Dolphin Interconnect Solutions

Dolphin Express IX Reflective Memory / Multicast

Whitepaper

Dolphin Engineering
6/19/2013

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 DOCUMENTS.

Table of Contents

Introduction

The Dolphin Express IX product family supports multicast operations as defined by the PCI Express Base Specification 2.1. Dolphin has integrated support for this functionality into the SISCI API specification.

The functionality of PCI Express multicast is to enable a single bus write transaction to be sent to multiple remote targets or in PCI Express technical terms **-** multicast capability enables a single TLP to be forwarded to multiple destinations.

The SISCI API enables customers to easily implement applications to directly access and utilize the reflective memory functionality without the need to write device drivers or spend time on studying PCI Express chipset specifications.

Dolphin reflective memory benchmarks included in the SISCI developers kit show end to end latencies as low as 0.99us and over 2,000 MegaBytes/sec dataflow at the application level.

Multicast implemented in hardware

Reflective memory systems (in computer literature also referred to as mirror memory systems, replicated shared memory, multicast or replicated memory systems) implements transparent and automatic updates of remote memory areas. Reflective memory is typically mapped into an embedded system application and enables similar applications on other nodes to share updated data without involving any traditional networking protocol and overhead. Data of any size is transmitted to all nodes directly by functionality implemented in hardware.

Typical applications can range from a two-node fail over pair to large DSM applications like aircraft, ship and submarine simulators, automated testing systems, industrial automation, control, online and high-speed data acquisition and distribution. Because of their inherent replication they are especially good for fault tolerance.

Traditional reflective memory

Other solutions typically implement reflective memory by providing a plug-in adapter card with onboard device memory. Applications can write to this memory and the data is automatically forwarded through to all other nodes connected. When data is needed, the applications need to read from the local board device memory. A typical 4 node configuration can be seen in the figure below.

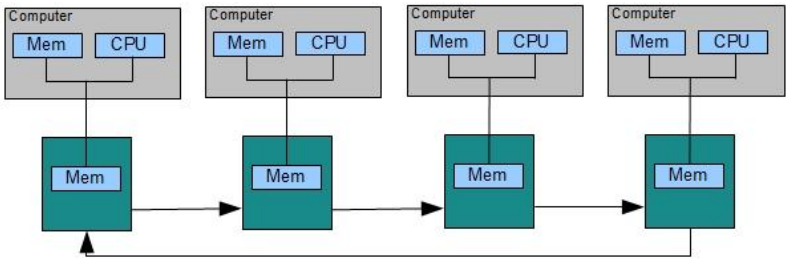

Figure 1 : Alternative types of reflective memory implementation

Dolphin Express reflective memory

The Dolphin solution is unique as it is able to utilize the computer system's standard main memory. This, combined with regular PCI Express technology running at wire speeds at 40Gbps gives significant performance improvements.

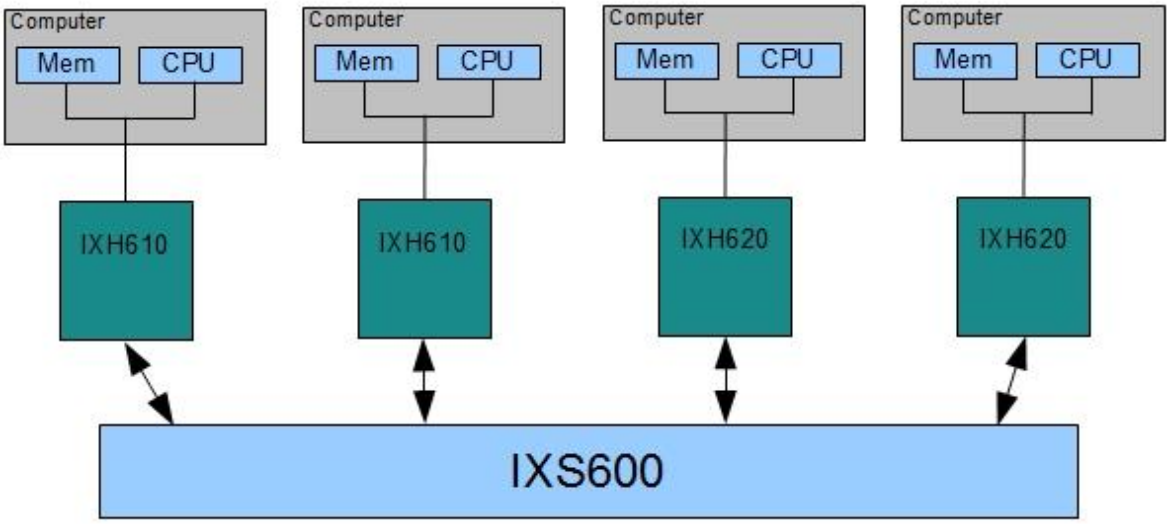

Figure 2 Dolphin Express IX reflective memory setup

The figure above visualizes a typical Dolphin Express setup. The Dolphin IXH610/620 card does not have any memory used for storing reflective memory data. This brings both significant performance and cost benefits. The IXS600 switch provides a mechanism for simultaneous multi-cast of data to all connected ports with a measured port to port latency less than 200 nanoseconds.

Transmitting data

Data can be written to other nodes using the reflective memory solution in the following ways:

- CPU: Data can be sent to reflective memory using one or more CPU posted write instructions. Using SISCI, applications can just do a standard memcopy() using the reflective memory as a target or just do a regular pointer assignment. The fully hardware based memory mapped data transmission does not rely on any operating system service or kernel driver functionality and provides the best possible deterministic data transmission latency and jitter.
- DMA: The Dolphin Express IX adapter card includes an efficient scatter / gather DMA engine that can be engaged to send small or larger amounts of data to reflective memory.
- PCIe device: Starting with the DIS 4.3.0 software release from Dolphin, customers can use the SISCI API to configure and enable GPUs, FPGAs etc (any PCIe master device) to send data directly to reflective memory. (Avoiding the need to first store the data in local memory).

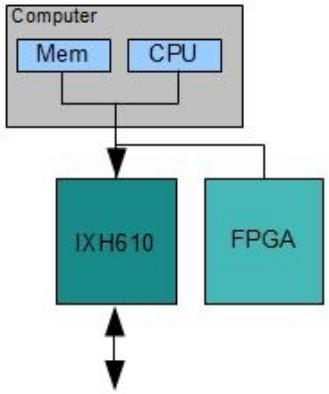

Figure 3: FPGA direct transmission

Data will be transmitted to all remote nodes. If a local reflective memory update is needed, application programmers need to copy the sent data to the local buffer as well. This is a very low cost operation as the data is already in the CPU cache.

The PCI Express based reflective memory solutions provides significant improvements over alternative solutions:

- Data in main memory: The Dolphin Express IX reflective memory solutions utilize main memory to store data. This has several significant benefits:
    - Reading data in main memory is significantly faster than solutions storing data in specialized PCIe device memory located in the computer IO system.
    - Main memory is cached: This means that the solution will benefit from the standard CPU cache when reading data. Reflective memory updates from remote will automatically invalidate the CPU cache and ensure full data consistency.
    - Specialized device memory is normally very expensive vs main memory modules.
    - You don't need to specify the reflective memory size when buying hardware. The size of Dolphin Express IX reflective memory is user configurable – a property set by the application during initialization of the system.
- Data is multicasted by a centralized switch. Each IXS600 switch will send data out on all connected ports simultaneously. This means that data will be received by all nodes virtually simultaneously when connected to a single switch. When multiple switches are used, each switch hop will add less than 200 nanoseconds delay to the distribution of the data. Alternative solutions using a ring topology to distribute data may have significant delays between when the first and the last node in the network receives the data. The minimal delay introduced by Dolphin Express IX reflective memory enables real-time applications to benefit from a significantly reduced total communication time.
- Hardware based CRC and retransmission. PCI Express implements a reliable data transmission by calculating a CRC for every data packet. Correctable link errors will automatically cause a hardware retransmit.
- Fair arbitration and sharing of bandwidth. Hard real-time systems should normally be configured to avoid narrow bottlenecks in the network. PCI Express uses a fair, round robin allocation of resources and provides a very deterministic data transmission even under maximum load.

Multicast groups

Currently up to 4 independent global multicast groups are supported. This enables SISCI programs to operate up to 4 independent reflective memory regions and control which nodes that will receive the multicast data. The default size of each multicast group is 2 Megabytes. As main memory is used, larger segments, up to 128 Megabytes or larger can be configured, please contact Dolphin support for instructions.

Performance

The Dolphin IXH610 adapter and IXS600 switch utilizes standard x8 PCI Express link enabling customer applications to take advantage of the exceptional 40Gb/s link bandwidth.

Dolphin reflective benchmarks included in the SISCI developer's kit can be used to measure the reflective memory performance of your system. The actual performance will slightly vary dependent on the computers IO system, but typically you should expect end to end latencies as low as 0.99us and over 2,650 Mega Bytes per second dataflow at the application level as shown on the figure below.

The SISCI reflective memory example 'reflective_bench' can be used to measure the throughput vs message block size. The program is included in the Dolphin software distribution package.

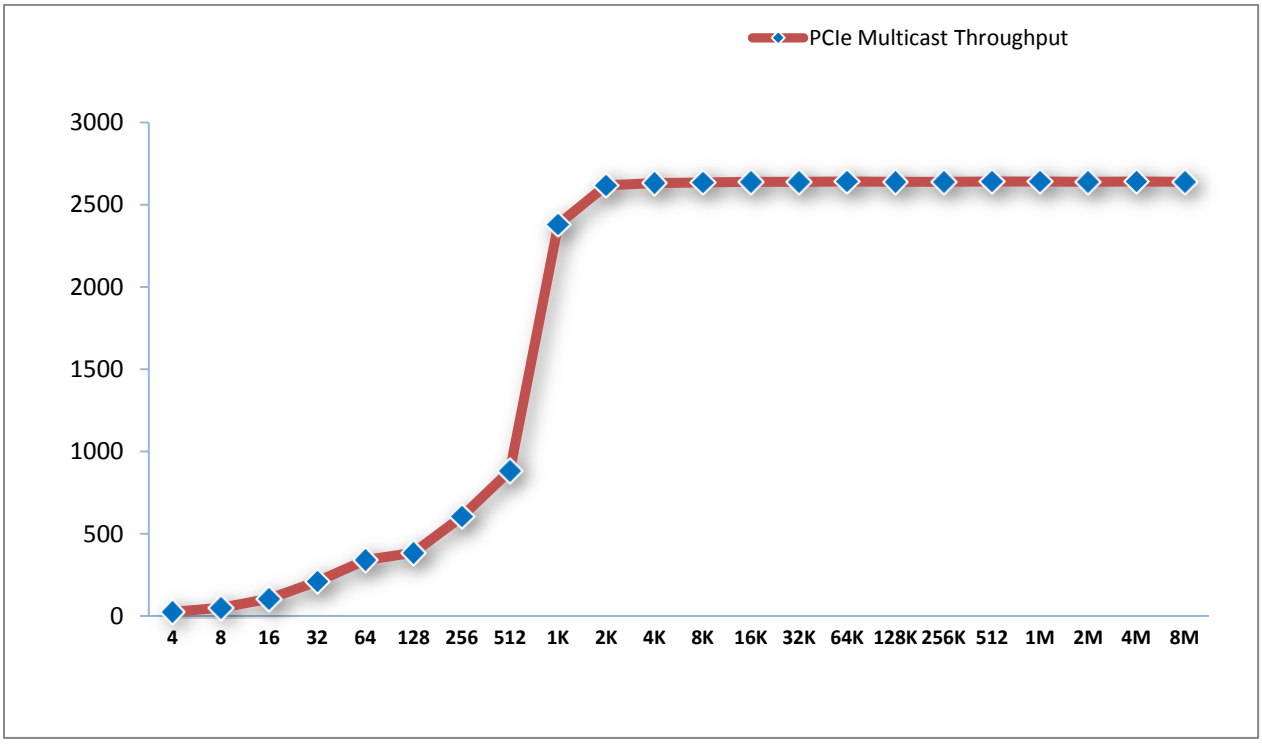

Figure 4: Reflective_bench results

Hardware configuration and installation

Each node has a Dolphin Express IXH610, IXH611 or Dolphin Express IXH620 XMC adapter card in NTB mode installed. Up to 8 systems can be connected to the Dolphin IXS600 8 port PCI Express Gen2 switch. More switches can be cascaded to create larger reflective memory systems, initially up to 20 nodes. Up to 56 nodes are supported for pure reflective memory functionality – general SISCI support is limited to 20 nodes. Please refer to the actual software release note for configuration details. The reflective memory functionality is only available when IXS600 switch is connected; two adapter cards can communicate using a direct cable using the standard SISCI unicast functionality (write to only one remote node).

The Dolphin software for Dolphin Express IX release 4.3.1 or newer should be installed on all computers in the network. The functionality is available through the SISCI API using Linux, Windows or RTX operating systems. The nodes can run any of the above operating systems – inter-communication between Linux, Windows and RTX is fully supported.

Customers that would like to use Dolphin Express reflective memory to distribute data from a PCIe attached GPUs, FPGAs etc can simply attach the PCIe device to a regular PCI Express slot to any of the computers. Additional information can be found in the ‘reflective_device.c’ example program included in the Dolphin software distribution package.

SISCI API

The SISCI API (Software Infrastructure Shared-Memory Cluster Interconnect) consists of driver and API software, tools, documentation and source needed to develop your own embedded application utilizing the low latency and high performance of a PCI Express Cluster. The SISCI API provides a C system call interface to ease customer integration of PCI Express over cable solutions.

SISCI enables customer applications to easily and safely bypass the limitations of traditional network solutions, avoiding time consuming operating system calls, and network protocol software overhead. SISCI resources (memory maps, DMA engines, Interrupts etc) are identified by assigned IDs and managed by a resource manager enabling portability and independent applications to run concurrently on the same system.

The SISCI API has been defined in the European Esprit project 23174 as a de facto industry standard Application Programming Interface (API) for shared memory based clustering.

In addition to the reflective memory/multicast functionality, the SISCI API provides functionality to access remote memory for unicast (single remote read or write), Direct Remote DMA (RDMA) using the onboard DMA engine. The API also includes support for sending and receiving remote interrupts and error checking. SISCI also support PCIe peer to peer communication over the PCIe cable.

Please consult the SISCI API reference manual for more details.

Reference and more information

Please visit www.dolphinics.com for additional information on the Dolphin Express IX product family.

Additional information including the online SISCI API reference manual can be found at http://www.dolphinics.com/products/embedded-sisci-developers-kit.html

Additional white papers on the Dolphin Express technology are currently available:

- SuperSockets for Linux
- SuperSockets for Windows
- Dolphin Express Reflective Memory / Mulicast (This document)
- Dolphin Express Peer to Peer communication – Direct PCIe

Please contact pci-support@dolphinics.com if you have any questions.