Packet-Level Telemetry in Large Datacenter Networks

Yibo Zhu, Nanxi Kang, Jiaxin Cao, Albert Greenberg, Guohan Lu, Ratul Mahajan, Dave Maltz, Lihua Yuan, **Ming Zhang**, Ben Zhao, Haitao Zheng

UCSB

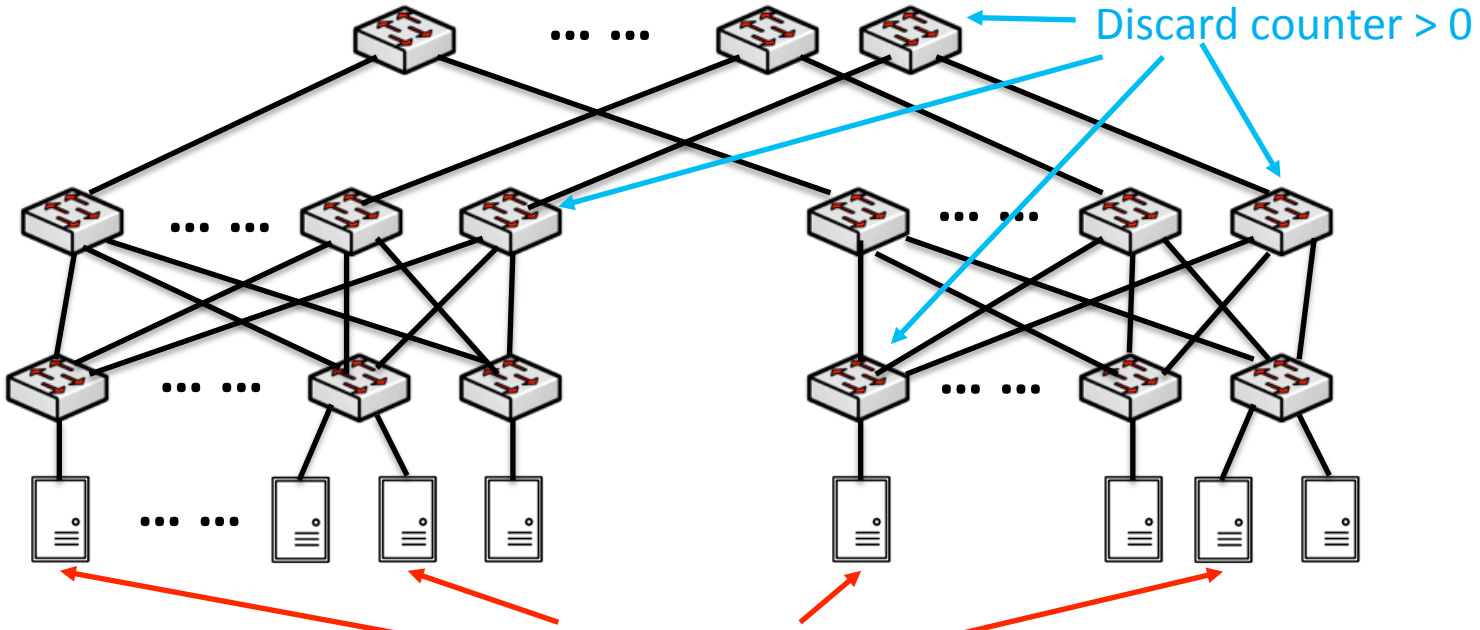
Princeton

Datacenter networks (DCN)

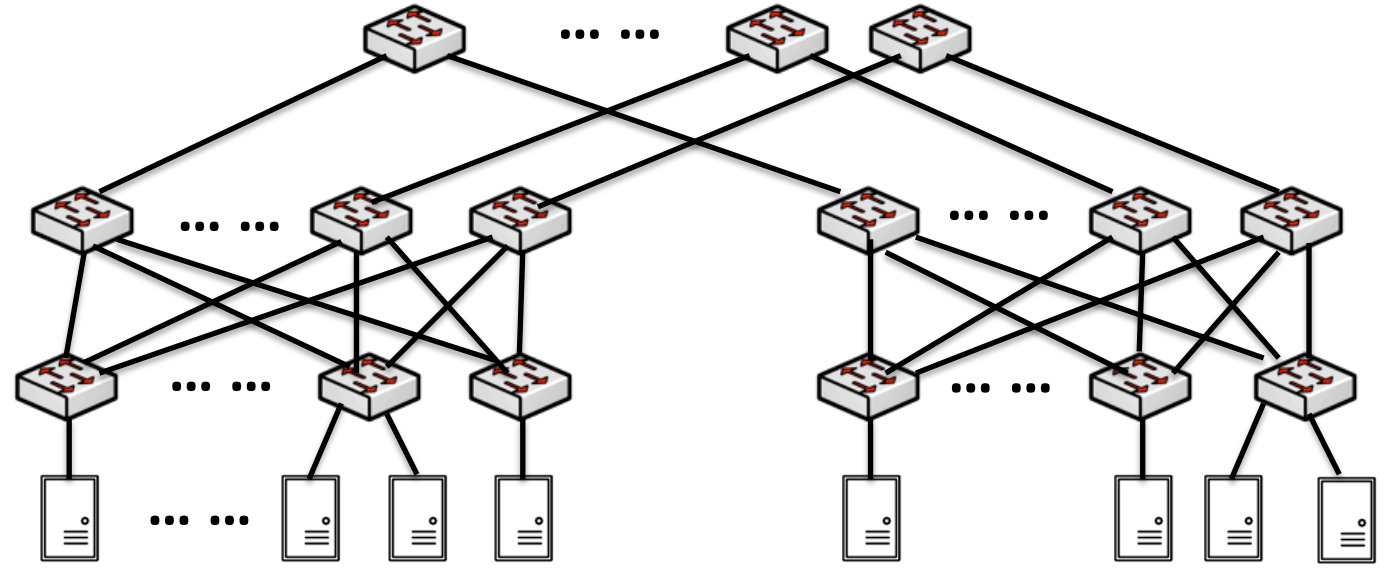
- Datacen complex
  - O(10K l O(100K) servers
  - Low-c vare stack
- Network
  - Packet ughput, load imbala
- Debuggi
  - Why a

Example #1: silent packet drops

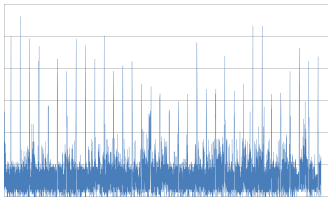
Too many switches/links: hard to localize using ping/traceroute



Example #2: latency spikes



Sender

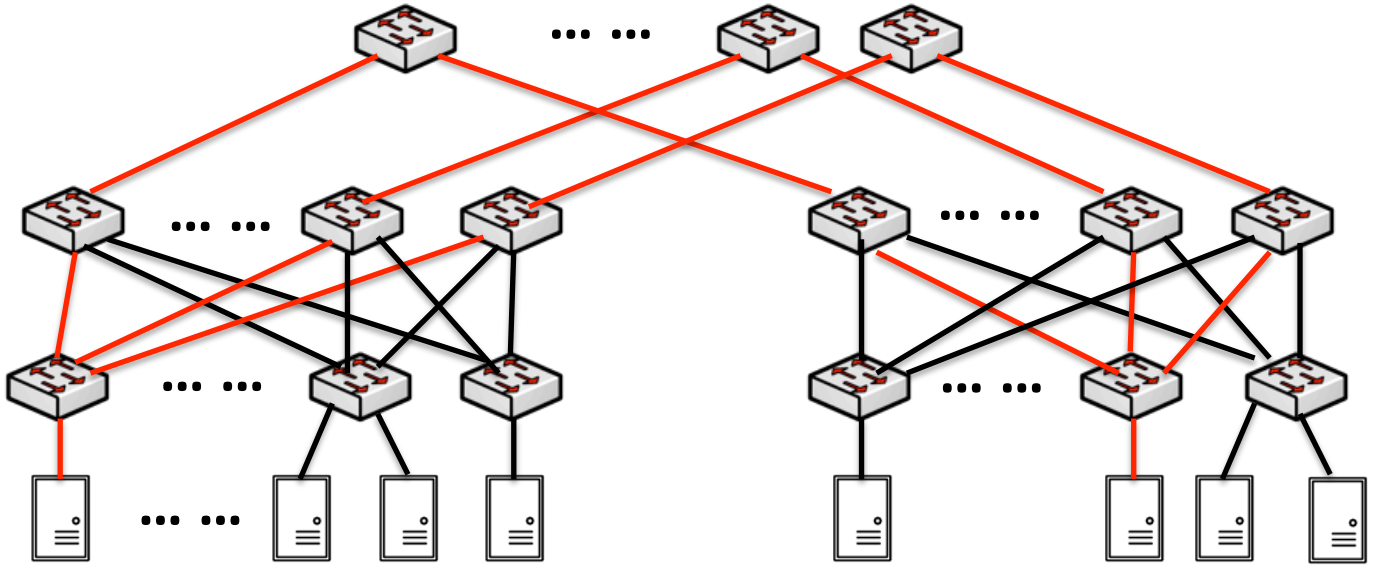


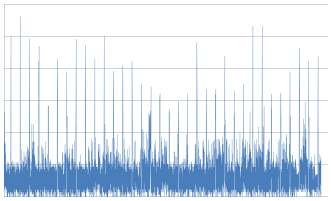
Receiver

Example #2: latency spikes

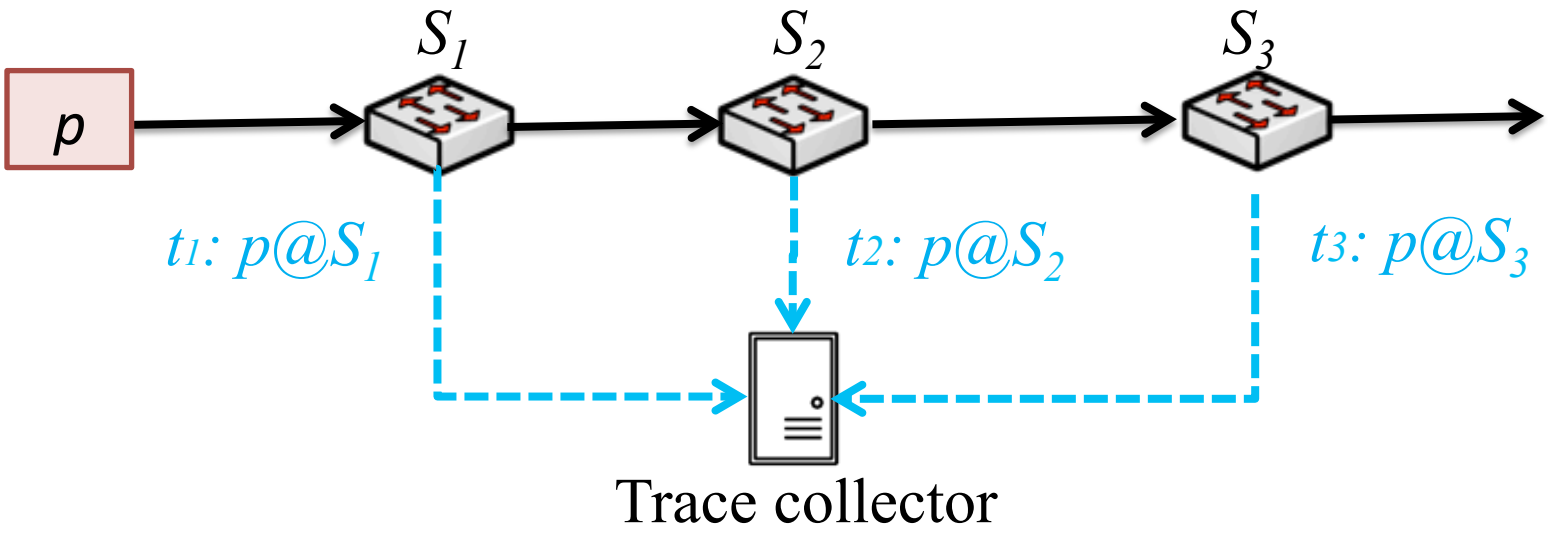
Interface counters: too coarse-grained

Ping/traceroute: cannot measure per link latency



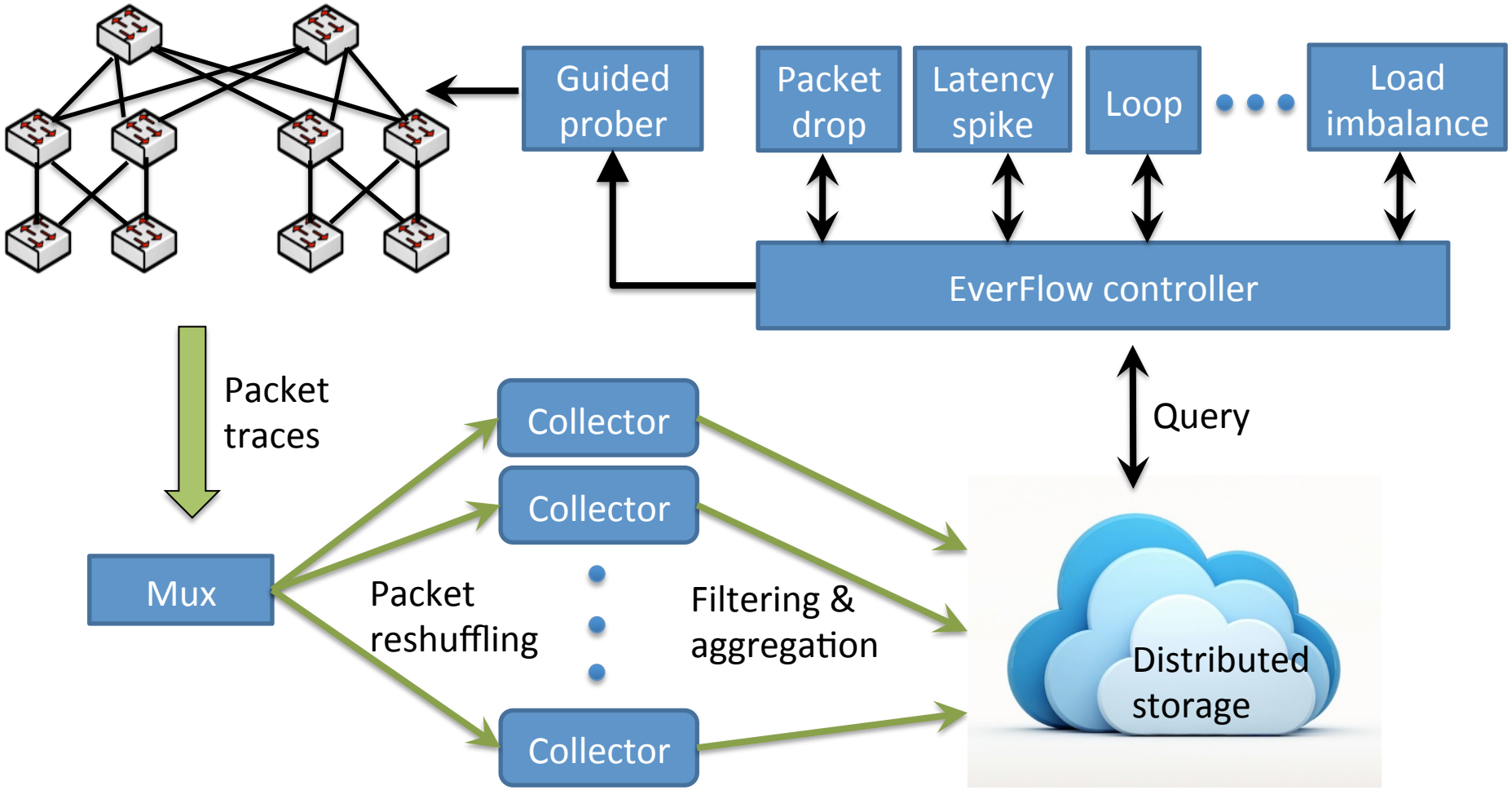


Solution: packet tracing

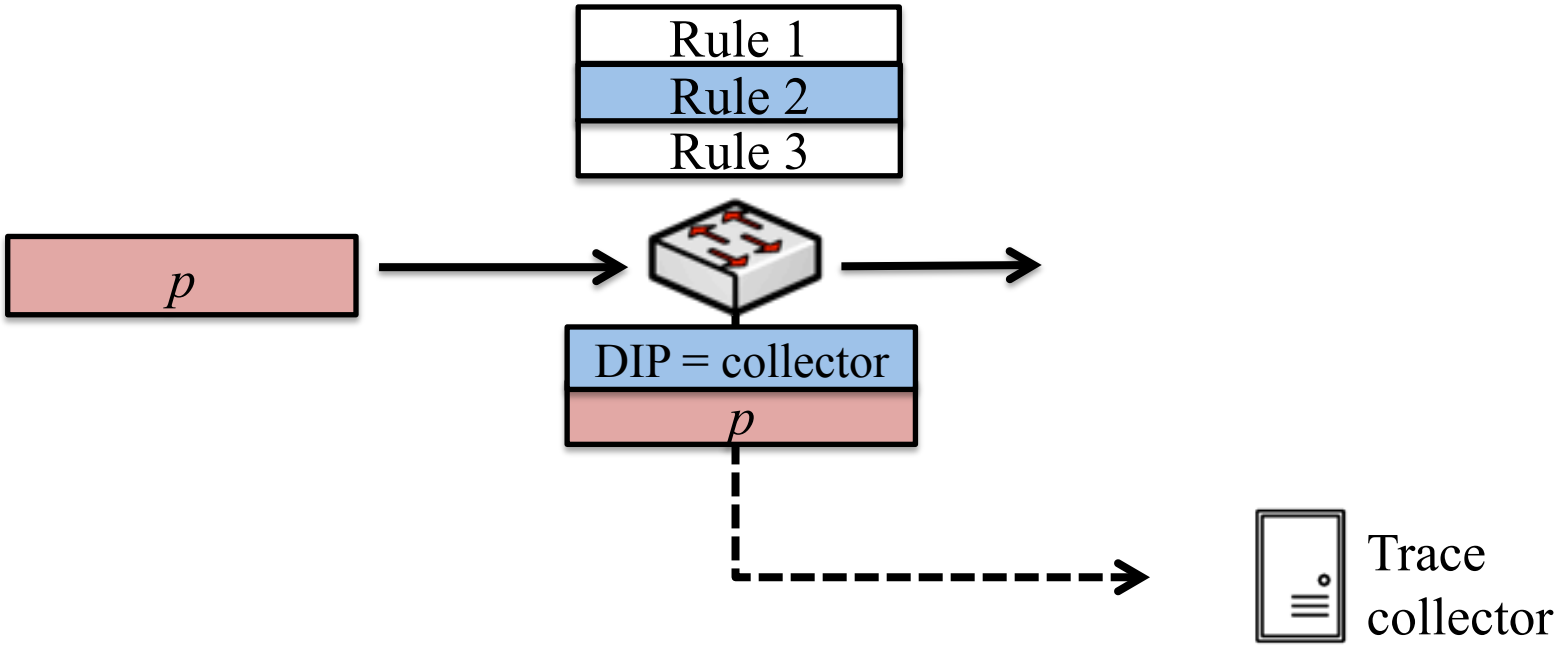


- Tracing packet *p* at every hop
  - Locate drop from *p*'s last appearance
  - Identify bottleneck from per-hop latency

Packet-level network telemetry



Data-plane match & mirror

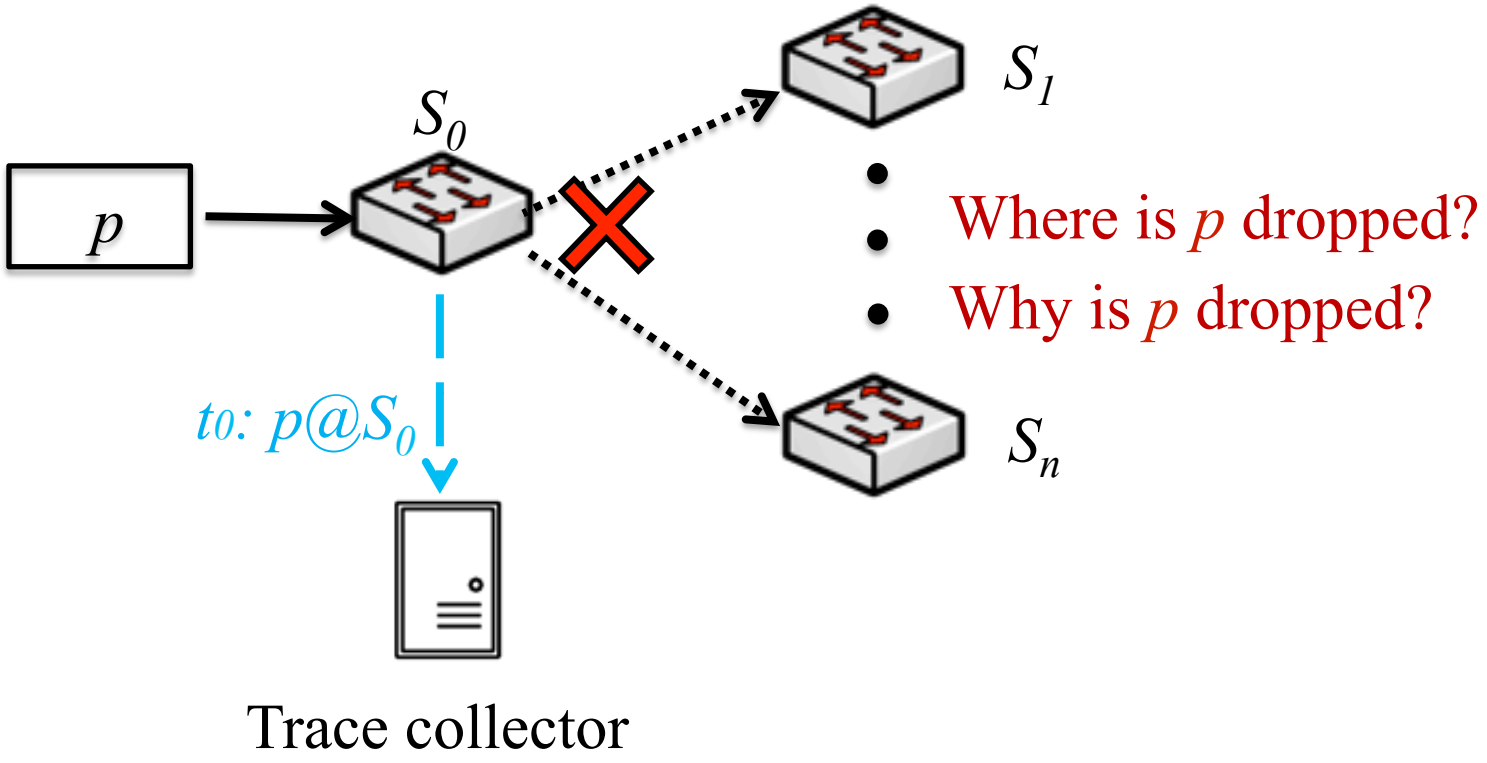


- Rule-based match & mirror for scalability
- Huge capacity with zero control plane overhead

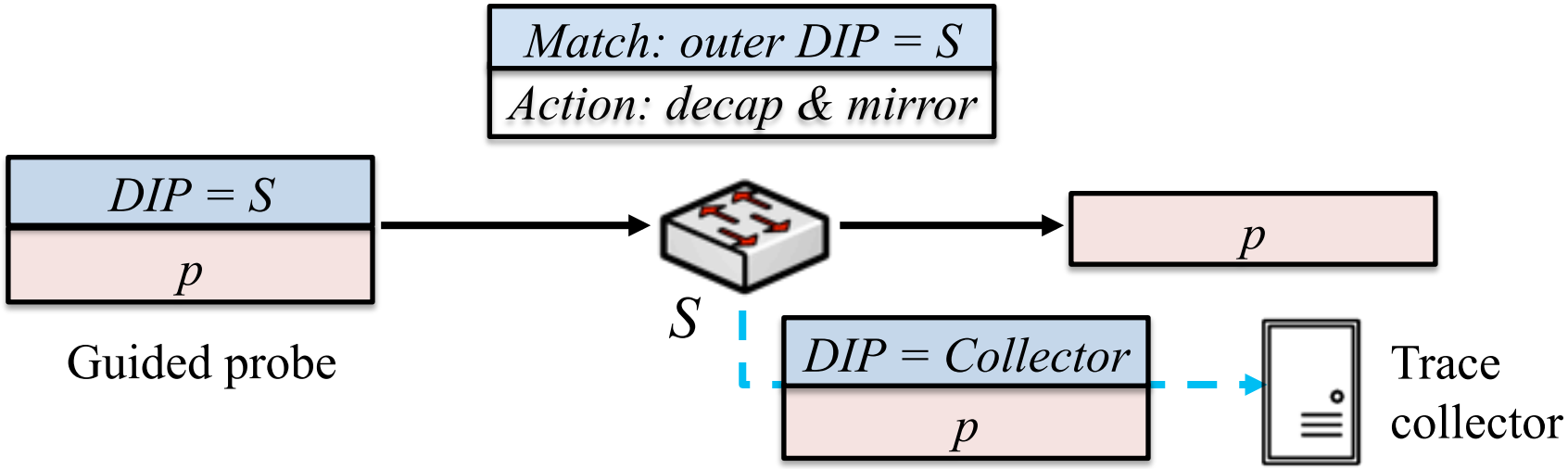
Match & mirror rules

- Rules based on existing chip
  - IPID-based random sampling
  - One bit in DSCP field: selective packet tracing
  - TCP SYN/FIN/RST: every TCP flow
  - Protocol traffic: BGP, PFC, RDMA
- Support needed from P4 programmable chip
  - Match on (hash value of) certain packet fields
  - Truncate mirrored packet

Challenges with packet drops



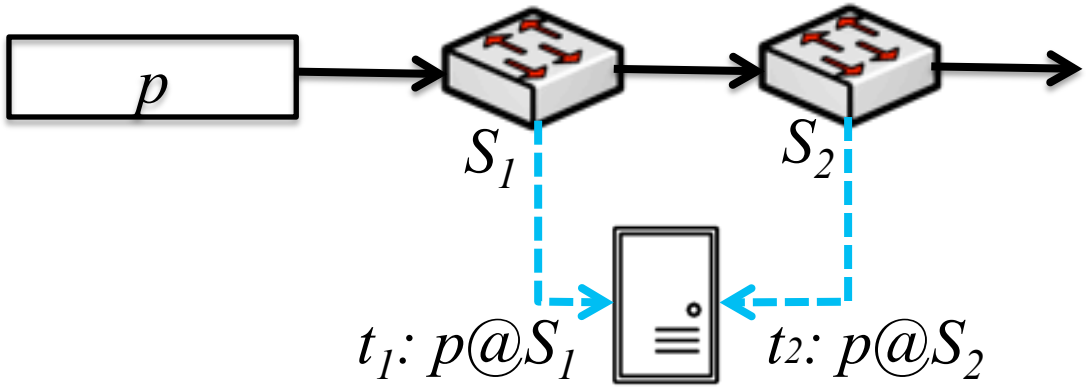
Debugging packet drops



- Current solution:
  - Inject guided probe into suspect switches
- P4 programmable chip/NIC:
  - Export metadata: incoming/outgoing port, matched rule
  - Mirroring triggered by metadata: packet drop reason

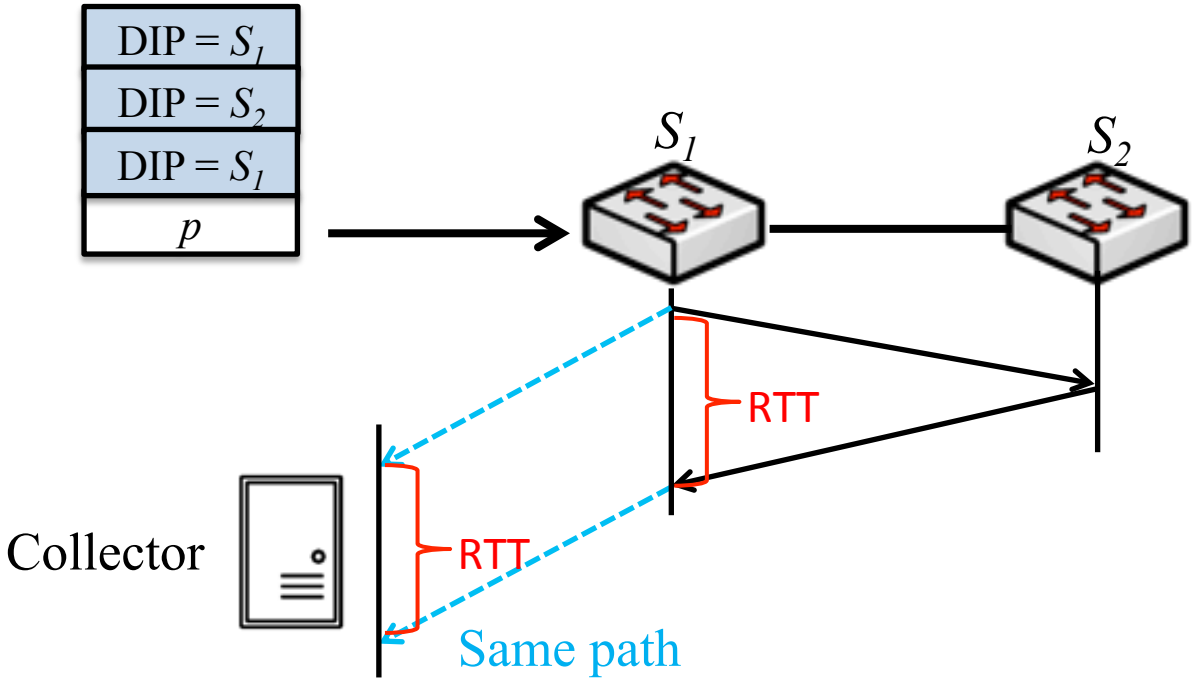
Challenges with link latency

- Switch does not provide timestamp



$t_2 - t_1$!= latency of $S_1 \rightarrow S_2$

Measuring link latency



- Current solution
  - Inject guided probe to bounce between *S1* and *S2*
- P4 programmable chip
  - Attach switch timestamp to mirrored packet

Conclusion

- Packet-level telemetry is both *crucial* and *practical* in large-scale DCNs
  - Packet drop, latency spike, load imbalance…
- P4 programmable chip/NIC will greatly enhance the utility of EverFlow
  - Export packet metadata in mirrored packet
  - Mirroring triggered by packet metadata