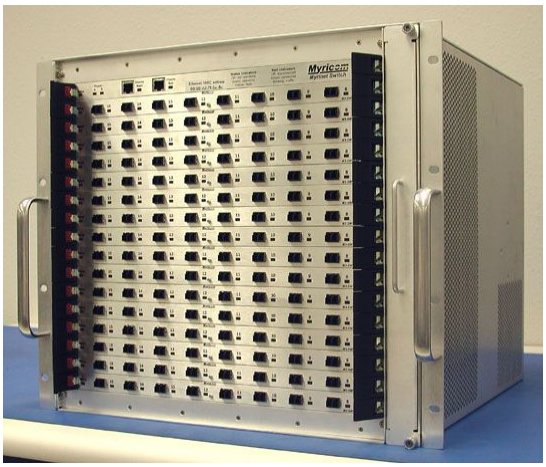
MYRINET

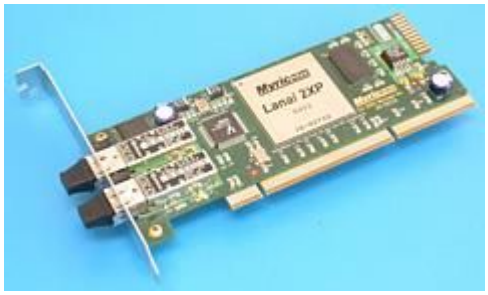
Myrinet [1] è una tecnologia per realizzare reti d'interconnessione ad alta performance utilizzata per creare clusters di workstations, servers, o anche sigoli computers.
E' particolarmente adatta per calcoli paralleli strettamente connessi, dove sono necessari alti data-rate, basse latenze di trasferimento, e il rilevamento dinamico e l'isolamento dei guasti.
Più in dettaglio quello che Myrinet fornisce sono:

-Full-duplex 2+2 Gigabit/second data-rate links.

-Switch di varia grandezza, dai più piccoli da 16 porte ai più grandi da 256 porte. Sono comunque tutti modulari e possono essere combinati tra loro; al loro interno utilizzano una topologia crossbar.

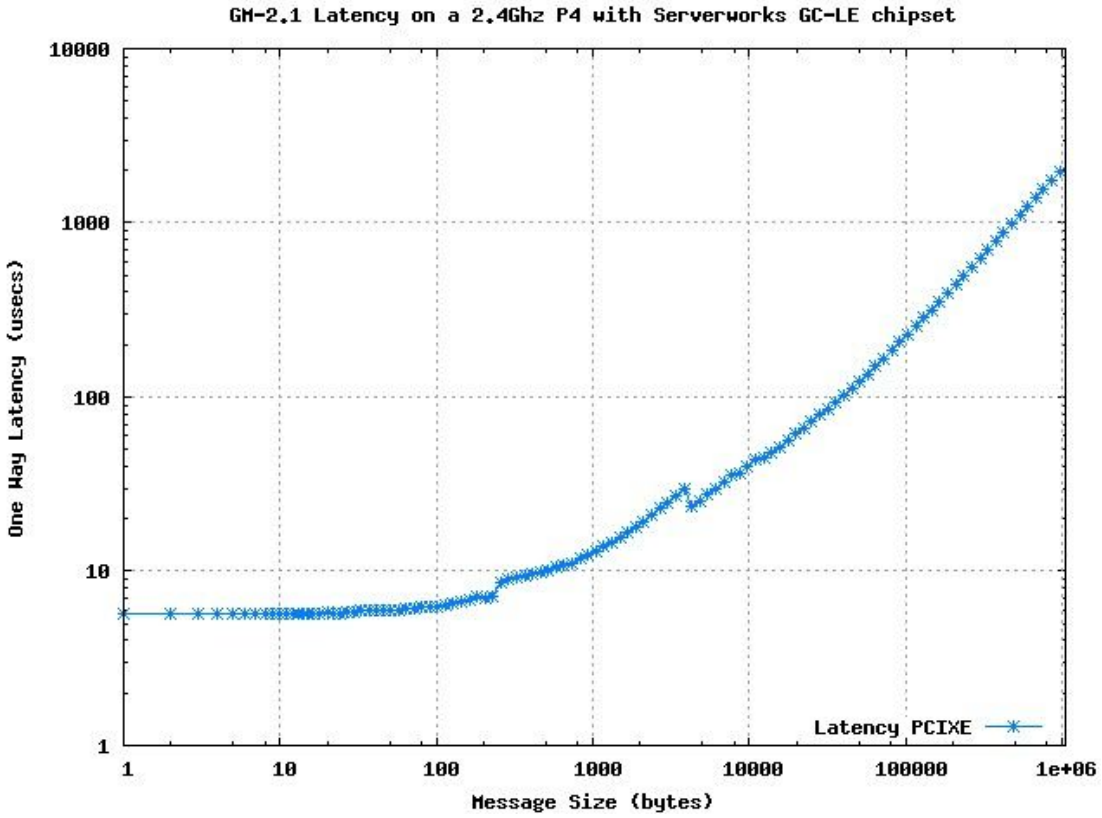


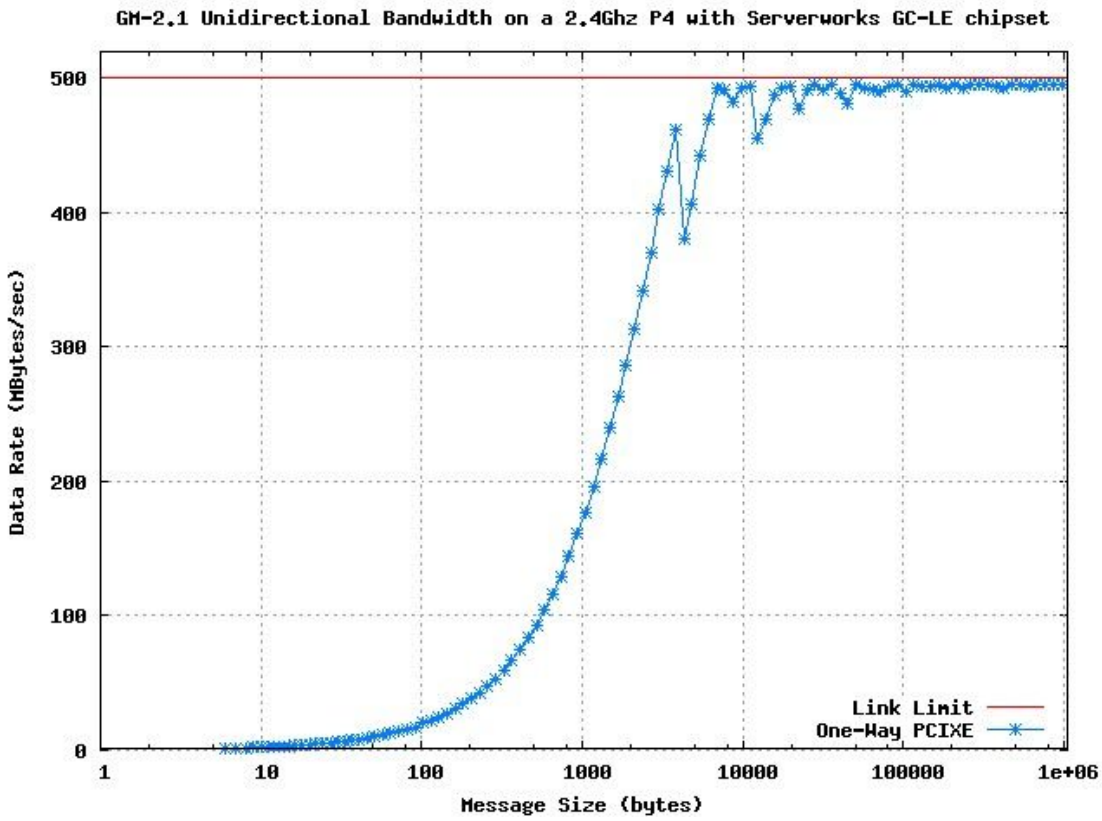
-Network Interface Cards (NICs) a porta singola o doppia , ovvero schede per slots PCI-X da inserire sui calcolatori.

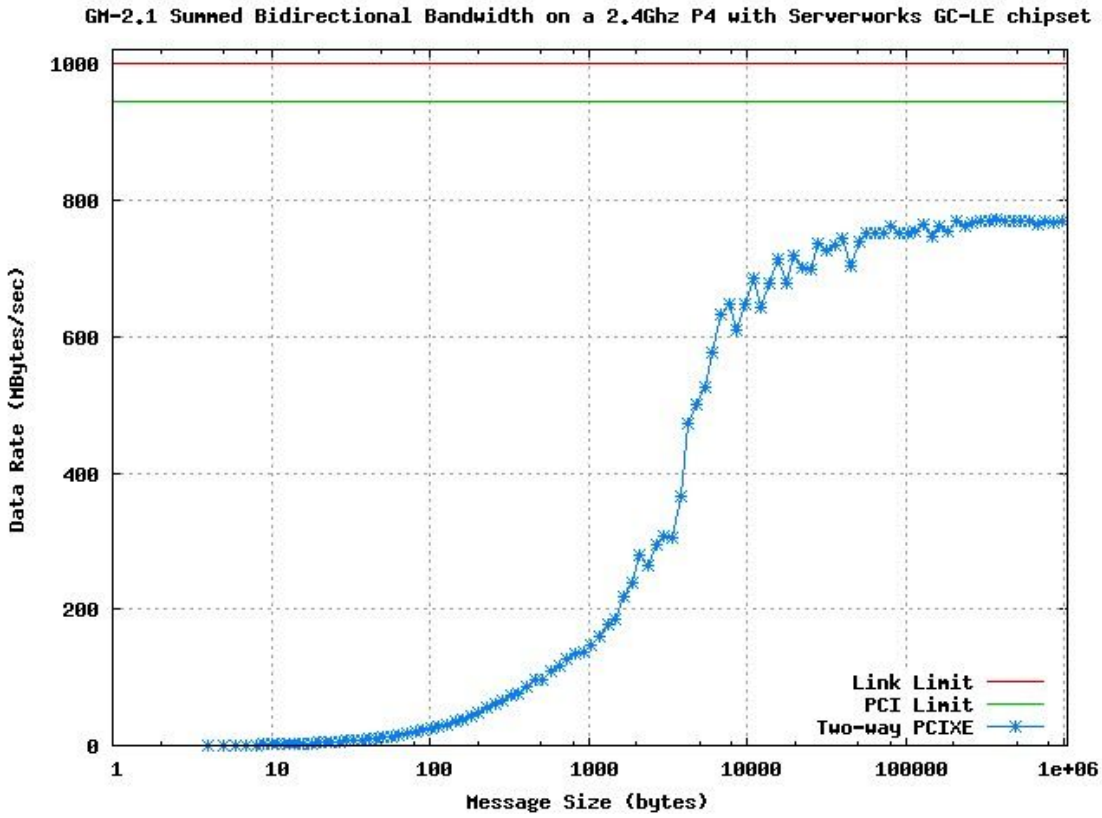


-Software open source per i maggiori sistemi operativi, che include i necessari drivers per le interfaccie di rete ma anche middleware per migliorare le prestazioni delle applicazioni.

Nel complesso questi componenti hardware e software permettono di realizzare una rete con controllo di flusso e degli errori, e con continuo monitoraggio di ogni link in modo da aggirare un qualsiasi collegamento difettoso. La maggior parte delle operazioni utili ai nodi sono direttamente implementate nel firmrware delle schede di rete, in modo tale che per inviare e ricevere messaggi non ci sia bisogno di chiamare le lente system call del sistema operativo ma si possa effettuare l'intera operazione in user space.
In particolare Myrinet offre supporto hardware nativo per i sistemi di message-passing di basso livello GM e MX, attraverso i quali possono essere realizzati anche altri sistemi di comunicazione come MPI e Sockets.
Le prestazioni dichiarate sono di latenze di 2.7µs utilizzando schede due porte con MX, con un overhead sulla CPU di soli 0.15µs per messaggio.
I data-rates sono invece di ~1.98 Gb/s (248 Mbytes/s) sulla singola direzione, che diventa ~3.9 Gb/s (~490 GBytes/s) visto dal processo host che fa uso del firmware e di schede a doppia porta.
I seguenti grafici mostrano benchmarks eseguiti utilizzando interfaccie Myrinet M3F2-PCIXE con CPU dual-2.4-GHz Pentium-4 Xeon con il chipset Serverworks "Grand Champion" (GC-LE):




GM-2.1 Unidirectional Bandwidth on a 2.4Ghz P4 with Serverworks GC-LE chipset
Data Rate (MBytes/sec)
Message Size (bytes)
Link Limit
One-Way PCIXE
0
100
200
300
400
500
1
10
100
1000
10000
100000
1e+06


GM-2.1 Summed Bidirectional Bandwidth on a 2.4Ghz P4 with Serverworks GC-LE chipset
Data Rate (MBytes/sec)
Message Size (bytes)
Link Limit
PCI Limit
Two-way PCIXE
0
200
400
600
800
1000
1
10
100
1000
10000
100000
1e+06

QsNet[II]

QsNet [2] è la soluzione di Quadrics per l'interconnessione di supercomputers; cerca di realizzare reti con un'alta banda passante, una bassa latenza e un'ottima scalabilità.
Utilizza interfaccie di rete per slots PCI-X per raggiungere una banda di 900Mbytes/s, anche in questo caso senza bisogno di passare attraverso il sistema operativo per effettuare l'invio o la ricezione di un messaggio.
La rete utilizza una topologia "fat tree", che permette di scalare fino a 4096 nodi per i sistemi di seconda generazione.

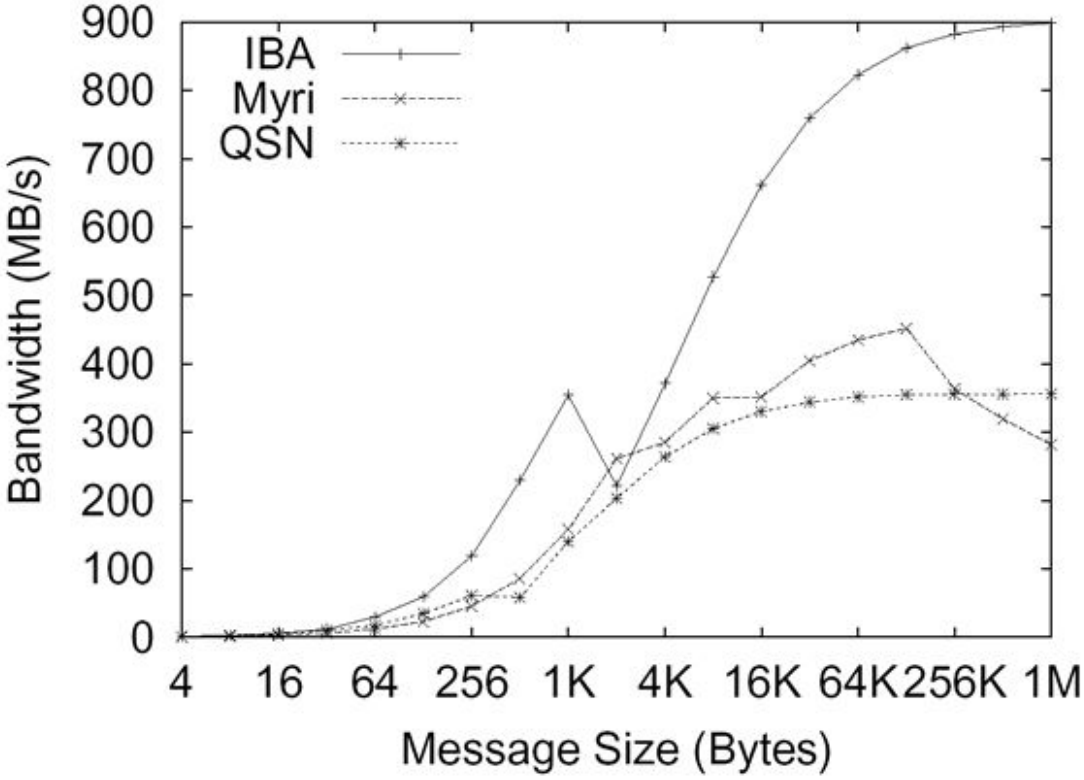
Inoltre ogni nodo può essere a sua volta un sistema multiprocessore SMP, in grado quindi di sfruttare appiendo le interfaccie di rete QsNet che sono state ottimizzate per essere utilizzate concorrentemente da più processi contemporaneamente, sempre in user space, senza alcuna necessità di task switching.
Ogni scheda di rete inoltre è fornita di processori I\O separati in modo tale che in ogni momento sia possibile sfruttare completamente l'intera banda disponibile, rendendo in pratica la prestazione dipendente dalla velocità di trasferimento del BUS PCI-X. E' possibile quindi scalare ulteriormente la performarce della rete utilizzando sistemi multi-rail, ovvero con più PCI BUS utilizzati in contemporanea.
Dal punto di vista dell'architettura di rete, come già accennato, la soluzione consigliata è quella di utilizzare le componenti QsNet per realizzare un fat tree; si ottiene in questo modo un sistema altamente scalabile e con una larghezza di banda che agli alti livelli dell'albero aumenta anch'essa con la dimensione della rete.
Inoltre data la ridondanza di connessione che ne deriva si ottiene anche una buona resistenza ai guasti. L'uso efficiente della rete è garantito dall'implementazione di algoritmi di routing adattativi in grado di scegliere dinamicamente il percorso più veloce per raggiungere la destinazione, pacchetto per pacchetto.
Il componente di base per la realizzazione di una rete di questo tipo, magari di grandi dimensione, è uno switch a 64 porte implementato come un fat tree a 3 livelli; ne servono più di uno contemporaneamente per realizzare gli switch degli stages più elevati.

Anche QsNet fornisce a corredo tutto il software necessario, a partire dai drivers fino a librerie MPI 1.2 ottimizzate per l'hardware a disposizione.
Un beanchmark effettuato sul sistema ASCI-Q presso il "Los Alamos National Laboratory" (un supercomputer basato sull'AlphaServer SC di HP con più di 2000 nodi), mostra che con QsNet si possono ottenere 2µs di latenza per messaggi fino a 256 bytes, e 300 MB/sec di banda passante.

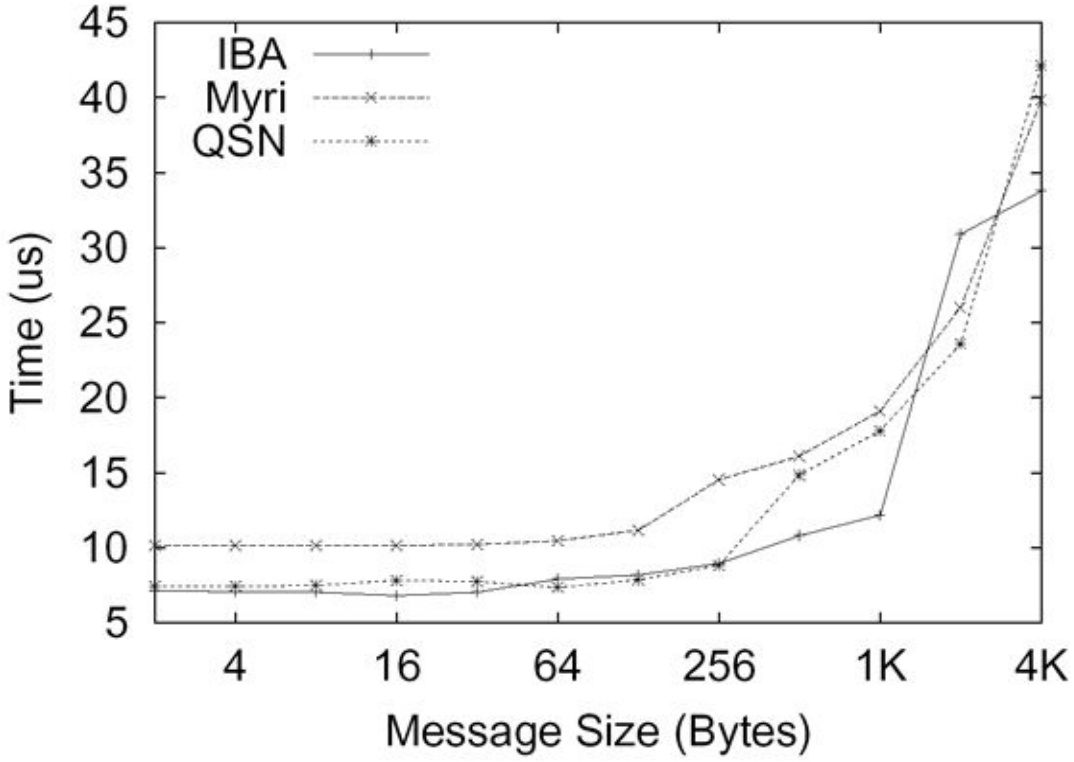
Conclusioni

I seguenti grafici mostrano le prestazioni comparate di QsNet, Myrinet e InfiniBand (qui non trattato) [3]:

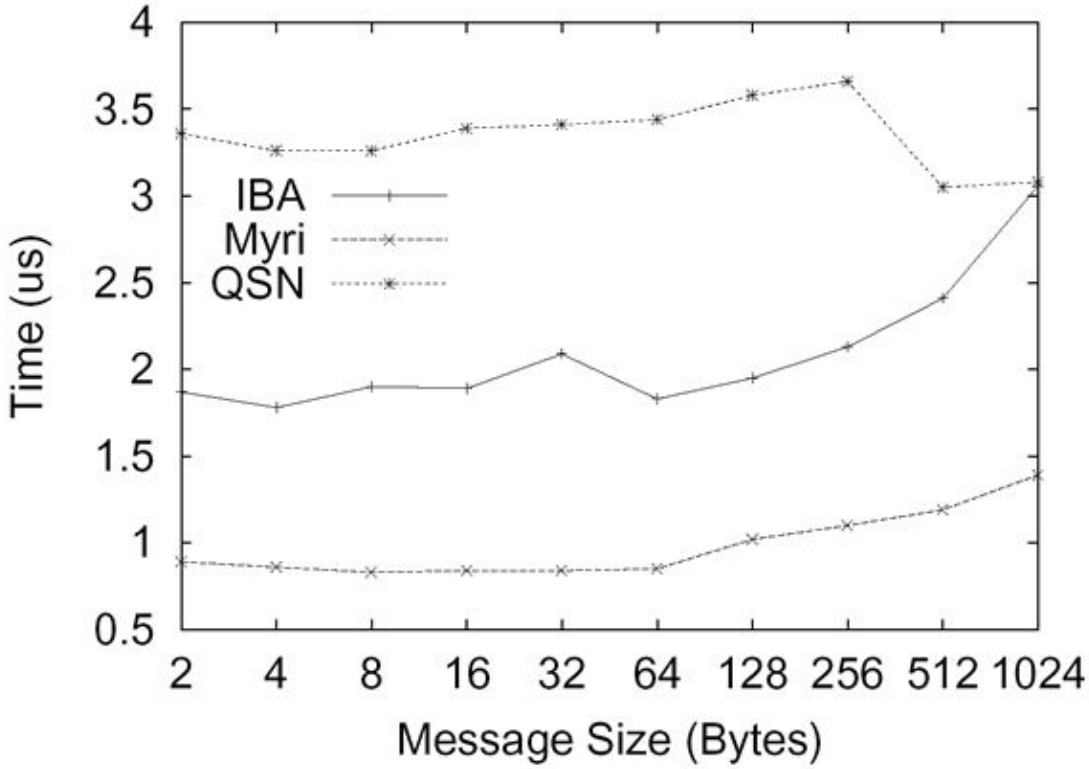
Bidirectional Bandwidth



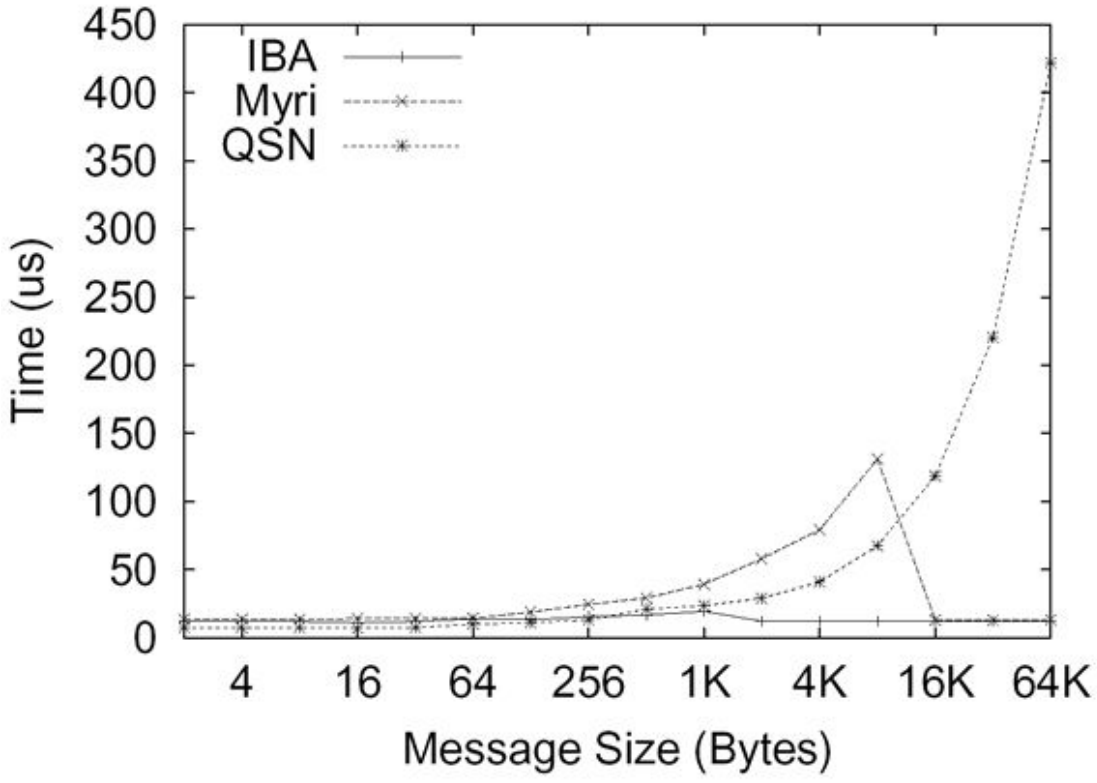
Bidirectional Latency



Host Overhead in latency tests



Overlap potential



Entrambe le soluzioni è evidente che offrono ottime prestazioni, è difficile dire con precisione quale delle due sia effettivamente la migliore, tant'è vero che al momento ben il 38% dei supercomputers presenti nella TOP500 sono equipaggieati con reti d'interconnesione Myrinet, ma il sistema più veloce utilizza QsNet.
Dal punto di vista dei costi invece Myrinet sembra riuscire ad offrire netti vantaggi sul concorrente, in termini di prezzo dei singoli componenti.

BIBLIOGRAFIA

[1] Myrinet - http://www.myri.com

[2] Quadrics - http://www.quadrics.com

[3] Jiuxing Liu, Balasubramanian Chandrasekaran, Jiesheng Wu, Weihang Jiang, Sushmitha Kini, Weikuan Yu, Darius Buntinas, Peter Wyckoff, D K Panda. Performance Comparison of MPI Implementations over InfiniBand, Myrinet and Quadrics.