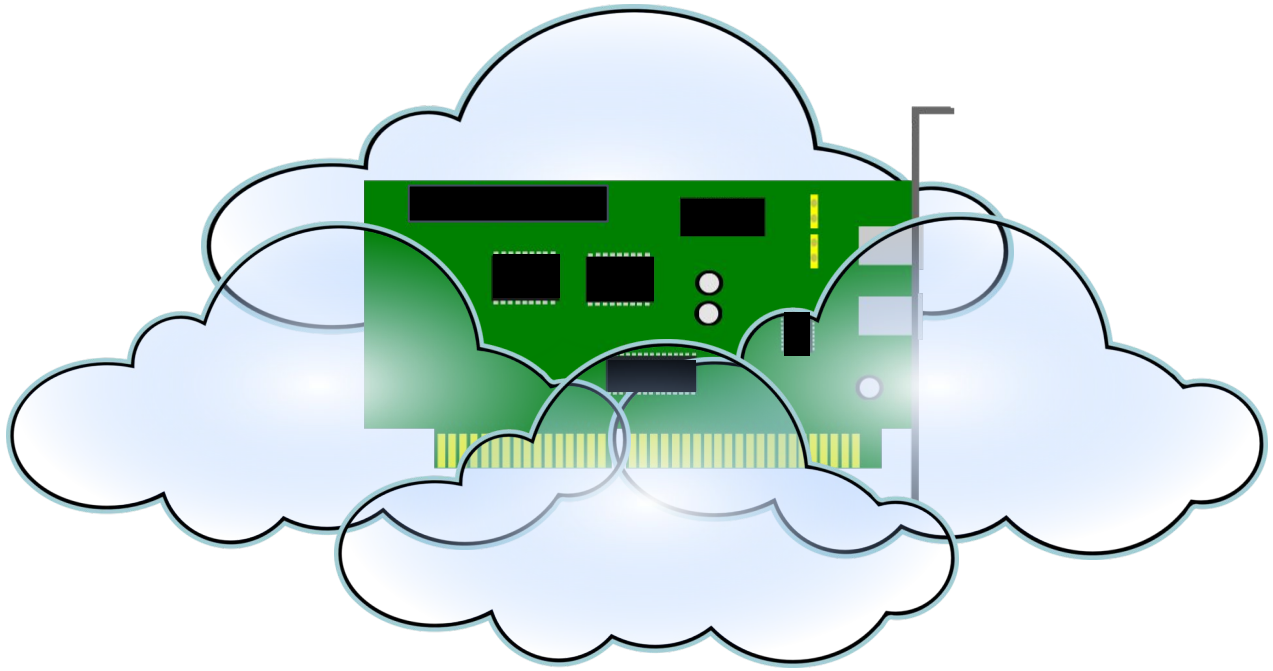
PENSANDO

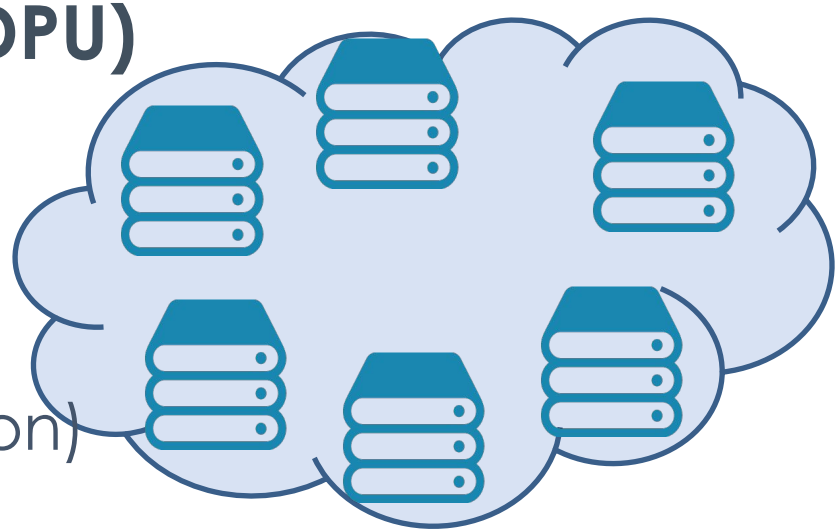
# Designing Cloud-grade DPUs



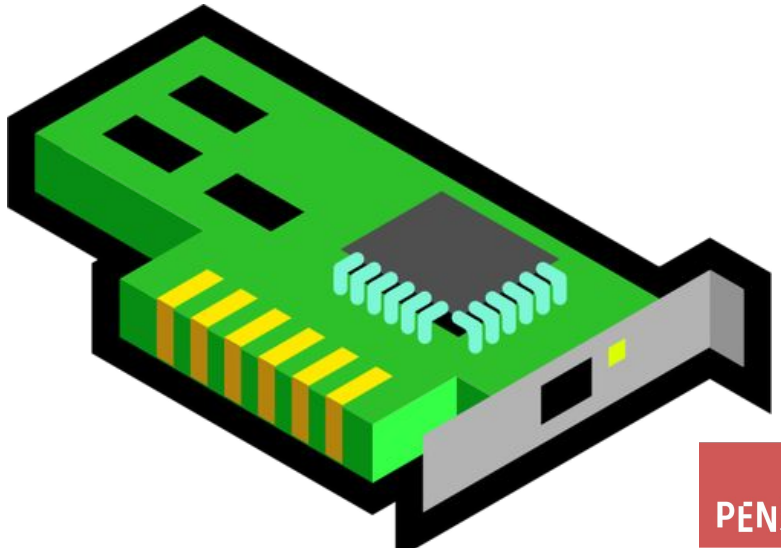
Mario Baldi

Fellow, Pensando Systems, Inc.

# The Cloud and the Data Processing Unit (DPU)



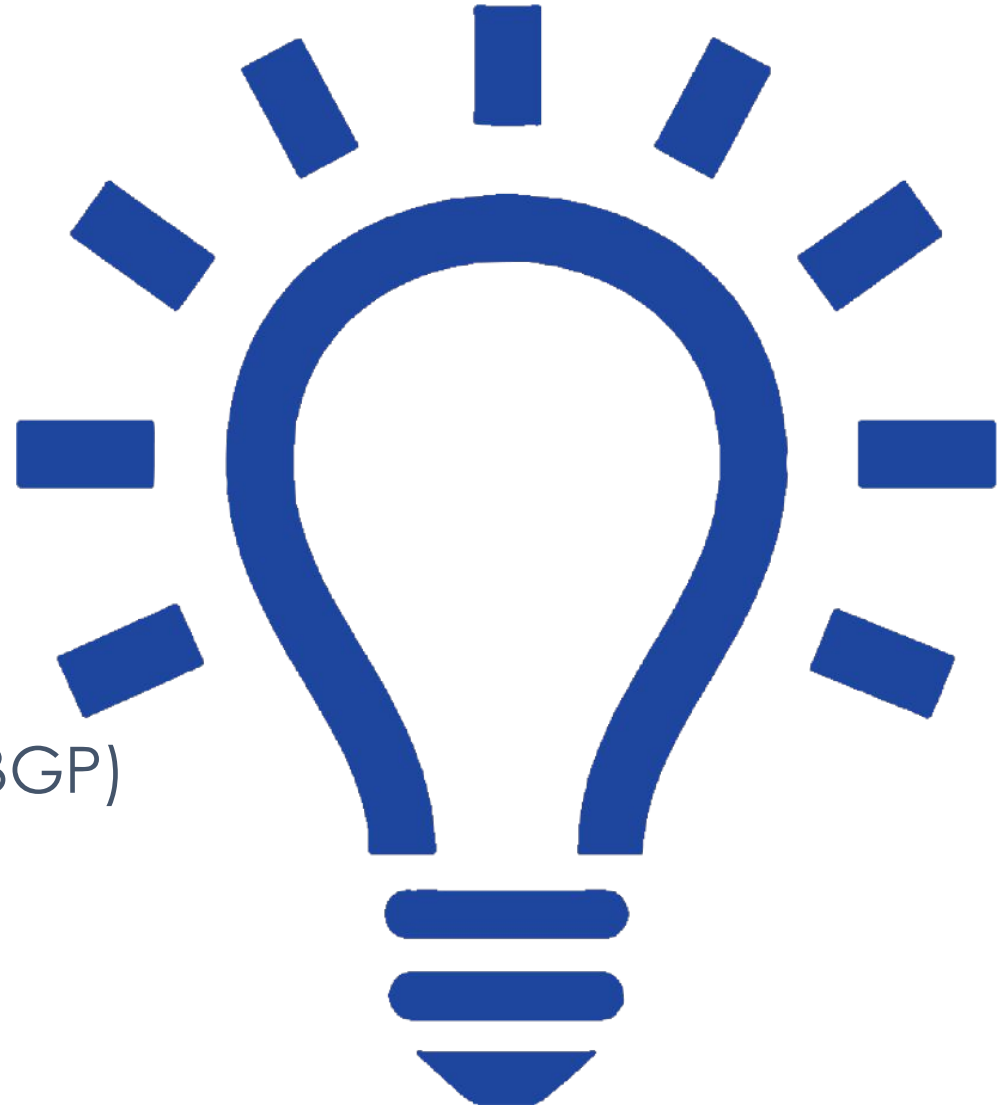
- Infrastructure services in addition to computing services
  - Networking (SDN)
  - Security (firewalling, IPsec tunnel termination)
  - Storage (NVMe-oF)
  - Observability (telemetry, packet capture)
- Traditionally executed on the CPU
  - Using valuable resources
- Offloading is appealing
  - Custom hardware
  - Data Processing Units (DPU)



PENSANDO

# New Use Cases and Opportunities

- Bare Metal as a Service (BMaaS)
- PCIe IO Emulation
  - Network devices (e.g., VirtIO)
  - Storage devices (e.g., NVMe)
- Tenant routing protocol offload (e.g., BGP)

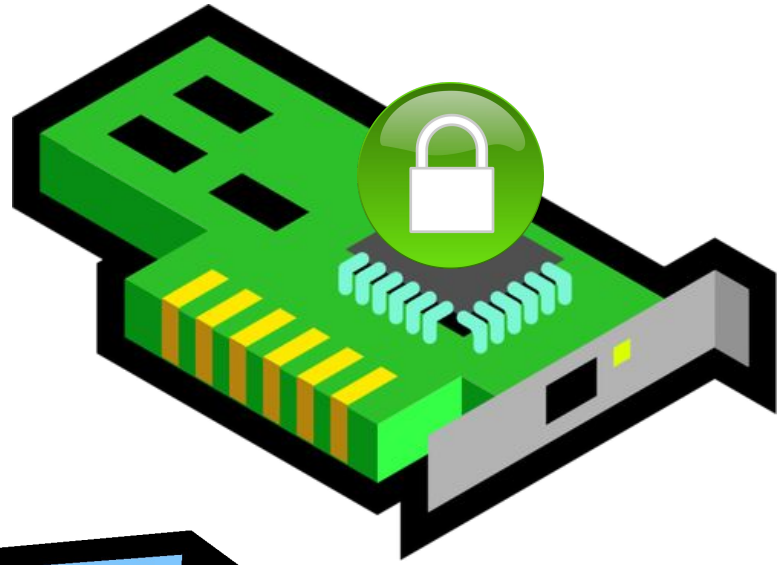




PENSANDO

# Cloud DPU Requirements

- Security
  - Secure execution environment
  - Protected from the network
  - Protected from the host
- Independence (from the host)
  - Operation and management
  - Control
  - Operation in bare metal scenarios





PENSANDO

# Cloud DPU Requirements

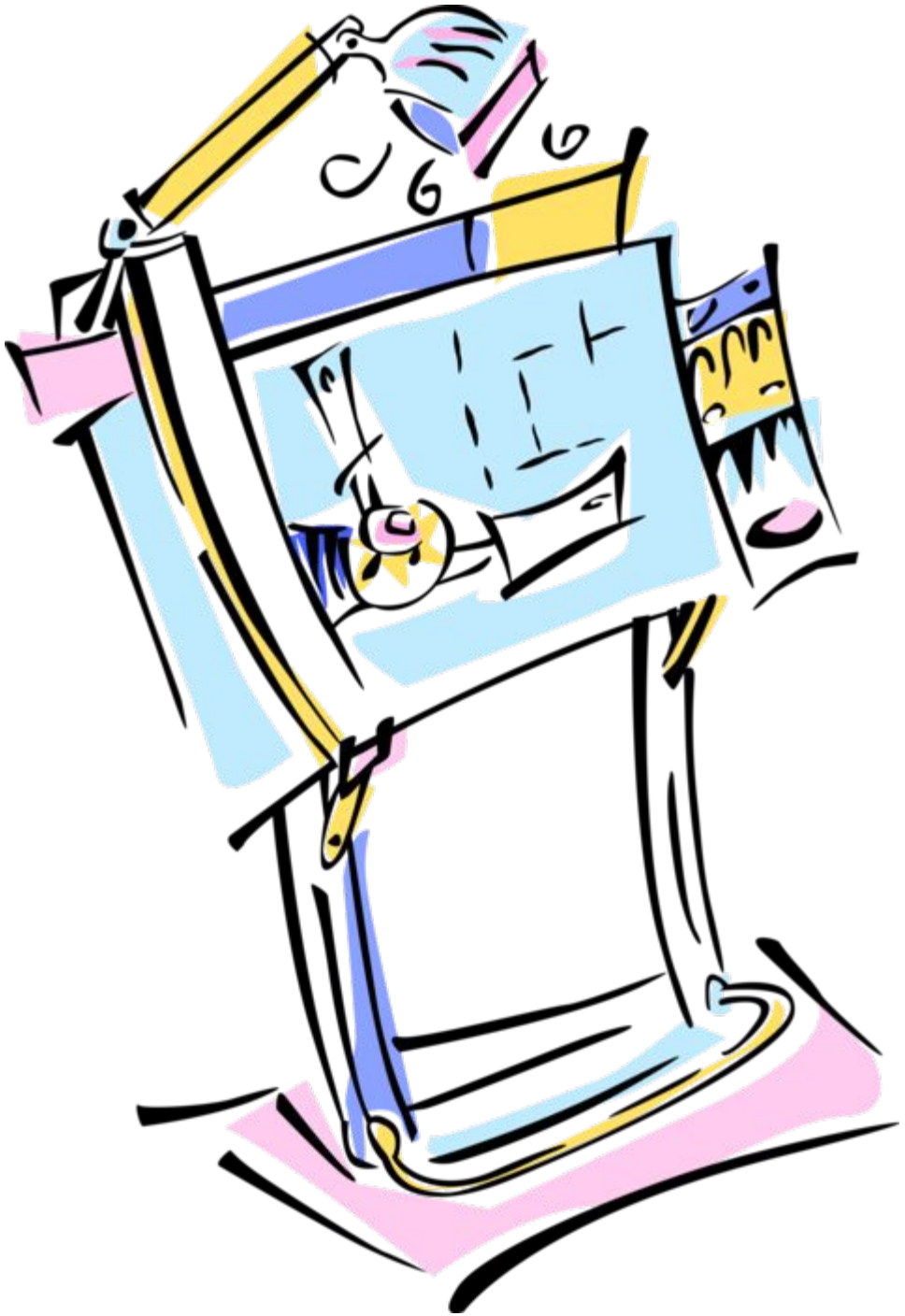
- Performance
- Scale
- Compute virtualization support
  - E.g., SRIOV
- Multi-tenancy with per-tenant service isolation
  - Traffic from one tenant cannot affect service to others
  - Network side and host side
- Network virtualization
- Flexibility - capability to evolve and adapt



PENSANDO

PENSANDO

# Design Implications and Options



# Programmability

Flexibility - Evolve and adapt

- General purpose processor(s)
- Application specific processors
  - Pipelined processing
- Combination
  - Also specialized hardware

Performance



PENSANDO

# Built-in Security

## Execution

- Root-of-trust
- Secure boot
- Trusted software execution
- Secure key store

## Isolation

- Controlled communication with the host
- Driven by DPU
- Limited agency by host

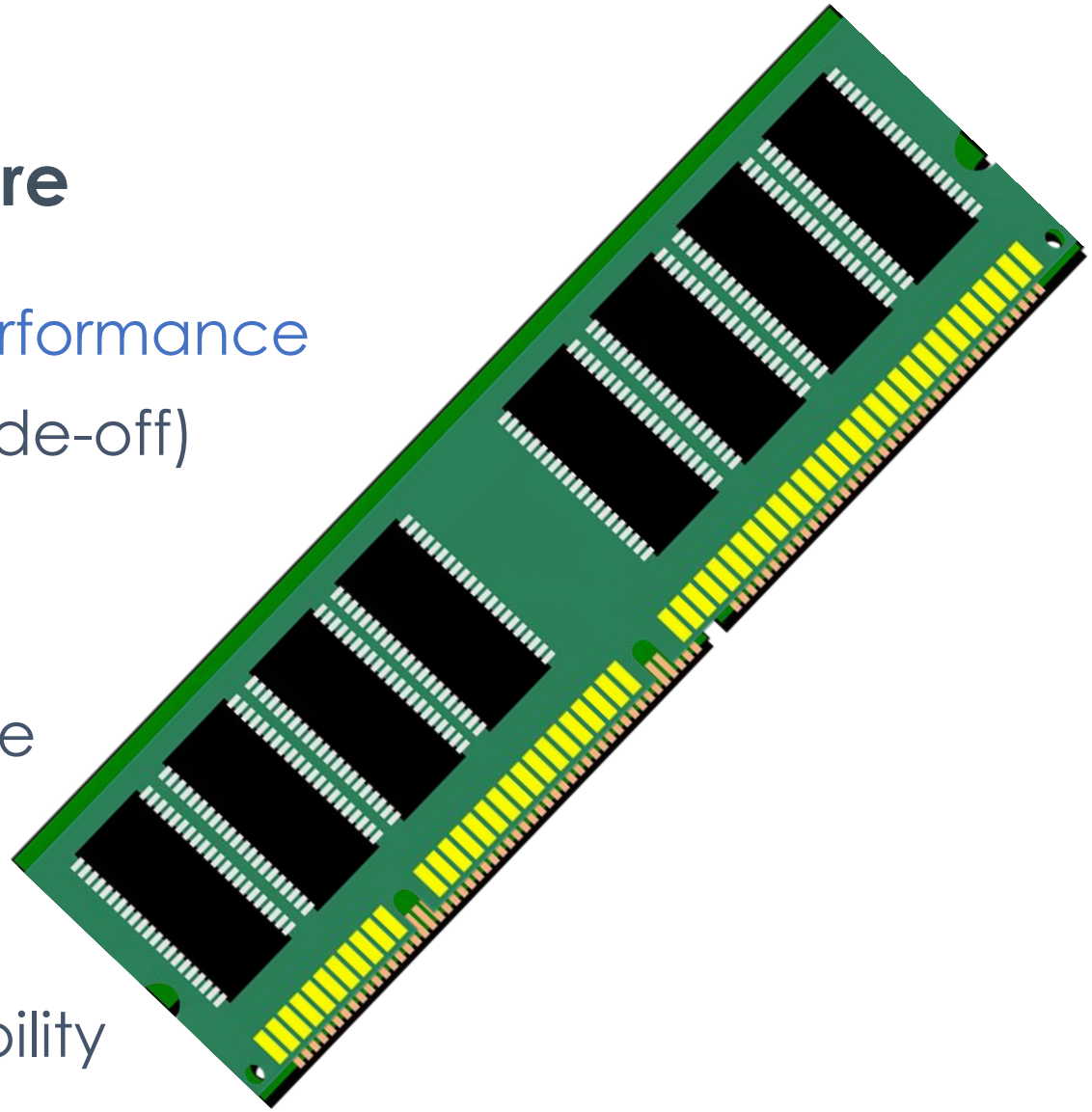


PENSANDO

# Sophisticated Memory Architecture

Scalability and Performance

(possibly a trade-off)



- Fast local memory → performance
  - E.g., SRAM at each pipeline stage
  - Small size → ~~scalability~~
  - Low efficiency → ~~scalability~~
- Large centralized memory → scalability
  - Access latency → ~~performance~~

**Multi-stage Memory Hierarchy**



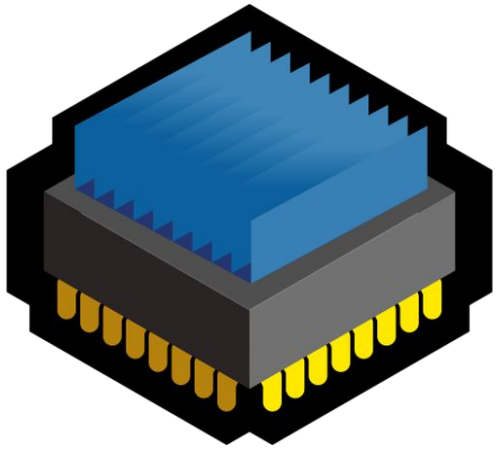
PENSANDO

# Central Memory with Local Cache

Mask access latency and limit contention

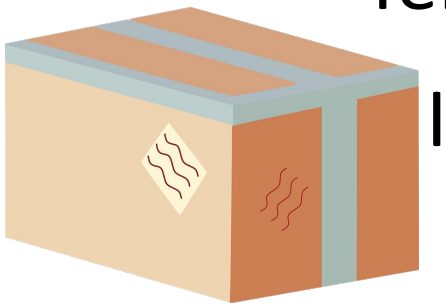
↓

possibly best of breed



In traditional, general purpose computing

- Spatial locality
- Temporal locality



In pipelined packet processing, also

**Functional locality**

↓ ↓

Limited amount of data  Small number of instructions

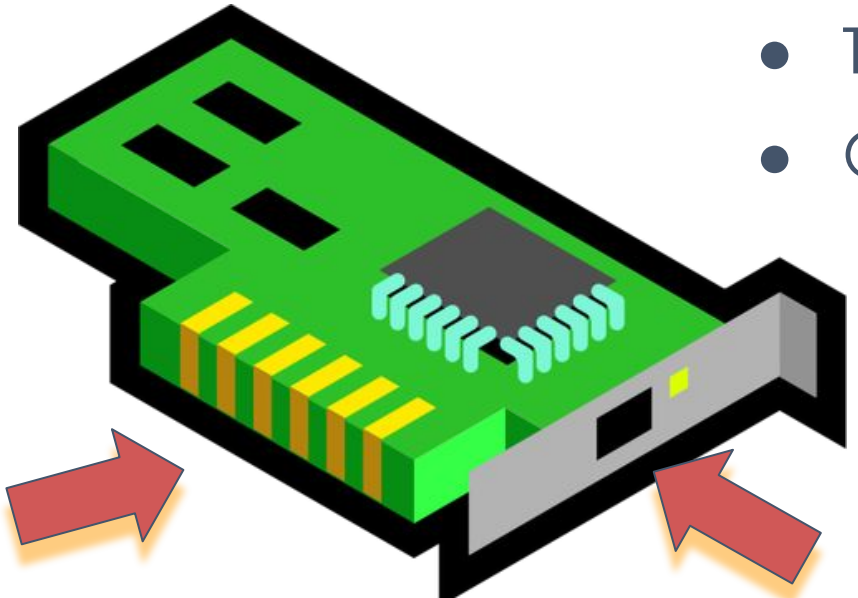
PENSANDO

# Input Differentiation and Scheduling

Virtualization and multi-tenancy

## PCIe

- Multiple logical interfaces
- Complex scheduling



## Ethernet

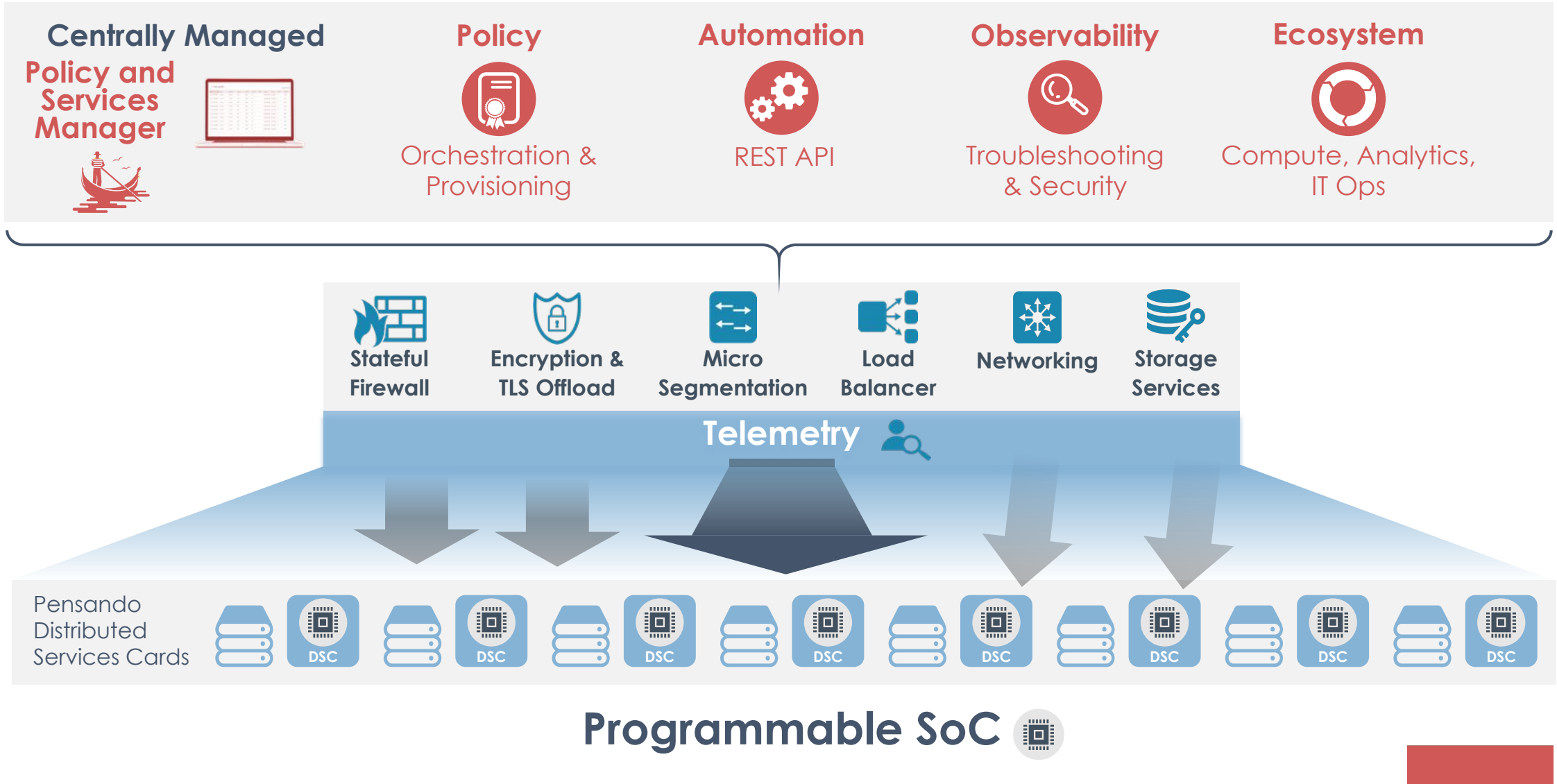
- Traffic classification
- Differentiated queuing
- Traffic policing
- Operation at line rate
  - Bit rate
  - Packet rate



PENSANDO


# The Pensando Solution

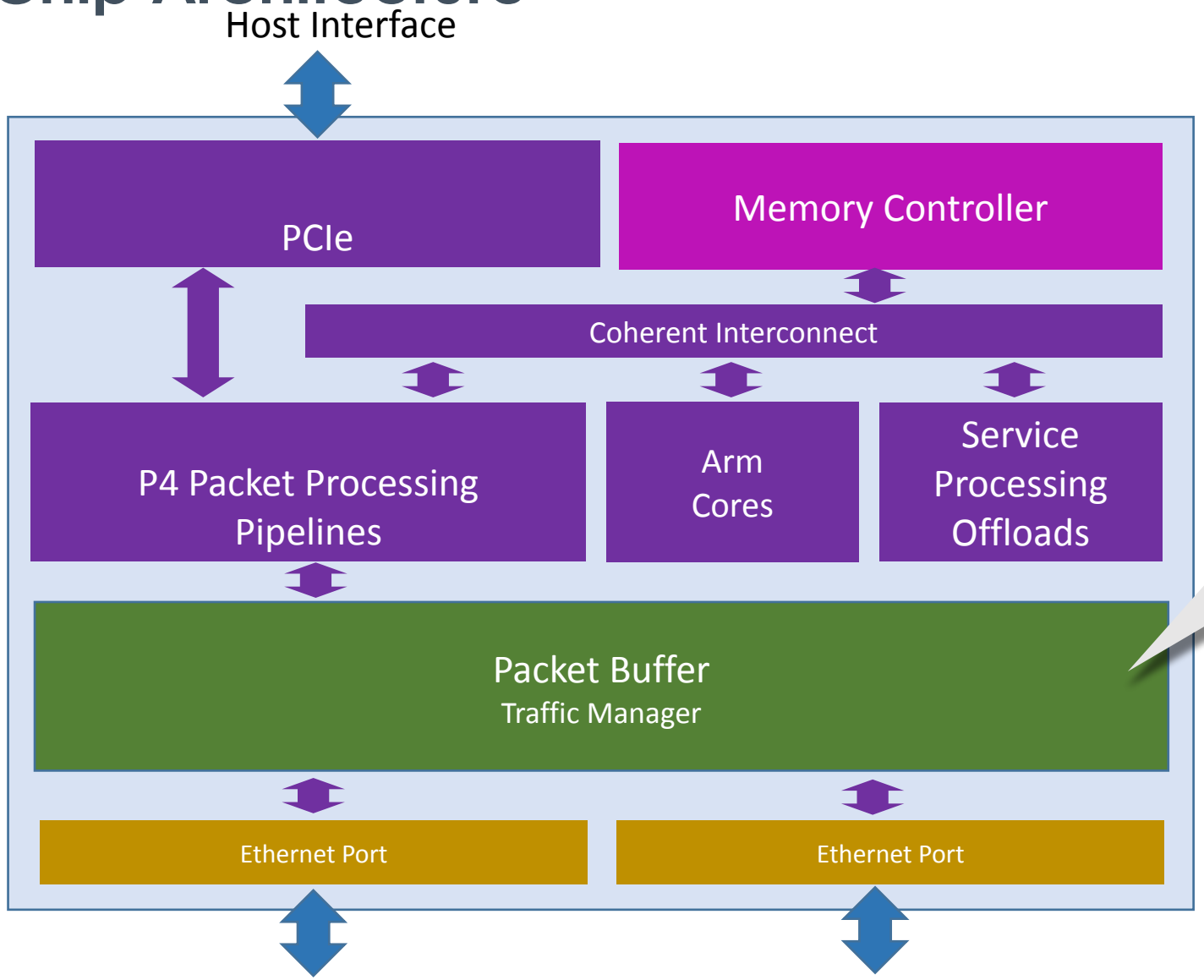
# Pensando Distributed Services Platform






PENSANDO

# System On Chip Architecture

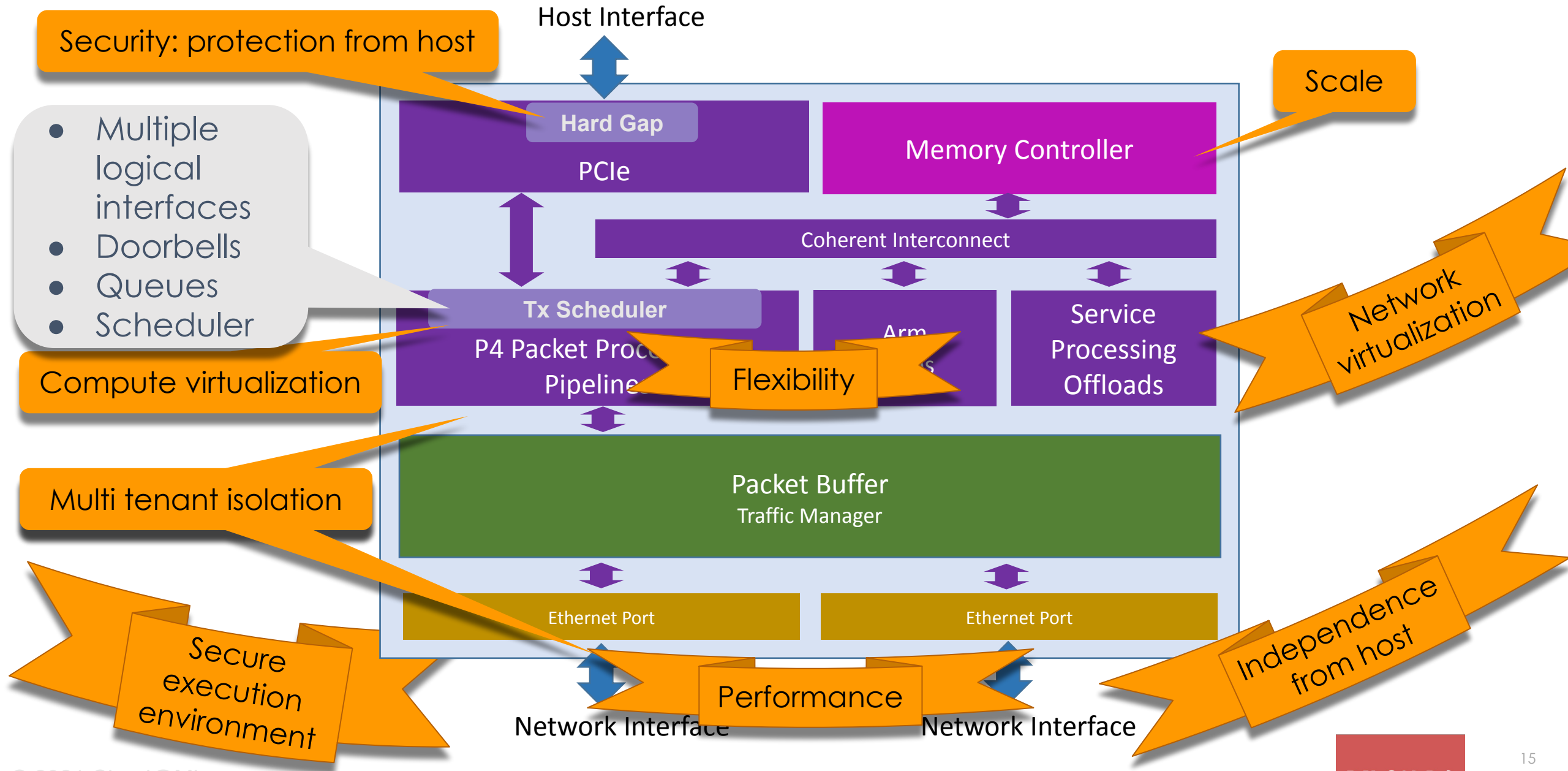



- Multiple queues
- Packet scheduling



PENSANDO

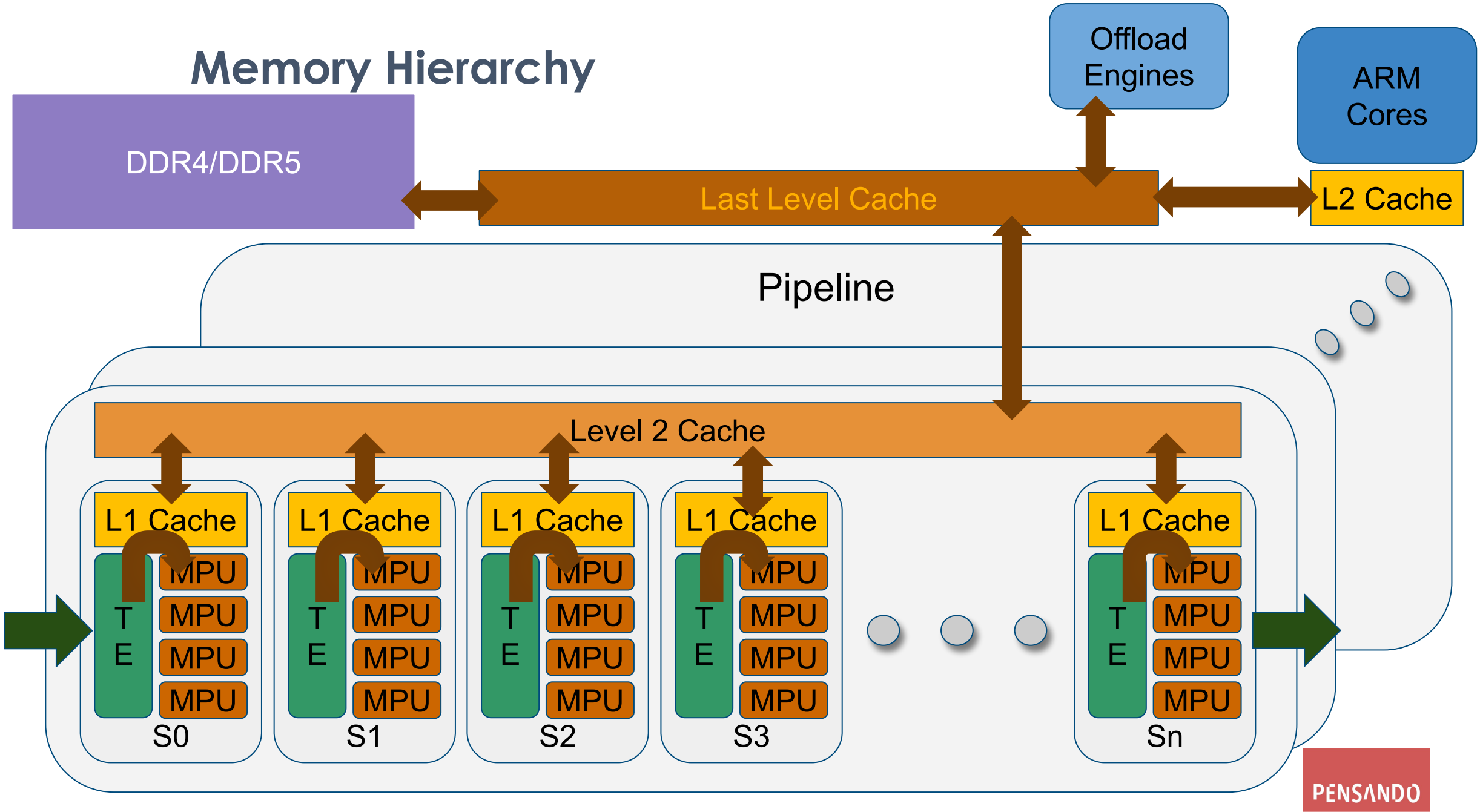
# Requirement Mapping





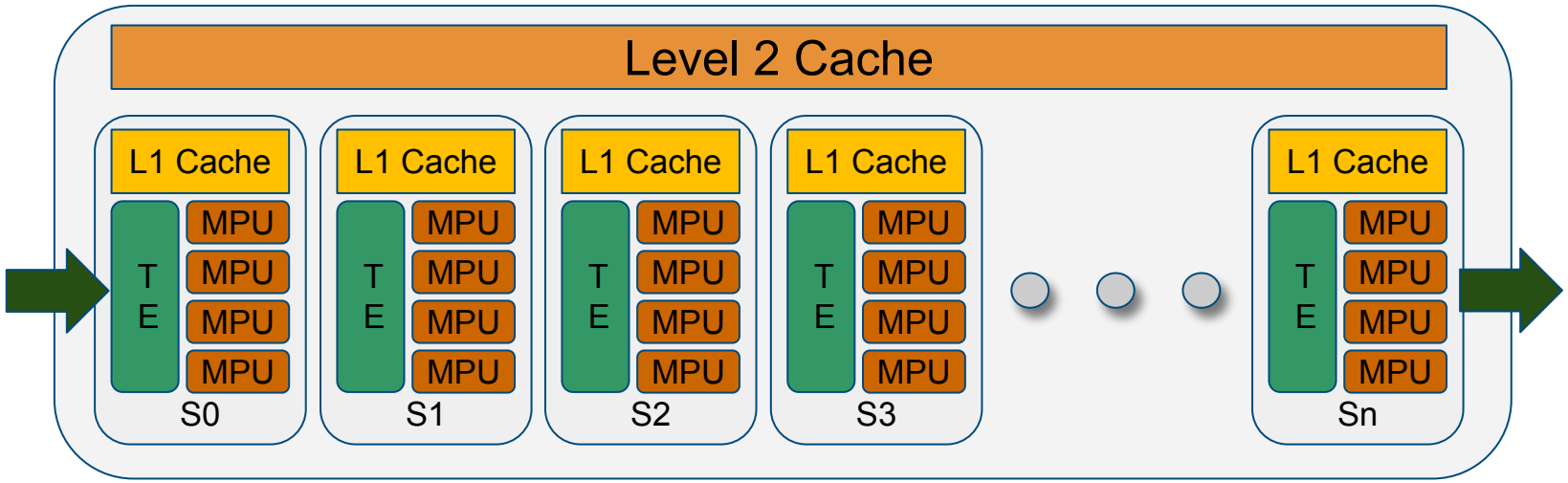

PENSANDO

# Memory Hierarchy



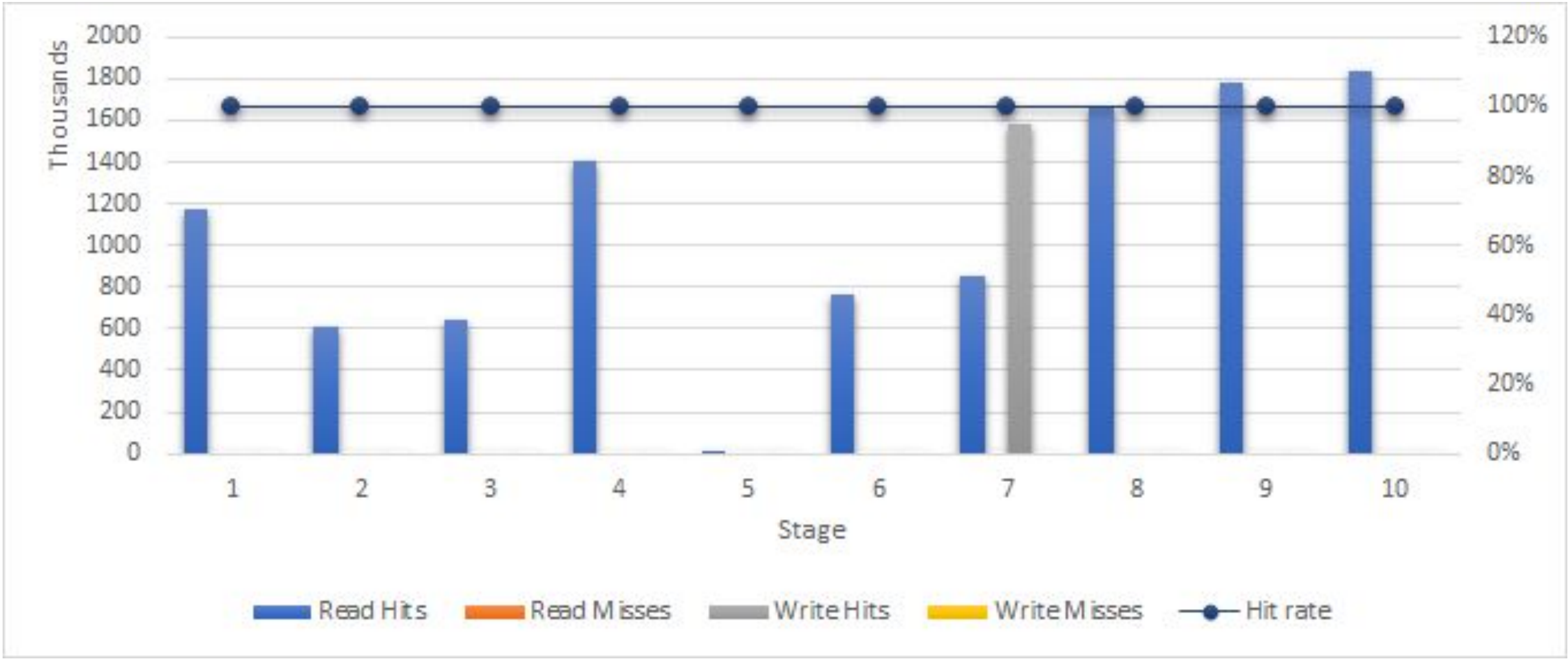
# Experimental Setup

- Production P4 code for cloud data center
  - Two different P4 programs
- Various configuration and traffic scenarios
- L1 data cache hits and misses per stage
- Instruction cache hits and misses per MPU



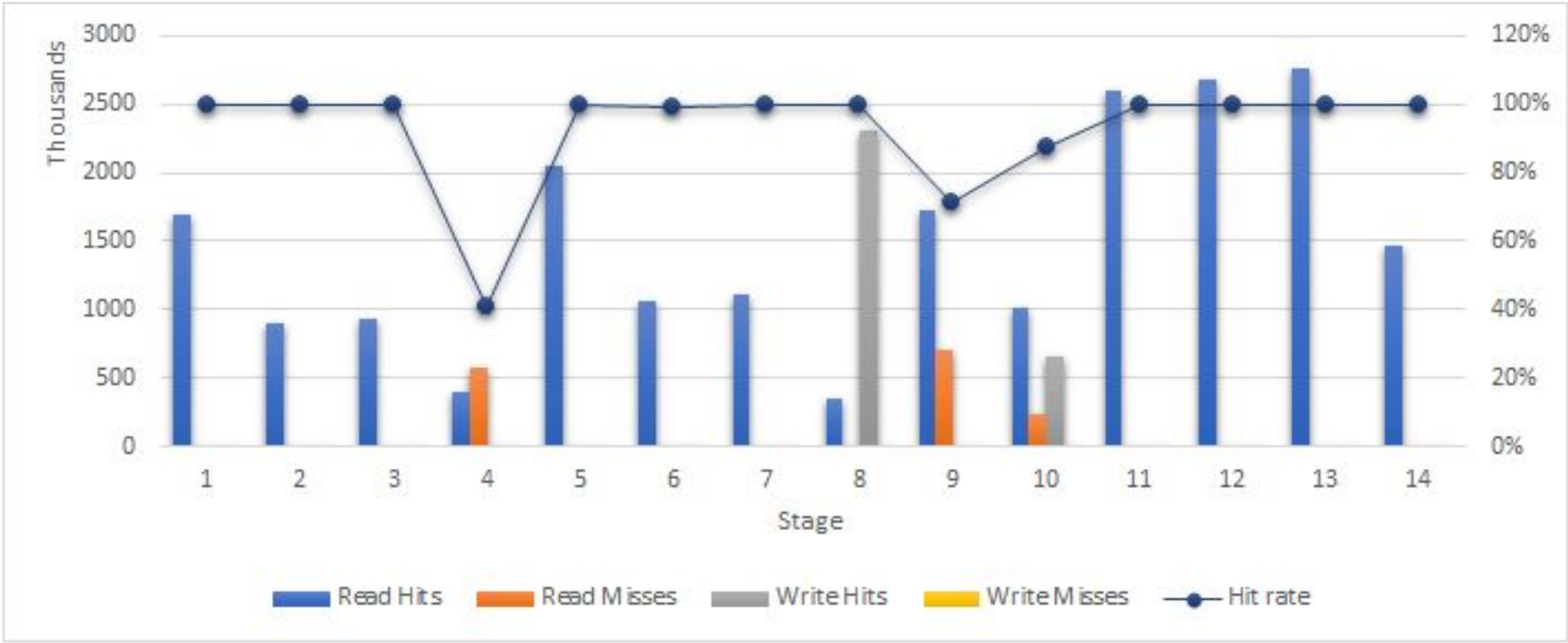

PENSANDO

# 1,000 active flows




PENSANDO

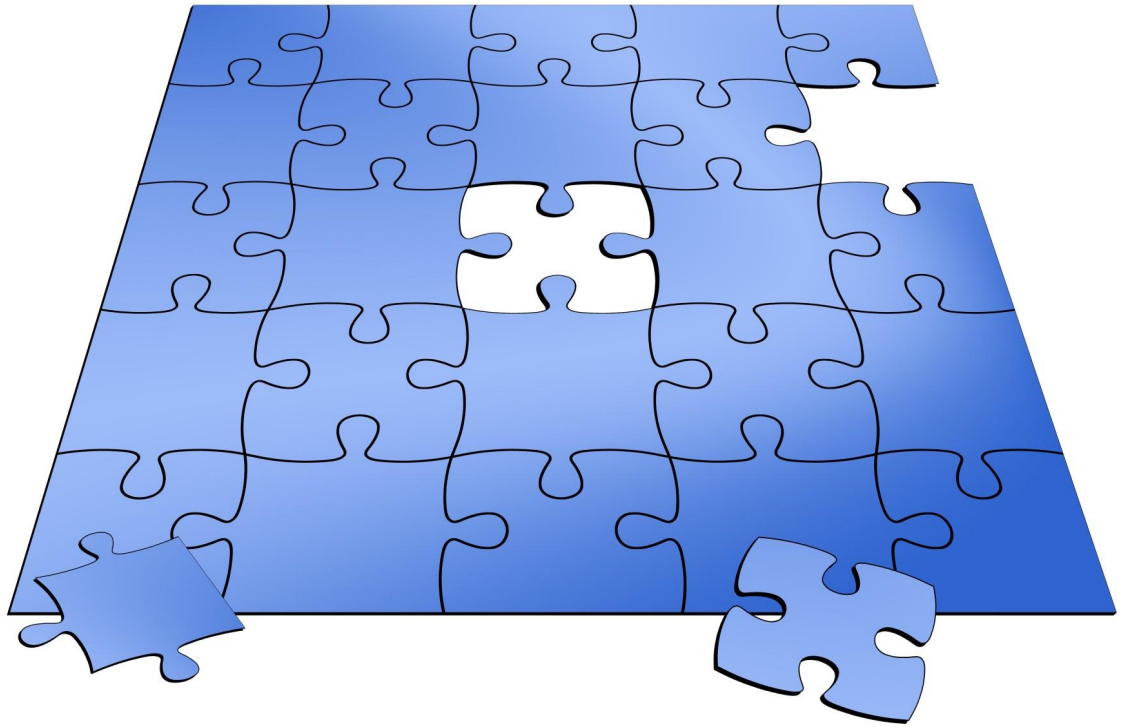
# 1,800,000 active flows




PENSANDO

# Concluding Remarks

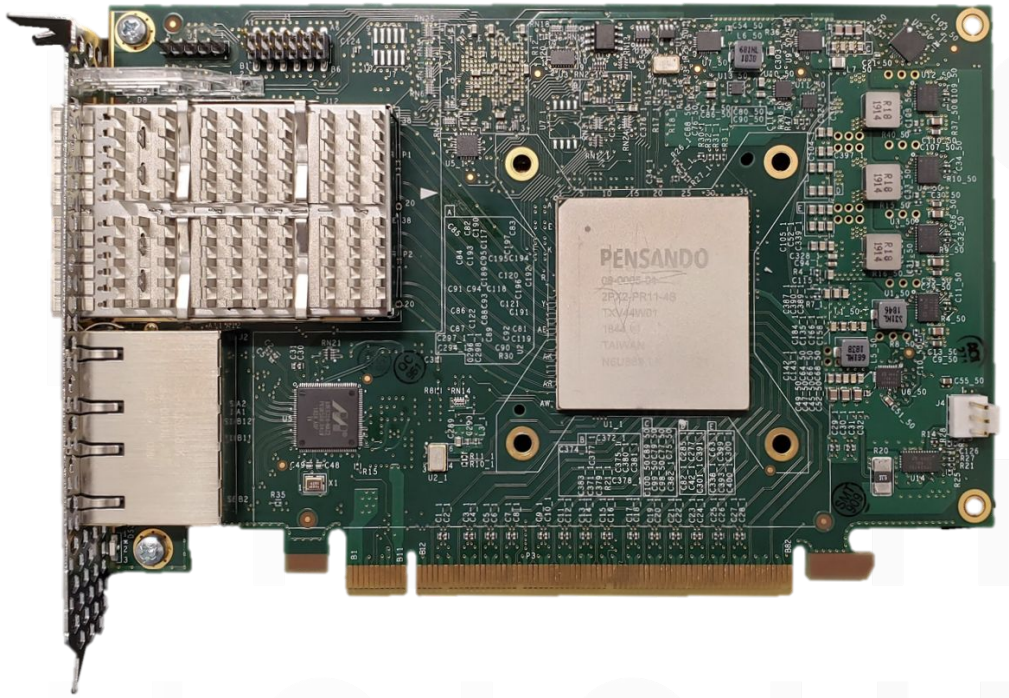
- Data Processing Units (DPUs) can play a key role in the cloud
  - Free up valuable resources by offloading services
  - Open up new opportunities
- Specific cloud requirements
- Design implications
- The Pensando Solution
  - How specific design choices address cloud requirements on DPUs





PENSANDO

# THANK YOU




baldi at pensando dot io
www.pensando.io
pensando.io/blog/