Product Brief

PEX89000 Series

Managed PCI Express 5.0 Switches Ranging from 144 to 24 Lanes

Machine Learning and Artificial Intelligence Systems Using PCIe

The Broadcom® PEX9700/8700/88000 series of switches running at PCIe Gen 3.0 and Gen 4.0 speeds are broadly used in ML/AI and storage applications today. Broadcom is now introducing the PEX89000 family of PCIe Gen 5.0 (32 GT/s) switches, allowing 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 supporting ML/AI and Server/Storage applications.

The PEX89000 switches offer up to 1024 Gb/s (128 GB/s) raw bandwidth per port (x16). The largest PEX89000 switch with 144 PCIe Gen 5.0 lanes allows user to achieve up to 9,216 Tb/s (1,152 GB/s) of raw bandwidth through the device.

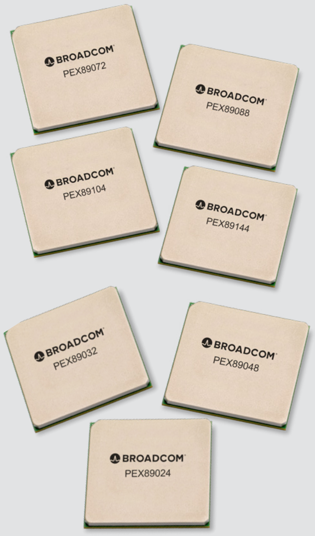
The PEX89000 switch series enables designers to:

- Create basic switch topologies at PCIe 5.0 rate (32 GT/s) 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 hyper-converged, AI/ML/DL applications.

Multi-Host Connectivity

Typically, PCIe connectivity allows single-host tree topology where all I/O devices connected to the switch are allocated to a single host. With multi-host functionality introduced by Broadcom in early 2010, users can connect multiple host devices to a PCIe switch and allocate I/O devices to different hosts. Broadcom switches, such as PEX89000, support multi-host connectivity and also allow dynamic removal, addition, and allocation of the I/O devices from one host to the other.

The PEX89000 switches incorporate advanced security features such as attestation and hardware secure boot. Hardware secure boot, which permits only authenticated firmware to execute, enables a controller to boot from an Internal Boot ROM (IBR) to establish the initial Root of Trust (RoT). Hardware secure boot authenticates and builds a Chain of Trust (CoT) with succeeding software using this RoT (Implicit Trust). The PEX89000 also supports attestation, the next generation in security (Explicit Trust).



Key Features

- Available in 144, 104, 88, 72, 48, 32, and 24-lane configurations
- On-chip best-in-class Broadcom 32-GT/s SerDes
- Each port speed (Gen1/2/3/4/5) independent of others
- Choice of link width: x1, x2, x4, x8, or x16
- On-chip PCIe analyzer with GUI support
- Embedded ARM CPU for management
- ExpressFabric™ PCIe switching architecture
- Sharing I/Os among multiple hosts
- Any port can be a host port or downstream (device) port
- Works with standard PCIe endpoints, hosts, and software
- MSI-X support

Key Features (cont.)

- Allows flexible fabric topologies
- 8 non-transparent bridging (NTB) ports
- Embedded MPT endpoint
- Low-power SerDes with PCIe sleep/ power-saving modes
- HW Secure Boot and Attestation

Key Advantages

- Device-specific relaxed ordering
- Port reconfiguration without impacting other ports
- Configurable with serial EEPROM, embeddedCPU, and/or host software
- Designate any port as the upstream port
- Standards compliant PCI Express base specification: r5.0, r4.0, r3.0, r2.0, and r1.0
- PCI power management spec r1.2
- Full line rate on all ports
- Cut-through packet latency of less than 115 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 I^{2}C
- SSC isolation on all ports
- SRIS/SRNS support
- ECRC and poison bit support
- Port status bits and GPIO available

Enhanced Non-Transparent Bridging 2.0 (NT2.0)

PEX89000 switches are equipped with the 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 PEX89000 switch (144-lane device) is equipped with eight NT2.0-capable ports, enabling a large number of hosts/servers to be connected to the switch.

Embedded ARM CPU

Each PEX89000 PCIe switch is equipped with an embedded ARM 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 allocation, I/O management, Hot add/remove, and interrupt handling.

Shared I/O Using Standards

PEX89000 switches enable mapping or assignment of the Virtual Functions (VFs) of SR-IOV endpoints (such as NVMe SSDs, NICs, GPGPUs) and multifunction devices to the host/s. Customers can use the SDK to allow sharing of VFs or PFs among multiple hosts. Once assigned, hosts can enumerate their assigned functions using standard BIOS and OS software.

Switch Operation Modes

PEX89000 switches can be configured in two modes – Base mode and Synthetic mode.

- Base mode: The switch functions without any firmware involvement. In this mode, the embedded CPU is disabled and the device will operate as a standard PCIe fan-out switch. The switch may be programmed to provide PCIe fanout using a set of chassis management capabilities with MPT endpoints embedded in the switch.
- Synthetic Mode: In this mode, the embedded CPU becomes the host and allocates I/O devices and internal resources to external host devices connected to the switch in single or multi-host environment. The embedded CPU synthesizes the hierarchy for each connected host based on firmware loaded in the embedded RAM.

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. The solution offers an innovative approach to set up and control the PEX89000 switches, configure the routing tables, handle errors, Hot-Plug events, and enable the solution using an embedded CPU without impacting data flowing through the switch.

Flexible Topologies

PEX89000 switches eliminate the topology restrictions of PCIe. The switch allows multiple hosts to connect to a single or multiple PCIe switch complex to enable topologies for hyper-scale systems.

Downstream Port Containment (DPC/eDPC)

Most servers have difficulty handling serious errors in I/O devices, especially when a device disappears from the system. PEX89000 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. SRIS (Separate Refclk Independent SSC Architecture) and other clocks such as SRNS and constant CLK are also supported.

Debug, Bring-up, and Monitoring

Broadcom PCIe switches support a rich set of features for debugging the system in initial bring-up and monitoring the performance during run time. These features include on-chip PCIe analyzer with GUI support, packet generation, eye-scope, error monitoring, port utilization, error counting, loop-back, header, and TLP logging. Additional features for telemetry applications are also being included.

Applications

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

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 machine learning, artificial intelligence, medical imaging, financial trading, data analytics, image processing, and so on. Broadcom PCIe switches are broadly used in machine learning and artificial intelligence applications to interconnect GPUs, FPGAs, Accelerators, NVMe SSDs, and NICs. The PEX89000 switch family doubles the connection bandwidth between these processing elements compared to previous PCIe 4.0 devices.

Composable systems can be built using PEX89000 switches with pools of compute, storage, and networking resources that can be dynamically configured and allocated to applications or clients based on service level agreements in public and private Cloud Computing environment.

NVMe JBOF

PCIe is broadly used in SAS-based storage subsystems and NVMe JBOFs. The PEX89000 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 for this application.

Server and Storage CPU to I/O Connectivity

PEX89000 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.

Software Development Kit (SDK) and Software Packages

All PEX89000 PCIe switch and bridge products come with the Broadcom SDK that includes drivers, source code, APIs and GUI interfaces to aid in configuring, debugging and running switches in the lab and the field. Software drivers and APIs are provided to help customize dynamic allocation of I/Os to hosts, hot add/remove, chassis management, LED management, error handling, and other key functions. Additionally, software packages are available for NVMe JBOF and complex multi-host multiswitch topologies through third-party vendors.

PEX89144 RDK

This evaluation kit would allow system designers to test and evaluate PEX89000 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 interoperability and performance testing.

PEX89144 HIB

This host interface board (HIB) would allow system designers to either connect a server to PEX89144 RDK through cable or use it independently to connect to an I/O device using the PCIe slot available on the board.

Product Ordering Information

Manufacturing Part Number		
Secure Part Number	Non-Secure Part Number	Description
SS26-0B00-02	SS26-0B00-03	PEX89144, 144-lane PCIe 5.0 Switch
SS24-0B00-02	SS24-0B00-03	PEX89104, 104-lane PCIe 5.0 Switch
SS23-0B00-02	SS23-0B00-03	PEX89088, 88-lane PCIe 5.0 Switch
SS22-0B00-02	SS22-0B00-03	PEX89072, 72-lane PCIe 5.0 Switch
SS29-0A00-02	SS29-0A00-03	PEX89048, 48-lane PCIe 5.0 Switch
SS28-0A00-02	SS28-0A00-03	PEX89032, 32-lane PCIe 5.0 Switch
SS27-0A00-02	SS27-0A00-03	PEX89024, 24-lane PCIe 5.0 Switch

BROADCOM®
connecting everything®

For more product information: broadcom.com