



Falcon: A Reliable and Low Latency Hardware Transport



Nandita Dukkupati, Prashant Chandra, Amber Huffman
@Open Compute Project (OCP) Steering Committee Meeting
7 March, 2024

Contribution Overview

Falcon Overview

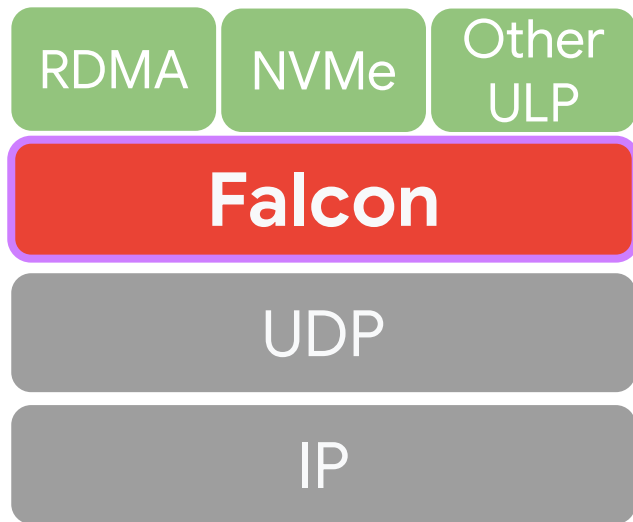
Falcon is a low latency and efficient hardware transport for RDMA and NVMe workloads and serves as an important ecosystem piece for AI/ML workloads.

Falcon was presented at the 2023 OCP Global Summit [[slides](#)] and at the February 12 Networking Project call.

Specification Contributions

- Falcon Transport Protocol, v0.9
- RDMA over Falcon Transport, v0.9
- NVMe over Falcon, v0.9

v1.0: Will follow up with fixes for errors/clarity based on 0.9 feedback.



Tail Latency in Ethernet networks

- HW assisted **delay-based Congestion Control**
- **Selective ACKs** for fast loss recovery
- **Multipath** capable connections

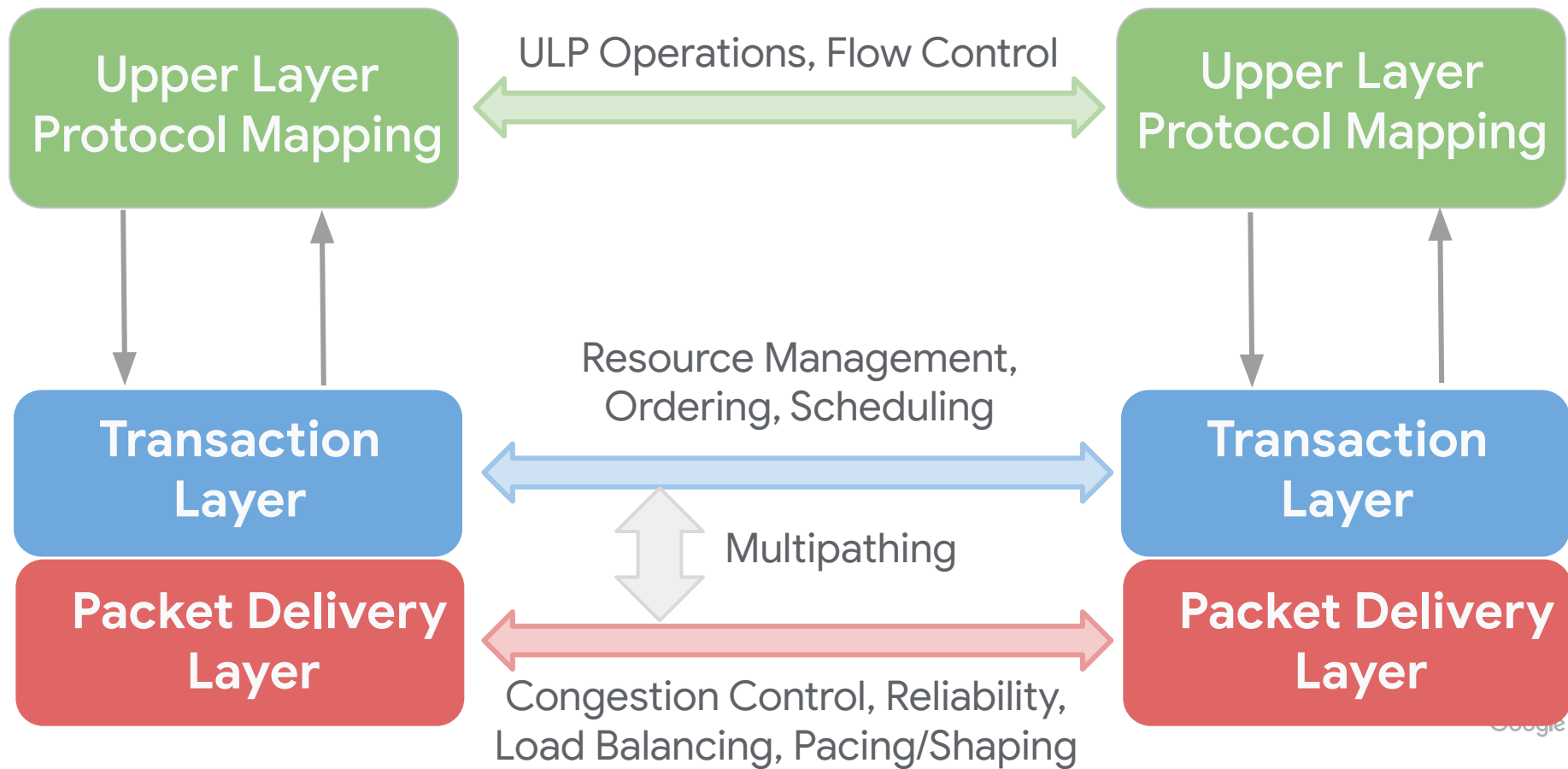
Isolation and Visibility at scale

- μ s-granularity per-flow **Traffic Shaping**
- Fine-grained Stats for **Debuggability, Software Defined Network control**

Efficiency and Security

- **Implemented in HW** for Low Latency, High Op Rate using Industry-standard Interfaces, and PSP encryption

Falcon Layers



Compliance with OCP Tenets

Openness

The specification complies with the tenet of Openness by empowering the Community with Google's production learnings to help modernize Ethernet. This includes leveraging production-proven technologies at scale including [Carousel](#), [Snap](#), [Swift](#), [PLB](#), and [CSIG](#) that have been openly published previously.

Efficiency

The specification complies with the tenet of Efficiency. Falcon achieves high performance by combining three key insights that achieve low latency in high-bandwidth, yet lossy, standard Ethernet data center networks. Fine-grained hardware-assisted round-trip time (RTT) measurements with flexible, per-flow hardware-enforced traffic shaping, and fast and accurate packet retransmissions, are combined with multipath-capable and PSP-encrypted Falcon connections. On top of this foundation, Falcon has been designed from the ground up as a multi-protocol transport capable of supporting Upper Layer Protocols (ULPs) with widely varying performance requirements and application semantics. The ULP mapping layer not only provides out-of-the-box compatibility with Infiniband Verbs RDMA and NVMe ULPs, but also includes additional innovations critical for warehouse-scale applications such as flexible ordering semantics and graceful error handling. Last but not least, the hardware and software are co-designed to work together to help achieve the desired attributes of high message rate, low latency, and high bandwidth, while maintaining flexibility for programmability and continued innovation.

Compliance with OCP Tenets

Impact

The specification complies with the tenet of Impact by introducing a new technology that helps the industry modernize Ethernet. Falcon provides a helpful solution to address demanding workloads that have high burst bandwidth, high Operations per second, and low latency in massive scale AI/ML training, High Performance Computing, and real-time analytics.

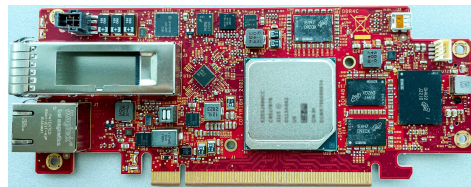
Scale

The specification complies with the tenet of Scale by being designed from the ground up to deliver high bandwidth and low latency in high-bandwidth, yet lossy, Ethernet data center networks. Additionally, it is composed of production-proven technologies delivered at scale including Carousel, Snap, Swift, Protective Load Balancing (PLB), and Congestion Signaling (CSIG).

Sustainability

The specification complies with the tenet of Sustainability by delivering an efficient Ethernet transport technology that minimizes retransmissions and other wasted effort and energy within an Ethernet network. Additionally, the technology allows a wide range of high performance workloads to be run on standard Ethernet networks.

Summary



Predictable performance for warehouse-scale: low tail latency, massive application bandwidth, mitigating congestion and efficient network utilization.

Efficiency @scale: hardware acceleration enables high message rate, high bandwidth, low latency, and connection scaling.

Need of the day: meets requirements of critical workloads, HPC and AI; also good for offloading Storage and RPC.

Questions?

Backup

Why we must rethink Transport in NICs

New and demanding workloads - high burst bandwidth, high Op rate, low latency.

We need solutions for new workloads (**massive scale AI/ML training, High Performance Computing, Real-time Analytics**) and satisfy demands of the existing ones (Storage, RPCs).

Incremental gains of software stack optimizations.

Need **order of magnitude improvements** over highly optimized Software Transports*.

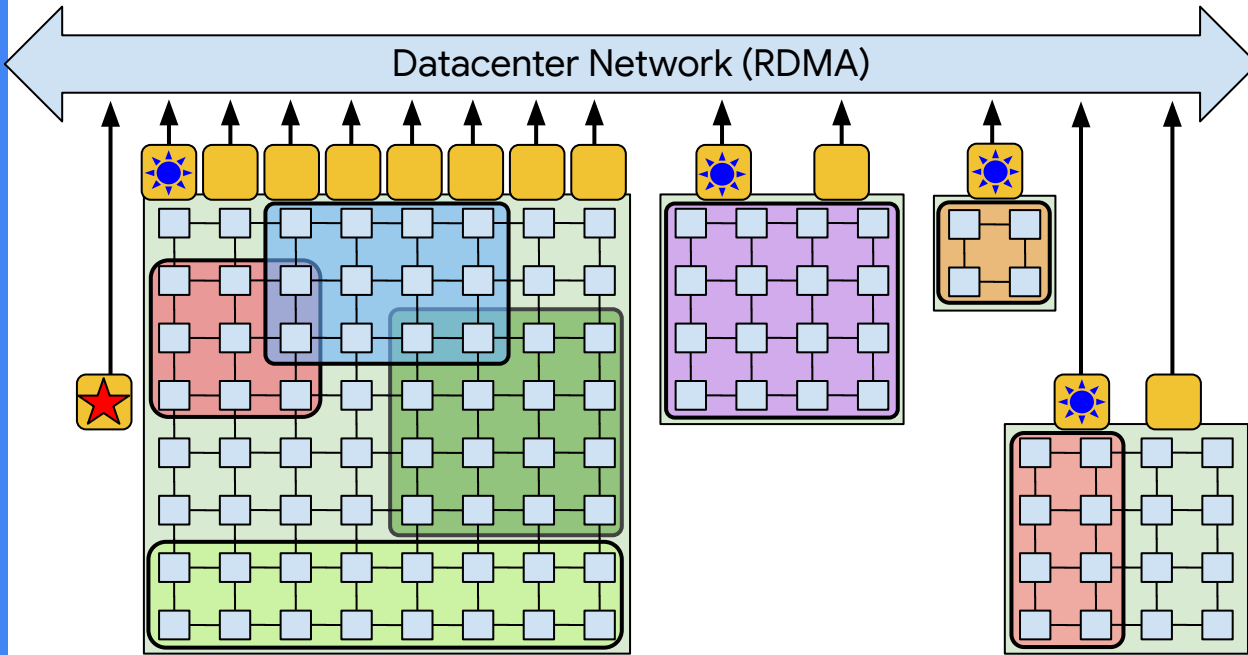
Modernising Ethernet for low latency and high bandwidth.

Deployment Experiences at scale give us a glimpse of the possibilities.

Hybrid networks are a reality.

Specialized networks **connected to Ethernet for scale**, e.g., Inter-chip Interconnect (ICI) and NVLink provide islands of ultra high performance.

***Snap**: a Microkernel Approach to Host Networking, SOSP 2019
80 Gbps/core unidirectional; 4.7M IOPS/core; 10us small message latency



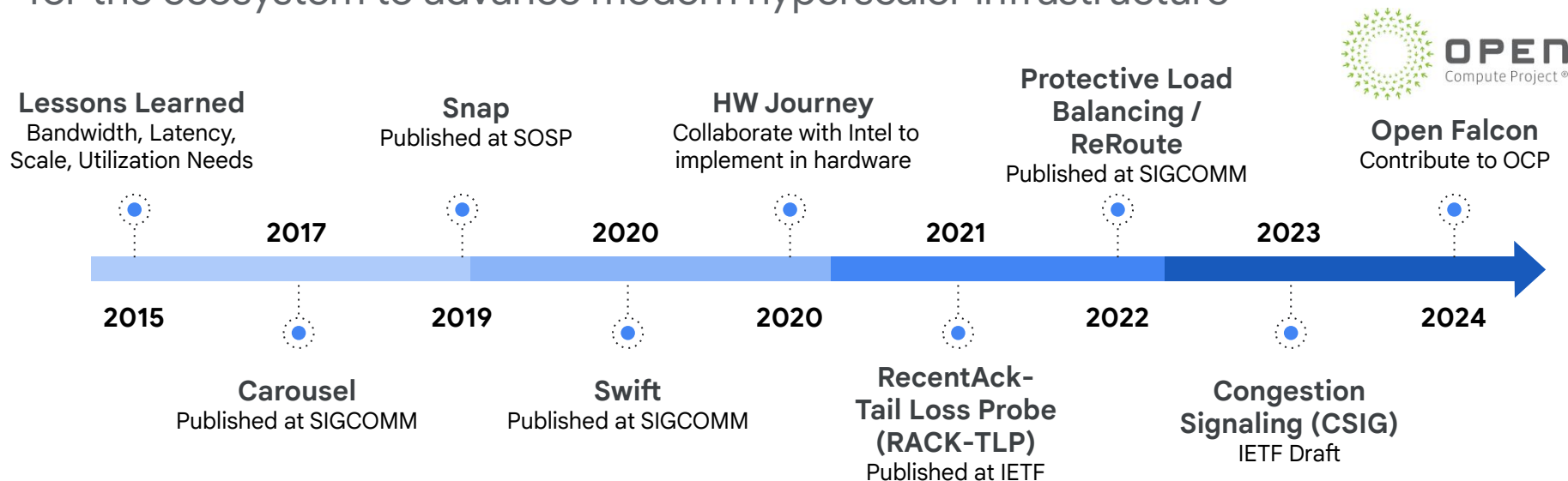
Accelerator Slices (with tightly coupled interconnects), Compute Servers, Storage

WHAT WOULD IT TAKE...

... to achieve
Performance, **Isolation**
and **Efficiency** at scale
for High Bandwidth, Low
Latency Workloads on
**today's Datacenter
Ethernet Networks?**

-  Host (many per island)
-  Resource Manager (global)
-  Scheduler (per island)

Falcon technology: Google opens reliable, low latency hardware transport for the ecosystem to advance modern hyperscaler infrastructure



Bringing 10 years of advances in low latency, isolation and efficiency to hardware

Unfolding Falcon

Mapping Upper Layer Protocols (ULP) to Falcon

Industry Standard Interface

RDMA InfiniBand Verbs Compatible ULP:

- Transparent to legacy Verbs applications.
- Supports Reliable Connected (RC) and Unreliable Datagram (UD) Queue Pair types.
- **Strictly Ordered:** in-order data placement, in-order completions.
- **On-NIC reorder buffer** to support OOO delivery from the network.

Enhanced Interface

Enhanced Verbs Mode:

- Introduces **Relaxed Ordering Modes:** weakly and unordered.
- **Graceful Error Handling with Complete-in-Error and continue (CIE):** signals errors to applications without tearing down the connection.

Industry standard interfaces are extended in support of warehouse-scale applications.

Falcon Transaction Layer

Exposes request/response interface to ULPs.

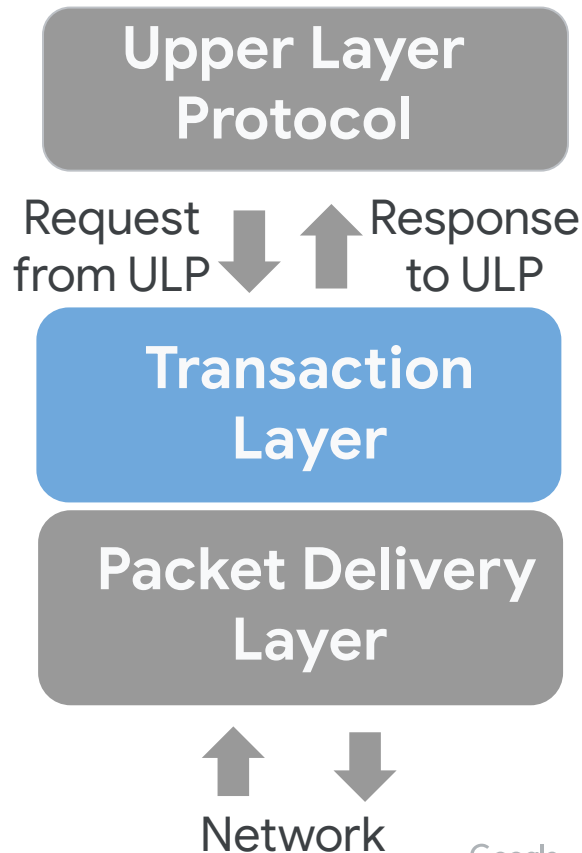
IB Verbs Reads, Atomics, Writes, Sends are mapped to request-response transactions.

Orders Transactions due to out-of-order network arrivals.

- Ensures ordering semantics expected by the ULP.

Schedules transactions on the wire per QoS-policies.

Manages finite Falcon resources for isolation and deadlock prevention.



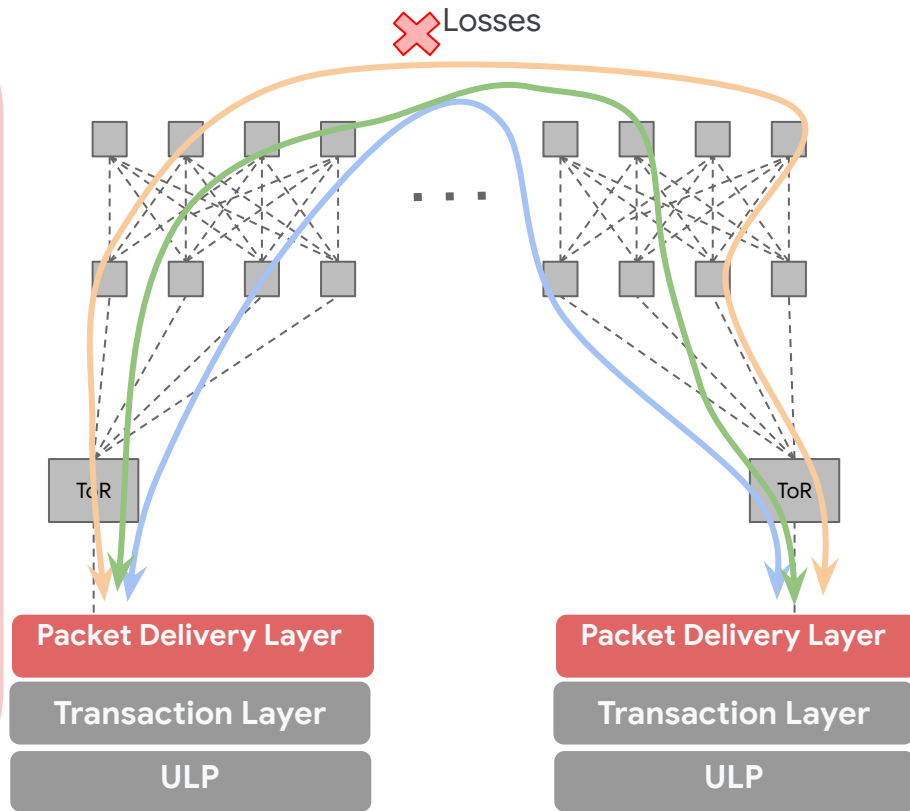
Falcon Packet Delivery Layer

Delay-based Congestion Control for low latency and high utilization.

Leverages multiple paths in the network fabric transparently to applications.

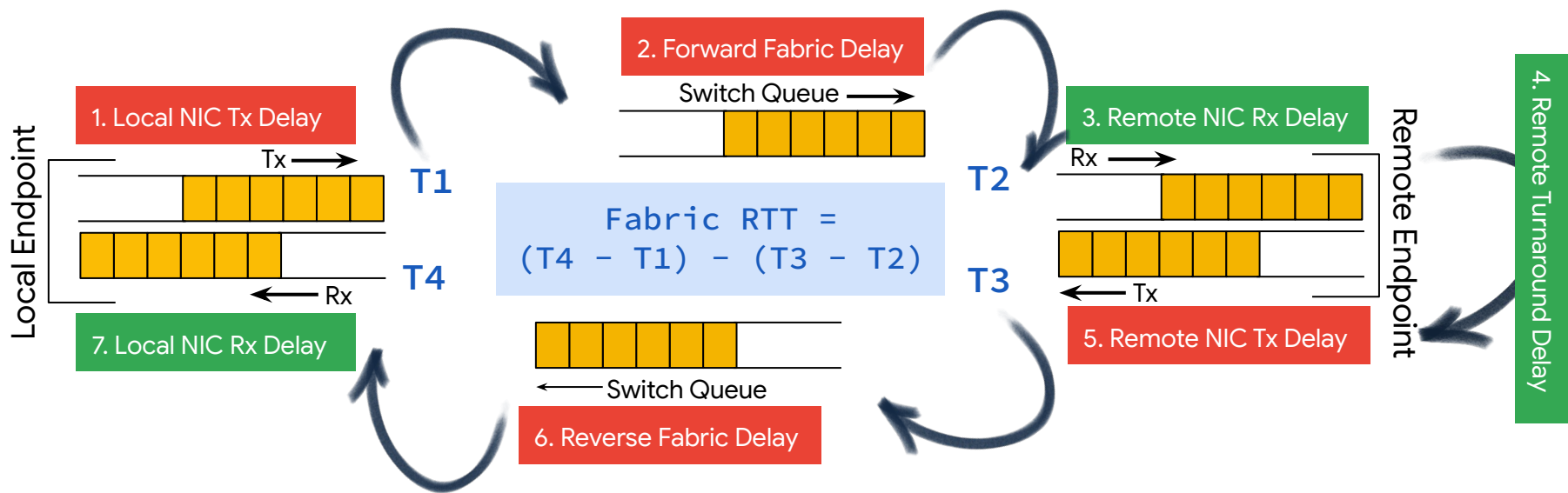
End-to-end reliable delivery

- Timely retransmission of lost packets.
- Hardware based retransmission.
- Ack coalescing/piggybacking for high Op rate.



Congestion Control in Falcon

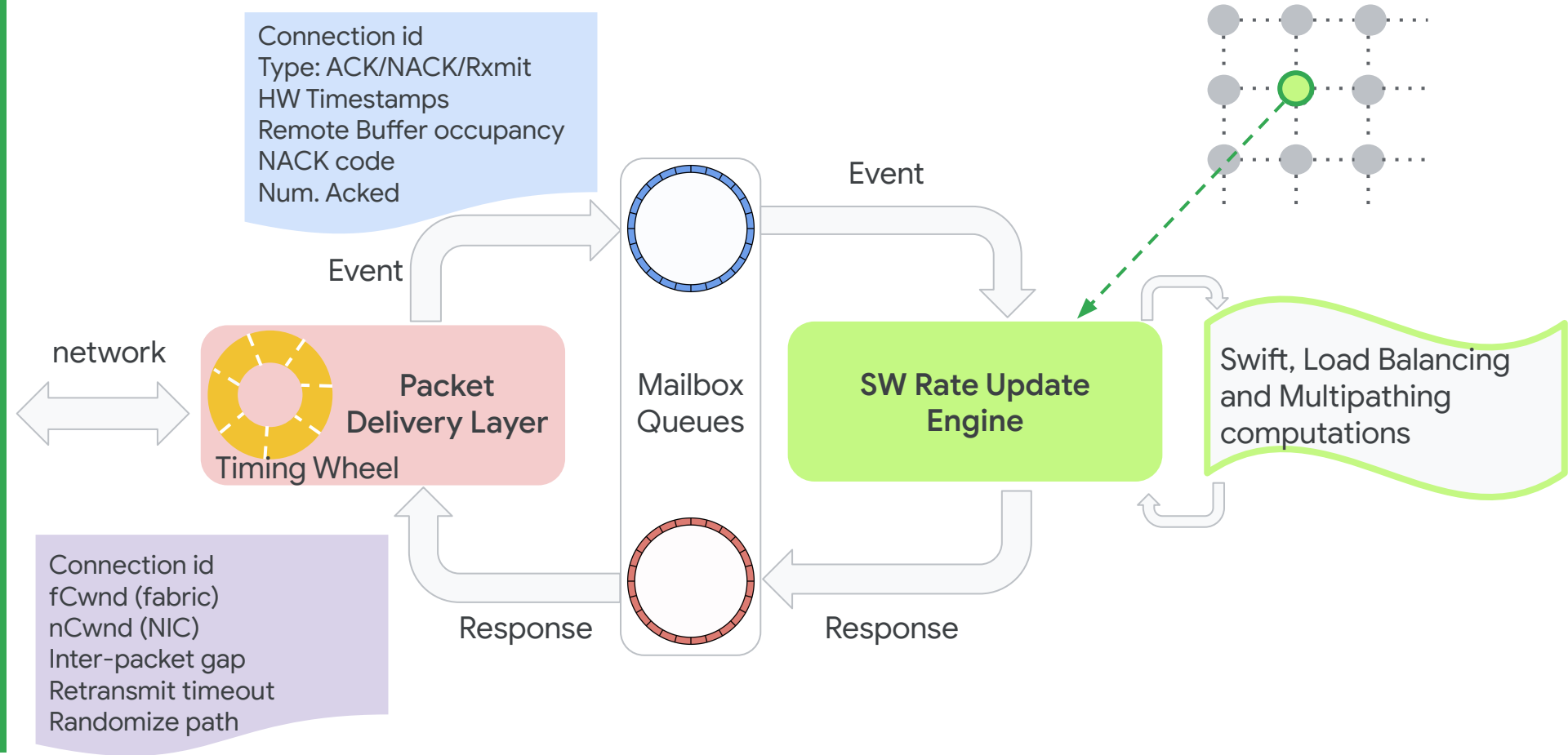
Swift Congestion Control as Baseline



Swift* is a **delay based congestion-control** for Datacenters that achieves **low-latency, high-utilization, near-zero loss** implemented completely at end-hosts and NICs **supporting diverse workloads** like large-scale incast across latency-sensitive, bursty and IOPS-intensive applications working seamlessly in heterogeneous datacenters.

*[Swift](#): Delay is simple and effective for congestion control in the Datacenter, SIGCOMM 2020.

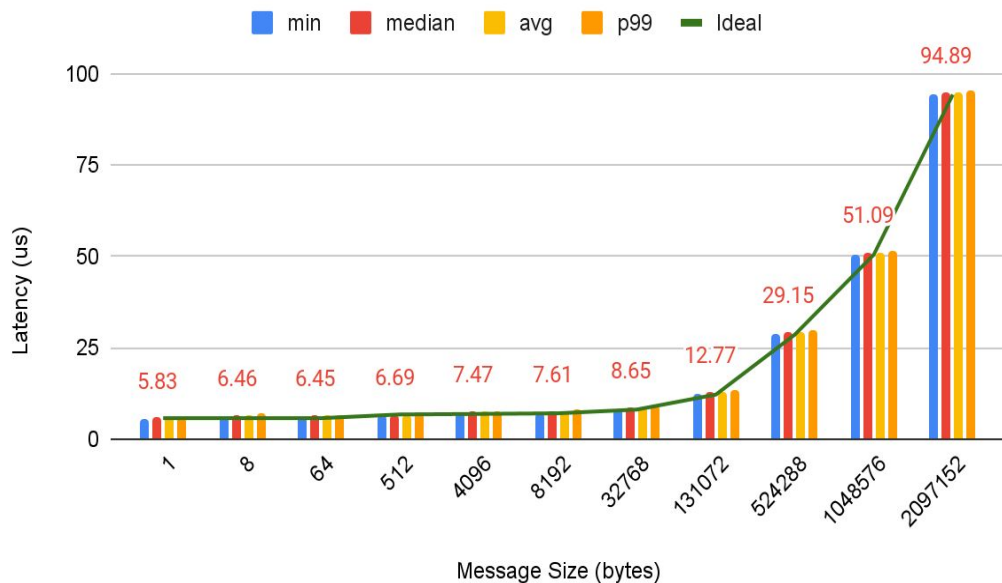
Programmability in Rate Update Engine (RUE)



RDMA/Falcon Performance

Single Queue Pair Latency

WRITE latency over message size



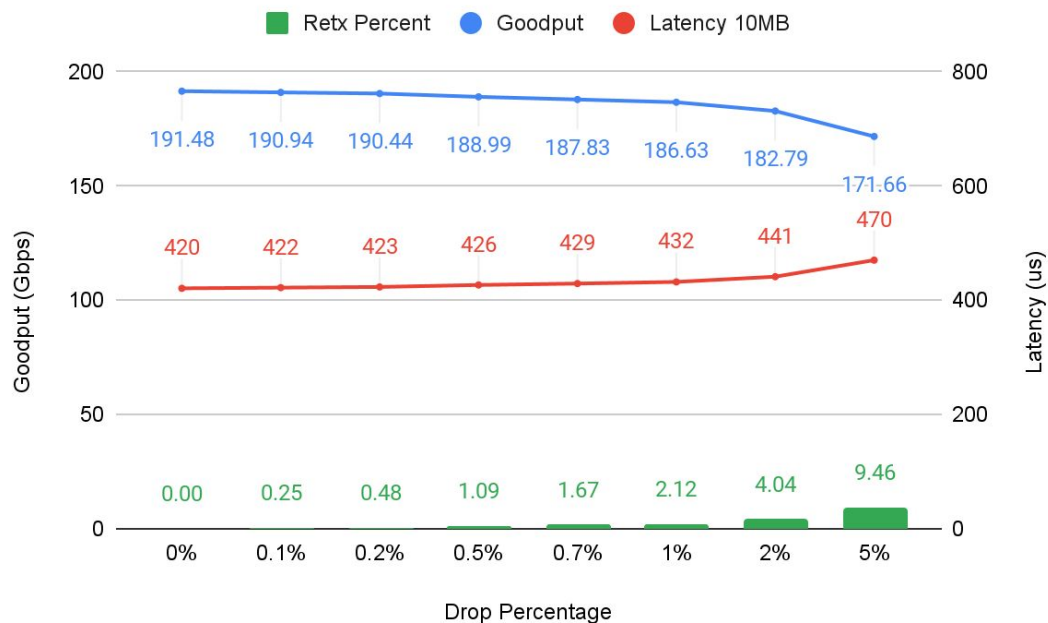
Setup:

Measure message completion time (round-trip time from application posting Op to receiving completion) over message sizes.

Takeaway:

Tight tail latency: **p99 latency**, **median** and **ideal latency** match across message sizes

Scalability under Packet Losses



Setup: induced packet losses on one Queue Pair from 0 to 5%.

Takeaways:

Stable goodput and message latency as loss rate increases; graceful degradation at higher losses.

Low retransmission overhead even under high packet loss rate

Call to Action

Industry Perspectives

"We welcome Google's contribution of Falcon as it shares the Ultra Ethernet Consortium's vision to drive Ethernet as the best data center fabric for AI and HPC, and look forward to continuing industry innovations in this important space."

Dr. J Metz, Chair, Ultra Ethernet Consortium



(led by AMD, Arista, Broadcom, Cisco, Eviden, Hewlett Packard Enterprise, Intel, Meta, Microsoft, and Oracle)

"Falcon is first available in the Intel IPU E2000 series of products. The value of these IPUs is further enhanced as the first instance of an Ethernet transport to add low tail latency and congestion handling at scale. Intel is a Steering Member of Ultra Ethernet Consortium, which is working to evolve Ethernet for high performance AI and HPC workloads. We plan to deploy the resulting standards-based enhancements in future IPU and Ethernet products."



Sachin Katti, SVP & GM, Network and Edge Group, Intel

"We are pleased to see a high-performance transport protocol for critical workloads such as AI and HPC that works over standard Ethernet/IP networks and enables massive application bandwidth at scale."

The Arista logo, which is a dark blue square with the word "ARISTA" in white, uppercase, sans-serif font.

ARISTA

Hugh Holbrook, Group VP, SW Eng. Arista

Industry Perspectives

"Cisco is pleased to see the contribution of Falcon to the OCP. Cisco has long supported open standards and believes in broad eco-systems. The rate and scale of modern datacenter networks and particularly AI/ML networks is unprecedented, presenting a challenge and opportunity to the industry. Falcon addresses many of the challenges of these networks, enabling efficient network utilization."



Ofer Iny, Cisco Fellow, Cisco

"Juniper is a strong supporter of open ecosystems, and therefore we are pleased to see Falcon being opened to the OCP community. Falcon allows Ethernet to serve as the Datacenter network-of-choice for demanding workloads, providing high-bandwidth, low tail-latency and congestion mitigation. Falcon provides the industry with a proven solution today for demanding AI & ML workloads."



Raj Yavatkar, Chief Technology Officer, Juniper

"Marvell strongly supports and is committed to the open Ethernet ecosystem as it evolves to support emerging, demanding workloads such as AI. We applaud the contribution of Falcon to OCP and welcome Google sharing practical experiences with the industry."



Nick Kucharewski , SVP & GM, Network Switching Group at Marvell

Learn more at the blog:

<https://cloud.google.com/blog/topics/systems/introducing-falcon-a-reliable-low-latency-hardware-transport>