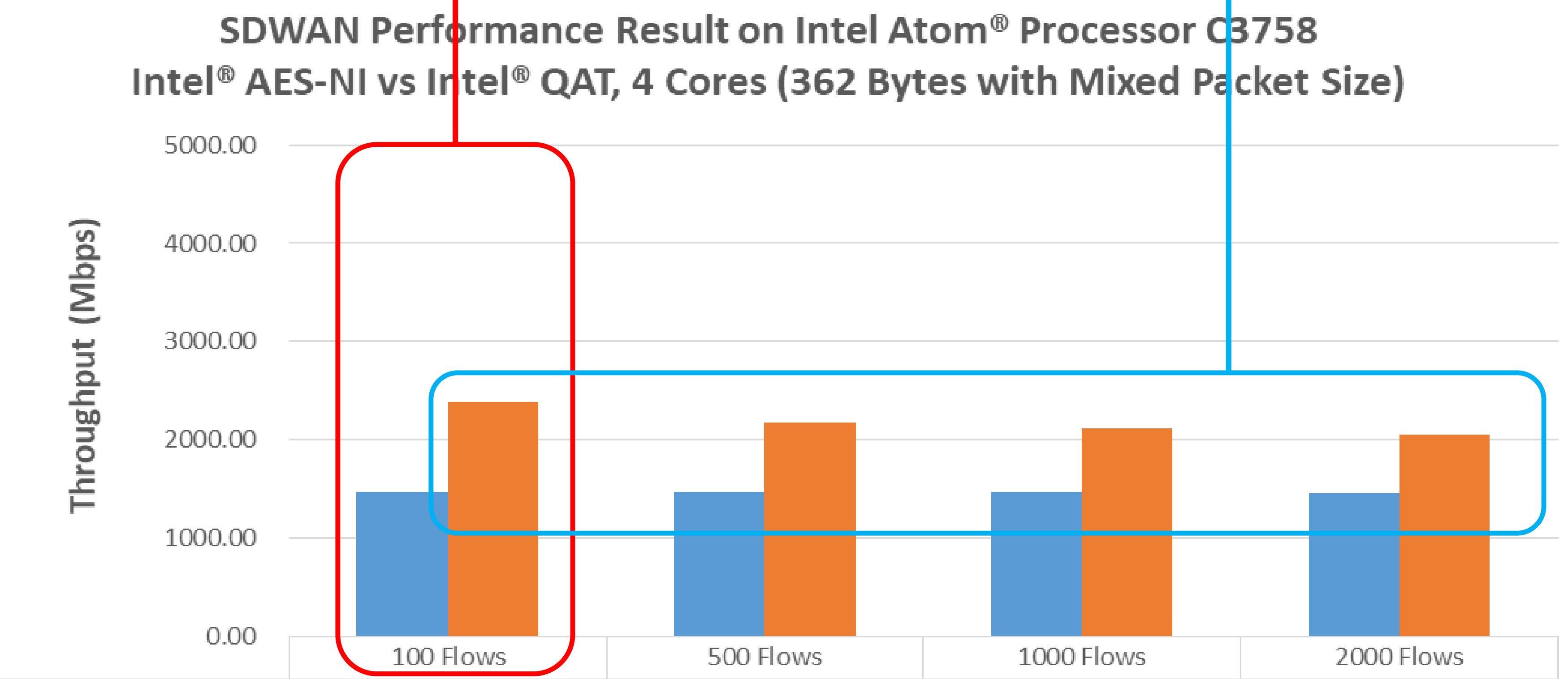
4-Core Configuration	
Host OS	Core 0
Hypervisor (KVM/Qemu)	Core 0
OVS Mgmt Threads	Core 0
OVS DPDK PMD	Core 1
DPI VNF (vCPU threads)	Core 0 (VM vCPU 0), Core 2 (VM vCPU 1)
IPsec VNF (vCPU threads)	Core 0 (VM vCPU 0), Core 3 (VM vCPU 1)



SDWAN Performance on 4 Cores Environment

Use QAT accelerator can improve IPSec performance

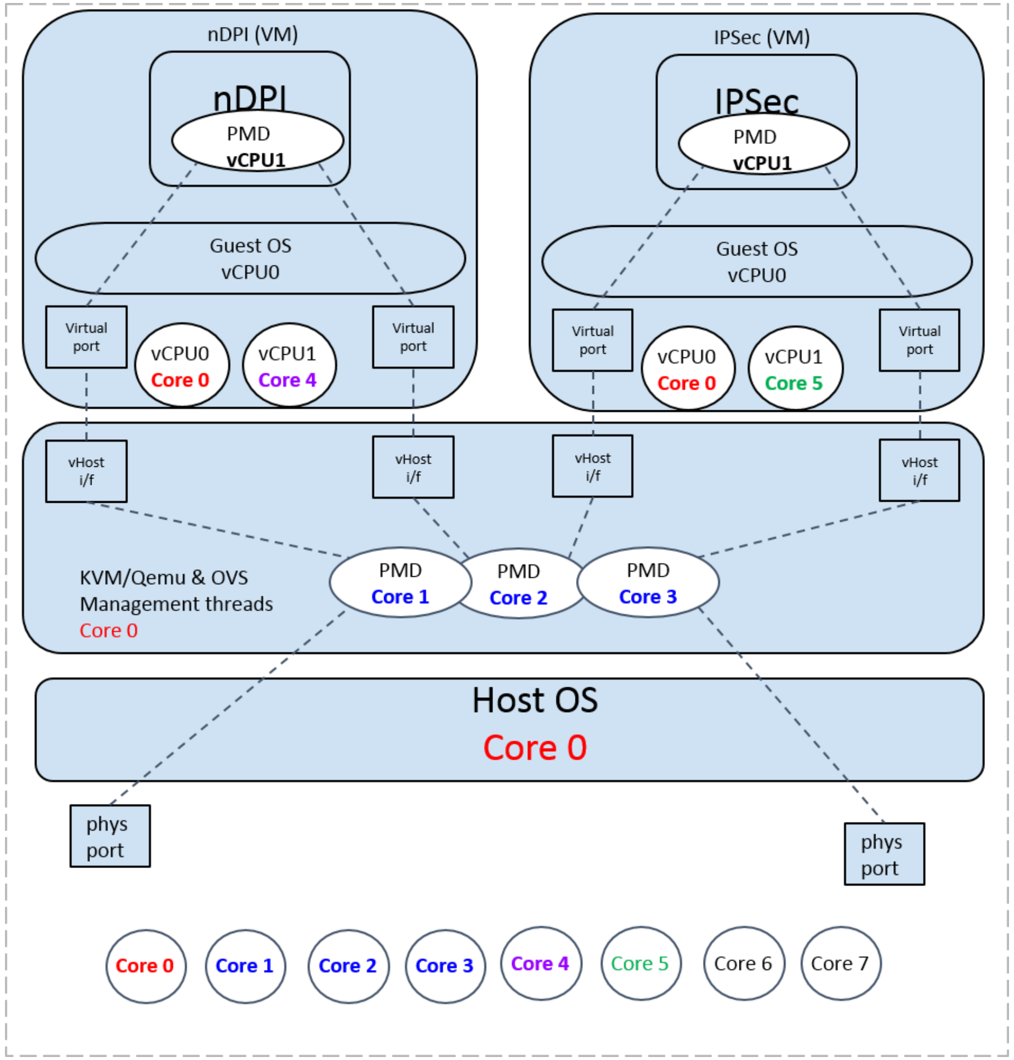
QAT performance will be decrease when flow number is increase, OVS is the performance bottleneck.



	100 Flows	500 Flows	1000 Flows	2000 Flows
Intel® AES-NI, 4 Cores	1469.20	1469.20	1466.45	1454.09
Intel® QAT, 4 Cores	2378.46	2176.69	2107.14	2049.09

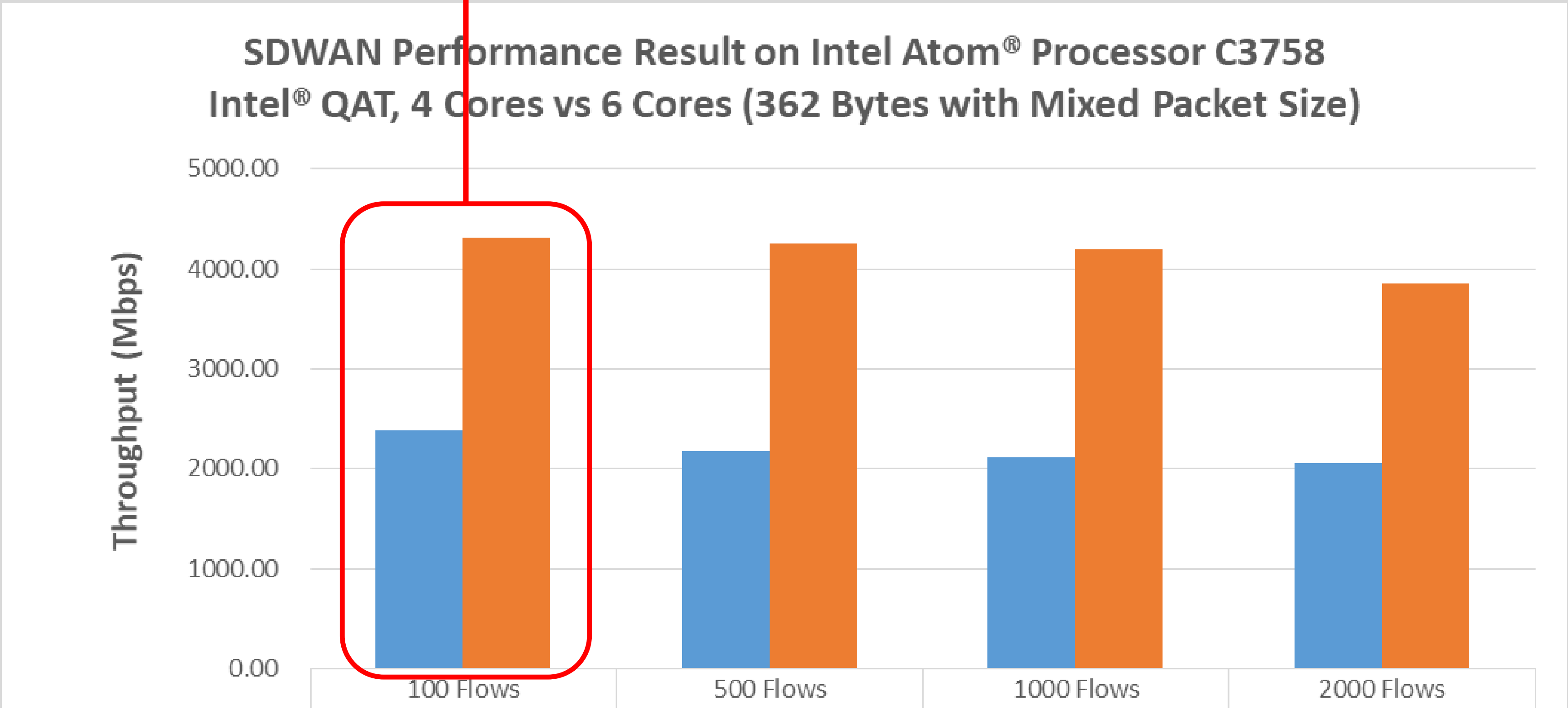
CPU Core Assignment – 6 Cores, 3 PMD Thread

6-Core Configuration	
Host OS	Core 0
Hypervisor (KVM/Qemu)	Core 0
OVS Mgmt Threads	Core 0
OVS DPDK PMD	Core 1 Core 2 Core 3
DPI VNF (vCPU threads)	Core 0 (VM vCPU 0), Core 4 (VM vCPU 1)
IPsec VNF (vCPU threads)	Core 0 (VM vCPU 0), Core 5 (VM vCPU 1)



SDWAN Performance of QAT Scenario

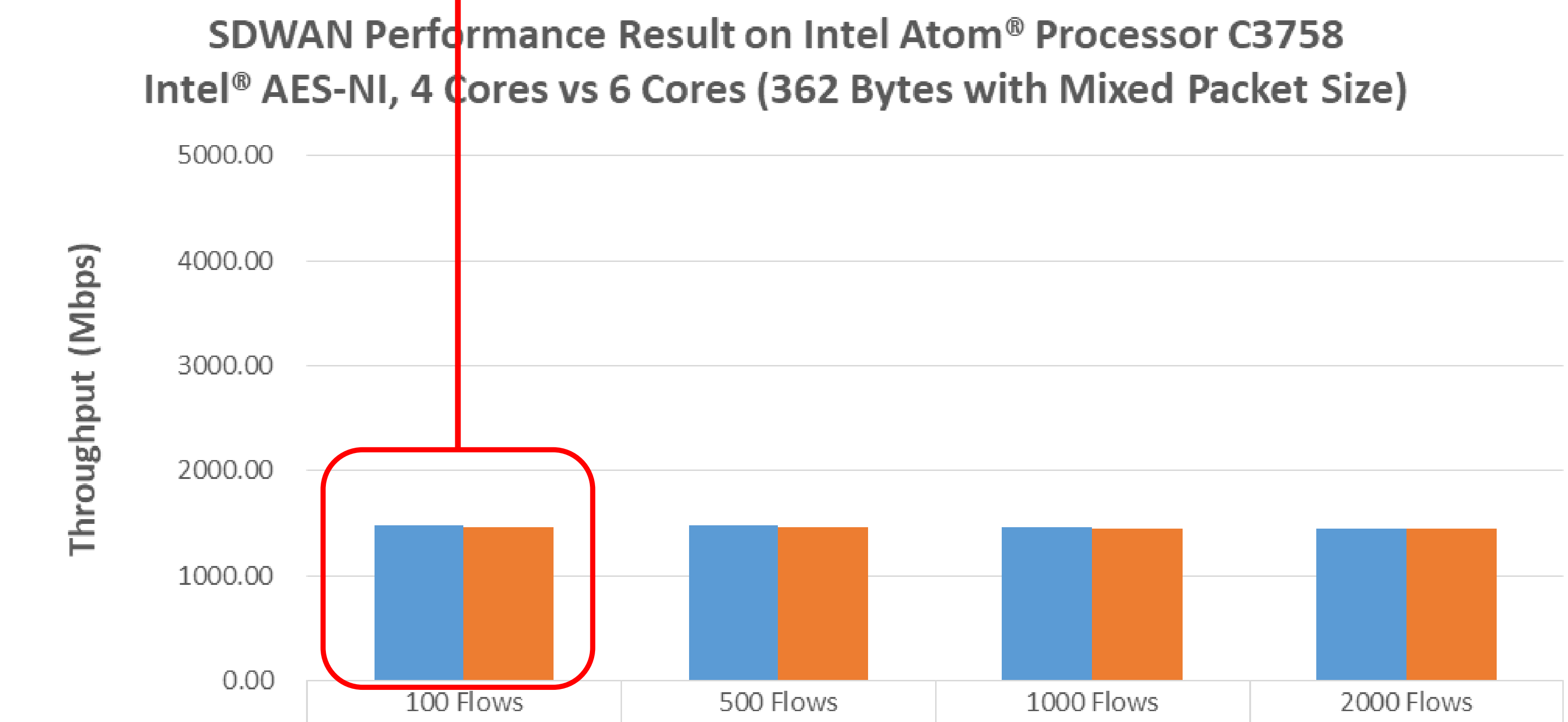
QAT performance will be increase when we allocate more CPU core to OVS.



	100 Flows	500 Flows	1000 Flows	2000 Flows
Intel® QAT, 4 Cores	2378.46	2176.69	2107.14	2049.09
Intel® QAT, 6 Cores	4310.94	4252.93	4194.91	3846.96

SDWAN Performance on 6 Core Environment

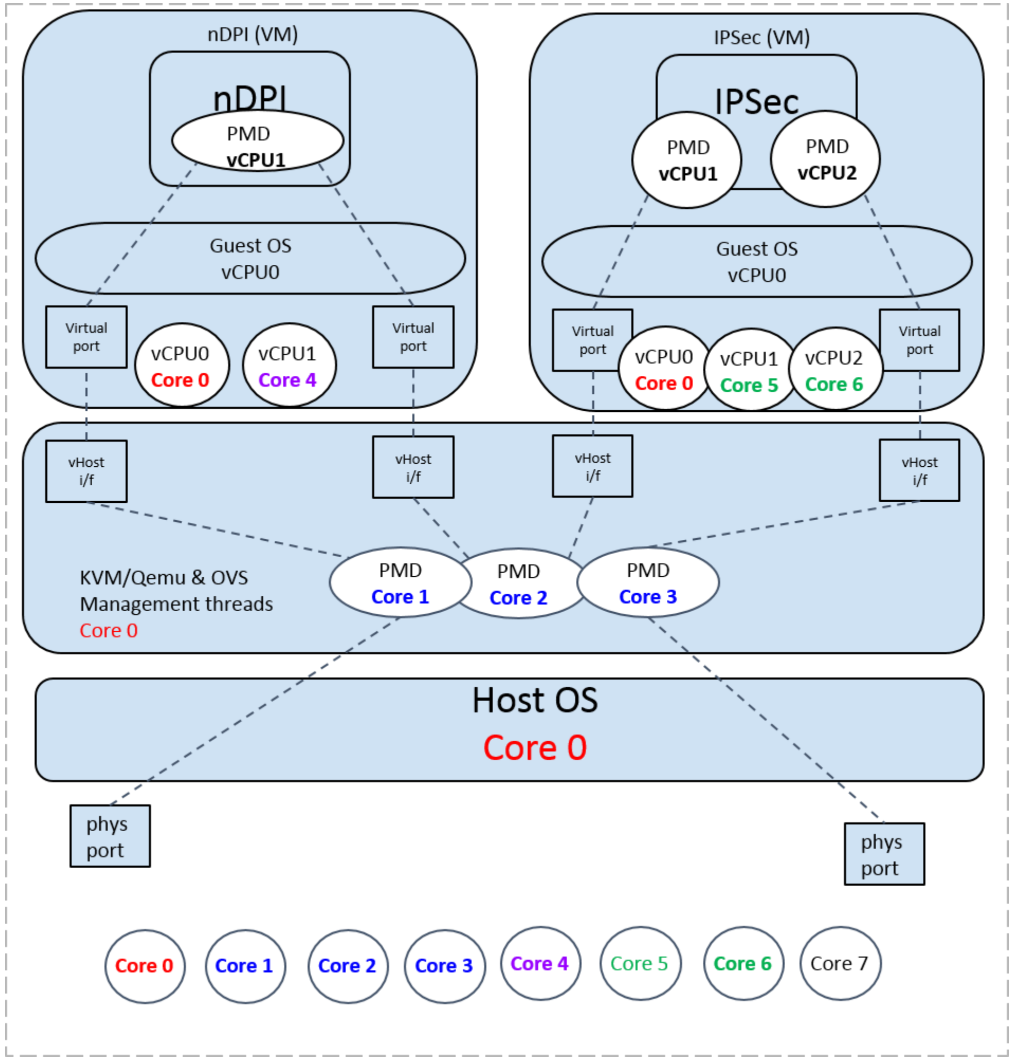
The performance of IPSec without QAT is almost the same with 4 core scenario, bottleneck is IPSec itself.



	100 Flows	500 Flows	1000 Flows	2000 Flows
Intel® AES-NI, 4 Cores	1469.20	1469.20	1466.45	1454.09
Intel® AES-NI, 6 Cores	1460.38	1459.55	1447.28	1453.39

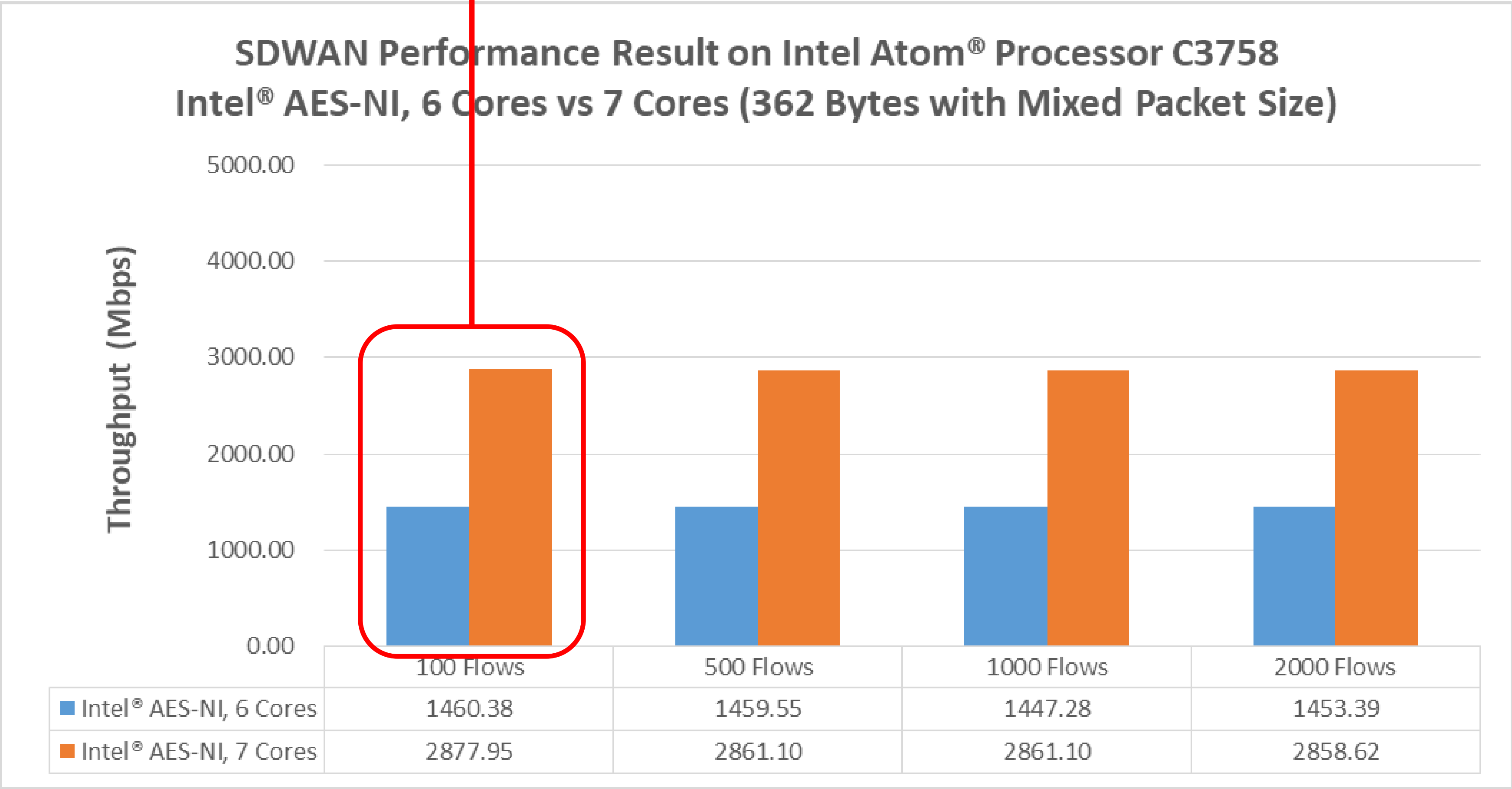
CPU Core Assignment – 7 Cores, 3 PMD Thread

7-Core Configuration	
Host OS	Core 0
Hypervisor (KVM/Qemu)	Core 0
OVS Mgmt Threads	Core 0
OVS DPDK PMD	Core 1 Core 2 Core 3
DPI VNF (vCPU threads)	Core 0 (VM vCPU 0), Core 4 (VM vCPU 1)
IPsec VNF (vCPU threads)	Core 0 (VM vCPU 0), Core 5 (VM vCPU 1), Core 6 (VM vCPU 2)



SDWAN Performance of AES-NI Scenario

The performance of IPSec AES-NI will be increase when we allocate more CPU core to IPSec VNF.



	100 Flows	500 Flows	1000 Flows	2000 Flows
Intel® AES-NI, 6 Cores	1460.38	1459.55	1447.28	1453.39
Intel® AES-NI, 7 Cores	2877.95	2861.10	2861.10	2858.62

Intel Xeon Processor E5-2695 v4
Performance Report: 2-8 VMs Service Chain with SR-IOV, OVS-DPDK, VPP and SPP

Produced by ITRI Performance Lab

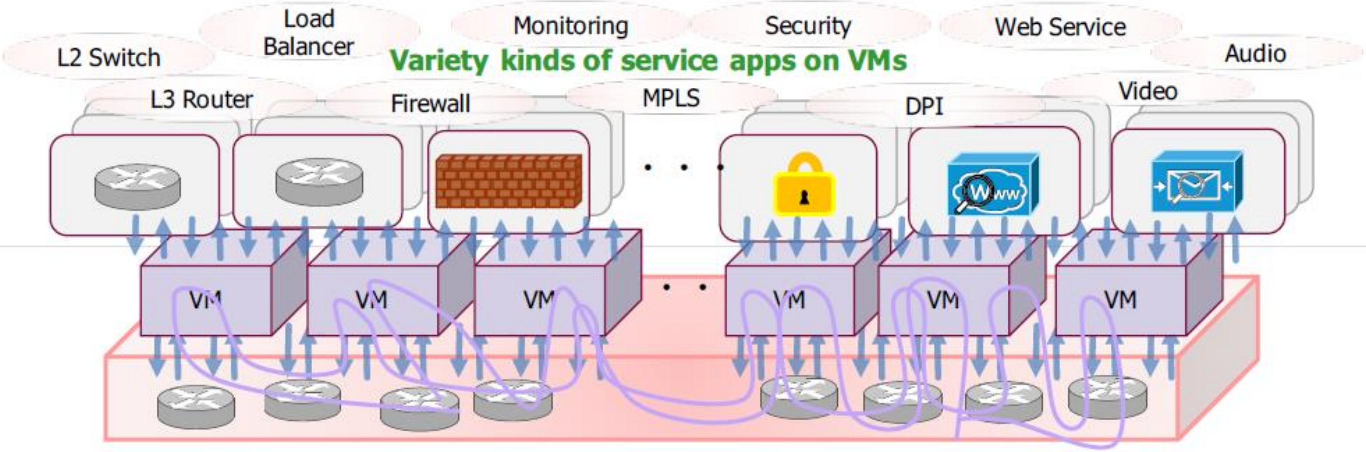
NTT’s Presentation on DPDK Summit

VM Chaining Scenario

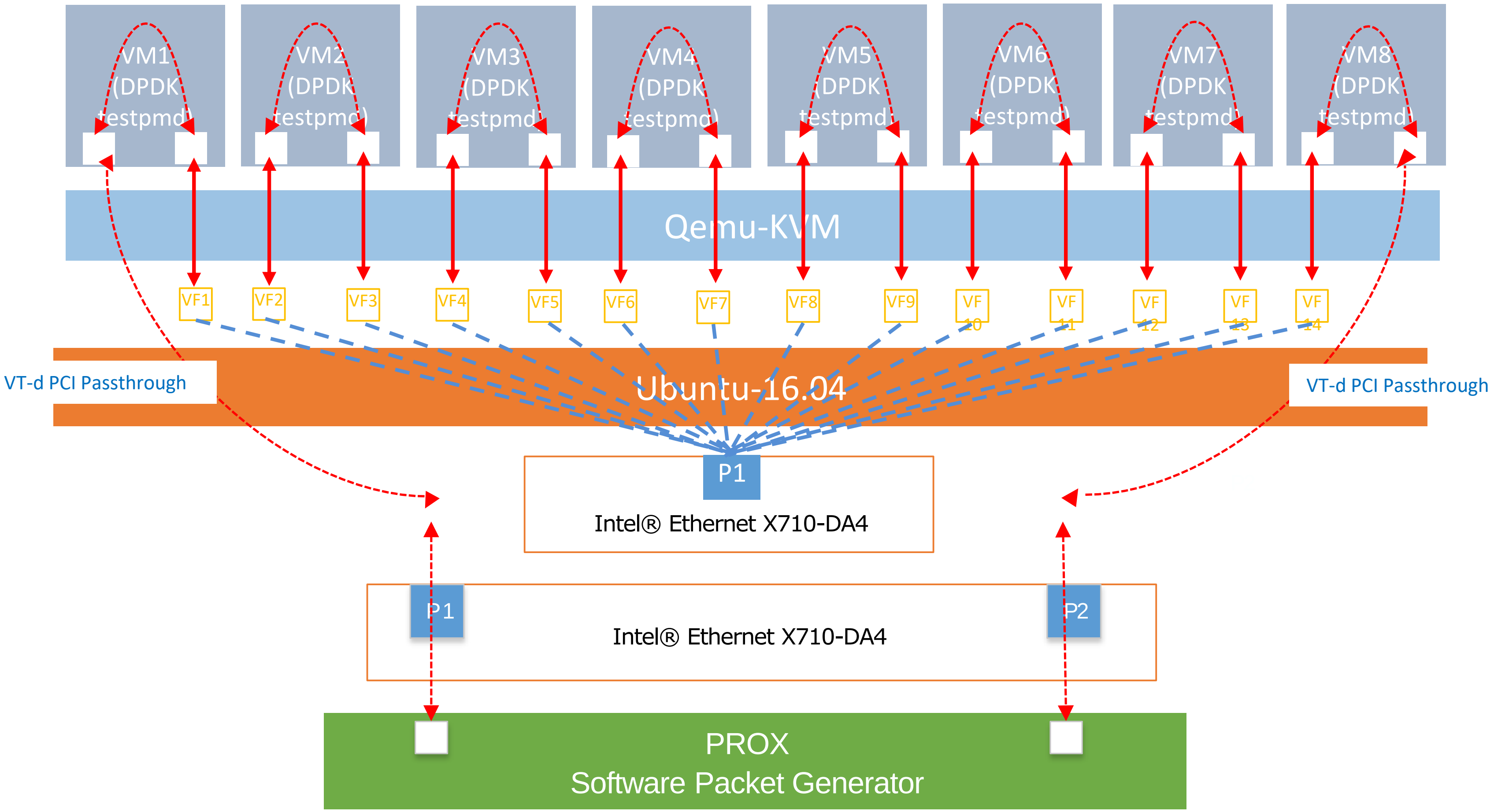
Motivation

DPDK

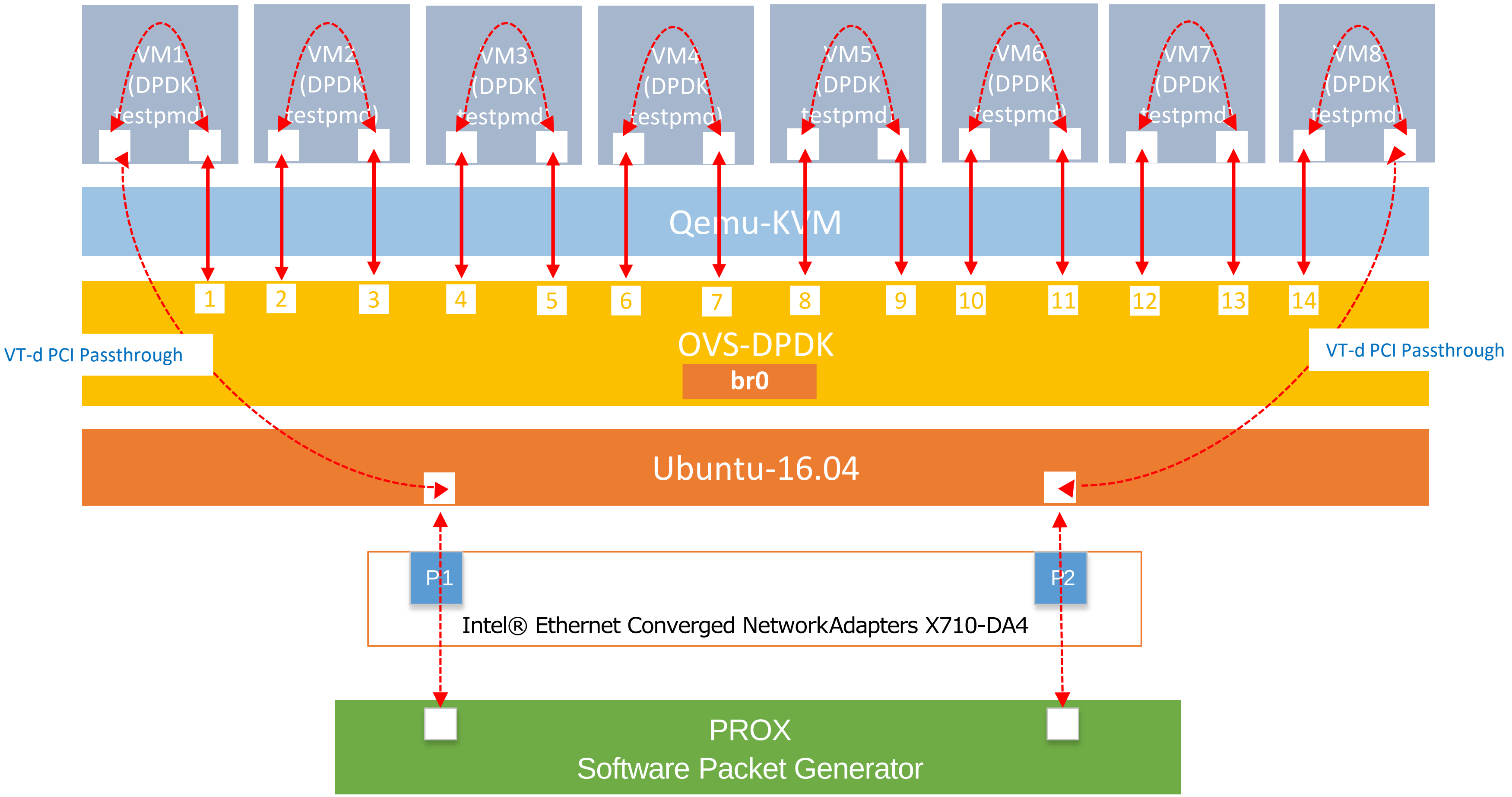
- Large-scale cloud for telecom services
- Service Function Chaining for virtual network appliances
- Flexibility, Maintainability and High-Performance



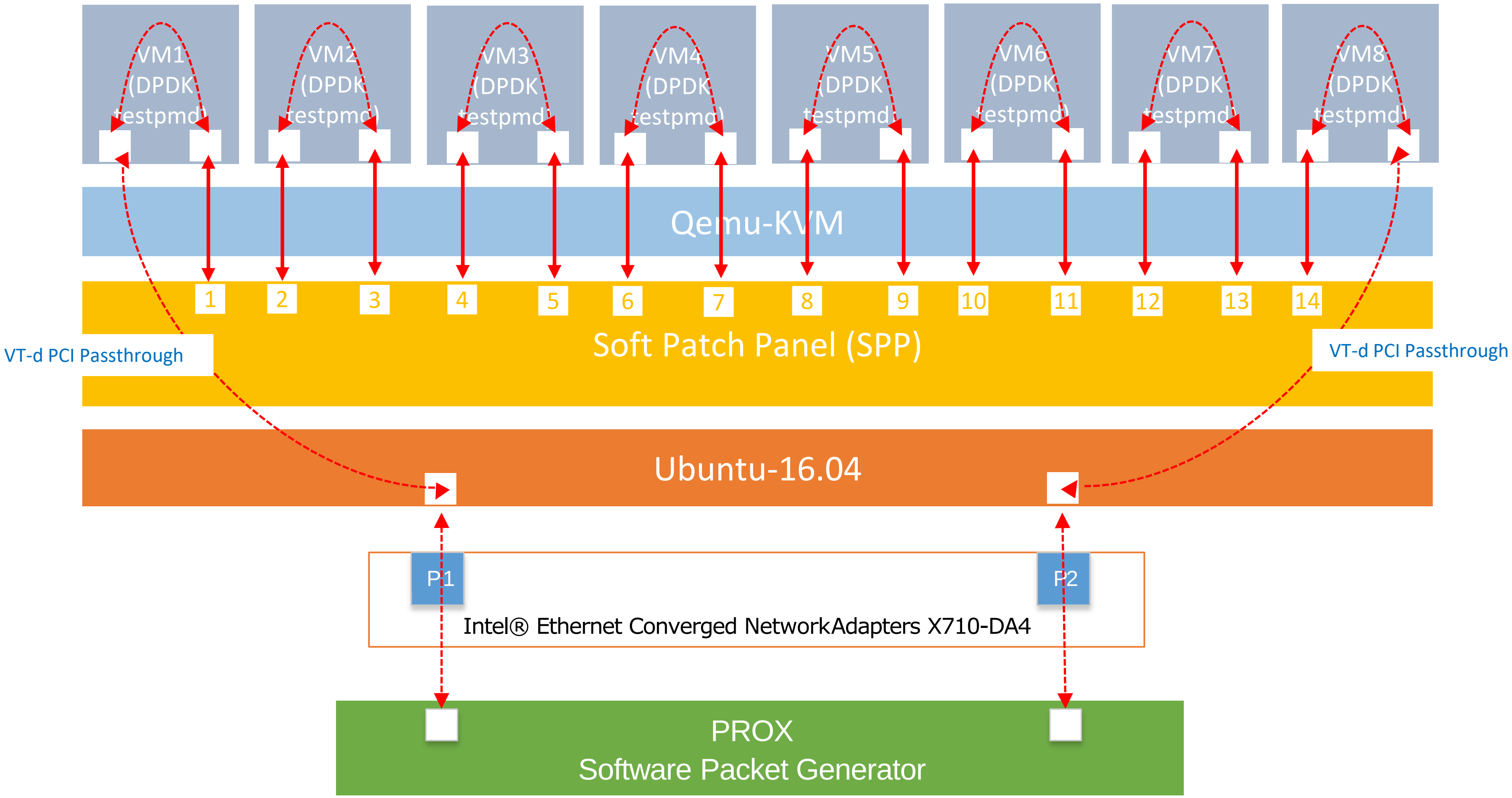
8 VMs Service Chain Test Setup Diagram (SRIOV)



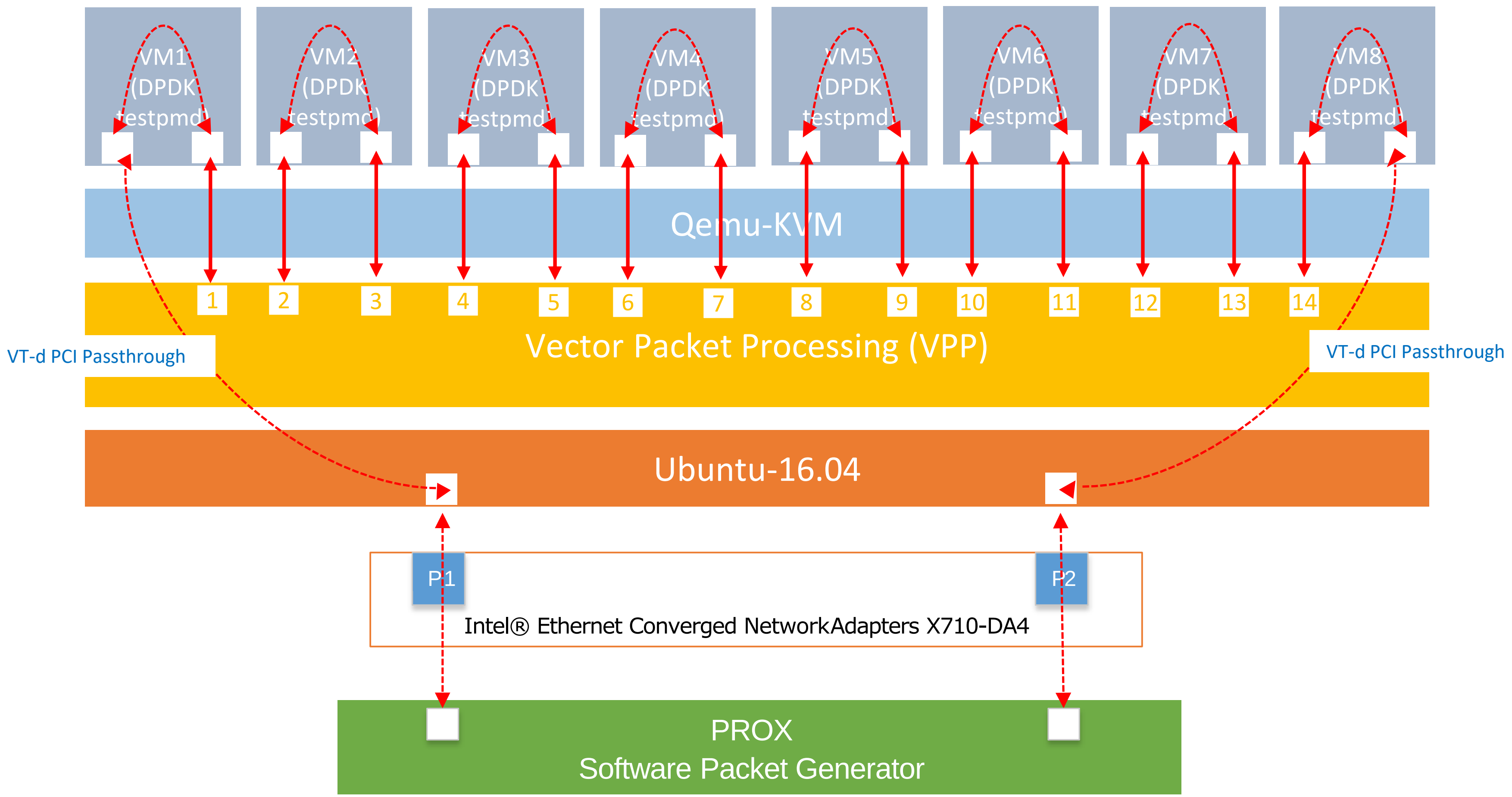
8 VMs Service Chain Test Setup Diagram (OVS-DPDK)



8 VMs Service Chain Test Setup Diagram (SPP)



8 VMs Service Chain Test Setup Diagram (VPP)



Performance Result : Summary

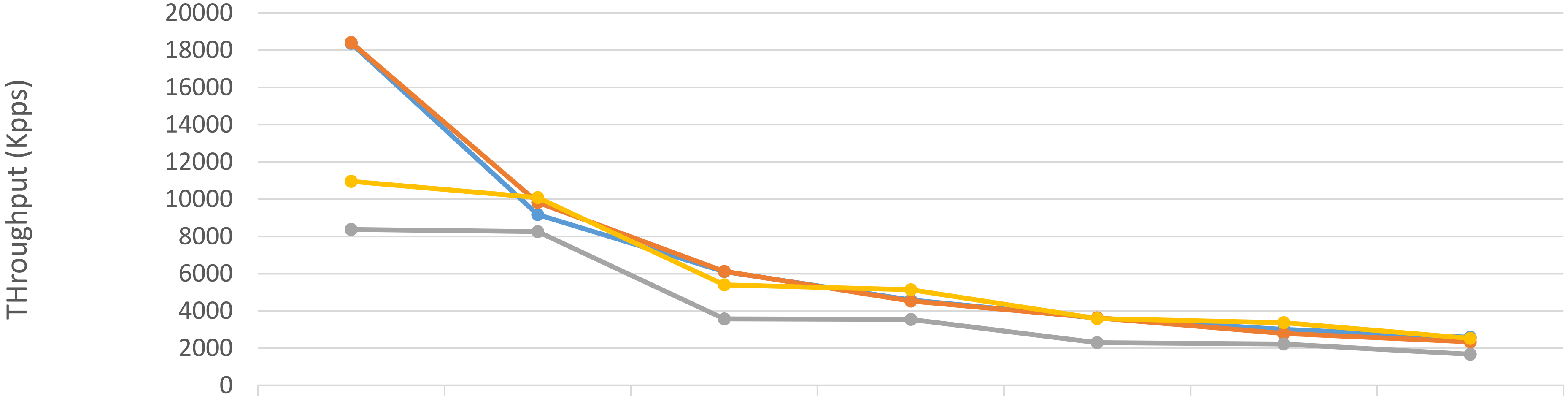
2 to 8 VMs Service Chain Performance
64 Byte, 10K Flow
(Bi-Direction, total 20G)

Binary-Search Pktgen

2 to 8 VMs Service Chain Performance: 64 Bytes, 10K Flow, Bi-Direction, Packet per second

SRIOV's performance is not good in VM chaining scenario, the bottleneck is NIC's limitation.

Phy-VM-Phy Core Scalability Performance running 10K Flows on
Intel® Xeon® Processor E5-2695 v4
Multiple VM Service Chain (64Byte, 10K Flow, Bi-direction)



	2 VMs	3 VMs	4 VMs	5 VMs	6 VMs	7 VMs	8 VMs
SRIOV	18370.39	9176	6106.92	4582.13	3610.25	3004.99	2580.51
SPP-vhost-4PMD	18405.19	9829.09	6117.86	4523.16	3624.33	2786.65	2343.46
VPP-vhost-4PMD	10959.9	10083.83	5393	5143.81	3581.41	3363.37	2522.53
OVS-DPDK-4PMD	8367.34	8263.44	3566.33	3537.34	2290.57	2228.31	1669.36

Number of Chaining VM

Performance Result : OVS-DPDK

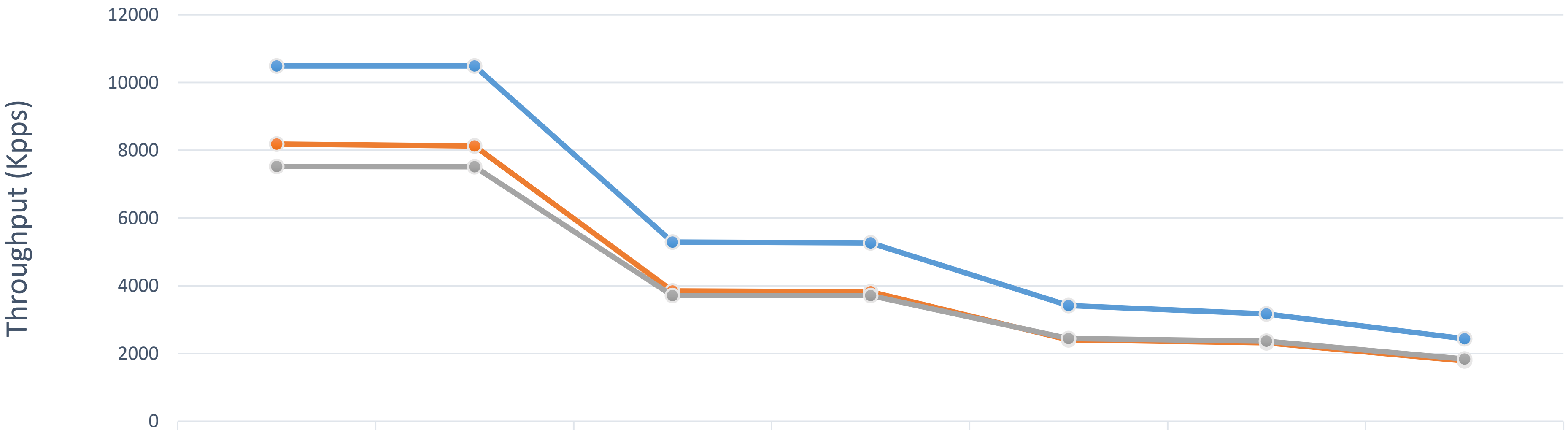
2 to 8 VMs Service Chain Performance
64 Byte, OVS-DPDK, 4 PMD
with Different number of Flow
(Single Direction 10G / Bi-Direction 20G)

Full-Speed Pktgen

2 to 8 VMs Service Chain Performance:
64 Bytes, OVS-DPDK, 4 PMD, Single Direction

OVS performance will be effected by flow number, it may cause 20% performance decrease.

Phy-VM-Phy Core Scalability Performance running Different Flows on Intel® Xeon® Processor E5-2695 v4
Multiple VM Service Chain (64 Bytes, OVS-DPDK, 4 PMD, Bi-direction)



	2 VMs	3 VMs	4 VMs	5 VMs	6 VMs	7 VMs	8 VMs
1 Flow	10491	10491	5292	5264	3416	3168	2438
10K Flows	8184	8127	3844	3824	2404	2316	1782
1M flows	7523	7513	3716	3716	2448	2368	1838

Packet Size (Bytes)

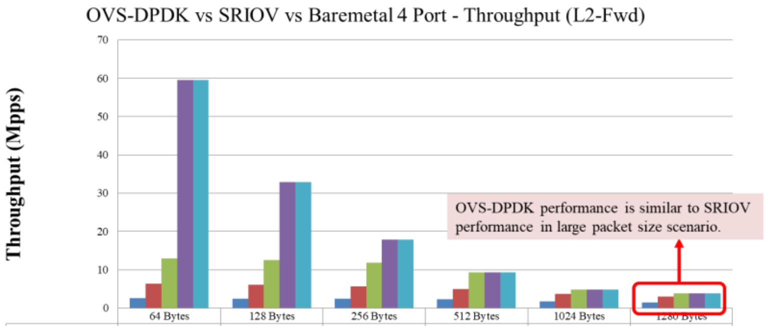
NFV Performance Lab

Features

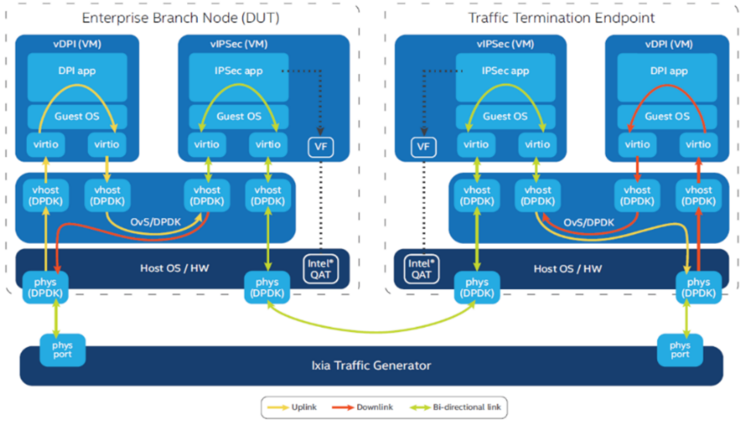
- VM/Container Performance Tuning
  - ➢ CPU Pining
  - ➢ NUMA configuration
  - ➢ BIOS configuration
- Data Plane Acceleration
  - ➢ SRIOV, SmartNIC
  - ➢ Enable DPDK (OVS-DPDK, VPP, SPP)
  - ➢ QAT, Intel AES-NI
- Scenario / Use Cases
  - ➢ NFVI / VNF performance characterization
  - ➢ SDWAN scenario (uCPE, DPI, IPSEC)
  - ➢ VM Chaining with different data plane

Contribution

- Help ODM/OEM vendors build their own NFV performance lab, enhance their NFV performance optimization and testing skills.
- Help hardware and software vendors optimize their NFV product performance and publish performance white papers.
- Help operators evaluate and analysis performance bottleneck on specific NFV scenarios.



	64 Bytes	128 Bytes	256 Bytes	512 Bytes	1024 Bytes	1280 Bytes
OVS-DPDK-1-Core	2.51	2.46	2.38	2.26	1.74	1.47
OVS-DPDK-2-Cores	6.31	6.1	5.64	4.98	3.71	3.04
OVS-DPDK-4-Cores	12.96	12.5	11.82	9.32	4.77	3.83
SRIOV	59.52	32.89	17.85	9.32	4.77	3.83
BM	59.52	32.89	17.85	9.32	4.77	3.83



總結

- 影響NFV網路效能的因素眾多，也涵蓋多個領域，包含加速卡、系統架構、資源配置、軟體效能等，在本課程中，主要進行NFV效能實驗室執行過程的經驗分享，也介紹如何透過標準化的測試流程，來比較不同網路加速方案的效能差異。
- 使用正確的網路架構與加速方案來實現網路功能的虛擬化，並滿足商用上的效能需求只是第一步，如何做好虛擬化網路功能的管理與維運才是NFV長期且重要的課題。
- 本課程是根據講師在工研院執行NFV效能實驗室的經驗分享，後續如果有任何問題需要討論或交流，可與講師聯絡，聯絡資訊如下：
  - 工研院資通所技術經理 李育緯
  - **Email: rayinlee@itri.org.tw**