

Supercomputer Affinity on HPE Systems

Edgar León and Jane Herriman

Cray User Group

Livermore Computing
Lawrence Livermore National Laboratory

May 6, 2024



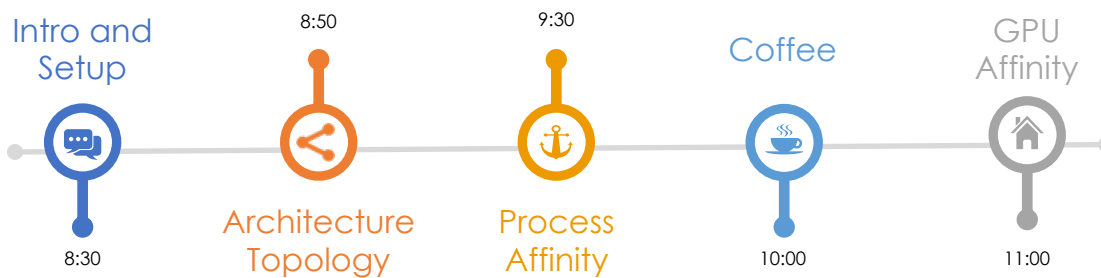
Lawrence Livermore
National Laboratory

Prepared by LLNL under Contract DE-AC52-07NA27344.

LLNL-PRES-862463.

NASA

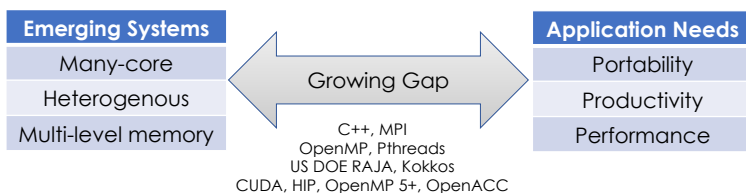
Agenda



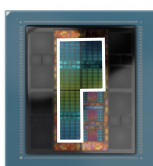
Lawrence Livermore
National Laboratory

NASA

Harder to meet application needs as hardware complexity grows



Our focus:
Mapping applications
to the machine



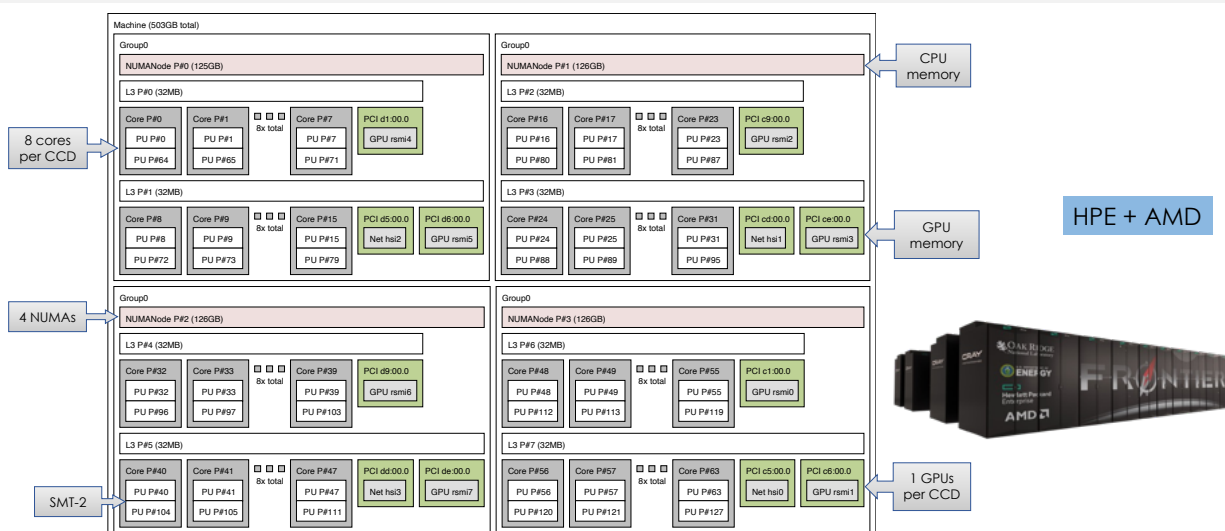
AMD Instinct MI300A APU
Figure courtesy of AMD

AMD EPYC Optimized 3rd Gen EPYC CPU
+ AMD Instinct MI250X GPUs



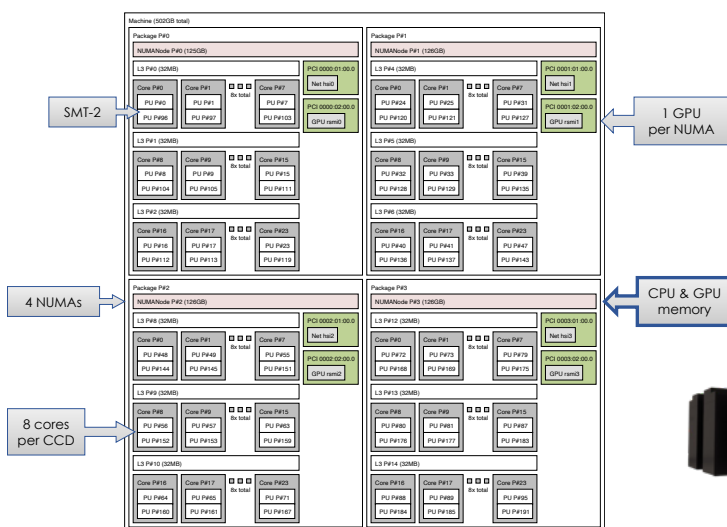
https://docs.olcf.ornl.gov/systems/crusher_quick_start_guide.html

The top super has a complex node architecture!



How about El Capitán?

HPE + AMD

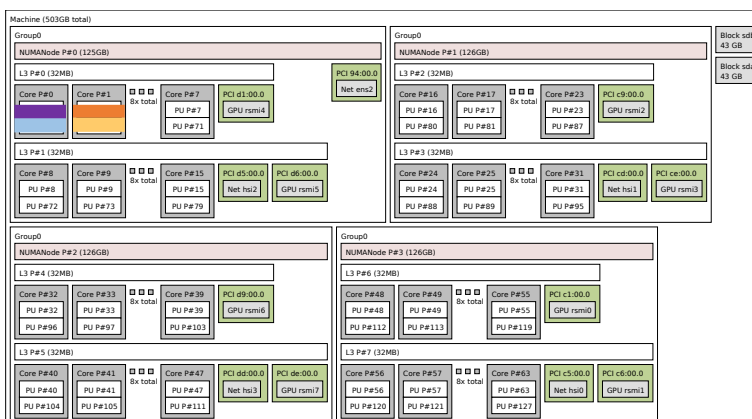


- 4 AMD Instinct MI300A APUs
- MI300A
 - 8x3 Zen4 cores
3 CCDs
 - 1 CDNA3 GPU
6 XCDs
 - 1 NUMA node
8 HBM3 stacks



What could go wrong?

- Multiple threads running on a single core
 - While other cores idle
- Multiple GPU kernels sharing a GPU
 - While other GPUs idle
- MPI tasks launching kernels on non-local GPUs
- Threads accessing memory on a remote NUMA domain
- ...



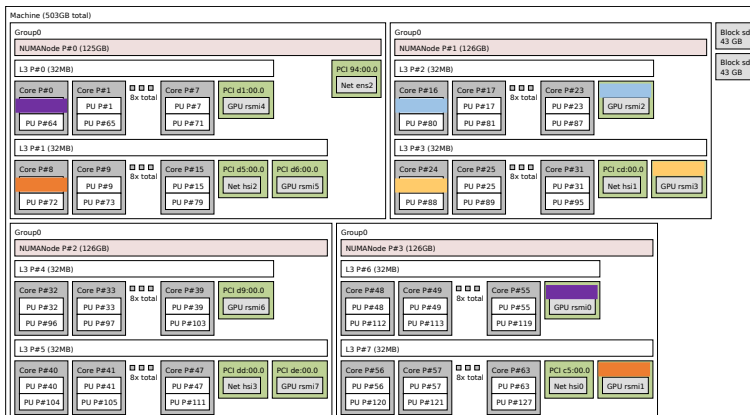
- **Multiple threads running on a single core**

-
- Figure 1 illustrates the node architecture of the HPC system, showing four distinct node configurations (Group0 NUMANode P#0, P#1, P#2, and P#3) and their connection to a network switch and storage.
- Group0 NUMANode P#0 (125GB):** Features a L3 P#0 (32MB) cache, 8x total cores (Core P#1 to P#8), 8x total processors (PU P#1 to PU P#8), and 8x total GPUs (GPU rsmi0 to GPU rsmi7). It is connected to a network switch (Net swi) and a storage unit (Block sd 43 GB).
 - Group0 NUMANode P#1 (126GB):** Features a L3 P#1 (32MB) cache, 8x total cores (Core P#8 to P#15), 8x total processors (PU P#8 to PU P#15), and 8x total GPUs (GPU rsmi8 to GPU rsmi15). It is connected to a network switch (Net swi) and a storage unit (Block sd 43 GB).
 - Group0 NUMANode P#2 (126GB):** Features a L3 P#2 (32MB) cache, 8x total cores (Core P#17 to P#24), 8x total processors (PU P#17 to PU P#24), and 8x total GPUs (GPU rsmi16 to GPU rsmi23). It is connected to a network switch (Net swi) and a storage unit (Block sd 43 GB).
 - Group0 NUMANode P#3 (126GB):** Features a L3 P#3 (32MB) cache, 8x total cores (Core P#24 to P#31), 8x total processors (PU P#24 to PU P#31), and 8x total GPUs (GPU rsmi24 to GPU rsmi31). It is connected to a network switch (Net swi) and a storage unit (Block sd 43 GB).
- The diagram also shows a central network switch (Net swi) and a storage unit (Block sd 43 GB) connected to all nodes.

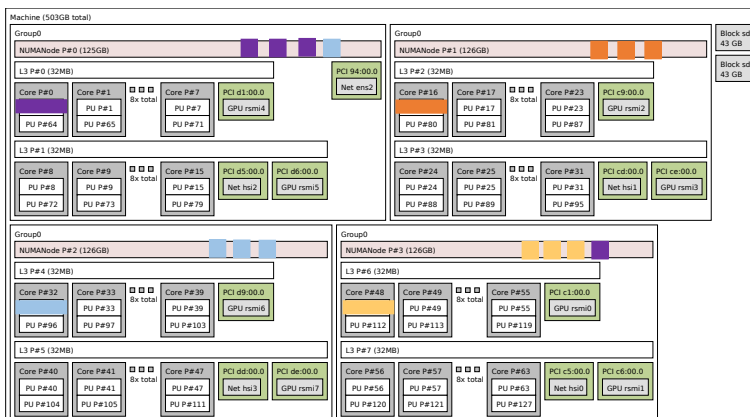
- Multiple threads running on a single core

- [illegible]

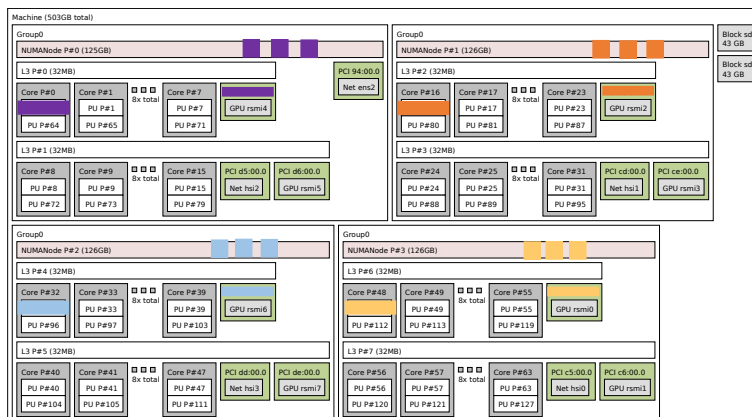
- Multiple threads running on a single core
 - While other cores idle
- Multiple GPU kernels sharing a GPU
 - While other GPUs idle
- **MPI tasks launching kernels on non-local GPUs**
- Threads accessing memory on a remote NUMA domain
- ...



- Multiple threads running on a single core
 - While other cores idle
- Multiple GPU kernels sharing a GPU
 - While other GPUs idle
- MPI tasks launching kernels on non-local GPUs
- **Threads accessing memory on a remote NUMA domain**
- ...



- Distribute workers based on the memory system
 - Maximize resources per worker
 - Memory and cache
 - Floating point units (SMT)
- Avoid remote memory accesses
 - Single NUMA per worker, when possible
- Leverage CPU-GPU-Memory locality
- Leverage SMT support
 - Thread specialization



<https://github.com/LLNL/mpibind/tree/master/tutorials/cug24>

