На правах рукописи

Семенов Александр Сергеевич

РАЗРАБОТКА И ИССЛЕДОВАНИЕ АРХИТЕКТУРЫ ГЛОБАЛЬНО АДРЕСУЕМОЙ ПАМЯТИ МУЛЬТИТРЕДОВО-ПОТОКОВОГО СУПЕРКОМПЬЮТЕРА

специальность 05.13.15 – Вычислительные машины, комплексы и компьютерные сети

АВТОРЕФЕРАТ

диссертации на соискание ученой степени
кандидата технических наук

Москва – 2010

Работа выполнена в отделе архитектуры и программного обеспечения суперкомпьютеров ОАО “НИЦЭВТ”.

Научный руководитель: кандидат физико-математических наук
Эйсымонт Леонид Константинович.

Официальные оппоненты: доктор физико-математических наук, профессор
Томилин Александр Николаевич;

кандидат физико-математических наук
Ефимкин Кирилл Николаевич.

Ведущая организация: ОАО “Институт Электронных Управляющих Машин им. И. С. Брука”.

Защита диссертации состоится “13“ декабря 2010 г. в 10:00 часов на заседании диссертационного совета Д 212.125.01 в Московском авиационном институте (государственном техническом университете) по адресу 125993, г. Москва, А-80, ГСП-3, Волоколамское шоссе, д. 4.

С диссертацией можно ознакомиться в библиотеке Московского авиационного института.

Автореферат разослан “12“ ноября 2010 г.

Ученый секретарь
диссертационного совета Д 212.125.01
к.т.н. Корнеенкова А. В.

Общая характеристика работы

Актуальность темы

Важнейшими проблемами современных высокопроизводительных вычислений являются развиваемая на многих задачах низкая реальная производительность суперкомпьютеров в сравнении с заявленной пиковой производительностью, а также низкая продуктивность параллельного программирования суперкомпьютеров. Эти проблемы существенно усиливают ряд других серьезных трудностей, например: высокое энергопотребление и чрезмерная стоимость систем.

Главные работы по решению проблем низкой реальной производительности и низкой продуктивности программирования в настоящее время ведутся фирмами Cray (проект Cascade) и IBM (проект PERCS) с целью создания суперкомпьютеров стратегического назначения (СКСН) транспетафлопсного уровня реальной производительности с перспективной мультитредово-векторной архитектурой и глобально адресуемой памятью. Эти проекты ведутся в рамках крупной американской программы DARPA HPCS, начатой в 2002 году. Всего лишь с задержкой в 4 года к работам по созданию СКСН такого же типа приступили ведущие фирмы и государственные организации Японии (проект Keisoku-Keisan-Ki) и Китая (проект "программа 863"). При этом наиболее близок к основным решениям проектов программы DARPA HPCS проект создания китайской перспективной СКСН, который предполагает создание собственного микропроцессора и собственной специальной коммуникационной сети с большой пропускной способностью и малым диаметром. В России работы по перспективной СКСН ведутся в рамках проекта "Ангара" по созданию мультитредово-потокового суперкомпьютера с глобально адресуемой памятью.

Главной задачей, решаемой в этих проектах, является преодоление так называемой проблемы «стены памяти». Суть этой проблемы состоит в сложившемся на сегодняшний день значительном отставании (сотни раз) времени выполнения операций с оперативной памятью от времени выполнения арифметико-логических операций в процессоре. Эта проблема обусловлена особенностями развития микроэлектронной компонентной базы – задержка обращения к внекристальной DRAM-памяти с учетом промахов в кэш-память дескрипторов сегментов и/или страниц может составлять 300-500 тактов процессора. Задержка обращения к коммуникационной сети может составлять десятки тысяч тактов процессора. Значимость этой проблемы усилена качественными изменениями требований перспективных прикладных программ – увеличение необходимого задаче логически неделимого объема памяти, рост доли команд обращений к памяти,

нарастающее ухудшение пространственно-временной локализации обращений к памяти (нерегулярность). Следует отметить важность приложений с интенсивной нерегулярной работой с памятью (сокращенно DIS-задачи), особенно в областях обеспечения национальной безопасности.

Наличие проблемы "стены памяти" приводит к тому, что на DIS-задачах из-за простоев процессора при ожидании данных из памяти реальная производительность (sustained performance) СКСН может деградировать до 0.1-5 % от пиковой производительности. Эта проблема также не позволяет существенно масштабировать реальную производительность при увеличении количества процессоров, используемых при выполнении задачи. В результате многие DIS-задачи практически невозможно решить за приемлемое время на существующих суперкомпьютерах.

Решение проблемы "стены памяти" позволит реально работать с глобально адресуемой памятью и перейти на одностороннее взаимодействие параллельных процессов, что важно для эффективного выполнения программ на языках класса PGAS (Partitioned Global Address Space – языки UPC, CAF), а также для перспективных языков с иерархическим описанием параллельных программ (языки Chapel, X10 и Fortress). В целом, это позволит повысить продуктивность параллельного программирования, по оценкам мировых экспертов, в 10-40 раз.

Повышение реальной производительности и продуктивности программирования повысит коэффициент полезного использования как оборудования, так и человеческого ресурса, поэтому повлияет на снижение энергопотребления, снизит стоимость систем и сократит сроки разработки приложений.

Решение проблемы «стены памяти» в создаваемых перспективных СКСН производится за счет комплексного использования новых архитектурных принципов построения процессоров, памяти и коммуникационной сети, а также применения новых вычислительных моделей программ и соответствующего системного и прикладного программного обеспечения. Архитектура глобально адресуемой памяти определяет организацию виртуального адресного пространства, методы его защиты, отображения на физическую память, алгоритмы трансляции виртуальных адресов в физические. Вопрос выбора этой архитектуры является одним из основных в перспективных СКСН. Выбор этой архитектуры должен быть компромиссным. С одной стороны, она не должна ограничивать функциональность и эффективность ее использования в приложениях. С другой стороны она не должна быть слишком сложной, чтобы ее реализация была эффективной в контексте применяемых в СКСН других решений.

Данная диссертационная работа посвящена разработке архитектуры глобально адресуемой памяти СКСН «Ангара», исследованию ее

функциональности, эффективности и ее реализации в базовом для этой СКСН мультитредово-потоковом микропроцессоре с учетом выбранной коммуникационной сети и памяти, а также исследованию применяемых программ с разной мультитредовой организацией и используемыми моделями памяти.

Цель и задачи диссертационной работы

Целью диссертационной работы является разработка архитектуры глобально адресуемой памяти для суперкомпьютера мультитредово-потокового типа. Для достижения этой цели решались следующие задачи:

1. Разработка архитектуры глобально адресуемой памяти мультитредово-потокового суперкомпьютера “Ангара”, которая является развитием вариантов организации глобально адресуемой памяти известных суперкомпьютеров заказного типа.

2. Реализация глобально адресуемой памяти в параллельной имитационной модели суперкомпьютера “Ангара” для исследования эффективности применения разработанной архитектуры глобально адресуемой памяти при решении прикладных задач.

3. Исследование возможностей использования разработанной архитектуры глобально адресуемой памяти в сочетании с другими особенностями суперкомпьютера “Ангара” для достижения высокой реальной производительности.

4. Исследование возможности использования разработанной архитектуры глобально адресуемой памяти при реализации перспективных языков параллельного программирования PGAS-класса.

Научная новизна работы

Полученные результаты являются новыми и состоят в следующем:

1. Разработана архитектура глобально адресуемой памяти мультитредово-потокового суперкомпьютера.

2. Разработаны и исследованы блоки, имитирующие работу глобально адресуемой памяти в составе параллельной программной потактовой имитационной модели суперкомпьютера «Ангара».

3. Показано, что применение разработанной глобально адресуемой памяти в сочетании с другими возможностями мультитредово-потокового суперкомпьютера “Ангара” в сравнении с опубликованными данными по существующим суперкомпьютерам позволяет:

— повысить реальную производительность на задачах с интенсивным нерегулярным доступом к памяти до одного-двух порядков при совпадении реальной производительности на остальных задачах;

— использовать при получении высокой реальной производительности короткие односторонние коммуникационные передачи, при помощи которых

проще создавать параллельные программы по сравнению с традиционным программированием с использованием MPI.

4. Продемонстрирована возможность использования разработанной архитектуры глобально адресуемой памяти при реализации перспективных языков параллельного программирования PGAS-класса на примере языка UPC.

Разработанная архитектура защищена патентом Российской Федерации № 2396592 "Способ организации глобально адресуемой общей памяти в многопроцессорной ЭВМ" от 21 октября 2008 года и обладает следующими основными свойствами:

— имеет сегментно-страничную организацию с двухуровневой виртуализацией адресов и широким диапазоном размеров сегментов и страниц, а также возможностью работы с сегментами сверхбольшого размера;

— обладает расширенными возможностями отображения виртуальной памяти на физическую за счет использования блочного и блочно-циклического методов, а также зашумления адресов;

— использует теги состояния в ячейках памяти и теги доступа в адресах, что позволяет выполнять мелкозернистую синхронизацию непосредственно на ячейках памяти без участия процессора;

— включает набор атомарных операций для односторонних взаимодействий параллельных процессов, векторных операций чтения-записи, операций синхронизации обращений к памяти.

Практическая ценность работы

Разработанная архитектура глобально адресуемой памяти является одним из основных архитектурных решений, принятых при разработке мультитредово-потокового суперкомпьютера "Ангара".

Программная параллельная имитационная потактовая модель мультитредово-потокового суперкомпьютера используется в ОАО "НИЦЭВТ" для отработки принципов работы суперкомпьютера "Ангара" и вариантов их аппаратной реализации, создания и отработки системного программного обеспечения, построения моделей прикладных задач.

Практическую ценность работы подтверждает акт №83/6-4328 от 26.10.2010 о внедрении архитектуры глобально адресуемой памяти мультитредово-потокового суперкомпьютера в ОАО "НИЦЭВТ".

Полученные результаты исследования возможностей реализации средств продуктивного параллельного программирования используются при реализации интерфейсов и языков PGAS-класса для суперкомпьютера "Ангара" и суперкомпьютеров поколения СКИФ-4.

Апробация работы и публикации

Основные положения работы докладывались на второй международной конференции “Параллельные вычислительные технологии” (ПАВТ-2008), на XIV международной конференции студентов, аспирантов и молодых ученых Ломоносов-2007, на научных семинарах в НИВЦ МГУ под руководством член-корр. РАН д.ф.-м. н. Вл. В. Воеводина, в ОАО “НИЦЭВТ” под руководством к. ф.-м. н. Л. К. Эйсымонта, а также в ИПМ РАН (направление – “Программирование”).

По результатам работы имеется восемь публикаций [1, 2, 3, 4, 5, 6, 7, 8], в том числе три статьи в журналах из списка ВАК [1, 6, 7] и патент РФ [5].

Объем и структура работы

Диссертация состоит из введения, четырех глав, заключения, списка литературы (92 наименования). Общий объем работы составляет 224 страницы, работа содержит 89 рисунков и 23 таблицы.

Основное содержание работы

Во введении обосновывается актуальность темы исследования, ставится цель и перечисляются решаемые в диссертации задачи. Приводятся основные результаты работы, их научная новизна и практическая значимость.

В первой главе формулируются требования к архитектуре глобально адресуемой памяти перспективных суперкомпьютеров и анализируются основные решения по организации такой памяти в известных суперкомпьютерах.

В первом разделе формулируются требования к созданию подсистемы оперативной памяти перспективных СКСН, выполнение которых необходимо для комплексного решения проблемы “стены памяти”.

Первое требование – большой (петабайтового уровня) объем оперативной памяти, доступной через логически единое (глобальное) адресное пространство. Следствие – такая память может быть составлена только из физических памятей вычислительных узлов мультипроцессорной СКСН.

Второе требование касается схемы организации отображения логически единого адресного пространства на распределенную по узлам физическую память. Эта схема должна обеспечивать защиту одной задачи от другой, допускать экономное использование физической памяти, должна обеспечивать возможность эффективной реализации параллельных алгоритмов.

Третье требование – обеспечение эффективности доступа к памяти, обладающей очень большими задержками выполнения операций с ней, причем в разных режимах пространственно-временной локализации обращений. С точки зрения блока трансляции адресов глобально адресуемой

памяти требование означает, что сложность трансляции виртуальных адресов и связанные с этим накладные расходы должны быть не существенны.

Четвертое требование – предоставление мелкозернистых средств взаимодействия и синхронизации процессов, работающих над общей памятью. В создаваемых перспективных СКСН количество таких процессов может достигать миллионов, поэтому объектов взаимодействия и синхронизации также должно быть много.

Пятое требование – поддержка низкоуровневых интерфейсов GASNet и ARMCI, используемых в runtime-системах языков программирования UPC, CAF и других, которые используют модель PGAS. Модель PGAS является более перспективной в плане обеспечения высокой продуктивности программирования и возможной эффективности работы программ по сравнению с моделью передачи сообщений MPI. Параллельные процессы в модели PGAS взаимодействуют посредством односторонних коммуникаций через общую для них память, обладающую свойством видимого пользователем неоднородного по эффективности доступа к разным ее областям.

Главная цель представленного далее обзора разных вариантов архитектур глобально адресуемой памяти состоит в отслеживании динамики изменений при переходе от одного поколения систем к другому. Знание этой динамики позволяет понять тенденции изменения архитектуры глобально адресуемой памяти в соответствии с опытом решения прикладных задач и возможностями реализации.

Во втором разделе первой главы рассматривается семейство суперкомпьютеров с мультитредовой архитектурой процессора Cray MTA-2 и Cray XMT (Eldorado). Мультитредовая архитектура позволяет выдавать большое количество независимых и одновременных обращений к памяти. Виртуальная память Cray XMT является сегментной, с возможностью распределения сегмента по всем узлам, присутствует скремблирование. Скремблирование (зашумление) любой последовательности адресов, следующих с постоянным шагом, позволяет перевести ее в последовательность адресов с равномерным случайным распределением. Это позволяет снизить конфликты обращений к вычислительным узлам и банкам памяти. Для эффективного взаимодействия параллельных процессов каждая 64-разрядная ячейка памяти имеет дополнительные теговые биты состояния.

Преимуществами организации глобально адресуемой общей памяти Cray XMT является простота, возможность организации мелкозернистого взаимодействия параллельных процессов при помощи теговых битов, а также возможность работы с сегментами большого объема без TLB-промахов. К недостаткам относятся: сегменты можно распределять только

по всем узлам системы сразу, что препятствует управлению локализацией размещения данных; сегментная организация памяти с двумя уровнями адресации – виртуальными адресами задач и физическими адресами влечет необходимость использования сложных подсистем централизованного управления памятью для задач, что может влиять на производительность исполняемых программ.

В третьем и четвертом разделах первой главы рассматриваются суперкомпьютеры Cray X1 и Cray BlackWidow (BW) с векторной архитектурой процессора. Главной особенностью виртуальной памяти Cray X1 и Cray BW является двухуровневая организация с виртуальным адресом, состоящим из номера виртуального узла, номера страницы на этом виртуальном узле и смещения в странице. Эта особенность значительно упрощает реализацию глобально адресуемой памяти, что позволяет транслировать виртуальные адреса с высокой пропускной способностью. Недостаток Cray X1 и Cray BW с точки зрения поддержки PGAS-языков – слабые возможности синхронизации команд работы с удаленной памятью. Главным недостатком является общее свойство векторных суперкомпьютеров – зависимость от векторизуемости приложений. В Cray BW по сравнению с Cray X1 ввели режим, эффективно поддерживающий хранение данных задачи на узлах с несмежными номерами, однако отказались от работы нескольких задач на одном узле.

В пятом и шестом разделах рассмотрены примеры организации глобально адресуемой памяти посредством подключения к коммерческим процессорам специальных схем разной сложности. Такие решения требуют меньше затрат для реализации поддержки глобально адресуемой памяти по сравнению с проектированием процессора с нуля.

В пятом разделе первой главы рассматривается суперкомпьютер Cray T3E. К его преимуществам относится децентрализованная организация управления памятью, так как сегментно-страничная организация глобально адресуемой памяти Cray T3E позволяет разделить вопросы защиты, распределения данных по узлам и размещения страниц на узлах. Универсальный механизм распределения данных обеспечивает аппаратную поддержку для реализации пользовательских способов распределения данных. Большое количество E-регистров, при помощи которых происходит работа с удаленными узлами, позволяет обеспечить максимальную пропускную способность и толерантность к задержкам обращений в глобально адресуемую память. Низкие накладные расходы на синхронизацию процессов обеспечиваются в Cray T3E аппаратной поддержкой работы с очередями сообщений MQCW и барьерных операций. Недостатком Cray T3E является отсутствие возможности запускать задачу на узлах с не подряд идущими номерами.

В шестом разделе первой главы в качестве примера программно-аппаратной поддержки глобально адресуемой памяти рассматривается вариант реанимирования подхода Cray T3E и предлагается современный дешевый способ построения суперкомпьютера с глобально адресуемой общей памятью, суть которого состоит в соединении коммерческого процессора и маршрутизатора коммуникационной сети через Network Interface Controller (NIC), в котором реализуется минимальное управление обращениями в глобальную память. Отличие данного решения от Cray T3E заключается в упрощении процедуры трансляции адреса. По мнению авторов решения, вычислительные возможности современных процессоров превосходят возможности коммуникационных сетей, поэтому определением номера узла и подсчетом смещения внутри этого узла может заниматься процессор, выполняя, таким образом, алгоритм распределения. Недостаток данного конкретного решения заключается в отсутствии поддержки мелкозернистой синхронизации, недостатки подхода – в отсутствии комплексности решения, зависимости от существующей коммерческой аппаратуры.

Анализ показывает, что глобально адресуемая память ни одного из рассмотренных суперкомпьютеров полностью не удовлетворяет всем требованиям. При разработке архитектуры такой памяти в СКСН “Ангара” предпринята попытка выполнить все требования при помощи выработки комплексного решения, с учетом недостатков и преимуществ, выявленных при анализе.

Во второй главе описывается разработанная архитектура глобально адресуемой памяти, а также приводится общее описание СКСН “Ангара”.

Одной из главных задач разработки суперкомпьютера “Ангара” являлось повышение реальной производительности на задачах с плохой пространственно-временной локализацией. В суперкомпьютере “Ангара” выбран подход мультитредовости, при котором сокрытие задержек обращения в память (обеспечение толерантности, нечувствительности процессора к задержкам) происходит за счет построения вычислительного процесса, процессора, коммуникационной сети и модулей памяти таким образом, чтобы одновременно выполнялось множество обращений к ней с очень высоким темпом.

Суперкомпьютер “Ангара” содержит вычислительные узлы, соединенные коммуникационной сетью, оптимизированной для достижения большой суммарной пропускной способности при передаче коротких пакетов. В качестве вариантов реализации такой сети рассматриваются 4D- и 5D-торы с адаптивной бездедлоковой передачей пакетов. Вычислительный узел (далее просто узел) суперкомпьютера “Ангара” содержит: многоядерный мультитредово-потоковый микропроцессор (вариант J7 или J10) с

аппаратной поддержкой трансляции адресов глобально адресуемой памяти и передачи реализующих такие обращения коротких системных пакетов; локальную память с большим расслоением на базе DRAM-модулей; сетевой маршрутизатор.

В одном ядре микропроцессора может одновременно выполняться несколько задач. Каждой задаче ставится в соответствие один домен защиты, причем одна из задач ядра – обязательно операционная система.

Физическая и виртуальная память данных адресуется до байта, однако основным рабочим объектом памяти является 64-разрядное слово (ячейка), базовый адресуемый элемент памяти, выровненный по 8-байтовой границе.

Каждой 64-разрядной ячейке памяти поставлены в соответствие два внешних теговых бита состояния ячейки, один – это full/empty – бит (fe-бит), а другой extag-бит – который имеет единичное значение, если в младших трех разрядах соответствующей тегу ячейки в качестве тегового бита fwd, trap0 или trap1 установлен хотя бы один разряд. Выполнение операций с памятью зависит от значения теговых битов состояния ячейки и от поля управления доступом из адреса, на этом построена низкоуровневая синхронизация при работе с памятью.

Например, рассмотрим доступ к адресуемой ячейке памяти в одном из режимов синхронизации Synchronize. Операция чтения в режиме Synchronize выполняется только для состояния full адресуемой ячейки памяти. Если к этой ячейке не подключены “треды-писатели” (см. далее), то микропроцессор сначала выполняет ожидание, то есть производится повторное выполнение соответствующей команды до тех пор, пока либо не выполнится условие выполнения команды, либо количество повторов этой команды не превысит значение, установленное в специальном регистре. Повтор выполнения команды производится автоматически функциональным устройством LSU для работы с командами обращений в память без выдачи тредом дополнительных команд. Если количество повторов команды оказалось больше установленного значения в специальном регистре, то возбуждается исключительная ситуация. Обработка этой ситуации состоит в установке в адресуемой ячейке trap0-бита и организации привязанного к ней списка структур-ссылок на “треды-читатели” этой ячейки. Факт наличия этого списка распознается далее уже по выставленному trap0-биту. Активация тредов этого списка произойдет только после записи значения в эту ячейку. Операция записи в режиме Synchronize выполняется только для состояния empty адресуемой ячейки. Если же состояние этой ячейки full, то также производится сначала ожидание, которое может закончиться записью или установкой trap0-бита и подключением к ячейке списка тредов, но это уже будут “треды-писатели”. В конечном итоге происходит успешная запись в ячейку и full/empty-бит ее состояния устанавливается равными full.

Физическое адресное пространство узла составляет 256 Гбайт. В системе допускается использовать до 32768 узлов, соответственно, максимальный объем адресуемой памяти – 8 Пбайт. Физический адрес состоит из номера узла, смещения в локальной памяти узла и поля управления доступом, в общей сложности 64 разряда.

Задача оперирует виртуальными адресами (VA) данных. VA содержит номер сегмента задачи, смещение в сегменте, а также служебную информацию. При помощи виртуальных V-сегментов и их свойств реализована защита данных задачи и их распределение по физической памяти узлов. Свойства V-сегмента описываются его дескриптором, который хранится в специальной таблице дескрипторов V-сегментов в памяти каждого узла, занятого задачей. Для ускорения доступа к таблице дескрипторов V-сегментов используется кэш дескрипторов (VTLB). Имеются два типа V-сегментов – обычные (размер от 128 байт до 256 Гбайт) и суперсегменты (размер от 128 Кбайт до 256 Тбайт). Суперсегменты введены для работы с данными очень большого объема как с единым целым, а также чтобы уменьшить промахи при обращении к VTLB.

Способ отображения V-сегментов на физическую память задается дескриптором V-сегмента. V-сегмент отображается на 2^P узлов со смежными номерами, начиная с узла с номером First Node Number (P и First Node Number заданы в дескрипторе).

Через аппарат виртуальных номеров узлов возможно отображение на узлы не обязательно со смежными номерами, что можно задать специальным битом дескриптора V-сегмента.

Распределение (отображение) на физические адреса может быть блочно-циклическим и блочным, в зависимости от поля Distribution Type дескриптора V-сегмента. В случае блочного распределения размер блока z, отведенного в узле для данного сегмента, вычисляется как z = размер сегмента / 2^P – в каждом узле только один блок сегмента. При блочно-циклическом распределении сегмент циклически распределяется по узлам блоками размером B = $4^{\mathrm{DistributionType}+3}$ (от 64 байт до 256 Кбайт). Причем сначала между N=2^P узлами распределяются первые N блоков, затем – вторые N и т.д., в одном узле может быть несколько блоков сегмента.

Смещение VA перед распределением может подвергаться дополнительной процедуре – скремблированию узлов, на что указывает бит ScrambleNodes дескриптора. Скремблирование узлов позволяет следующие с регулярным шагом обращения к памяти размешать случайным равномерным образом по узлам, на которые распределен сегмент. После распределения получившееся смещение сегмента в узле в случае установленного бита ScrambleBanks дескриптора сегмента подвергается дополнительному скремблированию, необходимому для снижения вероятности конфликтов по

доступу к банкам памяти.

Независимо от типа распределения размещение области для данных сегментов в физической памяти узлов возможно двумя способами – централизованно и децентрализованно.

При централизованном способе в каждом узле область данных сегмента начинается с физического адреса из поля Segment Base дескриптора V-сегмента. В этом случае VA транслируется сразу в физический адрес. Так как задача централизованного планирования физической памяти в узлах сложна, то введен децентрализованный способ.

При децентрализованном способе размещения область памяти под сегмент выделяется не в физическом, а в глобальном виртуальном адресном пространстве. Это пространство состоит из виртуальных R-сегментов, их нумерация общая для всех узлов, а в каждом узле хранится фрагмент такого R-сегмента. VA транслируется при этом в глобальный виртуальный адрес (GVA). В дескрипторе V-сегмента хранится номер R-сегмента, который вместе с полученным при трансляции виртуального адреса номером узла и смещением во фрагменте R-сегмента на этом узле образуют GVA, который окончательно транслируется на целевом узле.

Схема процесса трансляции VA в физический адрес или GVA изображена на рис. 1.

В отличие от V-сегментов, номера R-сегментов являются общими для всей системы. В каждом узле фрагмент R-сегмента образует локальную виртуальную память, состоящую из страниц размером по 16 Кбайт, 1 Мбайт, 64 Мбайт или 4 Гбайт. Размер страниц указан в дескрипторе V-сегмента и при трансляции переносится в GVA. GVA содержит достаточно информации для формирования физического адреса в целевом узле – для этого в этом узле остается только определить начальный адрес страницы R-сегмента, который хранится в дескрипторе страницы R-сегмента.

Дескрипторы страниц R-сегмента кэшируются в специальной кэш-памяти RPTLB. При промахе в RPTLB необходимо обратиться к таблице дескрипторов страниц R-сегмента в памяти узла. Ее адрес указан в дескрипторе R-сегмента. Хотя каждый R-сегмент имеет уникальный номер в пределах всей системы, но в узле, как правило, используется ограниченное число R-сегментов, поэтому для хранения их дескрипторов используется хешированная таблица дескрипторов R-сегментов RST (Real Segment Table) с разрешением конфликтов методом цепочек. Схема трансляции GVA в физический адрес на целевом узле изображена на рис. 2.

Сравним разработанную архитектуру глобально адресуемой памяти с исследованными ранее вариантами и рассмотрим выполнение требований.

Организацией глобально адресуемой памяти выполняется первое требование: через логически единое адресное пространство возможен доступ

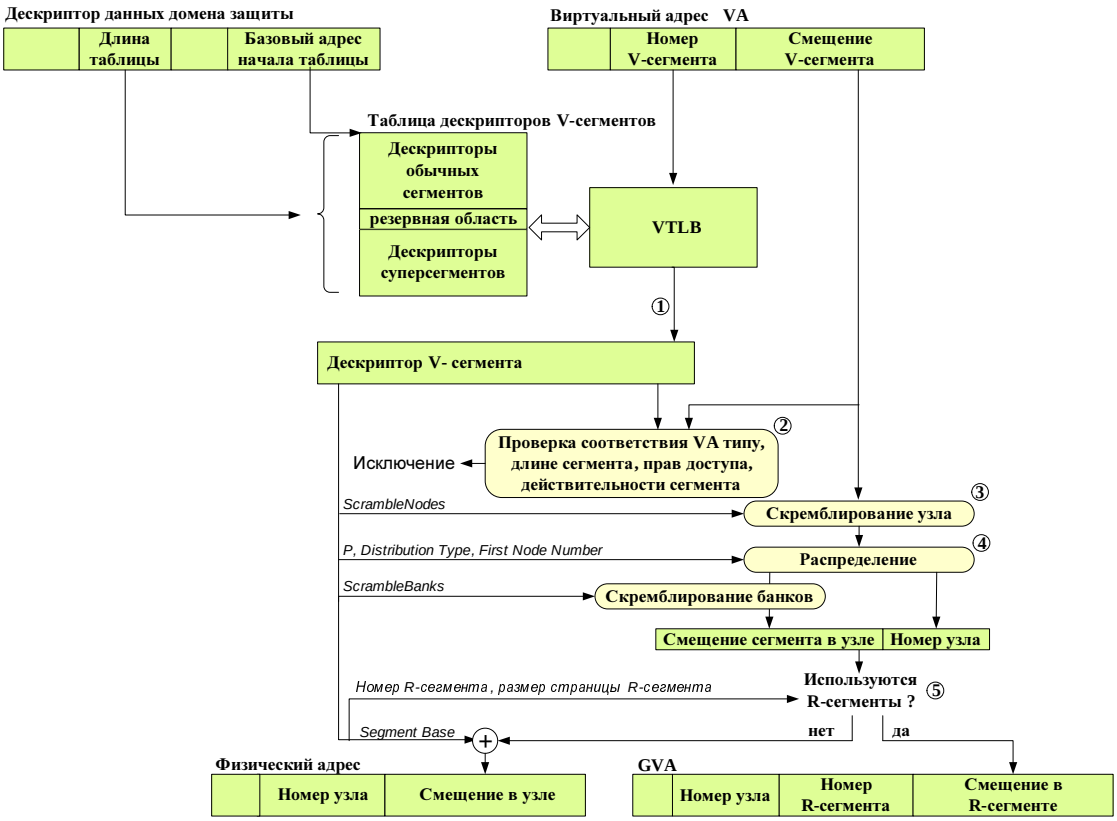

Рис. 1. Схема трансляции виртуального адреса в глобальный виртуальный или физический адреса в суперкомпьютере “Ангара”.

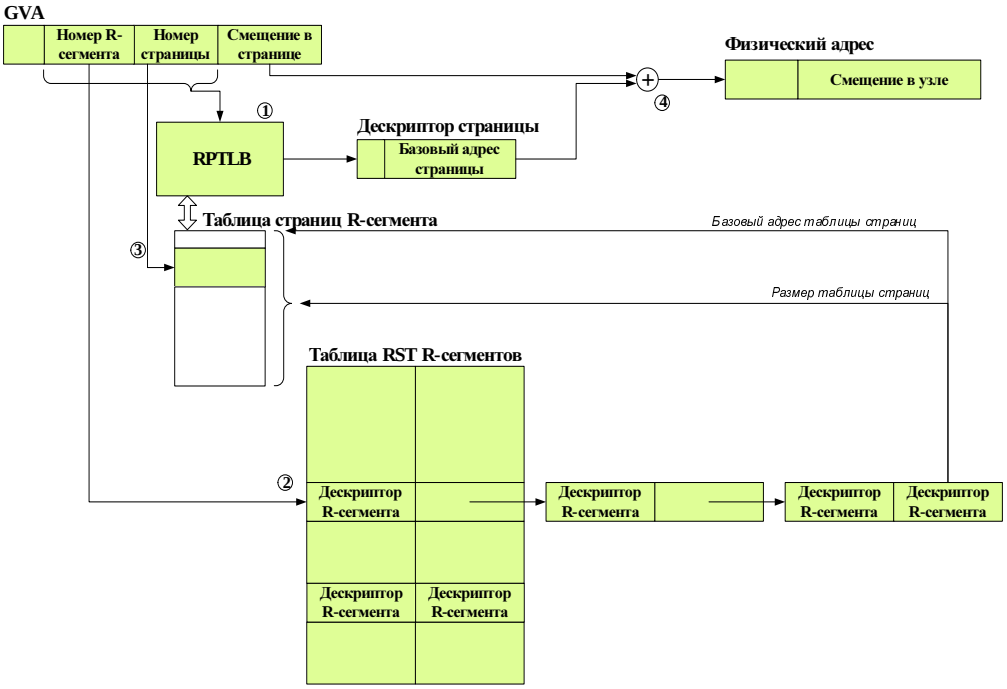

Рис. 2. Схема трансляции глобального виртуального адреса в физический адрес в суперкомпьютере “Ангара”.

к 8 Пбайт распределенной памяти.

V- и R- сегменты разработанной архитектуры глобально адресуемой памяти обеспечивают виртуализацию адресного пространства и высокогранулированную сегментную защиту по сравнению с организацией памяти в Cray BlackWidow. По сравнению с Cray XMT введенный уровень глобальных виртуальных адресов вводит возможность децентрализованного управления локальной памятью каждого из узлов за счет разделения вопросов защиты и отображения данных от вопросов их размещения в локальной памяти узлов, что вместе с введением страниц разного размера выполняет требование об экономном управлении памятью.

В глобально адресуемой памяти "Ангара" по сравнению с Cray XMT усилено управление отображением данных: при отображении на физическую память адресного пространства сегмента можно в дескрипторе этого сегмента задавать последовательность вычислительных узлов, в модулях памяти которых он может располагаться. Также при отображении на памяти выбранного множества вычислительных узлов можно использовать либо блочное, либо блочно-циклическое отображение. Оставлена возможность Cray XMT скремблирования номеров узлов, но появилась отдельная возможность скремблирования банков памяти. По аналогии с Cray XMT у каждой 64-разрядной ячейки памяти существуют дополнительные теговые биты, хорошо зарекомендовавшие себя для осуществления мелкозернистой синхронизации, что вместе с расширенным набором атомарных операций обеспечивает выполнение требования о мелкозернистой синхронизации параллельных процессов.

Для большей гибкости в СКСН "Ангара" по аналогии с Cray BlackWidow введена возможность использования для отображения данных сегмента задачи на не обязательно подряд идущую последовательность номеров узлов.

В третьей главе описывается параллельная программная имитационная потактовая модель СКСН "Ангара", а также разработанные блоки, имитирующие работу глобально адресуемой памяти в этой модели, необходимые для исследования возможностей применения разработанной архитектуры глобально адресуемой памяти при решении прикладных задач. Модель является параллельной программой, написанной на языке Charm++, позволяет моделировать СКСН "Ангара" в полном объеме и масштабируется по производительности до 256 двухпроцессорных (восьмиядерных) узлов суперкомпьютера МВС-100К в МСЦ РАН.

В модели отслеживаются реальные временные диаграммы обращения к памяти и работы с коммуникационной сетью. Схема выполнения команды обращения к глобально адресуемой памяти состоит в следующем. После выдачи каким-либо тредовым устройством узла А команды обращения к памяти, она попадает в функциональный блок LSU подготовки и

отслеживания обращений к памяти. Из LSU виртуальный адрес обращения к памяти передается в блок MMU для трансляции. Если адресуемая память – это локальная память узла A, то запрос на выполнение передается в контроллер памяти EDI и далее во внекристальную DRAM-память. Если адресуемая память принадлежит другому узлу, то запрос передается сначала в блок MSU, а потом в NI, в котором аппаратно формируется короткий пакет с командой обращения к памяти, который через коммуникационную сеть передается в целевой узел B. В узле B команда обращения к памяти поступает через блоки NI и MSU в блок MMU, в котором, если необходимо, происходит трансляция глобального виртуального адреса в физический, после чего выполняется обращение к EDI и DRAM-памяти узла B. Результат обращения возвращается через блоки MSU и NI узла B и далее через коммуникационную сеть в узел A в обратном порядке.

Таблица 1. Задержки различных обращений к памяти в тактах процессора с учетом накладных расходов на трансляцию виртуальных адресов в тактах микропроцессоров J7 (500 МГц) и J10 (2 ГГц).

Вид адресации и вариант трансляции	J7	J10
Обращение по физическому адресу в кэш данных	21	21
Обращение по физическому адресу в DRAM-память (промах в кэш данных)	64	97
Обращение по виртуальному адресу в кэш данных	34	34
Обращение по виртуальному адресу в DRAM-память	77	110
Обращение по виртуальному адресу в DRAM-память с промахом в VTLB и RPTLB	271	370
Обращение по виртуальному адресу в DRAM-память соседнего узла, расстояние 1	426	593
Обращение по виртуальному адресу в DRAM-память узла на расстоянии 7	832	1789

Блок трансляции адресов, внутрикристальная сеть, кэш данных, внекристальная память, маршрутизатор сети и сетевые линки имеют реальные характеристики по задержкам и пропускной способности, их влияние отражено в табл. 1. В этой таблице приведены полученные на модели для вариантов микропроцессора J7 (500 МГц) и J10 (2 ГГц) и внекристальной DRAM памяти DDR2 (800 МГц) задержки обращений к памяти своего и удаленного узла при разных вариантах адресации и выполнении процессов

трансляции.

В четвертой главе приводятся результаты проведенных исследований возможности достижения на суперкомпьютере с предложенной архитектурой глобально адресуемой памяти высокой реальной производительности, а также возможности реализации языков параллельного программирования PGAS-класса, что должно обеспечить высокую продуктивность параллельного программирования.

Исследования развиваемой реальной производительности были проведены на общепринятых тестах из пакета HPCChallenge с граничными значениями пространственно-временной локализации обращений к памяти: dgemm, HPL – умножение матриц и тест Linpack (лучшие значения временной и пространственной локализации); FFT – одномерное преобразование Фурье (хорошая временная, плохая пространственная локализация); STREAM – пересылка векторов (плохая временная, хорошая пространственная локализация); RandomAccess – тест нерегулярных обращений к памяти (худшие значения временной и пространственной локализации). В исследованиях также использовались тесты APEX-MAP и PUT/GET.

Первый этап исследований связан с оценкой влияния накладных расходов при трансляции виртуальных адресов на производительность программ. Требование об эффективности доступа к памяти означает, чтобы производительность задач с разной пространственно-временной локализацией при работе с памятью по виртуальным адресам и по физическим адресам отличалась незначительно.

Проведенные исследования показывают, что несмотря на задержки, вносимые блоком MMU при трансляции виртуального адреса, производительность однопроцессорных программ dgemm, STREAM Triad, FFT и RandomAccess, работающих по виртуальным адресам на небольшом числе тредов незначительно отличается от производительности программ, работающих по физическим адресам. При большом числе тредов выдается большое количество одновременных обращений в память, режимы работы по виртуальным и физическим адресам эквивалентны из-за толерантности мультитредового процессора к задержкам, причем это свойство верно даже при небольшом количестве каналов (конвейеров) трансляции адреса (1 для J7 и 2 для J10).

При исследовании процедуры трансляции адреса особое внимание уделялось случаям промахов в кэши дескрипторов сегментов и страниц, соответственно VTLB и RPTLB. Эти промахи могут серьезно повлиять на производительность, что и происходит в современных коммерческих микропроцессорах. При выполнении тех же четырех программ из пакета HPCChallenge для одного узла было установлено, что использование большого количества тредов даже в случае большого количества промахов

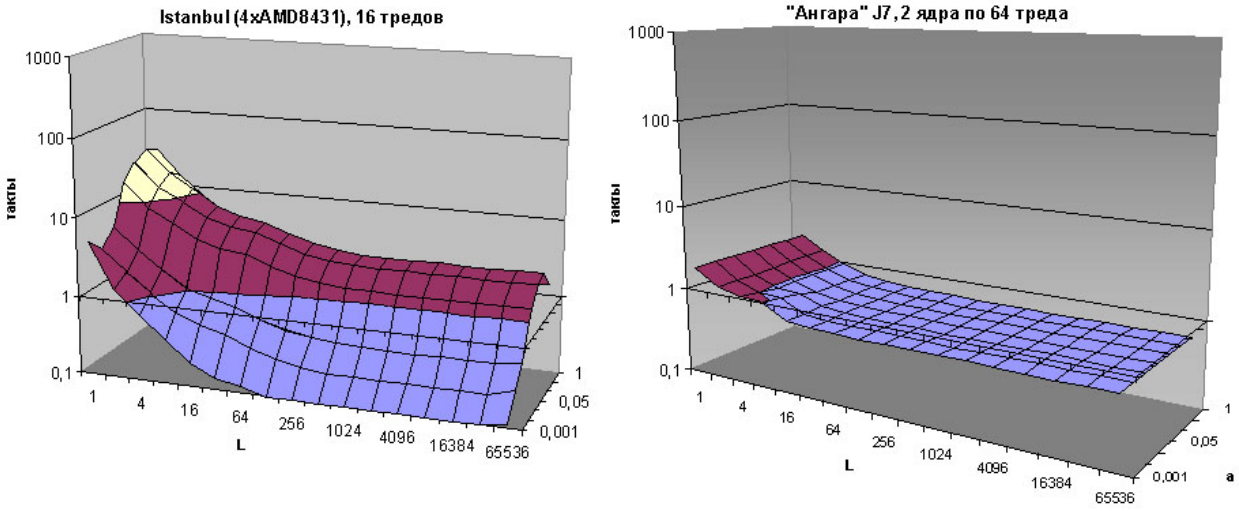

Рис. 3. APEX-поверхности платы с 4 процессорами AMD Istanbul и процессора “Ангара” J7.

в кэши дескрипторов помогает достичь или приблизиться к уровню производительности, достигаемому когда промахов нет. Введение кэшей дескрипторов оправдано в силу оптимизации количества используемых в программе тредов и ресурсов физической подсистемы памяти. Также для сокращения промахов в кэши дескрипторов имеется еще два средства – суперсегменты и большие страницы.

Исследования возможной эффективности подсистемы памяти также были проведены с использованием синтетического теста APEX-MAP. Тест APEX-MAP строит зависимость характеристики доступа к памяти в зависимости от параметров пространственной (L) и временной (a) локализации. Полученные на этом тесте интегральные оценки в виде APEX-поверхностей показывают резкое улучшение показателей эффективности доступа к памяти в СКСН “Ангара” для режимов плохой пространственно-временной локализации. APEX-поверхность для одного узла имеет вид практически горизонтальной плоскости, а не “горки”, характерной для коммерческих процессоров, см. рис. 3.

Таким образом, при работе на одном узле накладные расходы на организацию глобально адресуемой памяти не препятствуют достижению высокой производительности, что удовлетворяет требованию об эффективности доступа к памяти.

Второй этап исследований был связан с рассмотрением работы множества узлов.

APEX-поверхности для многоузлового варианта теста APEX-MAP представлены на рис. 4. Также, как и в одноузловом случае, видно резкое улучшение характеристик доступа к памяти в СКСН “Ангара” для режимов

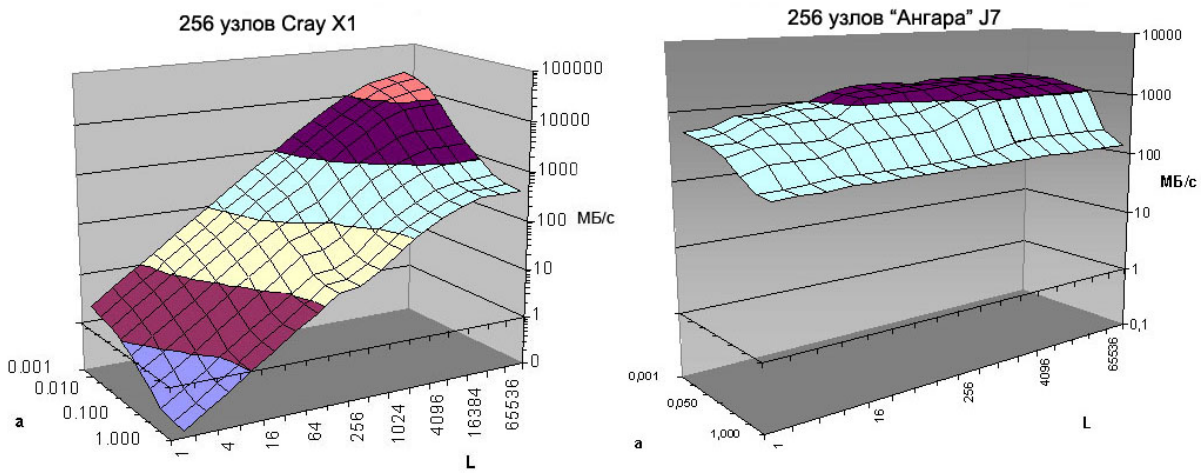

Рис. 4. APEX-поверхности 256 процессоров Cray X1 и "Ангара" J7.

плохой пространственно-временной локализации, причем даже в сравнении с суперкомпьютером с векторной архитектурой Cray X1. Таким образом, тест APEX-MAP демонстрирует достижение основной цели разработки СКСН "Ангара" – эффективной работы с глобально адресуемой памятью большого объема с приблизительно равными показателями в разных режимах пространственно-временной локализации.

В табл. 2 представлены результаты выполнения выбранных тестов из пакета HPCChallenge и задачи BFS поиска вширь в графе, на самых мощных на октябрь 2010 года суперкомпьютерах и СКСН "Ангара". Тесты для СКСН "Ангара" написаны на языке ассемблера этой системы, а взаимодействие параллельных процессов производится командами считывания и записи из оперативной памяти. Тесты для остальных суперкомпьютеров написаны на языке C с использованием библиотеки MPI.

На примере задач умножения плотнозаполненных матриц и HPL, обладающих близкой и хорошей пространственно-временной локализацией данных, что допускает сравнение результатов на этих тестах, показано, что реальная производительность СКСН "Ангара" находится на уровне существующих суперкомпьютеров.

На DIS-задаче FFT реальная производительность (по отношению к пиковой) СКСН "Ангара" лучше примерно от 3 (J10) до 19 (J7) раз. Абсолютная производительность только для J10 превышает производительность Cray XT5. Для другой DIS-задачи BFS доступны результаты только для суперкомпьютера IBM BlueGene/L, и если приводить производительность к одному узлу, то СКСН "Ангара" лучше примерно от 14 (J7) до 63 (J10) раз. Таким образом, разработанная глобально адресуемая память позволяет развивать реальную производительность на DIS-задачах до 1-2 порядков большую, чем на существующих суперкомпьютерах.

Таблица 2. Сравнение суперкомпьютеров на тестах с разной локализацией обращений к памяти. ПС – пропускная способность, * – умножение матриц, @ – используется 32768 узлов.

Система	IBM BG/P	IBM BG/L	Cray XT5	Ангара J7	Ангара J10
Тактовая частота микропроцессора, ГГц	0.85	0.7	2.6	0.5	2
Количество ядер (тредов в ядре)	4	1	6	2 (64)	8 (128)
Пиковая производительность микропроцессора, Гфлопс	13.6	5.6	62.4	4	128
ПС DRAM-памяти, Гбайт/с	5.6	13.6	12.8	25.6	204.8
Дуплексная ПС линка сети, Гбайт/с / тип сети (тор)	0.425 / 3D	0.175 / 3D	4.8 / 3D	4 / 4D	12 / 4D
Дуплексная ПС интерфейса микропроцессора с сетью, Гбайт/с	5.1	2.1	3.2	8	48
Число используемых процессоров	32768	65536	32768	8192	4096
HPL, Тфлопс	450.3 (81%)	259.2 (71%)	1533.7 (66%)	24.8* (83%)	380.9* (73%)
FFT, Гфлопс	5080 (0.9%)	2311 (0.6%)	10699 (0.5%)	6140 (19%)	15131 (3%)
RandomAccess, GUPS	117.13	35.47	37.69	150.7	602.9
BFS, Миллионов дуг в секунду	–	7655@	–	26675	63251

Тесты умножения матриц, FFT и BFS для СКСН “Ангара” имеют гетерогенную мультитредовую вычислительную модель. Использование распределенных в глобально адресуемой памяти данных сочетается в ней с локализацией данных в памяти узлов. Если изначальной (статической) локализации данных не хватает, то она обеспечивается подкачками, что выполняется специально выделенными в программе тредами. Тест BFS использует также локализацию вычислений при данных. Она осуществляется посредством команд запуска тредов на удаленных узлах и позволяет добиться рекордных результатов.

Особую роль при достижении результатов играла возможность блочного распределение сегментов, которое применялось, когда был необходим учет локализации распределенных данных и более удобно, когда данные

лежат большими последовательными порциями на разных узлах. Блочно-циклическое распределение использовалось совместно со скремблированием узлов. Возможность скремблирования позволяет организовать равномерно-случайный траффик в сети, который обеспечивает ее равномерную загрузку, что выгодно в случае, когда трудно организовать локализацию данных в узлах. Скремблирование банков памяти дает очень большой эффект при работе с памятью узлов: без скремблирования мультитредовый процессор простаивал бы из-за конфликтов по банкам памяти.

Высокая производительность для СКСН "Ангара" получена на программах с короткими односторонними коммуникационными обменами, обращения к памяти чужого узла происходят в любом месте программы, отсутствует программная обработка приема этих обращений. Такой стиль программирования значительно проще (продуктивнее), чем использование библиотеки MPI, для использования которой необходимо собирать большие коммуникационные пакеты для эффективной передачи, а также программировать прием сообщений.

Таким образом, сравнение с современным уровнем производительности, достигнутым на существующих суперкомпьютерах с использованием библиотеки MPI, показало, что применение мультитредовых моделей с использованием разработанной архитектуры глобально адресуемой памяти позволило повысить уровень реальной производительности для задач DIS-класса до 1-2 порядков и сохранить уровень производительности на остальных задачах. Причем высокая реальная производительность была получена с использованием коротких односторонних коммуникационных передач, при помощи которых проще создавать параллельные программы по сравнению с традиционным программированием с использованием MPI. Возможность получения высокой производительности на прикладных задачах при помощи разработанной архитектуры выполняет соответствующее требование к архитектуре глобально адресуемой памяти суперкомпьютеров.

Рассмотрим требование о поддержке разработанной глобально адресуемой памятью реализаций низкоуровневых интерфейсов GASNet и ARMCI, а также языка PGAS-класса UPC. Анализ набора функций GASNet и ARMCI и их семантики показал, что низкоуровневые интерфейсы удобно реализуются с использованием большого количества тредов и глобальной адресации. Основу оценки составил стандартный PUT/GET-тест операций записи/чтения с памятью удаленного узла. Полученная для СКСН реальная пропускная способность особенно для коротких пакетов не достижима в системах на базе коммерчески доступных компонентов.

Реализация языка UPC для СКСН "Ангара" обладает преимуществами по сравнению с реализацией UPC, например, на Cray X1. Во-первых,

существует возможность эффективной трансляции указателей UPC на глобальную память в виртуальный адрес разработанной глобально адресуемой памяти. Во-вторых, для достижения высокой производительности на СКСН “Ангара” требуется распараллеливание программы на треды, что гибче и проще, чем векторизация программ, необходимая для достижения высокой производительности на Cray X1. В третьих, СКСН “Ангара” позволяет эффективно использовать короткие коммуникационные передачи, что позволит достигать высокой производительности при использовании удобных поэлементных обращений к памяти удаленных узлов. Глобально адресуемая память СКСН “Ангара” обладает всеми функциональными свойствами организации памяти Cray X1, что с учетом известного положительного опыта реализации UPC на Cray X1 позволяет говорить о том, что разработанная глобально адресуемая память поддерживает реализацию языка UPC на СКСН “Ангара”.

Таким образом, все сформулированные в работе требования к архитектуре глобально адресуемой памяти суперкомпьютеров можно считать выполненными в разработанной архитектуре глобально адресуемой памяти мультитредово-потокового суперкомпьютера “Ангара”.

В заключении диссертационной работы перечисляются ее основные результаты.

Основные результаты работы

1. Разработана архитектура глобально адресуемой памяти мультитредово-потокового суперкомпьютера.

2. Разработаны и исследованы блоки, имитирующие работу глобально адресуемой памяти в составе параллельной программной потактовой имитационной модели суперкомпьютера «Ангара».

3. Показано, что применение разработанной глобально адресуемой памяти в сочетании с другими возможностями мультитредово-потокового суперкомпьютера “Ангара” в сравнении с опубликованными данными по существующим суперкомпьютерам позволяет:

— повысить реальную производительность на задачах с интенсивным нерегулярным доступом к памяти до одного-двух порядков при совпадении реальной производительности на остальных задачах;

— использовать при получении высокой реальной производительности короткие односторонние коммуникационные передачи, при помощи которых проще создавать параллельные программы по сравнению с традиционным программированием с использованием MPI.

4. Продемонстрирована возможность использования разработанной архитектуры глобально адресуемой памяти при реализации перспективных языков параллельного программирования PGAS-класса на примере языка UPC.

Работы автора по теме диссертации

[1] *Фролов А.С., Семенов А.С., Корж А.А., Эйсымонт Л.К.* Программа создания перспективных суперкомпьютеров // Открытые системы. – 2007. – №9. – С. 21–29.

[2] *Семенов А.С.* Распределенная общая память суперкомпьютера, предназначенного для решения задач большой размерности // Ломоносов - 2007: XIV Международная конференция студентов, аспирантов и молодых ученых; секция "Вычислительная математика и кибернетика": Тезисы докладов. – М.: – Издательский отдел факультета ВМиК МГУ, МАКС Пресс, 2007. – С. 73.

[3] *Семенов А.С., Эйсымонт Л.К.* Параллельное умножение матриц на суперкомпьютере с мультитредово-потоковой архитектурой // Программные системы и инструменты. – М.: – Издательство факультета ВМиК МГУ, 2007. – №8. – С. 106–117.

[4] *Семенов А.С.* Одномерное быстрое преобразование Фурье на суперкомпьютере с мультитредово-потоковой архитектурой // Параллельные вычислительные технологии (ПаВТ'2008): Труды международной научной конференции (Санкт-Петербург, 28 января - 1 февраля 2008 г.). – Челябинск: – Издательство ЮУрГУ, 2008. – С. 224–231.

[5] Патент РФ №2396592. Способ организации глобально-адресуемой общей памяти в многопроцессорной ЭВМ. *Семенов А.С., Слуцкин А.И., Соколов А.А., Эйсымонт Л.К.* 2008.

[6] *Аверичева Д.Л., Семенов А.С., Фролов А.С.* Поиск вширь в графе на суперкомпьютере с мультитредово-потоковой архитектурой // Информационные технологии. – М.: – Издательство "Новые технологии", 2009. – №7. – С. 7–12.

[7] *Семенов А.С., Соколов А.А., Эйсымонт Л.К.* Архитектура глобально адресуемой памяти мультитредово-потокового суперкомпьютера // ЭЛЕКТРОНИКА: Наука, Технология, Бизнес. – 2009. – №1. – С. 50–61.

[8] Моделирование российского суперкомпьютера "Ангара" на суперкомпьютере / *Эйсымонт Л.К., Семенов А.С., Соколов А.А. [и др.]* // В сборнике "Суперкомпьютерные технологии в науке, образовании и промышленности" под редакцией: академика В.А. Садовничего, академика Г.И. Савина, чл.-корр. РАН Вл.В. Воеводина. – М.: Издательство Московского университета, 2009. – С. 145–150.

Тираж 100 экз.
Отпечатано в Московском авиационном институте
(государственном техническом университете)
г. Москва, А-80, ГСП-3, Волоколамское шоссе, д. 4.
http://www.mai.ru/