Dolphin

INTERCONNECT SOLUTIONS

# SuperSockets for Linux Overview

# Table of Contents

## 1.0 Introduction

Since 1992, Dolphin Interconnect Solutions, www.Dolphinics.com, has delivered high speed networking and clustered computing systems. Dolphin focuses on the development and support of real time performance systems that benefit from low latency and high throughput. Dolphin's interface adapters and switches meet the needs of customers seeking high performance in the technical, HPC, financial and military markets. The latest generation of Dolphin's products is based on standard PCI Express technology and delivery the lowest latency while supporting standard interfaces.

This paper is an introduction to Dolphin SuperSockets software. SuperSockets enable standard applications to take advantages of Dolphin's extremely low latency and very efficient PCI Express network without application changes. Network communication is accelerated by eliminating communication bottlenecks allowing applications requiring fast communication to see dramatic performance improvements   SuperSockets is also available for Windows and is discussed in a separate paper -SuperSockets on Windows on our website.

## 2.0 What are Sockets?

From Wikipedia, Berkeley sockets originated with the 4.2BSD Unix operating system (released in 1983) as an API. In 1989 UC Berkeley released versions of its operating system and networking library free from the licensing constraints of AT&T Corporation's proprietary Unix. This interface implementation is the original API of the Internet Protocol Suite (TCP/IP).

All modern operating systems now have some implementation of the Berkeley socket interface, as it became the standard interface for connecting to the Internet. Even the Winsock implementation for MS Windows, developed by unaffiliated developers, closely follows the Berkeley standard.

The BSD sockets API is written in the programming language C. Most other programming languages provide similar interfaces, typically written as a wrapper library based on the C API.

### 2.1 BSD vs. POSIX

As the Berkeley socket API evolved over time, and ultimately into the POSIX socket API certain functions were deprecated or even removed and replaced by others. The POSIX API is also designed to be reentrant. These features now set the classic BSD API apart from the POSIX API.

| Action | BSD | POSIX |
|---|---|---|
| Conversion from text address to packed address | inet_aton | inet_pton |
| Conversion from packed address to text address | inet_ntoa | inet_ntop |
| Forward lookup for host name/service | gethostbyname, gethostbyaddr, getservbyname, getservbyport | getaddrinfo |
| Reverse lookup for host name/service | gethostbyaddr, getservbyport | getnameinfo |

### 2.2 Alternatives

The MPI API and IB verbs provide alternative solutions to socket communication, but are mostly only used with High Performance Applications running over specialized hardware. Existing applications using sockets requires a significant re-coding to utilize MPI or IB verbs communication. SuperSockets solves the performance issue with Ethernet, still offering the standard interfaces and therefore makes it unnecessary to invest in costly engineering to re-writing the application.

## 3.0 History of SuperSockets

Dolphin recognized that utilizing a high speed reliable network for sockets communication could eliminate the TCP/IP stack bottleneck and improve application performance.  SuperSocket is implemented on several high speed reliable low latency networks.  With SuperSockets, standard applications can take advantage of the performance improvements of these high performance interconnects without modification.

SuperSockets for Linux was introduced in 2001 with the Dolphin Express SCI interconnect and expanded for Dolphin Express DX Gen 1 PCI Express interconnect and switch technology in 2006. SuperSocket support for Dolphin Express IX - Gen 2 PCI Express networks was added in 2010 and continues to evolve.

SuperSockets is still supported on all Dolphin networks, new customers are recommended to use the Dolphin Express IX product family for performance reasons.

## 4.0 Why SuperSockets

SuperSockets is a plug and play driver module enabling any networked user space application(s) and Linux kernel services compliant to the Kernel Socket API to utilize the 40 Gigabit performance of PCI Express Network. SuperSockets takes advantage of the reliability of the PCI Express Network to eliminate layered system overhead and performance costs of TCP. The resulting performance is socket latency < 1.27 microseconds.

Dolphin developers identified the TCP/IP stack as a performance bottleneck many years ago and developed SuperSockets to bypass the stack and provide low latency, high performance socket communication. The basic latency problem with TCP/IP communication is not improved by wire speed. 10G Ethernet provides significantly higher throughput over 1G Ethernet, but latency is still a major overhead delay. The majority of the latency is caused by TCP/IP protocols receiving data from the Ethernet hardware and forwarding it to the right application while executing the communication protocols. TCP IP Offload Engines (TOE) exist to try to address the protocol execution problem, but they do not provide acceleration on forwarding the data to the receiving side of the application and must still process the TCP/IP stack. Hardware interrupts in this forwarding process are also a significant cause for delays and overhead.

The Dolphin SuperSockets API accelerates Berkeley compliant sockets applications without the need for application changes. The software is highly optimized to reduce system load (e.g. system interrupts) and supports applications that use both TCP and UDP. To further improve communication performance, SuperSockets takes advantage of the fast data transfer methods within PCI Express. It uses both Programmed Input/ Output (PIO) and Data Memory Access (DMA) operations to implement the most efficient transfers for all message sizes. PIO is highly efficient for small packet data transfers, while DMA is well suited for large pipelined data transfers. SuperSockets™ utilizes both methods and thresholds between methods can be tuned based on application requirements.

Communication within multi-processor systems can also be a source of latency. Applications utilizing local communication between processors (loopback communication) can be accelerated by SuperSockets™. This results in a significant performance improvement also for applications running on popular multi-core systems.

Overall, Dolphin SuperSockets delivers a fast and transparent way for any networked applications to dramatically improve performance. In combination with Dolphin Express hardware, application implemented with a sockets interface can achieve extremely low latency and high bandwidth communication.

## 4.1 SuperSockets in a nutshell

Major benefits are plug and play, high bandwidth, high availability, and much lower socket latency than network technologies like 1G and 10G Ethernet. Dolphin SuperSockets uses Dolphin Express Hardware remote memory access to implement a fast and reliable connection.

- Removes the need for time consuming networking protocols
- Reducing CPU Overhead
- PCI Express provides built-in data integrity
- Utilizes CPU Direct Remote Memory Access
- Reduces CPU cycles to transfer small messages to remote nodes
- Important for short transfers
- Significantly reduces small message latency
- Remote Direct Memory Access (RDMA)
- Reduces CPU cycles for large transfers
- DMA engine running in the background
- Adaptive select and Poll algorithms
- Multiplex socket calls, adapts to application characteristics
- Reduces number of interrupts
- High Availability Solution
- Instant Fail-over to alternate Dolphin card or Ethernet

## 5.0 How does it Work?

SuperSockets™ supports both zero-copy networking (RDMA) and remote memory cpu store operations (PIO) to transfer data. RDMA enables the network adapter to transfer data directly to or from application memory, eliminating the need to copy data between application memory and the data buffers in the operating system. SuperSockets engages the adapter DMA engine to transmit large messages and PIO (CPU load/store operations) to transfer small files. SuperSockets™ takes advantage of both methods to minimize latency for small messages and saving CPU cycles for larger transfers.

The actual data transport and protocols are implemented by a loadable Linux kernel device driver. No Linux patches are required. The driver offers a new transport family AF_SSOCKS compliant with AF_INET. Application programmers may use AF_SSOCKS directly but Dolphin recommends using the included SuperSockets LD_PRELOAD library to automatically convert AF_INET to AF_SSOCKS at run-time. The benefits of this are that SuperSockets can support applications without any modifications and also automatic fail-over to standard Ethernet communication in the case of problems with PCI Express network. PCI Express is by design a very reliable, error free protocol, but hardware may fail or cables may be disconnected.

One of the major bottlenecks with regular TPC/IP communication over Ethernet is the task of getting the user data extracted from the TCP/IP protocols. Using Ethernet, data is typically received by the Ethernet hardware and processed after the card has issued a system interrupt. The interrupt handler manages the queue and sends the user data up into the networking stack for further processing. Applications waiting for data are not notified until the networking stack has verified the data and identified and copied the data into the relevant user buffers. SuperSockets is built on the ideas of the Socket Direct Protocol (SDP) initially defined for Infiniband. PCI Express enables SuperSockets to establish a number of parallel queues for direct peer to peer communication between prioritized applications. Data is typically placed into dedicated application receive-buffers directly accessible from the application receive thread. Shared memory is used internally by SuperSockets to minimize communication protocol overhead and the number of system interrupts.

SuperSockets needs to be explicitly activated by the application or service to be accelerated. Performance counters can be retrieved using tools included in the SuperSockets bundle. A Dolphin Express communication channel is enabled by a configuration file named dishosts.conf. This file contains the host names of the machines in the cluster and their associated nodeId in the network. Sockets that are configured for a Dolphin Express endpoint will be routed through the low latency Dolphin SuperSockets module.

Basic CPU load and store instructions is extremely efficient for small messages. Application data is normally located in the CPU cache and it is therefore just a single write posted CPU store instruction to a memory mapped IO address to send 8 bytes of data. There is no need to lock down or register memory as with regular DMA transfers. The cost of sending 8 bytes of data is normally only a single CPU tick with worst case send latency for 8 bytes 210 nanoseconds. Using separate, distinct communication buffers enables multiple processes to send and receive data in parallel without any inter process synchronization or buffer contention.

The Dolphin PCI Express hardware includes CRC check-summing and hardware re-transmit so the CPU does not need to execute the full TCP/IP protocol for safe data transmission. Data can still be lost due to malfunctioning hardware, e.g. cable is unplugged. All abnormal events will be reported by the PCI Express IO system and SuperSockets utilizes this information with its light weight communication protocols to ensure reliable data delivery, fail-over to alternative path or proper error reporting to the applications.

SuperSockets™ implements a streamlined and lock-free messaging protocol on top of shared memory to provide a efficient combination of polling and interrupt / interrupt coalescing to minimize system overhead.

Receive operations become more efficient as data is pulled directly out of dedicated socket receive buffers – this operation causes the data to be cached and immediately available for use by the application. This reduces system interrupts by 75% over traditional TCP/IP interfaces.

SuperSockets provides optimized poll() and select() that easily can be enabled or disabled for performance tuning.

## 6.0 What advantages does SuperSockets give me?

Dolphin has invested years of development, testing, and support of this technology in support of their real time, embedded, and commercial customers. Users benefit from a well proven solution that can accelerate their application in minutes with no changes to the application and a very modest investment.

### 6.1 SuperSockets provides:

**Transparency .**

- Dolphin’s SuperSockets requires no changes to your applications or system configuration.
- Any application using TCP/UDP/RDS/IP will run without modification

**Performance .**

- Dolphin’s Solution is the fastest solution for latency sensitive applications 1.26 µs user level socket latency
- Lowest overhead - single CPU store to send 8 bytes to remote node
- Automatic switch to DMA operations for larger transfers to save CPU cycles

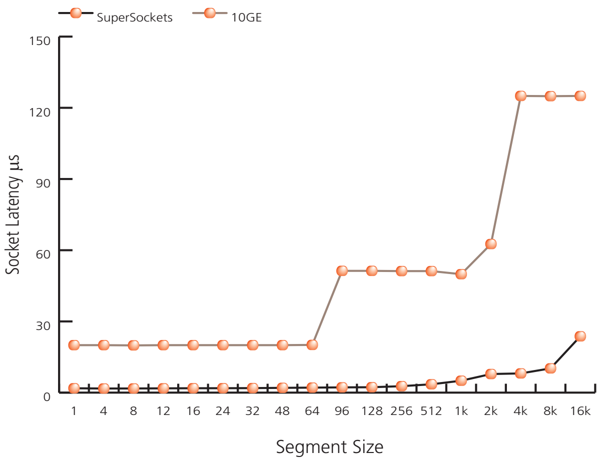
- Smart algorithms to minimize number of systems interrupts without adding system overhead or latency.

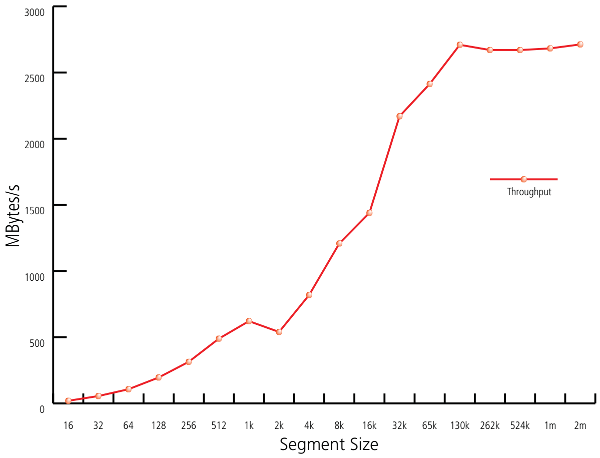
**Reliability .**

- Built on top of a reliable PCI Express protocol. Hardware check summing and error retransmit in the case of transient errors. PCI Express is a reliable network and errors do not happen unless there is a physical defect to the network.
- SuperSockets in critical reference installations for 5 years
- Built-in Fault Tolerance
- Automatic fail-over to alternate high speed network path, alternatively to native Ethernet if all high speed connections are down.

## 7.0 Why do I care?

Dolphin SuperSockets significantly reduces communication latency and provides close to the 40 Gbps wire-speed to applications. Actual application speedup will depend on the application and its communication bottlenecks. Typical applications run from 1.5 times to 10 times faster compare to current 10Gb Ethernet or Infiniband.





## 8.0 Want to go even faster?

Dolphin provides a easy to use shared memory API that provides direct access to remote memory, Interrupt and RDMA management. The SISCI Developer's Kit maximizes the performance of application using the PCI Express hardware. It delivers the lowest latency and highest throughput for application performance. End to end latency starts at 0.74us and throughput is over 3.5Gigabyte/s. Please find more information on the SISCI api at www.dolphinics.com.

## 9.0 Availability

SuperSockets™ is available for all popular Linux 2.6 - 3.x distributions and kernels for Intel and AMD x86 and x64 systems (both 32 and 64 bit installations).

All software is provided as an unified shell-installer which will install and configure the software cluster-wide as rpm or deb packages depending on the target system. The software comes with advanced monitoring capabilities.

More information about Installation, usage and interpretation of the data can be found under www.dolphinics.com/support

## 10.0 Summary

Dolphin SuperSockets, developed for high speed low latency clustered server applications, has evolved into a multidimensional software addition across a broad spectrum of applications from real time instrumental and data acquisition applications to multi-node cluster database applications providing financial analytics, medical device data collection, High Performance CFD and simulation, and telecommunications.

You should consider SuperSockets as an alternative to Ethernet or other networking methods if low latency and fault tolerant data movement across networks is required.

### 10.1 Key Product Features

- All applications will benefit from Dolphin Express without modification
- 100% compliant with Linux Socket library, Berkeley Socket compliant
- No OS patches or application modifications required. Just install and run
- Supports multiple adapters per host for increased fault tolerance and speed
- Both TCP and UDP supported
- Includes local loopback socket acceleration up to 10 times faster than standard Linux networking
- Automatic fail-over to redundant adapter in the case of network failure
- Transparent Fail-over to Ethernet if all Dolphin Express connection is down. Fail forward to Dolphin Express when the problem has been corrected
- Supports hot-pluggable links for high availability operation
- Supports both user space and kernel space clients
- Full support for socket inheritance/duplication
- Supports rolling software upgrades. NO need to stop your clustered application.
- Field proven for various MPI libraries, MySQL, DRBD and Oracle RAC

# 11.0 Dolphin Hardware Overview

## 11.1 Dolphin PCI Express IX Family

Dolphin has for more than two decades provided a variety of high performance, low latency products to solve the needs of financial, e-commerce, entertainment, high performance, and embedded customers.  These products enable customers to build and deploy faster more reliable systems.

Products range from high speed PCI Express Gen2 board and switch products to chip level PCI interconnect products with StarFabric. All products are designed as complete solutions including both hardware and software. The Dolphin Express IX family was introduced in 2010 and consists of two PCI Gen2 Adapters and a PCI Gen 2 8 Port Switch and associated Copper and Fibre cable connectors.

### 11.1.1 IXH610 Host Adapter

The Dolphin Express IXH610 is based on a Gen2 PCI Express non-transparent bridging architecture. This low profile PCI Express x8 adapter card provides 40 Gbits/s performance over a standard PCIe external cabling system. The IXH610 features transparent and non-transparent bridging (NTB), along with clock isolation. These powerful features make the adapter an ideal interconnect for applications such as test and measurement, medical equipment, and scientific computing.

Used in transparent mode, the card will run without the need for special drivers, while supporting standard PCI Express devices and device drivers. In NTB mode, the IXH610 offers Remote Direct Memory Access (RDMA) for high performance system to system communication. The IXH610 enables system clock isolation. By isolating the system clock and transmitting a very low jitter high quality clock to downstream devices, the IXH610 offers users improved signal quality and increased cable distances. The card also includes an EEPROM to support device re-configuration.

### 11.1.2 IXH620  XMC Adapter

The Dolphin Express IXH620 XMC Adapter brings 40Gbit/s connectivity and advanced connection features to embedded computers that support standard XMC slots or XMC VPX, VME or PCI carrier boards. The x8 Gen2 PCI Express adapter expands the capabilities of embedded systems by enabling very low latency, high throughput cabled expansion using standard PCI Express over cable.

Embedded application using SuperSockets can connect standard single board computers (SBC) or chassis to create a high speed network.

### 11.1.3 IXS600 High Speed PCI Express Switch

Today's high speed applications require both high performance and flexibility. Dolphin's IXS600 High Speed PCI Express switch delivers with a powerful and flexible PCI Express solution. This Gen2 PCI Express switch supports I/O expansion and high throughput clustering by combining Gen2 x8 PCI Express speeds of 40Gbps with Dolphin's clustering technology and IDT's standard transparent bridging capabilities. The integration of these features gives IXS600 users the ability to scale applications by connecting multiple cabled PCI Express devices in transparent mode or to cluster high speed processing solutions. As a key component of Dolphin's IX product family, the IXS600 provides the scalability for all types of cluster computing and I/O expansion applications.

The IXS600 is the switching component of the IX product family. This 8 port 1U cluster switch delivers 40 Gbps of non-blocking bandwidth per port at ultra low latency. Each connection is fully compliant with PCI Express Gen1 and Gen2 I/O specifications. Standard iPass™ connectors are used to connect copper or fiber-optic cabling.

For Non Transparent Bridging (NTB) operations, the IXS600 is combined with Dolphin's IXH610 PCI Express Host Adapter or IXH620 XMC Adapter. Operating in NTB mode, Dolphin has solved the strict power on requirements normally associated with PCI Express. Hosts, cables, and the switch can be hot-swapped and power cycled in any order, providing real plug and play as normally expected from networking components. The IXS600 supports Dolphin's extensive PCI Express Software libraries to create a powerful and reliable communication solution for applications that benefit from an ultra-low latency, high performance cluster interconnect. Dolphin's PCI Express Software enables standard Linux and Windows network applications to benefit from PCI Express performance without modification. This comprehensive solution is ideal for real-time, technical and high performance computing, cloud computing, and enterprise business applications.

## 12.0 More information

Go to www.dolphinics.com for more information on SuperSockets and other Dolphin technology products.

- SuperSockets for Linux
- SuperSockets for Windows
- Dolphin Express Reflective Memory / Multicast
- Dolphin Express Peer to Peer communication – Direct PCIe

Please contact your Dolphin sales representative or email sales@dolphinics.com if you have any questions.

## More information