



# FINANCIAL ANALYST DAY 2022

together we advance\_

## Building on Data Center Leadership

---

**Forrest Norrod**

Senior Vice President and General Manager, Data Center Solutions Business Group

# Cautionary Statement

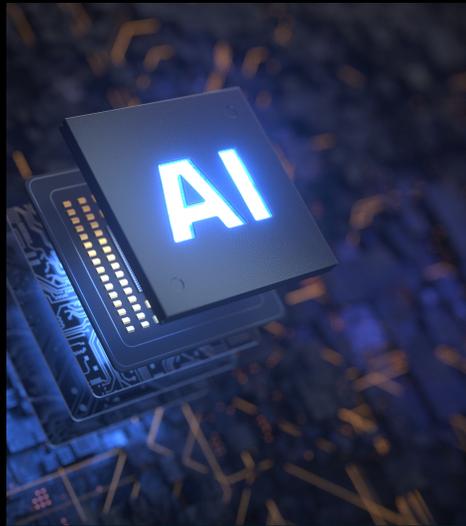
This presentation contains forward-looking statements concerning Advanced Micro Devices, Inc. (AMD) including, but not limited to, the timing, availability, features, functionality and expected benefits of AMD's data center products; trends and TAM of the data center market; and AMD's data center roadmap, which are made pursuant to the Safe Harbor provisions of the Private Securities Litigation Reform Act of 1995. Forward-looking statements are commonly identified by words such as "would," "may," "expects," "believes," "plans," "intends," "projects" and other terms with similar meaning. Investors are cautioned that the forward-looking statements in this presentation are based on current beliefs, assumptions and expectations, speak only as of the date of this presentation and involve risks and uncertainties that could cause actual results to differ materially from current expectations. Such statements are subject to certain known and unknown risks and uncertainties, many of which are difficult to predict and generally beyond AMD's control, that could cause actual results and other future events to differ materially from those expressed in, or implied or projected by, the forward-looking information and statements. Investors are urged to review in detail the risks and uncertainties in AMD's Securities and Exchange Commission filings, including but not limited to AMD's most recent reports on Forms 10-K and 10-Q.

AMD does not assume, and hereby disclaims, any obligation to update forward-looking statements made in this presentation, except as may be required by law.

# MEGATRENDS DRIVING COMPUTE IN DATA CENTERS



**Cloud**



**AI**



**HPC**



**Data Analytics**



**Video Streaming**

# POWERING THE MODERN DATA CENTER

AMD  
EPYC

Server CPUs

AMD  
INSTINCT

GPU Accelerators

AMD  
PENSANDO

XILINX  
ALVEO™

SmartNICs and DPUs

XILINX  
ALVEO™

XILINX  
VERSAL™

FPGAs and  
Adaptive SoCs

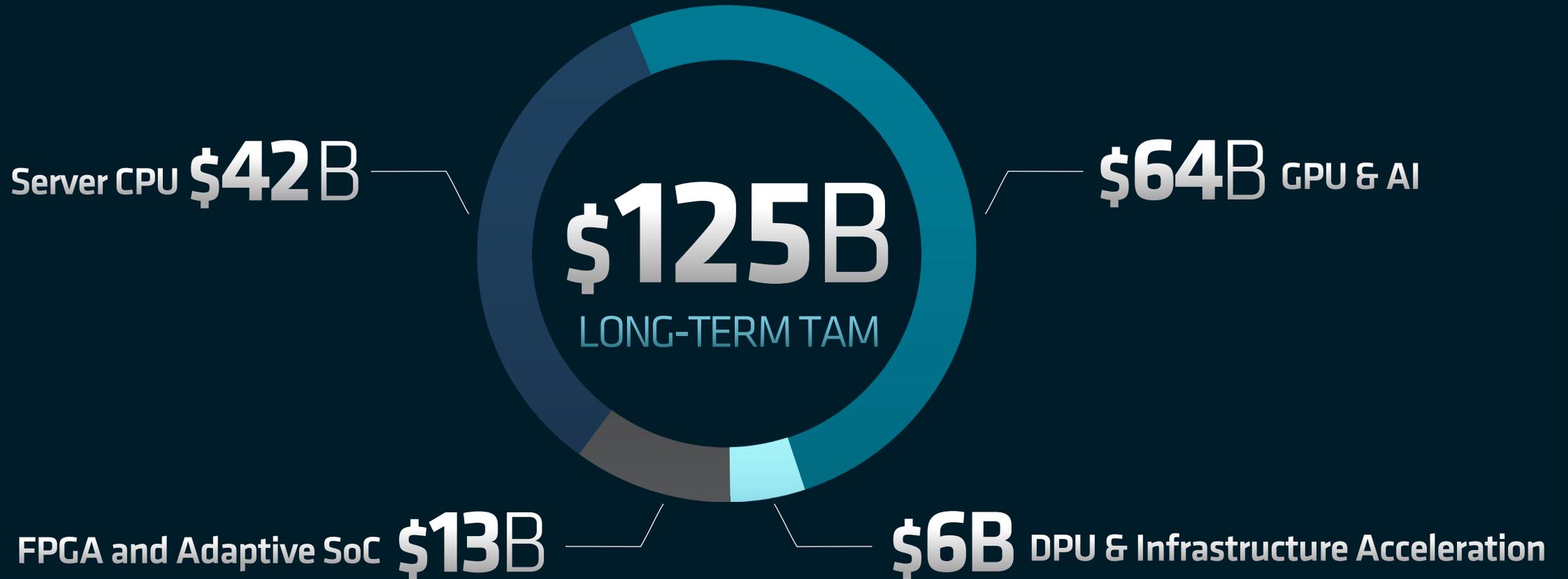
PERFORMANCE

EFFICIENCY

SECURITY



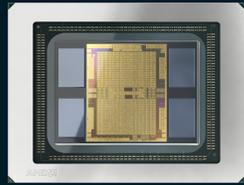
# EXPANDING DATA CENTER TAM



Based on AMD internal data



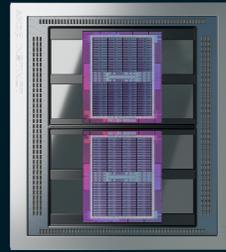
# OUR JOURNEY IN GPU ACCELERATION



AMD Instinct™ **MI100**  
AMD CDNA™

## Ecosystem Growth

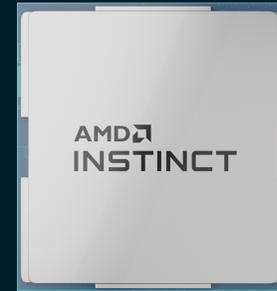
First purpose-built GPU architecture for the data center



AMD Instinct™ **MI200**  
AMD CDNA™ 2

## Driving HPC and AI to a New Frontier

First multi-die data center GPU expands scientific discovery and brings choice to AI training



AMD Instinct™ **MI300**  
AMD CDNA™ 3

## Data Center APU

Breakthrough architecture designed for leadership efficiency and performance for HPC and AI

2020

2023



# AMD INSTINCT™ MI200

## DEFINING LEADERSHIP IN HPC

| Performance Metric | MI250X Advantage<br>vs A100 |
|--------------------|-----------------------------|
| FP64 Vector        | 4.9X                        |
| FP64 Matrix        | 4.9X                        |
| FP32 Vector        | 2.5X                        |
| FP16, BF16         | 1.2X                        |
| Memory Size        | 1.6X                        |
| Memory Bandwidth   | 1.6X                        |
| OpenMM             | ~ 2.4X                      |
| HPL                | ~ 2.8X                      |



### 3<sup>RD</sup> GEN AMD INFINITY ARCHITECTURE

- CPU & GPU Memory Coherence
- Exceptional System Bandwidth & Performance

See Endnotes MI200-01, MI200-07, MI200-24A, MI200-26B



# LEADING THE EXASCALE ERA

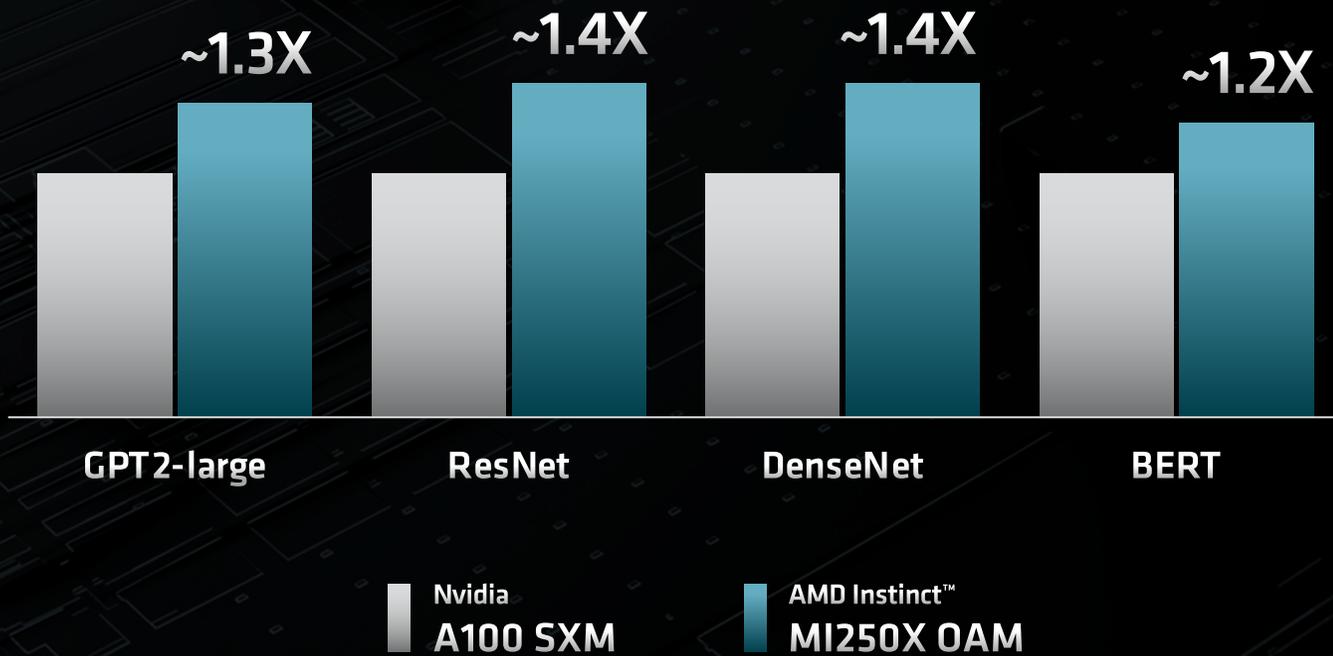
- Powering World's **#1 Supercomputer**  
First to break Exascale barrier
- Powering World's **#1 Green Supercomputer**  
8 of top 10 most efficient systems rely on AMD
- Powering World's **#1 AI Supercomputer**  
More than 3X the previous record holder
- 95% growth in TOP500 systems Year-over-Year  
Powering more than half of all new systems



**AMD**  
**INSTINCT**

BRINGING CHOICE  
TO AI TRAINING

AI Training Performance



See Endnotes MI200-57, MI200-59, MI200-61, MI200-63



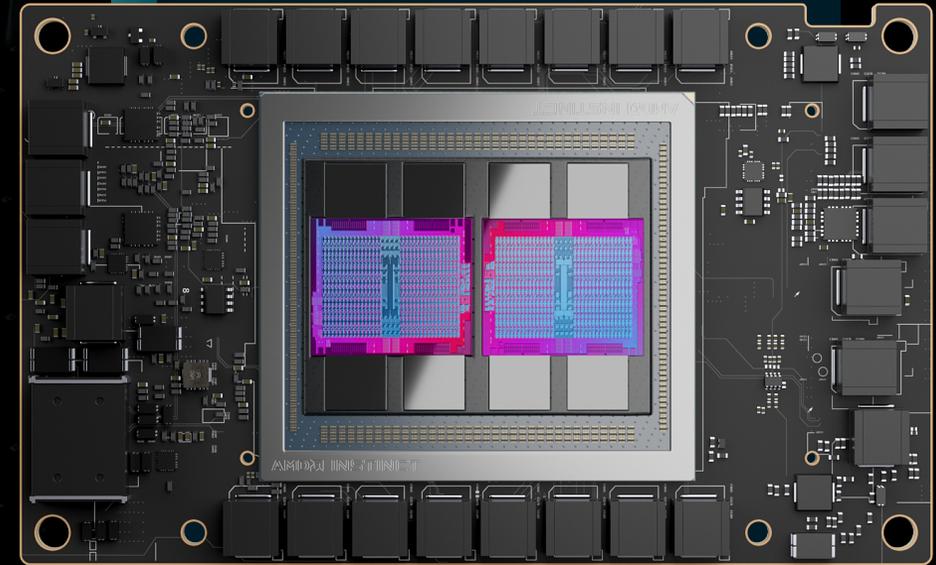
AMD  
INSTINCT

Microsoft

“As part of our long-term partnership with AMD, I’m excited to share that **Azure will be the first public cloud to deploy clusters of AMD’s flagship MI200 GPUs for large scale AI training.** We’ve already started testing these clusters using some of our own AI workloads with great performance.”

**Kevin Scott**

Executive Vice President and Chief Technology Officer, Microsoft



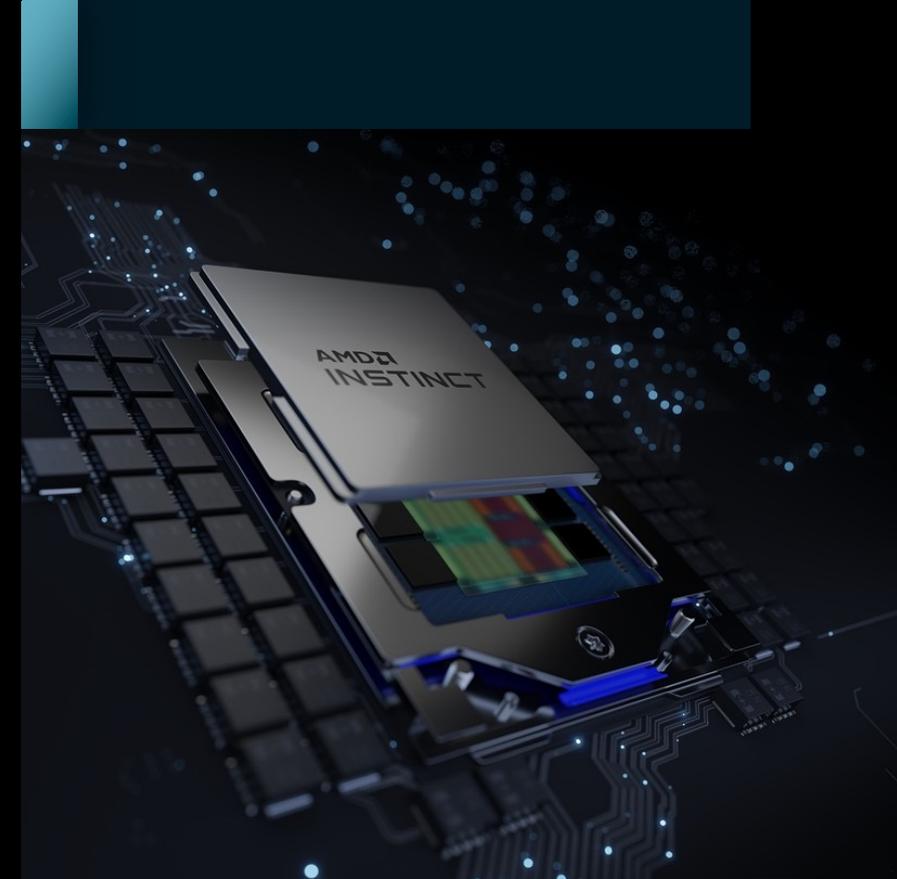
# AMD INSTINCT™ MI300

## THE WORLD'S FIRST DATA CENTER APU

- 4<sup>th</sup> Gen AMD Infinity Architecture: AMD CDNA™ 3 and EPYC™ CPU “Zen 4” Together  
CPU and GPU cores share a unified on-package pool of memory
- Groundbreaking 3D Packaging  
CPU | GPU | Cache | HBM
- Designed for Leadership Memory Bandwidth and Application Latency
- APU Architecture Designed for Power Savings Compared to Discrete Implementation

---

Available **2023**



**>8X**

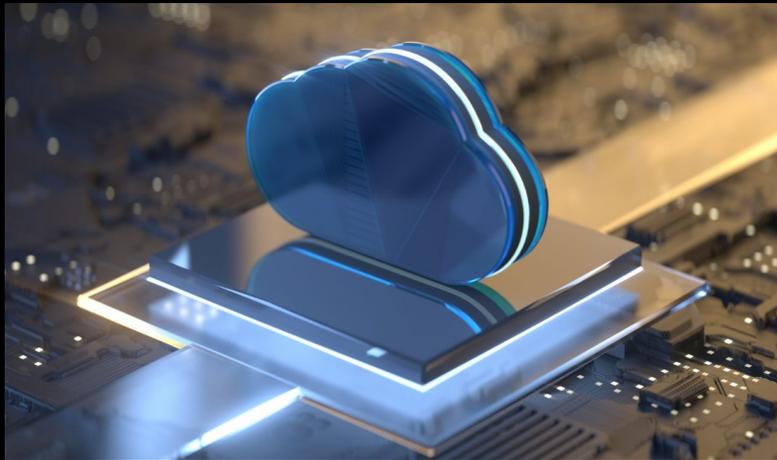
**Expected AI Training Performance  
vs. MI250X**

See Endnote MI300-03. Preliminary data and projections, subject to change.



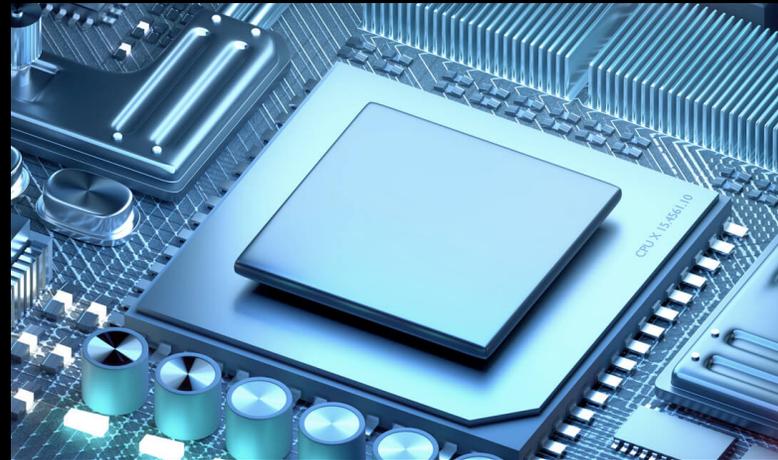
# SCALING THE SOFTWARE DEFINED DATA CENTER

## INFRASTRUCTURE ACCELERATION AND NETWORKING ARE ESSENTIAL



### Cloud and Virtualization

Cloud services overhead can consume up to 30% of CPU cores



### Shift of Compute to Edge

Extend visibility and management beyond the data center boundary



### Security of Data Traffic

Secure data at rest and in flight at all endpoints

# AMD NETWORKING TECHNOLOGY



## Solarflare™

Ultra Low Latency NICs

Ubiquitous in Financial Exchanges

Crucial for High Frequency Trading

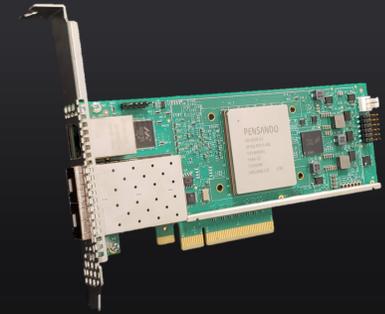


## Alveo™

FPGA-Based Adaptive Acceleration

Programmable Hardware for  
Custom Network Functions

Extreme Packet Processing Rates



## Pensando™

P4 Packet Processing Engines

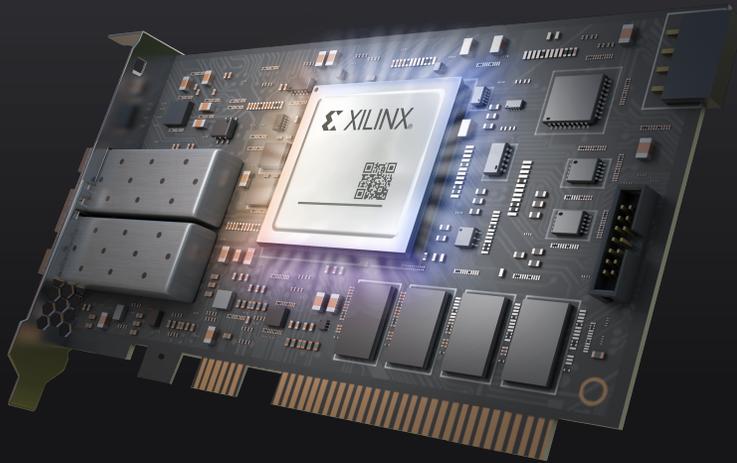
Tens of Millions of Network Flows

Multiple Concurrent Services at Line Rate



# ALVEO™

## ADAPTIVE NETWORK ACCELERATION



- Shipping to Hyperscale Customers
- Accelerates Custom and Evolving Network Functions
- Extends Confidential Computing to Network Interface
- 2 x 200G | 400M Packets per Second

**Next Generation in 2024**

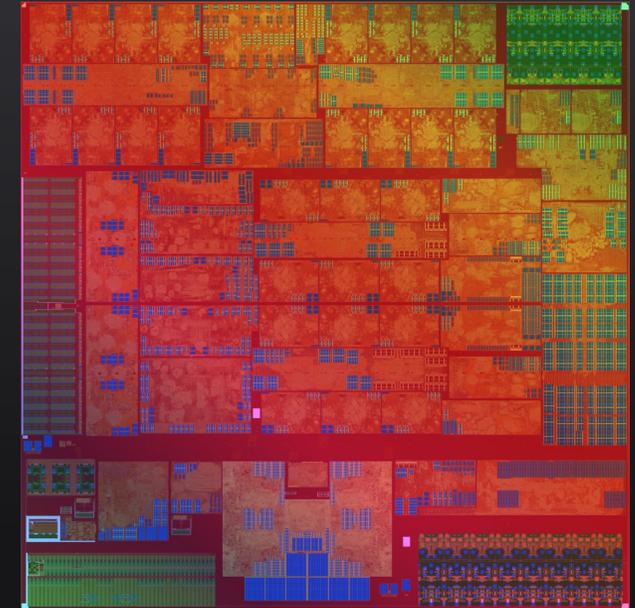
# PENSANDO™

## WORLD'S MOST INTELLIGENT DPU

- 144 P4 Packet Processors
- Fully Programmable Control, Data, and Management Planes
- Supports Tens of Millions of Network Flows
- Concurrent Services at Line Rate Performance  
Network | Security | Storage | Telemetry

---

2<sup>nd</sup> Generation | 2 x 200G | 7nm  
In Production Today



**“ELBA” DPU**

# PENSANDO™ SOFTWARE

TURNKEY INFRASTRUCTURE SOLUTION FOR  
THE SOFTWARE DEFINED DATA CENTER

- Based on Open Standards and APIs
- Easily Incorporate and Accelerate New Services
- “Zero Trust Security” Throughout
- Works with Existing Management Tools
- Solutions Deployed Today



# DPU ACCELERATION ACROSS THE DATA CENTER



P4 SmartNIC

aruba



Smart Switch

NetApp



Storage

**DELL** Technologies

  
Hewlett Packard  
Enterprise

**IBM.**

 Microsoft

**ORACLE**

**vmware**<sup>®</sup>

# AMD SECURING THE DATA CENTER

## COMPREHENSIVE SECURITY FEATURES FROM CORE TO EDGE

AMD  
EPYC

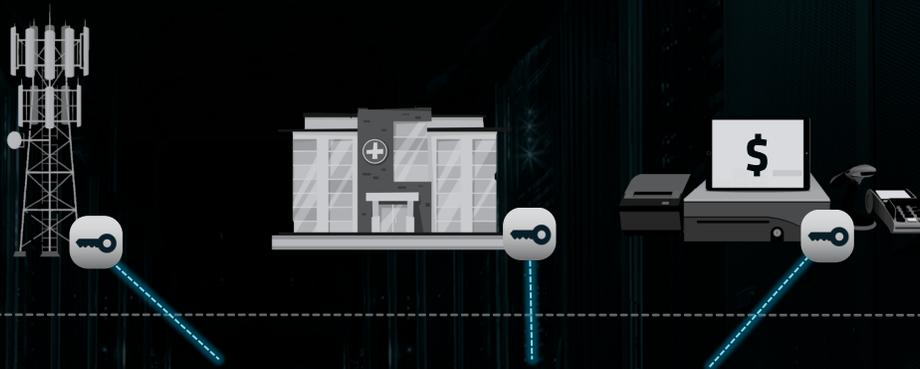


**Distributed Security Services**  
Extend the perimeter

AMD  
PENSANDO



**Edge and Beyond**



AMD  
EPYC



**Smart Switch & Appliance**  
Help secure east-west traffic  
and legacy endpoints

AMD  
PENSANDO



Appliance

Smart Switch



**Infrastructure**

AMD  
EPYC



**Advanced SmartNIC**  
Help secure every endpoint

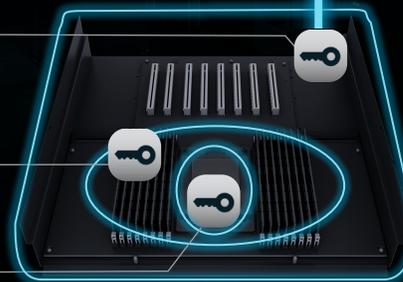
AMD  
PENSANDO



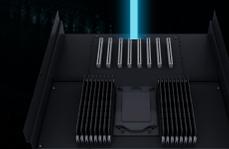
**Confidential Compute**  
Help secure application environment

AMD  
INSTINCT

**Root of Trust**  
Help secure CPU at boot



Modern Server



Legacy Server

**Compute Node**

0 1 1 0 1 1 0 1 1  
1 1 0 1 1 0 1 1 0  
0 0 1 0 0 1 0 0 1  
1 0 0 1 0 0 1 0 0  
0 1 1 0 1 1 0 1 1  
1 1 0 1 1 0 1 1 0  
0 0 1 0 0 1 0 0 1  
1 1 0 1 1 0 1 1 0  
0 0 1 0 0 1 0 0 1  
0 1 1 0 1 1 0 1 1  
1 1 1 1 1 1 1 1 1  
0 0 1 0 0 1 0 0 1

**Advanced Telemetry**  
Monitor and detect  
threats in real time

0 1 0 0 1 0 0 1 0  
1 1 0 1 1 0 1 1 0  
0 0 1 0 0 1 0 0 1  
0 0 1 0 0 1 0 0 1  
1 1 1 1 1 1 1 1 1  
0 1 1 0 1 1 0 1 1  
0 0 1 0 0 1 0 0 1  
1 1 1 1 1 1 1 1 1





OUR PATH FORWARD

# UNMATCHED DATA CENTER TECHNOLOGY

- Leadership Performance
- World-Class Efficiency
- Comprehensive Security Features

AMD  
EPYC

AMD  
INSTINCT

AMD  
PENSANDO

XILINX  
ALVEO™

XILINX  
VERSAL™



# Endnotes

- MI200-01 - World's fastest data center GPU is the AMD Instinct™ MI250X. Calculations conducted by AMD Performance Labs as of Sep 15, 2021, for the AMD Instinct™ MI250X (128GB HBM2e OAM module) accelerator at 1,700 MHz peak boost engine clock resulted in 95.7 TFLOPS peak theoretical double precision (FP64 Matrix), 47.9 TFLOPS peak theoretical double precision (FP64), 95.7 TFLOPS peak theoretical single precision matrix (FP32 Matrix), 47.9 TFLOPS peak theoretical single precision (FP32), 383.0 TFLOPS peak theoretical half precision (FP16), and 383.0 TFLOPS peak theoretical Bfloat16 format precision (BF16) floating-point performance. Calculations conducted by AMD Performance Labs as of Sep 18, 2020 for the AMD Instinct™ MI100 (32GB HBM2 PCIe® card) accelerator at 1,502 MHz peak boost engine clock resulted in 11.54 TFLOPS peak theoretical double precision (FP64), 46.1 TFLOPS peak theoretical single precision matrix (FP32), 23.1 TFLOPS peak theoretical single precision (FP32), 184.6 TFLOPS peak theoretical half precision (FP16) floating-point performance. Published results on the NVidia Ampere A100 (80GB) GPU accelerator, boost engine clock of 1410 MHz, resulted in 19.5 TFLOPS peak double precision tensor cores (FP64 Tensor Core), 9.7 TFLOPS peak double precision (FP64), 19.5 TFLOPS peak single precision (FP32), 78 TFLOPS peak half precision (FP16), 312 TFLOPS peak half precision (FP16 Tensor Flow), 39 TFLOPS peak Bfloat 16 (BF16), 312 TFLOPS peak Bfloat16 format precision (BF16 Tensor Flow), theoretical floating-point performance. The TF32 data format is not IEEE compliant and not included in this comparison. <https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/nvidia-ampere-architecture-whitepaper.pdf>, page 15, Table 1. MI200-01
- MI200-07 - Calculations conducted by AMD Performance Labs as of Sep 21, 2021, for the AMD Instinct™ MI250X and MI250 (128GB HBM2e) OAM accelerators designed with AMD CDNA™ 2 6nm FinFet process technology at 1,600 MHz peak memory clock resulted in 128GB HBM2e memory capacity and 3.2768 TFLOPS peak theoretical memory bandwidth performance. MI250/MI250X memory bus interface is 4,096 bits times 2 die and memory data rate is 3.20 Gbps for total memory bandwidth of 3.2768 TB/s ((3.20 Gbps\*(4,096 bits\*2))/8). The highest published results on the NVidia Ampere A100 (80GB) SXM GPU accelerator resulted in 80GB HBM2e memory capacity and 2.039 TB/s GPU memory bandwidth performance. <https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/a100/pdf/nvidia-a100-datasheet-us-nvidia-1758950-r4-web.pdf> MI200-07
- MI200-24A - Testing Conducted by AMD performance lab as of 10/12/2021, on a single socket Optimized 3rd Gen AMD EPYC™ CPU server with 1x AMD Instinct™ MI250X OAM (128 GB HBM2e) 560W GPU with AMD Infinity Fabric™ technology using benchmark OpenMM\_amoebagk v7.6.0, (converted to HIP) and run at double precision (8 simulations\*10,000 steps) plus AMD optimizations to OpenMM\_amoebagk that are not yet upstream resulted in a median score of 387.0 seconds or 223.2558 NS/Day Vs. NVidia DGX dual socket AMD EPYC 7742@2.25GHz CPU server with 1x NVIDIA A100 SXM 80GB (400W) using benchmark OpenMM\_amoebagk v7.6.0, run at double precision (8 simulations\*10,000 steps) with CUDA code version 11.4 resulted in a median score of 921.0 seconds or 93.8111 NS/Day. Information on OpenMM: <https://openmm.org/> Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations MI200-24A
- MI200-26B - Testing Conducted by AMD performance lab as of 10/14/2021, on a single socket Optimized 3rd Gen AMD EPYC™ CPU (64) server, with 1x AMD Instinct™ MI250X OAM (128 GB HBM2e, 560W) GPU with AMD Infinity Fabric™ technology using benchmark HPL v2.3, plus AMD optimizations to HPL that are not yet upstream. Vs. Nvidia DGX dual socket AMD EPYC 7742 (64C) @2.25GHz CPU server with 1x NVIDIA A100 SXM 80GB (400W) using benchmark HPL Nvidia container image 21.4-HPL. Information on HPL: <https://www.netlib.org/benchmark/hpl/> Nvidia HPL Container Detail: <https://nec.nvidia.com/catalog/containers/nvidia:hpc-benchmarks> Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations MI200-26B
- MI200-57- Testing Conducted by AMD performance lab as of 25/5/2022 using [SuperBench v 0.4.0](#) benchmark, GPT2-Large, EPYC/Instinct system: Dual socket, 64 core, 2nd Gen AMD EPYC™ 7002 Series CPU powered server with 8x AMD Instinct™ MI250X OAM (128 GB HBM2e) 500W GPUs with AMD Infinity Fabric™ technology. Benchmark: GPT2-Large with AMD | Microsoft optimized batch sizes tuned for GPT2-Large results for system configurations that are not yet available upstream. Benchmark Results: GPT2-large resulted in a median throughput of 8x MI250X = 761.08 Samples (Throughput)/ sec. Training model separates copies of model on each GPU; total system throughput obtained by calculating the sum of the throughput obtained on each GPU. Vs. EPYC/Nvidia system: NVIDIA DGXA100, Dual AMD EPYC 7002 Series CPUs with 8x NVIDIA A100 SXM 80GB (400W) Benchmark: GPT2-Large Commit(Container): [superbench/superbench:v0.4.0-cuda11.1.1](#) from here: (<https://hub.docker.com/r/superbench/superbench>) Benchmark Results: GPT2-large resulted in a median throughput of 8x A100 = 589.435 Samples(Throughput)/ sec. Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations MI200-57
- MI200-59 - Testing Conducted by AMD performance lab as of 25/5/2022 using [SuperBench v 0.4.0](#) benchmark, DenseNet 169/201, Framework PyTorch 1.9, EPYC/Instinct system: Dual socket, 64 core, 2nd Gen AMD EPYC™ 7002 Series CPU powered server with 8x AMD Instinct™ MI250X OAM (128 GB HBM2e) 500W GPUs with AMD Infinity Fabric™ technology, ROCm™ 5.1.0 Benchmark: DenseNet model (Median scores of DenseNet169, DenseNet201 datasets) with AMD | Microsoft optimized batch sizes tuned for DenseNet results for system configurations that are not yet available upstream. Commit(Container): [computeqc/superbench:rocm5.1.3\\_superbench04](#) from [here](#) Benchmark Results: DenseNet testing resulted in median throughput scores of 8x MI250X: DenseNet169 = 6567.769, DenseNet201 = 5254.561 Samples (Throughput)/ sec. Training model separates copies of model on each GPU; total system throughput obtained by calculating the sum of the throughput obtained on each GPU. Vs. EPYC/Nvidia system: NVIDIA DGXA100, Dual AMD EPYC 7002 Series CPUs with 8x NVIDIA A100 SXM 80GB (400W), CUDA 11.6 and Driver Version 510.47.03, Commit(Container): [superbench/superbench:v0.4.0-cuda11.1.1](#) from here: (<https://hub.docker.com/r/superbench/superbench>) Benchmark Results: DenseNet testing resulted in median throughput of 8x A100: DenseNet169 = 4712.705, DenseNet201 = 3877.668 Samples (Throughput)/ sec. Details on SuperBench found [here](#) Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations MI200-59
- MI200-61 - Testing Conducted by AMD performance lab as of 5/25/2022 using [SuperBench v 0.4.0](#) benchmark, Bert-Base, EPYC/Instinct system: Dual socket, 64 core, 2nd Gen AMD EPYC™ 7002 Series CPU powered server with 8x AMD Instinct™ MI250X OAM (128 GB HBM2e) 500W GPUs with AMD Infinity Fabric™ technology, ROCm™ 5.1.0, PyTorch 1.9 Benchmark: Bert-Base with AMD | Microsoft optimized batch sizes tuned for Bert-Base results for system configurations that are not yet available upstream. Commit (Container): [computeqc/superbench:rocm5.1.3\\_superbench04](#) from [here](#) Benchmark Results: Bert-Base resulted in a median throughput of 8x MI250X = 6230.021 Samples (Throughput)/ sec. Training model separates copies of model on each GPU; total system throughput obtained by calculating the sum of the throughput obtained on each GPU. Vs. EPYC/Nvidia system: NVIDIA DGXA100, Dual AMD EPYC 7002 Series CPUs, CUDA 11.6 and Driver Version 510.47.03, Commit(Container): [superbench/superbench:v0.4.0-cuda11.1.1](#) from [here](#) Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations MI200-61

# Endnotes (Cont.)

- MI200-63 - Testing Conducted by AMD performance lab as of 25/5/2022 using [SuperBench v 0.4.0](#) benchmark, Resnet 50/101/152 datasets. All configurations used Framework: PyTorch 1.9 EPYC/Instinct system: Dual socket, 64 core, 2nd Gen AMD EPYC™ 7002 Series CPU powered server with 8x AMD Instinct™ MI250X OAM (128 GB HBM2e) 500W GPUs with AMD Infinity Fabric™ technology, ROCm™ 5.1.0 Benchmark: Resnet model (Median scores of Resnet50, Resnet 101, Resnet152 datasets) with AMD | Microsoft optimized batch sizes tuned for Resnet results for system configurations that are not yet available upstream. Commit(Container): `computeqcq/superbench:rocm5.1.3_superbench04` from [here](#) Benchmark Results: Resnet testing resulted in median throughput scores of 8x MI250X: resnet50 = 9708.873, resnet101 = 6705.041, resnet152 = 5250.635 Samples (Throughput)/ sec. Training model separates copies of model on each GPU; total system throughput obtained by calculating the sum of the throughput obtained on each GPU. Vs. EPYC/Nvidia system: NVIDIA DGXA100, Dual socket 64-core AMD EPYC 7742 Series CPUs, with 8x NVIDIA A100 SXM 80GB (400W), CUDA 11.6 and Driver Version 510.47.03, OS: Benchmark: Resnet model (Median scores of Resnet50, Resnet 101, Resnet152 datasets) Commit(Container): `superbench/superbench:v0.4.0-cuda11.1.1` from [here](#) Benchmark Results: Resnet testing resulted in median throughput of 8x A100: resnet50 = 7057.052, resnet101 = 4968.324, resnet152 = 3806.318 Samples (Throughput)/ sec. Details on SuperBench found [here](#) Server manufacturers may vary configurations, yielding different results. Performance may vary based on use of latest drivers and optimizations MI200-63
- MI300-03 - Measurements by AMD Performance Labs June 4, 2022. MI250X FP16 (306.4 estimated delivered TFLOPS based on 80% of peak theoretical floating-point performance). MI300 FP8 performance based on preliminary estimates and expectations. Actual results based on production silicon may vary. MI300-03