ISCA 23

A Research Retrospective on the AMD Exascale Computing Journey

Gabriel H. Loh, Michael J. Schulte, Mike Ignatowski, Vignesh Adhinarayanan
Shaizeen Aga, Derrick Aguren, Varun Agrawal, Ashwin M. Aji
Johnathan Alsop, Paul T. Bauman, Bradford M. Beckmann, Majed Valad Beigi
Sergey Blagodurov, Travis Boraten, Michael Boyer, William C. Brantley
Noel Chalmers, Shaoming Chen, Kevin Cheng, Michael L. Chu
David Cownie, Nicholas Curtis, Joris Del Pino, Nam Duong
Alexandru Duţu, Yasuko Eckert, Christopher Erb, Chip Freitag
Joseph L. Greathouse, Sudhanva Gurumurthi, Anthony Gutierrez, Khaled Hamidouche
Sachin Hossamani, Wei Huang, Mahzabeen Islam, Nuwan Jayasena
John Kalamatianos, Onur Kayiran, Jagadish Kotra, Alan Lee
Daniel Lowell, Niti Madan, Abhinandan Majumdar, Nicholas Malaya
Srilatha Manne, Susumu Mashimo, Damon McDougall, Elliot Mednick
Michael Mishkin, Mark Nutter, Indrani Paul, Matthew Poremba
Brandon Potter, Kishore Punniyamurthy, Sooraj Puthoor, Steven E. Raasch
Karthik Rao, Gregory Rodgers, Marko Scrbak, Mohammad Seyedzadeh
John Slice, Vilas Sridharan, Rene van Oostrum, Eric van Tassell
Abhinav Vishnu, Samuel Wasmundt, Mark Wilkening, Noah Wolfe
Mark Wyse, Adithya Yalavarti, Dmitri Yudanov

Advanced Micro Devices, Inc.

Frontier: Exploring Exascale

The System Architecture of the First Exascale Supercomputer

Scott Atchley
Oak Ridge National Laboratory
Oak Ridge, TN, USA
atchleyes@ornl.gov

Chris Zimmer
Oak Ridge National Laboratory
Oak Ridge, TN, USA
zimmercj@ornl.gov

John R. Lange
Oak Ridge National Laboratory
Oak Ridge, TN, USA
University of Pittsburgh
Pittsburgh, PA, USA
langejr@ornl.gov

David E. Bernholdt
Oak Ridge National Laboratory
Oak Ridge, TN, USA
bernholdtde@ornl.gov

Verónica G. Melesse Vergara
Oak Ridge National Laboratory
Oak Ridge, TN, USA
vergaravg@ornl.gov

Thomas Beck
Oak Ridge National Laboratory
Oak Ridge, TN, USA
becktl@ornl.gov

Michael J. Brim
Oak Ridge National Laboratory
Oak Ridge, TN, USA
brimmj@ornl.gov

Reuben Budiardja
Oak Ridge National Laboratory
Oak Ridge, TN, USA
reubendb@ornl.gov

Sunita Chandrasekaran
University of Delaware
Newark, DE, USA
schandra@udel.edu

Markus Eisenbach
Oak Ridge National Laboratory
Oak Ridge, TN, USA
eisenbachm@ornl.gov

Thomas Evans
Oak Ridge National Laboratory
Oak Ridge, TN, USA
evanstm@ornl.gov

Matthew Ezell
Oak Ridge National Laboratory
Oak Ridge, TN, USA
ezellma@ornl.gov

Nicholas Frontiere
Argonne National Laboratory
Lemont, IL, USA
nfrontiere@anl.gov

Antigoni Georgiadou
Oak Ridge National Laboratory
Oak Ridge, TN, USA
georgiadoua@ornl.gov

Joe Glenski
Hewlett Packard Enterprise
Bloomington, MN, USA
glenski@hpe.com

Philipp Grete
Universität Hamburg
Hamburg, Germany
pgrete@hs.uni-hamburg.de

Steven Hamilton
Oak Ridge National Laboratory
Oak Ridge, TN, USA
hamiltonsp@ornl.gov

John Holmen
Oak Ridge National Laboratory
Oak Ridge, TN, USA
holmenjk@ornl.gov

Axel Huebl
Lawrence Berkeley National Laboratory
Berkeley, CA, USA
axelhuebl@lbl.gov

Daniel Jacobson
Oak Ridge National Laboratory
Oak Ridge, TN, USA
jacobsonda@ornl.gov

Wayne Joubert
Oak Ridge National Laboratory
Oak Ridge, TN, USA
joubert@ornl.gov

Kim McMahon
Hewlett Packard Enterprise
Bloomington, MN, USA
kim.mcmahon@hpe.com

Elia Merzari
Pennsylvania State University
University Park, PA, USA
ebm5351@psu.edu

Stan G Moore
Sandia National Laboratory
Albuquerque, NM, USA
stamoor@sandia.gov

Andrew Myers
Lawrence Berkeley National Laboratory
Berkeley, CA, USA
atmyers@lbl.gov

Stephen Nichols
Oak Ridge National Laboratory
Oak Ridge, TN, USA
nicholsss@ornl.gov

Sarp Oral
Oak Ridge National Laboratory
Oak Ridge, TN, USA
oralhs@ornl.gov

Thomas Papatheodore
Oak Ridge National Laboratory
Oak Ridge, TN, USA
papatheodore@ornl.gov

Danny Perez
Los Alamos National Laboratory
Los Alamos, NM, USA
danny_perez@lanl.gov

David M. Rogers
Oak Ridge National Laboratory
Oak Ridge, TN, USA
rogersdm@ornl.gov

Evan Schneider
University of Pittsburgh
Pittsburgh, PA, USA
eschneider@pitt.edu

Jean-Luc Vay
Lawrence Berkeley National Laboratory
Berkeley, CA, USA
jlvay@lbl.gov

P.K. Yeung
Georgia Institute of Technology
Atlanta, GA, USA
pk.yeung@ae.gatech.edu

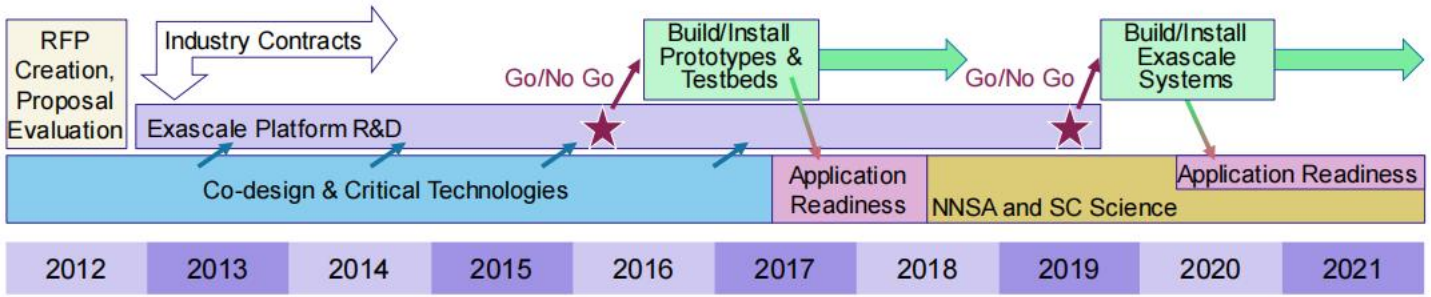
Oak Ridge National Laboratory
The world's premier research institution

- **Energy**
- **Biology**
- **Neutron science**
- **Materials**
- **Security**
- **High-performance computing**

[1]https://www.ornl.gov/
[2]https://en.wikipedia.org/wiki/Oak_Ridge_National_Laboratory

2011 the United States Department of Energy (DOE)

- **Request for Information (RFI) & Request for Purpose (RFP)**
- **Codesign : Technology providers collaborate closely with scientists and technologists from DOE**
- **Application Readiness : To ensure the applications are compatible and functional**
- **Power-performance Efficency needed to be a top priority**



(a)

Exascale System	Goal
Delivery Date	2019-2020
Performance	1000 PF LINPACK 300 PF on to-be-specified applications
Power Consumption	20 MW
Memory Capacity (incl. NVRAM)	128 PB
Node Memory Bandwidth	4 TB/s
Node Interconnect Bandwidth	400 GB/s

(b)

Figure 1. (a) Exascale timeline and (b) system objectives from the 2011 U.S. DOE exascale research and development Request for Information.

2023 Retrospective view

To innovate and accelerate the necessary exascale technologies, a series of programs between the DOE and technology companies covering processors, memory, storage, networking, and software are funded.

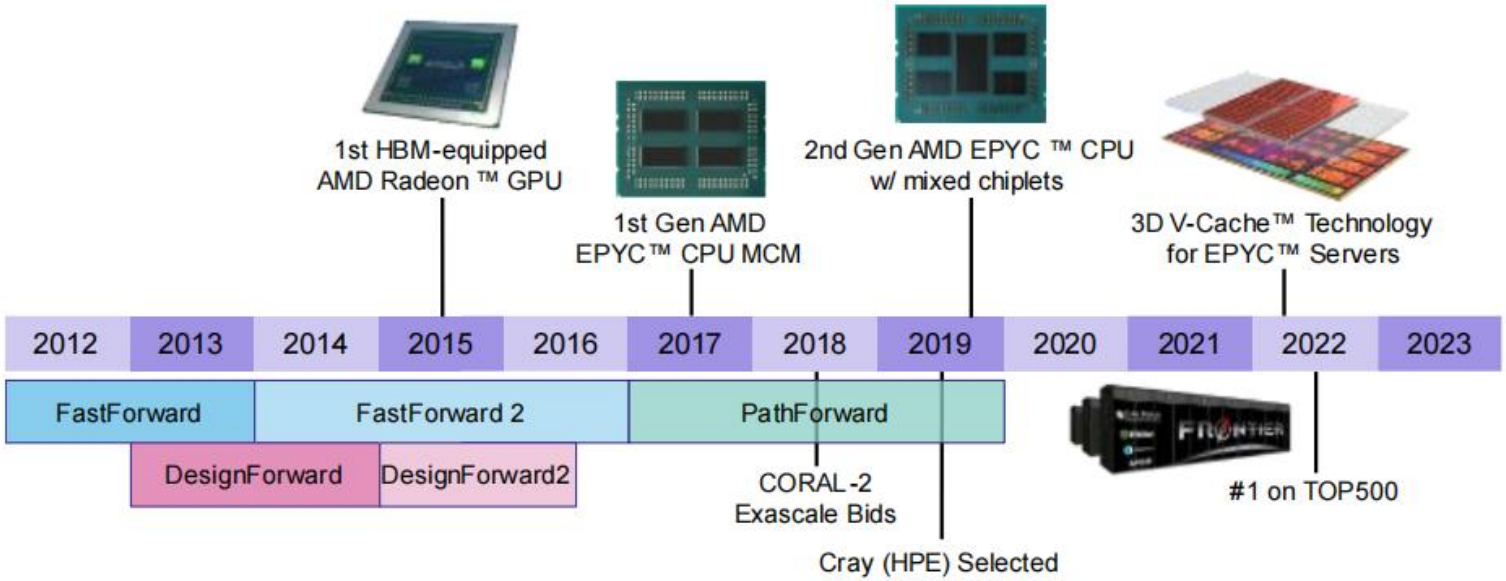
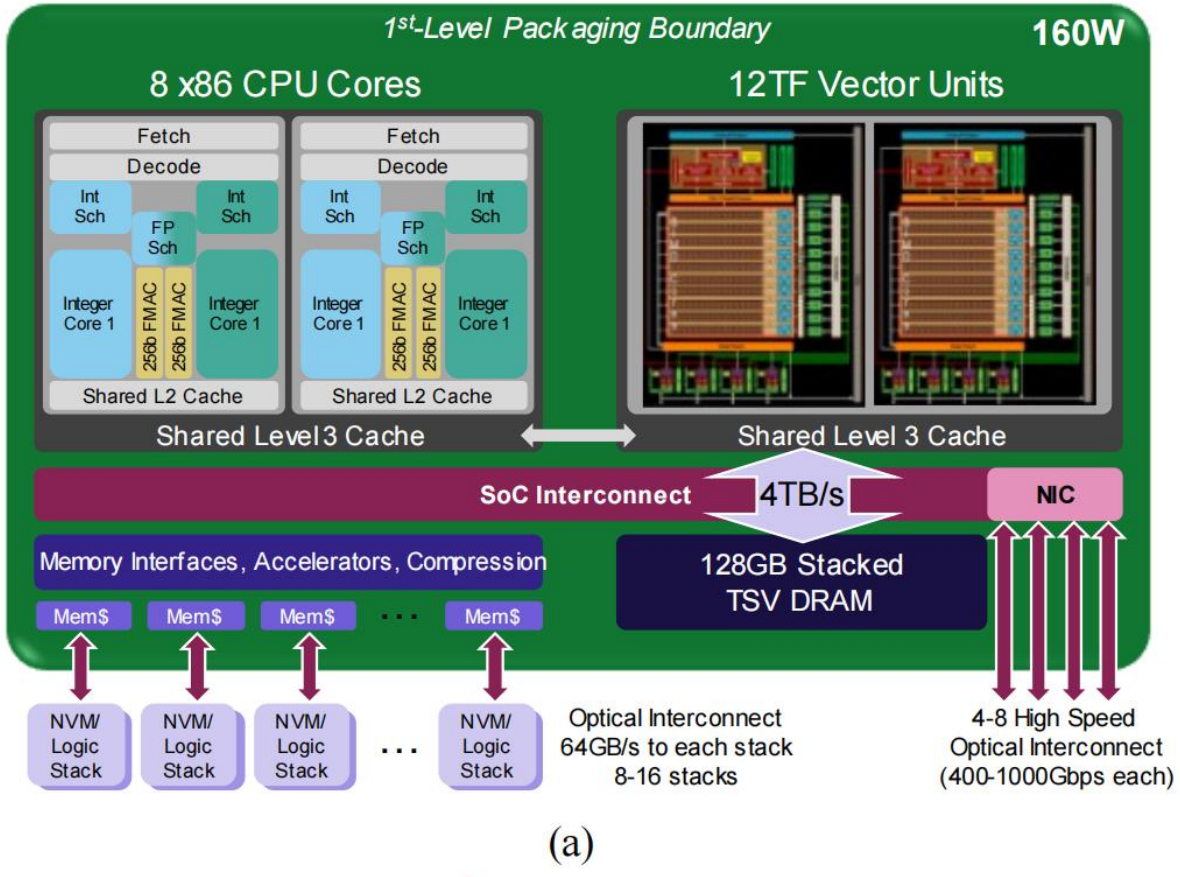


Figure 2. Timeline illustrating U.S. DOE exascale R&D programs and milestones (bottom) and key AMD technology introductions (top).

2012 Exascale Heterogeneous Processor (EHP) V1

Research community was highly concerned with the (fear to be) imminent end of DRAM scaling.

To maximize the compute and minimize the cost of data movement, 3D stacking is adopted.



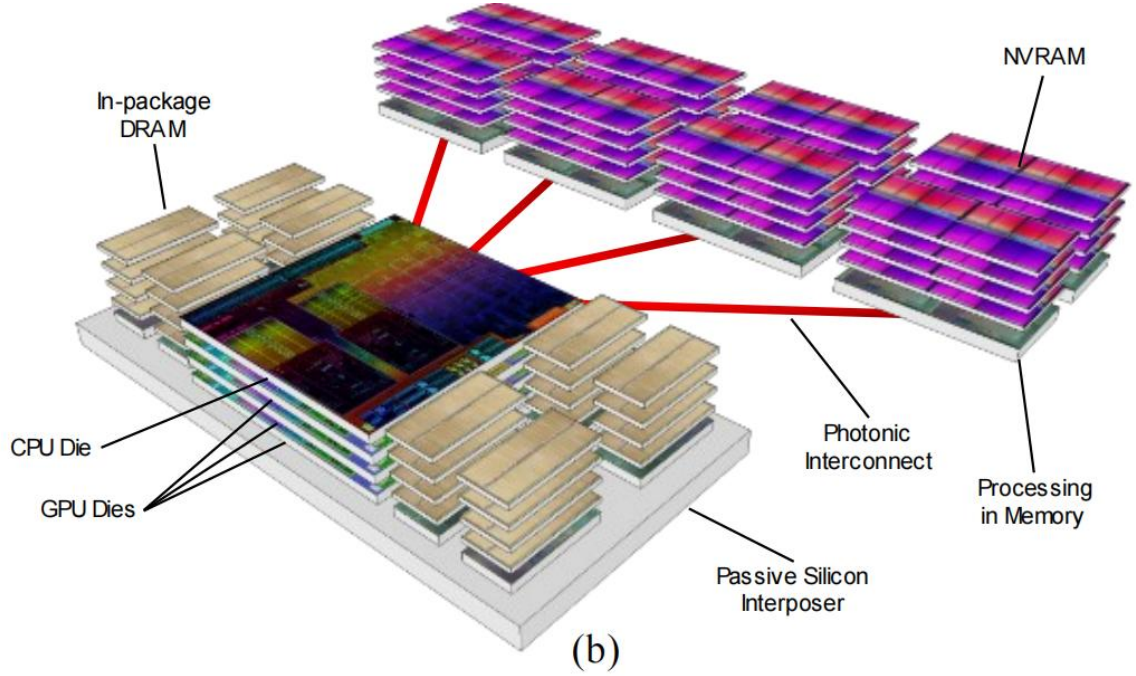
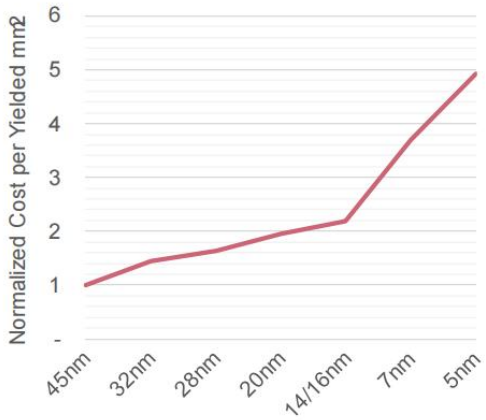


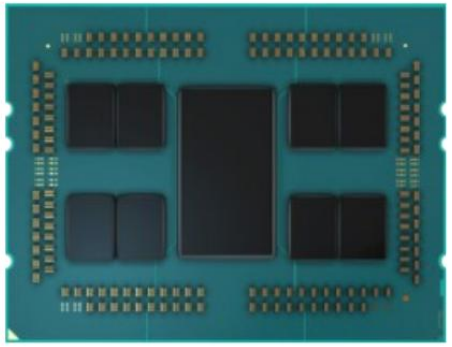
Figure 3. (a) Block diagram of the Exascale Heterogeneous Processor (EHP) concept from the original FastForward program circa 2012, (b) illustrative packaging view of the EHP.

2014 EHP V2

To reduce silicon cost, AMD adopts chiplet technology to reuse silicon components in multiple product configurations.



(a)



(b)

Figure 4. Silicon cost trends over time and (b) an AMD EPYC™ processor utilizing chiplets.

AMD packaging engineers also raised concerns about the asymmetry of the overall package (all CPU on one side, all GPU on the other).

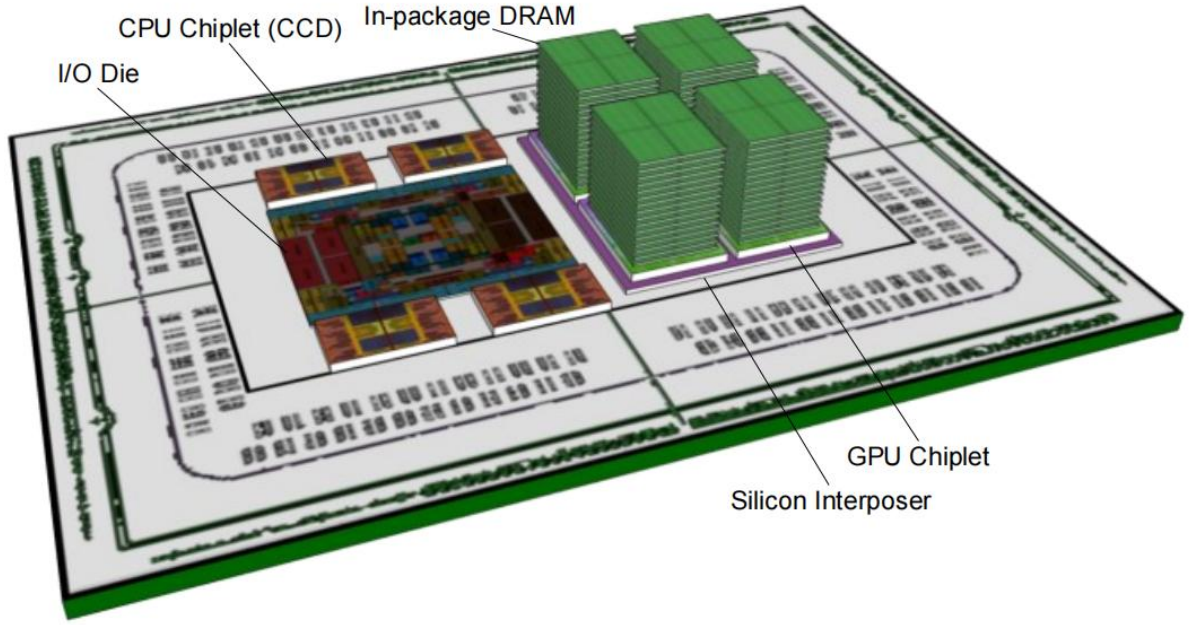


Figure 5. Refinement of the EHP (v2), circa 2014.

2016 EHP V3

- The power density of the GPU regions still present thermal challenges
- While technically feasible, the “triple stack” of DRAM on GPU on active interposer also significantly increases the manufacturing complexity

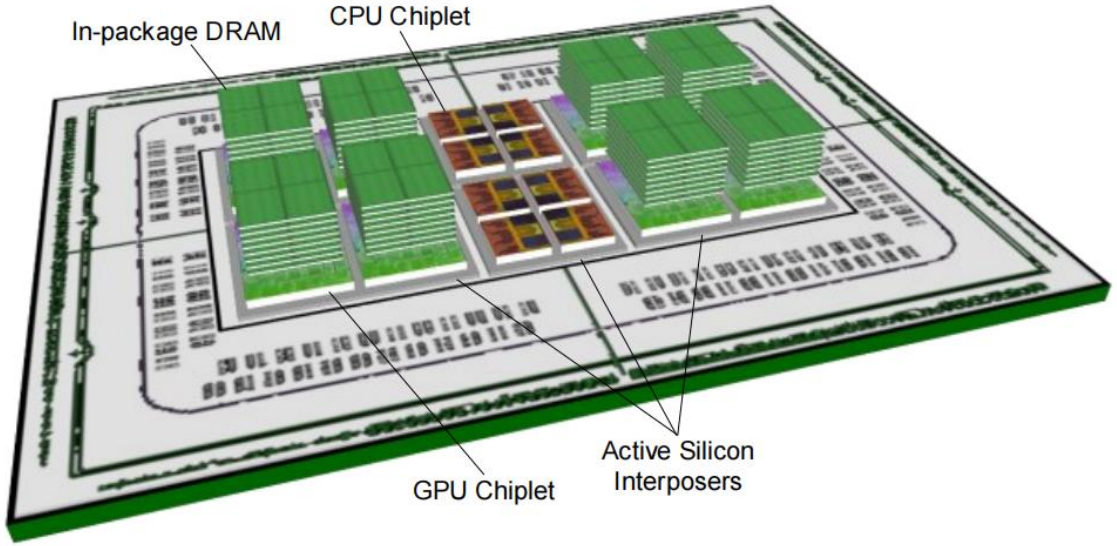


Figure 6. Refinement of the EHP (v3), circa 2016.

2018 EHP V4

The higher bandwidth required to support data movement and work distribution among the GPU compute units would be far less efficient to route among the larger number of chiplets

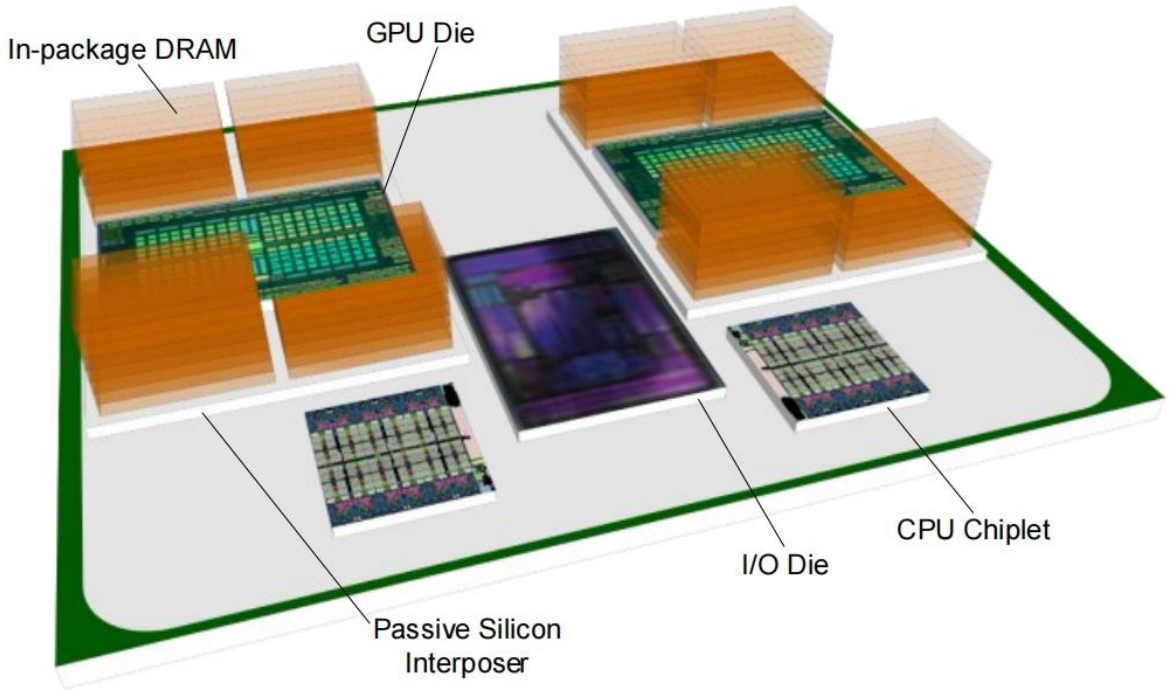


Figure 7. Refinement of the EHP (v4), circa 2018.

APU VS Discrete Node Architecture

APU:Combine a general-purpose CPU and a GPU on a single die

- **Enable faster data transfer and communication between the CPU and GPU**
- **Reduce the power consumption since both share the same die and memory**

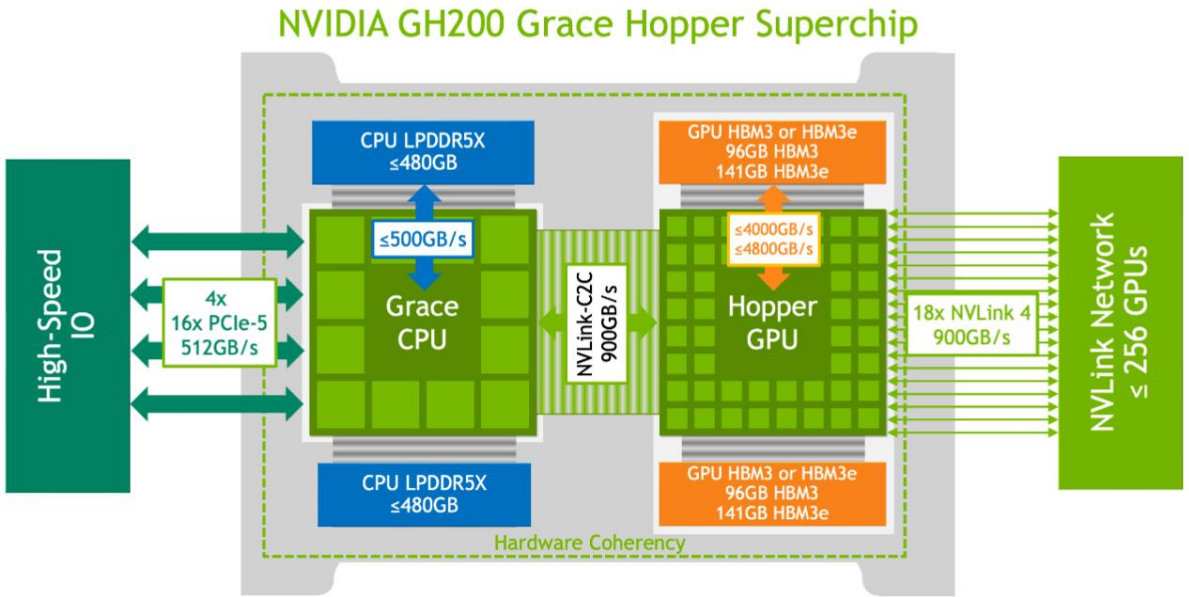


Figure 1. NVIDIA GH200 Grace Hopper Superchip Logical Overview

More scalability and customizable:

- **Customize platforms to provide different CPU-to-GPU ratios as well as to interoperate with other components**

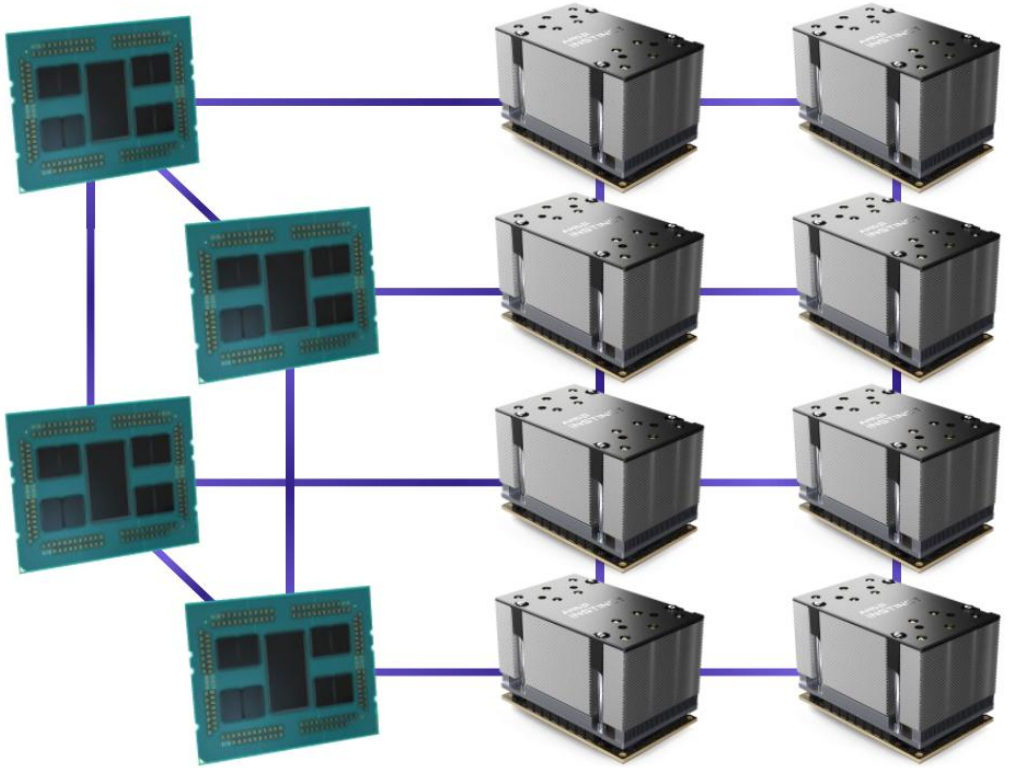


Figure 8. Discrete Node Architecture consisting of inter-connected CPUs (left) and accelerators (right).

Overview of one Frontier Computer Node

9408 compute nodes housed in 74 cabinets

64-core EPYC™ 7A53 CPU(3rd EPYC)

MI250X (CDNA2)

Infinity Fabric link

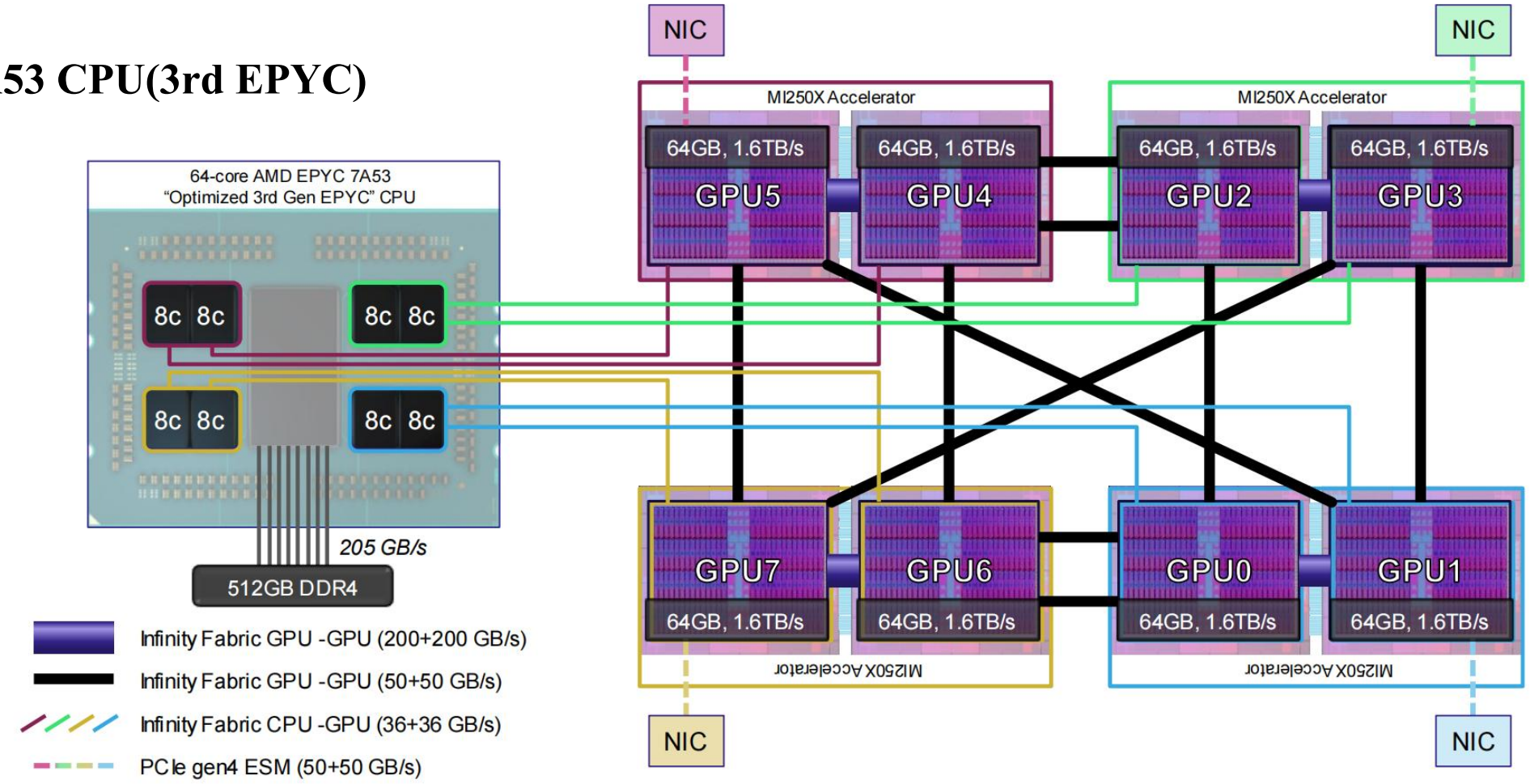


Figure 10. Block diagram of one Frontier Compute Node with peak theoretical memory and interconnect speeds. The "X+X GB/s" notation indicates X GB/s of bandwidth each for send and receive.

EPYC 7003 System on Chip (SoC)

CCD:
Core complex die

GMI:
Global memory interface

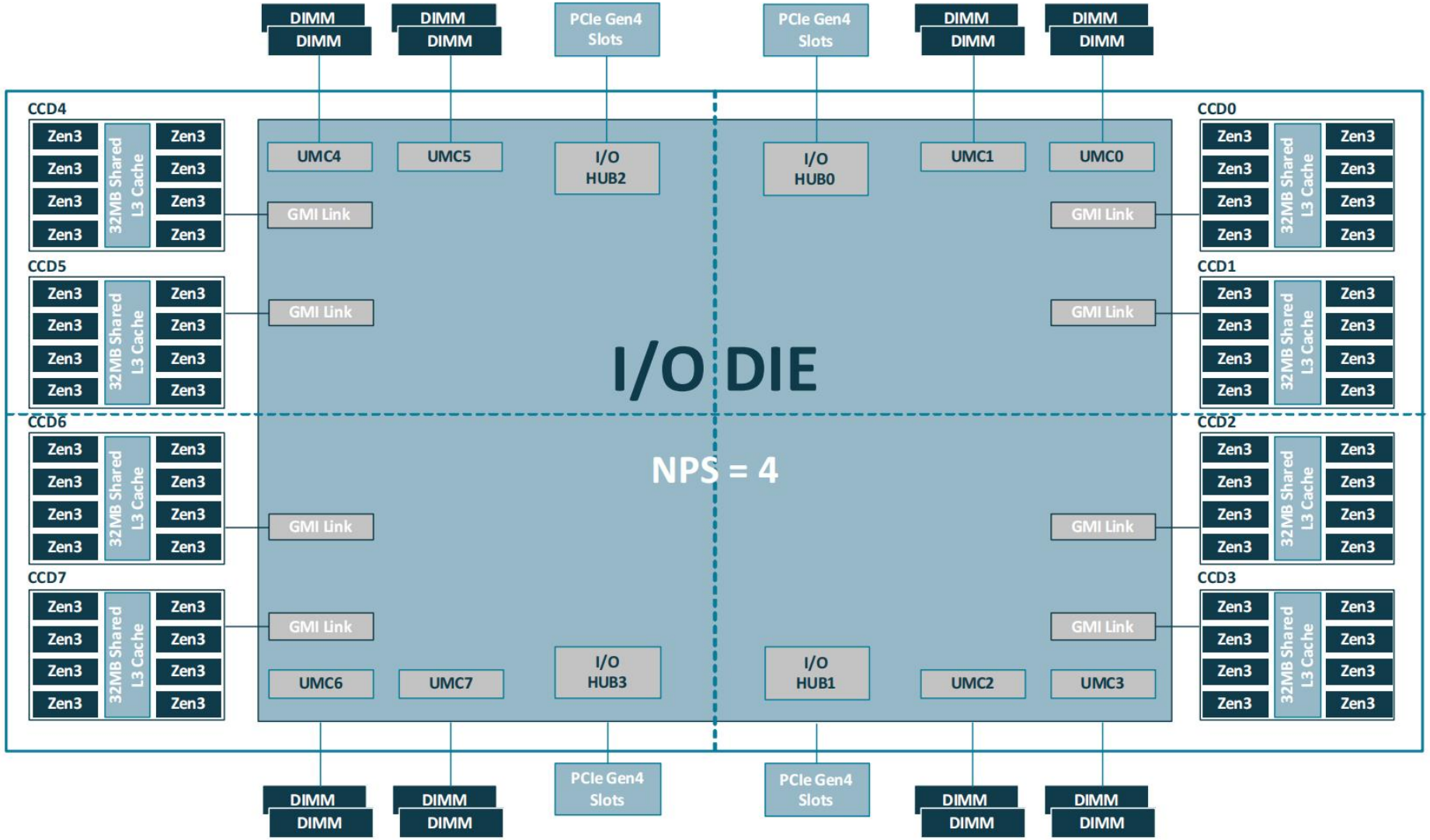


Figure 1-5: EPYC 7003 System on Chip (SoC): 8 CCDs and central IOD

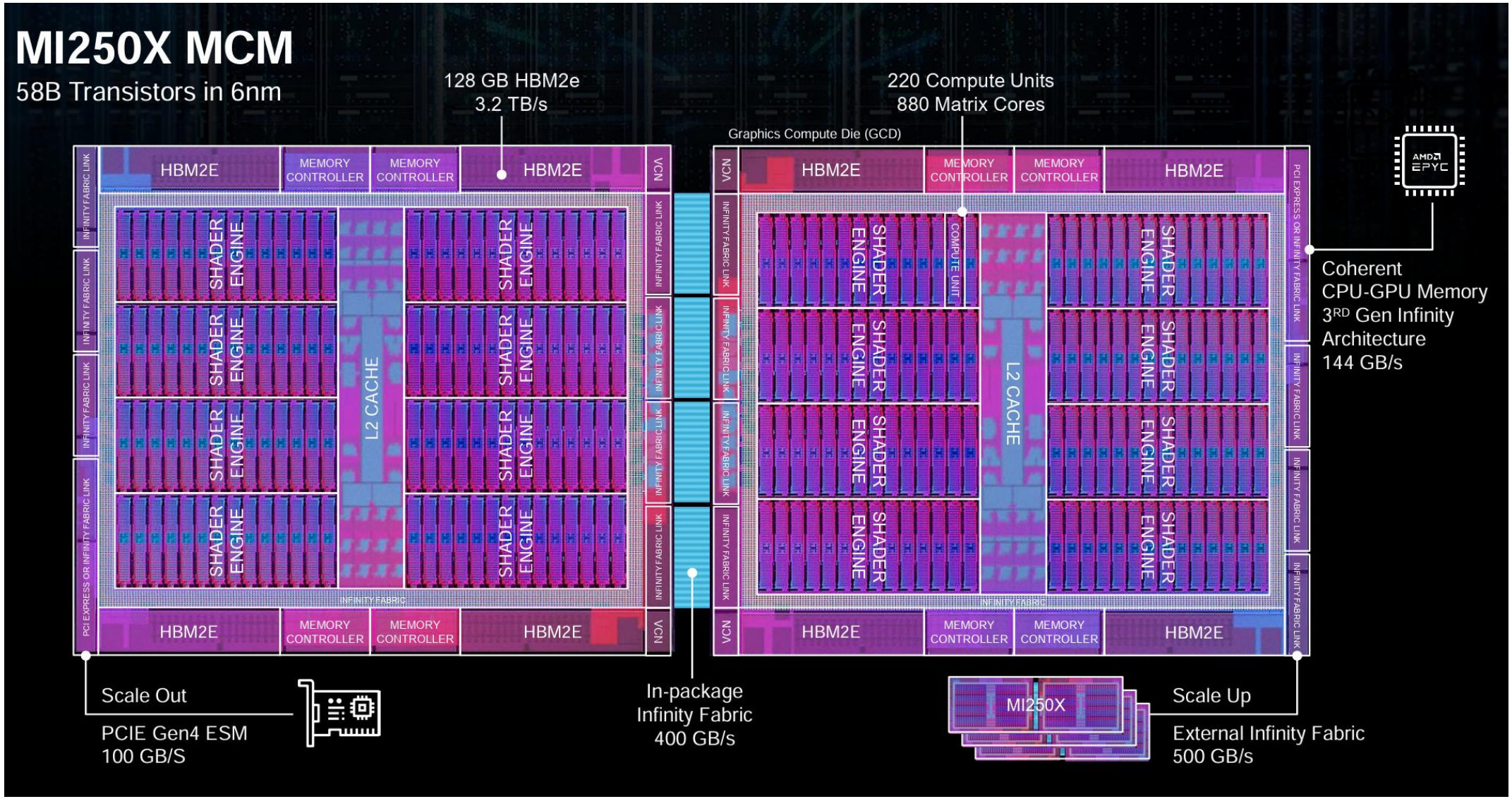
Evaluation of AMD EPYC 7A53 “Trento” CPU.

- **Trento is able to achieve up to 180 GB/s using non-temporal loads and stores in NPS-4 mode. When operating in NPS-1, that rate drops to ~ 125 GB/s.**
- **Table 3 illustrates how caching can negatively affect bandwidth when data are not expected to fit into cache.**

Function	Temporal (MB/s)	Non-Temporal (MB/s)
Copy	176780.4	179130.5
Scale	107262.2	172396.2
Add	125567.1	178356.8
Triad	120702.1	178277.0

Table 3: CPU STREAM bandwidth results using temporal and non-temporal stores.

MI250X


MI250X MCM
58B Transistors in 6nm
128 GB HBM2e
3.2 TB/s
220 Compute Units
880 Matrix Cores
Graphics Compute Die (GCD)
HBM2E
MEMORY CONTROLLER
SHADER ENGINE
COMPUTE UNIT
L2 CACHE
INFINITY FABRIC
INFINITY FABRIC LINK
VCN
PCI EXPRESS OR INFINITY FABRIC LINK
AMD EPYC
Coherent
CPU-GPU Memory
3RD Gen Infinity
Architecture
144 GB/s
Scale Out
PCIE Gen4 ESM
100 GB/S
In-package
Infinity Fabric
400 GB/s
MI250X
Scale Up
External Infinity Fabric
500 GB/s

Comparison between AMD and NVIDIA GPUs MI300X is released TODAY !!!

Manufacturer	AMD			NVIDIA	
Product	MI100	MI250X	MI300X	A100	H100
Release Time	2020.11	2021.11	2023.12	2020.5/11	2022.3
FP64	11.5 TF	95.7 TF	163.4 TF	19.5 TF	66.9 TF
TF32	N/A	N/A	490.3 TF	156 TF	494.7 TF
BF16	92.3 TF	383 TF	1307.4 TF	312 TF	989.4 TF
FP16	184.6 TF	383 TF	1307.4 TF	312 TF	989.4 TF
FP8	N/A	N/A	2614.9 TF	N/A	1978.9 TF
INT8	184.6 TF	383 TF	2614.9 TF	624 TF	1978.9 TF
Memory Size	32 GB	128 GB	192 GB	40/80 GB	80 GB
Memory Bandwidth	1.2 TB/s	3.2 TB/s	5.3 TB/s	1.5/2.0 TB/s	3 TB/s

TF:TFLOPS stands for matrix/tensore core computation(dense) throughput

MI300 Series

Block diagram of the AMD Instinct MI300A and MI300X

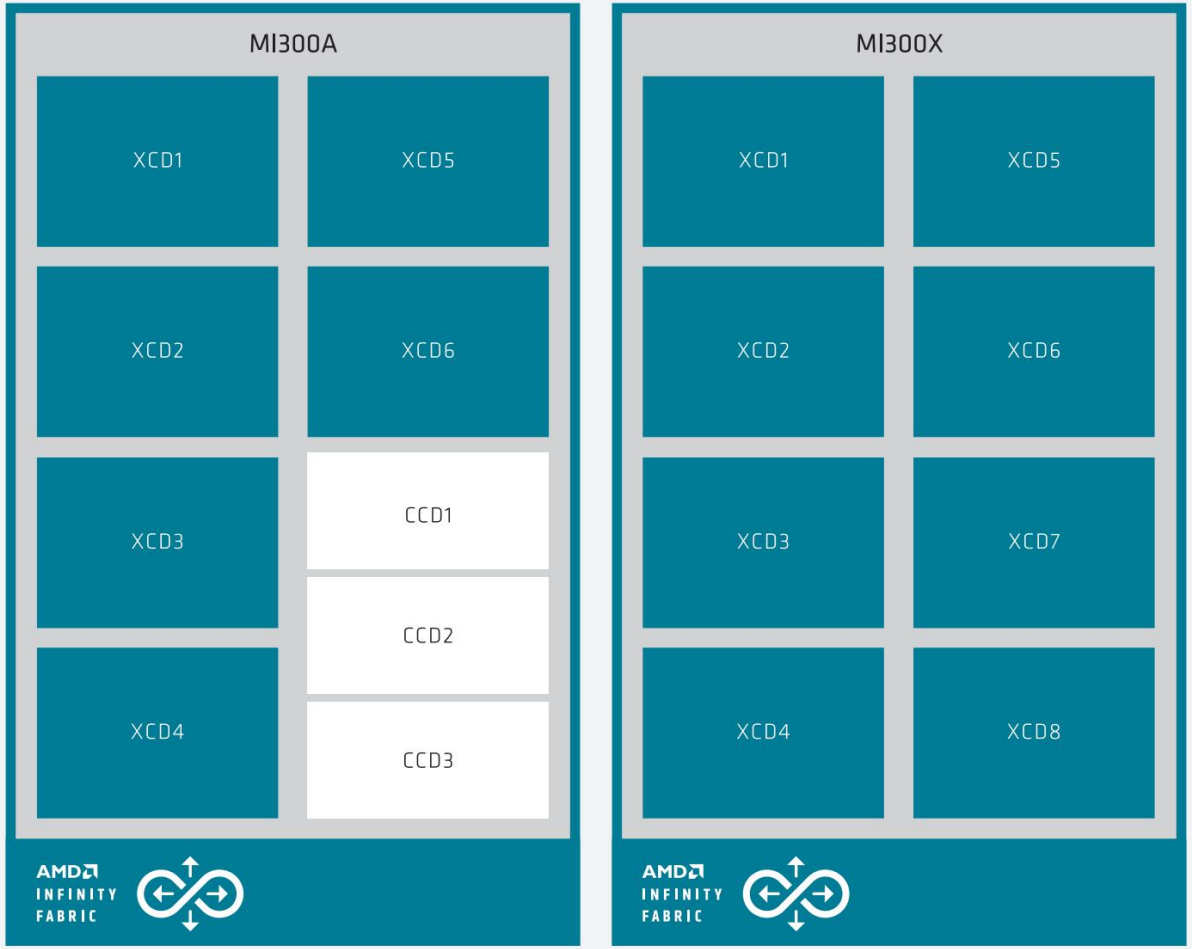
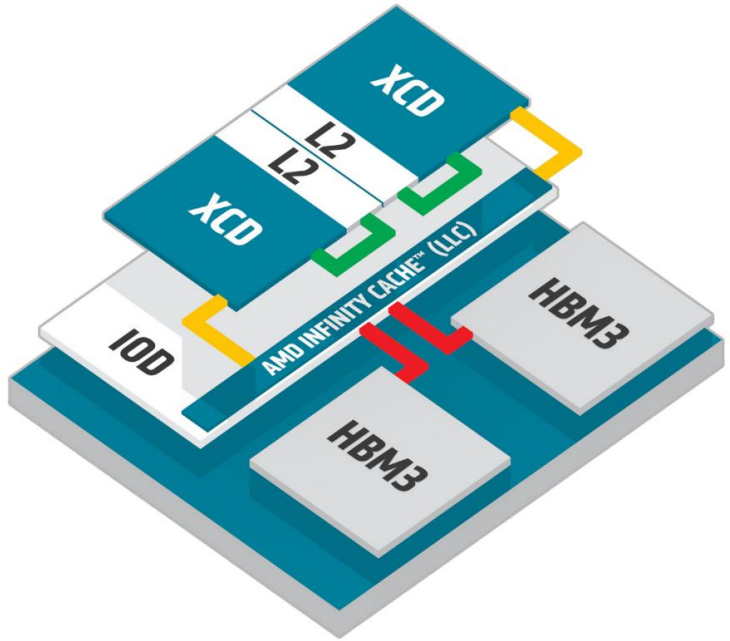


Figure 6. AMD CDNA™ 3 architecture memory architecture diagram



LEGEND

- Green line is **L2 to XCD** 51.6 TB/s (Aggregate)
- Yellow line is **LLC to XCD** 17.2 TB/s (Aggregate)
- Red line is **L2 to XCD Bandwidth** 5.3 TB/s (Aggregate)

THANKS & QA