Создание вычислительной системы для моделирования суперкомпьютера с производительностью экзафлопсного уровня

Development of Computer System for simulation of supercomputer with exaflops level of throughput

Елизаров С.Г., Лукьянченко Г.А., Корнеев В.В.

Sergey Elizarov, Georgy Lukyanchenko, Victor Korneev

Аннотация. Приведено обоснование необходимости эмуляции архитектур суперкомпьютеров экзафлопсного уровня производительности на кластерах из узлов, включающих сопроцессоры на ПЛИС.

Annotation. The substantiation of necessity of emulation architectures of supercomputers with exaflops level of throughput on clusters with nodes, including coprocessors on FPGA.

Ключевые слова: эмуляции архитектур суперкомпьютеров экзафлопсного уровня производительности, ПЛИС, модель параллельного программирования

Key words: emulation architectures of supercomputers with exaflops level of throughput, FPGA, model of parallel programming

1. Мотивация

Создание суперкомпьютеров с производительностью на уровне 1 экзафлопса (10^{18} операций с плавающей точкой в секунду) к 2018 – 2020 годам является стратегической научно-технической задачей для США, Китая, Японии, стран ЕС, России, Индии.

При этом производительность и отказоустойчивость такого суперкомпьютера будет определяться не только техническими характеристиками созданной к тому времени элементной базы (отдельных микропроцессоров, микросхем памяти и др.), но и пониманием общих принципов взаимодействия большого количества отдельных вычислительных устройств, составляющих такой суперкомпьютер. Эти принципы реализуются в виде системного и прикладного программного обеспечения, разрабатывать которое желательно уже сейчас, до того как появится элементная база 2018 года.

Исходя из современных представлений о развитии элементной базы СБИС и архитектур, суперкомпьютер экзафлопсного уровня производительности должен обрабатывать большое количество потоков (порядка 10^9 при производительности одного потока порядка 10^9 flops) над общей памятью (distributed shared memory - DSM) или секционированной общей памятью (partitioned global address space - PGAS). Могут использоваться как асинхронные потоки (треды), протекающие в универсальных процессорах и имеющие собственный отдельный счётчик команд у каждого потока, так и синхронные потоки, выполняемые по одному счётчику команд в разных АЛУ, например в GPU NVidia Fermi синхронно в каждом такте может протекать до 512 потоков. Синхронные потоки составляют аппаратно экономную альтернативу асинхронным, но более сложны в программировании.

В конечном итоге, программа для суперкомпьютера экзафлопсного уровня производительности представляет собой описание порождения, межпотоковых коммуникаций и синхронизации потоков, которое компилятор (статически) и библиотеки (динамически в ходе вычислений) преобразуют в совокупность асинхронных и синхронных потоков, определяя ресурсы, на которых будут протекать эти потоки. В этом смысле программа будет исполняться на гетерогенных ресурсах.

Сегодня компьютерная индустрия столкнулась с колоссальным вызовом: освоением технологии создания программ, эффективно загружающих многоядерные процессорные кристаллы и высокопроизводительные суперкомпьютеры на их базе. В связи с недостаточной исследованностью вопросов исполнения программ на гетерогенных ресурсах, организации памяти как всей общедоступной или частично общедоступной, а частично локальной, доступной например, только синхронным потокам с одним счётчиком команд, пропорции синхронных и асинхронных потоков, а также архитектуры суперкомпьютера экзафлопсного уровня производительности, в целом, **необходимо выполнить анализ и моделирование архитектур посредством натурного макетирования**.

2. Цели и задачи

2.1. Цель проекта – Создание ВС для моделирования суперкомпьютеров со сверх высоким уровнем распараллеливания в рамках работ по созданию в РФ вычислительных систем с производительностью на уровне 10^{18} оп/сек.

2.2. Содержание проекта – разработка и создание аппаратных и программных средств **моделирующей гетерогенной ВС (МГВС)**, включающей 100 – 1000 микропроцессоров и до 10000 ПЛИС. Моделирующая гетерогенная ВС предназначена для моделирования работы суперкомпьютера, объединяющего до 10^5 микропроцессоров, до 10^6 – 10^7 процессорных ядер и порождающего в ходе вычислений до 10^8 – 10^9 потоков обработки данных.

Создаваемая моделирующая гетерогенная ВС должна содержать наиболее важные технические и программные решения моделируемой системы. МГВС это суперкомпьютер, «наилучшим образом» воплощающий архитектурные идеи ВС экзафлопсного уровня производительности и позволяющей оценивать их эффективность, в том числе путем написания, отладки и оценки производительности системного и прикладного ПО.

Суперкомпьютер предлагается строить из вычислительных узлов (ВУ), объединяемых сетью FDR 56Gb/s InfiniBand.

ВУ будет состоять из наиболее производительной на момент выбора серийно выпускаемой многосокетной серверной платы с многоядерными процессорными кристаллами, общей разделяемой всеми процессорами одноблочной или многоблочной памяти большого объема (100 Гбайт и более) и несколькими (2-4) каналами PCI Express, используемым для:

- подключения адаптеров FDR 56Gb/s InfiniBand,
- подключения многопортового коммутатора PCI Express,
- подключения ускорителей на ПЛИС.

Встроенные в процессорные кристаллы контроллеры памяти должны поддерживать высокую степень расслоения локального блока памяти, а процессорные ядра допускать большое число незавершённых обращений к памяти. На уровне суперкомпьютера должно программно формироваться глобальное

адресное пространство разделяемой памяти, состоящей из блоков локальной памяти узлов, имеющих достаточно большой объем.

В каждом ВУ к портам коммутатора PCI Express подключаются ускорители на ПЛИС. Ускоритель представляет собой четыре или более ПЛИС, к которым подключены блоки памяти. В каждой ПЛИС реализуются контроллеры блоков памяти, подключенных к ней, набор процессорных ядер и контроллер PCI Express. Функциональность контроллера памяти ПЛИС расширена для обеспечения работы процессорных ядер “по готовности данных” на аппаратном уровне.

Коммутатор обеспечивает взаимодействие по связям PCI всех устройств, то есть обмены между ПЛИС и передачу данных между ПЛИС и процессорами серверной платы. При этом совокупность памятей ПЛИС и универсальных процессоров частично общедоступна, а частично локальна для каждой ПЛИС и используется как буферная память, в которую данные заносятся, обрабатываются и пересылаются в конвейерном режиме с совмещением ступеней по времени, что должно уменьшить потери производительности, обусловленные пересылкой данных.

В ПЛИС может быть реализована:

– совокупность процессоров для асинхронного исполнения потоков,
– процессоры для синхронного исполнения потоков,
– непосредственно создана схема, аппаратно реализующая вычисления потоков.

С архитектурной точки зрения, ВМ представляет собой структуру SMP, а суперкомпьютер в целом – параллельную ВС с мощными узлами.

Модель программирования использует явное задание потоков, синхронизация которых реализуется с помощью дополнительных признаков данных в общей памяти и оригинального контроллера памяти в ПЛИС. Для порождения и завершения асинхронных потоков в программах на традиционных языках программирования, например на языке С, использованы специальные библиотечные функции, позволяющие соответственно породить заданное количество потоков, завершить потоки, а также выполнять атомарные операции. В совокупности с общей памятью, слова которой снабжены дополнительными признаками и соответствующей функциональностью контроллера памяти, это позволяет задать синхронизацию и практически любые межпотоковые коммуникации.

Планируется создать используемую в процессе вычислений библиотеку формирования в ПЛИС ресурсов для исполнения асинхронных или синхронных потоков и распределения потоков по предварительно сформированным в ПЛИС ресурсам.

Ожидается, что на старших моделях ПЛИС семейства Virtex7 будет реализован многоядерный многопотоковый процессор с общей памятью, содержащий порядка тысячи предельно компактных 32-х разрядных RISC-ядер, аппаратная часть которых позволяет порождать, манипулировать и завершать потоки, не обращаясь с системных вызовам ОС. Контроллер памяти будет поддерживать работу процессорных ядер “по готовности данных” на аппаратном уровне и механизмы межпотоковой синхронизации. Микропроцессорные ядра, не будут содержать аппаратные блоки для выполнения операций над числами с плавающей точкой, но будут позволять полноценно реализовывать функции системного ПО, в том числе работу с памятью и другие функций, характерные для многопроцессорных ВС. На указанный процессор будет портирована posix-совместимая операционная система и все стандартные инструменты для включения такого процессора в гетерогенную вычислительную систему.

Таким образом, МГВС, включающая 10000 ПЛИС, сможет достаточно всесторонне моделировать работу многопроцессорной ВС, содержащей несколько миллионов ядер, что соответствует представлению о будущей экзафлопсной ВС.

3. Имеющийся научно-технический задел

К настоящему времени выполнена первая фаза проекта и развернуты работы по второй фазе.

В рамках первой фазы на ПЛИС семейства Virtex-5 (xc5vlx50t) с 256 Мб DDR2 реализован многоядерный процессор (10 ядер), основанный на модифицированной открытой Microblaze-совместимой архитектуре, рабочая частота ядра 50 МГц. Адаптирован для работы с матрицей ядер контроллер внешней памяти с открытым исходным кодом. Написаны и оптимизированы кэши команд. Локальная память реализована с помощью оригинального алгоритма подстановки локальных стеков. Создан загрузчик с внешней flash памяти. Разработана двухуровневая иерархия ядер и модуль запуска потоков на свободных ядрах, в том числе создан коммуникационный протокол для распределения заданий между ядрами без участия ОС. Произведено портирование ОС Minix3 и ключевых сервисов на описываемую выше архитектуру. Адаптирован под указанную архитектуру эмулятор, позволяющий полноценно производить отладку ПО. Разработан контроллер памяти ПЛИС с расширенной для обеспечения работы процессорных ядер “по готовности данных” функциональностью. Продемонстрирована возможность работы системных драйверов на такой архитектуре, в том числе разработан Ethernet контроллер и соответствующее системное ПО. Создано три рабочих экземпляра системы.

В рамках второй фазы отработаны механизмы связи ПЛИС по интерфейсу PCI Express, в том числе соответствующие интерфейсные ядра и механизмы перенастройки магистралей PCI Express для организации общей памяти у группы ПЛИС. Предложены и апробированы решения, позволяющие объединить более 100 ПЛИС с сохранением всех преимуществ высокоскоростной сети PCI Express.

В рамках второй фазы на двух ПЛИС семейства Virtex-6 реализован многоядерный процессор (100 ядер), основанный на модифицированной открытой Microblaze-совместимой архитектуре, рабочая частота ядра 100 МГц. Создан PCIe загрузчик. Разработана трехуровневая иерархия ядер и модуль запуска потоков на свободных ядрах, в том числе доработан коммуникационный протокол для распределения заданий между ядрами без участия ОС. Произведена доработка ОС Minix3 и ключевых сервисов. Адаптирован под указанную архитектуру эмулятор, позволяющий полноценно производить отладку ПО. Разработан контроллер памяти ПЛИС с расширенной для обеспечения работы процессорных ядер “по готовности данных” функциональностью. Продемонстрирована возможность работы системных драйверов на такой архитектуре, в том числе разработан Ethernet контроллер и соответствующее системное ПО. Создано три рабочих экземпляра системы. Производятся работы по отладке совместной работы как нескольких ПЛИС на одной несущей плате, так и нескольких содержащих ПЛИС модулей в составе одного или разных ВУ.

Литература

1. Корнеев В.В. Подход к программированию суперкомпьютеров на базе многоядерных мультитредовых кристаллов. Москва, НИВЦ МГУ им. М.В. Ломоносова. Научный журнал «Вычислительные методы и программирование» Том 10. 2009. С. 123-128
2. Корнеев В.В. Модель программирования – смена парадигмы. Москва. Открытые системы. №3. 2010.с. 29-31
3. Корнеев В.В., Будник А.В. Эффективность модели программирования на базе явного задания легких тредов. Вестник Южно-уральского государственного университета. № 18 (277). 2012. Серия «Математическое моделирование и программирование». Выпуск 12.С. 107-111
4. Елизаров Г.С., Горбунов В.С., Левин В.К., Лацис А.О., Корнеев В.В., Соколов А.А., Андрюшин Д.В., Климов Ю.А. Коммуникационная сеть МВС-ЭКСПРЕСС. Москва, НИВЦ МГУ им. М.В. Ломоносова. Научный журнал «Вычислительные методы и программирование» Том 13. 2012. С. 103-109

Елизаров Сергей Георгиевич
Физический Факультет МГУ имени М.В. Ломоносова, старший научный сотрудник
2002, Физический Факультет МГУ имени М.В. Ломоносова
К.ф.-м.н.
41 печатная работа
Компьютерные науки
elizarov@phys.msu.ru, (499)343-5624

Лукьянченко Георгий Александрович
Аспирант, НИЦ "Курчатовский институт"
2012, Физический Факультет МГУ имени М.В. Ломоносова
Не имею
1 печатная работа
Компьютерные науки
egorxe@yandex.ru, (499)343-5624

Корнеев Виктор Владимирович
ФГУП «НИИ «Квант», зам. директора по научной работе
1970, Алтайский политехнический институт им. И.И. Ползунова
Д.т.н., профессор
125 печатных работ и 5 монографий
Компьютерные науки
korv@rdi-kvant.ru, (499)153-4700