



HIGH-PERFORMANCE TCP FLOWS AND METHODS

SESSION RST-4512

RST-4512
9812_05_2004_c1

© 2004 Cisco Systems, Inc. All rights reserved.

1

Goals

Cisco.com

- **Provide insight as to why high-performance networks often have low per-flow throughput**
- **Discuss diagnostic tools available**
- **Examine common problems, and methods/proposals for increasing TCP throughput**
- **Review experimental results and case studies**
- **Take-away: Understand and be able to improve TCP performance; relate to developers and/or end users who might assert “there’s a problem with your network...”**
- **Audience assumption: Have some knowledge of how TCP works (but we’ll review basics)**

RST-4512
9812_05_2004_c1

© 2004 Cisco Systems, Inc. All rights reserved.

2

Agenda

Cisco.com

- **Problem Statement**
- **Instrumentation: Web100, Net100**
- **Path MTU Discovery**
- **TCP Dynamics Review (Sawtooth)**
- **TCP Dynamics: Upper-Bound Math**
- **Larger MTU (and Renewed Pathmtu)**
- **Change TCP Dynamics Approaches**
- **Experimental Results**
- **Case Studies**

RST-4512
9812_05_2004_c1

© 2004 Cisco Systems, Inc. All rights reserved.

3

Problem Statement (1)

Cisco.com

- **Over the past 15 years, many components impacting end-2-end performance have roughly tracked “Moore’s Law” (doubling each 18–24 months):**
 - End-station interface speed (shared 10Mbps → switched 1Gbps)**
 - Campus access link speed (1.5Mbps → 600+Mbps)**
 - WAN/backbone links (45Mbps → 10Gbps)**
 - Processor speed (40Mhz → 2Ghz), memory and disk bandwidth, etc.**
- **But not end-to-end throughput, for high bandwidth* delay product paths...why?**

RST-4512
9812_05_2004_c1

© 2004 Cisco Systems, Inc. All rights reserved.

4

Problem Statement (2)

Cisco.com

- 100Mbps end-to-end hosts often get 5–10Mbps throughput
- 1st issue: Transmit/receive buffers need to accommodate full bandwidth*delay product (web100, autotuning are addressing this)
- 2nd issue: TCP dynamics, behavior in presence of congestion- or BER-induced loss, time required to achieve “full utilization” on high-speed links
(related-path MTU discovery broken; L3/L4 revisions in progress, re-start of IETF PMTUD w.g., multiple implementations exist)

RST-4512
9812_05_2004_c1

BER: Bit-Error Rate

© 2004 Cisco Systems, Inc. All rights reserved.

5

INSTRUMENTATION, TOOLS



RST-4512
9812_05_2004_c1

© 2004 Cisco Systems, Inc. All rights reserved.

6

Instrumentation, Tools

Cisco.com

- **Help probe/solve at Least “issue #1”
(send/receive buffers)**
- **web100**
- **net100**

RST-4512
9812_05_2004_c1

© 2004 Cisco Systems, Inc. All rights reserved.

7

Web100

Cisco.com

- **Project to make it common for the average workstation to achieve 80–100Mbps e2e without requiring a network expert**
- **www.web100.org**
- **Phase 1: Kernel instrumentation for 100+ TCP variables on Linux (initially)**
- **Then add: Buffer autotuning**
- **Also makes good instructional tool; GUI and API**
- **(Note: Linux 2.4 also adds a variant of autotuning)**
- **Other tools, e.g.: miranda.ctd.anl.gov:7123/
Web100-enabled server, analyzes throughput to you;
Credits: Rich Carlson ANL/Internet2**

RST-4512
9812_05_2004_c1

© 2004 Cisco Systems, Inc. All rights reserved.

8

Bandwidth-Delay Product Quick Review

Cisco.com

- TCP sender and receiver buffers need to hold $\text{bandwidth} \times \text{delay}$ bytes to be able to “fill the pipe” (Why? Retransmission)
 - Quick example: $b/w=100\text{mbits/sec}$, $c=300,000\text{km/sec}$, $c_{\text{in_fiber}} \approx 180,000\text{km/sec}$, $\text{delay}=1800\text{km}$
 - $100\text{Mbits/sec} \times (1\text{B}/8\text{bits}) \times 1\text{s} / 180000\text{km} \times 1800\text{km} = 125,000\text{bytes}$
 - Many stock TCP's are 32kB, 64kB
- Why not make all buffers huge? Each TCP session reserves physical memory matching the buffer size

RST-4512
9812_05_2004_c1

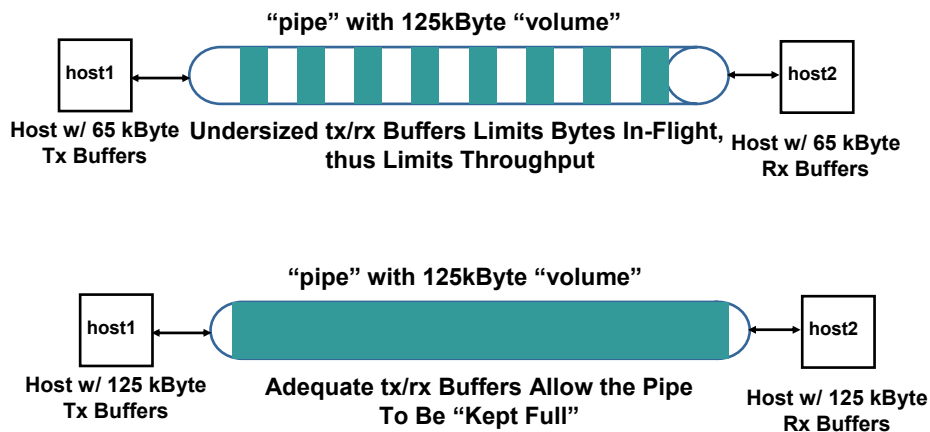
© 2004 Cisco Systems, Inc. All rights reserved.

9

Bandwidth-Delay Product Example

Cisco.com

Path bandwidth*delay: 100Mbps, 1800km → 125kBytes



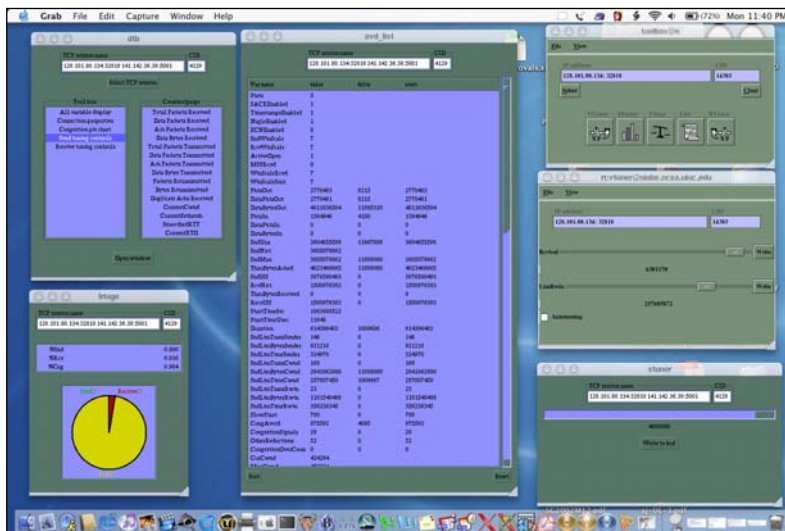
RST-4512
9812_05_2004_c1

© 2004 Cisco Systems, Inc. All rights reserved.

10

Web100 (Screenshot)

Cisco.com



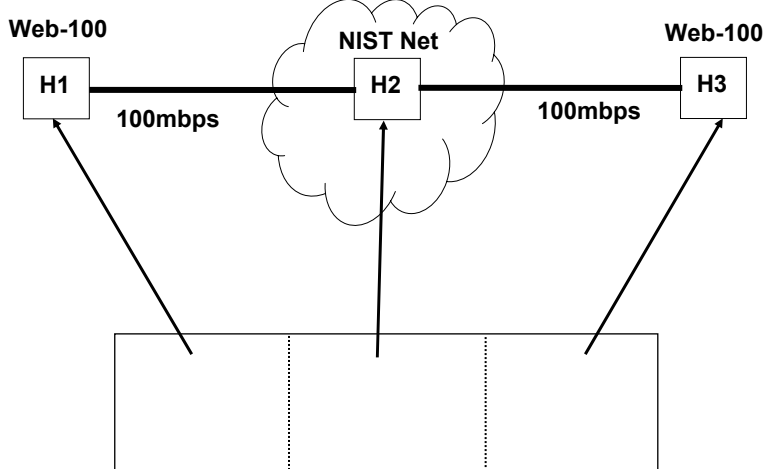
RST-4512
9812_05_2004_c1

© 2004 Cisco Systems, Inc. All rights reserved.

11

Experimental Setup (3x 1RU Linux 2.4.xx, 2.8GHz, 1GB, 2x100BT)

Cisco.com



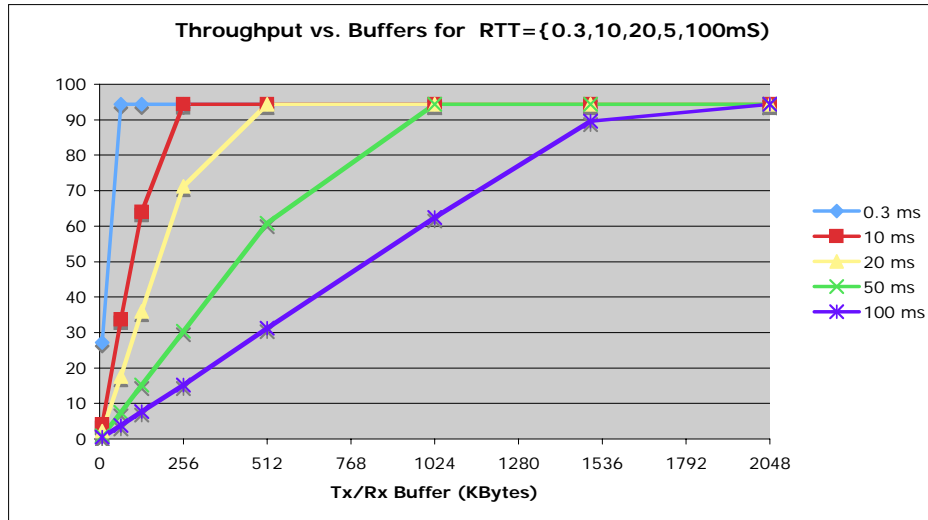
RST-4512
9812_05_2004_c1

© 2004 Cisco Systems, Inc. All rights reserved.

12

So what? Understand 3-5 parameters... {buffers, delay, loss, MTU} --> throughput

Cisco.com



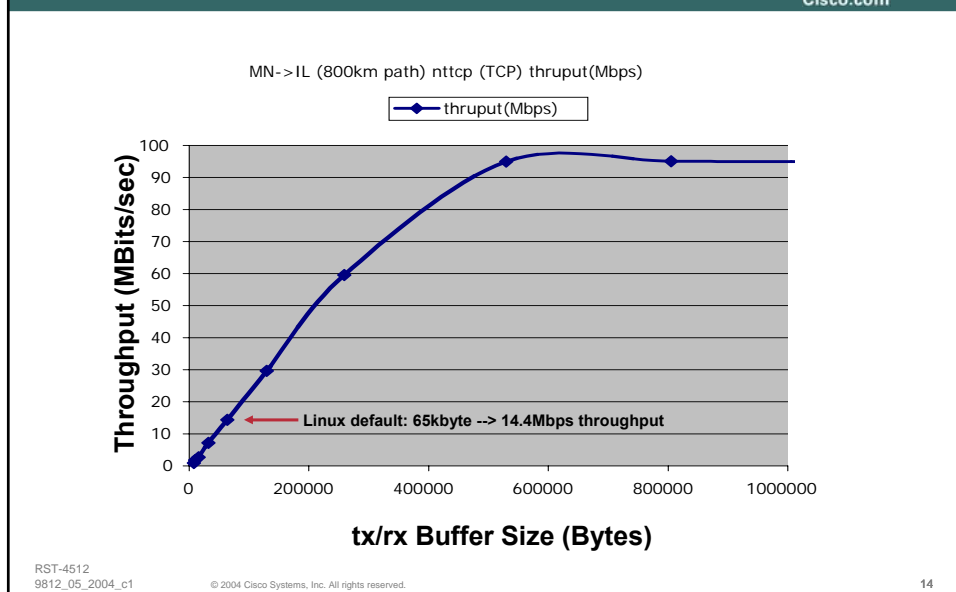
RST-4512
9812_05_2004_c1

© 2004 Cisco Systems, Inc. All rights reserved.

13

Impact of Tx/Rx Buffers (Example: 28msec RTT)

Cisco.com



RST-4512
9812_05_2004_c1

© 2004 Cisco Systems, Inc. All rights reserved.

14

Net100

Cisco.com

- Builds on top of web100 infrastructure
- www.csm.ornl.gov/~dunigan/net100
- www.net100.org/
- Makes it easy to control several aspects of TCP behavior
- Includes “WAD” (workaround daemon), which can make TCP decidedly unfair, if desired

RST-4512
9812_05_2004_c1

© 2004 Cisco Systems, Inc. All rights reserved.

15

PATH MTU DISCOVERY, TCP-BASICS REVIEW, AND TCP-BEHAVIOR LIMITS



RST-4512
9812_05_2004_c1

© 2004 Cisco Systems, Inc. All rights reserved.

16

Path MTU Discovery

Cisco.com

- After buffers, a potential source of e2e throughput improvement
- PMTUD essentially broken now—Firewalls, disabled ICMP responders, etc.
- Mathis et.al.—Propose TCP-level insertion of probes to periodically discover e2e path capabilities, and adjust session MTU to match
- ietf.org/internet-drafts/draft-mathis-plpmtud-00.txt

RST-4512
9812_05_2004_c1

© 2004 Cisco Systems, Inc. All rights reserved.

17

TCP: Basics Review

Cisco.com

- **Session initialization**
 - 3-way handshake, options negotiation
- **Slow-start**
 - Add one unacknowledged segment per new ACK received, leads to 1,2,4,8,16,...segments in-flight
- **Congestion avoidance/control**
 - Additive-Increase, Multiplicative-Decrease (AIMD)
- **Session close**
- **We'll focus mostly on congestion-avoidance phase**

RST-4512
9812_05_2004_c1

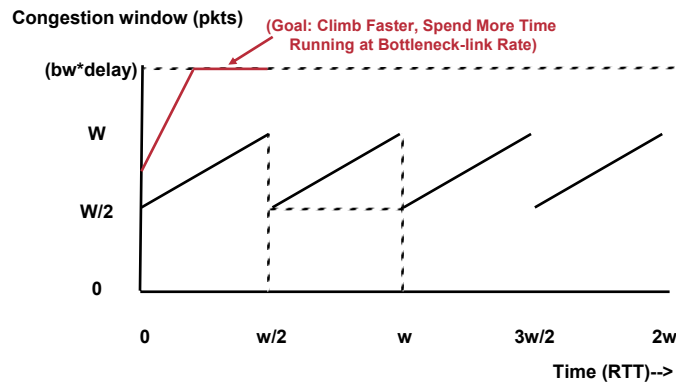
© 2004 Cisco Systems, Inc. All rights reserved.

18

Sawtooth Basics: TCP Dynamics (Congestion Control)

Cisco.com

Throughput Upper-Bound Developed From:



RST-4512
9812_05_2004_c1

© 2004 Cisco Systems, Inc. All rights reserved.

19

TCP: Nominal Upper-Bounds on Rate

Cisco.com

- $\text{Rate} \leq (0.7 \cdot \text{MSS}) / (\text{RTT} \cdot \sqrt{\text{loss}})$
(for constant pkt-loss probability)
- $\text{Rate} \leq (k \cdot \sqrt{\text{MSS}}) / (\text{RTT} \cdot \sqrt{\text{loss}})$
(for Bit-error-rate (BER) induced loss)
(src: Matt Mathis; supporting math follows)

MSS: TCP Max. Segment Size; RTT: Round-Trip Time

RST-4512
9812_05_2004_c1

© 2004 Cisco Systems, Inc. All rights reserved.

20

TCP Dynamics: Upper-Bound Math

Cisco.com

- Following 3 slides are quick summary of 1997 paper math, and some recent (BER-bound) updates
www.psc.edu/networking/papers/model_ccr97.ps
Matt's homepage: www.psc.edu/~mathis/
- (Ref. Graph on slide #19)

RST-4512
9812_05_2004_c1

© 2004 Cisco Systems, Inc. All rights reserved.

21

TCP Congestion Control: The Math

Cisco.com

- Loss probability p , so $(1/p-1)$ pkts delivered, then 1 loss (e.g. $p=0.01$: 99 delivered, then 1 loss)
- Area per cycle = $(w/2)^2 + (1/2)(w/2)^2$
= $(3/8)*(w)^2$ pkts
- This is $1/p$ pkts (by assumption); solve for w :
- $w = \sqrt{8/(3p)}$
- $BW = \text{data_per_cycle}/\text{time_per_cycle}$
= $(MSS*(3/8)w^2)/(RTT * (w/2))$
= $(MSS/p) / (RTT*\sqrt{2/(3p)})$
= $(k*MSS) / (RTT*\sqrt{p})$

Source: (src: mathis et.al ccr'97)

RST-4512
9812_05_2004_c1

© 2004 Cisco Systems, Inc. All rights reserved.

22

TCP Dynamics: Assumption Flaws?

Cisco.com

- Assumed constant pkt-loss-probability “p”
- But if we raise MTU by 100x, the pkts are 100-times “longer”
- If we’re working against BER loss limits, this is a “big deal”
- Example: 1500B MTU at $p=0.01$, then 150,000B MTU $\rightarrow p=1.0$ (!!)

RST-4512
9812_05_2004_c1

© 2004 Cisco Systems, Inc. All rights reserved.

23

TCP Dynamics: Reworked Assumptions/Math

Cisco.com

- Replace p by $8 \cdot \text{MTU} \cdot \text{BER}$
- then $\text{Rate} = (\text{MTU}/\text{RTT}) \cdot C/\text{sqrt}(8 \cdot \text{MTU} \cdot \text{BER})$
- or
- $\text{Rate} = C/\text{RTT} \cdot \text{sqrt}(\text{MTU}/(8 \cdot \text{BER}))$
- New result: Throughput varies as $\text{sqrt}(\text{MTU})$, rather than linear; so 2x throughput via 4x MTU increase (instead of previous 2x MTU)
- (This is a new (2003) result, and somewhat disturbing, since the linear-with-MTU model has been accepted/quoted/used since 1997...)

RST-4512
9812_05_2004_c1

© 2004 Cisco Systems, Inc. All rights reserved.

24

TCP DYNAMICS ISSUES AND POSSIBLE APPROACHES FOR IMPROVEMENT



RST-4512
9812_05_2004_c1

© 2004 Cisco Systems, Inc. All rights reserved.

25

TCP Dynamics: Issues

Cisco.com

- **Standard TCP can't easily distinguish between congestion-induced loss, and corruption/error-induced loss**
- **For congestion, must back-off to maintain fairness**
- **For corruption/errors, would *like* to just keep transmitting, or recover more quickly**
- **Current reality: React to both congestion- and error-induced loss by backing-off**

RST-4512
9812_05_2004_c1

© 2004 Cisco Systems, Inc. All rights reserved.

26

Possible Approaches

Cisco.com

- To (re-) climb throughput curve more quickly
- By larger MTU (climb by more bytes-per-RTT because packets are larger)
- By changing TCP dynamics (change how we climb the curve, but maintain fairness)
- Maybe combination of both

RST-4512
9812_05_2004_c1

© 2004 Cisco Systems, Inc. All rights reserved.

27

Raise-the-MTU Approach

Cisco.com

- Observation that MTU is in numerator of upper-bound formula
- So, if we can raise the MTU, the bytes-per-RTT will climb faster (e.g. by 9k per ACK, vs. 1.5k/Ack)
- Not to mention reduced CPU load, etc.
- Issue: Routers trending away from large data-path buffers

RST-4512
9812_05_2004_c1

© 2004 Cisco Systems, Inc. All rights reserved.

28

Raise-MTU Possible Experimental Approach

Cisco.com

- Difficult to test >9kbyte MTU across real networks
- Perhaps construct experimental fragmentation/reassembly and Forward-Error-Correction (FEC) device?
- Present 65k MTU to host (which physical I/f?)
- And use standard 1500byte MTU for transmission
- Make device+internet a “black-box” that transparently(?) supports very large MTU
- Incremental testing/deployment
- FEC to attack Bit-error-rate loss
- Credits: Cisco colleagues (Fedorkow, Wakerly)

RST-4512
9812_05_2004_c1

© 2004 Cisco Systems, Inc. All rights reserved.

29

Change-TCP: Behavior Approach

Cisco.com

- Rather than (necessarily) change MTU, change the TCP algorithm to climb the rate-curve faster
- (when safe/fair to do so—i.e. co-exist with deployed TCPs)
- Also, perhaps change the reduction amount when congestion/loss is detected
- In other words, change/adapt the shape of the “sawtooth”
- Same/similar end-result as large MTU?

RST-4512
9812_05_2004_c1

© 2004 Cisco Systems, Inc. All rights reserved.

30

Change-TCP: Behavior Examples

Cisco.com

- **High-Speed-TCP (table of values for AIMD increments in different regimes)**
www.icir.org/floyd/hstcp.html
- **Scalable TCP (formula approach to modify AIMD)**
www-lce.eng.cam.ac.uk/~ctk21/scalable/
- **FAST (different transfer function when controlling for queuing)**
netlab.caltech.edu/FAST/
netlab.caltech.edu/pub/papers/draft-jwl-tcp-fast-01.txt
- **XCP (add in-pkt data, routers can modify; cisco has prototype in development w/ Braden et.al.)**
www.ana.lcs.mit.edu/dina/XCP/

RST-4512
9812_05_2004_c1

© 2004 Cisco Systems, Inc. All rights reserved.

31

MEASUREMENTS AND EXPERIMENTAL RESULTS



RST-4512
9812_05_2004_c1

© 2004 Cisco Systems, Inc. All rights reserved.

32

Measurements/Results (1) Resources

Cisco.com

- **Sally Floyd**
www.icir.org/floyd/hstcp.html
- **Tom Dunnigan—ORNL**
www.csm.ornl.gov/~dunigan/net100/
www.csm.ornl.gov/~dunigan/net100/auto.html
- **SLAC**
www-iepm.slac.stanford.edu/monitoring/bulk/fast/
www-iepm.slac.stanford.edu/monitoring/bulk/fast/stacks.png

RST-4512
9812_05_2004_c1

© 2004 Cisco Systems, Inc. All rights reserved.

33

Measurements/Results (2)

Cisco.com

- **Dunnigan**
9k yields 6x improved throughput for stock TCP
- **SLAC**
e.g. 1gbps trans-atlantic: Stock climbs slowly to 200mbps, HS-TCP get to 900mbps quickly
Also observe 5x improvement w/ 9k vs. 1.5k in stock TCP (single stream); light (9k) vs. moderate (1.5k) cpu loading; striping effective, also

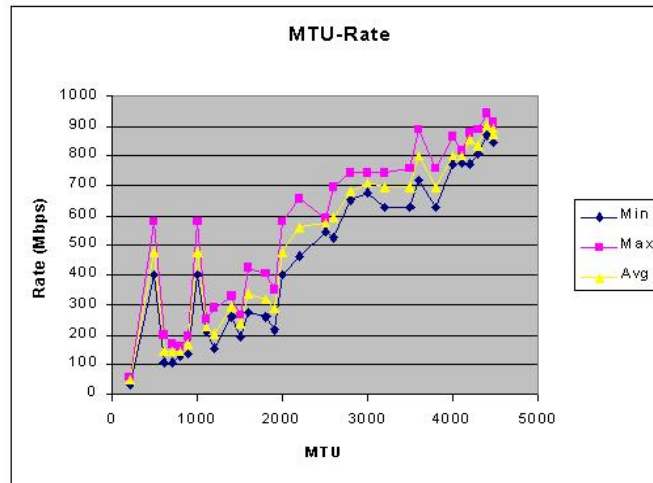
RST-4512
9812_05_2004_c1

© 2004 Cisco Systems, Inc. All rights reserved.

34

Throughput Sample Results (rreddy/PSC)

Cisco.com



Dual 1GHz pentium
1Gbps 10msec path
Linux 2.4.19 w/
web100

(Note: Not stated
whether this is
CPU/bus-bound, or
TCP cong. avoid
phenomenon)

Source: www.psc.edu/~rreddy/networking/mtu.html

RST-4512
9812_05_2004_c1

© 2004 Cisco Systems, Inc. All rights reserved.

35

Comparisons Among Methods

Cisco.com

- Following slides compare scalable, FAST, HS-TCP, “standard” TCP, and large-MTU
- Credits: From folks at Stanford Linear Accelerator (SLAC)
- Transatlantic (Sunnyvale, CA to CERN)
- See also:
www-iepm.slac.stanford.edu/monitoring/bulk/fast/
- (Caveat: Doug Leith, co-author of H-TCP, cautions that these may test driver efficiency more than congestion algorithm behavior)

RST-4512
9812_05_2004_c1

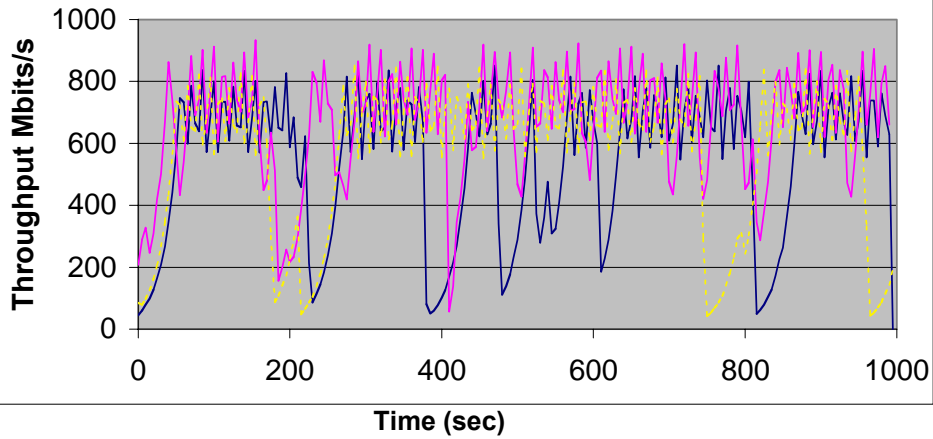
© 2004 Cisco Systems, Inc. All rights reserved.

36

Scalable_TCP results

Cisco.com

SNV-GVA: Scalable, txq=100



RST-4512
9812_05_2004_c1

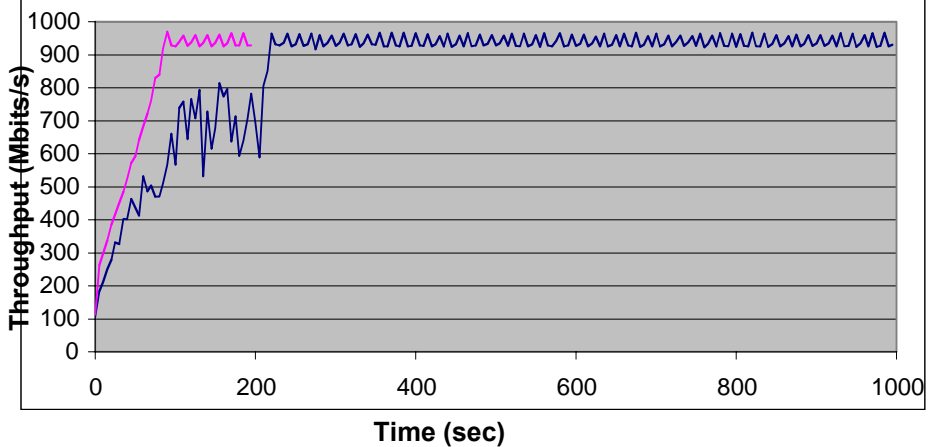
© 2004 Cisco Systems, Inc. All rights reserved.

37

FAST results

Cisco.com

FAST, txq=100, Repeat:



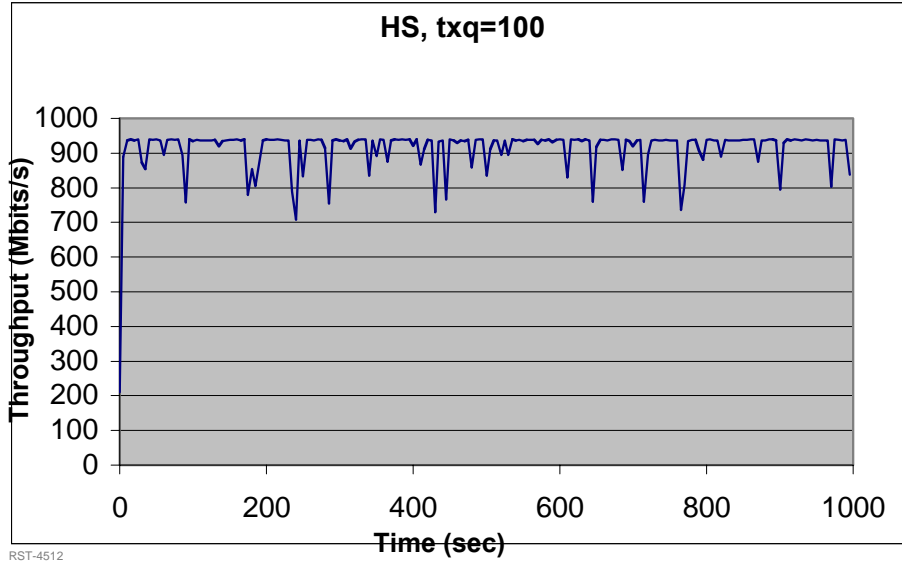
RST-4512
9812_05_2004_c1

© 2004 Cisco Systems, Inc. All rights reserved.

38

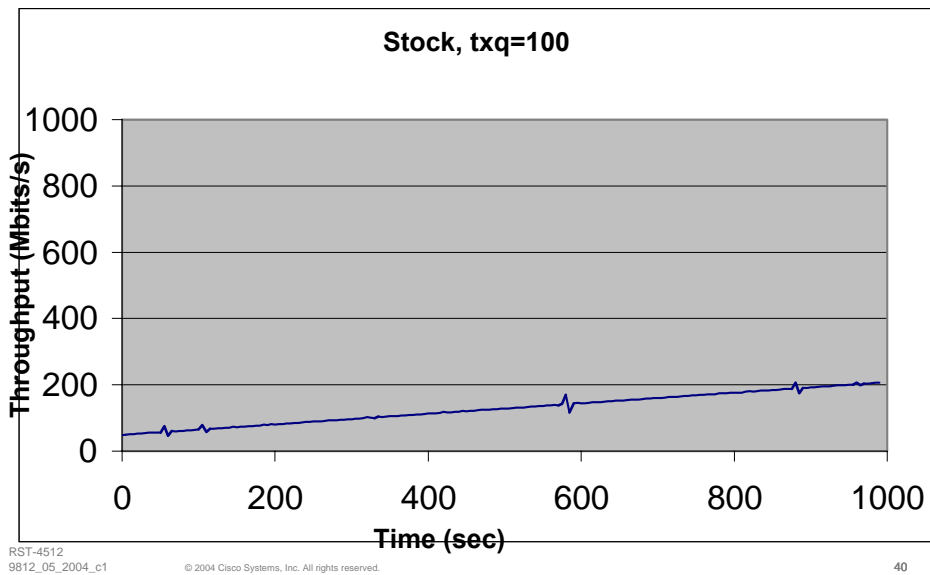
HS-TCP results

Cisco.com



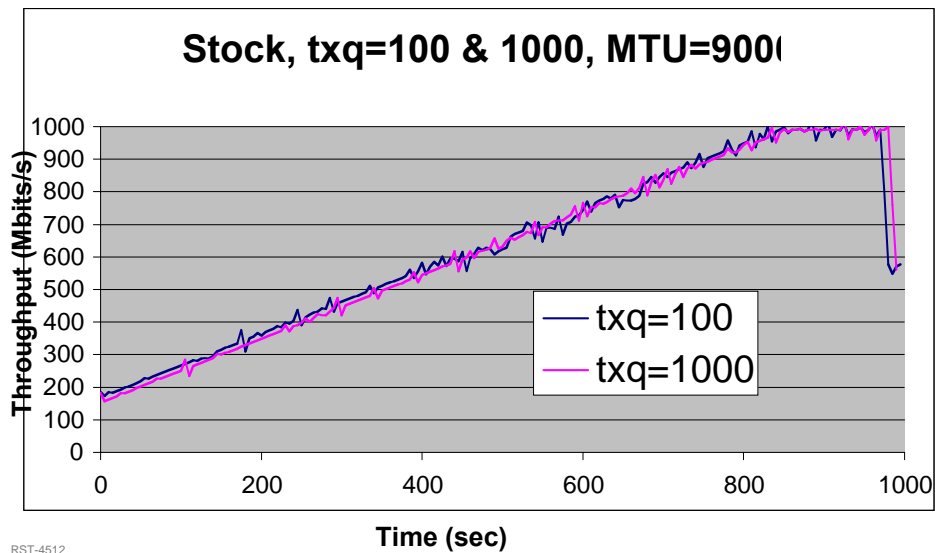
Standard TCP results

Cisco.com



Large-MTU results

Cisco.com



Non-TCP Methods

Cisco.com

Example: Tsunami

- Developed by Indiana-U folks when they were having trouble w/ standard TCP
- www.anml.iu.edu/anmlresearch.html
- Parallel UDP streams, TCP for backfill
- Effective, but very TCP-unfriendly (unfair)

RST-4512
9812_05_2004_c1

© 2004 Cisco Systems, Inc. All rights reserved.

42

Recent Results

Cisco.com

- In addition to FAST, HS-TCP, Scalable-TCP, XCP
- Several other experimental variants (H-TCP, x-LP)
- Also renewed interest in rate-based transport over UDP
- Integration of experimental stacks into web100 kernels (/proc variable to select stack)
- Results/presentations from Feb-2004 “Protocols for Long-Distance Nets” (PFLDNet) workshop at: [www-didc.lbl.gov/PFLDnet2004/program.htm](http://www.didc.lbl.gov/PFLDnet2004/program.htm)

RST-4512
9812_05_2004_c1

© 2004 Cisco Systems, Inc. All rights reserved.

43

CASE STUDIES



RST-4512
9812_05_2004_c1

© 2004 Cisco Systems, Inc. All rights reserved.

44

Case Studies (1): Hong Kong ↔ Los Angeles

Cisco.com

- Thanks to Charles Choy (HK)
- Application: Medical images from LA → HK
- 45Mbps pipe, getting 1–3mbps
- Added web100 in HK, diagnosed via host in Chicago
- Result: 25–35mbps easily achieved
- Root cause: Receive buffers too small; also, one FTP server did not negotiate window_scaling, so although buffers had been set “large enough,” TCP was not using them

RST-4512
9812_05_2004_c1

© 2004 Cisco Systems, Inc. All rights reserved.

45

Case Studies(2): Renater (France) ↔ FermiLab (Chicago)

Cisco.com

- Thanks to Francois-Xavier Andreu (Renater)
- Application: SDSS (astronomy) images from FermiLab (Chicago) → France; using “rsync” server
- 100mbps end-to-end, getting 4–7mbps
- Added web100 in France, diagnosed via host near Chicago
- Result: 90+mbps easily achieved (nttcp, mem → mem)
- Root cause: Rsync server in Chicago set up with “small” TCP buffers; client has same issue; secondary: looking at host i/o, and application (L7) buffering issues
- Side issues: Folks running data servers are not necessarily “data network” experts
- (This one still in progress, but diagnosis is solid)

RST-4512
9812_05_2004_c1

© 2004 Cisco Systems, Inc. All rights reserved.

46

Conclusions

- **Take-away tools/methods**

Practical, free tools exist to help diagnose and improve performance (e.g. web100)

Memory-memory tests (e.g. iperf, nttcp) can help establish baseline TCP performance

After addressing TCP bottlenecks (e.g. buffers), may need to look at system i/o, application-buffer issues

Need continued focus on bridging gaps among {application, networking, data_center} folks

- **Other items to consider:**

Improvements from increased MTUs and modified TCPs are real, measured

Some schemes require router participation (e.g. MTU → buffers, XCP → multipliers); consider deployability and fairness if you're involved in stack development/selection

Q AND A



Complete Your Online Session Evaluation!

Cisco.com

- WHAT:** Complete an online session evaluation and your name will be entered into a daily drawing
- WHY:** Win fabulous prizes! Give us your feedback!
- WHERE:** Go to the Internet stations located throughout the Convention Center, or <http://www.networkers04.com/desktop> from your laptop
- HOW:** Winners will be posted on the onsite Networkers Website; four winners per day

RST-4512
9812_05_2004_c1

© 2004 Cisco Systems, Inc. All rights reserved.

49

CISCO SYSTEMS



RST-4512
9812_05_2004_c1

© 2004 Cisco Systems, Inc. All rights reserved.

50