TCP Rapid: From Theory to Practice

Qianwen Yin Jasleen Kaur F. Donelson Smith
University of North Carolina at Chapel Hill

***Abstract*—Delay and rate-based alternatives to TCP congestion-control have been around for nearly three decades and have seen a recent surge in interest. However, such designs have faced significant resistance in being deployed on a wide-scale across the Internet—this has been mostly due to serious concerns about noise in delay measurements, pacing inter-packet gaps, and/or required changes to the standard TCP stack/headers. With the advent of high-speed networking, some of these concerns become even more significant.**

In this paper, we consider Rapid, a recent proposal for ultra-high speed congestion control, which perhaps stretches each of these challenges to the greatest extent. Rapid adopts a framework of continuous fine-scale bandwidth probing, which requires a potentially different and finely-controlled gap for every packet, high-precision timestamping of received packets, and reliance on fine-scale changes in inter-packet gaps. While simulation-based evaluations of Rapid show that it has outstanding performance gains along several important dimensions, these will not translate to the real-world unless the above challenges are addressed.

We design a Linux implementation of Rapid after carefully considering each of these challenges. Our evaluations on a 10Gbps testbed confirm that the implementation can indeed achieve the claimed performance gains, and that it would not have been possible unless each of the above challenges was addressed.

I. Introduction

Delay and rate-based alternatives to TCP congestion-control have seen a significant surge in interest [1], [2], [3], [4], [5], [6], [7], [8]—indeed, their performance gains (observed both in simulations and simple testbeds) seem quite promising, often exceeding TCP performance by several orders of magnitude. However, such designs have also faced significant reluctance in being adopted or deployed on a wide-scale across the Internet—this is primarily because the promised gains observed under controlled and simulated settings are not trusted to translate well to real-world settings.

Why? Firstly, because most of the alternatives rely on measurement of metrics such as end-to-end delay or available bandwidth as a measure of congestion, rather than on packet loss—these metrics can be fairly volatile, and their measurement can be quite prone to fine-scale buffering noise [9]. This is especially true in high-speed network environments. Secondly, several protocols rely on fine-scale pacing of inter-packet gaps (IPG), which are challenging to control predictably in interrupt-driven operating systems, especially at high speeds. Thirdly, stepping away from a conventional congestion-control framework that has been used and perfected for more than three decades, is resistance-worthy—why would an operator of production servers trust a new prototype? Given the promise of such congestion-control alternatives, especially in high-speed networks, it is important to address these challenges and open the real-world to their adoption.

Of all the proposed alternatives to the legacy TCP, RAPID [7] perhaps stretches the above challenges to the greatest extent. In simulations, this protocol shows outstanding gains in terms of scalability, adaptability, TCP-friendliness, and fairness. However, as described in Sections II-III, RAPID needs to create μs-precision inter-packet gaps for *all* data packets sent out. Further, it relies on observing fine-scale changes in inter-packet gaps for estimating end-to-end available bandwidth, which is fairly sensitive to the presence of fine-scale buffering noise. Finally, RAPID relies on a "gap-clocked" transmission of data packets, which is a significant departure from the conventional "ack-clocked" TCP framework. In this paper, we ask (in the context of TCP RAPID): *can these challenges be addressed in order to realize ultra-high speed real-world prototypes that perform as well as the promise delivered by simulations?* If the answer is a yes for a protocol as demanding as RAPID, then this would be a significant enabler for the practical adoption of delay, rate, and bandwidth-based protocol research.

Our Innovations This paper presents the following innovations:

- We tailor the state-of-the-art for creating inter-packet gaps in the Linux kernel, and show that it achieves μs accuracy at ultra-high speeds.[1]
- We adapt the state-of-the-art denoising schemes for alleviating the impact of fine-scale noise, and achieve robust bandwidth estimation for RAPID.
- We propose and evaluate the decoupling of the probing and adapting timescales used in RAPID congestion control, for alleviating the trade-off between responsiveness and stability in the presence of volatile available bandwidth.
- We design all the components of RAPID as pluggable kernel modules, which can be loaded/unloaded on the fly, without bringing down production servers. These work with standard TCP headers and the socket API.
- We evaluate the implementation design on 10/40Gbps testbeds in the presence of representative and bursty cross-traffic, and show that it lives up to the simulation-promised performance. Furthermore, we show that this performance can not be achieved without each of the above innovations.

In the rest of this paper, we summarize RAPID in Section II and identify challenges in Section III. We present our design

This work was supported in part by the National Science Foundation under Awards CNS–0347814, CNS–1018596, and OCI–1127413.

[1] In this paper, "ultra-high speed" refers to 10 Gbps and more.

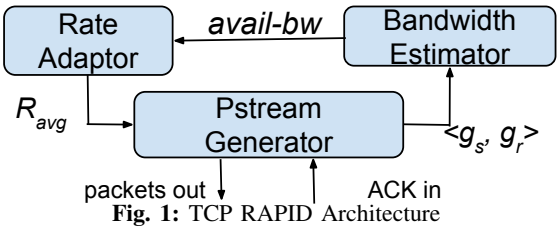


Fig. 1: TCP RAPID Architecture

of a Linux implementation in Section IV and our evaluations in Sections V-VI. We conclude in Section VII.

II. Background – TCP RAPID

Instead of simply relying on packet loss as congestion feedback, RAPID continuously estimates the end-to-end avail-bw, and uses the estimates to send packets out in logical groups referred to as p-streams. The transmission of p-streams is rate-based (and not ack-clocked), and is managed by three components operating in a closed loop (Fig 1). Given an average send rate R_{avg} informed by the rate adapter, the p-stream generator sends packet out in units of p-streams. Once the ACKs for a full p-stream return, the bandwidth estimator calculates the end-to-end avail-bw, based on which the rate adapter updates the sending rate for the next p-stream.

Pstream Generator RAPID uses each data packet sent for probing the end-to-end path for some rate R, by controlling the send-gap (g_s) from the previous packet sent as: $g_s = \frac{P}{R}$. Within a p-stream, packets are sent at N_r different exponentially-increasing rates (with N_p packets sent at each rate): $R_i = R_{i-1} \times s, i \in [2, N_r], s > 1$. Fig 2a plots the gaps in a p-stream with $N_r = 4$, and $N_p = 16$. The average send rate of the full p-stream is set to R_{avg}, informed by the rate adapter.

Bandwidth Estimator The RAPID receiver records the arrival time of data packets and sends back the timestamps in ACKs. Once the sender receives all ACKs for a p-stream, it extracts these timestamps, computes the receive gaps (g_r), and feeds them to the bandwidth estimator.

The send and receive gaps, g_s and g_r are used to compute an estimate for the end-to-end avail-bw ($ABest$), based on the principle of self-induced congestion—for the i-th packet in a pstream, $g_r^i > g_s^i$ indicates that it experienced queuing at the bottleneck (respective probing rate was higher than the avail-bw). $ABest$ is computed as the largest probing rate *beyond* which packets *consistently* experience bottleneck queuing (e.g. the second probing rate in Fig 2a).

Rate Adapter R_{avg} is initialized to 100Kbps. Thereafter, every time $ABest$ is updated, the average sending rate of the next p-stream is also updated by applying a conditional low-pass filter as:

$$R_{avg} = \begin{cases} R_{avg} + \frac{l}{\tau} \times (ABest - R_{avg}), & ABest \geq R_{avg} \\ R_{avg} - \frac{1}{\eta} \times (R_{avg} - ABest), & ABest < R_{avg} \end{cases}$$

where l is the duration of the p-stream , and τ and η are constants. The effect of the above filter is that it takes about τ *time units* for R_{avg} to converge to an increased avail-bw, and η p-streams to converge to a reduced avail-bw.

Note that at p-stream timescale, packets are sent at an R_{avg} no higher than the network can currently handle. However, the exponentially-spaced p-streams are able to simultaneously probe for rates both higher and lower than R_{avg} at smaller timescales —this gives the protocol excellent agility in the presence of dynamic cross-traffic. In fact, simulation-based evaluations in [10] show that the protocol has close-to-optimal performance along several dimensions, most notable of which are: (i) discovering and adapting quickly to changes in avail-bw (due to continuous probing at sub-pstream timescales); (ii) negligible impact on co-existing TCP traffic (due to an extremely low queuing footprint); and (iii) RTT-fairness, with no bias against long-RTT transfers (due to shedding of ack-clocking).

III. TCP RAPID: Challenges in Practice

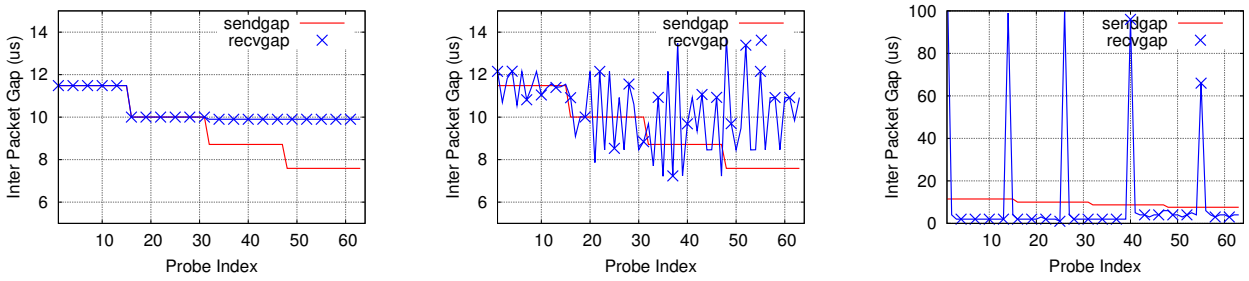
All of the performance gains reported in [10] have been observed solely in the NS-2 simulator environment. Three types of challenges can be identified in realizing the same performance in the real world.

A. Fine-scale Inter-packet Gap Creation

Challenge RAPID requires the TCP sender to send packets out with *high-precision* and *fine-grained* inter-packet gaps (IPG) for the purpose of bandwidth estimation. For instance, in order to probe for an avail-bw of 10 Gbps with even jumbo-sized frames, packets have to be sent with spacing as small as a few microseconds—and this value reduces proportionally as we consider higher network speeds. A spacing inaccuracy of even 1-2μs, can lead to a bandwidth-estimation error of 50%! Several other protocol proposals rely on fine-scale IPG creation and face a similar challenge in scaling up to ultra-high speeds [11], [12], [3]—the confounding aspect of RAPID is that it uses *every* packet for fine-scale probing, whereas these others rely on bandwidth probing only intermittently.

Sate of the art Existing bandwidth estimators [13], [14], [15] and transport protocols [11], [12] create gaps for bandwidth probing, by staying in a busy-waiting loop (often in user space) until the desired time gap elapses. Unfortunately, ensuring the fine-grained and high-precision IPGs needed for ultra-high speeds can be fairly challenging in current software-based end-systems, mainly for two reasons. First, most operating systems are interrupt-driven, and the process sending out packets of a p-stream may lose control of the CPU at any time while "waiting" for the time-gap between two consecutive packets. The resultant send gaps are unpredictable, and lack high precision.

Second, before those packets get transmitted by the NIC, they can get buffered at several places—in the protocol layer buffers as they are being handed down the kernel protocol stack, at the NIC interface queue when the kernel directs them to the corresponding NIC, and on the NIC outgoing queue. Such buffering can completely destroy the intended inter-packet gaps. Some of the upper-layer buffering can be avoided by using in-kernel support that relies on software timers (e.g., **qdisc_watchdog_timer** used in Linux FQ scheduler and



(a) Noise free **(b)** After bottleneck **(c)** At receiver

Fig. 2: Probe streams $N_r = 4, N_p = 16$

tasklets used in TRC-TCP [16]). However, such interrupt-driven transmissions will lead to increased overhead of context switch and consequently will slow down the system when multiple high-speed flows coexist [17]. Besides, it can not prevent packet buffering at the sending host.

Goal Our first objective in this paper is to address this challenge and: *consider and evaluate high-speed techniques that enable fine-grained gap creation with high precision.*

B. Noise Removal from Receive Gaps

Challenge Bandwidth estimation is key to RAPID, which relies on the assumption that any fine-scale changes in inter-packet gaps are indicative of the bottleneck avail-bw, and can be used to robustly estimate it. However, even when p-streams are sent out with accurate spacing, there are two types of noise sources that can challenge this assumption:

- *Burstiness in cross-traffic at bottleneck resources:* In a packet-switched network, traffic arrival can be fairly bursty at short timescales [18]. As illustrated in Fig 2b, frequent arrival of short-scale traffic bursts introduces noise in the persistent queuing signature of Fig 2a, and can lead to low estimates of avail-bw.
- *Transient queuing at non-bottleneck resources:* Even a non-bottleneck resource can induce short-scale transient queues when it is temporarily unavailable while servicing competing traffic. This can happen, for instance, while accessing high-speed cross-connects at the switches, or while waiting for CPU processing after packets arrive at the receiver-side NIC. In fact, *interrupt coalescence* can force packets to wait at the NIC before being handed to the OS for timestamping, even if the CPU is available — this can introduce noise worth *hundreds* of microseconds in inter-packet gaps [19]. Fig 2c plots the receive gaps observed when packets are delivered by the receiver NIC to the operating system — the inter-packet gap signatures are unrecognizable after interrupt coalescence.

It is important to note that such noise impacts not only other protocols that rely on bandwidth estimation, but also a myriad of protocols that rely on delay-measurements, with the performance of the former being more sensitive to noise [20], [11], [3], [21].

State of the art While heuristics for smoothing out noise have been designed for bandwidth estimators (e.g., [22]), most need fairly long p-streams in order to scale to 10 Gbps and beyond [19] — the longer the p-streams, the less are the performance gains of a protocol like RAPID. The recently-proposed denoising technique, BASS [19], aims for shorter p-streams—it can help estimate bandwidth on 10 Gbps paths with less than 10-15% error, using 96-packet p-streams[2]. While BASS is a promising technique, it has been mostly evaluated for bandwidth estimators — p-streams are sent far apart and assumed to be independent, and the average rate of each p-stream is not influenced by current $ABest$. Both of these aspects do not hold within a congestion-control protocol like RAPID.

Goal Our second objective in this paper is to: *consider and evaluate such highspeed techniques for reducing the impact of fine-scale noise in inter-packet gaps.*

C. Alleviating the Stability/Adaptability Trade-off

Challenge Noise is a particularly significant concern for delay and bandwidth-based transport protocols due to the finer time scales at which their respective congestion metrics are probed for. The larger the timescales at which these protocols choose to probe the network path, the smoother (and less impacted by small-scale noise) their measurements will be. However, the resulting protocol will be less quick in responding to changes in network conditions. This trade-off between stability and adaptability should be carefully balanced.

RAPID uses the same timescale (given by the length of a p-stream) for *probing* for end-to-end avail-bw, as well as for *adapting* the data sending rate to changes in avail-bw. Longer p-streams allow the protocol to react to avail-bw changes only at timescales at which the avail-bw is stabler and less noisy, however, shorter p-streams allow the protocol to sample avail-bw more frequently and track it more closely.

Goal Our third goal is to: *alleviate the trade-off between stability and adaptability by tracking changes in avail-bw closely, while only adapting to it at stabler timescales.*

D. Deployability Within the TCP Stack

Challenge TCP is the dominant transport protocol used by most applications. In order to achieve widespread impact and allow applications and network edge devices to work seamlessly, RAPID should be implemented such that: (i) it works with existing TCP protocol headers and the socket API; and (ii) it is a pluggable module within widely-deployed TCP stacks— in the context of servers, this caters to the requirement

[2] In [23], a machine-learning approach is used to denoise inter-packet gaps—however, no kernel-friendly implementation exists.

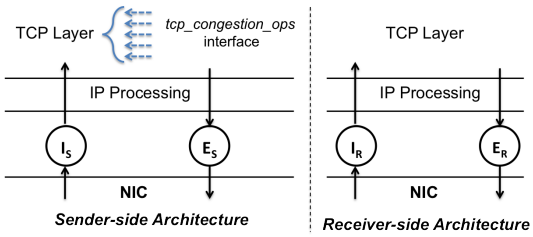


Fig. 3: Architecture of RAPID Implementation

of system administrators that the protocol implementation can be loaded (or unloaded) on the fly without bringing down a production server.

In widely-deployed TCP stacks, sending of data packets is ACK-clocked and window-controlled—pluggable congestion-control modules are supported in operating systems like Linux, that allow changes to the amount by which window growth occurs when ACKs are received (how many packets are eligible to be sent out). However, these do not allow changes to *when* a packet gets sent out. In contrast, the sending of packets in RAPID is "gap-clocked", in which the IPG (and not ACK arrival) determines when the next packet should be sent out. In fact, the receiving of ACKs and sending of data packets are completely asynchronous of each other. Clearly, supporting gap-clocking needs modules beyond the TCP congestion-control framework.[3]

In addition, the RAPID receiver needs to observe in high-precision, and communicate back to the sender, the gaps between packets it receives. Delay-based protocols TCP-LP [8] and LEDBAT [24] rely on TCP timestamping option for computing one-way delays. However, this option carries only milli-second resolution timestamps—timestamping with micro-second precision, which is needed for high-speed networking, is not available. Furthermore, timestamps are produced when ACKs are generated, and not when the respective data segments are received (subjecting gaps to variable ACK processing times). Fig 5 plots the time difference between the actual arrival of a data packet and the time recorded in TCP timestamps (as observed on our 10 Gbps testbed)—we find that the two can differ by up to 60μs.

Goal Our third goal is to: *consider mechanisms that enable RAPID to be loaded as a pluggable module on widely-deployed TCP stacks, while supporting its high-precision and fine-scale gap-clocking and timestamping requirements.*

IV. Our Implementation Design

We present our design of a RAPID implementation for addressing the challenges identified above.

A. Realizing Gap-clocking in a Standard TCP Stack

Current Linux kernel provides a congestion handler interface, **tcp_congestion_ops**, which allows different congestion control algorithms to be implemented in a pluggable manner. Implementing RAPID boils down to: *removing the dependency of packet transmissions on ACK arrival* and *scheduling these based on intended IPGs*. We elaborate how both of these can be achieved using the above interface and loadable modules.

[3]Other transport protocols that rely on bandwidth estimation, even intermittently, require gap-clocking as well.

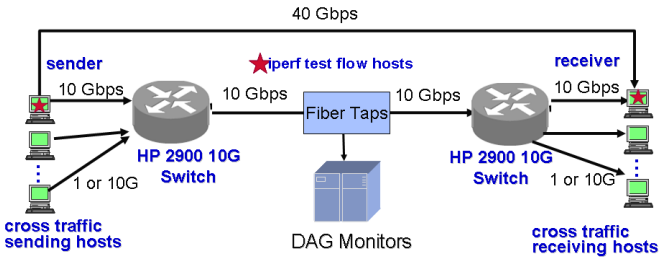


Fig. 4: Test-bed Topology

1) Removing ack-clocking: In RAPID, steady-state packet transmissions in large transfers are not triggered by the arrival of ACKs. ACK-clocking can be turned off with relative ease by simply using the tcp_congestion_ops interface to fix *cwnd* to a value much larger than the bandwidth delay product.[4] As a result of doing this, the TCP layer would send segments down for lower layer processing as soon as data is made available by the application.

2) Incorporating gap-clocking: The tcp_congestion_ops interface does not allow control over *when* a segment is sent by TCP. For scheduling packet transmission, we instead create a new Linux Qdisc module (E_s in Fig 3) which is attached to a given network interface. Link-layer frames containing TCP segments are processed by E_s before being delivered to the NIC device driver. E_s responds to two calls—**enqueue** and **dequeue**. It maintains a FIFO queue for each TCP connection for buffering packets received from **enqueue**. Upon each **dequeue**, it chooses the next packet that will be transmitted.

E_s is responsible for enforcing IPGs within each TCP connection. It (i) groups packets into units of p-streams; (ii) computes per-packet gaps according to R_{avg} (using a data structure shared with the TCP layer); (iii) assigns each packet an intended transmission time (t_s) according to the computed gaps; (iv) schedules the departure of the head of queue at its appointed t_s (mechanisms discussed below).

3) Interleaving packets from multiple flows: To schedule packets in the order of their t_s from multiple RAPID connections that use the same NIC, we maintain a minimum heap data structure—the elements of the heap are the packets at the *head* (with the earliest t_s) of each per-connection queue. The **dequeue** function in E_s removes and sends the top packet of the heap to the NIC **only if** its intended t_s time has passed.

With the presence of multiple RAPID flows, it may not be possible to respect t_s for every packet of every flow (for instance, when the transmission time of two packets from different flows are nearly the same). It is important to realize that the resultant short-scale queuing is expected by the bandwidth estimator. Indeed, the sender outbound link represents the first shared link for those flows—the intended send gaps control only the times at which the packets within each flow arrive at the NIC, and not when they depart the NIC.

B. Creating Accurate Inter-packet Gaps

Several research projects have relied on hardware support for fine-scale control of inter-packet gaps—for instance, the Comet-TCP [16] protocol stack is fully implemented on a

[4]In our 10Gbps experiments, we set *cwnd* to 16000.

programmable NIC to create gaps with sub-nano precision. However, the requirement of such specialized hardware seriously limits the deployability of a new transport protocol, which is one of our prime goals.

[25] employs a novel approach for fine-scale control of inter-packet gaps—it inserts appropriately-sized Ethernet PAUSE frames to occupy the desired gap between two TCP data packets. These special control frames are specified as part of the IEEE 802.3x for flow control between two ends of a link. They are discarded by a receiving NIC and thus consume bandwidth only on the first link (typically, from the sender to the first switch) on the path. As a result, the intended gaps are preserved between successive TCP packets arriving at the first outbound queue. [25] uses this Ethernet feature for implementing paced TCP (constant gaps within a given flow).

Inspired by this approach, we design E_s to send PAUSE frames and data packets in an interleaved back-to-back manner to the outgoing NIC, to be transmitted at line speed. For time-keeping and fine-scale gap-control, E_s relies on a $link\_clock$ (instead of the kernel clock), which tracks the transmission time that would be consumed by all outbound packets that have been sent so far to the NIC. The intended send time t_s assigned to each packet, is also compared to the $link\_clock$ (and not the kernel clock). Once **dequeue** is called, E_s checks whether the send time of the packets at the top of our heap is less than or equal to the current $link\_clock$. If true, that packet is dequeued and sent to the NIC; otherwise, E_s creates and sends a PAUSE frame of size $(t_s - link\_clock) \times C$, where C is the link capacity, and t_s corresponds to the packet at the top of the heap. $link\_clock$ is incremented by the expected transmission time of the frame being sent to the NIC ($\frac{framesize}{C}$). Creating a PAUSE frame of the above size ensures that the next time **dequeue** is called, the packet at the top of the heap would be eligible to be sent, and would have the desired gap from the previous data packet. Thus, inter-packet gaps are achieved by finely controlling the size of PAUSE frames.

We evaluate the accuracy of gap-creation using E_s by generating a large number of p-streams, covering a wide range of probing rates from 100Mbps to 10Gbps. The actual gaps are recorded by collecting traces using the DAG monitor in Fig 4, immediately after packets traverse the 1st switch. For comparison, we generate gaps with a user-level application modified from pathChirp [13]. We also implement a Qdisc that enforces t_s by registering a software timer (with qdisc_watchdog interface) for every packet departure. Fig 6 plots the distribution of the difference between actual gaps and intended ones—E_s limits the error within 1μs, which is significantly more accurate than using software interrupts or a user-level application!

C. Timestamping Packet Arrivals

In order to record inter-packet receive gaps with μs precision, and communicate back to the sender using standard header timestamp options, we create two Qdiscs E_R and I_R at the receiver, and one I_S ingress Qdisc at the sender (see Fig 3). I_R receives packets as they are delivered by the NIC to

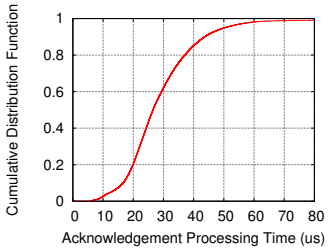


Fig. 5: ACK Processing Time

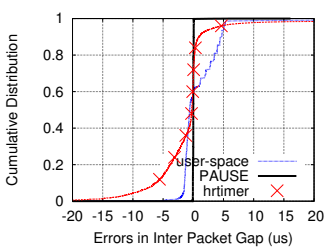


Fig. 6: IPG Creation Error

the kernel and uses $ktime\_to\_ns$ to timestamp packet arrivals with μs precision—these are recorded in a table shared with E_R. Once an ACK is generated and sent by TCP to E_R, it looks up the table for the arrival time of the corresponding packet that triggered this ACK, and substitutes it for the TCP timestamp value in header timestamp (TSval) field—TCP checksum is recomputed and updated accordingly.

When the ACK segment reaches the sender, the ingress Qdisc I_S saves the μs timestamp. To ensure correct TCP processing (which expects monotonically increasing millisecond timestamps), I_S restores TSval field with the local millisecond timestamp before handing the packet for upper-layer protocol processing. The saved μs timestamp is shared with E_S and the TCP layer—it is used by the bandwidth estimator implemented with **tcp_congestion_ops**, for computing $ABest$ once a p-stream is completely ACKed.

D. Denoising for Bandwidth Estimation

The recently-proposed Buffering-aware Spike Smoothing (BASS) technique has been shown to work well in denoising receive gaps within short multi-rate p-streams [19]. Below, we briefly summarize the technique and then evaluate it in the context of RAPID.

1) Buffering-Aware Spike Smoothing: BASS is based on the observation that even though buffering events like interrupt coalescence can completely destroy gaps for *individual* packet within p-streams (Fig 2c), the *average* receive gap within a single *complete* buffering event can still be recovered. BASS recovers this quantity by first carefully identifying boundaries of buffering events by analyzing receive gaps g_r —each "spike-up" and following dips in Fig 2c correspond to packets within the same buffering event (packets queued in the receiver NIC before generation of the next interrupt). BASS looks for sudden changes in receive gaps to detect these. After identifying buffering event boundaries, within each event BASS replaces both g_s and g_r with their respective averages. Such "spike-removal" is repeated up to three times until a robust signature of persistent queuing delay is restored in the smoothed p-stream. Fig 7 plots the BASS-smoothed gaps for the p-stream in Fig 2c—the spikes are successfully eliminated. The smoothed gaps for the p-stream are then fed into the bandwidth estimator.

2) Re-evaluating BASS: [19] evaluates BASS for several different settings of p-stream length—it shows that BASS can estimate bandwidth with less than 10-15% error, using p-streams with just 96 packets. For several reasons, however, it is necessary to re-evaluate BASS in the context of RAPID: **(i)** Unlike RAPID, [19] introduces a significant gap between

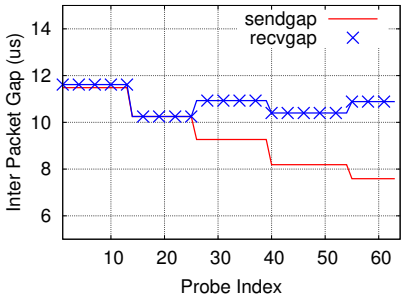


Fig. 7: BASS-denoised Gaps

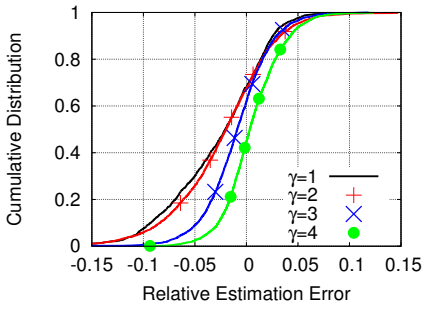


Fig. 8: Decoupling Probing/Adapting

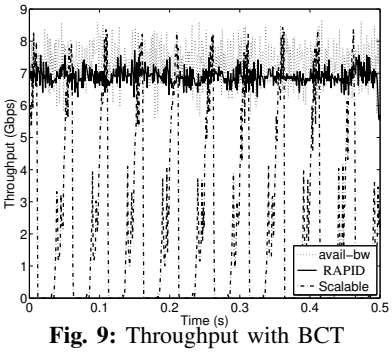


Fig. 9: Throughput with BCT

p-streams to ensure independence. RAPID, however, sends p-streams back-to-back—any queue buildup on the path caused by the previous p-stream may not drain out before the next one arrives. **(ii)** Unlike [19], in which p-streams were generated with pre-determined and controlled average send rates, the R_{avg} for a p-stream depends on the $ABest$ estimated by a recent one. In [19], in fact, p-streams were filtered out if the actual avail-bw did not fall within the range of their probing rates. **(iii)** Unlike [19], standard TCP implementations rely on delayed ACKs, and consequently, only every other packet gets a timestamp.

We incorporate BASS within the RAPID bandwidth estimator, to process receive gaps derived from ACK timestamps, before computing bandwidth estimates (in the presence of delayed ACK and p-stream-generation by the RAPID control loop). We measure the bandwidth estimation accuracy of our implementation (using the methodology in Section V) with several different choices of p-stream length, including 32, 48, 64, 96. We find that $N = 64$ achieves the best bandwidth estimation accuracy (errors less than 10% for over 80% p-streams).

E. Alleviating the Stability/Adaptability Trade-off

The stability/adaptability trade-off discussed in Section III-C is controlled by a single parameter—the p-stream length (which represents both the probing and adapting timescale). In order to alleviate this trade-off, we propose to decouple the probing and adapting timescales of RAPID. We achieve this by not requiring the sender to update R_{avg} upon each $ABest$ computation, but rather do it at a lower frequency (γ). Specifically, we adapt the R_{avg} of the transfer only once every γ p-streams, and set it to the mean of all γ $ABest$s collected since the last update.

Such decoupling naturally leads to a question: if we fix the rate-adapting timescale ($N \times \gamma$), do we get more accurate bandwidth estimates using longer p-streams (large N, $\gamma = 1$), or by using the mean bandwidth estimate of several smaller p-streams (small N, $\gamma > 1$)? To study this, we fix the rate-adapting timescale at 192 packets, and vary the probing timescale per N=48,64,96,192 ($\gamma = 4, 3, 2, 1$, respectively). Fig 8 plots the relative bandwidth estimation error observed across several p-streams sent over our 10 Gbps testbed. We find that using shorter p-streams (but the same rate-adapting timescale) reduces the estimation errors from 15% to 5%! However, when p-streams are shorter than 64 packets, the estimation errors do not reduce with larger γ—this is in agreement with our previous observation of the performance with $N = 64$. In the remaining evaluations, we adopt $N = 64$ with the rate-adapting timescale of 192 packets ($\gamma = 3$).

V. Evaluation of RAPID Implementation

In this section, we experimentally study how close our implementation gets to achieving the simulation-based performance reported in [10]. The key performance gains reported in [10] for RAPID were in terms of scaling to high-speed throughput, adapting quickly to changes in avail-bw, co-existing peacefully with low-speed TCP traffic, and inter-protocol fairness. We attempt to recreate similar experimental settings in our testbed, with some key differences: (i) We focus on contemporary ultra high-speed paths of 10/40Gbps capacity, while [10] focused mostly on 1Gbps paths, (ii) we use a bursty, representative traffic aggregate as cross-traffic for studying adaptability of RAPID, while [10] used a simple synthetic cross-traffic stream, (iii) we use shallow-buffered switches, while [10] provisions much larger buffers.

We also evaluate the Linux implementations of several protocols for comparison—New Reno, Bic, Cubic, Scalable, Highspeed, Hybla, Illinois, Vegas, Westwood, LP, Yeah and Fast.[5] For space constraints, we only present the results of Cubic and Scalable in this paper—the former is the default congestion control in Linux, the later consistently gives best link utilization.[6] Unless specified otherwise, we use the following settings for RAPID transfers: $\tau = 10ms, \eta = 3$, and $RTT = 30ms$ (representative of the medium US continental RTT).

A. Testbed Topology

The dumbbell testbed of Fig 4 consists of two HP 2900 switches with multiple 1Gbps and 10Gbps ports. The 10Gbps switch-to-switch path is used to connect two pairs of 10Gbps TCP senders and receivers. These hosts are Dell PowerEdge R720 servers with four cores on 8 logical processors running at 3.3GHz. The 10Gbps adapters on the two sending hosts are PCI Express x8 MyriCom NICs, on the two receiving hosts are PCIe Intel 82599ES NICs. The other 10 pairs of hosts

[5]Fast implementation is not publicly available. We implement it in Linux based on its Linux-emulating pluggable NS2 simulation code. With default parameters, it performs poorly in our testbed. Compound is no longer supported in recent Linux kernels.

[6]Other protocols are included in [26].

with 1Gbps NICs are used to generate cross traffic sharing the switch-to-switch link. The testbed also includes a 40Gbps direct fiber-attachment over QSPF+ ports (not a switched path) between one pair of the Dell servers. The 40Gbps adapters are PCIe x8 Mellanox NICs. All hosts in the testbed run the latest RedHat Linux 6 with 2.6.32 kernel.

For emulating path RTTs and loss properties for RAPID transfers, we use extended versions of our Qdiscs I_R and E_R on the receiver. The extended I_R drops packets randomly according to the required loss rate; E_R delays the transmission of ACKs to emulate RTT latency. For other TCP variants, we use *netem* to randomly drop packets at the sender, and to delay ACKs for RTT emulation at the receiver.

Limitations One limitation of using a real switch on our testbed is that we are unable to log or finely monitor the bottleneck queue size—instead, we must rely on indirect measures, such as packet losses and their impact on throughput. Second, our switches are fairly shallow-buffered—this implies that in all of our evaluations the bottleneck buffers are much smaller than the bandwidth-delay product (shallow buffers have been recommended widely for ultra-high speed networks [27]). Finally, the CPUs on our Dell servers are unable to keep up with 40Gbps throughput, and reach 100% utilization when the transfer rates reach 20Gbps—this is the maximum achievable throughput on the 40Gbps path.

1) Cross Traffic: We evaluate the RAPID implementation against two types of cross-traffic—responsive traffic from emulation of web users, and replayed traces of traffic aggregates (with different levels of burstiness).

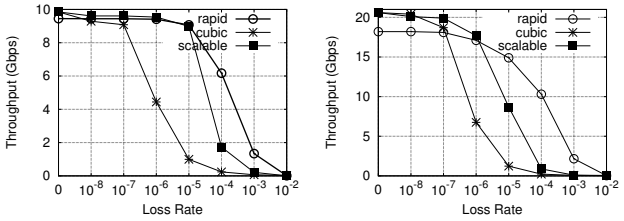
Responsive Web Traffic In order to generate bursty traffic loads on the switch-to-switch bottleneck, we use 10 pairs of hosts, each emulating thousands of web users by running a locally-modified version of the SURGE [28]—they establish "live" TCP connections with a diverse set of RTTs and inter-arrival times, and produce representative and responsive HTTP traffic. The average throughput of such traffic is 2.42Gbps. We record the flow completion times for each TCP connection.

Replayed Web Traffic Aggregates We also generate bursty, but non-responsive cross-traffic—this helps ensure repeatability and burstiness control across experiments. For this, we record a packet trace for each SURGE data source from the responsive web traffic generation above. We then replay the trace using *tcpreplay*. The average rate of the aggregated replayed traffic from the ten traffic generators is around 2.5Gbps.

To obtain cross-traffic with different levels of burstiness, we generate a smoothed version of the replayed traffic by running a token bucket Qdisc on each sending host. We also generate constant-bit-rate traffic from each sender using a UDP flow. We denote these three burstiness levels as **BCT** (the most bursty, raw replayed traffic), **SCT** (smoothed version of BCT) and **CBR** (constant bit-rate). As a measure of burstiness, the 5-95% ranges of the bit-rates (observed at a 1ms timescale) are 2.62Gbps, 1.40Gbps, and 0.49Gbps, respectively.

B. Sustained Throughput in Presence of Error-based Losses

Our first set of experiments evaluates achieved throughput in the presence of random bit error-based loss rates, ranging from



(a) 10Gbps **(b)** 40Gbps

Fig. 10: Steady-state Throughput with Error-based Losses

TABLE I: Throughput and Loss Ratio of TCP flow

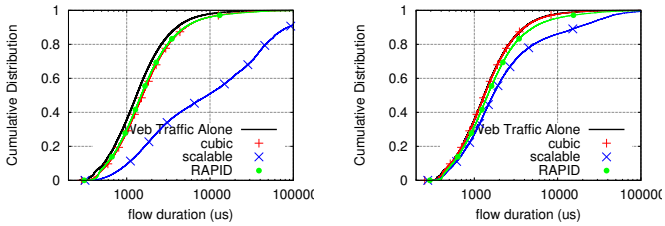
(Gbps) (%)	with replayed traffic			with responsive web traffic	
	UDP	SCT	BCT	RTT=5ms	RTT=30ms
RAPID	6.86	6.61	6.09	6.62	6.61
	0.000	0.014	0.060	0.000	0.001
cubic	3.58	3.25	2.51	6.18	2.79
	0.002	0.002	0.002	0.002	0.004
scalable	4.77	4.38	4.20	7.35	4.23
	0.072	0.046	0.040	0.036	0.037

10^{-2} (very high) to 10^{-8} (very low). Fig 10 plots the steady-state throughput achieved by each TCP protocol on 10/40Gbps paths. We find significant throughput loss when error rates exceed 10^{-6} for all protocols—however, RAPID scales much better than others and yields more than double throughput than most protocols. RAPID performs poorly with loss rate 10^{-2}—losses in practically every other p-stream prevent it from estimating avail-bw at all! The performance trends on the 40Gbps path are similar to those on 10Gbps—both agree with the scalability trends reported in [10]. For the remaining evaluations, we use only the switched 10Gbps path.

C. Adaptability to Bursty Traffic

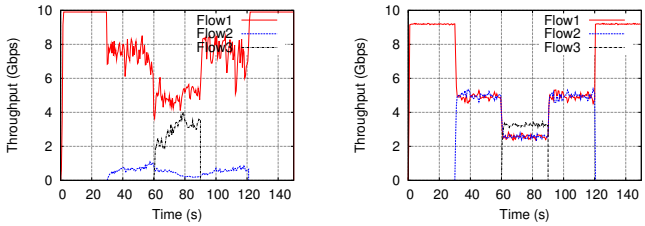
Next we evaluate the ability of the RAPID implementation to adapt to non-responsive, but bursty cross-traffic. We generate and experiment with cross-traffic with different levels of burstiness (as described in Section V-A1), and instantiate a high-speed transfer using different protocols to share the bottleneck link for 120 seconds. Table I shows the average bottleneck link utilization and loss rates observed within the high-speed transfer. We find that:

- The more bursty is the cross-traffic, the lower is the throughput (and link utilization) achieved by a high-speed transfer. This is true for all protocols and is to be expected—finite-buffered switches suffer more losses in the presence of more bursty traffic.
- RAPID significantly outperforms other protocols in its ability to adapt to burstiness—it consistently utilizes a much higher fraction of the bursty avail-bw than other protocols. Fig 9 illustrates this for BCT cross-traffic.
- With UDP and SCT, despite the higher utilization, RAPID incurs much lower packet loss rates than Scalable—this is indicative of the negligible queuing expected of the protocol [10]. With BCT, RAPID yields 1.8Gbps more throughput than Scalable, but also a higher loss rate. With less aggressive rate-adapting parameters ($\tau = 50ms$,$\eta = \frac{1}{4}$), however, RAPID yields a 0.013% loss rate, while still maintaining high throughput.



(a) RTT=5ms **(b)** RTT=30ms

Fig. 11: Flow Duration of Web Traffic



(a) scalable **(b)** RAPID

Fig. 12: Intra-Protocol Fairness

D. TCP Friendliness with Responsive and Bursty Web traffic

Web traffic transferred using conventional TCP continues to dominate the Internet. A high-speed protocol can be deployed over the Internet only if it has minimal impact on co-existing conventional low-speed TCP transfers. To study this, we generate responsive web traffic, as described in Section V-A1, and instantiate a high-speed transfer that shares the bottleneck link. We repeat the experiment by using different protocols for the high-speed transfer, as well as with RTT=5ms (to emulate increasing use of content distributions caches).

Note that each of the web transfers sharing the bottleneck link with a high-speed transfer, is responsive to any increased delays and losses. One important metric of web traffic performance is flow duration—increased flow duration strongly reduces user satisfaction. We use this as the primary metric in this section to study the impact on web traffic.

The more quickly RAPID grabs spare bandwidth, the higher is its throughput—however, the more transient queuing it causes in bottleneck buffers, and the more it impacts the performance of cross-traffic. In RAPID, this trade-off is controlled by the rate-adaptation parameters (τ, η) [10]. We first study the influence of these two parameters and briefly summarize our findings below:[7]

- *RAPID throughput:* With fixed τ, RAPID throughput first increases with $\frac{1}{\eta}$ due to its more aggressive behavior, and then decreases with it due to more induced losses. Identical $\tau \times \eta$ yields comparable throughput, which agrees with the simulation results in [29]. As long as $\tau \times \eta \geq 5$, RAPID experiences negligible losses.
- *Web traffic performance:* Smaller $\tau \times \eta$ increases the duration of co-existing low-speed web transfers. Although identical $\tau \times \eta$ yields similar RAPID throughput, a larger τ helps to reduce the median and the tail of the flow-duration distribution for web traffic.

Thus, for network operators targeting minimal impact on web traffic, a more conservative RAPID configuration with larger $\tau \times \eta$ and a larger τ is recommended. For our experiments in this paper, we use $(\tau, \eta) = (50, 4)$.

For $(\tau, \eta) = (50, 4)$, Fig 11 depicts the cumulative distribution of flow duration for web traffic, when they share the bottleneck link with a high-speed transfer (the corresponding throughput and loss rate of the high-speed transfer is listed in Table I). We notice that RAPID impacts web flow duration similarly to Cubic, while yielding much higher throughput

[7]For space constraints, we include detailed results only in [26].

TABLE II: Necessity of Implementation Mechanisms

RTT=30ms	No cross traffic(Gbps)	With 2.42 Gbps web traffic	
		Throughput(Gbps) Loss(%)	median low duration(ms)
Full RAPID	9.44	7.00, 0.001	1.45
V1 (no PAUSE)	8.37	6.77, 0.004	1.28
V2 (no μs timestamp)	9.83	7.64, 0.217	520.64
V3 (no arrival timestamp)	7.91	6.43, 0.001	1.24
V4 (no BASS)	9.82	7.83, 0.151	8705.24
V5 (no γ) N=64	8.99	6.45, 0.002	1.23
V5 (no γ) N=192	9.49	6.33, 0.007	1.45

(especially with RTT=30 ms). All other protocols either fail to grab a good share of bandwidth (e.g. Cubic), or behave so intrusively so as to more than double the median of flow duration and significantly lengthen the tail distribution (e.g. Scalable). We conclude that RAPID best addresses the trade-off between link utilization and TCP-friendliness—it achieves considerable link utilization while least starving conventional TCP traffic.

E. Intra-protocol Fairness

We next evaluate the intra-protocol fairness properties yielded by our implementation. We initiate three iperf flows between two pairs of end hosts. Each transfer emulates RTT=30ms and is active during different time intervals.

Fig 12a depicts the time-series of throughput obtained by the three transfers while we use Scalable as the underlying protocol—the protocol fails to yield any notion of fair share of the avail-bw. However, RAPID yields much greater fairness among co-existing flows in Fig 12b. This experiment and results are very close to the fairness observed under simulations in [10].

VI. How Critical Are the Added Mechanisms?

We have introduced several mechanisms to realize our RAPID implementation on a real system (vs. a simulator, as in [10])—these include: (i) inserting PAUSE frames to ensure precise gaps; (ii) implementing Qdiscs (I_R, E_R, and I_S) for higher accuracy in timestamping packet arrivals at the receiver; (iii) adapting BASS for accurate bandwidth estimation with short p-streams of $N = 64$; and (iv) decoupling the probing and adapting timescales to alleviate the stability-vs-adaptability trade-off. In this section, we ask: *are each of these necessary for achieving RAPID performance gains in high-speed settings?*

To answer this, we first run a RAPID flow with the complete set of mechanisms, under two experimental conditions—without any cross-traffic, and with the responsive web traffic

of 2.42Gbps burstiness on the 10 Gbps switch-to-switch link. Then, we reduce each of these mechanisms individually, and repeat the two experiments above (results in Table II):

- V1: Instead of inserting PAUSE frames for gap creation, E_S registers an hrtimer (using **qdisc_watchdog**) to schedule transmission. Fig 6 illustrated that there are significant errors in the intended send gaps in V1—the performance is naturally impacted. However, the impact of inaccurate g_s is lower than expected—this is due to alleviation by BASS, which is good at handling buffering related noise.
- V2 gets rid of I_R and E_R, but relies purely on the TCP timestamp option for estimating receive gaps.We find that V2 persistently over-estimates avail-bw as full link capacity,[8] starving cross-traffic and causing considerably high packet losses. This is because, with V2, the ms granularity of TCP timestamps obscures any fine-scale queuing delays within p-streams.
- V3 does not timestamp packet arrivals with I_R, but it does replace ACK timestamps with a μs precision value when each ACK is generated. E_R gets a timestamp by calling ktime_to_ns, and writes it in the TSval header field of returning ACKs. We find that V3 is influenced by ACK processing delays—it under-estimates the avail-bw, and under-utilizes the path even when it is idle.
- V4 gets rid of BASS denoising, and uses the raw receive gaps of p-streams for bandwidth estimation. We find that V4 persistently over-estimates the avail-bw as 10 Gbps—the spike-dips pattern (Fig 2c) in the p-streams wipes out any underlying trend of persistent queuing delays.
- V5 does not decouple the probing/adapting timescales—both are the p-stream length. We consider N=64 and N=192. (recall that in RAPID, probing timescale is N=64 and rate-adapting timescale is N=192). We find that neither a shorter, nor a longer timescale outperforms the decoupled RAPID. A short timescale (N=64) suffers from noisy $ABest$—it fails to fully utilize the empty path; while a long timescale (N=192) increases the duration for which each p-stream overloads the bottleneck, causing more losses. Also, bandwidth estimation using N=192, γ=1 are less accurate than using N=64, γ=3. Consequently, it yields less goodput.

To sum up, we find that *each* of the design components added in this paper is critical for ensuring that RAPID achieves its promised performance in practice.

VII. Concluding Remarks

This paper presents an ultra-high speed implementation of TCP RAPID. We conduct evaluations to show that the implementation successfully tackles several real-world challenges faced by the protocol and meets the performance bar set by simulation-based evaluations. Even more fundamentally, the networking community is generally skeptical about the practicality of delay, rate, or bandwidth-based congestion control—this paper takes a significant step in presenting evidence to convince them otherwise.

[8]This is why V2 and V4 deceitfully offer higher goodput than RAPID on idle paths (full RAPID is expected to under-estimate avail-bw within 10%).

For future work, we will continue to conduct intensive evaluations—we hope to upgrade our end hosts and conduct evaluations at truly 40 Gbps speeds. We also plan to deploy our implementation within scientific applications, especially in a wide-area setting. Besides, we are planning on evaluating our implementation in environments other than the large bandwidth-delay product networks considered here—specifically, data center environments, wireless environments, as well as for wide-area streaming media applications.

References

[1] Keith Winstein et al. Stochastic forecasts achieve high throughput and low delay over cellular networks. In *NSDI*, 2013.

[2] Radhika Mittal et al. Timely: Rtt-based congestion control for the datacenter. In *SIGCOMM*. ACM, 2015.

[3] Mayutan Arumaithurai et al. NF-TCP: A Network Friendly TCP Variant for Background Delay-insensitive Applications. In *International Conference on Research in Networking*. Springer, 2011.

[4] Shao Liu et al. Tcp-illinois: A loss-and-delay-based congestion control algorithm for high-speed networks. *Performance Evaluation*, 65, 2008.

[5] Kun Tan et al. A Compound TCP Approach for High-speed and Long Distance Networks. In *Proc. IEEE INFOCOM*, 2006.

[6] Andrea Baiocchi et al. YeAH-TCP: Yet Another Highspeed TCP. In *Proc. PFLDnet*, volume 7, pages 37–42, 2007.

[7] Rebecca Lovewell et al. Packet-scale congestion control paradigm. *IEEE/ACM Transactions on Networking*, 2016.

[8] Aleksandar Kuzmanovic et al. Tcp-lp: Low-priority service via end-point congestion control. *IEEE/ACM TON*, 2006.

[9] Guojun Jin et al. System capability effects on algorithms for network bandwidth measurement. In *Proc. SIGCOMM*. ACM, 2003.

[10] Vishnu Konda et al. RAPID: Shrinking the Congestion-control Timescale. In *Proc. INFOCOM*. IEEE, 2009.

[11] Thomas E Anderson et al. Pcp: Efficient endpoint congestion control. In *NSDI*, 2006.

[12] Yunhong Gu et al. Udt: Udp-based data transfer for high-speed wide area networks. *Computer Networks*, 2007.

[13] Vinay Joseph Ribeiro et al. pathchirp: Efficient available bandwidth estimation for network paths. In *Proc. PAM*, 2003.

[14] Manish Jain et al. Pathload: A measurement tool for end-to-end available bandwidth. In *Proc. PAM*, 2002.

[15] Jacob Strauss et al. A measurement study of available bandwidth estimation tools. In *Proc. SIGCOMM*. ACM, 2003.

[16] Hiroyuki Kamezawa et al. Inter-layer coordination for parallel tcp streams on long fat pipe networks. In *Proc. ACM/IEEE conference on Supercomputing*. IEEE, 2004.

[17] Antony Antony et al. Microscopic examination of tcp flows over transatlantic links. *Future Generation Computer Systems*, 2003.

[18] Z-L Zhang et al. Small-time scaling behaviors of internet backbone traffic: an empirical study. In *Proc. INFOCOM*. IEEE, 2003.

[19] Qianwen Yin et al. Can bandwidth estimation tackle noise at ultra-high speeds? In *Proc. ICNP*. IEEE, 2014.

[20] David X Wei et al. Fast tcp: Motivation, architecture, algorithms, performance. *IEEE/ACM ToN*, 14, 2006.

[21] Saverio Mascolo et al. Tcp westwood: Bandwidth estimation for enhanced transport over wireless links. In *Proc. MobiCom*. ACM, 2001.

[22] Seong-Ryong Kang et al. Characterizing tight-link bandwidth of multi-hop paths using probing response curves. In *Proc. IWQoS*. IEEE, 2010.

[23] Qianwen Yin et al. Can machine learning benefit bandwidth estimation at ultra-high speeds? In *Proc. PAM*. Springer, 2016.

[24] Sea Shalunov et al. Low extra delay background transport (ledbat). Technical report, 2012.

[25] Ryousei Takano et al. Design and evaluation of precise software pacing mechanisms for fast long-distance networks. *Proc. PFLDnet*, 2005.

[26] Q. Yin et al. TCP Rapid: From Theory to Practice. Technical Report 17-001, Department of Computer Science, UNC Chapel Hill, 2017.

[27] Yashar Ganjali et al. Update on buffer sizing in internet routers. *SIGCOMM*, 2006.

[28] P. Barford et al. Generating representative web workloads for network and server performance evaluation. *SIGMETRICS*, 1998.

[29] Rebecca Lovewell et al. Impact of cross traffic burstiness on the packet-scale paradigm. In *LANMAN*. IEEE, 2011.