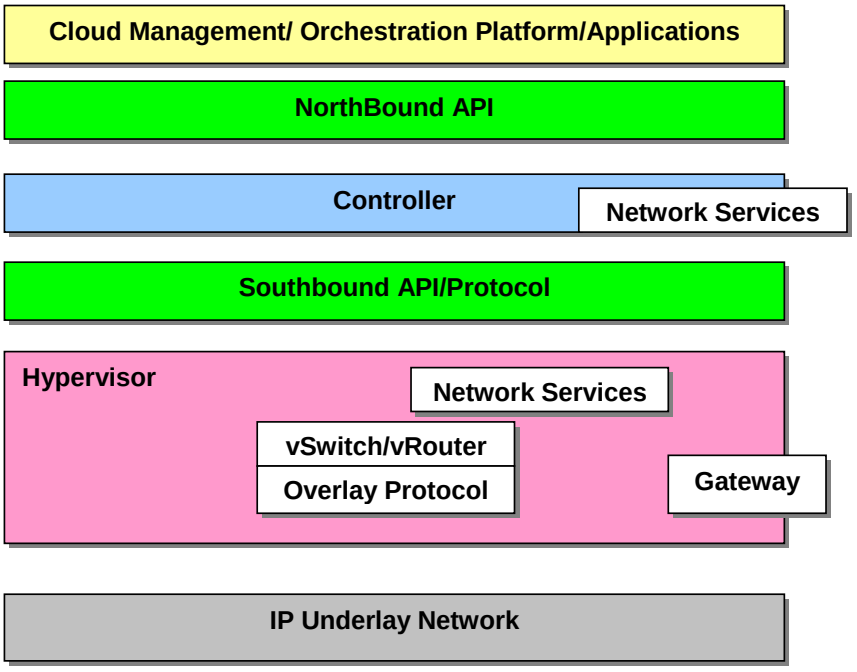
Architectural Criteria to Evaluate Overlay-Based SDN Solutions

By ***Jim Metzler, Ashton Metzler & Associates***
Distinguished Research Fellow and Co-Founder,
Webtorials Editorial/Analyst Division

Introduction and Goal

An overlay-based network virtualization solution typically has an architecture similar to the one shown in **Figure 1**. The main components of the solutions are the controller, hypervisor-resident vSwitches/vRouters and gateways that provide connectivity from virtual networks to traditional network segments; e.g., VLANs, non-virtualized servers, or Internet routers.

Figure 1: Network Virtualization Architecture



Within this class of solution, virtualization is performed at the network edge, while the remainder of the L2/L3 network remains unchanged and doesn't need any configuration change in order to support the virtualization of the network. The most common approach is to perform the encapsulation at the hypervisor vSwitch, which acts as the network virtualization edge (NVE). As a result, overlay NV solutions can generally be implemented over existing networks without requiring any changes.

The controller function is generally supported by a high availability (HA) cluster or another HA configuration. Controller functionality may be comprised of a number of sub-functions running on different servers. Cloud Management/Orchestration is typically obtained from a third party and network services may be integrated with the controller, integrated via virtual appliances, or possibly integrated via physical appliances through the gateway.

The primary goal of this document is to describe criteria that IT organizations can use to evaluate the architecture of an overlay-based network virtualization solution. There are other criteria that IT

organizations should also utilize with evaluating these solutions that are mentioned in this document but not described in detail. That includes the performance, scalability and manageability of the solution.

General Evaluation Criteria

One of the primary criterion that IT organizations should use relative to evaluating an overlay-based network virtualization (NV) solution is how well it solves the problem(s) that the IT organization is looking to solve. For example, can the solution enable the IT organization to move workloads between data centers? Between an enterprise data center and a cloud provider? Between a data center and a branch office?

Other solution level criteria that IT organizations should evaluate include:

- To what degree does the solution federate and hence interoperate with other solutions?
- What interaction, if any, is there between the virtual networks and the physical networks?
- What management functionality is provided into both the virtual and physical networks?
- To what degree does the solution support service chaining?
- What high availability options are possible?
- How scalable is the solution?
- What other functionality (e.g., security) does it provide?

Architecture-Based Evaluation Criteria

The main technical differences between the various overlay-based NV solutions that IT organizations should evaluate fall into the following categories:

- **Encapsulation formats**. Some of the tunneling/encapsulation protocols that provide network virtualization of the data center include VXLAN, NVGRE, STT, and SPB MAC-in-MAC (SPBM). Both the IEEE and the IETF have already standardized SPB. It is unclear as to whether or not all of the other proposals will become standards.

- **Tunnel control plane functionality** that allows ingress (encapsulating) devices to map a frame to the appropriate egress (decapsulating) device. The first-hop overlay device implements a mapping operation that determines where the encapsulated packet should be sent to reach its intended destination VM. Specifically, the mapping function maps the destination address (either L2 or L3) of a packet received from a VM into the corresponding destination address of the egress NVE device. The main differences here are whether a controller is used and the functionality of the controller.

  Some of the initial, controller-less approaches to network virtualization relied on IP multicast as a way to disseminate address mappings. A more common solution is based on a central repository of address mappings housed in a controller.

- **vSwitches supported.** A number of vSwitches are based to some degree on the open source Open vSwitch (OVS)[1], while other vSwitches are of proprietary design. Another point of differentiation is whether the vSwitch is a virtual router as well as being an encapsulating Layer 2 switch. With Layer 3 functionality, a vSwitch can forward traffic between VMs on the same hypervisor that are in different subnets and can be used to implement Layer 3 VNs. Where the

[1] While based on OVS, many vSwitches have implemented proprietary extensions to OVS.

tunneling vSwitch has full Layer 3 functionality, the majority of intelligence can be implemented at the edge of network, allowing the underlay network to be implemented as a simple Layer 2 fabric.

- **Broadcast/Multicast delivery** within a given virtual network. NVEs need a way to deliver multi-destination packets to other NVEs. There are three different approaches that can be taken:

  - The multicast capabilities of the underlay network can be used
  - The NVEs can replicate the packets and unicast a copy across the underlay network to each NVE currently participating in the VN.
  - The NVE can send the packet to a distribution server which replicates and unicasts the packets on the behalf of the NVEs.

- **Protocols.** Another characteristic of controller-based solutions is the choice of the southbound protocols/APIs that are used between the controller and the NVE and the choice of the northbound protocols/APIs that are used between the controller and cloud management systems and hypervisor management systems. If the southbound protocols are standardized, the NVE can potentially communicate with different types of controllers or controllers from different vendors. Some the alternatives here include OpenFlow, BGP, and CLI shell scripts.

  If the northbound protocols are standardized, the controller can be integrated with network services from ISVs or different types of third party orchestration systems. Most overlay-based controllers support a RESTful Web API for integration with cloud management and orchestration systems. With both southbound and northbound APIs the most important question becomes which third party switches, applications, virtual appliances, and orchestration systems have been certified and are supported by the overlay-based network virtualization vendor.

- **VN Extension over the WAN.** VN extension over the WAN can generally be accomplished with most NV solutions. However, in some cases the encapsulation used over the wide area may differ from that used within the data center. Some of the encapsulation techniques used for VN extension over the WAN include MPLS VPNs and two proprietary protocols from Cisco: Overlay Transport Virtualization (OTV) and Locator/ID Separation Protocol (LISP). OTV is optimized for inter-data center VLAN extension over the WAN or Internet using MAC-in-IP encapsulation. It prevents flooding of unknown destinations across the WAN by advertising MAC address reachability using IS-IS routing protocol extensions. LISP is an encapsulating IP-in-IP technology that allows end systems to keep their IP address (ID) even as they move to a different subnet within the network (Location). By using LISP VM-Mobility, IP endpoints such as VMs can be relocated anywhere regardless of their IP addresses while maintaining direct path routing of client traffic. LISP also supports multi-tenant environments with Layer 3 virtual networks created by mapping VRFs to LISP instance-IDs. Inter-data center network virtualization could also potentially be based on Layer 3 vSwitches that support MPLS VPNS and implement network virtualization using RFC 4023 MPLS over IP/GRE tunnels through an IP enterprise network to connect to an MPLS VPN service. SPBM is unique in that it offers extensions over the WAN natively without requiring additional protocols such as OTV or MPLS VPNs.

Appendix: Tunnel Encapsulation

Below is a detailed discussion of some of the primary ways that vendors are implementing tunnel encapsulation.

VXLAN: Virtual eXtensible LAN (VXLAN) virtualizes the network by creating a Layer 2 overlay on a Layer 3 network via MAC-in-UDP encapsulation. The VXLAN segment is a Layer 3 construct that replaces the VLAN as the mechanism that segments the data center LAN for VMs. Therefore, a VM can only communicate or migrate within a VXLAN segment. The VXLAN segment has a 24-bit VXLAN Network identifier. VXLAN is transparent to the VM, which still communicates using MAC addresses. The VXLAN encapsulation is performed through a function known as the VXLAN Tunnel End Point (VTEP), typically a hypervisor vSwitch or a possibly a physical access switch. The encapsulation allows Layer 2 communications with any end points that are within the same VXLAN segment even if these end points are in a different IP subnet. This allows live migrations to transcend Layer 3 boundaries. Since MAC frames are encapsulated within IP packets, there is no need for the individual Layer 2 physical switches to learn MAC addresses. This alleviates MAC table hardware capacity issues on these switches. Overlapping IP and MAC addresses are handled by the VXLAN ID, which acts as a qualifier/identifier for the specific VXLAN segment within which those addresses are valid.

As noted, VXLANs use a MAC-in-UDP encapsulation. One of the reasons for this is that modern Layer 3 devices parse the 5-tuple (including Layer 4 source and destination ports). While VXLAN uses a well-known destination UDP port, the source UDP port can be any value. As a result, a VTEP can spread all the flows from a single VM across many UDP source ports. This allows for efficient load balancing across link aggregation groups (LAGs) and intermediate multi-pathing fabrics even in the case of multiple flows between just two VMs.

Where VXLAN nodes on a VXLAN overlay network need to communicate with nodes on a legacy (i.e., VLAN) portion of the network, a VXLAN gateway can be used to perform the required tunnel termination functions including encapsulation/decapsulation. The gateway functionality could be implemented in either hardware or software.

STT: Stateless Transport Tunneling **(STT)** is a second overlay technology for creating Layer 2 virtual networks over a Layer 2/3 physical network within the data center. Conceptually, there are a number of similarities between VXLAN and STT. The tunnel endpoints are typically provided by hypervisor vSwitches, the VNID is 24 bits wide, and the transport source header is manipulated to take advantage of multipathing. STT encapsulation differs from VXLAN in two ways. First, it uses a stateless TCP-like header inside the IP header that allows tunnel endpoints within end systems to take advantage of TCP segmentation offload (TSO) capabilities of existing TOE server NICs. The benefits to the host include lower CPU utilization and higher utilization of 10 Gigabit Ethernet access links. STT generates a source port number based on hashing the header fields of the inner packet to ensure efficient load balancing over LAGs and multi-pathing fabrics. STT also allocates more header space to the per-packet metadata, which provides added flexibility for the virtual network tunnel control plane. With these features, STT is optimized for hypervisor vSwitches as the encapsulation/decapsulation tunnel endpoints.

NVGRE: Network Virtualization using Generic Router Encapsulation (NVGRE) uses the GRE tunneling protocol defined by RFC 2784 and RFC 2890. NVGRE is similar in most respects to VXLAN with two major exceptions. While GRE encapsulation is not new, most network devices do not parse GRE headers in hardware, which may lead to performance issues and issues with 5-tuple hashes for traffic distribution in multi-path data center LANs. With GRE hashing generally involves the GRE key. One

initial implementation of NVGRE from Microsoft relies on Layer 3 vSwitches whose mapping tables and routing tables are downloaded from the vSwitch manager. Downloads are performed via a command-line shell and associated scripting language.

SPBM: IEEE 802.1aq/IETF 6329 Shortest Path Bridging MAC-in-MAC uses IEEE 802.1ah MAC-in-MAC encapsulation and the IS-IS routing protocol to provide Layer 2 network virtualization and VLAN extension in addition to a loop-free equal cost multi-path Layer 2 forwarding functionality. VLAN extension is enabled by the 24-bit Service IDs (I-SIDs) that are part of the outer MAC encapsulation. Unlike other network virtualization solutions, no changes are required in the hypervisor vSwitches or NICs and switching hardware already exists that supports IEEE 802.1ah MAC-in-MAC encapsulation. For SPBM, the control plane is provided by the IS-IS routing protocol.

SPBM can also be extended to support Layer 3 forwarding and Layer 3 virtualization as described in the IP/SPB IETF draft using IP encapsulated in the outer SPBM header. This specification identifies how SPBM nodes can perform Inter-ISID or inter-VLAN routing. IP/SPB also provides for Layer 3 VSNs by extending VRF instances at the edge of the network across the SPBM network without requiring that the core switches also support VRF instances. VLAN-extensions and VRF-extensions can run in parallel on the same SPB network to provide isolation of both Layer 2 and Layer 3 traffic for multi-tenant environments. With SPBM, only those switches that define the SPBM boundary need to be SPBM-capable. Switches not directly involved in mapping services to SPB service IDs don't require special hardware or software capabilities. SPBM isn't based on special vSwitches, data/control plane separation, or centralized controllers.

About the Webtorials® Editorial/Analyst Division

The Webtorials® Editorial/Analyst Division, a joint venture of industry veterans Steven Taylor and Jim Metzler, is devoted to performing in-depth analysis and research in focused areas such as Metro Ethernet and MPLS, as well as in areas that cross the traditional functional boundaries of IT, such as Unified Communications and Application Delivery. The Editorial/Analyst Division's focus is on providing actionable insight through custom research with a forward looking viewpoint. Through reports that examine industry dynamics from both a demand and a supply perspective, the firm educates the marketplace both on emerging trends and the role that IT products, services and processes play in responding to those trends.

For more information and for additional Webtorials® Editorial/Analyst Division products, please contact Jim Metzler or Steven Taylor .

Published by Webtorials Editorial/Analyst Division
www.Webtorials.com

Division Cofounders:
Jim Metzler
jim@webtorials.com
Steven Taylor
taylor@webtorials.com

Professional Opinions Disclaimer

All information presented and opinions expressed in this publication represent the current opinions of the author(s) based on professional judgment and best available information at the time of the presentation. Consequently, the information is subject to change, and no liability for advice presented is assumed. Ultimate responsibility for choice of appropriate solutions remains with the reader.