Explaining BGP Slow Table Transfers

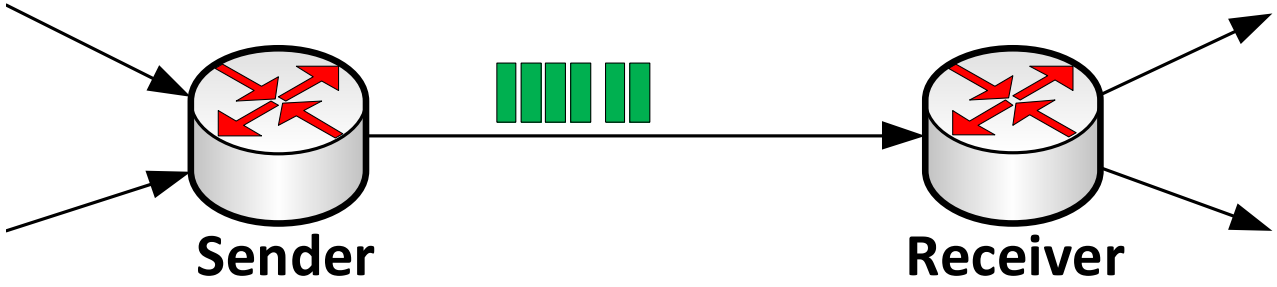
- Implementing a TCP delay analyzer

Pei-chun Cheng, Jong Han Park, Keyur Patel[+], Shane Amante[*], Lixia Zhang

UCLA, [+]Cisco, [*]Level3

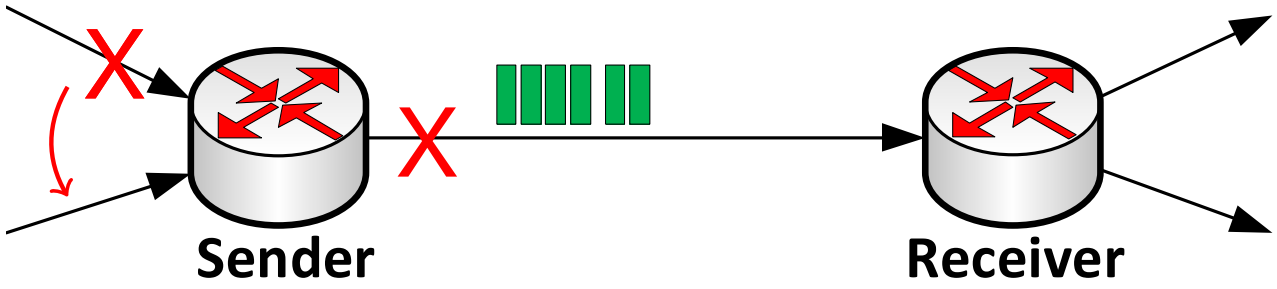
BGP table transfer

- Massive routing updates . . .
  - Affect a large portion of a router's BGP table
  - Session resets or major route changes



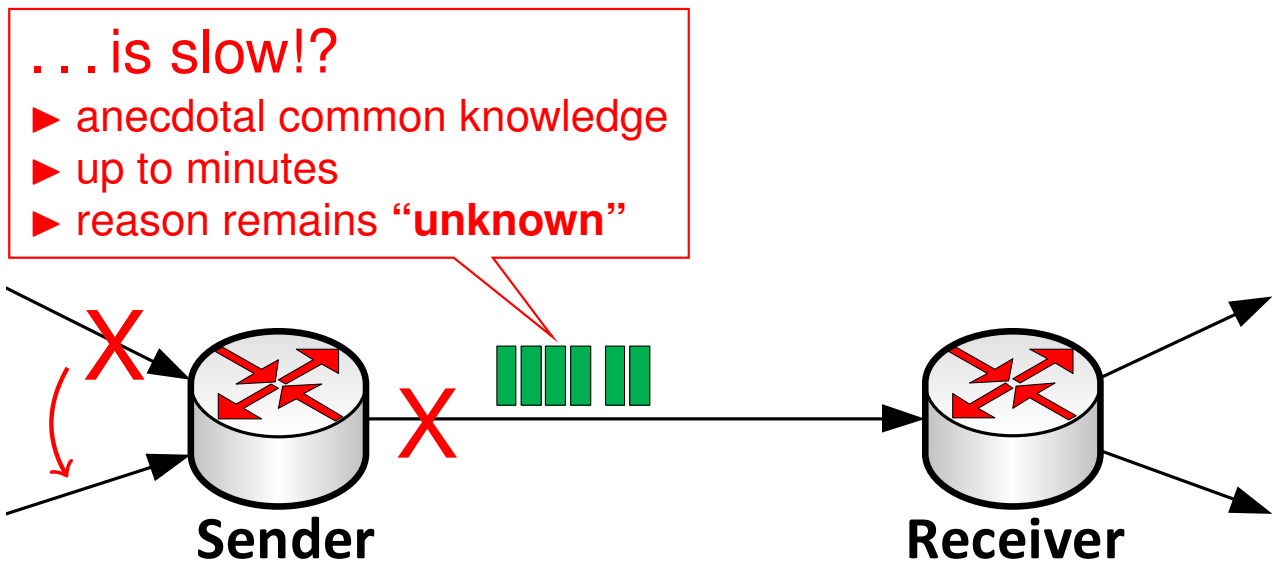
BGP table transfer

- Massive routing updates . . .
  - Affect a large portion of a router's BGP table
  - Session resets or major route changes



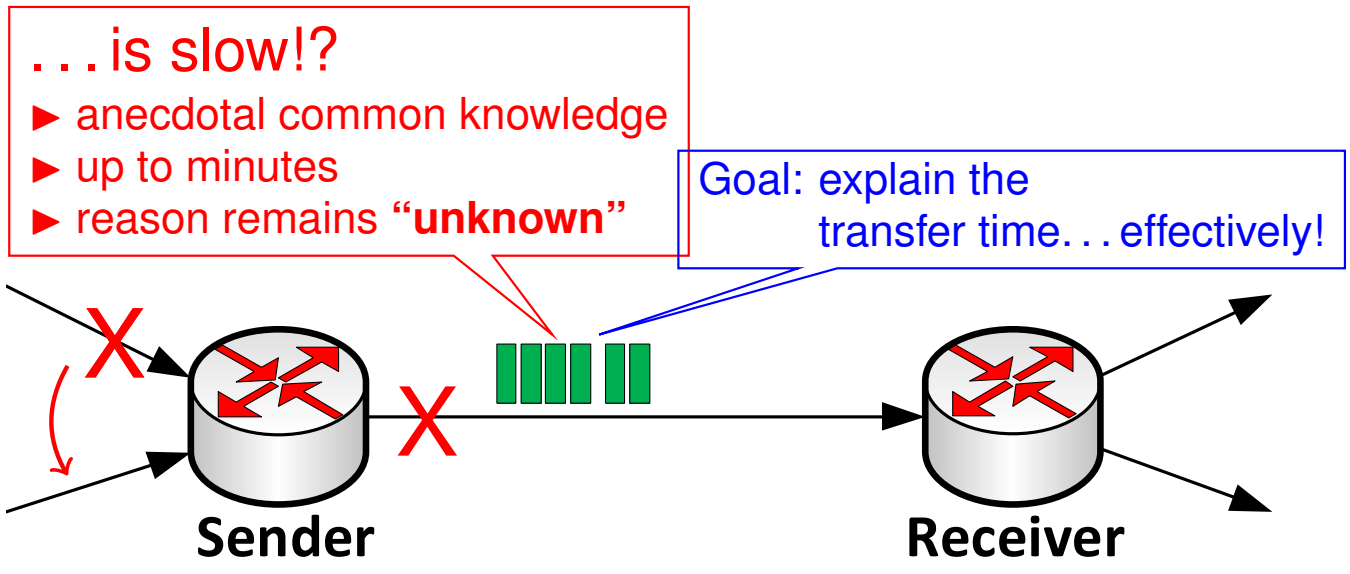
BGP table transfer

- Massive routing updates . . .
  - Affect a large portion of a router's BGP table
  - Session resets or major route changes



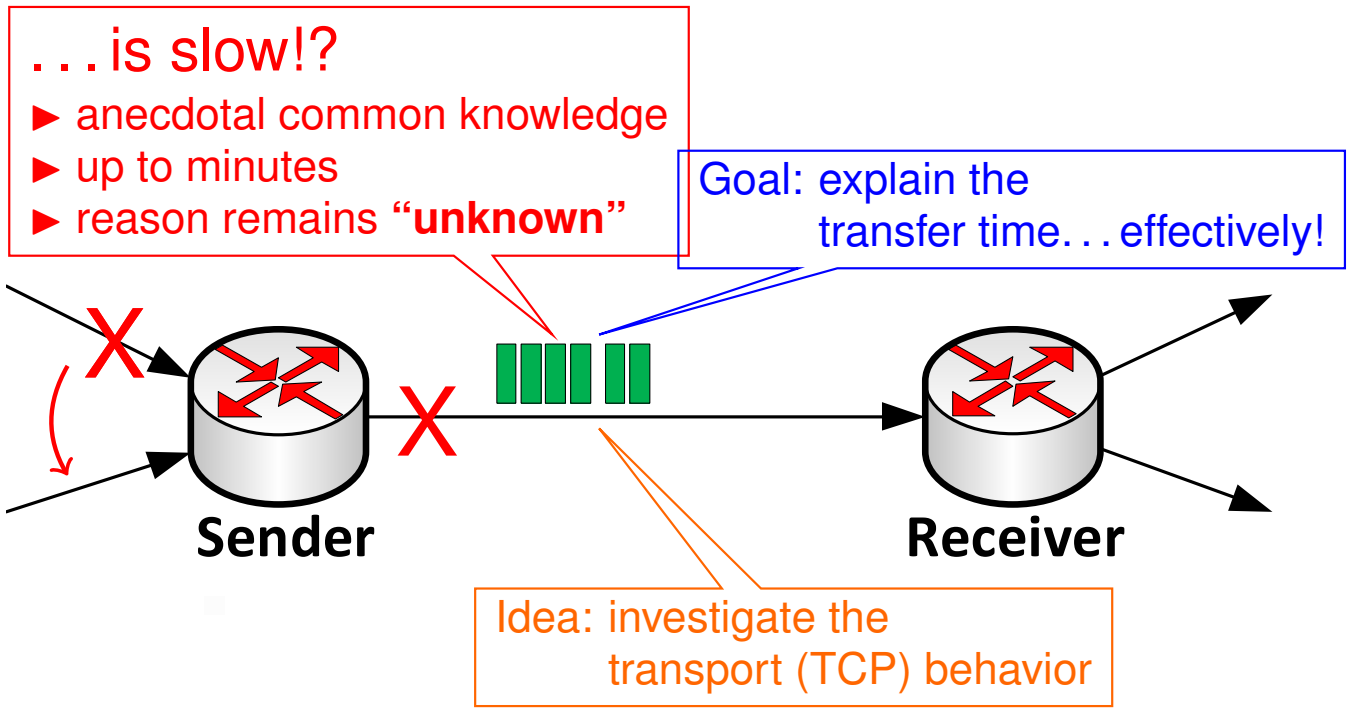
BGP table transfer

- Massive routing updates . . .
  - Affect a large portion of a router's BGP table
  - Session resets or major route changes



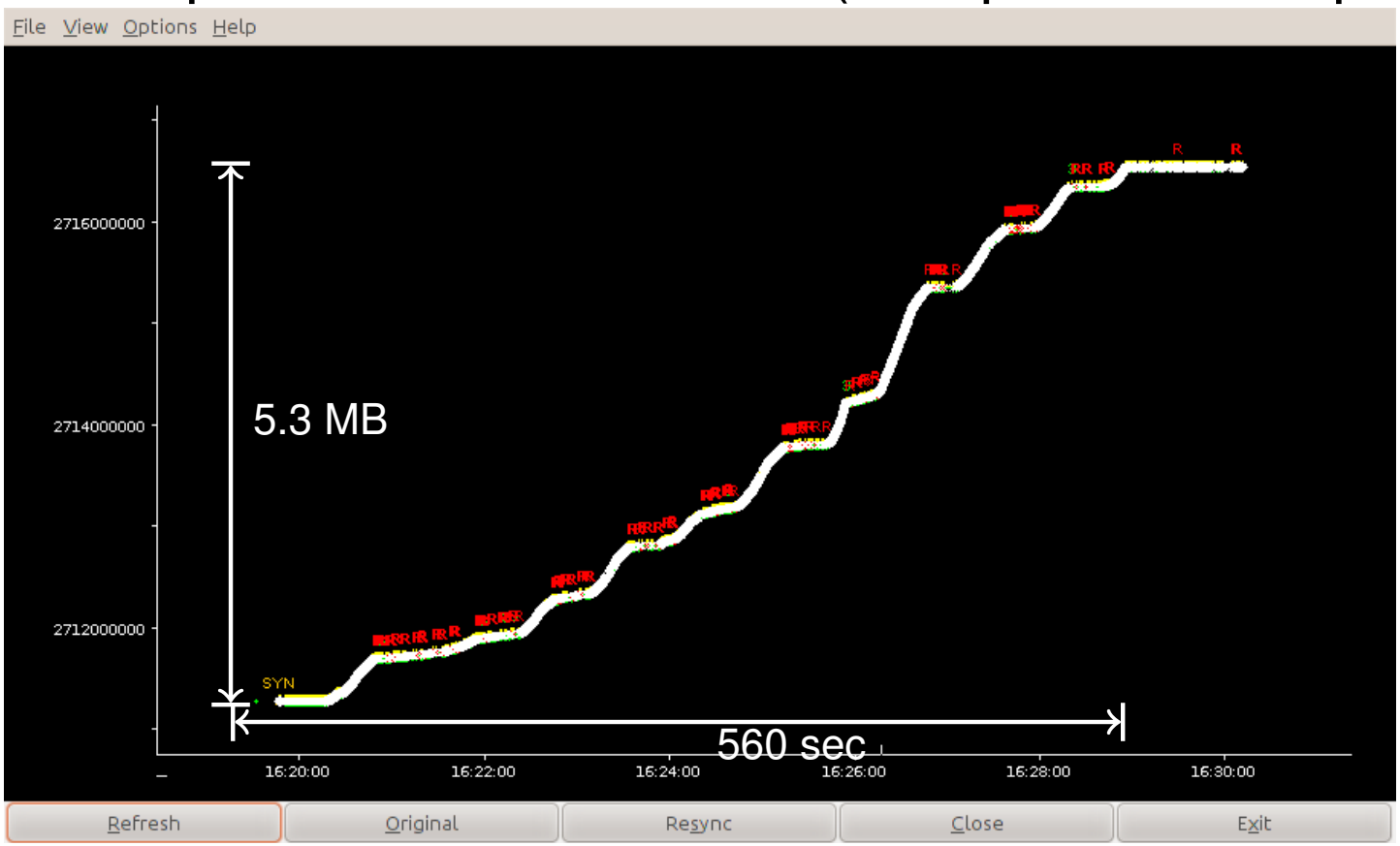
BGP table transfer

- Massive routing updates . . .
  - Affect a large portion of a router's BGP table
  - Session resets or major route changes



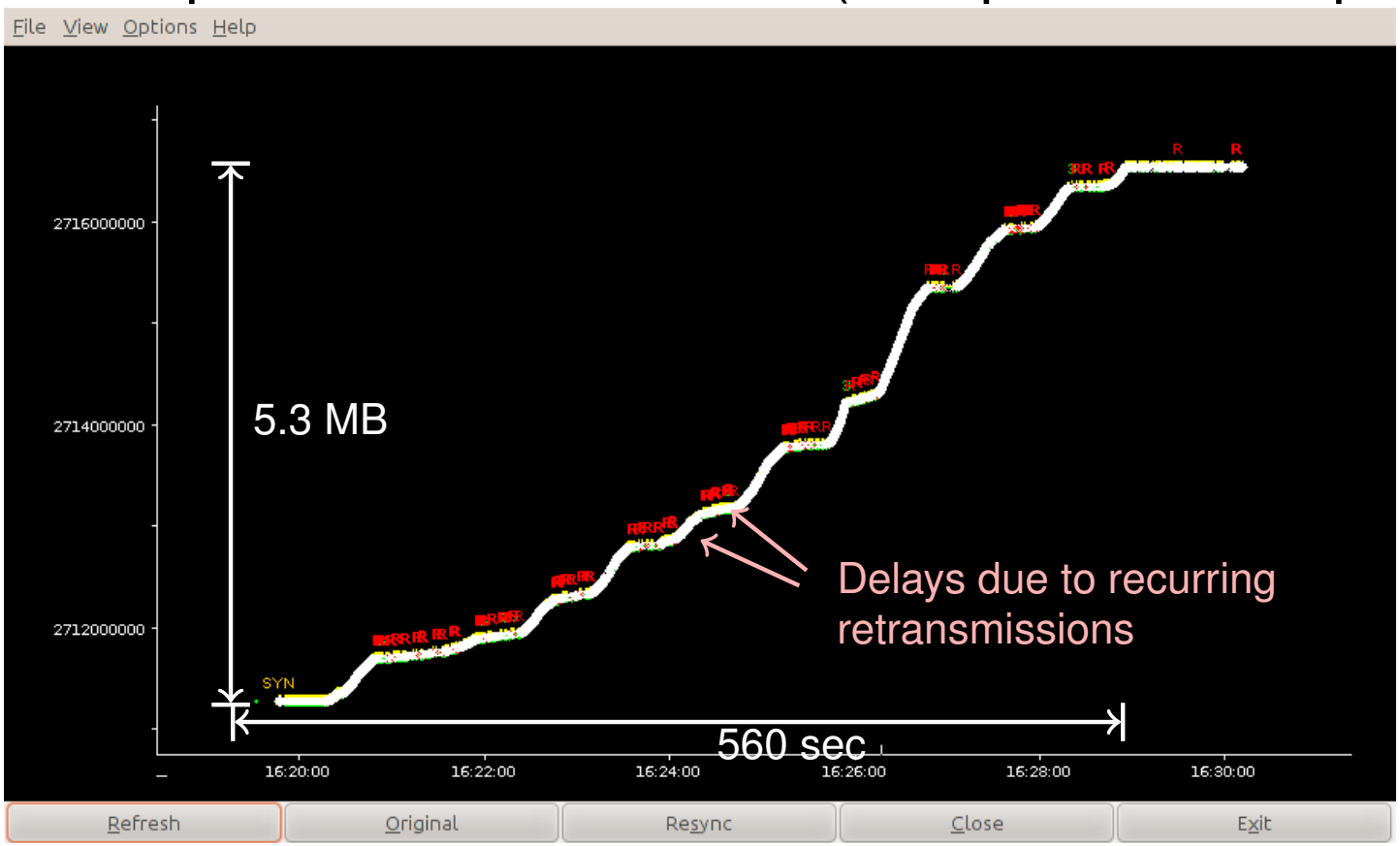
Why TCP trace?

- A sample BGP table transfer (low tput: $\sim$ 76 kbps)



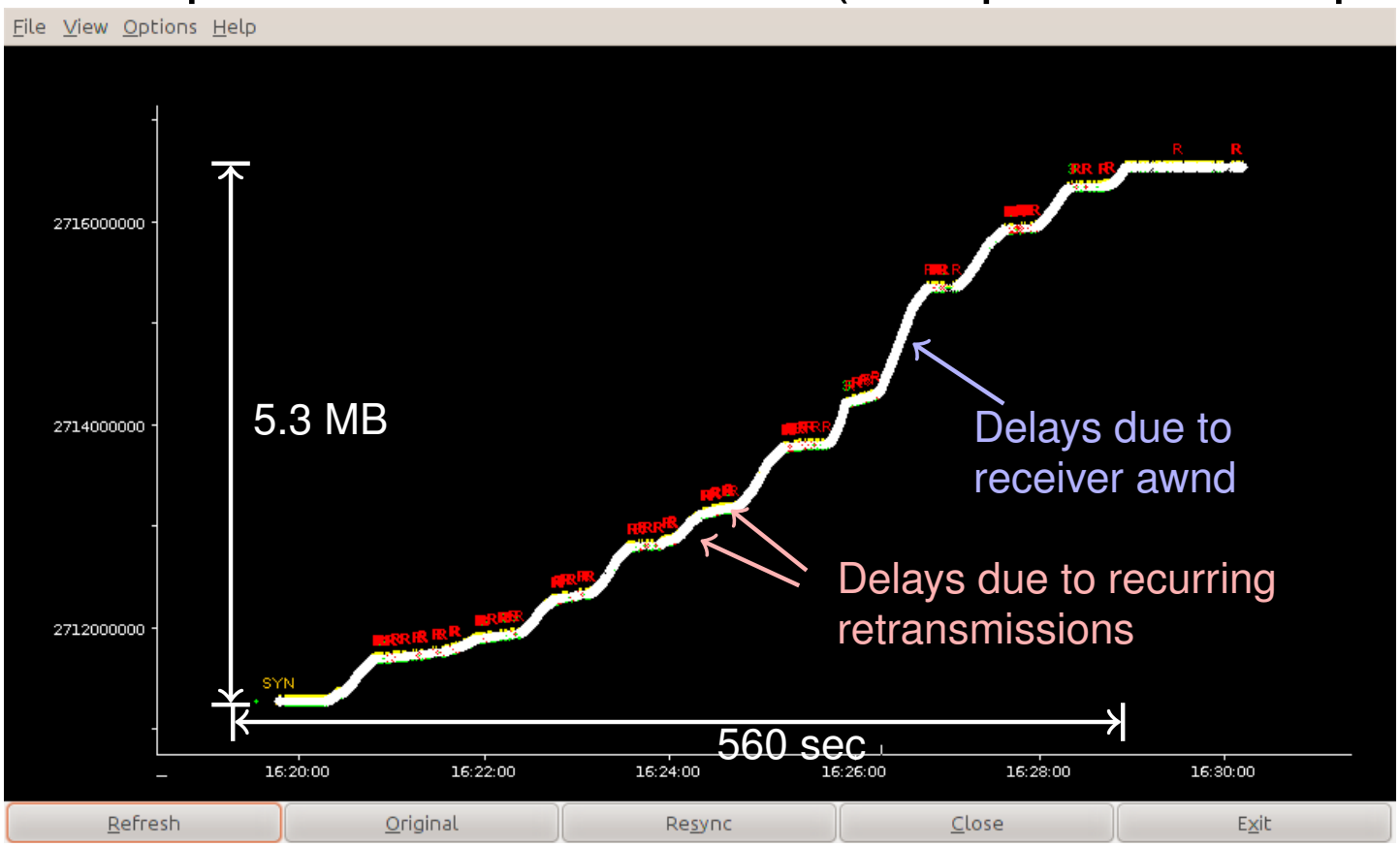
Why TCP trace?

- A sample BGP table transfer (low tput: $\sim$ 76 kbps)



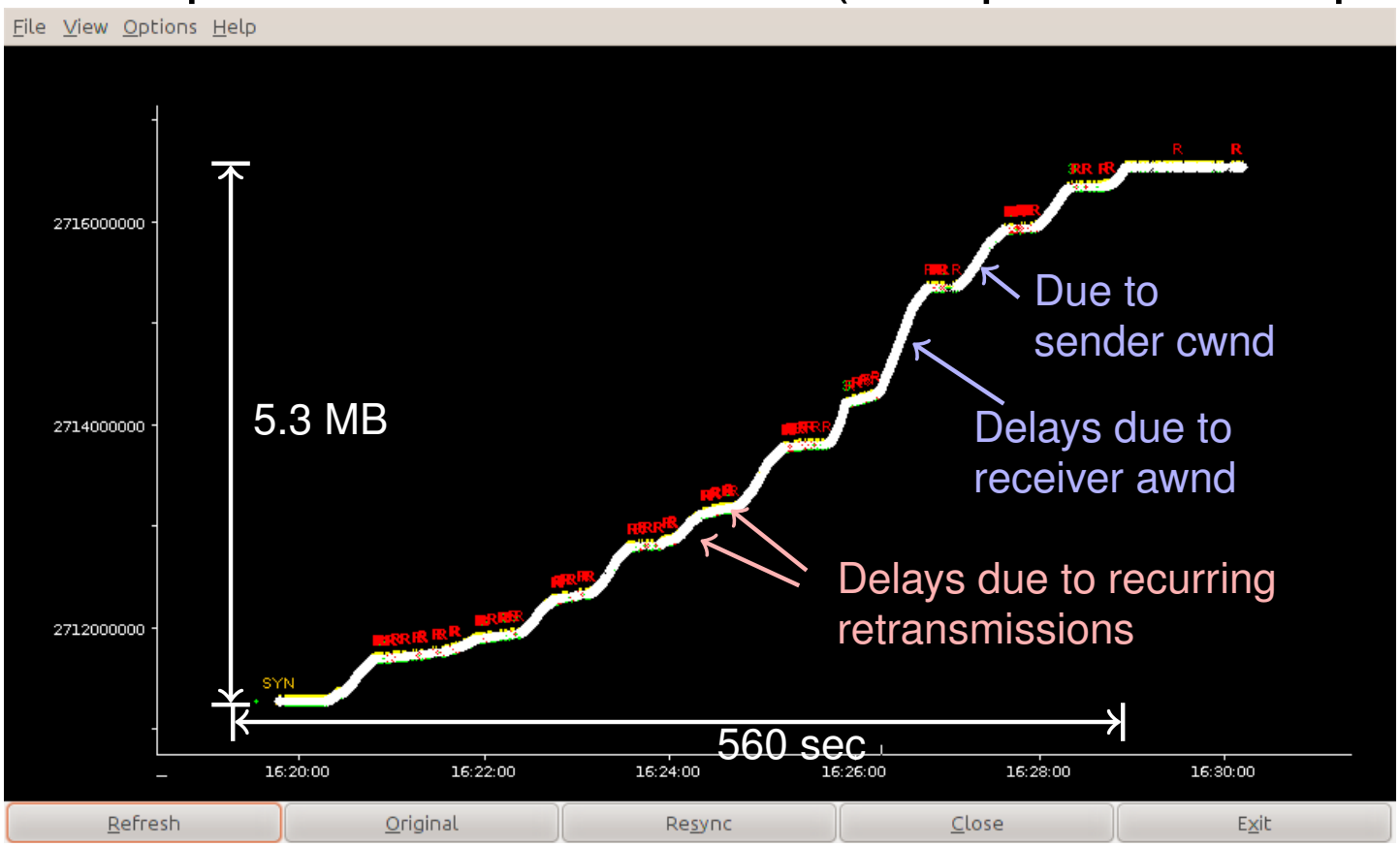
Why TCP trace?

- A sample BGP table transfer (low tput: $\sim$ 76 kbps)



Why TCP trace?

- A sample BGP table transfer (low tput: $\sim$ 76 kbps)



TCP induced delay

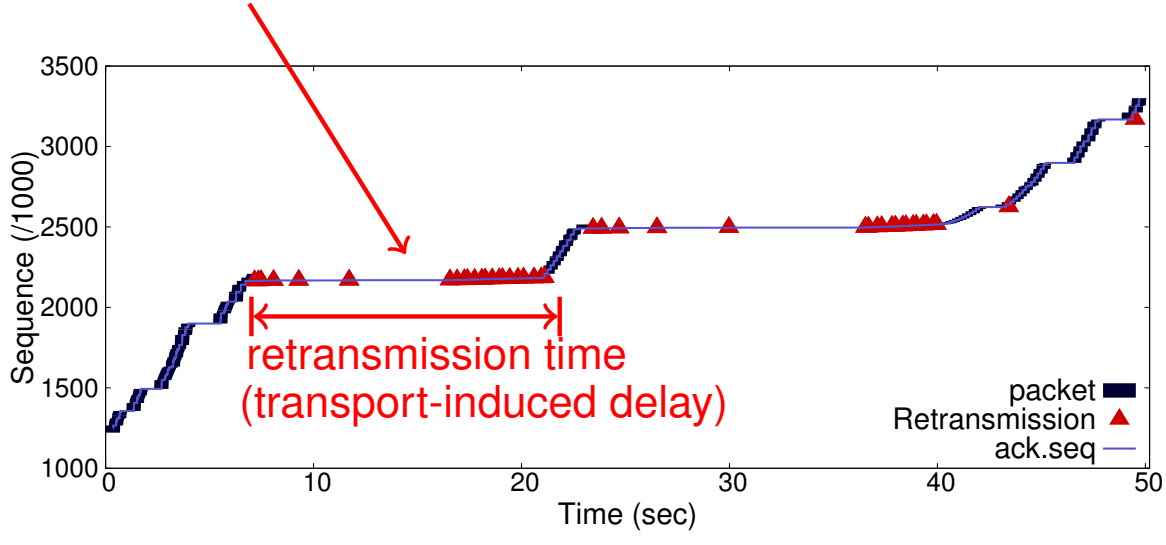
- Message delay seen at the BGP level

Timestamp	Delay	Prefix	Path
1235728588	1	66.154.112.0/24	19080 22298 30092
1235728588	1	66.154.104.0/22	19080 22298 30092
...			
1235728592	4	138.247.0.0/16	1239 13576 14263 23122
1235728592	4	205.151.56.0/24	174 16532
...			
1235728597	9	206.209.232.0/21	7018 16910
1235728597	9	219.239.44.0/23	10026 7497 7497 7497 17964
...			
1235728601	13	92.255.72.0/22	8342 20632 47168

TCP induced delay

- Message delay seen at the BGP level

Timestamp	Delay	Prefix	Path
1235728588	1	66.154.112.0/24	19080 22298 30092
1235728588	1	66.154.104.0/22	19080 22298 30092
...			
1235728592	4	138.247.0.0/16	1239 13576 14263 23122
1235728592	4	205.151.56.0/24	174 16532
...			
1235728597	9	206.209.232.0/21	7018 16910
1235728597	9	219.239.44.0/23	10026 7497 7497 7497 17964
...			
1235728601	13	92.255.72.0/22	8342 20632 47168

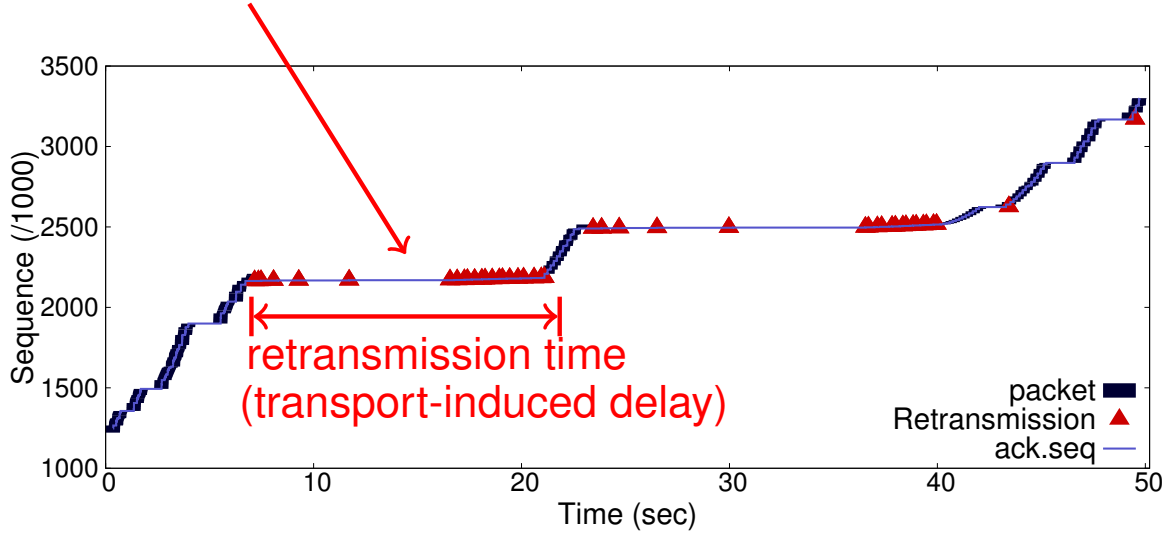


TCP induced delay

- Message delay seen at the BGP level

Timestamp	Delay	Prefix	Path
1235728588	1	66.154.112.0/24	19080 22298 30092
1235728588	1	66.154.104.0/22	19080 22298 30092
...			
1235728592	4	138.247.0.0/16	1239 13576 14263 23122
1235728592	4	205.151.56.0/24	174 16532
...			
1235728597	9	206.209.232.0/21	7018 16910
1235728597	9	219.239.44.0/23	10026 7497 7497 7497 17964
...			
1235728601	13	92.255.72.0/22	8342 20632 47168

- False attribution?
- Overlook transport issues
- ...



TCP induced delay

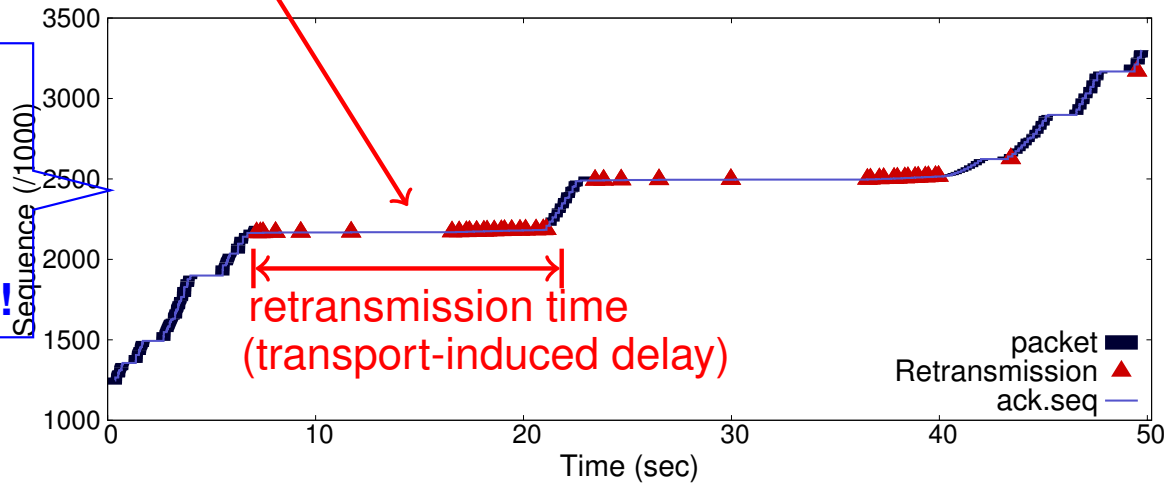
- Message delay seen at the BGP level

Timestamp	Delay	Prefix	Path
1235728588	1	66.154.112.0/24	19080 22298 30092
1235728588	1	66.154.104.0/22	19080 22298 30092
...			
1235728592	4	138.247.0.0/16	1239 13576 14263 23122
1235728592	4	205.151.56.0/24	174 16532
...			
1235728597	9	206.209.232.0/21	7018 16910
1235728597	9	219.239.44.0/23	10026 7497 7497 7497 17964
...			
1235728601	13	92.255.72.0/22	8342 20632 47168

- False attribution?
- Overlook transport issues
- ...

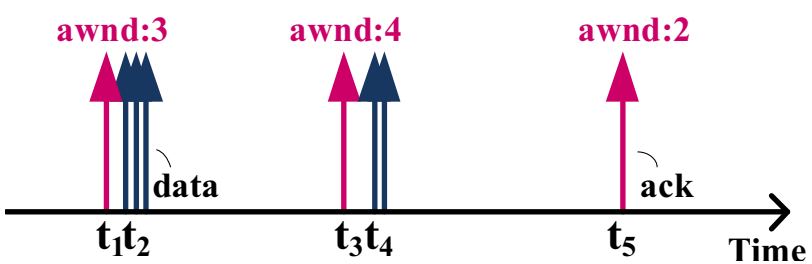
- How to detect?
- Visual inspection
- Hard to quantify

Call for a delay analysis tool!!



T-DAT: TCP Delay Analysis Tool

- Automatically identify delay contributors
  - Which $\sim$ Application (BGP), TCP, Network
  - Where $\sim$ Sender, Receiver, Network path
- Series-based approach
  - Convert TCP trace to various event series
    - Represent TCP transmission, retransmission, ACK, etc
    - 2-tuple: [start_time, end_time], event_data
  - Infer delay contributors from the series



Series: Advertised window

event duration	event data
[t1, t3]	3
[t3, t5]	4
[t5, -]	2

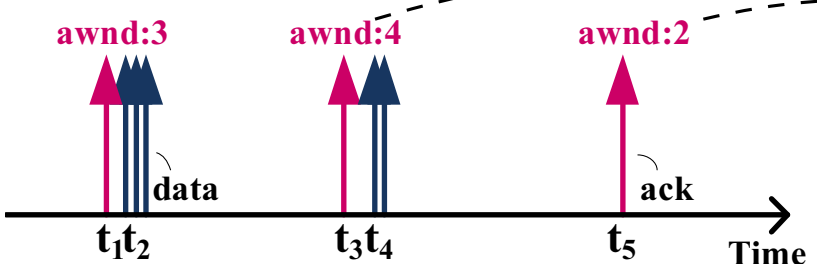
Series: Outstanding packets

event duration	event data
[t2, t3]	3
[t4, t5]	2

...

T-DAT: TCP Delay Analysis Tool

- Automatically identify delay contributors
  - Which $\sim$ Application (BGP), TCP, Network
  - Where $\sim$ Sender, Receiver, Network path
- Series-based approach
  - Convert TCP trace to various event series
    - Represent TCP transmission, retransmission, ACK, etc
    - 2-tuple: [start_time, end_time], event_data
  - Infer delay contributors from the series



Series: Advertised window

event duration	event data
[t1, t3]	3
[t3, t5]	4
[t5, -]	2

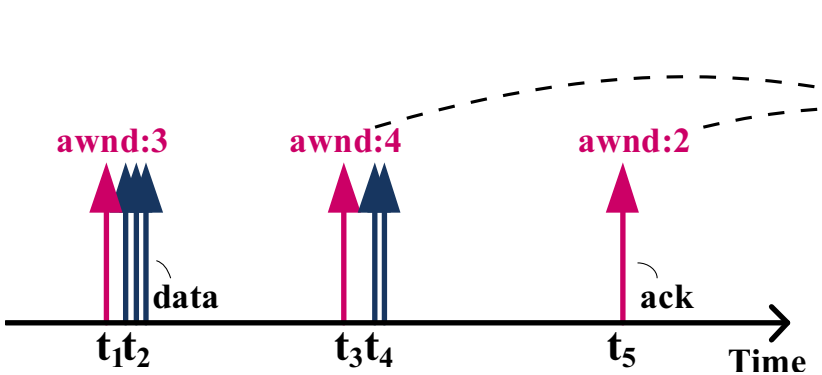
Series: Outstanding packets

event duration	event data
[t2, t3]	3
[t4, t5]	2

...

T-DAT: TCP Delay Analysis Tool

- Automatically identify delay contributors
  - Which $\sim$ Application (BGP), TCP, Network
  - Where $\sim$ Sender, Receiver, Network path
- Series-based approach
  - Convert TCP trace to various event series
    - Represent TCP transmission, retransmission, ACK, etc
    - 2-tuple: [start_time, end_time], event_data
  - Infer delay contributors from the series



Series: Advertised window

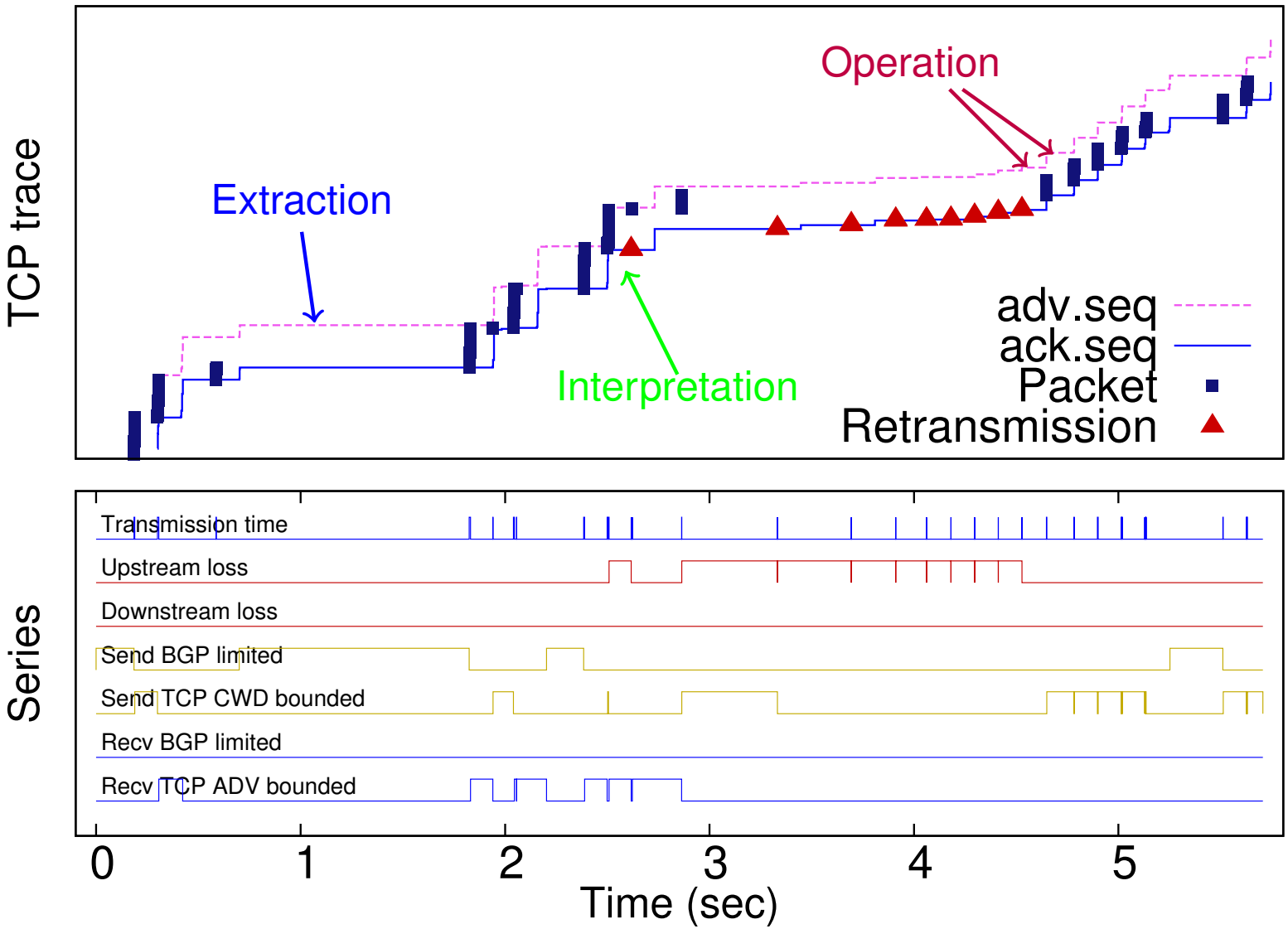
event duration	event data
[t1, t3]	3
[t3, t5]	4
[t5, -]	2

Series: Outstanding packets

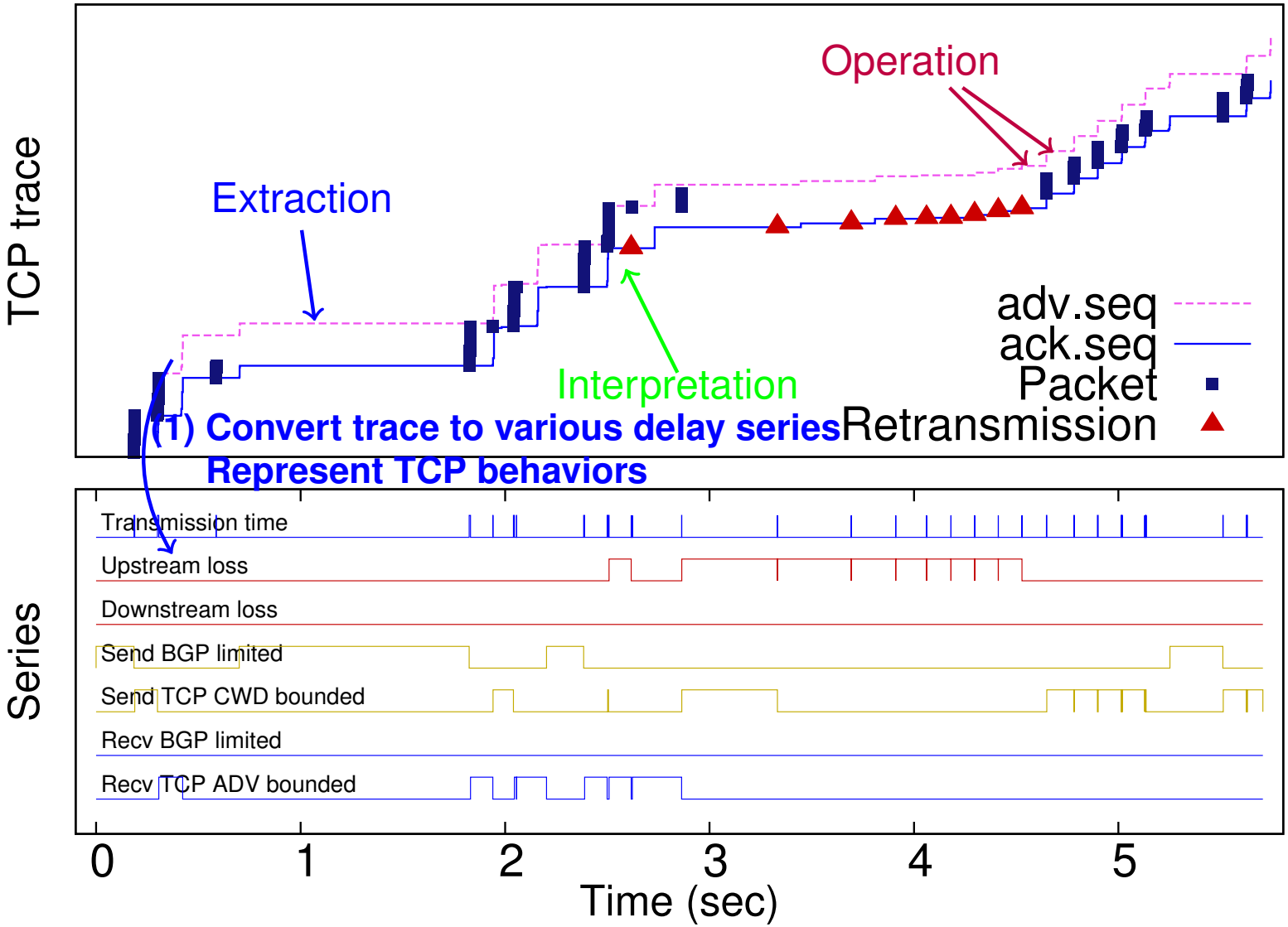
event duration	event data
[t2, t3]	3
[t4, t5]	2

...

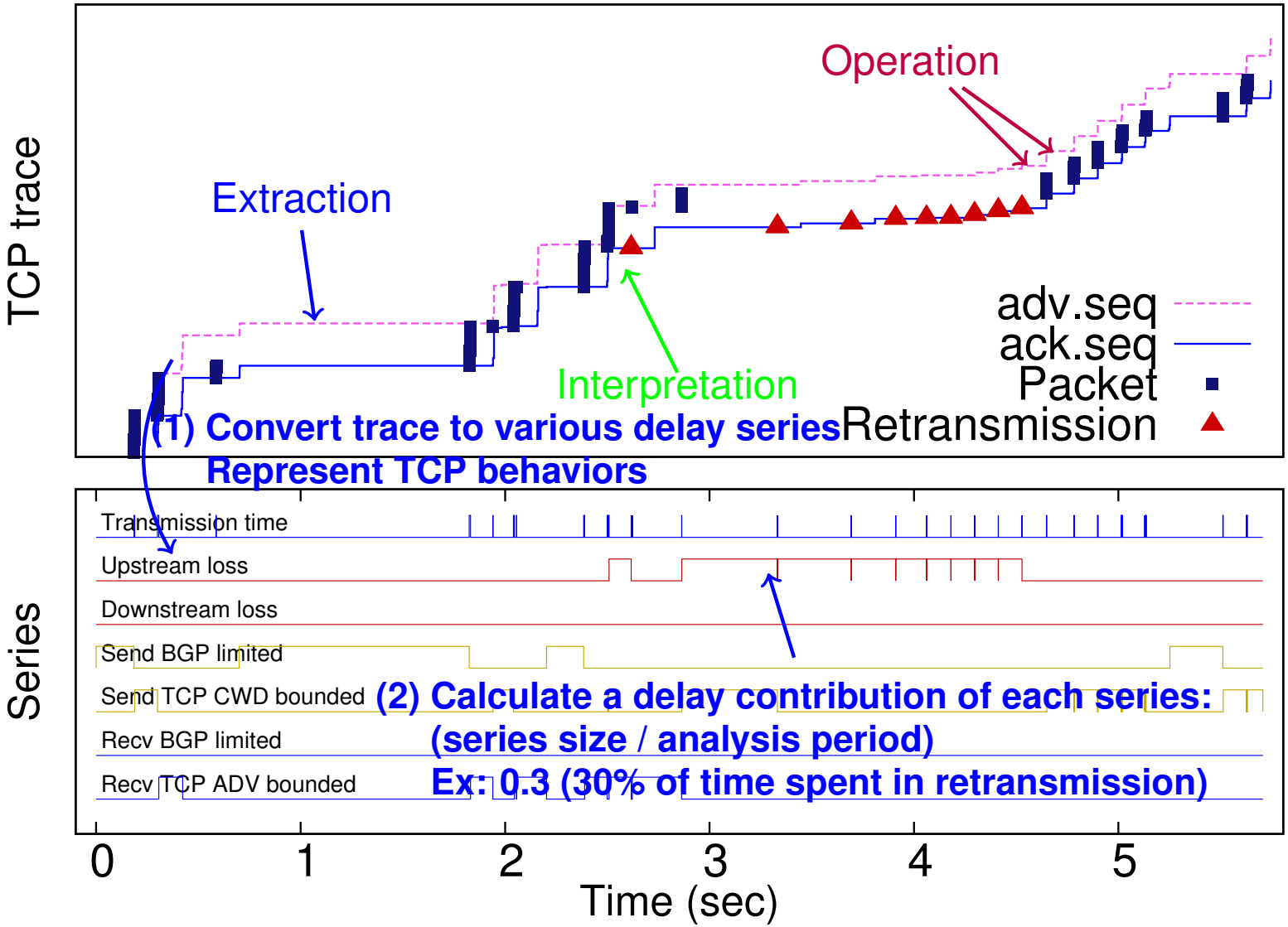
T-DAT operations in brief



T-DAT operations in brief



T-DAT operations in brief



Output delay contributors

- 8 delay contributors

$$\textit{Delay vector}: \overrightarrow{V} = (r_1, r_2, \ldots, r_8), \quad r_i = \frac{size(Series_i)}{AnalysisPeriod}, \quad i = 1 \ldots 8$$

  - Ex: ($\underbrace{0.27, 0.43,}_{\text{BGP(2)}}$ $\underbrace{0.11, 0.01,}_{\text{TCP(2)}}$ $\underbrace{0.00, 0.01, 0.01, 0}_{\text{Network(4)}}$)

- 3 contributing groups

$$\textit{Group vector}: \overrightarrow{G} = (R_s, R_r, R_n), \quad R_g = \frac{size(\bigcup Series_i)}{AnalysisPeriod}, \quad g \in \{s, r, n\}$$

  - Ex: ($\underbrace{0.38,}_{\text{Send}}$ $\underbrace{0.43,}_{\text{Recv}}$ $\underbrace{0.01}_{\text{Net}}$)

Output delay contributors

- 8 delay contributors

$$\textit{Delay vector}: \overrightarrow{V} = (r_1, r_2, \ldots, r_8), \quad r_i = \frac{size(Series_i)}{AnalysisPeriod}, \quad i = 1 \ldots 8$$

  - Ex: (0.27, 0.43, | 0.11, 0.01, | 0.00, 0.01, 0.01, 0)
    - BGP(2)
    - TCP(2)
    - Network(4)

- 3 contributing groups

$$\textit{Group vector}: \overrightarrow{G} = (R_s, R_r, R_n), \quad R_g = \frac{size(\bigcup Series_i)}{AnalysisPeriod}, \quad g \in \{s, r, n\}$$

  - Ex: (0.38, 0.43, 0.01)
    - Send
    - Recv
    - Net

Output delay contributors

- 8 delay contributors

$$\textit{Delay vector}: \overrightarrow{V} = (r_1, r_2, \ldots, r_8), \quad r_i = \frac{size(Series_i)}{AnalysisPeriod}, \quad i = 1 \ldots 8$$

  - Ex: (0.27, 0.43, 0.11, 0.01, 0.00, 0.01, 0.01, 0)
    - BGP(2): 0.27, 0.43
    - TCP(2): 0.11, 0.01
    - Network(4): 0.00, 0.01, 0.01, 0

- 3 contributing groups

$$\textit{Group vector}: \overrightarrow{G} = (R_s, R_r, R_n), \quad R_g = \frac{size(\bigcup Series_i)}{AnalysisPeriod}, \quad g \in \{s, r, n\}$$

  - Ex: (0.38, 0.43, 0.01)
    - Send: 0.38
    - Recv: 0.43
    - Net: 0.01

Output delay contributors

- 8 delay contributors

$$\textit{Delay vector}: \overrightarrow{V} = (r_1, r_2, \ldots, r_8), \quad r_i = \frac{size(Series_i)}{AnalysisPeriod}, \quad i = 1 \ldots 8$$

  - Ex: (0.27, 0.43, 0.11, 0.01, 0.00, 0.01, 0.01, 0)
    - BGP(2): 0.27, 0.43
    - TCP(2): 0.11, 0.01
    - Network(4): 0.00, 0.01, 0.01, 0

- 3 contributing groups

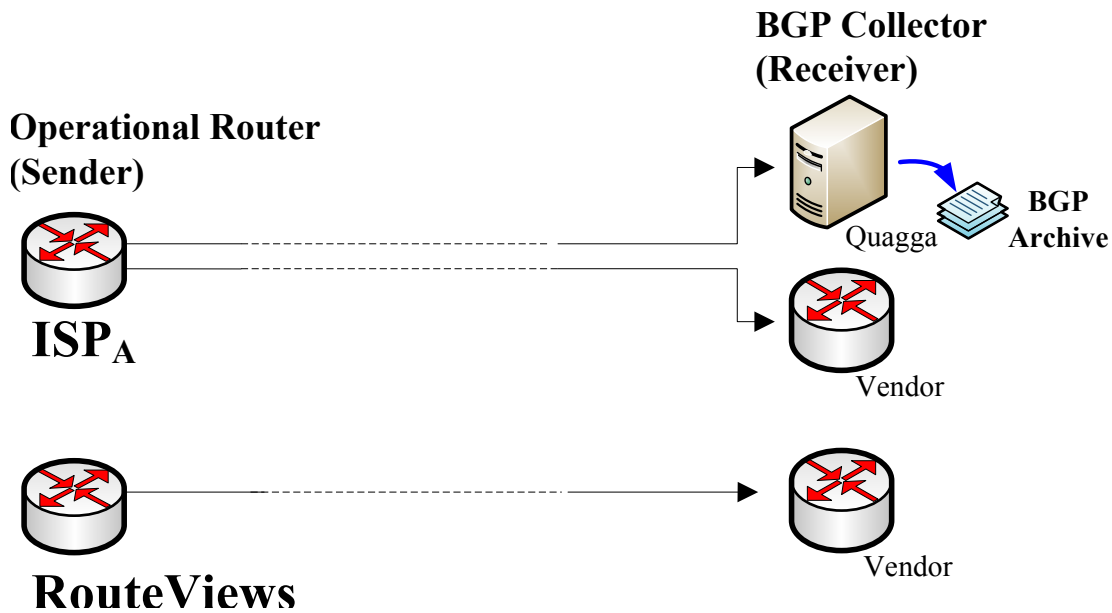
$$\textit{Group vector}: \overrightarrow{G} = (R_s, R_r, R_n), \quad R_g = \frac{size(\bigcup Series_i)}{AnalysisPeriod}, \quad g \in \{s, r, n\}$$

  - Ex: (0.38, 0.43, 0.01)
    - Send: 0.38
    - Recv: 0.43
    - Net: 0.01

Apply to BGP table transfers

Apply to BGP Data (ISP_A and RouteViews)

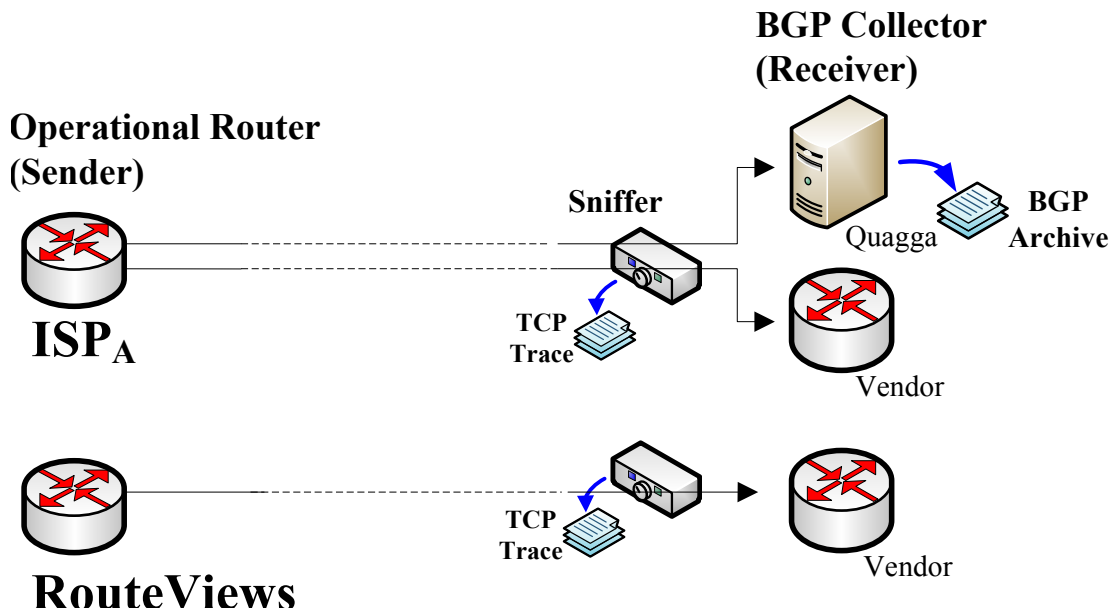
Trace Name	Duration	Collector	# Pkts/Bytes (M/GB)	# Rtrs	TCP	BGP	# BGP Tables Transfers
ISP_A-1	08.05 ∼ 09.04	Vendor	1023 / 218	24	Yes	-	10471
ISP_A-2	08.05 ∼ 09.04	Quagga	909 / 138	27	Yes	Yes	180
	09.09 ∼ 10.09		1296 / 219				219
	10.11 ∼ 11.01		492 / 81				37
RV	10.11 ∼ 11.01	Vendor	176 / 47	59*	Yes	-	94



DISCLAIMER: Monitoring sessions!!
NOT actual operational BGP sessions

Apply to BGP Data (ISP_A and RouteViews)

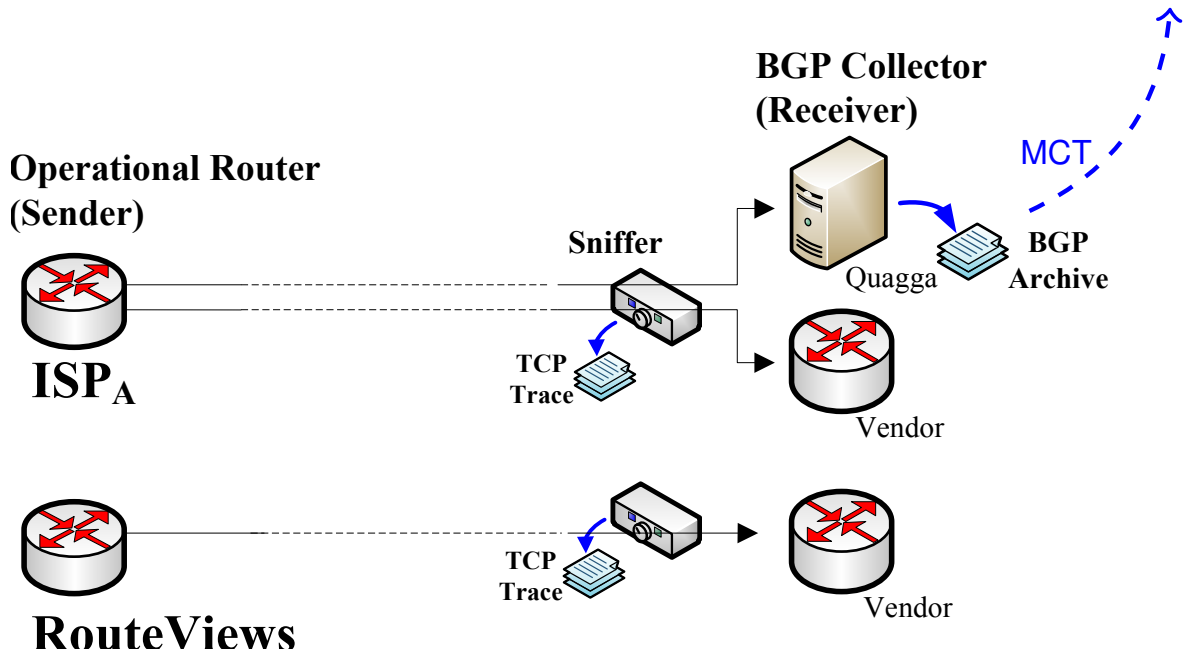
Trace Name	Duration	Collector	# Pkts/Bytes (M/GB)	# Rtrs	TCP	BGP	# BGP Tables Transfers
ISP_A-1	08.05 ∼ 09.04	Vendor	1023 / 218	24	Yes	-	10471
ISP_A-2	08.05 ∼ 09.04	Quagga	909 / 138	27	Yes	Yes	180
	09.09 ∼ 10.09		1296 / 219				219
	10.11 ∼ 11.01		492 / 81				37
RV	10.11 ∼ 11.01	Vendor	176 / 47	59*	Yes	-	94



DISCLAIMER: Monitoring sessions!!
NOT actual operational BGP sessions

Apply to BGP Data (ISP_A and RouteViews)

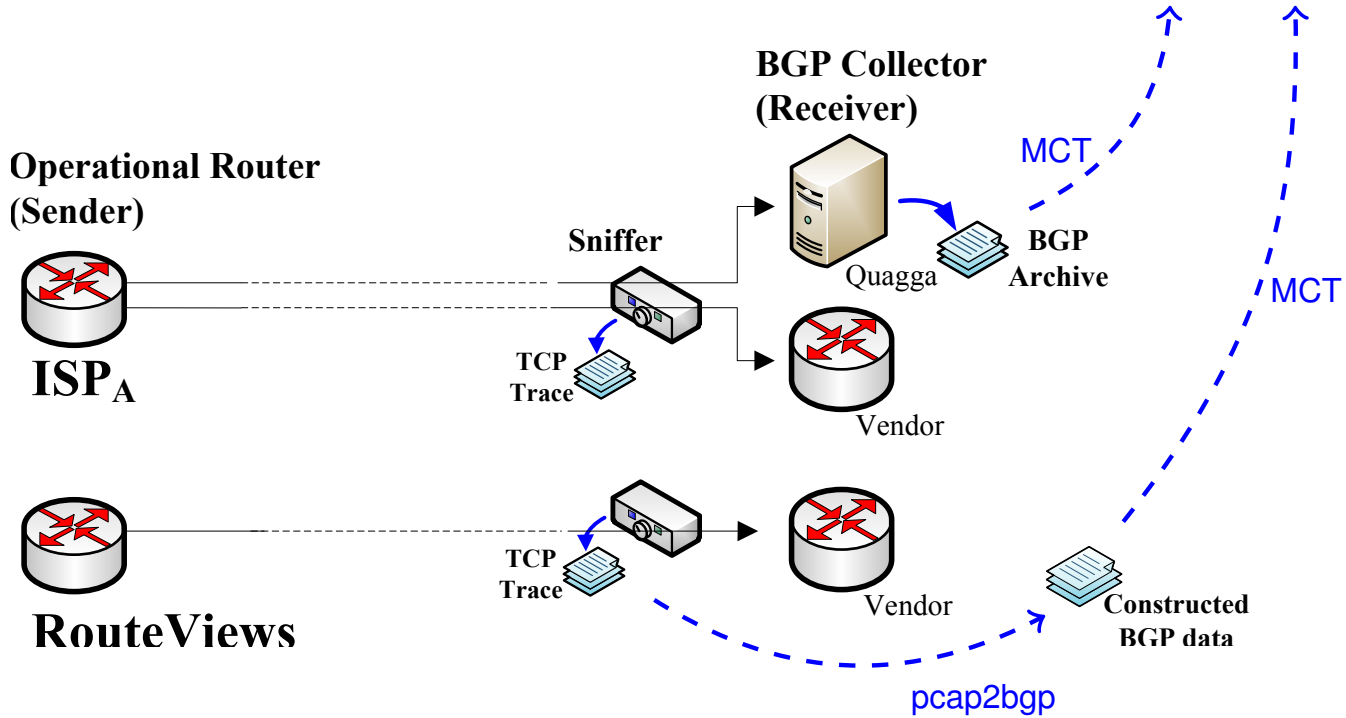
Trace Name	Duration	Collector	# Pkts/Bytes (M/GB)	# Rtrs	TCP	BGP	# BGP Tables Transfers
ISP_A-1	08.05 ∼ 09.04	Vendor	1023 / 218	24	Yes	-	10471
ISP_A-2	08.05 ∼ 09.04	Quagga	909 / 138	27	Yes	Yes	180
	09.09 ∼ 10.09		1296 / 219				219
	10.11 ∼ 11.01		492 / 81				37
RV	10.11 ∼ 11.01	Vendor	176 / 47	59*	Yes	-	94



DISCLAIMER: Monitoring sessions!!
NOT actual operational BGP sessions

Apply to BGP Data (ISP_A and RouteViews)

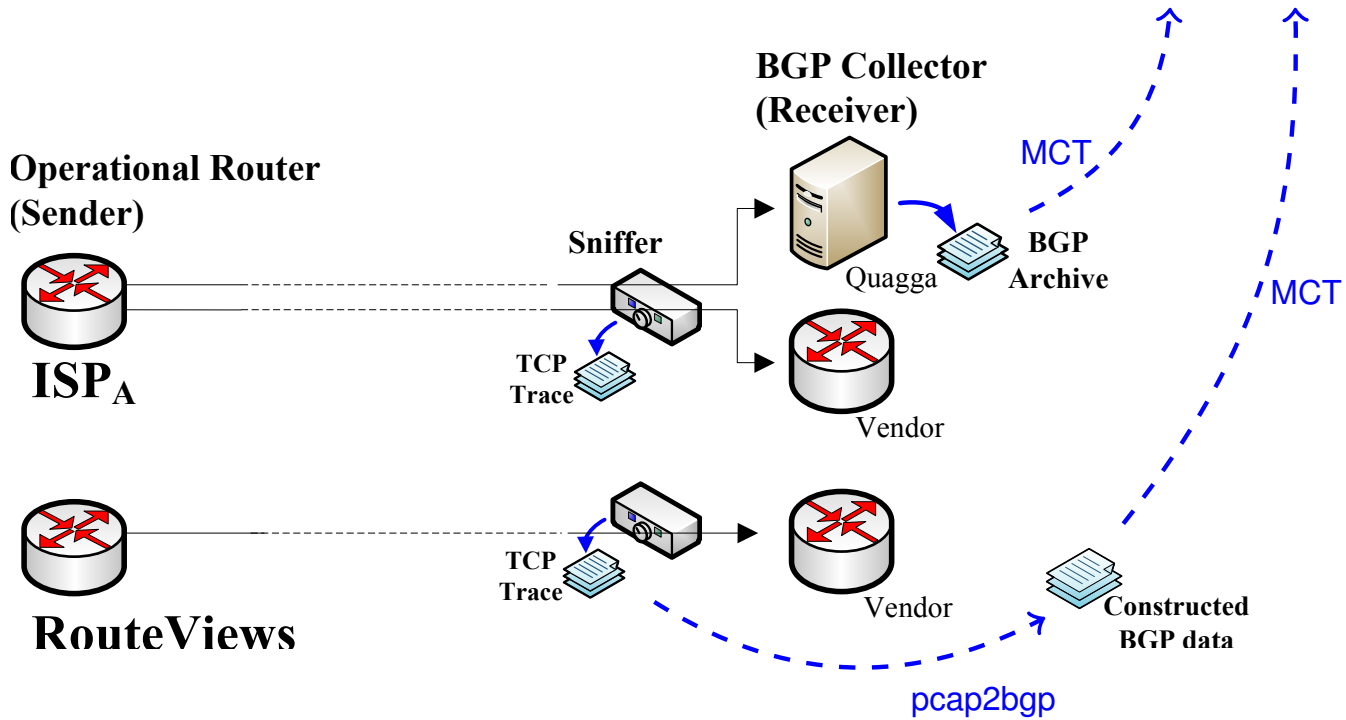
Trace Name	Duration	Collector	# Pkts/Bytes (M/GB)	# Rtrs	TCP	BGP	# BGP Tables Transfers
ISP_A-1	08.05 ∼ 09.04	Vendor	1023 / 218	24	Yes	-	10471
ISP_A-2	08.05 ∼ 09.04	Quagga	909 / 138	27	Yes	Yes	180
	09.09 ∼ 10.09		1296 / 219				219
	10.11 ∼ 11.01		492 / 81				37
RV	10.11 ∼ 11.01	Vendor	176 / 47	59*	Yes	-	94



DISCLAIMER: Monitoring sessions!!
NOT actual operational BGP sessions

Apply to BGP Data (ISP_A and RouteViews)

Trace Name	Duration	Collector	# Pkts/Bytes (M/GB)	# Rtrs	TCP	BGP	# BGP Tables Transfers
ISP_A-1	08.05 ∼ 09.04	Vendor	1023 / 218	24	Yes	-	10471
ISP_A-2	08.05 ∼ 09.04	Quagga	909 / 138	27	Yes	Yes	180
	09.09 ∼ 10.09		1296 / 219				219
	10.11 ∼ 11.01		492 / 81				37
RV	10.11 ∼ 11.01	Vendor	176 / 47	59*	Yes	-	94

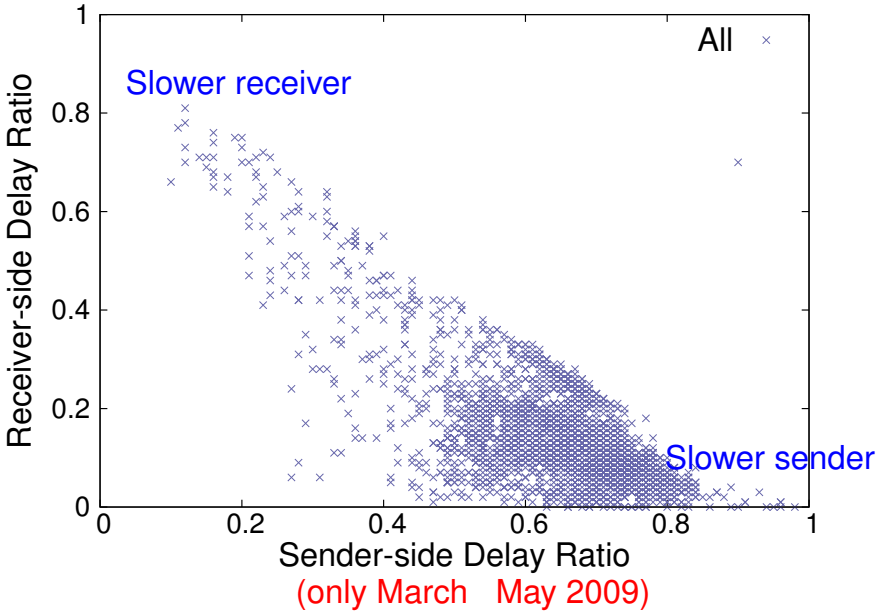


DISCLAIMER: Monitoring sessions!!
NOT actual operational BGP sessions

Find delay contributors - ISP_A

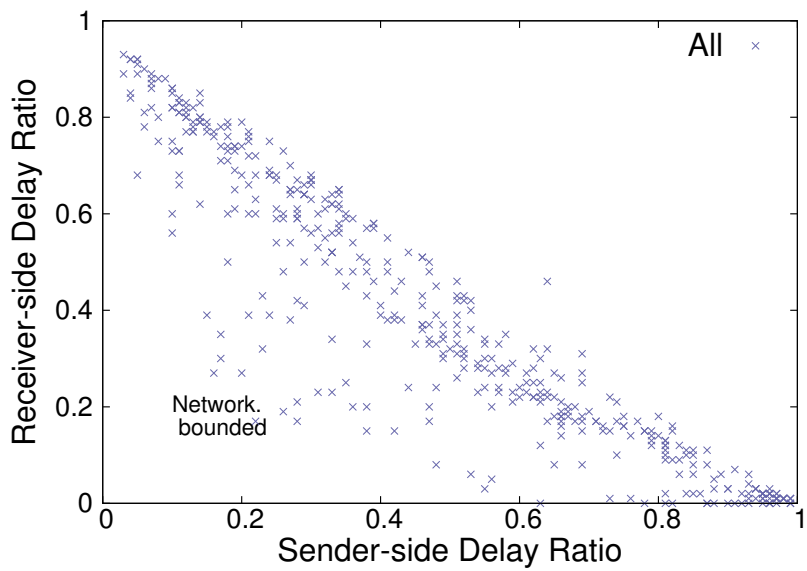
- ISP_A table transfers
  - Plot the group delay ratio, Ex: (0.7, 0.3, 0.0)
  - Network delay is usually low
  - The failing side of a session tends to have more impact

- Vendor collector



(only March May 2009)

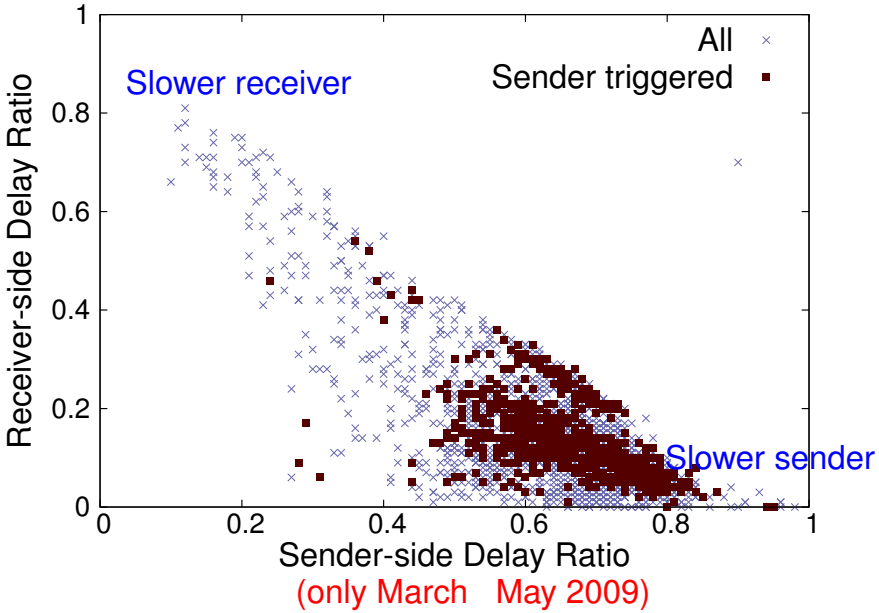
- Quagga collector



Find delay contributors - ISP_A

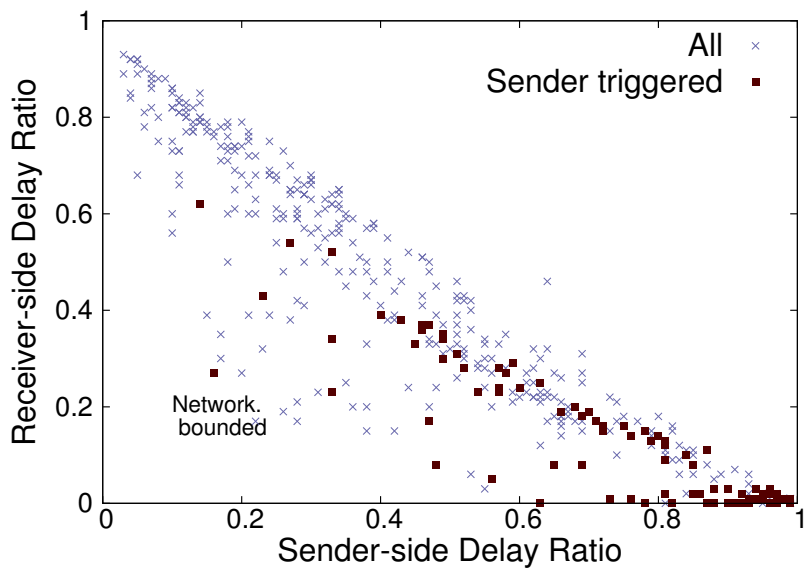
- ISP_A table transfers
  - Plot the group delay ratio, Ex: (0.7, 0.3, 0.0)
  - Network delay is usually low
  - The failing side of a session tends to have more impact

- Vendor collector



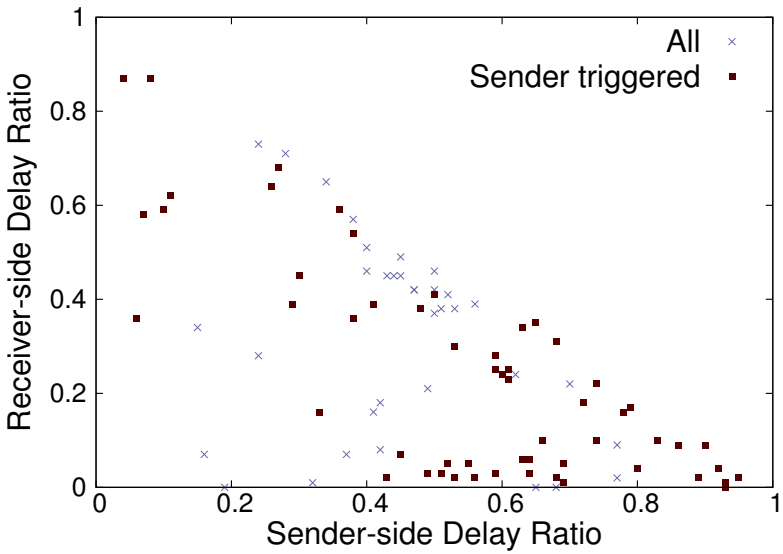
(only March May 2009)

- Quagga collector



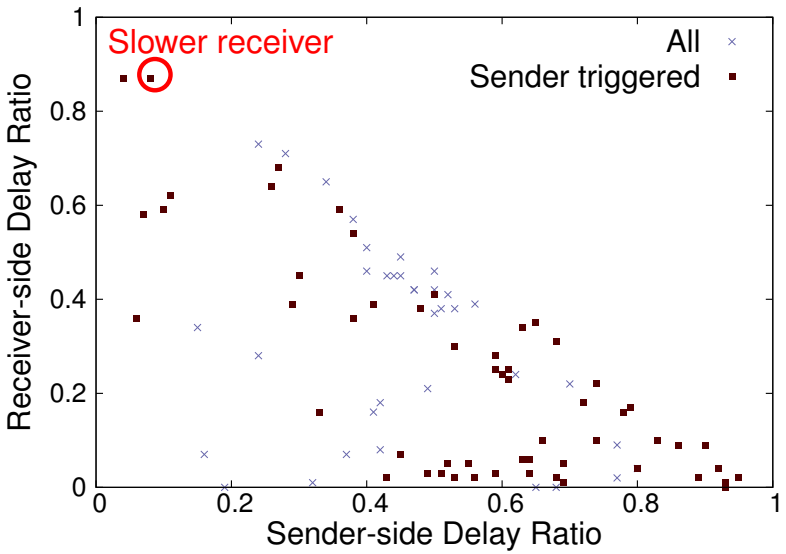
Find delay contributors - RouteViews

- RouteViews table transfers
  - From heterogeneous routers
  - Across Internet

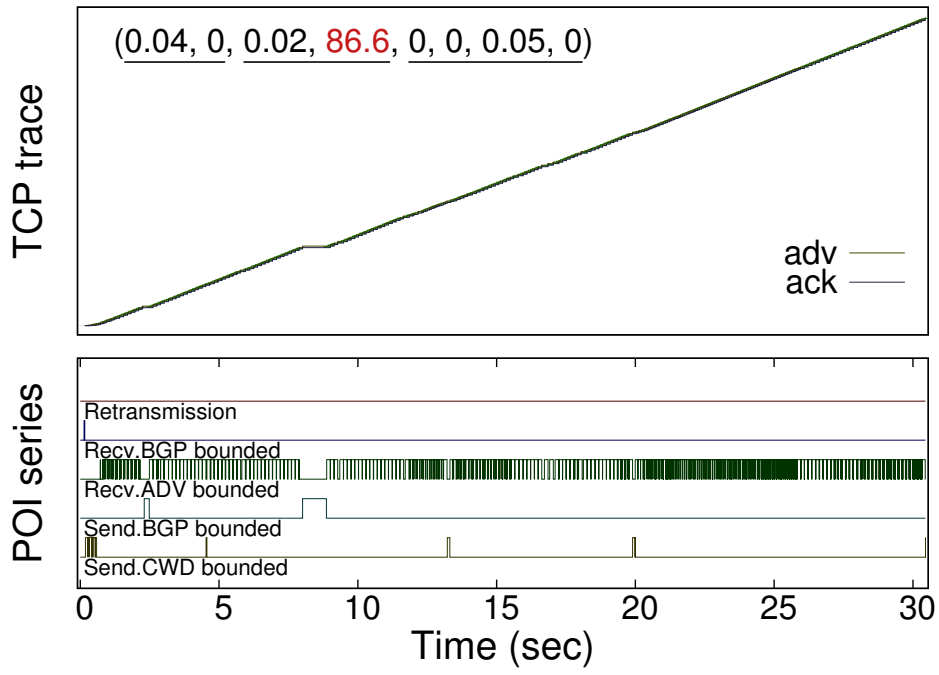


Find delay contributors - RouteViews

- RouteViews table transfers
  - From heterogeneous routers
  - Across Internet

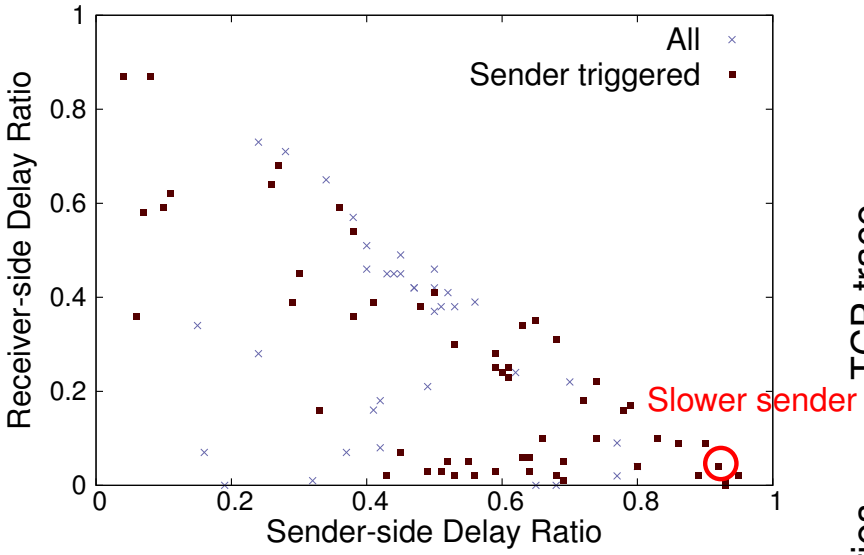


- Slower receiver (30 seconds)
  - 16KB adv. window for 86% of time
  - tput $\sim$ adv. window / RTT

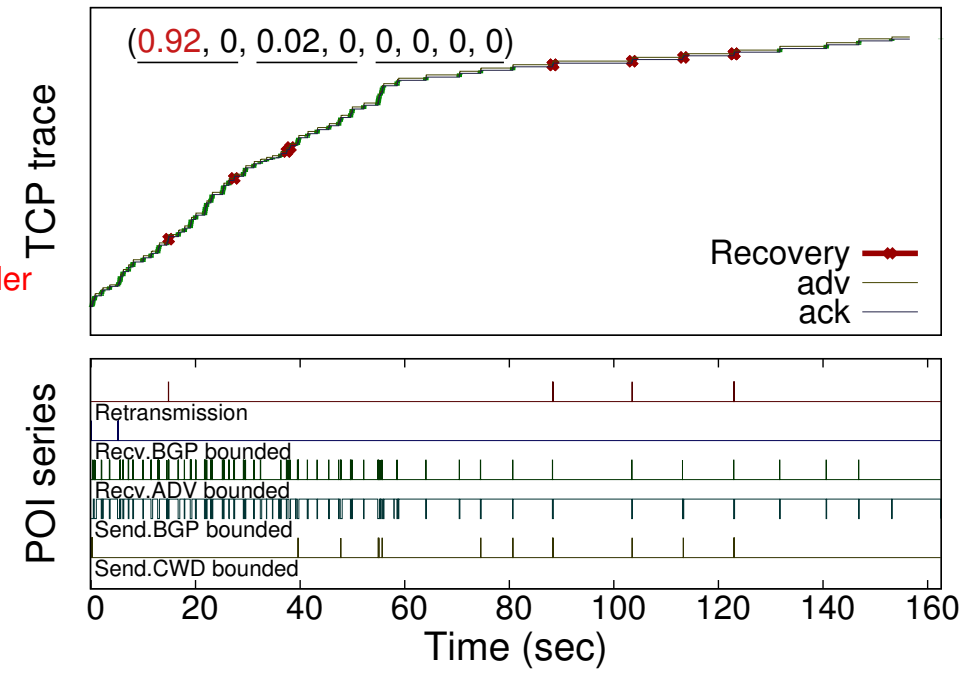


Find delay contributors - RouteViews

- RouteViews table transfers
  - From heterogeneous routers
  - Across Internet

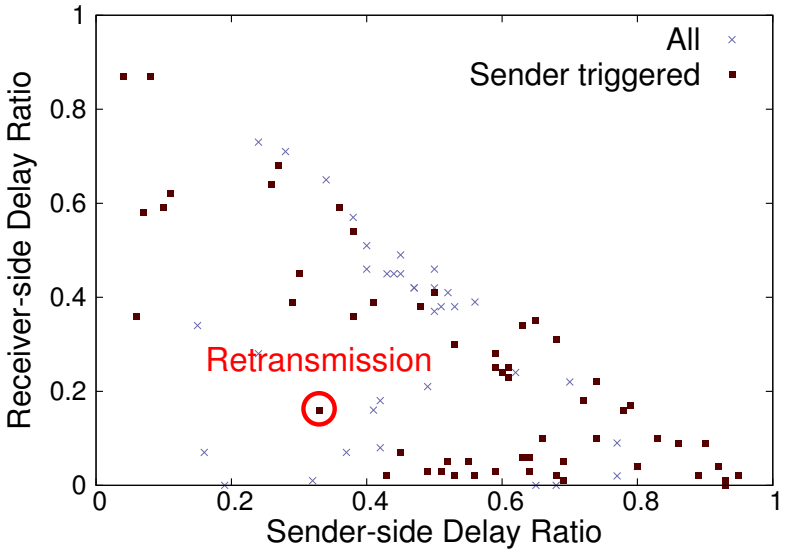


- Lazy sender (160 seconds)
  - The sender often pauses transmission
  - Remains idle for 92% of time

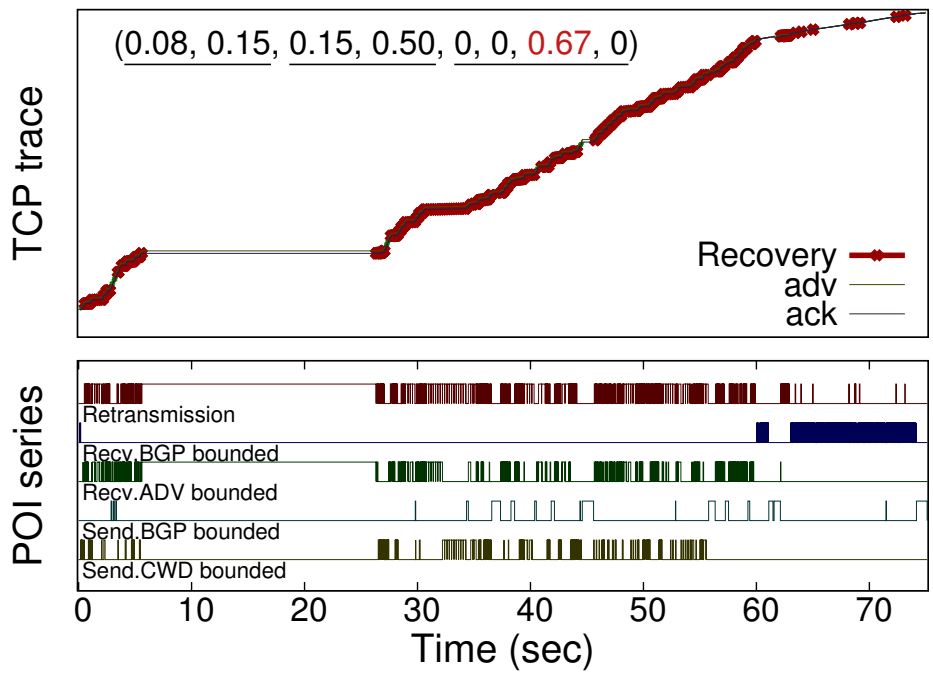


Find delay contributors - RouteViews

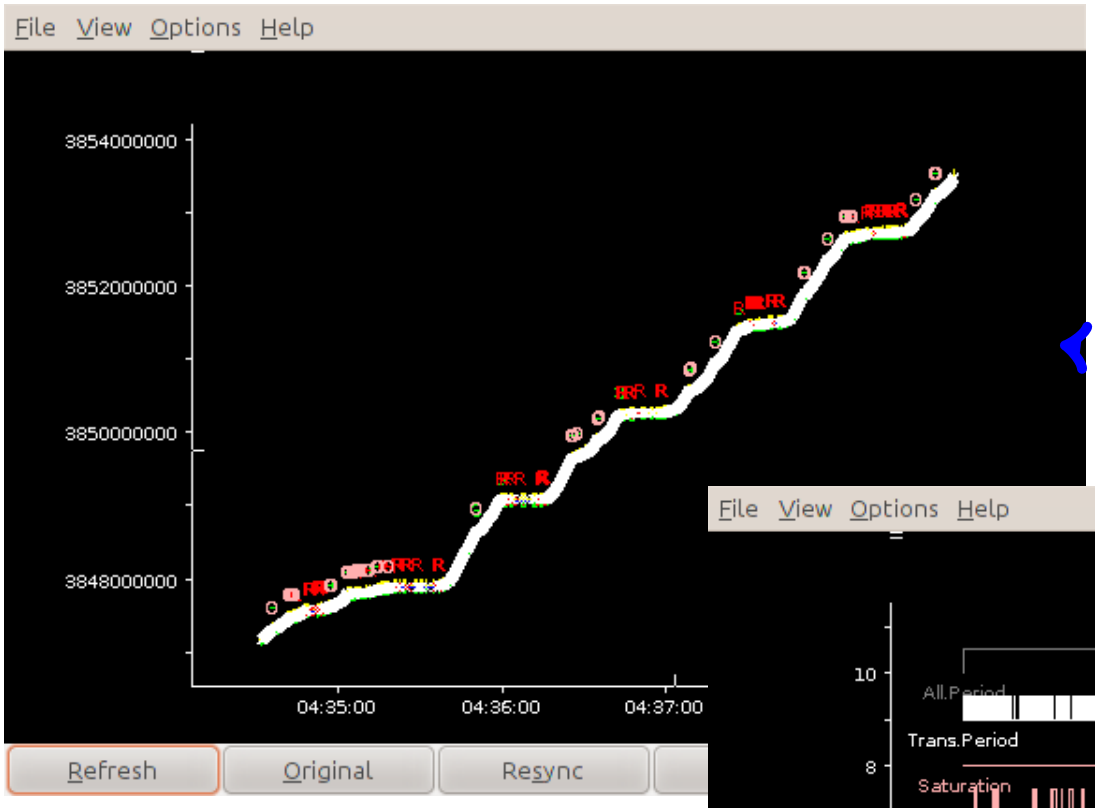
- RouteViews table transfers
  - From heterogeneous routers
  - Across Internet



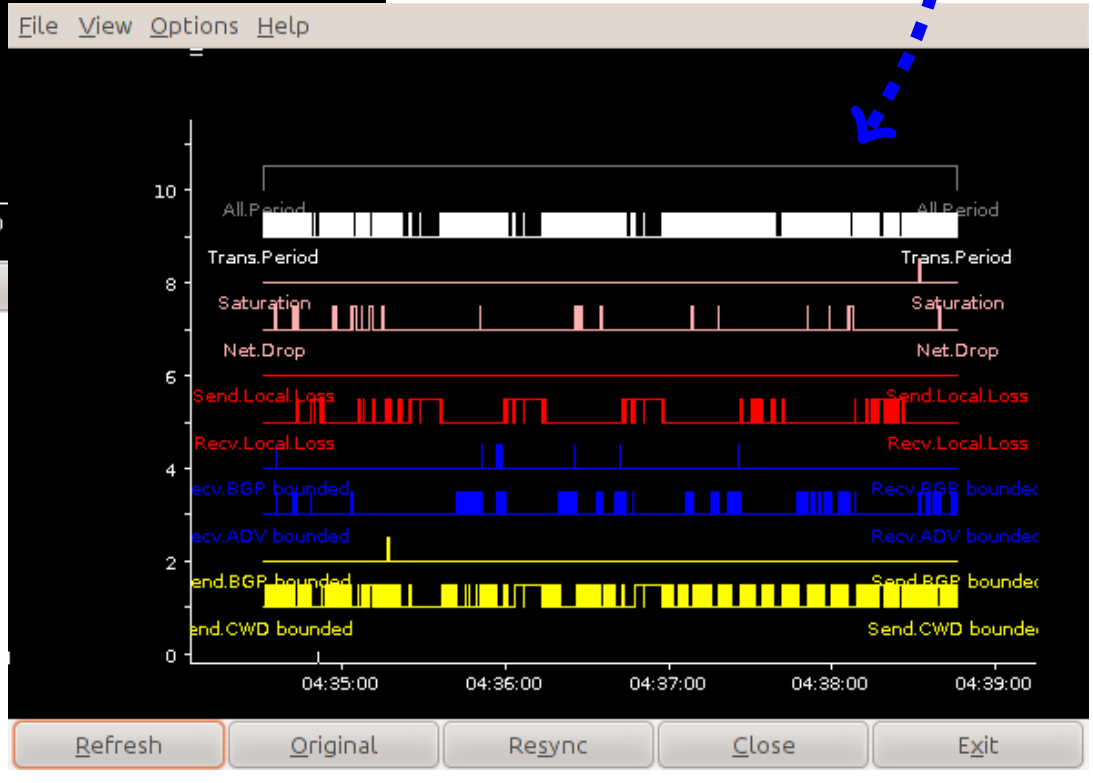
- Congestion limited (70 seconds)
  - 67% time in recovering the packet loss



Example Screenshot



$ tcptrace -G **tcp.pcap** > tcp_tsg.xpl
$ t-dat tcp_tsg.xpl > tcp_ser.xpl
$ BGPlot tcp_tsg.xpl tcp_ser.xpl

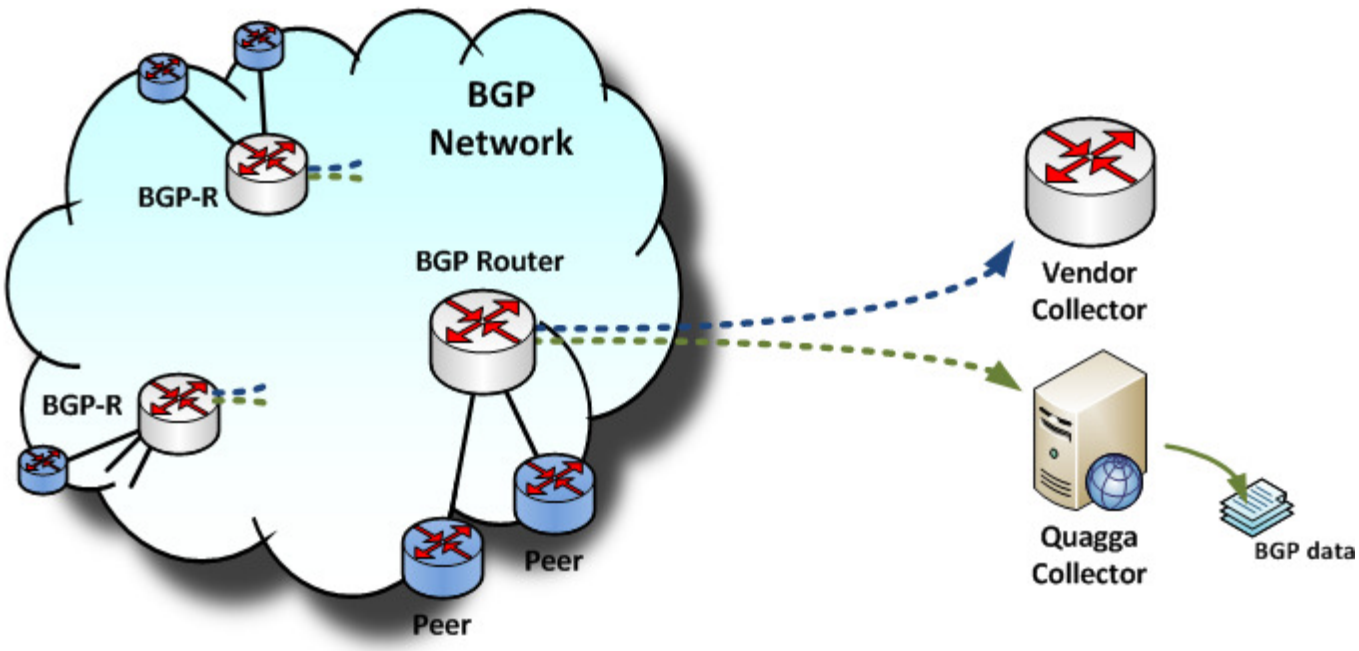


Summary

- A TCP delay analysis tool
  - Require TCP trace (tcpdump)
  - Automatically suggest the reasons of transfer delay
- Demonstrate the usage using BGP monitoring data
  - One ISP and RouteViews
  - Can apply to operational sessions
- The tool and detail report will be available at (http://irl.cs.ucla.edu/bgpmicro)

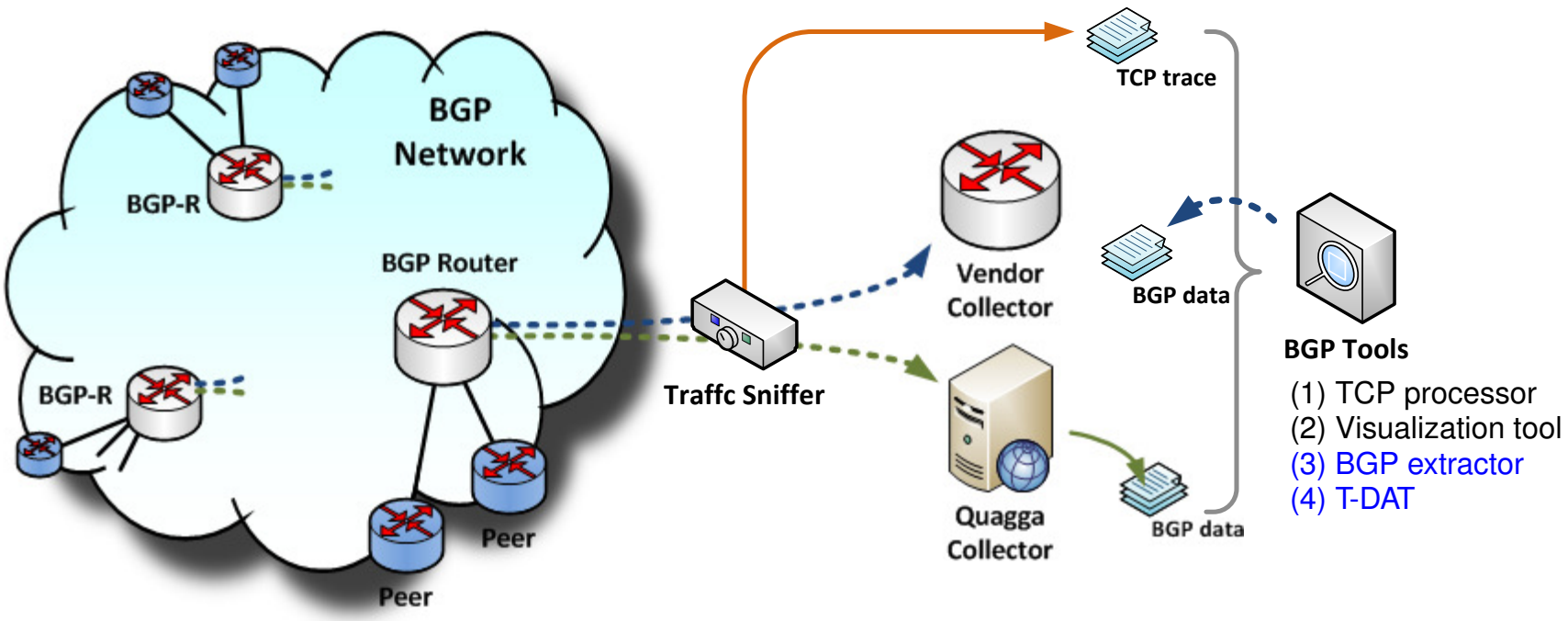
Potential Usage

- Augment the existing BGP monitoring settings
  - Enable TCP investigation
  - Reduce transport-induced artifact
- Allow potential in-band monitoring



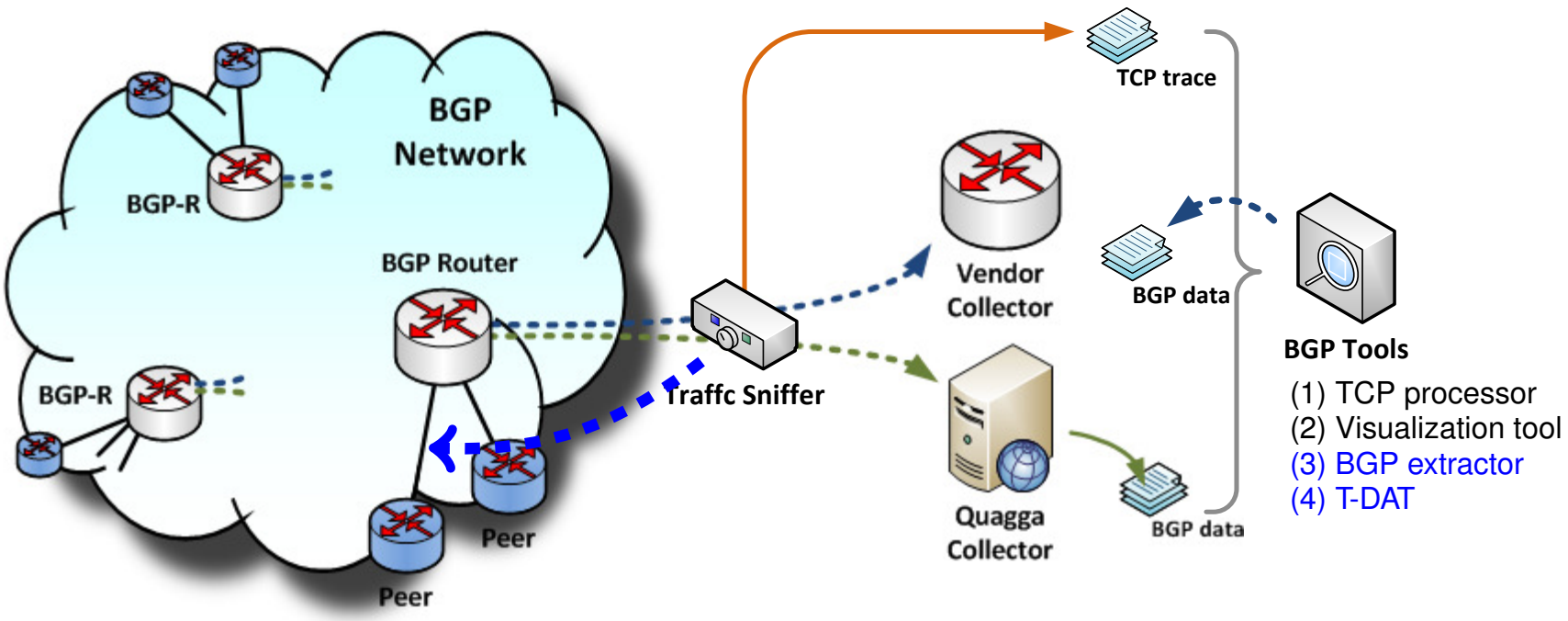
Potential Usage

- Augment the existing BGP monitoring settings
  - Enable TCP investigation
  - Reduce transport-induced artifact
- Allow potential in-band monitoring



Potential Usage

- Augment the existing BGP monitoring settings
  - Enable TCP investigation
  - Reduce transport-induced artifact
- Allow potential in-band monitoring



On-going Work

- Improve the tool
  - Parameter / threshold settings
  - Support other TCP variations?
- Collect BGP/TCP traces
  - ISP_A
  - RouteViews
  - UCLA
- Explore the tool usage for other TCP applications

Thanks!