FOSDEM 2022

Exascale PMI on non-exascale Slurm cluster

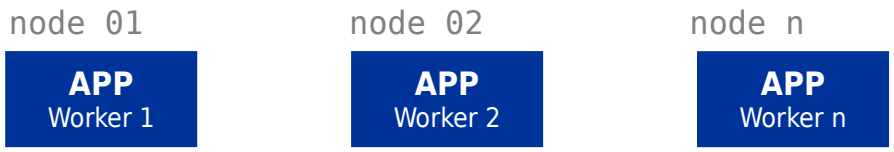
Alex Domingo

WHO AM I?

Hi! I'm Alex (github: @lexming)

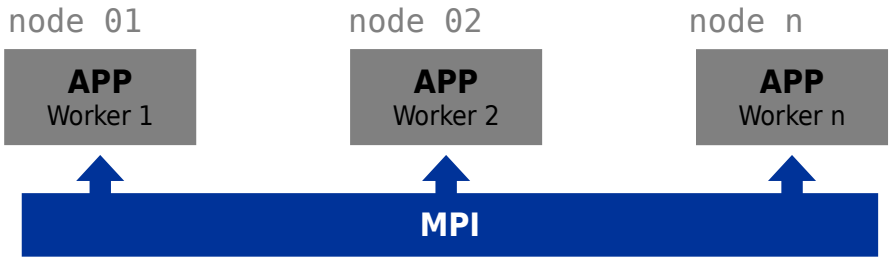
- Background
  - PhD in computational chemistry
  - User of Linux and FOSS in general since mid 2000's
- **Present time: HPC team of VUB** since 2019 (hpc.vub.be)
  - Horizontal team: Linux sysadmin, software optimization, direct user support
  - Maintainer of EasyBuild (easybuild.io): open source software build and installation framework for HPC
- Disclaimer: not a developer (except for EasyBuild, a little)
- Free time: eats chocolate and plays with raspberry pis at home

DISTRIBUTED PARALLEL COMPUTING



- Application executing in parallel on multiple CPU cores
- Cores can be located in multiple nodes

DISTRIBUTED PARALLEL COMPUTING



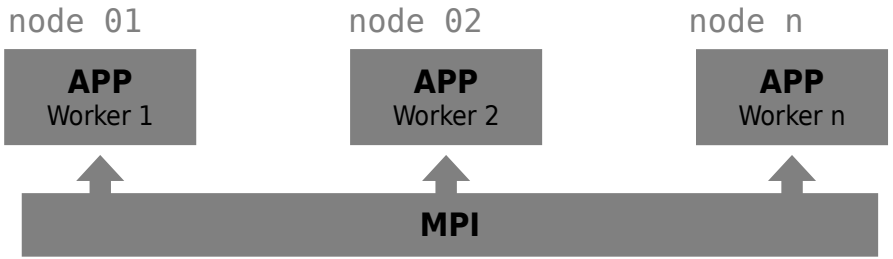
Message Passing Interface (MPI)

- Communication between application workers

More information on MPI

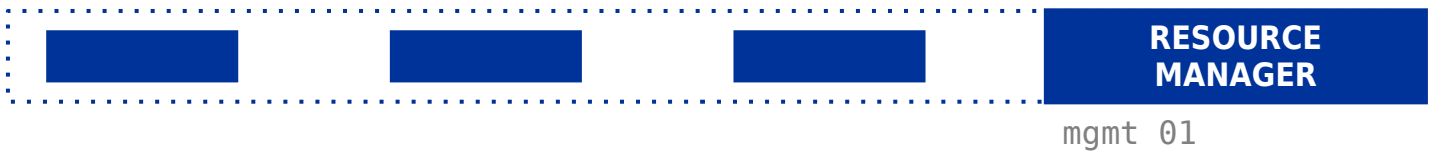
- EasyBuild Tech Talk: **The ABCs of Open MPI**
  https://github.com/easybuilders/easybuild/wiki/EasyBuild-Tech-Talks-I%3A-Open-MPI

DISTRIBUTED PARALLEL COMPUTING

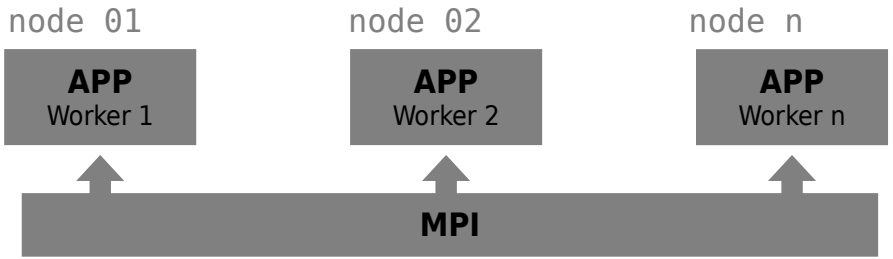


Resource Manager / Job scheduler

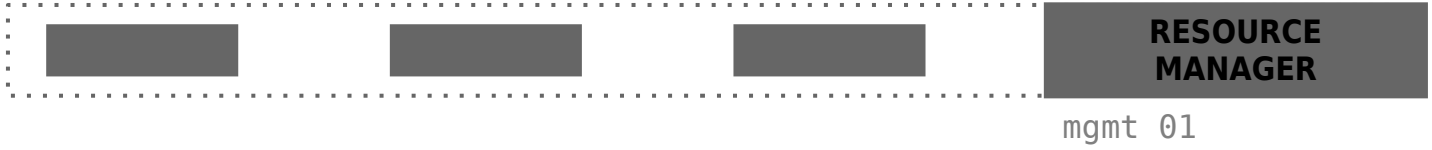
- Reserves and allocates hardware resources to execute application
- It can also organize execution of multiple applications (jobs)



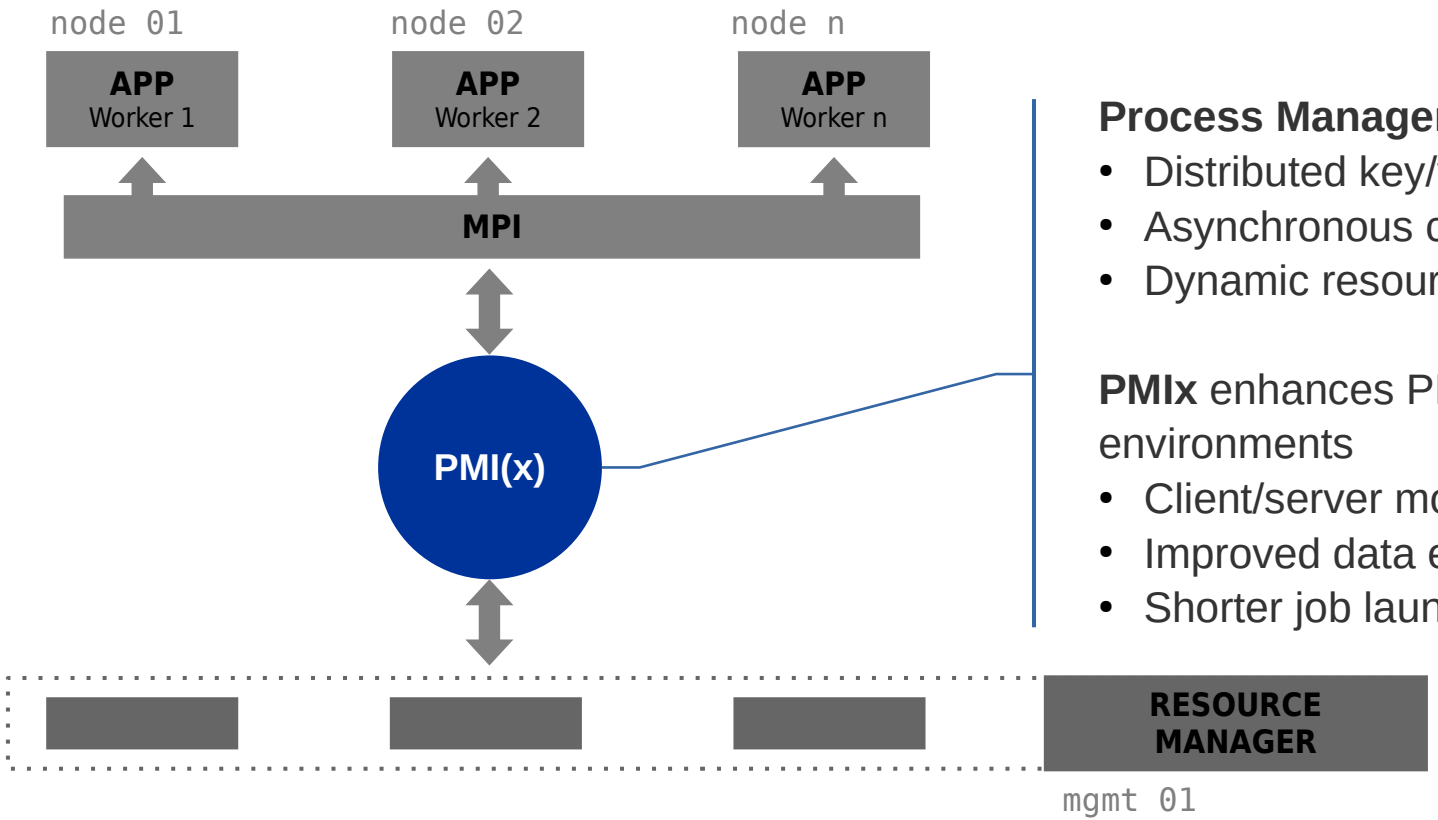
DISTRIBUTED PARALLEL COMPUTING



How do they interact?



PROCESS MANAGEMENT INTERFACE



Process Management Interface (PMI)

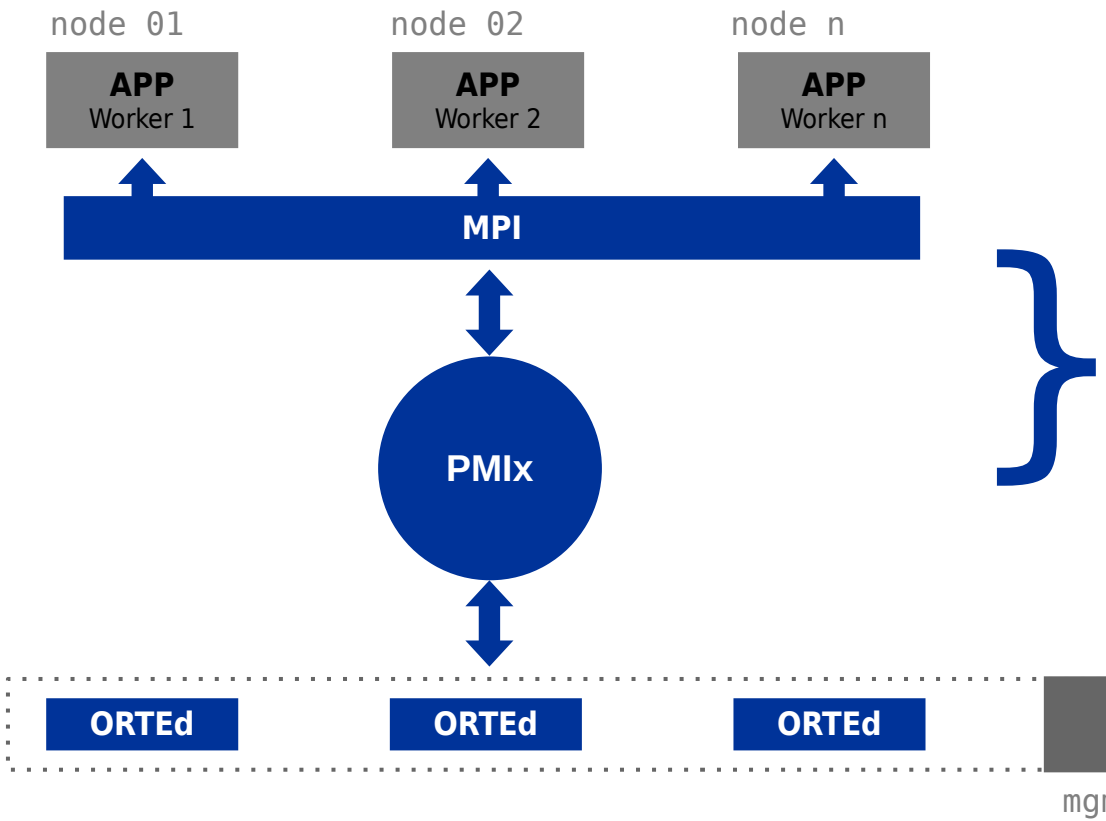
- Distributed key/value store
- Asynchronous communication
- Dynamic resource management

PMIx enhances PMI for exascale environments

- Client/server model
- Improved data efficiency
- Shorter job launch times

OpenPMIx
openpmix.github.io

PMIX IN OPENMPI



OpenMPI is the main open source MPI implementation

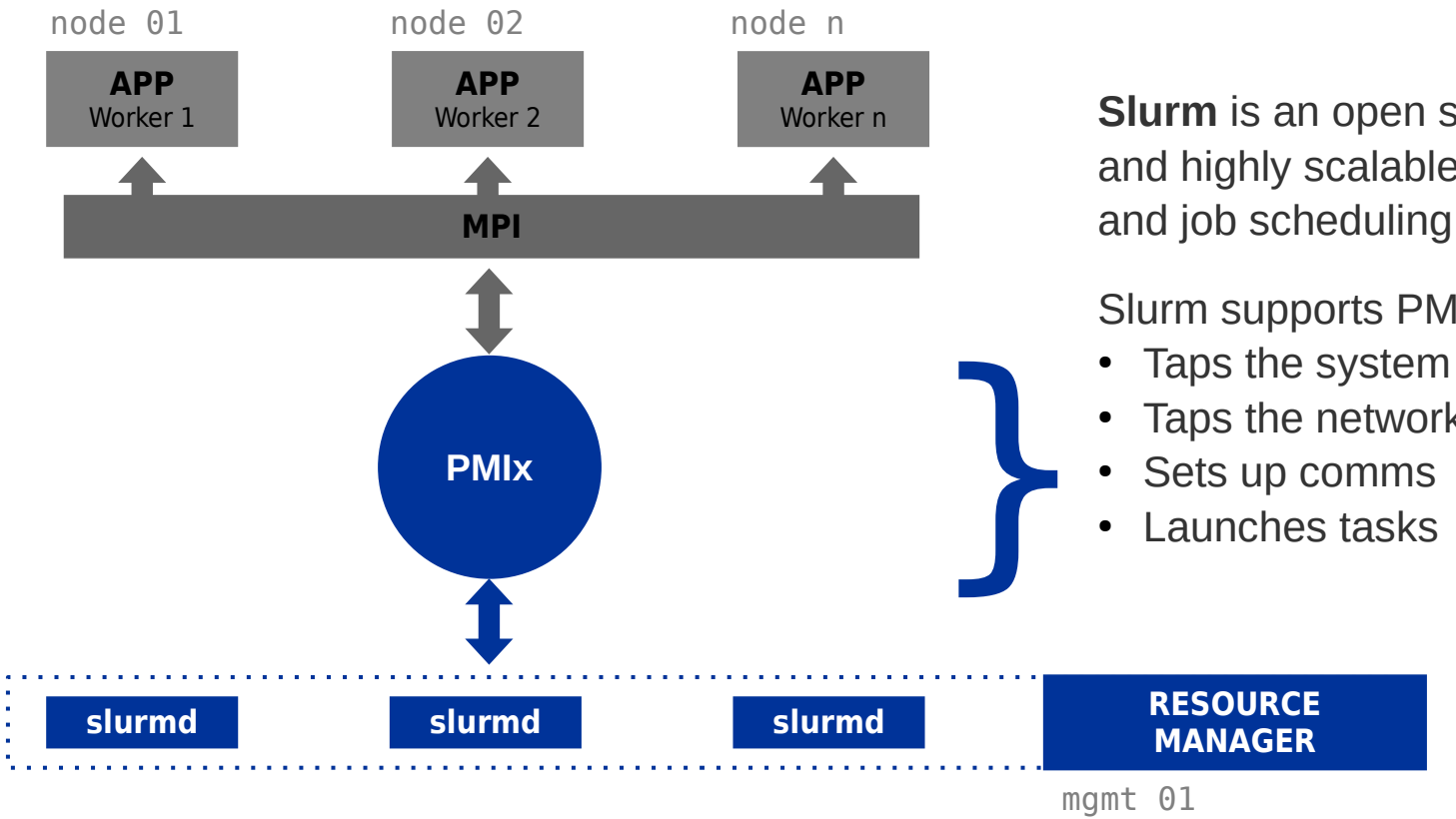
Not just the MPI API, it has its own runtime environment (ORTE)

- Taps the system hardware
- Taps the network
- Sets up comms
- Uses PMIx internally

You might already be using PMIx without knowing!

OpenMPI
www.open-mpi.org

PMIX IN SLURM



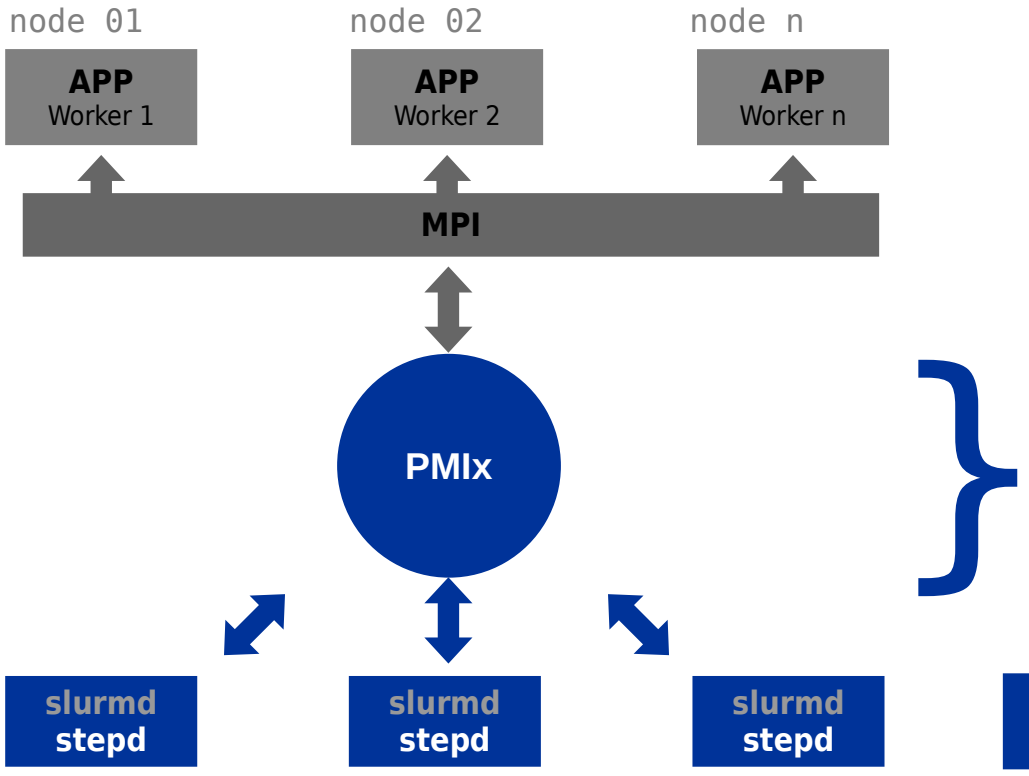
Slurm is an open source, fault-tolerant, and highly scalable cluster management and job scheduling system.

Slurm supports PMIx

- Taps the system hardware
- Taps the network
- Sets up comms
- Launches tasks

Slurm
slurm.schedmd.com

PMIX IN SLURM



Slurm is an open source, fault-tolerant, and highly scalable cluster management and job scheduling system.

Slurm supports PMIx

- Taps the system hardware
- Taps the network
- Sets up comms
- Launches tasks
- Direct communication between slurmstepd daemons

Slurm
slurm.schedmd.com

BUILDING SLURM WITH PMIX

Slurm supports PMIx since version 16.05 and PMIx-3 since v18.08

```
$ configure --with-pmix
```

Plugins for multiple PMIx versions can be installed at the same time

```
$ configure –with-pmix=/path/to/pmix-2:/path/to/pmix-3
```

USING SLURM WITH PMIX

The plugins can be selected at runtime with the `--mpi` option of **`srun`**

```
$ srun --mpi=list
srun: MPI types are...
srun: pmix_v3
srun: pmix
srun: none
srun: pmi2
```

Very nice, but **why would you want to use multiple versions of PMIx?**

BACK TO PMIX IN OPENMPI

Support for PMIx in OpenMPI (and other MPI implementations) varies from version to version

- If your system only provides a single OpenMPI version, choose the corresponding PMIx version and use that in Slurm
- Systems supporting multiple versions of OpenMPI are more tricky to set up

OpenMPI	Internal PMIx	External PMIx
1.x	x	
2.x	v1	v1
3.x	v2	v1, v2, v3
4.x	v3	v2, v3, v4
5.x	v4	v2, v3, v4

EXAMPLE: VUB TIER-2 HPC

Slurm MPI support

- Build with PMIx-3.2.3 from CentOS repos
- Install Slurm PMI libs (`slurm-pmi`) out of `$PATH`

```
/usr/lib64/slurmpmi
```

- Set default Slurm MPI type to **none** in `slurm.conf`

```
MpiDefault=none
```

Different OpenMPI versions are handled with software modules that set the appropriate MPI type on load

OpenMPI	Slurm MPI
1.x	x
2.x	PMI2
3.x	PMIx-3
4.x	PMIx-3
5.x	???

EXAMPLE: VUB TIER-2 HPC

- OpenMPI v2 is built with Slurm PMI2 libs

```
$ configure --with-slurm --with-pmi --with-pmi-libdir=/usr/lib64/slurmpmi
```

- OpenMPI v3 is built with external PMIx-3 libs
  - External PMIx libraries are different than those used by Slurm (different version and different dependencies)

We use **EasyBuild** to build each pair of OpenMPI and PMIx versions with the exact same dependencies

- EasyBuild is an open source build and installation system specifically designed to
  - Manage the huge software libraries of an HPC cluster
  - Build software optimized to the hardware executing it

```
1) GCCcore/10.3.0
2) zlib/1.2.11-GCCcore-10.3.0
3) binutils/2.36.1-GCCcore-10.3.0
[...]
10) hwloc/2.4.1-GCCcore-10.3.0
11) libevent/2.1.12-GCCcore-10.3.0
12) UCX/1.10.0-GCCcore-10.3.0
13) libfabric/1.12.1-GCCcore-10.3.0
14) PMIx/3.2.3-GCCcore-10.3.0
15) OpenMPI/4.1.1-GCC-10.3.0
```

EasyBuild
easybuild.io

EXAMPLE: VUB TIER-2 HPC

We use **Lmod** to handle all that software that can be loaded on-demand by our users

- Open source module system with modules written in Lua
- Modules modify the system environment to dynamically add software to the active shell environment

Multiple versions of OpenMPI are distributed with such software modules

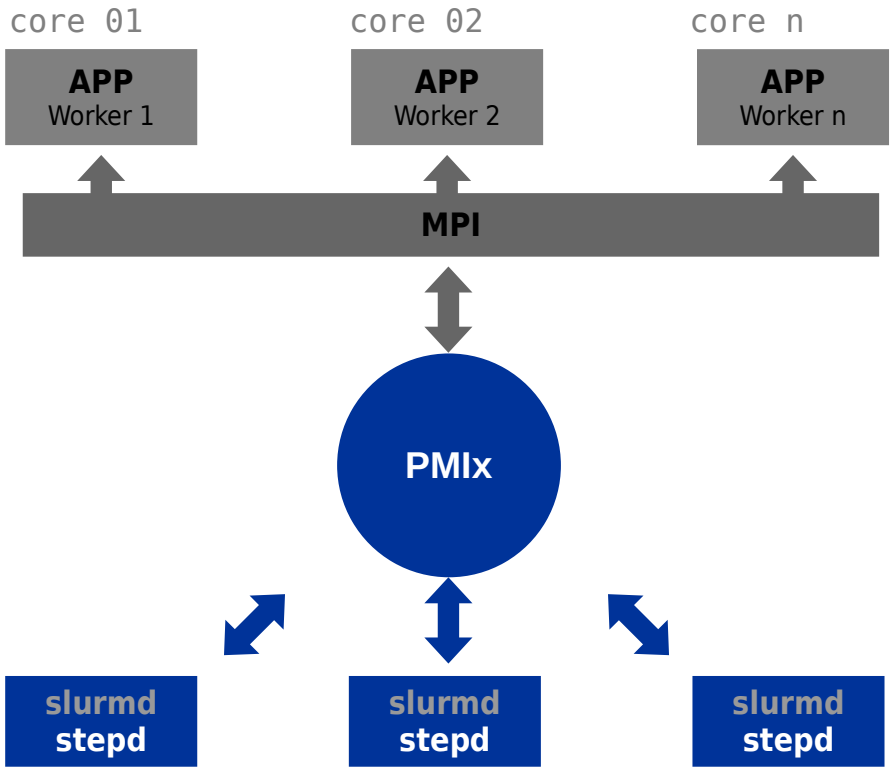
- Modules set **SLURM_MPI_TYPE** env variable on load

```
setenv("SLURM_MPI_TYPE", "pmix")
```

```
1) GCCcore/10.3.0
2) zlib/1.2.11-GCCcore-10.3.0
3) binutils/2.36.1-GCCcore-10.3.0
[...]
10) hwloc/2.4.1-GCCcore-10.3.0
11) libevent/2.1.12-GCCcore-10.3.0
12) UCX/1.10.0-GCCcore-10.3.0
13) libfabric/1.12.1-GCCcore-10.3.0
14) PMIx/3.2.3-GCCcore-10.3.0
15) OpenMPI/4.1.1-GCC-10.3.0
```

Lmod
lmod.readthedocs.io

DIRECT POINT-TO-POINT CONNECTIONS

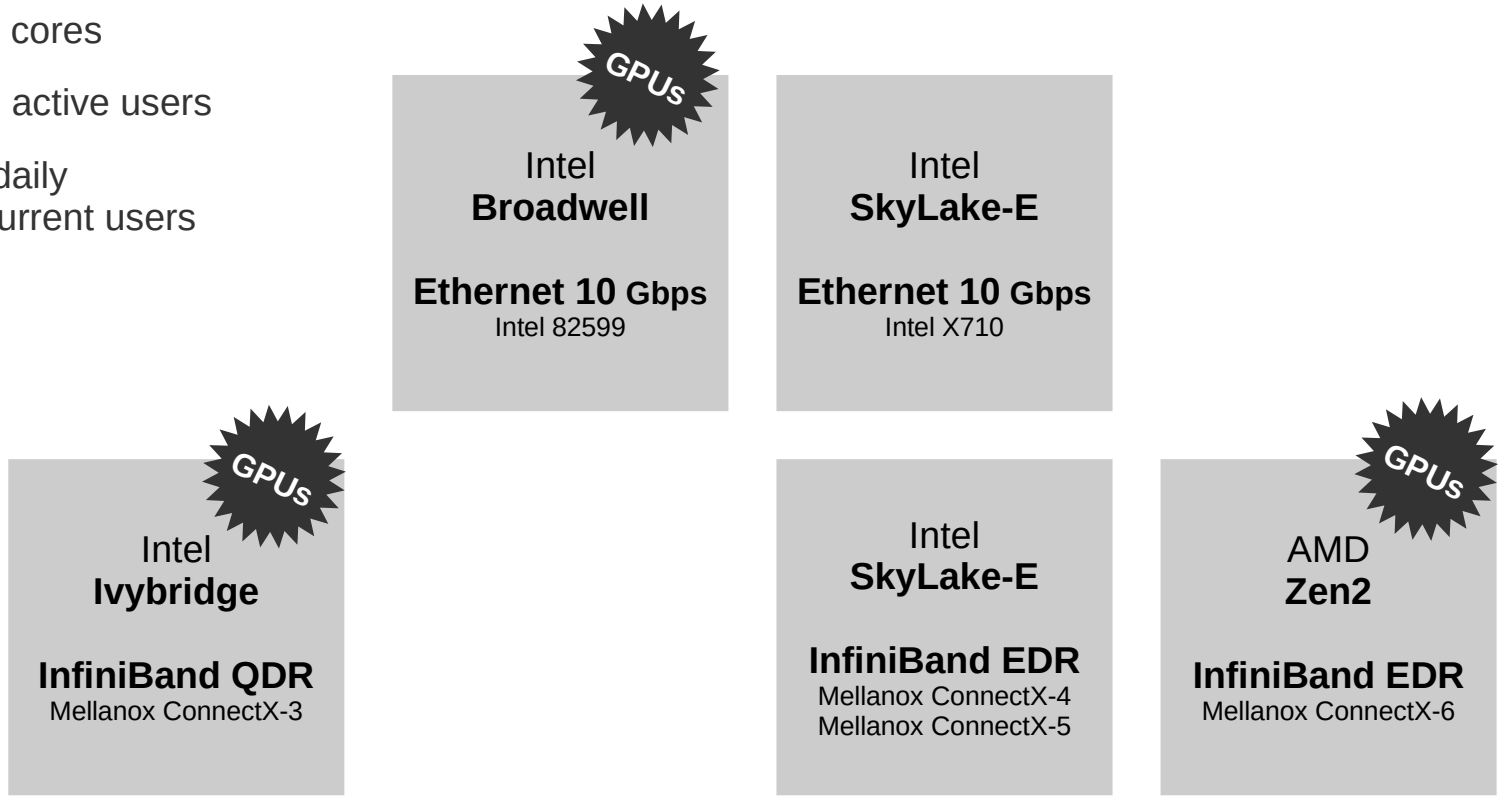


Direct communication between slurmstepd daemons

- `SLURM_PMIX_DIRECT_CONN` uses TCP (default)
- `SLURM_PMIX_DIRECT_CONN_UCX` uses UCX, but not as you might expect

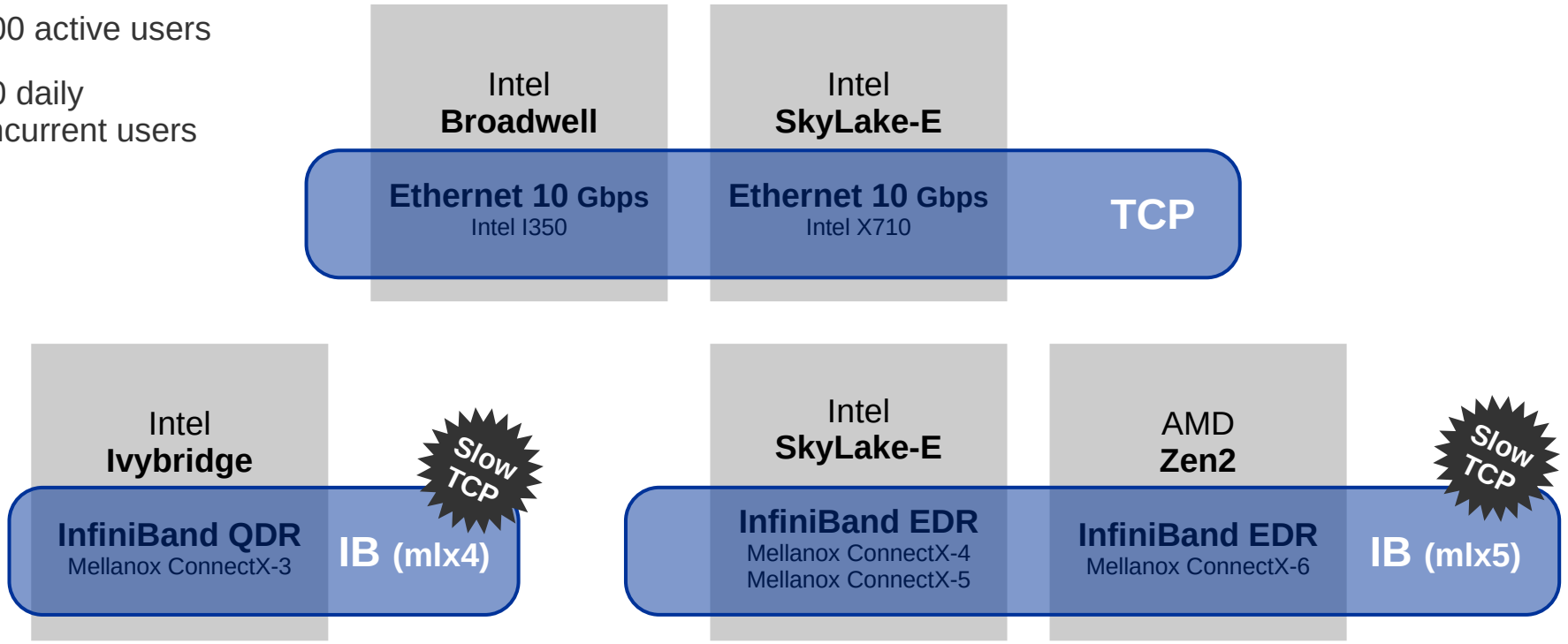
VUB TIER-2 HPC (HYDRA)

- 3648 cores
- ~500 active users
- ~50 daily concurrent users



VUB TIER-2 HPC (HYDRA)

- 3648 cores
- ~500 active users
- ~50 daily concurrent users



SLURM_PMIX_DIRECT_CONN_UCX

Slurm can be build with UCX to enable UCX in direct-connect

- This allows to leverage IB networks
- **But it is limited to Mellanox cards with the mlx5 driver and the dc transport !**

```
[node350:239329:0:239379] ib_mlx5_log.c:145  Transport retry count exceeded on mlx5_0:1/IB
(synd 0x15 vend 0x81 hw_synd 0/0)
[node350:239329:0:239379] ib_mlx5_log.c:145  RC QP 0x10546 wqe[3]: SEND --e [va
0x2b8fbfbf9500 len 568 lkey 0x116252]
```

```
[node351:101945:0:101954] Caught signal 11 (Segmentation fault: address not mapped to
object at address 0x18614b3240)
srun: error: _server_read: fd 16 error reading header: Connection reset by peer
srun: error: step_launch_notify_io_failure: aborting, io error with slurmstepd on node 1
```

SLURM_PMIX_DIRECT_CONN_UCX

Slurm can be build with UCX to enable UCX in direct-connect

- This allows to leverage IB networks
- **But it is limited to Mellanox cards with the mlx5 driver and the dc transport !**

InfiniBand QDR
Mellanox ConnectX-3

Ethernet 10 Gbps
Intel I350

InfiniBand EDR
Mellanox ConnectX-4
Mellanox ConnectX-5

default

```
SLURM_PMIX_DIRECT_CONN=true
SLURM_PMIX_DIRECT_CONN_UCX=false
```

task prolog

```
SLURM_PMIX_DIRECT_CONN=true
SLURM_PMIX_DIRECT_CONN_UCX=true
UCX_TLS=dc,sm,self
```

CONCLUSIONS

We configured a Tier-2 HPC cluster to execute distributed parallel jobs with a fully open source stack based on:

Slurm, OpenMPI and PMIx

- Use of IB networks where possible
- Low data footprint and data exchange on mid-tier networks
- Adopt de facto standard technologies
- Faster job starting times (theoretically)

ACKNOWLEDGEMENTS

- **Ward Poelmans and Sam Moors (colleagues in VUB-HPC)**
- EasyBuild community for its openness
- VUB for hosting us and feeding new users to our cluster
- VSC for financial support