Improving Support of MPI+OpenMP Applications

Geoffroy Vallée, David Bernhold
Computer Science Research Group, Computer Science and Mathematics Division
Oak Ridge National Laboratory

Overview

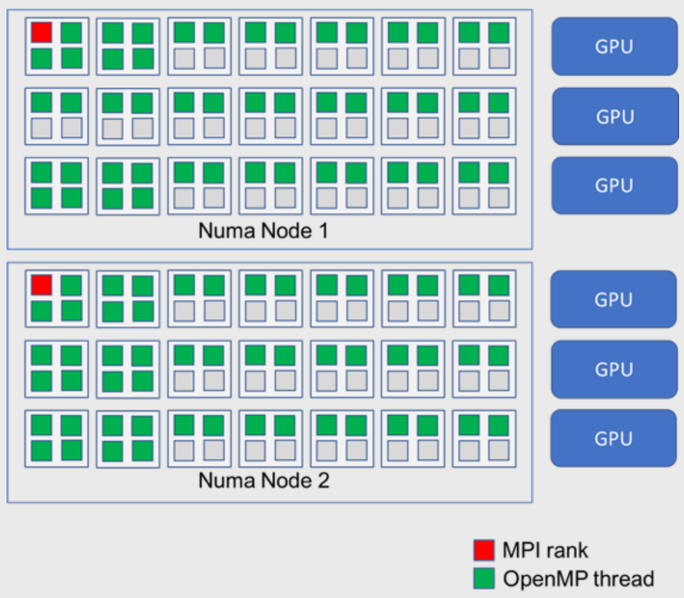
- MPI+OpenMP is a privileged option for implementing hybrid applications in the context of the U.S. DOE Exascale Computing Project (ECP)
- MPI and OpenMP implementations are not designed to work together
- It is difficult for users to precisely control the deployment of MPI+OpenMP applications
  - By default, the OpenMP runtime will assume that all local resources can be used
  - The definition of OpenMP places limits how an hybrid application can be deployed
  - Impossible to support complex "layouts"

Key Challenges

- How can users describe complex MPI+OpenMP *layouts* to run on compute nodes?
- Can users change the layout of an application at runtime?
  - *OpenMP limitations*: places are set during initialization and cannot be changed
  - *MPI limitations*: the number and placement of MPI ranks can be changed; however, how can we set a new OpenMP layout on a per-MPI rank basis

Layout – Example

- Goal: applications take full benefit of the underlying hardware
- Example based on the architecture of OLCF Summit's compute nodes:



Layout Description

- Definition of the layout illustrated above
- Passed through the mpirun command line:
  mpirun --mca rmaps explicit --mca rmaps_explicit_layout "[MPI, ….]" --np 128 myapp.exe

```
[MPI,-,Cores,[[0, -, -, -, -, -, -], [-, -, -, -, -, -, -], [-, -, -, -, -, -, -], [1, -, -, -, -, -, -],
      [-, -, -, -, -, -, -], [-, -, -, -, -, -, -]]
[OpenMP,MPI-0,HT,[[0,1,2,3],[4,5,6,7],[8,9,--],[10,11,-,-],[12,13,-,-],
      [14,15,-,-],[16,17,-,-], [18,19,-,-],[20,21,-,-],[22,23,-,-],[24,25,-,-],
      [26,27,-,-],[28,29,-,-],[30,31,-,-],[32,33,34,35],[36,37,38,39],
      [40,41,-,-],[42,43,-,-],[44,45,-,-],[46,47,-,-],[48,49,-,-]]
[OpenMP,MPI-1,HT, [[0,1,2,3],[4,5,6,7],[8,9,-,-],[10,11,-,-],[12,13,-,-],
      [14,15,-,-],[16,17,-,-],[18,19,21,22],[23,24,25,26],[27,28,-,-],
      [29,30,-,-],[31,32,-,-],[33,34,-,-],[35,36,-,-],[37,38,39,40],[41,42,43,44],
      [45,46,-,-],[47,48,-,-],[49,50,-,-],[51,52,-,-],[53,54,-,-]]
```

Implementation

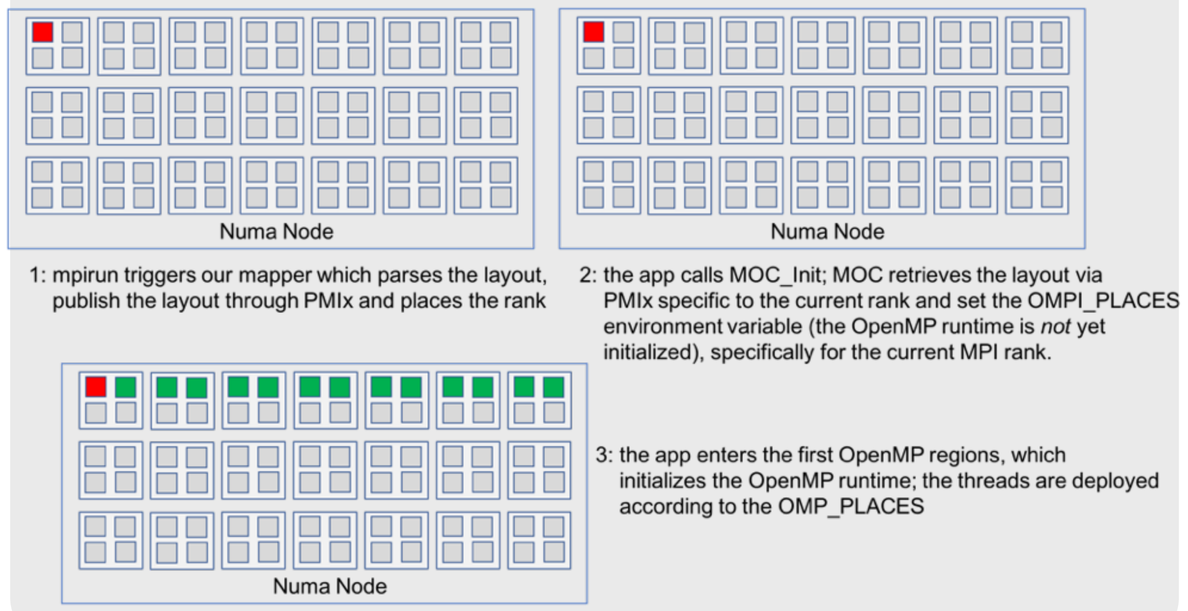
- A new Open MPI mapper for MPI rank placement: *explicit* module for the *rmap* framework
- A helper library: MPI OpenMP Coordination library (MOC)
  - Set OpenMP places for static layouts
  - Avoid requiring to modify the OpenMP runtime

```
int main (int argc, char **argv) {
(…)
   MPI_Init (&argc, &argv);
   MOC_Init (&argc, &argv);
   (…)
   MOC_Finalize ();
   MPI_Finalize ();
}
```

```
$ mpirun –mca rmaps explicit –mca_rmaps_explicit_layout "[MPI, …]" (…) -np 128 app.exe
```

Execution

- Example focusing on the context of a single rank



OLCF Oak Ridge Leadership Computing Facility
ECP Exascale Computing Project
Oak Ridge National Laboratory