Efficient RNIC Cache Side-channel Attack Detection through DPU-driven Architecture

Yunkun Liao, Jingya Wu, Wenyan Lu, Xiaowei Li, Guihai Yan

SKLP, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

2024/12/3

EURO-PAR
CONFERENCE 2024

ADAPT Architecture for Data Analytics and Processing Technology

SKLP 处理器芯片全国重点实验室 STATE KEY LAB OF PROCESSORS, ICT, CAS

ICT 中国科学院计算技术研究所 INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

中国科学院大学 University of Chinese Academy of Sciences

中科驭数 YUSUR

Contens

Background and Motivation 01

Our Design: DPU-driven RCSCA Detection 02

Evaluation 03

Conclusion 04

Background and Motivation

Datacenter Networking Demands

Tensorflow

Nccl

Spark

Memcached

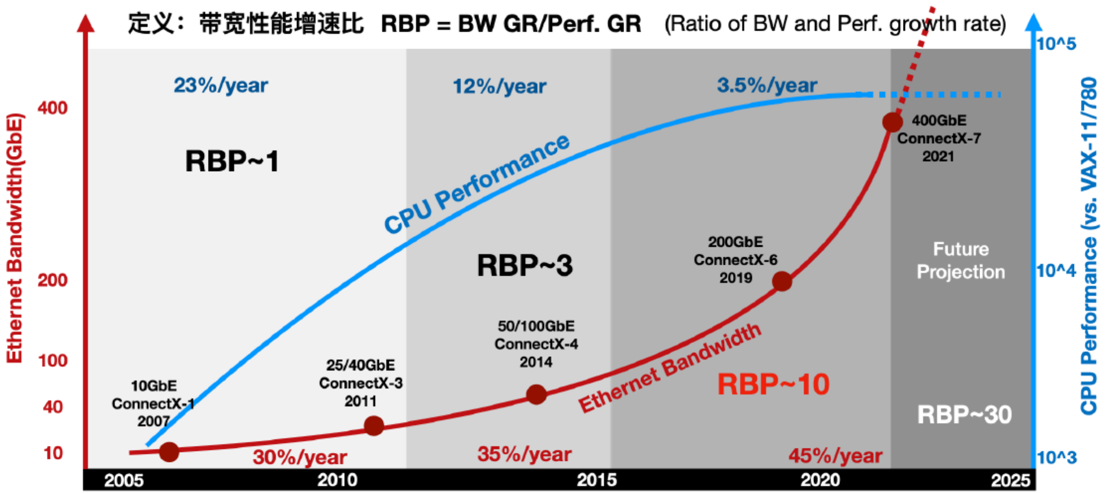
Distributed Transactions

Disaggregated Memory

Redis

RPC

Microservice



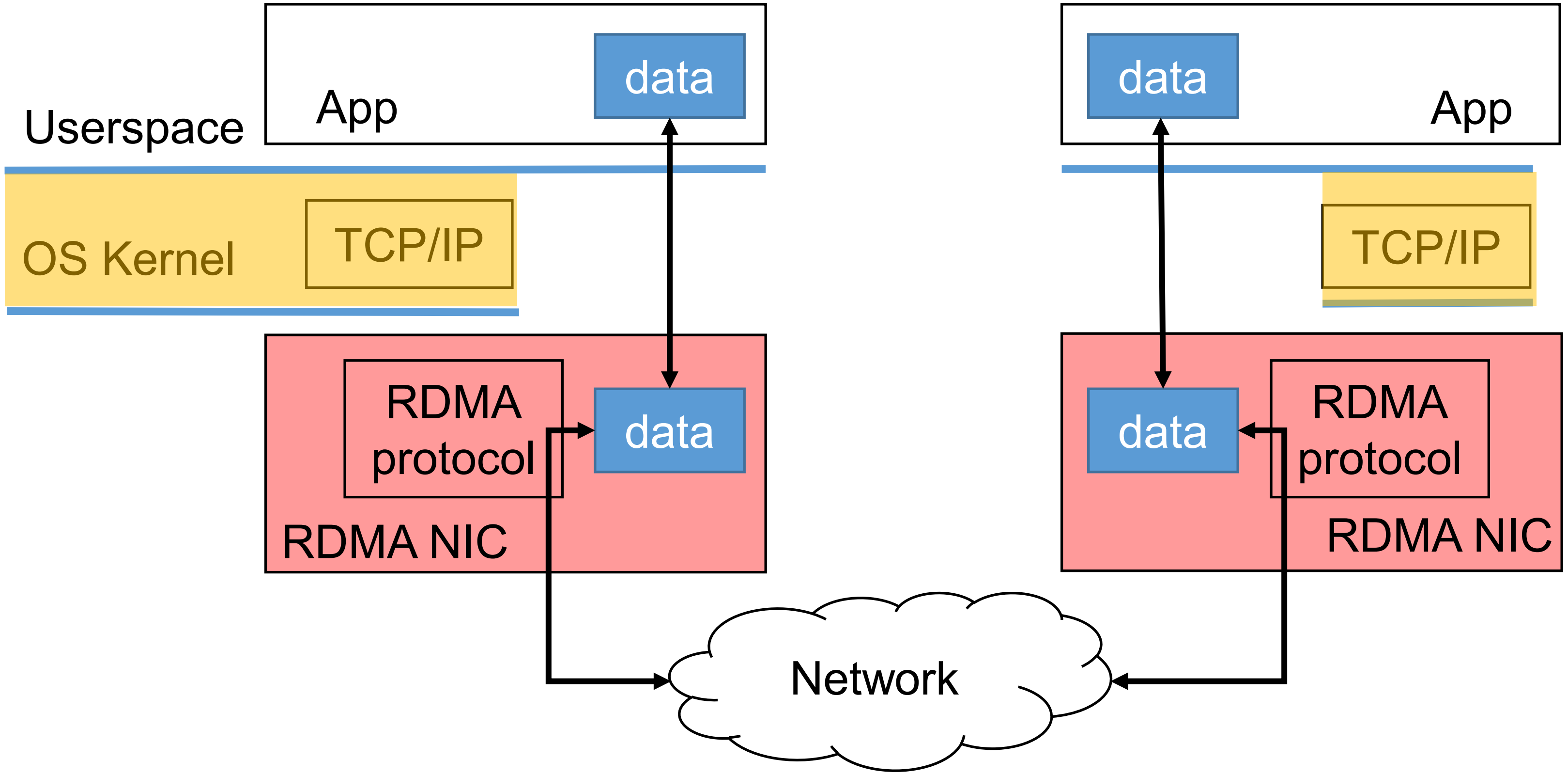
Low-latency and high-throughput network communication to support distributed applications.

Growing mismatch between increasing network speed and stalling CPU speedups.[1]

Hardware-based kernel bypass network for μs-scale latency.

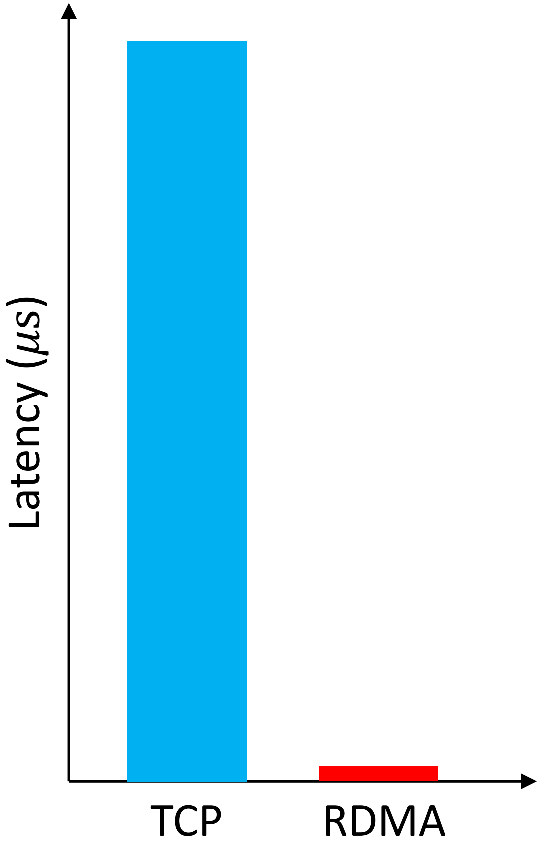
[1] https://yusur.tech/whitepaper/whitePaper2021

RDMA Fits the Datacenter Networking Demands



RDMA: Remote Direct Memory Access

- Bypass kernel
- Hardware offload
- Zero copy



- Lower latency & Higher throughput & Reduced CPU consumption

RDMA Becomes Essential in the Cloud

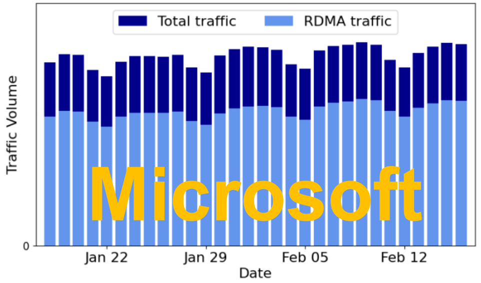
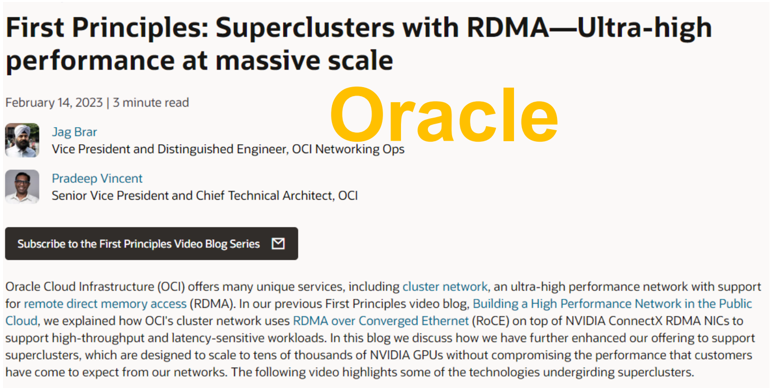


Figure 1: Traffic statistics of all Azure public regions between January 18 and February 16, 2023. Traffic was measured by collecting switch counters of server-facing ports on all Top of Rack (ToR) switches. Around 70% of traffic was RDMA.



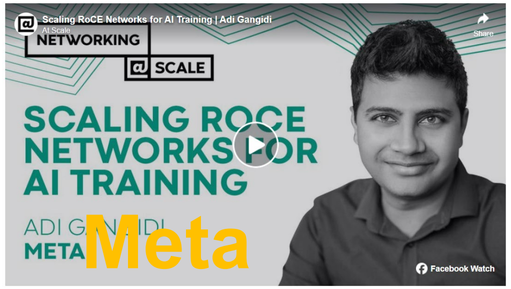
First Principles: Superclusters with RDMA—Ultra-high performance at massive scale

February 14, 2023 | 3 minute read

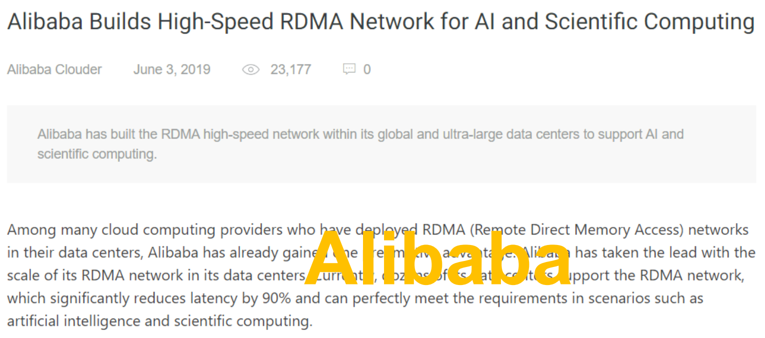
Jag Brar
Vice President and Distinguished Engineer, OCI Networking Ops

Pradeep Vincent
Senior Vice President and Chief Technical Architect, OCI

Oracle Cloud Infrastructure (OCI) offers many unique services, including cluster network, an ultra-high performance network with support for remote direct memory access (RDMA). In our previous First Principles video blog, Building a High Performance Network in the Public Cloud, we explained how OCI's cluster network uses RDMA over Converged Ethernet (RoCE) on top of NVIDIA ConnectX RDMA NICs to support high-throughput and latency-sensitive workloads. In this blog we discuss how we have further enhanced our offering to support superclusters, which are designed to scale to tens of thousands of NVIDIA GPUs without compromising the performance that customers have come to expect from our networks. The following video highlights some of the technologies undergirding superclusters.


Scaling RoCE Networks for AI Training | Adi Gangidi
At Scale
Share
NETWORKING
@ SCALE
SCALING ROCE NETWORKS FOR AI TRAINING
ADI GANGIDI
META
Meta
Facebook Watch

In this talk we provide an overview of Meta's RDMA deployment based on RoCEV2 transport for supporting our production AI Training infrastructure. We will shed light on how we designed our infrastructure to both maximize raw performance and consistency that is fundamental for the workload. We will talk about the challenges we solved in Routing, Transport and Hardware layers we solved along the way to scale our infrastructure. We will also touch on opportunities that remain in this space to make further progress over the next few years.



Alibaba Builds High-Speed RDMA Network for AI and Scientific Computing

Alibaba Clouder   June 3, 2019   23,177   0

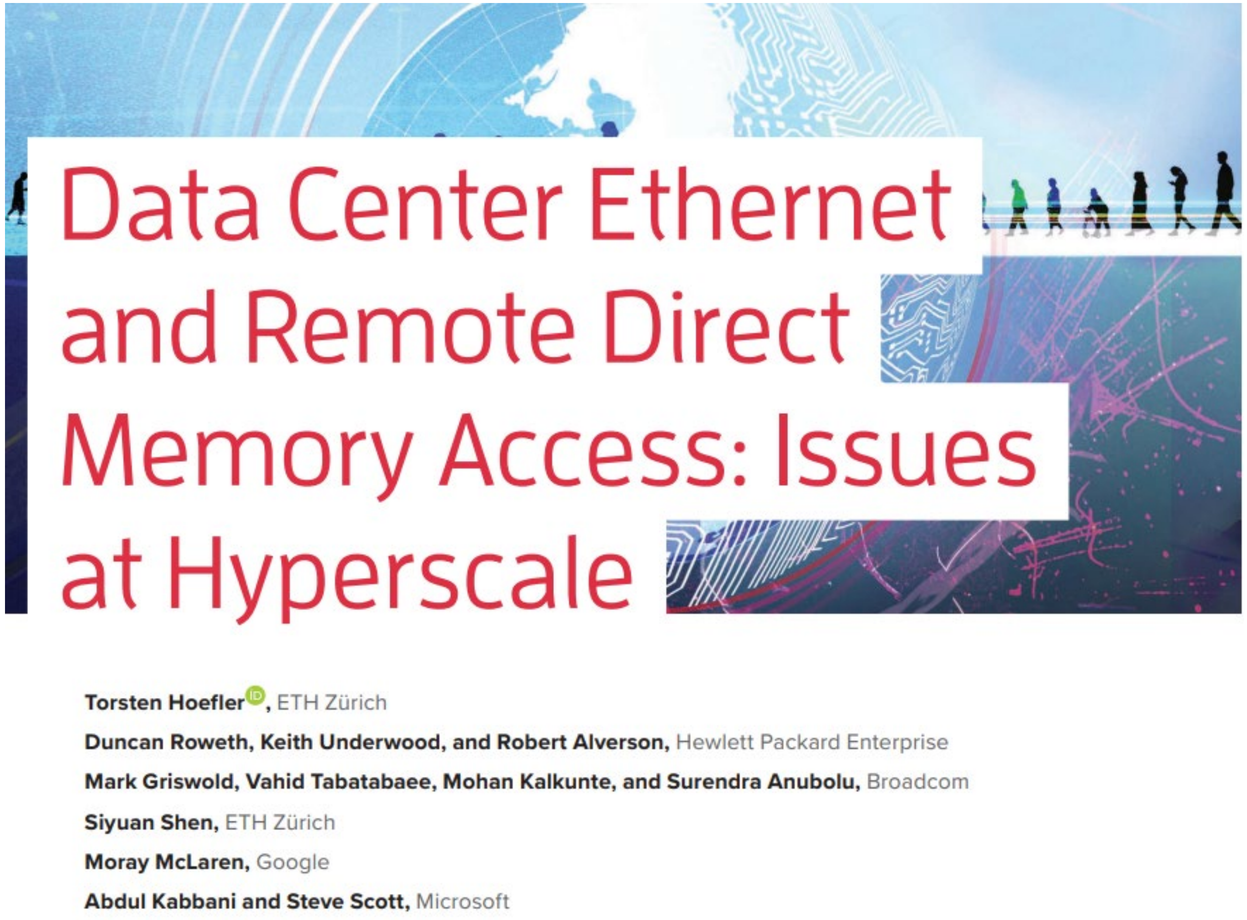
Alibaba has built the RDMA high-speed network within its global and ultra-large data centers to support AI and scientific computing.

Among many cloud computing providers who have deployed RDMA (Remote Direct Memory Access) networks in their data centers, Alibaba has already gained the preemptive advantage. Alibaba has taken the lead with the scale of its RDMA network in its data centers. Currently, dozens of its data centers support the RDMA network, which significantly reduces latency by 90% and can perfectly meet the requirements in scenarios such as artificial intelligence and scientific computing.

Cloud service providers heavily deploy RDMA in their datacenters.

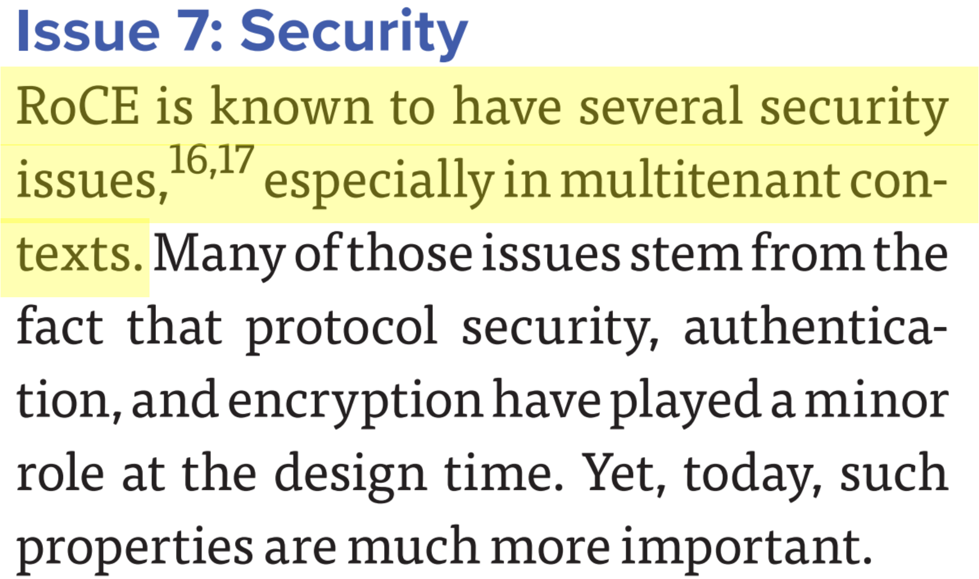
However, RDMA is originally designed for HPC.

RDMA Has Security Issues



Data Center Ethernet and Remote Direct Memory Access: Issues at Hyperscale

Torsten Hoefler, ETH Zürich
Duncan Roweth, Keith Underwood, and Robert Alverson, Hewlett Packard Enterprise
Mark Griswold, Vahid Tabatabaee, Mohan Kalkunte, and Surendra Anubolu, Broadcom
Siyuan Shen, ETH Zürich
Moray McLaren, Google
Abdul Kabbani and Steve Scott, Microsoft



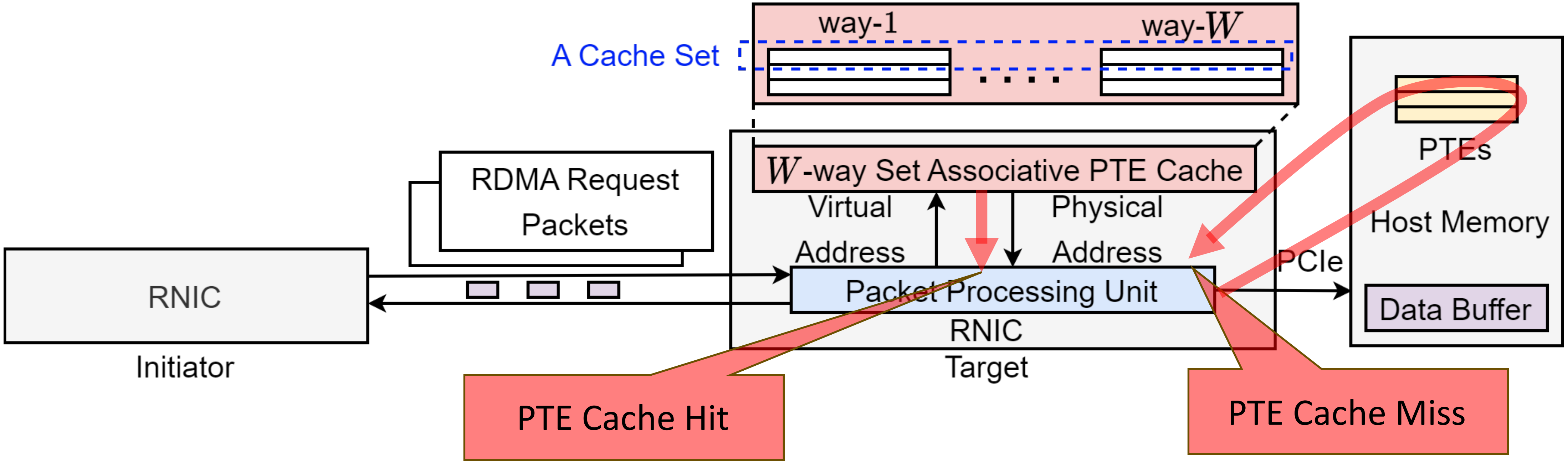
Issue 7: Security

RoCE is known to have several security issues,[16,17] especially in multitenant contexts. Many of those issues stem from the fact that protocol security, authentication, and encryption have played a minor role at the design time. Yet, today, such properties are much more important.

RDMA exposes **new attack surface** to malicious tenants.

An Attack Surface: PTE Caches in RDMA NIC

- RDMA NIC (RNIC) maintains page table entries (PTEs) for zero-copy data transfer.
- Frequently-accessed PTEs are cached on the chip.



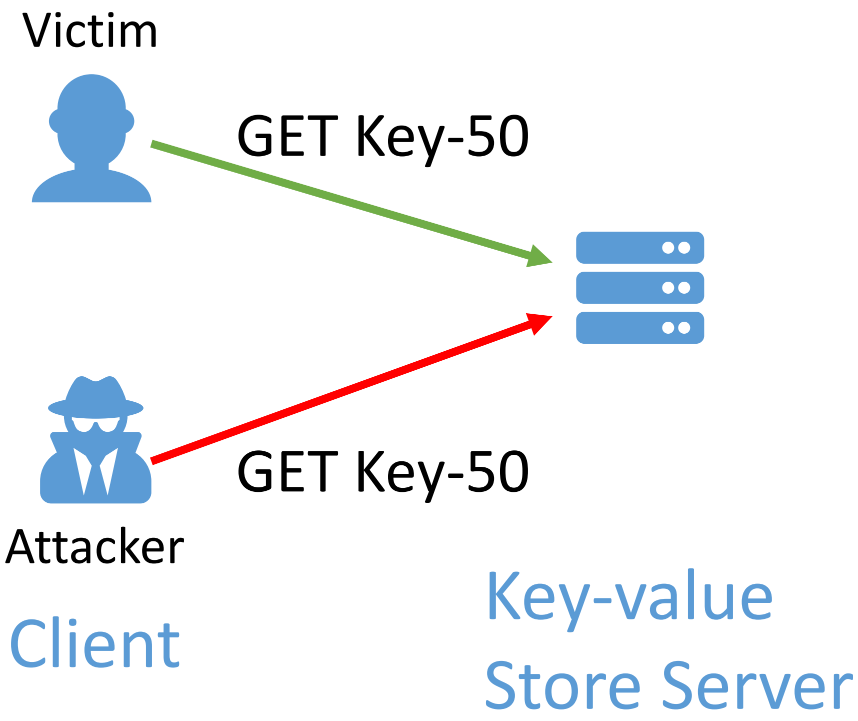
PTE Cache Hit: one on-chip memory read latency.

PTE Cache Miss: at least one PCIe round-trip latency.

Can be used for **timing-based cache side-channel attack**.

RNIC Cache Side-channel Attack Threat Model

- ❑ Server hosts data in memory exported via RDMA
- ❑ The attacker and the victim can be on different machines
- ❑ The attacker wants to learn the access pattern of the victim
- ❑ The attacker cannot observe the network traffic

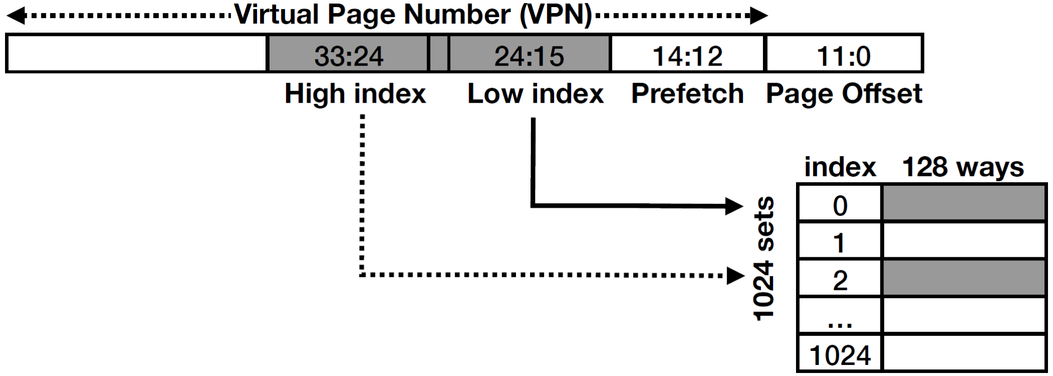


Attacker: Did the victim access Key-50?

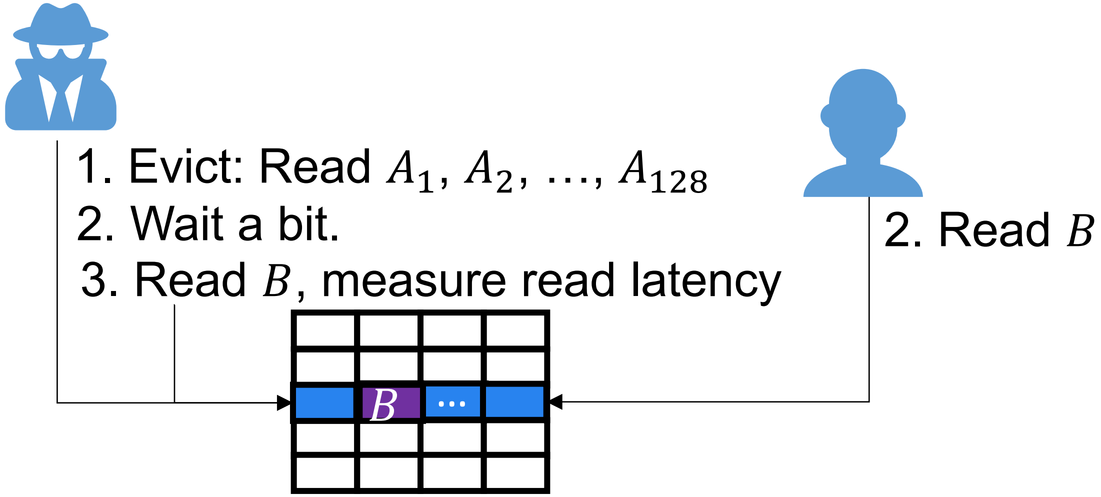
RCSCA Procedure

1. Reverse-engineer RNIC PTE Cache Structure.

2. Evict+Reload Side-channel Attack.



PTE cache structure of Mellanox CX-4 RNIC

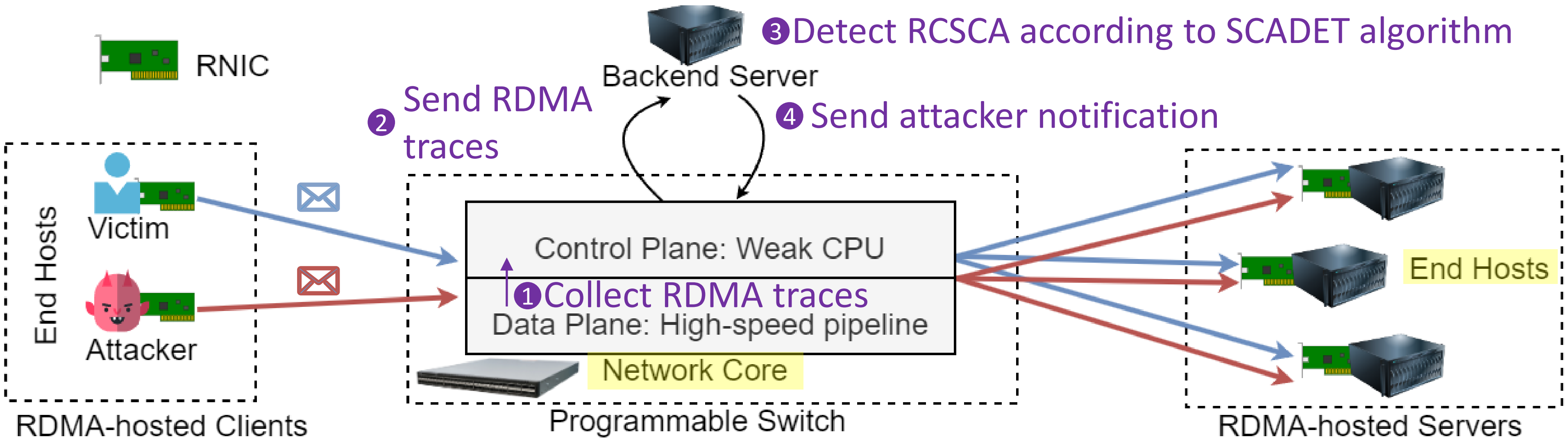

1. Evict: Read $A_1, A_2, \ldots, A_{128}$
2. Wait a bit.
3. Read B, measure read latency
2. Read B
B
...

Server-side PTE Cache

If read latency < Threshold, Victim accessed B.

Existing Switch-centric RCSCA Detection

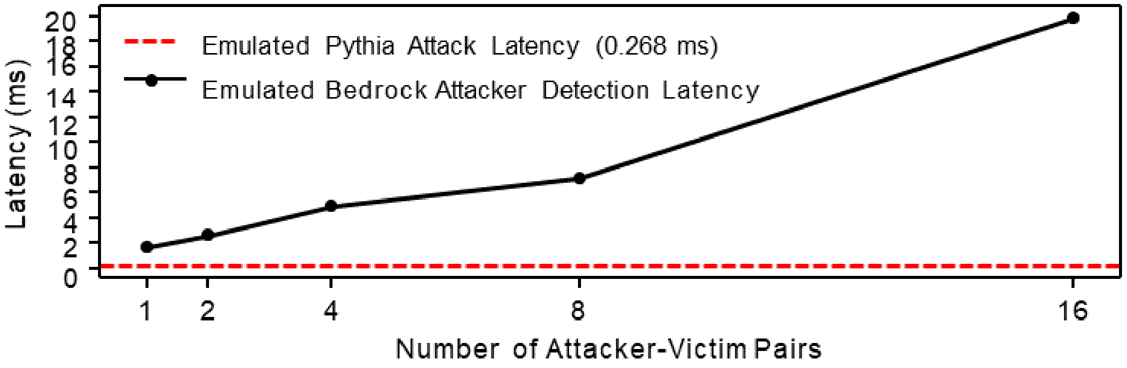
- End-host CPU cannot detect RCSCA due to CPU bypass.
- End-host RNIC cannot provide programmable compute to detect RCSCA.
- Network-core programmable switch relies on a backend server to detect RCSCA[1].



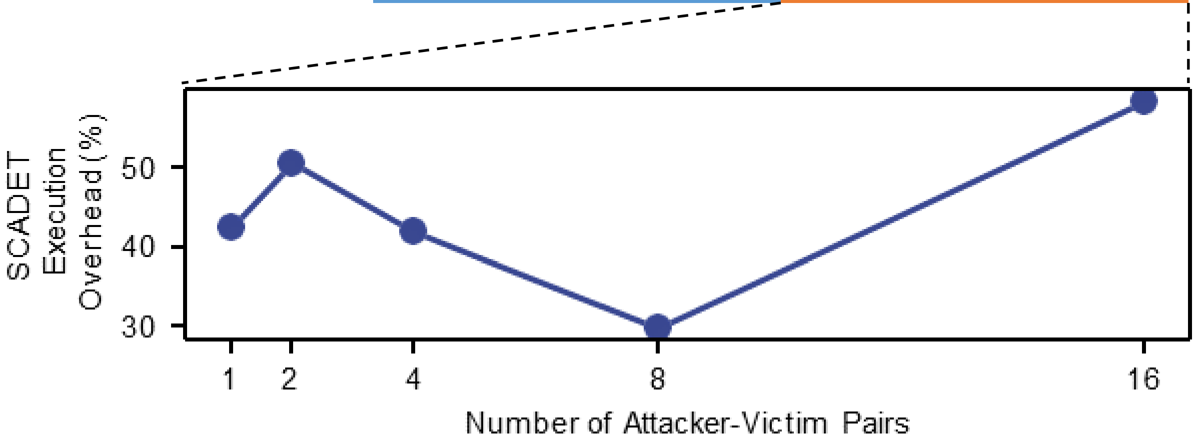
Is the switch-centric design good enough?

[1] Xing, Jiarong, et al. "Bedrock: Programmable Network Support for Secure {RDMA} Systems." 31st USENIX Security Symposium (USENIX Security 22). 2022.

Profiling Switch-centric RCSCA Detection



Detection latency: Network Transfer | SCADET Execution



- ❑ Switch-centric RCSCA detection (Bedrock[1]) is slower than the attacker (Pythia[2]).
  - ❑ Sensitive information may be leaked.
- ❑ Bedrock becomes much slower if there are many attacker-victim pairs.
- ❑ SCADET execution time contributes a lot to the detection latency.

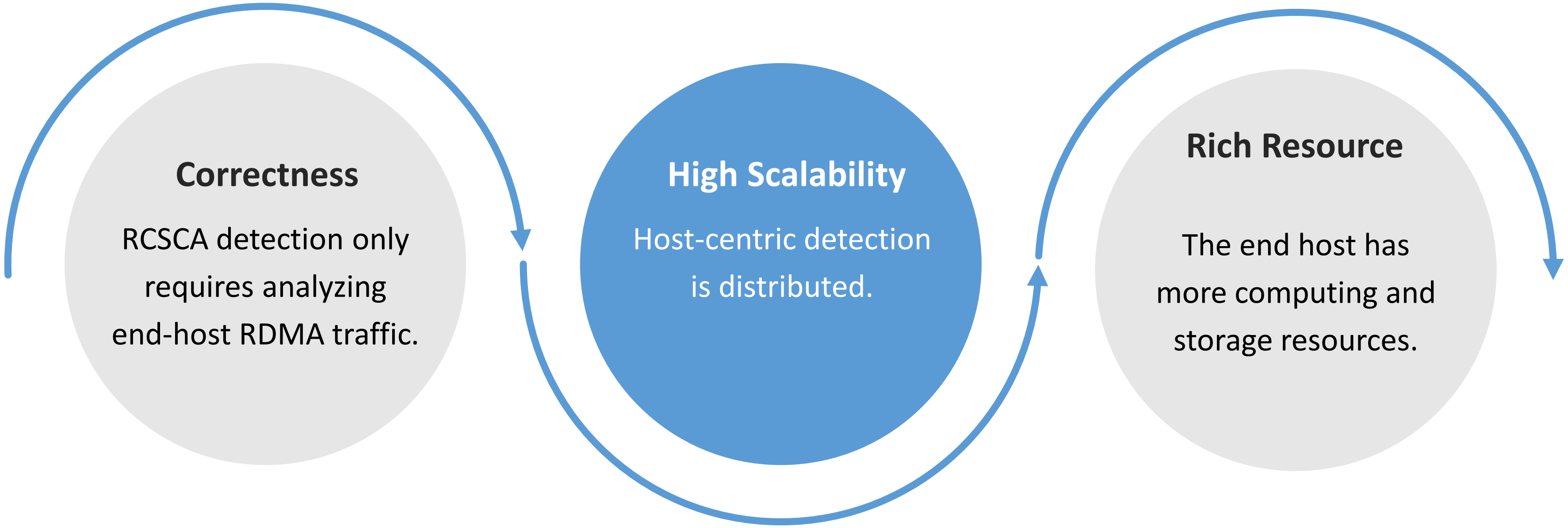
How to minimize the detection latency to minimize the information leakage?

[1] Xing, Jiarong, et al. "Bedrock: Programmable Network Support for Secure {RDMA} Systems." 31st USENIX Security Symposium (USENIX Security 22). 2022.

[2] Tsai, Shin-Yeh, Mathias Payer, and Yiying Zhang. "Pythia: remote oracles for the masses." 28th USENIX Security Symposium (USENIX Security 19). 2019.

Our Design: DPU-driven RCSCA Detection

Insight: Host-centric RCSCA Detection Is Better



Correctness

RCSCA detection only requires analyzing end-host RDMA traffic.

High Scalability

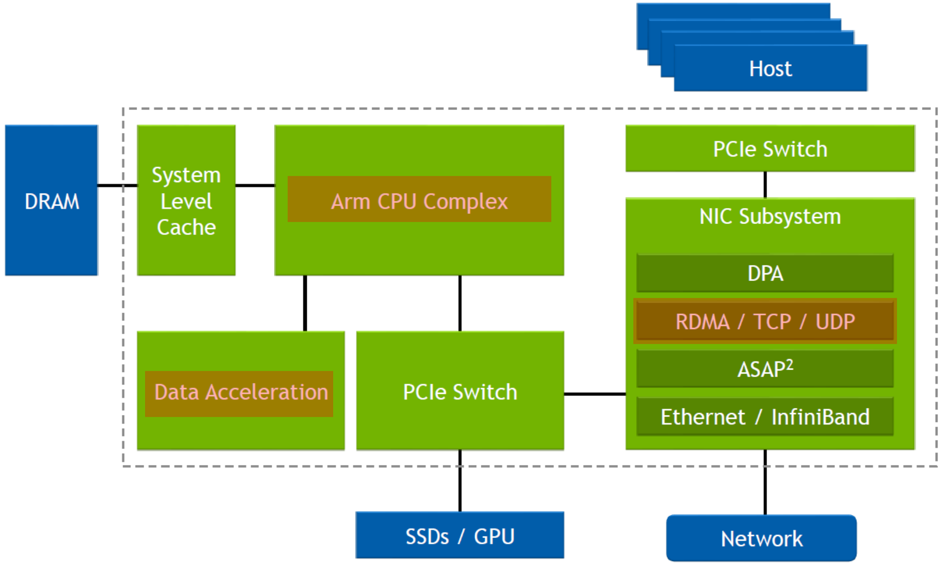
Host-centric detection is distributed.

Rich Resource

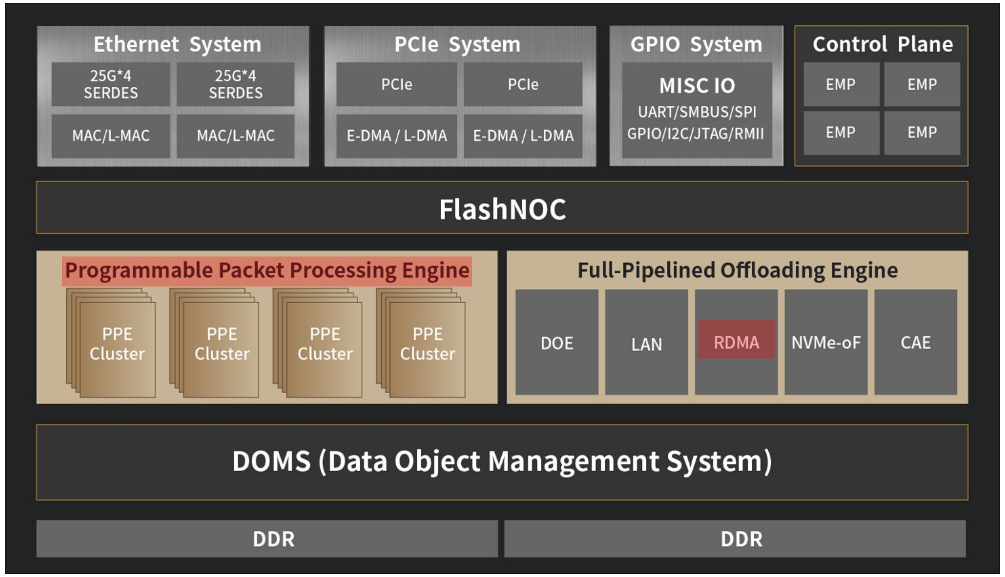
The end host has more computing and storage resources.

However, the traditional RNIC cannot be architected to detect RCSCA.

Key Opportunity: Data Processing Unit



NVIDIA DPU architecture[3]

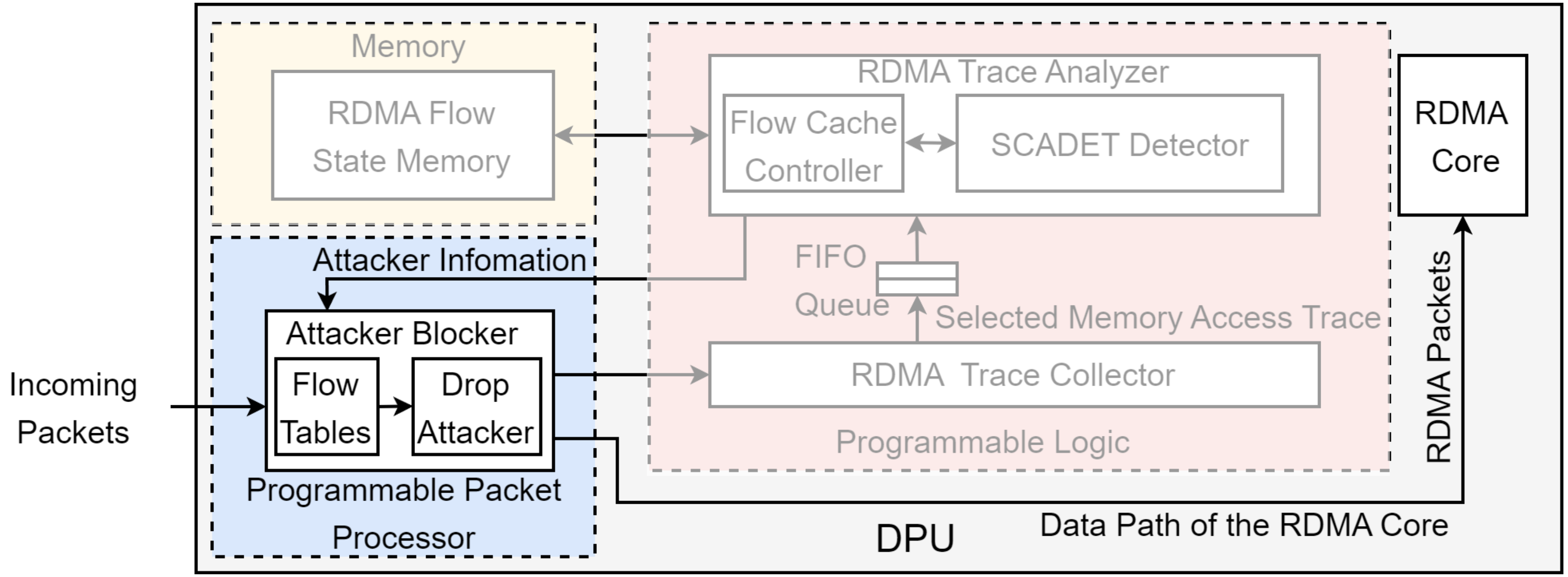


YUSUR K2Pro DPU architecture[2]

- Data Processing Unit (DPU)： combines the functionality of traditional RNICs with programmable compute and storage.
- We choose the FPGA-based DPU to demonstrate the DPU-driven RCSCA detector.

[3] Burstein, Idan. "Nvidia data center processing unit (dpu) architecture." 2021 IEEE Hot Chips 33 Symposium (HCS). IEEE, 2021.
[2] https://www.yusur.tech/dpu/K2-Pro

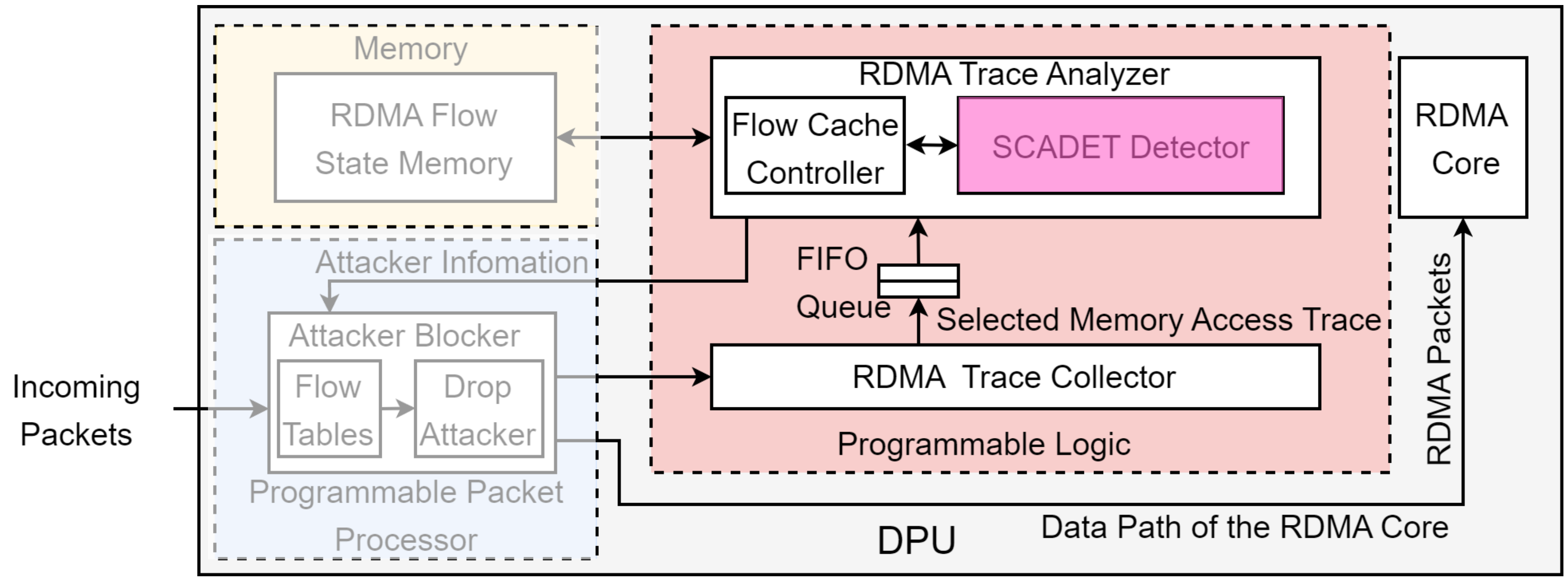
Design of DPU-driven RCSCA Detector



The attacker blocker is located in the programmable packet processor.

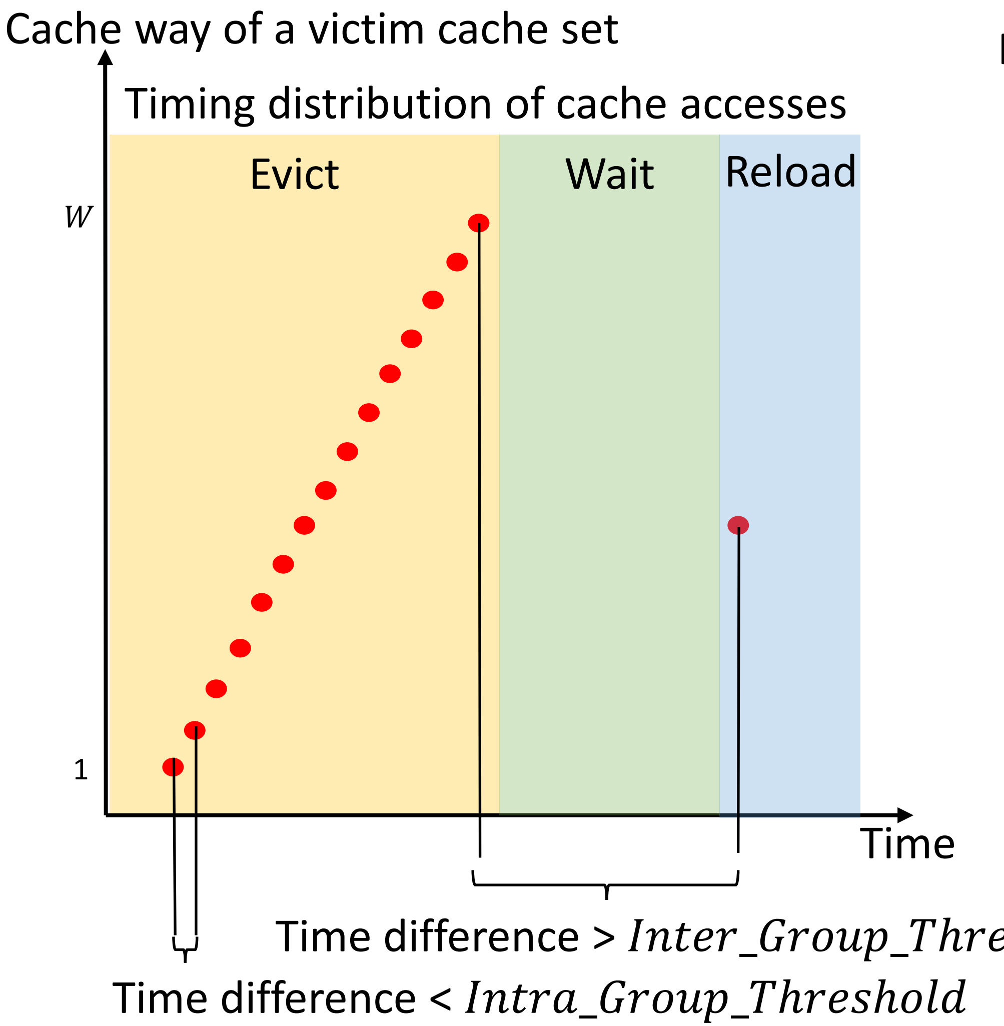
- Commodity programmable packet processor has a flow table.
- Commodity programmable packet processor supports the "Match+Action".

Design of DPU-driven RCSCA Detector

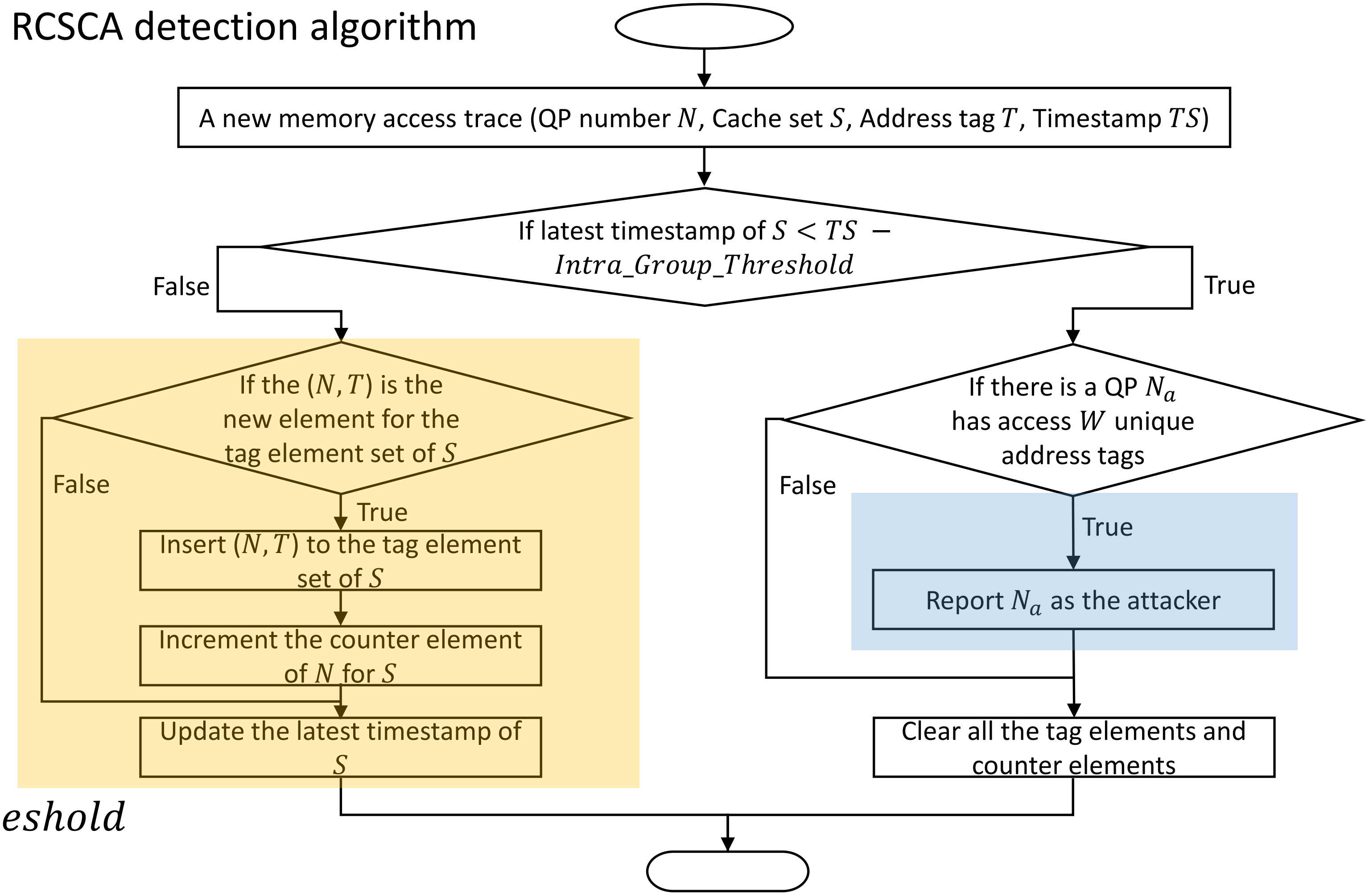


- We position the trace analyzer off the critical data path of the RDMA core to minimize RDMA performance overhead.
- The SCADET detector is highly optimized to match the incoming request speed.

SCADET Algorithm



RCSCA detection algorithm



A new memory access trace (QP number N, Cache set S, Address tag T, Timestamp TS)

If latest timestamp of $S < TS - Intra\_Group\_Threshold$

False

If the (N, T) is the new element for the tag element set of S

False

True

Insert (N, T) to the tag element set of S

Increment the counter element of N for S

Update the latest timestamp of S

True

If there is a QP N_a has access W unique address tags

False

True

Report N_a as the attacker

Clear all the tag elements and counter elements

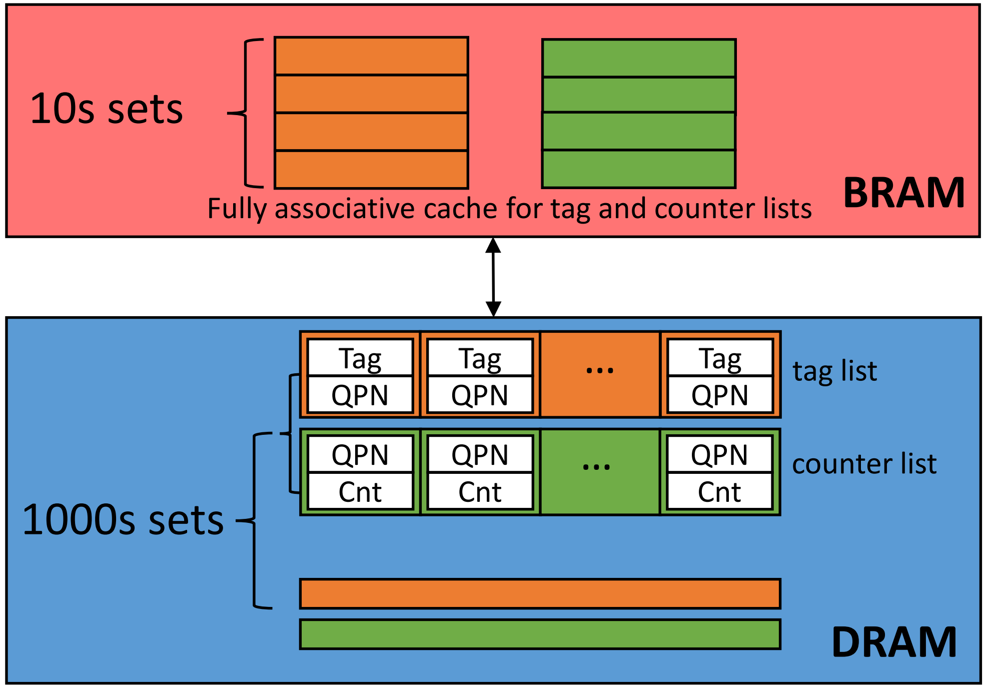
Accelerating the SCADET algorithm is required.

[4] Sabbagh, Majid, et al. "Scadet: A side-channel attack detection tool for tracking prime-probe." 2018 IEEE/ACM International Conference on Computer-Aided Design (ICCAD). IEEE, 2018.

SCADET Implementation： Storage Optimizations

Requirements:

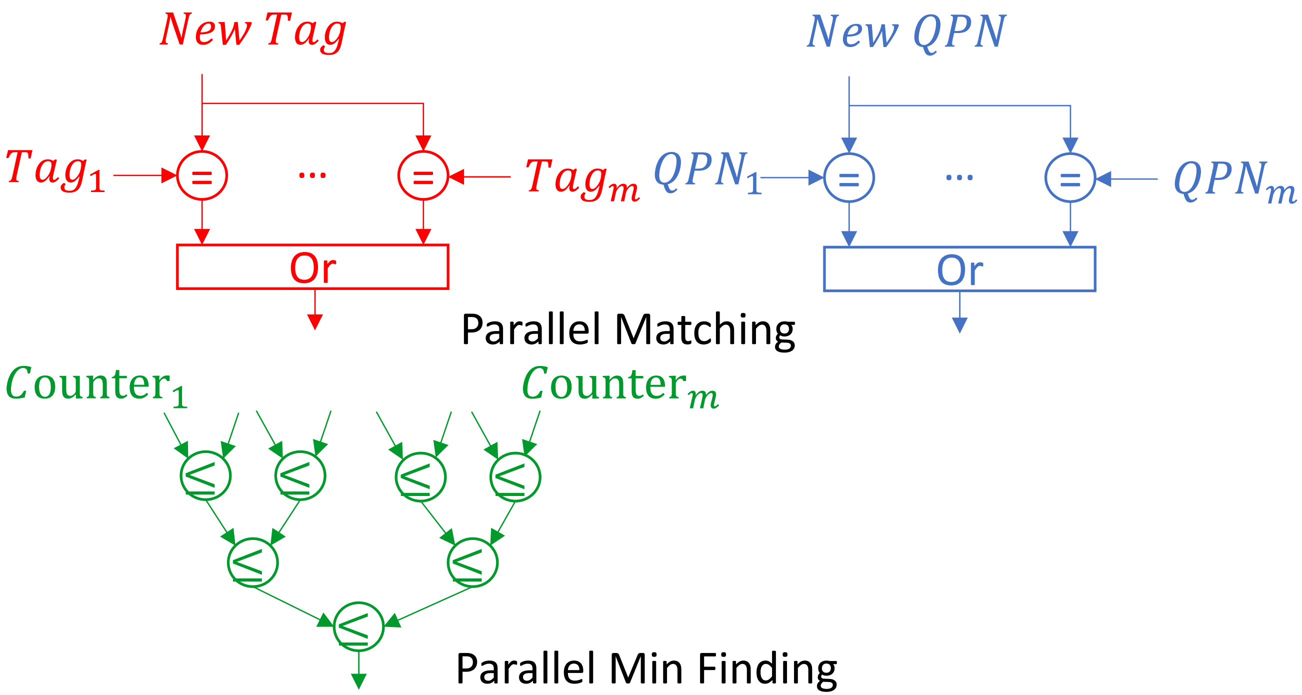
- Hardware-friendly data structure： fixed-length lists in continuous memory region.
- Large capacity for 1000s sets： DRAM.
- Fast access： BRAM-based flow cache.



SCADET Implementation： Computation Optimizations

```
If ($Time\_Diff < Intra\_Group\_Threshold$) {
    Search_Tag_List: For (…) {For (…)}
    Search_Counter_List: For (…) {For (…)}
    If ($Insert\_New\_Tag$ && $Tag\_List$ $is$ $Full$) {
        Search_QP_with_Minimum_Counter {
            For (…) {For (…)}
    }}
}
```

Critical path of the SCADET logic



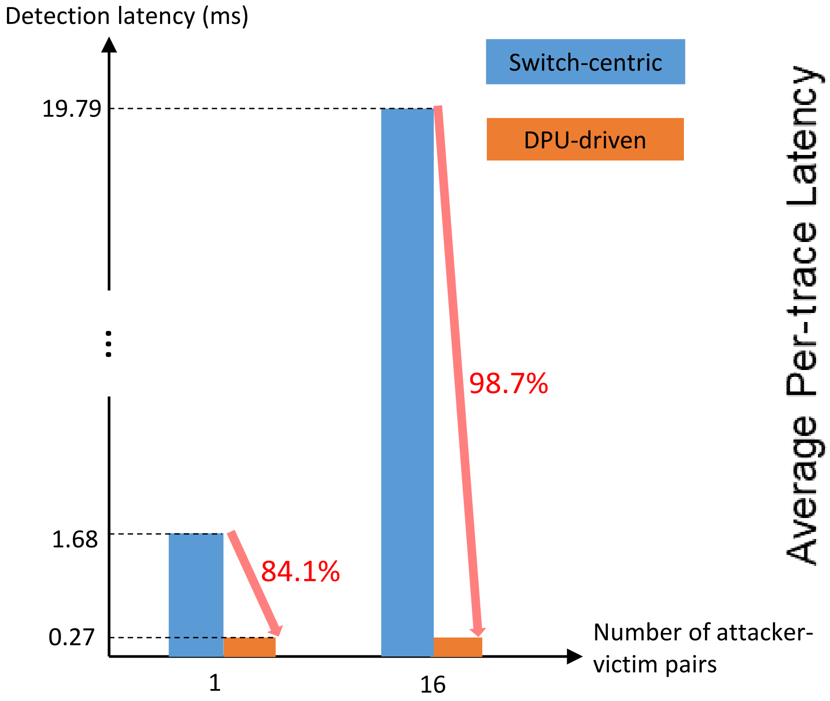
We use FPGA’s logic flexibility to parallelize the loops in the critical path.

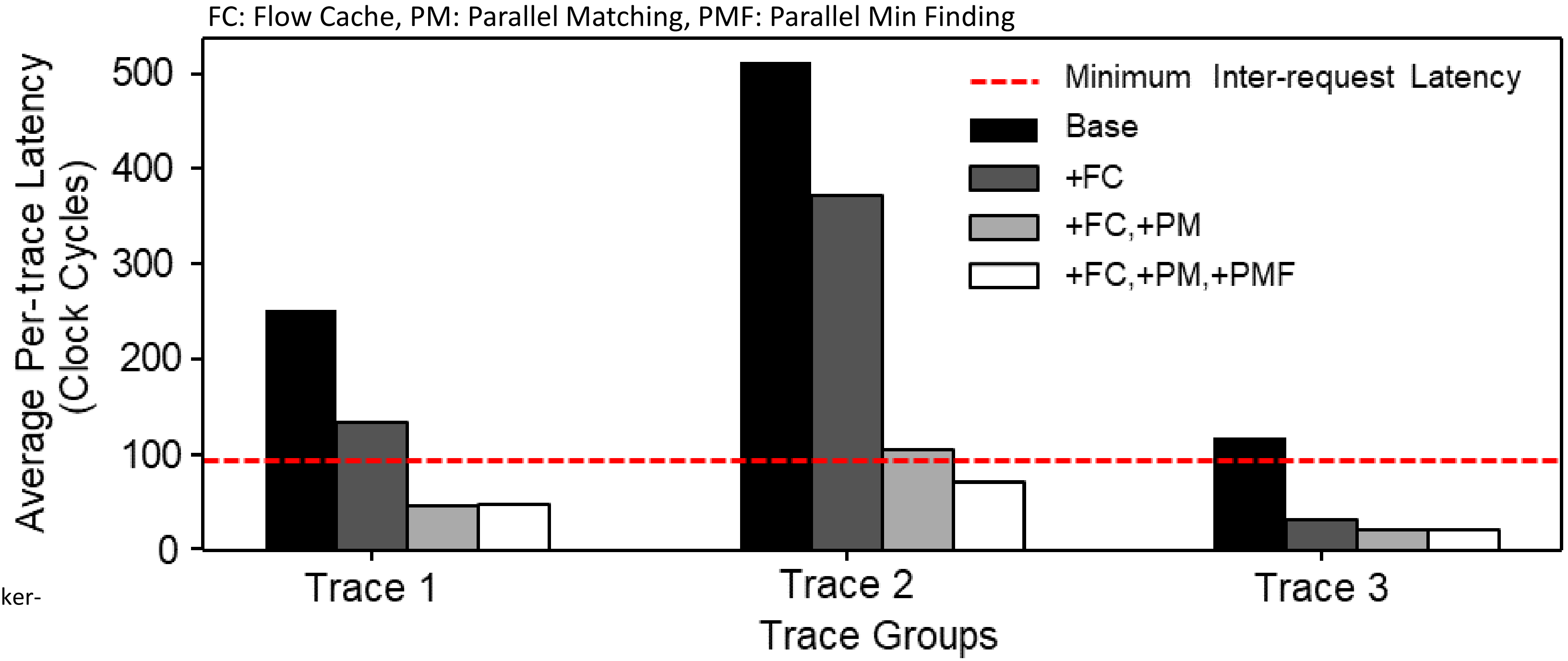
Evaluation

Experiments Setup

- **RNIC model**: We assume the RDMA PTE cache of the NVIDIA Mellanox ConnectX-4131A has 1024 sets and 128 ways per set according to the Pythia’s reverse engineering result.
- **Switch-centric baseline**: We set up a real system to emulate and evaluate Bedrock.
- **Our DPU-driven RCSCA detector:** We develop the trace analyzer with Xilinx HLS language and measure its cycle-accurate performance with Vitis 2022.1.
- **Traces**:
  - Trace-1： Evict+Reload memory traces
  - Trace-2： prefills the tag list before the injection of the Evict+Reload memory traces (Trace-1)， the worst workload for the DPU-driven detector.
  - Trace-3： memory access trace following uniform distribution that emulates a regular user

Low RCSCA Detection Latency

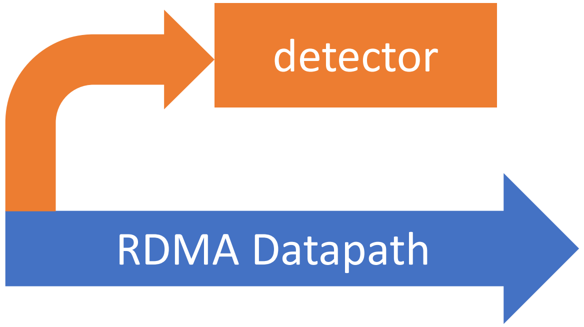




- DPU-driven design reduces the detection latency by more than 84.1%.
- DPU-driven design leaks zero sensitive information.
- The three proposed architectural optimizations for FPGA-based SCADET are essential.

Zero Performance Overhead and Small FPGA Resource Consumption

☺ Performance overhead: 0



The detector is off the RDMA datapath.

☺ FPGA resource overhead: small, affordable for datacenter-grade FPGA.

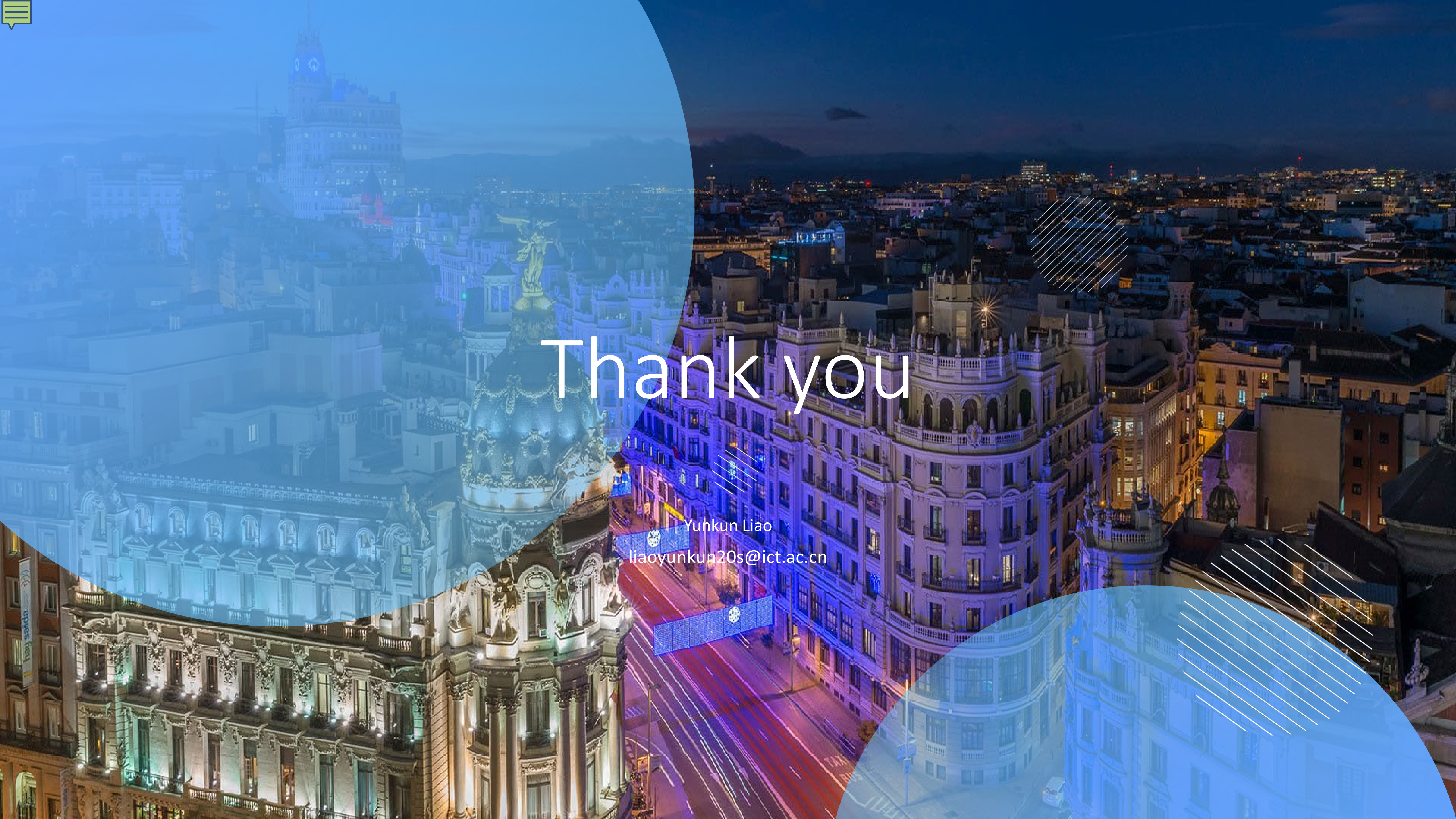
Resource type	Percentage (based on VU9P)
Flip Flop	1%
Lookup Table	6%
BRAM	~0
DSP	~0

Conclusion

Conclusion and Outlook

1. RDMA security is a critical concern, especially given the widespread deployment of RDMA in public cloud.
2. We have identified the weaknesses inherent in the current switch-centric designs for RDMA security.
3. We advocate for the benefits of a host-centric approach, driven by the capabilities of emerging Data Processing Units (DPUs).
4. Our demonstration of the substantial performance improvement in a DPU-driven design, exemplified by the detection of RNIC cache side-channel attacks, underscores its potential.

Takeaway: Smart DPU-driven Edge, Dumb Core.



Thank you

Yunkun Liao

liaoyunkun20s@ict.ac.cn