上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Network Service in OpenStack Cloud

Yaohui Jin

email: jinyh@sjtu.edu.cn

Sina Weibo: @bright_jin

(The slides will be shared in Sina Weipan & Slideshare)

Network & Information Center

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

Acknowledgement

- Team: Dr. Xuan Luo, Pengfei Zhang, Xiaosheng Zuo, Zhixing Xu, Xinyu Xu, Jianwen Wei, Baoqing Huang, etc.
- Prof. Hongfang Yu and team with UESTC
- Prof. Jianping Wang with CityU HK
- Engineers, discussion and slides from Intel, SINA, IBM, Cisco, Dell, VMware/EMC, H3C, Huawei, IXIA, …
- OpenStack Community
- China OpenStack User Group (COSUG)
- China OpenStack Cloud League (COSCL)
- Technical blogs such as blog.ioshints.info, ipspace.net, …

About me

- 上海交通大学 教授，以前做光通信的，现在改行做云计算了。。。 ☺
- 上海交通大学 网络信息中心 副主任，其实就是个苦逼的挨踢网管啊。。。 ☹
- 研究兴趣： 数据中心网络，海量流式数据分析，云计算架构

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

OpenStack in Academia for Research & Operation

- USC, Information Science Institute
- Purdue University
- University of Melbourne
- San Diego Supercomputer Center
- Brookhaven National Lab., DOE
- Argonne National Lab., DOE
- European Organization for Nuclear Research (CERN)
- Shanghai Jiao Tong University
- University of Science & Technology of China
- University of Electrical Science & Technology of China
- ……

Agenda

- Introduction
- SDN and OpenFlow
- Network Virtualization
- Network Virtualization in OpenStack
- Our Work

The Service Trend

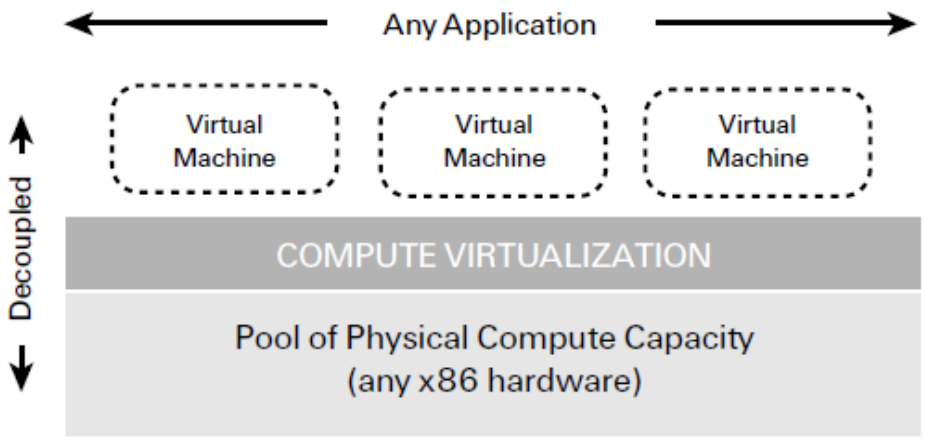
- "Decoupling infrastructure management from service management can lead to innovation, new business models, and a reduction in the complexity of running services. It is happening in the world of computing, and is poised to happen in networking.“

Jennifer Rexford

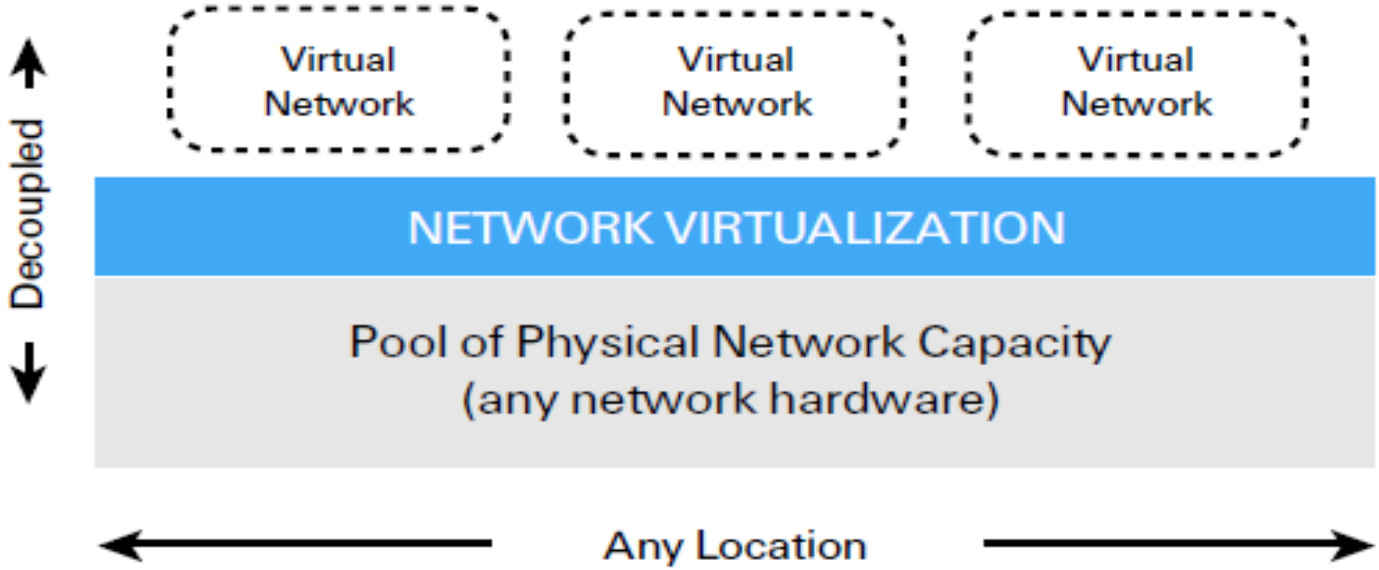
Professor, Princeton University

- Last month, VMware paid $1.2B to acquire Nicira for software defined networking (SDN).

Why is Nicira worth $1.2 billion?



vmware

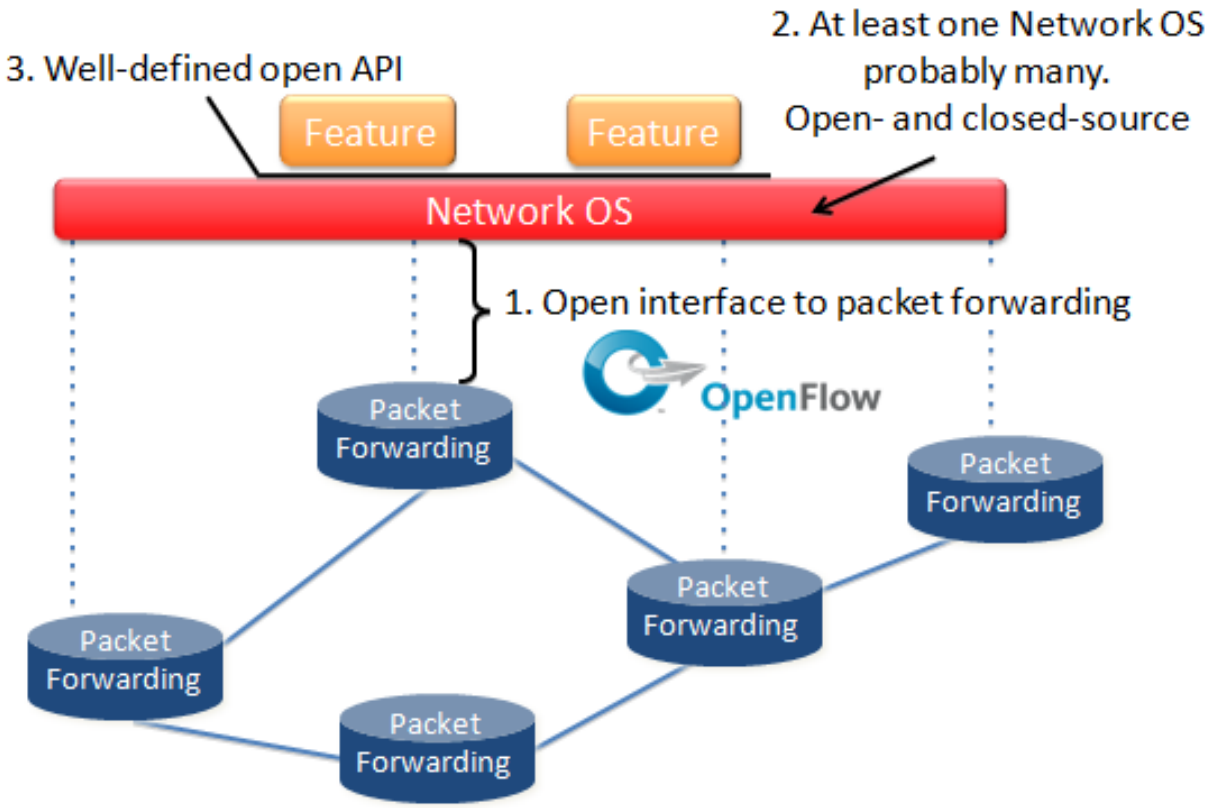


nicira

SDN and OpenFlow

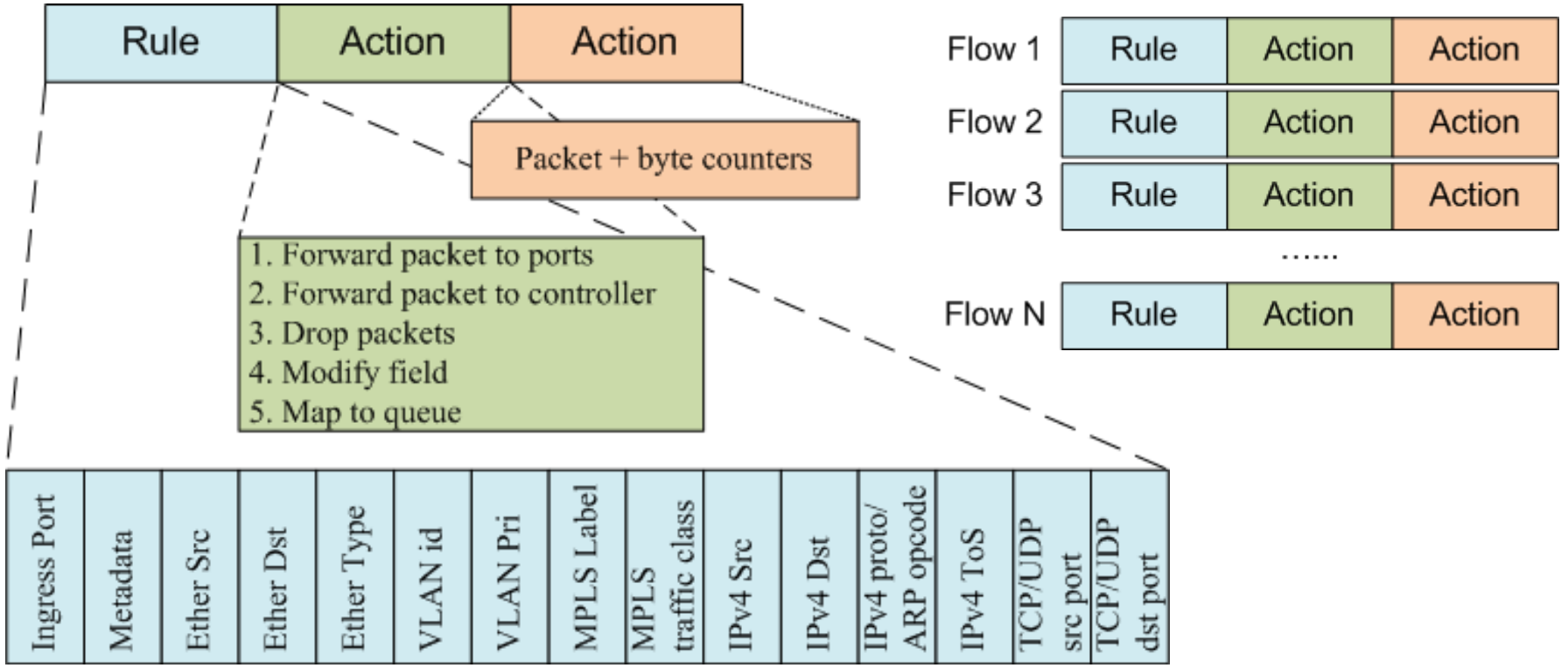
Software Defined Network (SDN)

- A network architecture in which the network control plane (OS) is decoupled from the physical topology using open protocols such as OpenFlow.



Flow Table (v1.1)

- Rules: Ethernet, IP, MPLS, TCP/UDP any combination, exact or wildcard
- Actions: Forward, Drop, Modify field (NAT)
- Statistics: Volume based billing, anti DDOS



OpenFlow Implementation

- Hypervisor Mode
  - Open vSwitch (OVS): XEN, KVM, …
  - OVS other features: security, visibility, QoS, automated control
- Hardware Mode
  - OpenFlow Switch
  - Hop by hop configuration

Reality Check

- "OpenFlow doesn't let you do anything you couldn't do on a network before" –Scott Shenker (Professor, UC Berkeley, OpenFlow co-inventor)
- Frames are still forwarded, packets are delivered to hosts.
- OpenFlow 1.3 was recently approved.
- Major vendors are participating - Cisco, Juniper, Brocade, Huawei, Ericsson, etc. It's still early stage technology but commercial products are shipping.
- OpenFlow led by large companies Google/Yahoo/Verizon and lack of focus on practical applications in the enterprise.

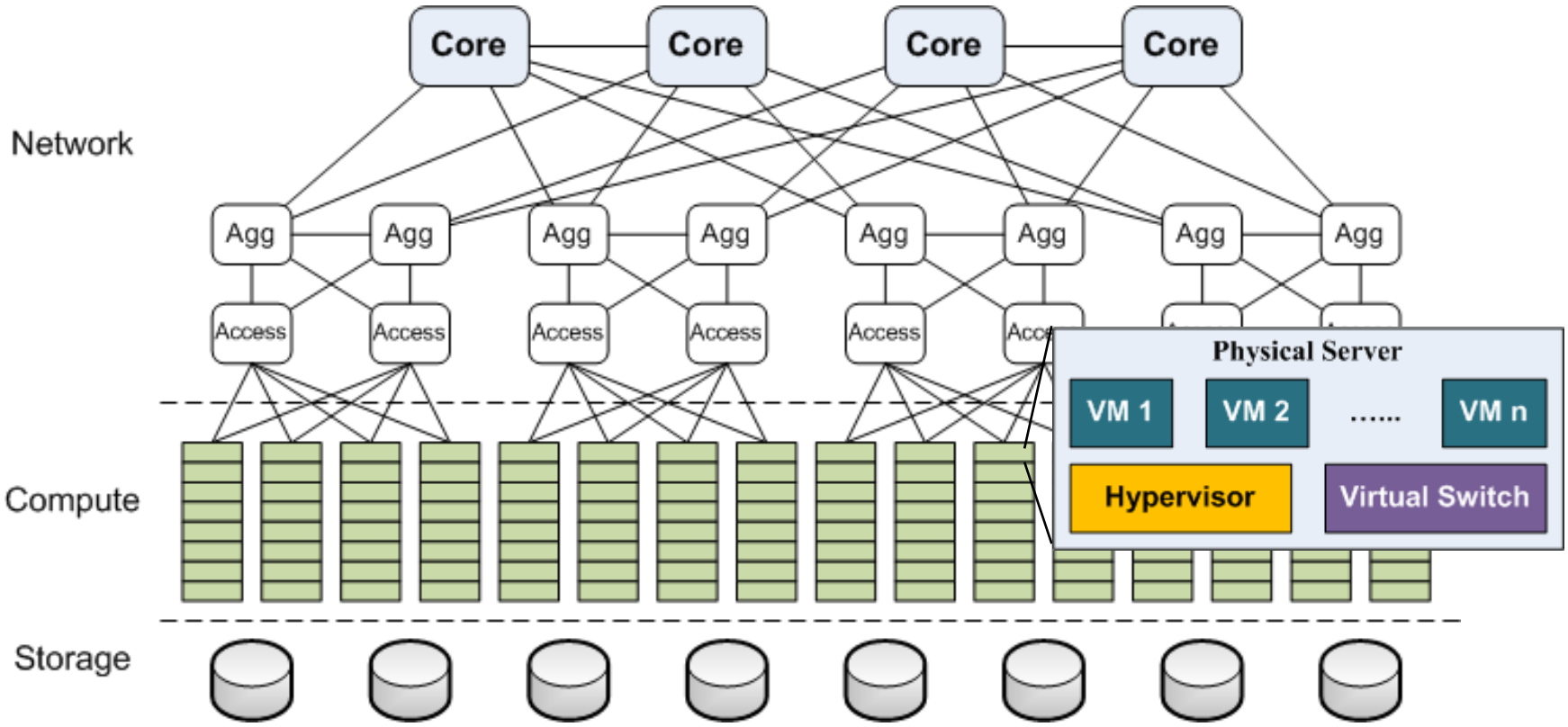
OpenFlow Interop

- Fifteen Vendors Demonstrate OpenFlow Switches at Interop (May 8-12, 2011)

big switch networks
IBM
BROADCOM.
Juniper NETWORKS
BROCADE
NEC
CITRIX
NETGEAR
DELL
NetOptics
extreme networks
FULCRUM microsystems
OPNET
hp
pronto

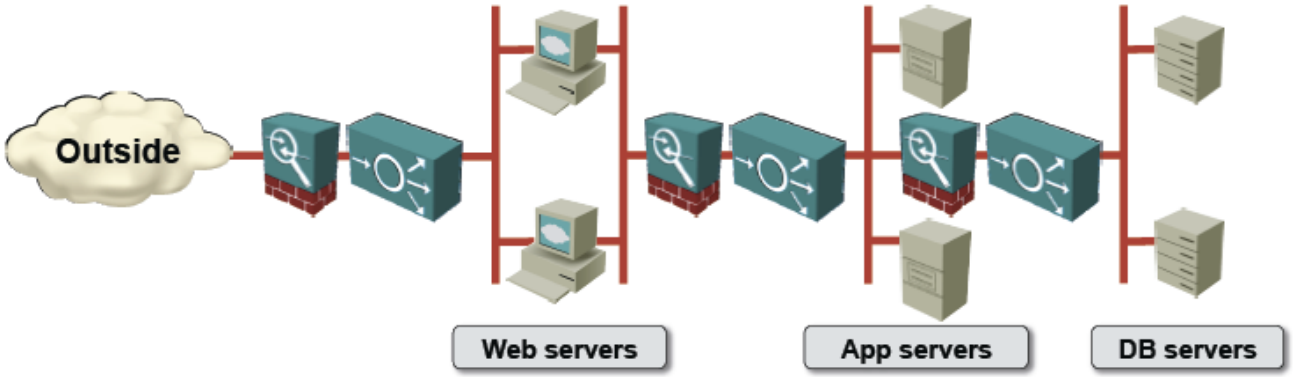
Network Virtualization

General Data Center Architecture

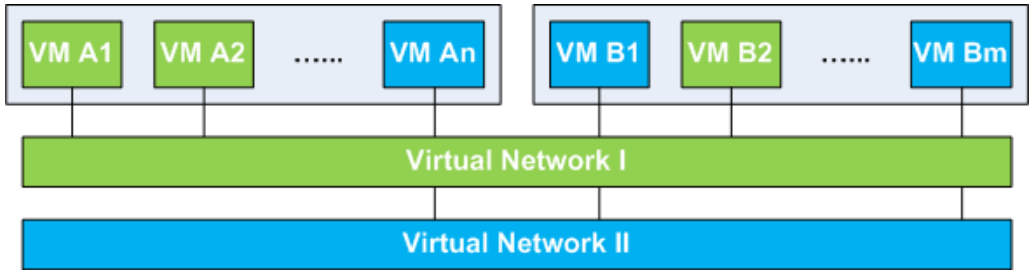


Cloud management system allows us dynamically provisioning VMs and virtual storage.

What customers really want?



Virtual Network



- Requirements
  - Multiple logical segments
  - Multi-tie applications
  - Load balancing and firewalling
  - Unlimited scalability and mobility

Multi-Tenant Isolation

- Making life easier for the cloud provider
  - Customer VMs attached to “random” L3 subnets
  - VM IP addresses allocated by the IaaS provider
  - Predefined configurations or user-controlled firewalls
- Autonomous tenant address space
  - Both MAC and IP addresses could overlap between two tenants, or even within the same tenant
  - Each overlapping address space needs a separate segment

Scalability

- Datacenter networks have got much bigger (and getting bigger still !!)
  - Juniper's Qfabric ~6000 ports, Cisco's FabricPath over 10k ports
- Tenant number dramatically increase as the IaaS experiences rapid commoditization
  - Forrester Research forecasts that public cloud today globally valued at $2.9B, projected to grow to $5.85B by 2015.
- Server virtualization increase demand on switch MAC address tables
  - Physical with 2 MACs -> 100 VMs with 2 vNIC need 200+ MACs!

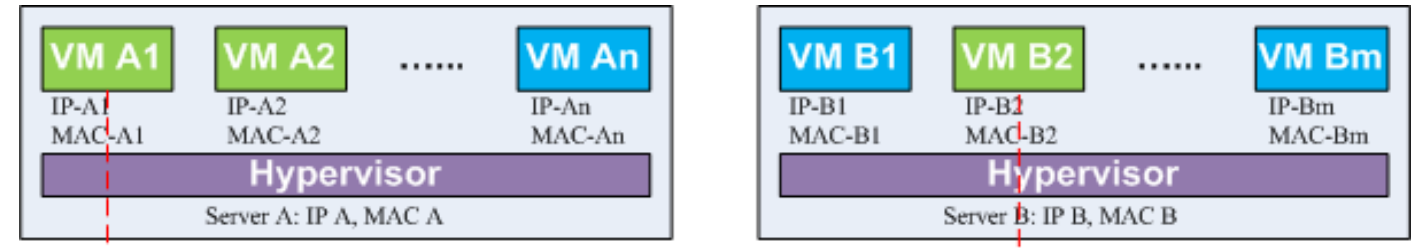
Possible Solutions (1)

- VLANs per tenant
  - limitations of VLAN-id range (Only 12bits ID = 4K)
  - VLAN trunk is manually configured
  - Spanning tree limits the size of the network
- L2 over L2
  - vCDNI(VMware), Provider Bridging(Q-in-Q)
  - Limitations in number of users (limited by VLAN-id range)
  - Proliferation of VM MAC addresses in switches in the network (requiring larger table sizes in switches)
  - Switches must support use of same MAC address in multiple VLANs (independent VLAN learning)

Possible Solutions (2): L2 over IP

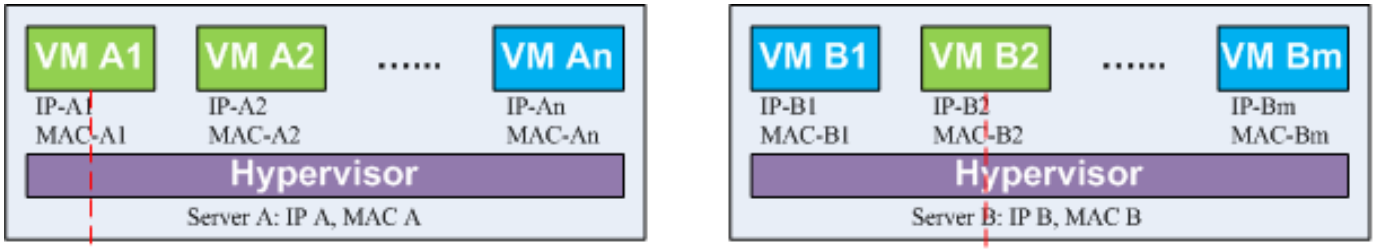
- Virtual eXtensible LAN (VXLAN)
  - VMware, Arista, Broadcom, Cisco, Citrix, Red Hat
  - VXLAN Network Identifier (VNI): 24 bits = 16M
  - UDP encapsulation, new protocol
- Network Virtualization Generic Routing Encapsulation (NVGRE)
  - Microsoft, Arista, Intel, Dell, HP, Broadcom, Emulex
  - Virtual Subnet Identifier (VSID): 24 bits = 16M
  - GRE tunneling, relies on existing protocol
- Stateless Transport Tunneling (STT)
  - Nicira
  - Context ID: 64 bits, TCP-like encapsulation

VXLAN/NVGRE: How it Works?



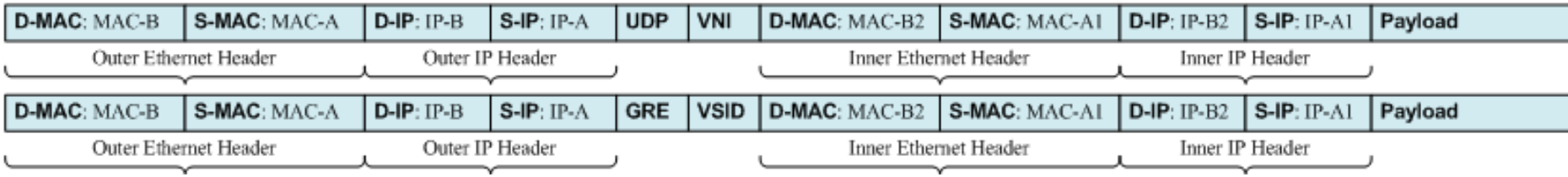
without overlay

Ethernet frame from VM-A1 to VM-B2



using VXLAN

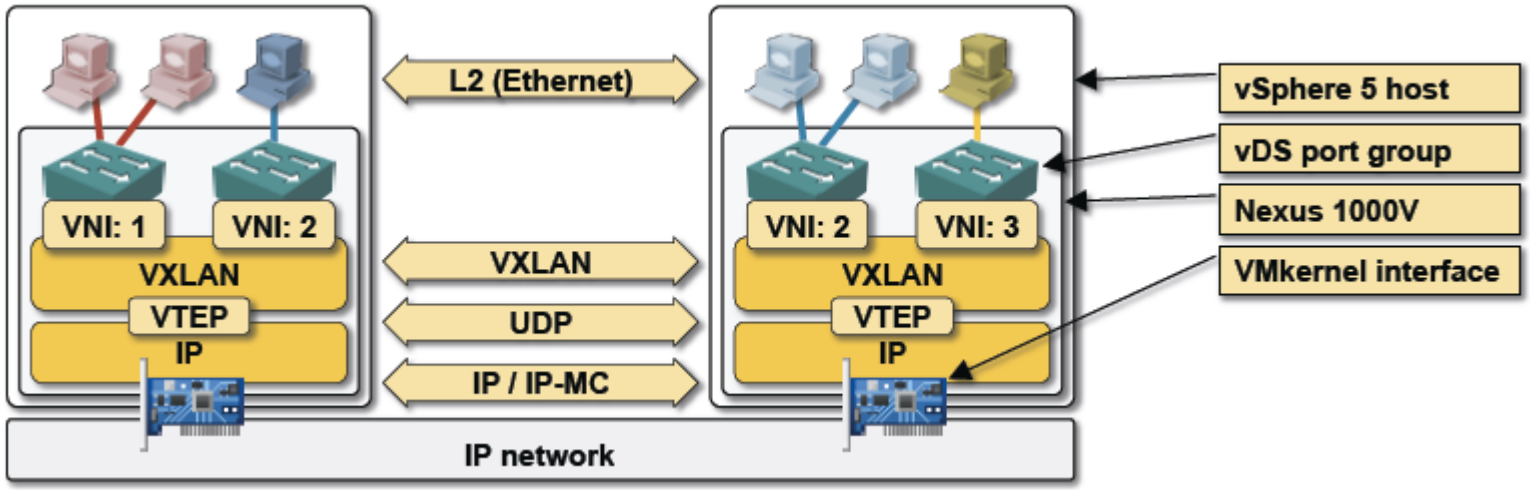
Ethernet frame from VM-A1 to VM-B2



using NVGRE

Dynamic MAC learning

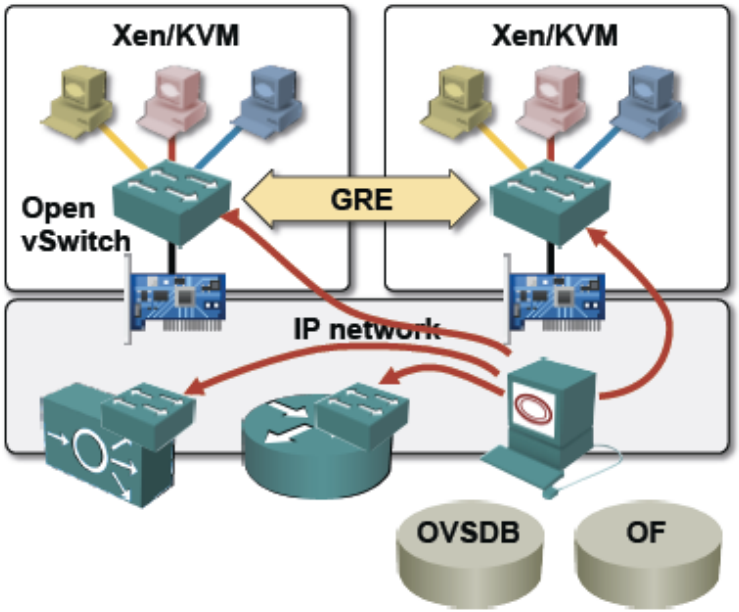
- Dynamic MAC learning with L2 flooding over IP multicasting



Flooding does not scale when fabric gets bigger.

Control Plane (Nicira)

- L2-over-IP with control plane
  - OpenFlow-capable vSwitches
  - IP tunnels (GRE, STT ...)
  - MAC-to-IP mappings by OpenFlow
  - Third-party physical devices
- Benefits
  - No reliance on flooding
  - No IP multicast in the core



Transitional Strategy Depends on Your Business

- 100s tenants, 100s servers: VLANs
- 1000s tenants, 100s servers: vCDNI or Q-in-Q
- Few 1000s servers, many tenants: VXLAN/NVGRE/STT
- More than that: L2 over IP with control plane

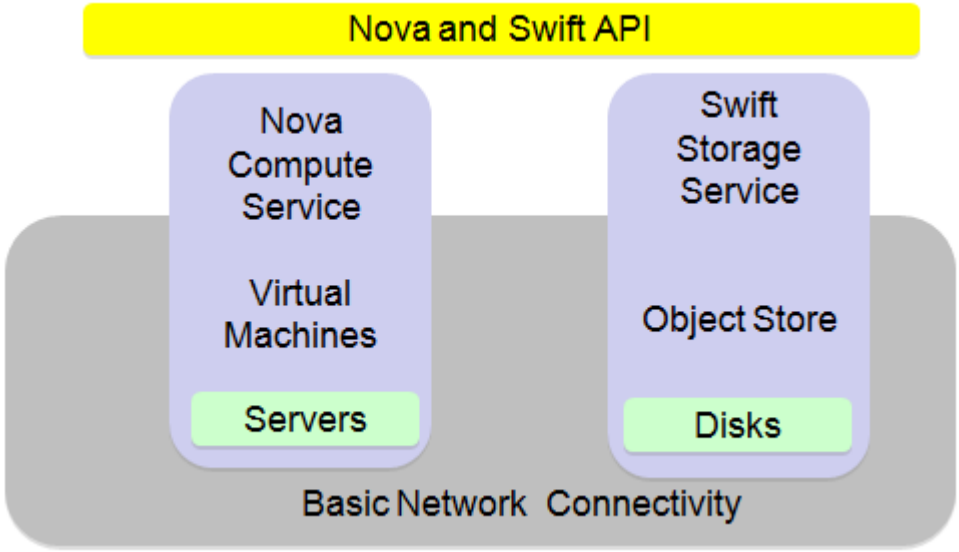
Open question: How to solve the co-existing scenarios in one cloud?

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Network Virtualization in Openstack

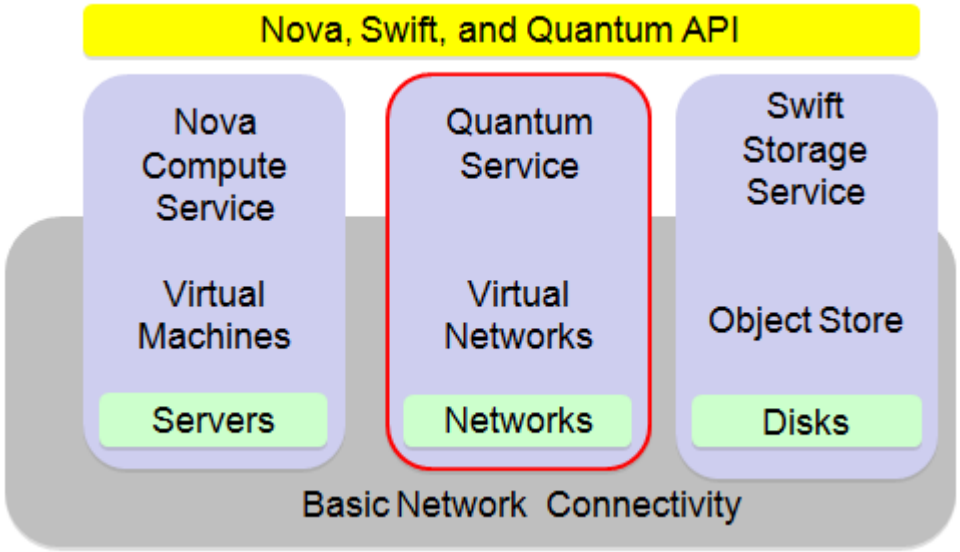
OpenStack Today

- Networking is embedded inside of Nova compute, and un-accessible to application developers
- Details and differences associated with network provisioning complicates a simple compute service
- Difficult to track changes in networking as Software-defined Networking (SDN) comes into play

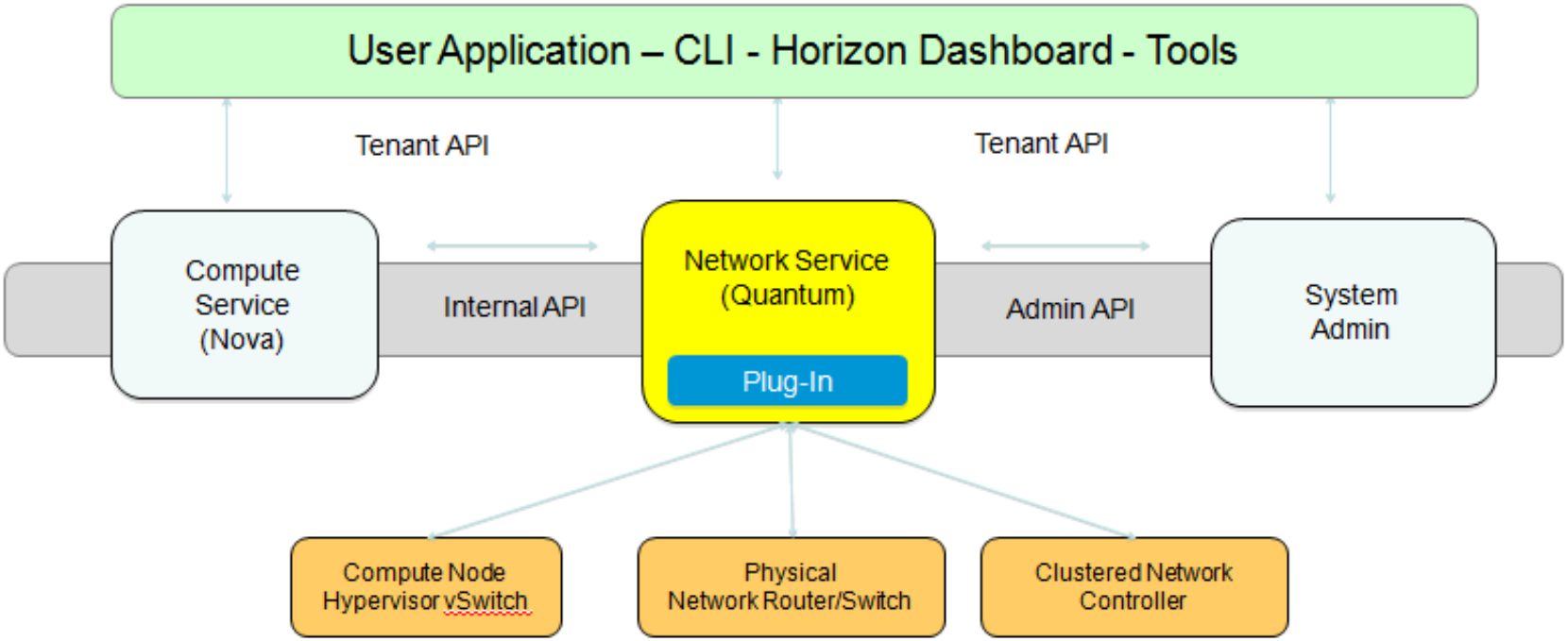


With Quantum – Networking becomes a Service

- Nova becomes simpler, easier to maintain and extend
- Developers have ability to create multiple networks for their own purposes (multi-tier apps)
- May support provisioning of both virtual and physical networks – differences captured through plugin's



Quantum API interactions

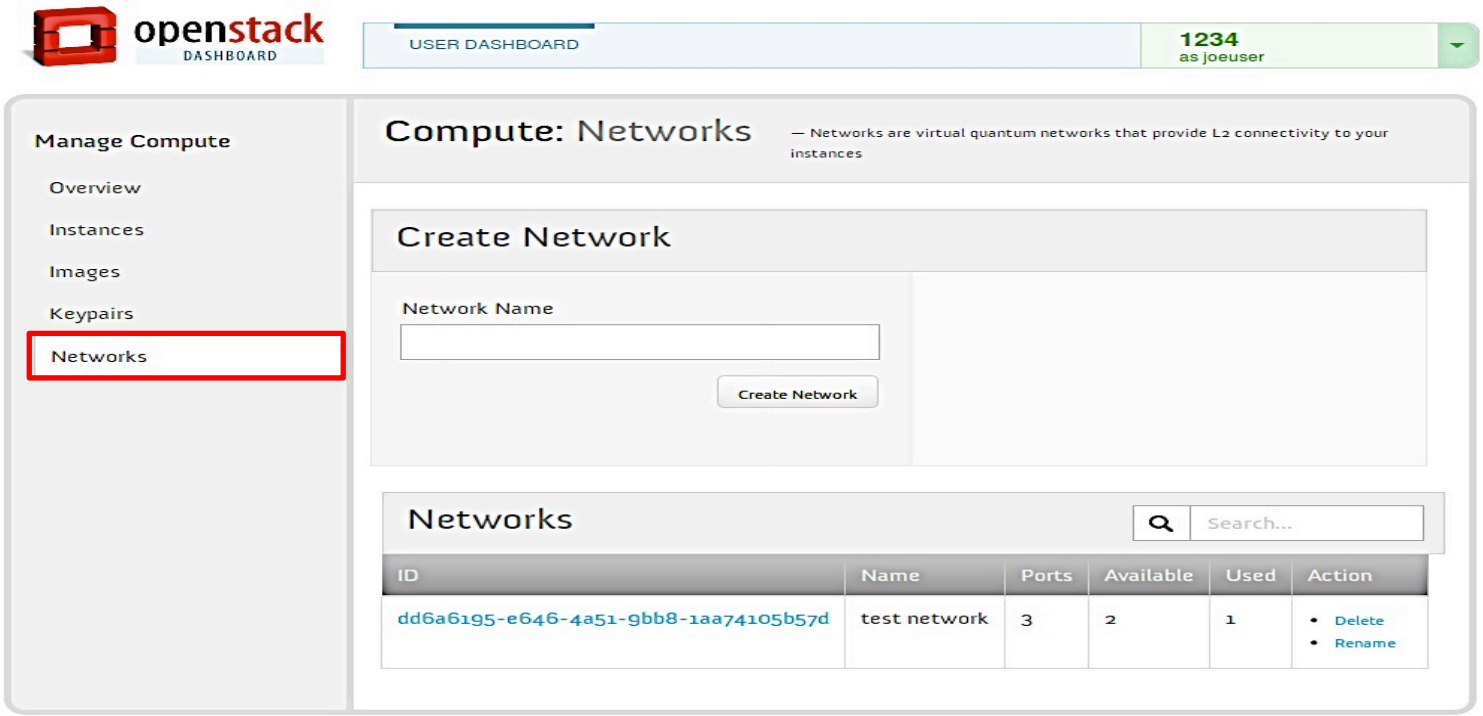


Plug-in's available today

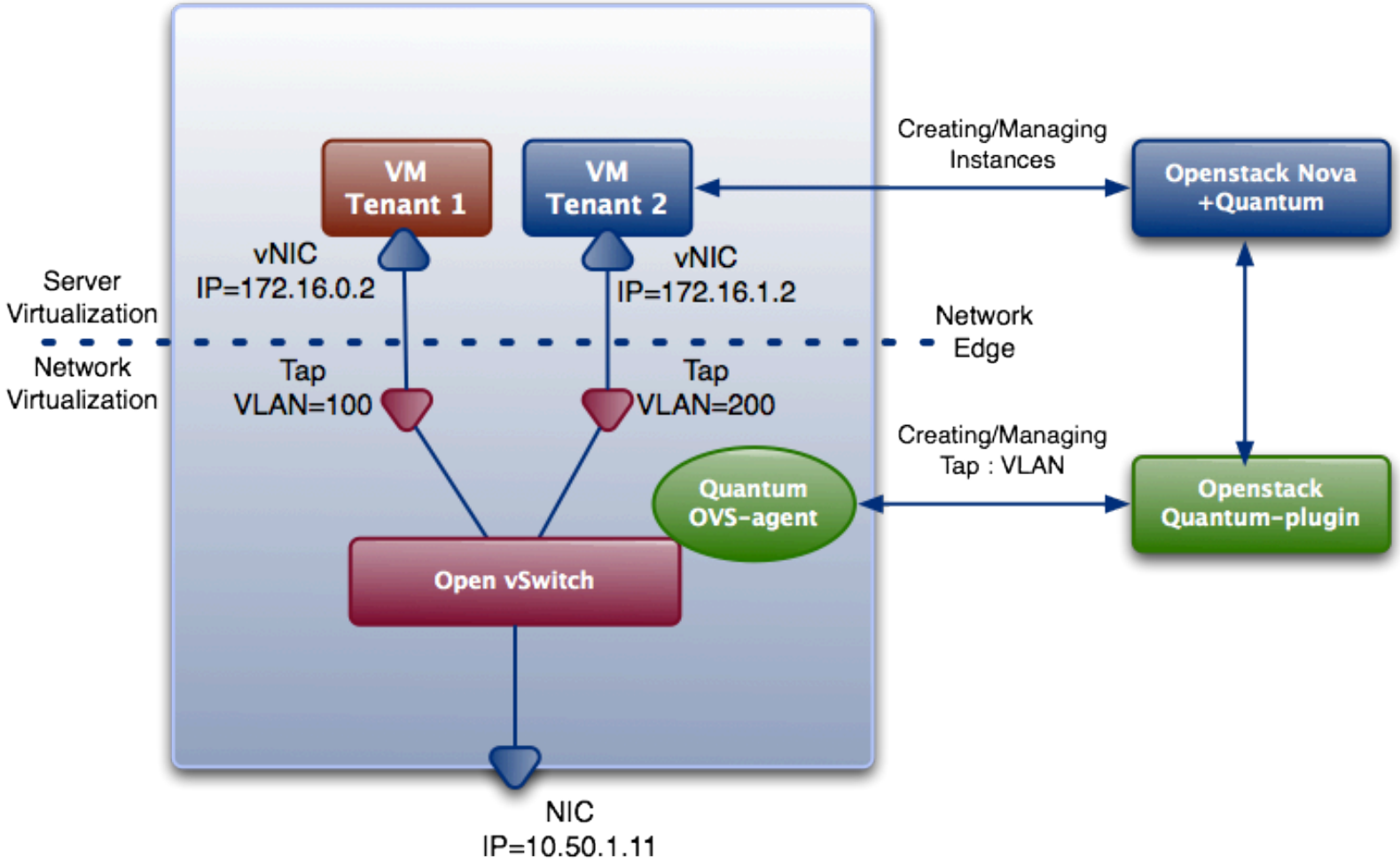
- Open vSwitch
- Linux bridge
- Nicira NVP
- Cisco (Nexus switches and UCS VM-FEX)
- NTT Labs Ryu OpenFlow controller
- NEC OpenFlow
- Big Switch Floodlight

Quantum in Horizon

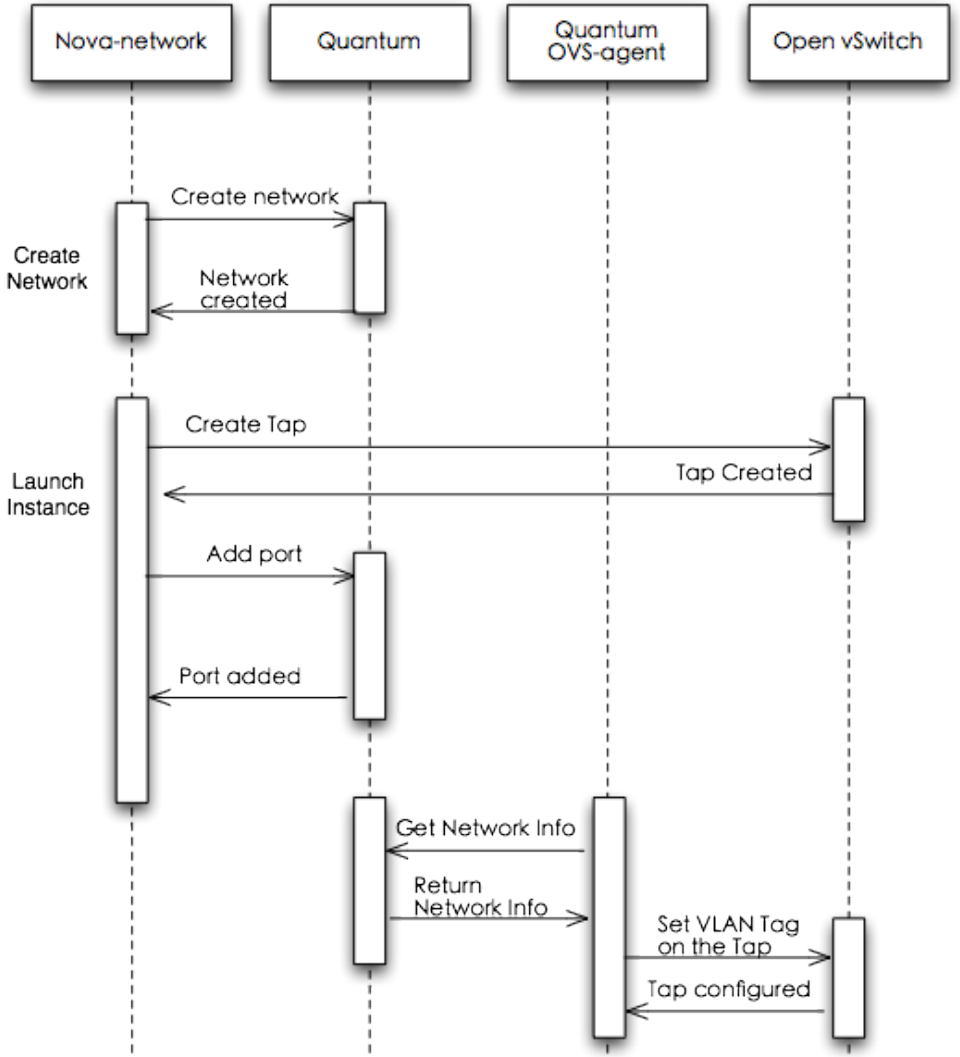
- Create/delete private network
- Create “ports” and attach VM’s
- Assign IP address blocks (DHCP)



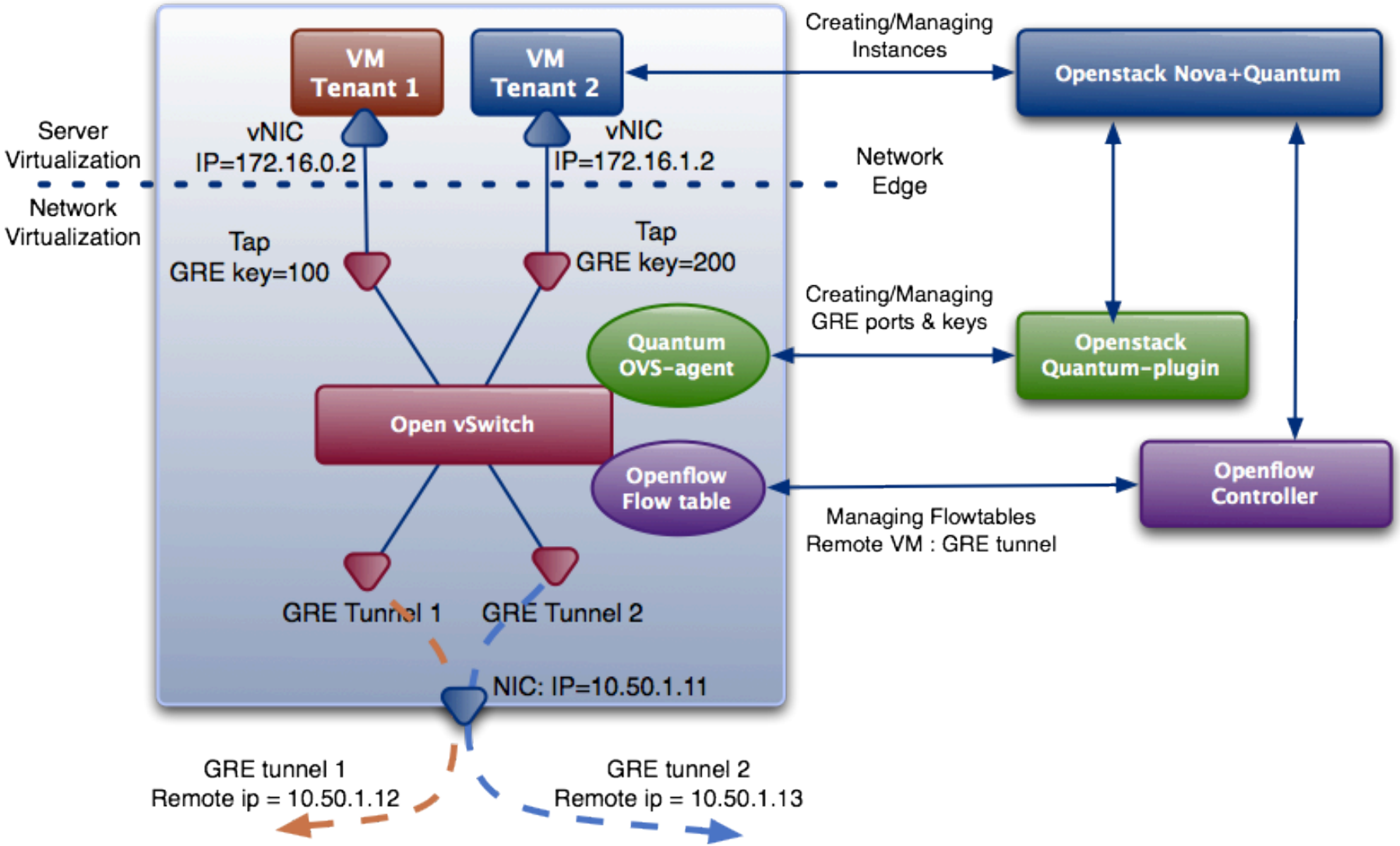
Quantum OVS Plugin: VLAN solution with Open vSwitch



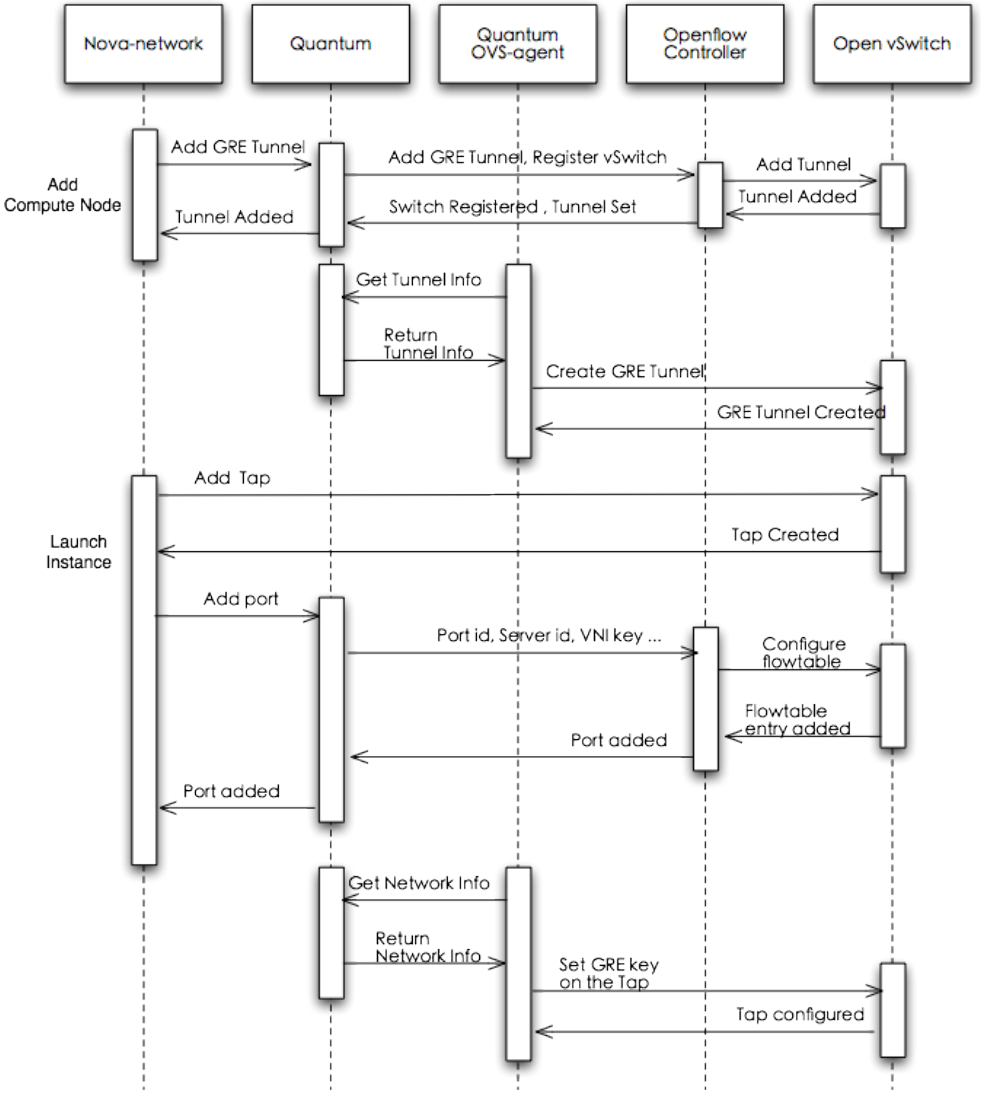
OVS Plugin Flow Chart



Ryu Plugin: Overlay solution with Openflow



Ryu Plugin Flow Chart



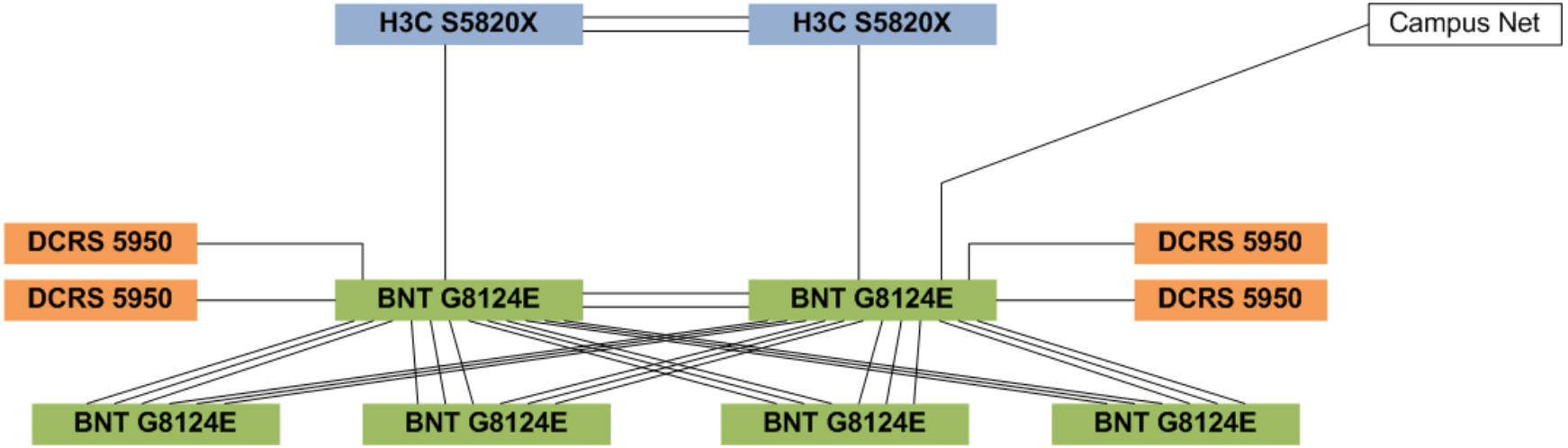
上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

vCube: Virtual, Versatile, Visible Network Service for OpenStack Cloud

© jinyh@sjtu

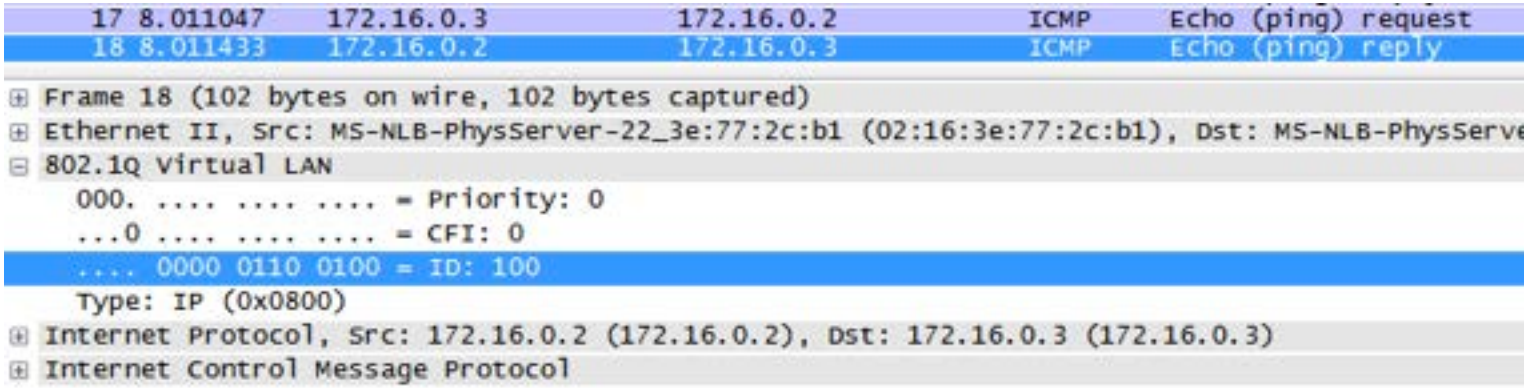
Network Environment

- Data Center Network: 10 GE Switch (BNT&H3C) in 2 domains
- Control and Manage: GE Switch (DCRS)
- 10GE connect to campus network
- Fat tree topology; L3: VRRP;
- L2: LACP+VLAG+MSTP
- Security control: SSH, NAT, ACL, VLAN
- NIC: Intel X520-DA2; Chelsio T420E-CR



Transition:Co-existing VLAN/GRE

- VLAN solution: Openstack + Open vSwitch



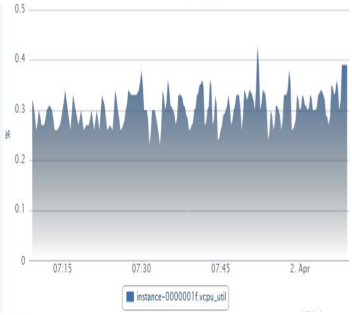
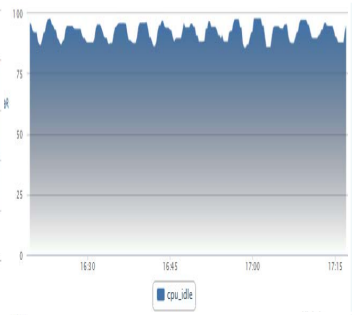
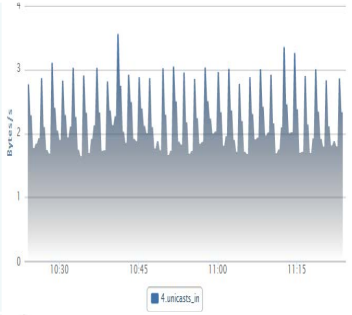
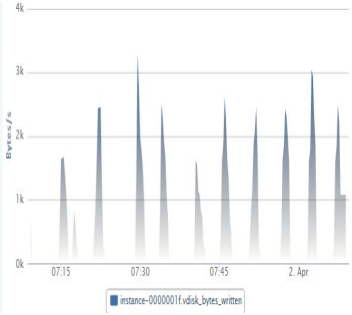
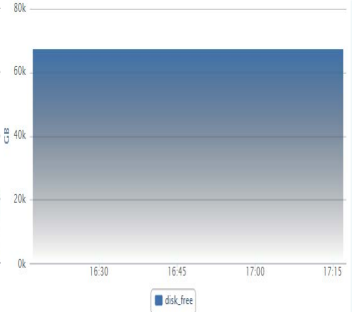
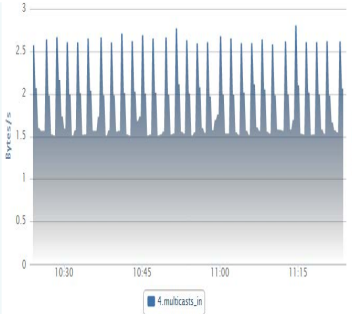
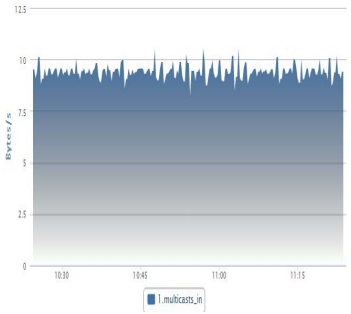
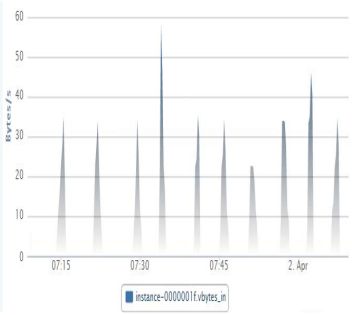
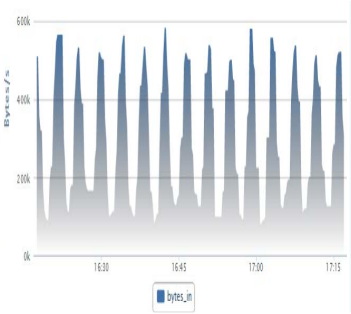
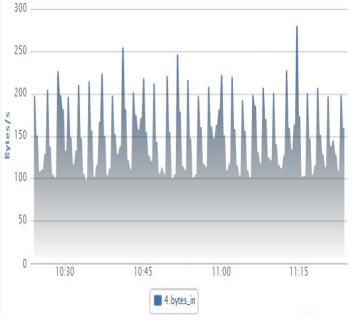
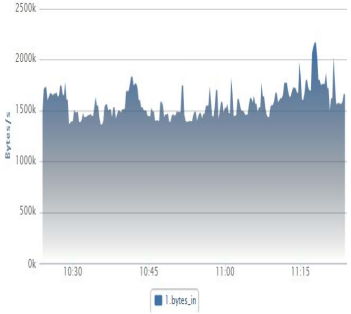
- GRE solution: Openstack + Ryu

```
NXST_FLOW reply (xid=0x4):
 cookie=0x0, duration=160852.878s, table=0, n_packets=334, n_bytes=30492, priority=1,in_port=18 actions=set_
 cookie=0x0, duration=160866.895s, table=0, n_packets=842, n_bytes=77812, priority=1,in_port=17 actions=set_
 cookie=0x0, duration=160818.522s, table=0, n_packets=48780, n_bytes=64444834, priority=0 actions=resubmit(
 cookie=0x0, duration=158078.155s, table=1, n_packets=49650, n_bytes=64526734, priority=1 actions=learn(tab
N_ID[],output:NXM_OF_IN_PORT[]),resubmit(,2)
 cookie=0x0, duration=8.37s, table=2, n_packets=5, n_bytes=434, hard_timeout=50,priority=1,tun_id=0x1,dl_dst
 cookie=0x0, duration=8.37s, table=2, n_packets=0, n_bytes=0, hard_timeout=50,priority=1,tun_id=0x2,dl_dst=0
 cookie=0x0, duration=8.371s, table=2, n_packets=6, n_bytes=476, hard_timeout=50,priority=1,tun_id=0x1,dl_ds
 cookie=0x0, duration=103396.557s, table=2, n_packets=349, n_bytes=14658, priority=0,arp actions=FLOOD
root@c9:~#
```

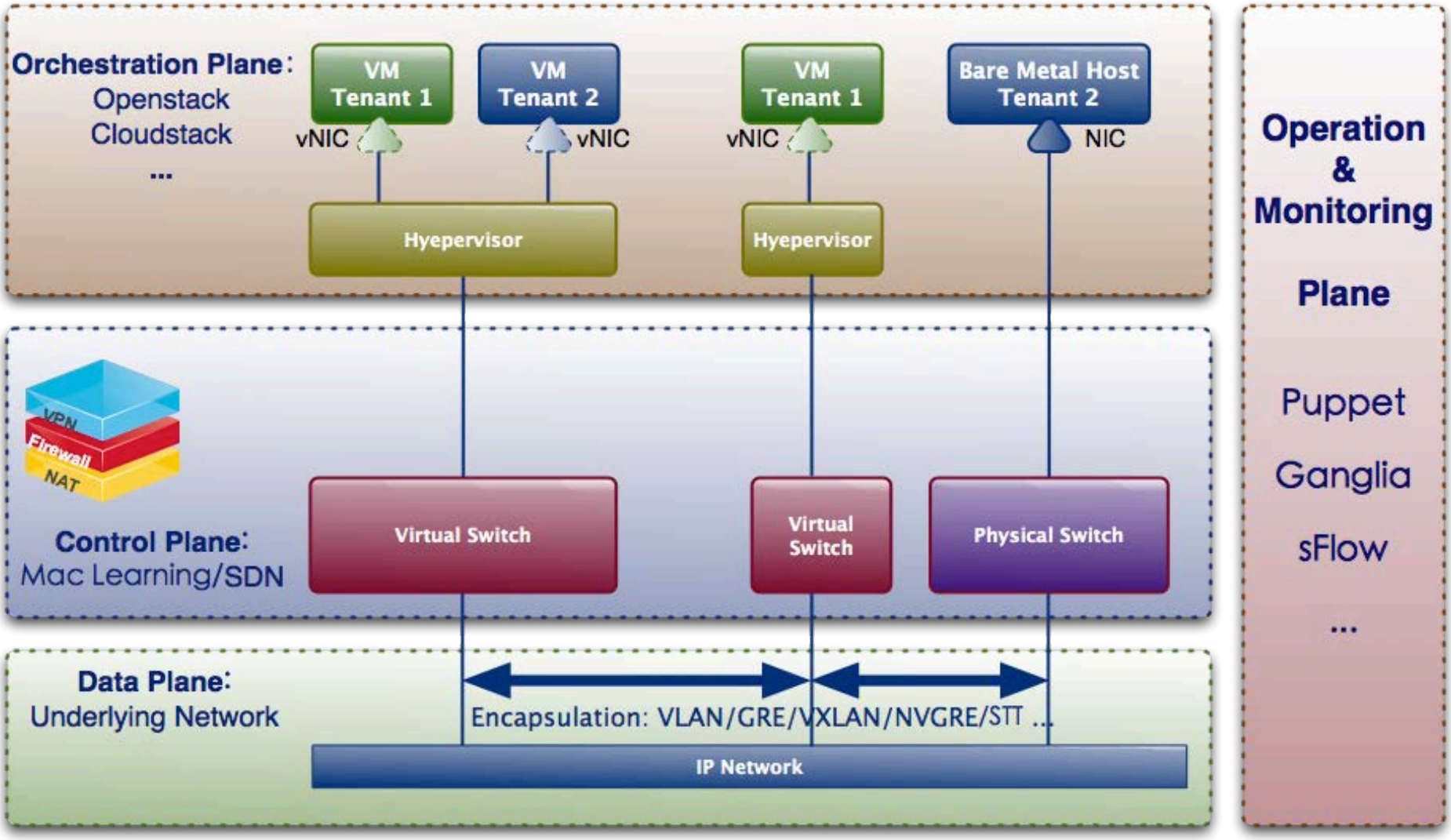
QoS in Virtual Network

- Bandwidth upper bound for VMs
  - With only OVS : 200Mbit/s
  - With OVS and virtio: 8Gbit/s
- Bandwidth guarantee with Openstack + OVS
  - User defined rate limitation
  - Differential service level for tenants
  - High bandwidth utilization
  - Stable performance under dynamic traffic

Visible Virtual Network by sFlow

	Virtual Machine	Physical Server		Virtual Switch	Physical Switch
CPU			Unicast		
Disk			Multicast		
Traffic			Port Traffic		

The Whole Picture



Thanks for your attention!

Weibo: @bright_jin