Artikel online:
www.el-info.de

Verkehrsfluss und Redundanz in Multiprozessor-PCIe-Backplane-Systemen

Ian Dobson, IDT

Viele Embedded- und Kommunikationsgeräte sind Chassis-basierte Systeme mit einer Mid- oder Backplane für Einsteckkarten. Diese Steckkarten enthalten oft Mikroprozessoren für die lokale Steuerung, die unter der Koordination eines zentralen Steuerelements arbeiten. Hot-Swap und hohe Systemverfügbarkeit erfordern zusätzlich den Einsatz redundanter Elemente auf Systemebene. Diese Vorgaben sind eine Herausforderung für ein PC-orientiertes Protokoll wie PCI Express.

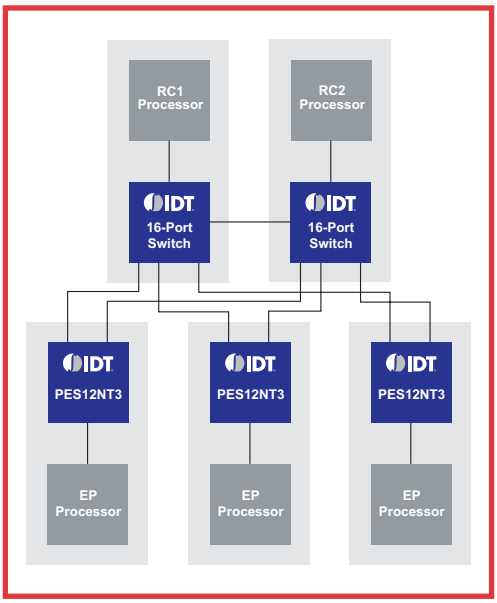

Bild 1. PCIe-Konfiguration für Upstream-Redundanz

Die Steuerungsebene einer Kommunikationsplattform besteht in der Regel aus einem Paar redundanter Systemsteuerungs-Karten, die untereinander über eine Direktschnittstelle verbunden sind und die mit den Line-Cards im restlichen System über eine Doppelstern-Konfiguration kommunizieren. In der Regel läuft der bidirektionale Datenverkehr zwischen der aktiven Steuerungskarte und den einzelnen Line-Cards.

Viele Plattformen müssen so ausgelegt sein, dass sie eine Mischung unterschiedlicher Line-Cards unterstützen können, von denen manche mit Embedded-CPUs ausgestattet sind und andere nicht. Dies und die Forderung, zahlreiche Bandbreiten-Optionen für unterschiedliche Line-Card-Kapazitäten kostengünstig bereitzustellen, prädestiniert PCI Express (PCIe) als Lösung für den Datenverkehr auf der Steuerungsebene bei Kommunikationstechnik-Anwendungen.

Dazu sollen zunächst die wichtigsten Merkmale von PCIe vorgestellt werden. Anschließend folgen eine Beschreibung der Probleme, die beim Aufbau einer Kommunikations-Plattform entstehen, sowie Überlegungen zur Lösung dieser Probleme.

Funktionen von PCIe für Kommunikationssysteme

Skalierbarkeit ist ein wesentliches Merkmal von PCIe, das es zu einer geeigneten Lösung für die systemweite Verbindung auf Steuerungsebene macht. Eine Kernfunktion ist die Unterstützung einer automatischen Linkbreiten-Verhandlung bei PCIe-Verbindungen. Diese Fähigkeit ist besonders nützlich für die Unterstützung verschiedener Plug-in-Module, die unterschiedliche Link-Breiten benötigen. Allerdings schreibt die PCIe-Spezifikation lediglich vor, dass ein Gerät für volle Busbreite oder eine einfache (x1) Breite verhandeln muss. Dazwischen liegende Breiten sind optional.

Auf einer physikalischen Ebene entspricht jede Verbindung innerhalb einer Lane einem unabhängigen Datenstrom. Das bedeutet, dass es keine Grenzwerte für Skew von einer Lane zur nächsten gibt, selbst wenn sie alle mit einer gemeinsamen Frequenz getaktet werden. Außerdem kann jeder Link mit seinem Link-Partner darüber verhandeln, ob im PCIe Generation 1 Modus bei 2,5 GBit/s oder im Generation 2 Modus bei 5 GBit/s übertragen werden soll.

Eine entscheidende Funktionen in Version 2.0 der PCIe-Spezifikation ist die Fähigkeit, die Übertragungsbandbreite eines Links im aktiven Betrieb dynamisch anzupassen. Befindet sich ein Link im Leerlauf, dann einigen sich die Datenlink-Ebenen an beiden Enden auf eine Verringerung der Link-Breite. Entsprechend können sie sich später darauf einigen, diese Breite wiederherzustellen, sobald der Datenverkehr zunimmt. So lassen sich Bandbreite und Stromaufnahme an die aktuelle Last auf dem Link anpassen.

Verbesserte Systemverfügbarkeit

Zur Verbesserung der Verfügbarkeit eines Kommunikationssystems kann der Entwickler eine Reihe von Funktionen auf der physikalischen bzw. der Datenlink-Ebene des PCIe-Protokolls verwenden. PCIe-Datenpakete werden auf ihrem Weg zwischen zwei Einheiten durch CRC (Cyclic Redundancy Check) auf der Link-Ebene sowie durch eine Sequenz-Nummer geschützt. Außerdem bietet der Bus eine optionale, die gesamte Strecke umfassende CRC-Funktion, um zusätzlichen Schutz für die Integrität der Daten zu bieten.

Jeder PCIe-Link, der aus mehreren PCIe-Lanes besteht (zum Beispiel ein Link mit x8 Breite), unterstützt zudem eine automatische Lane-Umkehr. Mit dieser Funktion kann der Benutzer festlegen, ob sich Lane 0 auf der logischen linken oder rechten Seite des Links befindet. Dadurch lässt sich eine „Überkreuzung" von Leiterbahnen beim Entwickeln der Leiterplatte verhindern. Dies ist bei den in der Kommunikationstechnik verbreiteten modularen Systemen besonders wichtig, weil dadurch optimale Layouts bei individuellen Modulen möglich werden. Als zusätzliche Option unterstützt PCIe auch eine automatische Polaritäts-Umkehr.

Der gemeinsame Einsatz der automatischen Linkbreiten-Verhandlungsfunktion zusammen mit einer automatischen Lane-Umkehr ermöglicht einen kontinuierlichen Systembetrieb selbst bei Ausfall einer Lane innerhalb eines Links. Der Link lässt sich zurücksetzen und kann eine automatische Verhandlung durchführen, um die nutzbare Hälfte des Links weiter zu verwenden. Auch wenn diese Funktion nur die Hälfte der Bandbreite nutzt,

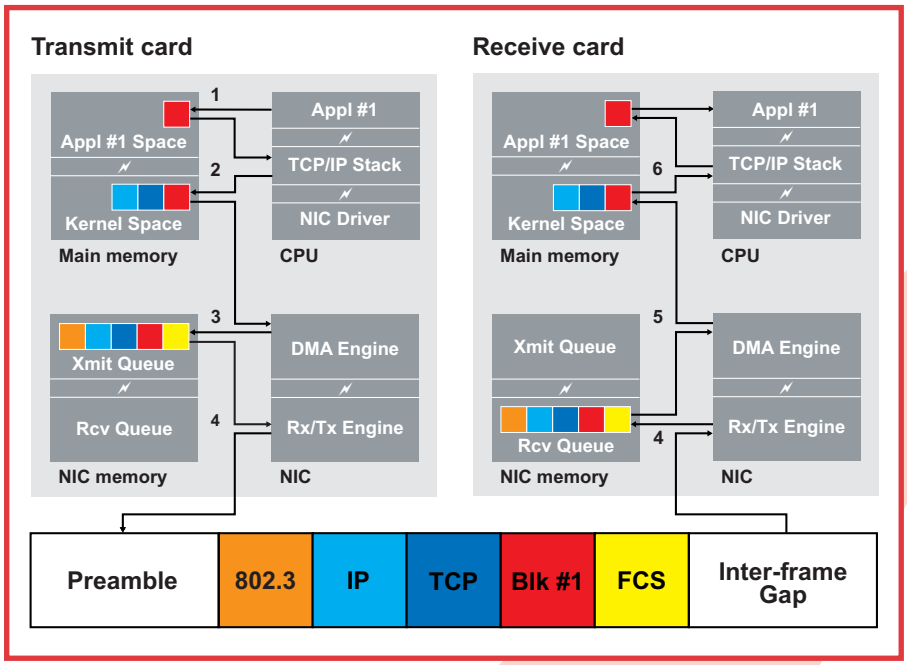

Bild 2. Beim Einsatz von Ethernet als Datenübertragungsmechanismus werden zahlreiche Daten-Kopien benötigt

bietet sie einen offenen Kanal für die Wiederherstellung wichtiger Informationen oder für die partielle Funktion des Moduls.

Karteninterne Steuerungsebene

Steuerungs-Zugriffe auf Bausteine sind in der Regel mit dem Zugriff auf Register oder der Verschiebung von Datenblöcken in den Speicher oder aus ihm heraus verbunden. Diese Funktion eignet sich sehr gut für das in PCIe verwendete speichergestützte Funktionsprinzip. Außerdem neigen Entwickler von Karten-internen Steuerfunktionen zum Einsatz eines hierarchischen Modells, bei dem eine einzige Housekeeper-CPU jedes einzelne Bauteil auf der Karte verwaltet. Auch dieses Konzept passt gut zum PCIe-Modell. Schließlich verläuft der Datenverkehr auf der Steuerungsebene einer einzelnen Karte fast exklusiv zwischen der Housekeeper-CPU und den einzelnen Bausteinen, und nicht im „Peer-to-Peer" Modus zwischen den Bauteilen.

Backplane-Steuerungsebene zwischen intelligenten Steckkarten

Verteilte Datenverarbeitungskonzepte sind schon immer die Architektur der Wahl für größere Kommunikationssysteme im Rand- oder Kernbereich des Netzwerks gewesen, und sie setzen sich immer mehr auch im Zugangsbereich durch. In einer solchen Architektur bleibt der Adressbereich jeder Karte klar von dem anderer Karten getrennt, sodass sich die gleiche Software ohne Änderungen auf mehreren Instanzen des gleichen Kartentyps ausführen lässt.

Auf Grund der komplexeren Datenverarbeitungs-Anforderungen von Edge- und Core-Knoten muss jede Schnittstellenkarte einen Großteil oder alle ihrer Daten selbst verarbeiten. Angesichts der Anzahl und Komplexität der auf jeder Karte befindlichen Bausteine enthalten solche Karten in der Regel einen eigenen Housekeeping-Prozessor. Oft verarbeiten diese Prozessoren auch die vom Datenpfad erzeugten Ausnahmen. Der Einsatz solcher Prozessoren auf den meisten oder allen Schnittstellenkarten verändert in der gesamten Backplane die Art der Kommunikation auf der Steuerungsebene. Diese Kommunikation beruht jetzt auf dem Modell einer Interprozessor-Kommunikation, bei der meist Messages und Interrupts an Stelle direkter Registerzugriffe ausgetauscht werden.

Bei diesem Modell übernimmt der Housekeeping-Prozessor auf jeder Karte die Verwaltung aller lokal auf der Karte vorhandenen Komponenten (sowie aller damit verbundenen I/O-Module). Der Prozessor verwaltet außerdem alle Messages, die mit Housekeeping-Prozessoren auf anderen Karten ausgetauscht werden, sowie die Kommunikation zwischen der Karte und dem zentralen Steuerungs-Prozessor im System.

Interprozessor-Kommunikation über Endpoint-Prozessoren

Immer mehr Prozessoren unterstützen Adress-Remapping für ihre Transaktionen. Bei diesen Bausteinen kann der Entwickler Daten direkt aus dem Speicher eines Moduls in den Speicher eines anderen mit Hilfe eines speichergestützten Adressierungskonzepts verschieben, obwohl diese Module getrennte Adressdomänen verwenden.

Dank dieser Technik kann der Entwickler PCIe mit seinem speichergestützten Adressierungskonzept als systemweites Verbindungsmedium für Transaktionen auf Steuerungsebene nutzen. Dieses Konzept bietet eine Reihe von Vorteilen gegenüber heute üblichen Lösungen. Als erstes vereinfacht es die Aufgaben der Hardware- und Software-Entwickler, weil sie kein proprietäres Protokoll mehr erlernen müssen. Zweitens vereinfacht es auch die Systementwicklung, da Paketcodierung, Prioritätsklassifizierungen und Datenfluss-Steuerung nicht mehr zwischen zwei Protokollen übersetzt werden müssen. Außerdem verspricht der systemweite Einsatz von PCIe als Verbindungsmedium für Transaktionen auf Steuerungsebene eine Verringerung der Transferlatenzzeiten und der erforderlichen Speicherbandbreite, weil die beim Message Queuing erforderlichen Übersetzungsschritte und das mehrfache Handling von Daten entfallen.

Interprozessor-Kommunikation über nichttransparente Bridges

Ein nicht-transparenter PCIe-Bridge-Baustein erscheint für die Domäne an jedem Port als PCIe-Endpoint. Er nimmt sämtliche an ihn gerichteten Transaktionen auf und erzeugt neue, veränderte Transaktionen an seinem anderen Port für viele weitere Ports. Weder die Manipulationen zur Änderung dieser Transaktionen, noch die Register zur Konfiguration dieser Manipulationen sind standardisiert, obwohl die meisten Bauteil-Hersteller ziemlich ähnliche Transpositionen sowie andere Interprozessor-Kommunikationsdienste wie Scratchpad-Register und Doorbell-Interrupts anbieten.

Welche Probleme stellen sich bei der Entwicklung einer Kommunikationssystem-Plattform und wie lassen sich diese mit PCIe lösen?

Redundanz

Jede Diskussion von Switching-Vorgängen auf einer Backplane in einer Kommunikationstechnik-Umgebung muss sich mit dem Thema Redundanz befassen. Sämtliche Redundanzkonzepte im Zusammenhang mit PCIe müssen eine Regel in der Spezifikation berücksichtigen: Es darf nur ein einziger Port in einem PCIe-Switch oder einer -Bridge als Upstream-Port benutzt werden. Der Zweck dieser Regel ist es, eine versehentliche oder mutwillige Rekonfigurierung von PCIe-Switches und -Bridges durch Peripheriekomponenten zu verhindern.

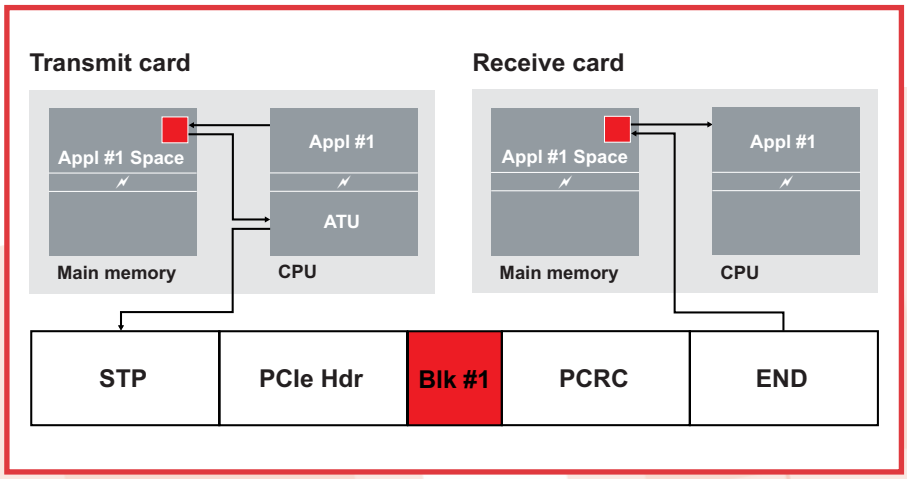

Bild 3. Direkte Speicherzugriffskonzepte wie PCIe erzeugen keine zusätzlichen Datenkopien

Über eine einfache Umprogrammierung des Switches lässt sich ein ausgefallenes, hierarchisch tiefer angeordnetes Element in einer Anwendung deaktivieren und der Datenverkehr vom ausgefallenen Element auf ein Reserveelement umleiten. Dabei sind N+1 und 1:1 Redundanzkonzepte möglich. Einige Beispiele für dieses Vorgehen sind z. B. eine Line-Prozessorkarte, die mit mehreren I/O-Modulen kommuniziert, oder der Austausch einer „dummen" Linecard in einem System. Erfordert eine Situation den Ersatz eines ausgefallenen, hierarchisch höher angeordneten Elements wie z. B. einer zentralen Steuereinheit, einer Prozessorkarte oder eines Switch Fabric Elements, dann muss man den Downstream-Elementen vortäuschen, dass sie ihre Befehle und Daten nach wie vor vom gleichen Upstream-Element erhalten. Dies lässt sich durch den Einsatz eines für PCIe-Signale ausgelegten 2:1 Multiplexers erreichen, der bestimmt, welches der beiden Upstream-Elemente den Datenverkehr an die Downstream-Elemente weitergibt. Eine solche Struktur kann allerdings nur das 1:1 Redundanzkonzept unterstützen.

Die bevorzugte Methode zur Implementierung einer Upstream-Redundanz (**Bild 1**) ist die Anordnung eines nicht-transparenten 3-Port PCIe-Switches auf den intelligenten Linecards, wobei der Upstream-Port mit dem aktiven Fabric/Steuerungskarten-Komplex und der nicht-transparente Port mit dem inaktiven Partner verbunden ist. In dieser Architektur kann der Datenverkehr zwischen den aktiven und den nicht-aktiven Steuerungskarten, zwischen dem inaktiven Steuerungs-Komplex sowie allen Linecards fließen und damit die Integrität der Links gewährleisten. Einfache Umschaltmechanismen innerhalb der nicht-transparenten Switches ermöglichen ein Umschalten der Aktivität mit minimaler Verzögerung und geringstem Traffic-Verlust.

Integrierte Architektur für Steuerungs- und Datenebene

Als Konsequenz aus den hohen Verfügbarkeitsanforderungen der meisten Kommunikationssysteme müssen System-Steuerfunktionen unabhängig vom Geschehen auf der Datenebene Zugriff auf jedes Modul im System haben.

Zur Erfüllung dieser Anforderungen nutzte man bisher physikalisch getrennte Datenpfade über die System-Backplanes beziehungsweise innerhalb der hochkomplexen einzelnen Karten. PCs und Server benutzten traditionell nur eine einzige Ebene für Steuer- und Datenverschiebungs-Funktionen. Obwohl die Kommunikationsindustrie nach wie vor engagiert eine physikalische Trennung der Steuerungs- und Datenebene vertritt, kann sich eine integrierte Architektur für die Steuerungs- und Datenebene aus einer Reihe von Gründen immer mehr durchsetzen:

- Backplane-Switches können heute eine Trennung des Datenverkehrs gewährleisten.
- Backplane-Switches können sicherstellen, dass Traffic mit hoher Priorität selbst beim Auftreten eines Denial-of-Service Angriffs auf niedrigere Prioritäten durchgeschleust wird.
- Mit dem weiteren Vordringen von PC-Architekturen in den Kommunikationstechnikbereich wird es mehr System- und Chip-Architekturen geben, die keine Unterstützung für eine Ebenentrennung bieten.

Engpass durch Mehrfachkopien

Forschungs- und Entwicklungsarbeiten der letzten Zeit konzentrierten sich auf eine Entschärfung des Engpasses bei mehrfachen Datenkopien. Dieses Problem lässt sich am besten anhand eines beispielhaften Datenflusses (**Bild 2**) verdeutlichen. Will man einen Datenblock von einer Anwendung über ein Message-gestütztes Übertragungsmedium wie Ethernet zu einer Anwendung auf einem anderen Prozessor verschieben, so werden diese Daten während eines solchen Vorganges oft sechsmal oder häufiger kopiert.

Solche Mehrfachkopien sind typisch für Datenbewegungen in vielen Ethernet-basierten Servern. Einige dieser Schritte sind erforderlich, weil die Speicherräume für Anwendung und Kernel separat gehalten werden müssen. Manche von ihnen werden aber nur zur Unterstützung mehrerer Schichten für Framing, Sequenzierung und Schutz für Message-gestützte Verbindungen benötigt.

In letzter Zeit tendieren Entwickler von Server- und Speichersystemen zu einem Shared-Memory-Modell, wobei in der Regel eine Adressübersetzung zum Einsatz kommt. In diesem Modell geben Anwendungen Daten-Pointer an intelligente I/O-Prozessoren weiter, die die Daten direkt in den Speicher der Ziel-Anwendung verschieben, wobei die Adressen während der Übertragung jedes Datenwortes übersetzt werden. Zwar ignoriert diese Beschreibung den Austausch der Pointer zwischen den verschiedenen Verarbeitungselementen, der Datenblock selbst aber wird nur einmal bewegt (**Bild 3**).

An Stelle eines Message-gestützten Vorgehens arbeitet dieses spezielle Datenverschiebungsmodell mit einem Speicheradressen-orientierten Verbindungskonzept. Daher nutzt man in Blade-Server-Systemen an Stelle von Ethernet-Backplanes zunehmend PCIe-Backplanes. Da sich Systementwickler mehr und mehr mit den Effizienz- und Latenzzeitvorteilen dieses Datenverschiebungskonzeptes vertraut machen, könnte ein ähnliches Konzept bald auch in Kommunikationssystemen zum Einsatz kommen.

Fazit

Auf Grund seiner Möglichkeiten zur Erweiterung der Signalreichweite und der verringerten Pin-Anzahl wird sich PCIe schnell zum bevorzugten CPU-Portprotokoll entwickeln. Weil schon heute viele CPUs mit PCIe-Ports erhältlich sind, und dazu die meisten in Kommunikationssystemen bevorzugt eingesetzten CPU- und Housekeeper-Typen zählen, wird sich PCIe auf der karteninternen Steuerungsebene immer mehr durchsetzen. Überall dort, wo eine CPU auf einer Karte Bausteine auf einer anderen Karte direkt steuert, kann PCIe zur Erweiterung des Steuerprotokolls über die Backplane eingesetzt werden. *(dar)*

- **IDT**

- www.el-info.de ➤ Webcode: 03204