



Hewlett Packard
Enterprise

HLRS Hunter – Architecture

Christian Simmendinger (HPE)

Feb 2024

Overview

Overview

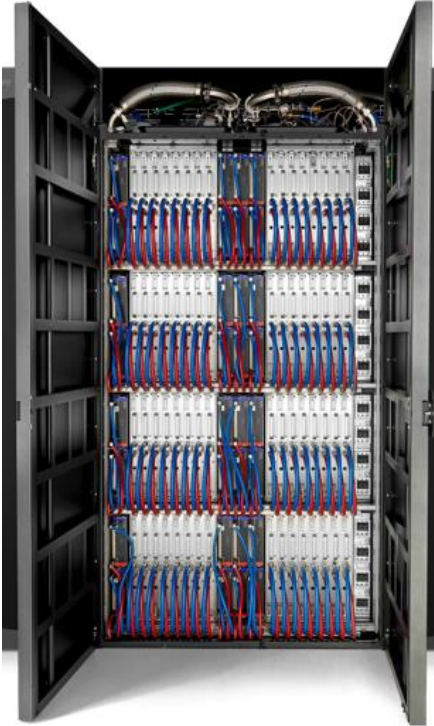
HPE Cray
EX 255a

HPE Cray
Slingshot

AMD
MI300A



Hunter – Stepping Stone System



- Hunter will be based on the HPE Cray EX4000 platform
 - HPE Cray EX255a (El Capitan blade architecture, MI-300A)
 - HPE Cray Slingshot Interconnect
- Work File Systems
 - HPE Cray ClusterStor E2000 Lustre Appliance
 - FS1: 13PB
 - FS2: 13PB
- Home File System: 540TB



HPE Cray EX 255a

Overview

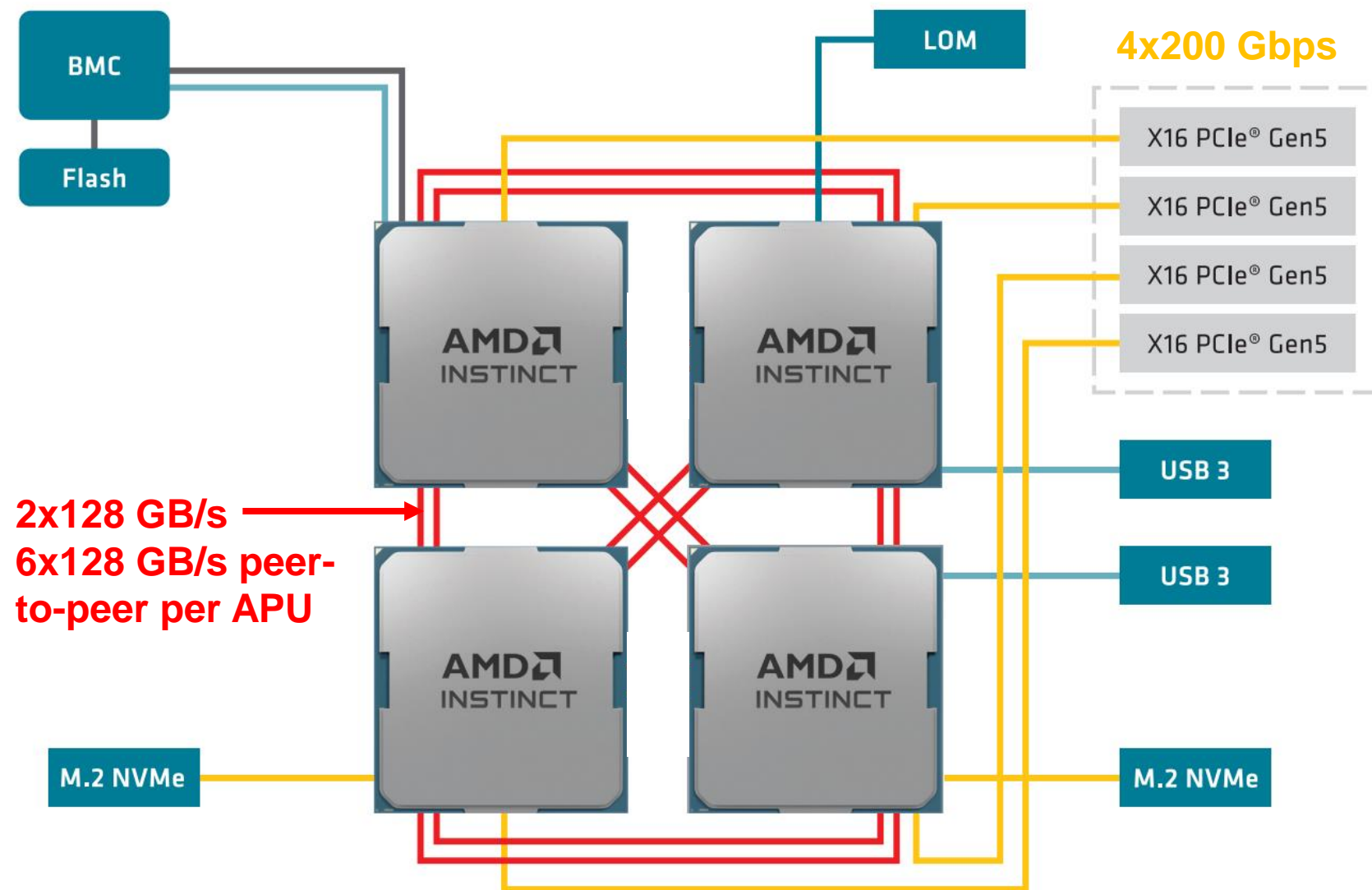
HPE Cray
EX 255a

HPE Cray
Slingshot

AMD
MI300A



HPE Cray Supercomputing EX255A Node Architecture



HPE Cray Supercomputing EX255a Specs

HPE Cray Supercomputing EX255a		Hawk Apollo 9000
Form Factor	1U blade for EX4000 and EX2500	1U blade Apollo 9000
Processors	AMD MI300A APU	EPYC 7742 CPU
Compute Blade	Two 4-socket MI300A APU nodes	Four 2-socket AMD Rome nodes
Core Count	24 CPU Cores and 228 Compute Cores per APU 96 CPU Cores and 912 Compute Cores per node	64 CPU Cores per CPU, 128 CPU Cores per node
Memory / blade	128GB HBM3 per MI300A APU; 512GB HBM3 per node	128GB DDR4 per socket, 256GB per Node
Memory Technology	HBM3 ~5,3 TB/s per MI300A APU	DDR4 ~205 GB/s per CPU socket
Intra Node	6x 128GB/s per APU, 2x 128GB/s Peer-to-Peer	96 GB/s Peer-to-Peer
Local Storage	0 or 1 local NVMe M.2 SSD per node	-
Fabric Option	HPE Slingshot 11 (4 injection ports per node, 4x 200 Gbps)	Infiniband HDR200 Socket-direct (1 injection port per node, 1x 200 Gbps)



HPE Cray Slingshot

Overview

HPE Cray
EX 255a

HPE Cray
Slingshot

AMD
MI300A



HPE SLINGSHOT

Dragonfly Network Architecture

- Packet-by-packet routing of unordered traffic (e.g. MPI/Lustre bulk data) optimally routed at each hop
- Adaptive routing of ordered traffic (e.g. Ethernet)
Each new flow can take an optimal new path

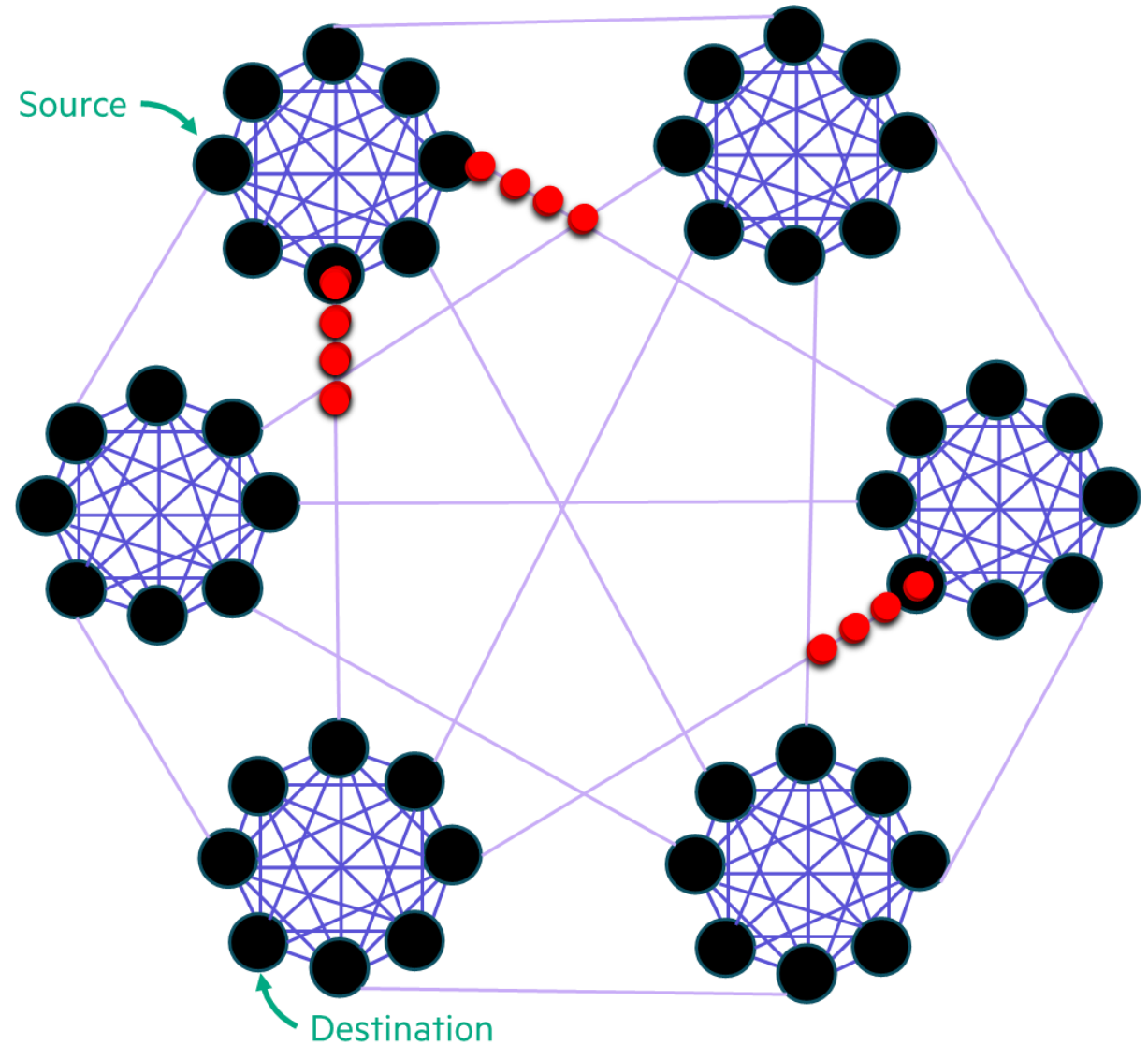
Rosetta Switch

64 port switch, 200 Gb/s

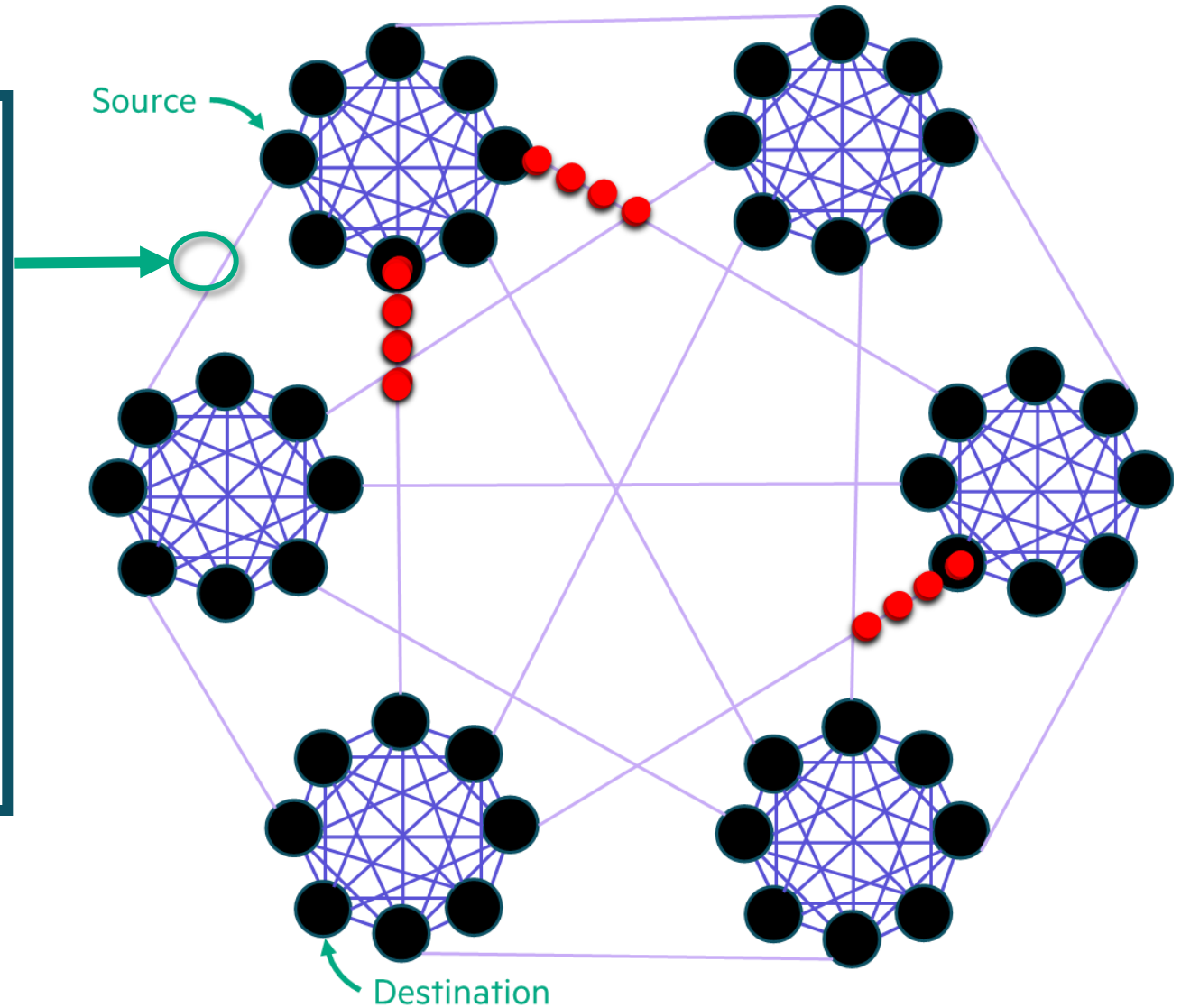
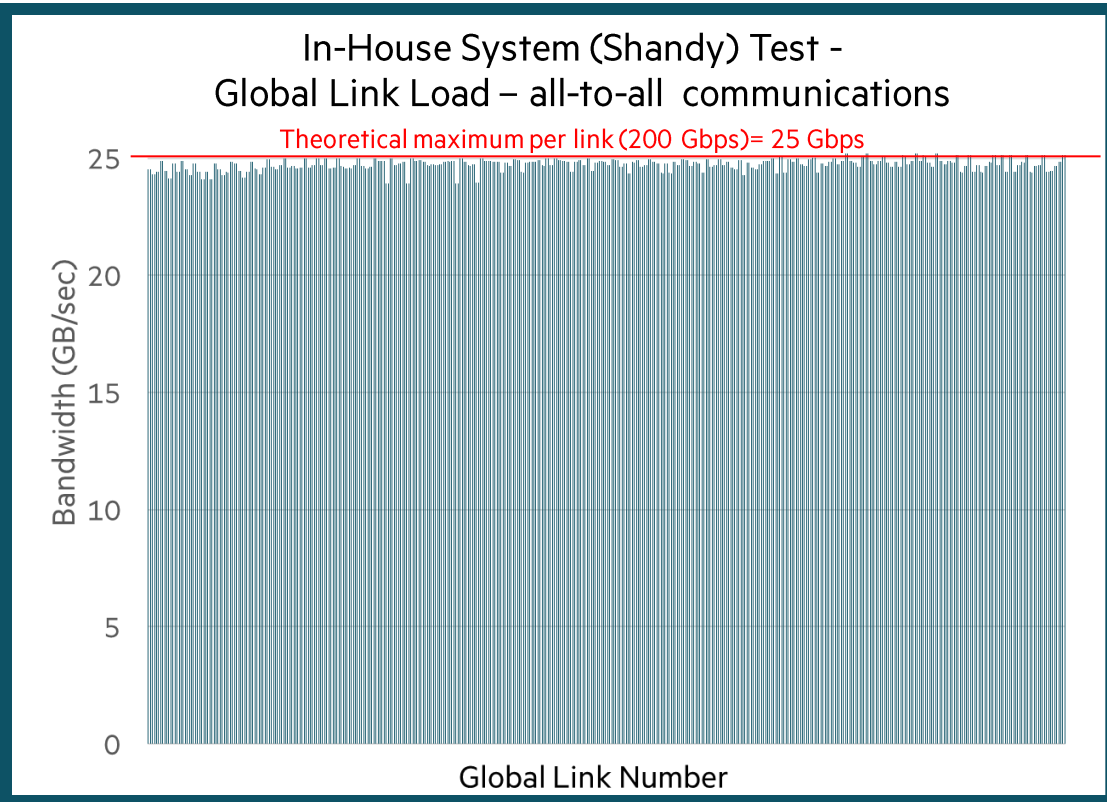
- Advanced adaptive routing
- Congestion control, QoS

Cassini NIC

- MPI hardware tag matching
- MPI progress engine
- Hardware support for one-sided operations
- Hardware support for collective operations
- 200 Gb/s



Achieving near maximal Bandwidth with fine grained adaptive routing



“Shandy” in-house system

- 8 groups, 1024 nodes
- Dual CX5 injection per node
- 25 TB/s aggregate injection BW
- 50% global bandwidth taper
- 12.5 TB/s aggregate global BW



AMD MI300A

Overview

HPE Cray
EX 255a

HPE
Slingshot

AMD
MI300A



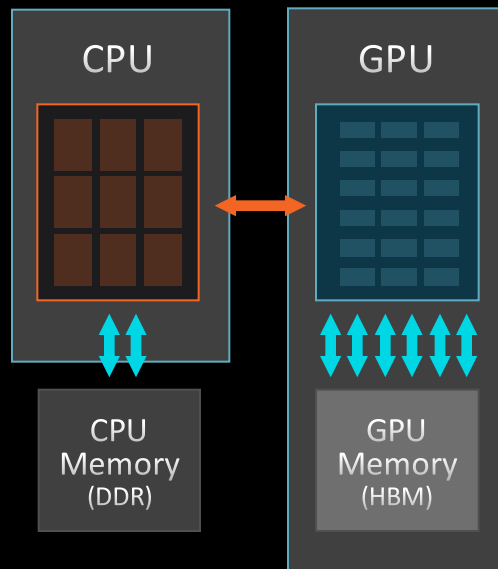
AMD MI300A

UNIFIED MEMORY APU ARCHITECTURE BENEFITS

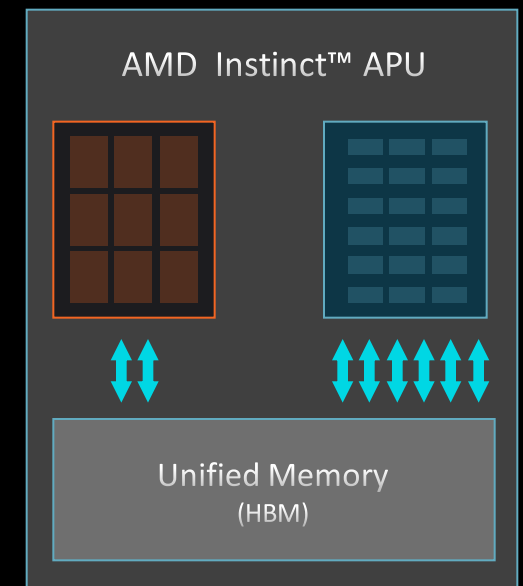
AMD CDNA™ 2 Coherent Memory Architecture



AMD CDNA™ 3 Unified Memory APU Architecture



- **Eliminate Redundant Memory Copies**
- **No programming distinction between host and device memory spaces**
- **High performance, fine-grained sharing between CPU and GPU processing elements**
- **Single process can address all memory, compute elements on a socket**



AMD Instinct™ MI300A

I/O Die (IOD)

256MB AMD Infinity Cache™
4 x16 4th Gen Infinity Fabric™ Links
4 x16 PCIe® 5

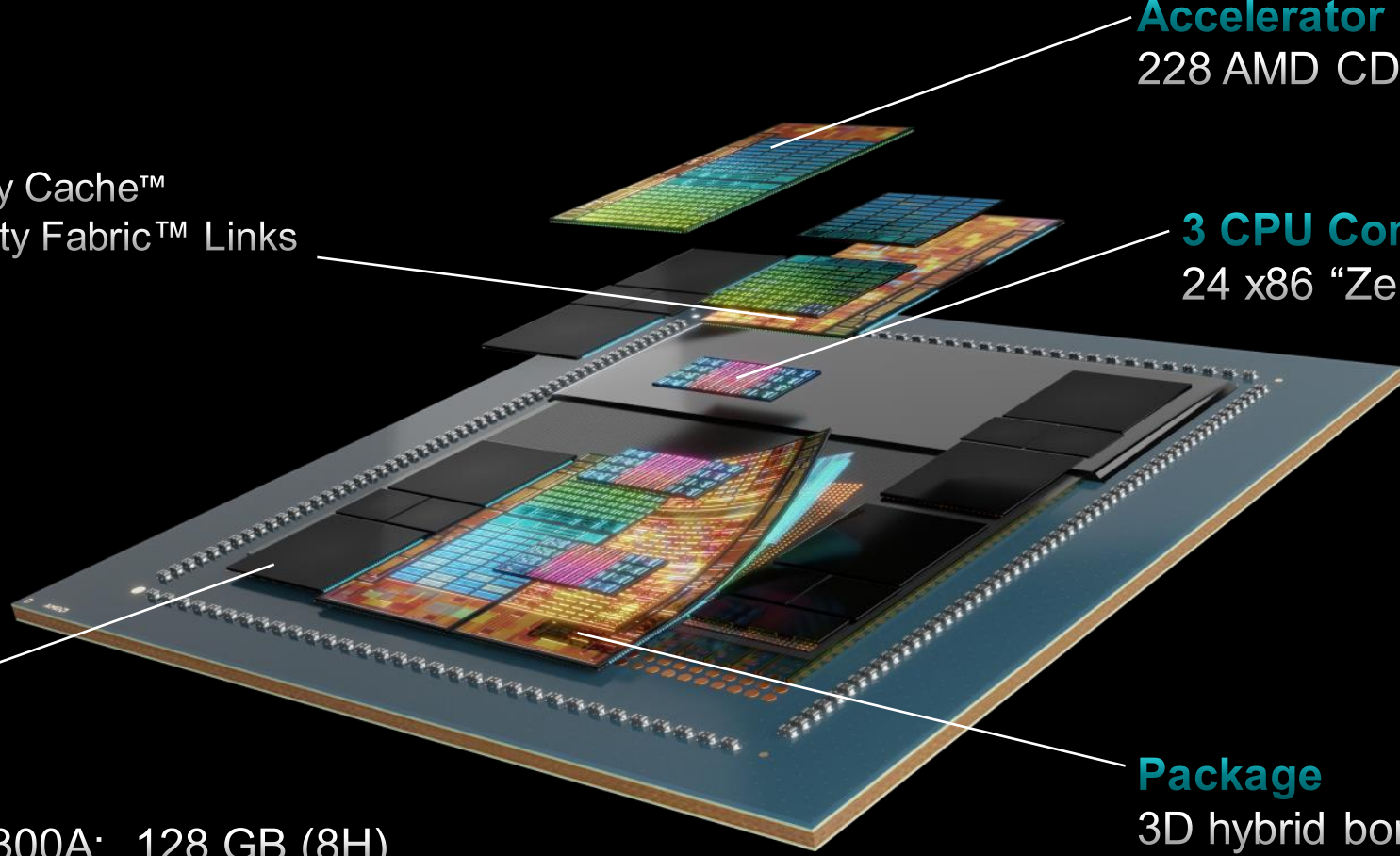
Accelerator Complex Die (XCD)
228 AMD CDNA™ 3 Compute Units

3 CPU Complex Die (CCD)
24 x86 “Zen 4” Cores

HBM3

8 physical stacks
AMD Instinct™ MI300A: 128 GB (8H)
~5.3 TB/s Bandwidth

Package
3D hybrid bonded
2.5D silicon interposer



Thank you!

christian.simmendinger@hpe.com

