Best-Path vs. Multi-Path Overlay Routing

David G. Andersen, Hari Balakrishnan
MIT Laboratory for Computer Science
{dga,hari}@lcs.mit.edu

Alex C. Snoeren
University of California, San Diego
snoeren@cs.ucsd.edu

Paper #E00-1345536670 — 13 pages.

Abstract

Time-varying congestion on Internet paths and failures due to software, hardware, and configuration errors often disrupt packet delivery on the Internet. In recent years, optimizations at the IP routing layer as well as overlay network architectures have been proposed to improve packet delivery in the face of these fundamental problems by taking advantage of multiple paths between two network locations. These proposals rely on a path-independence assumption in order to work well; i.e., they work best when the problems on different paths between two locations are uncorrelated in time.

This paper examines the extent to which this assumption holds on the Internet by analyzing 14 days of data collected from 30 nodes in the RON testbed. We examine two problems that manifest themselves—congestion-triggered loss and path failures—with an eye toward understanding the degree to which they are uncorrelated on network paths. In so doing, we also compare two different ways of taking advantage of path redundancy proposed in the literature: mesh routing based on packet replication, and reactive routing based on adaptive path selection.

1 Introduction

The routing infrastructure in the Internet does not attempt to provide loss-free packet delivery between end points. End-to-end transfers observe packet losses due to several reasons, including network congestion, path failures, and routing anomalies. As a result, applications and transport protocols have to cope with these packet losses. This is often done using packet retransmissions, coupled with a reduction in sending rate to react to congestion, resulting in degraded throughput and increased latency. Long outages on paths lasting several minutes occur in practice [1, 18], and end-to-end connections that are in the middle of data transfers usually end up aborting when such outages happen.

Over the past few years, both routing optimizations at the IP layer [25, 30, 34, 35] and overlay networks layered on top of the Internet routing substrate [1, 33, 24] have been proposed as ways to improve the resilience of packet delivery to these problematic conditions. These approaches either probe to find a single best path through the Internet, or duplicate packets down multiple paths as a form of path selection and redundancy.

To work well, these reactive routing require that a fundamental property hold, which is that *losses and failures on different network paths be uncorrelated with each other.* A failure or loss on one path from a source to a destination must not overlap in time with the failure of all other paths from the source to the destination.

Mesh routing is the simplest way to add redundant packets to the data stream by duplicating all of the packets along different paths [33]. In this scheme, the overhead is due to redundant packets, but does not require additional probing. When its paths are disjoint, mesh routing is resilient to the failure of a subset of its component paths. In this paper, we examine the behavior of mesh routing when its packets are sent over the Internet as an overlay network, and examine the degree to which its packets are actually lost independently.

In reactive routing implemented with overlay networks, the overlay nodes constantly probe the $O(N^2)$ paths between them, and send packets either directly over the Internet, or forward them via a sequence of other nodes in the overlay when the latter path provides better performance. The overhead in this approach comes from both probes and overlay routing traffic. The probes are required to ensure that when a problem occurs with the current path or when a better path presents itself, traffic is rerouted appropriately to reduce the observed loss rate. Inspired by the approach used in RON [1], we focus on a simple but effective overlay routing method that uses at most one intermediate node in the overlay network to forward packets.

We analyze fourteen days of probes between 30 geographically diverse nodes of the RON testbed. These probes include packets sent back-to-back via various mechanisms to help determine the degree to which failures and losses on the Internet are correlated. Using this data, we examine the performance of reactive routing and mesh routing, and compare their loss rate and latency reduction to the direct Internet path between pairs of nodes.

Our major findings are that:

- The conditional loss probability of back-to-back packets (the probability of losing the second when the first was lost) is high both when sent on the same path (70%) and when sent via different paths (60%).
- The likelihood of multiple paths between a source and a destination simultaneously failing is high, and seems higher in 2003 than in our 2002 data.
- The overall packet loss rate between our hosts is a low 0.42%. Reactive routing reduces this to 0.33%, and mesh routing reduces it to 0.26%. These improvements come primarily from reducing the loss during higher-loss periods of time; there are many hours of the day when the Internet is mostly quiescent and loss rates are low.
- Mesh-based routing and reactive routing appear to exploit different network properties, and can be used together for increased benefits.

The remainder of this paper is organized as follows: Section 2 surveys related work. In Section 3 we discuss the design of a the simple probe-based overlay routing protocol and replication-based simple multi-path protocols that we study empirically in Section 4. In Section 5, we examine the implications of our results on the design of improved routing schemes, and we conclude in Section 6 with a summary of our findings.

2 Background and related work

The Internet was designed as a best-effort medium; hence, it is not surprising that Internet paths often exhibit packet loss. Congested routers, packet collisions, and noisy data links combine to cause various levels of packet loss. Severe burst losses or outages may be exacerbated by link failures, routing problems, or both. Labovitz *et al.* show that routers may take tens of minutes to stabilize after a fault, and that packet loss is high during this period [18]. They also note that route availability is not perfect, causing sites to be unreachable some fraction of the time [19]. Paxson notes that packets are often subject to routing loops and other pathologies [26].

2.1 Reliable transmission

The traditional way to counteract losses in packetized data transfer is to use packet diversity, either through retransmissions, forward error correction (FEC), or a combination of the two. Retransmissions are appropriate for end-to-end protocols, but adding this functionality at the network level can cause problems for TCP's retransmission timers [21]. Furthermore, not all applications require this functionality, and may not be able to bear its cost in latency and bandwidth [31]. Inspired by such applications, we examine loss-optimized routing strategies that do not dramatically increase end-to-end round-trip latencies.

Hop-by-hop ARQ schemes can reduce the delay for certain topologies [3], but require buffering and network support at intermediate nodes. Furthermore, many ARQ schemes are tuned for certain loss characteristics, and function poorly over channels outside of their design space. While these schemes benefit links—such as wireless links—with high bit-error rates, they are not universally applicable in the general Internet context. In particular, these schemes do not apply in the case of congestive losses or link failures, the major causes of loss in the wired Internet.

In contrast, FEC adds redundant information to a stream, allowing the stream to be reconstructed at the receiver even if some of the information is corrupted or missing [15]. They are widely used in wireless systems to protect against bit corruption [23], and more recently in multicast and content distribution systems to protect against packet loss [9, 29]. The latter applications require packet-level—as opposed to bit-level—FEC. We consider packet-level FEC in this paper.

Unfortunately, sending redundant data along the same

Internet path is often not as effective as one might hope due to high packet loss correlation. Bolot examined the behavior of packet trains on a single link between INRIA and the University of Maryland in 1992 [6]. He found that the conditional loss probability of back-to-back packets is high when the packets are closely spaced (~ 8 ms), but returns to the unconditional loss probability when the gap is ~ 500 ms. Similarly, Paxson examined TCP bulk transfers between 35 sites in 1997 [27]. In this work, he found that the conditional loss probability of data packets that were queued together was 10–20 times higher than the base loss rate. We compare our loss probabilities with those of Paxson and Bolot in Section 4.

2.2 Improved routing

While ARQ and FEC schemes can reduce the perceived loss rate of a particular Internet path, there may exist alternative paths that provide lower loss rates. Early ARPANET routing attempted to optimize path selection for congestion [17], but this was removed for scalability and stability reasons. Today, a wide variety of traffic engineering approaches are employed to refine path selection in an attempt to decrease congestion, packet loss, and latency, and increase available bandwidth [2]. Unfortunately these techniques generally operate over long time-scales. As a result of current backbone routing's ignorance of short-term network conditions, the route taken by packets is frequently sub-optimal, a fact noted by the Detour study [32]. Recent network path selection products [25, 30, 34, 35] attempt to provide more fine-grained, measurement-based path selection for single sites.

Recent research in overlay networks has attempted to improve path conditions through indirect routing. The RON project uses active measurements to take advantage of some of these alternate paths [1]. Various Content Delivery Networks (CDNs) use overlay techniques and caching to improve the performance of content delivery for specific applications such as HTTP and streaming video. Overcast and other application level multicast projects attempt to optimize routes for bandwidth or latency [16].

2.3 Multi-path routing

The success of traffic engineering and overlay routing indicates the presence of redundant routes between many pairs of Internet hosts. A variety of approaches have been developed to leverage the existence of multiple, simultaneous paths through multi-path routing. Dispersity routing [22] and IDA [28] split the transfer of information over multiple network paths to provide enhanced reliability and performance. Simulation results and analytic studies have shown the benefits of this approach [4, 5]. Chen evaluated the use of parallel TCP flows to improve performance, but did not examine failures, or real Internet paths [11]. In addition, researchers have suggested combining redundant coding with dispersity routing to improve the reliability and performance of both parallel downloads [8] and multicast communication [33]. Akamai is reported to use erasure codes to take advantage of multiple paths between sites [20], and the designers of the Opus overlay system have proposed the future use of redundant transmission in an overlay, but, to our knowledge, have not yet evaluated this technique [7].

2.4 Sources of shared failures

Multi-path and alternate-path routing schemes make generous assumptions about path independence that may not hold when considering typical Internet paths, as we show in Section 4. Single-homed hosts share the same last-mile link to their provider, creating an obvious shared bottleneck and non-independent failure point. Even multi-homed hosts may have unexpected sources of shared failures. Many providers have some degree of shared physical infrastructure. In 2001, a single train derailment in the Howard Street tunnel in Baltimore, MD, impacted Internet service for at least 4 major US backbone carriers, all of whom used fiber running through the same physical location [13]. We also recently observed that many failures manifest themselves near the network edge, where routing protocols are less likely to be able to route around them [14].

Network failures are not only caused by exogenous factors, but may be the result of network traffic or other failures. These cascading logical failures can cause widespread outages that affect multiple paths or providers [10]. Finally, denial of service attacks or other global Internet problems such as worms and viruses can cause correlated, concurrent failures. For instance, the "Code Red" worm, as a side-effect of its scanning, could crash certain Cisco DSL routers and other products, causing correlated failures based solely upon net-

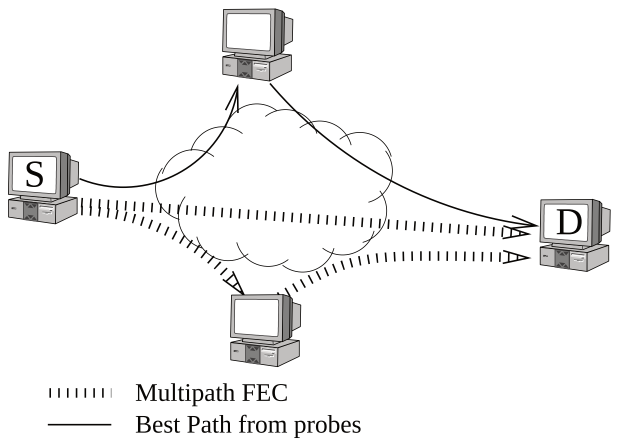

Figure 1: Multi-Path FEC routing and best path routing. In this diagram, probing has determined that the best path is to travel indirectly via the top node. Multi-Path FEC routing sends a packet down the direct path and via a random alternate hop, in this case, the bottom node.

work access technology [12]. We provide an empirical evaluation of the independence of losses on a particular set of Internet paths in Section 4.

3 Design

For much of the paper, we study two mechanisms for enhanced packet routing: probe-based reactive overlay routing, and multi-path redundant routing. These techniques would usually not be used independently. For instance, it's necessary to choose which intermediate nodes to use in redundant routing. The logical way to choose these nodes is via network measurements. The difference is the degree to which resources are allocated to measurements vs. redundant data, a trade-off that we consider further in Section 5.

3.1 Probe-based reactive overlay routing

RON-like systems send frequent probes to determine the availability, latency, and loss rate of the paths connecting the nodes in the overlay. A reactive routing scheme must choose a probing rate R, and a network size N. A generalized scheme would also need to choose the sets of nodes that probe each other. Higher probing rates permit quicker reaction to network change, with more overhead. Larger networks have more paths to explore, but create scaling problems. In the system we evaluate, every node probes every other node once every 15 seconds. When a probe is lost, the node sends an additional string of up to four probes spaced one second apart, to determine if the remote host is down. The paths are selected based upon the average loss rate over the last 100 probes. These are similar to the parameters used in an earlier evaluation of reactive overlay routing [1], but the interval between probes is five seconds longer.

3.2 Redundant multi-path routing

Redundant multi-path routing sends redundant data down multiple, ideally independent, paths, such that a certain fraction of lost packets can be recovered. In this study, we consider 2-redundant mesh routing [33], in which each packet is sent to the receiver twice via different paths. In the most basic scheme, the first packet is sent directly over the Internet, and the second is sent through a randomly chosen intermediate node. We discuss the implications of our results on more complex FEC schemes in Section 5.2. When packet losses are independent, redundant data transmissions can effectively mask even high packet loss rates, but when losses are correlated, FEC schemes lose their effectiveness. We turn to an empirical study of this correlation to understand how common FEC schemes might fare in practice.

4 Evaluation

We evaluate the correlation of losses and failures on a deployed Internet testbed. Table 1 lists the 30 hosts used in our experiments. The hosts are concentrated in the US, but span five countries on three continents. More importantly, the testbed hosts have a variety of access link technologies, from OC3s to cable modems and DSL links. We do not claim that this testbed is representative of the Internet as a whole. However, the nearly nine hundred distinct one-way paths between the hosts do provide a diverse testbed in which to evaluate routing tactics and packet loss relationships.

Table 2 lists the three datasets we examine. The first two, taken in 2002, were measured between 17 hosts on the RON testbed. The third was measured in 2003 between 30 hosts. RON_{wide} measured all combinations of mesh routing and probe-based routing to identify which combinations were most effective at reducing the probability of simultaneous losses. RON_{narrow} measures the three most promising methods with frequent

Name	Location	Description
Aros	Salt Lake City, UT	ISP
AT&T	Florham Park, NJ	ISP
CA-DSL	Foster City, CA	1Mbps DSL
CCI	Salt Lake City, UT	.com
*** CMU**	Pittsburgh, PA	.edu
Coloco	Laurel, MD	ISP
*** Cornell**	Ithaca, NY	.edu
Cybermesa	Santa Fe, NM	ISP
Digitalwest	San Luis Obispo, CA	ISP
GBLX-AMS	Amsterdam, Netherlands	ISP
GBLX-ANA	Anaheim, CA	ISP
GBLX-CHI	Chicago, IL	ISP
GBLX-JFK	New York City, NY	ISP
GBLX-LON	London, England	ISP
Intel	Palo Alto, CA	.com
Korea	KAIST in Korea	.edu
Lulea	Lulea, Sweden	.edu
MA-Cable	Cambridge, MA	AT&T
Mazu	Boston, MA	.com
*** MIT**	Cambridge, MA	.edu in lab
MIT-main	Cambridge, MA	.edu data center
NC-Cable	Durham, NC	RoadRunner
Nortel	Toronto, Canada	ISP
*** NYU**	New York, NY	.edu
PDI	Palo Alto, CA	.com
PSG	Bainbridge Island, WA	Small ISP
* UCSD	San Diego, CA	.edu
*** Utah**	Salt Lake City, UT	.edu
Vineyard	Cambridge, MA	ISP
VU-NL	Amsterdam, Netherlands	Vrije Univ.

Table 1: The hosts between which we measured network connectivity. Asterisks indicate U.S. universities on the Internet2 backbone. Hosts in bold were used in the 2002 data.

Dataset	Samples	Dates
RON_{narrow}	4,763,082	8 Jul 2002 – 11 Jul 2002
RON_{wide}	2,875,431	3 Jul 2002 – 8 Jul 2002
RON_{2003}	32,602,776	30 Apr 2003 – 14 May 2003

Table 2: The three datasets used in our experiments. The RON_{narrow} dataset contains one-way samples for three routing methods. The RON_{wide} dataset has round-trip samples for eleven methods. The RON_{2003} dataset uses a larger number of probing hosts to measure six routing methods.

loss	loss optimized path (via probing)
lat	loss optimized path (via probing)
direct	direct Internet path
rand	indirectly through a random node

Table 3: The types of routes between measurement nodes. Probes consisted of one or more packets of these types, such as *direct rand* (one packet directly, one via a random intermediate node).

one-way probes, sampling each path (for each method) every 45 seconds on average. RON_{2003} measures a few additional routing types between more nodes, and over a longer period of time.[1] Table 3 lists the routing tactics for individual packets; probes consist of one or two packets sent via various routing methods.

We focus primarily on the RON_{2003} dataset, but highlight interesting differences from the prior datasets. This data set focuses on seven routing methods, collected from six sets of probes:

- *Loss*: Probe-based reactive routing that attempts to minimize loss. Requires only probing overhead.

- *Lat*: Probe-based reactive routing that minimizes latency and avoids completely failed links.

- *Direct rand*: 2-redundant mesh routing, with no probing overhead. One copy of each packet is transmitted on the direct Internet path; the second over a random indirect overlay path. There is no delay between the packet transmissions. We use the first packet to predict the behavior of *direct* packets.

- *Lat loss*: Probe-based 2-redundant multi-path routing. In theory, this combination should be able to achieve the best of both worlds. It sends the first copy of each packet over a path selected to minimize loss, and the second over a path selected to minimize latency. We also use this to infer the *lat* packet.

- *Direct direct*: 2-redundant routing with back to back packets on the same path.

- *DD 10 ms*: Two packets sent directly with a 10 ms gap.

- *DD 20 ms*: Two packets sent directly with a 20 ms gap.

[1] Our datasets will be made available on-line.