Arista 7800R3 Platform Architecture

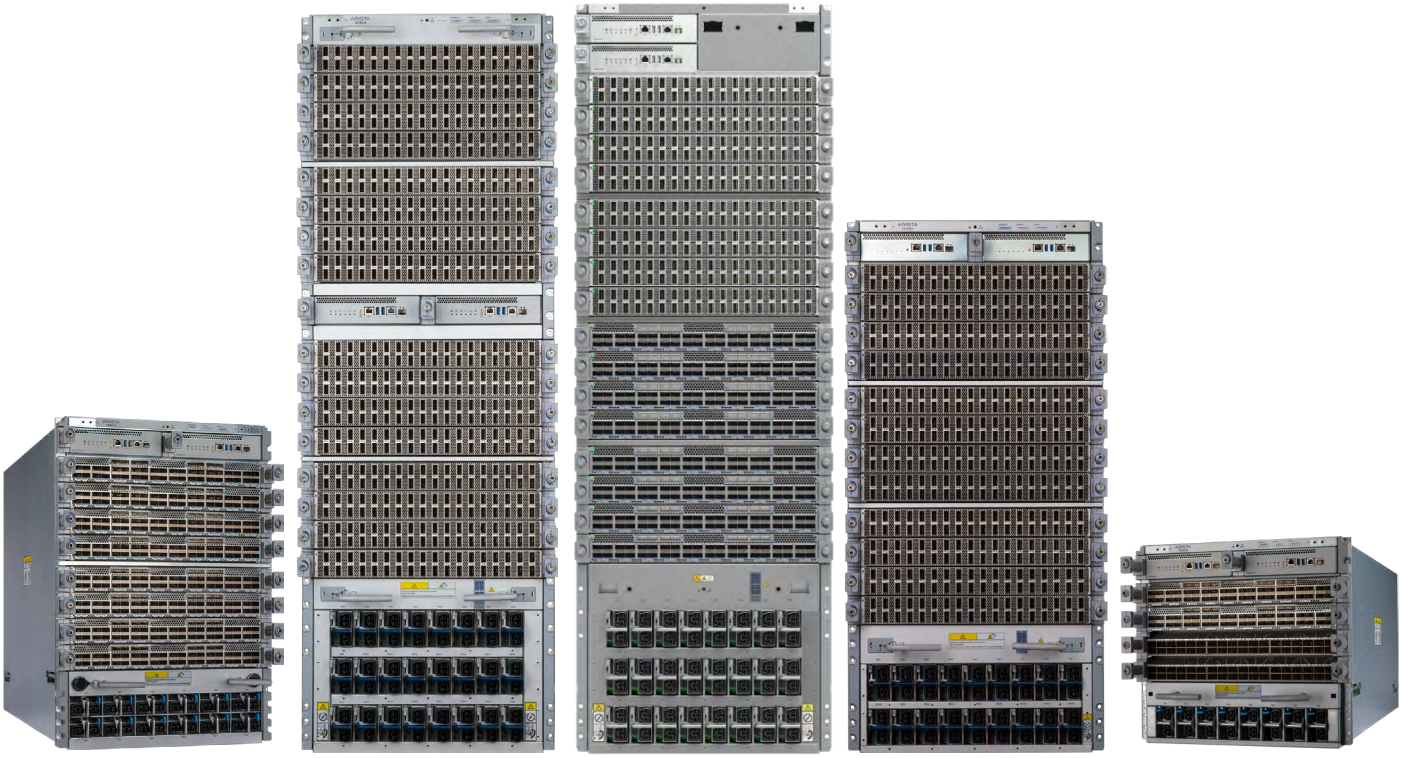


Figure 1: Arista 7800R3 Universal Spine platform.

Arista Networks' award-winning Arista 7500 Series was introduced as a high density non-blocking modular 10G system in April 2010 and established itself as a revolutionary switching platform, which maximized data center performance, efficiency and overall network reliability. Since 2010 the Arista 7500 Series established the benchmark for performance and reliability delivering continuous innovations and seamless upgrades. Nearly 10 years later, Arista is introducing the next generation of modular system, the 800G ready 7800R3 Universal Spine platform, continuing the long heritage of investment protection, scalability and industry leading density and performance. This white paper provides an introduction to the architecture of the Arista 7800R3 Universal Spine platform.

Arista 7800R3: Platform Overview

The Arista 7800R3 Universal Spine Platform is the evolution of the R-Series family of modular switches, with a consistent architecture of deep buffers, VOQ and non-blocking lossless forwarding. The 7800R Series are initially available in a choice of 4-, 8-, 12- and 16-slot systems that support a rich range of line cards providing high density 100G and 400G with choice of forwarding table scale and MACsec, IPsec and Arista TunnelSec™ encryption options.

At a system level, the 16-slot Arista 7816LR3 and 7816R3, with fabrics that scale to 460 Tbps, enable 576 x 400G or 768 x 100G (QSFP100) in a 32RU front to rear power efficient form factor, providing industry-leading performance and density without compromising on features and functionality. The 7816LR3 and 7816R3 provide platforms for service providers and cloud operators to deliver the next generation of rich services. The 7812R3, 7808R3 and 7804R3 are 12, 8 and 4-slot systems that provide the same capabilities in smaller form factors:

Table 1: Arista 7800R3 Key Port and Forwarding Metrics

Characteristic	Arista 7804R3	Arista 7808R3	Arista 7812R3	Arista 7816R3 and 7816LR3
Chassis Height (RU)	10 RU	16 RU	23 RU	32 RU
Linecard Module slots	4	8	12	16
Supervisor Module slots	2	2	2	2
50G Maximum Density using 8x50G breakout	1152	2304	3456	4608
100G Maximum Density (QSFP100)	192	384	576	768
400G Maximum Density (OSFP or QSFP-DD)	144	288	432	576
System Usable Capacity (Tbps)	115.2 Tbps (FD)	230.4 Tbps (FD)	345.6 Tbps (FD)	460.8 Tbps (FD)
Max forwarding throughput per Linecard (Tbps)	14.4 Tbps (DCS-7800R3-36P - 36 x 400G per LC)			
Max forwarding throughput per System (Tbps)	115.2 Tbps (FD)	230.4 Tbps (FD)	345.6 Tbps (FD)	460.8 Tbps (FD)
Max packet forwarding rate per Linecard (pps)	6 Billion pps (7800R3-36P)			
Max packet forwarding rate per System (pps)	24 Bpps	48 Bpps	72 Bpps	96 Bpps
Maximum Buffer memory/ System	96 GB	192 GB	288 GB	384 GB
Virtual Output Queues / System	More than 2.2 million			

The 7800R Series is consistent with the 7500R, using a common architecture. The 7800R Series provides higher density and forwarding on a per system basis as shown below.

Table 2: Arista 7800R3 and 7500R3 System Capacity

Characteristic	7504R3	7508R3	7512R3	7804R3	7808R3	7812R3	7816R3 and 7816LR3
Line card Slots	4	8	12	4	8	12	16
100G Maximum Density (QSFP100)	144	288	432	192	384	576	768
400G Maximum Density (OSFP or QSFP-DD)	96	192	288	144	288	432	576
Max forwarding throughput per System (Tbps)	76.8 Tbps	153 Tbps	230 Tbps	115 Tbps	230 Tbps	345 Tbps	460 Tbps
Max packet forwarding rate per System (pps)	16 Bpps	32 Bpps	48 Bpps	24 Bpps	48 Bpps	72 Bpps	96 Bpps

Designed for the Future

The Arista 7800R3 Universal Spine platform leverages the heritage of the Arista 7500 Series platform which has delivered continuous capacity increases for cloud scale networks for 10 years and driven the industry in new directions for power and efficiency.

The Arista 7800R3 platform builds on the R-Series innovations with system level capacity and environmental enhancements for the next 10 years of data center and core networking.

- Leveraging backplane and midplane less orthogonal connections allowing greater airflow and easy migration to higher performance
- Cooling up to 2.4kW per line card slot using high efficiency fans and datacenter optimized airflow.
- SerDes lanes supporting 112G rates for the next generation of I/O
- Up to 36kW of redundant power providing sufficient capacity for future needs
- 1.3RU height modules for airflow, cooling and power delivery

Arista 7800R3 - Cloud Scale And Features

Arista networks has always leveraged best in class merchant silicon packet processors for leaf and spine systems. The network silicon within the Arista 7800 Series is no exception and utilizes the latest high capacity multi-chip system packet processor, the Jericho2C+ from Broadcom. Clocking in at 16.8Tbps (18 x 400G network interfaces and 9.6Tbps fabric) per chip, it delivers the highest overall full feature system capacity and scale available in the market.

In addition to delivering port density and performance, forwarding table capacity are designed to provide operators with Internet route scale through Arista's innovative FlexRoute™ Engine, extending forwarding table capacity beyond the capability that merchant silicon natively offers. The 7800R3-series Modular Database (MDB) enables flexible allocation of forwarding resources to accommodate a wide range of network deployment roles.

The MDB provides a common database of forwarding and lookup resources to the ingress and egress stages in the 7800R3 platform. These resources are allocated using a set of forwarding profiles that ensure the optimal allocation of resources to different tables for a wide range of networking use-cases. The L3 optimized profile expands the routing and next-hop tables to address large scale networks where route table capacity is required, while the balanced profile is suited for leaf and spine datacenter applications.

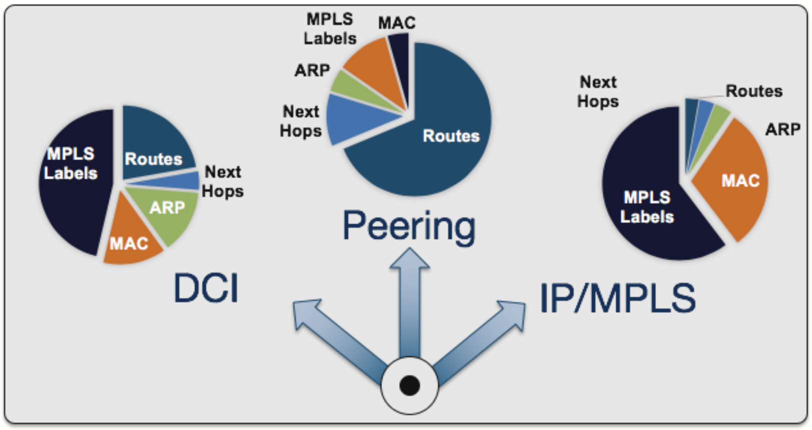


Figure 2: MDB enables a flexible range of deployment profiles.

The fungible nature of the resources within the MDB ensure that operators have the flexibility they need to standardize on a common platform across a wide range of roles with the confidence that the specific resource requirements can be achieved according to the needs of any given role. There is no need to have separate platforms for core network roles and edge roles in today's service provider networks. This enables cloud and service providers to streamline deployments, simplify sparing and consolidate testing.

Arista 7800R3: System Components

Chassis and Mid-Plane-less system

All Arista 7800R3 systems share a common architecture with identical fabric bandwidth and forwarding capacity per slot. Line cards, supervisor modules and power supplies are common across systems; the only differences are the physical size, fabric/fan module size, number of line card slots and quantity of power supplies. Airflow is always front-to-rear and all data cabling is at the front of the chassis.

Chassis design and layout are key aspects that enable high performance per line card slot: the fabric modules are directly behind line card modules and oriented orthogonally to the line card modules. This design alleviates the requirement to route high speed signal traces on a midplane or backplane, reducing trace lengths between system elements and enabling high speed signals to operate more efficiently with high signal integrity by being shorter lengths. The 7800R system eliminates the passive midplane or backplane, allowing the direct connection of line cards to fabric modules improving airflow and providing investment protection for future higher capacity systems.

Supervisor Modules

The Supervisor modules on the Arista 7800R Universal Spine platform are used for control-plane and management plane functions only. Two redundant supervisors can be installed in the chassis, each capable of managing the system. All data-plane forwarding is performed on line card modules and forwarding between line card modules is always via the fabric modules.

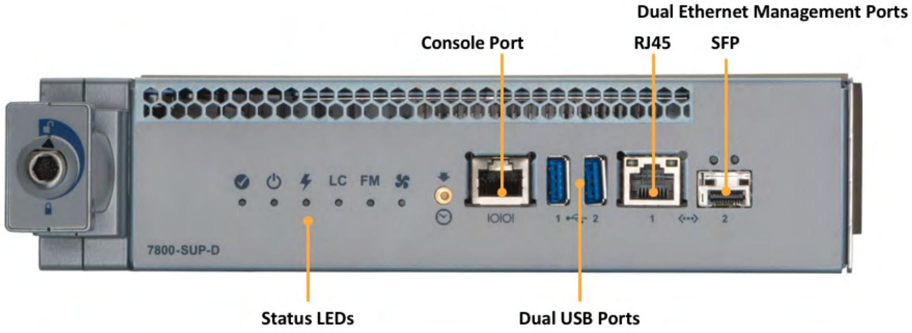


Figure 3: Arista 7800 Series Supervisor Module.

The Arista 7800 Series Supervisor provides a 6-core Intel Xeon Broadwell DE CPU running at 1.9GHz with 64GB RAM and is used with both the 7804R3 and 7808R3. The 7816L, 7816 and 7812 require a specific supervisor, which is equipped with an 8-core CPU.

Arista's Extensible Operating System (EOS®) makes full use of multiple cores due to its unique multi-process state sharing architecture that separates state information and packet forwarding from protocol processing and application logic. The multi-core CPU and large memory configuration provides headroom for running third party software within the same Linux instance as EOS, within a guest virtual machine or within containers. An enterprise-grade SSD provides additional flash storage for logs, VM images or third party software packages.

Out-of-band management is available via a serial console port and/or dual 10/100/1000 Ethernet interfaces (SFP and RJ45 ports are provided). There are two USB 2.0 interfaces that can be used for transferring images/logs or many other uses. A pulse-per-second clock input is provided for accurate clock synchronization.

There is more than 126 Gbps of inband connectivity from data-plane to control-plane and 10G connectivity between redundant Supervisor modules. Combined, these enable very high performance connectivity for the control-plane to manage and monitor the data-plane as well as replicate state between redundant Supervisors.

Arista 7800R3: Distributed Packet Forwarding

Distributed Data-Plane Forwarding

All Arista 7800R3 systems share a common architecture with identical fabric bandwidth and forwarding capacity per slot. Line cards, supervisor modules and power supplies are common across systems; the only differences are the physical size, fabric/fan module size, number of line card slots and quantity of power supplies. Airflow is always front-to-rear and all data cabling is at the front of the chassis.

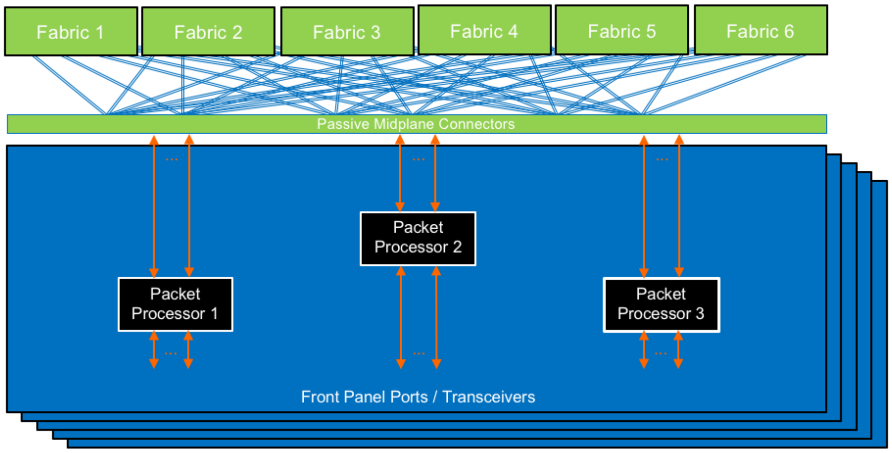


Figure 4: Distributed Forwarding within an Arista 7500R3 Series

Arista 7800R3 Universal Spine platform line card modules utilize packet processors to provide distributed dataplane forwarding. Forwarding between ports on the same packet processor utilizes local switching and no fabric bandwidth is used. Forwarding across different packet processors uses all fabric modules in a fully active/active mode. There is always Virtual Output Queuing (VOQ) between input and output, for both locally switched and nonlocally switched packets, ensuring there is always fairness even where some traffic is local.

Fabric Modules

Within the Arista 7800R3 multiple fabric modules are utilized in an active/active configuration. The 7816 system is equipped with 12 fabric modules, while the 7816L, 7812, 7808 and 7804 use 6 modules. Each fabric module provides up to 6 Tbps fabric bandwidth full duplex (3 Tbps transmit and 3 Tbps receive) to each line card slot, and with all fabric modules active in a 7816LR3/7816R3 system, 576 Tbps (288 Tbps transmit and 288 Tbps receive) of bandwidth is available. The fabric capacity is overprovisioned with respect to the usable line card capacity to ensure that if a fabric module were to fail, the throughput of the system degrades gracefully. Fabric modules support hot swap and may be inserted or removed while the system is in operation.

Each fabric module contains multiple field serviceable fan modules comprised of dual counter-rotating fans. Fans are connected to two independent fan controllers to increase fault tolerance. To ensure that there is sufficient cooling capacity for line cards and fabric modules, fans are powered on before any active forwarding elements and continue to run even if the fabric chips are disabled.

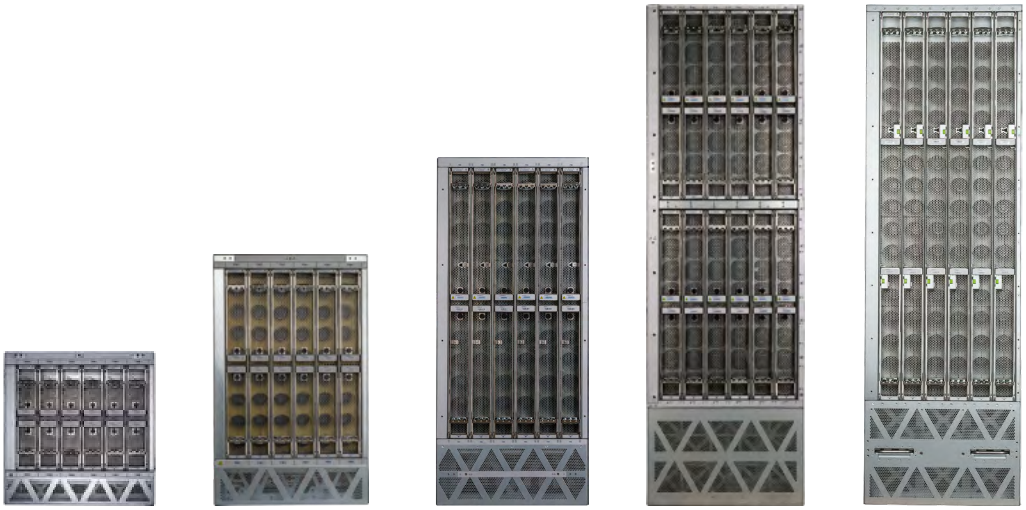


Figure 5: Arista DCS-7800R3 Series Fabric/Fan modules

Packets are transmitted across fabric modules as variable sized cells of up to 256 bytes. Serialization latency of larger frames is amortized via the parallel cell distribution that utilizes all available paths in an active/active manner, preventing hot spots or blocking that can occur with packet-based fabrics.

Besides data-plane packets, the fabric modules are also used for a number of other functions:

- **Virtual Output Queuing (VOQ):** a distributed scheduling mechanism is used within the switch to ensure fairness for traffic flows contending for access to a congested output port. A credit request/grant loop is utilized and packets are queued in physical buffers on ingress packet processors within VOQs until the egress packet scheduler issues a credit grant for a given input packet.
- **Hardware-based distributed MAC learning and updates:** when a new MAC address is learned, moves or is aged out, the ingress packet processor with ownership of the MAC address will update other packet processors.
- **Data-plane health tracer:** all packet processors within the system send continuous health check messages to all other packet processors, validating all data-plane connectivity paths within the system.

The 7800R3 series is designed to ensure growth for the future. As SerDes speeds increase from 50G per lane to 100G per lane, there is sufficient capacity to double the system capacity overall.

Table 3: Arista 7800R3 Fabric Bandwidth

	7800R3 50G SerDes	7800R3 100G SerDes
Max Physical Lanes per Line card	384	384
Fabric bandwidth per Line card (Full Duplex)	38.4T	76.8T
Fabric bandwidth per 8-Slot system (Full Duplex)	273.6T	614.4T

7800R3 Series Power Supplies

The 7800R3 series supports AC or DC power supplies, with a common grid-redundant dual-input 3kW model supported across all form-factors. The AC power supply's dual inputs operate in an active/standby configuration. The power supply implements Auto-Transfer Switchover (ATS) to switch between inputs for redundancy. DC power supplies provide load sharing/redundancy across both inputs.

7800R3 series platforms utilize a single internal power domain, so there is no need to allocate specific power supply units to individual power zones. This enables operators to achieve N+2 power supply redundancy as well as balance power feeds across the data center power grids.

Arista 7800R3 / 7800R3A: Line Card Architecture

Arista 7800R3 Universal Spine Platform Line Card Layout

Arista 7800R3 line card modules utilize the same Jericho2 family packet processors, with the number of packet processors varied based on the number and type of ports on the module. The packet forwarding architecture of each of these modules is essentially the same: a group of front-panel ports (different transceiver/port/speed options) is connected to a packet processor with connections to the fabric modules.

Each Jericho2 packet processor supports network interface speeds ranging from 10G to 400G for up to 4.8Tbps of total network capacity. In addition, 5.6Tbps of capacity provides connectivity to the fabric modules.

The 4.8Tbps of capacity per packet processor is delivered over a total of 96 50G PAM SerDes lanes that can be run from 10G to 50G and individually or combined in groups to allow flexible 10G, 25G, 40G, 50G 100G, 200G, and 400G interfaces.

As there are 96 PAM4 lanes, each packet processor supports up to a maximum of 96 logical or physical interfaces per chip, which defines the maximum possible port density for a given product form factor.

Some line cards employ gearboxes to increase the front panel interface density and maximize the capabilities by converting the 50G PAM4 SerDes lanes to more lanes at lower speeds and different encoding. For example converting 2 x 50G PAM4 lanes to 4 x 25G NRZ lanes.

Gearboxes enable an increased choice of interfaces without requiring additional packet processors, reducing overall system power consumption and heat generation. As the number of physical interfaces and supported breakout options is flexible, EOS provides tools to enable both configuration and analysis of the available port combinations for each platform.

Jericho2C is a member of the Jericho2 silicon family providing 2.4Tbps of front panel bandwidth and 2.4Tbps of inter-chip fabric capacity per chip. It is designed for lower capacity systems with a focus on 1G to 100G network connectivity with 100G to 400G uplinks.

Each Jericho2C chip supports a maximum of 32 50G PAM4 and 96 25G NRZ SerDes lanes. As with Jericho2, the 50G lanes support speeds from 10G to 50G individually or may be combined in groups to support interfaces up to 400G. The 25G NRZ lanes support 1G, 10G or 25G individually or 40G/100G when combined.

Arista 7800R3A Universal Spine Platform Line Card Layout

Arista 7800R3A line card modules utilize the next generation Jericho2C+ family packet processors, with the number of packet processors varied based on the number and type of ports on the module.

Jericho2C+ is an enhanced version of Jericho2 based on a 7 nm fabrication process. Jericho2C+ implements a pipeline and memory design that is fully compatible and consistent with Jericho2, while increasing the front panel bandwidth from 4.8Tbps to 7.2Tbps and offering fully integrated wire speed AES-256-GCM bulk encryption enabling MACsec, IPsec and VXLANsec.

Each Jericho2C+ packet processor supports network interface speeds ranging from 10G to 400G delivered over a total of 144 50G PAM SerDes lanes that can be run from 10G to 50G individually or combined in groups to allow flexible 10G, 25G, 40G, 50G 100G, 200G, and 400G interfaces.

While there are 144 PAM4 lanes, each packet processor supports up to a maximum of 118 logical or physical interfaces per chip, which defines the maximum possible port density for a given product form factor. Device specific front panel layout, availability of suitable transceivers and EOS support will govern the actual amount of possible breakouts.

Table 4: Arista 7800R3 / 7800R3A Series Line card Module Port Characteristics

Line card	Port (type)	Interfaces						Port Buffer	Fowarding Rate	Switching Capacity
		25G	50G	40G*	100G	400G	Max §			
7800R3A-36P 7800R3A-36PM 7800R3AK-36PM	36 OSFP	232	232	36	144	36	232	16GB	5.4 Bpps	14.4T
7800R3A-36D 7800R3A-36DM 7800R3AK-36DM	36 QSFP-DD	232	232	36	144	36	232	16GB	5.4 Bpps	14.4T
7800R3-36P	36 OSFP	288	288	36	144	36	288	24GB	6.0 Bpps	14.4T
7800R3-36D 7800R3K-36DM	36 QSFP-DD	288	288	36	144	36	288	24GB	6.0 Bpps	14.4T
7800R3-48CQ 7800R3K-48CQ	48 QSFP100	96	96	48	48	-	96	8GB	2.0 Bpps	4.8T
7800R3-48CQM	48 QSFP100	-	-	48	48	-	48	8GB	2.0 Bpps	4.8T
7800R3-48CQ2 7800R3-48CQM2 7800R3-48CQMS	48 QSFP100	96	96	48	48	-	96	8GB	2.0 Bpps	4.8T
7800R3K-72Y	72 SFP25	72	32	-	-	-	72	4GB	1.0 Bpps	2.4T

** Maximum port numbers are uni-dimensional, may require the use of break-outs, and are subject to transceiver/cable capabilities. 40G assumes use of QSFP+*
§ Where supported by EOS, each system supports a maximum number of interfaces. Certain configurations may impose restrictions on which physical ports can be used.

DCS-7800R3A-36P-LC

All stages associated with packet forwarding are performed in integrated system on chip (SoC) packet processors. Each packet processor provides both the ingress and egress packet forwarding pipeline stages for packets that arrive or are destined to the ports serviced by that packet processor. Each packet processor can perform local switching for traffic between ports on the same packet processor.

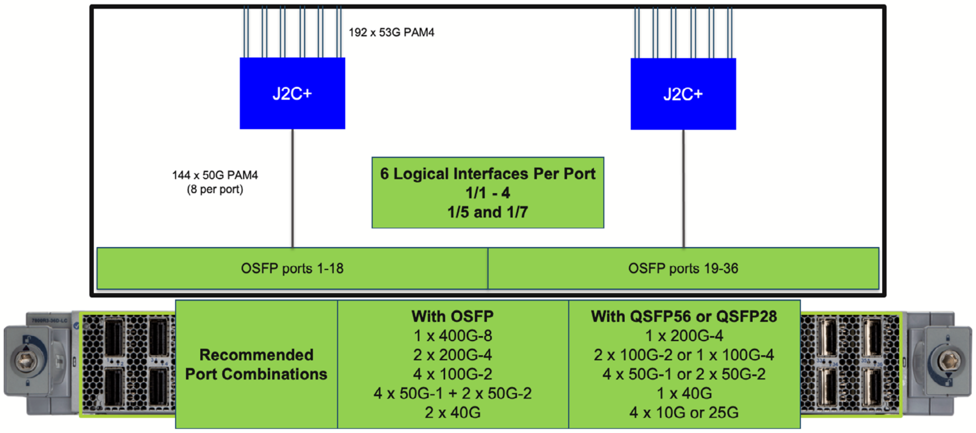
The architecture of a line card, in this case the DCS-7800R3A-36P-LC, a 36-port 400G OSFP module, is shown below in Figure 6. Each of the packet processors on the line card services a group of front panel ports. Each port is capable of supporting copper, AOC as well as the range of optics available in the OSFP form-factor subject to the capabilities of the cable or transceiver.

In the default configuration, each port is designed to offer up to 4-way breakout when using either a 100/200G QSFP or a 400G OSFP. 6 logical interfaces are available on each physical port, corresponding to lanes 1,2,3,4,5 and 7. Of these 6 interfaces, up to 4 logical interfaces can be active on each port, which provides the following connectivity choices:

- When used with a 400G OSFP:
  - 1 x 400G-8
  - 2 x 200G-4
  - 4 x 100G-2
  - 2 x 100G-4
  - 4 x 50G-2
- When used with a 200G QSFP:
  - 1 x 200G-4
  - 2 x 100G-2
  - 4 x 50G-1
  - All 40/100G QSFP speeds subject to transceiver support
- When used with a 100G QSFP:
  - 1 x 100G-4
  - 2 x 50G-2
  - 4 x 25G

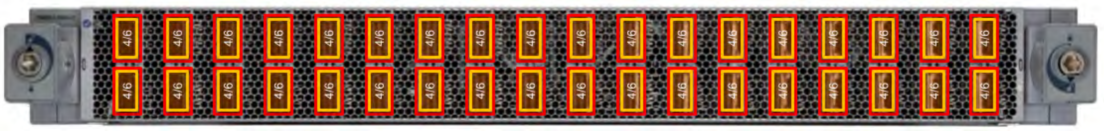
- » All 40 QSFP speeds subject to transceiver support
- When used with a 40G QSFP:
  - » 1 x 40G
  - » 4 x 10G

EOS provides alternative configuration profiles which enable different breakout capabilities, up to a maximum of 232 logical ports.



Recommended Port Combinations	With OSFP	With QSFP56 or QSFP28
	1 x 400G-8 2 x 200G-4 4 x 100G-2 4 x 50G-1 + 2 x 50G-2 2 x 40G	1 x 200G-4 2 x 100G-2 or 1 x 100G-4 4 x 50G-1 or 2 x 50G-2 1 x 40G 4 x 10G or 25G

Figure 6a: Arista DCS-7800R3A-36P-LC module architecture

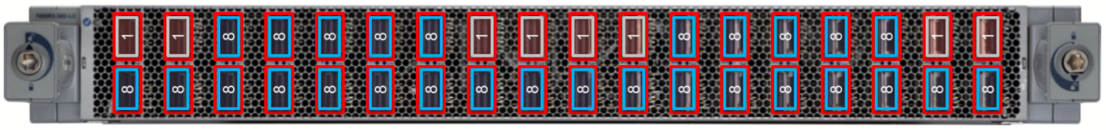


4/6 6 Logical Ports (Max 4 Active) – e.g. Eth x/1,3,5,7 or Eth x/1,2,3,4

🟥 Speed Group (Max 2 SerDes Speeds)

Port Type	Max Density	Comment
10/25G	144*	Using all ports
50G-2	144	Using all ports
50G-1	144*	Using all ports
100G-4	72	Using breakout QDD
100G-2	144	Using all ports
200G-4	72	Using all ports
400G-8	36	Using all ports

Figure 6b: Arista DCS-7800R3A-36P-LC default breakout capabilities (indicates not all signaling lanes are used)*



1 1 Logical Port – Eth x/1

8 8 Logical Ports – Eth x/1,2,3,4,5,6,7,8

🟥 Speed Group (Max 2 SerDes Speeds)

Port Type	Practical / Max Density	Comment
10/25G	224/232*	Using 8-way ports only / all ports
50G-2	112/120*	Using 8-way ports only / all ports
50G-1	224/232*	Using 8-way ports only / all ports
Any 50G	232	224 x 50G-1 using 8-way ports & 8 x (1 x 50G-1 or -2) using 1-way ports
100G-4	64*	Using QSFP28 in 1 way ports and breakout QDD on 8 way
100G-2	112/120*	Using 8-way ports only / all ports
Any 100G	120*	28 x (4 x 100G-2) on 8-way ports & 8 x (1x 100G-2 or -4) on 1-way ports
200G-4	56/64*	Using 8-way ports only / all ports
400G-8	36	Using all ports

Figure 6c: Arista DCS-7800R3A-36P-LC maximum breakout profile (indicates not all signaling lanes are used)*

DCS-7800R3A-36D-LC

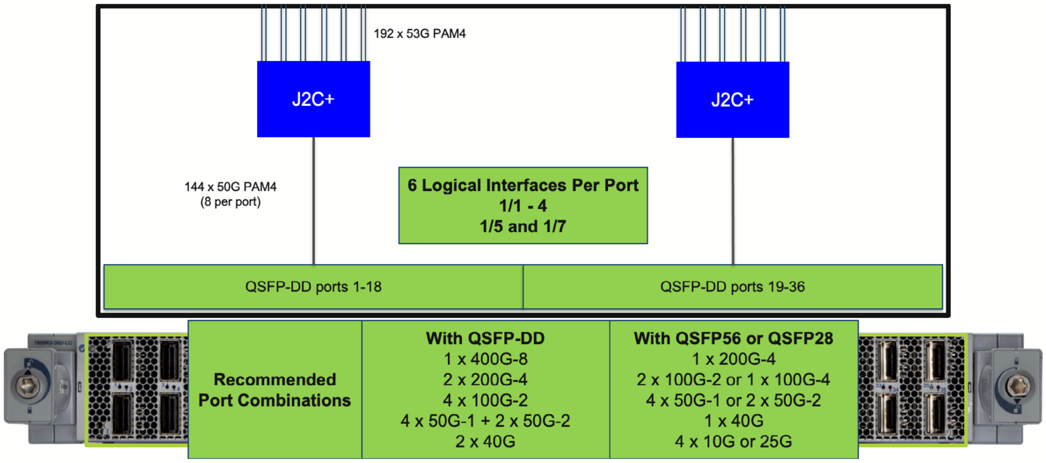


Figure 7: Arista DCS-7800R3A-36D-LC module architecture

The DCS-7800R3A-36D-LC is the QSFP-DD version of the previous line card. The packet processor to port assignment and default logical port assignments are identical. Further, each port is capable of supporting copper, AOC as well as the range of optics available in the QSFP-DD form-factor.

DCS-7800R3-36P-LC

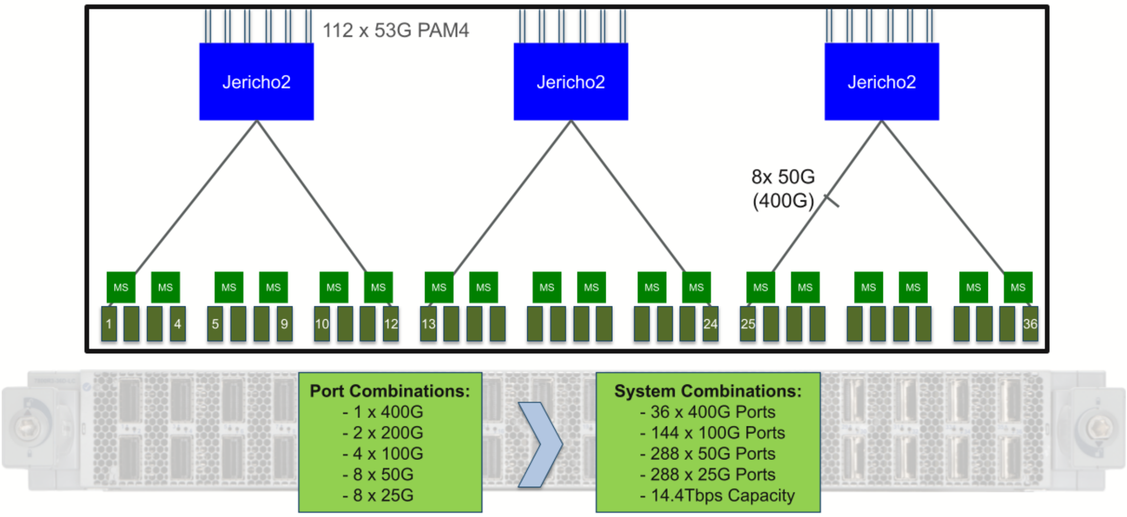


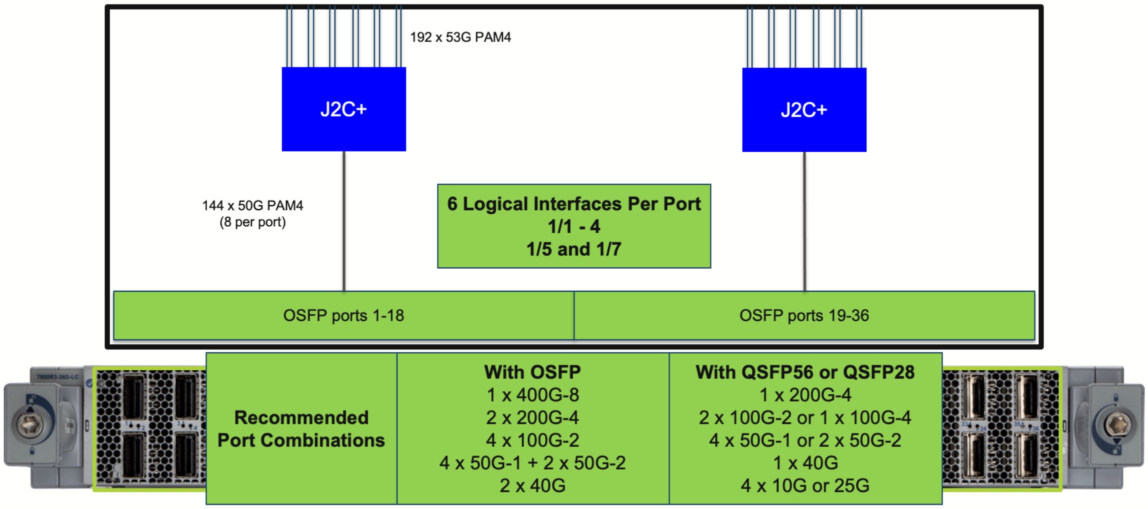
Figure 8: Arista DCS-7800R3-36P-LC module architecture

The DCS-7800R3-36P-LC is, a 36-port 400G OSFP module. Each of the packet processors on the line card directly services a group of front panel ports. Each of the 36 ports can operate as either 1 x 400G, 2 x 200G, 4 x 100G, 8 x 50G or 8 x 25G interfaces. Each port is capable of supporting copper, AOC as well as the range of optics available in the OSFP form-factor subject to the capabilities of the cable or transceiver.

DCS-7800R3-36D-LC

The DCS-7800R3-36D-LC is the QSFP-DD version of the previous line card. The packet processor to port assignment is identical. Further, each port is capable of supporting copper, AOC as well as the range of optics available in the QSFP-DD form-factor.

DCS-7800R3A-36PM-LC and DCS-7800R3A-36DM-LC

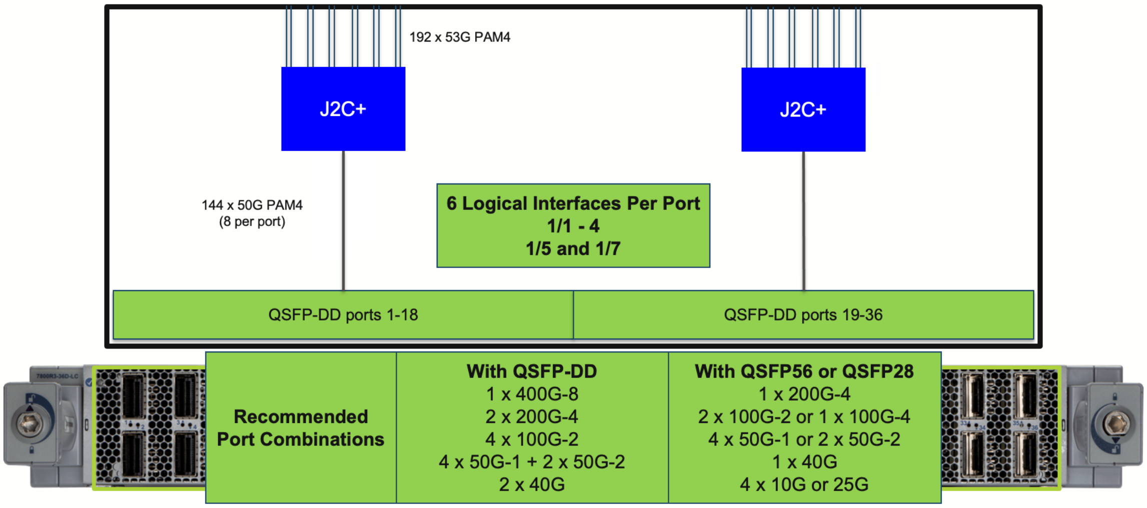


Recommended Port Combinations	With OSFP	With QSFP56 or QSFP28
	1 x 400G-8 2 x 200G-4 4 x 100G-2 4 x 50G-1 + 2 x 50G-2 2 x 40G	1 x 200G-4 2 x 100G-2 or 1 x 100G-4 4 x 50G-1 or 2 x 50G-2 1 x 40G 4 x 10G or 25G

Figure 9: Arista DCS-7800R3A-36PM-LC module architecture

The DCS-7800R3A-36PM-LC and DCS-7800R3A-36DM-LC are encryption capable variants of the 7800R3A-36P (OSFP) and 7800R3A-36D (QSFP-DD) 36 port 400G line cards. As encryption is directly integrated into the Jericho2C+ packet processor, the hardware layout is identical and other attributes of the line card including port assignments are identical. MACsec, IPsec and VXLANsec are supported at all speeds.

DCS-7800R3AK-36PM-LC and DCS-7800R3AK-36DM-LC



Recommended Port Combinations	With QSFP-DD	With QSFP56 or QSFP28
	1 x 400G-8 2 x 200G-4 4 x 100G-2 4 x 50G-1 + 2 x 50G-2 2 x 40G	1 x 200G-4 2 x 100G-2 or 1 x 100G-4 4 x 50G-1 or 2 x 50G-2 1 x 40G 4 x 10G or 25G

Figure 10: Arista DCS-7800R3AK-36DM-LC module architecture

The DCS-7800R3AK-36PM-LC and DCS-7800R3AK-36DM-LC are large scale, encryption capable variants of the 7800R3A-36P (OSFP) and 7800R3A-36D (QSFP-DD) 36 port 400G line cards. As both large scale resources and encryption are directly integrated into the Jericho2C+ packet processor, the hardware layout is identical and other attributes of the line card including port assignments are identical. MACsec, IPsec and VXLANsec are supported at all speeds.

DCS-7800R3K-36DM-LC

The DCS-7800R3K-36DM-LC adds large table scale and MACsec encryption engines to the DCS-7800R3-36D line card. MACsec is supported at all speeds.

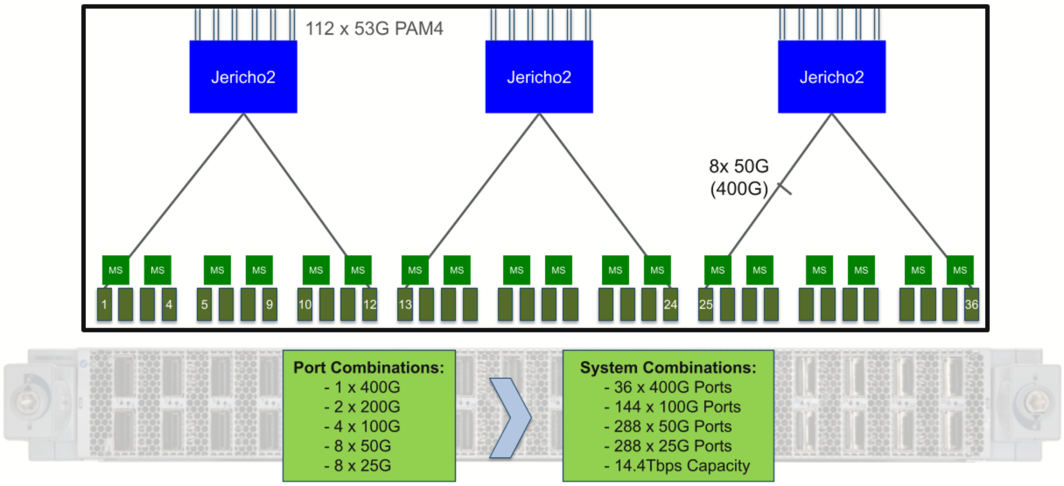


Figure 11: Arista DCS-7800R3K-36DM-LC module architecture

DCS-7800R3-48CQ-LC

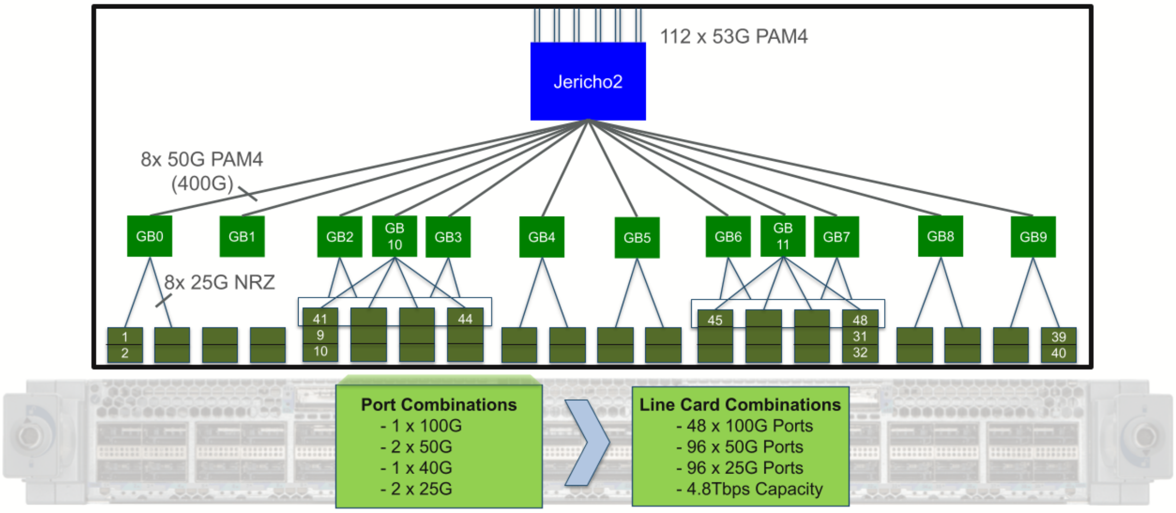


Figure 12: Arista DCS-7800R3-48CQ-LC module architecture

The DCS-7800R3-48CQ line card utilizes a single Jericho2 chip and provides 12 gearboxes to support a diverse range of optics and cables for up to 96 individual interfaces when breakouts are used.

Four logical interfaces are assigned to each odd+even pair of QSFP ports. The even numbered port is considered the primary port from each pair. This allows for combinations including:

- Both ports running in 100G mode: 2 x 100G-4
- Both ports running in 40G mode: 2 x 40G
- Breaking both ports into 2 x 50G or 2 x 25G (Total: 4 x 50G-2 or 4 x 25G-1)
- Running the even-numbered port as a breakout to 25G: 4 x 25G (odd-numbered port disabled)
- Running the even-numbered as a breakout to 10G: 4 x 10G (odd-numbered port disabled)

CLI tools provide further platform-specific details on the combinations of interface speeds available across the system. In order to operate a port in breakout mode (i.e., run a 100G port as 4 x 25G or 2 x 50G) the underlying transceiver must enable this.

The DCS-7800R3K-48CQ-LC is the large table version of the previous line card, and the 7800R3-48CQM and 7800R3-48CQMS modules add support for MACsec.

DCS-7800R3-48CQM-LC and DCS-7800R3-48CQMS-LC

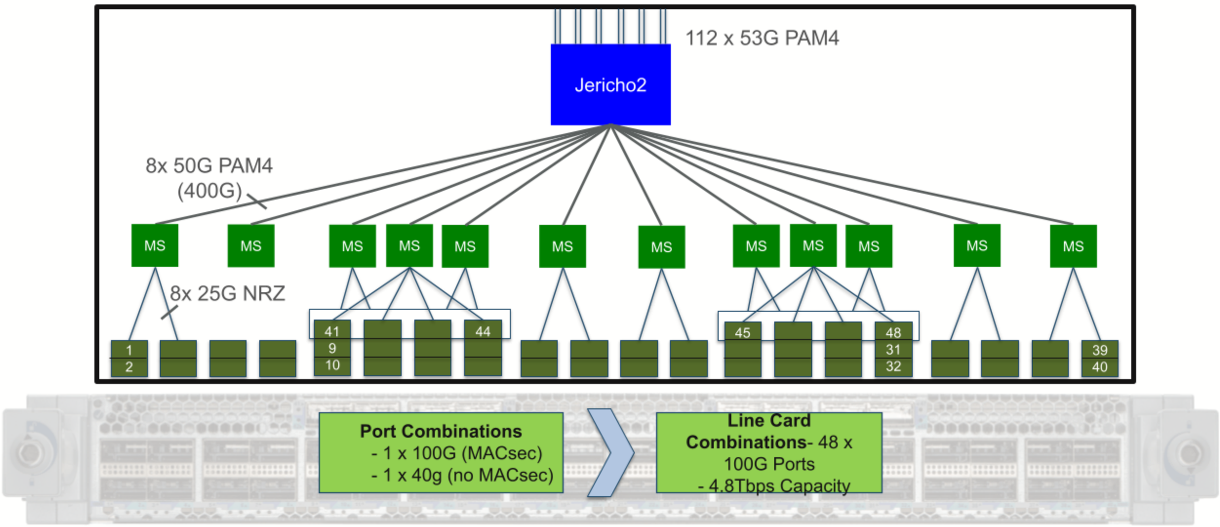


Figure 13: Arista DCS-7800R3-48CQM-LC module architecture

The 7800R3-48CQM enhances the 7800R3-48CQ line card with the addition of encryption support in the gearbox chips. 100G MACsec is supported on all QSFP ports, while 40G is also supported without MACsec.

The 7800R3-48CQM does not support breakout of QSFP ports to lower speed interfaces.

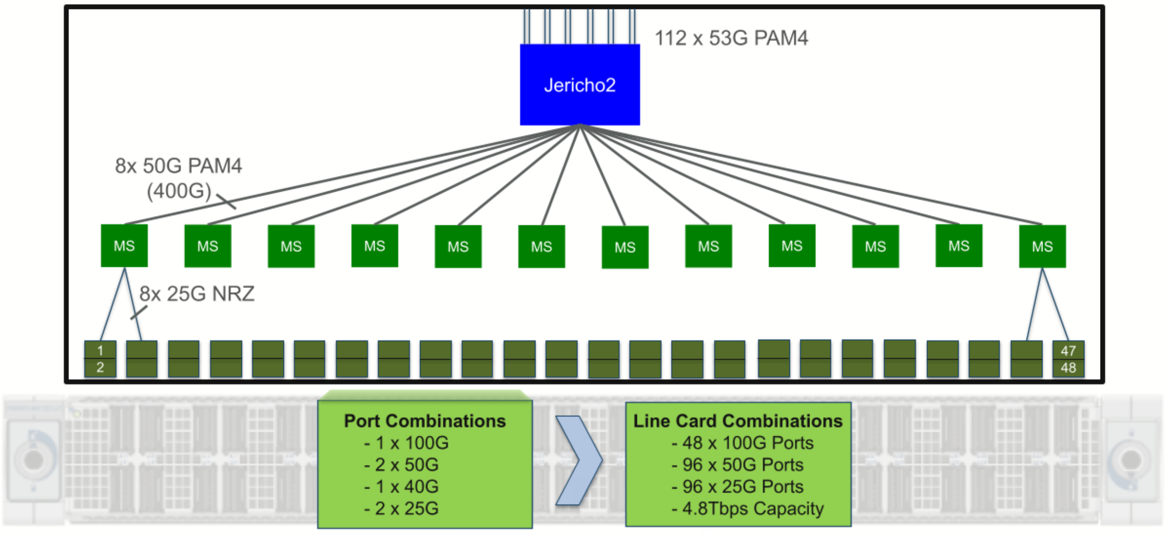


Figure 14: Arista DCS-7800R3-48CQMS-LC module architecture

The 7800R3-48CQMS is a further enhanced version of the 7800R3-48CQM line card with the addition of break-out support and MACsec support at speeds below 100G.

Ports are rotated 90˚ to improve cooling, the top row of ports are numbered 1, 3, 5 .. 47, while the bottom row of ports are even numbered.

Four logical interfaces are assigned to each horizontal pair of QSFP ports (e.g. 1 and 3, 2 and 4). In each pair, the rightmost port is considered the primary port - e.g. for the port block 1 - 4, ports 3 and 4 are primary.

This allows for combinations including:

- Both ports running in 100G mode: 2 x 100G-4
- Both ports running in 40G mode: 2 x 40G
- Breaking both ports into 2 x 50G or 2 x 25G (Total: 4 x 50G-2 or 4 x 25G-1)
- Running the rightmost port as a breakout to 25G: 4 x 25G (leftmost port disabled)
- Running the rightmost as a breakout to 10G: 4 x 10G (leftmost port disabled)

CLI tools provide further platform-specific details on the combinations of interface speeds available across the system. In order to operate a port in breakout mode (i.e., run a 100G port as 4 x 25G or 2 x 50G) the underlying transceiver must enable this.

DCS-7800R3K-72Y-LC

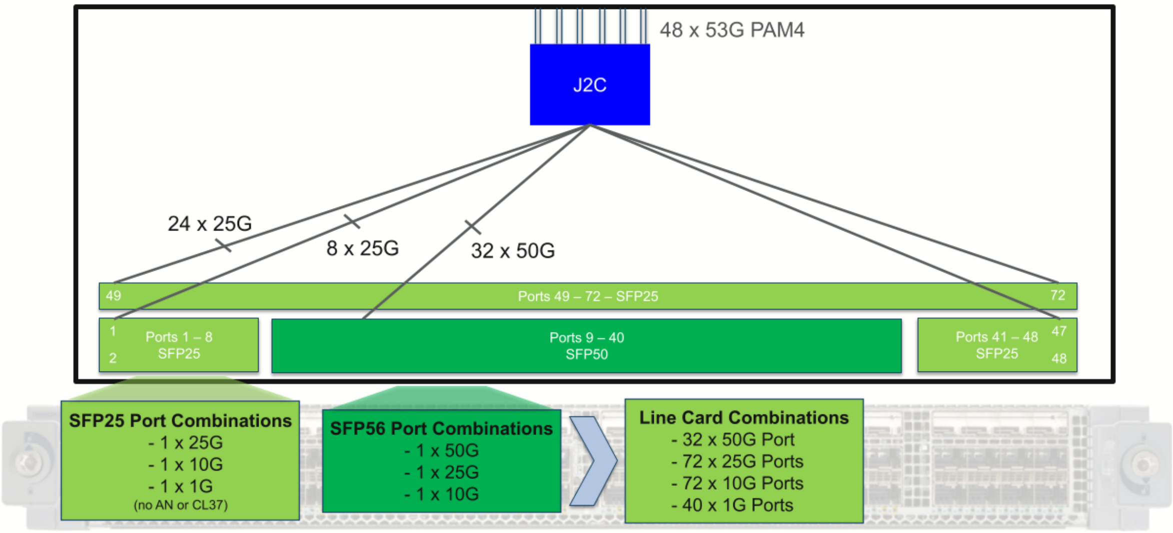


Figure 15: Arista DCS-7800R3K-72Y-LC module architecture

The DCS-7800R3K-72Y line card utilizes a single Jericho2C chip to provide a mixture of native 25G and 50G SFP ports.

The 72 SFP ports can be used as follows:

- Each of the 40 SFP25 ports, 1-8, 41-48 and 49-72 supports 1, 10 and 25G SFPs.
- The 32 SFP50 ports (9-40) support 10G, 25G and 50G-1 SFPs. The ports are divided into 4 speed groups. Each speed group supports up to two unique speeds (e.g. 10/25G, 10/50G, 25/50G) across the 8 member ports.

Each port is capable of supporting copper, AOC as well as the range of optics available in the SFP form-factor subject to the capabilities of the cable or transceiver.

7800R3 and 7800R3A Scale and Performance

The following tables detail the key hardware attributes and logical scalability of the 7800R3 and 7800R3A families:

Table 5: Arista 7800R Series Logical Resources, Scale and Performance

Packet Processor	7800R3 / 7800R3K (Jericho2)	7800R3A / 7800R3AK (Jericho2C+)
Bandwidth	4.8T	7.2T
Density	96 x 25G 48 x 100G-2 12 x 400G-8	144 x 25G 72 x 100G-2 18 x 400G-8
Performance	2.0 Bpps	2.8 Bpps
Buffer	8GB - HBM2	8GB - HBM2

The table below shows the key scale metrics of the 7800R3 and 7800R3K Series.

Table 6: Arista 7800R3 / 7800R3A Key L2, L3 Scale Metrics[1]

	7800R3 / 7800R3A Series		7800R3K / 7800R3AK Series			
	L3 Profile (default)	Balanced Profile	L3-XL Profile (default)	L3-XXL Profile	L3-XXXL Profile	Balanced-XL Profile
ARP Entries	88k	80k	112k	112k	80k	96k
MAC Addresses	224k	224k	256k	192k	384k	256k
IPv4 Unicast Routes	1450k	800k	2250k	2850k	3950k	1850k
Additional IPv4 Unicast Routes with FlexRoute	+ 1,792k	+ 1,792k	+ 2,048k	+ 1,536k	+ 3,072k	+ 2,048k
IPv6 Unicast Routes	433-483k	250-267k	683-750k	833-950k	1100-1317k	567-617k
Multicast Routes	128k	128k	128k	128k	128k	128k
TCAM ACL Entries (Per chip)	24k	24k	24k	24k	24k	24k
Traffic Policy ACL IPv4 Prefixes	30k	30k	430k	296k	30k	430k
Traffic Policy ACL IPv6 Prefixes	10k	10k	150k	100k	10k	150k
ECMP	512-Way	512-Way	512-Way	512-Way	512-Way	512-Way

[1] Unidimensional scaling maxima with currently available MDB profiles. Available resources depend on user configuration.

Arista 7800R3: Packet Forwarding Pipeline

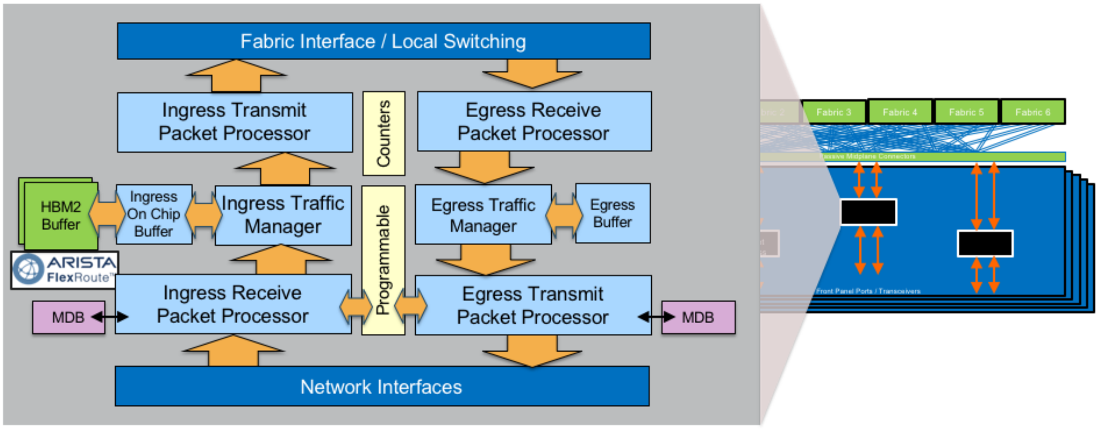

Fabric Interface / Local Switching
Ingress Transmit Packet Processor
Counters
Egress Receive Packet Processor
HBM2 Buffer
Ingress On Chip Buffer
Ingress Traffic Manager
Programmable
Egress Traffic Manager
Egress Buffer
ARISTA FlexRoute
MDB
Ingress Receive Packet Processor
Egress Transmit Packet Processor
MDB
Network Interfaces
Fabric 3
Fabric 4
Fabric 5
Fabric 6
Front Panel Ports / Transceivers

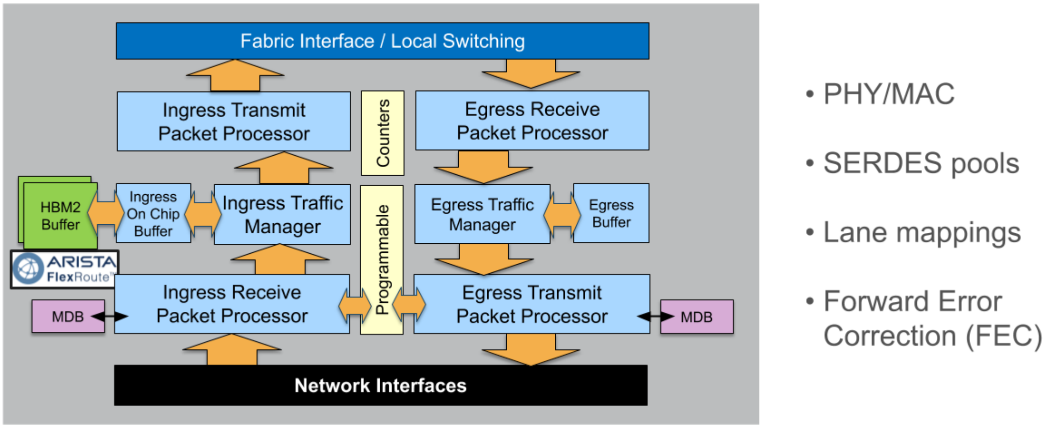
Figure 16: Packet forwarding pipeline stages inside a packet processor on an Arista 7800R3 line card module

Each packet processor on a line card is a System-on-Chip (SoC) that provides all the ingress and egress forwarding pipeline stages for packets to or from the front panel input ports connected to that packet processor. Forwarding is always hardware-based and never falls back to software/CPU forwarding.

The steps involved at each of the logical stages of the packet forwarding pipeline are outlined below.

Stage 1: Networking Interface (Ingress)

When packets/frames enter the switch, the first block they arrive at is the Network Interface stage. This is responsible for implementing the Physical Layer (PHY) interface and Ethernet Media Access Control (MAC) layer on the switch and any Forward Error Correction (FEC).


Fabric Interface / Local Switching
Ingress Transmit Packet Processor
Counters
Egress Receive Packet Processor
HBM2 Buffer
Ingress On Chip Buffer
Ingress Traffic Manager
Programmable
Egress Traffic Manager
Egress Buffer
ARISTA FlexRoute
MDB
Ingress Receive Packet Processor
Egress Transmit Packet Processor
MDB
Network Interfaces

- PHY/MAC
- SERDES pools
- Lane mappings
- Forward Error Correction (FEC)

Figure 17: Packet Processor stage 1 (ingress): Network Interface

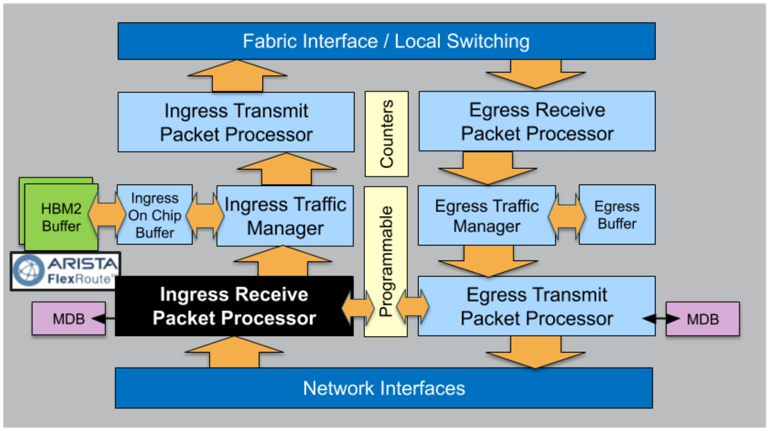
The PHY layer is responsible for the transmission and reception of bitstreams across physical connections including encoding, multiplexing, synchronization, clock recovery, and serialization of the data on the wire for whatever speed/type Ethernet interface is configured.

Programmable lane mapping is used to map the physical lanes to logical ports based on the interface type and configuration. Lane mapping is used for breakout of 4x25G and 2x50G on 100G ports.

If a valid bitstream is received at the PHY then the data is sent to the MAC layer. On input, the MAC layer is responsible for turning the bitstream into frames/packets: checking for errors (FCS, Inter-frame gap, detect frame preamble), and finding the start of frame and end of frame delimiters.

Stage 2: Ingress Receive Packet Processor

The Ingress Receive Packet Processor stage is responsible for forwarding decisions. It is the stage where all forwarding lookups are performed.


Fabric Interface / Local Switching
Ingress Transmit Packet Processor
Counters
Egress Receive Packet Processor
HBM2 Buffer
Ingress On Chip Buffer
Ingress Traffic Manager
Programmable
Egress Traffic Manager
Egress Buffer
ARISTA FlexRoute
MDB
Ingress Receive Packet Processor
Egress Transmit Packet Processor
MDB
Network Interfaces

- Packet Parsing
- SMAC/DMAC/ DIP lookups
- Forwarding table lookups
- Tunnel Decap
- Ingress ACL
- Resolution of forwarding action

Figure 18: Packet Processor stage 2 (ingress): Ingress Receive Packet Processor

After parsing the relevant encapsulation fields, the DMAC is evaluated to see if it matches the device's MAC address for the physical or logical interface. If it's a tunneled packet and is destined to a tunnel endpoint on the device, it is decapsulated within its appropriate virtual routing instance and packet processing continues on the inner packet/frame headers. If it's a candidate for L3 processing (DMAC matches the device's relevant physical or logical MAC address) then the forwarding pipeline continues down the layer 3 (routing) pipeline, otherwise forwarding continues on the layer 2 (bridging) pipeline.

In the layer 2 (bridging) case, the packet processor performs SMAC and DMAC lookup in the MAC table for the VLAN. SMAC lookup is used to learn (and can trigger a hardware MAC-learn or MAC-move update), DMAC (if present) is used for L2 forwarding and if not present will result in the frame being flooded to all ports within the VLAN, subject to storm-control thresholds for the port.

In the layer 3 (routing) case, the packet processor performs a lookup on the Destination IP address (DIP) within the VRF and if there is a match it knows what port to send the frame to and what packet processor it needs to send the frame to. If the DIP matches a subnet local to the switch for which there is no host route entry, the switch will initiate an ARP request to learn the MAC address for where to send the packet. If there is no matching entry at all the packet is dropped. IP TTL decrement also occurs as part of this stage. Additionally, VXLAN Routing can be performed within a single pass through this stage.

For unicast traffic, the end result from a forwarding lookup match is a pointer to a Forwarding Equivalence Class (FEC) or FEC group (Link Aggregation, Equal Cost Multipathing [ECMP] or Unequal Cost Multipathing [UCMP]). In the case of a FEC group, the fields which are configured for load balancing calculations are used to derive a single matching entry. The final matching adjacency entry provides details on where to send the packet (egress packet processor, output interface and a pointer to the output encapsulation/MAC rewrite on the egress packet processor).

For multicast traffic, the logic is similar except that the adjacency entry provides a Multicast ID, which indicates a replication requirement for both local (ingress) multicast destinations on local ports, as well as whether there are packet processors in the system that require packet replication via multicast replication in the fabric modules. By default, the Arista 7800R3 Series operates in egress multicast replication but can be configured for ingress multicast replication as well.

The forwarding pipeline always remains in the hardware data-plane. There are no features that can be enabled that cause the packet forwarding to drop out of the hardware-based forwarding path. In cases where software assistance is required (e.g. traffic destined within a L3 subnet but for which the switch has not yet seen the end device provides an ARP and doesn't have the L3-to-L2 glue entry), hardware rate limiters and Control Plane Policing are employed to protect the control-plane from potential denial of service attacks.

In parallel with forwarding table lookups, there are also Ingress ACL lookups (Port ACLs, Routed ACLs) for applying security and QoS lookups to apply Quality of Service. All lookups are ultimately resolved using strength-based resolution (some actions are complementary and multiple actions are applied, some actions override others) but ultimately the outcome of this stage is a resolved forwarding action.

Counters available within this stage provide accounting and statistics on ACLs, VLAN, and sub-interfaces, as well as a range of tunnel and next-hop group types. The R3-series line cards provide significant gains in overall counter scale and flexibility in allocation over previous generations, providing a 5X increase in scale in some dimensions. The criticality of flexibility in counter scaling cannot be overstated as operators migrate to next-generation technologies such as Segment Routing and the use of various overlay tunnel technologies that rely upon fine-grained network utilization information to accurately place network workloads.

Data plane counters are available in real-time via streaming telemetry using NetDB to export using gRPC with OpenConfig.

Arista FlexRoute™ Engine

One of the key characteristics of the Arista 7800R3 Universal Spine platform is the FlexRoute™ Engine, an Arista innovation that enables Internet-scale L3 routing tables with significant power consumption savings over legacy IP routing longest prefix match lookups. This in turn enables higher port densities and performance with power and cooling advantages when compared to legacy service provider routing platforms.

Arista's FlexRoute Engine is used for both IPv4 and IPv6 Longest Prefix Match (LPM) lookups without partitioning table resources. It is optimized around the Internet routing table, its prefix distribution and projected growth. FlexRoute enables scale to millions of prefixes, providing headroom for internet table growth for many years.

In addition to large table support, FlexRoute enables very fast route programming and reprogramming (tens of thousands of prefixes per second), and does so in a manner that is non-disruptive to other prefixes while forwarding table updates are taking place.

All Arista 7800R3 Series line cards take advantage of the multi-stage programmable forwarding pipeline to provide a flexible and scalable solution for access control, secure policy-based networking and telemetry in today's cloud networks. ACLs are not constrained by the size of fixed hardware tables but can leverage the forwarding lookup capabilities of the packet processor to trigger a wide range of traffic management actions.

sFlow

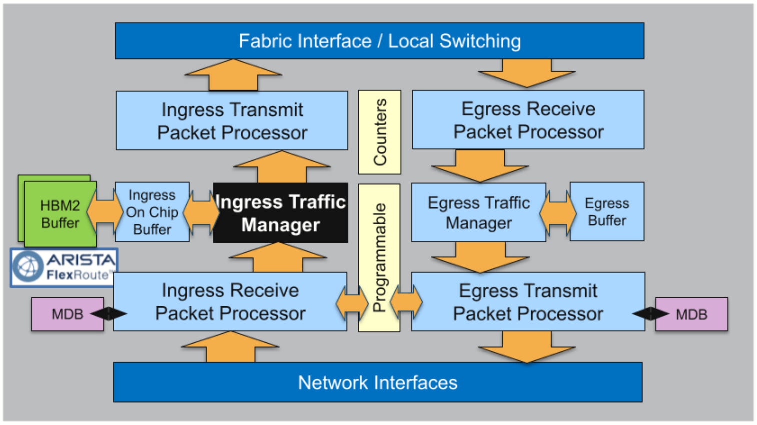
The programmable packet processing pipeline on the 7800R3 platform enables a range of new telemetry capabilities for network operators. In addition to new counter capabilities, flow instrumentation capabilities are enhanced through the availability of hardware-accelerated sFlow. As network operators deploy various tunnel overlay technologies in their network, sFlow provides an encapsulation independent means of getting visibility into high-volume traffic flows and enables operators to more effectively manage and steer traffic to maximize utilization. The programmable pipeline provides these capabilities inline without requiring an additional coprocessor. Sampling granularity of 1:100 on 100G and 400G interfaces can be realized on all interfaces.

Inband Network Telemetry (INT)

As a complement to sFlow, INT provides operators with a standards-based means of getting insight into per-hop latency, paths, congestion, and drops. This information can be correlated to allow an analysis of hotspots, path topology to influence traffic engineering decisions. INT provides operators with a data plane aware complement to standard IP/MPLS troubleshooting tools. Where ping and traceroute cannot necessarily confirm whether or not a flow traverses a specific interface in a port-channel, INT provides operators with a path and node traversal details by processing inband OAM frames and annotating these frames with metadata to provide detailed path and transit quality details. The programmable pipeline in the R3-series line cards provides the ability to facilitate this packet processing inline.

Stage 3: Ingress Traffic Manager

The Ingress Traffic Manager stage is responsible for packet queuing and scheduling.


Fabric Interface / Local Switching
Ingress Transmit Packet Processor
Counters
Egress Receive Packet Processor
HBM2 Buffer
Ingress On Chip Buffer
Ingress Traffic Manager
Programmable
Egress Traffic Manager
Egress Buffer
ARISTA FlexRoute
MDB
Ingress Receive Packet Processor
Egress Transmit Packet Processor
MDB
Network Interfaces

- Virtual Output Queuing (VOQ) subsystem
- Credit request
- On-chip buffer for uncongested output queues
- External buffer for queuing
- Shaping/Queuing
- PFC

Figure 19: Packet Processor stage 3 (ingress): Ingress Traffic Manager

Arista 7800R3 Universal Spine platforms utilize Virtual Output Queuing (VOQ) where the majority of the buffering within the switch is on the input line card. While the physical buffer is on the input packet processor, it represents packets queued on the output side (hence, the term virtual output queuing). VOQ is a technique that allows buffers to be balanced across sources contending for a congested output port and ensures fairness and QoS policies can be implemented in a distributed forwarding system.

When a packet arrives into the Ingress Traffic Manager, a VOQ credit request is forwarded to the egress port processor requesting a transmission slot on the output port. Packets are queued on ingress until such time as a VOQ grant message is returned (from the Egress Traffic Manager on the output port) indicating that the Ingress Traffic Manager can forward the frame to the egress packet processor.

While the VOQ request/grant credit cycle is underway, the packet is queued in input buffers. A combination of on-chip memory (up to 64 MB) and external memory (8 GB) per packet processor is used to store packets while awaiting the VOQ grant. The memory is used such that traffic destined to uncongested outputs (egress VOQ is empty) will go into on-chip memory (head of the queue) otherwise external buffer memory is utilized. The external buffer memory is used because it’s impractical to build sufficiently large buffers on-chip due to the very large chip die area that would be consumed.

While there is up to 384 GB buffer memory per system, the majority of the buffer is allocated in a dynamic manner wherever it is required across potentially millions of VOQs per system:

- ~30% buffer reserved for traffic per Traffic Class per Output Port
- ~15% buffer for multi-destination traffic
- ~55% available as a dynamic buffer pool

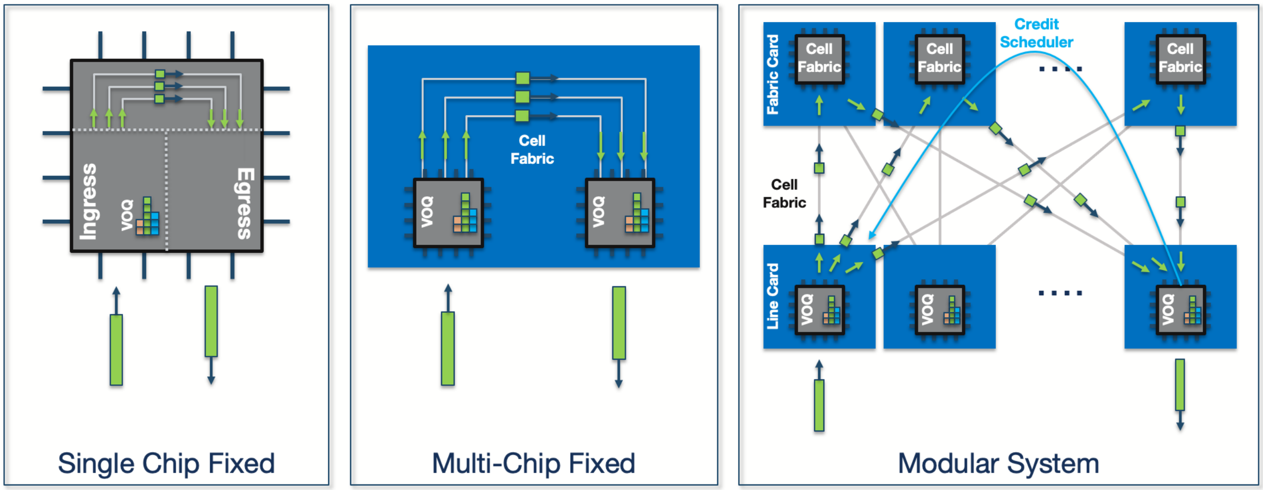


Figure 20: Physical Buffer on Ingress allocated as Virtual Output Queues

The dynamic pool enables the majority of the buffer to be used in an intelligent manner based on real-time contention and congestion on output ports. While there are potentially hundreds of gigabytes of buffer memory, individual VOQ limits are applied such that a single VOQ doesn't result in excess latency or queuing on a given output port. The default allocations (configurable) are as per Table 7:

Table 7: Default per-VOQ Output Port Limits

Output Port Characteristic	Maximum Packet Buffer Depth (MB)	Maximum Packet Buffer Depth (msec)
VOQ for a 10G output port	50 MB	40 msec
VOQ for a 25G output port	125 MB	40 msec
VOQ for a 40G output port	200 MB	40 msec
VOQ for a 50G output port	250 MB	40 msec
VOQ for a 100G output port	500 MB	40 msec
VOQ for a 400G output port	500 MB	10 msec

The VOQ subsystem enables buffers that are dynamic, intelligent, and deep so that there are always packet buffer space available for new flows, even under congestion and heavy load scenarios. There is always complete fairness in the system, with QoS policy always enforced in a distributed forwarding system. This enables any application workload to be deployed – existing or future – and provides the basis for deployment in Content Delivery Networks (CDNs), service providers, internet edge, converged storage, hyper-converged systems, big data/analytics, enterprise, and cloud providers. The VOQ subsystem enables maximum fairness and goodput for applications with any traffic profile, be it any-cast, in-cast, mice or elephant flows, or any flow size in between.

7800R3 Deep Packet Buffers

As with the 7500R Series generations, the 7800R3 series line cards utilize on-chip buffers (32MB with Jericho2, 64MB with Jericho2C+) in conjunction with flexible packet buffer memory (8GB of HBM2 per packet processor). The on-chip buffers are used for non-congested forwarding and seamlessly utilize the HBM2 packet buffers for instantaneous or sustained periods of congestion. Buffers are allocated per VOQ and require no tuning. It's further worth noting that during congestion, packets are transmitted directly from the HBM2 packet buffer to the destination packet processor.

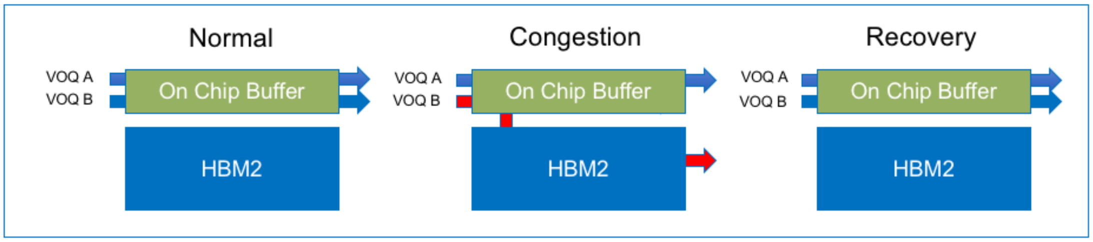


Figure 21: Packet buffer memory access

HBM2 memory is integrated directly into the Jericho2 packet processor this provides a reliable interface to the Jericho2 packet processor and eliminates the need for additional high-speed memory interconnects as does HMC or GDDR. This results in upwards of a 43% reduction in power utilization than the equivalent GDDR memory.

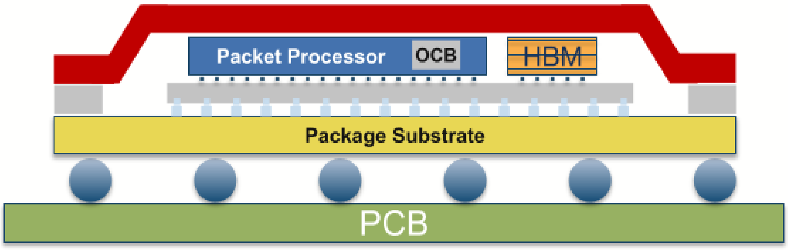


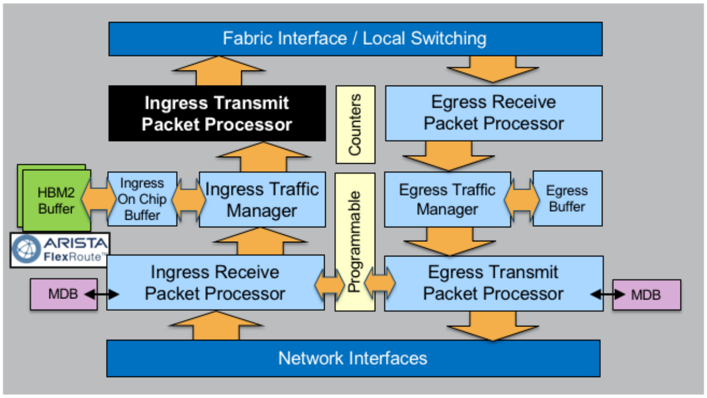
Figure 22: HBM memory packaging integration

Stage 4: Ingress Transmit Packet Processor

The Ingress Transmit Packet Processor stage is responsible for transferring frames from the input packet processor to the relevant output packet processor. Frames arrive at this stage once the output port has signaled, via a VOQ grant message, that it is the allocated slot for a given input packet processor to transmit the packet.
All available fabric paths are used in parallel to transfer the frame or packet to the output packet processor, with the original group of packets on the same VOQ are packed into 256-byte cells which are forwarded across up to 192 fabric links simultaneously. This mechanism reduces serialization to at most 256 bytes at 50Gbps and ensures there are no hot spots as every flow is always evenly balanced across all fabric paths. Since a packet is only transferred across the fabric once there is a VOQ grant, there is no queuing within the fabric and there are guaranteed resources to be able to process the frame on the egress packet processor.

Each cell has a header added to the front for the receiving packet processor to be able to reassemble and maintain in-order delivery. Forward Error Correction (FEC) is also enabled for traffic across the fabric modules, both to correct errors (if they occur) but also to help monitor data-plane components of the system for any problems.

Packets destined to ports on the same packet processor are switched locally and do not use fabric bandwidth resources, but otherwise aren't processed any differently in terms of the VOQ subsystem.

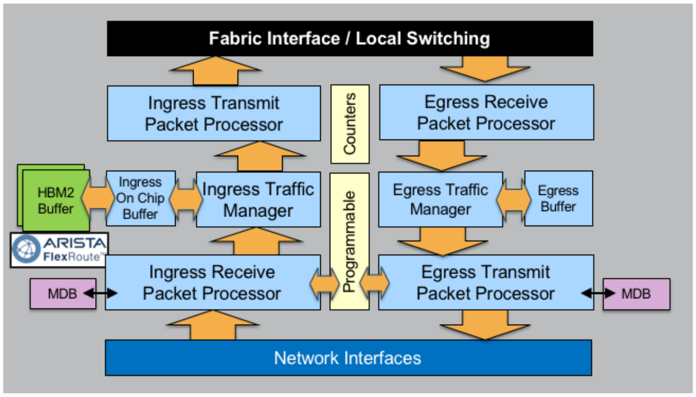

Fabric Interface / Local Switching
Ingress Transmit Packet Processor
Counters
Egress Receive Packet Processor
HBM2 Buffer
Ingress On Chip Buffer
Ingress Traffic Manager
Programmable
Egress Traffic Manager
Egress Buffer
ARISTA FlexRoute
Ingress Receive Packet Processor
Egress Transmit Packet Processor
MDB
MDB
Network Interfaces

- Maps OutLIF to egress packet processor
- Segments packets into cells across fabric

Figure 23: Packet Processor stage 4 (ingress): Ingress Transmit Packet Processor

Stage 5: Fabric Modules

There are 6 fabric modules in the rear of the 4, 8 and 12 slot chassis while the 16 slot chassis uses 12 fabric modules. ll fabric modules operate in an active/active manner. These provide connectivity between all data-plane forwarding packet processors inside the system.


Fabric Interface / Local Switching
Ingress Transmit Packet Processor
Counters
Egress Receive Packet Processor
HBM2 Buffer
Ingress On Chip Buffer
Ingress Traffic Manager
Programmable
Egress Traffic Manager
Egress Buffer
ARISTA FlexRoute
MDB
Ingress Receive Packet Processor
Egress Transmit Packet Processor
MDB
Network Interfaces

- Unicast cells flow to egress destination packet processor
- Multi-destination frames are replicated to each output packet processor

Figure 24: Fabric modules

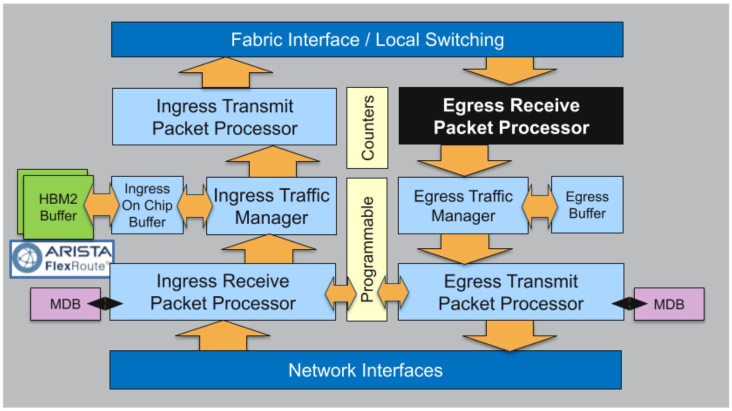
The fabric modules forward based on cell headers indicating which of the 48 possible output packet processors (3 per line card) to send the cell to.

For multi-destination packets such as multicast or broadcast, there is a lookup into a multicast group table that uses a bitmap to indicate which packet processors should receive replicated copies of the cell. Note: if there are multiple multicast receivers on an output packet processor, there is only one copy delivered per output packet processor as there is optimal egress multicast expansion inside the system. Control-plane software maintains the multicast group table based on the fan-out of multicast groups across the system. IP multicast groups that share a common set of output packet processors reuse the same fabric Multicast ID.

For destinations on the same packet processor traffic is locally sent to the local egress receive packet processor.

Stage 6: Egress Receive Packet Processor

The Egress Receive Packet Processor stage is responsible for reassembling cells back into packets/frames. This is also the stage that takes a multicast packet/frame and replicates it there are multiple locally attached receivers on this output packet processor.


Fabric Interface / Local Switching
Ingress Transmit Packet Processor
Counters
Egress Receive Packet Processor
HBM2 Buffer
Ingress On Chip Buffer
Ingress Traffic Manager
Programmable
Egress Traffic Manager
Egress Buffer
ARISTA FlexRoute
MDB
Ingress Receive Packet Processor
Egress Transmit Packet Processor
MDB
Network Interfaces

- Reassemble cells back into frames
- Egress multicast replication for local interfaces
- Egress ACL application

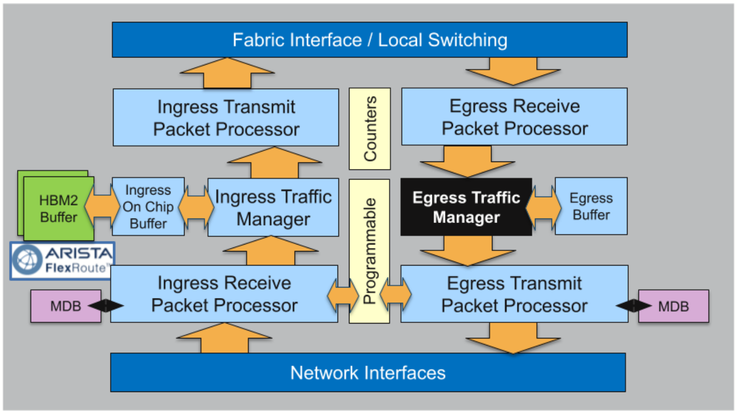
Figure 25: Packet Processor stage 6 (egress): Egress Receive Packet Processor

This stage ensures that there is no frame or packet reordering in the system. It also provides the data-plane health tracer, validating reachability messages from all other packet processors across all paths in the system.

Egress ACLs are also performed at this stage based on the packet header updates, and once the packet passes all checks, it is transmitted on the output port.

Stage 7: Egress Traffic Manager

The Egress Traffic Manager stage is responsible for the granting of VOQ credit requests from input packet processors and managing egress queues.


Fabric Interface / Local Switching
Ingress Transmit Packet Processor
Counters
Egress Receive Packet Processor
HBM2 Buffer
Ingress On Chip Buffer
Ingress Traffic Manager
Programmable
Egress Traffic Manager
Egress Buffer
ARISTA FlexRoute
MDB
Ingress Receive Packet Processor
Egress Transmit Packet Processor
MDB
Network Interfaces

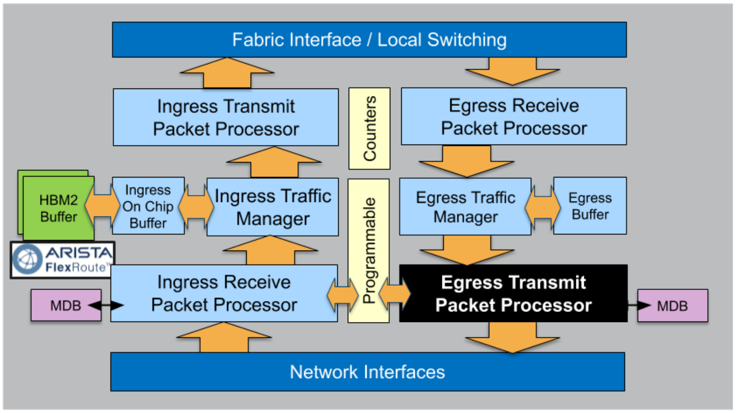
- Manage Egress Queues (unicast & multicast)
- Grant VOQ requests from Ingress
- PFC/ETS traffic scheduling

Figure 26: Packet Processor stage 7 (egress): Egress Receive Packet Processor

When an ingress packet processor requests to schedule a packet to the egress packet processor it is the Egress Traffic Manager stage that receives the request. If the output port is not congested then it will grant the request immediately. If there is congestion it will fairly balance the service requests between contending input ports, within the constraints of QoS configuration policy (e.g. output port shaping) while also conforming to PFC/ETS traffic scheduling policies on the output port. Scheduling between multiple contending inputs for the same queue can be configured to weighted fair queuing (WFQ) or round-robin.

The Egress Traffic Manager stage is also responsible for managing egress buffering within the system. There is an additional 32 or 64 MB on-chip buffer used for egress queuing. This buffer is primarily reserved for multicast traffic as unicast traffic has a minimal requirement for egress buffering due to the large ingress VOQ buffer and fair adaptive dynamic thresholds are utilized as a pool of buffer for the output ports.

Stage 8: Egress Transmit Packet Processor


Fabric Interface / Local Switching
Ingress Transmit Packet Processor
Counters
Egress Receive Packet Processor
HBM2 Buffer
Ingress On Chip Buffer
Ingress Traffic Manager
Programmable
Egress Traffic Manager
Egress Buffer
ARISTA FlexRoute
MDB
Ingress Receive Packet Processor
Egress Transmit Packet Processor
MDB
Network Interfaces

- Application of egress packet header rewrite actions
- TCP ECN marking
- Tunnel Encapsulation

Figure 27: Packet Processor stage 8 (egress): Egress Transmit Packet Processor

In this stage, any packet header updates such as updating the next-hop DMAC, Dot1q updates and tunnel encapsulation operations are performed based on packet header rewrite instructions passed from the Input Receive Packet Processor stage. Decoupling the packet forwarding on ingress from the packet rewrite on egress provides the ability to increase the next-hop and tunnel scale of the system as these resources are programmed in a distributed manner.

This stage can also optionally set TCP Explicit Congestion Notification (ECN) bits based on whether there was contention on the output port and the time the packet spent queued within the system from input to output. Flexible Counters are available at this stage and can provide packet and byte counters on a variety of tables.

Stage 9: Network Interface (Egress)

Just as packets/frames entering the switch went through the Ethernet MAC and PHY layer with the flexibility of multi-speed interfaces, the same mechanism is used on packet/frame transmission. Packets/frames are transmitted onto the wire as a bitstream in compliance with IEEE 802.3 standards.

Arista EOS: A Platform for Scale, Stability and Extensibility

At the core of the Arista 7800R3 Universal Spine platform is Arista EOS® (Extensible Operating System). Built from the ground up using innovative core technologies since our founding in 2004, EOS contains more than 8 million lines of code and years of advanced distributed systems software engineering. EOS is built to be open and standards-based and its modern architecture delivers better reliability and is uniquely programmable at all system levels.

EOS has been built to address two fundamental issues that exist in cloud networks: the need for non-stop availability and the need for high feature velocity coupled to high-quality software. Drawing on our engineers' experience in building networking products over more than 30 years, and on the state-of-the-art in open systems technology and distributed systems, Arista started from a clean sheet of paper to build an operating system suitable for the cloud era.

At its foundation, EOS uses a unique multi-process state-sharing architecture that separates system state information from packet forwarding and from protocol processing and application logic. In EOS, system state and data is stored and maintained in a highly efficient System Database (SysDB). The data stored in SysDB is accessed using an automated publish/subscribe/notify model. This architecturally distinct design principle supports self-healing resiliency in our software, eases software maintenance, and enables module independence. This results in higher software quality overall and accelerates time-to-market for the new features that customers require.

Arista EOS contrasts with the legacy approach to building network operating systems developed in the 1990s that relied upon embedding system state within each independent process, relying on extensive use of inter-process communications (IPC) mechanisms to maintain state across the system, with a manual integration of subsystems. These legacy system architectures lack an automated structured core like SysDB. In legacy network operating systems, as dynamic events occur in large networks or in the face of a system process failure and restart, recovery can be difficult if not impossible.

Additionally, as legacy network operating systems attempt to adapt to industry demands, such as streaming telemetry, individual subsystems must be manually extended to support state export into a system that was never designed to facilitate cloud-scale export mechanisms. As such, stabilizing and adapting to a wide range of telemetry and control protocols remains an ongoing challenge complicating integration and delaying migration to next-generation management interfaces for operators.

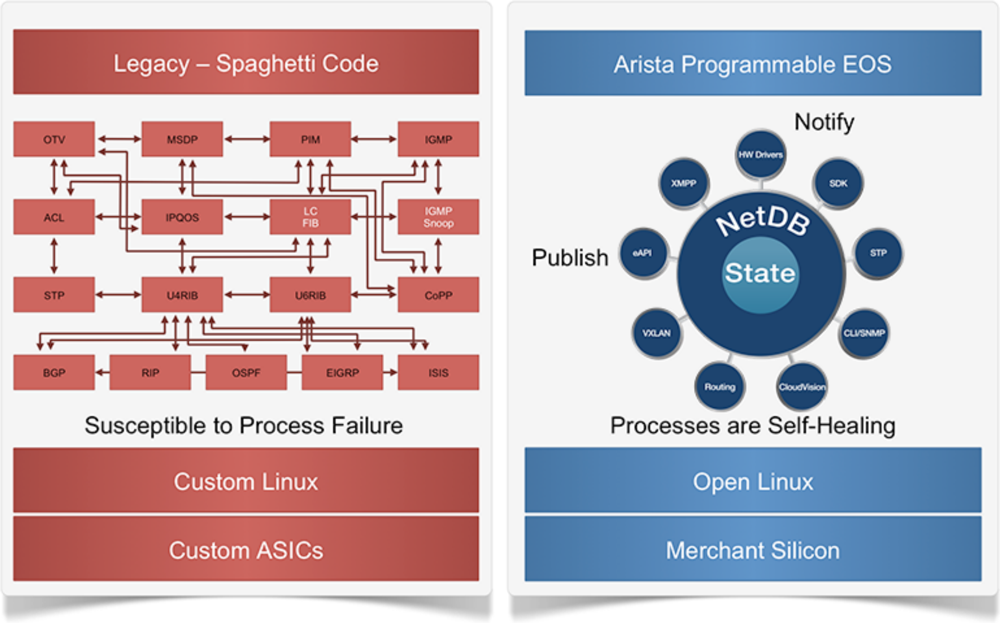


Figure 28: Legacy approaches to network operating systems (left), Arista EOS (right)

Arista took to heart the lessons of the open-source world and built EOS on top of an unmodified Linux kernel maintaining full, secured access to the Linux shell and utilities. This allows EOS to utilize the security, feature development, and tools of the vibrant Linux community on an on-going basis. This is in contrast to legacy approaches where the original OS kernel is modified or based on older and less well-maintained versions of Unix. This has made it possible for EOS to natively support things like Docker Containers to simplify the development and deployment of applications on Arista switches. Arista EOS represents a simple but powerful architectural approach that results in a higher quality platform on which Arista is able to continuously deliver significant new features to customers.

EOS is extensible with open APIs available at every level: management plane, control-plane, and data-plane. Service-level and application-level extensibility can be achieved with access to all Linux operating system facilities including shell-level access. Arista EOS can be extended with Linux applications and a growing number of open-source management tools to meet the needs of network engineering and operations.

Open APIs such as EOS API (eAPI), OpenConfig and EOS-SDK provide well-documented and widely used programmatic access to configuration, management, and monitoring that can stream real-time network telemetry, providing a superior alternative to traditional polling mechanisms.

The NetDB evolution of SysDB extends the core EOS architecture in the following ways:

- NetDB NetTable enables EOS to scale to new limits. It scales the routing stack to hold millions of routes or tunnels with sub-second convergence.
- NetDB Network Central enables system state to be streamed and stored as historical data in a central repository such as CloudVision, HBase, or other third-party systems. This ability to take network state and efficiently and flexibly export it, is crucial for scalable network analysis, debugging, monitoring, forensics, and capacity planning. This simplifies workload orchestration and provides a single interface for third party controllers.
- NetDB Replication enables state streaming to a variety of telemetry systems in a manner that automatically tolerates failures, and adapts the rate of update propagation to match the capability of the receiver to process those updates.

The evolution of SysDB to NetDB builds on the core principles that have been the foundation of the success of EOS: openness, programmability, and quality on a single build of EOS runs across all of our products.

System Health Tracer and Integrity Checks

Just as significant engineering effort has been invested in the software architecture of Arista EOS, the same level of detail has gone into system health and integrity checks within the system. There are numerous subsystems on Arista 7800R3 Universal Spine platform switches that validate and track the system health and integrity on a continual basis:

- All memories where code executes (control-plane and data-plane) are ECC protected; single bit errors are detected and corrected automatically, double bit errors are detected.
- All data-plane forwarding tables are parity protected with shadow copies kept in ECC protected memory on the control-plane. Continual hardware table validation verifies that the hardware tables are valid and truthful.
- All data-plane packet buffers are protected using CRC32 checksums from the time a packet/frame arrives, and at the time it leaves the switch. The checksum is validated at multiple points through the forwarding pipeline to ensure no corruption has happened, or if there has been a problem, rapidly facilitate its isolation.
- Forward Error Correction (FEC) is also utilized for traffic across the fabric modules, both to correct errors (if they occur) but also to help monitor data-plane components of the system for problems.
- Data-plane forwarding elements are continually testing and checking reachability with all other forwarding elements in the system. This is to ensure that if there are issues they can be accurately and proactively resolved.

Conclusion

Designed to address the demands of the world's largest cloud and service providers the Arista 7800R3 Series modular switches continue to provide operators with a proven, industry-leading, platform to evolve their network capabilities. By combining industry-leading 400G density with Internet-scale service capabilities and next-generation packet processing functionality at the optimum intersection of performance and power utilization.

The 7800R3 leverages the proven architecture that has made the previous generations of the product so successful; focus on efficient system design, reliability, and flexibility. This trend continues with innovations in the packet processors powering the R3-series, enabling operators to use the 7800R3 in an ever wider range of roles with a single hardware platform.

Arista's EOS network operating system continues to lead the industry in openness, extensibility, and software quality. EOS has been leading the industry in telemetry innovations through the availability of NetDB and enabled operators to truly automate their network deployments through rich programmatic interfaces and support for industry standards such as OpenConfig.

Given the cloud-scale hardware and software capabilities of the 7800R3, it makes the ideal platform for a range of applications. The 7800R3 is ideally suited for cloud-scale data centers, Service Provider WAN backbones, and Peering edges as well as large enterprise networks.

Santa Clara—Corporate Headquarters
5453 Great America Parkway,
Santa Clara, CA 95054

Phone: +1-408-547-5500
Fax: +1-408-538-8920
Email: info@arista.com

Ireland—International Headquarters
3130 Atlantic Avenue
Westpark Business Campus
Shannon, Co. Clare
Ireland

Vancouver—R&D Office
9200 Glenlyon Pkwy, Unit 300
Burnaby, British Columbia
Canada V5J 5J8

San Francisco—R&D and Sales Office
1390 Market Street, Suite 800
San Francisco, CA 94102

India—R&D Office
Global Tech Park, Tower A, 11th Floor
Marathahalli Outer Ring Road
Devarabeesanahalli Village, Varthur Hobli
Bangalore, India 560103

Singapore—APAC Administrative Office
9 Temasek Boulevard
#29-01, Suntec Tower Two
Singapore 038989

Nashua—R&D Office
10 Tara Boulevard
Nashua, NH 03062