

Measurement and Classification of Out of Sequence Packets in a Tier-1 Backbone

Jim Kurose

U. Massachusetts, Amherst

Joint work with:

G. Iannaccone, C. Diot (Sprint ATL)

S. Jaiswal, D. Towsley (UMass)

Outline

- Introduction: network measurement
- measurement-in-the-middle: methodology
 - ❖ out-of-sequence classification
 - ❖ RTT estimation
- measurement results: out-of-sequence study
- future work

Motivation

- “health” of TCP connection indicated by amount of out-of-sequence packets indicates
- understand **extent** and **cause** of out-of-sequence phenomenon:
 - ❖ retransmissions
 - ❖ duplicates
 - ❖ re-orderings
- develop new measurement methodology: “measurement-in-the-middle”

Measurement methodologies

taxonomy of approaches:

- **active** versus **passive**?
 - ❖ active: inject traffic, measure
 - ❖ passive: observe, measure existing traffic
- **where** measurements taken:
 - ❖ network edge (host, servers)
 - ❖ routers ("within" network)
- **what** metrics?
 - ❖ delay, loss, rate, traffic type (http, p2p, udp, ...)
 - ❖ per-hop (local) or per-path (end-to-end)

Measurement methodologies

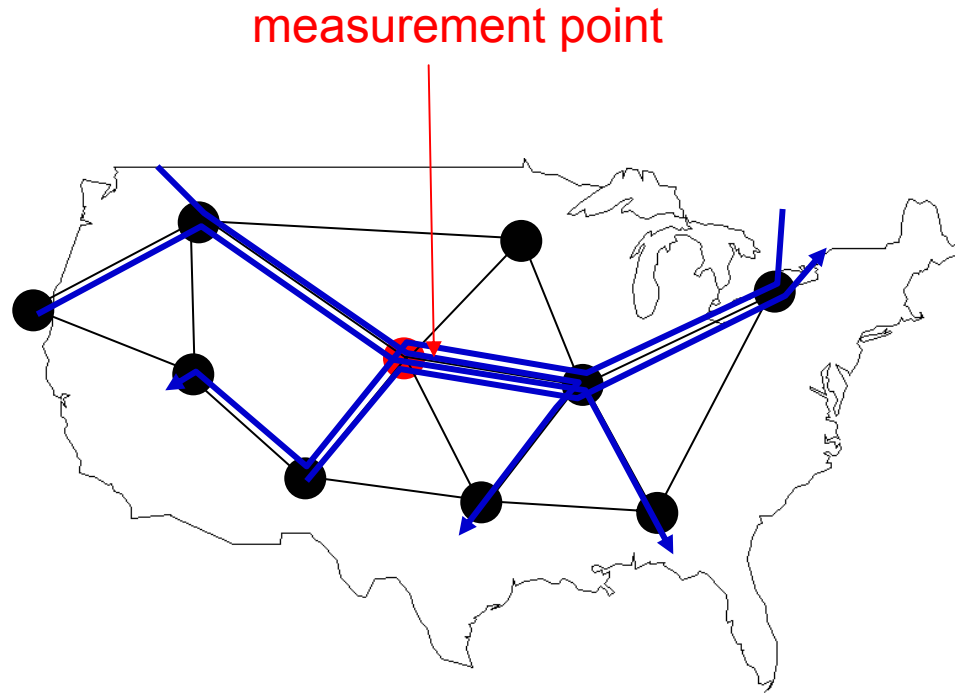
| | local, per-node performance | end-end path performance |
|--|--|---|
| at network edge (active) | <i>ping</i> <i>traceroute</i> <i>pathchar</i> (delay, loss, link bw) | <i>ping</i> <i>traceroute</i> pathchar (delay, loss, link bw) |
| in “middle” of network (passive) | <i>SNMP</i> <i>Netflow</i> traffic/flow types, rates, loss | <i>measurement-in- the-middle</i> per-flow loss, delays |

Outline

- Introduction: network measurement
- **measurement-in-the-middle**: methodology
 - ❖ out-of-sequence classification
 - ❖ RTT estimation
- measurement results: out-of-sequence study
- future work

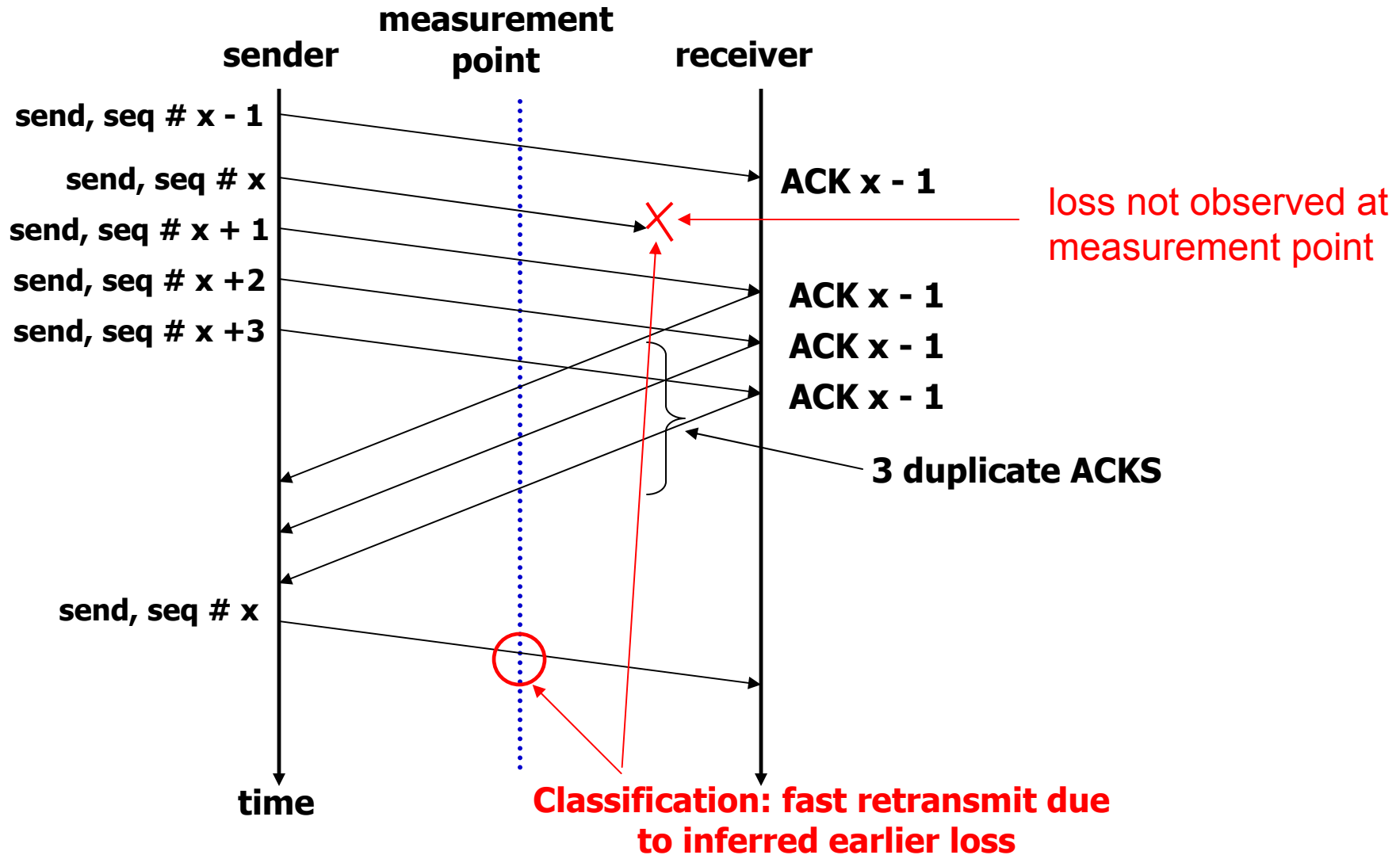
Measurement in the middle

- observe, measure flows passing through measurement point
- backbone measurement point allows *many* flows to be measured (19 hrs):
 - ❖ 18M TCP flows
 - ❖ 4300 ASes (33% of Internet total)



- *fundamental challenge*: limited observability
 - ❖ *can not observe* what happens upstream, downstream
 - ❖ but we *can infer* what happens!

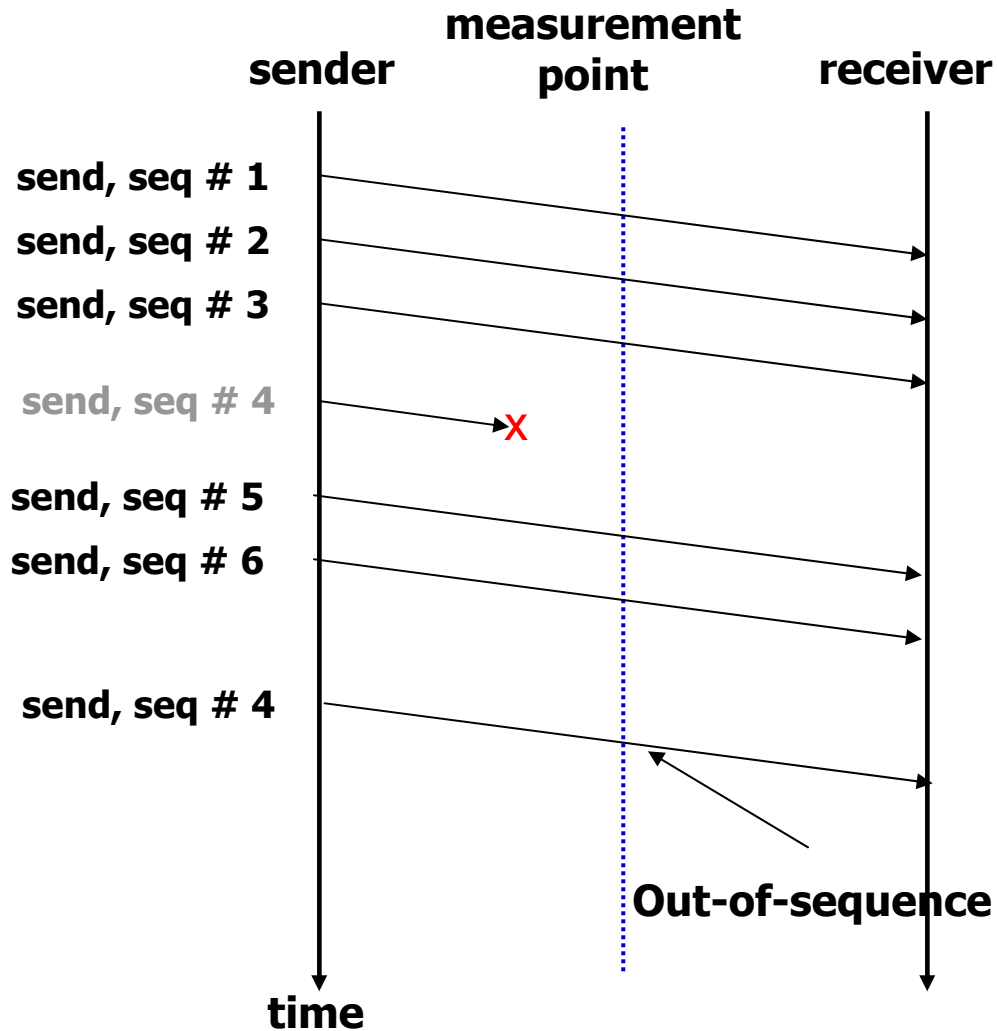
Measurement-in-the-middle: TCP example



Key observation:

- ❑ **exploit measurements, protocol knowledge** (and perhaps probabilistic model of system behavior) **to infer outcomes of unobserved events**
- ❑ ***TCP out-of-sequence behavior***: construct classification heuristics using:
 - ❖ estimates of connection characteristics
 - RTT estimate to compute RTO timer
 - ❖ information in packet headers
 - IP ID field, seq #, addressing info
 - ❖ knowledge of TCP behavior
 - e.g. 3 duplicate ACKS -> fast retransmit

Out-of-sequence packets



Out-of-sequence packet:
packet with sequence #
smaller than that of previously
observed packet

Classification:

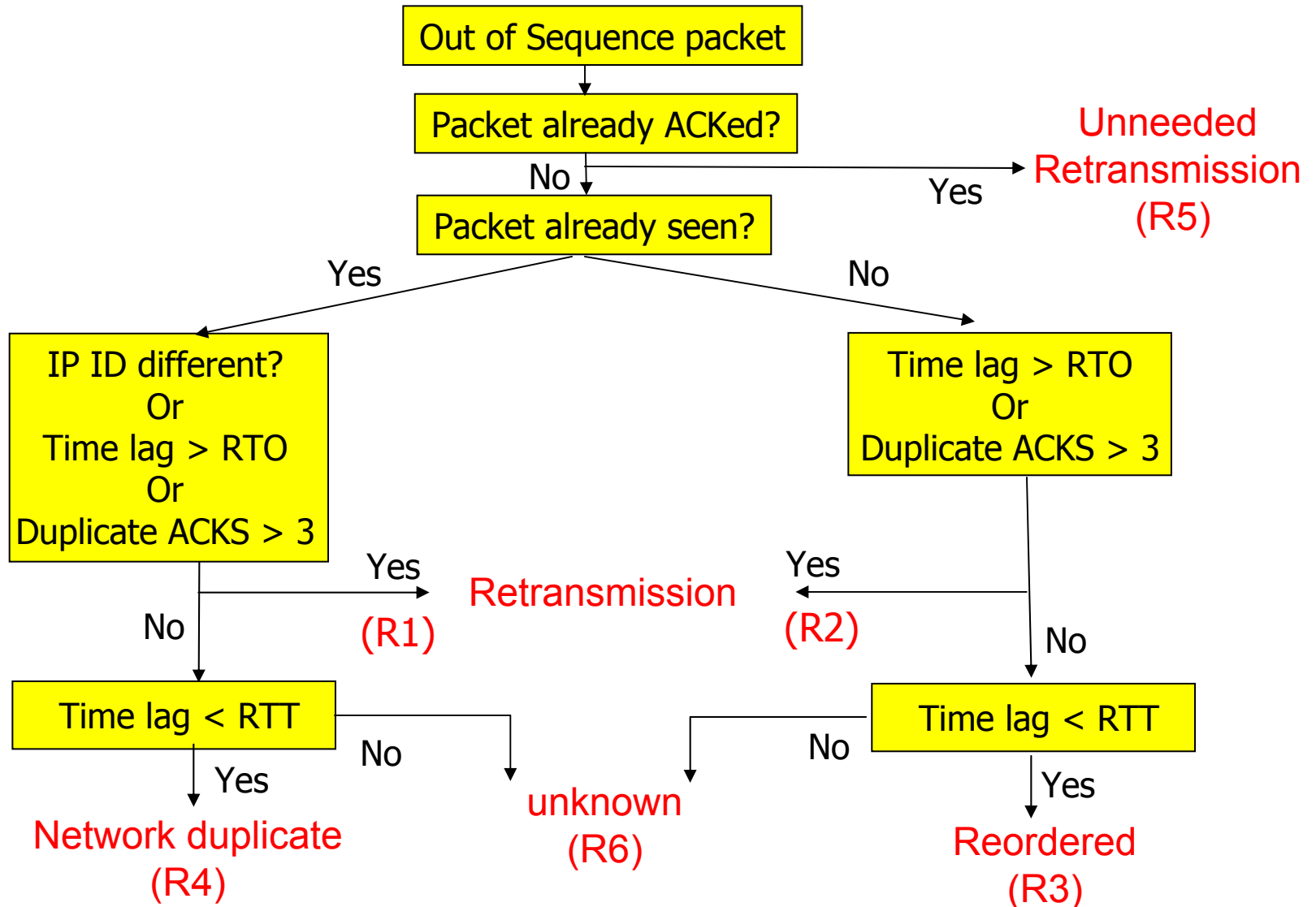
- retransmission
- network duplication
- in-network reordering

TCP/IP Headers

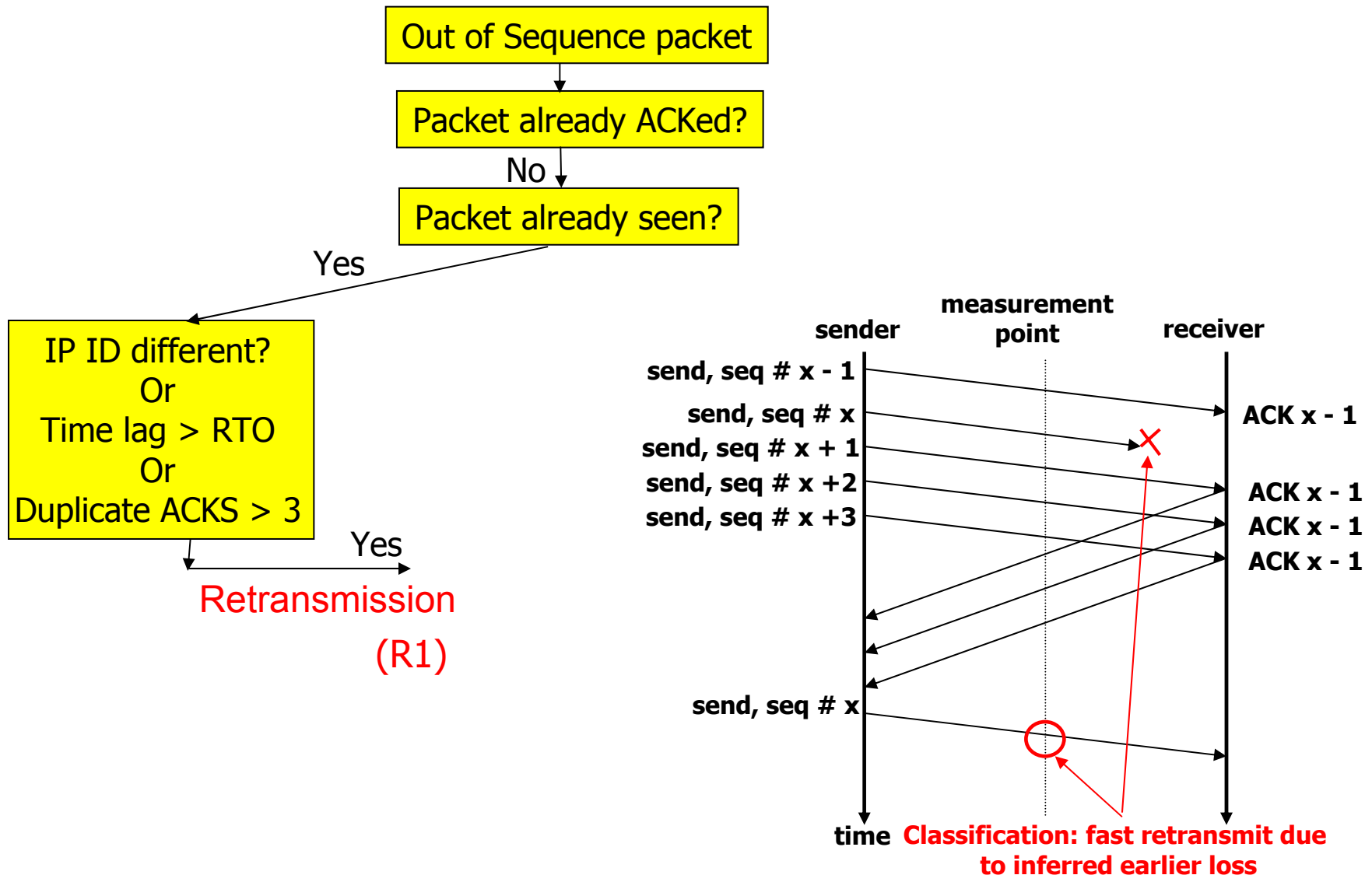
First 44 bytes of TCP packet recorded at measurement point

| | | | | | | | |
|------------|------------------------|----------|-----------|----|-------------------------|-----------------|----|
| IP Header | 0 | | | 15 | 16 | | 31 |
| | Version | HLEN | ToS | | Total Length | | |
| | IP Identification | | | | flags | Fragment Offset | |
| | Time to Live | | Protocol | | Header Checksum | | |
| | Source IP Address | | | | | | |
| | Destination IP Address | | | | | | |
| TCP Header | Source Port Number | | | | Destination Port Number | | |
| | Sequence Number | | | | | | |
| | Acknowledgement Number | | | | | | |
| | Header | Reserved | TCP Flags | | Window Size | | |
| | TCP Checksum | | | | Urgent Pointer | | |

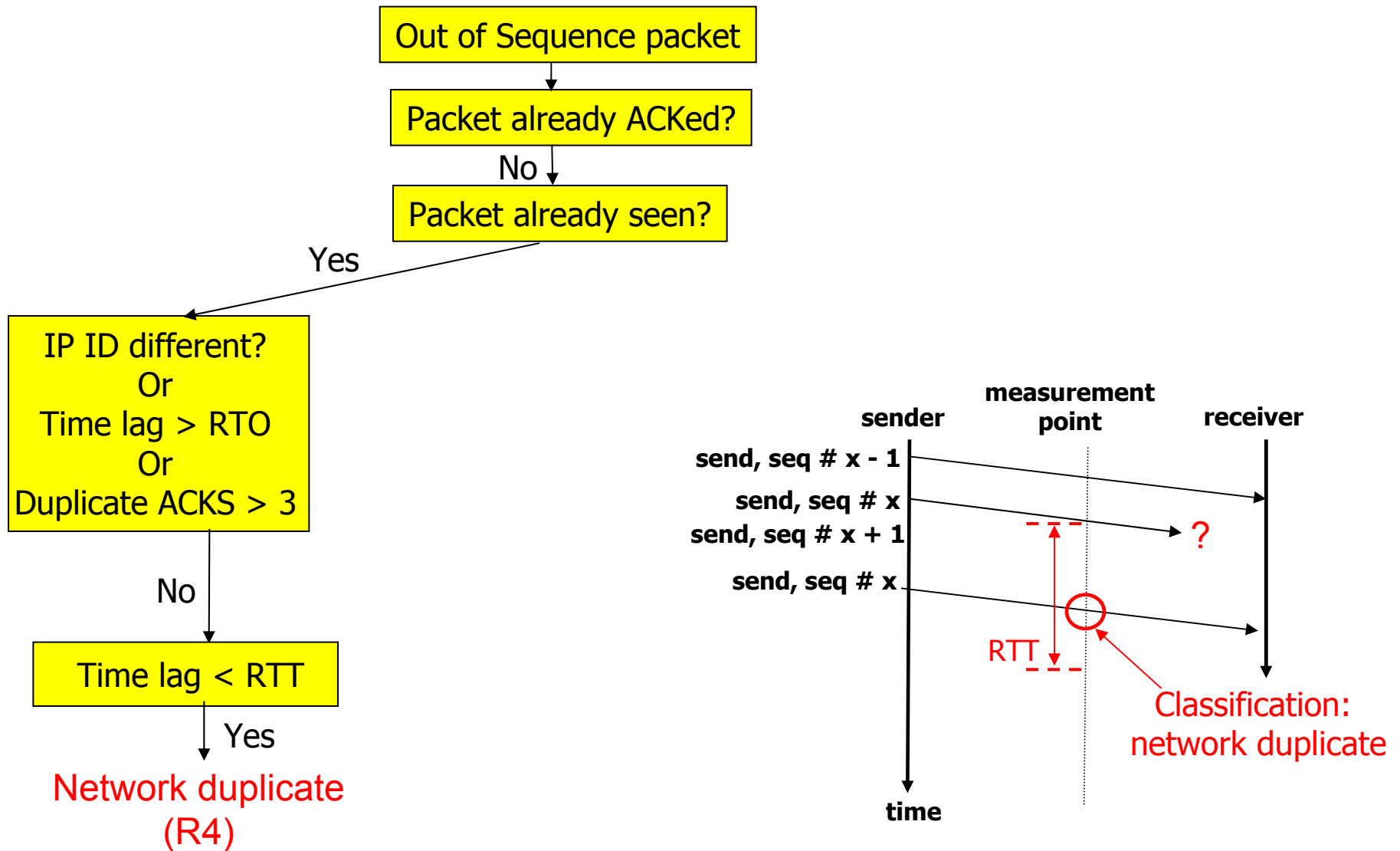
Six Classification Rules



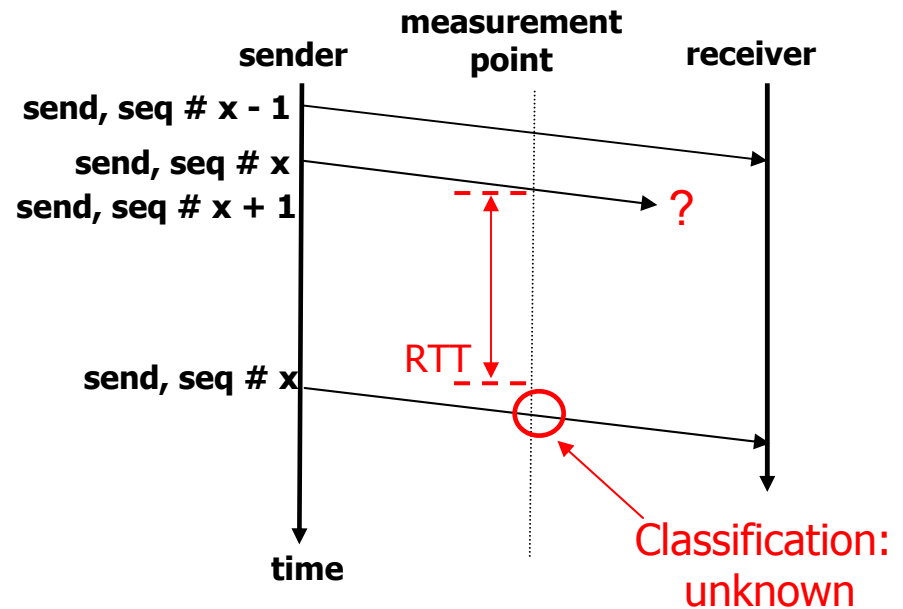
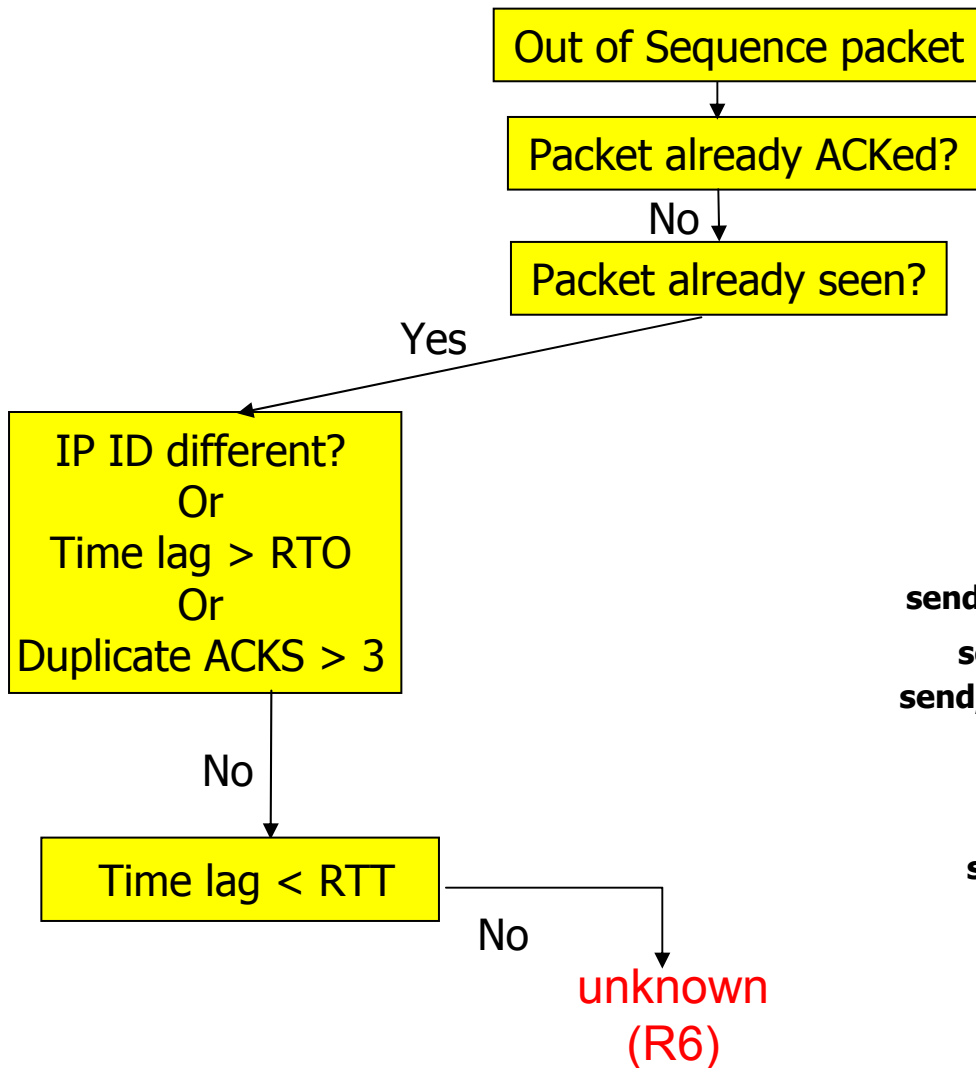
R1: Retransmission (of previously observed packet)



R4: Network duplicate (of previously observed packet)



R6: unknown (of previously observed packet)



R3: network reordering

Out of Sequence packet

Packet already ACKed?

No

Packet already seen?

No

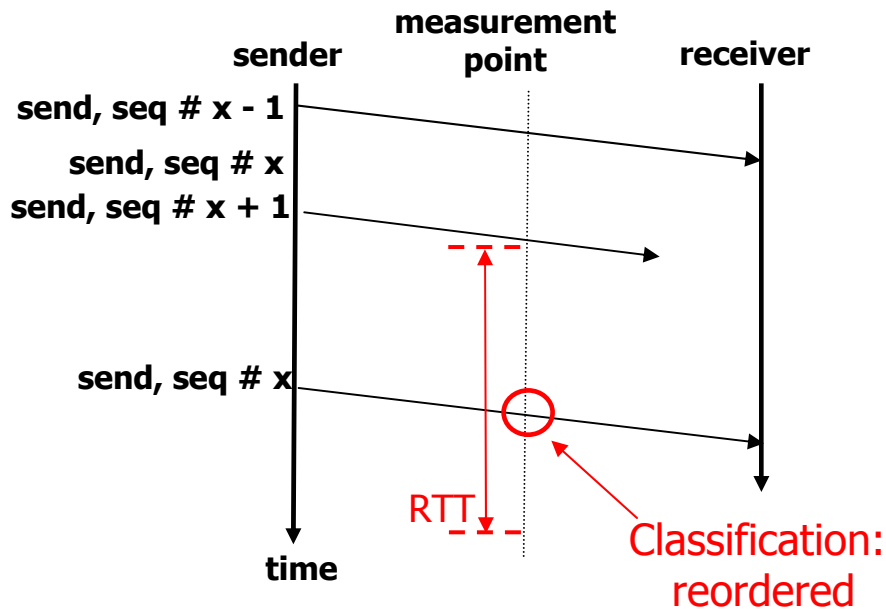
Time lag > RTO
Or
Duplicate ACKS > 3

No

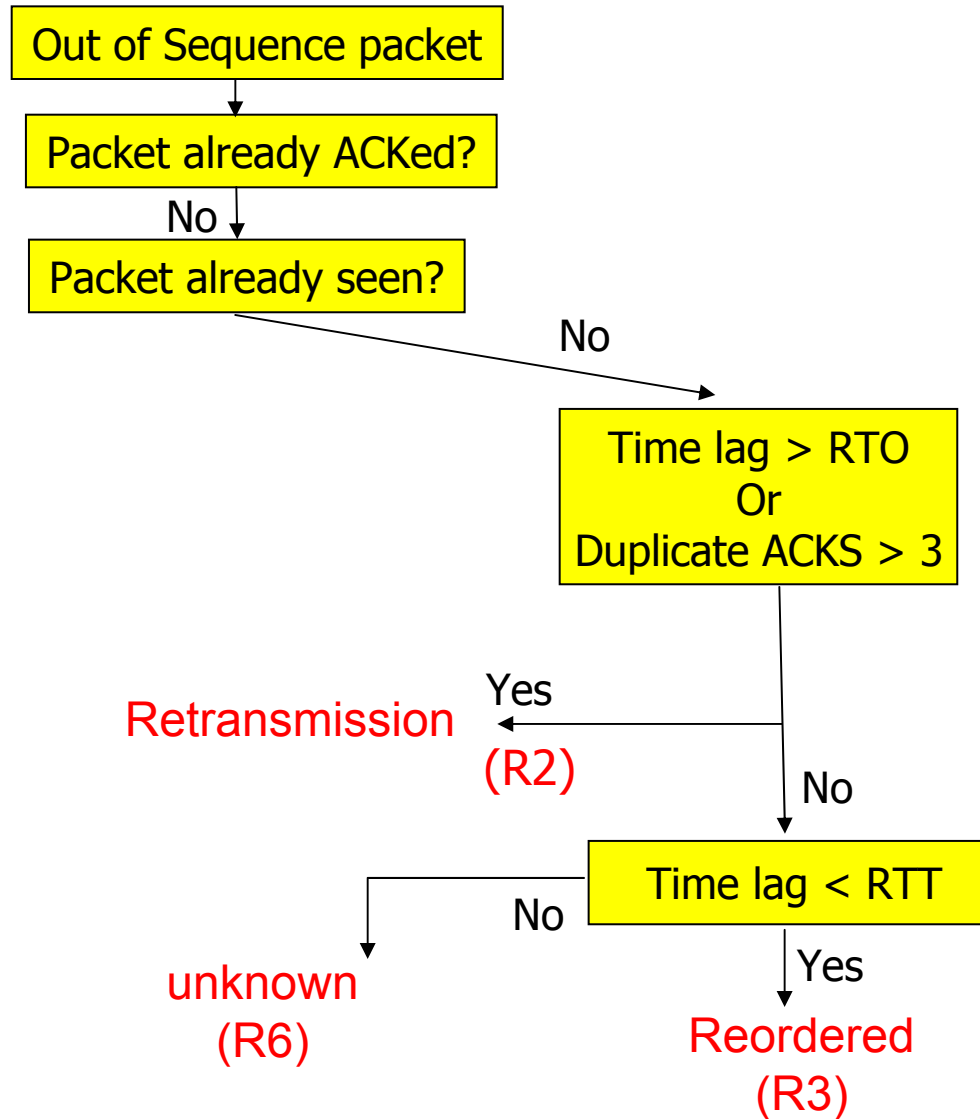
Time lag < RTT

Yes

Reordered
(R3)



R3, R6: retransmissions, unknown



A closer look at RTT estimation

□ RTT estimation crucial:

- ❖ events less than RTT apart: second event can *not* be caused by sender reaction to first event
- ❖ RTO: retransmission time = $RTT + 4 * RTT_deviation$

□ How to estimate RTT?

- ❖ sender RTT estimation complicated:
 - single packet being timed for RTT at any time
 - retransmissions not timed
- ❖ inexact: measurement point distant from sender

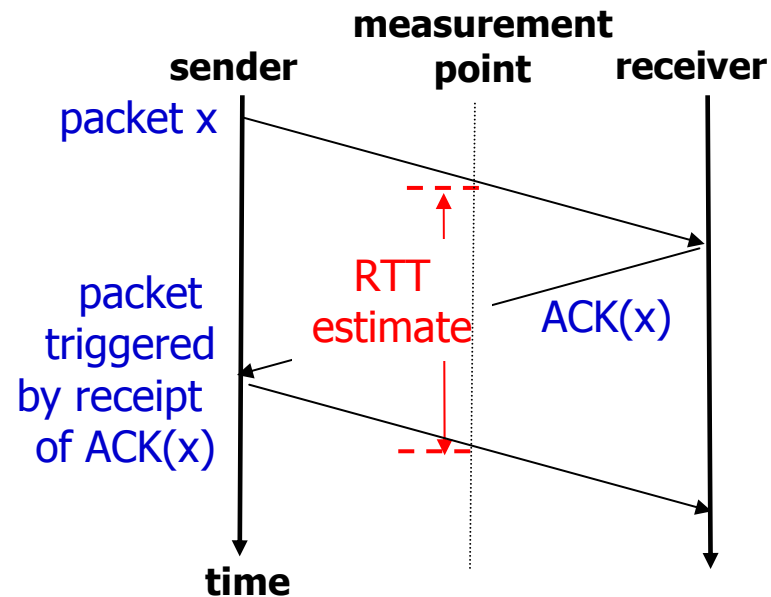
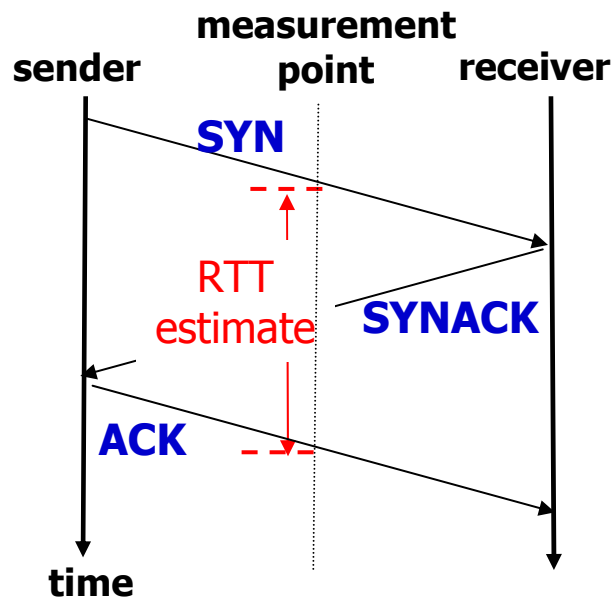
RTT-estimation-in-the-middle

Previous approaches to RTT estimation [4,5]:

- at beginning of session

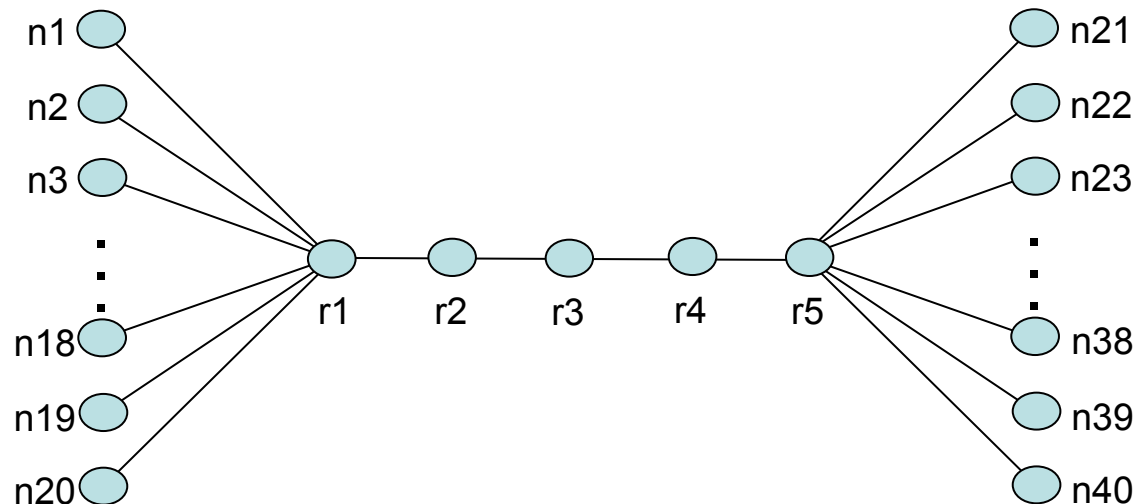
Running RTT estimation:

- update once per RTT

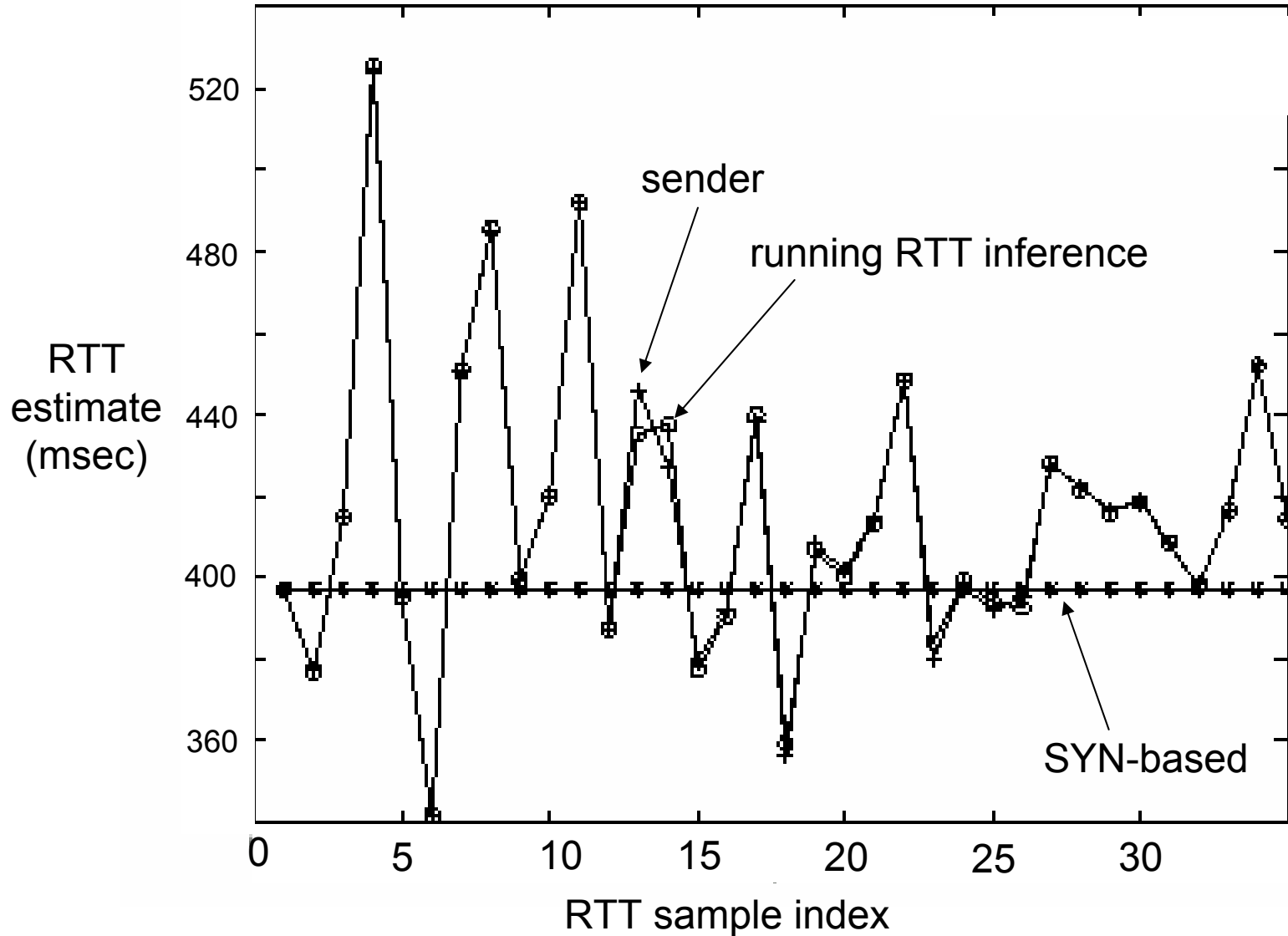


Evaluating RTT-estimation-in-the-middle

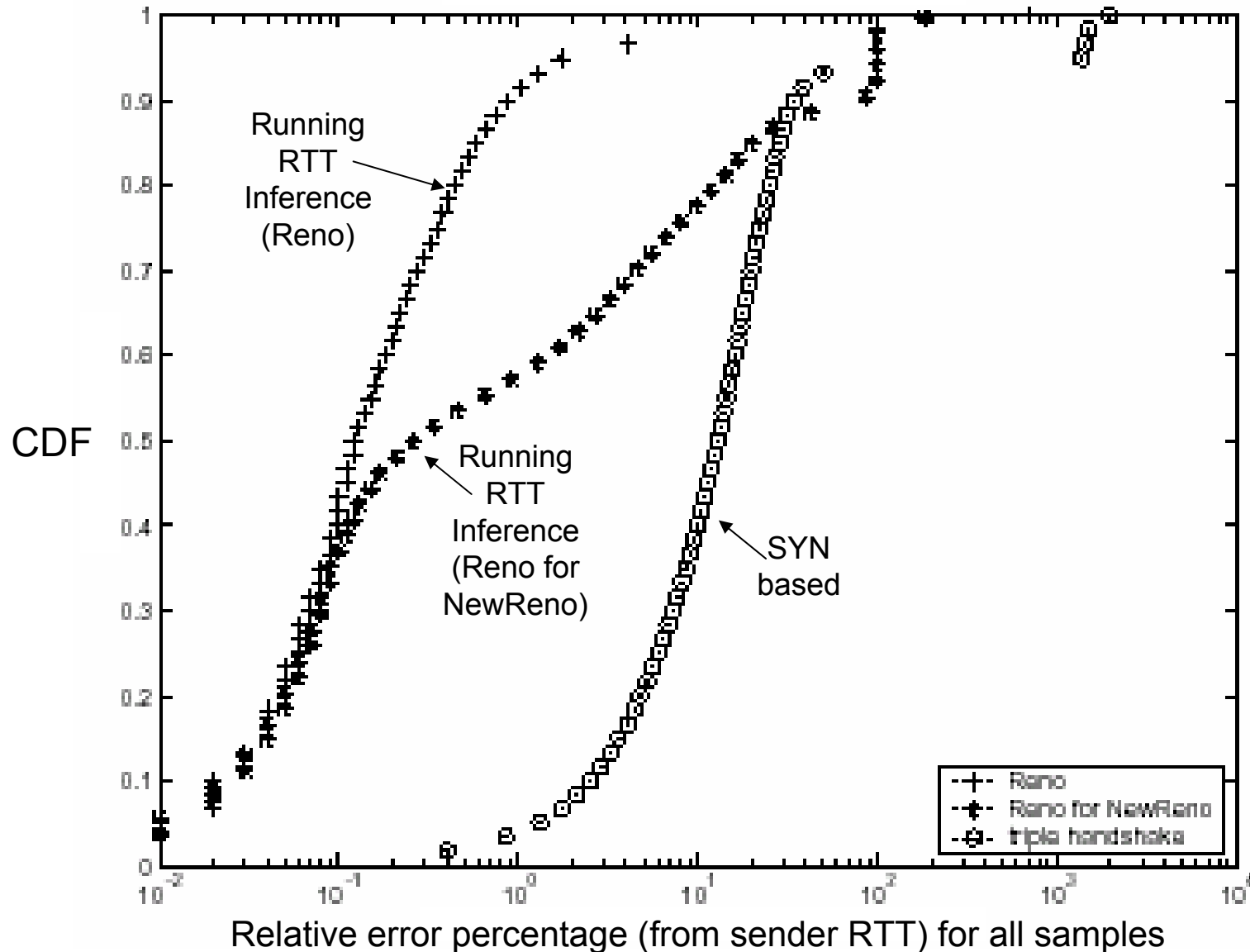
- ❑ ns simulation
- ❑ 3500 short lived (15 pkt/lifetime), 200 long-lived (30 sec.) end-end flows
- ❑ cross traffic at r2, r3, r4, r5
- ❑ r3-to-r4 link is bottleneck, and measurement link
- ❑ compare sender RTT, SYN-based RTT, running RTT



Evaluating RTT-estimation-in-the-middle



Evaluating RTT-estimation-in-the-middle

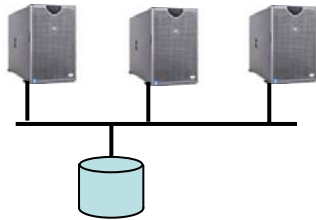


Outline: where are we?

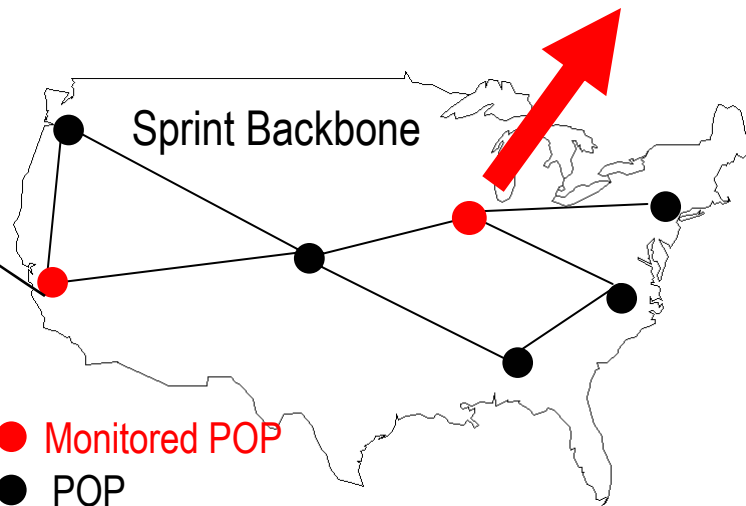
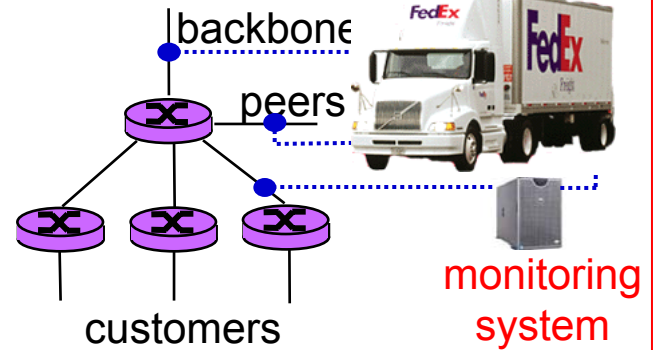
- introduction
- measurement-in-the-middle: methodology
 - ❖ out-of-sequence classification
 - ❖ RTT estimation
- measurement results: out-of-sequence study
- future work

IP Monitoring Infrastructure

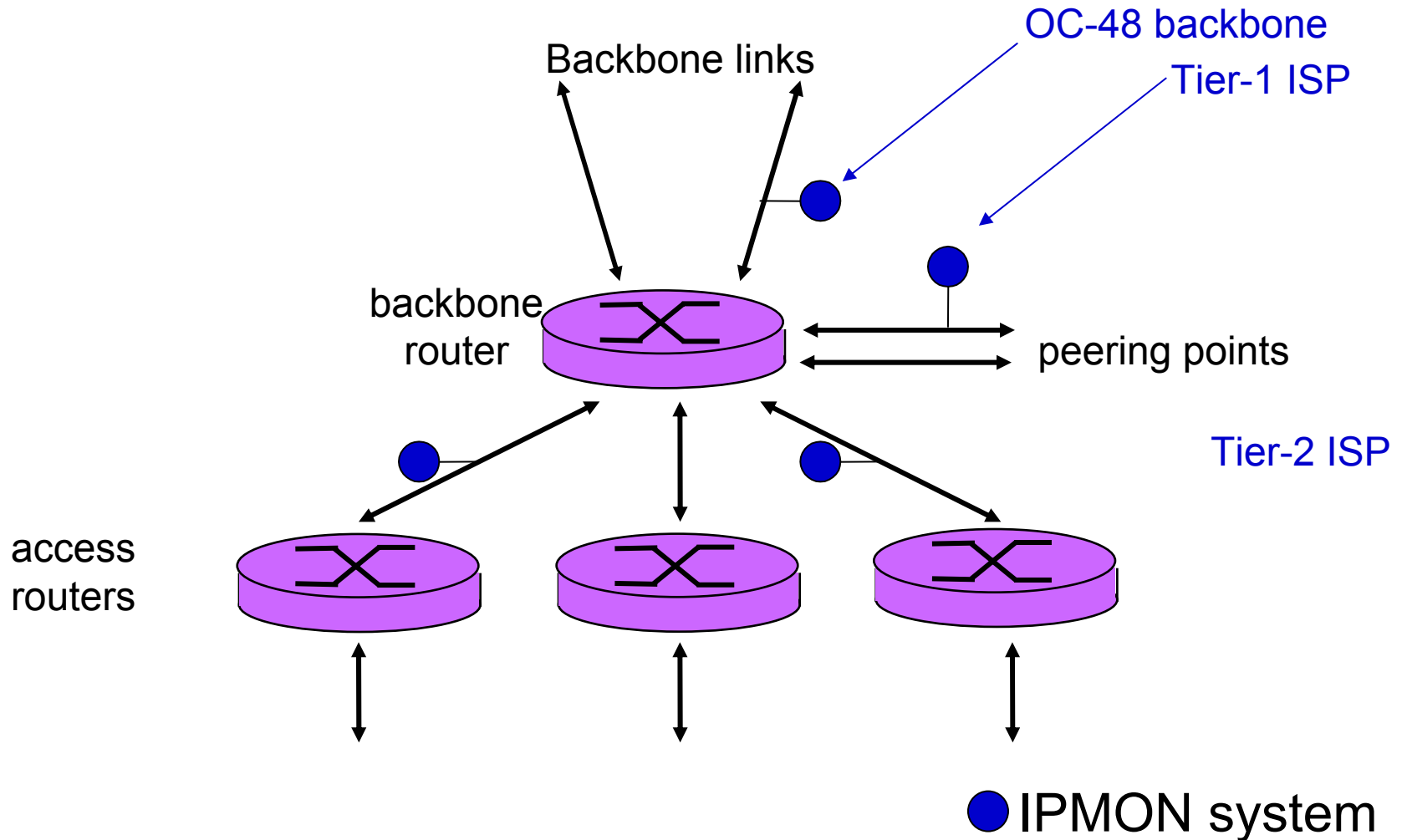
Data Repository and Monitoring Platform (Sprint ATL)



Configuration at Monitored POP



4 types of POP measurement points



4 types of measurement points

| | CDN | Tier-1 ISP | Tier 2 ISP | Backbone |
|-------------------|----------|------------|------------|----------|
| Link speed | 622 Mbps | 622 Mbps | 622 Mbps | 2.5 Gbps |
| Duration (hrs) | 6 | 6 | 6 | 1 |
| Unique source ASs | 1587 | 408 | 1196 | 2532 |
| # TCP conn. | 4.8M | 2.1M | 4.7M | 6.6M |
| % all TCP conn. | 99.46% | 15.68% | 12.79% | 27.21% |
| # TCP data pkts | 91M | 39M | 245M | 153M |

□ trace dates: Feb. 2002, Oct. 2002

Results for all 4 measurement points

| | CDN | Tier-1 ISP | Tier-2 ISP | Backbone |
|-------------------|-------|------------|------------|----------|
| # packets | 90.9M | 30.4M | 245.5M | 153.1M |
| out-of-seq. | 1.6% | 4.7% | 5.7% | 5.1% |
| retrans | 86.8% | 62.4% | 70.4% | 72.3% |
| unnneeded retrans | 9.4% | 12.3% | 19.5% | 14.2% |
| network dups | 0.01% | 0.04% | 0.05% | 0.03% |
| network reorder | 1.9% | 16.7% | 6.67% | 8.4% |
| unknown | 1.79% | 8.55% | 3.3% | 5.1% |

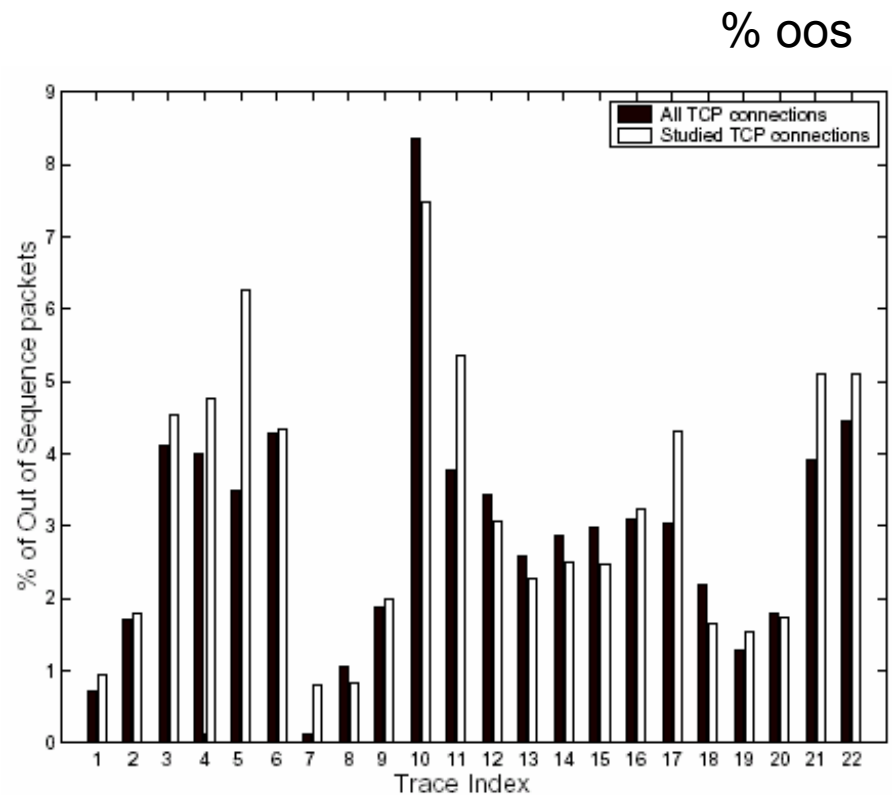
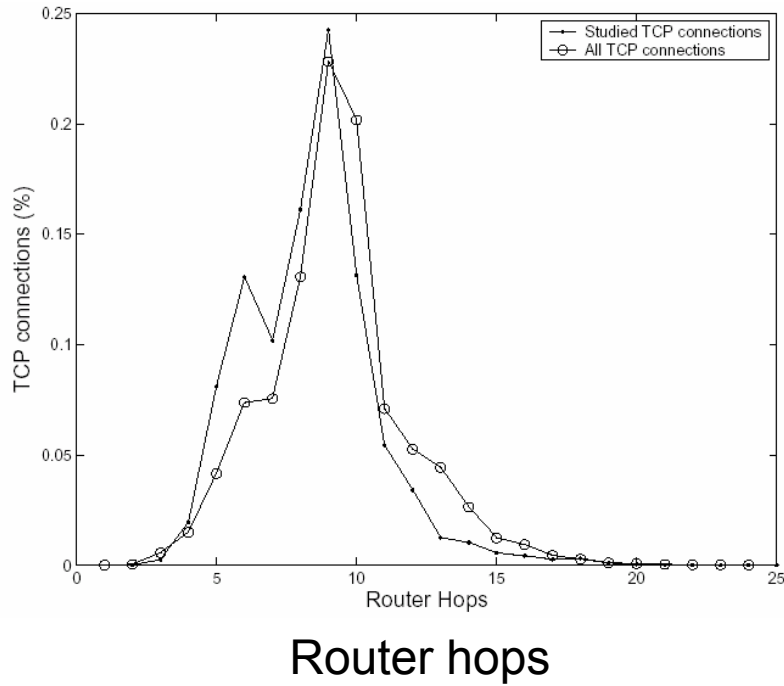
Packet reordering: observations

- little reordering
 - ❖ 0.03 - 0.7% of all data pkts
 - ❖ connections with at least 1 reordering: < 5%
 - ❖ 87% of lost pkts have "lag" < 3 pkts
- comparison with previous end-end studies (Bennett99, Paxson97)
 - ❖ previously reported numbers higher
 - ~ 0.5 - 5% pkts reordered
 - connections with at least one reordering
 - 90% (Bennett),
 - 36%, 12% (Paxson)
 - ❖ several methodological differences (back-back ICMP) also

Symmetric path at measurement point?

- requirement:
 - ❖ **symmetry**: data, ACK packets both pass through measurement point
 - ❖ **filtering**: only consider such symmetric traces from population of traces
- question: does filtering introduce bias?
 - ❖ filtered traces show same hopcount distribution as population
 - ❖ can compute out-of-sequence fraction for *all* traces
 - ❖ out-of-sequence fraction from 22 traces:
 - 10 traces < 10% relative error
 - 15 traces < 20% relative error

Symmetric path at measurement point?

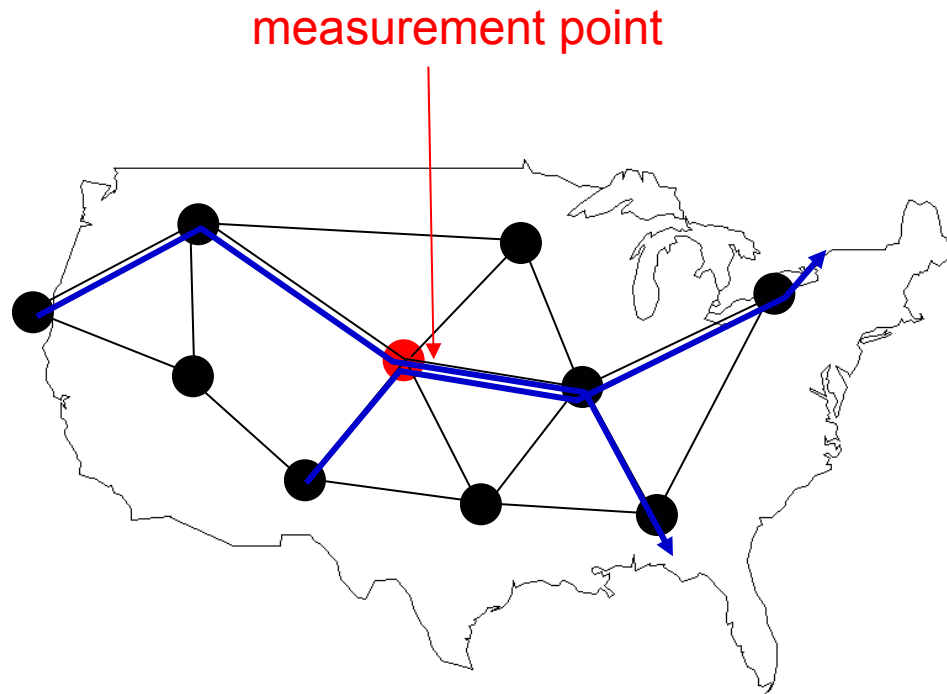


Outline: where are we?

- introduction
- measurement-in-the-middle: methodology
 - ❖ out-of-sequence classification
 - ❖ RTT estimation
- measurement results: out-of-sequence study
- future work

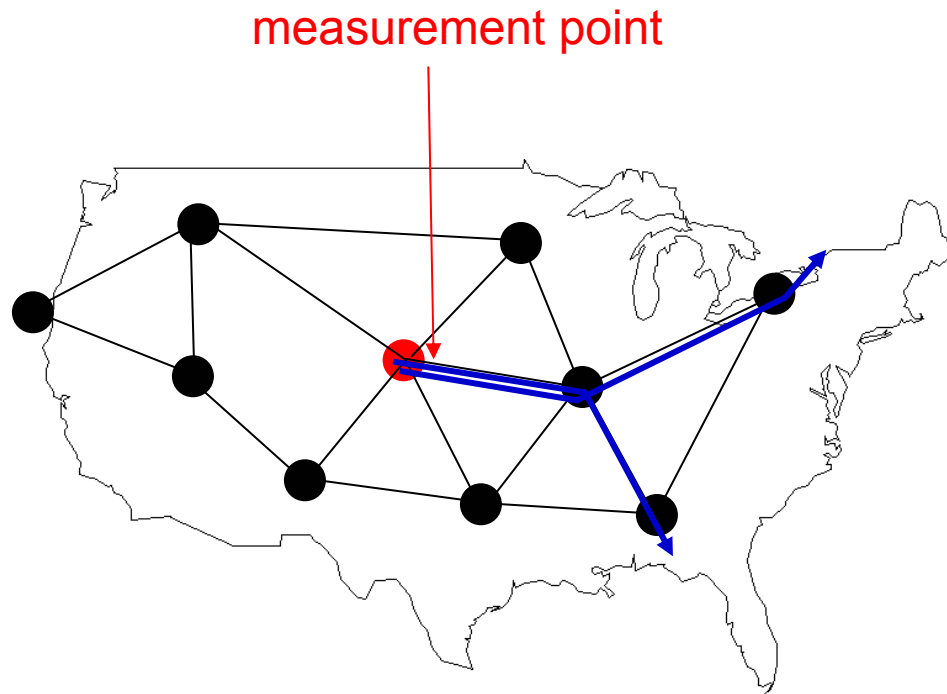
Future: where does loss occur?

- measurement-in-the-middle can be used to infer *where* in network loss occurs!



Future: where does loss occur?

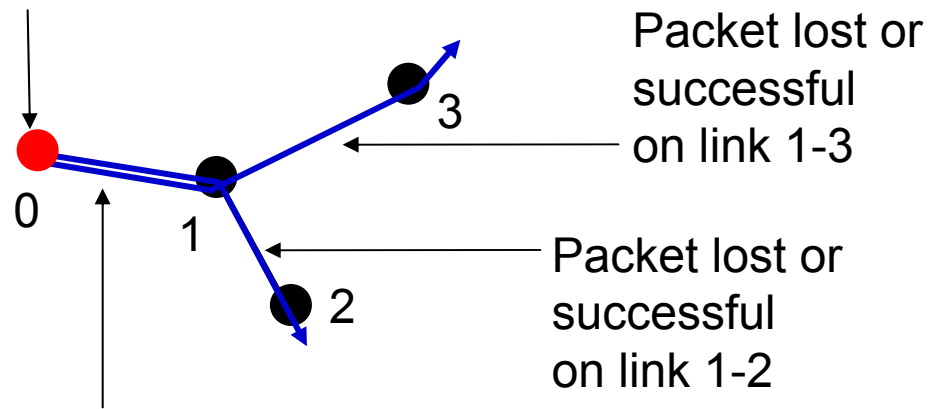
- focus on packets traveling downstream from measurement point



Future : where does loss occur?

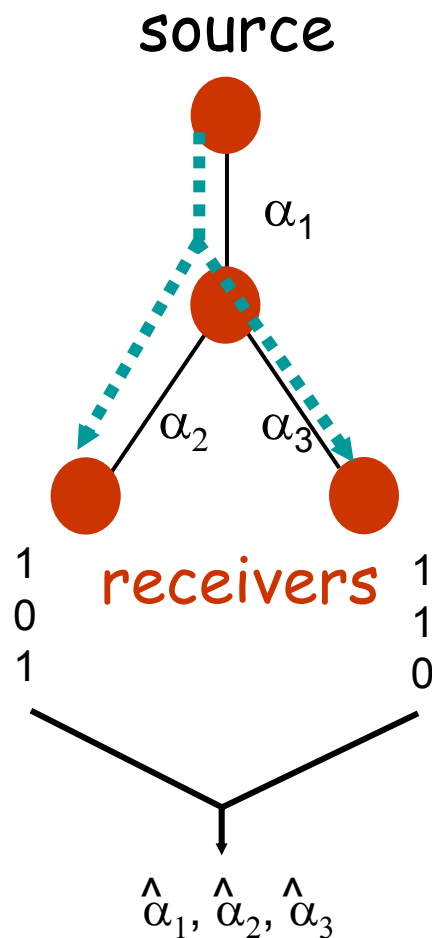
- packets from measurement point to destination viewed as “probes”

Two packets arrive back-to-back at 0, one destined for 2, other destined for 3



Two packets experience same fate on link 0-1

Future : where does loss occur?

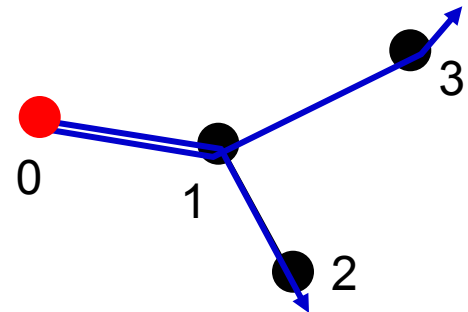


Exploiting correlation:

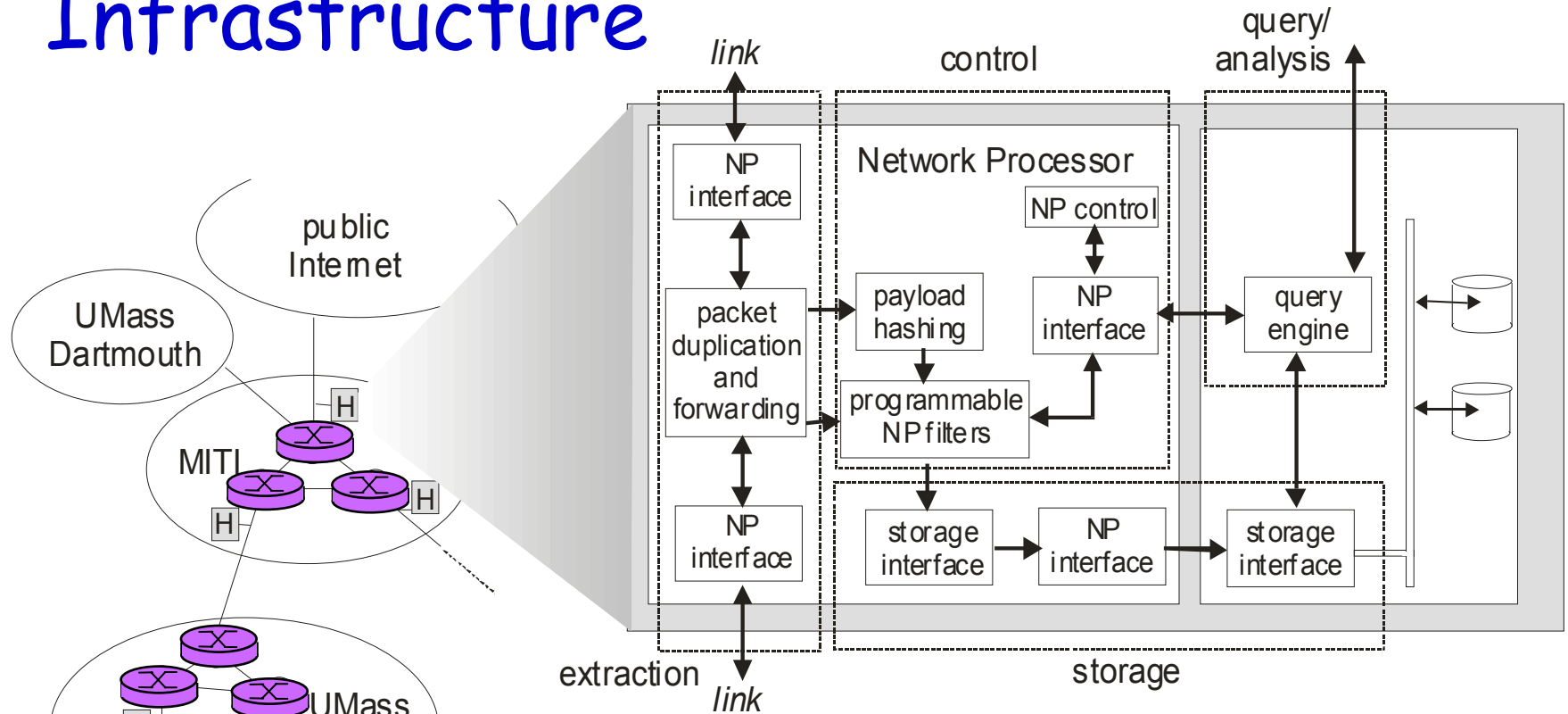
- ❑ Loss model: Bernoulli, α_k on link k
- ❑ N packet "trains": X_n per recv record of loss/success
- ❑ *Goal:* estimate link probabilities $\alpha = \{\alpha_k\}$ from X
- ❑ *Inference:*
 - ❖ given $\{\alpha_k\}$ can compute $\text{Prob}(X_n; \alpha)$.
 - ❖ Find $\{\alpha_k\}$ most likely to have produced observations $\{X_n\}$

Future Work: Inference

- how closely spaced must packets be?
- what is effective probe rate in practice?
- effects of lost ACKs?
- inference of *upstream* loss



Future Work: Measurement Infrastructure



- use of network processors
- indexing on the fly
- query engine
- multiple measurement points

Summary & Future Directions

- identification & classification of out of sequence phenomenon
- **measurement-in-the-middle**: methodology, architecture
- starting point to examine additional issues:
 - ❖ Identifying **TCP flavors**
 - ❖ **Where** in network do these occur?
(e.g. access points, peering points)
 - ❖ **Why** do these phenomenon occur?
(e.g., congestion, link failures, route updates)

Thanks!

“Measurement and Classification of Out of Sequence Packets
in a Tier-1 Backbone,”

S. Jaiswal, G. Iannaccone, C. Diot, J. Kurose, D. Towsley
to appear in *IEEE Infocom 2003*

<http://gaia.cs.umass.edu>

Symmetric path at measurement point?

- requirement:
 - ❖ **symmetry**: data, ACK packets both pass through measurement point
 - ❖ **filtering**: only consider such symmetric traces from population of traces
- question: does filtering introduce bias?
 - ❖ filtered traces show same hopcount distribution as population
 - ❖ can compute out-of-sequence fraction for *all* traces
 - ❖ out-of-sequence fraction from 22 traces:
 - 10 traces < 10% relative error
 - 15 traces < 20% relative error

Symmetric path at measurement point?

