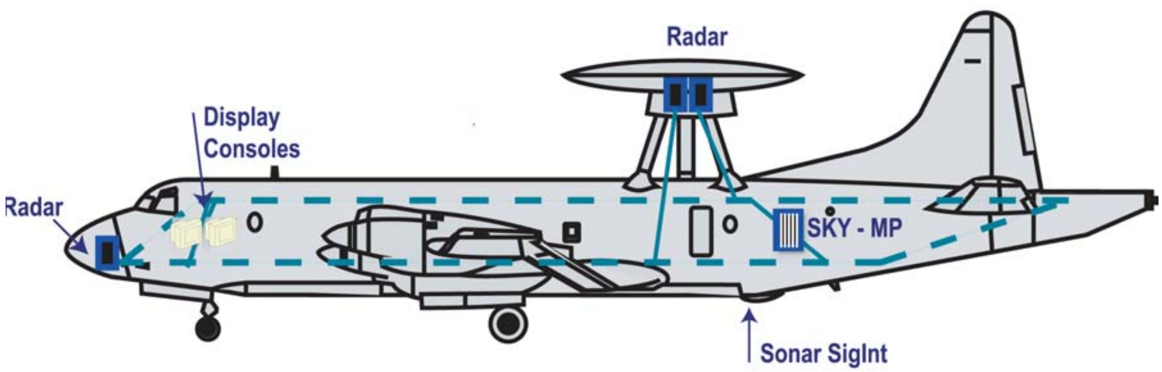
A SKY Computers White Paper

InfiniBand™ Architecture Delivers Enhanced Reliability

By: Steve Paavola, SKY Computers, Inc.



SKY Computers, Inc.
8 Centennial Drive
Peabody, MA 01960
978 977 3000
www.skycomputers.com

InfiniBand Architecture (IBA) is a revolutionary new, high-performance interconnect technology that not only provides high speed, but its intelligent fabric actually off loads the compute processing system from executing complex overhead and housekeeping tasks. IBA incorporates a number of new hardware and software technologies and processes designed to address the weaknesses of previous interconnect strategies and provide a growth path for future needs. It can be used inside a system to connect together multiple processors and local I/O, replacing older technologies such as SKYchannel and RACEway. It can also be used to tie together multiple systems, potentially running different applications. In this way, it is able to unify the interconnect across all of the applications, replacing and simplifying legacy technologies such as Ethernet, ATM, Fibre Channel and many proprietary interconnects.

While other articles have focused on the hardware performance aspects of IBA, this article will discuss how InfiniBand hardware and software work together to enhance the reliability and security of systems built with InfiniBand Architecture. Specifically, system protection, fault isolation and recovery, and CPU efficiency are all facilitated and enhanced by InfiniBand Architecture.

Independent Address Domains

InfiniBand Architecture supports an independent memory address space for each device in the system. The single, flat, address space supported by traditional protocols is replaced by multiple, protected, address domains. Each software application decides what memory locations it wants to make accessible to other nodes via the InfiniBand hardware. An application can grant remote access rights with byte level granularity for any window size within a registered virtual memory region. This feature is available on a Queue Pair by Queue Pair basis and prevents a misbehaving remote process from clobbering memory beyond its assigned memory window. Other processes, including the all important kernel space, are fully protected from unauthorized access. The act of registering a memory window pins down the memory to insure that the data is available when the channel adapter tries to access it, and also assures that the channel adapter can translate the virtual address into the correct physical address without assistance from the operating system. This is a key capability to enable security from accidental damage to unintended areas of memory, while simultaneously minimizing processor overhead. Since the hardware does the virtual to physical address conversion, the application can build and start DMA transfers without a system call. InfiniBand maintains process integrity by preventing other processes from accessing its memory. This yields both robustness and reliability for multiprocessor applications because errant processes won’t bring down the entire system.

Natural Scalability and Redundancy

At the physical level, InfiniBand Architecture is implemented as a point-to-point packet switched fabric. This provides a basic scaling advantage over shared bus architectures such as PCI and VME. The addressing capabilities of the InfiniBand protocol allow as many as 48,000 nodes to be supported in a single subnet. In addition, the global addressing ID field provides 64 bits for subnet identification. This allows routers to recognize up to a whopping 2^80 independent subnets. There are essentially no theoretical limits on the size of an InfiniBand system.

Extensive reliability features are built into the lower levels of the InfiniBand protocol. The hardware automatically checks for data integrity via CRC at every step in the transfer of a packet. An automatic retry mechanism is utilized to mask any transient errors that may occur. All service types, including the so-called "unreliable" classes of service, guarantee that the destination rejects corrupted data, and that a corrupted operation doesn't corrupt any adjacent data.

Finally, each device in the fabric maintains its own separate clock and power domain. Thus each component in the system is its own fault domain and may be implemented as an independent field replaceable unit. The channelized nature of InfiniBand I/O insures that a failed device only affects its own attached hardware link and the queue pairs associated with that device. Communications between other devices in the system is unaffected.

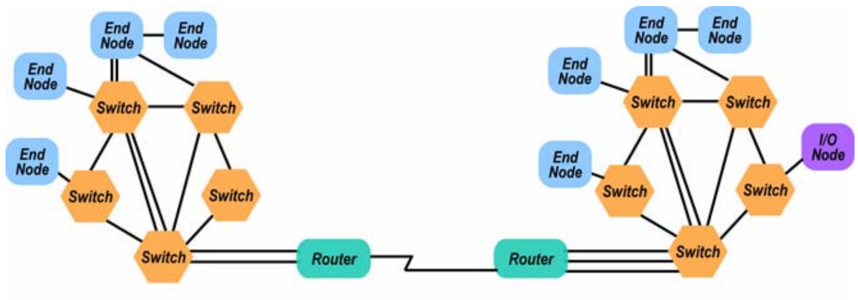
Fabric Management

The natural scalability and redundancy of InfiniBand fabrics allows designers to implement systems of varying topology, performance and failure resistance. Many of these functions are handled by a piece of software called a Subnet Manager. Using special, high-priority, Subnet Management Packets (SMP) the Subnet Manager is able to perform the following functions:

- Discover the fabric topology
- Discover nodes attached to the fabric
- Configure switch and end node parameters
- Configure switch forwarding tables
- Receive traps from subnet management agents

Upon detection of a fabric failure, the Subnet Manager can notify the application or system management software that the error has occurred and reconfigure switch and end node hardware to effect fail-over by rerouting I/O traffic over alternative paths.

This software is referred to as a Subnet Manager because there may be multiple management domains within the InfiniBand fabric. This flexibility provides further fault isolation capability and allows for the fabric to tolerate the failure of even the Subnet Manager itself.



This graphic is an example of two InfiniBand subnets connected by routers. The natural scalability and redundancy of InfiniBand fabrics allows designers to implement systems of varying topology, performance and failure resistance.

The creation of Subnet Management software would be a daunting task if every designer of an InfiniBand fabric based system had to write the code themselves. Fortunately, there is an open source project, called Open IB, which provides one.

Channel Based I/O Architecture

InfiniBand Architecture is a message passing, I/O channel-based architecture rather than register based. A channel-based architecture offloads the interprocessor communication overhead from the CPU to provide maximum available performance. For large multiprocessor applications the performance gain is significant.

The primary architectural element is the Queue Pair (QP). Each QP consists of an outbound queue, and an inbound queue. A connection is established by linking a local QP to a remote QP. Queue Pairs are not shared between applications; therefore, once they are set up, they can be managed at the application level without incurring the overhead of system calls. The QP is the mechanism by which quality of service, system protection, error detection and response, and allowable services are defined.

Creating a QP does require support from the operating system. Associated with each QP is a region of memory to be used for buffers, a key that must match on each incoming packet (the Q_Key) to verify the validity of the packet, and potentially a partition key specifying which portion of the fabric may be accessed by this QP.

The following operations are possible between Queue Pairs:

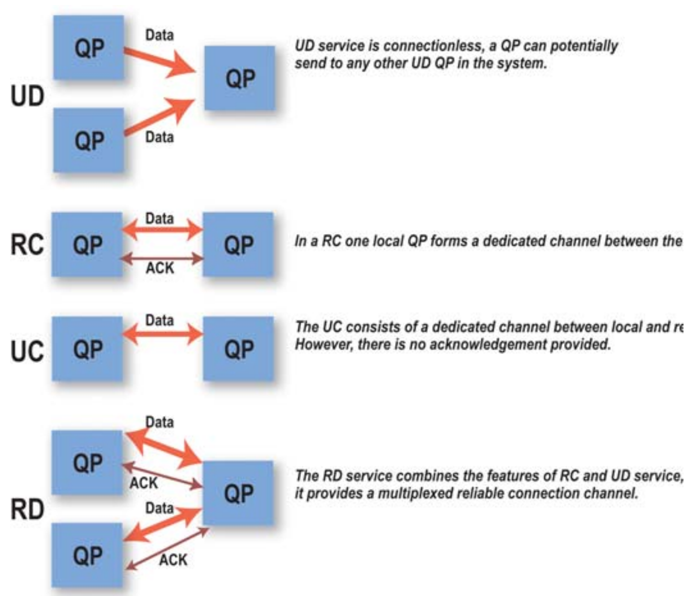
- Send – Includes a scatter/gather capability.
- RDMA-Write – Remote Direct Memory Access from local to remote memory.
- RDMA-Read – Remote Direct Memory Access from remote to local memory.

- Atomic Operations – Compare & Swap and Fetch & Add in remote memory.

Transport Service Types

Each QP is independently configured for a particular type of service. These service types provide different levels of service and different error recovery characteristics. The available transport service types include:

- **Reliable Connection (RC)** – Key to mission critical applications, the Reliable Connection provides the highest level of reliability and predictability available today. In a RC the local QP is associated with exactly one remote QP forming a dedicated channel between the QPs. The hardware protocol between those QPs provides reliable transport. It detects missing, corrupted or invalid messages and automatically suspends further activity between the two QPs, resends the lost information, and then resumes operation. Every queue operation is acknowledged and every operation completes exactly once, in order. The maximum message size is not limited by packet size or by the maximum transfer unit (MTU) size defined for the channel. Segmentation and re-assembly of long messages happens in hardware and is transparent to the application. The Reliable Connection supports all InfiniBand services – SEND, RDMA-Read, RDMA-Write and Atomic Operations.



- **Unreliable Connection (UC)** – An Unreliable Connection also consists of a dedicated channel between local and remote QPs. In this case, however, there is no acknowledgement provided. Send queue operations are marked as complete as soon as the QP transmits them. A missed message or message received in error is not automatically retried. It is dropped and the QP does not provide the sender with any indication as to whether the message was successfully delivered or not. Like RC service, operations complete in order, and the maximum size of a message is not limited by the packet size or path MTU. SEND and RDMA-Write are supported, but neither RDMA-

Read nor Atomic Operations are permitted over a UC channel. The UC provides more efficient communications, especially in streaming data applications, where a transient failure is not critical to the application.

- **Unreliable Datagram (UD)** – Unreliable Datagram service is connectionless. A QP can potentially send to any other UD QP in the system. The sending QP must send a key, called a Q_Key, which must match the receiver's Q_Key, or the message is dropped. (This prevents writing to unintended locations.) The only valid operation is SEND. A QP configured for Unreliable Datagram service cannot detect messages delivered out of order nor can it detect messages delivered more than once. A message received in error will be dropped without notification. The maximum message size is limited by the largest packet size supported by the path – the MTU (256 bytes to 4K bytes). This is the same class of service provided by most existing LANs, where upper layer software protocols perform the transport and reliability functions.
- **Reliable Datagram (RD)** – The Reliable Datagram service combines the features of RC and UD service. Essentially it provides a multiplexed reliable connection channel. RD is a key feature of InfiniBand Architecture as it combines the flexibility of multiple targets with the delivery guarantees of reliable connection. The QP is logically associated with a set of remote RD QPs. This service is most useful to an application having a number of different processes running on each node that need to communicate with each other in a reliable manner. Operations are acknowledged, complete exactly once, complete in order, and are automatically retried on error. All InfiniBand services are supported, i.e. SEND, RDMA-Write, RDMA-Read, and Atomic Operations.

DMA Scheduling

The use of DMA engines at every InfiniBand node is critical to offloading the CPU. In a traditional load/store model, the data must pass through all levels of the memory hierarchy on its way to and from a CPU register. The device driver is responsible for maintaining the queue of transfers for the device. In an InfiniBand based system, I/O operations are scheduled as queued DMA operations and are handled by the InfiniBand Host Channel Adapter (HCA) rather than the CPU. The HCA keeps track of which QPs have outstanding transfers, and completing those transfers. In addition to preserving memory bandwidth and precious cache resources, this architectural feature also minimizes the number of interrupts that must be serviced by the CPU during a data transfer.

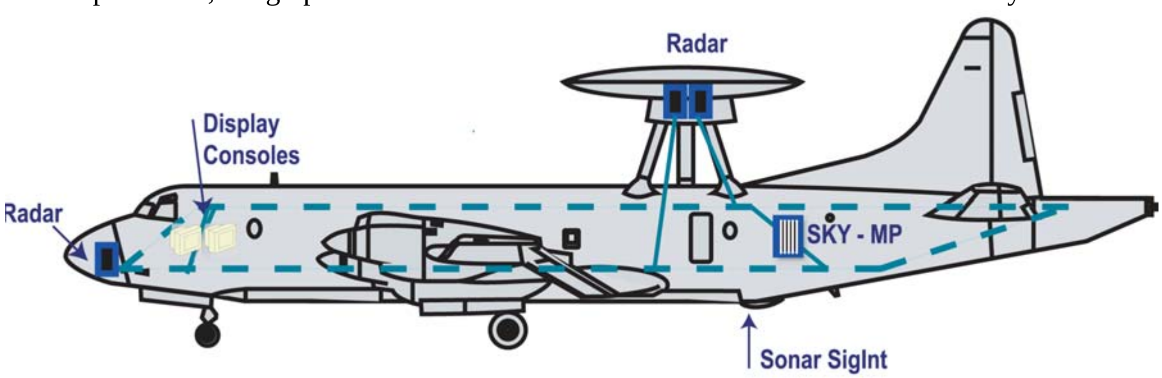
Summary

As interconnect speeds increase it is important to offload application processors from the housekeeping tasks. For example, management of the TCP/IP stack is not a significant load on a modem processor for 100BaseT, but can be a significant burden in 1 Gigabit/10Gigabit Ethernet systems. With an intelligent fabric architecture, InfiniBand systems can support the high speed interconnects of today as well as those of tomorrow.

As we have seen, hardware and software work together in the InfiniBand Architecture to provide greatly increased reliability and protection features while off-loading the application CPU from having to perform these tasks during the actual execution of the application. This concept of an "intelligent fabric" is one of the major revolutionary ideas embodied in the InfiniBand Architecture. Once the fabric connections are discovered, queue pairs and protection domains established, and the type and quality of service defined for each queue pair; the fabric operates reliably and securely at full performance without impact on system hardware or software resources.

InfiniBand on Board

We have described how InfiniBand technology provides a high-bandwidth interconnect within a compute intensive or I/O intensive application. As important to its role in a multiprocessor application such as a radar or sonar system, InfiniBand is able to tie together all of the applications in a complex environment such as is found on today's military platforms. In the example below, the graphic shows InfiniBand as the fabric that unites all onboard systems.



InfiniBand technology connects the radar or sonar sensors, the processing, and the command and control systems, while simultaneously maintaining redundancy and reliability.

The InfiniBand specification describes how to implement a router, a special purpose leaf node on the InfiniBand fabric, responsible for bridging between InfiniBand and other interconnect technologies to implement a cross-platform fabric. Typically, the other interconnects will be using TCP/IP for the protocol. The router can be used to connect together two InfiniBand subnets. It can also be used to connect InfiniBand Architecture applications to other applications. One simple, and potentially important, application is to provide a remote console. InfiniBand interfaces to Ethernet and Fibre Channel fabrics are already being demonstrated.

When considering platform or system expansion and re-configuration, InfiniBand technology provides an easy implementation and upgrade path. InfiniBand technology can be installed a sub-system at a time, and components can be updated independently in the future without redesigning the entire interconnect. This has not always possible with previous technologies.

The InfiniBand Architecture enables I/O to be located remote from processing. Once they are separated, they can be upgraded separately. Processors can easily be added to a multi-processor system. As faster processors become available, the processing complex can be replaced to provide the same capability with more performance and reliability. Similarly, a new radar could be added with the receiver located at the antenna, sending the digitized signal over a fiber optic InfiniBand cable to pre-existing processing complex(s).

Modularity, reliability, availability, serviceability and performance add up to a significant new capability for today's applications.