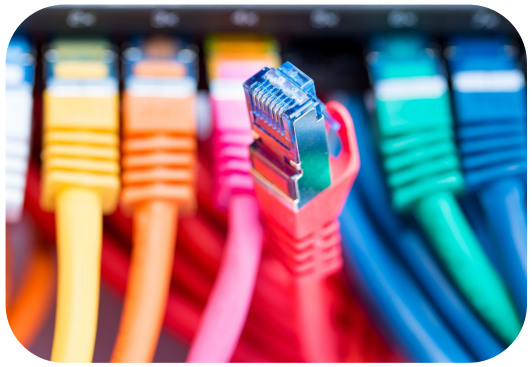
FlexDriver: A Network Driver for Your Accelerator

HAGGAI ERAN, MAXIM FUDIM, GABI MALKA,
GAL SHALOM, NOAM COHEN, AMIT HERMONY,
DOTAN LEVI, LIRAN LISS, MARK SILBERSTEIN

Motivation: accelerators network access

Distributed accelerated computing

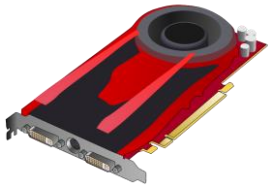
Accelerator disaggregation



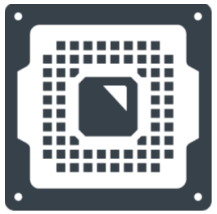
Packet processing acceleration

Motivation: distributed accelerated computing

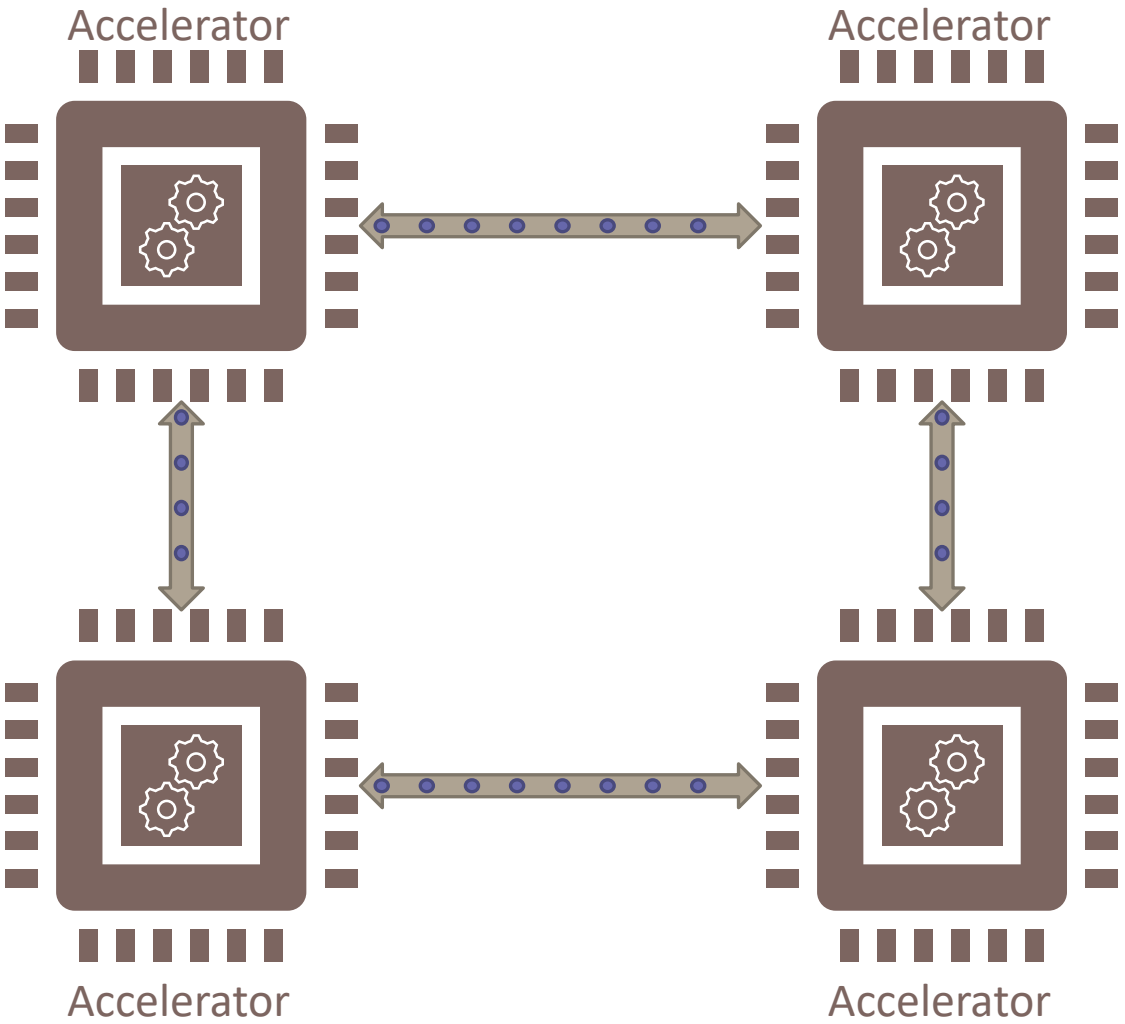
Fixed-function accelerators

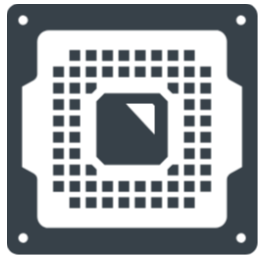


GPUs



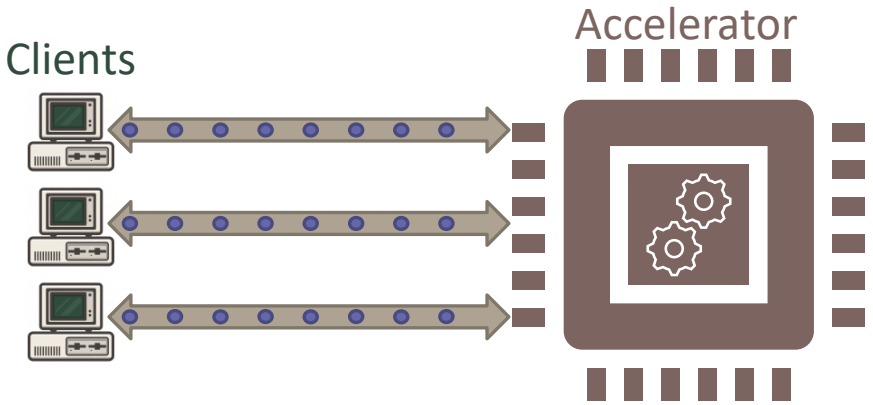
AI accelerators



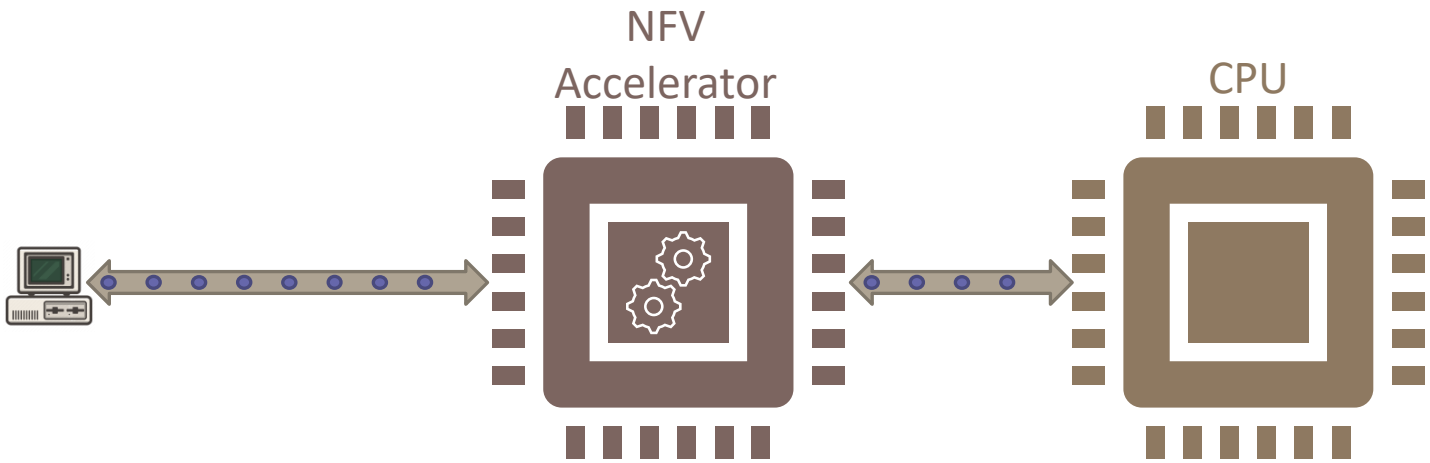


FPGAs

Motivation: accelerator disaggregation



Motivation: packet processing acceleration

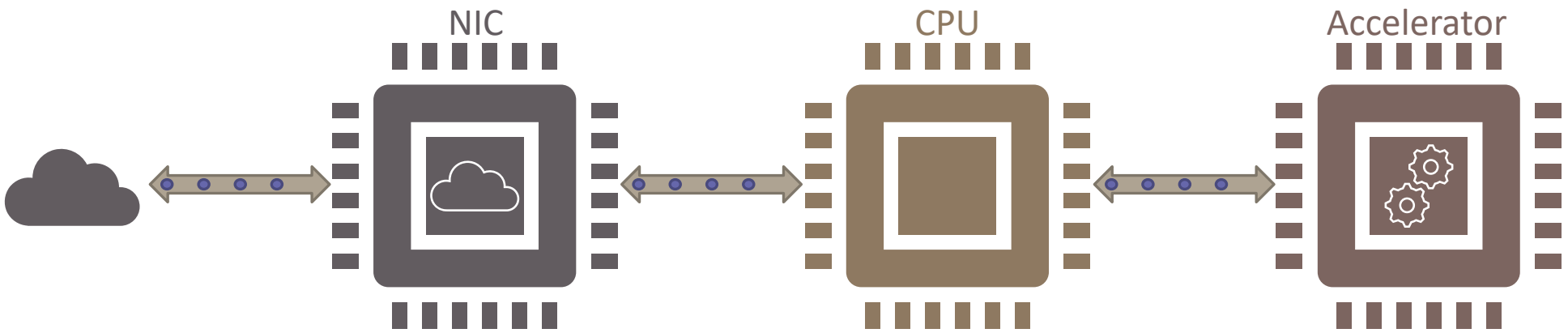


Efficient accelerator networking is important

Existing alternatives

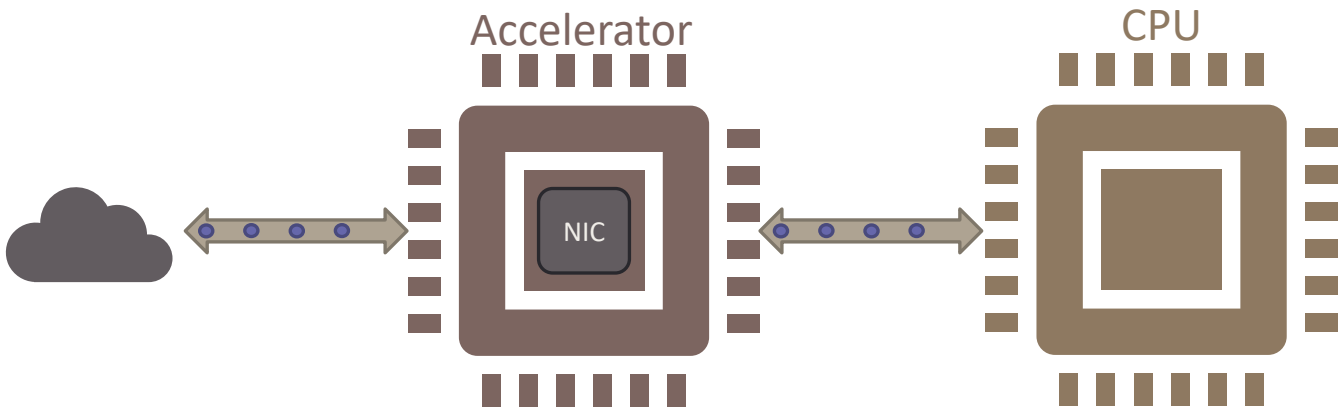
1. CPU-mediated
2. Accelerator-hosted
3. Bump-in-the-wire

Alternative 1: CPU-mediated design



- 👎 High CPU usage
- 👎 Potential bottleneck

Alternative 2: accelerator-hosted design

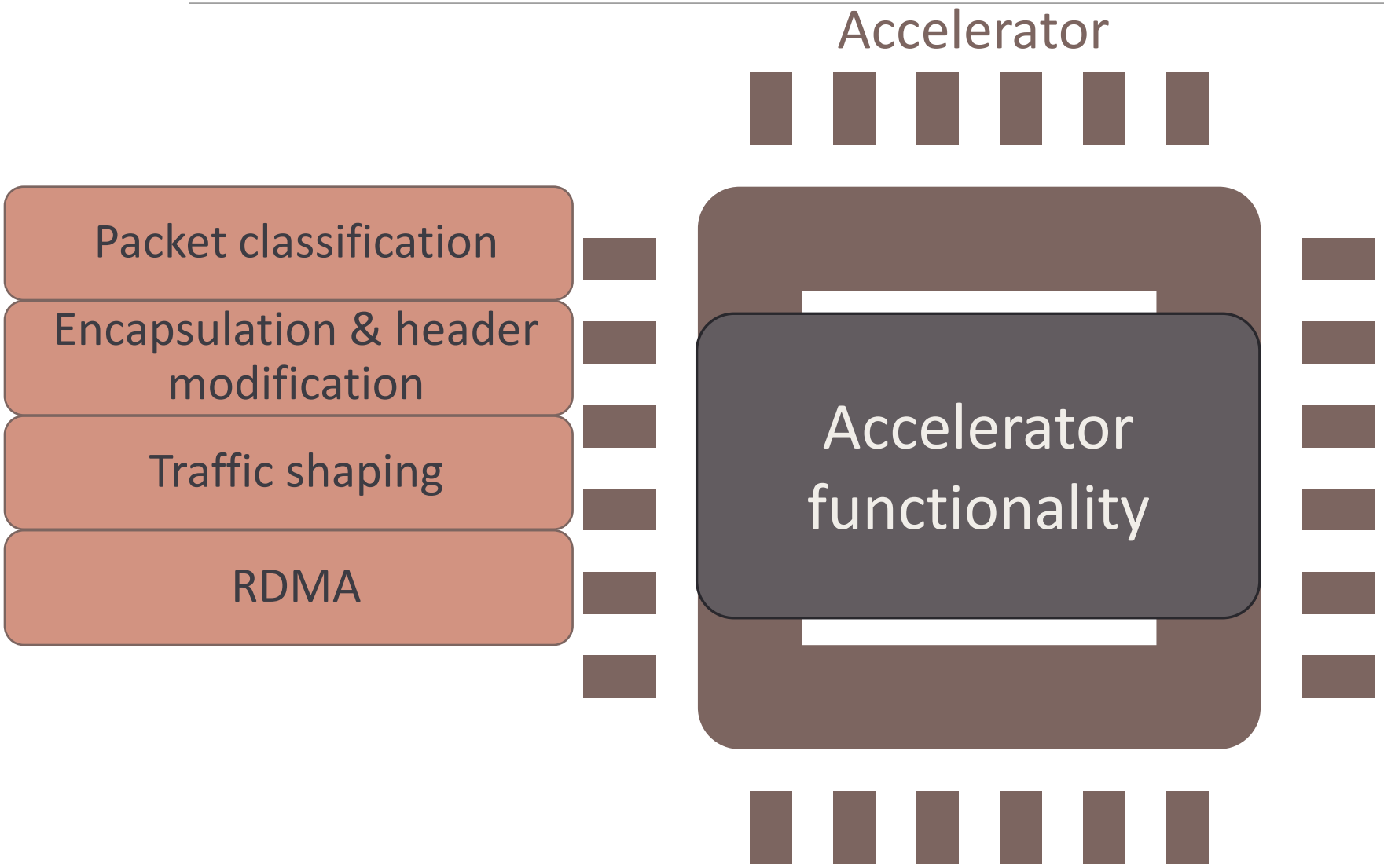


StRoM: Smart Remote Memory, Sidler et al., EuroSys ‘20.
Corundum: An Open-Source 100 Gbps NIC, Forencich et al., FCCM ‘20.

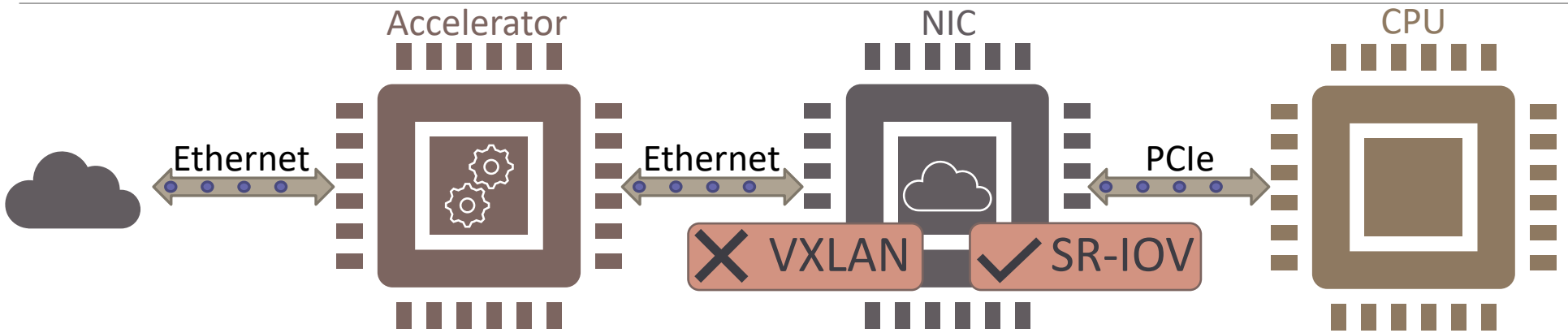
Google Cloud TPU

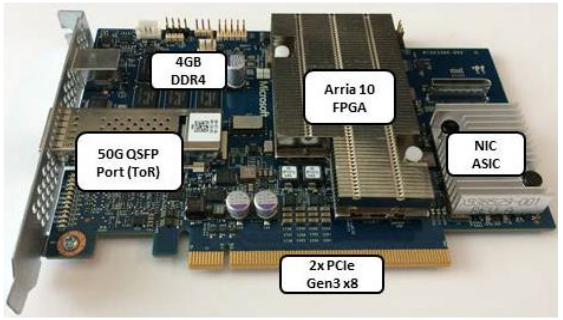
Intel Habana Gaudi

Accelerator-hosted design: Area—features trade-off



Alternative 3: bump-in-the-wire design



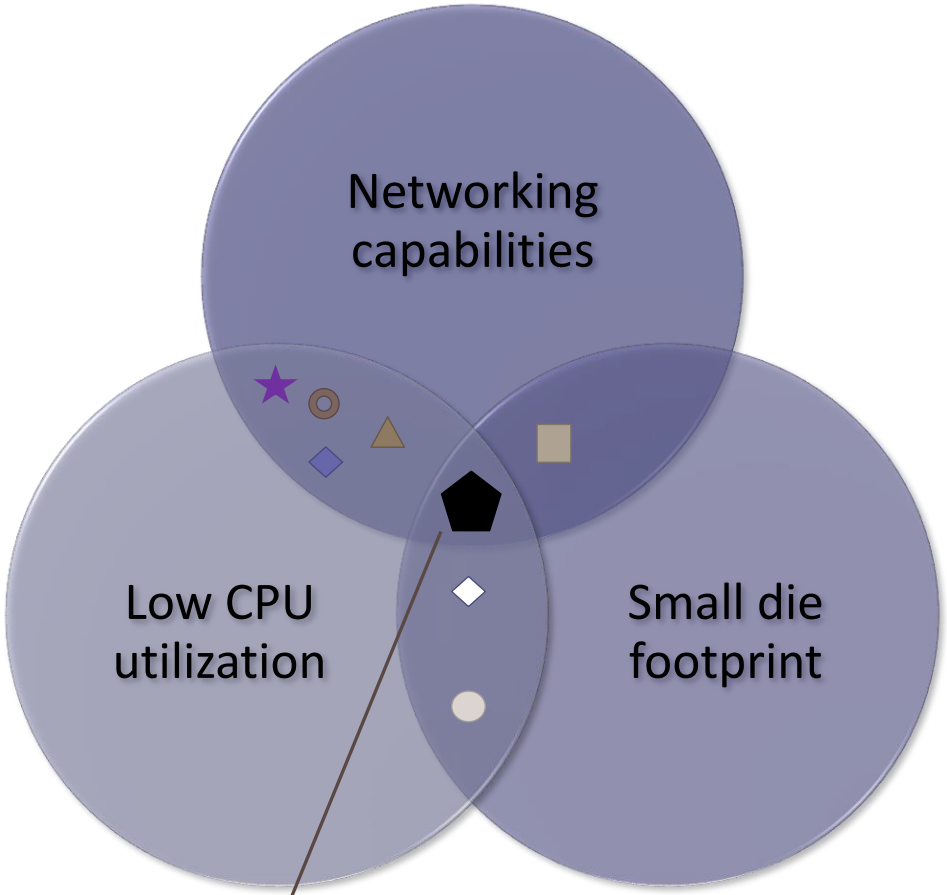


Microsoft Azure [NSDI'18]

NVIDIA Innova (1st gen.)

Intel PAC N3000

CPU usage vs. area vs. features



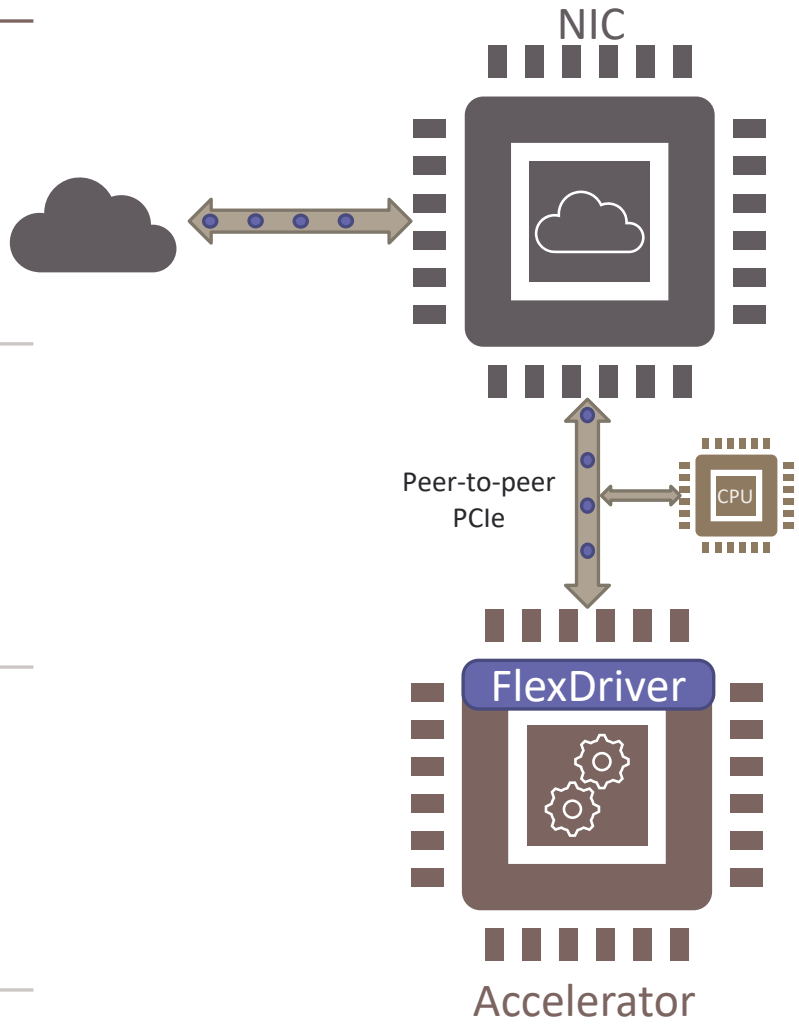
1. CPU-mediated	VN2F
2. Accelerator-hosted	Brainwave
	Corundum
	StRoM
3. Bump-in-the-wire	Innova-1 shell
	NICA
	AccelNet
Can we do better?	FlexDriver

FlexDriver: a new and better design point

FlexDriver design

Goals:

- Offload to NIC
- Low area requirements
- No CPU on data-path



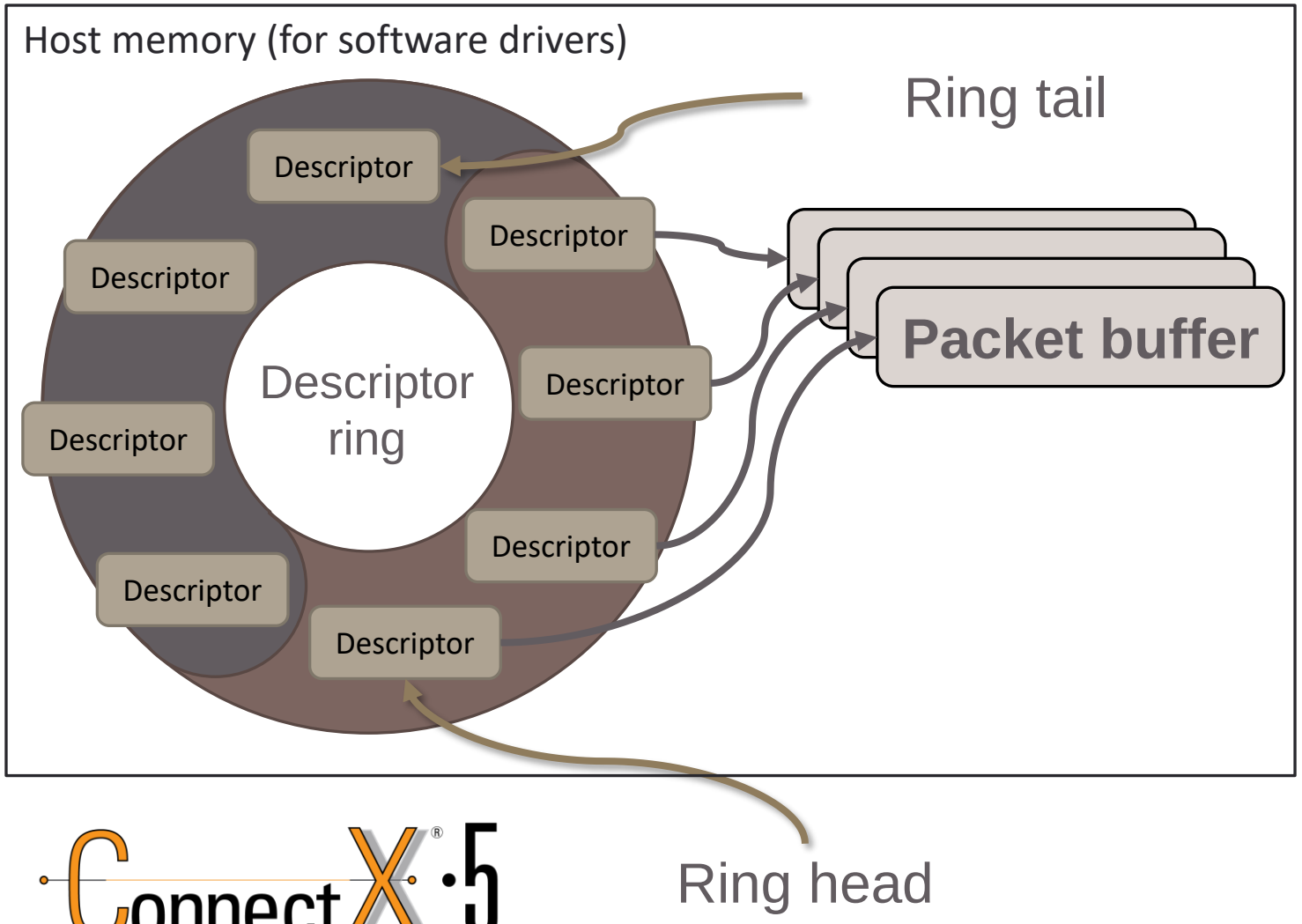
Agenda

FlexDriver design — Memory constraints

Evaluation — Area & throughput

Use-cases — Utilizing offloads

Background: NIC transmit interface



Free buffers

Driver can post new packets to be transmitted there.

Posted buffers

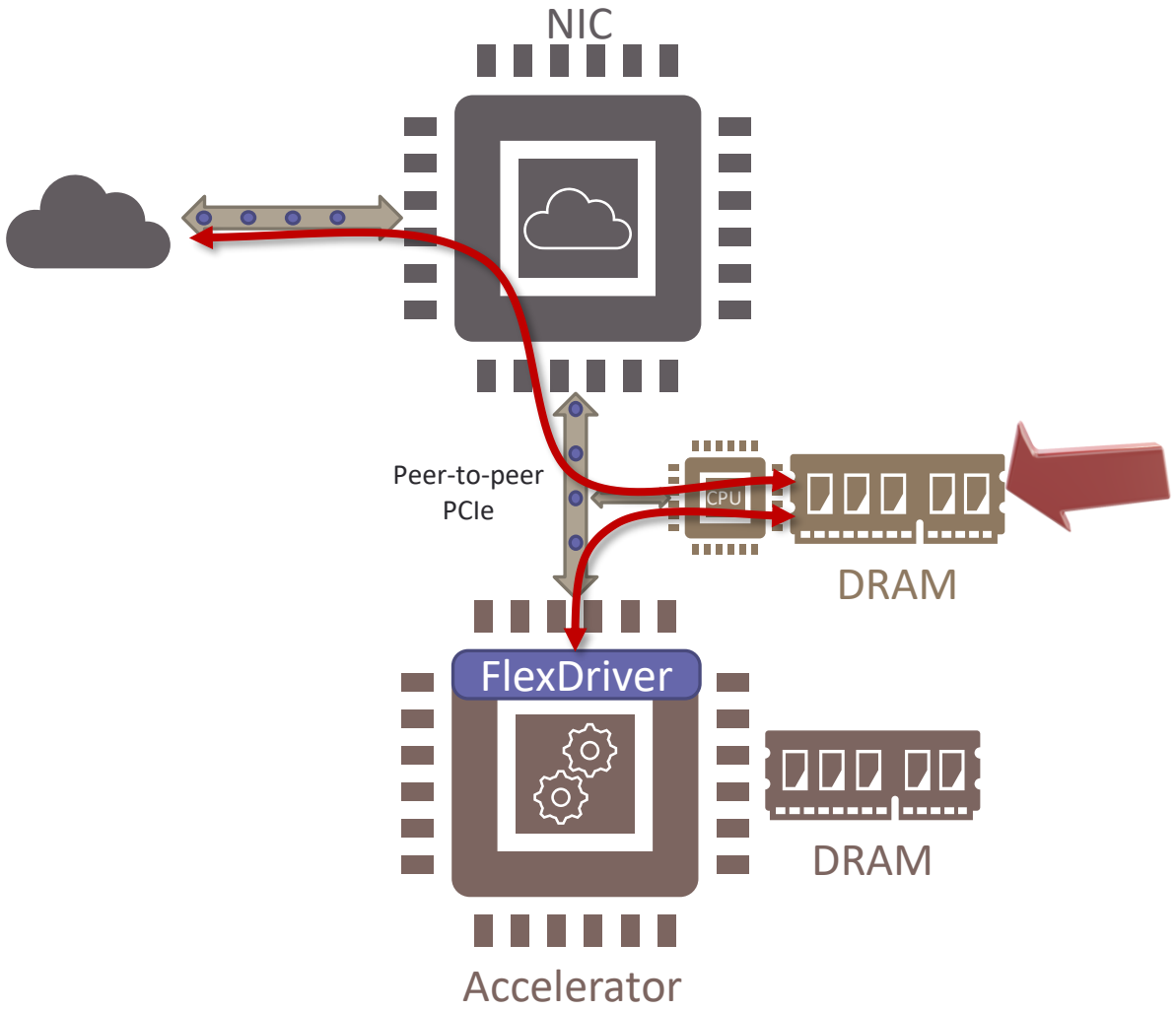
NIC can transmit these buffers.

ConnectX®-5

Where to place rings and buffers?

Host memory

- ✕ PCIe congestion
- ✕ Limit scalability



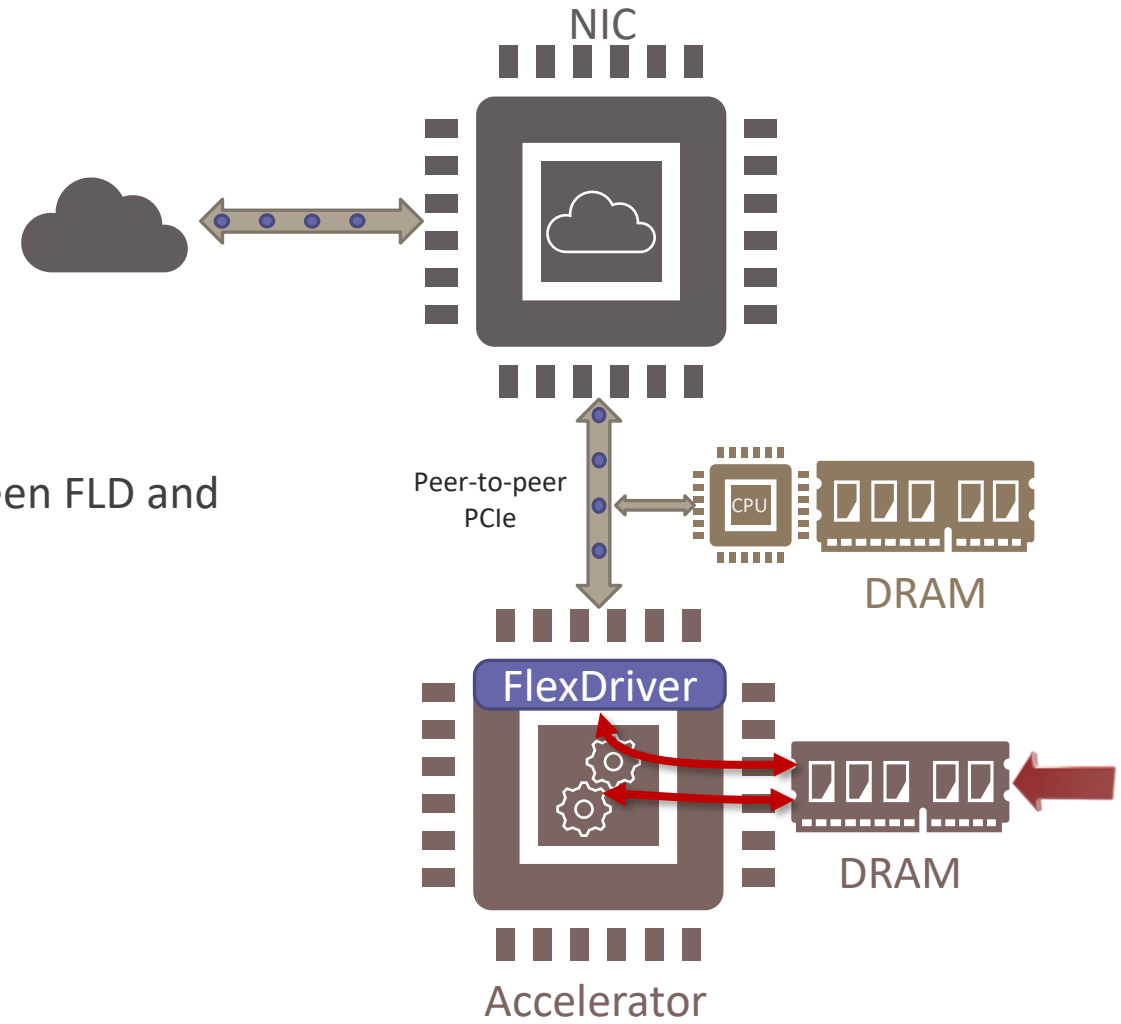
Where to place rings and buffers?

Host memory

- ✕ PCIe congestion
- ✕ Limit scalability

Accelerator-DRAM

- ✕ interference between FLD and accelerator

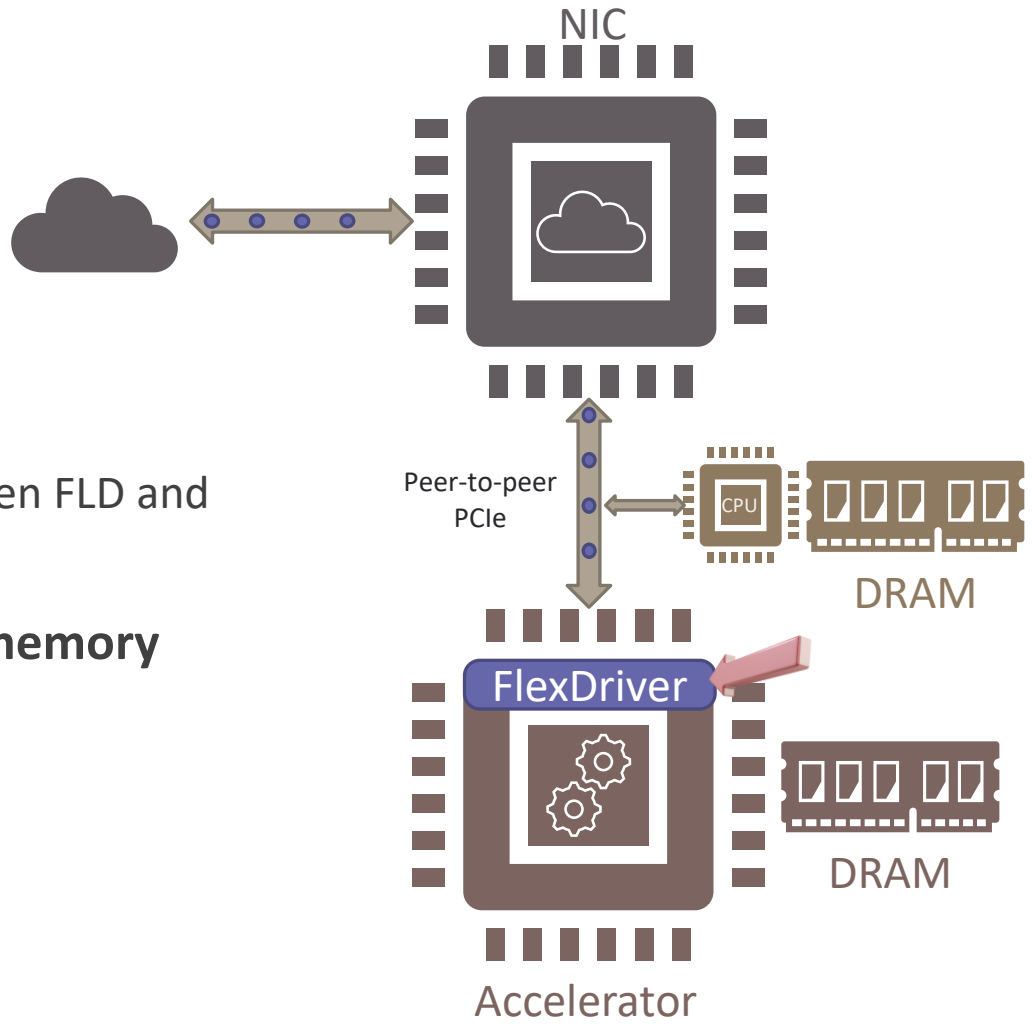


Where to place rings and buffers?

Host memory
- ✕ PCIe congestion
- ✕ Limit scalability

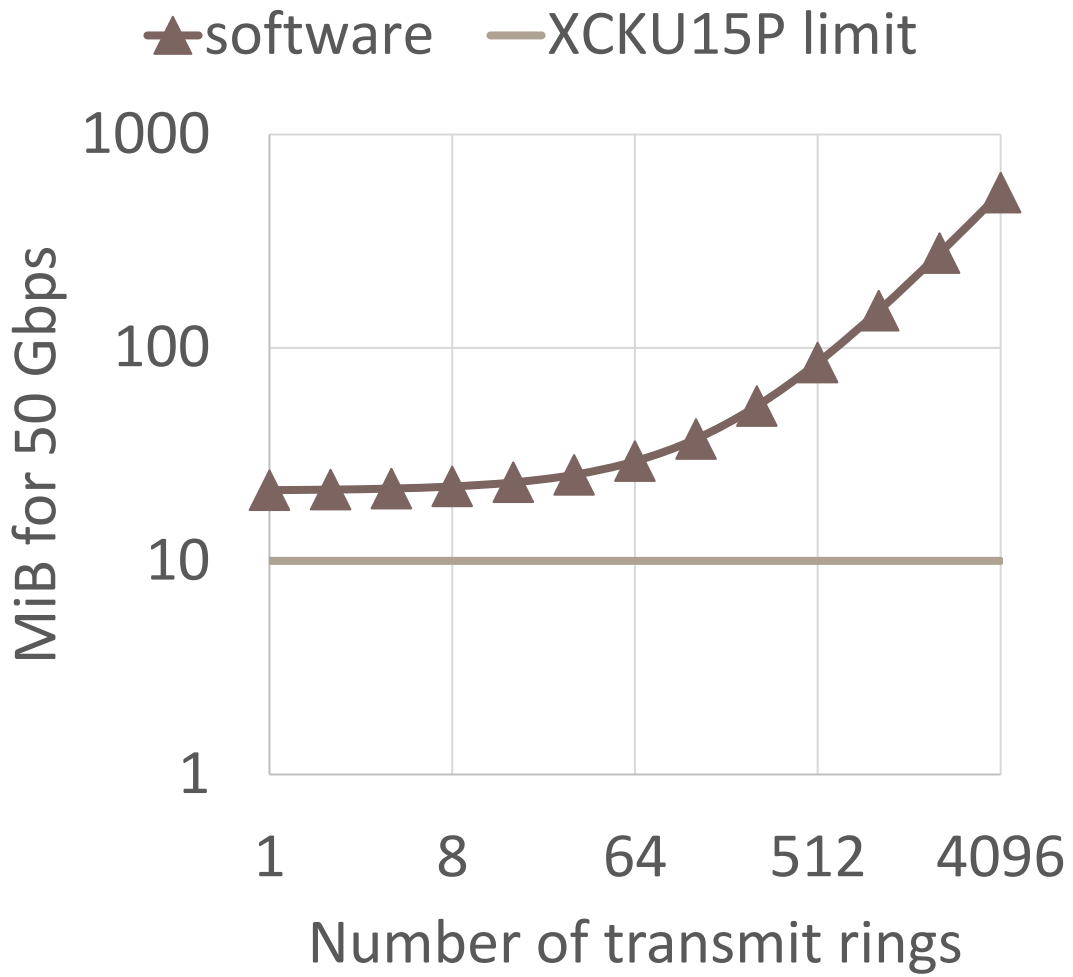
Accelerator-DRAM
- ✕ interference between FLD and accelerator

Dedicated on-chip memory
- ✓ No interference
- ◦ Challenge: limited area



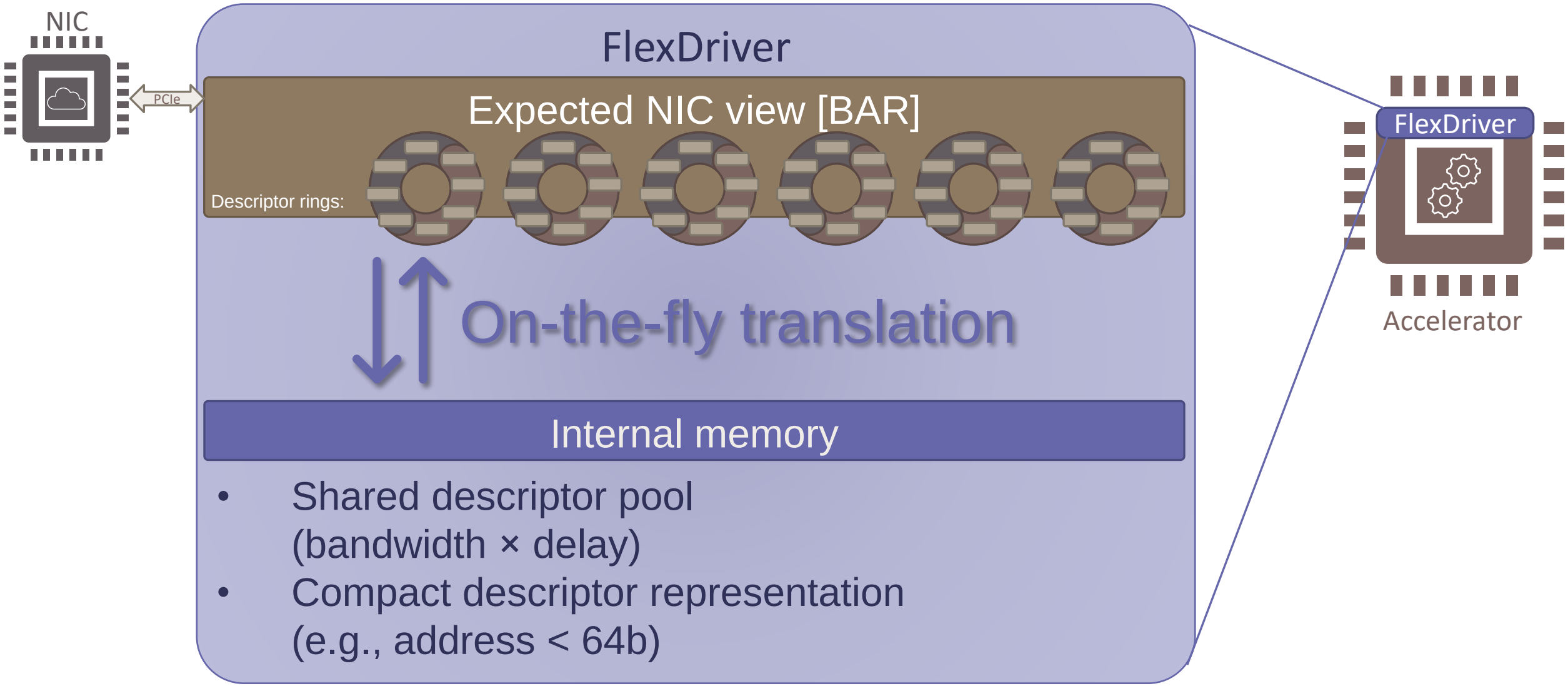
Challenge: memory for NIC—device interface

Ring size determined for latency hiding and throughput.
Multiple rings are needed for RDMA, QoS.

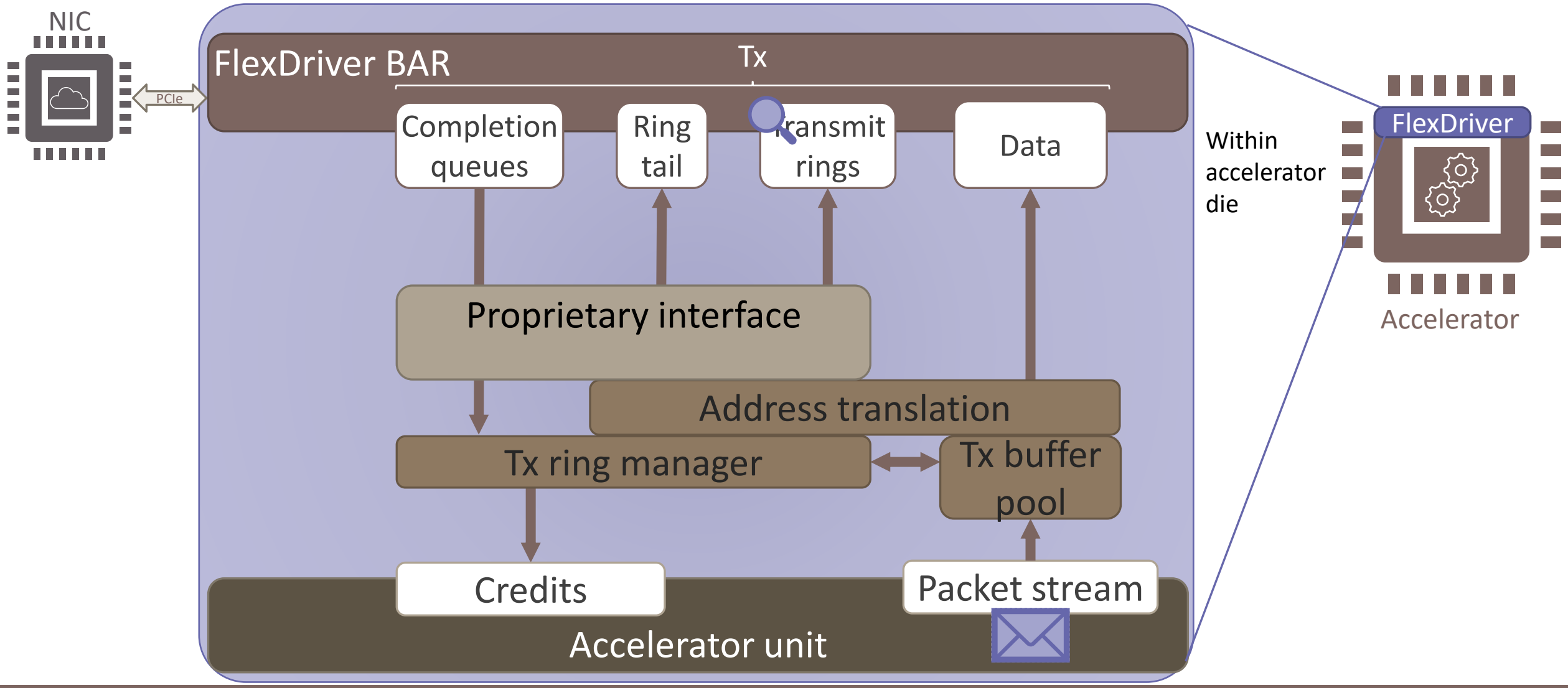


See paper for details

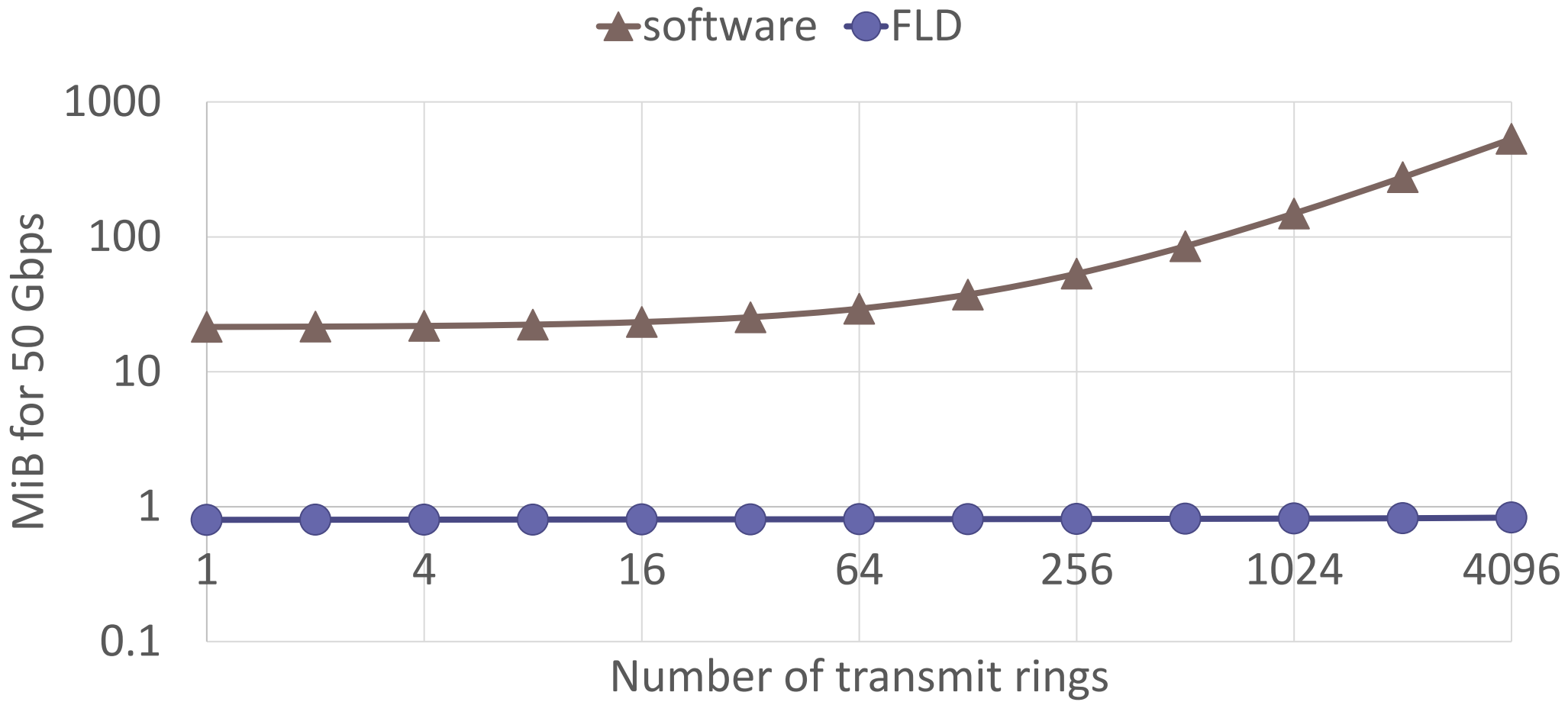
Solution: on-the-fly memory optimizations



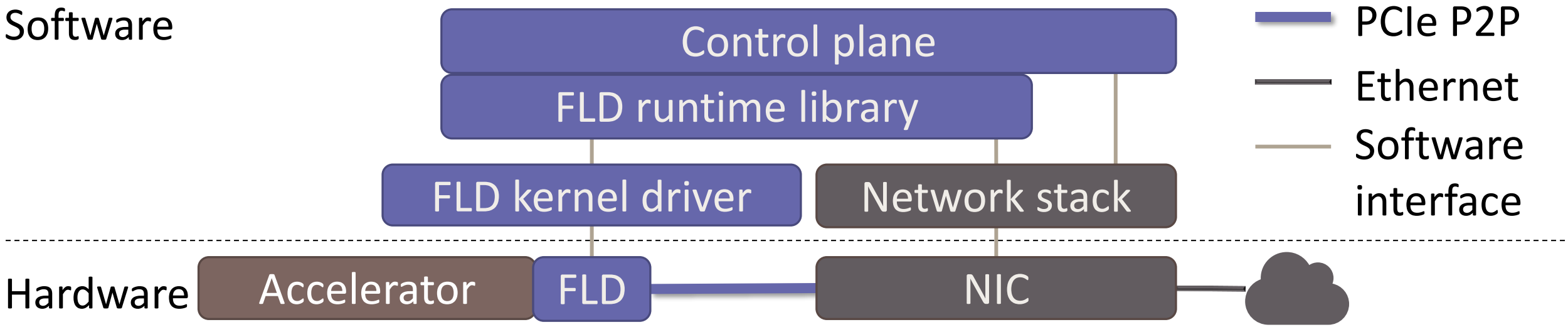
Hardware transmit block diagram



FlexDriver dramatically improves scaling



Software design



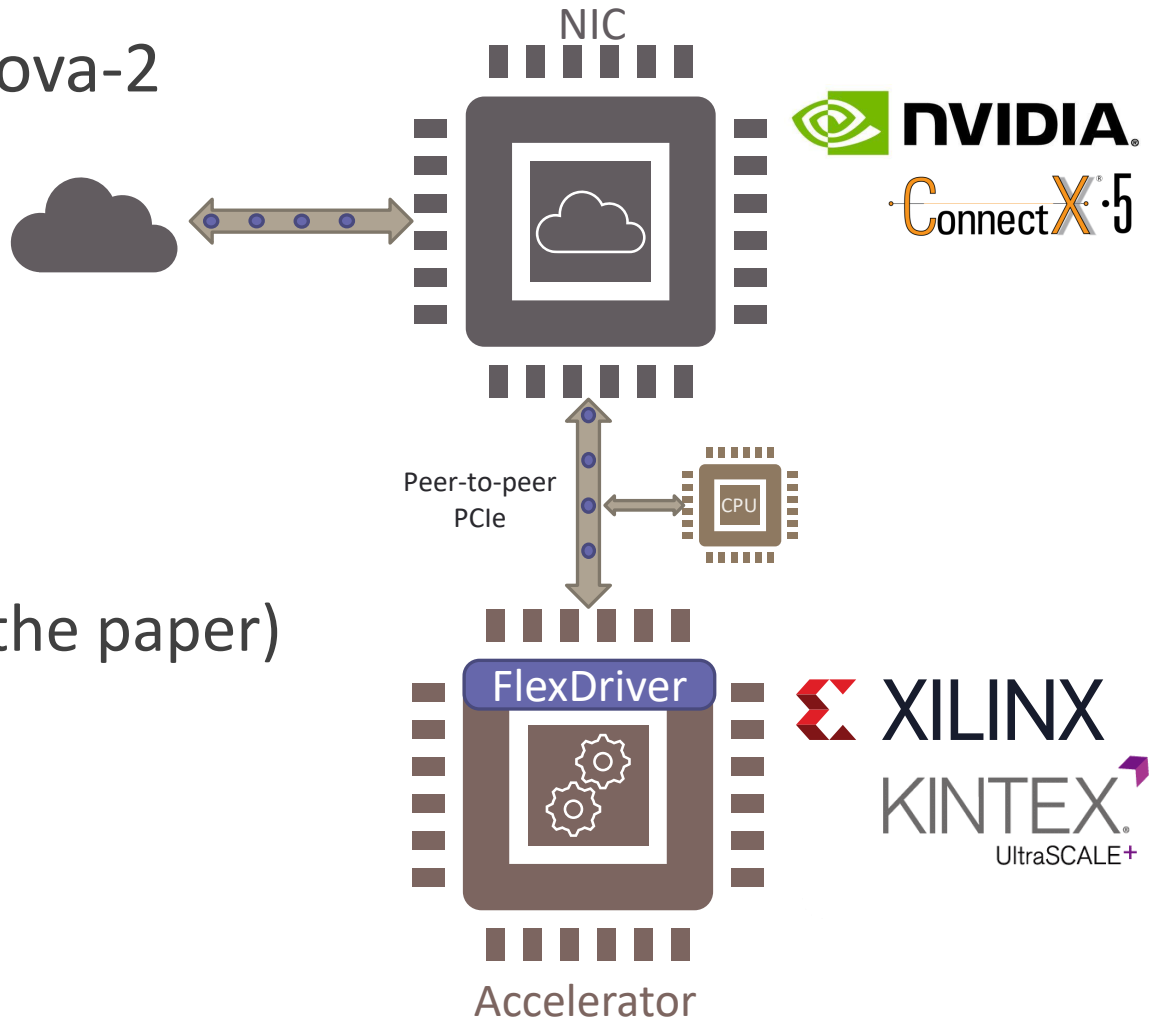
FLD-E: Ethernet DPDK, Linux TC

FLD-R: RDMA RoCE, RDMA CM

More details in the paper

Evaluation

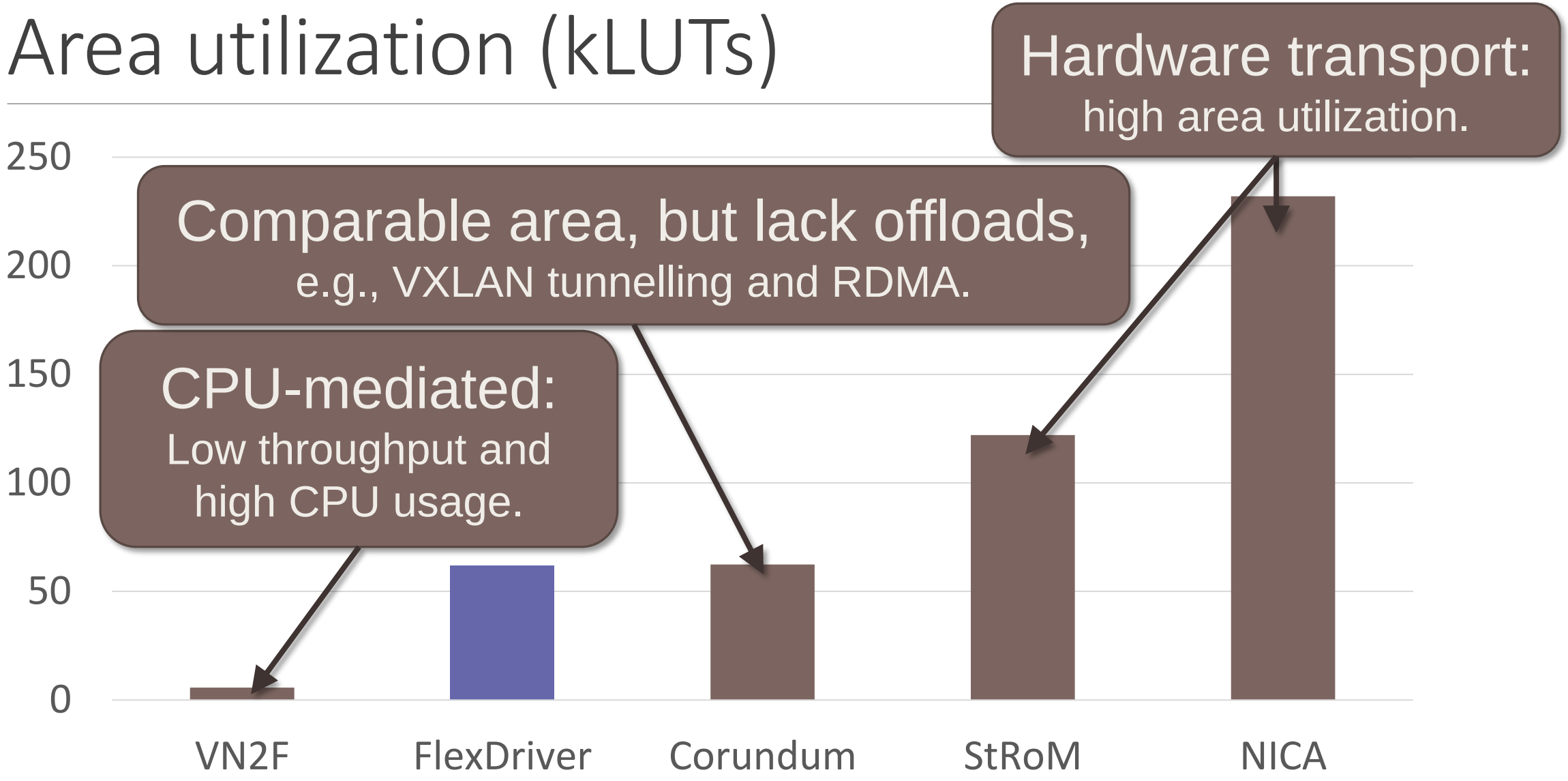
Implemented on NVIDIA Innova-2



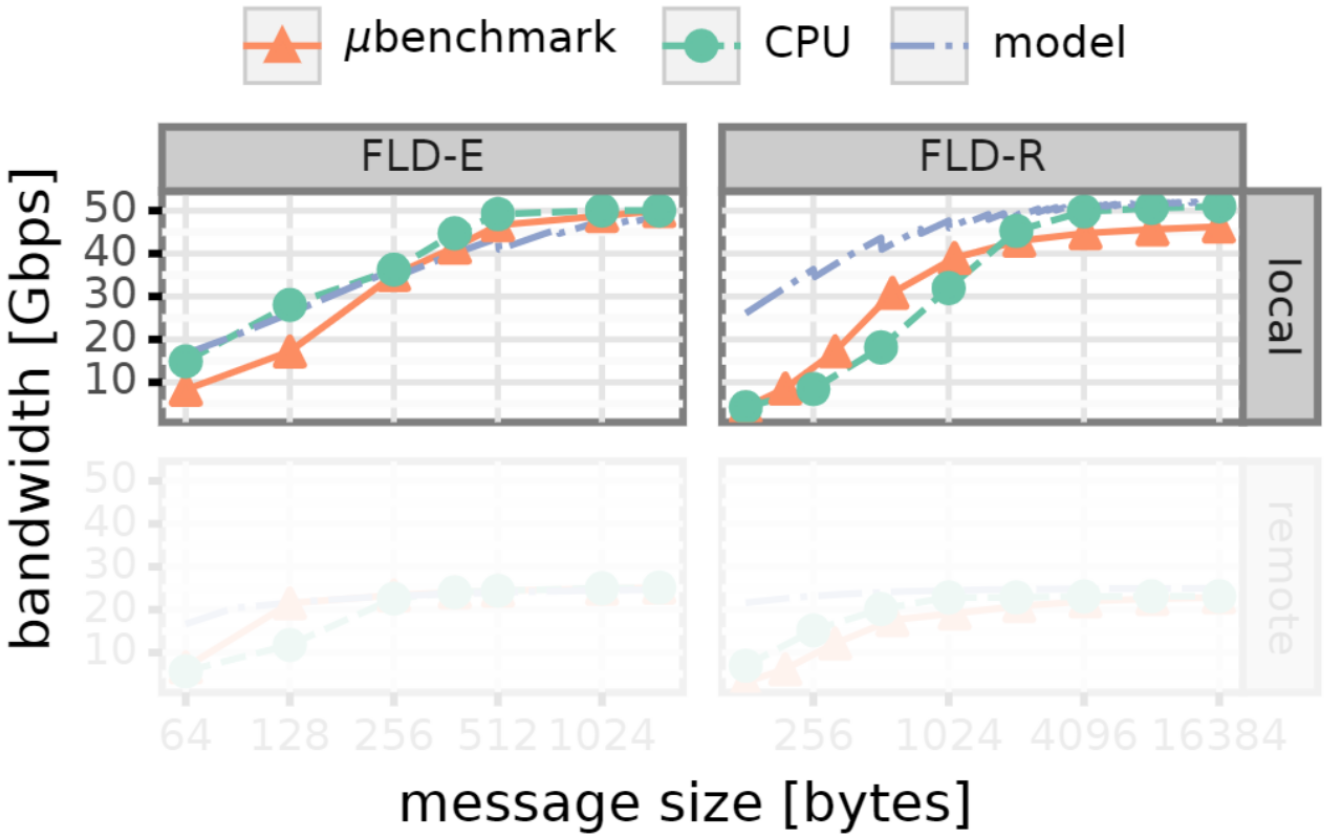
Evaluation

- Performance model (see the paper)
- Microbenchmarks
- Example use-cases

Area utilization (kLUTs)

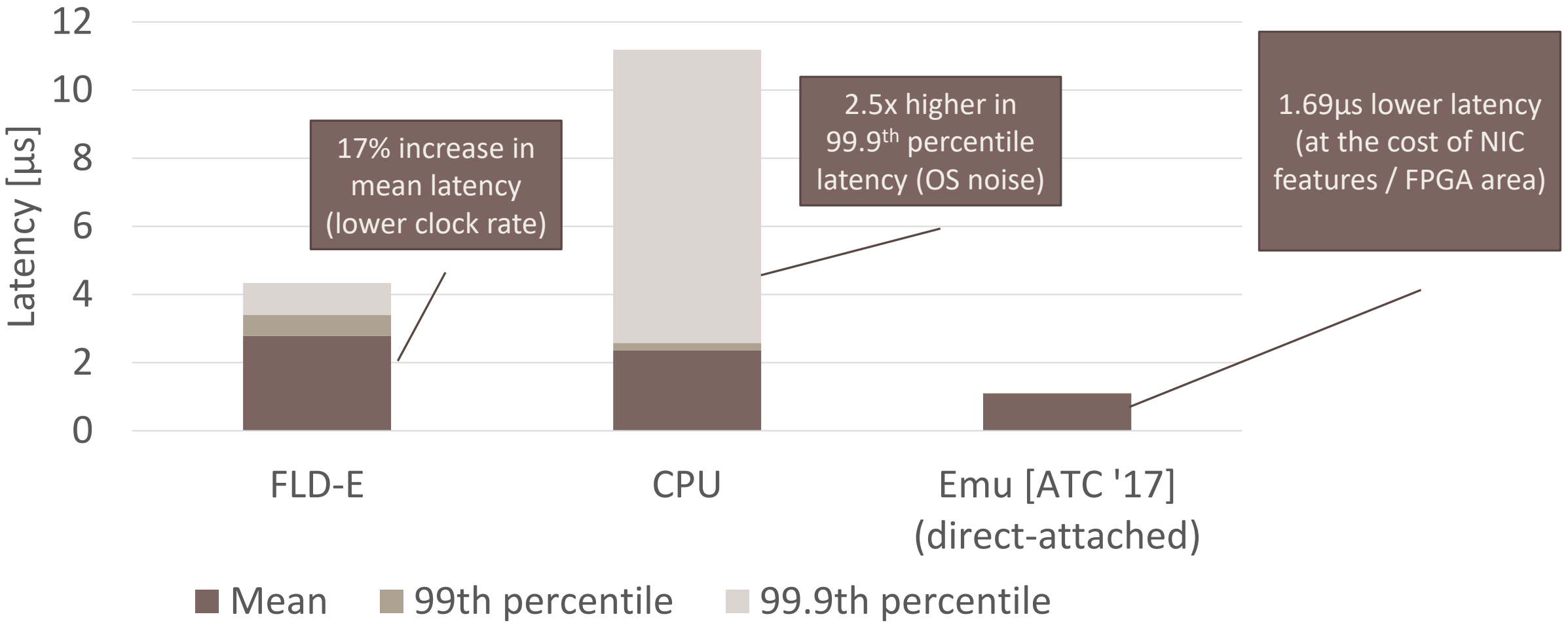


Evaluation: microbenchmark throughput



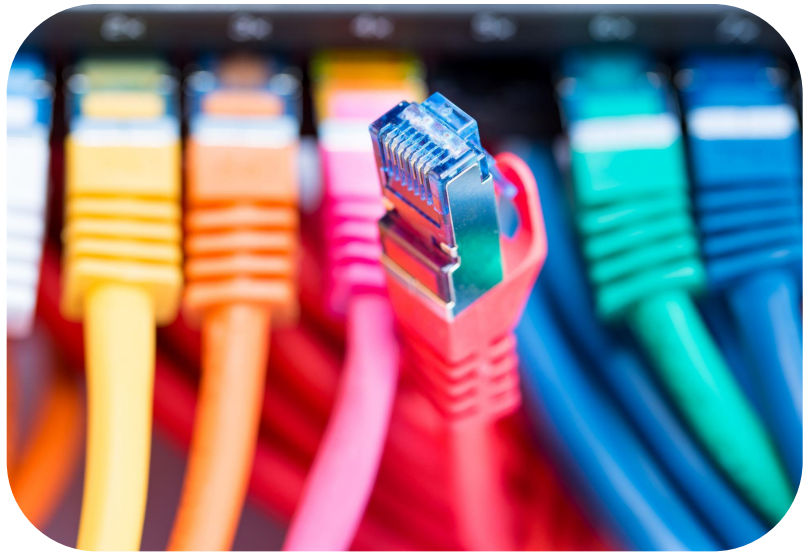
Large packets: FlexDriver reaches network and PCIe limits
Small-packet: remaining optimizations

Evaluation: microbenchmark latency



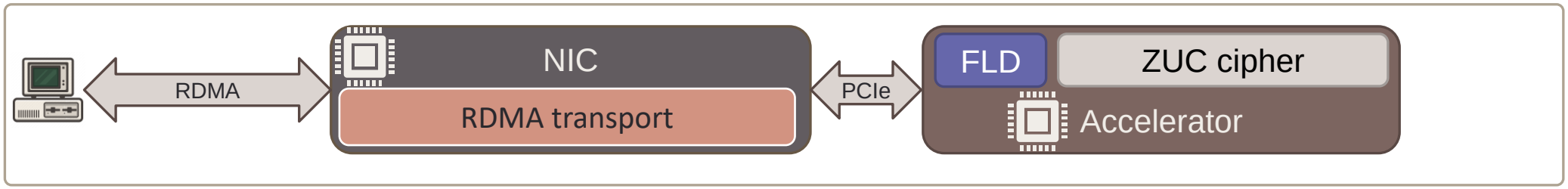
Use-cases

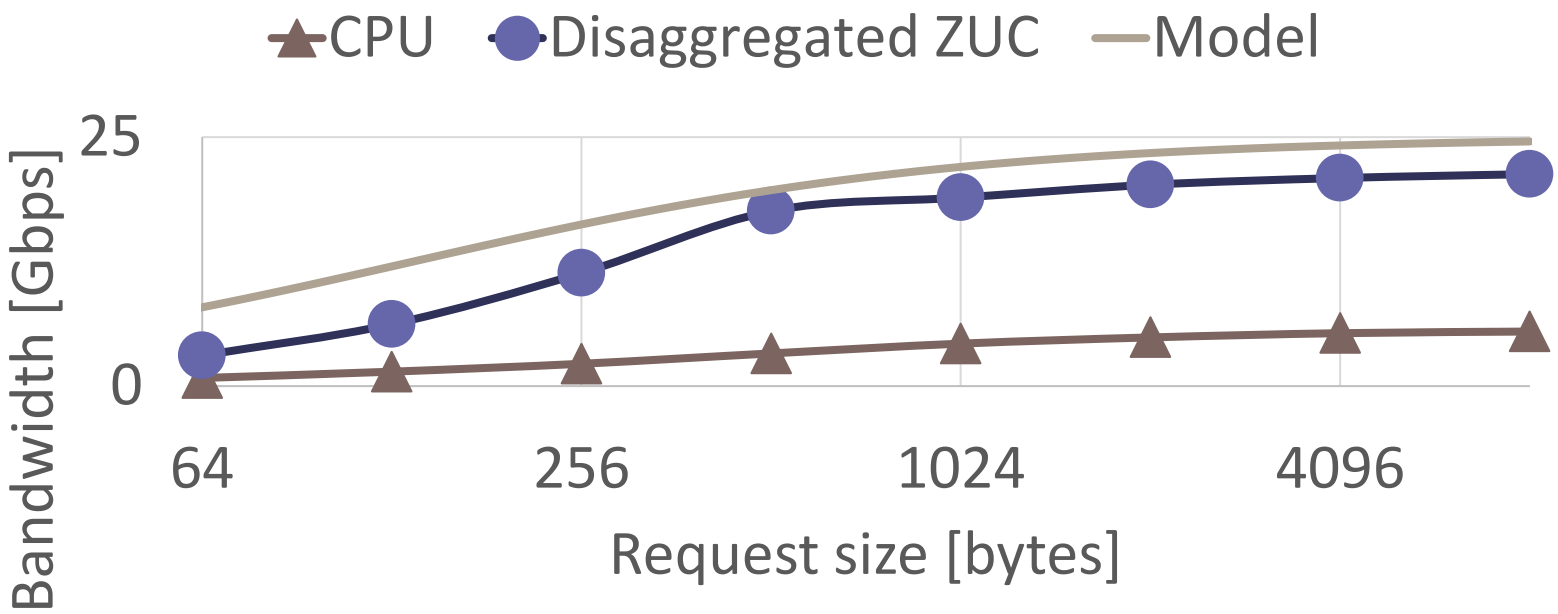
Disaggregated accelerator using RDMA



Two packet processing accelerators using the NIC offloads

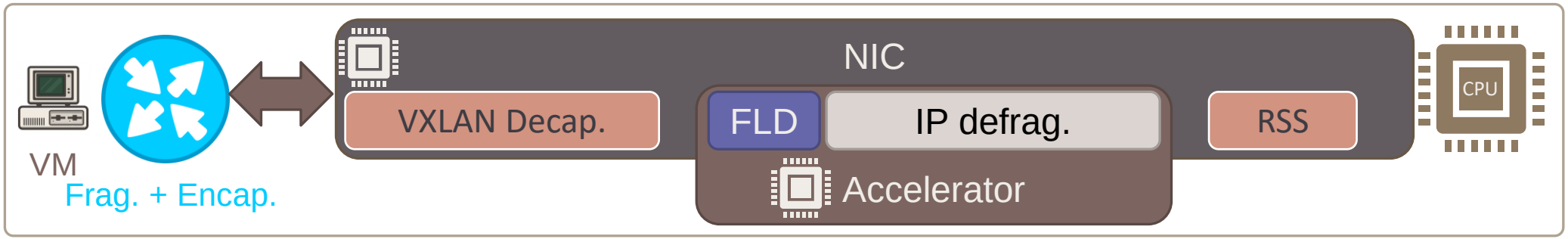
ZUC mobile cipher disaggregated accelerator





Single-core: 4× performance of the CPU.
Near the performance model's maximum.

IP defragmentation inline accelerator



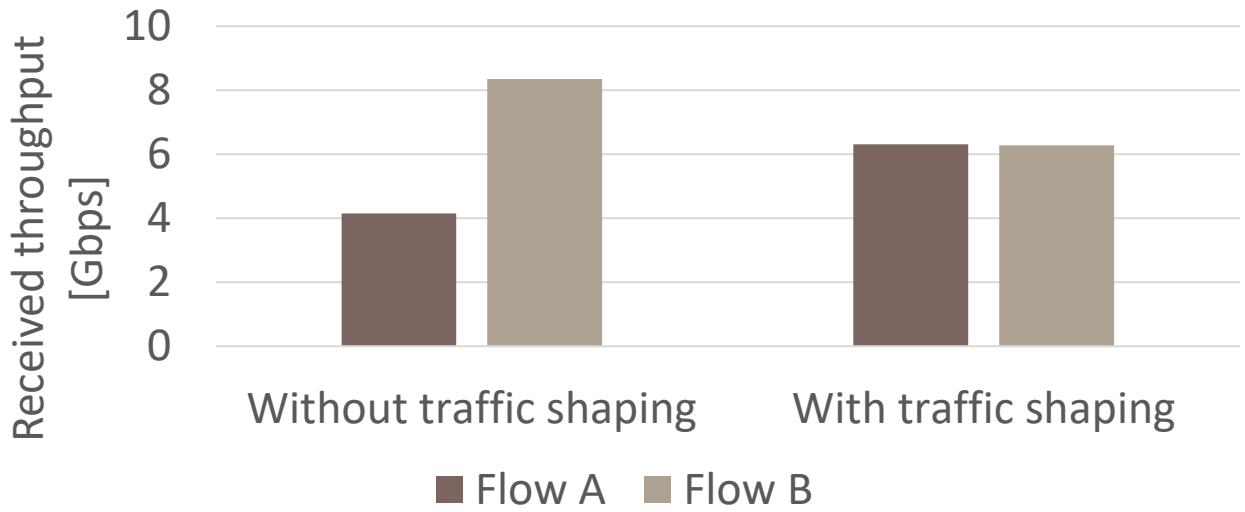
Configuration	Software	Hardware
No frag.	23.2 Gbps	23.2 Gbps
Frag.	3.2 Gbps	22.4 Gbps **(×7)**
VXLAN + Frag.	3.2 Gbps	16.8 Gbps **(×5.25)**

Better load balancing with hardware defragmentation.

IoT message authentication accelerator

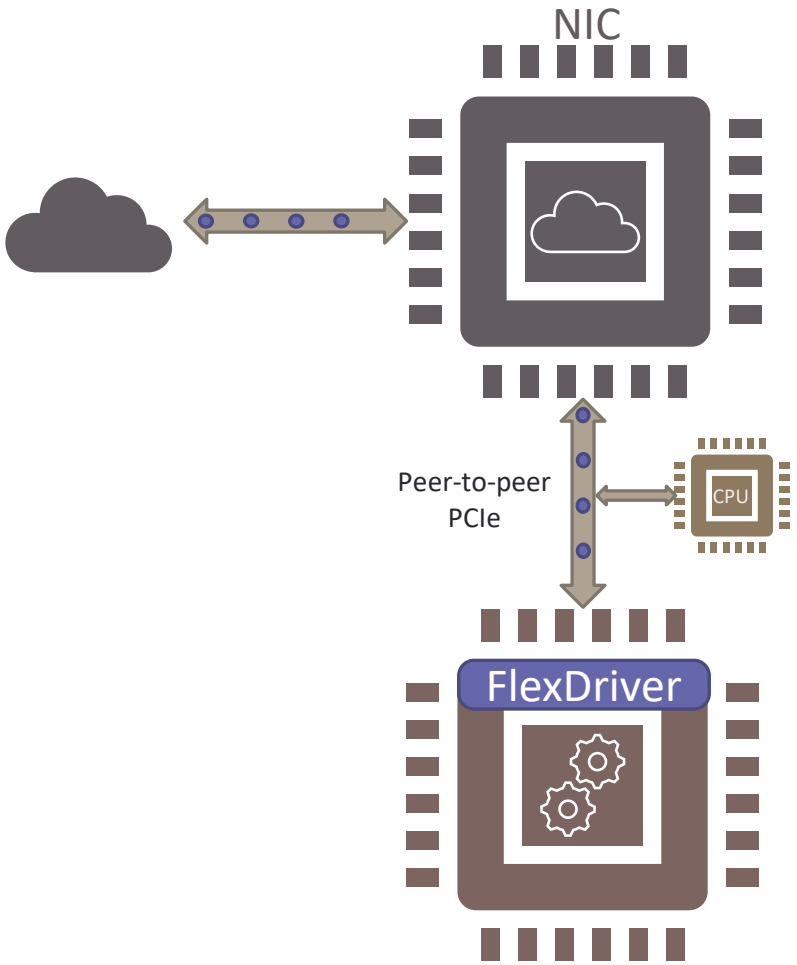
25 Gbps @256B packets

Fairness experiment:
12 Gbps accelerator
Flow B receives
twice the traffic



With traffic shaping: flow A gets its fair share

Conclusion



FlexDriver provides accelerators with **efficient network access**, relying on **existing ASIC NIC features** to **reduce area utilization** in the accelerator.

Thank you for listening!

haggaie@nvidia.com

NVIDIA

ACSL Accelerated Computing Systems Lab

TECHNION Israel Institute of Technology