Dolphin Interconnect Solutions

Dolphin Express IX Reflective Memory / Multicast

Whitepaper

Dolphin Engineering

3/24/2015

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.

Notes

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

Trademarks

SCRAMNet is a registered trademark of Systran Corporation.
GE FANUC is a registered trademark of GE Fanuc Automation Inc.
Windows is a registered trademark of Microsoft Corporation.

Table of Contents

DISCLAIMER 1
Notes 1
Table of Contents 2
Introduction 3
Multicast implemented in hardware 3
Traditional reflective memory 3
PCI Express reflective memory 4
Multicast memory and multicast groups 4
Using PCI Express reflective memory 5
Transmitting data to reflective memory 5
Reading Data from reflective memory 6
Interrupts 6
Significant benefits provided by PCI Express 6
Performance 7
Hardware configuration and installation 8
Reflective Memory Comparison 8
Roadmap and future plans 9
SISCI API 9
SISCI API Code examples 9
Reference and more information 10

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 to make it easily available to application programmers. The combination of Dolphin PCI Express hardware and the SISCI API creates a solution for customers seeking multi-cast or reflective memory type functionality.

The first Dolphin Express product line was introduced in 1994 and has been followed by several generations of shared memory solutions. The Dolphin Express IX product is our second generation of interconnect products supporting a real hardware based multicast implementation. PCI Express multicast enables 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.

Dolphin combines PCI Express multicast with our SISCI API. The combination allows customers to easily implement applications that directly access and utilize PCI Express' reflective memory functionality. Now, applications can be built without the need to write device drivers or spend time studying PCI Express chipset specifications.

The advantage of the PCI Express reflective memory approach is lower latency and higher bandwidth. Dolphin benchmarks show end-to-end latencies as low as 0.99 micro seconds and over 2,650 Megabytes /sec dataflow at the application level. These benchmarks are included in the SISCI developer's kit. By using PCI Express based reflective memory functionality, customers can easily solve their real time, distributed computing performance requirements.

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) implement 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 automatically transmitted to all nodes directly by functionality implemented in hardware.

Typical applications can range from a two-node fail over pair to large distributed shared memory applications like aircraft, ship and submarine simulators, automated testing systems, industrial automation, electronic trading, 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 reflective memory type 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. Applications reads data from the local adapter card device memory. A ring network topology connects the systems together. A typical 4 node configuration can be seen in the figure below.

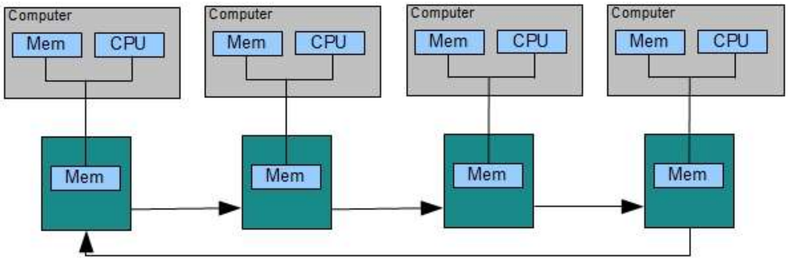


Figure 1 : Alternative types of reflective memory implementation

PCI 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 of 40Gbps gives significant performance improvements.

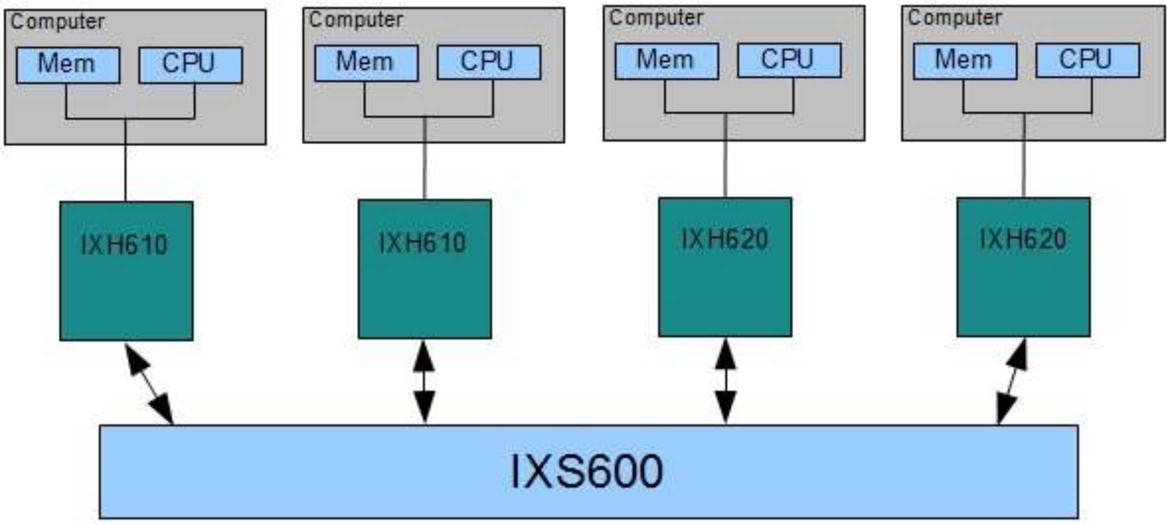


Figure 2 Dolphin Express IX reflective memory setup

The figure above visualizes a typical Dolphin Express setup. Dolphin IXH610 and IXH620 cards are connected through an IXS600 switch. Neither the IXH610 nor IXH620 card has any memory used for storing reflective memory data, resulting in 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.

Multicast memory and multicast groups

Dolphin Express IX supports up to 4 independent global multicast groups / memory segments. This enables SISCI programs to use up to 4 independent reflective memory regions and control which nodes receive the multicast data. This differs from other reflective memory solutions from other vendors which only support a single multicast group.

If a multicast group segment does not exist in a system, the multicast data will be silently dropped without any notification. Multicast data filtering is done by each connected adapter. Nodes can be rebooted and multicast segments can be added or removed at any time without any synchronization with the other nodes. Each multicast segment can be up to 2 Gigabytes with a total reflective memory size of 8 Gigabytes.

The current driver release - DIS 4.4.3 - supports a max segments size of 64 Megabytes and a total reflective memory size of 256 Megabytes. The Dolphin driver allocates the reflective memory segment memory from main memory during driver startup. The upcoming DIS 5.0 software distribution will support the 8 Gigabyte option.

A PC server with large PCI BARS is required to support the 8 Gigabyte option. If you would like to use reflective memory segments larger than 256 Megabytes, you should ask your system vendor to confirm the system BIOS supports memory mapped I/O above 4GB (large Base Address Register support per the PCIe specification).

Using PCI Express reflective memory

The major difference between traditional reflective memory solutions and PCI Express' approach to reflective memory is that the PCI Express solution utilizes two different addresses, one for reading and a different address for writing data. The SISCI API provides these addresses during initialization. The write address is inside the PCI Adapter address space. Any write to this address space will typically trigger an address translation inside the PCI adapter and cause PCIe transactions to be sent to the IXS600 switch and other nodes. The result of reading this address is undefined.

Transmitting data to reflective memory

Data can be transferred 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 the flexibility to do a standard memcopy() using the reflective memory as a target or 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.
- PCIe device: 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).
- Onboard 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. This functionality is available with the DIS 4.4.4 or newer software release from Dolphin.

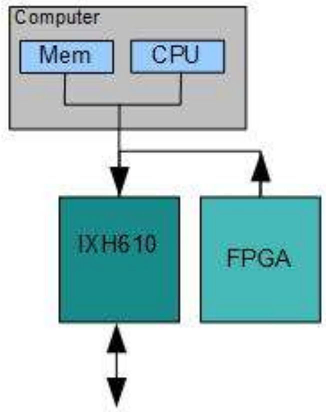
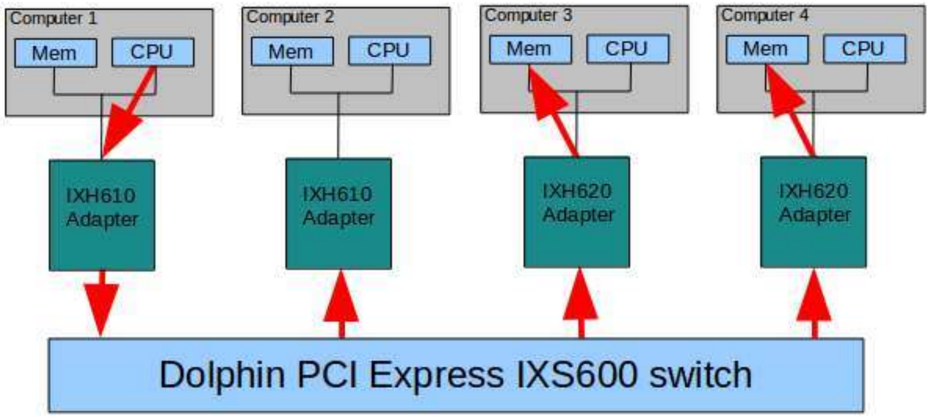


Figure 3: FPGA direct transmission

The figure below shows the flow of data (indicated by the red arrow) – from the CPU of computer 1 - to a local memory address allocated for a specific reflective memory group ID. Data will be transmitted by the PCI Express hardware into the main memory of all other nodes in the network that has allocated a reflective memory segment for the same group ID. All of this is easily managed through the SISCI API. In this example group ID includes computers 3 and 4, does not include computer 2.



Reading Data from reflective memory

To read data received from other nodes, the application needs to use the read address, this points to the allocated segment in local main memory.

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.

Interrupts

The SISCI API provides functionality to register and trigger application interrupt's into a remote node. Please consult the SISCI Users guide for details on using SISCI interrupts.

Significant benefits provided by PCI Express

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 multicast by a centralized switch.
    - Each IXS600 switch will send data out on all connected ports simultaneously. This means that all nodes will receive data 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 have significant delays between when the first and the last node in the network receives the data. Each node will typically introduce a fixed delay; the total delay in the network varies depending on the number of nodes.
- The minimal delay introduced by Dolphin Express IX reflective memory enables real-time applications to benefit from a significantly reduced total communication time – allowing the application to run at a faster simulation frequency or spend more time on computation.
- Dead nodes or unplugging cables will not stop the entire network; all nodes that remain connected to the network will be able to communicate without interruption.

- 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.

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.

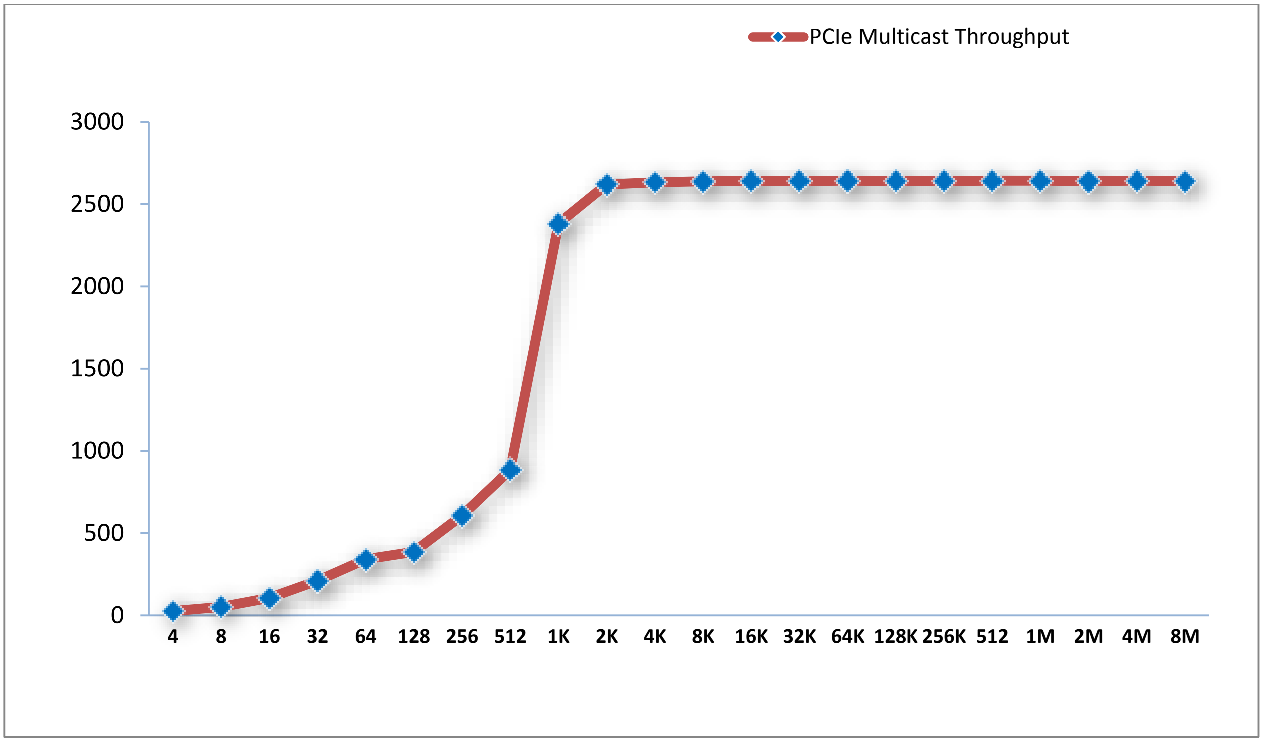


Figure 4: Reflective_bench results

Hardware configuration and installation

To create a reflective memory system with Dolphin products, each node must have a Dolphin Express IXH610/ IXH611 or Dolphin Express IXH620 XMC adapter card in NTB mode installed. A Dolphin IXS600 can be used to connect multiple systems. Up to 8 systems can be connected to a single IXS600 8 port PCI Express Gen2 switch. For larger reflective memory systems, IXS600 switches are cascaded to create larger topologies. Initially up to 20 nodes are supported with complete SISCI functionality. Up to 56 nodes are supported when limited to just reflective memory functionality. Please refer to the actual software release note for configuration details. The reflective memory functionality is only available when an 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).

SISCI reflective memory support is targeted at Linux, Windows or RTX operating systems. The nodes can be running any of the above operating systems and inter-communication between Linux, Windows and RTX systems is fully supported. VxWorks 6.9 support is planned for Q2 2015.

PCI Express reflective memory is not limited to server nodes. PCI Express devices are also supported. Data from attached GPUs, FPGAs etc can be distributed to multiple remote nodes simultaneously by simply attaching the PCIe device to a regular PCI Express slot in any of the computers. Additional information can be found in the 'reflective_device.c' example program included in the Dolphin software distribution package.

Reflective Memory Comparison

The various reflective memory systems available have different properties. Generally, PCI Express based reflective memory has significant lower latency, higher throughput but currently supports less nodes and distance. Details for some popular reflective memory solutions can be found in the table below.

Feature	**Dolphin Express IX**	**GE Fanuc**	**SCRAMNet GT**
Standard	PCI Express	Proprietary	Proprietary
Network speed	40 Gigabit/s	2.12 Gigabit/s	2.5 Gigabit/s
Network topology	Switch	Ring	Ring
Max nodes	56 / 20 *3	256	256
Max distance end to end	600 meter	Up to 10 km	Up to 30 km
Cables	iPass Copper or fiber	Fiber	SFP copper or fiber
Data Deliver Jitter	200 ns pr switch hop	1 us pr node	Less than 1 us pr node
8 nodes	1us	8 us	< 8 us
20 nodes	1.4 us	20 us	< 20 us
56 nodes	1.4 us	56 us	< 56
Transfer methods	PIO, DMA *1, PCIe master	PIO, DMA	PIO
Write performance PIO	2650 Megabytes/s	26 Megabytes/s	210 Megabytes/s
Write performance DMA	*1	170 Megabytes/s	NA
Read performance PIO	20 Gigabytes/s *2	6 Megabytes/s	
Read performance DMA	3400 Megabytes/s *1	408 Megabytes/s	NA
Number of multicast groups	4	1	1
Max Memory configuration	4 x 2 Gigabytes	256 Megabytes	128 Megabytes
Type of Memory	System main memory	Device memory	Device memory
Fixed memory settings	No, software configurable	Yes, card is ordered with a specific memory size	Yes, card is ordered with a specific memory size
Memory is cacheable	Yes	No	No
Remote interrupts	Yes	Yes	Yes

The data in the table is found by googling for "reflective memory" and SCRAMNet. Please let us know if the data is incorrect. 1)The DIS 4.4.3 supports up to 4 x 64 Megabytes, 256 Megabyte reflective memory segments. The 4x 2 Gigabyte option and DMA operations are available with the DIS 5.0 or newer software release. Please contact Dolphin for more information. 2) Actual throughput depends on the local system memory to memory bandwidth. 3) Scalability, the IX hardware limits the number of nodes that can be used for general purpose, unicast, interrupts to 20. The reflective memory functionality only is limited to 56 nodes. Dolphin is working with PCIe chip vendors to ensure future solutions will scale to 256 nodes or more.

Roadmap and future plans

Dolphin's reflective memory solution utilizes the standard multicast functionality as defined by the PCI Express Base Specification 2.1. Upcoming PCI Express Gen3 and future PCI Express Gen 4 chipsets will further increase the performance and scalability for applications utilizing PCI Express multicast.

Dolphin is committed to maintain a stable SISCI API to enable customers an easily upgrade to new future PCI Express based multicast solutions.

SISCI API

The SISCI API (**S**oftware **I**nfrastructure **S**hared-Memory **C**luster **I**nterconnect) 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.

SISCI API Code examples

The SISCI Developers kit contains several basic code examples to demonstrate the use of SISCI and the reflective memory functionality. A good starting point for reflective memory is "reflective.c" (click to open the source).

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 SISCI Users guide and 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 / Multicast (This document)
- Dolphin Express Peer to Peer communication – Direct PCIe

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