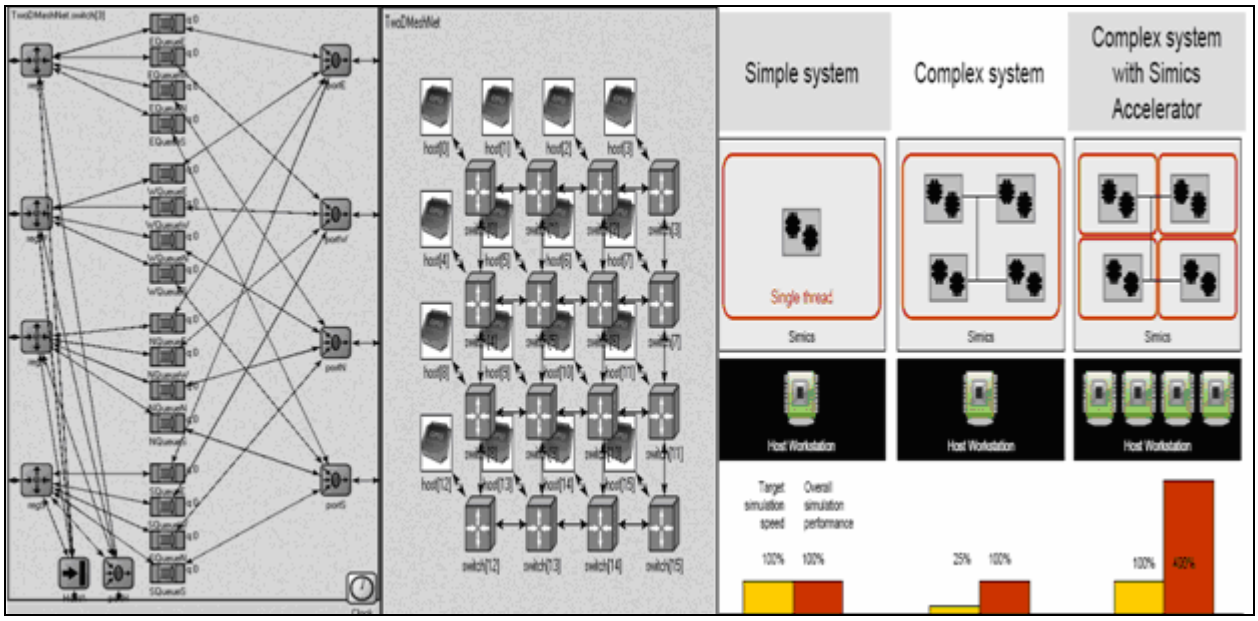
WP2: Multiprocessors communication networks for PetaFLOPS supercomputers

1. Main activities and results

Task 2.1: Development of collective network topology: network simulation and estimation of parameters. Based on mould of networks with couple thousands inputs and outputs, we have made comparative analyses of the most common networks - 3D Torus, High Radix Clos Network, Fat Tree and Flattened Butterfly. The conclusion was that most perspective networks for building up a supercomputer systems are High Radix, Clos Network and its modifications, as well as the combination of Fat Tree and hierarchical crossbar.These are the networks we have been examined. "Low-radix" networks, as the k-power n cubes, have their advantages and disadvantages. They can not take advantage of router bandwidth privilege, which force using of high-radix router – with a high number of tight connections. With modern technologies, based on Folded-Clos (or fat-tree) topology, high-radix networks ensure low latency and low working expenses, in contrast of conventional low-radix switches (commutators). The next generation Cray BlackWidow vector multiprocessor is going to be one of the first systems, which uses high-radix switches and implemented modified version of Folded-Clos network. Different type of switch architectures have been examined – all the way up from high-speed 4x4 (four by four) switch to a high performance switches, that implement High Radix Clos Network, hierarchical crossbar architecture, extended and improved main architectures, to High-Radix architectures of switches, as well as the distribution of virtual channels. Along with this, detailed analyzing was made of the accessory system communication networks and topologies. Routing algorithms, which are comparative and for which we have made the conclusion that global-adaptive routing for balanced topology is not necessary. Implementation of the UGAL routing algorithm, which overcome the great load-balance and uses the CLOS AD routing algorithm. By using the last, the implementation removes temporary load disbalance, defines flattened butterfly as similar to folded-Clos networks. Routing in "flattened butterfly" networks requires "jumping" around the nodes, node by node to the appropriate router, inside for the router jumps and jumps from the router to the designated destination at the final step. The basic and most commonly used routing algorithm for network management – STP (802.1D - spanning tree protocol and algorithm) and its improved successor RSTP (802.1w). (the last sometimes referred as PVST+ in cisco terminology). These routing algorithms` purpose is to prevent network loops, which often leads to partial or full collapse of the network, caused by flooding of messages. Compared to the STP, RSTP has the ability to restore much faster topologies, whose stability was damaged by disconnecting a link (link down event) or general change of the topology (insertion (or removing) of a network routing device for example). Modeling of meshed topologies has been developed to demonstrate the working mechanism of described protocols. Determination of bridge and port roles, according to the protocols behavior, has also been simulated. Multi-core processors TILE64 (be Tilera) and PC102/205 (by picoChip) have been examined and on the basis of profound comparative analysis was defined components array, based on which we are going to develop switch for the binding network. This switch will ensure low latency and high bandwidth, especially in the scope of streaming multimedia, network and graphic applications.As a result of the work in the frame of the project according to WP2, "Blade Center" is designed and implemented with the hardware platform, based on three Blade servers, HS21, Xeon Quad Core E5405 80w 2.00GHz/1333MHz/12MB L2, 2x1GB Chk, O / Bay SAS to disk subsystem IBM System Storage DS3400 Single Controller and hard disk drive subsystem for IBM 750GB Dual Port HS SATA HDD chassis specialist for Blade Center, IBM eServer BladeCenter (tm) H Chassis and recorder 2x2900W PSU UltraSlim, network switch Blade Center Chassis , BNT Layer 2 / 3 Copper Gb Ethernet Switch Module, Optical Switch chassis specialist for Blade Center, Brocade (R) 10-port 4 Gb SAN Switch Module with Optical Switch Module for IBM Short Wave SFP Module, together with the necessary wiring, special cabinet Blade Center, NetBAY S2 42U Standard Rack Cabinet and specialist-arrester Power Ultra Density Enterprise C19/C13 PDU Module (WW), is used to test the prepared models and to estimate the communication parameters of the developed models and to present the results.

With the partnership of professor Wolfgang Drzel, the WP2 team`s working plan is discussed. The initial idea was to investigate the 3D mesh network that connects the processor modules in the new generation supercomputers, but the new suggestion is to simulate the whole system with two networks – 3D-mesh and Fat tree with the methods of IBM End – to end Simulation of High performance Computing System. The IBM`s lab works with the "OMNET++ professional" network simulation software. We received the

management authorization to include our specialist into PhD Wolfgang Drzel`s team, after training them for the purpose. We negotiated with the colleagues from Research Laboratory, Zurich, to research in details the modules and blocks of the supercomputer system we are offering, composed by standard multi-core processors, while using the same simulation software – OMNET++ 4.0, but the free license for the academic purposes.
We acquired an academic license for the "Simics" platform as well, which was given up to us from the "Virtutech" consortium, which consist of the following components: Virtutech Simics Hindsight, Simics Accelerator, Simics Model Builder, Virtutech Simics Ethernet Networking, Simics Model Library and which is going to be used in a later stage of the study.



a) OMNET simulation view b) Simics concept for multicore systems

Figure 1: View in OMNET++ (a) and Simics (b)

The team working on the WP2 tasks have the necessary skills and preparation to work with both software packages installed on the "Blade Center", with parameter, described above, and located at the Technical University of Sofia, Bulgaria.
From the results we have received until now, and the analyses we have made, the necessity from development of new hybrid communication network and switch design, which are going to meet the requirements of high-performance computer systems becomes obvious! As well they are going to ensure the necessary low latency and high bandwidth, including PetaFlops systems, based on standard multi-core processor, for example AMD, Intel etc. During the first one year of the project, several meetings have been conducted with representatives of WP1 headed by Prof. Vladimir Lazarov, with discussions on the tasks placed in the appropriate WP and the possibilities for their realization are remarked on. The regular meetings of the team responsible for WP2 at the Technical University of Sofia are organized, with the discussions, powerful presentation on different topics according to WP tasks and comments on the results achieved so far. The results on this task are published in [NBKA_09], [BNL_09], [BNIR_09a], [BIIG_09a], [BGG_09a], [BGG_09a], [LPPMI_08].

Task 2.2: FPGA network realization. Functional and architecture designing of switch (with 4x4 gates/ports) have also been made along with designing of information packets, which can be exchanged by high-speed serial channels between computers in a multi-computer system with parallel architecture (cluster). For this purpose was used software development environment (IDE) for automated designing – WebPack and high-level programming language – VHDL, for the initial description of the switch.
The switch is intended to be multi-core system with shared RAM memory for I/O queues for packets. With the assistance of development environment, the switch is implemented on an extra large FPGA chip. The basic parameters of the implementation were given:

1) The part of used resources by the (chip) integral circuit (it should be noted that unused resources could be used for future upgrade/extension of the commutator).
2) The delay of the signals in the switch, which can be used to define its performance and latency.
With the purpose of logical check of functionality was made simulation at the RTL stage (stage of description – inter-register transfers) on the implemented switch with the assistance of ModelSIM. Results of the done research in the field of architectural characteristics of the modern extra-large FPGA chips are described in three articles, accepted to be published at the international science conference "Computer Science `09". These results were used in the designing, implementing and simulating of the developed switch on FPGA chip, [NZMK_09a], [MDK_09a], [K_09a].

2. Publications on the subject of the project, where the project DO 02-115/08 is quoted

a) published:

[NBKA_09] O. Nakov, P. Borovska, N. Kuchmova, D. Andreeva, Multiprocessor-based real-time control of a moving object, 8th WSEAS Int. Conf. on Applied Computer and Applied Computational Science (ACACOS '09), 20-22 May 2009, Zhejiang University of Technology, Hangzhou, China, Proceedings, 495-499

[BNL_09] P. Borovska, O. Nakov, M. Lazarova, PARMETAOPT – Parallel Metaheuristics Framework for Combinatorial Optimization Problems, IEEE International Workshop on Intelligent Data Acquisition and Advanced Computing Systems, Technology and Applications, 21-23 September 2009, Rende (Cosenza), Italy, Proceedings, 225-230

[LPPMI_08] L. Litov, P. Petkov, P. Petkov, S. Markov, N. Ilieva, Understanding of Human Interferon-Gamma Binding, Proc. of Fourth International Conference "ComputerScience'2008" and International Workshop on BioComputing'2008, Kavala, Greece, 18-19 Sept. 2008, pp.37÷42, ISBN: 978-954-580-254-6.

b) accepted:

[BNIR_09a] P. Borovska, O. Nakov, D. Ivanova, A. Ruzhekov, A Comparative Analysis of Next Generation High-End Switch Architectures, Fifth International Conference "Computer Science" 5-6 November 2009, International Workshops "Supercomputers Architecture and Applications", Technical University – Sofia, Bulgaria, Proceedings

[BIIG_09a] P. Borovska, D. Ivanova, K. Ivanov, G. Georgiev, Multi-core Architectures and Streaming Applications – trends, innovations and perspectives, Fifth International Conference "Computer Science" 5-6 November 2009, International Workshops "Supercomputers Architecture and Applications", Technical University – Sofia, Bulgaria, Proceedings

[BGG_09a] P. Borovska, G. Georgiev, I. Georgiev, 4x4 Switch Design and Simulation Analysis with OMNeT++, Fifth International Conference "Computer Science" 5-6 November 2009, International Workshops "Supercomputers Architecture and Applications", Technical University – Sofia, Bulgaria, Proceedings

[BGG_09a] P. Borovska, I. Georgiev, G. Georgiev, Modelling and Simulation Environments for Network on Chip Architectures: Survey, Fifth International Conference "Computer Science" 5-6 November 2009, International Workshops "Supercomputers Architecture and Applications", Technical University – Sofia, Bulgaria, Proceedings

[NZMK_09a] I. Nikolova, G. Zapryanov, P. Manoilov, E. Kucidimova, FPGA-based Architecture for Digital Image Visualization, Fifth International Conference "Computer Science" 5-6 November 2009, International Workshops "Supercomputers Architecture and Applications", Technical University – Sofia, Bulgaria, Proceedings

[MDK_09a] P. Manoilov, B. Delijska, P. Krivoshieva, FPGA Parallel DSP realized by Multiprocessor System on FPGA-Chip, Fifth International Conference "Computer Science" 5-6 November 2009, International Workshops "Supercomputers Architecture and Applications", Technical University – Sofia, Bulgaria, Proceedings

[K_09a] A. Kuncheva, DSP algorithms in modern programmable architecture - parallelisms of implementation, Fifth International Conference "Computer Science" 5-6 November 2009, International Workshops "Supercomputers Architecture and Applications", Technical University – Sofia, Bulgaria, Proceedings

3. Presentations and reports within the internal meeting held at the Technical University – Sofia

[1] S. Markov, PetaFlops Supercomputer Networks – one example. Workshop in IBM Research Laboratory Zurich, June 15, 2009

[2] Bulgarian Blue Gene/P – Overview of architecture, topologies, interface, SAN, I/O, switch ports, throughput

[3] System Area Network and I/O network – Omega topology

[4] Switches for Supercomputers – networks (SAN, I/O), interfaces, topologies, characteristics, communication parameters, ports, specifics, full overview and analysis

[5] Comparative Analysis of High-End switches all over the world, such as Voltaire, Grid Director, Myrinet

[6] Multicore processor "Tile64" – full overview, implementation in the projects all over the world, specifics, characteristics, communication parameters

[7] Multicore processor "picoArray" – full overview, implementation in the projects all over the world, specifics, characteristics, communication parameters

[8] Comparative Analysis of multicore processors Tile64 and PicoArray

[9] Streaming Applications - specifics and full overview

[10] "Simics" simulator – presentation of the functionality, features, realization of simple test examples

[11] Architectural Switch Design with picoArray – first steps

[12] Estimation of communication characteristics of system area networks based on OMNET++ simulations, topologies Fat tree and Omega for: a) Voltaire Switch; b) Myrinet c) BlackWidow

[13] System Area Network for BluGene/P using Radix switch

[14] OMNET++ - presentation of the product and its capabilities, installing software on "Blade Center", implemented examples

[15] Switch 4x4 Architectural Design, model and simulation results in the environment (OMNET + +) and analysis of statistics

[16] Realization of models (OMNET + +), for Fat Tree Topology and comparisons on parameters such as latency and throughput (Radix) in different trafficking and distributions

[17] YARC chip, models in OMNET++ and simulation results

4. Others

[1] Organizational and financial activities: Contract co-financing No: 091-CH-001-09 from 10.06.2009.

[2] Purchase of equipment under WP2.

[3] Additional activities to promote and disseminate the results within:

a) International scientific workshop "Supercomputer Architectures and Applications" and the departure of reports within the Fifth International Scientific Conference on Computer Science'2009 - 05-06.11.2009g., which is organized under the direct management of the Computer Systems Department at the Technical University – Sofia.