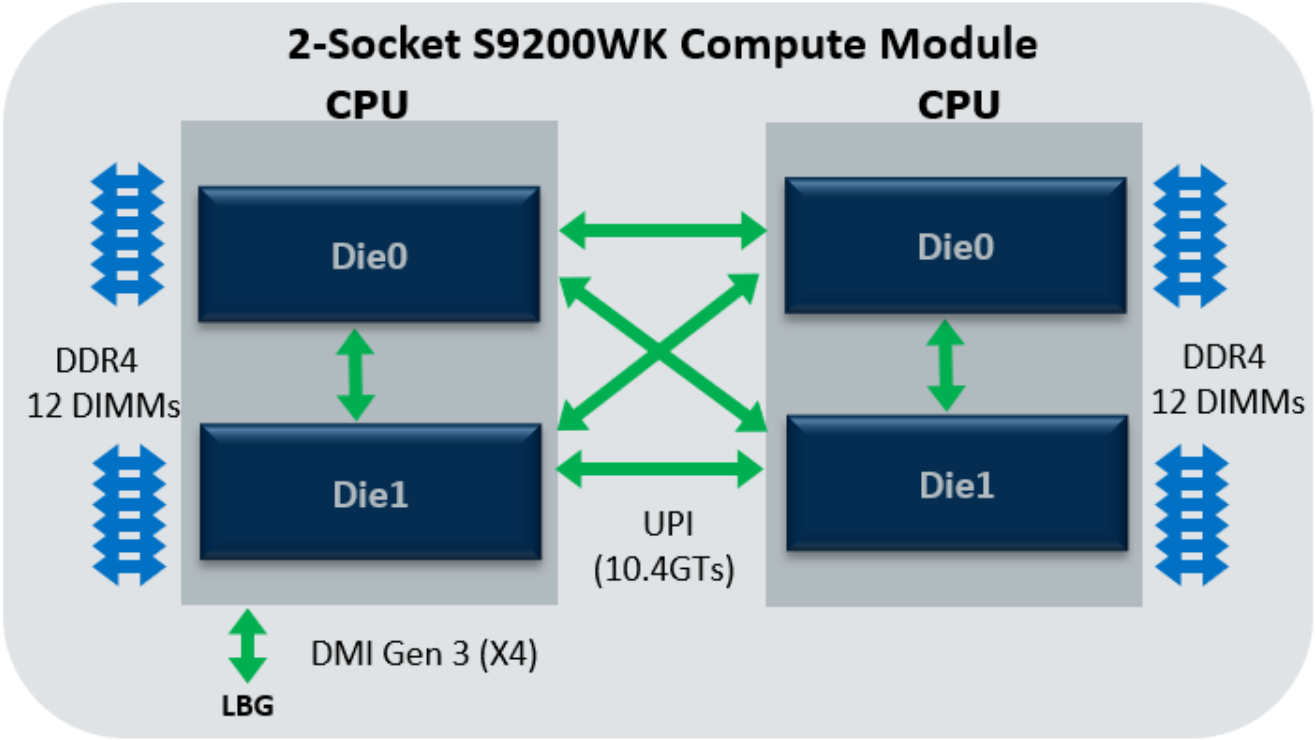
Binding with SLURM wrt OPA

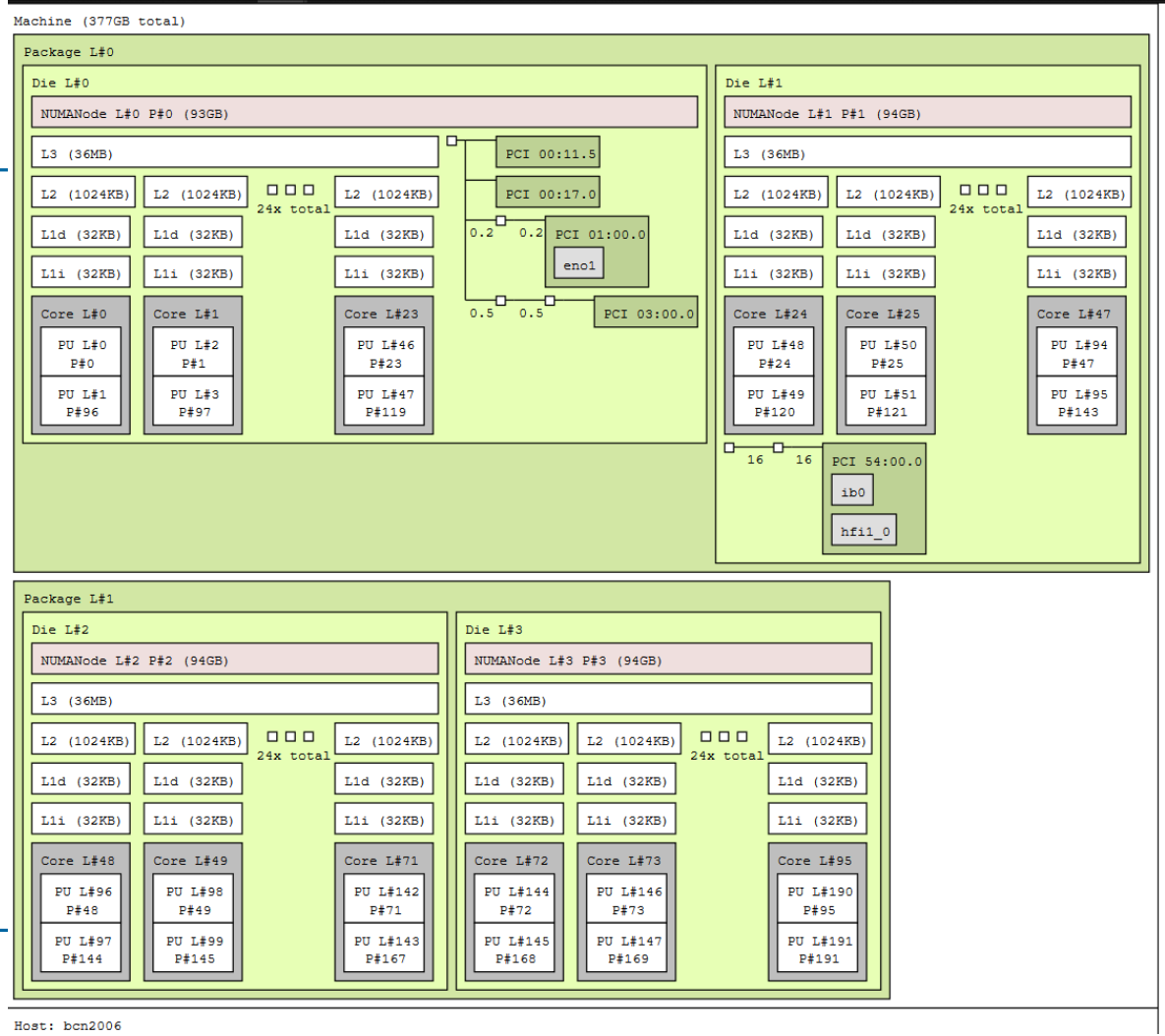
2020

Atos

CLX-AP Architecture



hwloc-ls --of svg



MPI communication

- Ranks on Die 1 have direct access to the Omni-path Adapter
- MPI communication for the ranks on other dies also passes the UPI bridge.
  - This increases latency for MPI communication
  - MPI traffic over the UPI bridge competes with traffic between NUMA domains.

- Ranks with more MPI communication than other ranks should preferentially be placed on the cores of Die/NUMAnode 1.

Does my application benefit?

- Some applications concentrate MPI communication on rank 0
  - MPI_Bcast from rank 0
  - MPI_Reduce to rank 0
- The first ranks on each node have increased MPI communication
  - PALMfft uses Alltoall between the first ranks on all nodes
- Put the first ranks somewhere on rank 24-47 (NUMAnode1)

How to change binding of the first rank to NUMAnode 1?

- Using CPU map (only the first hyperthread)

```
map=$(seq -s, 24 95),$(seq -s, 0 23)
srun --cpu-bind=map_cpu:$map
```

- Using CPU masks (includes both hyperthreads)

```
module load hwloc/2.1.0
mask=$(for n in {24..95} {0..23};do echo -n $(hwloc-calc --taskset core:$n),;done)
srun --cpu-bind=mask_cpu:$mask
```

What if I use Intel‘s mpirun/mpiexec?

- By default, it will put the first rank of each node on core 24
  - No further action required
  - Make sure that SLURM binding is disabled.

Impact on PALMfft

	Benchmark time (s)
First rank on core 0	181.4
First rank on core 24	170.1

Improvement is 6.2%

Thanks for your attention

john.donners@atos.net

Atos