

Evaluación de la Tecnología ATM en Clusters de Alto Rendimiento para aplicaciones en Física de Altas Energías

Rafael Ángel García Leiva*
Departamento de Física Teórica
Universidad Autónoma de Madrid

27 de Enero de 2004

Abstract

En el presente documento se analiza la viabilidad de utilizar la tecnología de red ATM, en modo de emulación de redes LAN, para la construcción de un cluster de PCs. La tecnología ATM es comparada con las tecnologías de red Fast Ethernet y Gigabit Ethernet. Los criterios de evaluación están basados fundamentalmente en consideraciones de rendimiento y de precio, y han sido diseñados teniendo en cuenta las necesidades típicas de las aplicaciones en el campo de la Física de Altas Energías.

1 Introducción

Actualmente se encuentra en construcción en el Laboratorio Europeo de Física Nuclear (CERN) en Ginebra, un nuevo acelerador de partículas llamado LHC (del inglés *Large Hadron Collider* [Eva-96]). El acelerador LHC colisionará protones e iones utilizando niveles de energía hasta ahora nunca conseguidos. Este nuevo colisionador permitirá a los científicos penetrar en la estructura de la materia, y recrear las condiciones que existían en el universo primitivo, justo después del *Big Bang*.

Los requerimientos computacionales del LHC, es decir, las necesidades de cálculo, de almacenamiento y de comunicación de datos, son extremadamente elevados. Está previsto que el cuando el nuevo acelerador entre en

* Dirección de correo electrónico: angel.leiva@uam.es

funcionamiento, en el año 2007, genere del orden de 12 a 14 PetaBytes de datos anuales. Se estima que para el procesamiento y análisis de esta información se requerire del equivalente a 70.000 PCs de hoy día. El laboratorio europeo CERN se ha comprometido a proporcionar un tercio de esta capacidad de cálculo y de almacenamiento. Siendo responsabilidad de los más de 150 institutos, de 34 países diferentes, que colaboran en el proyecto LHC, proporcionar el resto.

Gestionar y coordinar un entorno de computación tan complejo, representa un problema tecnológico sin precedentes en la historia de la informática. El grupo de computación LCG (*LHC Computational Group*), responsable de proporcionar la metodología y las herramientas necesarias para el análisis de los datos del LHC, apostó por las tecnologías *Grid* como solución a estos problemas computacionales y de organización. Los entornos Grid nos proporcionan las herramientas necesarias para realizar tareas de cálculo intensivo y de gestión de grandes volúmenes de datos en entornos distribuidos. Una introducción a los sistemas Grid y sus características puede encontrarse en [Fos-98].

El software elegido por LCG para el análisis y procesamiento de los datos que serán generados en por el LHC es el proporcionado por *DataGrid*. DataGrid [Gal-02] es un proyecto financiado por la Unión Europea con el objetivo de construir la siguiente generación de infraestructuras de cálculo, proporcionando computación intensiva y análisis de bases de datos compartidas de gran tamaño, desde cientos de TeraBytes hasta PetaBytes, a lo largo de comunidades científicas ampliamente distribuidas ¹.

1.1 Almacenamiento y Cálculo en DataGrid

El la Figura 1 se puede ver un esquema simplificado de la arquitectura de un entorno Grid según el proyecto DataGrid. Los usuarios del Grid acceden a los recursos disponibles a través de las máquinas Interfaz de Usuario (*User Interface*). Los trabajos enviados al Grid son gestionados por un Gestor de Recursos (*Resource Broker*), que en base a la información proporcionada por los Servicios de Información (*Information Services*), decide en que equipos han de ejecutarse. Junto con estos elementos, existe un Servicio de Autenticación y de Autorización (*Authentication and Authorization*) de usuarios, y un Registro de Eventos (*Logging and Bookkeeping*) con información sobre los trabajos procesados en el Grid. Para más información sobre DataGrid y su arquitectura consultar [EDG-03].

Cada instituto colaborador con LCG ha de proporcionar, dentro de sus

¹El software DataGrid puede ser descargado libremente desde www.edg.org

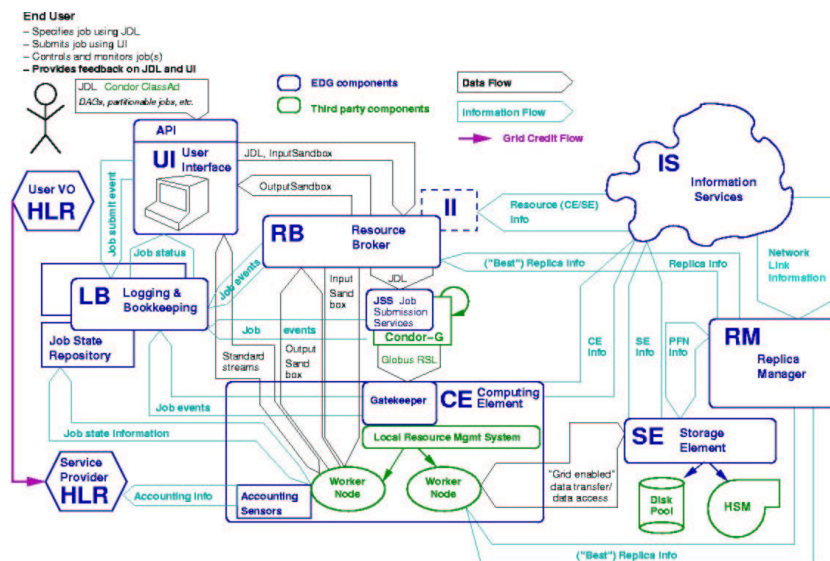


Figure 1: Arquitectura de DataGrid

posibilidades, un entorno de computación compuesto por uno o más Elementos de Cálculo (*Compute Element*), con sus correspondientes Nodos de Trabajo (*Worker Nodes*) asociados, y uno o más Elementos de Almacenamiento (*Storage Elements*).

Los Elementos de Computación, o CE, son los responsables de la ejecución final de los trabajos enviados al Grid. Un CE debe de proporcionar capacidad de cálculo, a través de un conjunto de Nodos de Trabajo (o WN) asociados. Los WN pueden ser supercomputadoras tradicionales, o bien, clusters de PCs. Un CE proporciona, además de la capacidad de cálculo a través de sus WN, una puerta de acceso (*gatekeeper*), un software de autenticación y autorización, y un gestor local de trabajos.

Los Elementos de Almacenamiento (o SE) proporcionan la capacidad de almacenamiento necesaria para realizar los trabajos en el Grid, guardando tanto los ficheros de entrada de los trabajos, así como los ficheros de salida, y todos aquellos otros ficheros temporales que sean necesarios para completar el trabajo. Un SE suele disponer de un gran número de discos de almacenamiento, generalmente en forma de *arrays*, y opcionalmente, de algún mecanismo de almacenamiento masivo, por ejemplo, cintas magnéticas u otros dispositivos similares. Los SE también disponen de un software de alto nivel, que oculta las complejidad y las diferencias entre los distintos SE, proporcionando un acceso homogéneo a la información en todo el Grid.

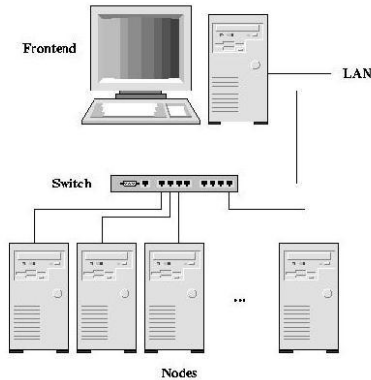


Figure 2: Cluster de PCs

1.2 Elementos de Computación basados en Clusters de PCs

Actualmente, y debido a su excelente relación rendimiento vs. precio, la práctica totalidad de los CE que forman parte del entorno de pruebas DataGrid, y el entorno de desarrollo de LCG, proporcionan las capacidades de calculo comprometidas, mediante el uso de *Clusters de PCs*.

Por cluster entendemos un grupo de ordenadores independientes, interconectados mediante una red de área local privada, y que trabajan coordinados para resolver un mismo problema [Gar-03]. En la Figura 2 podemos ver un ejemplo de la configuración más general y simple de cluster de PCs. En esta configuración tenemos dos o más PCs interconectados utilizando una red LAN de alta velocidad. Uno de los equipos, llamado *nodo frontal*, dispone de dos tarjetas de red, una conectada a la red de comunicaciones del centro donde se encuentre, y a otra conectada a la red local privada del cluster. El nodo frontal dispone de teclado, de monitor y de ratón, y es utilizado por el resto de los nodos como pasarela al mundo exterior, como puerta de acceso para los usuarios, y como servidor de ficheros. El resto de los equipos, llamados *nodos de cálculo*, o simplemente nodos, normalmente son “descabezados”, es decir, no disponen de teclado, ratón o monitor, y se utilizan exclusivamente para la ejecución de los trabajos.

Actualmente, la mayoría de los CE que existen en DataGrid, utilizan Linux como sistema operativo, y una combinación de red Fast Ethernet para la interconexión entre los nodos del cluster, y Gigabit Ethernet para las conexiones del nodo frontal y de los servidores de ficheros.

En el presente trabajo se estudia la viabilidad y la conveniencia de uti-

lizar la tecnología ATM, en su modo de emulación de redes locales, para interconectar los nodos de los clusters de PCs en los que se basan los CE de DataGrid y LCG.

2 Tecnologías para redes LANs

En esta sección se repasa, muy brevemente, las características de las tecnologías de red Ethernet y ATM, mencionando cuales son los tipos de redes objetivo para las que fueron diseñadas. A continuación se propone un conjunto de criterios que nos permitan comparar ambas tecnologías de red. Y por último se mencionan cuales son las características y necesidades propias de las aplicaciones en Física de Altas Energías (o HEP, *High Energy Physics*), objetivo último de los clusters de PCs de LCG.

2.1 Redes Ethernet

Ethernet es una tecnología para la interconexión de ordenadores en redes de área local. Ethernet fue inventada por Bob Metcalfe [Met-76] en 1973 en el laboratorio PARC de Palo Alto, propiedad de Xerox Corporation. Ethernet define un conjunto de protocolos, que según el modelo descrito por el estándar ISO/OSI de interconexión de sistemas abiertos [Day-83], se corresponden con la capa de enlace de datos (es decir, cómo la información se transmite), y la capa física (cableado y circuitería). Ethernet se basa en el sistema CSMA/CD (*Carrier Sense, Multiple Access, with Collision Detection* - detección del medio, accesos múltiples, con detección de colisiones). Las tramas Ethernet tienen una longitud de 1500 bytes, junto a con una cabecera de 14 bytes, y una cola de 4 bytes, y son transmitidas a través de un medio físico compartido por todos los equipos conectados a la red. Actualmente se dispone de redes Ethernet a velocidades de 10Mbit/sec, 100Mbit/sec (también conocida como *Fast Ethernet*), y 1000Mbit/sec (conocida como *Gigabit Ethernet*).

2.2 Redes ATM

ATM (*Asynchronous Transfer Mode* - Modo de Transferencia Asíncrona) es una tecnología para la interconexión de redes de área global, especialmente diseñada para la transmisión de voz, datos y vídeo. ATM es un estándar propuesto por un consorcio internacional, formado por empresas, organizaciones gubernamentales, y centros de investigación. Para más información sobre la tecnología ATM consultar [ATM-Forum]. ATM se caracteriza por la

necesidad de establecer una conexión (*connection oriented*) antes de iniciar la transmisión de los datos, y por la transferencia de información mediante la conmutación de paquetes. Resulta difícil describir el estándar ATM utilizando como base el modelo ISO/OSI. Las celdas transmitidas por ATM tienen una longitud de 53 bytes, de los cuales 48 bytes se dedican a la transmisión de los datos, y el resto es una cabecera con la información sobre la celda. ATM puede trabajar a velocidades que sean múltiplos del llamado OC-1, es decir, múltiplos de 51.48Mbits/seg. Velocidades de transferencia populares para ATM son OC-3 (155.52 Mbits/seg.), y OC-12 (633.08Mbits/seg.).

2.2.1 Emulación de redes LAN con ATM

ATM se diseñó inicialmente como estándar para la interconexión de redes de área global. Sin embargo, y debido a las limitaciones del ancho de banda ofrecidas por algunas las tecnologías LAN, son muchos los centros que optaron por implementar tecnologías ATM en sus redes locales. Para más información sobre la tecnología ATM aplicada a redes LAN, consultar [New-94] y [Thu-95].

Para utilizar ATM nativo en una LAN, lo único que hay que hacer es considerar la LAN como si fuese una red WAN, basándose en los protocolos WAN. El mayor inconveniente de esta solución se encuentra en las aplicaciones que utilicen la red, ya que, normalmente, suelen requerir para la comunicación de información protocolos basados en TCP/IP, como Sockets BSD, llamadas a RPC de Sun Microsystems, o el paso de mensajes a través de librerías como MPI o PVM. En el caso de optar por ATM nativo, habría que reescribir estas aplicaciones para utilizar los nuevos interfaces proporcionados por ATM.

Alternativamente, y con el objetivo de no tener que reescribir ninguna aplicación, puede utilizarse el estándar LANE (*LAN Emulation*) de emulación de redes LAN bajo ATM. LANE es un estándar propuesto por el ATM-Forum [LANE], que nos proporciona LANs virtuales conmutadas, simulando el medio compartido en los niveles de comunicación superiores. Es decir, LANE lo que hace es simular el medio físico y su acceso, dejando invariables la pila de protocolos que hubiese implementada sobre este medio, y por tanto, los interfaces de acceso a la red. Eso nos garantiza la compatibilidad software disponible en la nueva arquitectura.

2.3 Criterios de Evaluación

Los criterios tecnológicos fundamentales con los que podemos comparar las tecnologías de redes LAN utilizadas para la interconexión de los nodos que

forman un cluster de PCs son dos: *latencia* y *ancho de banda*. Otro tipo de criterios, no menos importantes, a tener en cuenta a la hora de evaluar cualquier tecnología de red, son los de tipo económico. Todos estos criterios son definidos y analizados en esta sección.

2.3.1 Criterios Tecnológicos: Latencia y Ancho de Banda

Por latencia entendemos el tiempo que transcurre desde que una aplicación solicita un dato residente en un equipo remoto, hasta que este dato está disponible. Por ancho de banda entendemos la cantidad de información transmitida por una red en una unidad de tiempo. La latencia se suele medir en milisegundos o microsegundos, y el ancho de banda en megabits por segundo (Mbits/seg).

La latencia entre dos nodos se ve afectada fundamentalmente por el retraso introducido por el hardware y software de comunicaciones. Dentro de los retrasos por hardware tenemos aquellos que se deben al retraso del interfaz de red, del conmutador, la propagación de la señal y la velocidad del bus de datos de los equipos. El retraso debido al software está causado fundamentalmente por las iteraciones del software de aplicación con la pila de protocolos de comunicaciones, y con el driver que controla el interfaz de red [Lin-95].

El ancho de banda viene determinado por la tecnología de red utilizada, y por las características del interfaz de red del equipo. Nótese que en el caso de tecnologías de red como Ethernet, donde el medio de transmisión es compartido, el ancho de banda disponible debe ser dividido entre todos los equipos conectados a la red.

La importancia de la latencia y del ancho de banda a la hora de evaluar una tecnología de red, depende de las características de las aplicaciones que vayan a ejecutarse. Las aplicaciones en red se caracterizan por la relación existente entre el tiempo que pasan realizando cálculos, y el tiempo que se necesita para la transmisión de datos. Aquellas aplicaciones que necesiten comunicar pequeños volúmenes de datos, pero de manera muy frecuente, necesitarán tecnologías de red que proporcionen una latencia muy baja. Por el contrario, las aplicaciones que necesitan transmitir grandes volúmenes de datos, en intervalos relativamente separados en el tiempo, requieren de una tecnología de red que proporcione un gran ancho de banda.

2.3.2 Criterios Económicos

Los elementos de carácter económico más importantes a tener en cuenta en la evaluación de una tecnología de red son: el precio de los interfaces (tarjetas)

de red, precio de los elementos de interconexión, como conmutadores y concentradores, y precio del medio físico de transmisión (cables de cobre o fibra óptica). También hay que considerar la disponibilidad de los componentes en el mercado, en cuanto al número de fabricantes que proporcionan hardware de una determinada tecnología, así como el número de distribuidores que nos permitan adquirirla.

Finalmente, la disponibilidad de drivers para el sistema operativo que se esté utilizando, sobre todo si se trata de Unix en alguna de sus variantes (Linux, Solaris, ...), puede llegar a ser un criterio de evaluación decisivo.

2.4 Características de las Aplicaciones HEP

Las aplicaciones en física de altas energías pueden ser clasificadas, muy genéricamente, en tres grandes grupos: simulación de sucesos, reconstrucción de señales, y otros tipos de análisis específicos.

La simulación de sucesos suele realizarse con el programa *Geant4*. Geant4 es un conjunto de utilidades para la simulación del paso de partículas a través de la materia. Geant4 nos proporciona herramientas para simular todos los aspectos de los detectores: geometría, trayectorias, respuesta del detector, gestión de eventos y trayectorias, y visualización del detector. Para más información sobre Geant4 consultar [Ago-03].

La reconstrucción de señales se suele realizar con la utilidad *Root*. Root es un entorno de programación orientado a objetos, que nos proporciona toda la funcionalidad necesaria para el manejo y análisis de grandes volúmenes de datos de manera eficiente. Root ha sido especialmente diseñado para aprovechar las ventajas que nos proporcionan las consultas a bases de datos distribuidas, y puede ejecutarse tanto en máquinas multiprocesador como en clusters de PCs. Para más información sobre Root consultar [Bru-96].

Además de estos dos tipos de aplicaciones, cada uno de los grupos que colaboran en el proyecto LHC, suele desarrollar un conjunto propio de programas más específicos, orientados a resolver los problemas asociados a la investigación concreta realizada por el grupo. Por ejemplo, el Grupo Experimental de Altas Energías de la Universidad Autónoma de Madrid, que es responsable de la construcción del tapón del calorímetro electromagnético de argón líquido del detector ATLAS del LHC, ha desarrollado un software para generar los coeficientes de calibración para el análisis de la respuesta del calorímetro, así como para realizar un control exhaustivo de la calidad del mismo (más información en [Rod-03]).

2.4.1 Método de Monte-Carlo

Los programas de física de altas energías consisten, generalmente, en simulaciones y análisis basados en el *Método de Monte-Carlo*. Una introducción al método computacional de Monte-Carlo, y sus aplicaciones en física, puede encontrarse en [Keo-97]. Estos programas de simulación se caracterizan por ejecutar un elevado número de sucesos Monte-Carlo utilizando semillas diferentes. Los nodos de cálculo suelen ejecutar el mismo programa, utilizando los mismos ficheros de entrada, produciendo ficheros de salida diferentes, y sin necesidad de interactuar entre ellos. Es decir, el tiempo de cálculo es mucho mayor que el tiempo de comunicación. La comunicación entre los nodos de cálculo es prácticamente inexistente, limitándose las comunicaciones dentro del cluster a las realizadas entre nodos de cálculo y servidor de almacenamiento.

Bajo estas circunstancias, resulta más fácil hacer una estimación del rendimiento del cluster, teniendo en cuenta el tamaño de los ficheros a transmitir, la tecnología de red utilizada, y el número de nodos de que se compone el cluster.

2.4.2 Retos de Computación

Para garantizar la viabilidad de las tecnologías elegidas, y su aplicación a los requerimientos futuros del acelerador LHC, LCG ha propuesto un conjunto de retos de computación [DC1-03] o DC (*Data Challenges*). Estos DC ponen a prueba las capacidades de la tecnología (CPU, disco, comunicaciones, etc) de los equipos seleccionados, el modelo de software utilizado y los modelos de almacenamiento de datos. Los retos computacionales, de complejidad creciente, son llevados a cabo en entornos de computación Grid.

3 ATM vs. Ethernet en Aplicaciones HEP

En esta sección se analizan las ventajas e inconvenientes de las redes locales basadas en ATM y Ethernet. También se estudia la idoneidad de estas tecnologías para la construcción de clusters de PCs. Por último se analiza qué tecnología resulta la más adecuada para HEP, teniendo en cuenta los requerimientos de los distintos tipos de aplicaciones que pueden ejecutarse en el cluster.

3.1 ATM y Ethernet para Clusters de PCs

Una de las características más importantes de la tecnología ATM es que propone una arquitectura de red basada en conmutadores. Todos los equipos que componen la red están conectados a los conmutadores, y la comunicación entre cualquiera dos equipos, se realiza a través de éstos. Esta característica contrasta con tecnologías como Ethernet donde todos los equipos comparten el mismo medio físico para comunicarse. La ventaja es que una red basada en conmutadores puede mantener varias conexiones simultáneas entre diferentes equipos [Lin-95].

Sin embargo, la aparición en los últimos años de conmutadores basados en tecnologías Fast Ethernet y Gigabit Ethernet, que nos garantiza el ancho de banda aun cuando más de un equipo acceda simultáneamente al medio, hacen que esta característica de ATM no sea ya tan relevante.

La principal desventaja de la tecnología ATM es que al estar basada en protocolos orientados a la conexión, se incrementa la latencia asociada a toda comunicación, ya que es necesario establecer una conexión antes de iniciar la transmisión de los datos. Esta desventaja limita la utilidad de los clusters basados en ATM a aquellas aplicaciones que no requieran del paso frecuente de mensajes (por ejemplo aplicaciones que no estén basadas en las librerías como MPI o PVM) entre los nodos del cluster.

3.2 ATM y Ethernet para Aplicaciones HEP

Dada las características de las aplicaciones HEP, en cuanto a la relación entre tiempo de computación vs. tiempo de comunicación que presentan, las hacen candidatas ideales para su uso en clusters de PCs basados en ATM y emulación LANE. Sin embargo, hoy día, la mayoría de los clusters construidos se basan en tecnología Ethernet. En esta sección se analiza si las futuras aplicaciones del LHC, llegarán a colapsar los interfaces de red Ethernet, obligándonos a considerar otras alternativas, en concreto, ATM.

Nótese, que una mejora importante del rendimiento de las aplicaciones HEP, se conseguiría si en lugar de utilizar emulaciones LAN como LANE, se consiguiese que las aplicaciones hablasen directamente con la capa ATM. En [Lin-95] se analiza el impacto sobre el rendimiento de las comunicaciones ATM según el API de interconexión entre procesos utilizado, incluyendo interfaces abiertos como BSD sockets, Sun's RCP y PVM, habiéndose encontrado diferencias de rendimiento significativas según el interfaz utilizado. También se analiza un interfaz propietario sobre ATM AAL 5 directo, que da, con diferencia, los mejores resultados (un tercio de latencia, y doble de ancho de banda). Evidentemente, estos resultados dependen del tamaño de

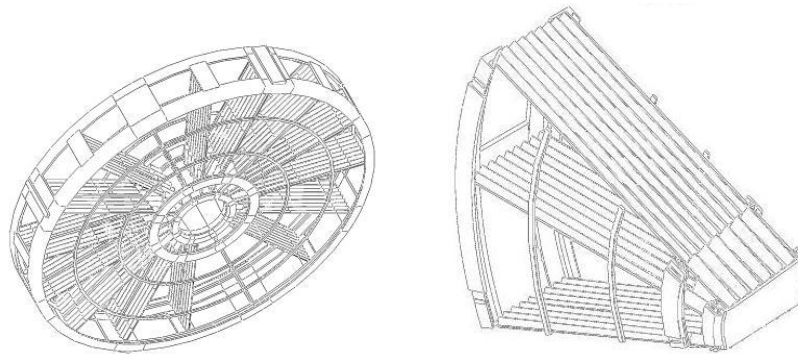


Figure 3: Tapón del Calorímetro Electromagnético

los datos a comunicar. Sin embargo, las aplicaciones utilizadas en HEP, como por ejemplo GEANT4 o ROOT, no están capacitadas para trabajar directamente con interfaces ATM, utilizando en su lugar, sockets basados en TCP/IP. Re-implementar GEANT4 o ROOT para utilizar ATM nativo es algo impensable hoy día.

3.2.1 Estudio de las Necesidades Computacionales de la Calibración del Calorímetro Electromagnético de ATLAS

El tapón del calorímetro electromagnético del detector ATLAS, está compuesto de dos ruedas divididas en 8 sectores cada una (ver Figura 3). Cada sector circular se compone de 4224 celdas detectoras de energía. Durante el proceso de calibración, se envía un haz de electrones a cada una de las celdas, y se mide su respuesta. La información sobre las respuestas de las celdas ha de ser posteriormente analizada para obtener un conjunto de coeficientes de calibración del detector.

El análisis de la respuesta de las celdas se realiza mediante la ejecución de un programa especialmente diseñado para tal fin. Este programa ha de ejecutarse 4 veces consecutivas, pero utilizando en cada ejecución parámetros de entrada distintos, para cada celda. Esto hace un total de:

$$16\text{sectores} * 4224\text{celdas}/\text{sector} * 4\text{ejecuciones}/\text{celda} = 270336\text{ejecuciones}$$

Dependiendo de los parámetros de entrada, el tiempo de ejecución puede variar ligeramente. Como promedio, una ejecución del programa suele durar aproximadamente 10 minutos, en un ordenador Pentium IV a 2GHz. Utilizando un único equipo, la calibración total de los dos tapones del calorímetro nos llevaría 1408 días de cálculo.

Dada la naturaleza del proceso de calibración, donde cada ejecución del programa es independiente de las restantes, lo hace un candidato ideal para su paralelización. A continuación vamos a estudiar las necesidades de tiempo de cálculo y de comunicaciones de entrada/salida, para ver cual es el máximo número de equipos que se pueden utilizar, dependiendo de el interfaz de red utilizado.

Cada ejecución del programa requiere de 5 ficheros de entrada, que en total suman aproximadamente 1.5MBytes. Produciendo otros 4 ficheros de salida, sumando aproximadamente 1.6MBytes. Estos ficheros están almacenados en un servidor, y son montados via NFS por los nodos de cálculo. Suponiendo un servidor de ficheros con un interfaz de red Ethernet a 10Mbits/seg, o lo que es equivalente, 1.25MBytes/seg, podríamos dar cabida a un máximo ideal de 300 equipos ejecutando los análisis simultáneamente antes de agotar las posibilidades del interfaz de red Ethernet.

En este tipo de aplicaciones, la latencia del interfaz de red no es importante, al no estar basadas en el paso de mensajes. Como hemos visto, el ancho de banda, tampoco es crítico, apenas agotándose las posibilidades de los interfaces de red Ethernet. Así que el único criterio importante que puede marcar la diferencia entre una tecnología u otra, es el económico, en cuanto a precios y disponibilidad de distribuidores.

3.2.2 Requerimientos de los Retos Computacionales DC

La primera fase del DC1 se centró en la generación de los eventos de simulación de ATLAS, utilizando las utilidades Geant4 y Root. En esta primera fase (ver [DC1-03] para una descripción completa de los resultados), participaron 40 institutos de 19 países diferentes, aportando en total 3200 equipos. En esta sección se re-analizarán los requerimientos computacionales de DC1, pero suponiendo que estos 3200 equipos corresponden a un único cluster situado en un único hipotético centro de investigación. Este tamaño de cluster se aproxima bastante a la realidad de los requerimientos del futuro LHC, en tanto en cuanto, para el año 2007 se espera que los centros regionales proporcionen al menos el doble de esta cantidad de equipos.

Durante la fase 1 del DC1 se requirió la generación de un gran número de sucesos monte-carlo que sirvieran como base para la generación de estadísticas de respuesta del detector. En concreto, se generaron 41 millones de sucesos. Estos sucesos requirieron de 44000 NCU-días de cálculo (NCU significa *Unidades Normalizadas CERN*, y equivalen a un Pentium III a 500MHz), y 15 TBytes de almacenamiento en disco.

Suponiendo que un equipo ASUS Terminator (Pentium IV a 2GHz) realizase el trabajo, se tardarían aproximadamente 28160 días. Por tanto, una

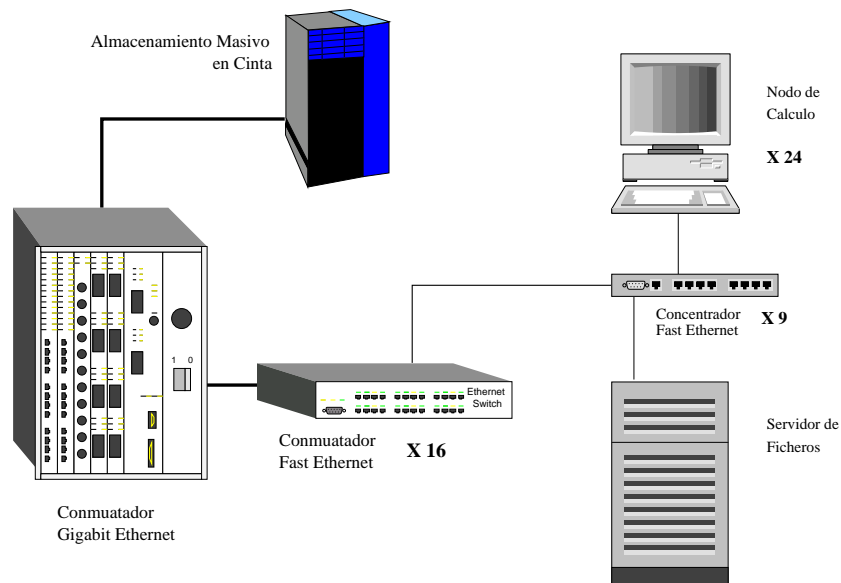


Figure 4: Propuesta de Cluster Ethernet para DC1

granja compuesta por 3200 ASUS Terminator, en condiciones óptimas de rendimiento, completaría el trabajo en algo menos de 9 días.

Al tratarse de sucesos monte-carlo independientes, podemos dividir el trabajo entre los distintos equipos según mejor nos convenga. Para este estudio, haremos una división en grupos de 60 sucesos, que son enviados a cada equipo, y que producen al cabo de 1 hora de cálculo un fichero de salida de 22 MBytes. Suponiendo interfaces de red Ethernet, se tardaría 17.6 segundos para transferir este fichero. Esto supone que podríamos conectar bajo este esquema, en una situación ideal 204 equipos antes de colapsar el medio.

Para solucionar la totalidad del problema DC1 se podría optar por utilizar una cascada de concentradores Ethernet, interconectados con conmutadores Fast Ethernet, que eviten el colapso del medio, y que a su vez estarían conectados a un conmutador Gigabit Ethernet que recopilase todos los datos y los guardase en un servidor de almacenamiento masivo. Por ejemplo, en la Figura 4 se muestra un ejemplo de topología de cluster, basado en Ethernet, que podría satisfacer las necesidades de cálculo de los CE de LCG.

3.3 Estimación de Costes

Según hemos visto en las secciones anteriores, los clusters basados en tecnología Ethernet son suficientes para satisfacer las necesidades computa-

Concepto	Prec. Unidad	Unidades	Total
Concentrador	1290	4	5160
PC	616	100	61600
Total:			66760

Table 1: Cluster basado en Ethernet

cionales del futuro LHC. Siendo difícil que las aplicaciones llegen a colapsar las comunicaciones en un cluster basado en una combinación de concentradores FastEthernet conectados en cascada, junto con un concentrador GigabitEthernet.

El último criterio que nos quedaría, por tanto, por analizar son los costes económicos asociados a ambas tecnologías.

Para hacer una estimación de la variación de los costes de un cluster si lo basamos en tecnología ATM o en tecnología Ethernet, nos basaremos en el estudio de un caso concreto: estimaremos el coste aproximado de un cluster compuesto por 100 nodos de trabajo. Por coste aproximado nos referimos a que obviaremos muchos de los costes que supone la construcción de un cluster real, como son los coste de cableado, instalación eléctrica, mano de obra, etc.

La solución de cluster basada en Fast Ethernet incluiría los siguientes elementos:

- **Nodos:** ASUS Terminator, sistema básico (*barebone*), con procesador y memoria adquirida por separado (nótese que estos equipos incluyen un interfaz de red Fast Ethernet integrado en placa madre).
- **Conmutadores:** 3Com Superstack 3 Switch 3300TM de 24 puertos Fast Ethernet + 1 puerto Gigabit Ethernet, junto con un puerto propietario Matrix - 3Com para la interconexión entre conmutadores.

Siendo el coste total de esta solución 66760 euros (ver Tabla 1 para un desglose detallado de costes).

La solución de cluster basada en ATM incluiría los siguientes elementos:

- **Nodos:** ASUS Terminator, sistema básico (*barebone*), con procesador y memoria adquirida por separado.
- **Catalyst:** Catalyst MSR 8540 de CISCO.
- **Módulos ATM:** Módulo 8540 de 16 puertos ATM OC-3.

Concepto	Prec. Unidad	Unidades	Total
Catalyst	28000	1	28000
Módulo	12800	6	76800
PC	616	100	61600
NIC	160	100	16000
Total			182400

Table 2: Cluster basado en ATM

- **NIC:** 3Com ATM Link PCI.

Siendo el coste total de esta solución 182400 euros (ver Tabla 2 para un desglose detallado de costes).

Según se puede observar, la solución basada en ATM triplica en precio a la solución basada en Ethernet. Nótese que ambas soluciones no son del todo iguales. Por ejemplo, el conmutador ATM de Cisco es capaz de mantener una transferencia sostenida de 40Gbit/seg, mientras que los conmutadores FastEthernet de 3Com conectados en cascada, sólo pueden llegar a los 5Gbit/seg. Sin embargo, desde el punto de vista de las aplicaciones HEP, ambos clusters nos proporcionan el ancho de banda necesario para no colapsar el cluster, y por tanto, podemos considerarlas soluciones equivalentes.

3.4 ATM bajo Linux

Existen un grupo de trabajo, el llamado *ATM on Linux*, alojado en el servidor de proyectos de software libre SourceForge [SF-ATM], que tiene como misión proporcionar soporte para la tecnología ATM bajo Linux. Este grupo proporciona software que nos permiten conexiones directas de las aplicaciones a la pila de protocolos ATM, el envío de paquetes IP bajo ATM, la emulación LANE y MPOA de redes LAN, además de otras tecnologías basadas en ATM. Sin embargo, el software hasta hoy liberado por este grupo, se encuentra en versión pre-alfa, es decir, aun muy inmaduro para su uso en entornos de producción.

4 Conclusiones

Hoy en día, en el mundo del PC, se puede decir que la tecnología Ethernet es prácticamente gratuita, ya que son muchas las placas madre que traen integrados uno o dos puertos Fast Ethernet. Además, la existencia de los

conmutadores Ethernet nos garantizan rendimientos similares a la tecnología ATM.

Basándonos en los criterios económicos, los conmutadores FastEthernet, siempre que sean de menos de 24 puertos, suelen encontrarse a un precio asequible. Además, no resulta difícil interconectar varios conmutadores de 24 puertos entre sí, aumentando así el máximo número de nodos que pueden conectarse a un mismo conmutador. En cambio, en el caso de tecnología ATM, resulta difícil encontrar un interfaz red a más de 155 Mbps a un precio razonable, y los correspondientes conmutadores ATM suelen alcanzar precios prohibitivos.

Hace una década existía un gran interés y expectación en la comunidad científica por los clusters de PCs basados en tecnología ATM. Sin embargo, hoy en día, y debido a la relación rendimiento / precio, y a la amplia disponibilidad de soluciones basadas en la tecnología Ethernet, prácticamente no existen implementaciones de clusters basadas en ATM.

Por último, el soporte de las tecnologías ATM bajo Linux es más bien escaso, no siendo recomendable el uso del software disponible en entornos de producción.

Sin embargo, y teniendo en cuenta los giros de 180 grados a los que nos tienen acostumbrados las tecnologías en informática y en comunicaciones, no conviene dar por totalmente descartada la tecnología ATM en su aplicación en redes LAN, y en particular, a clusters de PCs. Por tanto, se recomienda seguir de cerca los futuros desarrollos en ATM, y las publicaciones del ATM-Forum.

References

- [Eva-96] Lyndon R. Evans. The Large Hadron Collider Project, *CERN-LHC-Project-Report-53*, CERN, Septiembre 1996.
- [Fos-98] Ian Foster (ed.). *The Grid: Blueprint for a New Computing Infrastructure*, Morgan-Kaufmann, Julio 1998.
- [Gal-02] Fabrizio Galiardi, Bob Jones, Mario Reale y Stephen Bruke, European DataGrid Project: Experiences of deploying a large scale Testbed for e-Science applications, en Performance 2002 Tutorial Lectures Book "*Performance Evaluations of Complex Systems: Techniques and Tools*" por Maria Carla Calzarossa, Salvatore Tucci (eds.).

- [EDG-03] EDG Users Guide, *DataGrid-96-TED-0109-2-3*, Octubre 2003.
- [Gar-03] Rafael García Leiva y Jose del Peso. Analysis and Evaluation of OpenSource Solutions for the Installation and Management of Clusters of PCs under Linux, *ATL-SOFT-2003-001*, Enero 2003.
- [Met-76] R. M. Metcalfe y D. R. Boggs. Ethernet: Distributed packet switching for local computer networks. *Commun. ACM* 19, 7 (July 1976), 395-404.
- [Day-83] Day, J. D., and Zimmermann, H., The OSI Reference Model, Proc. of the IEEE, vol 71, pp. 1334-1340, Dic. 1983.
- [ATM-Forum] ATM Forum: <http://www.atmforum.org>.
- [New-94] P. Newman: ATM Local Area Networks, *IEEE Commun. Magazine*, vol. 32, pp. 86-98, Marzo 1994.
- [Thu-95] H. L. Thuong, et al. LAN Emulation on an ATM Network, *IEEE Commun. Magazine*, vol. 33, pp. 70-85, Mayo 1995.
- [LANE] LANE: LAN Emulation over ATM, *ATM-Forum af-lane-0021.000*, Enero 1995.
- [Lin-95] M. Lin et al. Distributed Network Computing over Local ATM Networks. En *IEEE Journal on Selected Areas in Communications*, 13(4), 1995.
- [Ago-03] S. Agostinelli et al. Geant4 - A Simulation Toolkit. *Nuclear Instruments and Methods*, A 506 (2003) 250-303.
- [Bru-96] Rene Brun Y Fons Rademakers. ROOT - An Object Oriented Data Analysis Framework, *AIHENP conference*, Lausanne 1996.
- [Rod-03] Stephane Rodier. The ATLAS Liquid Argon Electromagnetic EndCap Calorimeter: Construction and Tests. *Tesis Doctoral Univ. Autónoma de Madrid*. Diciembre 2003.
- [Keo-97] P. Kevin MacKeown. *Stochastic Simulation in Physics*. Springer Verlag, Diciembre 1997.

- [DC1-03] ATLAS Data Challenges, *DC1 Project Report*, CERN, Septiembre 2003.
- [Cav-94] J. D. Cavanaugh y T. J. Salo. Internetworking with ATM WANs. En *Advances in Local and Metropolitan Area Networks*. William Stallings, IEEE Computer Society Press, 1994.
- [SF-ATM] Source Forge: <http://linux-atm.sf.net>