Authors

Marc Lasserre
Florin Balus
Thomas Morin
Nabil Bitar
Yakov Rekhter
Yuichi Ikejiri
Purpose of the draft

- This document provides a framework for Data Center Network Virtualization over L3 tunnels. This framework is intended to aid in standardizing protocols and mechanisms to support large scale network virtualization for data centers:
  - Reference model & functional components
  - Help to plan work items to provide a complete solution set
  - Issues to address
Generic DC Network Architecture

DC GW: Gateway to the outside world (e.g. for inter-DC, VPN and/or Internet connectivity)

Intra-DC Network: High capacity network composed of core switches/routers aggregating multiple ToRs

Top of Rack (ToR): Aggregation switch (can also provide routing, VPN, tunneling capabilities)

End Device: DC resource such as a compute resource (server or server blade), storage component, a network appliance or a virtual switch
Reference model for DC network virtualization over a L3 Network

NVE: Network Virtualization Edge node providing private context, domain, addressing functions

Tenant End System: End system of a particular tenant, e.g. a virtual machine (VM), a non-virtualized server, or a physical appliance
NVE Generic Reference Model

Overlay module: tunneling overlay functions (e.g. encapsulation/decapsulation, Virtual Network identification and mapping)

VNI: Virtual Network Instance providing private context, identified by a VNID

VAP: Virtual Attachment Points e.g. physical ports on a ToR or virtual ports identified through logical identifiers (VLANs, internal VSswitch Interface ID leading to a VM)

The NVE functionality could reside solely on End Devices, on the ToRs or on both the End Devices and the ToRs
Virtual Network Identifier

• Each VNI is associated with a VNID
  – Allows multiplexing of multiple VNIs over the same L3 underlay

• Various VNID options possible:
  – Globally unique ID (e.g. VLAN, ISID style)
  – Per-VNI local ID (e.g. per-VRF MPLS labels in IP VPN)
  – Per-VAP local ID (e.g. per-CE-PE MPLS labels in IP VPN)
Control Plane Options

• Control plane components may be used to provide the following capabilities:
  – Auto-provisioning/Auto-discovery
  – Address advertisement and tunnel mapping
  – Tunnel establishment/tear-down and routing

• A control plane component can be an on-net control protocol or a management control entity
Auto-provisioning/Service Discovery

• Tenant End System (e.g. VM) auto-discovery

• Service auto-instantiation
  – VAP & VNI instantiation/mapping as a result of local VM creation

• VNI advertisement among NVEs
  – E.g. can be used for flood containment or multicast tree establishment
Address advertisement and tunnel mapping

• Population of NVE lookup tables
  – Ingress NVE lookup yields which tunnel the packet needs to be sent to.

• Auto-discovery components could be combined with address advertisement
Tunnel Management

• A control plane protocol may be used to:
  – Exchange address of egress tunnel endpoint
  – Setup/teardown tunnels
  – Exchange tunnel state information:
    • E.g. up/down status, active/standby, pruning/grafting information for multicast tunnels, etc.
Overlays Pros

• Unicast tunneling state management handled only at the edge (unlike multicast)
• Tunnel aggregation
  – Minimizes the amount of forwarding state
• Decoupling of the overlay addresses (MAC and IP) from the underlay network.
  – Enables overlapping address spaces
• Support for a large number of virtual network identifiers.
Overlays Cons

• Overlay networks have no control of underlay networks and lack critical network information

• Fairness of resource sharing and co-ordination among edge nodes in overlay networks
  – Lack of coordination between multiple overlays on top of a common underlay network can lead to performance issues.

• Overlaid traffic may not traverse firewalls and NAT devices

• Multicast support may be required in the overlay network for flood containment and/or efficiency

• Load balancing may not be optimal
Overlay issues to consider

• Data plane vs Control Plane learning
  – Combination (learning on VAPs & reachability distribution among NVEs)
  – Coordination (e.g. control plane triggered when address learned or removed)

• BUM Handling
  – Bandwidth vs state trade-off for replication:
    • Multicast trees (large # of hosts) vs Ingress replication (small number of hosts)
    • Duration of multicast flows
Overlay issues to consider
(Cont’d)

• Path MTU
  – Add’l outer header can cause the tunnel MTU to be exceeded.
  – IP fragmentation to be avoided
  – Path MTU discovery techniques
  – Segmentation and reassembly by the overlay layer

• NVE Location Trade-offs
  – NVE in vSw, hypervisor, ToR, GW:
    • Processing and memory requirements
    • Multicast support
    • Fragmentation support
    • QoS transparency
    • Resiliency
Next Steps

• Terminology harmonization
• Add’l details about hierarchical NVE functionality