Product Brief

PEX9700 Series Switch Chips

Managed PCI Express Switches Based on ExpressFabric® Technology

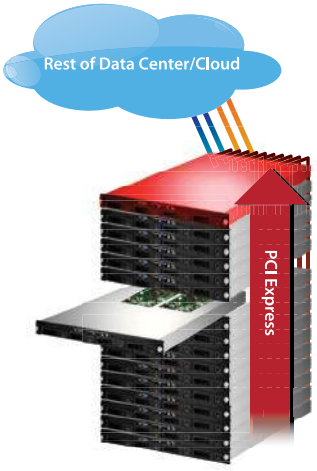
General Features

- State-of-the-art switch fabric
  - Sharing I/Os among multiple hosts
  - Host-to-host DMA
  - Low latency TWC
- Any port can be a host port or Downstream (device) Port
- Works with standard PCIe end-points and hosts – and software, as well as with existing application software
- MSI-X support
- Allows flexible fabric topologies

Key Advantages

- PCI Express Switches
  - 12 to 97 Lanes with Integrated on-chip SerDes
    - 5 to 25 Independent ports
    - Designate any Port as the Upstream Port
  - Low-power SerDes (under 90 mW per Lane)
  - Device-Specific Relaxed Ordering
  - Port configuration
  - Dedicated management port for mCPU
  - x4, x8, or x16, depending on Port configuration; x4 can down-train to x1 and x2 width
  - Configurable through serial EEPROM, I2C, SMBus, and/or Host port
- Standards Compliant
  - PCI Express Base Specification, r3.1 (backward compatible w/ PCIe r2.0, & r1.0a/1.1)
  - PCI Power Management Spec, r1.2
- High Performance
  - Full line rate on all ports
  - Cut-Thru packet latency of less than 150ns (x16 to x16)
  - 2KB Max Payload Size
  - Multicast through DMA

Converge Servers and IO controllers with PCIe



- Create cost-effective high-availability hyperscale systems by enabling communication between in-rack hosts and endpoints using PCIe
- Simplify connectivity while providing the highest PCIe switching performance available for data center servers, storage, and networks
- Reduce latency, system complexity, and power consumption by up to 50% in data-intensive environments
- Take advantage of industry-first features for most demanding hyper-converged, NVMe and rack scale systems

Avago PEX9700 switches allow customers to build high performance, low latency, scalable, cost-effective PCI Express-based fabrics. The switches enable I/O sharing with standard SR-IOV or multifunction capability, allowing multiple hosts or Nodes to reside on a single PCIe-based network. Hosts communicate through Ethernet-like DMA (NIC DMA) with other hosts and end-points using application software. Hosts may also communicate using Tunneled Window Connection (TWC), a special low latency host-to-host communication capability for short packets.

Shared I/O Using Standards

PEX9700 switches allow the Virtual Functions (VFs) of SRIOV endpoints (such as an Avago MegaRAID® SAS controller) to be shared and assigned to multiple hosts concurrently. Each host can enumerate its assigned functions using standard BIOS and OS software and use them with unmodified vendor-supplied drivers. The use of standard system software minimizes software support costs.

General Purpose Host-to-Host DMA

Ethernet is used almost universally for server to server communications. PEX9700 switches contain a virtual Ethernet NIC at each host port that allows Ethernet to be tunneled transparently through the fabric to any and all servers that are connected. Internal Ethernet communications using virtual Ethernet NICs and NIC DMA are complemented by the ability to share a physical SRIOV NIC, thus providing compatibility with the vast library of applications that leverage Ethernet communications.

Software-Defined Fabric

The switches are built on a hybrid hardware/software platform that offers high configurability and flexibility in regards to the number of hosts, end-points, and PCIe slots. The critical pathways have direct hardware support, enabling the fabric to offer non-blocking, line speed performance with features such as I/O sharing and DMA. The solution is completed by management processor that communicates with platform management via API and/or CLI. The solution offers an innovative approach to setup and control, making use of an off-chip management CPU (mCPU) to initialize the PEX9700 switch, configure the routing tables, handle errors, Hot-Plug events, and enable the solution to extend the capabilities without modifying the system software.

Key Advantages (continued)

- Quality of Service (QoS)
  - 8 Traffic Classes (TC) supported
- Reliability, Availability, Serviceability
  - visionPAK™
  - performance PAK™
  - DPC/eDPC Support
  - Read Tracking for surprise removal
  - All ports Hot-Plug capable thru I2C –SSC isolation on all ports
  - SRIS support
  - ECRC and Poison bit support
  - Port Status bits and GPIO available

Tunneled Window Connection (TWC)

The DMA or TWC approaches are two ways hosts can communicate. TWC allows short messages to be sent from one host to another in a very low latency manner, and without the overhead associated with DMA.

Downstream Port Containment (DPC/eDPC)

Most servers have difficulty handling serious errors, especially when a PCIe end-point disappears from the system. DPC/eDPC allows a downstream link to be disabled after an uncorrectable error, making recovery possible in a controlled and robust manner.

Flexible Topologies

PEX9700 switches eliminate the topology restrictions of PCIe. The switch allows other topologies such as mesh, I/O Expansion Box with Multiple Hosts, and many others. And it does this while allowing the components to remain architecturally and software compatible with standard PCIe.

Improved SSC Isolation

The switches offer several mechanisms for supporting multi-clock domains that include spread spectrum clocking; eliminating the need to pass a common clock across a backplane. In addition to the standard Avago approach to the problem, a new PCI-SIG approach called SRIS (Separate Refclk Independent SSC Architecture) is now available.

Applications

Products based on ExpressFabric technology can help deliver an outstanding solution for designing a heterogeneous system with a requirement for a flexible mix of processors, storage elements, and communication devices.

HPC Clusters

HPC clusters are made up of high-performance processing elements that communicate through high bandwidth, low latency pathways in order to execute applications such as medical imaging, financial trading, data warehousing, etc. PEX9700 switches can be used in switch fabric applications for HPC clustering. The processing subsystems can be connected to the PCIe fabric while running the same application software. PCIe switch based clustering eliminates expensive protocol bridging devices resulting in lower cost and power. And clustering systems can be built with I/O sharing as an additional native capability when needed.

Software Development Kit (SDK)

The SDK for the PEX9700 series includes drivers, source code and GUI interfaces to aid in configuring and debugging. Both the performancePAK™ and visionPAK™ are exclusive to Avago and are supported by its RDK and SDK, which are the industry's most advanced hardware-and software development kits.

performancePAK

The performancePAK is a suite of unique and innovative performance features that allows Avago Gen 3 switches to be the highest performing switches in the market today.

visionPAK

The visionPAK is a debug diagnostics suite of integrated hardware and software instruments that allows users to help bring their systems to market faster.

PEX9700 Series

Part Number	Lanes	Ports	Latency (ns)	HPC*	Aggregate Bandwidth	SSC*	Dedicated x1 mCPU Port	DMA Multicast	Package Size (mm^2)	Typical Power Modes			
										Power Typ. (W)	Peer-to-Peer	Fanout	Fabric
PEX9797	97	25	150	6	1536GT (8.0 GT/s/Lane x 96 SerDes x2 (full-duplex))	24	Yes	Yes	35x35	23.9	24.3	20.6	25.0
PEX9781	81	21	150	5	1280GT (8.0 GT/s/Lane x 80 SerDes x2 (full-duplex))	20	Yes	Yes	35x35	21.5	22.5	19.6	23.3
PEX9765	65	17	150	4	1024GT (8.0 GT/s/Lane x 64 SerDes x2 (full-duplex))	16	Yes	Yes	35x35	15.9	16.2	13.9	16.9
PEX9749	49	13	150	4	768GT (8.0 GT/s/Lane x 48 SerDes x2 (full-duplex))	12	Yes	Yes	27x27	13.5	14.5	12.8	15.2
PEX9733	33	9	150	2	512GT (8.0 GT/s/Lane x 32 SerDes x2 (full-duplex))	8	Yes	Yes	27x27	7.9	8.1	7.2	8.9
PEX9716	16	5	154	1	256GT (8.0 GT/s/Lane x 16 SerDes x2 (full-duplex))	4	No	No	19x19	4.0	4.0	3.8	4.8
PEX9712	12	5	158	1	192GT (8.0 GT/s/Lane x 12 SerDes x2 (full-duplex))	4	No	No	19x19	3.5	3.7	3.4	4.4

Product Ordering Information

Switch Part Numbers	Description	Rapid Development Kit (RDK) Part Number
PEX9797-AA80BC G	97-Lane, 25-Port ExpressFabric Device (35 × 35 mm2)	PEX9797-AARDK
PEX9781-AA80BC G	81-Lane, 21-Port ExpressFabric Device (35 × 35 mm2)	PEX9797-AARDK
PEX9765-AA80BC G	65-Lane, 17-Port ExpressFabric Device (35 × 35 mm2)	PEX9797-AARDK
PEX9749-AA80BC G	49-Lane, 13-Port ExpressFabric Device (27 × 27 mm2)	PEX9749-AARDK
PEX9733-AA80BC G	33-Lane, 9-Port ExpressFabric Device (27 × 27 mm2)	PEX9749-AARDK
PEX9716-AA80BC G	16-Lane, 5-Port ExpressFabric Device (19 ×19 mm2)	PEX9716-AARDK
PEX9712-AA80BC G	12-Lane, 5-Port ExpressFabric Device (19 × 19 mm2)	PEX9716-AARDK
PXF55033-AA	32-Port ExpressFabric top-of-rack switch box with QSFP+ connections	
PXF51003-AA	2-Port PCIe bus extender card with redrivers and QSFP+ connections	

Acronym Guide

DMA.....Direct Memory Access
HPC.....Hot-Plug Controllers
TWC.....Tunneled Window Connection
SSC.....Spread Spectrum Clock Isolation
MSI-X.....Message Signaled Interrupts
SRIS.....Separate Refclk Independent SSC Architecture
DPC.....Downstream Port Containment
eDPC.....Enhanced DPC
Commercial Temperature Range.....0+70 (Celsius)

For more information, visit www.avagotech.com