Design Considerations for Classical Ethernet Integration of the Cisco Nexus 7000 M1 and F1 Modules

White Paper

Last Updated: October 4, 2011

About the Authors

Nimish Desai, Technical Lead, Systems Development Unit (SDU), Cisco Systems

Nimish Desai

Nimish currently works as a Technical Leader in the Data Center Application group within SDU. He was a lead architect on the Virtual Switching System Solution CVD and as well as the best practices designs for Cisco Campus networks. Before his work on the SDU Campus solutions team, Nimish worked with Cisco Advanced Services providing design consultation and technical escalation for large Enterprise customers.

Nimish has been working on inter-networking technology for the last 17 years. Before joining Cisco, Nimish developed expertise with large financial institution supporting trading floor, large-scale design of enterprise networks with logistics and insurance companies, and product development experience with IBM. Nimish hold a MSEE from the New Jersey Institute of Technology. Nimish enjoys fishing and outdoor activities including RVing the National Parks.

Roney Daniel, Technical Lead, Systems Development Unit (SDU), Cisco Systems

Roney Daniel

Roney Daniel is a Technical Leader in the Systems Development Unit (SDU). He joined the Cisco Technical Assistance Center in 2000 and moved to the Financial Test Lab (FTL) in 2002 doing customer-focused testing for large enterprise and financial customers. In the FTL, he also worked on several internal Early Field Trial programs from various business units to validate the Cisco Catalyst 4000, Cisco Catalyst 6000, and Cisco Nexus 7000 family of products. He is currently working on validating Virtualized Multi-tenant Data Center (VMDC) architectures and also doing pre-qualification design work for newer designs with a focus on the Cisco Nexus product line. Prior to joining Cisco, Roney worked in IBM NHD from 1996 to 1999 as a System Test engineer. He holds a Bachelor's degree in Electronics and Communication Engineering.

Design Considerations for Classical Ethernet Integration of the Cisco Nexus 7000 M1 and F1 Modules

Design Considerations for Classical Ethernet Integration of the Cisco Nexus 7000 M1 and F1 Modules

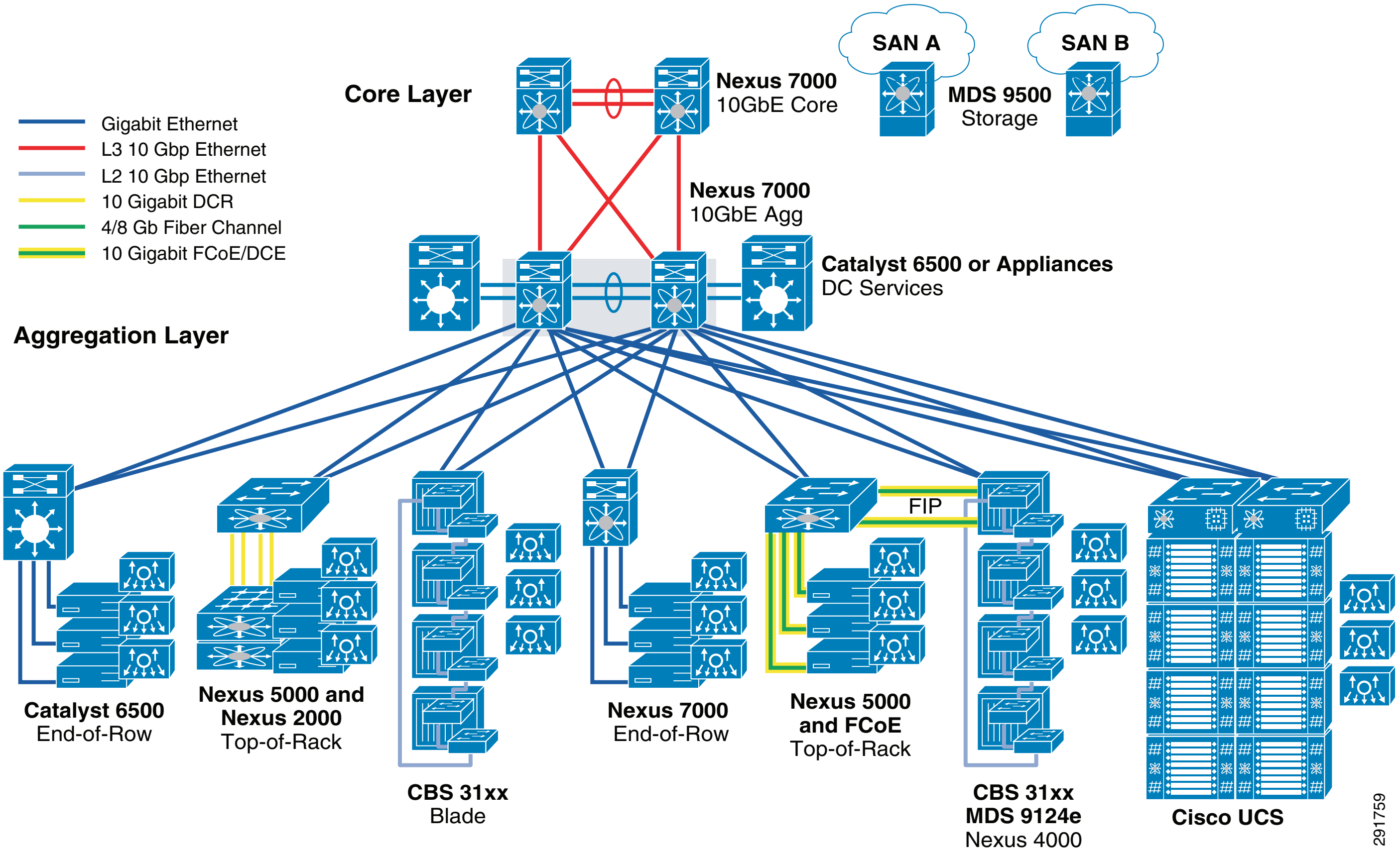
Introduction

Data Center Architecture

As part of Cisco's® ongoing commitment to develop Architectures for Business Transformation, Cisco has developed a fully functional data center cloud reference architecture inclusive of compute, storage, and networking. The modern data center is evolving to meet the new business requirements emanating from Web 2.0 and social network changes to the Internet and corporate networks. Figure 1 depicts the diversity of compute, storage, and networking elements that compose the modern data center. Virtualization and elasticity are two key enablers allowing the modern data center to reach maximum business flexibility at a reasonable cost. Environmental considerations, such as floor space, power, and cooling, must be balanced against new technology adoption and maximizing existing infrastructure investments using compatible technology advances. The next generation modules for the Cisco Nexus® 7000 Series switches enable the data center to efficiently balance existing investments with the adoption of next-generation technologies, such as FabricPath and FCoE. These technologies enable a unified fabric that supports cross-domain network and storage functions and provides higher availability and flexibility of the compute function.

Corporate Headquarters:
Cisco Systems, Inc., 170 West Tasman Drive, San Jose, CA 95134-1706 USA

Figure 1 Diversity in the Data Center



Audience

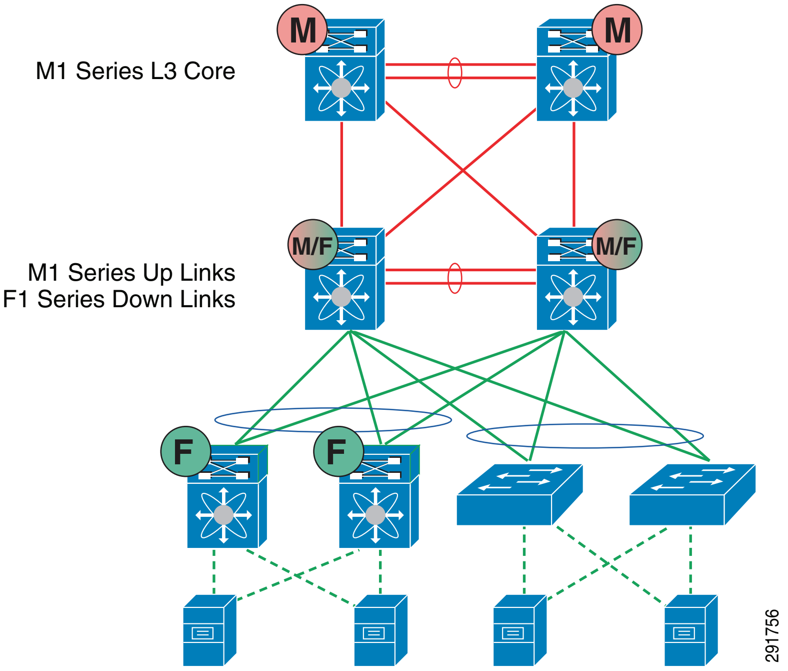
The target audience for this document includes sales engineers, field consultants, professional services personnel, IT managers, Cisco channel partner engineering staff, and customers who have requirements for a private data center and who wish to achieve savings through the use of automation or multi-tenancy or who need to ensure that the design of their data center does not preclude the use of these technologies in the future.

Cisco Nexus 7000 Architecture

Cisco Nexus 7000 I/O Module Families—M-Series and F-Series

The Cisco Nexus 7000 Series I/O modules are built to support the flexible deployment of features in the most demanding of data center environments. The Cisco Nexus 7000 supports two families of I/O modules, the M1 family and the F1 family. Each I/O module family offers different, complimentary levels of performance, scalability, and features. This section discusses the architecture and operation of each I/O module family, identifying the typical use cases for each.

Figure 2 Topology with F1 and M1 I/O Modules



M1 I/O Module Family

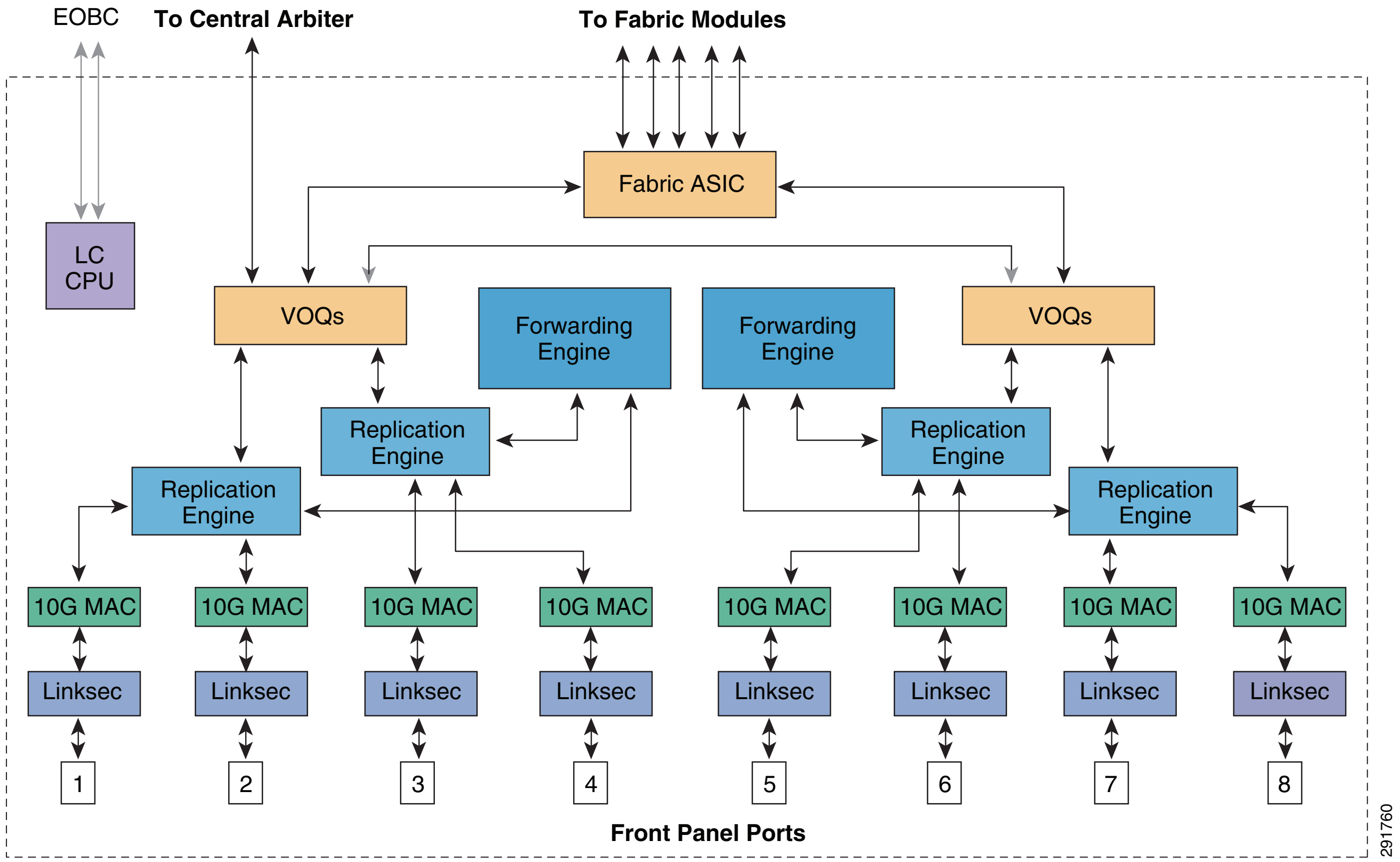
M1 modules support highly-scalable and rich Layer 2 and Layer 3 IPv4 and IPv6 features and are recommended for core, aggregation, and access network environments that benefit from IP-based services and secure segmentation. M-Series XL modules support larger forwarding tables for storing routes, ACLs, etc., as summarized in Table 1. M-Series modules are frequently required at the network core, peering, and aggregation points. When used with the F1-Series, the M-Series modules provide inter-VLAN services and form a pool of Layer 3 resources for the system.

Table 1 Table Sizes for M-Series Modules

	M1 Modules	M1 Modules XL (With Scalable Features License)
Layer 2 MAC Address Table	128,000	128,000
Layer 3 FIB Table	128,000	1,000,000
ACL TCAM	64,000	128,000
NetFlow Table	512,000 entries	512,000 entries

The M1 10 Gbps modules provide up to 80 Gbps of bandwidth to the switch fabric and up to 512 10 Gbps ports (4:1 oversubscribed) in a single 18-slot chassis, providing a high-density, compact solution for large 10Gbps Ethernet networks.

Figure 3 8-port M1-XL I/O Module Architecture



Every M1 I/O module contains one or more integrated forwarding engines. This architecture scales the forwarding performance of the chassis linearly by the number of the I/O modules employed. Each M1 forwarding engine delivers 60 million packets per second (Mpps) of Layer 2 and Layer 3 forwarding. The 8-port 10 Gbps I/O module carries two such engines, providing 120 Mpps. Thus, an 18-slot chassis with 16 8-port 10 Gbps M1 I/O modules processes nearly two billion packets per second (Bpps). The fabric interface on M1 family modules delivers 80 Gbps of bandwidth in each direction, providing up to 2.5 terabits per second (Tbps) system bandwidth in a Cisco Nexus 7018 chassis.

The M1 forwarding engine also delivers access control list (ACL) filtering, marking, rate limiting, and NetFlow with no effect on performance. Powerful ACL processing supports as many as 128,000 entries per module and multicast forwarding is built into each I/O module, providing high-bandwidth egress Layer 3 multicast replication.

The M1 I/O modules are deployable in all network environments because they support Layer 2 and Layer 3 forwarding, large forwarding tables (MAC table, FIB TCAM, ACL TCAM), and advanced features such as policing, NetFlow, and 802.1ae LinkSec. To match the deployment requirements, M1-based Cisco Nexus 7000 switches can serve as pure Layer 2 systems, combined Layer 2/Layer 3 systems, or pure Layer 3 systems. Typical deployment scenarios include:

- End- or middle-of-row 1 GE access with 10 Gbps uplinks
- Aggregation of 1 G or 10 Gbps access switch uplink ports at the distribution layer
- 10 Gbps data center or campus core
- M-Series modules, with XL modules and the scalable services license, may be used at the Internet edge.

F1 I/O Module Family

The F1 modules support Layer 2 switching and services with high performance, high density, low latency, and reduced power. The F1-Series also supports Cisco FabricPath technology for up to 16-way multipathing for scalable Layer 2 networks and IEEE Data Center Bridging (DCB) for Fibre Channel over Ethernet (FCoE). The Cisco Nexus 7000 F1-Series 32-Port 1 and 10-Gigabit Ethernet Module offers outstanding flexibility and performance with extensive fabric virtualization and multipath capabilities.

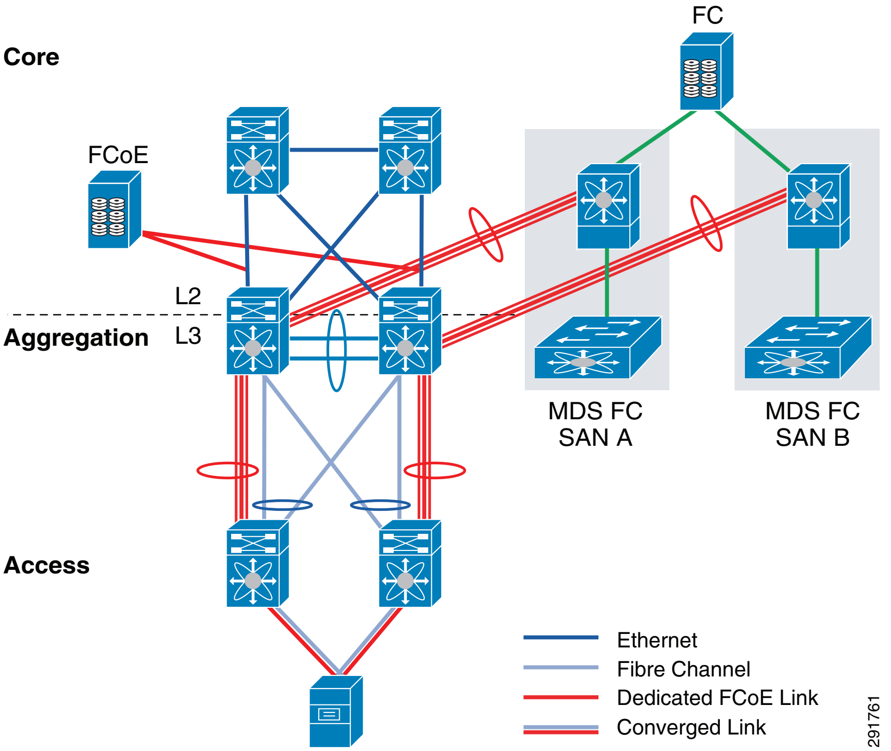
For economical performance, the Cisco Nexus F-Series can be used in the access and aggregation layers. The F-Series modules are being deployed by customers in performance sensitive environments to provide low latency line-rate Layer 2 switching for high-performance workloads. High-density 10 GE server access deployments can be built using the F-Series module in end-of-row server access topologies.

Powered by the F1 Forwarding Engine Switch on Chip (SoC), the 32-port 1G/10G F1 module delivers 480 million packets per second (Mpps) of distributed Layer 2 forwarding and up to 320 Gbps of data throughput. A Cisco Nexus 7000 18-slot switch fully populated with F1 I/O modules can deliver up to 10.2 Tbps of switching performance with a typical power consumption of less than 10 watts (W) per port.

Powerful ACL processing supports 32,000 entries per module in both ingress and egress. Classification and policy is enforced on criteria in Layer 2, Layer 3, and/or Layer 4 fields with no impact on performance. The F1 forwarding engine also supports applications that require port mirroring with integrated hardware support for 16 simultaneous unidirectional switched-port analyzer (SPAN) sessions per module.

The F1-Series delivers integrated hardware support for FCoE and IEEE DCB protocols, as well as Cisco FabricPath, which enables the creation of scalable, flexible networks that efficiently use all available bandwidth between nodes. With the availability of software to support FCoE, the Cisco Nexus 7000 Series switch with F1 I/O modules can be deployed in the server access layer to provide both LAN and storage connectivity via Converged Network Adapters (CNAs). The Cisco Nexus 7000 can also support multi-hop FCoE, which facilitates the use of the Cisco Nexus 7000 as a director-class FCoE aggregation switch, with connectivity to both FCoE access switches and either FC SANs or FCoE storage arrays as shown in Figure 4.

Figure 4 Multi-hop FCoE Topology with Cisco Nexus 7000



While the F1 I/O modules provide high bandwidth and throughput capabilities, they differ from the M1 I/O modules. Because the F1 modules leverage an integrated SoC design, the forwarding tables are smaller than on M1 modules and some features, such as Layer 3 forwarding, NetFlow, traffic policing, and 802.1ae LinkSec, are not supported by the F1 I/O modules.

However, you can combine M1 and F1 I/O modules in a Cisco Nexus 7000 chassis to leverage the complimentary feature sets of both I/O module families: the large-scale forwarding tables and Layer 3 switching capability of the M1 modules, along with the high bandwidth, high density east-west bridging performance of the F1 modules.

Deployment options for F1 family modules include:

- End- or middle-of-row access
- Aggregation of 10 Gbps access switch uplink ports at the distribution layer, typically combined with M1 I/O modules to provide the Layer 3 termination
- 10 Gbps Layer 2 core network and grid computing environments

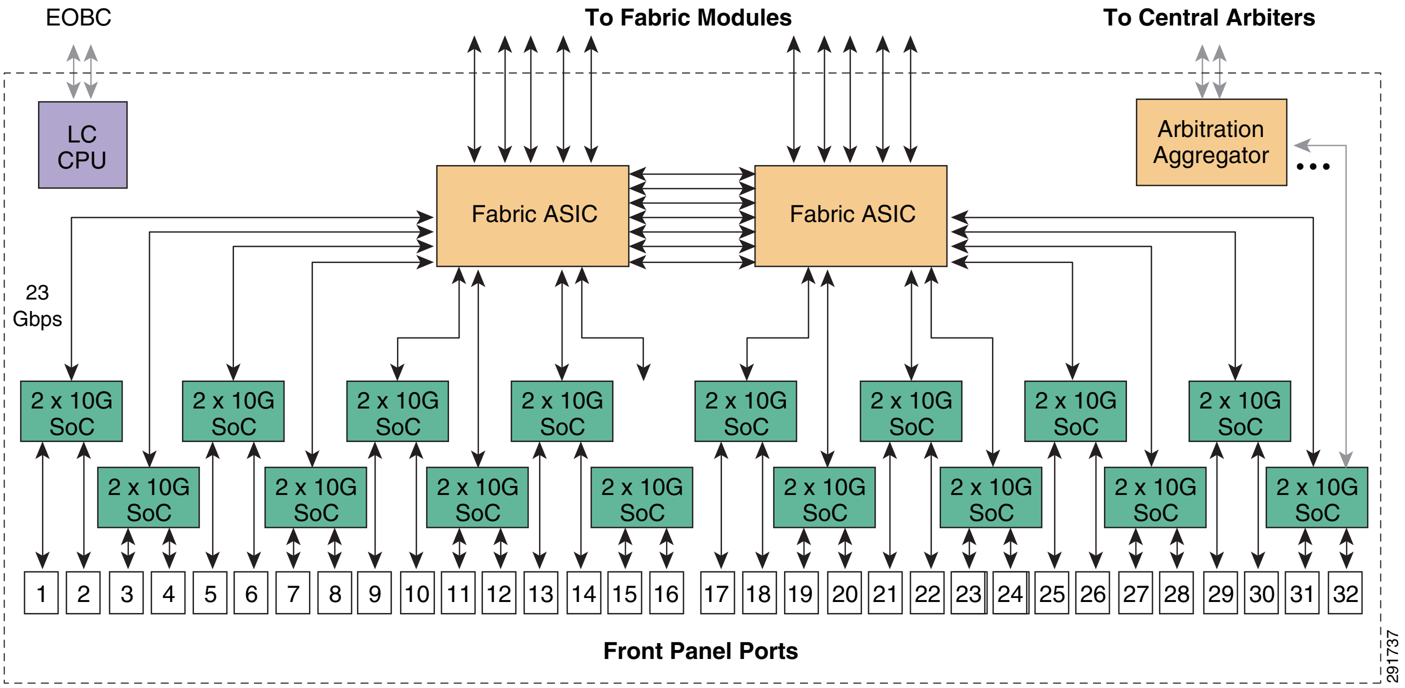
F1 Module Architecture

As depicted in Figure 5, the F1 I/O module has the following primary components:

- The forwarding engine SoC implements all I/O module functions, such as ingress buffering and forwarding look-ups. It also has on-chip memory for MAC tables, ACL TCAM, and Virtual Output Queuing. Each SoC has two front panel 10 GbpsE ports and a 23 G interface to the Fabric ASIC. The SoC design integrates functionality typically offered by purpose built ASICs on M1-Modules into a single chip. The SoC ASIC in the F1 module offers lower latency (4.7 µsec.), advanced features, such as FCoE and FabricPath, and lower power and cost. The main optimizations are in the areas of MAC address table sizes, ACL TCAM size, and Layer 3 functionality.

- The fabric ASIC on the module provides connectivity to the fabric module cross-bar.
- The I/O Module CPU (LC CPU) runs NX-OS microcode, which controls module functionality.
- The Arbitration Aggregator chip aggregates all the arbitration signals and sends them to the central arbiter chip on the supervisor.

Figure 5 F1 I/O Module Architecture



Cisco Nexus M1 and F1 and Proxy Routing

When combining M1 and F1 modules in the same chassis, the system automatically configures the F1 modules to send traffic requiring unicast or multicast routing over the switch fabric to the available M1 modules, a technique called proxy routing.

Proxy routing consists of three key steps:

1. A F1 I/O module sends a packet requiring routing over the fabric to a M1 module. F1 modules know which packets require routing based on the destination MAC address, which for routed traffic is the MAC address of the gateway (either the burned-in MAC address or an HSRP/VRRP/GLBP virtual MAC address).
2. A M1 I/O module receives a packet that requires proxy routing from a F1 module and performs the necessary ingress and egress forwarding decisions to derive the correct output port.
3. A M1 I/O module sends the packet to the correct output port (possibly sending it back across the fabric).

By default, all M1 I/O modules in the system share the load of proxy routing. Proxy routing consumes bandwidth and forwarding engine throughput on the M1 modules that participate, sharing bandwidth with other traffic that might be traversing the M1 modules. Therefore, a configuration option is provided allowing you to specify which M1 I/O modules participate in proxy routing.

When deploying proxy routing with M1 and F1 modules, consider the following:

- The number of M1 modules required depends on the routing requirements; both inter-VLAN and routed (VLAN to routed M1 interface) traffic require proxy routing if the packet enters the switch on a F1 interface.
- F1 modules provide the optimal benefit by increasing the network capacity for east-west bridged traffic in the access or aggregation layer of a network. In the network core layer, where the majority of the traffic is routed, M-Series modules should be used.
- The front panel M1 uplink ports can be used for proxy routing; however, the available bandwidth on the M1 modules is shared between proxy routing and other traffic.
- The M1 I/O module a particular flow uses for proxy routing is based on a hash function similar to the global port-channel load-balance configuration. Traffic is spread among all M1 modules participating in proxy routing on a per-flow basis.
- Every packet in every flow that requires proxy routing traverses the fabric to reach a M1 module. No caching of such flows or similar mechanisms exist on the F1 modules.

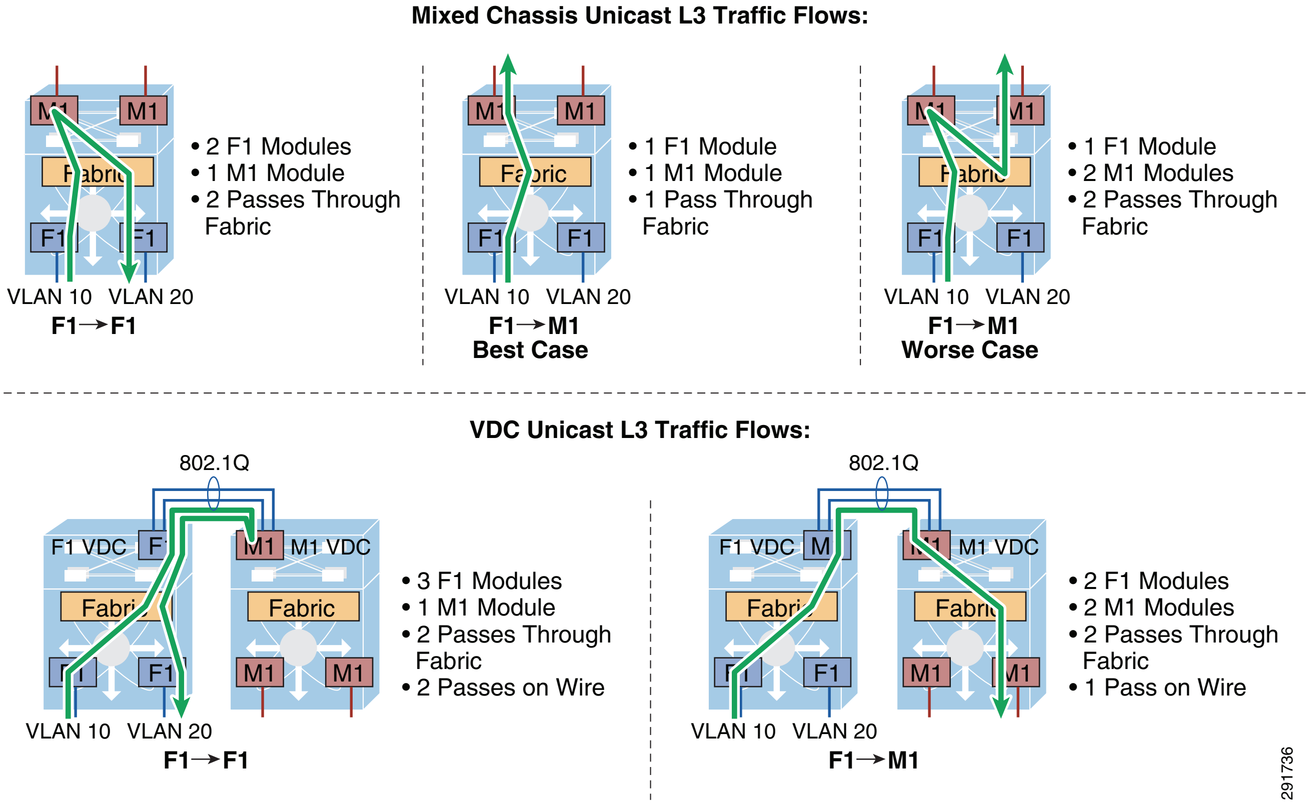
Cisco Nexus 7000 M1/F1 Chassis Modes

The Cisco Nexus 7000 Series can be configured using one of two modes that integrate M-Series and F-Series cards in the same chassis. You can configure M-Series modules and F-Series modules in their own distinct Virtual Device Contexts (VDCs). VDCs enable complete separation of control plane and data plane functionality so the system operates as two logical devices with no feature interactions between the two logical contexts. In this mode, the F-Series module context provides functions, such as FCoE and FabricPath, but it is not able to route Layer 3 traffic. Instead, the M-Series only routes VDC Layer 3 traffic.

Alternatively, the system can be configured using a mixed chassis VDC where both M-Series and F-Series modules are included in the same chassis. In this configuration, the capabilities of both modules exist in the same system. The F1 modules provide low latency for switched traffic while the M-Series modules provide proxy Layer 3 functionality for F1 modules.

Depending on the configuration, we can observe multiple traffic flow scenarios as shown in Figure 6.

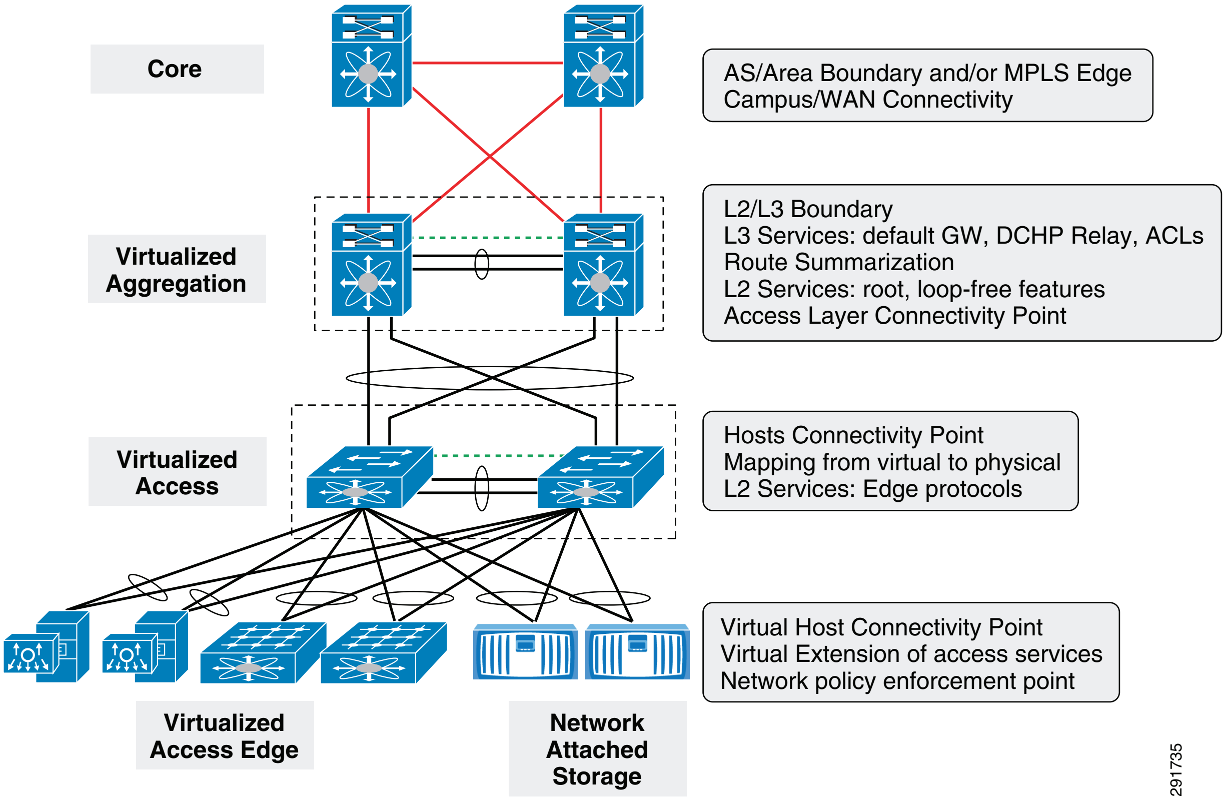
Figure 6 Traffic Flow



M1 and F1 Integration in Classical Ethernet

The traditional three-tier architecture (see Figure 7) is detailed in the following reference design guide: http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/DC_3_0/DC-3_0_IPInfra.html.

Figure 7 Three Tiered Architecture


Core
AS/Area Boundary and/or MPLS Edge
Campus/WAN Connectivity
Virtualized Aggregation
L2/L3 Boundary
L3 Services: default GW, DCHP Relay, ACLs
Route Summarization
L2 Services: root, loop-free features
Access Layer Connectivity Point
Virtualized Access
Hosts Connectivity Point
Mapping from virtual to physical
L2 Services: Edge protocols
Virtual Host Connectivity Point
Virtual Extension of access services
Network policy enforcement point
Virtualized Access Edge
Network Attached Storage
291735

A highly-available infrastructure is the fundamental backbone of any virtualized data center architecture. The infrastructure design is based on a three-tier model (core, aggregation, and access) as depicted in Figure 7. Cisco network platforms enable the consolidation of various functions at each layer with innovation in access technology, creating a single architectural platform for optimized resource use. From a hierarchical perspective, two key layers are discussed in this document:

- Aggregation layer—Traditionally, the aggregation layer is designed using a pair of hardware switches, enabling network connectivity at various speeds and functionality. With the Cisco Nexus 7000 Series, the VDC capability enables the consolidation of multiple aggregation topologies consisting of multiple distribution blocks, represented as a pair of Cisco Nexus 7000 switches. This document addresses one such distribution block. The distribution block primarily addresses the aggregation density and Layer 3-to-Layer 2 services, as well as inter-VLAN routing.
- Access layer—The access layer in a virtualized data center consists of a physical access layer and virtual access layer. The physical access layer provides connectivity to physical hosts, storage, and back-up devices and maps the virtual access layer devices to physical infrastructure. The traffic flows at this layer are more confined to virtual machine (VM)-to-VM and VM-to-storage (localized). This layer also provides policy enforcement for the traffic flow localized to this layer.

This document focuses on classical Ethernet deployments with the aggregation layer as shown in Figure 7. The F1 module supports FabricPath and FCoE technologies, however the validation and application of these features are outside the scope of this document. Virtualization of compute, network, and storage resources enables unprecedented flexibility. The classical Ethernet deployment and F1 module integration should consider the following critical areas of the three-tier architecture for successful integration:

- Topology Considerations

- Capability Selection
- Aggregation and Scalability

Topology Considerations

The hardware must be flexible enough to support a multitude of topologies, ranging from a classical Ethernet design (looped and non-looped STP topology) to FabricPath (non-STP-based topology), as well as a topology based on storage (FCoE) supporting a separate fabric. In this document, the topology refers to the integration of classical Ethernet loop-free vPC-based topologies. This document uses four distinct combinations of vPC-based topologies to describe the design criteria when using F1 modules.

Capability Selection

Module usage defines the ability of a data center to support various port densities, as well as feature and technology integration, at various price points per port. The appropriate application and integration of M1 and F1 modules defines the sustainable price and feature requirements for a given configuration. The proper combination of M1 and F1 I/O modules in a chassis enables one to design a scalable aggregation layer topology. This design considers the following capabilities, which require specific design considerations when combining M1 and F1 modules in a given system:

- Layer 3 capability for uplink and interconnects at the aggregation layer
- Layer 2 capability for east-west traffic flows
- Peer-link capability for supporting scalable loop-free topologies
- Proxy routing capability for traffic from Layer 2 to Layer 3 for both inter-VLAN (east to west) and server to users (south to north)

The capacity planning for each of the above requirements depends on aggregation layer scalability design considerations.

Aggregation and Scalability

In a typical data center design, the aggregation layer requires maximum flexibility, scalability, and feature integration because aggregation devices constitute the Layer 3 and Layer 2 boundary, which requires both routing and switching functionality. Access layer connectivity defines the total forwarding capability, port density, and Layer 2 domain flexibility. To increase aggregation capacity using F1 modules, the design must consider the extent of the Layer 2 domain. In turn, the Layer 2 domain design must plan for access layer device density and workload mobility (for example, VMware® vMotion™). The resulting design requirements define the pod. Multiple pods can be enabled to support variable workloads. Designs that center on F1 modules limit the pod to 14,000 unicast MAC addresses. This upper limit does not necessarily dictate the limits of the F1 module, as multiple 14,000 MAC domains (pods) can co-exist within F1 modules. The 14,000 MAC Layer 2 domain satisfies most data center workload requirements.

The design options proposed are validated using unicast traffic flows. However, the behavior and implication of technologies, such as multicast, quality of service (QoS), and services module integration, may alter the design and topology selections. This document considers neither these technology features nor their implications.

Design and Validation

This section describes the practical design options available with M1 and F1 modules in a mixed chassis and considers the scope and the criteria defined in Cisco Nexus 7000 I/O Module Families—M-Series and F-Series. It only addresses commonly deployed scenarios. The validation consists of four pods with the technology, topology, and the scale factors described in the following sections.

POD Topologies

The pod topologies evaluated in this document are based on vPC technology and, as such, any possible combination of topologies constitutes a loop-less STP configuration. The F1 module capabilities as well as link redundancy are two key factors in topology selection. As described in F1 Module Architecture, a pair of 10 Gbps ports constitute the SoC. A single SoC supports up to 16,000 MAC addresses. When selecting from the four proposed topologies, the primary design consideration prioritizes SoC efficiency over availability. SoC efficiency is governed by optimal use of the resources available for a given SoC (MAC address, TCAM resources for ACL, etc.). Availability requirements define the failure domain, for example, single versus dual links between a pair of access layers connected to same F1 module.

Considering the capability of SoC and the possible connectivity combinations (link redundancy) between access layer links to F1 modules, the following four pod configurations were validated:

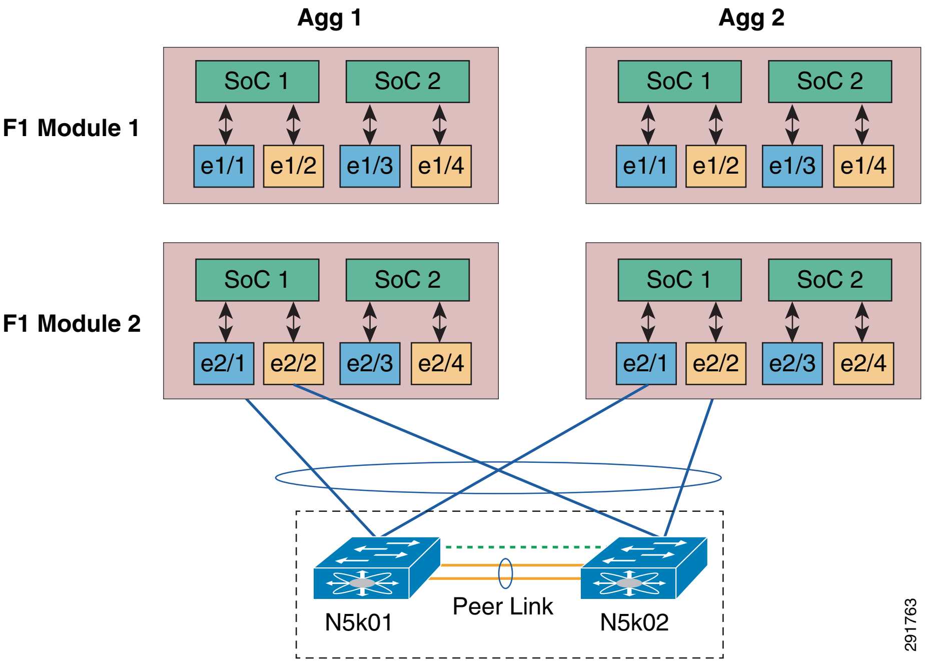
- Pod 1—Two Uplinks from a Cisco Nexus 5000 Pair to the Same SoC
- Pod 2—Two Uplinks from a Cisco Nexus 5000 Pair to Different SoC
- Pod 3—Four Uplinks from a Cisco Nexus 5000 Pair to Inter-Card, Same SoC
- Pod 4—Four Uplinks from a Cisco Nexus 5000 Pair to Inter-Card and Inter-SoC

Note In this validated topology, Cisco Nexus 5000 switches were used, however one can deploy Cisco Nexus 7000 switches with F1 modules instead.

Pod 1—Two Uplinks from a Cisco Nexus 5000 Pair to the Same SoC

In this topology, a pair of Cisco Nexus 5000 switches connect to ports in the same SoC. These connections define the MAC domains for that pod.

Figure 8 Two Uplinks from Cisco Nexus 5000 Pair to Same SoC

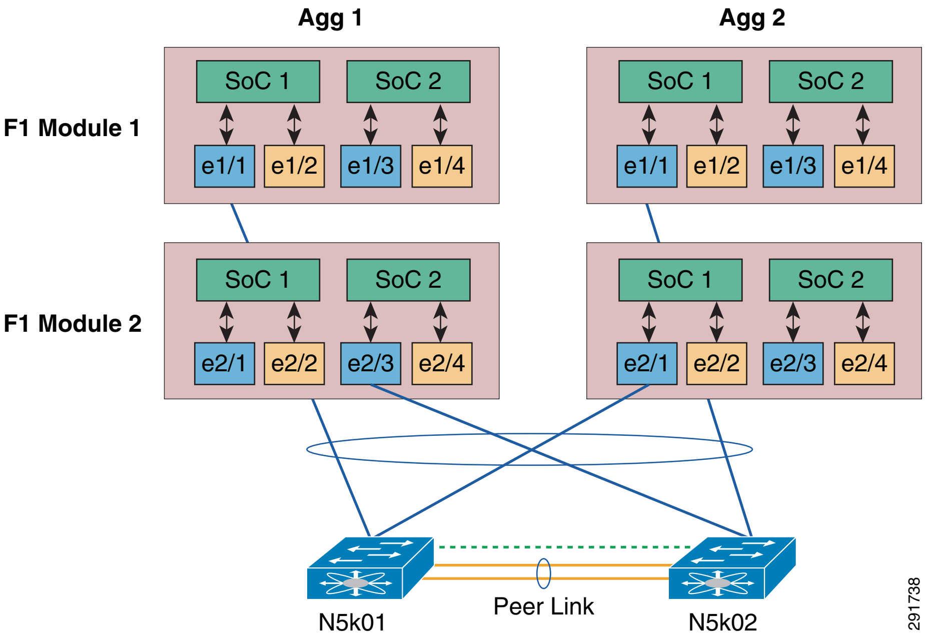


Two links to the same SoC provide for better MAC address usage because it limits the possibility of oversubscribing MAC addresses through workload growth within the pod. However, the SoC represents a single point of failure for both of the access layer vPC uplinks. This limitation forces server-to-user traffic (south-north) to rehash on the remaining PortChannel ports of both Cisco Nexus 5000s (access layer devices), while the north-south traffic at the aggregation traffic is redirected over the vPC peer-link.

Pod 2—Two Uplinks from a Cisco Nexus 5000 Pair to Different SoC

In this topology, a pair of Cisco Nexus 5000 switches connect to two different SoC ports on different modules.

Figure 9 Two Uplinks from Cisco Nexus 5000 Pair to Different SoC



With two links connected to different SoCs, availability improves as a single F1 module failure does not force traffic to redirect over the peer-link at the aggregation layer. However, this topology leaves an unused port on the SoC open to use by other pod connectivity, which can result in oversubscribing the SoC as it can be unaware that existing connectivity has already committed all available resources. If planned properly, an additional pair of access layer switches belonging to the same pod can reside within the same Layer 2 domain and can connect to the open ports of the SoC to enable better port usage planning.

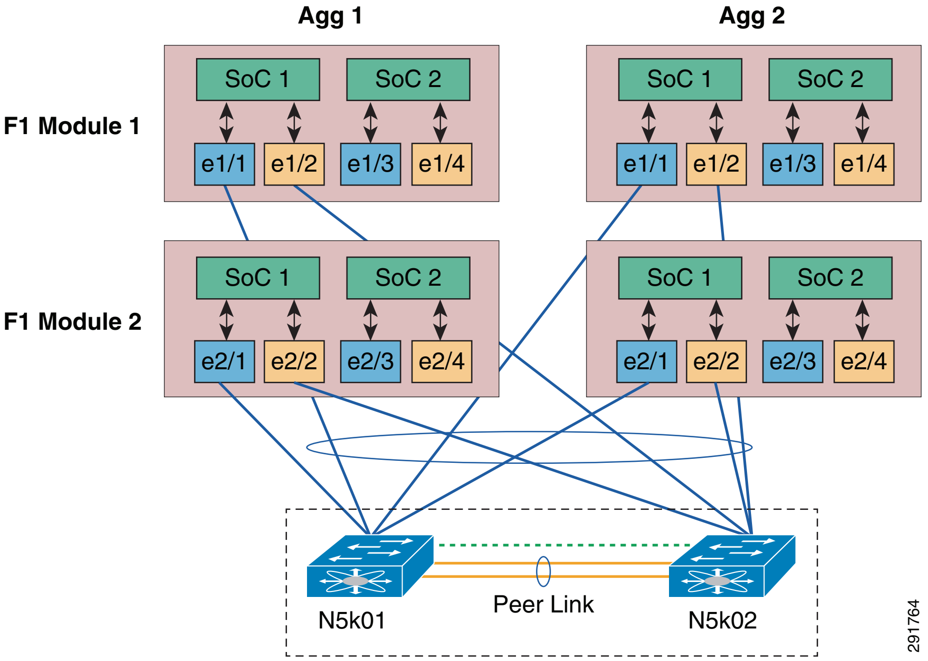
If a F1 module fails, server-to-user traffic (south-north) must rehash over the existing PortChannel ports of both access layer devices, while user-to-server traffic (north-south) selects the remaining F1 module port to the access layer devices.

The two remaining pod options discuss two uplinks from each Cisco Nexus 5000 to a Cisco Nexus 7000, totaling 40 G per aggregation layer.

Pod 3—Four Uplinks from a Cisco Nexus 5000 Pair to Inter-Card, Same SoC

In this topology, each access layer switch in the vPC pair has two links that connect to two F1 modules in a PortChannel configuration. This configuration enables full allotment of SoC resources on both modules.

Figure 10 Four Uplinks from Cisco Nexus 5000 Pair to Inter-Card, Same SoC



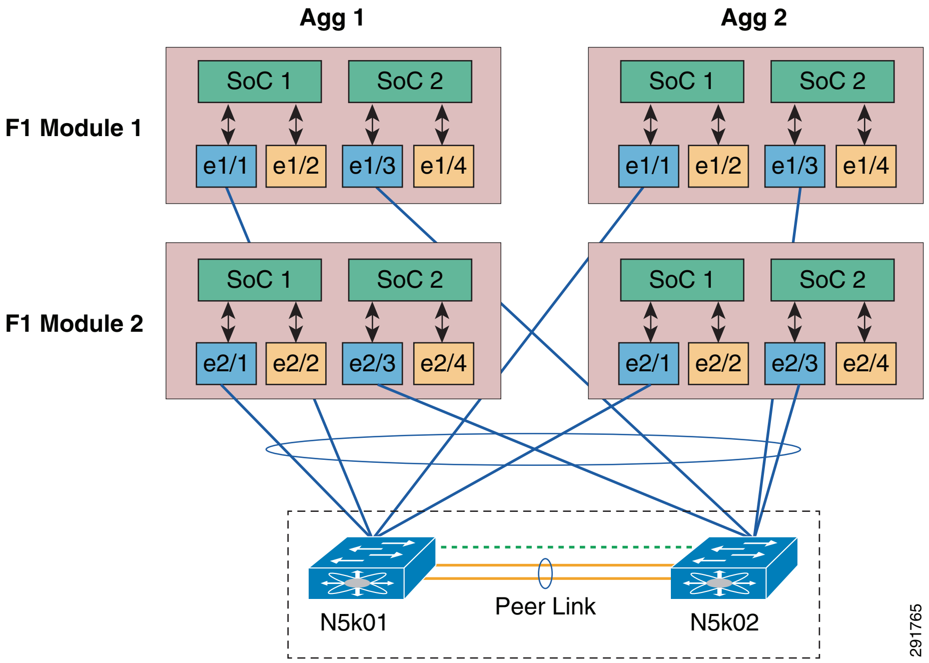
This configuration improves availability compared to Pods 1 and 2 as no single point of failure exists. It also constrains the possibility of oversubscribing the SoC MAC addresses within the connected pod, which improves efficiency.

As with Pod 2, if a F1 module fails, server-to-user traffic (south-north) must rehash over the existing PortChannel ports of both access layer devices, while user-to-server traffic (north-south) selects the remaining F1 module port to the access layer devices. Using four links from each access layer device also localizes the failure of any F1 module to the aggregation layer, allowing both aggregation layer devices to still forward traffic.

Pod 4—Four Uplinks from a Cisco Nexus 5000 Pair to Inter-Card and Inter-SoC

The Pod 4 topology is similar to that of Pod 3 where the uplinks on both access layer switches connect to different SoCs on different modules.

Figure 11 Four Uplinks from a Cisco Nexus 5000 Pair to Inter-Card and Inter-SoC



However, this topology does not increase availability as it introduces the possibility of oversubscribing each of the SoC pairs with future connectivity to additional Layer 2 domains. That said, you can connect an additional pair of access layer switches under the same layer two domain using the open ports of each SoC, which improves port usage.

As with Pod 3, if a F1 module fails, server-to-user traffic (south-north) must rehash over the existing PortChannel ports of both access layer devices, while user-to-server traffic (north-south) selects the remaining F1 module port to the access layer devices. Using four links from each access layer device also localizes the failure of any F1 module to aggregation layer, allowing both aggregation layer devices to still forward traffic.

Final Topology Considerations

Additional topologies are possible based on the desired availability, efficiency, and convergence. Considering the four pod topologies discussed, the final topology selection depends on the uplink configuration of the following types:

- With a single uplink for each access layer vPC pair, the decision hinges on whether to prioritize convergence or oversubscription:
  - If convergence is most important, use Pod 2 and plan the MAC addresses usage for the remaining open ports.
  - Otherwise, Pod 1 limits the oversubscription of MAC addresses to a single pod as no other pods can be connected.
- With two uplinks for each access layer vPC pair, Pod 3 is optimal as it offers high availability and efficient use of MAC addresses.

Aggregation, Scaling, and Capacity Planning

Typically, aggregation design considers several factors to develop a scalable model that is optimal for application service-level agreements, such as traffic types and response time. The number of ports used for connectivity is determined by the required bandwidth and oversubscription ratios. Key connectivity points affecting the port roles of modules at the aggregation layer are:

- Layer 3 ports for connectivity to the core
- Layer 3 ports interconnectivity between aggregation layer devices
- M1 module ports for proxy routing bandwidth
- Layer 2 peer-link connectivity for vPC topology
- Layer 2 switch ports for access layer device vPC connectivity

When deploying the M1 module without the F1 module, consider the following best practices when planning interconnectivity to the Layer 3 core devices at the aggregation layer:

- Use dedicated hardware port for 10 Gbps connectivity to the core to reduce oversubscription and do not share a single module to multiple VDC instance to achieve maximum availability and flexibility in operational management.
- To increase the bandwidth and reduce convergence times, deploy a full mesh topology with additional Layer 3 PortChannel ports.
- Use redundant Layer 3 interconnects between aggregation layer devices to reroute traffic when the uplinks of an aggregation layer device fail.
- Use summarization of the access layer subnets to enable stub networking for the access layer region and employ faster convergence timers for OSPF.

When deploying F1 modules with M1 modules in a mixed chassis deployment, all of these best practices apply. However, the hardware capability of F1 and M1 modules differs considerably. A F1 module can only enable Layer 2 functionality, while a M1 module can enable either Layer 2 or Layer 3. Therefore, the M1 card must be used for Layer 3 functionality and use either M1 or F1 modules for Layer 2 capabilities. When integrating F1 modules with M1 modules, consider the following additional best practices and design considerations:

- The F1 module can only support more than 16,000 MAC addresses when multiple distinct Layer 2 workload domains exist and each domain comprises fewer than 16,000 MAC addresses. However, to support multiple Layer 2 domains in a vPC topology, the peer-link must be able to carry the MAC addresses of all Layer 2 domains. Carrying the MAC addresses of all Layer 2 domains is not possible if peer-link is configured on the F1 module ports. Therefore, the peer-link for the vPC must be configured on the M1 module when topologies have more than 16,000 MAC addresses.
- Plan for the M1 port resources for the following:
  - Routing (Layer 3 uplink and Layer 3 interconnect)—Best practices call for dedicated Layer 3 uplink ports. The Layer 3 interconnect can be shared with other functionality, such as Layer 2 and peer-links, because it is typically unused as it is a non-optimal path for Layer 3 routing protocols.
  - Proxy routing of traffic coming from F1 ports—The requirements of proxy routing are described in Cisco Nexus M1 and F1 and Proxy Routing. The M1 port resources are required to route between F1 ports (inter-VLAN) and from F1 ports to the Layer 3 uplink (from server to user).

- Another design decision is whether to dedicate the M1 port to a specific function (such as routing, proxy routing, or a peer-link) or to share that port among more than one function. Cisco recommends using dedicated M1 ports for uplink functionality and proxy routing. Dedicated ports reduce oversubscription and reduce the failure domain, which can invoke several functional dependencies, influencing convergence within a reasonable time.
- Plan the oversubscription ratio. To plan for oversubscription, consider bandwidth requirements for the following distinct traffic flows:
  - Server-to-User Traffic (South to North)
  - User-to-Server Traffic (North to South)
  - Server-to-Storage and Server-to-Server Traffic (East to West)
  - Layer 3 and Layer 2 Interconnect Bandwidth

Server-to-User Traffic (South to North)

Server-to-user traffic oversubscription planning requires resource allocation for Layer 3 uplink and proxy routing. In a typical design, two 10 Gbps links are configured from each aggregation layer device to core devices using a full mesh topology. For less oversubscription, add additional links from each aggregation layer device. This decision affects proxy routing resources and the Layer 3 interconnects during failure. Cisco recommends using the same bandwidth allocated to the uplinks from each aggregation device for the proxy routing and Layer 3 interconnects. This minimum requirement reduces packet loss and manages congestion during normal and failure conditions. This document validates two PortChannel bundled 10 Gbps links between each core device, representing 40 G of uplink bandwidth. Additional bandwidth planning is required for proxy routing resources as described in the following sections.

User-to-Server Traffic (North to South)

User-to-server oversubscription planning follows similar design rules because the maximum traffic that can be forward to the aggregation layer is determined by the number of links connected from two core devices. The most notable exception is that this traffic bandwidth planning does not require resources for the proxy routing. In other words, the traffic from the user (core switches) does not use proxy routing tunnels because M1 ports forward it to Layer 2 ports.

Server-to-Storage and Server-to-Server Traffic (East to West)

The amount of server-to-storage and server-to-server traffic depends on where storage is attached in the network (aggregation or access layer), the mobility of virtual machines, the virtual machine configuration (separate interfaces for front-end and back-end traffic), and the size of the Layer 2 domain. If the pod design has a single pair of access layer devices and storage is attached at the access layer, all VM-to-VM and VM-to-storage traffic remains local to those access switches. However, inter-VLAN traffic must be routed at the aggregation layer (Layer 3 FHRP gateways) and that traffic must be forwarded using proxy routing. As a rule of thumb, the minimum network oversubscription for this traffic matches the bandwidth allocated for uplink capacity. In this document, 40 G worth of traffic is assumed to be server-to-server and the guidance matches that provided in Server-to-User Traffic (South to North).

Layer 3 and Layer 2 Interconnect Bandwidth

In a traditional best practice design, both aggregation layer devices are connected through redundant Layer 3 links to avoid black holing the traffic during a failure of all uplinks from one of the aggregation devices. During failover, all traffic originating from server to user that reaches failed devices traverses the Layer 3 interconnect and is forwarded by aggregation devices with uplinks connected to core switches. This bandwidth requirement is primarily controlled by the number of vPC-enabled access layer links connected to the aggregation devices. The same requirement is true for Layer 2 interconnect (peer-links in a vPC-based topology) bandwidth planning. Both connectivity requirements are critical for redundant design and improve availability. It is highly uneconomical to match the bandwidth of the entire access layer to links between aggregation devices. Therefore, when estimating the minimum bandwidth to deploy for interconnect links (Layer 3 and peer-link), the recommendation is to at least match the bandwidth offered from each aggregation to Layer 3 upstream core switches. In addition, decide whether you want separate or combined Layer 3 and peer-link functionality between aggregation layer devices. This design consideration requires a trade-off between convergence/availability and complexity/economics. The mixed use of interconnect links between the aggregation layer devices (peer-link and Layer 3 functionality on the same PortChanneled links) invokes multiple functions and dependencies, which can create longer convergence offset by a more economical use of 10 Gbps ports on M1 I/O cards.

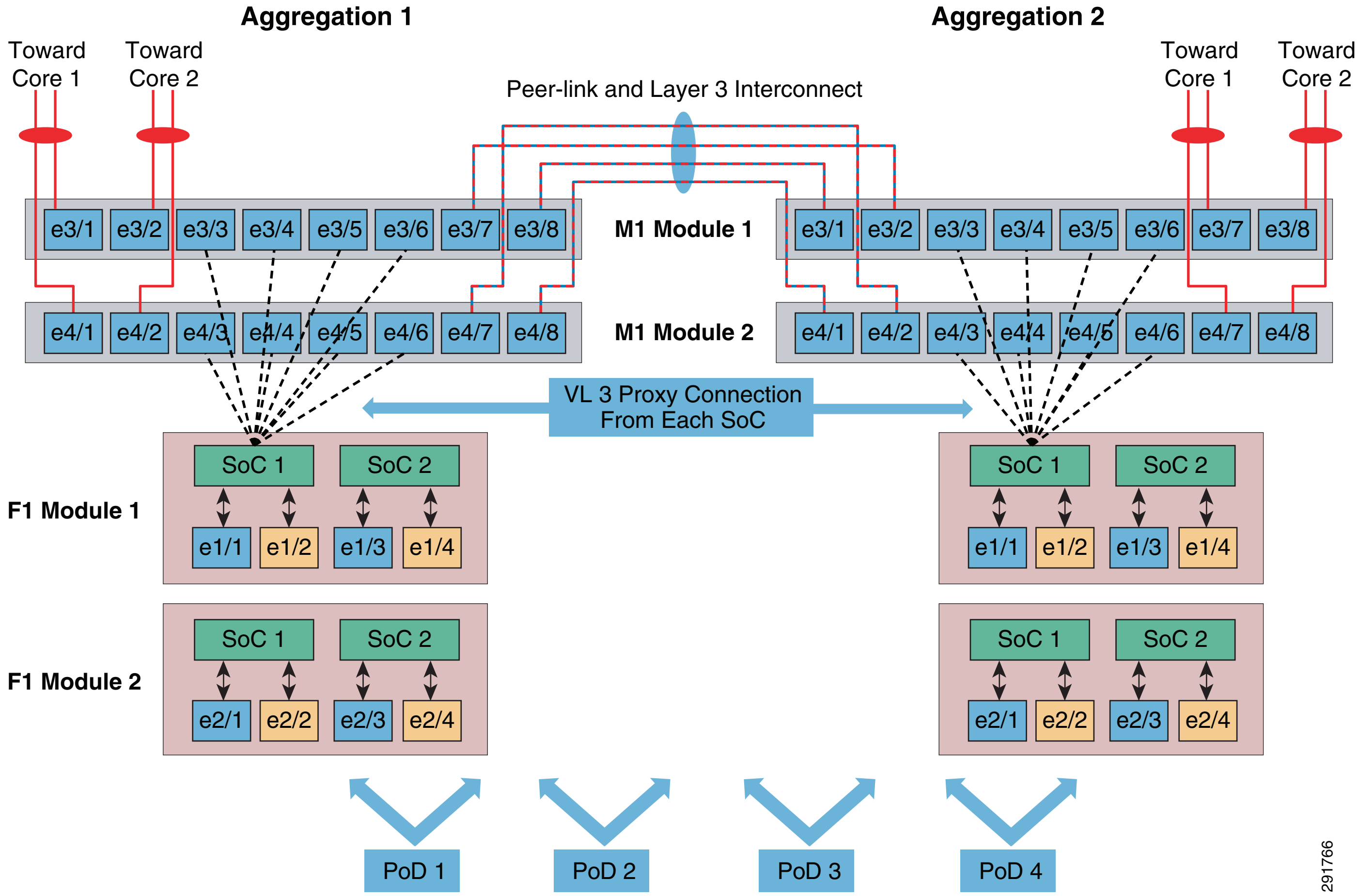
In this document, four links (40 G PortChannel bundled) between aggregation layer devices served as Layer 3 interconnect (SVI-based routing between aggregation layer) and peer-link (vPC topology) represented by two PortChannel routed links from aggregation layer devices to the core switches.

Note The multicast traffic design consideration requires further estimates of bandwidth consumed for the traffic traversing on peer-link.

Layer 3 Proxy Routing Design, Configuration, and Monitoring

Cisco Nexus M1 and F1 and Proxy Routing describes why the Layer 3 proxy routing function is essential for Layer 2 traffic to be routed to the rest of the network.

Figure 12 Cisco Nexus M1 and F1 Modules with Multiple SoC-to-M1 Connections



The design considerations for proxy routing are:

- Redundancy of hardware—Number of M1 modules
- Bandwidth required to be routed for two types of traffic—East-west and south-north

In a typical design, redundant M1 modules are recommended to avoid a single point of failure for vPC peer-link and uplink connectivity to core switches. Any M1 port can be used for multiple functions (shared resources), including routed traffic, Layer 2 switching, peer-link as well as proxy routing. The decision to dedicate hardware resources to a specific function depends on the design choice between oversubscription/economics and efficiency/simplicity.

Each SoC (two ports) forms an internal virtual Layer 3 connection to the M1 port that is designated for proxy routing Layer 2 traffic. The M1 port designated as the shared resource for multiple functions can lead to oversubscription of the hardware resources. When using multiple M1 modules, recirculation of traffic cannot be avoided if you dedicate a single port on one module to proxy routing and one port on another module to Layer 3 forwarding because the flow of traffic can be hashed such that internal re-circulation is required between the two modules (see Figure 6, which is the F1-M1 worst case). In a typical design, diversification of modules reduces the number of failure points. Module diversification is also recommended when designing redundancy for proxy routing. The recommendation is to designate an equal number of ports on two modules for proxy routing.

The number ports required for proxy routing is governed by a second factor, bandwidth. Server-to-User Traffic (South to North) recommends allocating as much bandwidth as is supported by the uplinks configured from each aggregation layer device. Additional bandwidth planning is required to allow for inter-VLAN routed traffic. In this design, 40 G of additional bandwidth was allocated for east-west traffic. Therefore, a total of 8 ports (80 Gbps of bandwidth) is allocated for proxy routing.

An internal PortChannel, called the router PortChannel or Virtual Layer 3 (VL), is used to send Layer 3 proxy traffic from a F1 module to a M1 module and to distribute the load among the available M1 forwarding engines. The characteristics and configuration of M1 ports designated as proxy routing ports is based on the following considerations:

- You cannot restrict the designated proxy routing ports that are used by other functions (Layer 3 or Layer 2 switching, peer-link port, etc.). The status on a configured port does not show its use status. Therefore, you must know the configuration specifics that apply to proxy routing. A physical port should be labeled with text that clarifies its intended use.
- You cannot restrict which F1 ports use a given M1 port. If the M1 port is designated to proxy routing, then all F1 ports in the system use it.
- One router PortChannel is created per VDC and it includes all VLANs enabled within the VDC.
- Only the interfaces on M1 modules within the VDC are defined as members of router PortChannel.
- The router PortChannel has a new interface type identified by iftype (0x23).
- Hashing on the router PortChannel is flow-based and is the same as the global PortChannel load balancing configuration.

Proxy Layer 3 Configuration

By default, proxy Layer 3 forwarding is enabled when the chassis is brought up with a M1 and F1 module installed. The system uses up to 128 available proxy forwarders as default. However, you can configure specific front panel ports on a M1 module with proxy routing or uplink functionality. Similarly, you can configure replicators for egress multicast replication.

The command provides two arguments that restrict the use of proxy forwarders:

```
N7k1-AGG1(config)# hardware proxy layer-3 forwarding ?
  exclude  All Available Members except the following
  use      Specify Members
```

Note Use caution, as these arguments are functionally opposite in nature.

The **hardware proxy layer-3 forwarding exclude** command excludes the port from proxy routing so the 10 Gbps of available bandwidth can be used by other features.

The **hardware proxy layer-3 forwarding use** command configures the port for proxy routing. This command has the following noteworthy restrictions:

- This command does not disable the front panel port, so it does not prevent the user from using it for an alternate purpose.
- At least one port in the use list must be available for proxy routing to work.

The following example configures proxy forwarders and replicators using an "exclude" list:

```
N7K2-AGG2(config)# hardware proxy layer-3 forwarding exclude interface Ethernet3/1-4, 
Ethernet4/1-4
N7K2-AGG2(config)# hardware proxy layer-3 replication exclude interface Ethernet3/1-4, 
Ethernet4/1-4
```

The following example shows how the same functionality can be achieved using a “use” list:

```
N7K2-AGG2(config)# hardware proxy layer-3 forwarding use interface Ethernet3/5-8,
Ethernet4/5-8
N7K2-AGG2(config)# hardware proxy layer-3 replication use interface Ethernet3/5-8,
Ethernet4/5-8
```

Configuration Verification and Statistics Monitoring

The following **show** command can be used to verify proxy Layer 3 configuration in a mixed chassis:

```
N7K2-AGG2(config)# show hardware proxy layer-3 detail
Global Information:
        F1 Modules:     Count: 2         Slot: 7-8
        M1 Modules:     Count: 2         Slot: 3-4

        Replication Rebalance Mode:             Manual
        Number of proxy layer-3 forwarders:     8
        Number of proxy layer-3 replicators:    4

Forwarder Interfaces                     Status     Reason
------------------------------------------------------------------------------
Eth3/5                                   up         SUCCESS
Eth3/6                                   up         SUCCESS
Eth3/7                                   up         SUCCESS
Eth3/8                                   up         SUCCESS
Eth4/5                                   up         SUCCESS
Eth4/6                                   up         SUCCESS
Eth4/7                                   up         SUCCESS
Eth4/8                                   up         SUCCESS

Replicator Interfaces                    #Interface-Vlan   Interface-Vlan
------------------------------------------------------------------------------
Eth3/5-6                                 51                225-250,627-650,801
Eth3/7-8                               51                1,427-450,601,827-850,1001
Eth4/5-6                                 50                201-224,401,602-626
Eth4/7-8                                 50                402-426,802-826
N7K2-AGG2(config)#
```

The following **show** command displays statistics for the router PortChannel. The command has two arguments. The **detail** argument shows the statistics for each F1 module separately. The **brief** argument shows the aggregate statistics for all F1 modules.

```
N7K2-AGG2(config)# sh hardware proxy layer-3 counters brief
Summary:
Proxy packets sent by all F-series module:
-----------------------------------------------------------------------
Router Interfaces             Tx-Pkts             Tx-Rate (pkts/sec approx.)
-----------------------------------------------------------------------
Eth3/5                                   190555840555        733182
Eth3/6                                   180744907105        593936
Eth3/7                                   183671843810        670341
Eth3/8                                   173977142036        641040
Eth4/5                                   175874031852        682123
Eth4/6                                   174859410868        680121
Eth4/7                                   174294668393        657679
Eth4/8                                   176616676560        687077
==========================================================
Total                                    2675186105338       5345499
==========================================================
N7K2-AGG2(config)#
```

The following command clears the statistics of the internal router PortChannel:

```
N7K2-AGG2(config)# clear hardware proxy layer-3 counters
```

The following syslog messages are generated when the last M1 module is removed in a mixed chassis:

```
2011 May 10 14:21:56.106 N7k1-AGG1 %MCM-5-MCM_REPLICATION_DISABLED: Proxy layer-3 modules
are not available for replication. Layer-3 Multicast Replication is Disabled.

2011 May 10 14:21:56.106 N7k1-AGG1 %MCM-5-MCM_ROUTING_DISABLED: Proxy layer-3 modules are
not available for routing. Routing is Disabled.
```

In the 5.1 release, this syslog message is not printed until the logging level for **proxy layer-3** is set to 5 or higher. The logging facility used with **proxy layer-3** is mcm (Multi Channel Manager). It has a severity of 2 by default, while the syslog message itself has a severity of 5. Therefore, the following configuration is required for the message to be printed:

```
N7k1-AGG1(config)# logging level proxy layer-3 5
```

Note Starting in the 5.2.1 release, the running configuration displays the command as it is.

Starting in the 5.2 release, the syslog message has a severity of 2 and is printed by default when the last M1 module is removed:

```
2011 May 10 12:23:35.482 AGG2-LAN proxy %MCM-2-MCM_REPLICATION_DISABLED: Proxy layer-3
modules are not available for replication. Proxy layer-3 multicast replication is
disabled.

2011 May 10 12:23:35.482 AGG2-LAN proxy %MCM-2-MCM_ROUTING_DISABLED: Proxy layer-3 modules
are not available for routing. Proxy layer-3 forwarding is disabled.
```

Prescriptive Physical Topology and Configured Limits

The topology selection, capacity planning, and function of proxy routing design considerations and recommendation discussed so far in this document optimize the resources (number of M1 and F1 cards and bandwidth capacity planning) in selecting the role of F1/M1 ports. The following proposed configuration does not imply that M1 ports need to be dedicated in all conditions; instead, it is a prescriptive design adopting those discussions and assumptions.

- Redundant M1-XL modules with a total of 16 ports
- 40 G of uplink bandwidth—Four 10 Gbps ports on each aggregation layer device, with each aggregation layer device with one diversified port on each M1 I/O module, PortChannel to two core devices. Hence south-to-north traffic requires four dedicated proxy routing ports. An equal amount of traffic is assumed for east-west (40 Gbps) and therefore requires four dedicated proxy routing ports.
- An equal amount of traffic is assumed for east-west direction (40 Gbps) and therefore requires four dedicated proxy-routing ports.
- Divide proxy routing and uplink forwarding across two M1 modules, using four ports on each module so that module failure equally reduces the capacity.
- 40 G of interconnect traffic for Layer 3 re-route and peer-link traffic. The four ports are divided between two M1 modules.

MAC Address and VLAN Scalability

The four pod topologies were evaluated using 9,000 MAC addresses for each pod with MAC addresses evenly distributed among 50 VLANs per pod. The consistent number of MAC addresses and VLANs enables better comparisons among the topologies and may yield better insight if the topologies behave differently.

Validated Topology and Configuration

The goal of the validation is to assess certain design limits and validate the proposed topology. The goal is not to validate every failure or configuration combination. The validation results depend on many variables, including the software version running on a device. Using a different software version may yield different results even with all other conditions being the same.

Table 2 Software and Hardware Summary

Platform	Software Release	Hardware Configuration	Part Number	Device Role
Cisco Nexus 7010	5.1(3)	Two Supervisors	N7K-SUP1	Aggregation
		Two M1 Cards	N7K-M108X2-12L	
		Two F1 Cards	N7K-F132XP-15	
Cisco Nexus 5010, Cisco Nexus 5020, Cisco Nexus 5548	5.0(3)N1.1a	—	—	Access

Table 3 Layer 2 M1 and F1 Integration Environment

Configuration	Validated Configuration
Peer-link	Diversified on two M1 ports
Layer 2 topology	Virtual PortChannel
ECMP load-sharing	Yes
Layer 3 graceful restart capability configured	Yes
PortChannel load-share	src-dst-ip-port-vlan

Table 4 Layer 3 Domain

Data Center Technology	Validated Data Center Environment	Comments
Routing protocol	OSPF	
NSF awareness in the core	Yes	
OSPF hello and hold timers	Default	10/40
OSPF SPT and LSA Timer	Tuned based on Campus Design	SPF—10-100-5000 LSA—10-100-5000 LSA Arrival—80

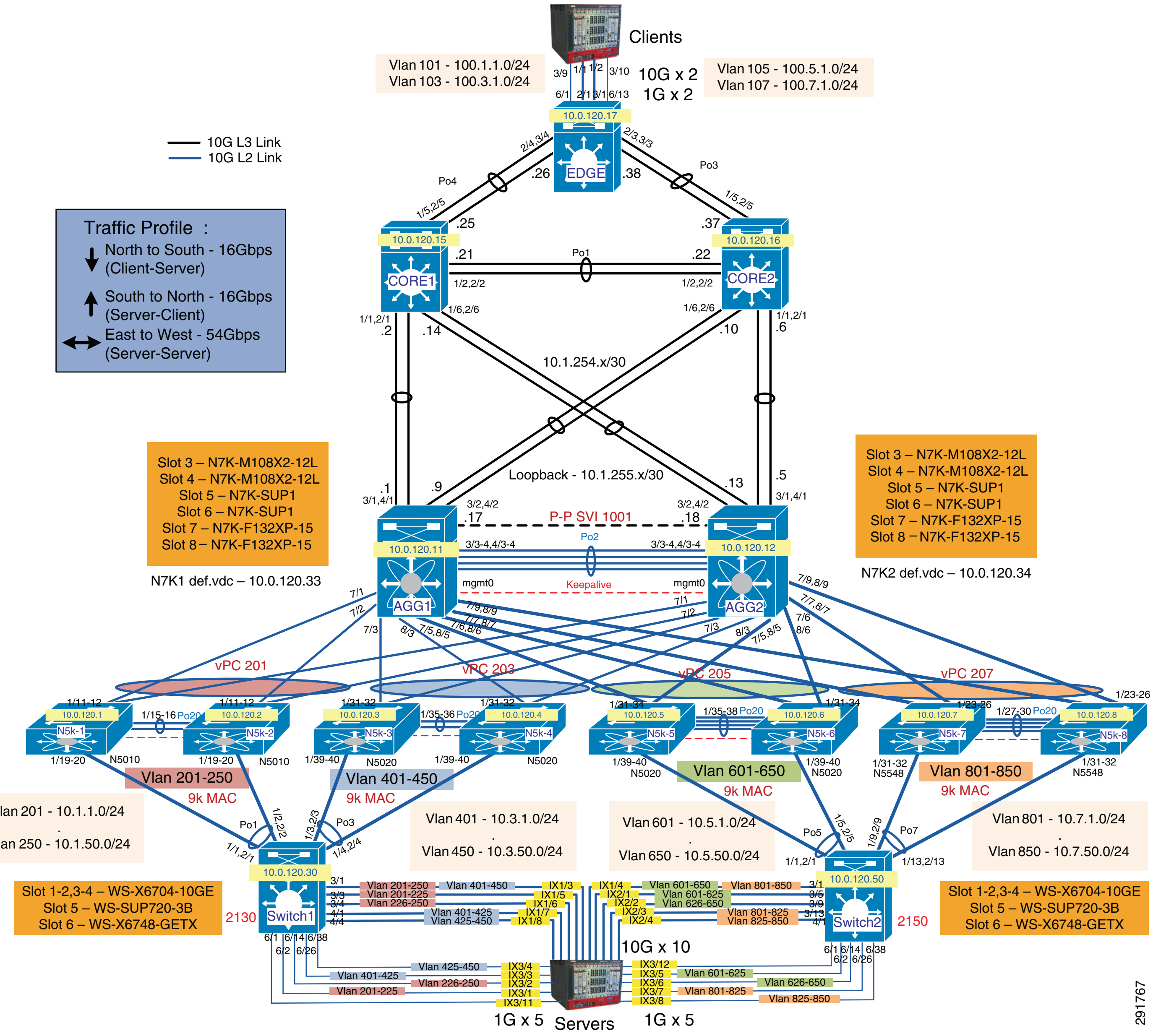
Table 4 Layer 3 Domain

Data Center Technology	Validated Data Center Environment	Comments
Topology	ECMP	
Number of routes	204	
Route summarization	Yes	
CEF load-sharing	Yes	
Core connectivity	Layer 3 Port-channel	
Core devices	Standalone Cisco Catalyst 6500	

Table 5 Layer 2 Domain

Data Center Technology	Validated Data Center Environment	Comments
STP	RPVST+	
Pods	4	
Total VLANs	200	
VLAN spanning	Within pod	Multiple switches
Peer-link	Diversified on two M1 modules	
MAC address per Pod	9,000	
Unique IP application flows per Pod	18,000	
PortChannel mode-LACP	Active-Active	
LACP timers	Defaults	
UDLD mode	Normal	

Figure 13 End-to-End Topology[1]



The following describes the topology configuration illustrated in Figure 13:

- The aggregation layer is defined using Cisco Nexus 7000 F1 modules for Layer 2 access layer connectivity, M1 modules for Layer 3 connectivity, and vPC peer-link connectivity between aggregation layer devices.
- The access layer uses three pairs of Cisco Nexus 5000 series switches and one pair of Nexus 5500 series switches, with each pair defined as a separate Layer 2 MAC domain or pod.
- Loop-free topology using virtual PortChannel architecture; specific configurations include:
  - Dual-sided vPC with Cisco Nexus 5000 and Cisco Nexus 7000.

[1] IXIA® is a registered trademark.

  - Enable peer switch capability at Cisco Nexus 7000 for better convergence and streamlining of control plane.
  - Use vPC delay restores to enable better Layer 3 convergence.
  - Enable reload restore to enhance the vPC behavior when one of the vPC peers comes back up after a data center outage.
  - ARP Syncing between aggregation devices.
- vPC peer-link on aggregation layer must be on M1 modules to achieve greater than 16,000 MAC address scalability.
- The pod topology connectivity combination is chosen to show the impact and design choices when connected to the access layer with F1 modules while supporting more than 16,000 MAC address per module.

For vPC-related design guidance, see: http://www.cisco.com/en/US/docs/switches/datacenter/sw/5_x/nx-os/interfaces/configuration/guide/if_vPC.html.

Summary

Resource allocation of M1 and F1 modules is key to an extensible design. When integrating F1 modules in a classical Ethernet architecture, the key design recommendations are:

- The number of M1 modules required depends on the routing requirements; both inter-VLAN and routed (VLAN to routed M1 interface) traffic require proxy routing if the packet enters the switch on a F1 interface.
- The M1 front panel port can be used for proxy routing and uplink connectivity; however, the available bandwidth on the M1 port will be shared between proxy routing traffic and egress traffic on the front panel port.
- Use 8-port M1-XL modules when there is a need for higher Layer 3 throughput and better performance.
- The 8-port M1-XL module has two forwarding engines. Depending on the design, one or both forwarding engines can be dedicated for proxy Layer 3 processing.
- Use multiple Layer 2 domains to extend the F1 module beyond 16,000 MAC addresses.
- Enable vPC peer-link on M1 module ports when vPC topologies require greater than 16,000 MAC addresses per aggregation layer device. Mixing M1 and F1 module ports for the peer-links is not supported.
- Deploy a minimum of two M1 modules for proper redundancy for proxy routing, vPC peer-link, and uplink to the core devices.
- The recommended pod topology is where access layer uplinks are diversified on two F1 modules to improve the resilience and convergence during line card failures.

References

- M1 Module Data Sheets
  http://www.cisco.com/en/US/products/ps9402/products_data_sheets_list.html

- Transceiver Compatibility Matrix http://www.cisco.com/en/US/partner/docs/interfaces_modules/transceiver_modules/compatibility/matrix/OL_6974.html
- Cisco Nexus 7000 Family: http://www.cisco.com/en/US/products/ps9402/index.html
- Cisco Design Zone: http://www.cisco.com/en/US/netsol/ns742/networking_solutions_program_category_home.html