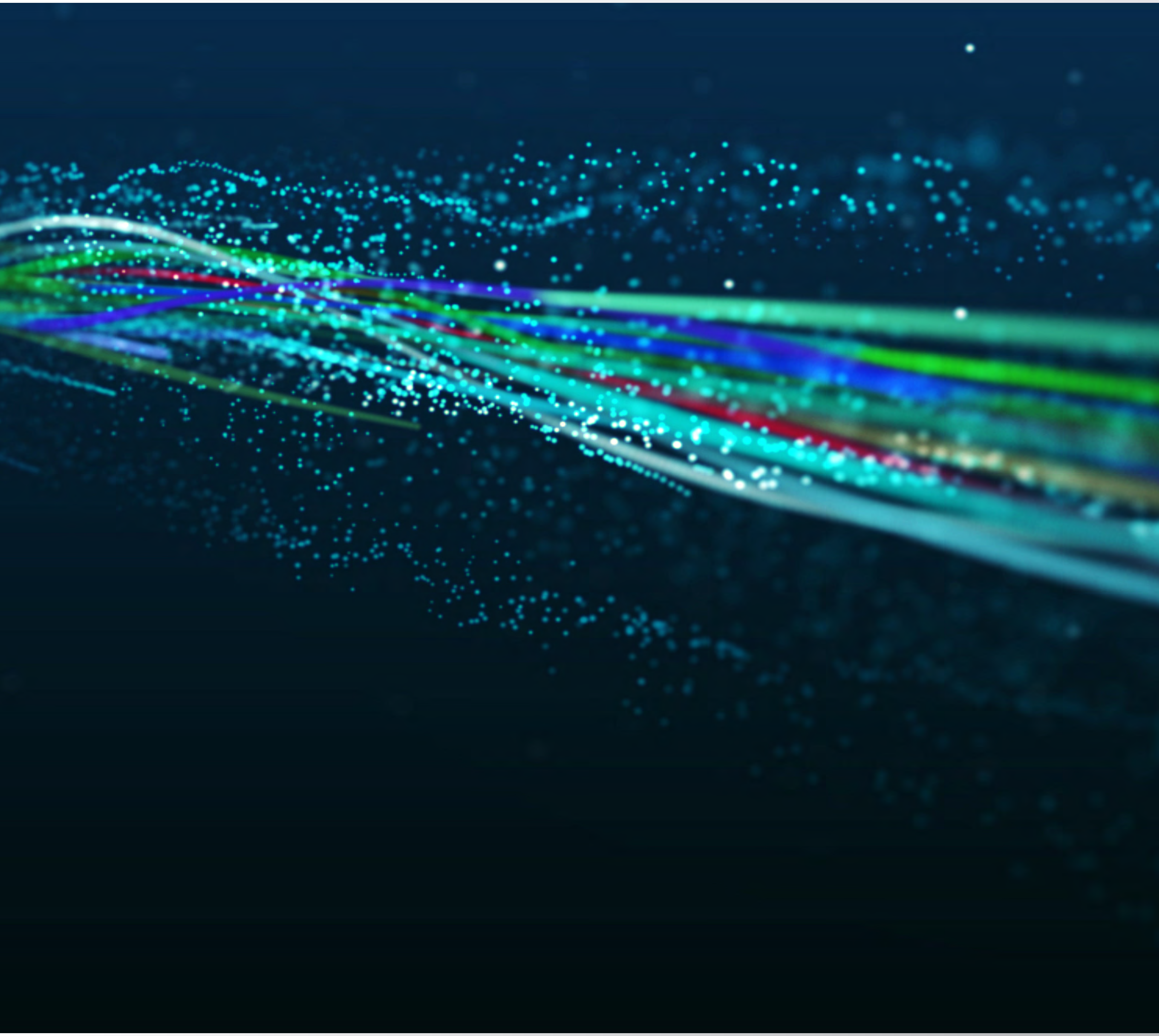
White Paper

Demystifying Ultra Ethernet

Tom Emmons and John Peach



In the early era of AI/ML, clusters were something of a science project, treated as specialist technology islands, developed independently from traditional services and networks, with the primary goal to serve as a proving ground for the future utility of this new form of compute.

Today AI/ML is becoming business critical and requires a common technology paradigm to fit into the established fiscal, operational and security parameters of an enterprise.

To make this a reality, we need a model that supports the needs of accelerated compute for multiple workload types; built, operated, monitored and secured using tried and trusted technologies, benefiting from the same skill sets, economics and openness that have given us the extensive networks that underpin the global economy.

Ethernet and IP have proven their versatility in every major application over the last 50 plus years, adapting to first replace and then improve upon legacy and proprietary technologies. History is repeating itself with Ethernet already the interconnect technology of choice for the majority of AI accelerators (XPUs). Advanced networking solutions - such as Arista's Etherlink ™ portfolio of AI focused platforms - already outperform legacy proprietary interconnect technologies.

Ultra Ethernet is the next natural evolution in the journey to ubiquitous networking.

The Ultra Ethernet Consortium, of which Arista is a founding member, is a standards organization that was formed with the goal of enhancing Ethernet for the needs of AI and HPC. Over 100 member companies and 1000 participants collaborated with a shared vision to make Ethernet even better. The recent publication of the 1.0 specification has fired the starting gun on hardware implementations that will take cluster performance to the next level, all while leveraging the same advanced Ethernet switching platforms.

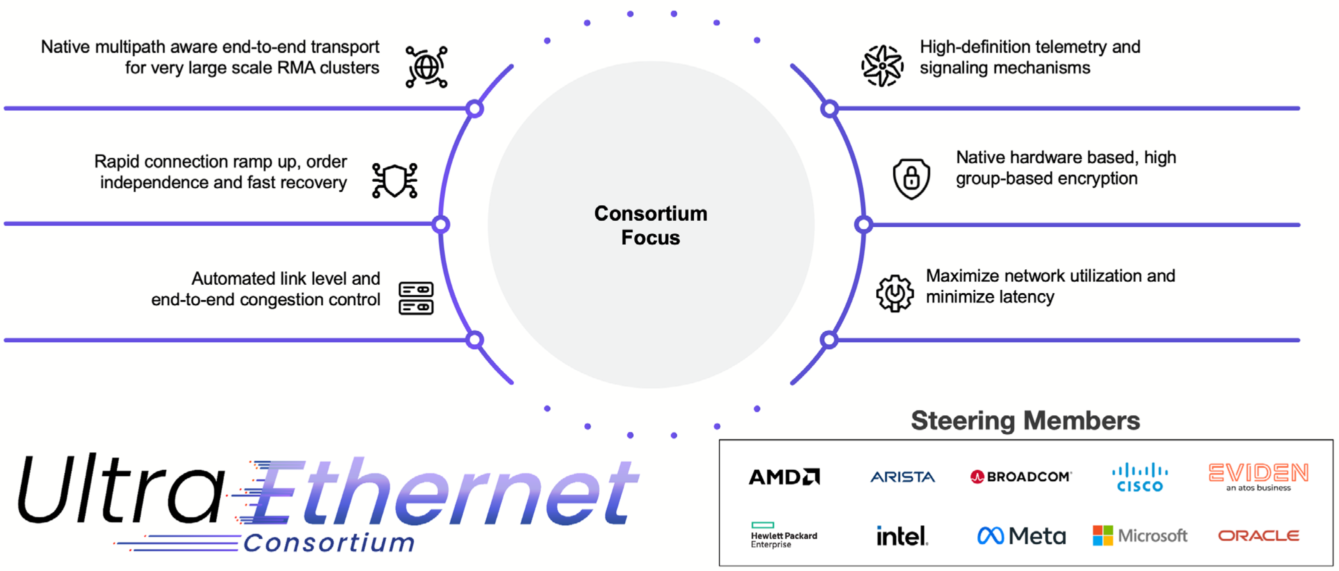

Native multipath aware end-to-end transport for very large scale RMA clusters
Rapid connection ramp up, order independence and fast recovery
Automated link level and end-to-end congestion control
Consortium Focus
High-definition telemetry and signaling mechanisms
Native hardware based, high group-based encryption
Maximize network utilization and minimize latency
Ultra Ethernet Consortium
Steering Members
AMD
ARISTA
BROADCOM
CISCO
EVIDEN an atos business
Hewlett Packard Enterprise
intel.
Meta
Microsoft
ORACLE

Figure 1: UEC Goals and Founding Members

Core to the UEC's vision is a reimagining of Remote Direct Memory Access (RDMA). RDMA is critical to the success of both AI and HPC applications. It allows systems and processors to exchange data at unprecedented speeds, with applications bursting data onto the network at rates of 400 Gbps today and 800 Gbps in the near future. This efficient communication makes distributing workloads across multiple servers and processors possible with minimal performance cost, enabling parallel computation of models spanning many thousands of accelerators.

The high flow rates and synchronized large volume flows common to RDMA traffic in AI/ML applications historically created challenges for Ethernet networks. Large flows create hashing nightmares as the network needs nearly perfect traffic distribution in order to avoid congestion. Additionally, RDMA flows that start up instantly and end just as quickly give traditional congestion control algorithms little time to react. While Arista's Etherlink enhancements already provide substantial improvement over vanilla Ethernet platforms, the next step in optimization requires rethinking of how applications interact with the network.

The UEC is focused on this next level optimization, making RDMA a native Ethernet application and adding new traffic distribution semantics and modern congestion control on top of standard Ethernet and IP layers.

Enter Ultra Ethernet Transport (UET) - RDMA reimagined to meet the demands of modern workloads, without the requirement of proprietary infrastructure.

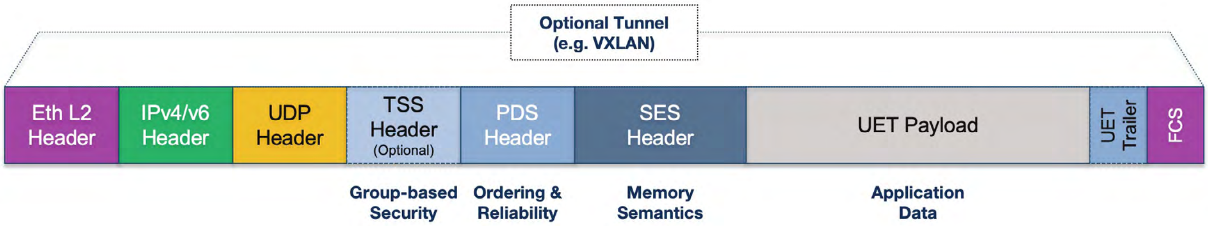


Figure 2: UET Packet Format

Native Libraries

Maximizing performance requires a closely coupled interaction between applications and transport to eliminate the compromises and workarounds of simply transporting an existing protocol without optimization. For this, UEC makes use of the mature and ubiquitous libfabric 2.0 API, standardized by the Open Fabrics Alliance and implemented broadly in HPC environments. UET's implementation of a native libfabric transport protocol ensures the most efficient interaction between application, API and transport.

Libfabric itself is extremely versatile and widely adopted; the API has done the hard work of taking the many different RDMA semantics and centralizing them under a single central interface. It makes porting applications across different systems and accelerator architectures much easier than traditional Verbs-based programs which must be compiled against specific libraries. Best of all, if your application sits on top of PyTorch and *CCL libraries, all of this can be hidden through the use of network plugins with minimal or no application changes making the transition to UET straightforward for a wide range of use cases.

Traffic Forwarding

As a native transport for all types of RDMA workloads, including traditional HPC as well as AI/ML, UET offers multiple forwarding paradigms suited to a range of high performance applications. Taking a high-level perspective, a fundamental concept of UET is the evolution from flow-based traffic distribution to packet spraying from the source NIC. Traditional transport layers require every packet to arrive in order, requiring every packet of a flow to follow the same path through the network. Allowing packets to take a different path through the network eliminates the need for flow hashing in the network and reduces the potential for imbalanced traffic distribution that leads to congestion. To handle this, the destination NIC must be able to receive packets out of order and reassemble the conversation to hand off to the application.

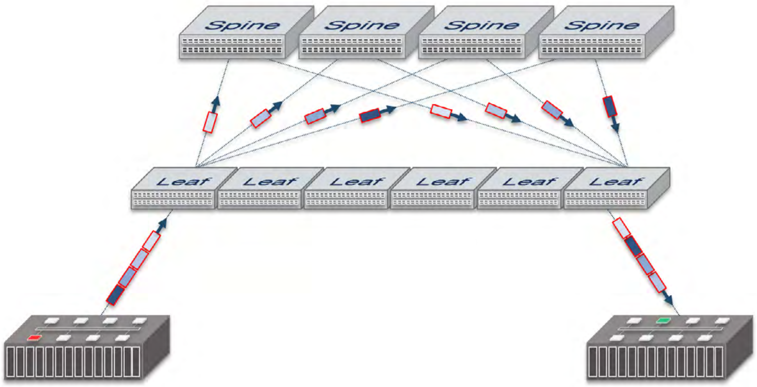


Figure 3: Packet Spraying Overview

Existing approaches to out-of-order packet arrival are proprietary implementations bolted on top of ROCEv2 standards. This prevents interoperability and fundamentally limits flexibility as they must map to existing ROCEv2 semantics and may not support all operations.

UET is built from the ground up for packet spraying, ensuring optimal efficiency at every layer. Every packet from every RDMA operation carries the necessary information to be written directly to memory. As the NIC only has to maintain packet metadata, this is a high throughput, low latency and low cost operation.

UET's native implementation ensures reordering support for all message types without having to store costly payload data in a reorder buffer. Efficiency is further improved by new drop detection mechanisms that allow for retransmission of individual lost packets instead of a full round trip of potentially hundreds of packets when using traditional transport technologies, considerably improving latency and good-put while minimizing network load.

Congestion Management

Traditionally, a packet sequence number is used to detect dropped packets. If packet 2 is received, followed by packet 4 then packet 3 is assumed to have been lost, triggering a go-back-N operation that requires many packets to be re-sent, just to receive a new copy of packet 3. This logic is slow and inefficient and no longer works with packet spraying since packet arrival times could be skewed by transient congestion or even the difference in propagation time across multiple network paths caused by a difference in fiber lengths.

Efficient detection and recovery from packet loss is therefore critical to aligning maximum performance with a packet-spraying architecture. UET natively solves this problem by providing selective acknowledgment and retransmission of individual packets. The receiver can specify exactly which packets have been received, which are missing, and which were dropped. Based on this information, the sender will resend only the specific packets that were dropped in the network. This reduces the time to completion and the network load of a dropped packet from a full round-trip time to only a single packet.

This still leaves the challenge of detecting which specific packets are dropped. UEC provides for heuristics using how much reordering is happening on the network to try to proactively detect dropped packets. However, for congestion, the most common form of packet loss, UET optionally makes use of a powerful new technology called packet trimming which is supported by all Arista Etherlink platforms.

At a high level, trimming functions as follows; when a packet arrives at a congested switch, rather than dropping it, the packet can instead be truncated to a minimal size - say 64B (a 64x reduction on the typical 4kB packet size). The trimmed packet is then placed in a higher priority queue and forwarded to the original destination. The destination will then reflect the trimmed packet back to the sender in the form of a NAK. This accomplishes two very important objectives.

- First, the trimmed packet serves as an explicit notification of a dropped packet. Trimmed packets are small enough that the cost to buffer and forward them in the network is minimal. They more than make up for that cost by allowing individual dropped packets to be efficiently detected and retransmitted, ensuring an efficient recovery from oversubscribing the network.
- Second, the trimmed packet serves as a powerful congestion notification, quickly informing both the sender and the receiver that the network is congested and the rate of transmission needs to slow down. As the trimmed packet skips the queue in a high priority class, the trimmed notification can actually make it to the receiver before any data packets even make it out of the congested queue.

The trimming mechanism allows the sender to react to network congestion more quickly and with a degree of intelligence. It will both reduce its sending rate and attempt to route around any specific congested paths.

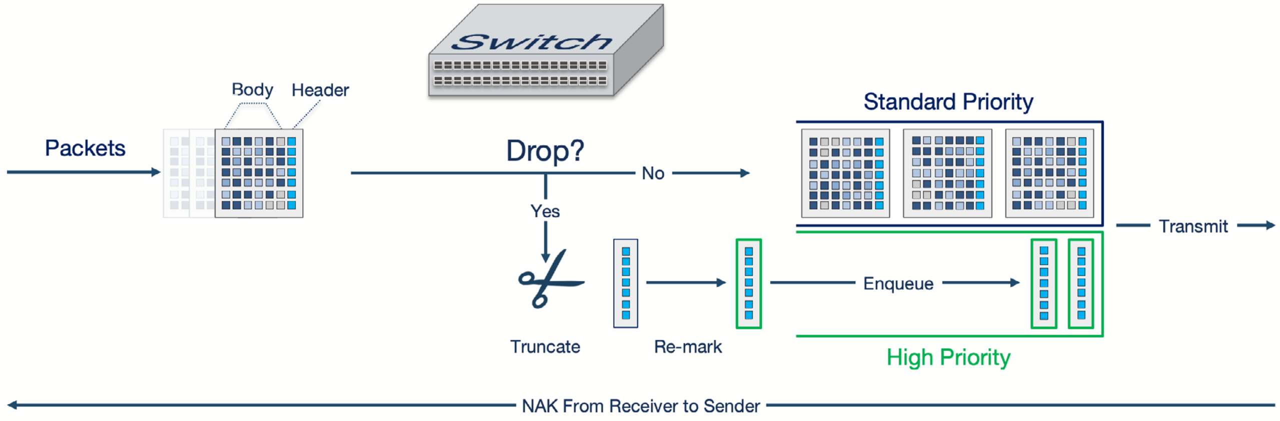


Figure 4: Packet Trimming Mechanism

Advanced Connection Setup and Host-based Flow Control

Of course managing congestion when it occurs is only part of the story, better yet is to avoid congestion in the first place, without introducing incremental overheads that inhibit application performance. To accomplish this, UEC introduces "Ephemeral Connections" and two new congestion control schemes.

Ephemeral Connections enable fast connection startup by eliminating the delay of a round-trip handshake before data begins to flow. Connections are established on demand by the first data packet and do not require explicit termination. This efficient mechanism both reduces the latency experienced by the application as well as reduces the need to maintain costly connection state on the NIC.

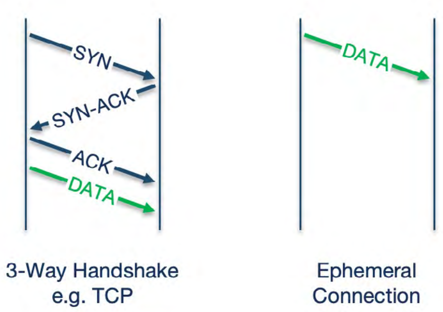


Figure 5: Comparing 3-Way Handshake to Ephemeral Connections

Network Signal Congestion Control (NSCC) is a sender-based method that makes use of several metrics to pace transmission rates upon detection of congestion including:

- Network delay as a fine-grain signal indicating the congestion level of the network.
- Trimmed packets that indicate congestion on certain paths
- ECN notifications which provide a leading indication of building congestion in a switch.

By combining these inputs, NSCC can quickly react to any congestion in the network by throttling transmission rates at the source

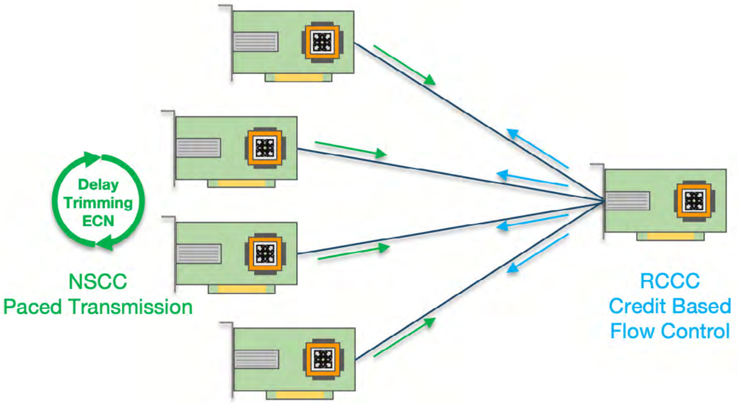


Figure 6: NSCC and RCCC Operation

UEC also includes an optional receiver-based mechanism. Receiver Credit Congestion Control (RCCC) operates to efficiently manage in-cast (where multiple packets arrive in parallel on several network interfaces but must be serialized to a single receiver). With RCCC, each receiver generates credits and allocates them fairly across all senders. This eliminates in-cast queue buildup in the last-hop switch while maximizing receiver throughput. NSCC and RCCC can each be used independently or together to optimize performance.

Security

The growing value of AI models, particularly when combined with business intellectual property or personalized data, makes security of data in-flight essential, especially for multi-tenant environments. UET makes security a first-class objective, rather than an afterthought.

Optional end-to-end encryption and authentication leveraging proven technologies like AES-GCM, Post-Quantum Cryptography (PQC) Key Derivation Functions (KDF) and replay prevention operates between UET hosts. A central pillar of UET's encryption specification is a novel group keying scheme optimized for the group computations common to AI and HPC.

A single group key is shared across all members of a singular job, for example all XPUs operated by a single tenant. Each individual NIC then uses the group key to derive a unique key for each connection.

With encryption enabled, everything inside the IP header is encrypted, protecting important model data from prying eyes and ensuring that no other tenants on the network have access to exposed application memory, preventing data injection as well as exfiltration.

Additional Future Capabilities

Within the network the UEC has standardized two other optional hardware based features to improve performance on a hop by hop basis.

Link Level Retry (LLR) is a retransmit mechanism on an individual link basis by implementing a small buffer on each switch port. If packets are dropped due to uncorrectable FEC errors on the link, LLR will respond to the drops by retransmitting packets without involvement of the hosts. This recovery prevents the need for more costly end-to-end transmission of these packets by avoiding a full round trip time of overhead. When application performance is dependent on the time to completion of the last packet in a collective, LLR has the potential to substantially increase performance reliability.

The second feature is Credit-Based Flow Control (CBFC), which provides a modern alternative to Priority Flow Control (PFC) on links where completely avoiding drops is desired. Unlike PFC which requires per-link tuning and only offers coarse granularity, CBFC allows the receiving switch to request exactly as many packets as it has space to reserve. As such, CBFC avoids the complexity of link specific tuning that requires detailed knowledge of packet sizes, link length, and response times. Additionally, CBFC provides the opportunity to schedule and load-balance based on CBFC state information, enabling a fixed set of buffers to be used more efficiently and allowing for a large number of virtual traffic classes abstracted from the limitation of 8 imposed by the 802.1p header used by PFC.

These new features require new logic design within switching silicon and will become available in future next-generation systems.

Summary

Taken together, UEC brings the relationship between AI and HPC applications and networks up to date. Tight integration between application semantics and network behaviors creates a native transport mechanism that retains the best parts of RDMA and combines them with best in class Ethernet solutions to create a potent package to build the next generation of applications on top of an Ethernet Transport.

Figure 7: Arista's Etherlink Portfolio

Here at Arista, as a founding member of the UEC, we're committed to making this vision a reality, by laying the ground work today for best-in-class, open standards based infrastructure with a diverse set of platforms that provide both freedom of choice and the flexibility to re-architect and redeploy as requirements evolve to maximize long-term investment protection. Today's Etherlink portfolio is UET ready today and we're already working hard on future generations of systems and partnering with other pioneers to build the best Ethernet networks for high performance computing.

Reference

- The Ultra Ethernet Consortium Launches Specification 1.0
- Ultra Ethernet Consortium
- Ultra Ethernet Specification
- Ultra Ethernet Whitepaper
- Arista 800G Portfolio
- Arista Blog Microsite

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