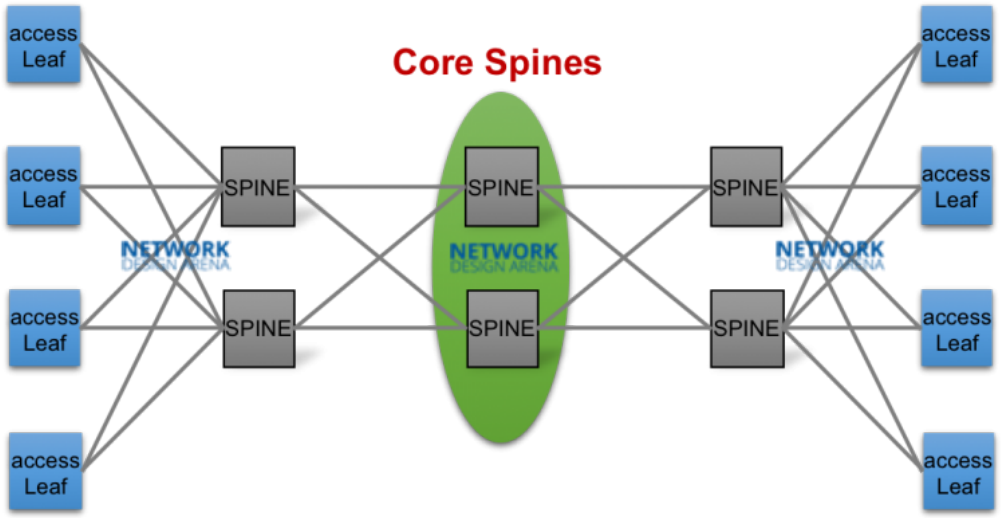
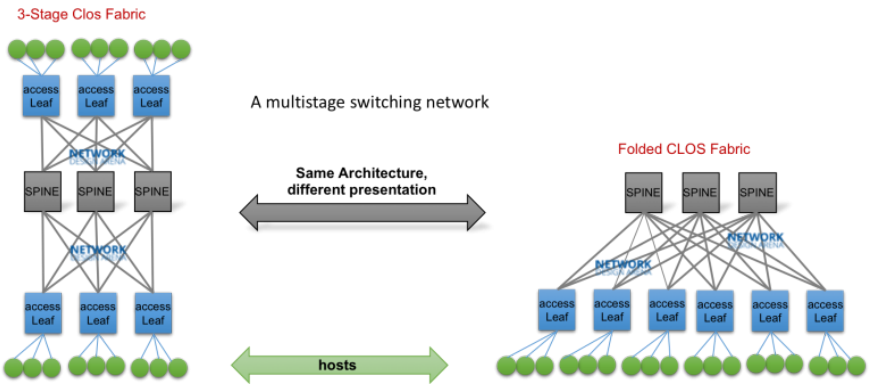
Clos (Spine & Leaf) Architecture – Overview



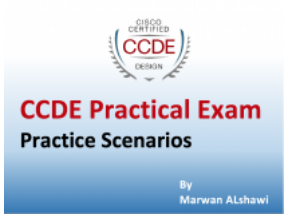
Part-1 of this blog, analyzed and discussed the different aspects of the classical 3-tier network architecture in a data center environment, along with the drivers and needs to adopt a new architecture. this blog focuses on Clos architecture also commonly known as “Spine and Leaf” architecture and why today, it’s becoming the most wildly used data center architecture.

First lets define what Clos architecture and why its called “Clos”.

Clos architecture offers a nonblocking architecture based on multistage topology, which can maximize the overall network performance by increasing the available bandwidth to higher than what a single node can provide. The mathematical theory of this architecture was initially created by Charles Clos in 1953, hence the reason it is called Clos. In a nutshell, this architecture is based on associating each output with an input point. In other words, the number of outputs equals the number of inputs, or there is precisely one connection between nodes of one stage with those of another stage. Below Figure illustrates this architecture that offers zero oversubscription (1:1) where a 3-stage Clos topology is used to connect 18 host (edge ports) using 3x of 6 port switches in each stage



Order Now



Featured Posts

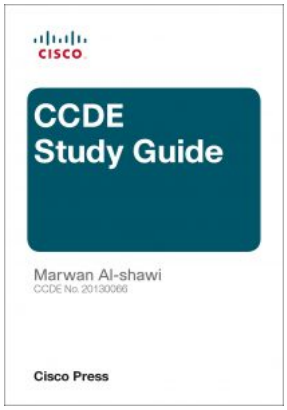
Certification Selection Dilemma: CCDE, De Cloud Certs

Cisco SDWAN Design Series-Part-4- The M Cisco SDWAN Policies

Cisco SDWAN Design Series-Part-3- Contro Planes Logic

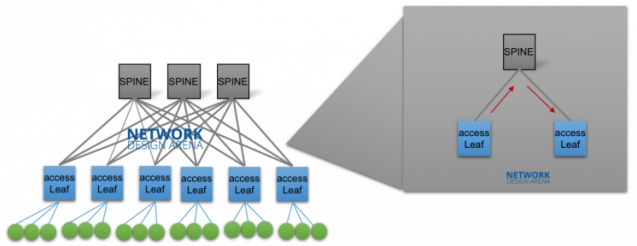
Cisco SDWAN Design Series-Part-2- Archite Components

Cisco SD-WAN – Network Evolution for the and Digital Era –



This architecture can be considered a two-tier hierarchy, in which the access layer switches are called leaves and the aggregation layers are called spines. With this two-tier leaves and spines architecture, each leaf switch is connected to every spine switch on the aggregation layer, which makes the DCN architecture flatten and eliminate bandwidth aggregation/oversubscription, because of bandwidth being the same at every tier. As a result, this architecture offers a high degree of network redundancy and a simplified design, and supports a large volume of bisectional traffic (east-west).

As illustrated below, there is one connection between each ingress stage switch and each middle stage switch. And each middle stage switch, is connected exactly once to each egress stage switch, as a result there is always one hope between any ingress and egress stage switches.



Although the Clos architecture can offer scalable 1:1 (zero) oversubscription, this level of oversubscription is rarely needed or used in such manner. Typically, 4:1 or 3:1 offers an efficient level of oversubscription with modern data centers that are built on 10/40/100G Ethernet. In fact, application and services requirements should drive the required level of oversubscription during the planning phase, taking into account the anticipated traffic load.

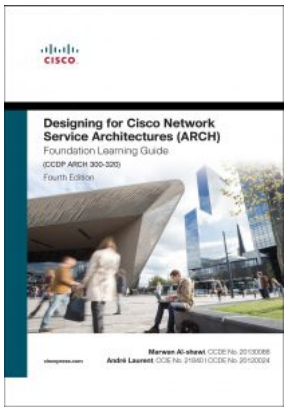
Even with Facebook Data Center Fabric scale, 4:1 fabric oversubscription from rack to rack, was used with the ability to quickly jump to 2:1 oversubscription, or even full 1:1 non-oversubscribed state at once. That being said, **this is a WebScale DC, and not an enterprise scale**.

https://code.fb.com/production-engineering/introducing-data-center-fabric-the-next-generation-facebook-data-center-network/

If you plan to support large scale data center, modular switches at the Spine layer, offers more efficiency and scalability as it will be matter of adding new line card when the number of nodes increased (scaling up at the Spine layer and scaling out at the leaf layer) . Moreover, by considering multipathing techniques such as using IGP or BGP multipathing , the design can efficiently utilizes network resources to support more leaf switches.

What if the size of the data center grows significantly with large number of Spines, how to optimize the design to reduce the significant increased Leaves' uplinks ?

This is where the five-stage Clos topology can be useful for a significantly larger scale DC networks, where a Core Spine layer can be introduced, as illustrated below.



Recent Posts

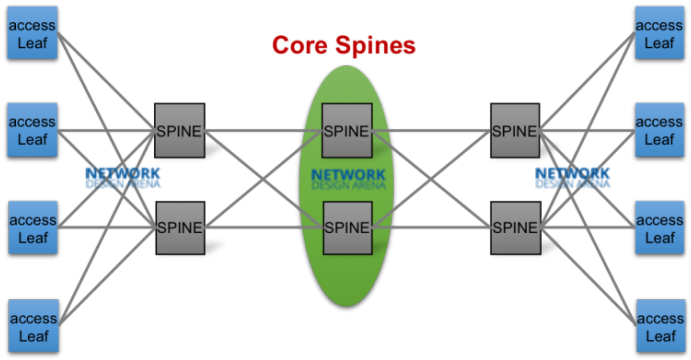
Certification Selection Dilemma: CCD DevNet or Cloud Certs

SDDC (ACI) and SDWAN -Better Toge

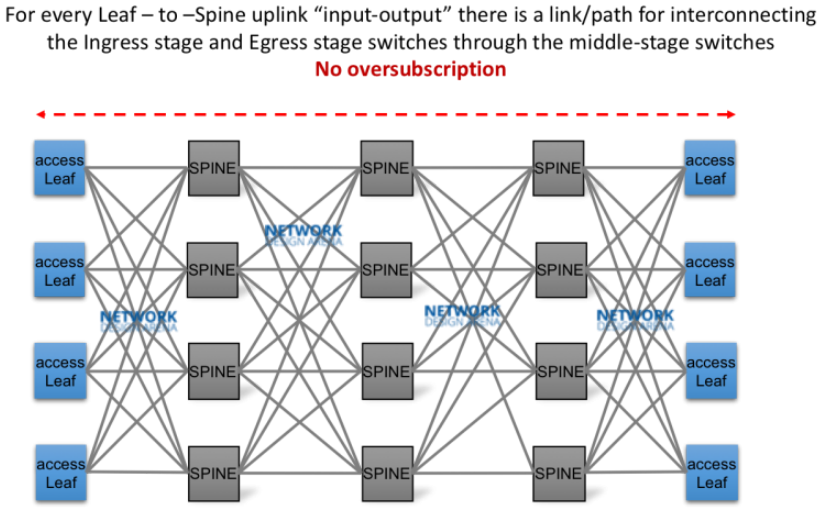
Cisco SDWAN Design Series-Part-4- T Magic of Cisco SDWAN Policies

The Road to HybridCloud

Cisco SDWAN Design Series-Part-3- C Data Planes Logic



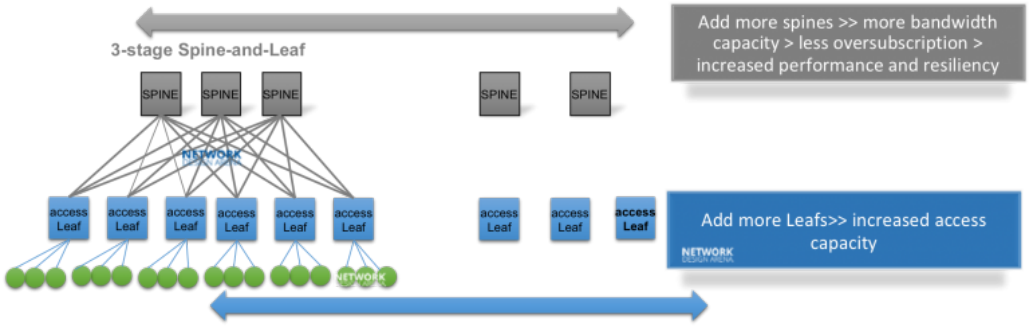
Or even, it can be scaled using none oversubscribed five-stage Clos as shown below.



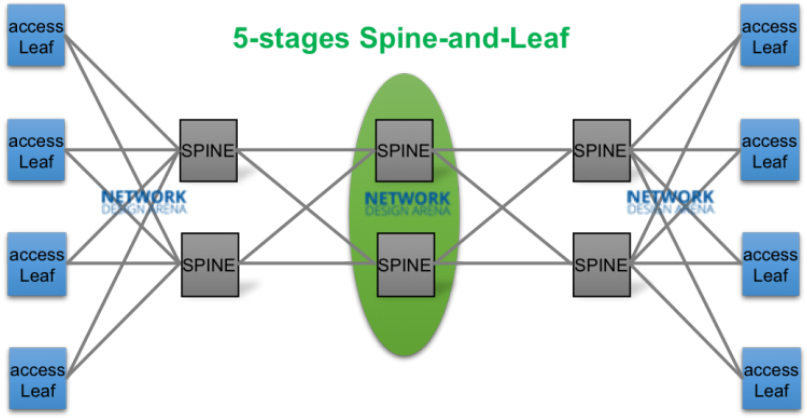
Let's be realistic, as discussed above, when do you really need 1:1 non-oversubscribed links with 10/40/100G ! even large scale content providers they use between 4:1 and 10:1 oversubscription. Also, practically this non- oversubscribed may introduce cabling limitations and high control plane complexity at scale.

Keep it simple and realistic

3-stage Spine-and-Leaf with reasonable oversubscription (based on applications demands) proven to support medium to large scale modern data center networks with low latency and high bandwidth demands.

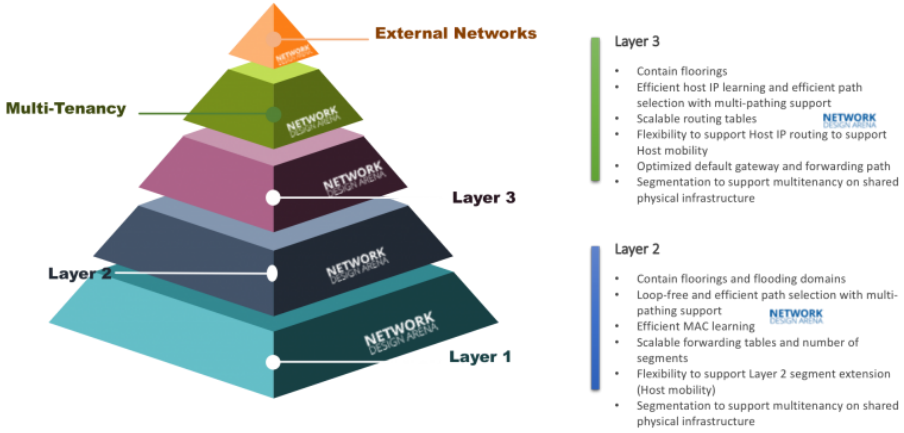


In rare situations in enterprises, If there a need for extremely large scale DC Network you can scale it using 5-stage Spine-and-Leaf architecture (with reasonable oversubscription) e.g. 8:1 leaf-to-Spine(bundled 40G interfaces), 4:1 Spine-to-Spine (40/100G interfaces)

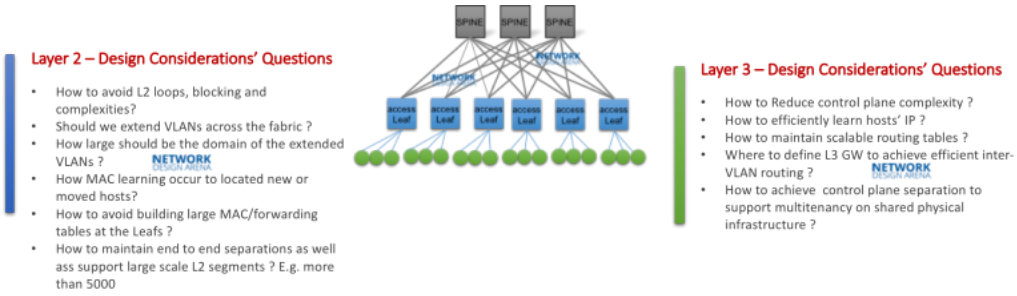


As designer you must not stop here. You need must also look at the higher layers than the physical architecture, by considering Layer 2 and Layer 3 traffic handling and communications, to meets applications requirements. Because you may have the best physical architecture and when you apply certain layer 2 or layer 3 control plane protocols you break it or reduce its efficiency and value e.g. blocking links, large flooding scope etc.

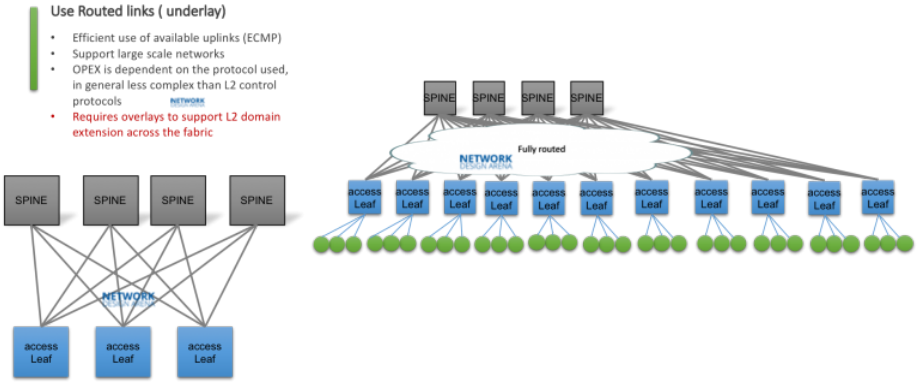
The following figure provide a summary of these considerations



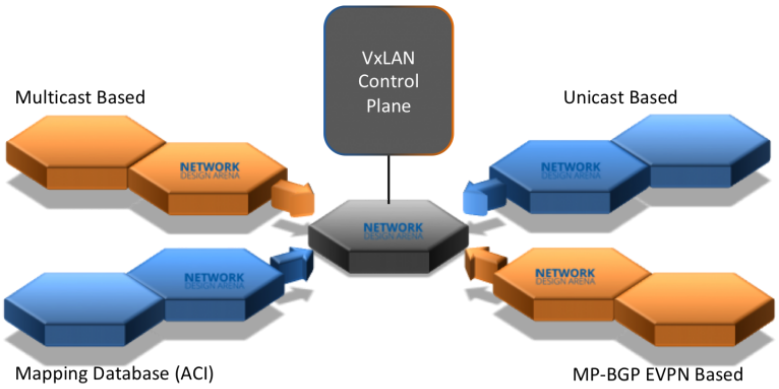
So should we use L3 or L2 fabric?



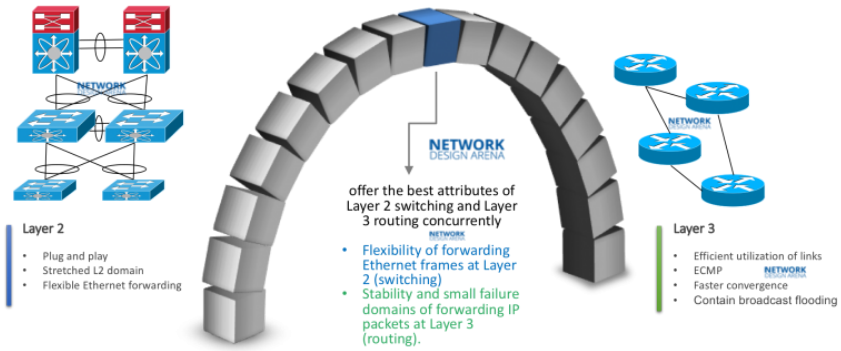
The proven reliable and flexible design model is by using routed (L3) Spine and Leaf architecture.



For the DCN overlay, to provide L2 segment extension for hosts' mobility VxLAN is the most commonly used tunneling mechanism. However, VxLAN control plane support multiple options as illustrated below



The MP-BGP EVPN is a proven, reliable and efficient control plane protocol, when used in a none ACI environment. In ACI environment the mapping DB uses proprietary approach that is similar to the MP-BGP EVPN operation with high degree of efficiency and scale.



The subsequent blog, will discuss the design approaches of the VxLAN tunneling along with the design consideration of the underlay routed Spine and Leaf (Clos) fabric.

Categories :

Data Center and Cloud

Marwan Al-shawi – CCDE No. 20130066, Google Cloud Certified Architect, AWS Certified Solutions Architect, Cisco Press author (author of the Top Cisco Certifications' Design Books "CCDE Study Guide and the upcoming CCDP Arch 4th Edition"). He is Experienced Technical Architect. Marwan has been in