«Направления развития суперкомпьютерных технологий в ОАО «НИЦЭВТ».

УДК 004.01

А.С.Симонов, А.И.Слуцкин, А.Е.Леонова

Аннотация

В статье рассматриваются основные направления развития суперкомпьютерных технологий в ОАО «НИЦЭВТ».

Ключевые слова: суперкомпьютеры, многопроцессорные вычислительные системы, коммуникационная сеть.

Суперкомпьютерные технологии (СКТ) играют важнейшую роль в инновационном развитии крупнейших мировых держав - США, Японии, Китая, стран Евросоюза. Рост производительности суперкомпьютерных систем позволяет решать совершенно новые задачи в области науки и промышленности. В Российской Федерации развитию СКТ также уделяется большое внимание. Создана Национальная суперкомпьютерная технологическая платформа, в которую входят несколько десятков организаций, осуществляющих разработку и внедрение СКТ в России и за рубежом; СКТ включены в перечень поддерживаемых государством критических технологий; Минпромторг РФ осуществляет финансирование отдельных проектов в области разработки СКТ отечественными предприятиями, правительство Союзного государства России и Беларуси профинансировало несколько союзных научно-технических программ по разработке и внедрению СКТ в науку и промышленность.

ОАО «НИЦЭВТ» - это одно из ведущих научных учреждений в России, в котором осуществляется развитие современных СКТ. ОАО «НИЦЭВТ» принимало активное участие в программах «СКИФ», «СКИФ-ГРИД» Союзного государства России и-Беларуси, а в программе «ТРИАДА» являлось ответственным исполнителем с российской стороны. В ОАО «НИЦЭВТ» выполняются научные исследования и разработки, обеспечивающие создание современных СКТ, сопоставимых с лучшими мировыми достижениями в этой области.

Работы в области СКТ в ОАО «НИЦЭВТ» ведутся по следующим направлениям:

- Информационно-аналитическая работа.
- Исследования в области архитектур многопроцессорных вычислительных систем (МВС).
- Создание и оценочное тестирование МВС.
- Разработки в области коммуникационных сетей суперкомпьютеров.
- Разработка вычислительных платформ для суперкомпьютеров.
- Разработка системного и промежуточного ПО.
- Разработка эффективных алгоритмов решения задач.

Информационно-аналитическая работа проводится с целью изучения и систематизации зарубежного опыта в области СКТ, выявления тенденций развития СКТ и использования полученных знаний при проведении разработки технических и программных средств. Наибольшее внимание уделяется работам таких признанных лидеров в данной области, как CRAY, SGI, IBM, Bull, Fujitsu, NEC, AMD, Intel, Mellanox, QLogic, Extoll. Результаты информационно-аналитической работы доводятся до инженеров-разработчиков на проводимых раз в квартал специальных обзорных

семинарах, что обеспечивает возможность развития разработок в русле мировых тенденций в области СКТ.

Исследования в области архитектур МВС призваны обеспечить возможность развития в ОАО «НИЦЭВТ» перспективных технологий создания массово-параллельных вычислительных систем транспетафлопсного класса для решения задач, которые могут быть эффективно распараллелены на большое число средне- и мелкогранулярных программных тредов.

Наибольший успех в данной области достигла фирма CRAY, создавшая линейку вычислительных систем TeraMTA/XMT с мультитредовой архитектурой, которая, в отличие от обычных суперскалярных процессоров, предполагает одновременную работу множества тредовых устройств, разделяющих общие функциональные устройства. Таким образом, удаётся, с одной стороны, обеспечить более эффективную загрузку функциональных устройств, а с другой стороны, обеспечить толерантность к задержкам при обращении к оперативной памяти за счёт низких, близких к нулю, накладных расходов на переключение тредовых устройств с контекста одного программного треда на другой.

Проведённые фирмой CRAY исследования показали, что на некоторых классах задач, для которых характерна интенсивная нерегулярная работа с памятью (шаблон Random Access), эффективность мультитредовых систем на несколько порядков выше, чем систем на основе суперскалярных процессоров. Особенно ярко это проявляется при работе с графами большой размерности, для которых характерны нерегулярная структура, низкая локализация данных в оперативной памяти, превалирование доступа к данным над вычислениями. Так, задача поиска кратчайшего пути в графе с 1 млн. вершин за одно и то же время была решена 20 000 узлов суперкомпьютера IBM Blue Gene /L и 4 узлами суперкомпьютера CRAY MTA 2!

В исследованиях в области архитектур активно применяются современные технологии имитационного моделирования. На рис. 1 приведены структура и показатели масштабируемости параллельной имитационной модели, разработанной в ОАО «НИЦЭВТ» в рамках проекта «Ангара».

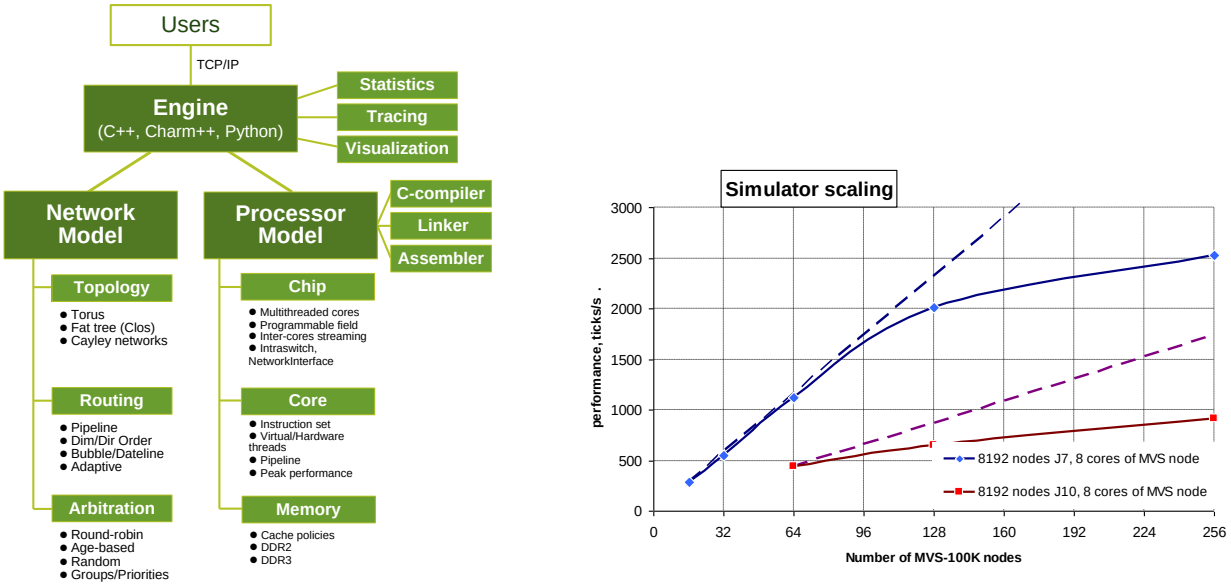

Рисунок 1 - Структура и показатели масштабируемости параллельной имитационной модели микропроцессора и маршрутизатора межузловой коммуникационной сети «Ангара»

При разработке модели использован язык Charm++, являющийся расширением языка C++ и обеспечивающий возможность неявного распараллеливания программ и балансировки нагрузки при решении задачи на МВС. График показывает хорошую масштабируемость модели при её вы-

полнении на вычислительной системе МВС-100К, установленной в Межведомственном суперкомпьютерном центре РАН.

На сегодня в ОАО «НИЦЭВТ» выполняется исследовательский проект, в рамках которого был разработан и в настоящее время проходит отладку на макете с FPGA собственный микропроцессор с мультитредовой архитектурой (рис. 2). В данном микропроцессоре реализовано одно вычислительное ядро с четырьмя независимыми конвейерами команд и 16 тредовыми устройствами. Система команд построена на базе MIPS с расширениями, обеспечивающими эффективное решение средне- и мелкогранулярных задач.

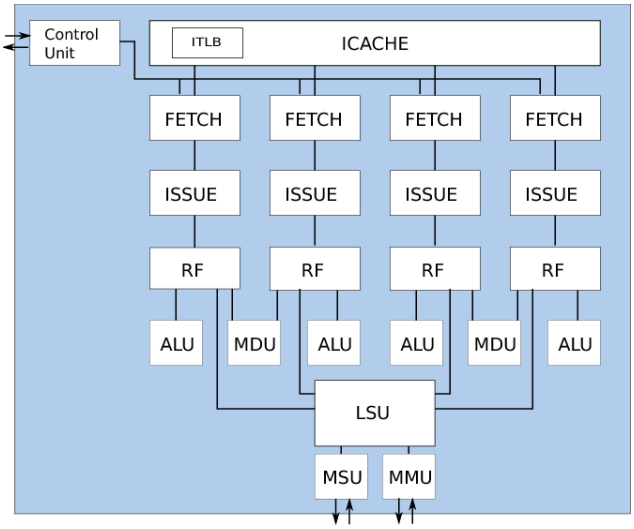

Рисунок 2 – Структура микропроцессора «Ангара-МТП» с мультитредовой архитектурой.

В дальнейшем в ОАО «НИЦЭВТ» планируется интегрировать данный микропроцессор с маршрутизатором коммуникационной сети в единую СБИС с целью использования его в качестве сопроцессора, перераспределения на него части вычислительной нагрузки - лёгких тредов, работающих с общими данными, и повышения, таким образом, эффективности решения прикладных задач.

Работа по созданию и оценочному тестированию МВС необходима для получения объективной информации о профилях решения задач на МВС, построенных как с использованием коммерчески доступных, так и заказных компонентов (рис. 3).

Рисунок 3 – МВС ЕС1720.05, созданная в ОАО «НИЦЭВТ» из коммерчески доступных компонентов с использованием макетных образцов маршрутизаторов высокоскоростной межузло-

вой коммуникационной сети ЕС8430 «Ангара», объединяющие вычислительные узлы в сеть «2D-тор» (3 × 3).

Основная цель состоит в выявлении факторов, существенно влияющих на эффективность решения на МВС прикладных задач конкретных потребителей и определении возможностей для совершенствования или адаптации технических и программных средств МВС. Полученный таким образом опыт используется при разработке технических решений перспективных продуктов СКТ - коммуникационной сети и вычислительной платформы.

Разработки ОАО «НИЦЭВТ» в области создания оригинальной коммуникационной сети для объединения узлов суперкомпьютеров на сегодня являются наиболее передовыми в нашей стране. Данная сеть по основным характеристикам, коммуникационной задержке и пропускной способности, существенно опережает имеющиеся на рынке коммерчески доступные решения, например, 10G Ethernet, Infiniband, и сопоставима с заказными коммуникационными сетями зарубежных суперкомпьютеров.

При разработке концепции коммуникационной сети был принят ряд решений относительно модели исполнения прикладных задач на вычислительной системе. Одно из наиболее важных состоит в том, что на вычислительном узле исполняются процессы только одной прикладной задачи. Это позволяет, во-первых, сократить накладные расходы на переключение между процессами, а во-вторых, необходимо поддерживать всего два виртуальных адресных пространства — решаемой задачи и операционной системы. Последняя необходима, помимо своих основных функций, для реализации параллельной файловой системы, функционирующей поверх коммуникационной сети и обеспечивающей возможность выполнения не только стандартных операций ввода/вывода, но и реализации контрольных точек.

Для объединения узлов МВС используется топология «4D-тор» (рис. 4). При этом в каждый вычислительный узел или узел ввода/вывода МВС устанавливается маршрутизатор и специализированное программное обеспечение, обеспечивающие поддержку распределённой глобально адресуемой памяти.

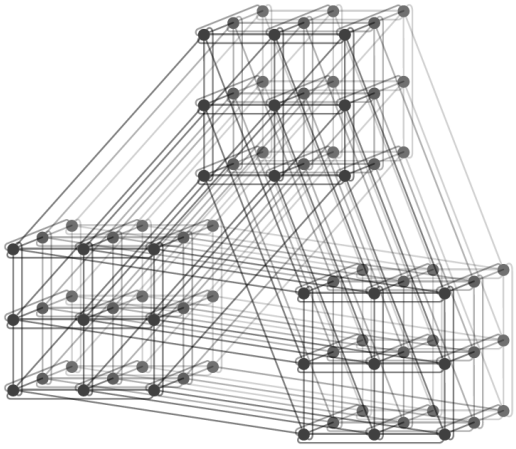

Рисунок 4 - Топология 4D-тор (3 × 3 × 3× 3)

Выбор топологии «многомерный тор» сделан, исходя из особенностей шаблонов обмена сообщениями по сети на интересующем классе задач. По сравнению с топологией Fat tree, используемой в сети Infiniband, данная топология значительно более толерантна к неравномерной загрузке, что подтверждается не только результатами имитационного моделирования, но и тестовыми расчётами на существующих суперкомпьютерах, например IBM Blue Gene/P, который имеет основную сеть с топологией 3D-тор. Кроме того, обеспечивается хорошая масштабируемость производительности МВС при решении сильно связанных задач.

Разработка данной коммуникационной сети ведется в ОАО «НИЦЭВТ» с 2006 года [1,2]. На подготовительном этапе было проведено тщательное изучение результатов зарубежных исследований и разработок, в том числе работ Уильяма Дэйли [3] и Хосе Дуато [4], а также архитектур IBM Blue Gene [5] и Cray SeaStar/Gemini [6].

За это время было создано три поколения макетных образцов маршрутизаторов на FPGA (рис. 5) и вычислительные кластеры на их основе (рис.3), отработаны решения по передаче сообщений и взаимодействию с коммерчески доступными процессорами вычислительных узлов с использованием современного интерфейса PCI Express.

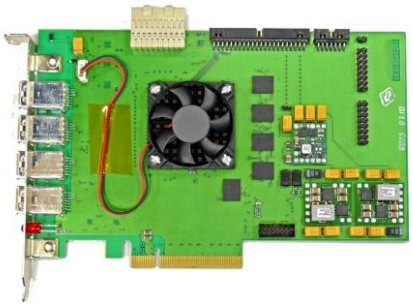

Рисунок 5 – Макетный образец маршрутизатора ЕС8430 «Ангара»

В настоящее время выполняется разработка заказной СБИС маршрутизатора на технологических нормах 65 нм, прототипы планируется изготовить на фабрике TSMC в первой половине 2012 г. Структурная схема маршрутизатора ЕС8430 «Ангара» приведена на рис. 6.

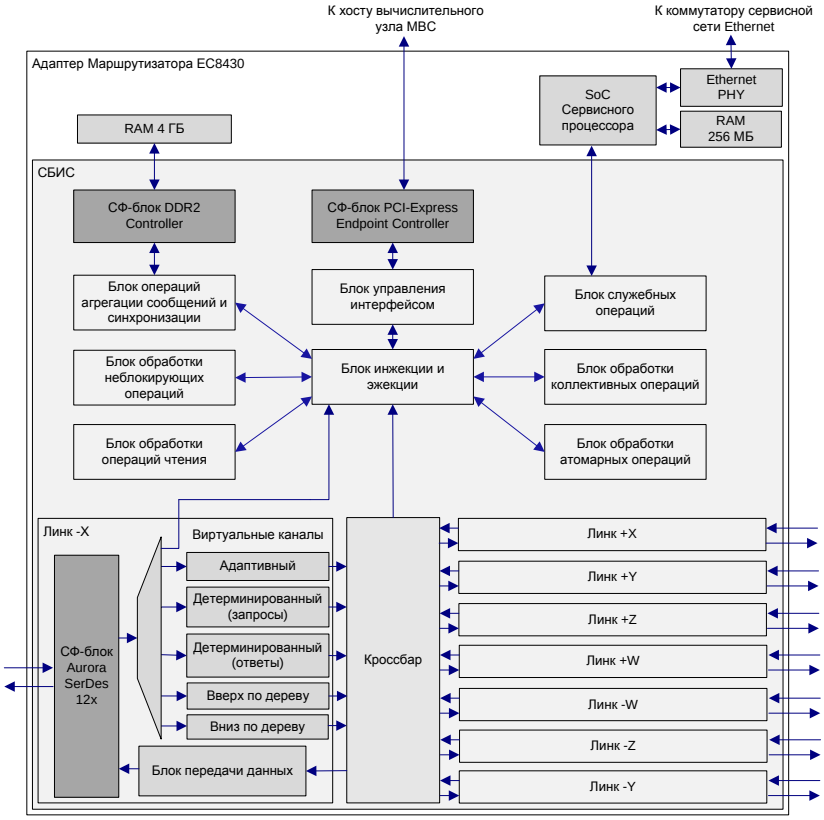

Рисунок 6 – Структурная схема маршрутизатора ЕС8430 «Ангара»

В маршрутизаторе ЕС8430 «Ангара» реализованы следующие функциональные возможности:

- Протокол надёжной доставки пакетов.
- Детерминированная и адаптивная передача данных.
- Аппаратная поддержка многопоточности.
- Аппаратная реализация операций:
  - Запись в память удалённого узла.
  - Запись в память удалённого узла с сохранением концептуальности.
  - Атомарные операции в памяти удалённого узла.
  - Чтение из памяти удалённого узла.
  - Неблокирующая запись больших массивов в память удалённого узла.
  - Чтение больших массивов из памяти удалённого узла.
- Аппаратная поддержка операций барьерной синхронизации.
- Реализация сборки массивов в памяти маршрутизатора с последующим копированием в память узла.
- Сеть коллективных операций (операции broadcast, reduce).
- Система обеспечения отказоустойчивости.

Основным рекомендуемым режимом работы разрабатываемой коммуникационной сети, на котором обеспечиваются наилучшие показатели производительности МВС, является режим с использованием библиотеки SHMEM. При этом общий объём глобально адресуемой маршрутизатором памяти МВС составляет до 128 ПБайт, а объём памяти, адресуемой на каждом узле, — до 2 ТБайт.

В ОАО «НИЦЭВТ» большое внимание уделяется разработке собственных универсальных и специализированных вычислительных платформ. За последние 5 лет создано целое семейство компактных вычислительных модулей для промышленной автоматизации и встраиваемых решений (рис. 7).

Полученный опыт позволил сделать качественный шаг – перейти к разработке суперкомпьютерной платформы-лезвия ЕС1740 «Ангара» на основе современных многоядерных микропроцессоров AMD Magny-Cours / Bulldozer в сокете G34 и коммуникационной сети ЕС8430 «Ангара». Структурная схема платформы приведена на рис. 8. Данная платформа представляет собой законченный полнофункциональный вычислительный узел, который может быть использован при создании МВС, масштабируемых до транспетафлопсного уровня производительности и обеспечивающих высокие показатели эффективности при решении реальных задач.

В ходе разработки платформы инженерами ОАО «НИЦЭВТ» анализируется опыт создания наиболее передовых суперкомпьютерных систем: Fujitsu K-Computer, CRAY XT5/XT6/XMT, IBM Blue Gene/P, /Q, СКИФ-АВРОРА, применяются современные средства проектирования, моделирования и инженерного анализа, позволяющие на ранних стадиях выявить возможные проблемы целостности сигналов, рассчитать и оптимизировать тепловые режимы работы наиболее теплонагруженных элементов и системы в целом, реализовать компоновку, обеспечивающую высокую плотность установки вычислительных узлов в стойки. Всё это позволяет получить конкурентоспособный продукт, привлекательный для использования при создании МВС высшего диапазона производительности и энергоэффективности.

«а» «б» «в»

Рисунок 7 – Встраиваемые вычислительные модули, разработанные и серийно изготавливаемые в ОАО «НИЦЭВТ» («а» - AMD Geode LX800, «б» - на основе VortexDX, «в» - на основе Intel Atom N270/330).

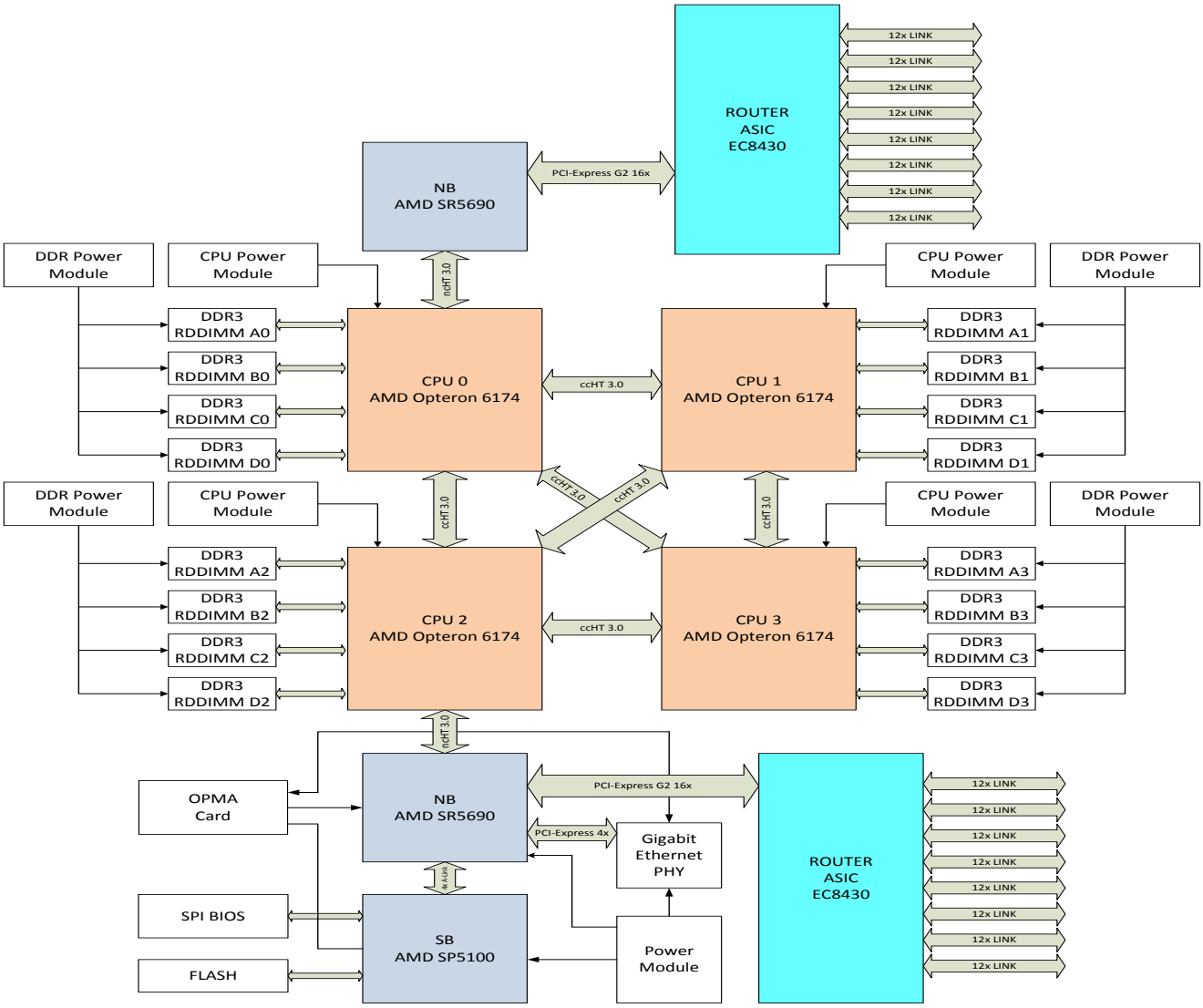

Рисунок 8 - Структурная схема разрабатываемой в ОАО «НИЦЭВТ» суперкомпьютерной платформы на основе процессоров AMD и коммуникационной сети EC8430 «Ангара».

В дополнение к основной линии развития, проекту «Ангара», в ОАО «НИЦЭВТ» ведутся проекты по созданию компактных суперкомпьютерных систем, проект «Ангара-TCCluster», и технических средств для распределённых бортовых вычислительных систем, проект «Онега».

Основная идея проекта «Ангара-TCCluster» состоит в использовании задела, нарабатываемого в ходе разработки вычислительной платформы ЕС1740 «Ангара» на основе процессоров AMD, для создания компактных МВС, объединение узлов в которой будет осуществляться в топологию «2D-тор» с использованием встроенного в микропроцессоры фирмы AMD интерфейса HyperTransport. Такое решение позволит обеспечить беспрецедентно низкие, менее 100 нс, коммуникационные задержки при обращении в память соседних узлов.

В рамках проекта «Онега» планируется использовать задел ОАО «НИЦЭВТ» в области СКТ с целью разработки комплекта модулей в форматах EPIC-Express и PC/104-Express, пригодного для создания на его основе средств промышленной автоматизации и распределённых бортовых информационно-управляющих систем различного назначения и включающего компактную вычислительную платформу, коммуникационное оборудование и модули прецизионных АЦП/ЦАП.

В качестве основы для вычислительной платформы для проекта «Онега» выбрана линейка микропроцессоров Atom фирмы Intel, специально разработанная для использования во встраиваемых системах и имеющая низкие показатели энергопотребления. При этом, в отличие от применяющихся в данном сегменте микропроцессоров с архитектурой ARM и PowerPC, данная линейка имеет ставшую стандартом архитектуру x86, что позволяет существенно сократить затраты на разработку программного обеспечения.

Для обеспечения возможности создания высоконадёжных отказоустойчивых распределённых систем под проект «Онега» проводится адаптация разработанного в ОАО «НИЦЭВТ» маршрутизатора межузловой коммуникационной сети, который позволит объединять узлы в топологию «2D-тор» с использованием проводных и оптоволоконных каналов связи.

При создании модулей прецизионных аналого-цифровых и цифро-аналоговых преобразователей будет использован опыт разработки подобных модулей для бесплатформенных инерциальных навигационных систем, где специалистами ОАО «НИЦЭВТ» были разработаны оригинальные методики обеспечения высокой, более 20 эффективных разрядов, точности преобразования в расширенном диапазоне рабочих температур.

Также в ОАО «НИЦЭВТ» развивается направление создания многопроцессорных серверов на основе линейки процессоров Intel Atom. В настоящее время на основе разработанных и уже запущенных в серийное производство модулей в формате EPIC и коммуникационной сети ведётся разработка сервера с высокой плотностью размещения процессоров с низким энергопотреблением. Такие решения, по сравнению со стандартными серверами-лезвиями, являются более предпочтительными для крупных ЦОД, одной из проблем в которых является необходимость подвода больших электрических мощностей и отвода тепла. По данному направлению ОАО «НИЦЭВТ» начал сотрудничество с одним из крупнейших московских дата-центров M1, в рамках которого на основе опыта, полученного при его создании и в процессе эксплуатации формируется технический облик перспективного энергоэффективного сервера, обеспечивающего решение задач хостинга интернет-ресурсов.

Разрабатываемые в ОАО «НИЦЭВТ» программные средства призваны обеспечить высокий уровень производительности МВС при решении реальных задач, в том числе с интенсивным нерегулярным доступом к памяти (DIS-класс, data intensive system). В рамках развития СКТ в ОАО «НИЦЭВТ» разработаны библиотеки интерфейса низкого уровня и реализация библиотеки SHMEM для маршрутизатора межузловой коммуникационной сети ЕС8430 «Ангара», проведена адаптация библиотеки MPI 2.0, проводится апробация перспективных языков параллельного про-

граммирования PGAS-класса (UPC, CAF). Для разработки прикладных программ поддерживаются стандартные математические библиотеки, такие как BLAS, LAPACK, SCALAPACK, FFTW и др.

Важное место в разработке системного ПО занимает создание собственной версии параллельной файловой системы LUSTRE, функционирующей поверх низкоуровневого ПО маршрутизатора межузловой коммуникационной сети ЕС8430 «Ангара». Это позволит обеспечить высокую пропускную способность между вычислительными узлами и узлами ввода/вывода при выполнении операций загрузки исходных данных, выгрузки результатов счёта и, что особенно важно, при сохранении и восстановлении с контрольных точек.

Для привлечения заказчиков при помощи демонстрации интересующих задач на разрабатываемом оборудовании в ОАО «НИЦЭВТ» создана лаборатория DISLab. Основными задачами данной лаборатории являются исследование задач DIS-класса, реализация и оптимизация эффективных алгоритмов их решения на создаваемых в ОАО «НИЦЭВТ» технических и программных средствах. Отдельное внимание в лаборатории уделяется оценочному тестированию суперкомпьютеров и коммуникационных сетей, а также перспективным подходам к параллельному программированию с использованием односторонних коммуникаций и языков PGAS-класса.

Первые результаты работы лаборатории DISLab показали перспективность выбранных в ОАО «НИЦЭВТ» направлений развития СКТ. Деградация реальной производительности при масштабировании целого ряда задач на вычислительных системах, оснащённых коммуникационной сетью ЕС8430 «Ангара», значительно ниже, чем при использовании коммерчески доступных решений. Особенно ярко это проявляется на задачах, для которых характерен интенсивный нерегулярный доступ ко всей распределённой по вычислительным узлам памяти, например задачах из теории графов.

С целью привлечения внимания к задачам DIS-класса полученные в лаборатории результаты в большинстве случаев будут открытыми и позволят специалистам из различных отраслей использовать наработанный опыт при выборе вычислительных систем и решении собственных прикладных задач.

Литература

1. А.И. Слуцкин, А.С. Симонов. «Развитие суперкомпьютерных технологий в ОАО «НИЦЭВТ», Научно-техническая конференция «Перспективные направления развития средств вычислительной техники»: Сборник тезисов докладов (г. Москва, 28 июня 2011 г.).

2. А. Симонов, И. Жабин, Д. Макагон. «Разработка межузловой коммуникационной сети с топологией «многомерный тор» и поддержкой глобально адресуемой памяти для перспективных отечественных суперкомпьютеров», Научно-техническая конференция «Перспективные направления развития средств вычислительной техники»: Сборник тезисов докладов (г. Москва, 28 июня 2011 г.).

3. William Dally and Brian Towles. Principles and Practices of Interconnection Networks. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2003.

4. Jose Duato, Sudhakar Yalamanchili, and Ni Lionel. Interconnection Networks: An Engineering Approach. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2002.

5. N.R. Adiga, M.A. Blumrich, D. Chen, P. Coteus, A. Gara, M.E. Giampapa, P. Heidelberger, S. Singh, B.D. Steinmacher-Burow, T. Takken, M. Tsao, P. Vranas. Blue Gene/L torus interconnection network, IBM J. RES. & DEV. VOL. 49 NO. 2/3 MARCH/MAY 2005.

6. Robert Alverson, Duncan Roweth, and Larry Kaplan. The Gemini System Interconnect, In Proceedings of the 2010 18th IEEE Symposium on High Performance Interconnects (HOTI ’10), IEEE Computer Society, Washington, DC, USA, 83-87.