SNIA DEVELOPER CONFERENCE

SDC 24

BY Developers FOR Developers

September 16-18, 2024
Santa Clara, CA

Accelerating GPU Server Access to Network-Attached Disaggregated Storage using Data Processing Unit (DPU)

Eriko Nurvitadhi, MangoBoost, Inc.

Craig Carlson, AMD

Agenda

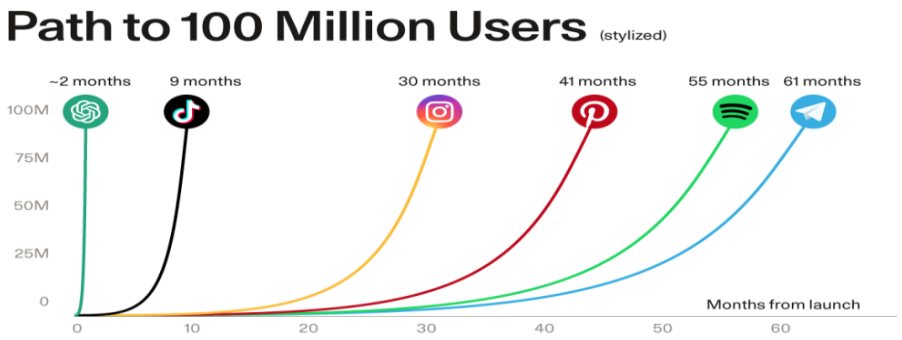
- Trends in AI & Implications on Storage Systems
- AMD’s GPU Ecosystem for AI
- Opportunity for Data Processing Unit (DPU) in AI
- Case Study: LLM Training with DPU-Accelerated Storage

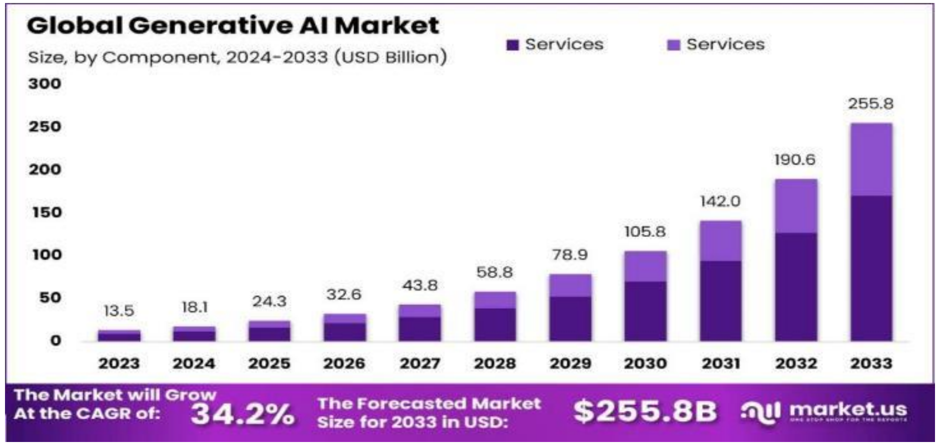
Trends in AI & Implications on Storage Systems

The Rise of AI

The first real AI “killer app”?

ChatGPT (source: https://chatgptguide.net/)



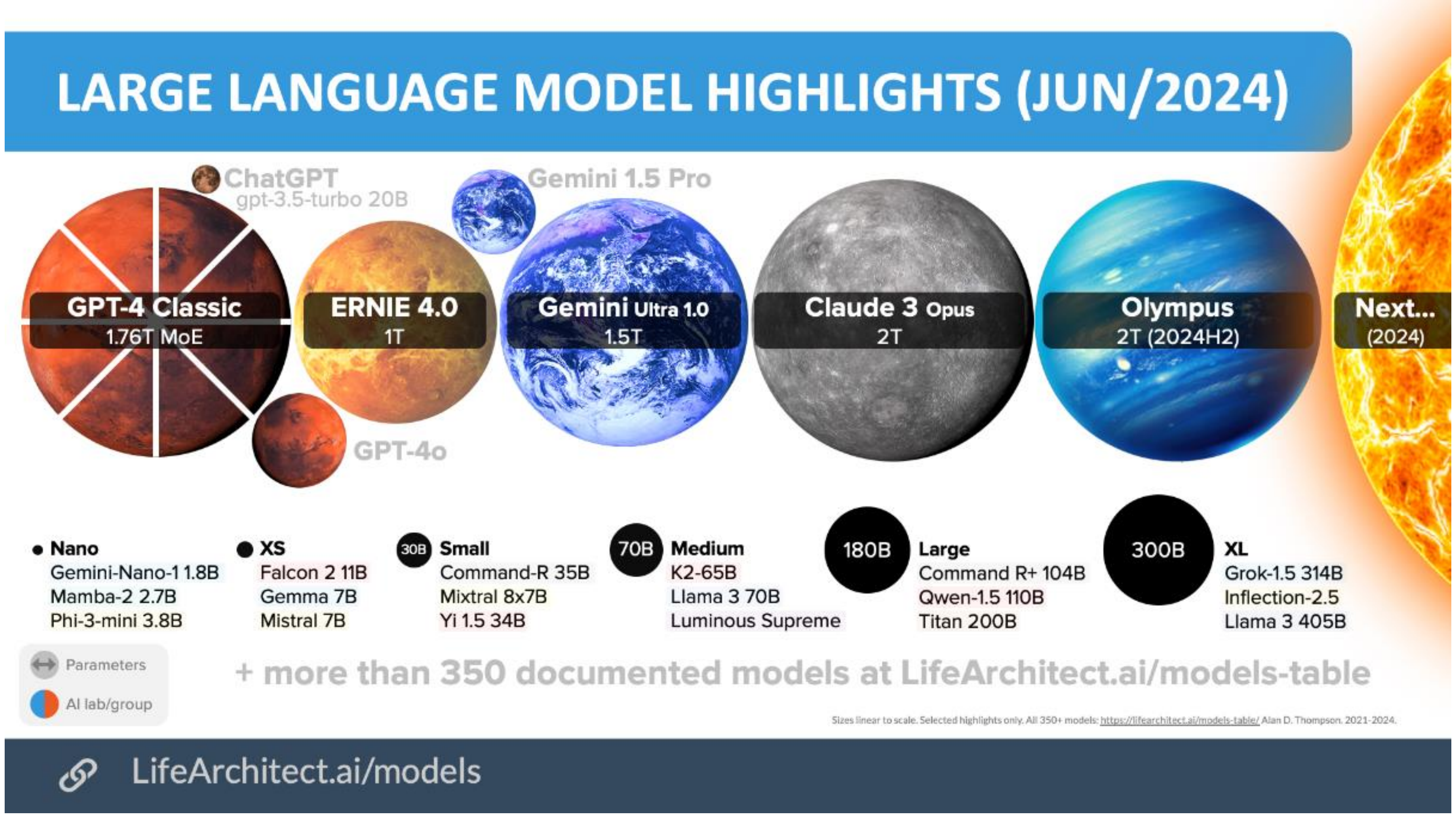


More AI capabilities are becoming available, with growing adoption in various domains (e.g., ChatBot, ImageGen, VideoGen, etc)

** source: video generated by AI, openai sora.*

Driven by Large Language Models (LLMs)

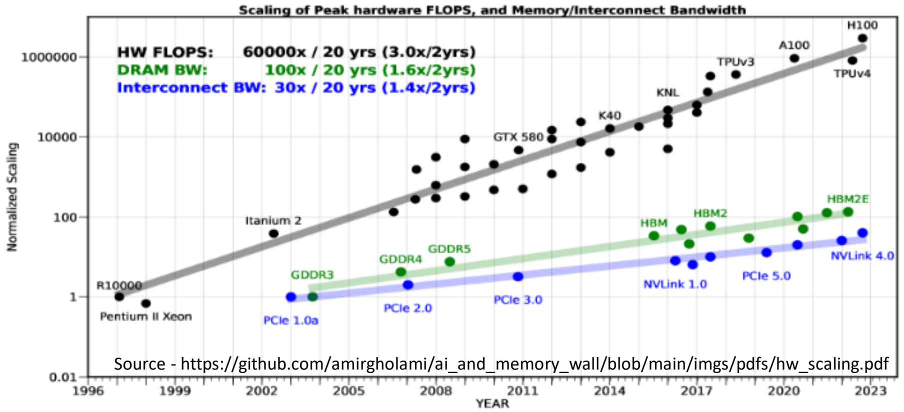
- The most popular AI LLM models these days are super big, complex, and many variations



How do we build efficient LLM AI systems as model sizes continue to grow?

AI Systems Increasingly Challenged by Data-Oriented Tasks

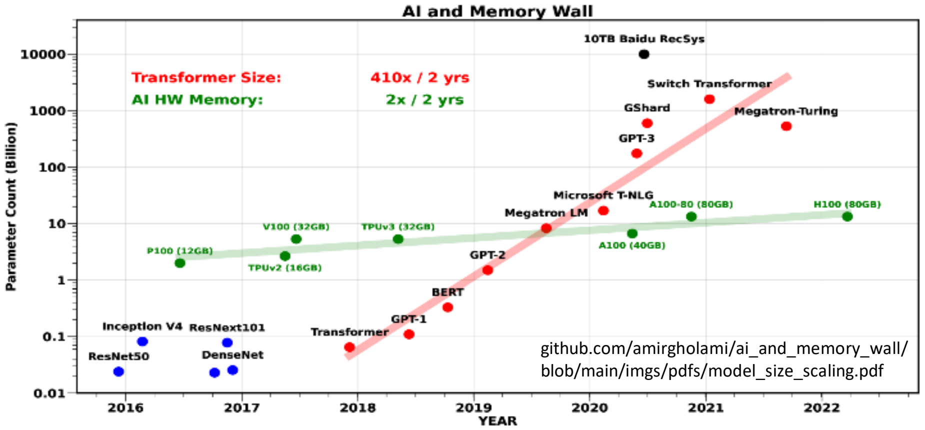
- Rapid growth in HW compute in past decade, but slower growth in data move & store

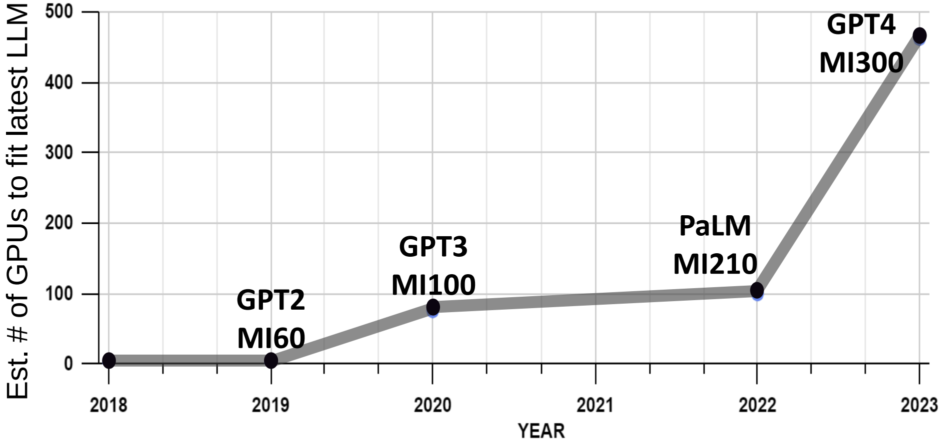


60,000x HW compute peak OPs growth in the last 20 years!

However, off-chip memory & interconnect bandwidth grow only by 100x & 30x.

LLMs don't fit in 1 card anymore





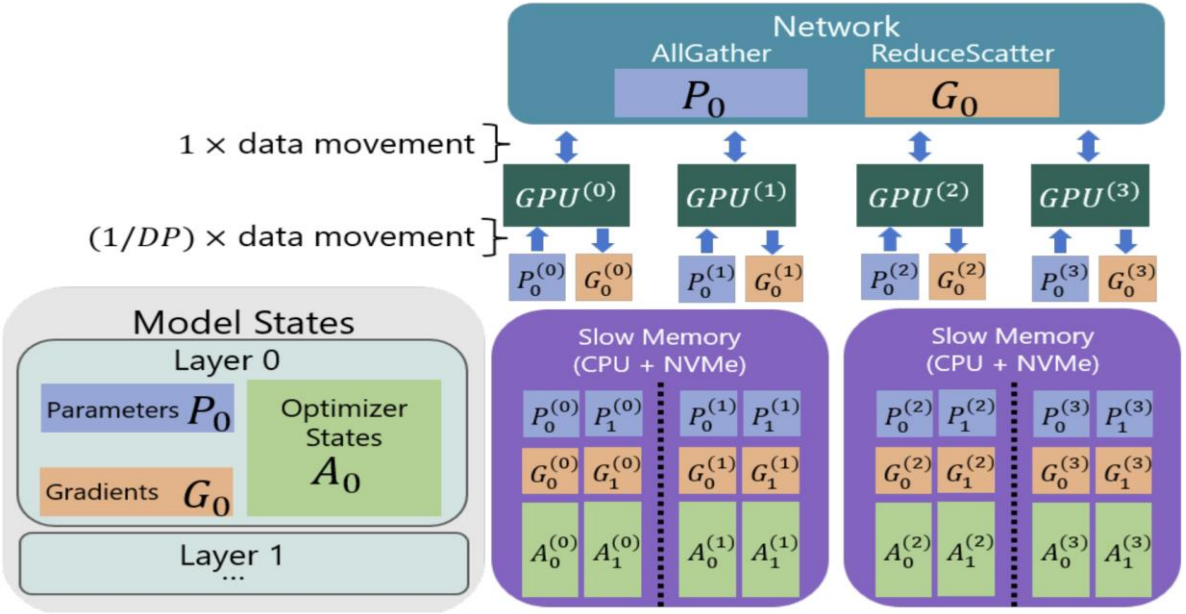
GPU mem size grow much slower vs LLM size

Need to consider larger storage options

Emergence of Storage-Optimized AI Frameworks & Benchmarks

Emergence of storage-optimized AI frameworks to allow “spilling” large AI model & intermediate data to mem & SSDs

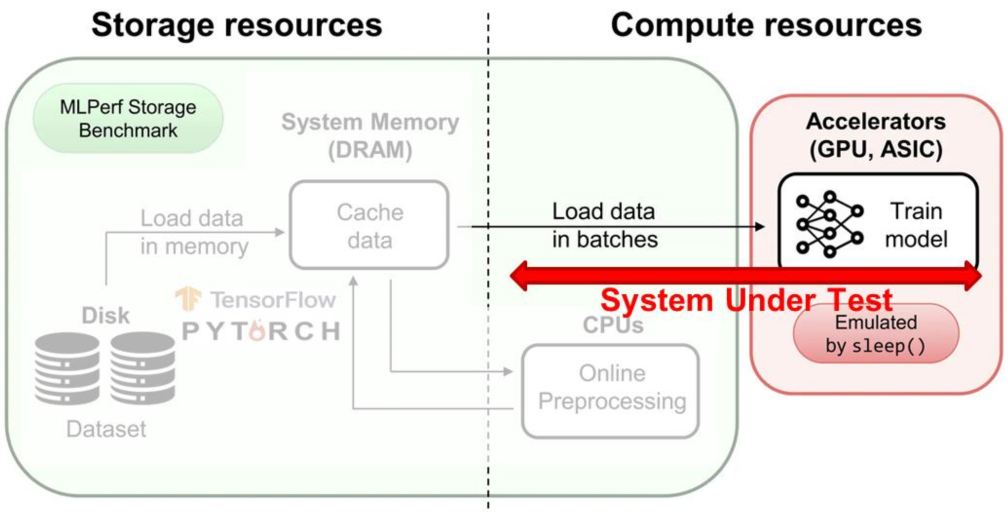
E.g., DeepSpeed Zero-Infinity training framework



Microsoft DeepSpeed ZeRO-Infinity *[https://arxiv.org/abs/2104.07857]*

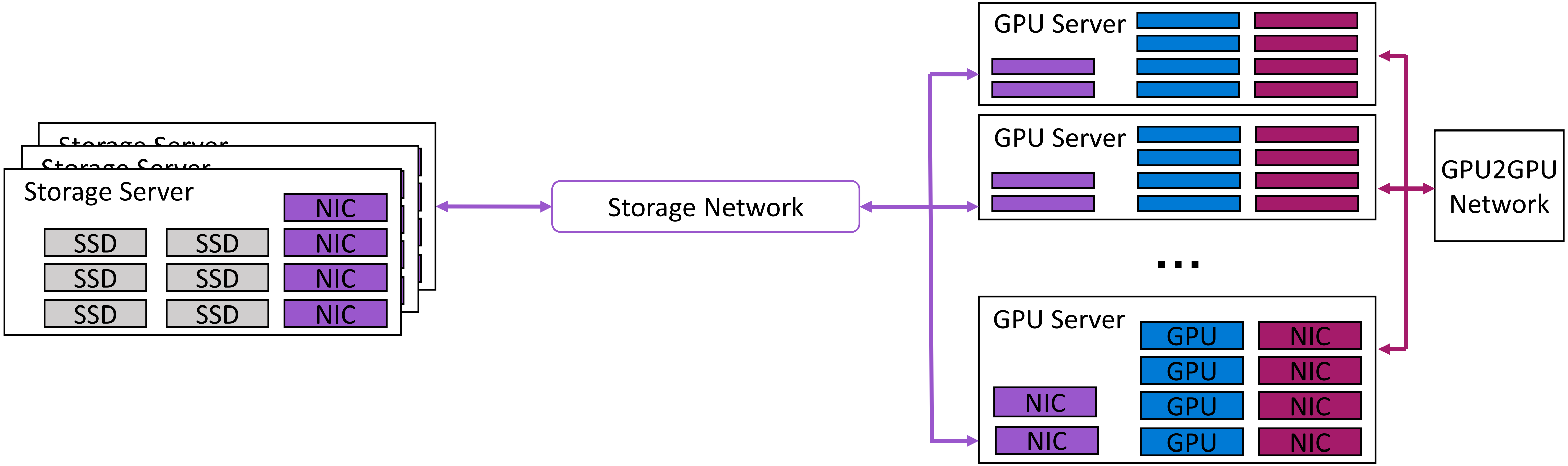
Emergence of AI benchmarks to evaluate storage

E.g., MLPerf storage



Mlperf Storage - *source: https://mlcommons.org/2023/06*

Need for Disaggregated Storage in AI Systems



High throughput and capacity: Storage server with many SSDs

→ to access & store large AI model/data/parameters

Eliminate over-provisioning: flexible configuration

→ can assign appropriate storage resources to varying AI workloads need

Save space + improve bandwidth util in GPU server:

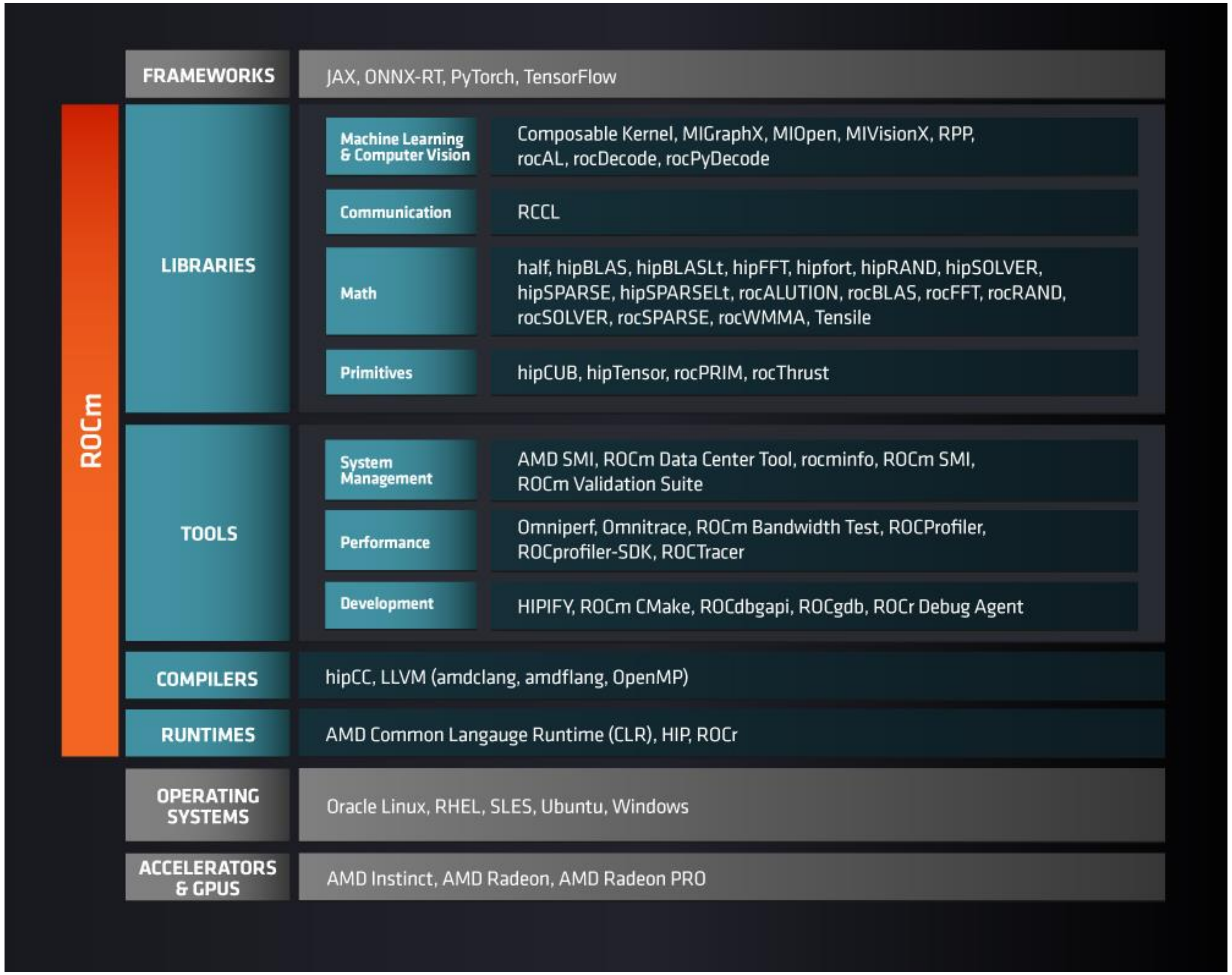
→ single NIC can offer higher BW than a single SSD

AMD’s GPU Ecosystem for AI

GPU

- AMD GPUs come in two classes
  - Radeon – Consumer GPUs
    - Primarily used for gaming, but can be used for AI/HPC
  - Instinct – Data center GPUs
    - CDNA – Architecture Designed for AI and HPC applications
    - HBM – Includes High Bandwidth Memory
    - Infinity Fabric – High speed interconnect
- ROCm Development Platform
  - Open source
  - Supports Instinct and Radeon GPUs

ROCm



- ROCm is an open source software stack that provides access to GPU computation
- ROCm supports both AMD Radeon and Instinct GPUs

ROCm Developer tools

- HIP Environment
  - Runtime Libraries and Kernel Extensions
  - Compilers
    - HIPCC – Frontend to C++ and Perl
    - FLANG – Fortran compiler for LLVM
  - Hipify – Convert Cuda software to ROCm
  - ROCm CMAKE – Simplify the building of ROCm applications
  - ROCgdb – ROCm Debugger

ROCm Libraries

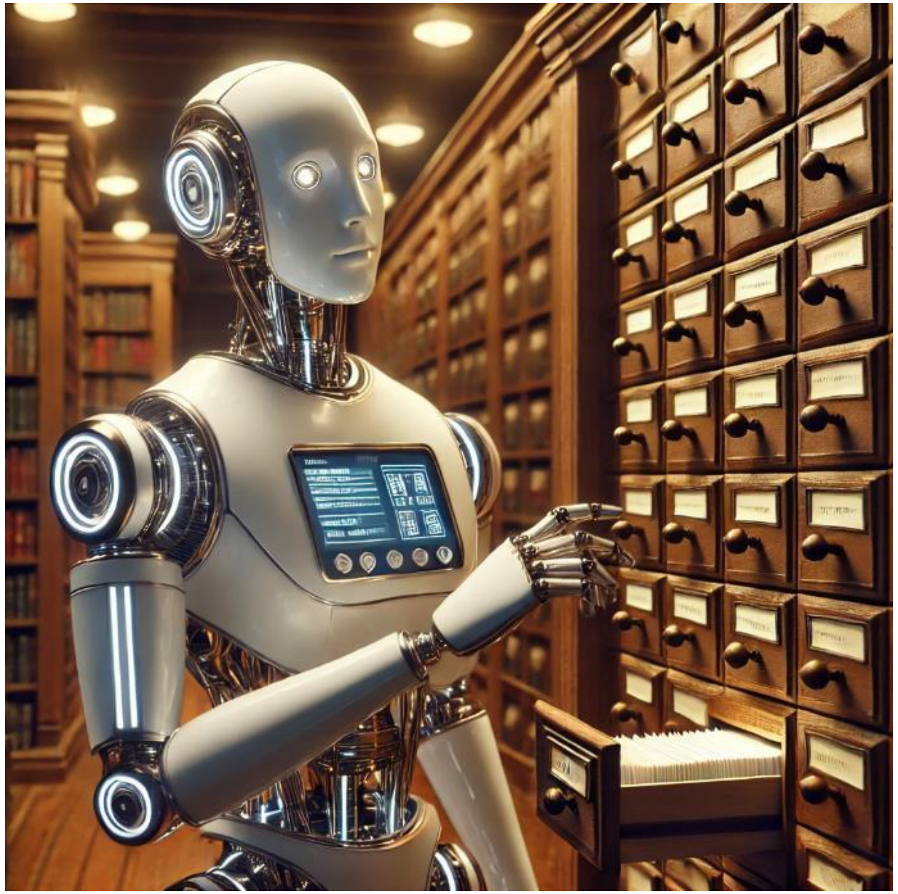
CUDA Library	ROCm Library	Description
cuBLAS	rocBLAS	Basic Linear Algebra Subroutines
cuFFT	rocFFT	Fast Fourier Transfer Library
cuSPARSE	rocSPARSE	Sparse BLAS + SPMV
cuSolver	rocSolver	Lapack Library
AMG-X	rocALUTION	Sparse iterative solvers & preconditioners with Geometric & Algebraic MultiGrid
Thrust	rocThrust	C++ parallel algorithms library
CUB	rocPRIM	Low Level Optimized Parallel Primitives
cuDNN	MIOpen	Deep learning Solver Library
cuRAND	rocRAND	Random Number Generator Library
EIGEN	EIGEN	C++ template library for linear algebra: matrices, vectors, numerical solvers
NCCL	RCCL	Communications Primitives Library based on the MPI equivalents

Continuous open source development

- Continuous development since initial release in 2016
  - 1.0 released in 2016 – Focused on HPC
  - 3.0 released in 2019 – Focused on AI
  - 5.0 released in 2022 – Full support for MI200
  - 6.1 released in 2024 – Support for MI300
  - 6.2 is the current release – Addition of new profiling tools

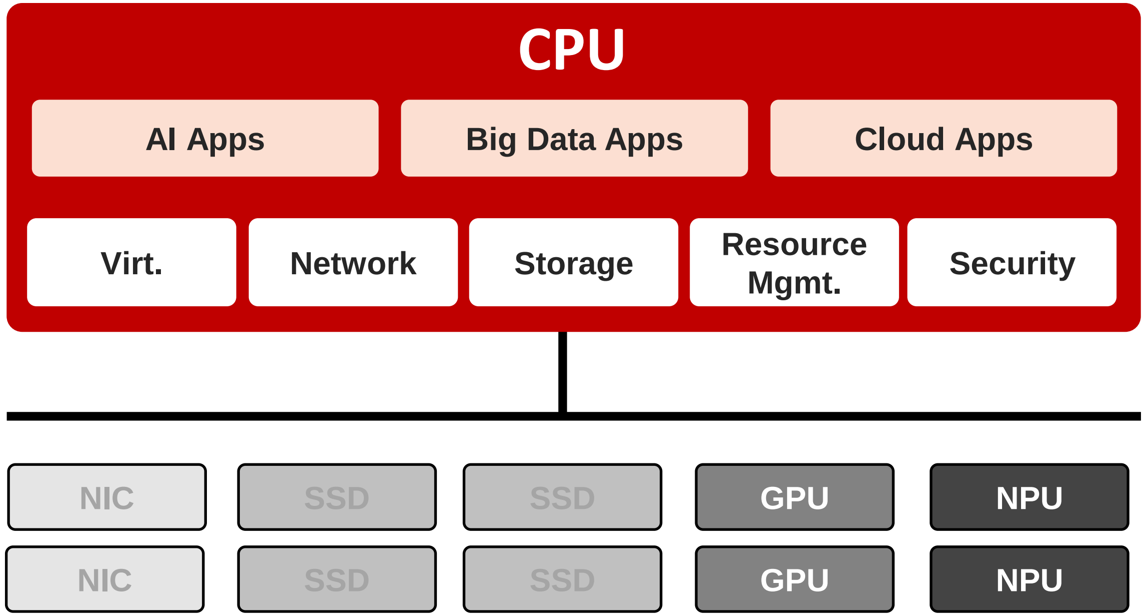
References

- ROCm Documentation
  https://rocm.docs.amd.com/en/latest/
- ROCm Github
  https://github.com/ROCm/ROCm
- ROCm Community
  https://github.com/ROCm/ROCm/discussions



Opportunity for Data Processing Unit (DPU) in AI

Modern Datacenters are No Longer Scalable



SW complexity

Growing software stack
(e.g., virtualization, NVMe-oF)
for emerging applications

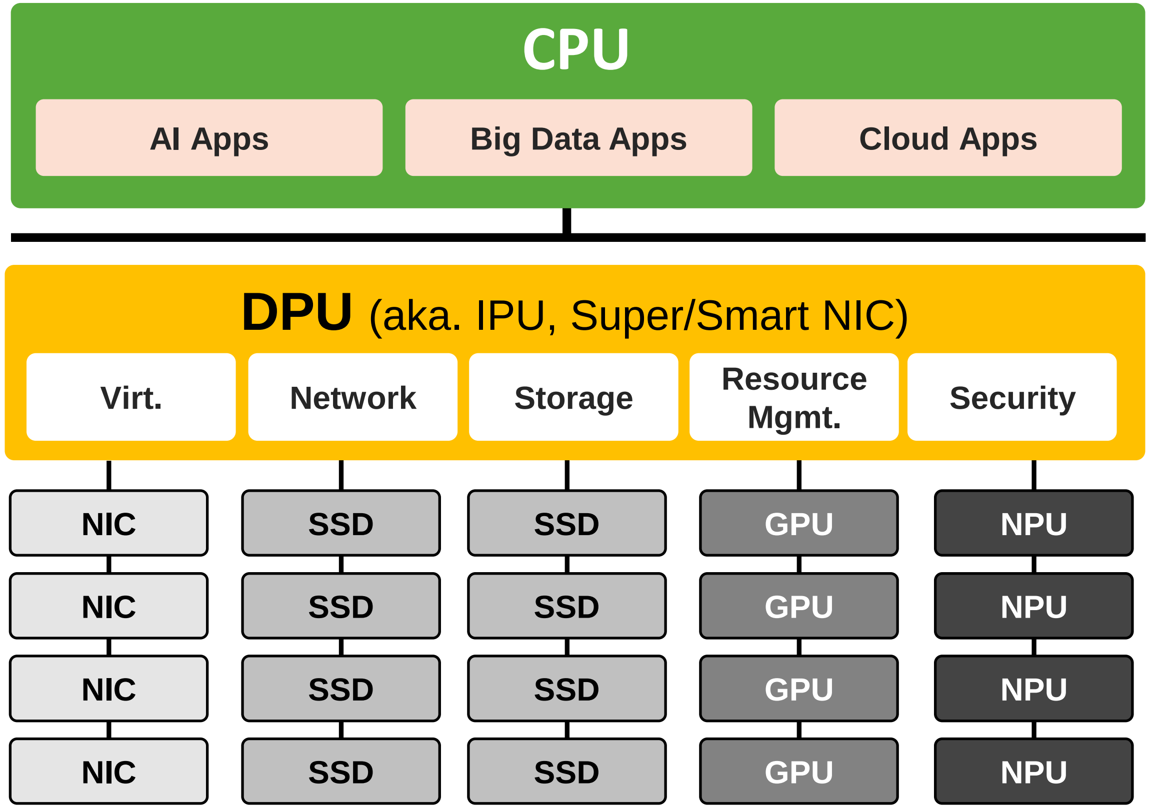
HW complexity

More devices to move data (NICs)
More devices to store data (SSDs, Mems)
More devices to process data (GPUs, NPUs)

Growing "Datacenter Tax" →

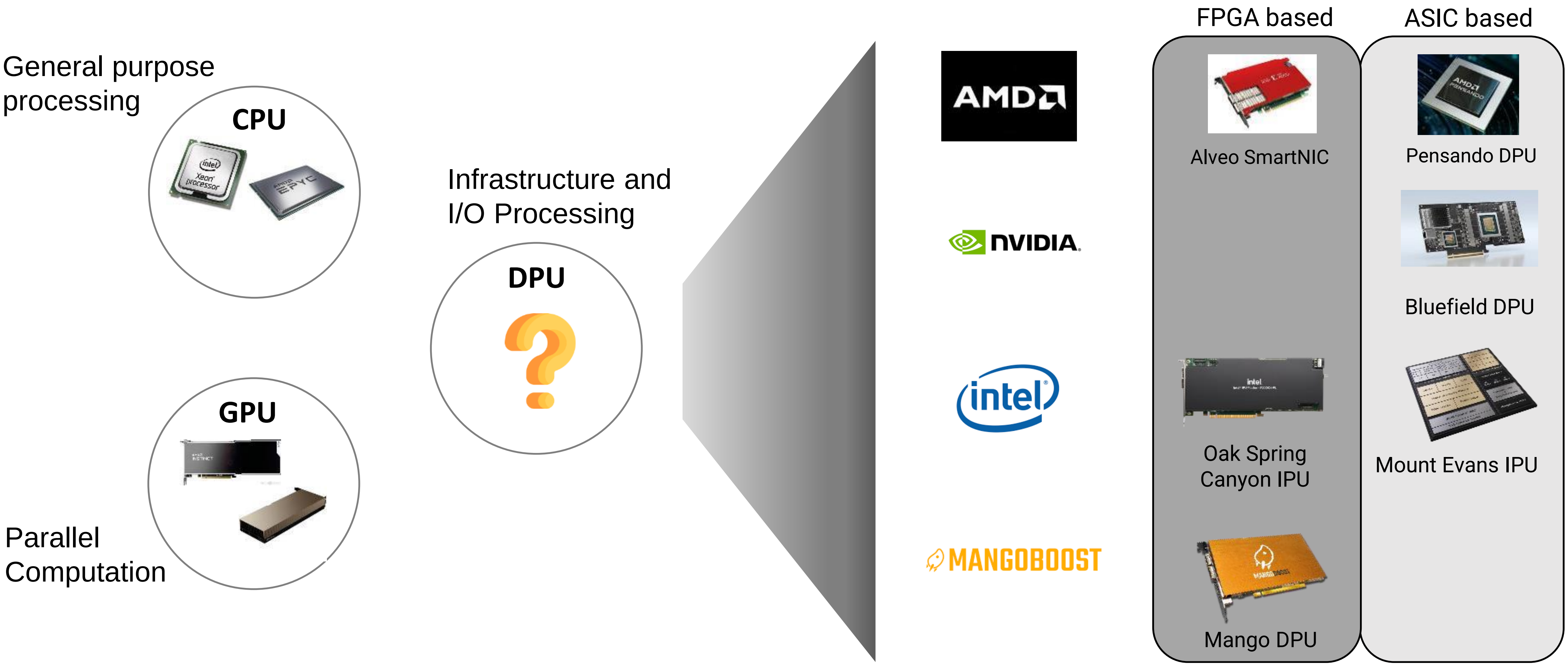
22-27% CPU overhead @ Google (ISCA 2015)
31-83% CPU overhead @ Facebook (ASPLOS 2020)

DPU Accelerates Various Infrastructure Data Processing



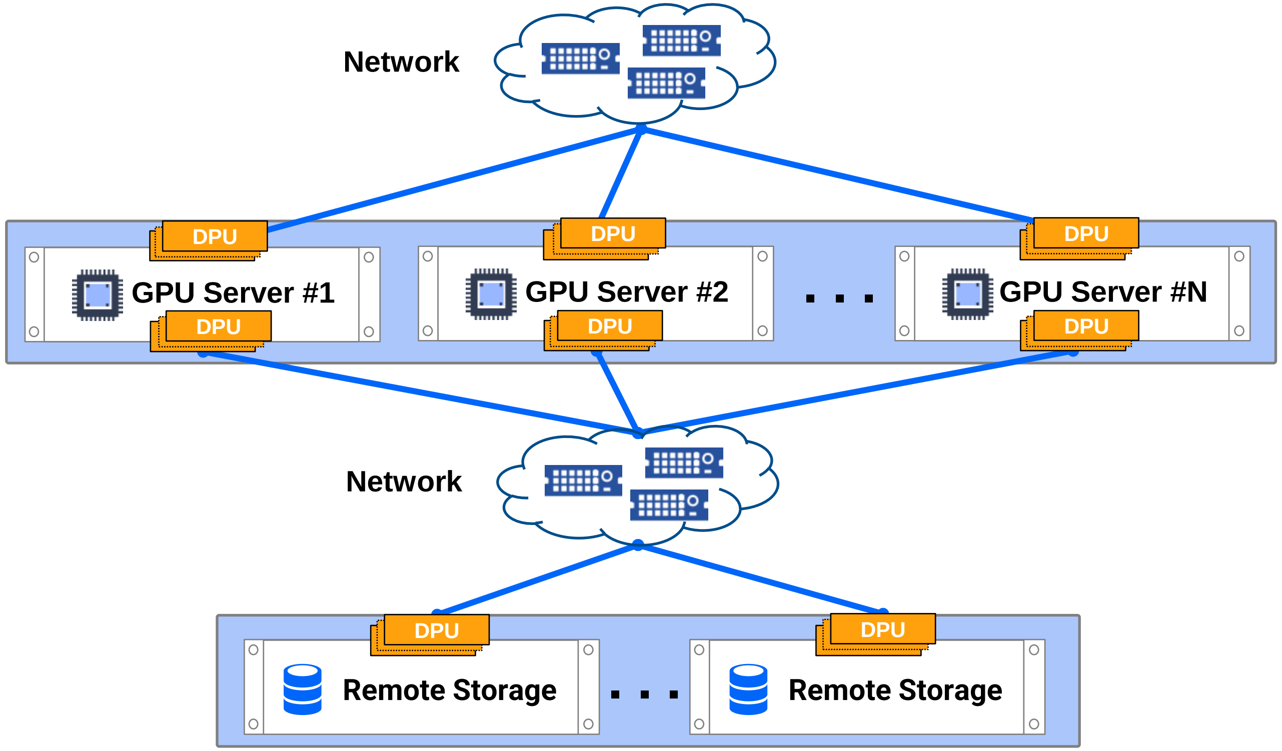
DPU enables more scalable, faster, and cheaper datacenter

Multiple DPU Products Entering the Market



Some examples of DPUs in the market

Opportunities for DPUs in AI systems



Inter-GPU Network
- **RDMA (e.g., RoCEv2, Infiniband)**
- **Collective Communication**

Intra-node communication
- **Peer-to-peer communication**

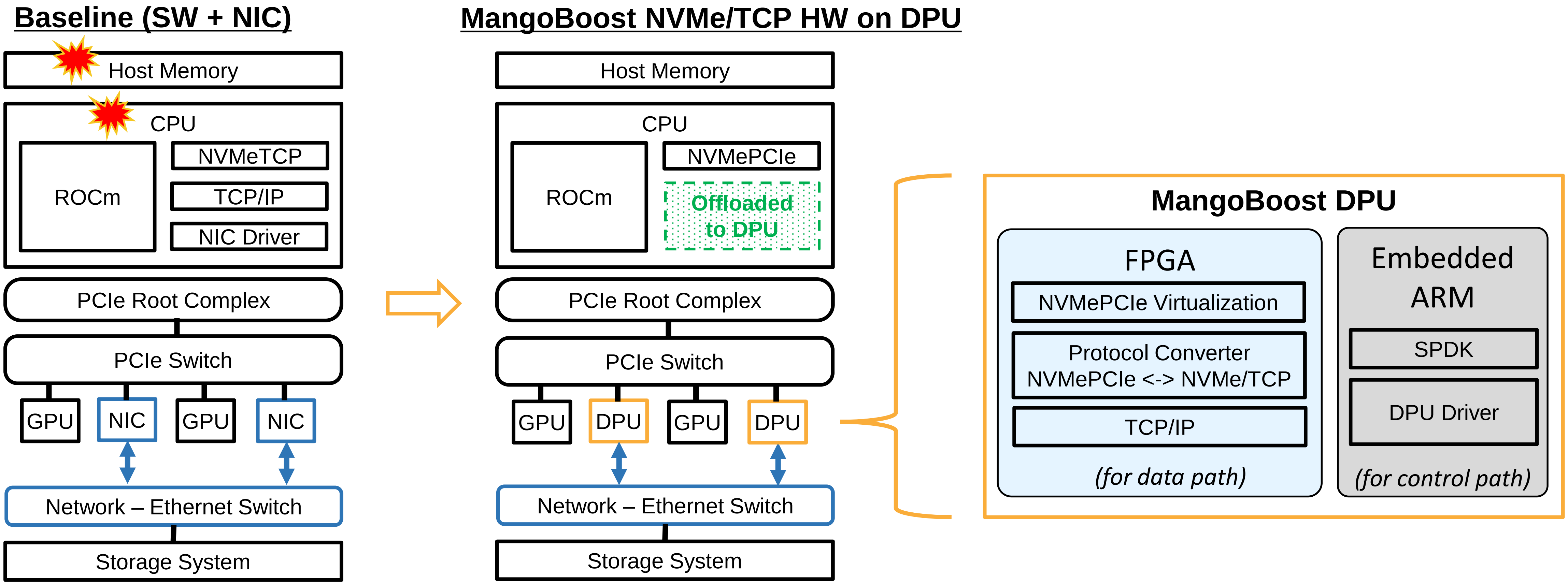
Storage Network
- **NVMe over Fabric**

Case Study: Optimize GPU to Storage communication using AMD GPU and MangoBoost DPU solution

GPU servers require high bandwidth I/O processing from network and storage
Many opportunities exist for DPU to improve AI system performance

Case study: LLM Training with DPU-Accelerated Storage

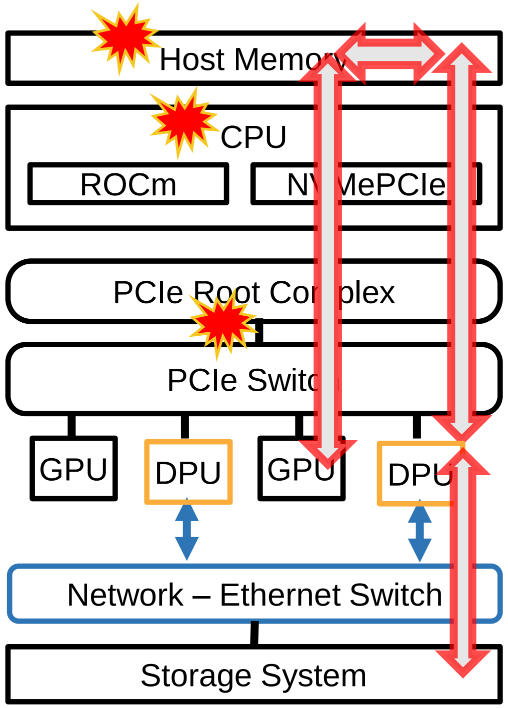
MangoBoost GPU Storage Boost (GSB) – (1) NVME/TCP HW



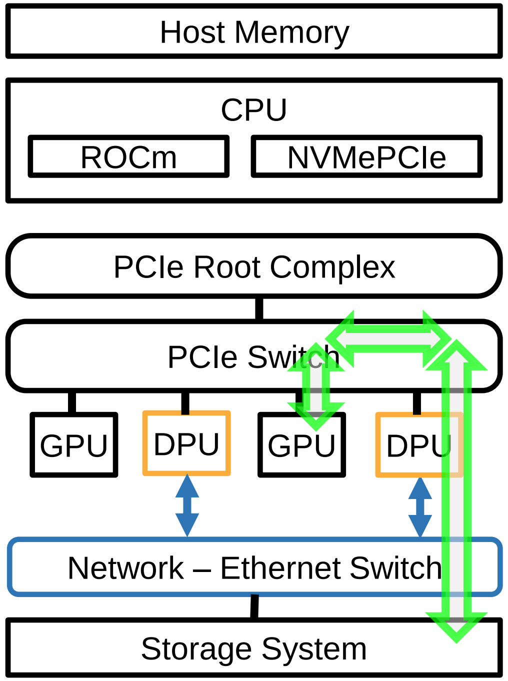
Simplify software stack and accelerate NVMe/TCP by full hardware acceleration of NVMe/TCP stack on DPU

GPU Storage Boost (GSB) – (2) Peer-to-Peer Comm.

MangoBoost NVMe/TCP HW on DPU



MangoBoost GPU Storage Boost

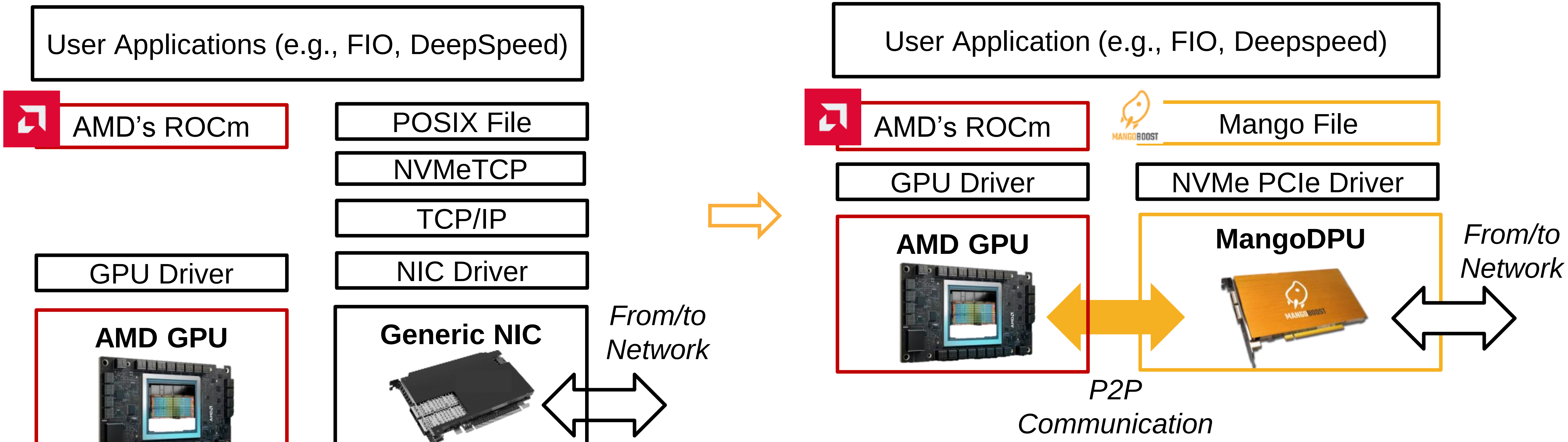


PCIe
Peer-to-peer communication

Optimize datapath of GPU and resolve resource (CPU, Memory, and PCIe) contentions by enabling peer-to-peer communication between DPUs and GPUs

 |

GPU Storage Boost (GSB): File APIs

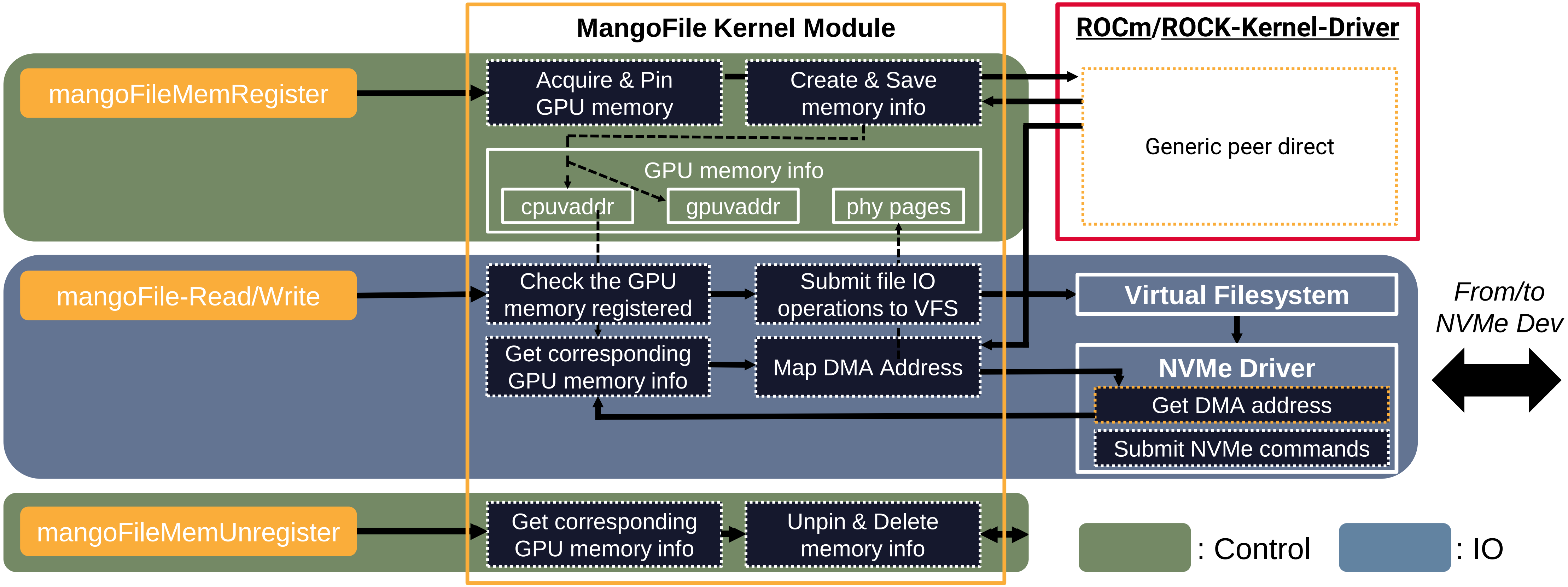


Baseline **GPU Storage Boost (GSB)**

Details of Mango File

- Enabling direct data movement between AMD GPU memory and storage device

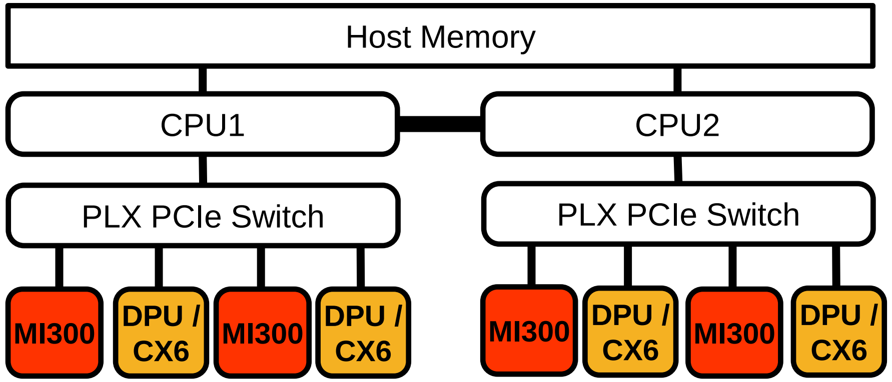
Note: SW made/modified by MangoBoost depicted in Orange



MangoFile library provides file-based IO operations that can leverage GSB.

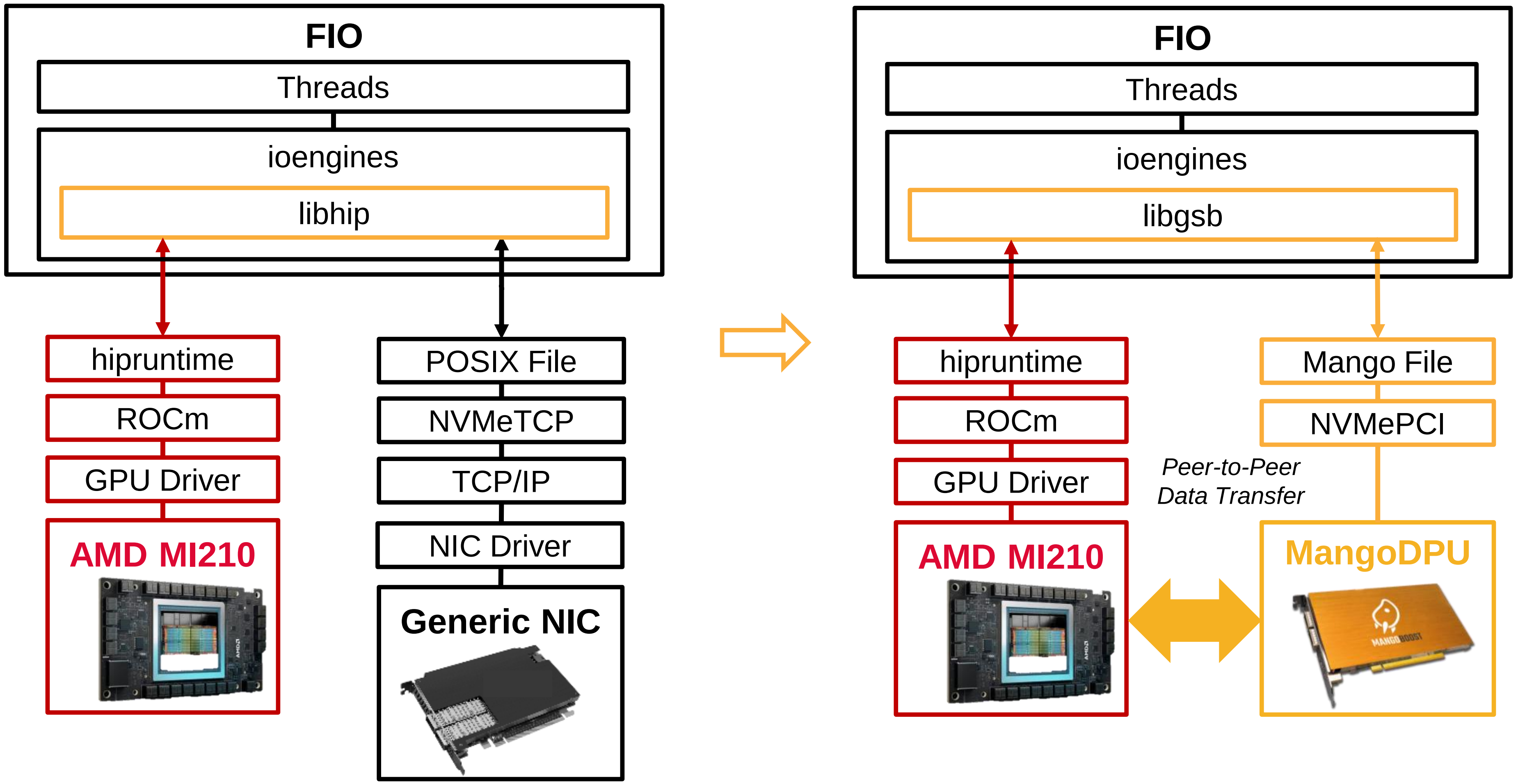
System Testbed (with AMD MI300 GPUs & MangoBoost DPUs)

	System Under Test (SUT)
Server	Supermicro GPU A+ System AS -8125GS-TNMR2
CPU	AMD EPYC™ 9534 64-Core Processor
Memory	2,377,705,704 kB (96GiB x 24 channels)
GPU	AMD MI300X x4
CPU-GPU	PCIe Gen5 16 lanes
NIC / DPU	2x100Gbps ConnectX-6 NIC x 4 = 800Gbps aggregate → baseline configuration 2x100Gbps MangoBoost DPU (on AMD U45N) x 4 = 800Gbps aggregate → DPU-enabled configuration
OS	Ubuntu 22.04.3 LTS
GRUB	GRUB_CMDLINE_LINUX_DEFAULT="ipv6.disable=1" GRUB_CMDLINE_LINUX="intel_iommu=on iommu=pt intremap=no_x2apic_optout intel_pstate=disable default_hugepagesz=1g hugepagesz=1g hugepages=128 irqpoll"



System Under Test (SUT)

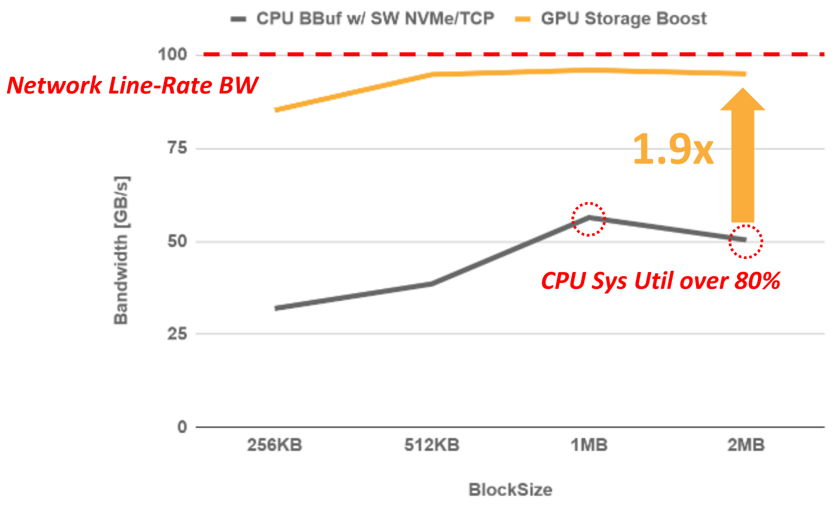
Eval 1: FIO Microbenchmark – Software Setup



FIO back-end (ioengines) was modified to enable ROCm and GSB

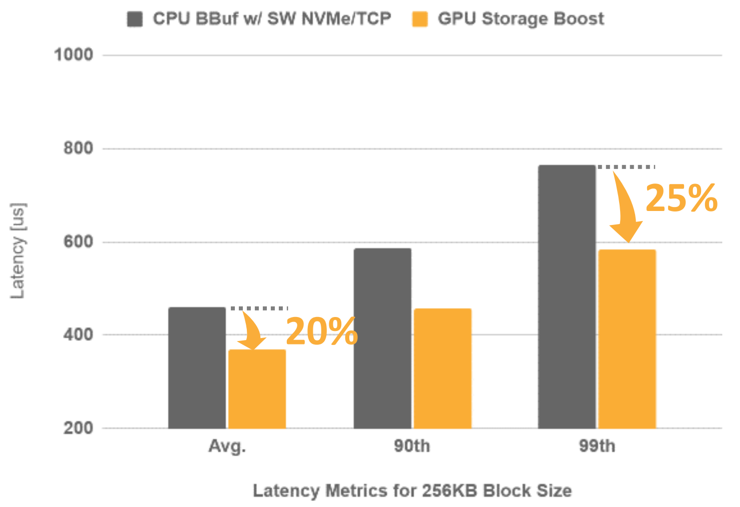
Eval 1: FIO Microbenchmark – Results

Data Movement Bandwidth



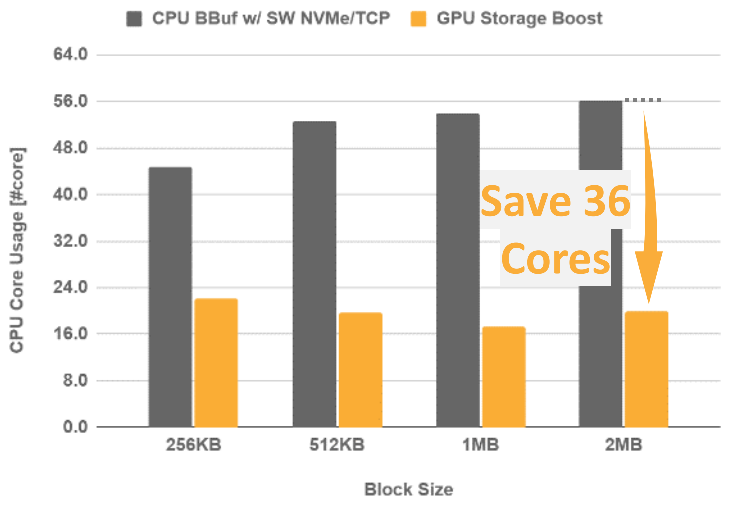
1.7x~2.6x Higher Bandwidth

Data Movement Latency



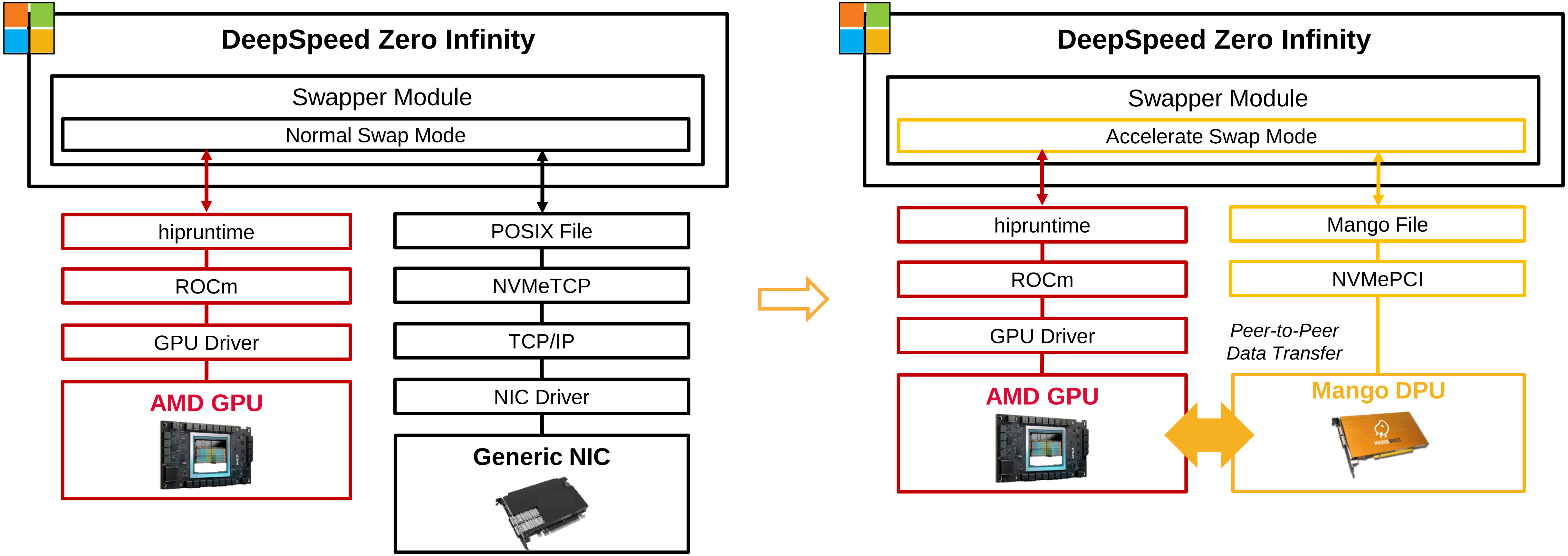
25% Lower Latency

CPU Cores Used



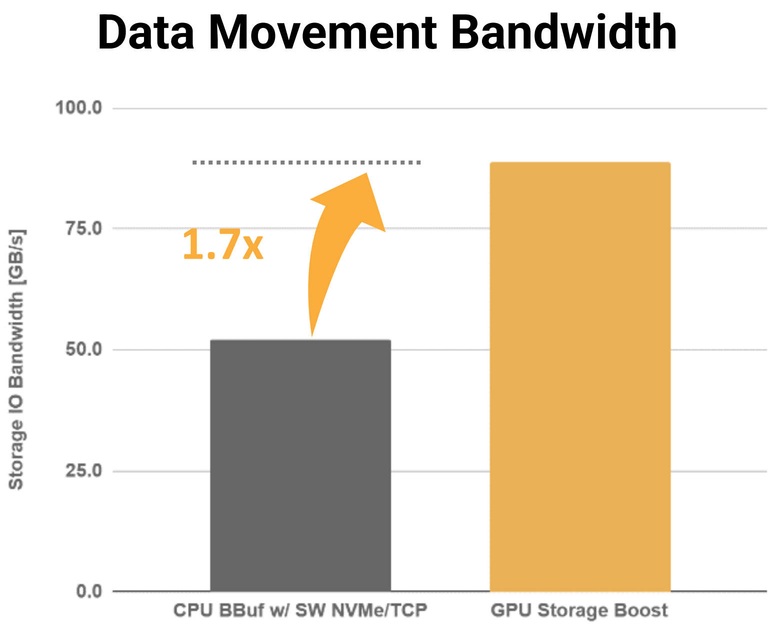
22~36 CPU cores Saved

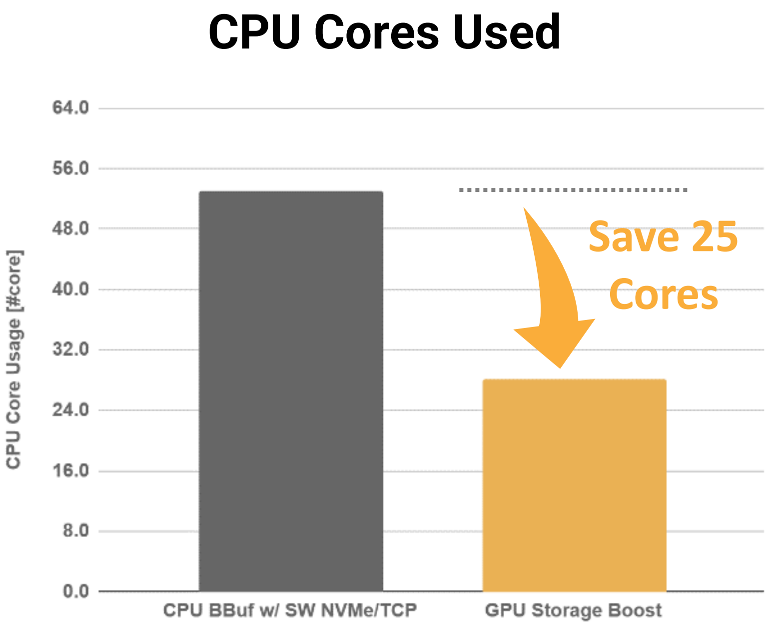
Eval 2: DeepSpeed Workload – Software Setup



DeepSpeed back-end (swapper module) was modified to enable GSB

Eval 2: DeepSpeed Workload – Results





Provide higher bandwidth and lower cpu utilization in the state-of-the-art AI training framework

Note: we also have results with another AMD GPU (MI210), showing similar benefits. Contact us for detail

Summary

- **Efficient storage system** is becoming a key-factor in AI system
  - Need to keep GPU compute busy, but not enough local device memory to keep large AI models/data/params
- AMD provides state-of-the-art AI ecosystems: **AMD Instinct™ GPU and AMD ROCm™ Software**
- Data Processing Unit (DPU) can improve storage system efficiency and performance
  - MangoBoost offers comprehensive DPU solutions, such as **GPU-storage-boost**
- Case study: Llama training with MangoBoost's storage solution (i.e., GPU Storage Boost)
  - Improve MicroBenchmark **throughput by 1.7x-2.6x** and **save 22-37 CPU cores**
  - Improve AI training storage access **throughput by 1.7x** and **save 25 CPU cores**
- **Demo is available upon request** (contact@mangoboost.io)

Disclaimer

All of the information contained in this document is considered confidential. The performance claims in this document are based on the internal cluster environment. Actual performance may vary depending on the server configuration. Software and workloads used in performance tests may have been optimized for performance only on MangoBoost products. Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. Results that are based on pre-production systems and components as well as results that have been estimated or simulated using MangoBoost reference platform for informational purposes only. Results may vary based on future changes to any systems, components, specifications, or configurations. Statements in this document that refer to future plans or expectations are forward-looking statements. These statements are based on current expectations and involve many risks and uncertainties that could cause actual results to differ materially from those expressed or implied in such statements. MangoBoost does not guarantee any specific outcome. Nothing contained herein is, or shall be relied upon as, a promise or representation or warranty as to future performance of MangoBoost or any MangoBoost product. The information contained herein shall not be deemed to expand in any way the scope or effect of any representations or warranties contained in the definitive agreement for MangoBoost products.

The information contained herein may not be reproduced in whole or in part without prior written consent of MangoBoost. The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors. The information contained herein is subject to change and may be rendered inaccurate for many reasons, including but not limited to product and roadmap changes, component and motherboard version changes, new model and/or product releases, product differences between differing manufacturers, software changes, BIOS flashes, firmware upgrades, or the like. MangoBoost assumes no obligation to update or otherwise correct or revise this information and MangoBoost reserves the right to make changes to the content hereof from time to time without any notice. Nothing contained herein is intended by MangoBoost, nor should it be relied upon, as a promise or a representation as to the future.

MANGOBOOST MAKES NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE CONTENTS HEREOF AND ASSUMES NO RESPONSIBILITY FOR ANY INACCURACIES, ERRORS OR OMISSIONS THAT MAY APPEAR IN THIS INFORMATION.

Please take a moment to rate this session.

Your feedback is important to us.

