

QFX10000 Switches

System Architecture

Table of Contents

Executive Summary	3
Introduction: QFX10000 Line of Fixed and Modular Switches.....	3
QFX10002-72Q and QFX10002-36Q Fixed Ethernet Switches	3
QFX10008 Modular Ethernet Switch	3
QFX10016 Modular Ethernet Switch.....	4
QFX10000 Modular Switch Line Cards	4
Innovations in Chassis Design.....	5
Control Plane.....	5
The Q5 ASIC	6
Hybrid Memory Cube.....	6
System Design	8
Cell-Based Forwarding.....	9
Virtual Output Queueing	9
System Scale.....	10
Software Architecture.....	11
Data Center Architectures.....	11
Multichassis LAG	12
Junos Fusion	12
IP Fabric.....	13
High Scale Network Virtualization with Overlays.....	13
MPLS	14
Conclusion.....	14
About Juniper Networks.....	15

List of Figures

Figure 1: QFX10002-72Q front and rear view	3
Figure 2: QFX10002-36Q front and rear view.....	3
Figure 3: QFX10008 front and rear view.....	3
Figure 4: QFX10016 front and rear view	4
Figure 5: Midplane-less design of the QFX10000 switch	5
Figure 6: Routing Engine on the modular QFX10000 platform.....	6
Figure 7: The Juniper Q5 ASIC	6
Figure 8: Hybrid Memory Cube	6
Figure 9: 1 Tbps forwarding ASIC with 4 GB external memory using DDR4 technology.....	7
Figure 10: Juniper Q5 1 Tbps forwarding ASIC with 4 GB external memory using HMC technology	7
Figure 11: System architecture of the QFX10002-72Q switch.....	8
Figure 12: System architecture of the QFX10002-36Q switch.....	8
Figure 13: System architecture of the QFX10000-36Q I/O card for modular QFX10000 switches.....	8
Figure 14: Cell-based forwarding in QFX10000 switches resulting in over 95% system utilization	9
Figure 15: Virtual Output Queues on the QFX10002-72Q system	10
Figure 16: Modular software architecture of QFX10000 switches	11
Figure 17: Diverse data center architectures with Juniper QFX10000 switches	11
Figure 18: MC-LAG architecture with QFX5100 in data center access and QFX10000 in aggregation role	12
Figure 19: Junos Fusion technology with QFX5100/EX4300 as satellites and QFX10000 in aggregation role	12
Figure 20: IP fabric with QFX10000 switches as spines and QFX5100 switches as leafs.....	13
Figure 21: Overlay networking examples with Juniper Contrail and VMware NSX	14

Executive Summary

The Juniper Networks® QFX10000 line of switches is a series of fixed and modular high-performance and low-latency platforms purpose-built for data center spine applications, as well as data center core and Data Center Interconnect (DCI) deployments. QFX10000 switches provide high-density 10GbE, 40GbE, and 100GbE aggregation options with migration path to 400GbE. Running the carrier-grade Juniper Networks Junos® operating system, the QFX10000 switches provide a flexible building block for supporting a number of different data center architectures. The QFX10000 switches not only offer flexible fixed and modular form factors, they also include rich, flexible software capabilities that allow them to form a solid foundation for everything from traditional data centers to the largest private and public cloud data centers. The Junos OS running on the QFX10000 switches has also been enhanced for greater scalability, modularity, and programmability, ensuring that these high-performance switches deliver an unparalleled pace of innovation.

Introduction: QFX10000 Line of Fixed and Modular Switches

The QFX10000 line of switches features a number of fixed and modular platforms, including the QFX10002-72Q, the QFX10002-36Q, the QFX10008, and the QFX10016.

QFX10002-72Q and QFX10002-36Q Fixed Ethernet Switches



Figure 1: QFX10002-72Q front and rear view

The QFX10002-72Q is a 2 U fixed form factor switch that provides scalability on multiple vectors. The switch offers 72 wire-speed 40GbE ports, providing 5.76 Tbps of throughput; each port can also be converted into 4X10GbE ports, giving customers as much as 288 ports of 10GbE density in a 2 U platform. Alternatively, every third port can be used as a 100GbE port, giving customers as many as 24 100GbE ports in a compact form factor. Customers can use the QFX10002-72Q as a 10GbE aggregation switch with a path to varying densities of 40GbE and 100GbE aggregation. Highly reliable and resilient, the QFX10002-72Q features redundant power supplies that can provide feed as well as power supply redundancy. The switch also offers multiple fan trays to provide resiliency in the event of a fan failure. Front-to-back cooling means the switch is optimized for spine, core, and edge applications in the data center.

The QFX10002-36Q fixed-configuration switch provides half the 10GbE, 40GbE, and 100GbE port densities of the QFX10002-72Q.



Figure 2: QFX10002-36Q front and rear view

QFX10008 Modular Ethernet Switch

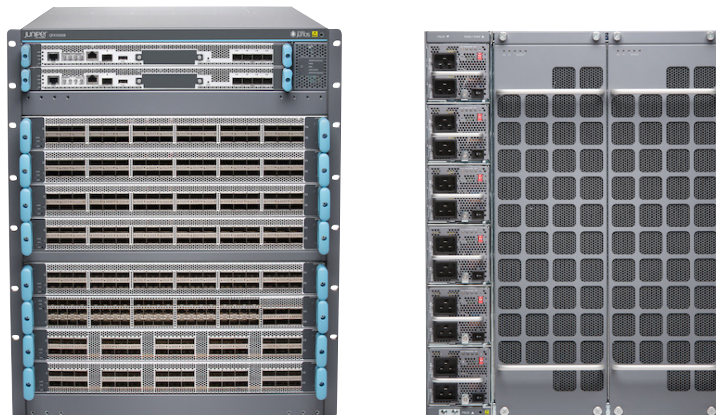


Figure 3: QFX10008 front and rear view

The eight-slot QFX10008 switch is the industry's highest density modular platform for its size and form factor. It supports up to 240 wire-speed 100GbE, 288 wire-speed 40GbE ports, and 1,152 wire-speed 10GbE ports; the crossbar switch fabric is designed to provide up to 6 Tbps performance per slot for a total 48 Tbps capacity in a 13 U form factor. The QFX10008 features a resilient design with redundant power supplies, Routing Engines (REs), fan trays, and fabrics. The QFX10008 switch can accommodate up to six power supplies and requires all six when fully loaded, providing full feed redundancy due to its dual input power design. This switch also offers redundant fan trays and holds up to six switch fabric cards with 5+1 redundancy.

QFX10016 Modular Ethernet Switch

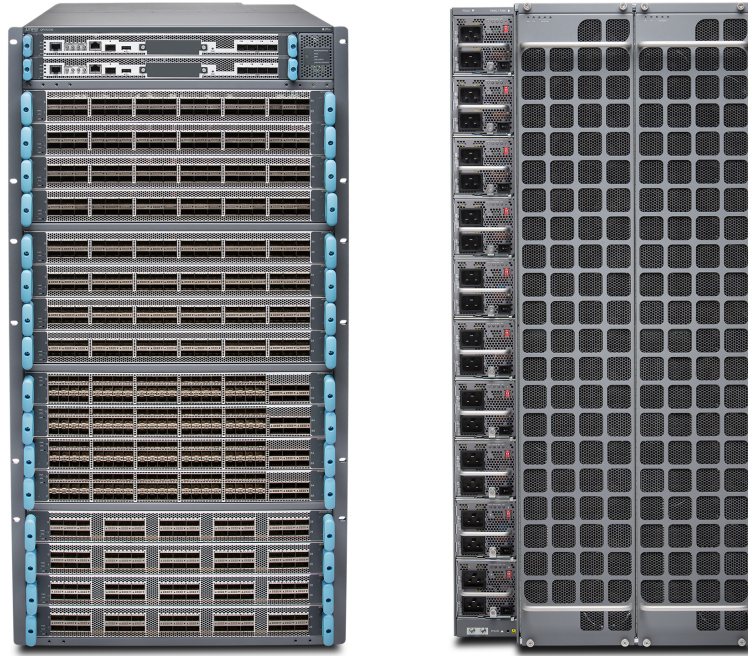


Figure 4: QFX10016 front and rear view

The 16-slot QFX10016 breaks new ground in the networking industry with unprecedented performance and scale. It is the industry's only platform to support 480 nonblocking 100GbE ports in a single chassis, as well as double the 40GbE and 10GbE density of the QFX10008, all in a 21 U form factor. The QFX10016 can accommodate up to 10 power supplies and requires all 10 when fully loaded, providing full feed redundancy due to its dual input power design. The switch also offers redundant fan trays and holds up to six switch fabric cards with 5+1 redundancy. The crossbar switch fabric is designed to provide up to 6 Tbps of bandwidth per slot.

QFX10000 Modular Switch Line Cards

The modular QFX10000 switches support three different line cards, each optimized for 10GbE, 40GbE, and 100GbE densities.

- **36x40GbE I/O Line Card:** The 36x40GbE I/O line card provides 36 ports of 40GbE, 12 ports of 100GbE, or, using breakout cables, 144 ports of 10GbE, offering unprecedented flexibility with multiple connectivity options of varying densities.



- **30x100GbE I/O Line Card:** The 30x100GbE I/O module provides 30 ports of 100GbE or a combination of 24 ports of 40GbE and six ports of 100GbE. The 30x100GbE is the industry's highest density 100GbE module—more than twice the number of existing line cards.



- **60x10GbE+6x40GbE I/O Line Card:** The 60x10GbE line card provides not only 60 ports of 10GbE, but also offers six uplink ports that support an additional 6x40GbE, 2x100GbE, or 24x10GbE ports.



Innovations in Chassis Design

The modular QFX10000 switches have been designed to provide maximum power and cooling efficiencies as well as easy migration to the next level of performance without having to upgrade the chassis to accommodate multi-generational I/O card upgrades. The modular QFX10000 switches accomplish this with a midplane-less design that allows direct connection between the line cards and fabric cards. This design significantly reduces overall system power consumption and optimizes cooling and airflow efficiency through the system while providing an easy upgrade path to next-generation interconnect technology. The midplane-less design also delivers a higher mean time between repairs and failures, as well as a more robust and resilient overall design.



Figure 5: Midplane-less design of the QFX10000 switch

Control Plane

QFX10000 fixed and modular platforms use the same control plane; the only difference between the two is that, while the modular switches have two Routing Engines, the fixed platforms have only a single RE with multiple CPU cores. In addition, an innovative software architecture brings virtualization to the control plane, enabling features such as topology-independent in-service-software upgrades (TISSU) to be offered on both platforms. The RE is based on a quad-core Intel Ivy Bridge X86 processor. In addition to providing interfaces for management connectivity, dedicated ports provide connectivity to a time source over an Ethernet interface.



Figure 6: Routing Engine on the modular QFX10000 platform

The Q5 ASIC



Figure 7: The Juniper Q5 ASIC

QFX10000 switches are powered by the Juniper Q5 forwarding ASIC, which provides up to 1 Tbps of forwarding capacity (12x40GbE, 5x100GbE, or 48x10GbE ports) along with unprecedented logical scale. This unique blend of high I/O capacity and high logical scale—never before achieved in the industry—is what truly differentiates the QFX10000 switches, enabling them to solve a variety of data center challenges requiring architectural flexibility as customers move from more traditional data center designs to multitenant cloud networks.

The QFX10000 fixed platforms use multiple Q5 chips (six on the QFX10002-72Q and three on the QFX10002-36Q), all of which are connected using a variable length cell-based fabric for maximum performance and system utilization. Similarly, the I/O cards on the modular QFX10000 switches have different numbers of forwarding ASICs, depending on the type of line card. The Q5 chip can connect to a dense external memory that is uniquely designed to keep pace with a high-scale 1 Tbps system. This external memory allows the system to provide packet buffering as well as augment the chip's logical scale to support a variety of tables such as forwarding information base (FIB), media access control (MAC), and MPLS, providing unprecedented flexibility in terms of data center architectures.

Typically, the slow speeds of external memories are gating factors when designing these systems. Traditionally, DDR3 and DDR4 memories have been used in networking systems, limiting the I/O bandwidth to about 200 Gbps. The memory technology, called Hybrid Memory Cube used alongside the Juniper Q5 ASIC in the QFX10000 switches, is the first external memory to be used in an ASIC that can scale to 1 Tbps of I/O bandwidth.

Hybrid Memory Cube

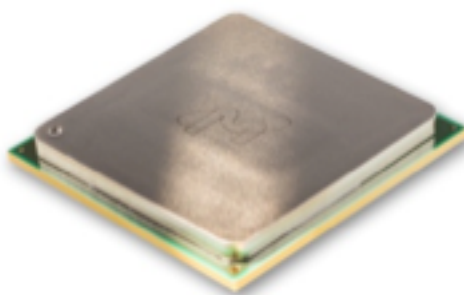


Figure 8: Hybrid Memory Cube

Hybrid Memory Cube (HMC) is a new class of three-dimensional memory, purpose-built for massive scale data center systems. Previously, it was simply not possible to design systems that could scale on multiple vectors, including I/O bandwidth and logical system scale. Due to the limited memory technology, only one of the system’s dimensions could be scaled, either I/O or the memory reserved for packet buffering and FIB tables. For systems using external memory, I/O bandwidth would typically be limited to about 200 Gbps. If customers wanted to scale I/O bandwidth, they would have to sacrifice packet memory as well as FIB and MAC lookup capacity. The combination of Juniper’s Q5 ASIC and the HMC technology breaks that barrier for the first time, enabling both I/O and packet buffer and lookup memory to be scaled at the same time.

When used instead of DDR4, HMC provides 17% greater power efficiency, an 84% improvement in system board design, and scalability up to 1 Tbps of ASIC performance. It would require 45 DDR4 memories to match the memory of a single HMC for a system of this scale. The following figures draw a comparison between two ASICs, one designed with DDR4 memory (Figure 9) and the other with HMC technology (Figure 10).

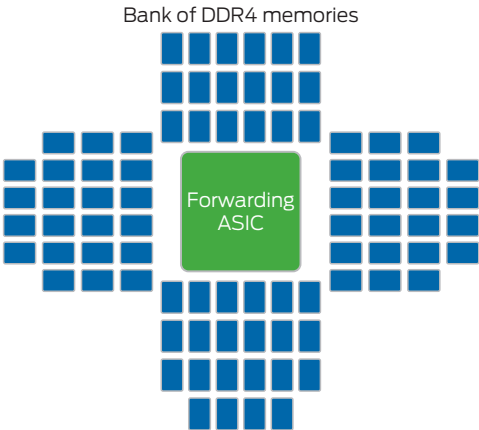


Figure 9: 1 Tbps forwarding ASIC with 4 GB external memory using DDR4 technology



Figure 10: Juniper Q5 1 Tbps forwarding ASIC with 4 GB external memory using HMC technology

The following table provides a comparison between a chip designed with DDR4 and HMC memories (4 GB) for a 1 Tbps forwarding chip.

Table 1: DDR vs. HMC

	DDR4	Hybrid Memory Cube (HMC)
Number of memory devices	90	2
Total number of pins between ASIC and memory	2,400	422
Power	61 W	49 W
Memory surface area	12,750 mm ²	1,922 mm ²

When coupled with the Q5 ASIC in the QFX10000, HMC provides the following functions:

- Packet buffering and virtual queue memory
- Improved logical system scale by augmenting local tables such as FIB, MPLS, and MAC on the Q5 ASIC, providing significant improvements and architectural flexibility

System Design

As discussed earlier, QFX10000 switches are multichip systems in which all chips are connected via a variable size cell-based fabric, providing the most efficient system utilization. The following diagram shows the internal design of the QFX10002-72Q and QFX10002-36Q systems.

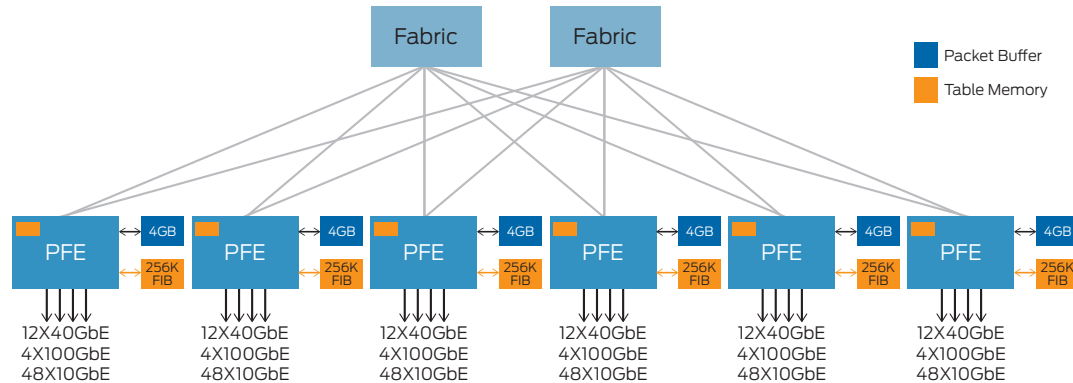


Figure 11: System architecture of the QFX10002-72Q switch

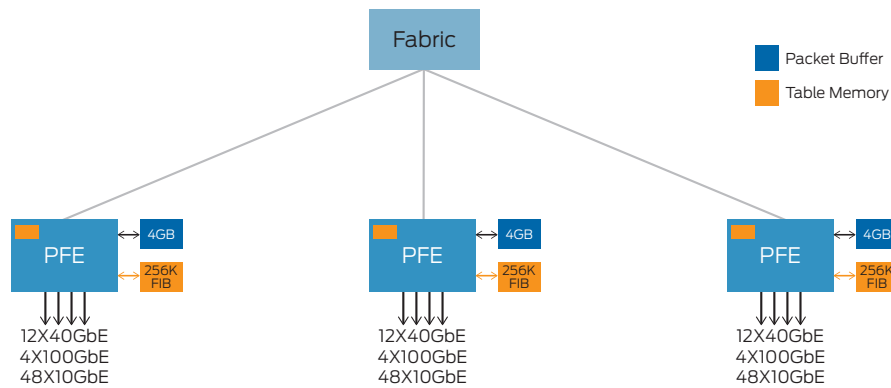


Figure 12: System architecture of the QFX10002-36Q switch

Just as with the fixed systems, the I/O modules in the modular QFX10000 switches are also multichip systems, providing I/O scale without sacrificing system logical scale. The following diagram shows the internal architecture of the QFX10000 36x40GbE I/O card.

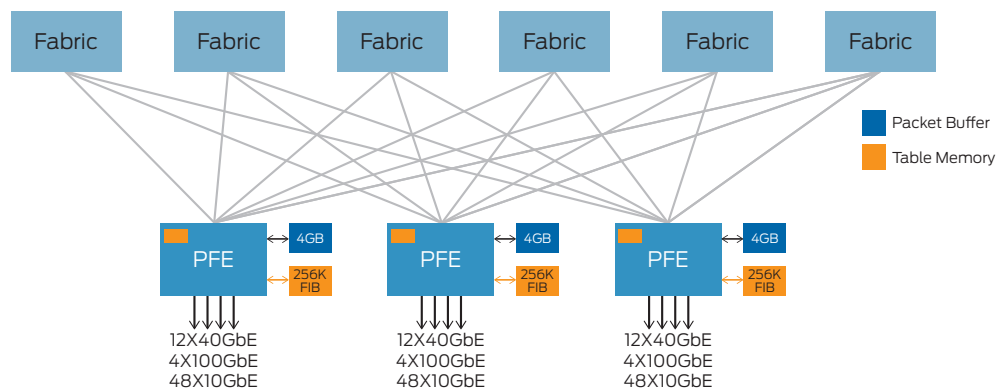


Figure 13: System architecture of the QFX10000-36Q I/O card for modular QFX10000 switches

Cell-Based Forwarding

A unique attribute of QFX10000 switches is their use of cell-based forwarding between Packet Forwarding Engines (PFEs) and line cards. When a packet arrives on the input port in the system and is transmitted over the switch fabric to the egress port, instead of sending an entire packet over the fabric, it is broken into smaller sized cells, which are then forwarded across all available fabric links. The benefits of a cell-based system versus sending an entire packet over the fabric are twofold:

- The fabric utilization in packet-based fabric systems peaks at about 65-70% utilization due to the unpredictable nature and size of packets. For example, a 9 KB packet might hash to one link of the fabric while a 64 byte packet could hash to another link, resulting in uneven system utilization.
- By reducing each packet to small, variable length cells (chosen based on the input packet size, which can be between 96 and 176 bytes), each packet can then be sprayed evenly across all fabric links, resulting in close to 95% effective system utilization.

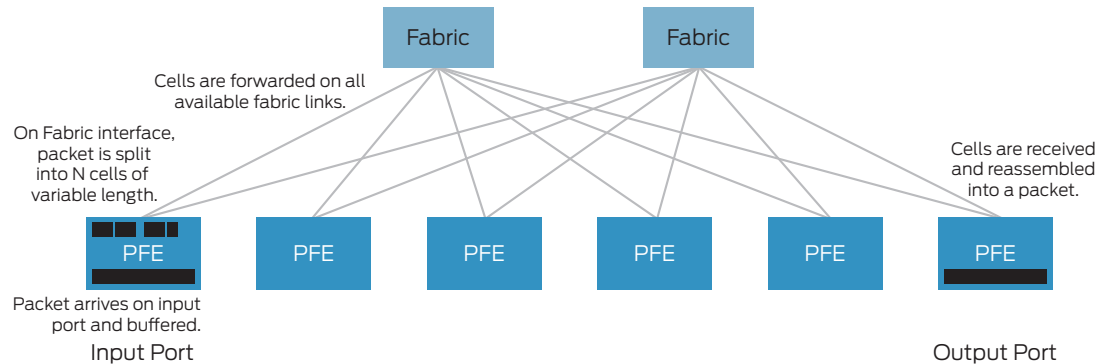


Figure 14: Cell-based forwarding in QFX10000 switches resulting in over 95% system utilization

Virtual Output Queueing

QFX10000 switches provide virtual output queueing. Virtual Output Queue (VOQ) systems differ from traditional egress transmit queue-based systems in two ways. In an egress transmit queue-based system, all transmit queues are maintained on the egress forwarding engine. This generally has two side effects: First, the system buffers packets at ingress and egress; second, it can result in head-of-line blocking if the egress port is congested.

In a head-of-line blocking scenario, an egress congested port (p2) would typically send a backpressure signal to the input port (p1) that is sending the traffic, causing port p1 to buffer traffic until port p2 is ready. If port p1 receives traffic that needs to be sent to another port (say, port p3) at the same time, it won't be able to send it until the data it is holding for port p2 is cleared out. As a result, a completely unrelated (and uncongested) port (p3) will end up having to wait for traffic from p1.

VOQ solves this problem by creating dedicated virtual queues between each input port and each output port. With this approach, no communication between any two ports can affect a third port in the system. Additionally, all queues are kept on ingress to avoid two buffering points in the system, as well as to improve the efficient utilization of the fabric. By queuing the ports on the ingress, the system makes sure not to send traffic over the fabric to egress ports that are not ready to receive that traffic.

In the QFX10000 switches, each ingress PFE maintains the VOQs for all egress ports—a total of eight between each ingress and egress port. Figure 15 shows the VOQ distribution on the QFX10002-72Q switch.

When you have applications that are unresponsive to network congestion events by flow controlling themselves, it helps to have a system that can provide appropriate levels of buffering during the time window when application traffic may be bursty.

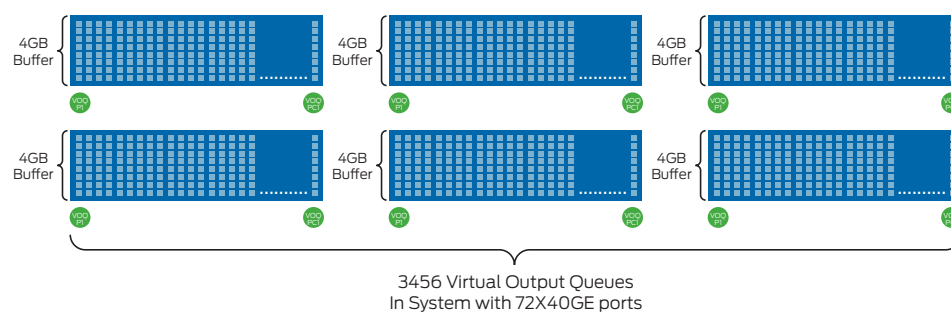


Figure 15: Virtual Output Queues on the QFX10002-72Q system

System Scale

As discussed earlier, QFX10000 switches scale on multiple vectors, including I/O as well as logical and packet buffering, making them a unique fit for diverse data center architectures ranging from spine roles all the way to edge and DCI without sacrificing port density.

The following tables provide more information on QFX10000 scaling.

Table 2: QFX10000 System Scale

	QFX10002-36Q	QFX10002-72Q	QFX10008	QFX10016
System throughput	2.88 Tbps	5.76 Tbps	48 Tbps	96 Tbps
Forwarding capacity	1 Bpps	2 Bpps	16 Bpps	32 Bpps
10GbE density	144	288	1,152	2,304
40GbE density	36	72	288	576
100GbE density	12	24	240	480

Table 3: QFX10000 Logical/Feature Scale

	QFX10002-36Q	QFX10002-72Q	QFX10008	QFX10016
MAC	256,000	512,000	1 million	1 million
ARP	256,000	256,000	256,000	256,000
Jumbo frames	9,216 bytes	9,216 bytes	9,216 bytes	9,216 bytes
VLANs	16,000	16,000	16,000	16,000
IPv4 FIB	256,000	256,000	256,000	256,000
IPv6 FIB	256,000	256,000	256,000	256,000
Host scale	2 million	2 million	2 million	2 million
Equal-cost multipath (ECMP)	64	64	64	64
Multicast groups	128,000	128,000	128,000	128,000
Filters	8,000	8,000	8,000	8,000
Filter terms	64,000	64,000	64,000	64,000
Policers	8,000	8,000	8,000	8,000
LAGs	144	288	1,000	1,000
Mirroring sessions	48	48	48	48
Generic routing encapsulation (GRE)	4,000	4,000	4,000	4,000
L3VPN	4,000	4,000	4,000	4,000
VXLAN (VNID scale)	32,000	32,000	32,000	32,000
Lossless queues/port	6	6	6	6

Software Architecture

QFX10000 switches not only feature many hardware innovations for data center-class switching, they also provide a robust, open, and flexible software platform.

- The QFX10000 is designed to make network management and provisioning simple and automated, with support for a highly diverse set of automation and orchestration tools including Junos Space Network Director as well as industry tools such as Puppet, Chef, Ansible, and OpenStack.
- Integration with Juniper's Cloud Analytics Engine automates network troubleshooting and analysis to reduce operational expenses and maximize efficiency of the network.

The following graphic shows the software architecture of QFX10000 switches, which has been optimized for data centers.

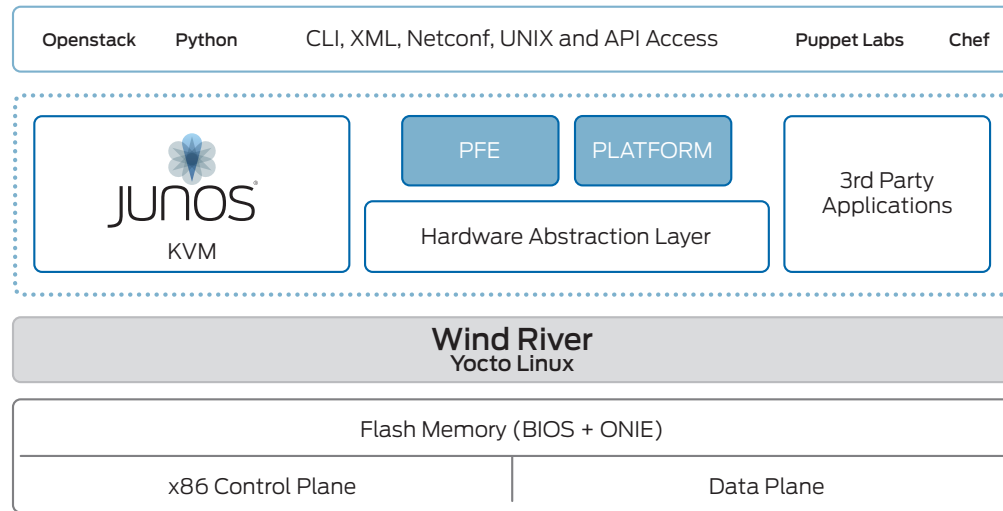


Figure 16: Modular software architecture of QFX10000 switches

The modular software architecture of QFX10000 switches provides the following specific benefits:

- Multicore Linux kernel (based on Windriver Yocto Linux)
- Higher control plane performance (running on four cores)
- Topology-independent in-service-software upgrades (TISSU)
- Hosting third-party apps in a virtual machine
- Zero touch provisioning (ZTP)
- Automation with Puppet, Chef, Ansible, and Python

Data Center Architectures

QFX10000 switches are flexible and scalable building blocks that can be used in a variety of data center architectures. Customers across different verticals have different requirements when it comes to how data center networks are built, provisioned, and operated, depending on the scale of the network, the applications, and whether the data center is purpose-built for a specific application or is a multitenant cloud (private or public) data center. Most customers fall into the following four categories for data center environments; Juniper can support all four by simply configuring different software features without having to use different hardware across each architecture.

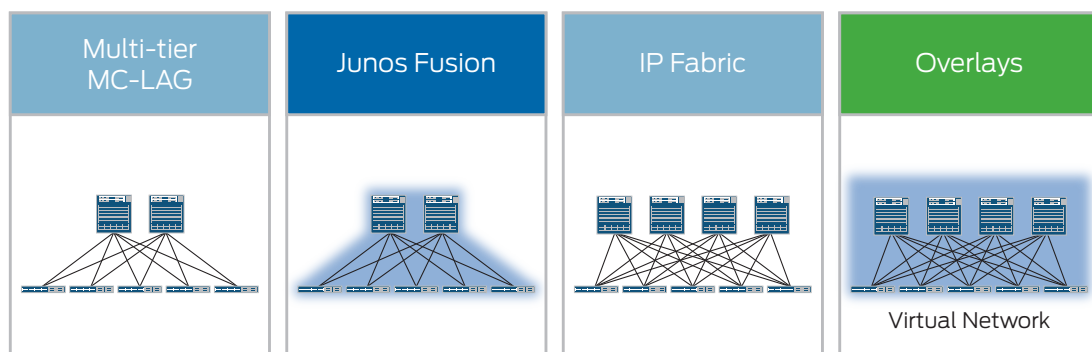


Figure 17: Diverse data center architectures with Juniper QFX10000 switches

Multichassis LAG

QFX10000 switches support the multichassis link aggregation (MC-LAG) protocol, which helps customers build L2 multipath data centers while eliminating Spanning Tree Protocol, maintaining full bandwidth utilization, and enabling virtualized applications that require ample east-west bandwidth and workload mobility. By combining QFX10000 spine or aggregation switches with Juniper Networks QFX5100 leaf or access switches, customers can build large data centers with L2 multipathing. Additionally, since QFX10000 switches support DCI protocols such as GRE, MPLS L3VPN, and Ethernet VPN, customers can use the same building blocks for several tiers within the data center (spine, core, and DCI) or collapse them into a single logical device.

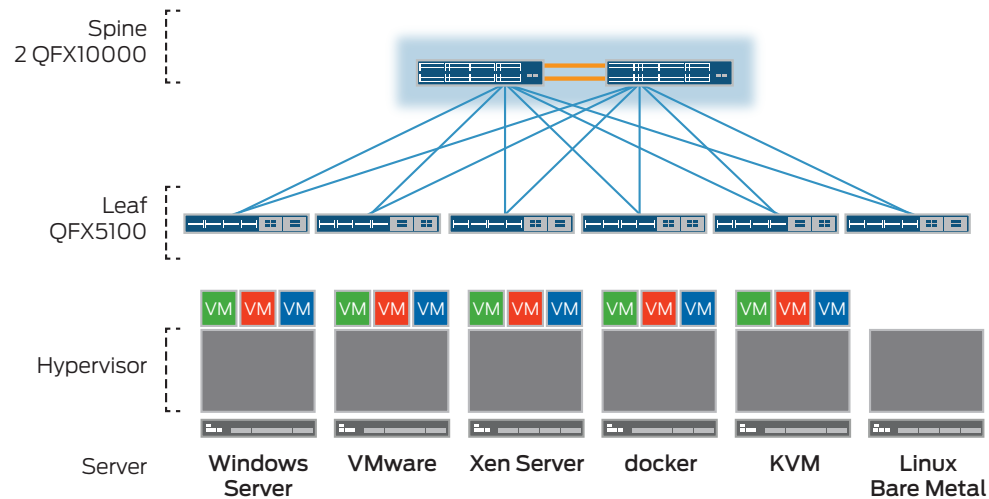


Figure 18: MC-LAG architecture with QFX5100 in data center access and QFX10000 in aggregation role

Junos Fusion

Junos Fusion is a groundbreaking data center technology from Juniper that provides simplicity of operations at scale, workload mobility, and any type of access in the data center from 100 Mbps to 40GbE. The Junos Fusion technology is comprised of aggregation devices (QFX10000 switches) and satellite devices (EX4300 switches or QFX5100 switches, or both). In a Junos Fusion deployment, satellite devices are managed by a single or a pair of aggregation devices, simplifying operations by enabling users to centrally manage the entire data center through the aggregation devices. The Junos Fusion technology is very resilient, enabling aggregation devices to be loosely coupled while having active control planes and running different operating systems.

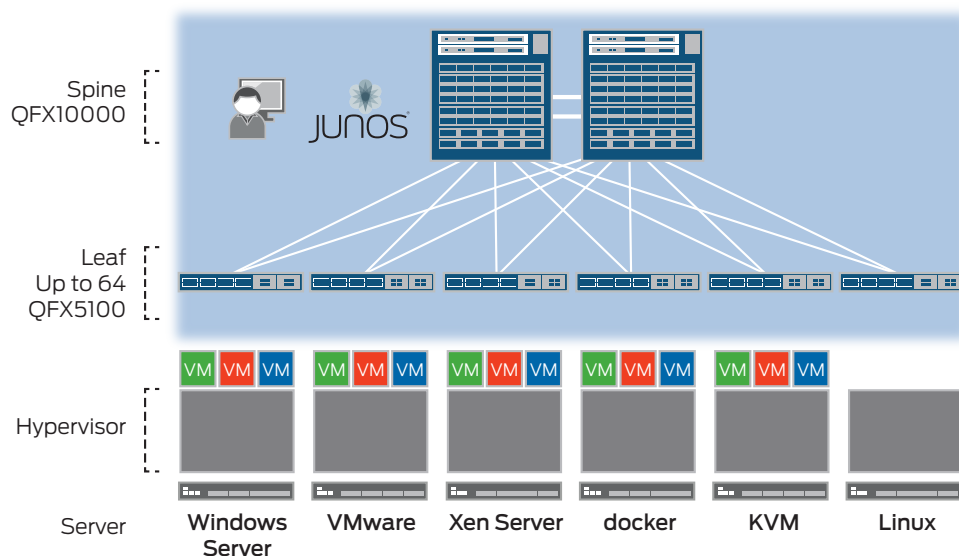


Figure 19: Junos Fusion technology with QFX5100/EX4300 as satellites and QFX10000 in aggregation role

IP Fabric

For hyper-scale data centers, two factors are very important. First, the network must be able to scale tens of thousands of ports; second, applications are not tied to the underlying network topology and do not require Layer 2 segments within the data centers. Each rack represents a subnet, and a routing protocol such as BGP is used to share endpoint reachability information by routing a prefix to other switches within the data center. In effect, every switch acts as a BGP (or OSPF) router within the data center. QFX5100 and QFX10000 switches running Junos OS support various routing protocols including BGP, IS-IS, and OSPF to build a Layer 3 data center. Furthermore, mechanisms such as OpenClos and ZTP automate the deployment of Layer 3 fabrics.

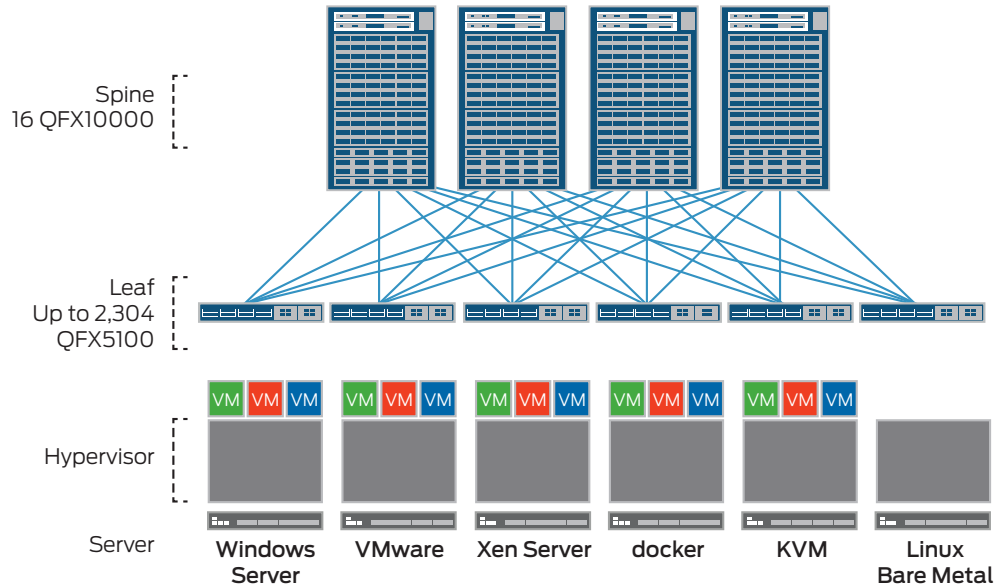


Figure 20: IP fabric with QFX10000 switches as spines and QFX5100 switches as leafs

High Scale Network Virtualization with Overlays

While IP fabrics have tremendous benefits when it comes to scale, the fact that applications must be cloud-enabled or topology-agnostic present distinct challenges. Many applications today are not topology agnostic and require any Layer 2 segment to be available anywhere in the network. By using network-based overlays such as Virtual Extensible LAN (VXLAN), coupled with a BGP control plane protocol of Ethernet Virtual Private Network (EVPN), customers can get the benefits of a scale-out IP fabric as well as provide L2 connectivity within the data center using an L2-over-L3 overlay such as VXLAN. QFX5100 and QFX10000 switches will support EVPN-VXLAN overlays, providing the benefits of scale-out networks for cloud-ready as well as enterprise applications. In addition to supporting network-based overlays, QFX Series switches also work with SDN controllers such as Juniper Contrail and VMware NSX using Open vSwitch Database (OVSDB) protocol.

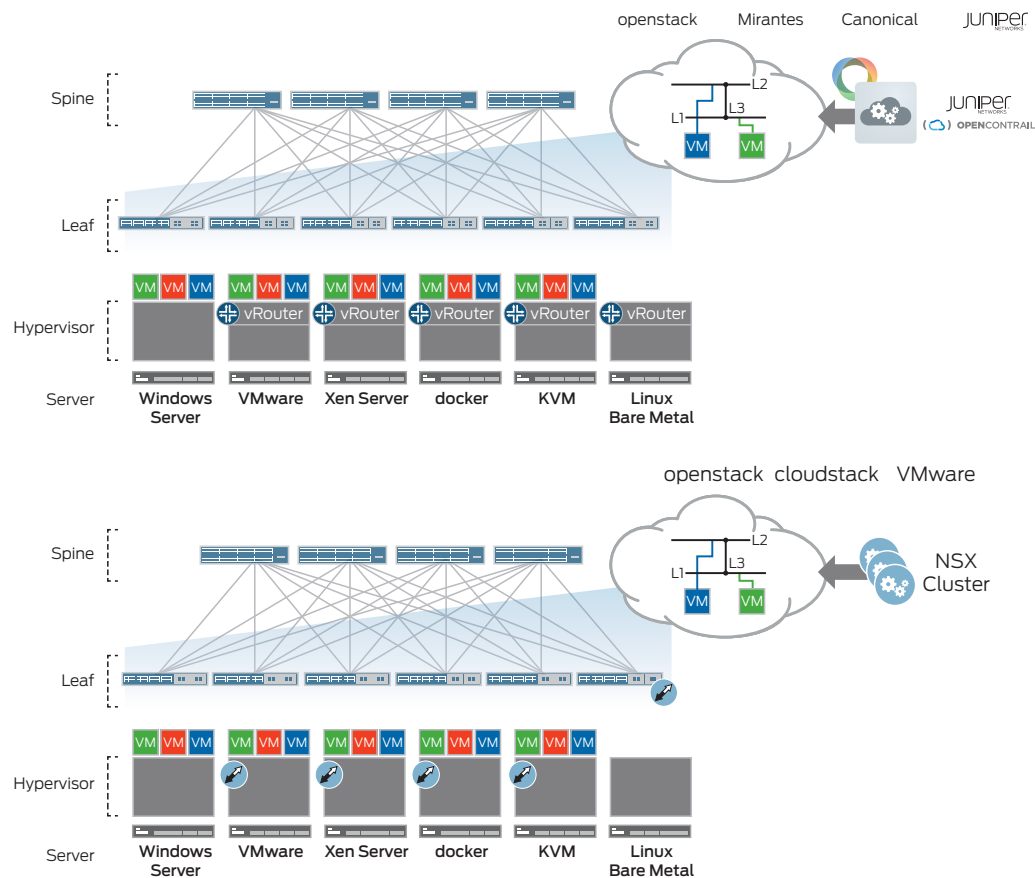


Figure 21: Overlay networking examples with Juniper Contrail and VMware NSX

MPLS

MPLS has many applications both inside and outside the data center. Traditionally, MPLS has been used to provide inter-data center connectivity using services such as L3VPNs, pseudowires, and virtual private LAN service (VPLS). However, new applications of MPLS are emerging in carrier service providers and multitenant large scale cloud data centers. QFX10000 switches provide comprehensive MPLS feature support for both DCI applications as well as new and emerging applications inside data centers.

Conclusion

Juniper Networks QFX10000 switches are designed to provide the foundation for next-generation data centers. This line of high-performance switches delivers industry-leading scalability, density, and flexibility, helping cloud and data center operators build automated data center networks that provide superior long-term investment protection. Together with the openness and automation capabilities present in the Junos operating system, the QFX10000 switches enable Juniper customers to scale their clouds and data centers as fast as their businesses grow, while extracting maximum value from their network infrastructure well into the future.

About Juniper Networks

Juniper Networks is in the business of network innovation. From devices to data centers, from consumers to cloud providers, Juniper Networks delivers the software, silicon and systems that transform the experience and economics of networking. The company serves customers and partners worldwide. Additional information can be found at www.juniper.net.

Corporate and Sales Headquarters

Juniper Networks, Inc.
1133 Innovation Way
Sunnyvale, CA 94089 USA
Phone: 888.JUNIPER (888.586.4737)
or +1.408.745.2000
Fax: +1.408.745.2100
www.juniper.net

APAC and EMEA Headquarters

Juniper Networks International B.V.
Boeing Avenue 240
1119 PZ Schiphol-Rijk
Amsterdam, The Netherlands
Phone: +31.0.207.125.700
Fax: +31.0.207.125.701

Copyright 2015 Juniper Networks, Inc. All rights reserved. Juniper Networks, the Juniper Networks logo, Junos and QFabric are registered trademarks of Juniper Networks, Inc. in the United States and other countries. All other trademarks, service marks, registered marks, or registered service marks are the property of their respective owners. Juniper Networks assumes no responsibility for any inaccuracies in this document. Juniper Networks reserves the right to change, modify, transfer, or otherwise revise this publication without notice.

