Measuring TCP Round-Trip Time in the Data Plane

Xiaoqi Chen, Hyojoon Kim, Javed M Aman, Willie Chang, Mack Lee, Jennifer Rexford
Princeton University
{xiaoqic,hyojoonk,javeda,whchang,mackl,jrex}@cs.princeton.edu

ABSTRACT
We present a data-plane algorithm that passively and continuously monitors the Round-Trip Time of TCP traffic, by matching data packets with their associated acknowledgments and calculating a time difference. Compared with traditional measurement systems based on active probing or measuring only SYN/ACK packets, our algorithm passively produces many samples for long-running connections. This enables network operators to observe abnormal RTT increases, which signal possible security or performance issues in the network, in real-time. To satisfy the stringent memory size and access constraints of programmable switches, our algorithm uses a multi-stage hash table data structure to maintain records for in-flight packets; the records not receiving their acknowledgments are lazily expired and overwritten. We implement our algorithm on a Barefoot Tofino programmable switch. Evaluation using a real-world traffic trace from a 10 Gbps campus network link demonstrates that our solution can accurately capture 99% of available RTT samples, using only 4 MB of data-plane memory.

CCS CONCEPTS
• Networks → Data path algorithms; Network measurement;

KEYWORDS
Network Monitoring, Data Plane, P4, TCP RTT, Hash Tables

ACM Reference Format:

1 INTRODUCTION
Round-Trip Time (RTT) is a key metric for network latency. An increasing RTT not only affects user’s Quality of Experience, but also indicates possible performance or security issues in the network, such as congestion or routing changes. Although RTT statistics are often readily available at end hosts [14, 15, 28], an Internet Service Provider (ISP), such as a consumer broadband provider or an enterprise network operator, does not have direct visibility into the latency experienced by its customers. Even in a data center, continuously monitoring RTTs at all hosts is costly. Yet, continuous RTT monitoring would allow an ISP to better understand both the security and performance of its network traffic:

• BGP Routing Security: When an attacker uses BGP routing attacks [3, 19] to detour and intercept traffic, that traffic would likely experience higher-than-normal RTT. Thus, unexpected changes in the RTT to a remote host may signal a reroute, either due to equipment failure, or a routing attack. Continuous RTT monitoring can help the ISP discover reroutes, even if the re-routing happens further downstream.

• Detecting IP Spoofing: A spoofed IP address exhibits discordant RTT values than the legitimate traffic from the same address, therefore RTT can be used to improve the accuracy for IP spoofing detection [13, 16].

• Service-Level Agreement (SLA): An ISP usually has RTT requirements in its SLA with customers. For example, Verizon agrees to set 45 ms and 30 ms as the maximum RTT of intra-North-America and intra-Europe traffic, respectively [26, 27]. Monitoring RTT in real-time allows an ISP or its customer to verify the RTT is within limits, or be notified about an upcoming breach of the SLA.

• Quality of Experience (QoE): An ISP may want to measure the QoE for customers using a variety of applications. Some applications such as video live-streaming are sensitive to high latency and jitter [4], which can be captured in RTT measurements. An increase in RTT may reflect persistent congestion and queuing on peering links [8], which can inform an ISP to upgrade its equipment for those links to better accommodate customers’ demand.

To measure RTT, network operators today rely on active measurement tools such as NDT [10], PingMesh [12], and perfSONAR [22], sometimes after a client reports a degradation in service quality. Meanwhile, passive performance measurement tools (e.g., Ruru [7]) mostly report RTT samples based only on the three-way TCP connection handshake. Such tools cannot capture the latency change during long-running TCP connections such as video streaming. Also, they may be biased when SYN/SYN-ACK packets are processed differently than regular TCP packets; for example, SYN/SYN-ACK packets might go through a middlebox or get delayed by the remote server before accepting new connections.

Figure 1: Match data and ACK packets to measure RTT.
In this paper, we present an algorithm to continuously measure RTT for all outgoing TCP packets, on a programmable switch at an ISP vantage point, passively and continuously. Our algorithm does this by matching an outgoing TCP packet using its sequence number with an incoming packet that has the corresponding acknowledgment number. As illustrated in Figure 1, our algorithm captures RTT samples beyond the three-way handshake, allowing continuous monitoring throughout a TCP session. Running in the data plane of commodity programmable switches gives us the opportunity to measure per-packet RTT in real-time, at a higher line rate. This enables potential future work on real-time mitigation directly in the data plane (e.g., reroute for further inspection) when RTT anomalies are detected (e.g., IP spoofing).

Continuously monitoring RTT in the data plane has several unique challenges. To achieve high throughput and constant-time processing, the programmable switch imposes strict constraints, including limited memory size and memory access pattern. We need to store records for outgoing packets in the memory, then efficiently look up these records to calculate RTT upon seeing incoming acknowledgments (ACKs). Due to the TCP delayed ACK mechanism, some packets never receive their corresponding ACKs, so we need to clean up their records; some ACKs are delayed, which inflates RTT, so we also need to filter them. Finally, as memory is limited in the data plane, sometimes we cannot record every packet and inevitably lose some RTT samples; we want to ensure the subset of RTT samples we indeed measure are unbiased: a high-RTT packet with a late-arriving ACK shall not be discriminated against, and it should have an equal chance to be reported as the samples with lower RTT.

Our solution is to use a multi-stage hash table data structure that performs “lazy garbage collection”, by assigning an expiration time for each record and overwriting expired records only upon hash collisions. For each outgoing packet, we record a fingerprint (a hash of 5-tuple flow ID and expected ACK number) and a timestamp in the hash table. The records matching with incoming packets produces RTT samples and are deleted, while those never matched with incoming packets are considered expired based on their timestamps, and are overwritten when hash collisions occur. When the data structure runs out of memory, it randomly rejects new records, thus automatically achieves unbiased sub-sampling.

We have implemented our algorithm on a commodity programmable switch using the P4 language [21]. We are in the process of deploying it in our local campus network. Our deployment will provide researchers with valuable measurement data about RTTs “in the wild,” while also giving the local network operators a useful tool for diagnosing end-user performance problems in real-time.

The remainder of this paper is structured as follows. Section 2 introduces our RTT measurement algorithm based on multi-stage hash tables, as well as some considerations in measuring real-world TCP flows. In Section 3, we evaluate our algorithm for its accuracy and resource requirements. Section 4 discusses some related work on RTT monitoring, and we conclude the paper in Section 5.

2 MEASURING RTT IN THE DATA PLANE

In this section, we present our data-plane RTT monitoring technique using a multi-stage hash table data structure.

2.1 Overview of Measuring TCP RTT

A TCP connection carries bi-directional data streams between two end hosts. In our application scenario, one end host resides in our local network and the other is a remote host, similar to [1]. At the vantage point, we can see both incoming and outgoing TCP packets, and observe TCP sequence (SEQ) and acknowledgment (ACK) numbers. In particular, each outgoing TCP packet with non-zero payload may be acknowledged by a future ACK number sent from the remote host. We can then infer the round-trip time from the vantage point to the remote host using the time difference between the two packets. Note that we only consider the Internet leg of the RTT and ignore the local leg from our vantage point to the local host, which we consider negligible for a local ISP.

Thus, at our vantage point, we do the following:

1. For each outgoing TCP packet with a unique expected ACK number (eACK), we record its flow ID (IP address pair and port pair), eACK (calculated using the SEQ number plus the payload size), and a timestamp. Non-handshake packets that have no payload are not recorded.

2. For each incoming TCP packet, we look up our records using its flow ID and ACK number. If we find a match, we subtract the current time with the recorded outgoing timestamp to recover an RTT sample from this packet.

2.2 Lazily Expiring Records

In reality, many TCP packets do not receive a corresponding ACK for various reasons; for example, a remote host using the TCP “delayed ACK” mechanism may only send one ACK for every two consecutive data packets. A strawman solution that removes records only when they are matched will soon find its memory filled up by stale records. To efficiently use the limited memory space, we need to clean up the records for the unmatched packets. Yet, if we clean the records too aggressively, ACKs arriving long after the data packets may fail to match their records, so we cannot produce RTT samples for high RTTs.

We set an expiration threshold for all records: a record that was not matched after a predetermined interval $T_{\text{Expire}}$ will be considered stale and get evicted. Fortunately, as a record includes a timestamp already, we do not need any extra memory to implement this expiration mechanism. This threshold is set to be reasonably larger than the RTT samples observed in a network to avoid prematurely removing records. For example, in Section 3 we set this to 500 ms as it corresponds to the 99\textsuperscript{th}-percentile of the RTT samples observed in our network, so we rarely under-sample high RTTs.

When we set $T_{\text{Expire}}$ too small, an outgoing packet’s record may get overwritten before the incoming packet can match it. When $T_{\text{Expire}}$ is too large, the algorithm’s memory fills up with useless records, preventing the tracking of new packets that would produce RTT samples later. Both cases result in missing a lot of RTT samples unnecessarily. When $T_{\text{Expire}}$ is set appropriately, the algorithm uses its memory efficiently to store records and is not under memory pressure.

It is, however, expensive to track all the records and actively remove a record from the data structure once it expires. Therefore, we lazily expire such records: if a record’s timestamp becomes too old, it is overwritten by a future attempted insertion into the same
We first compute its expected acknowledgment number
works that utilize hash table-based data structure to aggregate data,
+  
\( h(2) \)
+  
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
\( h \)
was found across all 4 stages, no RTT sample is produced for this incoming packet.

In Section 3, we evaluate the effect of changing the number of tables and their sizes on the algorithm’s success rate.

2.4 Analyzing and Reporting

In our prototype implementation, once we obtain an RTT sample, we report it alongside the packet’s flow ID to the switch control plane. It is also possible to aggregate the RTT samples based on destinations (defined by IP blocks or domain names), and subsequently calculate the RTT distribution for each remote destination in the data plane in real-time. The switch can report only the aggregated statistics for each IP block, saving the throughput required for reporting each and every RTT sample.

Network operators may specify an expected RTT distribution for a particular destination, and get notified when the distribution deviates significantly. This allows timely detection of SLA violations or BGP hijacks. Detecting these violations in the data plane opens up the possibility of the switch taking immediate action, e.g., to re-route the traffic to a faster or more secure alternative path. We leave this as future work.

2.5 Challenges with Measuring RTT

Bidirectional traffic. Our vantage point must see both directions of the data stream. This assumption is true for most local ISPs, however at Internet scale the outgoing and incoming traffic may traverse different paths due to asymmetric routing.

Outgoing traffic. Each RTT sample requires a unique outgoing SEQ number, thus the outgoing packets cannot have zero payload length. Therefore, we need some amount of data sent in the outgoing direction: a pure incoming TCP flow, such as downloading a large file from a remote server, does not produce RTT samples beyond the initial handshake.

However, we should note that many modern user applications like web apps, video streaming, etc., include two-way traffic for tracking or control purpose. In particular, web-based video playback (such as Netflix) are often chunk-based, with the browser requesting 5-second or 15-second chunks periodically, thus we can expect outgoing data (and hence RTT samples) every 5 or 15 seconds.

Delayed ACK. Delayed ACK is an optimization used by some TCP implementations to combine an acknowledgment packet with response traffic. By not immediately sending back an ACK packet for incoming data, the host has an opportunity to piggyback future response data with this acknowledgment. When there is no response to send, a delayed ACK timer will timeout, usually after 50 ms, and an ACK packet with no piggybacked data will be sent. The hosts also immediately send out the ACK after receiving two consecutive full-sized packets. A packet receiving a delayed ACK produces an artificially higher RTT sample since it includes the delay timeout.

To avoid producing biased RTT samples, we need to filter packets that experience a delayed ACK. Rather than track the TCP state machine for each flow, we use a very simple heuristic: the full-sized packets typically do not suffer from delayed ACKs, as end hosts are not allowed to delay ACKs when receiving two consecutive full-sized packets. To further ease implementation, we avoid tracking Maximum Transmission Unit (MTU) or TCP’s negotiated Maximum Segment Size (MSS) for each flow, but rather assume a packet is full-sized if its length is one of several commonly used MTUs (e.g. 1440, 1500, etc.); the user can choose to only report RTT samples produced by outgoing packets with these sizes.

Stretch ACK. Due to delayed ACK, we expect observing one incoming ACK for every two data packets. Yet, the server may send even fewer ACK packets, and this practice is referred to as Stretch ACK [17]. As we discussed earlier, packets not receiving corresponding ACKs create stale hash-table entries, which are automatically removed when they expire. Meanwhile, the effect of Stretch ACKs on RTT measurement accuracy has been studied previously by [9].

Selective ACK and retransmissions. When a packet is lost, TCP will re-transmit the packet after seeing duplicated ACKs; we may observe two identical outgoing packets in this case. If there are packets with larger SEQ numbers already delivered, the acknowledgment for the re-transmitted packet will directly jump to a much later ACK number than its eACK. The re-transmitted packet will not produce an incorrect RTT sample, however the resulting ACK packet may produce an inflated RTT sample, as it could be matched with an earlier data packet. TCP implementations may also send Selective ACK (SACK) upon packet drops to acknowledge subsequent packets; the SACK packet will share the same ACK number as an earlier normal ACK packet, which would have erased the matching record. Thus, our algorithm will not produce an incorrect RTT sample for these SACK packets.

Sampling under memory pressure. In Section 3, we show that our prototype tracks >99% of RTT samples using a moderate amount of data-plane resources. However, if the monitored link rate grows faster and average RTT grows higher, our algorithm needs more memory to save in-flight records and achieve adequate accuracy. Also, data-plane memory may be shared among other measurement applications running in the data plane, further limiting the memory available for RTT measurement. When memory is insufficient, records for new outgoing packets cannot be inserted into the data structure, which is filled up by unmatched and unexpired records. However, since records are naturally expiring and the location for insertion is pseudo-random (determined by hash functions), some records will be inserted successfully when their randomly chosen location aligns with a just-expired record. In effect, a random fraction of outgoing packets are automatically sampled, and the algorithm produces an unbiased sample subset of RTT measurements.

As an alternative to the packet-level sampling, we can also randomly sample a small fractions of flows or IP addresses, and only insert their records into the hash table data structure. This way, we can accurately measure the RTT distribution for the fraction of flows or IP addresses sampled.

Security. As a proof of concept, our current data plane implementation uses simple hash functions and fixed expiration threshold. An attacker who can inject traffic into the measured network may manipulate the outgoing and incoming packets, to either deliberately cause hash collisions and evade RTT measurement, or maximize the algorithm’s memory consumption by controlling the RTT to be just below the fixed threshold. To defend against such adversarial traffic, we should use a more secure hash function implementation in the
We implement our RTT measurement algorithm and multi-stage hash table data structure on a Barefoot Tofino programmable switch using P4 languages [21]. Our implementation has approximately 600 lines of code, and is open-sourced on GitHub.

We use different variants of the CRC16 function (with different polynomials) to calculate indices in hash tables, and use the CRC32 function to calculate packet fingerprints. Since each record consists of a 32-bit nanosecond-precision timestamp and a 32-bit fingerprint hash, a 3+8 tables, 64k records-per-table configuration uses 3+1=9 hash function computations and 8 $\times$ 2 $\times$ 32bit $\times$ 64k=$4096$KB of data-plane memory, both less than 50% of total capacity. To verify our algorithm can report RTT samples under a realistic workload, we collected a bi-directional traffic trace from a vantage point in a university campus network, which is also a future deployment site of the algorithm. We subsequently use the trace to evaluate the effectiveness of our RTT monitoring algorithm using the simulator, under various table sizes and number of stages.

### 3.1 Dataset and Method

We captured a bi-directional traffic trace from a 10 Gbps peering link between a border router of a university campus network and a local ISP. The traffic trace has been anonymized and sanitized to obfuscate personal data before being used by researchers, and our research has been approved by the university’s institutional review board.

The trace contains 1 million TCP packets and lasts 1.10 second. It contains 11,085 unique TCP flows, with a mean and median IP packet size of 1100 and 1500 bytes, respectively; about 58% of packets are likely MTU-sized (longer than 1450 bytes).

After tagging packets as incoming or outgoing based on IP prefix, we calculated the ground truth RTT samples by matching TCP sequence and acknowledgment numbers. The trace contains 0.6 million outgoing packets, 0.4 million incoming packets, and 71K pairs of RTT samples. The median RTT for all samples is 44 ms. We plot the RTT distribution we observed in the trace in Figure 4. As a back-of-envelope calculation, reporting all 71K RTT samples to the control plane in real-time (using 32-bit timestamps) requires 2.07 Mbps of additional throughput; reporting the RTT samples in lieu with the 12-byte flow ID requires 8.28 Mbps, a moderate fraction compared with the 10 Gbps line rate. We expect the throughput required for processing samples to grow proportionally when we scale up the measurement effort at a vantage point with multiple 100 Gbps links.

In the following experiments, we use 500ms as the stale threshold (corresponding to 99$^{th}$ percentile of all RTT samples), and investigate our algorithm’s success rate under various table size configurations. The success rate is determined by how many incoming packets are matched with a record (out of those having ground truth RTTs).

---


---
We now investigate the optimal number of tables to use in the multi-active/host-based measurements. Several works have explored previous work shows the presence of significant variability in RTTs in TCP handshake packets at ISP vantage points. Yet Ruru does not measure RTTs for subsequent packets in long-running TCP connections. Veal et.al. [25] proposed a method to measure RTT beyond going TCP packets with corresponding acknowledgments. This pre-processing scheme by deleting expired entries upon hash collision, which addresses the issue caused by the delayed ACK mechanism. To the best of our knowledge, we are the first to implement a multi-stage hash table data structure for computing RTT in the data plane. Also, our approach is unique in the sense that entries in the hash table are automatically expired and lazily cleaned.

**4 RELATED WORK**

Active/host-based measurements. Several works have explored measuring RTT from end-hosts. PingMesh [12] and NetBouncer [20] monitors the health of data center networks, including RTT, by using end hosts (or VM hypervisors) as vantage points to send and receive probe packets. However, unlike data center networks, ISPs do not have control over end-hosts, thus cannot run agents on them. Furthermore, active measurement tools add extra probing traffic into the network.

**RTT measurement at ISP vantage points.** Aikat et.al. [1] measured the RTT experienced by campus network users by capturing traffic at a border link and later analyzing the traffic to match outgoing TCP packets with corresponding acknowledgments. This previous work shows the presence of significant variability in RTTs in a TCP connection, motivating our work to monitor this variability in real-time. Ruru [7] is a system that passively measures the RTT of TCP handshake packets at ISP vantage points. Yet Ruru does not measure RTTs for subsequent packets in long-running TCP connections. Veal et.al. [25] proposed a method to measure RTT beyond handshakes at an intermediate vantage point. However, it requires a modification to a TCP packet to add a timestamp option. It also depends on the recipient host to echo this timestamp.

**Measuring RTT on a programmable switch.** Dapper [11] is a TCP monitoring tool that tracks various metrics, including RTT, in the data plane. Dapper produces accurate measurements for the tracked flows, but it can only track a single outgoing packet per congestion window for RTT measurement, and must wait until that packet’s acknowledgment arrives before recording another outgoing packet for the flow. Our algorithm does not limit the number of outgoing packet records stored for each flow; a flow can produce as many RTT samples as possible as long as the memory space permits.

**Hash table data structure.** Our multi-stage hash table data structure is motivated by prior works on data-plane algorithms for programmable switches. Count-Min Sketch [6] is a data structure made of several hash-indexed counter arrays, and is often used for estimating flow sizes and detecting heavy hitters. HashPipe [18] and PRECISION [2] designed more sophisticated multi-stage hash tables for heavy-hitter detection, with each record storing a hashed flow ID and a counter. Such data structures, however, are not suitable for continuous RTT monitoring as they are. An RTT sample is produced only when a corresponding acknowledgment arrives, thus a naive implementation of such data structure quickly gets populated by useless entries. Our multi-stage hash table implements an efficient garbage collection scheme by deleting expired entries upon hash collision, which addresses the issue caused by the delayed ACK mechanism. To the best of our knowledge, we are the first to implement a multi-stage hash table data structure for computing RTT in the data plane. Also, our approach is unique in the sense that entries in the hash table are automatically expired and lazily cleaned.

**Non-TCP traffic.** Google proposed QUIC [5], a UDP-based transport alternative to TCP. QUIC encrypts its packet header fields, which prevents an ISP from performing RTT measurement based on SEQ/ACK matching. The QUIC standardization body is planning to add a “spin bit” [23] specifically for RTT measurement at ISP vantage points.

**5 CONCLUSION**

We present an algorithm to track the per-packet Round-Trip Time of TCP traffic in a commodity programmable switch using a multi-stage hash table data structure. Our algorithm successfully reports over 99% of all RTT samples, in a traffic trace collected from a 10 Gbps peering link of a campus network. Our evaluation also shows that using three to four stages in our hash table structure achieves the best performance for RTT monitoring, given the same amount of total memory. We are currently deploying continuous RTT monitoring on our university campus network. For future work, we plan to integrate real-time RTT samples with anomaly detection and other routing change detection techniques.

**6 ACKNOWLEDGMENTS**

This research is supported by NSF Awards CNS-1704077. We sincerely thank the anonymous reviewers, as well as David Walker, Liang Wang, Shir Landau Feibish, Robert MacDavid, Mary Hogan, and Ross Teixeira, for their helpful comments and feedback.
REFERENCES


