Aspects of the InfiniBand™ Architecture

Extended Abstract
Invited Paper

Gregory F. Pfister
IBM Server Technology and Architecture, Austin, TX, USA
pfister@us.ibm.com

Abstract

The InfiniBand Architecture (IBA) is a new industry-standard architecture for server I/O and inter-server communication. It was developed by the InfiniBand^{SM} Trade Association (IBTA) to provide the levels of reliability, availability, performance, and scalability necessary for present and future server systems, levels significantly better than can be achieved with bus-oriented I/O structures. This paper provides a brief description of IBA.

1. The IBTA and the IBA

The InfiniBand Architecture (IBA) is a new industry-standard architecture for server I/O and inter-server communication. It was developed by the InfiniBand Trade Association (IBTA), a group of over 220 companies (as this is written) which was founded in August 1999. IBTA is lead by a Steering Committee of representatives from Dell, Compaq, HP, IBM, Intel, Microsoft, and Sun, co-chaired by IBM and Intel.

The team which developed the architecture and its specification is composed of about 100 individuals from those and other member and Sponsor companies. The resulting specification [1, 2] is large—approximately 1500 pages long. Clearly this article can only introduce the concepts and features involved; the reader should refer to the specification itself for a complete description, which is freely downloadable from the IBTA web site (see references).

While approximately 50 companies have already announced over 100 products supporting IBA, and a number of those products are expected to be available before the end of 2001, IBA's full penetration into the server market is expected to be relatively long-term as products and plans in the computer industry are often publicized. The reason is that full exploitation of IBA requires integration into system processor nests and memory subsystems, which change less rapidly than other aspects of systems. Nevertheless, IBA's use is expected to be very significant: IDC has estimated that by 2004, about 80% of all servers sold will support IBA.

The remainder of this extended abstract provides a short overview of IBA characteristics and a list of additional facilities now being defined by the IBTA.

2. The InfiniBand Architecture

In order to provide scaling, performance, reliability, and availability at levels that would be difficult or impossible to achieve using buses, IBA defines a switched, point-to-point communications network to which I/O devices and servers attach. The network transfers messages, not load/store memory-oriented commands. It differs in key ways from existing generally similar architectures, such as Fibre Channel or IP over Ethernet, in that it was designed to be implemented in hardware as the primary I/O attachment directly connected to a server's memory subsystem. Initial implementations with the key goal of time to market will be adapters which attach to existing I/O busses, particularly the PCI family; but IBA is ultimately intended to replace such busses as the primary industry-standard I/O attachment point to servers.

The key elements of the architecture are illustrated in Figure 1. They are:

- IB link: This is a point-to-point link connecting IBA elements. Its endpoint attach to IBA ports (not explicitly illustrated).
- Host Channel Adapter (HCA): This has ports which connect a host to one or more InfiniBand links. The host memory organization shown in the figure is for illustration only; IBA does not specify how the HCA is attached inside the host. The functions of the HCA visible to software on the hosts are defined by IBA verbs.
- Target Channel Adapter (TCA): This has ports which connect devices to IBA links. TCAs differ from HCAs only in that TCAs do not have defined verb support.

Proceedings of the 2001 IEEE International Conference on Cluster Computing (CLUSTER'01)
0-7695-1116-3/02 $17.00 © 2002 IEEE

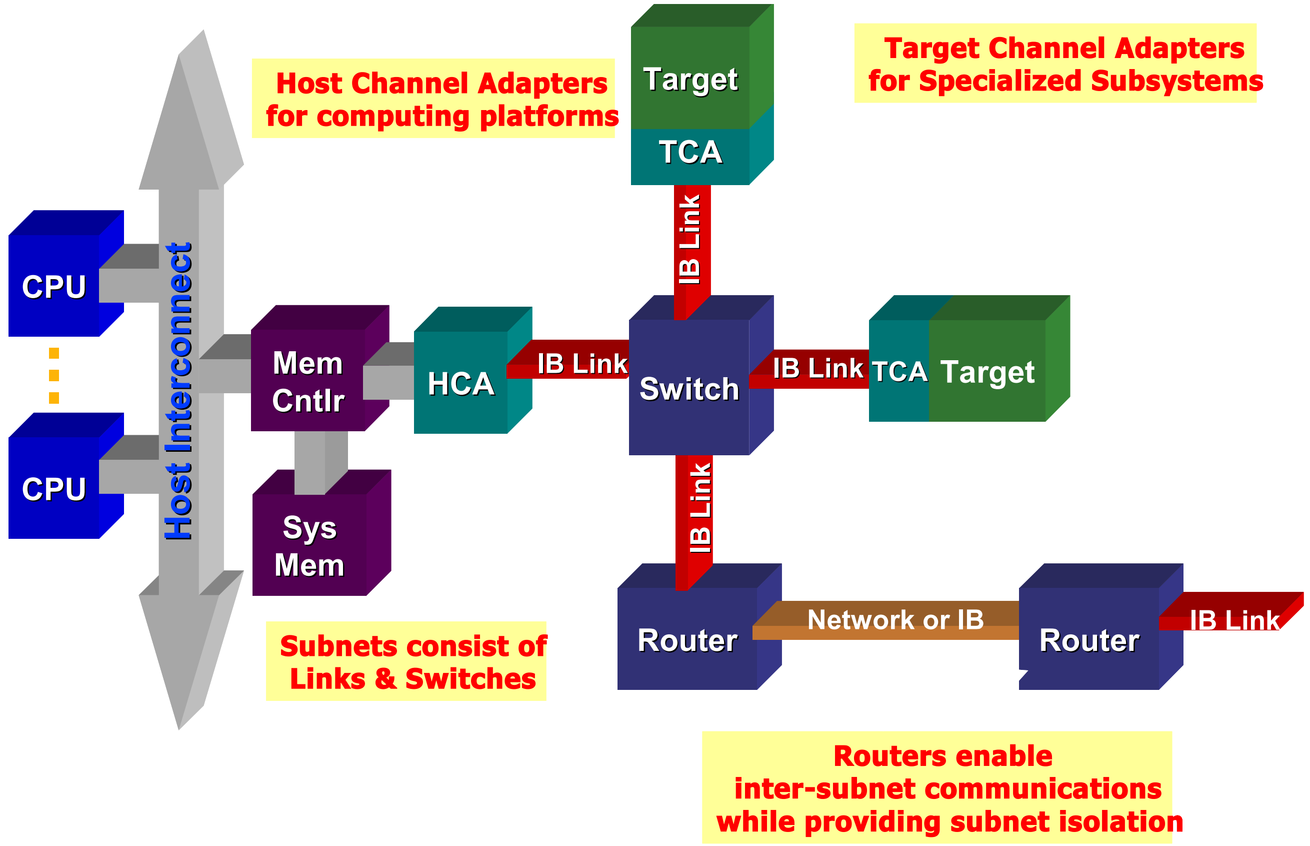

Figure 1: InfiniBand Architectural Elements

- Switch: This routes packets between links on an IBA subnet.
- Router: This connects separate IBA subnets together, providing both very large scaling and isolation. The transport used between routers may be IBA or may be some other mechanism, for example a MAN or WAN.

HCAs, TCAs, and Routers are collectively referred to as endnodes; they are potential destinations of normally used data-transferring packets.

Like all modern communications architectures, IBA is divided into layers; for IBA, physical, link, network, transport, and higher layers are used.

At the physical layer, IBA defines both copper and optical fibre links, with connectors and physical form factors. All links are bidirectional and use a signalling rate of 2.5 Gbaud, enabling an aggregate bidirectional data transfer rate of 500 MBytes/second. Four or 12 physical connections can be used in parallel as a single link (with a single connector), allowing bidirectional transfer rates of up to 6 GBytes/second. The basic copper link consists of four wires: two differential pairs, one for each direction. While link length is not defined—only an attenuation budget of 15dB—copper links are expected to be up to 17m and optical links to 10Km.

At the link layer, IBA provides packet definitions, with minimum transfer unit sizes accommodating data of 256 bytes, 1KB, 2KB, or 4KB. Virtual lanes are provided, usable both for deadlock prevention in routing and in traffic prioritization for quality of service implementation. Subnet-local routing is defined among up to 48K endnodes; the remainder of a 64K Local Identifier (LID) space is reserved for identifiers of multicast groups (an optional feature). Switch routing is destination-based, with forwarding tables in switches that are initialized by subnet management. The network topology and routing algorithms used are explicitly undefined by IBA; they are vendor-specific.

The transport layer of IBA provides for both unreliable and reliable communication. The sequencing and retransmission required for reliable communication is designed to be performed in hardware. Remote DMA operations, directly transferring data to or from memory locations in other endnodes, along Atomic operations on remote memory are also provided (optional). This layer also provides partitioning: A key-based matching mechanism allows isolate endnodes from each other so that multiple systems (including multiple types of operating systems) can share a single IBA fabric without necessarily being aware of each others' existence.

At the network layer, facilities for globally routing packets through IBA routers are provided. This includes an extended transport header that uses IPv6-like 128-bit Global Identifiers (GIDs) to designate endnodes. All IBA facilities, including reliable transmission, remote DMA, and partitioning, are defined to work globally as well.

At higher levels, the verb support defines operations to create and control send/receive Queue Pairs that are the source and targets of data sent on the links. Several communication services between queue pairs are also defined, including unreliable datagram, reliable connected, unreliable connected, reliable datagram (optional). Scatter/gather of data in work requests placed on queue pairs is also provided. Many of the key functions specified by the verbs are defined to operate without requiring priviledged execution modes, i.e., they can be performed directly by applications without the overhead of invoking an operating system kernel. These operations include sending and receiving messages, remote DMA, establishing memory protection bounds within already allocated regions, etc. Necessary memory mapping to allow user-mode virtual addresses in remote DMA operations also appear. Explicit APIs are not, however, defined by the IBTA.

IBA also defines in-band management facilities that include network initialization and maintenance, route selection by applications, and the definition of multicast groups. Management Datagrams (MADs) performing these operations, and the actions of managers sending and agents processing these MADs are specified, along several separate managers (such as performance management) and a framework that can be extended to include other management functions.

3. Current and Future IBTA Activities

The IBTA will continue to remain in operation for the foreseeable future, with tasks such as defining and administering specification compliance and interoperability; fixing errata in the specification; and also creating specification annexes which define necessary additional functions that did not appear in the 1.0 specification due to time constraints.

A complete test suite to define IBTA compliance is being developed and tested. As part of this activity, “plugfests” are being organized and operated which allow vendors to actively test their interoperability with other vendors’ products. A measure of the importance of interoperability to this effort is the fact that a well-attended plugfest has already happened, many months prior to the general availability of any IBA products. That these efforts are being successful was demonstrated at the June 2001 InfinBand Developers’ conference, where the vendors present connected their demonstration products into a single IBA subnet of over 70 endnodes that was successfully initialized and operated by yet other vendors’ management software. A demonstration of IBM’s DB2 EEE database has also been performed running in parallel across several IBA-connected hosts, using low-overhead IBA inter-node communication and performing using IBA to access disks without invoking the operating system.

Additions to the architecture in the form of annexes which do not obsolete 1.0-compliant implementations are under development. These cover a wide range of functions, including:

- further initialization and operational facilities such as console operation, booting, and I/O configuration management;
- a wire-level protocol providing standard lightweight support for sockets;
- other important facilities such as congestion control, routing protocols, management of multiple subnets, interoperable quality of service, and a change notification system to minimize outages from changes to IBA fabrics.

Assuming that IBA is successful, there are three very important things it will provide to the industry:

1. It will be a standard, high-volume enterprise-class server fabric, providing the reliability, availability, serviceability, manageability, performance, and scalability that implies. Such capabilities have heretofore only been available in proprietary systems.

2. It will, for the first time, provide non-proprietary low-overhead inter-host communication. This will result in new cluster multi-tier server solutions/markets that have previously been impossible.

3. It will allow complete separation of I/O devices from processing elements. This, will enable new form factors with much higher density of packaging than is now possible. It will also encourage new ways of looking at systems as a whole, e.g., data-centric views where the processing is considered peripheral to the data, which is central.

Separately, each of those three elements would be a very significant factor that could change the landscape of computing in the large. Together, they may well presage the widespread adoption of new, previously untried hardware and software structures for server computing.

References

[1] *InfiniBand Architecture Specification Volume 1, Release 1.0.a,* June 19, 2001, available from the InfiniBand Trade Association, http://www.infinibandta.org

[2] *InfiniBand Architecture Specification Volume 2, Release 1.0.a,* June 19, 2001, available from the InfiniBand Trade Association, http://www.infinibandta.org