nsdi '24

LuoShen: A Hyper-Converged Programmable Gateway for Multi-Tenant Multi-Service Edge Clouds

Tian Pan, Kun Liu, Xionglie Wei, Yisong Qiao, Jun Hu, Zhiguo Li, Jun Liang, Tiesheng Cheng, Wenqiang Su,

Jie Lu, Yuke Hong, Zhengzhong Wang, Zhi Xu, Chongjing Dai, Peiqiao Wang, Xuetao Jia, Jianyuan Lu,

Enge Song, Jun Zeng, Biao Lyu, Ennan Zhai, Jiao Zhang, Tao Huang, Dennis Cai, and Shunmin Zhu.

Presenter: Yisong Qiao

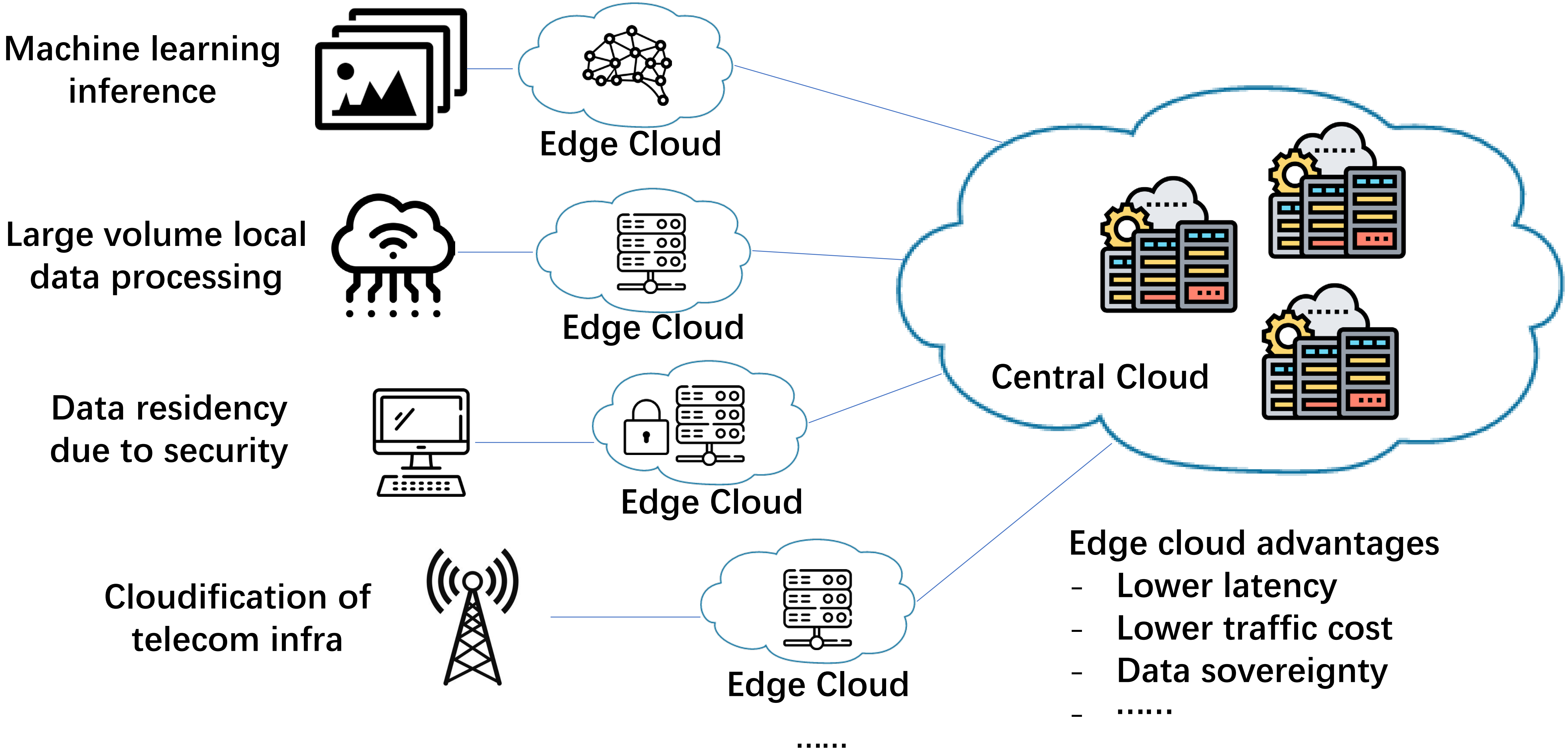
Alibaba Cloud

浙江大學
ZHEJIANG UNIVERSITY

紫金山实验室
Purple Mountain Laboratories

清華大學
Tsinghua University

Background: Extending Public Cloud to the Edge



Machine learning inference

Edge Cloud

Large volume local data processing

Edge Cloud

Data residency due to security

Edge Cloud

Cloudification of telecom infra

Edge Cloud

......

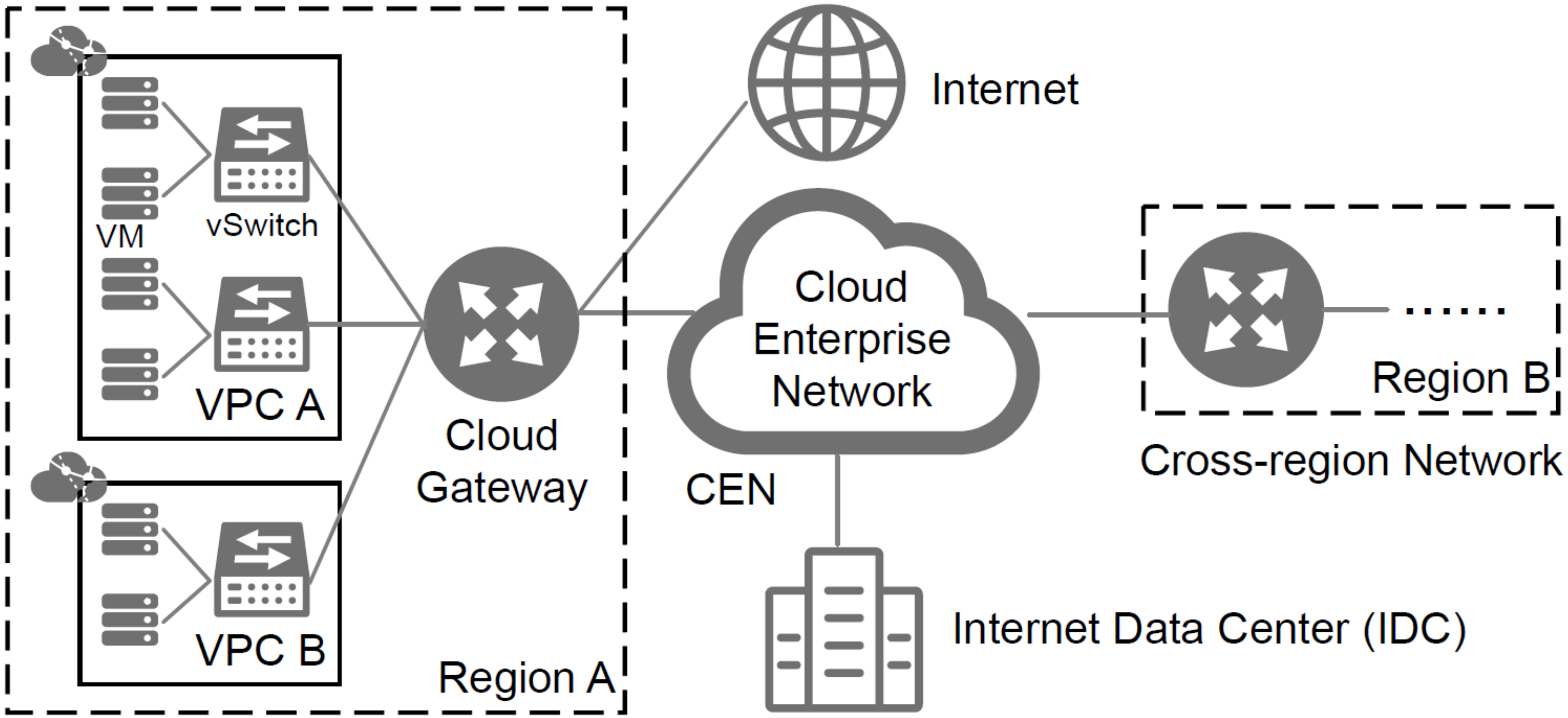
Central Cloud

Edge cloud advantages

- Lower latency
- Lower traffic cost
- Data sovereignty
- ......

Background: VPC Network Infra in Alibaba Cloud

——Networking requirements in the public cloud



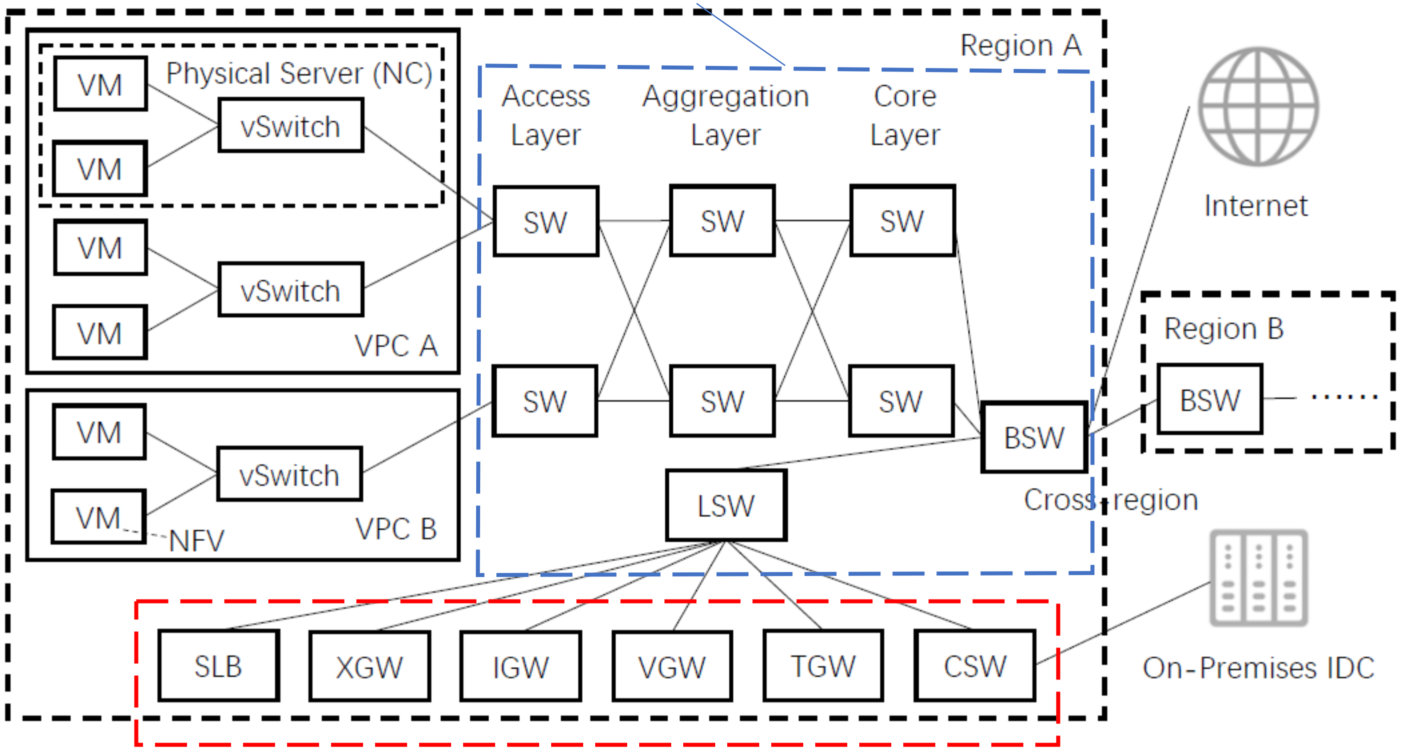
Cloud Services
VM-VM (same VPC)
VM-VM (different VPCs)
VM-Cross-region-VM
VM-IDC
IDC-VM
VM-Internet
Internet-VM
Internet-LB-Service
VM-LB-Service

- ✓ A tenant's VMs communicate with each other, and they may reside in the same VPC, in different VPCs of the same region, or in different VPCs across the globe.
- ✓ A tenant's VMs may also need to communicate with IDCs and Internet.
- ✓ To handle the growth of cloud traffic, either from Internet or from within the cloud, horizontal scaling of VMs is needed and load balancers are required.

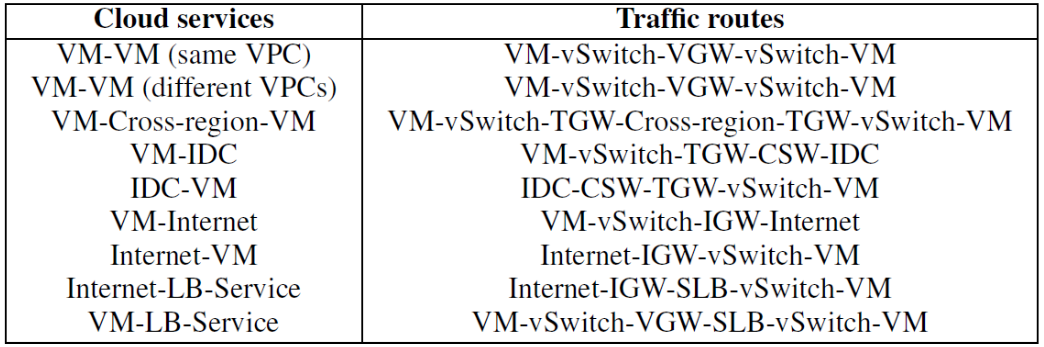
Background: VPC Network Infra in Alibaba Cloud

——Designs for scalable VPC networking

Underlay devices



“Role-splitting” gateway architecture



Cloud services	Traffic routes
VM-VM (same VPC)	VM-vSwitch-VGW-vSwitch-VM
VM-VM (different VPCs)	VM-vSwitch-VGW-vSwitch-VM
VM-Cross-region-VM	VM-vSwitch-TGW-Cross-region-TGW-vSwitch-VM
VM-IDC	VM-vSwitch-TGW-CSW-IDC
IDC-VM	IDC-CSW-TGW-vSwitch-VM
VM-Internet	VM-vSwitch-IGW-Internet
Internet-VM	Internet-IGW-vSwitch-VM
Internet-LB-Service	Internet-IGW-SLB-vSwitch-VM
VM-LB-Service	VM-vSwitch-VGW-SLB-vSwitch-VM

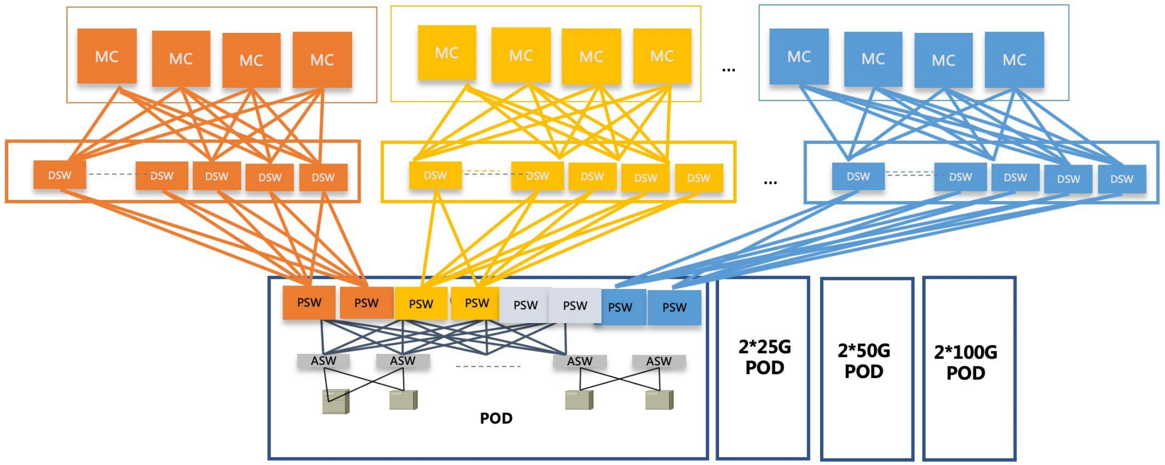
Design 1：Separation of underlay and overlay network devices.

- Rapid cloud service iteration without reconstructing underlay infrastructure

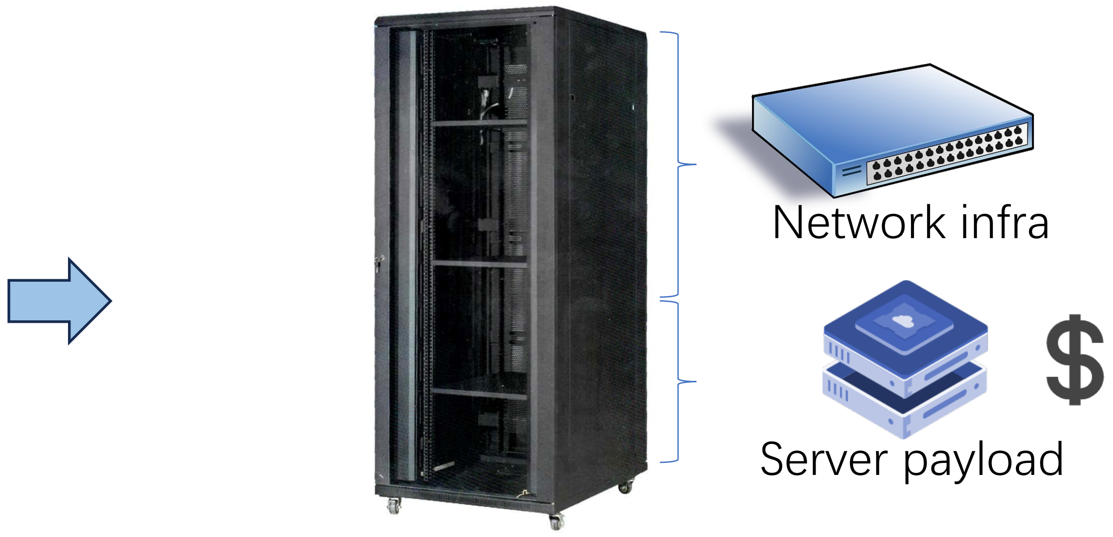
Design 2：Deploying different roles of gateway clusters for different cloud services.

- Horizontal table/traffic splitting among gateway clusters
- Different teams manage different gateways
- Good for failure isolation

Issues of Mirroring VPC Network Infra at the Edge



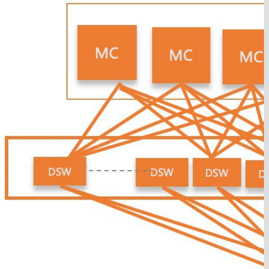
Public cloud infra

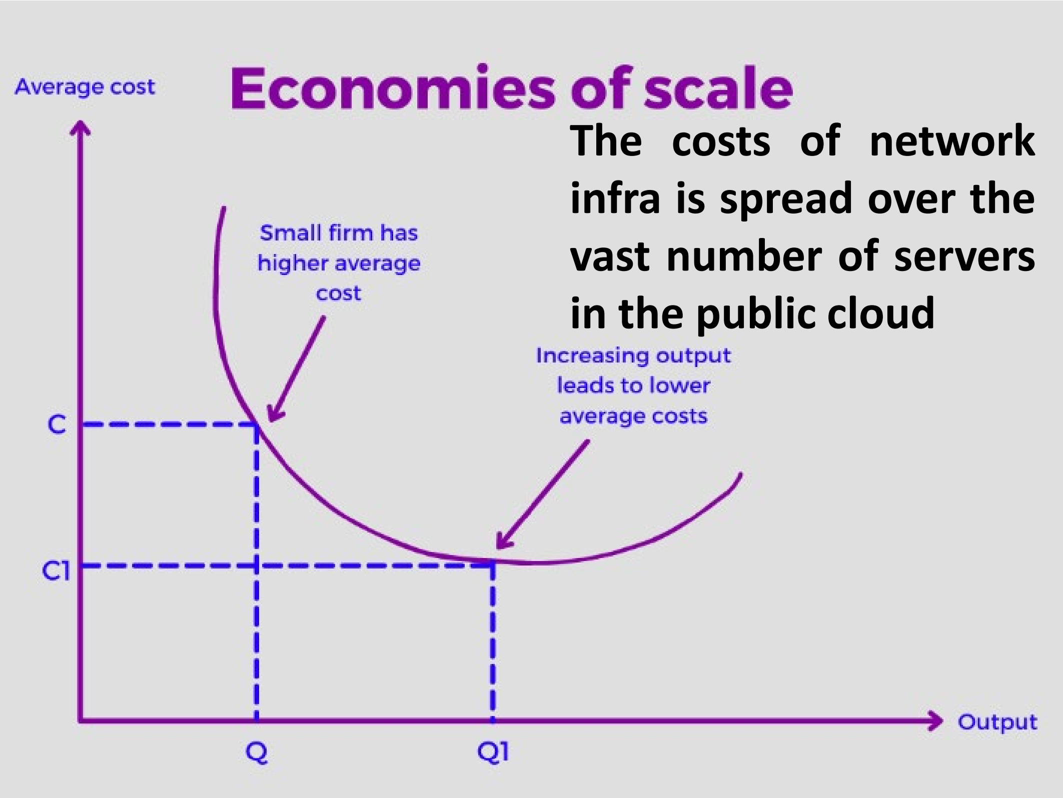


42U server cabinet at the edge

- ✓ How to fit the entire cloud network infrastructure within a constrained space, and leave as much space as possible to server payload which carries VMs for sale?

Issues of Mirroring VPC Network Infra at the Edge





The costs of network infra is spread over the vast number of servers in the public cloud

......

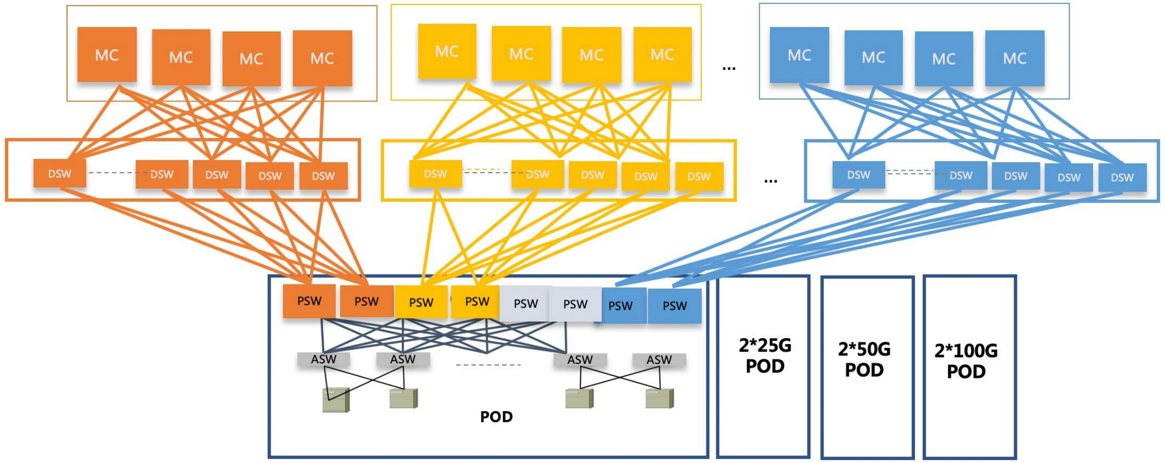
42U server cabinet at

The cost inefficiency will be magnified with increased edge clouds

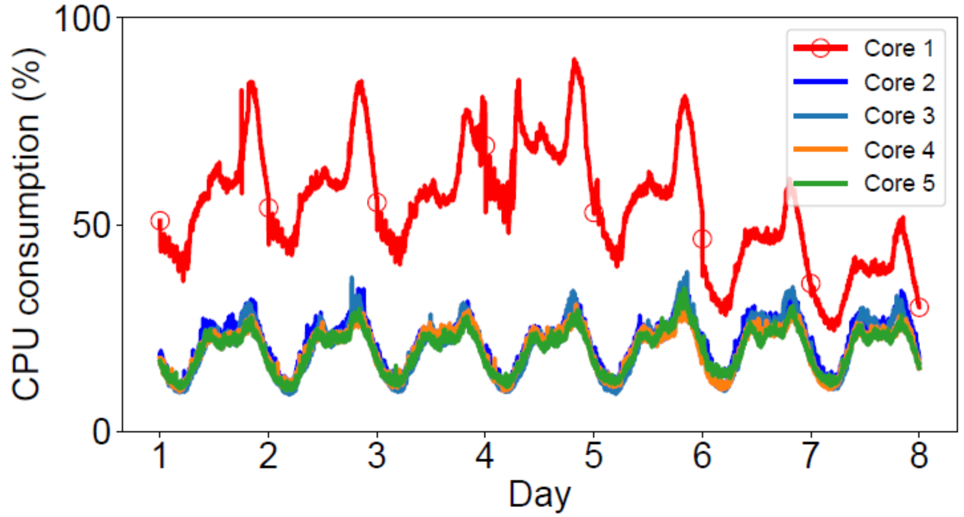
- ✓ Ho ure within a constrained space, and leave as much space as possible to server payload which carries VMs for sale?

- ✓ How to save upfront and operational costs without economies of scale?

Issues of Mirroring VPC Network Infra at the Edge



Public cloud infra



42U server cabinet at the edge

- ✓ How to fit the entire cloud network infrastructure within a constrained space, and leave as much space as possible to server payload which carries VMs for sale?

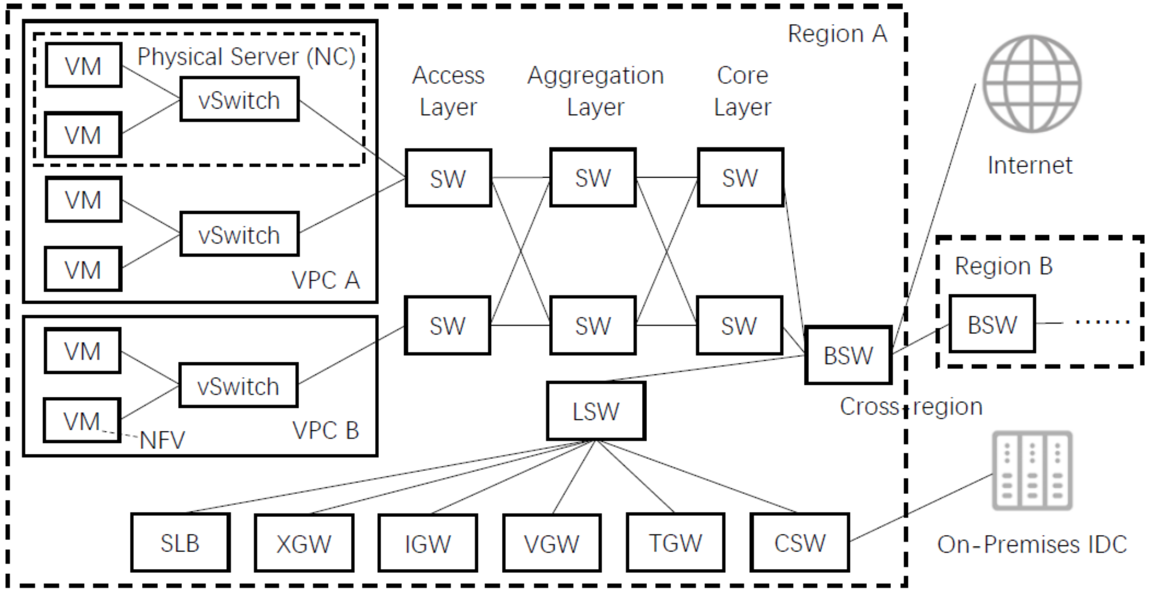
- ✓ How to save upfront and operational costs without economies of scale?

- ✓ How to provide the required stable performance in extreme cases?

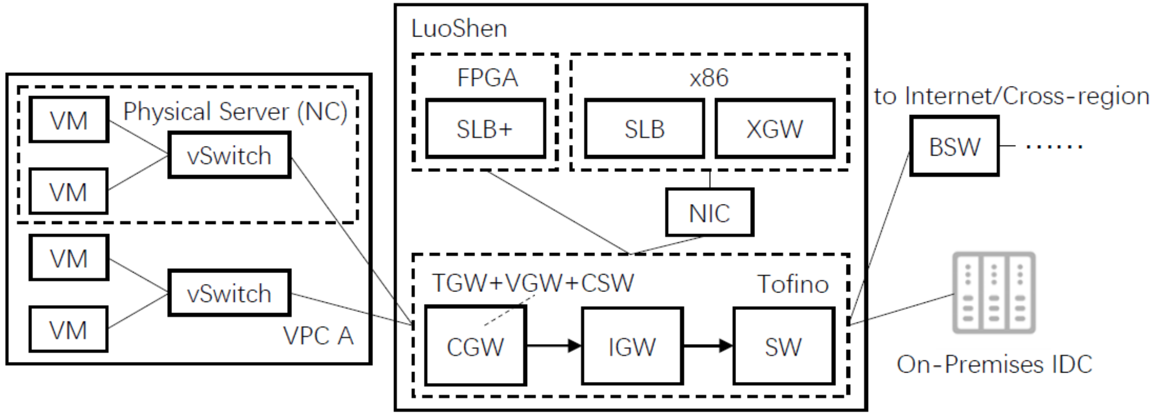
Design Goals and Overview of LuoShen

Design Goals

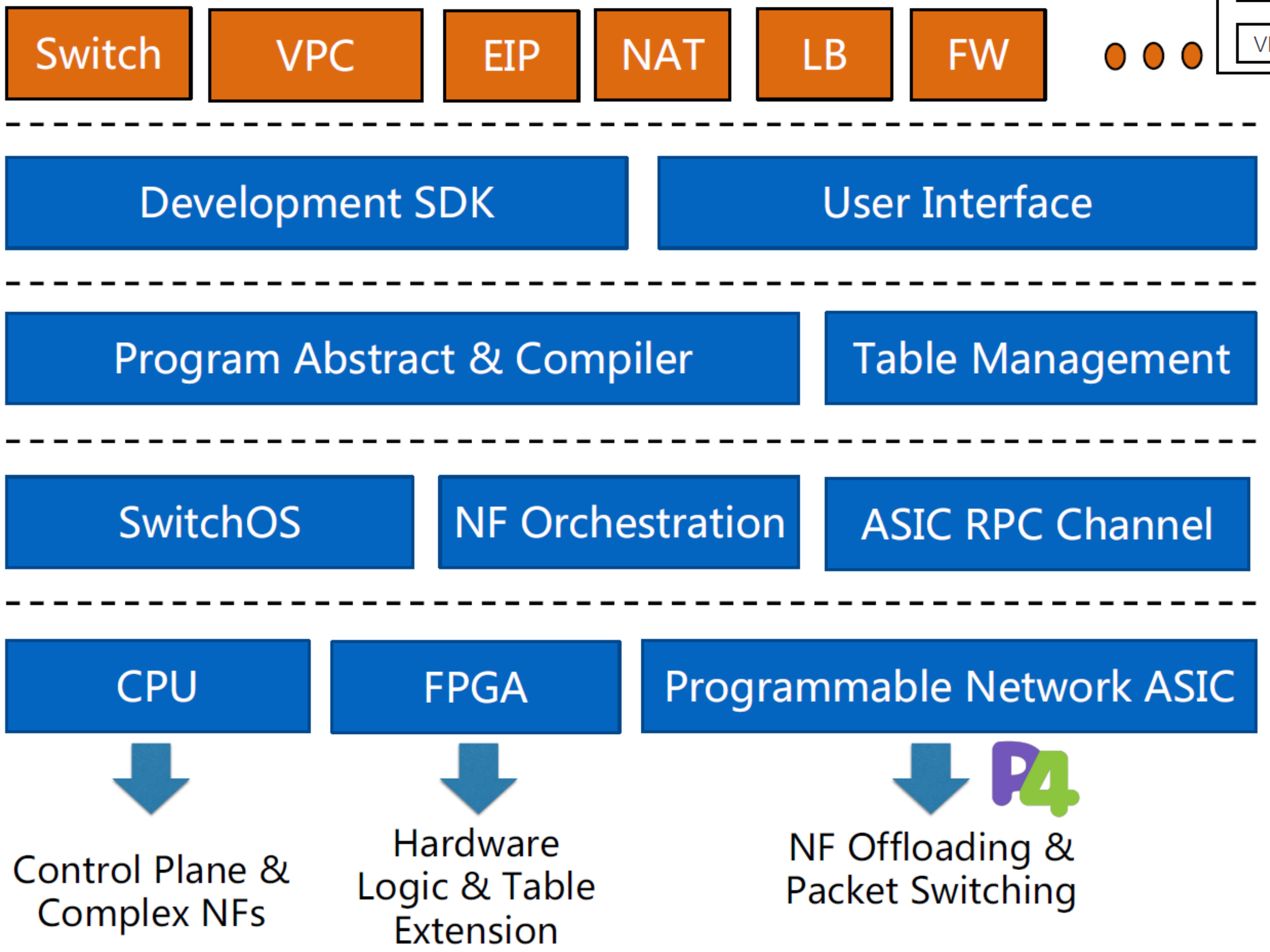
- Small deployment footprints
- Complete VPC network functions
- Cost efficiency
- Performance stability
- Elasticity and flexibility
- Avoid reinventing the wheel



LuoShen fits the entire VPC network infra into a 2U server switch with full functionality retained



The Server-Switch Hardware



Switch | VPC | EIP | NAT | LB | FW | ● ● ●

Development SDK | User Interface

Program Abstract & Compiler | Table Management

SwitchOS | NF Orchestration | ASIC RPC Channel

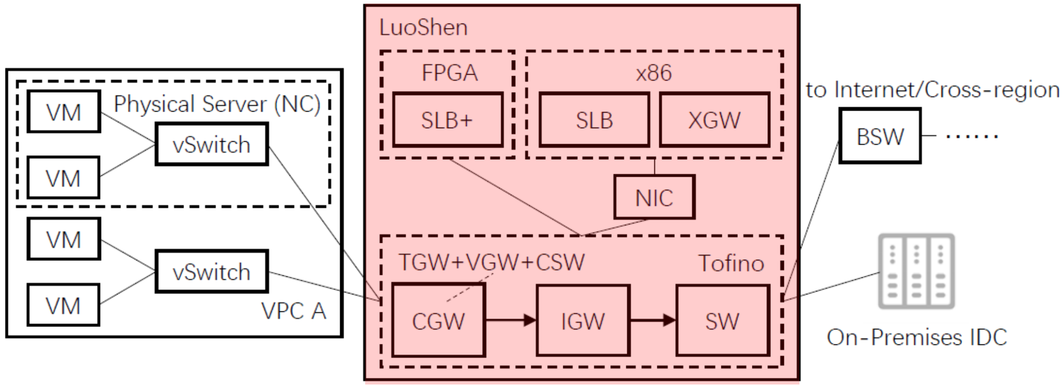
CPU | FPGA | Programmable Network ASIC

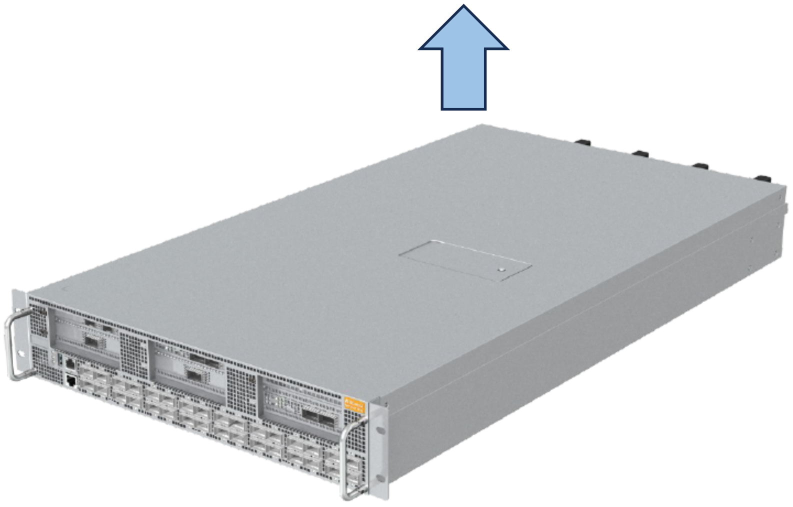
Control Plane & Complex NFs

Hardware Logic & Table Extension

P4

NF Offloading & Packet Switching

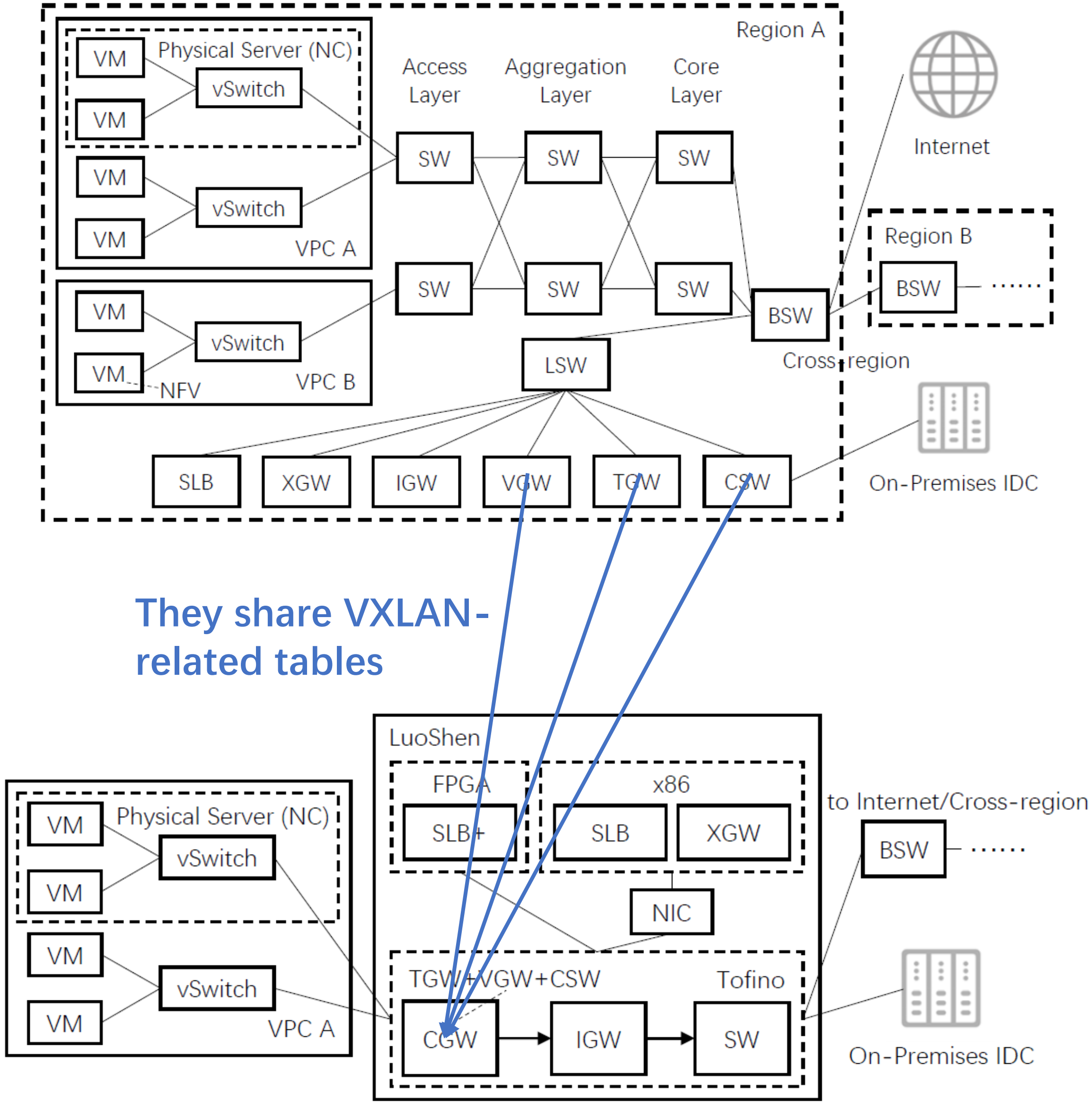




* Smart Network Appliance

Infrastructure Convergence Steps

#1: Converge different gateway functions sharing the same table

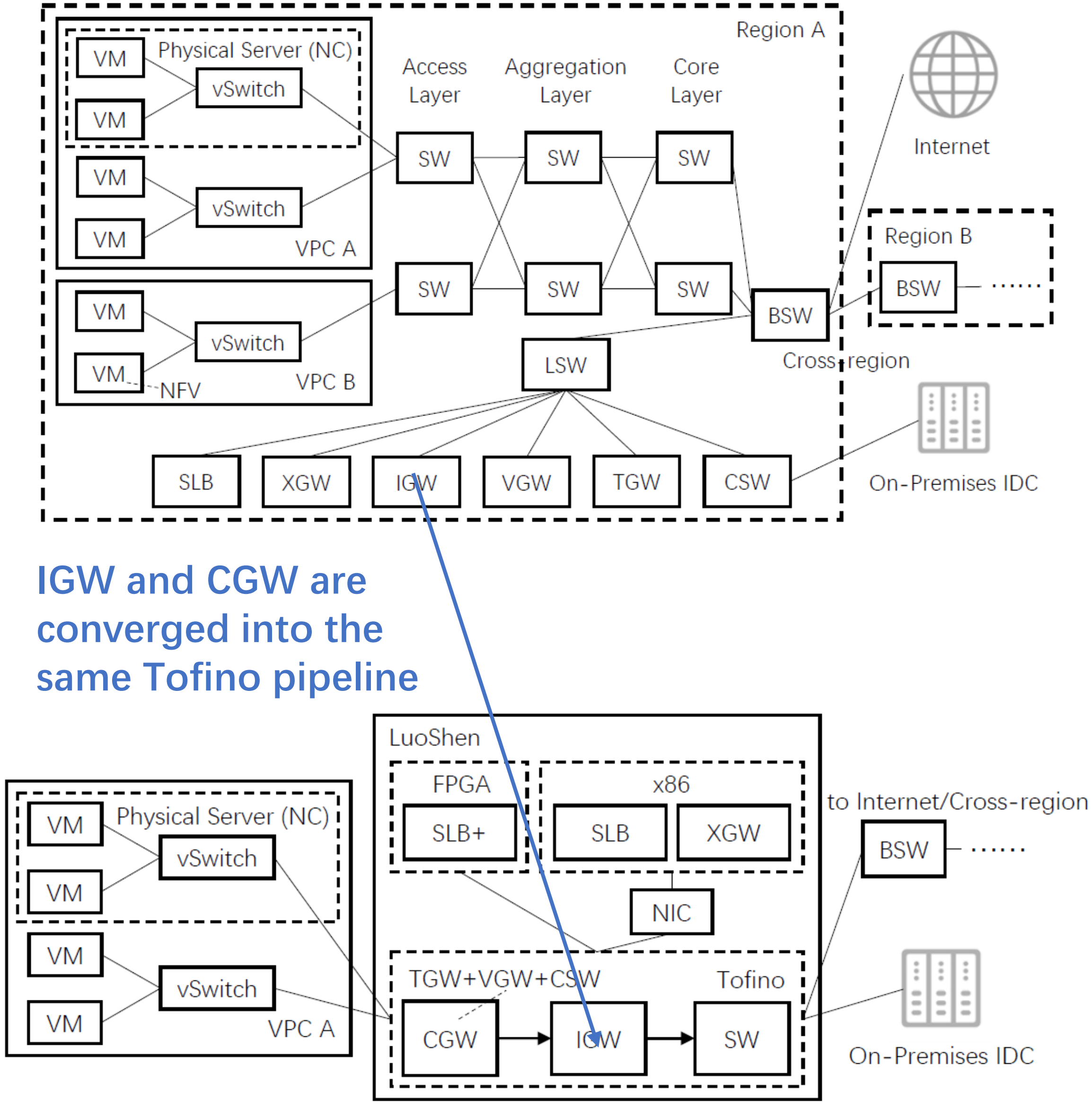


They share VXLAN-related tables

Infrastructure Convergence Steps

#1: Converge different gateway functions sharing the same table

#2: Converge different gateway functions without table overlapping



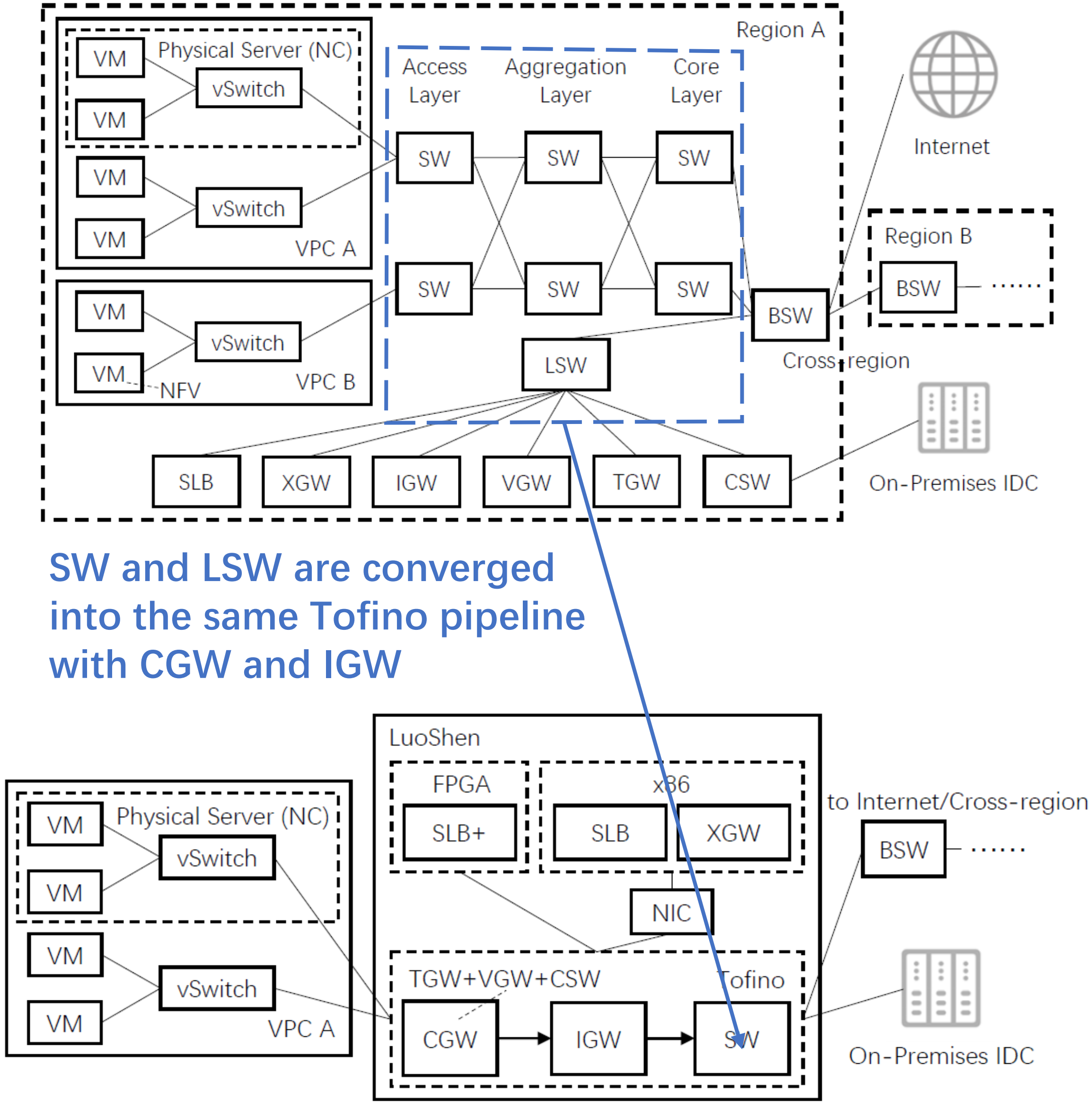
IGW and CGW are converged into the same Tofino pipeline

Infrastructure Convergence Steps

#1: Converge different gateway functions sharing the same table

#2: Converge different gateway functions without table overlapping

#3: Converge underlay and overlay devices



SW and LSW are converged into the same Tofino pipeline with CGW and IGW

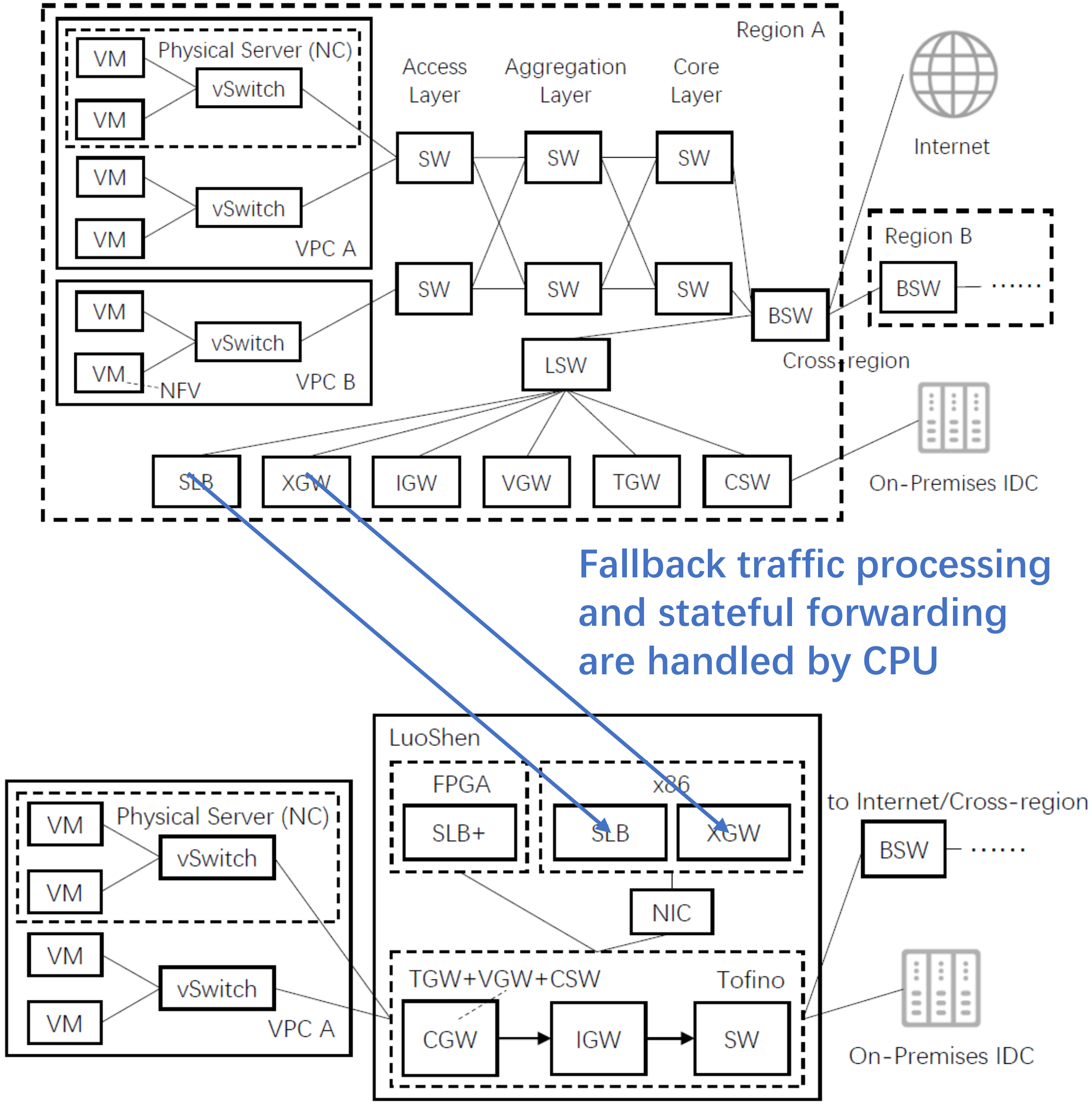
Infrastructure Convergence Steps

#1: Converge different gateway functions sharing the same table

#2: Converge different gateway functions without table overlapping

#3: Converge underlay and overlay devices

#4: Process fallback traffic and stateful forwarding at the CPU



Fallback traffic processing and stateful forwarding are handled by CPU

Infrastructure Convergence Steps

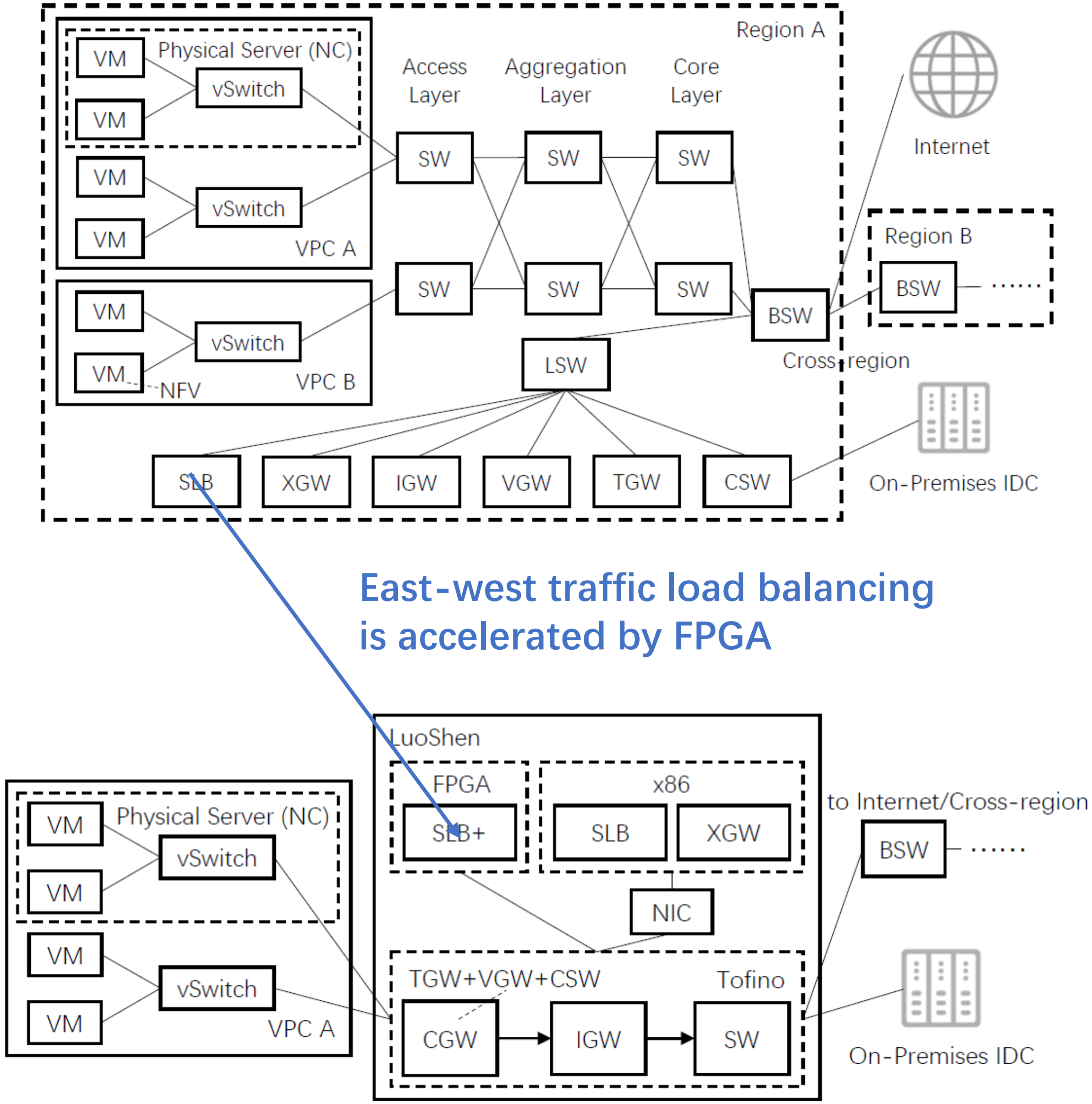
#1: Converge different gateway functions sharing the same table

#2: Converge different gateway functions without table overlapping

#3: Converge underlay and overlay devices

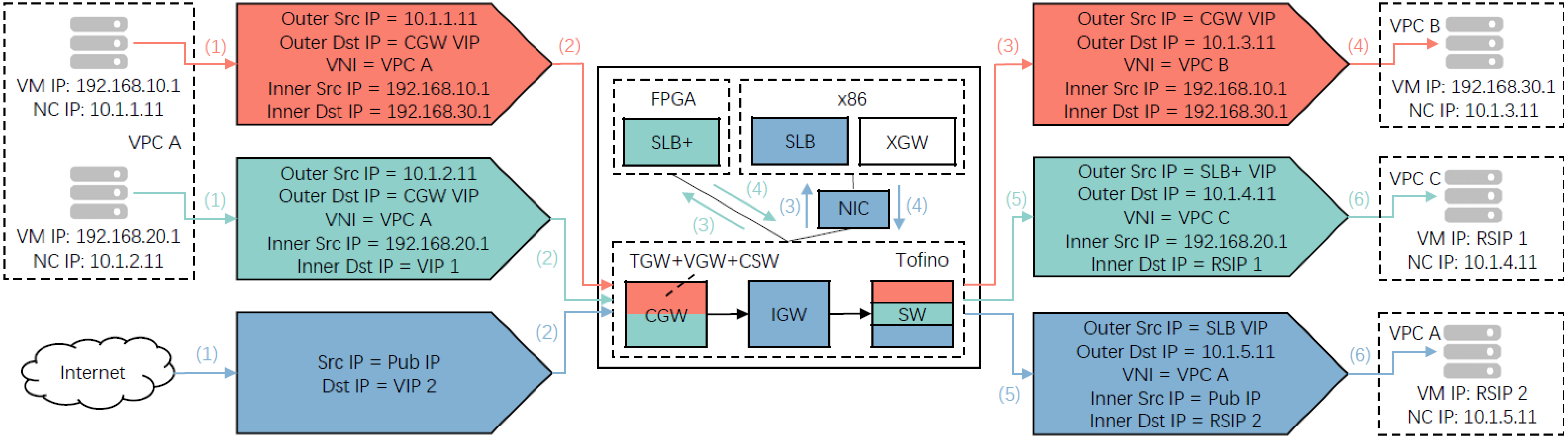
#4: Process fallback traffic and stateful forwarding at the CPU

#5: Offload high-bandwidth stateful forwarding to the FPGA



East-west traffic load balancing is accelerated by FPGA

Packet Journeys in LuoShen



VM IP: 192.168.10.1
NC IP: 10.1.1.11

VPC A

VM IP: 192.168.20.1
NC IP: 10.1.2.11

Internet

(1)

Outer Src IP = 10.1.1.11
Outer Dst IP = CGW VIP
VNI = VPC A
Inner Src IP = 192.168.10.1
Inner Dst IP = 192.168.30.1

Outer Src IP = 10.1.2.11
Outer Dst IP = CGW VIP
VNI = VPC A
Inner Src IP = 192.168.20.1
Inner Dst IP = VIP 1

Src IP = Pub IP
Dst IP = VIP 2

(2)

FPGA

SLB+

x86

SLB

XGW

NIC

(3) (4)

TGW+VGW+CSW

Tofino

CGW

IGW

SW

(3)

Outer Src IP = CGW VIP
Outer Dst IP = 10.1.3.11
VNI = VPC B
Inner Src IP = 192.168.10.1
Inner Dst IP = 192.168.30.1

(5)

Outer Src IP = SLB+ VIP
Outer Dst IP = 10.1.4.11
VNI = VPC C
Inner Src IP = 192.168.20.1
Inner Dst IP = RSIP 1

Outer Src IP = SLB VIP
Outer Dst IP = 10.1.5.11
VNI = VPC A
Inner Src IP = Pub IP
Inner Dst IP = RSIP 2

(4) (6)

VPC B
VM IP: 192.168.30.1
NC IP: 10.1.3.11

VPC C
VM IP: RSIP 1
NC IP: 10.1.4.11

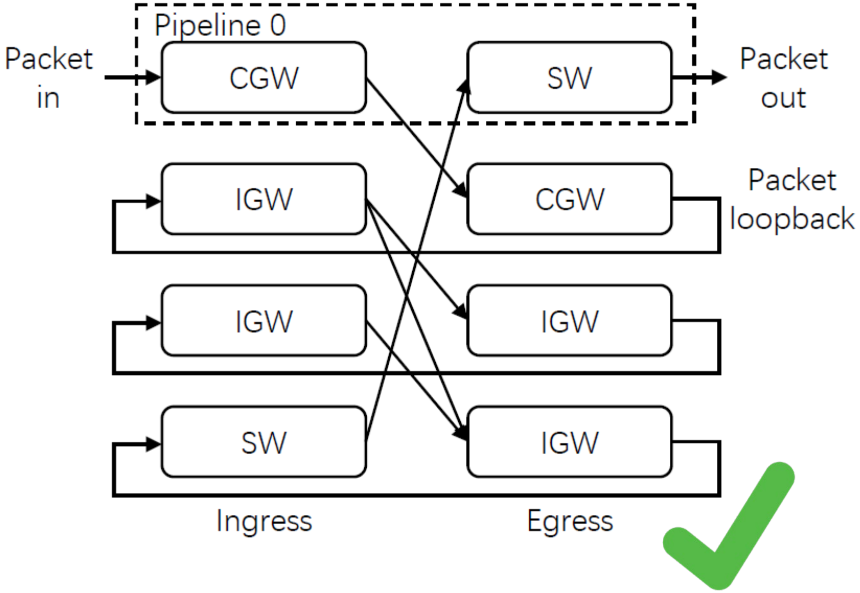
VPC A
VM IP: RSIP 2
NC IP: 10.1.5.11

🟥 VM-VM (different VPCs) 🟩 VM-LB-Service 🟦 Internet-LB-Service

VM-VM (different VPCs): CGW -> SW
VM-LB-Service: CGW -> SW -> SLB+ -> SW
Internet-LB-Service: IGW -> SW -> SLB -> SW

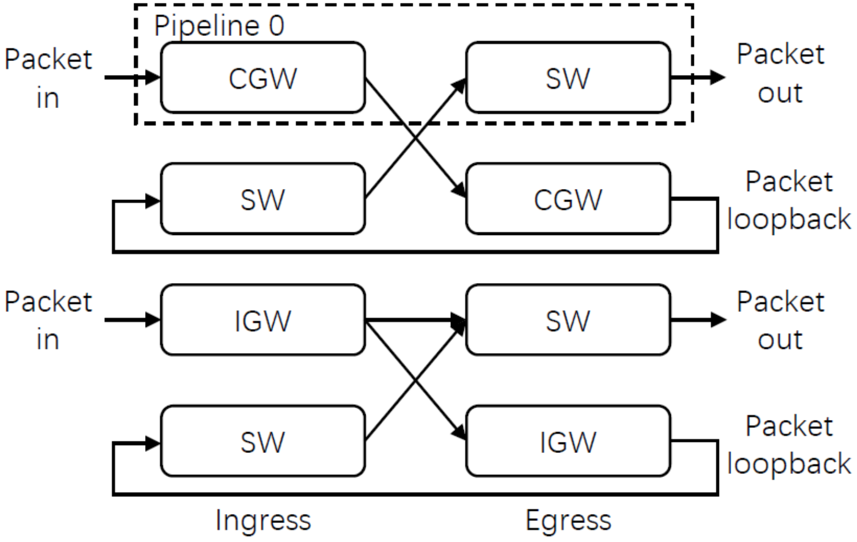
In LuoShen's P4-centric architecture, SW in Tofino is responsible for traffic distribution to CPU/FPGA; While traffic processed by CPU/FPGA needs to be looped back to Tofino for further processing.

Data Plane: Tofino Pipeline Layout



(a) LuoShen’s pipeline layout.

Pros	Cons
Traffic balance	1.2Tbps
Table balance	Less ports exposed
Flexible server attachment	
IGW code reuse	

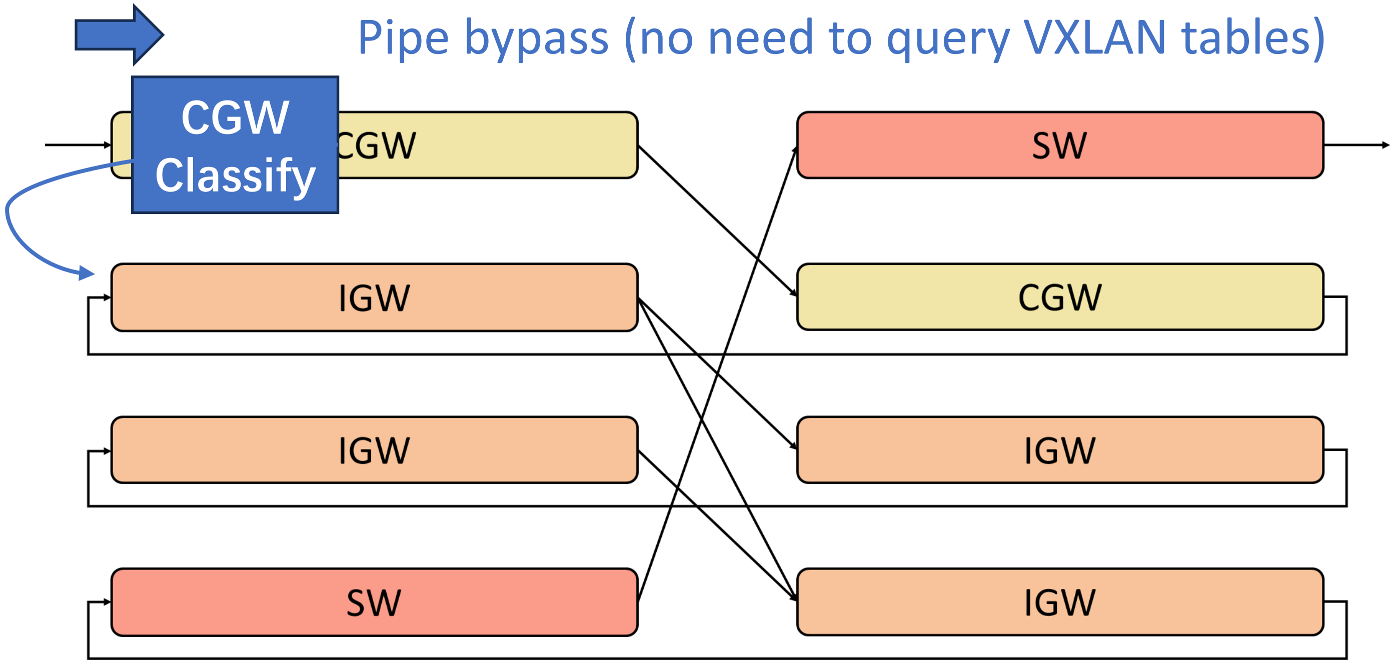


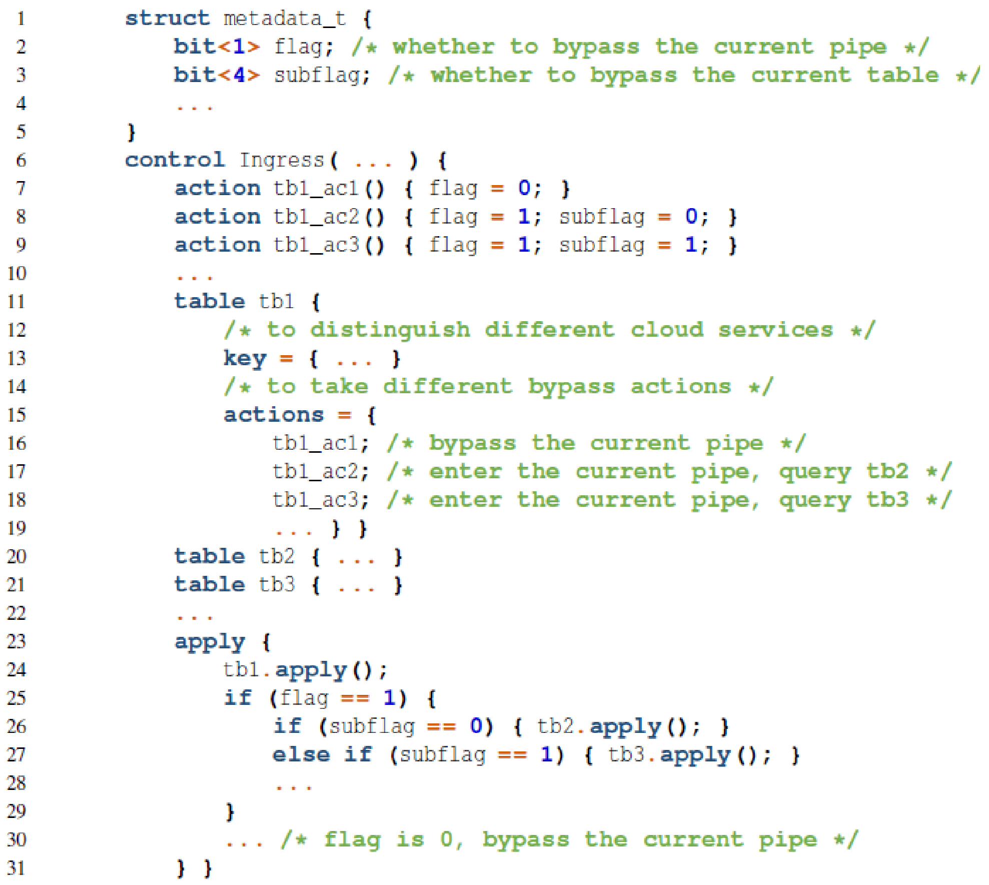
(b) Another layout option.

Pros	Cons
3.2Tbps	Traffic imbalance
More ports exposed	Table imbalance
	Constrained server attachment
	Less code reuse

Data Plane: Pipe/Table Bypass Logic

Internet traffic


Pipe bypass (no need to query VXLAN tables)


```
struct metadata_t {
    bit<1> flag; /* whether to bypass the current pipe */
    bit<4> subflag; /* whether to bypass the current table */
    ...
}
control Ingress( ... ) {
    action tbl_ac1() { flag = 0; }
    action tbl_ac2() { flag = 1; subflag = 0; }
    action tbl_ac3() { flag = 1; subflag = 1; }
    ...
    table tbl {
        /* to distinguish different cloud services */
        key = { ... }
        /* to take different bypass actions */
        actions = {
            tbl_ac1; /* bypass the current pipe */
            tbl_ac2; /* enter the current pipe, query tb2 */
            tbl_ac3; /* enter the current pipe, query tb3 */
            ... } }
    table tb2 { ... }
    table tb3 { ... }
    ...
    apply {
        tbl.apply();
        if (flag == 1) {
            if (subflag == 0) { tb2.apply(); }
            else if (subflag == 1) { tb3.apply(); }
            ...
        }
        ... /* flag is 0, bypass the current pipe */
    } }
```

P4 code framework for pipe/table bypass

In LuoShen, although each packet will sequentially pass through CGW, IGW and SW, not all tables have to be queried. We can make an early judgment to determine whether the packet will be processed by the local pipe or even the local table to reduce unnecessary processing overhead.

Control Plane: Isolation and Configuration

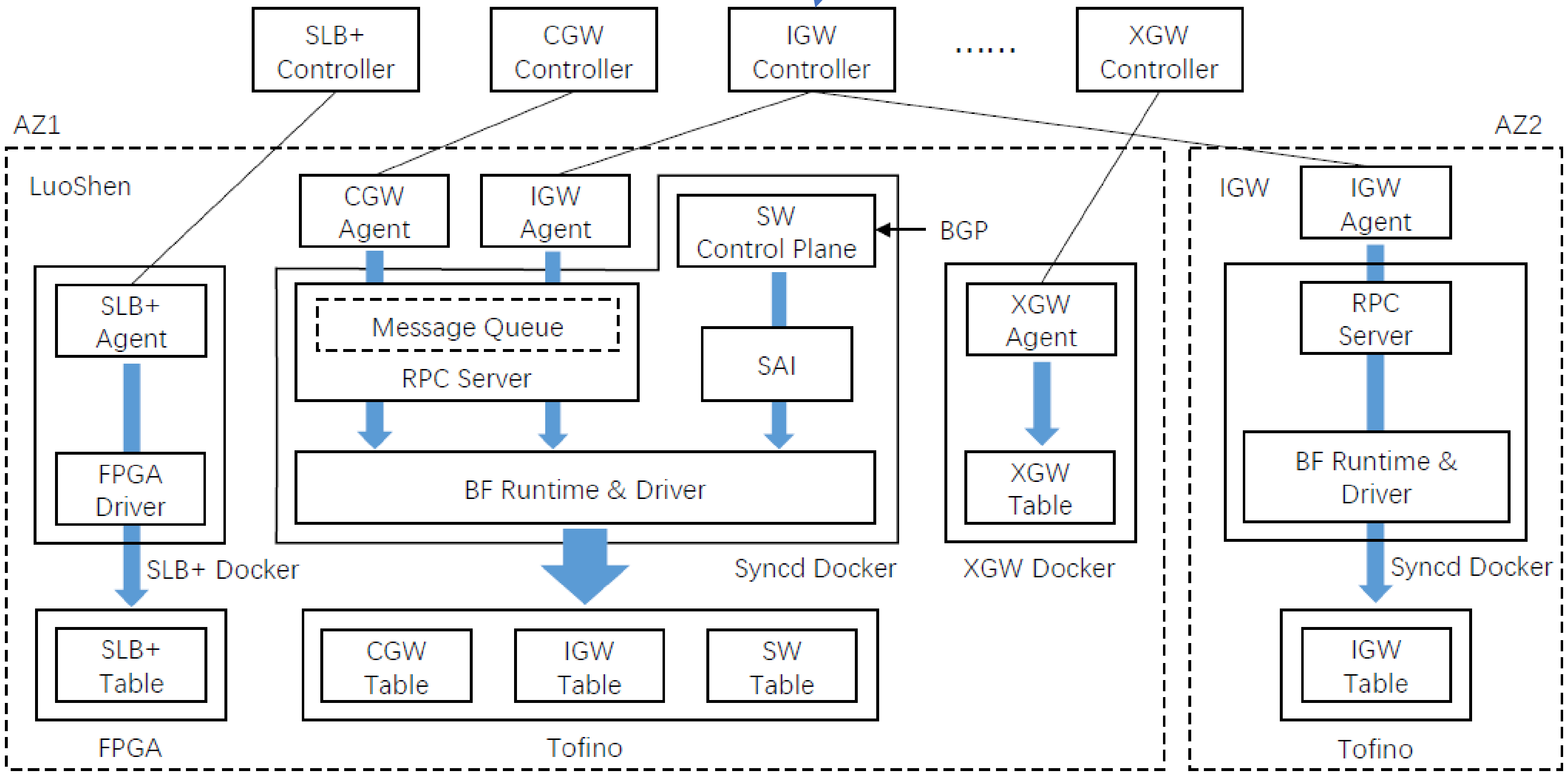
Resource isolation

- Resource dockerization
- CPU binding
- Isolation via cgroups
- Contain memory leaks
- Contain core dump files

Multi-component configuration

- Control plane code reuse
- Separated channels to CPU/FPGA/Tofino
- Separated channels to CGW/IGW and SW
- Batch configuration through BF Runtime

The same IGW controller for both LuoShen and IGW with code reuse

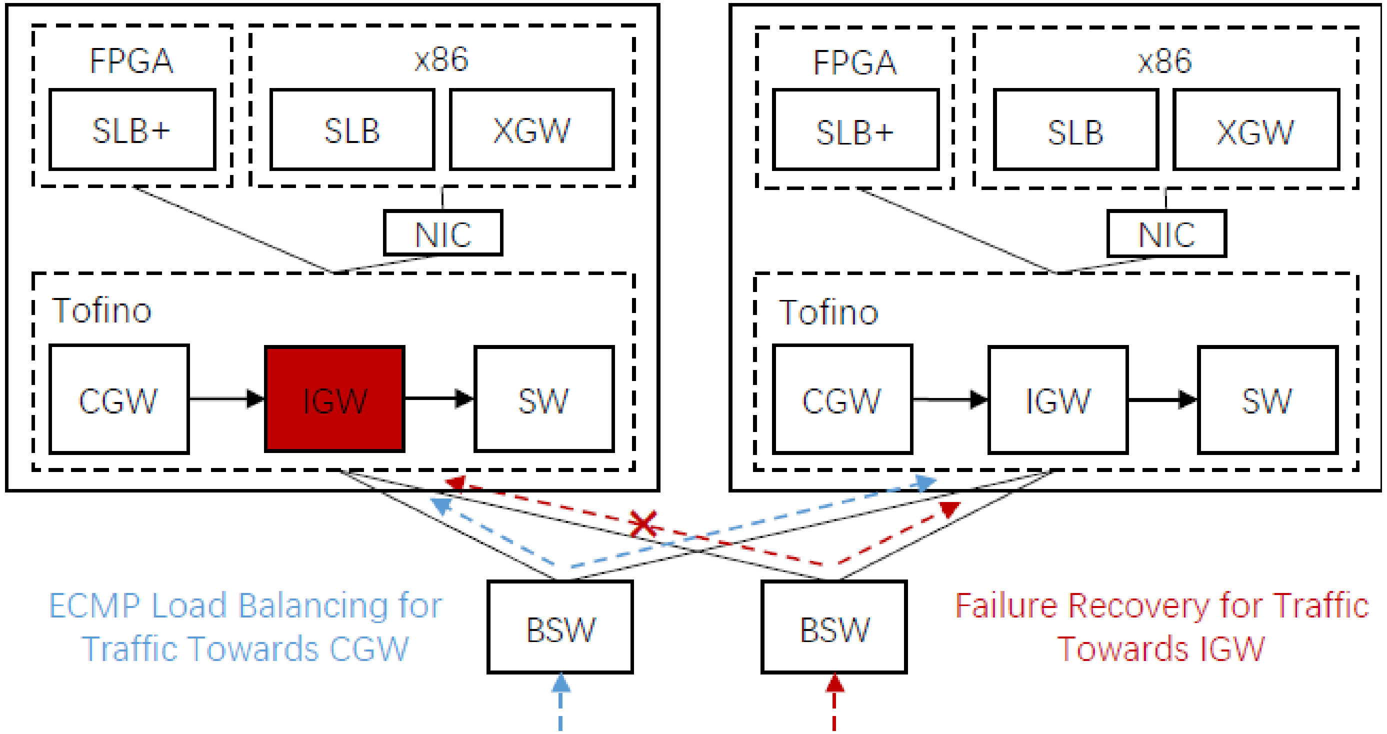


Multi-component table configuration channels

Control Plane: Inter-Component BGP Peering

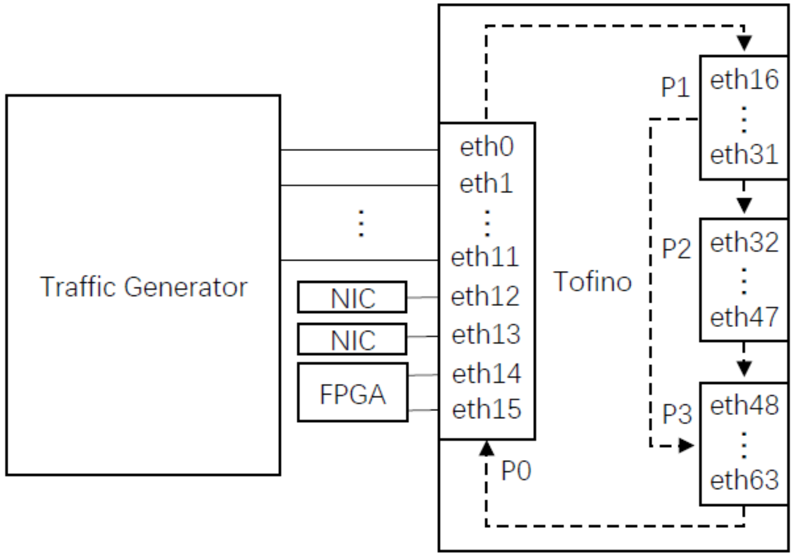
To exchange the reachability information between components, we set up BGP speakers at the control plane for inter-component BGP peering so that a component can learn the routes to others and it can also advertise its reachability to others.

With BGP peering, LuoShen achieves high availability based on component-level ECMP load balancing and fast failure recovery.

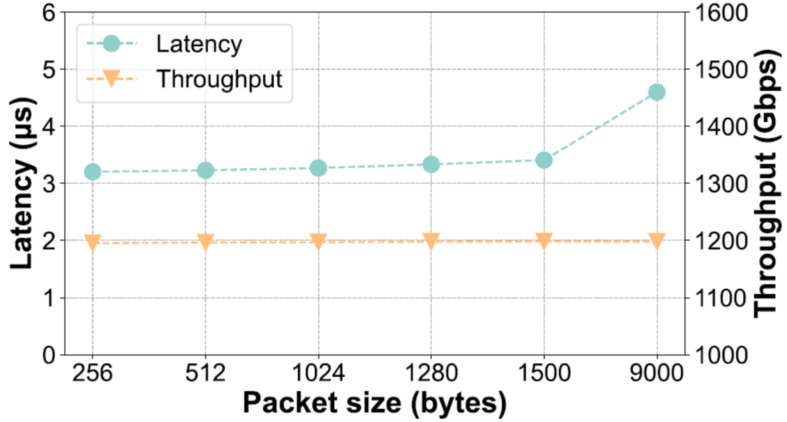

FPGA
SLB+
x86
SLB
XGW
NIC
Tofino
CGW
IGW
SW
FPGA
SLB+
x86
SLB
XGW
NIC
Tofino
CGW
IGW
SW
ECMP Load Balancing for Traffic Towards CGW
BSW
BSW
Failure Recovery for Traffic Towards IGW

Hot-standby deployment of LuoShen in production

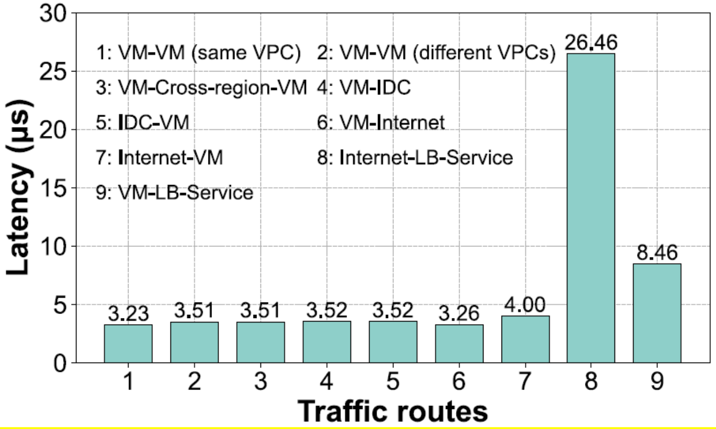
Evaluation#1: Forwarding Performance



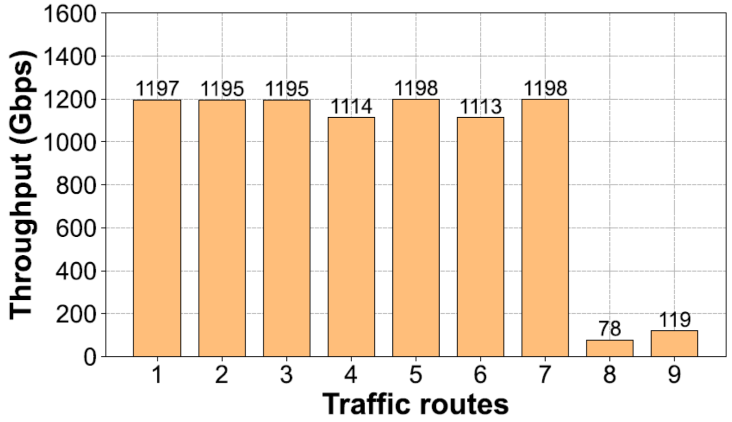
Performance test topology



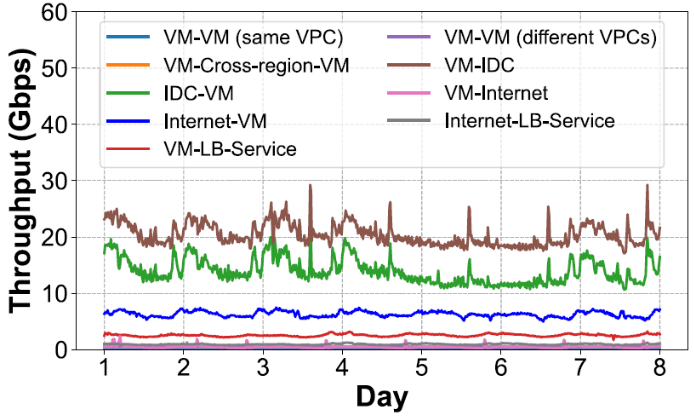
1.2Tbps throughput is achieved with maximum latency bounded in 5μs.



Latency of different traffic routes varies (512B pkts).



Throughput of different traffic routes varies (512B pkts).



Traffic throughput of different cloud services in production.

Evaluation#2: Advantages of LuoShen at the Edge

Basic of Calculation:

LuoShen --- 1 * 2U device with Tofino * 1, CPU * 2, FPGA * 1

Role-splitting --- 1 * 2U x86 server with FPGA for SLB+; 2 * 2U x86 servers for XGW and SLB; 3 * 2U Tofino switches for VGW, IGW and TGW; 3 * 1U switches for CSW, LSW and original SW

Cost of FPGA, x86 server, Tofino switch, LuoShen ≈ 1 : 10 : 10 : 15

Power of FPGA, x86 server, CPU, Tofino switch ≈ 100W : 500W : 200W : 300W

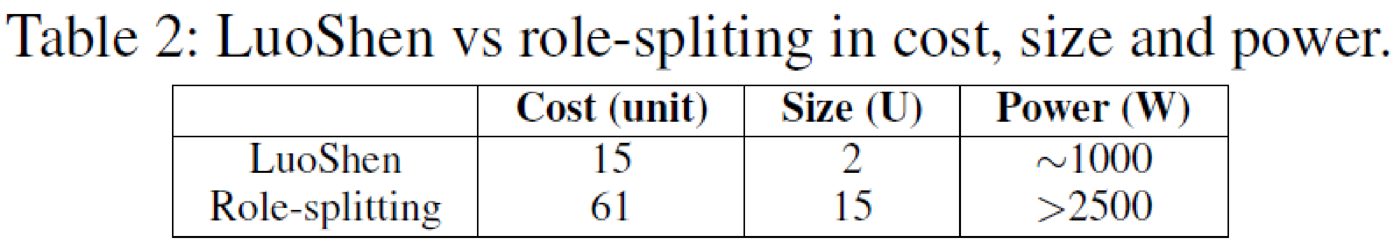


Table 2: LuoShen vs role-spliting in cost, size and power.

	Cost (unit)	**Size (U)**	**Power (W)**
LuoShen	15	2	∼1000
Role-splitting	61	15	>2500

LuoShen reduces the upfront cost, deployment footprints and power consumption by 75%, 87% and 60%, respectively.

Experiences

- □ How to deploy LuoShen in a step-by-step way in production?
- □ How to make hot upgrade of components in LuoShen?
- □ How to achieve performance and table size scaling in LuoShen?
- □ How to achieve failure isolation/failsafe in LuoShen?
- □ How to conduct fine-grained telemetry and debugging in LuoShen?
- □ How can elastic NFV deployment be achieved in LuoShen by utilizing server resources external to LuoShen during peak workloads?

Summary

- LuoShen is Alibaba's hyper-converged gateway for multi-tenant multi-service edge clouds. It follows a p4-centric architecture and achieves a good balance of performance, costs and deployment footprints.
- At the data plane, we propose techniques such as pipeline folding, pipe/table bypass, on-chip resource optimizations to maximize the table convergence density inside the Tofino.
- At the control plane, we achieve resource isolation, reserve multiple configuration channels, and conduct BGP peering with hot standby for inter-component reachability and high availability.
- LuoShen achieves 1.2Tbps throughput and reduces the upfront cost, deployment size and power usage by 75%, 87%, 60%, compared with the original role-splitting architecture.

LuoShen: A Hyper-Converged Programmable Gateway for Multi-Tenant Multi-Service Edge Clouds

Q & A