Tecnologías Grid

gLite

Curso de Doctorado 2008-2009

Área de Arquitectura y Tecnología de Computadores
Universidad de Oviedo

gLite

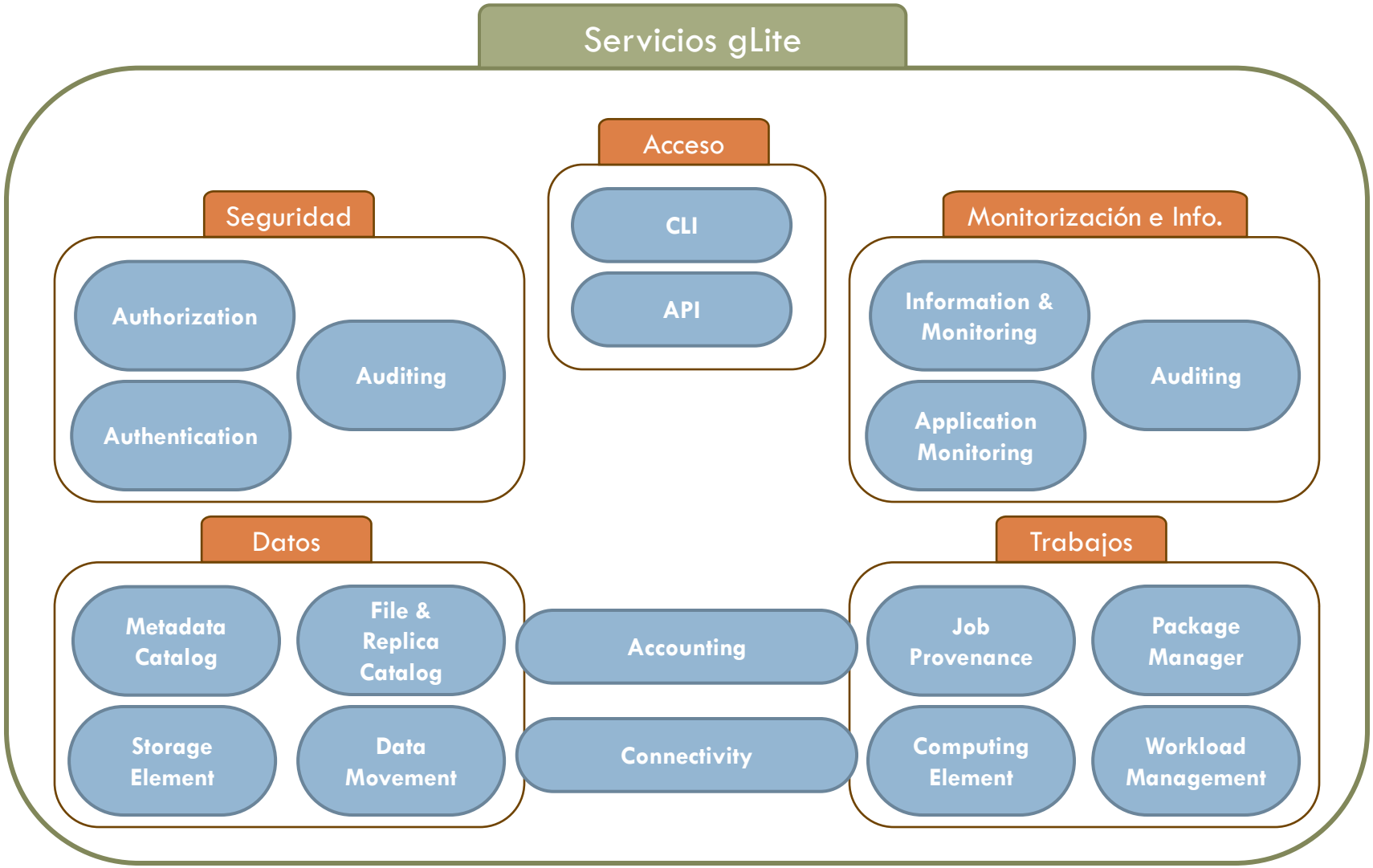
Introducción

Introducción

- EGEE (Enabling Grids for E-sciencE)
  - Mayor infraestructura grid del mundo
  - 120 organizaciones europeas, 47 paises, 68000 CPUs, 8000 usuarios, 150000 trabajos al día
- gLite
  - Middleware que da soporte a EGEE
  - Integra un conjunto de componentes para habilitar la compartición de recursos entre múltples organizaciones
  - Se basa en otros proyectos: Globus, Condor, LCG, …

Inicialmente la última E de EGEE significaba Europa

Introducción



gLite

Servicios de seguridad

Servicios de seguridad

- Autenticación basada en X.509
  - Las autoridades de certificación (CA) emiten certificados a los individuos
  - Para evitar vulnerabilidades, la identificación de los individuos se realiza mediante proxies
- Un proxy
  - Puede delegarse a otro servicio
  - Puede almacenarse externamente (MyProxy)
  - Puede incluir atributos adicionales (pertenencia a organizaciones)

Servicios de seguridad

- VOMS (Virtual Organization Membership Service)
  - Los certificados no son suficientes para definir las capacidades de un usuario del grid
  - VOMS proporciona un mecanismo para añadir atributos adicionales a un proxy
    - Los atributos proporcionan capacidades adicionales

```
subject   : /C=IT/O=GILDA/OU=Personal Certificate/L=GIJON/CN=GIJON01/CN=proxy
issuer    : /C=IT/O=GILDA/OU=Personal Certificate/L=GIJON/CN=GIJON01
identity  : /C=IT/O=GILDA/OU=Personal Certificate/L=GIJON/CN=GIJON01
type      : proxy
...
=== VO gilda extension information ===
VO        : gilda
subject   : /C=IT/O=GILDA/OU=Personal Certificate/L=GIJON/CN=GIJON01
issuer    : /C=IT/O=INFN/OU=Host/L=Catania/CN=voms.ct.infn.it
attribute : /gilda/Role=NULL/Capability=NULL
timeleft  : 11:18:08
```

gLite

Servicios de información

Servicios de información

- Objetivos de los servicios de información (IS):
  - Descubrir los recursos
  - Recopilar información del estado de los recursos
  - Proporcionar datos para gestionar la carga computacional y de datos de forma eficiente
- IS en gLite:
  - El modelo de datos se basa en el esquema GLUE (Grid Laboratory Uniform Environment)
  - La arquitectura utiliza BDII (Berkeley DB Information Index)

Servicios de información

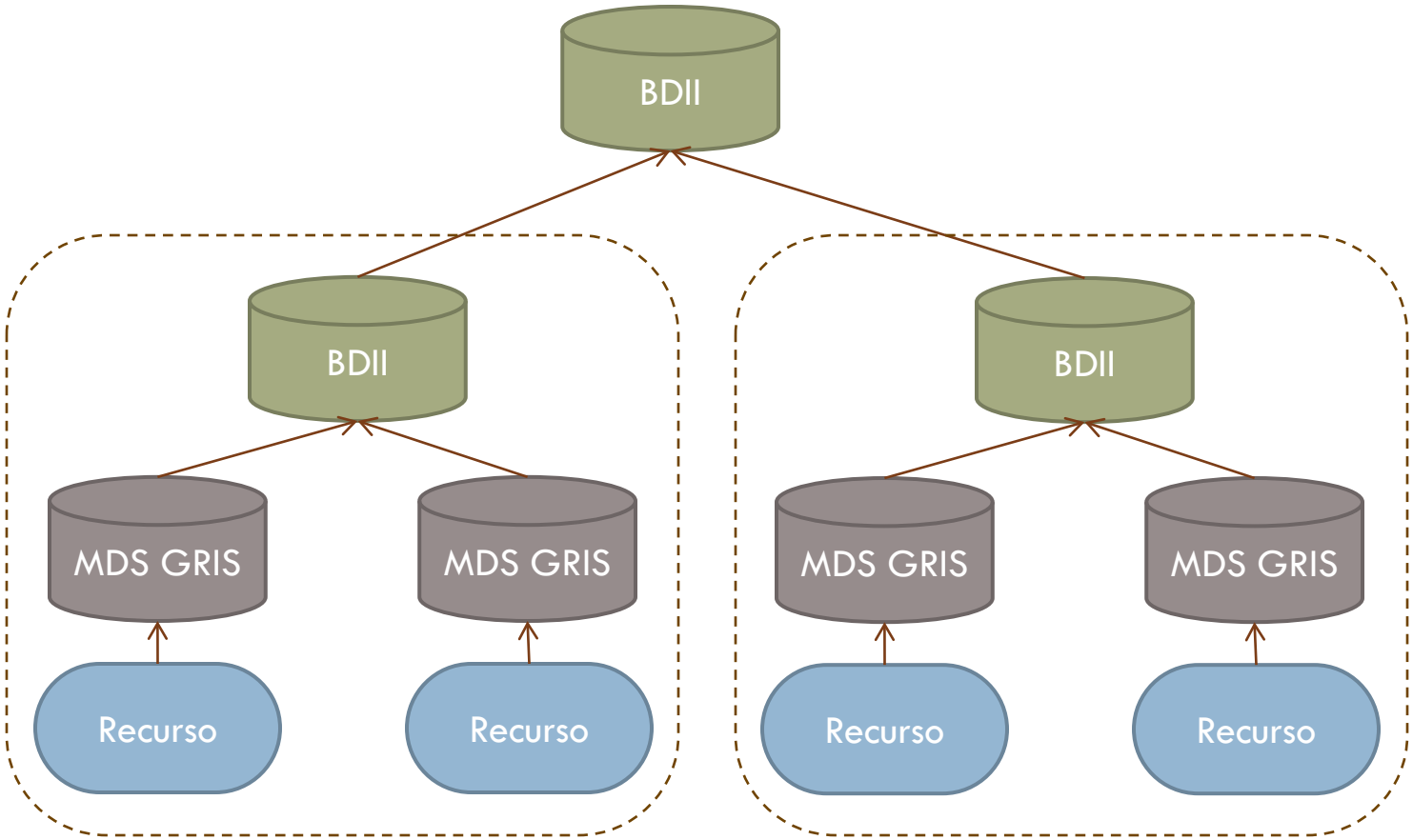
- Esquema GLUE:
  - Especificación sobre la información que puede ser publicada acerca de un grid
    - El objetivo es describir los recursos grid y sus atributos
  - La información se expresa de diversas formas:
    - LDAP, SQL, XML, ClassAd
  - Los elementos se organizan jerarquicamente:
    - Site, Cluster, Computing Element, Storage Element, etc.
  - Ejemplos de atributos:
    - GlueCEStateTotalCPUs, GlueCEStateFreeCPUs, GlueHostMainMemoryRamSize, etc.

Servicios de información

- Sistema de información:
  - Es una evolución del Globus MDS
    - Se basa en servidores Lightweight Directory Access Protocol (LDAP)
  - Componentes:
    - GRIS (Grid Resource Information Server): recopila información sobre los recursos locales
    - BDII: recopila información proporcionada por los GRIS
      - De forma periódica (cron) la información se transfiere entre BDIIs
  - Los usuarios u otros servicios pueden consultar al BDII de más alto nivel sobre el estado de cualquier recurso del grid

Servicios de información

- Arquitectura del sistema de información:



gLite

Servicios de gestión de datos

Servicios de gestión de datos

- Elementos de almacenamiento (SE):
  - Servicio que permite a los usuarios almacenar y acceder a información
    - SE = SRM + GridFTP + E/S
  - Protocolos de transferencia utilizados por los SE:
    - GSIFTP
      - ~GridFTP (estrictamente es un subconjunto del GridFTP)
    - GSIDCAP (GSI dCache Access Protocol)
      - Versión del protocolo dcap (nativo de dCache) que utiliza la seguridad GSI
    - RFIO/GSIRFIO (Remote File Input/Output protocol)
      - Para acceder a los archivadores de cinta

Grid de datos europeo

- Tipos de SE:
  - CASTOR
    - Consiste en un frontend que proporcionar una caché en disco a un sistema de almacenamiento masivo en cinta.
    - El proceso "stager" realiza la transferencia entre el disco y la cinta
  - dCache y DPM
    - Gestionan el almacenamiento distribuido en varios servidores de forma centralizada
    - Los discos se combinan formando un único sistema de ficheros virtual

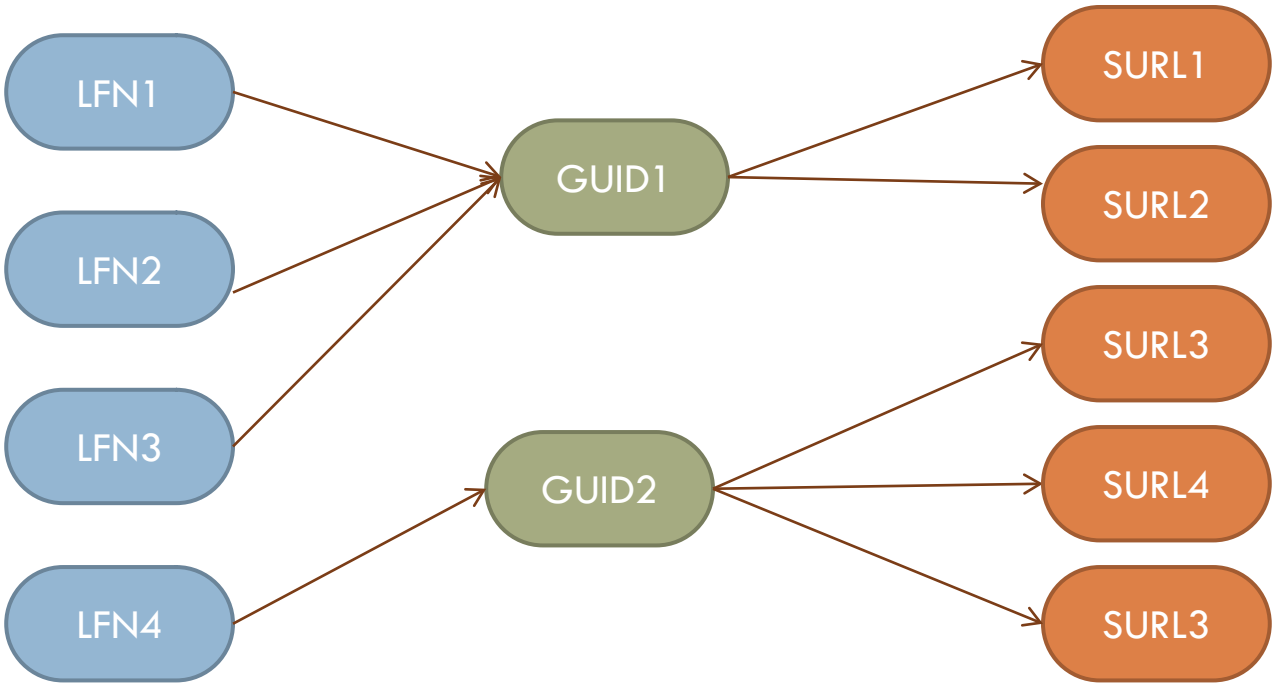
Todos proporciona un interfaz SRM

Grid de datos europeo

- Nombres de ficheros:
  - GUID (Grid Unique Identifier)
    - Identifica de forma univoca a un fichero
      - guid:38ed3f60-c402-11d7-a6b0-f53ee5a37e1d
  - LFN (Logical File Name)
    - Alias para referirse a un fichero (evita utilizar el GUID)
      - lfn:/grid/gilda/Datos/Dato1.txt
  - SURL (Storage URL)
    - Identifica una replica en el SE
      - srm://srm.cern.ch/castor/cern.ch/grid/dteam/doe/file1
  - TURL (Transport URL)
    - Punto de acceso temporal para un replica
      - gsiftp://tbed0101.cern.ch/data/dteam/doe/file1

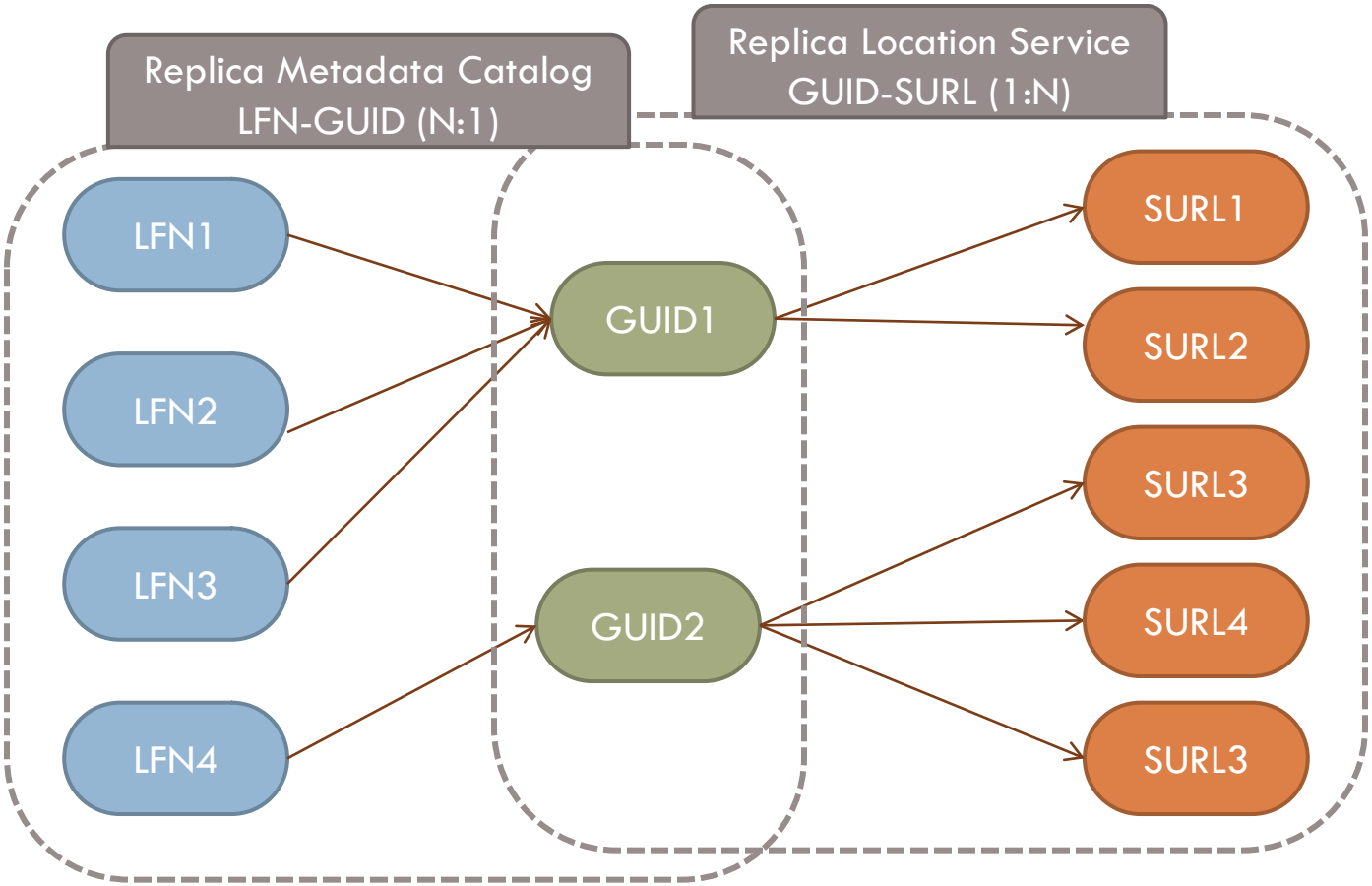
Servicios de gestión de datos

- Relaciones entre nombres de ficheros:



Servicios de gestión de datos

- Relaciones entre nombres de ficheros (LCG-2):



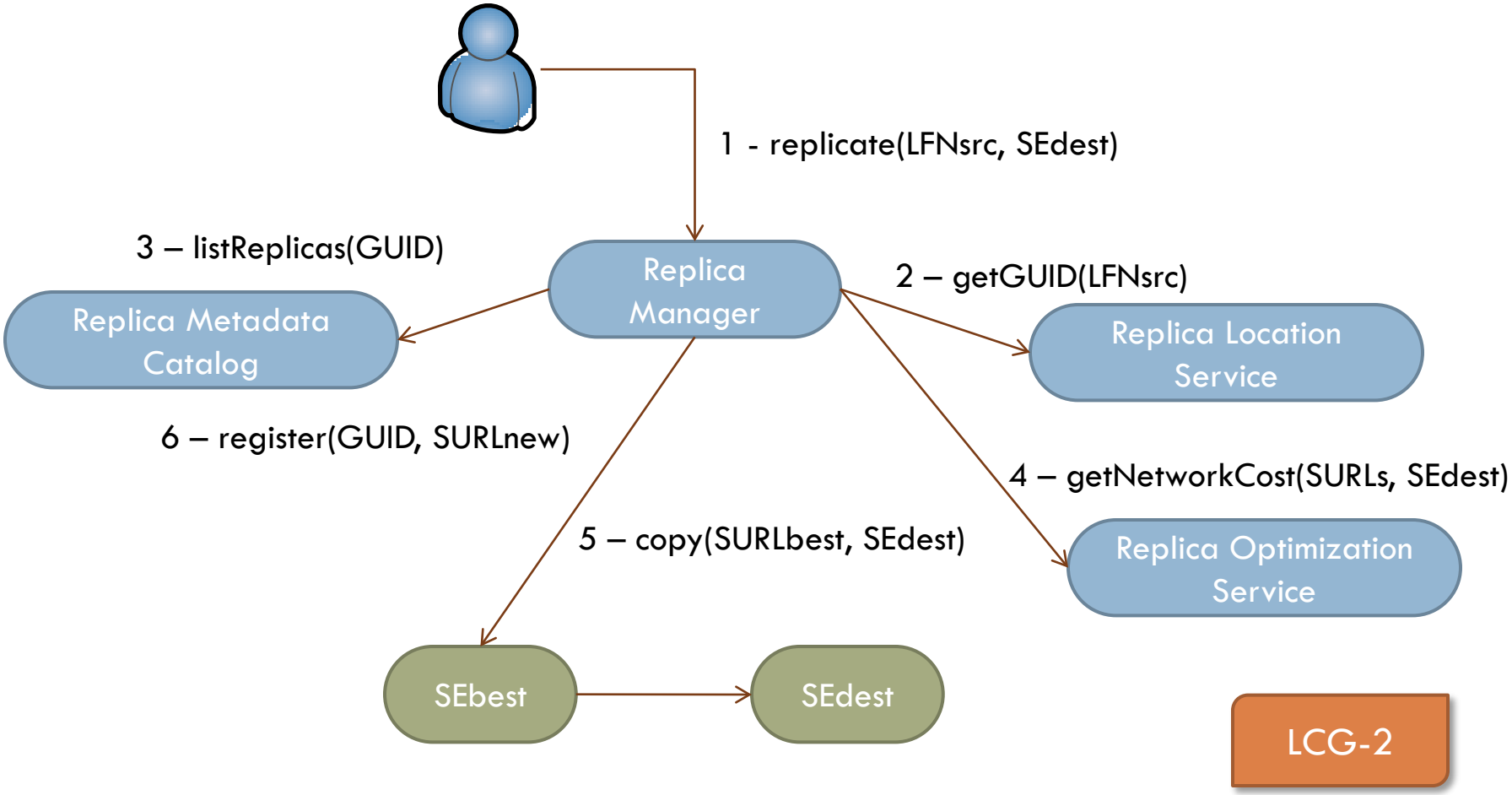
Servicios de gestión de datos

- Transferencia de un fichero (LFNsrc) a un SEdest:
  - Se pregunta al RMC (Replica Metadata Catalog) por el LFNsrc
    - Responde con su GUID
  - Se pregunta al RLS (Replica Location Service) por el GUID
    - Responde con una lista de SURLs
  - Se pregunta al ROS (Replica Optimization Service) por el coste de transferir de los SURLs a SEdest
    - En función de la respuesta se elige el mejor SURL
  - Se transfiere LFNsrc desde SURLbest a SEdest
  - Se registra el nuevo SURL en SEdest
    - Se añade un nuevo mapeo al GUID

LCG-2

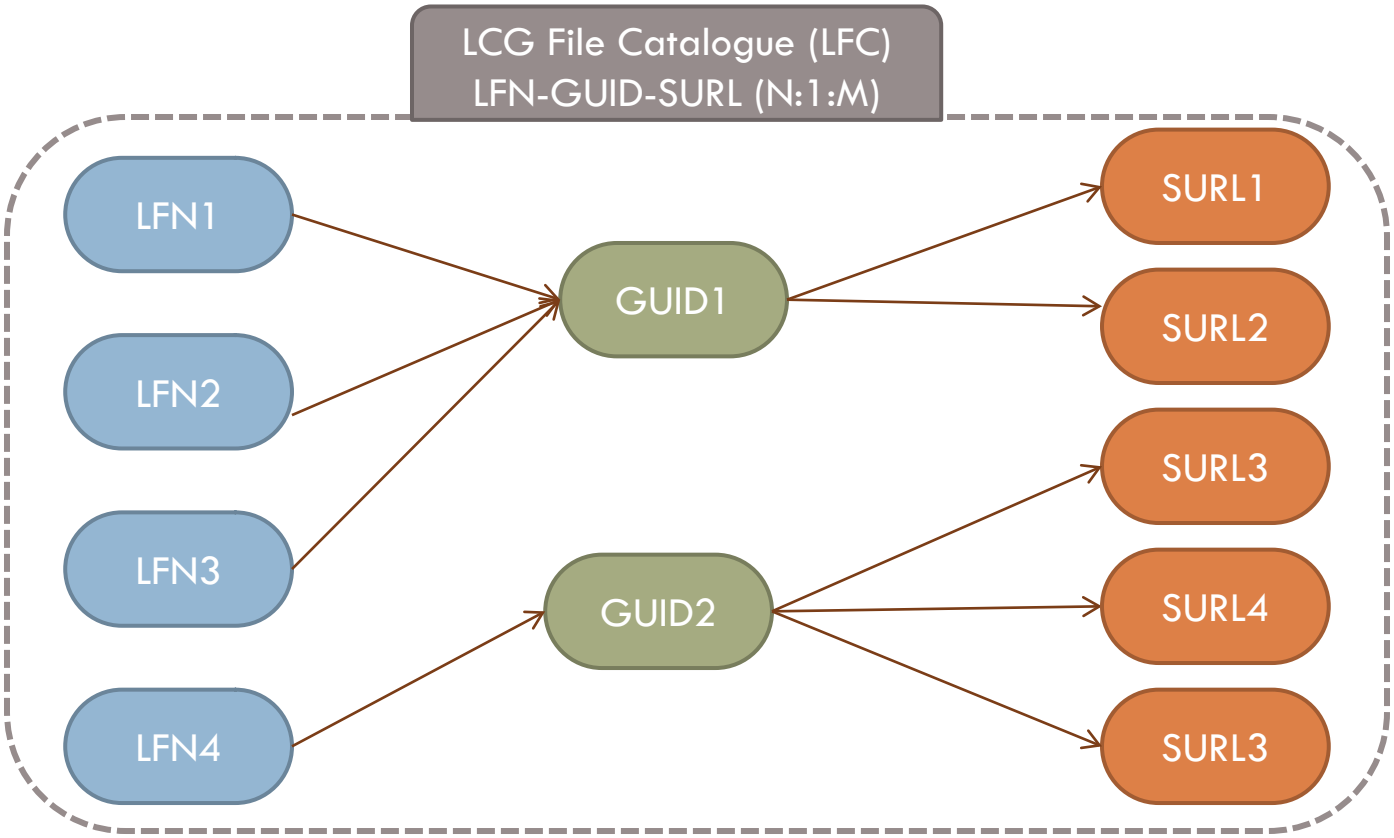
Servicios de gestión de datos

- Transferencia de un fichero (LFNsrc) a un SEdest



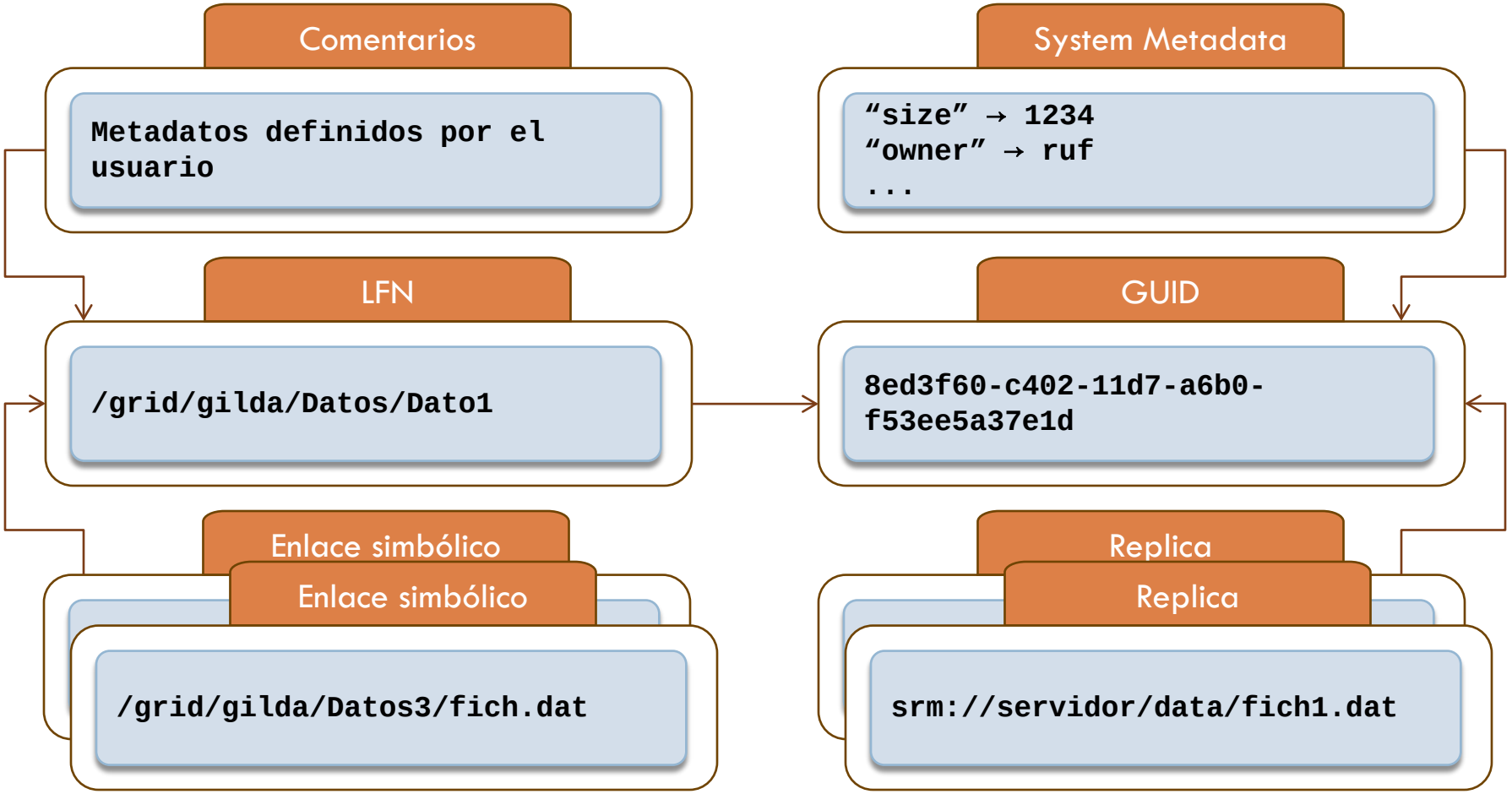
Servicios de gestión de datos

- Relaciones entre nombres de ficheros (LCG-3):



Servicios de gestión de datos

- Arquitectura del LFC (LFN como clave primaria):



Servicios de gestión de datos

- Interfaz del LFC:
  - Comandos lcg-* más APIs lcg_*
    - Proporcionan la funcionalidad necesaria para acceder a la información y manipularla
  - Ejemplos de comandos:
    - Listar el contenido de un directorio

```
$ lfc-ls /grid
```

    - Copiar un fichero a un SE y registrarlo

```
$ lcg-cr -d servidorsrm.atc -l lfn:/grid/fich.dat file:$PWD/file1.txt
```

    - Replicar un fichero en otro SE

```
$ lcg-rep -d servidorsrm2.atc lfn:/grid/gilda/fich.dat
```

gLite

Sistema de gestión de carga de trabajo

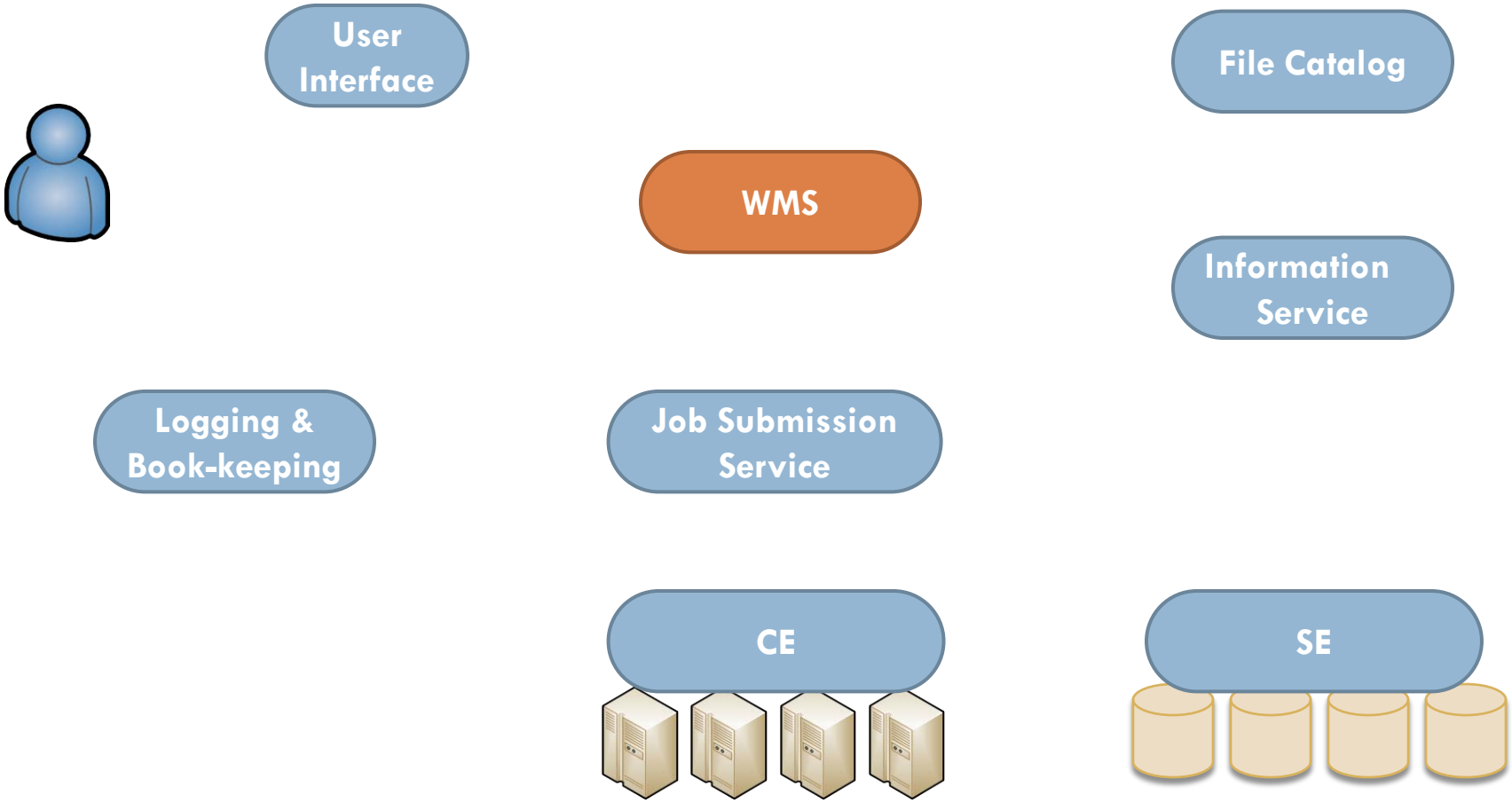
Sistema de gestión de carga de trabajo

- Workload Management System (WMS):
  - Conjunto de componentes responsables de la distribución de los trabajos sobre los recursos computacionales
  - Recibe trabajos de los usuarios y los dirige a los elementos de computación (CE)
  - Se encarga de realizar el matchmaking:
    - Estado de los recursos, requisitos, preferencias, etc.
  - Utiliza WMProxy + Condor-G

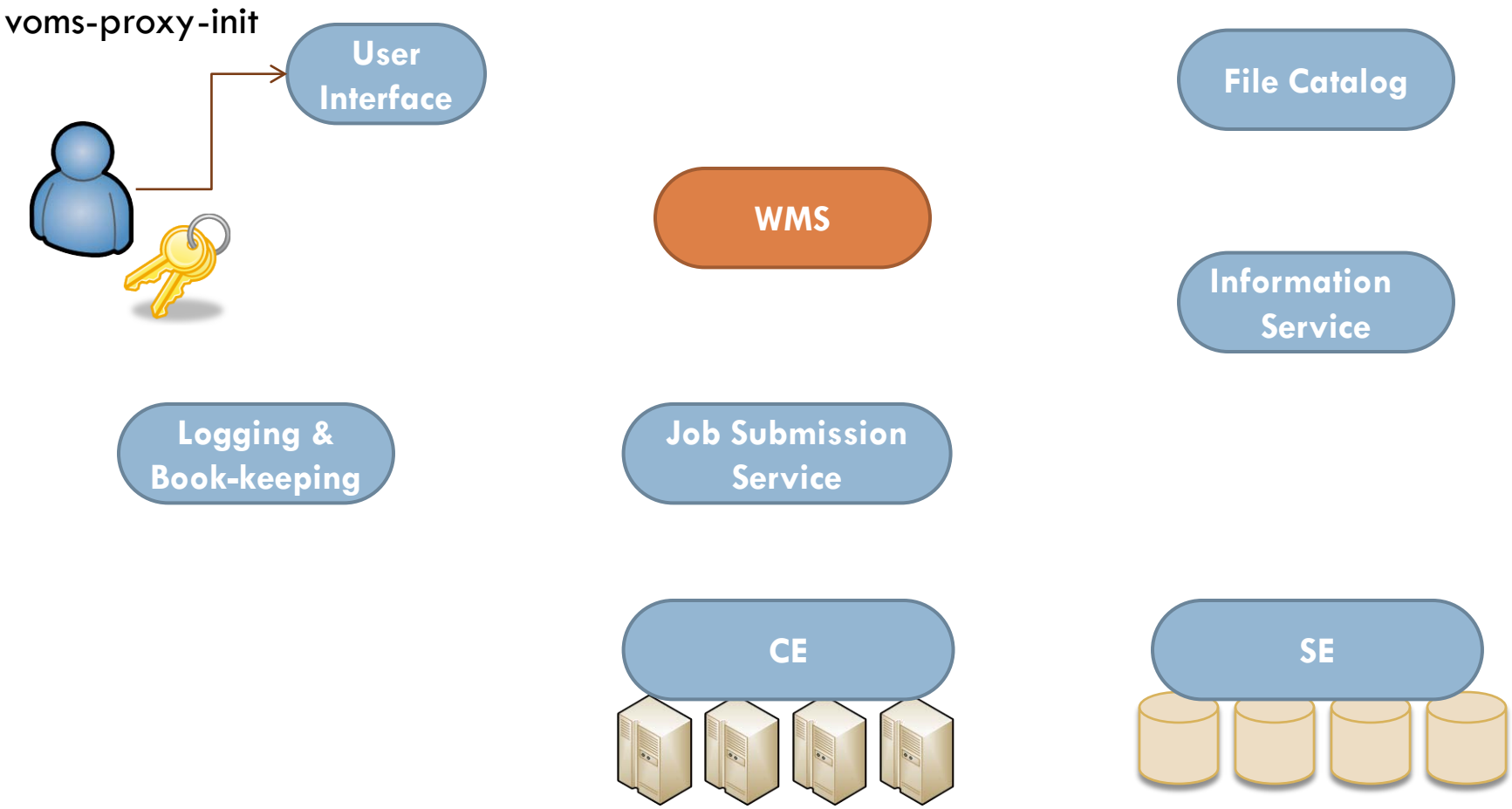
Sistema de gestión de carga de trabajo

- Elemento de computación (CE):
  - Frontend a un cluster
    - El cluster es gestionado por un LRMS: Condor, LSF, PBS, SGE
  - El CE recibe trabajos del WMS y los envía al LRMS
    - El LRMS los envía un nodo de ejecución (WN)
    - Cuando el WN termina la ejecución del trabajo, el CE devuelve los resultados al WMS
  - Versiones de gatekeeper o Grid Gate (GG):
    - LCG-CE (GT2 + GSI-enabled Condor)
    - Glite-CE (GSI-enabled Condor-C)
    - Cream (en desarrollo)

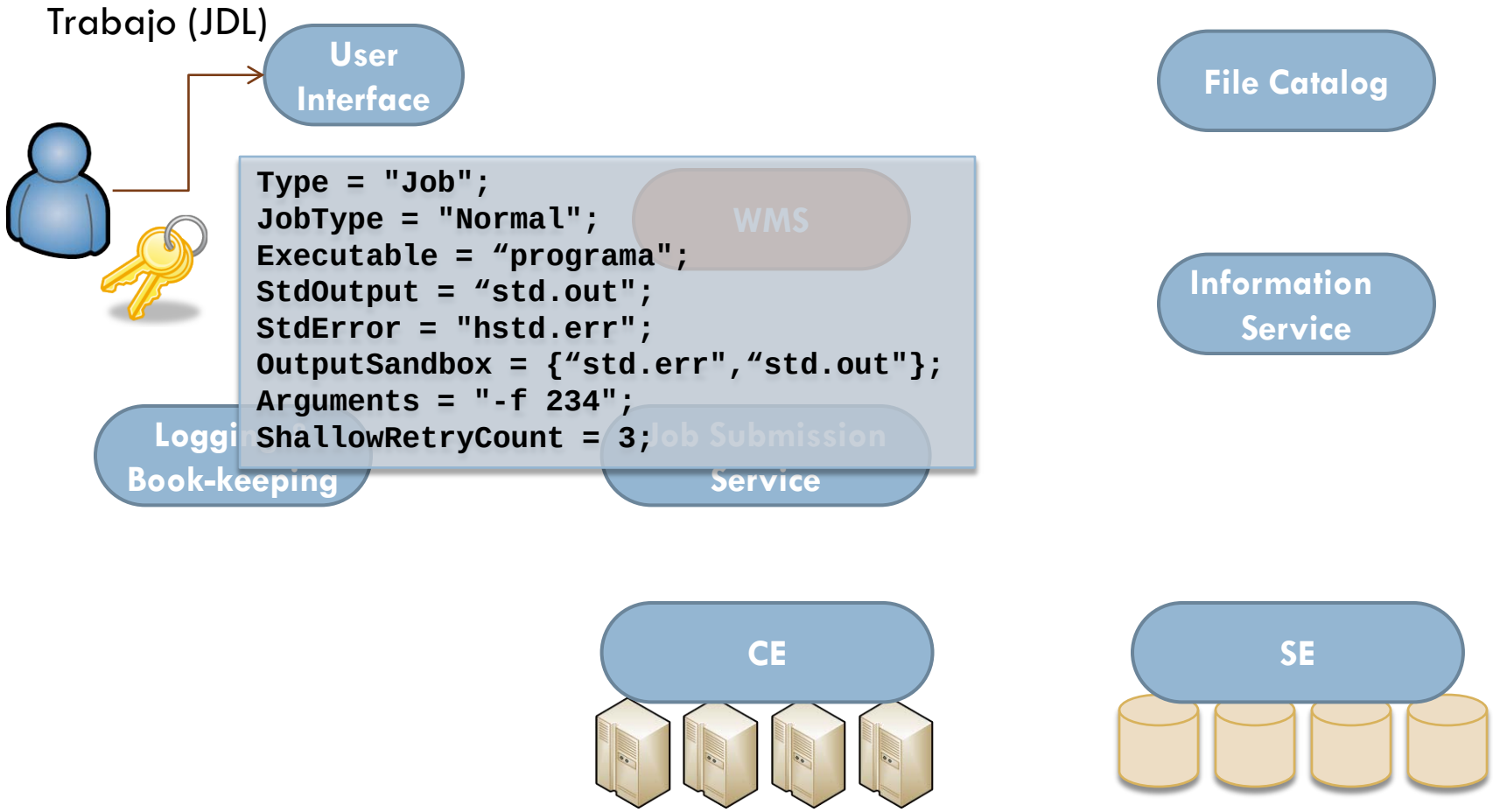
Sistema de gestión de carga de trabajo



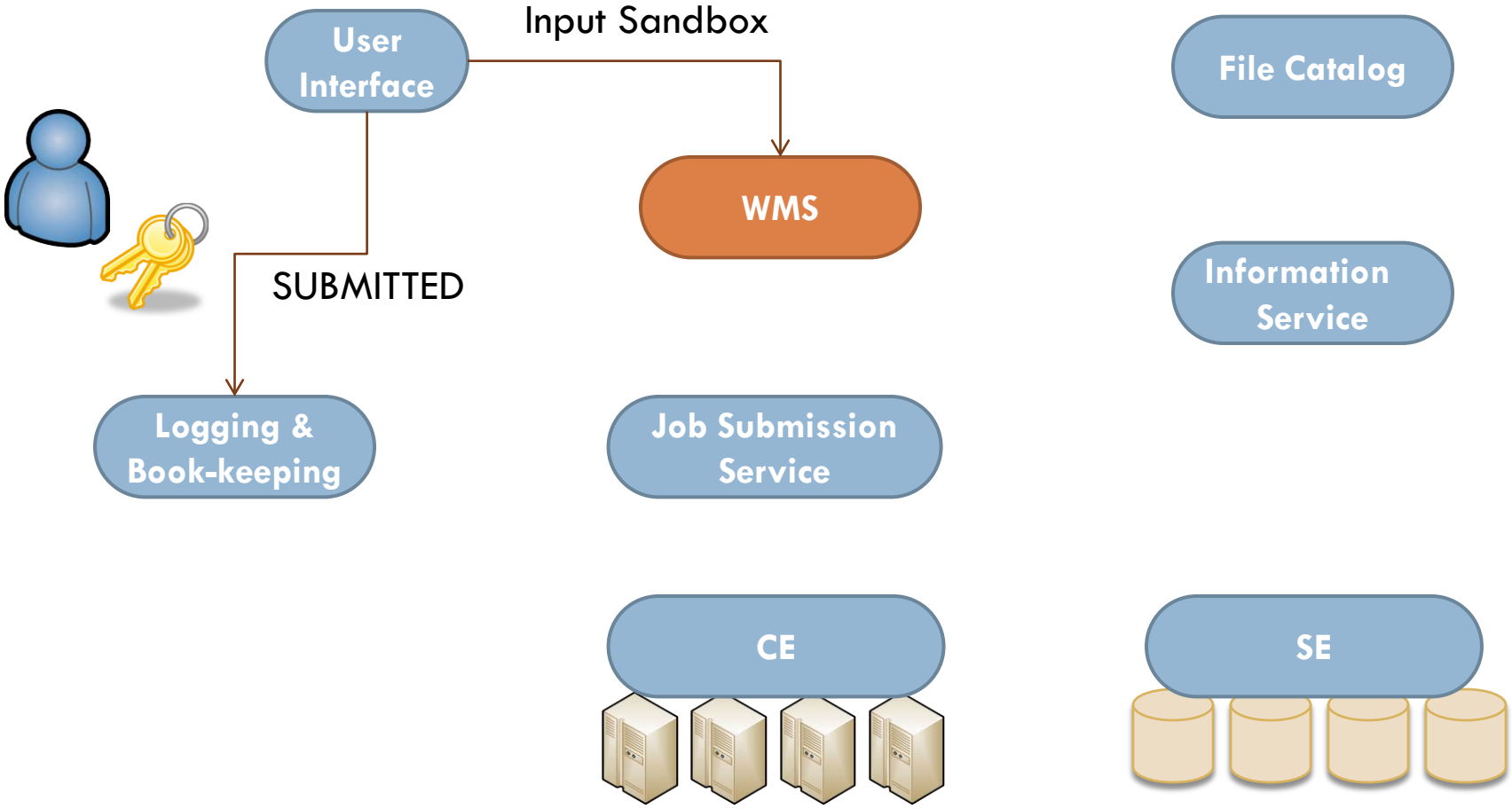
Sistema de gestión de carga de trabajo



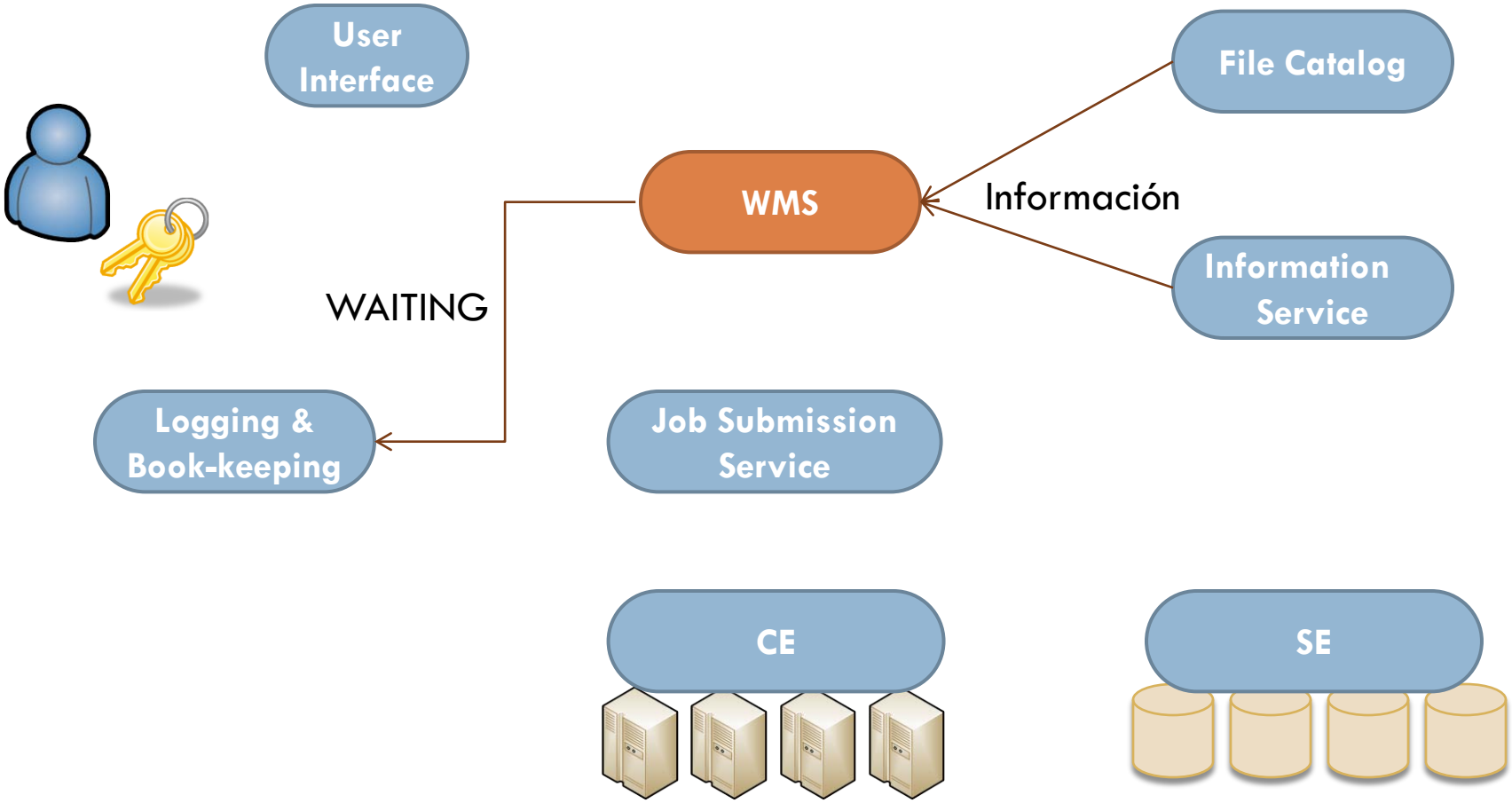
Sistema de gestión de carga de trabajo



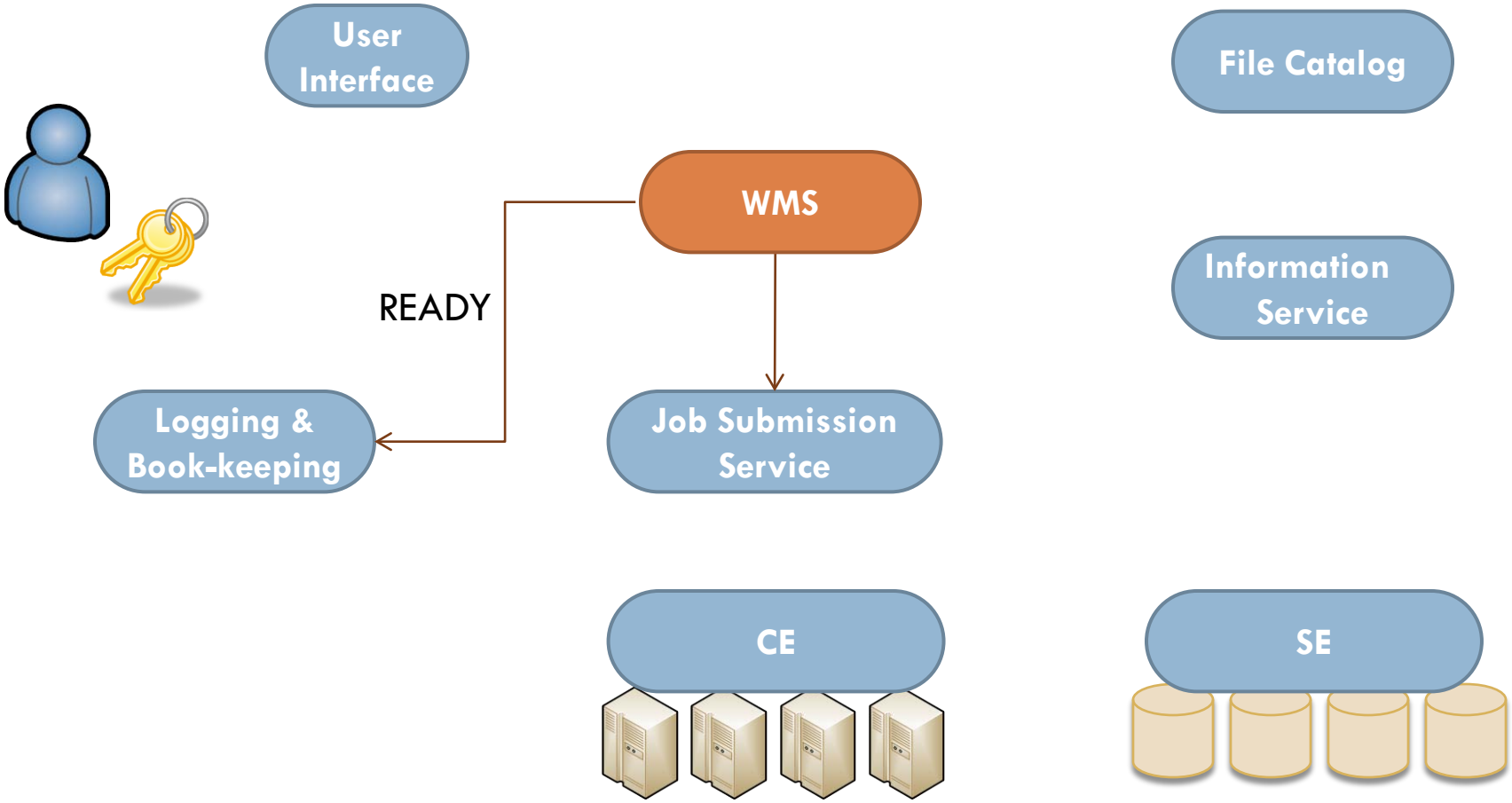
Sistema de gestión de carga de trabajo



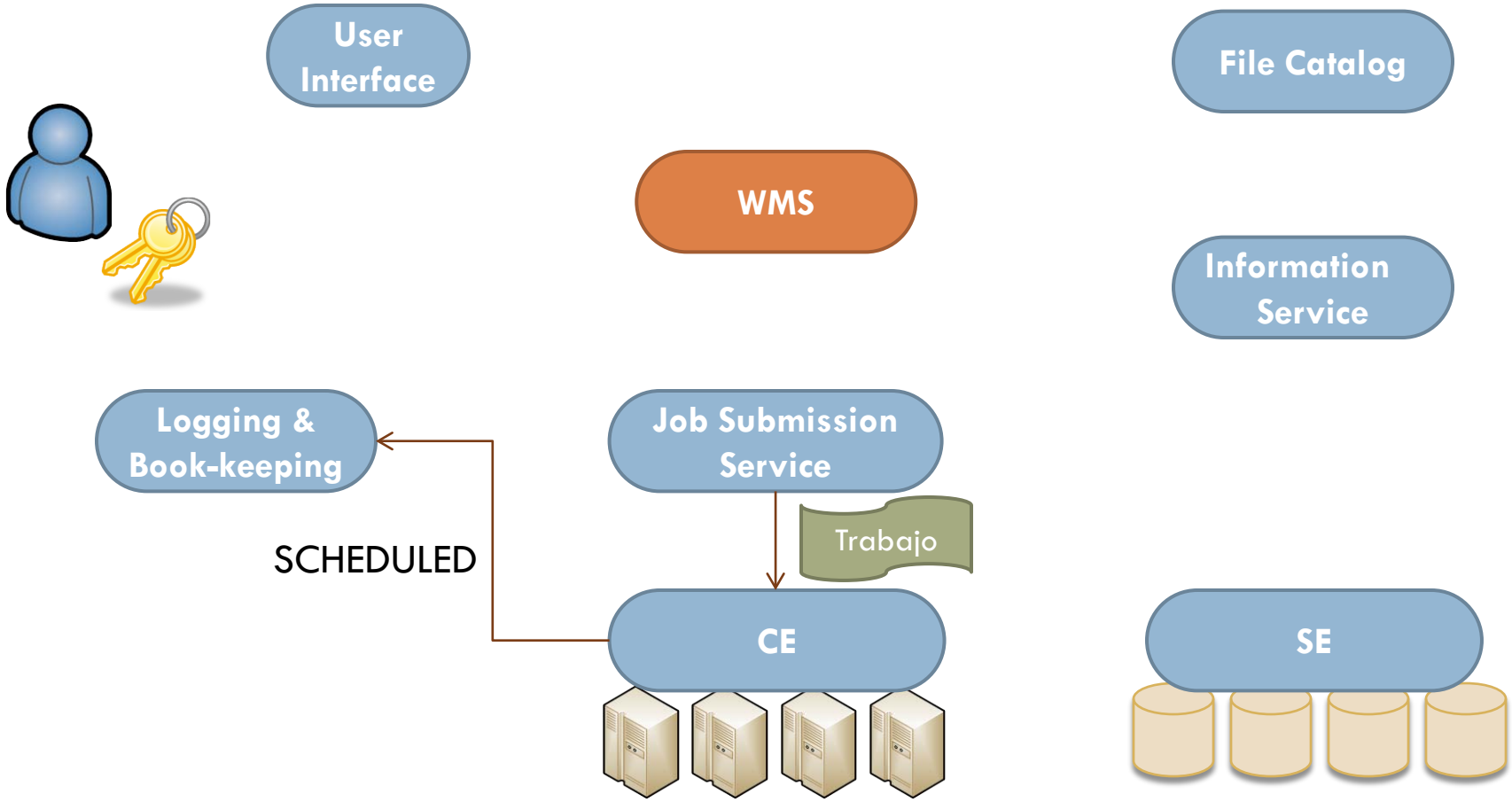
Sistema de gestión de carga de trabajo



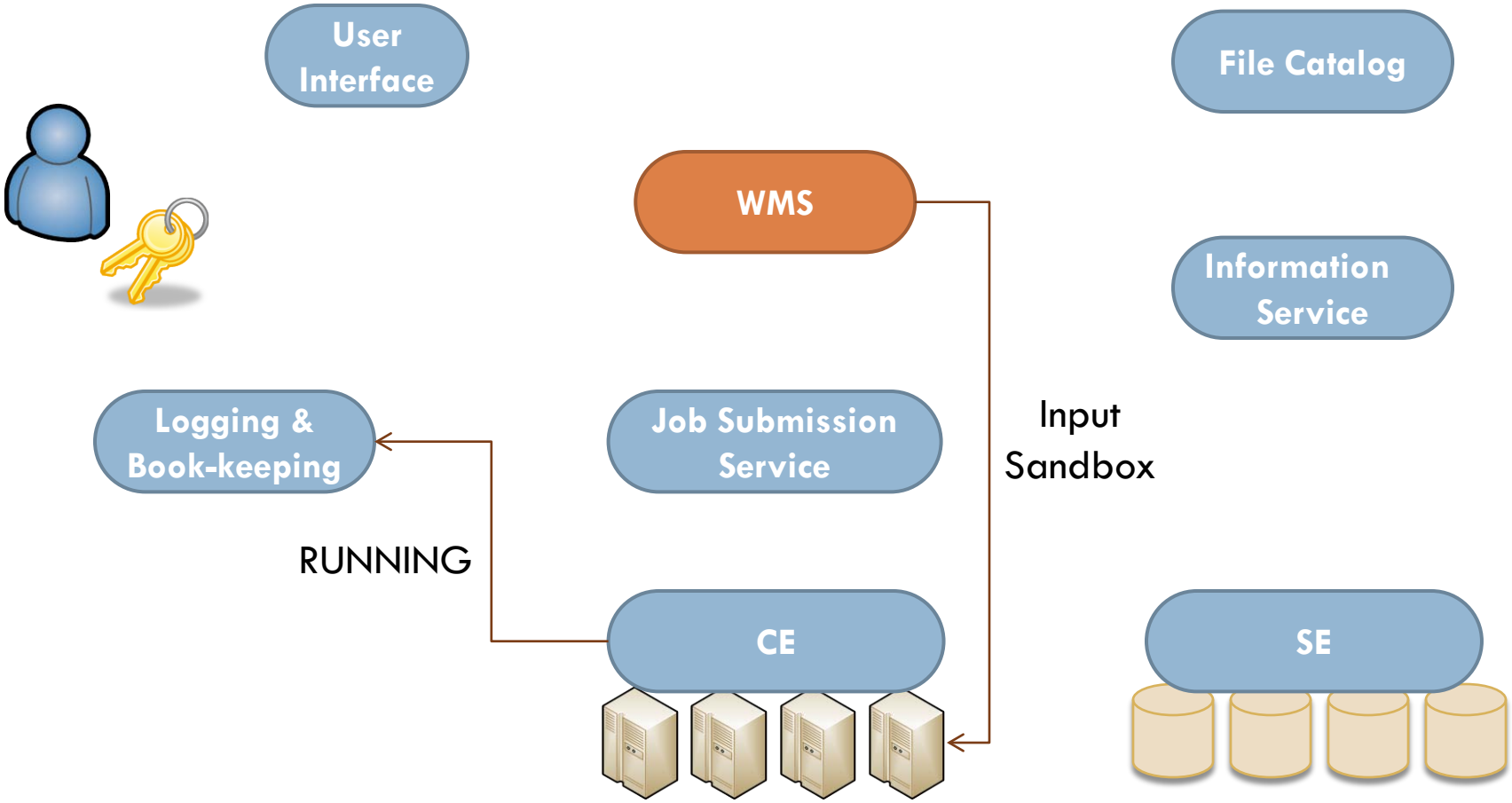
Sistema de gestión de carga de trabajo



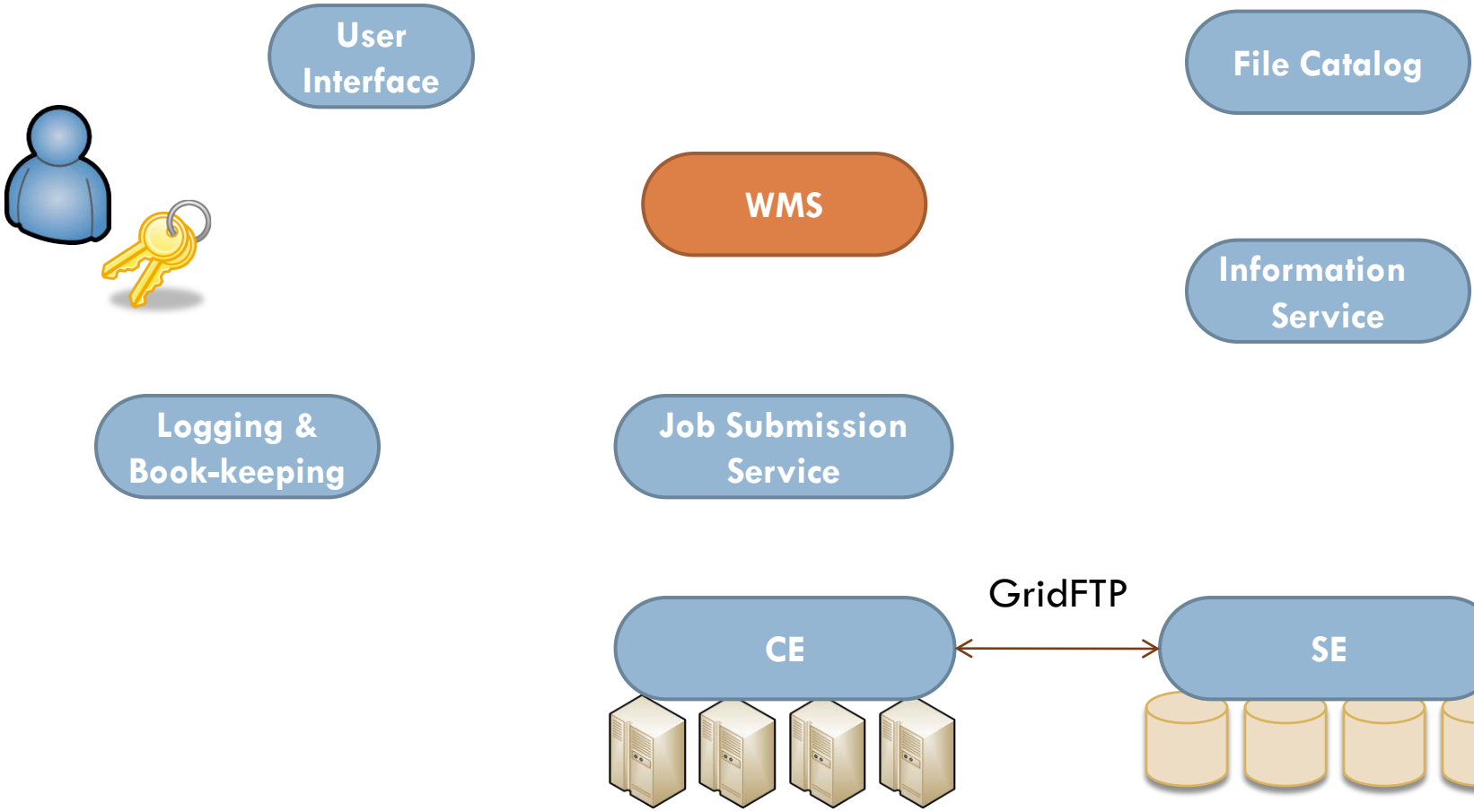
Sistema de gestión de carga de trabajo



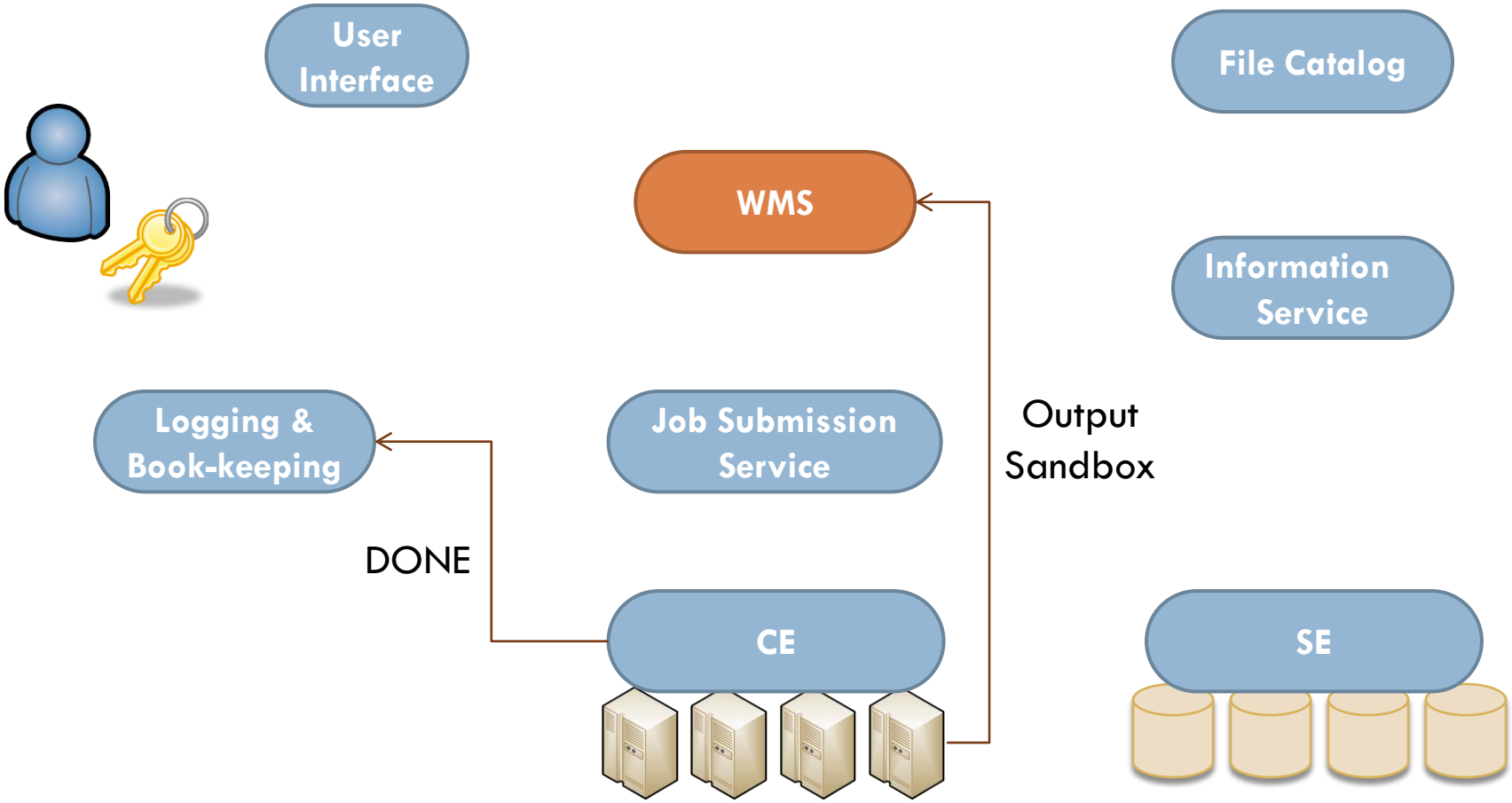
Sistema de gestión de carga de trabajo



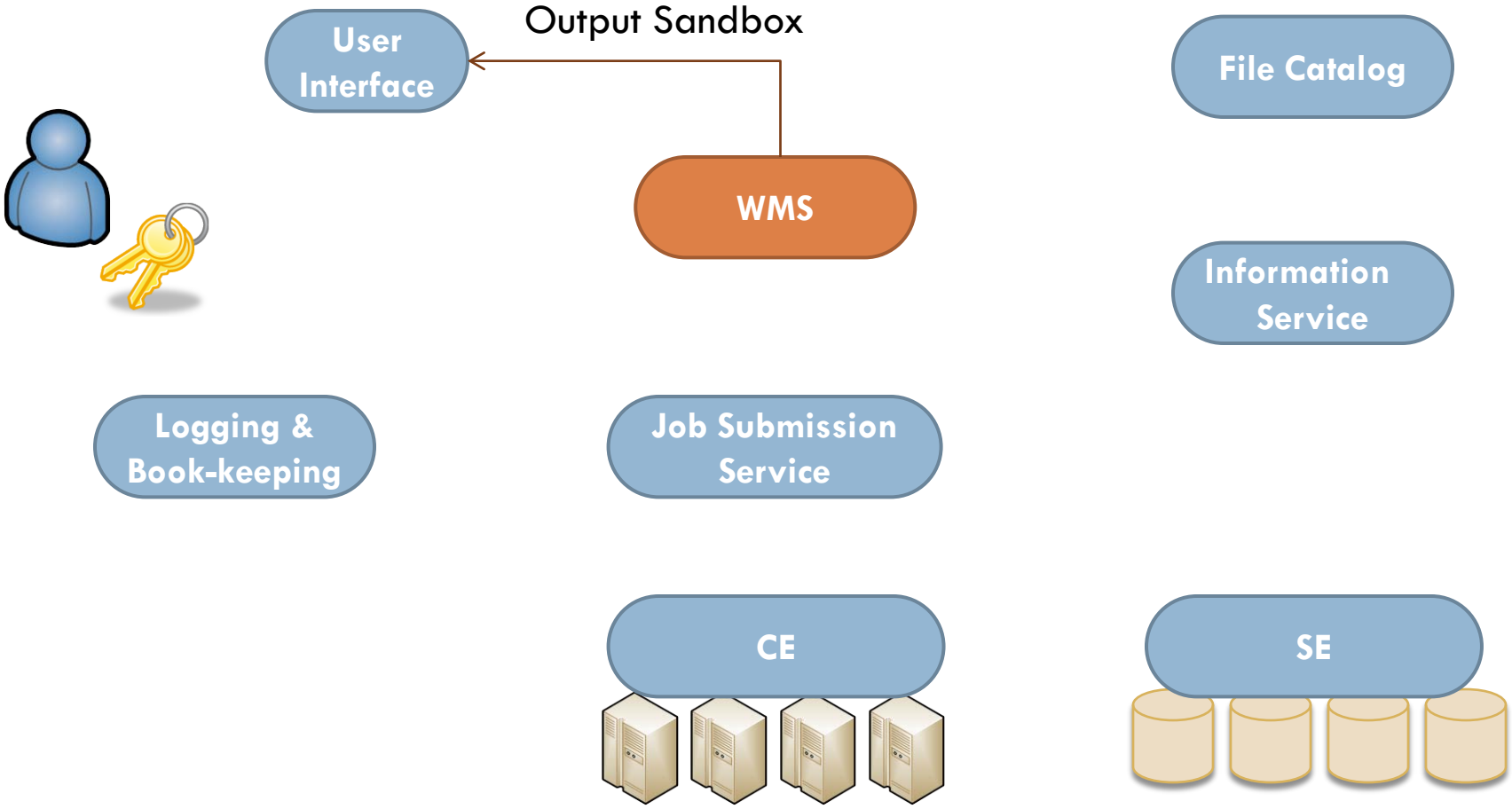
Sistema de gestión de carga de trabajo



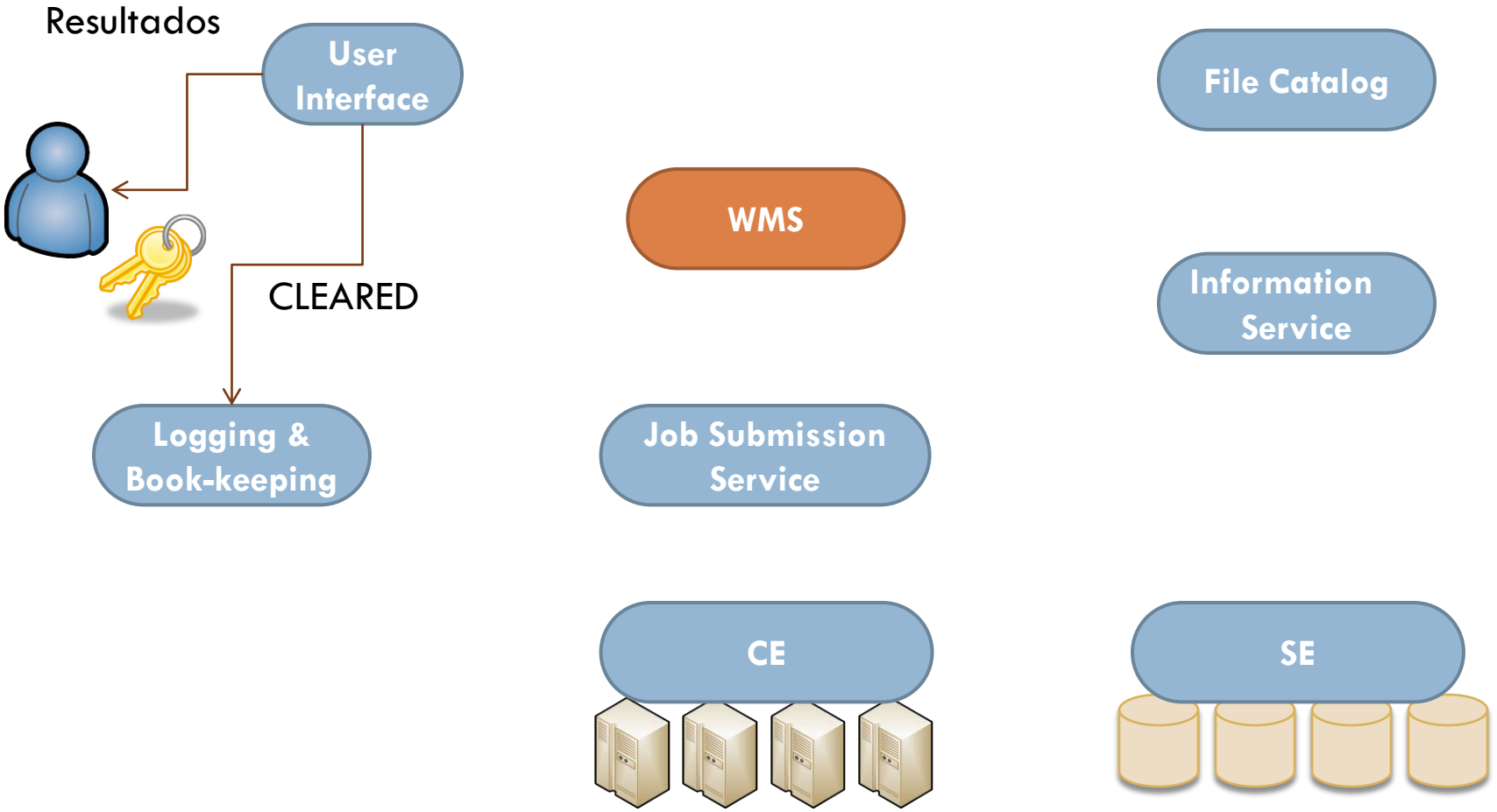
Sistema de gestión de carga de trabajo



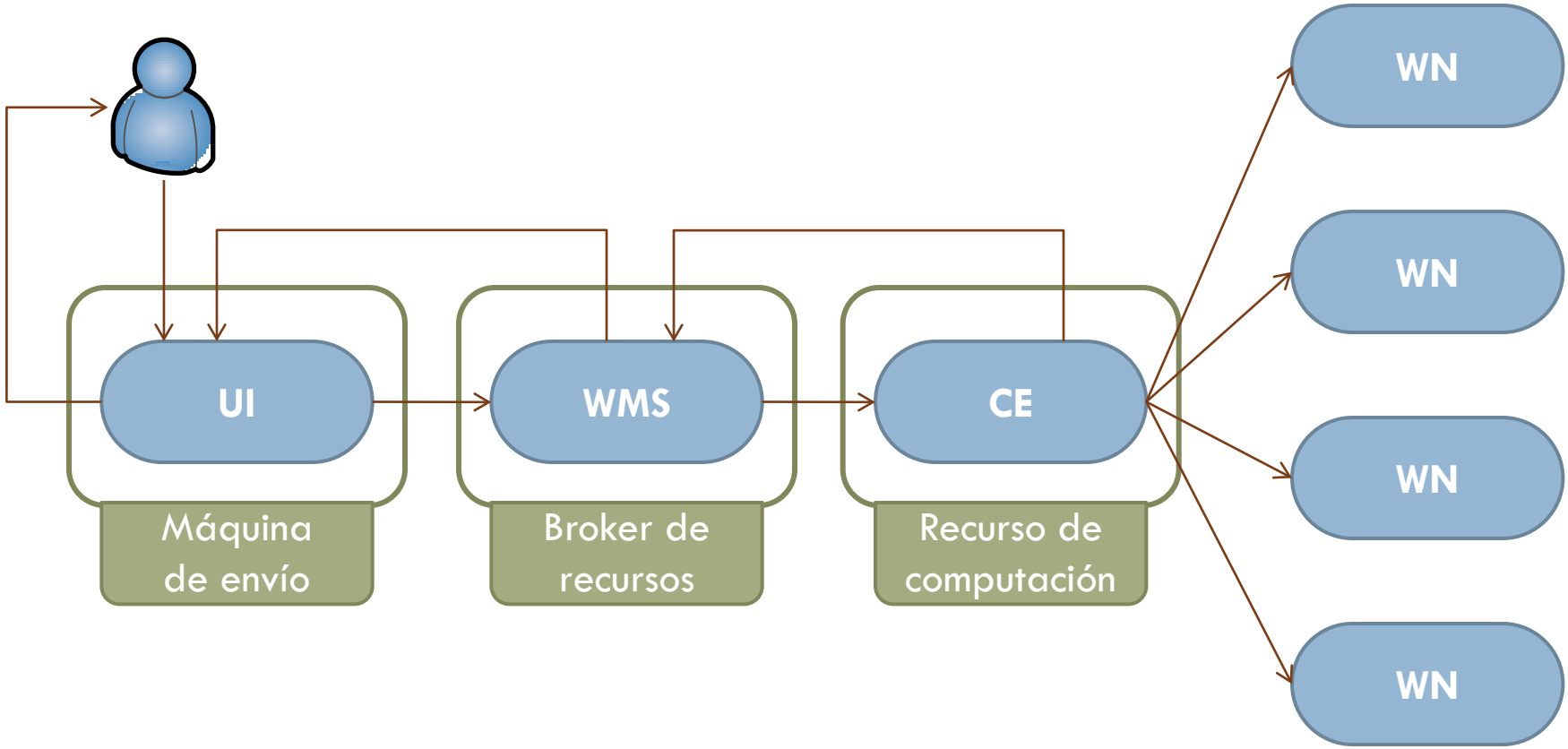
Sistema de gestión de carga de trabajo



Sistema de gestión de carga de trabajo



Sistema de gestión de carga de trabajo

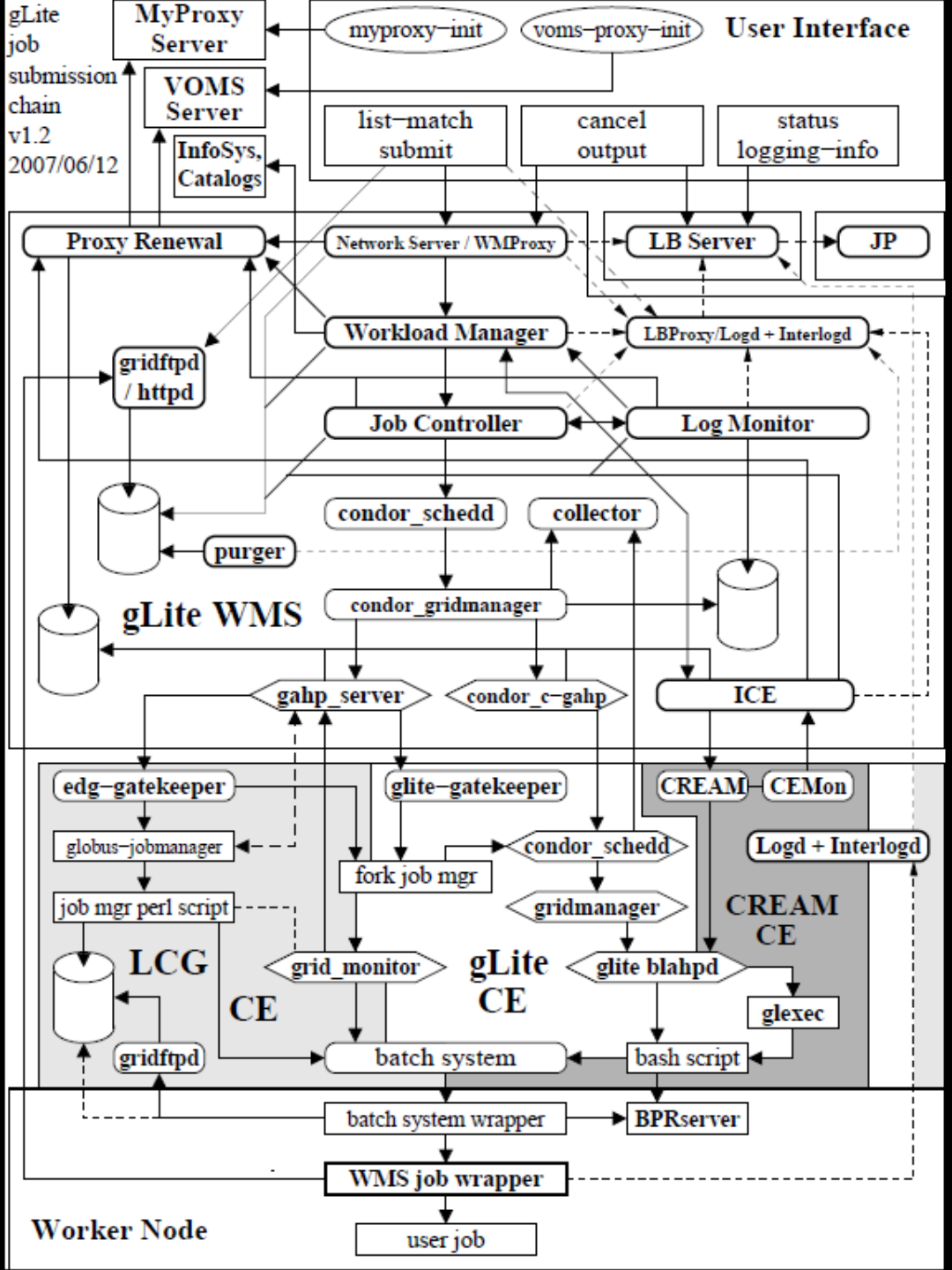


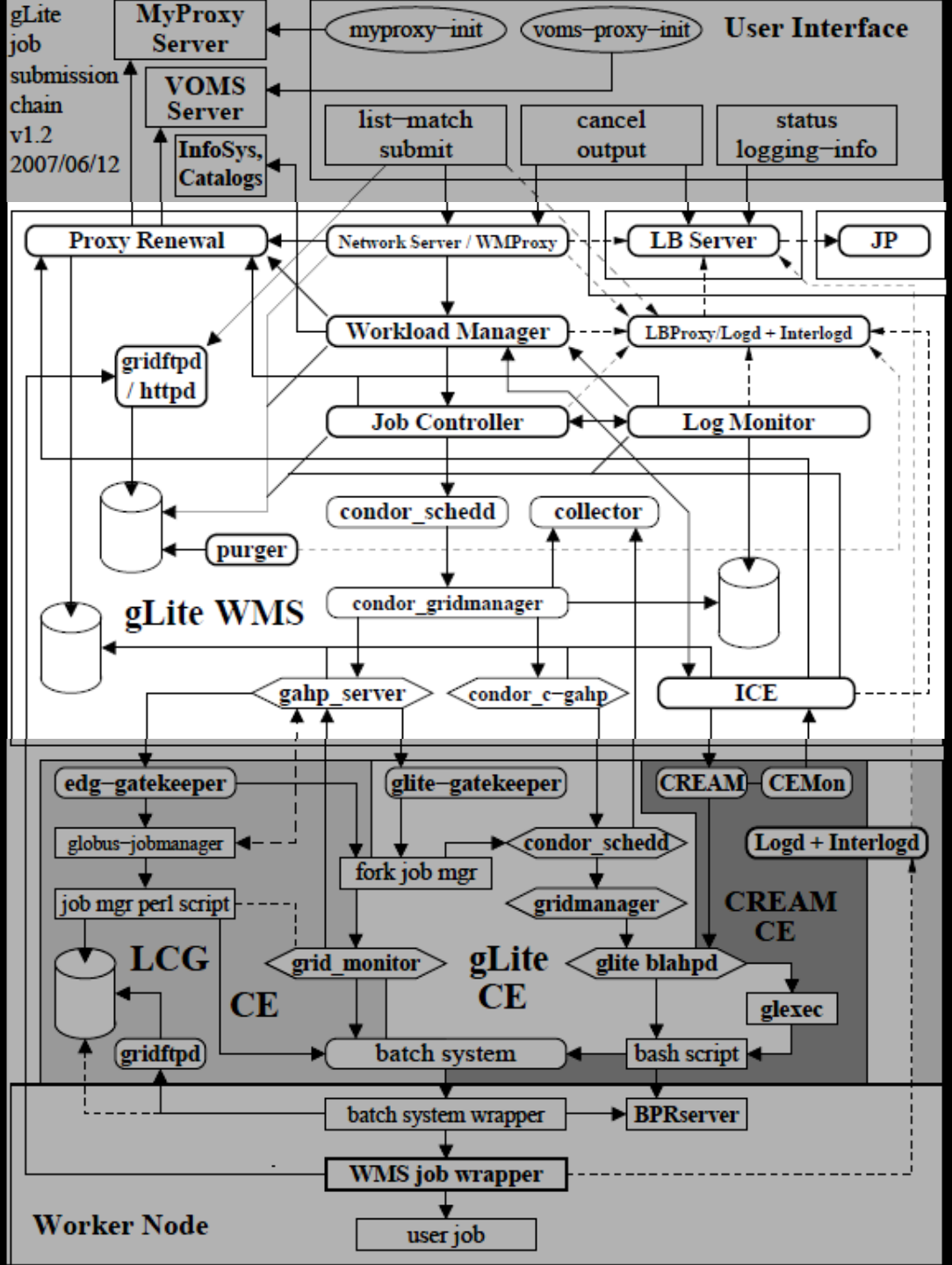
Sistema de gestión de carga de trabajo

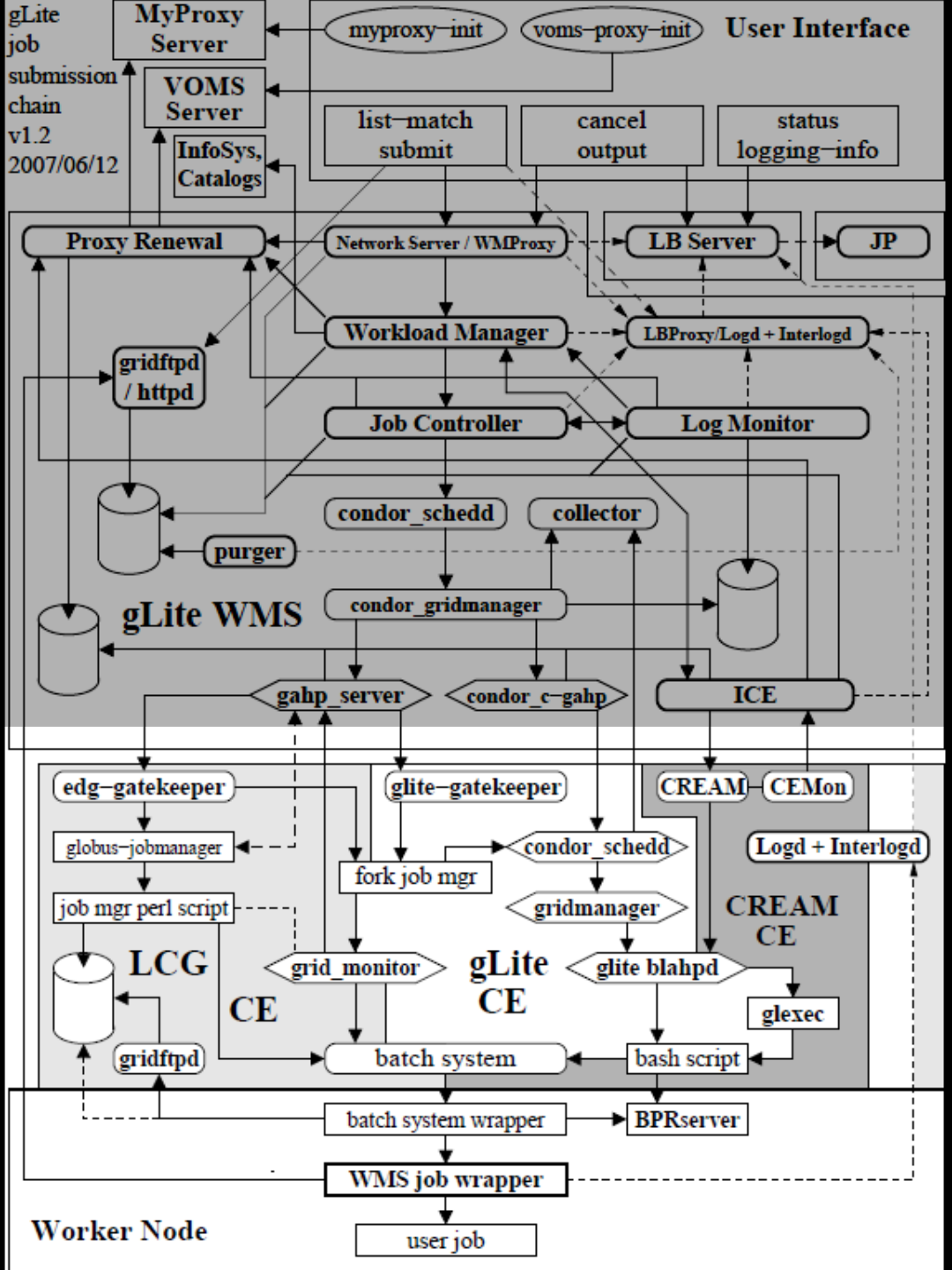
- Algunos componentes del WMS:
  - WMProxy
    - Recibe peticiones de los usuarios a través del UI y las valida
    - Crea el Input Sandbox
  - WM (Workload Manager)
    - Nucleo del WMS
    - Procesa las peticiones de trabajos
    - Realiza el matchmaking
  - JC (Job Controller)
    - Prepara el fichero de envío de Condor
    - Lo envía a Condor

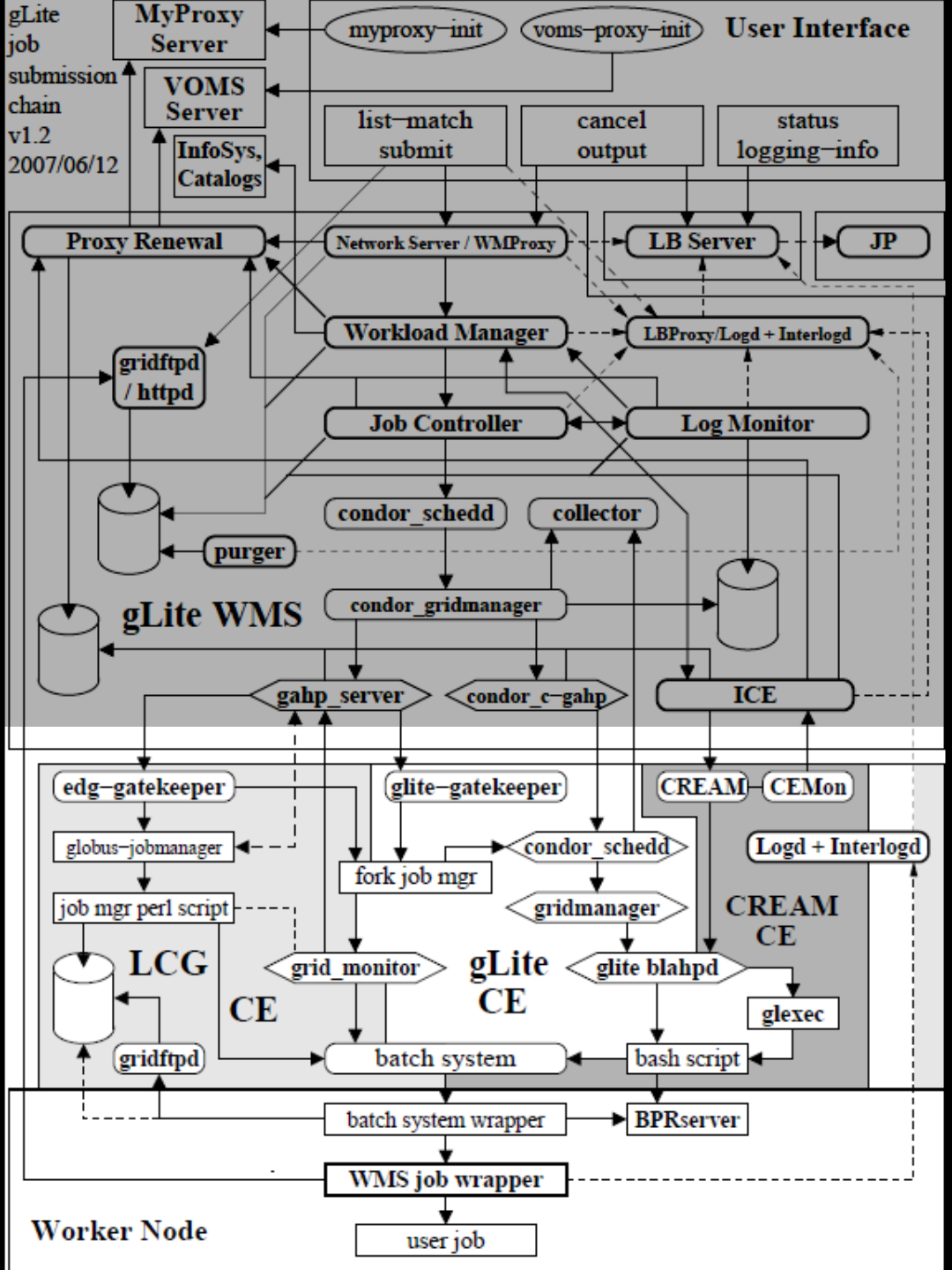
Sistema de gestión de carga de trabajo

- Algunos componentes del WMS:
  - Condor
    - Realiza la gestión del trabajo
  - DAGMan
    - Gestiona los trabajos con dependencias
  - LM (Log Monitor)
    - Monitoriza el log de Condor
    - Intercepta eventos sobre el estado del trabajo


gLite job submission chain v1.2 2007/06/12
MyProxy Server
VOMS Server
InfoSys, Catalogs
myproxy-init
voms-proxy-init
User Interface
list-match submit
cancel output
status logging-info
Proxy Renewal
Network Server / WMProxy
LB Server
JP
Workload Manager
LBProxy/Logd + Interlogd
gridftpd / httpd
Job Controller
Log Monitor
condor_schedd
collector
purger
gLite WMS
condor_gridmanager
gahp_server
condor_c-gahp
ICE
edg-gatekeeper
glite-gatekeeper
CREAM
CEMon
globus-jobmanager
condor_schedd
Logd + Interlogd
fork job mgr
job mgr perl script
gridmanager
CREAM CE
LCG CE
grid_monitor
gLite CE
glite blahpd
glexec
gridftpd
batch system
bash script
batch system wrapper
BPRserver
WMS job wrapper
Worker Node
user job


gLite job submission chain v1.2 2007/06/12
MyProxy Server
VOMS Server
InfoSys, Catalogs
myproxy-init
voms-proxy-init
User Interface
list-match submit
cancel output
status logging-info
Proxy Renewal
Network Server / WMProxy
LB Server
JP
Workload Manager
LBProxy/Logd + Interlogd
gridftpd / httpd
Job Controller
Log Monitor
condor_schedd
collector
purger
gLite WMS
condor_gridmanager
gahp_server
condor_c-gahp
ICE
edg-gatekeeper
glite-gatekeeper
CREAM
CEMon
globus-jobmanager
condor_schedd
Logd + Interlogd
fork job mgr
job mgr perl script
gridmanager
CREAM CE
LCG CE
grid_monitor
gLite CE
glite blahpd
glexec
gridftpd
batch system
bash script
batch system wrapper
BPRserver
WMS job wrapper
Worker Node
user job


gLite job submission chain v1.2 2007/06/12
MyProxy Server
VOMS Server
InfoSys, Catalogs
myproxy–init
voms–proxy–init
User Interface
list–match submit
cancel output
status logging–info
Proxy Renewal
Network Server / WMProxy
LB Server
JP
Workload Manager
LBProxy/Logd + Interlogd
gridftpd / httpd
Job Controller
Log Monitor
condor_schedd
collector
purger
gLite WMS
condor_gridmanager
gahp_server
condor_c–gahp
ICE
edg–gatekeeper
glite–gatekeeper
CREAM
CEMon
globus–jobmanager
condor_schedd
Logd + Interlogd
fork job mgr
job mgr perl script
gridmanager
CREAM CE
LCG CE
grid_monitor
gLite CE
glite blahpd
glexec
gridftpd
batch system
bash script
batch system wrapper
BPRserver
WMS job wrapper
Worker Node
user job


gLite job submission chain v1.2 2007/06/12
MyProxy Server
VOMS Server
InfoSys, Catalogs
myproxy–init
voms–proxy–init
User Interface
list–match submit
cancel output
status logging–info
Proxy Renewal
Network Server / WMProxy
LB Server
JP
Workload Manager
LBProxy/Logd + Interlogd
gridftpd / httpd
Job Controller
Log Monitor
condor_schedd
collector
purger
gLite WMS
condor_gridmanager
gahp_server
condor_c–gahp
ICE
edg–gatekeeper
glite–gatekeeper
CREAM
CEMon
globus–jobmanager
condor_schedd
Logd + Interlogd
fork job mgr
job mgr perl script
gridmanager
CREAM CE
LCG CE
grid_monitor
gLite CE
glite blahpd
glexec
gridftpd
batch system
bash script
batch system wrapper
BPRserver
WMS job wrapper
Worker Node
user job

Sistema de gestión de carga de trabajo

- Algunas siglas:

EGEE: Enabling Grids for EScienceE
VO: Virtual Organisation
JDL: Job definition Language
BDII: Berkeley Database Information Index
GRAM: Globus Resource Allocation Manager
MDS: Metadata Directory Service
GRIS: Grid Resource Information Service
GSI: Grid Security Infrastructure
GUID: Globally (Grid) Unique Identifier
IS: Information System
GAHP: Grid ASCII Helper Protocol
RGMA: Relational Grid Monitoring Architecture
GLUE: Grid Laboratory Uniform Environment
CG: Grid Gate
LRMS: Local Resource Management System

LM: Log Monitor
JC: Job Controller
LB: Logging and Bookkeeping
LB: Logging and Bookkeeping
PRS: Proxy Renewal Service
PS: Proxy Server
VOMS: Virtual Organisation Membership
RB: Resource Broker
UI: User Interface
WM: Workload Manager
WMS: Workload Management System
WN: Worker Node
CE: Computing Element
SE: Storage Element
ICE: InterfacetoCrEam