Índice

Índice de Figuras....1
Myrinet....2
Antecedentes....2
Características Principales de Myrinet....3
Formato de paquetes Mytinet....8
Autodetección....8
Protocolos de ruteo....9
Ruteo Wormhole....9
Protocolo Up/Down....9
Ruteo Fuente (source)....11
Paquete ITB....11
Middleware 12
Señales de control de Myrinet....14
Productos Myrinet....15
Myrinet 2000....15
MAC: Tarjetas de Red....15
Switches....18
Myri-10G....23
MAC....24
Importancia del protocolo iSCSI....29

Índice de Figuras

Figura 1: Ejemplo real de Myrinet en el supercomputador SUPERNET....2
Figura 2: Tabla resumen sobre Myrinet....3
Figura 3: Ancho de banda respecto al tamaño del paquete. Escala logarítmica....4
Figura 4: Latencia (microsegundos) respecto al tamaño del paquete. Escala logarítmica. 4
Figura 5: Probabilidad de interbloqueo para redes de 8 switchs....5
Figura 6: Probabilidad de detección incorrecta de la topología para una red con 16 switchs....5
Figura 7: Probabilidad de interbloqueo para redes de 16 switchs....6
Figura 8: Probabilidad de detección incorrecta de la topología para una red con 32 switchs....6
Figura 9: Probabilidad de interbloqueo para redes de 32 switchs....7
Figura 10: Ejemplo de ciclo en la red....9
Figura 11: Diagrama de la red a explorar....10
Figura 12: Grafo UP*/Down* para la red....10
Figura 13: Paquete original de Myrinet....11
Figura 14: Paquete con la inclusión de ITB....12
Figura 15: Ejemplo de uso de ITB....12
Figura 16: Soporte software de Myrinet....13
Figura 17: Latencia de mensajes sobre distintas librerías....13
Figura 18: Ancho de banda para mensajes enviados a través de distintas librerías....14

Figura 19: Señales de control de la interfase MAC.....14
Figura 20: Esquema básico de los componentes de una MAC.....15

Myrinet

Antecedentes

Myrinet es una red de interconexión de clusters de altas prestaciones. Sus productos han sido desarrollados por Myricom www.myri.com desde 1994, y desde entonces han ido mejorando en rendimiento. Su uso se extiende a 50 países y ha formado parte de las máquinas de más alto rendimiento publicadas en el TOP500 (www.top500.org). Myricom también proporciona soporte software y hardware a compañías como IBM, HP, Dell, Sun y muchos otros fabricantes.

En los últimos productos desarrollados se busca la compatibilidad con otros estándares (Ethernet).

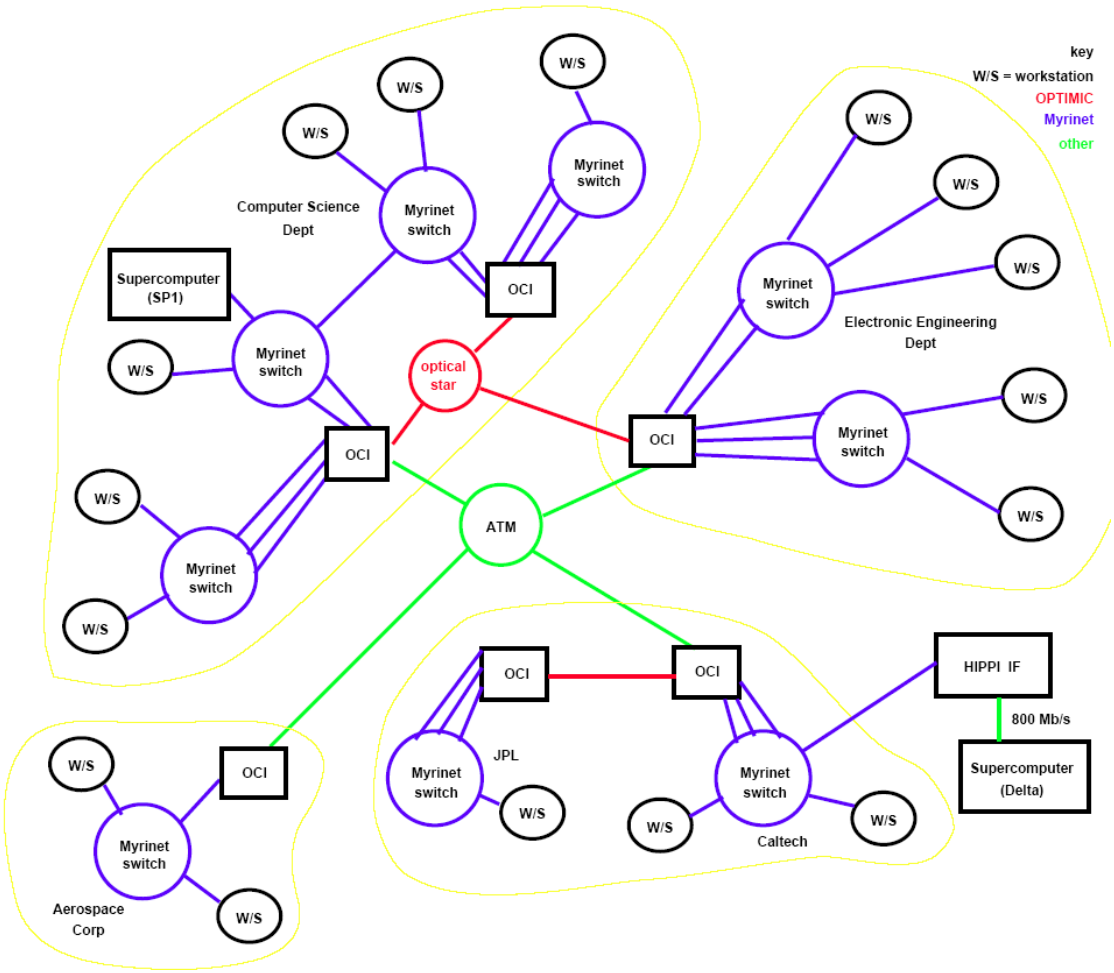


Figura 1: Ejemplo real de Myrinet en el supercomputador SUPERNET

Normalmente los supercomputadores tienen asociados varios tipos de redes de interconexión y no en exclusividad un tipo determinado, es el caso de Supernet, Mare Nostrum y otras.

Características Principales de Myrinet

Baja Latencia.
Escalabilidad-Autodetección de Topología.
Monitorización de cada enlace.
Software Libre.
Tratamiento Interbloqueo (Deadlock)

Dispositivos/Características	Myrinet-2000	Myri-10G
Envío Full-duplex de las MAC y enlaces de la red	2+2 Gigabits/s	10+10 Gigabits/s
Enlaces de los cables	Conector LC duplex para fibras de hasta 200m	Selected 10-Gigabit Ethernet cables, copper and fiber
NIC slot	Simple y doble puerto de la PCI-X	Puerto simple PCI-Express, soportadando los protocolos 10G Myrinet y 10G Ethernet
Switches	Basados en 16 y 32 puertos.	Basados en 16 puertos.
Switch networks	Se tienen hasta 256 puertos como máximo en un único armario de la red. Se llega hasta diez mil combinando componentes.	Se tienen hasta 128 puertos como máximo en un único armario de la red. Se llega hasta diez mil combinando componentes.
Opera con otros	Gigabit Ethernet	10-Gigabit Ethernet
Soporte software	Myrinet Express (MX-2G) o GM-2	Myrinet Express (MX-10G)
Latencia MX o MPI	2.6µs–3.2µs	2µs
Ratio de envío unidireccional con MX	247 MBytes/s (MAC con un puerto) 495 MBytes/s (MAC de 2 puertos)	1.2 GBytes/s
Ratio de envío TCP/IP (Emulando Ethernet con MX)	1.98 Gbits/s (MAC con un puerto) 3.95 Gbits/s (MAC de 2 puertos)	9.6 Gbits/s

Figura 2: Tabla resumen sobre Myrinet.

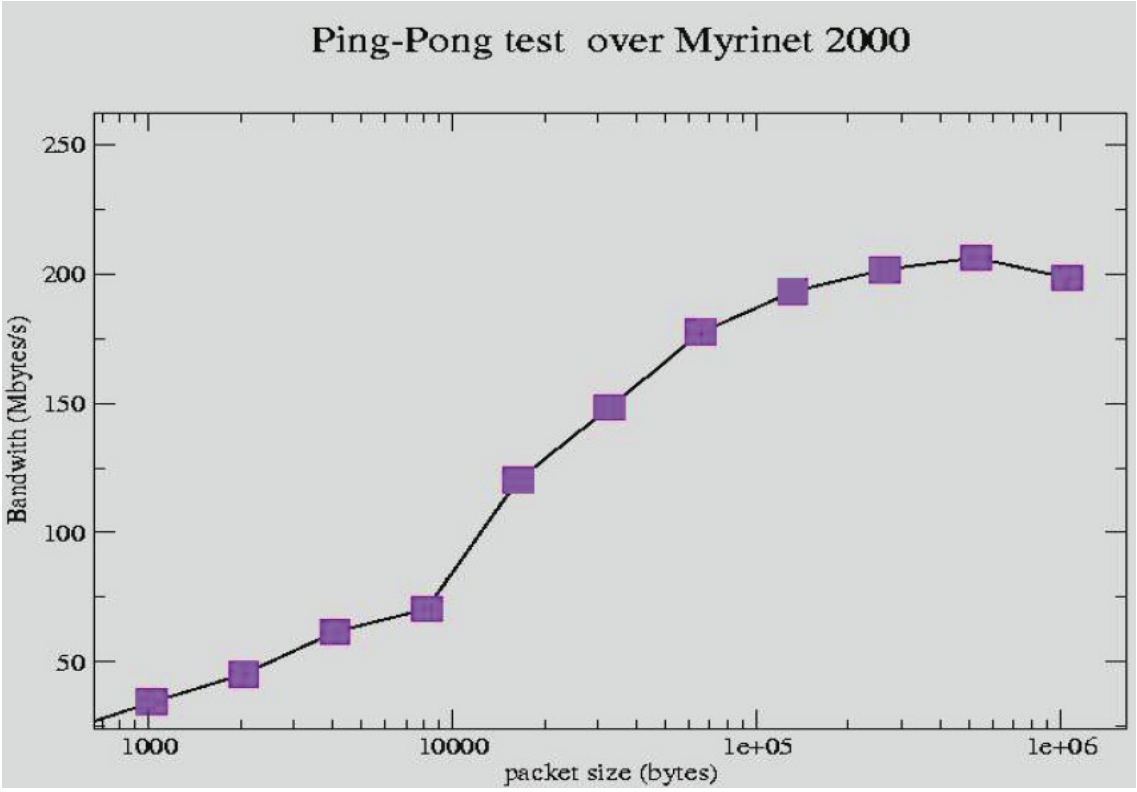


Figura 3: Ancho de banda respecto al tamaño del paquete. Escala logarítmica.

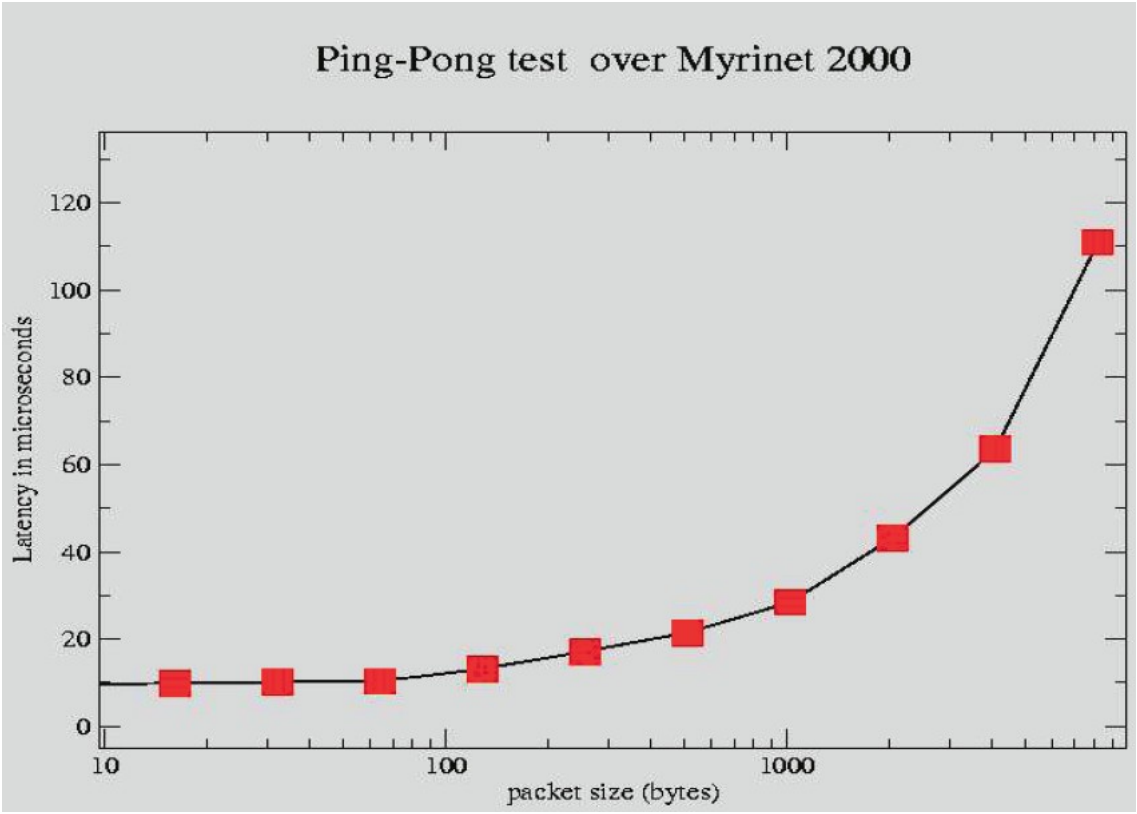


Figura 4: Latencia (microsegundos) respecto al tamaño del paquete. Escala logarítmica.

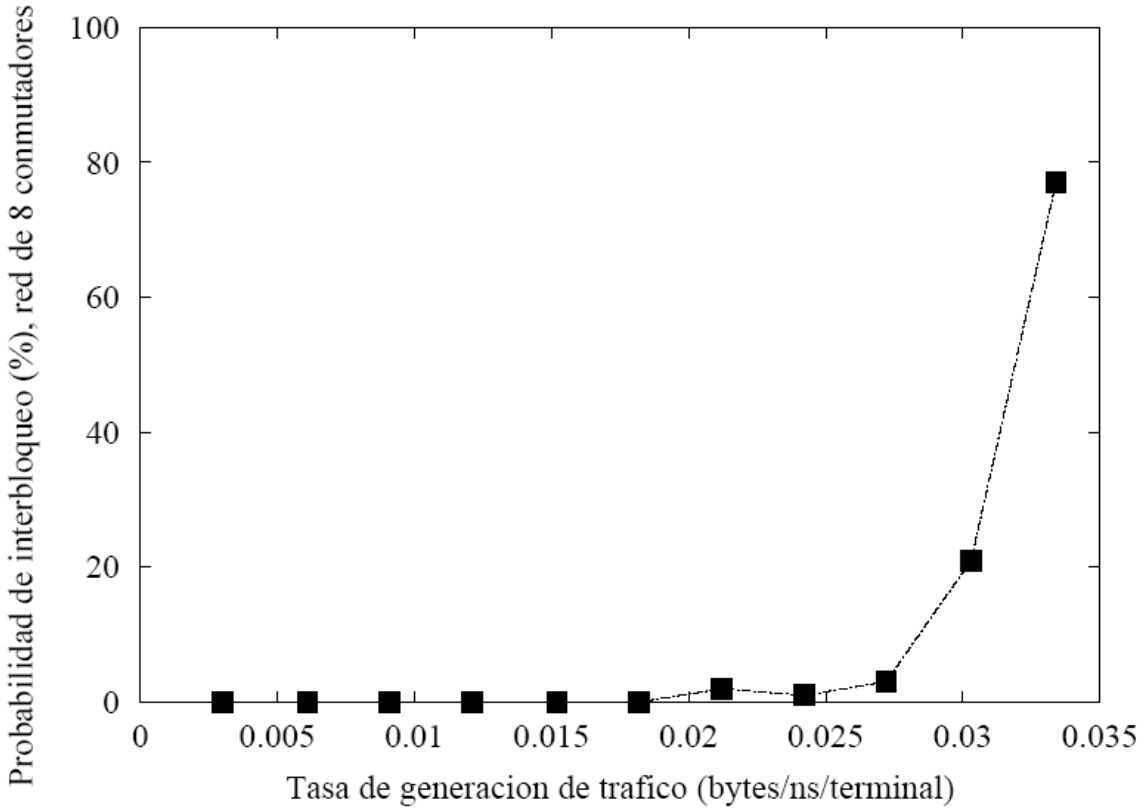


Figura 5: Probabilidad de interbloqueo para redes de 8 switchs.

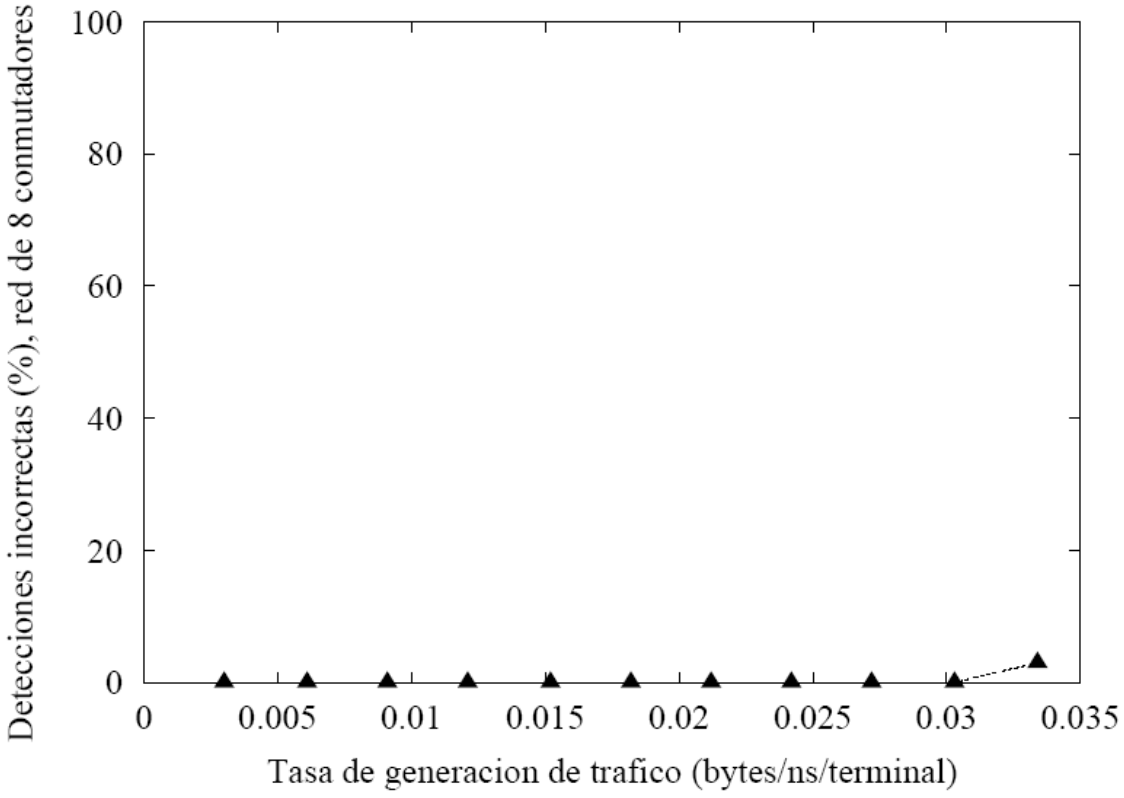


Figura 6: Probabilidad de detección incorrecta de la topología para una red con 16 switchs.

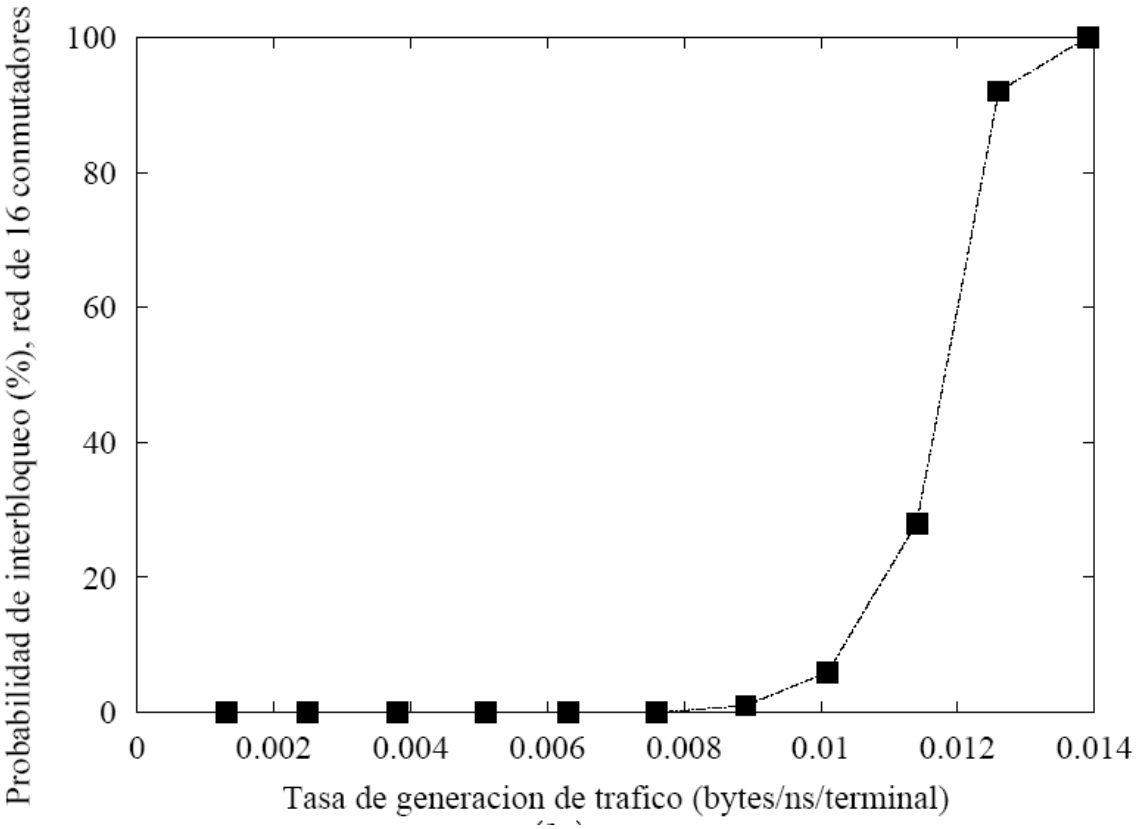


Figura 7: Probabilidad de interbloqueo para redes de 16 switchs.

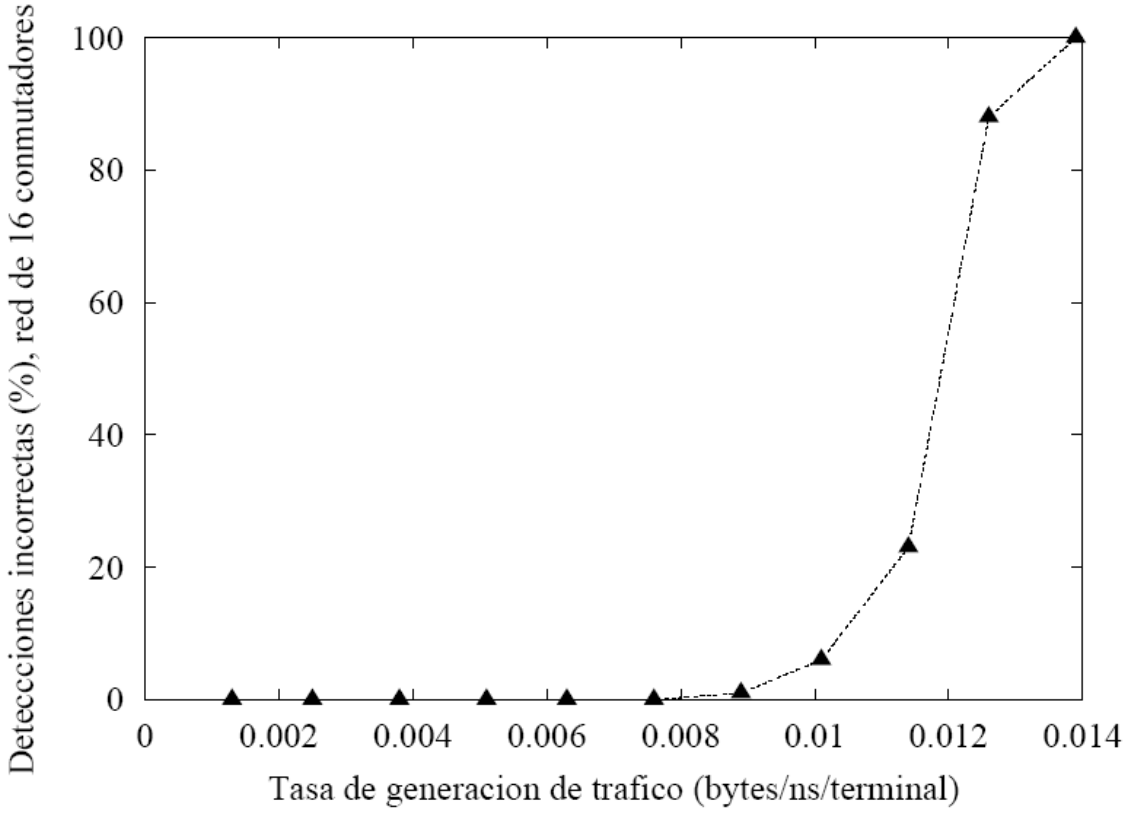


Figura 8: Probabilidad de detección incorrecta de la topología para una red con 32 switchs.

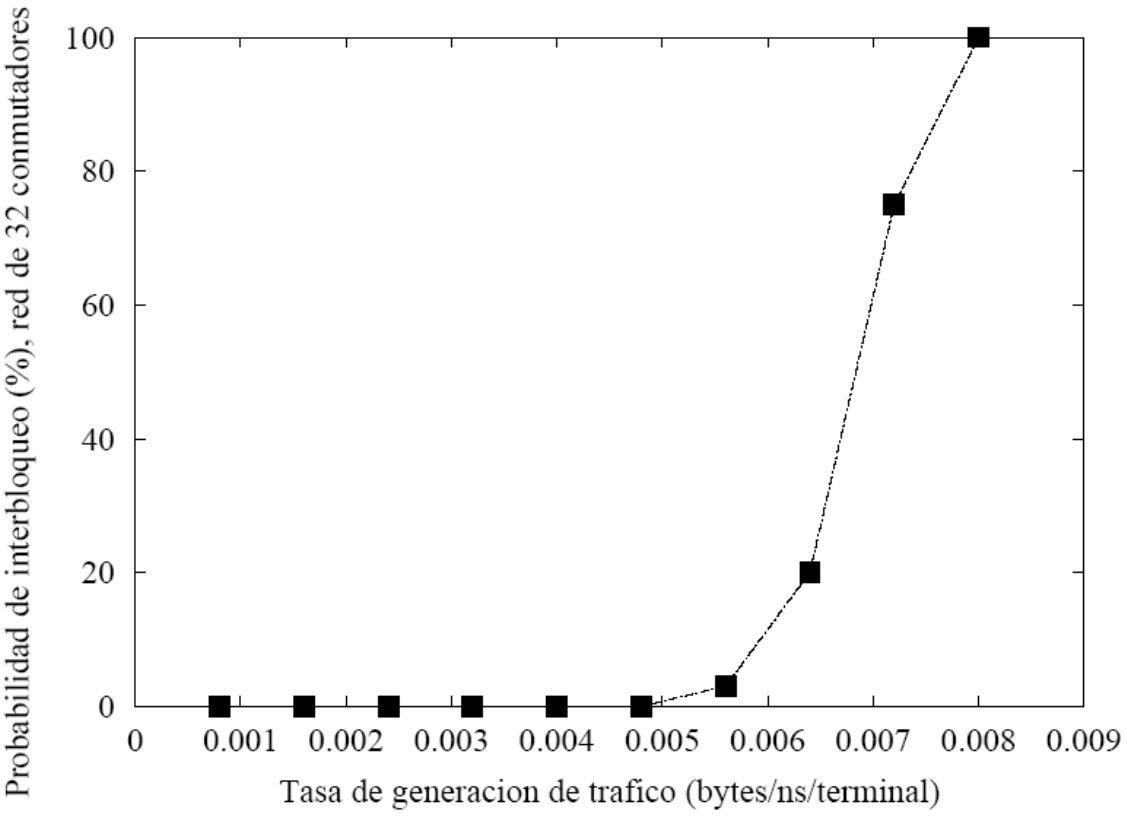
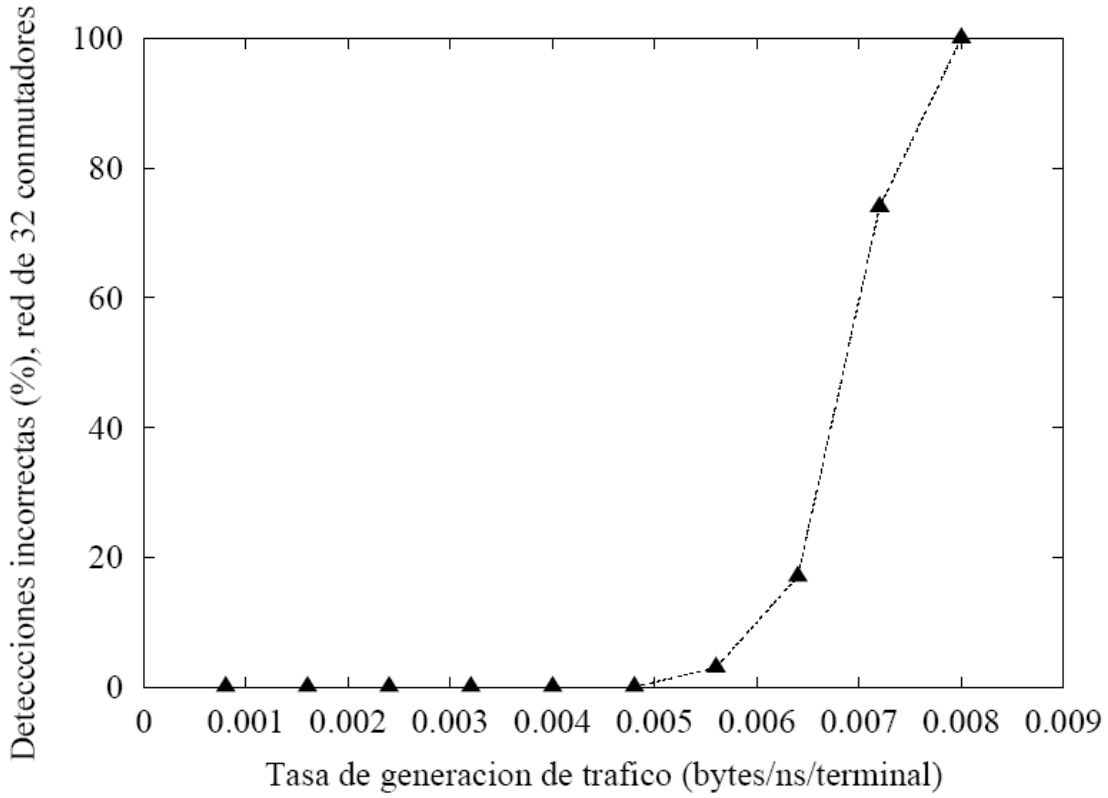
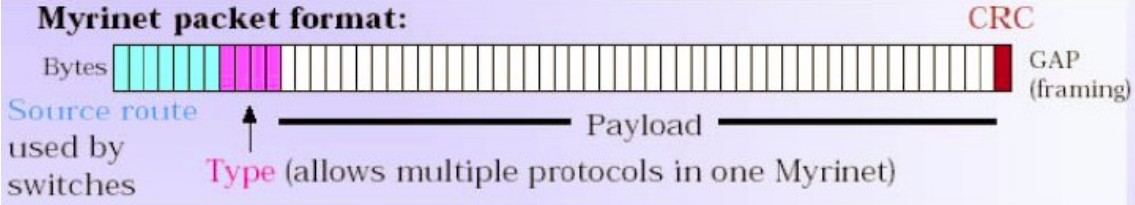


Figura 9: Probabilidad de interbloqueo para redes de 32 switchs.



Formato de paquetes Mytinet



Bits iniciales indica la dirección de la fuente, a continuación se especifica el tipo del protocolo (de los que permite Myrinet), la carga útil del paquete y un bit para el CRC.

Autodetección

El mecanismo de exploración implementado en Myrinet se realiza desde un único terminal de la red, denominado “mapper" ya que su tarea consiste en elaborar un “mapa" de la misma. Para adaptarse a posibles cambios en la red, las exploraciones deben repetirse periódicamente. De este modo los cambios pueden detectarse comparando el “viejo" mapa con el recién elaborado. Es por ello que en cada exploración el nuevo mapa se construye desde cero, sin tener en cuenta los mapas anteriores.

Básicamente, la elaboración de un mapa consiste en detectar qué componente (si es que hay alguno) se encuentra conectado a cada uno de los puertos de la red. Para llevar a cabo esta detección, el mapper envía mensajes de exploración por cada puerto y procesa las posibles réplicas a los mismos.

Para explorar un puerto, el mapper envía a dicho puerto, en primer lugar, un mensaje de detección de terminal. Si el componente “desconocido" es realmente un terminal, este enviará una réplica al mapper. Los mensajes de detección de terminal incluyen la ruta que debe seguir la réplica para llegar al mapper, por lo que el terminal detectado no necesita consultar su tabla de encaminamiento. En la réplica, el terminal detectado incluye su identificador único. Al recibir una replica a un mensaje de detección de terminal, el mapper añadirá el terminal detectado al mapa, donde el identificador único indicado permitirá distinguirlo de otros terminales. El mapper esperará la llegada de la réplica a cada mensaje durante cierto tiempo, pasado el cual reenviará el mensaje. Mientras no reciba réplica, el mapper reenviará cada mensaje hasta dos veces. Si tras esto sigue sin recibirse réplica, el mapper asumirá que el componente “desconocido" no es un terminal, y enviará al puerto un mensaje de detección de conmutador.

Los mensajes de detección de conmutador son distintos a los usados para detectar terminales debido a que un conmutador, en una red con las características de Myrinet, no puede generar mensajes y por tanto no puede enviar réplicas al mapper. Para solucionar esto, el mapper asigna a los mensajes de detección de conmutador una ruta de ida y vuelta.

Siguiendo esta ruta, el mensaje sale del mapper, llega hasta el puerto a explorar y, sólo si existe un conmutador conectado a dicho puerto, el mensaje entrará al conmutador, saldrá de él por el mismo puerto de entrada, y regresará al mapper recorriendo en dirección opuesta el camino de llegada. De este modo, el mismo mensaje puede ejercer de su propia réplica al seguir una ruta “circular". Si no existiera un conmutador conectado al puerto, el mensaje se descartará. Tras la recepción de un mensaje de detección de conmutador se añadirá el conmutador detectado al mapa,

donde será identificado mediante un número asignado por el mapper. Los mensajes de detección de conmutador enviados por el mapper que no “regresan" tras cierto tiempo también son reenviados hasta dos veces. Si tras esto sigue sin detectarse un conmutador, el mapper considerará que el puerto explorado está desconectado, y así lo reflejará en el mapa.

En definitiva, si existe un componente conectado a un determinado puerto se detectará por la recepción en el mapper de algún tipo de réplica, y la ausencia de las mismas permitirá considerar al puerto desconectado. En cualquier caso, una vez determinado el estado de la conexión, se dejará de explorar el puerto.

Téngase en cuenta que cada nuevo mapa se elabora sin considerar los anteriores y, por tanto, al principio del proceso sólo existe un puerto por explorar: el propio puerto del mapper. Al añadir cualquier conmutador al mapa, se añaden sus puertos a una lista de puertos por explorar1. Los puertos de un mismo conmutador son explorados simultáneamente. Una vez explorados todos los puertos de un conmutador, se pasa a explorar los puertos de otro. La exploración acaba cuando no existen en el mapa puertos conectados a componentes desconocidos.

Protocolos de ruteo

Ruteo Wormhole

En este protocolo la unidad de transferencia se denomina “*worm*”, que puede variar desde unos pocos bytes hasta varios cientos de bytes. En Myrinet el tamaño máximo del ‘*worm*’ es de 9 KB (bytes), este límite es impuesto por el procesador de la tarjeta de red (LANai, ver en la sección de Productos de Myrinet). Cada switch intermedio envía el ‘*worm*’ por el puerto disponible deseado tan pronto como la cabeza del fragmento es recibido, sin esperar a que el ‘worm’ sea completamente ensamblado. Lo que puede hacer que un mismo fragmento se encuentre siendo enviado por varios puntos de la red a la vez. Cuando el puerto de recepción no es alcanzable se bloquea el envío y es retornado por donde vino. Este tipo de ruteo permite a Myrinet una baja latencia en los envíos (en el mejor de los casos la propagación entre los extremos es justo el retardo en la propagación más la latencia mínima a nivel de ruter) y un almacenamiento menor en los puntos de conexión.

No se garantiza la ausencia de interbloqueos, pueden ocurrir cuando existe un ciclo entre los nodos. Para garantizar que esto no se produzca se usa en combinación con el enrutamiento ‘*Up/Down’*

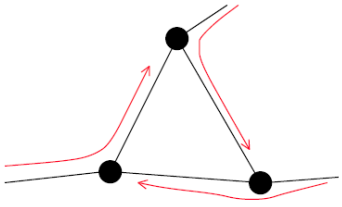


Figura 10: Ejemplo de ciclo en la red.

Protocolo Up/Down

Un nodo es escogido arbitrariamente para que sea la raíz del árbol y los enlaces a éste se considerarán como ‘*up’* (arriba) o ‘*down’* (abajo) respecto a dicho nodo raíz. El

estado de un enlace (‘*up/down*’) dependerá de cómo el algoritmo de distribución recorra el árbol. Un enlace estará ‘*up’* si se pasa de un nodo a otro con un mayor nivel en el árbol, esto se traduce en que se está accediendo a un nodo más cercano al nodo raíz. Por el contrario un enlace ‘*down’* es cuando descendemos un nivel en el árbol. Es necesario que cuando se genere el árbol se considere que los nodos que se encuentren en el mismo nivel deben de romperse su conexión (no físicamente, pero inutilizarse) con el fin de que no se creen ciclos.

El envío desde el nodo fuente hacia el destino requiere hacer primero los recorridos ‘*up’* de los enlaces hacia la raíz del árbol (siempre que fuera necesario) antes de hacer cualquier movimiento ‘*down’*. Myrinet funciona con la ejecución en background de un algoritmo de mapeo de la red (cálculo de la topología).

La ventaja de este protocolo es la simplicidad de la implementación tanto software como hardware, así como un ruteo libre de bloqueos.

Las desventajas son que las trayectorias seleccionadas no son generalmente las trayectorias más cortas y que existe el riesgo de que los enlaces cercanos a la raíz se congestionen, ocasionando un bajo rendimiento del procesamiento. Se asegura la ausencia de interbloqueos sólo sobre mensajes convencionales. Ahora bien, durante las exploraciones, el mapper envía los mensajes de exploración sin las restricciones que ‘*Up*/*Down’* impone a los mensajes de usuario. Por tanto, no puede garantizar que los mensajes de exploración sigan rutas legales. En definitiva, la coexistencia de mensajes que siguen rutas legales y otros que pueden seguir rutas ilegales impide que durante la exploración pueda asegurarse la ausencia de interbloqueos.

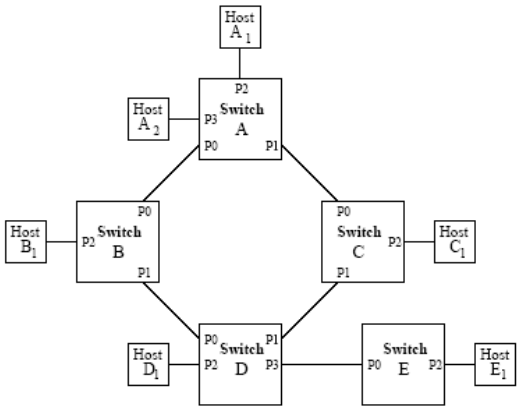


Figura 11: Diagrama de la red a explorar

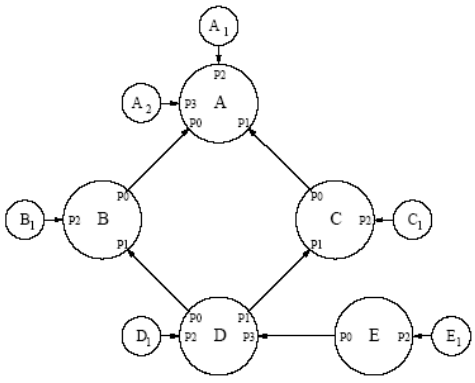


Figura 12: Grafo UP*/Down* para la red.

Ruteo Fuente (source)

Una vez se ha construido el árbol 'up/down' y se ha enviado a todos los hosts de la red, el nodo fuente que realizará un envío usa éste para calcular la ruta (un conjunto de puertos de salida de los switch) deseada hasta el host destino. Esta ruta es incluida, por emisor, en la cabecera del paquete a enviar. Así cuando se recibe el paquete en el conmutador ya se sabe por que puerto del mismo tiene que salir.

Antes de enviar el paquete el switch extrae la información relevante al puerto por donde debe hacerse el envío de la cabecera. Esto supone esperar la recepción, almacenar el total del paquete en el conmutador y realizar el cálculo del checksum (CRC). Así como un recálculo de la ruta cuando se produzca un error.

La ventaja principal de este método es que facilita los envíos broadcast simplemente añadiendo a la cabecera una representación lineal del árbol, mientras que en los anteriores protocolos hay que usar herramientas más complejas para generar este tipo de envíos.

Paquete ITB

Myricom la ha incorporado en su conocida red Myrinet mediante un tipo especial de paquete ITB (In-Transit Buffer). Técnicas dinámicas de reconfiguración de la red para redes sin pérdidas. Se pensaba que la reconfiguración dinámica de la red no era posible en una red sin pérdidas como las usadas en la mayoría de clusters de altas prestaciones. La razón es que las tablas de encaminamiento para los diferentes routers o conmutadores no pueden ser actualizadas síncronamente, y por lo tanto, las tablas de encaminamiento para las viejas y nuevas configuraciones de red pueden coexistir, generalmente conduciendo a interbloqueos. Myrinet se basa en una reconfiguración estática, llevando así a pérdidas muy significativas de prestaciones cada vez que hay un cambio en la topología. Se demuestra que aunque las tablas de encaminamiento no pueden ser actualizadas asíncronamente sin introducir bloqueos, es posible realizar de manera asíncrona varias etapas de reconfiguraciones parciales de las tablas de encaminamiento de manera tal que los bloqueos puedan ser evitados. Esta investigación abrió la puerta a nuevas y más potentes estrategias de reconfiguración de la red.

Básicamente esta técnica intenta eludir ciertas restricciones impuestas por los protocolos de ruteo, almacenando en buffers algunos paquetes en determinados nodos para permitir el tránsito de otros, una vez finalicen éstos últimos se recuperan aquellos que habían sido almacenados. Un ejemplo podría ser sobre el protocolo '*Up*/*Down*', el poder permitir una transición down-up para obtener una ruta más corta.

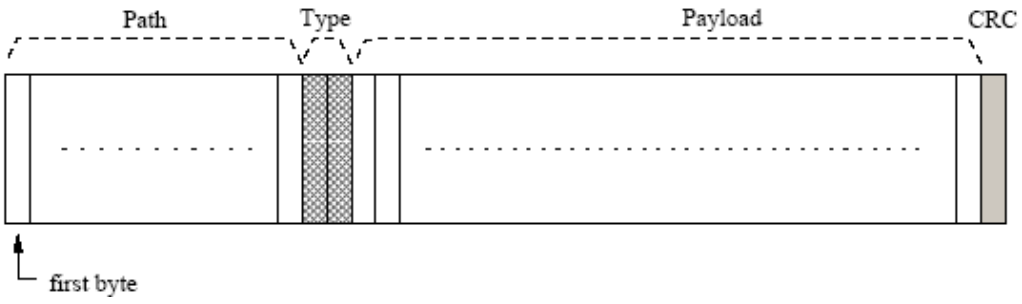


Figura 13: Paquete original de Myrinet.

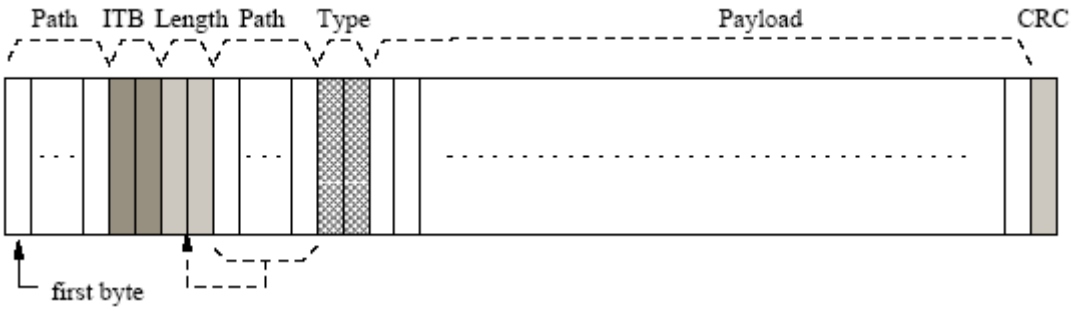


Figura 14: Paquete con la inclusión de ITB.

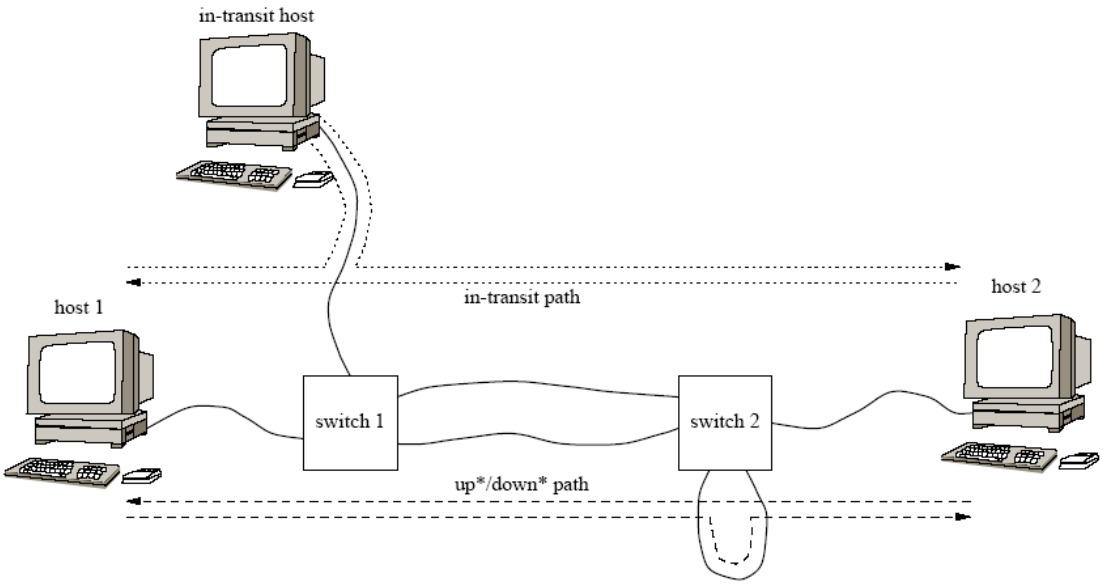


Figura 15: Ejemplo de uso de ITB.

Middleware

La infraestructura Myrinet puede ser utilizada para soportar herramientas de programación paralela tales como MPI o, en general, cualquier aplicación basada en protocolos IP ya que es posible asignar direcciones IP a una interface Myrinet. Por ejemplo, se puede soportar sistemas de archivos distribuidos como el "General Parallel File System (GPFS)" de IBM.

En cuanto al middleware de comunicación, la inmensa mayoría está desarrollado por Myricom, y distribuido bajo la fórmula de Software Libre. Destacan las librerías a bajo nivel GM y MX, las implementación de MPI, MPICH-GM y MPICH-MX y las implementaciones de Sockets de alto rendimiento Socktes-GM y Sockets-MX.

Se puede acceder desde su página web (Software & Customer Support). Las MAC requieren del uso del software GM-2 o el MX.

Firmware/ Driver/ API	Middleware					
	MPI-1	VIA	PVM	Sockets	DAPL	ClusterTools
MX (MX-2G)	MPICH-MX	n/a	n/a	Sockets-MX	En desarrollo	En desarrollo
GM	MPICH-GM	VI-GM	PVM-GM	Sockets-GM	DAPL-GM	ClusterTools

Figura 16: Soporte software de Myrinet

Además se proporcionan programas de control de Myrinet, documentación del para el programador incluyendo información sobre el procesador de la tarjeta Lanai-X (http://www.myri.com/vlsi).

Veamos unas estadísticas sobre distintas bibliotecas de paso de mensajes:

1. MpiJava (Java wrapper) sobre bibliotecas nativas (en C): MPICH, MPICH-GM y SCI-MPICH.

2. ScaMPI CCJ (Java puro).

3. JMPI (Java puro).

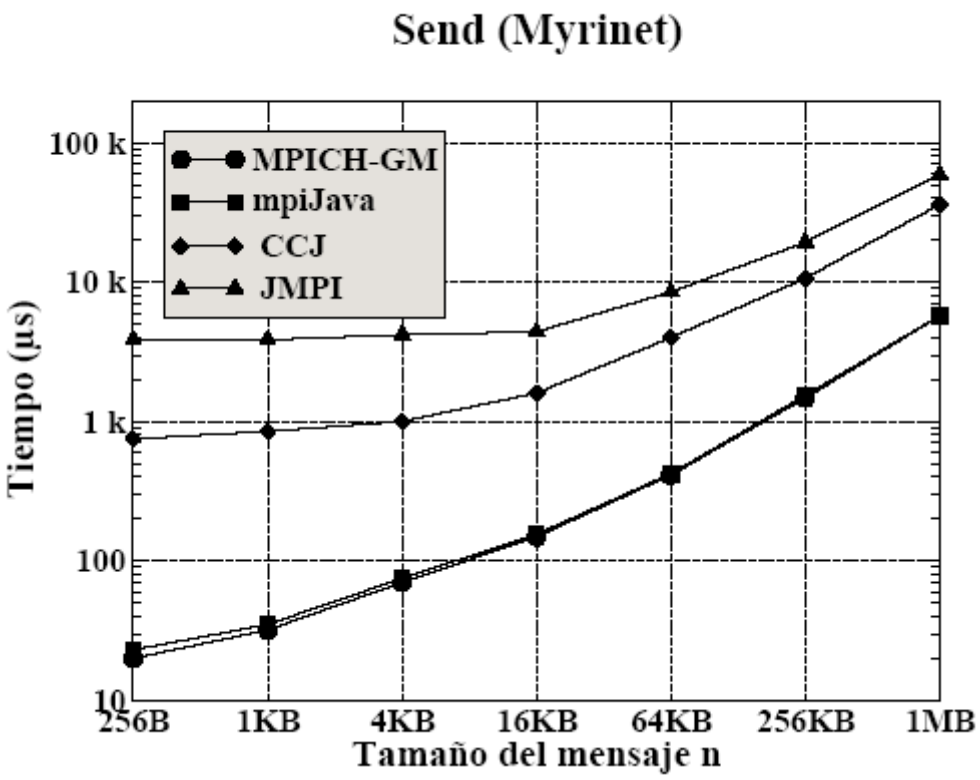


Figura 17: Latencia de mensajes sobre distintas librerías.

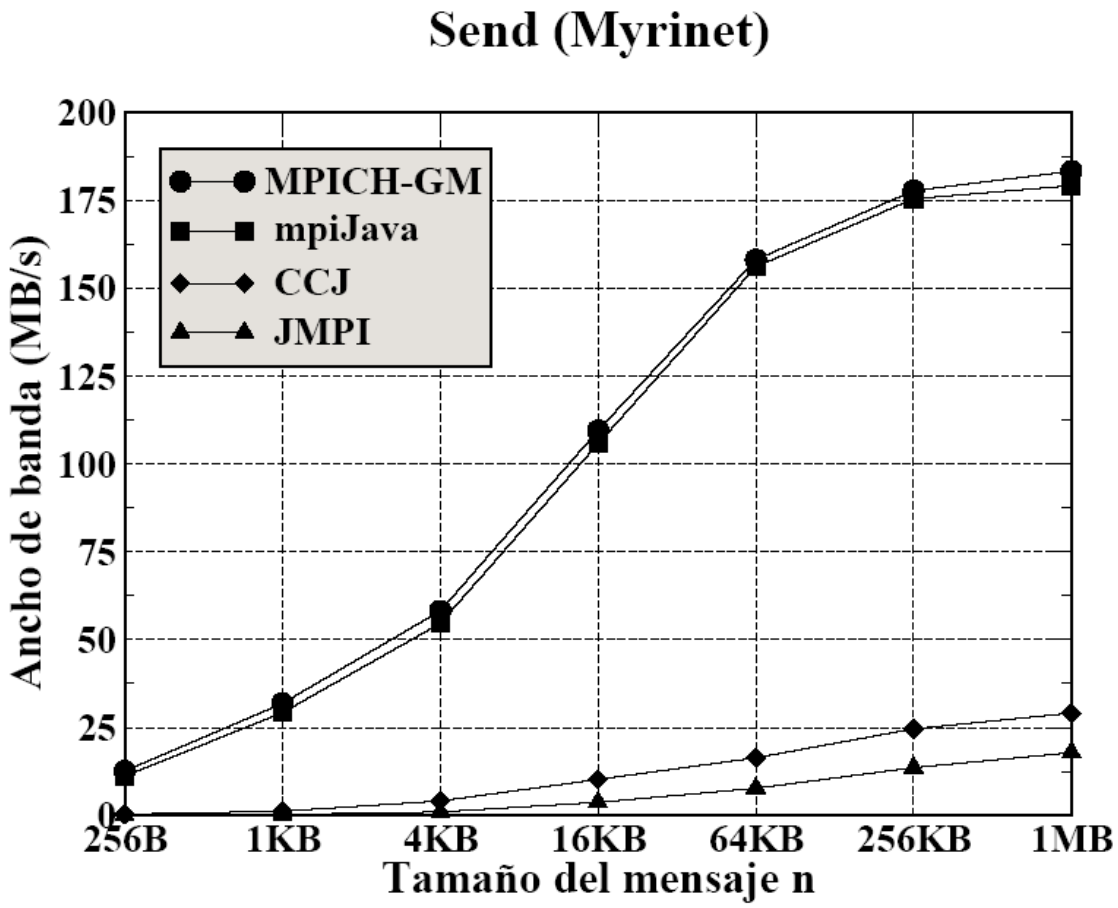


Figura 18: Ancho de banda para mensajes enviados a través de distintas librerías.

Señales de control de Myrinet

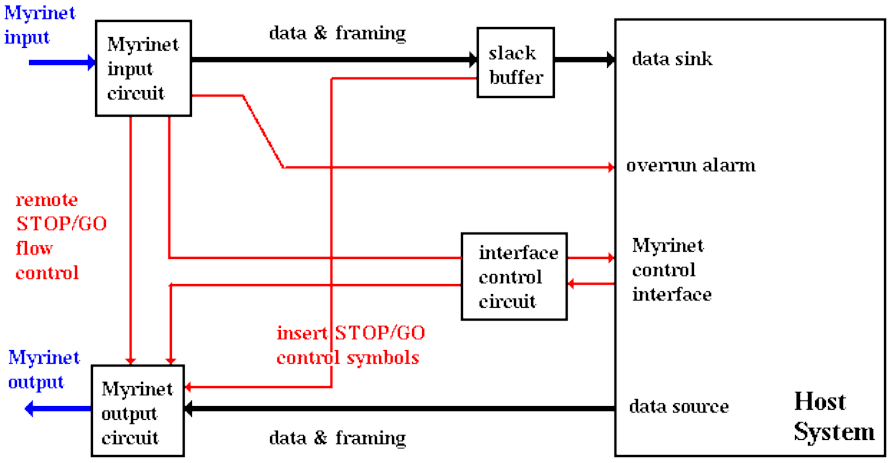


Figura 19: Señales de control de la interfase MAC

Myrinet usa un reducido número de señales de control para el envío y recpción de la información. Símbolos como STOP, GO (para manipular la FIFO del buffer), FRES (para resetear la red), etc.

Productos Myrinet

Los productos que Myrinet van desde Switches, software, cables y fibra óptica. Con respecto a la fibra óptica indicar que físicamente tiene limitaciones, como por ejemplo el ángulo de torsión, las temperaturas a las que puede funcionar, tipos de fibra (monomodo y multimodo),… hay que considerar que lo más complicado cuando se trabaja con fibra es que los enlaces tienen que estar perfectamente logrados (para que se propague correctamente la luz) y el costo que tienen los generadores de la información que viajará por el cable, es decir, cuanto mejor sea un láser más caro resultará.

Dispositivos suelen tener capacidades de tolerancia a fallos, con control de flujo, control de errores y monitorización de la red.

En cuanto al middleware de comunicación, la inmensa mayoría está desarrollado por Myricom, y distribuido bajo la fórmula de Software Libre. Destacan las librerías a bajo nivel GM y MX, las implementación de MPI MPICH-GM y MPICH-MX y las implementaciones de Sockets de alto rendimiento Socktes-GM y Sockets-MX. Los drivers y firmware de las tarjetas están disponibles para Linux, Windows, Solaris 10, Mac OS X y FreeBSD. Para apreciacines más específicas sobre la plataforma véase http://www.myri.com/scs/performance/Myri10GE/ .

Todos los productos cumplen la directiva RoHS para la restricción de ciertas sustancias peligrosas en aparatos eléctricos y electrónicos. Entró en vigor el 1 de julio de 2006, pero no es una ley, es simplemente una directiva. Esta directiva restringe el uso de seis materiales peligrosos en la fabricación de varios tipos de equipos eléctricos y electrónicos.

Myrinet 2000

MAC: Tarjetas de Red

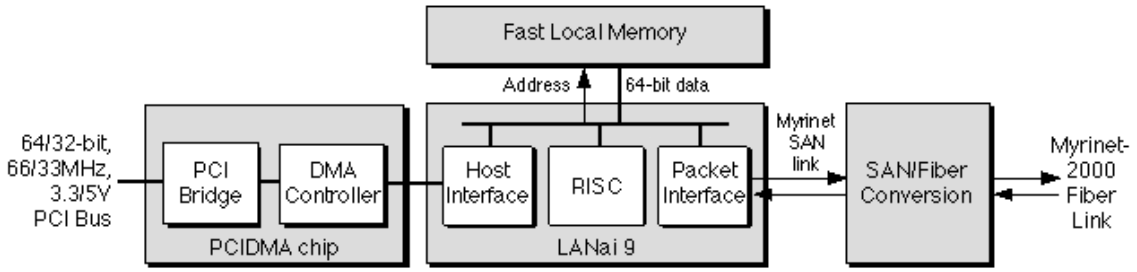


Figura 20: Esquema básico de los componentes de una MAC.

Las MAC[1] se conectan al bus del sistema para lograr mejores ratio de velocidad, evidentemente esto supone un límite a las MAC. Myrinet trabaja sobre PCI y PCI-X (express), a frecuencias entre 48MHz a 67MHz y 75MHz a 133MHz respectivamente. Por ejemplo para arquitecturas de 64 bits para 66MHz (PCI) tenemos unos ratios en torno al 533 MB/s o para 133MHz (PCI-X) unos 1067MB/s.

[1] También se suele denominar NIC o tarjeta de red.

Sin embargo también hay que considerar el modo en el que están operando estos buses y la memoria que los PC disponen.

TARJETA REDUCIDA D: M3F-PCIXD-2 Y M3F-PCIXD-4

Sus precios rondan los 419.7 € y 589.3€ según la cantidad de memoria local que se disponga (4MB). Puede llegar a lograr latencias de hasta 3.5µs sobre MPI y dispone de un puerto.

Esta tarjeta soporta los protocolos PCI y PCI-X pudiendo ser usada sobre cualquiera de estos slots, que deben de funcionar con 3.3V. Se respeta la detección de paridad del bus.

Se incluye en la MAC un procesador (225MHz), Lanai-XP o Lanai-2XP (http://www.myri.com/vlsi/ para más información), memoria local de entre 2MB a 4MB, memoria EEPROM de 512KB, en la cual se incluye datos de configuración del bus, inicialización de programas, seriales, siendo reprogramable desde el procesador. De hecho la única diferencia entre M3F-PCIXD-2 y M3F-PCIXD-4 es el procesador Lanai que disponen, siendo respectivamente Lanai-XP y Lanai-2XP.

Indispensable es el puerto serial de fibra óptica "Mirinet-2000" que es capaz de transmitir de 2 Gb/s (par un máximo de 200 m) por cada uno de los enlaces de la fibra, uno destinado al envío y el otro a la recepción. El puerto trabajará con fibra multimodo (múltiples señales luminosas a la vez por el hilo) la cual precisa de un láser de altas prestaciones (http://www.myri.com/open-specs/serial.pdf).

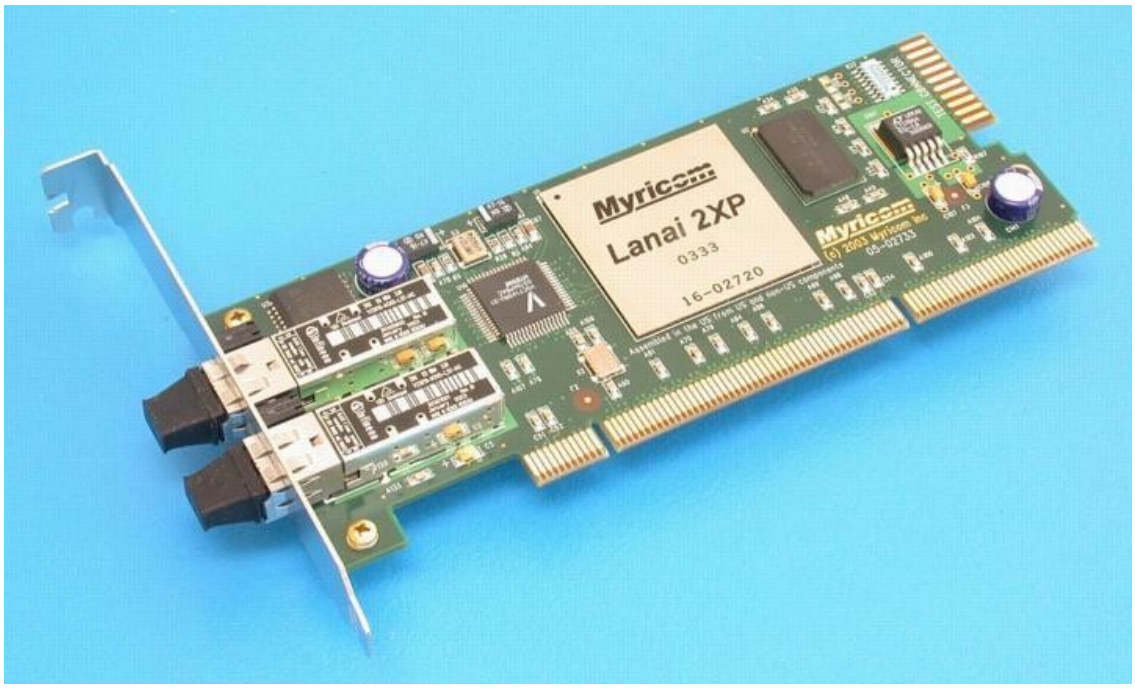
Físicamente ocupa 6.4cm de ancho, 16.4 cm de largo y 2.2 cm de grosor, con un peso de 82 g. Se tiene hasta dos tipos de frontales para su adaptación a carcasas pequeñas. Requiere una alimentación de 3.3V (que proviene del bus) con un consumo máximo de unos 6W. Soporta hasta un máximo de 55ºC.

TARJETA RÁPIDA F: M3F-PCIXF-2 Y M3F-PCIXF-4

Sus precios rondan los 504.5 € para 2MB de memoria local y 758.9 € para 4MB, pudiendo llegar a lograr latencias de hasta 2.6µs sobre MPI, con un único puerto.

El resto de características son similares a las tarjetas reducidas M3F-PCIXD-2 y M3F-PCIXD-4, exceptuando que estas tarjetas son más rápidas (operan a 333MHz).

TARJETA E: M3F2-PCIXE-2, M3F2-PCIXE-4



Sus coste aproximado es de 674.1 € (2MB) y 928.5 € (4MB) y puede lograr latencias de hasta 2.7µs sobre MPI, con dos puertos.

Estas tarjetas son capaces de operar en los límites de funcionamiento de los buses PCI (inclusive la versión 2.2) o PCI-X, pero siempre funcionando a 3.3V. Se mantiene la detección de paridad que tiene el bus. Los procesadores son equivalentes a las otras tarjetas (Lanai-XP y Lanai-2XP) operando a 333MHz.

Se incluyen dos puertos, en cada uno de ellos se establecen comunicaciones de hasta 2Gb/s (par un máximo de 200 m), lo que llega a ser un ratio de 4Gb/s, considerando que hay dos hilos para envío y otros dos para recepción. Se requiere además de lásers de calidad que operen bajo estas especificaciones.

Físicamente mantiene las mismas características que las otras MAC que distribuye esta empresa. La diferencia más notable es el aumento del consumo que requiere 9.3W para la MAC M3F2-PCIXE-2 y 8W para M3F2-PCIXE-4.

Número de hosts del cluster según las MAC

Las MAC de un solo puerto suelen usarse en clusters con más de 2048 hosts a través del driver GM-2 o bien 4096 sobre MX. Las MAC de dos puertos suelen soportar más de 1024 hosts con GM-2 o 2048 sobre MX. Como cabe la posibilidad de que los usuarios puedan crearse sus propios drivers el número de hosts pueden incrementarse.

Switches

Los switches se clasifican según el número de bocas que tienen, o que es lo mismo según el número máximo de puertos que pueden conectarse. Además de la conmutación punto a punto también son capaces de proveer otros servicios.

SWITCHES PEQUEÑOS

Todos ellos ocupan 2 unidades dentro de los armarios:

M3F-SW8: Switch de 8 puertos de fibra con un precio de 3434 €.

M3F-SW8M: Switch de 8 puertos de fibra y capacidad de monitorización, cuesta con aproximadamente 4261 €.

M3F-SW16: Switch de 16 puertos de fibra con un precio de 4769 €.

M3F-SW16M: 16 puertos de fibra con capacidad de monitorización 5596 €.

SWITCHES MEDIANOS

Para switch de más bocas se han creado otros productos altamente modulares. Entre otros componentes se adjuntan turbinas para disipar el calor, fuentes de alimentación, racks o bahías de anclaje de los switch y tarjetas en blanco para evitar interferencias electromagnéticas[2] (EMI) y facilitar el flujo de aire.

Cuando se requiere switches de mayores dimensiones es necesario recurrir a las denominadas "tarjetas en línea", consiste en una tarjeta que puede soportar 8 puertos de

[2] Interferencias conducidas o EMI: este tipo de interferencias está compuesta por el ruido común y aquellas inducidas por la cercanía de aparatos electromagnéticos como por ejemplo cables.

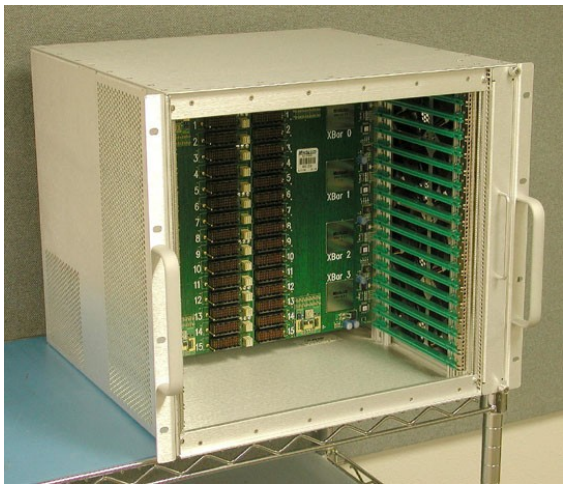
fibra. Myricom ofrece documentación sobre como deben realizarse las conexiones entre las tarjetas en línea para conseguir switches de mayor capacidad. También se ofrecen carcasas donde poder montar estas tarjetas, los modelos son los siguientes:

M3-E16: 16 puertos, ocupa 2 unidades del armario, permite 2 tarjetas en línea, cuesta 1356 €

M3-E32: 32 puertos, ocupa 3 unidades, 4 tarjetas en línea, costo 2713 €

M3-E64: 64 puertos, 5 unidades, 8 tarjetas, 5426 €

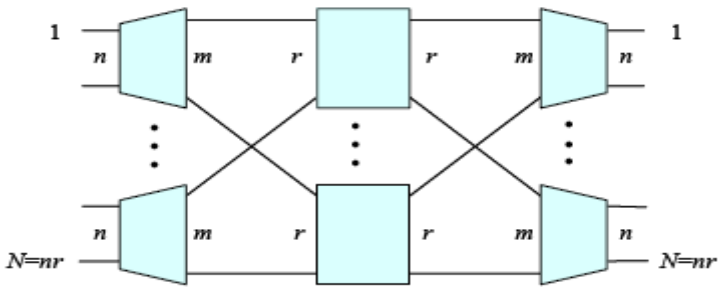
M3-E128: 128 puertos, ocupa 9 unidades del armario, 16 tarjetas, 10853 €



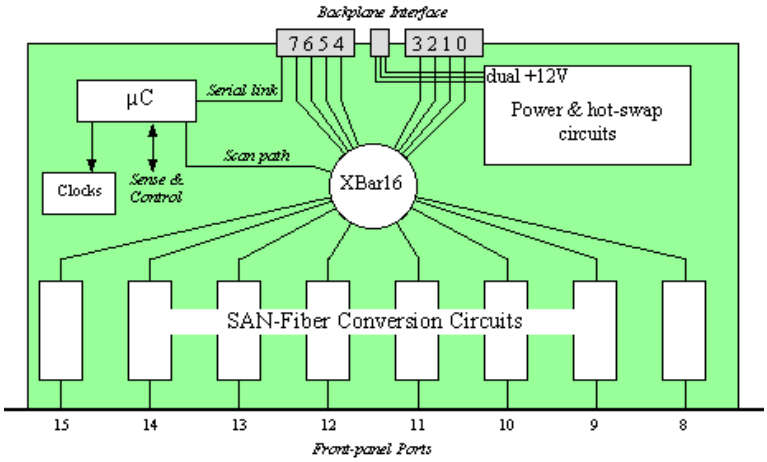
Estos recintos son completos con fuente de alimentación, el ventilador, una placa madre. Las carcasas del M3-E32, del M3-E64 y del M3-E128 incluyen 16 puertos crossbar con el fin de recrear una red de Clos[3]. La ranura superior se reserva para tarjeta en línea de monitorización, el resto se usan para las tarjetas con hasta 8 puertos en el panel delantero y 8 puertos SAN en su parte trasera, que conectan con la placa madre de la carcasa.

Las tarjetas en línea permiten 2Gb/s+2Gb/s (por hilo) en cada puerto y mantienen un microcontrolador para realizar las tareas de autotesteo, monitorización de voltaje y temperatura, identificación de tarjeta e información sobre el estado del circuito conversor de SAN-Fibra (en las tarjetas que lo tengan). La información es servida al host gracias a la tarjeta de monitorización M3-M. Su consumo es de 40.8W y su temperatura máxima no puede exceder los 40ºC.

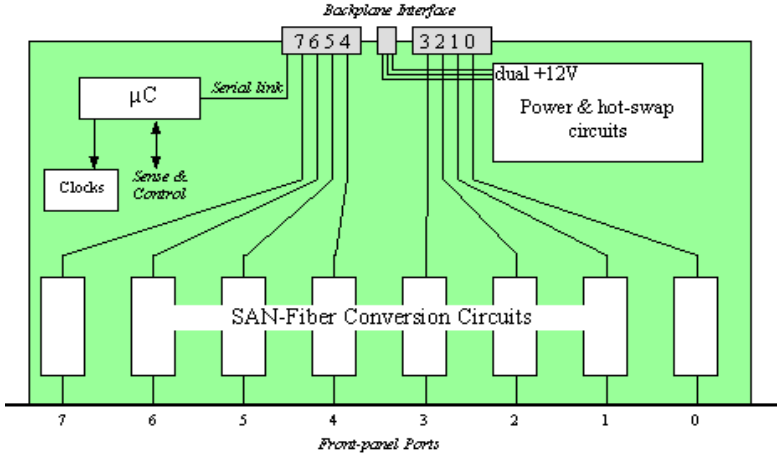
M3-SW16-8F: switch con 8 puertos de fibra y 8 puertos SAN en su parte trasera. Tiene un coste de 2035 €. Convirtiendo los hosts conectados al "Front-panel" en hojas de la red Clos.



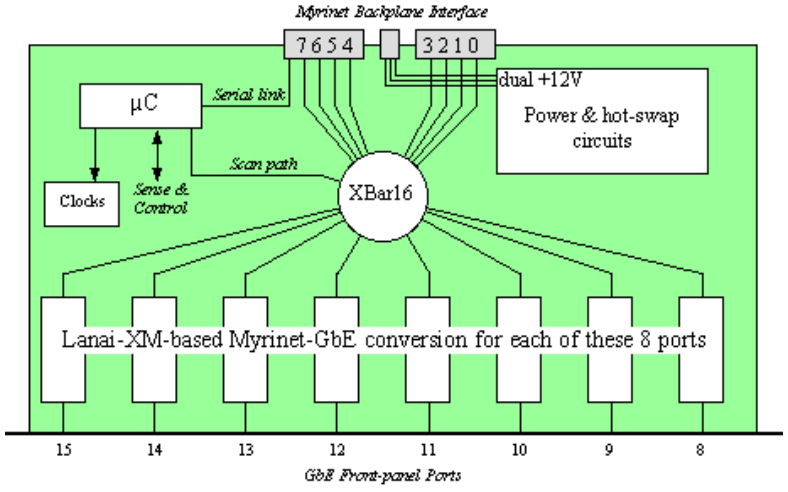
[3] Ejemplo de red Clos



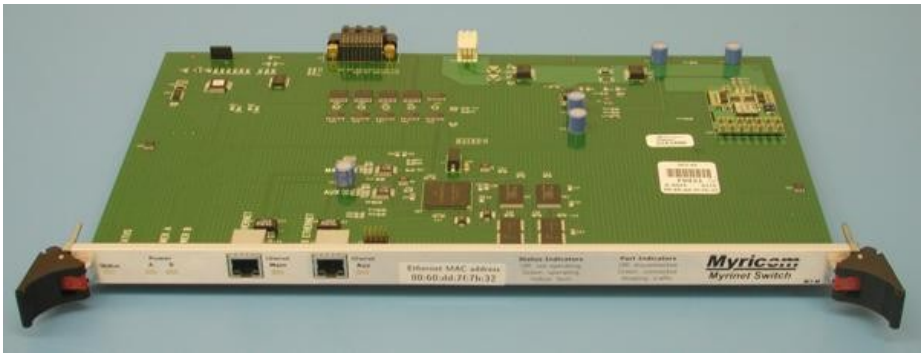
M3-SPINE-8F: Convierte a nivel físico 8 puertos de fibra y los 8 puertos SAN. Su precio está en torno a los 1356 €. Permiten una conexión directa con el backplane para permitir implementar otras topologías (distinta a la Clos). Esta tarjeta es la más usada para la expansión de más de 128 puertos, junto a la M3-SW16-8F en un recinto M3-E16.



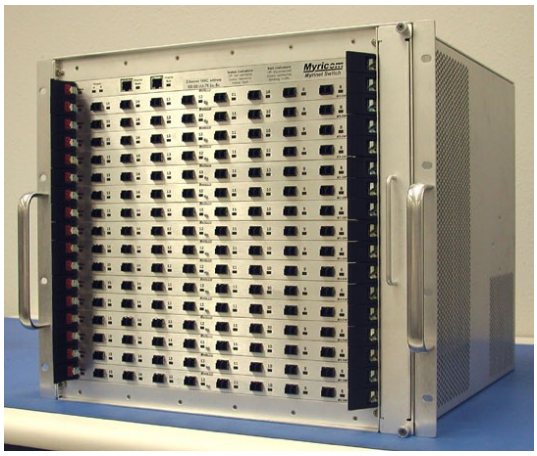
M3-SW16-8E: Tarjeta que conecta 8 puertos Gigabit-Ethernet a los 8 puertos SAN, 2713 €. Esta tarjeta es similar a la M3-SW16-8F, exceptuando que el front-panel está formado por puertos 1000BASE-T Gigabit-Ethernet.



Tarjeta en línea de monitorización M3-M, permite la dualidad con puertos Ethernet, tiene un precio de 847.9 €. Incluye un microprocesador que se comunica por la red interna de la carcasa con todas las tarjetas en línea. Proporciona servidor web, monitorización SNMP/Ethernet y control del switch.



Este es el resultado del ensamblado de todos los componentes:



SWITCHES DE MAYORES DIMENSIONES

Permite hasta 21 tarjetas en línea, se reserva el slot más a la izquierda para la tarjeta de monitorización para la supervisión de las mismas. Las tarjetas pueden ser intercambiadas *en caliente*, es decir, no es necesario apagar la fuente de alimentación. También puede retirarse en caliente una de las cuatro fuentes de alimentación (una es redundante) y los ventiladores. Se recomienda el uso de al menos una tarjeta muda (vacía) con el fin de evitar las EMI y favorecer la ventilación. Se proporcionan tarjetas en línea de hasta 16 puertos.

M3-SW32-16F: 16 conectores de fibra en el front-panel y el back-end 16 puertos seriales, sus características son similares al M3-SW16-8F (salvo el número de puertos). Cuesta 3052 €.

Cuatro virutas de Myricom XBar32 con los puertos 16-31 conectaron con los puertos de la placa madre, y los puertos 0-15 conectado con los puertos de fibra óptica del cuadrángulo del panel de delante. Cada transmisor-receptor del cuadrángulo conecta con un puerto dado en las 4 virutas XBar32. Los puertos del panel de delante se etiquetan 0-15 que corresponde al número de acceso XBar32.

M3-4SW32-16Q: un cuarteto de 32 puertos seriales (SW) y 16 cuarteros de puertos de fibra para el front. Cuesta 12210 €

M3-2SW32: dos líneas de 32 puertos seriales. Se usa como tarjeta para crear la res Clos256 (para un máximo de 256 puertos). Su precio es de 2035 €.

M3-THRU-16Q: Línea para la conexión "verdadera" de 16 cuarteros de fibra óptica. Cuesta 8140 €.

- **Clos256:** 256 host ports, no inter-switch ports. The line cards in an M3-CLOS-ENCL, left to right, are 1 M3-MONITOR, 8 M3-SW32-16F, 4 M3-2SW32, 8 M3-SW32-16F. The list price of the components for this configuration is $80,000 ($312.50 per host port). For fewer than 256 host ports, substitute M3-AIRDAM blank line cards for M3-SW32-16F line cards.
- **Clos256+256:** 256 host ports, 256 interswitch ports on 64 quad-fiber ports (see photo to the right). The line cards in an M3-CLOS-ENCL, left to right, are 1 M3-MONITOR, 8 M3-SW32-16F, 4 M3-4SW32-16Q, 8 M3-SW32-16F. The list price of the components for this configuration is $128,000.
- **Spine768:** 6 quad (24) 32-port switches presented on 192 quad-fiber ports (768 ports total). This configuration may be used as the spine for 3 Clos256+256 units. The line cards in an M3-SPINE-ENCL, left to right, are 1 M3-MONITOR, 6 repetitions of 1 M3-4SW32-16Q and 1 M3-THRU-16Q, and 8 M3-AIRDAM. The list price of the components for this configuration is $157,600.
- **Spine1024:** 8 quad (32) 32-port switches presented on 256 quad-fiber ports (1024 ports total). This configuration may be used as the spine for 4 Clos256+256 units. The line cards in an M3-SPINE-ENCL, left to right, are 1 M3-MONITOR, 8 repetitions of 1 M3-4SW32-16Q and 1 M3-THRU-16Q, and 4 M3-AIRDAM. The list price of the components for this configuration is $205,200.
- **Spine1280:** 10 quad (40) 32-port switches presented on 320 quad-fiber ports (1280 ports total). This configuration may be used as the spine for 5 Clos256+256 units. The line cards in an M3-SPINE-ENCL, left to right, are 1 M3-MONITOR, and 10 repetitions of 1 M3-4SW32-16Q and 1 M3-THRU-16Q. The list price of the components for this configuration is $252,800.

Para la configuración de 256 a 512 puertos se requiere contacto con la empresa.

Myri-10G

Incorporación de10-Gigabit Ethernet a Myricom

Los nuevos productos de Myrinet están orientados a complementar funcionalidades de las redes de 10Gb Ethernet. Ethernet tiene una gran cantidad de drivers para muchas plataformas (Windows, Linux, Mac, Solares,…) con el máximo rendimiento, el mínimo coste y cumpliendo con los estándares Ehternet. Sobre este tipo de MAC funcionan TCP/IP y UDP/IP con un rendimiento de procesamiento entorno a los 9.8 Gb/s.

La intención de unificar los criterios de baja latencia de 10-Gigabit Ethernet y la gran variedad de soporte software que proporciona Myrinet hace que se cree MX (Myrinet Express). MX permite que la conexión de las MAC a los switch 10-Gigabit Myrinet o 10-Gigabit Ethernet se haga como si se tratase de una interfase Ethernet con una menor latencia, comunicaciones MPI y las API's de los sokets.

Características de los Enlaces físicos de 10-Gigabit Ethernet:

Los puertos de Myri-10G cumplen con XAUI[4], establecido como regulación en IEEE 802.3ae, donde entre otras características, su ancho de banda tiene que ser de 10+10Gb/s en conexión full-duplex.

El cableado puede ser sobre fibra o bien sobre cobre, deben de tener una distancia máxima 200m y 15 metros, respectivamente. Evidentemente sobre fibra se consiguen mayores ratios en la velocidad de transmisión.

Características básicas de funcionamiento

Latencia de 2.3µs con MPICH-MX, ancho de banda de 1.2 GB/s (al hacer un ping) en una dirección y 2.2 GB/s en las dos direcciones (enviar y recibir). A través de los switch 10-Gigabit Ethernet se consigue el mismo ancho de banda, pero la latencia está en un rango de 2.6 a 2.8µs.

En el modo de trabajo 10-Gigabit Ethernet o la emulación de Ethernet MX-10G se consiguen ratios 9.6 Gbits/s (TCP/IP).

MAC

Pueden operar bajo dos tipos de protocolos de red 10-Gigabit Ethernet y 10-Gigabit Myrinet. Como el resto de NIC que Myrinet produce incluyen su propio procesador y soporte para el procesamiento de la información a fin de liberar a la CPU de trabajo.

Las MAC se conectan a los ordenadores a través del slot PCI-X x8[5] que opera a 2+2 GB/s en una conexión full-duplex, suficiente para el límite en el que trabajan los puertos de la MAC (1.25+1.25GB/s) para lograr una red a 10Gb. Las tarjetas Myri-10G son las únicas en trabajar con velocidades tan grandes sobre un soporte físico basado en cobre.

En vistas a sus funcionalidades en servidores se puede instalar físicamente sobre los mismos sin necesidad de que se incorporen ningún elemento físico (raíles) como soporte.

[4] XAUI: estándar donde se especifican todas las características que los puertos deben cumplir para realizar conexiones a 10 Gigabit Ethernet.

[5] El PCI Express x4 / x8 / x12 estaban inicialmete destinados para operar en servidores, en concreto, el x8 puede llegar a alcanzar un ancho de banda de unos 40Gbps. Para más información:
http://www.dell.com/content/topics/global.aspx/vectors/en/2004_pciexpress?c=us&l=en&s=corp

Los distintos tipos de tarjetas se establecen según el tipo de medio físico por el que se envía la señal: 10GBase-CX4[6], 10GBase-R[7] o con fibra bajo especificaciones de XAUI.

Las diferencias entre los modelos 10G-PCIE-8A-** y 10G-PCIE-8AL-** es el tipo de especificaciones PCI-X que se requiere.

10GBASE-CX4: 10G-PCIE-8A-C Y 10G-PCIE-8AL-C

Tiene un precio de 674.1€. Su puerto trabaja a velocidades de 10+10 Gbit/s (full-duplex) bajo las especificaciones IEEE 802.3ak para 10GBase-CX4 (tipo de cable coaxial). Cuando opera en modo Ethernet el flujo de control que la tarjeta se comporta según las especificaciones definida por IEEE 802.3x. El puerto conecta con 10GBase-CX4 y no debe de sobrepasar los 15m. Pueden usar slots PCI Express x8 o x16.

En los host la información se envía a 2 GB/s (250 MB/s por cada hilo del cable) en conexión full-duplex.

Se incorpora el procesador Lanai-Z8E, el cual opera como mínimo a 300MHz (arquitectura RISC). Una memoria de 2MB (proporciona la información al procesador a 2,400 MB/s), EEPROM de 512KB con la configuración mínima necesaria (configuración del PCI-X, firmware entre otros) y reprogramable por el procesador. En el futuro se espera incorporar a esta memoria el driver de arranque de la tarjeta (etherboot).

Sus dimensiones físicas son alto 68.9mm, longitud 147.3mm y 22mm de anchura, su peso es de 88g. Requiere una alimentación de 3.3V (que proviene de slot del bus) con un consumo máximo de unos 8.3W. Soporta hasta un máximo de 55ºC.

Cableado 10GBase-CX4

Distancia	*Código del producto*	*Precios*
1m, 30AWG[8]	10G-CX4-01M	127.1 €
2m, 30AWG	10G-CX4-02M	135.6 €
3m, 30AWG	10G-CX4-03M	144.1 €

[6] Cable coaxial especial, 4 cables en cada sentido a 3.125 Gbaud cada uno, distancia debe de ser menor a 15 m, sobre XAUI, codificación 4B5B.

[7] 10GBASE-R (en función del tipo de fibra: SR, LR, ER). Varios medios físicos (PMD):
- Fibra MMF, 850 nm, distancia < 65 m. (“Short”)
- Fibra SMF, 1310 nm, distancia < 10 km (“Long”)
- Fibra SMF, 1550 nm, distancia < 40 km. (“Extended”)

Codificación 64B/66B. Modulación a 10.3 Gbaudios.

[8] Mediad de clasificación de los cables atendiendo a la dependencia entre el diámetro y el área del conductor. Calibre 4/0 (4 ceros), al que corresponde el mayor diámetro, el número de ceros disminuye hasta alcanzar el valor 1/0. A partir de este valor el calibre del cable está asociado a un valor numérico creciente (2, 4, 6, etc). Es importante recordar que para estos calibres el diámetro del conductor se reduce cuando el valor numérico asignado aumenta.

5m, 30AWG	10G-CX4-05M	161.1 €
7m, 28AWG	10G-CX4-07M	169.5 €
10m, 26AWG	10G-CX4-10M	195 €
15m, 26AWG	10G-CX4-15M	228.9 €
Dispositivo de testeo Loopback	10G-CX4-LOOPBACK	50.87 €

NOTA: Los cables que InfiniBand utiliza los mismos conectores Fujitsu CX4 que los cables 10GBase-CX4, pero los cables que InfiniBand utiliza no cumplen el estándar sobre 10GBase-CX4.

10GBASE-R: 10G-PCIE-8A-R Y 10G-PCIE-8AL-R

Ronda los 758.9 € y esto sin incluir el tranceptor XFP, que se vende por separado. Puerto opera a 10+10 Gbit/s (full-duplex) sobre cada hilo de fibra la información se transmite a 10.3125 GBaud en una codificación 64b/66b. Dependiendo del transistor XFP[9] que se encuentre ubicado la placa frontal el puerto conectará a

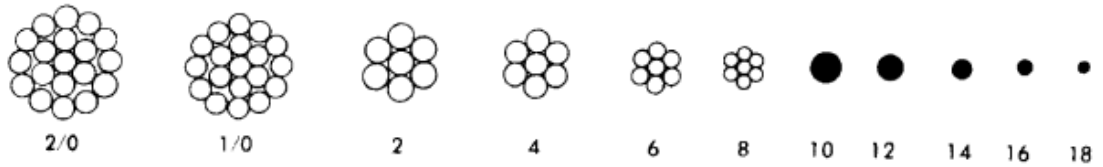


Fig. 8.1- Diámetros Relativos de Varios Calibres AWG

[9] Los transceptores XFP de 3Com son dispositivos hot-swappable estándar de la industria que se conectan a una ranura 10-Gigabit Ethernet, uniéndola con una red basada en fibra óptica o cobre. Véase la página:
http://www.3com.com/prod/es_LA_AMER/prodlist.jsp?tab=cat&cat=209632&subcat=229567

10GBase-SR (850nm de longitud de onda, 26-300m en fibra multimodo), a 10GBase-LR (1310nm de longitud de onda, 10km en fibra monomodo) o 10GBase-ER (1550nm de longitud de onda, más de 40 km de fibra monomodo). Puede operar con datos definidos en la capa Ethernet o Myrinet, en el primer modo las señales de control son la especificadas en IEEE 802.3x. La longitud permitida de la fibra depende del transmisor receptor de XFP y de la calidad de la fibra, pero en el modo Myrinet no debe exceder de los 200m.

Precio de los transceptores XFP para 10GBase-R

Distancia	*Código del producto*	*Precios*
10GBase-SR	10G-XFP-SR	423.9 €
10GBase-LR	10G-XFP-LR	763.1 €

Cables de fibra multimodo(50/125), conexión duplex para 10GBase-SR

Distancia	*Código del producto*	*Precios*
1m	M3F-CB-1M	59.35€
3m	M3F-CB-3M	63.59 €
5m	M3F-CB-5M	67.83 €
10m	M3F-CB-10M	76.31 €
25m	M3F-CB-25M	127.1 €
50m	M3F-CB-50M	169.5 €

El resto de características es igual a la 10GBase-CX4, procesador, memoria… exceptuando que su longitud y peso es mayor 165.1 mm y 91g respectivamente (sin incluir el XFP) y su consumo máximo es de 10W.

XAUI SOBRE FIBRA: 10G-PCIE-8A-Q Y 10G-PCIE-8AL-Q

Con un precio de 928.5 €. Los puertos cumplen XAUI sobre la fibra a través de la interconexión con MTP/MPO (pequeño conector de fibra capaz de conectar hasta 24 fibras) que permite ratios de 10+10 Gbit/s, full-duplex. El puerto es compatible con conectores XAUI o 10GBase-CX4, para la fibra tiene un "conversor hardware" externo, pero internamente funciona un POP4[10] (4 dispositivos MTP/MPO).Los conectores de la fibra que se usen en el puerto deben de cumplir las especificaciones del estándar MTP/MPO fibra multimodo (50/125 hilos) y no superar los 200m.

Físicamente sus dimensiones son similares a las 10GBase-CX4, pero su peso es superior (94g), pero con un menor consumo 6.3W.

Precio de cables de fibra bajo XAUI (sobre el cuarteto de puertos)

Distancia	*Código del producto*	*Precios*
1m	M3Q-CB-1M	127.1 €
3m	M3Q-CB-3M	148.3 €
5m	M3Q-CB-5M	169.5 €
10m	M3Q-CB-10M	211.9 €
25m	M3Q-CB-25M	339.1 €
50m	M3Q-CB-50M	466.3 €
100m	M3Q-CB-100M	635.9 €
150m	M3Q-CB-150M	847.9 €

[10] http://www.popoptics.org/
http://www.popoptics.org/POP4_Specification.pdf

200m	M3Q-CB-200M	1102 €

Importancia del protocolo iSCSI

Para poder realizar los test en diversas máquinas y plataformas con características propias la empresa se decidió por realizar pruebas sobre este tipo de protocolo.

El protocolo iSCSI utiliza TCP/IP para sus transferencias de datos. Al contrario que otros protocolos de red diseñados para almacenamiento, como por ejemplo Fibre Channel (que es la base de la mayor parte de las SANs), solamente requiere un simple y sencillo interfaz Ethernet (o cualquier otra red compatible TCP/IP) para funcionar. Esto permite una solución de almacenamiento centralizada de bajo coste sin la necesidad de realizar inversiones costosas ni sufrir las habituales incompatibilidades asociadas a las soluciones Fibre Channel storage area networks.

Los críticos de iSCSI argumentan que este protocolo tiene un peor rendimiento que el Fibre Channel ya que se ve afectado por la sobrecarga que generan las transmisiones TCP/IP (cabeceras de paquetes, por ejemplo). Sin embargo las pruebas que se han realizado muestrans un excelente rendimiento de las soluciones iSCI SANs, cuando se utilizan enlaces Gigabit Ethernet