Arista 7280R Switch Architecture (‘A day in the life of a packet’)  

The Arista Networks 7280R Series is a series of purpose built high performance fixed configuration 1RU and 2RU form factor Universal Leaf switches with a deep buffer, virtual output queue architecture combined with rich features, and support for 10/25/40/100G ports and standards-based 25G and 50G for long term investment protection. The 7280R Series includes all models in 7280R, 7280RA, 7280R2, 7280R2K and 7280R2A.

The 7280R capabilities address the demands of modern networking and rich multimedia content delivery, requiring a lossless forwarding solution. They are ideal for Universal Cloud Networks, where deep buffers and wire speed L2 and L3 forwarding are combined with advanced features for network virtualization, open and programmable monitoring and network analysis, resiliency and architectural flexibility.

Additionally, the deep packet buffers and 7280R series’ support for highly scalable IPv4 and IPv6 tables allow for a wide range of open networking solutions, including Cloud WAN aggregation, Service Provider NFV, Internet Peering and Peering Exchanges, Overlay Networks, Content Delivery and evolution of the network edge, in addition to flexible scale out of the Arista UCN.

The broad range of interfaces and density choices provide comprehensive deployment choices.

Figure 1: Arista 7280R Universal Leaf Platform
The 7280R has a number of differentiators compared to other fixed configuration systems. The major advantages come from the VoQ and deep buffer architecture (up to 32GB per system) combined with the rich EOS feature set.

- Standards-based high density 100G switch ideal for future proof designs and high bandwidth needs, such as in High Performance Storage or Content Delivery Networks (CDNs).
- Ultra-deep buffers in a fixed form factor makes for an ideal top-of-rack switch where lossless performance and in-cast problems are expected, such as in big data analytics, search and IP storage.
- Directly connected 25GbE, 40GbE and 50GbE attached storage systems, requiring high performance and predictable latency.
- Streaming network state for advanced analytics with Arista CloudVision®
- Network-wide virtualization platform for next generation cloud bursting with wire-speed VXLAN routing.
- Hardware-assisted PTP enables accurate timing solutions across Ethernet-based networks, without costly investment in separate timing networks.
- Unique monitoring and provisioning features – LANZ, DANZ, AEM, IEEE 1588 PTP, ZTP, VM Tracer, VXLAN, and eAPI
- Comprehensive L2 and L3 feature set for open multi-vendor networks with no proprietary lock-in
- Large scale L2 and L3 table resources allow deployment flexibility in both large L2 and L3 environments with any-workload suitability.
- NEBS compliance and DC power supplies designed for service provider environments
- MACsec encryption for simple, reliable and scalable datacenter interconnect and for securing links between tiers in leaf and spine datacenter designs.

The 7280R platform is designed for lossless behavior, in environments where large scale routing, VXLAN Routing, DANZ and enhanced LANZ are needed. It is ideal for networks where 100G uplinks are planned, for investment protection, and any place where system scalability is a concern. In addition, all members of the 7280R Series support full internet scale peering and have advanced features such as Precision Time Protocol (PTP) and Audio Video Bridging (AVB).

- **DANZ TAP Aggregation** - 10/25/40/100G TAP Aggregation with best-in-class performance and high density up to 60 ports of 100G TAP/Tool ports
- **Latency Analyzer (LANZ)** - Microsecond granularity on port utilization using buffering watermarks to provide immediate feedback and precision monitoring
- **Wirespeed VXLAN Routing** - Seamless integration between VXLAN and L2/L3 environments, physical and virtualized networks
- **IEEE 1588 PTP** - Build and scale accurate timing solutions with sub-microsecond accuracy
- **Network Wide Virtualization** - Multi-vendor API Support with eAPI, VXLAN and NSX, and other encapsulation techniques
- **AlgoMatch™** – Efficient packet matching algorithm that enables flow matching for access control, policy and visibility
- **sFlow Acceleration** - Advanced network telemetry, capacity planning, security analysis and quality of experience monitoring
- **The 100G QSFP interfaces on the 7280R can be used for various port speed options, depending on the transceivers used, where the 100G ports can be used to break out into 4 x 25G or 2 x 50G. The use of 40G QSFP+ optics can provide 4 x 10G ports.**
System Specifications
The tables below provide system specifications for all the 7280R platforms.

<table>
<thead>
<tr>
<th>Table 1: Arista 7280R and 7280RA Port Combinations and Forwarding Metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max 100GbE Ports</td>
</tr>
<tr>
<td>Max 50GbE Ports</td>
</tr>
<tr>
<td>Max 40GbE Ports</td>
</tr>
<tr>
<td>Max 25GbE Ports</td>
</tr>
<tr>
<td>Max 10GbE Ports</td>
</tr>
<tr>
<td>L2/3 Throughput</td>
</tr>
<tr>
<td>L2/3 PPS</td>
</tr>
<tr>
<td>Latency</td>
</tr>
<tr>
<td>Typical Power Draw</td>
</tr>
<tr>
<td>Total System Buffer</td>
</tr>
<tr>
<td>Rack Units</td>
</tr>
<tr>
<td>Airflow</td>
</tr>
<tr>
<td>Power Supplies</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2: Arista 7280R2 and 7280R2A Port Combinations and Forwarding Metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max 100GbE Ports</td>
</tr>
<tr>
<td>Max 50GbE Ports</td>
</tr>
<tr>
<td>Max 40GbE Ports</td>
</tr>
<tr>
<td>Max 25GbE Ports</td>
</tr>
<tr>
<td>Max 10GbE Ports</td>
</tr>
<tr>
<td>L2/3 Throughput</td>
</tr>
<tr>
<td>L2/3 PPS</td>
</tr>
<tr>
<td>Latency</td>
</tr>
<tr>
<td>Typical Power Draw</td>
</tr>
<tr>
<td>Total System Buffer</td>
</tr>
<tr>
<td>Rack Units</td>
</tr>
<tr>
<td>Airflow</td>
</tr>
<tr>
<td>Power Supplies</td>
</tr>
</tbody>
</table>
Arista 7280R - Router Table Scale, Features and Functionality

In addition to increasing port numbers and system performance, forwarding table sizes have continued to grow. Arista’s innovative FlexRoute™ Engine enables more than a million IPv4 and IPv6 route prefixes in hardware, significantly beyond what merchant silicon enables natively. Arista EOS NetDB™ evolution of SysDB allows for increased scale and performance with industry leading routing convergence, creating the first fixed switch system to truly be called a router. Extensions to FlexRoute for switches with larger on-chip tables increases the capability to over 2M routes, of both IPv4 and IPv6, with the ability to contain multiple full-route table copies and ensure many years of investment protection.

The table below shows the key scale metrics of the 7280R, 7280RA, 7280R2, 7280R2A and 7280R2K Series.

Table 3: Arista 7280R Key L2, L3 Scale Metrics

<table>
<thead>
<tr>
<th></th>
<th>7280R Series (Jericho)</th>
<th>7280RA Series (Jericho)</th>
<th>7280R2 Series (Jericho+)</th>
<th>7280R2A Series (Jericho+)</th>
<th>7280R2K Series (Jericho+)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAC Table Size</td>
<td>768K</td>
<td>768K</td>
<td>768K</td>
<td>768K</td>
<td>768K</td>
</tr>
<tr>
<td>v4/v6 Host Routes</td>
<td>768K</td>
<td>768K</td>
<td>768K</td>
<td>768K</td>
<td>768K</td>
</tr>
<tr>
<td>IPv4 Route Prefixes</td>
<td>Up to 1.2M</td>
<td>Up to 1.2M</td>
<td>Up to 1.3M</td>
<td>Up to 1.3M</td>
<td>2M+</td>
</tr>
<tr>
<td>IPv6 Route Prefixes</td>
<td>Up to 1M</td>
<td>Up to 1M</td>
<td>Up to 1.3M</td>
<td>Up to 1.3M</td>
<td>2M+</td>
</tr>
<tr>
<td>Multicast Groups</td>
<td>Up to 768K</td>
<td>Up to 768K</td>
<td>Up to 768K</td>
<td>Up to 768K</td>
<td>Up to 768K</td>
</tr>
<tr>
<td>LAG Groups</td>
<td>1152</td>
<td>1152</td>
<td>1152</td>
<td>1152</td>
<td>1152</td>
</tr>
<tr>
<td>LAG Members</td>
<td>64 Ports</td>
<td>64 Ports</td>
<td>64 Ports</td>
<td>64 Ports</td>
<td>64 Ports</td>
</tr>
<tr>
<td>ECMP Fanout</td>
<td>128-way</td>
<td>128-way</td>
<td>128-way</td>
<td>128-way</td>
<td>128-way</td>
</tr>
<tr>
<td>ACL Entries</td>
<td>N/A</td>
<td>24K</td>
<td>N/A</td>
<td>24K</td>
<td>24K</td>
</tr>
</tbody>
</table>

Airflow

The 7280R Series offers a choice of airflow direction (front-to-rear or rear-to-front). Power and fan modules are color-coded to show the airflow direction through the switch, making it easy to identify how the switch should be installed in a rack where Hot/Cold aisle separation is required.

Red handles shown below denote hot aisle placement with Airflow exit from the switch.

Figure 2: 7280R 1RU & 2RU Switch Chassis Rear View
Blue handles shown below denote cold aisle placement with Airflow intake to the switch.

![Figure 3: 7280R 1RU & 2RU Switch Chassis Rear View](image)

**Arista 7280R Universal Leaf Platform Architecture**

All 7280R Series switches share a common system design built around a high performance x86 CPU for the control plane. The CPU is connected to system memory, internal flash, SSD, bootflash, power supplies, fans, management I/O and peripherals, as shown below.

![Figure 4: 7280R High Level Architecture](image)

The Arista 7280R switches were designed for continuous operations with system wide monitoring of both hardware and software components, simple serviceability and provisioning to prevent single points of failure.

Key high availability features include:

- 1+1 hot-swappable power supplies and four hot-swap fans provide dynamic temperature control combined with N+1 redundancy
- Color coded PSU's and fans that deliver platinum level power efficiency
- Live software patching
- Self healing software with Stateful Fault Repair (SFR)
- Smart System Upgrade (SSU)

The x86 CPU is also connected over PCIe to the 7280R Packet Processor(s) that run all the data plane forwarding and have all the directly connected front panel ports.

All stages associated with packet forwarding are performed in integrated system on chip (SoC) packet processors. The packet processor provides both the ingress and egress packet forwarding pipeline stages for packets that arrive or are destined to 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 switch (in this case, the 48 port SFP25 w/ 6 port QSFP100 switch 7280SR2-48YC6) is shown in the figure below. Each of the packet processors on the switch services a group of 24 x 10/25G SFP and 3x 100G QSFP100 front panel ports.

![Figure 5: Arista 7280SR2-48YC6/7280SR2A-48YC6 switch architecture](image)

In the case of the 7280SR2-48YC6, each of the 48 ports supports 1G SFP, 10G SFP+ or 25G SFP transceivers. The 6 QSFP ports can operate as either 1 x 100G or breakout to 2 x 50G or 4 x 25G when populated with a 100G QSFP transceiver. If populated with a 40G QSFP+ transceiver, each port can operate as either 1 x 40G or breakout to 4 x 10G. All port speed configuration changes are hitless and no ports are disabled or unavailable.

In order to operate a port in breakout mode (e.g. run a 100G port as 4 x 25G or 2 x 50G, or a 40G port as 4 x 10G), the underlying transceiver must enable this. In the case of 25G/50G, SR4 and PSM4 optics support breakout (as they are optically 4x25G in parallel). Similarly, Direct Attach Copper (DAC) QSFP100 to 4xSFP25 cables provide breakout support. The same holds true for 4x10G breakout of a 40G port, which is available for transceivers and cables that are based on 4x10G parallel signaling.

**Port Identification**

The 40GbE QSFP+ and 100GbE QSFP100 transceivers are the same physical size and can be inserted into any QSFP based ports. It’s important that switches indicate the port capabilities on the front panel so that an engineer installing transceivers inserts them into the correct port types.

The figure below shows the two types of QSFP ports. Ports on the right are highlighted with a lilac line to identify they are QSFP100 capable; these ports will support either QSFP+ or QSFP100 transceivers. The ports on the left without the lilac decal are QSFP+ only and will not support QSFP100. The LED indicators below the ports are grouped in sets of four. The left set are for the top port and the right for the lower port. When ports are configured for either 40G or 100G, the leftmost LED indicator will show green when connected, with the others indicating orange. When the ports are configured for 10G or 25G, each of the 4 LEDs will be green when the associated link is connected. For 50G connectivity on the QSFP100 ports, LEDs 1 and 3 will indicate the link state with 2 and 4 remaining orange.

![Figure 6: QSFP+ & QSFP100 Ports](image)
The pluggable transceiver used for 25GbE is an SFP28 and is the same size as both the 1GbE SFP and 10GbE SFP+. To facilitate easy identification of 25GbE capable ports a green color outline is used to indicate 25GbE capable SFP ports as shown below.

**Arista 7280R Universal Leaf Platform Layout**

Arista 7280R series switches utilize high performance packet processors, with the number of packet processors varied, based on the number and type of ports on the system. The packet forwarding architecture of each of these systems is essentially the same: a group of front-panel ports (different transceiver/port/speed options) are connected to the packet processor. The following diagrams show the layout of each 7280R system.

As shown above, the 7280SR-48C6 and 7280SRA-48C6 use a single packet processor and provide customers with 48 x 10G SFP+ ports allowing for 1/10G on each port and 6 x QSFP100, for connection to spine switches, located in the center of the front panel. The QSFP ports will accept either QSFP100 or QSFP+ optics and cables and can be configured for a choice of 10/25/40/50/100G depending on the transceiver. Out of band management of the switch is available through the use of either a 100M/1G RJ45 Ethernet interface or a Serial port on the rear of the switch. Two USB ports are also available, one on the rear and another on the front. Each system has a choice of 8GB or 32GB of DRAM, and the larger memory comes with an SSD for local storage. AlgoMatch is provided on the 7280SRA model.

The 7280TR-48C6 and 7280TRA-48C6 also use a single packet processor and provide customers with a 10GBASE-T solution with 48 x 100/1000/10GbE RJ45 ports and 6 x QSFP100 ports for connection to spine switches, located in the center of the front panel. The QSFP ports will accept either QSFP100 or QSFP+ optics and cables and can be configured for a choice of 10/25/40/50/100G, depending on the transceiver. Out of band management of the switch is available through the use of either a 100M/1G RJ45 Ethernet interface or a Serial port on the rear of the switch. Two USB ports are also available, one on the rear and another on the front. Each system has a choice of 8GB or 32GB of DRAM, and the larger memory comes with an SSD for local storage. AlgoMatch is provided on the 7280SRA model.
Similarly, the 7280SRAM-48C6 is built with single packet processor and provides 48 x 10G SFP+ ports allowing for 1/10G on each port and 6 x QSFP100. All the QSFP ports support built-in wire speed 100G MACsec encryption that is purpose built for open and secure high performance cloud environments and data center interconnects (DCI). The QSFP ports can also be used as 40G QSFP+ ports without MACsec. Each system is built with 32GB of DRAM and AlgoMatch support.

The 7280QR-C36, shown on the left, is a 36 port QSFP-based switch, using two packet processors with 8GB of buffer. All ports are 40G capable, with 12 ports also capable of 100G mode. All ports are capable of operating at the lower speeds of 4x10G and the 100G ports enable both 25G and 50G modes. Ports are arranged with the QSFP+ ports in two 12-port banks numbered from 1-12 and 13-24. The 12 QSFP100 ports are labeled from 25-36 and are located in the center of the front panel. Other ports on the front panel include a 100M/1G Ethernet interface for out of band management, a serial RJ45 connection and a USB port.

Shown in the right of figure 10 above, the 7280QRA-C36S is also a 36 port QSFP based switch using two packet processors and 8GB of buffer. All ports are 40G capable, and the system is optimized for 40/100G deployment with up to 12 ports capable of 100G operation. In a common configuration with 30 ports of 40G and 6 ports of 100G, all ports are active. Additional QSFP100 ports can be enabled in 100G mode, with the adjacent QSFP port disabled, allowing up to a total of 12 ports of 100G and 18 ports of 40G, for a total of 30 active ports. In a 10G deployment scenario, 30 of the QSFP ports can be enabled in 4 x 10G mode with the remaining 6 ports of 100G (or 40G) active. Additional 100G ports are enabled when the adjacent QSFP ports are disabled.

All the 100G (QSFP100) capable ports support 10/25/50G modes in addition to 40G and 100G speeds. In the set of 24 QSFP ports numbered from 7-30, all the odd numbered ports are capable of QSFP100. Ports 7, 9 and 11 together with 25, 27 and 29 are always 100G capable. Enabling 100G mode on an odd port in the range 13-23 disables the adjacent (higher) even port in the range 14-24, as shown below.

Out of band management of the switch is available through the use of either a 100M/1G RJ45 Ethernet interface or a Serial port on the rear of the switch. Two USB ports are also available, one on the rear and another on the front.

Each of the systems has a choice of 8GB or 32GB of DRAM, and the larger memory comes with an SSD for local storage.
The 7280QR-C72 is a 72 QSFP port 2U switch, incorporating 4 packet processors and a single fabric chip. All 72 ports are 40G capable, for a high density aggregation switch or edge router. Up to 16 of the 72 ports are capable of 100G, and support is provided for high density 10G with up to 160 ports. Enabling a QSFP port in 100G mode automatically disables the associated neighbor 40G port such that when all 16 100G ports are active, the system has a balance of 40 ports of 40G also active. Breaking out 10G mode is supported on 24 of the QSFP+ ports and all 16 of the 100G capable ports, a total of 40 ports, for up to 160 x 10G per system. The sixteen QSFP100 ports also support both 25G and 50G modes.

Port numbering is arranged with the 1-36 on the upper half and 37-72 on the lower half. Ports capable of being operated in a 4X mode have 4 LEDs, and ports that only support a 1X mode have a single LED. The 16 QSFP100 capable ports are labeled and identified by front panel color marking and are the odd numbered ports 11-25 and 47-61. Other ports on the front panel include a 100M/1G Ethernet interface for out of band management, a serial RJ45 connection and a USB port.

The 7280CR-48 is a 48 QSFP100 port 2U switch with an additional 8 ports of 40G QSFP+, for a total of 56 ports. The 7280CR-48 incorporates 8 packet processors and dual fabric chips. All 48 QSFP100 ports are 40G/100G capable, and the 8 additional 40GbE QSFP+ ports allows up to 56 ports of 40G in total. All ports can be used in 4X mode allowing up to 224 ports of 10G in total and a flexible combination with 10/25/40/100G modes.
All 56 ports are capable of breaking out to 4x10G with the 48 QSFP100 ports capable of 25/50G modes. Port numbering is arranged with the QSFP100 ports 1-48, and the QSFP+ are numbered from 49-56. Other ports on the front panel include a 100M/1G Ethernet interface for out of band management, a serial RJ45 connection and a USB port. The system includes 16G of DRAM 32GB of packet buffer and an optional SSD.

**Arista 7280R2 Universal Leaf Platform Layout**

The Arista 7280R2 series of switches utilize the higher performance 7280R2 packet processor, and the total number varies based on the quantity and type of ports on the system. The packet forwarding architecture of each of these modules is essentially the same as the 7280R Series.

The 7280SR2-48YC6 and 7280SR2A-48YC6 provide customers with 48 x 10/25G SFP ports allowing for 1G/10G/25G on each port and 6 x QSFP100, for connection to spine switches, located in the center of the front panel. The systems use a pair of packet processors and 8GB of packet buffer. The QSFP ports will accept either QSFP100 or QSFP+ optics and cables and can be configured for a choice of 10/25/40/50/100G depending on the transceiver. Out of band management of the switch is available through the use of either a 100/1G RJ45 Ethernet interface or a Serial port on the rear of the switch. Two USB ports are also available, one on the rear and another on the front.

The 7280CR2-60, 7280CR2A and 7280CR2K-60 are all 60 port QSFP100 2U switches, with all ports 40G capable. Each system leverages 6 packet processors and 2 fabric chips for non-blocking performance.

All QSFP100 ports are capable of breaking out to 4x10G or 4x25G, 2x50G modes. Ports are numbered with the 1-30 on the top section and 31-60 on the lower section. Other ports on the front panel include a 100M/1G Ethernet interface for out of band management, a serial RJ45 connection and a USB port.

Figure 16: 7280SR2-48YC6 and 7280SR2A-48YC6 Architecture

Figure 17: 7280SR2-48YC6 and 7280SR2A-48YC6 Front Panel

Figure 18: 7280CR2-60, 7280CR2A-60 and 7280CR2K-60 Architecture
The 7280CR2A-30 and 7280CR2K-30 are 30 port QSFP100 1U switches, with all ports 40G capable. Each system leverages 3 packet processors and a single fabric chip for non-blocking performance.

All QSFP100 ports are capable of breaking out to 4x10G or 4x25G, 2x50G modes. Ports are numbered 1-30 from left to right. Other ports on the front panel include a 100M/1G Ethernet interface for out of band management, a serial RJ45 connection and a USB port.

**Arista 7280R Port Layout**

All members of the Arista 7280R series utilize high performance packet processors, with the number of packet processors based on the quantity and type of ports on the module. Each packet processor has dedicated high capacity external packet buffer memory, available for the ports attached to the packet processor. The tables below identify how the ports are arranged across the forwarding engines.

**Table 4: 7280R Port Configurations**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Chip-1</td>
<td>1-6 (100G)</td>
<td>1-18</td>
<td>1-12 (40G)</td>
<td>1-6 (40G)</td>
<td>1-48 (10G)</td>
<td>1-48 (10G)</td>
</tr>
<tr>
<td></td>
<td>49 (40G)</td>
<td></td>
<td>25-30 (100G)</td>
<td>Odd 7-17 (100G)</td>
<td>49-54 (100G)</td>
<td>49-54 (100G)</td>
</tr>
<tr>
<td>Chip-2</td>
<td>7-12 (100G)</td>
<td>19-36</td>
<td>13-24 (40G)</td>
<td>31-36 (40G)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>50 (40G)</td>
<td></td>
<td>31-36 (100G)</td>
<td>Odd 19-29 (100G)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Even 20–39 (40G)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chip-3</td>
<td>13-18 (100G)</td>
<td>37-54</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>51 (40G)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chip-4</td>
<td>19-24 (100G)</td>
<td>55-72</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>52 (40G)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chip-5</td>
<td>25-30 (100G)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>53 (40G)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chip-6</td>
<td>31-36 (100G)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>54 (40G)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chip-7</td>
<td>37-42 (100G)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>55 (40G)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chip-8</td>
<td>43-48 (100G)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>56 (40G)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
A central CPU complex on the 7280R Series switches is used exclusively for control-plane and management functions; all data-plane forwarding logic occurs at the packet processor level.

Arista EOS®, the control-plane software for all Arista switches, executes on multi-core x86 CPUs with multiple gigabytes of DRAM. As EOS is multi-threaded, runs on a Linux kernel and is extensible, the large RAM and fast multi-core CPUs provide for operating an efficient control plane with headroom for running 3rd party software, either within the same Linux instance as EOS or within a guest virtual machine.

Out-of-band management is available via a serial console port and/or the 10/100/1000 Ethernet management interface. The 7280R Series also offers USB2.0 interfaces that can be used for a variety of functions including the transferring of images or logs.

**Arista 7280R Universal Leaf Platform Packet Forwarding Pipeline**

Each packet processor on a switch 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.

**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).
The PHY layer is responsible for transmission and reception of bit streams 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 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 decision. It is the stage where all forwarding lookups are performed.

Before any forwarding can take place, first 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 7280R packet parser supports many tunnel formats (MPLS, IPinIP, GRE, VXLAN, MPLS0GRE) including 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 methods of forwarding.
Following parsing, 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 through to. If the DIP matches a subnet local on this switch but 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. 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, whatever fields are configured L2/L3/L4 load balancing 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, indicating expansion for both local (ingress) multicast destinations on local ports, whether there are other packet processors that require a copy and if so, what packet processors they are (via multicast expansion in the fabric modules). By default, Arista 7280R Series operates in egress multicast expansion but can be configured for ingress multicast expansion too.

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 silicon (hardware) data-plane 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.

Flexible Counters within this stage provide accounting and statistics on ACLs, VLAN and sub-interfaces, as well as next hop groups. Counters are updated sub-second via a scalable batch update and is available as telemetry that can be streamed in real-time.

---

Due to pending litigation, the operation of ACLs as described in this document is not implemented for systems distributed within the US after July 3rd 2017. For more information and alternative options, please liaise with an Arista representative.
Arista FlexRoute™ Engine

One of the key characteristics of the Arista 7280R Universal Spine platform is the FlexRoute™ Engine, an Arista innovation which enables more than a million IPv4 and IPv6 route prefixes in hardware. This enables deployments in Internet edge/peering use-cases where historically traditional edge routers would have been required. Arista FlexRoute enables large L3 routing tables with significant power consumption savings over existing ways that IP routing longest prefix match lookups are performed. This in turn enables higher port densities and performance with power and cooling advantages due to more efficient transistor count and activity factor reduction compared to alternatives.

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.

Arista AlgoMatch™ Engine

Arista 7280RA and 7280R2A Series switches include support for AlgoMatch. AlgoMatch is a unique Arista innovation that enables more flexible and scalable solutions for access control, secure policy based networking and telemetry in today’s cloud networks. By combining general purpose memory technology with advanced software algorithms, Arista AlgoMatch is able to provide more scale, performance and efficiency compared to standard implementation of merchant silicon based solutions.

AlgoMatch provides a more efficient packet matching algorithm for flow matching. The net benefits are a high performance network policy engine with both increased functionality and scale that is delivered in a cost and power efficient manner.

Figure 20: Packet Processor stage 2 (ingress): Ingress Receive Packet Processor with AlgoMatch
The AlgoMatch Engine resolves both ingress and egress ACLs in a single policy engine extending policy flexibility.

Stage 3: Ingress Traffic Manager

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

Arista 7280R Universal Spine platforms utilize Virtual Output Queuing (VoQ) where the majority of the buffering within the switch is on the ingress port. While the physical buffer is on the input, it represents packets queued on the output side (hence the name virtual output queuing). VoQ is a technique that allows buffers to be balanced across sources contending for a congested output port. It ensures fairness and allows for QoS policies to 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 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.

The operation of AlgoMatch is implemented as described for all 7280R systems. For more information on AlgoMatch, please liaise with an Arista representative.
While the VoQ request/grant credit loop is under way, the packet is queued in input buffers. A combination of onboard memory (16MB) and external memory (4GB) per packet processor is used to store packets while waiting for their VoQ grant. The memory is used such that traffic destined to uncongested outputs (egress VoQ is empty) will go into onboard memory (head of the queue). Otherwise, external buffer memory is utilized. External buffer memory is used because it’s not feasible to build sufficiently large buffers “on-chip”, as that would require a large area and number of transistors.

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

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

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 in Table 6:

<table>
<thead>
<tr>
<th>Output Port Characteristic</th>
<th>Maximum Packet Queue Depth</th>
<th>Maximum Packet Buffer Depth (MB)</th>
<th>Maximum Packet Buffer Depth (msec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VoQ for a 100Mbps output port</td>
<td>5,000 packets</td>
<td>1.25 MB</td>
<td>12.5 msec</td>
</tr>
<tr>
<td>VoQ for a 1G output port</td>
<td>12,500 packets</td>
<td>12.5 MB</td>
<td>12.5 msec</td>
</tr>
<tr>
<td>VoQ for a 10G output port</td>
<td>50,000 packets</td>
<td>50 MB</td>
<td>5 msec</td>
</tr>
<tr>
<td>VoQ for a 25G output port</td>
<td>125,000 packets</td>
<td>125 MB</td>
<td>5 msec</td>
</tr>
<tr>
<td>VoQ for a 40G output port</td>
<td>200,000 packets</td>
<td>200 MB</td>
<td>5 msec</td>
</tr>
<tr>
<td>VoQ for a 50G output port</td>
<td>250,000 packets</td>
<td>250 MB</td>
<td>5 msec</td>
</tr>
<tr>
<td>VoQ for a 100G output port</td>
<td>500,000 packets</td>
<td>500 MB</td>
<td>5 msec</td>
</tr>
</tbody>
</table>

Individual queues are configurable with queue depths between 0 and 1.57M packets and 0 bytes to 2.1 gigabytes.
The VoQ subsystem enables buffers that are dynamic, intelligent and deep such that there is always 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 Neworks (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 anything in between.

**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 here once the output port has signaled via a VoQ grant message that it is the allocated slot for a given input.

The original packet is sliced into variable-sized cells of at most 256 bytes. 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 can be switched locally and bypass using any fabric bandwidth resources, but otherwise aren’t processed any differently in terms of the VoQ subsystem.

**Stage 5: 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 into multiple output packets/frames if there are multiple receivers on this output 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 6: 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 26: Packet Processor stage 6 (egress): Egress Receive Packet Processor](image)

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 service requests in a fair manner 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 6MB on-chip buffer used for egress queuing, which is allocated as 64K packet descriptors. This buffer is mostly 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 7: Egress Transmit Packet Processor**

![Figure 27: Packet Processor stage 7 (egress): Egress Transmit Packet Processor](image)
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. Decoupling the packet forwarding on ingress from the packet rewrite on egress enables larger numbers of next-hops and tunnels 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 8: 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 the IEEE 802.3 standards.

**Multi-Chip Architecture**

Many of the switches in the 7280R series use more than one packet processor to support higher port density. The architecture of these switches uses a pair of packet processors directly connected with each other, or when they contain more than two packet processors, they are connected through a fabric chip. The fabric chip provides connectivity between all data-plane forwarding packet processors inside the system.

Refer to the 7500R architecture white paper for the role of the fabric chip and the interaction with the multiple packet processors.

**Arista EOS: A Platform for Scale, Stability and Extensibility**

At the core of the Arista 7280R Series is Arista EOS® (Extensible Operating System). Built from the ground up using innovations in core technologies since our founding in 2004, EOS contains more than 8 million lines of code and over 1000 man-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 nonstop 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 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 where there is separation of state information from packet forwarding, protocol processing and application logic. In EOS, system state data is stored and maintained in a highly efficient, centralized 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, easier software maintenance and module independence, higher software quality overall, and faster time-to-market for new features that customers require.

Arista EOS contrasts with the legacy approach to building network operating systems developed in the 1980’s that relied on embedding system state held within each independent process, extensive use of inter-process communications (IPC) mechanisms to maintain state across the system, and manual integration of subsystems without 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.

---

Due to pending litigation, the operation of ACLs as described in this document is not implemented for systems distributed within the US after July 3rd 2017. For more information and alternative options, please liaise with an Arista representative.
Arista took to heart the lessons of the open source world and built EOS on top of an unmodified Linux kernel. We have also maintained full, secured access to the Linux shell and utilities. This allows EOS to preserve the security, feature development and tools of the Linux community on an ongoing basis, unlike legacy approaches. This has made it possible for EOS to natively support things like Docker Containers, simplifying 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 faster to deliver significant new features to customers.

EOS is fully extensible, with open APIs at every level: management plane, control-plane, data-plane, services-level extensibility, application-level extensibility and with access to all Linux operating system facilities including shell-level access. Arista EOS can be extended with unmodified 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.
Conclusion
The Arista 7280R series universal leaf switches offer a selection of network switches that are designed for large, virtualized cloud networks, while meeting the performance requirements of modern applications. The 7280R Series also provides the capability to add support for new protocols, offers easy network migration and achieves these goals cost effectively. Combining Arista’s open and extensible EOS with the flexibility and programmability offered by the architecture, the Arista 7280R Series is a great addition to Arista’s product portfolio of comprehensive data center networking solutions.