HW Topology WG

05/04/18

Working Group Organization

- General status and current roadmap
- Explicit query interface
- Unifying launch method and binding
- User-Defined Software Stack Scenario

Working Group Organization

- Three timeslots :
  - 1 EU / US (5PM GMT +2) **<— Yes spring time sorry**
  - 1 EU / JP (9PM GMT +2)
  - 1 JP / US (1AM GMT +2)
- **Is someone willing to chair the US / JP call ?**

Current Topics

- **Explicit topological query** with for example MPI-T. This track aims at providing a way to describe the actual topology to the end-user;
- **Implicit topological query** with the communicator proposal. In this second proposal, the topology is more abstract (not directly exposed) but it can still be manipulated with compact primitives. It may be simpler and more generalizable than the explicit approach;
- **Mapping / Binding / Launch** the way mpirun arguments differ have to be questioned and the work-group asks the questions of the possibility of a common interface for binding/launching.

Current Priorities

- Get as much input as possible
- Present prototype implementations
- Prepare plausible interfaces for the standard

Next Steps

- Setup « technical » tickets matching each topics (explicit, implicit, launch);
- Form groups of interest and critalize them to bring reasonable proposals to the Forum and get feedback at this point.

Explicit Query

- Use MPI-T to bring topological informations
- Do not add any new functions to the standard (solely rely on MPI-T capabilities)
- How : by attaching informations to communicator objects

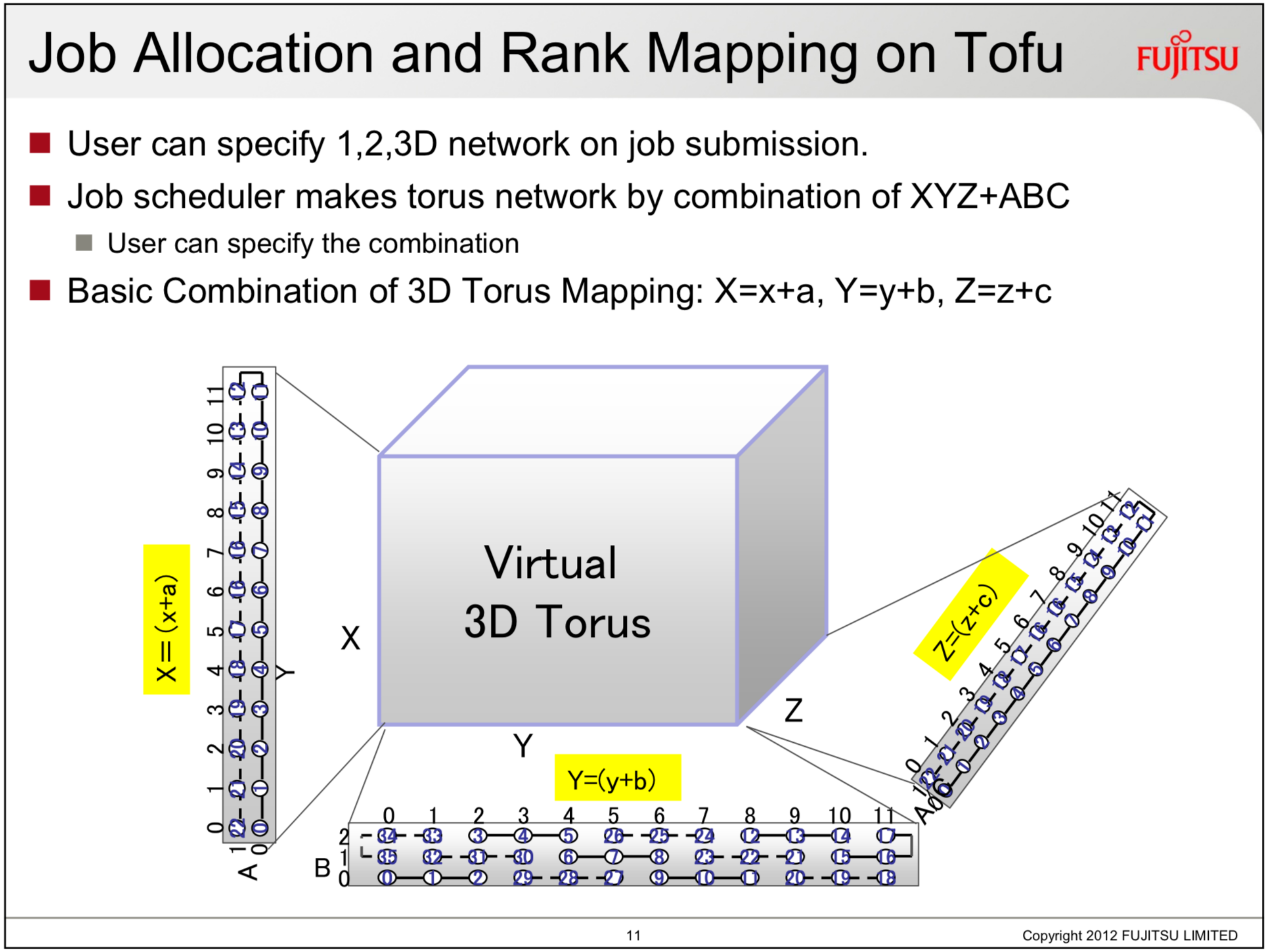
What is Job Topology ?

We have several referential to handle :

- Machine has a topology (network, intra-node)
- Allocation has a topology (dedicated cores)
- Computation has a topology (communication graph)

In the end, we want the computation to match as closely as possible machine's topology within allocation's constraints.

What is Job Topology ?



Job Allocation and Rank Mapping on Tofu

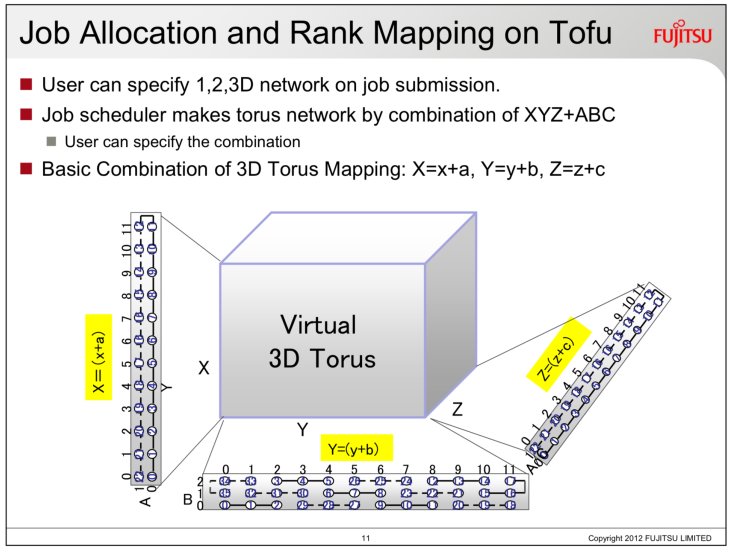
FUJITSU

- User can specify 1,2,3D network on job submission.
- Job scheduler makes torus network by combination of XYZ+ABC
  - User can specify the combination
- Basic Combination of 3D Torus Mapping: X=x+a, Y=y+b, Z=z+c

What is Job Topology ?

- Machine has a topology (network, intra-node)
- Allocation has **FIXED** topology (1D, 2D, 3D)
- Computation has a topology

- Can MPI ranks be aware of machine constraints (distance ?)
- Is it possible to query communicator topology ?
- Virtual topology VS physical topology



Job Allocation and Rank Mapping on Tofu

FUJITSU

- User can specify 1,2,3D network on job submission.
- Job scheduler makes torus network by combination of XYZ+ABC
  - User can specify the combination
- Basic Combination of 3D Torus Mapping: X=x+a, Y=y+b, Z=z+c

11

Copyright 2012 FUJITSU LIMITED

Distance Metric

```
0   1   2   3
4   5   6   7
8   9   10  11
```

1D rank

```
5   4   3   2
1   0   1   2
3   4   5   6
```

1D rank

```
4 - 5 - 6
```

2-neighbor

```
(0,0) (0,1) (0,2) (0,3)
(1,0) (1,1) (1,2) (1,3)
(2,0) (2,1) (2,2) (2,3)
(3,0) (3,1) (3,2) (3,3)
```

2D rank

```
1 0 1 2
2 1 2 3
3 2 3 4
```

2D rank

```
    1
    |
4 - 5 - 6
    |
    9
```

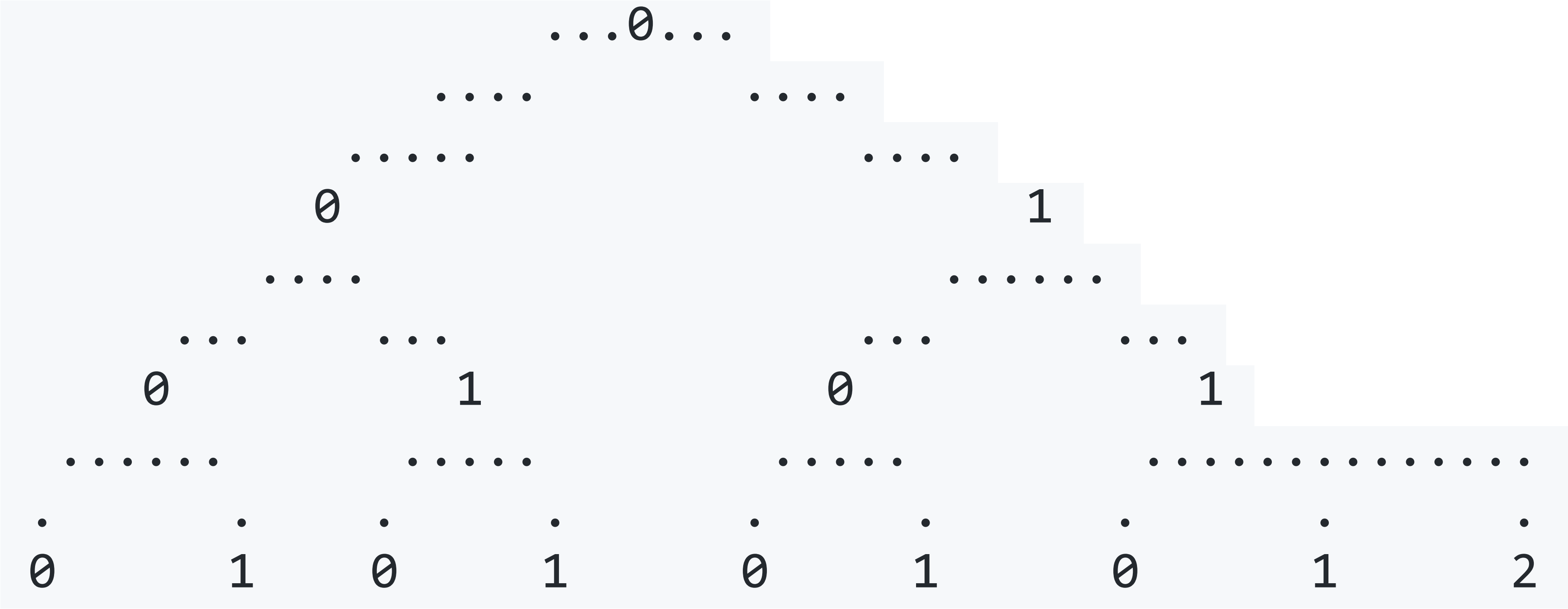
4-neighbor

Distance Metric

How do we expose it ?

- Direct query in MPI-T means N^2 matrix
  - Can be generated on the fly
- Can we get function pointers from MPI-T (edgy ?)
- Or a new MPI function
- We can also define a rank object

More on Coordinates



```
                  ...0...
              ....       ....
           .....             ....
          0                        1
        ....                    ......
     ...    ...              ...      ...
    0          1            0            1
 ......      .....        .....      ..............
.      .    .     .      .     .      .      .      .
0      1    0     1      0     1      0      1      2
```

- Should a given core in the machine have fixed « topological » coordinates ?

- These coordinates would be the rank which would then be mapped later on in virtual topologies.

Launch Method

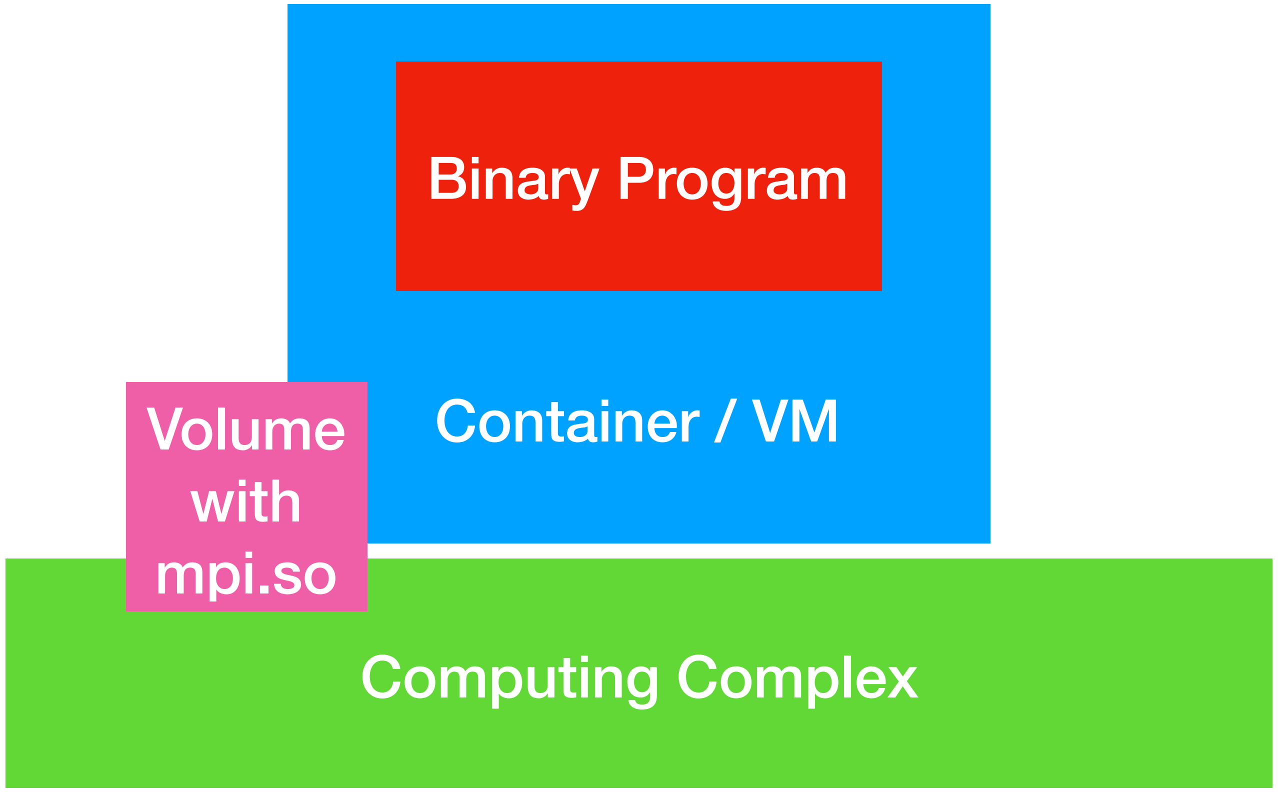
- Is it possible to replace mpirun with a portable launch algorithm « abstracting » the topology
- Can we imagine portable MPI launch scripts ?
- Can we get rid of « -np » which is hard to parse ?
- What does -c means and who decides for pinning ?
  - Slurm ? The PMI ? Mpirun ?
  - In the case of model-mixing (MPI+Y) who decides ?

It is a very HARD problem !

User-Defined Software Stacks

- Can we launch containers indifferently from the underlying MPI layer ?

- **MPI has no ABI**
- **How can we gess the underlying PMI ?**
  -> PMI1, PMI2, PMIX, PMIXvX, …
- **What is the minimal set of functions to launch an MPI program in a nested context ?**



Binary Program

Container / VM

Volume with mpi.so

Computing Complex

PMI_Init
PMI_Get_rank
PMI_Get_size
PMI_KVS_Get_value_length_max
PMI_KVS_Get_key_length_max
PMI_Finalize
PMI_Barrier PMI_KVS_Get
PMI_KVS_Put
PMI_KVS_Commit

Sufficient for MPICH 3.0.4 to MPICH 3.2
Not anymore in master (switch to PMIX)

Is it discussed somewhere ? Can it be in the WG ?