Initiatives and Technologies

PCI Express Provides Enterprise Reliability, Availability, and Serviceability

PCI EXPRESS TECHNOLOGY*

get in the express lane

desktop • enterprise • mobile • communications

Contents

Overview 4

The components of PCI Express RAS 4

Reliable protocol architecture 4

Standard error handling and reporting 5

Standard hot-plug usage model 6

Summary 7

For more information 7

Author biography 7

Overview

PCI Express* technology is the industry standard I/O interconnect expected to provide local I/O connectivity across desktop, mobile, enterprise and communications platforms. It resides at the center of enterprise interconnect innovations anticipated across storage, networking, and clustering. While developers are becoming increasingly aware of the performance-related capabilities of PCI Express — the higher bandwidth it brings to server platforms and its scalability for a range of devices — they may still be learning about the architecture's outstanding capabilities in reliability, availability, and serviceability, also known as RAS. This article provides a comprehensive starting point for developers wanting to learn more about PCI Express RAS.

The components of PCI Express RAS

PCI Express is rich in RAS capabilities, which are vital to customers in corporate IT and the data center. This means that by incorporating PCI Express RAS into their early product designs, platform OEMs and adapter developers can position themselves well ahead of their competition. Moreover, they can do so without having to make changes in operating systems or drivers because PCI Express defines an evolutionary bus model that is fully compatible with existing software infrastructures. Software developers also can position themselves ahead of the competition by taking advantage of the advanced RAS features of PCI Express adapters and promoting those features into standard usage models.

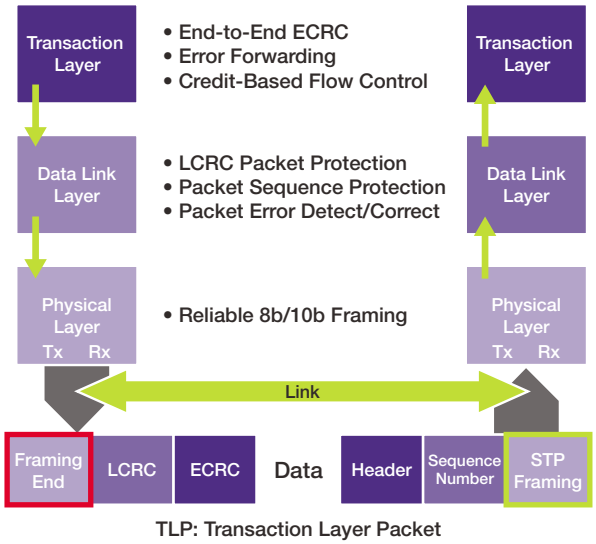
PCI Express RAS consists of three fundamental components: (1) a reliable protocol architecture; (2) device-level protocol error detection, correction, and reporting; and (3) device-level requirements for support of a hot-plug usage model (SHPC 1.0) based on current industry standards.

Reliable protocol architecture

To enable the smooth implementation of RAS, PCI Express provides reliable protocol error detection, correction, and reporting capabilities at three cooperative functional layers of a device architecture: physical, data link, and transaction. In addition, PCI Express defines the base unit of interdevice bus communication as the transaction layer packet (TLP). Within a PCI Express–based design, one or more TLPs combine to form a transaction, which is transmitted over a link from one device on the bus to another. For a given TLP, a device provides device-level error detection and correction through analysis of reliability mechanisms present in every TLP. These error detection/correction responsibilities span the three functional device layers.

From a transaction perspective, PCI Express protocol reliability begins at the physical layer and is based on the ANSI X3.230-1994 standard. Following this standard, PCI Express provides special symbols that demarcate TLPs, identify basic link-management functions, and enable the physical layer to quickly and reliably identify and differentiate TLPs from link-management functions and forward TLPs directly to the data link layer for further processing.

Reliable Protocol Architecture



Data Reliability is at the Foundation of PCI Express

Figure 1: PCI Express architecture provides reliable protocol error detection, correction, and reporting across all three functional layers of a device architecture: physical, data link, and transaction.

At the data link layer, the primary responsibility of PCI Express is to detect and correct protocol errors. Within the PCI Express architecture, each TLP contains a full 32-bit link cyclical redundancy check (LCRC) field as well as a transaction sequence number to preserve transaction reliability and data integrity. An LCRC and sequence number are generated and applied to every TLP by the data link layer of the transmitting device and checked by the data link layer of the receiving device. The transmitting data link layer is capable of correcting most protocol errors automatically through TLP transmission retries.

The uppermost layer (from a software perspective) in the PCI Express device hierarchy is the transaction layer. This layer has two primary responsibilities: (1) to receive TLPs from the data link layer and create TLPs for transmission through it and (2) to implement a transaction flow-control mechanism. Flow control ensures that a transmitter and receiver cooperate to ensure that no TLP is transmitted to a receiving device unless that device has a place to "hold" the TLP—thereby boosting overall data reliability by limiting the opportunity for "lost" transactions.

Additionally, in a PCI Express device the transaction layer can be configured to apply an end-to-end CRC (ECRC) to every transmitted TLP and to check the ECRC for every received TLP. Unlike the LCRC, which can be regenerated (by switches/bridges) during TLP transmission and interdevice routing, the ECRC remains unmodified during this process. This approach greatly increases reliable data transfer by including end-to-end transaction reliability checking in addition to the link cyclical redundancy checking already provided by the data link layer.

Standard error handling and reporting

Complementing its reliable protocol architecture, PCI Express provides comprehensive standards for improved device error reporting. Device error reporting consists of two elements: (1) standard hardware error detection and recording (which is enabled by and required of all PCI Express devices through the reliable protocol architecture) and (2) the capability for software to configure devices to report protocol errors (to software) in a standard fashion.

To enhance server reliability, all PCI Express devices are designed to record and report protocol errors in a common, vendor-independent fashion through standard register requirements. This standard for hardware error reporting enables system software to implement vendor/device-independent server models for error detection and correction. To ensure the preservation of PCI Express standards, the PCI Express architecture specifications impose device requirements that are used to measure device implementations through PCI SIG workshop testing. PCI Express standard device error reporting enables designers of server system software to deliver a higher reliability baseline than what is available through existing PCI- and PCI-X–based servers.

Standard Error Reporting

Error Scope

Transaction Layer	Correctable	Fatal
Poisoned TLP Receiver	NO	NO
ECRC Check Failure	NO	NO
Unsupported Request	NO	NO
Completion Timeout	NO	NO
Completer Abort	NO	NO
Unexpected Completion	NO	NO
Receiver Overflow	NO	YES
Flow Control Error	NO	YES
Malformed TLP	NO	YES
Data Link Layer	**Correctable**	**Fatal**
Bad TLP	YES	N/A
Bad DLLP	YES	N/A
Replay Timeout	YES	N/A
Replay NUM Rollover	YES	N/A
Protocol Error	NO	YES
Physical Layer	**Correctable**	**Fatal**
Receiver Error	YES	N/A
Training Error	NO	YES

Robust Error Detection/Correction Exists at Each Layer

Figure 2: **PCI Express provides robust error detection and reporting at all three layers of the bus architecture.**

If software chooses to gather more detail regarding device errors, PCI Express configuration options provide flexibility in error-reporting detail. These options include uncorrectable error-severity adjustment, error filtering or masking, layer-specific error type logging (with reporting of detailed error codes), identification of correctable and uncorrectable error source (source-device identification), and logging of the TLP headers of error messages received from devices.

PCI Express specifies two major classifications of errors: (1) correctable by the hardware, with no data loss, and (2) uncorrectable by the hardware, with potential data loss. Uncorrectable errors are further classified as either non-fatal or fatal. With the combination of its reliable protocol services and standard error reporting, PCI Express delivers a key advantage over existing PCI and PCI-X bus models: enabling system software to perform predictive failure analysis, which can greatly increase server uptime. By introducing device standards for detailed error reporting that includes both correctable and non-fatal errors, PCI Express devices provide system software the information it needs to monitor frequency of anomalous protocol activity on a per-device basis, before data loss occurs. Problem devices can be identified (predicted) and swapped out of a PCI Express Server before a fatal error occurs.

Standard hot-plug usage model

Corporate IT and data center managers face a classic serviceability problem with the multitude of service models presented by the diverse configurations of server/adapter products constituting a given server platform. It stands to reason that diverse service models translate into complexity of service and an increased likelihood of downtime.
To address this problem, PCI Express enables system software to present a standard service model for server hot-plug capability — replacement of a slotted device following a fatal error while the server remains operational. This service model is based on current industry hot-plug standards, namely the Standard Hot Plug Controller (SHPC) 1.0.

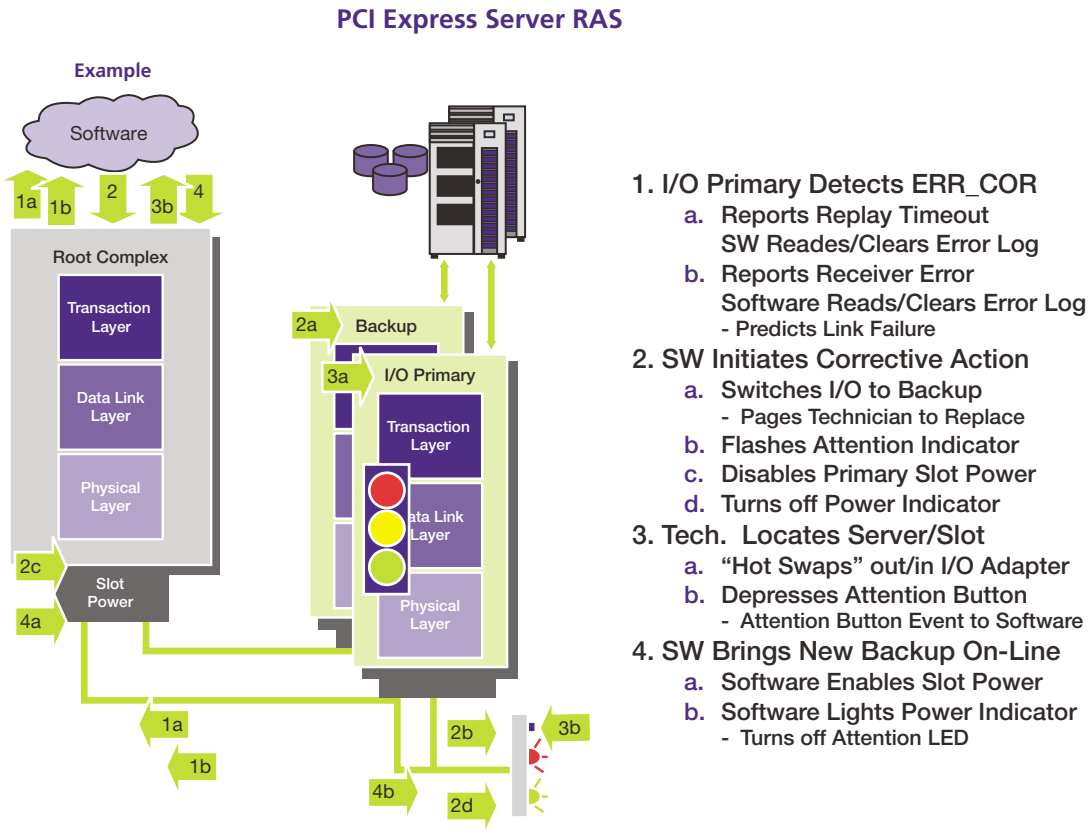


Figure 3: **By supporting predictive failure analysis, PCI Express enables proactive error correction within corporate IT or the data center.**

A standard hot-plug usage model defines capabilities that hardware must implement to provide a platform-independent means of implementing hot plug. This is required so that standard distribution system software can be developed to "know" how to check whether hot-plug capabilities exist at the slot and device level. By supporting the SHPC 1.0 usage model, PCI Express Native Hot Plug can deliver a true slot-based FRU (Field Replaceable Unit) schema to server racks. This enables data-center technicians, and the managers who train them, to focus on a consistent, vendor-independent formula for floor service across diverse adapters and vendor systems.

The PCI Express Native Hot Plug model is defined through standard register requirements at two functional levels: module/card (device) and chassis/slot. At the module/card level, PCI Express specifies that slotted endpoints must declare yes/no support for key elements of native hot plug, such as power indicator present, attention indicator present, and attention button present. At the chassis/slot level, PCI Express enables the operating system or management software to "discover" whether a given chassis capability exists on a given slot and then to enable it within software for that chassis/vendor configuration. Standard device requirements enable standard distribution system software to implement vendor-independent driver-support models, which translate into server platform standards, which in turn translate into common models for serviceability within corporate IT and the data center.

Summary

PCI Express opens the door for device, platform, and software developers to deliver outstanding levels of server RAS — reliability, availability, and serviceability — to the lowest level of hardware, the highest level of enterprise management, and all points in between. Through PCI Express, RAS enters the foundation of next-generation servers in three fundamental ways. First, it becomes integral to the bus and system components through hardware design requirements in support of reliable protocol services. Second, it enters devices and platforms through device requirements for standard error handling and reporting and a standard hot-plug model. Third, it enters the server platform and beyond, into the enterprise, through propagation of standard models for error reporting, predictive correction, and serviceability. This translates into a clear message for developers seeking to differentiate their server products by implementing superior RAS: now is the time to begin incorporating PCI Express.

For more information

PCI Express is the latest development in the evolutionary PCI bus architecture, which was initiated by Intel and since adopted by a broad spectrum of industry leaders. For more information on PCI Express and its support for RAS capabilities, visit the Intel developer site at http://developer.intel.com/technology/pciexpress/devnet/ or the PCI SIG site at http://www.pcisig.com/home. You also can learn more detail through a slide set available at http://www.intel.com/technology/pciexpress. For greater technical depth, see the *PCI Express Technology Primer* to be published through the Intel Press; pre-release information is available at http://www.intel.com/intelpress.

Author biography

Steve Krig is a staff software engineer in the Enterprise Platform Group at Intel Corporation. In his 15 years with the company, he has played significant roles in the introduction of emerging technologies to the industry such as DMI (Desktop Management Interface), AGP (Accelerated Graphics Port), InfiniBand, IPCO (Industry BKMs for Design Compliance), and other major projects and initiatives. He also has received division recognition for his work on these projects and is the co-author of the compliance and interoperability section in the *PCI Express Technology Primer*. Steve holds a B.S. in computer science from Seattle Pacific University.

Copyright© 2003 Intel Corporation. All rights reserved.
*Other names and brands may be the property of others.