White paper
Cisco public

Cisco's Massively Scalable Data Center Network Fabric Design and Operation

Contents

Massively scalable data center network fabric

Massively Scalable Data Centers (MSDCs) are large data centers, with thousands of physical servers (sometimes hundreds of thousands), that have been designed to scale in size and computing capacity with little impact on the existing infrastructure. Environments of this scale have a unique set of network requirements, with an emphasis on application performance, network simplicity and stability, visibility, easy troubleshooting, and easy life-cycle management, etc. Examples of MSDCs are large web/cloud providers that host large distributed applications such as social media, e-commerce, gaming, Software as a Service (SaaS), Artificial Intelligence and Machine Learning (AI/ML) workloads, etc. These large web/cloud providers are often also referred to as hyperscalers or cloud titans.

Cisco's MSDC Layer 3 IP fabric architecture is based on Cisco Nexus® 9000 Series Switches, which are designed to meet the requirements of such networks. In this white paper, we will first discuss the MSDC Layer 3 IP fabric design options that enable network scalability, simplicity, and stability. Next, we will talk about extensive automation and programmability features that can be leveraged to enable fast, zero-touch network deployment and real-time on-demand device provision. Subsequently, this white paper covers considerations on the unique telemetry features and related visibility applications Cisco provides to enable more visibility and faster root-cause analysis. Given the close relation of MSDC, IP-storage, and AI networks, we will talk about features that enable RDMA over Converged Ethernet (RoCEv2)[1] deployments for better application performance with best throughput, low latency, and lossless network, to achieve a better user experience. Last but not least, this white paper looks at operational tasks to address how to upgrade a Layer 3 IP fabric without service impact.

MSDC Layer 3 IP fabric design evolution

Historically, web-scale providers have built their MSDC networks with a spine-leaf Clos-based Layer 3 IP fabric architecture. These basic designs started with a two-tiered spine-leaf Clos design with 1G servers and 10G uplinks between spine and leaf, as shown in Figure 1.

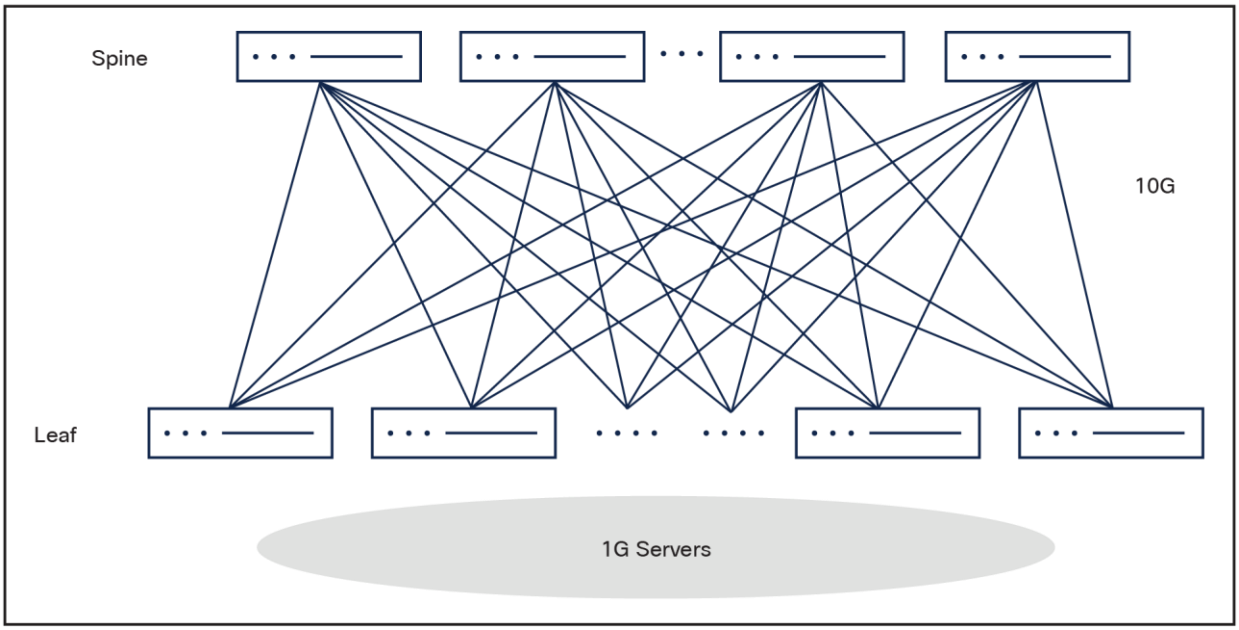

Figure 1.
Two-tiered spine-leaf design

In this two-tiered Clos architecture, every leaf layer switch is connected to each of the spine layer switches in a full-mesh topology. The leaf layer switch 1G downlink ports connect to 1G servers, and 10G uplink ports connect to spine layer switches. The spine layer is the backbone of the network and is responsible for interconnecting all leaf switches. The leaf layer and spine layer are connected with Layer 3 IP connectivity to provide Equal-Cost Multi-Path (ECMP) routing for both east-west server-to-server traffic and south-north server-to-user traffic. The subnet where the servers reside is local to a single or a pair of leaf switches. Such networks do not consider Layer 2 connectivity between multiple racks or live workload mobility. Connectivity to external network domains can be achieved either at the spine or leaf layer, depending on design requirements. Most deployments leverage a dedicated pair of border leaf switches.

As Ethernet network speeds evolved and standards were defined, the design evolved to support 10G servers and 40G uplinks; then, from the year 2017, most web-scale providers started to deploy MSDC with 25G servers and 100G uplinks. While in 2019 it was still common to look into deployments with 50G servers and 100G uplinks, in recent years, the adoption of 25G and 100G servers with 100G and 400G uplinks respectively became more predominant as shown in Figure 2.

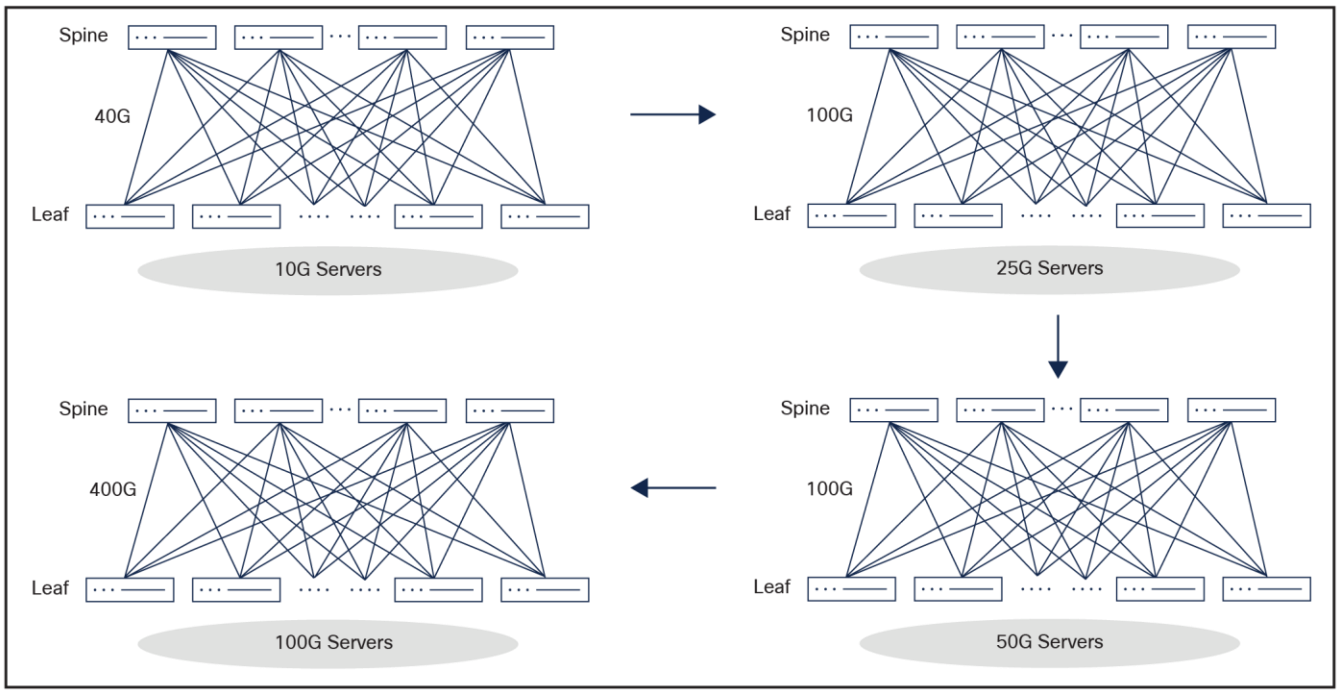

Figure 2.
MSDC server speed evolution

The technology transition with MSDC was not just about speed; as the number of users grew, the number of servers in MSDC grew exponentially to host large distributed applications, to a point where these two-tiered spine-leaf designs were required to scale out to support more servers. As a result, more evolved generations of MSDC designs were introduced to handle these increased scalability requirements; for example, a three-tiered spine-leaf design or hyperscale fabric plane Clos design. However, there are some common characteristics shared by these designs:

- The minimum MSDC network building blocks are defined as “server pods.”

  Each server pod is a two-tiered spine-leaf topology; it has a group of spine switches, leaf switches, and servers.

- Servers are grouped in server pods.

  Servers are connected to leaf switches. The maximum number of servers that each leaf switch can connect to is decided by the number of leaf switch downlink ports. Other factors may also contribute to it such as rack space and power.

- The maximum number of spine switches in a server pod is decided by the number of uplinks per leaf switch.

  For example, if a leaf switch has eight uplinks, then the maximum number of spine switches it can connect to is eight. If each leaf switch connects two uplinks to a single spine switch, then the maximum number of spine switches it can connect to is four.

- The maximum number of leaf switches a server pod can support is determined by the number of ports on each spine switch.

  For example, if a spine switch has 288 ports, then the maximum number of leaf switches a server pod can support is 288. This is the case for a single server pod. For multiple server pods, some interfaces on the spines must be reserved for inter-pod connectivity through a third tier as discussed later, lowering the number of leaf switches in a server pod.

- The leaf switches in the server pod are ToR (Top-of-Rack) fixed form-factor switches.
- It is a best-practice recommendation to use ToR fixed form-factor leaf switches in MSDC server pod design, as it provides more predictable latency and faster switching performance, in addition to space, power, and cooling savings.
- To scale the network to support more servers, more server pods are added.
- Different design options offer different ways to interconnect the server pods.

The section below describes the different design options for MSDC using Cisco Nexus 9000 Series Switches as examples. Based on your use case, you can choose the design that meets your requirements.

Cisco MSDC design example 1: Two-tiered spine-leaf topology

Cisco Nexus 9000 Series Switches provide a rich portfolio of 10G/25G/40G/50G/100G/400G/800G switches. Figure 3 shows an example of a two-tiered spine-leaf MSDC design with 25G servers and 100G uplinks using Cisco Nexus 9000 switches.

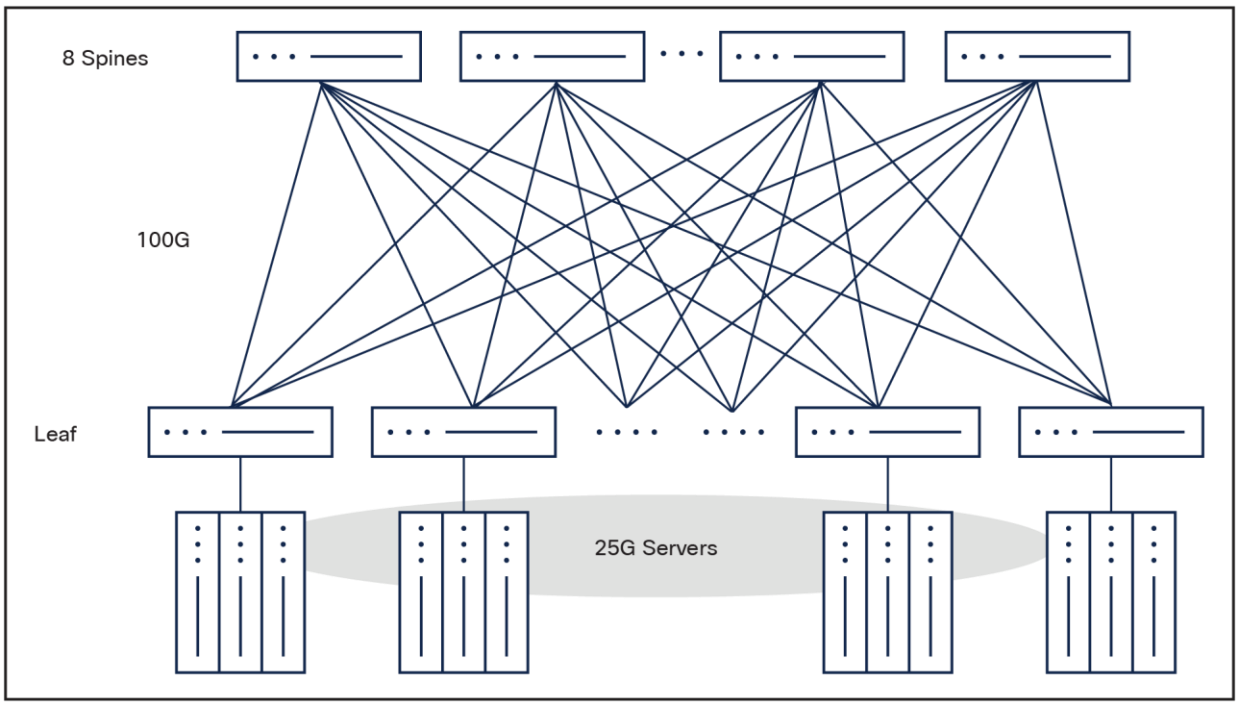

Figure 3.
Cisco MSDC design example 1

In this example, the leaf switches are Cisco Nexus 93400LD-H1 fixed ToR switches; each has 48x10/25/50G fiber ports and 4x400G QSFP-DD ports (or 16x100G ports using breakout cables). In this design, each leaf switch has 8x100G uplinks (breakout mode) connecting to spine layer switches. The number of spine switches can be two, four, or eight, depending on spine-layer high availability, load balancing, and scale requirements. In this example, each leaf switch has eight uplinks connecting to eight spine switches, one link per spine switch. All 48x25G downlink ports connect to 25G servers; each server is single-homed to one leaf switch or, dual-homed using a virtual Port-Channel (vPC) to two leaf switches. Some web-scale providers implement server high availability from the application layer and connect servers to a single ToR switch. In this example, a single-homed server is used for the sake of simplicity.

The spine switches in this example are modular Cisco Nexus 9808 switches, fully loaded with N9K-X9836DM-A (36x400G ports) line cards to support a total of 288x100/400G ports natively (or 1152x100G ports if using breakout cables) with 7+1 Fabric Module (FM) redundancy. Each 100G downlink connects to one leaf switch; each spine switch can support up to 288 leaf switches using native 100G ports or up to 1152 leaf switches if using breakout cables.

This two-tiered spine-leaf design is one server pod. The characteristics of this server pod are summarized in Table 1.

Table 1. Cisco MSDC design example 1

Two-tiered spine-and-leaf design example								
	Maximum number of uplinks	**Number of uplinks to each spine**	**Uplink speed (Gbps)**	**Number of downlinks**	**Downlink speed (Gbps)**	**Oversubscription ratio**	**Maximum number of single-homed server to each leaf**	**Maximum number of switches per server pod**
Leaf (Cisco Nexus 93400LD-H1)	8	1	100	48	25	1.5	48	288
Spine (Cisco Nexus 9808 fully loaded with N9K-X9836DM-A, total: 288x100/400G native ports, 7+1 FM redundancy)	N/A	N/A	N/A	288	100	N/A	N/A	8
Maximum number of single-homed servers per server pod	13824							

Based on Table 1, some of the key characteristics of this design are:

- Leaf switch: 8x100G uplinks, 48x25G downlinks, the oversubscription ratio at leaf switch is 1.5:1 (48x25G: 8x100G). Each leaf switch has one uplink to each spine switch.
- Spine switch: 288x100G downlinks.
- There are 8 spine switches in this server pod.
- There are 288 leaf switches in this server pod.
- The maximum number of single-homed physical servers per pod is 13,824.

Cisco provides a very rich portfolio of Cisco Nexus 9000 fixed and modular switches with different port densities and speeds (up to 800G). As an example, the choices of switches for different layers include the following (please note that the choices are not limited to the examples given below):

- Leaf switches: Cisco Nexus 9300-FX3 and Cisco Nexus 9300-GX2 series switches.

- Spine switches: Cisco Nexus 9300-GX2 and Cisco Nexus 9300-H2R fixed switches or Cisco Nexus 9500 and Cisco Nexus 9800 modular switches with different line-card options.

For more information, please refer to the Cisco Nexus 9000 series switches data sheet.

Cisco MSDC design example 2: Three-tiered spine-leaf topology

The second MSDC design example is a three-tiered spine-leaf topology. Tier-1 is made up of leaf switches that connect to servers. Tier-2 is made up of fully meshed spine switches connecting to leaf switches. Tier-1 and Tier-2 together form a server pod; the capacity of each server pod (that is, the maximum number of servers per server pod) depends on the choice of Tier-1 and Tier-2 switches. Tier-3 is made up of super-spine switches, which interconnect the server pods. The Tier-2 spine switches in each server pod are fully meshed, connecting to the Tier-3 super-spine switches. With a three-tiered spine-leaf design, in order to scale the number of servers, more server pods are added. To scale the bandwidth interconnecting the server pods, more super-spine switches are added. The maximum number of server pods that this design can support depends on the super-spine switch capacity.

Figure 4 shows an example of a three-tiered spine-leaf MSDC design with 25G servers and 100G uplinks using Cisco Nexus 9000 switches.

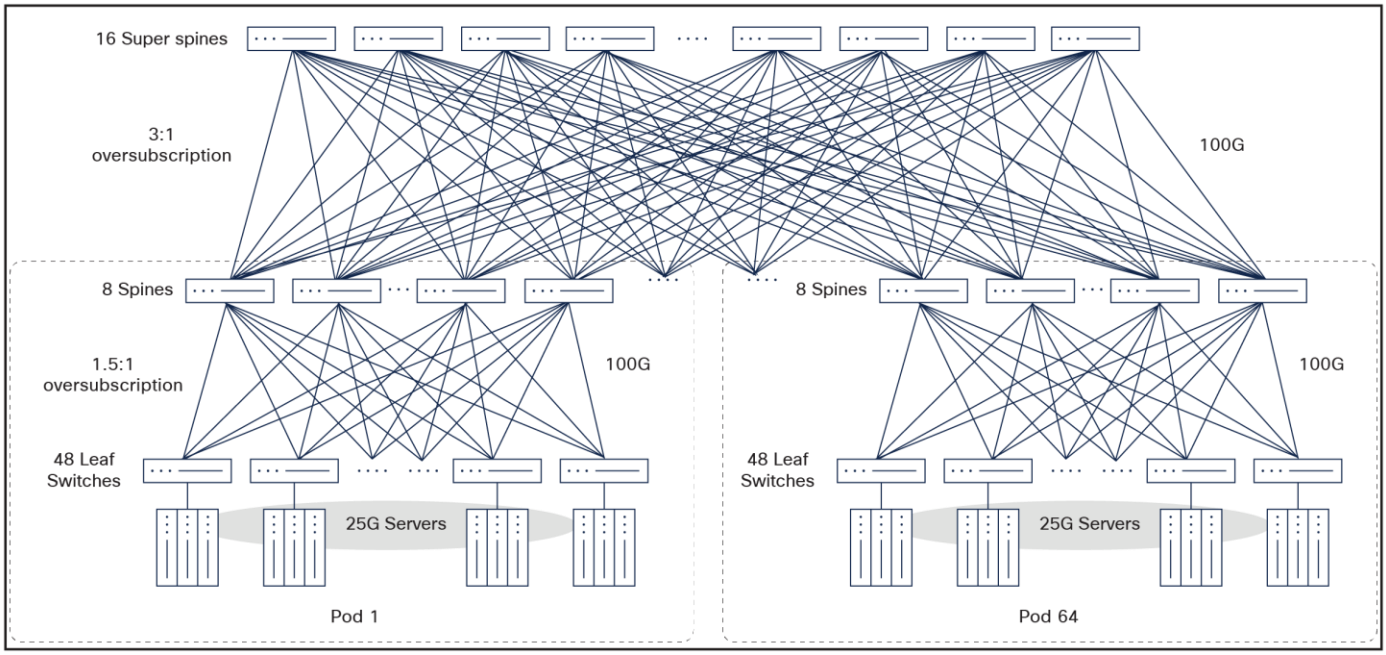

Figure 4.
Cisco MSDC design example 2

In this example, the leaf switches are also based on Cisco Nexus 93400LD-H1 and servers are single-homed as in the previous example for consistency. The spine switches are Cisco Nexus 9364D-GX2 switches, which support a total of 64 ports of 100G/400G natively or 256x100G ports using breakout cables. In each spine, 48x100G downlinks are connected to leaf switches, and 16x100G uplinks are connected to super-spine switches. In this example, there are 16 super-spine switches. Therefore, the spine-to-super-spine oversubscription ratio is 3:1 (48x100G: 16x100G). Or 1:1.3 undersubscribed if using 16x400G instead.

The super-spine switches are modular Cisco Nexus 9808 switches fully loaded with N9K-X9836DM-A. There is a total of 288x100/400G native ports or 1152x100G using breakout. While this configuration can support up to 144 server pods using breakout, this example illustrates 64 server pods only. On each super-spine switch, 512x100G downlinks connect to each spine switch fully meshed. For this design, every eight spine switches in Tier-2 plus 48 leaf switches in Tier-1 form a server pod; all of the server pods plus the super-spine switches in Tier-3 form a cluster, then each super-spine in this design can support 64 server pods for the cluster and scalable to 144 server pods.

The characteristics of this design example with the number of switches at each layer, the oversubscription ratio at each layer, the number of uplinks and downlinks, the number of server pods, and the maximum number of supported servers per cluster are summarized in Table 2.

Table 2. Cisco MSDC design example 2

Three-tiered spine-and-leaf design example with fully meshed Clos at each tier									
	Maximum number of uplinks	**Number of uplinks to each spine**	**Uplink speed (Gbps)**	**Number of downlinks**	**Downlink speed (Gbps)**	**Oversubscription ratio**	**Maximum number of single-homed servers to each leaf**	**Maximum number of switches per server pod**	**Maximum number of switches per cluster**
Leaf (Cisco Nexus 93400LD-H1)	8	1	100	48	25	1.5	48	48	3072
Spine (Cisco Nexus 9364D-GX2: 64 ports of 100/400G)	16	1	100	48	100	3	N/A	8	512
Super spine (Cisco Nexus 9808 fully loaded with N9K-X9836DM-A, total: 512x100G ports (breakout), 7+1 FM redundancy)	N/A	N/A	N/A	512	100	N/A	N/A	N/A	16
Maximum number of single-home servers	2304								

Three-tiered spine-and-leaf design example with fully meshed Clos at each tier									
per server pod									
Maximum number of server pod per cluster	64								
Maximum number of single-home servers per cluster	147456								

Based on Table 2, some of the key characteristics of this design are:

- For each server pod, there are 8 spine switches, 48 leaf switches, and 2304 single-home servers.
- The maximum number of server pods per cluster is 64 for this specific design choice, as it is only using 512 ports out of 1152 from the super spines in breakout mode.
- There are 16 super spines in the cluster to interconnect the 64 server pods.
- The maximum number of single-homed servers per cluster is 147,456, more than 10 times the number of servers supported in example 1.

Cisco provides a very rich portfolio of Cisco Nexus 9000 fixed and modular switches with different port densities and speeds (up to 800G). Similarly to the previous example, the choices of switches for different layers are as follows, and not limited to the examples given below:

- Leaf switches: Cisco Nexus 9300-FX3 and Cisco Nexus 9300-GX2 series switches.
- Spine switches: Cisco Nexus 9300-GX2 and Cisco Nexus 9300-H2R fixed switches or Cisco Nexus 9500 or Cisco Nexus 9800 modular switches with different line-card options.
- Super-spine switches: Cisco Nexus 9500 or Cisco Nexus 9800 modular switches with different line-card options.

For more information, please refer to the Cisco Nexus 9000 Series Switches data sheet.

Cisco MSDC design example 3: Hyperscale fabric plane Clos design

Over the years, several very large web providers redefined their Data Center design by introducing a new MSDC design: hyperscale fabric. This may also be referred to as a “butterfly” scaling fabric. Some of the key characteristics of the hyperscale fabric are the following:[2,3]

- Hyperscale is about scalability and elegant simplicity.
- Network switches are layered and disaggregated.

The principal network building block is a server pod. In our example, the size of a server pod is limited to 64 server racks; each server rack has one ToR switch. The 64 switches connect fully meshed to four upper-level devices called fabric switches.

- Server pods are cross-connected.

Each fabric switch is numbered (1, 2, 3, 4, etc.), and each number corresponds to a higher-level layer of switches, which forms a spine plane. Each fabric switch connects to 64 spine switches in the plane design. The number of spine planes in this case is 4. The number of spine switches in each spine plane is up to 64. This is shown in Figure 5.

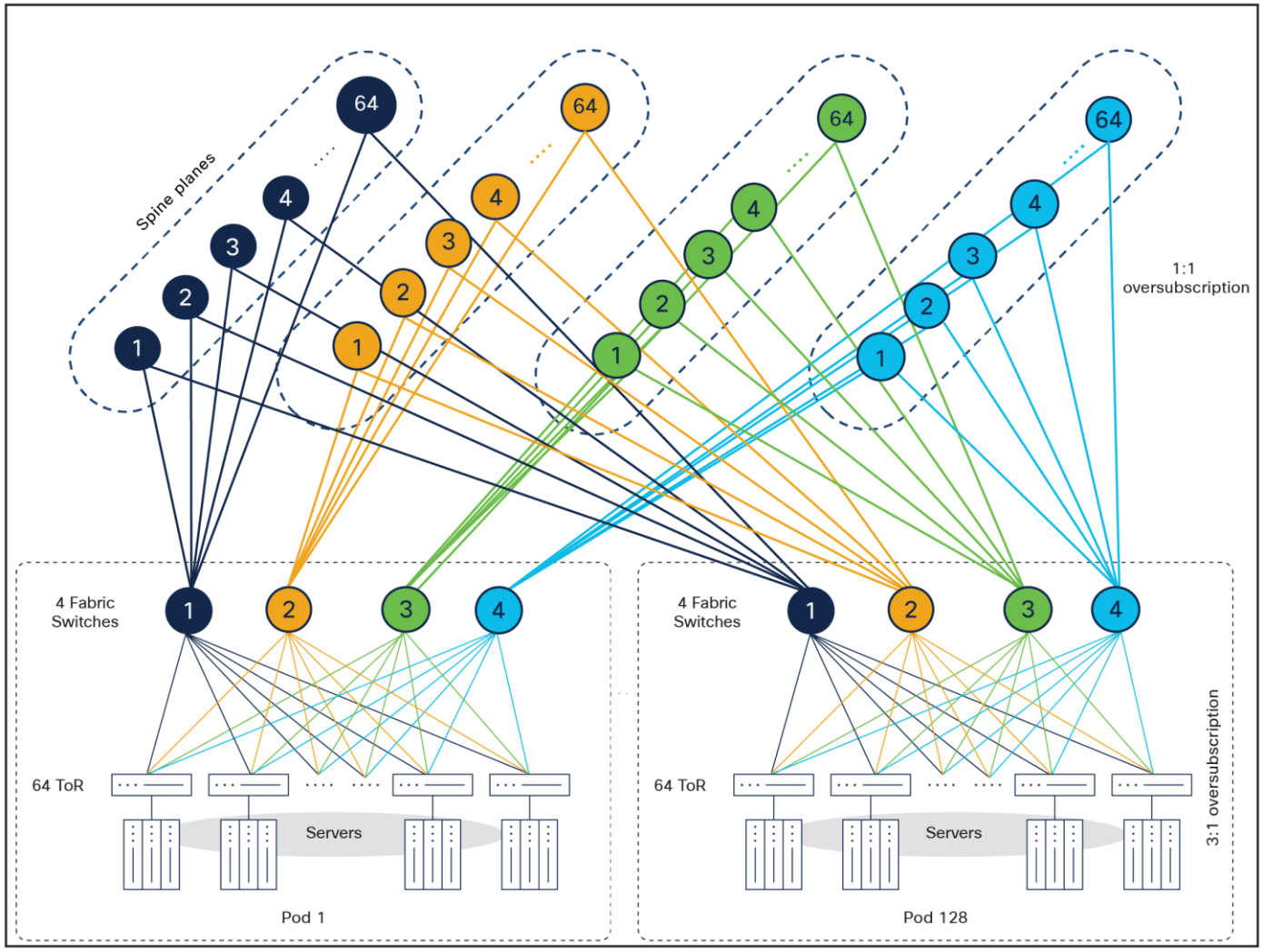

Figure 5.
Cisco MSDC design example 3

The benefits of this design are:

- It is highly modular and scalable in all dimensions. When more compute capacity is needed, more server pods are added. When more intra-fabric network capacity is needed, more spine switches can be added on all of the spine planes.
- Each pod has only 64 server racks; it requires only basic mid-size fabric switches to aggregate the ToRs.
- Each pod is served by a set of four fabric switches or more depending on scaling requirements coming from the ToRs; the oversubscription ratio of each pod is easily controlled by the number of spine switches on the spine planes. For example, if the starting point were 16 spine switches in each plane, the fabric oversubscription from pod to pod would be 4:1 (64x100G:16x100G). This network capacity can be increased in granular steps or quickly jump to a 2:1 oversubscription ratio with 32 spines, or even a full 1:1 non-oversubscribed state at once with a total of 64 spines per plane as illustrated.
- The application flow path is selected at a ToR switch; once the path is chosen (for example, in Figure 5, if a blue path is chosen), then that flow will stay on the blue path in the network. It is easy to troubleshoot.

- The fabric plane design provides many parallel paths between servers, and each path has equal performance. The load in the fabric is distributed and different flows can take different paths coinciding on this large distributed infrastructure. The fabric can survive multiple device failures, component failures, and link failures without any production impact.
- Typically deployed using fixed format switches on all layers, which provide deterministic latency with single-stage devices. Generally, modular switches operate in store-and-forward mode with an additional forwarding layer inside the chassis with fabric modules and line cards, and that increases latency.

The section below shows an example of a hyperscale fabric plane MSDC design using Cisco Nexus 9000 Series Switches.

In this example, the leaf switches use Cisco Nexus 93400LD-H1; the same as in previous examples. In this design, each leaf switch has 4x100G uplinks (breakout mode) connecting to fabric layer switches, with one uplink to each fabric switch. The number of fabric switches per server pod is four. All of the downlink 48x25G ports connect to servers; each server is single-homed.

The fabric switches are Cisco Nexus 9332D-GX2 switches. Each fabric switch supports a maximum of 128 ports of 100G using breakout or native 32x100/400G. There are 64x100G downlinks connected to leaf switches. The number of leaf switches per server pod is 64. There are 64x100G uplinks connected to spine plane switches, with one uplink to each spine plane switch. The number of spine plane switches per plane is 64. The oversubscription ratio at the fabric switch is 1:1 (64x100G: 64x100G).

The spine-plane switches are also based on Cisco Nexus 9332D-GX2 switches. Each spine-plane switch supports a maximum of 128 ports of 100G using breakout or native 32x100/400G. There are 128x100G downlinks connected to fabric switches. There are four spine planes.

The characteristics of this example of a hyperscale fabric plane Clos design (that is, the number of switches at each layer, the oversubscription ratio at each layer, the number of uplinks and downlinks, the number of server pods, and the maximum number of supported servers per hyperscale fabric) are summarized in Table 3.

Table 3. Cisco MSDC design example 3

Hyperscale fabric plane Clos design example										
	Maximum number of uplinks	**Number of uplinks to each fabric per spine switch**	**Uplink speed (Gbps)**	**Number of downlinks**	**Downlink speed (Gbps)**	**Number of spine planes**	**Oversubscription ratio**	**Maximum number of single-homed servers to each ToR**	**Number of switches per server pod**	**Maximum number of switches per hyperscale fabric**
ToR (Cisco Nexus 93400LD-H1)	8	1	100	48	25	N/A	3	48	64	8192
Fabric switch (Cisco Nexus 9332D-GX2 with 128 ports of 100G (breakout))	64	1	100	64	100	N/A	1	N/A	4	512
Spine plane switch (Cisco Nexus 9332D-GX2 with 128 ports of 100G (breakout))	N/A	N/A	N/A	128	100	4	N/A	N/A	N/A	256
Maximum number of single-homed servers per	3072									

Hyperscale fabric plane Clos design example										
server pod										
Maximum number of server pod per hyperscale fabric	128									
Maximum number of single-homed servers per hyperscale fabric	393216									

Based on Table 3, some of the key characteristics of this design are:

- For each server pod, there are four fabric switches, 64 leaf switches, and 3072 single-home servers.
- The maximum number of server pods per hyperscale fabric is 128.
- There are four spine planes, with 64 spine switches in each spine plane to interconnect the 128 server pods.
- The maximum number of single-homed servers per hyperscale fabric is 393,216, which is about 25 times the number of servers supported in example 1.

Cisco provides a very rich portfolio of Cisco Nexus 9000 Series Fixed Switches with different port densities and speeds (up to 800G). Similarly to the previous examples, the choices of switches for different layers include the following, and are not limited to the examples given below:

- Leaf switches: Cisco Nexus 9300-FX3 and Cisco Nexus 9300-GX2 Series Fixed Switches.
- Fabric switches: Cisco Nexus 9300-GX2 and Cisco Nexus 9300-H2R switches
- Spine-plane switches: Cisco Nexus 9300-GX2 and Cisco Nexus 9300-H2R switches

For more information, please refer to the Cisco Nexus 9000 Series Switches data sheet.

Cisco MSDC design example 4: AI Networking with Rail and Plane design

With the explosion of Large Language Models (LLM) and the increasing number of AI workloads in recent years, web-scale and cloud providers redesigned their large scalable data center fabrics to host AI clusters running these specialized workloads[4]. This led to another variation of MSDC design to support dedicated AI networking fabrics using the concept of front-end and back-end networks. The need for both an AI front-end and back-end network arises from the distinct roles and requirements they fulfill within an AI Data Center. The front-end network manages user interactions, data ingestion, checkpointing, and general-purpose computing tasks. The back-end network is designed for high-speed, lossless Ethernet, and high-volume data transfer between AI accelerators such as Graphics Processing Units (GPU). It handles the intensive computational tasks required for distributed training and inference of large AI models. RoCEv2 support is a foundational requirement in this network for exclusive GPU-to-GPU communication.

All previous designs covered in this document fall under the front-end network category; therefore, this section covers only the uniqueness of an AI back-end network whose designs may vary depending on AI accelerator vendors. The Cisco Data Center Networking Blueprint for AI/ML Applications white paper provides examples of a flat design supporting AI clusters. In this example, we illustrate a rail-optimized design mostly recommended by NVIDIA[5] and how it can scale with a plane design. Another variation is a three-ply design as recommended by Intel®[6] where each node is dual connected to three different leaf switches with each leaf mapping to a dedicated set of spine switches. These design choices strictly depend on Collective Communication Library (CCL) optimizations to improve overall performance.

In our example, the principal network building block is an AI pod. Each AI pod has 32 compute nodes with an 8-way GPU architecture connecting to 8-leaf switches in the back-end network. Each set of 8-leaf switches connects fully meshed to 32-spine switches composing an AI zone with 4 AI pods. The AI zones can be subsequently scaled and interconnected using a hyperscale fabric plane design as described in example 3 and shown in Figure 6.

- Using a 1:1 ratio, the network interface and GPU pair from each compute node within an AI Pod are connected to the same leaf switch. For example, GPU1:NIC1 connects to Leaf 1, GPU2:NIC2 connects to Leaf 2, and so on, defining a rail-optimized design in a back-end network.

The compute nodes are also connected to the front-end network for management, storage, and external access.

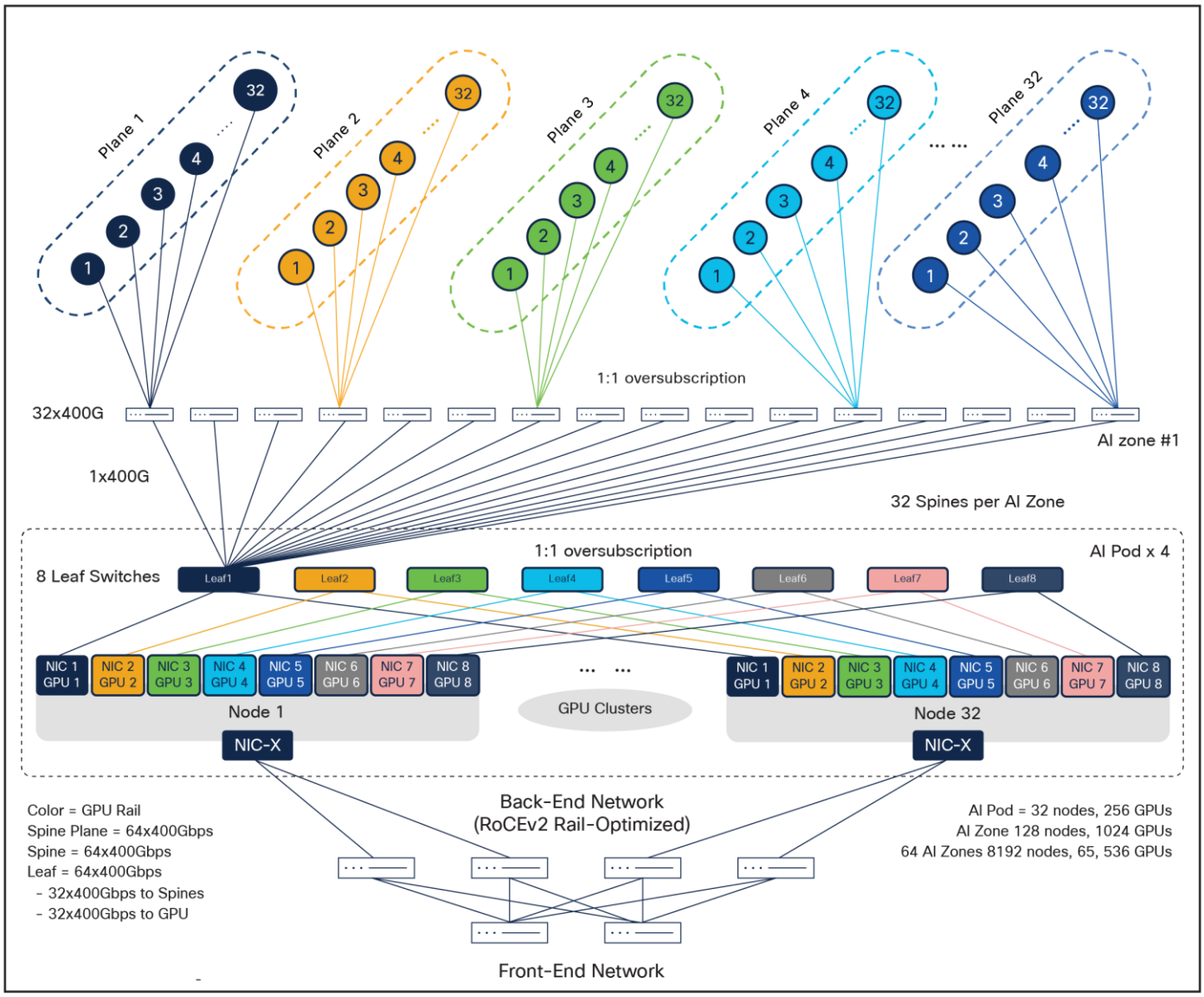

Figure 6.
Cisco MSDC design example 4

The benefits of this design are:

- It is highly modular and scalable in all dimensions. When more compute capacity is needed, more AI pods/zones are added to expand the AI cluster.
- AI zones are interconnected using a plane design, the same concept as the hyperscale fabric plane design with another tier, or directly connected spine-to-spine in a mesh plane depending on the desirable scale.
- Each AI pod has 32 compute nodes and is served by a set of 8 leaf switches in a rail-optimized fabric design.
- Rail-optimized network topology helps maximize performance while minimizing network interference between flows.

The section below shows an example of a rail and plane MSDC design using Cisco Nexus 9000 Series Switches.

In this example, the leaf switches use Cisco Nexus 9364D-GX2. Each leaf switch has 32x400G uplinks connecting to the spine switches. The number of spines per AI zone is 32. The remaining 32x400G ports are downlinks connected to compute nodes; per GPU:NIC pair at 400Gbps. The oversubscription ratio at the leaf switch is 1:1 (32x400G: 32x400G), forming a non-blocking fabric.

The spines are also based on Cisco Nexus 9364D-GX2 switches. In this example, 32x400G ports are downlinks connected to leaf switches. The other 32x400G ports are reserved for a plane design. The same switch model is used for the plane switches with 64x400G ports interconnecting different AI zones. This design is scalable to 32 planes with 32 switches each allowing for a 1:1 oversubscription ratio end to end.

The characteristics of this example are summarized in Table 4.

Table 4. Cisco MSDC design example 4

Rail-Optimized design with Hyperscale fabric plane example										
	Maximum number of uplinks	**Number of uplinks to each spine switch or plane switch**	**Uplink speed (Gbps)**	**Number of downlinks**	**Downlink speed (Gbps)**	**Number of spine per AI zone**	**Oversubscription ratio**	**Maximum number of compute nodes on each leaf switch**	**Number of leaf switches per AI pod**	**Maximum number of switches per rail-optimized fabric**
Leaf switch (Cisco Nexus 9364D-GX2 with 64 ports of 400G	32	1	400	32	400	N/A	1:1	32	8	2048
Spine switch	32	1	400	32	400	32	1:1	N/A	N/A	2048

Rail-Optimized design with Hyperscale fabric plane example										
(Cisco Nexus 9364D-GX2 with 64 ports of 400G										
Plane switch (Cisco Nexus 9364D-GX2 with 64 ports of 400G	N/A	N/A		64	400	N/A	N/A	N/A	N/A	1024
Maximum number of compute nodes per AI pod	32									
Maximum number of AI pods per rail-optimized fabric	256									
Maximum number of compute nodes per rail-optimized fabric	8192									

Based on Table 4, some of the key characteristics of this design are:

- The maximum number of compute nodes per AI pod is 32, scalable to 128 within an AI zone.
- For each AI zone, there are 4 AI pods with a total of 32 spines and 32 leaf switches.
- The reserved 32x400G ports from each spine can connect to 32 plane switches within its dedicated plane. For example, all spines 1 from each AI zone are connected to Plane 1, and all spines 2 from each AI zone are connected to Plane 2, and so on.
- Each spine plane can interconnect 64 AI zones using 32 planes to achieve a 1:1 oversubscription ratio.
- With 64 AI zones, there are 8192 compute nodes with 65,536 GPUs at 400Gbps.

Cisco provides a very rich portfolio of Cisco Nexus 9000 Series Fixed Switches with different port densities and speeds (up to 800G). Similarly to the previous examples, the choices of switches for different layers include the following, and are not limited to the examples given below:

- Leaf switches: Cisco Nexus 9300-GX2 and Cisco Nexus 9300-H2R series switches
- Spine switches: Cisco Nexus 9300-GX2 and Cisco Nexus 9300-H2R series switches for fixed or modular Cisco Nexus 9800 series for higher density

For more information, please refer to the Cisco Nexus 9000 Series Switches data sheet.

MSDC routing design considerations

For an MSDC, since the number of devices in the network is very large, there is a requirement to design the network with a minimal feature set and to select a routing protocol that is simple, stable, and scalable and that supports some types of traffic engineering. EBGP has been chosen by many MSDCs as the only routing protocol; an extensive description is available as part of RFC 7938 – Use of BGP for Routing in Large-Scale Data Centers.[7] In recent years, there have been multiple IETF working groups that have formed to work on routing solutions specific to MSDCs to further improve Data Center routing efficiency. Some of these solutions are listed below:

- Dynamic Flooding on Dense Graphs: https://tools.ietf.org/html/draft-ietf-lsr-dynamic-flooding
- Link State Vector Routing (LSVR): https://datatracker.ietf.org/wg/lsvr/about/
- Routing In Fat Tree (RIFT): https://datatracker.ietf.org/wg/rift/about/

The goal of the working groups is to employ the best attributes of link-state and distance-vector protocols while eliminating some of the negatives; including minimizing routes, fast convergence, dynamic topology detection while reducing flooding, wide propagation of updates, etc. Cisco has been actively participating in some of those working groups and has shipping solutions as part of Link-State Routing (IS-IS/OSPF) as well as the Distance/Path Vector Routing Protocol (BGP). Further enhancements as part of RFC 5549 - Advertising IPv4 Network Layer Reachability Information with an IPv6 Next Hop[8] exist to allow the usage of IPv6 addressing simplification while still being able to have IPv4 applications present.

The two examples below show an MSDC design with Layer 3 IP fabric, running only EBGP. It illustrates how EBGP single-hop sessions are established over direct point-to-point Layer 3 links between different tiers, and how to allocate EBGP Autonomous System Numbers (ASN).

EBGP ASN design examples for MSDC

Figure 7 shows an example of an EBGP ASN allocation scheme for an MSDC, also known as Multi-AS. A single ASN is assigned on all super-spine switches, a unique ASN is assigned to all spine switches in each pod, and a unique ASN is assigned to each leaf or pair of ToR switches in each pod, assuming that the server is dual-homed to a pair of ToR switches. Private ASNs from 64512 through 65534 are used in the example. Cisco NX-OS supports 4-byte AS numbers to provide enough numbers of private ASNs for MSDC deployments.

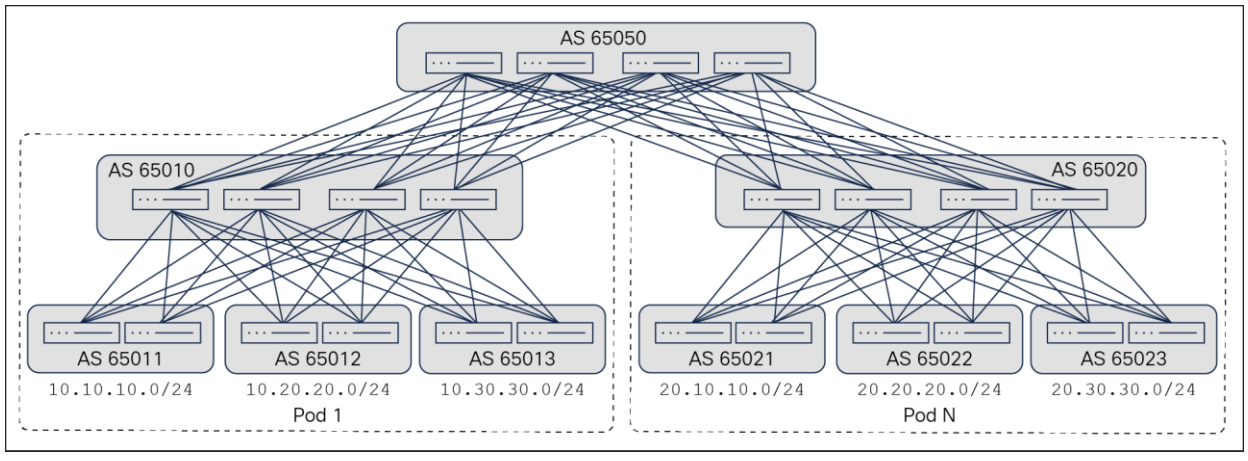

Figure 7.
EBGP ASN design example 1

Figure 8 shows another example of an EBGP ASN allocation scheme for an MSDC, also known as Dual-AS. A single ASN is assigned on all super-spine switches, a unique ASN is assigned to all spine switches in each pod, and a unique ASN is assigned to all ToR switches in each pod. Private ASNs from a range of 64512 to 65534 are used in this example.

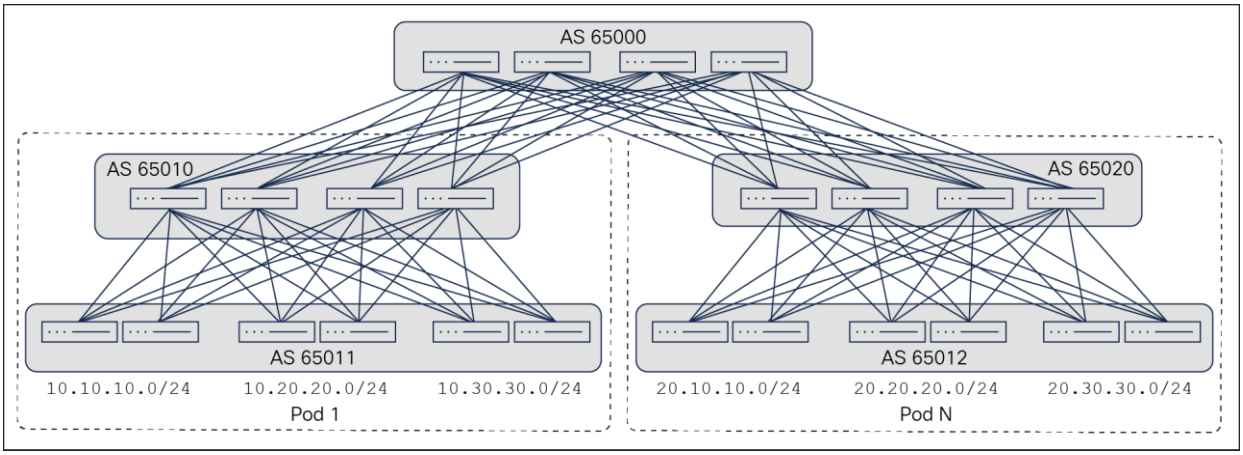

Figure 8.
EBGP ASN design example 2

For a Multi-AS design, each leaf or pair of leaf switches has a unique individual AS, while all the spines or super spines exist in a common AS. The path starts in a source AS and goes over an intermediate AS to the destination AS without the need to modify the default BGP behavior. The dual-AS design, however, breaks how eBGP works, as the AS-path would start and end in the same AS, which violates the BGP AS-path loop prevention mechanism.

To overcome communication issues, additional considerations need to be implemented for BGP to ignore the AS-path violation. One option is to disable the AS-path loop prevention on the spine with “as-override”; or a second option is to configure the leaf with “allowas-in.” Either option will yield the same result but through different techniques. In “as-override” the remote AS is replaced with its own AS in the AS-path. This is different in “allowas-in,” which allows its own AS to be present in the AS-path as many times as configured.[9]

ECMP in an MSDC

ECMP is used between different tiers for load sharing in an MSDC design. Using Figure 9 as an example, for each ToR switch, there are four uplinks connected to the four fabric switches, with one uplink per fabric switch. The links are configured as Layer 3 point-to-point IP connections with EBGP neighbors.

Cisco Nexus 9000 Series Switches support Layer 3 ECMP Dynamic Load Balancing (DLB) starting with NX-OS release 10.5(1). When a switch has several equal-cost paths to a destination, it uses a hashing algorithm, considering factors like source and destination IP addresses, port numbers, and protocol type, to decide the path for each packet. Traditional load balancing keeps the path constant unless the network topology changes or an administrator reconfigures it. In contrast, Layer 3 ECMP DLB on Cisco Nexus 9000 switches dynamically adjust path selection based on current network conditions, monitoring traffic loads to distribute traffic efficiently across all available paths. The Layer 3 ECMP DLB is supported on RoCEv2 with leaf and spine topologies that is especially used in back-end AI/ML training networks.

The number of ECMPs at each tier in the example shown in Figure 9 is as follows:

- ToR: 4 ECMP uplinks
- Fabric switch: 40 ECMP downlinks, 64 ECMP uplinks
- Spine plane switch: 64 ECMP downlinks

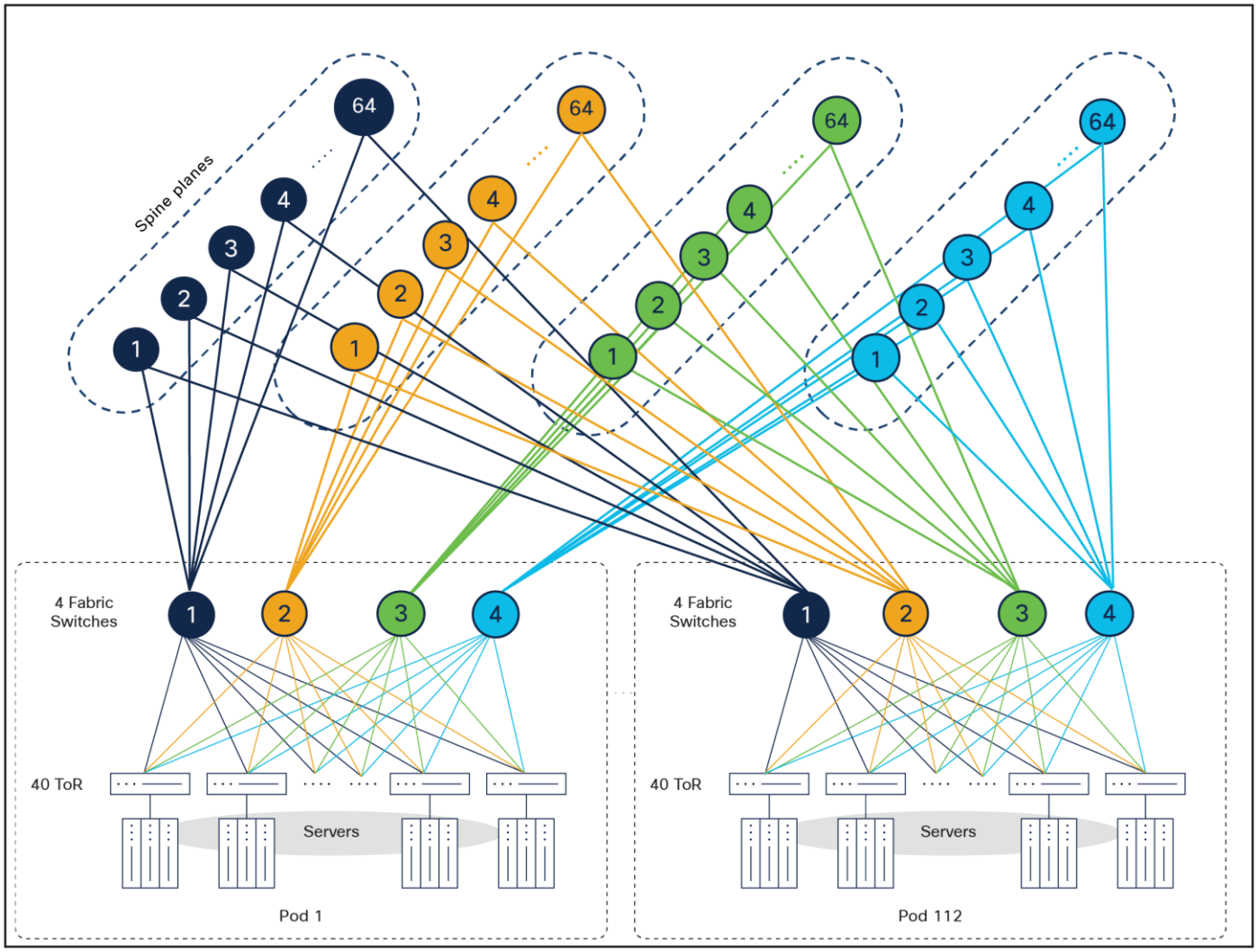

Figure 9.
ECMP in Cisco MSDC design example 3

Prefix advertisement

The prefix advertisement design depends on application communication requirements. The general principle is that ToR layer switches store more host routes. The other layer of switches is responsible for learning infrastructure routes and host-route summarization; they store more Longest Prefix Match (LPM) routes. Cisco Nexus 9000 Series Switches support different routing-scale profiles; each layer of switches boots up with different routing profiles. For example, ToR layer switches run host-route heavy-mode routing profiles, spine layer switches run with LPM heavy-mode routing profiles, etc.
For details, please refer to the Cisco Nexus Switches NX-OS Verified Scalability Guide for the corresponding release.

IPv4 and IPv6 dual stack and RFC 5549 in MSDC

The other trend in MSDC design is that MSDC fabrics are moving toward becoming IPv6-only. Many MSDCs are enabled with a dual-stack configuration, supporting both IPv4 and IPv6. This enables an easy migration to an IPv6-only MSDC in the future. Cisco Nexus 9000 switches fully support dual-stack IPv4 and IPv6, and also support RFC 5549 – Advertising IPv4 Network Layer Reachability Information with an IPv6 Next Hop. The primary motivation for RFC 5549 is to enable IPv6 devices to efficiently forward IPv4 packets over an IPv6 infrastructure without requiring the presence of dedicated IPv4 routing mechanisms. This capability is especially beneficial during the transition from IPv4 to IPv6, as it allows IPv6 devices to manage IPv4 traffic without requiring dual-stack (IPv4 and IPv6) configurations. It also provides more operational benefits as it requires less configuration to manage (single BGP peering session), less computational requirements, and simplifies network management while facilitating the transition to IPv6.

MSDC automation

For MSDCs, because the number of devices in the fabric is very large, managing configurations manually is not operationally efficient. MSDC customers typically use software-based approaches to introduce more automation and modularity into the network. The automation tools are used to handle different fabric topologies and form factors, creating a modular solution that can adapt to different-sized data centers. Cisco Nexus 9000 Series Switches support numerous capabilities to aid automation, for example:

- Shells and scripting: bash, guest shell, Python API, etc.
- Development and operations (DevOps) tools: Puppet, Chef, Ansible, SaltStack, Terraform, NX-SDK, etc.
- Model-driven programmability: OpenConfig YANG model, YANG model from Cisco, NETCONF agent, RESTConf agent, gRPC agent, etc.
- Docker container support with Cisco NX-OS

Below are some use cases that are enabled with Cisco Nexus 9000 Series Switches in MSDC environments:

- Power On Auto Provisioning (POAP): This feature is used widely in MSDC to automatically discover attached network devices, install software images, and base configuration files on the devices.
- Puppet/Chef/Ansible: They are widely used by MSDC operators to manage the servers for application deployment, configuration management, etc. With the support of Puppet, Chef, and Ansible on the Cisco Nexus 9000 Series Switches, MSDC operators can use the same tools to manage both network devices and servers consistently.
- Model-driven programmability: Model-driven programmability of Cisco NX-OS software devices allows you to automate the configuration and control of those devices. Data models are written in a standard, industry-defined language. Data modeling provides a programmatic and standards-based method of writing configurations to network devices, replacing the process of manual configuration. Cisco Nexus 9000 Series NX-OS supports both OpenConfig YANG models and native YANG models. The YANG model from Cisco is defined in the YANG data-modeling language but specific to NX-OS. With a Cisco Nexus 9000 MSDC network, native YANG models can be leveraged to provide more feature support and more flexibility. In a multivendor MSDC environment, an OpenConfig YANG model can be used to provide consistent vendor-neutral configurational and operational management but with less flexibility.

- Docker container on NX-OS: Customers develop their own applications packaged by Docker and install the applications in a Docker container running directly on Cisco Nexus switches. Use cases of Docker applications include telemetry streaming switch counters, collecting statistics from switches, monitoring switches, etc.

For more information on automation and programmability, please refer to the Cisco Nexus 9000 Series NX-OS Programmability Guide for the corresponding release.

MSDC telemetry and visibility

For MSDC customers, network monitoring, auditing, capacity planning, and troubleshooting are very important aspects of day-to-day operations. Cisco Nexus 9000 Series Switches support rich software and hardware capabilities to fulfill these requirements. For software features, Cisco NX-OS supports streaming telemetry features to continuously stream data out of the network, providing near-real-time network monitoring. For hardware capabilities, Cisco Nexus 9000 Cloud Scale switches support Flow Table Events (FTEs). The flow table from a Cisco Nexus 9000 Cloud Scale ASIC can generate notifications or events when certain conditions are detected in a flow packet. The detected events are then streamed to a collector for further analysis. MSDC customers can either build their own applications to process the data from the collector or use Cisco's Nexus Dashboard[10] platform for event analytics, resource utilization, flow and traffic analytics, etc.

Cisco Nexus Dashboard allows operators to minimize downtime by turning hardware and software telemetry into insights (including anomalies and advisories) to identify potential issues and recommendations to fix them, gathering years of experience under a single network operations platform. It can also take advantage of its analytics to learn more about energy usage, generate sustainability reports, detect configuration changes, and identify traffic behavior (including flow records, drops, congestion, latency, AI/ML RoCEv2 counters, and more). It also minimizes risk by providing pre and post-upgrade assistance and can enhance visibility by integrating tools from vendors such as VMware, Splunk, ServiceNow, Panduit, and many more. It incorporates a set of advanced alerting, baselining, correlation, and forecasting algorithms to provide a deep understanding of the behavior of the network.

For more information, please refer to the Cisco Nexus Dashboard data sheet.

MSDC with RoCEv2

During the past few years, RoCE technology has been widely adopted by MSDC customers because it enables low latency, low CPU utilization, and higher utilization of network bandwidth. Common RoCE use cases are distributed storage/database applications, gaming, Augmented Reality (AR), Virtual Reality (VR), machine learning and deep learning applications, etc. By adopting RDMA, applications have reported impressive performance improvement. Cisco Nexus 9000 Series Switches fully support RoCEv2 with rich features; for example, Quality of Service (QoS), Priority Flow Control (PFC), PFC watchdog, and Explicit Congestion Notification (ECN). Cisco Nexus 9000 switches have been widely deployed by hyperscale web providers in MSDCs running RDMA applications.

For more information regarding RDMA testing and configurations in Cisco Nexus switches, please refer to the white papers "Benefits of Remote Direct Memory Access Over Routed Fabrics[11]" and "RoCE Storage Implementation over NX-OS VXLAN Fabrics.[12]"

MSDC switch upgrade

For MSDCs, due to the large number of switches in the network, how to quickly and smoothly upgrade switch software without impacting data center services is very critical. Software upgrade needs to go through very strict procedures for MSDC customers: prepare very detailed upgrade steps, submit upgrade tickets, plan change windows, etc. Once an upgrade ticket is approved, during the change window, the upgrade is fully automated with automation tools. For example, it first identifies which switches need to be upgraded, then downloads the software onto the switches to provision the upgrade, and finally performs the upgrade.

Below are several examples of upgrades performed by MSDC customers:

Example 1: Customer A designs very good redundancy at the application layer among the server pods. If all the switches in a server pod need to be upgraded, the traffic is steered away from the corresponding server pod, and the application traffic is handled by other server pods. Automation tools will then upgrade all of the switches in the corresponding server pod.

Example 2: Customer B chooses to upgrade the switches in the corresponding server pod in batches. For example, for the fabric switches in the server pod, one or two fabric switches will be isolated using the Cisco Graceful Insertion and Removal (GIR) feature, and then a software upgrade is performed in the isolated switches without affecting the application service. Similarly, for the leaf layer switches, the GIR feature can also be used to isolate the corresponding leaf switches for software upgrades. At this layer, it is also common to perform In-Service Software Upgrades (ISSU) or Enhanced ISSU with no impact in the data plane during software upgrades. Typically, MSDC customers create custom maintenance mode profiles to ensure the corresponding switches stop receiving production traffic during upgrades.

For more information on GIR and ISSU, please refer to the Nondisruptive Data Center Network System Maintenance Best Practices with Cisco Nexus 9000 Series Switches[13] white paper.

Conclusion

Cisco's MSDC Layer 3 IP fabric architecture based on Cisco Nexus 9000 Series Switches is designed to meet the requirements of MSDC networks: network simplicity, repeatability, distributed architecture, easy troubleshooting, automation, easy life-cycle management, and related features. In this white paper, we have discussed the Cisco MSDC Layer3 IP fabric evolution and design options, including its design tools, extensive automation and programmability support, unique telemetry features, rich visibility applications, RoCEv2 features, etc. Cisco's MSDC solution enables scalability, simplified operation, and improved application performance with higher throughput, low latency, and a better user experience for your data center.

For more information

For additional information, see the following references:

[1] RDMA over Converged Ethernet

[2] Introducing the data center fabric and the next-generation Facebook data center network

[3] Project Altair: The Evolution of LinkedIn's Data Center Network

[4] RoCE networks for distributed AI training at scale

[5] Training Deep Learning Models at Scale: How NCCL Enables Best Performance on AI Data Center Networks

[6] Intel Gaudi 3 AI Accelerator Technical Paper

[7] RFC 7938 – Use of BGP for Routing in Large-Scale Data Centers

[8] RFC 5549 – Advertising IPv4 Network Layer Reachability Information with an IPv6 Next Hop

[9] VXLAN "eBGP" EVPN - the incarnation of a hybrid

[10] Cisco Nexus Dashboard

[11] Benefits of Remote Direct Memory Access Over Routed Fabrics

[12] RoCE Storage Implementation over NX-OS VXLAN Fabrics

[13] Nondisruptive Data Center Network System Maintenance Best Practices with Cisco Nexus 9000 Series Switches

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV Amsterdam,
The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at **https://www.cisco.com/go/offices**.

Printed in USA C11-743245-01 11/24