

Uso de supercomputación (HPCs) en ATLAS

Jornadas Técnicas del IFIC 2020
11/03/2020

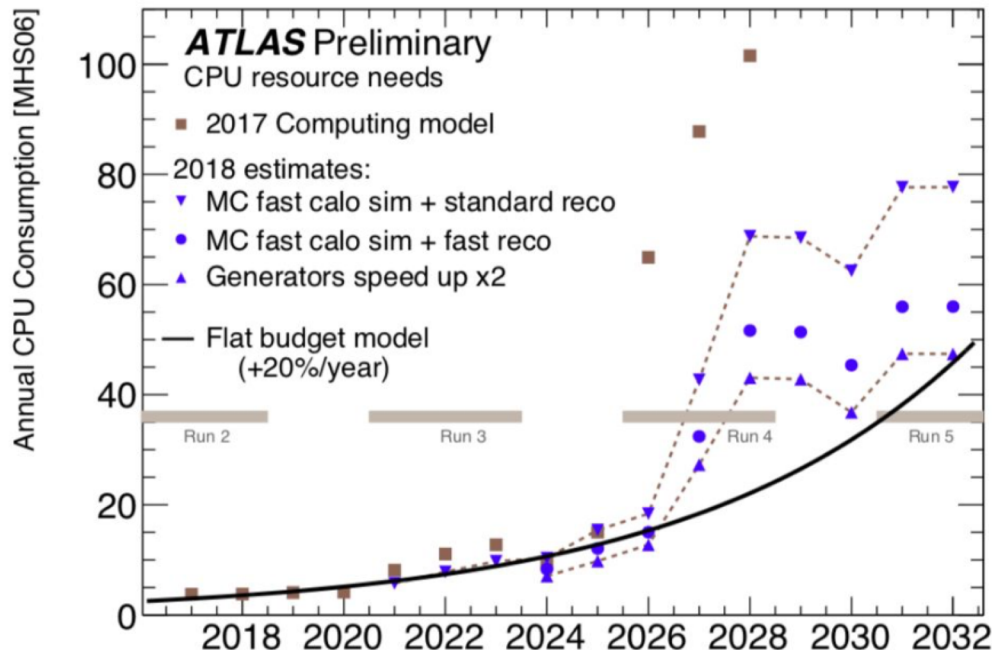
Santiago González de la Hoz
Fco. Javier Sánchez Martínez
Esteban Fullana Torregrosa

Motivación

-LHC de alta luminosidad es un reto para los recursos de computación

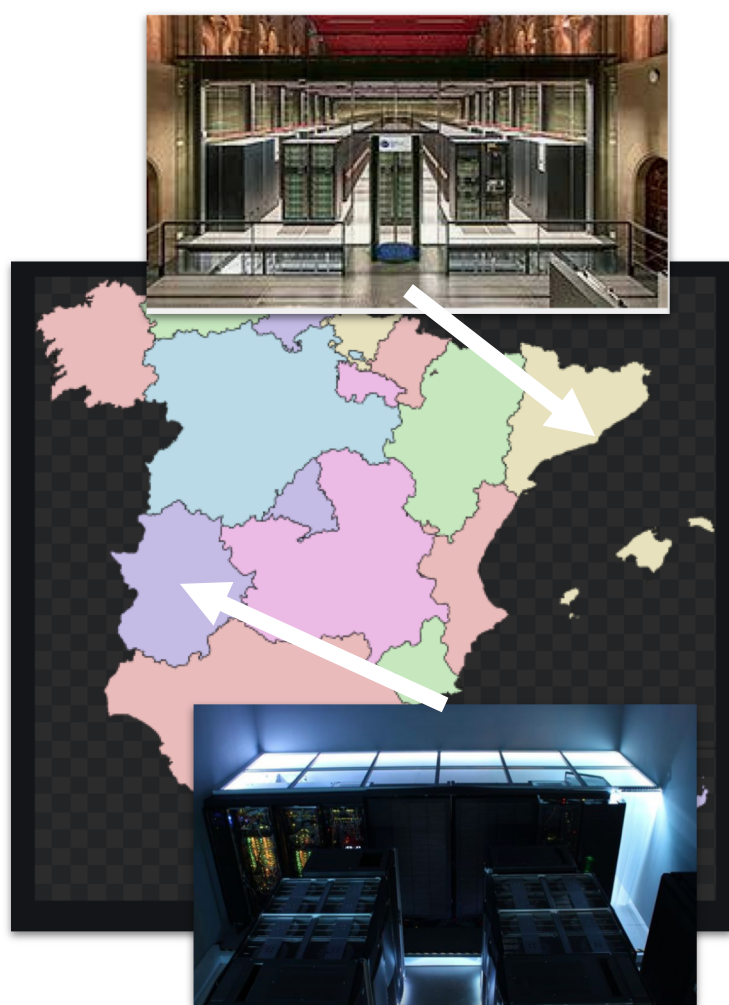
-El abaratamiento del HW no es suficientemente para compensar nuestras necesidades de crecimiento

-¿Pueden los superordenadores (HPCs) ayudarnos?



El reto

- Nuestro objetivo era usar los HPCs de la Red Española de Supercomputación (**RES**) para generar sucesos de simulación de ATLAS.
- Hemos estado haciendo peticiones de horas de CPU a la RES y a PRACE indicando la física de ATLAS como el objetivo científico.
- En total más de **7 millones de horas de CPU** nos han sido concedidas en Lusitania II y Mare Nostrum 4 en un total de 5 peticiones.



Detalles Técnicos de los HPCs

-LUSITANIA II:

Situado en el [CENITS](#)



20 Fujitsu Primergy CX2550 nodos con 2 procesadores Intel Xeon E5-2660v3 (10 cores cada uno) **sumando 400 cores** corriendo a 2,6 GHz.

Site connectivity: RedIRIS and GEANT

MARE NOSTRUM 4:

Situado en el [BSC](#)



3456 Lenovo ThinkSystem SD530 nodos. Cada nodo tiene 2 Intel Xeon Platinum con 24 cores cada uno **sumando 165.888 cores** corriendo a 2,1 GHz.

Site connectivity: RedIRIS and GEANT

Problemática

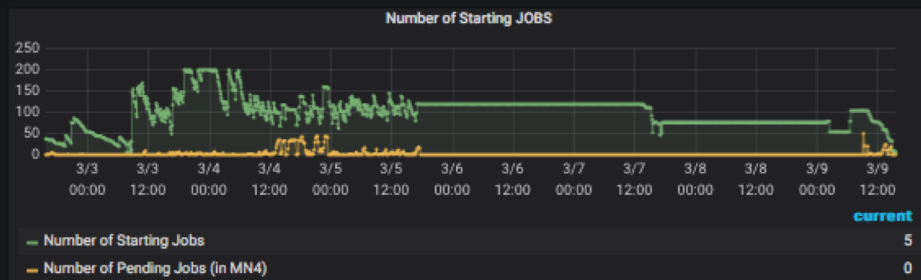
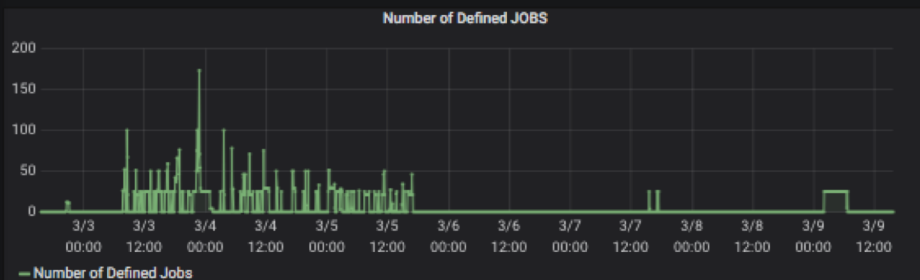
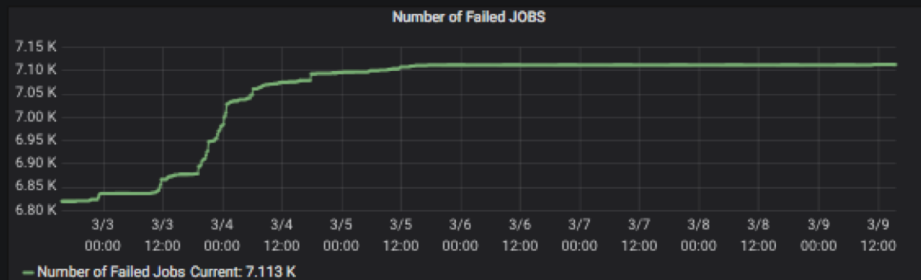
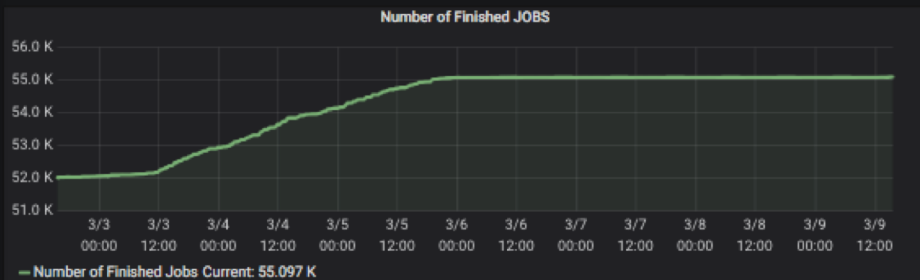
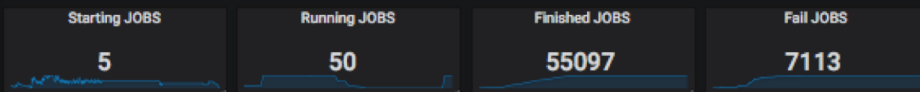
- Los HPCs no tienen (quieren) conexiones remotas en los trabajos
- El software ha de estar disponible localmente
- Tenemos que usar el sistema que nos encontramos
- La preparación de los ficheros de entrada y salida lo tenemos que gestionar localmente (IFIC)

Soluciones

- Instalar un gestor de trabajos GRID como **ARC-CE** localmente (en el IFIC) para interaccionar con el experimento ATLAS (Panda).
- Hacer creer a éste que el sistema remoto (HPC) es local:
 - Montar el directorio remoto (HPC) con **sshfs** localmente
 - Copiar inputs/outputs del sistema remoto como si fueran locales
 - Usar el batch system (**SLURM**) del sistema remoto (HPC) como si fuera local redirigiendo los comandos locales al sistema remoto con un wrapper usando ssh
 - Crear un entorno de trabajo que contiene lo necesario del experimento usando un contenedor (gigante) de **singularity**
 - Permite correr en un entorno de SW aislado dentro del SO:
 - <https://singularity.lbl.gov/>
 - Ocupa 450 GB que han de estar accesibles para los trabajos en el sistema HPC

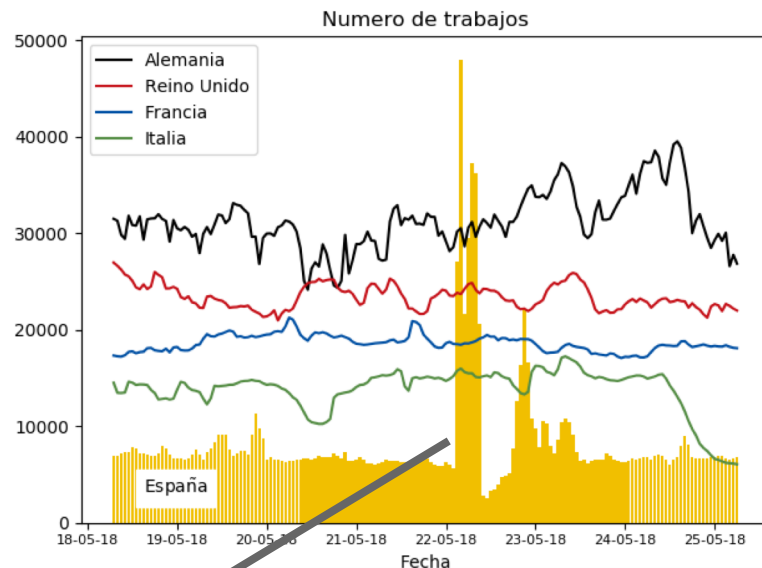
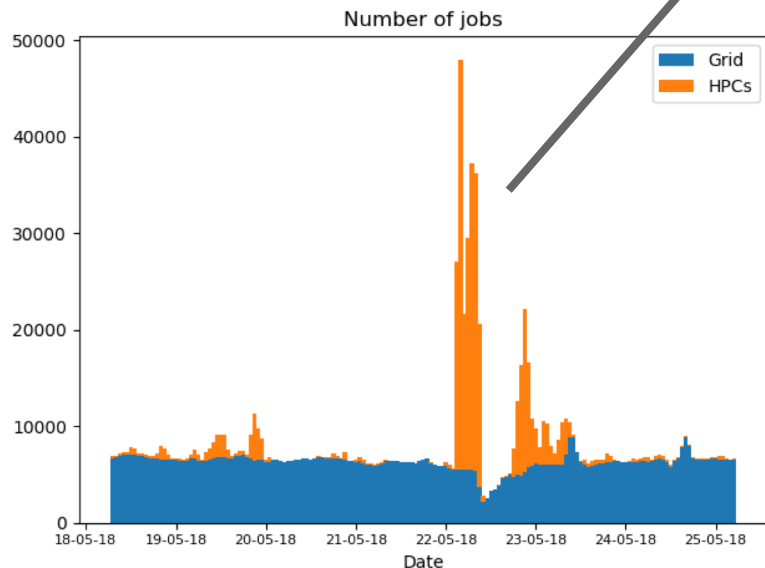
IFIC HPCs : MareNostrum4

Monitoring de los jobs de los HPCs



Logros

Impacto de la conexión MN4 al sistema de producción de ATLAS comparado con los recursos Grid (aquí sólo recursos españoles)



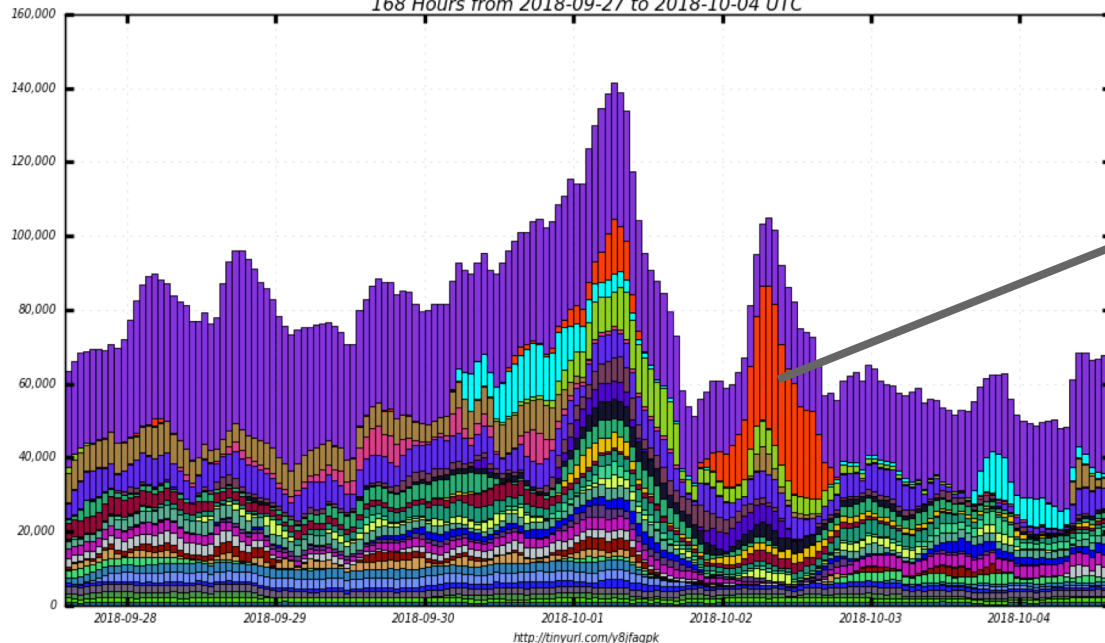
Número de trabajos corriendo en España comparado con los principales países europeos en el momento de conectar MN4 al sistema de producción de ATLAS

Cuando hemos tenido recursos HPCs, ha supuesto un **incremento alrededor del 30% en el número de “slots”** corriendo trabajos en el Tier2 del IFIC

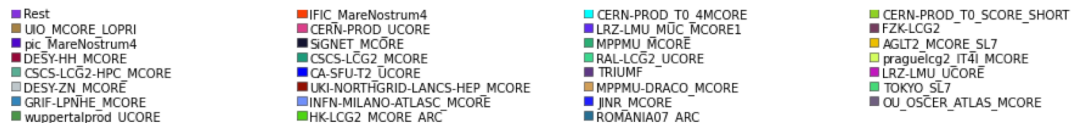


Slots of Running Jobs

168 Hours from 2018-09-27 to 2018-10-04 UTC



Números de slots
corriendo trabajos de
ATLAS. En rojo
IFIC_MN4



Maximum: 141,661, Minimum: 47,981, Average: 78,377, Current: 67,890

Conclusiones

- Los HPCs de la red española de supercomputación han simulados **más de 100 millones de eventos** de simulación completa del detector.
- Mucho trabajo *under de curtain* de desarrollo y de “*babysitting*” de los trabajos ha sido necesario con contacto periódico con los técnicos de los HPCs.
- Un trabajo que ha dado su recompensa con un éxito rotundo en el uso de los HPCs para ATLAS.

Fin

ARC-CE

- ARC-CE y el HPC comparten el directorio “session” mediante “sshfs”. Este directorio pertenece físicamente al almacenamiento de HPC, pero es visible desde el ARC-CE
 - Este directorio ha de ser accesible por el trabajo en el sistema de batch local del HPC
- Para que sea un proceso automático sin password, creamos clave de ssh y la registramos en el sistema remoto
- Configuramos el ARC-CE como si el sistema de batch (SLURM) del HPC fuera local
- Creamos las herramientas impostoras de los comandos de SLURM de manera que ejecuten comando remotos en el HPC
- Creamos un script llamado sshslurm y el resto de comandos SLURM con enlaces simbólicos a éste
- Nos aseguramos que el comando puede ejecutarse sin password mediante la clave de ssh
- Ahora el ARC-CE es capaz de enviar trabajos al HPC y recibir el output (ver flujo en transparencia 13)

Flujo de trabajo del arc-ce

1-El cliente envía el trabajo al ARC-CE

2-El ARC-CE prepara el trabajo en un directorio de “session” obteniendo los ficheros de input

3-El ARC-CE submite el trabajo al sistema BATCH local, pero en realidad es un ssh que ejecuta el comando en el sistema remoto

4-El sistema SLURM del HPC ejecuta el trabajo que tiene sus scripts en un directorio común llamado “session”

5-El ARC-CE pregunta al sistema local de colas por el estado del trabajo, pero en realidad lo está haciendo en el sistema remoto por ssh

6-Cuando acaba el trabajo, el output que está en “session” es “local” para el ARC-CE al estar montado por sshfs

ARC-CE

- Se añadió un directorio “runtime” que es compartido entre el ARC-CE y el HPC que contiene variables de entorno necesarias para los trabajos de ATLAS
 - `arc/runtime/APPS/HEP/ATLAS-SITE`
- Lo siguiente fue que los trabajos en el HPC se ejecuten dentro de singularity y que dispongan del entorno de SW apropiado para ATLAS
- Se modificó, para ello, el script que genera los trabajos para SLURM del ARC-CE `/usr/share/arc/submit-SLURM-job`; de forma que el trabajo se ejecuta dentro de singularity desde el primer momento
- Después de todo esto, se creó una cola de PANDA en AGIS, de esta forma ATLAS pudo enviar trabajos correctamente a los HPCs españoles

Monitoring de los jobs de los HPCs

