

# LUMI

A white wolf is the central focus, standing in a futuristic, blue-toned digital environment. The background is filled with vertical lines, glowing particles, and a grid-like pattern, creating a sense of depth and technology. The wolf is looking slightly to the right of the viewer.

**LUMI Architecture**

**Kurt Lust**  
LUMI User Support Team (LUST)  
University of Antwerp

October 2023

# Why do I need to know this?

- **Q:** “I only want to run some programs, why do I need to know about system architecture?”
- **A:** An HPC system is not a big smartphone or big PC
  - Scaling is not for free
  - Instead in many cases an application needs to be mapped properly on the hardware to run efficiently
  - Which requires an understanding of
    - The hardware: This section
    - The middleware between the application and the hardware: Several sessions during this course
    - The application: Very domain-specific and hence not the topic of a general course
  - 20% unused performance = 32M EURO (or 50M EURO including expected exploitation costs)

# LUMI is...

LUMI

... a (pre-)exascale supercomputer ...

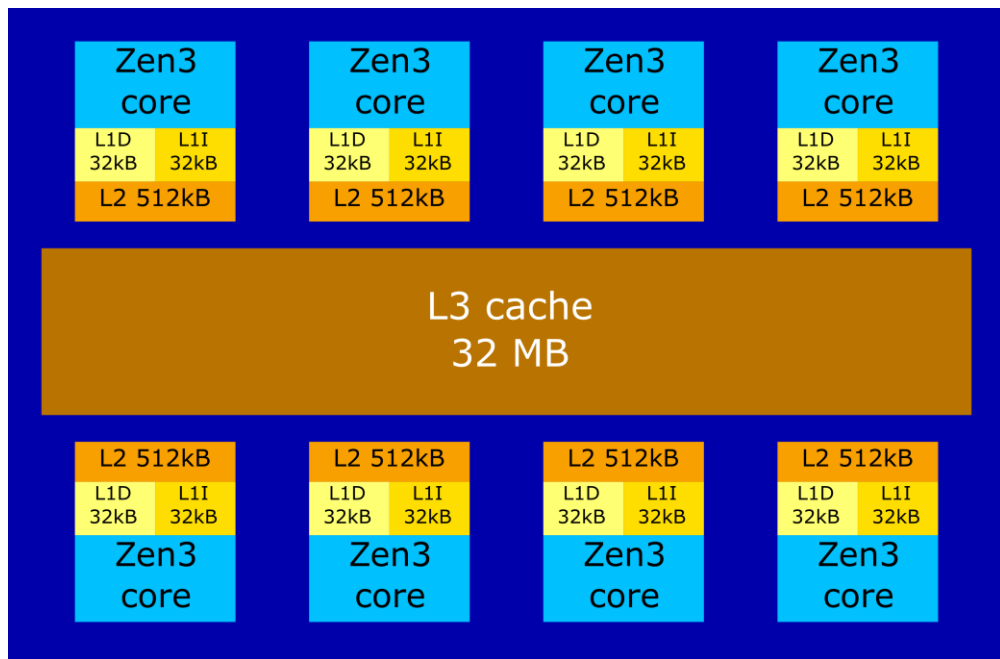
- ... and not a superfast PC ...
- ... and not a compute cloud infrastructure.
  
- Each of these infrastructures have their own trade-offs
- A supercomputer is
  - Build for scalable parallel applications
  - A shared infrastructure with lightweight management
  - Principle: Reduce hardware costs by clever software
  - Build for streaming data through the machine at all levels, not for random access to small bits of data

# LUMI spec sheet: A modular system

L U M I

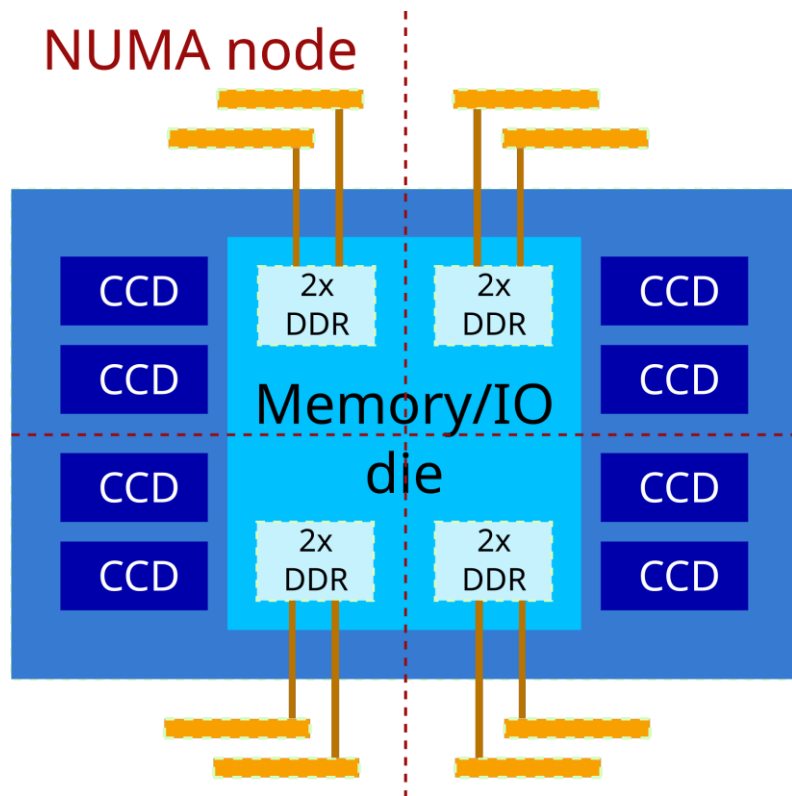
- LUMI-G: 2928 nodes with 1 AMD EPYC 7A53 CPU and 4 AMD MI250x accelerators (512 GB + 4x128 GB RAM)
- LUMI-C: 1536 nodes with 2 64-core AMD EPYC 7763 CPUs (1376x 256GB, 128x 512 GB and 32x 1TB)
- Nodes for interactive data analytics: 8 4TB CPU nodes and 8 nodes with 8 GPUs each for visualisation
- LUMI-F: 8 PB Lustre flash-based file storage (1740 GB/s)
- LUMI-P: 4 20 PB hard disk based Lustre file systems (4x 240 GB/s)
- Object based file system
- Currently 4 user access nodes with two AMD Rome CPUs each
- All linked together with a HPE Cray Slingshot 11 interconnect
- Coming up: Open OnDemand environment

# The AMD EPYC 7xx3 (Milan/Zen3) CPU L U M I



- Building block: a Core Complex Die (CCD)
- 8 cores
  - Each core has private L1 and L2 caches
  - L3 cache shared
- Instruction set equivalent to Intel Broadwell generation
  - AVX2+FMA, no AVX-512

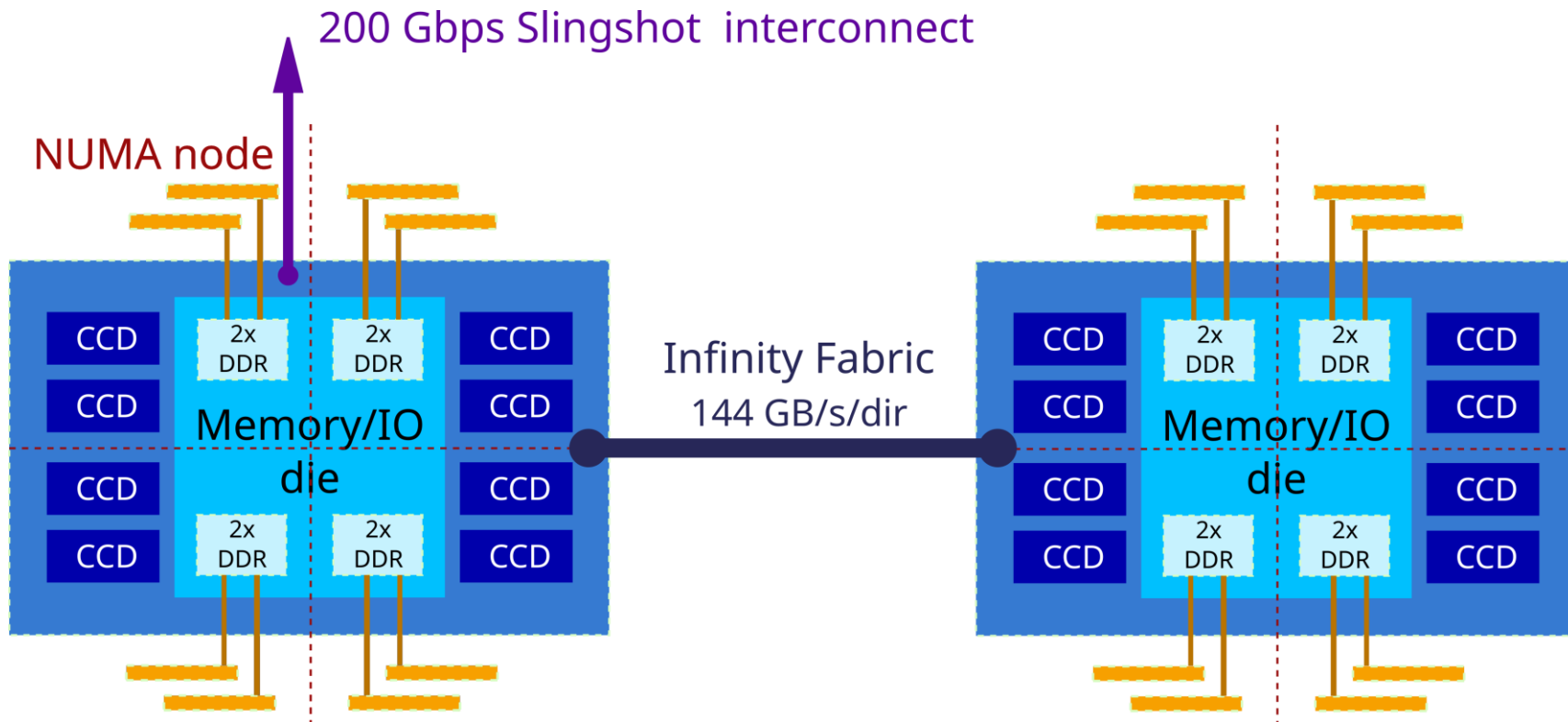
# The AMD EPYC 7xx3 (Milan/Zen3) CPU LUMI



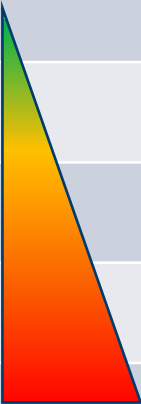


- 8 CCDs or 8 L3 cache regions
- Memory/IO die logically split into 4 NUMA domains with
  - 2 CCDs (16 cores)
  - 2 DDR4 controllers
- Memory/IO die also provides the PCIe links and intersocket links

# LUMI-C node

LUMI



# Strong hierarchy

hierarchy layer		per	sharing	distance	data transfer delay	data transfer bandwidth
1	2 threads	core	L1I, L1D, L2, execution units, rename registers			
2	8 cores	CCD	L3 Link to I/O die			
3	2 CCDs	NUMA node	DRAM channels (and PCIe lanes)			
4	4 NUMA nodes	socket	inter-socket link			
5	2 sockets	node	inter-node link			



# Delays in numbers

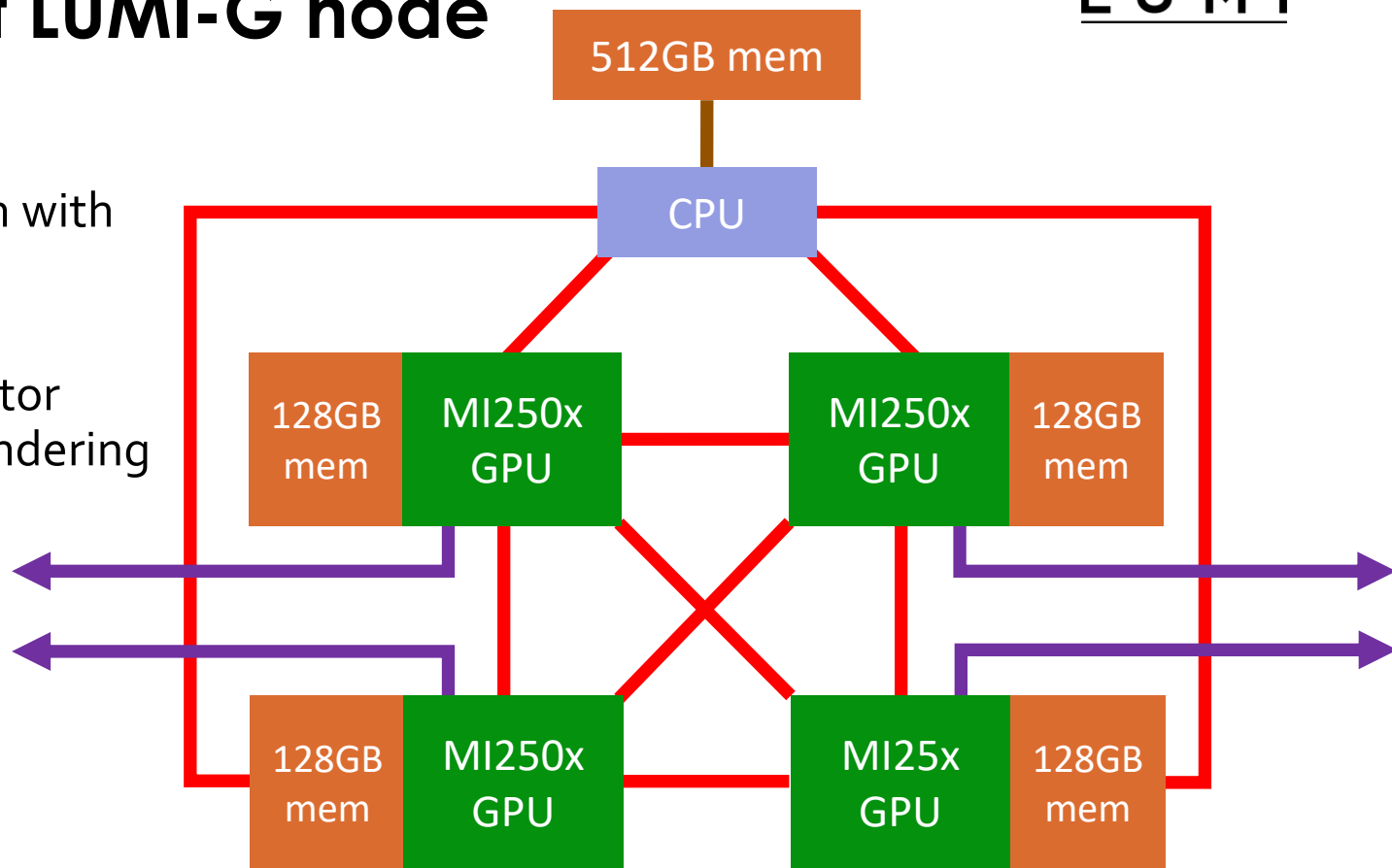
		NUMA nodes CPU 1				NUMA nodes CPU 2			
		0	1	2	3	4	5	6	7
NUMA nodes CPU 1	0	10	12	12	12	32	32	32	32
	1	12	10	12	12	32	32	32	32
	2	12	12	10	12	32	32	32	32
	3	12	12	12	10	32	32	32	32
NUMA nodes CPU 2	4	32	32	32	32	10	12	12	12
	5	32	32	32	32	12	10	12	12
	6	32	32	32	32	12	12	10	12
	7	32	32	32	32	12	12	12	10

- NUMA behaviour not that pronounced within a socket
- but definitely something to take into account between sockets

# Concept LUMI-G node

LUMI

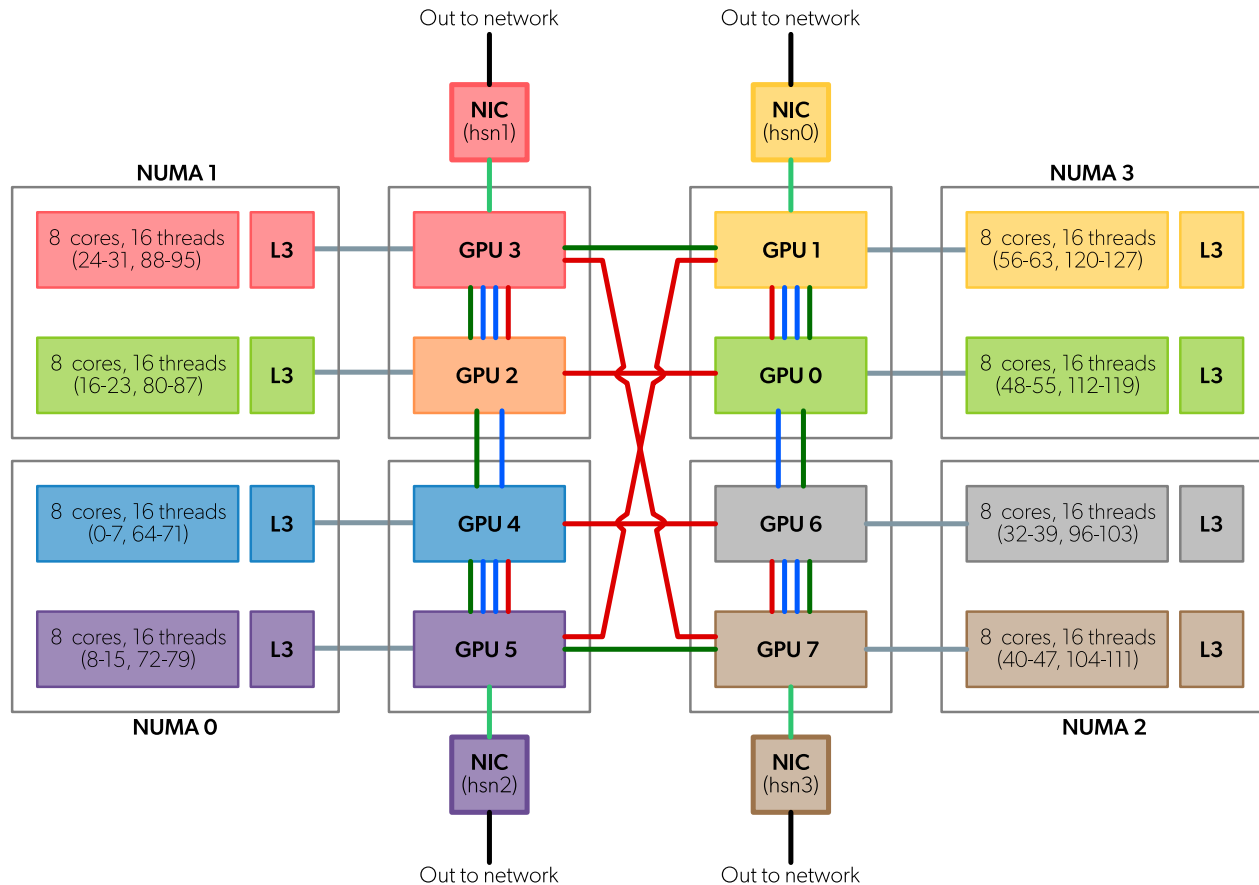
- “GPU first” system with coherent unified memory
- Compute accelerator (CDNA2), not a rendering GPU (RDNA architecture)!



# Real LUMI-G node

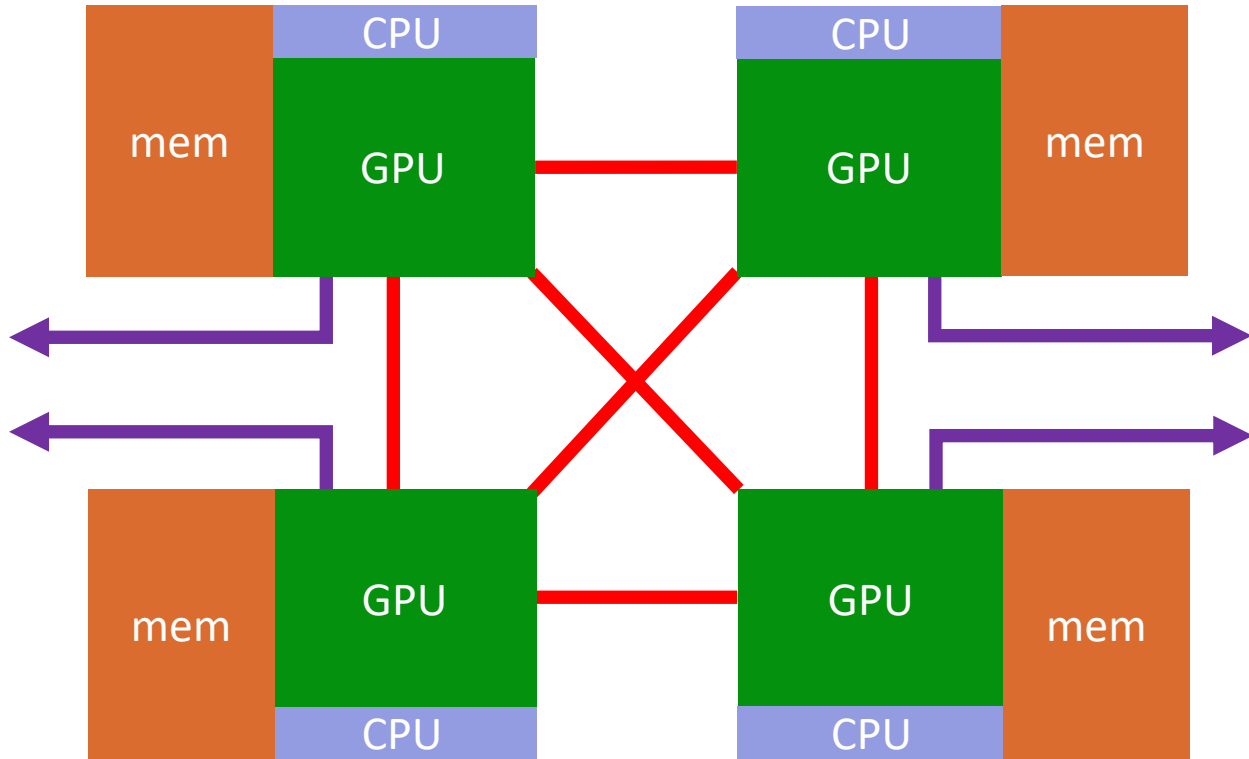
LUMI

- 4 GPUs behave as 8 with 64GB each
- Bandwidth between the dies is low
- Binding to the CCDs is important for performance: Each GPU die closely associated to an L3 cache region



# The future we're preparing for...

L U M I



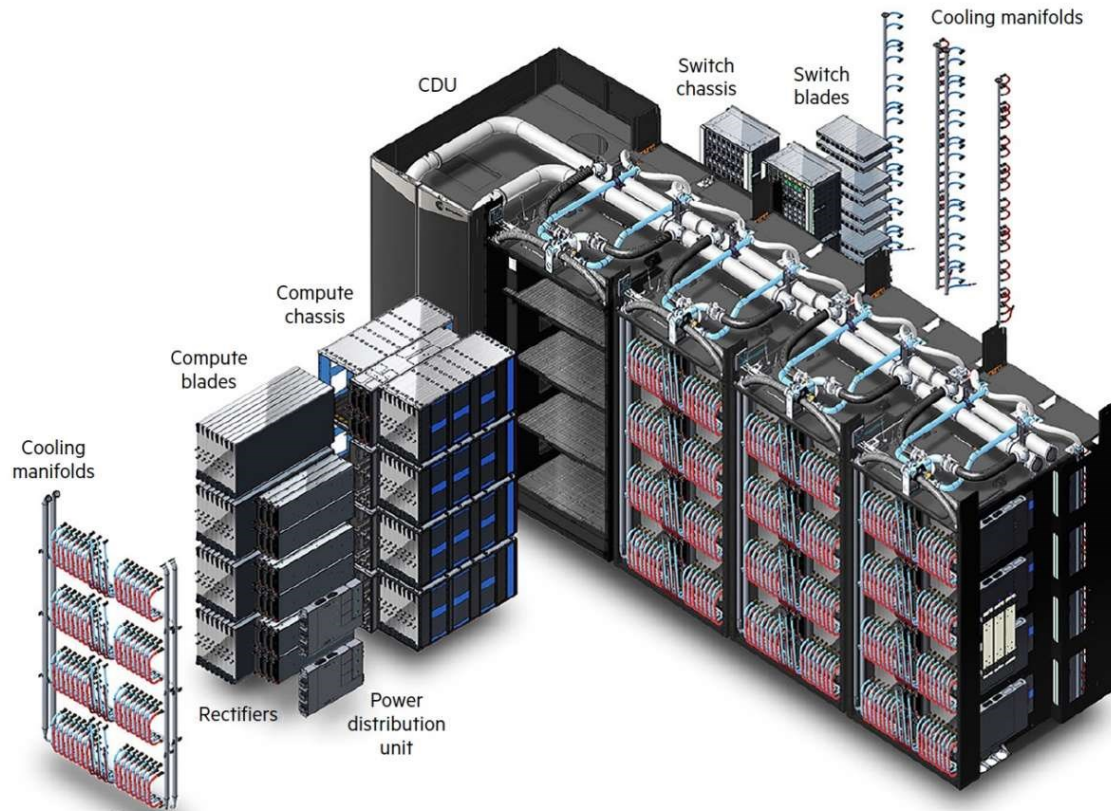
- AMD MI300A (and future Intel GPU?)
- AMD: 13 chiplet design
- CPU and GPU share the memory controllers
- Memory capacity may be a bit disappointing
- Upcoming El Capitan system in the USA

# Slingshot interconnect

- 200 Gb/s (25 GB/s/dir) interconnect based on Ethernet but with proprietary extensions for better HPC performance
  - Adapts to Ethernet devices in the network
  - Lot of attention to adaptive routing and congestion control
  - MPI acceleration
- Not your typical Mellanox/NVIDIA software stack with ucx but libfabric...
- Dragonfly topology
  - 16 switch ports connect to nodes, 48 to other switches
  - 16 or 32 switches in a group with all-to-all connection between the switches in a group
  - Groups are then also connected in an all-to-all way
  - Possible to build large networks where nodes are only 3 hops between switches away on an uncongested network

# HPE Cray EX system

L U M I



- LUMI-C
  - 1 network port/node
  - 4 nodes/compute blade
  - 2 switch blades/chassis
  - 4 nodes on a blade distributed over 2 switches!
- LUMI-G
  - 4 network ports/node
  - 2 nodes/compute blade
  - 4 switch blades/chassis
  - 2 nodes on blade on other switch pair!

# LUMI

# LUMI

