Cisco Nexus 9800 Series Switches White Paper

Contents

With their new platform design, Cisco Nexus® 9800 Series Switches deliver a breakthrough in high-performance data-center switches in combination with either Cisco Silicon One™ Application Specific Integrated Circuits (ASICs), Cisco NX-OS (the Nexus operating system), or Cisco® Application Centric Infrastructure (Cisco ACI®) software. When they operate in ACI mode, Nexus 9800 series switches function as the fabric spine nodes to form the foundation of a transformational ACI architecture for a fully integrated and automated network fabric solution driven by an application-network profile. The Cisco Nexus 9800 Series modular switches expand the Cisco Nexus 9000 Series portfolio with a new chassis that supports extremely high port density 400 Gigabit Ethernet line cards.

This white paper focuses on the common hardware architecture of Cisco Nexus 9800 Series Switches and the packet-forwarding implementation in the classic NX-OS mode.

Cisco Nexus 9800 Series Switches utilize the second-generation Cisco Silicon One Q200 ASIC with increased performance to 12.8 Tbps in 7-nm process technology. The Q200 ASIC delivers high-scale routing and deep buffering, both of which typically require off-chip memories.

Cisco Nexus 9800 Series Switches are highly scalable, deep-buffered, on-chip High Bandwidth Memory (HBM), 400G-optimized data-center switches that range from 57 Tbps to 115.2 Tbps. The 8-slot Cisco Nexus 9808 Switch (Figure 1) is the first available platform in a family; it will be followed by 4-slot platforms. Each line card slot in the chassis can support line cards having 400GE, 100GE, or 10/25/50GE ports.

The Cisco Nexus 9800 Series Switch line cards and fabric modules are built with a power-efficient, high-performance, and high-capacity ASIC that supports fully shared on-die packet buffering. The ASIC provides these capabilities without compromising on features and power efficiency, which enables the Cisco Nexus 9800 Series Switches to be optimized for supporting high-bandwidth applications across data centers of varying sizes and scale.

Furthermore, the chassis architecture supports control-plane redundancy with dual supervisors, data-plane redundancy with up to eight fabric modules, fan-tray redundancy with four fan trays, and power-supply redundancy.

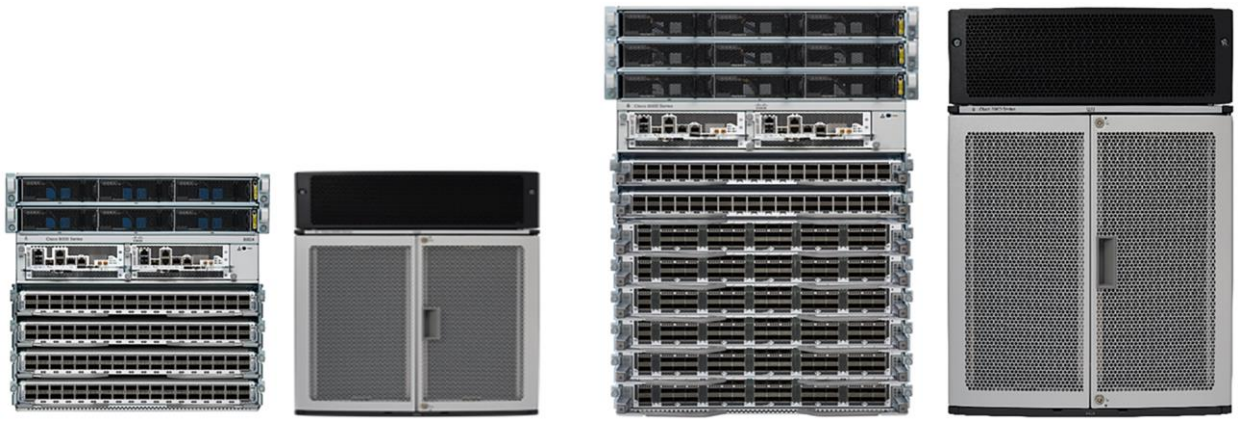

Figure 1.
Cisco Nexus 9808 Switch and Cisco Nexus 9804 Switch chassis

Table 1. Cisco Nexus 9804 Series Switch and Cisco Nexus 9808 Series Switch chassis and forwarding characteristics

Metric	Cisco Nexus 9804	Cisco Nexus 9808
Height	10 RU	16 RU
Supervisor slot	2	2
Fabric module slots	8 (N+1)	8 (N+1)
Fabric module ASICs	1 x Q200L	2 x Q200L
Line card slots	4	8
Maximum forwarding throughput per system (Tbps)	57.6 Tbps	115.2 Tbps
Air flow	Front to back (port side intake)	Front to back (port side intake)
Power shelves	2 HVAC/HVDC–6 (3 per tray) DC60–8 (4 per tray) DC100–8 (4 per tray)	3 HVAC/HVDC–9 (3 per tray) DC60–12 (4 per tray) DC100–12 (4 per tray)
Fan trays	4	4
Typical system power	4.1KW	8KW

The design of the Cisco Nexus 9800 Series chassis is a significant improvement on previous generation modular chassis design with better power distribution, connectors, fans, and thermal design that allow the chassis to scale up to higher Ethernet speed line cards and fabric modules in the future. These design principles allow the total system capacity to double, with next-generation line cards and fabric modules that support higher speed ports such as 800G at the same port density per slot as that of current generation line cards.

The chassis of the Cisco Nexus 9800 Series has an innovative midplane-free design (Figure 2). A midplane is commonly used in modular platforms to provide connectivity between line cards and fabric modules. Being an extra piece of hardware inside the switch chassis, it obstructs the cooling airflow. Hence, additional methods need to be applied to facilitate an airflow path – for example, cut-outs on the midplane or airflow redirection – which result in reduced cooling efficiency. With a precise alignment mechanism, Nexus 9800 Series switch line cards and fabric modules directly attach to each other with connecting pins. Line cards and fabric modules have orthogonal orientations in the chassis so that each fabric module is connected to all line cards and vice versa. Without a midplane blocking the airflow path, the chassis design delivers maximized cooling efficiency. It also allows a compact chassis design without the need for large cooling fans.

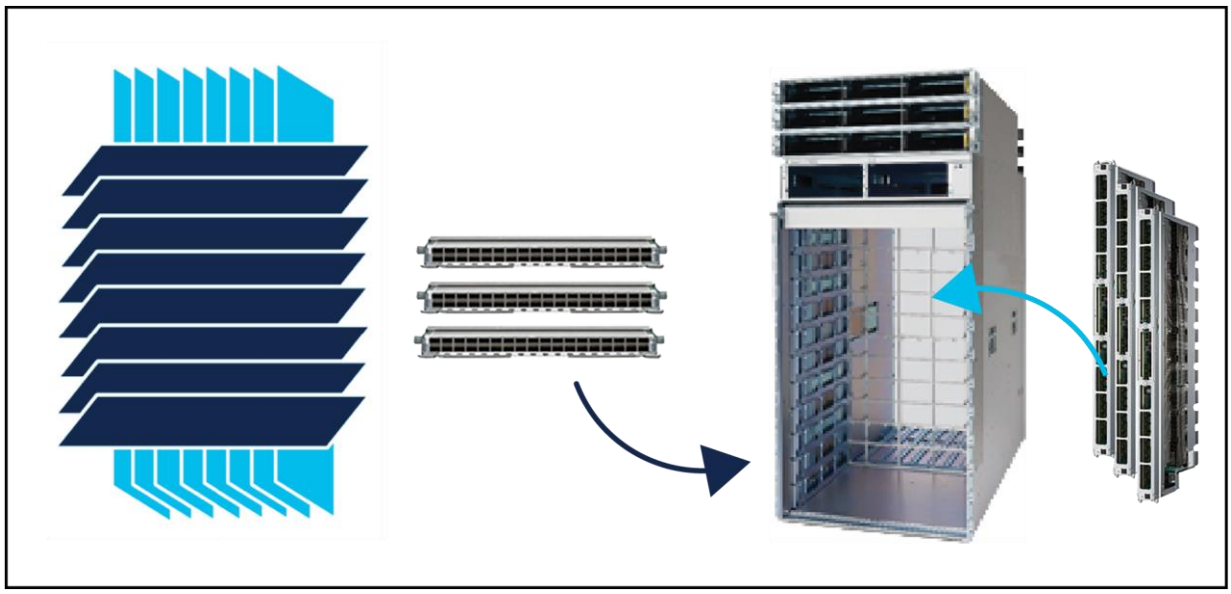

Figure 2.
Cisco Nexus 9800 Series midplane-free chassis

Midplane-free chassis design simplifies the switch platform deployment and hardware upgrade. In some cases where new components, such as new line cards or new fabric modules, are introduced, an upgrade to the midplane is required. This introduces complexity and more service disruption to the hardware upgrade process. The Cisco Nexus 9800 Series alleviates the need for midplane installation or upgrade. Another advantage of removing the midplane is significantly improved mean-time-to-repair. With a midplane, if you bend a pin on the midplane, the entire switch must be taken out of service and disassembled to replace that midplane. With Nexus 9800, the components that are damaged can be replaced without taking the other components of the chassis out of service.

The Cisco Nexus 9800 platform uses a folded Clos topology (often referred to as a fat-tree topology) internally to connect the fabric modules and the line cards. As shown in Figure 3, the ASICs on the fabric modules form the spine layer, and the ASICs on the line cards form the leaf layer.

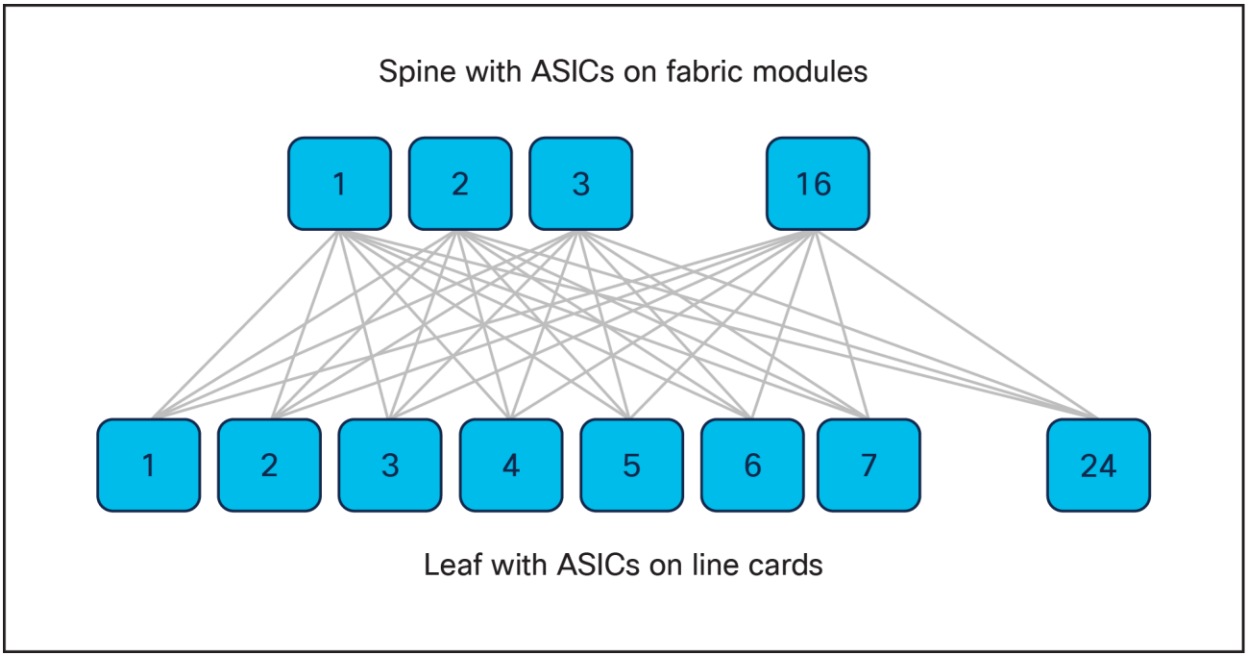

Figure 3.
Internal folded Clos architecture of Cisco Nexus 9800 platform switches

The Clos topology keeps the switch internal architecture simple and consistent with the overall data center network architecture. It eliminates the need for a switching fabric between line cards.

Cisco Nexus 9800 Series Switches supervisor

The Cisco Nexus 9800 Series supports redundant half-width supervisor engines that are responsible for control-plane functions (Figure 4). The switch software, Enhanced NX-OS, runs on the supervisor modules. The redundant supervisor modules take active and standby roles supporting stateful switchover in the event of supervisor module hardware failure.

The CPU complex of the Nexus 9800 supervisor is based on the 4-core 2.4 GHz Broadwell CPU. The default system memory size is 32 GB. There is a built-in 128 GB SSD (solid state disk) to provide additional onboard non-volatile storage. The combination of high-speed multicore CPU and large memory form the foundation of a fast and reliable control plane for the switch system. Control plane protocols will benefit from the ample computation horsepower and achieve fast initiation and instantaneous convergence after changes in the network state. Additionally, the expandable, large DRAM and multicore CPU provide sufficient computer power and resources to support cgroup-based Linux containers in which third-party applications can be installed and run in a well-contained environment. The on-board SSD provides extra storage for logs, image files, and third-party applications. Table 2 lists the Cisco Nexus 9800 supervisor specifications.

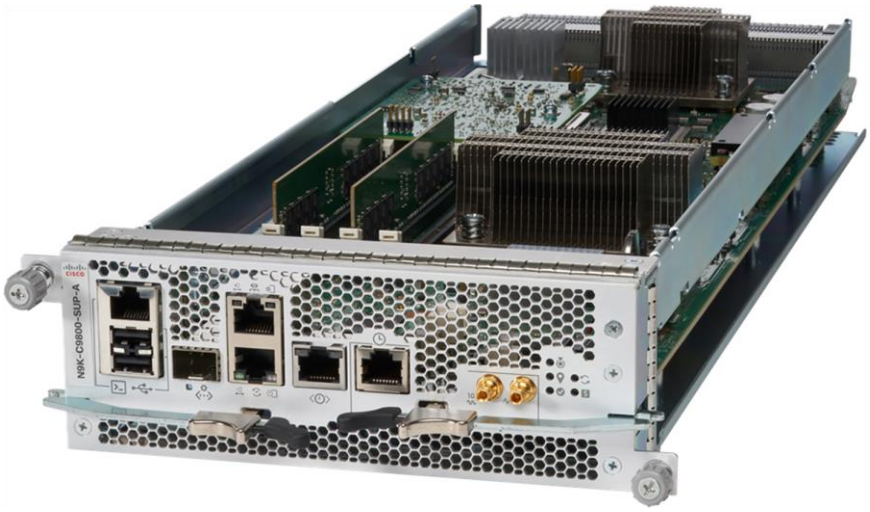

Figure 4.
Cisco Nexus 9800 supervisor

Table 2. Cisco Nexus 9800 supervisor specifications

Ports	• BMC and management Ethernet port • 2x USB 2.0 1A ports for flexibility
Processor	• 4-core 2.4 GHz Broadwell CPU
DRAM	• 32G
SSD	• 128G
Other features	• Timing class B • IEEE 1588 • SyncE • TOD • 10 MHZ / 1 PPS

Cisco Silicon One Q200/Q200L ASIC (Application Specific Integrated Circuit)

The Cisco Silicon One architecture ushers in a new era of networking, enabling one silicon architecture to address a broad market space, while simultaneously providing best-of-class devices.

At 12.8 Tbps, the Cisco Silicon One Q200 (Figure 5) builds on the groundbreaking technology of the Cisco Silicon One Q100 Processor. Q200 provides high-performance, power-efficient routing and switching utilizing 7nm silicon technology. Cisco Nexus N9K-X9836DM-A and N9K-X98900CD-A line cards are implemented with the Q200 ASIC whereas the N9K-C9804-FM-A and N9K-C9808-FM-A utilize Q200L.

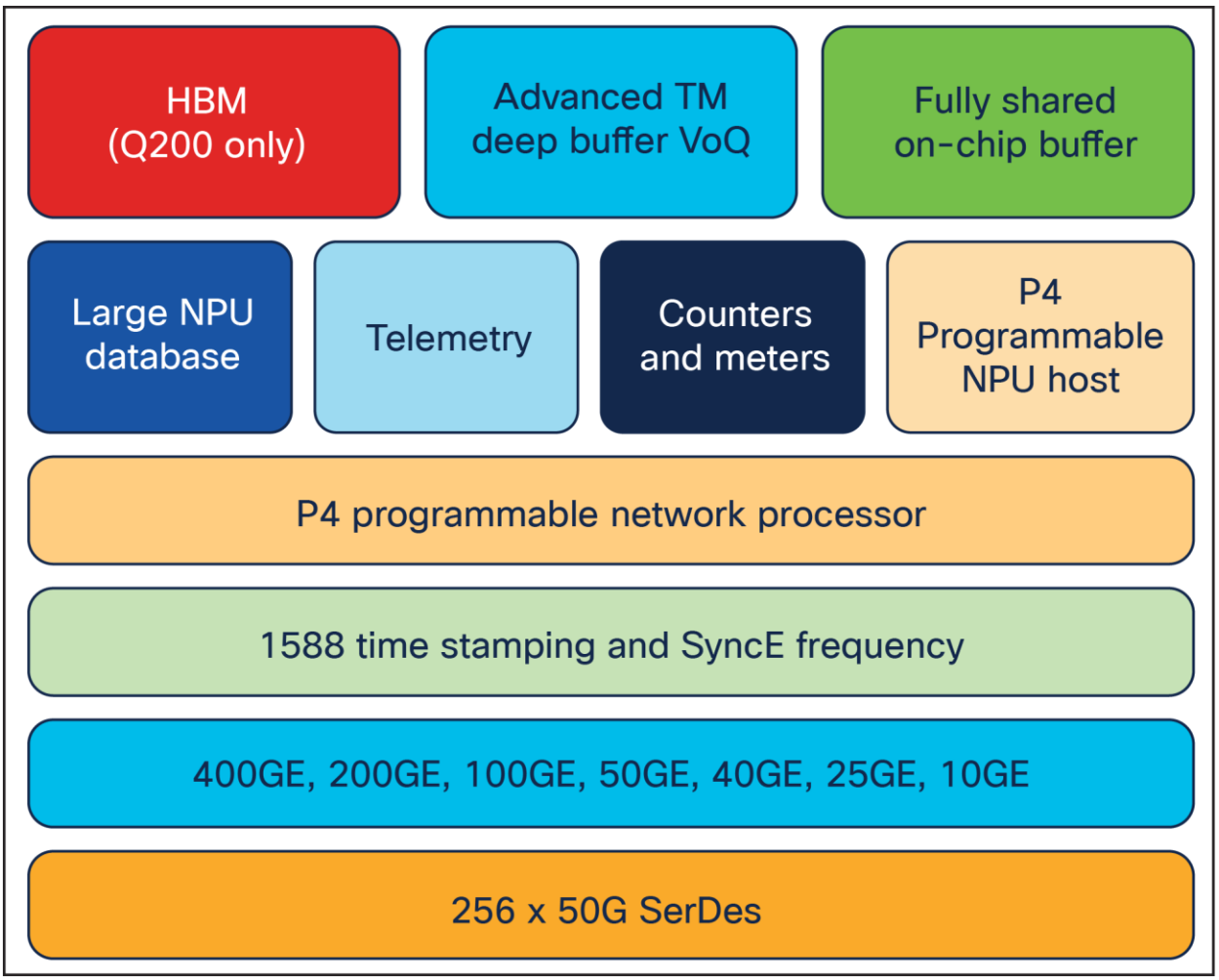

Figure 5.
Cisco Silicon One Q200/Q200L Processor block diagram

Both Cisco Silicon One Q200 and Q200L processors include six identical slices. In line card deployments such as Cisco Nexus N9K-X9836DM-A, half of the Q200 ASIC slices will be allocated for fabric module connectivity. The remaining three slices will face the front-panel ports. In case of the fabric modules, all six slices are deployed in fabric connectivity mode (Figure 6).

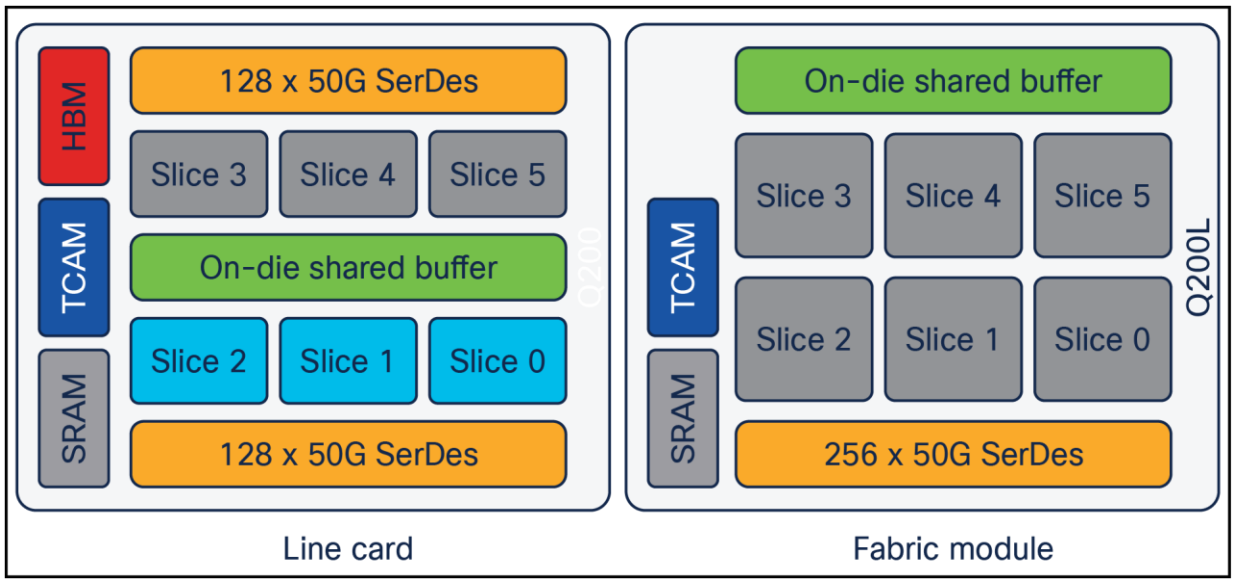

Figure 6.
Cisco Silicon One Q200/Q200L slices

Cisco Nexus 9800 line cards and fabric modules

The Cisco Nexus 9800 Series modular platform supports 400 GbE line cards. The line cards utilize multiple Cisco Silicon One forwarding ASICs to achieve high performance and bandwidth with line-rate forwarding. All ports support different breakout options for 100GbE, 40GbE, and 10GbE.

Table 3. Cisco Nexus 9800 line cards

Line card	N9K-X9836DM-A	N9K-X98900CD-A
Bandwidth	14.4 Tbps	9 Tbps
ASIC	3 x Q200	2 x Q200
Native 100 GE ports	NA	34
Native 400 GE ports	36	14
MACsec	All 36 ports	16 100GE ports
Performance	9.374 Bpps	5.858 Bpps
CPU	4-core 2.4 GHz Broadwell	4-core 2.4 GHz Broadwell
Memory	32GB DRAM	32GB DRMA
Packet buffer	324MB + 24GB	216MB + 16GB

Nexus N9K-X9836DM-A Line Card

The Cisco Nexus N9K-X9836DM-A is a Cisco Silicon One Q200-based, MACsec-capable, 36-port QSFP56-DD 400GbE line card that provides 14.4 Tbps of throughput with line-rate MACsec on all ports (Figure 7).

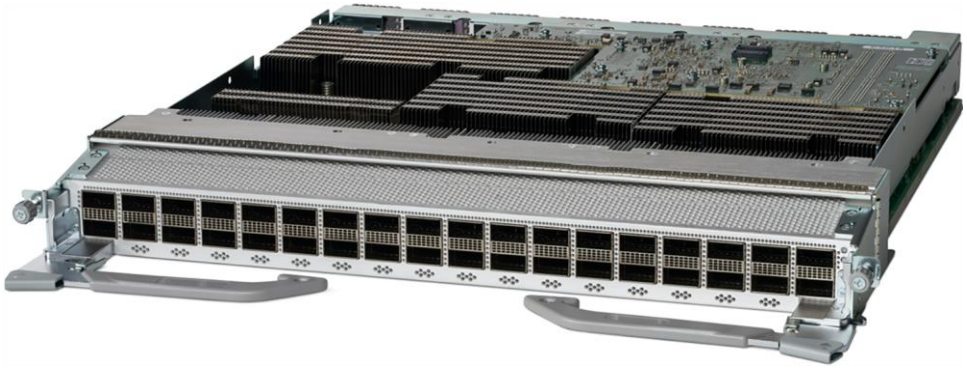

Figure 7.
Cisco Nexus N9K-X9836DM-A 36-port QSFP56-DD 400 GbE line card with MACsec

The Cisco Nexus N9K-X9836DM-A (Figure 8) line card utilizes three Cisco Silicon One Q200 ASICs, local CPUs, and DRAM. The first three slices of each individual Q200 are allocated for front-panel ports. The remaining three slices on each Q200 ASIC are connected to fabric modules. All front-panel ports support MACsec with Cisco Nexus N9K-X9836DM-A.

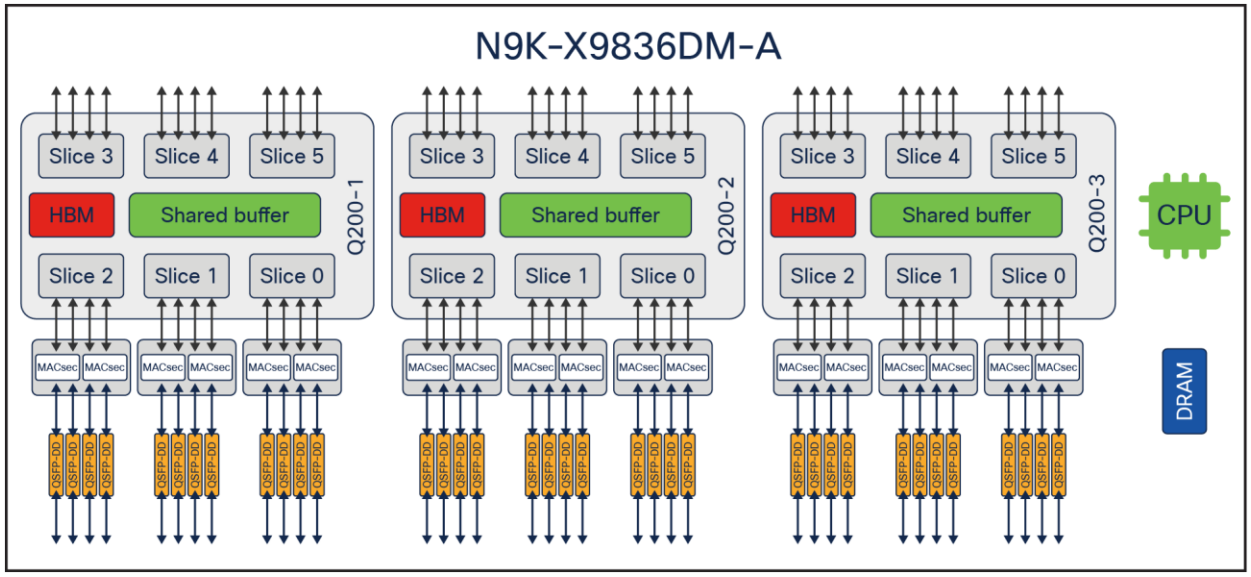

Figure 8.
Cisco Nexus N9K-X9836DM-A architecture

Cisco Nexus N9K-X9836DM-A fabric connectivity

With eight FM-A fabric modules, the chassis provides each N9K-X9836DM-A line card with 19.2-Tbps bandwidth, as shown in Figure 9. Note that each Cisco Silicon One Q200 processor on an N9K-X9836DM-A will have 16 x 50G connections to each FM-A fabric module for a total of 2400G per line card (Figure 10).

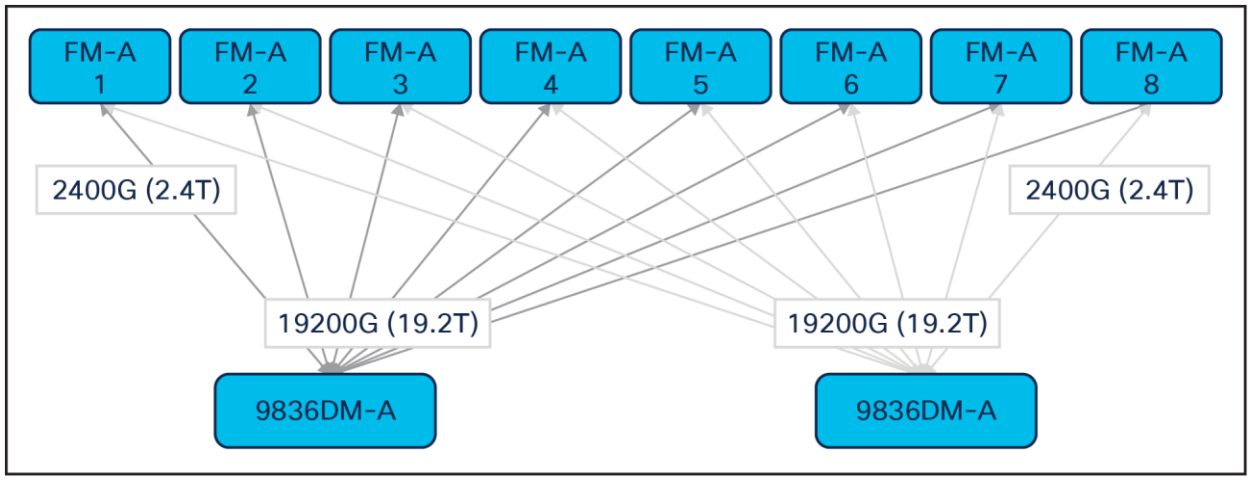

Figure 9.
Cisco Nexus N9K-X9836DM-A fabric connectivity

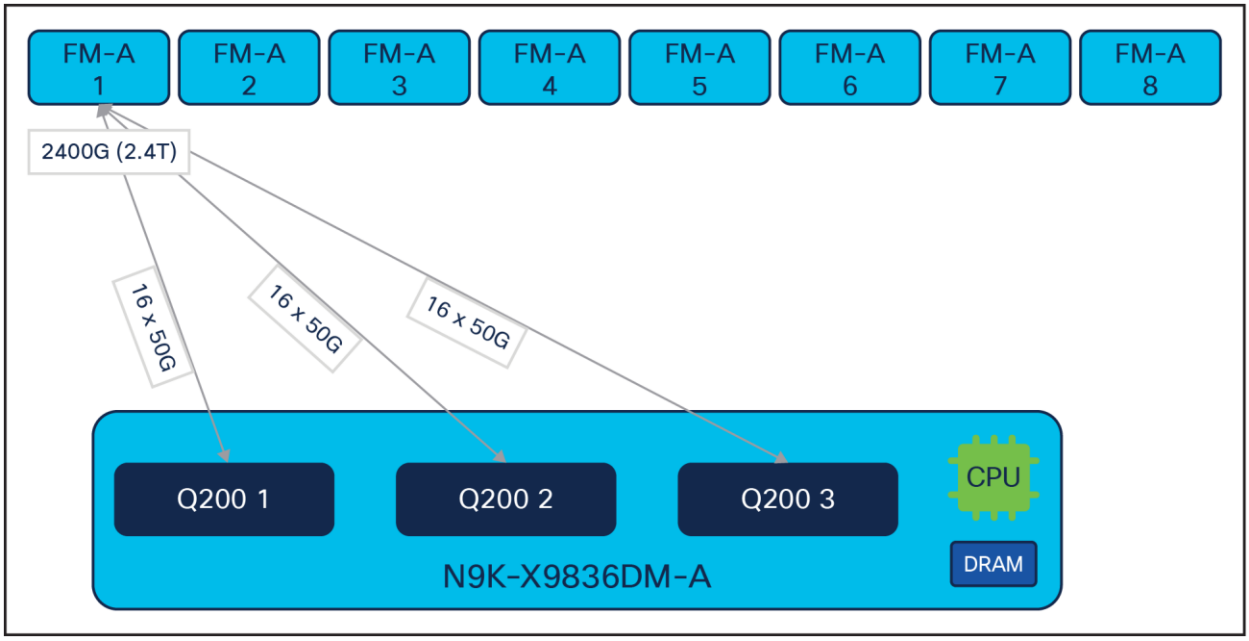

Figure 10.
Cisco Nexus N9K-X9836DM-A fabric connectivity from each Cisco Silicon One Q200 processer

Cisco Nexus N9K-X98900CD-A Line Card

Cisco Nexus N9K-X98900CD-A is a Cisco Silicon One Q200-based, 14-port QSFP56-DD 400GbE, 34-port QSFP-28 100GE line card that provides nine Tbps of throughput with line rate MACsec on sixteen QSFP-28 100GE ports (Figure 11).

Figure 11.
Cisco Nexus N9K-X98900CD-A Line Card

The Cisco Nexus N9K-X98900CD-A Line Card (Figure 12) utilizes two Cisco Silicon One Q200 ASICs, local CPUs, and DRAM. The first three slices of each individual Q200 are allocated for front-panel ports. The remaining three slices on each Q200 ASIC are connected to fabric modules. Sixteen QSFP-28 100GE ports support MACsec.

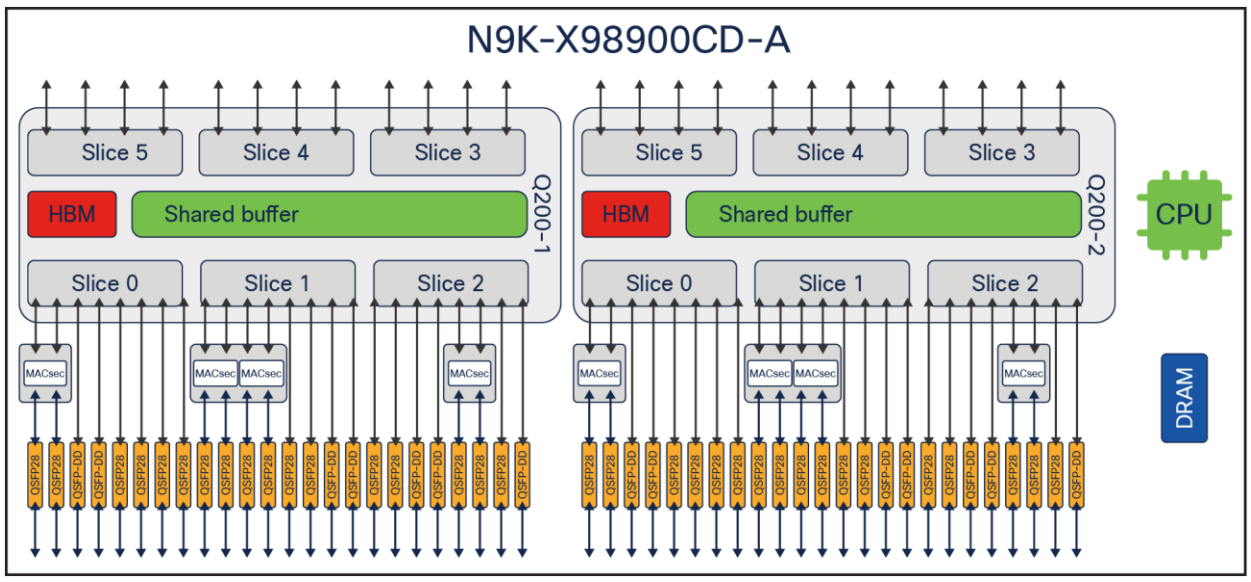

Figure 12.
Cisco Nexus N9K-X98900CD-A architecture

Cisco Nexus N9K-X98900CD-A port map

Figure 13 displays the port capabilities of the Cisco Nexus N9K-X98900CD-A.

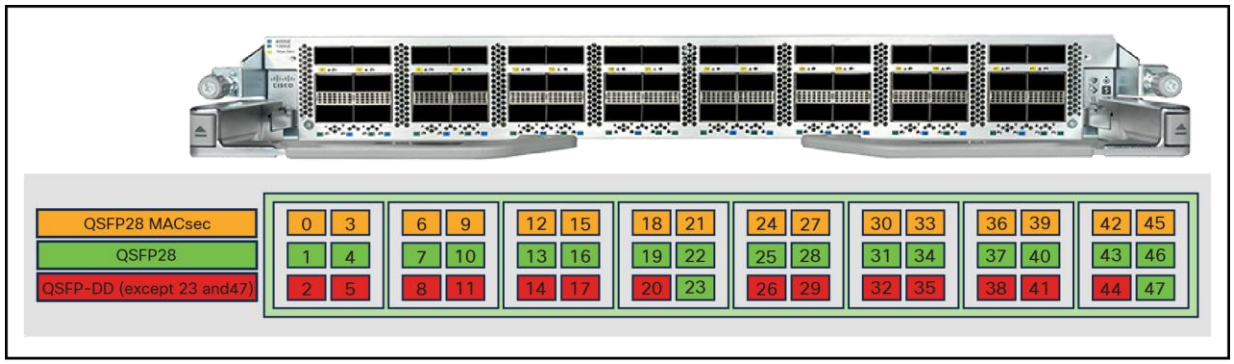

Figure 13.
Cisco Nexus N9K-X98900CD-A port map

Cisco Nexus N9K-X98900CD-A fabric connectivity

With eight FM-A fabric modules, the chassis provides each 98900CD-A line card with 12.8-Tbps bandwidth, as shown in Figure 14. Note that each Cisco Silicon One Q200 processor on N9K-X98900CD-A will have 16 x 50G connections to each FM-A fabric module for a total of 1600G per line card (Figure 15).

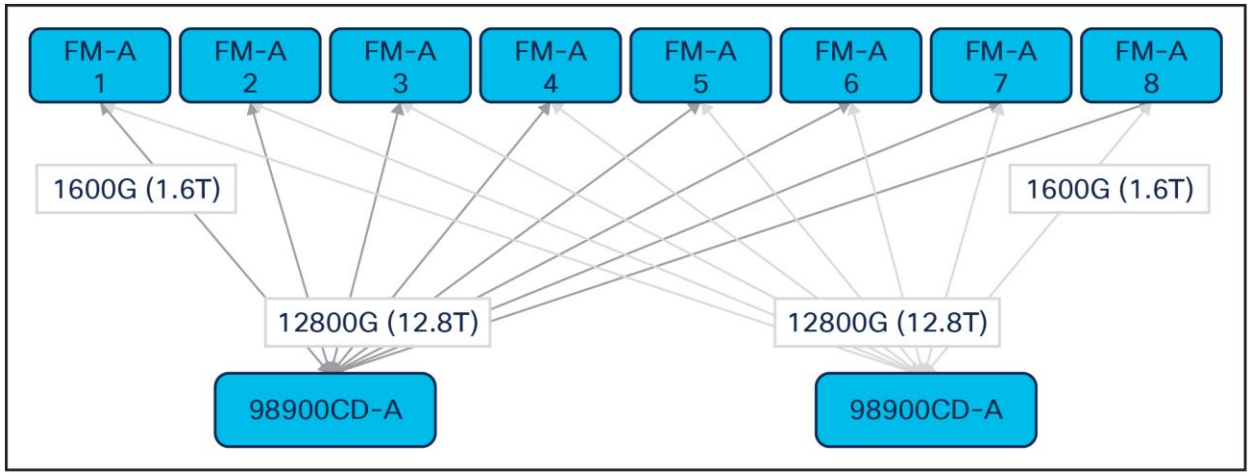

Figure 14.
Cisco Nexus N9K-X98900CD-A fabric connectivity

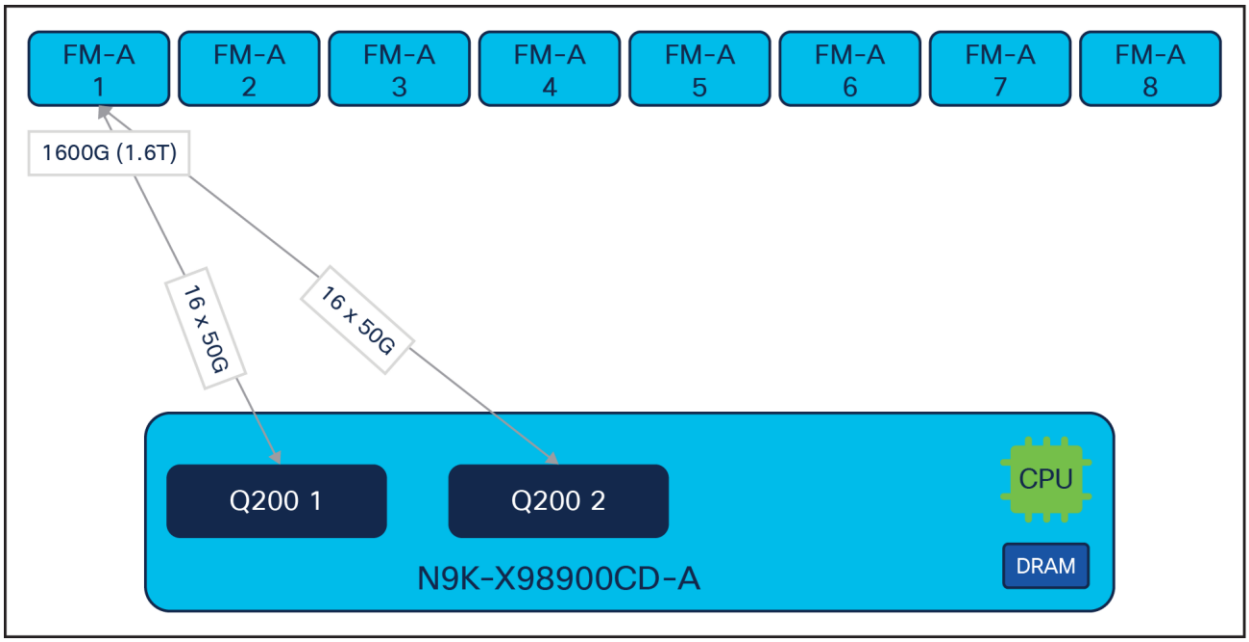

Figure 15.
Cisco Nexus N9K-X98900CD-A fabric connectivity from each Cisco Silicon One Q200 processer

Cisco Nexus 9800 Series fabric module

The Cisco Nexus 9800 Series switch fabric is powered by eight fabric cards that provide 7+1 -line rate redundancy. In addition, the fabric supports a separate operational model with 4+1 fabric-card redundancy to provide an entry-level option for systems with only the 48-port 100 GbE line card. This mode reduces cost and power for networks that want to take advantage of the latest platforms but are not yet ready to broadly deploy 400 GbE (Figure 16). Table 4 displays the datils for each fabric module.

Figure 16.
Cisco Nexus 9800 fabric module

Table 4. Cisco Nexus 9800 fabric modules

Fabric module	N9K-C9804-FM-A	N9K-C9808-FM-A
Number of FMs per chassis	8 (N+1) 400GE 5 (N+1) 100GE Only	8 (N+1) 5 (N+1) 100GE Only
Fabric module ASICs	1 x Q200L	2 x Q200L
Line card slots	4	8

Cisco Nexus 9800 Resiliency with 8 FM-As

Figure 17 displays the forwarding bandwidth provided with each fabric module. Cisco Nexus 9800 FM-A provides 7+1 fabric redundancy and graceful degradation on failure while delivering 14.4 Tbps per line card. The system continues to function even when only one plane is operational.

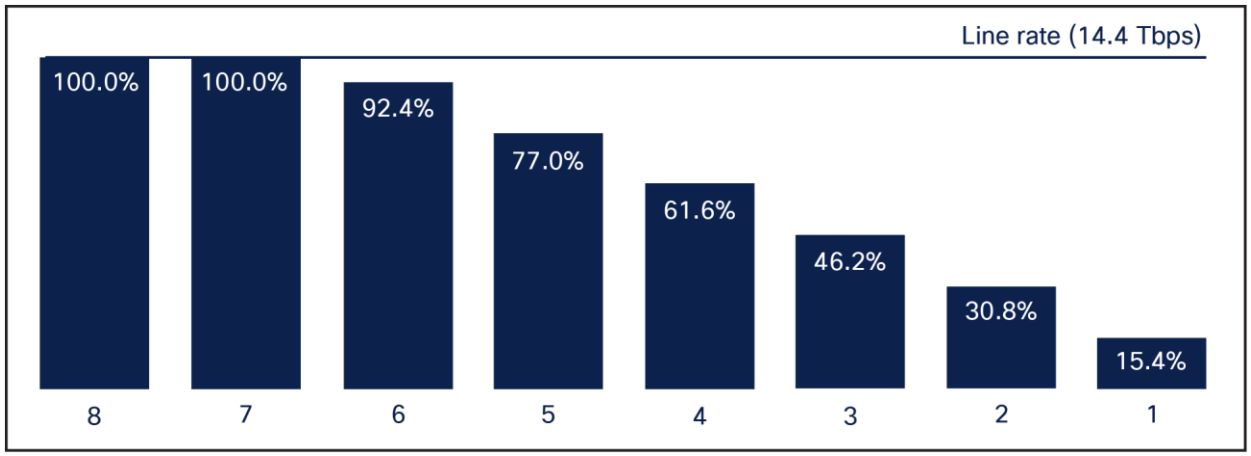

Figure 17.
Cisco Nexus 9800 resiliency with eight FM-As

Packet forwarding with Cisco Nexus 9800 line cards and fabric modules

This section describes the packet-forwarding process with the Cisco Nexus 9800 line cards and fabric modules.

The data-plane forwarding architecture of the Cisco Nexus 9800 platform switches includes the ingress pipeline on the ingress line card, fabric module forwarding, and the egress pipeline on the egress line card. The ingress and egress pipelines can be run on different line cards, the same line card, or even within a single ASIC on the line card if the ingress and egress ports are both on the same ASIC. As shown in Figure 16, the forwarding pipeline for the Cisco Silicon One Q200 ASICs consists of the ingress pipeline, Shared Memory Switch (SMS), and egress pipeline.

When a packet arrives to the ingress line card slice (Figure 17), it will be switched across the local SMS onto a slice that operates in fabric mode. Next, the packet is switched to the fabric devices that exist physically on the dedicated fabric module. On the fabric module, the packet is switched to the correct destination line card and Cisco Silicon One Q200 ASIC slice. Finally, the packet is received by the fabric-facing slice of the transmit side Q200. The packet is internally switched to the output Q200 slice. Output lookups/features are done in the transmit Q200 slice, and the packet is transmitted to the output interface. Figures 18 and 19 also display the intra-line card and intra-ASIC forwarding cases.

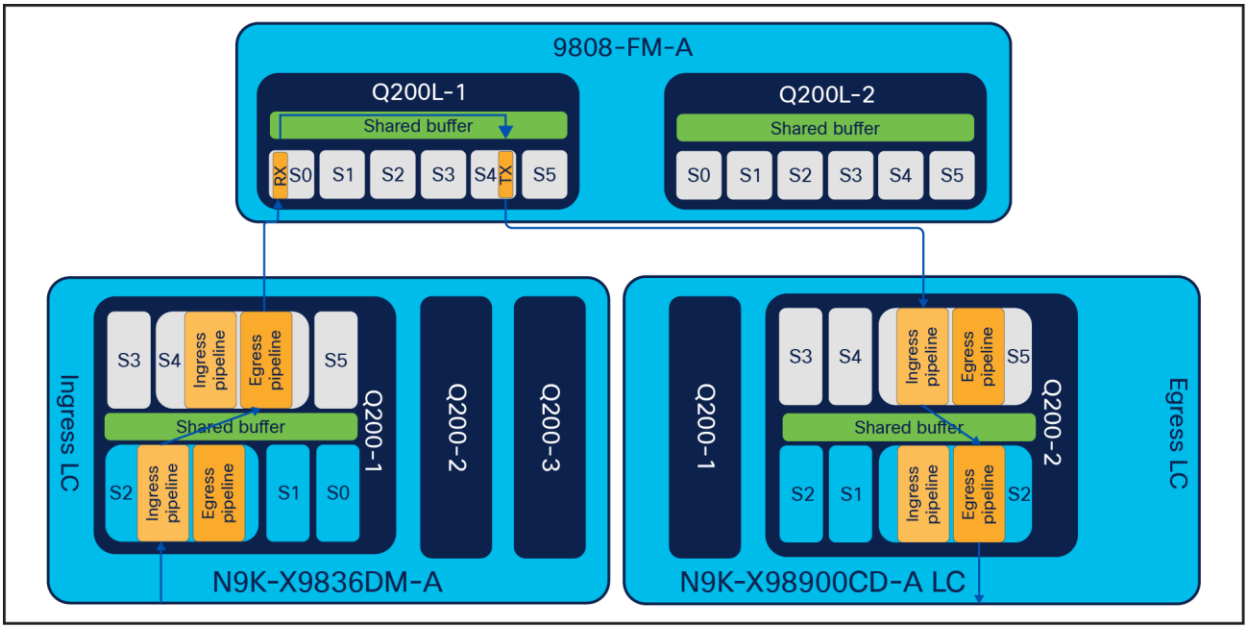

Figure 18.
Forwarding pipeline with new-generation line cards and fabric modules (inter-line cards)

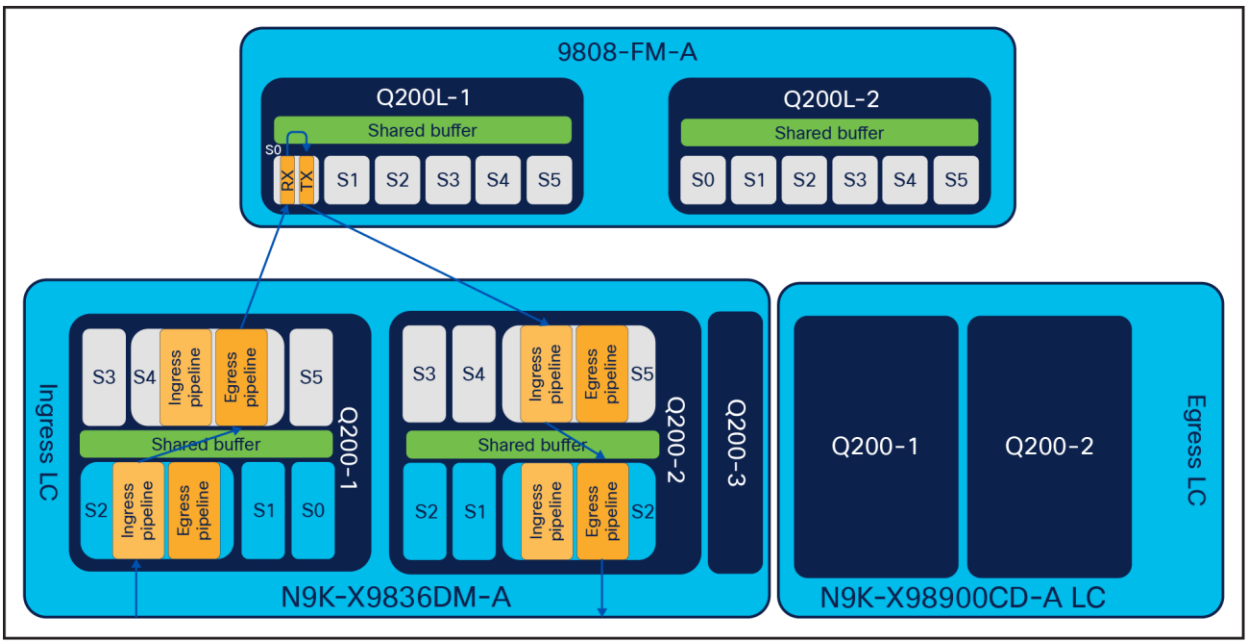

Figure 19.
Forwarding pipeline with new-generation line cards and fabric modules (intra-line cards)

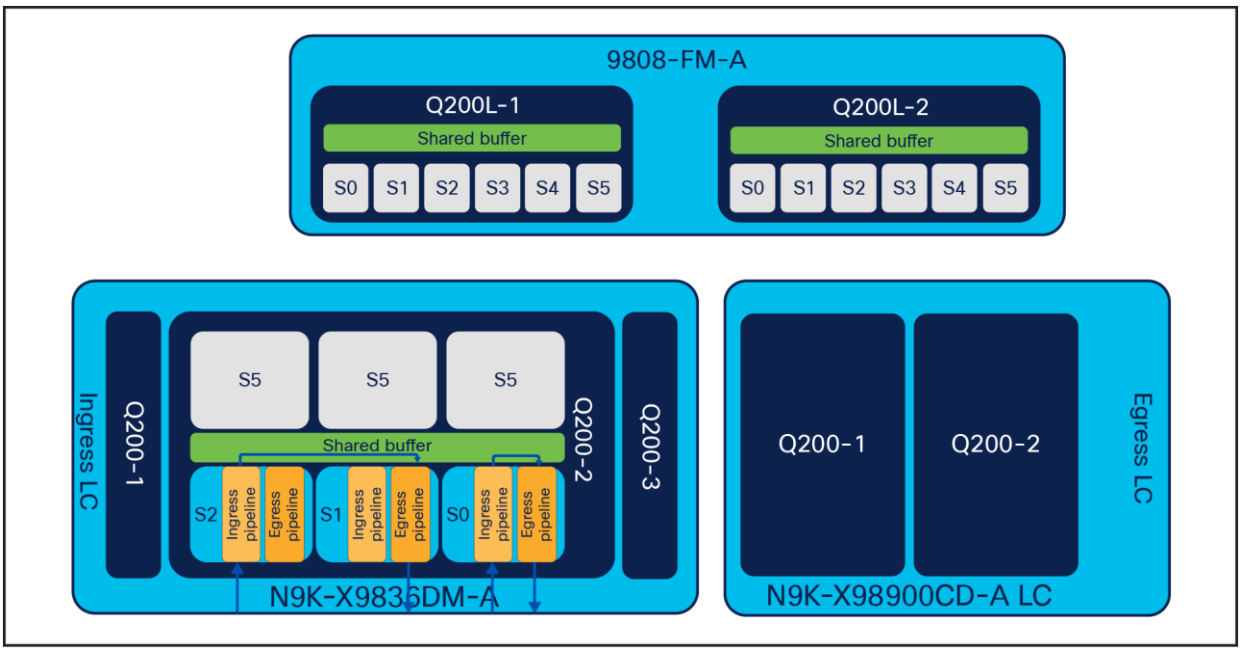

Figure 20.
Forwarding pipeline with new-generation line cards and fabric modules (intra-ASIC)

Multicast packet walk with Cisco Nexus 9800 line cards and fabric modules

In a modular system, the process of replication occurs as close to the point of data exit as possible. This approach aims to prevent the creation of unnecessary duplicate copies of the packet. Figure 20 illustrates the three-stage replication process using a Cisco Nexus 9800 Series switch.

In the first stage (Stage 0), unlike in standalone configurations, replication does not take place within the ingress ASIC. Instead, the process begins at the next stage.

Moving to the second stage (Stage 1), a multicast frame is directed to the fabric module. Here, the fabric ASIC responsible for receiving the frame generates multiple copies of the packet. Each copy is intended for a destination line card that has subscribed to the multicast traffic.

Finally, at the last stages (Stage 2 and 3), the egress ASIC comes into play. This stage involves two rounds of replication. In the first round, each slice with an active member receives a copy of the packet. The second round of replication caters to each (sub)interface present on the same slice that requires a copy of the packet.

In essence, this modular replication strategy in the Cisco Nexus 9800 system optimizes the replication process to minimize unnecessary duplication and efficiently deliver multicast packets to the intended recipients.

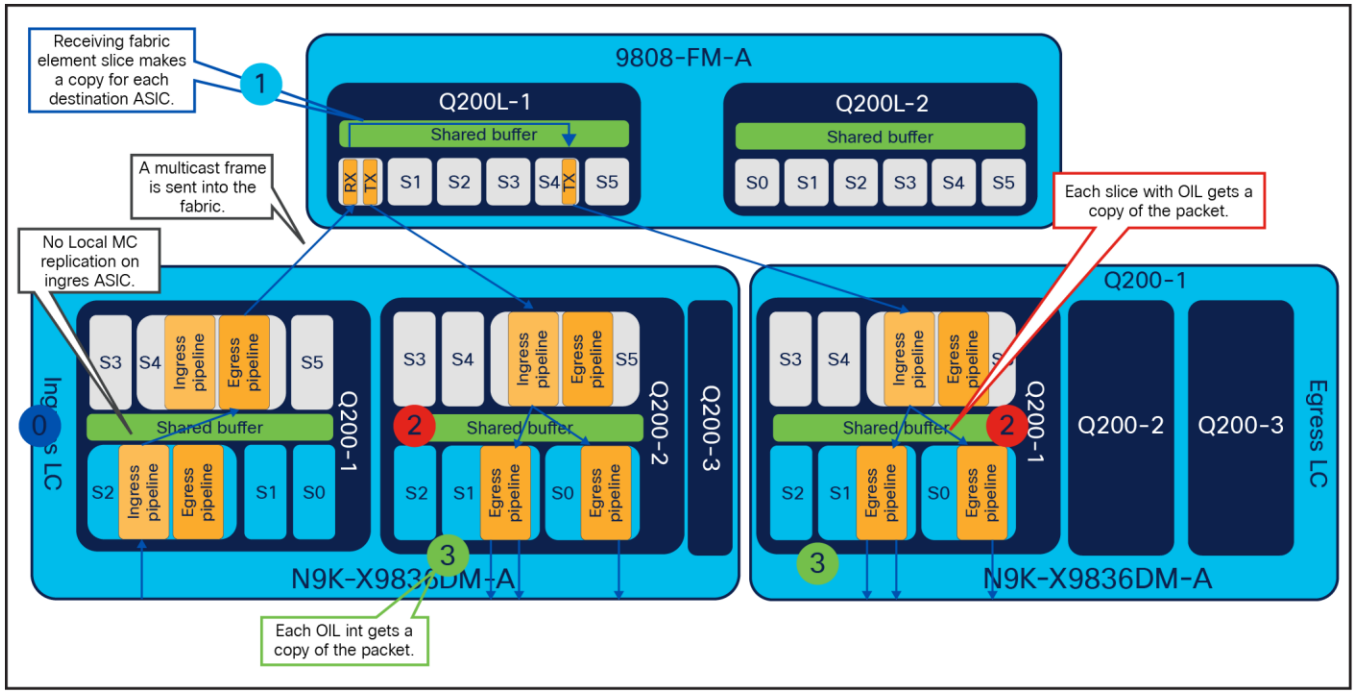

Figure 21.
Cisco Nexus 9800 multicast replication

Conclusion

With the addition of Cisco Nexus 9800 Series Switches, the Nexus 9000 family now offers a full breadth of 400G options that can meet the data-center and hybrid-cloud needs of enterprises and service providers of all sizes. Cisco Nexus 9800 Series Switches deliver density in form factors needed to operate efficiently in space-constrained facilities such as edge data centers with significant advancements in power and cooling efficiency, and are ideal for more sustainable data-center and hybrid-cloud operations.

For more information

https://www.cisco.com/c/en/us/solutions/data-center/high-capacity-400g-data-center-networking/index.html#~what's-new

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 C11-3051510-01 08/23