HyperTransport™ Technology and InfiniBand Architecture:

The Complete High Bandwidth I/O Solution

Eric Krause

ISV Relations

ADVANCED MICRO DEVICES, INC.
One AMD Place
Sunnyvale, CA 94088

Introduction

Without a doubt, the server has become one of the most influential assets in today's global economy. Data centers used for data mining, databases, and server networks providing web pages and clusters have all greatly changed the business world, allowing for unprecedented growth. However, as the world economy grows, it demands more of its IT infrastructure. Data must be communicated faster and more reliably as the world depends on servers more, and as development and growth become more complex.

This aggravates an already existing problem. I/O communication, like most technologies, has progressed over the years but not at a rate sufficient to keep up with increasing demands of industry. For more than a decade, microprocessor performance has followed Moore's Law, which states that microprocessor performance will double every 18 months. While this has kept pace with industry demands and spurred growth forward, nevertheless it is evident that I/O communication performance shows signs of lagging behind.

There are two main pieces to the overall I/O problem: internal (or in the server box I/O), and external I/O (the I/O between servers). A number of new technologies address either or both parts of the problem, although it is difficult to make sense of where they fit within the server network, and how they relate to each other. A number of proposed solutions are competing with each other to become the industry standard, but two appear to have emerged on top. These are HyperTransport™ technology, which addresses the internal I/O problem, and InfiniBand Architecture, which takes on the external I/O problem. These two solutions have gained the acceptance and support of many industry giants, and complement each other well to widen the overall I/O bottleneck. First examining HyperTransport technology and InfiniBand Architecture individually, then determining how they fit together and interact with each other, is the best way to understand their symbiotic relationship.

HyperTransport™ Technology Overview

AMD began developing HyperTransport technology in 1997. Designed as a high bandwidth solution for chip-to-chip communications within a server platform, it also has applications in other onboard technologies such as networking, telecommunications, and embedded systems. The proliferation of HyperTransport technology throughout so many segments provides developers with a source of high-volume, low-cost devices that offer outstanding performance. As a result, leaders from a variety of market segments have banded together to create an organization to promote the evolution and adoption of HyperTransport technology. AMD, API Networks, Apple Computer, Cisco Systems, NVIDIA, PMC-Sierra, Sun Microsystems, and Transmeta founded the HyperTransport Consortium on July 23, 2001. HyperTransport technology was designed to meet several goals, including:

- Increased Internal I/O Bandwidth
- Improved Scalability
- Minimal Software Support Required
- Straightforward Physical and Electrical Design
- Reduced Power Requirements

As CPUs advanced in terms of clock speed and processing power, the I/O subsystem that supports the processor could not keep up. In fact, different links developed at different rates within the subsystem. The basic elements found on a motherboard include the CPU, Northbridge, Southbridge, PCI bus, and system memory. Other components are found on a motherboard, such as network controllers, USB ports, etc., but most generally communicate with the rest of the system through the Southbridge. Figure 1 shows a layout of these components.

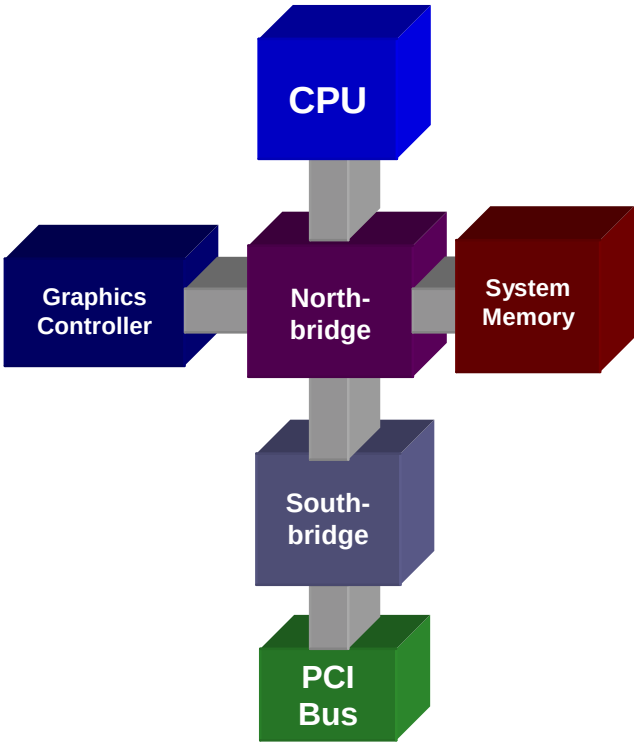

Figure 1: Common motherboard layout

Many of the links above have advanced over the years. They each began with standard PCI-like performance (33MHz 32-bit wide, for just over 1Gbps throughput), but each has developed differently over time:

- The link between the CPU and Northbridge has progressed to a 133MHz (effectively a 266MHz as it is sampled twice per clock cycle) 64-bit wide bus. This provides a throughput of close to 17Gbps.
- The Northbridge to system memory link has advanced to support PC2100 memory: it is a 64-bit wide 133MHz (also sampled twice per clock cycle) bus. This link also has a bandwidth of almost 17Gbps.
- The Northbridge to graphics controller connection has stayed at 32-bits wide and grown to a 66MHz bus, but with 4xAGP it is sampled four times per clock. 8xAGP (sampling the data eight times per clock) will pull the throughput of this link even with the other two at nearly 17Gbps.

Until recently, however, the Northbridge-Southbridge link has remained the same standard PCI bus. Although most devices connected to the Southbridge do not demand high bandwidth, their demands are growing as they evolve, and the aggregate bandwidth they could require easily exceeds the bandwidth of the Northbridge-Southbridge link. Many server applications, such as database functions and data mining, require access to a large amount of data. This requires as much throughput from the disk and network as possible, which is gated by the Northbridge-Southbridge link.

HyperTransport technology addresses this bottleneck by providing a point-to-point architecture that can support bandwidths of up to 51.2Gbps in each direction. Not all devices will require this much bandwidth, which is why HyperTransport technology operates at many different frequencies and widths. Currently, the specification supports a frequency of up to 800MHz (sampled twice per period) and a width of up to 32-bits in each direction. HyperTransport technology also implements fast switching mechanisms, so it provides low latency as well as high bandwidth. By providing up to 102.4Gbps aggregate bandwidth, HyperTransport technology enables I/O-intensive applications to use the throughput they demand.

A versatile platform needs to support a number of different devices including PCI buses, PCI-X buses, network controllers, etc. HyperTransport technology supports a scalable architecture so that a platform can support as many devices as necessary and with the appropriate bandwidth each device needs. HyperTransport technology supports tunneling, which means a device can be connected to a HyperTransport chain without terminating the chain. Many devices can be daisy-chained together to form one long HyperTransport chain. However, each device on the chain will share the available bandwidth. A device can create its own HyperTransport chain to form a branch from the original chain, therefore the device is able to support and manage the communications between its own devices separately from the initial chain. This allows the platform architect a great deal of latitude in a motherboard design so that the platform will appropriately support the necessary resources.

In order to ease the implementation of HyperTransport technology and provide stability, it was designed to be transparent to existing software and operating systems. HyperTransport technology supports plug-and-play features and PCI-like enumeration, so existing software can interface with a HyperTransport technology link the same way it

does with current PCI buses. This interaction is designed to be reliable, because the same software will be used as before. In fact it may become more reliable, as data transfers will benefit from the error detection features HyperTransport technology provides. Applications will benefit from HyperTransport technology without needing extra support or updates from the developer.

The physical implementation of HyperTransport technology is straightforward, as it requires no glue logic or additional hardware. HyperTransport technology specifications also stress a low pin count. This helps to minimize cost, as fewer parts are required to implement HyperTransport technology, and reduces Electro-Magnetic Interference (EMI), a common problem in board layout design. Because HyperTransport technology is designed to require no additional hardware, is transparent to existing software, and simplifies EMI issues, it is a relatively inexpensive, easy-to-implement technology.

Recently it has become more important for platforms to provide power-saving features. Cooling large numbers of servers is a significant cost factor for many businesses. HyperTransport technology helps platforms save power by transmitting signals based on enhanced Low Voltage Differential Signaling (LVDS). By using two signals for each bit, less voltage is needed per signal, and the noise effects for each signal are similar, making it easier to filter those effects out. In this way, LVDS supports a very high frequency signal, enabling HyperTransport technology to support clock speeds up to 800MHz. It also requires no glue logic or extra onboard hardware, and supports power saving protocols to power down extra components when not in use. This is important for a server or clustered platform, as there are usually large numbers of nodes networked. Saving a little power on each platform adds up to big savings quickly, both in terms of power consumption and cooling costs.

Figure 2 is a sample motherboard layout that implements HyperTransport technology and could support AMD's (code-name) Hammer processor.

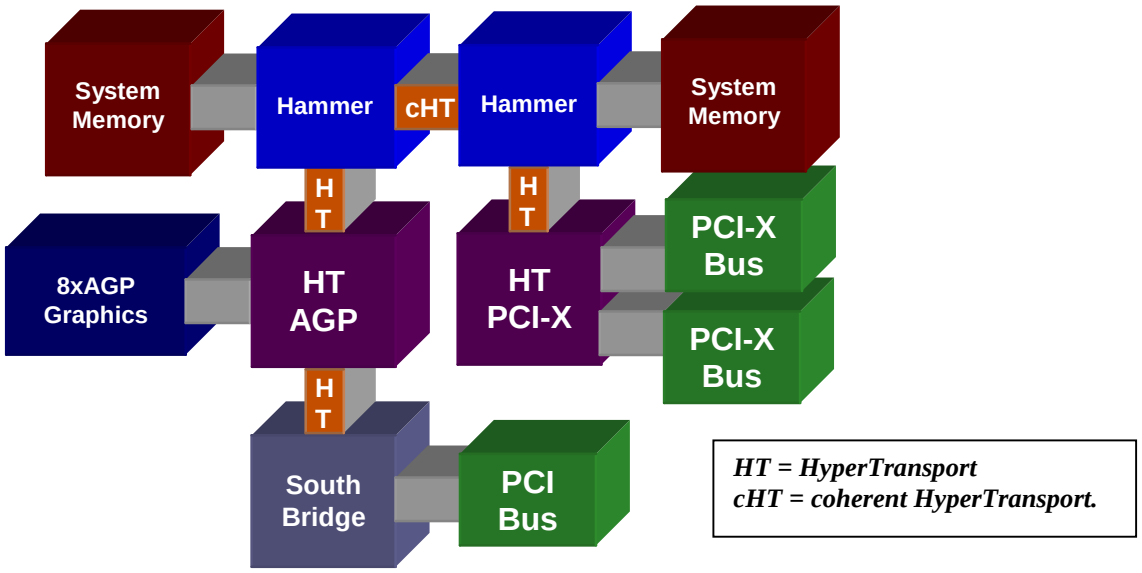

Figure 2: Possible platform layout

AMD Hammer processors will have an integrated memory controller, so the system memory will connect directly to the processor itself. Coherent HyperTransport is a proprietary implementation of HyperTransport technology developed by AMD, with added coherency features to properly enable the connection between the processors above. HyperTransport technology's tunneling feature is used to implement a HyperTransport-to-AGP bridge and a Southbridge onto the same HyperTransport chain. The other processor is simply connected to a HyperTransport-to-PCI-X bridge that supports two PCI-X buses. Theoretically, each of these chains can support the total bandwidth their components can demand.

HyperTransport technology is a solution to many of the constraints that exist in current internal I/O systems. Higher bandwidth and improved scalability, coupled with minimal software support, straightforward hardware implementation, and reduced power requirements will make it a key component in next-generation servers.

InfiniBand Architecture Overview

InfiniBand Architecture is a high-performance external I/O solution for servers and server clustering. When InfiniBand Architecture was consolidated from NGIO technology and Future I/O technology in 1999, Compaq, Dell, Hewlett-Packard, IBM, Intel, Microsoft, and Sun formed the InfiniBand Trade Association (IBTA). Now made up of more than 180 companies, the IBTA is tasked with defining InfiniBand Architecture specifications and promoting compatibility and interoperability across all InfiniBand Architecture developers' products. Networks can always utilize more bandwidth, but as server networks have advanced, questions of security, scalability, reliability, and standardization have become more important.

InfiniBand Architecture provides all of these features. It is a point-to-point architecture, which supports links that operate in 1X, 4X, and 12X versions. Respectively, these provide 2.5 Gbps, 10 Gbps, and 30 Gbps of bandwidth in each direction. InfiniBand Architecture defines four different classes of nodes that make up an InfiniBand Architecture network. Those classes are:

- **Host Channel Adapters**—HCAs reside on processor-based platforms, i.e. servers
- **Target Channel Adapters**—TCAs are typically used to support I/O devices
- **Switches**—Switches route packets from one link to another of the same subnet
- **Routers**—Routers transport packets between InfiniBand Architecture subnets

An InfiniBand Architecture subnet is serviced and maintained by a master Subnet Manager (SM) that resides on any single node in that subnet. This manager directs all traffic on the InfiniBand Architecture fabric. When a server requests access to information found on a storage array over the fabric, the master SM establishes that connection based on network performance, usage, and topology. Once connected, the two nodes are able to exchange data. That data is encoded with information about the path it is to take through the subnet, and an identifier of the intended recipient. When a node receives a packet of information, it examines the keys contained in the packet to make sure that it is the intended receiver and that the packet originated from an authorized sender. If this is not the case, the receiver will drop the packet. Security is always an issue

for servers, as confidential and important information is to be shared only amongst trusted nodes. Firewalls and software encoding are common mechanisms currently used in networks, although implementing security features into the network hardware is inherently more secure.

InfiniBand Architecture also provides easy scalability. The SM maintains the topology of the fabric, so that when a node goes out of service, the impact is minimal. If another node is added to the fabric, the SM will initialize and configure the new node, and add it to the fabric automatically. This allows a network administrator to remove a node for servicing or add nodes to expand the network. It also allows a company to expand its network with new servers without the need of removing the older ones. Simply add the new nodes, and the network is expanded. InfiniBand Architecture developers are also creating bridges to other network architectures, such as Gigabit Ethernet and Fibre Channel. These bridges allow a network administrator to expand an already existing network with InfiniBand Architecture-enabled nodes, regardless of the existing architecture. Scalability is also an important feature for server networks, as the usage of a network changes with time. A new business may begin with a small network, but as the business itself grows, its computing needs also grow. For many network architectures, however, adding nodes to the network is a complex task. The easier it is to add and remove nodes the less risk of downtime for the network.

In terms of reliability, InfiniBand Architecture provides a number of different service levels by which nodes may communicate. Each service level has its own virtual lanes and flow control associated with it, as well as a certain quality of service for the network connection. Depending on how fast or accurate the connection needs to be, the proper service level can be used. Also, if the master SM's node goes down, another SM on another node will take over the role of master SM. This assures that problems on one node are isolated to that node. InfiniBand Architecture's support for scalability allows for that node to be disconnected from the fabric, serviced, and reintroduced later. In the meantime, another node can be added to the fabric if necessary. Reliability is one of the most important characteristics of a network. Servers demand that data be transferred without error. Should there be a problem with one node, that problem must remain isolated to that node and network performance should be affected as little as possible. Ensuring quality of service is important, as server applications must be able to trust that the data they receive is correct.

The IBTA promotes InfiniBand Architecture specifications and proper adherence to the specifications. Establishing these standards is key to enabling the aforementioned characteristics of InfiniBand Architecture. Security, scalability, and reliability all depend on the nodes that make up the network to work together properly, regardless of the vendor. Standardization of a network architecture provides customers with a wide base of products which are much more likely to interact with each other properly. It also makes it easier to replace a node in the event of failure, and makes it much easier to service the network.

InfiniBand Architecture provides support for security, scalability, and reliability, and provides a standard that is supported by the IBTA. It addresses the major needs of server and storage area networks, and has already gained a great deal of support throughout the industry. Companies are already developing HCAs, TCAs, switches, routers, and fabric management software in anticipation of its success.

HyperTransport™ Technology and InfiniBand Architecture

Many large companies depend on enterprise-class servers to provide accurate and dependable computations on large amounts of data. This requires as much network bandwidth as possible, which in turn demands a fast, wide pipe from the network controller to system memory. HyperTransport technology provides a speedy, dependable solution for the internal link, while InfiniBand Architecture offers the secure, easily scalable, reliable external link. HyperTransport technology and InfiniBand Architecture complement each other well to form a complete and compelling high-bandwidth solution for the server market. Many features of HyperTransport technology and InfiniBand Architecture correspond closely, and allow a HyperTransport technology-based platform supporting an HCA to function more closely in concert with the network itself. The complementary characteristics of HyperTransport technology and InfiniBand Architecture include not only their bandwidth, but are found within their protocols, reliability features, and support for scalability as well.

HyperTransport technology provides the bandwidth support necessary to take full advantage of InfiniBand Architecture throughput. A 1X InfiniBand link could demand as much as 5Gbps bandwidth. While PCI 64/66, at over 4Gbps, may be adequate, it can hold back network performance. Even PCI-X, at just over 8.5Gbps, cannot handle a single 4X

InfiniBand Architecture channel. Figure 3 illustrates how InfiniBand Architecture's bandwidth needs compare to the support internal I/O technologies can provide.

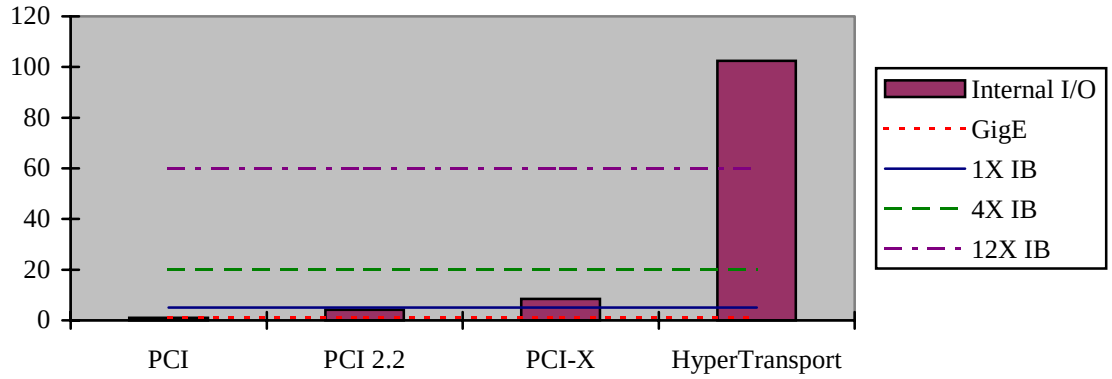

Figure 3: Bandwidth of HyperTransport™ technology and other I/Os compared to InfiniBand Architecture (in Gbps)

Regarding bandwidth, HyperTransport technology provides an appropriate framework for InfiniBand Architecture. A HyperTransport technology link can easily handle even a 12X InfiniBand Architecture channel, whereas PCI-based buses are unable to handle the slower links.

HyperTransport technology and InfiniBand Architecture have similar protocols. They both support atomic operations: Fetch and Add, and Compare and Swap. Atomic operations are not broken down into smaller operations by a protocol. This means that atomic operations are not slowed down by the latency required when using a series of commands. When a HyperTransport technology-based HCA receives one of these operations, it can pass the command on, ensuring that the atomic instructions in InfiniBand Architecture are truly atomic all the way to system memory. HyperTransport technology also supports I/O streams, which group data communiqués, in a similar fashion to InfiniBand Architecture's virtual lanes. Within a HyperTransport chain, each node's traffic is in a different I/O stream, but a single node may have more than one I/O stream. These I/O streams are treated as independent groupings of traffic by the fabric. The virtual lanes of InfiniBand Architecture provide a quality of service to the fabric, and a HyperTransport technology-based HCA can place packets from a virtual lane into a corresponding I/O stream, passing on some of the quality of service InfiniBand Architecture provides into the box.

HyperTransport technology and InfiniBand Architecture have similar means to enhance reliability. They both support the error-detecting ability of Cyclic Redundancy Checks (CRCs). Data is sent to the HCA via HyperTransport technology (which supports a 32-bit CRC), and then the HCA passes it on using the InfiniBand Architecture CRC implementation. Data is covered by this error detection from the system memory of the source server to the system memory of the destination server. An InfiniBand fabric comprised of HyperTransport technology-based platforms supporting the HCAs can also be made more reliable through redundancy, addressed as a part of scalability.

The ease of scalability that HyperTransport technology and InfiniBand Architecture provide makes it feasible to have additional components in the platform and nodes on the network to fill in should a failure occur. These features can be used together to implement two HCAs on a HyperTransport technology-based platform. This would enable increased bandwidth and redundancy (if one HCA should fail, the other keeps the system connected to the fabric). The impact of supporting extra hardware on a platform is mitigated by HyperTransport technology's power management implementation. An extra component, such as a redundant HCA, can be powered down. If the powered HCA fails, it can be powered down and the other HCA can be revived. This redundant support makes any InfiniBand Architecture fabric composed of platforms enabled with HyperTransport technology highly resistant to large-scale network failure.

HyperTransport technology and InfiniBand Architecture work well together in terms of bandwidth, protocol, reliability, and scalability. A server network based on InfiniBand Architecture and made up of HyperTransport technology-based server platforms will enjoy a significant performance improvement, while ensuring more secure protection, reliable data transfer, and ease of service and scalability. HyperTransport technology and InfiniBand Architecture will revolutionize the way enterprise business data centers perform.

Conclusion

HyperTransport technology was designed to alleviate the I/O subsystem bottleneck found on computers today. InfiniBand Architecture provides a reliable and easily scalable framework for storage and server networks. Platforms enabled with HyperTransport technology communicating over an InfiniBand Architecture fabric will

be able to take advantage of the complementary reliability features, protocols, and bandwidth these two technologies provide. This enables powerful I/O and computing ability. Data mining, database support, server clustering, and other I/O intensive applications will be able to take full advantage of a HyperTransport technology-based system over an InfiniBand fabric. HyperTransport technology and InfiniBand Architecture are two well-positioned technologies, designed to take the industry into new levels of computational ability.

For More Information

For more information about any of the topics below, please refer to the following Web sites:

- AMD's Hammer Platform Overview: http://www.amd.com/us-en/assets/content_type/DownloadableAssets/MPF_Hammer_Presentation.PDF
- AMD or AMD's initiatives: http://www.amd.com
- HyperTransport technology or the HyperTransport Consortium: http://www.hypertransport.org/
- InfiniBand Architecture or the InfiniBand Trade Association: http://www.infinibandta.org/

AMD Overview

AMD is a global supplier of integrated circuits for the personal and networked computer and communications markets with manufacturing facilities in the United States, Europe, Japan, and Asia. AMD, a Fortune 500 and Standard & Poor's 500 company produces microprocessors, flash memory devices, and support circuitry for communications and networking applications. Founded in 1969 and based in Sunnyvale, California, AMD had revenues of $3.9 billion in 2001. (NYSE: AMD).