Multicast PCI Express: ottimizzare le applicazioni embedded e di comunicazione

Matt Jones
Product Marketing manager
Divisione Enterprise Computing
IDT

Il recente inserimento nelle specifiche PCIe di nuove capacità multicast mette a disposizione dei progettisti le prestazioni necessaria per ottimizzare l'uso delle risorse di sistema, garantendo prestazioni superiori grazie alla riduzione delle latenze di sistema

PCI Express (PCIe), ovvero l'interconnessione seriale che rappresenta un aggiornamento degli standard di bus PCI e PCI-X, è stato definito per fornire prestazioni scalabili superiori e "alleggerire" le problematiche di integrità del segnale e di layout di scheda che si sono manifestate con l'ampliamento del parallelismo dei bus.
L'esigenza di disporre di uno standard di interconnessione aggiornato si è manifestata in modo più acuto nel campo dei sistemi desktop e aziendali e nel campo delle applicazioni di memorizzazione: tale necessità ha spinto PCI Special Interest Group (PCI-SIG) e i fornitori di componenti a rivedere le specifiche e le prime soluzioni di sistema per soddisfare le esigenze emergenti di queste applicazioni.
Grazie a fattori di forma ottimizzati per soddisfare specifiche esigenze, i primi utilizzatori operanti nei mercati server e storage caratterizzati da elevati volumi hanno contribuito a una rapida e diffusa adozione di PCIe come standard "de facto" per la connettività chip-to-chip.
A dispetto della rapida proliferazione di PCIe e della sua ubiquità nelle applicazioni di elaborazione e memorizzazione, la diffusione di questo standard nel campo delle applicazioni embedded e di comunicazione è stata piuttosto limitata. Dal punto di vista storico, l'adozione di nuove tecnologie di interconnessione in queste applicazioni precede l'adozione in altri mercati dove i prodotti hanno tempi di sviluppo e cicli vita più lunghi. Nel caso della transizione a PCIe, i mercati embedded e delle comunicazioni hanno proceduto con lentezza in quanto PCI e PCI-X, prevalentemente utilizzati nel piano di controllo, hanno continuato a soddisfare le esigenze, in termini di prestazioni, richieste a livello di sistema. Oggi, però, i progetti di nuova generazione o sottoposti a operazioni di "refre-

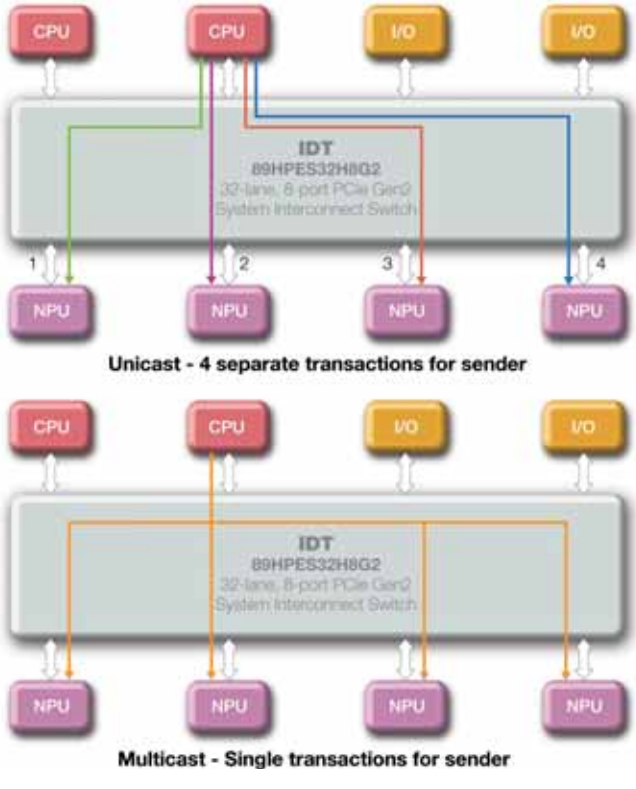

Fig. 1 - Riduzione del sovraccarico delle risorse nella trasmissione dati in multicast rispetto a una trasmissione Unicast in loop

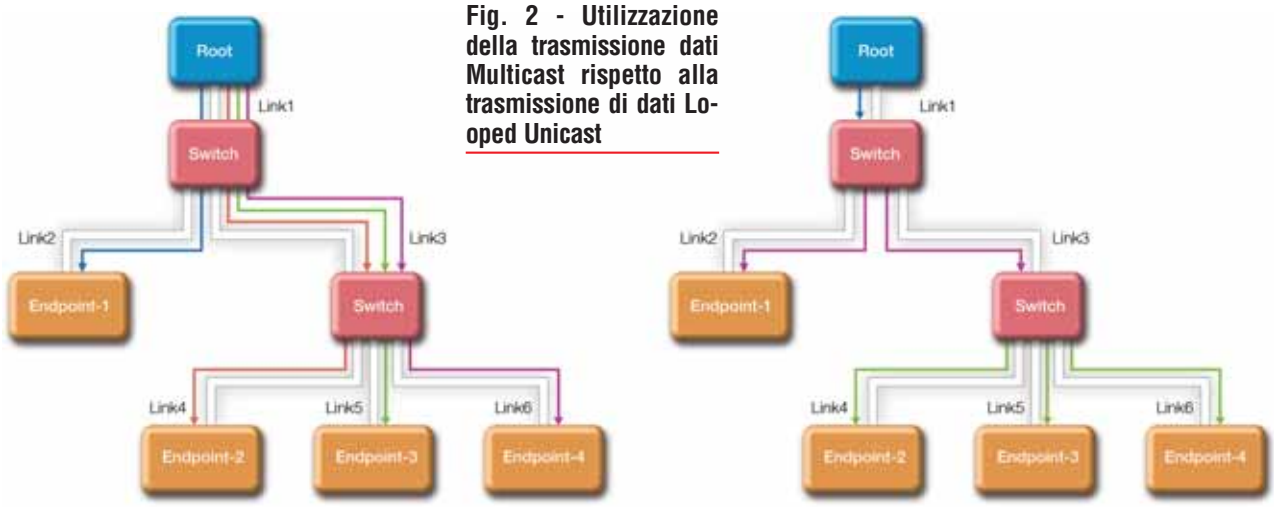

Fig. 2 - Utilizzazione della trasmissione dati Multicast rispetto alla trasmissione di dati Looped Unicast

sh" gravitano intorno a PCIe in quanto questo standard consentire di sfruttare l'ampio ecosistema formato da processori commerciali, periferiche e soluzioni di switching equipaggiati con PCIe come interfaccia nativa.
L'adozione di questo standard rimane però ancora limitata. Nell'ambito PCIe, i mercati legati alle applicazioni embedded e di comunicazione sono state in pratica "bypassate" ragion per cui le specifiche e i prodotti di supporto devono evolvere per indirizzare le nuove esigenze imposte da tali applicazioni.
Le specifiche PCIe definiscono una topologia basata su una struttura ad albero con un'unica fonte e una serie di diramazioni. Tale struttura è particolarmente indicata per garantire l'efficienza delle connessioni tra un singolo complesso di elaborazione e gli I/O locali associati. Una soluzione di questo tipo – perfetta per le applicazioni server e storage – non contempla o non si presta all'interconnettività dei sistemi multi-root. I sistemi di comunicazione e i sistemi embedded più avanzati spesso prevedono strutture distribuite di elaborazione e di intelligenza e – nel tempo – hanno adattato i costrutti PCI e PCI-X per garantirsi un adeguato supporto. L'adozione di PCIe come sistema primario di interconnessione richiede opportune estensioni per supportare i costrutti nati per ottimizzare l'uso delle risorse, per rendere efficiente la trasmissione e la condivisione dei dati e per garantire la coerenza tra "pari" nelle configurazioni multi-root.
Nell'ambito dell'ecosistema PCIe molto lavoro è stato fatto per estendere le specifiche PCIe in modo da soddisfare le esigenze delle applicazioni embedded e di comunicazione più impegnative. Questo lavoro sta proseguendo con uno sguardo critico sull'implementazione necessaria a garantire le estensioni desiderate senza imporre particolari oneri all'ampia base di utenti e senza imporre alcuna modifica all'attuale ecosistema e agli attuali modelli di impiego.
Nel maggio 2008, attraverso un "engineering change notice" (ECN) alla revision 2.0 delle specifiche base, il PCI-SIG ha aggiunto allo standard PCIe una serie di funzionalità multicast. In questo modo è stato possibile mettere a disposizione le prestazioni necessarie per il trasferimento e la condivisione dei dati tra più elementi di sistema distribuiti. Ciò ha permesso di eliminare un'importante barriera all'adozione di PCIe come principale sistema di interconnessione anche nelle applicazioni embedded e di comunicazione più impegnative. PCIe Multicast ottimizza le risorse di sistema e rende più efficiente la trasmissione dati tra più elementi di sistema, assicurando latenze ridotte e maggiori livelli di coerenza. Aspetto ancora più importante, l'implementazione di PCIe Multicast assicura questi significativi vantaggi sotto forma di semplice "estensione" delle specifiche PCIe esistenti, quindi senza aggravi e impatti sull'ecosistema o sui modelli già esistenti.

Vantaggi del Multicast

Con il termine Multicast si identifica un sistema utilizzato per inviare simultaneamente pacchetti dati quantificati a un gruppo di destinazioni: questo permette di gestire in modo più efficiente le risorse e la banda di sistema in quanto evita qualsiasi duplicazione superflua delle informazioni. In un sistema con intelligenza distribuita o replicata, caratteristiche tipiche delle applicazioni embedded e di comunicazione, le funzionalità multicast rappresentano un efficiente meccanismo di distribuzione "uno a molti" (one-to-many) ideale per l'espletamento di alcune task, come ad esempio l'invio simultaneo di comandi di boot o reset (necessari per ridurre la durata delle sequenze di reset e i tempi di fermo di sistema) o l'aggiornamento automatico delle informazioni di controllo dell'istradamento (necessario per assicurare la

coerenza dei dati di sistema). L'ottimizzazione delle risorse di sistema attraverso la riduzione dei sovraccarichi legati alla trasmissione dello stesso dato a più riceventi in multicast è illustrato in figura 1. In questo modello semplificato, uno switch PCIe senza multicast è messo a confronto con uno switch in grado di supportare il multicast. Il sistema/switch equipaggiato col multicast garantisce un uso più efficiente delle risorse sostituendo quattro transazioni sequenziali in loop con un'unica transazione multicast gestita direttamente dallo switch: questo approccio consente alla CPU di svolgere altri task più rapidamente. La maggiore efficienza può essere sfruttata per aumentare le prestazioni (le risorse di elaborazione di sistema possono infatti contare su task aggiuntive) o per ridurre costi e consumi (in quanto sono coinvolte meno risorse di elaborazione).
Oltre all'ottimizzazione delle risorse di sistema, il passaggio da più transazioni unicast in loop a un'unica transazione multicast riduce le latenze di consegna e di conseguenza aumenta il livello di coerenza tra i pari del sistema. Rivedendo il semplice modello di figura 1 e ipotizzando che la trasmissione sequenziale dei dati avvenga come indicato, il quarto endpoint non è allineato nei confronti del primo fino a quando tutte le iterazioni in corso non sono state completate. In un sistema single-root, questo gap è relativamente vantaggioso in quanto la maggior parte dei dati viene trasmesso dall'unico host e le quattro transazioni descritte verrebbero completate prima di poter intraprendere altre azioni. Invece, nei sistemi con intelligenza distribuita e traffico peer-to-peer significativo, il gap introdotto dalle transazioni iterative unicast origina potenziali problemi di ordinamento dei dati in quanto i pacchetti possono essere consegnati a ciascun endpoint in momenti relativamente diversi rispetto a quando l'endpoint riceve il pacchetto unicast in loop.

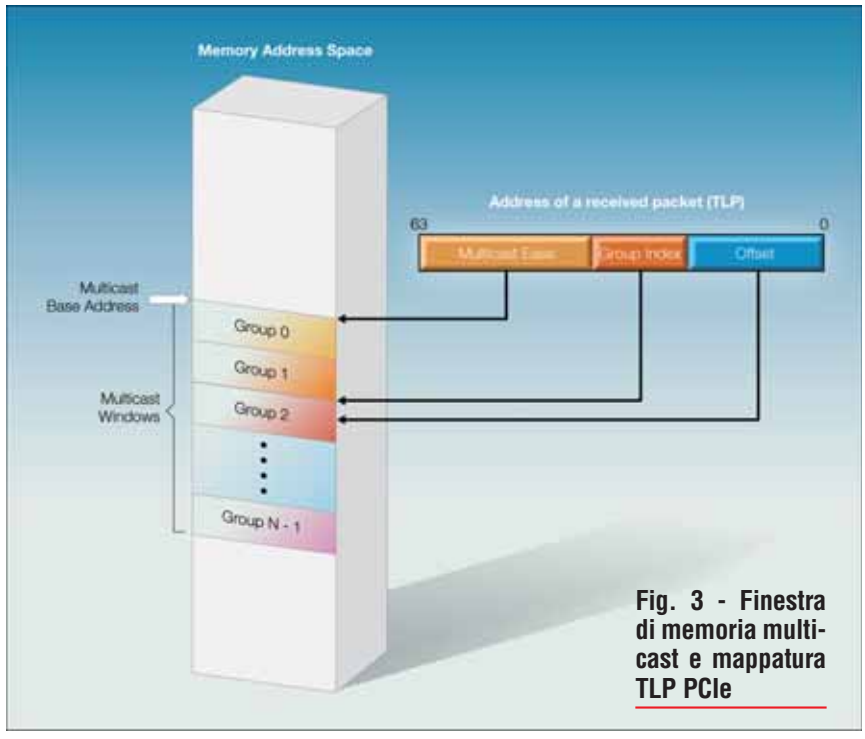

Fig. 3 - Finestra di memoria multicast e mappatura TLP PCIe

Si consideri uno scenario in cui le informazioni da inviare agli endpoint attraverso il loop siano necessarie per l'aggiornamento di una tabella di instradamento. Tali informazioni sono destinate a ciascuno delle quattro NPU (network processing unit) equipaggiati sulle linecard di elaborazione pacchetti del sistema di comunicazione. Gli aggiornamenti della tabella di ciascun endpoint sono soggetti a un incremento di latenza funzione dalla sequenza di consegna e completamento; i gap risultanti consentono alle schede di linea di "attendere" il loro aggiornamento continuando l'instradamento dei pacchetti ricevuti sulla base di una tabella obsoleta.
Il multicasting comporta inoltre un miglior utilizzo del link verso il sistema, spesso rimuovendo i colli di bottiglia o consentendo il ricorso a collegamenti più piccoli ma più efficienti che permettono di ridurre sia i consumi sia la complessità a livello progettuale. Il protocollo PCIe Multicast esegue delle copie del dato solo quando rileva dei "branch". La figura 2 descrive una struttura di interconnessione PCIe con salti multipli possibili attraverso la coppia di switch PCIe. Inviare in modalità unicast looped dati identici agli Endpoint 1, 2, 3 e 4 – come illustrato a sinistra - comporta che i Link 1 e 3 vengano attraversati più volte dagli stessi dati. Il sistema sulla destra sfrutta le capacità multicast di PCIe e, in tutti i casi in cui il dato deve essere copiato logicamente, lo trasmette solo una volta su ciascun link, lasciando la gestione alla funzionalità multicast.

Implementazione PCIe Multicast

Come osservato in precedenza, la rapida e diffusa adozione dello standard PCIe ha favorito la nascita di un esaustivo ecosistema e lo sviluppo di una vasta base di utenti che spesso sfruttano le precedenti versioni delle specifiche.
Pertanto PCIe Multicast, come qualsiasi estensione delle specifiche, non deve comportare impatti negativi oppure oneri aggiuntivi per l'ecosistema esistente o per i modelli in uso. A tale proposito, PCIe Multicast è stato definito per non richiedere modifiche hardware agli attuali apparati né ai formati TLP (transaction layer packet).
Per ottimizzare le funzionalità rispettando tali vincoli, PCIe Multicast è stato definito come funzionalità multicast address-based che utilizza un segmento dello spazio di memoria comune PCIe e un semplice modello di programmazione per l'instradamento dei normali TLP PCIe verso più destinatari, con un massimo di 64 gruppi multicast (MCG).
Benché un MCG possa contenere anche 0 o un solo membro, i sistemi usufrui-

scono dei vantaggi solo a partire da dimensioni di MCG superiori a 2: ciò comporta l'utilizzo di uno switch PCIe (o di una serie di switch) per fornire la connettività tra "l'iniziatore" e gli MCG. Il traffico multicast può essere avviato da qualsiasi dispositivo della gerarchia PCIe e trasmesso a qualsiasi numero di partecipanti collegati a una porta dello switch dotato di una "Multicast Capability Structure". Un interruttore equipaggiato con una Multicast Capability Structure per ogni porta è in grado di inviare in multicasting pacchetti da una qualsiasi delle sue porte verso qualsiasi altra delle sue porte. Gli apparati di origine e terminali possono trarre dei vantaggi dalla presenza di una Multicast Capability Structure ma, come osservato in precedenza, per i dispositivi questa è solo una funzione opzionale.
Successivamente all'enumerazione del sistema, che non viene influenzato dalla presenza di funzioni di multicasting, il software di sistema configura lo spazio di indirizzamento multicast aprendo una finestra multicast nello spazio di memoria PCIe a partire dal Multicast Window Base Address, come illustrato in figura 3. La finestra multicast è configurata come serie continua di indirizzi che parte dall'indirizzo base e che può essere suddivisa in sub-range di eguali dimensioni per ciascuno dei 64 MCG supportati. Non esistono limitazioni pratiche alle dimensioni della finestra multicast (per esempio la finestra multicast può avere fino a 2^63 byte). Il numero di gruppi nella finestra di multicast può variare da 1 a 64.
Ai dispositivi PCIe nel sistema che supportano il multicast è richiesto la presenza di una Multicast Capability Structure per ciascuna funzione PCIe abilitata al multicast. In uno switch PCIe, questo implica che ciascuna porta dello switch che trasmetta o riceva dati multicast debba supportare una Multicast Capability Structure nel ponte (bridge) virtuale PCI-to-PCI (P2P) associato. Nell'ambito di ciascuna Multicast Capability Structure, appositi registri di controllo configurati in modo identico mantengono le seguenti informazioni: indirizzo base finestra di multicast, numero di MCG e dimensioni finestra MCG. Oltre a questo, ciascuna Multicast Capability Structure possiede un vettore di controllo da 64 bit configurato in modo indipendente che abilita o disabilita il ricevimento dei TLP dagli MCG da 0 a 63 e che quindi governa l'appartenenza di ciascuna funzione PCIe a ogni singolo MCG. I registri e i bit di controllo sono leggibili e scrivibili in qualsiasi momento durante il funzionamento del dispositivo. Trasmissione e routing dei TLP PCIe Multicast variano notevolmente rispetto a quelli dei TLP unicast. Queste differenze possono essere meglio apprezzate con l'aiuto dello schema funzionale base di uno switch PCIe riportato in figura 4. Una volte ricevuto un TLP dall'origine, la decodifica dell'indirizzo alla porta di ingresso determina che il TLP è di tipo multicast (a livello logico, questo è il punto dove la transazione iniziale diventa transazione multipla). I TLP multicast decodificati senza errori sono inoltrati al bus PCI virtuale dello switch. A differenza del traffico unicast, che ha regole di instradamento differenti in funzione del fatto che il TLP sia stato ricevuto sul lato primario o secondario del bridge P2P, i TLP multicast sono inoltrati in modo simmetrico indipendentemente dal fatto che il bridge P2P sia associato a una porta in upstream o in downstream.
Tutte le porte dello switch (per esempio le funzioni bridge P2P) connesse al bus PCI virtuale ricevono il TLP multicast ed esaminano l'ID MCG nell'indirizzo in funzione dello stato del bit nel vettore multicast-receive-enable. Quest'ultimo indica, sulla base del singolo MCG, se a una funzione bridge P2P è consentito l'inoltro del TLP verso la sua destinazione. Questo consente a ciascuna funzione bridge P2P nello switch di registrarsi come ricevente di TLP multicast su base per-gruppo.
Una volta che la funzione bridge P2P all'interno dello switch accetta un TLP multicast, essa esegue il processo di uscita sul TLP. L'elaborazione di uscita di TLP multicast varierà in base alle funzionalità del partner di link sulla porta switch associa al bridge P2P. Nel caso in

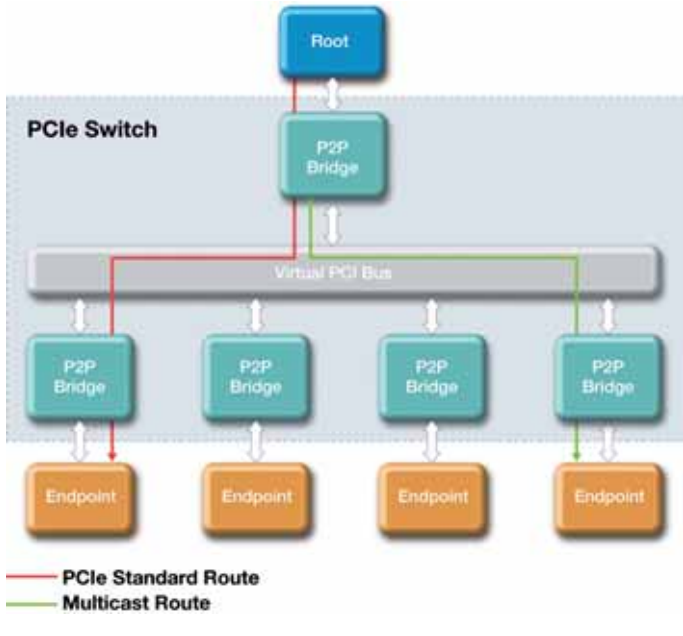

Fig. 4 - Esempio di Multicast PCIe-vista funzionale dello switch

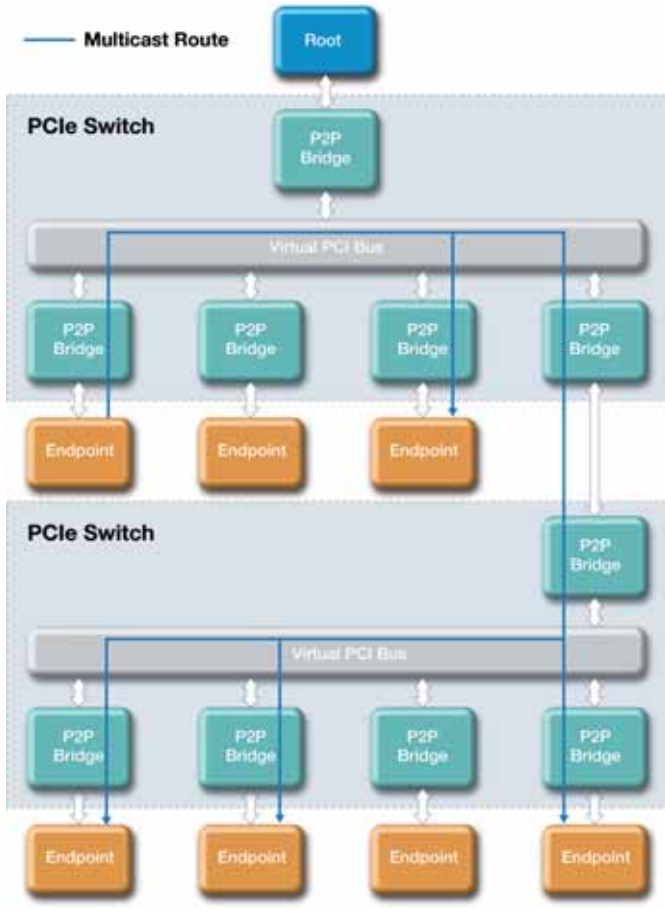

Fig. 5 - Esempio di Multicast PCIe – Instradamento nei sistemi multiswitch

cui il partner di link sia dotato di una struttura con capacità multicast, come la trasmissione tra switch PCIe riportata in figura 5, il TLP viene inoltrato per un ulteriore instradamento senza modifiche. Come notato in precedenza, per ricevere i TLP multicast con terminali PCIe non è richiesta l'implementazione di strutture con capacità multicast. Per supportare endpoint senza capacità multicast, il software di sistema deve garantire che i registri di indirizzo base dell'endpoint vadano a sovrapporsi ad alcune porzioni del range di indirizzi multicast o che lo switch PCIe impieghi un meccanismo di sovrapposizione multicast specificato opzionalmente.
Poiché il compito di garantire la sovrapposizione tra i range di indirizzi degli endpoint e il range di indirizzi multicast pone dei vincoli al progettista del sistema e può richiedere delle basi di codice uniche per ciascun SKU di prodotto, le implementazioni più efficaci di switch prevedono l'integrazione del meccanismo di overlay all'interno della Multicast Capability Structure di ogni porta dello switch, assicurando la massima flessibilità e consentendo di sfruttare gli endpoint già disponibili. La funzionalità di sovrapposizione degli indirizzi – come illustrato in figura 6 – è un meccanismo che può essere utilizzato per rimappare l'indirizzo di un TLP multicast ricevuto dalla finestra multi cast, alla finestra del registro indirizzi base (BAR) dell'endpoint. La sovrapposizione degli indirizzi viene eseguita dalle porte dello switch. Ciascuna porta dello switch può essere configurata con un valore di sovrapposizione differente per consentire la mappatura indipendente all'interno della finestra BAR associata a ciascun endpoint. Viene supportata anche la conversione tra indirizzi a 32-bit e indirizzi a 64-bit (per esempio, la regione multicast può essere posizionata al di sopra il limite dei 4GB e il BAR dell'endpoint può essere al di sotto del limite dei 4GB, o viceversa). In definitiva il recente inserimento nelle specifiche PCIe di nuove capacità multicast mette a disposizione dei progettisti le prestazioni necessaria per ottimizzare l'uso delle risorse di sistema, garantendo prestazioni superiori grazie alla riduzione delle latenze di sistema. A questo si aggiunge una trasmissione dati più efficiente e coerente tra pari nei sistemi embedded multi-root e nei sistemi di comunicazione.
Avendo avuto l'accortezza di non creare aggravi all'ecosistema PCIe esistente o all'attuale base utenti, le nuove funzionalità non richiedono variazioni all'hardware originale e non richiedono nuovi formati TLP. Grazie alla semplicità del suo modello di programmazione l'implementazione PCIe Multicast address-based può essere integrata all'interno dello switch PCIe, offrendo livelli di funzionalità e flessibilità ben superiori alle precedenti implementazioni proprietarie, come ad esempio il dualcasting.

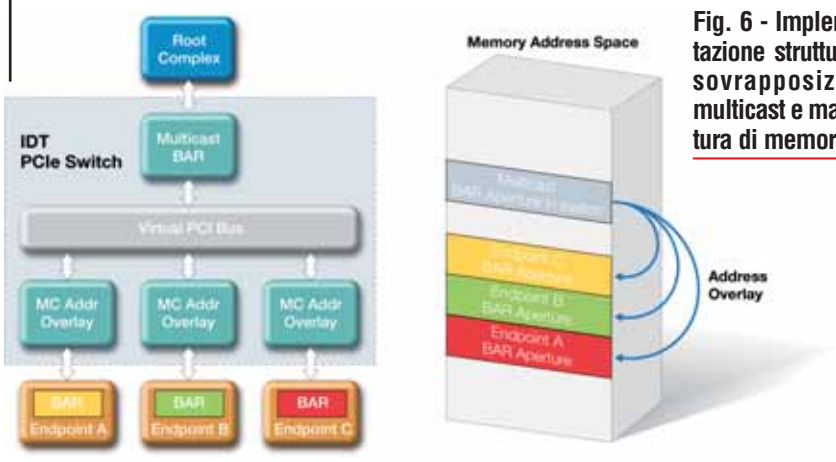

Fig. 6 - Implementazione struttura di sovrapposizione multicast e mappatura di memoria

IDT
readerservice.it n.2