MPI Forum 2018/06/13
Austin, TX

HWT-WG update

Three Directions

- The *implicit* access to the topology information
  - The topology can be accessed through MPI abstractions
- The *explicit* access to the topology information
  - The topology description can be accessed by the user directly
- Mapping and binding of MPI processes
  - Borderline point

Implicit access

- Current proposal based on hierarchical communicators
- Presented at the Forum in Portland (2017)
- Prototype implementation available: Hsplit
  - External library
  - hwloc/netloc-based
  - Positive feedback from users: CERFACS, Météo France

Hsplit Interface

- Comms creation functions:

```
-  MPI_Comm_split_type(MPI_Comm oldcomm, int split_type, int key,
                       MPI_Info info, MPI_Comm *newcomm)
```

With a new split_type value: MPI_COMM_TYPE_PHYSICAL_TOPOLOGY

```
- MPI_Comm_hsplit_with_roots(MPI_Comm oldcomm, MPI_Info info,
                             MPI_Comm *newcomm, MPI_Comm *rootscomm)

- MPI_Get_htopo_neighbours(MPI_Comm oldcomm, int hops, int metric
                           MPI_Comm *newcomm)
```

- Query Functions

```
- MPI_Comm_get_min_hlevel(MPI_Comm comm,
                          int      nranks,
                          int     *ranks,
                          char   **type)
- MPI_Comm_get_hlevel_info(MPI_Comm comm,
                           int     *num_comms,
                           int     *index,
                           Char   **type)
```

Explicit access

- Determination of processes coordinates and neighborhood
- Exemple: Fujitsu’s extensions

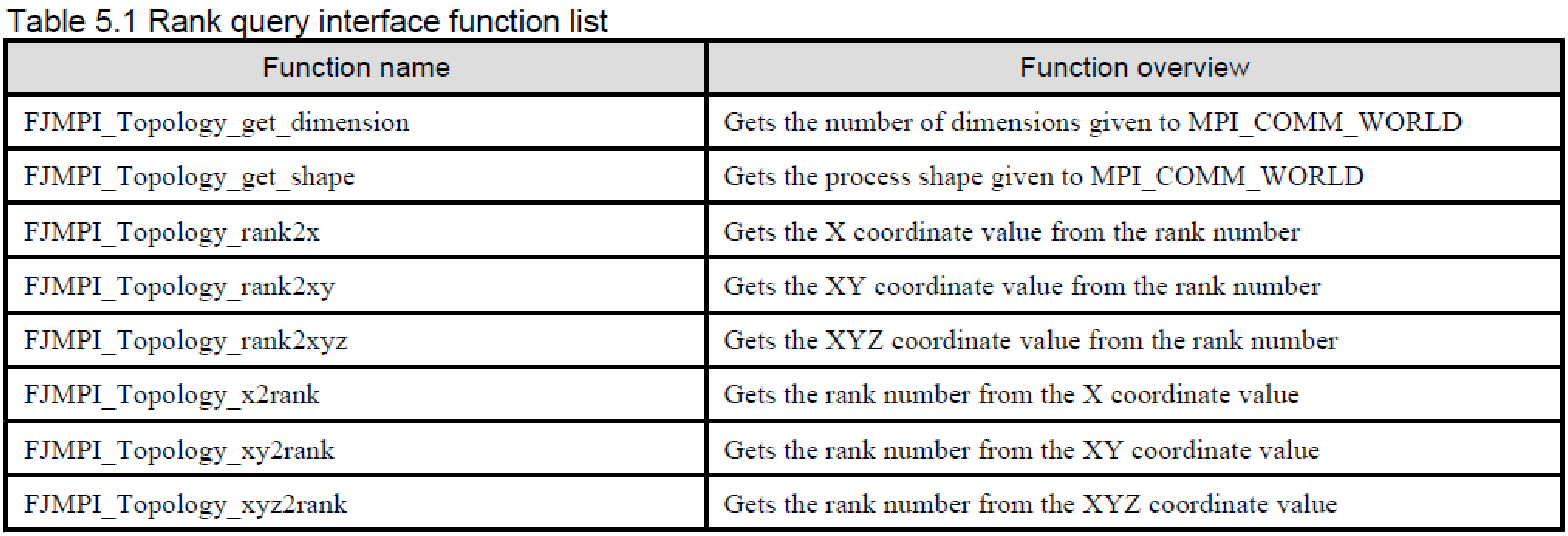

Table 5.1 Rank query interface function list

Function name	Function overview
FJMPI_Topology_get_dimension	Gets the number of dimensions given to MPI_COMM_WORLD
FJMPI_Topology_get_shape	Gets the process shape given to MPI_COMM_WORLD
FJMPI_Topology_rank2x	Gets the X coordinate value from the rank number
FJMPI_Topology_rank2xy	Gets the XY coordinate value from the rank number
FJMPI_Topology_rank2xyz	Gets the XYZ coordinate value from the rank number
FJMPI_Topology_x2rank	Gets the rank number from the X coordinate value
FJMPI_Topology_xy2rank	Gets the rank number from the XY coordinate value
FJMPI_Topology_xyz2rank	Gets the rank number from the XYZ coordinate value

Mapping/binding

- Difficult issue
  - “Outside the scope of the standard”
  - Involves RJMS, process managers, MPI applications
    - At what level (e.g MPI_Bind)?
    - Identify the possible interactions
  - Binding is easy, mapping not so
    - Even worse in hybrid dynamic cases
- Not very user-friendly
  - Changes from one implementation version to the other
  - Not portable
- Standardize mpiexec parameters?