LUMI Architecture

Emanuele Vitali
LUMI User Support Team (LUST)
CSC

March 2025
Slides updated from previous version,
authored by Kurt Lust (LUST, UAntwerp)

Why do I need to know this?

LUMI

- **Q:** “I only want to run, why do I need to know about system architecture?”
- **A:** An HPC system is not a big smartphone or big PC
  - Built for large parallel jobs
    - Scaling is not for free
      - Not in Hardware, neither in Software!
  - Often those application needs to be mapped 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
    - The specific problem you’re trying to solve
  - 20% unused performance = several millions of EUROs

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 in the first place
  - 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

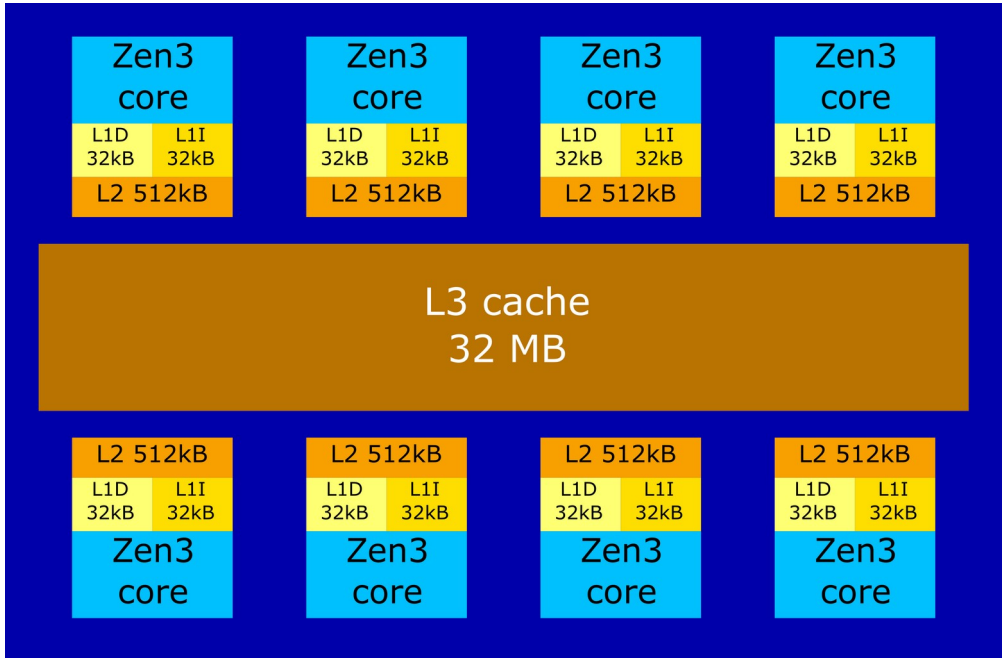
LUMI spec sheet: A modular system

LUMI

- **LUMI-G**: 2978 nodes with 1 AMD EPYC 7A53 CPU and 4 AMD MI250x accelerators (512 GB + 4x128 GB RAM)
- **LUMI-C**: 2048 nodes with 2 64-core AMD EPYC 7763 CPUs (1888x 256GB, 128x 512 GB and 32x 1TB)
- **LUMI-D**: 8 4TB CPU nodes and 8 nodes with 8 A40 GPUs each for visualization
- **LUMI-F**: 8 PB Lustre flash-based file storage (>2 TB/s)
- **LUMI-P**: 4 20 PB hard disk based Lustre file systems (4x 240 GB/s)
- **LUMI-O**: 30 PB object based file system
- **LUMI-L**: 4 user access nodes with two AMD Rome CPUs each for ssh access and some for web access via Open OnDemand
- All linked together with a HPE Cray Slingshot 11 interconnect

The AMD EPYC 7xx3 (Milan/Zen3) CPU

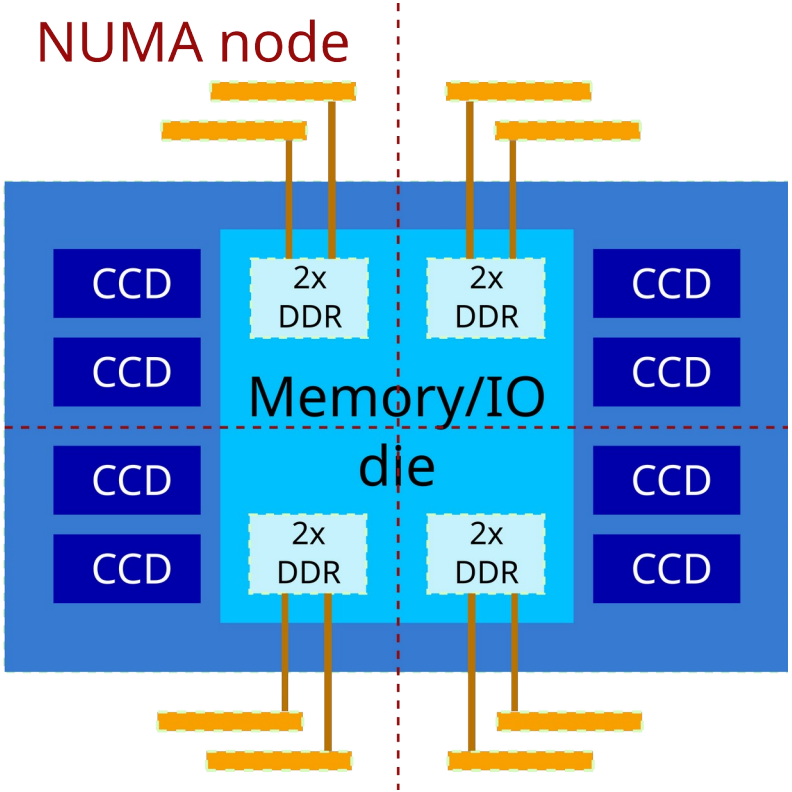
LUMI



- 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 and Rome/Zen2
  - AVX2+FMA, no AVX-512
- *Rome/Zen2 has 2 4-core core complexes per die.*
  - *Milan architecture simpler*

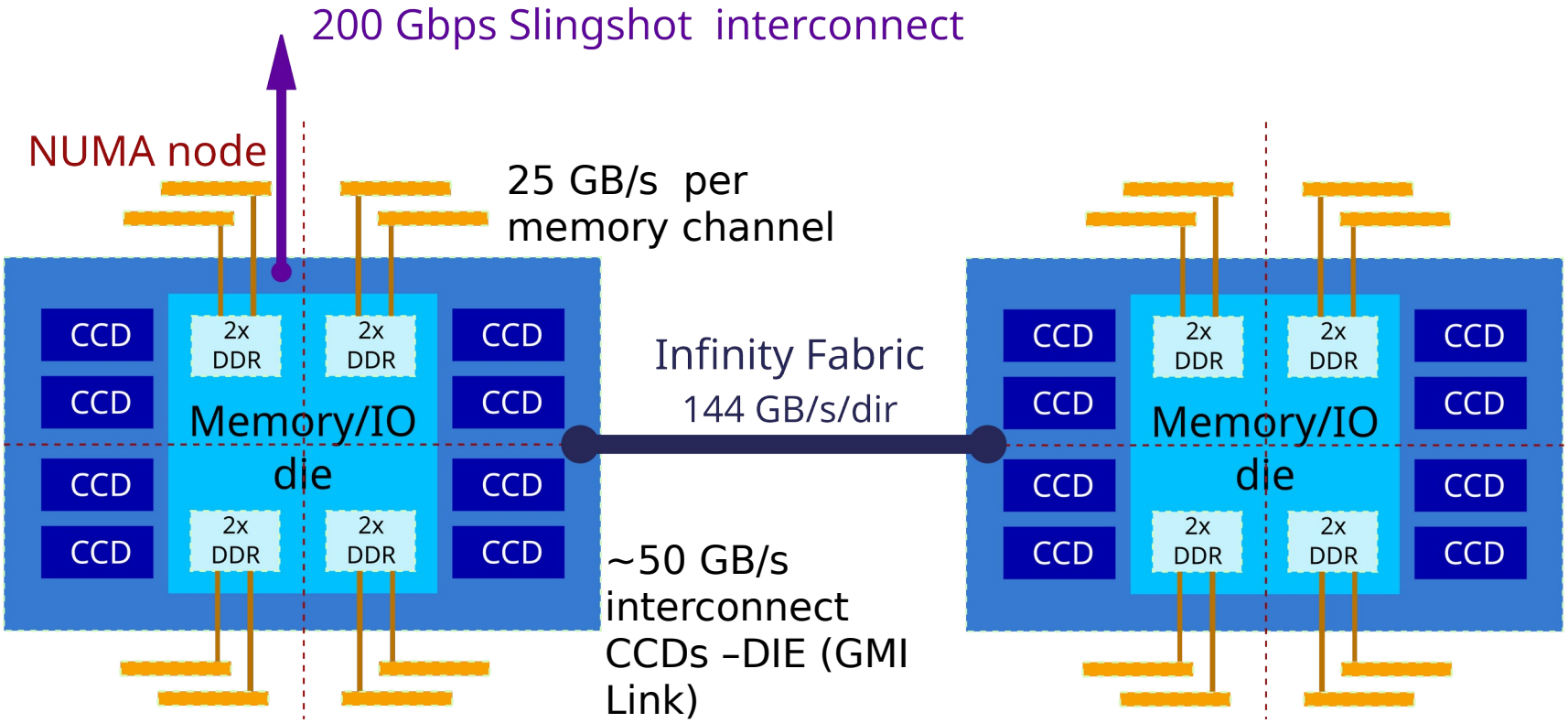
The AMD EPYC 7xx3 (Milan/Zen3) CPU

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
- *Comparision with Rome/zen2:*
  - *Rome/zen2 has 16 L3 cache regions with 4 cores each (in the 64-core variant)*

LUMI-C node


200 Gbps Slingshot interconnect
NUMA node
25 GB/s per memory channel
CCD
2x DDR
Memory/IO die
Infinity Fabric
144 GB/s/dir
~50 GB/s interconnect CCDs –DIE (GMI Link)

Strong hierarchy

LUMI

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

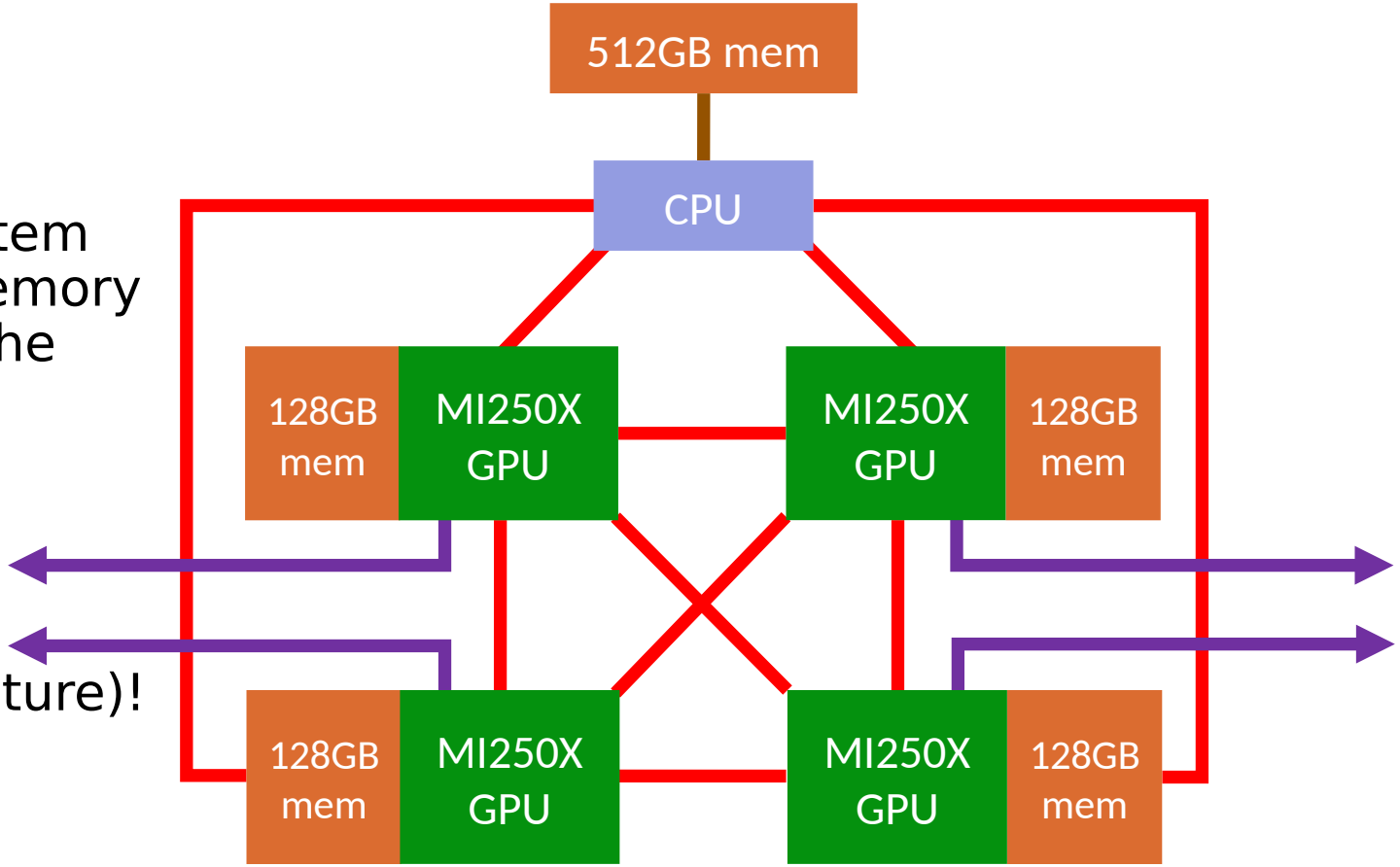
NUMA nodes	NUMA nodes CPU 1				NUMA nodes CPU 2			
	0	1	2	3	4	5	6	7
NUMA nodes CPU 1 0	10	12	12	12	32	32	32	32
1	12	10	12	12	32	32	32	32
2	12	12	10	12	32	32	32	32
3	12	12	12	10	32	32	32	32
NUMA nodes CPU 2 4	32	32	32	32	10	12	12	12
5	32	32	32	32	12	10	12	12
6	32	32	32	32	12	12	10	12
7	32	32	32	32	12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

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

Concept LUMI-G node

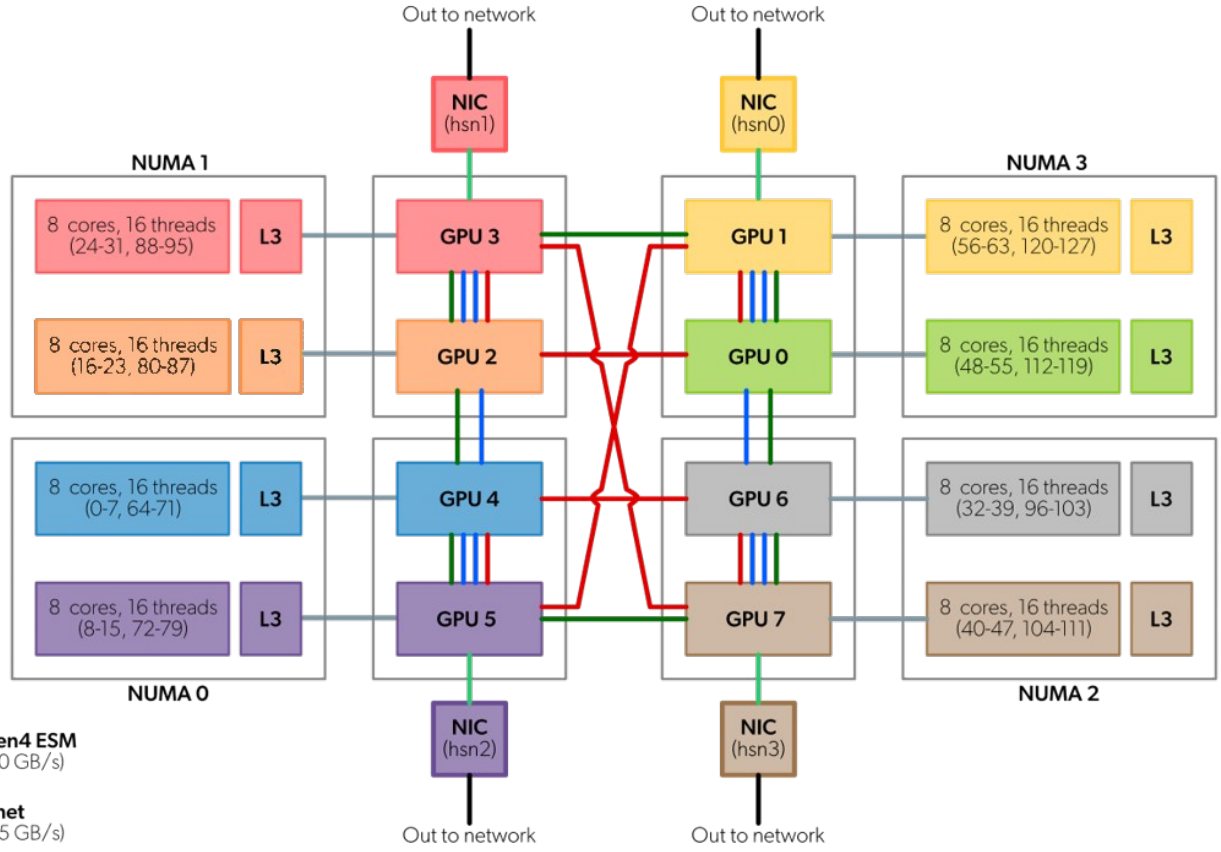
LUMI

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



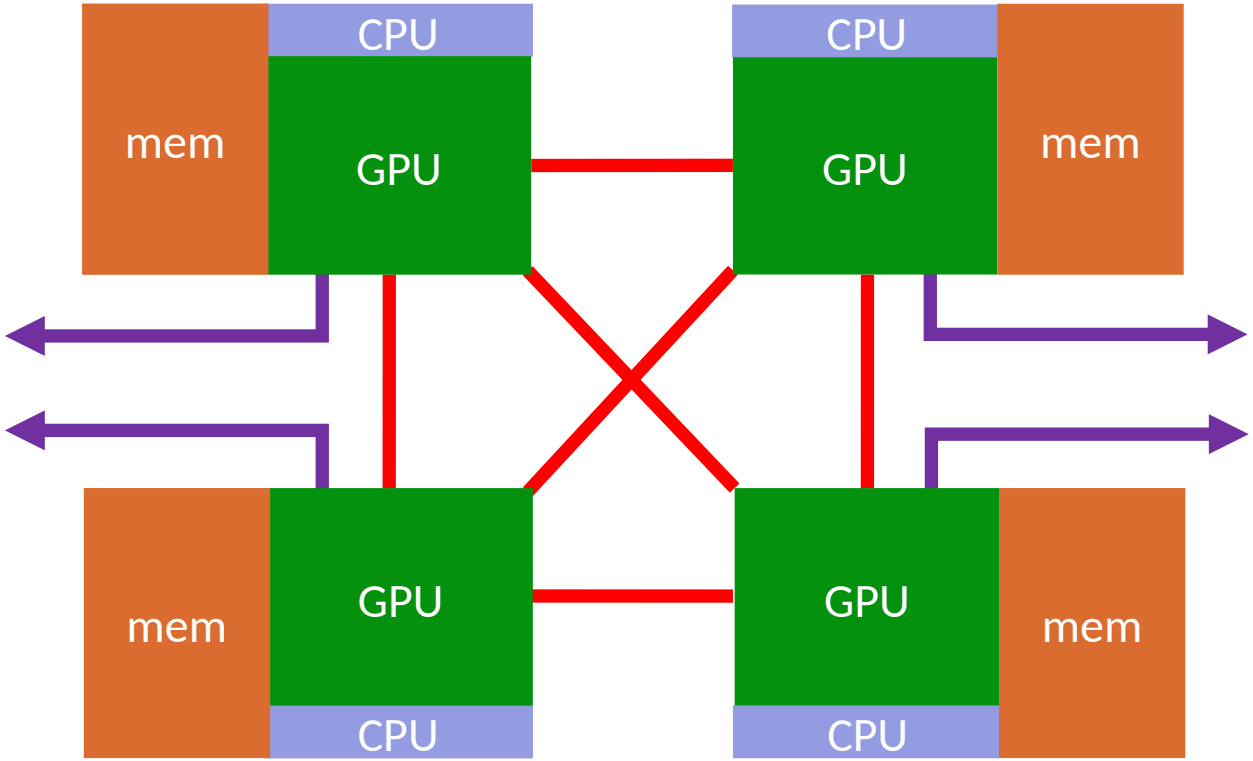
Real LUMI-G node

- MI250X is built with 2 chiplets
- Chiplet memory has 64GB each, 1.6TB/s Bandwidth
- Bandwidth between chiplets is only 200GB/s



The next gen: El Capitan

LUMI



- AMD MI300A
- 13 chiplet design:
  - 4 memory/IO,
  - 6 GPU,
  - 3 CPU
- CPU and GPU share the memory controllers
- 128 GB memory capacity per package may be a bit disappointing
- Current TOP500 #1

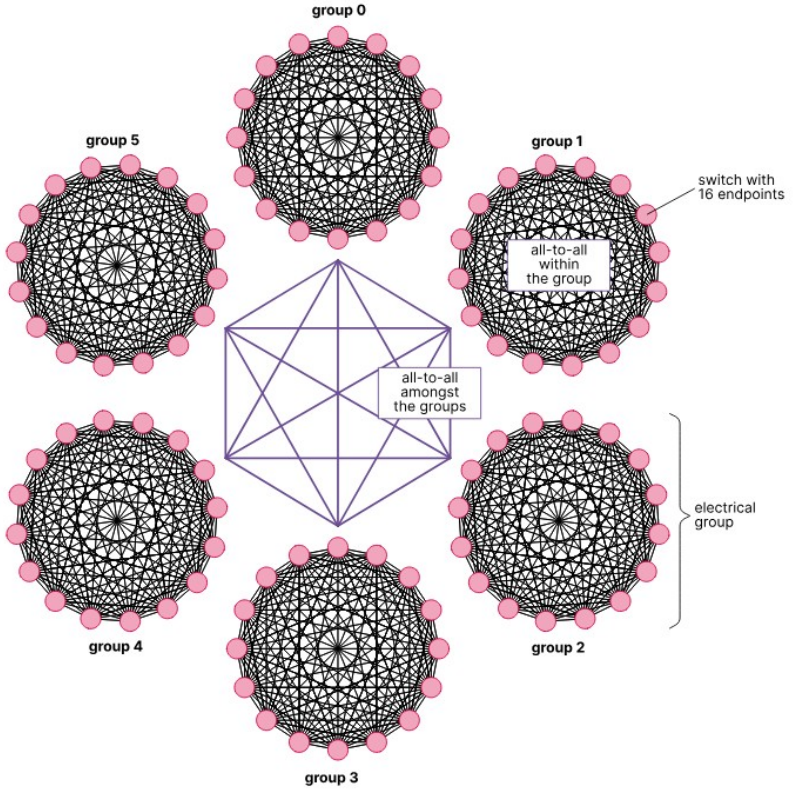
Slingshot interconnect

LUMI

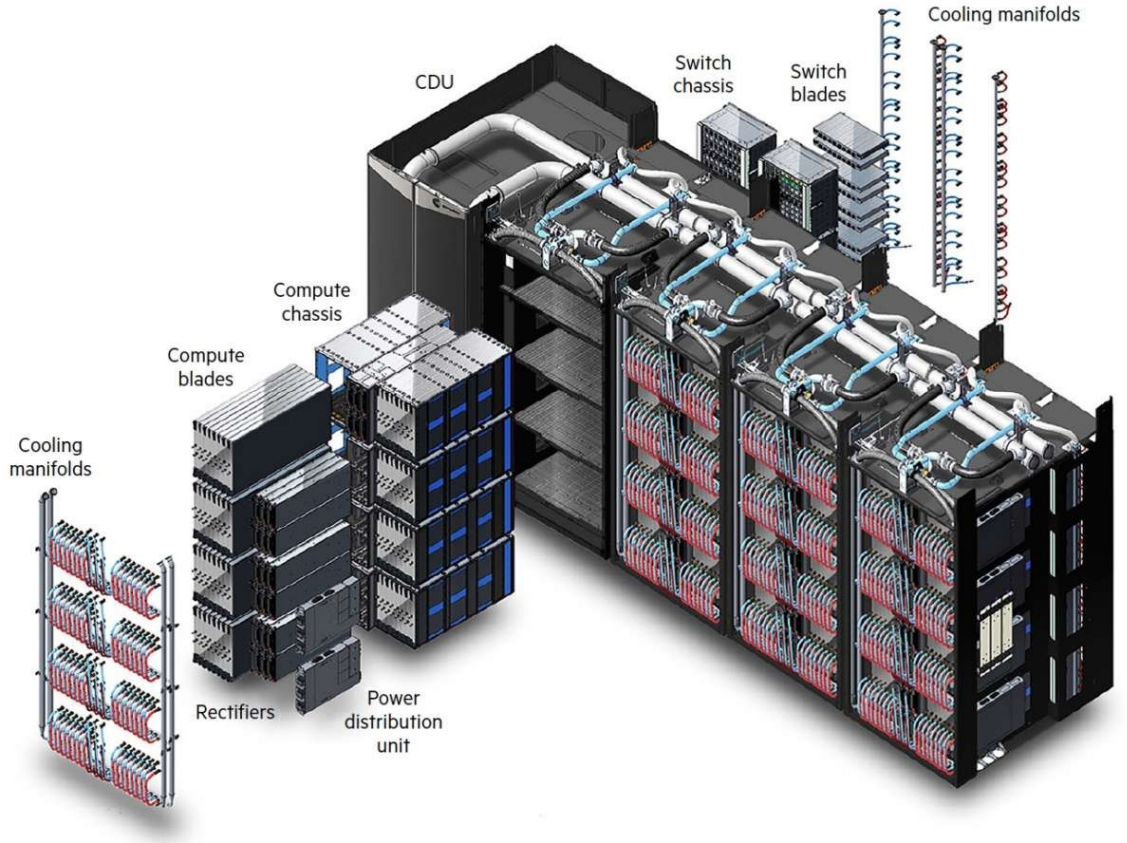
- 200 Gb/s (25 GB/s/dir) interconnect based on Ethernet but with proprietary extensions for better HPC performance
  - Adapts to Ethernet devices in the network
  - Lot of attention to adaptive routing and congestion control
  - MPI acceleration
- Not your typical Mellanox/NVIDIA software stack with ucx but libfabric…

Slingshot interconnect

- 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



- 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 (artist rendering)

LUMI


LUMI
Questions?