September 07, 2018

FLOATING/PERVASIVE LAYER 3 OUTSIDE TO PEER WITH VIRTUAL ROUTERS IN DATACENTER

Umamaheswararao Karyampudi

Himanshu Mehra

Rajagopalan Janakiraman

Nicolas Vermande

Minako Higuchi

FLOATING/PERVASIVE LAYER 3 OUTSIDE TO PEER WITH VIRTUAL ROUTERS IN DATACENTER

AUTHORS:
Umamaheswararao Karyampudi
Himanshu Mehra
Rajagopalan Janakiraman
Nicolas Vermande
Minako Higuchi

ABSTRACT

Techniques are described herein for a Floating Layer 3 Outside (L3Out) mechanism that enables an Application Centric Infrastructure (ACI) datacenter fabric to peer with Virtual Routers that can move across hypervisors. This may be performed without losing connectivity in protocol sessions, almost zero packet loss, and no extra configuration. These techniques save hardware resources with respect to Internet Protocol (IP) address and policy Content Addressable Memory (CAM) usage with no extra provisioning on the ACI.

DETAILED DESCRIPTION

Application Centric Infrastructure (ACI) today provides Wide Area Network (WAN) connectivity to datacenters via Layer 3 Outside (L3Out) on a Border Leaf in the ACI fabric. Peering is performed either with a Physical Router or a Virtual Router behind the Border Leaf. If the Virtual Router moves across hosts in a Virtual Machine (VM) mobility domain that may not be directly connected to the Border Leaf, WAN peering/L3Out functionality cannot be supported.

Currently, if the Next Hop (NH) from the fabric is a Virtual Router or Firewall (FW), an explicit External L3Out (L3Ext) path must be configured, pointing to the active uplink of the hypervisor where the virtual function resides. When hypervisor resources are aggregated into clusters, it is not guaranteed that the router VM always runs on the same host. Therefore, all possible L3Ext logical interfaces leading to all hypervisor facing ports must be configured so the routing function can be maintained upon VM move. This is not scalable and is very cumbersome to configure as new Virtual Network Functions (VNFs) are deployed.

VM move detection and provisioning Virtual Local Area Networks (VLANs) on ports is handled by the Virtual Mobility (VMM) scheme in ACI. However, if the VMs represent a router, there are other problems with respect to maintaining control sessions, extending the gateway functionality without burning additional addresses, and maintaining forwarding and policy states for destinations behind the Virtual Router.

Figure 1 below illustrates the traditional approach.

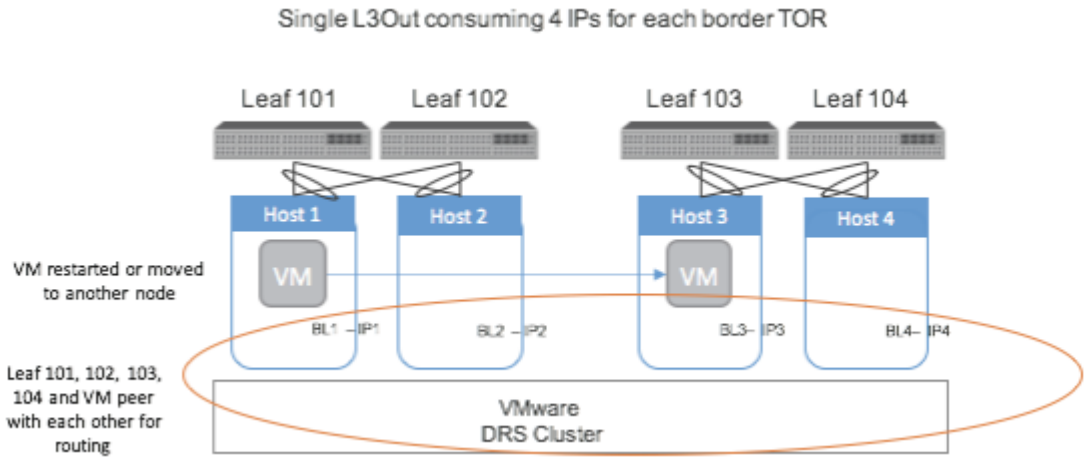

Figure 1

In a virtualized environment in which the WAN Router is a VM that can move from one host to another (from the ACI's point of view, it will move from one port on the Border Top-of-Rack (TOR) to another port on another Border TOR), this technique cannot be applied without wasting resources such as VLANs, Internet Protocol (IP) addresses, security policies, and route tables on all the TOR switches behind which the VM can potentially move. Routers as VMs can migrate to different hypervisors based on resource management algorithm decisions of the corresponding orchestrators. While the user can configure L3Out on each TOR port which is connected to the hypervisors where the VM can move, it will consume resources on each TOR unnecessarily and the configuration management becomes cumbersome as the dynamic event on the VM orchestrator needs to translate into a static/user L3Out configuration on the ACI fabric.

Accordingly, a method is described to support peering with a Virtual Router that is behind any ACI Leaf with a construct called pervasive/floating L3Out that saves IP addresses as well as enables an optimized traffic path while retaining the Routing Control Sessions as the Virtual Router moves across hosts. The hosts themselves may be connected

anywhere in the ACI fabric within or across PODs. This seamlessly allows Virtual Router movement without manually reconfiguring the ACI fabric.

The dynamic L3Out deployment scheme described herein anchors the Gateway functionality for a control plane session in a pair of Border TORs while extending the forwarding plane and security policies on demand to any other TOR in the ACI fabric based on the VM router discovery/motion. This scheme also provides shortest path forwarding to the VM from/to anywhere in the ACI fabric. This alleviates the configuration complexity involving static L3Out configuration and conserves resources effectively and on-demand.

Use cases for the L3Out tracking mechanism include virtual FW/routers hosted in a hypervisor cluster, virtual FW/router high availability, maintenance mode, disaster avoidance, and a Layers 4-7 (L4-7) service graph with a virtual security appliance. With respect to a virtual FW/router hosted in a hypervisor cluster resource, scheduling is dynamically managed. The VM hosting boundary is the cluster itself, not a single host. With respect to virtual FW/router high availability, on top of FW redundancy mechanisms, a hypervisor high availability mechanism may restart the FW VM on any available host within the hypervisor cluster.

With respect to maintenance mode, when the hypervisor needs to be upgraded, the VM administrators “evacuate” the host, performing live migration of the FW/router VM to another host in the hypervisor cluster. With respect to disaster avoidance, in a stretched cluster, outages are expected on some nodes. VMs are moved to hosts that are not expected to experience the outage. With respect to a L4-7 service graph with a virtual security appliance, the virtual security appliance may be “pinned” to a single L3Out path. L3Out tracking may allow the virtual security appliance to participate in a service graph and dynamically reconfigure its L3Out path attribute for all the above use cases.

Resources such as policy table entries (security contracts) and forwarding tables are required for Virtual Router as well as routes learnt behind the Virtual Router. Since the Virtual Router may represent a large routing domain (especially in a Service Provider (SP) datacenter environment), savings may be significant if the hardware is programmed only on the TOR to which the Virtual Router is currently connected. These features may benefit all cloud SPs as well as enterprise private clouds where VNF is a requirement.

The L3Out tracking feature may provide a high capability level when using East/West (E/W) or North/South (N/S) non-distributed router/FWs (i.e. supporting the VM movement). Currently, it is recommended to enable disaster recovery systems / high availability on clusters with anti-affinity rules so that pairs of Virtual Routers are never on the same host, but are still able to move or be restarted. The L3 boundary may not be at the TOR. Therefore, a single Switch Virtual Interface (SVI) at the Aggregator/Spine is sufficient to peer with the Virtual Router. This may give the false impression that the current way to implement NH VNF in ACI is complicated.

Figure 2 below illustrates two options for a solution. The first option is the possibility to link L3Out to a VMM domain to abstract the notion of static path and rely on VMM information to track VM location. The second option is to dynamically change the L3Out path attributes upon VM End Point (EP) move detection.

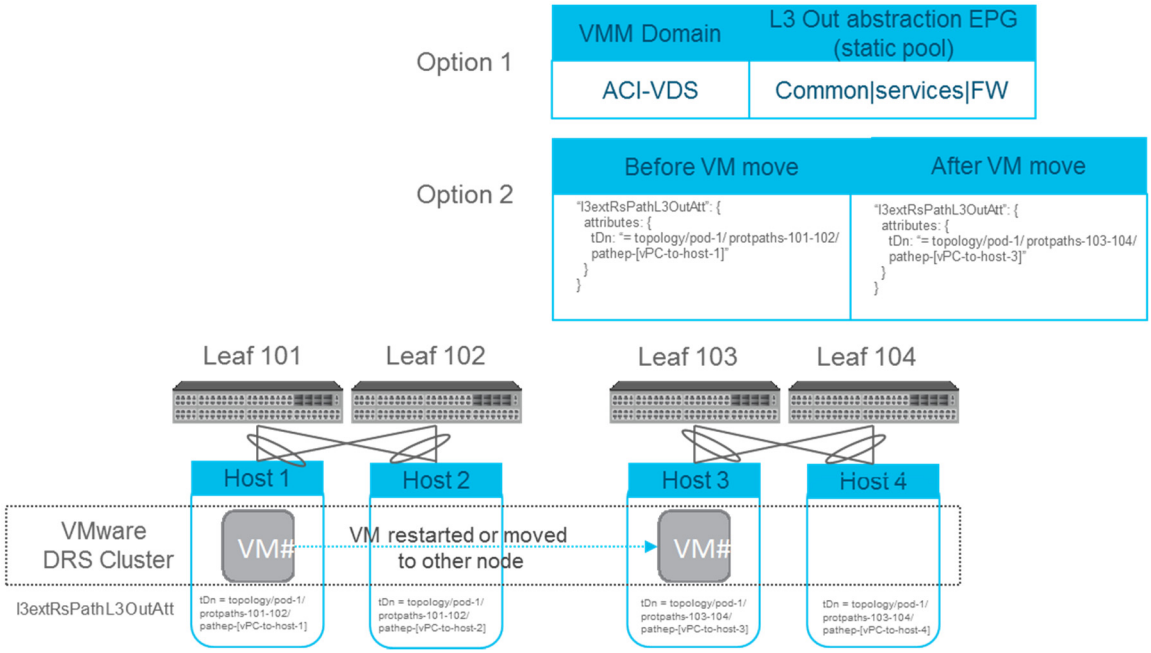

Figure 2

This solution may represent a Virtual Router domain by a L2 segment (e.g., VLAN, Virtual Extensible LAN (VXLAN), or a subnet that can span the entire ACI fabric or a particular set of nodes and PODs as defined by the user). Detection of the Virtual Router behind any Leaf in the ACI Fabric may be based on listening for VM creation/move events from the orchestration agent or based on traffic from the Virtual Router. For the scheme based on data traffic detection, the mismatch of VLAN in the data traffic and the VLANs

opened on the port results in traffic being punted to the Central Processing Unit (CPU) for inspection, which leads to Virtual Router detection.

A primary or anchor Border Leaf pair may be used in the ACI fabric for a given Virtual Router domain. The primary Leaf runs control plane sessions with the Virtual Router irrespective of where the Virtual Router connects in the ACI fabric. This ensures that control plane sessions such as Border Gateway Protocol (BGP) and Open Shortest Path First (OSPF) stay up during the move without any side effect.

When the virtual router is discovered on a new ACI TOR switch, the L2 segment and Gateway functionality (SVI) is stretched to the new TOR by use of a common "secondary IP address" for the span of the Virtual Router domain. The presence of the first virtual router on the new TOR may also create the forwarding domain or Virtual Route Forwarding (VRF) in the switch. The virtual router may reach other virtual routers and anchor Border Leafs as well other hosts in the same L2 segment/subnet through an optimized external bridge domain scheme. The scheme uses Address Resolution Protocol (ARP) based detection and gleans using the secondary IP address. Learning the virtual router may be synchronized with other TORs having membership in the Bride domain through this mechanism.

WAN routes may be synchronized between TORs within the ACI fabric through Multiprotocol BGP (MP-BGP). The NH of the routes learnt through MP-BGP is the VxLAN Tunnel EndPoint (TEP) address of the TOR switch, which peers with the WAN router. Shortest path forwarding is enabled from other TORs for traffic destined to destinations behind the Virtual Router, irrespective of the attachment of the Virtual Router. Detection of the Virtual Router behind a new TOR triggers the NH change in MP-BGP for the routes learnt from the Virtual Router so that traffic destined to the Virtual Router are not forwarded to the old Leaf. This may be performed through a method that involves the routes learnt from the Virtual Router.

In this method, the NH of the routes are changed to the TEP address of the new TOR (where the Virtual Router is attached) instead of the anchor TOR which actually runs the BGP session. This NH is advertised to the other TORs through MP-BGP. This scheme operates well, but may cause churn in BGP when the Virtual Router moves around, especially when a large number of routes are learnt behind the virtual router.

In the reverse direction, traffic originated from/behind the Virtual Router is forwarded from the new TOR itself, which acts as the first hop router through the use of a common router Media Access Control (MAC) address in the virtual router domain. Routes within the fabric may be distributed to all the switches which are part of the same VRF through MP-BGP. Since the detection of the Virtual Router triggers VRF creation dynamically, it also triggers the route pull mechanism for moving the other interesting routes into the new TOR through MP-BGP sessions.

Finally, the detection of the Virtual Router behind a new TOR also triggers an automatic policy pull of WAN IP prefix to End Point Group (EPG) mapping (or "External EPG") and the corresponding contract rules for the pulled EPGs. This is necessary for applying the security rules for traffic from destinations behind the Virtual Router on the new TOR itself to save on fabric bandwidth. In the "egress mode" option (where the policy table is saved by applying policy only in the egress TOR), the rules on the new TOR may be required for the reverse direction traffic as well. In any case, the policies are pushed for the Virtual Router and the destinations behind the Virtual Router based on the detection of the Virtual Router. When the Virtual Router moves, the policies are removed from the old router.

The set of External EPGs and contracts that are pulled into the new TOR may be decided by one of two schemes. The first scheme involves a user configuration of prefixes and EPGs in the L3Out. The L3Out itself is marked as a "floating L3Out," thereby making all the EPGs in the L3Out floating EPGs for the span of the Virtual Router domain. This scheme works well where one L3Out represents one Virtual Router domain where the user specifies the prefixes or the aggregates of the prefixes for the routes learnt behind that Virtual Router domain.

When an L3Out represents multiple Virtual Router domains, these techniques may be made even more efficient by pulling prefixes corresponding only to the Virtual Router which is currently attached to the new TOR. In this case, the IP prefix to EPG mapping may be pulled to the new TOR, only if the corresponding prefix is learnt through a routing session with the Virtual Router. This coupling of the policy pull with routing table results in significant savings when the prefixes can be learnt from one of multiple Virtual Routers.

For application EPs, there are no peering sessions that need to be moved. Furthermore, there is no need to represent and move routing and policy states of other EPs which are behind the application EP. These are problems unique to router EPs and are resolved through the Floating L3Out solution. When the EP is a router and is moving, these techniques solve the challenges of maintaining routing peer sessions and at the same time extending shortest path forwarding and optimizing forwarding table and policy usage with no additional user configuration.

Different sets of ACI Border Leaves may be anchored for different sets of Virtual Routers, thereby providing control plane load balancing. This may be provided natively in the ACI fabric and transparent to the user when the IP address of the Virtual Routers are provisioned dynamically through Dynamic Host Configuration Protocol (DHCP). The first hop TOR snoops on the DHCP request and updates the primary subnet and DHCP relay address to one of the Anchor TOR SVI addresses through round robin fashion or through a more sophisticated control plane session and routing scale based scheme.

In ACI fabric, the VMM domain is created where the Virtual Router can reside. This VMM domain may span many Leafs. Two Leafs may be pinned as Border Leafs in a redundant configuration and L3Out configuration may be deployed. This deploys the L3 bridge domain (L3 SVI) with primary and secondary addresses. Both Border Leafs have the same SVI extended. Each Leaf has its own primary IP address for the SVI but the same secondary IP address. Both Leafs have the same SVI MAC (router MAC) address for the SVI.

The Virtual Router can reside behind any ACI Leaf at any point. However, the Virtual Router may only peer with the Border Leaf to run routing protocols. On any other Leaf, when the controller determines where the Virtual Router is, it extends the L3Out of the Border Leaf with same secondary IP address but no primary IP address. However, the SVI MAC address is same as that of the Border Leaf SVI MAC address. This ensures that all the interfaces are part of the same bridge domain so that any broadcast packet is received by all the interfaces.

Since Peering protocol is run between the Border Leaf and the Virtual Router, all other ACI Leafs learn the routes via a route reflector running in the spine. All the security policies associated with L3Out of Border Leaf are deployed onto non-Border Leafs when

the Leafs are deployed. Since all SVI MAC addresses are same, any packet sent by the Virtual Router to the ACI fabric may be routed by the first node at which the packet is received. The same Leaf also applies the security policy and routes the traffic to its final destination. Each Leaf (either Border Leaf or non-Border Leaf) may use the secondary IP address whenever it needs to glean information for a neighbor in order to generate the information.

Upon receiving an ARP response from the neighbors, the response may be flooded in the same SVI VLAN so that all Leafs learn the neighbors at the same time. When the Virtual Router moves, it sends a Gratuitous ARP (GARP) message which may be processed by all Leafs, and updates its routing database for that neighbor with the new Leaf behind which it is moved. Also, when the routes are advertised from a peer which is currently attached to a different switch, the fabric BGP session changes the next hop for all the routes learnt to the switch to which the host is currently connected. This ensures that any traffic destined to the host is directly delivered to the current switch without having to make two hops in the fabric.

Figures 3-7 provide solution details by illustrating an example packet walk. Figure 3 below illustrates an example topology.

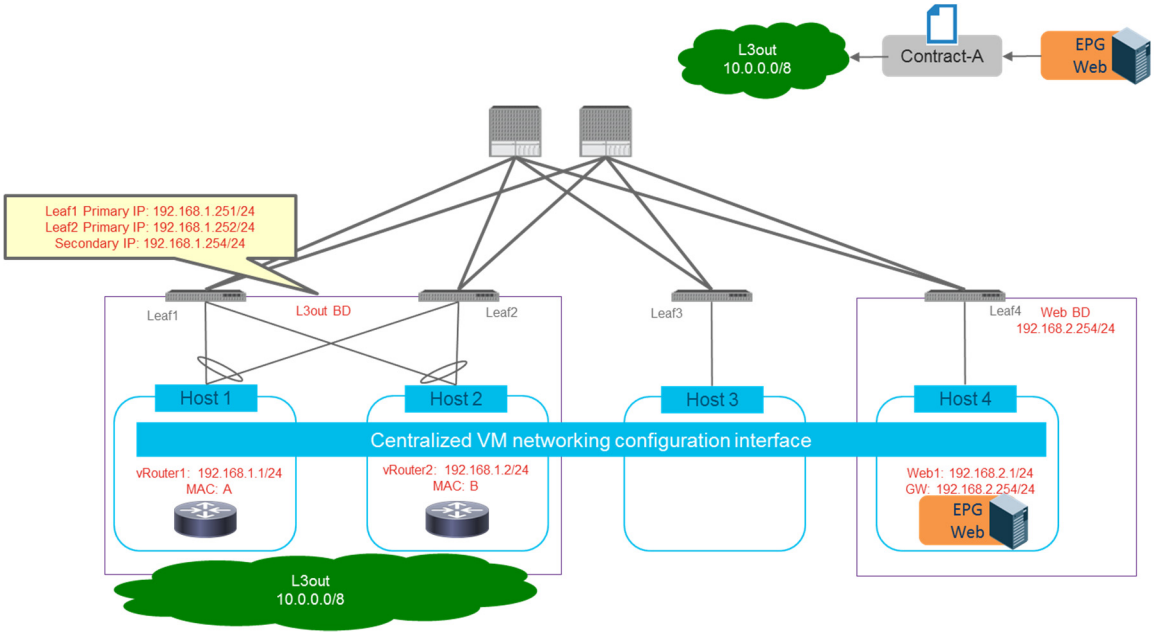

Figure 3

Figure 4 below illustrates an example peering status.

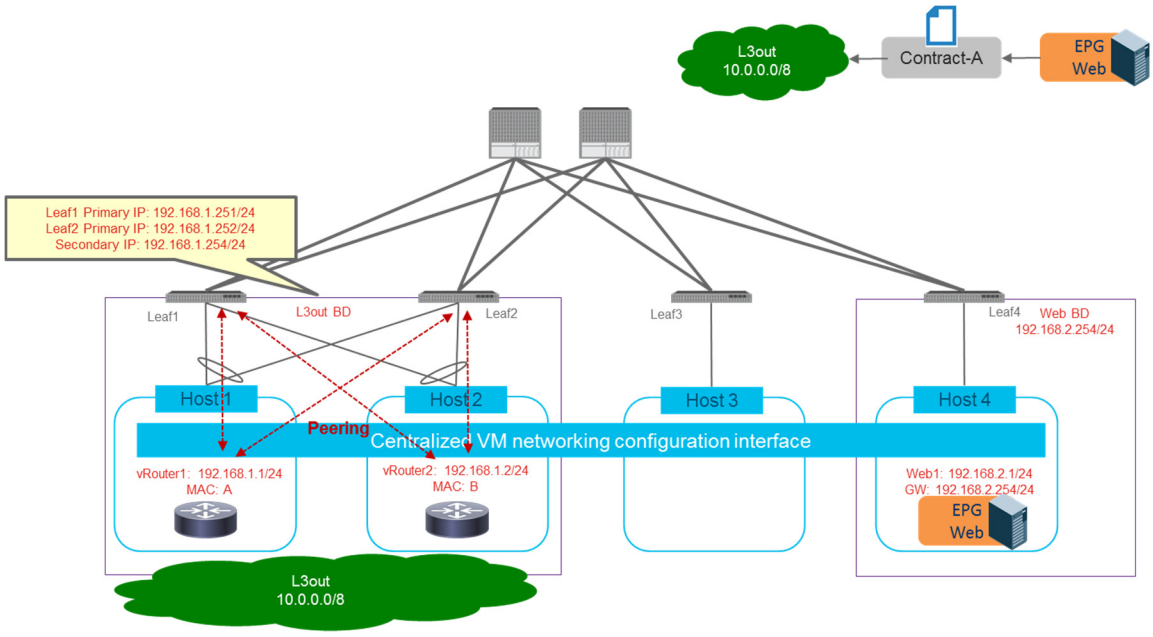

Figure 4

Figure 5 below illustrates an example traffic flow from L3Out to the Web.

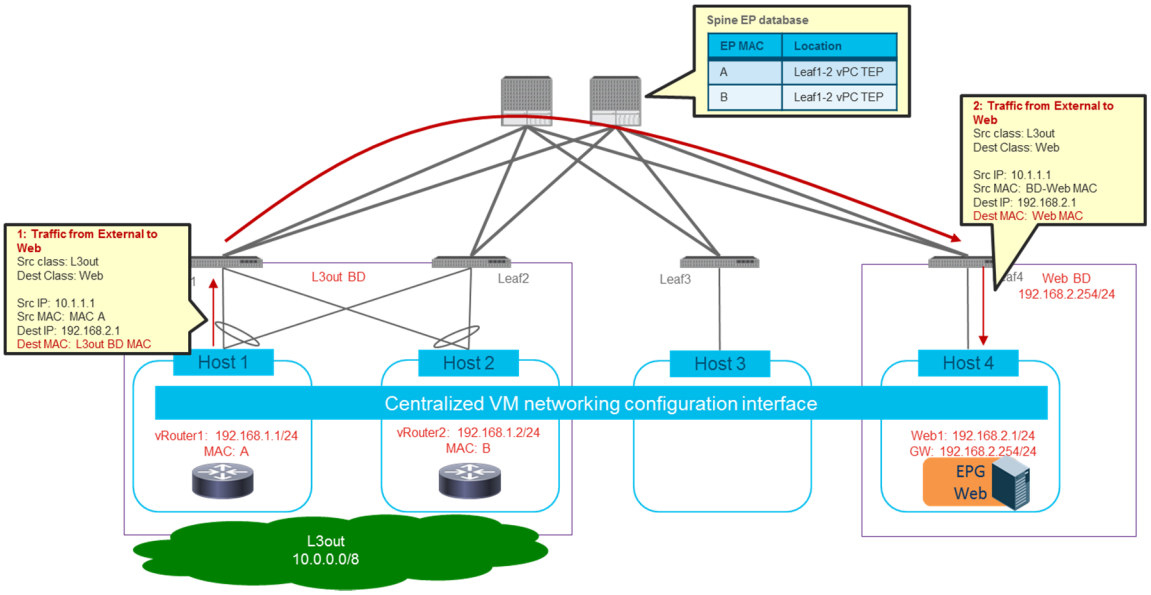

Figure 5

Figure 6 below illustrates an example traffic flow from the Web to L3Out.

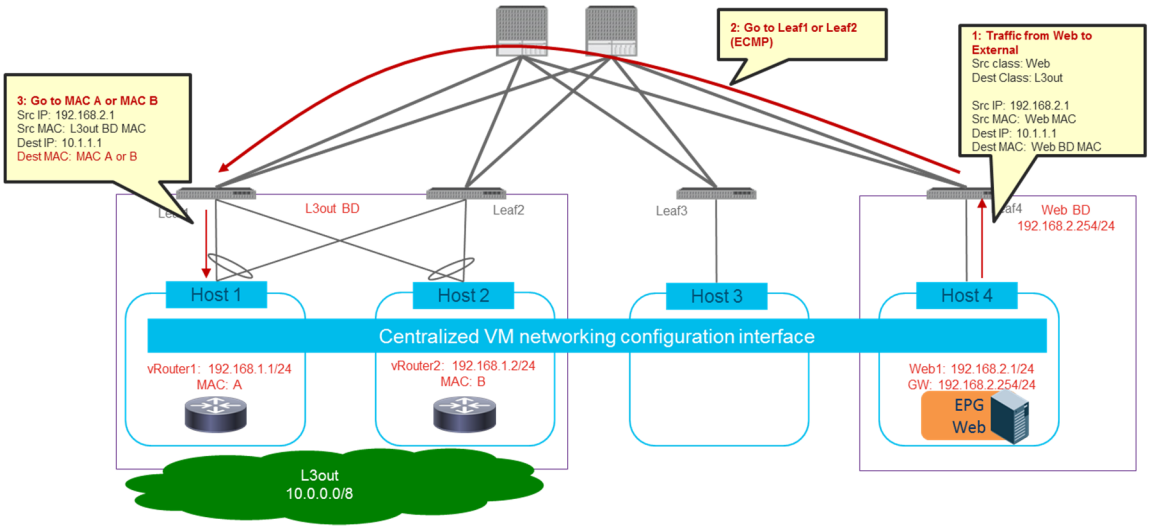

Figure 6

Figure 7 below illustrates an example in which vRouter1 moves to under a different Leaf EP via GARP. In this example, the immediate option is recommended. Before the VM moves, a notification is received from a virtual center. If there is no problem, the new location and switch program policy are known.

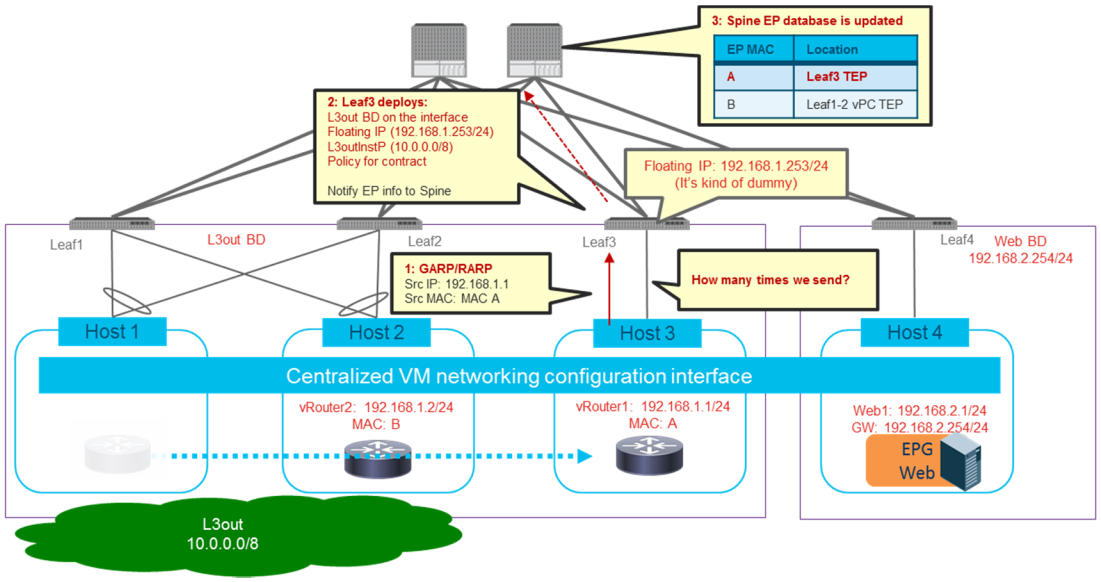

Figure 7

When the VM moves, switch deployment may occur. To program SVI, a node and at least one port are needed. Although immediate, if the TOR does not have the actual interface connected to the VM in the L3Out bridge domain, the SVI is not programmed.

The techniques described herein may utilize a non-floating IP option or a floating IP option. For the non-floating IP option (ARP flooding), flooding may occur, but probably

is not a major problem. Since there is no propagation to the front panel ports, it can be filtered and the move would not occur often. For the floating IP option (ARP glean), an IP lookup may be performed for ARP. There is no ARP flooding, and ARP glean is used. The IP address is needed, but using the secondary IP address may be a good option.

If GARP or Reverse ARP (RARP) is missed, no traffic may come from the vRouter. If there is traffic from L3Out (outside) to the Web, this traffic may cause leaf3 to be learned for MAC A. That said, if traffic is always initiated by the Web to L3Out, there is a possibility of a black hole until ARP tracking occurs again.

Figure 8 below illustrates a peering status example in which vRouter1 moves to under a different Leaf.

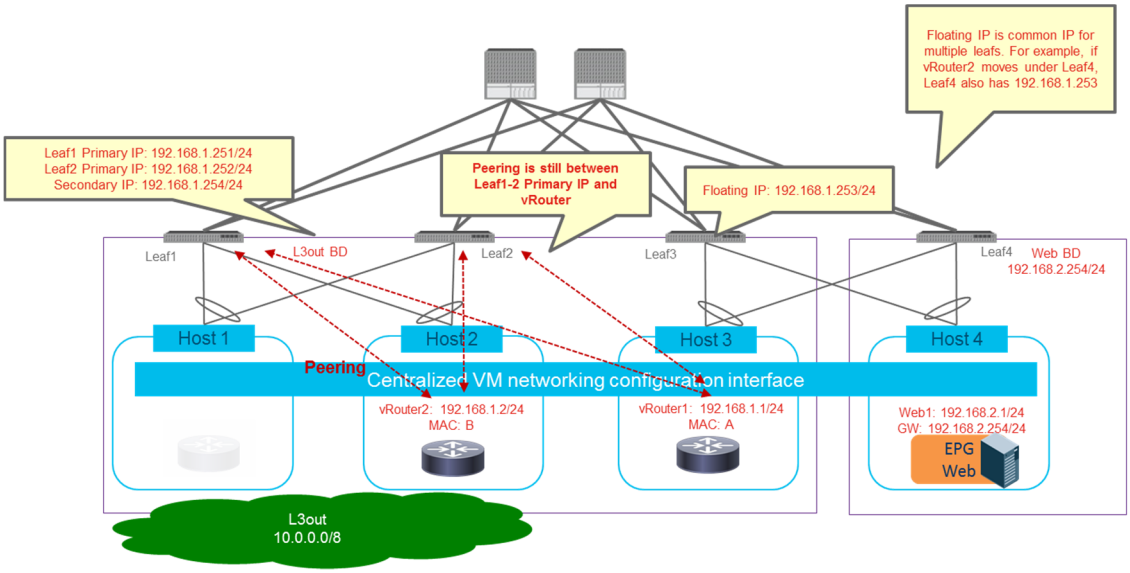

Figure 8

Figure 9 below illustrates a traffic flow (L3out to the Web) example in which vRouter1 moves to under a different Leaf.

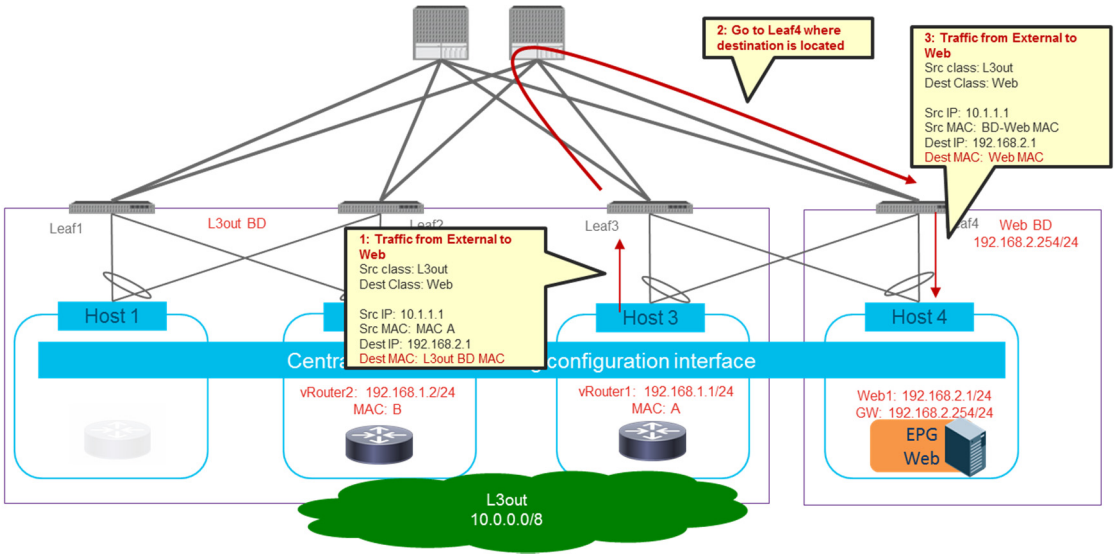

Figure 9

Figures 10 and 11 below illustrates a traffic flow (the Web to L3Out) example in which vRouter1 moves to under a different Leaf.

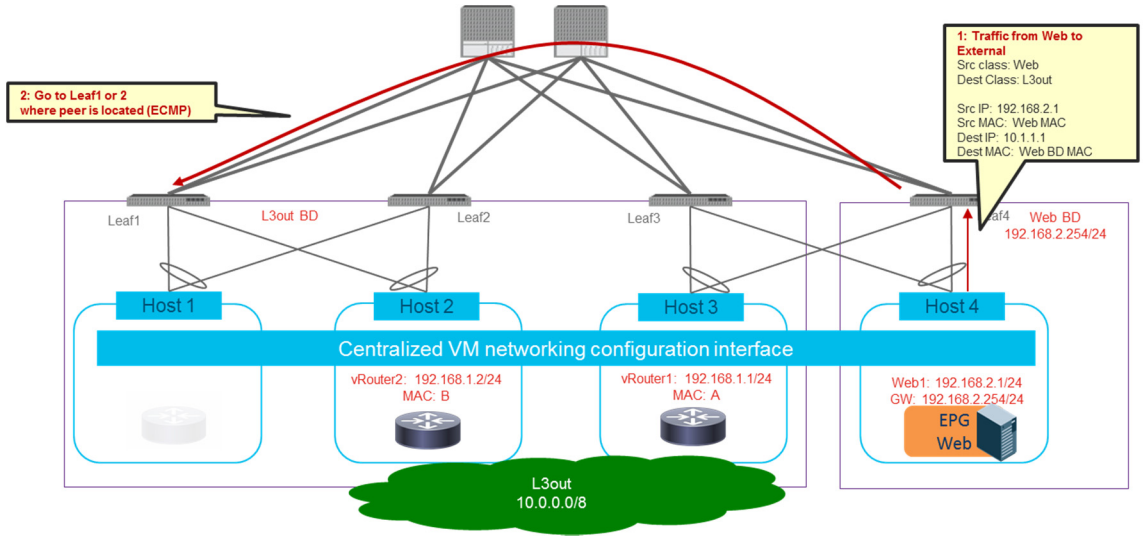

Figure 10

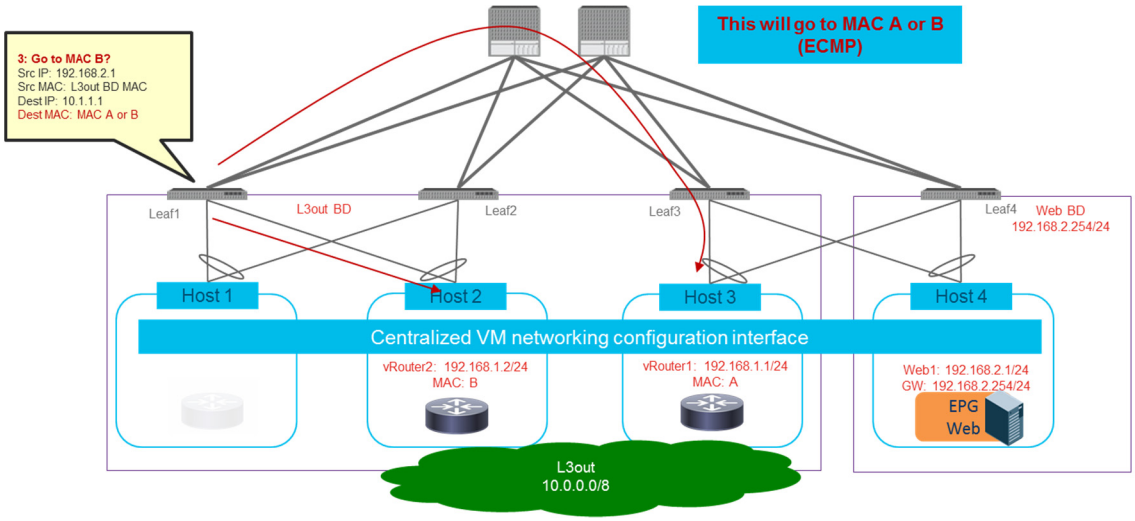

Figure 11

For L3Out, the domain may be a VMM domain. L3Out Node Leafs1-2 may be primary Border Leafs. The L3Out interface profile may be provided in certain examples. A floating IP configuration may be utilized.

These techniques provide several advantages. First, the solution is implementable with only software changes in the ACI fabric, and may work with any Virtual Router. Second, this solution is operable at high scales as it is realized in hardware with an optimized use of resources and shortest path forwarding. Third, this solution combines the fabric view of the ACI controller with orchestration and specific feature usage of components such as BGP next hop, ARP gleaning, and VxLAN data plane TEP, which are supported in any standards based implementation. This solution may be applied to an ACI fabric or any other overlay based data center fabric.

In summary, techniques are described herein for a Floating L3Out mechanism that enables an ACI datacenter fabric to peer with Virtual Routers that can move across hypervisors. This may be performed without losing connectivity in protocol sessions, almost zero packet loss, and no extra configuration. These techniques save hardware resources with respect to IP address and policy Content Addressable Memory (CAM) usage with no extra provisioning on the ACI.