Grid y Computación de Altas Prestaciones

GRyCAP

MPI en Infraestructuras Avanzadas Grid

Modelos de Programación en Grid (MPG)

Objetivos

Grid y Computación de Altas Prestaciones
GRyCAP

- Entender y comprender las dificultades para el lanzamiento de trabajos MPI en una infraestructura Grid Avanzada.
- Conocer los soportes que ofrece un Infraestructura Grid Avanzada para el lanzamiento de trabajos MPI.
- Diseñar e implementar trabajos MPI que puedan lanzarse en una infraestructura Grid Avanzada.
- Experimentar en una infraestructura Grid Real el lanzamiento de trabajos MPI.

Contenido

Grid y Computación de Altas Prestaciones
GRyCAP

- Introducción
  - Proyecto EGEE
    - Limitaciones gLite
  - HTC&HPC en el Grid
- Soporte a MPI en Grids

Introducción
Temas clave

Grid y Computación de Altas Prestaciones
GRyCAP

- Enabling Grids for e-Science (EGEE)
  - Middleware glite: http://glite.web.cern.ch/glite
- European Grid Infrastructure (EGI)
  - Infraestructura en producción http://www.egi.eu
- European Middleware Innitiative (EMI)
  - Integración de Arc, gLite y UNICORE en UMD, http://www.eu-emi.eu/
- Soporte a Aplicaciones Científicas en el Grid

Introducción

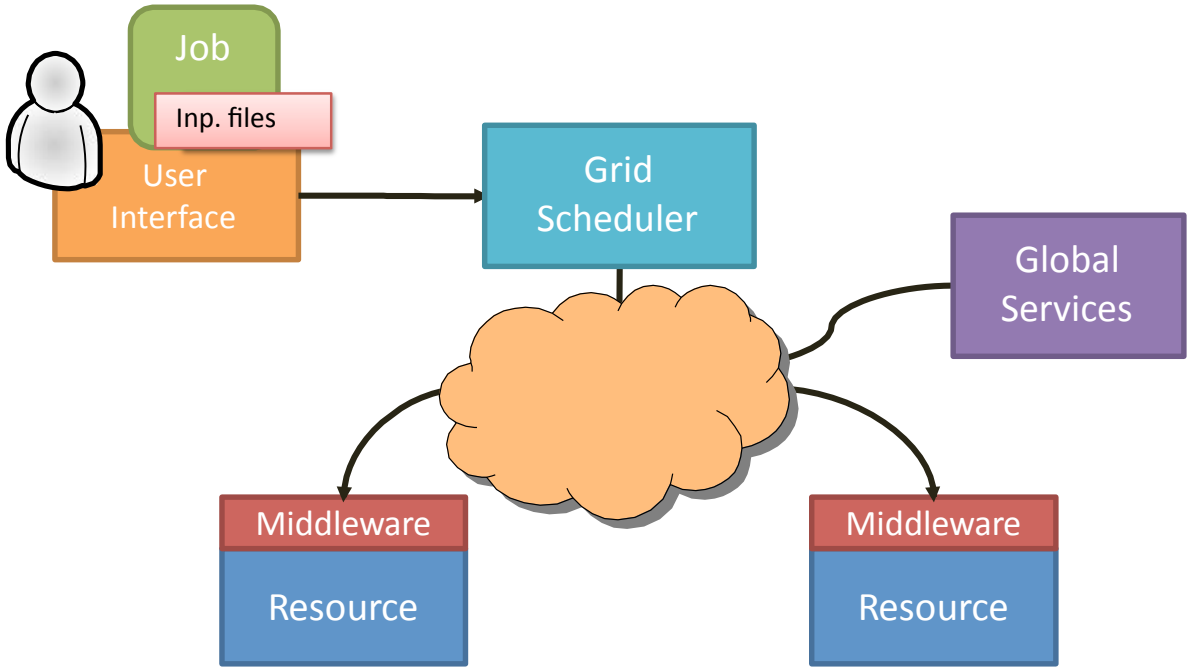
EGEE: Un Grid a la medida de HEP (High-energy physics)

Grid y Computación de Altas Prestaciones
GRyCAP

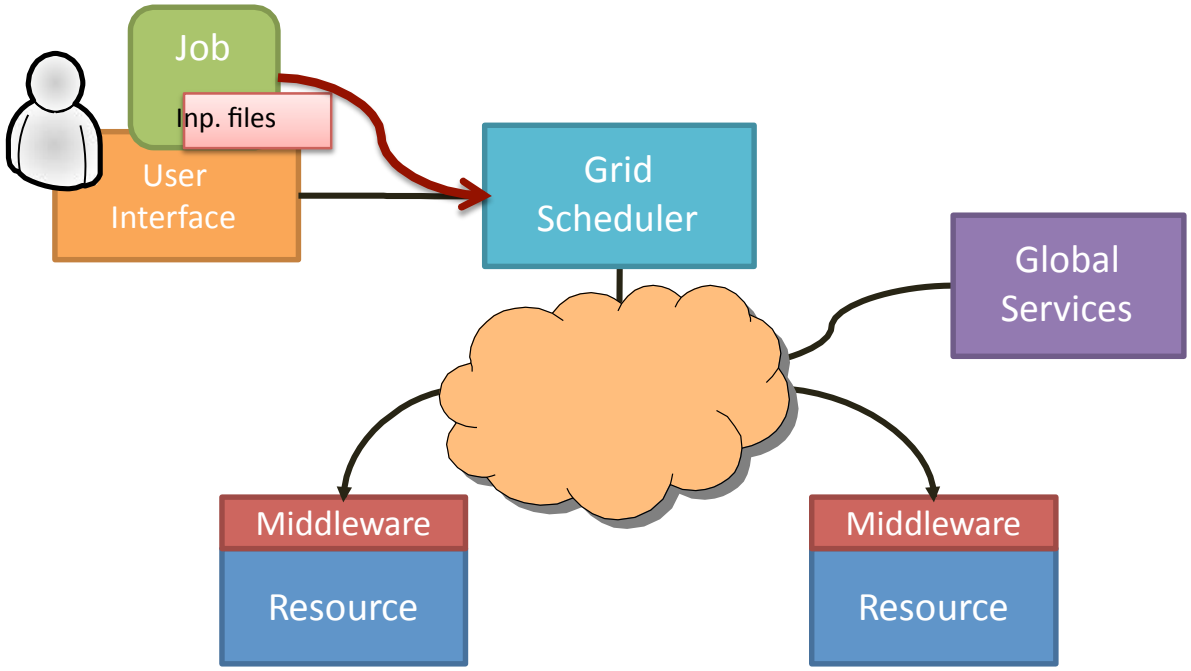
Claves

- Muchos (cientos) grupos de investigación distribuídos geográficamente
  - 20 años trabajando en modo “Grid” pero sin la tecnología Grid
- Análisis de grandes cantidades de datos provenientes de detectores en aceleradores de partículas
- Una comunidad muy motivada debido al exito de la *World Wide Web*

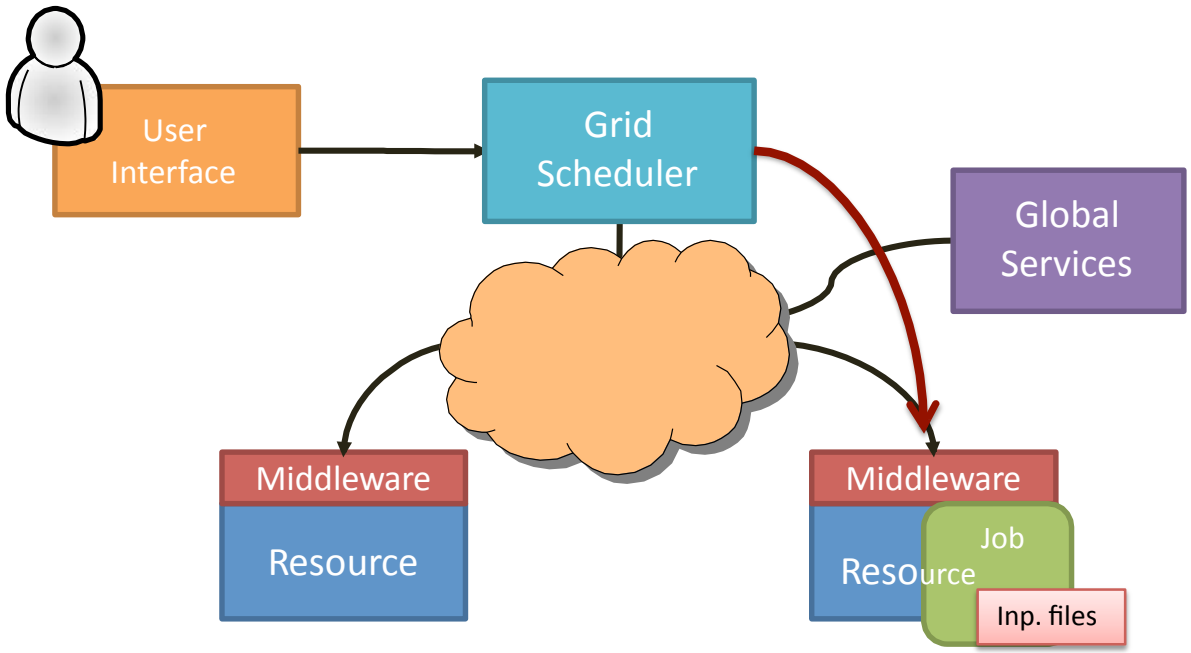
Introducción
Ejecución batch monoproceso en Grids (gLite)



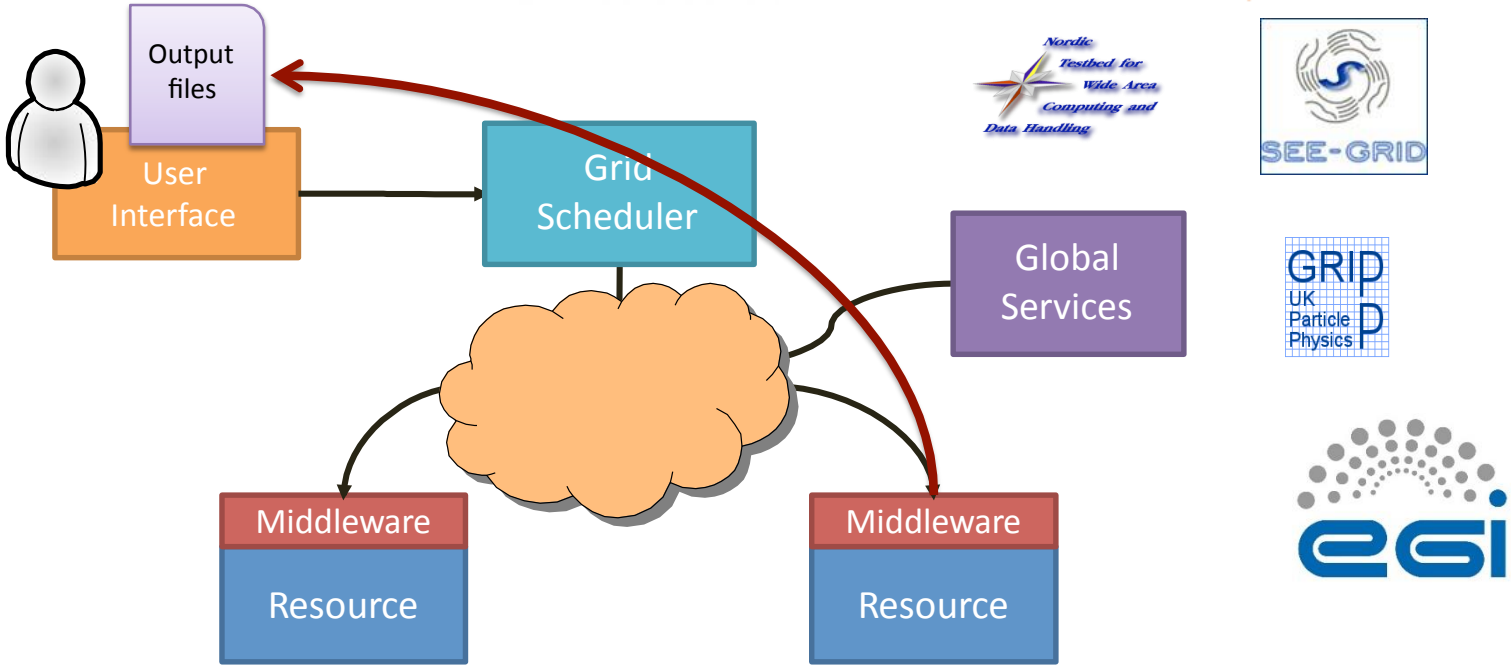
Introducción
Ejecución batch monoproceso en Grids (gLite)



Introducción
Ejecución batch monoproceso en Grids (gLite)



Introducción
Ejecución batch monoproceso en Grids (gLite)



Introducción
Limitaciones gLite

- **Limitaciones de la aproximación de glite**
  - Ejecución de **trabajos interactivos** con prioridad alta
    - EGEE está orientado a la producción masiva en modo batch
  - Falta **soporte a la ejecución MPI** (parallel computing)
    - Sólo trabajos mono-procesador en glite
  - Faltan **herramientas** visuali
- **Desa**
  - posbilidad de interactuar con
  - Integrar en el middleware Grid herramientas de visualización

Desarrollar las herramientas de middleware manteniendo la compatibilidad con gLite (!!)

Introducción
High Throughput Computing (HTC)

Grid y Computación de Altas Prestaciones
GRyCAP

- **El Grid** se diseñó con la idea de ser una fuente de HTC
- **High throughput computing (HTC)**

– Se utilizan muchos recursos computacionales durante largos periodos de tiempo

  - Acceso a mucho tiempo de cpu promedio durante largos periodos de tiempo (meses)
  - Optimizar el número de trabajos ejecutados por unidad de tiempo.
  - Computación en modo granja, o procesos independientes

Grid y Computación de Altas Prestaciones GRyCAP

- **Pero….**
- **Los usuarios necesitan más capacidad de computación**
  - Usando más de un core por ejecución
  - Usando más de un site por ejecución incluso
- **Ejecución paralela de aplicaciones**
  - Los **trabajos paralelos** usan más de un core
  - ¿Cómo usarlos de manera eficiente?
    - **Shared memory**: todos los cores acceden a un área común de la memoria para acceder a los datos
    - **Message Passing**: los cores se intercambian mensajes con los datos

Grid y Computación de Altas Prestaciones
GRyCAP

- **High performance computing (HPC)**
  - Disponer simultaneamente de una gran cantidad de recursos computacionales
  - Lo importante es que la aplicación se ejecute en el menor tiempo posible.
  - Para ello es necesario que los procesadores individuales que participan en el cálculo cooperen

MPI

Contenido

Grid y Computación de Altas Prestaciones GRyCAP

- Introducción
  - Proyecto EGEE
    - Limitaciones gLite
  - HTC&HPC en el Grid
- **Soporte a MPI en Grids**

Soporte a MPI en el Grid ¿Porqué?

Grid y Computación de Altas Prestaciones GRyCAP

- **Muchas áreas de aplicaciones requieren soporte a MPI**
  - Ciencias de la tierra, fusion, astrofísica, Química Computacional…
  - Se pueden obtener resultados significativos usando 10s-100s

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

- **Un soporte de calidad atraería comunidades al Grid**
  - Cómo una infraestructura en sí misma
  - Cómo testbed antes de ejecutar en máquinas HPC

Problemas a resolver

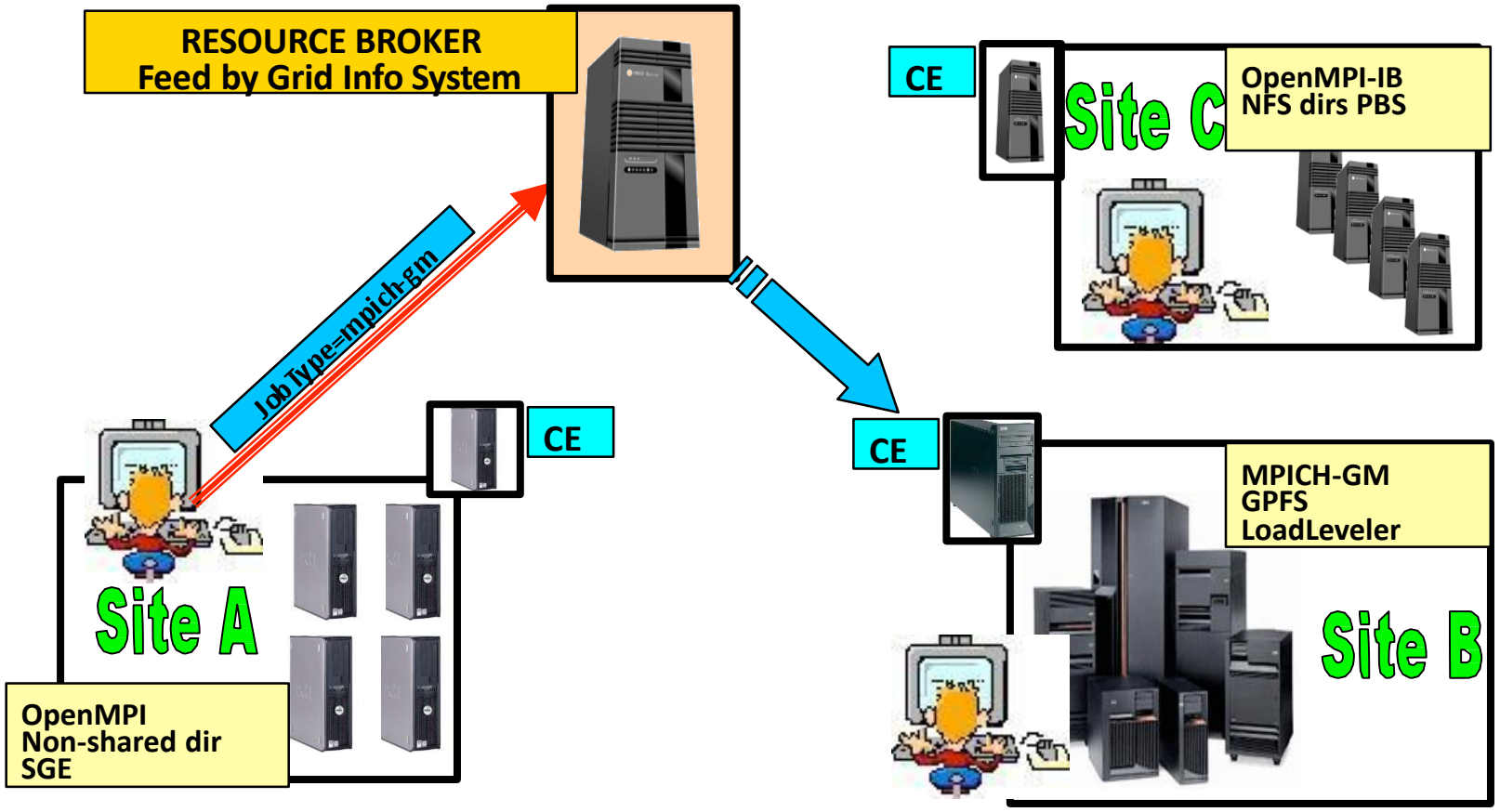
Grid y Computación de Altas Prestaciones GRyCAP

- El Grid No es un entorno homogéneo

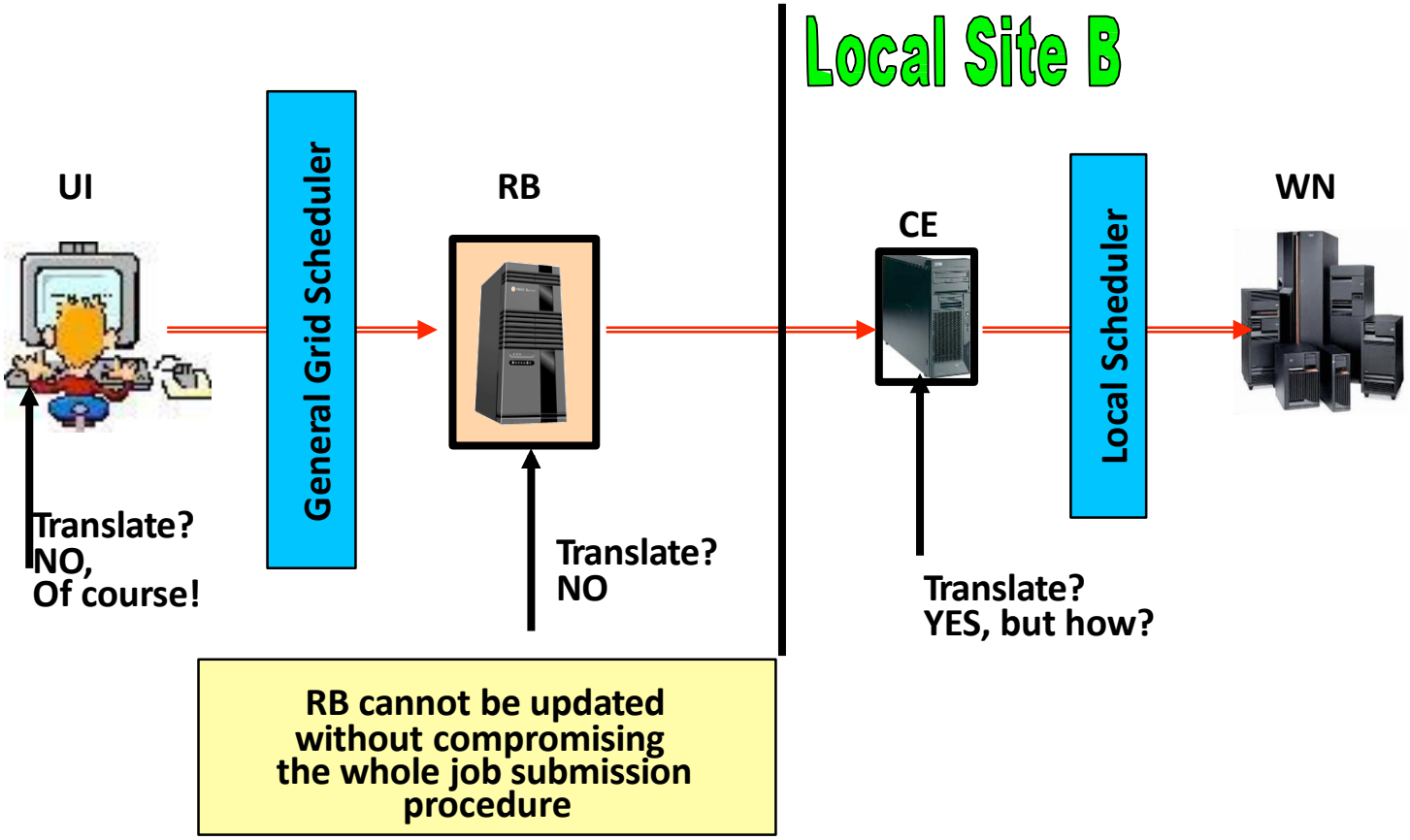
- Sistemas de ficheros no compartidos
  - Muchos sites no tienen soporte a sistemas de ficheros compartidos
  - Muchas implementaciones MPI esperan encontrar el ejecutable en el nodo donde se ejecuta el proceso
  - En general el setup es muy variado

- 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 tipica en el Grid



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



Ejemplo con Sun Grid Engine

Grid y Computación de Altas Prestaciones
GRyCAP

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
```

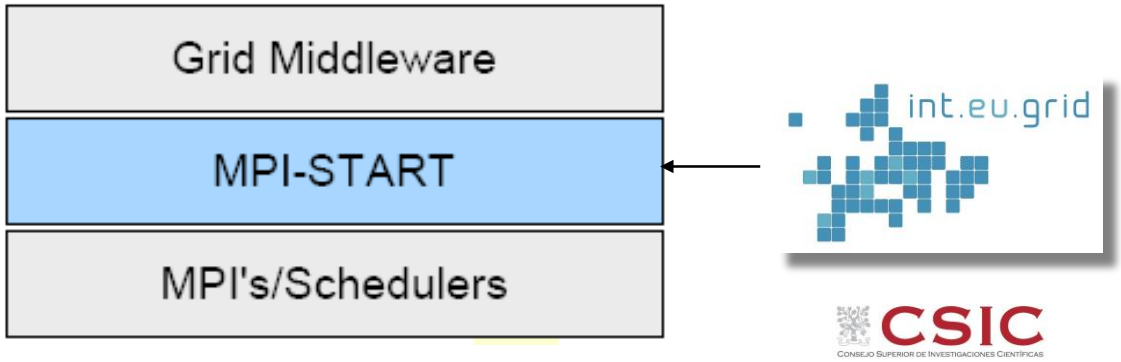
machinefile

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

Diseño de una capa de software intermedio: Objetivos

Grid y Computación de Altas Prestaciones GRyCAP

- 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 a usuario para manejar sus datos pre- y post-run

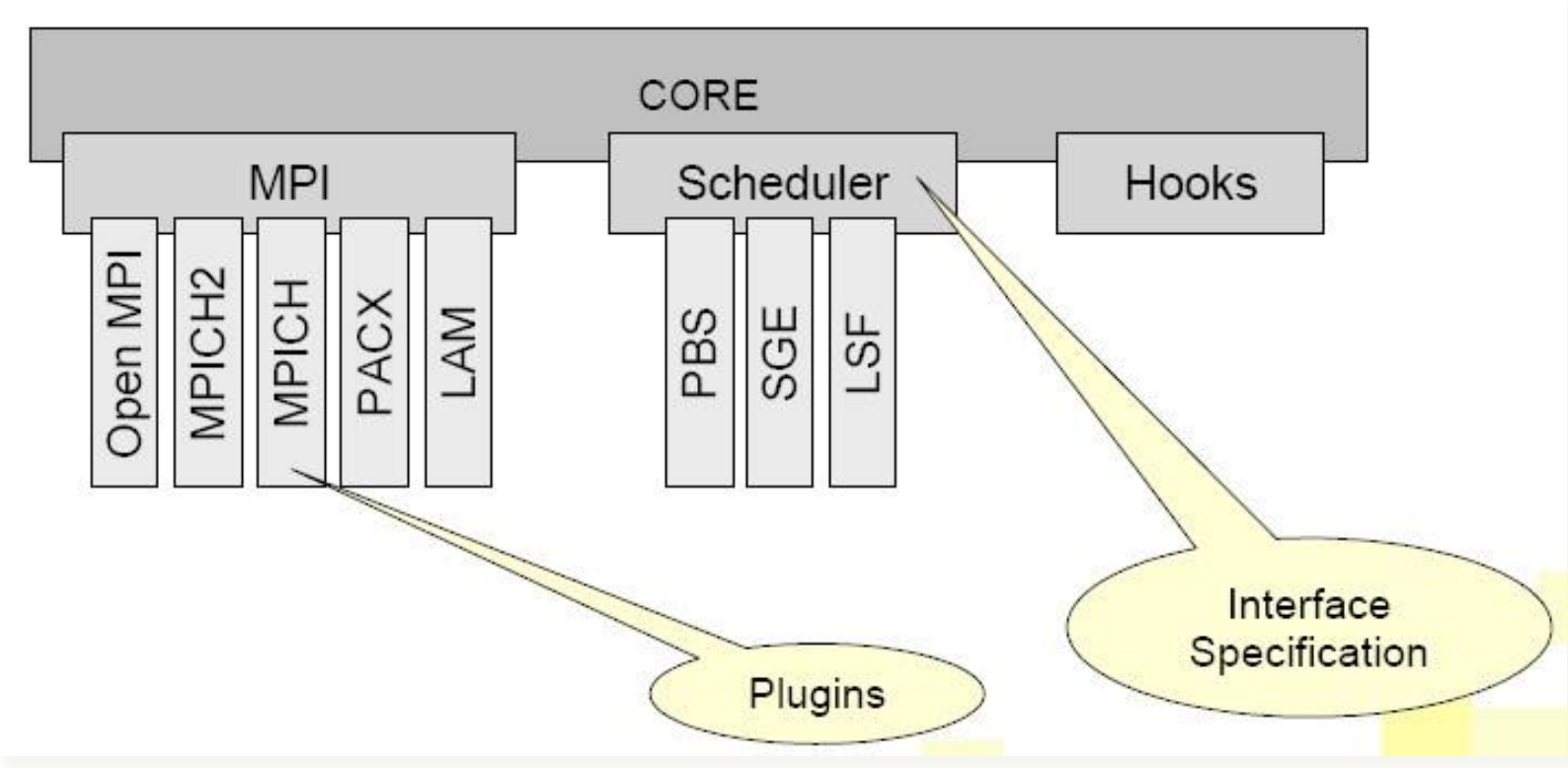


Consideraciones de diseño de mpi-start

Grid y Computación de Altas Prestaciones GRyCAP

- 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
- Opciones de debug remoto avanzadas

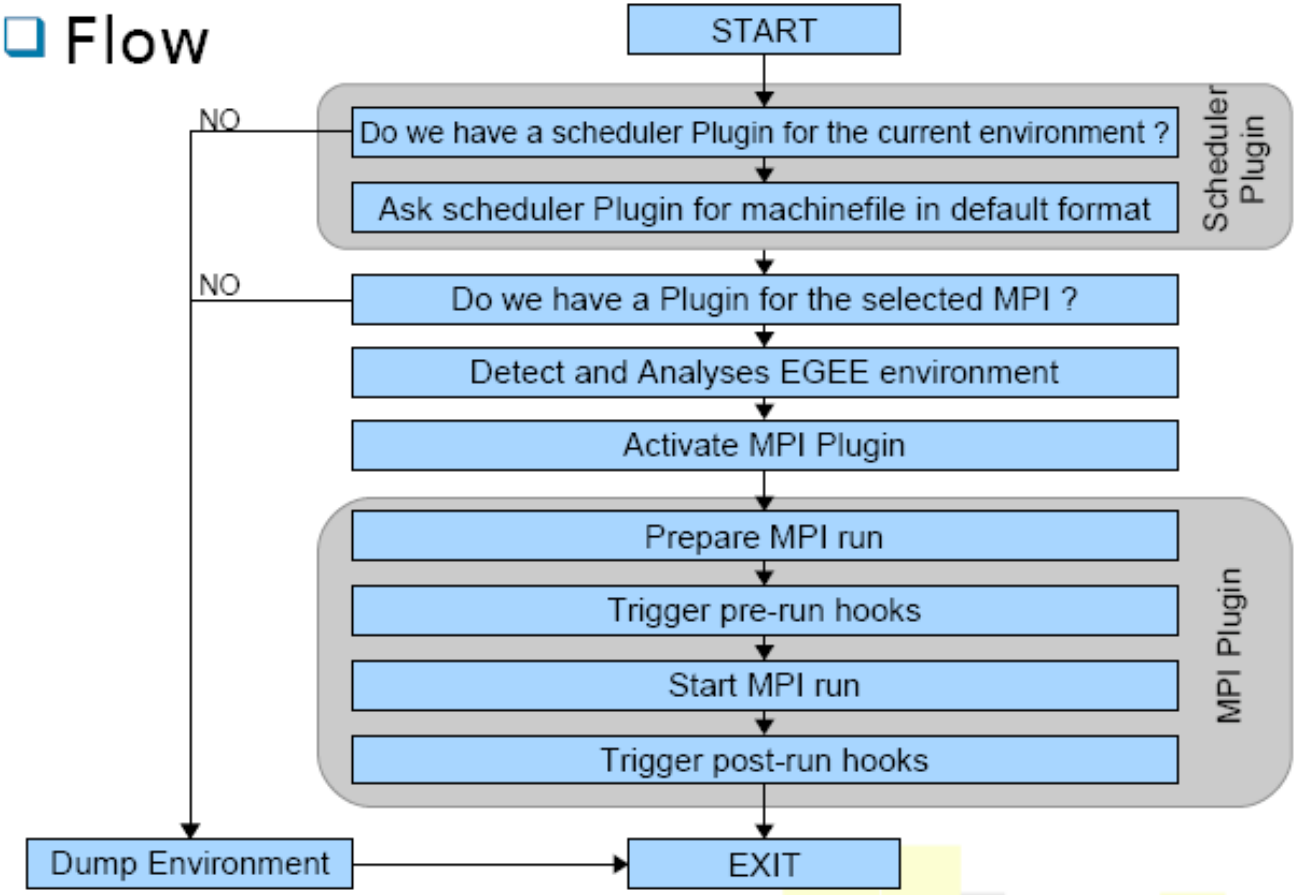
Arquitectura de mpi-start



Arquitectura de mpi-start

Grid y Computación de Altas Prestaciones
GRyCAP

- Flow



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

Grid y Computación de Altas Prestaciones
GRyCAP

- 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

hIp://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.GlueHostApplicationSoUwareRunTimeEnvironment)

Job submission

Grid y Computación de Altas Prestaciones GRyCAP

JDL

```
JobType = "normal";
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);
```

wrapper

```
# 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
```

hooks

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