Network Working Group
Internet-Draft
Intended status: Informational
Expires: September 18, 2016

M. Bagnulo
UC3M
D. Dolson
Sandvine

March 17, 2016

NFVI PoP Network Topology: Problem Statement
draft-bagnulo-nfvrg-topology-01

Abstract

This documents describes considerations for the design of the interconnection network of an NFVI PoP.

Status of This Memo

This Internet-Draft will expire on September 18, 2016.

Copyright Notice

Table of Contents

1. Introduction

An NFVI PoP is defined as a "single geographic location where a number of NFVI-Nodes are sited" where an NFVI-Node is "a physical device deployed and managed as a single entity providing the NFVI functions required to support the execution environment for VNFs" [ETSI_GS_NFV-INF_001]. In other words, an NFVI PoP is the premises where the processing, storage and networking resources (i.e., servers and switches) used to execute the network virtual functions (VNFs) are deployed. The servers and switches in a NFVI PoP will be interconnected forming the NFVI PoP interconnection network. The goal of this document is to explore the different design considerations for the NFI PoP interconnection network topology, including design goals and constrains.

The NFVI PoP is essentially a data center, and the NFVI PoP interconnection network is essentially a data center network. As such it is only natural to use the current state of the art in data

center networking as a starting point for the design of the NFVI PoP network.

2. Considerations for the design of the NFVI PoP network topology

This section describes different pieces of information that are relevant input for the design of the NFVI PoP network topology. In some cases, the information is known (and sometimes ready available), while in other cases, the information is not known at this stage.

2.1. External links

The NFVI PoP is part of the operator's infrastructure and as such it is connected to the rest of the operator's network. Information about the number of links and their respective capacity is naturally a required in order to properly design the NFVI PoP topology. Different types of PoPs have different number of links with different capacity to connect to the rest of the network. In particular, the so-called "local PoPs" that connect the links from end users (either DSL lines or FTTH or else) and also connect to the rest of the operator's network. The "regional" PoPs or "regional data centers" have links to the "local PoPs" and to other "regional PoPs" and other parts of the operator's infrastructure.

For instance, a local PoP in a DSL access network can have between 15.000 and 150.000 DSL lines with speeds between 10 Mbps and 100 Mbps and tens of links to the core network of the operator with links between 20 Gbps and 80 Gbps.

It would be useful to confirm these numbers and to have information about other types of PoPs.

2.2. Number of servers

While knowing the exact number of servers is not required to design the PoP network topology, knowing the order of the number of servers is at least useful. If the resulting topology have tens of servers, then the topology is likely to be be very simple (e.g., a tree-like topology with access/aggregation/core switches may be suitable). On the other hand, if the topology should encompass several hundreds of servers of even a few thousands of servers, then the problem is more challenging as we are likely to reach the available capacity of existing switches and more sophisticated topologies may be required.

The number of servers on a PoP depends on several factors, including the number and capacity of external links (i.e., the offered load to the PoP), the number and type of Virtual Network Functions that will be provided by the PoP, the performance of the VNF implementations

and the number and length of service function chains that will be provided.

The number of external links is discussed in the previous section. The number and capacity of the external links is relevant to determine the number of servers because they will carry the load offered to the PoP. In other words, traffic coming through the external links will require processing by the different VNF hosted in the servers, influencing the number of servers needed.

The number of different VFNs provided in the PoP as well as the number and length of service functions chains provided in the PoP will also influence the number of servers required in the PoP. The more demanding the VNFs provided, the more servers will be needed to provide it and the longer the service function chain a higher number of servers will be required to support it.

Finally, the performance of the NFV implementations also affects the number of servers required in a PoP. In particular, some VNF implementations are capable of processing at line speed, while other implementations of other VNFs are not capable of that, requiring additional servers to provide the VNF for the same line speed. While there is some initial work assessing the performance of the different VNFs (e.g., [swBRAS]), still more work is needed to have a full picture for the different VNFs at different line speeds.

Overall, we need to have a rough estimate of the range of the number servers that will be part of the PoP network in order to provide a successful design and we need to take into account the aforementioned considerations to obtain it.

2.3. Traffic patterns

The pattern of the expected traffic of the NFVI PoP network is of course essential to properly design the network topology. In this section we describe different characteristics of the traffic pattern that we believe are relevant and that it would be useful to have information about.

2.3.1. Macroscopic behaviour

There are essentially 4 types of traffic direction within a NVFI PoP network, namely, cross-through traffic, intra-PoP traffic, PoP-generated traffic and PoP-terminated traffic, as depicted in Figure 1.

```
           external   +---------------------------------+    external
            links     |          NFVI PoP               |     links
      ----------------|                                 |---------------
      ----------------|                                 |---------------
       >-->----->----->--cross-through traffic----->---->------>----->
      ----------------|                                 |---------------
       >---->----->----->PoP-terminated traffic         |
      ----------------|                                 |---------------
      ----------------|       PoP-generated traffic----->----->----->
      ----------------|                                 |---------------
                      |   <----Intra-PoP traffic--->    |
                      +---------------------------------+
```

Figure 1: Types of traffic in NFVI PoP network

The cross-through traffic is the traffic that reaches the PoP through an external link, it is processed by a service function chain (i.e., it is processed by a number of VNFs inside the PoP) and then is forwarded through an external link. Processing such type of traffic is one of the main purposes of the PoP since the PoP is part of the operator's infrastructure whose main purpose is to forward user's traffic.

The PoP-generated traffic is generated by VNFs located within the PoP. An example of such VNF would be a cache located inside the PoP which serves content to users. Similarly, PoP-terminated traffic is external traffic that is terminated by one VNF located inside the PoP for example a firewall.

Finally, Intra-PoP traffic is traffic generated and terminated inside the PoP that never leaves the PoP. This traffic includes much of the management traffic, deploying and moving virtual machines and VNFs across different servers and other signaling traffic (e.g., the one associated with voice calls).

In order to properly design the PoP network topology, it is relevant to know the distribution of the expected traffic in these categories.

2.3.2. Traffic pattern within the PoP

The traffic within the PoP will be composed of essentially two types of traffic:

1. The traffic served by the PoP. This is the traffic coming from and/or going to external links and that should traverse a number of servers where the different VNFs are placed. This includes the cross-through traffic, the PoP-generated traffic and the PoP-terminated traffic.

2. The operation and management traffic that includes all the traffic resulting from the management of the virtual machines and the VNFs, as well as signaling traffic required to provides the VNFs. This is the Intra-PoP traffic.

The traffic pattern of the traffic served by the PoP is basically determined by the location of the input link, the location of the output link and the mapping of the service function chain to servers.

2.3.2.1. Mapping of service function chains to servers

There are multiple possible strategies to deploy VNFs and SFCs in servers.

- Parallel SFC deployment strategy: One possible approach is to deploy all the VNFs of a given service function chain in a single server and deploy as many of these servers in parallel in order to server the different flows. When more flows arrive to the PoP, more servers are used in parallel.

- Sequential SFC deployment strategy: Another possible approach would be to deploy each VNF in a different server and have one (or more) servers dedicated to process this particular VNF for all the flows of the PoP. When the number of flows increases, the number of servers providing each VNF is also increased.

- Hybrid strategy: it is also possible to use a hybrid strategy, where several VNFs of the SFC are deployed together in a server and other VNFs of the SFC are deployed in separated servers.

There are many factors that influence this decision, including the performance of the implementation of the VNF (maybe the VNF is too demanding to be executed with other VNFs in the same server) or licensing conditions (maybe some VNF licenses are based on the number of servers deployed, while maybe others depend on the number of users served, or even the time the VNF is being executed).

In any case, to design the PoP topology it would be relevant to know:

The number of servers that the traffic served by the PoP will traverse (which is determined by the length of the SFCs and the deployment strategy of SFCs in servers).

The number of different SFCs that will be simultaneously available in the PoP at any point in time. At any point in time, different flows coming from a particular external link can be served by one or more different SFCs. These SFCs can be mapped to different sequences of servers. Depending on this, different flows coming

from any external links will have to traverse different sequences of servers, affecting the Intra-PoP traffic pattern.

2.3.2.2. Locality

There are two locality aspects that are affect the pattern of the traffic served by the PoP. First, whether the servers providing the different VNFs of each SFC can be assumed to be topologically close (e.g., in the same rack). If the SFCs that process the majority of the flows can be assumed to be topologically close, topologies that exploit locality can be useful.

The other locality related aspect that affects the topology design is the distribution of output links of the traffic arriving through the different input links. Consider the case of a Local PoP, which has links connecting to users (DSL, FTTH, etc) and links connecting to the rest of the provider's network. Let's call the first type of links user's links and the second type of links, core links. It is reasonable to assume that most of the traffic coming from a user's link will go to a core link and vice-versa. We can expect that the traffic between two user's links will be low and the same for the traffic between two core links. If we now consider the case of a regional PoP, it is not so clear we can make such assumption about the traffic between links. In case this assumption can be made, it would be possible to deign the topology to pair user's link with core link to optimize the transit between them.

2.3.2.3. Churn

There is also the question about how often the provided SFCs will change and frequently VNFs and virtual machines will be deployed in servers. This affects the amount of churn traffic in the PoP. There may be more to it...?

2.3.2.4. Growth

Another relevant aspect is the expected growth in terms of offered load to the PoP and also in terms of VNFs in the PoP. We should understand if the capacity of the PoP is expected to increase linearly or exponentially in time. Similarly, we need to understand if the number of VNFs and the length of the SFCs will remain more or less constant or will evolve. If does evolve, which is the expected pace. The reason for this is that different topologies support growth in different manners so depending on the expectation in this aspects, different topologies may be more or less suitable.

2.4. Technological considerations

2.4.1. Direct and Indirect networks

A network is called an Indirect network there are two types of nodes, nodes that source/sink traffic and nodes that forward traffic. A network is called a Direct network if every node plays both roles. Usually data center networks are Indirect networks, with switches that forward packets and servers that source/sink packets. While there have are proposals that use both switches are servers to forward packets (e.g., [BCube]), the main concern expressed against them is that the resources available in the servers should be used to execute applications (which is the final goal of the data center) rather than be used in forwarding packets.

In the case of an NFVI PoP network, the actual purpose of the servers is in many cases to forward packets through the VNFs provided by the server, so it may make perfect sense to use servers to forward packets. From this perspective, either direct networks or networks that use both switches and servers to forward packets may be attractive for NFVI PoPs.

2.4.2. SFC technology

Service Function Chaining can be accomplished using the IETF SFC protocol [I-D.ietf-sfc-architecture] or using a SDN approach, where a controller instructs the switches where to forward the different flows using Openflow. The two approaches have a different architecture with different components and it is possible that different topologies accommodate more naturally the elements of the different SFC architectures.

If using an OpenFlow approach, determine what form the rules take, consider when forwarding rules must be dynamically updated due to the arrival of new flows, and determine what peak update rate is required. Evaluate the SDN switches against all of these requirements.

2.4.3. Network Virtualization Technology

Technologies exist to improve performance of NFV functions, including PCI-passthrough, SRIOV and NUMA-aware process pinning. Other technologies are likely to become available in the future to offload network function.

Consider selecting infrastructure with these features if the NFV functions can utilize them and if the orchestration and control-plane infrastructure can configure them optimally.

Performance of individual NFV implementations may vary by an order of magnitude with or without the hardware features, so PoP planning and sizing must include consideration of how the functions fit with the hardware and whether the infrastructure can deploy virtual machines in a manner that allows the hardware to be used.

2.4.4. Software or Hardware Switching

Some network architectures require software switches (such as OpenVSwitch), whereas other architectures only require top-of-rack and backplane Ethernet switching.

Although software switches perform very well, they consume processing cores. PoP design must consider how many processor cores are required for software switches.

3. Design goals

In this section we describe the goals for the design of a NFVI PoP network topology. In broad terms, they include scalability, performance, costs, fault tolerance, operation, management and backward compatibility

3.1. Effective load

A first performance parameter that we should take into account when considering different topologies is the effective load supported by the network. The main goal of the NFVI PoP is to forward traffic between the different external links connected to the PoP. The performance of the PoP is measured by the traffic it is able to forward i.e., the more effective load it manages. In order to assess the effective load supported by the different topologies, we increase the offered load coming to the PoP through the different links and we measure the effective load that the PoP is able to deliver.

The effective load supported by a topology is likely to be affected by multiple factors, including the the different aspects we described in the traffic patterns section Section 2.3 (such as the traffic matrix between the different external links, the characteristics of the SFCs and so on), the routing inside the PoP, the different locality considerations, and the intra-PoP traffic. Moreover, in order for the comparison of two topologies to make sense, they need to be "equal" in some other dimension (e.g., cost, number of servers, number of links, number of switches or else).

For example, as a starting point, we can assume a purely random traffic matrix, i.e., every packet arriving through an external link is forwarded through n random servers in the topology and exits

through a randomly picked external link, and assume shortest-path, equal cost multi-path routing. We can compare different topologies with the same number N of servers. We perform the comparison by measuring the effective load when increasing the offered load and for different values of N and n. Of course, these conditions may greatly differ from the real operation condition, this is why it is useful to have information about the items described in section Section 2.

When performing this evaluation, it is useful to also measure the packet loss and to track the occurrences of hot-spots in the topology, in order to identify the bottlenecks of the topology which may be useful to improve it.

Related to this, it may be useful to consider the bisection bandwidth of the different topologies.

3.2. Latency

Another relevant performance indicator is the latency suffered by packets while traversing the PoP network. That is, for a topology of N servers, which is the latency for a packet that arrives through an external link, traverses n servers and exits through an external link. Since we only care about the latency cause by the topology itself (in order to assess the topology) we can measure the "latency" as the number of hops that the packet should traverse.

It is useful to measure the mean latency, but also the maximum latency, since an upper bound for the time a packet stays in the PoP is also relevant. Again, the latency/Hop count depends on the traffic matrix (i.e., the relation of the input and output links), the routing and the different locality aspects, hence it is useful to have information about these aspects. In any case, a purely random case as the one described for the effective load measurement could be used as a starting point.

Queuing between software elements can introduce latency, so it is important to include extra hops caused by software components (such as software switches) that may be required to deliver packets to virtual machines from physical interfaces, in contrast to technologies (e.g., SRIOV) that allow virtual machines to receive traffic directly from network interface cards.

3.3. Scalability

Scalability refers to how well the proposed topology supports the growth in terms of number of servers, line speed of the servers and capacity of the external links. there are some topologies that in order to support an increased number of servers require growing some

components beyond what is technically feasible (or what is economically efficient). For instance it is well known that tree topologies require the core switches to grow in order to support more servers, which is not feasible beyond certain point (or it becomes very expensive). That being said, we should consider scalability in the range of servers that we expect that a PoP will have to support in a reasonable time frame.

Another dimension to consider is the size of forwarding tables required by switches in the network. E.g., do the switches have capacity to learn the required number of MAC addresses? Some service chaining technologies utilize many private Ethernet addresses; is there capacity to learn the number that are required? The same reasoning should be applied to whichever types of forwarding tables are required, whether IP routing, MPLS, NSH, etc.

Another aspect somehow related to scalability is how well the different topologies support incremental growth. It is unclear at this point which will be the growth pace for the NFVI PoPs. In other words, given that we have a PoP with N servers operational, then next time we need to increase the number of servers, will it increase to N+1, to 2*N or to N*N? Different topologies have different grow models. Some support growing lineally indefinitely, others can be over-dimensioned in order to support some linear growth, but after a given number of additional servers, they need to grow exponentially.

3.4. Fault Tolerance

Fault tolerance is of course paramount for an NFVI PoP network. So, when considering topologies, we must consider fault tolerance aspects. We basically care about how well the topology handles link failures, switch failures and server failures.

We can assess the fault tolerance of topology by measuring the following parameters of the topology [DC-networks]:

- Node-disjoint paths: The minimum of the number of paths that share no common intermediate nodes between any arbitrary servers.

- Edge disjoint paths: The minimum of the total of number of paths that share no common edges between any arbitrary servers.

- f-fault tolerance: A network is f-fault tolerant if for any f failed components, the network is still connected.

- Redundancy level: A network has redundancy level of r if and only if after removing any set of r components, it remains connected

and exists a set of r+1 components such that after removing them, the network is no longer connected.

3.5. Cost

The cost of the resulting network is also a relevant aspect to be consider. In order to assess the cost, we can consider the number of switches and the number of interfaces in topology for the same number of servers. We should also take into account that type of switches required, as we know that the cost of a switch does not scale linearly with the number of interfaces of the switch and with the speed of the interfaces.

3.6. Backward compatibility

Another relevant aspect to consider is compatibility with existent hardware. It is unlikely that operators will throw away all their current infrastructure based on specialized hardware and replace it for VNFs running in COTS servers. It is more likely that there will be an incremental deployment where some functions will be virtualized and some function will be executed in hardware. It is then important to consider how the different topologies support such hybrid scenarios.

4. Topologies

In this section, we plan to describe different topologies that have been proposed for data centers and include some considerations about the different design goals described in section Section 3.

5. Security considerations

TBD, not sure if there is any.

6. IANA Considerations

There are no IANA considerations in this memo.

7. Acknowledgments

We would like to thank Bob Briscoe, Pedro Aranda, Diego Lopez, Al Morton, Joel Halpern and Costin Raiciu for their input. Marcelo Bagnulo is partially funded by the EU Trilogy 2 project.

8. Informative References

[I-D.ietf-sfc-architecture]
Halpern, J. and C. Pignataro, "Service Function Chaining (SFC) Architecture", draft-ietf-sfc-architecture-11 (work in progress), July 2015.

[ETSI_GS_NFV-INF_001]
., ETSI., "Network Functions Virtualisation (NFV); Infrastructure Overview", NFV ISG, 2015.

[swBRAS] Bifulco, R., Dietz, T., Huici, F., Ahmed, M., and J. Martins, "Rethinking Access Networks with High Performance Virtual Software BRASes", EWSDN 2013, 2013.

[BCube] Guo, C., Lu, G., Li, D., Wu, H., and X. Zhang, "BCube: A High Performance, Server-centric Network Architecture for Modular Data Centers", SIGCOMM 2009, 2009.

[DC-networks]
Liu, Y., Muppala, J., Veeraraghavan, M., Lin, D., and M. Hamdi, "Data Center Networks - Topologies, Architectures and Fault-Tolerance Characteristics", Springer Briefs in Computer Science Springer 2013, 2013.

Authors' Addresses

Marcelo Bagnulo
Universidad Carlos III de Madrid
Av. Universidad 30
Leganes, Madrid 28911
SPAIN

Phone: 34 91 6249500
Email: marcelo@it.uc3m.es
URI: http://www.it.uc3m.es

David Dolson
Sandvine
408 Albert Street
Waterloo, ON N2L 3V3
Canada

Phone: +1 519 880 2400
Email: ddolson@sandvine.com