Multiplanar Scale Out fabrics with MRC

Arista's support of MRC for building an open multi-plane Scale Out Ethernet AI fabric

Introduction

As AI clusters scale toward 100,000 XPUs and beyond, any underlying inefficiency in the Scale Out fabric can become a major contributor to the outliers affecting model training time and the overall return on infrastructure investment. An oversubscribed link, an asymmetric path latency, or a link or optic failure translates directly into XPU idle time, an inflation of the overall Job Completion Time (JCT), and consequently increased infrastructure cost.

In an effort to mitigate these concerns as demand for XPU scale continues to increase, interest has grown in combining multi-plane topologies with packet spraying. However, current deployments have been constrained by proprietary, single-vendor implementations that require tight coupling between the NIC and the switches of the Scale Out fabric. To eliminate this final barrier, OpenAI has collaborated with a vendor consortium to establish and validate a new **Multipath Reliable Connection (MRC)** OCP specification[2]. The specification aims to deliver XPU scale and bandwidth efficiency through a standardised multi-planar Scale Out fabric architecture.

This whitepaper explores the architectural drivers for the multi-planar approach, the key Arista features developed in our AI Etherlink™ platforms to enable MRC, and its deployment in a real world environment at OpenAI[1] using an SRv6 forwarding plane.

Two-tier at Scale

A fundamental requirement when designing the Scale Out fabric is the reduction of switching tiers, with the optimal goal being a two-tier or less topology regardless of the specific XPU cluster scale. A two-tier fabric guarantees a uniform, deterministic inter-switch latency, ensuring performance consistency for all XPUs when workloads traverse the fabric regardless of their physical location. The critical factor to avoiding the penalty of a third tier, is the radix (port density) represented by K of the switching nodes deployed within the fabric. With the XPU scale achievable within a two-tier topology calculated from the formula: $N = K^2/2$. In the case of the Arista AI platform portfolio, which provides support for both fixed and modular platforms, this can be more precisely defined as proportional to the product of the radix used at each of the two tiers; N = Kspine x (Kleaf /2), as the hardware platform and therefore port density may be different at each tier.

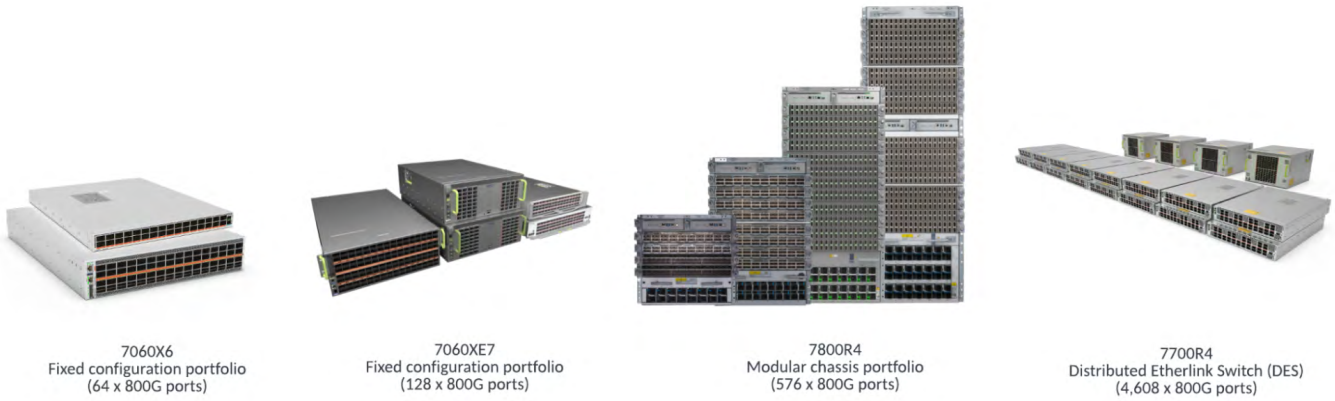

Figure 1: Arista's AI fixed and modular platform portfolio

This is made possible by Arista's consistent EOS AI feature set across the portfolio, allowing the fixed configuration 7060X6 and 7060XE7 platforms, the non-blocking modular 7800R4 chassis, and the distributed scheduled fabric DES-7700 solution to be deployed at either tier of the Scale Out fabric. This consequently presents two distinct design choices for achieving a two-tier architecture.

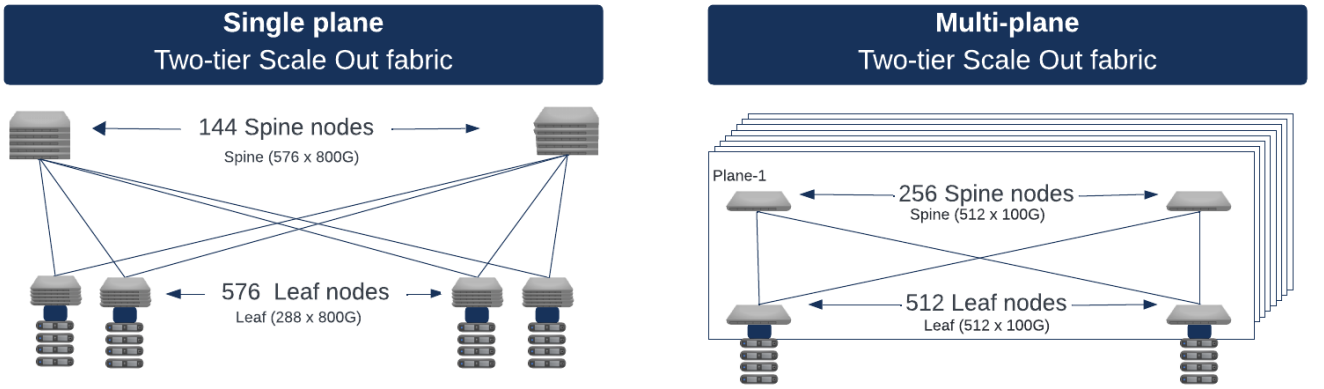

Figure 2: Two-tier single plane and multi-plane design models

Single plane with increased port density: To support a larger radix of switches and therefore XPU nodes, deploy higher density platforms at the Spine and if necessary the leaf layer. This is possible with Arista's AI Etherlink platforms, which includes the fixed 7060X6 (64 x 800G) and 7060XE7 (64 x 1600G or 128 x 800G) platforms and the higher density 7800R4 modular family, which scales to support 576 x 800G. The combination of a 7800R4 Spine and 7060X6 Leaf is a common design approach supporting up to 18,432 x 800G XPUs or 73,728 x 400G XPUs in a single-plane two-tier design. Alternatively, the modular 7800R4 can be deployed at the leaf and spine layer, to increase the XPU scale further (82,944 @ 800G XPUs) all within a single-plane two-tier design.

Multi-planes at a lower speed: Deploy multiple two-tier fabrics (planes) connecting the XPUs to each plane at a lower speed, but with the same overall aggregate bandwidth. This is made possible in Arista's Etherlink platforms which can flexibly break out each individual port, allowing a 800G port to operate as 2 x 400G ports, 4 x 200G or 8 x 100G. This allows a 7060X6 platform supporting 64 x 800G, to be deployed as a non-blocking leaf node connecting 128 XPU at 400G, 256 XPUs at 200G or 512 XPUs at 100G. In an eight-plane design with XPUs connecting at 100G to each plane, this provides support for 131K XPUs operating at 800G capacity, with each plane maintaining the optimal 2-tiers design using the same 7060X6 platform at the leaf and spine of each plane, The 7060X7 platform further extends this capability, enabling even greater overall XPU scale.

The adoption of a multi-plane approach does increase the number of Scale Out planes, but it optimizes the switch radix within each plane, enhancing the resiliency of the overall design, while maintaining an optimal two-tier architecture. However, to maximize the parallel planes, a plane-aware reliable transport is now required; packets need to be evenly distributed across the planes, paths need to be monitored, traffic needs to be rerouted in the event of congestion or a failure, all while maintaining deterministic load balancing across the planes. Enter OpenAI's Multipath Reliability Connection (MRC) specification and its integration with Arista's AI platform portfolio with SRv6 support.

Multipath Reliability Connection (MRC)

The MRC specification extends the RoCEv2 transport protocol preserving existing Verbs but restricting support to memory write operations (Write and Write-IMM) while incorporating functionality defined within the UEC (Ultra Ethernet Consortium) specification with regards to packet spraying, congestion notification, and packet trimming.

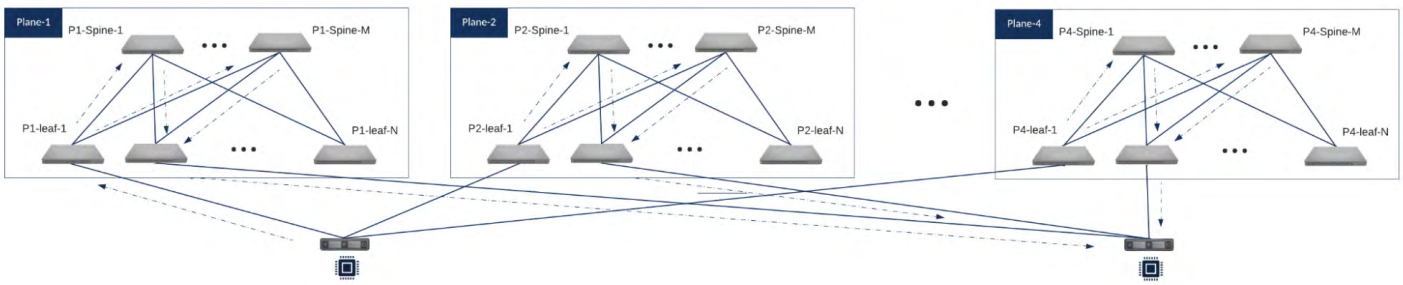

Figure 3: Diagram of the multi-planar transport of MRC

The aim of the specification is to provide a reliable transport connection between XPUs, but decouple it from any single physical path within the fabric, enabling the use of multiple paths across multiple planes of the topology. This means, instead of pinning a Queue Pair (QP) to a specific physical path, MRC allows a single connection to be reliably distributed across multiple active paths and planes, with any out-of-order packets being handled at the receiver.

Entropy Value (EV)

The spraying of packets within a QP flow across planes is governed by an assigned Entropy Value (EV). Each flow is mapped to an EV profile, containing a set of EV values, representing a set of unique paths through the planes to the destination, with the sending NIC rotating through the EV values for each subsequent packet within the QP flow.

Depending on whether an IPv4 or IPv6 underlay is deployed in the Scale Out fabric, the EV value is programmed as a 16-bit value in the UDP source port of the IPv4 packet, or programmed as a 32-bit value striped across the UDP source port and the IPv6 Flow Label of the IPv6 packet. Multiple forwarding planes are supported by the specification. In the case of an IP forwarding plane, the EV value in the UDP and IPv6 header introduce entropy allowing the packets to be efficiently hashed across the paths of a traditional Equal-Cost Multi-Path (ECMP) Scale Out fabric. In another, the EV values can be used by the MRC-enabled NIC to program an end-to-end SRv6 path for the packet, removing the need for ECMP in the fabric planes, relying instead on SRv6 forwarding within the fabric. It is this optimised operational model deployed at OpenAI that is discussed within the whitepaper.

Out-of-order packet delivery

The ability to reliably spray packets within a QP flow across multiple paths introduces the possibility for packets within the flow to be received out of order at the destination XPU. This is addressed by each packet carrying the RDMA virtual address and remote key, so the receiving NIC can write each packet directly to system memory, regardless of the order it arrives. The receiver then uses Selective ACKs (SACK) to signal precisely which packets have arrived, allowing the sender to identify gaps and trigger selective retransmission of only the missing packets rather than require a go-back-N model and the resultant retransmission of the full window.

Path selection and monitoring

The resilience of MRC is derived from its ability to independently monitor the health of every path of every plane within the fabric, and therefore every EV value within an EV profile. If a packet is lost on a specific path, detected through the reception of a Negative Acknowledgement (NACK) or Selective Acknowledgement (SACK) in combination with a timeout at the source, the EV and therefore path is transitioned to an "inactive" state. This removes the EV as a potential "active" path and is temporarily skipped before being put back into service. If packets are completely lost on a path, it will be marked as bad and taken fully out of service. To ensure paths are not permanently marked as bad, while minimizing human intervention during a failure event, a local agent on the MRC-enabled NIC sends background probes across any inactive path to test reachability. If a probe returns a successful acknowledgement, the EV value and path is transitioned to an "active" state and again a candidate path for packet spraying.

Packet trimming and congestion feedback

To distinguish between physical path failure, transient buffer congestion, and load-balancing inefficiency, MRC utilizes a combination of packet trimming as defined within the UEC specification and Explicit Congestion Notification (ECN).

ECN is used for load-balancing signalling rather than simply congestion notification. A scale-out fabric designed for non-blocking forwarding, under efficient load-balancing conditions, shouldn't exhibit congestion. An ECN marked packet therefore signifies a path in the fabric is more loaded than others and therefore a need to move flows off the path. To this end, the receiver echoes any ECN marking back to the sender with its associated EV value in a SACK message. The sender then temporarily avoids that specific EV path for forwarding, thereby relieving the load-balancing inefficiency. Critically in this model, ECN is disabled on the last-hop to the receiver because last-hop in-cast congestion is handled separately by packet trimming.

When a node in the fabric experiences extreme buffer congestion, in accordance with the UEC specification, it "trims" the payload and forwards only the header through a high priority queue to the destination. Dropping the entire packet would otherwise result in MRC detecting a path failure at the sender. The destination NIC receives the header, recognizes that the data was lost due to congestion rather than link failure, signalling to the sender via a NACK message to retransmit the packet. Because the header reached its destination, the sender can confirm the path is functional and avoid permanently removing the EV value from its active set.

Path failure

When a packet is not trimmed due to a congestion event but dropped and not received, this is signaled back to the sender in a SACK message and interpreted as a path failure. Consequently the sender interprets this as a path failure marking the EV value as bad. Background probes are subsequently sent to test the path and if a threshold of probes are successfully received across the path, the EV can transition back to an active state.

Signal	Cause	Sender reaction	EV state
ECN mark	Congestion on specific Path	Rebalance; Temporarily move flows to other EV values	Remains active, but skipped once
Packet Trim and NACK	In-cast congestion issue	Selective retransmit of trimmed packet	Remains active, but skipped once
Packet lost no trimming	Link or switch failure	Temp remove EV value from the EV set	Inactive (bad) - probe and reinstate on success
Link-state bitmap in SACK	NIC port down on remote end	Remap EV set to other NIC port/ plane at destination	Plane level failure

MRC deployment with SRv6

The integration of MRC within an SRv6-enabled forwarding plane, moves the path selection and SRv6 encapsulation and de-encapsulation to the XPU's MRC-enabled NIC. The Arista nodes within the Scale Out fabric now function exclusively as SRv6 transit nodes supporting micro-Node (uN) SIDs, providing deterministic low-latency forwarding based on the explicit path programmed at the source MRC-enabled NIC.

With the MRC transport enabled on the NICs handling rapid path re-selection and removal in response to congestion, drops, or failure events within the fabrics, the OpenAI implementation disabled dynamic routing for the SRv6 forwarding plane within the fabric. Preprogramming static IPv6 routes at each switch hop instead. This can complement the dynamic routing within the Scale Across fabric and any other workloads within the fabric which are not MRC aware, while maintaining a simplifying operational model for the SRv6 forwarding plane without adding any performance penalty to the convergence.

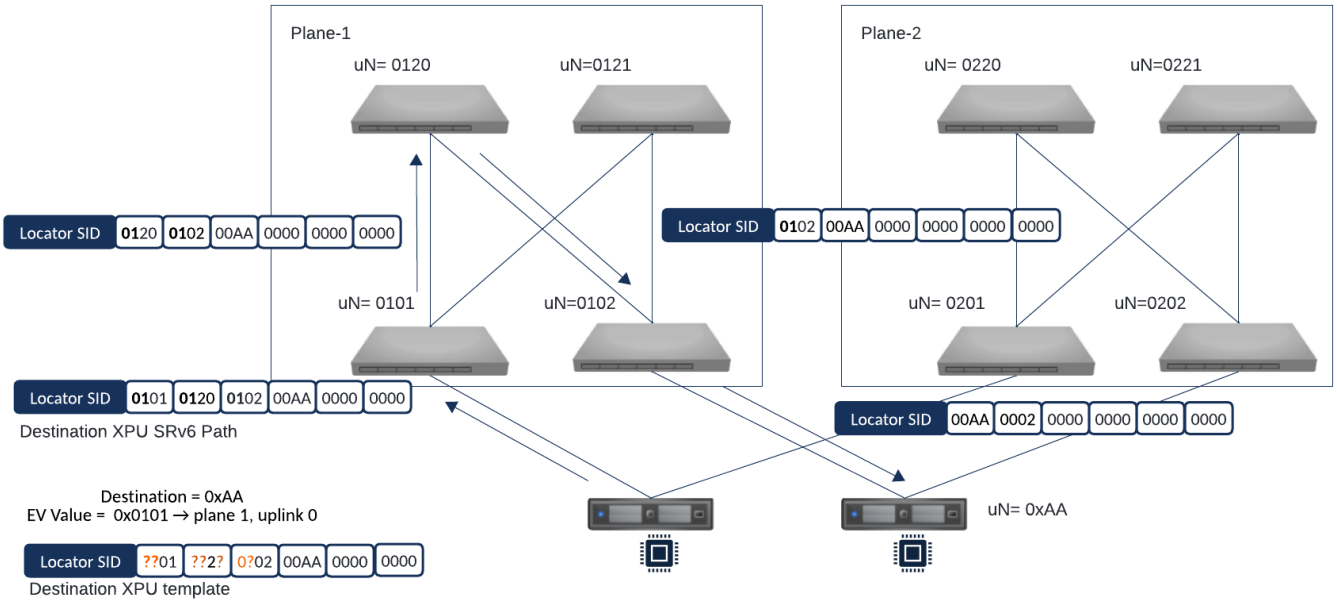

Figure 4: Diagram of the MRC multi-planar transport with SRV6

To further simplify the deployment model, the OpenAI implementation pre-provisioned the Arista fabric switches on startup via a centralized configuration element, configuring each node's IPv6 interface address and associated static routes to the adjacent upstream and downstream nodes within the fabric. The SRv6 addressing follows a strict schema, the nodes share the same locator SID with specific bits of a node's 16-bit uN (micro-Node) SID value, defining its plane and position within the overall topology. With this approach, a template SRv6 address to the destination XPU, in combination with the assigned 32-bit EV value, can be used to derive the unique end-to-end SRv6 path. The EV value is used to set the unique bits within each uN of the template address, defining the plane and uplink (Spine node) to be used for the path. The final uN SID in the path is derived from the IPv6 address of the destination XPU . This approach also has the added value of removing the need to hold both an SRv6 address and EV state per path,

as the SRv6 path can be derived directly from each unique EV value. It also means that the full original SRv6 path can be calculated from the EV value in any captured packet or at the destination.

While ECMP continues to play a critical role for load-balancing efficiency within the Cloud, AI (Frontend and Backend) and general data center designs, in the unique context of an MRC deployment the SRv6 forwarding plane can provide specific benefits over a standard ECMP approach. Each path across every plane of the fabric can now be verified and monitored with an explicit SRv6 probe. Secondly load-balancing across all active paths is now deterministic by explicitly programming the path at the source, with failover to a new active path provided immediately at the source without the requirement to wait for the network to reconverge and ECMP groups to shrink. However, even within an MRC deployment the two approaches shouldn't be seen as mutual exclusive, ECMP will still be deployed within the AI Frontend network, and as has been observed Scale Out fabrics required to support both MRC and non-MRC enabled workloads will require ECMP and SRv6 forwarding to co-existing within the same fabric.

Conclusion

As AI clusters scale towards 100,000 XPUs and beyond scale, the adoption of multi-plane topologies in combination with packet spraying has emerged as a compelling approach for achieving resiliency and scale within an optimal two-tier Scale Out fabric. However, such deployments have historically been constrained by proprietary, single-vendor implementations that necessitate tight integration between the NIC and the switching fabric.The collaborative development and open-sourcing of the MRC specification effectively eliminates this vendor lock-in, enabling a wider deployment of standard-based multi-plane topologies. This shift now provides network architects the freedom to select the optimal NIC and switching fabric combination for specific AI workloads.

Arista's AI Etherlink™ platforms optimize the Multipath Reliable Connection (MRC) protocol for scale-up, scale-out and across via network-accelerated packet trimming and intelligent buffering. Multiplanar traffic across independent fabric planes enhances deterministic performance and increased resiliency. At this massive scale our flagship AI leaf X series and AI Spine are offloading inter-cluster traffic with seamless routing and uncompromised performance.

Links

Multi-fabric blog

[1]MRC paper: https://cdn.openai.com/pdf/resilient-ai-supercomputer-networking-using-mrc-and-srv6.pdf

[2]MRC Open Specification V1.0: https://www.opencompute.org/documents/ocp-mrc-1-0-pdf

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

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

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

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

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

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

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

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