Supercomputer Affinity on HPE Systems

Edgar A. León and Jane E. Herriman

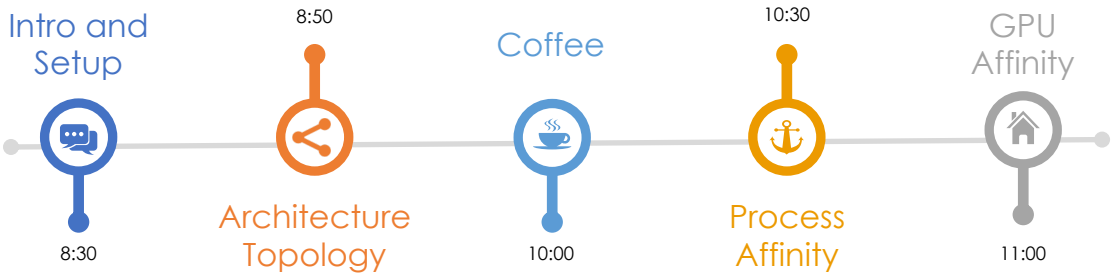
Livermore Computing
Lawrence Livermore National Laboratory

Cray User Group

May 8, 2023

Lawrence Livermore National Laboratory Prepared by LLNL under Contract DE-AC52-07NA27344. LLNL-PRES-847296.

Agenda



Lawrence Livermore National Laboratory

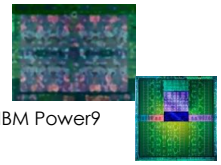
Harder to meet application needs as hardware complexity grows

Emerging Systems
Many-core
Heterogeneity
Multi-level memory

Growing Gap

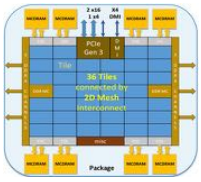
Application Needs
Portability
Productivity
Performance

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

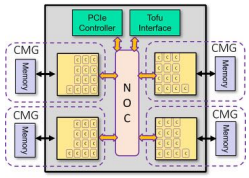


IBM Power9

NVIDIA Volta



Intel KNL

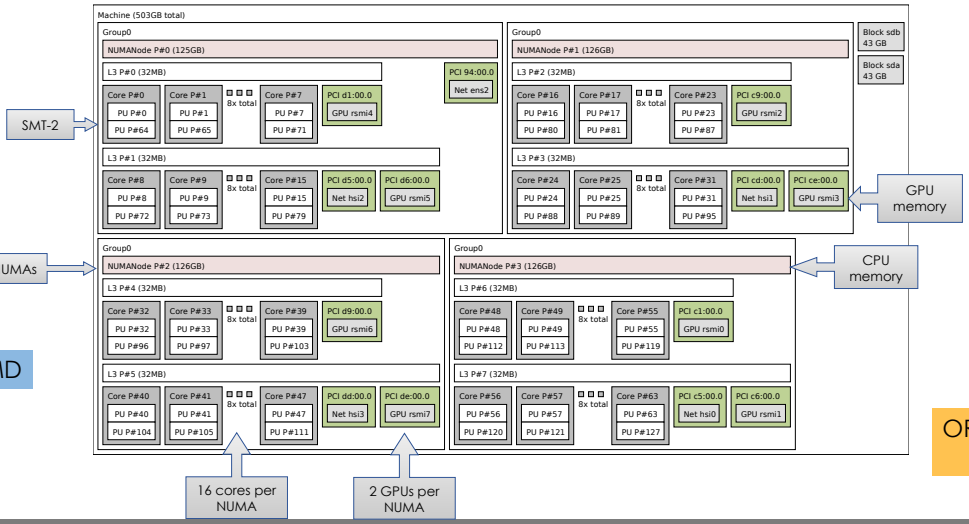


Fujitsu Armv8.2-A

Our focus:
Mapping applications to the machine

Lawrence Livermore National Laboratory

The top super has a complex architecture!



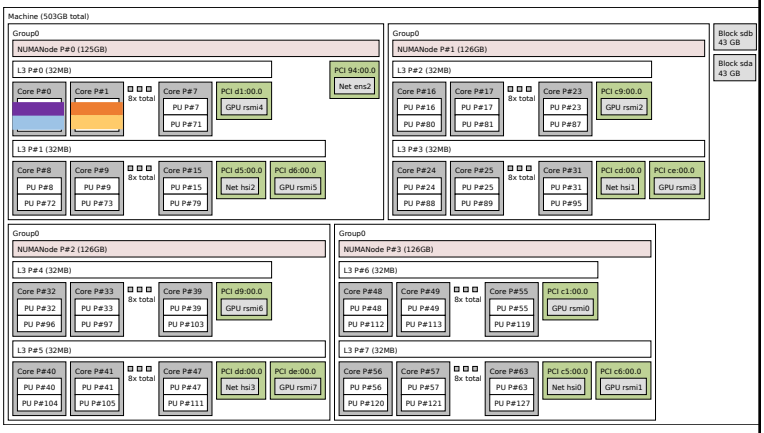
HPE/AMD

ORNL Frontier
LLNL EAS

Lawrence Livermore National Laboratory

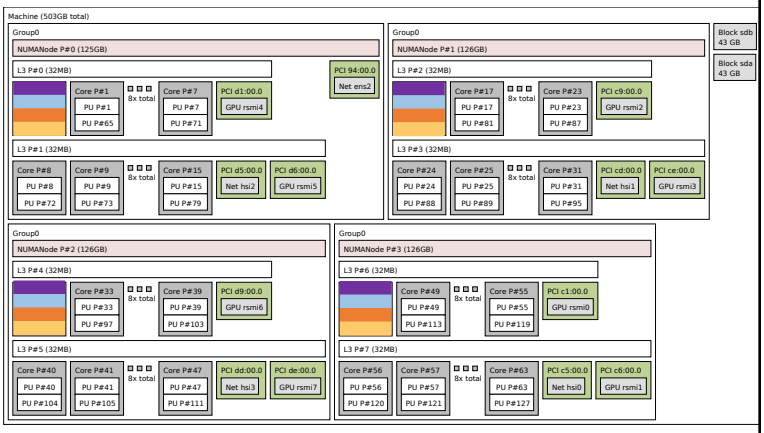
What could go wrong?

- **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?

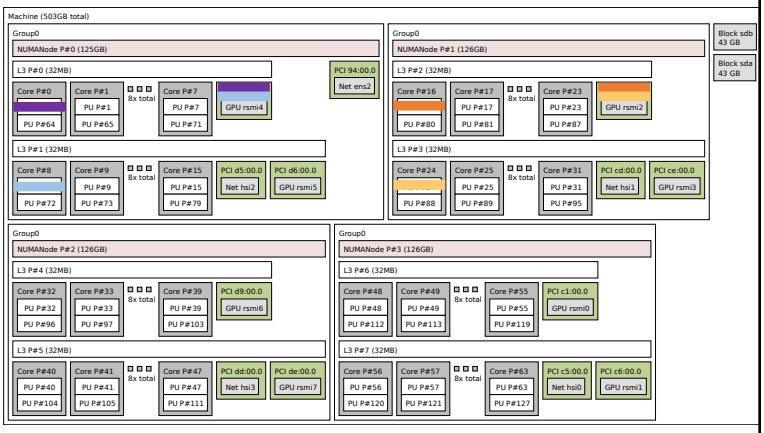
- **Multiple threads running on a single core**
  - **While other cores idle**
- Multiple GPU kernels sharing a GPU
  - While other GPUs idle
- MPI tasks launching kernels on non-local GPUs
- Threads accessing memory on a remote NUMA domain
- …



Lawrence Livermore National Laboratory NNSA

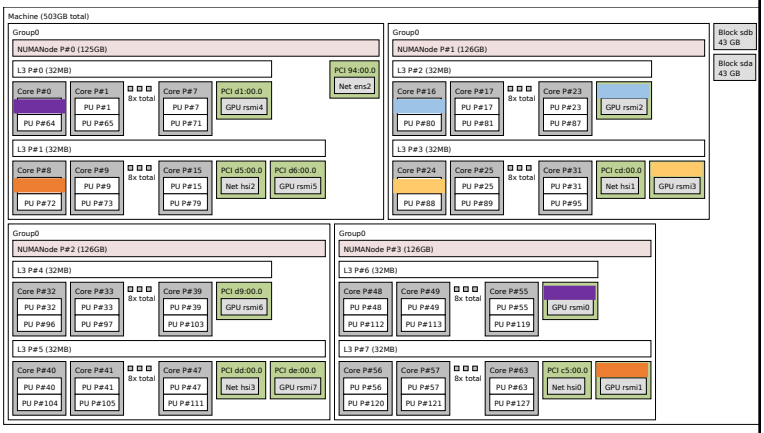
What could go wrong?

- 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
- …



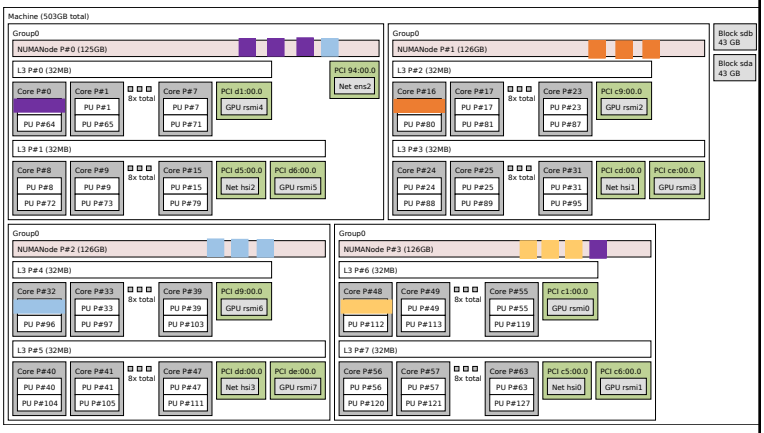
What could go wrong?

- 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
- …



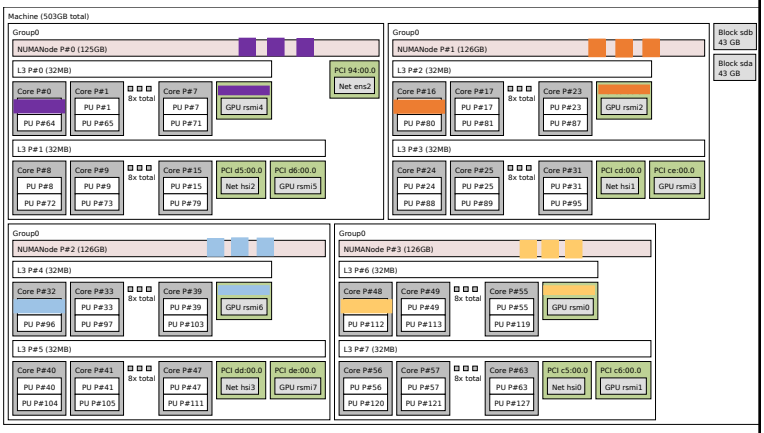
What could go wrong?

- 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**
- …



Leveraging *locality* is key to a good mapping

- General guidelines
  - Spread out to maximize resources
  - Single NUMA per worker, when possible
  - Maximize memory / cache
  - Leverage CPU locality
  - Leverage GPU locality
  - Leverage SMT support



You will learn how to control affinity on heterogeneous systems

https://github.com/LLNL/mpibind/tree/master/tutorials/cug23