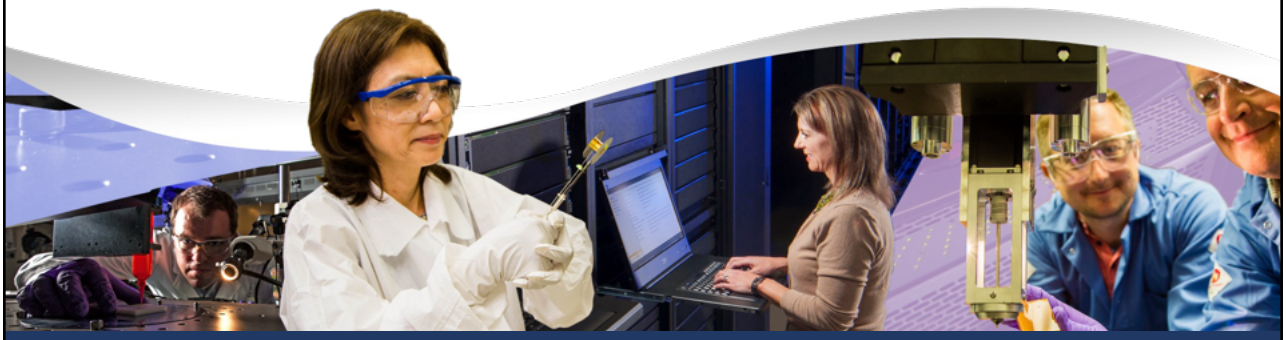
Supercomputer Affinity on HPE Systems

Edgar León and Jane Herriman

Livermore Computing
Lawrence Livermore National Laboratory

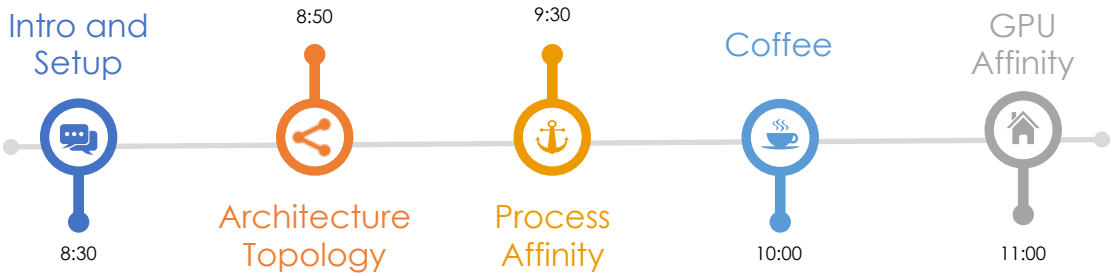
Cray User Group

May 6, 2024



Lawrence Livermore National Laboratory Prepared by LLNL under Contract DE-AC52-07NA27344. LLNL-PRES-862463. NNSA

Agenda



Lawrence Livermore National Laboratory NNSA

Harder to meet application needs as hardware complexity grows

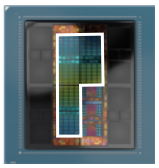
Emerging Systems
Many-core
Heterogenous
Multi-level memory

Growing Gap

C++, MPI
OpenMP, Pthreads
US DOE RAJA, Kokkos
CUDA, HIP, OpenMP 5+, OpenACC

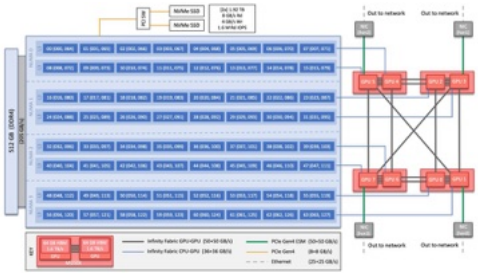
Application Needs
Portability
Productivity
Performance

Our focus:
Mapping applications to the machine



AMD Instinct MI300A APU
Figure courtesy of AMD

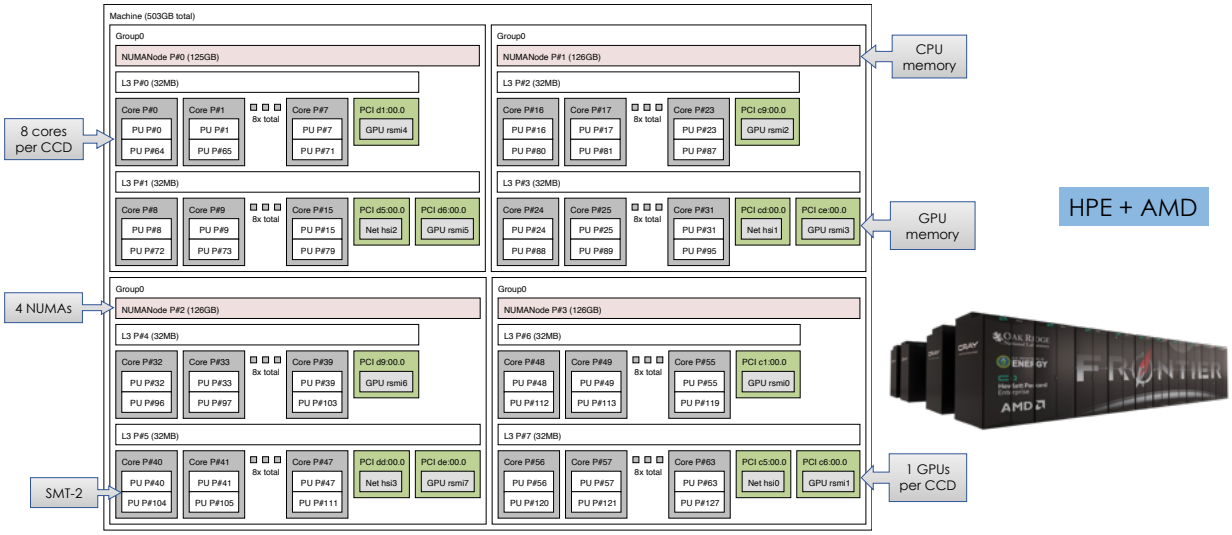
AMD EPYC Optimized 3rd Gen EPYC CPU
+ AMD Instinct MI250X GPUs



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

Lawrence Livermore National Laboratory NNSA

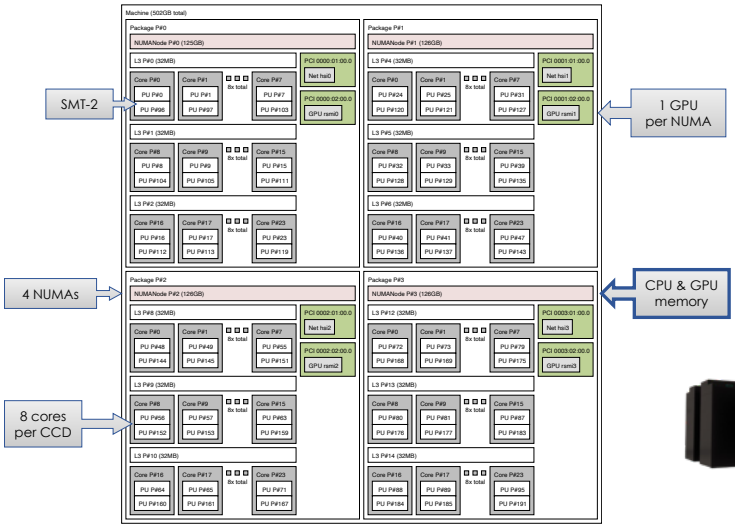
The top super has a complex node architecture!



Lawrence Livermore National Laboratory NNSA

How about El Capitán?

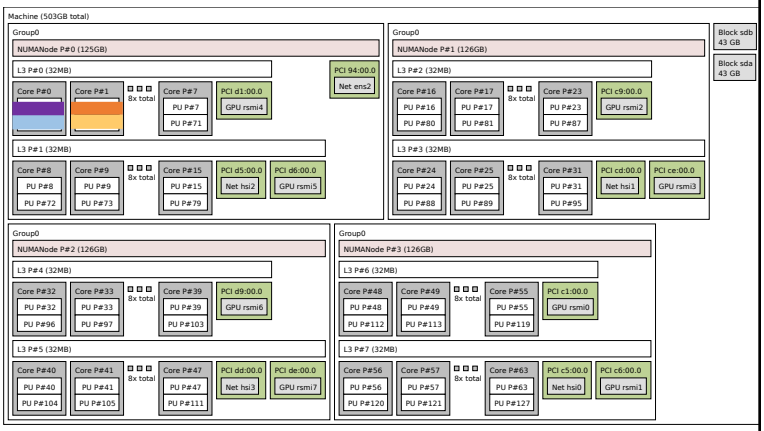
HPE + AMD



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

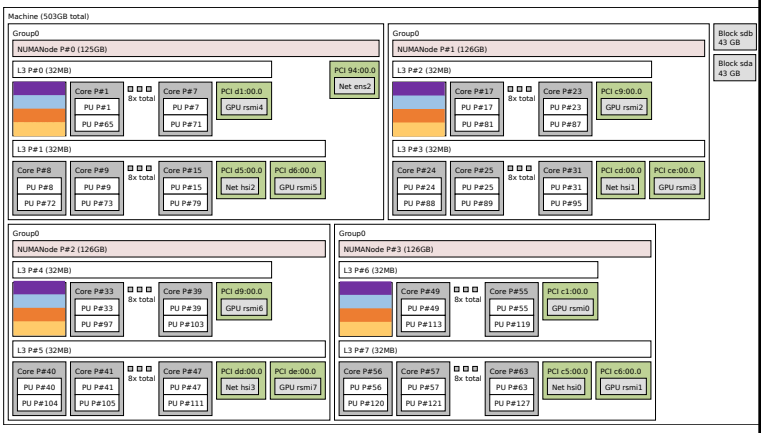
What could go wrong?

- **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
- …



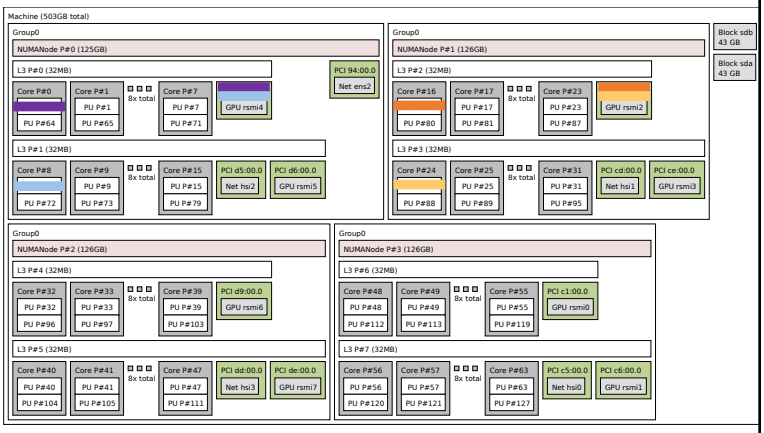
What could go wrong?

- **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
- …



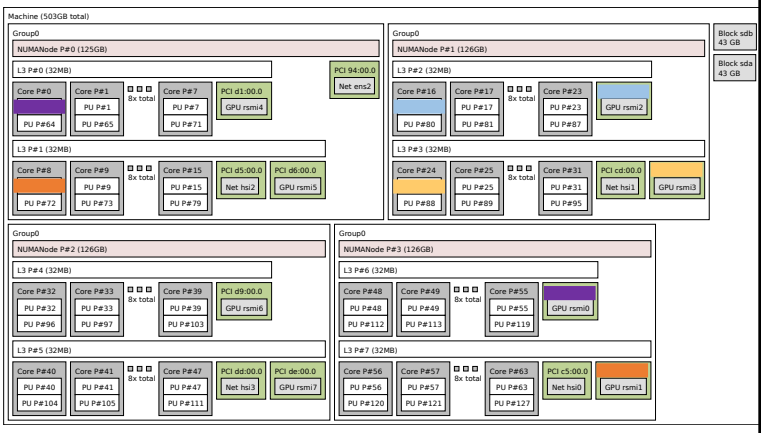
What could go wrong?

- 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
- …



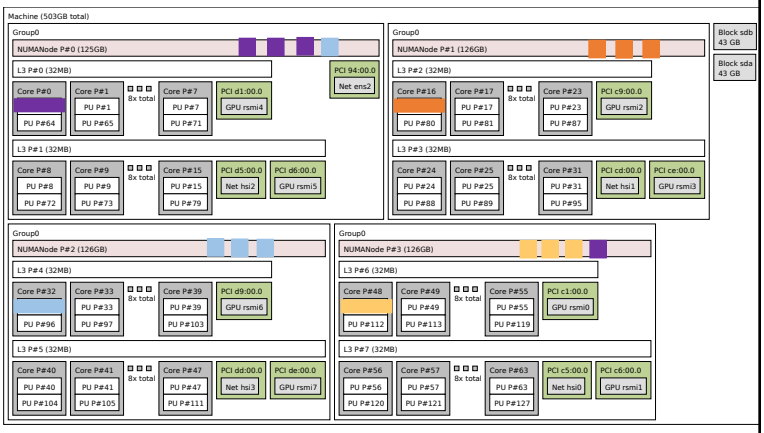
What could go wrong?

- 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
- …



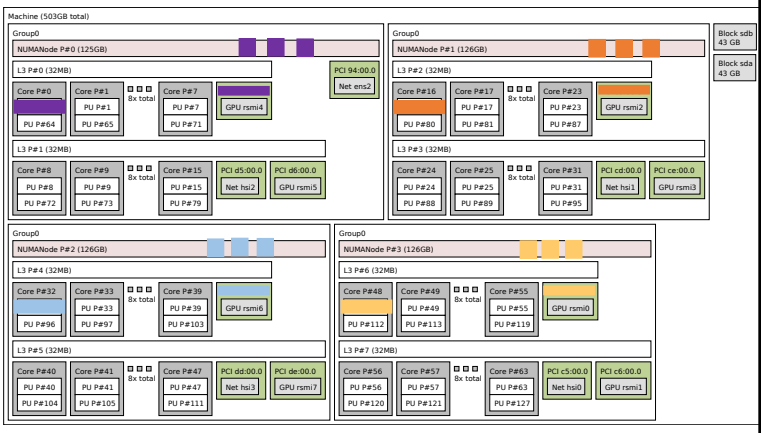
What could go wrong?

- 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**
- …



General guidelines for an efficient mapping

- 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



You will learn how to control affinity on heterogeneous systems

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