IP Fabric Design

December 2016

Americas Principal Systems Engineering

Overview

Data center networking architectures have evolved with the changing requirements of the modern data center and cloud environments. The traffic patterns in data center networks are shifting rapidly from north-south to east-west. Cloud applications are often multitiered and hosted at different endpoints connected to the network. The communication between these application tiers is a major contributor to the overall traffic in a data center.

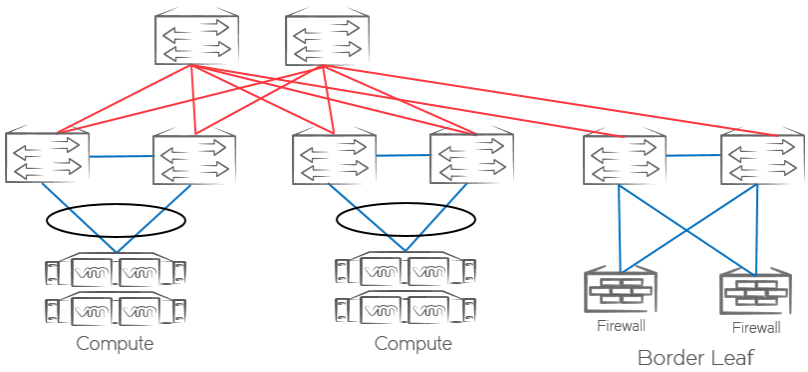
The goal is the creation of a cost effective, scalable and open data center. A key component to the next generation data center is the underlay network. This is the foundation upon which everything else within the data center will be built. A solid underlay network will have the flexibility to create overlay networks to support virtual environments required by the business. There are many facets to the underlay network that will contribute significant value over the lifetime of the infrastructure. The features of the underlay network and the corresponding business values will highlight the benefits and the expected return on investment for this new infrastructure.

The data center design centers on a vendor agnostic IP Fabric technology. A move from the old, legacy, hierarchical network infrastructure will dramatically improve the performance of the underlay network by eliminating the inefficiencies of a tiered network in terms of traffic flow and the limitations of scale and management. The solution design provides for a scale out network infrastructure that offers a pay as you grow philosophy. The underlay network provides the flexibility to accommodate any type of traffic flows and can adapt to meet the needs of applications in terms of east/west or north/south preference.

A main tenant of the solution is the open architecture of the underlay network. This concept encourages best of breed technologies and therefore provides excellent investment protection. Having a flexible infrastructure that can be changed dynamically via network programmability and orchestration ensures that the underlying network hardware need not be replaced to support new and innovative technologies and solutions.

The traffic patterns are the primary reasons that data center networks need to evolve into scale-out architectures built to maximize the throughput for east-west traffic. In addition to providing high east-west throughput, scale-out architectures provide a mechanism to add capacity to the network horizontally, without reducing the provisioned capacity between the existing endpoints.

The de-facto industry standard for implementing scale-out architectures is using Clos topologies. These topologies include the 3-stage folded Clos as shown below and the optimized 5-stage folded Clos for greater scalability. Because the infrastructure is built on IP, advantages such as loop-free communication using industry-standard routing protocols, equal cost multi-path, very high solution scale, and standards-based interoperability are leveraged.



Underlay Networking Requirements

The first question to answer when building an IP Fabric is what control plane protocol to use? There are several options to choose from including, OSPF, IS-IS and BGP. All are standards-based and interoperable protocols that are well understood and provide a solid foundation for a highly scalable solution. In addition, industry-standard overlay control and data plane protocols such as BGP EVPN and Virtual Extensible Local Area Network (VXLAN) are used to extend the Layer 2 domain and extend Layer 3 VRF tenancy domains by enabling communications via the overlay which also supports virtual machine (VM) mobility.

A key factor in moving from the old, monolithic architectures of the past 20 years is the ability to adapt to the ever-changing application requirements supporting the business. Optimizing the network is now a function of Dev/Ops automation that can be done quickly, efficiently and without the worry of human error.

With the IP Fabric architecture, there are critical services required to achieve an agile infrastructure:

- **Traffic Engineering** provides the ability to steer traffic across the IP Fabric. This can be extremely useful in times of troubleshooting or failure to move traffic around a potentially problematic switch. In cases of normal operations, traffic engineering can also be helpful to create specific network paths for mission critical applications (i.e. end of month payroll processing) along with helping to resolve issues with elephant and mice flows and the unequal distribution of traffic.

- **Prefix Filtering** offers control of routing in the case of what subnets see each other. This is particularly useful when leaking routes between VRFs in a multi-tenanted environment. It is critical to have control of what routes are leaked between VRFs. This is especially true when leaking between tenant VRFs and a common services VRF.

- **Prefix Distribution** is key for scaled networks. With the continued growth of server virtualization, the need for multi-tenancy and micro-segmentation, there are more prefixes per rack and in the overall network. The underlay protocol must scale to accommodate both today's requirements as well as what will be needed in the future.

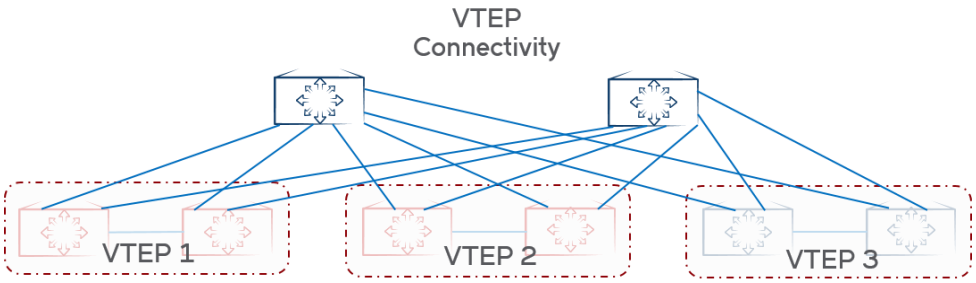
- **Traffic Tagging** can enhance the ability to identify specific traffic types or application flows within the network. In a highly scaled environment, the more visibility that can be achieved leads to more control and in many cases, help in troubleshooting specific issues that are attributed to the network (reducing mean time to innocence).

IP Fabrics afford a vendor agnostic architecture which is extremely appealing to customers in all verticals. In most cases, the IP Fabric is not built out 100% day one, but instead is grown at a pace that is needed by the business. This pace can be dictated by resource availability, budgets and application roll-out.

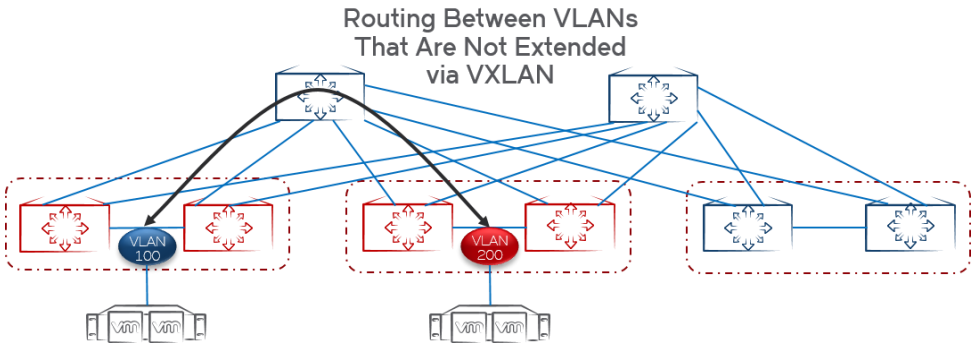
With this flexibility, customers may create an IP Fabric that is comprised of different vendor products, but all being fully interoperable. That being the case, it is imperative that the protocols being used are proven and stable ensuring that the IP Fabric can grow without requiring architectural changes down the road.

Underlay Networking Functions

The underlay network has two primary functions: provide connectivity between VTEPs (VXLAN Tunnel Endpoint) for VXLAN tunnels and to route traffic between VLANs that are not extended over VXLAN tunnels. Each of these examples are shown below:



The underlay network provides the routing capability for the VTEPs to see each other. Please note that BGP EVPN will be used for control plane learning of all the hosts that utilize the VTEPs and traverse the VXLAN tunnels, with VTEP discovery for the creation of VXLAN tunnels happening automatically via BGP EVPN. This functionality does not rely upon the underlay network..

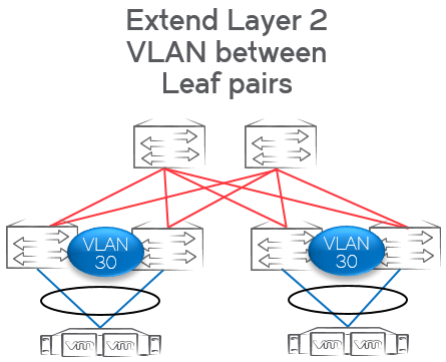


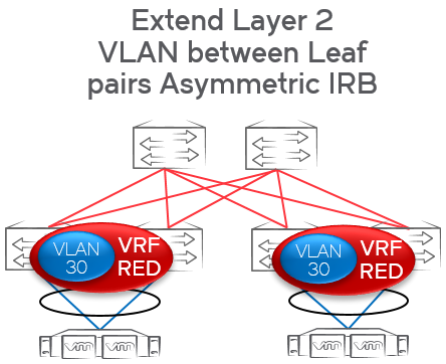
The example shown here assumes that VLAN 100 exists only in one rack and accordingly, VLAN 200 exists only in one rack. There is no need to extend either VLAN between racks and the VLANs are part of the default VRF. In this case, each VLAN will use a static anycast gateway at the Top of Rack and then the routing table in the default VRF to route to the other VLAN.

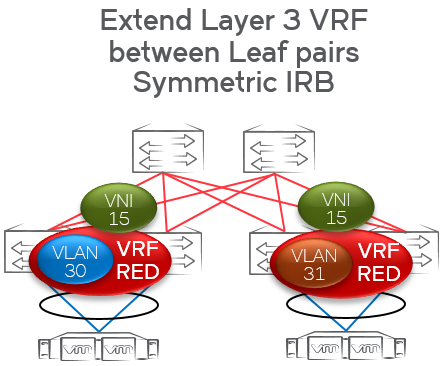
For use cases where VLANs are extended, or Integrated Routing and Bridging is required, BGP EVPN will be used as described in the next section.

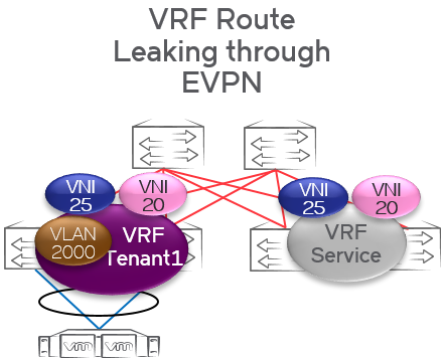
Overlay Networking Functions

The overlay network is responsible for providing connectivity in several different scenarios as described below. As shown, BGP EVPN will use the VTEPs (VXLAN Tunnel Endpoint) for the overlay network as well as routing of hosts that do not require Layer 2 extension or Layer 3 Integrated Routing and Bridging between VRFs. All traffic that uses VXLAN tunnels for L2 or L3 VNIs will use the BGP EVPN control plane learning.









Underlay Networking Selection

The underlay network selection comes down to one protocol versus two. eBGP will serve as a single protocol for the underlay and overlay, while for the two protocol solution, it is common to run a combination of iBGP + an IGP (OSPF or IS-IS). There are differences when comparing the two options. The following table and discussion will highlight these as it applies to configuration and on-going management/troubleshooting.

Requirement	OSPF	IS-IS	BGP
Traffic Engineering	Limited	Limited	Extensive
Prefix Filtering	Limited	Limited	Extensive
Prefix Distribution	Yes	Yes	Yes
Traffic Tagging	Limited	Limited	Extensive
Multi-Vendor	Extensive	Limited	Extensive

BGP is clearly ahead in all the major service categories when looking at the deployment of an IP Fabric. When looking at the above services, BGP provides:

- **Traffic Engineering** with proven BGP attributes such as MED (Multi-Exit Discriminator), AS (Autonomous System) Padding, Local Preference, AS Path Prepending and other features to provide granular control of certain traffic across the fabric.
- **Prefix Filtering** by matching on attributes or prefixes and allowing those matched to be pruned either inbound or outbound from the switch.
- **Traffic Tagging** with extended communities that allow the association of prefixes with multiple communities. This adds control and visibility to traffic flows.
- **Multi-Vendor** stability and support through extensive use for many years. The best use case reference for BGP is the Internet, which has massive scale, extreme stability and provides numerous tools for configuration, automation and management.

There are differences between eBGP and iBGP when it comes to configuration. Although both are BGP, they have different attributes and requirements for the IP Fabric. Some critical areas to understand is the ability to utilize ECMP (Equal Cost Multipath) for traffic distribution, the use of Autonomous Systems in the IP Fabric and the ability to do traffic engineering.

ECMP is critical to the distribution of traffic and providing a high-performance IP Fabric. In a traditional Clos architecture, each leaf is connected to each spine and therefore allows for a distribution of traffic from leaf to spine. With iBGP, ECMP is not inherent to the protocol and route reflectors at the spine will only propagate the best route to the leaf switches. To overcome this issue, BGP Add Path is a feature required to allow for ECMP across the fabric.

The use of Autonomous Systems differs in that iBGP will use one AS for the IP Fabric (all spines and leaf switches will be in the same AS). With eBGP, the spine switches will be in a single AS

while each rack (single or dual leaf) will have a unique AS. On the surface, this looks more complicated with eBGP, however, the configuration of each leaf will be identical and the ability to use peer groups greatly reduces configuration parameters. Also, the ability to use BGP dynamic neighbor discovery also greatly reduces configuration complexity.

One key area to consider is that iBGP does not support traffic engineering. If this is required today or potentially could be used in the future, eBGP will be required as the underlay networking protocol.

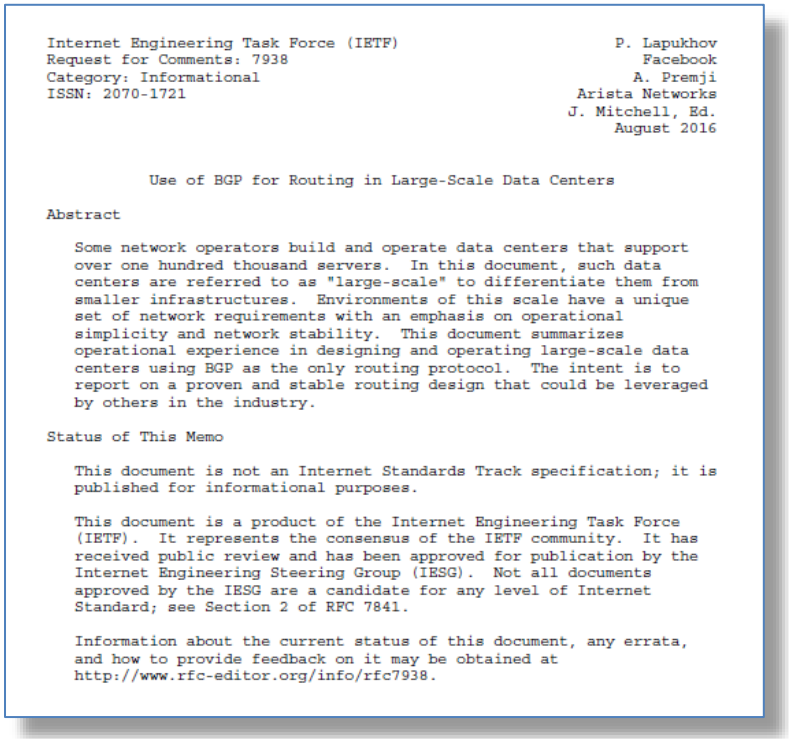
In summary, the use of eBGP in the IP Fabric clearly provides several advantages over using two protocols. More features are readily accessible with eBGP and the perceived complexities of eBGP have been mitigated by features within the protocol itself. To top it off, Dev/Ops tools for automation are available to minimize the amount of human intervention required for the configuration, validation and troubleshooting of the vendor agnostic IP Fabric.

Industry Adoption

The IP Fabric architecture continues to gain momentum throughout the industry. Beginning with the MSDCs (Massively Scalable Data Centers), service providers and cloud providers, the concepts behind the solution are attractive, scalable, vendor agnostic, and implementable with standards-based and interoperable protocols. Many traditional enterprise-class organizations have now started to gravitate toward this solution due to the above-mentioned attributes and the fact that the solution scales from the very large to the mid-sized and smaller environments.

Looking at the recommendations from the early adopters and understanding where the industry is moving ensures a solution that is mainstream and easily implemented. It also provides a community of interest that can be tapped to see what others are doing and how they have solved issues that may surface in your environment.

RFC 7938: Details the use of BGP for Routing in Large-Scale Data Centers. The RFC goes into detail on the use of routing protocols and the Clos architecture.



Internet Engineering Task Force (IETF)
Request for Comments: 7938
Category: Informational
ISSN: 2070-1721

P. Lapukhov
Facebook
A. Premji
Arista Networks
J. Mitchell, Ed.
August 2016

Use of BGP for Routing in Large-Scale Data Centers

Abstract

Some network operators build and operate data centers that support over one hundred thousand servers. In this document, such data centers are referred to as "large-scale" to differentiate them from smaller infrastructures. Environments of this scale have a unique set of network requirements with an emphasis on operational simplicity and network stability. This document summarizes operational experience in designing and operating large-scale data centers using BGP as the only routing protocol. The intent is to report on a proven and stable routing design that could be leveraged by others in the industry.

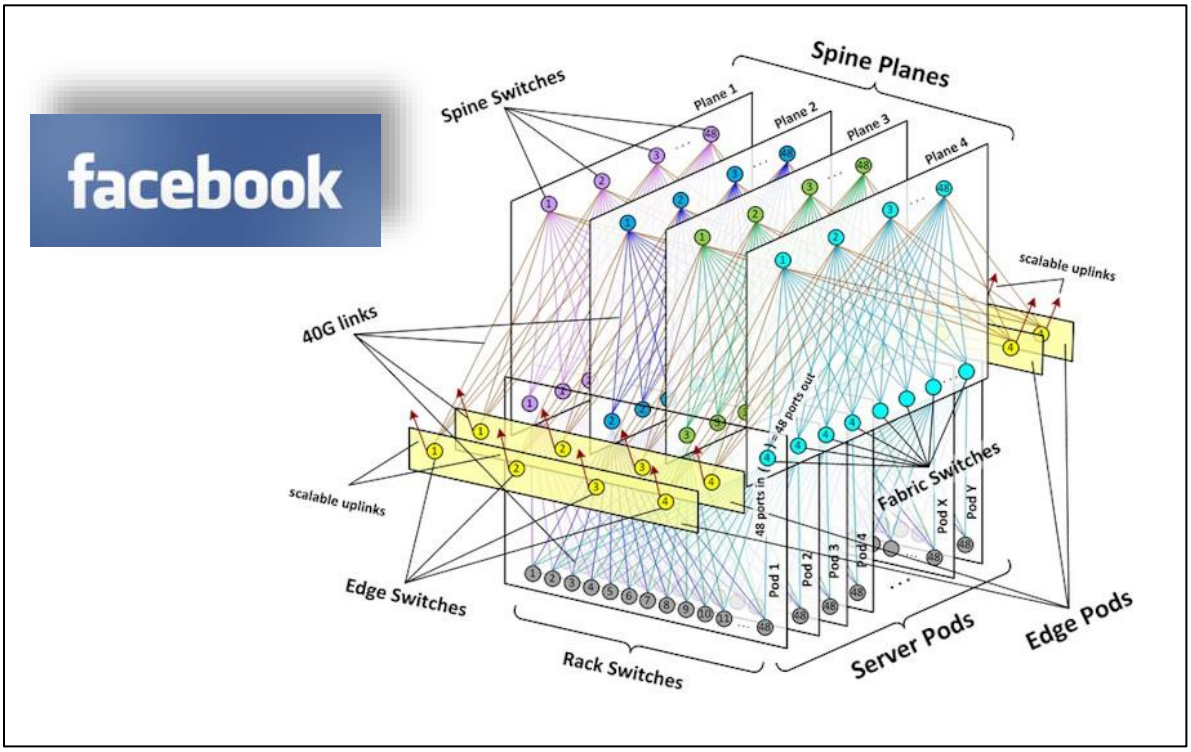
Status of This Memo

This document is not an Internet Standards Track specification; it is published for informational purposes.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Not all documents approved by the IESG are a candidate for any level of Internet Standard; see Section 2 of RFC 7841.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at http://www.rfc-editor.org/info/rfc7938.

Large network implementations using IP Fabrics:



http://www.enterprisetech.com/2014/11/18/facebook-gives-lessons-network-datacenter-design/

"The entire network fabric is implemented in 40 Gb/sec Ethernet (excepting the links between servers and top of rack switches), and it runs the Border Gateway Protocol (BGP) to do Layer 3 switching across the whole shebang, from the top of racks out to the edge switches, and Facebook can support both IPv4 and IPv6 protocols. The switches are based on merchant silicon (of course)."

"Fabric offers a multitude of equal paths between any points on the network, making individual circuits and devices unimportant – such a network is able to survive multiple simultaneous component failures with no impact," says Andreyev. "Smaller and simpler devices mean easier troubleshooting. The automation that Fabric required us to create and improve made it faster to deploy than our previous data center networks, despite the increased number of boxes and links. Our modular design and component sizing allow us to use the same mid-size switch hardware platforms for all roles in the network – fabric switches, spine switches, and edges – making them simple 'Lego-style' building blocks that we can procure from multiple sources."

Routing Design for Large Scale Data Centers: BGP is a better IGP!

Presented by: Parantap Lahiri, George Chen, Petr Lapukhov, Edet Nkposong, Dave Maltz, Robert Toomey, and Lihua Yuan

Global Networking Services Team, Global Foundation Services, Microsoft Corporation

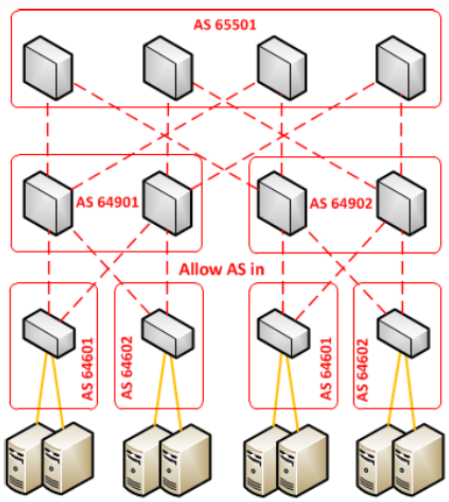
Microsoft

Build a topology providing significant amount of bisection bandwidth

The simpler the better

Design a scalable routing model for this topology

Single protocol
Simple behavior
Wide vendor support



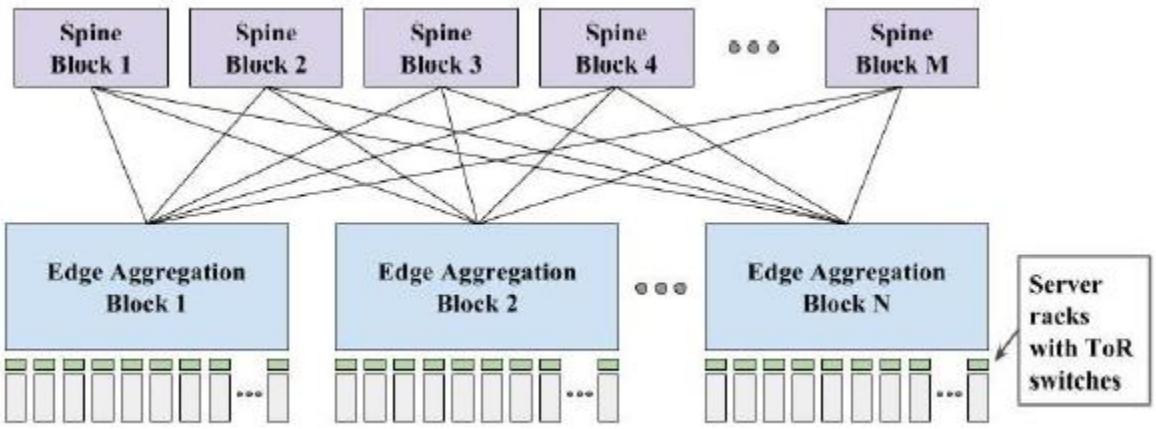
BGP has been thought as slow and suitable for inter-domain routing only…

With modern implementations, it might as well work as IGP!

BGP is simple, allows for per-hop traffic engineering and supported by practically all vendors

This made it perfect choice for us!

Google



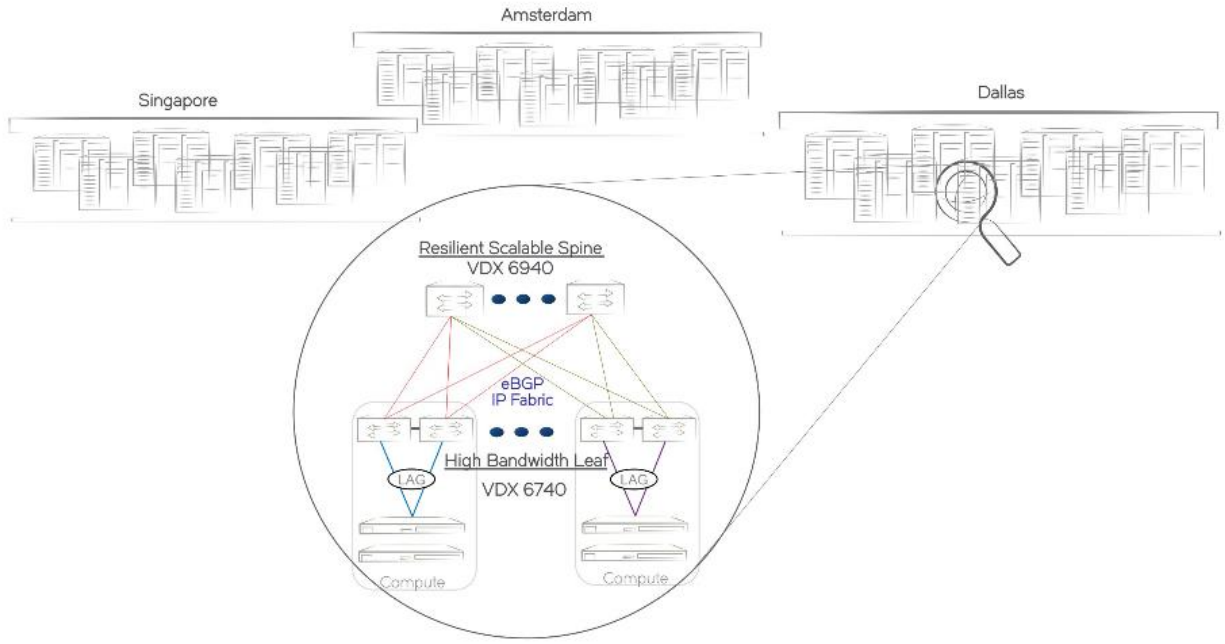
http://highscalability.com/blog/2015/8/10/how-google-invented-an-amazing-datacenter-network-only-they.html

The shift to East/West dominate communication patterns meant a different topology was needed for datacenter networks. The old traditional fat tree network designs were out and something new needed to take its place.

Google has been on the forefront of developing new rich service supportive network designs largely because of their guiding vision of seeing The Datacenter as a Computer. Once your datacenter is the computer then your network is equivalent to a backplane on a single computer, so it must be as fast and reliable as possible so remote disk and remote storage can be accessed as if they were local.

Google's efforts revolve around a three pronged plan of attack: use a Clos topology, use SDN (Software Defined Networking), and build their own custom gear in their own Googlish way.

servers.com



http://emiratespressreleases.com/servers-com-reveals-advanced-software-defined-network-for-cloud-and-hosting/

AMSTERDAM, THE NETHERLANDS-(Marketwired – Feb 9, 2016) – Servers.com has completed the roll out of a new software-defined network using Brocade IP Fabric. The network connects four data centres across two continents to ensure Servers.com can flexibly scale to meet the growing demand for its cloud, content delivery, and dedicated hosting services at competitive price-points. The highly automated, open standards-based network has enabled the company to join the small group of providers globally who offer 40 GbE server connectivity for data-intensive applications such as Big Data analytics.