Increasing TCP initial window

draft-hkchu-tcpm-initcwnd-01.txt

Nandita Dukkipati
Yuchung Cheng
Jerry Chu
Matt Mathis
{nanditad, ycheng, hkchu, mattmathis}@google.com

30 July, 2010
78th IETF, Maastricht

Overview of prior results for IW10

- Our proposal: increase TCP IW to 10 MSS
- IW10 improves average TCP latency by ~10%
- Large scale data-center experiments demonstrate latency improves across network and traffic properties:
  - Varying network BW, RTTs, BDP, HTTP response sizes, mobile networks
  - Small overall increase in retransmission rate (~0.5%), with most from multiple connections
- Prior work:
  - http://www.ietf.org/proceedings/10mar/slides/tcpm-4.pdf
  - http://ccr.sigcomm.org/online/?q=node/621

New contributions and the questions addressed

- A framework for running experiments with different IWs in the same data-center
- Primary concern from IETF-77: how does IW10 perform on highly multiplexed links such as in Africa and South America?
- What is the impact on latency due to losses in IW?
- Evaluated the impact of different IWs [3, 10, 16] on latency and retransmission rate
  - Reinforced the prior experiment results with IW10
- Testbed experiments for IW study in controlled environment
  - Preliminary results on fairness

Improved methodology for experiments

Previous methodology:

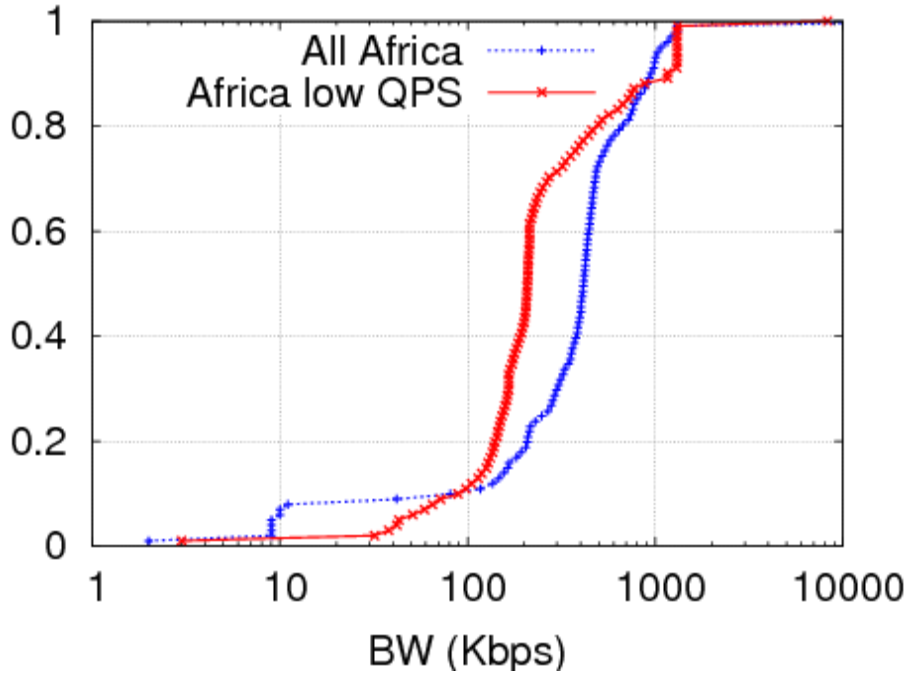
- Change IW for entire data-center every week
  - Less apples-to-apples: changes in server software and user base
  - Takes weeks to collect data

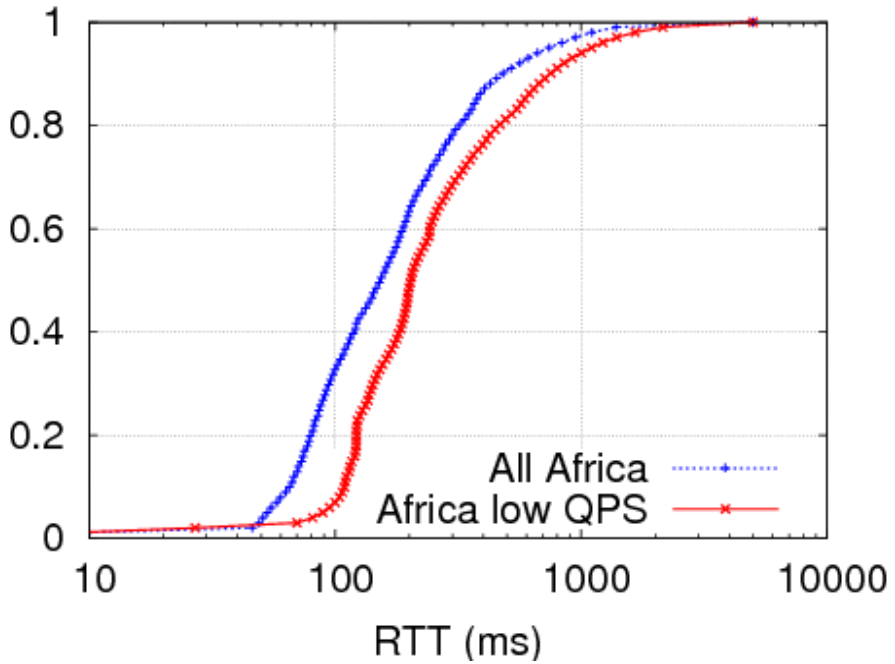
New methodology:

- Serve different IWs based on IP address in one data-center simultaneously for weeks
  - Same IW for connections from the same IP
  - More apples-to-apples: similar load across server software update and user churn

Analysis of IW10 on Africa traffic

Experiment for 1 week in June 2010





Impact of IW10 on Africa traffic

Web search latency (ms) and retransmission rate %

All of Africa

Percentile	Avg.	50	75	90	99
IW=10	988.4	503	795	1467	5042
IW=3	1123.9	538	878	1710	5923
Impr.	135.5	35	83	243	881
% Impr.	12%	6.5%	9.5%	14.2%	14.9%

	Retrans. %
IW=10	3.77%
IW=3	3.35%
Increase	0.42

Africa with low QPS

Percentile	Avg.	50	75	90	99
IW=10	1870.5	733	1363	3146	11579
IW=3	2340.7	857	1773	4110	14414
Impr.	470.2	124	410	964	2835
% Impr.	20.1%	14.5%	23.1%	23.5%	19.7%

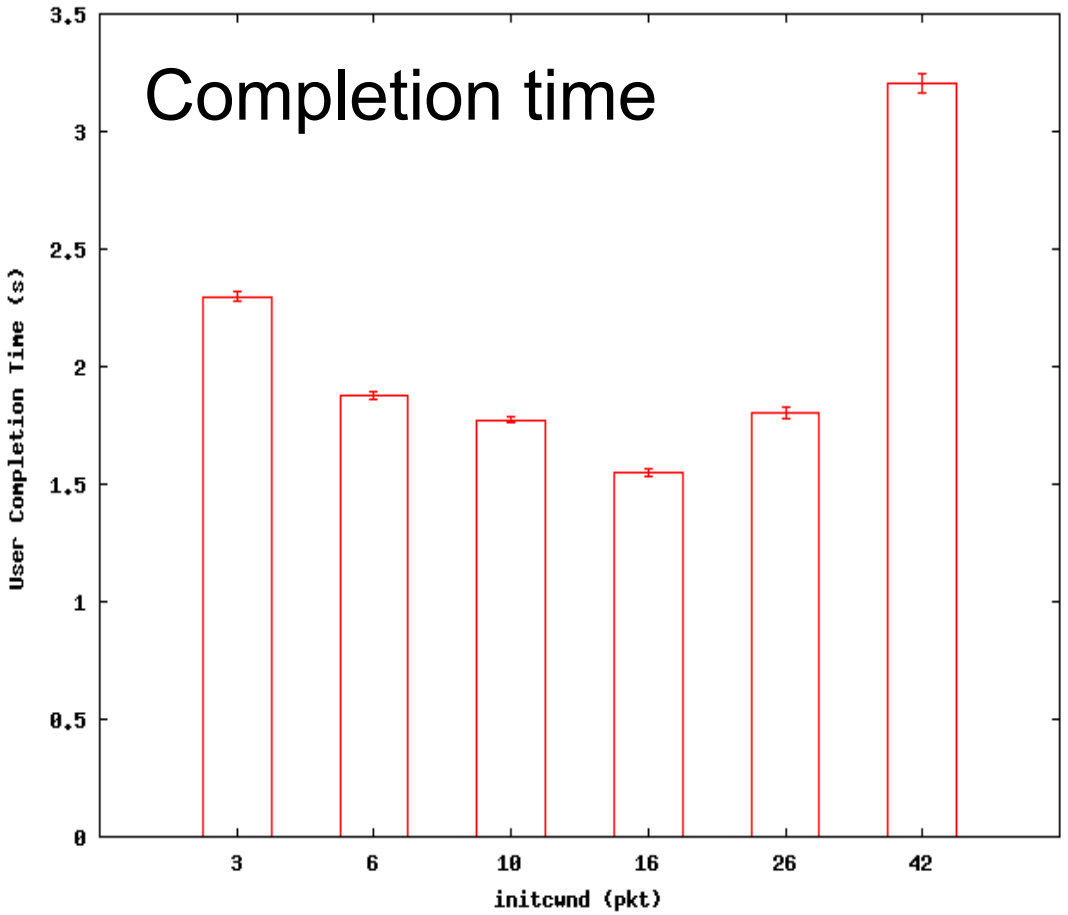
	Retrans. %
IW=10	6.71%
IW=3	5.83%
Increase	0.87

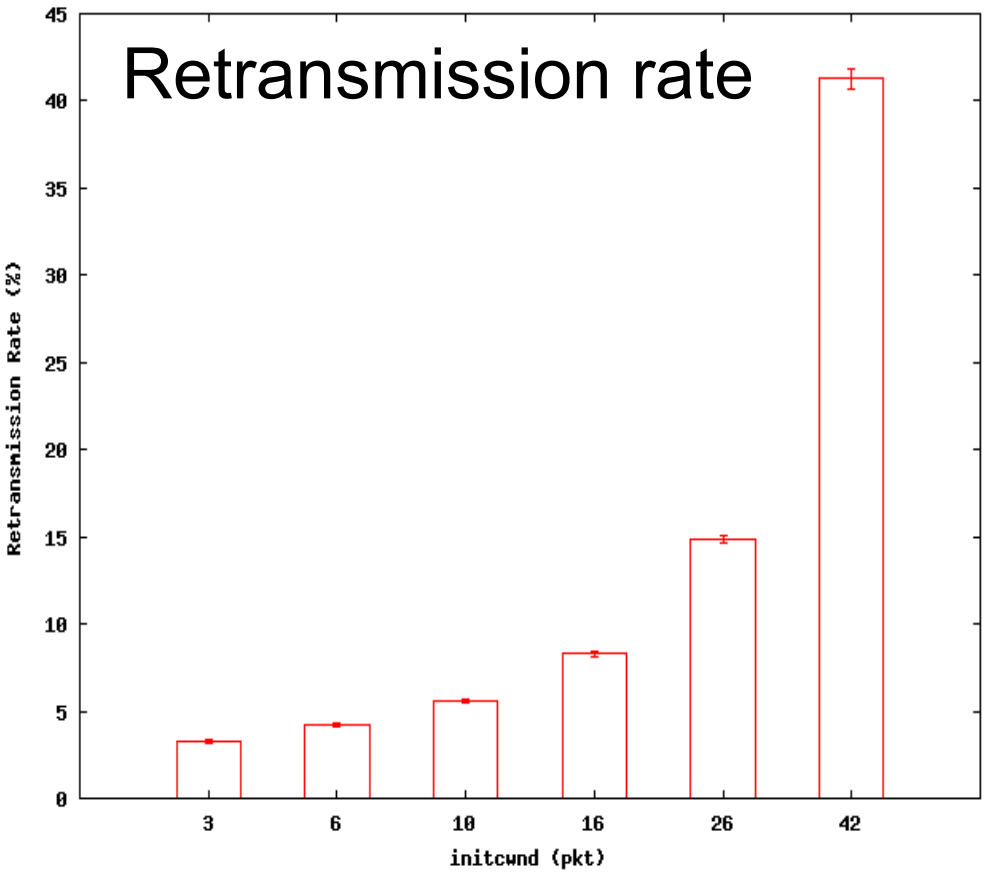
Why does latency improve in Africa?

- Large network round-trip time
- Larger IW helps faster recovery of packet losses
- Experiments on testbed demonstrate latency improves in spite of increased packet losses

Why does latency improve in Africa?

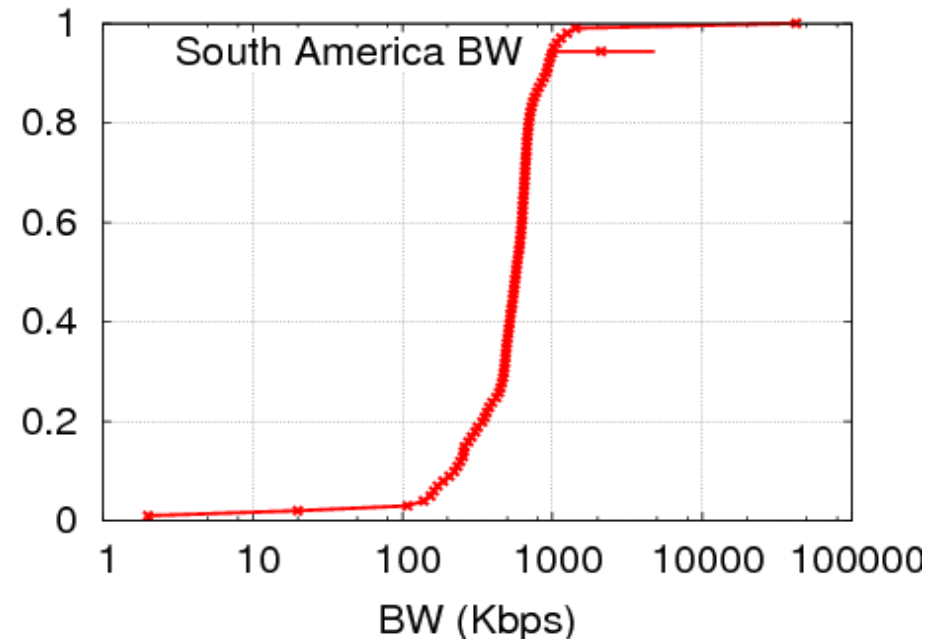
- Testbed experiment: 20Mbps, RTT 300ms, BDP buffer, offered load 0.95, 50KB response size
- Motivating example: Makerere University, Uganda

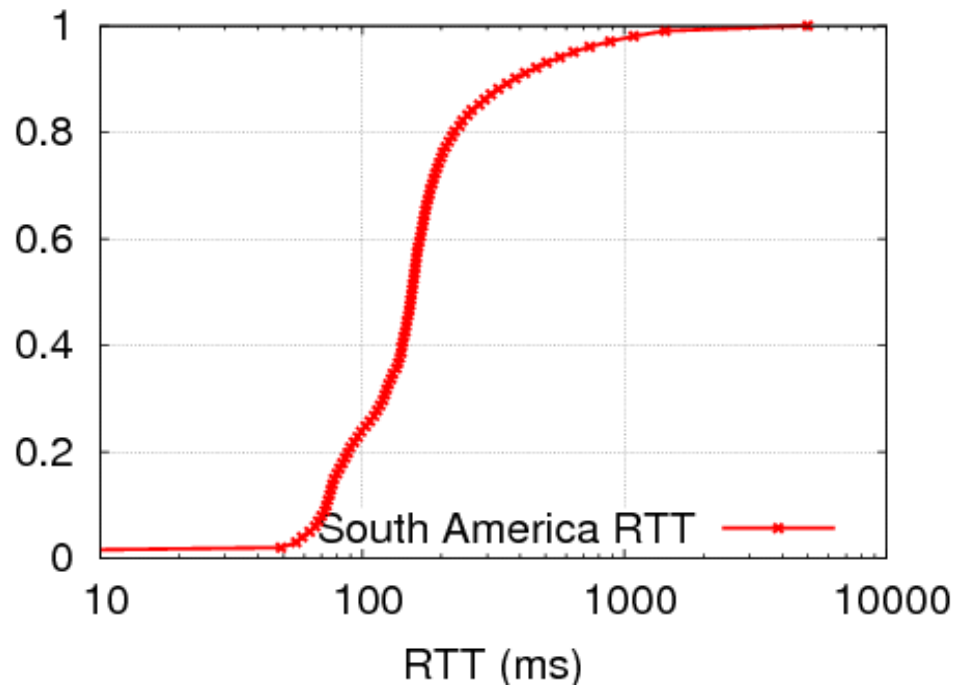




Analysis of IW10 on South America traffic

Experiment for 1 week in June 2010





Latency improvement across services in South America

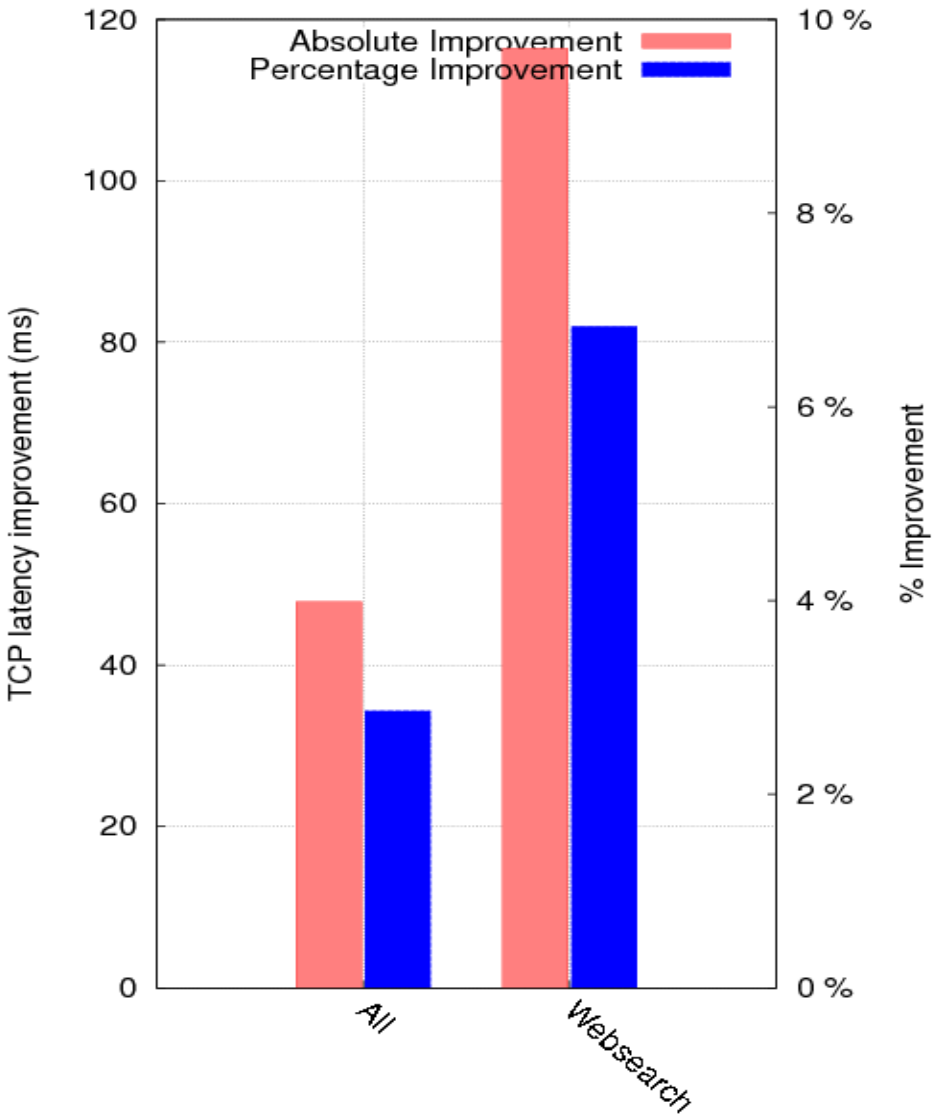
- Latency improves across a variety of services
- Services with multiple connections experience:
  - Least latency benefits
  - Most increase in retransmission rate

Percentile	Web	iGoogle	News	Blogger Photos (multiple connections)	Maps (multiple connections)
10	18 [6%]	30 [10%]	4 [2.5%]	2 [1.1%]	6 [3.8%]
50	38 [6.6%]	198 [26%]	45 [9.9%]	98 [12.7%]	12 [3.2%]
90	154 [11%]	430 [16%]	336 [15%]	251 [4.5%]	37 [2.6%]
99	561 [12%]	986 [9.7%]	1827 [19%]	691 [2.9%]	134 [2.9%]
Delta in Retrans %	0.51	0.52	0.35	2.93	1.28

entry: latency improvement (ms) [% improvement]

Impact of latency under packet losses

Latency of traffic with retransmissions > 0 improves with IW10 as compared to IW3



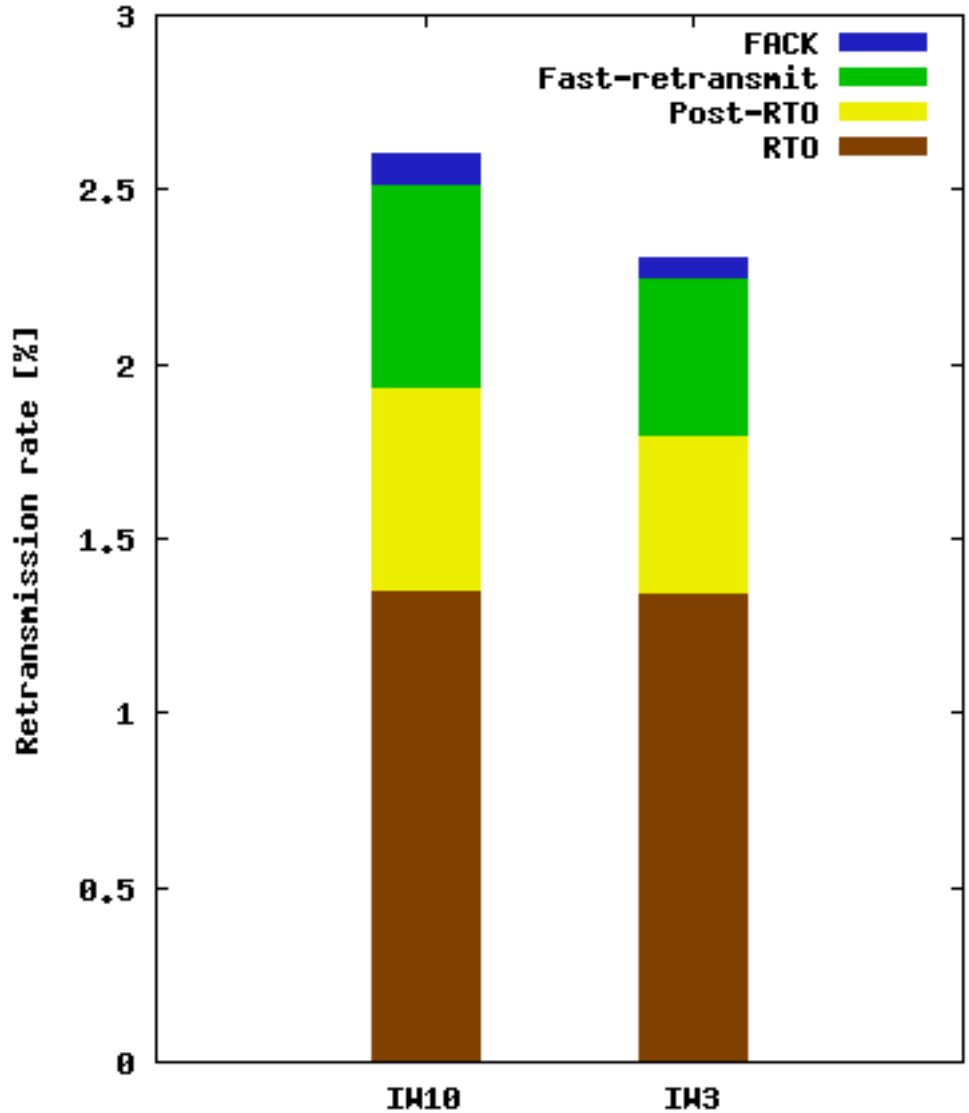
% traffic with rexmit > 0

	IW3	IW10
All	6.6%	6.8%
Web Search	6.11%	6.57%

Retransmissions of IW3 vs IW10

IW10 has ~0 increase in #timeouts, but has more

- fast-retransmit
- post-RTO retransmits



Experiments with higher IWs

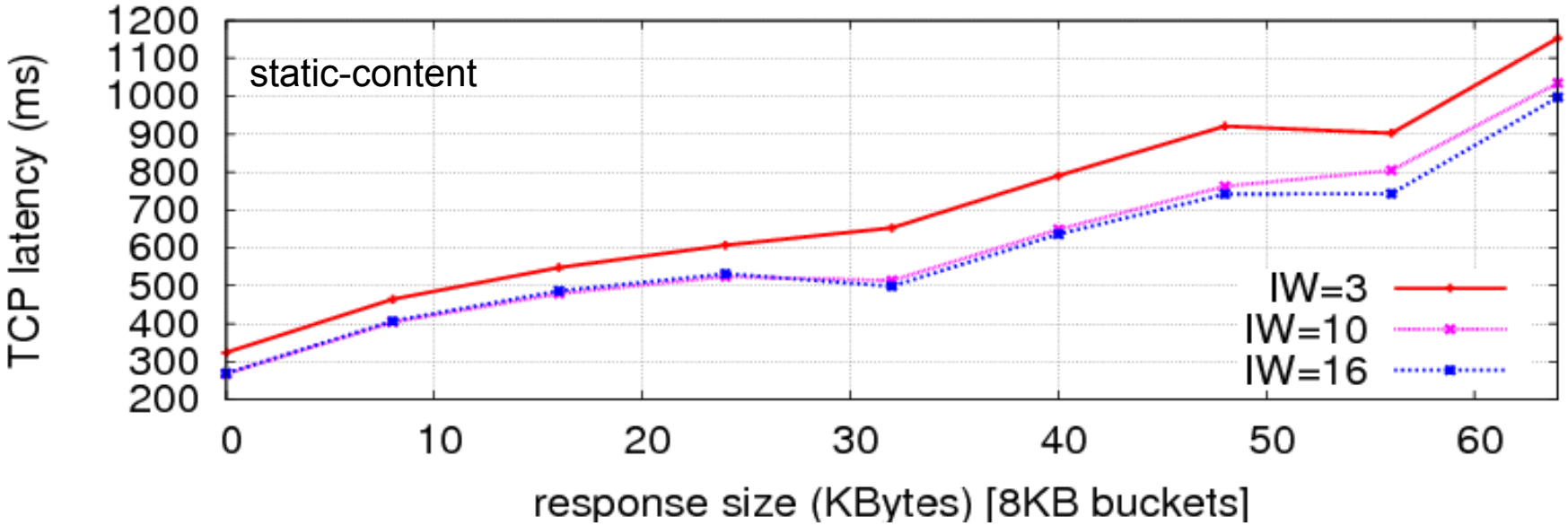
Does IW > 10 show better latency?

Try IW = {3, 10, 16} at

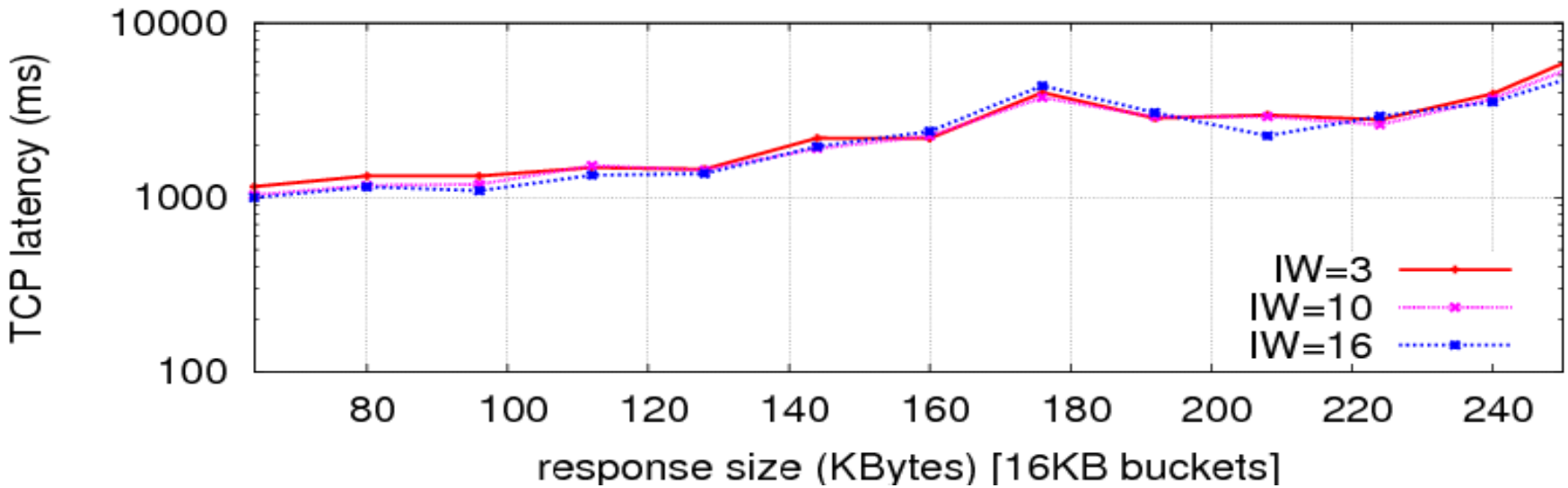
- DC 1
  - 20% in US east coast (RTT < 100ms)
  - 80% in south America (RTT > 100ms)
- DC 2
  - 97% in Europe (RTT < 100ms)

Comparison of IW = 3, 10, 16 (DC 1)

Small improvement for larger IWs (>10); mostly for mid-size flows



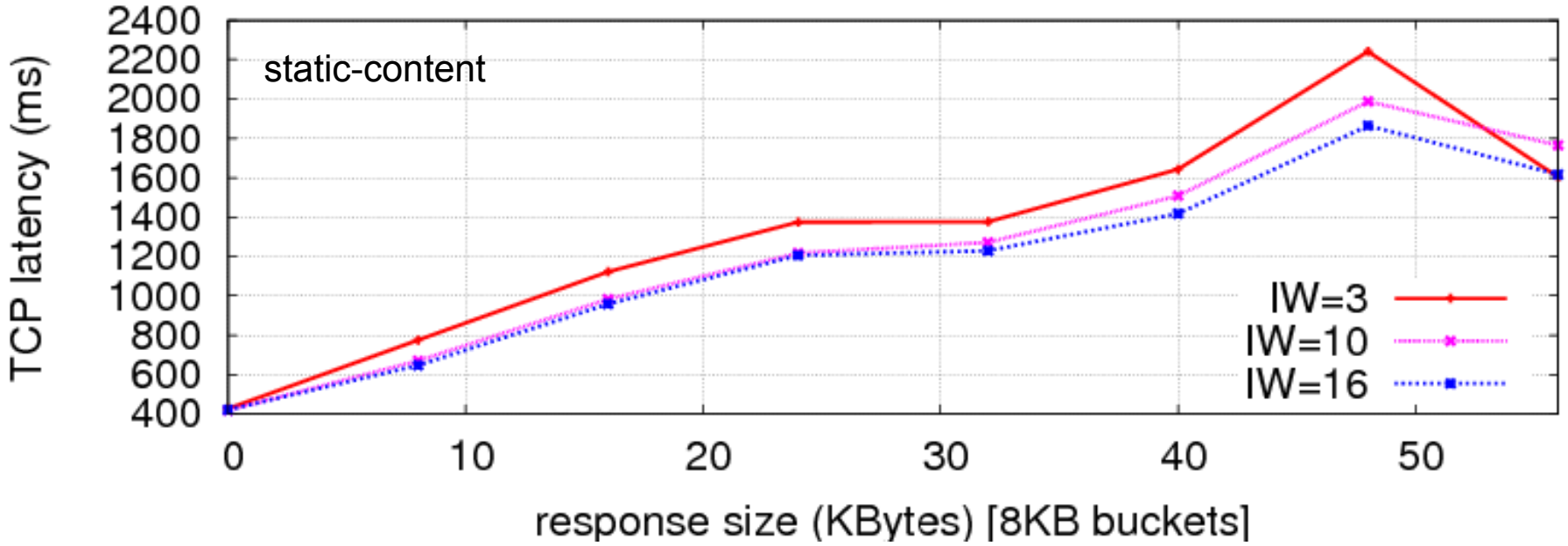
Small/Mid:
0-64KB

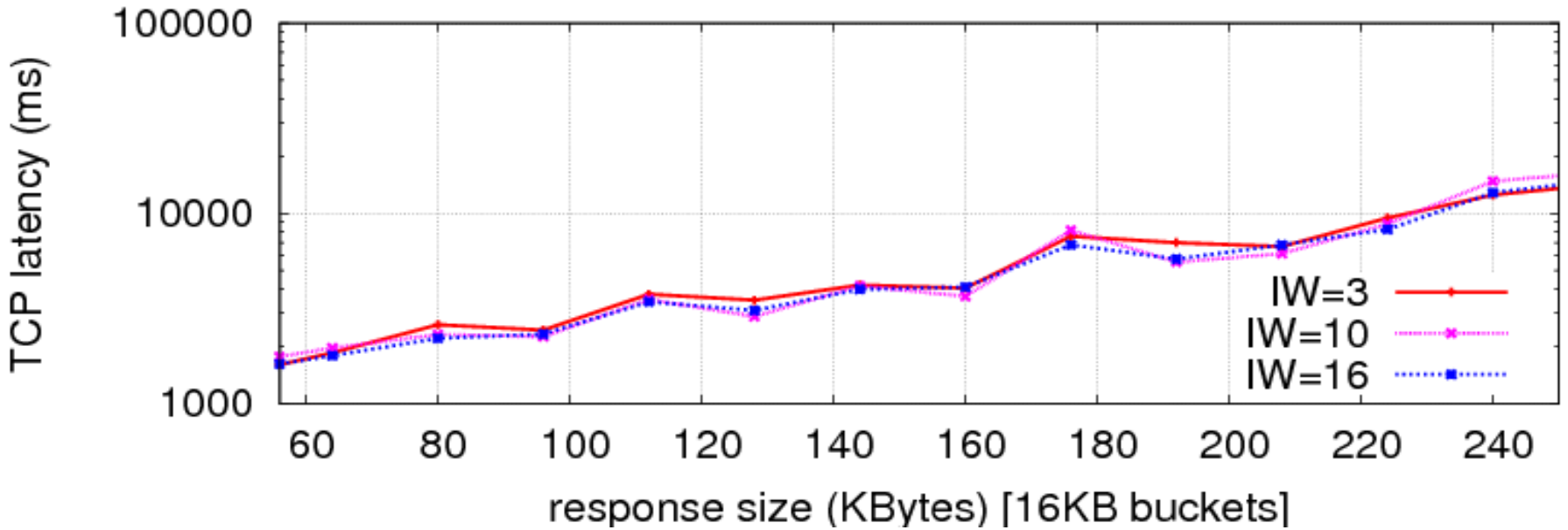


Large:
>64KB

Comparison of IW = 3, 10, 16 (DC 2)

Small improvement for larger IWs (>10); mostly for mid-size flows





Testbed topology

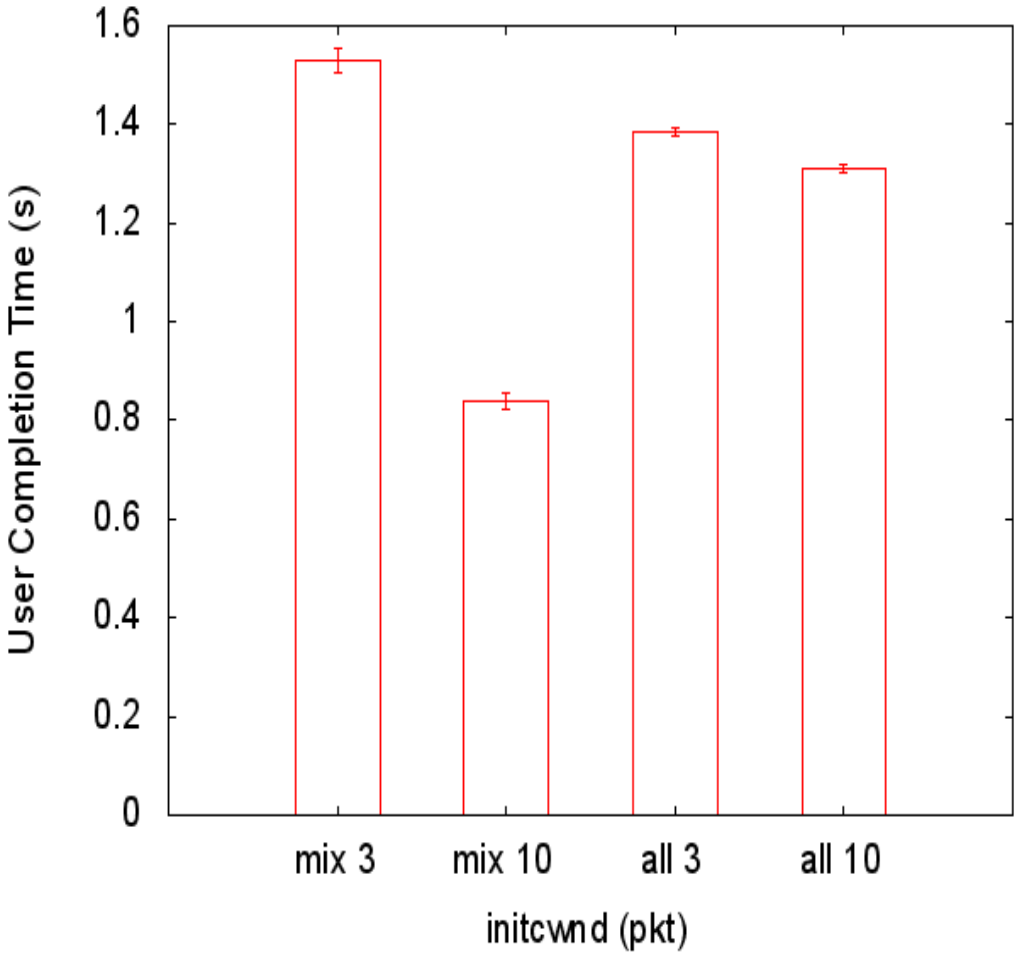
All results are preliminary!

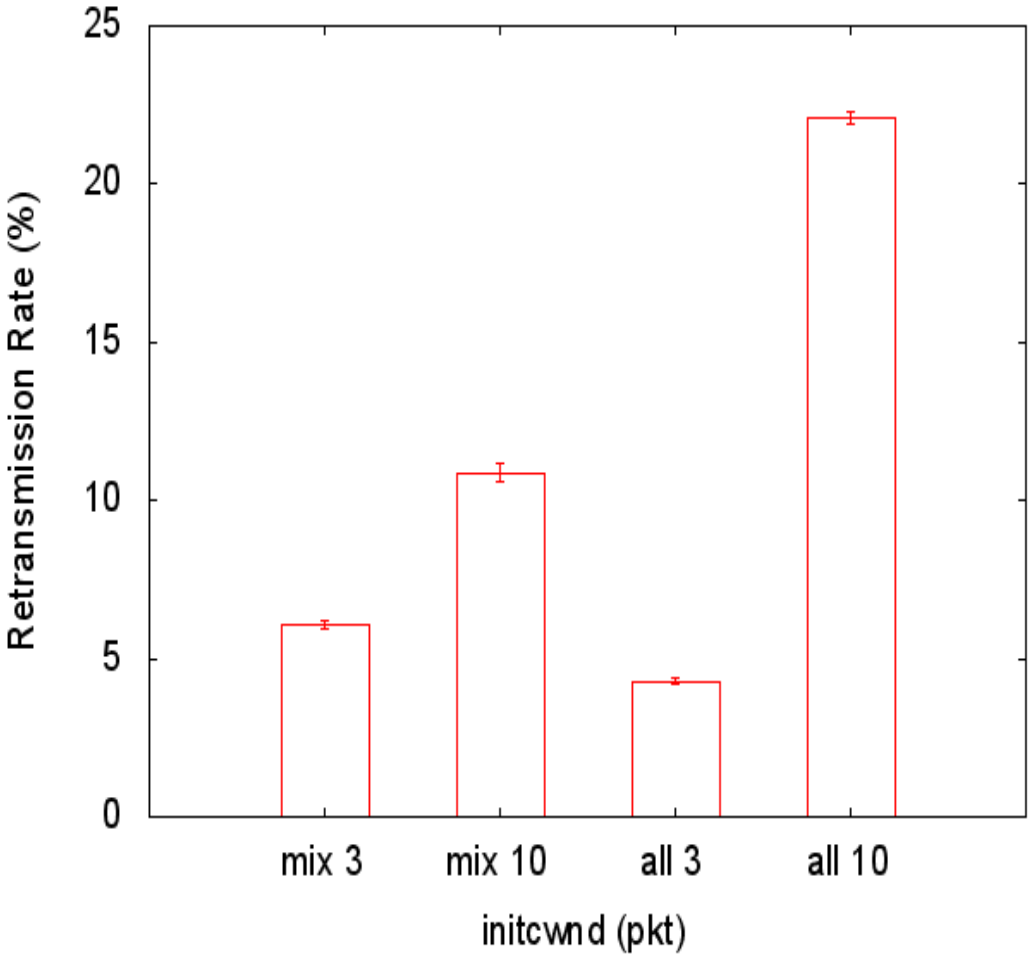
Linux box as a router
with netem to emulate
N buffers, 300ms RTT
20Mbps bottleneck

- Traffic generator – enhanced netperf dispatched based on poisson arrival
- Offered load - # of conn/sec (λ) with fixed response size, no pipelining
- Tests parameters - bottleneck b/w, RTT, buffer space, response size
- Test metrics - user completion time (UCT), retransmission rate, link utilization
- Measurement & Diagnosis tools

Fairness between IW10 and IW3 flows

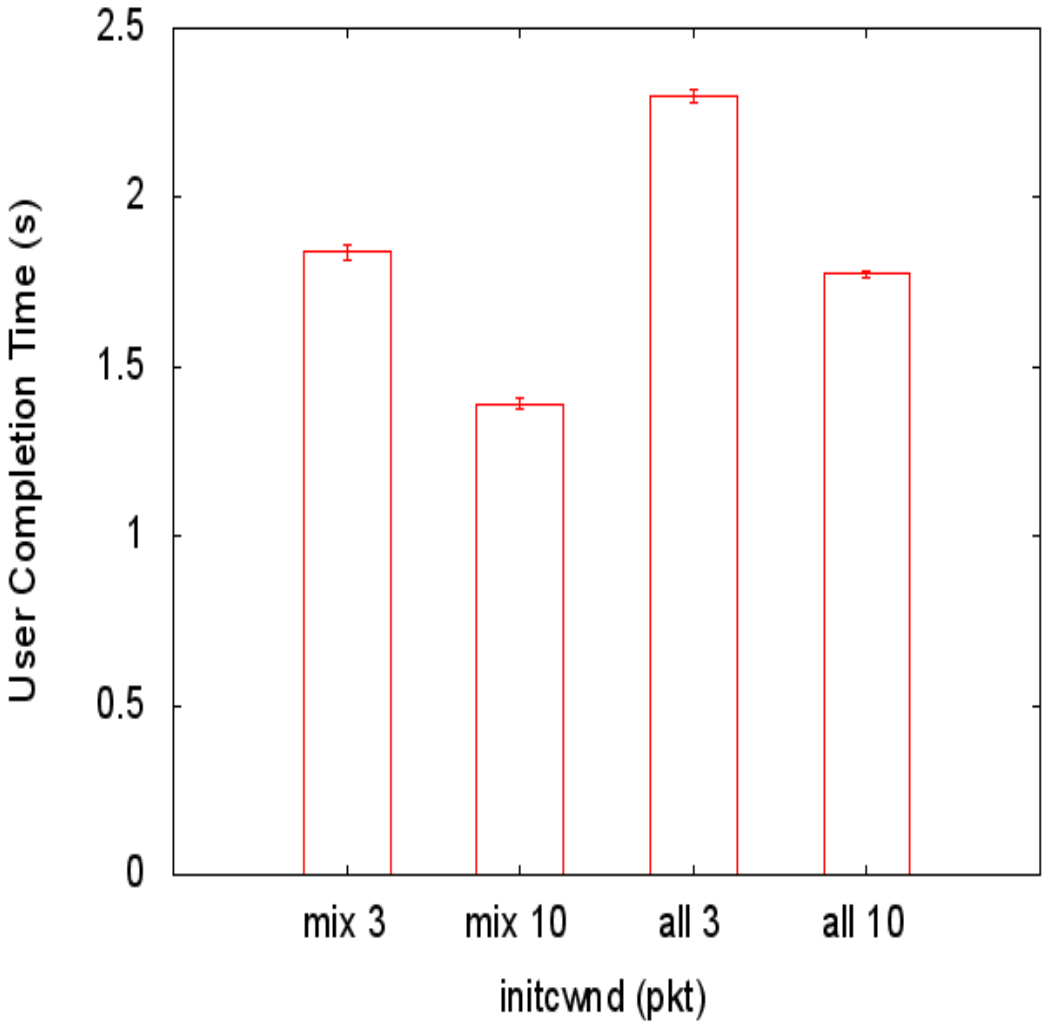
50/50 mix of IW3 and IW10 traffic
BDP buffer, load 0.95, 15KB response size

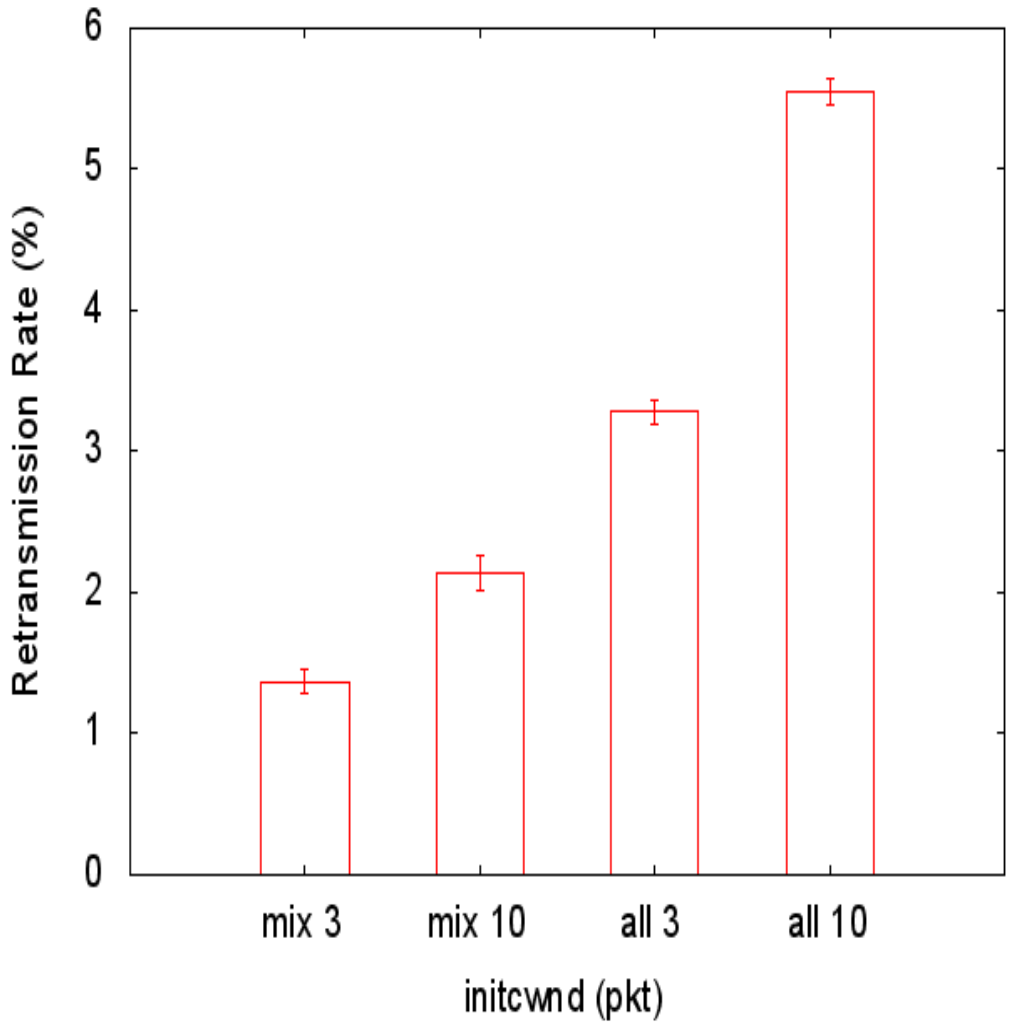




Fairness between IW10 and IW3 flows

Same as previous slide except response size is 50KB





Conclusion

- Take away summary
  - IW10 improves latency even in Africa and South America
  - IW10 helps in quicker recovery from packet losses
    - A higher retransmission rate does not necessarily increase latency
  - IW16 shows marginal latency improvement over IW10
- Next steps
  - Ongoing work: fairness between IW3 and IW10 in the transition phase
  - For any pending issues with IW10, join us in solving the problems!

Steps to configure IW on Linux

Changing TCP IW on Linux (kernel version >= 2.6.30)

On your server, do
$ ip route show

select the outgoing route then do
$ ip route change default via <gateway> dev eth0 initcwnd <iw>

If the server process explicitly set SNDBUF, then SNDBUF value >= IW*MSS. Otherwise increase the initial socket buffer if IW*MSS > /proc/sys/net/ipv4/tcp_wmem[1]

$ cat /proc/sys/net/ipv4/tcp_wmem
4096   16384   4194304
$ echo '4096 IW*MSS 4194304' > /proc/sys/net/ipv4/tcp_wmem

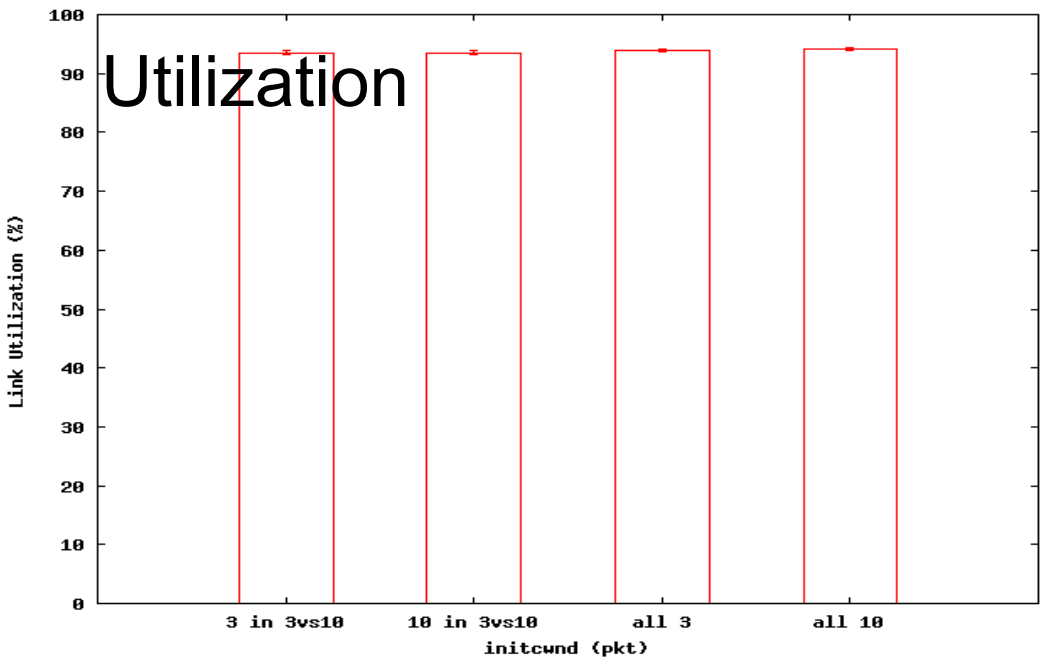
Must restart server process to use new tcp_wmem[1]

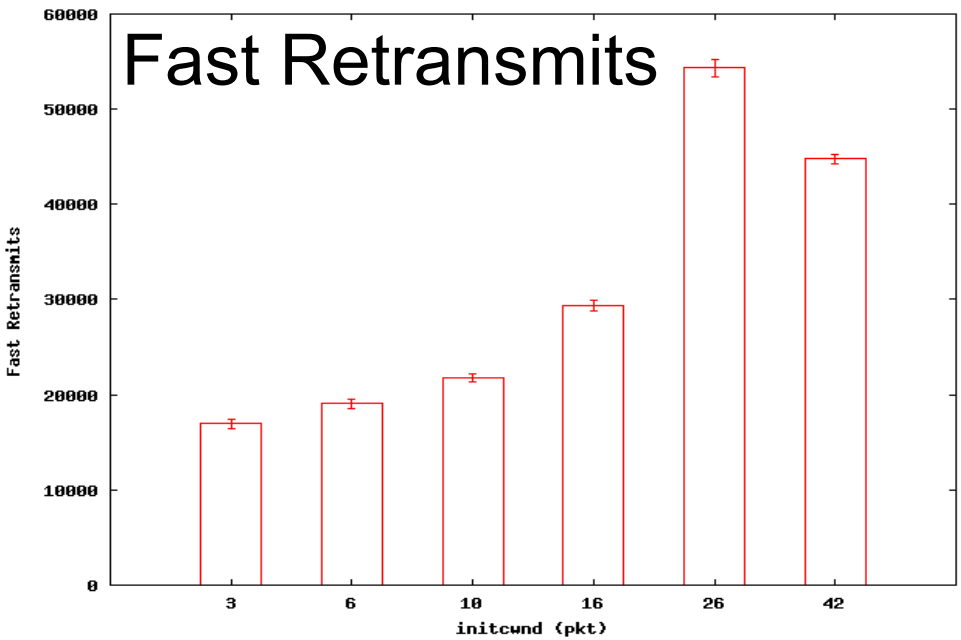
Acknowledgements

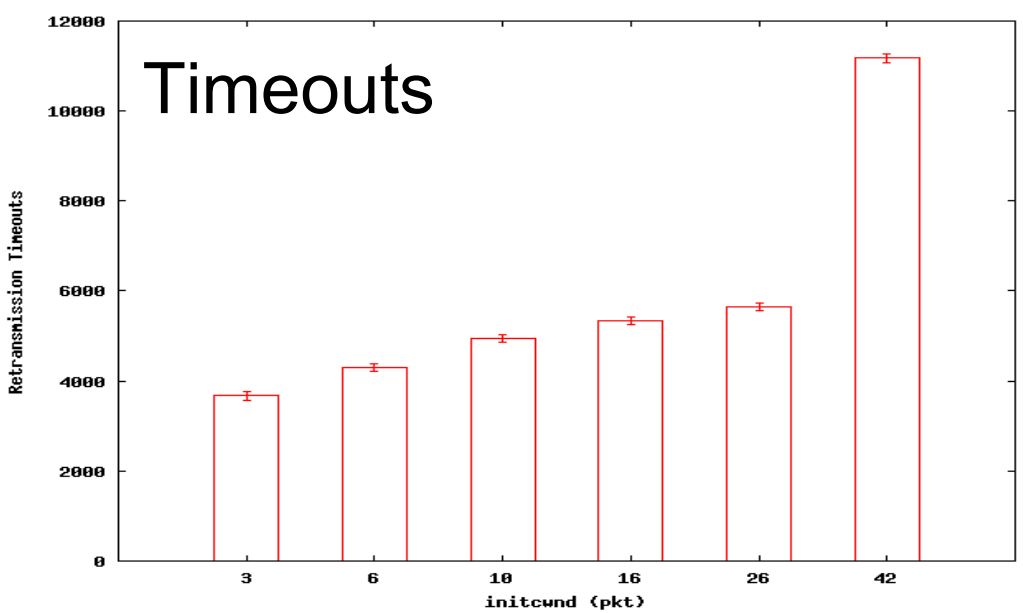
- We acknowledge the following people at Google for their contributions towards the large scale IW experiments:
  - Ethan Solomita
  - Elliott Karpilovsky
  - John Reese
  - Yaogong Wang
  - Roberto Peon
  - Arvind Jain

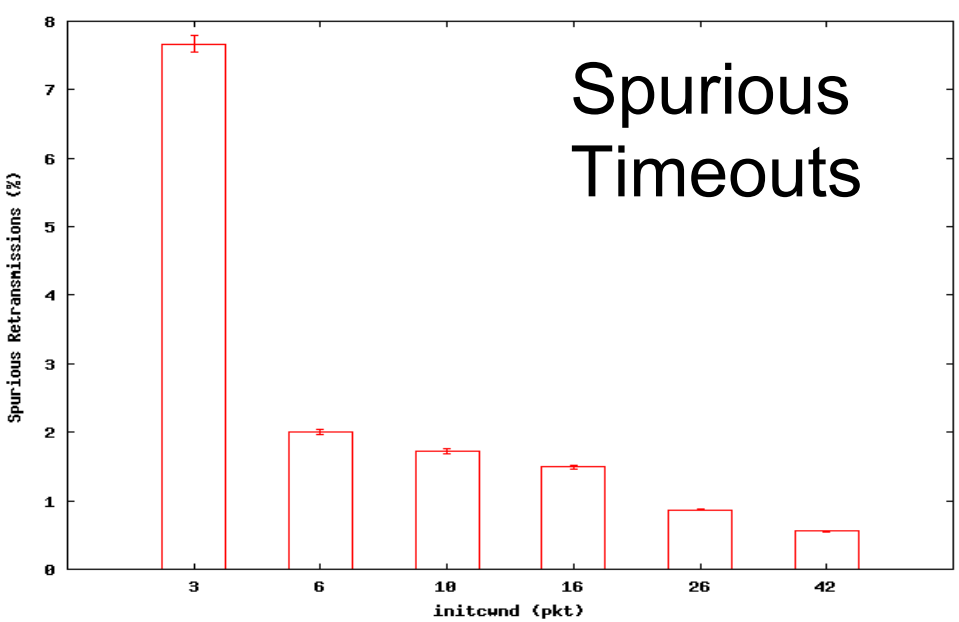
Why does latency improve in Africa?

- (from tesbed experiment results)



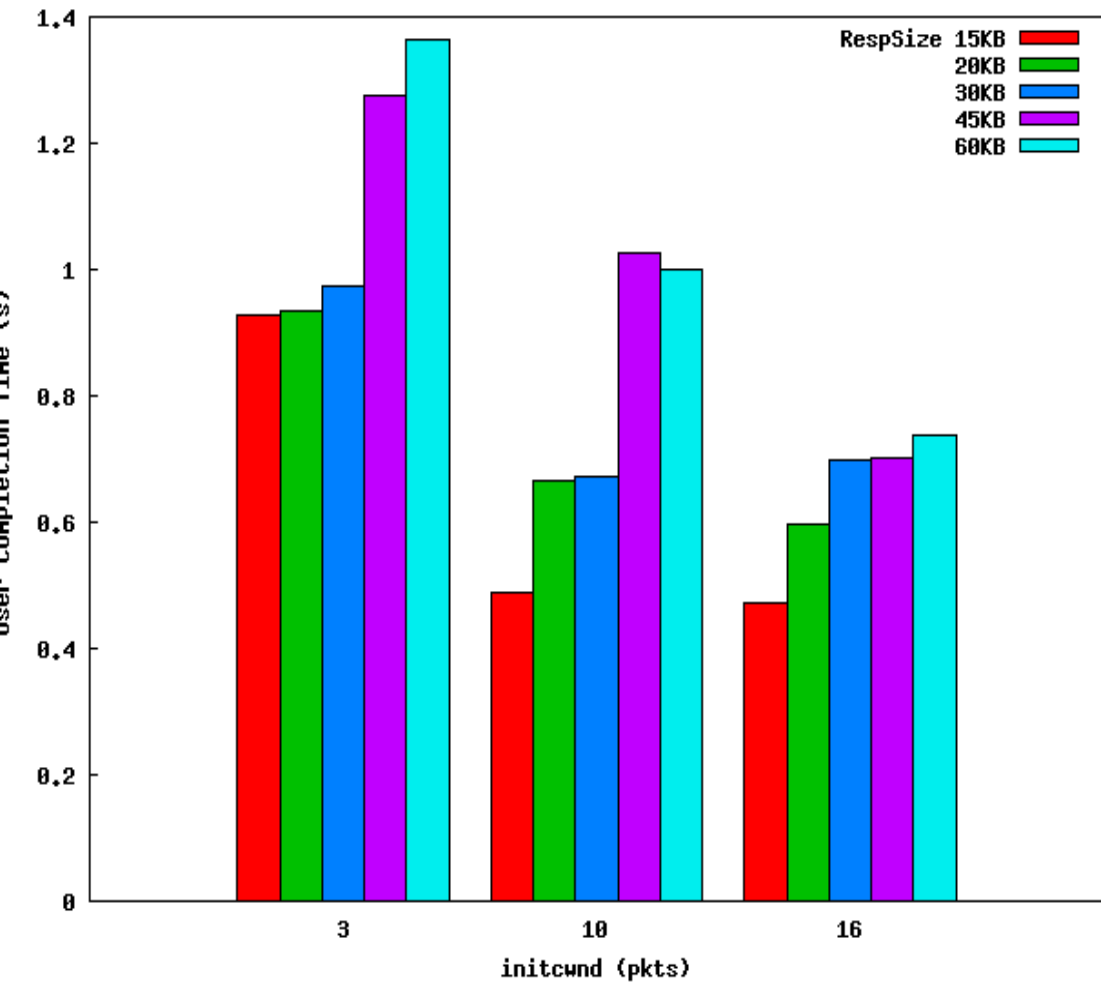


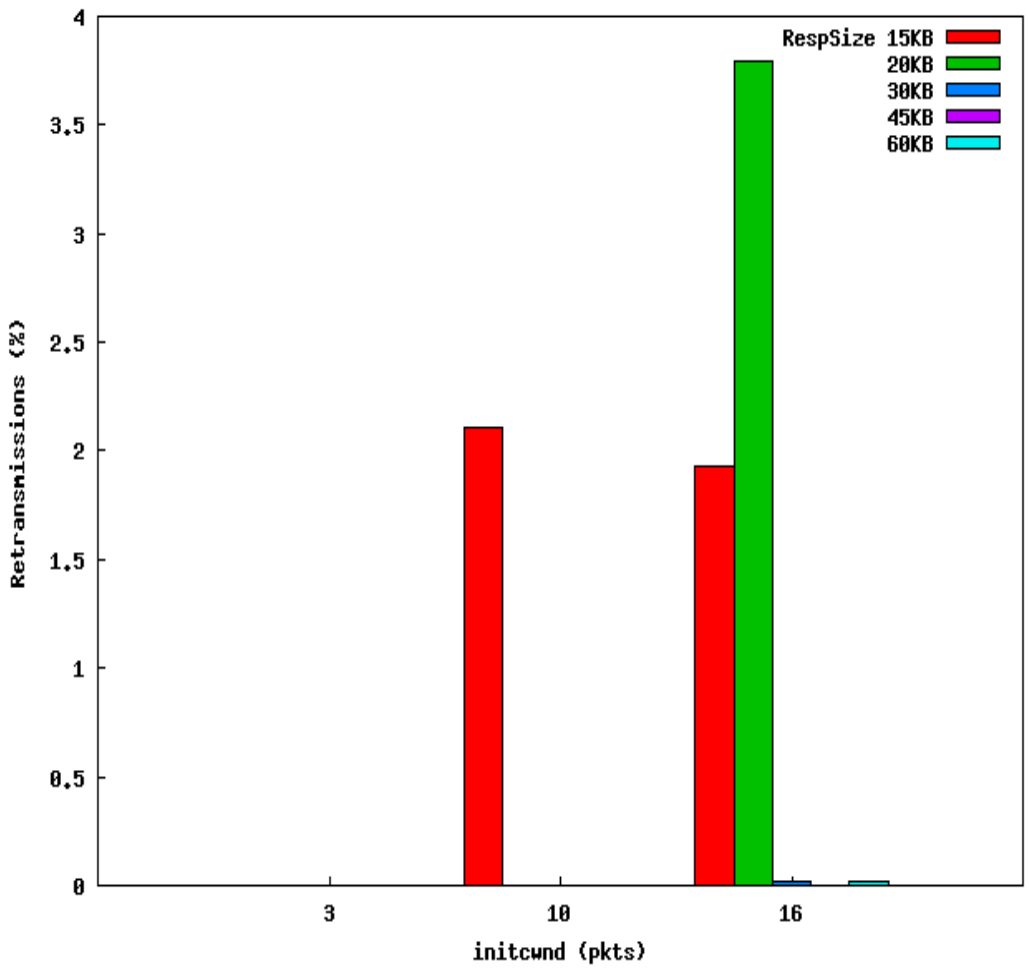




More preliminary results from testbed: latency improves across all transaction sizes

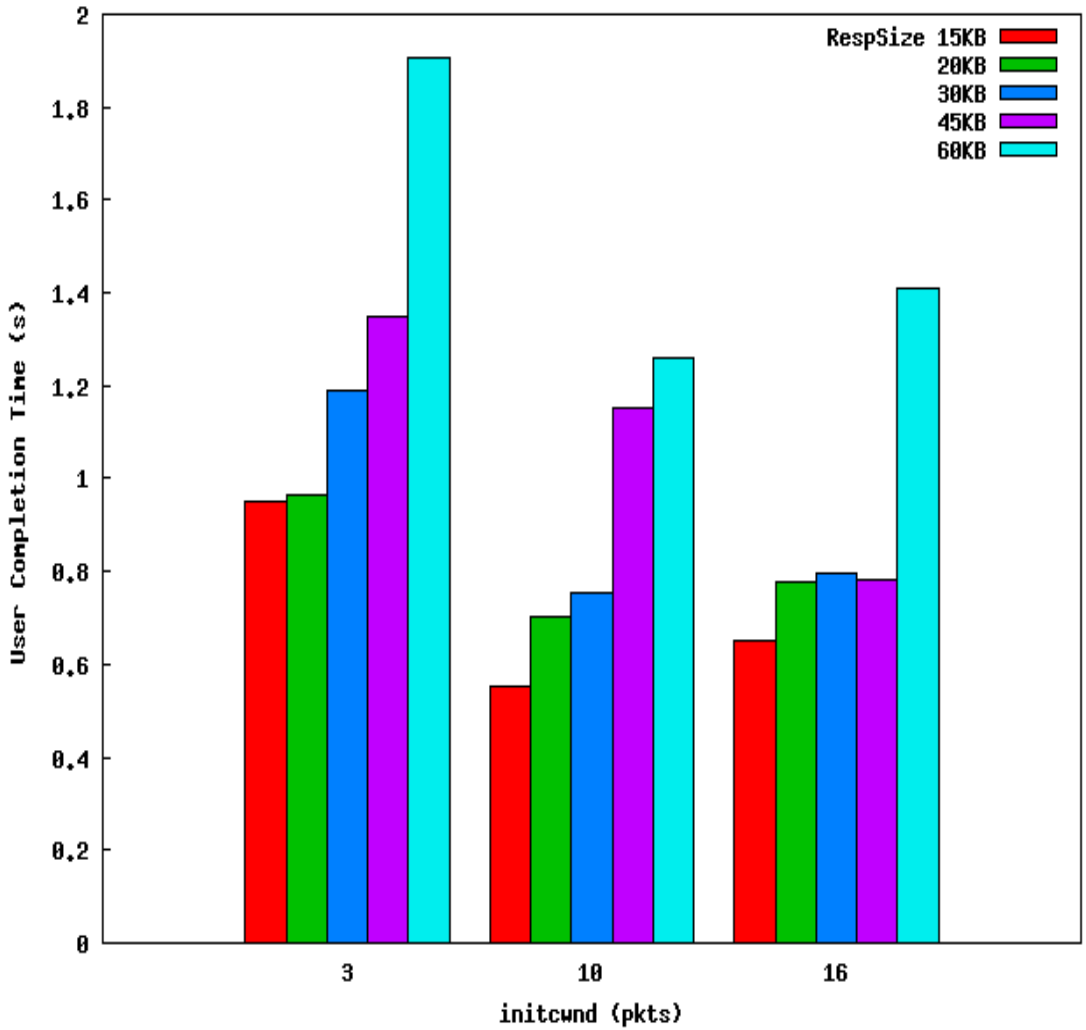
with BDP buffer & < 90% offered load

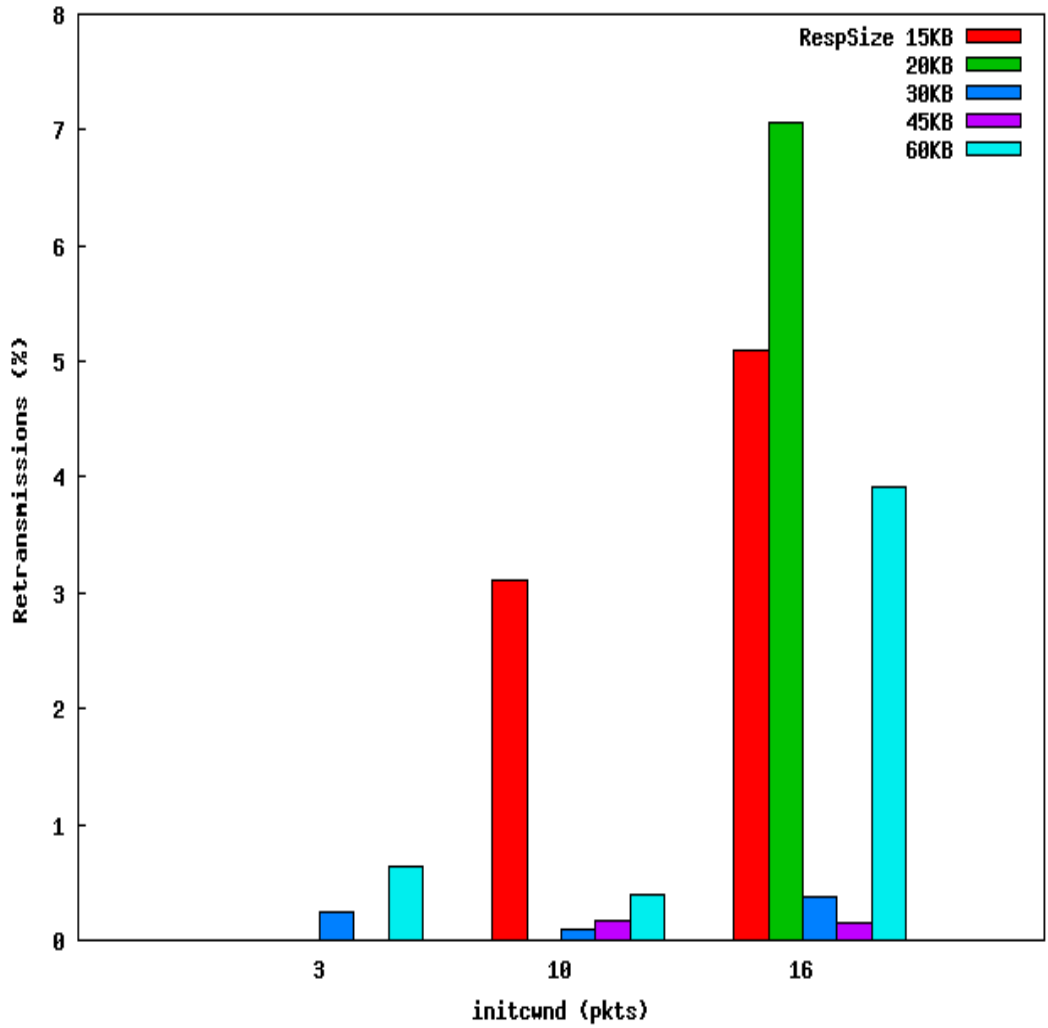




But retransmission rates can be quite different

with BDP buffer and > 95% offered load





Insufficient buffer can hurt IW10 latency

40% BDP buffer, 75% offered load