MPI y sus aplicaciones en infraestructuras Grid

Dr. Isabel Campos Plasencia

Instituto de Física de Cantabria-IFCA

Santander, Spain

CSIC

MPI en Infraestructuras Grid

CSIC

- Motivación
  - Soporte a computación paralela en el Grid
  - Problemas
  - Objetivos
- Soporte a MPI en el grid: mpi-start
  - Diseño
  - Arquitectura
- Ejemplos de uso
- Soporte a MPI en EGEE

Soporte a MPI en el Grid ¿Porqué?

- **Muchas áreas de aplicaciones requieren soporte a MPI**
  - Ciencias de la tierra, fusion, astrofísica, Química Computacional...
  - Se pueden obtener resultados significativos usando 10s-100s of CPUs
- **Muchos clusters de hecho están listos para usar MPI**
  - En modo local mediante envío directo
  - Sistemas de ficheros compartidos, intranets de alto rendimiento
- **Pero no pueden proveer ese acceso a través del Grid**
- **Un soporte de calidad atraería comunidades al Grid**
  - Cómo una infraestructura en sí misma
  - Cómo testbed antes de ejecutar en máquinas HPC

Soporte a MPI en el Grid
¿Porqué?

- **Muchas áreas de aplicaciones requieren soporte a MPI**
  - Ciencias de la tierra, fusion, astrofísica, Quími
  - Se pueden obtener resultados signifi
    CPUs
- **Muchos**
- **través del Grid**
- **idad atraería comunidades al Grid**
  - Cómo una infraestructura en sí misma
  - Cómo testbed antes de ejecutar en máquinas HPC

Hay muchos factores a la hora de dar soporte a trabajos MPI que se han solucionado a nivel de clusters individuales y SuperComputers, etc… que tienen que ser reanalizados cuando se quiere implementar MPI en el Grid

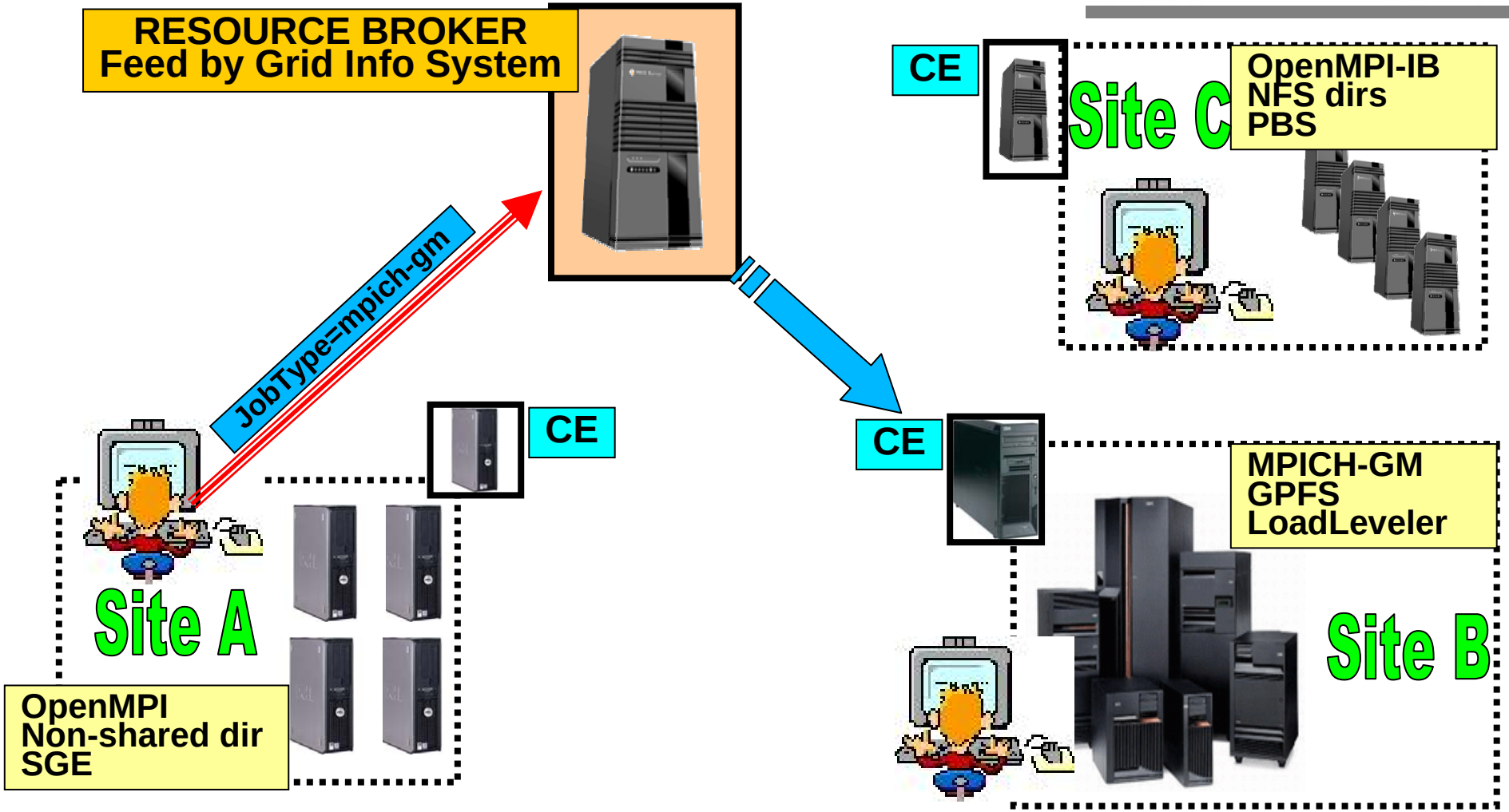
Problemas a resolver

No es un entorno homogéneo

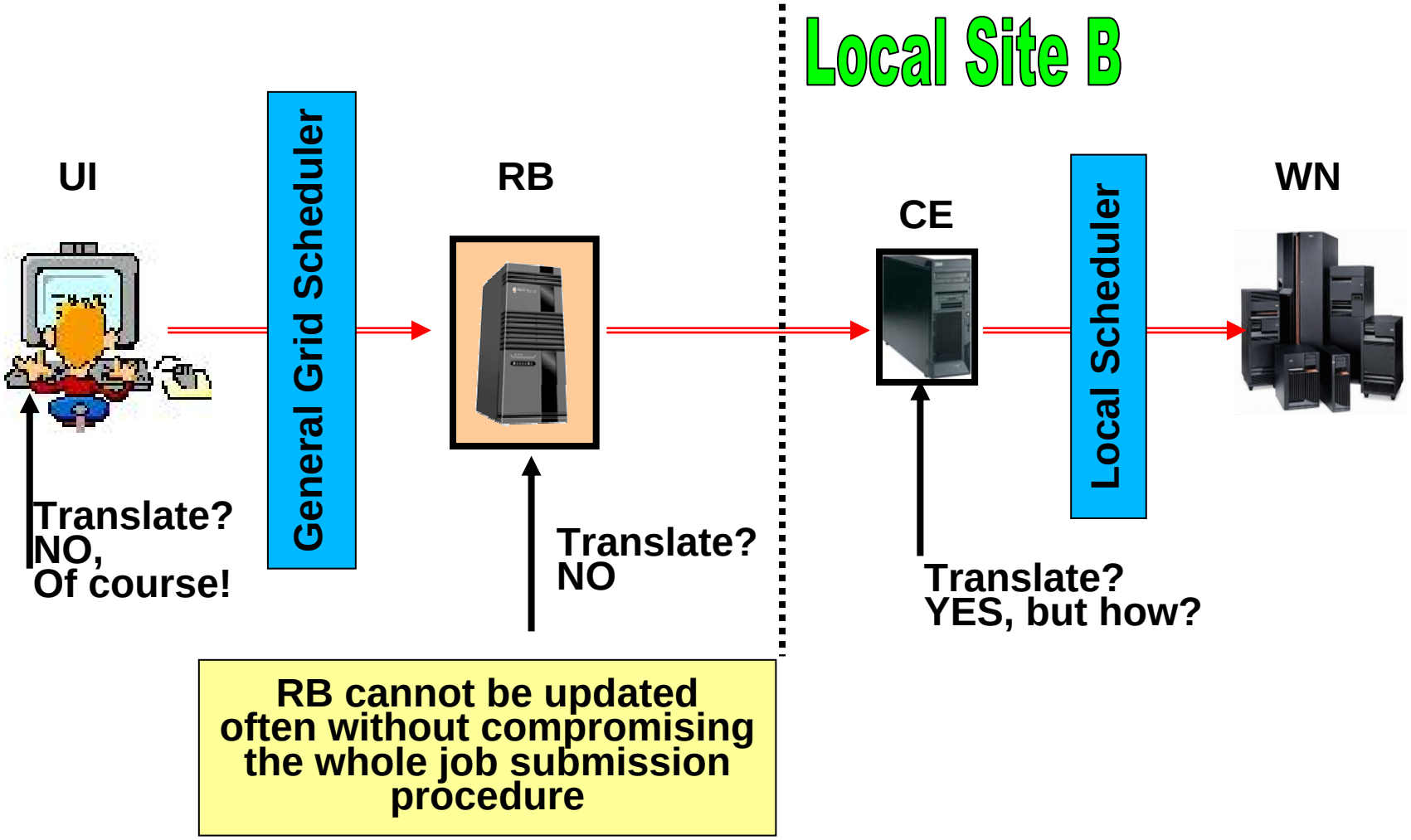
- **Sistemas de ficheros no compartidos**
  - Muchos sites no tienen soporte a sistemas de ficheros compartidos
  - Muchas implementaciones MPI esperan encontrar el ejecutable en el nodo donde se ejecuta el proceso
  - **En general el setup es muy variado**

- **MPI no establece un standard de cómo iniciar un programa**
  - No hay una sintaxis común para ***mpirun***
  - MPI-2 define mpiexec como mecanismo de lanzamiento, pero el soporte a mpiexec es opcional en todas las implementaciones
  - Los ***Brokers*** tienen que manejar distintas implementaciones MPI: MPICH, OpenMPI, LAMMPI,
  - ***Schedulers*** distintos (PBD, SGE,...) y distintas implementaciones MPI en cada site tienen distintas maneras de especificar el fichero ***machinefile***

Situación típica en el Grid



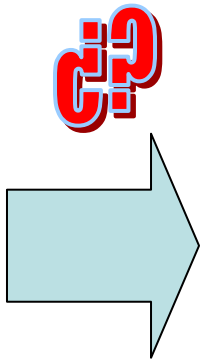
El lenguaje del Grid Scheduler tiene que ser traducido a la sintaxis del scheduler local



Ejemplo con Sun Grid Engine

RESOURCE BROKER

```
Executable = “myprog”;
Arguments  =  “arguments”;
JobType    =  “MPI”;
ProcNumber =  4;
StdOutput  = “std.out”;
StdError   = “std.err”;
InputSandBox = {“myprog”};
OutputSandBox = {“std.out”,
¨               “std.err”};
```



BATCH

```
#/bin/sh
#$ -o $HOME/mydir/myjob.out
#$ -N myjob
#$ -pe mpi 4
. /etc/profile.sge
. /etc/mpi.setup –e mpi
cd mydir
mpirun –np 4 ./myprog
```

machinefile

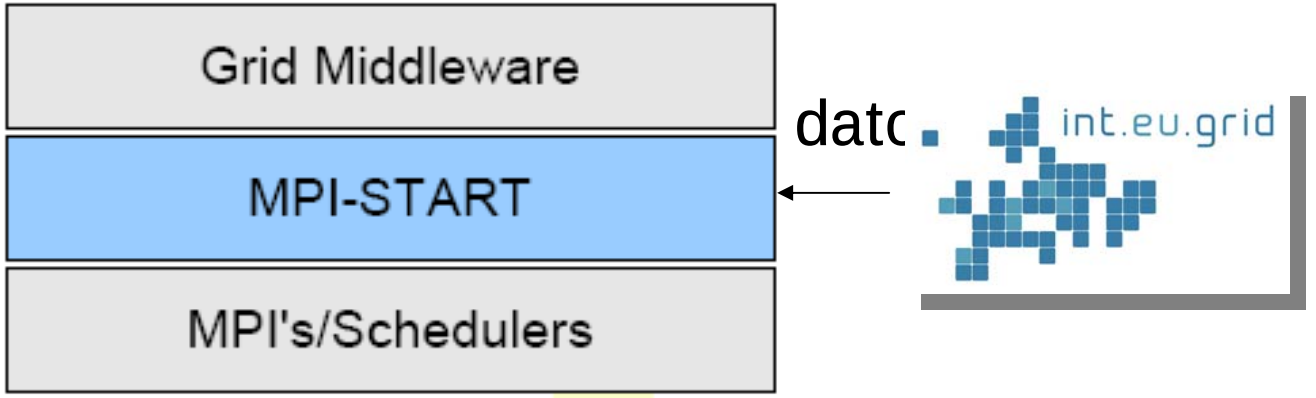
```
nodo1 1
nodo2 1
nodo3 1
nodo4 1
```

Diseño de una capa de software intermedio
Objetivos

CSIC

MPI-START

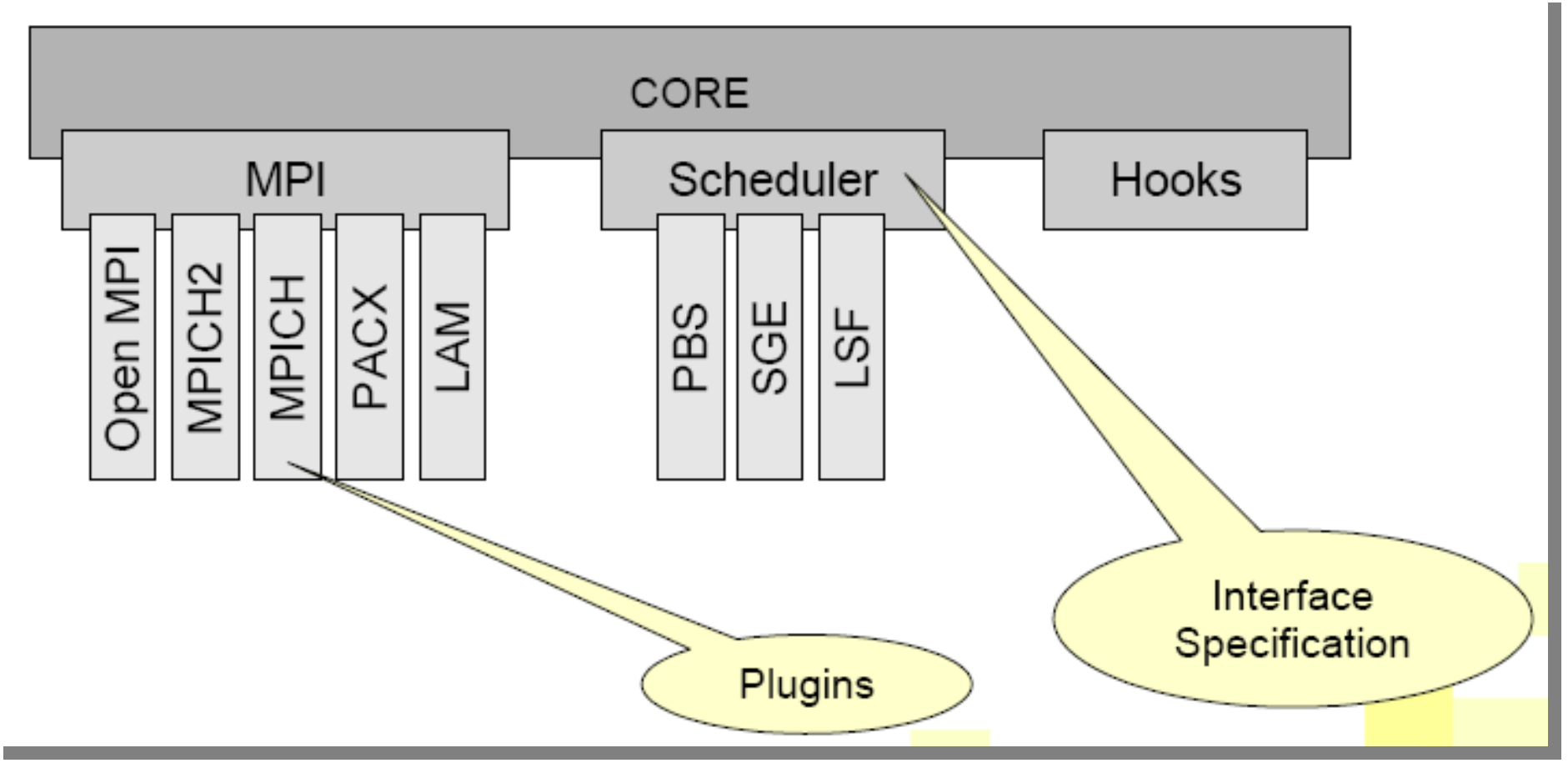
- Especificar un interface único a la capa superior de middleware para describir un trabajo MPI
- Ser capaz de dar soporte a implementaciones MPI distintas y nuevas, sin tener que cambiar el middleware del Grid
- Soportar las operaciones básicas de distribución de ficheros
- Dar soporte post-run



Consideraciones de diseño de mpi-start

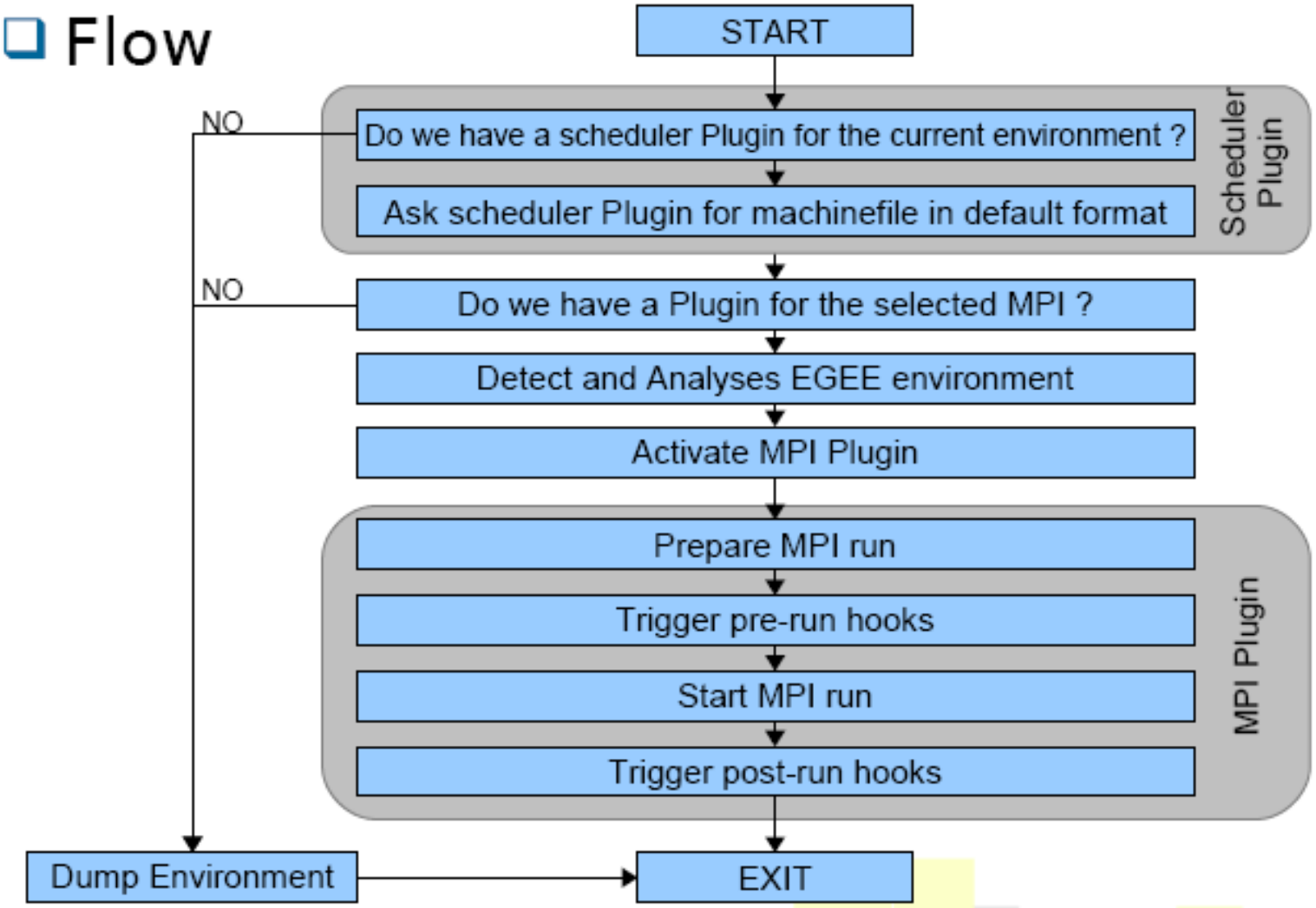
- Portable
  - MPI-START debe ser capaz de ejecutarse bajo cualquier sistema operativo que soporte el middleware
    - **Script en bash**
- Arquitectura modular y extensible
  - Instalable como un Plugin
  - Independiente de path absolutos para poder adaptarse a las distintas configuraciones locales de los site
- Posibilidad de “inyección remota” con el trabajo
  - Dar al usuario cierta potencia de trabajo independiente del site
- Opciones de debug remoto avanzadas

Arquitectura de mpi-start

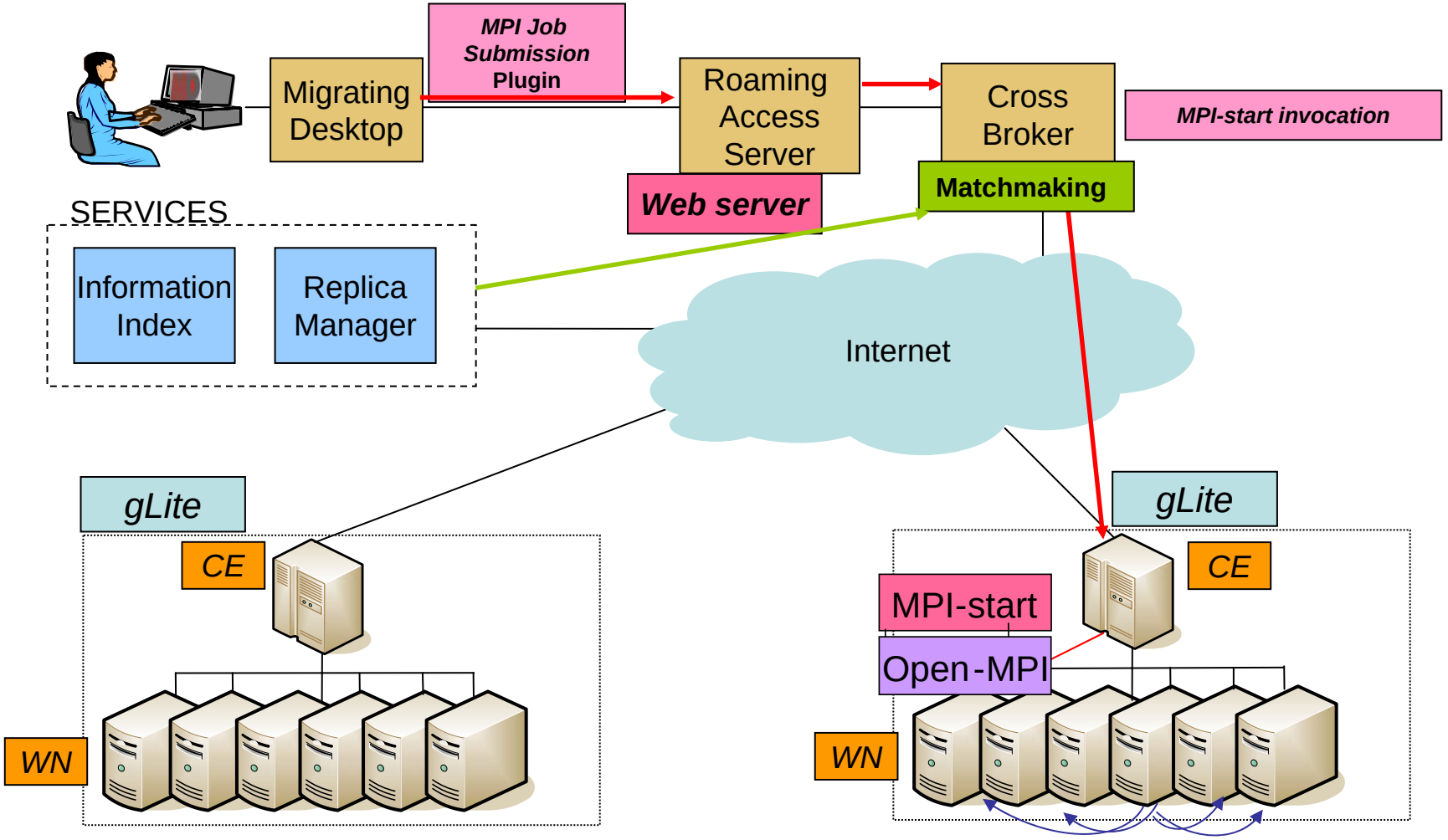


Arquitectura de mpi-start

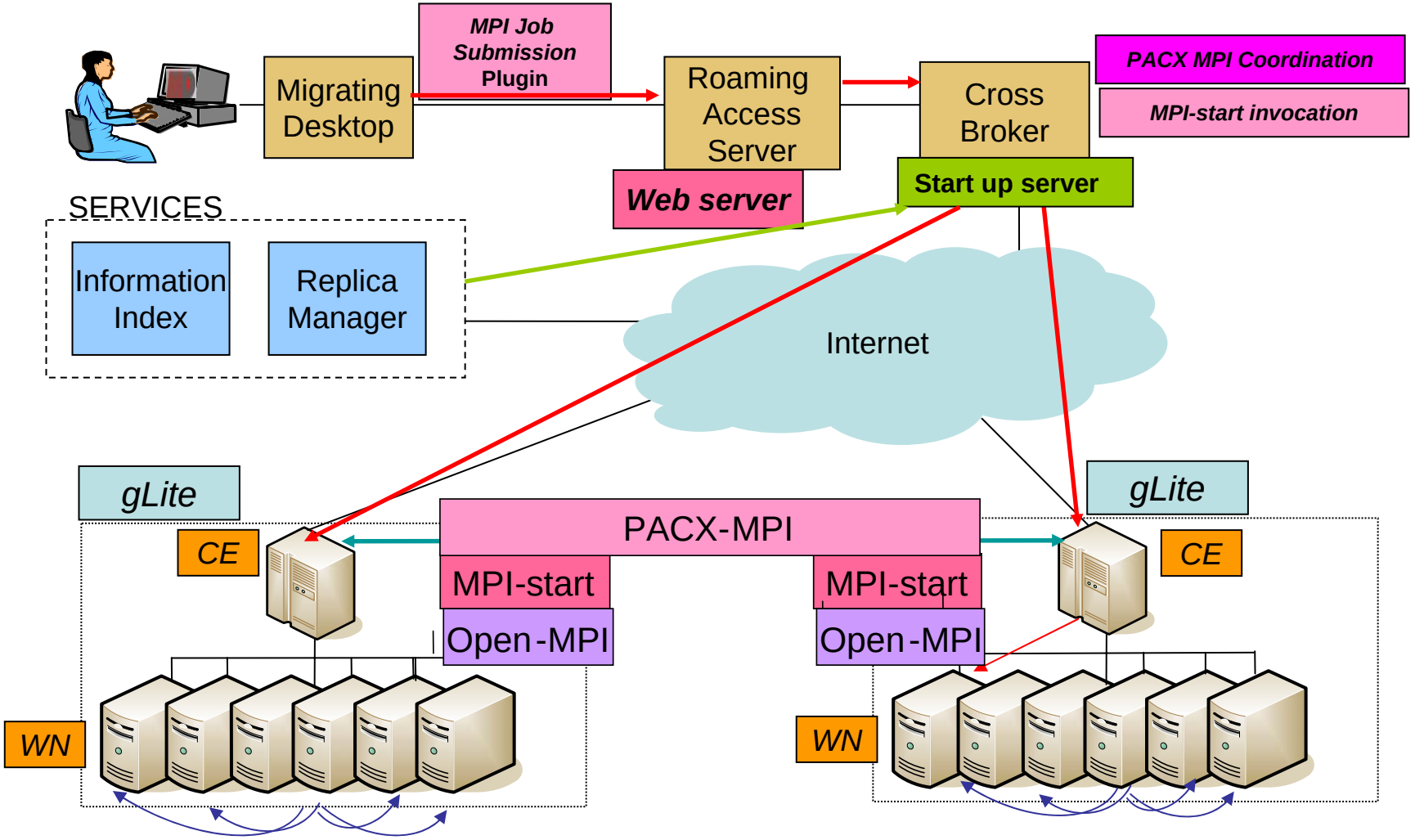
- Flow



Funcionamiento de MPI-START intra-cluster



Funcionamiento de MPI-START inter-cluster



Ejemplo básico

- hostname.jdl

```
Executable        = "/bin/hostname";
JobType           = "Parallel";
SubJobType        = "openmpi";
NodeNumber        = 4;
StdOutput         = "std.out";
StdError          = "std.err";
OutputSandbox     = {"std.out","std.err"};
```

stdout.txt

```
wn12-ieg.bifi.unizar.es
wn12-ieg.bifi.unizar.es
wn11-ieg.bifi.unizar.es
wn11-ieg.bifi.unizar.es
```

El Benchmark IMB

- imb.jdl

```
Executable          = "IMB-MPI1";
Arguments           = "barrier";
JobType             = "openmpi";
NodeNumber          = 4;
StdOutput           = "std.out";
StdError            = "std.err";
OutputSandbox       = {"std.out","std.err"};
InputSandbox        = {"IMB-MPI1"};
```

El Benchmark IMB

- stdout.txt

```
#---------------------------------------------------
#    Intel (R) MPI Benchmark Suite V2.3, MPI-1 part
#---------------------------------------------------
# Date        : Mon Aug 13 13:05:03 2007
# Machine     : i686# System      : Linux
# Release     : 2.4.21-47.0.1.EL.cernsmp
# Version     : #1 SMP Thu Oct 19 16:35:52 CEST 2006

...

#---------------------------------------------------
# Benchmarking Barrier
# #processes = 2
# ( 2 additional processes waiting in MPI_Barrier)
#---------------------------------------------------
 #repetitions  t_min[usec]  t_max[usec]  t_avg[usec]
         1000        11.43        11.43        11.43

#---------------------------------------------------
# Benchmarking Barrier
# #processes = 4
#---------------------------------------------------
 #repetitions  t_min[usec]  t_max[usec]  t_avg[usec]
         1000       187.78       187.88       187.83
```

Opciones de manejo de input/output

- o3.jdl

```
JobType              = "openmpi";
NodeNumber           = 8;
VirtualOrganisation = "imain";
Executable           = "o3sg_8";
StdOutput            = "std.out";
StdError             = "std.err";
InputSandbox         = {"o3sg_8","o3_hooks.sh","input.8"};
OutputSandbox        = {"std.out","std.err"};
Environment  = {"I2G_MPI_PRE_RUN_HOOK=./o3_hooks.sh",
                "I2G_MPI_POST_RUN_HOOK=./o3_hooks.sh"};
```

Opciones de manejo de input/output

MPI-START copia automaticamente el input necesario para la Simulación a todos los nodos con procesos MPI

```
pre_run_hook () {
}

# the first paramter is the name of a host in the
copy_from_remote_node() {

    if [[ $1 == `hostname` || $1 == 'hostname -f' || $1 == "localhost" ]]; then
        echo "skip local host"
        return 1
    fi

    # pack data
    CMD="scp -r $1:\"$PWD/$OUTPUT_PATTERN\" ."
    echo $CMD
    $CMD
}

...
```

Opciones de manejo de input/output

MPI-START copia automaticamente el output local
Si existe proveniente de los procesos individuales
Al nodo que contiene el proceso Master MPI

```
...

post_run_hook () {
    echo "post_run_hook called"

    if [ "x$MPI_START_SHARED_FS" == "x0" ] ; then
        echo "gather output from remote hosts"
        mpi_start_foreach_host copy_from_remote_node
    fi

    ls -al

    echo "pack the data"
    tar cvzf $OUTPUT_ARCHIVE $OUTPUT_PATTERN
    echo "upload the data"
    lcg-cr --vo $OUTPUT_VO -d $OUTPUT_HOST -l $OUTPUT_SE/$OUTPUT_ARCHIVE
file://$PWD/$OUTPUT_ARCHIVE

    return 0
}
```

El usuario puede indicar a que Storage Element quiere que se copie el resultado de la simulación

Ejemplo: GROMACS

GROMACS es un paquete Química Computacional

- Se puede compilar con **Open MPI** como implementación MPI

usar

```
JobType = "Parallel";
SubJobType = "OpenMPI";
NodeNumber = 4;
VirtualOrganisation = "icompchem";
Executable = "mdrun";
Arguments ="-v -s full -e full-o full -c after_full -g flog -N 4";
StdOutput = "std.out"; StdError = "std.err";
InputSandbox =
        {"speptide.top","after_pr.gro","full.mdp","gromacs_hooks.sh"};
OutputSandbox = {"std.out","std.err"};
Environment =
{"I2G_MPI_PRE_RUN_HOOK=./gromacs_hooks.sh","I2G_MPI_POST_
RUN_HOOK=./gromacs_hooks.sh"};
```

```
export OUTPUT_ARCHIVE=output.tar.gz
export OUTPUT_HOST=se.i2g.cesga.es
export OUTPUT_SE=lfn:/grid/icompchem/test
export OUTPUT_VO=icompchem

pre_run_hook ()
{
### Here comes the pre-mpirun actions of gromacs export
PATH=$PATH:/$VO_ICOMPCHEM_SW_DIR/gromacs-3.3/bin
grompp -v -f full -o full -c after_pr -p speptide -np 4
}

post_run_hook ()
{
echo "post_run_hook called"
echo "pack the data and bring it to an Storage Element"
tar cvzf $OUTPUT_ARCHIVE *
lcg-cr --vo $OUTPUT_VO -d $OUTPUT_HOST -l
$OUTPUT_SE/$OUTPUT_ARCHIVE file://$PWD/$OUTPUT_ARCHIVE
}
```

MPI support in EGEE

Soporte a MPI en gLite

CSIC

No se puede hacer MPI en el Grid!

¿Porqué la gente piensa esto?

1. Soporte en el WMS y API no es lo suficientemente flexible

   - Incompatible con la configuración standard de los sites

2. Falta de soporte para configurar un site con MPI
   - Muy pocos sites lo soportan
   - Los sites HEP tienen poco background MPI y no lo soportan

- No hay obstáculos técnicos inabordables, sólo pequeños problemas pero con un impacto elevado en lo que el usuario experimenta

Soporte a MPI en gLite WMS

- Hay un MPICH job type en el gLite WMS
  - Permite solicitar múltiples nodos para un mismo trabajo
  - Traduce el número de nodos al Globus RSL
  - Hace un wrap del binario del usuario para llamar a mpirun
  - Añade “MPICH” y no. cpus al ClassAd del trabajo
- Asume (hard-coded) site configuration
  - mpirun está ya fuera de servicio en muchos sites
- Conjunto restringido de jobmanagers
  - Sólo soporta “pbs” and “lsf”
  - No “lcgpbs”, “torque”, “lcglsf”, etc.
  - Descarta de partida el > 80% de EGEE

lcgpbs	16
pbs	47
lcglsf	27
sge	14
pbspro	8
lcgcondor	6
lcgsge	3
lsf	3
condor	2
Total	278

Jobmanager types en sitios que publican MPICH (9/5/08)

Prescripción de gLite para soportar MPI

- Si quieres soportar MPI, publica la “MPICH” tag
- Ya está !
- Los sites que quieren soportar MPI necesitan hacer el trabajo de los detalles por su cuenta
  - Coste muy alto para introducir recursos MPI en una infraestructura
  - Sólo los sites con una comunidad local muy interesada van a intentar dar soporte MPI

Problemas con las aproximaciones hard coded

- Para cada nueva implementación o versión el middleware tiene que ser modificado
- El middleware modificado hay que compilarlo de nuevo y construir el entorno de instalación adecuado (varios días de trabajo)
- El cambio en el middleware tiene que pasar todo el ciclo
  - Test + Validation
  - Testbed Release
  - Mínimo, 8 meses en EGEE
- Problemas no solubles, como combinaciones de schedulers Grid e implementaciones MPI que no funcionan juntos
  - Como encontrar el hostfilehow to find the hostfile
  - Ej. El formato en SGE del machinefile no esta soportado en mpich

EGEE/int.eu.grid reunión de integración (Dublin, Marzo 1007)

- En el marco del Technical Coordination Group on MPI
- Recomendaciones
  - Que NO haya ningun tipo especial de job para MPI
  - Solo jobs que admitan multiples nodos
  - El usuario inluye un wrapper script junto con el trabajo que hace el setup del MPI
  - Este wrapper puede incluso compilar la aplicación
  - Evitar asunciones sobre cual es el setup del site
- Acordar los principios de configuracion para los sites que quieran soportar MPI
  - **TAGS para el sistema de información**
- Usar mpi-start para simplificar la vida al usuario
  - Detalles de MPI setup, jobmanagers, etc

Cómo configurar un site de EGEE para soportar MPI en un site

CSIC

- Instrucciones: http://www.grid.ie/mpi/wiki
  - Instalar MPI-START
  - Publicar los TAGS necesarios en el infosys como GlueHostApplicationSoftwareRunTimeEnvironment
    - Software: MPI-START
    - Implementaciones: MPICH, MPICH2, OPENMPI ó LAM
    - Interconexion: MPI-Infiniband
- La receta definitiva para ejecutar MPI en EGEE

http://egee-uig.web.cern.ch/egeeuig/production_pages/MPIJobs.html

- Para encontrar los sites que soportan mpi-start añade esto a la los requerimientos en el JDL

Member("MPI-START", other.GlueHostApplicationSoftwareRunTimeEnvironment)

Job submission

```
JobType = "MPICH";
NodeNumber = 8;
Executable = "mpi-start-wrapper.sh";
Arguments = "mpi-test OPENMPI";
InputSandbox = {"mpi-start-wrapper.sh","mpi-hooks.sh","mpi-
test.c"};
Requirements = Member("OPENMPI",
other.GlueHostApplicationSoftwareRunTimeEnvironment);
```

JDL

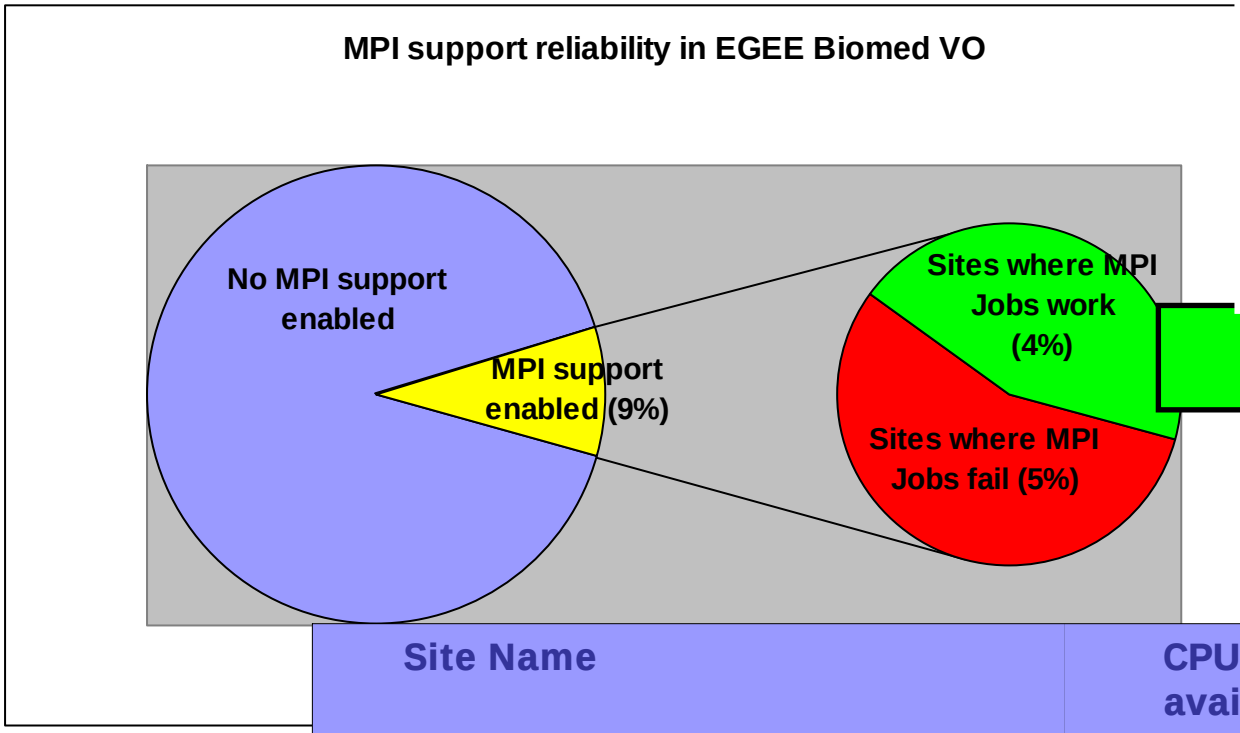
```
# Setup for mpi-start.
export I2G_MPI_APP=$MY_EXECUTABLE
export I2G_MPI_TYPE=$MPI_FLAVOUR
export I2G_MPI_PRE_RUN_HOOK=mpi-hooks.sh
export I2G_MPI_POST_RUN_HOOK=mpi-hooks.sh
# Invoke mpi-start.
$I2G_MPI_START
```

wrapper

```
pre_run_hook () {
mpicc -o ${I2G_MPI_APP} ${I2G_MPI_APP}.c
}
```

hooks

Soporte a MPI en EGEE



9% de los sites en Biomed publican el tag MPICH.
de ellos, sólo en la mitad funcionan los trabajos MPI, porque usan mpi-start

Site Name	CPUs available	Site Interconnect (from gridice)
egeece01.ifca.es	340	Gigabit Ethernet
ce.grid.rug.nl	120	Gigabit Ethernet
cirigridce01.univ-bpclermont.fr	60	Gigabit Ethernet
grid10.lal.in2p3.fr	900	Gigabit Ethernet
gridgate.cs.tcd.ie	770	Gigabit Ethernet
grive11.ibcp.fr	22	Gigabit Ethernet
marce01.in2p3.fr	210	Gigabit Ethernet

Ejercicios Grid MPI en EGEE

Isabel Campos
IFCA-CSIC

0: Sistema de Información

CSIC

1. Averiguar que CEs soportan actualmente mpi-start
2. Obtener una lista de CEs que soporten OPENMPI
3. Obtener una lista de implementaciones MPI instaladas en el CE gridgate.cs.tcd.ie
4. Averiguar que sites soportan MPI, pero no la nueva configuración con mpi-start

1: En los worker nodes

CSIC

1. Averiguar is las librerías de mpich2 están instaladas en el site de GRIF en LAL (grid10.lal.in2p3.fr)
2. Averiguar si el site del ifca admite el envío de trabajos usando mpiexec (egeece01.ifca.es)
3. Mirad a ver que más podeis encontrar sobre la instalación de MPI en el site!

2: Ejecutar un trabajo MPI con mpi-start

CSIC

1. Ejecutar la aplicación MPI mpi-test.c en el Grid usando MPI-START

http://wwwiscampos.ifca.es/users/iscampos/docs/GridsEciencia/mpi-job.tgz

- Para ello crear:
  - Un fichero JDL para enviar el trabajo a un site que soporte ambos, MPI-START y cualquier implementación MPI
  - Un fichero de hooks para MPI-START que compile la aplicación
  - Un wrapper MPI-STARTpara ejecutar la aplicación

3. MPI + almacenamiento Grid

1. Enviar un trabajo al grid que descargue el código fuente de la aplicación de un storage element del grid como primer paso. Posteriormente la compila y despues la ejecuta

PASOS:

- Copiar el fichero tar, mpi-job.tgz a un storage element
- Crear un script para descargar la aplicación del SE y ejcutarla
- Crear el JDL para enviar el trabajo
- Modificar la aplicacion MPI para escribir los datos a un fichero, y a continución copialros a un SE cuando la aplicación acaba
- Verificar que el resultado es el correcto copiandolo del SE a una máquina local.

4: Ejecutar un trabajo MPI en el grid sin tener MPI-START instalado en el site

CSIC

- MPI-START se puede usar en sites que no lo tienen instalado si se envía en el input sandbox junto con la aplicación
- Se puede descargar de aqui:

http://i2gui01.ifca.es/users/iscampos/docs/mpi-start.tar.gz

- Modificar el wrapper para desempaquetar mpi-start y usarlo para testear la aplicacón en sites sin mpi-start instalado