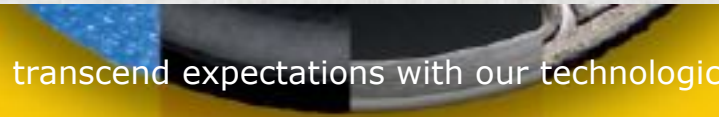
NTT Communications

Transform. Transcend.

100Gbps OpenStack
For Providing High-Performance NFV

Takeaki Matsumoto



Transform your business, transcend expectations with our technologically advanced solutions.

Agenda

- Background
- Goal / Actions
- Kamuee (Software router)
- DPDK application on OpenStack
- Benchmark
- Conclusion

Self-Introduction

Takeaki Matsumoto

takeaki.matsumoto@ntt.com

NTT Communications
Technology Development

R&D for OpenStack
Ops for Private Cloud

Background

- NTT Communications
  - A Global Tier-1 ISP in 196 countries/regions
  - Over 150 datacenters in the world
- Problems
  - Costs
    - spending 1M+ USD for each core router
  - Flexibility
    - long time to add router, orchestration, rollback...

Goal / Actions

- Goal
  - Cheaper and more flexible router with 100Gbps performamce
- Actions
  - Research & verify software router requirements
  - Check the OpenStack functions for NFV
  - Benchmark the software router performance with OpenStack

Kamuee

- Software router with 100Gbps+ (on Baremetal)
  - Developed by NTT Communications
  - 146Gbps with 610K+ IPv4 Routes and 128Byte packets
  - Using technologies
    - DPDK
    - Poptrie
    - RCU
  - Achieving 100Gbps Performance at Core with Poptrie and Kamuee Zero
    https://www.youtube.com/watch?v=OhHv3O1H8-w

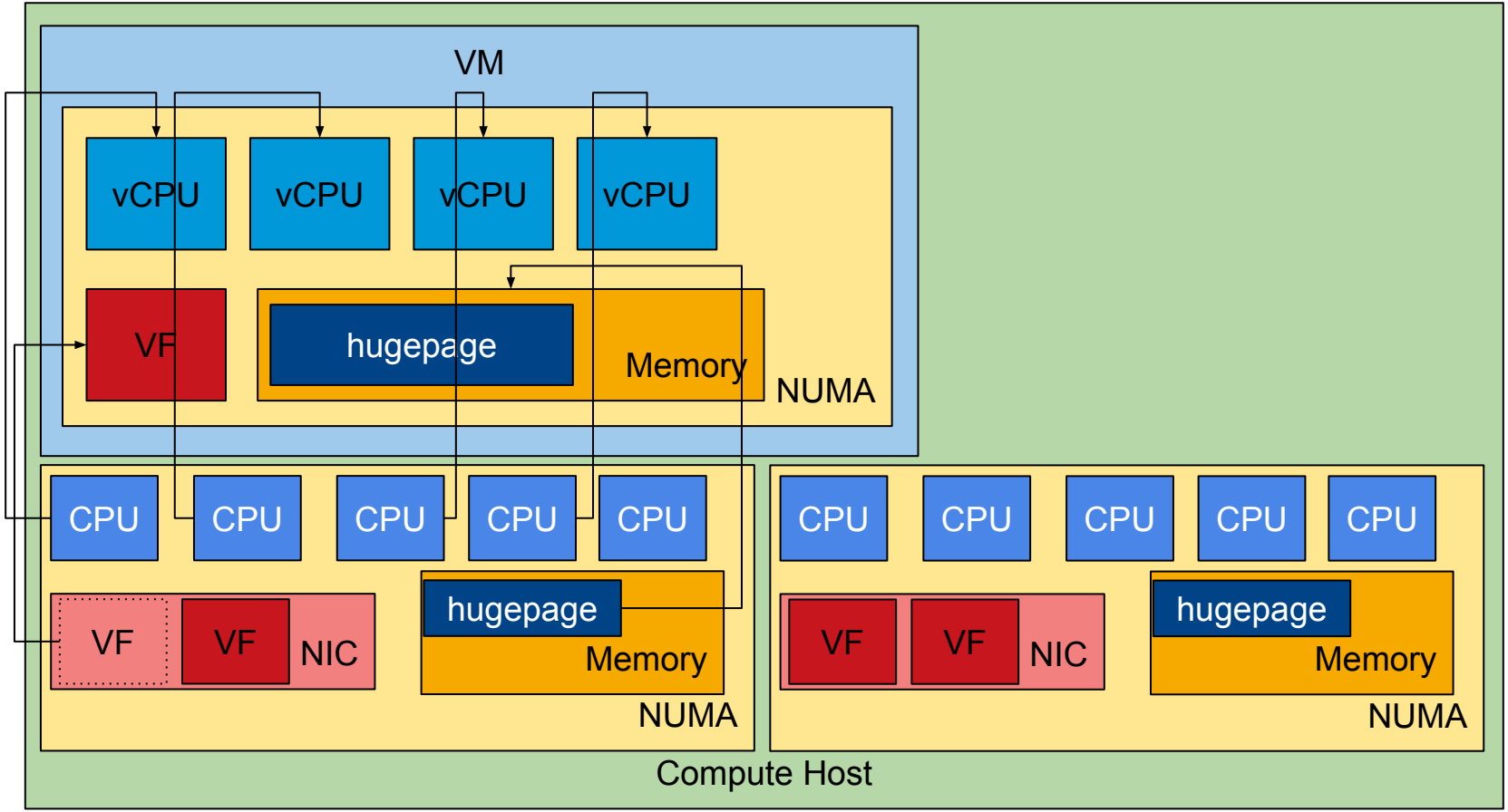
Requirements

- High-performance NFV requirements
  - High-bandwidth network port
  - Low-latency communication NIC-to-CPU
  - Dedicated CPU cores
  - Hugepages
  - CPU features

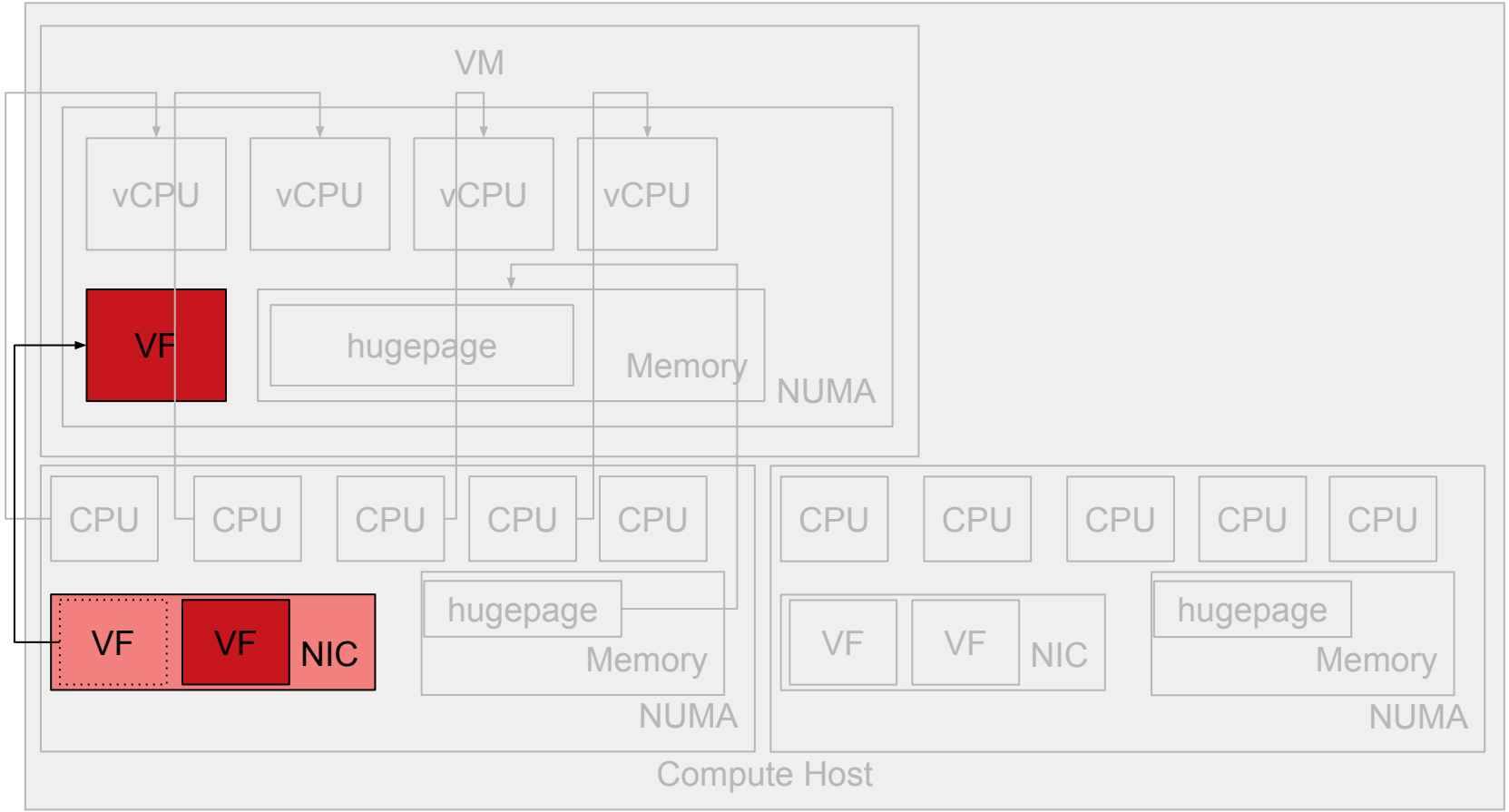
Agenda

- Background
- Goal / Actions
- Kamuee (Software router)
- DPDK application on OpenStack
  - SR-IOV
  - NUMA
  - vCPU pinning
  - Hugepages
  - CPU feature
- Benchmark
- Conslusion

DPDK application on OpenStack

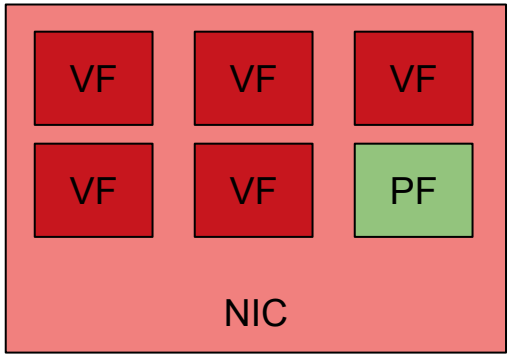


SR-IOV



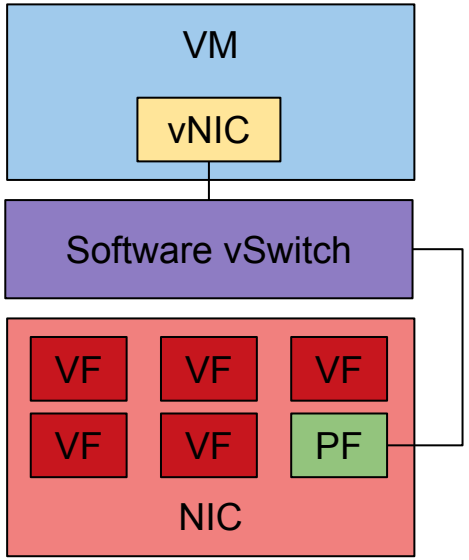
SR-IOV

- What is SR-IOV?
  - Hardware-level virtualization on supported NIC
  - SR-IOV device has
    - Physical Function (PF)
      - Normal NIC device (1 device/physical port)
    - Virtual Funtion (VF)
      - Virtual NIC device from PF
      - can be created up to NIC's limit

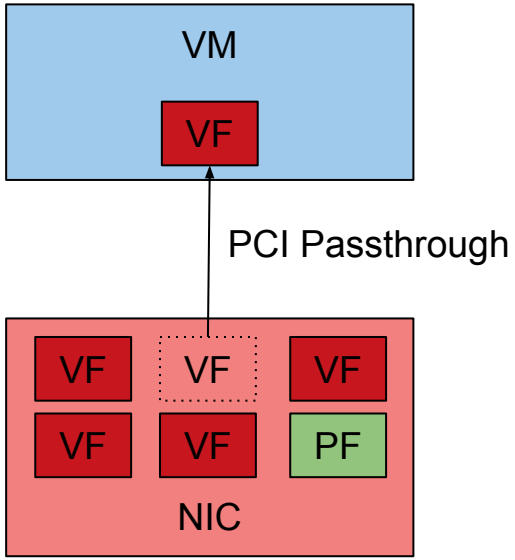


SR-IOV

- Why need SR-IOV?
  - vSwitch can be bottleneck on high-performance network
  - SR-IOV has no effect on Host CPU



Typical Implementation



SR-IOV

SR-IOV

- OpenStack supports SR-IOV
  - VF can be used as Neutron port

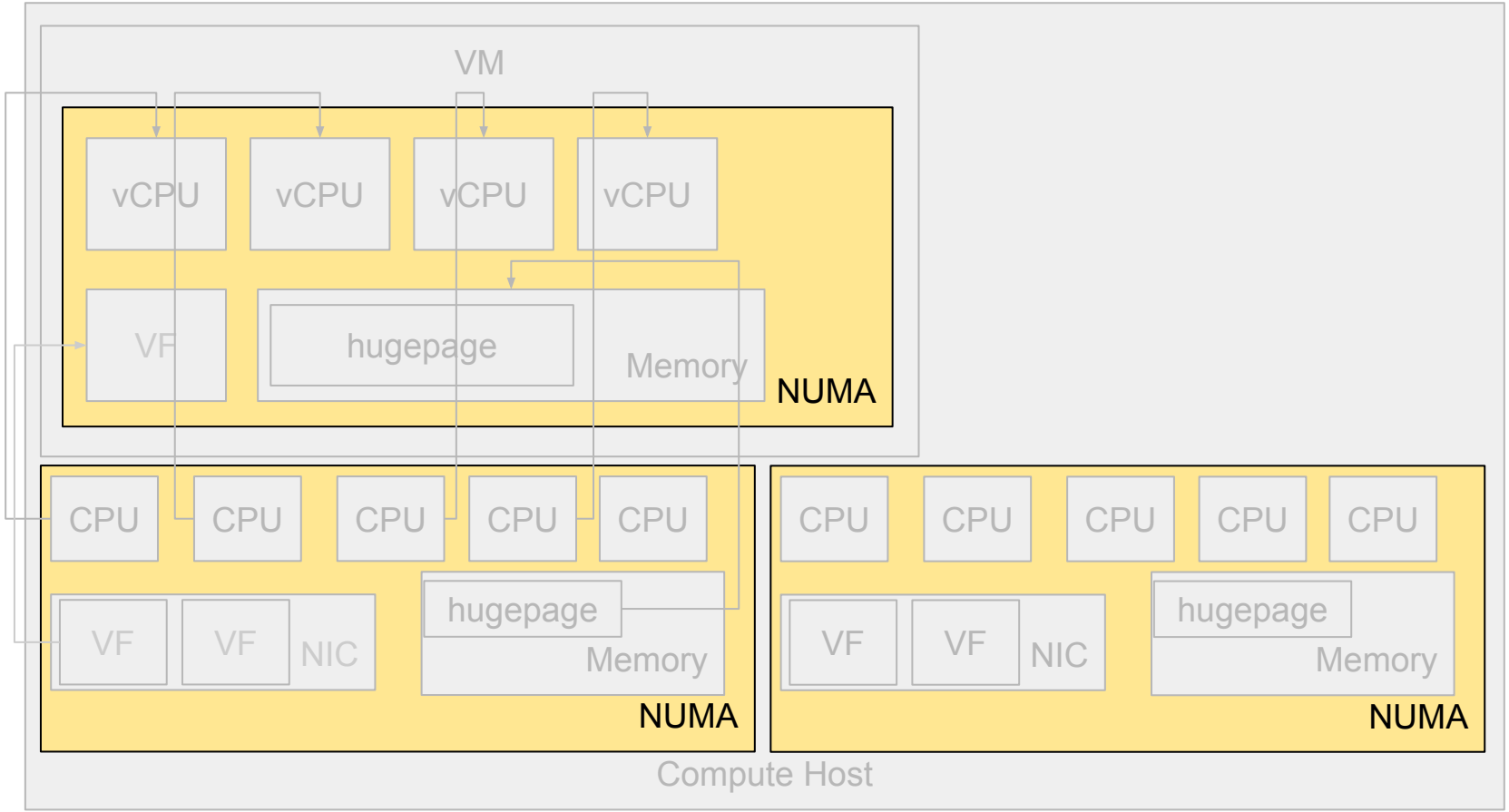
```
$ neutron port-create $net_id --name sriov_port --binding:vnic_type direct
```

```
$ openstack server create --flavor m1.large --image ubuntu_14.04 --nic
port-id=$port_id sriov-server
```

  - Instance get VF directly with PCI-Passthrough

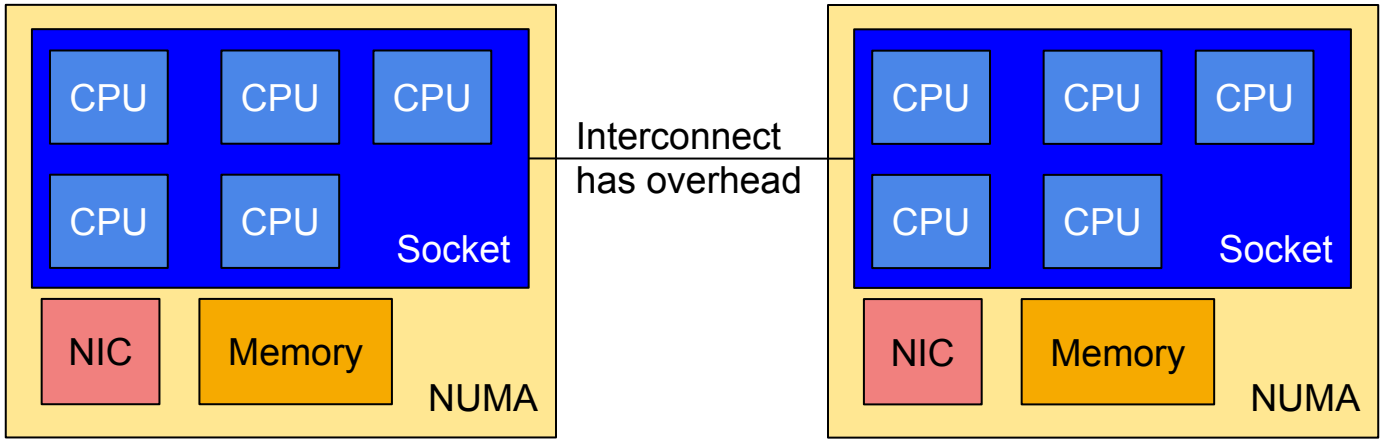
```
ubuntu@sriov-server $ lspci | grep Ethernet
00:05.0 Ethernet controller: Mellanox Technologies MT27700 Family [ConnectX-4
Virtual Function]
```

NUMA



NUMA

- What is NUMA?
  - Non-Uniform Memory Access
  - Server usually has multi NUMA nodes on each CPU socket
  - CPU cores, Memory, PCI devices belong to its NUMA nodes
  - For low-latency, we have to think about NUMA Topology



NUMA

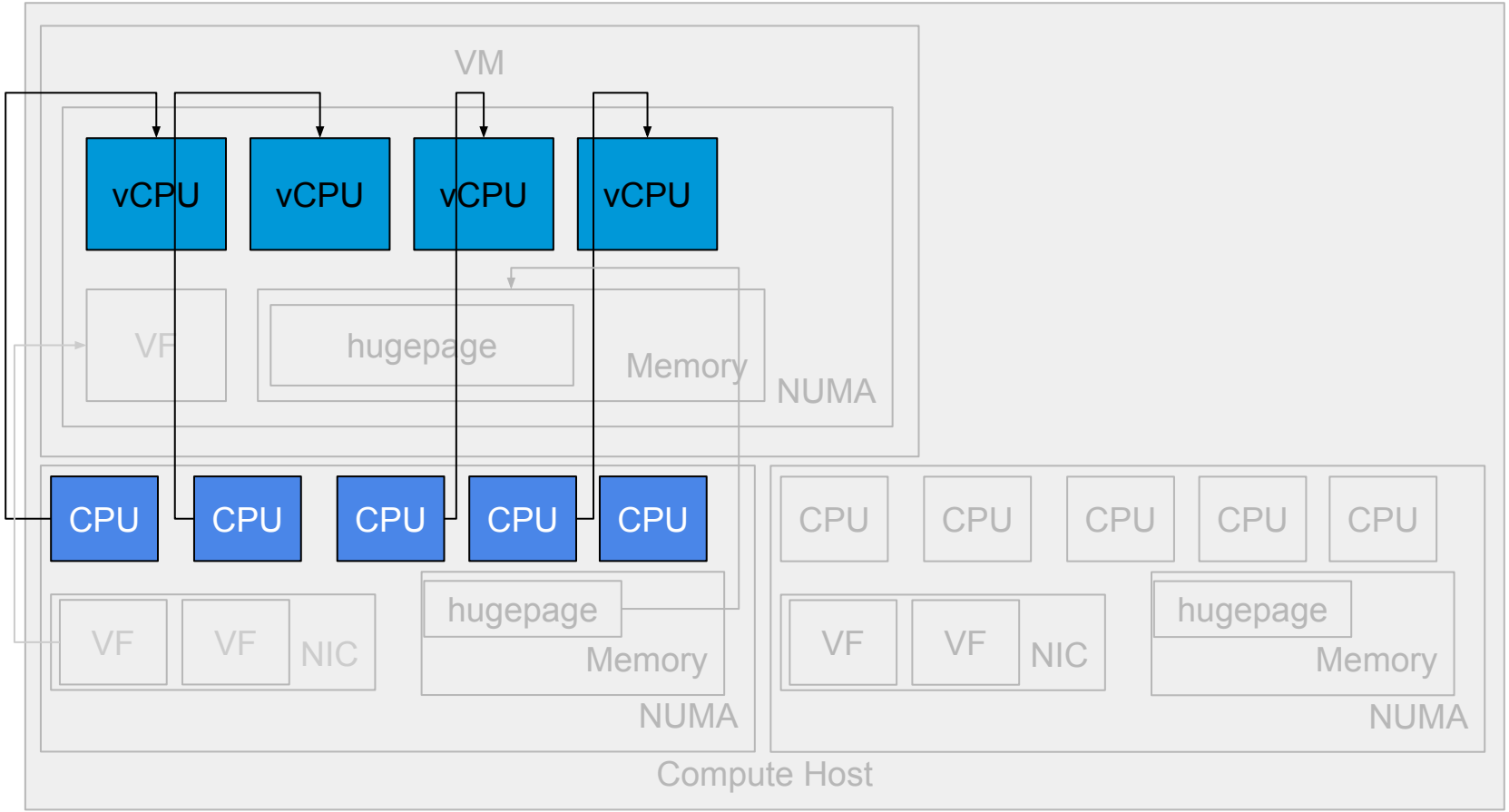
- OpenStack has NUMATopologyFilter
  - can schedule VM with thinking about NUMA topology

```
$ openstack flavor set m1.large --property hw:numa_nodes=1
```

  - When using hugepages or CPU-pinning, automatically launch on same NUMA node
  - 2 NUMA nodes also can be used

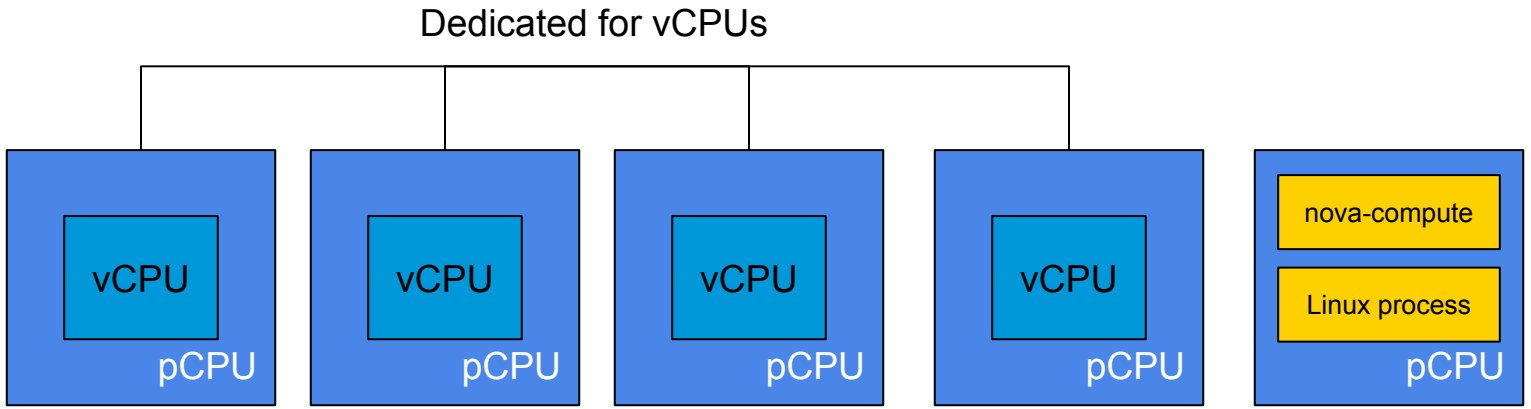
```
$ openstack flavor set m1.large --property hw:numa_nodes=2
```

vCPU pinning



vCPU pinning

- What is vCPU pinning?
  - vCPU:pCPU=1:1 dedicated allocation
    - Reduces context-switching



vCPU pinning

- OpenStack flavor has extra spec "hw:cpu_policy"
  - enables vCPU pinning

```
$ openstack flavor set m1.large --property hw:cpu_policy=dedicated
```

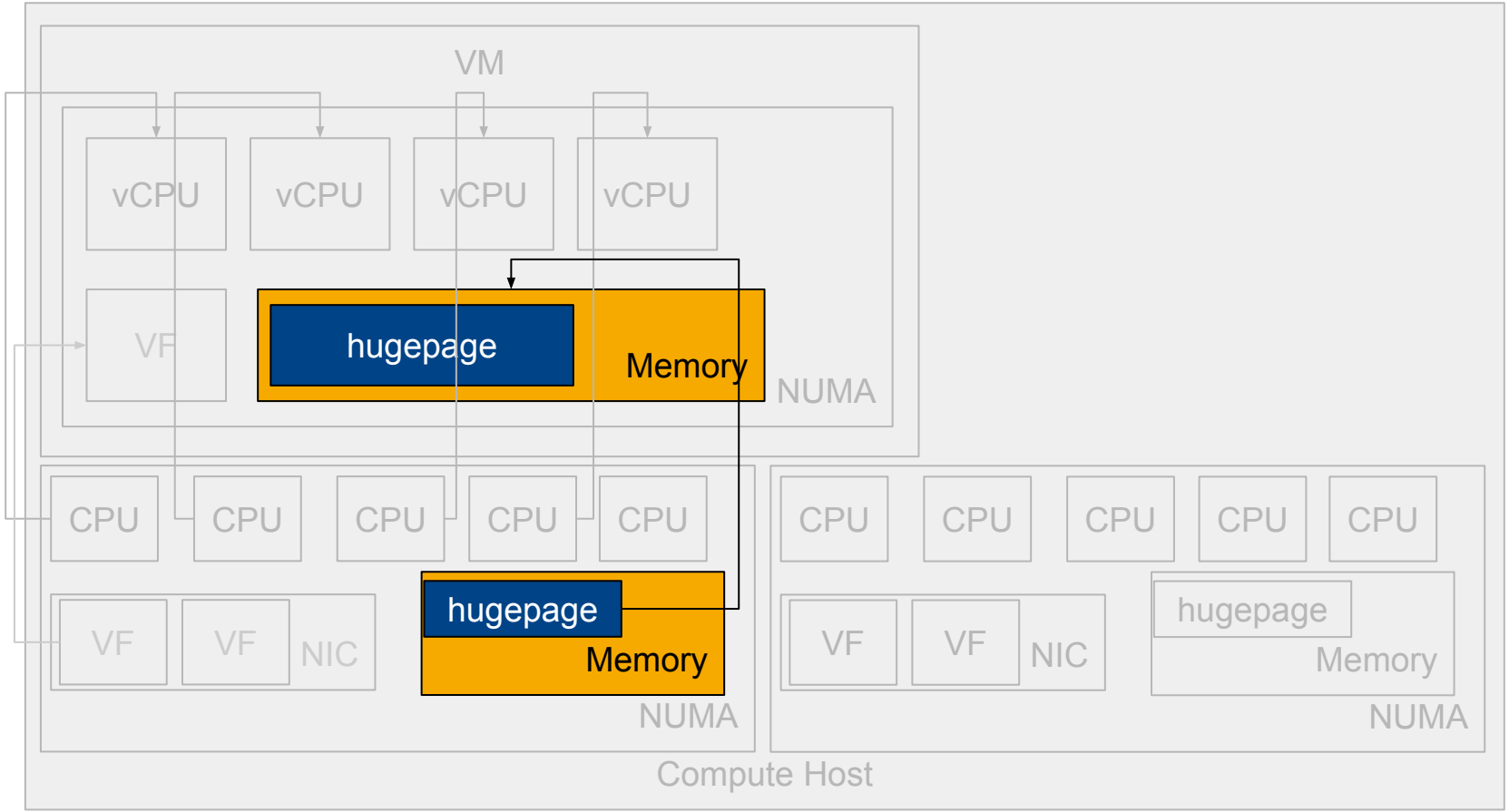
```
$ virsh vcpupin instance-00000001
VCPU: CPU Affinity
----------------------------------
   0: 0-31
   1: 0-31
   2: 0-31
   3: 0-31
   4: 0-31
   5: 0-31
   6: 0-31
   7: 0-31
   8: 0-31
```

Default allocation

```
$ virsh vcpupin instance-00000002
VCPU: CPU Affinity
----------------------------------
   0: 1
   1: 2
   2: 3
   3: 4
   4: 5
   5: 6
   6: 7
   7: 8
   8: 9
```

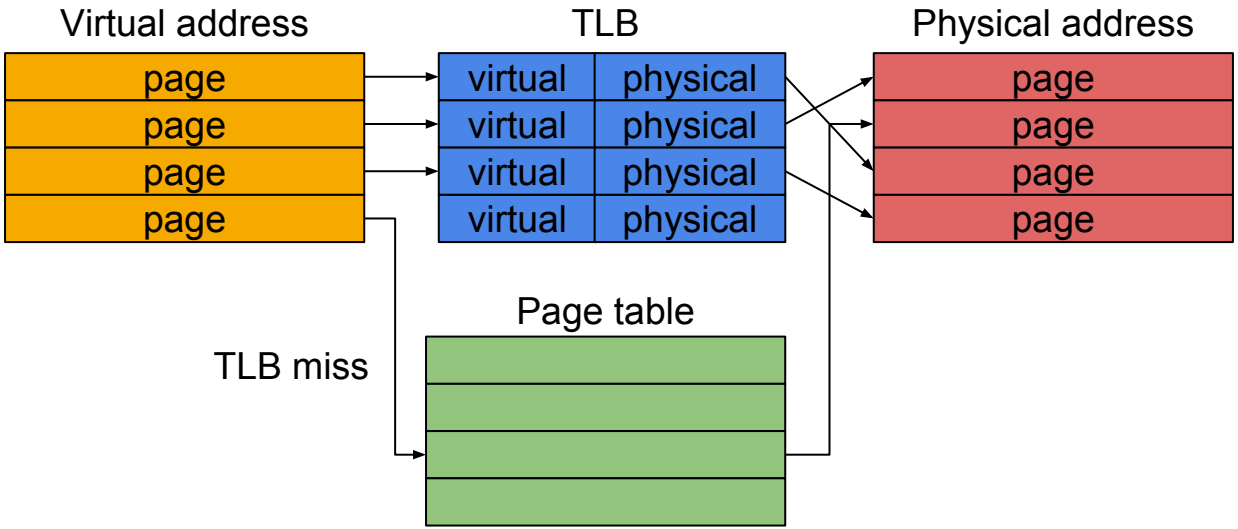
vCPU pinning

Hugepages



Hugepages

- What is Hugepages?
  - segmented pages in memory from 4KB to larger size
    - Reduces TLB misses
    - DPDK applications usually use Hugepages



Hugepages

- OpenStack flavor has extra spec "hw:mem_page_size"
  - Enables Hugepages and assign to guest

```
$ cat /proc/meminfo | grep Hugepagesize
Hugepagesize:    1048576 kB
```

```
$ openstack flavor set m1.large --property hw:mem_page_size=1048576
```

```
$ cat /etc/libvirt/qemu/instance-00000002.xml | grep hugepages -1
  <memoryBacking>
    <hugepages>
      <page size='1048576' unit='KiB' />
    </hugepages>
  </memoryBacking>
```

Other CPU features

- Optimization feature for DPDK
  - SSSE3, SSE4,...
- "[libvirt] cpu_mode" option in nova.conf
  - By default, none is set in some distribution
  - host-model, host-passthrough, or custom is required

```
$ cat /proc/cpuinfo | grep -e model\ name -e flags
model name      : QEMU Virtual CPU version 2.0.0
flags           : fpu de pse tsc msr pae mce cx8 apic sep
mtrr pge mca cmov pse36 clflush mmx fxsr sse sse2
syscall nx lm rep_good nopl pni vmx cx16 x2apic popcnt
hypervisor lahf_lm vnmi ept
```

cpu_mode=none

```
$ cat /proc/cpuinfo | grep -e model\ name -e flags
model name      : Intel Core Processor (Broadwell)
flags           : fpu vme de pse tsc msr pae mce cx8 apic
sep mtrr pge mca cmov pat pse36 clflush mmx fxsr sse
sse2 ss syscall nx pdpe1gb rdtscp lm constant_tsc
rep_good nopl xtopology eagerfpu pni pclmulqdq vmx
ssse3 fma cx16 pcid sse4_1 sse4_2 x2apic movbe popcnt
tsc_deadline_timer aes xsave avx f16c rdrand hypervisor
lahf_lm abm 3dnowprefetch tpr_shadow vnmi flexpriority
ept vpid fsgsbase tsc_adjust bmi1 hle avx2 smep bmi2
erms invpcid rtm rdseed adx smap xsaveopt arat
```

cpu_mode=host-model

Agenda

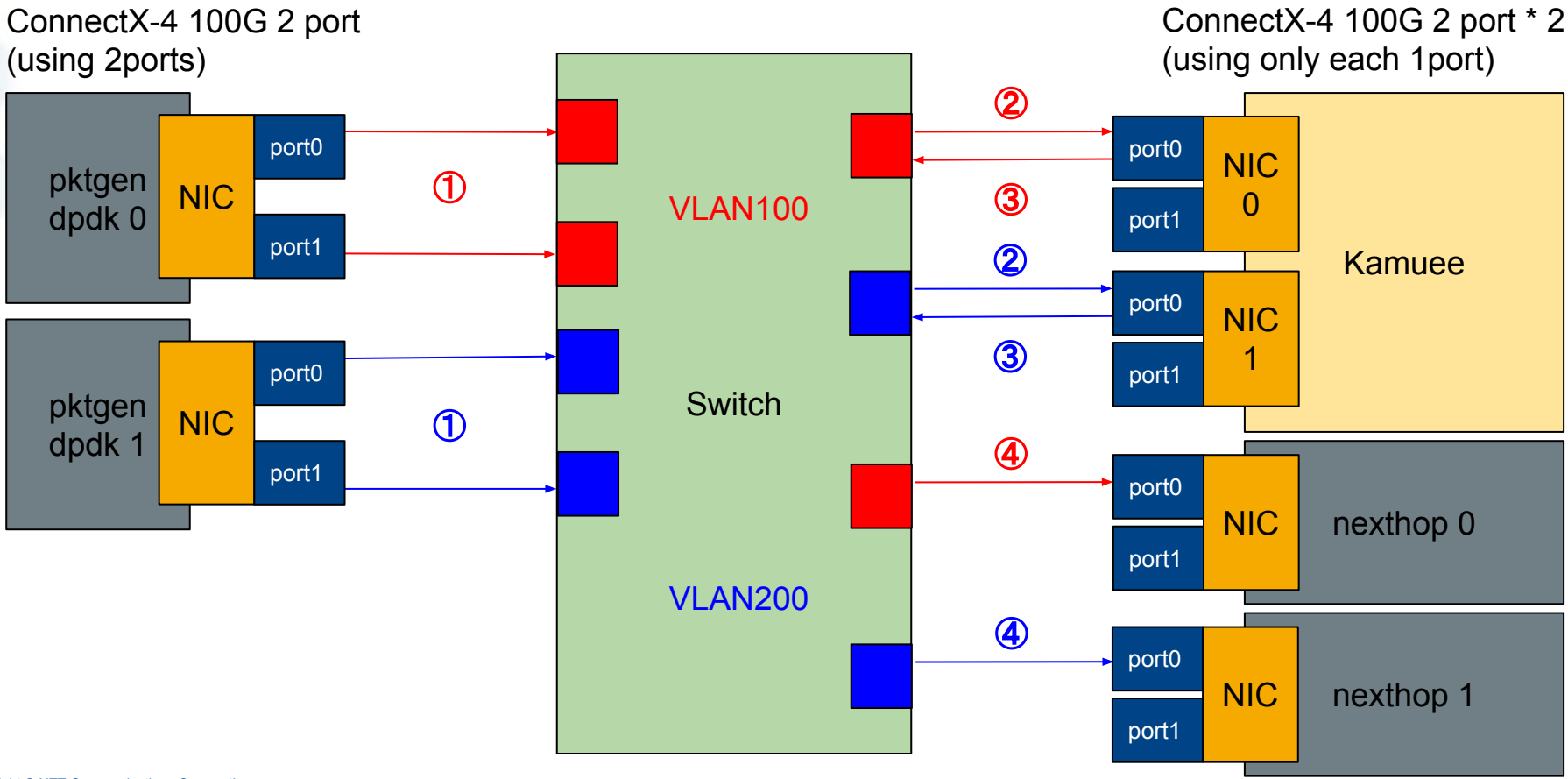
- Background
- Goal / Actions
- Kamuee (Software router)
- DPDK application on OpenStack
- Benchmark
  - Environment
  - Baremetal performance
  - VM + VF performance
  - VM +PF performance
  - Baremetal (VF exists) performance

Environment: Hardware

- Server
  - Dell PowerEdge R630
    - Intel® Xeon® CPU E5-2690 v4 @ 2.60GHz (14 cores) * 2
    - DDR-4 256GB (32GB * 8)
    - Ubuntu 16.04
- NIC
  - Mellanox ConnectX-4 100Gb/s Dual-Port Adapter
    - 1 PCIe Card, 100G Ports * 2
- Switch
  - Dell Networking Z9100
    - Cumulus Linux 3.2.0
    - 100Gbps Port * 32

Environment: Architecture

※Each line is 100G link



Environment: pktgen-dpdk

- Open source packet generator
  - Output: about 100Mpps≒67.2Gbps/server (64Byte packet)
    - 50Mpps/port

```
| Ports 0-1 of 2   <Main Page>  Copyright (c) <2010-2017>, Intel Corporation
  Flags:Port      :   P-----R--------:0   P-----R--------:1
Link State        :      <UP-100000-FD>      <UP-100000-FD>      ----TotalRate----
Pkts/s Max/Rx     :                 0/0                 0/0                 0/0
       Max/Tx     :  50040027/50007295  50054111/49993569  100001053/100000864
```

  - dst mac
    - kamuee NIC0 port0 (port0-1 on pktgen-dpdk 0)
    - kamuee NIC1 port0 (port0-1 on pktgen-dpdk 1)
  - dst ip (range)
    - 1.0.0.1-254 (port0 on each server)
    - 1.0.4.1-254 (port1 on each server)
  - dst TCP port (range)
    - 1-254 (port0 on each server)
    - 256-510 (port1 on each server)

Environment: Kamuee

- DPDK software router
- Spec configuration
  - 2 NUMA nodes
    - Using 26 cores
      - Forwarding: 12 cores/port * 2 (each NUMA)
      - Other functions: 2 cores
    - Using 16GB memory
      - 1GB Hugepages * 8 * 2 (each NUMA)
    - 2 NICs
      - only port 0 is used * 2 (each NUMA)

Environment: Kamuee

- Routing configuration
  - 518K routes (like Fullroute) loaded
    - Forwading to nexthop server

```
kamuee-console> show ipv4 route
1.0.0.0/24 nexthop: 172.21.4.105
1.0.4.0/24 nexthop: 172.21.3.104
...
```

```
kamuee-console> show ipv4 route 172.21.3.104
172.21.3.104/32 ether: 24:8a:07:4c:2f:6c port 0
```

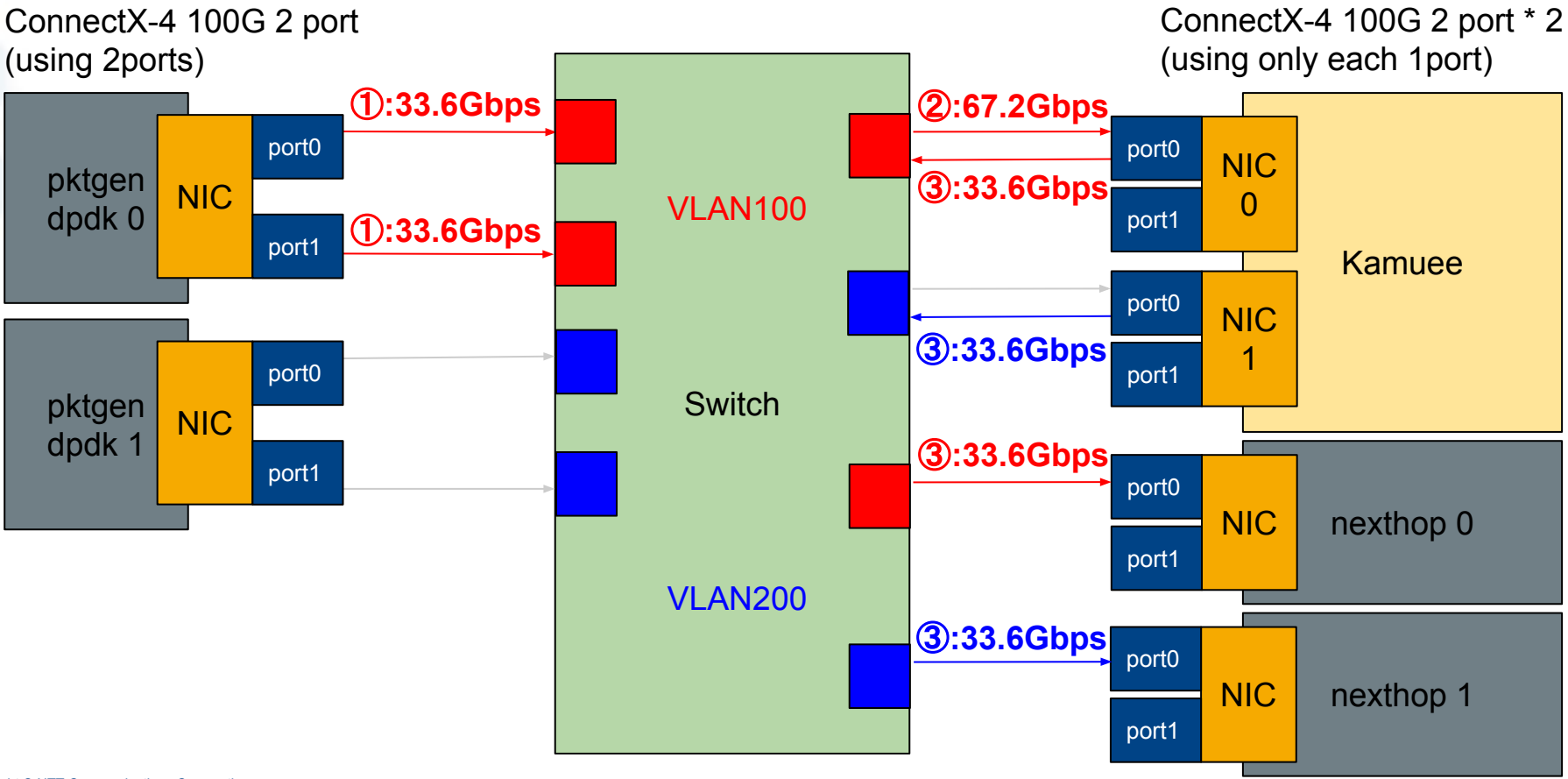
```
kamuee-console> show ipv4 route 172.21.4.105
172.21.4.105/32 ether: 24:8a:07:4c:2f:64 port 1
```

- DPDK EAL options
  - ./kamuee -n 4 --socket-mem 8192,8192 -w 0000:00:05.0,txq_inline=128 -w 0000:00:06.0,txq_inline=128

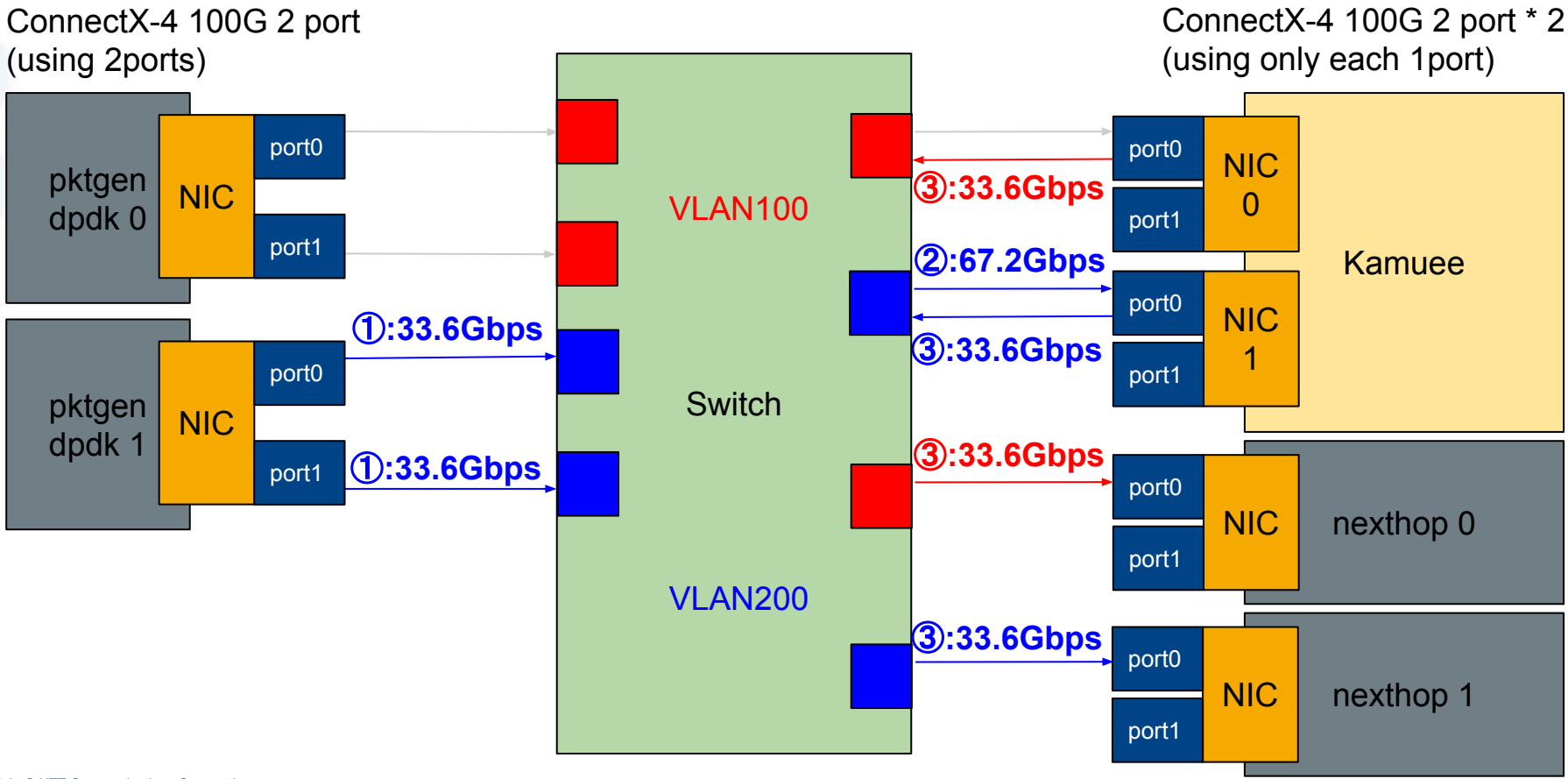
Environment: nexthop

- Measuring RX packets
  - Using eth_stat.sh
    - https://community.mellanox.com/docs/DOC-2506#jive_content_id_How_to_Measure_Ingress_Rate
    - using "rx_packets_phy" on ethtool
      - hardware-level packet counter

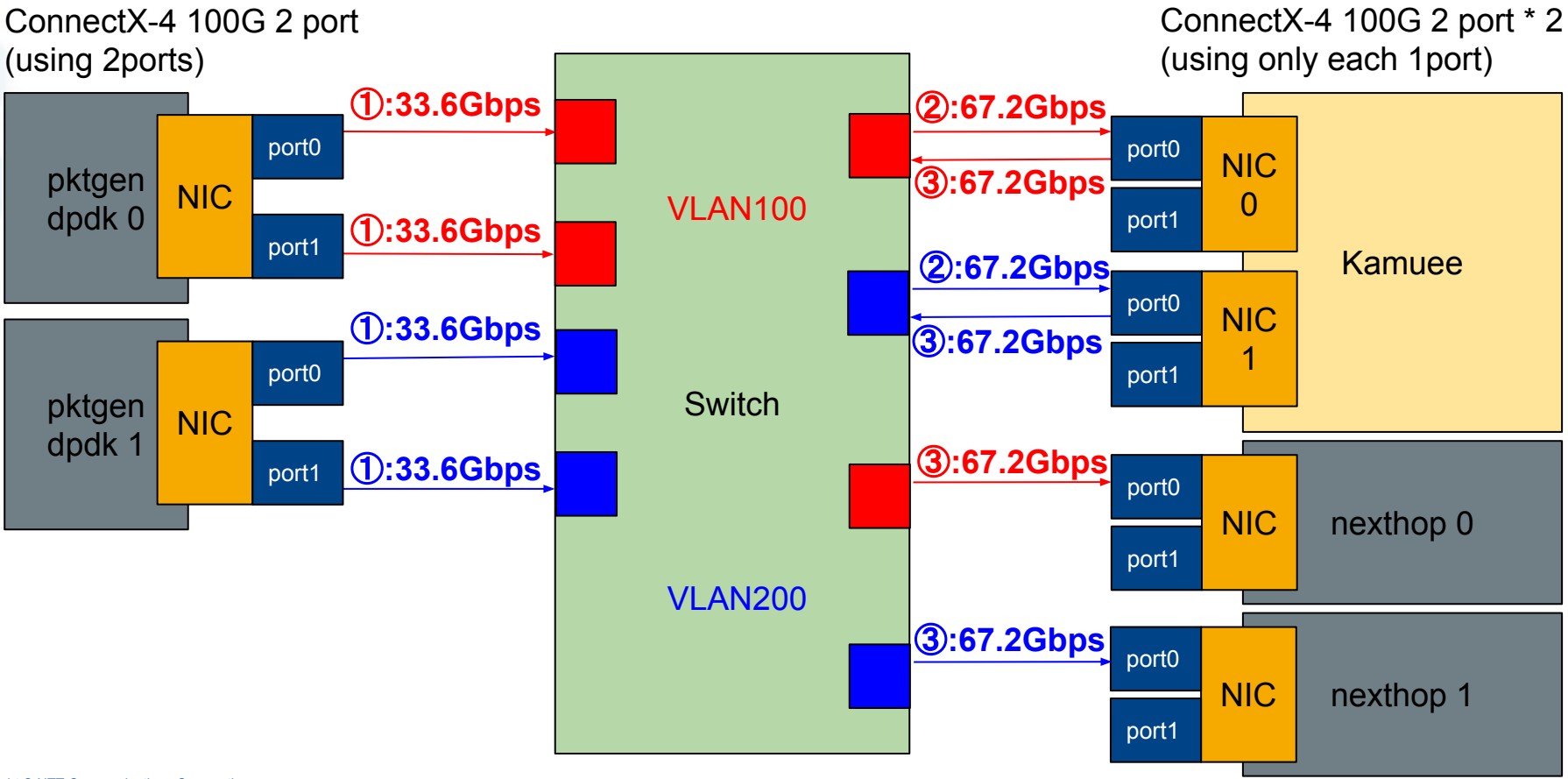
Environment: Ideal flow on each pktgen server (64Byte)



Environment: Ideal flow on each pktgen server (64Byte)



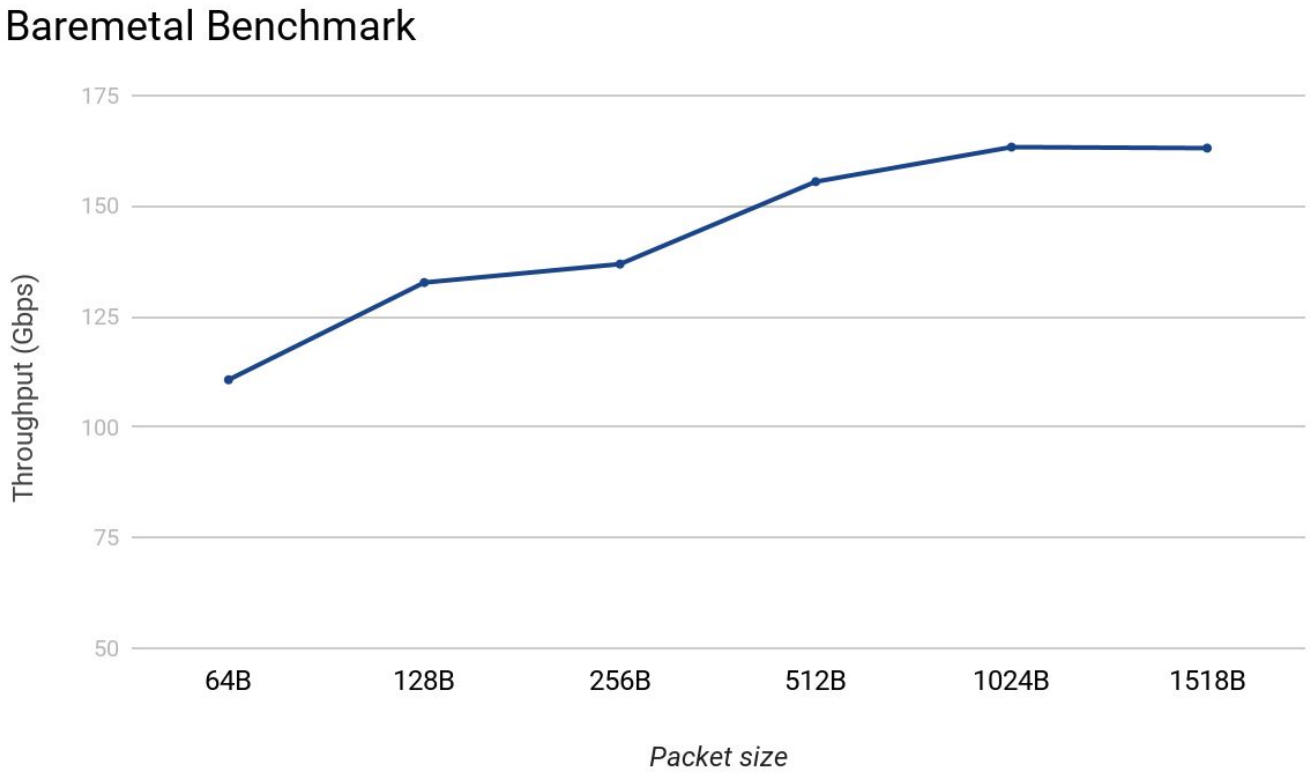
Environment: Ideal flow (64Byte)



Baremetal performance: Configuration

- BIOS
  - Hyper-Threading: OFF
- Boot parameters
  - intel_idle.max_cstate=0 processor.max_cstate=0 intel_pstate=disable nohz_full=1-27 rcu_nocbs=1-27 rcu_novb_poll audit=0 nosoftlockup default_hugepagesz=1G hugepagesz=1G hugepages=32 isolcpus=1-27
- Mellanox
  - CQE_COMPRESSION: AGGRESSIVE(1)
  - SRIOV_EN: False(0)
- Ports
  - 2 PFs (only port0 on each NIC)

Baremetal performance: Result



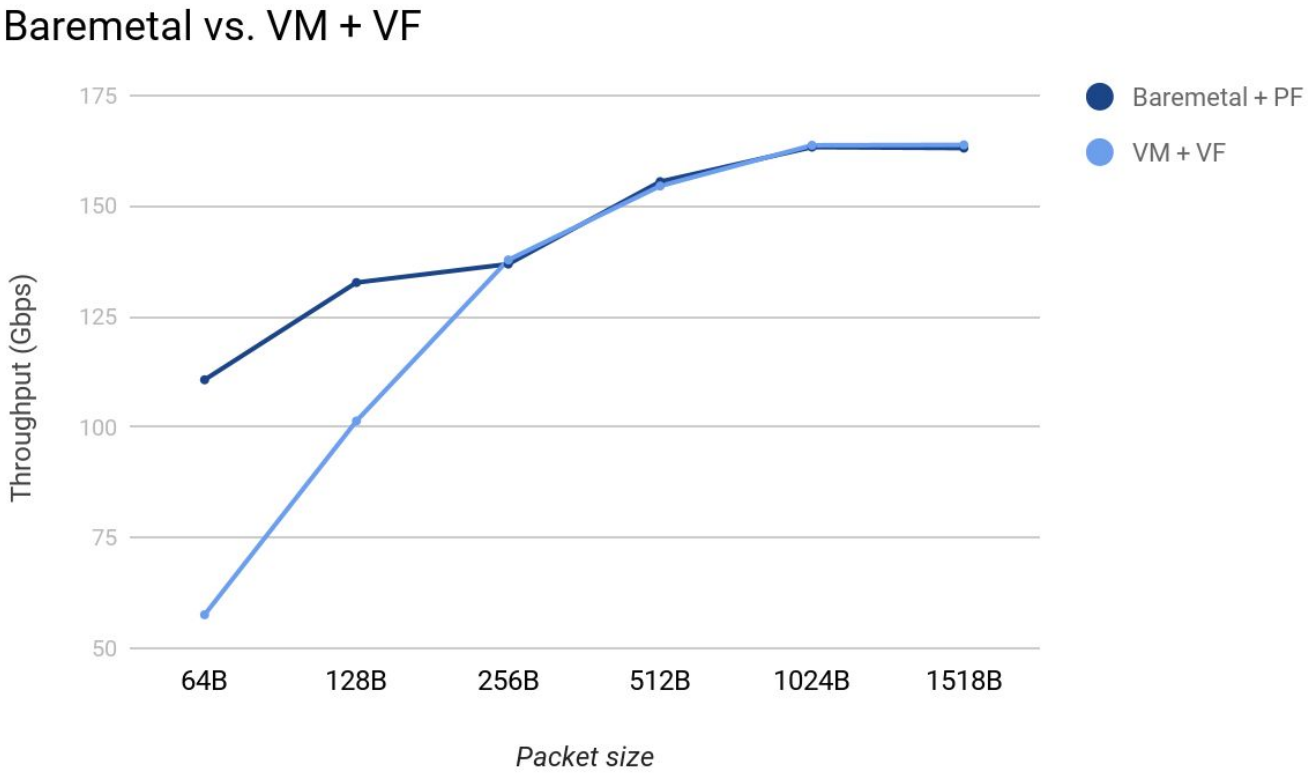
VM + VF performance: Host Configuration

- BIOS
  - Hyper-Threading: OFF
  - VT-d: ON
- Host boot parameters
  - intel_idle.max_cstate=0 processor.max_cstate=0 intel_pstate=disable nohz_full=1-27 rcu_nocbs=1-27 rcu_novb_poll audit=0 nosoftlockup default_hugepagesz=1G hugepagesz=1G hugepages=32 isolcpus=1-27 intel_iommu=on
- Mellanox
  - CQE_COMPRESSION: AGGRESSIVE(1)
  - SRIOV_EN: True(1)
  - NUM_OF_VFS: 1

VM + VF performance: Guest Configuration

- Flavor
  - vCPUs: 27
  - Memory: 32GB
  - extra_specs:
    - hw:cpu_policy: dedicated
    - hw:mem_page_size: 1048576
    - hw:numa_mem.0: 16384
    - hw:numa_mem.1: 16384
    - hw:numa_cpus.0: 0-13
    - hw:numa_cpus.1: 14-26
    - hw:numa_nodes: 2
- Ports
  - 2 VFs (vf 0 on each NIC port0)
- Guest boot parameters
  - intel_idle.max_cstate=0 processor.max_cstate=0 intel_pstate=disable nohz_full=1-26 rcu_nocbs=1-26 rcu_novb_poll audit=0 nosoftlockup default_hugepagesz=1G hugepagesz=1G hugepages=16 isolcpus=1-26

VM + VF performance: Result



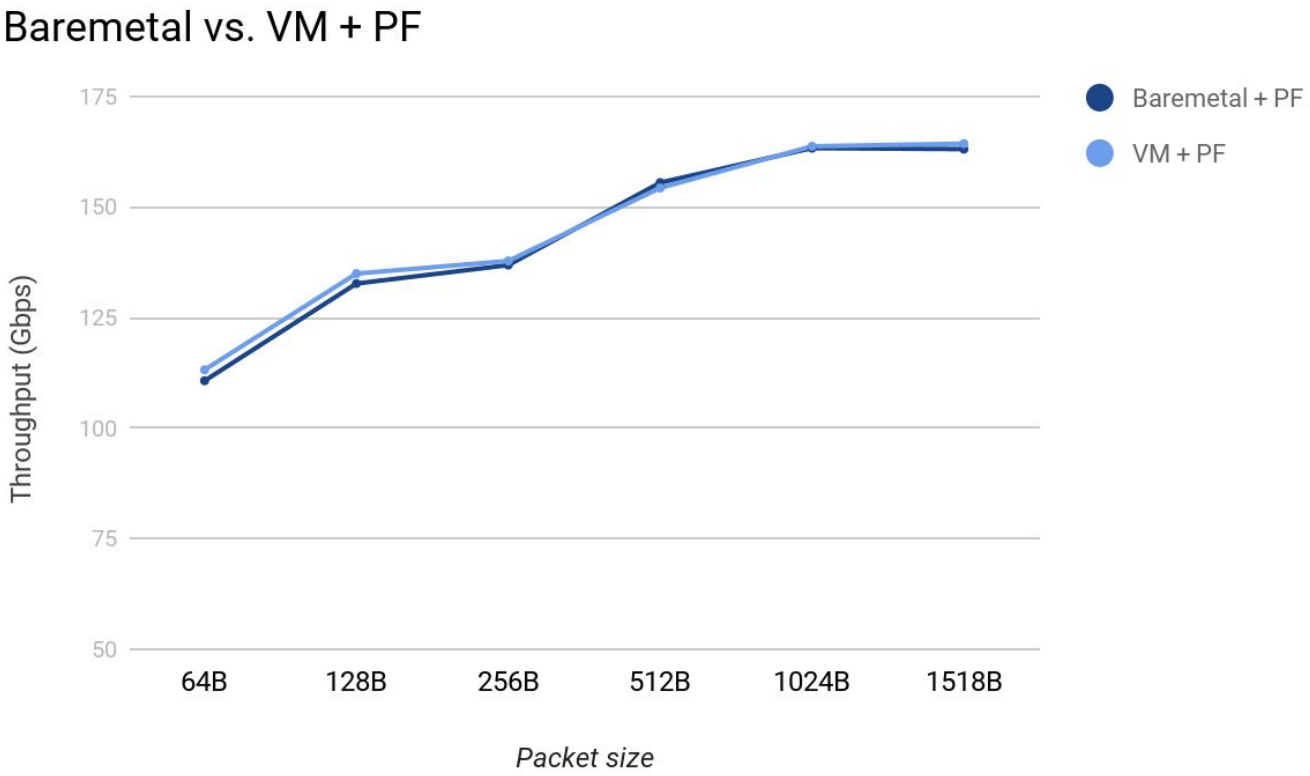
VM + PF performance: Host Configuration

- BIOS
  - Hyper-Threading: OFF
  - VT-d: ON
- Host boot parameters
  - intel_idle.max_cstate=0 processor.max_cstate=0 intel_pstate=disable nohz_full=1-27 rcu_nocbs=1-27 rcu_novb_poll audit=0 nosoftlockup default_hugepagesz=1G hugepagesz=1G hugepages=32 isolcpus=1-27 intel_iommu=on
- Mellanox
  - CQE_COMPRESSION: AGGRESSIVE(1)
  - SRIOV_EN: False(0)

VM + PF performance: Guest Configuration

- Flavor
  - vCPUs: 27
  - Memory: 32GB
  - extra_specs:
    - hw:cpu_policy: dedicated
    - hw:mem_page_size: 1048576
    - hw:numa_mem.0: 16384
    - hw:numa_mem.1: 16384
    - hw:numa_cpus.0: 0-13
    - hw:numa_cpus.1: 14-26
    - hw:numa_nodes: 2
- Ports
  - 2 PFs (only port0 on each NIC with PCI-Passthrough)
- Guest boot parameters
  - intel_idle.max_cstate=0 processor.max_cstate=0 intel_pstate=disable nohz_full=1-26 rcu_nocbs=1-26 rcu_novb_poll audit=0 nosoftlockup default_hugepagesz=1G hugepagesz=1G hugepages=16 isolcpus=1-26

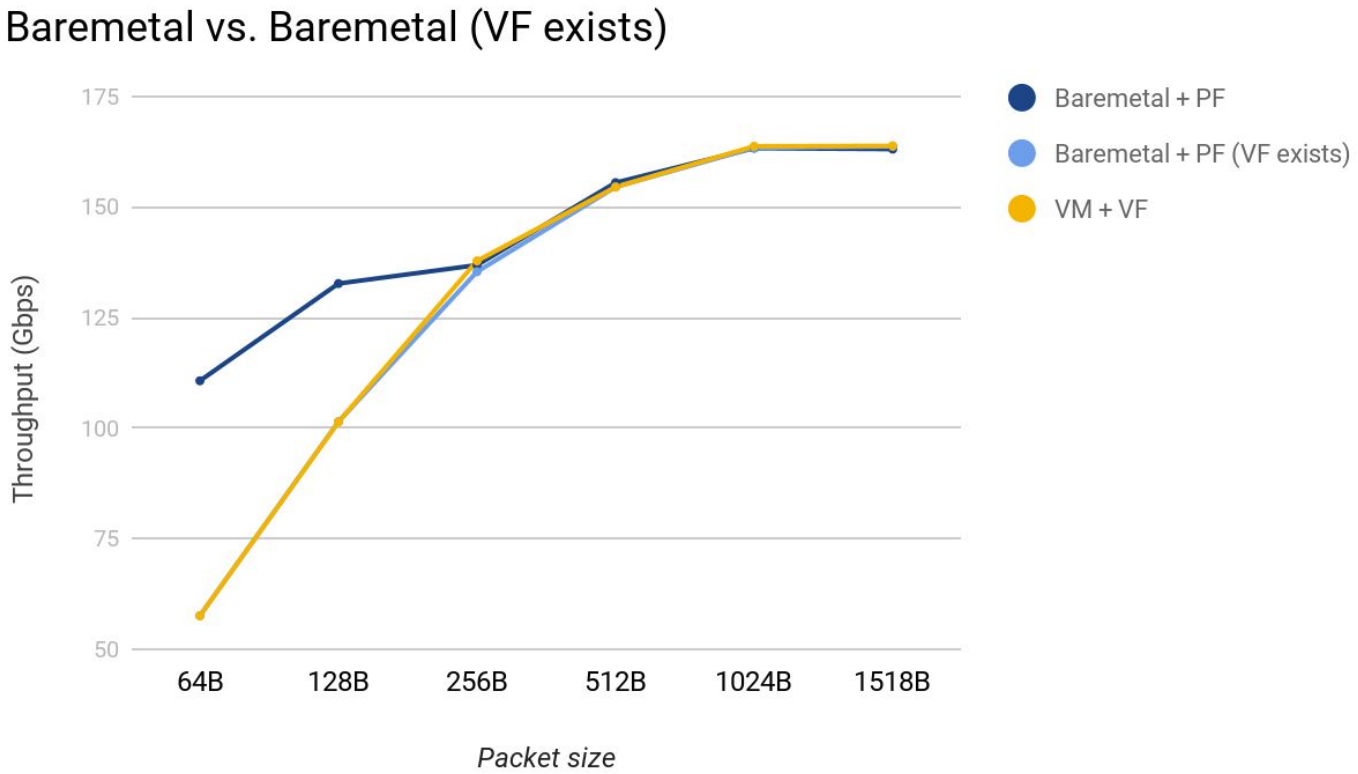
VM + PF performance: Result



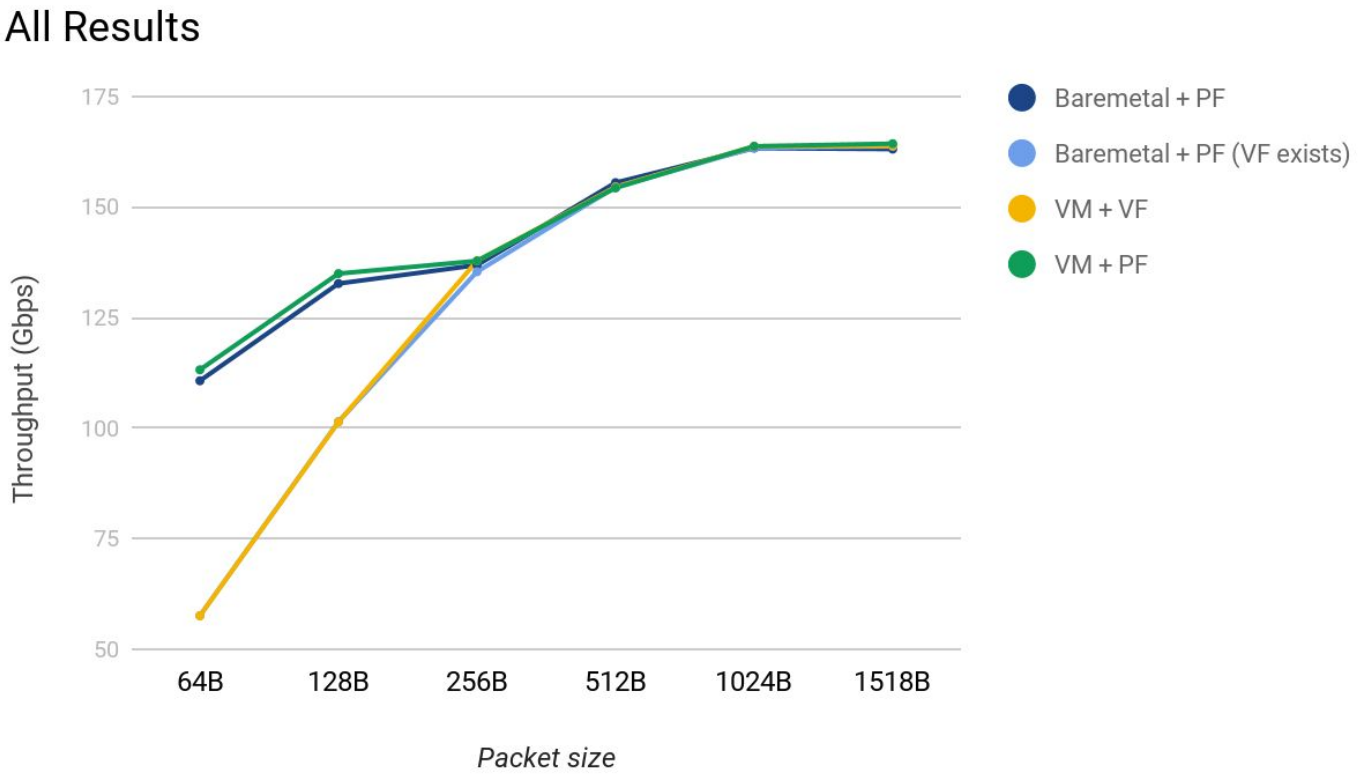
Baremetal (VF exists) performance: Configuration

- BIOS
  - Hyper-Threading: OFF
- Boot parameters
  - intel_idle.max_cstate=0 processor.max_cstate=0 intel_pstate=disable nohz_full=1-27 rcu_nocbs=1-27 rcu_novb_poll audit=0 nosoftlockup default_hugepagesz=1G hugepagesz=1G hugepages=32 isolcpus=1-27
- Mellanox
  - CQE_COMPRESSION: AGGRESSIVE(1)
  - SRIOV_EN: True(1)
  - NUM_OF_VFS: 1
- Ports
  - 2 PFs (only port0 on each NIC)
  - 2 VFs (vf 0 on each NIC port0) exists [not used]

Baremetal (VF exists) performance: Result



All Results



Conclusion

- OpenStack functions for NFV works fine
  - SR-IOV port assignment
  - NUMA awareness
  - vCPU pinning
  - Hugepages
  - CPU Feature
- KVM + Intel VT archive close to baremetal performance
- SR-IOV performance evaluation is required
  - SR-IOV device implementation depends on its vendor

Conclusion

- Our decision
  - VM + PF is powerful option
    - SR-IOV advantange
      - Multiple VF can be created
        - Router
        - Firewall
        - Load balancer
        - ...
    - 100G router consumes almost host resources
      - "1 Host: 1 VM" is realistic option
        - no need so many ports

Thank you!

References

- Kamuee Zero
  - https://www.internetconference.org/ic2016/PDF/ic2016-paper-03.pdf
- Poptrie
  - https://conferences.sigcomm.org/sigcomm/2015/pdf/papers/p57.pdf
- SR-IOV
  - https://docs.openstack.org/ocata/networking-guide/config-sriov.html
- How to enable SR-IOV with Mellanox NIC
  - https://community.mellanox.com/docs/DOC-2386
- Hugepages
  - https://www.mirantis.com/blog/mirantis-openstack-7-0-nfvi-deployment-guide-huge-pages/
- isolcpu & cpupinning
  - https://docs.mirantis.com/mcp/1.0/mcp-deployment-guide/enable-numa-and-cpu-pinning/enable-numa-and-cpu-pinning-procedure.html
- NUMA
  - https://docs.openstack.org/nova/pike/admin/cpu-topologies.html