Earth Science
Requirements and experiences with use of MPI in EGEE

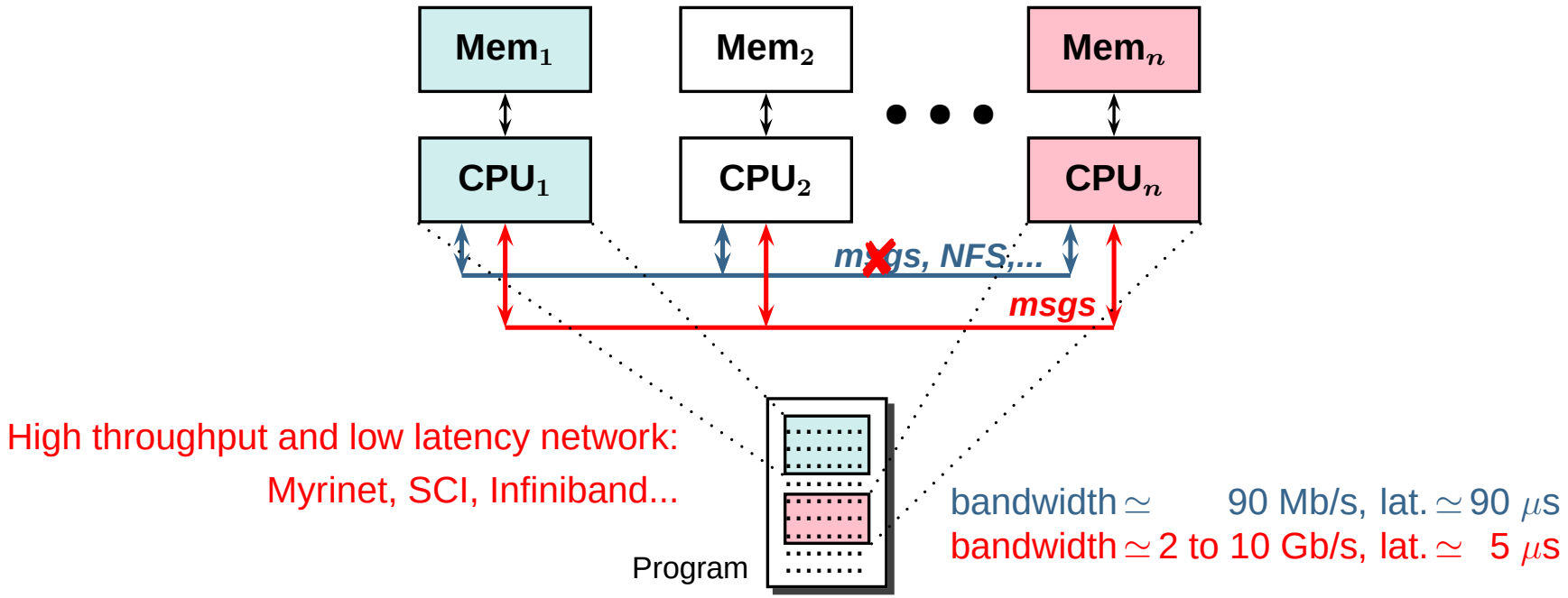
Jean-Pierre Vilotte and Geneviève Moguilny

Institut de Physique du Globe de Paris, CNRS (FR)

Contents

MPI for the dummies

- A program written with a traditional language (Fortran, C...).
- Each process receives the same instance of the program, but with conditional instructions, runs the same code or a different part of the code, on the same data or on different data.
- The variables of the program are private and stay in the memory of the processor allocated to the process.
- Data is exchanged between processes through calls to routines of a message passing library like **MPI** (Message Passing Interface).



The MPI message passing library

Review:

1992: Need to create portable applications with good performance $\Rightarrow$ creation of a Working Group (mainly from the US and Europe) to adopt the HPF methods.

1994: Version 1.0 of MPI.

1997: Version 1.2.
Version 2.0: definition of a standard for MPI-2 (dynamic control of tasks, parallel I/O...).

2007: MPI-2 version 2.1.

Main open source implementations:

- **MPICH**, with **`mpirun`** to launch jobs/ processes via ssh, currently MPICH-1.2.7 on EGEE.
  2009: **MPICH2** (MPI-1 + MPI-2 standards): implementation that efficiently supports different computation and communication platforms. MPICH2-1.1 on EGEE.
- **OpenMPI** (MPI-2 standard), combine FT-MPI, LA-MPI, LAM/MPI and PACX-MPI;
  with **`mpiexec`**. OpenMPI-1.3.2 on EGEE.

and based on,

- specific network libraries (MPI-GM and MPI-MX: MPICH over Myrinet...).

Very small example

```
 1: program HelloMPI
 2:
 3: implicit none
 4: include 'mpif.h'
 5: integer :: nb_procs, rang, ierr
 6:
 7: call MPI_INIT(ierr)
 8:
 9: call MPI_COMM_SIZE(MPI_COMM_WORLD, nb_procs, ierr)
10: call MPI_COMM_RANK(MPI_COMM_WORLD, rang, ierr)
11: print *, 'I am the process number ', rang,' among ',nb_procs
12:
13: call MPI_FINALIZE(ierr)
14:
15: end program HelloMPI
```

- **Compilation and link:**

  `mpif90 HelloMPI.f90 -o HelloMPI`

  ⇕

  `ifort -c -I`*includes_path* **`HelloMPI.f90`**
  `ifort -L`*libs_path* **`HelloMPI.o -static-libcxa -o HelloMPI -lmpichf90 -lmpich`**

- **Execution:**

  `mpirun -np 4 -machinefile mf HelloMPI`

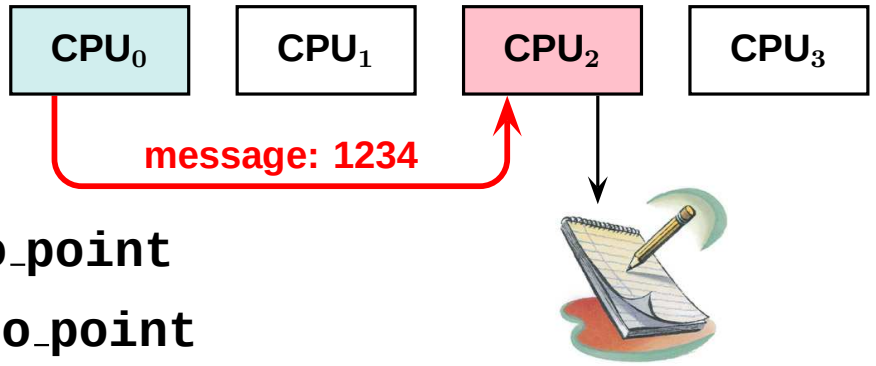
  Run **`HelloMPI`** (with `ssh` or `rsh`) on 4 processors (**`np`** = *number of processors*) of hosts listed in the file named **`mf`**.

- **Output:**

```
 I am the process number           0  among           4
 I am the process number           2  among           4
 I am the process number           3  among           4
 I am the process number           1  among           4
```

Simple example of point to point communication

```
 1: !! point_a_point.f90 : Exemple d'utilisation de MPI_SEND et MPI_RECV
 2: !! Auteur : Denis GIROU (CNRS/IDRIS - France) <Denis.Girou@idris.fr> (1996)
 3: program point_to_point
 4:   implicit none
 5:   include 'mpif.h'
 6:   integer, dimension(MPI_STATUS_SIZE) :: statut
 7:   integer, parameter                  :: etiquette=100
 8:   integer                             :: rank,value,ierr
 9:   call MPI_INIT(ierr)
10:   call MPI_COMM_RANK(MPI_COMM_WORLD,rank,ierr)
11:
12:   if (rank == 0) then
13:      value=1234
14:      call MPI_SEND(value,1,MPI_INTEGER,2,etiquette,MPI_COMM_WORLD,ierr)
15:   elseif (rank == 2) then
16:      call MPI_RECV(value,1,MPI_INTEGER,0,etiquette,MPI_COMM_WORLD,statut,ierr)
17:      print *,'Me, process ', rank, ', I received  ',value,' from the process 0.'
18:   end if
19:
20:   call MPI_FINALIZE(ierr)
21: end program point_to_point
```



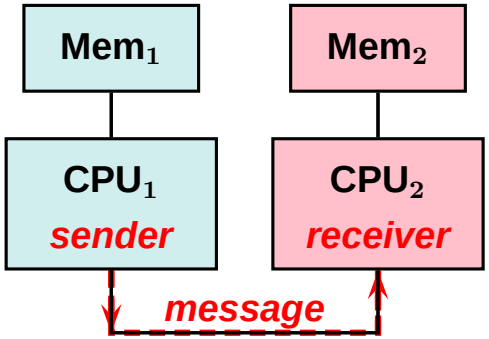
mpif90 point_to_point.f90 -o point_to_point

mpirun -np 4 -machinefile mf point_to_point

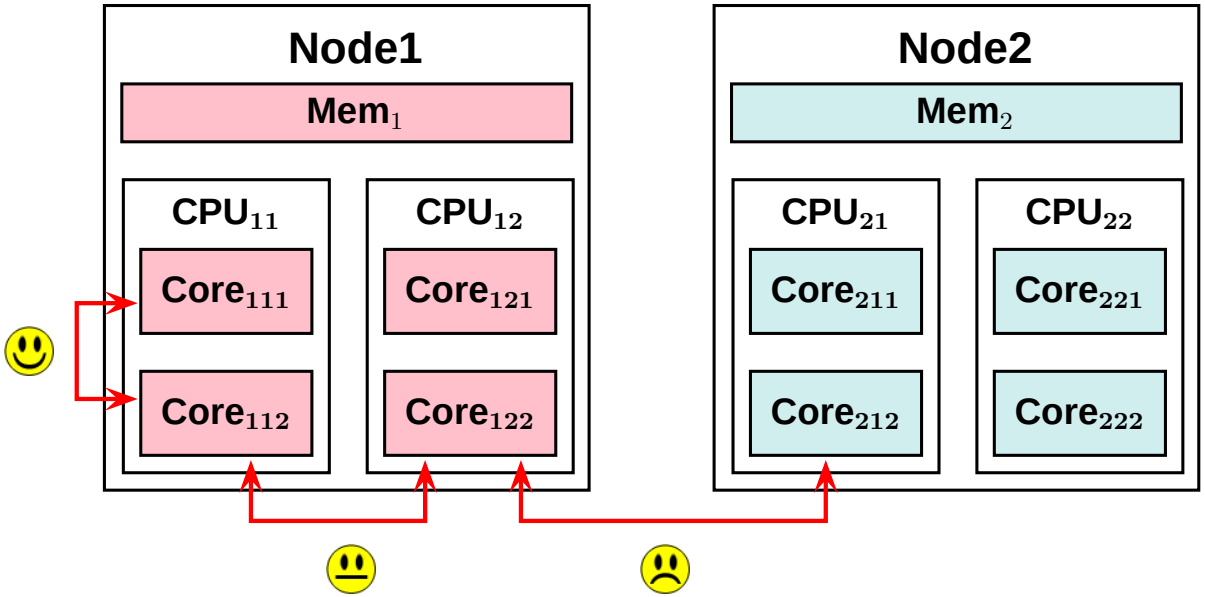
```
Me, process    2 , I received    1234 from the process 0.
```

Optimization

Previously, 1 Node = 1 CPU (= 1 Core)



Nowadays, 1 Node = n CPUs = $n \times m$ Cores



1 process runs on 1 core but too much processes on one node can cause lack of memory.

MPI application development at IPGP

IPGP (*Institut de Physique du Globe de Paris*, CNRS, FR) is one of the main research and educational institution in the domain of Geosciences in France and is in charge of several national observatories and of the volcanic hazard monitoring of the French volcanos in the Antillas and the Reunion Island.

Research and monitoring activities of IPGP involve the development of new data analysis and simulation methods. Some of its most important applications are parallelized with MPI and Fortran 90, and run on local resources or national computing centers.

IPGP is involved in

NA4, VO ESR (*Earth Sciences Research*) , *Solid Earth Physics* domain.

Why MPI?

MPI allows to change the order of magnitude of the problem to solve, by a possible quasi-unlimited increase of the available memory, with standard hardware and software.

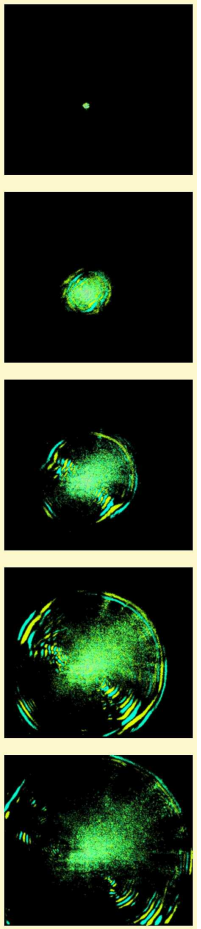
Well parallelized MPI application $\Rightarrow$
execution time inversely proportional to the number of used cores.

MPI-1: $\simeq$ 150 routines, but only about 10 really essential and used.

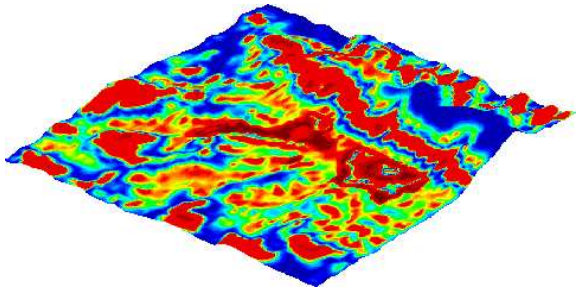
First MPI application on EGEE: SPECFEM3D

Resolution of regional and global scales seismic wave propagation for complex geological structures, with use of the spectral-element method (SEM).
First written by D. Komatitsch (IPGP then Université de Pau).

- $\simeq$ 20 000 lines of Fortran 90 / MPI.
- Very scalable application, ran on 1944 CPUs at the *EarthSimulator* (Japan). On EGEE, ran on 64 CPUs at Nikhef (NL), on 4 or 16 CPUs at SCAI (DE), LAL (FR), CPPM (FR), CGG (FR), SARA (NL), IISAS-Bratislava (SK), HG-01-GRNET (GR), TU-Kocise (SK), AEGIS01-PHY-SCL (YU) and ACAD (BG).
- **Constraints:**
  - Memory optimization $\Rightarrow$ necessary recompilation and update of the input files on SE at each input parameter change.
  - Heavy outputs ($\Rightarrow$ writes on the `/tmp` of each node), + outputs on shared directory to retrieve **($\Rightarrow$ `/home` NFS mounted)**.
  - Need to launch successively **two mpirun** that have to use the same nodes, allocated in the same order.



Other MPI application running on EGEE: SEMUM3D



Simulation code for 3D seismic wave propagation in elastic, heterogeneous media at local and regional scales. SEMUM3D is especially designed for the simulation of the seismic response of complex geological media such as sedimentary basins.

Complex geometries using unstructured hexahedra meshes.
Parallel implementation based on domain decomposition and mesh partitioning with **ParMETIS**.

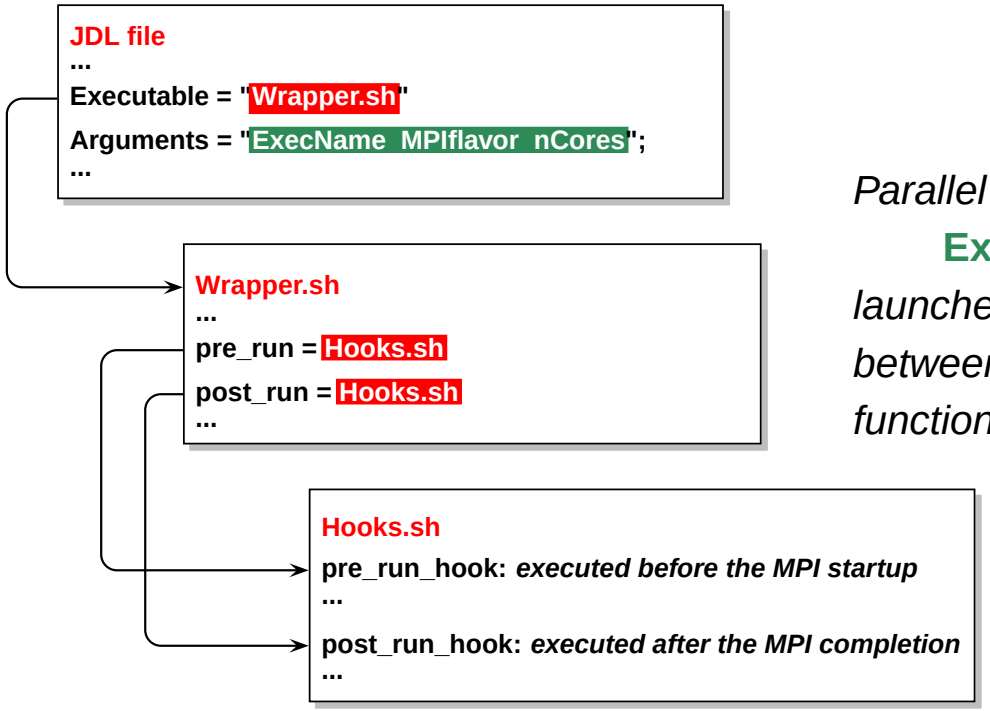
Main problems encountered

- Long Restitution time.
- Heterogeneity of MPI installations,
  solution recommended by the **TCG Working Group on MPI**:
  - Sites: configuration, publication of the implementation (flavor, version) of MPI, paths, homes sharing or not...
  - Users submission (MPI-START package):
    separation between environment definition / launch of the application.

Key requirements:

- **More CEs with MPI and shared homes.**
- **Real and complete implementation of these TCG WG on MPI recommendations.**

Functionality issues with MPI-START



Parallel execution of

ExecName *with* **MPIflavor** *on* **nCores** *cores launched by another script of the MPI-START package, between the execution of* ***pre_run*** *and* ***post_run*** *functions*.

Issues

- If a CE meets all requirements except the MPI-START package, users can download it from the MPI wiki, adapt it and send it in the `InputSandbox`. In this way a user can only have ONE set of scripts for all CEs (with or without MPI-START installed), but, he has to know the path of the adequate `mpirun`.
- The launch of the mpirun is done by another script of the MPI-START package. So, for SEMUM3D it works, but SPECFEM3D needs the modification of another file of the package (`mpich.mpi`).

Key requirement:

- **A more user friendly configurable MPI-START package.**

Deployment of SEMUM3D on EGEE-III with MPI-START (1/2)

On the 26 CEs satisfying the requirements (ESR, MPICH, PBS) on July 2009
SEMUM3D ran correctly on 7 CEs (sometimes after several attempts).

Among these 26 CEs, only 16 are publishing if homes are shared or not, and when not shared and published, the tag used is not uniform: MPI_NO_SHARED_HOME or MPI_HOME_NOT_SHARED.
Other problems encountered:

- on some CEs, although the MPICH tag is published, it's not possible to know the path of `mpirun` (no MPI-START, no `locate` command and now, it seems to be impossible to use interactive jobs to find it on the CEs);
- on some CEs, the job is systematically *aborted* with no error message;
- on other CEs, the application has a MPI problem (*MPI_SEND invalid rank*);
- finally on some other CEs, the application is never launched (at least, 72 h after the submission).

Hopefully, the application ran:
– always OK on **gridgate.cs.tcd.ie** (also in EELA), like on all **in2p3.fr** CEs,
– almost always OK on **ce.grid.rug.nl** and **afroditi.hellasgrid.gr**,
– sometimes OK on **ce1.egee.fr.cgg.com**.

Deployment of SEMUM3D on EGEE-III with MPI-START (2/2)

EGEE CE Hostname	Site	MPI-START	(No-)shared homes published?	Published implementations	SEMUM3D run with `glite-wms...` commands
ce.grid.rug.nl	RUG-CIT	Yes	MPI_SHARED_HOME	mpich1, openmpi	**OK**
ce.ngcc.acad.bg	BG03-NGCC	Yes	MPI_SHARED_HOME	mpich1	`Aborted`
ce02.grid.acad.bg	BG04-ACAD	Yes	MPI_SHARED_HOME	mpich1	`Aborted`
ce01.afroditi.hellasgrid.gr	HG-03-AUTH	Yes	MPI_SHARED_HOME	lam, mpich1, mpich2, openmpi	**OK**
ce01.ariagni.hellasgrid.gr	HG-05-FORTH	Yes		mpich1, mpich2, openmpi	`mpiexec: No such file or dir`
ce01.athena.hellasgrid.gr	HG-06-EKT	Yes		mpich1, mpich2, openmpi	`0-MPI_SEND: Invalid rank 1`
ce02.athena.hellasgrid.gr	HG-06-EKT	Yes		mpich1, mpich2, openmpi	`0-MPI_SEND: Invalid rank 1`
ce01.kallisto.hellasgrid.gr	HG-04-CTI-CEID	Yes		mpich1, mpich2, openmpi	`Aborted`
ce01.marie.hellasgrid.gr	HG-02-IASA	Yes		mpich1, mpich2, openmpi	`mpiexec: error while loading shared libraries`
ce02.marie.hellasgrid.gr	GR-06-IASA			mpich1 mpich2, openmpi	`Exit Code: 41`
ce01.isabella.grnet.gr	HG-01-GRNET	Yes		mpich1, mpich2, openmpi	`Aborted`
ce1.egee.fr.cgg.com	CGG-LCG2	Yes		mpich	**∼OK** but pb with tar
ce2.ui.savba.sk	IISAS-Bratislava	Yes	MPI_SHARED_HOME	mpich1, mpich2, openmpi	`Aborted`
gazon.nikhef.nl	NIKHEF-ELPROD			mpich1	`sh: mpirun: cmd not found`
trekker.nikhef.nl	NIKHEF-ELPROD			mpich1	`sh: mpirun: cmd not found`
grid001.ts.infn.it	INFN-TRIESTE		MPI_NO_SHARED_HOME	mpich1	`EC 127, executable not found`
prod-ce-02.pd.infn.it	INFN-PADOVA	Yes	MPI_HOME_NOT_SHARED MPI_NO_SHARED_HOME	mpich1	72h proxy expired
gridba2.ba.infn.it	INFN-BARI		MPI_HOME_NOT_SHARED	mpich	`EC 127 (Wrapper not found)` or OK
grid012.ct.infn.it	INFN-CATANIA	Yes	MPI_HOME_NOT_SHARED	mpich1, mpich2, mpich*	`EC 126` `Perm denied on executable`
gridce.sns.it	SNS-PISA		MPI_HOME_NOT_SHARED	mpich	`Aborted`
lapp-ce01.in2p3.fr	IN2P3-LAPP	Yes	MPI_SHARED_HOME	lam, mpich1, mpich2, openmpi	**OK**
grid10.lal.in2p3.fr	GRIF	Yes	MPI_SHARED_HOME	lam, mpich1, mpich2, openmpi	**OK**
marce01.in2p3.fr	IN2P3-CPPM	Yes	MPI_SHARED_HOME	lam, mpich1, mpich2, openmpi	**OK**
grid-eo-engine04.esrin.esa.int	ESA-ESRIN	Yes	MPI_NO_SHARED_HOME	mpich1	`Exit Code: 1`
gridgate.cs.tcd.ie	csTCDie	Yes	MPI_SHARED_HOME	lam, mpich1, mpich2, openmpi	**OK (also on EELA)**
egee-ce.csc.fi	CSC	Yes	MPI_SHARED_HOME	hpmpi, mpich	`Exit Code: 1`

ESR MPI needs on EGEE

- **Key requirements**
  - **More CEs (and/or more CPUs) supporting MPI.**
  - **CPUs reservation.**
  - **More MPI CEs with shared homes.**
  - **A more user friendly configurable MPI-START package.**
- Others requirements
  - Specific MPI queues with equivalent CPU power / CPU architecture.
  - CEs with longer `MaxCPUTime` than for sequential jobs.
  - 1 MPI process runs on 1 Core but 1 Node = n CPUs = $n \times p$ Cores
    - Specification of memory / process (to avoid lack of memory problem),
    - Specification of max process / node (to avoid network saturation).
  - High performance networks (Myrinet, Infiniband...).
  - Inter-sites MPI (**MPICH-G2**, **MPICH-V**)?

MPI WG suggestions

- **New JDL attributes**
  - `JobType = "Normal"` and `CpuNumber` $\Rightarrow$ run on a single cluster
  - Processes distribution: `SMPGranularity`
  - Reservation of whole nodes: `WholeNodes`
- **Information system variables** to use with `lcg-info` or in the JDL file
  - MPI-START support: `MPI-START`
  - MPI flavors: `MPICH, MPICH2, LAM, OPENMPI`
  - MPI versions: `OPENMPI-1.0.2, MPICH-1.2.7, MPICH-G2-1.2.7`...
  - MPI compilers
  - MPI interconnects
  - Shared Homes
- **Environment variables** to test/use in the Wrapper/Hooks user's scripts
  - Mandatory:
    `MPI_<flavor>_VERSION`, `MPI_<flavor>_PATH`, `MPI_MPICH_MPIEXEC`, `MPI_INTERCONNECT`, `MPI_SHARED_HOME`, `MPI_SHARED_AREA`.
  - Optional:
    `MPI_<flavor>_COMPILER`, `MPI_<flavor>_<version>_PATH`, `MPI_<flavor>_<version>_<compiler>_PATH`, `MPI_OPENMPI_COMPILER` (?).

Related projects

- **Int.eu.grid (Interactive European Grid Project)**

  Aim: deploying and operating an interoperable production-level e-Infrastructure for demanding interactive applications that will impact the daily work of researchers.

  Main features:

  - Distributed Parallel, Interactive Computing and Storage at the Tera level.
  - User Friendly Access through a Grid Interactive Desktop with powerful visualization.
  - VO support at all levels.

  Lots of developments done for **MPI** (MPI-START, cross-site MPI), some used on EGEE.

  http://www.i2g.eu (ended on 30th April 2008).

- **Dirac (Distributed Infrastructure with Remote Agent Control)**

  Aim: complete Grid solution for a community of users such as the LHCb Collaboration.
  Layer between a particular community and various compute resources (EGEE, EELA...)
  to allow optimized, transparent and reliable usage.

  MPI tests in progress.

  http://mareela.in2p3.fr/DIRAC
  https://twiki.cern.ch/twiki/bin/view/LHCb/DiracProject

Conclusion

Most of the important parallel IPGP applications

- use MPI(-1)/Fortran90 and
- need shared homes.

High performance network is in general less essential.

Two important parallel applications ran on EGEE: SPECFEM3D (without MPI-START) then, SEMUM3D (with MPI-START).

The implementation of MPI is much better than on EGEE-I or EGEE-II,
thanks to the work of the TCG WG on MPI.
Their **Web site** contains a lot of helpful information for Users and System administrators.

Still EGEE can't be used as a production tool because of

- the few sites satisfying the requirements (ESR, MPI, shared homes),
- the very long restitution time, even requiring few cores.