Instalación de CE y WN LCG

Jesús De Oliveira

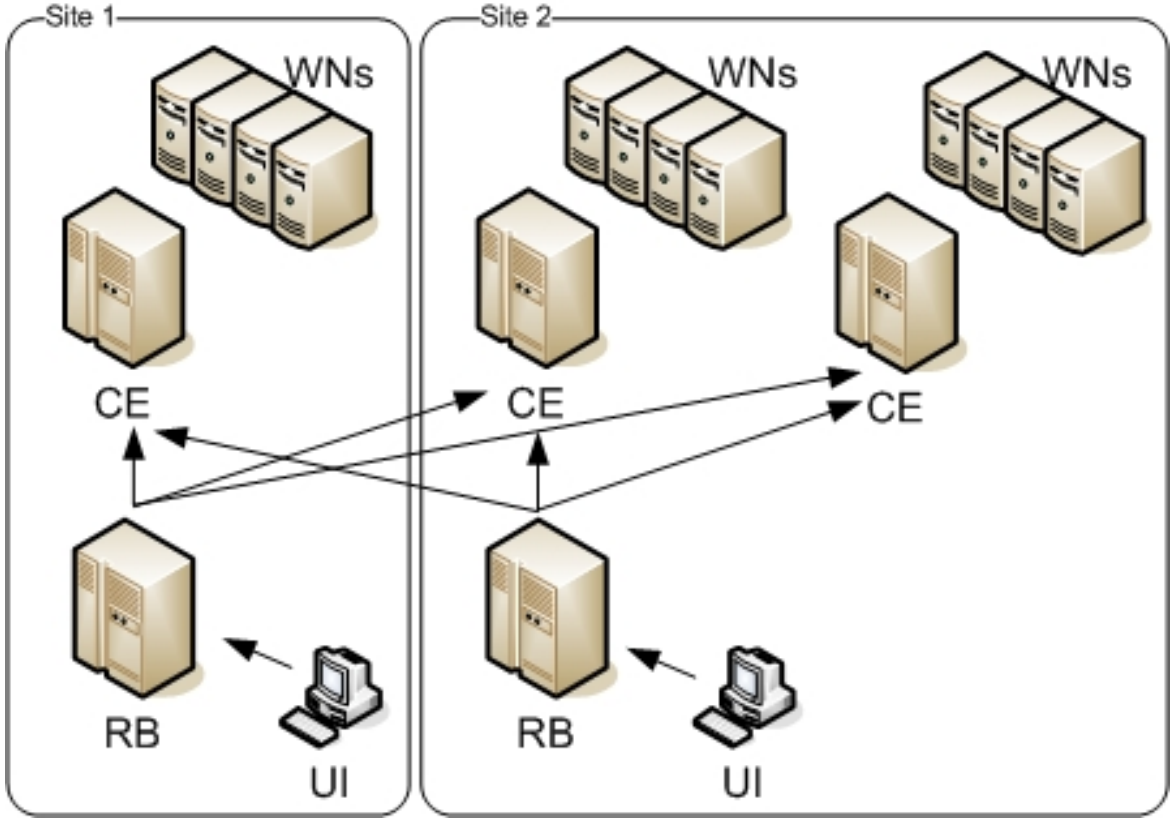
Universidad Simón Bolívar

Temario

- **Introducción**
  - Componentes del CE y el WN
  - Comunicación entre componentes
    - Envío de trabajos
    - Monitoreo y descubrimiento de recursos
- **Instalación con YAIM**
  - Pasos preliminares
  - Archivos de configuración
  - Script de instalación
  - Script de configuración
- **Diagnostico de la instalación**
  - Metodología del diagnóstico
  - Problemas frecuentes

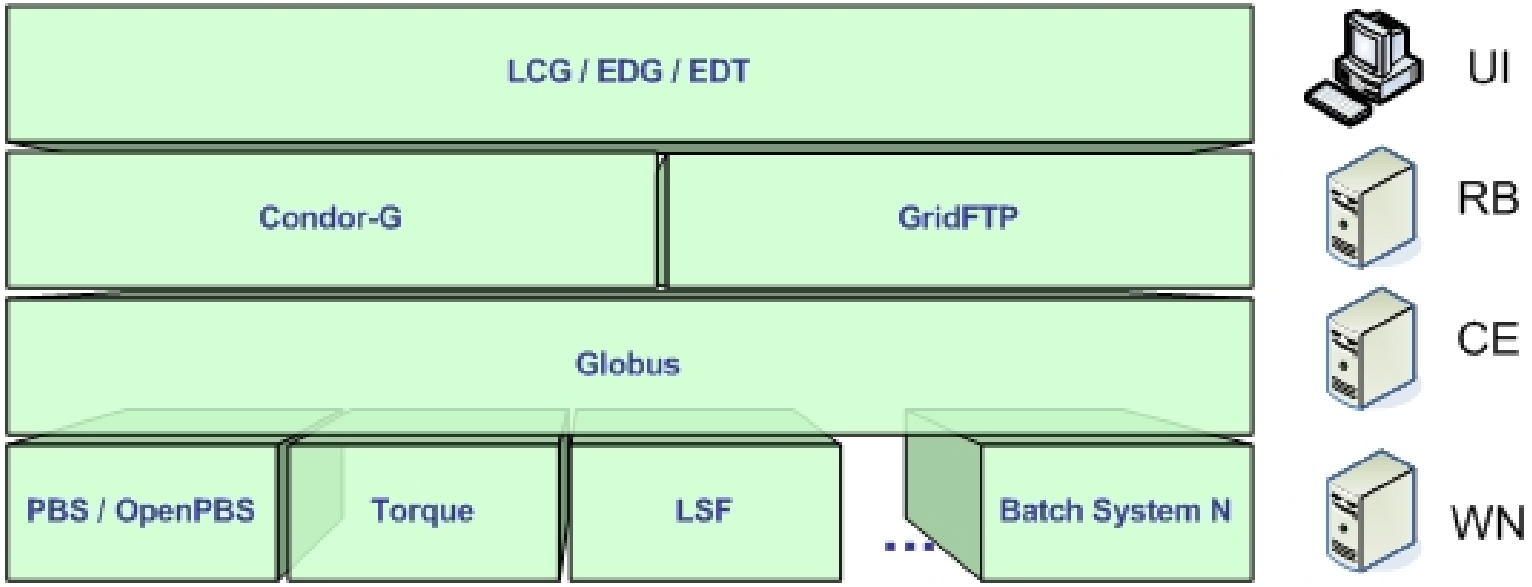
Introducción

- **CE: Computing Element**
  - Interfaz entre el GRID y el “recurso de cómputo” (cluster)
  - Recibe los trabajos desde el Resource Broker (RB) y los envía al manejador de colas local (*Batch System*)
- **WN: Worker Node**
  - Nodo de cómputo donde finalmente son ejecutados los trabajos



Introducción

- **Componentes mas importantes del CE:**
  - Globus Resource Allocation Manager (GRAM)
  - Cliente del manejador de colas
  - Master del manejador de colas (por lo general)
- **Componentes mas importantes del WN:**
  - Slave del manejador de colas
  - Software *globus* para transferencia de archivos (grid-ftp)



Introducción

- **Comunicación: Envío de trabajos**
  - El RB envía el trabajo a través de Condor-G al GRAM Gatekeeper en el CE
    - Job Wrapper
    - GRAM Sandbox
  - El GRAM Gatekeeper encola el trabajo en el Batch System
    - *globus-job-manager*: Interfaz entre globus y el manejador de colas
    - *grid-monitor*: Determina el estado del trabajo en el manejador de colas
  - El manejador de colas asigna el trabajo a uno o varios nodos (WN's)
  - El Job Wrapper comienza a ejecutarse en el WN
    - Transfiere el GRAM Sandbox desde el CE a través de scp (depende del manejador de colas)
    - Descarga desde el RB el InputSandbox del trabajo, a través de grid-ftp
    - Redirecciona la entrada/salida estándar y de error y ejecuta el programa del usuario

Introducción

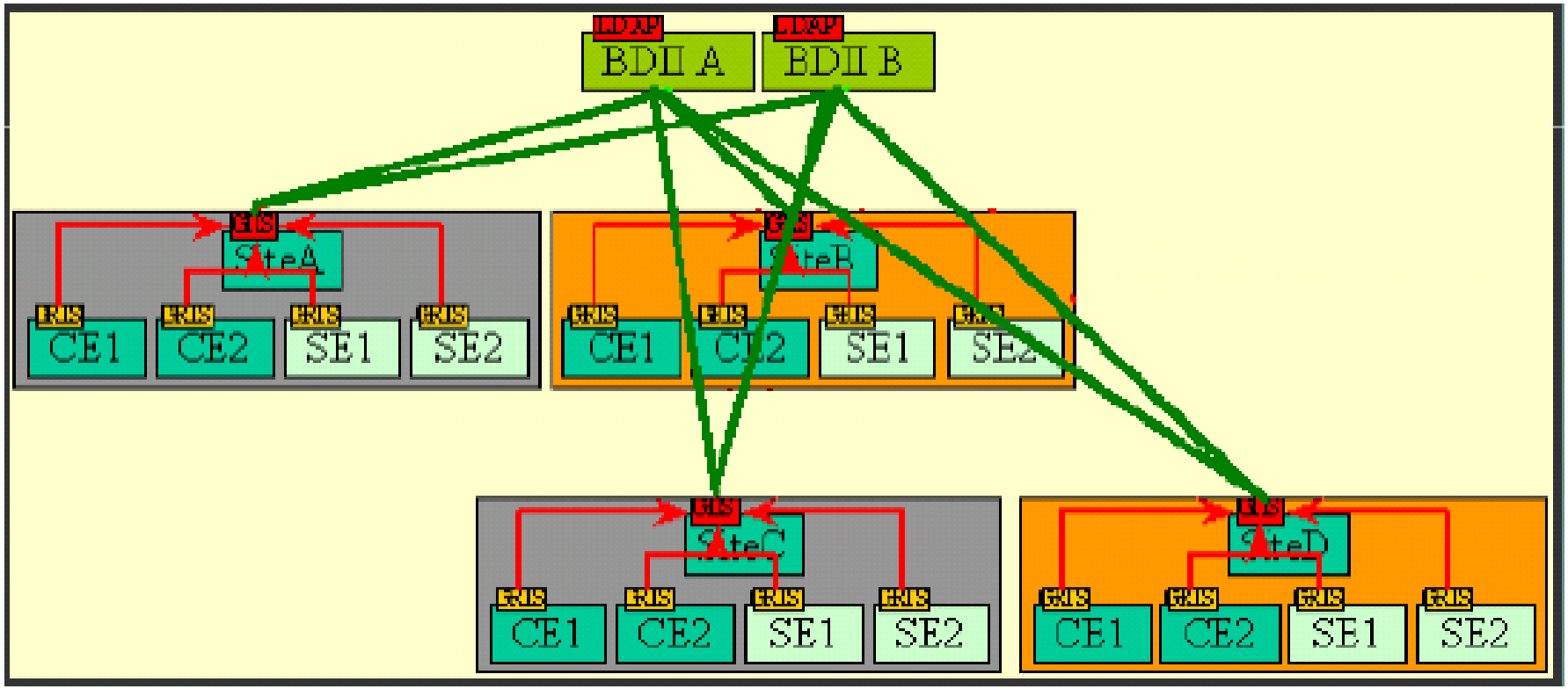
- **Comunicación: Envío de trabajos (cont.)**
  - Durante la ejecución del trabajo en el WN
    - El grid-monitor consulta el estado del trabajo en el Batch System y lo comunica al RB a través de callbacks
  - Al finalizar el trabajo (termina el ejecutable del usuario)
    - El Job Wrapper sube el Output Sandbox al RB vía grid-ftp
    - Termina el trabajo en el Batch System
    - El grid-monitor detecta y notifica la finalización del trabajo al RB
    - El RB reinicia en el CE el globus-job-manager para limpieza y para transferir el Output Sandbox al RB, también vía grid-ftp
- **En toda comunicación entre nodos se realiza la autenticación GSI**
  - Certificado Proxy del usuario
  - Certificado de host

Introducción

- **Comunicación: Monitoreo y descubrimiento**
  - Sistema de información basado en Globus MDS
  - Se usa el protocolo LDAP para el intercambio de información, de acuerdo al GLUE Schema
  - En el CE corre un GRIS (Grid Resource Information System) que reporta el estado del recurso de cómputo
    - Número de CPUs
    - Trabajos encolados y en ejecución
    - Software instalado en el recurso
    - SEs cercanos al recurso
    - etc...
  - Utiliza comandos específicos del manejador de colas local para obtener información de su estado (p.e. *pbsnodes*)

Introducción

- **Comunicación: Monitoreo y descubrimiento (cont.)**



Instalación con YAIM

- **Pasos preliminares**
  - Establecer el nombre *completamente calificado* del host en el archivo `/etc/hosts`
  - Los nodos deben estar sincronizados con NTP
    - `ntpq -p`
    - `ntpstat`
  - Firewall:
    - Es recomendable desactivar el servicio *iptables* en los nodos, y correr un firewall que proteja al site completo, en un nodo dedicado
      - `chkconfig iptables off`
      - `/etc/init.d/iptables stop`
  - Certificados X.509 de hosts
    - La clave privada debe estar desencriptada
      - `openssl rsa < criptedhostkey.pem > uncriptedhostkey.pem`
    - Asegurarse de poseer los certificados de las autoridades certificadoras en `/etc/grid-security/certificates`

Instalación con YAIM

- **Descripción del proceso de instalación**
  - YAIM: Conjunto de scripts para facilitar la instalación y configuración de los paquetes LCG
  - Descarga e instala los RPMs correspondientes a cada “meta-paquete” utilizando *apt*
  - Configura cada componente a partir de archivos de configuración general del site: *site-info.def, wn-list.conf y users.conf*
  - Para descargar e instalar YAIM:
    - `wget http://grid-deployment.web.cern.ch/grid-deployment/gis/yaim/lcg-yaim-2.4.0-4.noarch.rpm`
    - `rpm -ivh lcg-yaim-2.4.0-4.noarch.rpm`

Instalación con YAIM

- **Archivos de configuración: site-info.def**
  - Este archivo es “sourced” por los scripts de instalación y configuración
  - Variables requeridas para configurar el CE y WNs con el manejador de colas *torque*
    - MY_DOMAIN=<dominio de los nodos>
    - CE_HOST=<hostname del CE>
    - SE_HOST=<hostname del SE>
    - RB_HOST=<hostname del RB>
    - PX_HOST=<hostname del PX>
    - BDII_HOST=<hostname del BDII>
    - MON_HOST=<hostname del MON>
    - REG_HOST=<hostname del REG>
    - LFC_HOST=<hostname del LFC>
    - WN_LIST=<ubicación del archivo wn-list.conf>
    - USERS_CONF=<ubicación del archivo users.conf>
    - LCG_REPOSITORY=<ubicación del repositorio de paquetes>
    - INSTALL_ROOT=<directorio de instalación de paquetes>

Instalación con YAIM

- **Archivos de configuración: site-info.def (cont.)**
  - JAVA_LOCATION=<localización de java>
  - CRON_DIR=<localización de cron jobs>
  - GLOBUS_TCP_PORT_RANGE="<rango de puertos globus (callback)>"
  - GRID_TRUSTED_BROKERS="<DN del RB>"
  - GRIDMAP_AUTH="<url servidor ldap para autenticación> ..."
  - SITE_EMAIL=<email del admin>
  - SITE_NAME="<nombre del site>"
  - SE_TYPE=<tipo del SE>
  - JOB_MANAGER=<tipo de job-manager que usa GRAM>
  - CE_BATCH_SYS=<tipo del manejador de colas> (torque)
  - CE_CPU_MODEL=<modelo del CPU> (PIII: pentium III)
  - CE_CPU_VENDOR=<marca del CPU> (intel)
  - CE_CPU_SPEED=<velocidad del CPU>
  - CE_OS=<sistema operativo>
  - CE_OS_RELEASE=<versión del sistema operativo>
  - CE_MINPHYSMEM=<memoria física (RAM)>

Instalación con YAIM

- **Archivos de configuración: site-info.def (cont.)**
  - CE_MINVIRTMEM=<memoria virtual en Kb>
  - CE_SMPSIZE=<número de CPUs>
  - CE_SI00=<medida de performance en SpecInt 2000>
  - CE_SF00=<medida de performance en SpecFloat 2000>
  - CE_OUTBOUNDIP=<WNs con conectividad hacia “afuera”?>
  - CE_INBOUNDIP=<WNs alcanzables desde “afuera”?>
  - CE_RUNTIMEENV=“<tags de software soportado> ..."
  - CE_CLOSE_SE=“<etiqueta del SE cercano> ..."
  - CE_CLOSE_<etiqueta>_HOST=<hostname del SE>
  - CE_CLOSE_<etiqueta>_ACCESS_POINT=<punto de montaje de la partición de datos>
  - DPM_HOST=<hostname del DPM>
  - BDII_HTTP_URL=“<url de lista de BDIIs> ...”
  - BDII_REGIONS=“<tipo de nodo 1> ... <tipo de nodo N>”
  - BDII_<tipo de nodo>_URL=“<url del productor de información MDS”

Instalación con YAIM

- **Archivos de configuración: site-info.def (cont.)**
  - QUEUES="<cola 1> <cola 2> ..."
  - VO_SW_DIR=<directorio base para instalación de software de organizaciones virtuales>
  - VOS="<vo 1> <vo 2> ..." (organizaciones virtuales soportadas)
  - VO_<VO NAME>_SW_DIR=<directorio de software para la vo dada>
  - VO_<VO NAME>_DEFAULT_SE=<hostname del SE predeterminado para la vo dada>
  - VO_<VO NAME>_SGM="<url de entrada ldap donde se encuentran los Software Managers de la vo dada>"
  - VO_<VO NAME>_USERS="<url de entrada ldap o voms donde se encuentran los miembros de la vo dada>"
  - VO_<VO NAME>_STORAGE_DIR=<directorio de almacenamiento en el SE para la vo dada>
  - VO_<VO NAME>_QUEUES="<colas asociadas a la vo> .."
  - VO_ULA_VOMS_RB=<url del VOMS> (si existe)

Instalación con YAIM

- **Archivos de configuración: users.conf**
  - Lista de cuentas pool unix a donde son “mapeados” los usuarios (de acuerdo a la organización virtual a la que pertenecen)
    - UID:LOGIN:GID:GROUP:VO:SGM_FLAG
      - *UID: Id de la cuenta pool*
      - *LOGIN: nombre de usuario de la cuenta pool*
      - *GID: Id del grupo*
      - *GROUP: nombre del grupo*
      - *VO: nombre de la organización virtual*
      - *SGM_FLAG: es Software Manager? (si = “sgm”)*
- **Archivos de configuración: wn-list.conf**
  - Lista de hostnames de los WNs

Instalación con YAIM

- **Script de instalación**
  - `/opt/lcg/yaim/scripts/install_node <site-info.def> lcg-CE-torque`
  - `/opt/lcg/yaim/scripts/install_node <site-info.def> lcg-WN-torque`
  - Es posible instalar en un mismo host ambos tipos de nodo
  - Prestar atención a la salida de los scripts para detectar problemas en la instalación
  - Problema común: variables con espacios intermedios:
    - `MYDOMAIN = labf.usb.ve` (incorrecto)
    - `MYDOMAIN=labf.usb.ve` (correcto)
    - Se puede validar el archivo site-info.def:
      - `source site-info.def`

Instalación con YAIM

- **Script de configuración**
  - `/opt/lcg/yaim/scripts/configure_node <site-info.def> CE_torque`
  - `/opt/lcg/yaim/scripts/configure_node <site-info.def> WN_torque`
  - Prestar atención a la salida de los scripts para detectar problemas de configuración
  - IMPORTANTE: Configurar primero el CE y luego los WNs
    - Los esclavos de torque en los WNs deben poder comunicarse con el master durante la configuración
  - En caso de problemas es recomendable reconfigurar todo de nuevo

Diagnostico de la instalación

- **Metodología para el diagnóstico**
  - Examinar los logs de los diversos componentes:
    - `/var/log/messages`
    - `/var/log/globus-gatekeeper.log` en el CE
    - `/var/log/globus-gridftp.log` en los WNs
    - `/var/spool/pbs/server_logs` en el CE
    - `/var/spool/pbs/mom_logs` en los WNs
  - Probar el funcionamiento aislado de los componentes, a partir de las capas de mas bajo nivel:
    - SCP desde los WNs al CE y viceversa, desde las cuentas pool, sin password
      - `su - <cuenta pool>`
      - `scp <CE host>:/etc/redhat-release /tmp/prueba`

Diagnostico de la instalación

- **Metodología para el diagnóstico**
  - Envío de trabajos directamente a torque
    - Desde el CE:
      - `su - <cuenta pool>`
      - `qsub -q <cola> <script de prueba>`
    - Se debe ver el trabajo con *qstat*
    - Al finalizar, se deben generar los siguientes archivos:
      - `<script>.o<id>` (salida estándar)
      - `<script>.e<id>` (salida de error)
  - Envío de trabajos a través de globus
    - Desde el UI:
      - `grid-proxy-init`
      - `globus-job-run <hostname del CE>:2119/jobmanager-lcgpbs -q <cola> /bin/hostname`
    - Se debe ver el nombre de host donde se ejecutó el trabajo
  - Envío de trabajos con edg-job-submit
    - Usar el atributo `Requirements = ( other.GlueCEUniqueID == "<CE>:2119/jobmanager-lcgpbs-<cola>" );` en el JDL para obligar la ejecución en el CE.

Diagnostico de la instalación

- **Problemas frecuentes**
  - Problemas con la autenticación:
    - Por lo general los trabajos fallan por la siguiente razón:
      - *Failed to estabilish a security context*
    - El problema es que el certificado del usuario o del host no es válido en alguna fase de la comunicación
      - *Sincronización de los nodos (NTP!)*
      - *Falta el certificado de alguna autoridad certificadora en alguno de los extremos*
        - `openssl -CApath /etc/grid-security/certificates <certificado>`
      - *El DN del usuario no se encuentra en /etc/grid-security/gridmap-file*
        - `/opt/edg/sbin/edg-mkgridmap /etc/grid-security/gridmap-file --safe`
    - Los problemas de seguridad también causan que los trabajos fallen por otras razones

Diagnostico de la instalación

- **Problemas frecuentes**
  - Problemas con SSH entre los WN y el CE
    - En los logs de Condor-G en el RB se observa:

```
(693.000.000) 02/01 17:42:02 Job was held.
    Unspecified gridmanager error
    Code 0 Subcode 0
```

    - En /var/log/messages se observan los intentos de autenticación fallidos
    - Verificar si desde las cuentas pool se puede conectar entre el CE y los WN, *sin requerir password*
      - `su <cuenta pool>`
      - `ssh -vv <CE host>` (desde el WN)
      - `ssh -vv <WN host>` (desde el CE)
      - *Estar atentos a los mensajes de ssh en /var/log/messages*
    - Regenerar los archivos ssh-knowhosts y shosts.equiv
      - `/opt/edg/sbin/edg-pbs-knownhosts`
      - `/opt/edg/sbin/edg-pbs-shostsequiv`

Diagnostico de la instalación

- **Problemas frecuentes**
  - Problemas de transferencia de archivos
    - Los trabajos fallan por la siguiente razón:
      - *Couldn't stage out a file*
    - El CE no puede subir al RB el OutputSanbox
    - Verificar comunicación entre el CE y el RB con gridftp
      - `globus-url-copy -dbg file:///root/test.txt gsiftp://<rb host>/tmp/test.txt`
      - `globus-url-copy -dbg gsiftp://<rb host>/tmp/test.txt file:///root/test.txt`
      - `globus-url-copy -dbg file:///root/test.txt gsiftp://<ce host>/tmp/test.txt`
      - `globus-url-copy -dbg gsiftp://<ce host>/tmp/test.txt file:///root/test.txt`
    - Verificar firewalls: (iptables)

Diagnostico de la instalación

- **Problemas frecuentes**
  - Problemas de transferencia de archivos
    - Los trabajos fallan por la siguiente razón:
      - *submit-helper script running on host lxb1761 gave error: cache_export_dir (<some dir>) on gatekeeper did not contain a cache_export_dir.tar archive*
    - El WN no pueden bajar del RB el InputSanbox
    - Verificar comunicación entre el WN y el RB con gridftp
      - `globus-url-copy -dbg file:///root/test.txt gsiftp://<rb host>/tmp/test.txt`
      - `globus-url-copy -dbg gsiftp://<rb host>/tmp/test.txt file:///root/test.txt`
      - `globus-url-copy -dbg file:///root/test.txt gsiftp://<wn host>/tmp/test.txt`
      - `globus-url-copy -dbg gsiftp://<wn host>/tmp/test.txt file:///root/test.txt`
    - Verificar firewalls: (iptables)

Diagnostico de la instalación

- **Problemas frecuentes**
  - El recurso no se ve en el Information Index
    - `lcg-infosites –vo <vo>` no muestra el recurso
    - El envio de trabajos con edg-job-submit falla por la siguiente razón:
      - *Cannot plan: No compatible resources*
    - Intentar conexión al CE directamente a través de ldap
      - *ldapbrowser*
      - *URLs:*
        - `ldap://<CE>:2170/`
        - `ldap://<CE>:2135/`
- **Recomendaciones finales**
  - Al modificar cualquier variable del site-info.def, reconfigurar TODO.
  - Estar atento a los diferentes logs para rastrear los problemas

Preguntas

Gracias!