IETF 91
draft-heitzhe-tcpm-vm-rto

TCP Retransmission Timer for Virtual Machines

Jakob Heitz, Cisco (jheitz@cisco.com)
Chuan He, Ericsson (chuan.he@ericsson.com)

November, 2014

RTO Computation by RFC 6298

The top chart shows round trip time (RTT) samples that vary by 10%. The computed RTO is very good.

In the bottom chart, every 6 RTT samples is just 1.5 times higher than the average of the others.

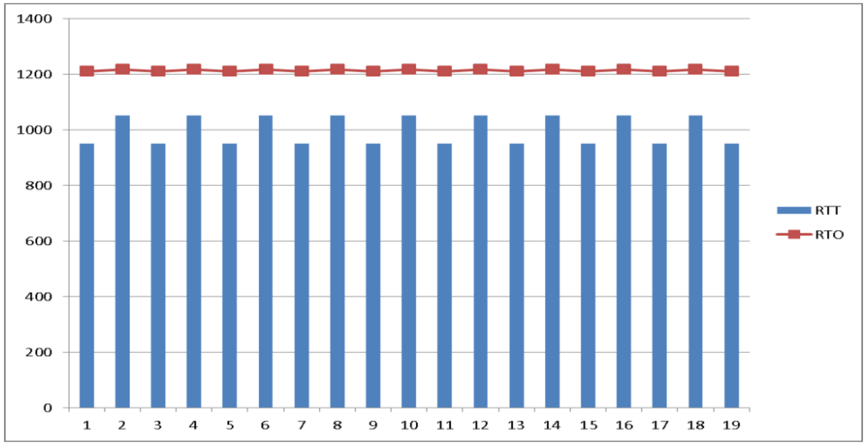
The computed RTO drops too quickly and causes a spurious retransmission every time. Fine clock granularity is assumed.

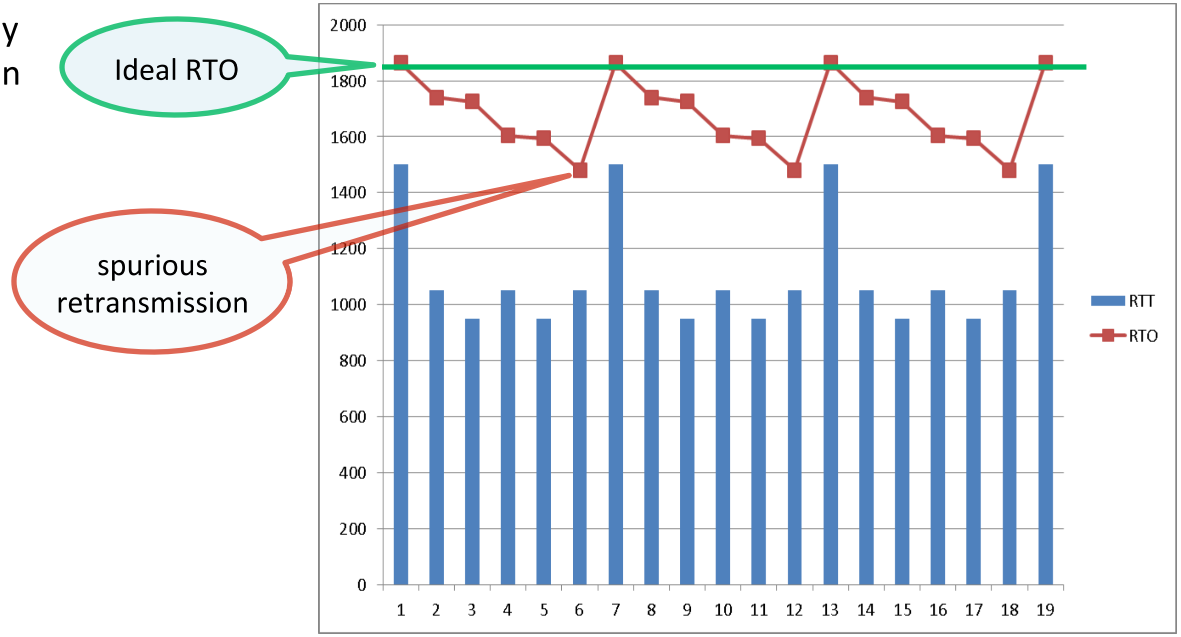
The ideal RTO should remain higher than the spikes,

at around 1.25 of the spikes.

VMs see RTT spikes 10 to 100 times the median RTT.

Median is often < 100uS.




Ideal RTO

spurious retransmission

The Problem

- Between VMs, RTT are highly variable, with median <1mS and maximums often >10mS when loaded.
- RFC6298 needs a minimum 1 Second, because it causes too many spurious retransmissions.
- The proposed algorithm has much fewer retransmissions, therefore needs no minimum.

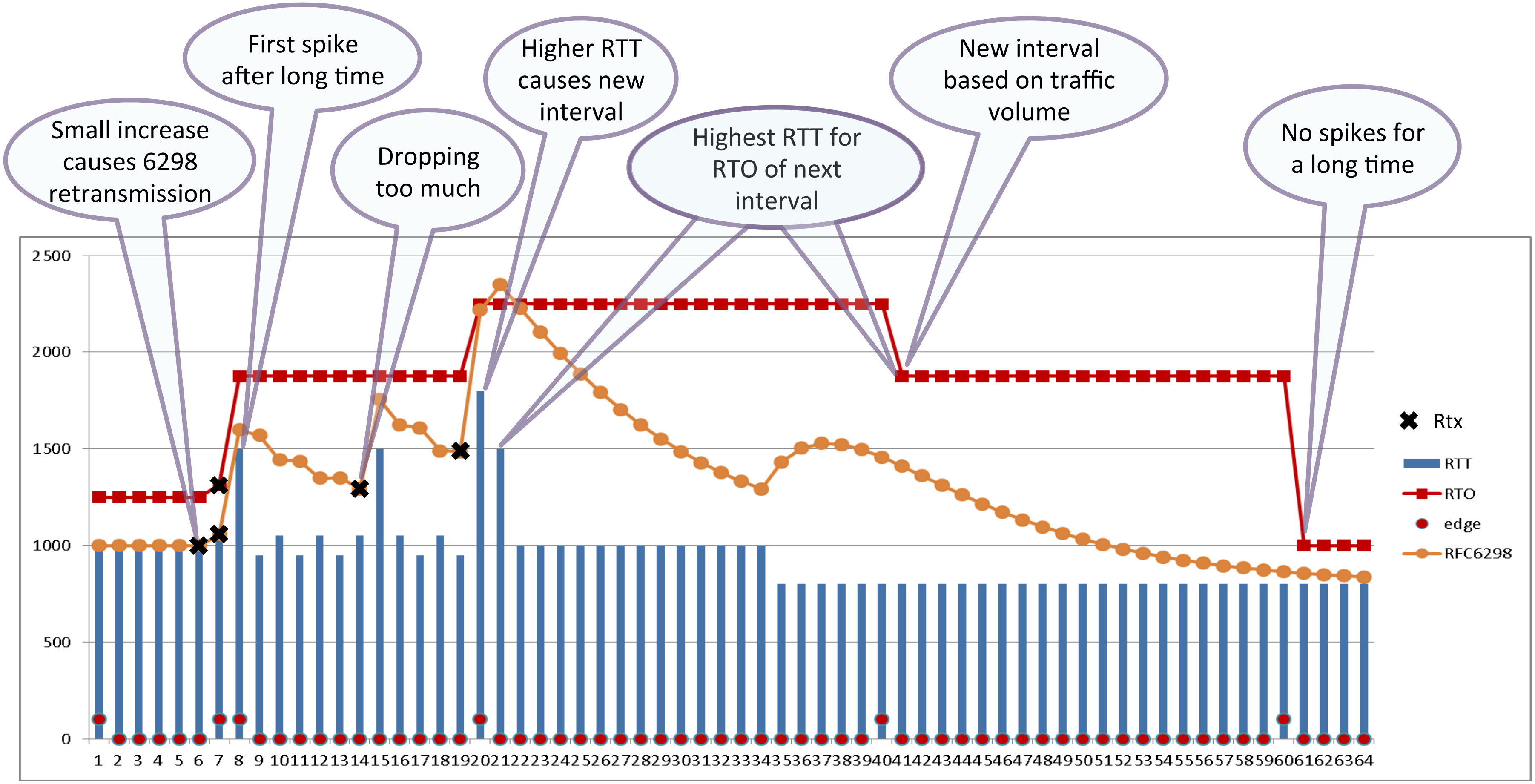
Comparison

RFC6298 experiences a retransmission at every spike.

VM-RTO experiences a retransmission at the first spike. It stays high after that to avoid further retransmissions. Eventually, it must come down.

VM-RTO is less aggressive, needs no minimum. Useable at very low RTT environments.

IMPORTANT: The RTO does not decrease during an interval.



Small increase causes 6298 retransmission

First spike after long time

Dropping too much

Higher RTT causes new interval

Highest RTT for RTO of next interval

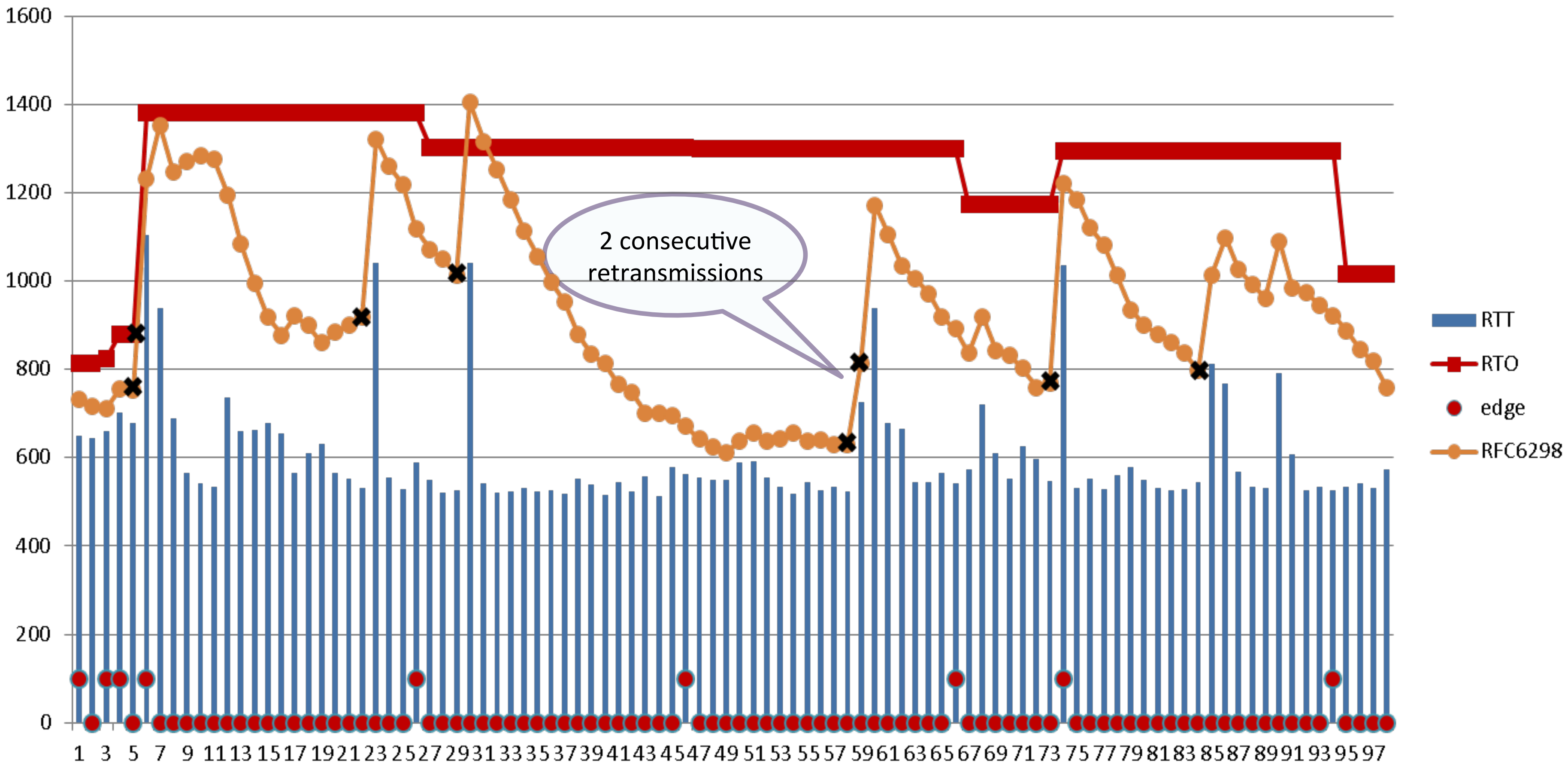
New interval based on traffic volume

No spikes for a long time

Comparison

Ping times between 2 processes in a lightly loaded VM in microseconds.

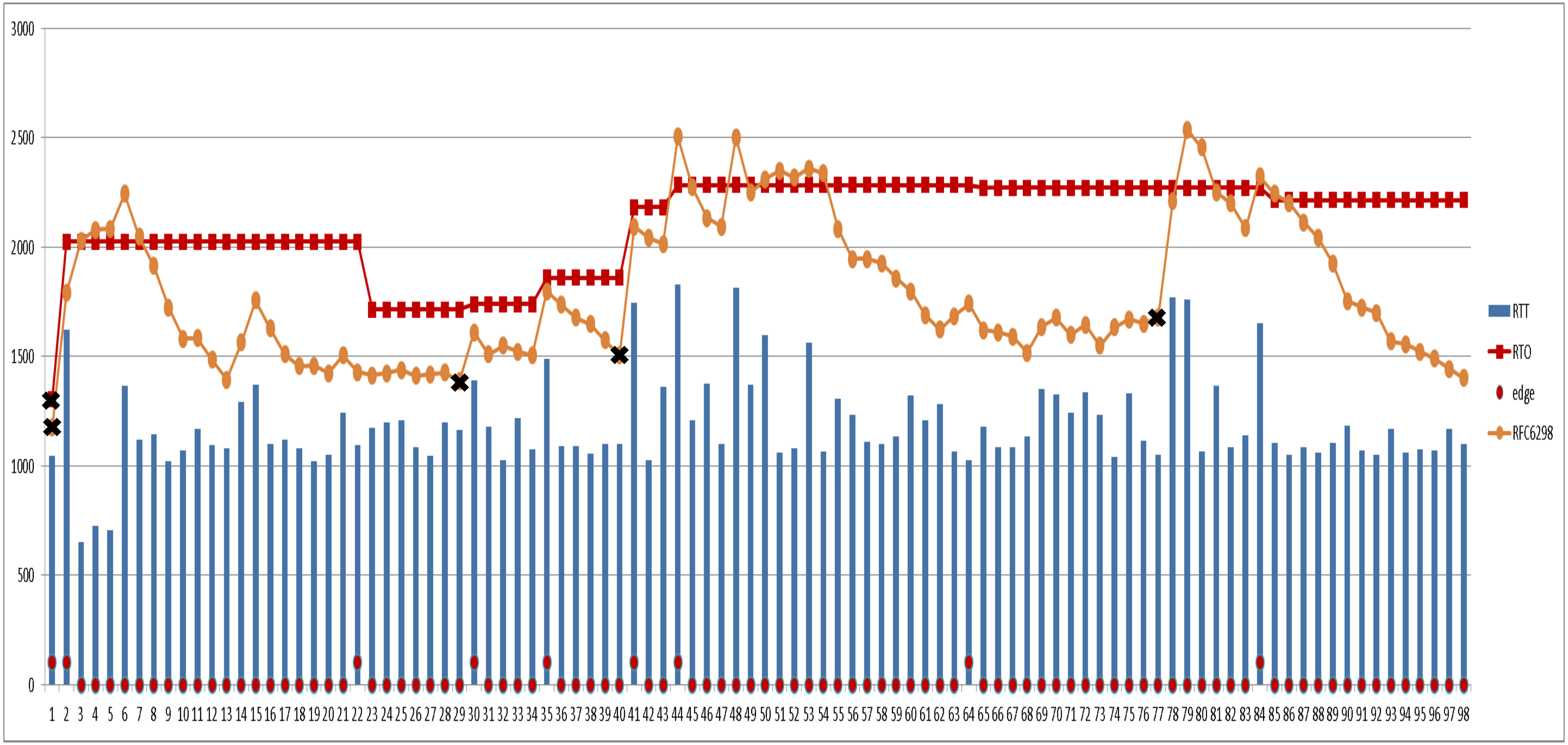
For illustration, the interval is defined as 20 samples. Actual interval is 20 windows.



2 consecutive retransmissions

Comparison

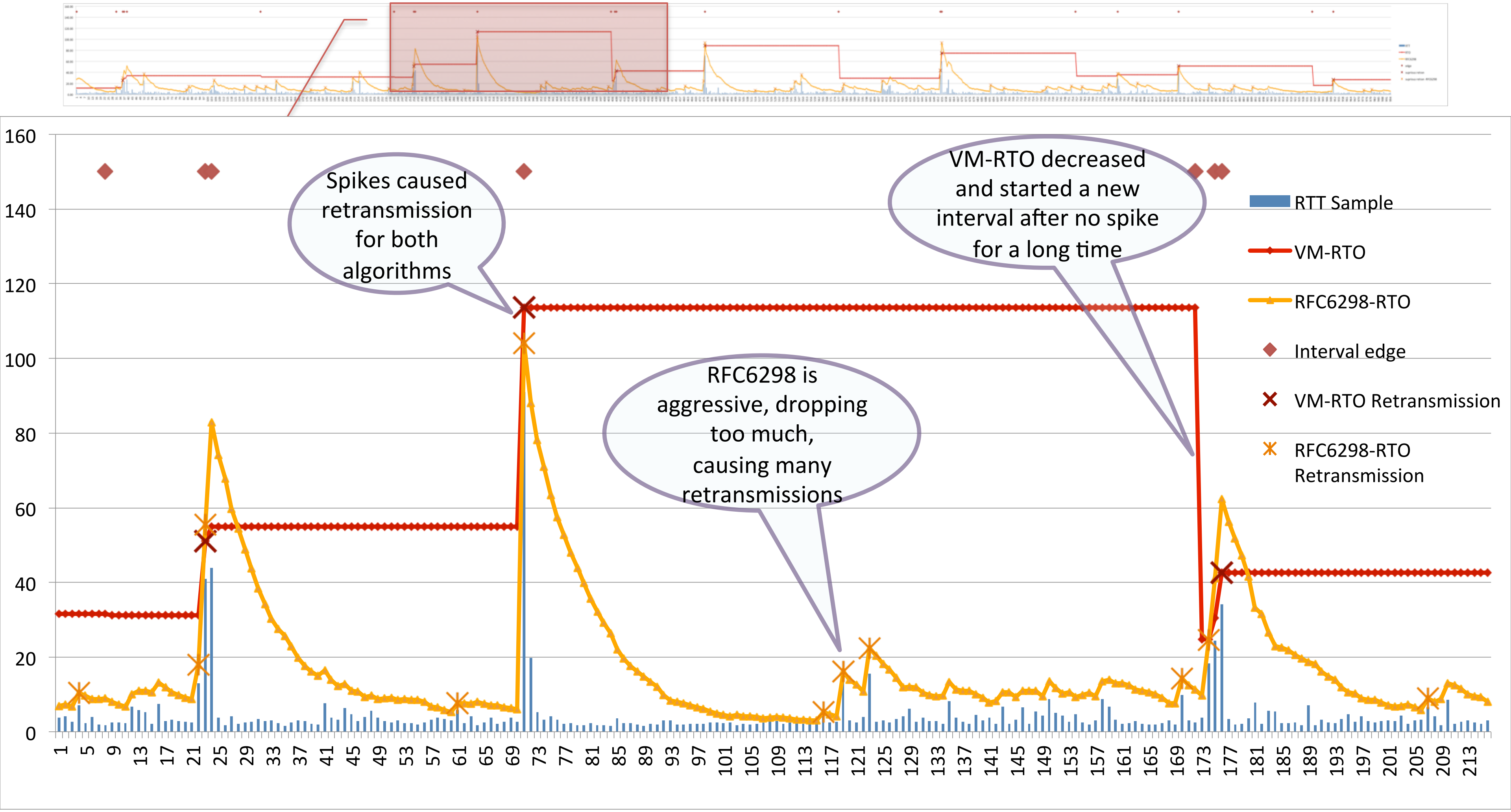
Ping times between 2 processes in a lightly loaded VM in microseconds



Comparison: Real RTT Samples from VM

994 RTT samples from real VM in production, median is 2.8 milliseconds, maximum is 113.72 milliseconds. The interval is 100 samples. No minimum RTO for both algorithms.

Retransmission rate: RFC6298 is 5.6% (56/994), VM-RTO is 0.9% (9/994)



Spikes caused retransmission for both algorithms

VM-RTO decreased and started a new interval after no spike for a long time

RFC6298 is aggressive, dropping too much, causing many retransmissions