White Paper

Cisco Nexus 9500 Cloud-Scale Line Cards and Fabric Modules

Contents

Introduction

Starting in 2016, the data center switching industry will begin the shift to new capacity and capabilities with the introduction of 25, 50, and 100 Gigabit Ethernet connectivity. This new Ethernet connectivity supplements the previous 10 and 40 Gigabit Ethernet standards, with the similar cost points and power efficiency, and represents a roughly 250 percent increase in capacity.

Cisco is releasing a number of new products in the Cisco Nexus® 9000 Series Switches product line to enable our customers to build higher-performance and more cost-effective data center networks. Using Cisco® cloud-scale intelligent application-specific integrated circuits (ASICs), the new Cisco Nexus 9000 Series products also include network innovations to address the new challenges of supporting cloud-scale data centers, converged and hyperconverged infrastructure, and virtualized and containerized applications.

A Cisco Nexus 9500 platform switch equipped with the new line cards and fabric modules can operate in both NX-OS mode, based on Cisco NX-OS Software, and ACI mode, based on Cisco Application Centric Infrastructure (Cisco ACI™). This flexibility enables customers to deploy Cisco Nexus 9500 platform switches in the mode that best fits their current operational model while keeping the option open to migrate to the other mode without the need for additional hardware investment or changes.

This document discusses the hardware architecture of the new generation of line cards and fabric modules and operations in NX-OS mode. This architecture is an extension to the overall hardware architecture of the Cisco Nexus 9500 platform that is published on Cisco.com and can be downloaded from the following link: http://www.cisco.com/c/en/us/products/collateral/switches/nexus-9000-series-switches/white-paper-c11-729987.html.

Overview: Next-Generation Cisco Nexus 9500 Line Cards and Fabric Modules

The Cisco Nexus 9500 platform was launched in November 2013. It set new industry records at the time for 1, 10, and 40 Gigabit Ethernet port density, performance, power efficiency, and cost effectiveness. Now the new generation Cisco Nexus 9500 line cards and fabric modules offer even more with the introduction of 25, 50, and 100 Gigabit Ethernet speeds, the increased forwarding scalability, and a set of network innovations that help organizations build cloud-scale data center networks with outstanding visibility and security.

At its first introduction, the new generation Cisco Nexus 9500 line cards and fabric modules include the products listed in Table 1.

Table 1. Cisco Nexus 9500 Line Cards and Fabric Modules

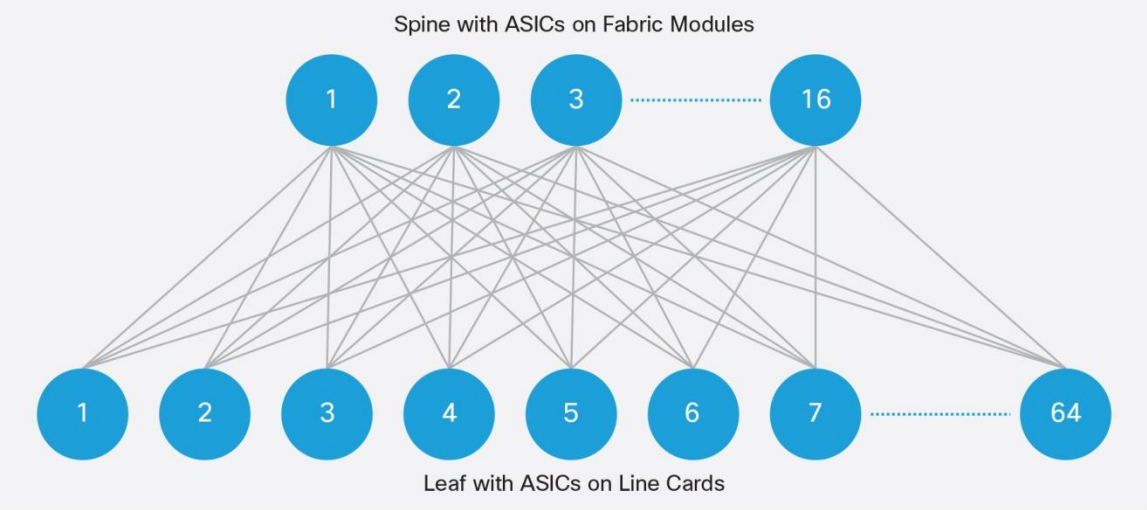
Product	Description
N9K-X9732C-EX	New line card providing 32 x 100-Gbps ports, or up to 512 x 100-Gbps ports per switch
N9K-C9504-FM-E	New fabric module for the Cisco Nexus 9504 Switch chassis
N9K-C9508-FM-E	New fabric module for the Cisco Nexus 9508 Switch chassis
N9K-C9516-FM-E	New fabric module for the Cisco Nexus 9516 Switch chassis*

* Support for Cisco Nexus 9508 and 9504 chassis comes first. Please check the latest release notes for support for the Cisco Nexus 9516 chassis.

The new Cisco Nexus 9500 line cards and fabric modules are fully supported in all existing Cisco Nexus 9500 platform switch chassis types, including Cisco Nexus 9504, 9508, and 9516 chassis. They do not require major chassis replacement or any changes or upgrades on the chassis common components, including the supervisors, system controllers, chassis fan trays, power supply modules, etc. This approach helps protect the investments of customers who want to maintain their current investment in the Cisco Nexus 9500 platform switches while adopting 25, 50, and 100 Gigabit Ethernet technologies.

The Cisco Nexus 9500 platform continues to use a folded Clos topology (often referred to as a fat-tree topology) internally to connect the new fabric modules and the line cards. As shown in Figure 1, the ASICs on the fabric modules form the spine layer, and the ASICs on the line cards form the leaf layer. With the largest Cisco Nexus 9500 switch platform, the 16-slot Cisco Nexus 9516, this internal Clos topology can grow to up to 16 spine switches and 64 leaf switches: four ASICs per fabric module on the 16-slot chassis, and four ASICs per line card.

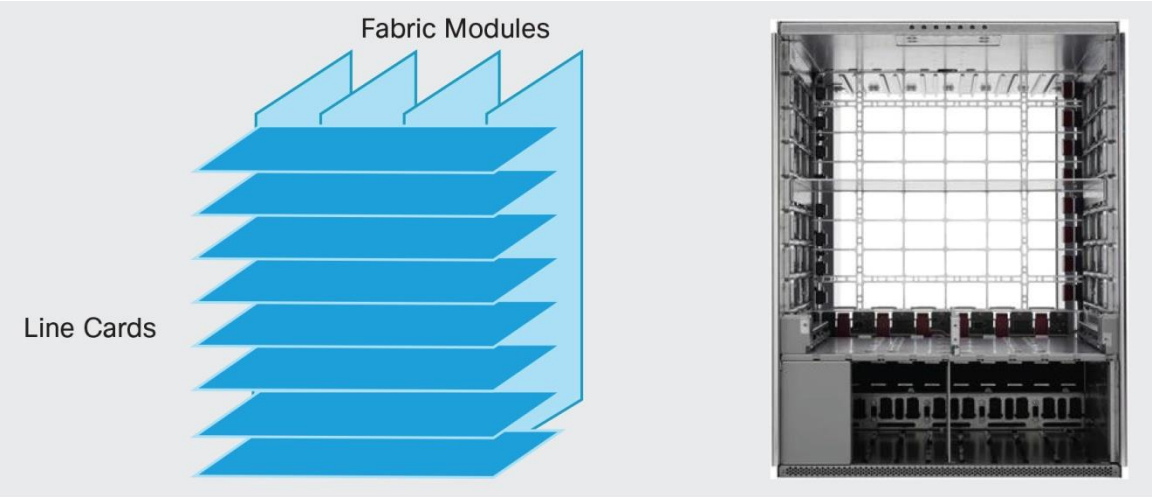
Figure 1. Internal Folded Clos Architecture of Cisco Nexus 9500 Platform Switches



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. Unlike switching fabric architecture that requires complicated and inefficient virtual output queue (VoQ) buffer management to avoid head-of-line blocking, the architecture for the Cisco Nexus 9500 platform is implemented with simple and intelligent egress queue management.

The Cisco Nexus 9500 fabric modules and line cards are physically interconnected through direct attachment with connecting pins. Line cards are inserted horizontally and fabric modules are inserted vertically, giving line cards and fabric modules orthogonal orientations in the chassis so that each fabric module is connected to all line cards, and all line cards are connected to all fabric modules. This direct attachment alleviates the need for a switch chassis midplane. Figure 2 depicts the orthogonal interconnection of line cards and fabric modules and the midplane-free chassis of a Cisco Nexus 9500 platform switch.

Figure 2. Cisco Nexus 9500 Line Cards and Fabric Modules Interconnection



The midplane-free chassis design offers several advantages. It allows the most direct and efficient cooling airflow. It increases chassis reliability by eliminating one moving component: the chassis midplane. It also makes upgrading to new-generation line cards and fabric modules easier, without the need to upgrade the midplane for a higher interconnection speed or the potential need to replace the chassis. This design enables the Cisco Nexus 9500 platform chassis to support all generations of line cards and fabric modules, providing a long lifespan for the switch chassis and excellent investment protection for the switch chassis and the common components.

New-Generation Cisco Data Center Switching ASICs

The new line cards and fabric modules for the Cisco Nexus 9500 platform switches are built with Cisco's cloud-scale technology ASICs. By taking advantage of a newer generation of semiconductor device fabrication, 16FF+, these ASICs are a full technology generation ahead of merchant silicon. Newer fabrication techniques enable higher transistor density and lower power consumption: features that are essential to build ASICs with more bandwidth, more ports, larger forwarding tables, larger buffers, and opportunities to implement new and more advanced capabilities.

Cisco's cloud-scale ASICs introduce 25, 50, and 100 Gigabit Ethernet (GE) speeds to data center networks at an optimal cost point and with increased performance. Table 2 shows the port capacities of the new ASICs that are used in the new Cisco Nexus 9500 fabric modules and line cards. The Application Spine Engine 2 (ASE2) ASIC is used in the fabric modules, and the leaf-and-spine engine (LSE) ASIC is used in the line cards.

Table 2. Cisco New ASIC Port Capacity

ASIC	10 GE Ports	25 GE Ports	40 GE Ports	100 GE Ports
ASE2	144	144	64	36
LSE	80	72	20	18

In addition to the need for cost-effective 25, 50, and 100 Gigabit Ethernet support, a large number of other changes taking place in the data center need to be addressed. Virtual machines, which have become common elements in most data centers, are being joined and in some cases replaced by Linux containers. The distributed storage functions built for big data are rapidly being used in other environments. Automation and dynamic resource allocation are resulting in a far more dynamic environment that complicates both security and day-two operations. These and other new requirements motivated Cisco to focus on delivering more capabilities and innovations in the next generation of switching.

Cloud-Scale Forwarding

The shifts in application development and the associated growth in the use of Linux containers and microservers are affecting many aspects of data center design, including the scaling requirements. As servers are disaggregated into their component services—for example, as each process or thread becomes an endpoint—the result will be an increase in the number of addressable endpoints by an order of magnitude or more. When aggregated across even a small number of racks, the increase in network scaling requirements will be substantially greater than the increase required by virtualization. The exponential increase in the number of endpoints per rack resulting from the use of containers and the overall increase in the total number of endpoints in the data center both contribute to the increased scaling requirements. Cisco is responding to these requirements by using some of the additional transistor capacity offered by next-generation switch ASICs for increased route and end-host scale.

Cisco's cloud-scale ASICs use a flexible forwarding table (FFT), which allows forwarding table resources to be shared among different forwarding entities. The forwarding table can be carved for various forwarding lookup operations such as lookups for Layer 2 (MAC addresses), IPv4/IPv6 host, IPv4/IPv6 longest-prefix match (LPM), next-hop adjacency information, Multiprotocol Label Switching (MPLS) labels, multicast entries, and reverse-path forwarding (RPF). The flexibility to program different resources as needed in the shared memory empowers customers to deploy next-generation products for a wide range of data center applications. In addition to the shared forwarding table, a 16,000-entry overflow ternary content-addressable memory (TCAM) table is available for Layer 2 and Layer 3 lookups.

Table 3 shows the FFT sizes of the ASE2 and LSE ASICs. Multiple forwarding entries can be referenced using a group of entries in the FFT. When this is performed with an efficient Trie algorithm, the FFT can accommodate about 1 million routes. Forwarding Table Size

Table 3. New-Generation ASIC Flexible Forwarding Table Sizes

ASIC	Flexible Forwarding Table Size
ASE2	352,000 entries (100 bits per entry)
LSE	544,000 entries (104 bits per entry)

Intelligent Buffering

Although cost-effective bandwidth and port density will always be critical requirements in data center design, these are not the only requirements associated with the mixing of various traffic types such as distributed storage and distributed application interprocess communication. Algorithm improvements plus the availability of more transistors for buffers are allowing Cisco to deploy intelligent buffering that provides both comparatively larger buffers than the last generation of merchant silicon switch-on-a-chip (SOC) designs and multiple enhanced buffer and queuing management features:

- Approximate fair discards (AFD)
- Elephant traps (ETRAP)
- Dynamic packet prioritization (DPP)
- Dynamic buffer management
- Eight classes of per-priority Pause
- Flowlet load balancing

The intelligent buffering functions, such as AFD, ETRAP, and DPP, add flow awareness to the active queue management. The switches can differentiate flows based on the data transfer sizes and treat them differently. Small flows can be prioritized over larger flows through the use of an express lane within the egress queue if they are more sensitive to packet drops or queue latency. The early discard mechanism can be applied to large flows first, with the discards applied to a flow proportional to the difference between the flow's data arrival rate and the system's computed fair rate. This approach achieves fairness among flows based on their data rate. The overall results of intelligent buffer management are higher application performance and better support for applications with microbursts/incast flows mixed with larger flows along the same network path.

Telemetry and Visibility

The shift to more efficient cloud-based provisioning requires new diagnostic and operational characteristics for the data center network. The operations team's knowledge of where a server was and what it was doing started to alter with early virtualization technologies and will change entirely as the use of Linux containers continues to expand. The need to understand far more about the state of the infrastructure and the state of the applications running on the infrastructure requires more telemetry information than has traditionally been available.

The new Cisco ASICs support a number of new sources of analytic information, such as an enhanced flow table, buffer monitoring, and expanded counters to complement the diagnostic functions that exist in the current generation of switches:

- Embedded logic analyzers
- 128,000-entry flexible (or flex) counters
- Atomic counters
- Dropped-flow capture
- 32 Cisco Switched Port Analyzer (SPAN) sessions
- SPAN on drop and SPAN on user-defined filter (UDF)
- Encapsulated Remote SPAN (ERSPAN) time stamp and termination
- IEEE 1588 latency measurements

Full flow visibility for every packet has not been possible in any data center switch for the past decade. The cost of providing both the required bandwidth and the table scalability made features such as full-flow monitoring too expensive. Cisco has shifted the model with its latest generation of ASIC. This new ASIC can provide full flow information. It also supports the collection of almost five times as much flow telemetry information as the standard NetFlow Version 9, with capacity and scale at reduced cost.

High-Performance VXLAN Routing

Virtual Extensible LAN (VXLAN) is becoming the new industry-standard overlay technology for building more scalable and reliable data center fabrics that support application workload mobility and provide network agility that matches the needs of applications. To move business-critical applications onto the VXLAN overlay fabric, high performance for both Layer 2 bridging and Layer 3 routing in the VXLAN overlay network is essential.

Most merchant-silicon ASICs are not built to support VXLAN routing natively. They have to sacrifice forwarding performance to provide VXLAN functions.

Cisco's new-generation ASCIs have built-in VXLAN routing capability. They provide uncompromised forwarding bandwidth and latency performance for VXLAN routed traffic. And they provide this support in both NX-OS mode and ACI mode.

Network Capabilities

Changing requirements for the network include more than increased capacity and better operational characteristics. Virtualization, containers, storage, and multitenancy also require basic improvements in the forwarding and security functions. More sophisticated tunneling and forwarding capabilities, which used to be limited to expensive backbone routers, have been made possible in high-speed data center networks through continued ASIC innovation. With the next generation of Cisco ASIC, features such as segment routing, group-based policy (GBP) for security, network service header (NSH), and full-featured VXLAN overlays are available, at greater scale and in more efficient devices:

- Single-pass VXLAN tunnel gateway
- GBP VXLAN
- VXLAN routing
- Bidirectional (Bidir) Protocol-Independent Multicast (PIM)
- NSH
- Push 5+2 (Fast Reroute [FRR]) MPLS labels
- Security group tag (SGT) and endpoint group (EPG) mapping
- Unified ports

Although many of these features are available at lower scale in merchant-silicon switches, Cisco takes advantage of the feature scale supported by the new ASIC designs to offer an additional set of features that may benefit more traditional data center customers, including Fibre Channel and unified ports and interoperation with Cisco TrustSec® SGTs.

Note: Cisco's cloud-scale ASICs are designed and built with the features described here, but the software support for some of the features was not available at the time of this writing. To verify the supported features in a given software release, refer to the corresponding software release note.

New Cisco Nexus 9500 Fabric Modules

The Cisco Nexus 9500 platform switches use four new fabric modules to provide non-blocking internal bandwidth between line-card slots. The new fabric modules are built with the ASE2 ASIC. The number of line card slots in the Cisco Nexus 9500 chassis determines the number of ASE2 ASICs used to build the fabric module. Table 4 summarizes the number of ASE2 ASICs and the number of 100 Gigabit Ethernet ports for each of the new fabric modules, and the total number of 100 Gigabit Ethernet (100 Gbps) fabric module ports in a Cisco Nexus 9500 platform switch.

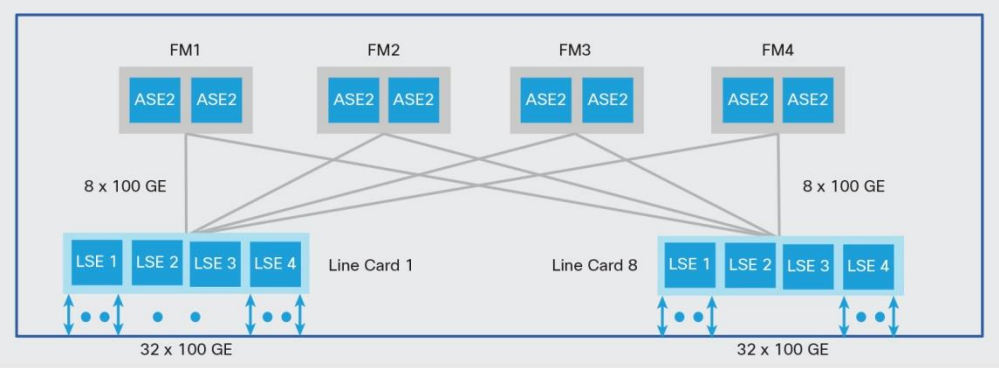
Table 4. Cisco Nexus 9500 Platform Switches Internal 100-Gbps Port Capacity in New Fabric Modules

Switch Model	Number of ASE2 ASICs per Fabric Module	Number of 100-Gbps Ports per Fabric Module	Number of 100-Gbps Ports per Chassis
Cisco Nexus 9504	1	32	128
Cisco Nexus 9508	2	64	256
Cisco Nexus 9516	4	128	512

The fabric module in the Cisco Nexus 9500 platform switches performs the following important functions in the modular chassis architecture:

- The module provides high-speed non-blocking data-forwarding connectivity for the line cards. All links on network forwarding engines are active data paths. Each fabric module can provide up to eight 100-Gbps links to every line-card slot. A Cisco Nexus 9500 platform chassis deployed with four fabric modules can provide 32 x 100-Gbps fabric paths to each line-card slot. This is equivalent to 3.2 terabits per second (Tbps) of bandwidth per slot per direction (6.4 Tbps bidirectional). Figure 3 shows the internal connectivity of a Cisco Nexus 9508 Switch as an example.
- The module can perform distributed unicast forwarding lookup.
- The module can perform distributed multicast lookup and packet replication to send copies of multicast packets to receiving egress ASICs on the line cards.

Figure 3. Cisco Nexus 9508 Switch Internal Connectivity



New-Generation Cisco Nexus 9500 Line Cards

All next-generation Cisco Nexus 9500 line cards consist of one or more LSE ASICs, depending on the required port types and density. The forwarding lookup operations can be partitioned between line cards and fabric modules.

Cisco Nexus 9500 line cards have a built-in dual-core CPU. This CPU is used to offload or speed up some control-plane tasks, such as actions to program the hardware table resources, collect and send line-card counters and statistics, and offload Bidirectional Forwarding Detection (BFD) protocol handling from the supervisors. These capabilities significantly improve system control-plane performance.

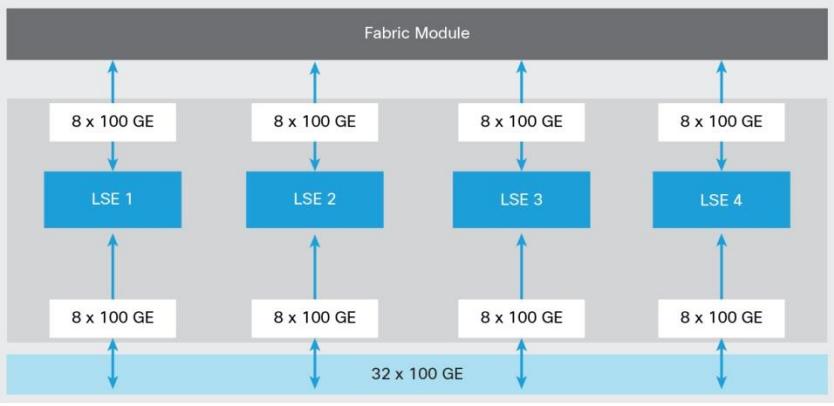
The Nexus N9K-X9732C-EX Line Card

The new Cisco Nexus N9K-X9732C-EX line card provides 32 x 100 Gigabit Ethernet Quad Small Form-Factor Pluggable 28 (QSFP28) ports. Figure 4 shows the front of the line card.

Figure 4. Front of N9K-X9732C-EX Line Card

The N9K-X9732C-EX line card is built with the Cisco LSE ASIC as its forwarding engine. Each LSE ASIC has 18 x 100 Gigabit Ethernet ports, but only 16 ports are used on the N9K-X9732C-EX line card to help guarantee full line-rate performance for all packet sizes with the folded Clos architecture of four fabric modules. Among the 16 active ports on each LSE, 8 ports are used as front-panel ports, and the other 8 ports are used internally to connect to the fabric modules. Therefore, four LSE ASICs are used to build the line card. Figure 5 illustrates the line card's internal architecture.

Figure 5. N9K-X9732C-EX Line-Card Architecture



This line card uses a PHY-less design. This design reduces data transport latency on the port by 100 nanoseconds (ns), decreases port power consumption, and improves reliability because fewer active components are used.

In addition to 100 Gigabit Ethernet, the front-panel ports can operate at 50, 40, 25, 10, and 1 Gigabit Ethernet speeds. When a port is populated with a QSFP28 transceiver, it can operate as a single 100 Gigabit Ethernet port, or it can break out to two 50 Gigabit Ethernet ports or four 25 Gigabit Ethernet ports. When a port has a QSFP+ transceiver plugged in, it can run as a single 40 Gigabit Ethernet port, or it can break out to four 10 Gigabit Ethernet ports. With an appropriate QSFP-to-SFP adaptor (QSA), a port can also operate at 10 and 1 Gigabit Ethernet and 100 Megabit Ethernet speeds.

Hierarchical Forwarding Scalability with New Line Cards and Fabric Modules

The forwarding lookup on the Cisco next-generation ASICs use a shared hash table memory known as the flexible forwarding table, or FFT, to store Layer 2 and Layer 3 forwarding information.

Because the new line cards and fabric modules for the Cisco Nexus 9500 platform are both built with the new-generation ASICs, they both have FFTs that can be used to program forwarding tables. As with the original network forwarding engine (NFE)–based line cards and fabric modules, a Cisco Nexus 9500 platform switch equipped with the new-generation line cards and fabric modules can perform hierarchical forwarding lookup: that is, it can use the FFTs on both the line cards and fabric modules to increase the system-wide forwarding scalability. For example, the FFT on the line cards can be used to store the Layer 2 MAC address table and Layer 3 host table, and the FFT on the fabric modules can be used for Layer 3 LPM routes. Or the FFT on both line cards and the fabric modules can be used to program Layer 3 host routes and LPM routes, with IPv4 and IPv6 entries partitioned between them. Also, both line cards and fabric modules can have multicast tables and take part in distributed multicast lookups and packet replication.

The flexibility to partition entries between line cards and fabric modules allows the Cisco Nexus 9500 platform switches to optimize table resource use on the line cards and fabric modules and to increase the Layer 2 and Layer 3 forwarding scalability of the system. It also enables Cisco Nexus 9500 platform switches to be deployed in data centers at a broad range of scales with a variety of application types.

Table 5 shows one example of FFT partitioning between line cards and fabric modules: using fabric modules for IPv6 entries and using line cards for IPv4 and MAC address entries. This is only one example of how the FFT can be partitioned. Other approaches are possible: for instance, the number of IP host routes or LPM routes can be increased by fully using the table resources on both the line cards and fabric modules for the selected forwarding route types.

Table 5. FFT Partition Example

Entry Type	Fabric Module (ASE2)	Line Card (LSE)	Cisco Nexus 9500 Platform
Layer 2 MAC addresses	Not used	16,000	16,000
IPv4 host routes	Not used	1 million*	1 million*
IPv4 LPM	Not used	1 million*	1 million*
IPv6 LPM/64	320,000	Not used	320,000
IPv6 host routes	4000	Not used	4000

* Shared entries

Packet Forwarding with New Line Cards and Fabric Modules

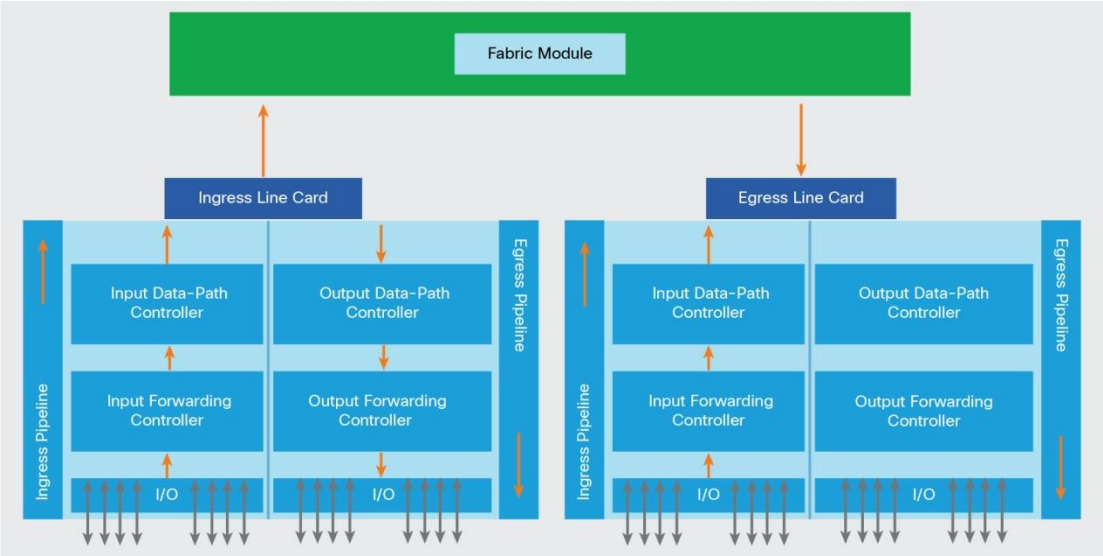
This section describes the packet-forwarding process with the new line cards and fabric modules.

Packet Forwarding Pipeline

The data-plane forwarding architecture of the Cisco Nexus 9500 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 the same line card, or even on the same ASIC if the ingress and egress ports are on the same ASIC.

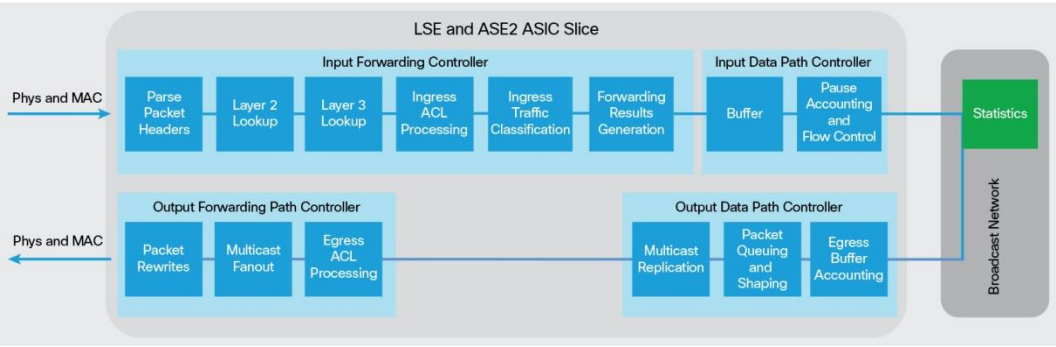
As shown in Figure 6, the forwarding pipeline for the new generation ASIC consists of the input forwarding controller, input data-path controller, egress data-path controller, and egress forwarding controller.

Figure 6. Forwarding Pipeline with New-Generation Line Cards and Fabric Modules



Each of the pipeline elements shown in Figure 6 can consist of multiple forwarding processing steps. Figure 7 illustrates the detailed steps in the new generation ASIC pipelines when the switch is operating in NX-OS mode.

Figure 7. Detailed Forwarding Steps in the ASIC Pipeline



Input Forwarding Controller

The input forwarding controller receives the packet from the MAC address, parses the packet headers, and performs a series of lookups to decide whether to accept the packet and how to forward it to its intended destination. It also generates instructions to the data path for the storage and queuing of the packet. Because the ASE2 and LSE switches are cut-through switches, input forwarding lookup is performed while the packet is being stored in the Pause buffer block.

Packet-Header Parsing

When a packet arrives through a front-panel port, it goes through the ingress pipeline, and the first step is packet-header parsing. The flexible packet parser parses the first 128 bytes of the packet to extract and save information such as the Layer 2 header, EtherType, Layer 3 header, and TCP IP protocols.

These parsed fields are used in a series of forwarding table and access control list (ACL) lookups to determine:

- Destination output interfaces (based on Ethernet learning, IP host route entries, LPM, etc.)
- Compliance of switching and routing protocols (spanning tree, VXLAN, Open Shortest Path First [OSPF], Fabric Shortest Path First [FSPF], Intermediate System–to–Intermediate System [IS-IS], redirects, IP packet checks, etc.)
- Policies (network access rights, storage zoning, permit or deny, security, etc.)
- Control-plane redirection and copying (Bridge Protocol Data Unit [BPDU], Address Resolution Protocol [ARP], Internet Group Management Protocol [IGMP], gleaning, etc.)
- System class-of-service (CoS) classification (input queue, output queue, IEEE 802.1p tagging, etc.)
- Service rates and policers
- SPAN (ingress, egress, drop, etc.)
- Statistics (flow and interface packets, byte counters, etc.)
- Network flow-based load balancing (multipathing, EtherChannels, etc.)
- Flow samplers (M of N bytes, M of N packets, etc.)
- Packet-header rewrites (next-hop addresses, overlay encapsulation, time to live [TTL], etc.)
- Flow table (to collect NetFlow and analytics information)

Layer 2 and Layer 3 Forwarding Lookup

As the packet goes through the ingress pipeline, it is subject to Layer 2 switching and Layer 3 routing lookups. First, the forwarding examines the destination MAC (DMAC) address of the packet to determine if the packet needs to be Layer 2 switched or Layer 3 routed. If the DMAC address matches the switch's own router MAC address, the packet is passed to the Layer 3 routing lookup logic. If the DMAC address doesn't belong to the switch, a Layer 2 switching lookup based on the DMAC address and VLAN ID is performed. If a match is found in the MAC address table, the packet is sent to the egress port. If no match is found for the DMAC address and VLAN combination, the packet is forwarded to all ports in the same VLAN.

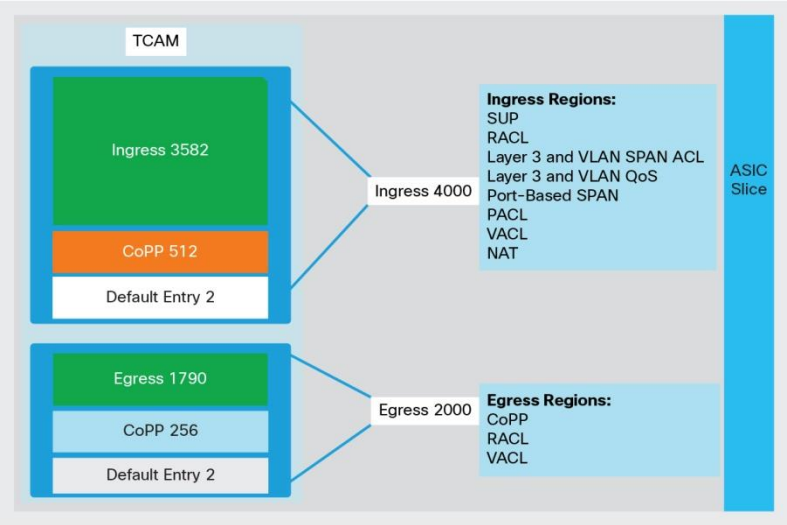
In the Layer 3 lookup logic on the line card, the destination IP (DIP) address is used for searches in the Layer 3 host table. This table stores forwarding entries for directly attached hosts or learned /32 host routes. If the DIP address matches an entry in the host table, the entry indicates the destination port, next-hop MAC address, and egress VLAN. If no match to the DIP address is found in the host table, the packet is forwarded to the fabric module, where an LPM lookup is performed in the LPM routing table.

When Layer 2 switching and Layer 3 host routing is performed, if the egress port is local to the ASIC, packets will be forwarded locally without going to fabric modules. If a packet needs to be routed using the LPM table on the fabric module, it will be forwarded to the fabric module even if the source and destination ports are on the same line-card ASIC.

Ingress ACL Processing

In addition to forwarding lookups, the packet undergoes ingress ACL processing. The ACL TCAM is checked for ingress ACL matches. As shown in Figure 8, each ASIC has an ingress ACL TCAM table of 4000 entries per ASIC slice to support system internal ACLs and user-defined ingress ACLs. These ACLs include port ACLs (PACLs), routed ACLs (RACLs), VLAN ACLs (VACLs), quality of service (QoS) ACLs, and Network Address Translation (NAT) ACLs. ACL entries are localized to each ASIC slice and programmed only where needed. This approach allows optimal use of the ACL TCAM in a Cisco Nexus 9500 platform switch.

Figure 8. ACL TCAM on the New-Generation ASIC



Ingress Traffic Classification

Cisco Nexus 9500 platform switches support ingress traffic classification. On an ingress interface, traffic can be classified based on the address fields, IEEE 802.1q CoS, and IP precedence or differentiated services code point (DSCP) in the packet header. The classified traffic can be assigned to one of the eight QoS groups. The QoS groups internally identify the traffic classes used for the subsequent QoS processes as packets go through the system: for instance, in egress queueing.

Input Forwarding Result Generation

The final step in the ingress forwarding pipeline is to collect all the forwarding metadata generated earlier in the pipeline and pass it to the downstream blocks through the data path. A 64-byte internal header is stored along with the incoming packet in the packet buffer. This internal header includes 16 bytes of iETH (Insieme Ethernet) header information, which is added on the top of the packet when the packet is transferred to another chip in the modular chassis. This 16-byte iETH header is stripped off when the packet finally is sent out the front-panel port. The other 48 bytes of the internal header are used only to pass metadata from input forwarding to output forwarding and are consumed by the output forwarding engine.

Input Data Path Controller

The input data-path controller has the following functions:

- It buffers packet data to handle the latency of the input forwarding controller pipeline.
- It pauses accounting- and flow-control generation.
- It manages Fibre Channel buffer-to-buffer credits.

After a packet's destination is known, the input data path controller sends the packet to the output data path controller.

Central Statistics

The central statistics module provides ASIC internal statistics per packet. The ASE2 and LSE ASICs can have up to eight counters per packet. The sources of the packet increments are:

- Input and output forwarding controllers: These controllers can flexibly count different types of forwarding events by passing up to eight indexes.
- Output data path controller. This controller uses the Central Statistics module to count input and drop statistics on a per-port per-class per-drop per-unicast/multicast/flood per- good/drop basis.

Output Data Path Controller

The output data path controller performs egress buffer accounting, packet queuing, scheduling, and multicast replication. All ports dynamically share the egress buffer resource. The output data path controller also performs packet shaping. The Cisco Nexus 9500 platform uses a simple egress queuing architecture. In the event of egress port congestion, packets are directly queued in the buffer of the egress line card. There are no virtual output queues (VoQ) on the ingress line cards. This design greatly simplifies the system buffer management and queuing implementation. A Cisco Nexus 9500 platform switch can support up to 10 traffic classes on egress, including 8 user-defined classes identified by QoS group IDs, a CPU control traffic class, and a SPAN traffic class. Each user-defined class can have a unicast queue and a multicast queue per egress port.

Output Forwarding Controller

The output forwarding controller receives the input packet and associated metadata from the buffer manager and is responsible for all packet rewrites and egress policy application. It extracts the internal header and various packet header fields from the packet, performs a series of lookups, and generates the rewrite instructions.

Fabric Module Lookup

When a packet is forwarded to a fabric module, the fabric module takes different actions depending on the lookup results on the ingress line card and the forwarding lookup functions the FFT table on the fabric module is programmed for. In the example shown in Table 5 in a previous section, the fabric module has both the IPv6 host route and the IPv6 LPM route table while the ingress line card has the IPv4 host and LPM route tables as well as the Layer 2 MAC tables. If the packet is a Layer 2 switched or IPv4 routed packet, the ingress line card resolves the egress port, the next-hop MAC address, and the egress VLAN information. The fabric module simply forwards the packet to the egress line card. If the packet needs an IPv6 routing lookup, the fabric module searches the IPv6 host route table and the IPv6 LPM route table and uses the best match for the destination IPv6 address to forward the packet. If no match for the destination IPv6 address or an IPv6 default route is found, the packet is dropped

Multicast Packet Forwarding

For multicast packets, the system needs to determine whether packets must undergo Layer 2 multicast forwarding or Layer 3 multicast forwarding. A packet will go through Layer 3 multicast forwarding if it meets the following criteria:

- It is an IP packet with a multicast address.
- It is not a link-local packet.
- Multicast routing is enabled on the bridge domain.

For Layer 2 IP multicast packets, forwarding lookup can be performed using either the Ethernet or IP header. By default, a Cisco Nexus 9500 platform switch uses the IP address for Layer 2 multicast forwarding lookup, but the fanout is limited to the same bridge domain.

For broadcast packets, the ASIC floods the packet in the bridge domain. The ASIC maintains separate per-bridge domain fanout lists for broadcast traffic and for unknown unicast traffic.

IP multicast forwarding in ASE2 and LSE ASICs relies on the forwarding information base (FIB) table. For both Layer 2 and Layer 3 IP multicast, the switch looks up the IP address in the FIB table. The source address is used for the RPF check, and the destination address is used to determine the outgoing interface list. If the FIB destination address search doesn't result in a match, the packet is classified as unknown multicast. If IGMP and Multicast Listener Discovery (MLD) snooping are enabled, the packet is forwarded to all router ports on which the incoming bridge domain is present. If snooping is disabled, the packet is flooded in the bridge domain.

Multiple-Stage Replication

Multicast packets go through the same ingress and egress processing pipelines as unicast packets. However, one difference in the packet lookup and forwarding process is that the Cisco Nexus 9500 platform switches perform three-stage distributed multicast lookup and replication. The multicast routing table is stored on all line cards and fabric modules. The ingress ASIC on the line card performs the first lookup to resolve local receivers. If any local receivers are present, a copy is sent to the local ports. The ingress ASIC then sends a copy of the incoming packet to the fabric module. On receiving the packet, the fabric module performs the second lookup to find the egress line cards. The fabric module replicates the packet to each egress line card on which there are receivers. The egress line card performs the third lookup to resolve its local receivers and replicates the packet on those ports. This multiple-stage multicast lookup and replication is the most efficient way of replicating and forwarding multicast traffic.

Conclusion

The new-generation line cards and fabric modules for the Cisco Nexus 9500 platform switches are built with Cisco's new-generation switching ASICs to lead today's data center switching transition. With the new line cards and fabric modules, the Cisco Nexus 9500 platform switches deliver high-density and high-performance 25, 50, and 100 Gigabit Ethernet ports at an effective cost point. Equally important, the new line cards and modules introduce additional capabilities to meet the new requirements in the data center for increased forwarding capacity, a high-performance VXLAN overlay solution, intelligent buffer management, network telemetry and visibility, etc. With the new-generation line cards and fabric modules, the Cisco Nexus 9500 platform switches provide the leading modular switch platform for building cloud-scale data center networks and outstanding support for converged and hyper-converged infrastructure.

For More Information

http://www.cisco.com/c/en/us/products/switches/nexus-9000-series-switches/white-paper-listing.html

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 www.cisco.com/go/offices.

Printed in USA C11-737199-00 06/16