Dolphin Interconnect Solutions

Transparent PCIe Hot-Add

Whitepaper

Dolphin Engineering

5/8/2019

DISCLAIMER

DOLPHIN INTERCONNECT SOLUTIONS RESERVES THE RIGHT TO MAKE CHANGES, WITHOUT FURTHER NOTICE, TO ANY OF ITS PRODUCTS AND DOCUMENTATION TO IMPROVE RELIABILITY, FUNCTION, OR DESIGN. DOLPHIN INTERCONNECT SOLUTIONS DOES NOT ASSUME ANY LIABILITY ARISING OUT OF THE APPLICATION OR USE OF ANY PRODUCT OR DOCUMENT.

Notes

This document is based on information available at the time of publication. While efforts have been made to ensure accuracy, the information contained herein does not purport to cover all details or variations in hardware and software.

Table of Contents

Introduction

PCI Express (PCIe) is the dominating technology used to connect various types of networking, storage, FPGA and GPGPU boards to servers and desktop systems. Normally, these add-on boards are placed into PCIe slots residing on the server baseboard. In cases where some distance between the add-on board and server is needed, or the number of add-on boards exceeds the number of free slots, or where the add-on board is a physically larger storage or computer unit, a transparent PCI Expansion system with cabled PCIe can be used. Cabled PCIe extends up to 9 meters using copper cables and up to 100 meters using PCIe fiber cables.

No special software is required when using Transparent PCIe expansion systems. Just place the add-on boards in the expansion system, power it on, and boot the server. The server BIOS will automatically enumerate the PCIe sub-system and all add-on boards will appear as if they were installed inside the servers.

This works great if the IO Expansion system is powered on before the server is powered on and booted. However, if the server boots before the IO Expansion system, the IO devices will not be detected by the BIOS and will not be available until the server is booted again. Modern operating systems and BIOS' do not support general PCIe hot-add operations.

Similarly, it is not possible to add or remove add-on boards from the expansion system without first powering down the system, adding or removing the boards, and then rebooting the server.

Dolphin eXpressWare

Dolphin has, over the course of more than two decades, developed a robust and rich software infrastructure for PCIe systems. Our software enables multiple servers connected by a PCIe fabric to communicate at native PCIe ultra-low latencies and x16 throughput. Our software includes standard TCP/IP drivers, SuperSockets, a socket accelerator compliant to Berkeley Sockets, and a remote memory programming interface named SISCI.

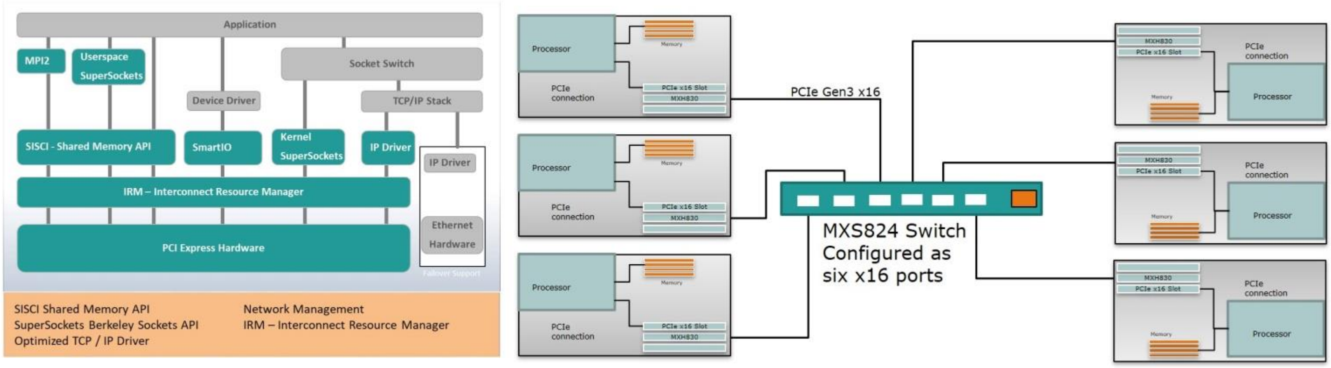


Figure 1: eXpressWare NTB s/w Suite and six node PCIe cluster

SmartIO API

The latest addition to Dolphin's eXpressWare software suite is a new concept called SmartIO. The main goal behind SmartIO has been to add flexibility to standard transparent PCIe IO, enabling standard IO devices to be shared and accessed from severs connected to the PCIe fabric. The latest extension to the SmartIO software is the ability to surprise hot-add an IO device or a complete PCIe expansion system with many PCIe IO boards. This functionality does not require any special BIOS or BIOS setting and works with most modern Linux kernels without any special patches.

This is made possible by eXpressWare software enumeration that takes place when the eXpressWare drivers are loaded or when new devices are detected. The solution requires that one of Dolphin's standard PCIe NTB enabled cards is installed in the server. There are no other special requirements for the PCIe sub-system.

Installing eXpressWare

Download the appropriate Linux eXpressWare installer from www.dolphinics.com. Please carefully follow the instructions found in the "getting started" guide shipped with each NTB adapter to complete the registration, and the login credentials will be automatically emailed to you.

Run the Dolphin installer with the command line options --enable-smartio --install-node.

No other eXpressWare modules are required.

Hardware setup

1. Install a supported Dolphin NTB adapter in the host. The hot add functionality is supported with MXH830, PXH810, PXH820, PXH824, PXH830, PXH840 adapter cards.
2. Connect the host adapter to the end-point, desired IO board, appliance or expansion system. If the configuration includes an MXS824 PCIe switch, ensure that the switch configuration is set to transparent mode. All standard PCIe IO systems and devices are supported.

The figures below show three different typical configurations:

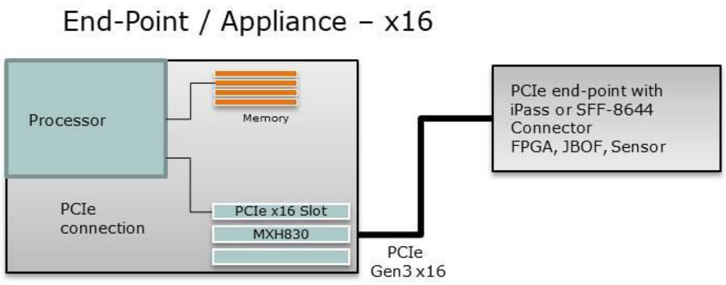


Figure 2 Directly attached end-point, IO board (FPGA),Unit Under Test, or appliance (JBOF)

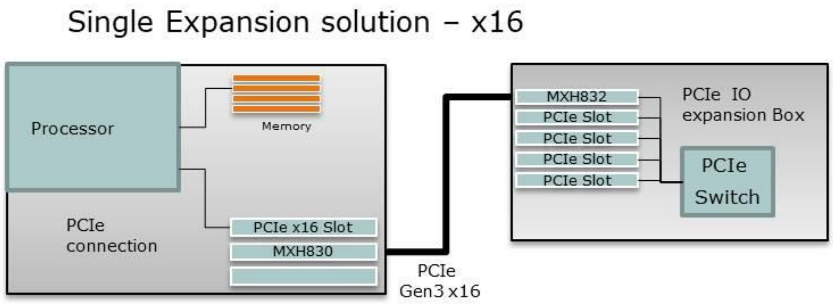


Figure 3 Expansion system with multiple PCIe expansion slots for IO boards (NVMe, GPU, FPGA, NIC)

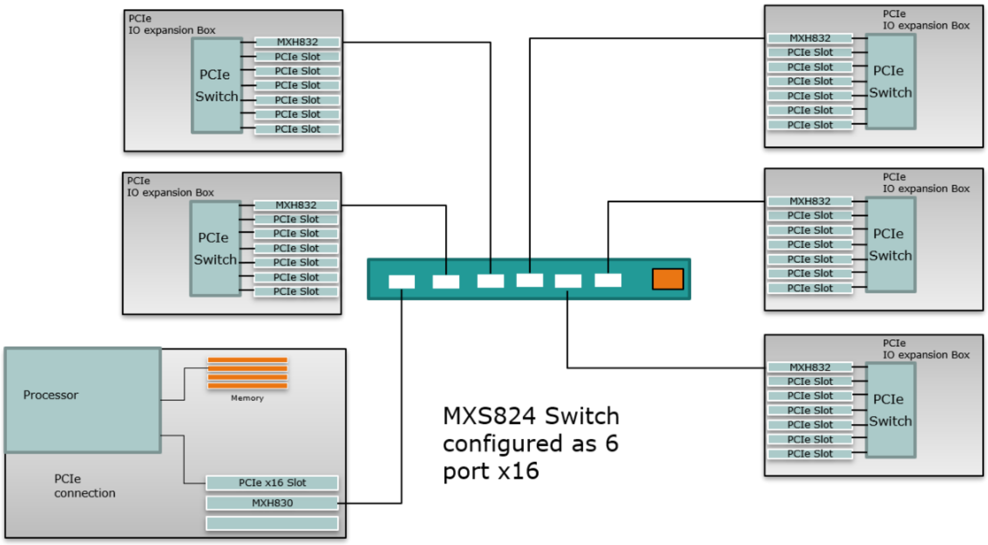
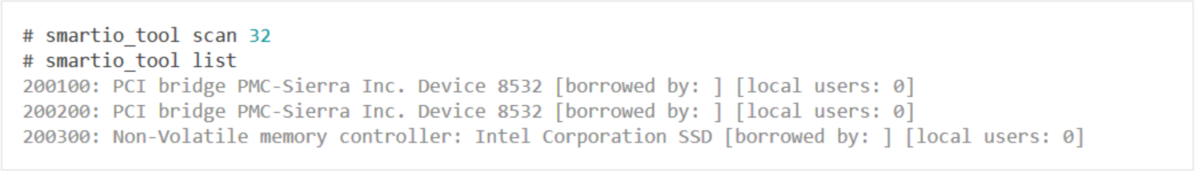


Figure 4 One server connected to MXS824 and 5 Expansion.

Hot-adding devices

SmartIO Hot-Add eliminates power on sequencing requirements. End-points and expansion systems may have their cables connected and power applied independent of the host system(s) or switch(s) on the fabric. This is ideal for hot-adding standard PCIe NVMe, JBOD, FPGA, units-under-test or any PCIe appliance without rebooting the host system.

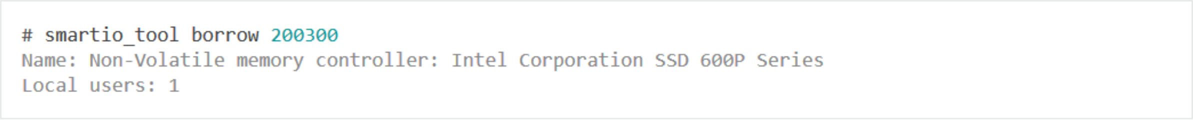
Initially, devices behind an NTB must be discovered by issuing a scan command with the smartio_tool command line utility. This command requires that a nodeId is specified. This nodeId must be different from the host's or other nodeId:



```
# smartio_tool scan 32
# smartio_tool list
200100: PCI bridge PMC-Sierra Inc. Device 8532 [borrowed by: ] [local users: 0]
200200: PCI bridge PMC-Sierra Inc. Device 8532 [borrowed by: ] [local users: 0]
200300: Non-Volatile memory controller: Intel Corporation SSD [borrowed by: ] [local users: 0]
```

Figure 5: Smartio_tool scan

After the scan, the devices are available for borrowing:



```
# smartio_tool borrow 200300
Name: Non-Volatile memory controller: Intel Corporation SSD 600P Series
Local users: 1
```

Figure 6: Smartio_tool borrow

If you would like to always borrow all devices after a scan, you can enable the automatic borrow feature. Just run the command “smartio_tool autoborrow_transparent 1” before the scan command.

Using the devices

After hot adding a device, it can be used as usual. There are no special requirements, just ensure the standard device driver for the device is loaded. There is no performance penalty, the PCIe transactions are routed as usual between the system root complex and device.

Removing and adding devices

If you need to remove a device, e.g. an NVMe, you should ensure the device is not in use by an application. An NVMe should be unmounted. If device supports hot removal with power on, just unplug it and do another smartio_tool scan as shown in Figure 5: Smartio_tool scan. If you later need to add it back, just repeat the scan operation. Adding and removing devices will not interrupt communication with other devices in use.

System requirements

If the add-on board requires large PCIe BARs, you may need to increase the Dolphin NTB board prefetch space. The sum of PCIe BAR sizes + natural alignment for all added devices must be smaller than the prefetch space allocated by the Dolphin NTB board.

Availability

The SmartIO hot-add functionality is available with the MX830, PXH810, PXH820, PXH830 and PXH840 cards using eXpressWare 5.12.0 or higher.

The hot-add functionality requires Linux kernel 3.10.0-514 or newer. The software is being qualified for CentOS 7.5 and Ubuntu 18.04.

Please carefully consult the software release notes and eXpressWare installation and usage manuals for details on installation, configurations, and limitations.

Roadmap and future plans

eXpressWare 6.0 will add support for multiple hosts and multiple add-on boards being selectively added to an arbitrary host, dynamic add-on board re-assignment.

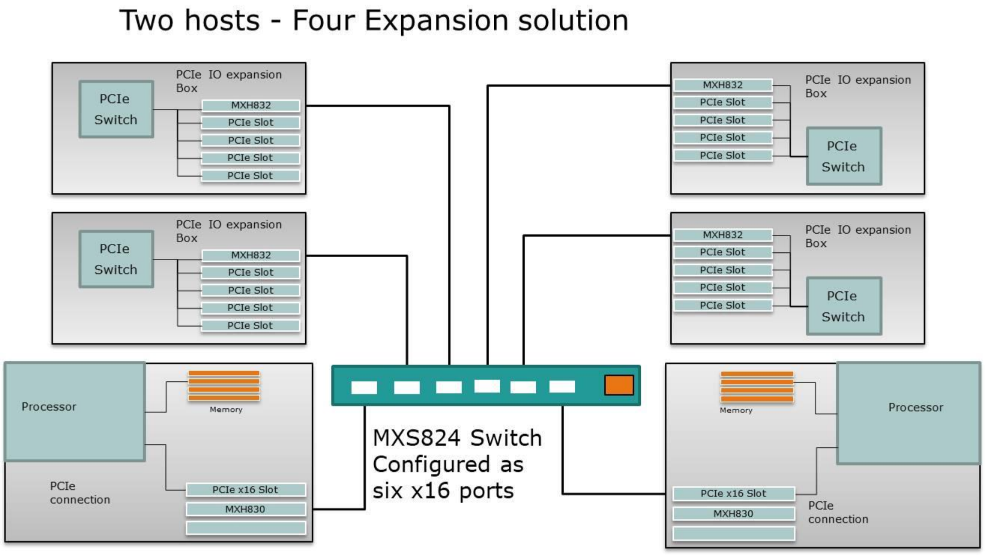


Figure 7 Multiple Servers and multiple IO expansion boxes using MXS824 to fan.

Windows will be supported with the new PCIe Gen4 MXH930 card. Please contact Dolphin for availability.

Reference and more information

Please visit www.dolphinics.com for additional information on products and solutions.

Additional information on Dolphin eXpressWare SmartIO can be found at https://www.dolphinics.com/smartio.html

Additional white papers on the Dolphin Express technology are currently available from http://www.dolphinics.com/support/whitepapers.html :

Whitepaper	Description
Flexible Device Sharing using Device Lending	How to access PCIe devices over a PCIe fabric. Virtualization and peer to peer transactions
PCI Express Device Lending	PCI Express Device Lending - borrow PCIe devices from remote Linux systems
Dolphin SuperSockets for Windows	Learn how Dolphin SuperSockets works on Windows platforms
Dolphin SuperSockets for Linux	Learn how Dolphin SuperSockets works on Linux platforms
Dolphin Reflective Memory Solution	Dolphin's high speed, low latency PCI Express reflective memory solution
PCI Express Peer to Peer Communication	PCI Express peer to peer communication solution made easy
Dolphin Shared Memory SISCI API	Dolphin SISCI API provides a high speed, shared memory solution for PCI Express
PCIe Fabric Hardware Architecture Part 1	Curtiss Wright white paper on using PCIe Fabrics with VPX single board computers part 1
PCIe Fabric Hardware Architecture Part 2	Curtiss Wright white paper on using PCIe Fabrics with VPX single board computers part 2

Please contact pci-support@dolphinics.com with any questions.