Building IP-Clos Data Centers: the GRNET case

Michalis Mamalis (mmamalis@noc.grnet.gr) Network Engineer	Dr. Christos Argyropoulos (cargious@noc.grnet.gr) Network Engineer
Greek Research & Technology Network (GRNET) NOC 7, Kifisias Av., Ampelokipi, 11523 Athens, Greece	

Keywords: Data Center, IP-Clos, EVPN, VXLAN. Ansible

Abstract: Layer 2 has traditionally been a weakness in data center design in terms of availability and scale. IP-Clos or IP Fabric based DCs answer those limitations by providing a design with nothing more than Layer 3 protocols while introducing in the process, new features like active active server multi-homing as well as addressing the 4k VLAN limitation. In this paper we will present the case of GRNET: building 3 DCs based on IP-Clos to house our computing and storage services while relying on Ansible for initial deployment and configuration management.

IP Clos provides a scalable solution for deploying Medium to Large Scale Data Centers. In GRNET we have used IP Clos technology to provide networking services to more than 60 racks distributed in 3 Data Centers. Before we dive into the magic of IP Clos let's start with the drive behind our decision to go with this particular solution over all other available alternatives.

The main drive behind our decision is to avoid the use of Spanning Tree in the DC, but by doing so to not end up locked in any vendor specific solution, a requirement that IP-Clos indeed meets. Another advantage in comparison to traditional, switching based architectures (i.e. Core, Distribution, Access), is that IP Clos solves the upper limit on available VLANs inside the DC; VLANs do not play a significant role in an IP Clos DC and that comes from the fact that traffic that is received on the first building block of an IP Clos topology, the leaf switch, is encapsulated into VxLAN packets identified by Virtual Network Identifiers (VNI). By adding a 24 bit VxLAN header to the packet IP Clos manages to move the theoretical limit to 16 million VxLANs inside the DC.

Instead of any vendor proprietary or vendor tweaked protocol, IP Clos relies on the use of MP-BGP, a well-known and established protocol, to build the control plane. To accomplish this, a new address family is introduced to the protocol: family EVPN that must be supported by all devices in the topology. In this way the newly learned MAC addresses are exchanged over BGP NLRIs in the control plane without putting any burden to the forwarding plane of the switches. MP-BGP also comes very handy in the case where we want to interconnect two or more DCs located in different geographical locations. So, by going through with this decision we have actually dropped any vendor specific solution for both intra and inter DC connectivity.

The network topology follows the Clos model, which is actually quite old and very popular in the field of telecommunications, for our specific needs the model consists of just two layers, namely the Spine and the Leaf layer that together with the access interfaces results in a 3-Tiered Clos model. The leaf layer provides the connectivity to the servers which can be either Bare Metal Servers (BMS), or hypervisor based servers. Also, any other networking device we wish to connect to the IP Clos can be connected through the leaf layers. The leaf layer in turn is connected to the Spine layer over multiple high-speed links that provide redundancy and equal cost multipath load balancing over L3 IP links. Based on the implementation and specifically on the chipset limitations, the leaf layer can either terminate the L3 GW for the servers or act as L2 only capable device, in which case the L3 is terminated on the Spine layer. The connections between the spine and leaf layer are high speed (40G) L3 only links.

Each leaf switch first has to establish VxLAN tunnels with every other leaf, with which it wishes to share a Broadcast Domain. To do that it needs to find a path towards the loopback address of the other leaf switch. This task is accomplished over the underlay network which is responsible for distributing the VxLAN Tunnel Endpoints (VTEP) to both the leaf and spine layers. In GRNET we chose EBGP as the underlay network routing protocol. In this way each leaf switch has one EBGP session with each spine switch over a /31 link. Over those two sessions leaf-1 advertises its loopback to the two spine devices that comprise the spine layer which in turn advertise the loopback over to leaf-2. In this way, leaf-2 ends up with 2 copies of leaf-1 loopback address which together with EBGP multipath that is enabled on the sessions enables us to fully utilize both high speed links of every leaf to the spines and achieve a first level of load balancing.

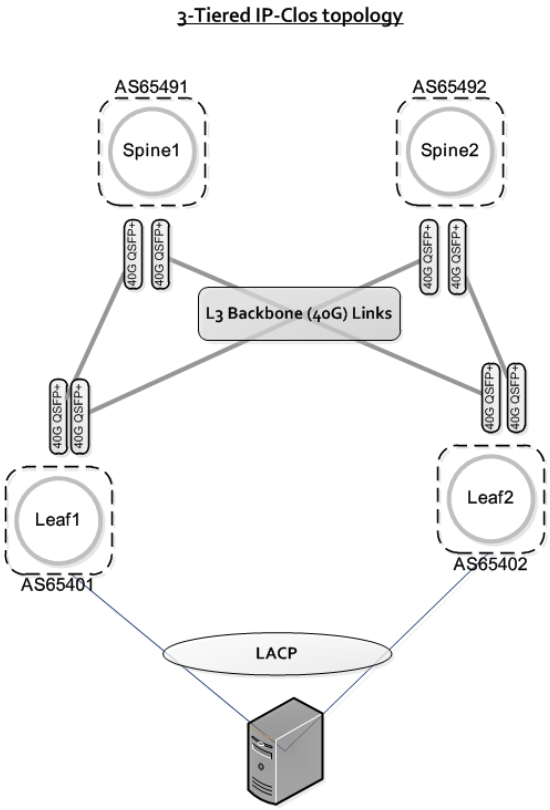

Figure 1:.3-Tiered IP-Clos topology depicting server multi-homing and leaf layer multiple links to spines.

Once the underlay network is in place and each leaf device knows a way to reach every other VTEP that is located either on a leaf or on a spine, the next step is to enable server to server communication. IP Clos relies for this task to the overlay network. As already mentioned the overlay network is actually a full mesh iBGP over all IP Clos devices. By exchanging several EVPN route types each iBGP speaker in essence translates a MAC address, as it arrives on the access ports of the leaf layer to an EVPN route type and distributes them to the other iBGP speakers.

The choice of EVPN in the control plane allows for another equally important design choice, each device/server connects to the IP Clos with active active multi-homing to two leaf switches without the leaf switches having any direct connection with each other. The server implements LACP with the switches that rely on their control plane via EVPN to synchronize all MAC addresses, avoid loops, handle BUM traffic and load balance traffic towards the server.

Another important building block to the DC is the location of the L3 gateway. Based on certain hardware choices that we have made, the L3 gateway lives on the spine layer and is actually shared between the two spine switches by enabling the statement: virtual-gateway. Virtual gateway statement signals to the EVPN control plane that we wish

to share the same MAC and IP address between the two spines. This MAC address is actually advertised by both spine switches to the server via ARP, thus providing another potential level of load balancing.

The 3-Tiered topology we have come up with comprises of a large number of network devices that share the same control plane but are absolutely distinct in terms of management plane. So, have we managed to get rid of one (or several) problem(s) while introducing a new and different one? That is indeed the Achilles's heal of the IP Clos solution and to tackle that we have turned to a relevant new management tool we use in GRNET and is no other than Ansible. We have built several new playbooks that not only build an IP Clos topology from the ground up, but can also provision an entire new service to the IP Clos or can just add a new VLAN to the network. All the relevant data, required to build a new service is mitigated to YAML files that are called upon each time we fire an Ansible playbook. Our new Data Centers already host a number of services delivering multiple Gigabits of traffic over the IP Clos topology.

Vitae

Michalis Mamalis received his Diploma in Electrical and Computer Engineering from the Democritus University of Thrace in 2002. His diploma thesis focused on the design and implementation of an MMIC board acting as the optical receiver part of an STM-4 optical link. His areas of interest include MPLS, routing protocols, IPv6, policy based networking, network virtualization. Currently he is with the network administration and operations team of GRNET NOC.

Dr. Christos Argyropoulos received the Diploma in Electrical Engineering and Computer Science from the University of Patras and the Ph.D. degree in Electrical Engineering from the National Technical University of Athens (NTUA). He worked as a research associate to the Network Management and Optimal Design Laboratory (NETMODE) at NTUA participating in several European FP7 projects (NOVI, GÉANT GN3 and GN3+ etc.). His main research interests lie in the area of computer networks with emphasis on network virtualization and software defined networking. Currently he is with the Greek Research and Education Network (GRNET).