Network Working Group	L. Andrew, Ed.	
Internet-Draft	CAIA, Swinburne University of	
Intended status: BCP	Technology	
Expires: January 9, 2010	S. Floyd, Ed.	
	ICSI Center for Internet Research	
	G. Wang, editor	
	NEC, China	
	July 08, 2009	

TOC

Common TCP Evaluation Suite
draft-irtf-tmrg-tests-02.txt

Status of this Memo

Abstract

This document presents an evaluation test suite for the initial evaluation of proposed TCP modifications. The goal of the test suite is to allow researchers to quickly and easily evaluate their proposed TCP extensions in simulators and testbeds using a common set of well-defined, standard test cases, in order to compare and contrast proposals against standard TCP as well as other proposed modifications. This test suite is not intended to result in an exhaustive evaluation of a proposed TCP modification or new

congestion control mechanism. Instead, the focus is on quickly and easily generating an initial evaluation report that allows the networking community to understand and discuss the behavioral aspects of a new proposal, in order to guide further experimentation that will be needed to fully investigate the specific aspects of a new proposal.

Table of Contents

1. Introduction

TOC

This document describes a common test suite for the initial evaluation of new TCP extensions. It defines a small number of evaluation scenarios, including traffic and delay distributions, network topologies, and evaluation parameters and metrics. The motivation for such an evaluation suite is to help researchers in evaluating their proposed modifications to TCP. The evaluation suite will also enable independent duplication and verification of reported results by others, which is an important aspect of the scientific method that is not often put to use by the networking community. A specific target is that the evaluations should be able to be completed in three days of simulations, or in a reasonable amount of effort in a testbed.

This document is an outcome of a ``round-table'' meeting on TCP evaluation, held at Caltech on November 8-9, 2007. This document is the first step in constructing the evaluation suite; the goal is for the evaluation suite to be adapted in response from feedback from the networking community.

2. Traffic generation

TOC

Congestion control concerns the response of flows to bandwidth limitations or to the presence of other flows. For a realistic testing of a congestion control protocol, we design scenarios to use reasonably-typical traffic; most scenarios use traffic generated from a traffic generator, with a range of start times for user sessions, connection sizes, and the like, mimicking the traffic patterns commonly observed in the Internet. Cross-traffic and reverse-path traffic have the desirable effect of reducing the occurrence of pathological conditions such as global synchronization among competing flows that might otherwise be mis-interpreted as normal average behaviours of those protocols [FK03] (Floyd, S. and E. Kohler, “Internet Research Needs Better Models,” .), [MV06] (Mascolo, S. and F. Vacirca, “The Effect of Reverse Traffic on the Performance of New TCP Congestion Control Algorithms for Gigabit Networks,” .). This traffic must be reasonably realistic for the tests to predict the behaviour of congestion control protocols in real networks, and also well-defined so that statistical noise does not mask important effects.

It is important that the same ``amount'' of congestion or cross-traffic be used for the testing scenarios of different congestion control algorithms. This is complicated by the fact that packet arrivals and even flow arrivals are influenced by the behavior of the algorithms. For this reason, a pure packet-level generation of traffic where generated traffic does not respond to the behaviour of other present flows is not suitable. Instead, emulating application

or user behaviours at the end points using reactive protocols such as TCP in a closed-loop fashion results in a closer approximation of cross-traffic, where user behaviours are modeled by well-defined parameters for source inputs (e.g., request sizes for HTTP), destination inputs (e.g., response size), and think times between pairs of source and destination inputs. By setting appropriate parameters for the traffic generator, we can emulate non-greedy user-interactive traffic (e.g., HTTP 1.1, SMTP and Telnet) as well as greedy traffic (e.g., P2P and long file downloads). This approach models protocol reactions to the congestion caused by other flows in the common paths, although it fails to model the reactions of users themselves to the presence of the congestion.

While the protocols being tested may differ, it is important that we maintain the same ``load'' or level of congestion for the experimental scenarios. To enable this, we use a hybrid of open-loop and close-loop approaches. For this test suite, network traffic consists of sessions corresponding to individual users. Because users are independent, these session arrivals are well modeled by an open-loop Poisson process. A session may consist of a single greedy TCP flow, multiple greedy flows separated by user ``think'' times, or a single non-greedy flow with embedded think times. The session arrival process forms a Poisson process [HVA03] (Hohn, N., Veitch, D., and P. Abry, “The Impact of the Flow Arrival Process in Internet Traffic,” .). Both the think times and burst sizes have heavy-tailed distributions, with the exact distribution based on empirical studies. The think times and burst sizes will be chosen independently. This is unlikely to be the case in practice, but we have not been able to find any measurements of the joint distribution. We invite researchers to study this joint distribution, and future revisions of this test suite will use such statistics when they are available.

There are several traffic generators available that implement a similar approach to that discussed above. For now, we are planning to use the Tmix [Tmix] (Weigle, M., Adurthi, P., Hernandez-Campos, F., Jeffay, K., and F. Smith, “Tmix: a tool for generating realistic TCP application workloads in ns-2,” .) traffic generator. Tmix represents each TCP connection by a connection vector consisting of a sequence of (request-size, response-size, think-time) triples, thus representing bi-directional traffic. Connection vectors used for traffic generation can be obtained from Internet traffic traces. By taking measurement from various points of the Internet such as campus networks, DSL access links, and IPS core backbones, we can obtain sets of connection vectors for different levels of congested links. We plan to publish these connection vectors as part of this test suite. A draft set of connection vectors is avilable at http://wil.cs.caltech.edu/suite/TrafficTraces.php.

2.1. Loads

TOC

For most current traffic generators, the traffic is specified by an arrival rate for independent user sessions, along with specifications of connection sizes, number of connections per

sessions, user wait times within sessions, and the like. For many of the scenarios, such as the basic scenarios in Section 3.1 (Basic scenarios), each scenario is run for a range of loads, where the load is varied by varying the rate of session arrivals. For a given congestion control mechanism, experiments run with different loads are likely to have different packet drop rates, and different levels of statistical multiplexing.

Because the session arrival times are specified independently of the transfer times, one way to specify the load would be as A = E[f]/ E[t], where E[f] is the mean session size (in bits transferred), E[t] is the mean session inter-arrival time in seconds, and A is the load in bps.

It is important to test congestion control in ``overloaded'' conditions. However, if A>c , where c is the capacity of the bottleneck link, then the system has no equilibrium. Such cases are studied in Section 3.4 (Transients: release of bandwidth, arrival of many flows). In long-running experiments with A>c , the expected number of flows would increase without bound. This means that the measured results would be very sensitive to the duration of the simulation.

Instead, for equilibrium experiments, we measure the load as the ``mean number of jobs in an M/G/1 queue using processor sharing,'' where a job is a user session. This reflects the fact that TCP aims at processor sharing of variable sized files. Because processor sharing is a symmetric discipline [Kelly79] (Kelly, F., “Reversibility and stochastic networks,” .), the mean number of flows is equal to that of an M/M/1 queue, namely rho/(1-rho) , where rho=lambda S/C , and lambda [flows per second] is the arrival rate of jobs/flows, S [bits] is the mean job size and C [bits per second] is the bottleneck capacity. For small loads, say 10%, this is essentially equal to the fraction of the capacity. However, for overloaded systems, the fraction of the bandwidth used will be much less than this measure of load.

In order to improve the traffic generators used in these scenarios, we invite researchers to explore how the user behavior, as reflected in the connection sizes, user wait times, and number of connections per session, might be affected by the level of congestion experienced within a session [RMC03] (Rossi, D., Mellia, M., and C. Casetti, “User Patience and the Web: a Hands-on Investigation,” .).

2.2. Equilibrium

TOC

In order to minimize the dependence of the results on the experiment durations, scenarios should be as stationary as possible. To this end, experiments will start with rho/(1-rho) active cross-traffic flows, with traffic of the specified load.

This is insufficient if the traces have very long pauses between bursts, because this initial loading has all finished by the time the number of actual tmix sessions builds up. It may be better to

start with several (many?) pre-existing connection vectors instead of greedy sources.

It is still an open issue whether to use tests with rho>1 . If such tests are used, the initial number of flows will need to be defined.

Note that the distribution of the durations of the active flows at a given time is (often significantly) different from the distribution of flow durations, skewed toward long flows. For simplicity, this will be ignored and the initial flow sizes will be drawn from the general flow size distribution.

2.3. Packet size distribution

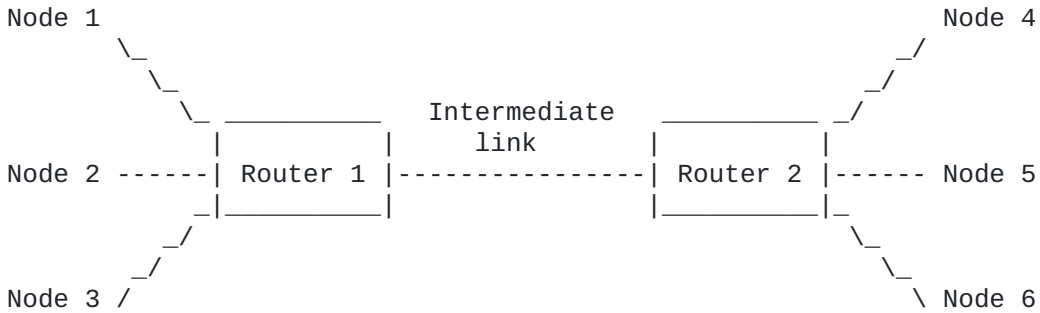
TOC

For flows generated by the traffic generator, 10% use 536-byte packets, and 90% 1500-byte packets. The packet size of each flow will be specified along with the start time and duration, to maximize the repeatability.

2.4. Round Trip Times

TOC

Most tests use a simple dumbbell topology with a central link that connects two routers, as illustrated in Figure 1. Each router is connected to three nodes by edge links. In order to generate a typical range of round trip times, edge links have different delays. On one side, the one-way propagation delays are: 0 ms, 12 ms and 25 ms; on the other: 2 ms, 37 ms, and 75 ms. Traffic is uniformly shared among the nine source/destination pairs, giving a distribution of per-flow RTTs in the absence of queueing delay shown in Figure 2. These RTTs are computed for a dumbbell topology with a delay of 0 ms for the central link. The delay for the central link is given in the specific scenarios in the next section.



A dumbbell topology

Figure 1

For dummynet experiments, delays can be obtained by specifying the delay of each flow.

Path	RTT	Path	RTT	Path	RTT
1-4	4	1-5	74	1-6	150
2-4	28	2-5	98	2-6	174
3-4	54	3-5	124	3-6	200

RTTs of the paths between two nodes, in milliseconds. **These RTTs are subject to change, based on comparison between the resulting packet-weighted RTT distribution and measurements I'd like to change the RTT 1-4 to 3ms or 5ms instead of 4ms... -- LA**

Figure 2

3. Scenarios

TOC

It is not possible to provide TCP researchers with a complete set of scenarios for an exhaustive evaluation of a new TCP extension; especially because the characteristics of a new extension will often require experiments with specific scenarios that highlight its behavior. On the other hand, an exhaustive evaluation of a TCP extension will need to include several standard scenarios, and it is the focus of the test suite described in this section to define this initial set of test cases.

3.1. Basic scenarios

TOC

The purpose of the basic scenarios is to explore the behavior of a TCP extension over different link types. The scenarios use the dumbbell topology of Section 2.4 (Round Trip Times), with the link delays modified as specified below.

This basic topology is used to instantiate several basic scenarios, by appropriately choosing capacity and delay parameters for the individual links. Depending on the configuration, the bottleneck link may be in one of the edge links or the central link.

3.1.1. Topology and background traffic

TOC

The basic scenarios are for a single topology, with a range of capacities and RTTs. For each scenario, traffic levels of uncongested, mild congestion, and moderate congestion are specified; these are explained below.

Data Center: The data center scenario models a case where bandwidth is plentiful and link delays are generally low. It uses the same configuration for the central link and all of the edge links. All links have a capacity of either 1 Gbps, 2.5 Gbps or 10 Gbps; links from nodes 1, 2 and 4 have a one-way propagation delay of 1 ms, while those from nodes 3, 5 and 6 have 10 ms [WCL05] (Wei, D., Cao, P., and S. Low, "Time for a TCP Benchmark Suite?," .), and the common link has 0 ms delay.

Uncongested: TBD Mild congestion: TBD Moderate congestion: TBD

Access Link: The access link scenario models an access link connecting an institution (e.g., a university or corporation) to an ISP. The central and edge links are all 100 Mbps. The one-way propagation delay of the central link is 2 ms, while the edge links have the delays given in Section 2.4 (Round Trip Times). Our goal in assigning delays to edge links is only to give a realistic distribution of round-trip times for traffic on the central link.

Uncongested: TBD Mild congestion: TBD Moderate congestion: TBD

Trans-Oceanic Link: The trans-oceanic scenario models a test case where mostly lower-delay edge links feed into a high-delay central link. The central link is 1 Gbps, with a one-way propagation delay of 65 ms. The edge links have the same bandwidth as the central link, with the one-way delays given in Section 2.4 (Round Trip Times). An alternative would be to use smaller delays for the edge links, with one-way delays for each set of three edge links of 5, 10, and 25 ms. **Implementations may use a smaller bandwidth for the trans-oceanic link, for example to run a simulation in a feasible amount of time. In testbeds, one of the metrics should be the number of timeouts in servers, due to implementation issues when running at high speed.**

Uncongested: TBD Mild congestion: TBD Moderate congestion: TBD

Geostationary Satellite: The geostationary satellite scenario models an asymmetric test case with a high-bandwidth downlink and a low-bandwidth uplink [HK99] (Henderson, T. and R. Katz, "Transport Protocols for Internet-Compatible Satellite Networks," .), [GF04] (Gurtov, A. and S. Floyd, "Modeling Wireless Links for Transport Protocols," .). The capacity of the central link is 40 Mbps with a one-way propagation delay of 300 ms. The downlink capacity of the edge links is also 40 Mbps, but their uplink capacity is only

4 Mbps. Edge one-way delays are as given in Section 2.4 (Round Trip Times). Note that ``downlink'' is towards the router for edge links attached to the first router, and away from the router for edge links on the other router.

Uncongested: TBD Mild congestion: TBD Moderate congestion: TBD

Wireless Access: The wireless access scenario models wireless access to the wired backbone. The capacity of the central link is 100 Mbps with 2 ms of one-way delay. All links to Router 1 are wired. Router 2 has a shared wireless link of nominal bit rate 11 Mbps (to model IEEE 802.11b links) or 54 Mbps (IEEE 802.11a/g) with a one-way delay of 1us connected to dummy nodes 4', 5' and 6', which are then connected to nodes 4, 5 and 6 by wired links of delays 2, 37 and 75 ms. This is to achieve the same RTT distribution as the other scenarios, while allowing a CSMA model to have realistic delay for a WLAN.

Note that wireless links have many other unique properties not captured by delay and bitrate. In particular, the physical layer might suffer from propagation effects that result in packet losses, and the MAC layer might add high jitter under contention or large steps in bandwidth due to adaptive modulation and coding. Specifying these properties is beyond the scope of the current first version of this test suite.

Uncongested: TBD Mild congestion: TBD Moderate congestion: TBD

Dial-up Link: The dial-up link scenario models a network with a dial-up link of 64 kbps and a one-way delay of 5 ms for the central link. **modems are asymmetric, 56k downlink and 33.6k or 48k uplink. Should we change this?** This could be thought of as modeling a scenario reported as typical in Africa, with many users sharing a single low-bandwidth dial-up link.

Uncongested: TBD Mild congestion: TBD Moderate congestion: TBD

Traffic: For each of the basic scenarios, three cases are tested: uncongested; mild congestion, and moderate congestion. All cases will use scaled versions of the traces available at <http://wil.cs.caltech.edu/suite>. **The exact traffic loads and run times for each scenario still need to be agreed upon. There is ongoing debate about whether rho>1 is needed to get moderate to high congestion. If rho>1 is used, note that the results will depend heavily on the run time, because congestion will progressively build up. In those cases, metrics which consider this non-stationarity may be more useful than average quantities.** In the default case, the reverse path has a low level of traffic (10% load). The buffer size at the two routers is set to the maximum bandwidth-delay-product for a 100 ms flow (i.e., a maximum queueing delay of 100 ms), with drop-tail queues in units of packets. Each run will be for at least a hundred seconds, and the metrics will not cover the initial warm-up times of each run. (Testbeds might use longer run times, as should simulations with smaller bandwidth-delay products.)

As with all of the scenarios in this document, the basic scenarios could benefit from more measurement studies about characteristics of congested links in the current Internet, and about trends that could help predict the characteristics of congested links in the future. This would include more measurements on typical packet drop rates, and on the range of round-trip times for traffic on congested links.

For the access link scenario, more extensive simulations or experiments will be run, with both drop-tail and RED queue management, with drop-tail queues in units of both bytes and packets, and with RED queue management both in byte mode and in packet mode. Specific TCP extensions may require the evaluation of associated AQM mechanisms. For the access link scenario, simulations or experiments will also include runs with a reverse-path load equal to the forward-path load. For the access link scenario, additional experiments will use a range of buffer sizes, including 20% and 200% of the bandwidth-delay product for a 100 ms flow.

3.1.2. Flows under test

TOC

For this basic scenario, there is no differentiation between ``cross-traffic'' and the ``flows under test''. The aggregate traffic is under test, with the metrics exploring both aggregate traffic and distributions of flow-specific metrics.

3.1.3. Outputs

TOC

For each run, the following metrics will be collected, for the central link in each direction: the aggregate link utilization, the average packet drop rate, and the average queueing delay, all over the second half of the run. **This metric could be difficult to gather in emulated testbeds since routers statistics of queue utilization are not always reliable and depend on time-scale.** Separate statistics should be reported for each direction in the satellite and wireless access scenarios, since those networks are asymmetric. **Should "over the second half of the run" be "starting after 50s"? Sally used the second half of the run, for 100s simulations, but we to get non-random results, we should run for longer. The warm-up time doesn't need to scale up with the run length.**

Other metrics of interest for general scenarios can be grouped in two sets: flow-centric and stability. The flow-centric metrics include the sending rate, good-put, cumulative loss and queueing delay trajectory for each flow, over time, and the transfer time per flow versus file size. **Testbeds could use monitors in the TCP layer (e.g., Web100) to estimate the queueing delay and loss. NS2 flowmon has problems, because it seems not to release memory associated with terminated flows.** Stability properties of interest include the standard deviation of the throughput and the queueing delay for the bottleneck link and for flows [WCL05] (Wei, D., Cao, P., and S. Low,

"Time for a TCP Benchmark Suite?," .). The worst case stability is also considered.

3.2. Delay/throughput tradeoff as function of queue size

TOC

Different queue management mechanisms have different delay-throughput tradeoffs. E.g., Adaptive Virtual Queue [KS01] (Kunniyur, S. and R. Srikant, "Analysis and Design of an Adaptive Virtual Queue (AVQ) Algorithm for Active Queue Management," .) gives low delay, at the expense of lower throughput. Different congestion control mechanisms may have different tradeoffs, which these tests aim to illustrate.

3.2.1. Topology and background traffic

TOC

These tests use the topology of Section 2.4 (Round Trip Times). This test is run for the access link scenario in Section 3.1 (Basic scenarios).

For each Drop-Tail scenario set, five tests are run, with buffer sizes of 10%, 20%, 50%, 100%, and 200% of the Bandwidth Delay Product (BDP) for a 100 ms flow. For each AQM scenario (if used), five tests are run, with a target average queue size of 2.5%, 5%, 10%, 20%, and 50% of the BDP, with a buffer equal to the BDP.

3.2.2. Flows under test

TOC

The level of traffic from the traffic generator will be specified so that when a buffer size of 100% of the BDP is used with Drop Tail queue management, there is a moderate level of congestion (e.g., 1-2% packet drop rates when Standard TCP is used). Alternately, a range of traffic levels could be chosen, with a scenario set run for each traffic level (as in the examples cited below).

3.2.3. Outputs

TOC

For each test, three figures are kept, the average throughput, the average packet drop rate, and the average queueing delay over the second half of the test.

For each set of scenarios, the output is two graphs. For the delay/bandwidth graph, the x-axis shows the average queueing delay, and

the y-axis shows the average throughput. For the drop-rate graph, the x-axis shows the average queueing delay, and the y-axis shows the average packet drop rate. Each pair of graphs illustrates the delay/throughput/drop-rate tradeoffs for this congestion control mechanism. For an AQM mechanism, each pair of graphs also illustrates how the throughput and average queue size vary (or don't vary) as a function of the traffic load. Examples of delay/ throughput tradeoffs appear in Figures 1-3 of [FS01] (Floyd, S., Gummadi, R., and S. Shenker, "Adaptive RED: An Algorithm for Increasing the Robustness of RED," .) and Figures 4-5 of [AHM08] (Andrew, L., Hanly, S., and R. Mukhtar, "Active Queue Management for Fair Resource Allocation in Wireless Networks," .).

3.3. Ramp up time: completion time of one flow

TOC

These tests aim to determine how quickly existing flows make room for new flows.

3.3.1. Topology and background traffic

TOC

Dumbbell. At least three capacities should be used, as close as possible to: 56 kbps, 10 Mbps and 1 Gbps. The 56 kbps case is included to investigate the performance using mobile handsets.

For each capacity, three RTT scenarios should be tested, in which the existing and newly arriving flow have RTTs of (74 ms, 74 ms), (124 ms, 28 ms) and (28 ms, 124 ms). **Was (80,80), (120,30), (30,120), but the above are taken from Table 1 to simplify implementation. OK?**

Throughout the experiment, there is also 10% bidirectional cross traffic, as described in Section 2 (Traffic generation), using the mix of RTTs described in Section 2.4 (Round Trip Times). All traffic is from the new TCP extension.

3.3.2. Flows under test

TOC

Traffic is dominated by two long lived flows, because we believe that to be the worst case, in which convergence is slowest.

One flow starts in ``equilibrium'' (at least having finished normal slow-start). A new flow then starts; slow-start is disabled by setting the initial slow-start threshold to the initial CWND. Slow start is disabled because this is the worst case, and could happen if a loss occurred in the first RTT. **Roman Chertov has suggested**

doing some tests with slow start enabled too. Will there be time? Wait until initial NS2 implementation is available to test

The experiment ends once the new flow has run for five minutes. Both of the flows use 1500-byte packets.

3.3.3. Outputs

TOC

The output of these experiments are the time until the 1500(10^n) th byte of the new flow is received, for n = 1,2,... . This measures how quickly the existing flow releases capacity to the new flow, without requiring a definition of when ``fairness'' has been achieved. By leaving the upper limit on n unspecified, the test remains applicable to very high-speed networks.

A single run of this test cannot achieve statistical reliability by running for a long time. Instead, an average over at least three runs should be taken. Each run must use different cross traffic, as specified in Section 2 (Traffic generation).

3.4. Transients: release of bandwidth, arrival of many flows

TOC

These tests investigate the impact of a sudden change of congestion level. They differ from the "Ramp up time" test in that the congestion here is caused by unresponsive traffic.

3.4.1. Topology and background traffic

TOC

The network is a single bottleneck link, with bit rate 100 Mbps, with a buffer of 1024 packets (120% BDP at 100 ms).

The transient traffic is generated using UDP, to avoid overlap with the scenario of Section 3.3 (Ramp up time: completion time of one flow) and isolate the behavior of the flows under study. Three transients are tested:

1. step decrease from 75 Mbps to 0 Mbps,

2. step increase from 0 Mbps to 75 Mbps,

3. 30 step increases of 2.5 Mbps at 1 s intervals, simulating a ``flash crowd'' effect.

These transients occur after the flow under test has exited slow-start, and remain until the end of the experiment.

There is no TCP cross traffic as described in Section 2 (Traffic generation) in this experiment. because flow arrivals/departures occur on timescales long compared with these effects.

3.4.2. Flows under test

TOC

There is one flow under test: a long-lived flow in the same direction as the transient traffic, with a 100 ms RTT.

3.4.3. Outputs

TOC

For the decrease in cross traffic, the metrics are (i) the time taken for the flow under test to increase its window to 60%, 80% and 90% of its BDP, and (ii) the maximum change of the window in a single RTT while the window is increasing to that value.

For cases with an increase in cross traffic, the metric is the number of packets dropped by the cross traffic from the start of the transient until 100 s after the transient. This measures the harm caused by algorithms which reduce their rates too slowly on congestion.

3.5. Impact on standard TCP traffic

TOC

Many new TCP proposals achieve a gain, G , in their own throughput at the expense of a loss, L , in the throughput of standard TCP flows sharing a bottleneck, as well as by increasing the link utilization. In this context a "standard TCP flow" is defined as a flow using SACK TCP (Floyd, S., Mahdavi, J., Mathis, M., and M. Podolsky, “An Extension to the Selective Acknowledgement (SACK) Option for TCP,” July 2000.) [RFC2883], but without ECN (Ramakrishnan, K., Floyd, S., and D. Black, “The Addition of Explicit Congestion Notification (ECN) to IP,” September 2001.) [RFC3168]. **What about:** Window scaling (Jacobson, V., Braden, B., and D. Borman, “TCP Extensions for High Performance,” May 1992.) [RFC1323] (yes), FRTO (Sarolahti, P. and M. Kojo, “Forward RTO-Recovery (F-RTO): An Algorithm for Detecting Spurious Retransmission Timeouts with TCP and the Stream Control Transmission Protocol (SCTP),” August 2005.) [RFC4138] (yes), ABC (Allman, M., “TCP Congestion Control with Appropriate Byte Counting (ABC),” February 2003.) [RFC3465] (no)? The intention is for a "standard TCP flow" to correspond to TCP as commonly deployed in the Internet today (with the notable exception of CUBIC, which runs by default on the majority of web servers). This scenario quantifies this tradeoff.

3.5.1. Topology and background traffic

TOC

The dumbbell of Section 2.4 (Round Trip Times) is used with the same capacities as for the convergence tests (Section 3.3 (Ramp up time: completion time of one flow)). All traffic in this scenario comes from the flows under test.

3.5.2. Flows under test

TOC

The scenario is performed by conducting pairs of experiments, with identical flow arrival times and flow sizes. Within each experiment, flows are divided into two camps. For every flow in camp A, there is a flow with the same size, source and destination in camp B, and vice versa. The start time of the two flows are within 2 s.

The file sizes and start times are as specified in Section 2 (Traffic generation), with start times scaled to achieve loads of 50% and 100%. In addition, both camps have a long-lived flow. The experiments last for 1200 seconds.

In the first experiment, called BASELINE, both camp A and camp B use standard TCP. In the second, called MIX, camp A uses standard TCP and camp B uses the new TCP extension.

The rationale for having paired camps is to remove the statistical uncertainty which would come from randomly choosing half of the flows to run each algorithm. This way, camp A and camp B have the same loads.

3.5.3. Outputs

TOC

The gain achieved by the new algorithm and loss incurred by standard TCP are given by G=T(B)_Mix/T(B)_Baseline and L=T(A)_Mix/ T(A)_Baseline where T(x) is the throughput obtained by camp x , measured as the amount of data acknowledged by the receivers (that is, ``goodput''), and taken over the last 8000 seconds of the experiment.

The loss, L , is analogous to the ``bandwidth stolen from TCP'' in [SA03] (Souza, E. and D. Agarwal, “A HighSpeed TCP Study: Characteristics and Deployment Issues,” .) and ``throughput degradation'' in [SSM07] (Shimonishi, H., Sanadidi, M., and T. Murase, “Assessing Interactions among Legacy and High-Speed TCP Protocols,” .).

A plot of G vs L represents the tradeoff between efficiency and loss.

3.5.4. Suggestions

TOC

Other statistics of interest are the values of G and L for each quartile of file sizes. This will reveal whether the new proposal is more aggressive in starting up or more reluctant to release its share of capacity.

As always, testing at other loads and averaging over multiple runs are encouraged.

3.6. Intra-protocol and inter-RTT fairness

TOC

These tests aim to measure bottleneck bandwidth sharing among flows of the same protocol with the same RTT, which represents the flows going through the same routing path. The tests also measure inter-RTT fairness, the bandwidth sharing among flows of the same protocol where routing paths have a common bottleneck segment but might have different overall paths with different RTTs.

3.6.1. Topology and background traffic

TOC

The topology, the capacity and cross traffic conditions of these tests are the same as in Section 3.3 (Ramp up time: completion time of one flow). The bottleneck buffer is varied from 25% to 200% BDP for a 100 ms flow, increasing by factors of 2.

3.6.2. Flows under test

TOC

We use two flows of the same protocol for this experiment. The RTTs of the flows range from 10 ms to 160 ms (10 ms, 20 ms, 40 ms, 80 ms, and 160 ms) such that the ratio of the minimum RTT over the maximum RTT is at most 1/16. **In case a testbed doesn't support up to 160 ms RTT, the RTTs may be scaled down in proportion to the maximum RTT supported in that environment.**

Intra-protocol fairness: For each run, two flows with the same RTT, taken from the range of RTTs above start randomly within the first 10% of the experiment. The order in which these flows start doesn't matter. An additional test of interest, but not part of this suite,

would involve two extreme cases - two flows with very short or long RTTs (e.g., the delay less then 1-2 ms represents communication happen in the data-center and the delay larger than 600 ms considers communication over satellite).

Inter-RTT fairness: For each run, one flow with a fixed RTT of 160 ms starts first, and another flow with a different RTT taken from the range of RTTs above, joins afterward. The starting times of both two flows are randomly chosen within the first 10% of the experiment as before.

3.6.3. Outputs

TOC

The output of this experiment is the ratio of the average throughput values of the two flows. The output also includes the packet drop rate for the congested link.

3.7. Multiple bottlenecks

TOC

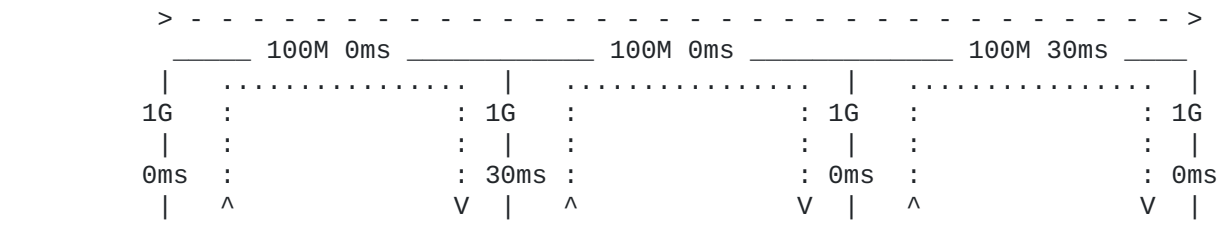
These experiments explore the relative bandwidth for a flow that traverses multiple bottlenecks, and flows with the same round-trip time that each traverse only one of the bottleneck links.

3.7.1. Topology and background traffic

TOC

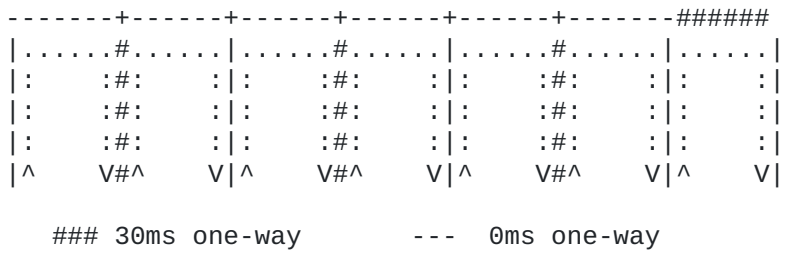
The topology is a ``parking-lot'' topology with three (horizontal) bottleneck links and four (vertical) access links. The bottleneck links have a rate of 100 Mbps, and the access links have a rate of 1 Gbps.

All flows have a round-trip time of 60 ms, to enable the effect of traversing multiple bottlenecks to be distinguished from that of different round trip times. This can be achieved as in Figure 3 by (a) the second access link having a one-way delay of 30 ms (b) the bottleneck link to which it does not connect having a one-way delay of 30 ms and (c) all other links having negligible delay. This can be extended to more than three bottlenecks as shown in Figure 4, by assigning a delay of 30 ms to every alternate access link, and to zero or one of the bottleneck links.) For the special case of three hops, an alternative is for all links to have a one-way delay of 10 ms, as shown in Figure 5. It is not clear whether there are interesting performance differences between these two topologies, and if so, which is more typical of the actual internet.



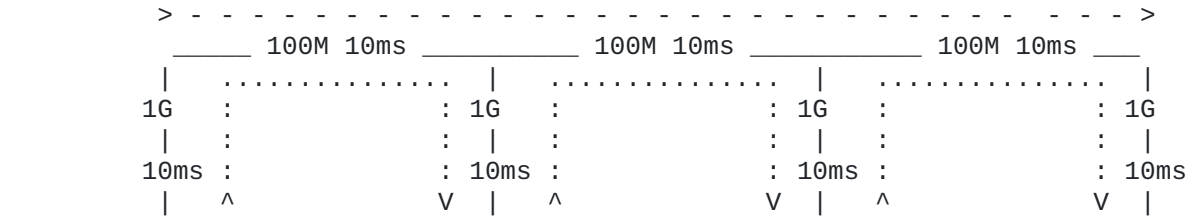
Basic multi-hop topology.

Figure 3



Extension to 7-hop parking lot. (Not part of the basic test suite.)

Figure 4



Alternative highly symmetric multi-hop topology.

Figure 5

Throughout the experiment, there is 10% bidirectional cross traffic on each of the three bottleneck links, as described in Section 2 (Traffic generation). The cross-traffic flows all traverse two access links and a single bottleneck link.

All traffic uses the new TCP extension.

3.7.2. Flows under test

TOC

In addition to the cross-traffic, there are four flows under test, all with traffic in the same direction on the bottleneck links. The multiple-bottleneck flow traverses no access links and all three bottleneck links. The three single-bottleneck flows each traverse two access links and a single bottleneck link, with one flow for each bottleneck link. The flows start in quick succession, separated by approximately 1 second. These flows last at least 5 minutes.

An additional test of interest would be to have a longer, multiple-bottleneck flow competing against shorter single-bottleneck flows.

3.7.3. Outputs

TOC

The output for this experiment is the ratio between the average throughput of the single-bottleneck flows and the throughput of the multiple-bottleneck flow, measured over the second half of the experiment. Output also includes the packet drop rate for the congested link.

3.8. Implementations

TOC

There are two on-going implementation efforts.

A testbed implementation is jointly being developed by the Centre for Advanced Internet Architectures (CAIA) at Swinburne University of technology and by Netlab at Caltech. It will eventually be available for public use through the web interface http://wil-ns.cs.caltech.edu/testing/benchmark/tmrg.php.

A simulation implementation in ns is being developed by NEC Labs, China. Contributions can be made via its source forge page, http://sourceforge.net/projects/tcpeval/.

3.9. Conclusions

TOC

An initial specification of an evaluation suite for TCP extensions has been described. Future versions will include: detailed specifications, with modifications for simulations and testbeds; more measurement results about congested links in the current Internet; alternate specifications; and specific sets of scenarios that can be run in a plausible period of time in simulators and testbeds, respectively.

Several software and hardware implementations of these tests are being developed for use by the community. An implementation is being developed on WAN-in-Lab [LATL07] (Lee, G., Andrew, L., Tang, A., and S. Low, “A WAN-in-Lab for protocol development,” .), which will allow users to upload Linux kernels via the web and will run tests similar to those described here. Some tests will be modified to suit the hardware available in WAN-in-Lab. An NS-2 implementation is also being developed at NEC. We invite others to contribute implementations on other simulator platforms, such as OMNeT++ and OpNet.

3.10. Acknowledgements

TOC

This work is based on a paper by Lachlan Andrew, Cesar Marcondes, Sally Floyd, Lawrence Dunn, Romaric Guillier, Wang Gang, Lars Eggert, Sangtae Ha and Injong Rhee.

The authors would also like to thank Roman Chertov, Doug Leith, Saverio Mascolo, Ihsan Qazi, Bob Shorten, David Wei and Michele Weigle for valuable feedback.

---

4. IANA Considerations

TOC

None.

---

5. Security Considerations

TOC

None.

---

6. Informative References

TOC

[AHM08]	Andrew, L., Hanly, S., and R. Mukhtar, “Active Queue Management for Fair Resource Allocation in Wireless Networks,” IEEE Transactions on Mobile Computing vol. 7, 2008.
[FK03]	Floyd, S. and E. Kohler, “Internet Research Needs Better Models,” SIGCOMM Computer Communication Review (CCR) vol. 33, no. 1, pp. 29-34, 2003.
[FS01]	Floyd, S., Gummadi, R., and S. Shenker, “Adaptive RED: An Algorithm for Increasing the Robustness of RED,” ICIR, Tech. Rep., 2001. [Online]. Available: http://www.icir.org/floyd/papers/adaptiveRed.pdf .
[GF04]	Gurtov, A. and S. Floyd, “Modeling Wireless Links for Transport Protocols,” SIGCOMM Computer Communication Review (CCR) vol. 34, no. 2, pp. 85-96, 2004.
[HK99]	Henderson, T. and R. Katz, “Transport Protocols for Internet-Compatible Satellite Networks,” IEEE Journal on Selected Areas in Communications (JSAC) vol. 17, no. 2, pp. 326-344, 1999.
[HVA03]	Hohn, N., Veitch, D., and P. Abry, “The Impact of the Flow Arrival Process in Internet Traffic,” Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03) vol. 6, pp. 37-40, 2003.
[KS01]	Kunniyur, S. and R. Srikant, “Analysis and Design of an Adaptive Virtual Queue (AVQ) Algorithm for Active Queue Management,” Proc. SIGCOMM'01 pp. 123-134, 2001.
[Kelly79]	Kelly, F., “Reversibility and stochastic networks,” Wiley 1979.
[LATL07]	Lee, G., Andrew, L., Tang, A., and S. Low, “A WAN-in-Lab for protocol development,” PFLDnet 2007.
[MV06]	Mascolo, S. and F. Vacirca, “The Effect of Reverse Traffic on the Performance of New TCP Congestion Control Algorithms for Gigabit Networks,” PFLDnet 2006.
[RFC1323]	Jacobson, V., Braden, B., and D. Borman, “TCP Extensions for High Performance,” RFC 1323, May 1992 (TXT).
[RFC2883]	Floyd, S., Mahdavi, J., Mathis, M., and M. Podolsky, “An Extension to the Selective Acknowledgement (SACK) Option for TCP,” RFC 2883, July 2000 (TXT).
[RFC3168]	Ramakrishnan, K., Floyd, S., and D. Black, “The Addition of Explicit Congestion Notification (ECN) to IP,” RFC 3168, September 2001 (TXT).
[RFC3465]	Allman, M., “TCP Congestion Control with Appropriate Byte Counting (ABC),” RFC 3465, February 2003 (TXT).
[RFC4138]	Sarolahti, P. and M. Kojo, “Forward RTO-Recovery (F-RTO): An Algorithm for Detecting Spurious Retransmission Timeouts with TCP and the Stream Control Transmission Protocol (SCTP),” RFC 4138, August 2005 (TXT).
[RMC03]	Rossi, D., Mellia, M., and C. Casetti, “User Patience and the Web: a Hands-on Investigation,” IEEE Globecom 2003.
[SA03]	Souza, E. and D. Agarwal, “A HighSpeed TCP Study: Characteristics and Deployment Issues,” LBNL, Technical Report LBNL-53215, 2003.
[SSM07]	Shimonishi, H., Sanadidi, M., and T. Murase, “Assessing Interactions among Legacy and High-Speed TCP Protocols,”

	Proc. Workshop on Protocols for Fast Long-Delay Networks (PFLDNet) 2007.
[Tmix]	Weigle, M., Adurthi, P., Hernandez-Campos, F., Jeffay, K., and F. Smith, “Tmix: a tool for generating realistic TCP application workloads in ns-2,” SIGCOMM Computer Communication Review (CCR) vol. 36, no. 3,pp. 65-76, 2006.
[WCL05]	Wei, D., Cao, P., and S. Low, “Time for a TCP Benchmark Suite?,” [Online]. Available: http://wil.cs.caltech.edu/pubs/DWei-TCPBenchmark06.ps 2006.

Authors' Addresses

TOC

	Lachlan Andrew
	CAIA, Swinburne University of Technology
	PO Box 218
	Hawthorn, Vic 3122
	Australia
Email:	landrew@swin.edu.au
	Sally Floyd
	ICSI Center for Internet Research
	1947 Center Street, Suite 600
	Berkeley, CA 94704
	USA
Email:	floyd@icir.org
	Gang Wang
	NEC, China
	Innovation Plaza, Tsinghua Science Park, 1 Zhongguancun East Road
	Beijing 100084
	China
Email:	wanggang@research.nec.com.cn

Full Copyright Statement

TOC

Intellectual Property

The IETF takes no position regarding the validity or scope of any Intellectual Property Rights or other rights that might be claimed to pertain to the implementation or use of the technology described in this document or the extent to which any license under such rights might or might not be available; nor does it represent that it has made any independent effort to identify any such rights. Information on the procedures with respect to rights in RFC documents can be found in BCP 78 and BCP 79.

Copies of IPR disclosures made to the IETF Secretariat and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementers or users of this specification can be obtained from the IETF on-line IPR repository at http://www.ietf.org/ipr.

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights that may cover technology that may be required to implement this standard. Please address the information to the IETF at ietf-ipr@ietf.org.