Architetture multihost con switch Pci Express

Gli switch PCIe di IDT consentono di adottare l'architettura multi-root ripartibile per risolvere i complessi problemi di sistema che in precedenza impedivano l'adozione di PCIe come standard di interconnessione primario nei sistemi multi-root.

di Matt Jones

Sul mercato sta emergendo un'architettura di switching totalmente nuova che permette di risolvere, a livello di sistema, alcuni problemi chiave che in precedenza impedivano l'adozione dello standard Pci Express come soluzione primaria di interconnessione nelle comunicazioni AdvancedTca, nelle comunicazioni basate su sistemi "blade" proprietari e nelle applicazioni embedded e server/storage. Benchè in queste applicazioni PCIe sia uno standard di fatto, nelle interconnessioni locali e chip-to-chip la sua portata come soluzione di interconnessione primaria per sistemi a intelligenza distribuita è sempre stata limitata a causa dei vincoli legati allo sviluppo delle architetture multi-root e all'efficienza di gestione delle risorse di sistema. Per quanto complesse siano le barriere, la soluzione proposta, un'architettura switch Pci multi-root ripartibile, permette di "spezzare" le varie task del protocollo e dello switching PCIe negli elementi base, ma anche di fare leva su transazioni più semplici e su istanze multiple, arrivando così a una soluzione di switching PCIe degna della nuova classificazione.

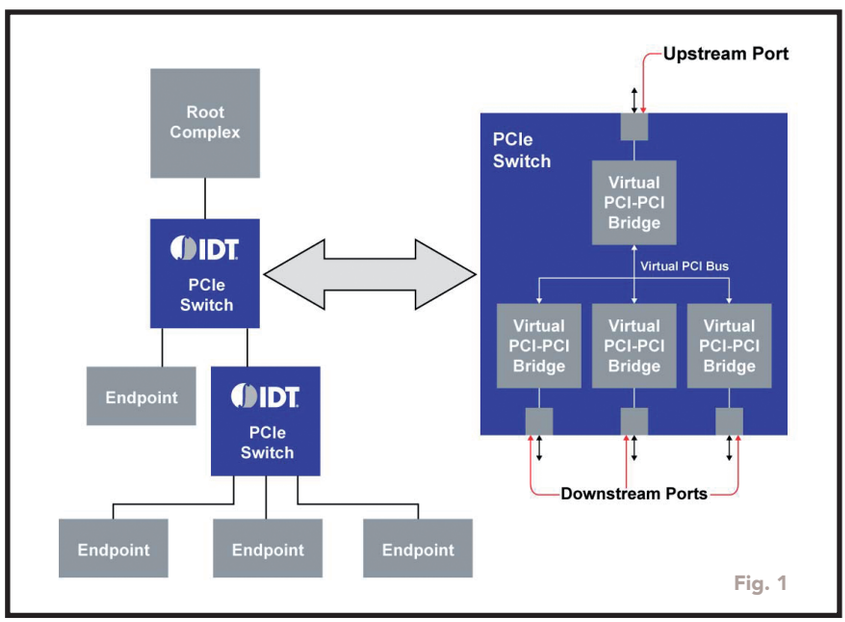

Fig. 1

Topologia sistema PCIe e panoramica sullo switching

Lo standard d'interconnessione seriale PCIe, largamente diffuso per i vantaggi che offre in termini di efficienza, scalabilità, consumo e costo rispetto agli standard concorrenti, è costruito sulla base del precedente standard Pci per garantire la compatibilità con i codici firmware e software di sistema esistenti. Benchè la precedente topologia bus-based di Pci e di Pci-X sia stata sostituita da una soluzione di connettività punto-punto che utilizza, per la distribuzione, degli switch a pacchetto, il risultato è sempre una semplice struttura ad albero con un'unica root complex (nel più dei casi una Cpu o un processore complesso) come riportato nella Fig. 1. La root complex è responsabile della configurazione di sistema e della classificazione delle risorse PCIe, ma anche della gestione di interrupt ed errori per l'intero albero PCIe. Per potenziare ulteriormente la semplicità e i costrutti originari, la root complex e i suoi endpoint condividono un unico spazio di indirizzamento, comunicando attraverso letture e scritture di memoria e interrupt. Internamente, i dispositivi PCIe implementano bus e bridge Pci-to-Pci virtuali e strutture logiche che

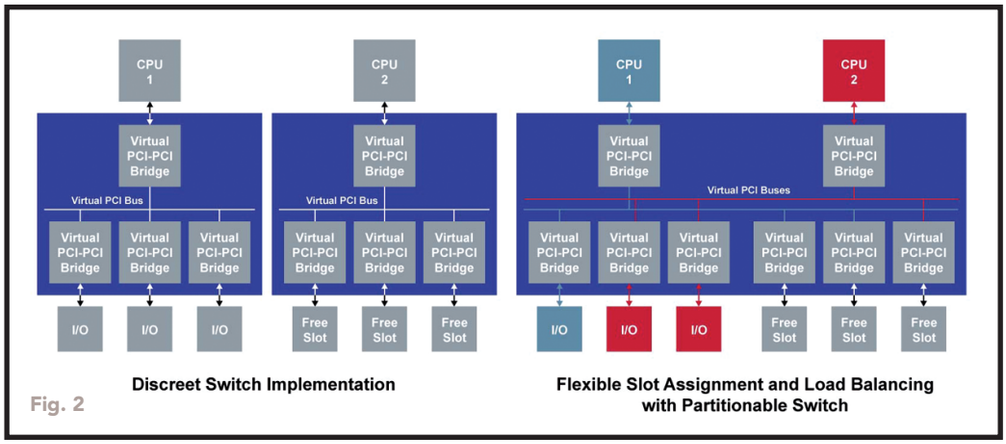

Fig. 2

ricalcano le funzioni di bus e di bridging fisico dei precedenti sistemi Pci, oltre ad assicurare il supporto dei messaggi e degli interrupt dei protocolli precedenti. Lo switch PCIe illustrato nella Fig.1 offre il dettaglio logico dell'implementazione di queste gerarchie bus e bridge virtuali. Replicando le strutture logiche Pci-based all'interno dei dispositivi PCIe è stato possibile facilitare la migrazione hardware e software tra standard ma ha lasciato al nuovo standard un livello limitato di estensibilità, in particolare per quanto riguarda le applicazioni multi-root, in quanto il protocollo è rimasto focalizzato sull'efficienza d'interconnessione per costrutti single root, con una relazione uno-a-uno tra una sorgente root e un endpoint.

Le problematiche del PCIe nelle interconnessioni di sistema multi-root

Per utilizzare PCIe come sistema principale di interconnessione nelle soluzioni multi-root, la sfida è stata se lavorare all'interno delle specifiche dello standard o estenderle per permettere agli architetti di sistema di sfruttare le doti di efficienza, scalabilità e basso consumo del PCIe e per capitalizzare la ricchezza e la forte crescita dell'ecosistema di processori e periferiche. Queste problematiche sono state analizzate e sono state proposte delle soluzioni che contemplano la possibilità di lavorare all'interno dello standard attraverso dei costrutti Ntb *(Non-transparent bridging)* o delle estensioni mr-IOV *(multi-root I/O virtualization)*. Entrambi gli approcci hanno dei vantaggi ma fondamentalmente sono gravate dal fatto di non utilizzare l'attuale ecosistema e di richiedere lo sviluppo di software proprietario. La soluzione di fornire supporto multi-root e di garantire un uso efficiente e una condivisione delle risorse di sistema all'interno delle specifiche PCIe può essere più semplice rispetto alle soluzioni offerte in precedenza. Rivisitando la classica definizione di "emergente", un nuovo approccio suggerisce che i problemi possono essere indirizzati all'interno dello standard e dell'ecosistema correnti focalizzandosi sugli elementi PCIe più di base, mantenendo le transazioni semplici e sviluppando delle soluzioni di switching che, attraverso la molteplicità, si combinano per risolvere problemi di switching più grandi a livello di sistema.

Architetture di switch multi-root ripartite

Facendo leva sugli elementi logici costituenti uno switch PCIe, i bridge Pci-to-Pci virtuali e il bus Pci virtuale, l'architettura di switch multi-root ripartibile offre dei controlli fisici per creare degli switch logici multipli o delle partizioni switch all'interno di un singolo dispositivo di switching, come illustrato nella Fig. 2. Ciascuna delle partizioni switch risultanti è logicamente discreta e aderisce alle specifiche Pci Express Base 2.0. Ciascuna ripartizione indipendente rappresenta una gerarchia PCIe la cui configurazione, le cui operazioni di switching e la cui logica di reset sono isolati dalle altre ripartizioni. Attraverso la replica delle strutture di gestione e controllo associate al bus PCIe virtuale è possibile supportare più bus Pci virtuali assicurando la coesistenza in un unico switch conforme PCIe di più root complex distinti. Ulteriori elementi di flessibilità architetturale sono garantiti dalla capacità di associare liberamente dei bridge Pci-to-Pci virtuali a qualsiasi bus virtuale costituito sia staticamente nel momento di un reset fondamentale dello switch sia dinamicamente mentre lo switch sta lavorando. L'architettura switch multi-root ripartibile consente inoltre ai progettisti di sistemi di ripartire uno switch fisico da n-porte su n-partizioni (ad esempio 16 partizioni per uno switch da 16 porte) e di fare leva sulla possibilità di assegnare flessibilmente qualsiasi porta switch a qualsiasi delle partizioni oltre che di cambiare la configurazione del sistema e l'assegnazione delle porte dinamicamente durante le operazioni di commutazione. Inoltre, nell'ambito di una determinata partizione, è possibile configurare qualsiasi porta come porta di upstream o di origine ed è possibile spostare tale porta di origine anche durante le attività di switching. A dispetto dell'indipendenza delle partizioni la logica di controllo condivisa dalle partizioni all'interno dello switch rimane una risorsa globale che può essere controllata attraverso il SMBus (System Management Bus) degli switch fisici o in banda da qualsiasi delle origini collegate a qualsiasi delle partizioni logiche. Ciò aumenta la flessibilità dell'architettura in quanto consente la riallocazione dinamica delle risorse inizializzate dall'origine nell'ambito o all'esterno di una determinata partizione. L'architettura switch supporta le architetture di sistema multi-root, consente l'applicazione avanzata di funzionalità che aumentano la configurabilità del sistema e permette di ottimizzare l'utilizzo delle risorse, la loro disponibilità e la loro sicurezza.

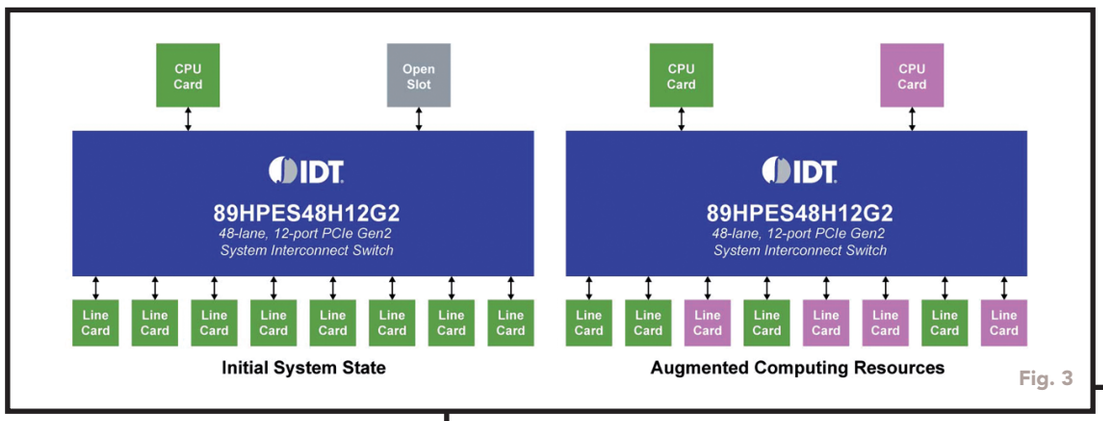

Fig. 3

Configurabilità di sistema e livello di riutilizzo hardware

L'applicazione più diretta dell'architettura per aiutare la configurabilità del sistema è di sostituire più switch PCIe fisici discreti con un unico switch ripartito. Questo offre dei notevoli benefici in termini di costi totali di ownership grazie alla riduzione dei consumi di energia, degli ingombri su scheda e degli oneri legati alle interconnessioni di sistema. Oltre a questo, un complesso di switching unico riduce i costi di sviluppo del sistema attraverso la possibilità di riutilizzare l'hardware: in tal caso un'unica piattaforma è in grado di servire più mercati finali e di posizionarsi su più livelli di prezzo/prestazioni.

Riassumendo, un sistema multi-root con un set fisso di risorse di elaborazione in grado di fare leva su un'architettura di commutazione ripartibile permette di dare vita a un ampio spettro di sistemi grazie alla capacità di mappare in modo efficiente le risorse di elaborazione su un numero variabile e su un ampio spettro di dispositivi o di schede periferici. Un esempio di tale flessibilità è proposto nella Fig. 3, dove si mette in paragone un'implementazione switch discreta. Una seconda soluzione è di utilizzare lo switch ripartibile per aumentare le risorse di elaborazione di un set fisso di periferiche e di I/O per assicurare un aumento di prestazioni globali di sistema riducendo il rapporto tra periferiche di I/O e risorse di elaborazione. La Fig. 4 offre un esempio basato su un'applicazione multi-card equipaggiata con una scheda di elaborazione aggiuntiva per ottenere migliori prestazioni.

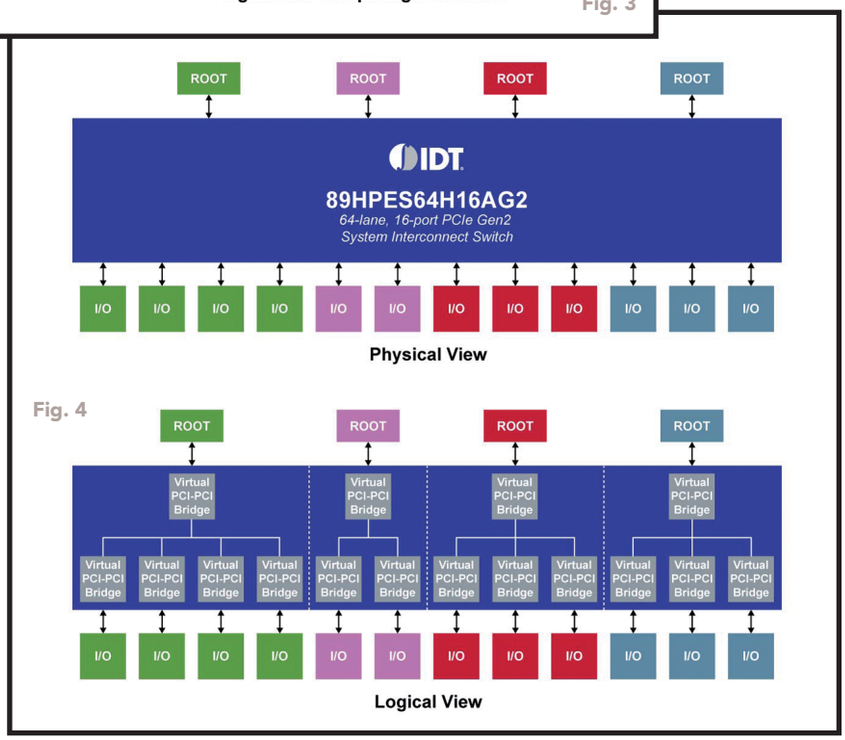

Fig. 4

Utilizzo ottimizzato delle risorse e gestione QoS

Oltre alla flessibilità delle configurazioni statiche sottolineata sopra, l'architettura multi-root ripartibile consente la configurazione dinamica durante le operazioni di commutazione. Questa capacità permette di gestire agilmente gli elementi periferici, di elaborazione e di I/O ottimizzandone l'impiego attraverso la condivisione delle risorse a livello di sistema. Oltre a ridurre i costi di sistema massimizzando l'uso delle risorse sotto forma di allocazione ottimizzata degli elementi di elaborazione e di I/O, lo switch consente la gestione diretta dei livelli di QoS (Quality of service) di sistema e garantisce gli adeguati livelli di Sla (service-level agreements) per gli utenti e il traffico chiave. La Fig. 5 offre un esempio base di un evento di riconfigurazione di sistema in cui le risorse globali sono ridistribuite mentre l'elemento di origine più a sinistra ha scaricato le periferiche per assicurare il massimo della banda al traffico o agli utenti ad alta priorità. L'allocazione dinamica delle risorse di commutazione che l'architettura supporta pone ulteriore enfasi

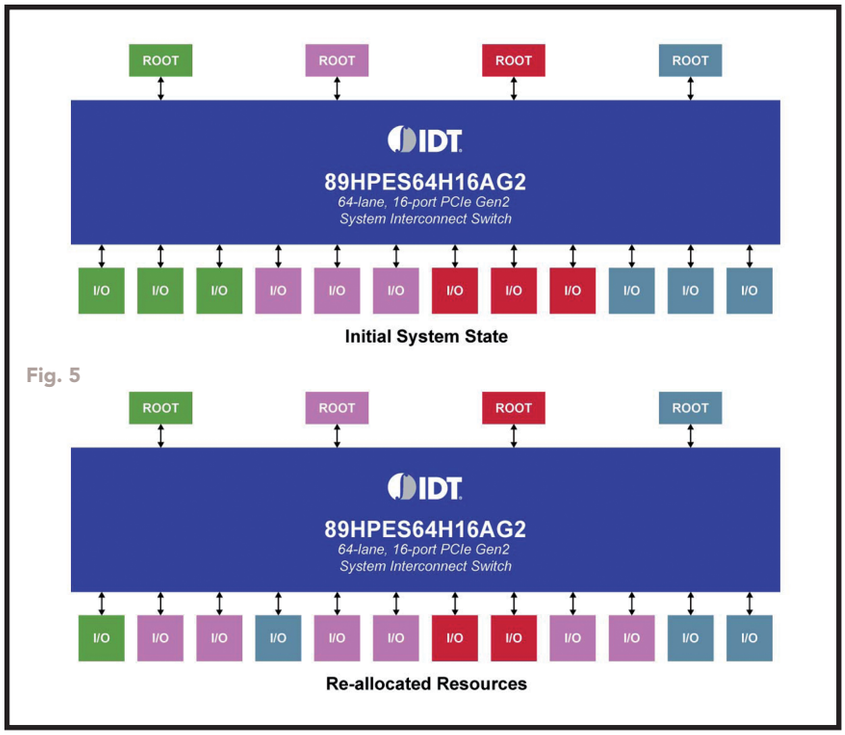

Fig. 5

Il supporto avanzato per il failover nelle architetture multi-root ripartite offre livelli di disponibilità delle risorse di affidabilità di sistema nettamente superiori rispetto a quelli offerti in precedenza dall'ecosistema di switching PCIe. I precedenti approcci a questo problema offrivano la possibilità di un failover basato su un doppio host attraverso algoritmi proprietari Ntb-based. Benchè sia efficace per i sistemi dual-root, questo metodo è ormai inadeguato per i sistemi multi-root in quanto aumenta gli oneri e i costi di progetto a causa della complessità del software richiesto per le funzionalità Ntb proprietarie. In caso di failover e per supportare le capacità di allocazione dinamica delle risorse, le architetture multi-root ripartibili qui illustrate utilizzano dispositivi di elaborazione e periferici standard commerciali basati su PCIe e sfruttano firmware e software esistenti.

sui due principi di configurazione statica illustrati sopra in quanto le periferiche o le schede di sistema aggiunte o rimosse a caldo possono essere ridistribuite dinamicamente senza dover mettere offline anche gli elementi non coinvolti.

Incremento della disponibilità delle risorse

Un'importante estensione dell'allocazione dinamica delle risorse riguarda la gestione avanzata del failover, cioè la sua applicazione per incrementare la disponibilità e l'affidabilità delle risorse stesse. Con le architetture multi-root ripartibili, le risorse associate a un'origine guasta possono essere dinamicamente riassegnate a origini funzionanti, come illustrato in Fig 6. L'architettura consente a qualsiasi numero di origini attive rimanenti di prendere il controllo delle risorse di sistema isolate per ristabilire il servizio con interruzioni e perdite di dati minime. La flessibilità architetturale garantisce ai progettisti di sistema la possibilità di scegliere tra innumerevoli strategie di failover, partendo dalla più semplice riassegnazione a un'origine predeterminata fino ad arrivare a soluzioni più eleganti che prevedono la riassegnazione basata sullo stato corrente dello switch e sul carico corrente delle origini attive.

Matt Jones
Product Marketing Manager
Enterprise Computing
IDT
www.idt.com

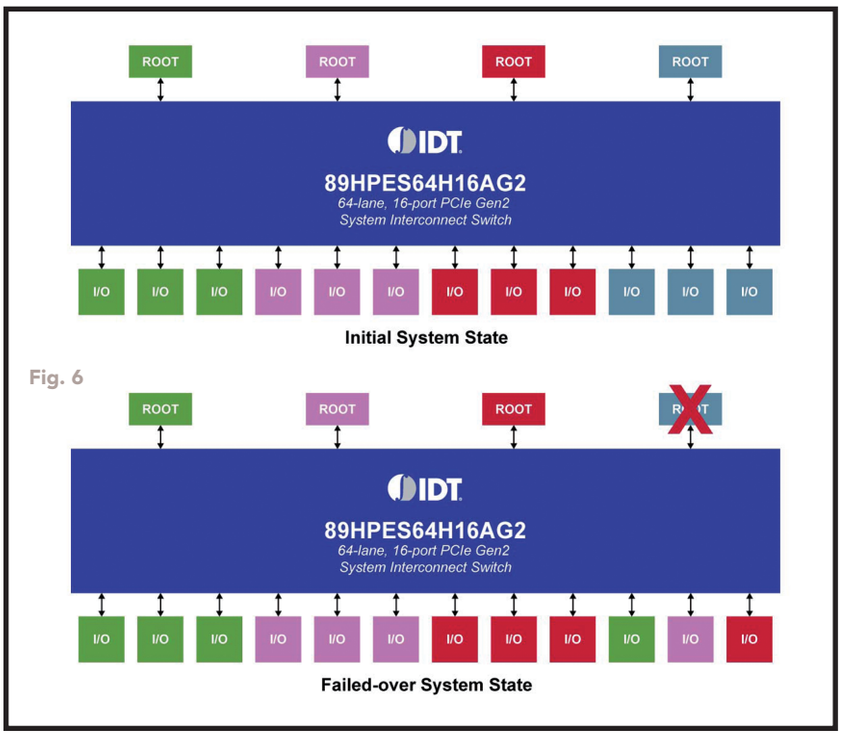

Fig. 6