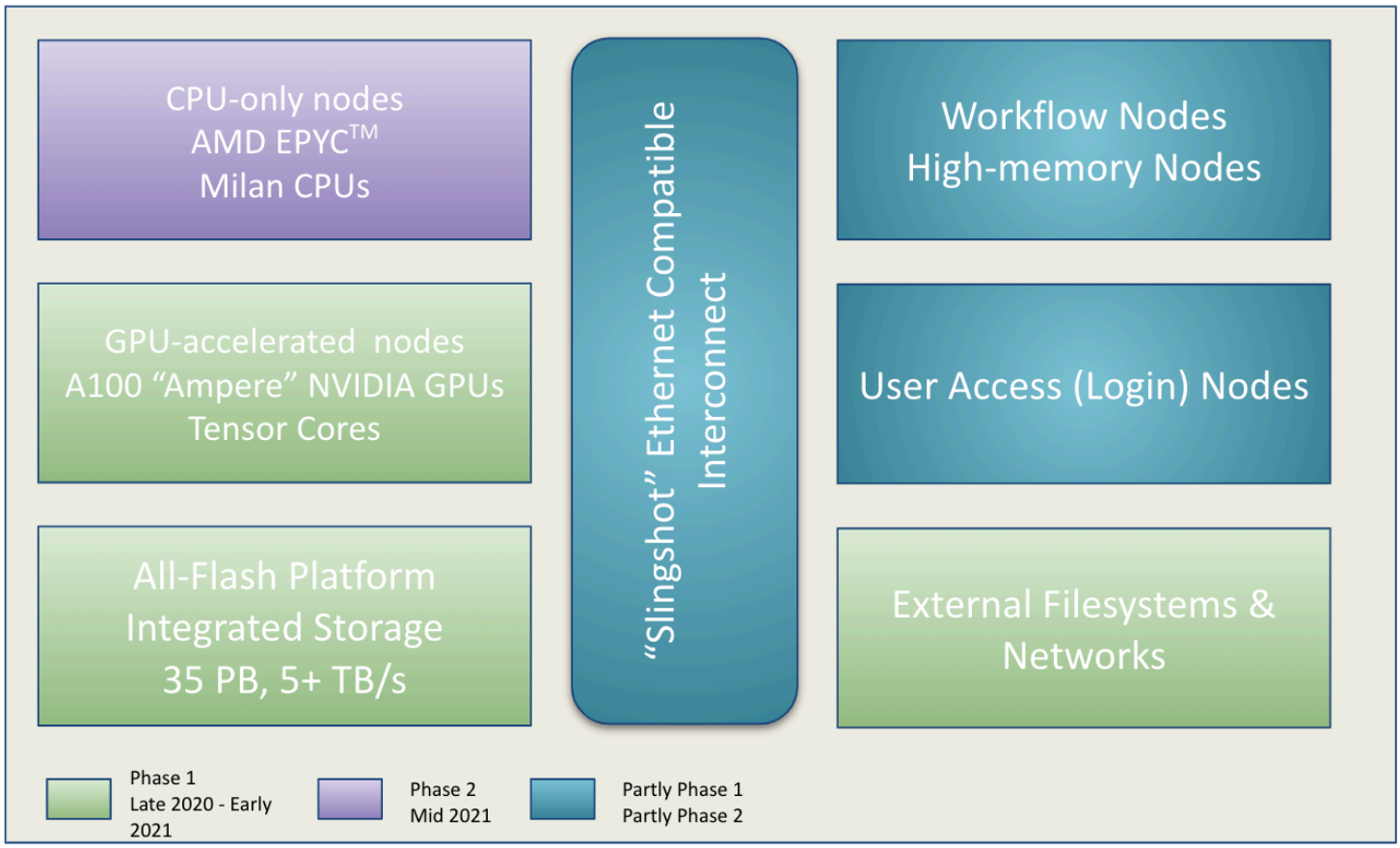
Perlmutter

Perlmutter is a HPE (Hewlett Packard Enterprise) Cray EX supercomputer, named in honor of Saul Perlmutter, an astrophysicist at Berkeley Lab who shared the 2011 Nobel Prize in Physics for his contributions to research showing that the expansion of the universe is accelerating. Dr. Perlmutter has been a NERSC user for many years, and part of his Nobel Prize-winning work was carried out on NERSC machines and the system name reflects and highlights NERSC's commitment to advancing scientific research.

Perlmutter, based on the HPE Cray Shasta platform, is a heterogeneous system comprising both CPU-only and GPU-accelerated nodes, with a performance of 3-4 times Cori when the installation completes. The system is scheduled to be delivered in two phases: Phase 1, with 12 GPU-accelerated cabinets housing over 1,500 nodes, and 35PB of all-flash storage, was delivered by early 2021, and Phase 2 with 12 CPU cabinets will be delivered later in 2021.



System Overview - Phase 1

System Partition	# of cabinets	# of nodes	CPU Aggregate Theoretical Peak (FP64 in PFlops)	CPU Aggregate Memory (TiB)	GPU Aggregate Theoretical Peak (PFlops)	GPU Aggregate Memory (TiB)
GPU-accelerated compute nodes	12	1,536	3.9	384	FP64: 59.9 TF32 Tensor: 958.1	240
Login Nodes	-	15	0.07	7.5	FP64: 0.1 TF32 Tensor: 2.3	0.6

System Specification - Phase 1

CPUs

System Partition	Processor	Clock Rate (MHz)	Cores per Socket	Threads/Core	Sockets per Node	Memory per Node (GiB)
GPU-accelerated compute nodes	AMD EPYC 7763 (Milan)	2450	64	2	1	256
Login Nodes	AMD EPYC 7742 (Rome)	2250	64	2	2	512

GPUs

System Partition	Processor	Clock Rate (MHz)	SMs per GPU	INT32, FP32, FP64, Tensor cores per GPU	GPUs per Node	Memory per Node (GiB)
GPU-accelerated compute nodes	NVIDIA A100 GPU	1410	108	6912, 6912, 3456, 432	4	160
Login Nodes	NVIDIA A100 GPU	1410	108	6912, 6912, 3456, 432	1	40

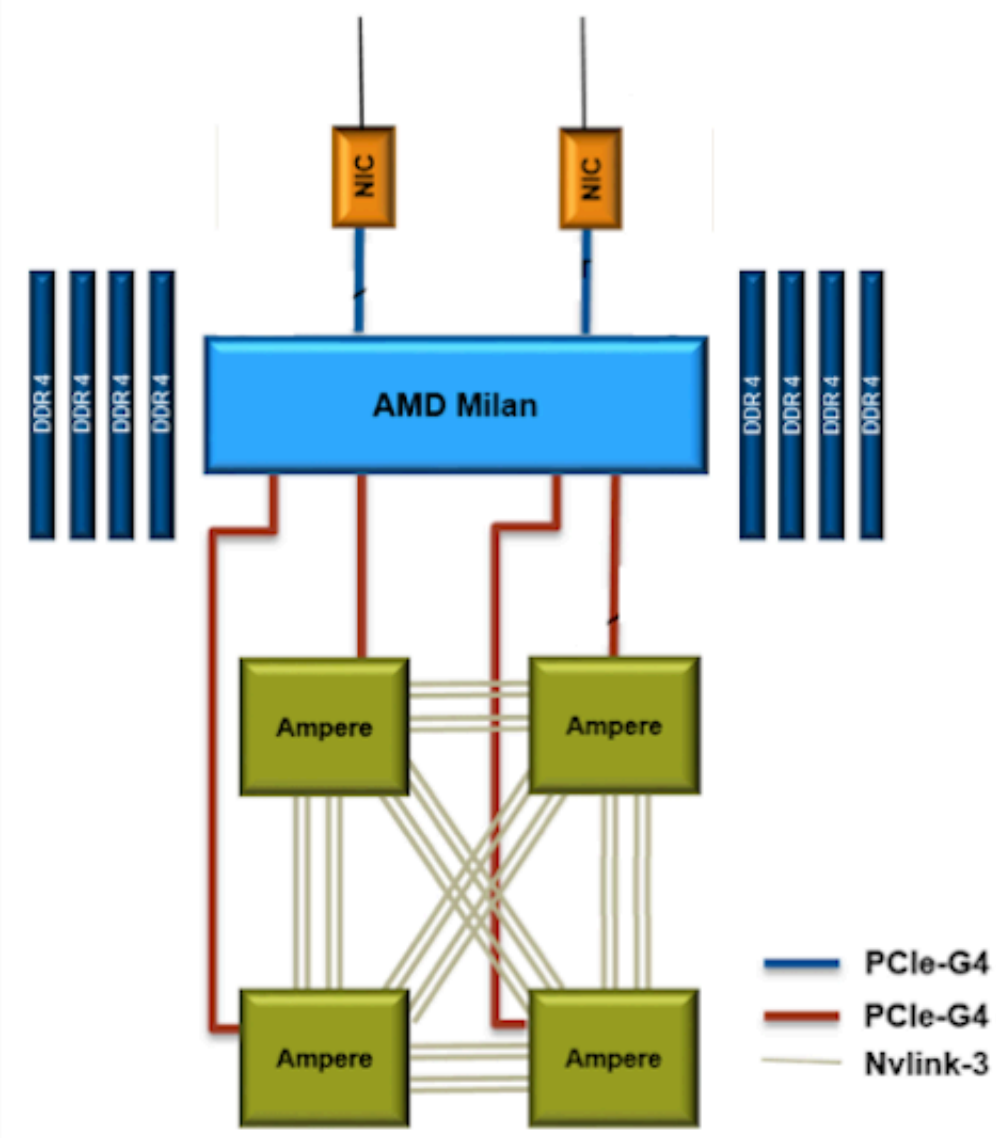
System Details - Phase 1

- The compute subsystem comprises of 12 compute cabinets of GPU accelerated nodes.

- Phase 1's compute cabinet is segmented into 8 chassis, each containing 8 compute blades and 4 switch blades. In total, a GPU cabinet contains 64 compute blades and 32 switch blades.
- A compute blade contains 2 GPU accelerated nodes, making a total of 128 nodes per cabinet or 1536 nodes for the system.
- The interconnect is HPE Cray Slingshot and consists of the network switches and the network interface cards (NICs). All node types within either compute or non-compute cabinets are connected via the Slingshot network fabric.
- Switch blades provide the high-speed network for the compute blades. Each switch blade connects to all 8 compute blades within the chassis.
- Nodes on the Slingshot network fabric can access networks external to the system via the high-speed network, suppporting the capability of services external to the system to utilize the compute nodes via the scheduler.
- The system supports a high-performance, parallel I/O itnerface for accessing external kernel-based file systems, allowing data transport between user applications on compute nodes and interfaces to such extgernal file systems.
- Compute cabinets are liquid-cooled. Redundancy in cooling units is provided as needed. All non-compute nodes housed in service cabnets are air-cooled.

Node Specifications

GPU-Accelerated Compute Nodes



- The accelerated nodes consist of a single socket of an AMD EPYC 7763 (Milan) processor and four NVIDIA Ampere A100 GPUs.
- The Milan CPU is connected to all GPUs and the NICs via PCIe 4.0.
- There are 2 NICs per node.

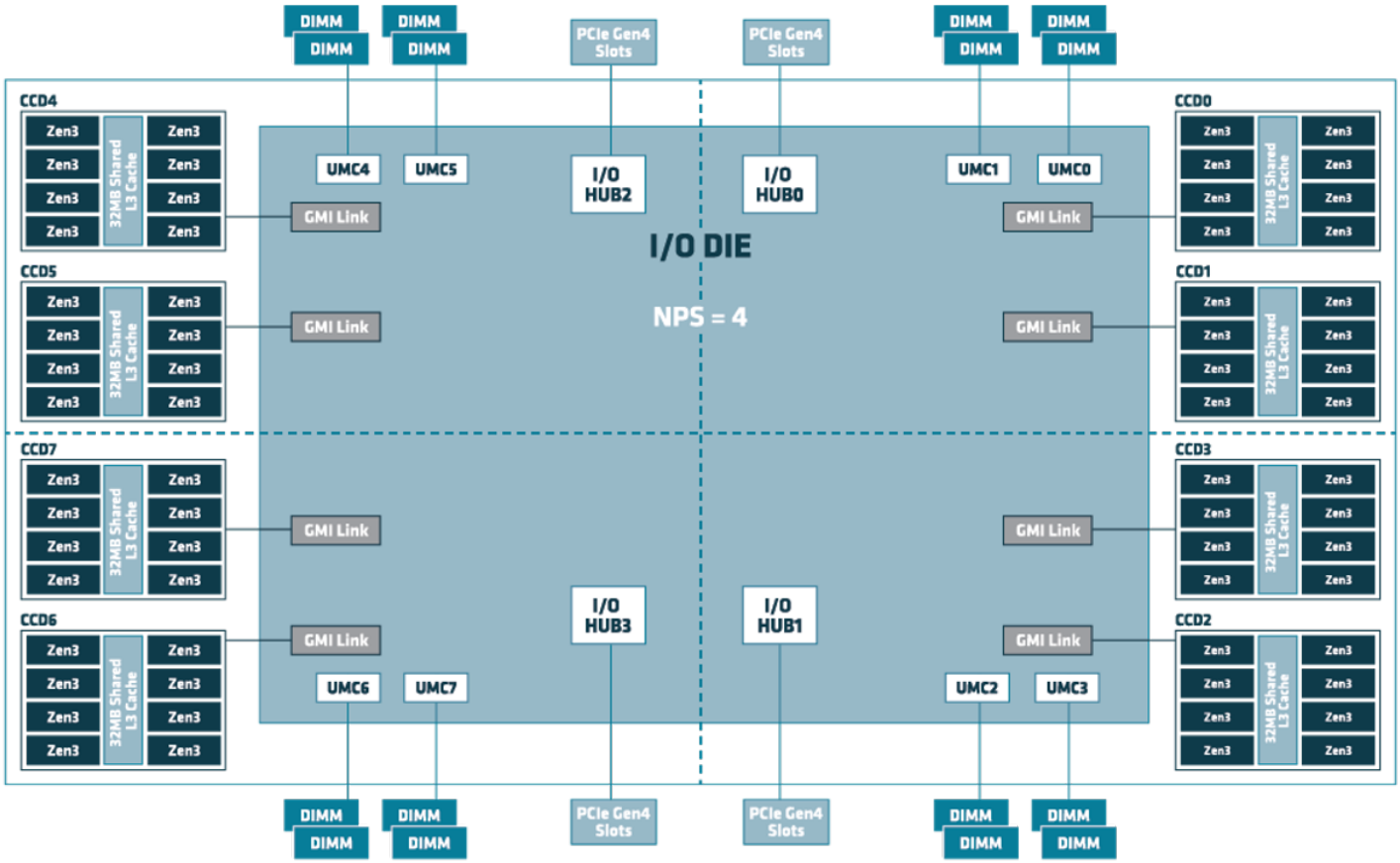
AMD EPYC 7763 (Milan) Processor

AMD EPYC processors use a multi-chip module (MCM) design where separate dies are provided for CPU and I/O components for easier scalability. The CPU dies are called CCDs (Core Complex Dies) and the I/O dies are often denoted as IODs.

The CCDs connect to memory, I/O, and each other through the IOD. Each CCD connects to the IOD via a dedicated, high-speed Global Memory Interconnect (GMI) link. The IOD also contains memory channels, PCIe Gen4 lanes, and Infinity Fabric links. All dies, or chiplets, interconnect with each other via AMD's Infinity Fabric

Technology.

Milan is the codename for AMD's third generation EPYC processor series, which was launched in March, 2021.



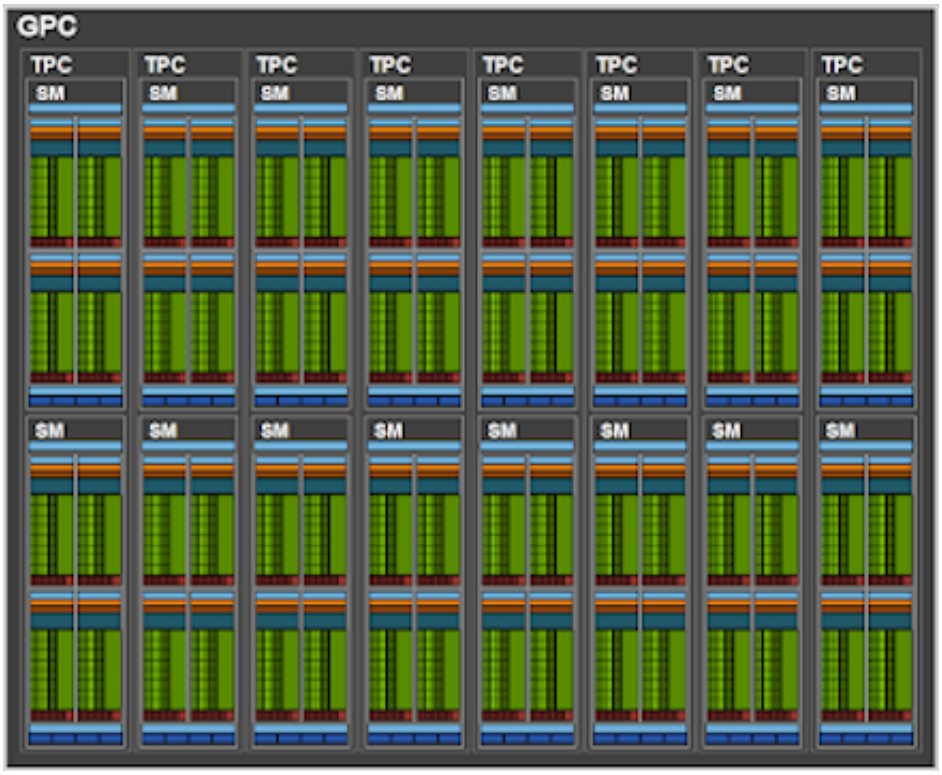
- An EPYC 7763 processor is a Milan processor, based on 64 AMD "Zen 3" compute cores.
- A Zen 3 core supports Simultaneous Multithreading (SMT), allowing 2 execution threads (hardware threads) to execute simultaneously per core. It supports AVX2 for 256-bit wide SIMD operations. Its clock rate is 2450 MHz.
- A core has a 32KiB L1 write-back data cache and a 512KiB unified (instruction/data) L2 cache.
- Eight cores share a single 32MiB L3 cache, and this grouping is referred to as a Core Complex (CCX). A single CCX is contained within a single CCD.
- An EPYC 7763 processor has eight CCDs and one IOD, as depicted by the diagram above.
- The theoretical peak performance values are as follows:
  - 39.2 GFlops per core
  - 2.51 TFlops per socket
  - 3.85 PFlops total for GPU accelerated nodes
- This Milan processor has 8 memory controllers, supporting 3200MHz DDR4, for 256GiB of memory and the

maximum bandwidth of 204.8 GB/s per socket. The total aggregate memory for the Phase 1 compute subsystem is 384TiB.

- The IOD can be configured for various NUMA node topologies. The current configuration is 4 NUMA nodes per socket (NPS), denoted as 'NPS=4.' This configuration is shown in the architecture diagram above.
- For more info, please check High Performance Computing (HPC) Tuning Guide for AMD EPYC 7003 Series Processors.

NVIDIA Ampere A100 GPU

The architecture diagram below (top) is for the full implementation of the NVIDIA GA100 GPU. The GPU is partitioned into 8 GPU Processing Clusters (GPCs). A GPC is made of 8 Texture Processing Clusters (TPCs), with 2 Streaming Multiprocessors (SMs) per TPC, as shown in the bottom diagram. The GPU has 12 memory controllers.



The NVIDIA A100 Tensor Core GPU implementation of the GA100 GPU has a slightly different configuration, as explained below.

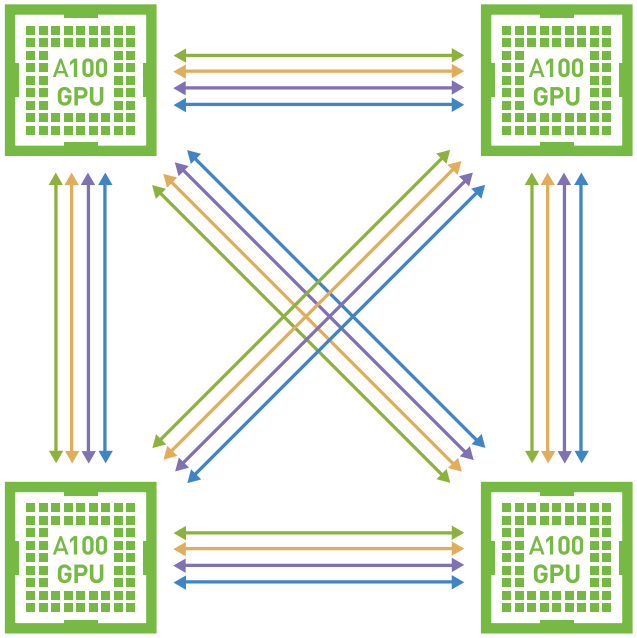
- The A100 GPU has 7 active GPCs. Two of them have 7 TPCs while the rest have 8, which leads to 108 SMs per GPU.
- The A100 GPU has 10 512-bit memory controllers, for 40 GiB HBM2 (High Bandwidth Memory, the 2nd generation) at the maximum bandwidth of 1555.2 GB/s. The aggregate memory for the entire compute subsystem is 240TiB from 4 x 1536 = 6144 GPUs.
- L2 data cache of 40 MiB is divided into 2 partitions to enable higher bandwidth and lower latency memory access. Each L2 partition localizes and caches data for memory accesses from SMs in the GPCs directly connected to the partition.
- The A100 GPU has a new feature called Multi-Instance GPU (MIG) that allows a GPU to be configured into up to seven separate GPU instances for executing multiple applications separately.

- As shown in the diagram above, each SM has 64 INT32, 64 FP32, 32 FP64, and 4 Tensor cores, totaling 6912 INT32, 6912 FP32, 3456 FP64, and 432 Tensor cores per GPU.
- The register file size is 256 KiB per SM.
- An SM has 192KiB of unified L1 data cache and shared memory.
- The GPU is running at 1410 MHz.
- A feature called Sparsity can exploit fine-grained structured sparsity in deep learning networks to double the throughput of Tensor core operations.
- Theoretical peak performance values for some operations are:

Operations	GPU (TFlops)	Node (TFlops)	System (PFlops)
FP32	19.5	78.0	119.8
FP64	9.7	39.0	59.9
TF32 Tensor	155.9 | 311.9*	623.7 | 1247.5*	958.1 | 1916.1*
FP16 Tensor	311.9 | 623.7*	1247.5 | 2495.0*	1916.1 | 3832.3*
FF64 Tensor	19.5	78.0	119.8

* With Sparsity

- Note that a GPU-accelerated node contains 4 GPUs. These GPUs are connected to each other with NVLink-3, the third generation NVLink.



- Each NVLink connection provides 25 GB/s/direction for a total aggregate of 100 GB/s/direction between 2 GPUs.
- A single GPU has a total of 12 links with the others, yielding 600 GB/s total bandwidth.
- For more info, please check the A100 whitepaper or the NVIDIA Developer blog post NVIDIA Ampere

Architecture In-Depth.

Login Nodes

A login node has:

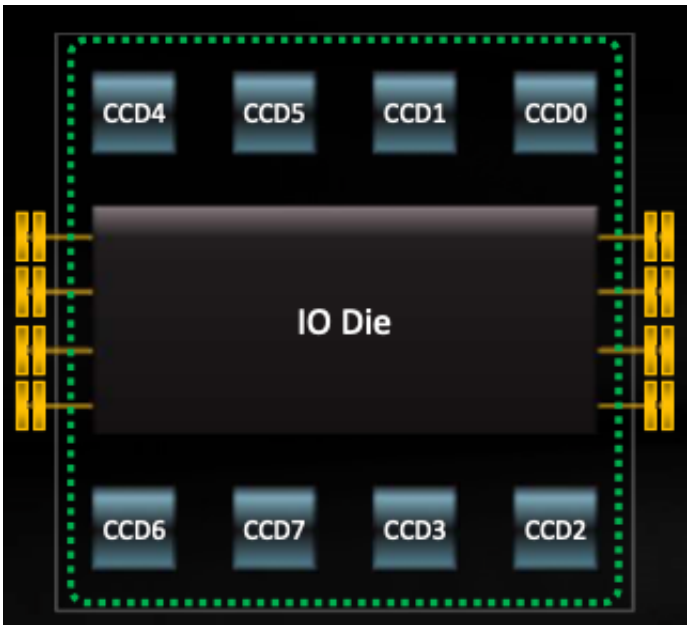
- Two sockets of AMD EPYC 7742 (Rome) processors, with 512 GiB of memory in total
- One NVIDIA A100 GPU with 40 GiB of memory
- Two NICs connected via PCIe 4.0
- 960 GB of local SSD scratch

AMD EPYC 7742 (Rome) Processor

Rome is the codename for AMD's second generation EPYC processor series. A 2.25 GHz AMD Zen2 core in a Rome processor can support SMT, allowing 2 execution threads (hardware threads) to execute simultaneously per core. Each core has its own 32-KiB L1 data and 512-KiB L2 caches. The theoretical peak is 36.0 Gflops per core or 2.3 Tflops per socket.

Four cores share a single 16-MiB L3 cache, and they are grouped as a modular unit called Core-Complex (CCX). A CCD contains two CCXs, as depicted in the diagram below.

The EPYC 7742 processor has eight CCDs for a total of 64 cores, and one IOD per socket, as shown below.



A Rome processor supports 8 memory controllers. Each memory controller supports 2 DIMMs (3200 MHz DDR4), for the maximum memory bandwidth of 409.6 GB/s per socket.

The current NUMA configuration is 1 NUMA node per socket (NPS=1).

NVIDIA Ampere A100 GPU

See the 'GPU-Accelerated Compute Nodes' section above.

File Systems

- Perlmutter scratch
- File systems at NERSC