Arista 7500R3 Platform Architecture

Figure 1: Arista 7500R3 Universal Spine platform

Arista Networks' award-winning Arista 7500 Series was introduced in April 2010 and established itself as a revolutionary switching platform, which maximized data center performance, efficiency, and overall network reliability. It raised the bar for switching performance, being five times faster, one-tenth the power draw, and one-half the footprint compared to other modular data center switches available at the time.

Since its introduction, the Arista 7500 Series has consistently delivered nearly a three-fold increase in density and performance with each generation, without sacrificing on features and functionality and with complete investment protection. Arista has delivered continuous improvements in performance and density with significant increases in features and functionality. Evolving from a best-in-class Cloud-scale switching to an Internet-scale, service provider routing platform. This white paper provides an overview of the architecture of the Arista 7500R3 Universal Spine platform.

Arista 7500R3: Platform Overview

The Arista 7500R3 Universal Spine Platform represents the evolution of the 7500R family of modular switches available in 4-slot, 8-slot and 12-slot form factors that support a range of line card options.

At a system level, the Arista 7508R3 with a 153Tbps fabric scales to 192 x 400G, 288 x 100G (QSFP100) in 13 RU providing industry-leading performance and density without compromising on features/functionality or investment protection.

Table 1: Arista 7500R3 Key Port and Forwarding Metrics

Characteristic	Arista 7504R3	Arista 7508R3	Arista 7512R3
Chassis Height (RU)	7 RU	13 RU	18 RU
Linecard Module slots	4	8	12
Supervisor Module slots	2	2	2
50G Maximum Density	288	576	864
100G Maximum Density (QSFP100)	144	288	576
400G Maximum Density (OSFP or QSFP-DD)	96	192	288
System Usable Capacity (Tbps)	76 Tbps	153 Tbps	230 Tbps
Max forwarding throughput per Linecard (Tbps)	9.6 Tbps (DCS-7500R3-24P) (24 x 400G per LC)		
Max forwarding throughput per System (Tbps)	76.8 Tbps (FD)	153.6 Tbps (FD)	230.4 Tbps (FD)
Max packet forwarding rate per Linecard (pps)	4 Billion pps (7500R3-24P)		
Max packet forwarding rate per System (pps)	16 Bpps	32 Bpps	48 Bpps
Maximum Buffer memory/ System	64 GB	128 GB	192 GB
Virtual Output Queues / System	More than 2.2 million		

Investment Protection

The Arista 7500R3 Universal Spine platform represents the fourth generation of the Arista 7500 Series – Arista has continued to provide investment protection to existing Arista 7500 Series deployments. Existing chassis, power supplies, fans, 7500R-series line cards and fabric modules can continue to be used alongside newer 7500R3 Series line cards.

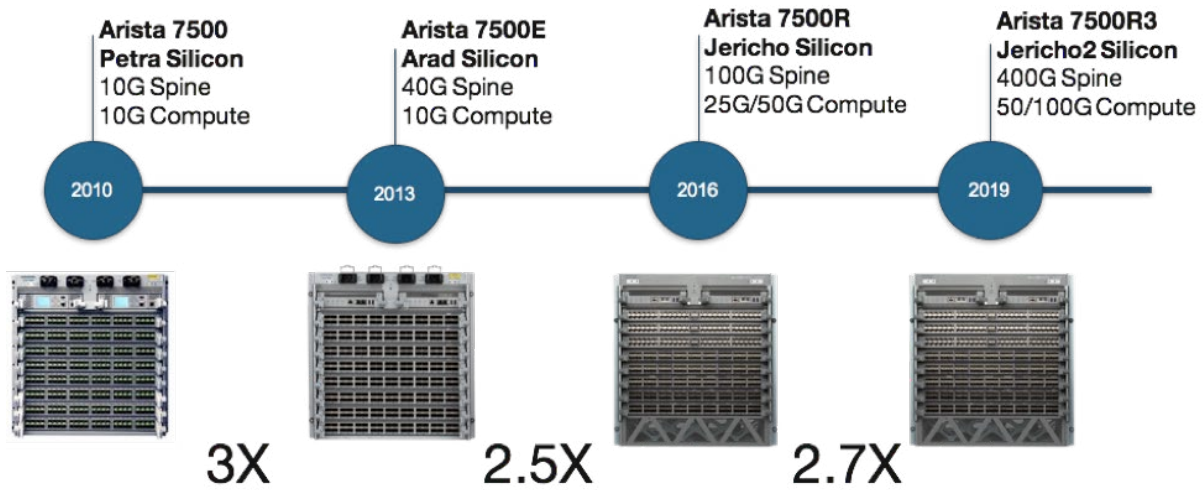

Figure 2: Arista 7500 Series Platform Generations

For new deployments, backward interoperability with existing components may appear less important. However, what it represents is a platform that is feature-rich with proven high-quality day one, leveraging thousands of person-years of software development on the single OS and platform, accelerating qualification and deployment as well as time to revenue for operators.

Arista 7500R3 - Cloud Scale And Features

Each iteration of the packet processor silicon in the Arista 7500 Series – from the first-generation in 2010 (Petra / 7500) to the latest (Jericho2 / 7500R3) has been riding Moore's Law. This observation, that on average, there will be two times more transistors available every two years has held true for decades. The network silicon within the Arista 7500 Series has used those additional transistors to more than double performance and density at each generation, from 80Gbps per packet processor (8 x 10G interfaces) to 4.8Tbps (12 x 400G interfaces).

In addition to delivering increased port scale and performance, forwarding table sizes have continued to increase. The 7500R family provides operators with the scale of Internet Arista's innovative FlexRoute™ Engine continues to expand forwarding capacity beyond what the merchant silicon enables natively and the Arista EOS NetDB™ evolution of SysDB enables increased scale and performance and industry leading routing convergence.

With the 7500R3 this innovation continues. The 7500R3-series introduces the Modular Database (MDB) to enable the 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 7500R3 platform. These resources are allocated using forwarding profiles that ensure the optimal allocation 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.

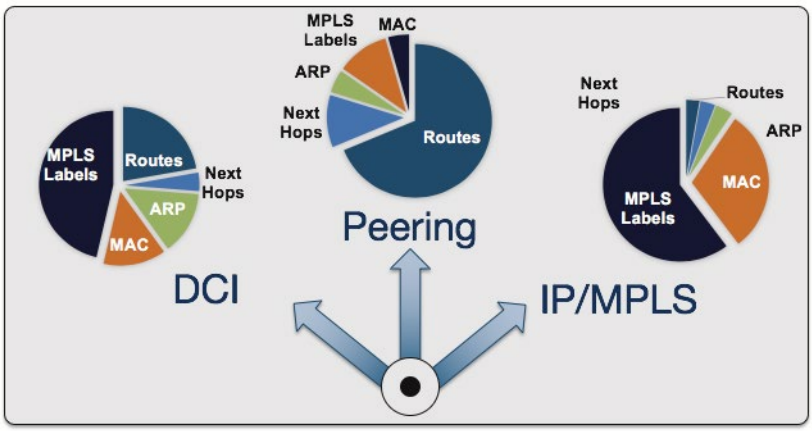

Figure 3: MDB enables a flexible range of deployment profiles.

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 allocated according to the needs of any given role. There is no need to have a separate platform for core network roles and edge roles in today's service provider networks. This enables cloud and service providers to streamline their deployments, simplify sparing and consolidate testing.

Arista 7500R3: System Components

Chassis And Mid-Plane

All Arista 7500R3 chassis (4-slot, 8-slot and 12-slot) share a common system architecture with identical fabric bandwidth and forwarding capacity per slot. Line cards and power supplies are common across all systems; the only differences are in the size of the fabric/fan modules, a half-width or full-width supervisor module, number of line card slots and power supplies. Airflow is always front-to-rear and all data cabling is at the front of the chassis.

Chassis design and layout are a key aspect that enables such high performance per slot: the fabric modules are directly behind line card modules and oriented orthogonal to the line card modules. This design alleviates the requirement to route high-speed signal traces on the midplane of the chassis, reducing the signal trace lengths and enabling more high-speed signals to operate at faster speeds by being shorter lengths. This characteristic has enabled Arista to scale the system from 10 Tbps with first-generation modules in 2010 up to 230 Tbps in 2019 - a 23X performance increase in 9 years.

Supervisor Modules

Supervisor modules on the Arista 7500R3 Universal Spine platform are used for control-plane and management plane functions only. There are two redundant supervisors in the chassis, each capable of managing the system. All data-plane forwarding is performed online card modules and forwarding between line card modules is always via the fabric modules.

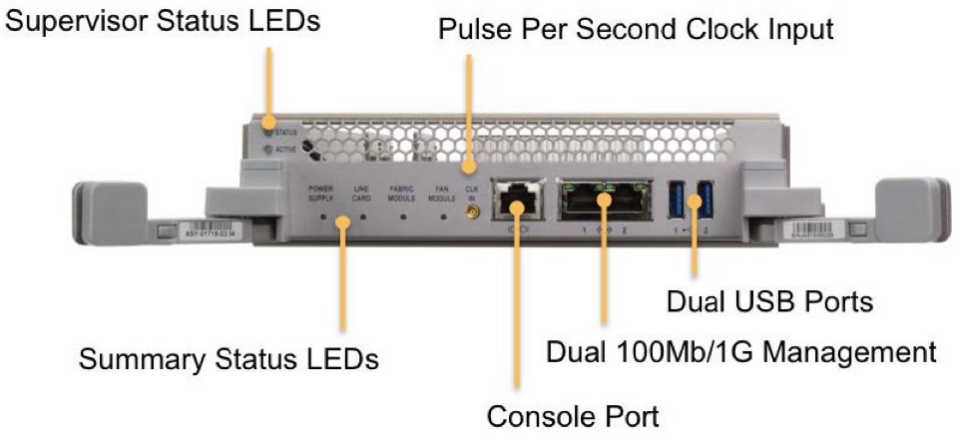

Figure 4: Arista 7500 Series Supervisor 2 Module.

The Arista 7500 Series Supervisor 2 provides a 6-core hyper-threaded Intel Xeon 'Broadwell' CPU with turbo boost frequency to 2.7 GHz and 32GB RAM and is used with the 7504R3, 7508R3, and 7512R3 systems.

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 optional 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. There are two USB2.0 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 40 Gbps of inband connectivity from data-plane to control-plane and more than 30 Gbps 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 7500R3: Distributed Packet Forwarding

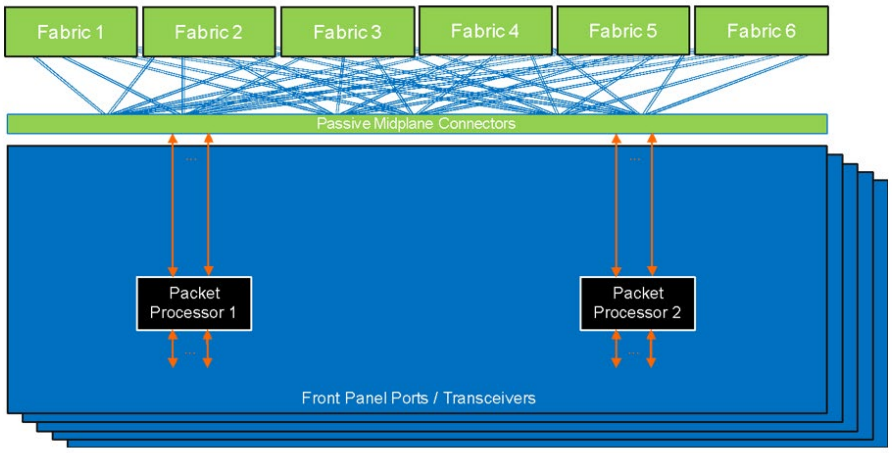

Figure 5: Distributed Forwarding within an Arista 7500R3 Series

Arista 7500R3 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 7500R3 up to six fabric modules are utilized in an active/active mode. Each fabric module provides up to 3.2 Tbps fabric bandwidth full-duplex (1.6 Tbps receive + 1.6 Tbps transmit) to each line card slot, and with six active fabric modules, there is 19.2 Tbps full duplex bandwidth (9.6 Tbps receive and 9.6 Tbps transmit) available. With six active fabric modules in a 7512R3 system, 230 Tbps (115 Tbps transmit and 115 Tbps receive) of fabric bandwidth is available. If a fabric module were to fail, the throughput of the fabric degrades gracefully. Fabric modules support hot swap and may be inserted or removed while the system is in operation.

Figure 6: Arista DCS-7500R3 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 of the update.
- **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.

Arista 7500R3: Line Card Architecture

Arista 7500R3 Universal Spine Platform Line Card Layout

Arista 7500R3 line card modules utilize the same Jericho2 packet processor, 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) are 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 interfaces that can run from 10G to 50G and individually or combined in groups from 10G to 50G to allow flexible 10G, 25G, 40G, 50G 100G, 200G and 400G interfaces.

There are 96 PAM4 lanes and each packet processor supports up to a maximum of 96 logical or physical interfaces per chip. The maximum logical or physical interfaces per packet processor set the maximum 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 additional lanes at lower speeds and with a different encoding.

Gearboxes enable systems to increase the choice of interfaces without requiring additional packet processors, reducing overall system power consumption and heat generation while also increasing reliability.

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.

Table 2: Arista 7500R3 Series Line card Module Port Characteristics[1]

Line card	Port (type)	Interfaces*							Port Buffer	Fowarding Rate	Switching Capacity
		10G	25G	50G	40G	100G	400G	Max§			
7500R3-24P	24 OSFP	192	96	192	-	96	24	192	16GB	4.0 Bpps	9.6Tbps
7500R3-24D	24 QSFP-DD	192	96	192	-	96	24	192	16GB	4.0 Bpps	9.6Tbps
7500R3-36CQ	36 QSFP100	96	96	72	36	36	-	96	8GB	2.0 Bpps	3.6Tbps

[1]The switching operations for previous generations of 7500R series line cards are very similar. Additional details regarding operation and logical system scale for these previous line cards can be found online in the Arista 7500R Switch Architecture white paper.

* Maximum port numbers are uni-dimensional, may require the use of break-outs and are subject to transceiver/cable capabilities.

§ 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-7500R3-24P-LC

All stages associated with packet forwarding are performed in an integrated system on chip (SoC) packet processor. 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-7500R3-24P-LC, a 24-port 400G OSFP module, is shown below in Figure 7. Each of the packet processors on the line card services a group of front panel ports.

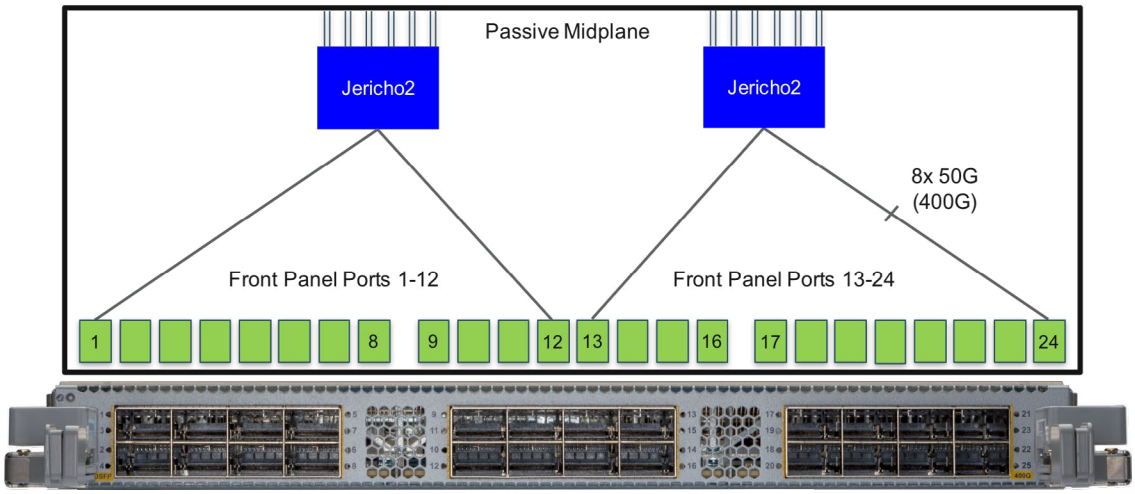

Figure 7: Arista DCS-7500R3-24P-LC module architecture

In the case of the DCS-7500R3-24P-LC, each of the 24 ports can operate as either 1 x 400G, 2x200G, 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-7500R3-24D-LC

The DCS-7500R3-24D-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-7500R3-36CQ-LC

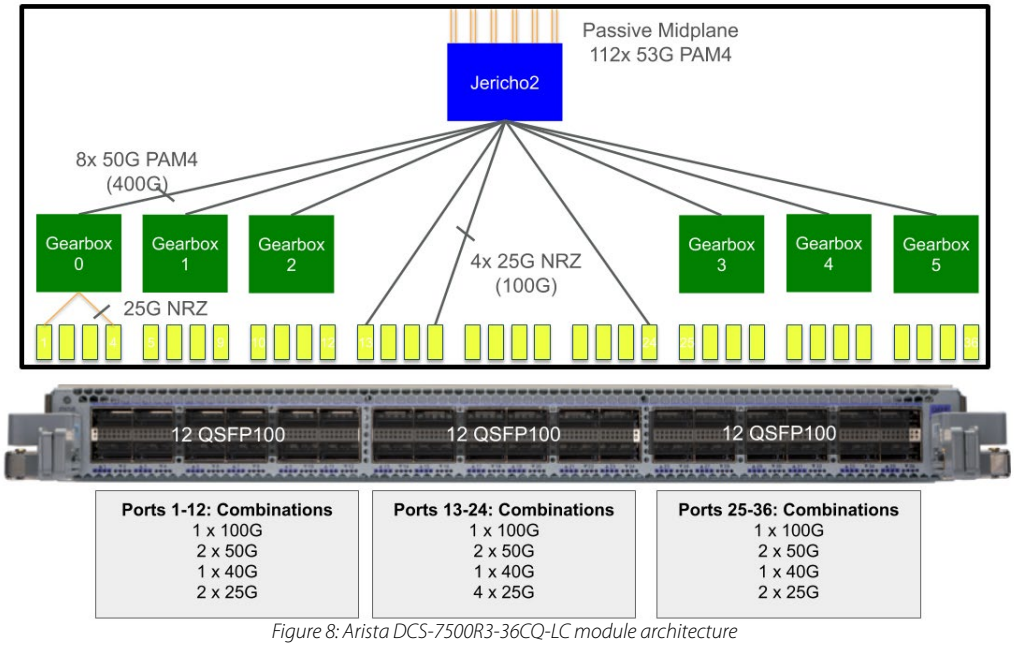

Figure 8: Arista DCS-7500R3-36CQ-LC module architecture

The DCS-7500R3-36CQ line card utilizes a single Jericho2 chip and provides 6 gearboxes to support a diverse range of optics and cables for up to 96 individual interfaces when breakouts are used.

All ports are capable of operating in 1 x 100G, 1 x 40G, 2 x 50G, 2 x 25G mode.

When using 4-way breakouts (e.g. 4 x 25G or 4 x 10G) the following limitations apply:

- Ports 13 - 24
  - » Both odd and even ports can operate in breakout mode.
- Ports 1- 12 and 25 - 36
  - » Odd-numbered ports support breakout (even-numbered port disabled)

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. In the case of 25G/50G, SR4 and PSM4 optics support breakout (as they are optically 4 x 25G in parallel), likewise Direct Attach Copper (DAC) QSFP100 to 4 x SFP25 cables provide breakout support.

Interoperability with previous line card generations

The 7500R3 Series Universal Spine supports a wide range of line cards including the 7500R, 7500R2 and 7500R3 Series. Each line card generation leverages a common architecture, with different generations of the packet processor, that offer consistently higher throughput and larger scale forwarding resources in combination with a deep packet buffer and VoQ architecture.

Table 3: Arista 7500R Series Logical Resources, Scale and Performance					
Packet Processor	7500R3K[2] (Jericho2)	7500R3[2] (Jericho2)	7500R2K (Jericho+)	7500R2 (Jericho+)	7500R (Jericho)
Bandwidth	4.8T	4.8T	900G	900G	600G
Density	96x25G 48x100G 12 x 400G	96x25G 48x100G 12 x 400G	36x25G 48x10G	36x25G 48x10G	24x25G / 60x10G
Performance	2.0 Bpps	2.0 Bpps	837 Mpps	837 Mpps	720 Mpps
Buffer	8GB - HBM2	8GB - HBM2	4GB - GDDR5	4GB - GDDR5	4GB - GDDR5
Resource Tables					
MAC Addresses	736K	448K	768K	768K	768K
IPv4 Host Routes	1.4M	896K	768K	768K	768K
IPv4 Unicast LPM Routes	1.2M	704K	2M	1.3M	1M+
IPv6 Host Routes	368K	224K	768K	768K	768K
IPv6 Unicast LPM Routes	411K	235K	2M	1.3M	1M
Multicast Routes	448K	448K	Up to 768K	Up to 768K	Up to 768K
ACL Entries	24K	24K	24K	24K	24K
Algorithmic ACLs	100K	100K	24K	24K	24K

[2] Represents a balanced profile for the partitioning of the MDB

Arista 7500R3: Packet Forwarding Pipeline

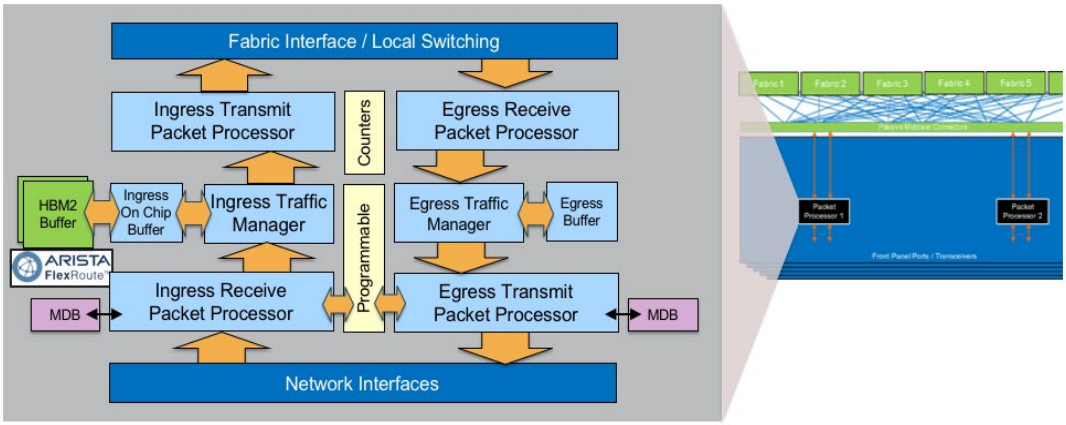

Figure 9: Packet forwarding pipeline stages inside a packet processor on an Arista 7500R3 line card module

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 for 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).

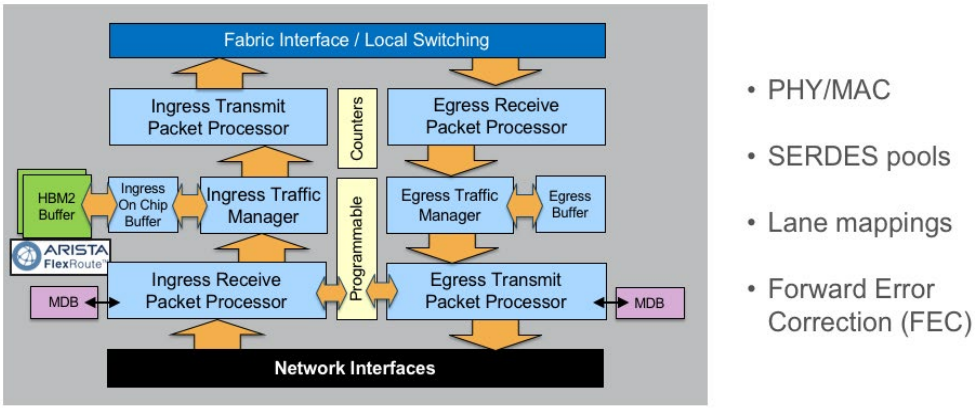

Figure 10: Packet Processor stage 1 (ingress): Network Interface

The PHY layer is responsible for 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 4x 25G and 2x 50G on 100G ports.

If a valid bit stream is received at the PHY then the data is sent to the MAC layer. On input, the MAC layer is responsible for turning the bit stream 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.

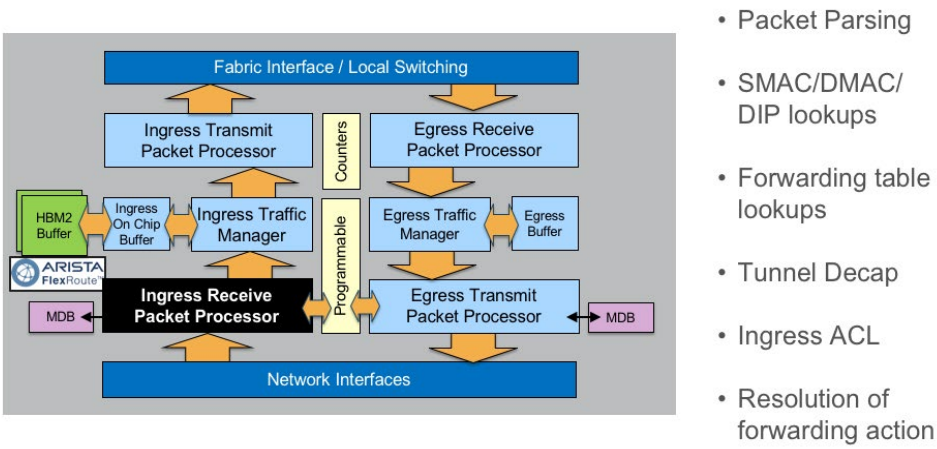

Figure 11: Packet Processor stage 2 (ingress): Ingress Receive Packet Processor

Before any forwarding can take place, packet or frame headers must be parsed and fields for forwarding decisions extracted. Key fields include L2 Source and Destination MAC addresses [SMAC, DMAC], VLAN headers, Source and Destination IP Addresses [SIP, DIP], class of service (COS), DSCP, and so on. The Arista 7500R3 packet parser supports many tunnel formats (MPLS, IPinIP, GRE, VXLAN, MPLSoGRE) as well as parsing Ethernet and IP headers under a multi-label stack. The parser is flexible and extensible such that it can support future protocols and new forwarding models. As IPv6 deployments grow and traffic shifts to being native IPv6 it is anticipated that more IPv4 traffic will be tunneled over the aforementioned tunnel formats using IPv6. Jericho2 provides the additional parsing flexibility necessary to enable this next generation of tunnel formats.

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 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 7500R3 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 provide 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 are 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 7500R3 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 beyond 1 million IPv4 and IPv6 prefixes combined, 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 7500R3 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 7500R3 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.

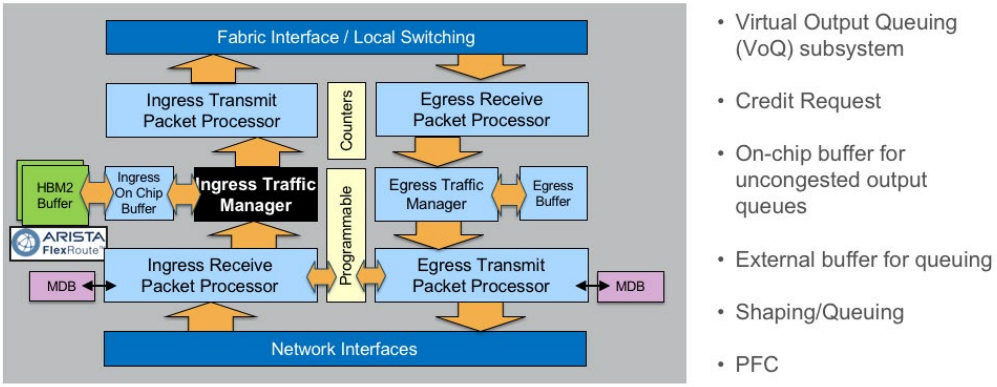

Figure 12: Packet Processor stage 3 (ingress): Ingress Traffic Manager

Arista 7500R3 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 (32MB) and external memory (8GB) 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 192GB 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

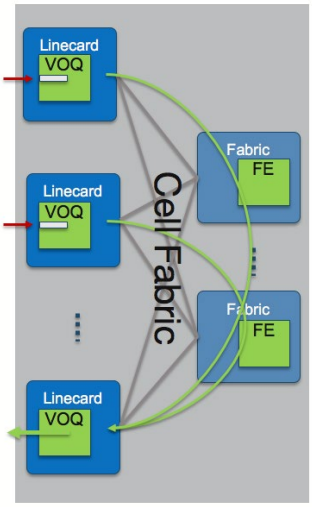

Figure 13: Physical Buffer on Ingress allocated as Virtual Output Queues

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 is 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 4:

Table 4: 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.

7500R3 Deep Packet Buffers

As with previous generations, the 7500R3 series line cards utilize on-chip buffers (32MB with Jericho2) 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.

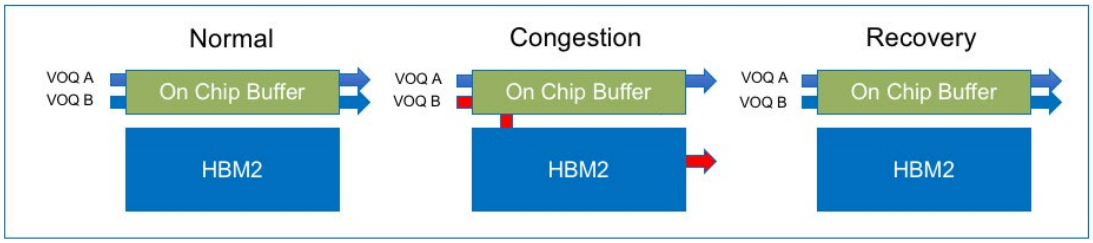

Figure 14: Packet buffer memory access

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.

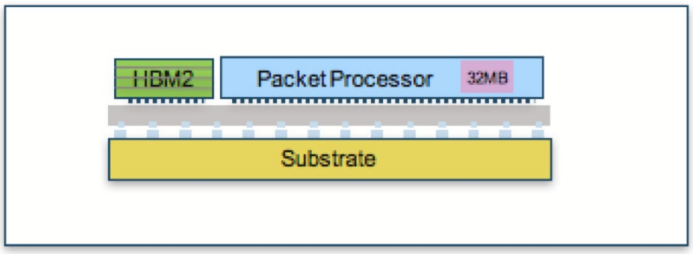

Figure 15: HBM memory packaging integration

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.

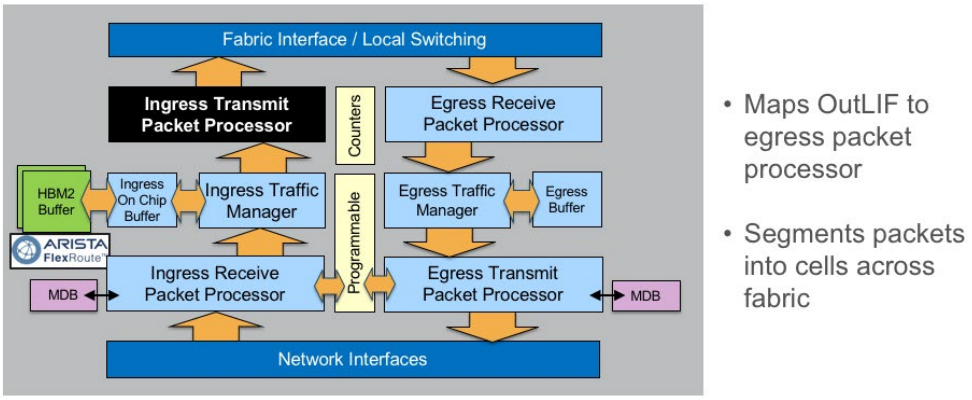

Figure 16: Packet Processor stage 4 (ingress): Ingress Transmit Packet Processor

All available fabric paths are used in parallel to transfer the frame or packet to the output packet processor, with the original packet segmented into variable-sized cells which are forwarded across up to 112 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.

Stage 5: Fabric Modules

There are 6 fabric modules in the rear of the chassis all operating in an active/active manner. These provide connectivity between all data-plane forwarding packet processors inside the system.

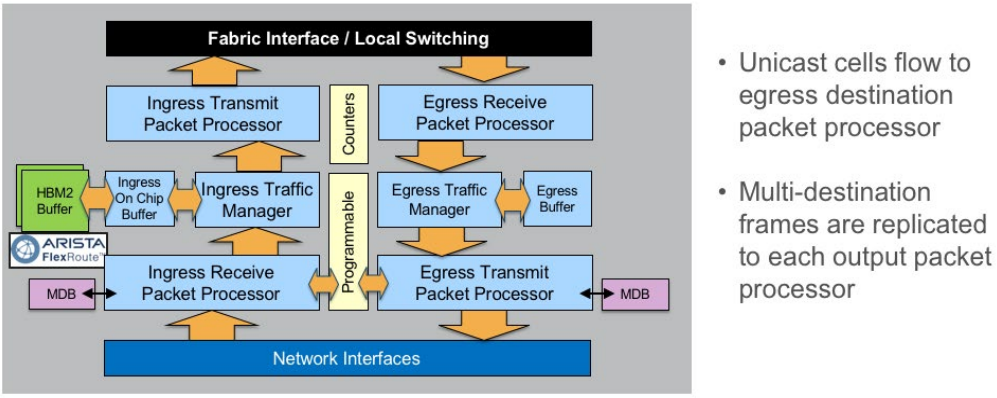

Figure 17: Fabric modules

The fabric modules forward based on cell headers indicating which of the 24 possible output packet processors 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.

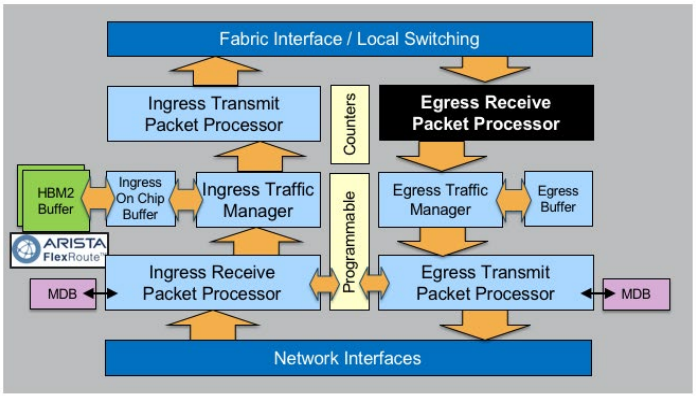

Figure 18: Packet Processor stage 6 (egress): Egress Receive Packet Processor

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.

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.

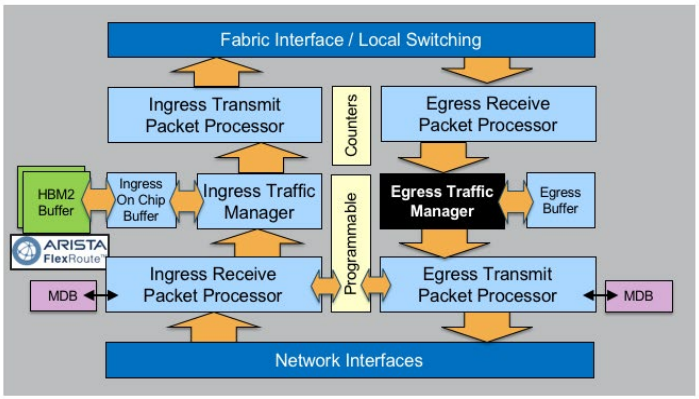

Figure 19: Packet Processor stage 7 (egress): Egress Receive Packet Processor

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 32MB 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

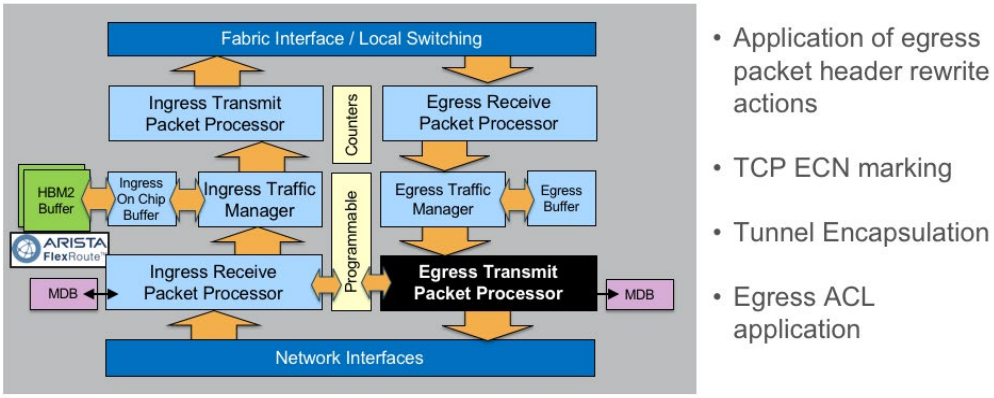

Figure 20: Packet Processor stage 8 (egress): Egress Transmit Packet Processor

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.

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 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 bit stream in compliance with IEEE 802.3 standards.

Arista EOS: A Platform for Scale, Stability and Extensibility

At the core of the Arista 7500R3 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 1990's 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 which 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.

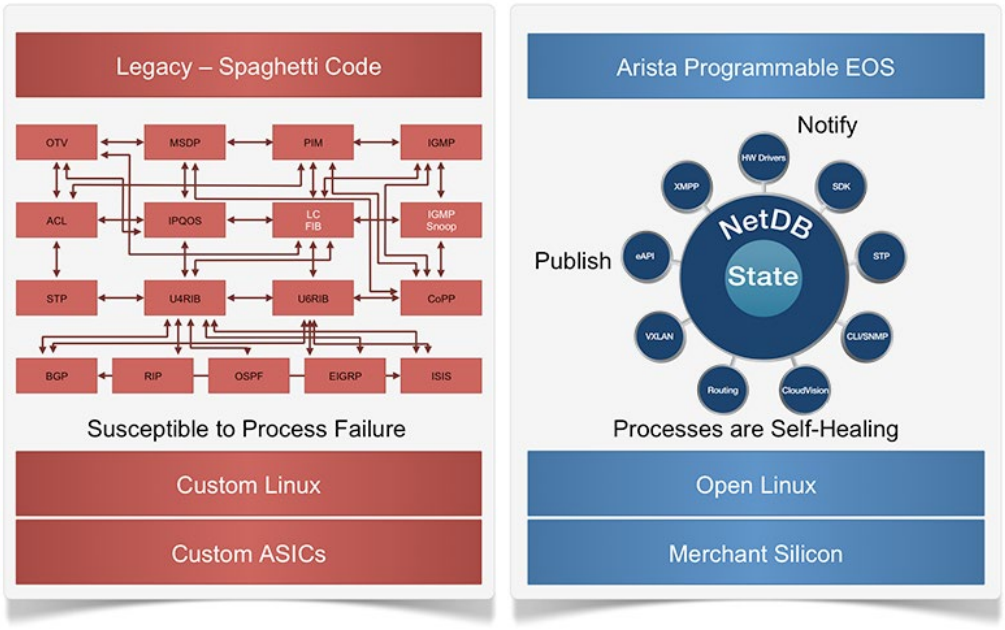

Figure 28: Legacy approaches to network operating systems (left), Arista EOS (right)

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 7500R3 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 to 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 7500R3 Series modular switches continues 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 7500R3 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 7500R3 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 7500R3, it makes the ideal platform for a range of applications. The 7500R3 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 & B, 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