Juniper Networks | Engineering Simplicity

White Paper

DESIGN CONSIDERATIONS FOR SPINE-AND-LEAF IP FABRICS

Properly Scaling and Sizing a Three-Stage Spine-and-Leaf Clos IP Fabric for the Data Center

TABLE OF CONTENTS

EXECUTIVE SUMMARY

The three-stage Clos fabric, invented in the 1950s for use in telephone switching networks, has been widely adopted in data center fabrics because of its simplicity, its ability to support the edge-to-edge (east-west) traffic flows required by modern applications, and its flexible scale-out characteristics. Properly sizing and scaling a three-stage spine-and-leaf IP fabric, however, can be a challenge for network designers.

This white paper covers background and definitions of Clos-based spine-and-leaf IP fabrics; contrasts scale-out and scale-up designs, including the use of chassis devices in spine-and-leaf fabrics; explains scaling steps and why there are size constraints on spine-and-leaf fabrics, and describes two basic scaling considerations: contention and oversubscription.

Scale steps are also revisited to provide examples of practical scaling steps and the aspects of wiring a scale-out fabric. The final two sections address topics of interest when designing spine-and-leaf IP fabrics: splitters and external fabric connections.

Introduction

When designing his original fabric concept, Charles Clos focused on one specific challenge: moving data from input ports to output ports. The fabric itself should, from the outside, appear to be a single, flat, forwarding domain; all traffic is carried east-west from input ports to output ports. While the fabric externally appears to provide a flat, undifferentiated forwarding service with consistent latency and delay (thanks to consistent hop counts through the fabric), devices have assigned roles depending on where they are attached to the fabric (see Figure 1).

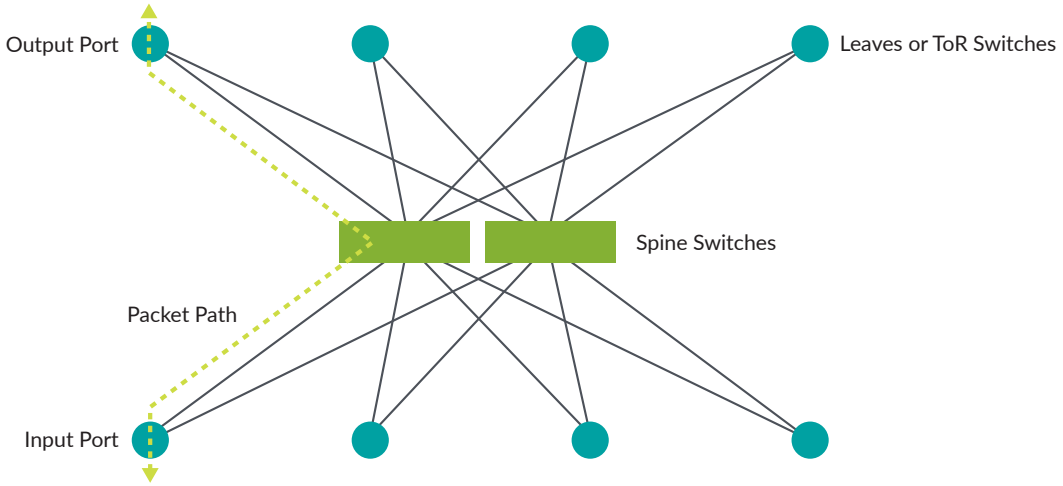

Figure 1: Traffic flow in a three-stage IP fabric

Devices along the edge of the fabric, which have input and output ports, are called leaves; in packet-switched networks, these devices are also called *top-of-rack (ToR)* switches (although they often route rather than switch packets). Devices used to interconnect the leaves in a packet-switched networks are called *spine switches* (although again, they often route rather than switch packets). Packets enter an input port at a leaf, are handed off to a spine switch, and then moved back to a leaf device to be transmitted through an output port. Hence, Clos fabrics have three stages.

Each of the two device types—leaves and spines—have a specific purpose. Leaves must accept traffic, apply any policies required (such as filtering out packets based on their source or destination addresses), and select which spine switch to forward each packet. Spine switches, on the other hand, are purely transit devices; in a pure Clos

design, spine switches do not apply policies, accept externally sourced traffic, or perform any other functions. Some modern "mixed hierarchical/ spine-and-leaf" designs do place intelligence or external ports in spine switches. The original intent, however, was for the intelligence to be at the edge to make the entire fabric appear to be a flat, undifferentiated forwarding domain.

Placing all intelligence at the fabric edge in the underlay complements the end-to-end principle often considered ideal in packet-based IP network and protocol design.

Scale-Out Design

In scale-out designs, system capacity increases by adding new devices rather than by increasing the capacity of existing devices. Figure 2 illustrates scale-up design in a five-stage spine-and-leaf (Benes) fabric. (Click here to learn more about the five-stage Benes fabric.)

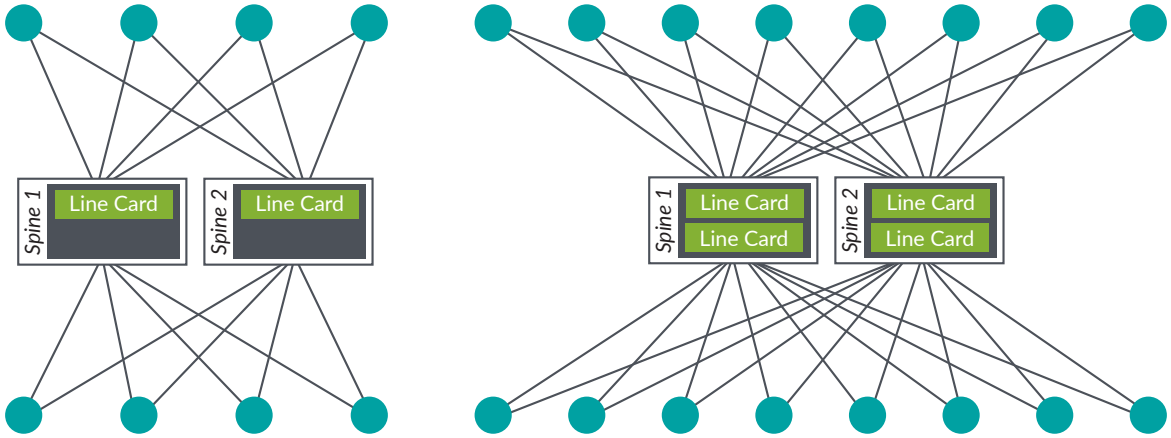

Figure 2: Increasing IP fabric size using a hybrid scale-out/scale-up design pattern

On the left side of Figure 2, each spine switch has one line card, and there are eight leaves total. To increase the number of leaves—that is, to scale the fabric—an additional line card is added to each spine. In this case, the size of the individual devices increases—adding a new line card to each spine—rather than increasing the number of devices in the network.

The primary advantage of the scale-up over the scale-out design pattern is the increased rack density possible when using chassis-style devices, plus the reduced amount of cabling required at larger scales. Figure 3 illustrates a scale-out design pattern using a Clos network.

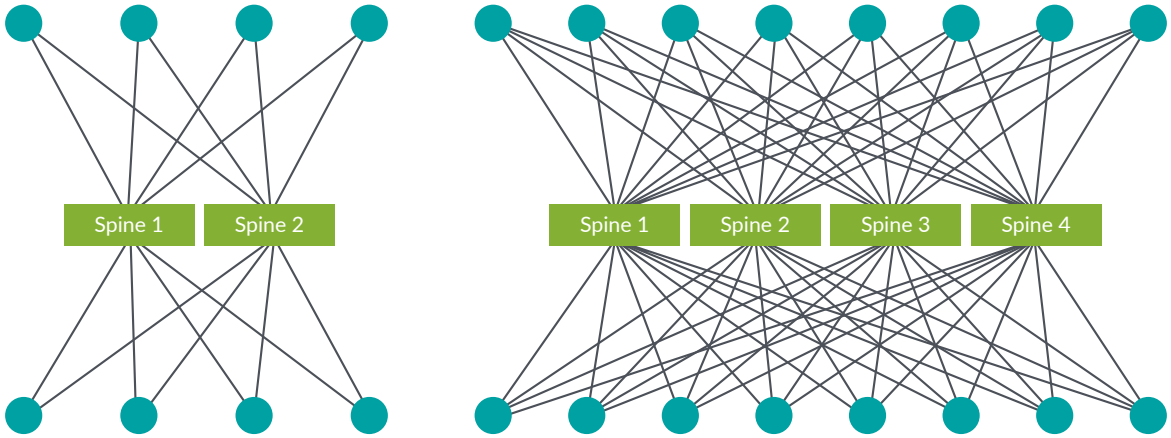

Figure 3: Increasing IP fabric size using fixed format devices in a scale-out design pattern

The left side of Figure 3 features two spines and eight leaves. On the right side, the number of leaves is increased to 16 by doubling the number of spine switches. As long as there are ports available on the leaf switches, a Clos fabric can scale out by adding more spine switches.

The scale-out design has four advantages over the scale-up approach. First, additional scale can be purchased as needed rather than during network design. Instead of investing in a large chassis to accommodate potential future growth by filling it with line cards, operators can purchase and deploy new switches in parallel when more capacity is needed.

Second, fabrics can be designed using a *single SKU*, or a single kind of device in each generation, using scale-out principles. If the spine-and-leaf devices are correctly chosen based on a maximum desired scale and a given oversubscription rate, *every device in the fabric can be identical* (except for border leaves and other "specialized" nodes). Single SKU designs simplify troubleshooting, spares, and planning for the fabric.

Third, exposing all the links to network management allows the operator to better monitor and understand the fabric operation. Chassis devices must have an internal fabric (network) to carry packets between line cards; the addition of this internal fabric converts the overall design from a Clos to a Benes, adding one or two more stages to the fabric (depending on the chassis design). While this internal network has quality of service (QoS), queuing, delay, jitter, and all the characteristics of any other network, these are typically not visible to the operator. Whether this is an issue of telemetry, troubleshooting, or operations depends on the fabric's purpose—the applications supported by the fabric.

Fourth, using fixed format devices built on commodity switching engines takes advantage of economies of scale in design and production.

Finally, increasing the scale in a scale-out design increases network resilience in parallel with increasing the number of workloads supported. In the scale-up design, the number of spines is two in both the original and the scaled-up design; if one of these two spines fails, the amount of bandwidth available to workloads is halved. In the scale-out design, the number of spines increases from two to four, so the failure of a single spine only reduces total bandwidth by one quarter.

Scaling Steps

Designers should use the following three rules of thumb to determine the maximum size of a Clos (or Benes) fabric:

- Every leaf should have the same number of fabric-facing ports (ports connecting the leaf to the spine)
- The number of fabric-facing ports on any leaf (not the sum of all fabric-facing ports on all leaves) limits the number of spines
- The aggregate bandwidth across all spine ports multiplied by the oversubscription rate determines the bandwidth available at the fabric edge

There are practical "scale steps" dictated by the number of ports on each device, as well.

Non-Blocking and Noncontending

When deployed in circuit-switched networks, Clos fabric are non-blocking, which means the fabric will carry any traffic accepted at an input port to the correct output port. If the fabric has no available bandwidth, attempting to set up a circuit across the fabric will fail.

Because packet-switched networks do not have the admission controls of circuit-switched networks, they are *noncontending rather than non-blocking*, as shown in Figure 4.

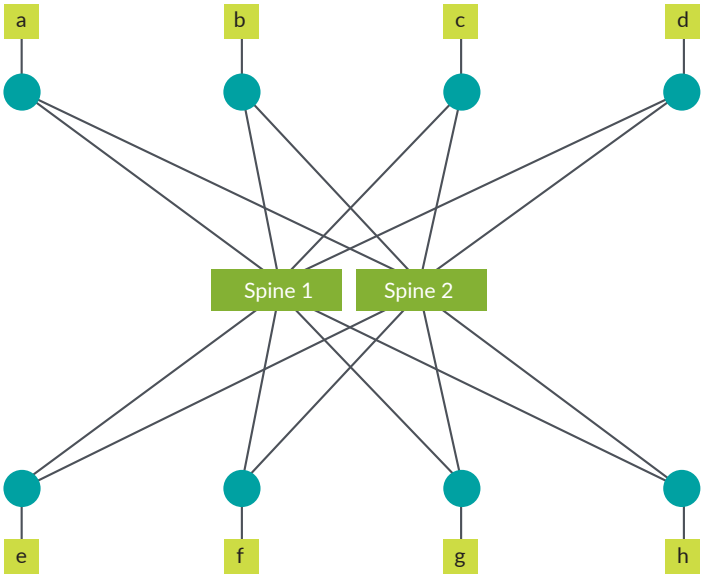

Figure 4: Non-blocking versus noncontending IP fabrics

Assume all links are 100 Gbps, so the fabric has 1:1 oversubscription on all ports. If a sends 100 Gbps of traffic to *e*, *b* sends 100 Gbps of traffic to *f*, *c* sends 100 Gbps of traffic to *g*, and d sends 100 Gbps of traffic to *h*, no packets will be dropped because traffic is equally distributed across the fabric. However, if both a and b send 100 Gbps of traffic to e, half of the traffic will be dropped somewhere in the fabric, most likely at the leaf node to which e is connected.

Oversubscription

The oversubscription rate is simply the amount of bandwidth at each leaf facing the fabric vs. the amount of bandwidth facing the workload. For instance, using a 32x100 Gbps leaf, some options are:

- 16x100 Gbps facing the fabric and 16x100 Gbps supporting 100 Gbps workloads, producing 1:1 oversubscription
- 12x100 Gbps facing the fabric and 20x100 Gbps facing workloads, each of which consumes 100 Gbps, producing 1:1.67 (20/12) oversubscription
- 8x100 Gbps facing the fabric and 24x100 Gbps facing workloads, each of which consumes 100 Gbps, producing 1:3 oversubscription
- 8x100 Gbps facing the fabric and 24x100 Gbps facing workloads, each of which consumes 40 Gbps, producing 1:1.2 oversubscription
- 8x100 Gbps facing the fabric and 24x100 Gbps facing workloads, each of which consumes 25 Gbps, producing 1:1 oversubscription

***Note**: It isn't possible to go lower than 1:1 oversubscription, as the workloads cannot consume the bandwidth available on the fabric.*

Oversubscribing allows the designer to build a larger fabric edge with a smaller number of spine switches, reducing CapEx. Oversubscription is usually considered safe when the application profiles are known to load the network at a much lower level than the server and ToR port speeds, or when traffic patterns are cyclical and nonoverlapping.

Oversubscription can cause problems when traffic loading reaches some significant number, or if traffic patterns push flows through one or two fabric hot spots (such as a database or backup server). In the extreme case, packets will overflow the smaller queues usually associated with data center switches, causing drops and ultimately impacting application performance. Even before reaching this extreme, however, increased queue depths can cause delay and jitter across the fabric, adversely impacting application performance. Problems related to drops, delay, and jitter are generally difficult to diagnose and resolve.

Not every leaf needs to be configured with the same oversubscription, as long as every leaf has the same number of spine-facing ports (connected to spines).

Scale Steps and Practical Scale-Out

Given the scale-out design guidelines and principles, it might seem possible to design a fabric that can scale up to virtually any number of ports. This is not the case. The number of ports at each spine device and the number of fabric-facing ports at each leaf limits the maximum size of the finished or completely scaled-out fabric. For instance, using only 32x100 Gbps devices and configuring each leaf so there are 8x100 Gbps fabric-facing ports and 24x100 Gbps workload facing ports means:

- The maximum number of spine devices possible is eight, because there are eight fabric-facing ports at each leaf.
- The largest possible number of leaves is 32, because each spine has 32x100 Gbps ports

Doubling the number of ports in each spine by moving to 64 port devices (using a 64x100 Gbps spine) will double the number of leaves and double the fabric scale without impacting the oversubscription rate at each leaf. Switching from eight to 16 fabric-facing ports on each leaf reduces the oversubscription rate and increases the maximum number of spines to 16, but does not change the total number of possible leaves.

Designing a Fabric Based on Workload Needs

Fabric design is a relatively well-defined process:

1. Begin by determining the maximum number of workloads the fabric will need to support when completely scaled out and add some "fudge factor" (because all designers should include a buffer).
2. Use this number to determine the correct number of leaves.
3. Determine the number of spines.

For instance, for a fabric requiring 800x100 Gbps workload ports at 1:1 oversubscription:

- 1:1 oversubscription requires the port count on a 100 Gbps device to be split, with half supporting workloads and half fabric facing.
- There are 32x100 Gbps and 64x100 Gbps configurations available, so each device must have either 16 or 32 workload ports respectively.
- Using 32x100 Gbps leaves, there must be 800/16 = 50 leaves, which means the spine devices must have at least 50x100 Gbps ports each—hence 64x100 Gbps is the only viable spine option.
- Using 64x100 Gbps spines, the fully scaled-out fabric can have 64 leaves, which means the final scale will be 64 leaves times 16 workload-facing ports per leaf, for a total of 1024x100 Gbps workload-facing ports.

While it might seem like any fabric size can be designed, practical device sizes mean there are some specific "scale steps" that make sense while others do not. Once the fabric has reached its maximum scale-out size, based on port counts, it cannot be scaled-out any further. Therefore, *the designer must know the maximum projected fabric size*.

According to scale-out principles, however, the fabric does not need to be *initially configured* to this maximum size. Instead, the fabric can be built at some smaller scale and devices added over the life of the fabric to increase the scale. While open ports may be left on spine devices, *all fabric-facing ports on each leaf must be connected to a spine device to maintain the oversubscription rate at every step of the scale-out process*. For instance, using the example above, the fabric can scale out along these lines:

- Two spines, eight leaves:
  - Each leaf has 16 fabric-facing ports and there are two spines, so each leaf must have 16/2 = 8 connections to each spine.
  - There are eight leaves, each with eight connections to each spine, so there are 8x8 = 64 total connections to each spine, matching the number of ports available on each spine.
  - The fabric has 8x16 = 128x100 Gbps ports at this scale.

- Four spines, 16 leaves:
  - Each leaf has 16 fabric-facing ports and there are four spines, so each leaf must have 16/4 = 4 connections to each spine.
  - There are 16 leaves, each with four connections to each spine, so there are 16x4 = 64 total connections to each spine, matching the number of ports available on each spine.
  - The fabric has 16x16 = 256x100 Gbps ports at this scale.
- Eight spines, 32 leaves:
  - Each leaf has 16 fabric-facing ports and there are eight spines, so each leaf must have 16/8 = 2 connections to each spine.
  - There are 32 leaves, each with two connections to each spine, so there are 32x2 = 64 total connections to each spine, matching the number of ports available on each spine.
  - The fabric has 32x16 = 512x100 Gbps ports at this scale.
- 16 spines, 64 leaves:
  - Each leaf has 16 fabric-facing ports and there are eight spines, so each leaf must have 16/16 = 1 connection to each spine.
  - There are 64 leaves, each with one connection to each spine, so there are 64x1 = 64 total connections to each spine, matching the number of ports available on each spine.
  - The fabric has 64x16 = 1024x100 Gbps ports at this scale.

Even factors within the number of ports on the spine devices will determine the possible scale steps; not all intermediate stages are possible. For instance, a scale step with 12 leaves would require:

- Withtwo2 spines, 16/2 = 8 connections from each leaf; 12x8 = 96 fabric-spine connections to each spine, which is more than 64—this is not possible.
- With three spines, 16/3 = 5.34 connections from each leaf—this is not possible.

Moving to four spines, of course, makes 16 leaves possible; to reach 12 leaves, the operator can simply install four spines, leaving 16 ports open on each spine. Using unequal connection counts between the leaves and spines allows a greater variety of designs, but these are left as an exercise for the reader.

Splitters

Splitters are either rack mount or “octopus cable” units that can divide a high-speed interface into multiple lower-speed ports. Both optical and copper (DAC) splitters are widely available. The most common sizes divide a 100 Gbps port into ten 10 Gbps connections, four 25 Gbps connections, or two 50 Gbps connections. Using splitters allows the operator to have the same 100 Gbps port along the entire fabric edge, while also allowing a variety of servers to connect to the fabric.

While optical splitters do provide the operator with more configuration flexibility using a smaller range of hardware (even to the point of building a single SKU fabric), they can create operational and cabling challenges. Operationally, optical splitters are optical devices, and are subject to the same failures and troubleshooting difficulties as other optical devices. Further, they can be difficult to manage in terms of cabling plant.

External Fabric Connections

Operators are often tempted to connect a fabric to the network core through a standard 10 Gbps, 40 Gbps, 25 Gbps, or 100 Gbps port on a standard ToR switch. Data center switches, however, are designed for low-oversubscription situations and consistent link speeds. As a result, they usually do not have deep queues, strong QoS feature sets, or strong security and filtering capabilities built into the hardware. Attaching a pair of 10 Gbps core connections to a standard 32x100 Gbps or 64x100 Gbps switch will generally result in large amounts of dropped traffic and poor application performance.

Data center fabric switches should not be used for external connections to the fabric. Instead, a router designed for moving traffic between lower and higher speed links, such as Juniper Networks® QFX Series Switches or Juniper Networks MX Series 5G Universal Routing Platforms, should be used to create a border leaf node on the fabric. There are two ways to connect a border leaf device into the fabric.

If the border leaf device has enough switch ports to connect to *all* spine switches, the border leaf can directly replace a ToR switch. If the border leaf device does not have enough switch ports to connect to all the spine switches, it should be connected to a ToR switch like any other workload. In this case, the border leaf switch, and any other externally facing services, should be the only workloads connected to the ToR switch.

A border leaf router or other device *should not* be connected to a subset of the spine switches, as this can unbalance the fabric traffic flows, resulting in poor performance.

Conclusion

Three-stage IP fabrics based on the Clos design, initially designed for use in telephone networks, have been widely adopted for use in data center fabrics because they provide many benefits, including simplicity and scale-out design patterns. Understanding how to properly design these fabrics to scale over time, however, can be difficult. By following the fundamentals of IP fabric design covered in this white paper, users can deploy scale-out fabric designs with confidence.

About Juniper Networks

Juniper Networks brings simplicity to networking with products, solutions, and services that connect the world. Through engineering innovation, we remove the constraints and complexities of networking in the cloud era to solve the toughest challenges our customers and partners face daily. At Juniper Networks, we believe that the network is a resource for sharing knowledge and human advancement that changes the world. We are committed to imagining groundbreaking ways to deliver automated, scalable, and secure networks to move at the speed of business.

Corporate and Sales Headquarters
Juniper Networks, Inc.
1133 Innovation Way
Sunnyvale, CA 94089 USA
Phone: 888.JUNIPER (888.586.4737)
or +1.408.745.2000
Fax: +1.408.745.2100
www.juniper.net

APAC and EMEA Headquarters
Juniper Networks International B.V.
Boeing Avenue 240
1119 PZ Schiphol-Rijk
Amsterdam, The Netherlands
Phone: +31.0.207.125.700
Fax: +31.0.207.125.701

JUNIPER NETWORKS | Engineering Simplicity