

Load Balancing for AI Training Workloads

Sarah McClure, Sylvia Ratnasamy, Scott Shenker
UC Berkeley

Abstract

We investigate the performance of various load balancing algorithms for large-scale AI training workloads that are running on dedicated infrastructure. The performance of load balancing depends on both the congestion control and loss recovery algorithms, so our evaluation also sheds light on the appropriate choices for those designs as well.

1 Introduction

The near-universal adoption of TCP has long imposed an implicit requirement on networks: *the stream of packets in a single TCP connection must be routed along a single network path*. This requirement arises because splitting a connection across multiple paths could lead to a significant reordering of packets; in turn, a TCP sender might mistake the reordering for loss (because of the resulting duplicate ACKs) and then unnecessarily decrease the connection’s congestion window; finally, cutting the congestion window may do little to decrease the reordering, so the sender may repeatedly decrease the window resulting in a severe impact on the connection’s throughput [20, 57]. Thus, in enterprise and WAN networks, sending a connection’s traffic along multiple paths was almost always avoided.

Datacenter networks are different, in that they typically have many equal-cost paths. Sending packets along multiple equal-cost paths would not induce the level of reordering that occurs in more heterogeneous networks, so one might have expected such networks to depart from the dominant single-path paradigm. However, most datacenter networks have continued to follow this paradigm in spirit. ECMP [26] routes on five-tuples, which keeps a connection’s traffic on a single path, and more recent extensions like PLB [43] and flowlet switching [30] only change a connection’s route during idle periods so as to avoid reordering. More specifically, during nonidle periods flows are still constrained to a single path, but those paths can be changed (*e.g.*, to avoid congested links) during idle periods, thereby giving the load-balancing scheme a bit more flexibility without incurring reordering.

This persistence of the single-path paradigm (with minor extensions) exists despite the fact that the single-path paradigm has serious drawbacks. In particular, requiring a TCP-connection (or a flowlet) to stay on a single path: (i) limits the capacity available to a connection, even if significant capacity exists on alternate paths [27, 29, 44], (ii) can create hotspots that impact tail latency if several long-lived flows are sent along the same link [5, 43], and (iii) can lead to highly sub-

optimal load balancing when flow sizes vary greatly (because some links are more heavily loaded than others) [5, 22].

Not only are there serious drawbacks, but the community has developed viable technical alternatives to the single-path paradigm. As far back as the 1990s, researchers were already exploring modifications to TCP that successfully exploited multiple paths [27, 44] and robustly tolerated packet reordering [57]. More recent research has explored transport designs for the datacenter that uses packet spraying [6] (in which a host or switch sends packets across all equal-cost paths; this is the antithesis to keeping connections on a single path) and packet trimming for loss detection [24] (in which a switch, when a packet arrives that would otherwise be dropped, strips the data from the packet and forwards the remaining header at high-priority).

The reason that these alternative designs were not widely adopted is presumably because the traditional solution of ECMP, particularly when combined with approaches like PLB, provided “good enough” performance for the then-current workloads. Thus, there was little need to depart from the single-path paradigm, even if this paradigm was not optimal. However, the advent of AI training workloads has greatly changed the performance requirements on AI-oriented datacenter networks, which in turn has opened the door to innovations that violate the single-path paradigm. In what follows, when we refer to AI workloads we mean large-scale AI training workloads that are run on their own dedicated infrastructure.

While we discuss AI workloads at greater length in §2, here we merely note that these workloads have two characteristics that accentuate the limitations of the single-path paradigm. First, the vast bulk of the AI traffic is in long-lived flows that do not have idle periods, so techniques like PLB do not ameliorate the rigidity of the single-path paradigm (*i.e.*, as mentioned above, if two long-lived flows are hashed by ECMP to the same link, PLB cannot reroute them because there are no idle periods, so tail-latencies can be strongly affected). Second, the performance requirements of AI workloads are extreme, requiring the full completion of all flows in a communication collective [3, 12, 21] (which we define in §2). Thus, the techniques that mitigated some of ECMP’s problems (*e.g.*, PLB) are no longer *effective*, and the performance degradation caused by strictly adhering to limiting flows (or flowlets) to a single path is no longer *acceptable*.

This need for better performance for AI workloads has led to active exploration, mostly in industry, of designs that go

beyond the single-path paradigm. Here we mention two efforts in particular. The Ultra Ethernet Consortium (UEC) [51] is considering new transport designs that include packet spraying and packet trimming [2, 3]. Nvidia has developed SpectrumX which uses a switch-based packet-spraying approach. These efforts suggest that we are finally seeing a broad appetite in industry to consider load balancing techniques that no longer restrict packet flows to a single path.

When designing load-balancing schemes for AI workloads, the goal is to minimize collective completion time (CCT) for AI traffic. Note that this performance goal is quite different from the goals for traditional datacenter workloads where the focus is on: (i) per-flow measures such as flow-completion time and per-flow fairness (as compared to AI’s focus on collective completion time) and (ii) maximizing utilization so that the network is being used efficiently (utilization is less relevant for AI workloads because the network is a small component of the overall cost of an AI cluster, so the network need not be cheap and fully utilized, it only needs to complete collectives quickly).

While progress has been made in developing specific solutions to meet this goal, such as the two previously mentioned efforts (UEC and SpectrumX), we were surprised to find that many basic design questions remain open.

For instance, while it seems clear that packet spraying is a robust technique for spreading load along all suitable paths, we were not able, in our literature search, to find answers to such key questions as: (i) How much better is packet spraying versus the state-of-the-art ECMP+PLB designs? (ii) Do the benefits of (various forms of) packet spraying remain when we introduce failures or the workload creates incast? (iii) Is it better to implement packet spraying in switches (as in SpectrumX and pFabric [6]) or in hosts (as in NDP)? (iv) Do we need packet trimming for loss detection or are more incremental changes to TCP-like protocols adequate? (v) And what forms of congestion control are appropriate in the context of spraying-like load balancing?

The ongoing industry efforts suggest there is a lack of consensus on these questions. For instance, Nvidia supports switch-based packet spraying, UEC uses host-based spraying, and Google’s Falcon design [35] uses host-implemented packet pacing but no packet-spraying.

Given the urgency to meet the performance of AI workloads, one might have expected that these questions would have been resolved quickly. However, answering such questions is challenging because it requires decoupling load balancing from the other aspects of a transport protocol (such as congestion control and loss recovery). Given the competitive and evolving nature of the AI market, we cannot expect complete agreement on all aspects of transport. However, it would be extremely valuable for the ecosystem if switch vendors knew what load-balancing primitives they should include in their switch designs. Providing the technical results for making such a decision is the focus of this paper. We should note that performance is not the only objective; some design choices involve changes in switch designs (and thus are harder to deploy immediately)

and others involve additional computation (for encoding) whose overhead must be weighed against the performance it enables. Thus, our goal is to provide the understanding of the performance tradeoffs, but we understand that other factors will play a role in what designs are eventually deployed.

We start the paper with, in §2, some background on the design space and workloads we will be considering. We then, in §3, discuss the various other related components – network design, loss recovery, congestion control – and how we deal with them in our work, and then quickly state the set of questions we address and describe related work. We describe our methodology in §4. We then present our simulation results for (i) load balancing in §5 and (ii) loss recovery in §6, before synthesizing our findings in §7.

2 An Overview of the Load Balancing Design Space and AI Workloads

This paper focuses on only a narrow slice of network functionality – load balancing – but there are many options to consider. So we start this section by reviewing the overall design space for load balancing, and then end with a description of AI workloads whose performance requirements we are trying to meet.

2.1 Load Balancing Design Space

The design space for load balancing has been outlined (at least in part) in several papers [5, 31, 52], so here we provide a fairly brief summary. At a high level, the main dimensions of the design space are:

- **Granularity:** Load balancing inevitably involves spreading load over multiple paths. Approaches differ greatly in terms of the granularity of traffic that is placed on different paths. The granularity varies from the smallest unit – *i.e.*, a packet [11, 24, 40] – to small sets of contiguous packets – *i.e.*, a flowlet [30, 43] – to ongoing subflows within a single connection – *i.e.*, parallel streams whose union reconstitutes the original connection [44] – to entire connections or flows themselves [26].
- **Location:** Local load balancing decisions – *i.e.*, over which equal-cost paths traffic is spread over – can be made at the host [8, 43] and at switches [6, 16], and different schemes choose to split these decisions differently (*e.g.*, ECMP is completely switch driven, packet spraying can be host or switch driven, *etc.*). All else being equal, having decisions be made in hosts makes it easier to deploy and evolve a load balancing solution.
- **State Used:** Load balancing must decide where to send traffic, and designs vary on what state is used to make these decisions. All designs use state about network connectivity to determine the current set of equal cost paths, but the differences arise in what state about the load on the network do they use. These choices include using no state about load (*e.g.*, [20, 24, 26]) or local state about the load (such as at the particular host or switch) (*e.g.*, [31, 43, 45]), or using global state about the load (*e.g.*, [5, 6]).

This is clearly a large design space, with many subtle design choices remaining within each particular design choice (*e.g.*, exactly what algorithm should be used to dynamically adjust the load balancing in response to measured load). To do a manageable search, whenever available for a given set of design choices we use a design used in practice today [16, 21, 43] since it has obviously undergone testing (and our focus here is on selecting among the overall set of choices, not micro-optimizing a given choice). And as we describe in our evaluation (in §5, §6, and §7) we remove designs from further consideration once their performance has been found lacking. The options in each dimension are shown in Table 1 with our explored options marked.

2.2 Characteristics of AI Workloads

We start with a quick summary. AI workloads have very specific traffic patterns that are quite different from traditional datacenter workloads, are highly performance-sensitive with metrics that are different from traditional datacenter workloads, and are commonly deployed as the sole workload running in the cluster [21, 34, 42]. These characteristics, taken together, provide strong motivation for designing a load balancing scheme specifically tailored to AI workloads.

More specifically, AI workloads involve rounds of computation interspersed with rounds of communication, where the communication is one of a few basic collectives, such as AllReduce, AllGather, and AlltoAll (see [17, 36] for a review). Thus, the communication workloads can be well-characterized by the traffic generated by these collectives, and the network-related performance metric is the collective completion time, since the next round of computation cannot start until the previous round of communication has completely finished. This results in workloads that are:

- **Uniform:** The nodes participating in a collective are typically using the same communication library (*e.g.*, [18]) and the same lower level networking support (*i.e.*, transport protocol and NICs). The collectives are generally symmetrical, in that each participating node has the amount of data to send, and generates traffic with the same traffic characteristics, with all packets being of the same size.
- **Synchronous:** All nodes participating in the collective begin at roughly the same time. This is particularly relevant to load balancing, because it synchronizes the traffic [3] which is both good for load balancing (no need to adapt to newly arriving flows while a collective is in progress) and bad (creating possible incast congestion).

In our evaluation of load balancing schemes, we consider an AlltoAll collective by default. We also consider permutation traffic matrices – where each node is sending traffic to exactly one other node, and receiving traffic from exactly one other node – with potentially many such traffic patterns simultaneously coexisting. While not mapped to a specific collective, permutation matrices are often used to implement collectives such as AllGather and AllReduce [12, 21, 49, 54]

and can be used to iteratively perform an AlltoAll ($n - 1$ permutation matrices iteratively).

AI workloads above a certain scale are often deployed in isolation. This is the case we assume in our paper, because it is in these large-scale isolated deployments that achieving high performance (*i.e.*, low CCTs) is so critical.

3 Approach

3.1 Other Components

In exploring the various options for load balancing, the main conceptual challenge is to isolate the merits of load balancing from the effects of other components of the network, all in the context of AI workloads. The relevant components are: the physical network (topology and provisioning), congestion control, and loss recovery (which includes both loss detection and packet retransmission). We now clarify our assumptions about the role of these components.

Physical network: Since the physical network is not the dominant factor in the cost of an AI cluster, and the AI workloads can be well understood in advance, we assume the network to be a fat-tree topology with full bisection bandwidth. For an all-to-all or permutation traffic pattern on such networks, all flows are only bottlenecked on the sending or receiving host’s link to its ToR. The rest of the network, which we will call the core, is sufficiently provisioned to handle the load emerging from the ToRs if (and only if) load balancing is working well.

Congestion control with ideal load balancing: Given a full-bisection-bandwidth and failure-free network and a static workload (defined by a given collective), and perfect load-balancing so there is no overloaded link in the core, there should be no need for congestion control. By this we mean that, in this ideal case, it is possible to avoid congestion entirely for given a particular workload by computing a schedule for which packets should be sent when and by which route. If every host sends according to the packet schedule, no packets will be dropped, and there would be no need for adaptive congestion control such as currently implemented in TCP.

Load balancing in the real world: However, reality is far from the ideal case we just described. First, current load balancing schemes do not schedule at such a fine granularity. Typically the application decides when the next collective should start, and then the load balancing scheme merely tries to spread the load over the network without any detailed knowledge of the traffic matrix. Moreover, rather than computing some global schedule for which routes are used for which flows or packets, load balancing schemes typically involve hosts and/or switches independently selecting (*e.g.*, randomly or round-robin or shortest queue) their choice of next-hops, with no coordination between different hosts and switches, and with no global knowledge of the overall load pattern. These independent actions can result in temporary overloads (whereas a perfect packet schedule would have avoided such overloads). We focus exclusively on this class of independent and coarse-

	Location	Host				Switch				
		Granularity	Packet	Flowlet	Subflow	Flow	Packet	Flowlet	Subflow	Flow
State	None	✓ [24, 40]	✗ [25, 52]	✗ [44]	✗	✗	✓ [6, 11, 20]	✗	✓	✓ [26]
	Local	✗ [8]	✓ [28, 31, 43, 56]	✗	✗	✗	✓ [16, 55]	✓ [30, 48]	✓ [45]	✓ [10]
	Global	✗ [41]	✗	✗	✗ [3, 19]	✗	✗ [5, 32, 53]	✗	✗ [4, 7]	✗

Table 1: A survey of existing load balancing approaches based on their considered state, granularity, and implementation location. We consider the design points marked with ✓. We note that some works allow splitting flows in time/packets in different amounts and with different strategies, thus we use a wide definition of “flowlet” that includes any splitting in these dimensions.

grained load balancing; we assume hosts and switches have no knowledge of the traffic matrix and engage in no global coordination, and only rely on local (but perhaps load-sensitive) actions.

In addition, links can fail, and in this case the goal of load balancing is to effectively use the remaining bandwidth to minimize the CCT. Moreover, we want load balancing to achieve this goal across a wide range of failure scenarios (*e.g.*, multiple failed links, partial failures of bonded links, *etc.*).

Loss recovery in the real world: Packet loss is inevitable because failures do happen and (as we argue below) load balancing is not perfect. The job of loss recovery (as opposed to congestion control) is to detect which packets have been lost and schedule their retransmissions. We examine a wide variety of loss recovery approaches, from what TCP uses (*e.g.*, dupACKs and timeouts) to rateless erasure coding [37, 38, 47] (which ignores losses and just transmits packets until the file is complete). The ideal form of loss recovery would be an erasure encoding with zero overhead; *i.e.*, as soon as you have received as much data as in the original file, the connection is complete.

Congestion control in the real world: We now arrive at the crux of why the setting we are considering is so different from traditional datacenters. For AI workloads, the goal is to minimize the CCT; that means we can ignore the more traditional goals – such as fully utilizing links or avoiding losses or controlling per-packet latencies – as long as the CCTs are minimized. The traditional goal of adaptive congestion control is to match a flow’s rate to the available bandwidth along its path, and adaption is needed to deal with unpredictable changes in the available bandwidth (due to changing loads or paths). This mission statement makes sense within the single-path paradigm (MPTCP [44] extends this paradigm to multiple paths, each with its own subflow) and when backing off is needed to fairly share the network with other flows.

We are considering AI workloads that are deployed in isolation; this means that we need not be careful about sharing the network with other workloads. And when packets are sprayed randomly by switches, hosts do not know what paths their packets are taking, so they cannot reason about the available bandwidth along a given path. Moreover, the goal of congestion control in this setting is not to match its rate to avoid loss but to minimize CCT. And when flows are spread across many paths, and this load balancing is done independently (so there can be temporary overloads) and/or some paths have failures and others don’t (and the paths may not have the same amount of available bandwidth for each flow), *there is no well-defined rate that both*

minimizes CCT and avoids packet loss. Thus, the traditional mission statement of congestion control does not apply here.

To make this point clearer, consider a single host sending data at rate $r > 0$ over a single bottleneck link with bandwidth $b > 0$. The throughput t is given simply by $t = \min[r, b]$, so we typically design congestion control algorithms to find the bottleneck rate b (as is most explicitly done in BBR [13]). The goodput g can never exceed the throughput t , but if the loss recovery algorithm deals badly with loss, then g as a function of r for $r > b$ can show a decrease. Note that to complete a file transfer of length L , it would take a time $\frac{L}{g}$. Thus, in this very traditional case there is a rate $r = b$ at which flows can send that avoids loss and minimizes flow-completion time.

With AI workloads, CCT is the metric of interest, and we see a very different result. While we delay the description of the simulation setup until §4, here we just describe a single result. We consider a set of nodes in a collective sending at rate r (with no adaptive congestion control, merely sending at a constant rate below their link rate b) using ideal loss recovery (*i.e.*, erasure encoding with no overhead), and plot the resulting CCT versus r/b for two load-balancing schemes: ECMP and host-based packet spraying. Both the spraying and ECMP are random (without any coordination), so loss can occur where random spraying overloads a link temporarily or causes incast. In Figure 1, we see that for packet spraying the CCT continues to improve up until the hosts are sending at rate b even though they are sustaining loss (the loss is shown in Figure 2). This graph also contains an “ideal” curve that describes the CCT that would be achieved if there were no loss. While hard to read on the graph, the distance between the ideal and host-based packet spraying is about 20%-30%. This represents the inability of such independent and randomized load-balancing decisions to achieve perfect load balancing.

ECMP shows even worse performance, with the CCT flattening out (within error) at a CCT value significantly above that of packet spraying. The packet loss rate is also significantly higher than with host-based packet spraying. This shows the unsurprising result that rigidly staying on a given path provides worse load balancing.

To compare this with a standard state-of-the-art congestion control algorithm, we simulated some scenarios with the Swift [33]. In cases where reordering triggered many spurious retransmissions the CCT with Swift was much greater than what was achieved by sending at a fixed sending rate at b and ideal loss recovery. More importantly, in cases where there were no spurious retransmissions, the CCT achieved by Swift

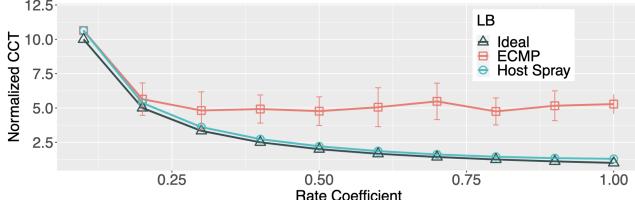


Figure 1: ECMP and host-based spraying collective completion time as flow rates increase to their full fair share. The x-axis is the ratio of the host’s sending rate to host-ToR link rate. As a comparison, we also plot the ideal case where the CCT is merely the total message size divided by the sending rate.

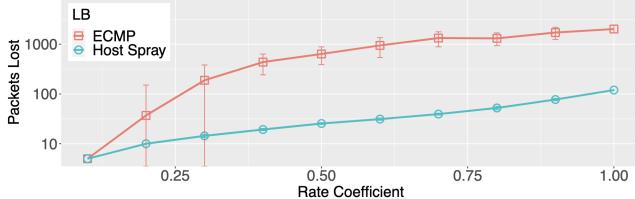


Figure 2: Packets lost in worst-hit flow for scenarios in Figure 1, where each flow has 500 packets. For each simulation run, we recorded which flow suffered the most losses (because the flow that takes the longest determines the CCT), and then averaged over simulation runs.

was around 1.75x what we observed for the fixed sending rate. This is not a comment about Swift’s particular congestion control; instead, it is merely confirming the observation that if you are designing a congestion control algorithm to avoid loss, and you are using less-than-ideal load balancing (which can cause loss), then you cannot achieve optimal CCTs.

One might question whether these results would still hold if the buffers were increased to greatly reduce loss. Recall that Swift, like most modern congestion control algorithms, tries to avoid loss *and* minimize per-packet latency by keeping buffer occupancy low (*i.e.*, to avoid buffer-bloat). So even on runs with larger buffers, using Swift is significantly worse ($\sim 23\%$) than sending at full rate.

3.2 Causes of Load Balancing Imperfection

Before continuing, we discuss why loss is inevitable in the setting we consider. As mentioned in §2, we will consider many different granularities of load balancing; here we will distinguish between those that load balance flows in units of rate (flow, subflow, *etc.*) and those that load balance in units of packets (flowlet, packet spraying, *etc.*). In principle, when load balancing in units of rate, an ideal load balancer only needs $O(N)$ subflows to perfectly use N potential paths (the reason it is not just N is to allow some splitting of flows to make load balancing perfectly balanced). In a setting where the load balancing decisions are made independently (and often randomly), perfect placement of (sub)flows is unlikely, resulting in collisions and performance degradation.

When load balancing in units of packets instead of rate, the main cause of loss is when too many packets arrive at a link at the same time and overflow the buffer. If switches have radix k and buffers of size q (measured in packets), then k packets

can arrive simultaneously. If all of these packets are destined to the same output port, then $(k - 1) - q$ packets can be lost. While flows sharing a destination link are unlikely to be fully synchronized at a per-packet level, some amount of overlap is expected and so loss will happen in practice. While this pattern is well-known for ToR downlinks (incast), traffic matrices like all-to-all make this possible throughout the network. And our simulations show that such loss is common (*e.g.*, 20% loss on the worst-hit flow with host-based packet spraying).

3.3 Our Approach

From these initial results, we came to three conclusions that drove our research agenda:

- In the setting we consider, with large-scale AI workloads running on their own infrastructure, we cannot use traditional notions of congestion control, because incurring loss in this setting may be required to minimize CCT.
- Even with ideal loss recovery, there is a difference in effectiveness in different load balancing schemes. Understanding this more thoroughly is one of the main goals of our study.
- The presence of loss in steady state means that effective loss recovery is crucial, so we must also explore various loss recovery options.

Given these conclusions, our approach is to adopt a congestion control philosophy that ignores loss and only focuses on minimizing CCT. In most cases, we just set the sending rate equal to the bandwidth of the host-ToR link. But in cases, such as we find later in our simulation results (in §5, §6, and §7), where a different value of r achieves the minimal CCT, we use that value.

We use this approach to ask two basic questions.

- *What form of load balancing works best, assuming perfect loss-recovery?* This means investigating the effectiveness of schemes that make different choices for granularity, location, and state (load-dependent or not). We first ask this question in the context of a network with no failures and a given buffer size, and then ask if our results change when we consider failures and when we greatly increase the buffer size. We find that packet spraying works best (whether at host or switch), and this advantage becomes even more extreme for networks with failures. With larger buffers, packet spraying further widens its performance gap against other schemes.
- *What form of loss recovery works best with spraying?* Our main finding is that coding, even with moderate overheads, is the best overall. We also propose a modification of TCP’s loss recovery mechanism that makes it more robust against reordering; while its performance is worse than coding in most cases, it is comparable or better than other alternatives we consider.

Because the design space is vast, and the interactions between component are delicate, we do not claim that our findings are definitive or exhaustive, merely that they are evidence pointing towards alternatives that deserve more exploration.

3.4 Related Work

There is a vast literature on load balancing in datacenters, and we cannot hope to summarize it in any depth. The citations in Table 1 provide a subset that covers the space along the dimensions we described in §2. Many of the individual papers evaluate specific designs (that make choices for each of the components), but we are not aware of work that tries to systematically evaluate each component in designing a solution. This is especially true for the unique setting of large AI training workloads. In addition, many of these proposed systems [21, 30, 43, 52] do not seriously consider packet spraying given its potential adverse effects on the transport due to reordering. Some systems do include spraying in their design [6, 11, 24, 40], but often introduce specialized mechanisms to make spraying work (*e.g.*, proposing an entirely new transport). Two exceptions are [20] and [8], where packet spraying is evaluated against ECMP but other load balancing granularities are not considered and one transport is assumed. There are two papers which make more general comments: [21] mentions (as we do) the possibility of not using loss-avoiding congestion control, and [3] questions the need for packet spraying.

4 Methodology

Simulator: We compare different load balancing schemes with simulations on Broadcom’s htsim [1], which we chose due to its ability to scale to large datacenter networks and its existing implementation of many design points in this space. Specifically, unless stated otherwise, we simulate a 128 node 3-tier fat-tree network with the simulator default 100 Gbps links, 1 us link latency, and 32 KB per-link buffers¹ ([1, 11]). We vary both the link rate and buffer size to determine if the results change as these values scale to higher values. We run all considered scenarios 10 times to determine the level of variability in the measured performance.

Traffic workload: We consider a small set of traffic matrices reflective of collectives performed in large AI model training [17]. Our default, an AlltoAll collective, has each host sending a message to all other nodes. Similarly, this collective can be performed as iterative permutation matrices which we consider as well. Other collectives, such as an AllGather can also be implemented as a series of (identical) permutation matrices (*e.g.*, ring algorithms) or naively sending a message to all receivers (the same traffic matrix as an all-to-all). We do not consider collectives that require compute between flows (*e.g.*, AllReduce), as the specifics of the message identities then becomes important (*i.e.*, a single, message-agnostic traffic matrix is insufficient to describe the class of workload). Lastly, in all workloads, we randomize the start time of flows by up to

¹While this is smaller than state-of-the-art per-link buffer size [10, 16], given that incast is a primary concern, we set the per-link buffer size consistent with the ability to absorb worst-case incast ($\sim 2 - 40$ KB per switching capacity per port). More intuitively, applying the per-port per-Gbps buffer size of a switch would allow an unrealistic ability to absorb incast for the few-port switches we consider.

the calculated time between packet sends (inverse pacing rate).

Load Balancing (LB) Schemes: As mentioned in §2, we consider all the load balancing options noted in Table 1. We provide additional detail about each option here:

- **Per-Packet LB:** We consider two main forms of packet spraying: in-switch round-robin placement of packets and in-host incrementing of the flow label (which is hashed in switches implementing ECMP). In the latter, this (most likely) causes a packet to be hashed onto a different path than the previous. We also consider an adaptive version of in-switch spraying which places the packet on the option with the shortest queue.
- **Per-Flow LB:** Our baseline for per-flow load balancing is ECMP: packets are hashed based on their five-tuple + flow label and placed on one path in the network. We consider an in-network load-aware option that upon initial hashing, checks the queue length of the initially chosen link and, if the length is sufficiently large, moves the flow to a link with one of the shortest queues. When considering subflows, we assume that each subflow is assigned a different flow label to ensure they are hashed differently.
- **Per-Flowlet LB:** Flowlet routing is inherently dynamic, as it changes the current path in gaps between sending packets. We use two options in this category. First, we consider an in-switch adaptive version which maintains state about the spacing between packets for all flows active in the previous x us and changes the link for the flow when there is a sufficient gap between packets ($> x$). Second, we consider a model that does not wait for gaps, but instead can change the flow label after a certain minimum number of packets, and does so depending on whether the current path is sufficiently “good.” Our implementation makes this goodness decision based on recent ECN markings as in [43]. Specifically, our configuration allows the flow to repath at most every ~ 10 packets, with an ECN marking threshold of 60% buffer size and a threshold for deeming a path “bad” of 40% of the recent ECN marks. These parameters were chosen after evaluating a range of different values.

Loss Recovery Schemes: When flows are paced at the correct rate, imperfect load balancing may still cause buffer overload despite the available capacity in the network. Thus, we consider several loss recovery mechanisms to determine which work effectively under the best load balancing. Further, some methods of load balancing such as packet spraying may have additional adverse effects on a loss recovery mechanism that depends on in-order packet delivery (*e.g.*, duplicate ACKs).

Accordingly, we consider several different methods for detecting losses and scheduling retransmissions:

- **TCP (dupACKs and RTOs):** As a well-known baseline, we use an off-the-shelf loss recovery protocol, TCP. Here, retransmissions are triggered by duplicate ACKs and timeouts, but as noted previously in our simulations these events do not change the pacing rate but only determine

which packet must be (re)transmitted next.

- **ROCE (PFC and sequence numbers):** We also consider a ROCE-like protocol which uses PFC² to prevent congestive loss. However, given that there may be non-congestive loss, hosts implement a basic reliability system which assumes in-order packet delivery [23, 39, 48]. If a packet is missing, the receiver sends a NACK and the sender retransmits either at the beginning of the message or at the given packet number. For non-congestive loss (or loss due to incorrectly set PFC thresholds), this is similar to dupACK mechanisms in TCP with a threshold of only 1.
- **Packet Trimming:** Packet trimming has been used in many works [9, 24, 40] to explicitly notify a host of congestive loss. As we briefly mentioned in §1, when there is not enough buffer space for a packet, the switch keeps the header (removes the payload) and queues it for transmission at high priority. Depending on the configuration, the trimmed packet may be sent to either the receiver (to be reflected back to the sender) or the original sender (return-to-sender, “RTS”). Thus, no metadata should be lost in the network unless there is so much congestion that headers must be dropped. We assume all switches in the network are enabled to perform trimming. Again, once loss is detected, this merely determines the next packet to be sent and does not change the rate.
- **Erasure Coding:** Lastly, we consider a loss recovery mechanism based on rateless erasure coding [37, 38, 47]. In this case, the sender encodes the data before sending and creates some number of recovery symbols which can be used to recover lost data. The receiver can decode the entire messages with high probability as long as a sufficient number of symbols have been received. Therefore, such a protocol comes with a constant overhead (the required recovery symbols), but does not need to manage specific packets/sequence numbers (all symbols are equally necessary and can be lost).

Failure models: We use the failure model provided in the simulator and consider failures as partial bandwidth loss of the selected links between pods and the core routers. This models the effect of failures within a link aggregation group (LAG).

For a failure scenario, we set the number of links per pod that are experiencing failures and the percentage of bandwidth lost between the two nodes. This bandwidth degradation exists for the entire duration of the simulation.

Similarly, we also consider a model of failure with *flaky* links which drop packets in random bursts. Again, we set the number of nodes per pods that experience the failure. We also parameterize the arrival pattern and burst duration distributions. During a burst of loss, the link fails at transmitting packets sent on the link.

We consider scenarios that both do and do not have enough

²For our PFC configuration, we set the pause threshold to the buffer size and leave additional headroom in the buffer for the inflight packets (our default buffer size would incur loss otherwise) and set the threshold for unpausing to 1 packet less. Thus, while the network is lossless, we only allow overload equivalent to our default queues.

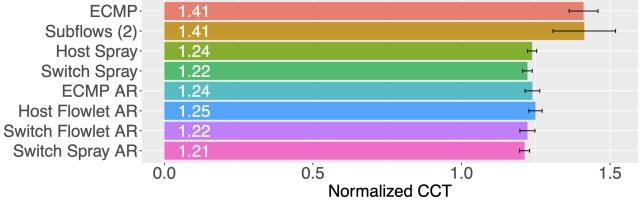


Figure 3: All-to-all completion times for different load balancing schemes with a perfect loss recovery protocol. Values are normalized to minimum possible completion time. Adaptive techniques are denoted with “AR” (adaptive routing).

capacity for the traffic matrix with the failures. In cases where the network can still handle the load, we set the failure values such that the flows in the traffic matrix are not bottlenecked on the failed links. We calculate this value in App. A. In this case, a “perfect” load balancing algorithm should be able to handle the traffic, and is not guaranteed to cause congestion due to the failures since there is sufficient capacity.

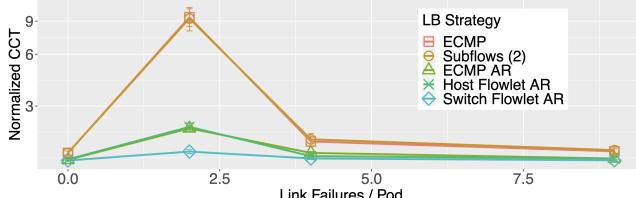
5 Results: Load Balancing Schemes

We begin by determining which granularity of load balancing performs the best with our topology and default all-to-all workload. To evaluate the load balancing scheme without the impacts of loss recovery, we use a minimal mechanism: erasure coding with no overhead. For this evaluation, we unrealistically assume that no additional decoding symbols are necessary to receive the message. Thus, the sender simply sends packets until the receiver has ACKed the full message and no other mechanisms are needed to recover from loss.³ We lift this assumption in the following sections.

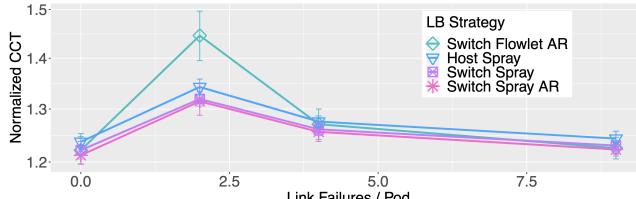
As shown in the top half of Figure 3, packet spraying achieves the lowest collective completion time for the all-to-all traffic matrix among the non-adaptive approaches. ECMP does significantly worse than all other options since any hash collisions will cause continuous overload for the duration of the flows. We also note that adding subflows in hopes of increasing the entropy and therefore the load balancing (we add two subflows for every flow) does not change the average completion time and instead only increases the variability. We experimented with higher numbers of subflows and, at least for this all-to-all workload, did not observe a performance gain (we revisit the role of subflows later in §7).

The values in the figure are normalized to the lowest possible completion time given the host-ToR link rate (*i.e.*, with no loss or queueing delay). In this setup, the minimum possible completion time is $\sim 22.5\text{ms}$. Thus a difference of 10% in normalized completion time is a difference of 2.25ms in which a GPU could perform billions of FLOPs [15]. Further, with a large enough difference (and a large message size), this could have a noticeable impact on training time as the

³Given the delay between when the sender transmits the final *necessary* packet and when it receives the corresponding ACK, some additional data may be transmitted. To avoid extra transmissions, an implementation may instead opportunistically stop, but we leave exploration of these options to future work.



(a) Normalized completion time for flow and flowlet-granularity load balancing schemes across failure counts.



(b) Normalized completion time for non-flow-based load balancing schemes across failure counts.

Figure 4: Completion times across failure counts of different load balancing schemes. Each link loses 90% of its bandwidth in the 2 failures case, 60% for 4 failures, and 20% for 9 failures. In all cases, there is sufficient capacity for the overall load. Note the different y-axes. In-switch flowlet AR is shown in both graphs for reference.

communication time compounds over iterations.

Given these basic results, we determine how adaptivity changes the results, if at all. We consider a few *in-switch* forms of adaptivity: per-packet adaptive routing (AR), flowlet AR⁴, and ECMP AR. In each of these cases, the given unit of flow is placed on a link with the shortest queue.

We note that there is a large space of potential in-host adaptive spraying policies (with some design points explored in [8]) which may make the performance even better. We leave the finding the best policy to future work and assume a naive approach that changes the flow label on every packet since it produces similar results to the more advanced in-network schemes.

We see that many of these approaches reach similar performance: packet spraying in all forms and adaptive approaches all achieve completion times within 3% of the best-performing option (adaptive in-switch packet spraying). This similarity is important as each of these options require in-network support (whether adaptive or not) *except* host-based packet spraying and flowlet routing. This leads to our first takeaway:

■ *Under the assumptions of our default scenario (ideal loss recovery, all-to-all workloads, no failures), host-based schemes based on packet spraying (as in NDP [24]) or adaptive flowlets (akin to PLB [43]) emerge as our leading options since they achieve comparable performance (within 4%) of the best-performing solutions without requiring specialized switch support.*

We now consider whether these results hold as the number of failures in the network is varied. These failure cases vary

⁴We note that with paced senders, flowlets [30] are unlikely to occur. In this setup, flowlet routing behaves more as per-packet load balancing since packets are sufficiently spaced in time in this workload. In higher-rate workloads considered later in the section, this is not the case.

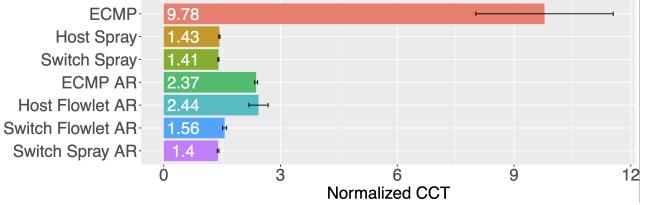


Figure 5: All-to-all completion times for different load balancing schemes in a failure scenario with insufficient bandwidth for the workload (3 failures per pod with 90% BW lost).

from very skewed (2 links with 90% of their bandwidth lost) to fairly uniform (9 out of the 16 links with only 20% of their bandwidth lost). For now, in each of these cases, we ensure that the total bandwidth leaving/entering the pod is sufficient to carry the workload (*i.e.*, these links do not become the bottleneck for the flows).

As shown in Figure 4, all methods of load balancing perform worse in failure scenarios (higher normalized completion times). In the highly-skewed scenario (2 failures), ECMP does particularly poorly as the flows mapped to that link are stuck on the low bandwidth links, even after the other flows have completed. Accordingly, even the modest amount of flexibility given by in-switch adaptive ECMP improves the completion time by about 77%. Meanwhile, our in-host flowlet implementation achieves very similar results. We note that the in-switch form of flowlet routing performs even better, though largely because for the configured gap between flowlets and the pacing rate in this specific workload, flowlets are often only 1 packet (in higher-rate workloads considered later, this is not the case). This represents a limitation of in-switch flowlet routing based on timing: depending on the workload, the given threshold may behave very differently.

In all failure cases, the packet-based schemes perform much better than all other granularities and the gains from adaptivity under failure are far more modest. Across all failure scenarios, adaptive routing does not significantly change the completion time and remains approximately 20-30% away from optimal. Closer examination of our simulation output reveals this is the price of only local state: the switch with the failed uplinks is still offered load as if it has full capacity, but cannot fully serve it. A solution based on global state might instead have steered some traffic away from the problematic switch earlier in the path. This leads to our second takeaway:

■ *Load balancing schemes that operate on a global view of network conditions (e.g., [5, 7, 41]) could offer significant performance gains under failure and is an important direction for future work in the very uniform scenarios of AI clusters. These potential gains (upto ~25%) outweigh the performance differences between the various flavors of existing load-balancing schemes.*

We now consider failures in which the lost bandwidth is significant enough that the calculated pacing rate for senders produces more load than the network can serve (see App. A). Specifically, we consider a scenario similar to the 2 failure

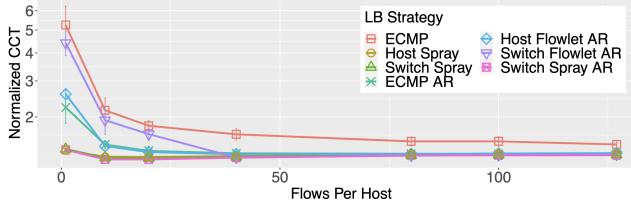


Figure 6: Completion times of different load balancing schemes across considered traffic matrices. The number of flows per host corresponds to the number of concurrent permutation matrices.

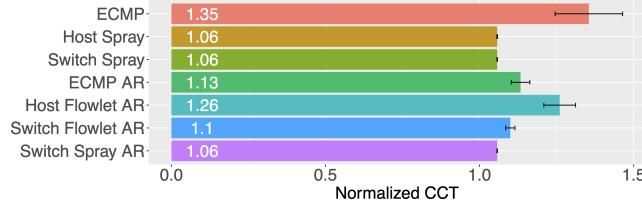


Figure 7: All-to-all completion times for different load balancing schemes with large switch buffers (400KB/port). With this buffer size, queues can easily absorb potential incasts (over 10x the switch radix \times packet size).

case in Figure 4, and increase the failed links to 3 while maintaining the failure percentage. As shown in Figure 5, the relative performance of all the load balancing options remains the same, though all achieve worse CCTs.

Next, we analyze the robustness of these results across other workloads. In particular, we consider permutation traffic matrices and vary the number of concurrent matrices (*i.e.*, varying the potential for incast). As shown in Figure 6, the overall trends remain the same as the number of simultaneous matrices (flows per host) in the workload increases. Spraying remains the best-performing granularity, while the specifics of the adaptivity matter far less. Notably, here flowlet routing that is based on packet timing behaves more like ECMP as the packets are paced more closely in time (left side of Figure 6), meaning that the switch never finds an opportunity to re-route.⁵ These results lead to an important takeaway:

■ *Host-based load balancing based on adaptive flowlet routing performs poorly for collectives in which each host has a small number of simultaneous flows (e.g., permutations) and in certain failure scenarios (2 failures/pod in Fig 4(a)). Hence, we eliminate it from consideration and are left with host-based packet spraying as our leading scheme.*

Lastly, we consider whether these results hold in a topology with large switch buffers. In this setup, switches have 400KB per-link buffers (>10 x the baseline). Therefore, the consequence of overload is primarily queueing delay rather than loss as switches are able to easily absorb incast. As shown in Figure 7, for the all-to-all traffic matrix, the relative performance remains mostly the same, though per-packet approaches increase their performance gap over other granularities, near-

⁵Of course, more opportunities for re-routing may appear in a non-paced flow with window-based sending, but it is unclear why adding bursts would have any other benefit in this setting.

Workload	Buffer Size	Packet Spraying			ECMP
		Host	Switch	Switch AR	
ATA	Large	1.06	1.06	1.06	1.35 (0.05)
	Small	1.24	1.22	1.22	1.41 (0.02)
Perm	Large	1.25	1.26	1.22	5.60 (0.58)
	Small	1.30	1.32	1.31	5.25 (0.51)

Table 2: Average CCT across runs for different packet spraying approaches and ECMP with small and large buffer configurations. The standard deviation (SD) is provided for ECMP in parentheses. For all spraying approaches, the SD was never more than 0.01.

ing optimal. A summary of the impact of large buffers for two extremes of workloads is shown in Table 2 with ECMP for reference. Large buffers result in better performance for spraying in both workloads (though the impact is smaller for the permutation matrix). With ECMP, however, the difference is within error. This suggests that a significant portion of the gap between spraying and optimal is due to temporary overloads, while for ECMP the sustained overload that comes with hash collisions results in worse performance regardless of whether that comes as queueing delay and loss, or just loss. Hence, our takeaway:

■ *Prevailing wisdom is that large-buffer switches are undesirable in datacenters since they inflate latencies and increase switch power and cost. This guideline should be re-examined for AI clusters, where the performance benefits they enable (up to $\sim 20\%$) may warrant their use.*

6 Results: Loss Recovery Schemes

Given our conclusions from §5, we now focus on host-based packet spraying and analyze the performance implications of different loss recovery protocols with packet spraying.

6.1 Setting

In a packet spraying scenario, reordering between packets in a flow is likely and a serious concern for reliability protocols [20, 52]. For the network configuration we consider, we would need a dupACK threshold of 32 to ensure that dupACK-based retransmissions were due to losses and not reordering (as can be derived using the analysis in §7). Of course, this would delay retransmitting in the cases where packets were lost (and perhaps having them only detected by timeouts). On the other hand, as we explore more in our results, leaving the threshold as-is can cause many spurious retransmits in some scenarios (up to $\sim 37\%$ of packets sent for TCP). For now, we leave TCP’s loss recovery unmodified, but explore options in this area in §7.

Clearly, reordering can have performance implications for a protocol which relies on ACK numbers to infer loss. Thus, we examine a set of alternatives to determine which are resilient to this reordering in the scenarios we consider. Fundamentally, one can detect loss in a few ways: (1) timeouts, (2) ACK-/sequence numbers, and (3) explicit loss notifications (*e.g.*, trimming [9, 24, 40]). Timeouts will detect all forms of loss, but depending on the RTT, may be costly or lead to spurious retransmissions. Duplicate ACKs are ambiguous and may be caused by packet reordering, but can allow for faster response times than a timeout (assuming the time between packets is significantly lower than the RTT). Trimming is explicit

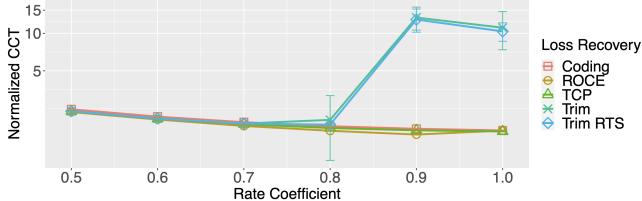


Figure 8: All-to-all completion times across flow rates ($x = \text{pacing rate} / \text{calculated rate}$) with different loss recovery mechanisms and host-based packet spraying load balancing. “RTS” denotes the trimming configuration in which trimmed packets are immediately returned to the sender rather than reflected (“Trim”).

about *congestive* loss, but requires resources (queue space, bandwidth) to transmit that information and does not help in the case of non-congestive loss (where other mechanisms such as timeouts will be necessary). In our setting, both sources of loss are important as flaky hardware may drop packets and faults in the load balancing may cause congestive loss.

We also note another common loss recovery approach in the AI infrastructure space: ROCE and PFC [21, 42]. As mentioned in §2, ROCE assumes in order deliver and requests retransmissions when there are gaps in the arriving sequence numbers. Therefore, ROCE is vulnerable to reordering in the same way as TCP. There have been industry efforts to avoid the impacts of packet spraying by reordering the packets in the NIC [14, 46]. But, fundamentally, NIC buffering solutions must exist either transparently to the transport (in which case sequence numbers cannot serve as a signal of loss, delaying necessary retransmissions) or above the transport, in which the reordering remains an issue.

Lastly, as noted in §4, we consider a loss recovery mechanism based on erasure coding [38, 47]. We now model this with an overhead of 5%; sources send packets until 105% of the messages size has been received. The code itself only requires a few extra symbols to recover the message with high probability, so this represents an especially pessimistic overhead to essentially guarantee message recovery. Again, we do not model the time to encode the symbols, which would pose an additional overhead [50].

6.2 Results

To begin, we compare our selected loss recovery mechanisms for an all-to-all workload on topologies with no failures. As noted in §3, the appropriate rate for a flow to send will depend both on the quality of the load balancing as well as the loss recovery mechanism. Thus, for each considered workload, we determine the best rate for the given loss recovery mechanism when host-based spraying is deployed by sweeping rates and determining which minimizes the CCT.

In Figure 8, the CCT for different loss recovery schemes is shown as the pacing rate of the flows increases to their full fair share (*i.e.*, “Rate Coeficient” = pacing rate / fair share rate). As shown in the figure, for the all-to-all workload, TCP loss recovery performs similarly to other options and achieves its

best CCT at full line rate (rate coefficient = 1.0) while both trimming-based protocols (return-to-sender (RTS) enabled vs. default reflection at the receiver) have dramatically worse performance at these rates. As the rate increases, the overhead of queueing and sending the trimmed packets only makes the congestion worse. This resource overhead is especially important in the high-incast scenarios in an all-to-all as buffer space becomes critical. For example, even with a ratio of data packet size to trimmed packet size of 100:1 [8], traffic matrices with incast patterns with over 100 flows to/from a given node can cause at least two⁶ additional drops and take the associated bandwidth as well. Further, we found that these senders had to have desynchronization mechanisms when receiving NACKs or trimmed packets as bursts of loss could continually occur, causing the collective to never complete (as acknowledged in [24]).

Surprisingly, in this setup, TCP-based senders experienced less than 5 spurious retransmits (for a message size of 500 packets) on average, even with packets taking all possible paths for all rate coefficients considered in Figure 8. Further inspection revealed this to be due to the interpacket delay set by the calculated per-flow rate. Due to the high number of flows in the all-to-all, for any given flow, the fair share rate was sufficiently low that packets in that flow are spaced out enough in time that reordering due to spraying is unlikely. In this case, dupACKs will only occur due to actual packet losses. Though, if the time between packets is sufficiently large, a timeout will occur before there are multiple duplicate ACKs. Accordingly, most flows in this regime experience no retransmissions from dupACKs and instead rely entirely on timeouts.

Since this behavior is caused by low per-flow rates, we experiment with higher-rate traffic matrices with fewer concurrent permutation matrices. The left-most point in Figure 9 shows the best-achieved CCTs for our loss recovery schemes for a single permutation matrix. As shown in the figure, in scenarios with low numbers of flows per link (and therefore higher per-flow rates), the potential degree of reordering is higher and the performance of protocols which rely on ordering to infer loss suffer. For example, with 1 flow per sender, a flow with TCP loss recovery sending at full rate experiences 99 spurious retransmits on average vs. less than 1 on average for 127 flows/sender (all-to-all). For ROCE, this impact is even more dramatic as its loss recovery mechanism expects in-order delivery. Despite this significant performance impact, it is worth noting that the overall performance of TCP in this setup is still better than TCP with other granularities of load balancing (~2x over ECMP and comparable to the best flowlet approach with perfect loss recovery in Figure 6). In addition, experiments with TCP without fast recovery enabled (no response to duplicate ACKs) only produced worse performance. In contrast to TCP, while trimming protocols

⁶This is because the number of lost packets, l is sufficiently high to need to evict more than one queued data packet ($\lceil l \cdot \frac{\text{size(header)}}{\text{size(packet)}} \rceil > 1$).

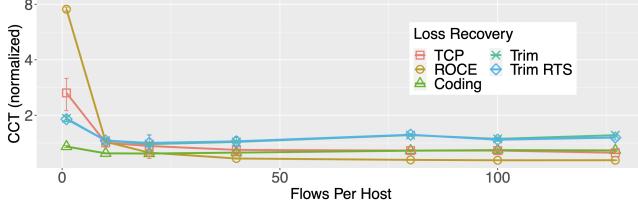


Figure 9: Completion times of loss recovery protocols across workloads (concurrent permutation matrices) with host spraying as the load balancing mechanism.

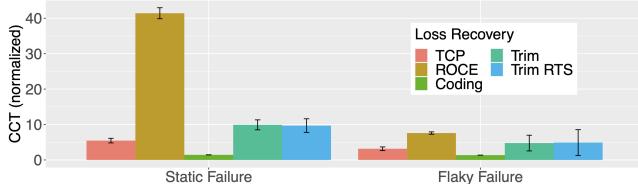


Figure 10: Completion times in failure scenarios for single permutation matrix. In both failures, 2 links per pod are affected. In the static case, the links lose 90% of their bandwidth for the entire simulation. In the flaky failure, random bursts of drops occur every 100us and last for 10us on average.

suffered in the all-to-all scenario, in the low flow/sender scenario in Figure 9, these protocols outperform TCP (by about 70% of optimal) as they do not have to react to reordering for fast responses to loss. This leads us to our next takeaway:

■ *Loss recovery based on sequential delivery (TCP and ROCE) experiences bad performance when per-flow rates are high (significant reordering) and trimming-based approaches degrade when there are many flows and incast is significant. This points to the potential value of a hybrid approach, which we leave to future work.*

We now consider failures as well to see how each form of loss recovery fares. As shown in Figure 10, here again the continuous congestion in a static failed link scenario causes trimming-based protocols to suffer an overhead and ultimately have a higher CCT. The right side of the figure shows the results for our other failure model: flaky links. Here, the two flaky links per pod experience bursts of loss with Poisson arrivals with a mean of 100us and an exponentially distributed duration of 10us. Importantly, these failures do not cause trimmed packets since the packets are lost on the wire. As shown in the graph, the relative ordering of the loss recovery schemes remains the same, though TCP and coding are the most robust to the more hidden mode of failure.

From the results thus far, one might conclude that erasure coding is the best approach, but it is important to note the role of buffer sizes. Assuming that hosts are sending at the correct rates for the traffic matrix and topology, the queue size determines the ability of the network to absorb the load imbalances that may come with ECMP-based packet spraying. Accordingly, we compare the loss recovery schemes in the large buffer scenario. In this scenario, we find that TCP and both trimming approaches can reach CCTs within 1% of optimal for an all-to-all while coding must pay its constant 5% overhead and

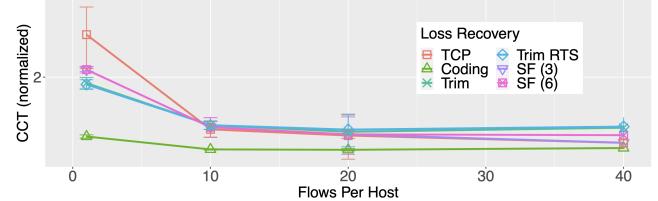


Figure 11: Our subflow approach (“SF”) against the other best-performing options across workloads in Figure 9. Only cases with few concurrent permutation matrices are shown as after 40, the calculated number of subflows is 1 and the behavior is the same as TCP loss recovery.

ROCE incurs a roughly 6% overhead from a small amount of reordering. These large buffers erase most differences in the performance of these protocols, as congestive loss disappears.

Notably, even under our default small-buffer regime an erasure-coding-based loss recovery scheme proves to have very stable performance across all workloads tested (Figure 9). For this evaluation, we assume a (large) 5% overhead for decoding symbols (*i.e.*, the receiver must ACK 2.1MB of data). While this overhead must be paid, no matter the loss rate, the approach has very stable performance. Though, the main open question is whether the encoding/decoding can be implemented efficiently. We conjecture that coding could be efficiently implemented in NIC hardware much as encryption and other operations are today, but this is an area that requires further investigation and prototyping experience. This leads us to our final takeaway:

■ *Of the schemes we evaluated, erasure coding-based loss recovery emerges as our leading option, though its practical implementation remains an open question.*

7 Putting the Pieces Together

Our results so far suggest that a promising transport solution for AI clusters would implement congestion control via pacing at a fixed rate (computed based on the topology and workload), host-based packet spraying for load balancing, and erasure coding for loss recovery.

Since erasure coding is not commonly supported today, one might ask whether there is a good alternative that relies on existing well-understood technologies? We address this next.

Our results showed that trimming-based approaches and TCP loss recovery both had (different) failure modes: trimming suffered in cases of incast while TCP struggled when reordering was likely. In light of our results, we now determine if we can create a loss recovery design point which achieves the performance of trimming with low flow/link scenarios and the performance of TCP-like protocols when the degree of potential incast is high.

As noted in §6, at high numbers of flows per link, the individual rate of any one flow was low enough that there is a very low chance of reordering (*i.e.*, the timing between packets was high enough that the chance of receiving three packets in the time that one is delayed is very low or zero given the queue sizes). While this was just a side effect of the

workload in our evaluations, it poses an opportunity to make an off-the-shelf loss recovery protocol robust to the reordering that packet spraying can cause. We note that prior work has addressed some concerns with reordering and TCP [20, 57], but makes very different assumptions about the workload, objectives, and congestion control.

Based on this observation, we have two variables available to minimize the chance of spurious retransmits: the dupACK threshold and the time between packets. While we should not send packets arbitrarily slowly to prevent reordering, we can split a flow into several subflows which have a lower pacing rate individually than the overall flow. In this approach a subflow only serves as a unit of accounting (for ACKs and therefore dupACKs) and can have its packets sprayed arbitrarily according to the load balancing scheme.

We can compute the number of dupACKs resulting from a single delayed packet for a given queue size (in bytes) q , link bandwidth (in Bps) b , and the topology (here, we assume a 3-tier fat-tree). In our topology, the maximum number of different queues that two packets in the same flow could encounter is 4 (the links to/from the source and destination do not have redundant options), so the maximum possible *difference* in queueing delay is: $D = \frac{4q}{b}$. If packets within the flow are sent at rate r (Bps), then the maximum number of dupACKs resulting from the delay of one packet is $\frac{Dr}{p}$ for packet size p . Therefore, a loss detection system based on some number of duplicate ACKs, t , can be triggered due to spraying-caused reordering if $t < \frac{Dr}{p}$. Simplifying this gives:

$$t < \frac{4qr}{b} \quad (1)$$

Therefore, we can set the number of subflows such that the per-subflow rate, r , is low enough that for a small threshold, t , reordering from spraying is unlikely (or impossible) to trigger a retransmission.

For dupACK threshold of 3 and one flow per host in our default topology, the number of subflows necessary is only 11. If the flows can tolerate the potential delay to detecting loss and set the threshold to 6, this value can be halved. Further, we note that this value does *not* depend on the link bandwidth as the rate of the subflows relative to the bandwidth (r/b) is only a function of the number of subflows and the traffic matrix (*i.e.*, for a traffic matrix with f flows per host and s subflows per flow, $r = \frac{b}{fs}$, so $\frac{r}{b}$ in Eqn. 1 is equivalent to $\frac{1}{fs}$). Accordingly, experiments with 400 Gbps links resulted in the same relative performance as 100 Gbps. Overall, the number of subflows is a function of the queue size and the hop count, both values that are generally kept low for performance/cost anyways.

Two configurations of our subflow approach are shown in Figure 11: one with the dupACK threshold used to calculate the number of subflows set to 3 and another at 6. As shown in the figure, at low levels of flows per host, this approach performs better than TCP (55% better) and is near the performance of trimming protocols (within 20%). Thus, while it does not quite reach the performance of trimming, it does not experience the

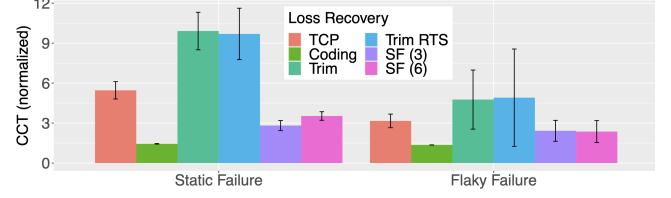


Figure 12: Our subflow approach (“SF”) under failure compared to the other best-performing options from Figure 10. The failure scenarios are the same as in Figure 10: a static loss of bandwidth in two links, and random bursts of loss in two links.

failure model of unmodified TCP loss recovery. At higher numbers of flows, the number of subflows necessary is 1, making the performance equivalent to TCP (not shown for redundancy). We experimented with higher dupACK thresholds, though as the value increases the performance degrades as more legitimate losses are ignored. In the case of failure (as shown in Figure 12) this approach outperforms both trimming and TCP.

We note that this is a similar approach work [21] which splits messages across multiple RDMA QPs in order to increase the entropy of the flows on the network for hashing. Each flow for each sub-message has a different five-tuple and therefore is hashed differently, resulting in better load balancing. In our scenario, however, we are motivated by the loss recovery protocol and use near-optimal (spraying) load balancing transparently.

In summary, our results suggest that, with this simple extension to leverage subflows, TCP loss recovery (or similar mechanisms such as ROCE) remains an attractive option for our ideal transport solution, though additional work remains to validate this design point in deployment.

8 Conclusion

In this paper, we compared the performance of different load-balancing schemes for AI training clusters. Our results make the case for host-based packet-granularity spraying as the load balancing scheme of choice but also exposes how its performance is sensitive to factors such as failure, workload, switch buffer sizing, and - perhaps most importantly - the particular choice of loss recovery scheme. On the latter, we identify erasure coding as a promising but (arguably) under-explored option for loss recovery. We also show - perhaps surprisingly - that there is no clear winner between TCP’s traditional ACK-based approach to loss recovery and more recent proposals for packet trimming. This leads us to sketch a new design point for loss recovery that combines known techniques (duplicate ACKs, subflows) to avoid the failure cases of both TCP and trimming.

We recognize that our results are indicative rather than conclusive and much work remains to confirm our findings. Perhaps most important would be to validate our findings through deployment at scale. While some of the individual components of the schemes we evaluate are widely implemented and deployed, their synthesis as we studied in this paper are less so. One exception to this is erasure coding which is less commonly implemented. Thus, implementing and

evaluating a hardware-based implementation of coding-based loss recovery is another topic for future work. Finally, while we attempted to capture common policies and design options, we recognize that the design space of policies and schemes is far larger than what we covered in this paper. A more exhaustive evaluation of policies for adaptive routing, setting duplicate ACK threshold, and so forth is another area for future work.

References

- [1] Tomahawk3 / bcm56980 series. <https://github.com/Broadcom/csg-htsim>, 2023.
- [2] Internet congestion control (iccr) meeting session at ietf122. <https://www.youtube.com/watch?v=6E4zwP078no&t=5799s>, 2025.
- [3] Vamsi Addanki, Prateesh Goyal, Ilias Marinos, and Stefan Schmid. Ethereal: Divide and conquer network load balancing in large-scale distributed training, 2025.
- [4] Mohammad Al-Fares, Sivasankar Radhakrishnan, Barath Raghavan, Nelson Huang, Amin Vahdat, et al. Hedera: dynamic flow scheduling for data center networks. In *Nsdi*, volume 10, pages 89–92. San Jose, USA, 2010.
- [5] Mohammad Alizadeh, Tom Edsall, Sarang Dhamapurikar, Ramanan Vaidyanathan, Kevin Chu, Andy Fingerhut, Vinh The Lam, Francis Matus, Rong Pan, Navindra Yadav, et al. Conga: Distributed congestion-aware load balancing for datacenters. In *Proceedings of the 2014 ACM conference on SIGCOMM*, pages 503–514, 2014.
- [6] Mohammad Alizadeh, Shuang Yang, Milad Sharif, Sachin Katti, Nick McKeown, Balaji Prabhakar, and Scott Shenker. pfabric: Minimal near-optimal datacenter transport. *ACM SIGCOMM Computer Communication Review*, 43(4):435–446, 2013.
- [7] Theophilus Benson, Ashok Anand, Aditya Akella, and Ming Zhang. Microte: Fine grained traffic engineering for data centers. In *Proceedings of the seventh conference on emerging networking experiments and technologies*, pages 1–12, 2011.
- [8] Tommaso Bonato, Abdul Kabbani, Ahmad Ghalayini, Michael Papamichael, Mohammad Dohadwala, Lukas Gianinazzi, Mikhail Khalilov, Elias Achermann, Daniele De Sensi, and Torsten Hoefler. Reps: Recycled entropy packet spraying for adaptive load balancing and failure mitigation, 2025.
- [9] Tommaso Bonato, Abdul Kabbani, Daniele De Sensi, Rong Pan, Yanfang Le, Costin Raiciu, Mark Handley, Timo Schneider, Nils Blach, Ahmad Ghalayini, Daniel Alves, Michael Papamichael, Adrian Caulfield, and Torsten Hoefler. Fastflow: Flexible adaptive congestion control for high-performance datacenters, 2024.
- [10] Broadcom. Tomahawk3 / bcm56980 series. <https://www.broadcom.com/products/ethernet-connectivity/switching/strataxgs/bcm56980-series>.
- [11] Qizhe Cai, Mina Tahmasbi Arashloo, and Rachit Agarwal. dcpim: Near-optimal proactive datacenter transport. In *Proceedings of the ACM SIGCOMM 2022 Conference*, pages 53–65, 2022.
- [12] Zixian Cai, Zhengyang Liu, Saeed Maleki, Madanlal Musuvathi, Todd Mytkowicz, Jacob Nelson, and Olli Saarikivi. Synthesizing optimal collective algorithms. In *Proceedings of the 26th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming, PPoPP ’21*, page 62–75, New York, NY, USA, 2021. Association for Computing Machinery.
- [13] Neal Cardwell, Yuchung Cheng, C Stephen Gunn, Soheil Hassas Yeganeh, and Van Jacobson. Bbr: Congestion-based congestion control: Measuring bottleneck bandwidth and round-trip propagation time. *Queue*, 14(5):20–53, 2016.
- [14] NVIDIA Corporation. Nvidia bluefield-3 networking platform. <https://resources.nvidia.com/en-us-accelerated-networking-resource-library/datasheet-nvidia-bluefield?lx=LbHvpR&topic=networking-cloud>.
- [15] NVIDIA Corporation. Nvidia h100 tensor core gpu. <https://www.nvidia.com/en-us/data-center/h100/>.
- [16] NVIDIA Corporation. Nvidia spectrum sn5000 series switches. <https://nvdam.widen.net/s/mmvbnpk8qk/networking-ethernet-switches-sn5000-datasheet-us>.
- [17] NVIDIA Corporation. Collective operations. <https://docs.nvidia.com/deeplearning/nccl/user-guide/docs/usage/collectives.html>, 2020.
- [18] NVIDIA Corporation. Nvidia collective communications library (nccl). <https://developer.nvidia.com/nccl>, 2025.
- [19] Andrew R Curtis, Wonho Kim, and Praveen Yalagandula. Mahout: Low-overhead datacenter traffic management using end-host-based elephant detection. In *2011 Proceedings IEEE INFOCOM*, pages 1629–1637. IEEE, 2011.
- [20] Advait Dixit, Pawan Prakash, Y Charlie Hu, and Ramana Rao Komella. On the impact of packet spraying in data center networks. In *2013 Proceedings IEEE INFOCOM*, pages 2130–2138. IEEE, 2013.

- [21] Adithya Gangidi, Rui Miao, Shengbao Zheng, Sai Jayesh Bondu, Guilherme Goes, Hany Morsy, Rohit Puri, Mohammad Riftadi, Ashmitha Jeevaraj Shetty, Jingyi Yang, et al. Rdma over ethernet for distributed training at meta scale. In *Proceedings of the ACM SIGCOMM 2024 Conference*, pages 57–70, 2024.
- [22] Albert Greenberg, James R. Hamilton, Navendu Jain, Srikanth Kandula, Changhoon Kim, Parantap Lahiri, David A. Maltz, Parveen Patel, and Sudipta Sengupta. VI2: a scalable and flexible data center network. In *Proceedings of the ACM SIGCOMM 2009 Conference on Data Communication*, SIGCOMM ’09, page 51–62, New York, NY, USA, 2009. Association for Computing Machinery.
- [23] Chuanxiong Guo, Haitao Wu, Zhong Deng, Gaurav Soni, Jianxi Ye, Jitu Padhye, and Marina Lipshteyn. Rdma over commodity ethernet at scale. In *Proceedings of the 2016 ACM SIGCOMM Conference*, pages 202–215, 2016.
- [24] Mark Handley, Costin Raiciu, Alexandru Agache, Andrei Voinescu, Andrew W Moore, Gianni Antichi, and Marcin 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*, pages 29–42, 2017.
- [25] Keqiang He, Eric Rozner, Kanak Agarwal, Wes Felter, John Carter, and Aditya Akella. Presto: Edge-based load balancing for fast datacenter networks. *ACM SIGCOMM Computer Communication Review*, 45(4):465–478, 2015.
- [26] Christian Hopps. Analysis of an Equal-Cost Multi-Path Algorithm. RFC 2992, November 2000.
- [27] Christian Huitema. Multi-homed TCP. Internet-Draft draft-huitema-multi-homed-01, Internet Engineering Task Force, November 1995. Work in Progress.
- [28] Abdul Kabbani, Balajee Vamanan, Jahangir Hasan, and Fabien 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*, pages 149–160, 2014.
- [29] Srikanth Kandula, Dina Katabi, Bruce Davie, and Anna Charny. Walking the tightrope: responsive yet stable traffic engineering. In *Proceedings of the 2005 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, SIGCOMM ’05, page 253–264, New York, NY, USA, 2005. Association for Computing Machinery.
- [30] Srikanth Kandula, Dina Katabi, Shantanu Sinha, and Arthur Berger. Dynamic load balancing without packet reordering. *ACM SIGCOMM Computer Communication Review*, 37(2):51–62, 2007.
- [31] Naga Katta, Aditi Ghag, Mukesh Hira, Isaac Keslassy, Aran Bergman, Changhoon Kim, and Jennifer Rexford. Clove: Congestion-aware load balancing at the virtual edge. In *Proceedings of the 13th International Conference on emerging Networking EXperiments and Technologies*, pages 323–335, 2017.
- [32] Naga Katta, Mukesh Hira, Changhoon Kim, Anirudh Sivaraman, and Jennifer Rexford. Hula: Scalable load balancing using programmable data planes. In *Proceedings of the Symposium on SDN Research*, pages 1–12, 2016.
- [33] Gautam Kumar, Nandita Dukkipati, Keon Jang, Hassan MG Wassel, Xian Wu, Behnam Montazeri, Yaogong Wang, Kevin Springborn, Christopher Alfeld, Michael Ryan, et al. Swift: Delay is simple and effective for congestion control in the datacenter. In *Proceedings of the Annual conference of the ACM Special Interest Group on Data Communication on the applications, technologies, architectures, and protocols for computer communication*, pages 514–528, 2020.
- [34] Jennifer Langston. Microsoft announces new supercomputer, lays out vision for future ai work. <https://news.microsoft.com/source/features/ai/openai-azure-supercomputer/>, 2020.
- [35] Dan Lenoski and Nandita Dukkipati. Google opens falcon, a reliable low-latency hardware transport, to the ecosystem. <https://cloud.google.com/blog/topics/systems/introducing-falcon-a-reliable-low-latency-hardware-transport>, 2023.
- [36] Shenggui Li and Siqi Mai. Paradigms of parallelism. https://colossalai.org/docs/concepts/paradigms_of_parallelism/.
- [37] Michael Luby. LT Codes . In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, page 271, Los Alamitos, CA, USA, November 2002. IEEE Computer Society.
- [38] Lorenz Minder, Amin Shokrollahi, Mark Watson, Michael Luby, and Thomas Stockhammer. RaptorQ Forward Error Correction Scheme for Object Delivery. RFC 6330, August 2011.
- [39] Yang Nie, Zheng Shi, Xinyi Chen, and Liguo Qian. An out-of-order packet processing algorithm of roce based on improved sack. In *2022 IEEE 6th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, pages 1402–1408, 2022.

- [40] Vladimir Olteanu, Haggai Eran, Dragos Dumitrescu, Adrian Popa, Cristi Baciu, Mark Silberstein, Georgios Nikolaidis, Mark Handley, and Costin Raiciu. An edge-queued datagram service for all datacenter traffic. In *19th USENIX Symposium on Networked Systems Design and Implementation (NSDI 22)*, pages 761–777, 2022.
- [41] Jonathan Perry, Amy Ousterhout, Hari Balakrishnan, Devavrat Shah, and Hans Fugal. Fastpass: A centralized “zero-queue” datacenter network. In *Proceedings of the 2014 ACM conference on SIGCOMM*, pages 307–318, 2014.
- [42] Kun Qian, Yongqing Xi, Jiamin Cao, Jiaqi Gao, Yichi Xu, Yu Guan, Bin Zhang Fu, Xuemei Shi, Fangbo Zhu, Rui Miao, et al. Alibaba hpn: A data center network for large language model training. In *Proceedings of the ACM SIGCOMM 2024 Conference*, pages 691–706, 2024.
- [43] Mubashir Adnan Qureshi, Yuchung Cheng, Qianwen Yin, Qiaobin Fu, Gautam Kumar, Masoud Moshref, Junhua Yan, Van Jacobson, David Wetherall, and Abdul Kabbani. Plb: congestion signals are simple and effective for network load balancing. In *Proceedings of the ACM SIGCOMM 2022 Conference*, pages 207–218, 2022.
- [44] Costin Raiciu, Dragos Niculescu, Marcelo Bagnulo, and Mark James Handley. Opportunistic mobility with multipath tcp. In *Proceedings of the sixth international workshop on MobiArch*, pages 7–12, 2011.
- [45] Siddhartha Sen, David Shue, Sunghwan Ihm, and Michael J Freedman. Scalable, optimal flow routing in datacenters via local link balancing. In *Proceedings of the ninth ACM conference on Emerging networking experiments and technologies*, pages 151–162, 2013.
- [46] Leah Shalev, Hani Ayoub, Nafea Bshara, and Erez Sabbag. A cloud-optimized transport protocol for elastic and scalable hpc. *IEEE Micro*, 40(6):67–73, 2020.
- [47] A. Shokrollahi. Raptor codes. *IEEE Transactions on Information Theory*, 52(6):2551–2567, 2006.
- [48] Cha Hwan Song, Xin Zhe Khooi, Raj Joshi, Inho Choi, Jialin Li, and Mun Choon Chan. Network load balancing with in-network reordering support for rdma. In *Proceedings of the ACM SIGCOMM 2023 Conference*, ACM SIGCOMM ’23, page 816–831, New York, NY, USA, 2023. Association for Computing Machinery.
- [49] Rajeev Thakur and William D. Gropp. Improving the performance of collective operations in mpich. In Jack Dongarra, Domenico Laforenza, and Salvatore Orlando, editors, *Recent Advances in Parallel Virtual Machine and Message Passing Interface*, pages 257–267, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg.
- [50] Rajeev Thakur and William D. Gropp. Improving the performance of collective operations in mpich. In Jack Dongarra, Domenico Laforenza, and Salvatore Orlando, editors, *Recent Advances in Parallel Virtual Machine and Message Passing Interface*, pages 257–267, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg.
- [51] UEC. Ultra ethernet consortium. <https://ultraethernet.org/>.
- [52] Erico Vanini, Rong Pan, Mohammad Alizadeh, Parvin Taheri, and Tom Edsall. Let it flow: Resilient asymmetric load balancing with flowlet switching. In *14th USENIX Symposium on Networked Systems Design and Implementation (NSDI 17)*, pages 407–420, 2017.
- [53] Peng Wang, Hong Xu, Zhixiong Niu, Dongsu Han, and Yongqiang Xiong. Expeditus: Congestion-aware load balancing in clos data center networks. In *Proceedings of the Seventh ACM Symposium on Cloud Computing*, pages 442–455, 2016.
- [54] Cliff Woolley. Nccl: Accelerated multi-gpu collective communications. <https://images.nvidia.com/events/sc15/pdfs/NCCL-Woolley.pdf>.
- [55] David Zats, Tathagata Das, Prashanth Mohan, Dhruba Borthakur, and Randy Katz. Detail: Reducing the flow completion time tail in datacenter networks. In *Proceedings of the ACM SIGCOMM 2012 conference on Applications, technologies, architectures, and protocols for computer communication*, pages 139–150, 2012.
- [56] Hong Zhang, Junxue Zhang, Wei Bai, Kai Chen, and Mosharaf Chowdhury. Resilient datacenter load balancing in the wild. In *Proceedings of the Conference of the ACM Special Interest Group on Data Communication*, pages 253–266, 2017.
- [57] Ming Zhang, Brad Karp, Sally Floyd, and Larry Peterson. Rr-tcp: a reordering-robust tcp with dsack. volume 2003, pages 95–106, 11 2003.

A Bottleneck Under Failures

In this appendix, we define an inequality to determine if a failure scenario will cause flows to be bottlenecked at/because of the failure. Accordingly, we seek to find the failure parameters such that flows remain bottlenecked on the ToR-host link. We assume failures that occur on links between pods and the core routers. For an all-to-all traffic matrix in a 3-tier clos with switch radix k , this requires setting f , the number of failed links, and p the percentage of bandwidth lost such that:

$$\frac{k^3}{4} - 1 > \frac{\frac{k^2}{4}(\frac{k^3}{4} - \frac{k^2}{4})}{\frac{k^2}{4} - f(1-p)} \quad (2)$$

The LHS represents the number of flows on the host-ToR link while the RHS is the number of flows (numerator) per link (denominator) leaving the pod. Thus, if the above inequality is not true, the failure scenario *does* impact the networks ability to serve the offered load and the calculated per-flow rates will cause overload, even with perfect load balancing.