

# SmartNICs and Data Processing Units (DPUs)

## Computer Networks

Sina Daneshgar  
Mohammadmohsen Abbaszadeh

Sharif University of Technology

Feb 2026

# Seminar Overview

- 1 Introduction and Global Context
- 2 The Infrastructure Tax
- 3 Detailed Architecture

- 4 Programming Paradigms
- 5 Real-World Applications
- 6 Performance Simulation
- 7 Summary and Future Outlook

# The Evolution of Compute

- CPU** General purpose, high flexibility, but high latency for simple repetitive tasks.
- GPU** Specialized for parallel processing (Graphics, AI).
- DPU** Specialized for **Data-Centric** tasks (Moving, Processing, Securing data).

## Why now?

Moore's Law is slowing down, while networking speeds are exploding (10G → 100G → 400G). The CPU can no longer keep up with the networking interrupt load.

# Quantifying the "Infrastructure Tax"

Data centers spend 20-30% of their total compute power just on "tax" tasks:

- **Virtualization:** OVS (Open vSwitch) overhead, encapsulation (VXLAN, Geneve).
- **Storage:** NVMe-oF (NVMe over Fabrics) protocol translation and management.
- **Security:** Distributed Firewalls, Micro-segmentation, and Wire-speed TLS/IPsec.

## The Bottleneck

Every packet processed by the CPU is a cycle stolen from the user application (Application Stall).

# Deep Dive: What's inside a DPU?

Unlike a standard NIC, a DPU (like NVIDIA BlueField or AMD Pensando) contains:

- **ARM/RISC-V Cluster:** Runs a standard Linux OS for management and control.
- **Network Acceleration Engine:** Programmable hardware for parsing and switching.
- **Hardware Engines:**
  - **Crypto:** Hardware-accelerated IPsec/TLS and Disk encryption.
  - **Storage:** VirtIO-blk acceleration and Compression engines.
  - **Timing:** PTP (Precision Time Protocol) for finance/telecom.
- **On-board RAM:** For local state, lookups, and buffering ( 16-32GB DDR4/5).

- **P4 (Programming Protocol-independent Packet Processors):**
  - Define custom headers and parsers.
  - Programmable Match-Action tables inside the NIC silicon.
  - Allows for "Protocol Independence."
- **Implementation:** Our repository (`examples/02-p4`) shows a basic L2 forwarding logic defined in P4<sub>16</sub>.

- **eBPF (Extended Berkeley Packet Filter):**
  - Runs JIT-compiled bytecode inside the Linux kernel.
  - High safety (verified at load time) and extreme speed.
- **XDP (eXpress Data Path):**
  - An eBPF hook placed as early as possible (the NIC driver).
  - Can DROP, FORWARD, or REDIRECT packets *before* the kernel allocates a socket buffer (`sk_buff`).
- **Demo:** See our `examples/03-ebpf-xdp` filter for ICMP-drop results.

# Application 1: Cloud Resource Disaggregation

- **Traditional:** Storage and Compute locked in the same physical box.
- **DPU-based:** Storage is remote (over NVMe-oF), but the DPU makes it look like a local NVMe drive to the host OS.

## Benefit

Storage performance is identical to local SSDs, but with the flexibility of network-attached storage.

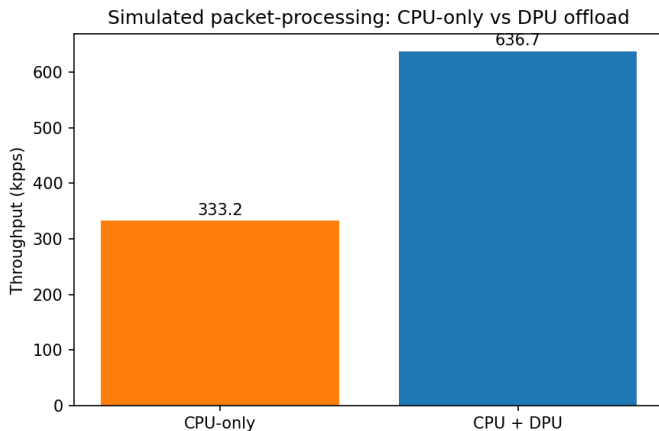


## Application 2: Bare Metal as a Service (BMaaS)

- Cloud providers (AWS, Azure) want to rent "Bare Metal" to users.
- **Problem:** How to monitor and isolate the user if they have full control?
- **Solution:** The DPU acts as the "Sidecar" manager. It manages networking and security *outside* the host, so the user cannot bypass it.

# Simulation: Offload Performance

We simulated the host CPU throughput gains when offloading work to a DPU.



Simulation results show  $> 50\%$  throughput increase in targeted workloads when logic is moved to the hardware pipeline.

# Current Challenges and Limitations

- **Vendor Lock-in:** No universal "Standard" for all DPUs yet (though P4 helps).
- **Complex Debugging:** Inspecting code running inside a NIC is harder than on a host.
- **Power Consumption:** High-end DPUs can consume over 75W-100W per card.

- **Offload is Mandatory:** At 100G+ speeds, host processing is no longer viable.
- **Isolation is Security:** DPUs create a "Hard Gap" between the user VM and the infrastructure provider.
- **The New Tier:** DPUs are becoming the third pillar of data center compute alongside CPUs and GPUs.

- DPUs recover "lost" CPU cycles for user applications.
- They provide hardware-level security isolation.
- Programming models (P4/eBPF) are maturing fast.

**Thank You!**

Questions?