Ordered Chaos: Towards Adaptive Multipath Load Balancing with Precise Fast Loss Detection

Tommaso Bonato*†[1], Ahmad Ghalayini†, Nadeen Gebara†, Benoit Pit-Claudel‡[2],
Jiayong Li*, Mikhail Khalilov*, Luca Piccolboni†, Waleed Reda†,
Pei-Huan Tsai§†, Tao Ji†, Abdul Kabbani†, Torsten Hoefler*†
*ETH Zurich, Zurich, Switzerland
{tommaso.bonato,jiayong.li,mikhail.khalilov,torsten.hoefler}@inf.ethz.ch
†Microsoft Corporation, Redmond, United States of America
{aghalayini,nadeengebara,lpiccolboni,waleedreda,taoji,abdulkabbani}@microsoft.com
‡MIT, Cambridge, United States of America
bpitcla@alum.mit.edu
§Columbia University, New York, United States of America
pt2630@columbia.edu

***Abstract*—Load balancing in best-effort datacenter fabrics faces a tradeoff. Order-preserving schemes such as Equal-Cost Multipath (ECMP) retain loss semantics but use path diversity coarsely, while packet spraying balances load well but removes the ordering signals needed for fast, precise loss detection. We present Ordered Chaos, an end-host design that preserves enough structure for reliable loss detection without giving up fine-grained multipath routing. Reactive Subflow Spraying (RSS) divides each flow into ordered, dynamically routed subflows. Precise Fast Loss Detection (PFLD) combines this ordering with probes to distinguish reordering from congestion, corruption, tail, and retransmission losses within one RTT. Across simulations, AI and HPC workloads, and a DPU prototype, RSS+PFLD improves tail flow completion time by up to 3×, eliminates up to 99% of timeout-based recoveries, and accelerates collective and training communication by up to 35%.**

***Index Terms*—data centers, computer networks, packet loss, routing**

I. Introduction

Large-scale AI training and tightly-coupled HPC workloads occupy tens of thousands of accelerators for weeks to months, making the transport layer's ability to detect and recover from packet losses a direct factor in time-to-result and cluster efficiency. At scale, these environments experience packet loss due to transient congestion, tail drops, failures, and persistent packet corruption arising from aging optics, hardware faults, and link degradation. Production measurements report that inter-switch and inter-card loss events, which include silent packet drops and packets discarded after link corruption, account for about 18% of packet drops associated with network performance anomalies [1], and end-to-end loss ratios in the range of 10^{-5} to 10^{-4} have been consistently reported within datacenters [2]–[5]. These losses compound across multi-hop paths, stalling collectives, desynchronizing bulk-synchronous phases, and increasing training costs [6]. Modern datacenter best-effort (lossy) fabrics expose multiple parallel paths between nodes [7]–[9] to meet throughput and resiliency demands, yet turning this path diversity into consistent application performance in the presence of loss and reordering, without introducing new reliability problems, remains an open problem [10]–[21].

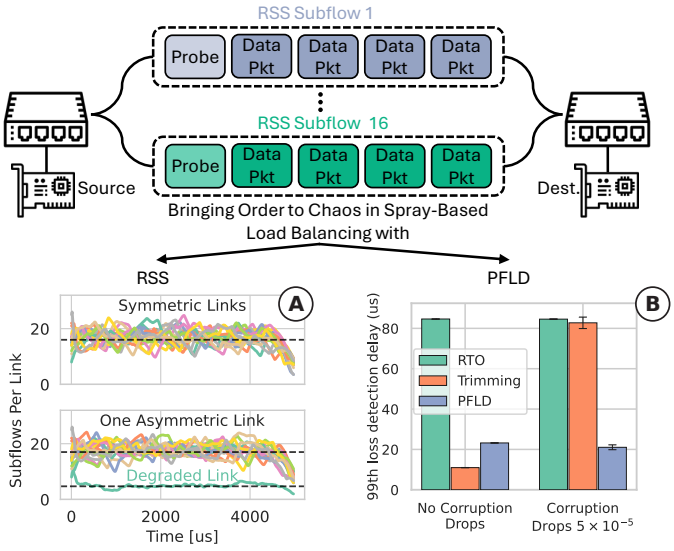


Figure 1: Scheme showcasing RSS+PFLD with 16 subflows. Ⓐ shows the performance of RSS during a tornado pattern with a symmetric and an asymmetric topology. Ⓑ shows the time it takes to detect a packet loss for different schemes.

In practice, most datacenter fabrics attempt to exploit this path diversity using per-flow load-balancing mechanisms, with Equal-Cost Multipath (ECMP) serving as the dominant starting point. ECMP hashes a flow's five-tuple to select a single path for its lifetime. While simple and order-preserving, ECMP's hash-based static mapping can lead to collisions that over-subscribe some links while under-utilizing others, and its lack of adaptivity prevents effective response to asymmetric conditions, congestion, or failures [11], [22], [23]. Reactive flow-level schemes such as PLB [10] reduce persistent collisions

[1]Tommaso Bonato has since left Microsoft.

[2]Benoit Pit-Claudel has since joined Amazon. This work does not relate to his position at Amazon.

SC26, November 15-20, 2026, Chicago, Illinois, USA
979-8-3195-4789-7/26/$31.00 ©2026 IEEE

through occasional rerouting, but still operate at coarse, per-flow granularity, limiting their ability to fully utilize available path diversity.

To overcome these limitations, the ecosystem has increasingly converged on multipath spraying, shifting from per-flow path selection to much finer-grained packet-level traffic distribution across paths. Correspondingly, industry efforts such as Ultra Ethernet [24], [25], Falcon [26], SRD [27], and the NVIDIA ConnectX-8 [28] and BlueField-3 [29] NICs all adopt out-of-order-capable transports with per-packet spraying or other fine-grained load balancing mechanisms.

However, packet spraying destroys the ordering signals that traditional transports rely on for fast loss recovery: an endpoint cannot immediately distinguish between a true loss and a delayed packet. To manage this without the prohibitive memory cost of per-packet timers, modern DPUs like BlueField-3 employ hardware-resident bitmaps and per-Queue Pair (QP) gap timers [29]. While these end-to-end timers eventually detect all losses, including those caused by corruption, they do so without the benefit of explicit network signaling. Consequently, some recent systems are starting to rely on switch-side packet trimming for near-instant congestion notification [30], [31]. Critically, because corruption drops typically occur at the PHY/MAC layer before a packet is queued, the switch cannot generate a trim notification. This leaves best-effort transports to rely on the expiration of conservative hardware timeouts, which, while functional for detection, introduce significant tail latency ($P99$) spikes compared to the rapid recovery seen in congestion-induced drops.

Based on these observations, we propose **Ordered Chaos**, a unified, host-based load balancing and loss-recovery system for AI training and tightly-coupled HPC collectives. The key insight behind Ordered Chaos is that preserving *just enough ordering structure* enables fast, deterministic loss signaling without sacrificing the utilization benefits of fine-grained multipathing. Building on this insight, Ordered Chaos introduces two *novel and synergistic contributions*: *Reactive Subflow Spraying (RSS)* and *Precise Fast Loss Detection (PFLD)*.

RSS introduces a new operating point between per-flow hashing (e.g., ECMP/PLB) and per-packet spraying (e.g., RPS) by deterministically splitting each flow into σ logical subflows via round-robin mapping, such that each subflow carries a predictable packet subsequence. Rather than reacting at flow granularity, RSS periodically reroutes only the worst-performing subflow based on end-host congestion signals (e.g., RTT or ECN), achieving fine-grained adaptivity under asymmetry and failures while intentionally preserving in-order delivery within each subflow. PFLD exploits this preserved structure to enable a novel, *receiver-driven* loss recovery mechanism: by combining per-subflow sequence gaps with lightweight probes, PFLD deterministically distinguishes reordering from true loss and enables prompt recovery from congestion, corruption, tail, and retransmission losses without switch support or unbounded per-packet state. This receiver-driven, deterministic loss identification enables precise recovery actions and substantially reduces spurious or duplicate retransmissions compared to time-based, heuristic approaches.

Figure 1 illustrates how RSS preserves partial in-order delivery across subflows and how PFLD leverages this structure for fast and precise loss recovery. Ⓐ shows RSS on a symmetric and an asymmetric fat-tree topology with 16 subflows. In the asymmetric case, one of the 16 output links is throttled to 20% capacity; RSS reacts within microseconds, redistributing subflows so the degraded link carries proportionally fewer, converging to the ideal distribution (dotted lines). Ⓑ compares the 99th-percentile loss detection delay during an incast workload. With corruption drops ($\approx 10^{-5}$), RTO-based recovery takes 60–80 μs, and packet trimming cannot detect corruption drops because no trimmed header is produced. PFLD detects both congestion and corruption losses within a single RTT ($< 20\,\mu s$). We define *corruption drops* as losses not caused by congestion.

Overall, without any switch modifications, Ordered Chaos reduces the non-overlapped communication time of LLaMA-7B training by 28%, speeds up ring all-reduce by 35%, and improves tail flow completion time by up to 3× across AI collectives, HPC applications (ICON, LAMMPS, MILC), and microbenchmarks. Our main contributions are:

- **Reactive Subflow Spraying (RSS).** We introduce RSS, a subflow-level reactive load-balancing mechanism that partitions each flow into a small number of strictly ordered subflows and periodically reroutes only the worst-performing subflow based on simple end-host signals (e.g., RTT or ECN). RSS achieves near-uniform link utilization under asymmetric and dynamic conditions while requiring no switch modifications.
- **Precise Fast Loss Detection (PFLD).** We present PFLD, a receiver-driven loss detection mechanism that leverages *subflow-level ordering* to precisely detect and recover from congestion, tail, retransmission, and corruption losses without packet trimming. PFLD combines lightweight probe packets with packet-sequence-number (PSN)-aware entropy encoding to deliver fast and deterministic loss signals with minimal memory requirements.
- **Unified transport co-design with minimal state.** We co-design RSS and PFLD so that ordered subflows provide the structural signal needed for precise end-host loss detection, while accurate and timely loss feedback enables RSS to react aggressively without hardware assistance. This tight integration keeps per-flow state small (a few per-subflow counters and a receiver bitmap), ensuring scalability and robustness in the presence of asymmetry, failures, and corruption drops.
- **Evaluation at scale and on hardware.** We evaluate RSS+PFLD at scale using large-scale packet-level simulations over a best-effort underlying fabric, spanning AI collectives, end-to-end LLM training traces, and replayed HPC workloads, and further validate implementation feasibility via a programmable DPU prototype. Ordered Chaos improves tail flow completion time by up to 3× and accelerates collective and training communication by up to 35%.

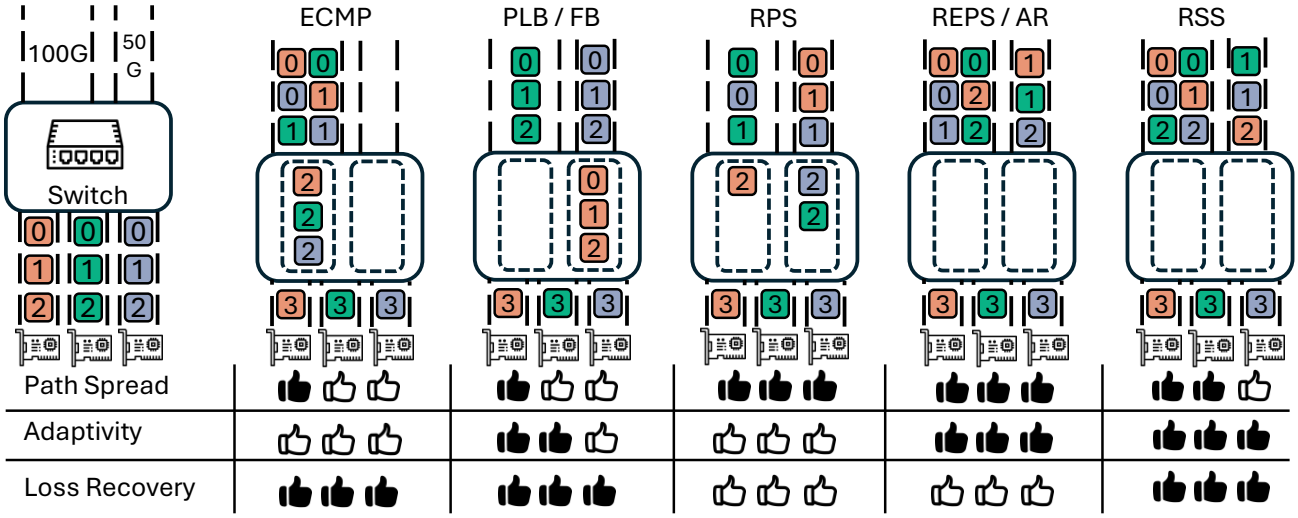


Figure 2: Routing decisions for different load balancing approaches using a sample switch with 3 50 Gbps input ports and 2 output ports (one running at 100 Gbps and the other at 50 Gbps). Packets flow upward, from the input ports at the bottom to the output ports at the top. *FB* stands for *FlowBender* and *AR* for *Adaptive Routing*.

II. Motivation

Multipath, out-of-order-capable transports are becoming the prevailing design choice, yet a fundamental tension remains unresolved. Existing multipath load balancing and transport schemes struggle to simultaneously deliver fine-grained load balancing, fast loss detection for both congestion- and corruption-induced drops, and end-host only deployability. Figure 2 summarizes the dominant multipath load-balancing approaches used in today's transports and highlights this tension: order-preserving designs such as ECMP and FlowBender (FB) underutilize path diversity and react slowly to asymmetry, whereas packet-spraying schemes such as Random Packet Spraying (RPS) and its adaptive variants (REPS/AR) achieve high utilization by sacrificing the ordering signals required for timely loss detection.

We further observe that, perhaps counter-intuitively, per-packet spraying is not necessarily needed to achieve high load-balancing effectiveness (Section V-B). RSS leverages this insight by operating at subflow granularity, retaining most of the utilization benefits of packet-level spreading while preserving controlled ordering. At the same time, although packet trimming can rapidly signal congestion-induced losses, it provides no notification for corruption drops, which occur before queueing. This limitation motivates the combination of ordered subflows (RSS) with fast, end-host loss detection (PFLD). In the remainder of this section, we characterize the prevalence and impact of corruption drops in production and explain why packet trimming, despite its effectiveness for congestion, is insufficient for fast loss recovery in this setting.

A. Corruption Drops in Production

Large-scale AI training and HPC jobs run for days to months across tens of thousands of links. At this scale, both systemic failures (link flaps, switch failures, routing changes) and random corruption drops are the norm: production AI clusters report 5K to 60K link-flap incidents per day [6], and corruption drops caused by aging optics, bit errors, connector degradation, and ASIC glitches occur continuously across such large deployments [9].

Production measurements report that inter-switch and inter-card loss events, which include silent packet drops and packets discarded after link corruption, account for about 18% of packet drops associated with network performance anomalies. These events take about 161 minutes to locate on average and account for half of the packet drops whose localization takes more than 180 minutes [1].

Corruption compounds across multi-hop paths. If the per-link corruption probability is p_{link} and a path traverses h links, the end-to-end failure rate is

$$p_{\text{end}} = 1 - (1 - p_{\text{link}})^h \approx h \cdot p_{\text{link}} \quad \text{for small } p_{\text{link}}.$$

For instance, with a per-link corruption rate of 5×10^{-5} (the rate we adopt for our evaluation, consistent with reported values [3], [32]) and a typical 6-hop datacenter path, the end-to-end loss probability is roughly 3×10^{-4}. This aligns with measurement studies that consistently report end-to-end loss ratios in the 10^{-5} to 10^{-4} range under normal operation [2]–[5], with localized degradations transiently pushing affected links into the 10^{-3} to 10^{-2} range before quarantine or repair.

These failure rates mean the network is almost always experiencing some fault or recovery event. Transports must remain effective under persistent change rather than only in steady state, making robust loss detection and recovery mechanisms critical for maintaining cluster efficiency [6].

B. Why Trimming Is Not Enough

To mitigate packet loss in the presence of reordering, packet trimming has been proposed [31], [33]: when a switch queue is full or reaches a threshold, the incoming data packet's payload is stripped and the remaining header is pushed into a high-priority queue. The receiver sends a NACK upon receiving a trimmed header, prompting the sender to retransmit.

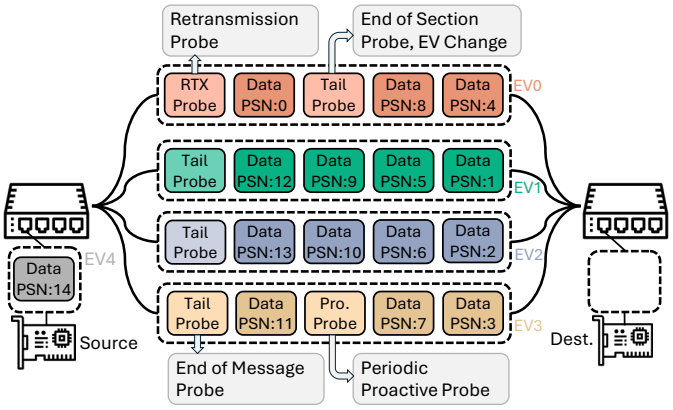


Figure 3: Detailed scheme of RSS+PFLD with 4 subflows. In the example we assume *PSN 0* is lost before the visualization. We also assume that each *EV* leads to a different path.

Although trimming is being adopted in several recent best-effort protocols [24], [27], due to its effectiveness in rapidly signaling congestion-induced losses, it has two fundamental limitations. First, many production deployments still lack trimming support entirely [34]. Second, and more critically, corruption drops occur *before* a packet is queued at the switch, leaving no trimmed header to signal the loss. In such cases, transports without end-host loss detection must fall back to retransmission timeouts, delaying recovery by orders of magnitude relative to the round-trip time and directly increasing job completion times.

Credit-based flow control similarly prevents congestion drops but not corruption drops that occur at the PHY/MAC layer before queueing. On such lossless fabrics, PFLD's role narrows to detecting corruption and link failures, while RSS remains useful when the fabric exposes endpoint-controlled path selection.

III. RSS Design

The RSS sender logic is detailed in Algorithm 1 and its defining characteristics are described next.

A. Subflow Spraying

The RSS mechanism deterministically partitions a single flow into σ parallel subflows each representing a $1/\sigma$ fraction of the aggregate bandwidth. This can be seen in Figure 3 which jointly illustrates the operation of RSS and PFLD with $\sigma = 4$. The parameter σ serves as a programmable tuning knob for load-balancing granularity: while $\sigma = 1$ reverts to static ECMP, higher values approach the high-entropy benefits of per-packet spraying. In this work, we fix $\sigma = 16$ to balance path diversity with reordering overhead in high-radix fabrics.

Unlike MPTCP [13], which necessitates independent state machines per subflow, RSS utilizes a shared flow state across all paths. Furthermore, while MPTCP achieves load balancing through fine-grained rate modulation ($bits/s$), RSS employs a *discrete steering* approach. By dynamically adjusting the number of subflows allocated to a specific physical path (see III-B), RSS provides reactive load balancing without the overhead of hardware rate-shapers or per-subflow timers.

To ensure compatibility with existing network fabric, RSS steers subflows by manipulating the entropy value (EV) in packet headers to leverage standard ECMP hashing at switches. Consistent with prior work [10], [12], we utilize the 16-bit source port, the IPv6 flow label [10], or the Ultra Ethernet entropy field [24] as the entropy source.

Algorithm 1 Pseudocode for an RSS sender using mean RTT.

Require: σ: number of RSS subflows; $N = \log_2(\sigma)$

```
1:  procedure updateEv(sid)
2:      if PSN-based PFLD is enabled then
3:          ev[sid] ← low16(nextPsn(sid))
4:      else
5:          ev[sid] ← (rand() ≪ N) + sid
6:      end if
7:  end procedure
8:  procedure onAck(ack)
9:      sid ← getSubflowId(ack)
10:     rttSum[sid] ← rttSum[sid] + ack.rtt
11:     samples[sid] ← samples[sid] + 1            ▷ RTT samples this interval
12:     if now ≥ nextRerouteTime then
13:         nextRerouteTime ← now +τ+ randomJitter()
14:         active ← {i | samples[i] > 0}
15:         if active ≠ ∅ then
16:             sid ← worstMeanRtt(rttSum, samples, active)
17:             updateEv(sid)
18:         end if
19:         reset(rttSum, samples)
20:     end if
21: end procedure

22: procedure onPacketSchedule(seqNo)
23:     return ev[seqNo mod σ]                     ▷ Round-robin subflow selection
24: end procedure
```

While subflow and flowlet abstractions are established in protocols like MPTCP [13] and adaptive routing schemes [19], RSS introduces a *deterministic mapping paradigm* that eliminates the ambiguity inherent in traditional grouping. Unlike FlowBender's [15] reactive, flow-level switching or the stochastic nature of gap-dependent flowlets, RSS provides a predictable, hardware-native sequence-to-path correspondence. By proactively partitioning the flow regardless of application-level "gaps," RSS creates the architectural foundation for PFLD, enabling fine-grained, low-latency loss detection without the overhead of tracking probabilistic reordering events.

In Algorithm 1, updateEv uses two EV encodings. For the power-of-two σ used here, let $N = \log_2(\sigma)$. Plain RSS randomizes the upper bits to select a new path while retaining the subflow ID in the low N bits. PSN-based PFLD instead uses the low 16 bits of the first PSN after a reroute; because subflow s carries PSNs p for which $p \bmod \sigma = s$, the low N bits still identify the subflow while the EV also marks the start of its new ordered section (Section IV).

B. Reactive Rerouting

While PLB [10] is limited to serial probing of a single path, RSS concurrently monitors σ paths to maintain a comprehensive view of network state. Unlike gap-dependent flowlet switching [14], which fails under the dense traffic typical of RDMA and AI collectives [35], RSS evaluates rerouting on a fixed schedule τ independent of traffic patterns. At the first feedback event at or after each deadline, the sender evaluates path-level metrics, such as average RTT or ECN marking fractions, and migrates the worst-performing subflow for which it collected

samples. Since the EV determines the switch-level routing, updating this field enables near-instantaneous rerouting to a new physical path. This parallel, scheduled approach ensures high-radix fabric utilization even during periods of maximum link saturation.

The update interval τ governs the stability and convergence of RSS as a control system. At flow creation, *nextRerouteTime* is initialized to the first update deadline. Algorithm 1 advances subsequent deadlines by τ plus a small random jitter. Path feedback is collected throughout each interval and reset after the rerouting decision. We recommend configuring τ as a function of the network RTT to ensure sufficient feedback, such as ECN marks or sample RTTs, is received before rerouting decisions are made. The jitter prevents global synchronization between independent flows, which can lead to periodic fabric oscillations. Setting τ too low risks rerouting from too few samples, while an excessively large τ increases convergence time and delays the response to transient hotspots. A deployment that must react to complete feedback loss can additionally schedule a no-feedback timer for each subflow. We demonstrate the impact of the choice of τ in Section V-F.

C. RSS's Deployability

RSS requires only end-host support; no changes to network devices are needed. Recent work likewise demonstrates endpoint-controlled, path-aware load balancing with standard Ethernet mechanisms in low-diameter networks [36]. It requires an end-host-settable path selector and endpoint support for out-of-order delivery across selectors while preserving order within each selector. Ethernet/RoCE exposes path selection through fields such as the UDP source port or IPv6 flow label, while Ultra Ethernet provides an entropy field [10], [12], [25]. Slingshot, InfiniBand, and Omni-Path can provide the same abstraction when the NIC exposes equivalent path handles or entropy controls. The design is not applicable when a fabric hides all path choice or enforces strictly in-order endpoint delivery. Out-of-order-capable stacks are now available from cloud providers [24], [27], [30] and in research prototypes [37].

RSS has a small, $O(\sigma)$ memory footprint. The implementation maintains an entropy value, feedback count, accumulated RTT, maximum RTT, and ECN count per subflow, plus the next-update timestamp. A hardware implementation can instantiate only the accumulators required by its selected metric; the mean-RTT configuration in Algorithm 1, for example, needs only the entropy, feedback count, and accumulated RTT vectors.

IV. PFLD Design

While RSS results in out-of-order delivery *across* subflows, it preserves *in-order delivery* within each subflow. Since packets within a subflow sharing an EV are guaranteed to traverse the same physical path (Section III-B), they maintain strict sequential ordering relative to their specific subflow. PFLD leverages this deterministic structure for high-fidelity loss detection.

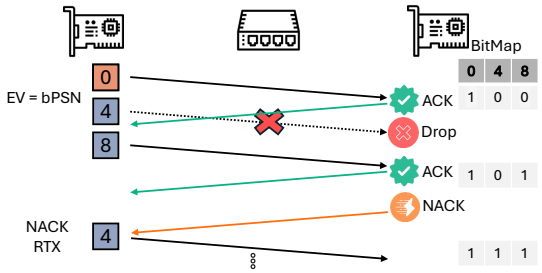


Figure 4: Operation principle of PSN-Based PFLD with 4 subflows. Only the first subflow is shown.

A. Design Space and PSN-Based PFLD

We considered several PFLD variants with progressively smaller state requirements, summarized in Table I. An *Infinite Memory* design tracks every in-flight PSN per EV, detecting gaps immediately but requiring state proportional to the bandwidth-delay product, which is prohibitive on NICs. A *Fixed Path* variant constrains each subflow to a static EV with an arithmetic PSN subsequence, reducing state to one expected-PSN counter per path plus a single 512-bit per-flow bitmap; however, it cannot support rerouting. A *Generation-Based* extension adds expected-PSN counters for each concurrent generation of every subflow while retaining one per-flow bitmap.

PSN-Based PFLD is our proposed state-optimal scheme that eliminates the need for explicit generation tracking. The key insight is to encode the start of each rerouted section directly in the EV. When RSS reroutes a subflow, it sets the EV to the low 16 bits of the next PSN scheduled on that subflow. The first packet on the new path carries this EV. Because the value is derived from that packet's PSN rather than chosen randomly, the receiver can recover the section's start PSN without a separate handshake or explicit generation state. A single 512-bit bitmap per flow records the received status of each packet relative to this start. Because round-robin spraying maps PSN p to subflow $p \bmod \sigma$, the lower $\log_2(\sigma)$ EV bits still identify the subflow. With $\sigma = 16$, the remaining 12 bits provide 4096 EVs per subflow. Generation PSNs vary across flows, message sizes, and reroutes; if two EVs collide on a poor path, RSS can reroute the affected subflow again.

Version	State to Store
Infinite Memory	#BDP Packets × 32b PSN
Fixed Path	$\#Paths$ × 32b ePSN + 512b bitmap
Gen-Based	$\#Paths \times \#Gen$ × 32b ePSN + 512b bitmap
PSN-Based	512b bitmap

Table I: PFLD variants and their memory footprint. Bitmaps enable selective NACK; dropping them implies go-Back-N.

In the receiver-driven variant (Figure 4), PSNs are global: subflow s carries exactly those PSNs for which $p \bmod \sigma = s$, and each EV identifies the start of one ordered section of that subsequence. The receiver lifts the 16-bit EV into its current absolute-PSN window, computes the packet's bitmap offset, and marks the corresponding bit. Packets between the maximum received PSN index and the first unset bit are identified as lost and NACKed immediately for retransmission. Upon

receiving NACKs, the sender enqueues the missing packets for retransmission.

Data placement is identical to RPS or another out-of-order transport and requires no separate reorder buffer: each packet carries its PSN and offset, so an out-of-order-capable RDMA stack places the payload directly at its final location. PFLD's per-flow bitmap is used only for loss detection. Running PFLD at the receiver provides two benefits. First, the receiver is the first endpoint with positive evidence of loss: a later packet or probe has arrived on the same ordered subflow while an earlier PSN has not. Second, it can immediately NACK the precise gap. A sender sees only missing ACKs, which may instead reflect ACK coalescing, ACK loss, or reverse-path delay and therefore require additional feedback or time-based thresholds such as RACK-TLP.

The required bitmap size is determined by the maximum unresolved PSN span in an active section and scales with rate, RTT, and packet size. At 400 Gb/s and 10 μs RTT, the NIC-wide BDP spans about 122 packets with a 4 KiB maximum transmission unit (MTU) and 489 packets with 1 KiB packets. We therefore use a 512-bit bitmap, which covers both cases; this choice sizes the tracking window rather than constraining PFLD correctness to a particular MTU. For 512-byte packets or larger BDPs, a deployment can use a larger bitmap, such as 1024 bits, or overlapping generations without changing the detection logic. When RSS reroutes a subflow, the sender emits a tail probe on the old EV to close the section as shown in Figure 3. The bitmap is indexed by absolute PSN and slides forward monotonically, so late-arriving packets on the old path are marked within the existing window. NACKs are scoped to each section's PSN range, preventing cross-section spurious retransmissions.

Figure 5 shows the memory occupancy per NIC for all variants assuming a 400 Gb/s network with a 10 μs RTT and a 4 KiB MTU. PSN-Based PFLD requires less than 100 KiB even at 1000 active flows, while the Infinite Memory variant reaches about 480 KiB. The bitmap size can be adjusted based on deployment requirements and is not tied to the core RSS or PFLD design.

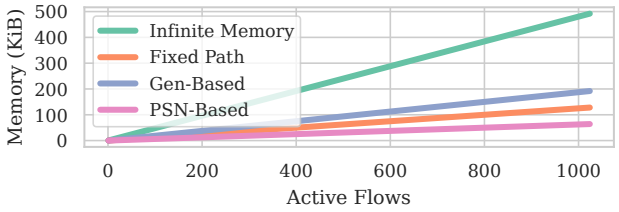
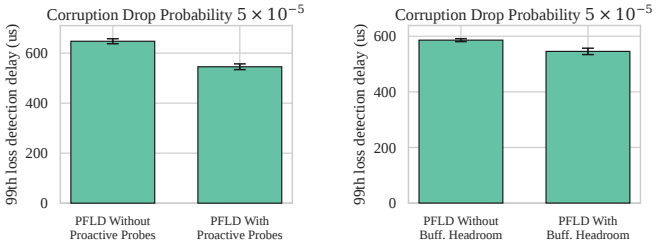


Figure 5: PFLD Memory Consumption Per-NIC.

B. Probing for Full Reliability

PSN-based loss detection identifies drops whenever a subsequent packet arrives on the same subflow and reveals a gap. However, three corner cases require additional mechanisms: tail losses (no subsequent packet exists), prolonged loss bursts (detection waits indefinitely), and retransmission losses (non-sequential PSNs break gap detection). We address all three with lightweight probe packets shown in Figure 3.



(a) PFLD with proactive probes on and off during an incast.

(b) PFLD with and without dedicated buffer headroom.

Figure 6: Impact of probes on PFLD at corruption probability 5×10^{-5}.

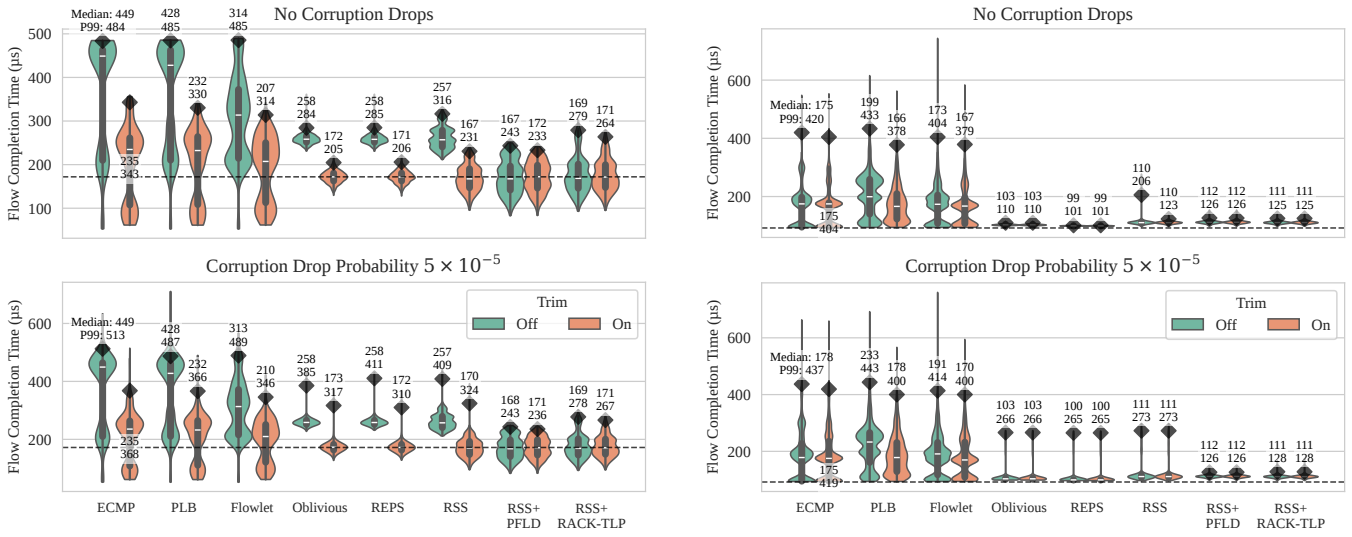
1) Probe Packet for Tail Loss: The PSN-based gap mechanism alone cannot detect the loss of the last packet before an EV change or the last packet of a subflow, because no subsequent packet exists to reveal the gap. PFLD handles these cases with tail probes. With RSS, each message has multiple tail packets (one per subflow, plus additional ones at each reroute), so the sender appends a header-only *tail probe* after the last data packet on each subflow. The probe carries the last PSN that should have been received before it, allowing the receiver to NACK any missing PSNs up to and including that value.

2) Proactive Probing: The gap-based mechanism only detects losses after receiving a packet. If a burst of consecutive packets is lost, detection is delayed until after the RTO, causing spurious retransmissions. The sender prevents this with periodic *proactive probes*: header-only, low-priority packets that ensure PSN synchronization. Upon receiving a probe, the receiver NACKs any unreceived PSNs with the same EV. Probes are small and less likely to be dropped, and increase the chance of detecting losses before the RTO fires (Figure 6a).

3) Retransmitted (RTX) Packet Probing: Retransmitted packets' PSNs don't respect expected sequential ordering. As a result, the receiver cannot identify if retransmitted packets are lost. This motivates accompanying each RTX packet with a per-packet *RTX probe*. This header-only probe carries the PSN of the retransmission. If the receiver receives the probe but not the corresponding data packet, it NACKs that PSN.

Probe loss affects latency, not correctness. If an RTX packet, probe, or NACK is lost, the sender falls back to the normal retransmission timeout. At a regular RSS update, Algorithm 1 considers only subflows that returned feedback during the current interval. A completely silent subflow is therefore rerouted only when the optional no-feedback timer described in Section III-B is enabled. Later packets or probes on an active section can expose earlier missing PSNs. Common losses are therefore detected before timeout, while pathological combinations remain RTO-bounded.

4) Headroom Buffer for Probe Packets: Probe packets play an essential role in PFLD: they detect tail losses, synchronize state, and protect retransmissions. Although they are header-only packets and thus less likely to be dropped, they are enqueued with data packets and remain exposed to congestion. As an



(a) 1 MiB 8:1 Incast

(b) 4 MiB Tornado Pattern

Figure 7: Microbenchmarks comparing different schemes with and without corruption drops. Positive-drop panels use corruption probability 5×10^{-5}. For each scheme, we also evaluate with and without packet trimming. The dashed line indicates the ideal FCT and the rhombus marker the 99th percentile.

optional optimization, a deployment may reserve a small switch-buffer headroom for probes; it is not required for correctness or for the end-host-only design. Our evaluation shows that less than 0.2% of the low-priority queue is sufficient to avoid probe drops entirely (Figure 6b).

V. Large Scale Simulations

We implement RSS and PFLD in the htsim simulator [31], and run large-scale per-packet simulations.

A. Methodology & Setup

We simulate a 2-tiered fat-tree topology with 1,024 nodes with 64-port switches. We set the MTU to 4 KiB and the latency of all links to 1.5 μs (*i.e.* a minimum RTT of 12 μs when traversing both tiers).

We compare several network settings: a symmetric case with all links having 400 Gbps capacity, and an asymmetric case with 4 out of the 16 aggregate switches having all of their links reduced to 100 Gbps. The congestion control algorithm used is MPRDMA [35], which manages window-based congestion control on a per-flow basis. We replace only its path-selection logic with each evaluated scheme. Under RSS, MPRDMA's window is shared across all subflows; each subflow transmits at $1/\sigma$ of the aggregate rate, so RSS controls *where* packets go while MPRDMA controls *how fast* they are sent. Under ECMP and PLB, MPRDMA operates on a single path; under RPS and REPS, each packet is placed on a path independently.

We configure the retransmission timeout (RTO) for each experiment based on the topology's maximum unloaded RTT, with additional allowance for queueing delay. Unless noted otherwise, we set the corruption drop rate to 5×10^{-5} (Sec. II-A). We verified that our results hold qualitatively across rates from 10^{-6} to 10^{-4}: lower rates reduce the absolute gap between schemes because fewer drops occur, while higher rates amplify PFLD's advantage because more losses require fast detection. The LLaMA experiment uses 5×10^{-4}, and the finite HPC trace replays use 5×10^{-3} so that enough losses occur to compare recovery behavior.

1) Simulation Workloads: We provide an extensive evaluation section with simulation-based results. We use ATLAHS [38] with htsim to simulate advanced workloads. The workloads are summarized below:

- **Microbenchmarks:** These are workloads classically used by previous works including Incast (many-to-one), and Tornado (one-to-one but all going across the tree to stress the load balancing) traffic patterns.
- **AI-collectives:** These include the most common collectives used for synchronous AI training: all-reduce using the ring algorithm [39] and all-to-all [40].
- **End-to-end AI traces:** We use ATLAHS [38] to collect end-to-end traces for a single training iteration of the LLaMA 7B model [41], an established benchmarking model [42].
- **HPC application traces:** We replay MPI traces from three representative HPC applications: ICON (climate modeling) [43], LAMMPS (molecular dynamics) [44], and MILC (lattice QCD) [45].
- **Data center traces:** We sample flow sizes from a Google RPC distribution at 80% offered load to evaluate production-like message-size diversity [10].

2) Schemes: We evaluate RSS with and without PSN-based PFLD. Section V-F shows the impact on RSS of the number of subflows and the update interval. For all other experiments, we set the number of subflows to 16, and the update interval to 15 μs, roughly equal to the RTT value. We compare it against the following baselines:

- **ECMP [46]:** Hashes flow-level information to pick a single path for all packets in a flow. The selected path is maintained

for the lifetime of the flow.

- **PLB [10]**: Operates similarly to ECMP but proactively reroutes flows that see high ECN marking rates. We choose the same parameters as advised in the Flowbender [15] paper (the ancestor of PLB); the ECN threshold is set to 0.05 and flows are rerouted every congestion period.
- **Flowlet Switching [14]**: Splits a flow into flowlets, bursts separated by an inter-packet gap above a threshold. Upon gap detection, the next flowlet is hashed to a different path.
- **RPS [18]**: Randomly selects a path per packet, providing fine-grained load balancing without congestion awareness.
- **REPS [12]**: A lightweight decentralized per packet adaptive load balancer. It sprays packets across multiple paths while caching high-quality paths and reusing that entropy to bias future choices. Compared to RPS, it can react to asymmetries and failures.
- **RACK-TLP [47], [48]**: The state-of-the-art sender-side loss recovery mechanism, standardized in RFC 8985 and used in Linux and Google's datacenter stack. We pair RACK-TLP with RSS for load balancing to approximate proprietary Falcon [26] architecture. This isolates the effect of PFLD's receiver-driven loss detection versus RACK-TLP's sender-side heuristics under identical load balancing.

B. Microbenchmark Evaluation

We use a set of standard microbenchmarks to assess both load balancing performance and loss-detection reactivity. For each workload, we evaluate the performance with and without corruption drops. Each scheme is run with and without packet trimming. To account for randomness, we run each configuration 100 times with different random seeds and report violin plots over all flows across runs. In grey we also report a dotted line with the ideal completion time assuming perfect fairness.

In Figure 7a, we showcase an 8:1 incast where each sender transmits 1 MiB to a receiver. As expected, with packet trimming enabled and no corruption drops, schemes with finer granularity, such as REPS, RSS, and RPS, perform best. However, once trimming is disabled, only RSS+PFLD maintains the same performance because it detects losses more quickly. With corruption drops enabled, the main advantage of PFLD becomes clear. Even with trimming, corruption drops remain problematic for the other schemes, and RACK-TLP has a delayed reaction, while PFLD reacts proactively to all kinds of drops. This makes the 99th-percentile FCT with RSS+PFLD roughly half that of ECMP. PFLD's probes add a small common-case overhead, which can make its median slightly higher than RACK-TLP's, but immediate receiver-side evidence provides the larger improvement at the tail.

Then, in Figure 7b, we focus on a Tornado pattern, where each node sends to and receives from its symmetric node in the other half of the tree. This ensures that all packets of all flows traverse the network, stressing the load balancer. The trends mirror the incast case: without corruption drops, RSS+PFLD is only slightly slower than per-packet schemes; when trimming is disabled and drops occur, RSS+PFLD gains a substantial lead, up to 2.5× at the 99th.

Across both workloads, we also measured the total number of retransmission timeouts triggered. With PFLD, timeouts are reduced by 99% compared to ECMP, both with and without trimming. In practice, the only remaining timeouts with PFLD are edge cases such as lost NACKs or simultaneous loss of a probe and all its associated data packets. This confirms that PFLD's probing mechanisms effectively eliminate RTO as the dominant recovery path, which is the root cause of the tail latency improvements observed above.

C. AI Workloads

We show simulated results for two commonly used collectives in AI training: the all-reduce implemented via the ring algorithm [39], and the all-to-all implemented using an algorithm where we limit the number of parallel connections per node [7], [40]. Later, we also use ATLAHS [38] to run end-to-end training workloads using open-source traces [49].

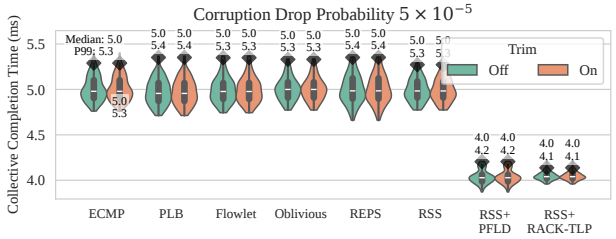


Figure 8: 8 MiB ring all-reduce with 5×10^{-5} corruption.

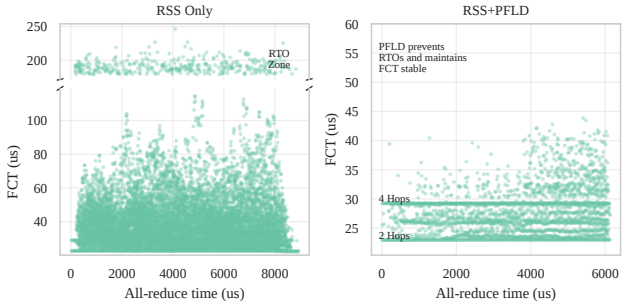


Figure 9: Individual flow completion time of all-reduce flows.

In Figure 8, we show all-reduce performance with corruption drops. We omit results without corruption drops because they are indistinguishable on our fully provisioned topology. We run the all-reduce on 128 nodes 100 times and plot the distribution of collective completion times. PFLD and RACK-TLP outperform all other approaches because they provide an accurate loss-detection signal, yielding about a 35% speedup. Their median completion times are essentially identical; at the 99th percentile, RACK-TLP is marginally faster (about 4.1 ms versus 4.2 ms for PFLD).

A more important benefit of PFLD is improved synchronization of the ring. Under ideal conditions, a ring all-reduce should not induce congestion. However, if a flow is slowed down, incast can arise and compound quickly. This is evident in Figure 9, which shows flow completion times over time for RSS with and without PFLD. With PFLD there is a clear pattern, with

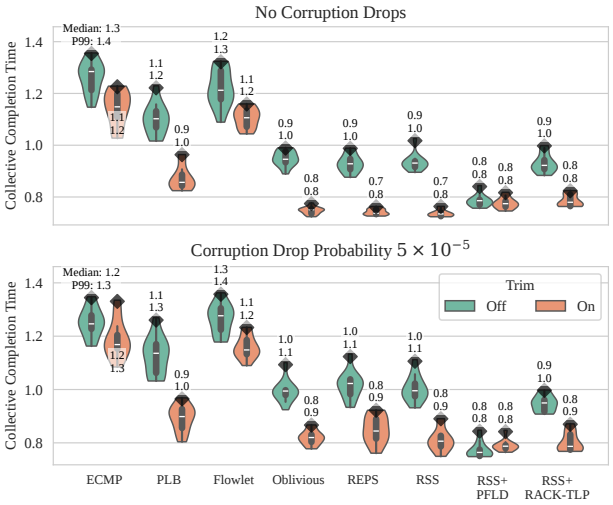


Figure 10: 1 MiB all-to-all. Positive-drop panels use corruption probability 5×10^{-5}.

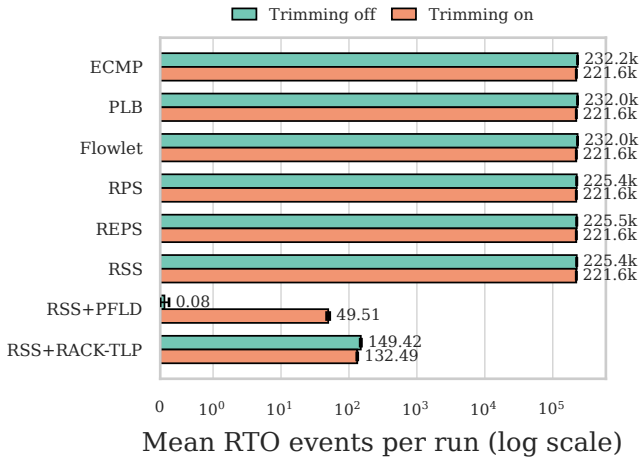


Figure 11: RTO events for a 128-node all-to-all with 4 MiB per directed flow and at most 32 active senders. Bars report means across 100 matched seeds; across-seed variation is negligible at this scale.

flows completing at a small set of values depending on hop count; without PFLD, the pattern is chaotic. Flows may time out, stalling part of the ring while other flows continue, which triggers incast and disrupts synchronization.

In Figure 10 we evaluate performance of all-to-all. Without corruption, per-packet load balancers perform slightly better than PFLD with trimming. However, without trimming, PFLD gets a 20% lead over the second-best algorithm with and without corruption drops.

We also take a deeper look at RTO counts for the all-to-all experiment in Figure 11. RSS+PFLD averages only 0.08 RTOs per run without trimming and 49.51 with trimming, versus 149.42 and 132.49 for RSS+RACK-TLP. All other schemes exceed 221,000 RTOs per run. Thus, RSS+PFLD makes timeouts exceptional without trimming and remains well below RSS+RACK-TLP with trimming.

We also evaluate LLaMA 7B using the traces available at [49]

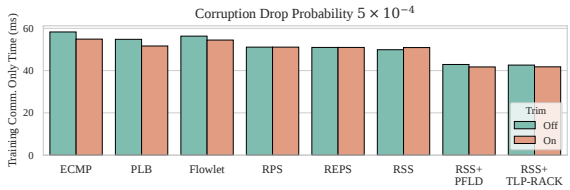


Figure 12: LLaMA 7B training: communication time per iteration at corruption probability 5×10^{-4}.

in Figure 12. We measure the time to complete one training iteration and plot only the exposed communication portion since the network cannot improve the compute-heavy parts. Compared to ECMP, RSS+PFLD reduces exposed communication time by 28% and achieves a 10% improvement over REPS and RPS. Notably, while trimming improves the performance of single-path schemes, it provides little benefit for multipath schemes.

D. HPC Workloads

We evaluate RSS+PFLD on execution traces from representative HPC applications, replayed end-to-end under random corruption drops. In contrast to the synthetic tests, these traces expose smaller headroom for load-balancing gains: the standard load-balancing baselines remain relatively close in communication-only and end-to-end runtimes, while the tail FCT is significantly lower with loss recovery schemes. We report mean FCT and p99 FCT to show tail behavior, communication-only runtime to isolate the network effect, and end-to-end trace runtime to quantify the overall application impact. Across all workloads in Figure 13, RSS+PFLD is the fastest, with RACK-TLP being second. In MILC, the most communication-heavy workload, PFLD improves tail FCT by about 1.6× over RACK-TLP and communication-only runtime by up to 15% over ECMP.

E. Data Center Traces

We evaluate a workload based on Google RPC size distributions [10]. We sample flow sizes from the empirical distribution, adjust interarrival times to reach 80% offered load, and simulate approximately 100,000 flows. Figure 16 groups results into 0 to 10 KB, 10 KB to 1 MiB, and larger than 1 MiB. For the smallest messages, medians and tails remain similar because a short flow rarely contains a corruption drop. With corruption drops, RSS+PFLD roughly halves the 99th-percentile FCT of 10 KB to 1 MiB messages relative to schemes without specialized recovery. For messages larger than 1 MiB, it lowers median FCT by about 41% relative to ECMP and remains competitive with RSS+RACK-TLP across both trimming settings.

F. Parameter Sensitivity

In Figure 14 we show the impact of RSS's update interval τ and number of subflows σ on the maximum FCT. In general, adding more subflows helps improve performance, but gains are limited after a certain point (e.g., 16/32 subflows). Counterintuitively, adding subflows can actually have a negative impact in certain cases: 1) More subflows require more state. 2) More subflows in RSS make the PFLD reaction time slower as each subflow might have few packets in-flight per RTT. 3)

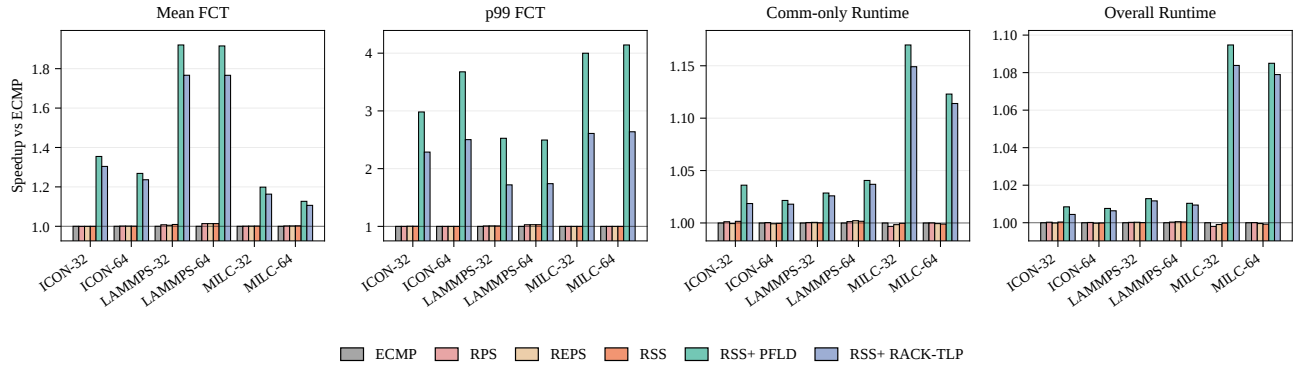
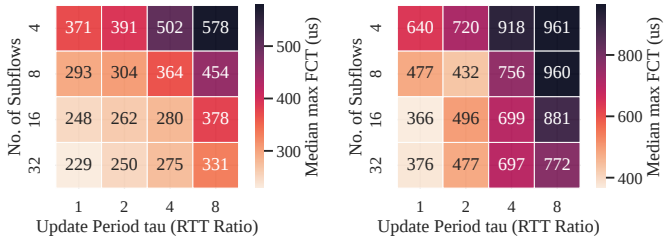


Figure 13: Performance on replayed HPC execution windows, normalized to ECMP. Real HPC traces expose limited headroom for pure load balancing, so the main gains of RSS+PFLD come from faster and more precise loss recovery, which sharply reduces FCT tails and shortens communication phases.



(a) Symmetric Topology (b) Asymmetric Topology

Figure 14: RSS's sensitivity to changes in σ and τ during a tornado workload to stress its load balancing capabilities.

Increasing the number of subflows means that each subflow gets less information and RSS might make uninformed routing decisions. For τ, a small but reasonable (multiple of RTT) value is preferred for frequent updates.

We then run an ablation study, shown in Figure 15, to evaluate the impact of PFLD's different probe types and frequencies. For tornado traffic, section and tail probes halve RTO events from 112.8 to 56.7; periodic probes add little in this case, while retransmission probes eliminate the remaining exposed RTOs. Under 32:1 incast, periodic frequency matters: $P = 16$ remains near 2,600 RTO events, while $P = 8$, $P = 4$, and $P = 1$ reduce

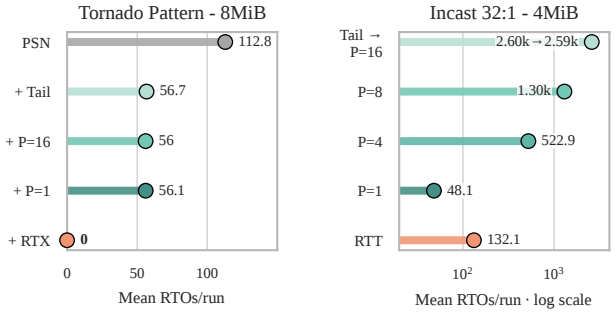


Figure 15: PFLD probe ablation. $P = k$ denotes one periodic proactive probe after every k data packets; RTT denotes at most one periodic probe per RTT. Tail → $P = 16$ summarizes the nearly identical section/tail-only and $P = 16$ settings, while RTX adds retransmission probes. Both workloads are run without trimming and with corruption; variation across seeds is negligible at this scale.

the mean to 1,300, 522.9, and 48.1, respectively; probing at most once per RTT yields 132.1. Each of the 16 subflows sends infrequently during this large incast, so ordinary data packets provide sparse ordering evidence and RTT probes are especially useful. Probe type and frequency could adapt dynamically to traffic pace and network conditions. Retransmission probes instead address the residual recovery failures in the tornado workload.

VI. Hardware Feasibility Validation

To complement the large-scale simulation study, we validate the practical feasibility of PFLD's probing mechanisms on a programmable Data Processing Unit (DPU; specifications withheld for confidentiality). The goal of this prototype is not to replicate the full-scale evaluation of Section V, but to answer three research questions: (1) does PFLD's probing logic introduce measurable overhead on real hardware? (2) does the implementation produce spurious retransmissions? and (3) do the different probing variants (tail, retransmission, proactive) behave on hardware as predicted by simulation?

a) Setup: We implement RSS and PFLD on the DPU's programmable data path, maintaining receiver-side loss-detection state and probe-generation logic in the DPU's memory. The prototype runs in loopback mode, continuously sending fixed-size messages (4 KiB to 1 MiB) for 20 seconds per configuration. This message range represents the lower end of typical data-plane traffic and is a conservative setting for overhead evaluation, since probing costs are amortized over fewer data packets.

b) Probing overhead is negligible: We compare a baseline RSS-only implementation against one incorporating PFLD's probing. As shown in Figure 17a, retransmission probes add zero overhead (as expected, since no drops occur in this experiment and hence no retransmission probes are generated). Tail probes add about 6% overhead for the smallest messages (4 KiB, where each message is a single MTU-sized packet), and converge to zero once messages exceed 128 KiB. This confirms that PFLD's probing is practical even on resource-constrained DPU hardware.

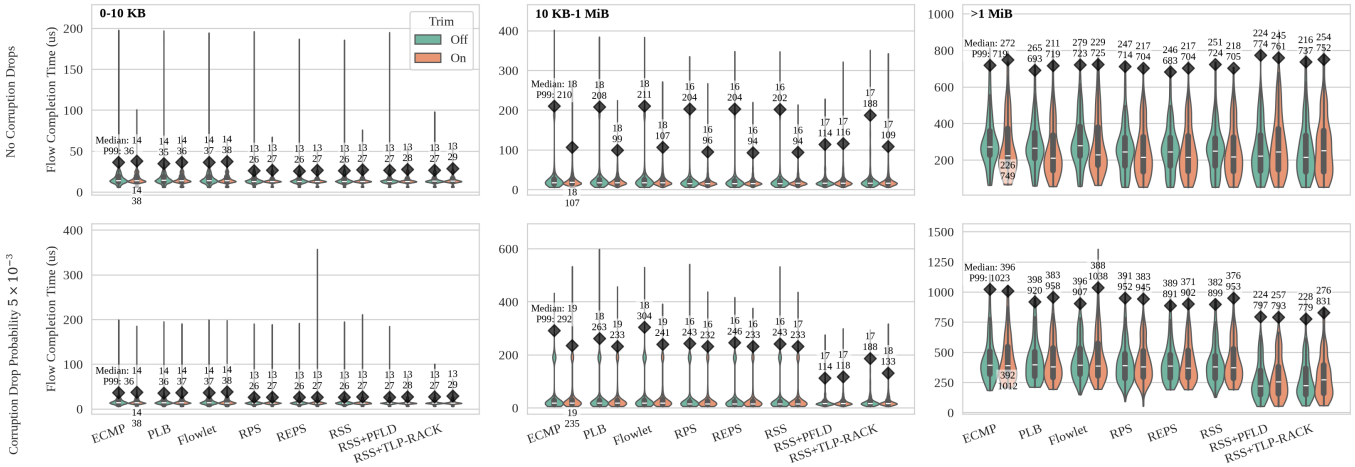
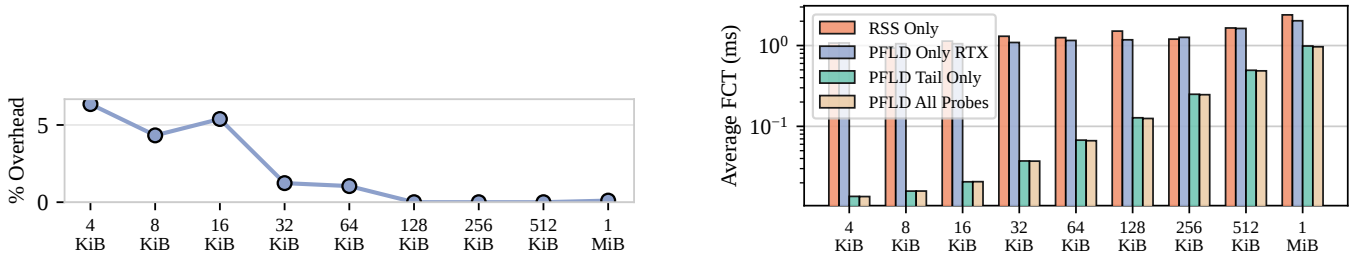


Figure 16: Flow completion times for a Google RPC workload at 80% offered load, grouped by message size. The top row disables corruption drops and the bottom row uses corruption probability 5×10^{-3}. Diamonds mark the 99th percentile.



(a) Performance overhead of probes.

(b) Observed average FCT with 1s timeout and 0.001 drop rate.

Figure 17: PFLD DPU hardware prototype results.

c) Functional correctness and probing benefits: We emulate a corruption drop rate of 0.001 at the receiver and verify that PFLD's retransmission count exactly matches the drop count across all variants, with no spurious retransmissions. Figure 17b compares the average FCT under a 1-second RTO timeout. RSS with only out-of-sequence detection suffers frequent timeouts for tail losses, as does PFLD with only retransmission probes. Adding tail probes eliminates these timeouts, reducing average FCT by up to 2.48× for 1 MiB messages and substantially more for smaller messages. The full PFLD configuration (all probe types) matches or improves on every variant, confirming the results from Section IV.

VII. Related Work

To maximize network utilization, the field has recently moved towards finer-grained load balancing. While flow-based schemes like Flowbender [15] and PLB [10] offer improvements over ECMP, their granularity is limited. Approaches that split flows, such as flowlet switching [14], trade load balancing precision for packet ordering, a compromise that is less effective for dense RDMA traffic which often lacks the idle intervals needed for rerouting [35]. Per-packet schemes like RPS [18] and DRB [50] offer the best granularity but falter in asymmetric topologies or demand global network knowledge. DRILL [16], Adaptive Routing (AR) switch-based solutions [8], [30], SGLB [17], and REPS [12], also proposed as part of UE [24], solve these issues to different degrees but introduce challenges: SGLB requires global switch coordination, AR demands switch modifications, and REPS struggles with ACK coalescing.

Ordered Chaos builds upon the evolution of these multipath transports by addressing key limitations in how these designs couple load balancing with loss detection. Early multipath approaches such as *MPTCP* primarily target bandwidth aggregation and resilience by operating multiple underlying TCP subflows, while relying on conventional per-flow path selection (e.g., ECMP) within the network [13]. Ordered Chaos extends beyond this model by introducing RSS subflows that share a single connection context and by explicitly structuring load balancing to preserve controlled ordering, allowing loss detection to be integrated directly into the multipath design rather than layered atop replicated transport instances.

Fine-grained load balancing inherently introduces out-of-order (OOO) packet arrivals. This breaks traditional RDMA reliability mechanisms like IRN [51], which conflate reordering with packet loss and trigger waves of spurious retransmissions. Recognizing this, some solutions rely on switch support for explicit loss notification. Works like NDP [31], CP [33], and more recently DCP [34] for RDMA, use packet trimming to

have the switch convert a dropped data packet into a header-only notification. Link Guardian [4] takes a complementary approach with link-local retransmission at the switch level, masking corruption losses before they reach the endpoint. While effective, these approaches require deploying new logic across the network fabric. More recently, MRC extends RoCEv2 with per-packet multipathing and selective reliability feedback over best-effort Ethernet [52]; Ordered Chaos explores a complementary subflow-ordered design point. Other end-host solutions present their own trade-offs. For instance, to avoid spurious retransmissions with its Adaptive Routing, the NVIDIA SuperNIC relies on timeouts for loss recovery. Falcon [26] pairs subflow-based multipathing with RACK-TLP [47], [48] for loss recovery, an architecture conceptually similar to RSS; however, RACK-TLP's sender-side heuristics degrade when subflows carry few packets. Other approaches either lack precision, like the predictive heuristics in STrack [53], or trade efficiency for reordering tolerance, like RACK-TLP itself, which delays retransmission by a full round-trip time and incurs significant memory overhead. MPRDMA [35] addresses multipath congestion control and is orthogonal to our work; we use it as the congestion control layer in our evaluation.

Ordered Chaos complements the RACK-TLP approach by reshaping multipath load balancing itself: RSS distributes traffic at subflow granularity rather than per packet, retaining most of the benefits of dynamic spraying while preserving partial flow order with in-order delivery guarantees within each subflow. This preserved structure enables PFLD to provide *receiver-driven*, deterministic loss identification and precise recovery with minimal duplicate retransmissions, even under aggressive multipathing, and without reliance on timers, heuristic thresholds, or special switch features.

VIII. Conclusion

The emergence of multipath-native transports, exemplified by specifications like Ultra Ethernet, signals an industry-wide shift towards best-effort out-of-order delivery to maximize network performance. However, this embrace of performance-oriented *chaos* creates a critical reliability gap, as fully out-of-order transports often lack robust mechanisms for fast and precise loss detection. In this work, we presented Ordered Chaos, a unified, end-host system that directly addresses this gap. Our design combines Reactive Subflow Spraying (RSS), which provides adaptive, fine-grained load balancing using ordered subflows, with Precise Fast Loss Detection (PFLD), which leverages order within flows and adds lightweight probing to rapidly identify losses from both congestion and corruption while being end-host only deployable. Our DPU-based prototype further confirms that PFLD's probing mechanism is practical for implementation on real hardware with negligible overhead. Additionally, our evaluation across microbenchmarks, AI collectives, end-to-end LLM training, and replayed HPC application traces demonstrates that this synergy significantly improves performance: RSS+PFLD improves tail flow completion time by up to 3×, reduces timeouts by up to 99%, and speeds up collective and training communication by up to 35%.

Acknowledgments

This research is supported by the National Research Foundation, Singapore (NRF), and the Ministry of Digital Development and Information (MDDI) under the AI Visiting Professorship (Award No. AIVP-2025-005). This work also received support from the ERC PSAP project (grant agreement No. 101002047). We also thank the Swiss National Supercomputing Center (CSCS) for providing the computational resources used in this work.

The authors used ChatGPT 5.4 and Claude Opus 4.6 for editing throughout the paper. All ideas and content are exclusively the original work of the authors.

References

[1] Y. Zhou, C. Sun, H. H. Liu, R. Miao, S. Bai, B. Li, Z. Zheng, L. Zhu, Z. Shen, Y. Xi, P. Zhang, D. Cai, M. Zhang, and M. Xu, "Flow event telemetry on programmable data plane," in *Proceedings of the 2020 Annual Conference of the ACM Special Interest Group on Data Communication on the Applications, Technologies, Architectures, and Protocols for Computer Communication*. Association for Computing Machinery, 2020, pp. 76–89. [Online]. Available: https://doi.org/10.1145/3387514.3406214

[2] C. Guo, L. Yuan, D. Xiang, Y. Dang, R. Huang, D. Maltz, Z. Liu, V. Wang, B. Pang, H. Chen, Z. Lin, and V. Kurien, "Pingmesh: A large-scale system for data center network latency measurement and analysis," in *Proceedings of the ACM SIGCOMM Conference*. London, United Kingdom: ACM, 2015, pp. 139–152. [Online]. Available: https://doi.org/10.1145/2785956.2787496

[3] D. Zhuo, M. Ghobadi, R. Mahajan, K. Förster, A. Krishnamurthy, and T. E. Anderson, "Understanding and mitigating packet corruption in data center networks," in *Proceedings of the ACM SIGCOMM Conference*. Los Angeles, CA, USA: ACM, 2017, pp. 362–375. [Online]. Available: https://doi.org/10.1145/3098822.3098849

[4] R. Joshi, C. H. Song, X. Z. Khooi, N. Budhdev, A. Mishra, M. C. Chan, and B. Leong, "Masking corruption packet losses in datacenter networks with link-local retransmission," in *Proceedings of the ACM SIGCOMM 2023 Conference*, ser. ACM SIGCOMM '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 288–304. [Online]. Available: https://doi.org/10.1145/3603269.3604853

[5] A. Roy, H. Zeng, J. Bagga, and A. C. Snoeren, "Passive realtime datacenter fault detection and localization," in *14th USENIX Symposium on Networked Systems Design and Implementation (NSDI)*. Boston, MA, USA: USENIX Association, 2017, pp. 595–612. [Online]. Available: https://www.usenix.org/conference/nsdi17/technical-sessions/presentation/roy

[6] K. Qian, Y. Xi, J. Cao, J. Gao, Y. Xu, Y. Guan, B. Fu, X. Shi, F. Zhu, R. Miao, C. Wang, P. Wang, P. Zhang, X. Zeng, E. Ruan, Z. Yao, E. Zhai, and D. Cai, "Alibaba hpn: A data center network for large language model training," in *Proceedings of the ACM SIGCOMM 2024 Conference*, ser. ACM SIGCOMM '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 691–706. [Online]. Available: https://doi.org/10.1145/3651890.3672265

[7] T. Hoefler, T. Bonato, D. De Sensi, S. Di Girolamo, S. Li, M. Heddes, D. Goel, M. Castro, and S. Scott, "Hammingmesh: A network topology for large-scale deep learning," *Commun. ACM*, vol. 67, no. 12, p. 97–105, Nov. 2024. [Online]. Available: https://doi.org/10.1145/3623490

[8] D. De Sensi, S. Di Girolamo, K. H. McMahon, D. Roweth, and T. Hoefler, "An in-depth analysis of the slingshot interconnect," in *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, 2020, pp. 1–14. [Online]. Available: https://doi.org/10.1109/SC41405.2020.00039

[9] A. Gangidi, R. Miao, S. Zheng, S. J. Bondu, G. Goes, H. Morsy, R. Puri, M. Riftadi, A. J. Shetty, J. Yang, S. Zhang, M. J. Fernandez, S. Gandham, and H. Zeng, "Rdma over ethernet for distributed training at meta scale," in *Proceedings of the ACM SIGCOMM 2024 Conference*, ser. ACM SIGCOMM '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 57–70. [Online]. Available: https://doi.org/10.1145/3651890.3672233

[10] M. A. Qureshi, Y. Cheng, Q. Yin, Q. Fu, G. Kumar, M. Moshref, J. Yan, V. Jacobson, D. Wetherall, and A. Kabbani, "Plb: congestion signals are simple and effective for network load balancing," in *Proceedings of the ACM SIGCOMM 2022 Conference*, ser. SIGCOMM '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 207–218. [Online]. Available: https://doi.org/10.1145/3544216.3544226

[11] K. He, E. Rozner, K. Agarwal, W. Felter, J. Carter, and A. Akella, "Presto: Edge-based load balancing for fast datacenter networks," in *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication*, ser. SIGCOMM '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 465–478. [Online]. Available: https://doi.org/10.1145/2785956.2787507

[12] T. Bonato, A. Kabbani, A. Ghalayini, M. Papamichael, M. Dohadwala, L. Gianinazzi, M. Khalilov, E. Achermann, D. D. Sensi, and T. Hoefler, "REPS: Recycled entropy packet spraying for adaptive load balancing and failure mitigation," in *Proceedings of the 21st European Conference on Computer Systems*, ser. EuroSys '26. ACM, Apr. 2026, pp. 225–246. [Online]. Available: https://doi.org/10.1145/3767295.3769320

[13] C. Raiciu, S. Barre, C. Pluntke, A. Greenhalgh, D. Wischik, and M. Handley, "Improving datacenter performance and robustness with multipath tcp," in *Proceedings of the ACM SIGCOMM 2011 Conference*, ser. SIGCOMM '11. New York, NY, USA: Association for Computing Machinery, 2011, p. 266–277. [Online]. Available: https://doi.org/10.1145/2018436.2018467

[14] E. Vanini, R. Pan, M. Alizadeh, P. Taheri, and T. Edsall, "Let it flow: Resilient asymmetric load balancing with flowlet switching," in *14th USENIX Symposium on Networked Systems Design and Implementation (NSDI 17)*. Boston, MA: USENIX Association, Mar. 2017, pp. 407–420. [Online]. Available: https://www.usenix.org/conference/nsdi17/technical-sessions/presentation/vanini

[15] A. Kabbani, B. Vamanan, J. Hasan, and F. Duchene, "Flowbender: Flow-level adaptive routing for improved latency and throughput in datacenter networks," in *Proceedings of the 10th ACM International on Conference on Emerging Networking Experiments and Technologies*, ser. CoNEXT '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 149–160. [Online]. Available: https://doi.org/10.1145/2674005.2674985

[16] S. Ghorbani, Z. Yang, P. B. Godfrey, Y. Ganjali, and A. Firoozshahian, "Drill: Micro load balancing for low-latency data center networks," in *Proceedings of the Conference of the ACM Special Interest Group on Data Communication*, ser. SIGCOMM '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 225–238. [Online]. Available: https://doi.org/10.1145/3098822.3098839

[17] C. Qi, W. Wu, Y. Wang, K. He, Y.-H. S. Kao, Z. He, C.-Y. Yen, Z. Jiang, F. Luo, S. Anubolu, Y. Gao, B. Lin, W. Ni, Y. Yang, D. Wei, B. Zhou, J. Wang, and S. Ding, "Sglb: Scalable and robust global load balancing in commodity ai clusters," in *Proceedings of the ACM SIGCOMM 2025 Conference*, ser. SIGCOMM '25. New York, NY, USA: Association for Computing Machinery, 2025, p. 626–644. [Online]. Available: https://doi.org/10.1145/3718958.3750527

[18] A. Dixit, P. Prakash, Y. C. Hu, and R. R. Kompella, "On the impact of packet spraying in data center networks," in *2013 Proceedings IEEE INFOCOM*, 2013, pp. 2130–2138. [Online]. Available: https://doi.org/10.1109/INFCOM.2013.6567015

[19] M. Alizadeh, T. Edsall, S. Dharmapurikar, R. Vaidyanathan, K. Chu, A. Fingerhut, V. T. Lam, F. Matus, R. Pan, N. Yadav, and G. Varghese, "Conga: distributed congestion-aware load balancing for datacenters," in *Proceedings of the 2014 ACM Conference on SIGCOMM*, ser. SIGCOMM '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 503–514. [Online]. Available: https://doi.org/10.1145/2619239.2626316

[20] C. H. Song, X. Z. Khooi, R. Joshi, I. Choi, J. Li, and M. C. Chan, "Network load balancing with in-network reordering support for rdma," in *Proceedings of the ACM SIGCOMM 2023 Conference*, ser. ACM SIGCOMM '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 816–831. [Online]. Available: https://doi.org/10.1145/3603269.3604849

[21] T. Bonato, D. D. Sensi, S. D. Girolamo, A. Bataineh, D. Hewson, D. Roweth, and T. Hoefler, "Flowcut switching: High-performance adaptive routing with in-order delivery guarantees," *IEEE Transactions on Networking*, vol. 34, pp. 1974–1987, 2026. [Online]. Available: https://doi.org/10.1109/TON.2025.3636209

[22] M. Al-Fares, S. Radhakrishnan, B. Raghavan, N. Huang, and A. Vahdat, "Hedera: Dynamic flow scheduling for data center networks," in *7th USENIX Symposium on Networked Systems Design and Implementation (NSDI 10)*. San Jose, CA: USENIX Association, Apr. 2010. [Online]. Available: https://www.usenix.org/conference/nsdi10-0/hedera-dynamic-flow-scheduling-data-center-networks

[23] Y. Xu, K. He, R. Wang, M. Yu, N. Duffield, H. Wassel, S. Zhang, L. Poutievski, J. Zhou, and A. Vahdat, "Hashing design in modern networks: Challenges and mitigation techniques," in *2022 USENIX Annual Technical Conference (USENIX ATC 22)*. Carlsbad, CA: USENIX Association, Jul. 2022, pp. 805–818. [Online]. Available: https://www.usenix.org/conference/atc22/presentation/xu

[24] Ultra Ethernet Consortium, "Ultra ethernet specification version 1.0," 2025, accessed on August 29, 2025. [Online]. Available: https://ultraethernet.org/wp-content/uploads/sites/20/2025/06/UE-Specification-6.11.25.pdf

[25] T. Hoefler, K. Schramm, E. Spada, K. Underwood, C. Alexander, B. Alverson, P. Bottorff, A. Caulfield, M. Handley, C. Huang, C. Raiciu, A. Kabbani, E. Opsasnick, R. Pan, A. Ran, and R. Sohan, "Ultra ethernet's design principles and architectural innovations," 2025. [Online]. Available: https://arxiv.org/abs/2508.08906

[26] A. Singhvi, N. Dukkipati, P. Chandra, H. M. G. Wassel, N. K. Sharma, A. Rebello, H. Schuh, P. Kumar, B. Montazeri, N. Bansod, S. Thomas, I. Cho, H. L. Seibert, B. Wu, R. Yang, Y. Li, K. Huang, Q. Yin, A. Agarwal, S. Vaduvatha, W. Wang, M. Moshref, T. Ji, D. Wetherall, and A. Vahdat, "Falcon: A reliable, low latency hardware transport," in *Proceedings of the ACM SIGCOMM 2025 Conference*, ser. SIGCOMM '25. New York, NY, USA: Association for Computing Machinery, 2025, p. 248–263. [Online]. Available: https://doi.org/10.1145/3718958.3754353

[27] L. Shalev, H. Ayoub, N. Bshara, and E. Sabbag, "A cloud-optimized transport protocol for elastic and scalable hpc," *IEEE Micro*, vol. 40, no. 6, pp. 67–73, 2020. [Online]. Available: https://doi.org/10.1109/MM.2020.3016891

[28] NVIDIA Corporation, *NVIDIA ConnectX-8 SuperNIC User Manual*, 2025, includes official specifications. [Online]. Available: https://docs.nvidia.com/networking/display/nvidia-connectx-8-supernic-user-manual.pdf

[29] ——, "Nvidia bluefield-3 dpu," Apr. 2021, datasheet. [Online]. Available: https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/documents/datasheet-nvidia-bluefield-3-dpu.pdf

[30] ——. (2025) Spectrum-x — ethernet networking platform for ai. Product page. [Online]. Available: https://www.nvidia.com/en-us/networking/spectrumx/

[31] M. Handley, C. Raiciu, A. Agache, A. Voinescu, A. W. Moore, G. Antichi, and M. Wójcik, "Re-architecting datacenter networks and stacks for low latency and high performance," in *Proceedings of the Conference of the ACM Special Interest Group on Data Communication*, ser. SIGCOMM '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 29–42. [Online]. Available: https://doi.org/10.1145/3098822.3098825

[32] T. Bonato, S. Abdous, A. Kabbani, A. Ghalayini, N. Gebara, T. Lam, A. Agarwal, T. Chen, Z. Yu, K. Taranov, M. Elhaddad, D. De Sensi, S. Ghorbani, and T. Hoefler, "Uno: A one-stop solution for inter- and intra-data center congestion control and reliable connectivity," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, ser. SC '25. New York, NY, USA: Association for Computing Machinery, 2025, p. 1195–1210. [Online]. Available: https://doi.org/10.1145/3712285.3759884

[33] P. Cheng, F. Ren, R. Shu, and C. Lin, "Catch the whole lot in an action: Rapid precise packet loss notification in data center," in *11th USENIX Symposium on Networked Systems Design and Implementation (NSDI 14)*. Seattle, WA: USENIX Association, Apr. 2014, pp. 17–28. [Online]. Available: https://www.usenix.org/conference/nsdi14/technical-sessions/presentation/cheng

[34] W. Li, X. Liu, Y. Zhang, Z. Wang, W. Gu, T. Qian, G. Zeng, S. Ren, X. Huang, Z. Ren, B. Liu, J. Zhang, K. Chen, and B. Liu, "Revisiting rdma reliability for lossy fabrics," in *Proceedings of the ACM SIGCOMM 2025 Conference*, ser. SIGCOMM '25. New York, NY, USA: Association for Computing Machinery, 2025, p. 85–98. [Online]. Available: https://doi.org/10.1145/3718958.3750480

[35] Y. Lu, G. Chen, B. Li, K. Tan, Y. Xiong, P. Cheng, J. Zhang, E. Chen, and T. Moscibroda, "Multi-path transport for rdma in datacenters," in *Proceedings of the 15th USENIX Conference on Networked Systems Design and Implementation*, ser. NSDI'18. USA: USENIX Association, 2018, p. 357–371. [Online]. Available: https://www.usenix.org/conference/nsdi18/presentation/lu

[36] T. Bonato, A. Kubicek, A. Kabbani, A. Ghalayini, M. Besta, and T. Hoefler, "Spritz: Path-aware load balancing in low-diameter networks," in *2026 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. IEEE, May 2026, pp. 527–543. [Online]. Available: https://doi.org/10.1109/IPDPS65963.2026.00052

[37] Y. Geng, V. Jeyakumar, A. Kabbani, and M. Alizadeh, "Juggler: a practical reordering resilient network stack for datacenters," in *Proceedings of the Eleventh European Conference on Computer Systems*, ser. EuroSys '16. New York, NY, USA: Association for Computing Machinery, 2016. [Online]. Available: https://doi.org/10.1145/2901318.2901334

[38] S. Shen, T. Bonato, Z. Hu, P. Jordan, T. Chen, and T. Hoefler, "ATLAHS: An application-centric network simulator toolchain for AI, HPC, and distributed storage," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, ser. SC '25. New York, NY, USA: Association for Computing Machinery, 2025, pp. 349–367. [Online]. Available: https://doi.org/10.1145/3712285.3759838

[39] S. Li, Y. Zhao, R. Varma, O. Salpekar, P. Noordhuis, T. Li, A. Paszke, J. Smith, B. Vaughan, P. Damania, and S. Chintala, "PyTorch distributed: Experiences on accelerating data parallel training," *Proceedings of the VLDB Endowment*, vol. 13, no. 12, pp. 3005–3018, 2020. [Online]. Available: https://doi.org/10.14778/3415478.3415530

[40] M. Naumov, D. Mudigere, H.-J. M. Shi, J. Huang, N. Sundaraman, J. Park, X. Wang, U. Gupta, C.-J. Wu, A. G. Azzolini *et al.*, "Deep learning recommendation model for personalization and recommendation systems," *arXiv preprint arXiv:1906.00091*, 2019.

[41] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," *arXiv:2302.13971*, 2023.

[42] MLCommons, "Mlcommons training benchmark," https://mlcommons.org/benchmarks/training/, 2024, accessed: 2025-09-17.

[43] T. V. Pham, C. Steger, B. Rockel, K. Keuler, I. Kirchner, M. Mertens, D. Rieger, G. Zängl, and B. Früh, "Icon in climate limited-area mode (icon release version 2.6.1): a new regional climate model," *Geoscientific Model Development*, vol. 14, no. 2, pp. 985–1005, 2021. [Online]. Available: https://doi.org/10.5194/gmd-14-985-2021

[44] A. P. Thompson, H. M. Aktulga, R. Berger, D. S. Bolintineanu, W. M. Brown, P. S. Crozier, P. J. In't Veld, A. Kohlmeyer, S. G. Moore, T. D. Nguyen *et al.*, "Lammps-a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales," *Computer physics communications*, vol. 271, p. 108171, 2022. [Online]. Available: https://doi.org/10.1016/j.cpc.2021.108171

[45] U. of Utah, "The mimd lattice computation (milc) collaboration," https://web.physics.utah.edu/~detar/milc/.

[46] C. Hopps, "Rfc2992: Analysis of an equal-cost multi-path algorithm," RFC Editor, USA, Tech. Rep., 2000. [Online]. Available: https://datatracker.ietf.org/doc/html/rfc2992

[47] Y. Cheng, N. Cardwell, N. Dukkipati, and P. Jha, "The rack-tlp loss detection algorithm for tcp," RFC 8985, Feb. 2021. [Online]. Available: https://www.rfc-editor.org/info/rfc8985

[48] N. Dukkipati, N. Cardwell, Y. Cheng, and M. Mathis, "Tail loss probe (tlp): An algorithm for fast recovery of tail losses," IETF, Internet-Draft draft-dukkipati-tcpm-tcp-loss-probe-01, Feb. 2013, work in Progress. [Online]. Available: https://datatracker.ietf.org/doc/draft-dukkipati-tcpm-tcp-loss-probe/

[49] Scalable Parallel Computing Lab (SPCL), ETH Zurich, "Spcl application traces repository," https://storage2.spcl.ethz.ch/traces/, 2025, directory of publicly hosted application traces (e.g., LLaMA, DLRM).

[50] J. Cao, R. Xia, P. Yang, C. Guo, G. Lu, L. Yuan, Y. Zheng, H. Wu, Y. Xiong, and D. Maltz, "Per-packet load-balanced, low-latency routing for clos-based data center networks," in *Proceedings of the Ninth ACM Conference on Emerging Networking Experiments and Technologies*, ser. CoNEXT '13. New York, NY, USA: Association for Computing Machinery, 2013, p. 49–60. [Online]. Available: https://doi.org/10.1145/2535372.2535375

[51] R. Mittal, A. Shpiner, A. Panda, E. Zahavi, A. Krishnamurthy, S. Ratnasamy, and S. Shenker, "Revisiting network support for rdma," in *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*, ser. SIGCOMM '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 313–326. [Online]. Available: https://doi.org/10.1145/3230543.3230557

[52] R. Sohan, E. Spada, E. Davis, M. Handley, I. Burstein, T. Hurson, J. Jose, V. Kashyap, R. Pan, S. Sur, S. Anantharamu, A. Barnea, A. Caulfield, E. Cohen, E. Edmunds, Y. Friedman, M. Ghazi, M. Guramali, T. Hoefler, V. Jain, A. Kabbani, N. Katz, Y. Le, C. Mbariky, G. Morandin, M. Moshref, S. O'Neil, M. Papamichael, J. Pfefferle, S. S. Pyla, C. Raiciu, D. Riddoch, K. Schramm, Y. Shpigelman, S. Shuler, S. Shyman, R. Sivaramu, A. Tootoonchian, and Y. Wang, "The multipath reliable connection (MRC) transport," Jun. 2026. [Online]. Available: https://arxiv.org/abs/2606.18170

[53] Y. Le, R. Pan, P. Newman, J. Blendin, A. Kabbani, V. Jain, R. Sivaramu, and F. Matus, "Strack: A reliable multipath transport for ai/ml clusters," 2024. [Online]. Available: https://arxiv.org/abs/2407.15266