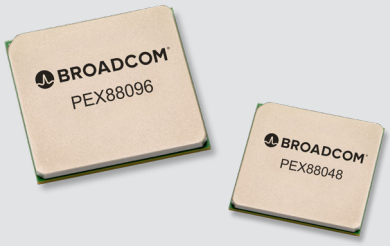
BROADCOM®

Product Brief

PEX88000 Series

Managed PCI Express 4.0 Switches



Key Features

- PCIe 4.0 r1.0 support
- Embedded ARM CPU for management
- ExpressFabric® PCIe switching architecture
- Sharing I/Os among multiple hosts
- 48 General purpose DMA functions/ channels
- Any-to-Any DMA transfer
- Any port can be a host port or downstream (device) port
- Low latency data transfers
- Works with standard PCIe endpoints, hosts and software
- MSI-X support
- Allows flexible fabric topologies
- 48 non-transparent bridging (NTB) ports
- Embedded MPT end-point
- Industry's best SerDes technology with extended reach

Converged Servers, Storage Systems, and I/O Systems Using PCIe

Broadcom PEX88000 switches allow customers to build systems from simple PCIe connectivity inside the box to high performance, low latency, scalable, cost-effective PCIe fabrics for composable hyper-scale compute systems.

- Create basic switch topologies at PCIe 4.0 rate to connect Hosts/CPUs to I/Os and peripherals in server and storage systems
- Create cost-effective high-availability hyper-scale 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 in data-intensive environments
- Take advantage of industry-first features for most demanding hyperconverged, Artificial Intelligence, Machine Learning, NVMe JBOFs, and rack scale systems

General Purpose DMA for Highly Efficient Any-to-Any Data Transfers

PEX88000 device offers up to 48 DMA channels/functions – one associated with each PCIe x2 port enabling data transfers from one host to another, host-to-I/O and I/O-to-I/O with very low latency, and without the overhead associated with data transfers using host DMA. Multiple DMA functions may be active simultaneously without interfering with other streams of data transitioning through other ports of the switch.

Enhanced Non-Transparent Bridging 2.0 (NT2.0)

PEX88000 switches are equipped with field-proven NT technology that Broadcom has been shipping since 2004. This multi-host enabling architecture has been enhanced to NT2.0 based on years of use and feedback by leading OEMs/ODMs. The largest PEX88000 switch (96-lane device) is equipped with 48 NT2.0-capable ports enabling a large number of hosts/servers to be connected to the switch.

Embedded ARM CPU

Each PEX88000 PCIe switch is equipped with an embedded ARM Cortex-R4 CPU, internal RAM, timer blocks, watchdog timer, and vectored interrupt controllers. The embedded CPU can be used to configure desired switch functionality, creation of synthetic hierarchy, I/O management, Hot add/remove, and interrupt handling.

Key Advantages

- 24 to 96 data lanes with integrated on-chip SerDes
- 96, 80, 60, 48, 32, 24-lane packages available
- Two additional management port for mCPU
- SerDes field-tested by Broadcom
- Low-power SerDes (under 90 mW per Lane) switches
- Device-specific relaxed ordering
- Port reconfiguration without impacting other ports
- Port configuration
  - 24 to 96 independent ports
  - Each port speed (Gen1/2/3/4) independent of others
  - Choice of link width – x1, x2, x4, x8, or x16
  - Designate any port as the upstream port
- Standards compliant
  - PCI Express base specification, r4.0, r3.0, r2.0, r1.0
  - PCI power management spec, r1.2
  - High performance
  - Full line rate on all ports
  - Cut-through packet latency of less than 100 ns (x16 to x16)
  - 2 KB max. payload Size
- Quality of service (QoS)
  - 8 traffic classes (TC) supported
- Reliability, availability, serviceability
  - VisionPAK – SerDes Eye capture
  - Performance PAK
  - DPC/eDPC support
  - Read tracking for surprise removal
  - All ports hot-plug capable via
  - SSC isolation on all ports
  - SRIS support
  - ECRC and poison bit support
  - Port status bits and GPIO available

Switch Operation Modes

- Base Mode: The switch functions without any FW involvement. In base mode, the embedded CPU can be disabled and the device will operate as a standard PCIe fan-out switch.
- Base with MPT Mode: The switch may be programmed to provide NVMe fanout, with a full set of chassis management capabilities with MPT endpoint.
- Synthetic Mode: Adds the ability to synthesize the hierarchy from the host point of view. This can be done with the embedded CPU.

Software Defined PCIe Switch Fabric

The switches are designed for hybrid hardware/software platforms that offer high configurability (the number of hosts, downstream ports, and assignment of the slots/ports with those hosts). Once the configuration is complete, the data flows directly between connected devices with hardware support, enabling the fabric to offer non-blocking, line-speed performance with features such as I/O sharing and DMA. The solution offers an innovative approach to set up and control the PEX88000 switches, configure the routing tables, handle errors, Hot-Plug events, and enable the solution using either embedded CPU or an external CPU without impacting data flowing through the switch.

Flexible Topologies

PEX88000 switches eliminate the topology restrictions of PCIe. The switch allows Multiple Hosts to connect to a single PCIe switch complex to enable topologies for hyper-scale systems. And it does this while staying compatible with standard PCIe protocol.

Downstream Port Containment (DPC/eDPC)

Most servers have difficulty handling serious errors in I/O devices, especially when a device disappears from the system. PEX88000 DPC/eDPC implementation allows a downstream link to be disabled after an uncorrectable error or time-out, making recovery possible in a controlled and robust manner.

Improved SSC Isolation

The switches offer multi-clock domains that include spread-spectrum clocking. PCI-SIG approach, called SRIS (Separate Refclk Independent SSC Architecture), is also supported.

Shared I/O Using Standards

PEX88000 switches enable the Virtual Functions (VFs) of SRIOV endpoints (such as SRIOV-enabled NVMe SSDs) and multifunction devices to be assigned to different physical hosts connected to the PCIe Fabric topology. Those hosts can enumerate their assigned functions using standard BIOS and OS software and use them with unmodified, vendor-supplied drivers.

Applications

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

Server and Storage CPU to I/O Connectivity

PEX88000 can be used to fan-out host PCIe ports to connect to a large number of I/Os or other subsystems in servers and storage systems. No software is required in this application. Figure 1 illustrates some simple and common applications of the PCIe switch.

NVMe JBOF

PCIe is broadly used in SAS-based storage subsystems and NVMe JBOFs. PEX88000 has been purpose-built to support NVMe All Flash Array (AFA) and hybrid (HDD/NVMe) systems. The embedded CPU in the switch provides capabilities to manage device configuration, chassis management, LED control, hot add/remove, and many other essential functions. The topologies in Figure 2 illustrate support for up to 32 x2 duel-port or 16 x4 single-port NVMe drives with two x16 links to hosts/CPUs.

HPC, Machine Learning, Artificial Intelligence, and I/O Sharing

HPC clusters are made up of high-performance processing elements that communicate through high bandwidth, low-latency pathways to support applications such as medical imaging, financial trading, data analytics, image processing, machine learning, and so on. PCIe is now broadly used in machine leaning and artificial intelligence applications to interconnect GPUs, FPGAs, Accelerators, and NICs. The PEX88000 family of products doubles the connection bandwidth between these processing elements compared to previous PCIe 3.0 devices. Clustering systems can be built with I/O sharing as an additional native capability when needed. The topologies in Figure 3 illustrate the use of PEX88000 switches in the above applications. An external management processor and software are required to implement these topologies.

Figure 1: Simple and Common PCIe Switch Applications

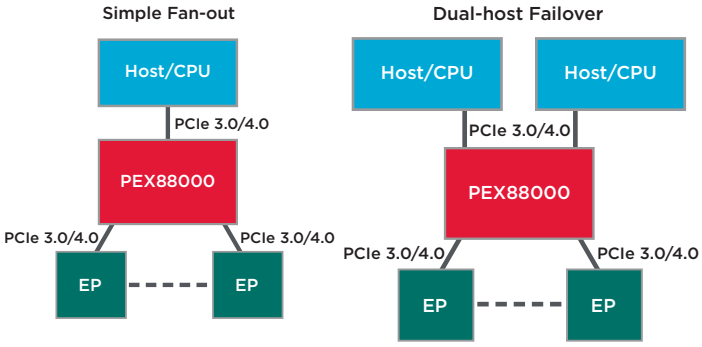


Figure 2: NVMe JBOF topologies

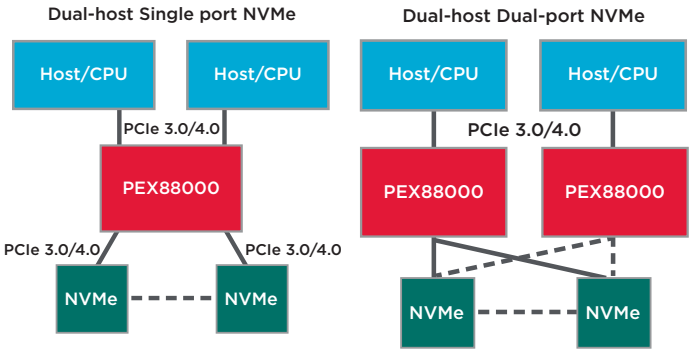
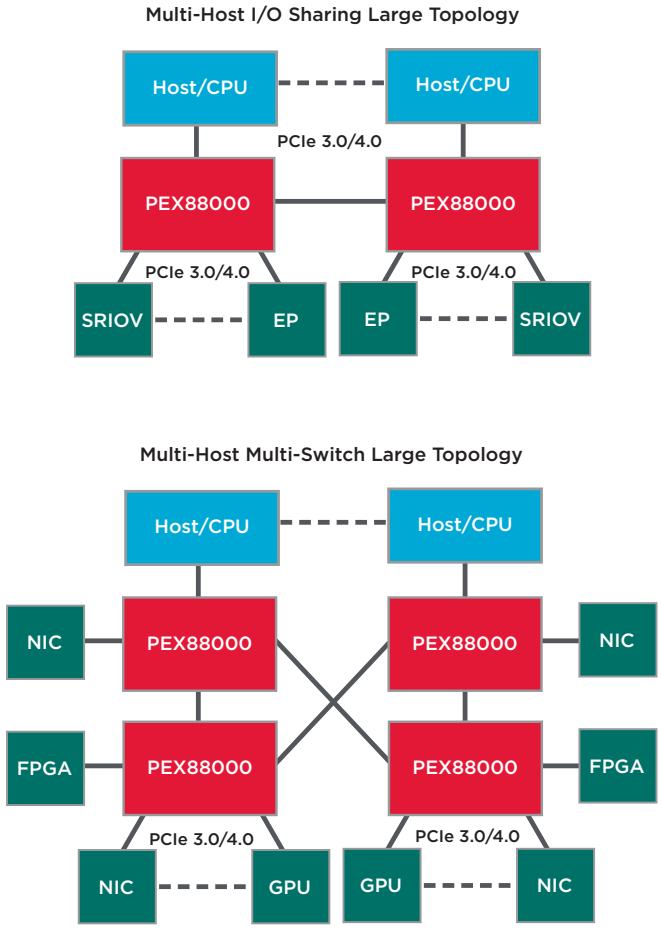


Figure 3: PEX88000 Switches in HPC, Machine Learning, Artificial Intelligence, and I/O Sharing Applications



Software Development Kit (SDK) and Software Packages

All PCIe switch and bridge products come with Broadcom SDK that includes drivers, source code, and GUI interfaces to aid in configuring and debugging. Various software drivers and APIs are provided to help implement dynamic allocation of I/Os to hosts, hot add/remove, chassis management, LED management, error handling, and other key functions as part of SDK. Additionally, software packages are available for NVMe JBOF, I/O Sharing, and complex multi-host multiswitch topologies through third-party vendors.

Acronym Guide

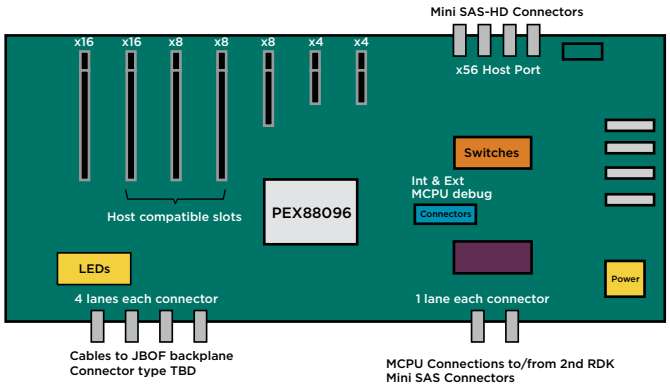
DMA	Direct Memory Access
HPC	Hot-Plug Controllers
TWC	Tunneled Window Connection
SSC	Spread Spectrum Clock Isolation
MSI-X	Message Signaled Interrupts
SRIS	Separate Refclk Independent SSC Architecture
DPC	Downstream Port Containment
eDPC	Enhanced DPC
Temperature Range	0°C to +70°C

Product Ordering Information

Standard Part No.	Description
SS02-0B00-00	PEX88096B0-DB, 98-lane PCIe 4.0 Switch
SS03-0B00-00	PEX88080B0-DB, 82-lane PCIe 4.0 Switch
SS04-0B00-00	PEX88064B0-DB, 66-lane PCIe 4.0 Switch
SS05-0B00-00	PEX88048B0-DB, 50-lane PCIe 4.0 Switch
SS06-0B00-00	PEX88032B0-DB, 34-lane PCIe 4.0 Switch
SS07-0B00-00	PEX88024B0-DB, 26-lane PCIe 4.0 Switch
SS08-0B00-00	PEX88T32, 16 up and 16 down retimer
Secure Part No.	**Description**
SS02-0B00-02	PEX88096B0-DB, 98-lane PCIe 4.0 Switch with Secure Boot enabled
SS03-0B00-02	PEX88080B0-DB, 82-lane PCIe 4.0 Switch with Secure Boot enabled
SS04-0B00-02	PEX88064B0-DB, 66-lane PCIe 4.0 Switch with Secure Boot enabled
SS05-0B00-02	PEX88048B0-DB, 50-lane PCIe 4.0 Switch with Secure Boot enabled
SS06-0B00-02	PEX88032B0-DB, 34-lane PCIe 4.0 Switch with Secure Boot enabled
SS07-0B00-02	PEX88024B0-DB, 26-lane PCIe 4.0 Switch with Secure Boot enabled

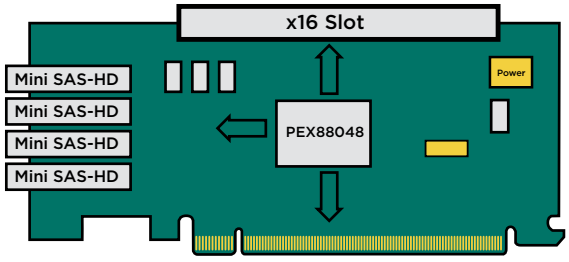
PEX88096B0 RDK

This eval kit would allow system designers to test and evaluate PEX88096 in the desired configuration. This RDK can be connected to a server through mini-SAS HD connectors and cascaded using slim-line connectors to create large topologies. Each RDK has seven standard PCIe ports for I/O devices for interop or performance testing.



PEX88048B0 RDK

This eval kit would allow system designers to either connect a server to PEX88096B0 RDK, use it independently to connect to another subsystem through its mini-SAS-HD ports, or connect to an I/O device using the PCIe slot available on the board.



BROADCOM®
connecting everything®

Visit the Broadcom Server Storage website at: www.broadcom.com/products/storage

BC-0484EN July 17, 2019