Week 30

\* Multi-Tenant Data Center (MTDC)

Tenants in a multi-tenant data center require access to the public Internet, while also being required to communicate among its VMs as if they were connected to the same LAN environment, regardless of location. Requirements for the MTDC network infrastructures include:

• Managing and isolating overlapping address spaces

• Support of VM migration for flexible provisioning

• Decoupling of virtual from physical network topology

• Ability to scale to large numbers of nodes

There are two existing general approaches to provide

Virtual Private Networks (VPN) for MTDC networks. One approach is to micro-manage all FIB (Forward Information Base) of all network entities in the data center by establishing link paths from source host to destination hosts, e.g. by using IEEE 802.1q (VLAN) or MPLS. These protocols use additional fields to route or to isolate sub- networks without intervention of conventional link layer or IP network protocol. However, VLAN incurs complexity of configuration and has scalability limitations.

The other approach is to encapsulate/tunnel virtual network packets within physical network headers. Arbitrary network topologies and addressing schemes of the MTDC network can be realized with this approach. Furthermore, it can be accomplished using commodity switches without the requirement of additional protocol features. However, this approach comes with the cost of encapsulation overhead: the virtual network header is encapsulated by physical network header, reducing the effective MTU (Maximum Transfer Unit), and packet-processing on general purpose processors incurs frequent context switching on the hypervisor.

\* Related Works of MTDC Network Architecture

Several MTDC network architectures and data center net- work virtualization techniques have been proposed. To name a few, NVP uses the Stateless Transport Tunneling proto- col (STT), VXLAN and GRE as tunneling protocols at a service node (which resides in the hypervisor) to provide network virtualization. To avoid the encapsulation overhead at hypervisor, NVP introduced STT, which amortizes the encapsulation overhead by coalescing multiple datagrams. Although NVP is able to achieve line rate of 10 Gbps, it still hogs up CPU resources at both end-hosts and requires special feature in NIC called TCP Segmentation Offload (TSO). Since STT disguises encapsulation header as regular TCP header, there is no handshake or acknowledgment; thus，middle boxes such as firewall or load balancers in the data center need to be modified. Furthermore, a single packet loss can lead to drop of the entire STT frame, and significant performance degradation. When NVP uses GRE tunneling, it only achieves 25% of line rate at 10 Gbps and incurs more than 80% load at CPU.

VL2 uses IP-in-IP and the tunneling process is done at the ToR(Top of Rack) switches, but its evaluation is based on 1 Gbps line rate. NetLord uses an agent on hypervisor for tunneling and defines their own encapsulation protocol.

PortLand uses packet rewriting instead of tunneling, by translating pseudo MAC and actual MAC at ToR switches. However, it assumes IP Core/aggregation switches have longest prefix match upon MAC address. It uses MAC addresses in a manner similar to IP addresses in a multi-root fat tree data center topology [13]. Although it is supported in OpenFlow-enabled switches, it is difficult to expect commodity switches to have this feature, since longest prefix match is only implemented upon IP address but not flat addresses such as MAC.

SecondNet uses “source based routing” by using MPLS label field. Every packet is encapsulated by MPLS field, which records every egress port of every switch on the path. Since data center physical network topology is fairly static with small number of hops, the 20-bit width MPLS label has enough space to record all egress port numbers of every hopping switch. Since it is not based on encapsulation it achieves 10Gbps line rate with fair throughput balance among VMs. However, it has scalability limitations in terms of the IP core switch hopping count, and requires MPLS feature on every switch at the core and aggregation layers.

Common aspects of aforementioned MTDC network architectures are that they assume low-cost commodity switches with simple routing protocol at IP core/aggregation, which use only a few FIBs at each switch.

Another common aspect is that MTDC architectures virtualize network at the boundary of the core fabric network and provides a mean to resolve “many-to-one” mapping of virtual address space to physical underlay address space either by multiplexing addresses or encapsulating packets. These functions reside in hypervisor, edge switches, or ToR switches. As examples, PortLand uses edge switches to map multiple actual MAC addresses to a pseudo MAC address. NetLord and NVP place agents at the hypervisor to perform encapsulation of virtual address by corresponding physical network address. VL2 uses IP-in-IP in ToR switches by encapsulating virtual IP address by physical IP address.