TCP Behavior of BGP

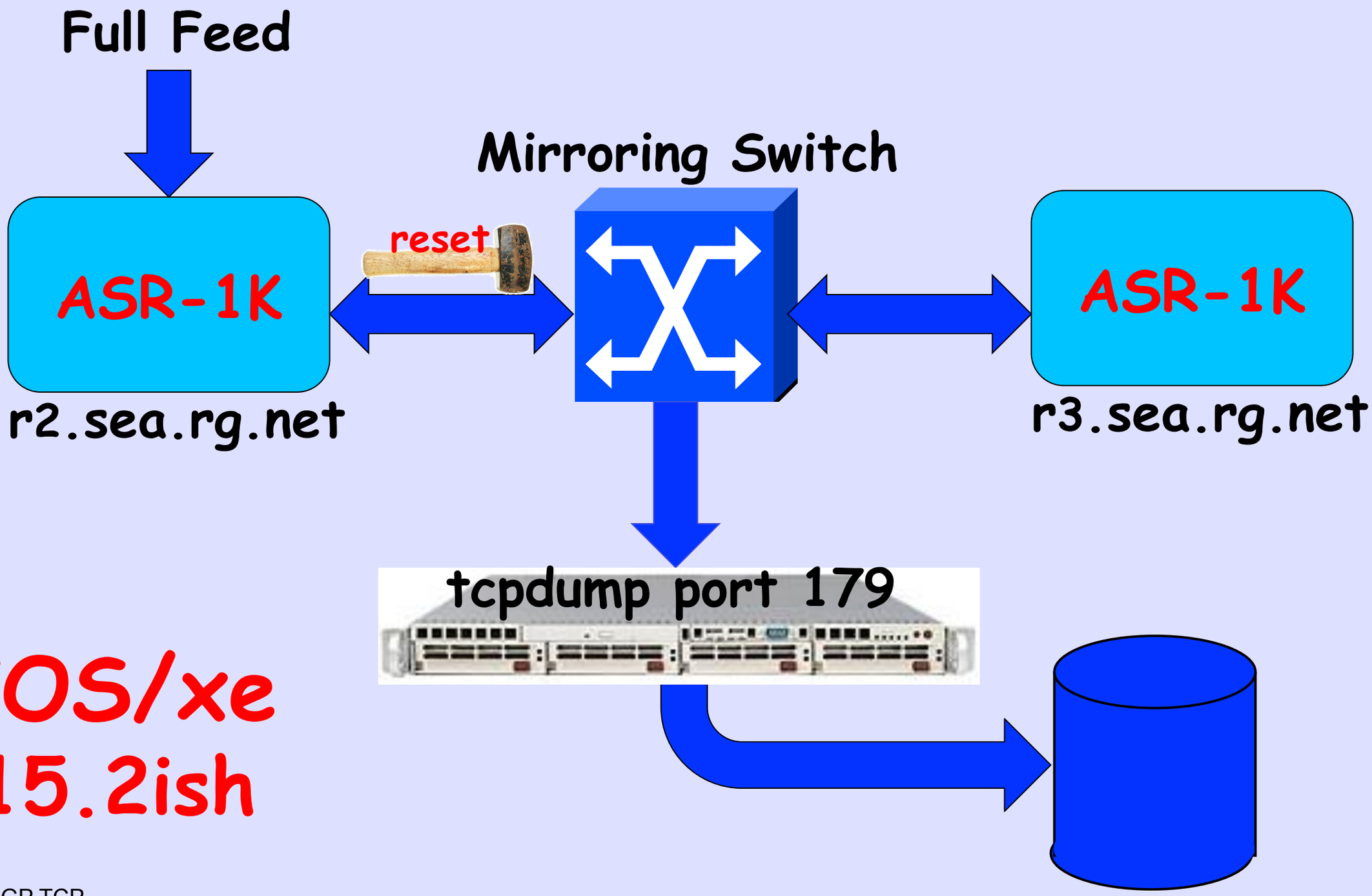
NANOG / Dallas

2012.10.22

Randy Bush <randy@psg.com>
Mark Allman <mallman@icir.org>
Keyur Patel <keyupate@cisco.com>
Balaji Pitta Venkatachalapathy <bvenkata@cisco.com>
Manoj Pandey <mpandey@cisco.com>
Vern Paxson <vern@icir.org>
Cristel Pelsser <cristel@iij.ad.jp>
Ed Kern <ejk@cisco.com>

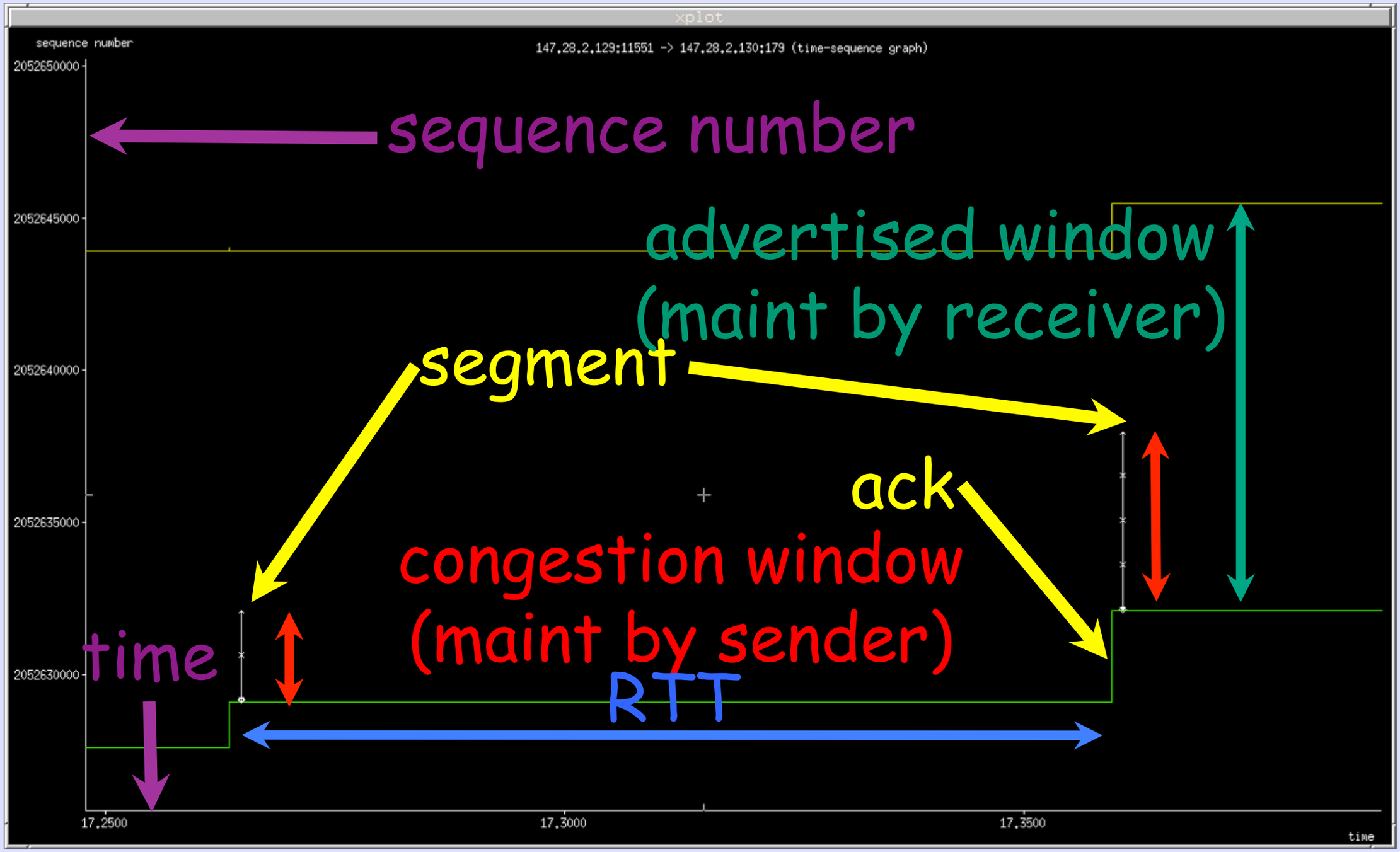
Goal:
Can We Make BGP Convergence Even Faster?

So We Measure



IOS/xe
15.2ish

Time Sequence Plots

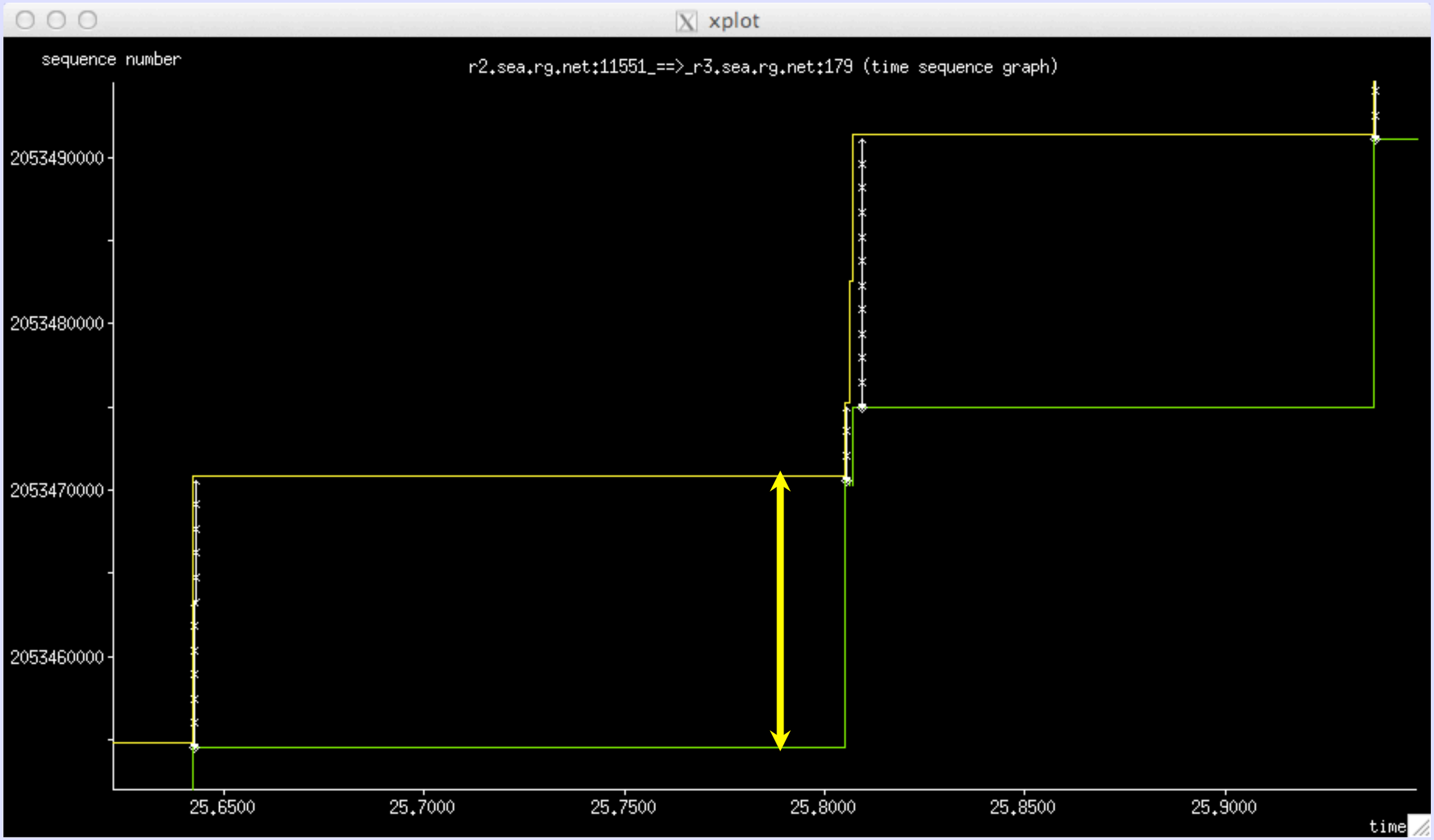


Warning: In xplot, relative scaling of axes is completely arbitrary, i.e., one can zoom one without the other and often does by accident.

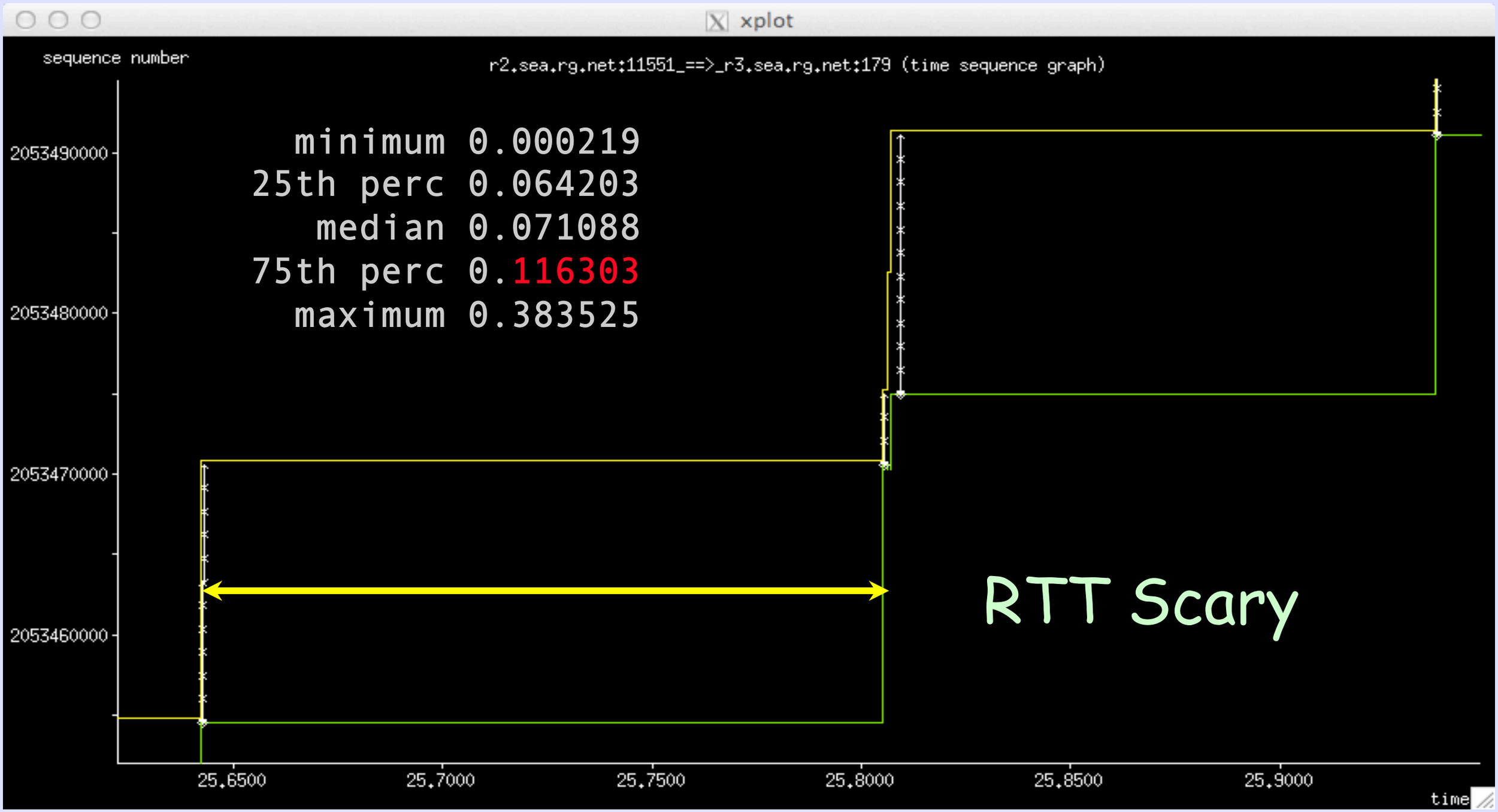
So viewers should not read too much into flat vs. steep curves, only shape patterns.

And there are no calipers, and the axis labels suck. But it's free ☺.

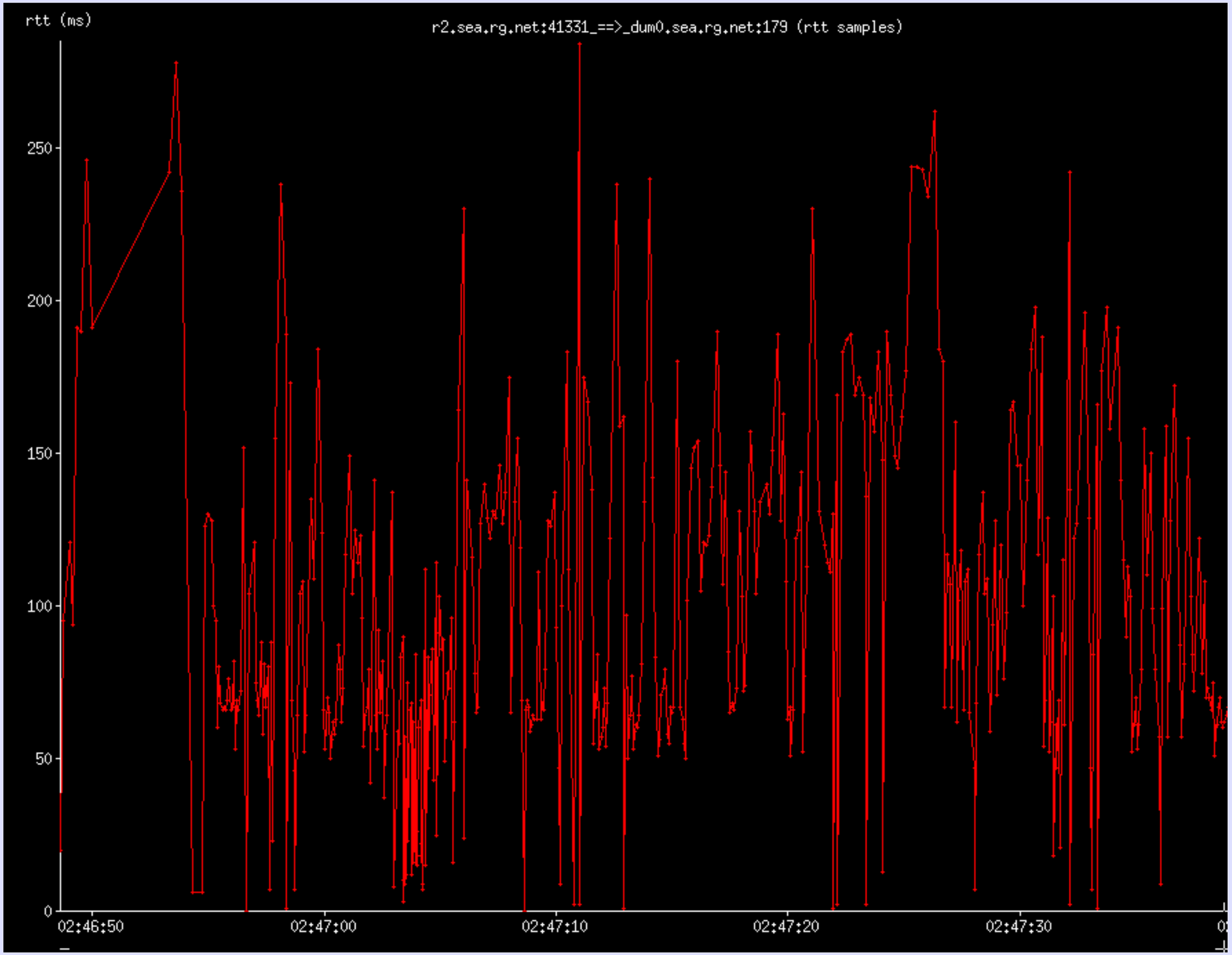
cwnd limited by adv.win



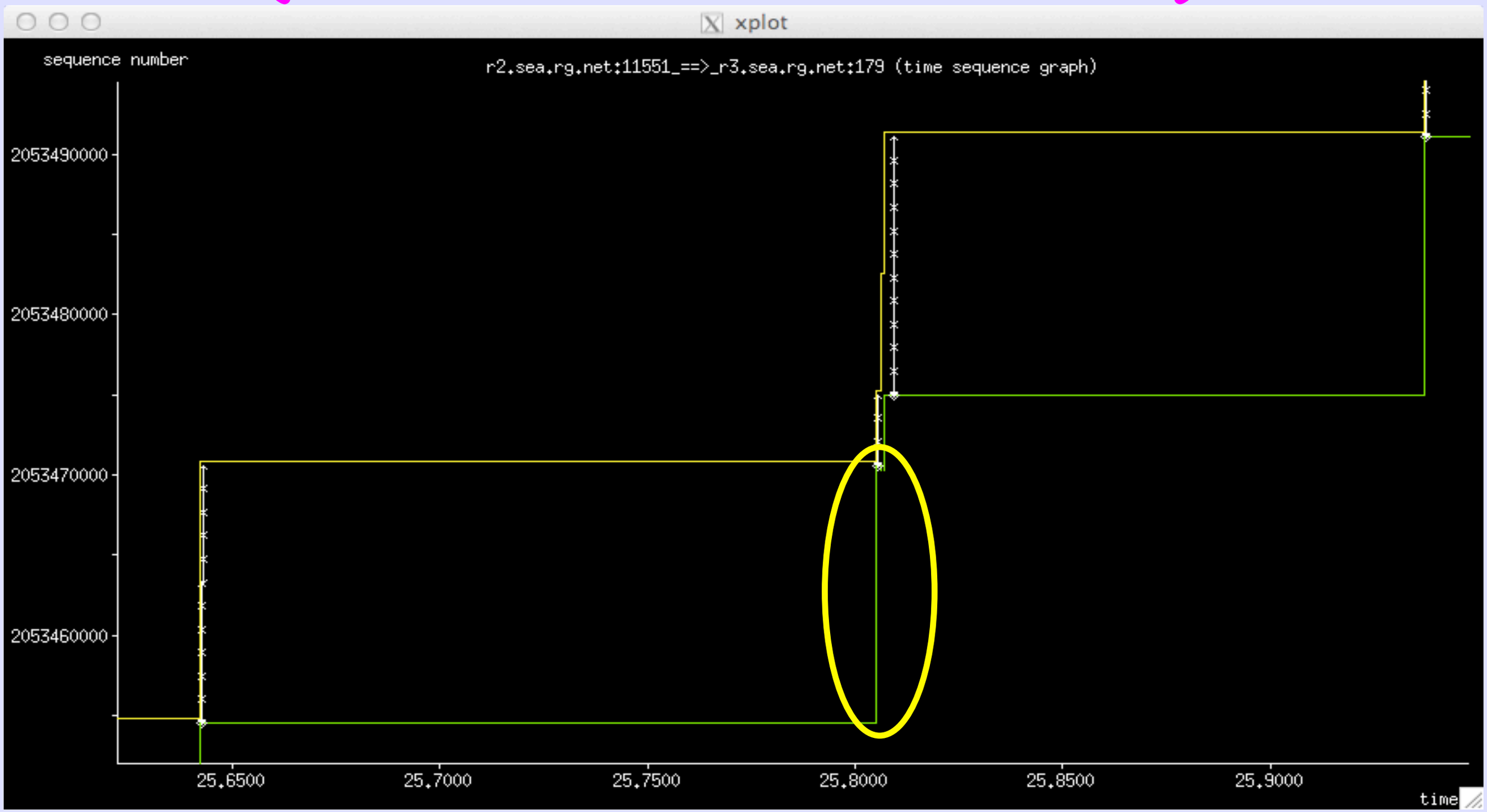
RTT~110ms-On a LAN!



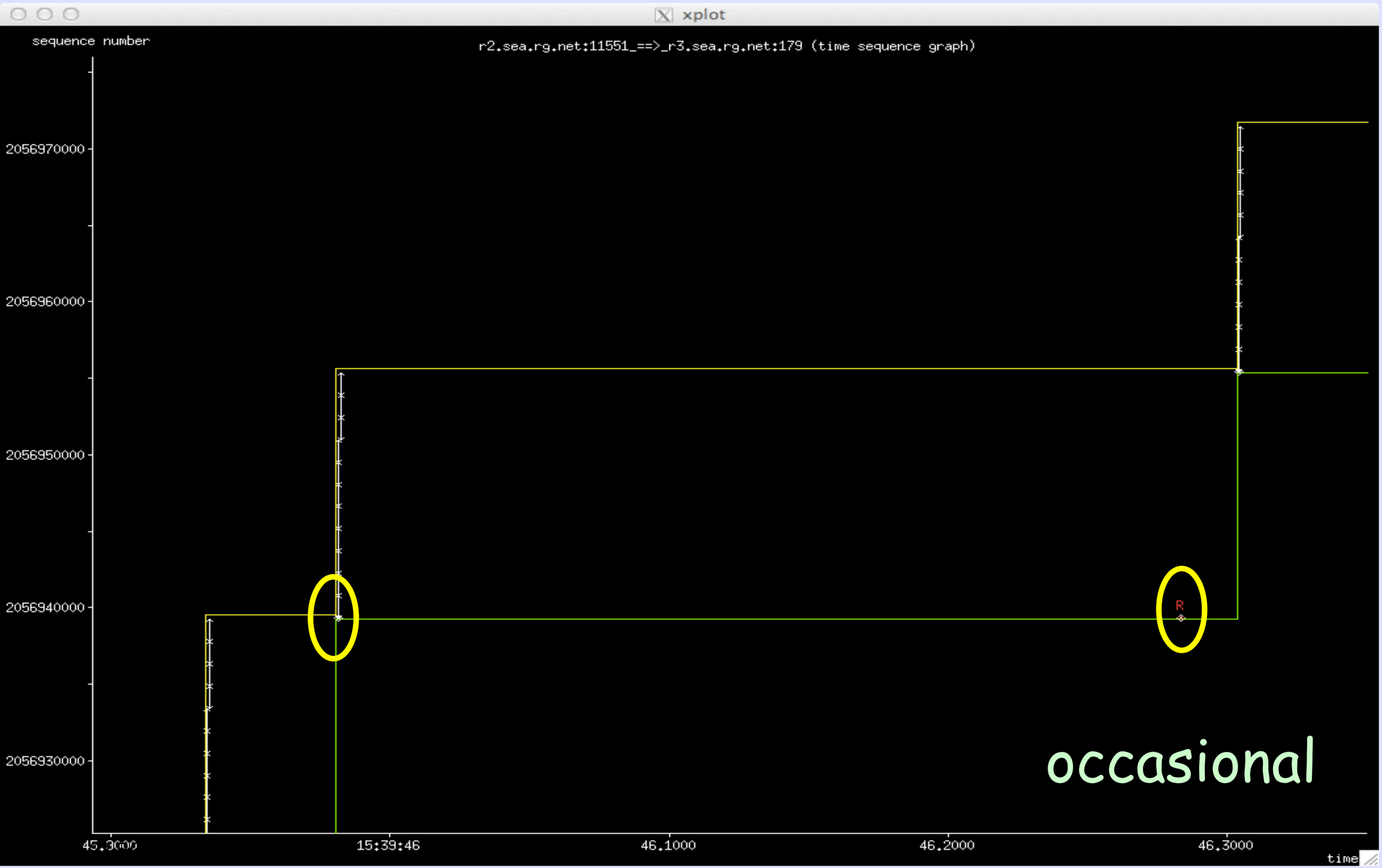
RTT over Full Session



One ACK for Entire Window (AKA 'Stretch' ACK)



Loss and Retransmit



Is Stretch ACK OK?

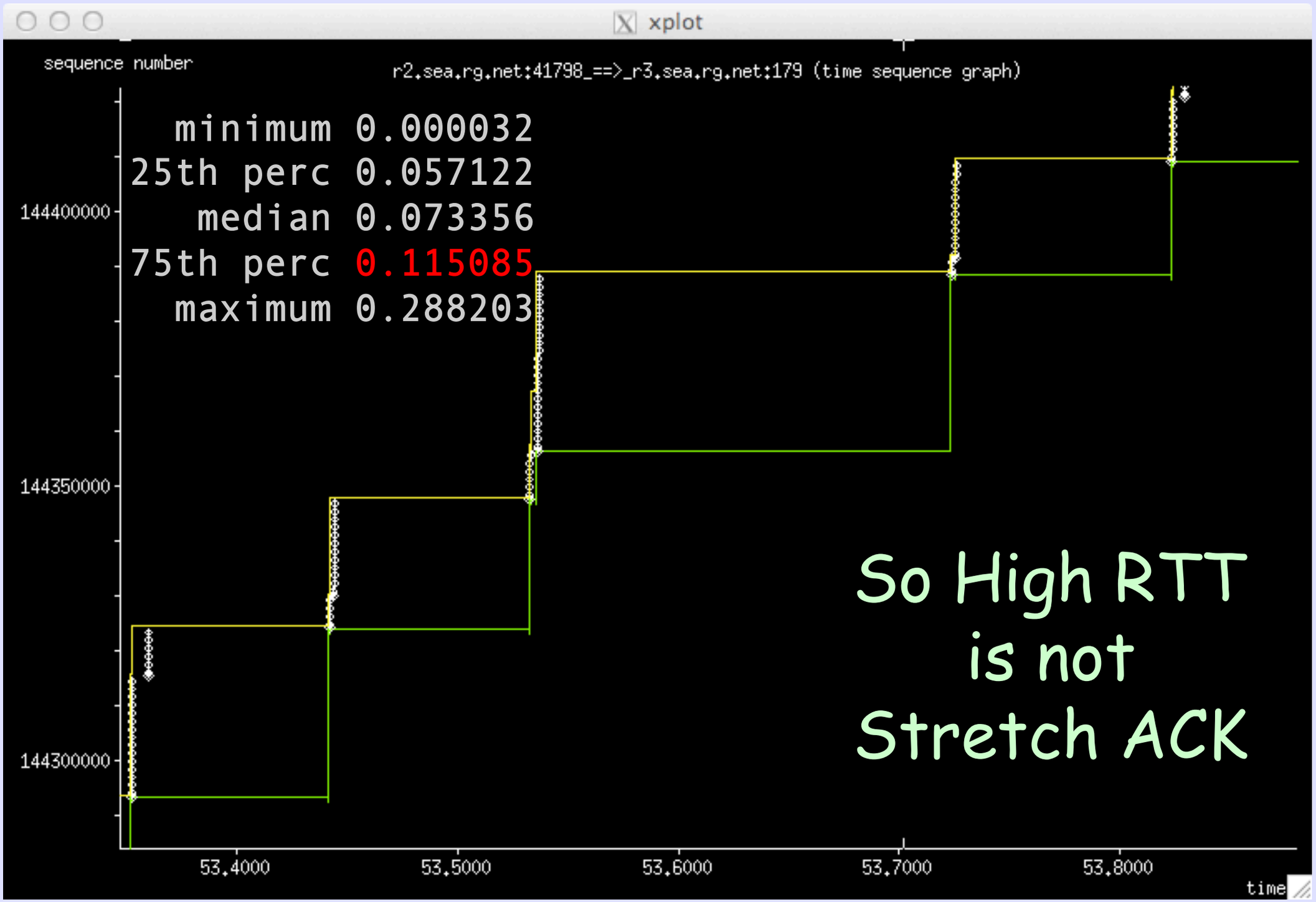
- Stretch ACK for the entire window
  - May contribute to long RTTs, as we wait to coalesce ACKs
  - Very Bursty, as big ACKs cause large window shifts
  - Loss, as bursts overwhelm a buffer (maybe NIC?)

Small Advertised Window

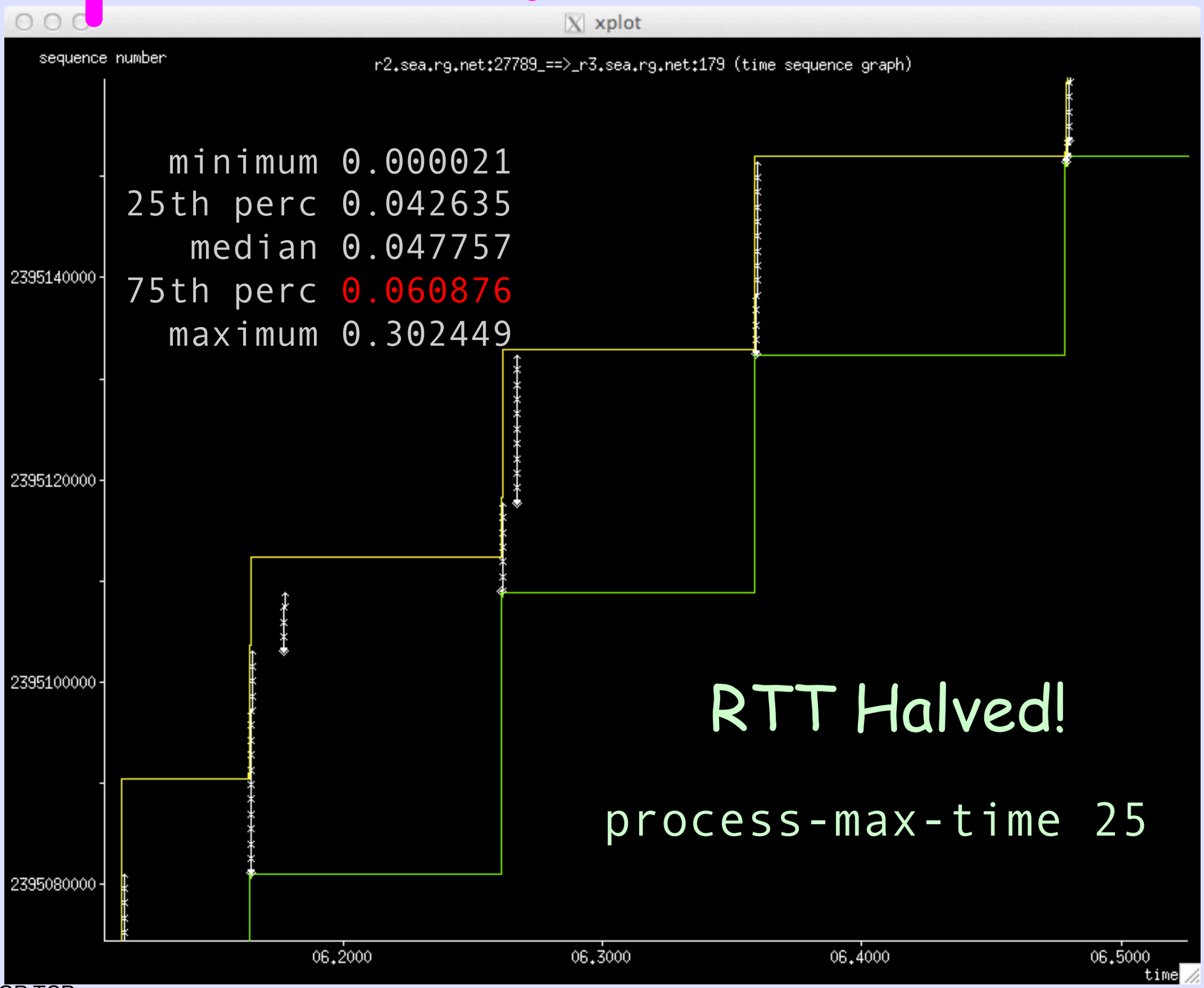
- Would like at least RTT * Bandwidth for TCP to fully utilize the path capacity
- Window size issue is exacerbated by artificially long RTTs
- We increased the advertised window size but it had no impact

Removed Stretch ACK from Code

Stretch ACK Removed



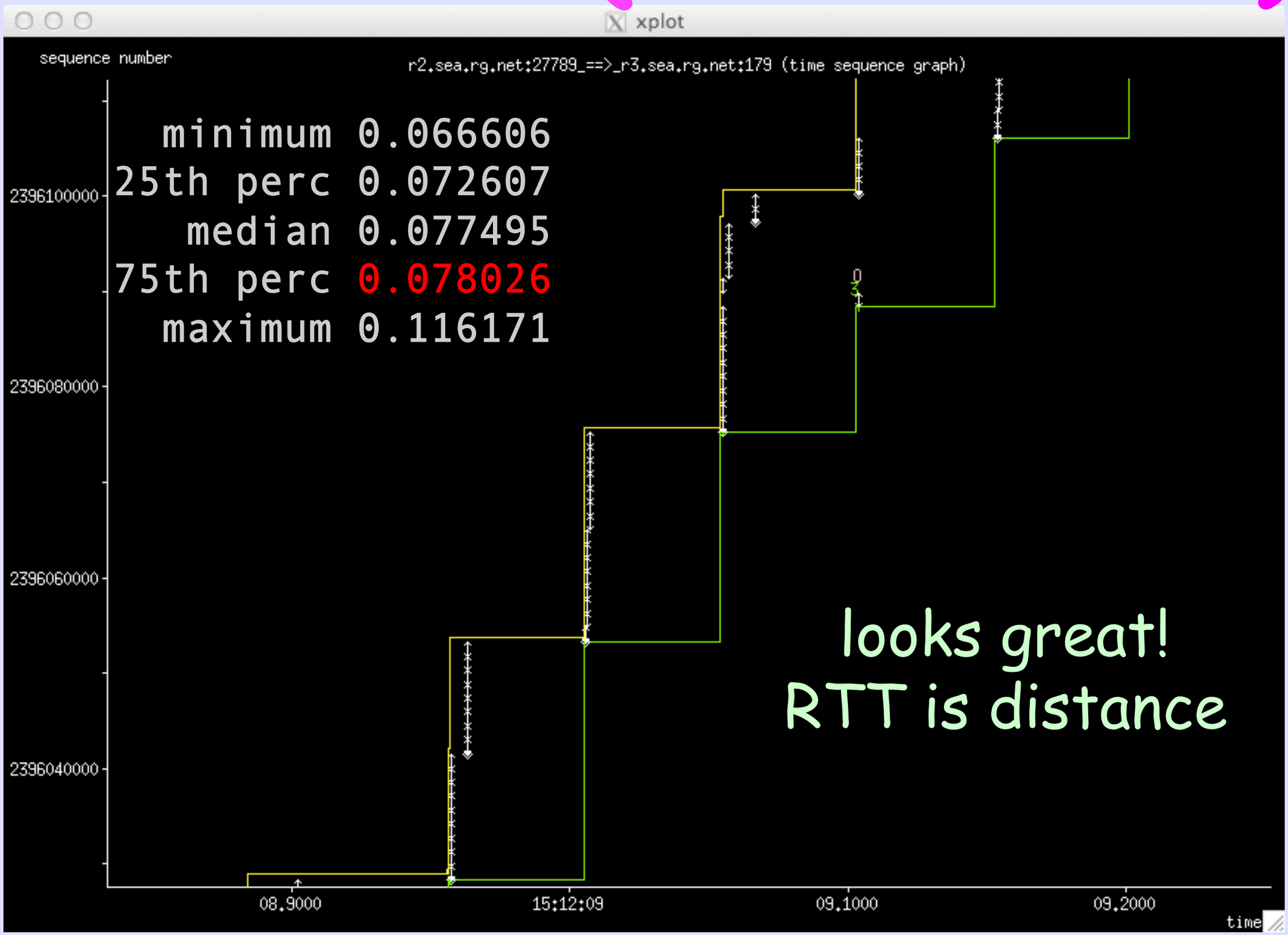
Drop RTC from 50 to 25



What if it is the TCP Stack?

So Let's Measure a Non-BGP Protocol

RPKI-Rtr (SEA-DFW)

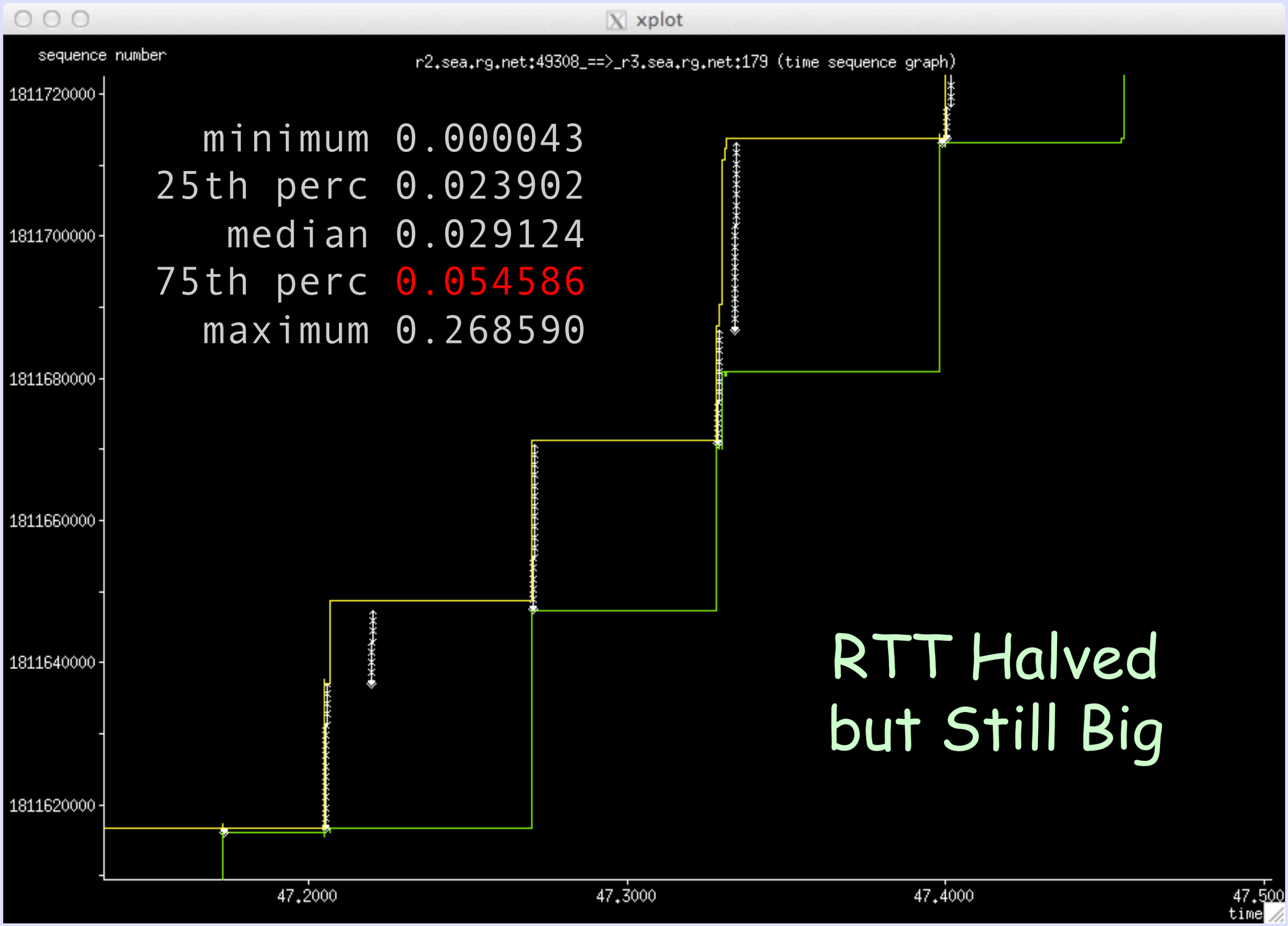


Stack Looks Good!

- Looks like what one would really expect TCP to look like
- ACKs are generated correctly
- Sender fills the window
- RTT looks to be roughly right for the underlying path, Dallas to Seattle
- Window is too small for the path, but buffer small as they're saving RAM

So is it BGP or Could it Be RIB->FIB?

BGP w/o RIB->FIB



Better, Not Yet Beautiful

- Get rid of Stretch ACK
- Open the Window to >= 32K
- RIB->FIB is a known 'opportunity'
- Is Run To Completion keeping the RTT high?
- The stack is not so bad. Yay!
- We really want to measure XR!!!

We got the RTTs Down

They are Still Too Long

We are Still

Chasing This

And

We Saw Something

Very Strange

in Dallas

RPKI (DFW-DFW) but XR