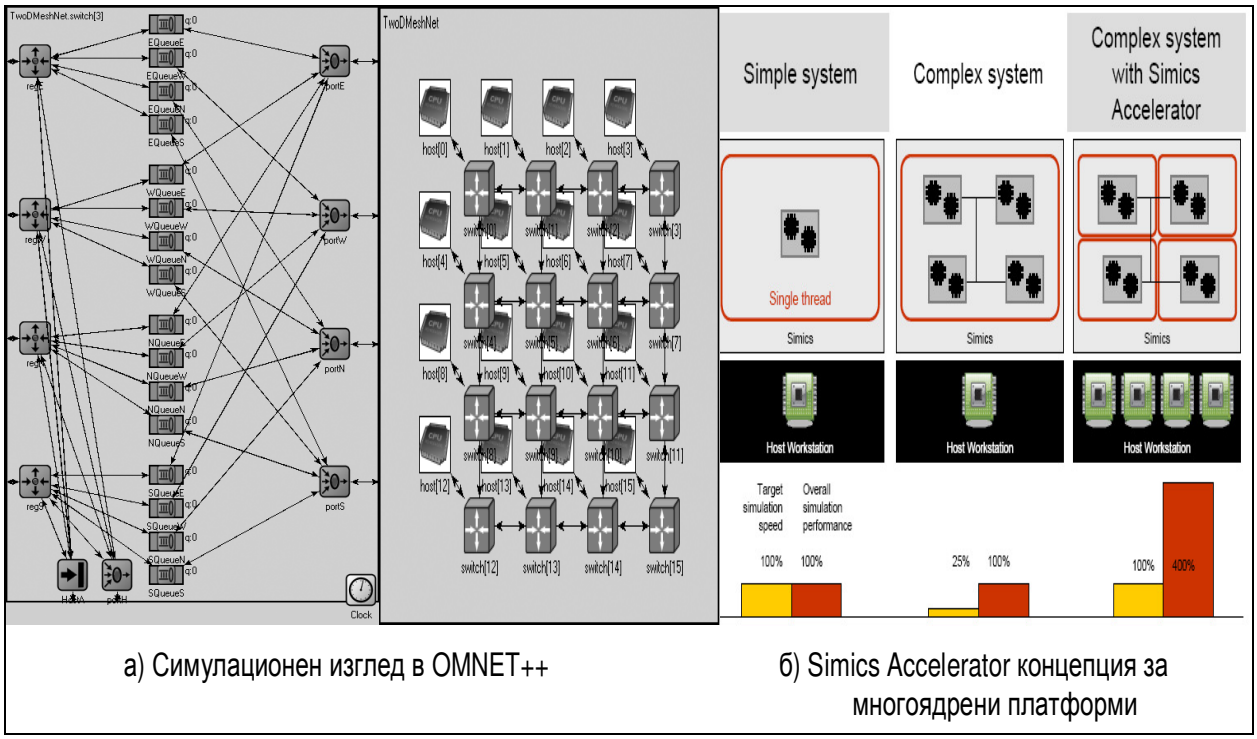
РП2: Многопроцесорни комуникационни мрежи за PetaFLOPS суперкомпютри

1. Основни дейности и резултати

<u>Задача 2.1:</u> Проектиране топологията на свързващата мрежа: мрежова симулация и измерване на параметрите. На базата на моделиране на мрежи на няколко хиляди входа и изхода, бе направен сравнителен анализ на основните параметрите на най-често използваните мрежи – 3D Torus, High Radix Clos Network, Fat Tree и Flattened Butterfly. Анализът показа, че перспективните мрежи за изграждане на суперкомпютърни системи са High Radix Clos Network и нейните модификации, както и комбинацията от Fat Tree и йерархичен управляем кросбар. Тези мрежи бяха изследвани. "Low-radix" мрежите, като k-ичните n-кубове имат свойте предимства и недостатъци. Те не могат да се възползват от предимството на "router bandwith", което налага използването на "high-radix router" – с голям брой тесни връзки. Със съвременната технология, базирани на “folded-Clos” (или fat-tree) топология, "high-radix" мрежите осигуряват ниска латентност и ниска цена в сравнение с мрежите с конвенционални "low-radix" комутатори. Следващото поколение “Cray BlackWidow” векторен мултипроцесор е една от първите системи, която използва "high-radix" комутатор и имплементира модифицирана версия на folded-Clos (вложена Clos) мрежа. Изследвани са широка област суич архитектури от висикоскоростен 4x4 комутатор до високопроизводителни комутатори, които реализират High Radix Clos Network, йерархична кросбар архитектура, усъвършенствани и разширени основни архитектури до High Radix архитектури на комутатори, както и разпределението на виртуалните канали, заедно с това е извършено и задълбочено анализиране и на съпътващите ги системни комуникационни мрежи и топологии. Изследвани са алгоритми за маршрутизация, които са съпоставими и за които правим извода, че не е необходимо наличие на глобално-адаптивна маршрутизация за балансирана топологията. Имплементация на UGAL маршрутизиращия алгоритъм, който преодолява това разпределяне на товара (load-balance) и внедрява CLOS AD маршрутизиращ алгоритъм, който премахва временния товарен (load) дисбаланс, определя “flattened butterfly” като сходна с “folded-Clos” мрежите. Маршрутизацията в мрежата “flattened butterfly” изисква „скок” от възел до съответния му локален рутер, нулеви или повече вътрешни за рутера „скокове” и накрая „скок” от рутера до определената дестинация. Изследвани са основните протоколи за управление на мрежи – STP (spanning tree protocol and algorithm) и подобрения нов стандарт Rapid Spanning Tree Protocol (RSTP). Целта на STP протокола е да предотвратява т.нар. LOOP (наводняване и сриване) на мрежата. Спрямо STP, RSTP предлага много по-бързо възстановяване на целостта на мрежата при промяна в състоянието на линк и изобщо при промяна в топологията на мрежата. Реализирани са модели за разпределение на ролите на всеки бридж в смесена топология с приложение на RSTP. Изследвани са и са анализирани многоядрените процесори, Tile64 и PicoArray и на база задълбочен сравнителен анализ е определена елементната база за изграждане на комутатор на свързваща мрежа, на основата на вграден многоядрен процесор, който да осигури ниска латентност и висока пропускателна способност, особено за потокови мултимедиини, мрежови и графични приложения. Като резултат от дейността по РП2 до момента е проектиран и имплементиран “Blade Center”, с хардуерна платформа, изградена на базата на: три броя високопроизводителни Blade сървъри, HS21, Xeon Quad Core E5405 80w 2.00GHz/1333MHz/12MB L2, 2x1GB Chk, O/Bay SAS с дискова подситема IBM System Storage DS3400 Single Controller и твърд диск за дискова подсистема IBM 750GB Dual Port HS SATA HDD с шаси за специализиран Blade Center, IBM eServer BladeCenter(tm) H Chassis и записващо устройство 2x2900W PSU UltraSlim, мрежов комутатор за шаси за Blade Center, BNT Layer 2/3 Copper Gb Ethernet Switch Module, оптичен комутатор за шаси за специализиран Blade Center, Brocade(R) 10-port 4 Gb SAN Switch Module с модул за оптичен комутатор IBM Short Wave SFP Module, заедно с необходимото окабеляване, специализиран шкаф за Blade Center, NetBAY S2 42U Standard Rack Cabinet и специализирано у-во за захранване Ultra Density Enterprise C19/C13 PDU Module (WW), която платформа се използва за тестване и оценка на комуникационните параметри на разработените до момента модели. С д-р Волганг Дрцел обсъдихме плана за работата на нашия екип, който работи по РП2. Първоначалната идея бе в новата генерация суперкомпютри да изследваме 3D-mesh мрежи за връзка между процесорните модули, но сега се предлага да се моделира цялата система с две мрежи 3D-mesh и Fat tree и да се симулира по методиката на IBM End – to end Simulation of High performance Computing System. Лабораторията на IBM работи с пакета за моделиране на

мрежи OMNET++ 4.0 professional. Имаме съгласието на Ръководството на лабораторията след обучение на наши специалисти да ги включим към екипа на д-р Волганг Дрцел. По принцип се договорихме с колегите от IBM Research Laboratory, Zurich подробно да изследваме модули и блоковете на предлаганата от нас суперкомпютърна система, изградена със стандартни многоядрени процесори, като използваме същия пакет програми за симулация OMNET++ 4.0, но безплатния вариант за академични звена.

Разполагаме също и с академичен лизенз за платформата "Simics" предоставена ни от консорциума "Virtutech", която обединяваа в себе си няколко компонента: Virtutech Simics Hindsight, Simics Accelerator, Simics Model Builder, Virtutech Simics Ethernet Networking, Simics Model Library и която се планира да бъде използвана на по-късен етап.



Фигура 1: Изглед в среда на OMNET++ (а) и на Simics (б)

Екипът отговорен за изпълнението по задачите на РП2 имат необходимите умения и подготовка за работа и с двата софтуерни пакета, които са инсталирани на “Blade Center” с цитираните по-горе параметри, локализиращ се в ТУ-София.
От получените до момента резултати и направените анализи се очертава ярко необходимостта от изграждането на нова хибридна комуникационна мрежа и дизайн на комутатор, които да удовлетворяват изискванията на високопроизводителните компютърни системи и да осигуряват ниска латентност и висока пропускателна способност, включително за PetaFlops системи изградени на база стандартни многоядрени процесори, напр. AMD или Intel.
През изтеклата първа година на проекта са проведени срещи с представители на РП1 оглавен от ст.н.с. II ст. Владимир Лазаров, с дискусии по задачите поставени в съответните РП и възможностите за тяхната реализация. Провеждат се регулярно срещи на екипа отговорен за РП2 към ТУ – София, с изнасяне на презентации, задълбочени дискусии и полезни коментари за постигнатите резултати. Резултати по тази задача са публикувани в [NBKA_09], [BNL_09], [BNIR_09a], [BIIG_09a], [BGG_09a], [BGG_09a], [LPPMI_08].

Задача 2.2: FPGA реализация на мрежата. Извършено е функционално и архитектурно проектиране на комутатор (с 4 входни и 4 изходни порта) на информационни пакети, които могат да бъдат обменяни по високоскоростни серийни канали между компютри в мултикомпютърна система с паралелна архитектура (cluster). За тази цел е използвана софтуерна развойна среда за автоматизирано проектиране WebPACK и език от високо ниво VHDL за входно описание на

комутатора. Комутаторът е проектиран като мултиядрена система с разпределени RAM – памети за входни и изходни опашки с пакети. С помощта на развойната среда комутаторът е имплементиран върху програмируема свръхголяма интегрална схема (FPGA- чип). Като основни параметри на имплементацията са отчетени:
1) процентът на заетите ресурси от интегралната схема (като свободните ресурси могат да бъдат използвани за бъдещо разширение на комутатора);
2) закъснението на сигналите в комутатора, от което би могло да бъде определено бързодействието и пропускателната му способност. С цел логическа проверка на функционирането му е извършено симулиране на RTL – ниво (ниво на описание – междурегистрови прехвърляния) на имплементирания комутатор с помощта на симулатора ModelSIM. Резултати от извършените изследвания върху особеностите на архитектурата на съвременните свръхголеми програмируеми FPGA - чипове са представени в две публикации, докладвани на Международната Научна Конференция "Computer Scence'09". Тези резултати са използвани при проектирането, имплементирането и симулирането на работата на разработения комутатор върху FPGA- чип. [NZMK_09a], [MDK_09a], [K_09a].

2. Публикации по темата на проекта, където е цитиран проект ДО 02-115/08

а) излезли от печат:

[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.

б) приети за печат:

[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, Proceeding

[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, Proceeding

[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, Proceeding

[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, Proceeding

[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, Proceeding

[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, Proceeding

[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, Proceeding

3. Презентации и доклади в рамките на проведени вътрешни срещи в ТУ – София

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

[2] Bulgarian Blue Gene/P – Общ преглед на архитектура, топологии, интерфейс, SAN, I/O, комутатор (switch) портове, пропускателна способност, т.н.

[3] Системна комуникационна мрежа и I/O мрежа – топология Омега

[4] Общ преглед на комутатори (switches) за суперкомпютри – мрежи (SAN, I/O), интерфейси, топологии, общи характеристики, комуникационни параметри, портове, специфики, пълен преглед и анализ

[5] Сравнителен анализ на "High-End" комутатори (switches) в световен мащаб, като Voltaire, Grid Director, Myrinet

[6] Многоядрен процесор "Tile64" – пълна информация и данни, включително цена, оценка, проекти реализирани с Tile64 в световен мащаб, специфики, интерфейси, комуникационни параметри,пълен преглед и анализ

[7] Многоядрен процесор PicoArray - пълна информация и данни, включителни цена, оценка, проекти реализирани с PicoArray в световен мащаб, специфики, интерфейси, комуникационни параметри, пълен преглед и анализ

[8] Преглед и оценка на всички проекти реализирани с многоядрени процесори Tile64 и PicoArray в световен мащаб

[9] Особености, специфики, пълен анализ и документация за Потоковите Приложения (Streaming Applications)

[10] Симулаторът Симикс "Simics"– представяне на функционални възможности, особености и ако е възможно реализация на прости тестови примери

[11] Архитектурно проектиране на комутатор с PicoArray – първи стъпки

[12] Оценка на комуникационните характеристики на системните мрежи на основата на симулации (OMNET++), топология „дебело дърво” и „Омега” за: а) Voltaire Switch; б) Myrinet в) BlackWidow

[13] Системна мрежа за BluGene/P с използване на Radix комутатор

[14] OMNET++ - представяне на продукта и неговите възможности, инсталиране на софтуера на работни стации и на “Blade Center”, реализирани на примери

[15] Архитектурата на 4x4 комутатор, съставяне на модел и симулационни резултати в среда на (OMNET++) и анализ на статистиките

[16] Реализация на модели в (OMNET++), за топология "дебело дърво" и сравнения по параметри като латентност и пропускателна способност (Radix) при различни трафици и различни разпределения (нормално, експоненциално и Гаусово)

[17] Архитектурата на YARC комутатор, модели в (OMNET++)

4. Други

[1] Организационно финансови дейности: Договор за съфинансиране No:091-CH-001-09 от 10.06.2009г.;

[2] Закупуване на техника по РП2;

[3] Допълнителни дейности по разпространяване и популяризиране на резултатите в рамките на:

а) Международен научен семинар “Суперкомпютърни архитектури и приложения” и изнасяне на доклади в рамките на Петата международна научна конференция „Компютърни науки’2009” – 05-06.11.2009г., която се организира под непосредственото ръководство на катедра “Компютърни системи” към Технически Университет – София.