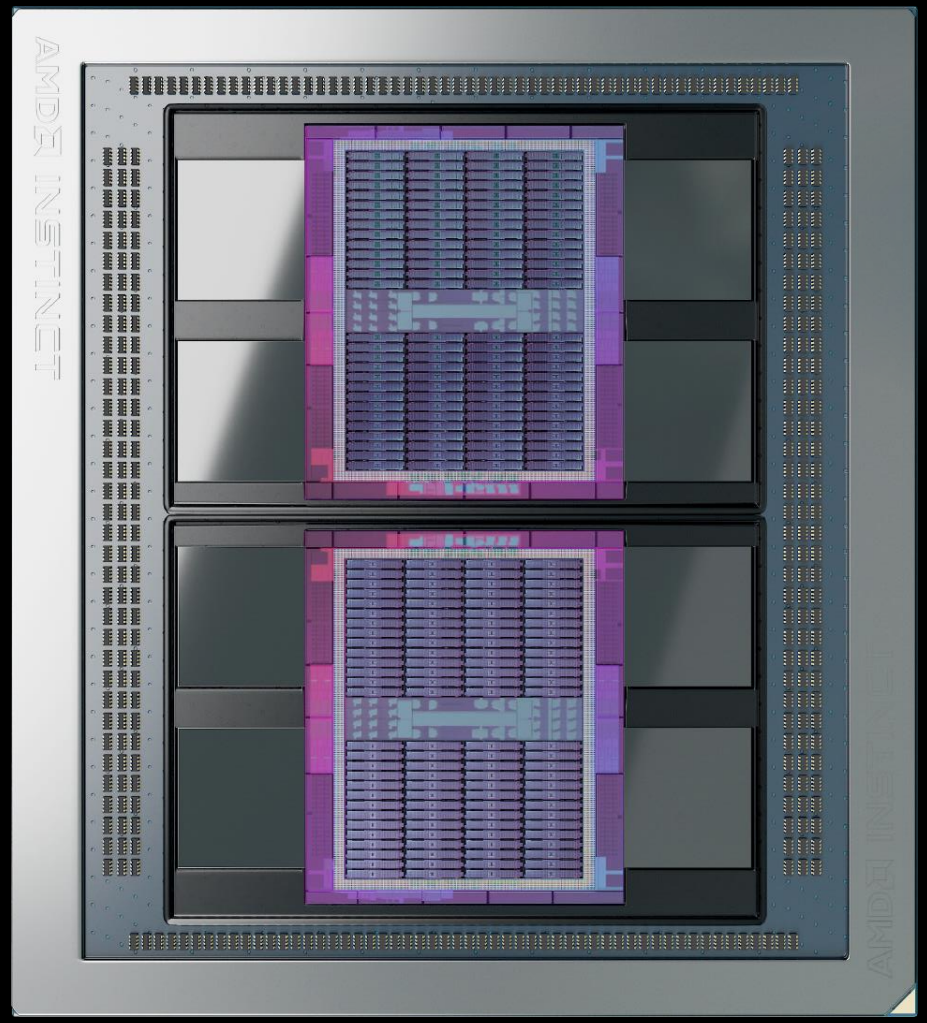
Introduction to the AMD CDNA™ 2 Architecture

Suyash Tandon, Justin Chang, Julio Maia, Noel Chalmers, Paul T. Bauman, Nicholas Curtis, Nicholas Malaya, Alessandro Fanfarillo, Jose Noudohouenou, Chip Freitag, Damon McDougall, Noah Wolfe, Jakub Kurzak, Samuel Antao, George Markomanolis, Bob Robey, Gina Sitaraman

AMD @HLRS
Sept 25-28th, 2023

AMD
together we advance_

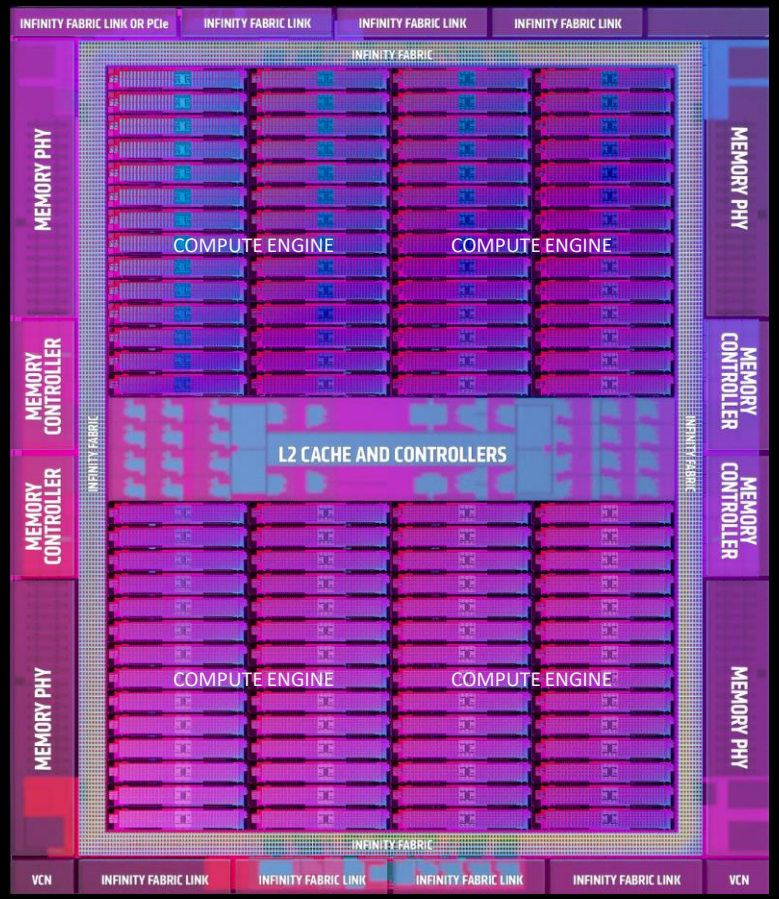


AMD INSTINCT™ MI250X

WORLD'S MOST ADVANCED DATA CENTER ACCELERATOR

58B Transistors in 6nm	**220** Compute Units
880 2nd Gen Matrix Cores	**128** GB HBM2E @ 3.2 TB/s

https://www.amd.com/system/files/documents/amd-cdna2-white-paper.pdf



2ND GENERATION CDNA ARCHITECTURE

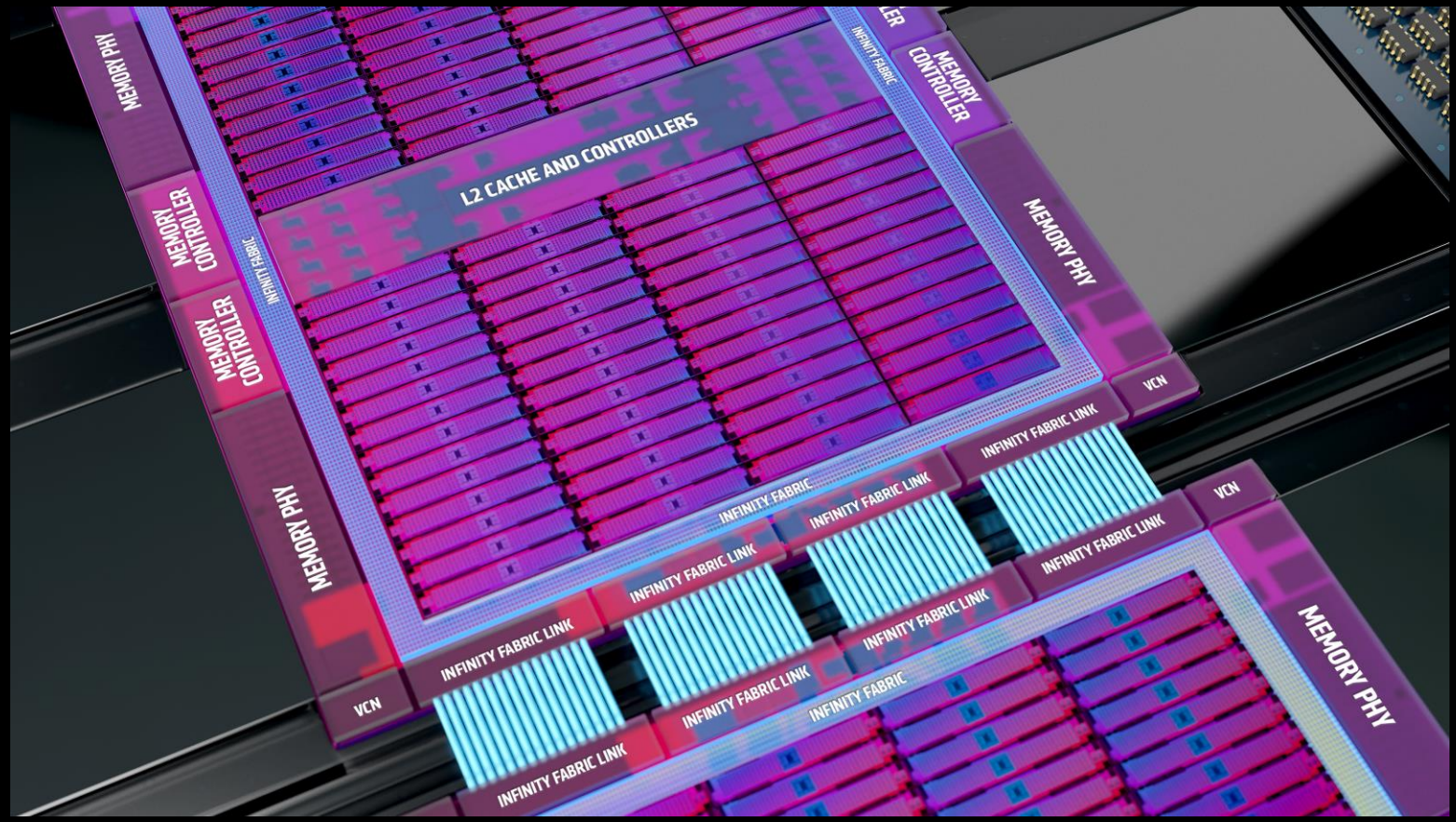
TAILORED-BUILT FOR HPC & AI

- TSMC 6NM TECHNOLOGY
- UP TO 110 CU PER GRAPHICS CORE DIE
- 4 MATRIX CORES PER COMPUTE UNIT
- MATRIX CORES ENHANCED FOR HPC
- 8 INFINITY FABRIC LINKS PER DIE
- SPECIAL FP32 OPS FOR DOUBLE THROUGHPUT

MULTI-CHIP DESIGN

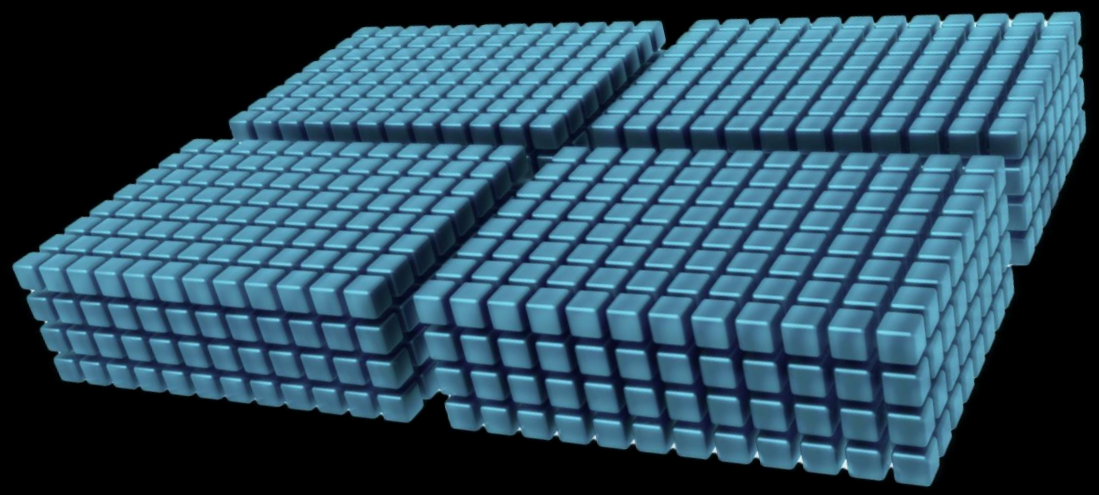
TWO GPU DIES IN PACKAGE TO MAXIMIZE COMPUTE & DATA THROUGHPUT

- INFINITY FABRIC FOR CROSS-DIE CONNECTIVITY
- 4 LINKS RUNNING AT 25GBPS
- 400GB/S OF BI-DIRECTIONALBANDWIDTH



2nd GENERATION MATRIX CORES

OPTIMIZED COMPUTE UNITS FOR SCIENTIFIC COMPUTING

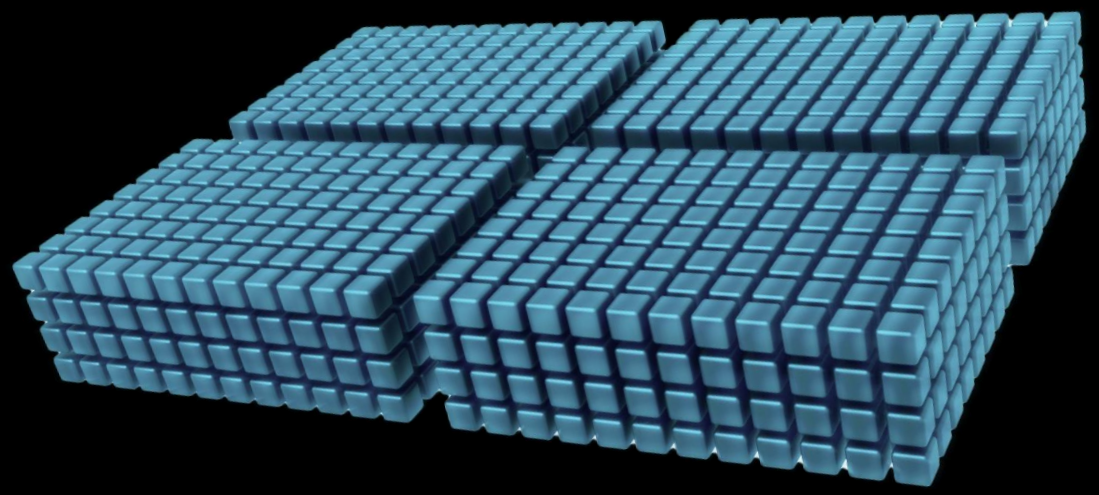


DOUBLE PRECISON (FP64)
MATRIX CORE THROUGHPUT
REPRESENTATION

MI100 MATRIX CORES OPS/CLOCK/COMPUTE UNIT	MI250X MATRIX CORES OPS/CLOCK/COMPUTE UNIT
No FP64 Matrix Core	256 FP64
256 FP32	256 FP32
1024 FP16	1024 FP16
512 BF16	1024 BF16
512 INT8	1024 INT8

2nd GENERATION MATRIX CORES

OPTIMIZED COMPUTE UNITS FOR SCIENTIFIC COMPUTING



- Current support for using MFMA instructions:
  - AMD libraries: rocBLAS
  - Intrinsics
  - Inline assembly
- Not currently supported:
  - Libraries of device functions, utilizing the matrix operations, that can be called from kernels
  - Abstraction frameworks (Kokkos, Raja, OCCA)
    - These would have to use one of the other mechanisms internally

AMD Matrix Cores Blog Post: https://gpuopen.com/learn/amd-lab-notes/amd-lab-notes-matrix-cores-readme/

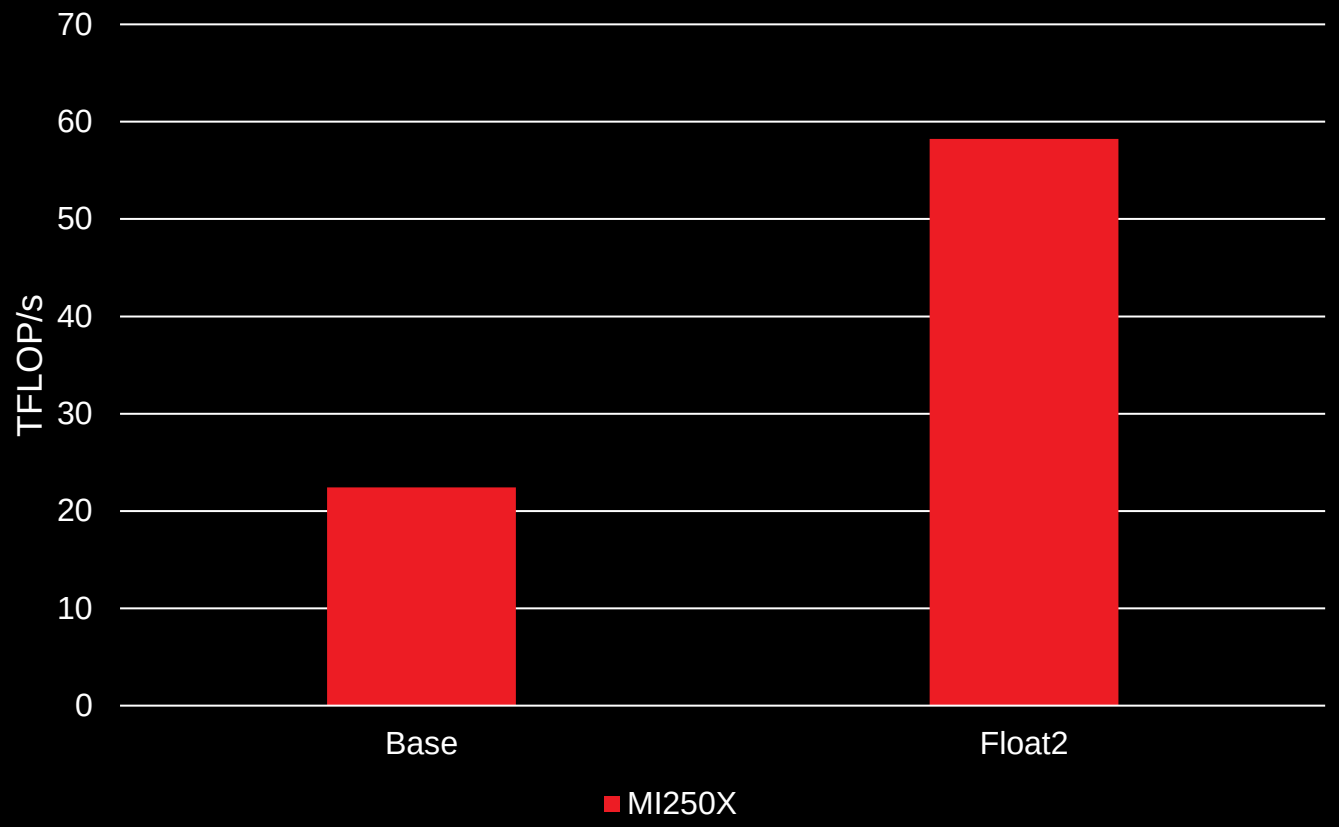
NEW IN AMD INSTINCT MI250X

PACKED FP32

FP64 PATH USED TO EXECUTE TWO COMPONENT VECTOR INSTRUCTIONS ON FP32

DOUBLES FP32 THROUGHPUT PER CLOCK PER COMPUTE UNIT

pk_FMA, pk_ADD, pk_MUL, pk_MOV operations



https://www.amd.com/en/technologies/infinity-hub/mini-hacc

From AMD MI100 to AMD MI250X

MI100

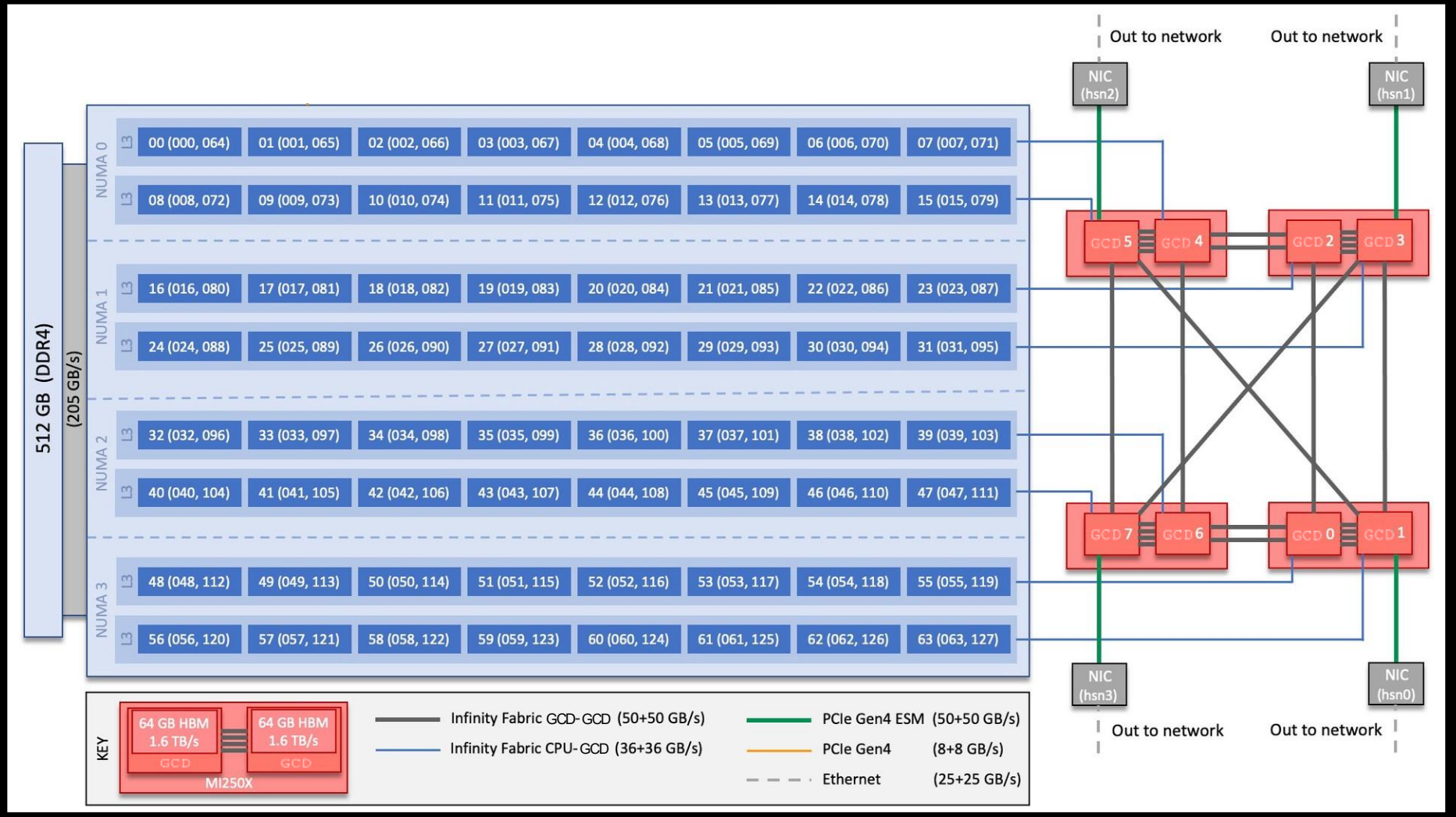
- One graphic compute die (GCD)
- 32GB of HBM2 memory
- 11.5 TFLOPS peak performance per GCD
- 1.2 TB/s peak memory bandwidth per GCD
- 120 CU per GPU
- The interconnection is attached on the CPU

AMD CDNA™ 2 white paper:
https://www.amd.com/system/files/documents/amd-cdna2-white-paper.pdf

MI250X

- Two graphic compute dies (GCDs)
- 64GB of HBM2e memory per GCD (total 128GB)
- 26.5 TFLOPS peak performance per GCD
- 1.6 TB/s peak memory bandwidth per GCD
- 110 CU per GCD, totally 220 CU per GPU
- The interconnection is attached on the GPU (not on the CPU)
- Both GCDs are interconnected with 200 GB/s per direction
- 128 single precision FMA operations per cycle
- AMD CDNA 2 Matrix Core supports double-precision data
- Memory coherency

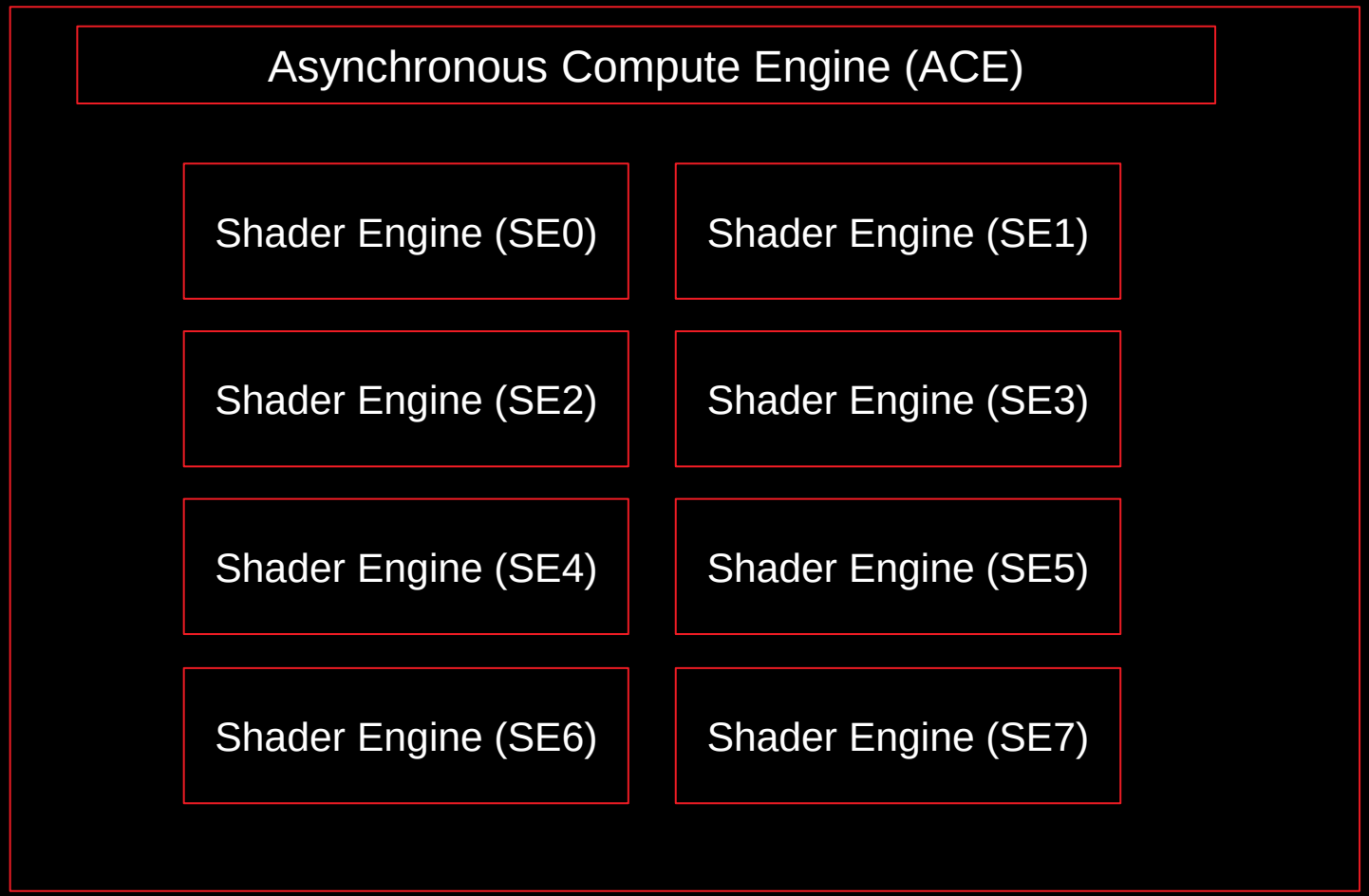
MI250X Node Architecture



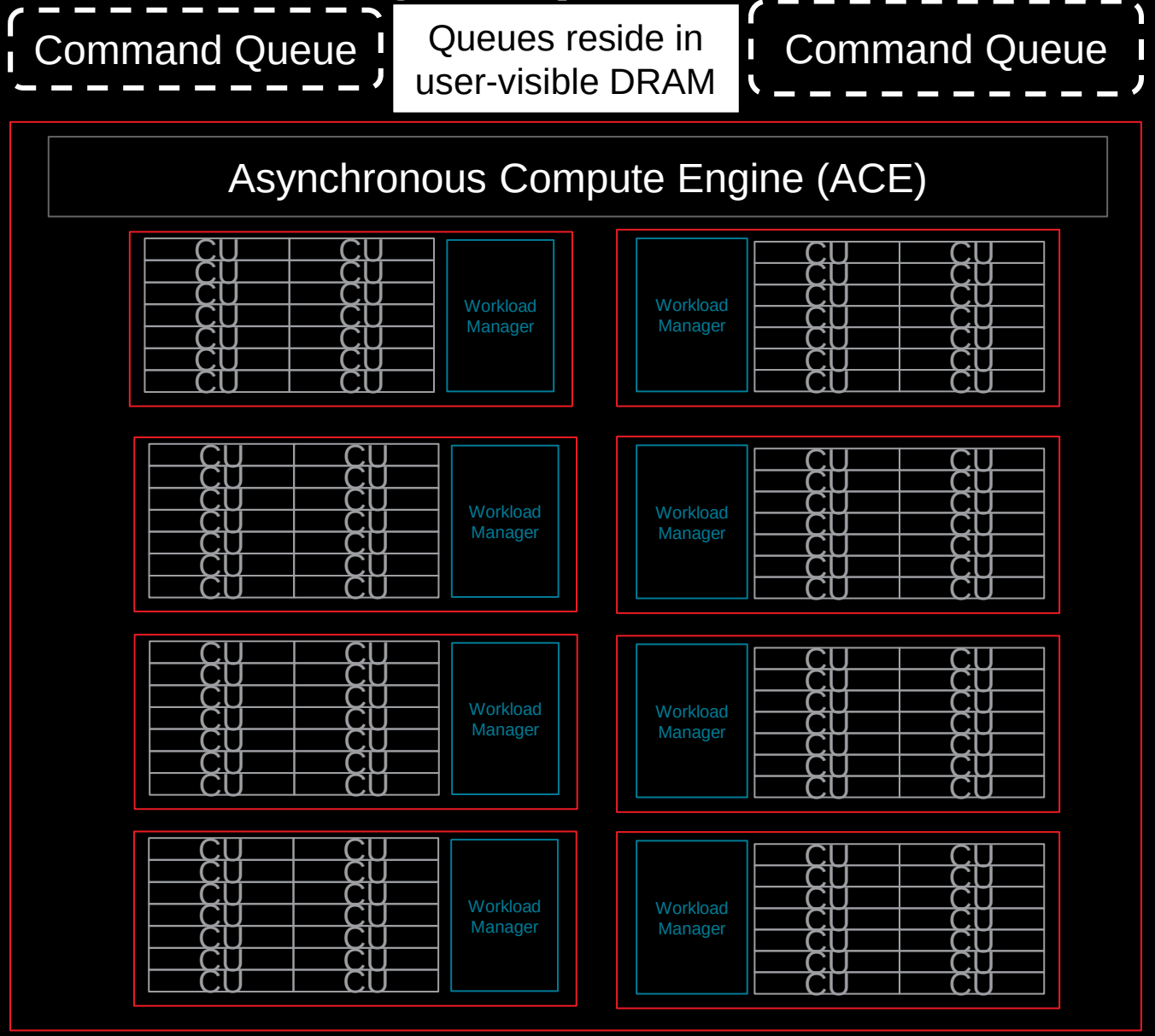
- 64 cores on a single socket CPU
- 4 MI250X GPUs, each with 2 GCDs
  - Each GCD is presented as a GPU device to `rocm-smi`
- 512 GB of DDR4 RAM
- Infinity Fabric™ links between GCDs and between GCDs and CPU cores
- 4 NICs attached to odd numbered GCDs

Courtesy: https://docs.olcf.ornl.gov/systems/frontier_user_guide.html#frontier-compute-nodes

AMD GCN GPU Hardware Layout (MI250X one GCD)



AMD GCN GPU Hardware Layout (MI250X one GCD)

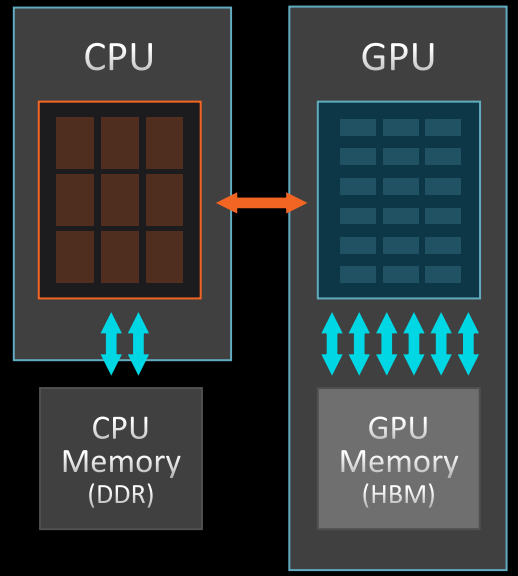


CPU+GPU unification – MI300 APU

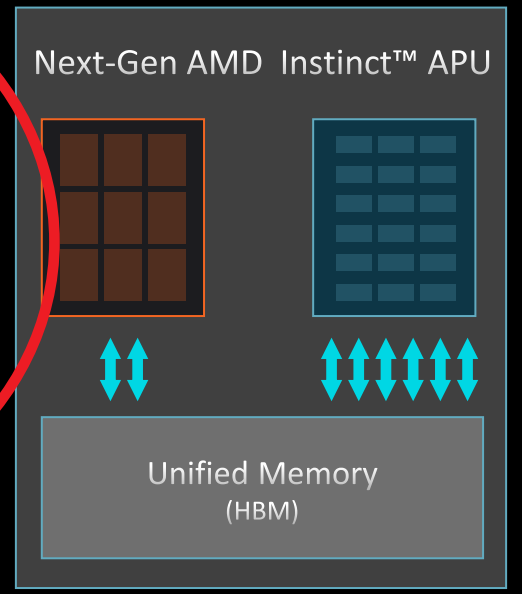
- Keep data in the same place – move the compute
- Abstraction lifts the need for user-defined copies
- Can we expand the abstraction beyond the APU?

AMD CDNA™ 2 Coherent Memory Architecture → AMD CDNA™ 3 Unified Memory APU Architecture

- **Simplifies Programming**
- **Low Overhead 3rd Gen Infinity Interconnect**
- **Modular System Design**

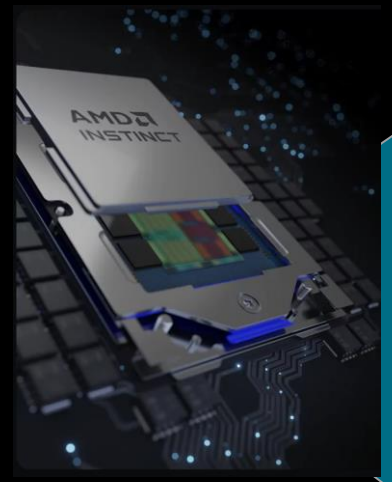


- **Eliminates Redundant Memory Copies**
- **High-Efficiency 4th Gen AMD Infinity Architecture**
- **Low TCO with Unified Memory APU Package**

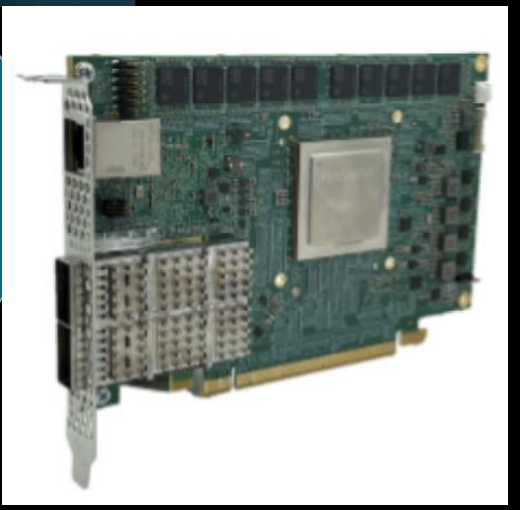


Co-designing across different classes of hardware

- Convergence of multiple flavors of technology
  - CPUs
  - GPUs
  - FPGAs
  - SW defined networks
  - Offloading model applied everywhere
  - From components to systems
  - From systems to a user abstraction
  - Co-design
    - With system integrator
    - Within the SoC
    - Across a network
    - Toolchains – software libraries and tooling



Co-design



Data Processing Unit (DPU) definition

- DPU as a new type of accelerator or integration/coupling of different accelerators?
- Definition
  - Controlling data streaming and collective processing
  - Managing network traffic, protocols and storage
  - Abstraction of a DPU as a collection of devices

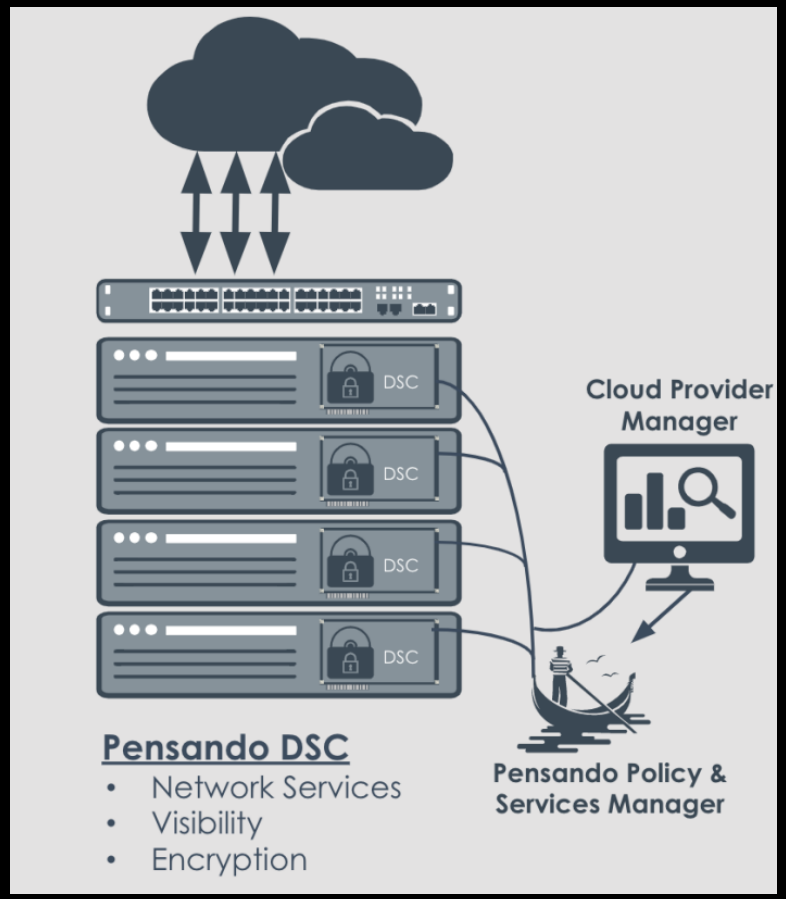
- Requirements clash and overlap: cloud vs. traditional HPC
  - Opportunities: what needs prioritizing for these different kinds of deployments
    - Cloud prioritizes data confidentiality and tenant isolation
    - HPC prioritizes low latencies
    - Both prioritize core count – more cores maps to more revenue (cloud) and performance (HPC)
  - Convergence of technologies: create technology that is transferable across HPC and cloud
    - Widespread modelling and simulation offerings
    - Deep learning frameworks
  - Variety of workloads in HPC which might be different from what is in the cloud
    - ML workloads showcasing different requirements than HPC apps
    - Requirement shift depending on the scale

Specialized use cases for DPUs

- Sparse Machine Learning communication challenges over the network
- Serialization/deserialization of data
- RDMA and shared memory abstractions for a collection of nodes
- Enabler for composability of resources
- High-performance distributed storage
- Offloading programming models (e.g., collectives or other distributed operations)
- Data-lake enabler
- Be smarter behind existing APIs
  - Evolve standards to accommodate user needs and HW limitations
  - Can we keep/simplify the same API and be smarter in their implementation?
    - Examples: MPI, RCCL/NCCL

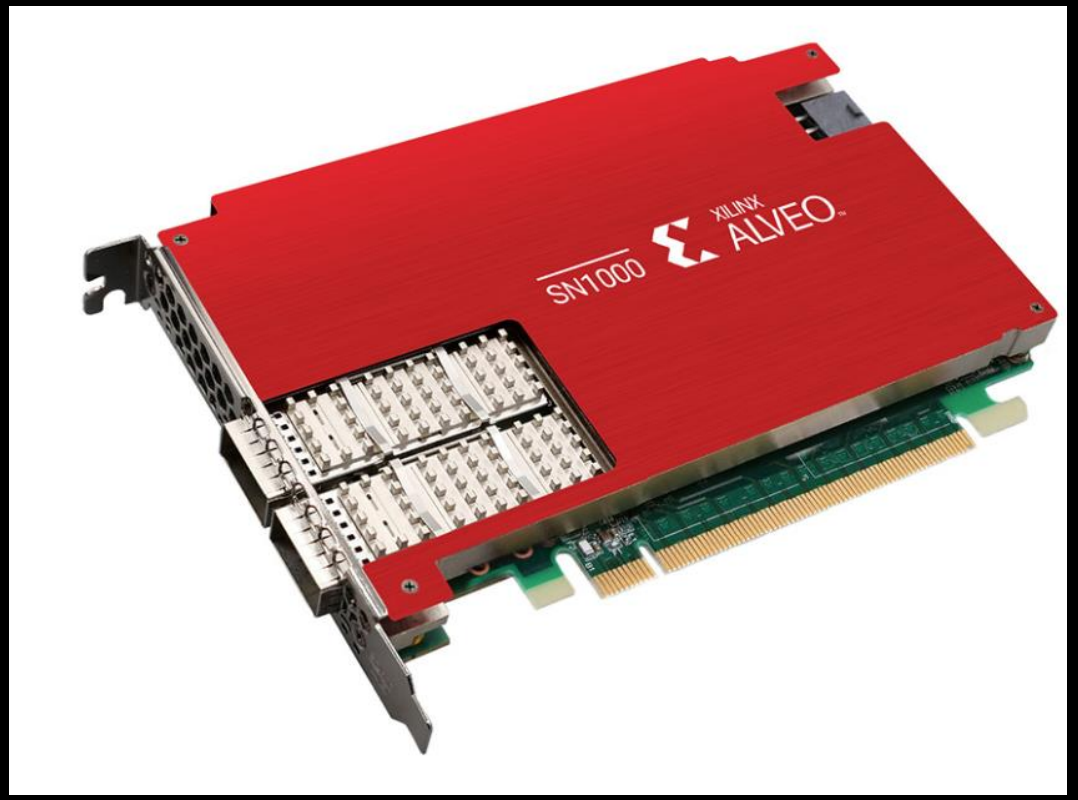
DPUs and Smart NICs – complementing capabilities

Distributed Services Card (DSC)



Highly manageable and fully programmable

AMD SmartNIC Accelerator



Complete function set tightly coupled with programmable logic (up to 400G)

Disclaimer

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. Any computer system has risks of security vulnerabilities that cannot be completely prevented or mitigated. AMD assumes no obligation to update or otherwise correct or revise this information. However, AMD reserves the right to revise this information and to make changes from time to time to the content hereof without obligation of AMD to notify any person of such revisions or changes.

THIS INFORMATION IS PROVIDED ‘AS IS.” AMD 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. AMD SPECIFICALLY DISCLAIMS ANY IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY, OR FITNESS FOR ANY PARTICULAR PURPOSE. IN NO EVENT WILL AMD BE LIABLE TO ANY PERSON FOR ANY RELIANCE, DIRECT, INDIRECT, SPECIAL, OR OTHER CONSEQUENTIAL DAMAGES ARISING FROM THE USE OF ANY INFORMATION CONTAINED HEREIN, EVEN IF AMD IS EXPRESSLY ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

Third-party content is licensed to you directly by the third party that owns the content and is not licensed to you by AMD. ALL LINKED THIRD-PARTY CONTENT IS PROVIDED “AS IS” WITHOUT A WARRANTY OF ANY KIND. USE OF SUCH THIRD-PARTY CONTENT IS DONE AT YOUR SOLE DISCRETION AND UNDER NO CIRCUMSTANCES WILL AMD BE LIABLE TO YOU FOR ANY THIRD-PARTY CONTENT. YOU ASSUME ALL RISK AND ARE SOLELY RESPONSIBLE FOR ANY DAMAGES THAT MAY ARISE FROM YOUR USE OF THIRD-PARTY CONTENT.

Questions?


AMD