LUMI

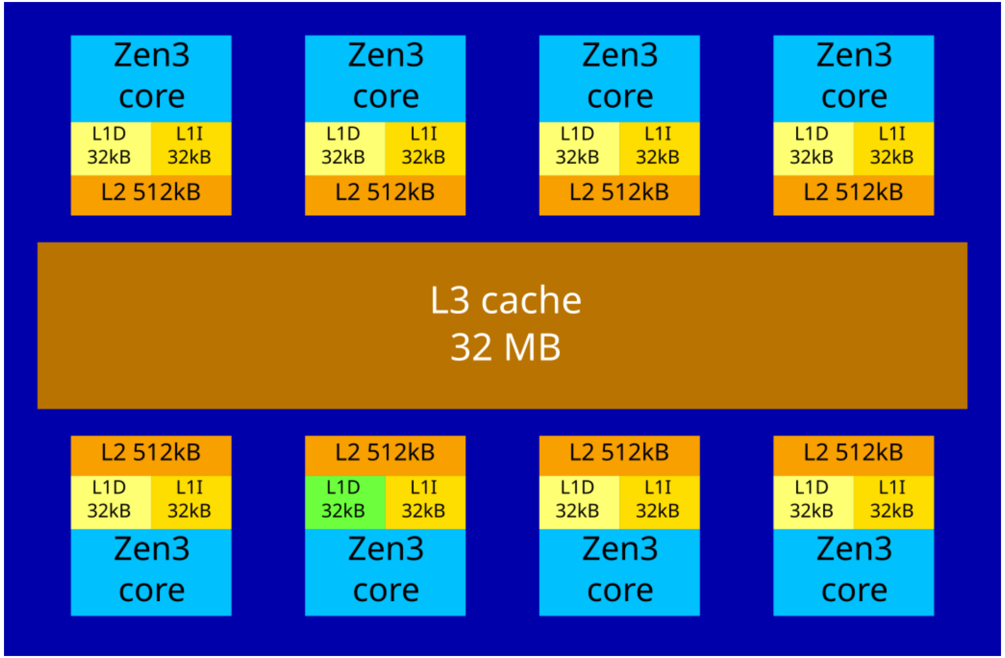
LUMI Architecture

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

LUMI is…

- LUMI-G: 2560 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)
- LUMI-F: 7 PB Lustre flash-based file storage (1740 GB/s)
- LUMI-P: 4 20 PB hard disk based Lustre file systems (4x 240 GB/s)
- Currently 4 user access nodes with two AMD Rome CPUs each
- All linked together with a HPE Cray Slingshot 11 interconnect
- Coming up:
  - Nodes for interactive data analytics: 8 4TB CPU nodes and 8 nodes with 8 GPUs each for visualisation
  - Object based file system
  - Open OnDemand environment

The AMD EPYC 7xx3 (Milan/Zen3) CPU



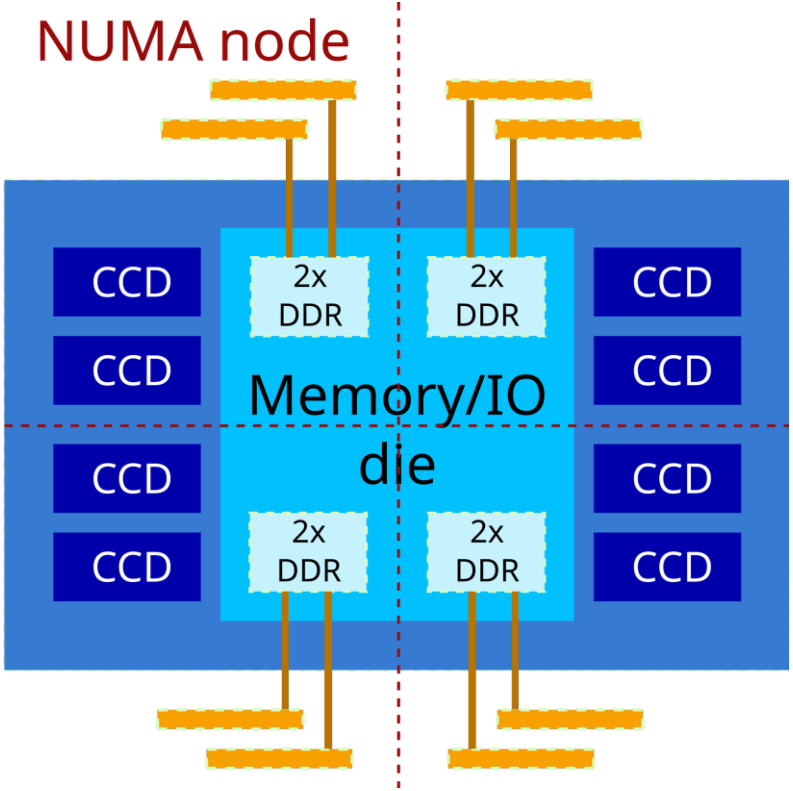
Zen3 core | L1D 32kB | L1I 32kB | L2 512kB
Zen3 core | L1D 32kB | L1I 32kB | L2 512kB
Zen3 core | L1D 32kB | L1I 32kB | L2 512kB
Zen3 core | L1D 32kB | L1I 32kB | L2 512kB

L3 cache
32 MB

L2 512kB | L1D 32kB | L1I 32kB | Zen3 core
L2 512kB | L1D 32kB | L1I 32kB | Zen3 core
L2 512kB | L1D 32kB | L1I 32kB | Zen3 core
L2 512kB | L1D 32kB | L1I 32kB | Zen3 core

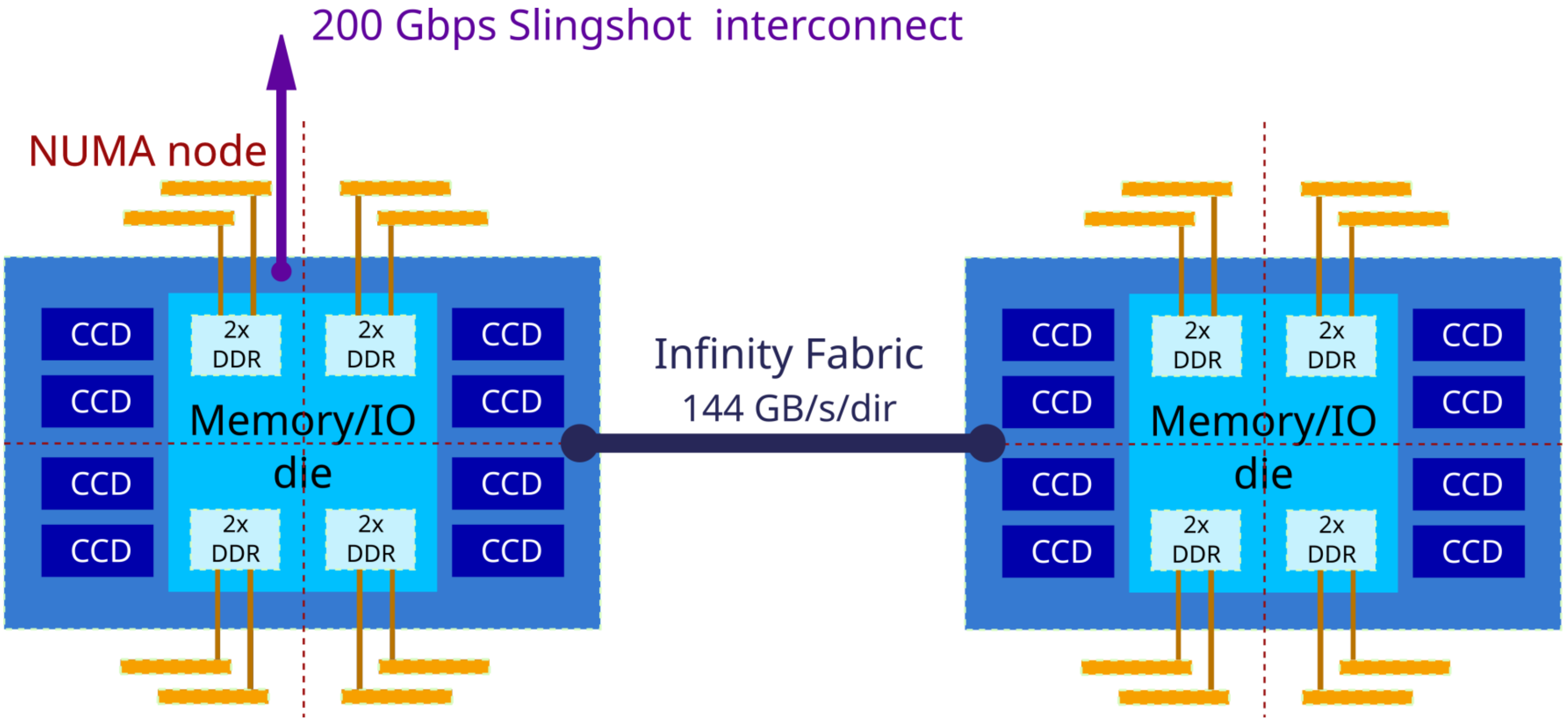
- Building block: a Compute Complex Die (CCD)
- 8 cores
  - Each core has private L1 and L2 caches
  - L3 cache shared
- Instruction set equivalent to Intel Broadwell generation
  - AVX2+FMA, no AVX-512

The AMD EPYC 7xx3 (Milan/Zen3) CPU

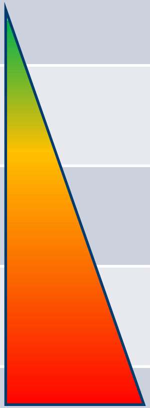
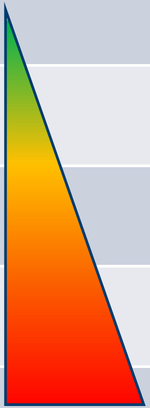
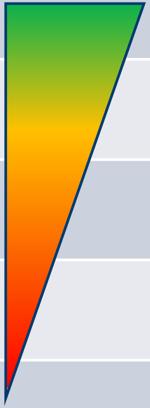


- 8 CCDs or 8 L3 cache regions
- Memory/IO die logically split into 4 NUMA domains with
  - 2 CCDs (16 cores)
  - 2 DDR4 controllers
- Memory/IO die also provides the PCIe links and intersocket links

LUMI-C node


200 Gbps Slingshot interconnect
NUMA node
CCD
CCD
CCD
CCD
2x DDR
2x DDR
Memory/IO die
2x DDR
2x DDR
CCD
CCD
CCD
CCD
Infinity Fabric
144 GB/s/dir
CCD
CCD
CCD
CCD
2x DDR
2x DDR
Memory/IO die
2x DDR
2x DDR
CCD
CCD
CCD
CCD

Strong hierarchy

hierarchy layer		per	sharing	distance	data transfer delay	data transfer bandwicth
1	2 threads	core	L1I, L1D, L2			
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			

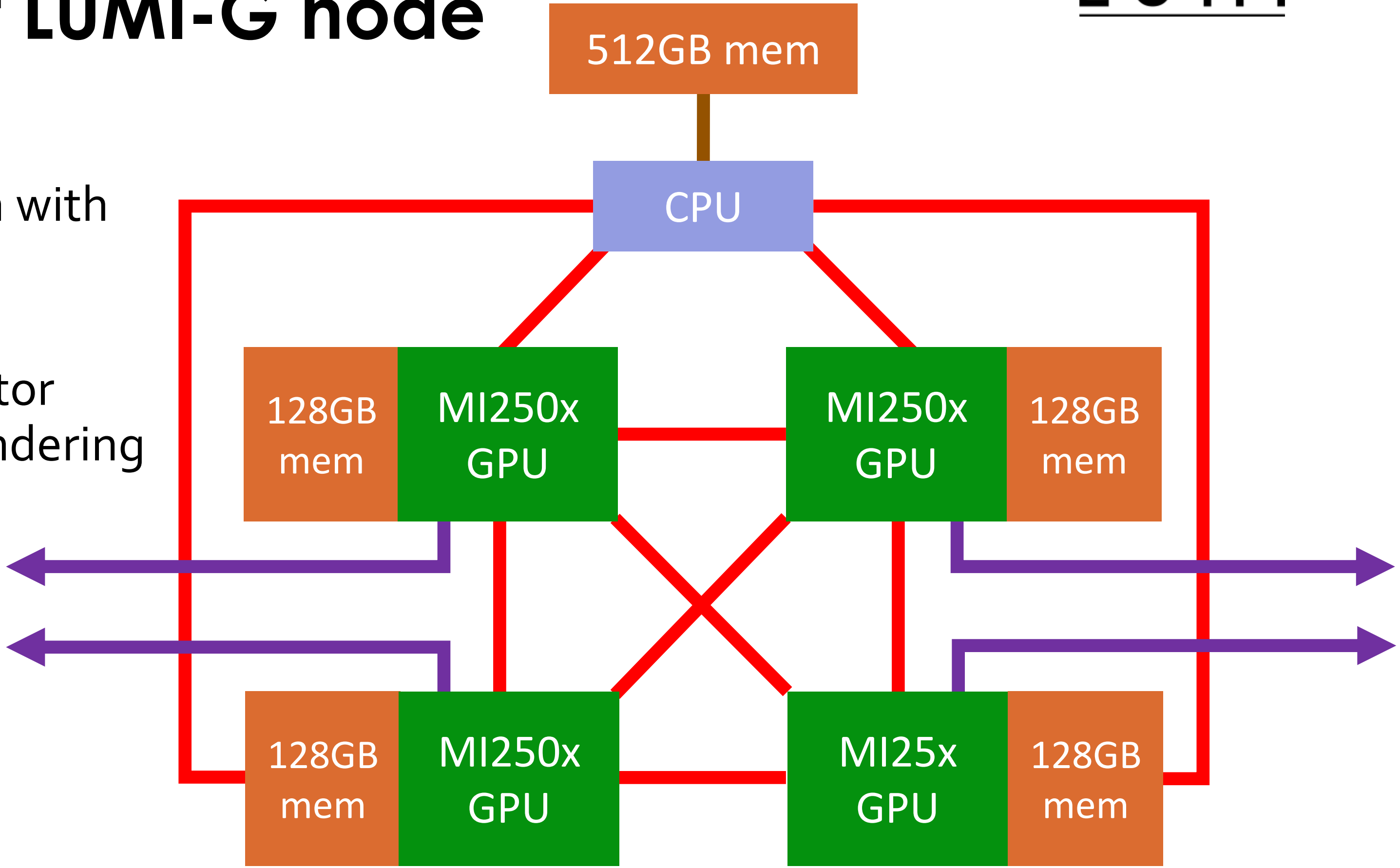
Delays in numbers

		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

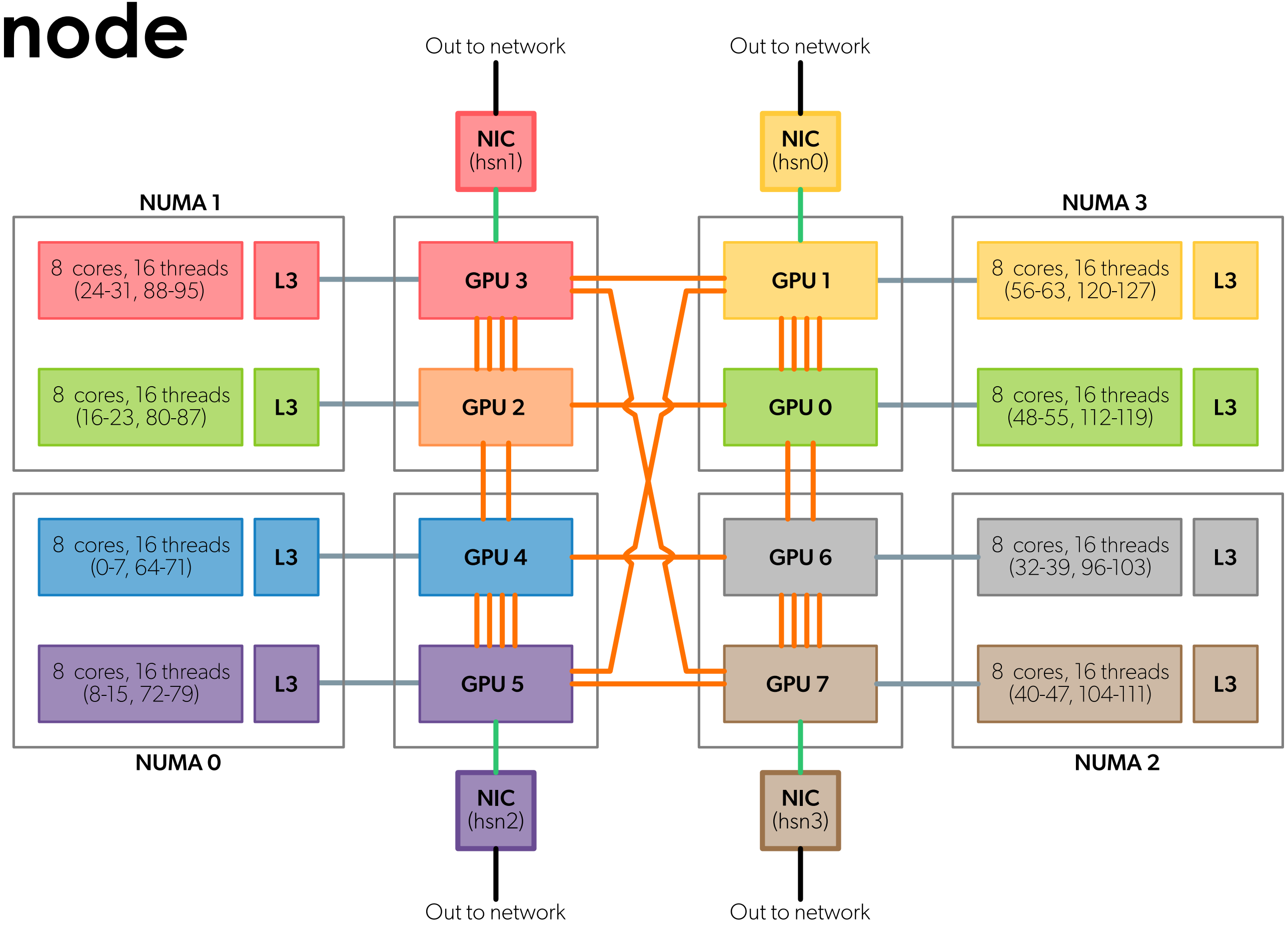
Concept LUMI-G node

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

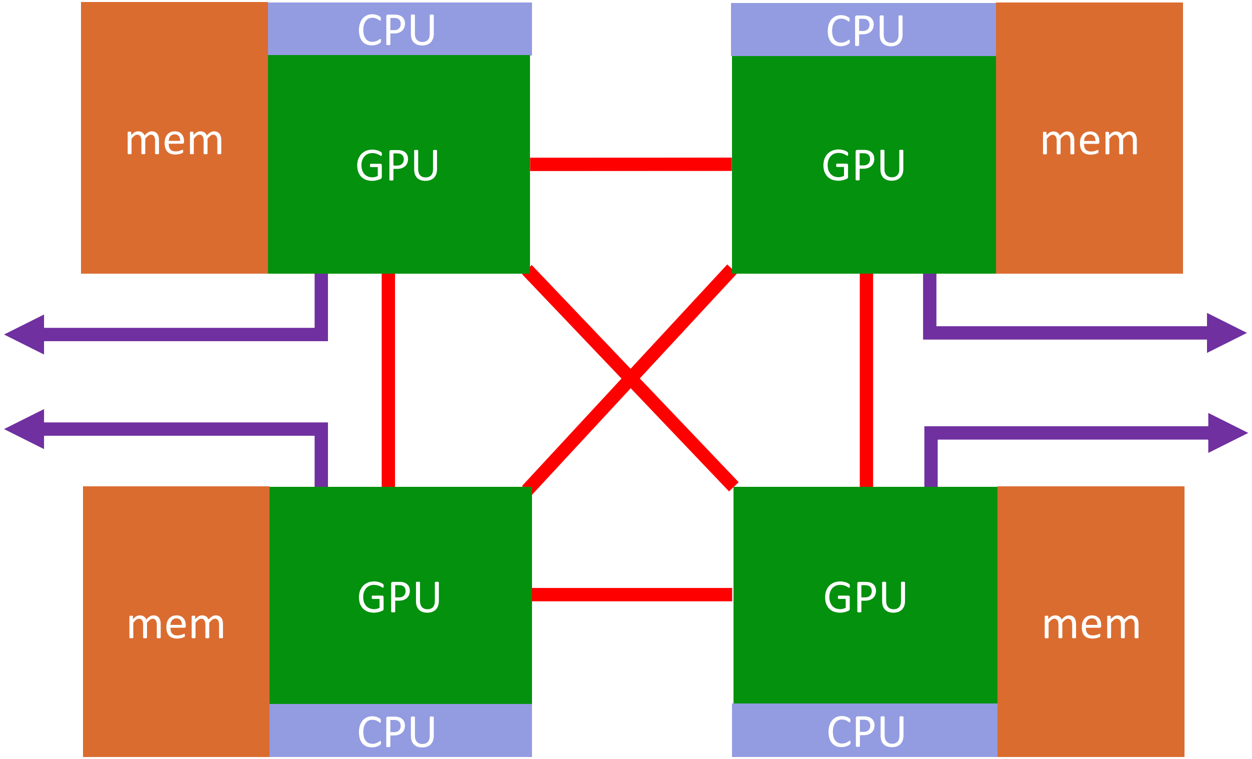


Real LUMI-G node

- 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


Out to network
Out to network
NIC (hsn1)
NIC (hsn0)
NUMA 1
NUMA 3
8 cores, 16 threads (24-31, 88-95)
L3
GPU 3
GPU 1
8 cores, 16 threads (56-63, 120-127)
L3
8 cores, 16 threads (16-23, 80-87)
L3
GPU 2
GPU 0
8 cores, 16 threads (48-55, 112-119)
L3
8 cores, 16 threads (0-7, 64-71)
L3
GPU 4
GPU 6
8 cores, 16 threads (32-39, 96-103)
L3
8 cores, 16 threads (8-15, 72-79)
L3
GPU 5
GPU 7
8 cores, 16 threads (40-47, 104-111)
L3
NUMA 0
NUMA 2
NIC (hsn2)
NIC (hsn3)
Out to network
Out to network

The future we're preparing for…

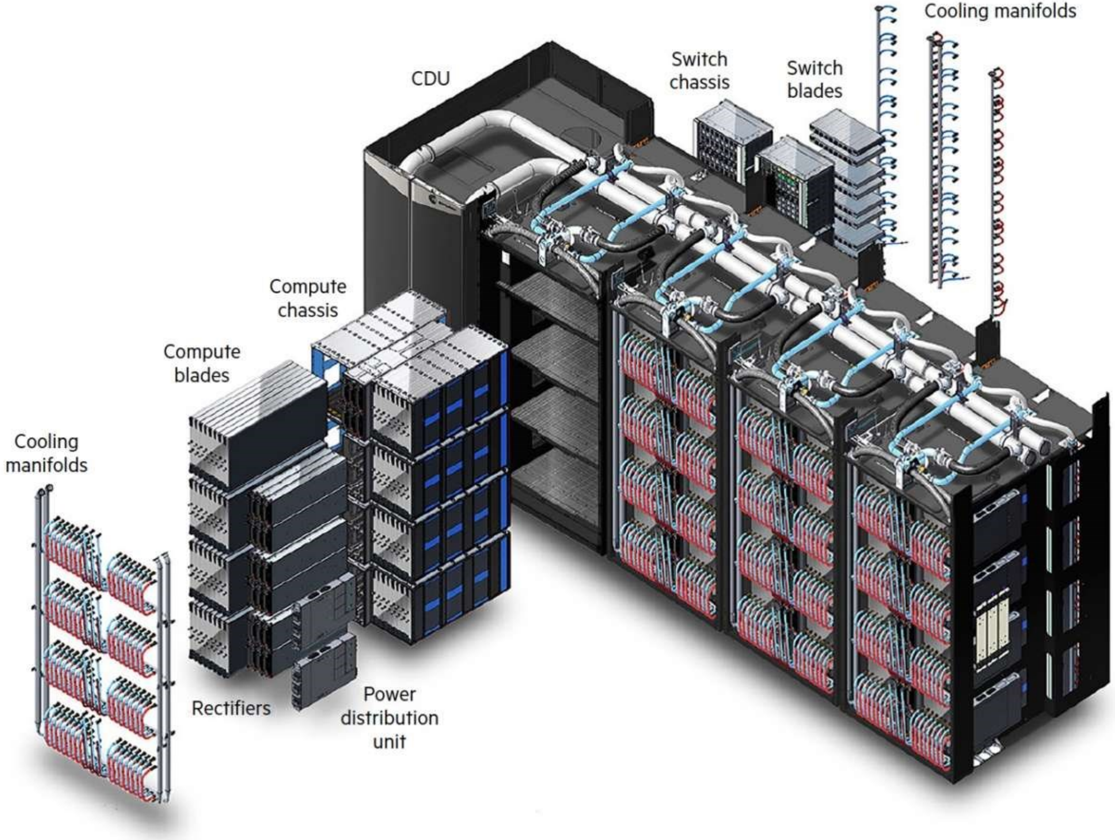


- AMD MI300 (and Intel Falcon Shores ?)
- CPU and GPU share the memory controllers
- Memory capacity may be a bit disappointing

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
  - 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



- 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