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 like social media, e-commerce, gaming, SaaS, 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 and 3000 series switches, which are designed to meet the requirements of such networks. In this white paper, we'll first discuss the MSDC Layer 3 IP fabric design options that enable network scalability, simplicity, and stability. Subsequently, we'll 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 and IP-storage networks, we'll talk about features that enable RoCEv2-like deployments for better application performance with best throughput, low latency, 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-like 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.

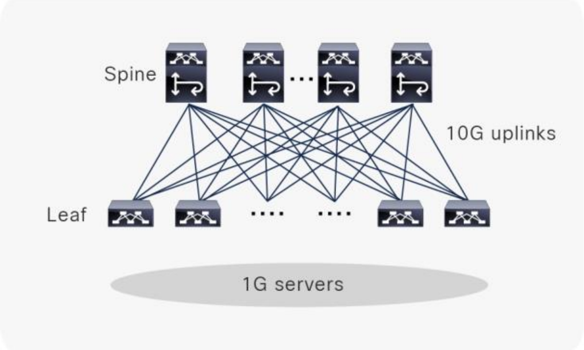

Figure 1.
Two-tiered spine-leaf design

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 ECMP (equal-cast multipath) paths for both east-west machine-to-machine traffic and south-north machine-to-user traffic. The subnet where the servers reside is local to a single or a pair of leaf switches. Such networks don't consider Layer 2 connectivity between multiple racks or live-host mobility.

As Ethernet network speeds evolved, 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, and, in 2019, some were looking into deployment with 50G servers and 100G uplinks, or even 100G servers with 400G uplinks, as shown in Figure 2.

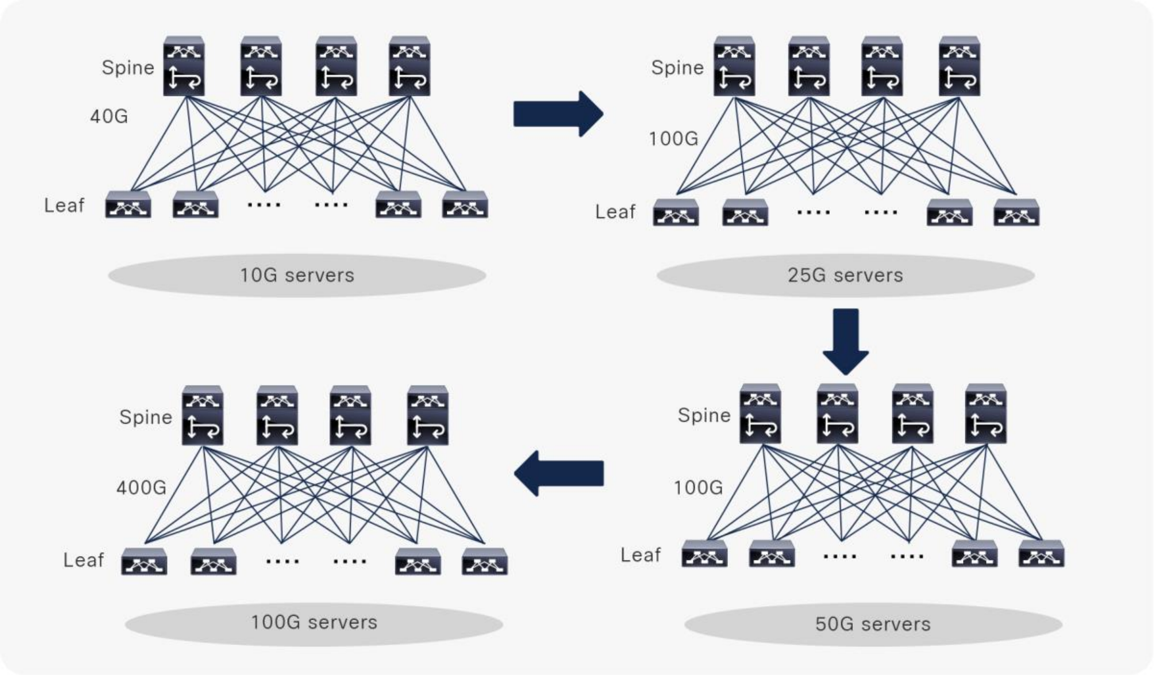

Figure 2.
MSDC server speed evolution

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 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 the scale; 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 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.

- 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 256 ports, then the maximum number of leaf switches a server pod can support is 256.

- 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.
- 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 Data Center Nexus 9000 and 3000 switches as examples. Based on your use case, you can choose the design that meets your requirements.

Cisco MSDC design example1: Two-tiered spine-leaf topology

Cisco Nexus 9000 and 3000 series switches provide a rich portfolio of 10G/40G/25G/100G/400G 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.

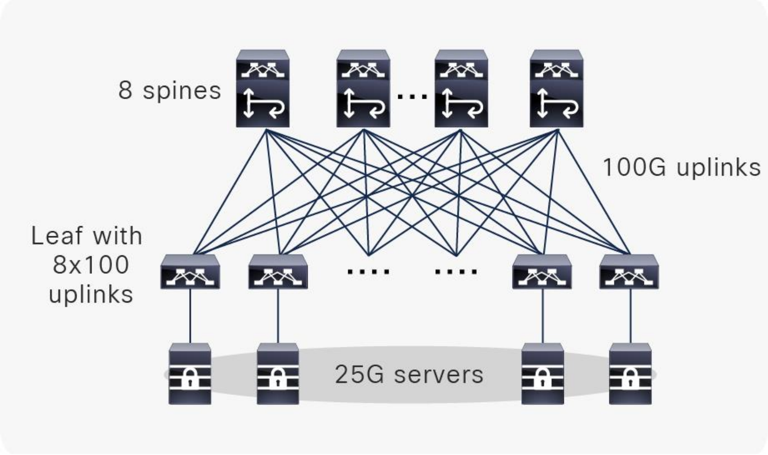

Figure 3.
Cisco MSDC design example 1

In this example, the leaf switches are Cisco Nexus 93240YC-FX2 fixed ToR switches; each has 48 x 10/25 Gbps fiber ports and 12 x 40/100 Gbps QSFP28 fiber ports. In this design, each leaf switch has 8 x 100 Gbps uplinks connecting to spine layer switches. The number of spine switches can be two, four, or eight, depending on spine-layer high-availability, load-balance, and scale requirements. In this example, each leaf switch has eight uplinks connecting to eight spine switches, one link per spine switch. All 48 x 25G downlink ports connect to 25G servers; each server single-home connects to one leaf switch or, using a virtual Port-Channel (vPC), dual-home connects 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-home server is used for the sake of simplicity.

The spine switches in this example are modular Cisco Nexus 9508 switches, fully loaded with N9K-X9732C-FX line cards to support total of 256 x 100G ports with 4+1 FM redundancy. Each 100G downlink connects to one leaf switch; each spine switch can support up to 256 leaf switches.

This two-tiered spine-leaf design is actually 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-home server to each leaf	Maximum number of switches per server pod
Leaf (Cisco Nexus 93240YC-FX2)	8	1	100	48	25	1.5	48	256
Spine (Cisco Nexus 9508 fully loaded with N9K-X9732C-FX, total: 256 x 100G ports, 4+1 FM redundancy)	N/A	N/A	N/A	256	100	N/A	N/A	8
Maximum number of single-home servers per server pod	12288							

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 (48 x 25G: 8 x 100G). Each leaf switch has one uplink to each spine switch.
- Spine switch: 256x100G downlinks.
- There are 8 spine switches in this server pod.
- There are 256 leaf switches in this server pod.
- The maximum number of single-home physical servers this server pod can support is 12,288.

Cisco provides a very rich portfolio of Cisco Nexus 9000 and Cisco Nexus 3000 fixed switches with different port densities and speed supports. Using Cisco Nexus 9000 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-FX2, Cisco Nexus 9300-FX, Cisco Nexus 9300-EX series switches, etc.
- Spine switches: Cisco Nexus 9332C and 9364C fixed spine switches, Cisco Nexus 9500 Series modular switches with different line-card options, etc.

For more information, please refer the data sheets on Cisco Nexus 9000 and Cisco Nexus 3000 series switches.

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.

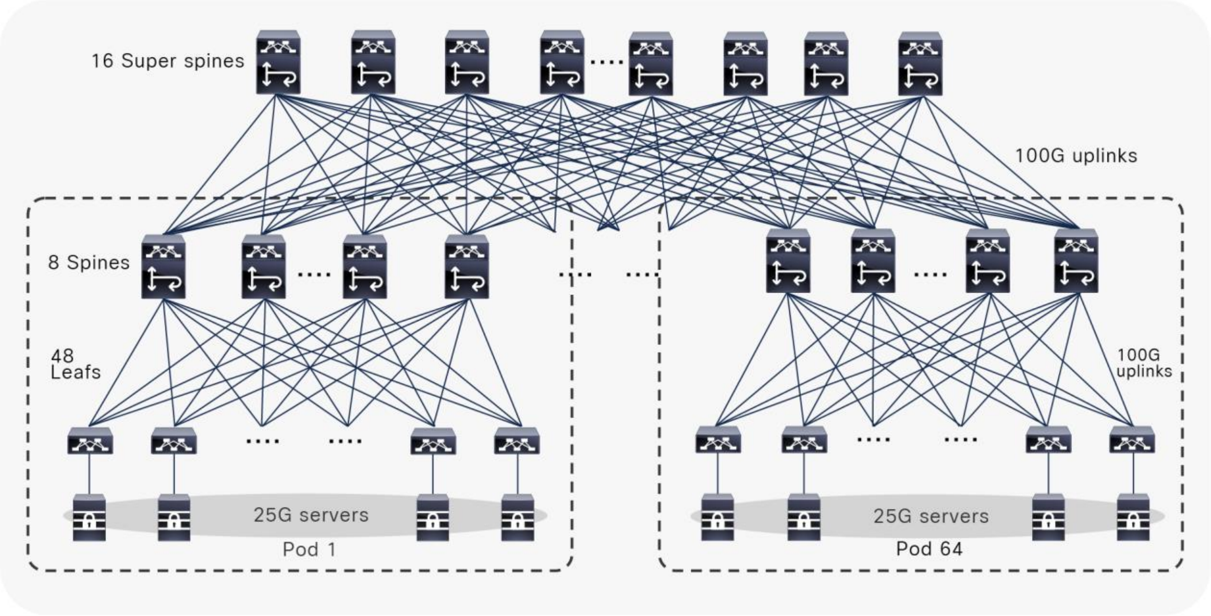

Figure 4.
Cisco MSDC design example 2

In this example, the leaf switches are using Cisco Nexus 93240YC-FX2; it has 48 x 10/25 Gbps fiber ports and 12 x 40/100 Gbps QSFP28 fiber ports. In this design, each leaf switch has 8 x 100 Gbps uplinks connecting to spine-layer switches. The number of spine switches can be two, four, or eight depending on spine-layer high-availability, load-balance, and scale requirements. In this example, eight spine switches are used. All of the downlink 48 x 25G ports connect to servers; each storage server single-home connects to one leaf switch or vPC dual-home connects 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-home server is used for the sake of simplicity.

The spine switches are Cisco Nexus 9364C switches. These support, each, 64 ports of 100G. There are 48 x 100G downlinks connected to leaf switches and 16x100G uplinks 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).

The super-spine switches are modular Cisco Nexus 9516 switches fully loaded with N9K-X9732C-FX. There is a total of 512 x 100G ports with 4+1 FM redundancy. The number of super-spine switches is 16. On each super-spine switch, the 512 100G downlinks connect to each spine switch in a fully meshed manner. Each super-spine switch can support up to 512 spine switches. 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 up to 64 server pods for the cluster.

The characteristics of this design example (for example, 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

Two-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-home servers to each leaf	Maximum number of switches per server pod	Maximum number of switches per cluster
Leaf (Cisco Nexus 93240YC-FX2)	8	1	100	48	25	1.5	48	48	3072
Spine (Cisco Nexus 9364C: 64 ports of 100G)	16	1	100	48	100	3	N/A	8	512
Super spine (Cisco Nexus 9516 fully loaded with N9K-X9732C-FX, total: 512 x 100G ports, 4+1 FM redundancy)	N/A	N/A	N/A	512	100	N/A	N/A	N/A	16
Maximum number of single-home servers per server pod	2304								
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.
- There are 16 super spines in the cluster to interconnect the 64 server pods.
- The maximum number of single-home servers per cluster is 147,456, which is more than 10 times the number of servers supported in example 1.

Cisco provides a very rich portfolio of Cisco Nexus 9000 and Cisco Nexus 3000 fixed switches with different port densities and speed supports. Using Cisco Nexus 9000 as an example, the choices of switches for different layers are as follows (please note that the choices are not limited to the examples given below):

- Leaf switches: Cisco Nexus 9300-FX2, Cisco Nexus 9300-FX, Cisco Nexus 9300-EX series switches, etc.
- Spine switches: Cisco Nexus 9332C and 9364C fixed spine switches, Cisco Nexus 9500 Series modular switches with different line-card options, etc.
- Super-spine switches: Cisco Nexus 9500 Series modular switches with different line-card options, etc.

For more information, please refer to the data sheets for Cisco Nexus 9000 and Cisco Nexus 3000 switches.

Cisco MSDC design example3: Hyperscale fabric plane Clos design

In recent years, several very large web providers redefined data-center design by introducing a new MSDC design: hyperscale fabric. Some of the key characteristics of the hyperscale fabric are the following:[1]

- Hyperscale is about scalability rather than bigness.
- Network switches are layered and disaggregated.

The principal network building block is a server pod. The size of a server pod is limited to 48 server racks; each server rack has one Top-of-Rack switch (ToR). The 48 switches fully mesh connect 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 48 spine switches. The number of spine planes is four. The number of spine switches in each spine plane is up to 48. This is shown in Figure 5.

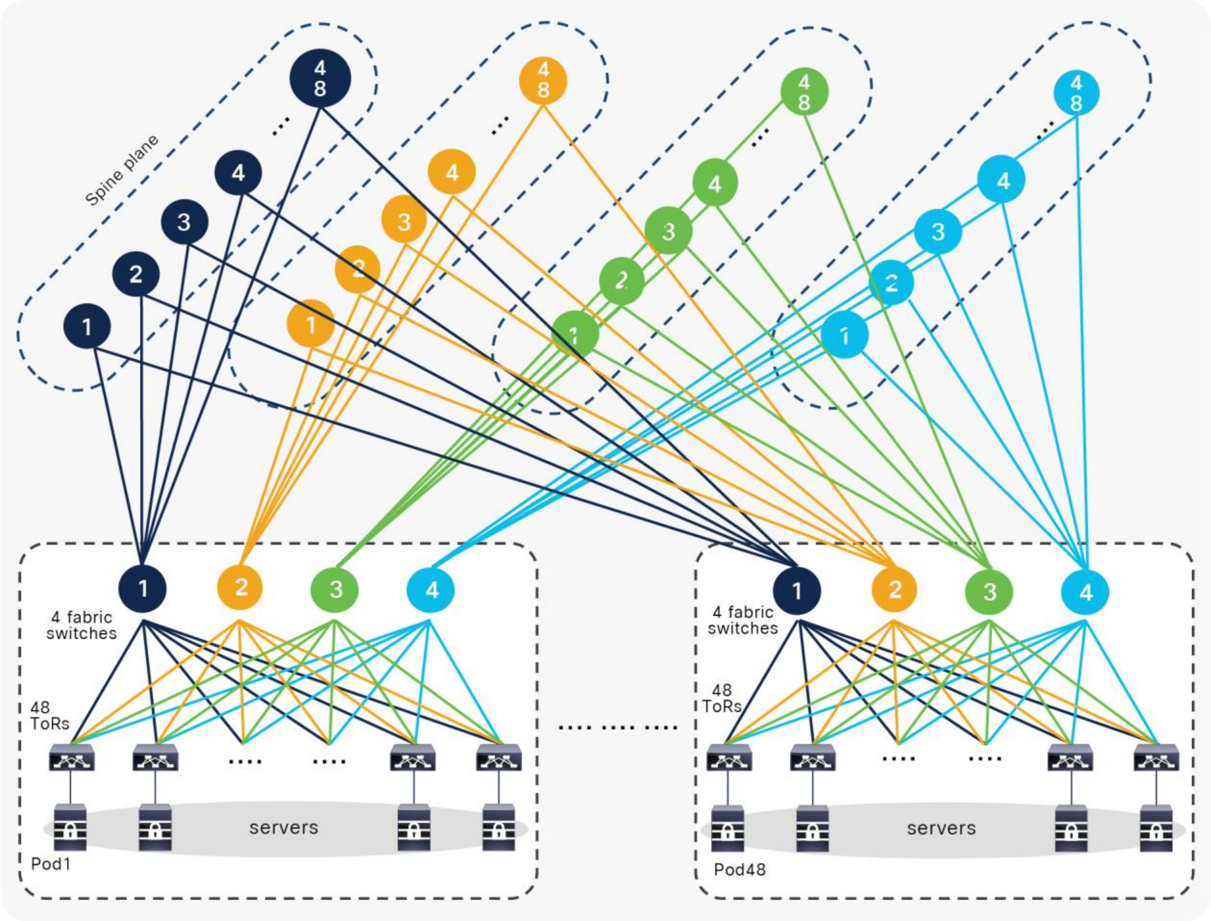

Figure 5.
Cisco MSDC design example 3

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 48 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; the oversubscription rate of each pod is easily controlled by the number of spine switches on the spine planes.
- 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 ensures that servers in each pod are redundantly connected to the entire network. When a server fails, it is taken offline and replaced.

The section below shows an example of a hyperscale fabric plane MSDC design using Cisco's switch.

In this example, the leaf switches use Cisco Nexus N34200YC-S; it has 48 x 10/25 Gbps fiber ports and 8 x 40/100 Gbps QSFP28 fiber ports. In this design, each leaf switch has 8 x 100 Gbps uplinks connecting to fabric layer switches, with two uplinks to each fabric switch. The number of fabric switches per server pod is four. All of the downlink 48 x 25G ports connect to servers; each server single-home connects to one leaf switch or vPC dual-home connects 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-home server is used for the sake of simplicity.

The fabric switches are Cisco Nexus3408-S switches with 8 100G LEM (a line-card Ethernet module). Each fabric switch supports a maximum of 128 ports of 100G. There are 80 x 100G downlinks connected to leaf switches. The number of leaf switches per server pod is 40. There are 48x100G uplinks connected to spine plane switches, with one uplink to each spine plane switch. The number of spine plane switches per spine plane is 48. The oversubscription ratio at the fabric switch is 1.67 (80x100G: 48x100G)

The spine-plane switches are Cisco Nexus3408-S switches with 8 100G LEM (a line-card Ethernet module). Each spine-plane switch supports a maximum of 128 ports of 100G. There are 112 x 100G 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**	**Oversubsc-ription ratio**	**Maximum number of single-home servers to each ToR**	**Number of switches per server pod**	**Maximum number of switches per hyperscale fabric**
ToR (Cisco Nexus 34200YC-S)	8	2	100	48	25	N/A	1.5	48	40	4480
Fabric switch (Cisco Nexus 3408-S with 128 ports of 100G)	48	1	100	80	100	N/A	1.666666667	N/A	4	448

Hyperscale fabric plane Clos design example										
Spine plane switch (Cisco Nexus 3408-S with 128 ports of 100G)	N/A	N/A	N/A	112	100	4	N/A	N/A	N/A	192
Maximum number of single-home servers per server pod	1920									
Maximum number of server pod per hyperscale fabric	112									
Maximum number of single-home servers per hyperscale fabric	215040									

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

- For each server pod, there are four fabric switches, 48 leaf switches, and 1920 single-home servers.
- The maximum number of server pods per hyperscale fabric is 112.
- There are four spine planes, with 48 spine switches in each spine plane to interconnect the 112 server pods.
- The maximum number of single-home servers per hyperscale fabric is 215,040, which is about 20 times the number of servers supported in example 1.

Cisco provides a very rich portfolio of Cisco Nexus 9000 and Cisco Nexus 3000 fixed switches with different port densities and speed supports. Using Cisco Nexus 9000 and 3000 switches as examples, 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-FX2, Cisco Nexus 9300-FX, Cisco Nexus 9300-EX, Cisco Nexus3400-S series switches, etc.
- Fabric switches: Cisco Nexus 9332C and 9364C fixed spine switches, Cisco Nexus3400-S series switches, etc.

- Spine-plane switches: Cisco Nexus 9332C and 9364C fixed spine switches, Cisco Nexus 3400-S series switches, etc.

For more information, please refer to the data sheets for Cisco Nexus 9000 and Cisco Nexus 3000 series switches.

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 "RFC7938 – Use of BGP for Routing in Large-Scale Data Centers".[3] In recent few years, there are multiple competing 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 in below:

- IS-IS Routing for Spine-Leaf Topology (LSR) https://tools.ietf.org/html/draft-ietf-lsr-isis-spine-leaf-ext
- Dynamic Flooding on Dense Graphs (LSR) https://tools.ietf.org/html/draft-ietf-lsr-dynamic-flooding (Note: Cisco implemented this feature based on the standard, please refer white paper "IS-IS Dynamic Flooding in Data Center Networks White Paper" [4])
- 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 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 Advertising IPv4 Network Layer Reachability Information with an IPv6 Next Hop (RFC5549) 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. EBGP single-hop sessions are established over direct point-to-point Layer 3 links between different tiers, and how to allocate EBGP ASNs.

EBGP ASN design examples for MSDC

Figure 6 shows an example of an EBGP ASN allocation scheme for an MSDC. 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 pair of ToR switches in each pod, assuming that the server is dual-home connected 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 deployment.

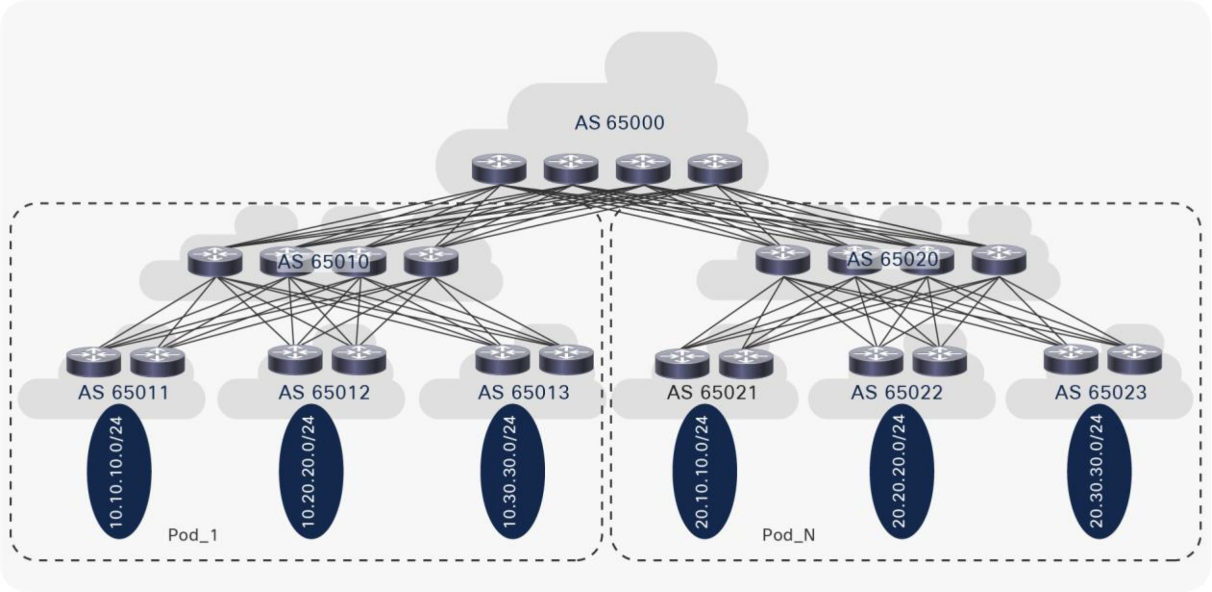

Figure 6.
EBGP ASN design example 1

Figure 7 shows another example of an EBGP ASN allocation scheme for an MSDC. 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.

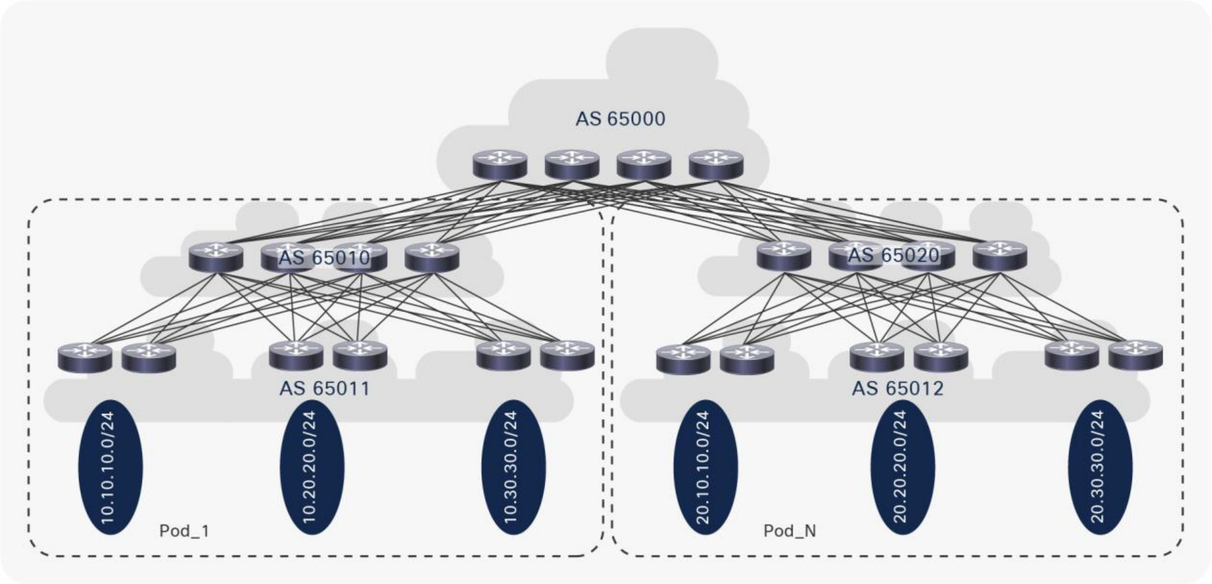

Figure 7.
EBGP ASN design example 2

ECMP in an MSDC

An Equal Cost Multipath Protocol (ECMP) is used between different tiers for load sharing in an MSDC design. Using Figure 8 as an example, for each ToR switch, there are eight uplinks connected to the four fabric switches, with two uplinks per fabric switch. The two links can be configured as Layer 3 point-to-point IP connections with EBGP neighbors or bundled as Layer 3 point-to-point port channel IP connections with EBGP neighbors.

Cisco Nexus 9000 and 3000 switches also support a resilient hashing feature. Resilient hashing maps flows to physical ports; in case a link fails, the flows assigned to the failed link are redistributed uniformly among the working links. Existing flows through the working links are not rehashed. Resilient hashing helps minimize flow rehashing caused by link failure, thus improving application performance.

The number of ECMPs at each tier in the example shown in Figure 8 is as follows:

- ToR: 8 ECMP uplinks or 4 ECMP port channel uplinks
- Fabric switch: 48 ECMP downlinks, 48 ECMP uplinks
- Spine plane switch: 48 ECMP downlinks

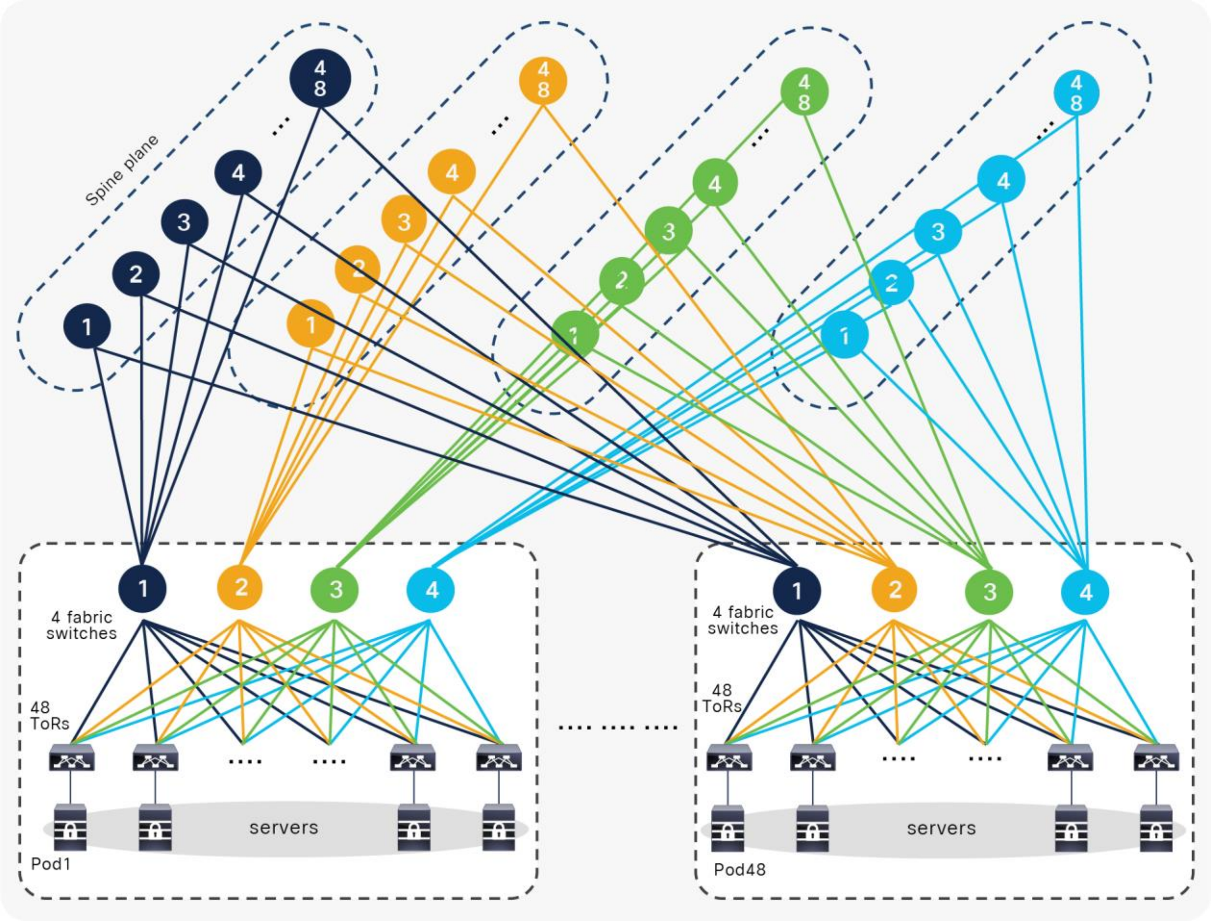

Figure 8.
ECMPs in Cisco MSDC design example 3

Prefix advertisement

The prefix advertisement design depends on application communication requirements. The general rule-of-thumb 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 LPM routes. Cisco Nexus 9000 and 3000 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 (longest path match) heavy-mode routing profiles, etc. For details, please refer to the Cisco Nexus switches NX-OS-verified scalability guide.

IPv4 and IPv6 dual stack in MSDC

The other trend in MSDC design is that MSDC fabrics are moving toward becoming IPv6-only. Right now, 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 future. Cisco Nexus 9000 and 3000 switches fully support dual-stack IPv4 and IPv6, and also support RFC5549 (Advertising IPv4 Network Layer Reachability Information with an IPv6 Next Hop).

Cisco MSDC automation

For MSDCs, because the number of devices in the fabric is very large, it can’t be configured and operated in a manual way. MSDC customers typically use software-based approaches to introduce more automation and more 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 and 3000 series switches support numerous capabilities to aid automation, for example:

- Shells and scripting: bash shell, guest shell, Python API, Power On Auto Provisioning (POAP), etc.
- Development and operations (DevOps) tools: Puppet, Chef, Ansible, SaltStack, 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 and 3000 series switches in MSDC environments:

- Power On Auto Provisioning (POAP): The POAP 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: Puppet, Chef, and Ansible 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 and 3000 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 and 3000 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 and 3000 series NX-OS programmability guides for the corresponding release.

Cisco 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 and 3000 switches support rich software and hardware capabilities to fulfill these requirements. For software features, Cisco Nexus 9000 and 3000 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 Streaming Statistics Export (SSX) and Flow Table Events (FTEs). The SSX feature reads statistics directly from a Cisco Nexus 9000 Cloud Scale ASIC and streams the statistics to a collector for analysis. For the FTE feature, 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 turnkey application “Network Insights Resources (NIR)” for event analytics, resource utilization and flow analytics, etc. The NIR application is integrated as a plugin into the Cisco ACI Application Policy Infrastructure Controller (APIC) and the Cisco DCNM GUI. For more information on NIR, please refer to the “Network Insights – Resources for Cisco ACI and Cisco DCNM Based Data Center Fabric White Paper.”[4]

Cisco Network Insights Advisor (NIA), a related application, can also be leveraged by MSDC customers. NIA monitors all the devices in a network and provides complete network lifecycle management for data center operations. It consists of the following main components:

- Advisories: This provides recommendations regarding Cisco software version upgrades, upgrade path and hardware refreshes, etc.
- Notices: This notifies customer regarding field notices, end-of-life and end-of-sale notices, etc.
- Issues: This notifies customers regarding bug reports, Product Security Incident Response Team (PSIRT) reports, configuration compliance recommendations, etc.
- TAC Assist: This can automatically collect logs and send them to Cisco’s Technical Assistance Center (TAC).

- Job Configuration: This can be scheduled as an automatic job or as an on-demand job to scan configurations, scrub bugs, etc.

For more information regarding NIA, please refer to the "Cisco Network Insights Advisor (NIA) Data Sheet."[5]

Cisco MSDC with RoCEv2

During the past few years, RDMA over Converged Ethernet (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/deep learning applications, etc. By adopting RDMA, applications have reported impressive performance improvement. Cisco Nexus 9000 and 3000 series switches fully support RoCEv2 with rich features; for example, Quality of Service (QoS), Priority Flow Control (PFC), PFC watchdog, Explicit Congestion Notification (ECN), etc. Cisco Nexus 9000 and 3000 switches have been widely deployed by hyperscale web providers in production MSDCs running RDMA applications.

For more information regarding RDMA testing and configurations in Cisco Nexus switches, please refer to the white paper "Benefits of Remote Direct Memory Access Over Routed Fabrics."[6] Benefits of Remote Direct Memory Access Over Routed Fabrics

Cisco 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 important. 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: first, to identify which switches need to be upgraded, then to download the software onto the switches to prestige the upgrade, and finally to perform 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 choses to upgrade the switches in the corresponding server pod in batches. For example, for the four fabric switches in the server pod, one or two fabric switches will be isolated using Cisco GIR (Graceful Insertion and Removal) feature, then upgrade the isolated switches without affecting application service. Similarly, for the leaf layer switches, the GIR feature is also used to isolate the corresponding leaf switches, then upgrade them. Typically, MSDC customers create custom maintenance mode profiles to ensure the corresponding switches stop receiving production traffic during upgrades. For more information on GIR, please refer to the Cisco Nexus 9000 and 3000 series switches Graceful Insertion and Removal configuration guides.

Conclusion

Cisco’s MSDC Layer 3 IP fabric architecture based on Cisco Nexus 9000 and 3000 series switches is designed to meet the requirements of MSDC networks: network simplicity, stability, visibility, easy troubleshooting, automation, easy life-cycle management, and related issues. In this white paper, we have discussed the Cisco MSDC Layer3 IP fabric design options, including its design tools, extensive automation and programmability features, unique telemetry features, rich visibility applications, RoCEv2 features, etc. Cisco’s MSDC solution enables scalability, simplified operation, 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] Introducing the data center fabric and the next-generation Facebook data center network:

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

[2] Project Altair: The Evolution of LinkedIn’s Data Center Network

https://engineering.linkedin.com/blog/2016/03/project-altair-the-evolution-of-linkedins-data-center-network

[3] RFC7938 – Use of BGP for Routing in Large-Scale Data Centers

https://tools.ietf.org/html/rfc7938

[4] IS-IS Dynamic Flooding in Data Center Networks White Paper

https://www.cisco.com/c/en/us/products/collateral/switches/nexus-9000-series-switches/white-paper-c11-743015.html

[5] Network Insights – Resources for Cisco ACI and Cisco DCNM Based Data Center Fabric White Paper

https://www.cisco.com/c/en/us/products/collateral/data-center-analytics/network-insights-data-center/white-paper-c11-742021.html

[6] Cisco Network Insights Advisor (NIA) Data Sheet

https://www.cisco.com/c/en/us/products/collateral/data-center-analytics/network-insights-advisor/datasheet-c78-742685.html

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

https://www.cisco.com/c/dam/en/us/products/collateral/switches/nexus-9000-series-switches/white-paper-c11-741091.pdf

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 C11-743245-00 01/20