Tuning TCP Parameters
for the 21st Century

H.K. Jerry Chu
hkchu@google.com

Parameters to Examine

- init RTO (for 3WHS and init data transmission)
- initcwnd (IW) and/or restart cwnd (RW)
- min RTO
- Delayed ack timer

InitRTO - RFC1122

…

The following values SHOULD be used to initialize the estimation parameters for a new connection:

(a) RTT = 0 seconds.

(b) RTO = 3 seconds. (The smoothed variance is to be initialized to the value that will result in this RTO

...

DISCUSSION:

Experience has shown that these initialization values are reasonable, and that in any case the Karn and Jacobson algorithms make TCP behavior reasonably insensitive to the initial parameter choices.

Proposed Change

The following values SHOULD be used to initialize the estimation parameters for a new connection:

(a) RTT = 0 seconds.
(b) RTO = 1 second.

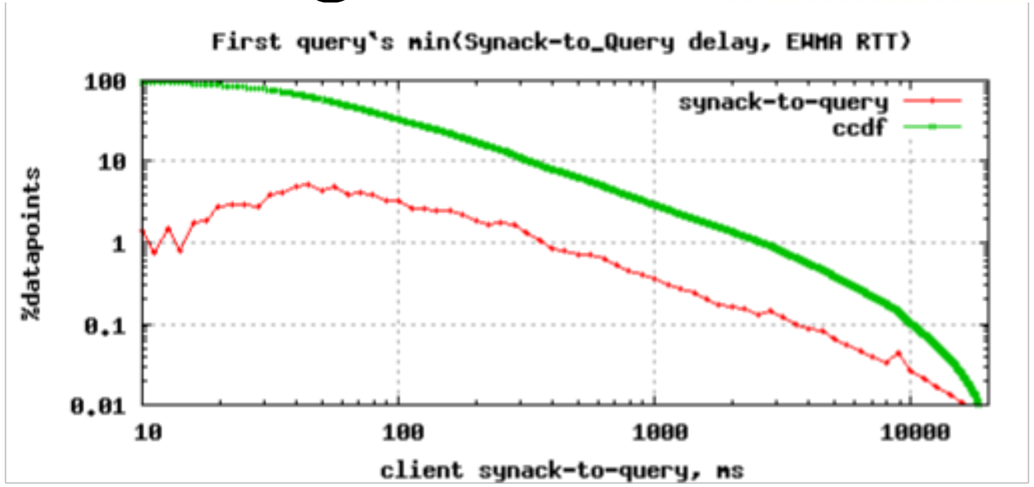
Before the three-way handshake is complete, upon the first retransmission timer expiration, the next RTO SHOULD remain as calculated above. Upon the second retransmission timer expiration, the RTO MUST be calculated per RFC 1122. Thus the retransmission timeout does not follow "exponential backoff" until the second retransmit. The pattern with an initial RTO of 1 second is,

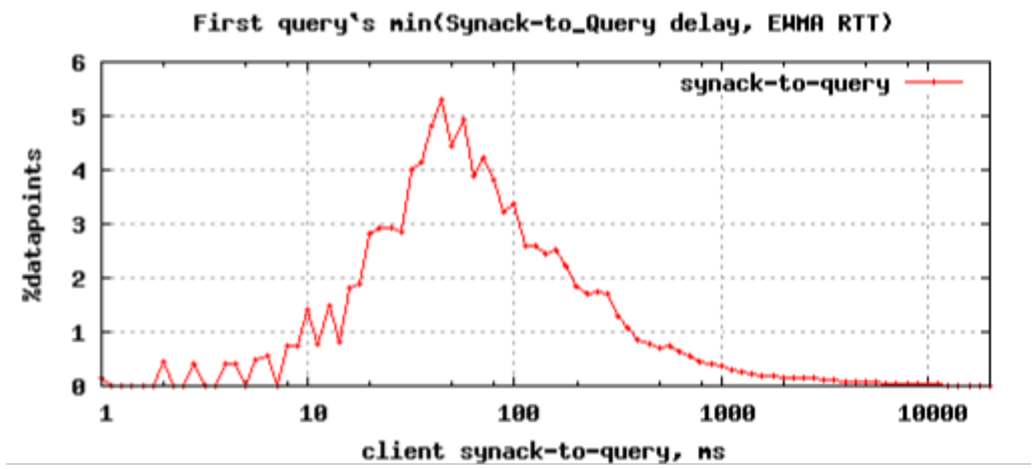
1s, 1s, 2s, 4s, ...

Init RTO in OSes

Operating System	SYN RTO (seconds)	SYN-ACK RTO (seconds)
FreeBSD 7.1	3, 6, 12, ...	3, 6, 12, ...
Solaris 10	3.38, 6.76, 13.52, ...	3.38, 6.76, 13.52, ...
Windows XP	3, 6	3, 6, 12, …
Windows Vista	3, 6	3, 6, 12, …
Windows 7	3, 6, 12, …	3, 6, 12, …
Linux (all versions)	3, 6, 12, …	3, 6, 12, …
Mac OS X 10.5.6	1, 2, 4, ...	1, 1, 1, 1, 1, 2, 4, …

Google's World-Wide RTT Distribution





A pessimistic estimate of query RTT distribution (including retransmissions): ~2.5% connections with RTT > 1sec

Regional data for connections with > 1sec RTT:

Asia: 2.57%

U.S. west coast: 0.31 - 0.53%

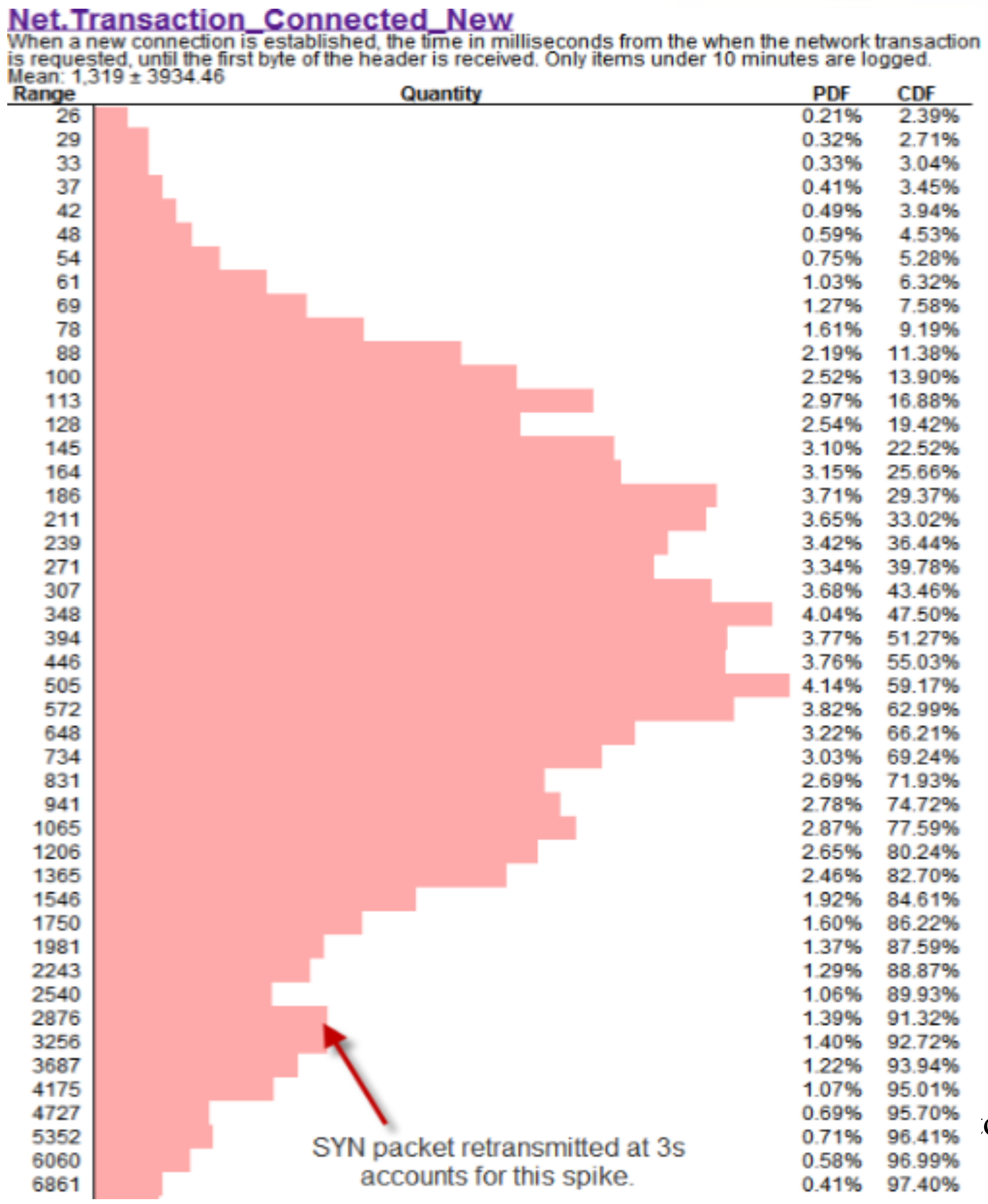
Europe: 0.79 - 1.37%

measured from client SYN to client ACK, excluding SYN but including SYN-ACK retransmissions

Packet Drop rate

- TCP retransmit rate: 0.8% - 2.4%
  - measured at Google's frontend servers
- SYN-ACK retransmit rate: 0.6% - 3.8%
  - measured from a different set of Google servers

SYN Retransmit rate



Net.Transaction_Connected_New

When a new connection is established, the time in milliseconds from the when the network transaction is requested, until the first byte of the header is received. Only items under 10 minutes are logged.

Mean: 1,319 ± 3934.46

Range	PDF	CDF
26	0.21%	2.39%
29	0.32%	2.71%
33	0.33%	3.04%
37	0.41%	3.45%
42	0.49%	3.94%
48	0.59%	4.53%
54	0.75%	5.28%
61	1.03%	6.32%
69	1.27%	7.58%
78	1.61%	9.19%
88	2.19%	11.38%
100	2.52%	13.90%
113	2.97%	16.88%
128	2.54%	19.42%
145	3.10%	22.52%
164	3.15%	25.66%
186	3.71%	29.37%
211	3.65%	33.02%
239	3.42%	36.44%
271	3.34%	39.78%
307	3.68%	43.46%
348	4.04%	47.50%
394	3.77%	51.27%
446	3.76%	55.03%
505	4.14%	59.17%
572	3.82%	62.99%
648	3.22%	66.21%
734	3.03%	69.24%
831	2.69%	71.93%
941	2.78%	74.72%
1065	2.87%	77.59%
1206	2.65%	80.24%
1365	2.46%	82.70%
1546	1.92%	84.61%
1750	1.60%	86.22%
1981	1.37%	87.59%
2243	1.29%	88.87%
2540	1.06%	89.93%
2876	1.39%	91.32%
3256	1.40%	92.72%
3687	1.22%	93.94%
4175	1.07%	95.01%
4727	0.69%	95.70%
5352	0.71%	96.41%
6060	0.58%	96.99%
6861	0.41%	97.40%

Connect data from Windows clients world-wide (collected through Google Chrome):

SYN retransmit rate is estimated at ~1.42% (extrapolating the curve and extracting the spike at 3secs)

Expected Gain

- Mainly benefit short-lived connections (e.g., HTTP/TCP) where 3WHS latency is significant
- For a route with packet drop rate of X%, average 3WHS completion time improves by 2*2000ms*X%
- E.g., a user accessing a web site 10ms away with packet drop rate of 1% will enjoy 40ms reduction in average latency!

Expected Cost

- Spurious SYN/SYN-ACK retransmissions
- May trigger early transition to congestion avoidance and fast retransmit (if > 2 rexmits, i.e. RTT > 1+1+2=4secs)
  - induce more duplicate packets
  - IW reduced to LW
  - ssthresh reduced to 1 or 2
  - no good RTT sample
- Need to detect spurious retransmission to undo the damage
  - TS or DSACK option can help filtering dupacks from spurious retransmissions

Related Ideas

- RTT history to the same destination (or subnet) may provide a better value than a blind 1 sec (see RFC2140)
  - only feasible on the server side
- Use RTT measured from 3WHS to set init data RTO
  - Difference in transmission delay among packets of different sizes may be significant for slow links

initcwnd/restart cwnd

- Increased from 1 to 2 after a much publicized specweb problem when sender and receive deadlock until delayed ack timer fires
- Increased again in RFC2414 (later RFC3390)

```
If (MSS <= 1095 bytes)
then win <= 4 * MSS;
If (1095 bytes < MSS < 2190 bytes)
then win <= 4380;
If (2190 bytes <= MSS)
then win <= 2 * MSS;
```

Pros and Cons of a Larger initcwnd

- Pros - cut down # of RTTs => improve user latency
  - increasing initcwnd from 3 to 4 reduces the network latency of Google's search queries by up to several percentage points
  - SDCH benefits more
- Cons – more congestion?
  - RFC3390 contains a detailed discussion
  - can base initcwnd on per-client history to mitigate some issue
  - will packet pacing help?
  - how far can we go?
- Any alternatives?
  - Fast Startup schemes still under research at iccrg

Change in HTTP Response Size

year	2000	2007
min	17B	85B
max	0.23GB	2.45GB
mean	12294	68275
median	2410	2780
SCV (squared coefficient of variation)	321	3425

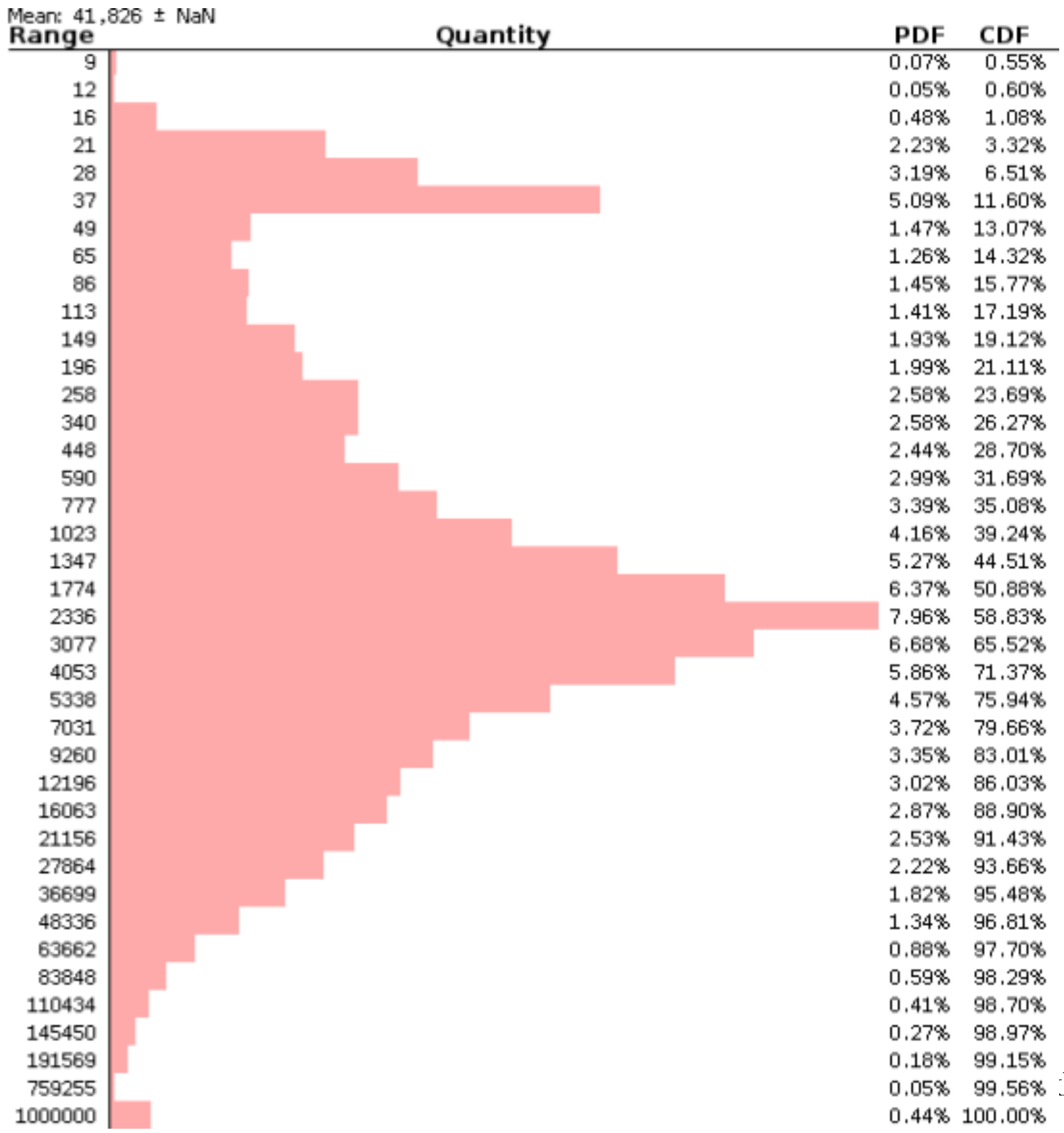
Data from www. websiteoptimization.com

Average size increased by 5.5x

Median grew only 15%

Long tail got even longer

HTTP Response Size Distribution



Mean: 41,826 ± NaN

Range	Quantity	PDF	CDF
9		0.07%	0.55%
12		0.05%	0.60%
16		0.48%	1.08%
21		2.23%	3.32%
28		3.19%	6.51%
37		5.09%	11.60%
49		1.47%	13.07%
65		1.26%	14.32%
86		1.45%	15.77%
113		1.41%	17.19%
149		1.93%	19.12%
196		1.99%	21.11%
258		2.58%	23.69%
340		2.58%	26.27%
448		2.44%	28.70%
590		2.99%	31.69%
777		3.39%	35.08%
1023		4.16%	39.24%
1347		5.27%	44.51%
1774		6.37%	50.88%
2336		7.96%	58.83%
3077		6.68%	65.52%
4053		5.86%	71.37%
5338		4.57%	75.94%
7031		3.72%	79.66%
9260		3.35%	83.01%
12196		3.02%	86.03%
16063		2.87%	88.90%
21156		2.53%	91.43%
27864		2.22%	93.66%
36699		1.82%	95.48%
48336		1.34%	96.81%
63662		0.88%	97.70%
83848		0.59%	98.29%
110434		0.41%	98.70%
145450		0.27%	98.97%
191569		0.18%	99.15%
759255		0.05%	99.56%
1000000		0.44%	100.00%

Data collected from Google Chrome (rough estimate with caveat!):

Median: ~2KB

Mean: ~41KB

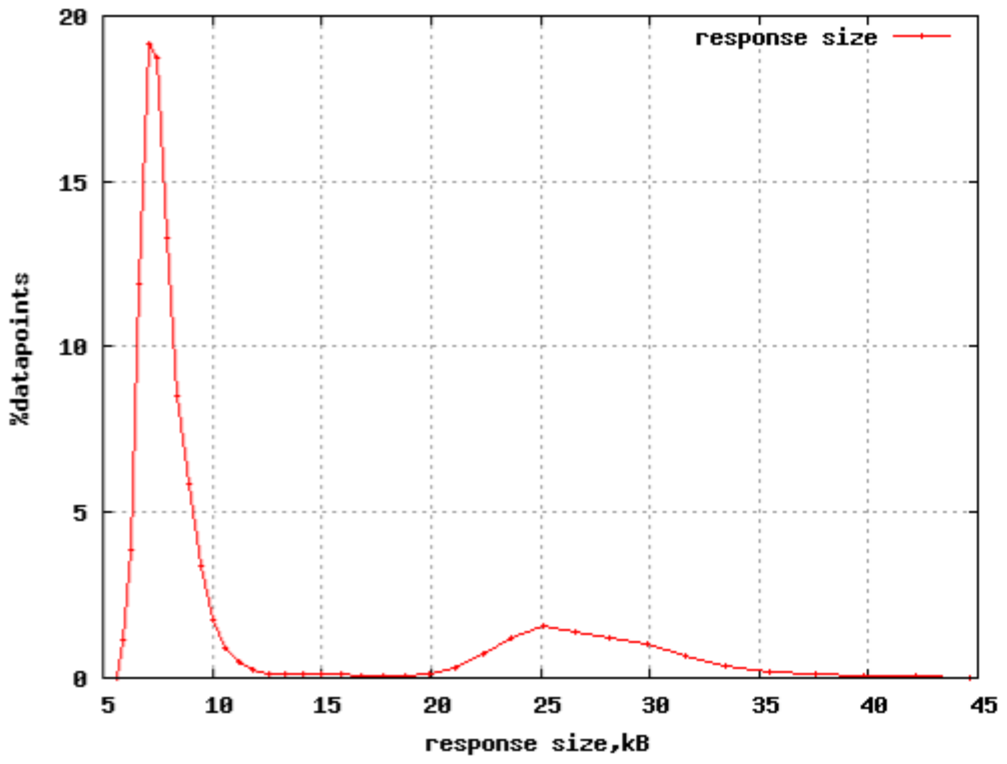
99th percentile mean: 8.1KB due to heavy tail (0.5% in the 1MB + bucket)

67.5% < 3*mss (4380)

73% < 4*mss

77% < 5*mss

Search Result Size Distribution



Data collected from one datacenter in Europe:

87% of search query results are < 10.5KB

(The 1st peak is ~7KB)

Acknowledgements

The following is a list of people who wrote the initial proposal and provided much of the precious data:
Mike Belshe, Andre Broido, Yuchung Cheng, Arvind Jain, Robert Love, Jim Roskind, Ricardo Vargas