LOOPS

Localized Optimizations over Path Segments

IETF 104 side meeting, 2019-03-27

Agenda

- 13:45 intro, chair’s overview (I’m playing chair :-)
- 13:50 LOOPS technical presentation (Yizhou)
- 14:00 overview over encapsulation opportunities (Tom)
- 14:10 technical discussion
- 14:20 charter proposal presentation
- 14:25 discussion, next steps
- 14:45 conclude

Objectives for this meeting

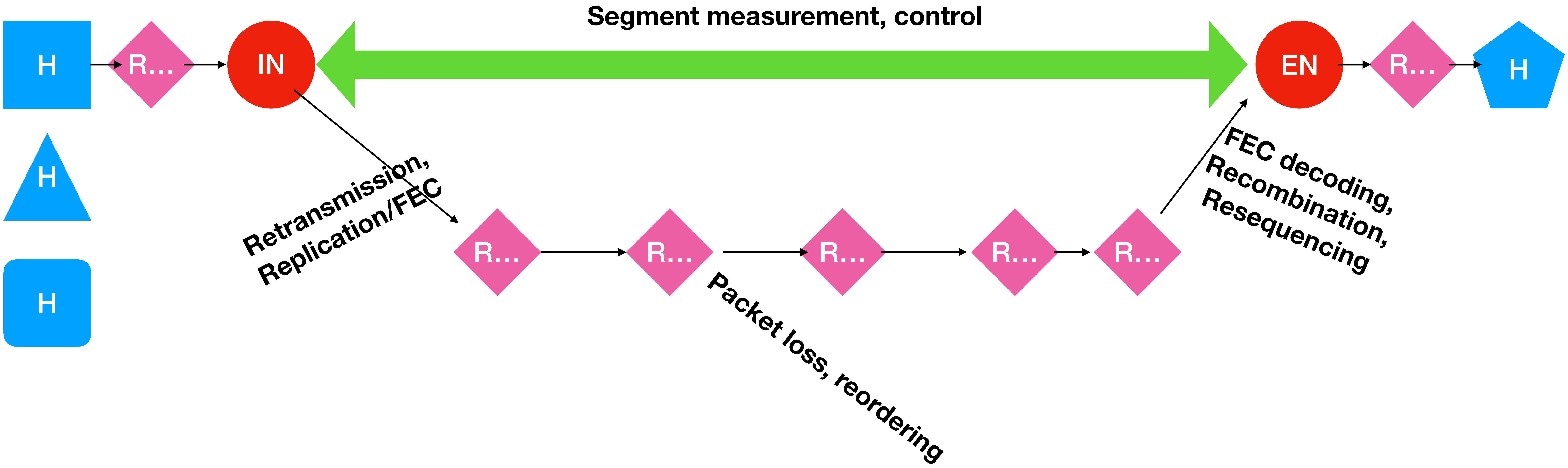
(1) check consensus that this is work worth doing

(2) check consensus that the WG we propose is the right way of starting that work

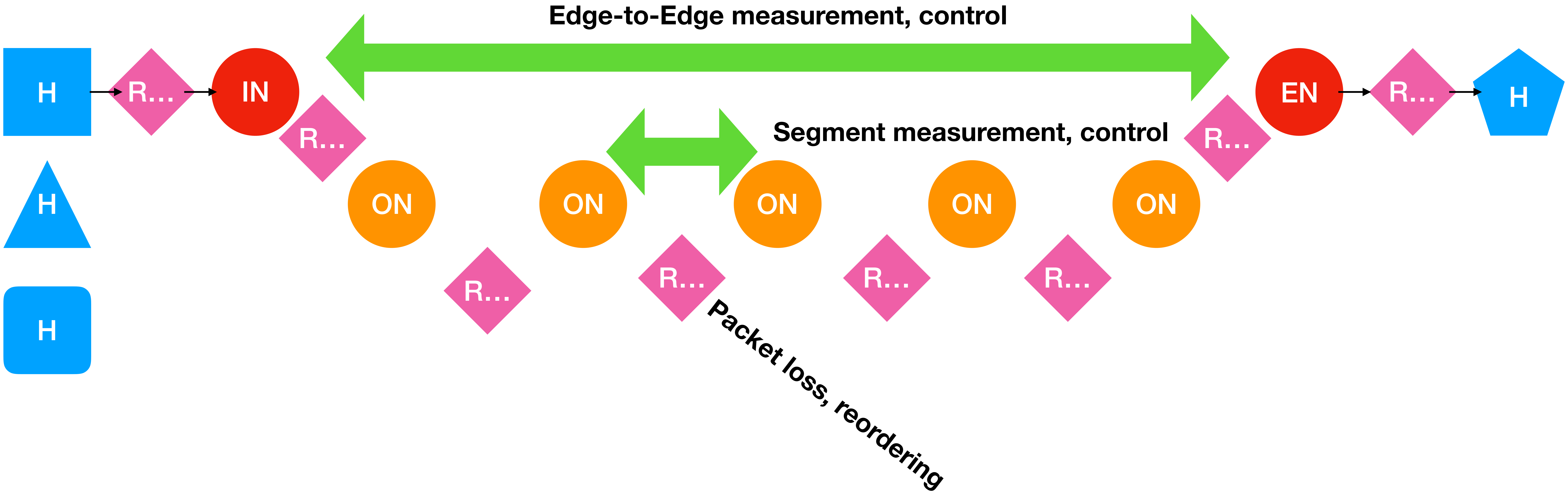
(3) get more feedback on charter proposal

(4) understand the roles of the individuals that will contribute to the effort (time frames: until WG is chartered, for WG work)

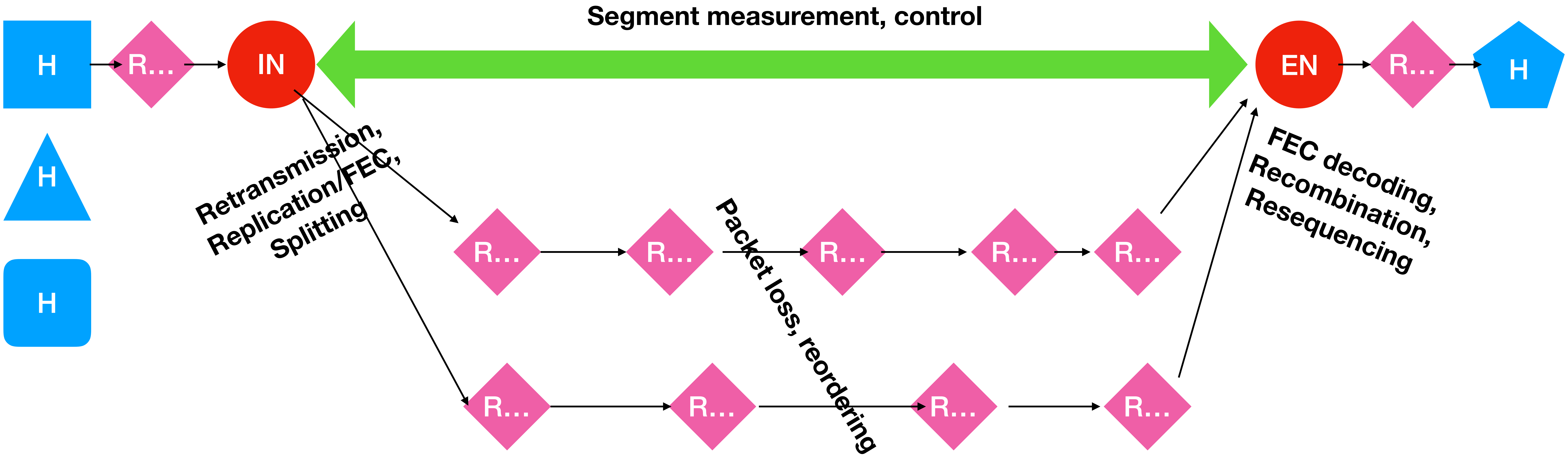
LOOPS Opportunity



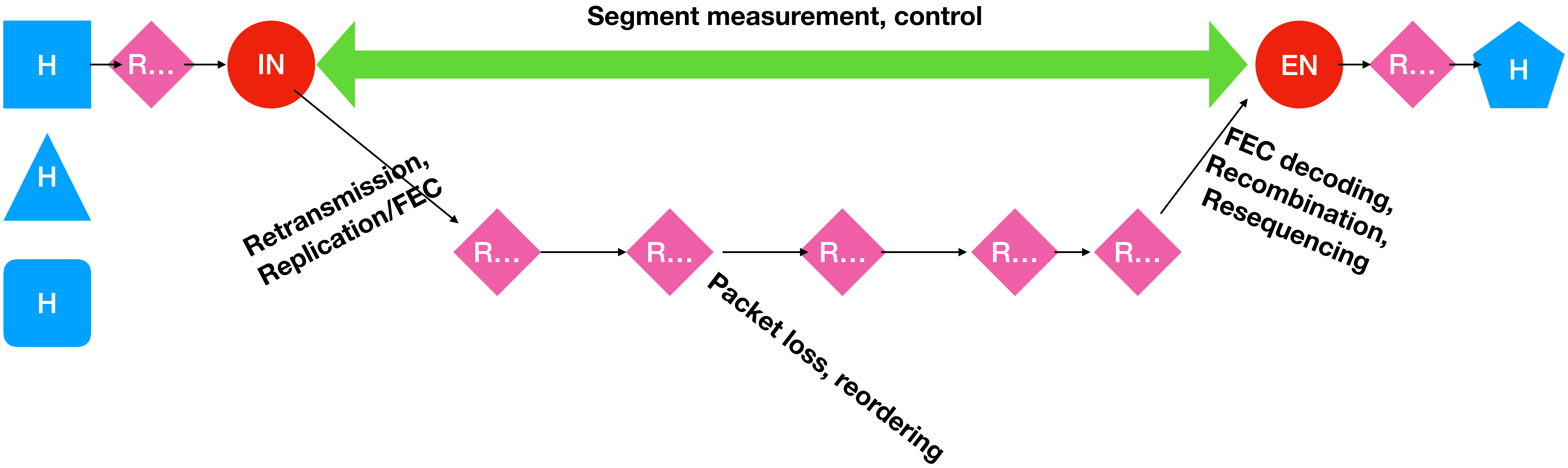
Opportunity (multi-segment)


Edge-to-Edge measurement, control
H
R…
IN
R…
ON
R…
ON
R…
Segment measurement, control
ON
R…
ON
R…
ON
R…
EN
R…
H
H
H
Packet loss, reordering

Opportunity (multi-path)



LOOPS Opportunity: First Step



Elements of LOOPS

- Information model for local recovery: in-network retransmission/FEC
  - Can be encapsulated in a variety of formats; define some of those
- Local measurement: e.g. segment forward delay/variation
  - To set recovery parameters
  - To determine if loss was caused by congestion
- Congestion relay:
  ECN (or drops) to inform end hosts about congestion loss

Freezer (not on agenda right now)

- Multipath
- Measurement across LOOPS pairs (“almost e2e”)
- MTU handling, fragmentation, aggregation, header compression
- Selection of one or more specific tunnel encapsulation or measurement format (beyond “sketches” showing it can be made to work)

Terms

- LOOPS: Local Optimization on Path Segments
- LOOP: Local Optimization Overlay Pair (one ingress, one egress)

Out of scope:

- LONG LOOP: L.O. Node Group (several LOOPs, chained) — out of scope
- InterLOOP probe: probe by one LOOP to find others on the path

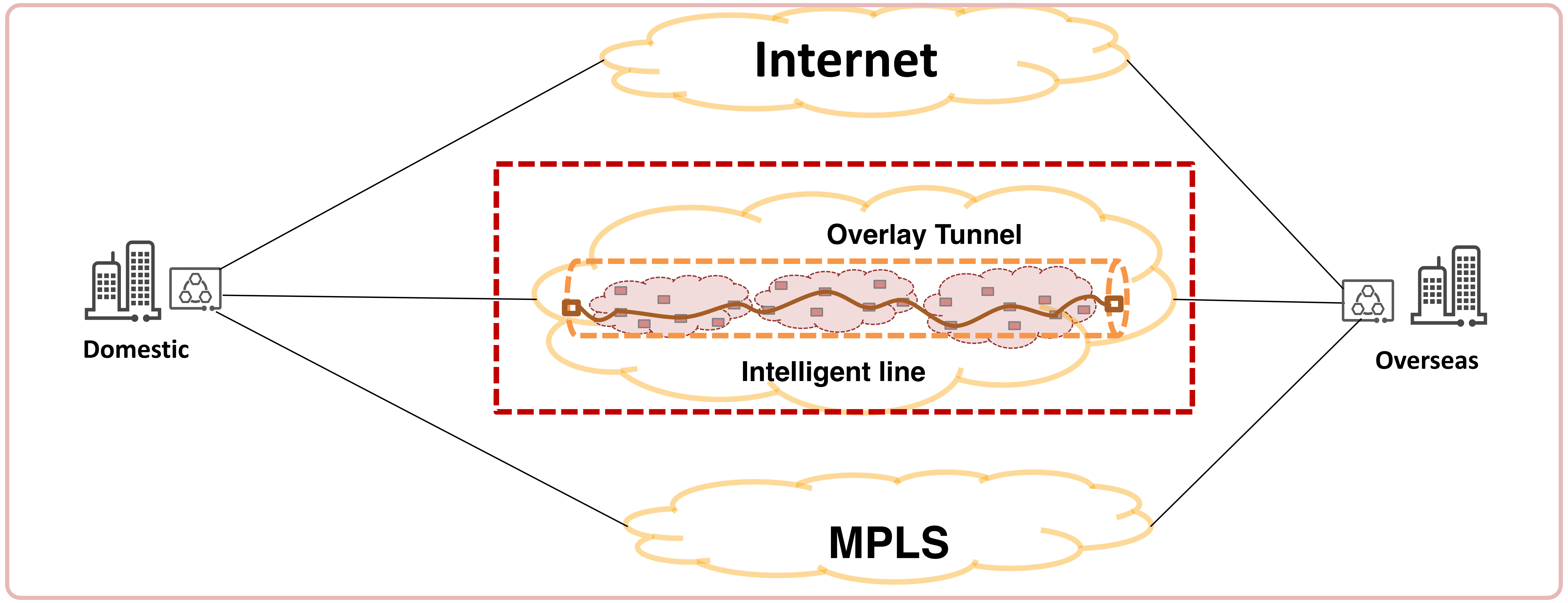
LOOPS Problem & Opportunities

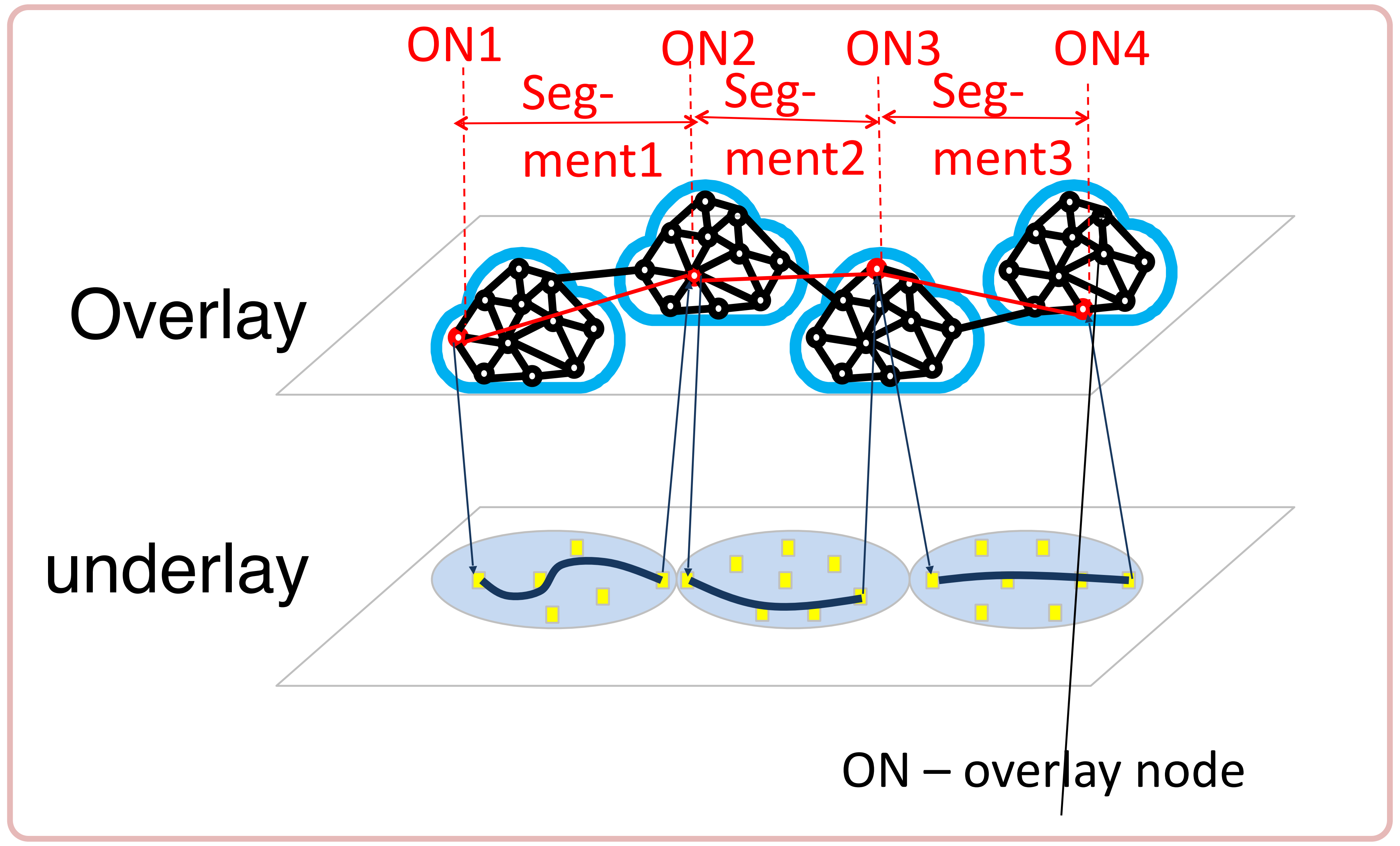
draft-li-tsvwg-loops-problem-opportunities

Yizhou Li

Xingwang Zhou

Usage Scenario & Motivations - 1

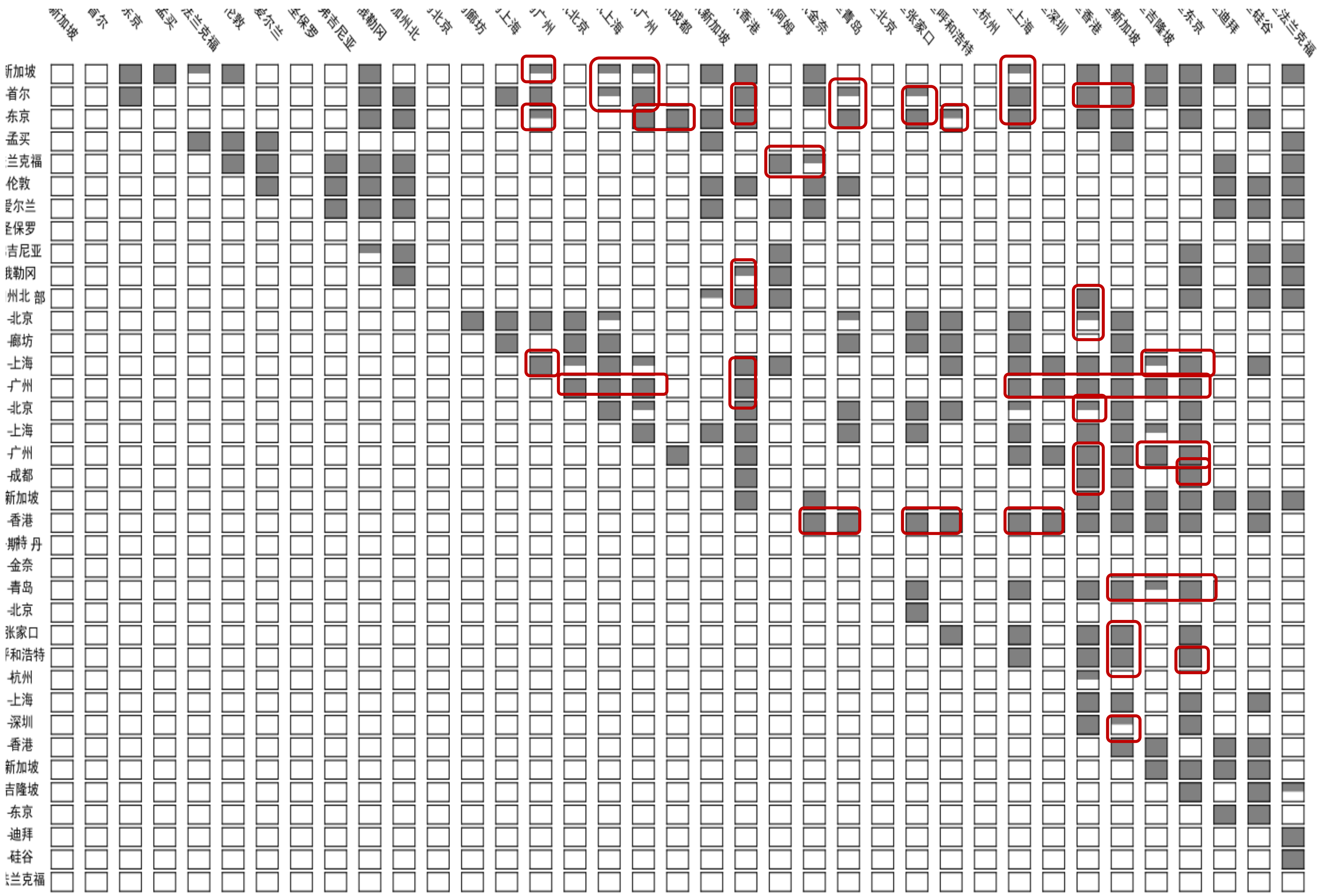




- Default path does not always give the best latency and throughput
- Cloud-Internet Overlay Network (CION): Build a better WAN path via a sequence of overlay nodes in different geographic sites in multiple clouds
- Experiments: 71% chance of finding a better overlay path based on 37 cloud routers globally

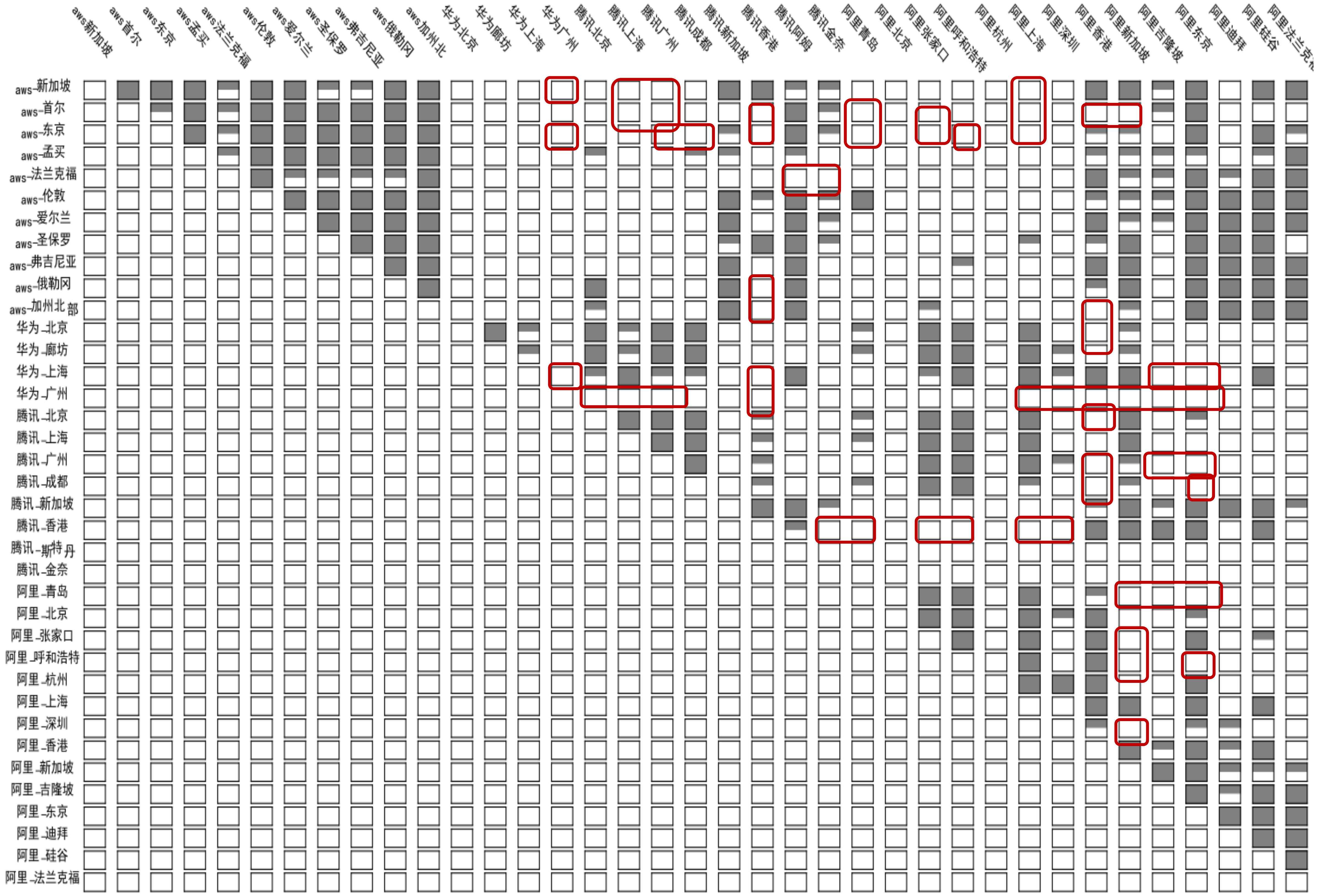
Usage Scenario & Motivations - 2

Delay satisfaction rate between overlay node pairs



- Requirement: delay <200ms (inter-regional), 100ms(regional), 40ms(domestic)
- Satisfaction rate between any pairs
  - 37.21% at 99% percentile
  - 32.81% at 99.9% percentile

Packet loss satisfaction rate between overlay node pairs



- Requirement: packet loss rate: <1% every 20 seconds
- Satisfaction rate between any pairs
  - 44.27% at 99% percentile
  - 29.51% at 99.9% percentile

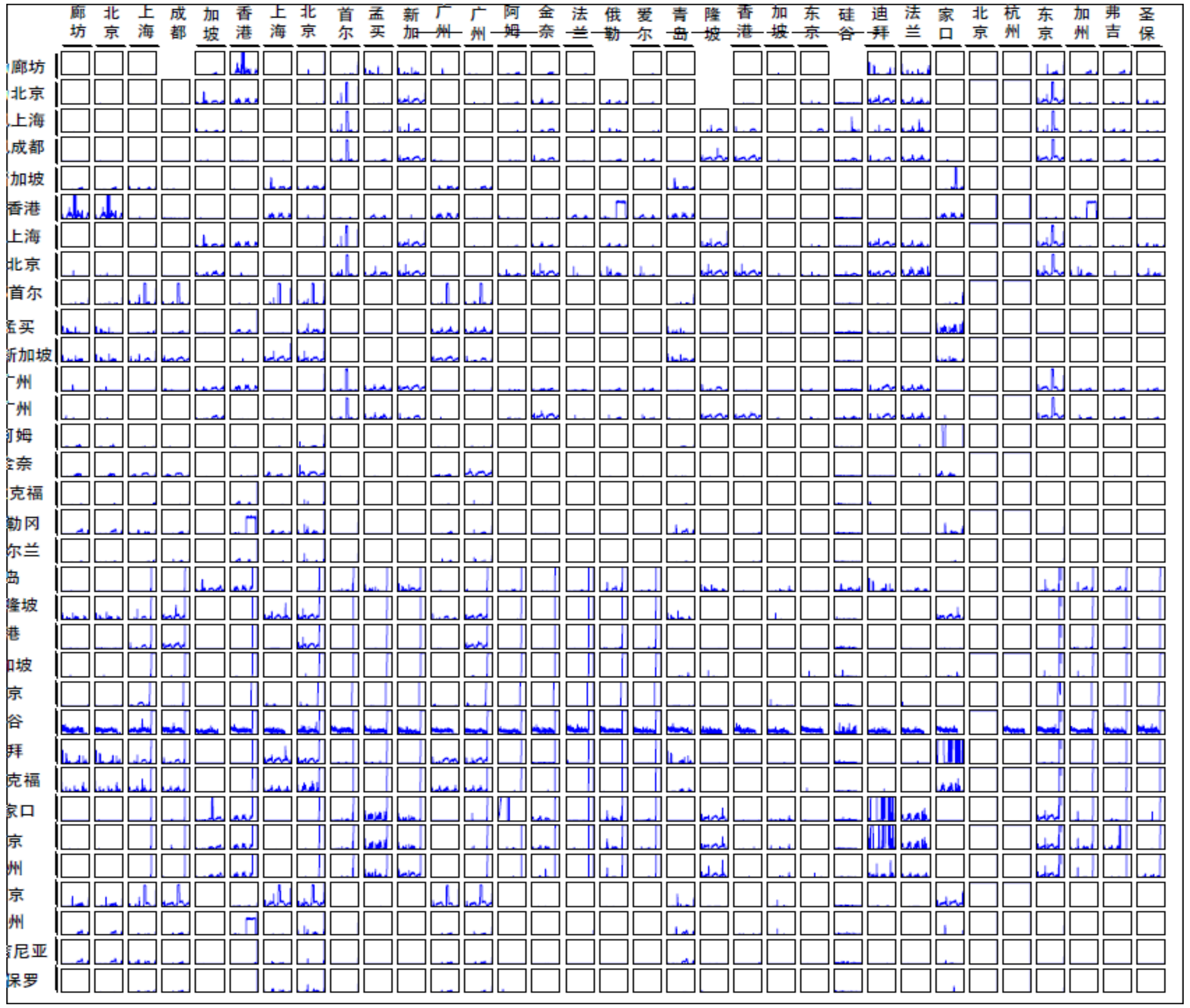
- **Top half of small square represents 99% percentile data, bottom half represents 99.9% percentile data**
- **Test every 20 seconds, 2000 ping packet, 55 hours**
- **Gray - satisfied, White - not satisfied**
- **Red rectangle – ON pairs meet delay requirement but not meet loss rate requirement**

Loss rate requirement may not be met even when delay requirement is met. Packet loss needs to be addressed independently

Packet loss in long haul network

- Tail loss or short flows:
  - Need to wait for timeout in tail loss.
  - E2e retransmission may take an additional RTT for non-tail loss for short flows.
  - Significant factors cause long FCT, esp for short flows.
- Packet loss in real time streams:
  - RTCP NACK based RTP retransmission takes additional e2e RTT which may miss the playout time.
- Packet loss in large flows like bulk data transfer:
  - Loss based congestion control at sender reduces the sender's sending rate even when the loss is not caused by a persistent congestion.
  - Throughput degradation.

Loss on a single segment may have significant effect on e2e path



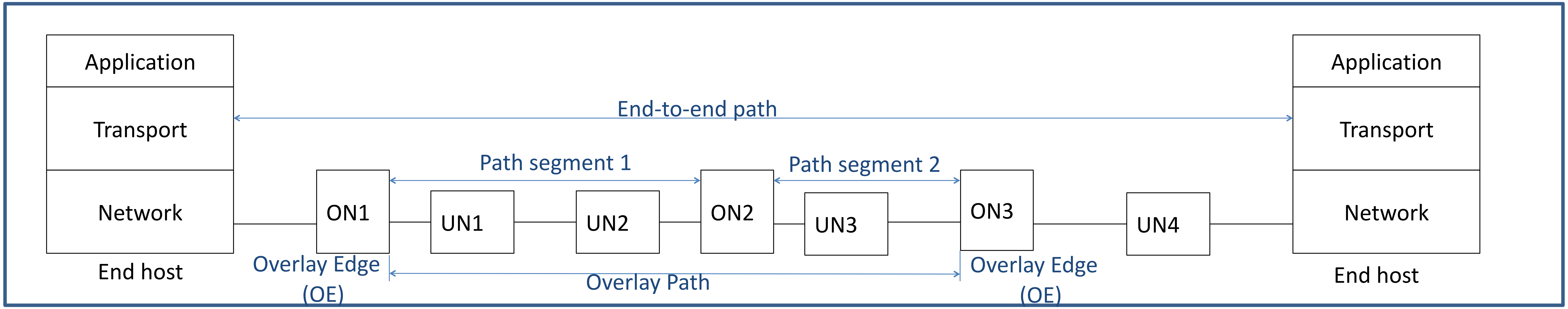
- Loss over path segments between node pairs has different characteristics and vary over time
- Loss over a short segment may affect end to end path loss rate significantly
- Goal: solve the shortest board of the bucket in terms of packet loss

New Opportunities in Solving the Problems

- Overlay nodes partitions the whole path to shorter segments
  - in better position to deduce the network status
  - more responsive to loss and local delay change
- Overlay nodes have computing and memory resources:
  - capable of providing complex functions like loss detection & recovery.
  - measurement between overlay nodes
  - ECN marking and overall increasing capability of ECN in network

Localized Optimizations On Path Segment (LOOPS) to provide local (ON to ON) best-effort reliability

ON - Overlay node
UN - Underlay node

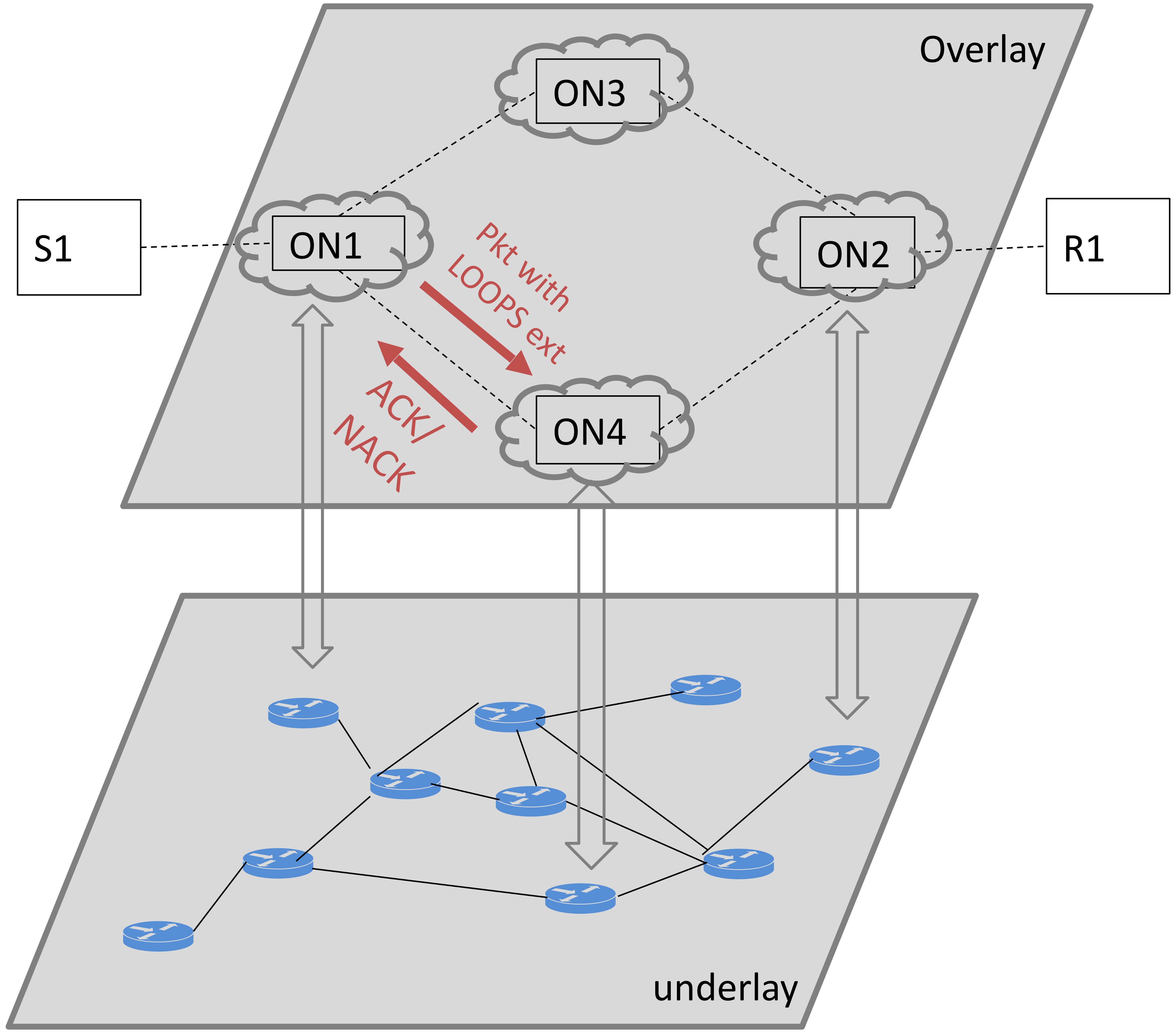


LOOPS main feature: Local in-network recovery, by retransmission and/or FEC

Potential impacts to be considered:

- Local recovery and end-to-end retransmission
- Interaction with end-to-end congestion control

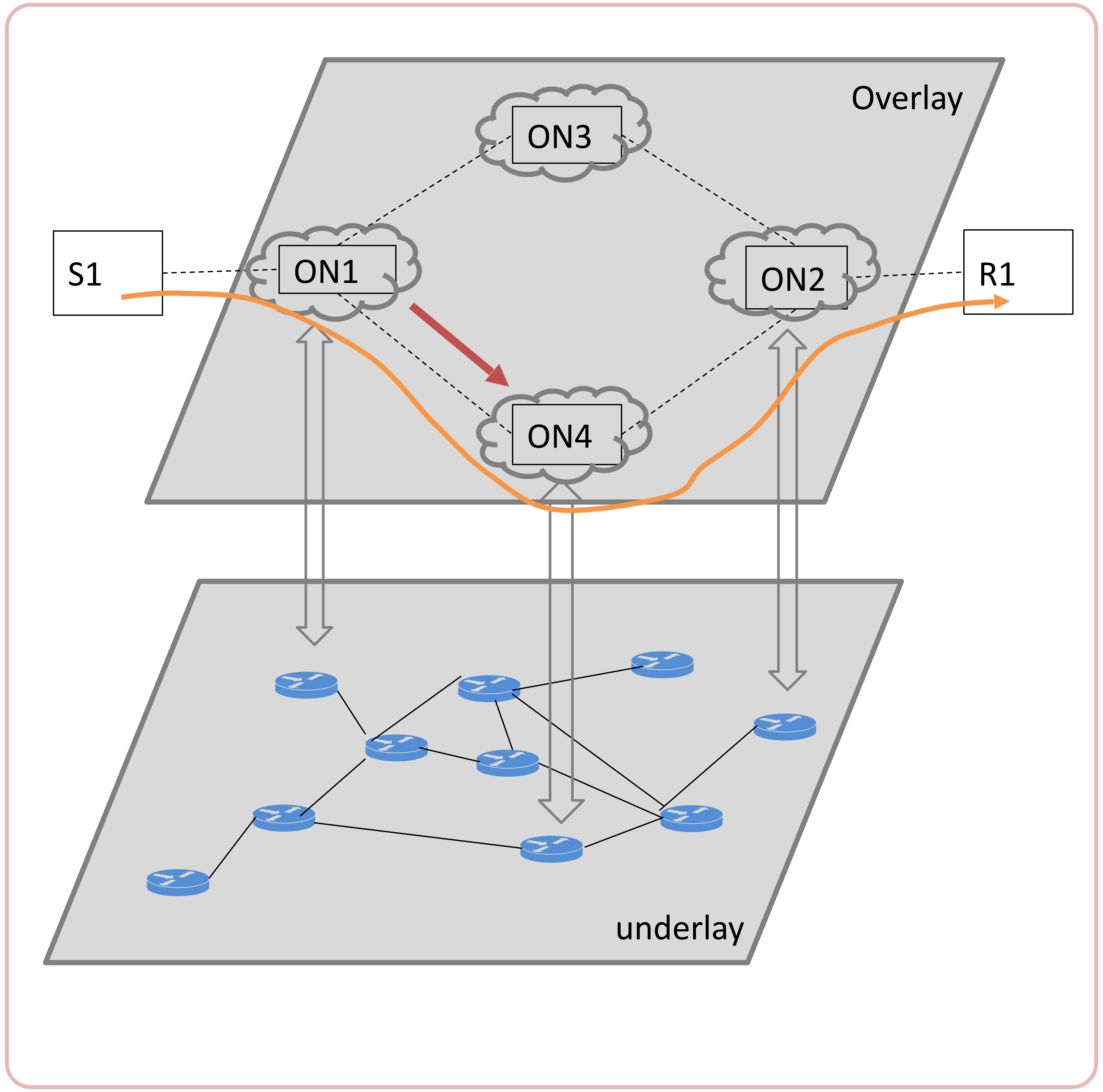
Elements of a solution 1/3 – Local recovery



Potential approaches:

1. Local retransmission
   - Separate packet number space
   - ACK and NACK list to notify the loss
2. FEC:
   - Fine tune redundancy and window size

Elements of a solution 2/3 – Congestion Control interaction



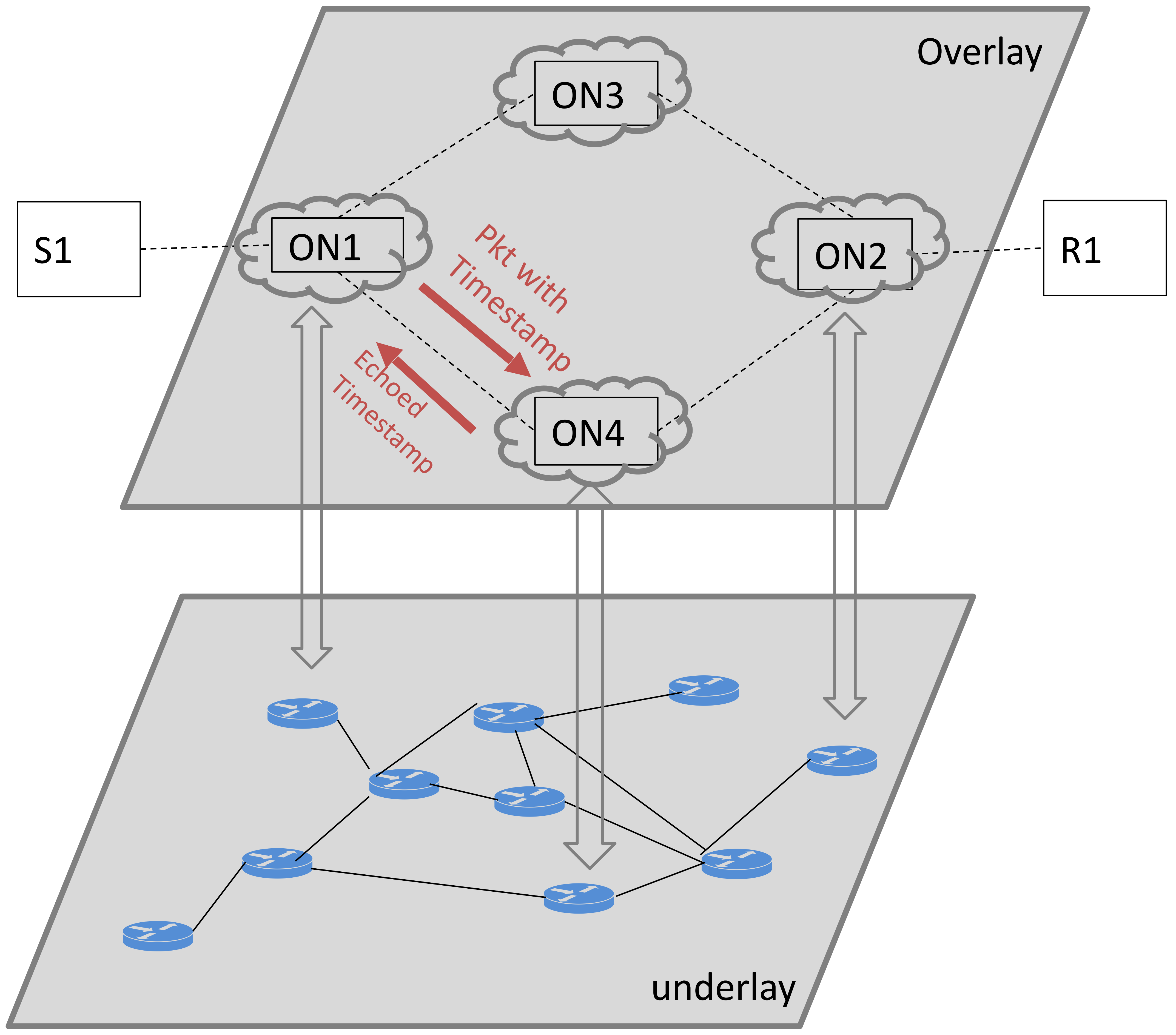
Why required?

- Local recovery may cause the loss based TCP CC not invoked correctly as the sender may be hidden from a loss event

Approach:

- ON helps determine if a packet was lost due to non-congestion
  - Local recovery makes a significant benefit to throughput (esp with re-ordering buffer at egress node)
  - Most TCP sender enables spurious retransmission detection, it can undo the previous unnecessary window shrinkage (if there was) to improve throughput without re-ordering buffer
- A packet was lost due to congestion
  - mark ECN/CE

Elements of a solution 3/3 – Local Measurement



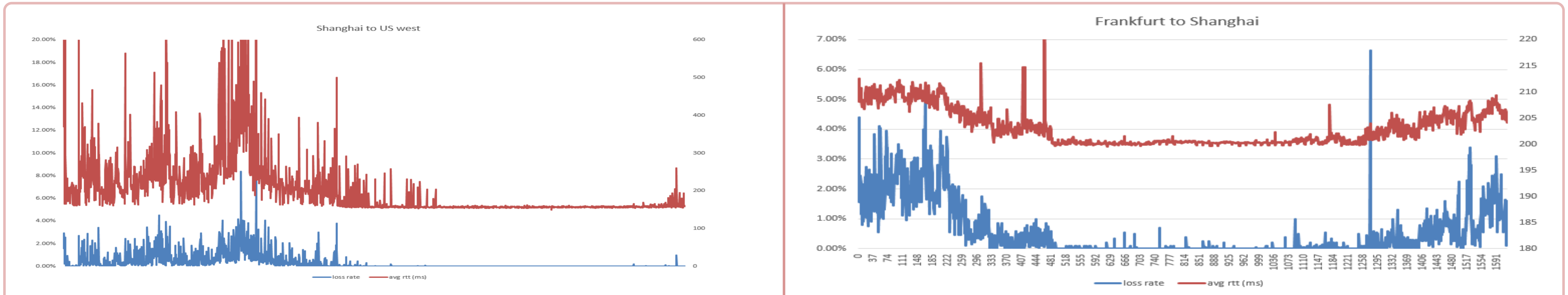
Why required?

- Help determine settings, like rtx timeout, FEC parameters.
- Help determine ECN/CE marking decision for CC interaction

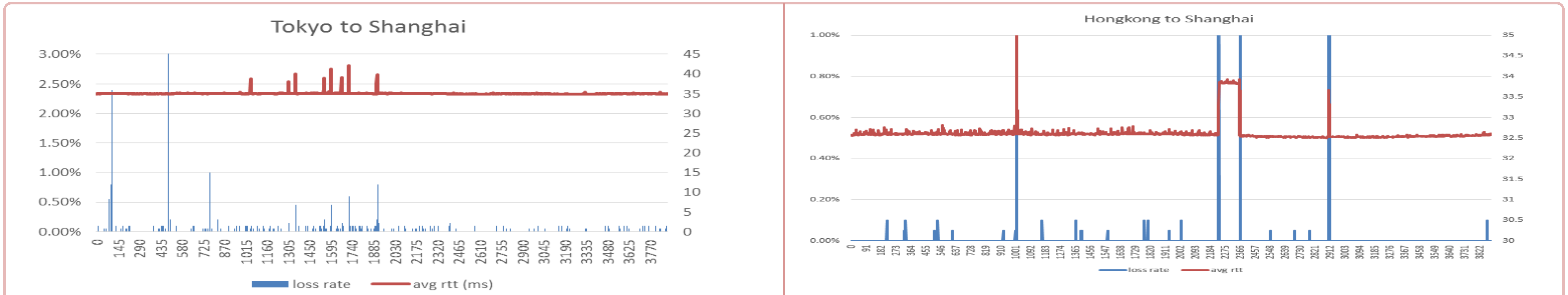
Approaches:

1. Calculate loss rate
2. Use timestamping to measure delay and its variation over a path segment
3. Use time based congestion detection to decide ECN/CE marking for recovered packet

Tests show cause of packet loss can be deduced by measurement in some cases



Delay and packet loss rate changes have strong correlation



Delay and packet loss rate changes have weak correlation

- In case of strong correlation, high delay means congestion loss, otherwise non-congestion loss.
- In case of weak correlation, non-congestion loss or some CAR(committed access rate) functions somewhere. Need more mechanism to identify.

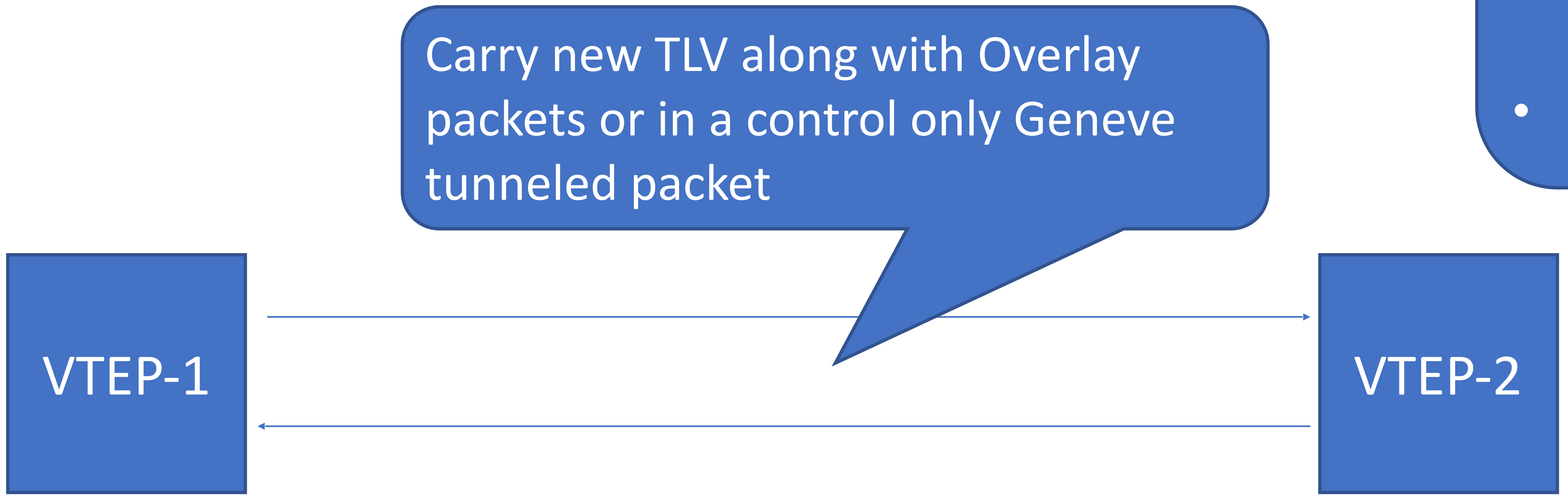
Summary

- LOOPS features:
  - Local recovery
  - Local measurement
  - Congestion control interaction
- Binding to current overlay encapsulations, do not invent a new encap.

Overlay network segment resiliency using a new Geneve Option

New Geneve Option TLV carrying:

- Send Packet seq #
- Last Rx packet seq #
- Timestamp (Optional to calculate RTT)
- Missing Seq#(s) (Optional)



Carry new TLV along with Overlay packets or in a control only Geneve tunneled packet

- If TimeStamp is present in the option TLV, then receiving VTEP is required to send back immediately the option TLV in either the next scheduled tunneled packet back to the sender VTEP, or if none scheduled in a control only geneve packet. This to allow the sender VTEP to calculate RTT.
- If Missing Seq #(s) is received then Received VTEP resends the missing tunneled packets.

Charter

Background

- The Internet has a long history of employing performance enhancing proxies (PEPs, RFC 3135) to improve performance over paths with links of varying quality. Today's PEPs often interact deeply and "transparently" (intrusively) with end-to-end transport and application layer protocols. This practice is coming to an end with increasing deployment of encryption.

- At the same time, network structures are becoming more complex, and network nodes are becoming more powerful. It is becoming more viable to trade processing power in network nodes against path quality, in particular for expensive path segments. Transport protocols and their implementations are moving towards playing better with forwarding node functions such as ECN marking and AQM.

LOOPS (1)

- LOOPS (Local Optimizations on Path Segments) attempts to capture opportunities opened by these developments, enabling optimizations within segments of an end-to-end path. Typically, these segments are located between overlay nodes (tunnel endpoints), which allows a local optimization protocol to run between these nodes. Many end-to-end flows can be aggregated into each such tunnel flow being optimized.

- Initially, LOOPS will focus on path segments that do not include either end host. Also, multipath forwarding will not be specifically addressed.

LOOPS (2)

- The functions to be addressed by LOOPS include:
- **Local recovery**. Packet losses on the path segments are recovered autonomously, removing the need to burden the entire end-to-end path with the recovery, and decreasing the latency by which these recoveries can be effected. Local recovery can be based on forward error correction (FEC) and/or (non-persistent) retransmission.
- **Local measurement**. To properly parameterize the LOOPS algorithms (e.g., RTO, FEC rate), measurements are continuously performed of the path segment between the tunnel endpoints.
- **Interaction with end-to-end congestion control**. Based on configuration and measurement information, losses can possibly be categorized into congestion and non-congestion losses. The losses that cannot be positively identified as non-congestion losses are relayed as congestion events to the end-to-end congestion control. Circuit breakers (RFC 8084) may be employed to further protect against congestion collapse.

LOOPS (3)

- The LOOPS protocol will need to run embedded into a variety of tunneling protocols. To this end, LOOPS will be defined on a generic information model level, and initial bindings will be defined for a small set of tunneling protocols to be selected by the working group.

Relationship to Other WGs and SDOs

- LOOPS will work closely with developments in TSVWG, TCPM, and in particular with QUIC as an example of a transport protocol that may more readily absorb features of interaction with LOOPS segments. However, there is no dependency — LOOPS is designed to optimize already in the presence of legacy transport protocols. LOOPS will interact with the homes of tunneling protocols to which bindings are being defined, which depending on the choices of the WG may include NVO3 (Geneve), intarea (GUE), spring/6man (SRv6). [Add more here. Any other SDOs?]

Milestones

- LOOPS generic information model and protocol, WG document adopted, October 2019
- LOOPS generic information model and protocol, PS to IESG, May 2020
- LOOPS binding candidates identified and WG documents adopted, February 2020
- LOOPS binding to tunneling protocol A, PS to IESG, July 2020
- LOOPS binding to tunneling protocol B, PS to IESG, September 2020

(F)AQ (1)

- So this is only about encrypted traffic?
  - Any traffic is welcome, we just don’t try to peek beyond L3 info
- So how do you know which packets are worth recovering?
  - Today we don’t. If more L3 marking becomes available, we’d use it.
- How do you transport your measurement-related information?
  - Forward info: In encapsulation extension (e.g., with sequence number). Reverse info: The same way we transport the ACK channel. Depends on encapsulation.

(F)AQ (2)

- How do you avoid spending more for LOOPS encapsulation than the performance enhancement is worth?
  - LOOPS will need some management that is weighing this (and doing the pair setup in the first place).
- How to relay congestion for non-ECN-capable transports?
  - Dropping. Or, actually, not even requesting a retransmission when a congestion event would be relayed anyway.

Discussion

Next Steps

- Set up mailing list for discussion (loops@ietf.org?); github: **loops-wg**
- Sketch out (no completeness required) a **solution** document:
  - Sketch a basic Information Model and a recovery protocol using that
  - Explain how measurement and congestion relay can work with that
  - Sketch a couple of encapsulation bindings
- Refine **charter** based on input
- If this works: do **BOF** request by end of April
  - Identify roles of contributors, find BOF chairs
- Run BOF in July (IETF105@Montreal), **form WG**