Uno switch fabric seriale ad alta velocità

SBS Technologies, partner americana di EuroLink Systems, rende disponibile la nuova tecnologia InfiniBand

Franco Guzzo Eurolink Systems

Fig. 1 - IB4X-CPIEXP-2 è un HCA Infiniband / PCI Express da 10Gbps, progettata da SBS Technologies

Fig. 2 - La scheda IB4X-LPCIX-2 è capace di sfruttare pienamente la velocità di trasmissione disponibile sulle connessioni InfiniBand

InfiniBand è una tecnologia emergente che offre elevate prestazioni, bassa latenza, buona scalabilità e basso costo. Progettata per i backplane dei grossi sistemi, questa tecnologia Switch Fabric seriale è stata inizialmente introdotta nel mercato dei computer di elevate prestazioni (HPC), ma sta oggi guadagnando mercato anche nell'area dei computer embedded. I grossi sistemi che supportano più di uno switch hanno, infatti, bisogno di trasportare i dati da un nodo di uno switch fabric ad un altro nodo su un altro switch in modo deterministico e abbastanza velocemente, senza attendere che le risorse diventino disponibili. L'ampiezza di banda disponibile diventa sempre più importante per InfiniBand come per tutte le altre tecnologie di connessione fra gli switch fabric nei server e nelle workstation. Oggi, sono i PCI-X host channel adapters (HCA) che hanno il ruolo di connettere i nodi degli Switch Fabric nei grandi sistemi InfiniBand. Essi operano a 133 MHz e hanno una banda teorica aggregata di 1 GBps (Giga Byte al secondo), mentre i sistemi che usano i bus PCI-X HCA senza InfiniBand possono arrivare al massimo a 830 MBps. Difatti, la banda teorica nei PCI-X è di 133 MHz x 64 bit, ovvero 8,512 Gbps (Giga bit al secondo) e cioè 1,064 GBps. Per contro, una connessione InfiniBand consiste di 4 linee con 2,5 Gbps in ogni direzione e, siccome ciascuna è codificata 4B/5B, allora la banda aggregata risulta di 2,5 Gbps x 4 x 2 x 4/5 e dunque è pari a 16 Gbps, ovvero 2 GBps.

Naturalmente, un sistema PCI-X è incapace di sfruttare appieno la banda disponibile nelle connessioni InfiniBand. Usando la definizione tradizionale di commutatori non bloccanti, ciascuno switch dovrebbe avere le porte connesse al backplane nello stesso numero rispetto a quelle rivolte al server, così che la larghezza di banda degli ingressi e delle uscite si equivalgono. In tal caso, il sistema risulta composto da "n" porte ed "m" server connessi a ciascuno switch in modo tale che m=n/2. Perciò, considerando la larghezza di banda attuale offerta dai bus PCI-X, la banda disponibile sugli switch fabric è significativamente sotto utilizzata e, infatti:

banda I/O server BWs = m x 830 MBps = n x 415 MBps [Equazione 1]

banda Switch Fabric BWf = n/2 x 2 GBps = n x 1 GBps [Eq. 2]

Evidentemente, utilizzando meno del 50% della banda disponibile sugli switch, l'efficienza è bassa, mentre i costi sono elevati. D'altra parte, utilizzare più switch rispetto a quelli che veramente servirebbero, renderebbe parimenti inefficiente l'architettura.

Il punto di vista ottimo

Un approccio più pratico prende in considerazione le reali limitazioni della banda disponibile per gli I/O sui server. In questo esempio, la larghezza di banda per gli I/O nel server e la banda dello switch sono uguali:

banda I/O server BWs <= banda switch fabric BWf [Eq. 3]

m x 830 MBps <= (n – m) x 2 GBps [Eq. 4]

830 MBps / 2 GBps <= (n-m) / m

n/m >= 1,415

m <= n / 1,415 [Eq. 5]

Questa equazione, per la granularità degli switch disponibili pari a 8, 24, 96 e 144 porte, si trasforma nelle configurazioni mostrate in tabella 1.

Inoltre, i server connessi agli switch potrebbero non raggiungere la banda di 830 MBps e, di conseguenza, la tabella 1 rappresenta il caso migliore. In questo caso gli 830 MBps assunti per l'equazione 4 vengono ridotti e il rapporto fra le porte del server rispetto alle porte dello switch fabric nell'equazione 5 aumenta. È bene tenere presente che in ogni architettura switch fabric, l'equazione 3 è quella che pesa di più sulle prestazioni dell'interfaccia. Si può fare un esempio considerando un'architettura switch fabric con 72 server connessi da 96 porte, dove 24 porte su ogni switch sono usate per l'interconnessione degli switch stessi. Analizziamo la bontà delle prestazioni di quest'architettura:

banda switch fabric BWf = 24 x 2 GBps = 48 GBps

banda I/O server BWs = 48 GBps / 72 = 670 MBps [Eq. 6]

Tuttavia, è lecito chiedersi se queste prestazioni siano adeguate. Per rispondere basta considerare che, assumendo una distribuzione di traffico simile sui 288 server, circa il 25% del traffico entrante in uno switch da un server andrà a un server connesso a quello switch. Quindi, possiamo ricalcolare l'equazione 6 in questo modo:

banda I/O server BWs = 48 GBps / (72 x 75%) = 890 MBps [Eq. 7]

In ogni caso, la capacità dei server di caricare lo switch fabric non può essere quantificata con precisione in modo empirico e dev'essere misurata nei test. Per questo motivo è probabile che il server non raggiungerà mai il limite pratico del bus PCI-X di 830 MB/s e quindi l'architettura in esempio è probabilmente adeguata.

Ci sarebbero da considerare alcune varianti riguardo alla rete di connessione che lega tutti gli switch attraverso il backplane. Tale rete potrebbe semplicemente connettere fra loro pochi switch, mentre in altri casi, la rete o "fabric" può essere più complessa e composta da molti switch che interconnettono a loro volta altri switch, ai quali sono connessi i server. Il tipo di architettura più complessa ora descritta può arrivare a comprendere centinaia di nodi.

Un'architettura ottimizzata tiene conto delle capacità dei server attuali ed è consistente con il bilanciamento del numero delle porte rispetto alle bande disponibili, così come si vede nella tabella 1. Un'architettura per un sistema formato da centinaia di nodi, naturalmente, richiede un maggior numero di switch e molti più cavi, i quali incrementano i costi in modo significativo senza per contro migliorare le prestazioni.

I sistemi reali

Sistemi ancora più complessi possono avere più di un livello di switch non connessi ai server direttamente, detti "Spine Switches". Nelle architetture multilivello gli Spine Switch che formano la rete Switch Fabric devono essere connessi in modo tale da utilizzare pienamente la banda di 2 GBps, disponibile sulle connessioni Infiniband verso tutti gli switch radice connessi ai server.

In conclusione, prima di decidere di usare un'architettura non bloccante tradizionale, è utile testare le prestazioni del server, tenendo conto che il concetto "non blocking" dev'essere applicato alla banda entrante nello switch dal server e non necessariamente alla velocità del link fisico relativo alla connessione. Solamente in un secondo tempo, qualora le prestazioni del server aumentino significativamente, allora potrà essere incrementato il numero degli switch, in modo da inseguire l'aumento di banda.

SBS Technologies offre svariate soluzioni all'avanguardia in proposito, fra cui la scheda IB4X-CPIEXP-2 e la scheda Infiniband 4 x Duale da 10 Gbps. La prima è un HCA Infiniband – PCI Express da 10 Gbps, progettata per fornire un'elevata velocità di trasporto e una ridotta utilizzazione della CPU in tecnologia High Performance Computing Clustering (HPCC). Questa HCA è configurata con 128 MByte di memoria DDR. L'architettura Infiniband definisce e supporta molte applicazioni, il più delle quali con capacità RDMA (Remote Direct Memory Access), come: video streaming, aerospazio, militare, controllo elettronico, interfacce MPI (Message Passing Interface), Mirror API.

Questa scheda è compatibile con PCI Express Revision 1.0a, mentre la scheda IB4X-CPCIX-2 e la scheda IB4X-LPCIX-2 sono basate sull'HCA prodotta da Mellanox InfiniHost, un sistema capace di sfruttare pienamente la velocità di trasmissione disponibile sulle connessioni InfiniBand.

TABELLA 1 - CONFIGURAZIONE DEGLI SWITCH CONFORMI ALL'EQUAZIONE [5]

porte switch (n)	porte server (m)	banda server (BWs)	porte switch fabric (n-m)	banda switch fabric (BWf)
8	5	5.15 GBps	3	6 GBps
24	16	13.28 GBps	8	16 GBps
96	67	55.61 GBps	29	58 GBps
144	101	83.83 GBps	43	86 GBps