SENIC: Scalable NIC for End-Host Rate Limiting

Sivasankar Radhakrishnan

Yilong Geng, Vimalkumar Jeyakumar,

Abdul Kabbani, George Porter, Amin Vahdat

UCSDCSE Computer Science and Engineering

STANFORD COMPUTER SCIENCE

Google

Consolidation of Servers

reddit

Instagram

NETFLIX

Network resource management and allocation is crucial

Network Resource Allocation

- Performance isolation: Oktopus, Seawall, EyeQ
- Congestion control: QCN, RCP, D3, DCTCP, HULL

Rely on programmable rate limiters

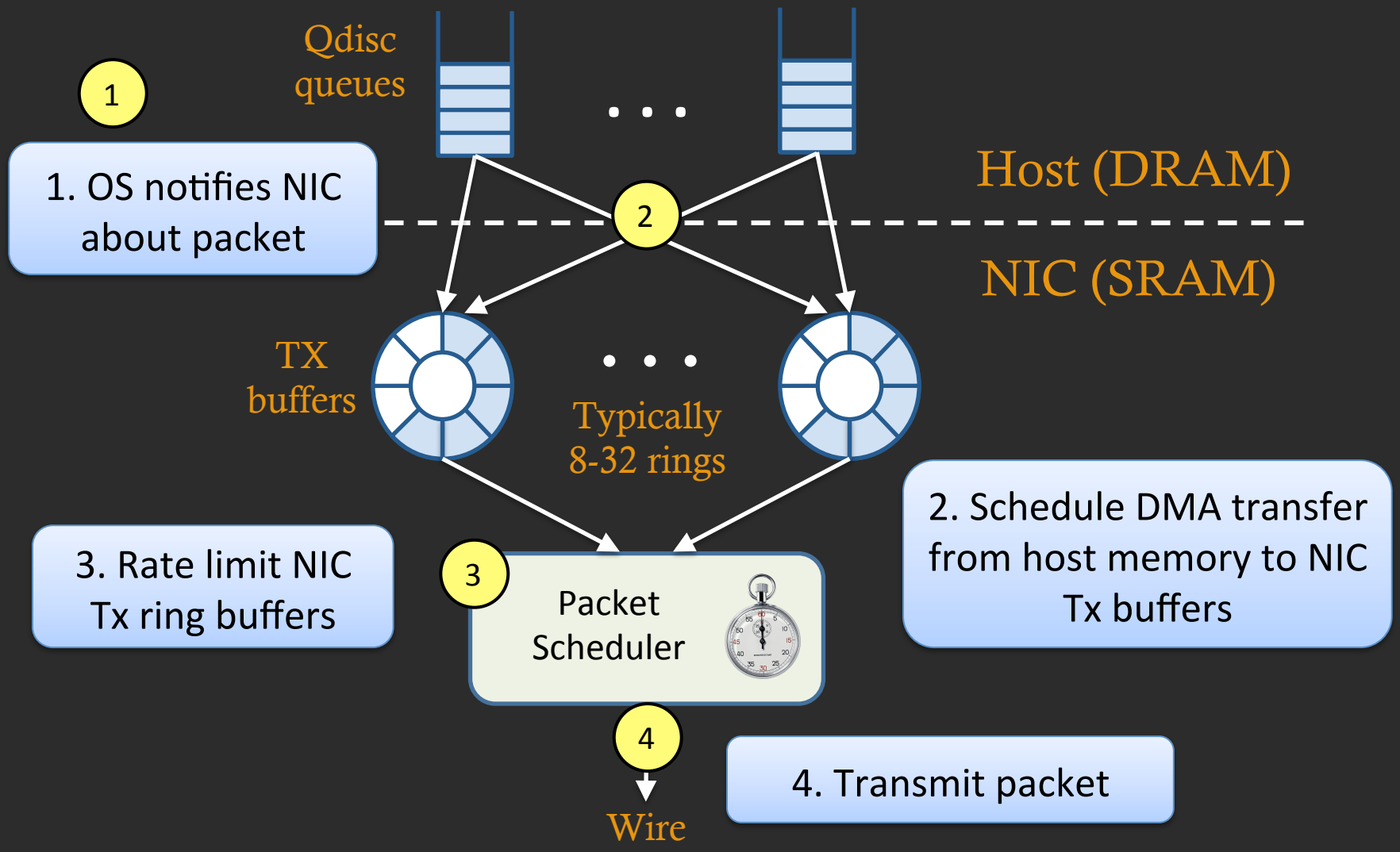
Scalable rate limiting is required
Thousands of rate limiters per server

Rate Limiter Options

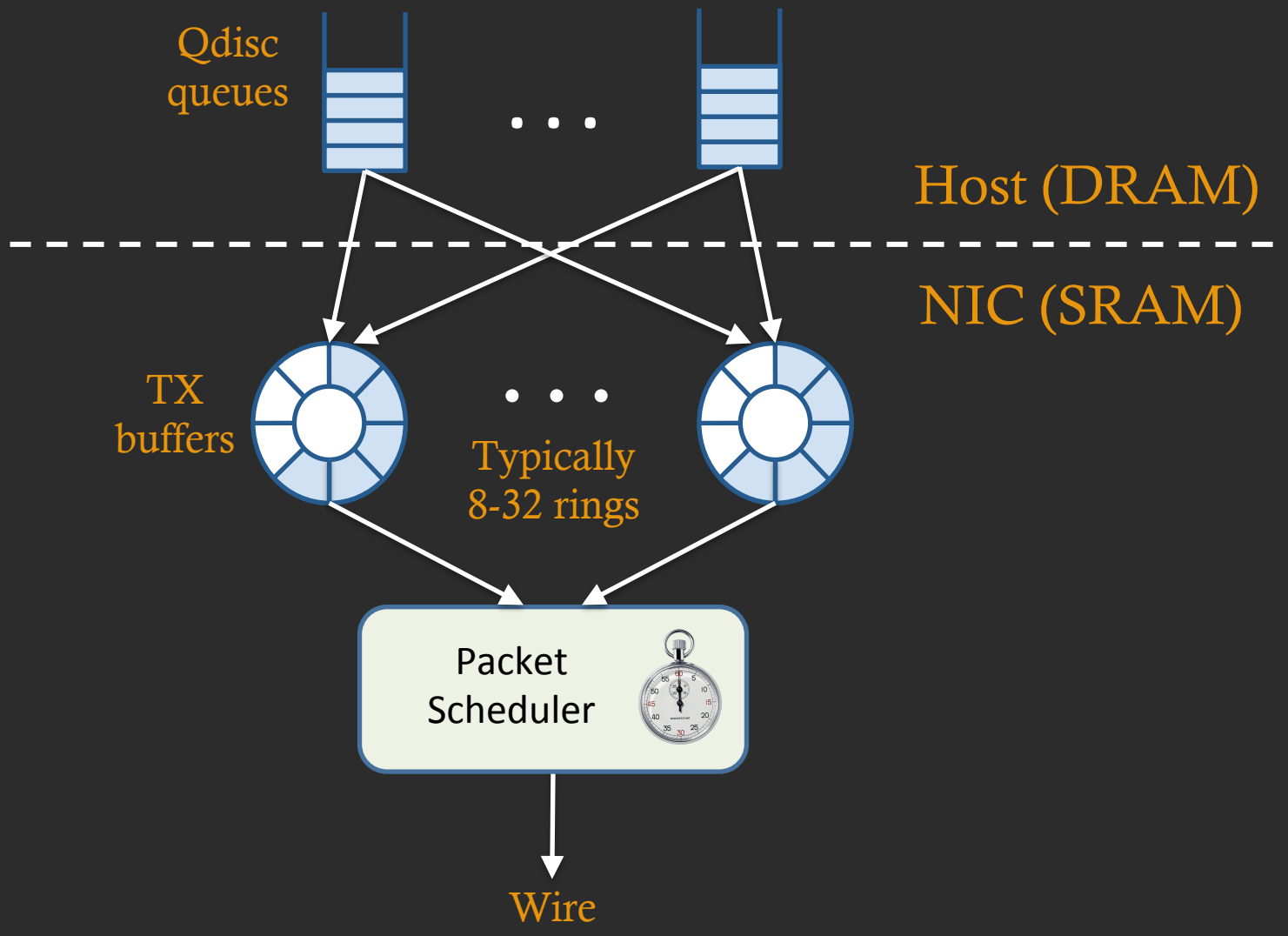
	Software	Hardware	SENIC
Scales to many classes	✔	✖	✔
Works at high link speeds	✖	✔	✔
Low CPU overhead	✖	✔	✔
Accurate and precise	✖	✔	✔
Supports hypervisor bypass	✖	✔	✔

Reorganize responsibilities of the NIC and operating system

Current NIC Design

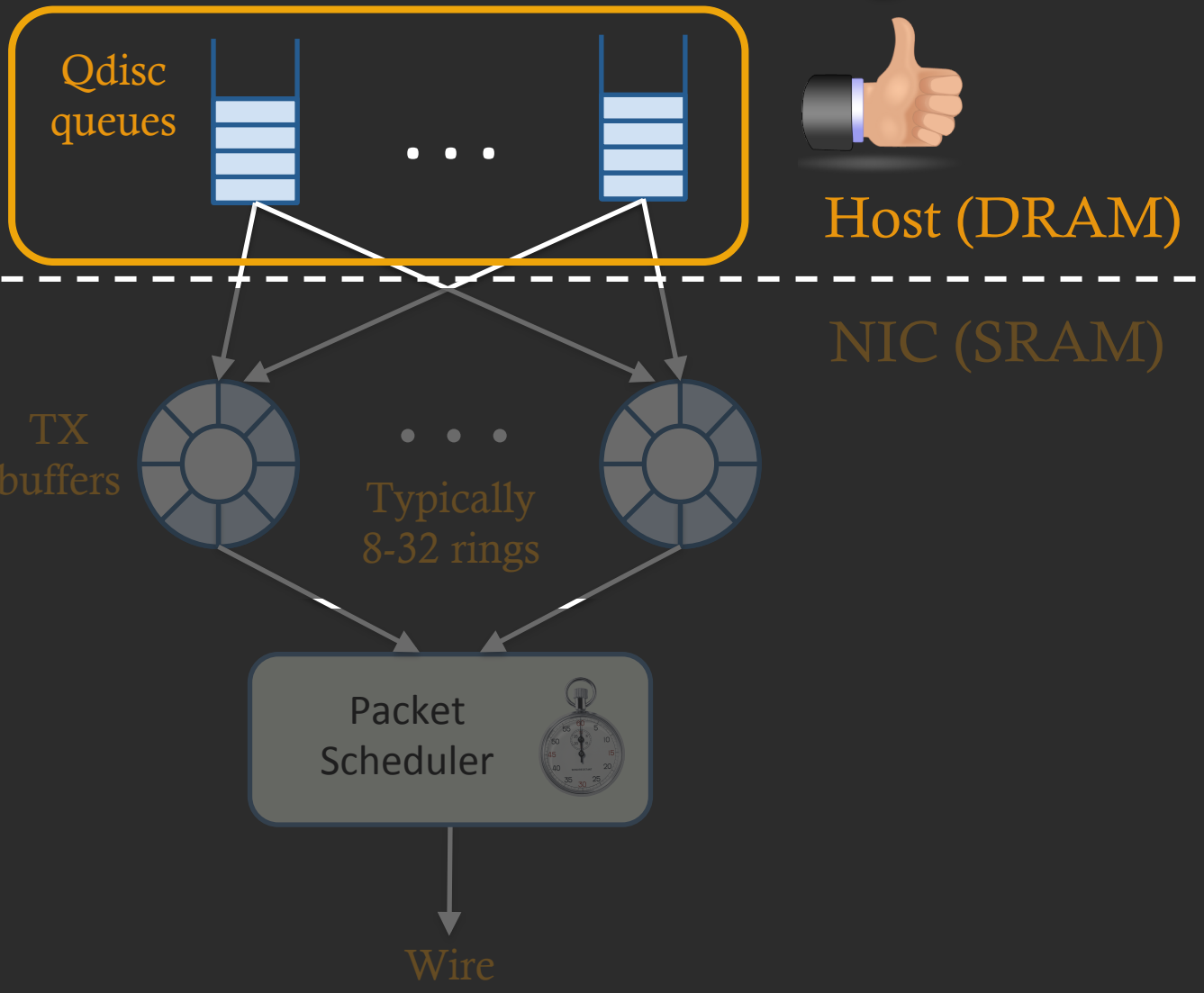


Current NIC Design

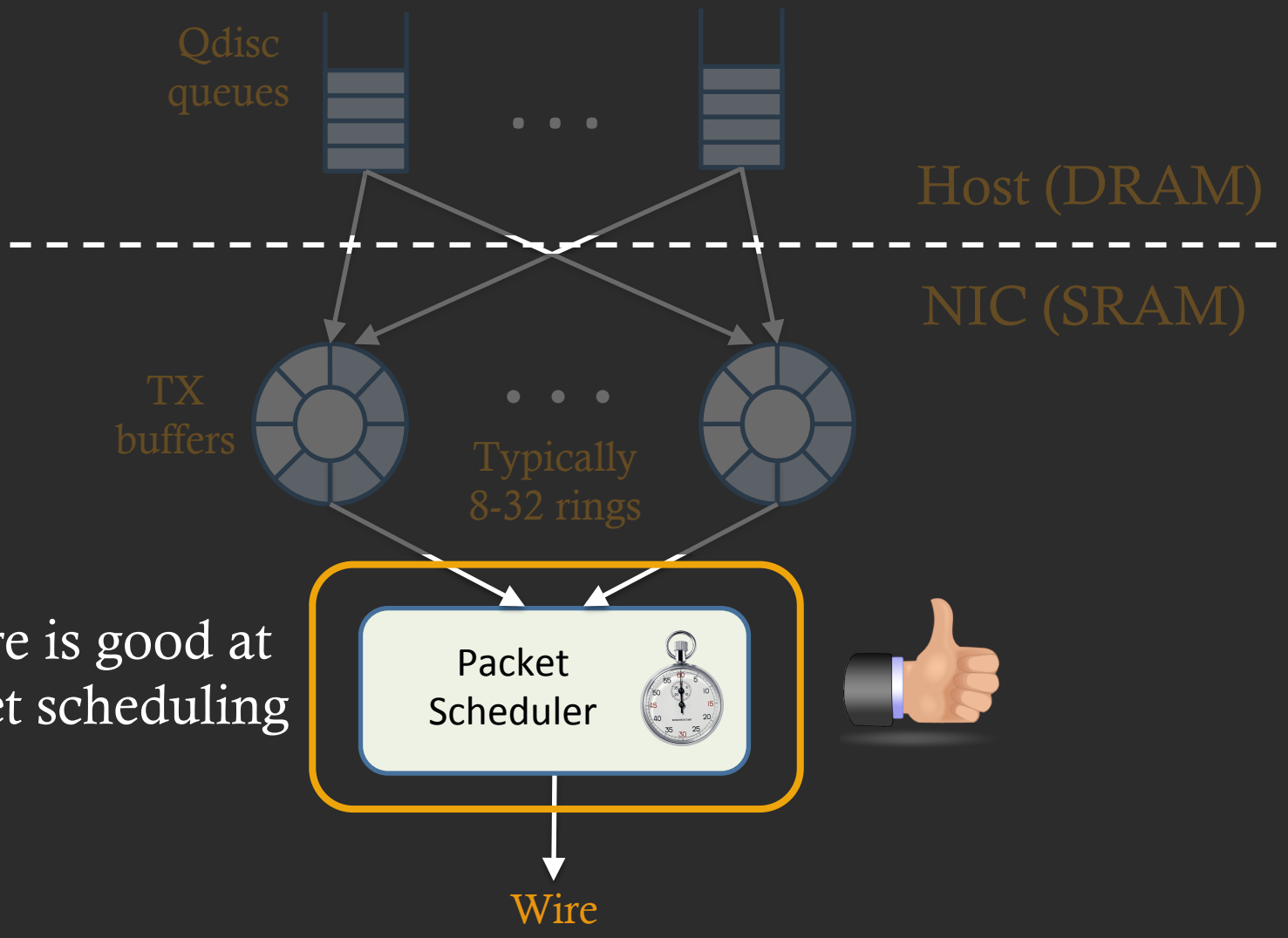


Current NIC Design

Host DRAM is cheap and abundant

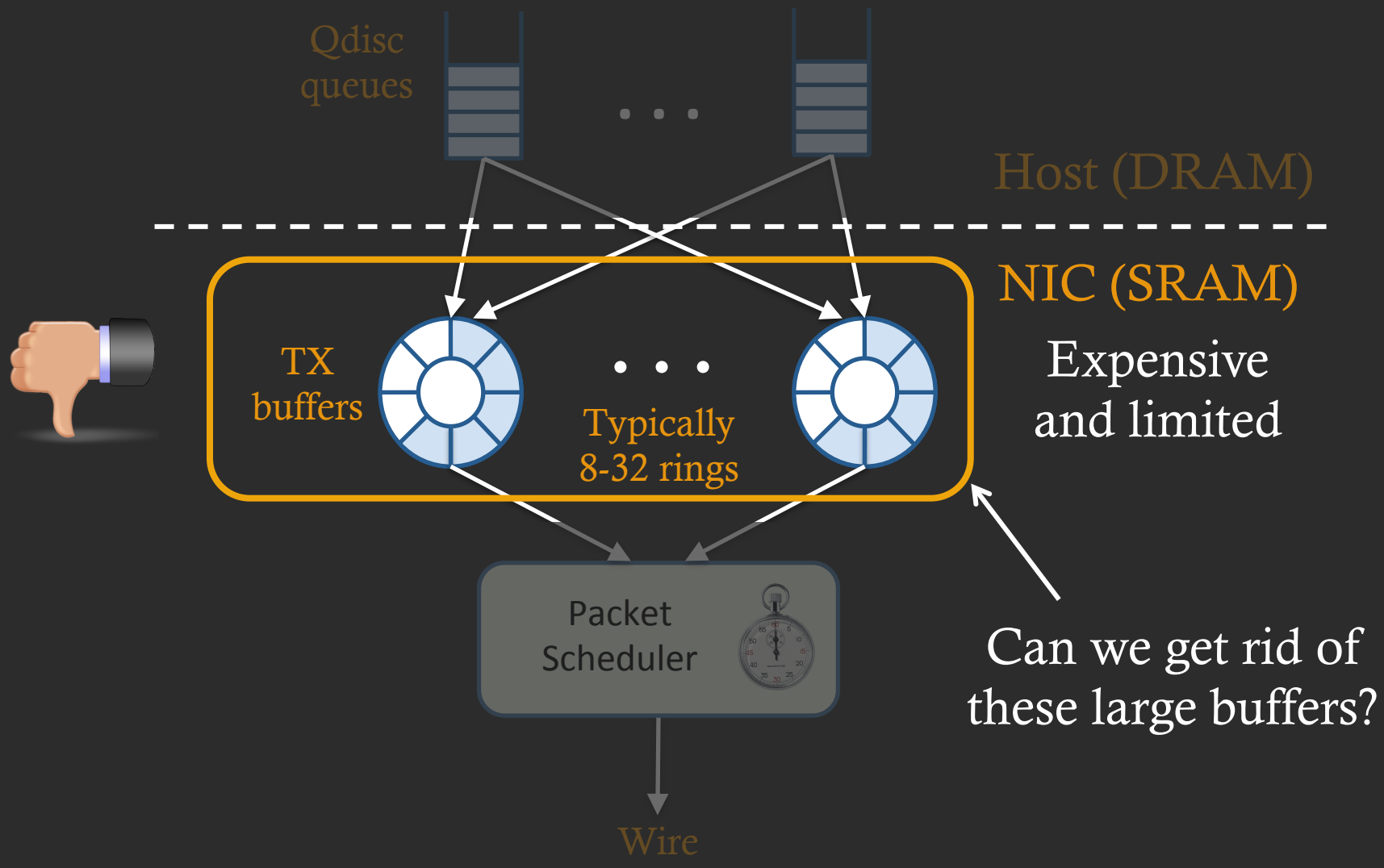


Current NIC Design

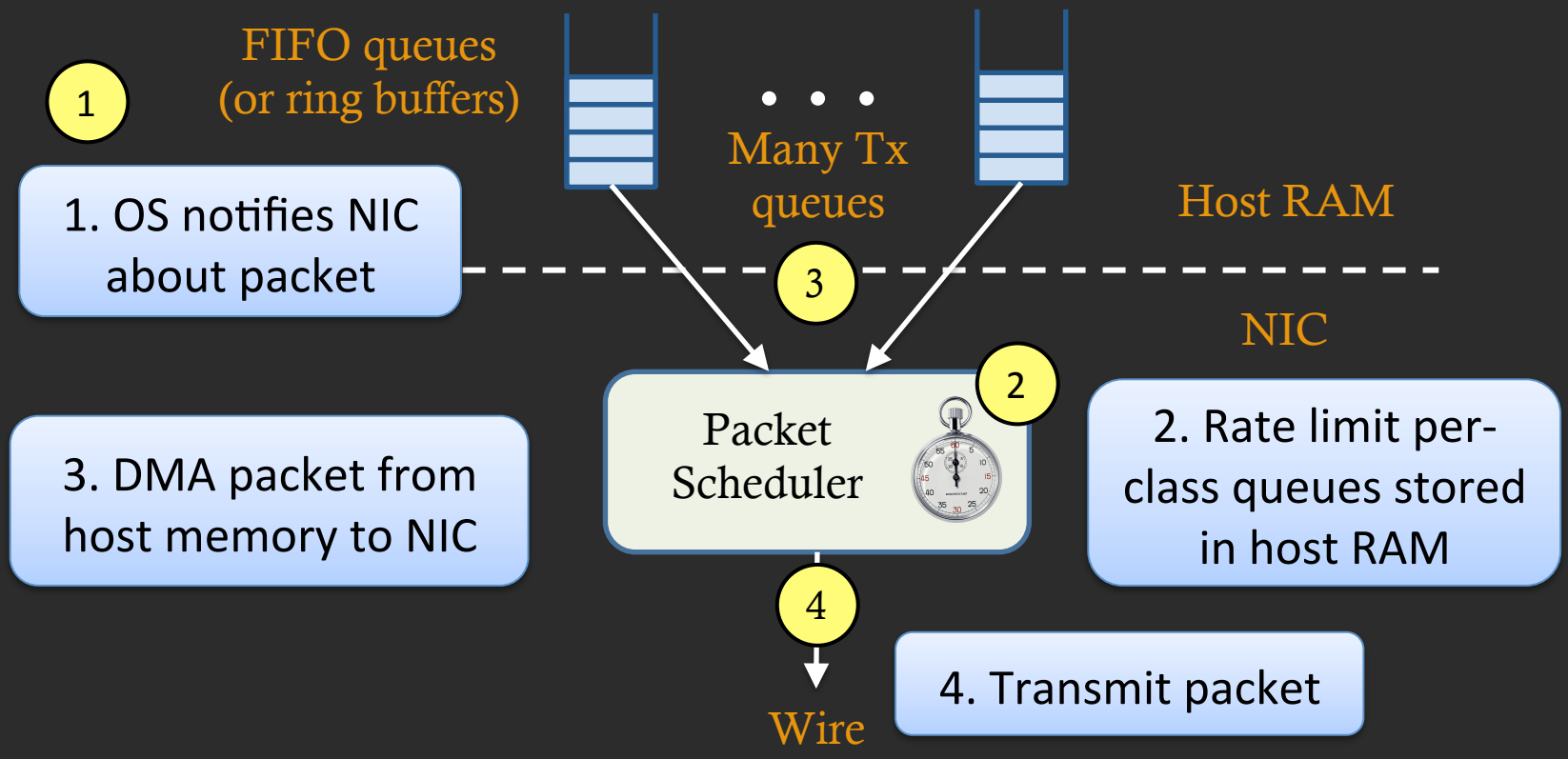


Hardware is good at per-packet scheduling

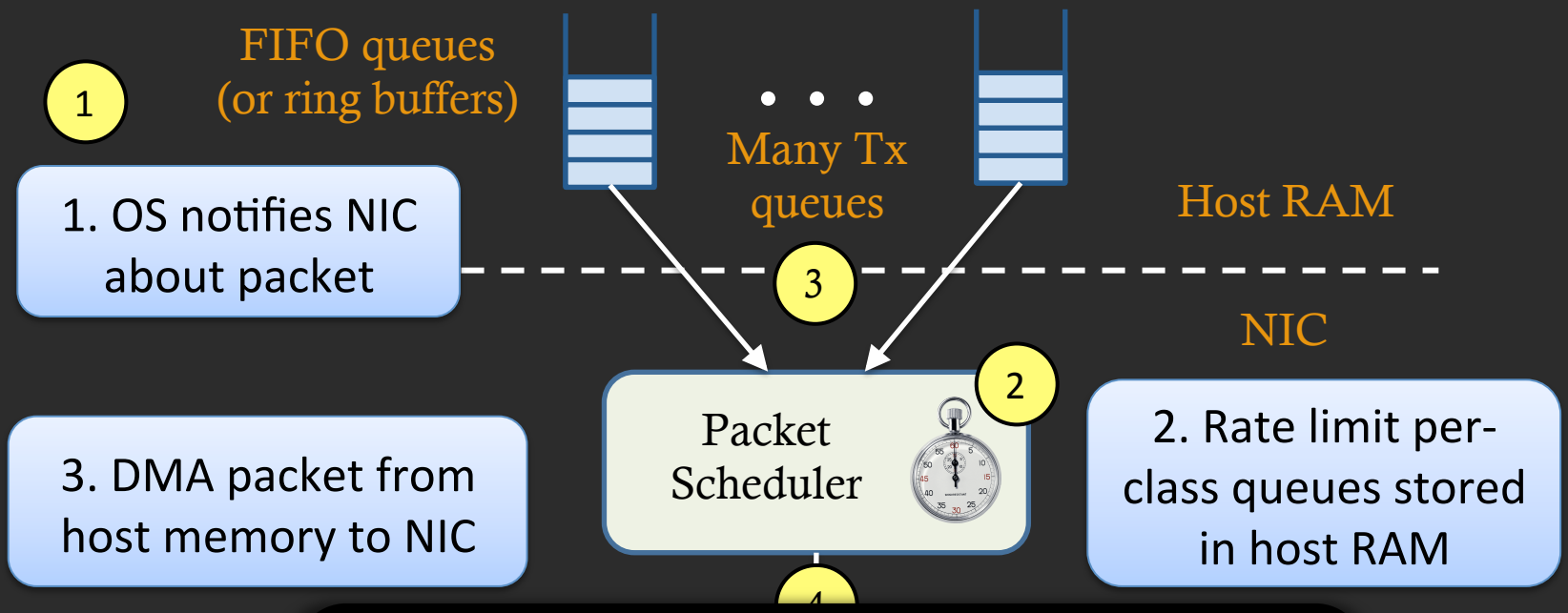
Current NIC Design



SENIC Design

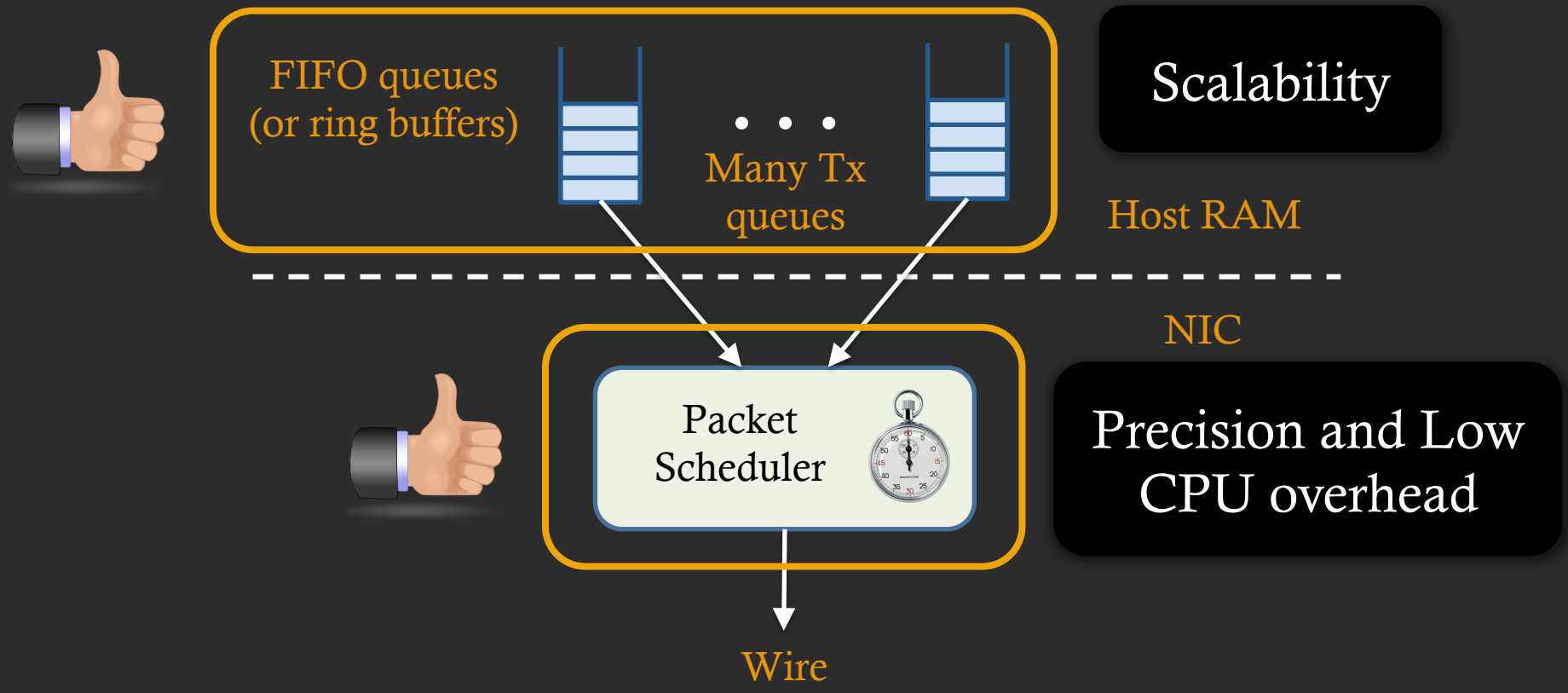


SENIC Design



Late binding of packet transfers to NIC

SENIC Design



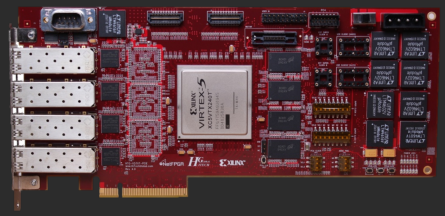
Scalability

Precision and Low CPU overhead

CPU handles control plane operations

(Configuring queues, rate limits, packet classification)

SENIC Prototypes



- NetFPGA 10G hardware prototype
  - Demonstrates feasibility
  - Implements simple token bucket scheduler
  - Late binding of DMA transfers from host memory

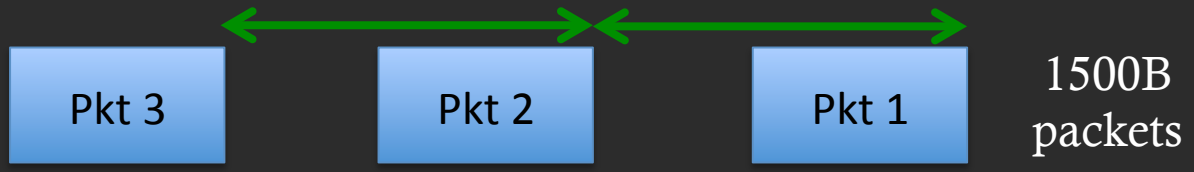
- Software prototype
  - Dedicated CPU core for network scheduling
  - Works with any existing NIC

NetFPGA 10G Microbenchmarks

- Synthesized at 100MHz with 1000 rate limiters

Is it Accurate?

- Synthesized at 100MHz with 1000 rate limiters
- Inter-packet delay for a traffic class

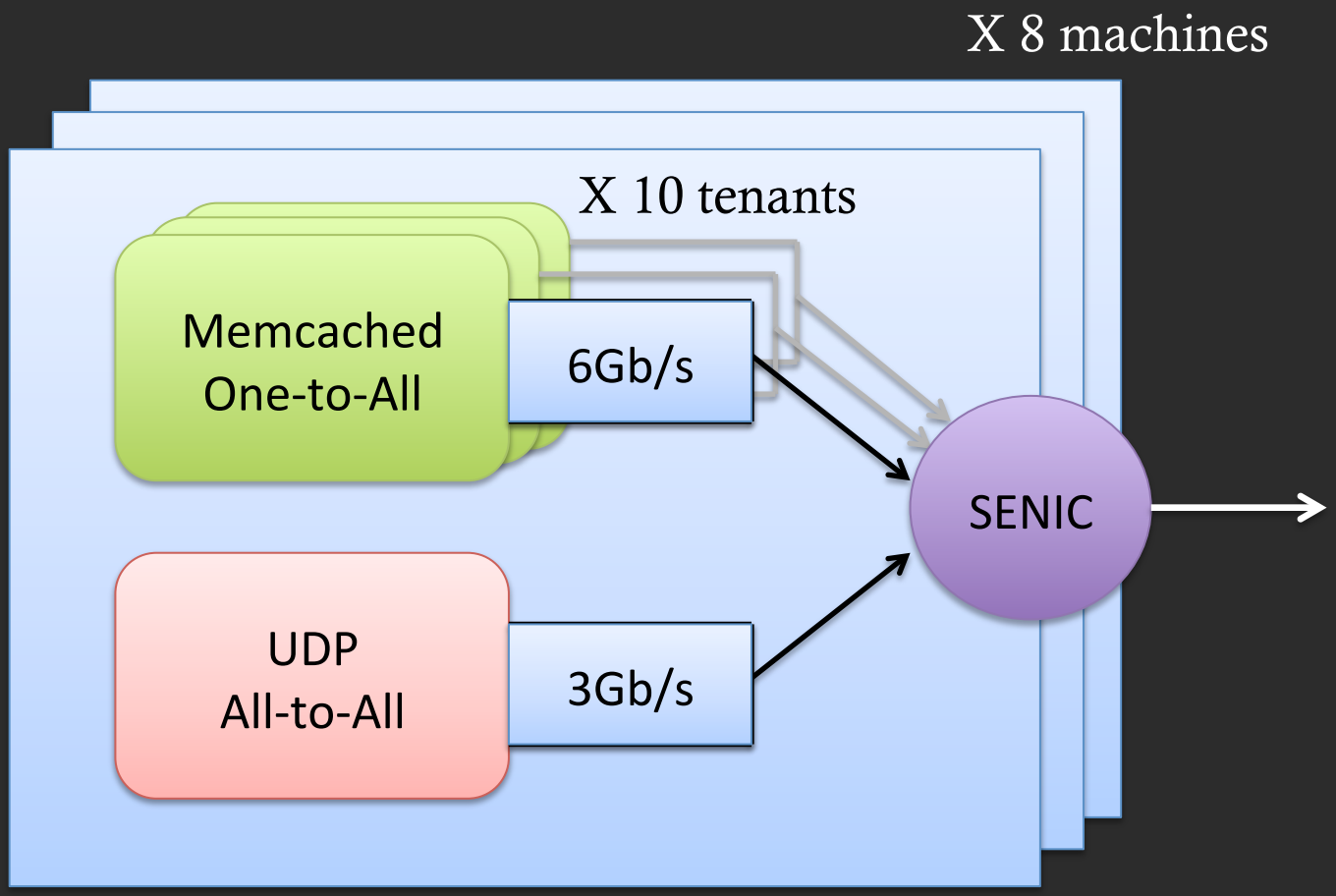


- Average: within 0.038% of ideal pacer delay
- Standard deviation: 1.7% of inter-packet delay

Is it Fast?

- Scheduling decision latency:
  - 5 SRAM lookups (50 ns)
- 1500B packet at 40Gb/s: 300ns budget
- Smaller packets: schedule a burst at a time

Macrobenchmark: Tenant Isolation

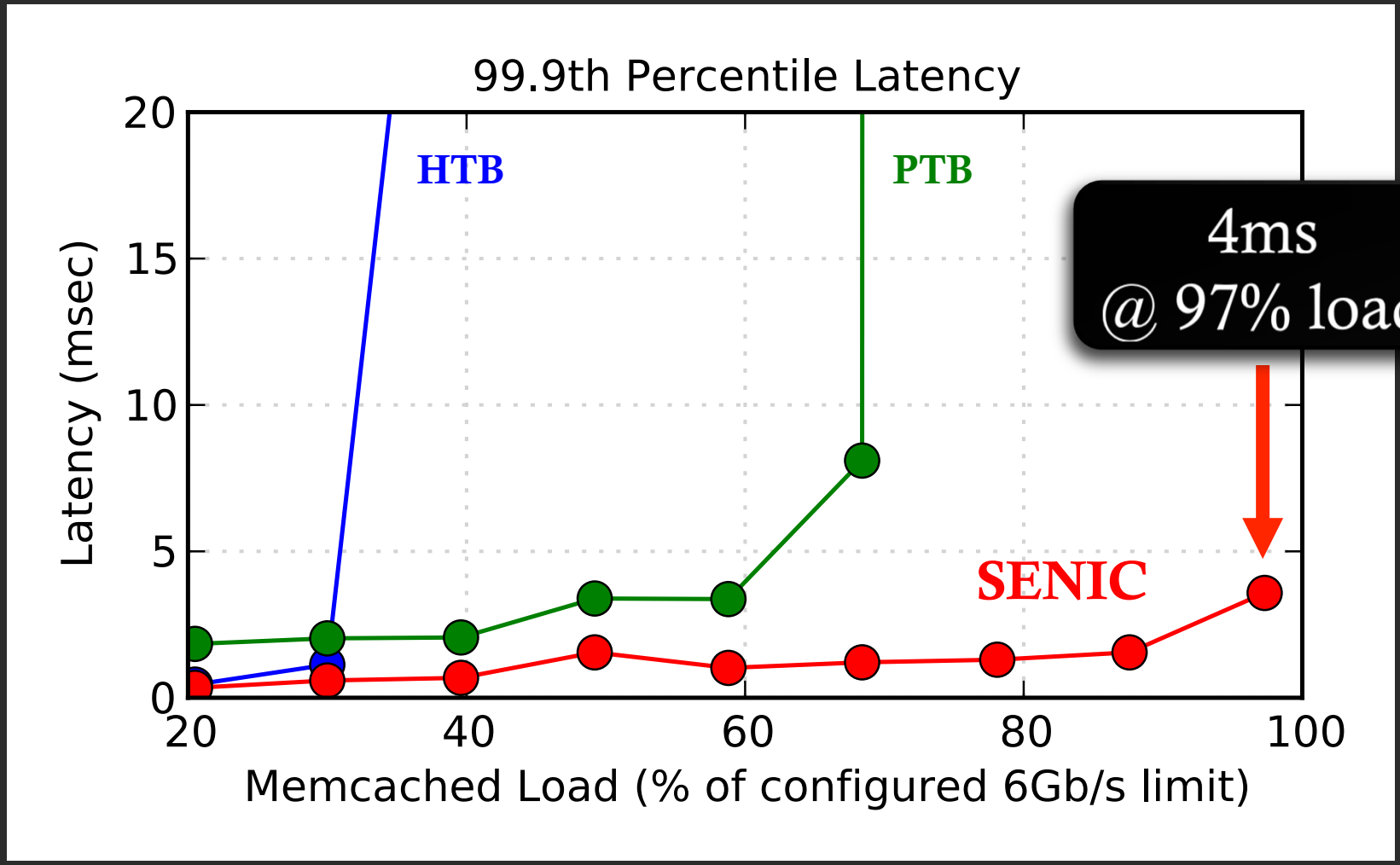


Macrobenchmark: Tenant Isolation

- Metrics:
  1. Memcached tail latency
  2. UDP throughput
- Compare SENIC to:
  1. Hierarchical Token Buckets (HTB)
  2. Parallel Token Buckets (PTB)
- Varying memcached tenant load

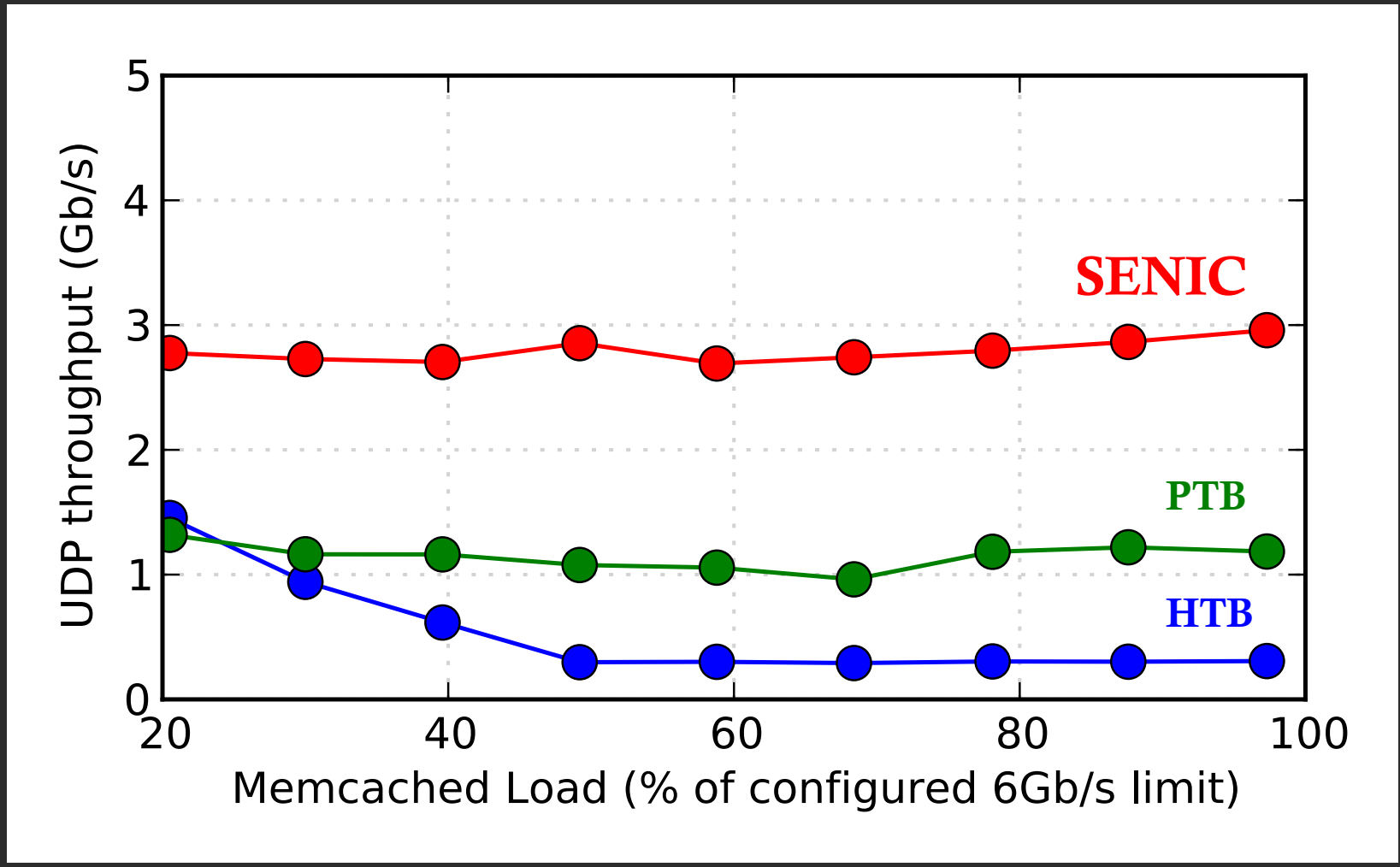
Memcached Tail Latency

(Lower is better)



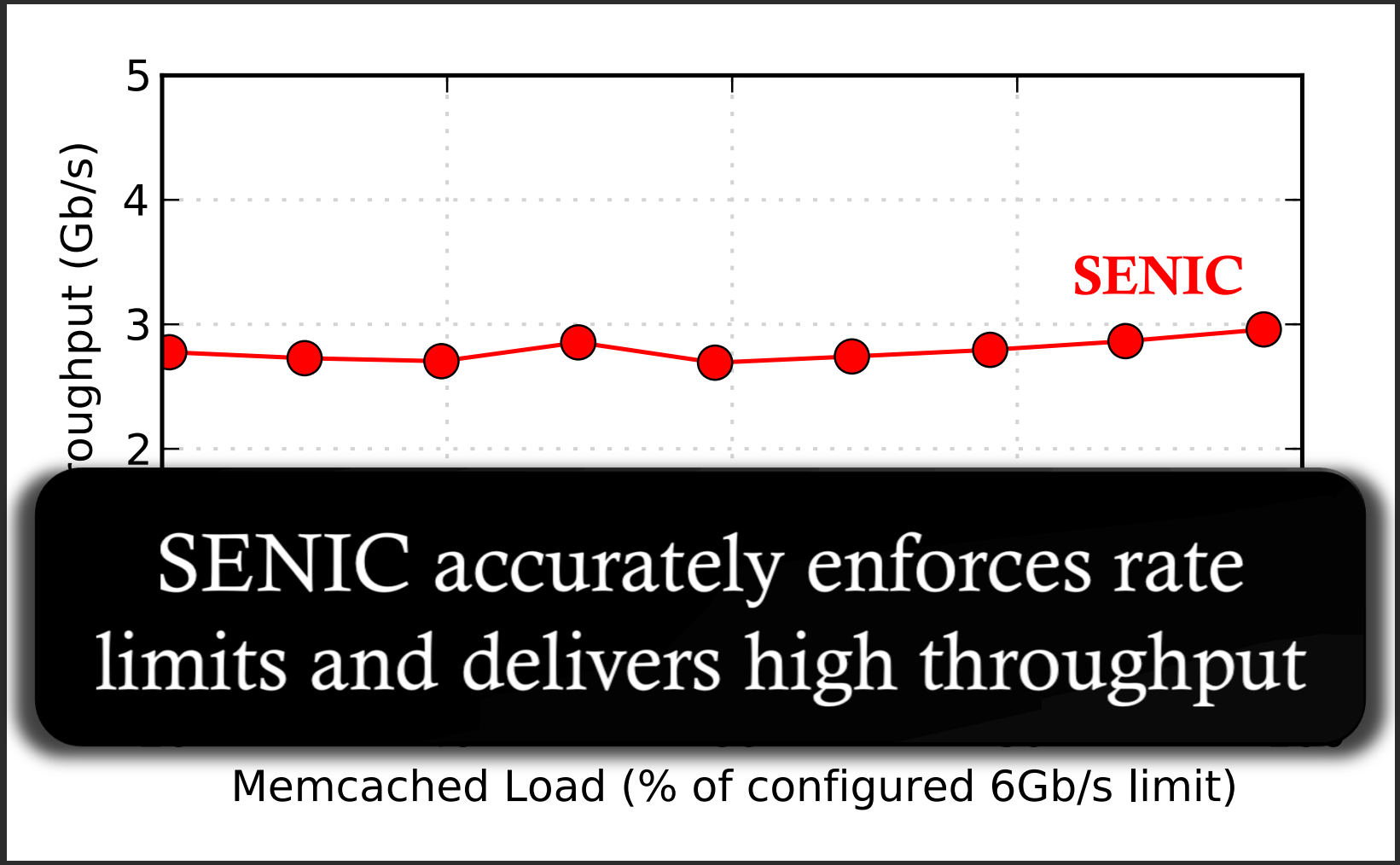
UDP Tenant Throughput

(Closer to 3Gb/s configured limit is better)



UDP Tenant Throughput

(Closer to 3Gb/s configured limit is better)

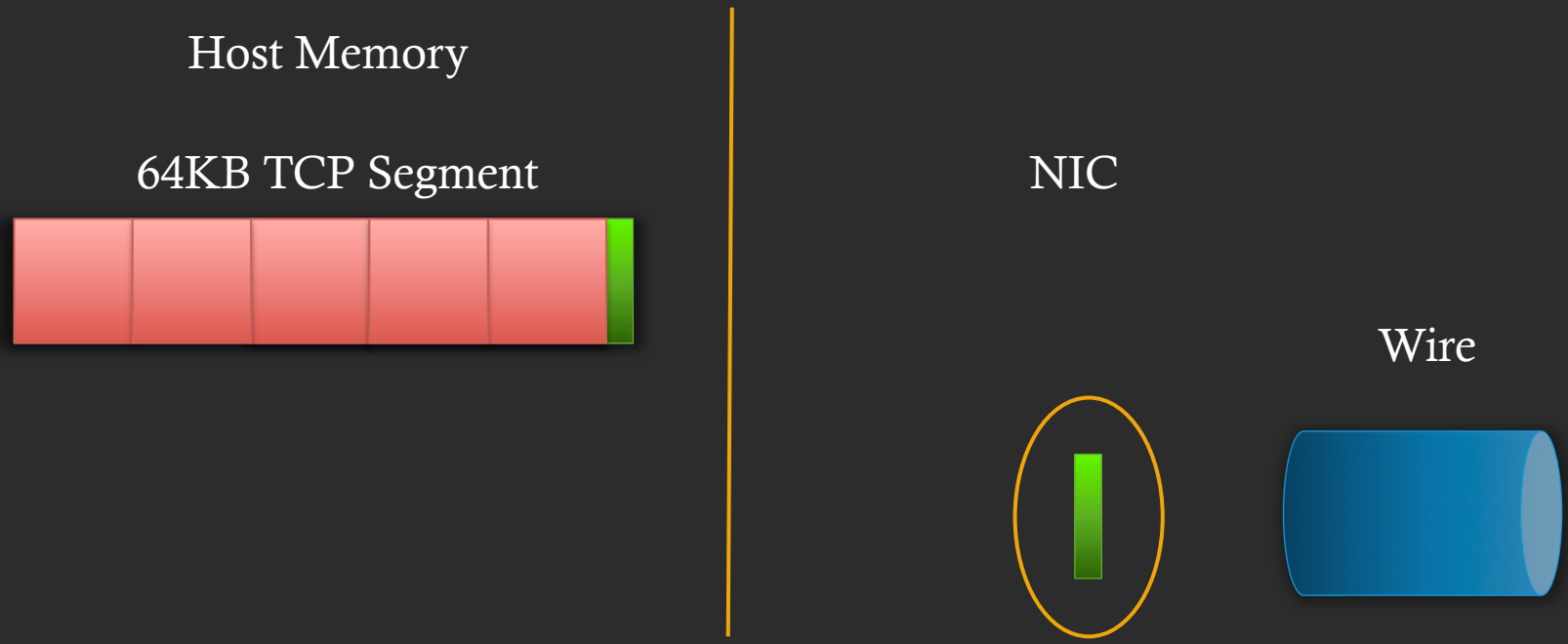


SENIC accurately enforces rate limits and delivers high throughput

SENIC Supports Other NIC Features

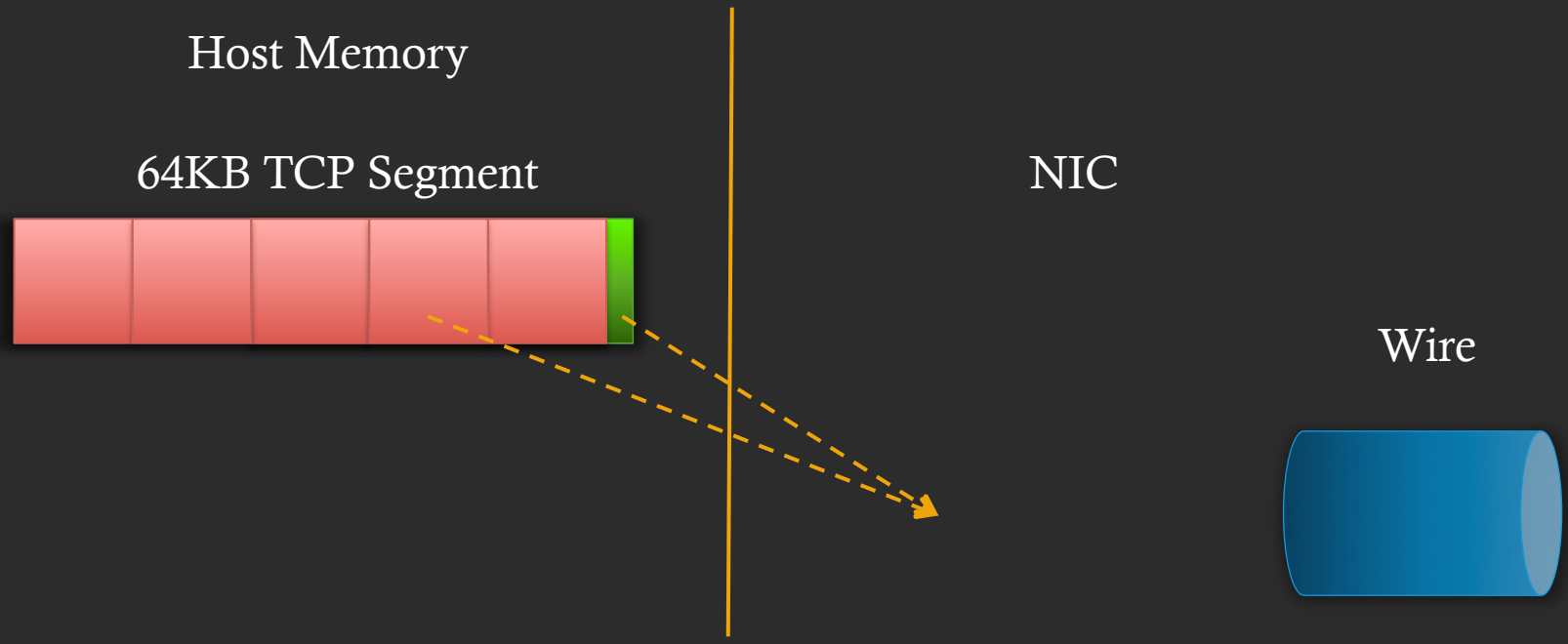
1. TCP Segmentation Offload
2. Hypervisor Bypass + Untrusted Guest VMs
3. Constant-Time Hierarchical Scheduler

TCP Segmentation Offload



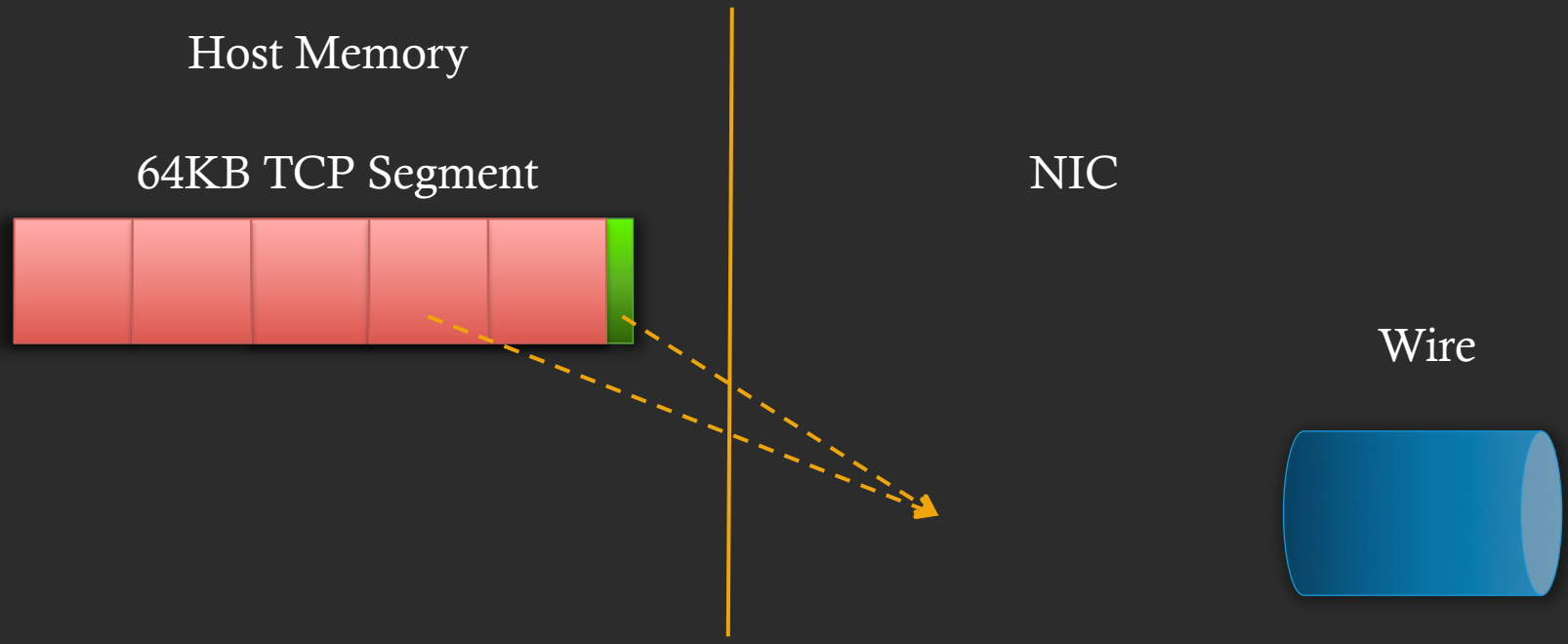
Header cached
on NIC

SENIC – TSO



DMA header and payload for each MTU sized packet

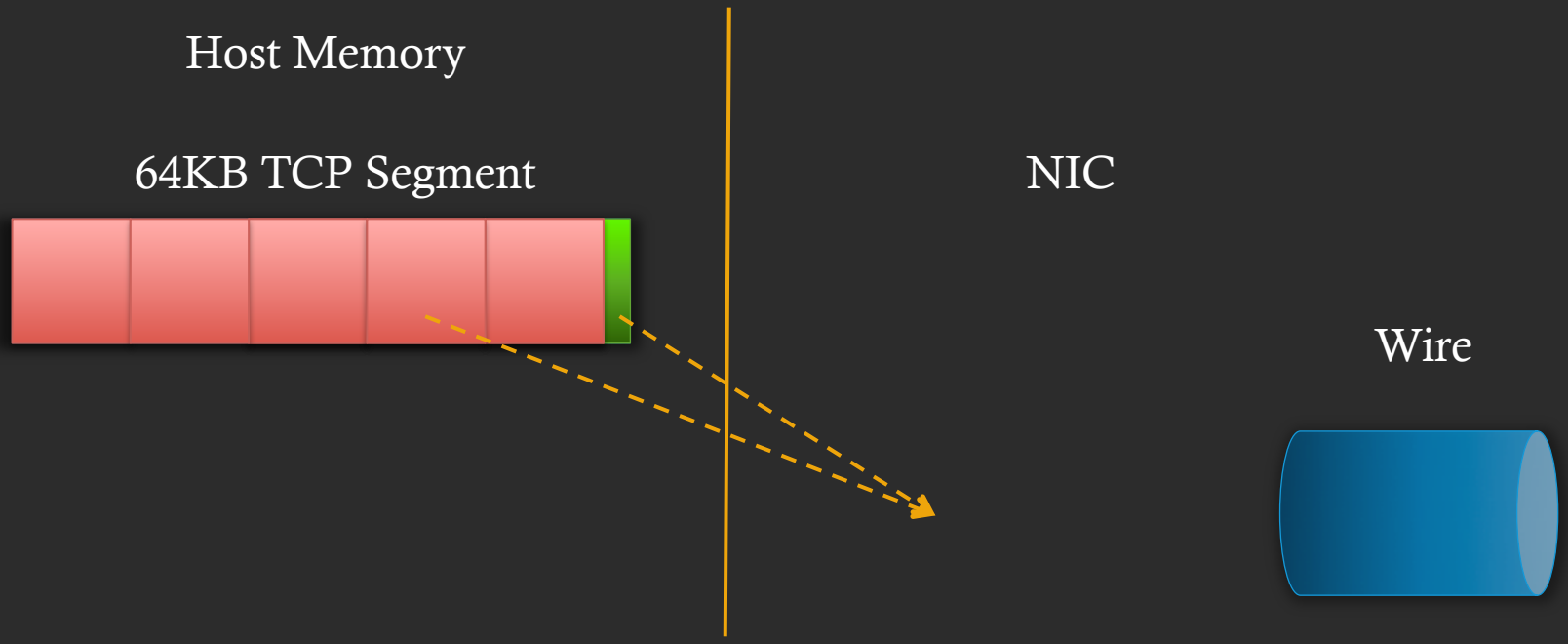
SENIC – TSO



2X DMA transfers?

No Problem!

SENIC – TSO



- 40Gb/s, 1500B MTU: 6.5M DMA transfers per second
- Measurement from a Mellanox Connect-X3 NIC:
  - 13 – 14M DMA transfers per second supported

Summary

- Delivers vision of scalable rate limiting
- Accurate and precise
- Easily implementable in hardware and software

Code @ http://sivasankar.me/senic/