



Using RDMA to increase processing performance

Applications are increasing the demand for CPU processing performance and the amount of data being transferred between subsystems.

Offloading data movement to I/O hardware increases the amount of CPU resources available for these applications, boosting the system's performance.

By William Lee

Remote Direct Memory Access (RDMA) optimizes data movement between servers, accelerating the performance of applications in today's clustered supercomputers. With a few simple handshakes, two servers' network adapter cards can move data between their respective system memory spaces with little to no CPU involvement (see Figure 1). Using InfiniBand Host Channel Adapters (HCAs) with RDMA, data is placed directly into target memory. The CPU is required only for setup and completion signaling.

When compared to CPU-intensive transport technologies such as TCP/IP running on top of high-speed links (see Figure 2), RDMA can reduce CPU overhead by up to a factor of 10. Using conventional Network Interface Cards (NICs), the CPU must move data on to and off of the network preventing it from working on applications. In most common systems, data is moved twice. In embedded connections, on the other hand, where data sources and destinations are known at boot time, RDMA can lower CPU overhead to near zero. Using RDMA in I/O-intensive embedded systems can increase the value of the system by optimizing overall performance.

RDMA moves data directly from one CPU's memory to another's, but when data moves off the local bus, data integrity and transport become an issue. For this reason, RDMA requires a reliable transport to ensure data is segmented and reassembled correctly and arrives uncorrupted. On the send side, transport offload in the I/O device takes care of data segmentation and packet assembly. On the receive side, the I/O device handles all the data integrity checking, header stripping, and data reassembly. Assured delivery acknowledgements and retransmissions are also processed automatically. Subsystems using RDMA and transport offload can transfer a sustained data stream with little impact on the other applications' performance.

InfiniBand HCAs are designed to implement RDMA and transport offload. HCAs at both ends manage all the data transfer, essentially moving entire blocks of data from one subsystem's memory to another's. The receiving side can be a computer subsystem, server, or storage array. Additionally, data source and destination can be set up well before data becomes available. In certain embedded systems with predefined source and destination, this can be done when the system initializes before the applications run.

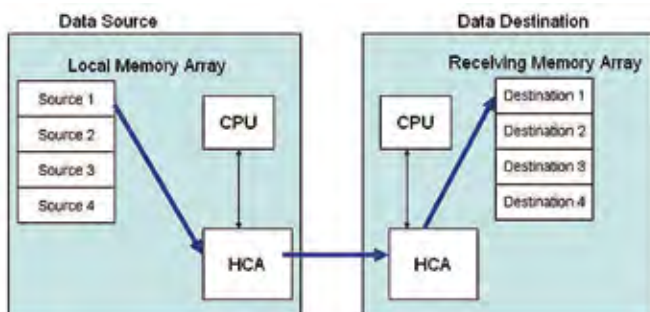


Figure 1

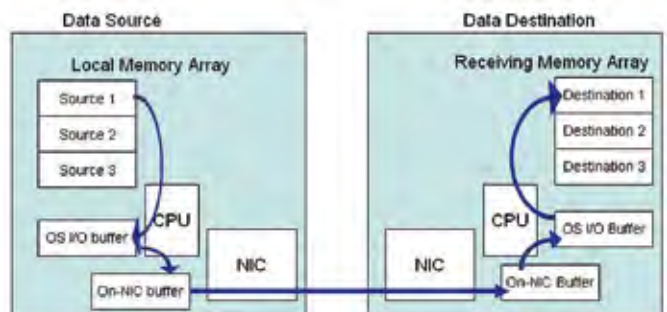


Figure 2

Data acquisition

High-bandwidth data acquisition systems generating a continuous stream of data, such as radar, satellite antennas, video cameras, or a collection of sensors aggregated onto one I/O channel, can take advantage of an InfiniBand HCA's RDMA capability. Blocks of memory can be allocated at boot time for local storage of the data being acquired. The HCA can be configured with the location of each local block of memory and the destination location for that block in the receiving subsystem.

Memory blocks are used to store data as it is acquired. When the first block is filled, the HCA is notified and transfers the data while the next block is filled. Data can be continuously acquired and transmitted in this manner. The CPU only has to notify the HCA as each block is filled, representing a significant reduction in CPU overhead compared to non-RDMA systems.

The receiving node can be a storage array, single server, or server within a cluster. Real-time rendering of video or radar images, for example, requires a significant amount of processing. Because the destination location was predefined and configured in the sending unit, CPU involvement in the data transfer is near zero; it will receive notification at the end of each data block reception for processing as needed. Using InfiniBand RDMA ensures that a continuous stream of data is available for the server and that the server has the maximum processing power required for the job.

Clustered storage array

Storage servers can be clustered with an InfiniBand fabric to create a high-performance, scalable array for Network Attached Storage (NAS) or Storage Area Network (SAN) units. Each storage server managing its own client connections appears as a monolithic block of storage when in reality it can share storage capacity with any number of its peers. IT managers can easily increase the capacity of a storage array by adding more storage servers to the embedded fabric.

Using InfiniBand RDMA ensures that a continuous stream of data is available for the server and that the server has the maximum processing power required for the job.

Relying on the storage server's CPU to manage data movement to and from its peers would severely restrict the number of clients the storage server could serve. Using RDMA between the storage servers offloads data transfers from the CPU, freeing it to serve more clients or to do more work for each client, such as running more I/O operations per second or providing more bandwidth. This improves scalability because the workload on the storage server does not increase dramatically with each added peer. Furthermore, connecting the backup system to the embedded fabric simplifies backups because data movement from each unit can be offloaded to the unit's HCA rather than involving its CPU.

Data transfers between storage servers wouldn't necessarily require predefined memory locations. With each client request, the storage server would communicate the source and destination locations with the appropriate peer. At this point the HCAs would take over, moving the data efficiently across the embedded fabric.

www.embedded-computing.com/search

Real-world applications

Mellanox Technologies' InfiniHost III InfiniBand HCA devices and cards have been implemented in systems similar to those described earlier. The single-chip architecture can provide RDMA and reliable transport without requiring external memory. Integrated SERDES supports one or two ports at speeds up to 20 Gbps, and transmission distances up to 20 meters or more over copper cables and 300 meters over fiber-optic cables. These reach capabilities allow the data source to be located in a separate room from the destination if necessary. These devices and cards have full open source driver support through the OpenFabrics Alliance (openfabrics.org). **ECD**

William Lee is a senior product marketing manager at Mellanox Technologies. Prior to joining Mellanox, he was a product line marketing manager at Zarlink Semiconductor and a strategic marketing manager at Marvell. He received his BSEL from the California Polytechnic State University in San Luis Obispo.



For more information, contact William at:

Mellanox Technologies

2900 Stender Way • Santa Clara, CA 95054

408-916-0022 • bill@mellanox.com

www.mellanox.com