Optimización de la red para la transferencia masiva de datos en entornos Grid

Optimization of Network Resources for Mass Data Transfer in Grid Environments

R. Marco y L. Serrano

Resumen

En este artículo se aborda desde un caso concreto la problemática que lleva a la necesidad de una optimización del protocolo de control de transmisión de red (TCP) y se presentan alternativas al TCP tradicional en fase de estudio. Asimismo se comenta una transferencia masiva de datos CERN-IFCA y cómo la red resulta ser un aspecto crítico para la investigación en el IFCA.

Palabras Clave: Grid, eCiencia, Red, TCP, Transferencia de datos.

Summary

In this article reasons for an optimization of the network transmission control protocol (TCP) and some alternatives to classical TCP, still in test phase are presented.

The experience of a massive data transfer CERN-IFCA shows why the network becames a critical resource for research activity at IFCA.

Keywords: Grid, eScience, Network, TCP, data transfer.

La participación del IFCA en iniciativas y proyectos Grid durante los últimos seis años ha sido intensa a nivel nacional e internacional

Se comenta una transferencia masiva de datos CERN-IFCA y cómo la red resulta ser un aspecto crítico para la investigación

1.- Problemática de red y computación en el Instituto de Física de Cantabria

El Instituto de Física de Cantabria [1] es un centro mixto del Consejo Superior de Investigaciones Científicas y de la Universidad de Cantabria. Sus principales áreas de investigación son Astrofísica y Estructura de la Materia (Altas Energías y Física Estadística). En todas sus líneas de investigación la intensa participación en grandes proyectos y colaboraciones internacionales (p.ej. DELPHI y CMS en el CERN, CDF en Fermilab, Misión Planck y XMM-Newton de la ESA, ...) ha puesto de manifiesto que la red y la computación son dos elemento críticos de trabajo.

Esto dio origen al nacimiento de una línea de investigación en Computación Distribuida Grid en el IFCA y al desarrollo de un centro de eCiencia [2], que además del interés por si misma tiene un valor añadido por su proximidad y trato directo con los grupos interesados en aprovechar esta tecnología. La participación en iniciativas y proyectos Grid durante los últimos seis años ha sido intensa a nivel nacional (IRISGrid, LCG-ES) e internacional (DataGrid, CrossGrid, LCG, EGEE), en los que se ha trabajado principalmente en tareas de coordinación, desarrollo de testbeds, aplicaciones y seguridad. En el 2006 se continuará en la misma línea con los proyectos INT.EU.GRID, EGEE2 y EELA. Si bien la iniciativa inicial surgió del grupo de Altas Energías no cabe duda que ha supuesto un gran beneficio para todos los grupos de investigación en el IFCA.

Con objeto de resaltar la importancia de ser el acceso a estos recursos de red y computación merece la pena comentar la forma de trabajo dentro de una gran colaboración científica que genera una enorme cantidad de datos como es CMS: se suelen fijar semanas en las que se presenta y discute todo el trabajo, los investigadores que trabajan en analizar los datos del experimento obtienen los datos con las últimas calibraciones pocos días antes de las reuniones y necesitan procesarlo rápidamente aplicando los métodos de análisis que han desarrollado. También es habitual realizar reuniones de integración en las que se reúnen varios grupos con el objetivo de trabajar de forma intensiva. En

cualquier caso si los recursos de red y computación no están a la altura necesaria suele ser casi imposible alcanzar los objetivos científicos y por tanto participar de forma eficiente en la colaboración.

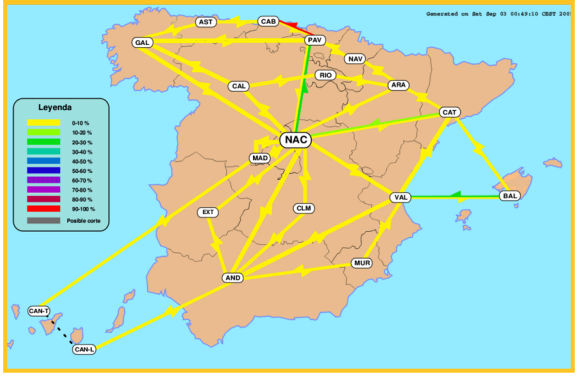
2.- Transferencia de Datos del experimento DELPHI (CERN-IFCA)

La importante participación del IFCA en el experimento DELPHI ha generado un gran interés en mantener los datos disponibles en el IFCA para futuros usos, a pesar de que el experimento terminó en 2001. Existe la posibilidad de que alguna idea novedosa o descubrimientos en futuros experimentos puedan generar interés en hacer un re-análisis de estos datos.

El volumen total de datos de DELPHI es aproximadamente 36TB distribuido en 300 cintas, en 500K ficheros, y se encuentra almacenado en la librería de cintas del CERN que es manejada por el software CASTOR y que está enganchada a la infraestructura Grid, de forma que es posible acceder a los datos a través del protocolo GridFTP. El proceso de copia duró aproximadamente dos semanas en las que se intentó sacar el máximo rendimiento del acceso a RedIRIS2 [3] del IFCA (622Mbps) tal y como se puede ver en la figura 1.

La optimización del uso de los recursos no es trivial. El mecanismo sencillo de copia fichero a fichero no resulta eficiente

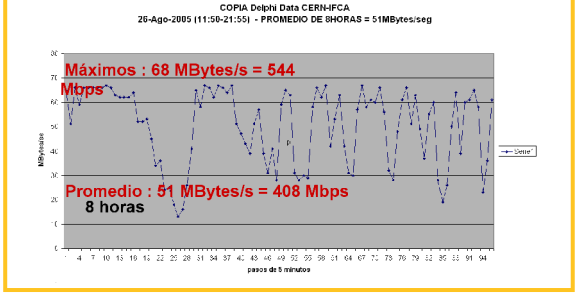
FIGURA 1: WHEATHER MAP DE RedIRIS DURANTE LA TRANSFERENCIA MASIVA DE DATOS CERN-IFCA



La optimización del uso de los recursos no es trivial. El mecanismo sencillo de copia fichero a fichero no resulta eficiente. Al copiar un fichero con GridFTP automáticamente se monta la cinta que contiene el fichero, este es copiado a un pool de disco, se inicia la copia y la cinta es desmontada. El tiempo dedicado a montar y desmontar la cinta y acceder a la posición de un fichero es demasiado grande para repetirlo medio millón de veces. Para solucionar esto se creó un gestor del proceso de copia que la organiza de forma que cada vez que se monta una cinta con un volumen suficientemente grande de datos es volcada a disco y copiada; las futuras versiones de CASTOR resolverán este problema a través de un scheduler mejorado. Por otra parte se dispuso de recursos limitados de almacenamiento en disco duro en el CERN que posiblemente hayan afectado al rendimiento sostenido. En el IFCA el almacenamiento se realizó en un pool de disco de forma que no se perdiese eficiencia por limitaciones de almacenamiento. La copia se realizó desde 15-30 nodos en paralelo en el IFCA, y desde cada nodo se realizaron copias fichero a fichero con 10 threads paralelos.

La velocidad de transferencia real se monitorizó. Se alcanzaban máximos de 68 MBytes/s (544 Mbps)

FIGURA 2: ESTADÍSTICA DE 8 HORAS DE TRANSFERENCIA DE DATOS CERN-IFCA



La velocidad de transferencia real se monitorizó contabilizando el volumen de datos transferidos cada cinco minutos. El resultado fue que se alcanzaban máximos de 68 MBytes/s (544 Mbps) durante periodos pero frecuentemente el rendimiento se reducía aproximadamente a la mitad. Uno de los motivos puede ser

la sobrecarga del sistema de almacenamiento y otro motivo puede ser los problemas de congestión de red tal y como aparecen en la figura 2.

3.- Problemática del protocolo de transporte TCP

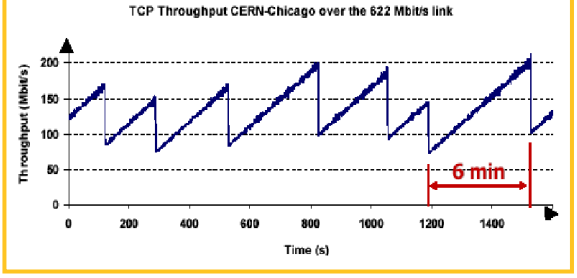
El protocolo TCP es el más usado en Internet. Está orientado a establecer una conexión en la que se garantiza la integridad de la transmisión. Incorpora un control de flujo y uno de congestión a través del uso de ventanas de transmisión. La ventana de transmisión define el volumen de datos enviado por cada mensaje de confirmación de la buena recepción de datos.

Estudios recientes muestran un rendimiento muy bajo en redes de alta capacidad. Así, por ejemplo, en enlaces de 622Mbps, usando flujos de al menos 10Mbytes, el 90% de los flujos usa menos de 5Mbps y el 99% menos de 20Mbps. En ocasiones una posible justificación puede encontrarse en la máquina extremo (limitaciones de procesamiento, velocidad de bus, configuración de parámetros TCP: tamaños de buffers). Pero otra posible justificación de la disminución del rendimiento está en el propio TCP y en concreto en su mecanismo de control de las congestiones.

Una posible justificación de la disminución del rendimiento de red está en el propio TCP y en concreto en su mecanismo de control de las congestiones

El control de congestión de TCP se realiza usando el mecanismo AIMD (Additive Increase and Multiplicative Decrease). y consiste en que partiendo de una tasa de transferencia baja se va aumentando 1 paquete por RTT (Round Trip Time) y al observarse pérdidas la tasa se reduce a la mitad. Este mecanismo funciona muy bien para redes de 10/100Mbps locales pero al aplicarlo a redes grandes con gran ancho de banda el tiempo de recuperación tras una pérdida por congestión es demasiado largo y se pierde mucho rendimiento (ver figura 3).

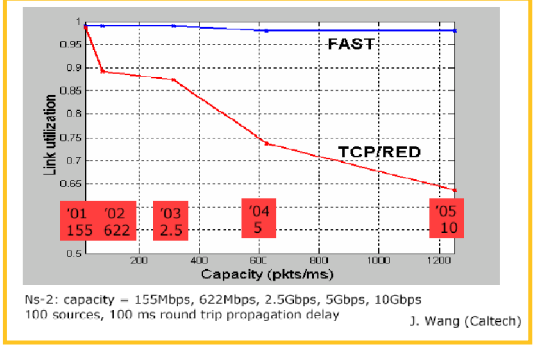
FIGURA 3: EL MECANISMO DE CONTROL DE TCP REDUCE LA TASA DE TRANSFERENCIA A LA MITAD



Dentro del proyecto DataTag [4] se analizaron varias alternativas para sustituir este mecanismo TCP tradicional. Como ejemplo comentaremos tres:

- El High-Speed TCP (HSTCP) usa un mecanismo de estabilización y decrecimiento que depende de la ventana de transmisión en vez de seguir un comportamiento constante. Se consigue un tiempo de estabilización más rápido y se disminuye el tiempo de decrecimiento.

FIGURA 4: MEJORA DE RENDIMIENTO



Dentro del proyecto DataTag se analizaron varias alternativas para sustituir este mecanismo TCP tradicional, como:
• El High-Speed TCP
• El Scalable TCP
• El FAST-TCP

- El Scalable TCP (S-TCP) es un refinamiento del HSTCP que modifica el valor de la ventana de transmisión en cada RTT, aumentándolo si se ha recibido correctamente y disminuyéndolo si hay pérdidas.
- El FAST-TCP usa el RTT, retardo del ACK de cada paquete, como primer parámetro para detectar una posible señal de congestión. La figura 4 muestra la espectacular mejora de rendimiento en redes de gran capacidad.

4.- Conclusiones

La Red jugará un papel fundamental en los próximos años. Actualmente estamos en un proceso de preparación para el gran salto de necesidades de computación y red que supondrá el inicio de los experimentos de LHC en el CERN. En concreto para el IFCA es probable que en el 2007-2008 se necesiten aprovechar de forma eficiente anchos de banda de 2.5-10 Gbps.

Será necesaria una optimización del uso de los recursos de red. Por una parte el buen aprovechamiento de la red puede verse limitado por la infraestructura de computación involucrada, por ejemplo por posibles problemas de gestión de recursos y almacenamiento y por otra el uso de grandes anchos de banda hará necesario el uso de protocolos de TCP mejorados para evitar grandes pérdidas de rendimiento por congestión de red.

Nuevos protocolos TCP que reducen las pérdidas por congestión de red han sido probados con éxito en el proyecto DataTag. La implementación generalizada de estos TCPs mejorados necesitará un acuerdo general y probablemente será incluido en los nuevos kernels.

Actualmente estamos en proceso de preparación para el gran salto de necesidades de computación y red que supondrá el inicio de los experimentos de LHC en el CERN

El proceso de investigación y pruebas continúa. En la actualidad en el marco de los Grupos de Trabajo de TERENA RedIRIS colabora en proyectos relacionados con estos protocolos que mejoran el rendimiento junto con otras redes académicas.

RedIRIS e IFCA están colaborando para la realización de pruebas similares aplicadas a necesidades de trabajo real como son las copias masivas entre el CERN y el IFCA.

Agradecimientos

Gracias a Jesús Marco, Iban Cabrillo del IFCA y al equipo de CastorGrid del CERN coordinado por Tony Osborne por su colaboración para la transferencia CERN-IFCA de los datos de Delphi.

Referencias

[1] Instituto de Física de Cantabria. http://www.ifca.unican.es

[2] Proyectos Grid en el IFCA. http://grid.ifca.unican.es

[3] Infraestructura de Red de RedIRIS. http://www.rediris.es/red

[4] Proyecto Datatag. Resultados y Figuras. http://www.datatag.org

En el marco de los Grupos de Trabajo de TERENA RedIRIS colabora en proyectos relacionados con estos protocolos que mejoran el rendimiento junto con otras redes académicas

Rafael Marco de Lucas
(rmarco@ifca.unican.es)
Instituto de Física de Cantabria - (CSIC-UC)
Laura Serrano
(laura.serrano@rediris.es)
Área de red
RedIRIS