

# Una panoramica su PCI Express

Caratteristiche e motivazioni, il confronto con PCI e altre tecnologie

Pietro Grillo

**I**n questo articolo si descrivono le caratteristiche principali di PCI Express, effettuando un confronto diretto con PCI (e le sue versioni Enhanced) e con altre tecnologie. L'obiettivo è comprendere cos'è PCI Express, le possibilità di utilizzo, le prospettive di diffusione nonché le effettive potenzialità. In un prossimo articolo si cercherà invece di approfondire l'argomento partendo dall'architettura e procedendo con la descrizione dello stack layer, del routing, del QoS (Quality of Service) e controllo di flusso per arrivare alla configurazione di PCI Express. L'articolo contiene anche un utile ripasso relativo a PCI, evidenziando caratteristiche e limiti di questa tecnologia, perché proprio il successo di PCI, insieme ai suoi limiti, hanno condotto allo sviluppo di PCI Express, che risulta tra l'altro software compatibile "a ritroso" con PCI.

## Un po' di storia

Alcune aree attinenti la tecnologia dei computer, dai processori alle schede video, hanno avuto interessanti sviluppi e hanno raggiunto bande e data rate estremamente elevate. La stessa

cosa non è avvenuta nell'ambito dei dispositivi di I/O. Il bus PCI è nato negli anni '90 e da allora i cambiamenti sono stati molto limitati. L'introduzione di nuovi bus seriali, tra cui appunto PCI Express, ha contribuito a vivacizzare questo panorama.

Hard disk, periferiche, schede LAN, schede audio, USB e Firewire devono "viaggiare" sul bus PCI, lo stesso nato con il primo PC 486 negli anni '90. Se si considera che un Pentium 4 con memoria DDR può operare a 2,1GB/s sul bus di memoria, mentre schede 8x AGP per la parte grafica lavorano sui

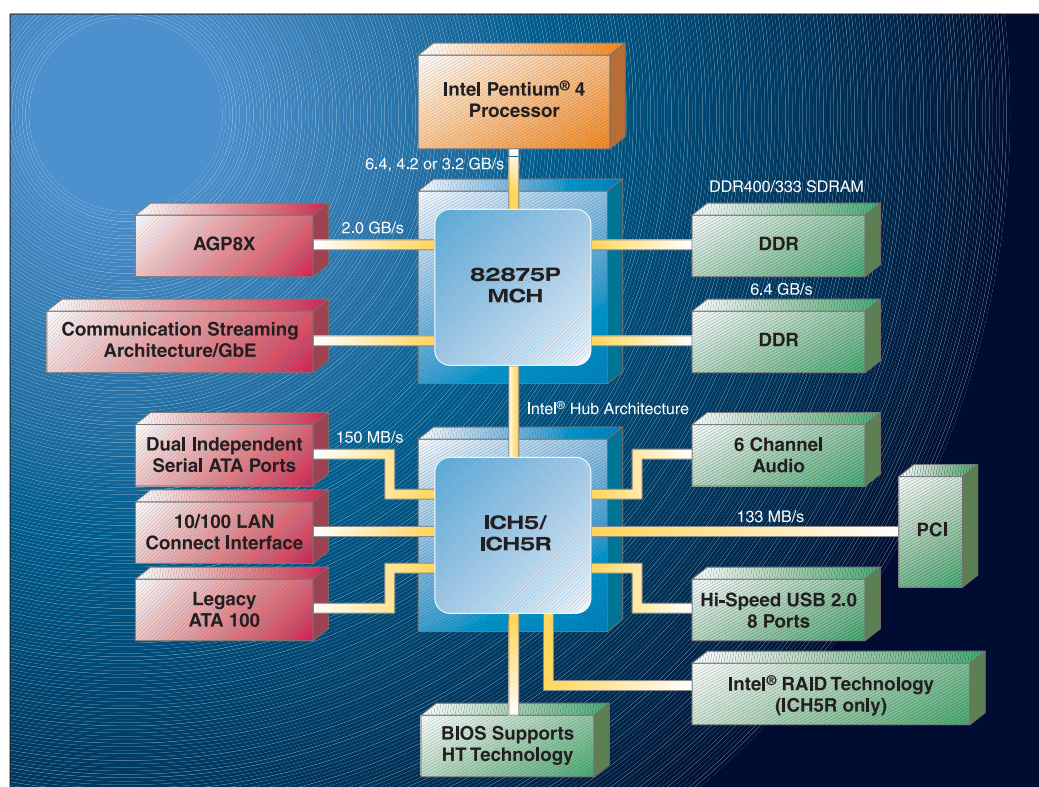
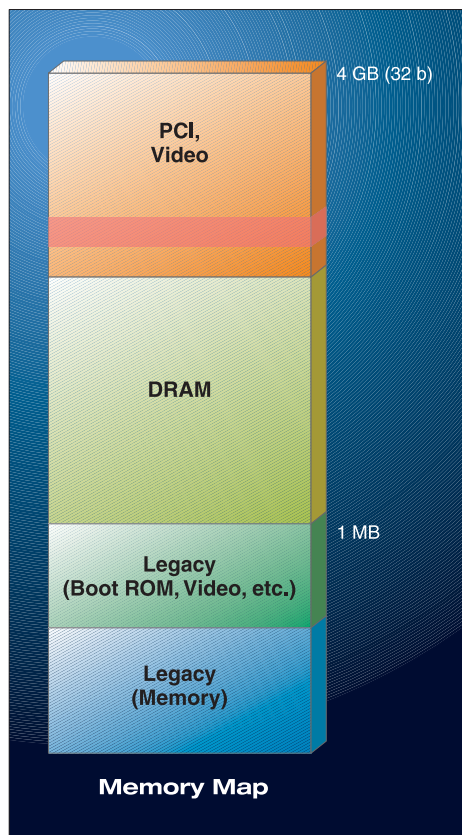


Fig. 1 - Schema del chipset di server Intel

**Fig. 2 - Mappa di memoria di PCI**



2,1GB/s, appare chiaro che i 33MHz e i 133MB/s di PCI rappresentano un limite per le nuove macchine.

PCI è un simbolo della diffusione del PC, in quanto ha reso disponibile una piattaforma stabile e flessibile che ha consentito ad hardwaristi e softwaristi di implementare dispositivi "Plug & Play", che hanno rappresentato la base della semplificazione dell'uso dei computer contribuendo all'esponenziale diffusione dello stesso.

Prima di PCI Express sono state proposte soluzioni per l'aggiornamento di PCI, che non hanno avuto l'esito sperato. Per esempio, considerando la banda come uno dei principali problemi di PCI, si è cercato di apportare dei miglioramenti in tal senso con l'introduzione di PCI-X: 64bit/133MHz e successivi. Alcuni limiti però ne consentono lo sviluppo solo al livello dei server: per esempio PCI-X richiede un controller per ogni slot.

Esaminando il percorso evolutivo dei bus, si possono distinguere tre generazioni.

La prima generazione è costituita da bus dedicati, come ISA (Industry Standard Architecture), MCA Microchannel Architecture), EISA (Extendend ISA), VESA (Video Electronics Standard Association).

La seconda vede l'affermazione di un bus comune, PCI (Peripheral Component Interconnect), in grado di gestire velocità comprese tra 132 e 528 MBps. Introdotto negli anni '90, al tempo dei primi processori a 33MHz, esso era destinato principalmente ai settori embedded, desktop e apparati mobili in applicazioni quali "communication access system" e "control plane". Una delle sue evoluzioni, PCI-X, che prevede velocità superiori (nell'intervallo di 1064-4096 MBps) e utilizzato in server e applicazioni di immagazzinamento dati può essere ritenuto appartenente alla seconda generazione, in quanto si tratta di una versione enhanced di PCI.

La terza è focalizzata su PCI Express, un duplex seriale con velocità comprese tra 250 e 8000 MBps che per le sue caratteristiche si adatta a qualsiasi applicazione, comprese quelle multiprocessor (peer-to-peer). Il multiprocessor è possibile grazie alla capacità di effettuare operazioni di bridging in

maniera non trasparente. Questa potenzialità è di sicuro interesse: ad esempio per poter implementare schede grafiche con CPU a bordo si potrebbe utilizzare PCI Express per il backplane o nei multiprocessor server e in piattaforme modulari.

L'implementazione di IOV è un'altra interessante caratteristica di PCI Express. Acronimo di Input/Output Virtuale, IOV rappresenta la possibilità di condividere un singolo I/O fisico da parte di più "immagini" del sistema. Per dispositivo fisico si intende la memoria, oppure hard disk, reti, server e così via. Quando si parla invece di immagine di sistema si intende un sistema operativo dedicato o di tipo general purpose al quale si possa assegnare dispositivi fisici o virtuali.

## La terminologia usata

**È utile riassumere la terminologia più usata in ambito di PCIe, facendo un breve elenco dei termini principali.**

**Lane:** Una linea è un paio di coppie di segnali differenziali tra due SERDES (serializzatore-deserializzatore) - comunicazione dual simplex

**Link:** Una o più linee logicamente connesse che determinano un singolo bus punto a punto

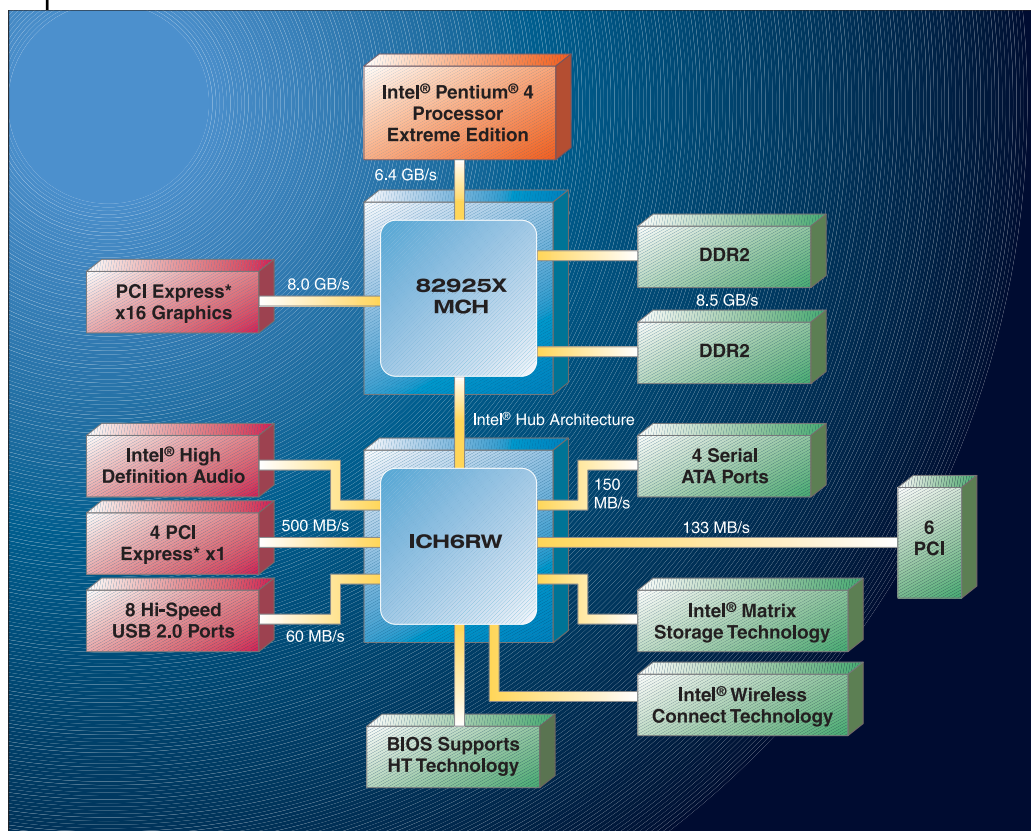
**Port:** L'endpoint dual simplex di un link

**Root Complex:** MCH/ICH/Memoria. Sostanzialmente il master dell'albero del bus PCIe

**Bridge:** tipicamente si riferisce alla logica che converte tra un PCIe e altri protocolli, come PCI, ma può anche riguardare la transazione a livello del transaction layer all'interno di un switch PCIe

**Switch:** È un bridge multiporta di tipo PCI-PCI (P2P), un Bridge per ogni porta

**Fabric:** la collezione di switch all'interno di un dominio



**Fig. 3 - Cambiamento dell'architettura del chipset di server Intel con PCIe**

esiste condivisione sul bus. A livello dei PC, PCIe soppianta lo slot AGP per le connessioni grafiche e permette il collegamento di chipset GB Ethernet. Inoltre aumenta anche la potenza trasportabile sul bus, con conseguente diminuzione di alimentazioni aggiuntive sulle schede, per esempio quelle grafiche.

### Le caratteristiche di PCI

PCI è un bus di tipo multi-drop parallelo che prevede una comunicazione sincrona

Ovviamente ci sarà un software o firmware intermediario (IOV) usato per supportare lo IOV.

Due parole sull'utilizzo di questa funzionalità. Se si calcola l'uso effettivo delle componenti di un sistema si scopre che spesso esse sono utilizzate in una percentuale molto bassa, in modo inefficiente. Spesso le risorse di I/O sono duplicate, perché servono su più fronti. È possibile aumentare l'efficienza condividendo su bus PCI Express le risorse implementando un meccanismo di I/O virtuale.

PCI-SIG, il comitato incaricato di trovare l'alternativa a PCI è l'organismo che ha dato il nome PCI Express a questo nuovo standard voluto da Intel in partnership con IBM, Dell, Compaq, HP e Microsoft.

PCI Express, nato come sviluppo di PCI in grado di far fronte ai cambiamenti in atto e alle esigenze dei sistemi, in realtà finisce per essere un bus completamente nuovo (pur mantenendo la compatibilità a ritroso con PCI). PCI Express non è più un bus parallelo ma un bus seriale, o meglio un assemblamento di fili che connettono dispositivi punto-a-punto, individuabili come linee. Una connessione base prevede due coppie di connessioni differenziali, quindi 4 punti di connessione, a differenza dei 32 di PCI. Inoltre ogni linea è utilizzata per connettere due punti e questo comporta lo sfruttamento totale della banda da parte dei dispositivi interessati e non

da 32b a 33MHz fino a 64b e 66MHz. I carichi possibili sono limitati, si parla di 10 carichi on board oppure 5 slot. Questo bus infatti è soggetto a problemi di riflessione su linea e skew, senza dimenticare che il bridging comporta un aumento della frammentazione. I segnali non sono organizzati a pacchetti o messaggi ma sono di tipo Out-of-band (Interrupt, errori, segnalazioni in genere).

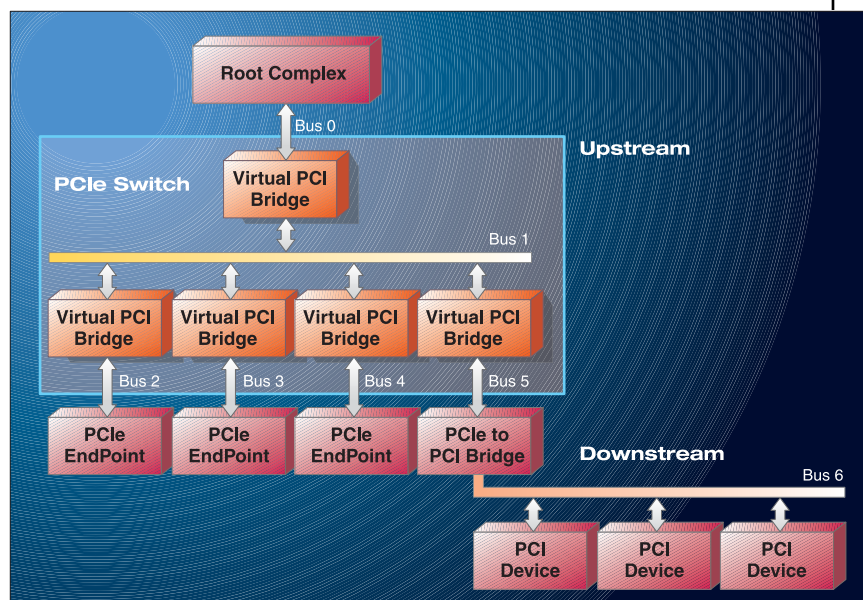
Il trasferimento dei dati avviene in modalità Burst, con meccanismi di accettazione della segnalazione. Esso si basa su un protocollo di ritardi e il target può forzare lo stato di wait e il retry. La comunicazione prevede un meccanismo di multiplexing basato su address e dato. La comunicazione è organizzata in transazioni di load/store, con tre diversi spazi di indirizzamento: quello della memoria, dell'I/O e lo spazio di indirizzi della configurazione. La transazione è di tipo gerarchico e avviene tra un master e un target. Il master inizia la transazione, il bus è gestito con un arbitraggio, e sono possibili transazioni di tipo peer-to-peer (ovvero alla pari).

### L'architettura di PCI

PCI è un bus che vive sull'architettura generica di PC, dove si trovano FSB/NorthBridge/South Bridge/AGP/Etc. e MCH/ICH.

Dall'esame di NB/SB/AGP/e così via, si evince la presenza di

**Fig. 4 - Topologia di uno switch PCI Express**



un collo di bottiglia rappresentato da PCI bus. Il north bridge pilota direttamente la grafica (AGP) e la memoria della CPU (DDR-SDRAM). Il south bridge invece arbitra l'accesso di dispositivi al PCI. Sia NB che SB funzionano su PCI.

Per quanto riguarda MCH e ICH, il primo pilota direttamente la memoria della CPU mentre il secondo connette PCI, IDE/ATA e USB. MCH e ICH sono Memory e I/O Hub link.

Questa architettura è raffigurata nella figura 1, dove è rappresentato uno schema del chipset di server Intel. Anche qui si comprende l'inadeguatezza di PCI in riferimento alla banda richiesta dai dispositivi collegati agli Hub di sistema.

La mappa di memoria di PCI è schematizzata in figura 2. Lo spazio degli indirizzi è piatto e organizzato in tre sottospazi, quello di memoria, di I/O e di configurazione. Il principio di mappatura della memoria prevede l'uso di aperture di indirizzi, e i registri BAR (Base Address Register) descrivono l'offset (e la dimensione) di questa apertura.

Mediante appositi bridge è possibile estendere il bus dal punto di vista elettrico ma non a livello di indirizzi. Questa estensione permessa dal Bridging PCI to PCI (anche noto come P2P) comporta un aggravio in termini di arbitraggio del bus e di latenza di una transazione.

È anche possibile effettuare operazioni di bridging non trasparente, in modo che due CPU poste per esempio su due processor card vedano il bridge come se fosse un end point. In pratica viene isolato lo spazio di memoria di due CPU sullo stesso PCI e consentito l'accesso alle aperture di memoria tra due CPU mediante meccanismi di handshaking (condivisione) basati su registri di doorbell (campanello, segnalazione).

### PCI enhanced: PCI-X

Per comprendere appieno l'evoluzione verso PCI-X è necessario fare alcune considerazioni. Innanzitutto il bus parallelo rappresenta un limite per il data rate, in quanto esistono limiti di larghezza massima del bus raggiunta e di dimensione fisica del bus che diminuisce esponenzialmente con l'aumentare del clock. Il bus condiviso consente di effettuare una transazione per volta e la transazione non è limitata, con conseguenti pro-

blemi di disconnessione del target in situazioni critiche: da qui la necessità di poter eseguire un nuovo tentativo (retry). L'introduzione di bridge multipli comporta un aumento della latenza nell'arbitraggio di accesso.

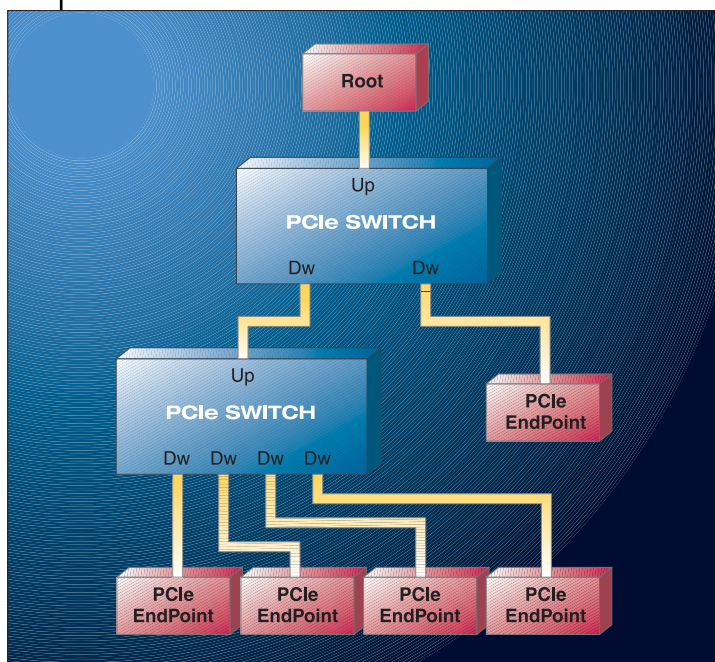
PCI-X è stato sviluppato in due versioni successive. La prima, PCI-X1.0, prevede una frequenza di 133 MHz e modifiche elettriche, il riposizionamento delle transazioni in ritardo, mediante lo slittamento del protocollo, il data transfer è specificato in dimensione, aumentando l'efficienza delle transazioni slittate. La seconda, PCI-X 2.0 prevede un aumento della frequenza (266 e 533 MHz) con clocking per DDR e QDR e comunicazione punto-punto. L'architettura passa da due Hub controller a un bridge Host/PCI-X.

### L'architettura PCI-Express

Dopo questa panoramica su PCI ed enhanced PCI-X è possibile spostare l'attenzione su il PCI Express: PCIe. Questo preserva il software scritto per PCI, possiede le medesime transazioni di load/store e la stessa mappatura della memoria: anche i 64 registri di configurazione sono mantenuti (anche se da 256B si passa a 4KB di registri di config).

I cambiamenti fondamentali rispetto a PCI si possono così riassumere:

- architettura organizzata in layer (Physical layer, Data Link Layer e Transaction Layer);
- il bus è caratterizzato da una comunicazione point to point, di tipo dual-simplex seriale. Il clock di riferimento è trasportato dal segnale serializzato e avviene una codifica 8B10B;
- la comunicazione basata su messaggi sostituisce i segnali out of band, e questo comporta l'introduzione di nuovo spazio degli indirizzi per i messaggi;



**Fig. 5 - Sviluppo dell'architettura PCIe**

- un protocollo di Requester/Completer richiede tutte le transazioni ripartite (split);
- introduzione di nuove caratteristiche in termini di QoS e di controllo di flusso;
- introduzione di funzionalità atte a garantire l'integrità, come CRC, ritrasmissioni, e così via.

In figura 3 è possibile osservare il cambiamento dell'architettura del chipset di server Intel con PCIe.

### PCIe Switch

Lo switch in particolare è un dispositivo molto importante nell'albero di un bus PCIe. Quest'ultimo riesce a raggiungere data rate molto elevati grazie al protocollo seriale punto-punto: in pratica si connette un dispositivo con un altro con un collegamento diretto, punto-punto. Il clock è trasportato dal segnale stesso e vengono eliminati problemi di crosstalk dovuti allo scaling, di skew e così via. Naturalmente in una motherboard, in un server, blade server o backplane è necessario poter realizzare questa comunicazione punto-punto mediante switch specializzati.

IDT ad esempio ha già a catalogo un certo numero di switch PCIe e ha annunciato la prossima introduzione di prodotti caratterizzati da prestazioni ancora più elevate.

Lo switch PCIe è un bridge virtuale multi porta tra PCI-PCI. Lo switch ha un unico upstream, orientato verso il root complex. Il numero di porte di downstream, verso gli end-point,

può essere ritenuto indefinito: il limite fissato a 256 è decisamente elevato per l'applicazione.

Ogni porta ha un unico PCI-PCI bridge, in grado di garantire la comunicazione punto-punto. PCIe prevede anche transazioni peer-to-peer, anche se l'architettura ad albero basata su unico root complex può fare ipotizzare diversamente. La gerarchia del bus è organizzata ad alberi, sicché ogni dispositivo è connesso a un dato root complex, compreso il root complex master del bus, e ogni link che parte da un root complex definisce uno specifico dominio gerarchico.

Si capisce come questa struttura elimini l'arbitraggio per il possesso del bus o link. L'arbitraggio si sposta a livello della trasmissione su una porta, e questo si traduce immediatamente in un aumento della banda.

La figura 4 mostra la topologia di uno switch PCI Express e in figura 5 è possibile osservare lo sviluppo dell'architettura PCIe. Si può notare una struttura ad albero, con un'unica radice che fa da master. La funzione del Root è quella di configurare il sistema, gestire gli interrupt e il processo degli errori. I root e gli end point condividono un singolo spazio di indirizzi comune. Lo scopo principale della comunicazione su questo bus è la lettura e scrittura di memoria e la gestione di interrupt (MSI e INTx).

### PCIe Bridge

Oltre agli switch esistono anche i Bridge, che però non devono essere confusi con i Bridge virtuali all'interno di uno PCIe Switch, i quali includono solo il transaction layer (o per meglio dire la capacità di routing).

Un Bridge (si tenga la definizione del Bridge PCIe include l'intero stack) connette un dispositivo non PCIe con un dominio PCIe. Per tale motivo avrà una porta verso un PCIe upstream verso il root complex e una o più porte verso uno o più downstream che non sono PCIe.

Un Bridge detto Reverse Bridge consente invece di connettere un dispositivo PCIe in una architettura PCI.

readerservice.it

|           |       |
|-----------|-------|
| Compaq    | n. 29 |
| Dell      | n. 30 |
| HP        | n. 31 |
| IBM       | n. 32 |
| IDT       | n. 33 |
| Intel     | n. 34 |
| Microsoft | n. 35 |