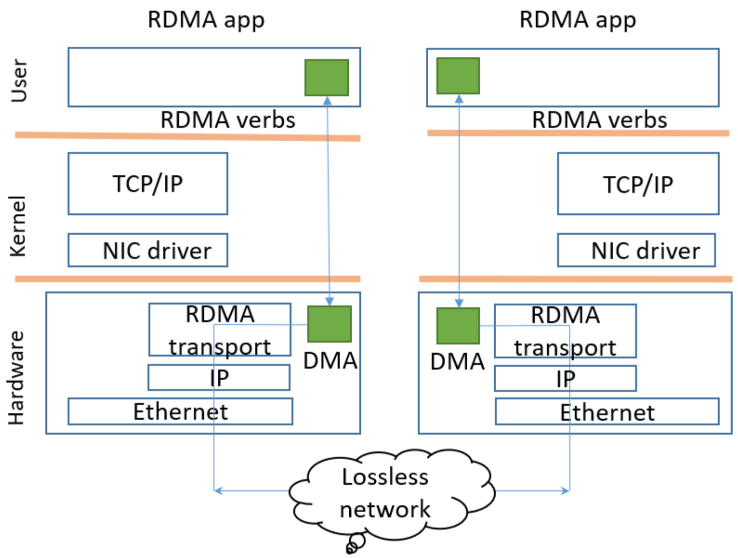
Rethinking Intra-host Congestion Control in RDMA Networks

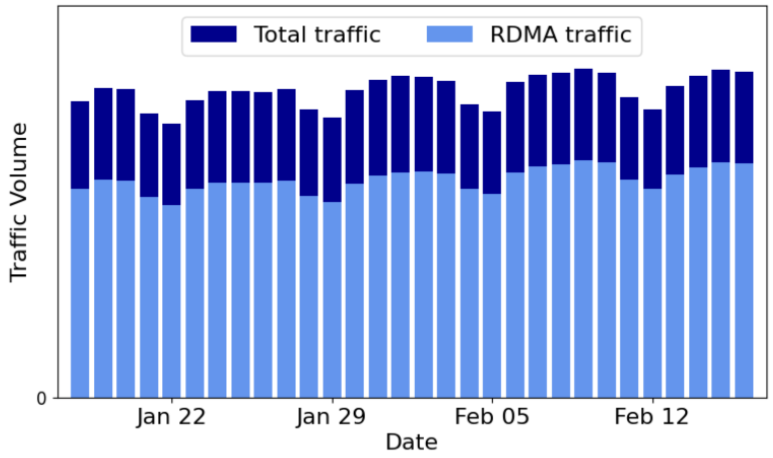
Zirui Wan, Jiao Zhang, Yuxiang Wang, Kefei Liu,
Haoyu Pan, Tao Huang

Beijing University of Posts and Telecommunications

Background

- **RDMA becomes the de-facto standard for high-speed networks in modern datacenters**





[1] Wei Bai, Shanim Sainul Abdeen, Ankit Agrawal et al. Empowering Azure Storage with RDMA. NSDI 2023

- RDMA achieves high performance using kernel-bypass and transport offload

- Wide adoption of RDMA
  - Around **70% traffic in Azure is RDMA**

Background

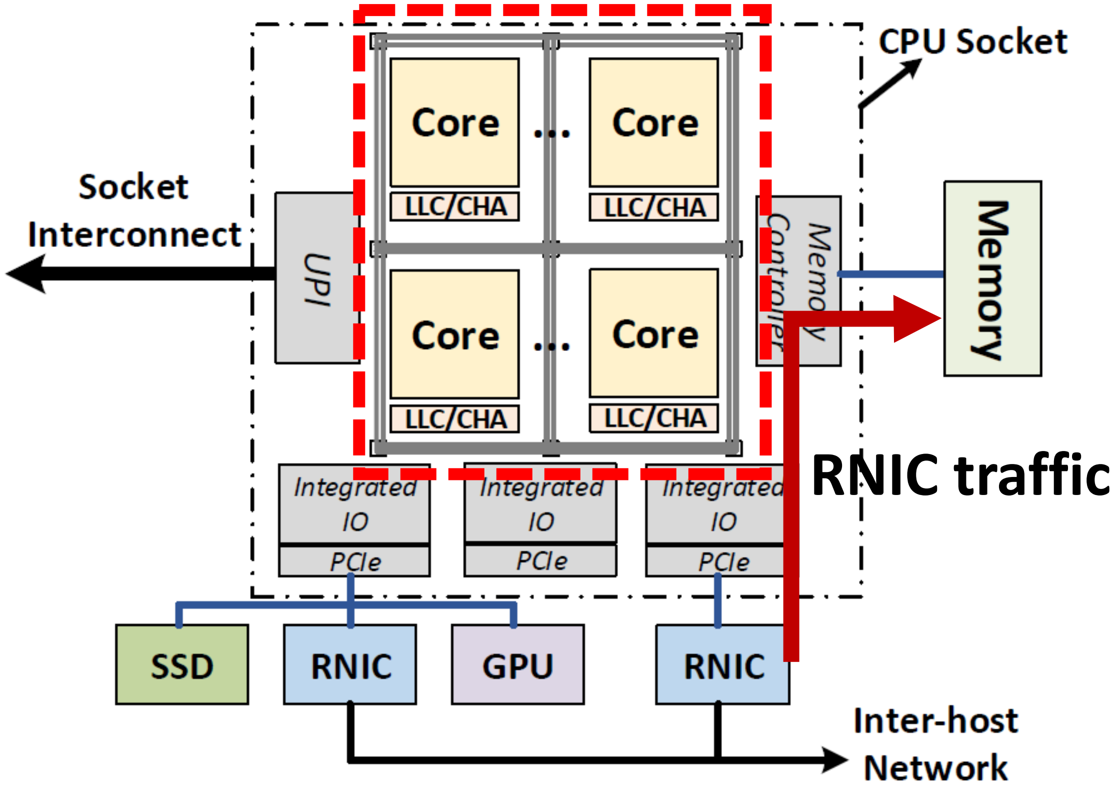


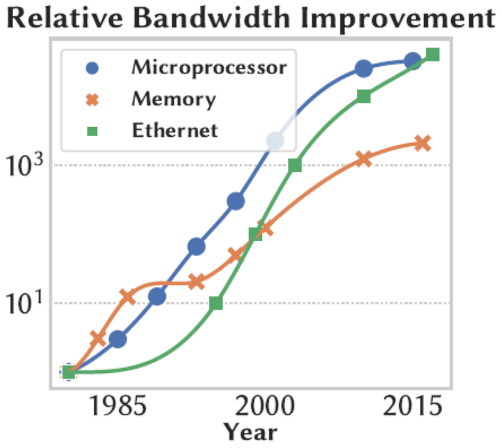
Illustration of intra-host network

- **RNIC traffic data path**
  - RNIC --> IIO Stack --> Memory Controller -> Memory
- **Ideally, RNIC traffic is guaranteed by intra-host network**
  - Sufficient bandwidth and high performance by **Mesh architecture**
  - **Lossless interconnect** fabric by credit-based flow control scheme

Background

■ Evolution of intra-host network

DGX A100 Architecture



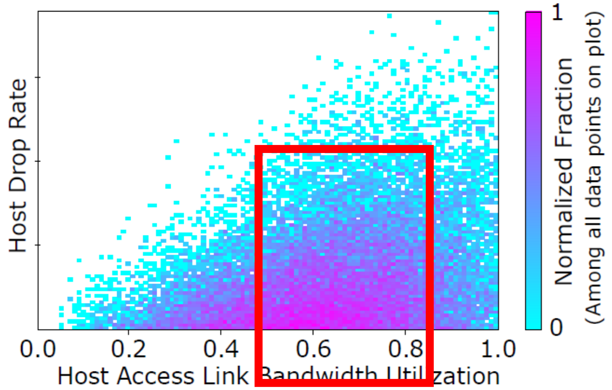
[1] Wang, M., Xu, M., & Wu, J. (2022). Understanding I/O direct cache access performance for end host networking. SIGMETRICs

- **More complicated intra-host network**
  - ➢ e.g., NVIDIA DGX can be equipped with up to 8 RNICs and 8 GPUs.
- **Stagnant technology with intra-host network**
  - ➢ e.g., RNIC from 25Gbps to 400Gbps
    Memory BW from 10GBps to 55GBps

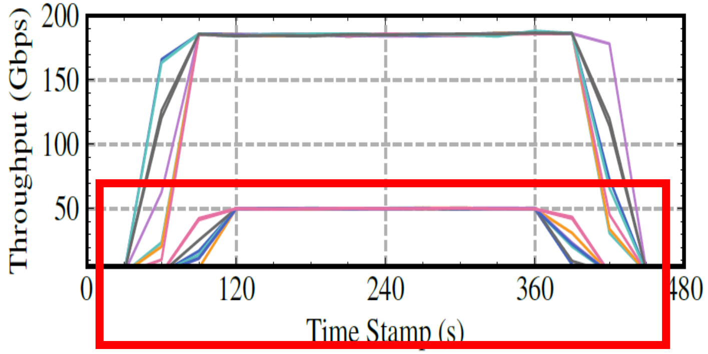
RNIC traffic may not get sufficient intra-host resources

Motivation

- **Large-scale production datacenter operators demonstrate**
  - **RNIC traffic suffers intra-host congestion**



- Drop rate increasement in Google



- Throughput degratation in Bytedance

Total bandwidth	TCP bandwidth ratio	TX pauses
25Gbps	40%	0
30Gbps	45%	1Kpps
32Gbps	50%	8Kpps
35Gbps	46%	15Kpps

Table 2: TX pauses in hybrid RDMA/TCP traffic.

- Tx pauses generation in Alibaba

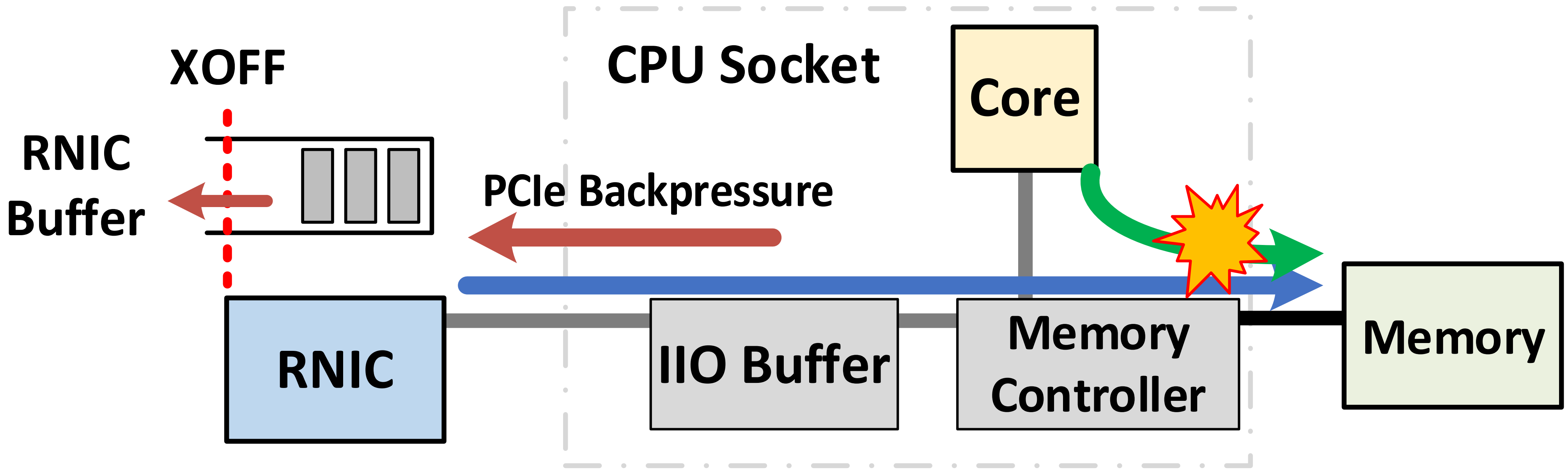
[1] Agarwal, Saksham, et al. "Understanding host interconnect congestion." HotNets, 2022.

[2] Liu, Kefei, et al. "Hostping: Diagnosing intra-host network bottlenecks in RDMA servers." NSDI 23. 2023.

[3] Gao, Yixiao, et al. "When cloud storage meets RDMA." NSDI 21. 2021.

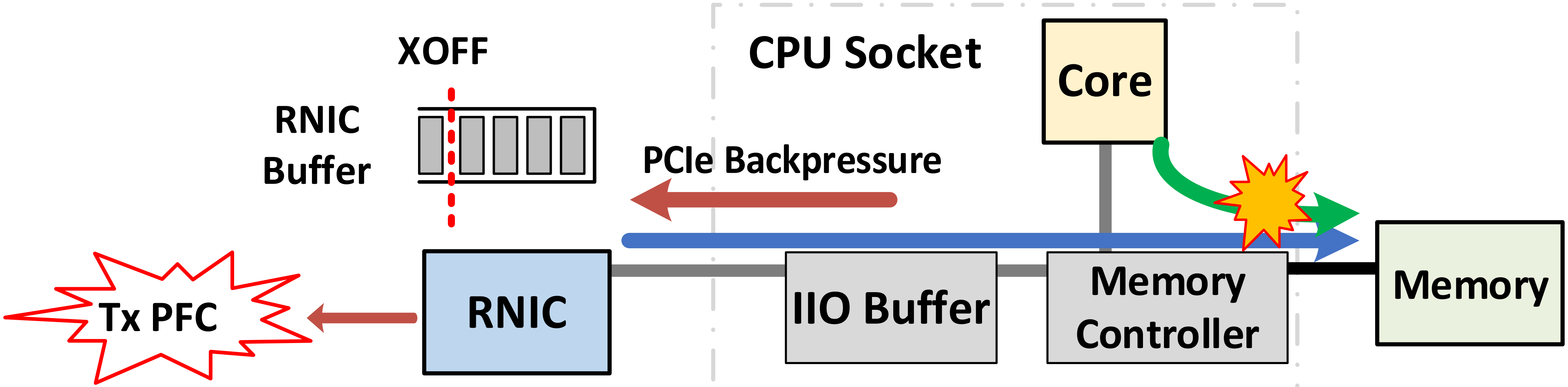
Motivation

- Illustration of RDMA intra-host congestion



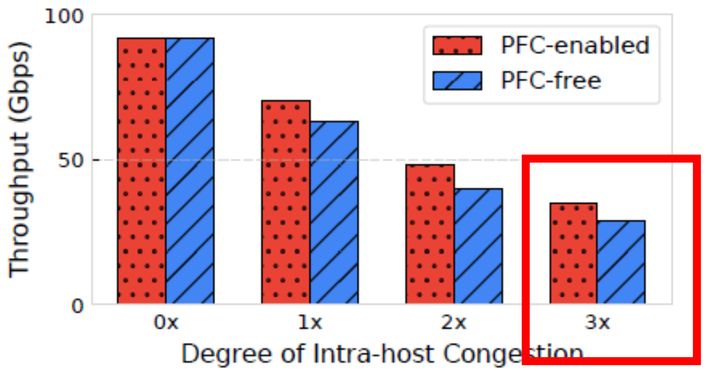
Motivation

- Illustration of RDMA intra-host congestion

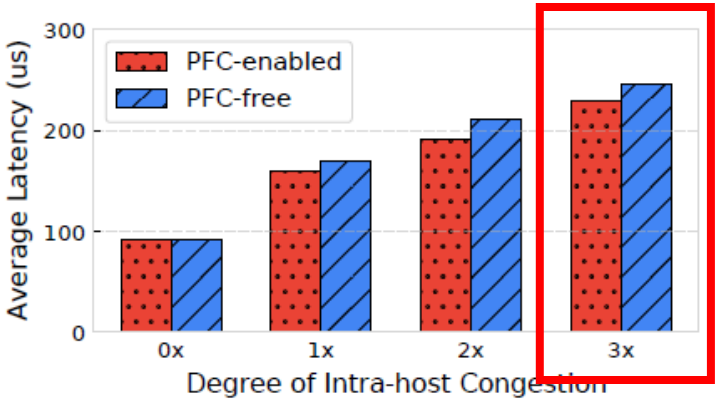


Motivation

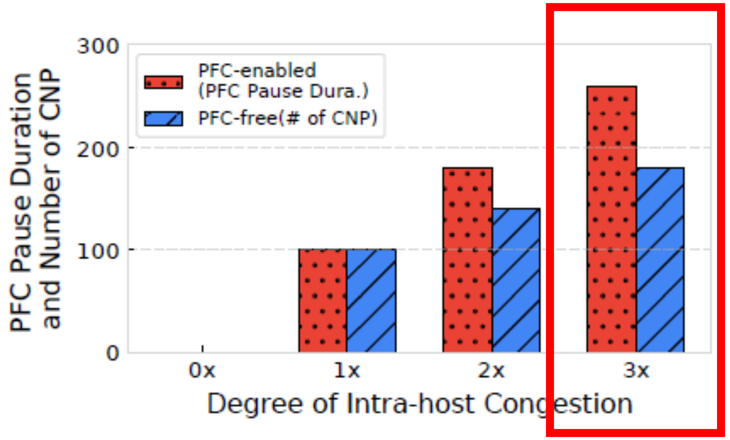
- Understanding impacts of RDMA intra-host congestion



Throughput



Latency



PFC pauses

RDMA intra-host congestion leads to performance loss

- ***68% throughput decreases***
- ***2.6X latency increases***
- ***PFC pauses***

Motivation

- Understanding impacts of RDMA intra-host congestion

We desire to design a new RDMA intra-Host Congestion Control mechanism

Throughput

Latency

PFC pauses

RDMA intra-host congestion leads to

- ***68% throughput decreases***
- ***2.6X latency increases***
- ***PFC pauses***

Motivation

■ **Challenges of RDMA intra-host congestion control**

Inter-host Network

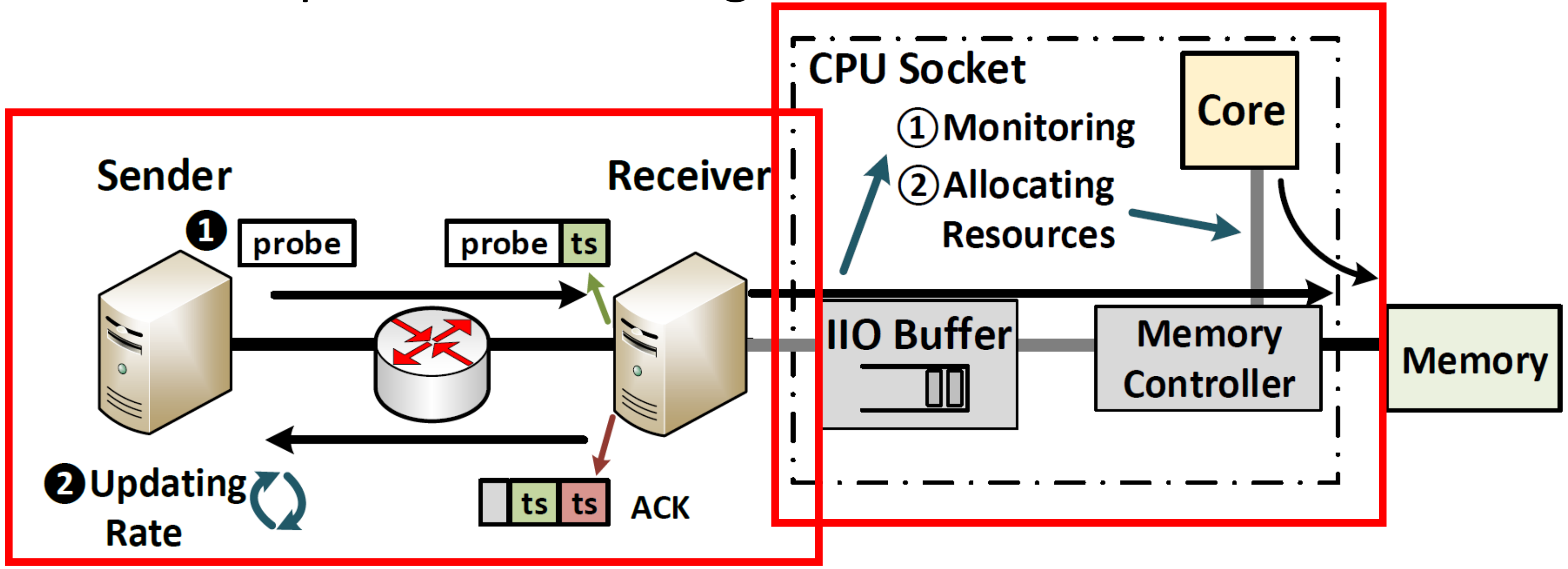
- Inter-host is mature with various CC signals and resource allocation mechanisms
- The receiver only receives packets and generates ACKs.

Intra-host Network

- Intra-host is **complicated and lossless**.
  - ➢ traditional CC signals fail
- Intra-host resource allocation tool is naïve.
- **Commercial RNICs do not provide interfaces for kernel to modify packets. (Different from TCP traffic)**
  - ➢ hostCC marks ECN-bit which is hard to deploy in RDMA networks.

RHCC Design

- **RDMA intra-Host Congestion Control (RHCC)**
  - The intra-host traffic monitors **IIO buffer occupancy** and allocate resources by multi-levels. (similar with hostCC)
  - The RNIC traffic uses **probe mechanism** (using NVIDIA PCC framework) to detect congestion and update the sending rate.



RHCC Design

■ Rationales of RHCC

Challenges

- Intricate intra-host network
- Simple intra-host resource allocation tool
- RDMA offloads transport protocol

Rationales

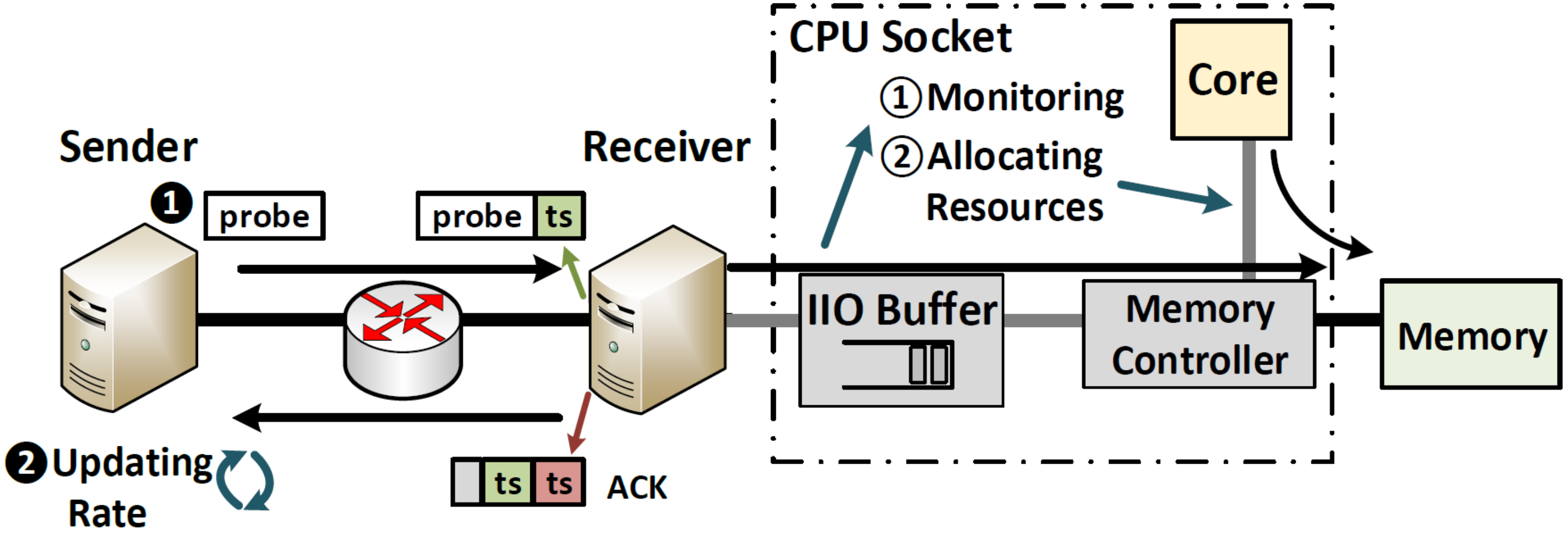
- IIO signal indicate **precise congestion information**
- **Simple** multi-level memory bandwidth allocation
- **Deployable RNIC traffic probe mechanism**

RHCC Design

Intra-host traffic congestion response

- The kernel monitors IIO buffer occupancy, I_{cur}, and compare it with threshold, I_{thr}, where $I_{cur} > I_{cur}$ means intra-host congestion happens.

- After congestion happens, the kernel allocates available memory bandwidth for intra-host traffic with multi-level.

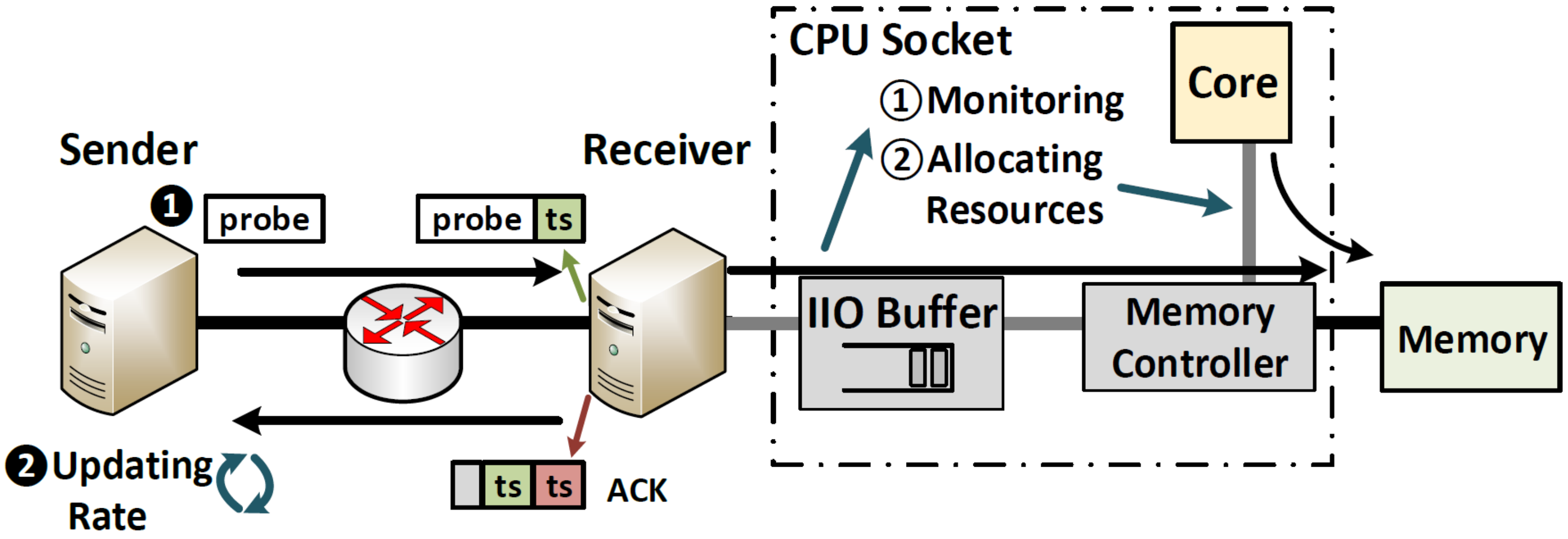


RHCC Design

■ RNIC traffic congestion response

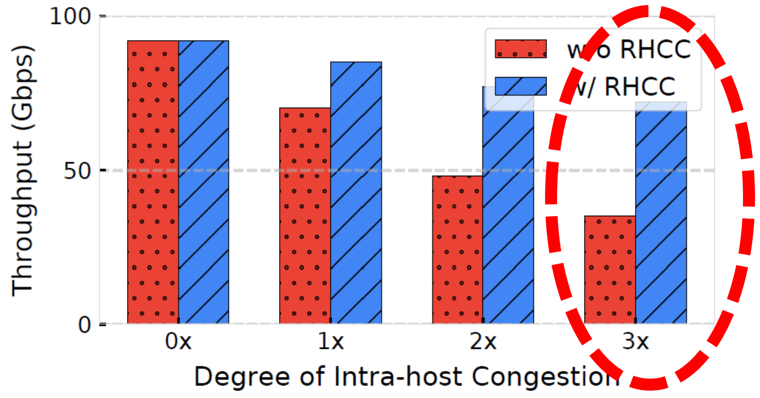
- The sender periodically transmits probe packets (using NVIDIA PCC framework) to detect receiver processing delay, D_{cur}, and updates the sending rate using

$$R = R\times[1 - \alpha\times(D_{cur} - D_{thr}) - \beta\times(D_{cur} - D_{old})]$$

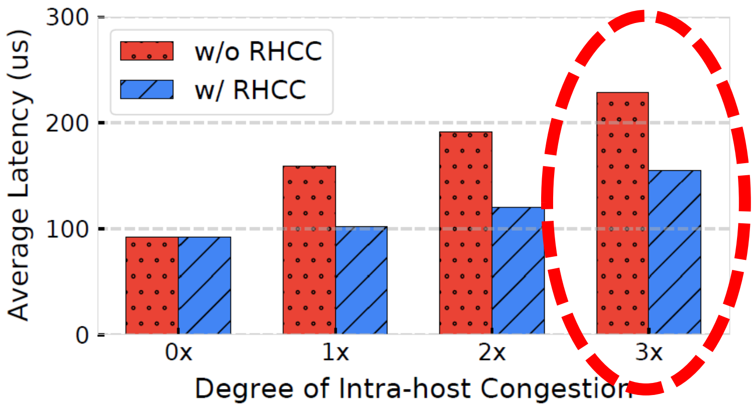


Evaluation

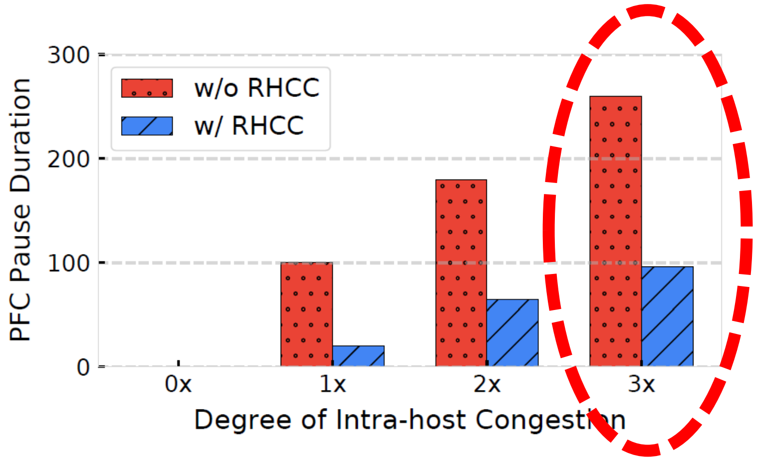
■ Testbed results



Throughput



Latency



PFC pauses

RHCC improves performance even with high degree of congestion

- ***2X throughput increases***
- ***1.4X latency decreases***
- ***2.7X PFC pauses decreases***

Discussion and Future Work

New intra-host architecture

- CXL enables high performance connectivity between CPU and CXL devices by maintaining a unified, coherent memory space.
  - Its benefits on intra-host congestion merit future research.

New programmable intra-host network

- It is difficult to implement complex intra-host adjustment functions.
- Commercial RNIC only provides simple counters.
  - Using diagnostic counters is limited

Conclusion

- We **analyze the requirements to design a new RDMA intra-host congestion control mechanism**.
- We present **RHCC, a novel RDMA intra-Host Congestion Control solution** that combining intra-host traffic congestion avoidance and proactive RNIC traffic adjustment.
  - ✓ ***RHCC reduces PFC pauses and strengthens the performance of intra-host network.***
- We hope RHCC can inspire congestion control in RDMA intra-host networks.

Thank you!

Zirui Wan, Jiao Zhang, Yuxiang Wang, Kefei Liu,
Haoyu Pan, Tao Huang

北京郵電大學
Beijing University of Posts and Telecommunications