



UNIVERSIDAD TÉCNICA
FEDERICO SANTA MARÍA

WLCG

Worldwide LHC Computing Grid

Juan Carlos Céspedes Gacitúa
20 de enero de 2016

Introducción

Los servicios están constantemente cambiando y necesitan nuevos paradigmas de comunicación. Estos pueden ser con el fin de acomodar procesos o mejorarlos, y a veces también se deben conjugar con las posibilidades reales con que se cuenta y los recursos que pueden limitar ciertas implementaciones. Estas limitaciones pueden significar la generación de una nueva forma en que se acoplen los recursos con el fin de obtener sinergia.

En el presente trabajo se habla de la worldwide LHC computing grid, que surge como respuesta a los problemas presentados anteriores, tanto de comunicaciones internacionales como limitación de recursos.

En la demostración se verá la actividad de la grid a tiempo casi real, a través del software Google Earth, y un vínculo de este a un servidor del CERN.

¿Qué es el CERN?

El CERN es la Organización Europea para la Investigación Nuclear, fundada en 1954 como una cooperación internacional de los países europeos, es uno de los principales centros de investigación de física de partículas del mundo, operando importantes laboratorios, en particular el LHC, Gran Colisionador de Hadrones. Cuenta con 21 estados miembros, sin embargo en tiene colaboraciones internacionales con gran parte del mundo y en sus instalaciones convergen investigadores de todas las regiones del planeta.

¿Qué es el LHC?

El LHC, Gran Colisionador de Hadrones, con 27 kilómetros de circunferencia es el acelerador de partículas más grande y poderoso de la historia. La función de este acelerador es guiar 2 rayos de partículas a través del anillo y aumentar su velocidad hasta niveles muy cercanos a la velocidad de la luz, 99.9999991% con lo cual una partícula demora una diezmilésima de segundo en completar el circuito, para luego colisionar en los detectores, recreando eventos de interés para los científicos del CERN.

El problema del LHC

Como es esperable para un laboratorio del tamaño del LHC, los detectores generarán mucha cantidad de datos de salida, y a mayor cantidad de datos el tratamiento de la información se vuelve más compleja y costosa en términos de procesamiento. Es por esto que el LHC filtra la información importante, esto sucede en 3 etapas, la primera etapa es distinguir las colisiones interesantes de aquellas que no serán capaces de ofrecer datos importantes. Luego de seleccionar estas colisiones interesantes se filtra a través de electrónica de alta velocidad para filtrar los datos que generan los detectores sobre una de las colisiones “interesantes”, luego de esto, a través de procesado, la información continúa su filtrado. Tras esta última etapa de filtrado sólo el 1% de información ha sido capaz de superar todas las etapas, esto significa aproximadamente 30 petabytes al año de información cruda que el CERN debe hacerse cargo. En etapas posteriores estos datos crudos deben ser procesados, analizados y generarán información y conocimiento que lleven a los científicos a nuevos descubrimientos. Pero esto significa otro problema, el poder de procesamiento requerido por el LHC es significativo.

Desde un punto de vista económico establecer un centro de datos centralizado en el que los científicos trabajen y procesen sus requerimientos no es viable para el CERN, no genera un incentivo financiero a la organización pues mucho trabajo en la información se realizará desde instituciones externas al CERN, pero asociadas al experimento. Todo eso lleva a una situación en la que se debe trabajar, pero el costo de debe repartir, a fin de que sea abordable por todos los colaboradores.

Solución

El sistema que adopte el CERN para el manejo de datos del LHC debe considerar que los datos crudos se deben respaldar centralizadamente, pues es necesario asegurar la salida cruda de los detectores para múltiples análisis y como resguardo por muchos años. El sistema también debe ser transparente para los científicos pues no es ideal que actividades, como la búsqueda de poder de cómputo, les distraiga de su labor. Mantener una colaboración internacional y poder llevar la información a todos los colaboradores. Otro

punto muy importante es que los potenciales costos que le signifiquen al CERN sean menores que la opción de montar un centro de datos propio.

La solución a este problema consiste en abrir la plataforma a las instituciones colaboradoras, pero mantener la posición central del CERN. Para esto se genera una red de colaboración, en la cual las instituciones asociadas aportan con almacenamiento y poder de cómputo al proyecto, generando un super computador distribuido, transparente para el científico, la meta de la plataforma es que el usuario no se preocupe por el origen de los recursos, sino que sólo estén al alcance de él.

La solución es la computación distribuida. En este nuevo paradigma de comunicaciones se juntan muchos centros de datos y se combinan generando un “único súper computador”, el LHC aplica 3 niveles a sus colaboradores, Tier0 – Tier1 – Tier2.

TIER 0

El “Tier 0” es el primer nivel de la WLCG, este nivel se encuentra ubicado en Geneva, cercano al LHC y aquí es el primer centro de datos al cual llegan los datos generados en el LHC, este Tier se encarga de recopilar la información y reconstruirla, luego de eso genera un duplicado de la data, que envía a almacenamiento en cinta magnética para resguardo de largo plazo, y reenvía los datos a los centros Tier 1.

TIER 1

El “Tier 1” es el segundo nivel de la WLCG, este nivel también se respalda la información, se reprocesa y se generan análisis. Estos centros son de gran cómputo y corresponden principalmente a centros nacionales. Entre el Tier 0 y los Tier 1 existen enlaces de aproximadamente de 10Gbps cada uno, pero esto puede variar, la topología es forma estrella

WLCG

Para cumplir con todos los objetivos propuestos, la red debe contar con un middleware que una todas las plataformas y sea capaz de trabajar en distintos entornos, para esto se utiliza EMI. También es necesario que el acceso a la red sea autenticado y para esto se generan organizaciones virtuales, VO, que se conecten, repartan la carga de trabajo y autenticuen a los usuarios, para esto existen estructuras como EGI. En cuanto a los servidores de la red, es importante que estén actualizados y tengan un buen manejo de archivos, para esto se pueden utilizar soluciones como QUATTOR y CASTOR, creados por CERN.

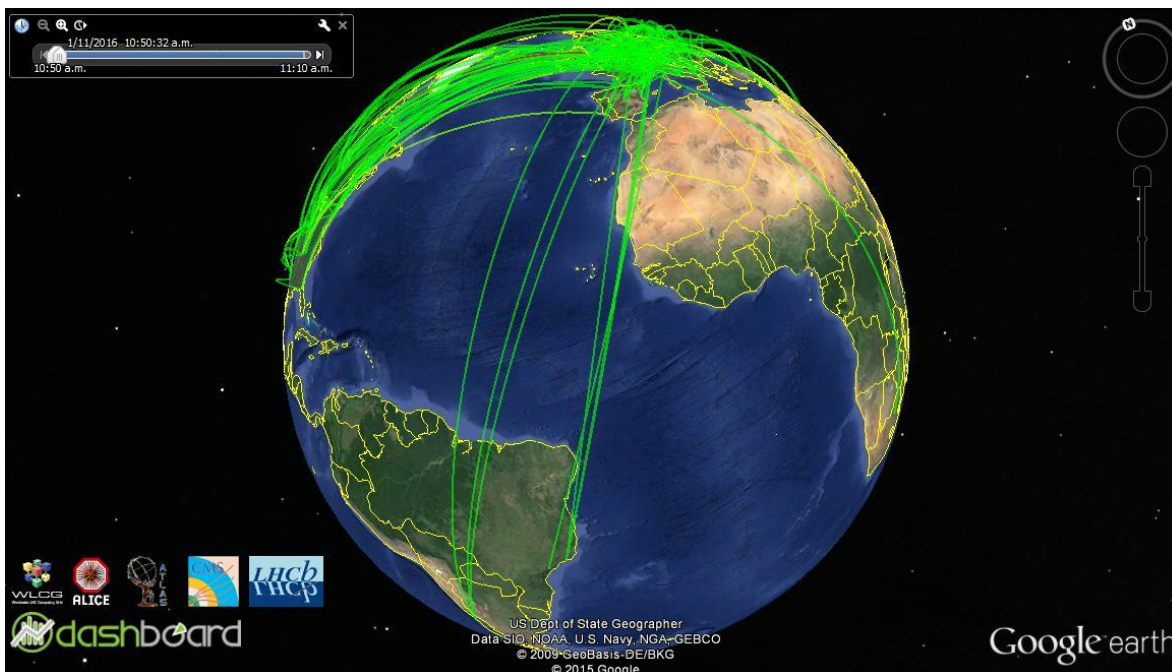
EELA-UTFSM

La Universidad Técnica Federico Santa María es la única universidad chilena, y una de las pocas de Latinoamérica, que está conectada a la WLCG, a través del proyecto EELA, E-science grid facility for Europe and Latin America.

En el siguiente sitio (<https://wlcg-rebus.cern.ch/apps/topology/federation/421/>) podemos ver la estructura del tier 2 "Latin America Federation", del cual es parte la Universidad Técnica Federico Santa María, y también es posible ver los datos de la institución. En el sitio se encuentra la información respectiva a todas las instituciones y TIER.

Visualización de la actividad de la red en Google Earth

Es posible visualizar la actividad de la red en Google Earth, para esto los servidores actualizan un complemento cada 10 minutos con la máxima información posible que pueda poner en el completo.



Para esto se debe ir a la pestaña “añadir” del menú de Google earth y seleccionar “Vinculo de red”. Se despliega una pantalla y es posible otorgar un nombre a elección, en el vinculo se debe escribir el enlace <http://dashb-earth.cern.ch/dashboard/dashb-earth-all.kmz> y ajustar el periodo de actualización a 1 minuto, en la pestaña actualizar de la ventana, menú “Actualización basada en la hora”.

Luego en la sección izquierda estará nuestro vinculo creado con el nombre, es importante que esté seleccionado.

Conclusión

A través del presente trabajo se investigó de la implementación de una red de computadores en malla por el CERN para su laboratorio LHC. Como se pudo ver, este sistema es robusto, permite a los científicos acceder a un alto poder de cómputo y es capaz de distribuir los costos entre todos los actores involucrados al proyecto. Es, además, eficiente en el uso de recursos, considerando que los recursos computacionales de los servidores puestos en la red se utilizan desde todo el mundo. Esta cooperación internacional, además de disminuir los costos al CERN permite acelerar los descubrimientos, pues hace llegar la información a una mayor cantidad de investigadores.

WLCG es una red amplia, que ha demostrado ser aplicable a otros proyectos que necesiten distribuir sus servidores de esta manera. Si bien, aun no queda claro que las redes grid sean el próximo paso en el internet, es anecdótico mencionar que el World Wide Web nació en los laboratorios del CERN con el fin de compartir información entre investigadores. En este caso se comparten recursos computacionales, veremos en el futuro si se adopta esta tecnología en el mercado para usuarios finales.