Network Overlays

Vivek Kashyap
vivk@us.ibm.com
Linux Technology Center
IBM
Cloud/DataCenter

- Virtual Machines (VM) deployed on physical nodes in the DataCenter
  - Domain: rack, VLAN, site
- Network isolation a must for a viable multi-tenant solution
  - Each tenant (company, department) has security and privacy requirements
  - Private view of network: load-balancers, firewalls
- Collaborative applications on same virtual network
Virtualization increases network complexity

- VM deployable anywhere in the data-center without subnet boundary, host, rack server, or VLAN constraints
  - VM IP and MAC addresses migrate with the VM
  - VM network security profile migrates with the VM
- Physical network must be dynamically configured
  - Network state must be coordinated between hypervisor and switch port
    > Physical network modified as the VM migrates
Layer-2 and massive scaling of data centers

- Broadcast/multicast is problematic in flat layer-2 networks
  - Every device in the network must process the packets
  - Broadcast storms may bring network down

- Address resolution
  - ARP: broadcast and periodic flushing of entries
  - Neighbour Discovery: Solicited node address reduces load significantly
  - Routers connected to L2 broadcast domains must handle ARP traffic

- Virtual machine placement is constrained
  - Limiting broadcast domain to TOR limits placement of workloads to the rack
  - Extending layer-3 to aggregation layer may cause broadcast on all VLANs
    - Layer-2 access switches enable VLAN on all ports to allow reachability
    - Layer-3 to core suffers maximum broadcast load

- Large number of Virtual Machines and the resultant broadcast/multicast frames require much larger table sizes in the switches
Multi-tenancy and large data center networks

- Each tenant expects features akin to a physical data center
  - Security and isolation, control, workload balancing and placement
- Network resources must be provisioned on-demand
  - New customer, additional resources
  - Workload/VM migration without being constrained by subnet boundaries
- Tenant virtual network spans multiple boundaries
  - For optimal utilization the VM's from one tenant may be supported in same rack or host
  - Tenant traffic must stay isolated
    > VLANs and broadcast domain
    > Firewalls
- Tenant virtual network must interact with non-virtualized resources where needed
  - Firewalls, intrusion detection, load-balancers, etc.
Solution: Create overlay network over the physical network
- Decouple physical and logical configuration
- VM's dynamically deployed without effecting switch/router table sizes

Overlay network == Tunnel
- Encapsulate the frame at first hop network device (vswitch, switch, router)
- Send packet to the target decapsulating device
Virtual networks

- Each tenant's traffic is completely isolated from other tenants
  - Frames carry a Virtual Network Identifier (24bit == 16 million networks)
    > For comparison: there are only 4K VLANs
- Virtual network address space is isolated from other virtual networks
  - Same IP and MAC may be used in different virtual networks
  - VM can be placed anywhere irrespective of layer 2 constraints
- Packets do not exit a virtual network except through controlled 'gateways'
Multi-tenancy with overlapping addresses
Overlay network structure

- The encapsulation and decapsulation of packets is done at the edge boundary of the overlay (Overlay Boundary Point)
  - The OBP may be a vswitch, access switch, or a network appliance
- OBP keeps a per VN mapping of end-station to remote OBP address
- OBP maintains a per-VN state for delivering multicast packet
Standardization required

- Define a header format
  - A VNID must be supported in each frame
  - Payload maybe Ethernet or IP packet
- Fragmentation?
  - Encapsulation may lead to exceeding the link MTU
    - Fragmentation of packet at IP layer, done in overlay, or prevented
- Checksum & FCS
  - Need not duplicate checksum or FCS in both inner and outer headers
- Control plane
  - How to populate the forwarding table of a virtual network instance?
  - How to handle multi-destination frames with a virtual network instance?
  - How to associate an end-point with a virtual network instance?
  - How to de-associate an end-point from a virtual network instance
IETF Proposals on overlay networks

- VxLAN: Virtual exTensible Local Area Network
  - 24-bit VN ID
  - Encapsulates frames in UDP (MAC in UDP)
  - Utilizes IP multicast for address resolution and broadcast

- NVGRE: Network Virtualization using Generic Routing Encapsulation
  - 24 bit VN ID
  - Encapsulates frames in GRE
  - Utilizes IP multicast for address resolution and broadcast

- Stateless Transport Tunneling Protocol for Network virtualization
  - Utilizes TCP segmentation offload capabilities of NIC
Address resolution

- VxLAN/NVGRE rely on IP multicast
  - Packets are flooded to all OBPs participating in the VN if target unknown
- Can we avoid flooding?
  - Suggests ToR/switch caching mappings and responding
  - Multiple caches in network may provide similar function
  - Extend to support an 'ARP reduction module'
    > Receives ARP requests from the bridge. If not in cache sends to ARM agent.
    > ARM agent consults “Address database” to return target OBP mapping
Fragmentation?

- Overlay encapsulation may cause packet to be fragmented
  - A loss of fragment implies loss of entire data-segment
- Proposal
  - vSwitch/OBP considers IP_DF bit to be set for each packet implicitly
  - If the packet is to be fragmented due to addition of overlay header
    » Generate an ICMP error “Datagram too big” to the VM
      » This is the case with IPv6 by default
  - VM will reduce its view of the MTU for that path as a result
    » And no need to fragment in OBP
  - If IP packet is larger than MTU, the VM's IP stack will fragment
    » Each fragment will be an encapsulated as a separate overlay packet
Additional references

1. Network Virtualization Overlay Control Protocol Requirements
   - http://www.ietf.org/id/draft-kreeger-nvo3-overlay-cp-00.txt

2. ARP Broadcast reduction for large data centers

3. Problem statement for ARMD

4. Problem Statement: Overlays for Network Virtualization
Questions?

Contact: vivk@us.ibm.com
Legal Statement

This work represents the view of the author and does not necessarily represent the view of IBM.

IBM is a registered trademark of International Business Machines Corporation in the United States and/or other countries.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Other company, product, and service names may be trademarks or service marks of others.