The HTX-Board: A Rapid Prototyping Station

Holger Fröning Mondrian Nüssle David Slogsnat Heiner Litz Ulrich Brüning
University of Mannheim
Computer Architecture Group
B6,26
68159 Mannheim, Germany
{froening,nuessle,slogsnat,heiner.litz,bruening}@uni-mannheim.de

Abstract

The Hypertransport technology as a chip-to-chip and board-to-board interconnect technology is a well established standard, used in various computing systems. New is the standardization of an expansion slot with direct Hypertransport connection, called HTX. The opportunity of such a direct I/O interface led to the development of a rapid prototyping station with a HTX connector, pluggable into any HTX-equipped system. The architecture and physical design of this HTX device is presented. Beside the HTX connection, the most remarkable parts in the architecture are a FPGA as main component and an array of six high speed serial transceivers. The intend of the transceivers is to build up custom direct interconnection networks. Beside the use for rapid prototyping and development of custom interconnects, possible applications are co-processing and CPU-offloading. The physical board design is shown with the most challenging problems like cost-efficient stack-up, the power distribution system and signal integrity for high speed signals.
The Hypertransport Consortium already adopted this development as a Reference Design in their portfolio. To our knowledge, until now there are no other comparable devices available.

1. Introduction

The Hypertransport (HT) technology [1] is a point-to-point link interconnect designed for chip-to-chip or board-to-board communication. This technology provides high bandwidth together with very low latencies, making this technology suitable for almost any application from embedded systems over PCs to high-performance computing systems. Nearly the complete CPU portfolio of AMD already compromises at least one HT link. One of the most important recent developments in the HT context is the introduction of the HTX connector [2]. In a very short time period the HTX technology was accepted by industry and lead to the launch of new mainboards from different vendors. The first add-in card designed for HTX is InfiniPath by PathScale [3], meantime several others are available. Still missing is a rapid prototyping station, which is mandatory for fast developments of HTX devices. The HTX-Board presented here is exactly designed for this purpose. It's main component is a high-performance FPGA, which is closely coupled to the main CPU over the HTX connector.

The application range of the HTX-Board is versatile. Beside the FPGA there are many other components embedded on the card, like DDR2-SDRAM, Gigabit-Ethernet, flash memory, mezzanine connectors, USB and six high-speed serial transceivers. This makes this board suitable for rapid prototyping, cpu-offloading, co-processing and custom interconnection networks. The FPGA is a Xilinx Virtex-4FX device [4]. This high performance state-of-the-art FPGA also contains two PowerPC blocks, fitting perfectly into the rapid prototyping and co-processing applications.

Immediately after publishing details about this rapid prototyping station for HTX, the Hypertransport Consortium (HTC) [5] promoted the project by including it as a reference design in their portfolio.

The remaining sections of this paper are organized as follows. The next chapter discusses the state of the art for peripheral interconnects. This leads to the motivation for a device with a HTX connector. Chapter 3 presents the basic architecture of HTX-Board, including a description of all major components and the sophisticated configuration methods. The chapter is finalized with a short overview of the board design, before the final conclusion and a short outlook on the next steps is presented.

2. Motivation

In today's Commodity-of-the-Shelf (COTS) computing systems the standard I/O interface is PCI, PCI-X or PCIe [6]. This implicates that there is no direct connection between peripheral device and main CPU. System and I/O interconnect are running complete different protocols and a bridge is required to convert these protocols.

The bridge introduces additional latency for device accesses. Typically, a bridge is not replicated so all I/O devices have to share the access to the system bus. This bottleneck limits the available bandwidth and closely-coupled systems are not possible.

For PCI and PCI-X [7] the I/O systems are bus-based, which introduces additional complexity for arbitration. A large protocol overhead is required to schedule the different connected devices, even if only one device is present.

PCI-Express (PCIe) [8][9] is a point-to-point interconnect and avoids the drawbacks of shared medium based systems. Instead it is based on high-speed serial links, which require serializers and de-serializers (SerDes) for communication. The source-clock is embedded in the data stream and a DC-balanced code like 8b/10b ensures clock recovery at the receiver. These requirements raise the complexity and thus increases the latency of accesses.

A complete different approach is the HTX interface. It makes the common Hypertransport protocol [10][11] available for peripheral devices by defining a standard expansion slot. Typically, HTX is only used in systems where the system interconnect also runs the Hypertransport protocol. Thus no protocol conversion is required and a bridge can be omitted. If the main CPU has only one HT link (e.g. the AMD Athlon64 [12]), a chain has to be set up with the HTX possibly at the end. The intermediate devices are tunnel devices, and the end of the chain is a cave device with reduced complexity. For CPUs with several HT links (like the AMD Opteron [13]) the HTX slot can be directly connected. This direct connection between device and CPU is optimal regarding latency, available bandwidth and overall performance. For SMP systems, the CPU contains a switch to redirect the traffic to the appropriate destination. Direct topologies like meshes are build up with these CPUs.

Summarized, the HTX system is a point-to-point interconnect with direct connection between peripheral device and CPU, avoiding any bridges or other intermediate protocol converters. HTX makes it possible for devices to connect directly to the system interconnect, an opportunity which was demanded by research and industry for a long time.

The goal of the work presented here is to design a reconfigurable device for HTX, enabling co-processing and rapid prototyping of applications which require close coupling to CPU and memory. Beside the basic idea of a HTX connection, the device architecture targets the area of interconnection networks. The HTX-Board is designed to be a first prototype for a next generation direct network, comparable to the ATOLL network [14] with more sophisticated features. Several possibilities for inter-node communication are embedded on the device, making the evaluation of different techniques possible. Another requirement is an embedded CPU core, which can be used as a network processor for certain applications. For an unconstrained use of such an embedded CPU core several additional auxiliary features are inevitable. For instance, an on-device DRAM, flash memory, Ethernet connection for loading bootstraps or an USB connection.

State of the art in reconfigurable logic devices are FPGAs, containing logic cells in the range of several 10.000 up to several 100.000. Other building block like embedded CPU cores, Ethernet Media-Access-Control (MAC) cores or other advanced logic blocks are also already included.

3. Basic Architecture

The key component of the system presented here is the FPGA, which is connected to the various components like communication devices, dynamic and flash memory or for auxiliary functions.

A Xilinx Virtex4-FX was chosen because it already contains CPU cores and a large number of high speed serial transceivers. Other features like support for dynamic reconfiguration, differential I/O with up to 1GHz or the embedded MAC cores are not mandatory, but fit very well in the architecture of the design and are not left unused. The FPGA is directly connected to the HTX interface with differential links. This connection is 16bit wide in each direction. Wider HT connection are not supported over HTX connectors. The complete power supply is also provided by the HTX connector, no external power is required.

Beside the HTX interface, the most important feature are the Small Form Factor Pluggable (SFP) Transceivers. Six of these SFPs are placed on the board, connected to the high speed serial links of the FPGA. These links can run speeds from 622Mbit/s up to 6.25GBit/s. One advantage of SFP is that the transceivers are pluggable. By exchanging the transceivers every kind of transmission is possible, electrical or optical over various connector types. Running all SFP transceivers at full speed, the bidirectional bandwidth is 60 GBit/s[1]. All six transceivers are accessible at the front panel of the board. The intend is to build up direct interconnection networks with a 3D-topology, for instance tori or meshes (or any other topology with a node degree [15] of not more than six).

1. Assuming an 8b/10b code and all six transceivers running at 6.25GBit/s.

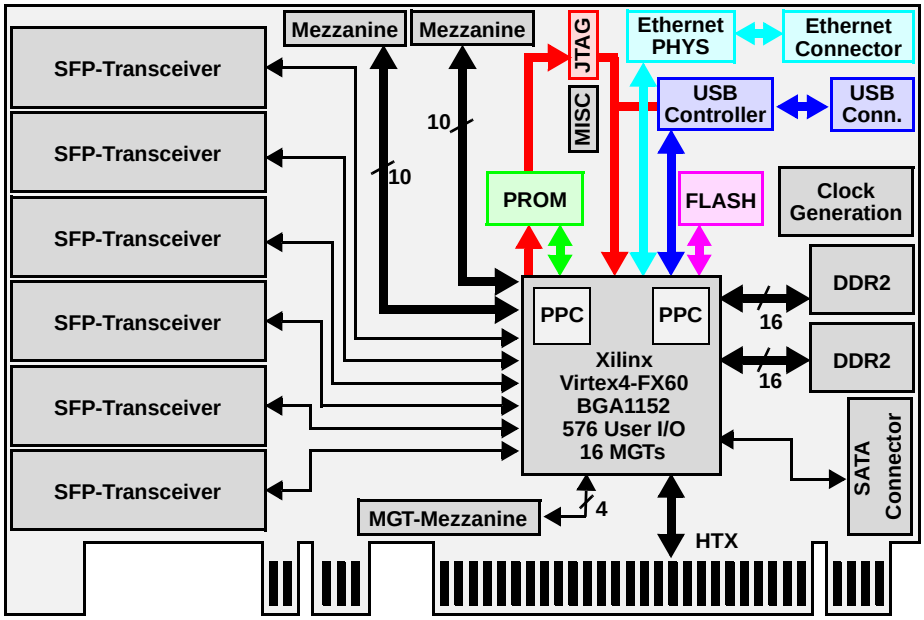

Figure 1.Block Diagram

The embedded CPU cores of the FPGA use a data width of 32bit, which is sufficient for the targeted applications. For unconstrained usage of the CPU cores, additional memory is required on the device. Furthermore, flash memory and an Ethernet interface is useful for loading the bootstrap of the CPU cores. Thus the FPGA is connected to auxiliary components like a DDR2-SDRAM, a flash memory and an Ethernet device. Regarding the data width of the DDR2-SDRAM interface it is optimal to match the data width of the CPU core, which is 32bit. Because the highest data width of available DDR2 devices is 16bit, two of them are placed on the board to match the data width.

The top-level block diagram is shown in Figure 1 on page 3 with all important components, mainly the FPGA, the HTX connector and the SFP array. These components are already shortly introduced. A deeper explanation of all major components is presented in the following. The block diagram also shows the placement of the different components on the board, with the FPGA in the middle surrounded by the others. Not shown in this block diagram is everything related to power, furthermore components of minor interest like the reset controller and voltage supervisor.

3.1 FPGA features

The requirements for the FPGA are basically at least six high speed serial links and at least one embedded CPU core. The high speed serial links should provide a bandwidth which is competitive, even in the next years.

The Xilinx Virtex4-FX60 (V4FX60) [16] device meets all the requirements and was available as engineering sample at the beginning of the project. The footprint of the FX100 device is the same. This allows to replace the FX60 with the larger one, with more logic cells and more block ram. The FPGA contains 16 Multi-Gigabit Transceivers (MGT), which is much more than required. The MGTs can operate at frequencies up to 6.25GBit/s, 10GBit/s is predicted for the future. From the 16 available MGTs, six are connected to the SFP array, four to a mezzanine connector and one to a SATA connector. The remaining ones are unused at the moment, but are available if applications demand a board layout change. The FX100 provides even 20 MGTs. Also included are two PowerPC 405 32bit RISC cores [17] (PPC), running with up to 450MHz. The PPC is well-known and provides state-of-the-art architecture with adequate performance. It has separate instruction and data caches (Harvard architecture [18]). Beside the PPC cores there are two Gigabit-Ethernet MAC cores. Together with a physical-layer device and an Ethernet connector it is possible to set up an Ethernet connection with up to 1Gbit/s. The I/O of the FPGA supports differential signalling, e.g. LVDS or LDT[1] compatible. Various other I/O standards are possible. For single-ended signalling, 576 user I/Os are available. This number is roughly bisected for differential signalling. The maximum speed of the I/O cells is 1GHz.

1. The Hypertransport protocol was formerly called Lightning Data Transfer Protocol (LDT).

3.2 HTX interface

The HTX standard [19] defines an interface between mainboard and peripheral device. This interface is bidirectional with differential signalling. The data width is limited to 16bit in each direction, but daughter cards can choose to implement only 8bits. The specification of the interface exactly follows the HT standard. This means that for every 8 data bits one clock signal is provided. This double-data rate clock is driven by the source of the data. For each direction there is one control signal to distinguish between control packets and data packets.

The HTX connector itself is composed of two standard PCI-Express connectors, but arranged in a reverse direction. The advantage is that widely available connectors can be used and that the backward mounting prevents from damage when inserting PCIe devices. Beside the HT signals, the connector provides power (12V and 3.3V), a 200MHz differential and a 66MHz single-ended reference clock, reset and power-ok signals. A system management interface is optional. Maximum specified operation frequency is 800MHz, but the PCIe connectors are known to work at higher speeds. So it could be chosen to operate the device at higher frequencies.

3.3 Auxiliary devices

If used in a mainboard, the direct connection between device and memory allows very fast and efficient access to the main memory system of the mainboard. But for stand-alone usage without being plugged into a HTX mainboard, DDR2-SDRAM memory is already included on the HTX-Board. Furthermore the embedded PPC cores can profit a lot from having dedicated memory. There are two DDR2 devices on the board, having together a data width of 32bit. This matches the data width of the PPC cores. DDR2 was chosen because of the guaranteed future availability, reduced power consumption, on-die termination, improved bus efficiency and migration to higher bus speeds and device densities.

An Ethernet connection is made possible by the Ethernet Physical Layer (PHY) device and a RJ-45 connector. Together with the MAC core embedded in the FPGA a Gigabit-Ethernet connection can be set up.

Beside the RAM and the PHY, a flash memory is available for non-volatile data storage. Again, the main target component is the PPC core. The flash memory allows to store a bootstrap loader. The loader only initializes the components needed to access another data source (e.g. the Ethernet interface). From this data source the remaining code is fetched to fully boot the PPC core.

Mezzanine extension boards can be connected to the HTX-Board over two mezzanine connectors. For differential signalling, the connection is 10bit wide in each direction. Additionally, an I^2C bus and a differential clock is available on the connector, together with 3.3V and 1.8V as power supply. The differential signals can also be used as single-ended, allowing up to 40 signals for both directions. The mezzanine connector is specified for frequencies up to 9.5GHz. A further connector is placed next to the FPGA and connected to four MGTs. This allows to easily access the FPGA over four bidirectional high speed serial links at maximum frequency. One MGT is routed to a SATA connector, e.g. to connect to a hard disk drive with an appropriate IP loaded into the FPGA.

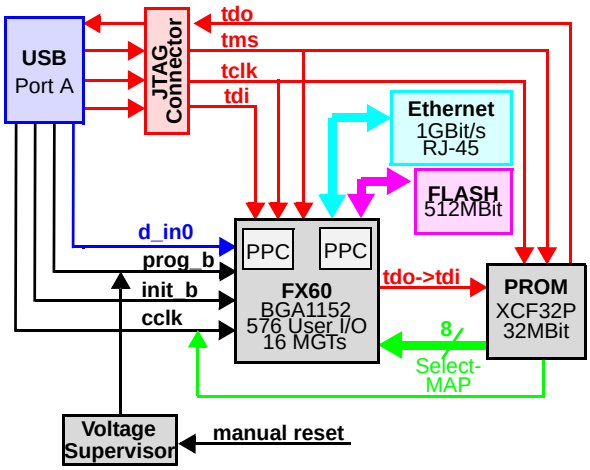

Figure 2.Configuration overview

3.4 FPGA configuration

A major task of each rapid prototyping station based on FPGAs is the configuration of these FPGAs. One demand is a robust programming mechanism to ensure re-configuration without hard reset. Especially for peripheral devices in computing environments another demand is the required time for configuration. In the time frame from applying power to the device until the first access over the I/O interconnect the FPGA must be loaded and configured. Otherwise the main system will not detect the device or the complete system hangs due to an incomplete I/O cycle. A computing system may also be composed by a large number of nodes, each equipped with a HTX-Board. For this case a widely available programming method must be available. The programming equipment typically has to be available for every single node, proprietary solutions are unwanted in such cases. Configuration should be possible with commodity parts, allowing the configuration process to run in parallel.

.

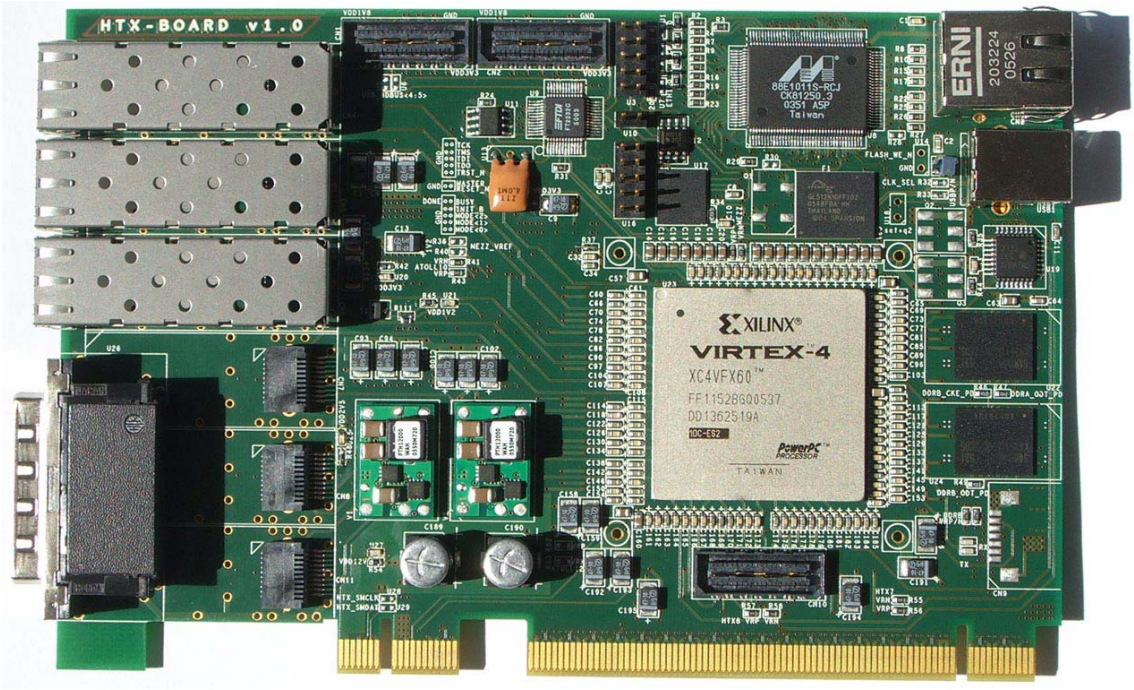

Figure 3.Photo of a prototype

These preliminary considerations led to the following programming system (Figure 2 on page 4). The FPGA itself provides various programming methods, only few of them are used here to meet the demands stated above.

The essential configuration method is JTAG [20]. JTAG has priority over every other method and works invariably. To access several devices using JTAG, the devices have to be connected in a daisy-chain fashion. Here, the other device in the chain beside the FPGA is a Platform Flash ROM (PROM) [21]. This device has a capacity of 32Mbit allowing to store a complete configuration bit stream in it. The PROM is configured using JTAG. Upon de-assert of the FPGA's and PROM's reset signal, the PROM starts to load the bit stream into the FPGA over a 8bit wide parallel interface, clocked at 40MHz. Compared to JTAG, this is a very fast configuration method, allowing the HTX-Board to be used as a peripheral device in a computing system. The configuration is performed before the device is accessed by the host system. Another advantage is that the configuration is stored non-volatile, rendering programming after each power-up unnecessary. The major drawback regarding programming over JTAG is that special hardware is required to generate a JTAG command stream. To allow configuration with commodity hardware, a USB UART [22] controller is connected to the JTAG chain. This controller has an interface designed for synchronous serial protocols like JTAG. This interface allows to access the JTAG chain using over the USB interface, which is a standard component of every modern computing system. The computing system is connected over a USB cable to the HTX-Board. A new developed software tool suite allows programming of the FPGA and the PROM over the JTAG chain.

Virtex4 devices offer the possibility to re-configure functional blocks during normal operation, using the Internal Configuration Access Port (ICAP) [23]. This dynamic reconfiguration is not affected by the configuration methods shown above. It is still fully available and provides the opportunity to change certain functional blocks of the FPGA without re-configuring the complete design.

3.5 Design of the Printed Circuit Board

This chapter covers the physical design of the Printed Circuit Board (PCB). Most interesting issues are the signal routing, the power distribution system and the finally resulting stack-up of the PCB. For a high speed design like this, special care has to be taken on signal integrity [24]. Signal speeds with up to 6.25Gbit/s require controlled impedance, and the Hypertransport signals must be routed length-matched. Next, the FPGA as the main component has a large power dissipation, which requires a sophisticated power distribution. It's large BGA footprint makes the fan-out of the footprint boundary a challenge. Beside all this, the total cost of the system always have to be kept in mind.

Sophisticated hand-routing of the design lead to a stack-up with only eight layers. This stack-up is optimal regarding the distribution of signal layers and reference planes. There are no adjacent signal layers, and the two innermost power/ground planes have a very small spacing. The core and auxiliary power supply for the FPGA (1.2V and 2.5V) have exclusive power planes for improved power distribution.

4. Conclusion

The advantages of rapid prototyping using reconfigurable logic like FPGAs are well-known. The only drawback of FPGAs is the limitation regarding speed and size of the contained logic. This is more than compensated by it's advantages for prototyping applications. The lead time and costs are much lower compared to an ASIC. Hardware/Software co-design is possible at an early design stage and various different solutions can be implemented in the FPGA. Errors can quickly be found and fixed. A system can be optimized by shifting the boundary between hard- and software to find the most appropriate partitioning.

The HTX-Board targets rapid prototyping applications, but is not limited to them. It is also suitable for co-processing, CPU-offloading and many other applications. Beside the FPGA it includes various building blocks, like DDR2 memory, flash memory, Gigabit Ethernet and SFP transceivers. Mezzanine cards can be directly connected to the HTX-Board if special functions are required.

Most important for this unique design is the HTX connector, allowing direct connection to CPU and memory of computing systems based on the AMD64 architecture. A connection from peripheral device to system interconnect without bridges has been demanded for a long time. The HTX expansion slot is an opportunity to connect devices directly and without intermediate bridges. The number of available mainboards with HTX expansion slots is growing rapidly, various devices for HTX are already available. The HTX-Board can accelerate the design of new devices by rapid protoyping.

This year AMD announced the Torrenza initiative. It capitalizes Hypertransport and the Direct Connect Architecture to enable the development of application specific co-processors. Developments already in progress include support for an HTX expansion slot. Third parties can develop custom-specific devices to include them in AMD64 platforms. In the next phase licensing of the coherent Hypertransport protocol will be available, allowing coherent HTX devices.

After publishing first details about the HTX-Board development, the Hypertransport Consortium decided to include it as a Reference Design within their product portfolio. This shows the ascribed importance to the HTX-Board and the need for such a design. Other members of the HTC can take advantage by using this Reference Design as a template for new developments. This re-use lowers required time and costs.

During the design of the HTX-Board, the total costs were always kept in mind. Sophisticated board design with a carefully chosen architecture lead to a PCB stack-up with only eight layers, which shows the effort invested in cost optimization.

Figure 3 on page 5 shows a photo of a prototype, currently the final design is on the way. On this prototype, a parallel SCSI connector replaces three SFPs for debugging purposes. For an unhindered view, the heat sink of the FPGA is left away. At the moment we are focussed on the completion and verification of the HT-Core. This core has a convenient interface towards FPGA applications allowing fast adoption by HTX-Board users. Developers of FPGA applications can rely on the HT core to connect towards the system.

5. Acknowledgements

We acknowledge the support of the European Community Research Infrastructure Activity under the FP6 "Structuring the European Research Area" Programme. (HadronPhysics: contract number RII3-CT-2004-506078)

6. References

[1] Hypertransport Consortium, *HyperTransport Technology I/O Link - White Paper*, www.hypertransport.org, July 2001.

[2] Hypertransport Consortium, *The Future of High Performance Computing: Direct Low Latency Peripheral-to-CPU Connections*, www.hypertransport.org, November 2005.

[3] Lloyd Dickman, *An Introduction to Pathscale InfiniPath*, www.pathscale.com.

[4] Xilinx Corporation, *Virtex-4 User Guide*, Document ug070 v1.5, www.xilinx.com, 2006.

[5] Hypertransport Consortium, http://www.hypertransport.org.

[6] Peter Sassone, *Commercial trends in off-chip communication*, Technical Report, Georgia Institute of Technology, May 2003.

[7] Edward Solari, George Solari, Willse Solari, *PCI & PCI-X Hardware and Software: Architecture and Design*, 5th Edition, Annabooks, San Diego, 2001.

[8] Peripheral Component Interconnect Special Interest Group (PCI SIG), *PCI Express base specification 1.0a*, 2002.

[9] Don Anderson, Ravi Budruk, Tom Shanley, *PCI Express System Architecture*, 1st Edition, Addison-Wesley Professional, 2003.

[10] Hypertransport Technology Consortium, *Hypertransport I/O Link Specification Revision 3.00*, Document #HTC20051222-0046-0008, 2006.

[11] Jay Trodden, Don Anderson, *HyperTransport System Architecture*, 1st Edition, Addison-Wesley Professional, 2003.

[12] Advanced Micro Devices (AMD), *AMD Athlon64 Product Data Sheet*, Publication #24659, 2006.

[13] .Advanced Micro Devices (AMD), *AMD Opteron Product Data Sheet*, Publication #23932, 2004.

[14] Holger Fröning, Mondrian Nüssle, David Slogsnat, Patrick R. Haspel, Ulrich Brüning, *Performance Evaluation of the ATOLL Interconnect*, IASTED Conference, Parallel and Distributed Computing and Networks (PDCN), Innsbruck, Austria, Feb. 15 - 17, 2005.

[15] Kai Hwang, Zhiwei Xu, *Scalable Parallel Computing*, 1st Edition, McGraw-Hill, 1998.

[16] Xilinx Corporation, *Virtex-4 Family Overview*, Document ds112 v1.5, www.xilinx.com, last access 2006.

[17] International Business Machines Corporation (IBM), *PowerPC 405 Embedded Cores*, http://www.ibm.com/chips/techlib/techlib.nsf/products/PowerPC_405_Embedded_Cores, last access 2006.

[18] J.L. Hennessy and D.A. Patterson, Computer Architecture, *A Quantitative Approach*, 2nd Edition, Morgan Kaufman, San Francisco, 1996.

[19] HTX standard document Hypertransport Consortium, *HyperTransport EATX Motherboard/Daughtercard Specification*, Document #HTC2004105-0040-0006, www.hypertransport.org, 2005.

[20] Institute of Electrical and Electronics Engineers, Inc. (IEEE), *IEEE 1149.1 Standard Test Access Port and Boundary Scan Architecture*, http://standards.ieee.org, 2001.

[21] Xilinx Corporation, *Platform Flash In-System Programmable Configuration PROMs*, Document ds123 v2.9, www.xilinx.com, 2006.

[22] Future Technology Devices International Ltd. (FTDI), *FT2232C Dual USB UART/FIFO IC Data Sheet*, www.ftdichip.com, last access 2006.

[23] Xilinx Corporation, *Virtex-4 Configuration Guide*, Document ug071 v1.4, www.xilinx.com, 2006.

[24] Howard Johnson, Martin Graham, *High-Speed Digital Design: A Handbook of Black Magic*, 1st Edition, Prentice Hall PTR, 1993.