Interactive European Grid

MPI Support in Int.Eu.Grid: Open MPI, PACX-MPI, MPI-Start, Marmot

Kiril Dichev, Rainer Keller

HLRS, Stuttgart

MPI Support
Overview

- Open MPI
  - An full-featured MPI-2 implementation
- PACX-MPI
  - MPI between clusters
- MPI-Start
  - Common layer to start MPI processes in EGEE and I2G
- Marmot
  - MPI application checking tool

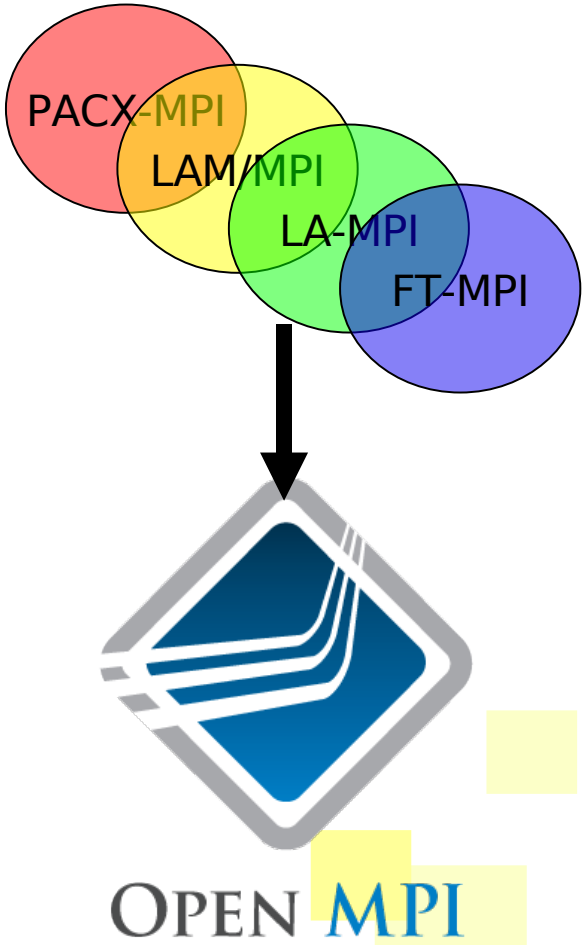
int.eu.grid

Open MPI Support

H L R S

Open MPI
History

- At SC03, the developers of FT-MPI, LA-MPI, LAM/MPI decided to focus their experience and efforts on one MPI implementation, in 2004 PACX-MPI joined.
- in 2004 the real designing and coding started
- 1st Release at SC 2005



Open MPI
Who is MPI?

- Current status:
  - Stable version v1.2.4 (as of Sept 2007)
  - Release v1.3 sometime late 2007 / early 2008
- 14 members, 6 contributors
  - 4 US DOE labs
  - 8 universities
  - 7 vendors
  - 1 individual

Open MPI
int.EU.Grid status

- Open MPI v1.2.2 RPM is available to int.EU.Grid
- Installed on all clusters and tested + validated
- Batch startup works on all sites
- Interactive startup works on most sites (local issues)
- Deviations from plan: none
- Next steps: Getting v1.2.4 up on all sites

int.eu.grid

PACX-MPI Support

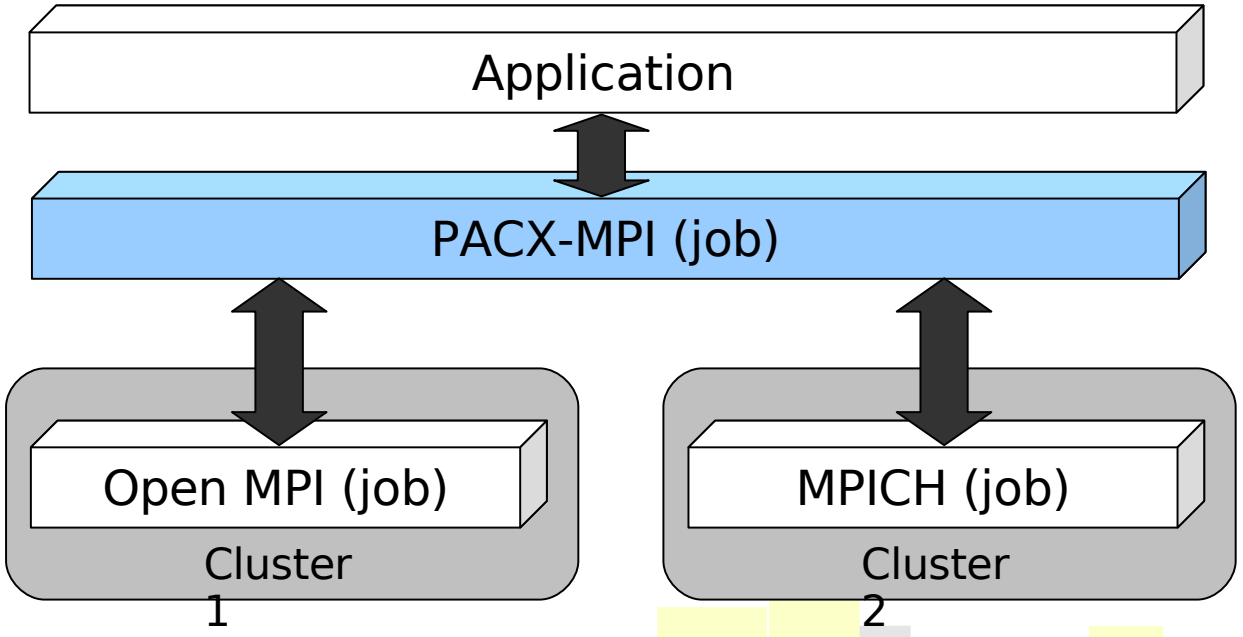
int.eu.grid

PACX-MPI
Overview

- A middleware to seamlessly run MPI-applications on a network of parallel computers (originally dev. in 1995 to connect Vector+MPP).
- PACX-MPI is an optimized standard-conforming MPI- implementation, applications just need **recompilation**(!)
- For C: pre-processor renaming: MPI_Send becomes PACX_Send.
- For Fortran: Function replacement @ link-step.

PACX-MPI
Overview

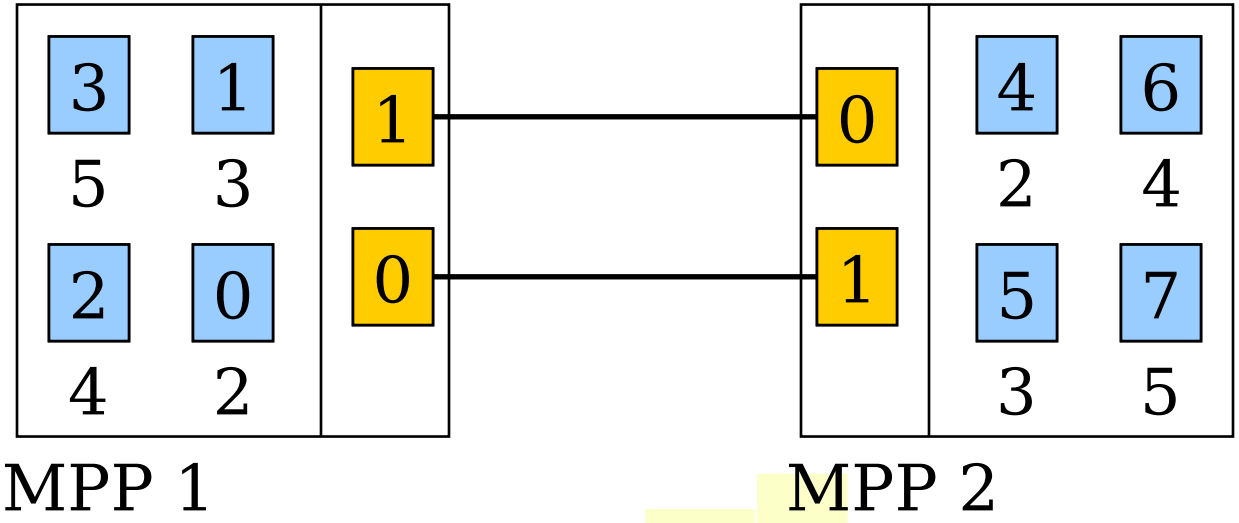
- PACX-MPI starts an MPI job in each cluster
- PACX-MPI “merges/manages” these MPI jobs internally and emulate transparently a bigger MPI job to the application



PACX-MPI
Usage

- Compiling with PACX
  - pacxcc -c hello.c
  - pacxcc -o **hello** hello.o

- Running requires 2 additional processes:



PACX-MPI
int.EU.grid status

- PACX-MPI v5-i2g RPM is available
- Installed on all clusters of I2G
- Interactive startup not supported on some sites
  Reason (h) in wiki – needs site-local fix.
- Deviations from plan: see above
- Few applications use it
- (Application DD_Filtre2 needs Fortran)

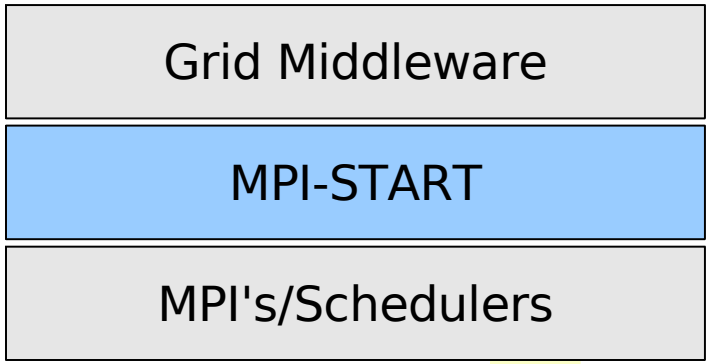
int.eu.grid

MPI-Start Support

H L R S

MPI-Start
Goals

- Goals of mpi-start:
  - Define a unique interface to the upper layer for MPI jobs
  - Support of a new MPI implementation doesn't require any change in the Grid middleware
  - Support of “simple” file distribution
  - Provide some support for the user to help manage his data.



MPI-Start
Design Goals

- Design Goals
  - Portable
    - The program must be able to run under any supported operating system
  - Modular and extensible architecture
    - Plugin/Component architecture
  - Relocatable
    - The program must be independent of absolute path, to adapt to different site configurations.
    - Remote "*injection*" of mpi-start along with the job
  - Very good "remote" debugging features

MPI-Start
int.EU.grid status

- Support of MPI in a single cluster
- Support of different MPIs simultaneously

- Remove all MPI implementation specific features from the middleware
- File distribution plugins being developed together with EGEE

- Deviations from plan: Not in Annex I
- Future work: Improvements with file-distribution

int.eu.grid

Marmot Support

HLRS

Marmot
Overview

- MPI checking tool for MPI-errors at runtime
- Developed in the frame of CrossGrid
- Library written in C++, linked to the application
- No source code modification required
- One Additional process working as debug server
- Implementation of C and Fortran language binding of MPI-1.2 standard

Marmot
Usage

- After linking to Marmot, start with +1 process:

```
9314 rank 2 performs MPI_Barrier
9315 rank 1 performs MPI_Sendrecv
9316 rank 2 performs MPI_Sendrecv
9317 rank 0 performs MPI_Sendrecv
9318 rank 1 performs MPI_Sendrecv
9319 rank 0 performs MPI_Sendrecv
9320 rank 2 performs MPI_Sendrecv
9321 rank 0 performs MPI_Barrier
9322 rank 1 performs MPI_Barrier
9323 rank 2 performs MPI_Barrier
9324 rank 1 performs MPI_Comm_rank
9325 rank 1 performs MPI_Bcast
9326 rank 2 performs MPI_Comm_rank
9327 rank 2 performs MPI_Bcast
9328 rank 0 performs MPI_Sendrecv
WARNING: all clients are pending!
```

Iteration step: Calculate and exchange results with neighbors

Communicate results among all procs

Marmot
int.EU.grid status

- Marmot RPM is build on top of Open MPI
- Test + Validation request pending (static libraries tested)
- Work in progress:
  - Dynamic libraries to be tested
  - Marmot usage only possible with newest MPI-Start