Towards Converged SmartNIC Architecture for Bare Metal & Public Clouds

Layong (Larry) Luo, Tencent TEG

August 8, 2018

Agenda

1 SmartNIC in Bare Metal Cloud

2 SmartNIC in Public Cloud

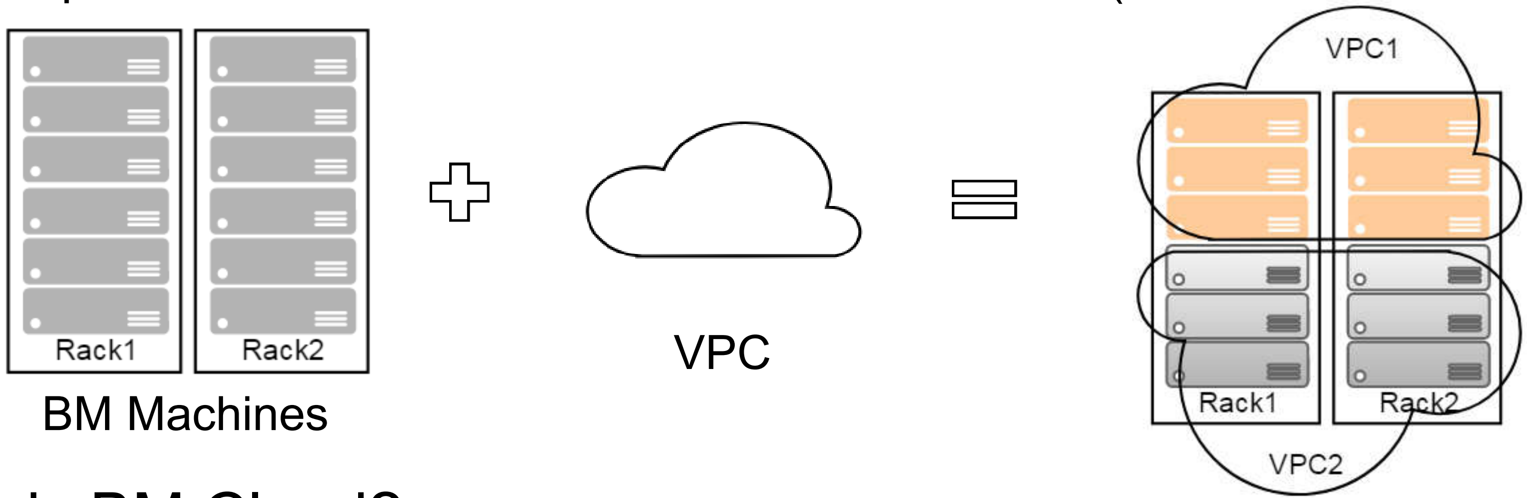
3 Converged SmartNIC Architecture

4 Tencent SmartNIC Experience

5 Future Challenges

Introduction to Bare Metal Cloud

- **What** is Bare Metal (BM) Cloud?
  - – data centers in which dedicated physical machines (aka ***bare metal machines***) are provided to customers via cloud service model (VPC)



- **Why** is BM Cloud?
  - – Addressed two big obstacles for cloud adaption
    - Performance degradation: No virtualization overhead in CPU
    - Migration cost: Exactly the same stacks, tools and experience as on-premises

Introduction to Bare Metal Cloud

- **Who** is using BM Cloud in Tencent: typical use cases

VIPKID

Hybrid Cloud inside Tencent
- Frontend services in Public Cloud (VMs)
- Backend big data services in BM Cloud (PMs)
  - IO intensive, CPU intensive

蘑菇街
mogujie.com

Custom Virtualization Stack
- Custom Portal & OpenStack: smooth migration
  - Customer has strong technical teams
- Consistent experience as on-premises

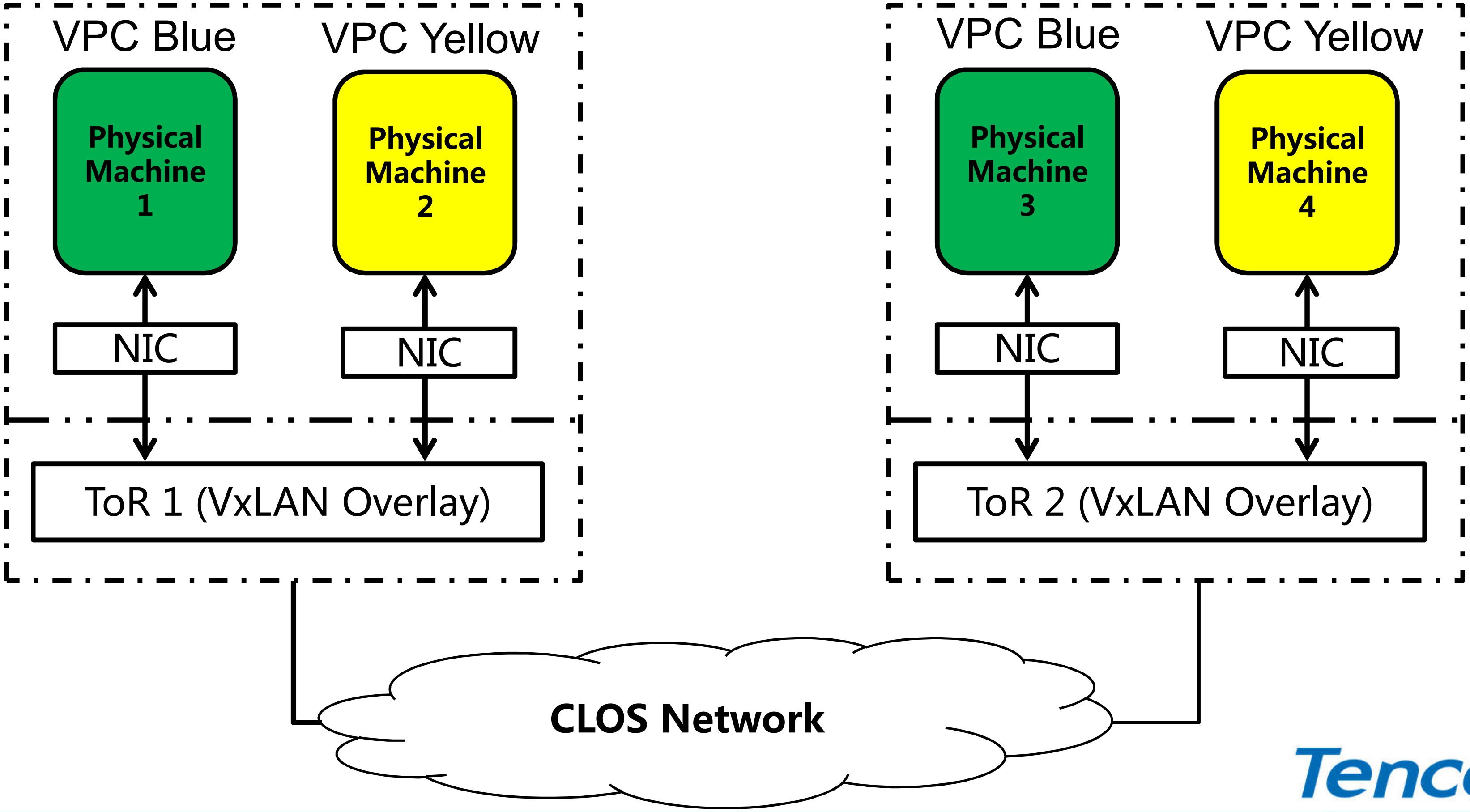
mobike
摩拜单车

Container Cloud
- Container cloud for serverless computing
- No virtualization overhead

Learning more: https://cloud.tencent.com/product/cpm

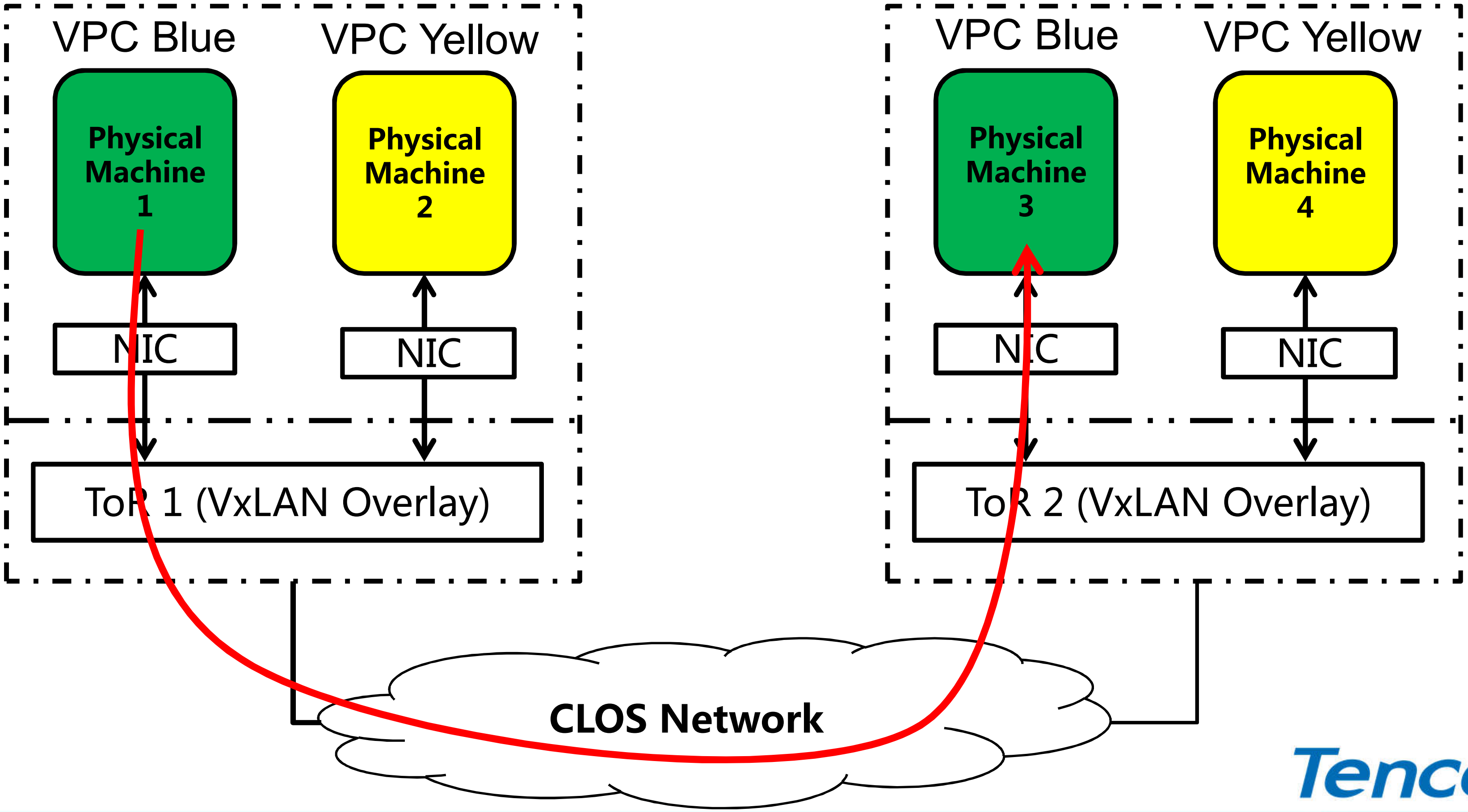
Introduction to Bare Metal Cloud

- **How** to implement BM Cloud: ToR based Virtualization
  - Requirements: any server from anywhere to any customer, BYO IP addresses



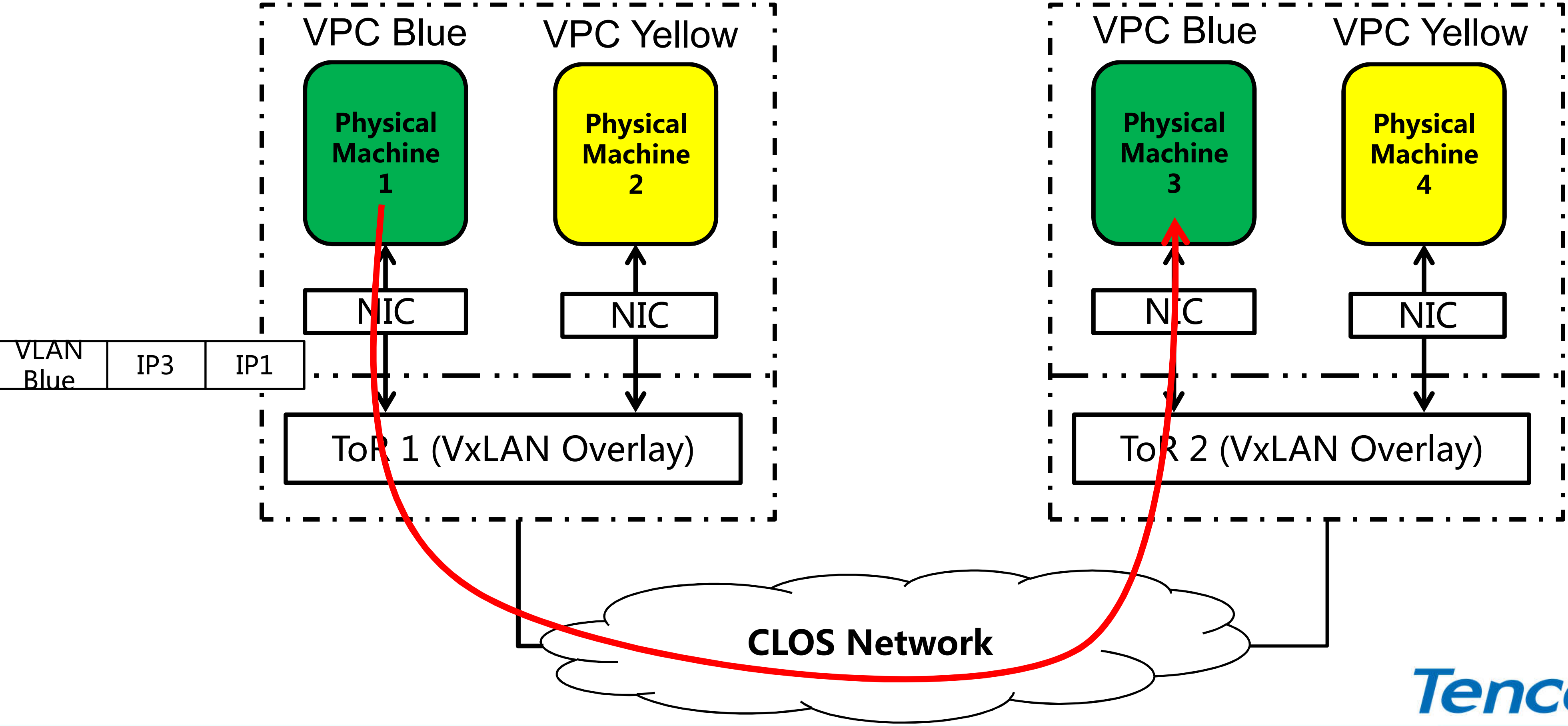
Introduction to Bare Metal Cloud

- **How** to implement BM Cloud: ToR based Virtualization
  - Requirements: any server from anywhere to any customer, BYO IP addresses



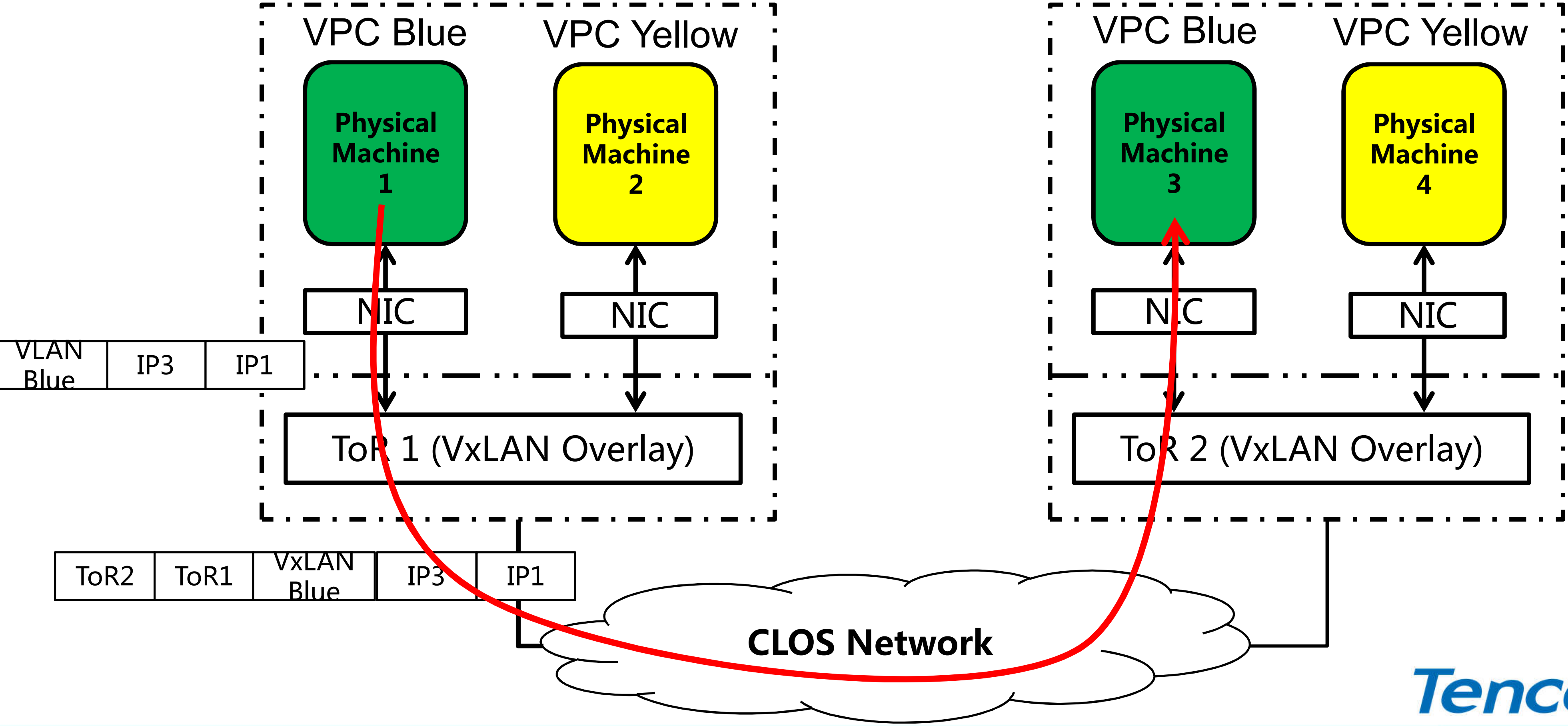
Introduction to Bare Metal Cloud

- **How** to implement BM Cloud: ToR based Virtualization
  - Requirements: any server from anywhere to any customer, BYO IP addresses



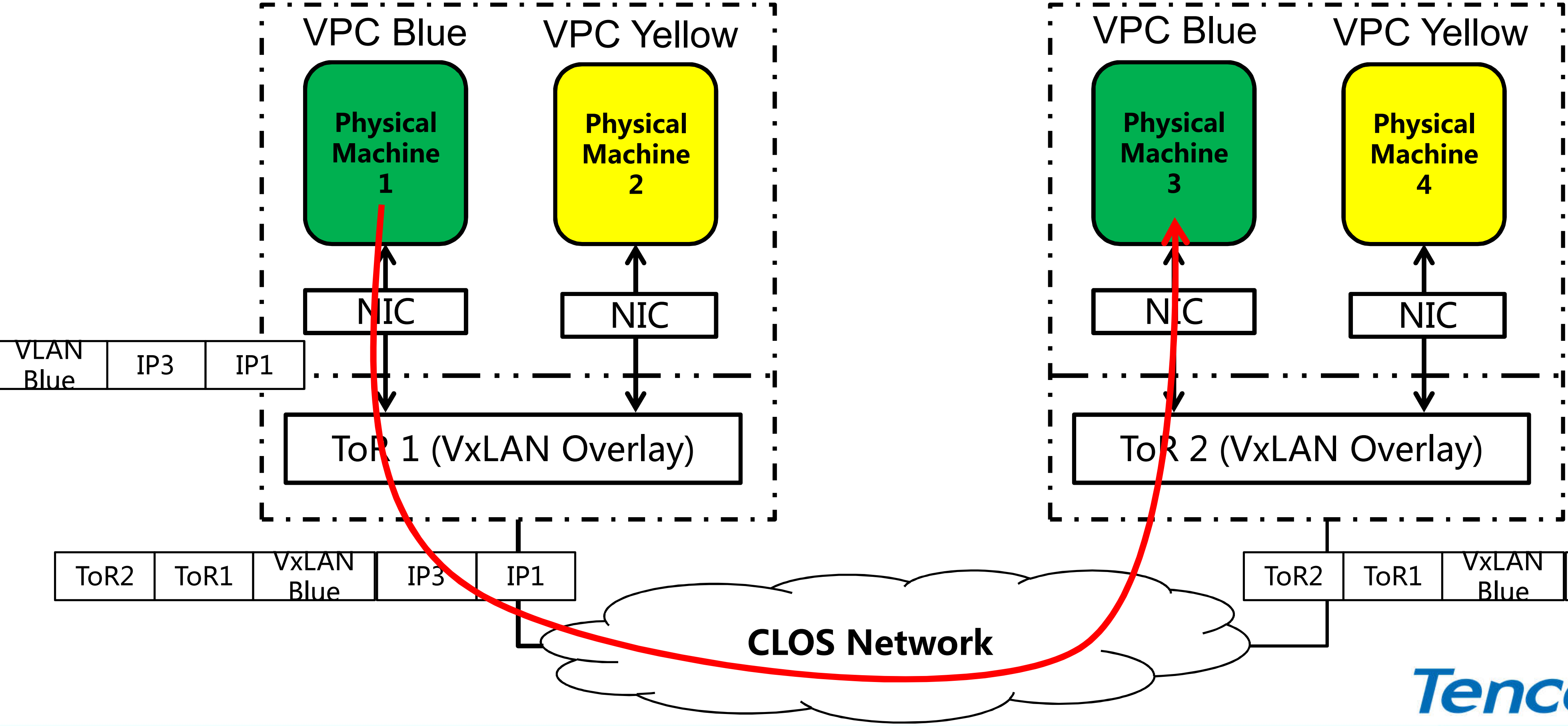
Introduction to Bare Metal Cloud

- **How** to implement BM Cloud: ToR based Virtualization
  - Requirements: any server from anywhere to any customer, BYO IP addresses



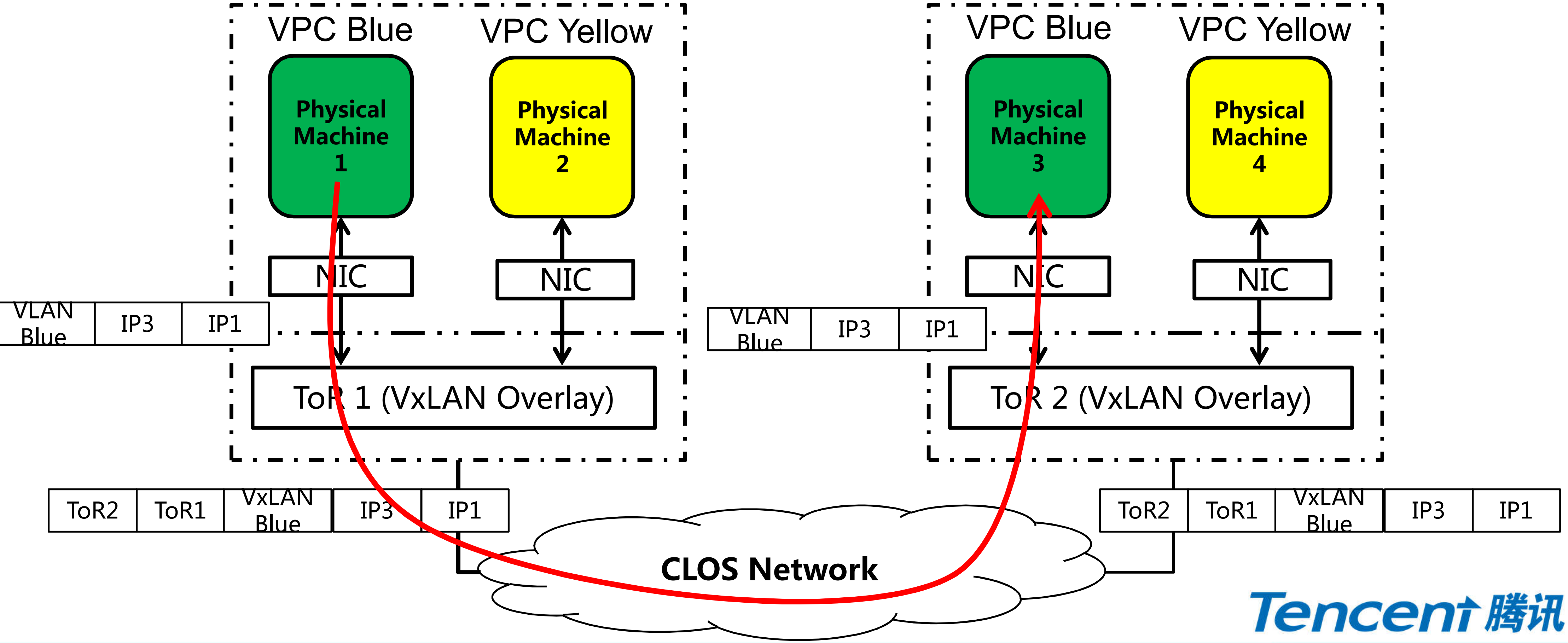
Introduction to Bare Metal Cloud

- **How** to implement BM Cloud: ToR based Virtualization
  - Requirements: any server from anywhere to any customer, BYO IP addresses



Introduction to Bare Metal Cloud

- **How** to implement BM Cloud: ToR based Virtualization
  - Requirements: any server from anywhere to any customer, BYO IP addresses



VLAN Blue	IP3	IP1

ToR2	ToR1	VxLAN Blue	IP3	IP1

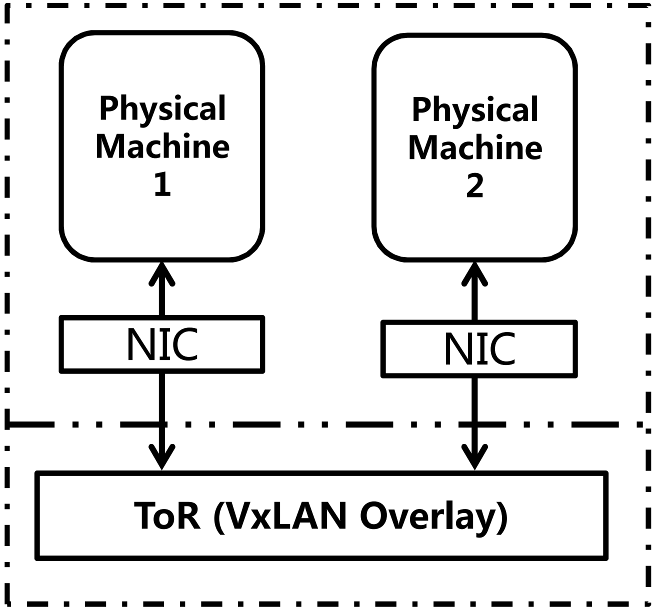
VLAN Blue	IP3	IP1

ToR2	ToR1	VxLAN Blue	IP3	IP1

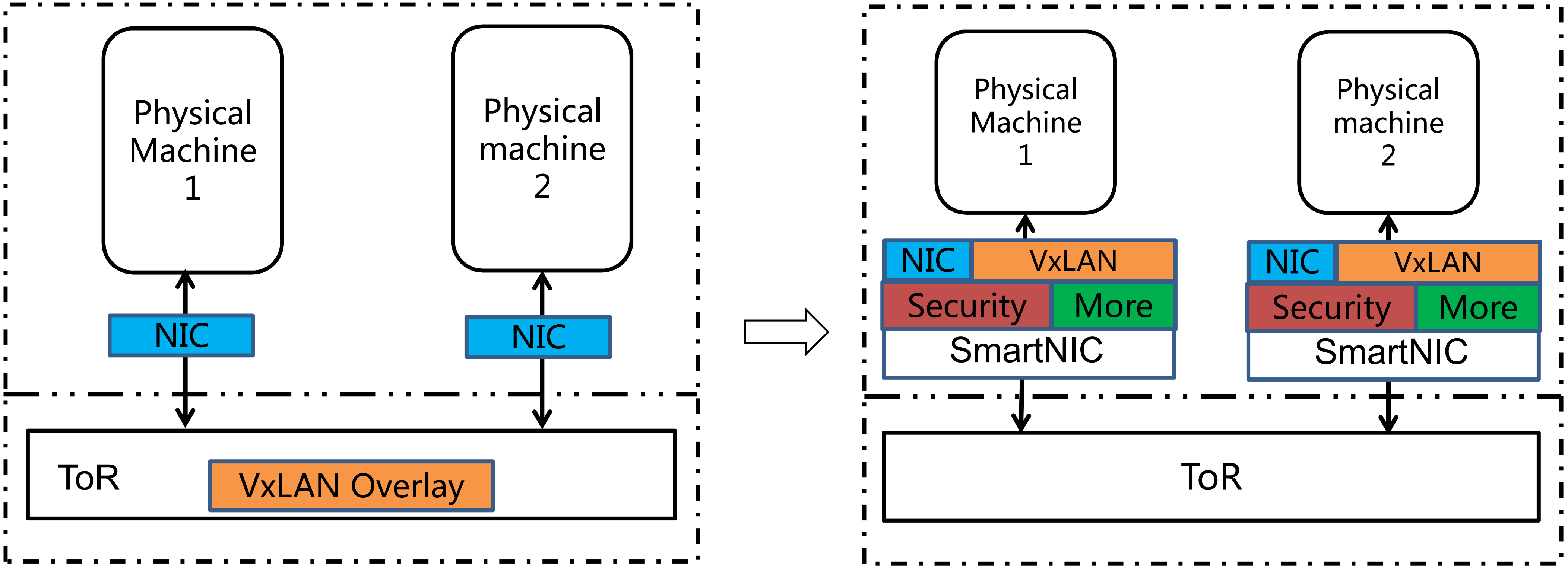
Challenges in Bare Metal Cloud

- **Scalability**
  - ToR switch table size is limited
    - 32-bit host routing table, VxLAN tunnel table
  - VPC network size is limited

- **Flexibility**
  - ToR switch limited programmability
  - Unable to support security group and more



SmartNIC in Bare Metal Cloud



1. ToR based Virtualization

2. SmartNIC based Virtualization

Challenges:

- **Scalability** : limited switch table size
- **Flexibility**: unable to support security group

Solutions:

- **Scalability:** ToR (centralized) -> multiple SmartNICs (distributed)
- **Flexibility:** Programmable chips (ARM & FPGA) to support advanced features (security group, network ACL, QoS…)

Agenda

1 SmartNIC in Bare Metal Cloud

2 SmartNIC in Public Cloud

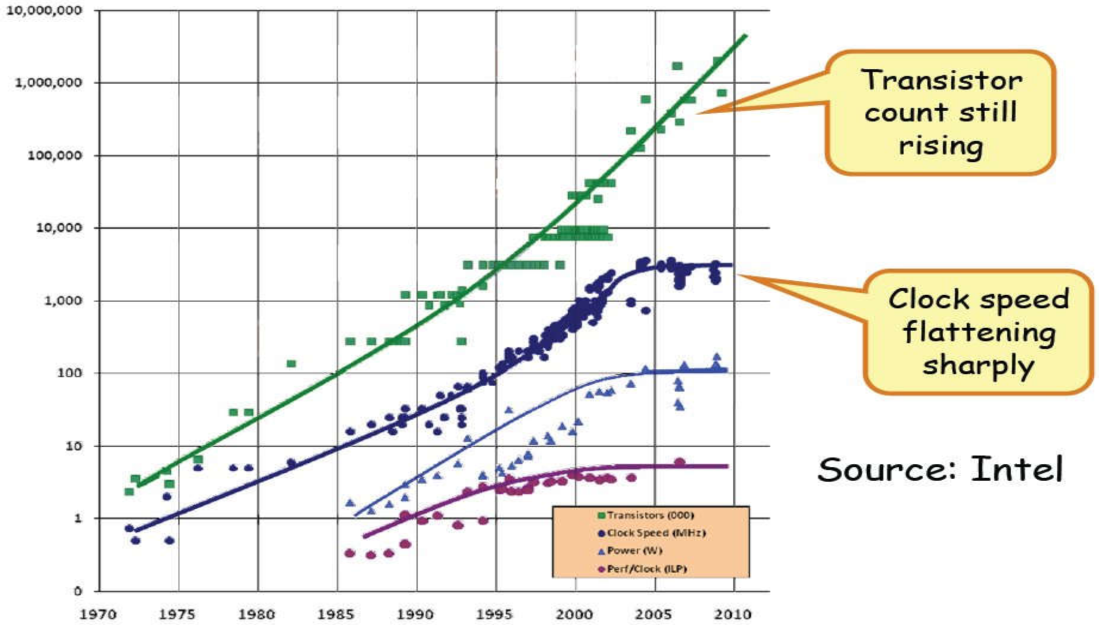
3 Converged SmartNIC Architecture

4 Tencent SmartNIC Experience

5 Future Challenges

Why SmartNIC in Public Cloud?

- **Performance Perspective**
  - Slow increase of CPU performance: double every 2 years, but not much longer
  - Fast increase of network speed (1G -> 50G) & host SDN policies



Transistor count still rising

Clock speed flattening sharply

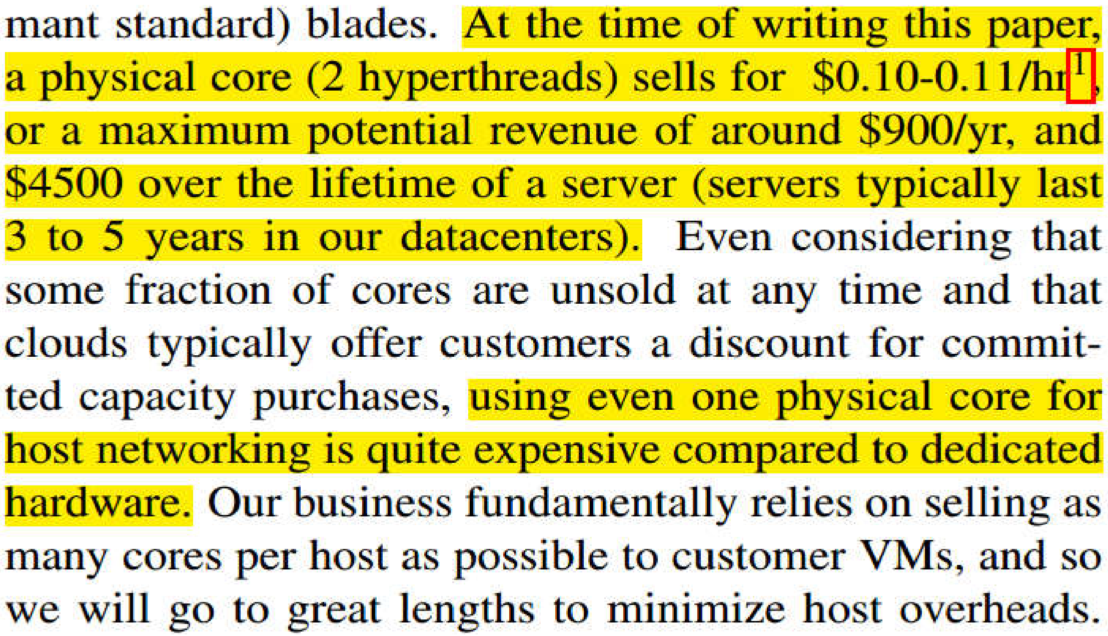
Source: Intel

Source: https://bertrandmeyer.com/2011/06/20/concurrent-programming-is-easy/intel/

- Specialization (HW acceleration) for efficiency (perf per watt)

Why SmartNIC in Public Cloud?

- **Revenue Perspective**
  - SmartNIC increases the NIC cost a bit
  - But the CPU savings/revenue increase could be very significant

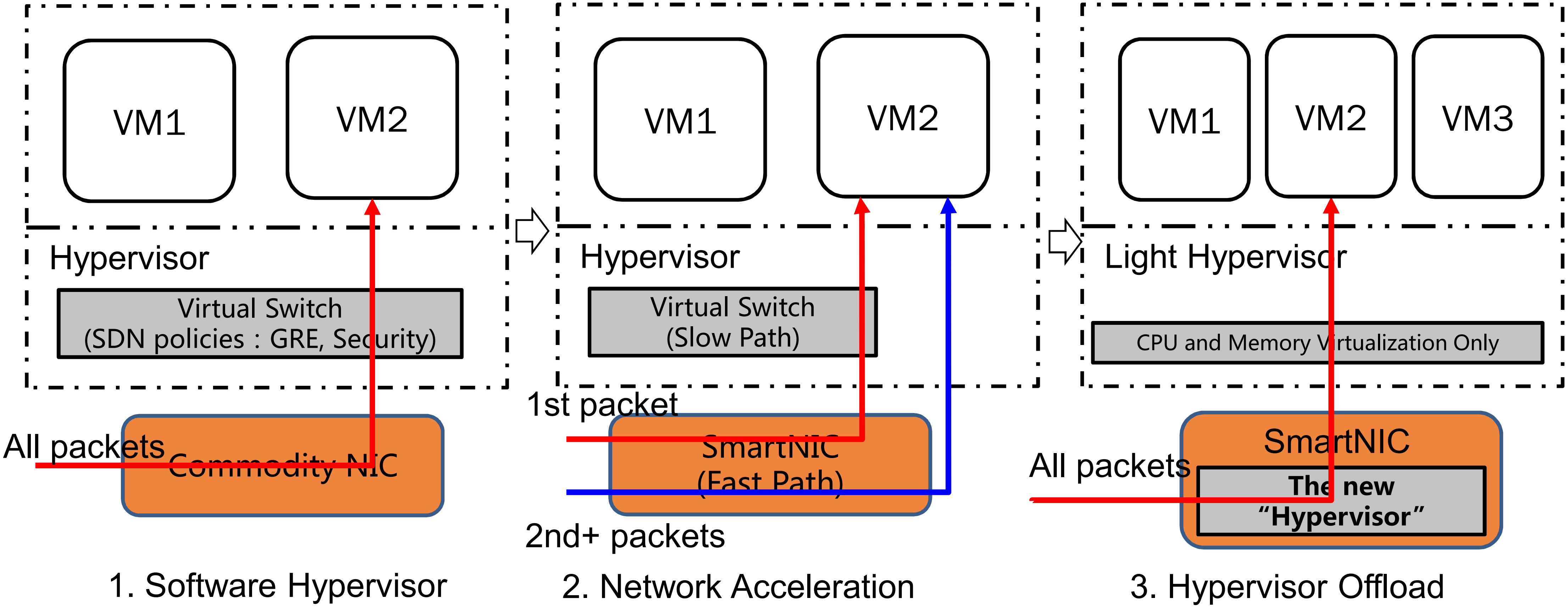


mant standard) blades. At the time of writing this paper, a physical core (2 hyperthreads) sells for $0.10-0.11/hr[1], or a maximum potential revenue of around $900/yr, and $4500 over the lifetime of a server (servers typically last 3 to 5 years in our datacenters). Even considering that some fraction of cores are unsold at any time and that clouds typically offer customers a discount for committed capacity purchases, using even one physical core for host networking is quite expensive compared to dedicated hardware. Our business fundamentally relies on selling as many cores per host as possible to customer VMs, and so we will go to great lengths to minimize host overheads.

Azure SmartNIC, NSDI 2018

- Maximize CPU savings by offloading infra workloads to SmartNIC

SmartNIC Evolution in Public Cloud



Performance Boost

CPU Savings/Revenue Increase

Push Performance Boost and CPU Savings to the limit!

Agenda

1 SmartNIC in Bare Metal Cloud

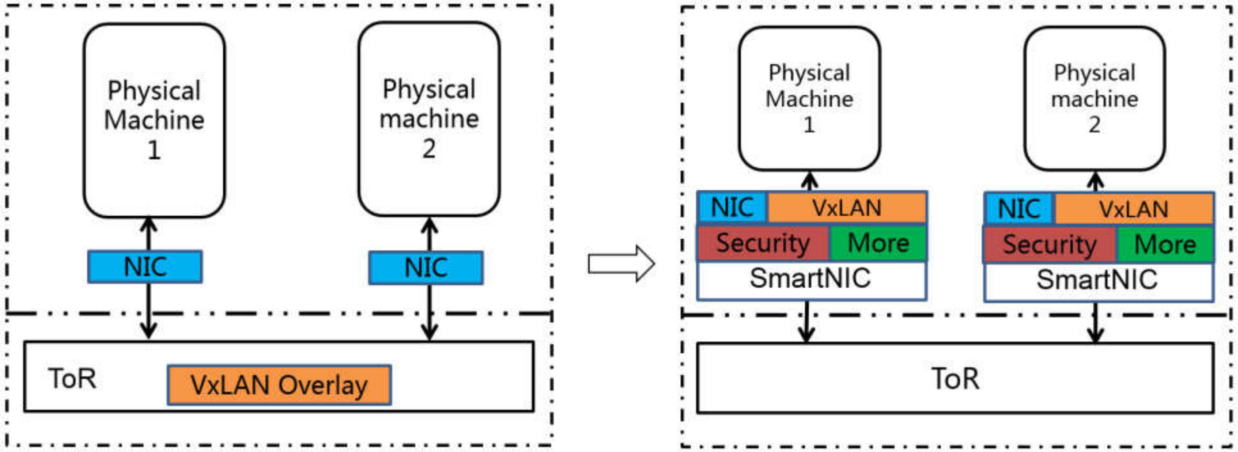
2 SmartNIC in Public Cloud

3 Converged SmartNIC Architecture

4 Tencent SmartNIC Experience

5 Future Challenges

Converged SmartNIC for Bare Metal and Public Cloud

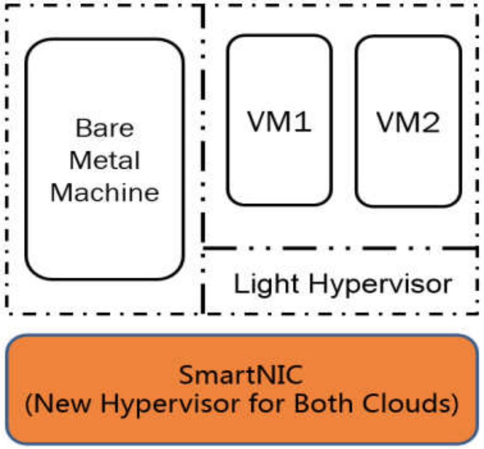


1. ToR based Virtualization

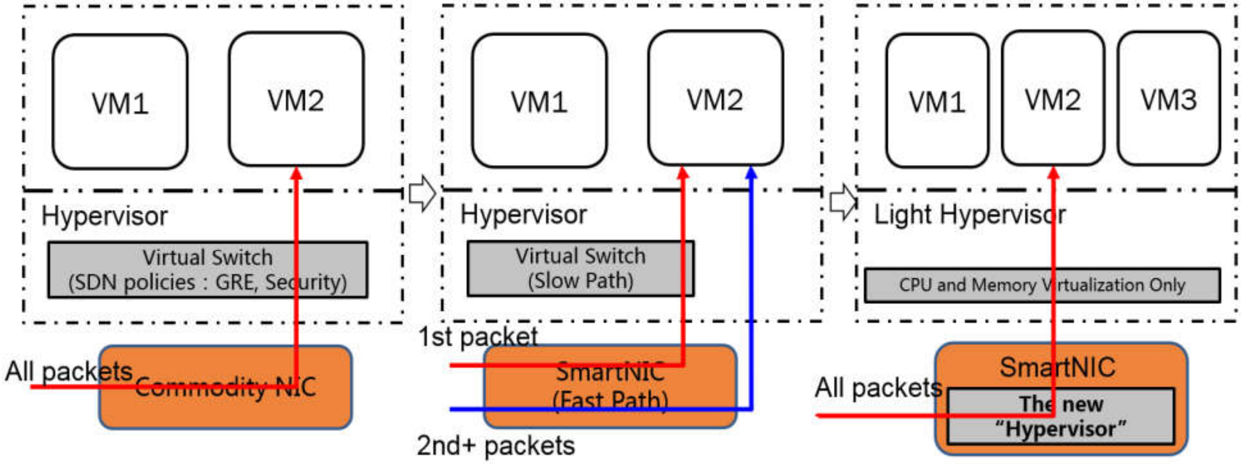
2. SmartNIC based Virtualization

SmartNIC Evolution in BareMetal Cloud

Convergence



Converged “Hypervisor” in SmartNIC

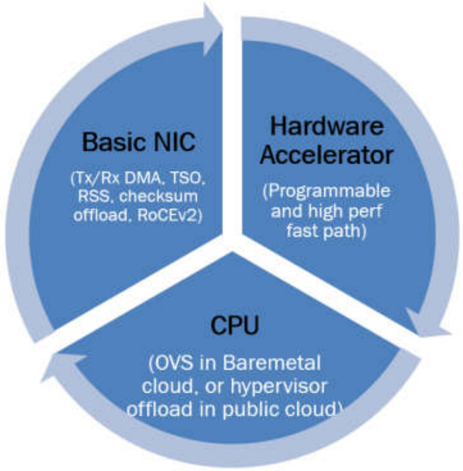


1. Software Hypervisor

2. Network Acceleration

3. Hypervisor Offload

SmartNIC Evolution in Public Cloud



Converged SmartNIC Platform

Agenda

1 SmartNIC in Bare Metal Cloud

2 SmartNIC in Public Cloud

3 Converged SmartNIC Architecture

4 Tencent SmartNIC Experience

5 Future Challenges

Tencent SmartNIC Experience

- **Hardware Selection: SoC vs. discreate chips, FPGA vs. ASIC/NP/ARM**
  - No simple right answer
  - Requirements and constraints vary in different companies at different time: time to market, feature set, requirement stability, chip availability, cost, power …

- **Agility: Tencent Speed**
  - Build a SmartNIC team (~10) in less than a year
  - Finish FPGA pipeline in 3 months (FPGA hard to program? Yes and No)
  - Build a SmartNIC board in 4 months, in just one iteration
  - Ship a SW-HW co-design project (from planning to deployment) in about 1 year

Agenda

1 SmartNIC in Bare Metal Cloud

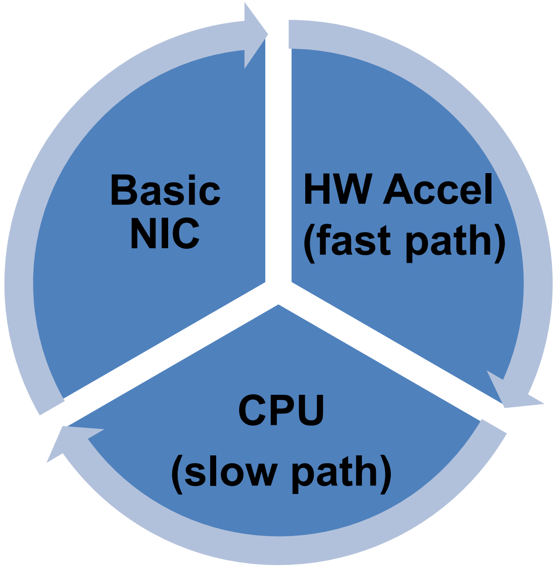
2 SmartNIC in Public Cloud

3 Converged SmartNIC Architecture

4 Tencent SmartNIC Experience

5 Future Challenges

Future Challenges on Hardware



Future Challenges on Hardware

Power, area and cost challenges

Future Challenges on Hardware

SoC
(all in one)

Power, area and cost challenges

Future Challenges on Hardware

Partner 1

🟩	HW Accel: FPGA
🟨	ARM CPU
🟨	Basic NIC (RoCEv2?)

Partner 2

🟨	HW Accel: ASIC (Programmability?)
🟩	ARM CPU
🟨	Basic NIC (RoCEv2?)

SoC
(all in one)

Partner 3

🟩	HW Accel: FPGA
🟨	ARM CPU
🟥	Basic NIC (?)

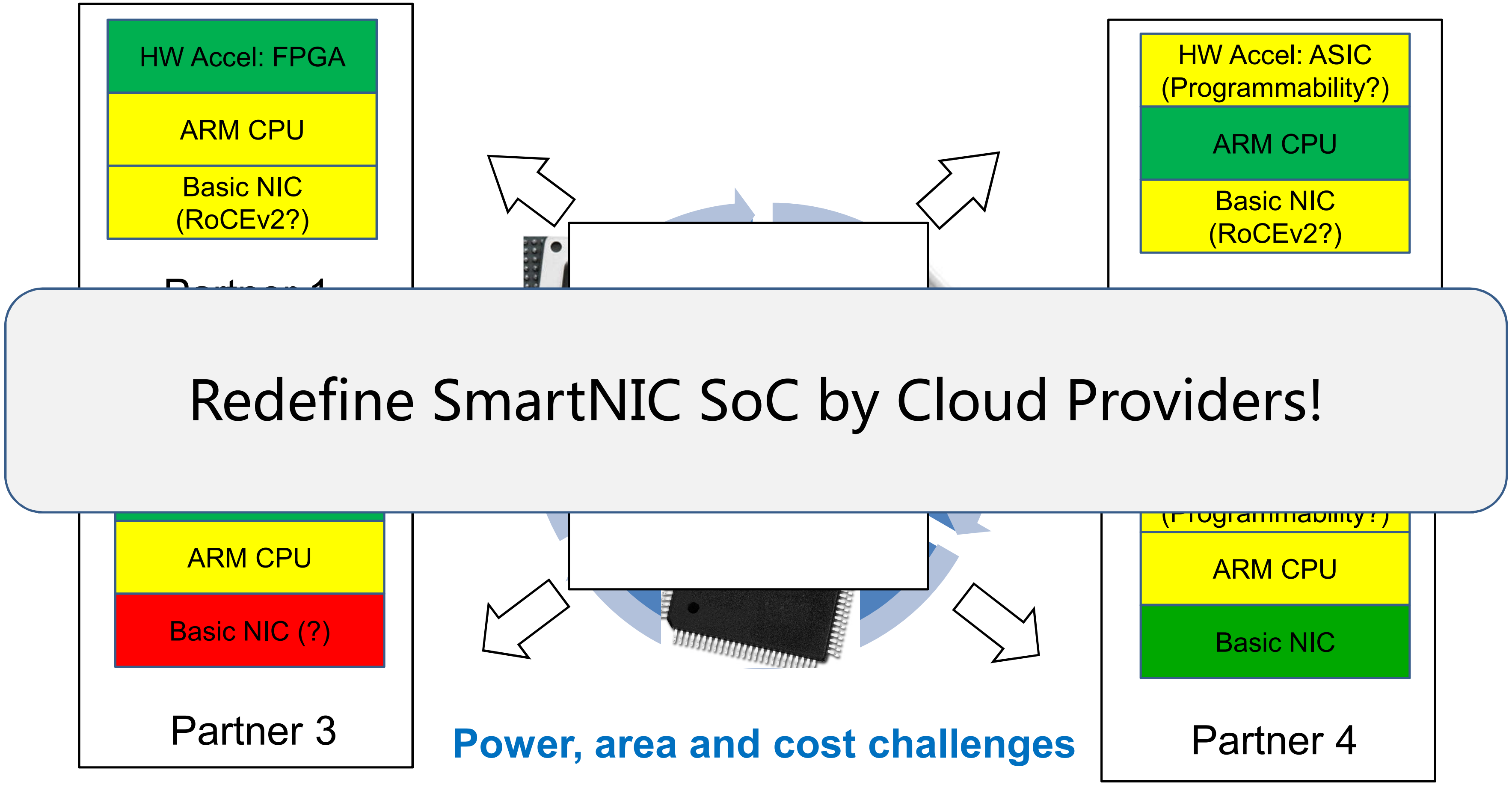
Power, area and cost challenges

Partner 4

🟨	HW Accel: ASIC (Programmability?)
🟨	ARM CPU
🟩	Basic NIC

🟩 Ready 🟨 Partial Ready 🟥 Not Ready

Future Challenges on Hardware



Redefine SmartNIC SoC by Cloud Providers!

Power, area and cost challenges

- 🟩 Ready
- 🟨 Partial Ready
- 🟥 Not Ready

Future Challenges on Architecture

- Task partition on heterogenous platform
  - Architectural boundaries between x86, FPGA and ARM for different workloads: host SDN, storage and NFV (IPSec VPN, LB, etc.)
- Hitless upgrade and reboot
  - Collaborative process between x86, FPGA and ARM
- Live migration with hypervisor offload
  - How to log dirty page if hypervisor is totally bypassed?

Thanks!