Network Replay and Consistency Across Testbeds

Alexander Wolosewicz (IL Tech), Vinod Yegneswaran (SRI),

Ashish Gehani (SRI), Nik Sultana (IL Tech)

ILLINOIS TECH
College of Computing

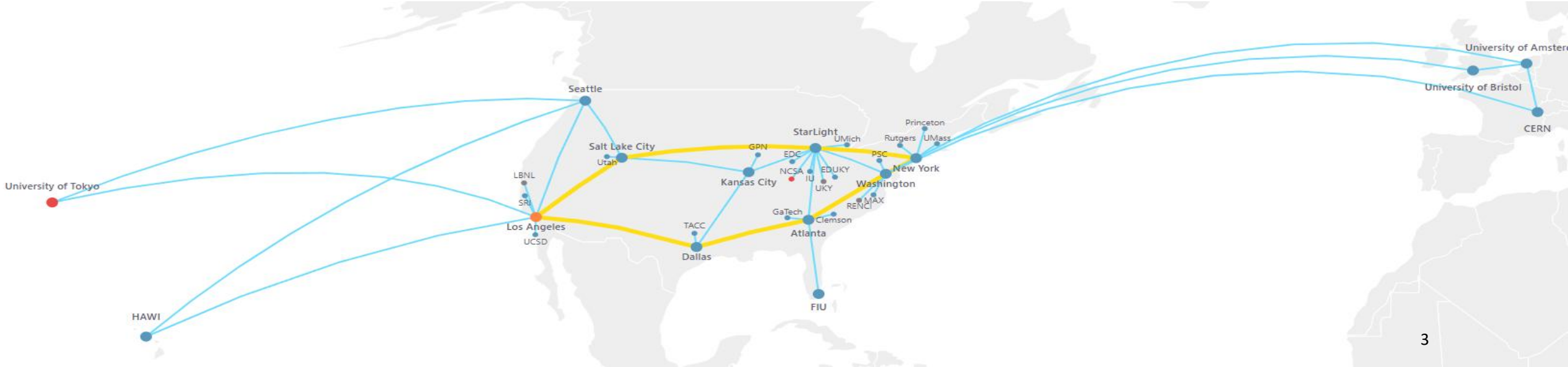
Introduction

- Network testbeds use virtualization to share resources among experimenters
- Shared infrastructure can introduce congestion, jitter, and loss which impact both artifact reproducibility and debugging
- Consistent replaying can assist, but existing techniques rely on non-commodity or non-shared hardware, or are low-bandwidth
- We built **Choir**, a consistent replayer which can function in virtual networks at 100 Gbps

ILLINOIS TECH
College of Computing

Background

- **FABRIC:** National testbed with 33 sites across the US and international partners, experimenters are virtual network tenants
  - 23 sites have PTP support for VMs
- **DPDK:** C library for high-performance packet processing



Network Consistency

- In evaluating **Choir**, we find variations in replay consistency in some environments
- Realize that running replays can allow for measuring overall network consistency
- Build metrics to quantize this

ILLINOIS TECH
College of Computing

Consistency Metrics

- Four sub-metrics for **U**niqueness, **O**rdering, **L**atency, **I**ATs
- Numerator a measure of distance, denominator max possible value (normalize)
- Combine into 4D vector, use magnitude, subtract from 1
  - 1 = complete consistency

$$U_{AB} = 1 - \frac{2 \times |A \cap B|}{|A| + |B|} = U_{BA}$$

$$O_{AB} = \frac{\sum_{i=0}^{|B|} d_i}{\sum_{n=0}^{|A \cap B|} n} = O_{BA}$$

$$L_{AB} = \frac{\sum_{i=0}^{|A \cap B|} \text{abs}(l_{Ai} - l_{Bi})}{|A \cap B| \cdot \max(t_{B|B|} - t_{A0}, t_{A|A|} - t_{B0})} = L_{BA}$$

$$I_{AB} = \frac{\sum_{i=0}^{|A \cap B|} \text{abs}(g_{Ai} - g_{Bi})}{(t_{B|B|} - t_{B0}) + (t_{A|A|} - t_{A0})} = I_{BA}$$

$$\kappa_{AB} = 1 - \frac{\sqrt{U_{AB}^2 + O_{AB}^2 + L_{AB}^2 + I_{AB}^2}}{2} = \kappa_{BA}$$

ILLINOIS TECH
College of Computing

Deriving Maximums

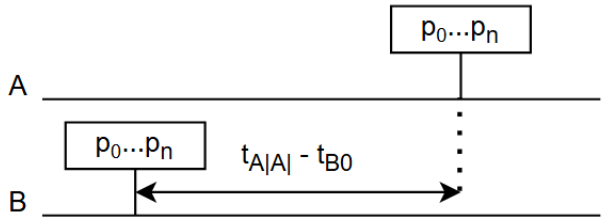


Figure 2: The maximum possible L situation.

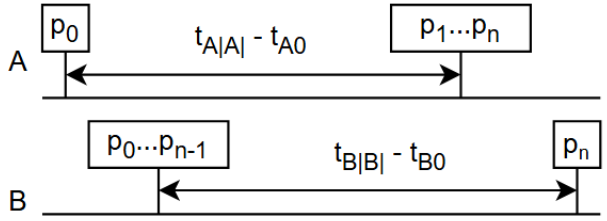


Figure 3: The maximum possible I situation.

- Max **L**: where all packets are at one end of A and the other end of B
- Max **I**: where the first gap is the entire time of A, and where another gap is the entire time of B

ILLINOIS TECH
College of Computing

Choir

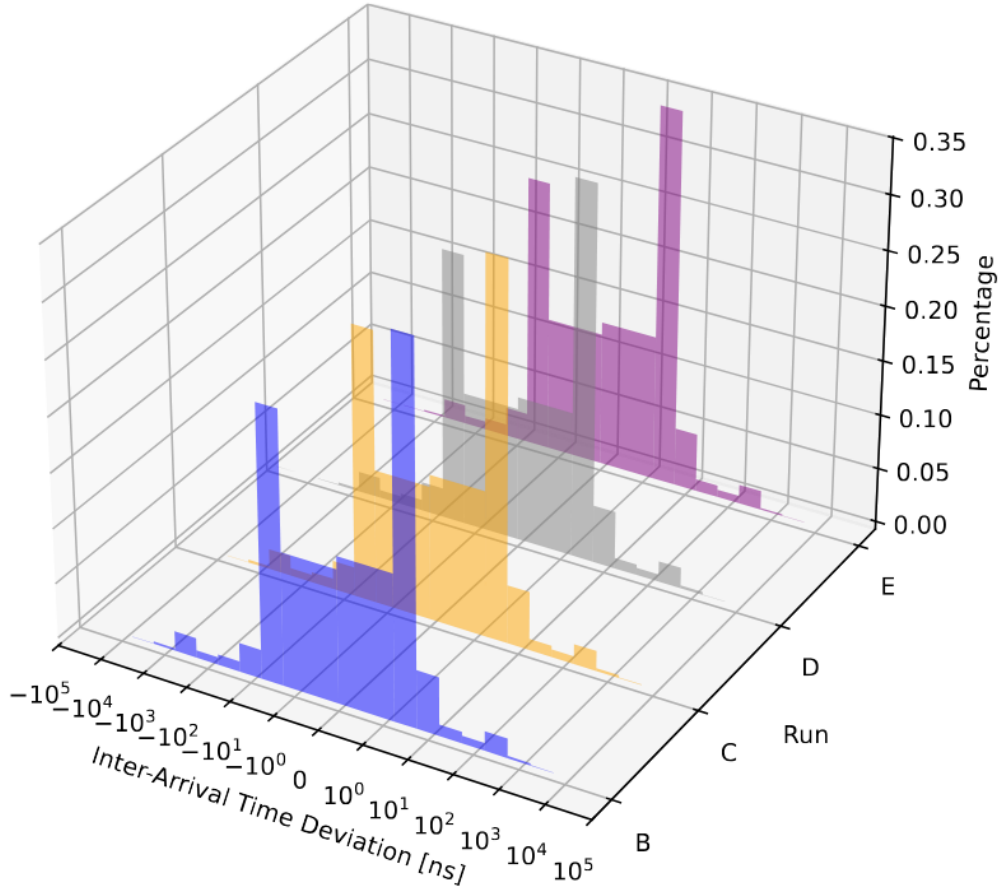
- Middleboxes, when inactive just forward traffic, are commanded to record and run replays
- When recording, hold packets in memory (don't free on TX), and store transmit times (TSC counts, constant-frequency on FABRIC)
- To run, calculate TSC delta, loop over precise TSC reads for new TX
- Accuracy bounded by NIC sending delay
  - Do not use state-of-the-art trick as in Moongen (IMC 2015) of invalid packets to fill TX queue due to virtual environment

ILLINOIS TECH
College of Computing

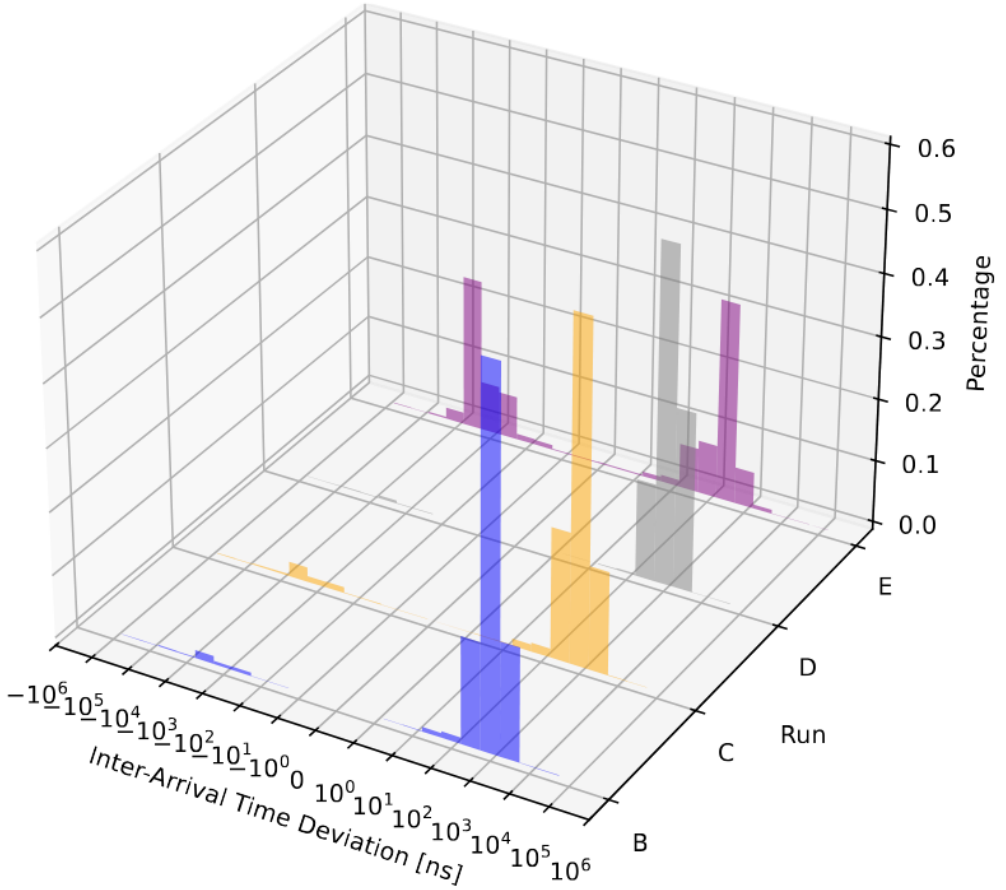
Evaluation (Local)

- Start by evaluating the consistency of the replay
  - 0.3 seconds of 40 Gbps (1.055M packets, 3.519 Mpps)
- **Setup:**
  - **Generator:** Xeon E5-2678 @ 2.5 GHz, Mellanox ConnectX-5
  - **Replayer:** Xeon E5-2670 @ 2.3 GHz, Mellanox ConnectX-5
  - **Recorder:** Xeon E5-4620 @ 2.2 GHz, Intel E810
  - **Connecting Switch:** AS9516-32D Tofino2
- 92.23-92.51% of packets had IAT deviations <= 10 ns
- Most between 500 ns – 5 µs latency variation

ILLINOIS TECH
College of Computing



The percentage of packets with a given IAT delta.



The percentage of packets with a given latency delta.

ILLINOIS TECH
College of Computing

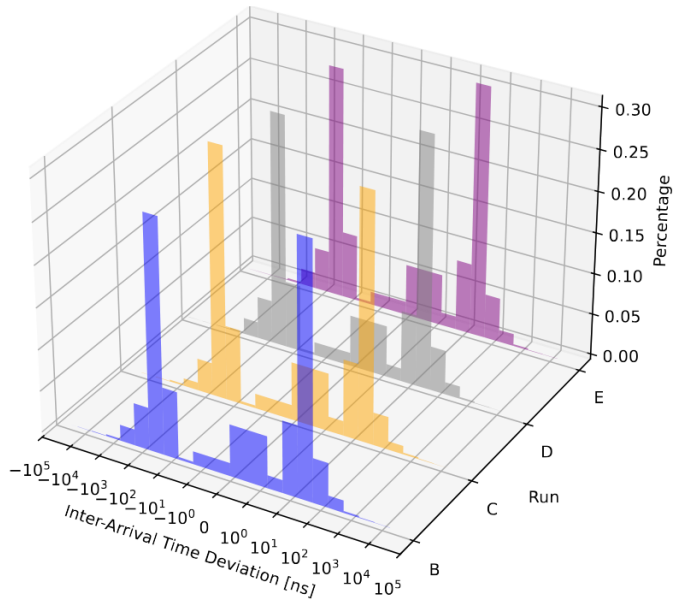
Evaluation (FABRIC)

- All nodes: 3 CPU threads
  - Shared: Mellanox ConnectX-6 SR-IOV Virtual Functions
  - Dedicated: Mellanox ConnectX-6
- Worse IAT deviation (~25-30% <= 10 ns)

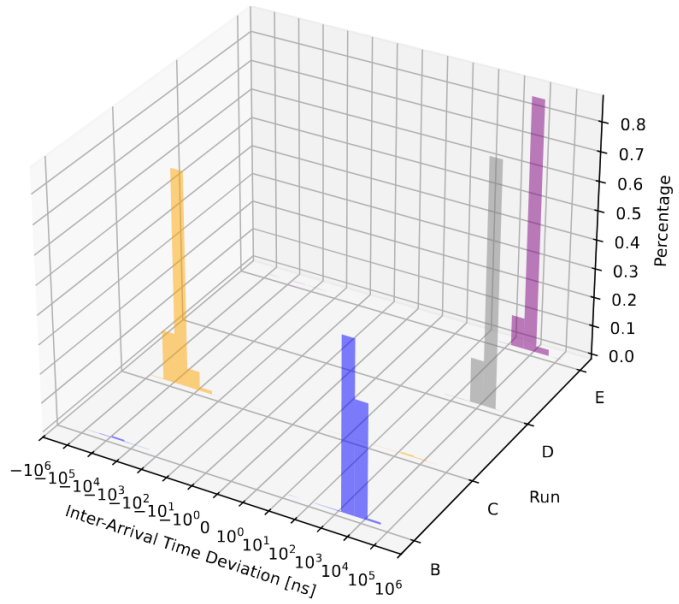
ILLINOIS TECH
College of Computing

Shared vs Dedicated (40 Gbps)

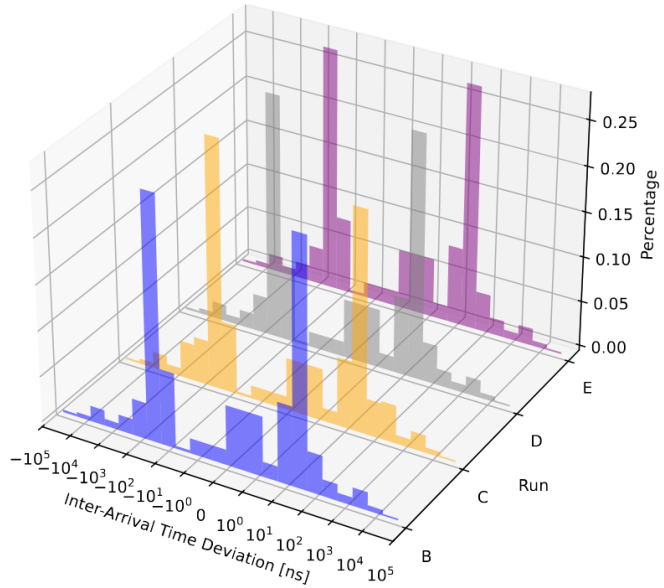
- Top: shared, bottom: dedicated
- Dedicated has worse IAT outliers, worse latency



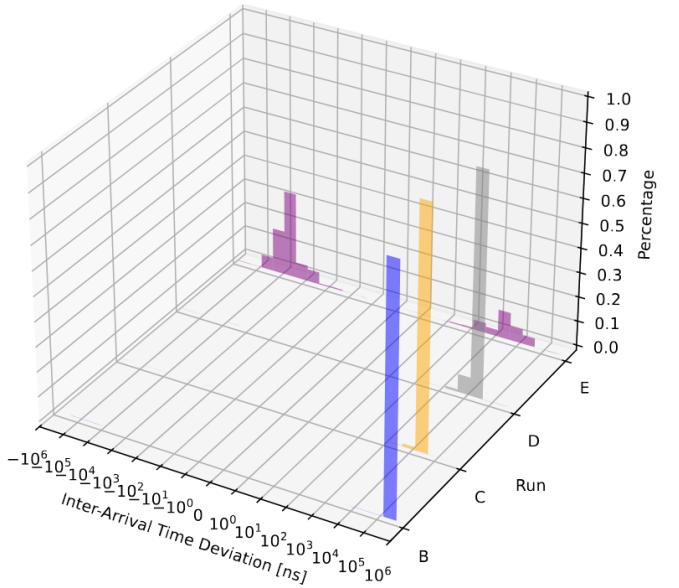
The percentage of packets with a given IAT delta.



The percentage of packets with a given latency delta.



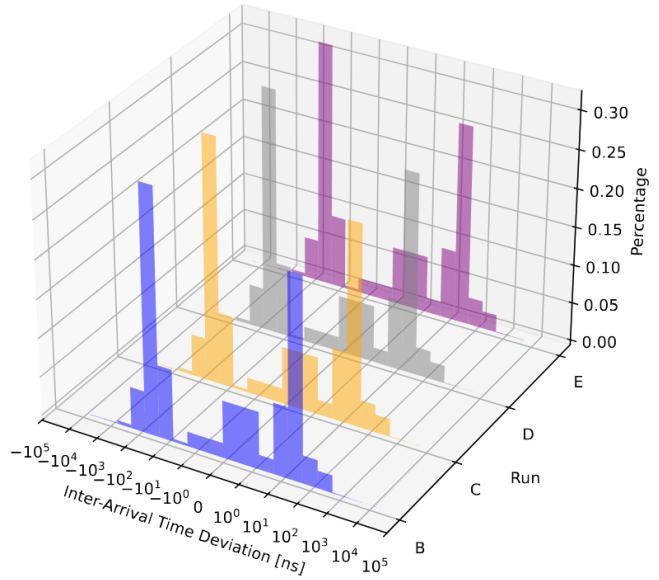
The percentage of packets with a given IAT delta.



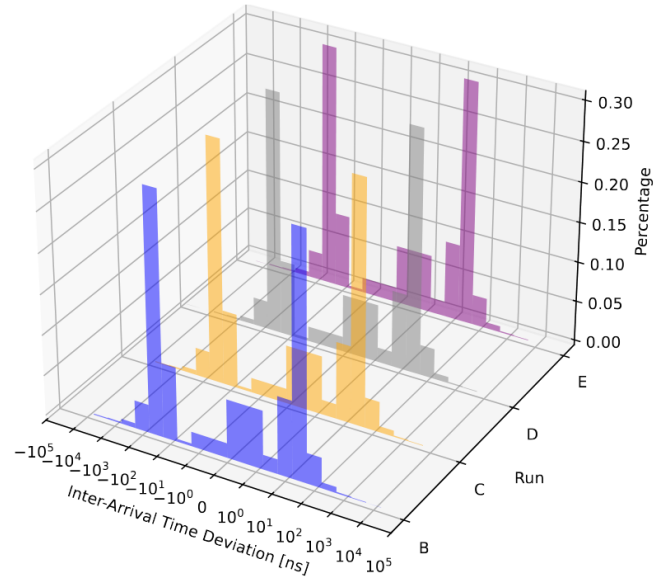
The percentage of packets with a given latency delta.

80 Gbps and Adding Noise

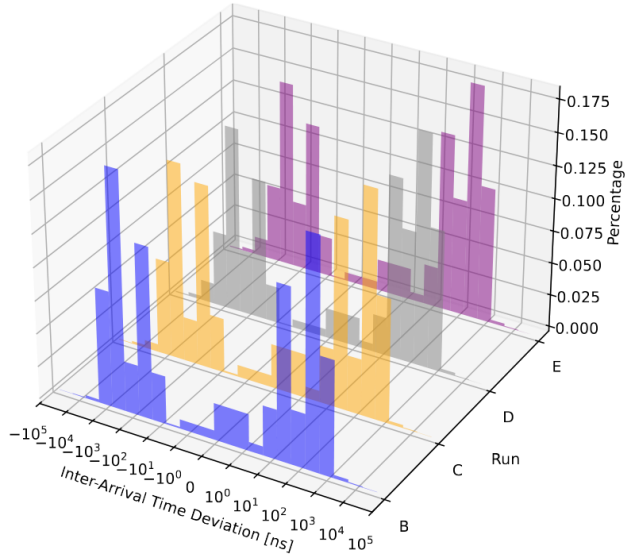
- Top: 80 Gbps, bottom: shared with co-tenant noise
- Still similar shared vs dedicated performance
- Noise observable



The percentage of packets with a given IAT delta using dedicated NICs.



The percentage of packets with a given IAT delta using shared NICs.



The percentage of packets with a given IAT delta.

ILLINOIS TECH
College of Computing

The Metrics

- Largely expected κ trends
- Non-IAT metrics potentially too small

Environment	U	O	I	L	κ
Local Single-Replayer	0	0	0.0294	4.27×10^{-6}	0.9853
Local Dual-Replayer	0	0.0259	0.2022	9.68×10^{-3}	0.9282
FABRIC Dedicated 40 Gbps 1	0	0	0.4996	3.07×10^{-5}	0.7426
FABRIC Shared 40 Gbps	0	0	0.0662	2.24×10^{-5}	0.9669
FABRIC Dedicated 40 Gbps 2	0	0	0.4998	4.20×10^{-4}	0.7502
FABRIC Dedicated 80 Gbps	0	0	0.1073	8.20×10^{-6}	0.9463
FABRIC Shared 80 Gbps	0	0	0.1105	2.26×10^{-5}	0.9448
FABRIC Ded. 80 Gbps Noisy	0	0	0.1085	1.37×10^{-5}	0.9458
FABRIC Shd. 40 Gbps Noisy	1.99×10^{-4}	0	0.5024	2.04×10^{-5}	0.7488

ILLINOIS TECH
College of Computing

Conclusion

- Built **Choir**, middlebox replayer for virtual networks
  - 100 Gbps, consistent, general
- Designed metric κ to quantize network consistency
  - Can concisely convey consistency, start a discussion on such measurement
  - Future work: improvement in sub-metric scaling (IAT dominates)
  - Future work: establishing a baseline and monitoring for notable divergence
- **Acknowledgement:** Mert Cevik and Komal Thareja (RENCI) for FABRIC assistance; Mami Hayashida, Hussamuddin Nasir, and Jim Griffoen (U. of Kentucky) for FABRIC PTP assistance; Nishanth Shyamkumar (IL Tech) for DPDK assistance

ILLINOIS TECH
College of Computing