УДК 004.415.2.031.43

УПРАВЛЕНИЕ ВЫЧИСЛЕНИЯМИ И ПОТОКАМИ ДАННЫХ В РВС ДЛЯ ЗАДАЧ СЖАТИЯ ДАННЫХ В ТЕМПЕ ПОСТУПЛЕНИЯ

Е.А. Дудников
ООО «Научно-исследовательский центр супер-ЭВМ и нейрокомпьютеров»
Россия, 347900, Таганрог, пер. Итальянский, 106
E-mail: everlast-83@mail.ru

Ключевые слова: реконфигурируемые вычислительные системы (РВС), метод сжатия Хаффмана, управление и синхронизация потоков данных; сильносвязанные задачи реального времени; FPGA; структурно-процедурная организация вычислений.

Аннотация: Сжатие данных методом Хаффмана в темпе их поступления - трудоемкая сильносвязная задача, так как алгоритм вычислений в процессе сжатия многократно обращается к исходному кодируемому сообщению. Из-за подобных интенсивных межпроцессорных обменов и обращения к системе распределенной памяти, реальная производительность традиционных систем зачастую не превышает 10÷15% от заявляемой пиковой производительности. Реконфигурируемые системы имеют преимущества в удельной производительности, энергоэффективности и вычислительной эффективности по сравнению с кластерными системами при решении подобного класса задач. Данная работа рассматривает различные архитектуры современных РВС. Анализируются системы управления вычислениями и организации потоков данных в этих архитектурах.

1. Введение

Объем производимой человечеством информации неуклонно растет. Ежедневно генерируется порядка 328 миллионов терабайт данных. Уже в текущем 2024 году, согласно оценкам аналитиков, объем произведенной информации превысит 147 зеттабайт. Более того, опробованные во время пандемии удалённые формы обучения работы, которые приводят к повышенному производству и потреблению информации, были приняты и закреплены в повседневной жизни. Основными отраслями генерации данных становятся промышленность, финансовый сектор, медиабизнес и здравоохранение. Уже сегодня возникает острая потребность обрабатывать огромные потоки данных, в том числе и в режиме реального времени. Решение данной проблемы требует реорганизации инфраструктуры, внедрения новых технологий, наращивания пропускной способности каналов передачи данных, вычислительных возможностей. Это непременно приведёт к значительным финансовым затратам. Облегчить подобный переход могут новые высокоскоростные системы сжатия данных, позволяющие более эффективно использовать существующий технический ресурс. Разработка систем сжатия плотных потоков данных в реальном времени становится также актуальной в свете перспективных разработок фотонных компьютеров, где световые скорости обработки данных кратно превышают скорости передачи данных. На основе вышеизложенного можно заключить, что разработка методов и средств эффективного сжатия высокоскоростных потоков разнородных

данных «на лету» является актуальной и востребованной задачей. Аппаратной базой для подобных вычислительно трудоемких сильносвязанных задач могут стать РВС на базе FPGA. Данные устройства позволяют адаптировать имеющийся аппаратный ресурс к алгоритму решения практически любой вычислительно трудоёмкой задачи, а также обеспечивают возможность перепрограммирования для работы с новыми или оптимизированными алгоритмами. Реконфигурация системы без дорогостоящего обновления оборудования особенно важна, поскольку алгоритмы сжатия постоянно развиваются, меняются, становятся более сложными. Обеспечить подобную гибкость на процессорах и ускорителях со статической архитектурой практически невозможно.

2. Архитектуры современных РВС на базе FPGA

Существующие современные РВС можно разделить на однокристальные и многокристальные (многоплатные). Разработкой однокристальных РВС различных конфигураций для промышленных систем, авиационной, аэрокосмической и оборонной промышленности занимаются изготовители FPGA и сторонние производители (Xilinx, Intel, Achronix, SouthEast Technical Sales, Extreme Engineering Solutions и т.д.). Хотя производители однокристальных РВС и позиционируют свои изделия как самодостаточные высокопроизводительные вычислительные устройства, их применение целесообразно лишь для решения задач, не обладающих большой вычислительной сложностью и громоздкостью. Небольшой объём оперативной памяти устройств не позволяет эффективно обрабатывать значительные массивы данных, характерных при решении сильносвязанных вычислительно трудоёмких задач, к числу которых относятся задачи по сжатию данных. Архитектура однокристальных РВС требует интенсивного информационного обмена между РВС и внешними устройствами памяти и управления. Типовые интерфейсы передачи данных не способны обеспечить подобные информационные обмены на должном уровне быстродействия. Вышеописанные недостатки приводят к тому, что для выполнения задач сжатия данных разработчикам приходится применять статистические таблицы, предварительную обработку данных и прочие ухищрения, что неизбежно приводит к падению производительности или плотности сжатия данных.

Многокристальные высокопроизводительные вычислительные системы, использующие FPGA в качестве элементной базы, принято разбивать на два типа. К первому типу относятся так называемые гибридные системы, представляющие собой классические кластерные вычислители, в которых FPGA используются в качестве ускорителей. В этих системах блоки программируемых сопроцессоров реализованы на базе FPGA, связанных скоростными магистралями с CPU и между собой. По способу подключения CPU и FPGA для управления вычислениями и обменом данными гибридные архитектуры многокристальных РВС можно разделить на четыре класса.

Первый класс (рис. 1, а) представляет собой кластер CPU и FPGA, связанных распределенной памятью или общей памятью. Это классический и распространенный кластер, в котором узлы ЦП в сочетании с FPGA соединены друг с другом системной сетью. Однако у него возникают трудности с доступом к FPGA на разных узлах ЦП, а также ограниченные возможности связи между FPGA.

Во втором классе (рис. 1, б) узлы CPU и FPGA эквивалентно подключены к одной системной сети, где произвольные CPU и FPGA могут взаимодействовать друг с другом (к этому классу относятся CloudFPGA от IBM [1], Microsoft Catapult v2 [2] и Galapagos [3]). CloudFPGA от IBM – кластерная система FPGA высокой плотности для облачных вычислений, в которой узлы FPGA напрямую подключены к сети центра обработки данных. Microsoft Catapult v2 имеет уникальную особенность: FPGA

вставляется между сетевой платой процессора и сетевым коммутатором, где FPGA напрямую подключается к сети для обработки входящих пакетов. Galapagos - это крупномасштабная гетерогенная вычислительная среда CPU и FPGA для использования в облачных вычислениях. Хотя этот класс кластеров FPGA позволяет любым процессорам и FPGA взаимодействовать друг с другом, перегрузка с различными и неожиданными шаблонами трафика в общей сети может привести к снижению производительности, пропускной способности и задержкам связи.

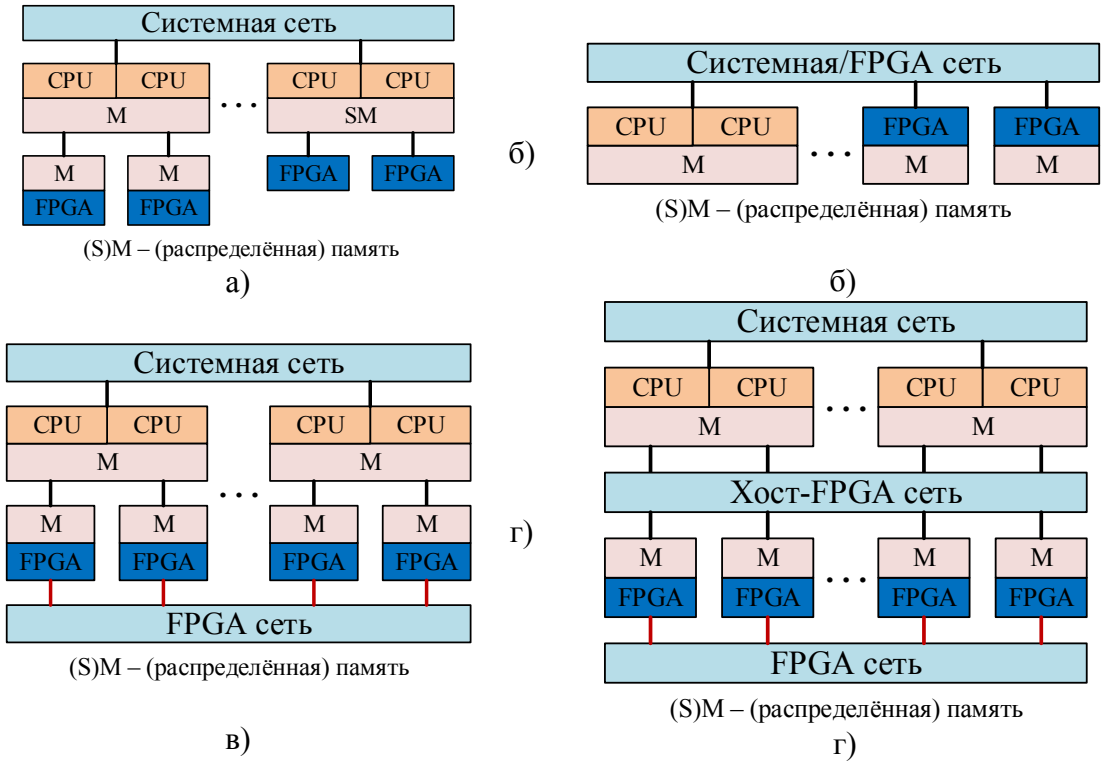

Рис. 1. Классификация различных архитектур гибридных РВС на базе ПЛИС.

Третий класс (рис. 1, в) – система CPU с FPGA, где напрямую FPGA, соединенные между собой сетью, образуют отдельной кластер. Эта конфигурация предназначена для улучшения ограниченных возможностей связи между FPGA первого класса за счет выделенной сети между FPGA. Это позволяет пользователям организовать межпроцессорные обмены FPGA отдельно от сети CPU. Типичные высокопроизводительные карты FPGA имеют высокоскоростные порты связи, такие как QSFP28, со скоростью 100 Гбит/с, их можно соединять друг с другом по прямой сети или с помощью коммутаторов. Существует несколько машин такого класса - Microsoft Catapult v1, Amazon EC2 F1, Novo-G/Novo-G# [4], Cygnus и Noctua. У этого класса есть проблема с доступом к FPGA для разных узлов CPU.

Четвертый класс (рис 1, г) представляет слабосвязанный кластер CPU и кластер FPGA, сочетающий преимущества второго и третьего классов за счет разделения системы на отдельные кластеры CPU и FPGA и свободного их соединения с помощью мостовой сети «хост-FPGA» для большей гибкости. Мостовая сеть позволяет расширить существующую систему с помощью кластера FPGA и управлять любыми FPGA локально или удаленно с помощью любого процессора, что важно для крупномасштабной системы, в которой несколько пользователей динамически совместно используют ресурсы РВС. К этому классу относится прототип кластера FPGA ESSPER на базе суперкомпьютера Fugaku. Тесты передачи данных между CPU и локальной FPGA показали, что копирование данных из виртуальной памяти в

циклические буферы является узким местом, а пропускная способность не превышает трети от пиковой [5].

В рамках европейского проекта по созданию экзафлопного суперкомпьютера было разработано несколько проектов (ExaNeSt, ECOSCALE и т.д. Все это привело к созданию прототипа EuroEXA [6]. Основой архитектуры РВС EuroEXA является вычислительный модуль (CRDB), содержащий две ПЛИС: Xilinx ZU9 рассчитан на интерконнект с верхним уровнем, VU9P является основным ускорителем вычислений. FPGA имеют тесные связи лишь в пределах вычислительного модуля. Межсоединение CRDB с блейдом по типу Dragonfly является узким местом, и накладные расходы на связь между CPU и аппаратным ускорителем являются одной из наиболее важных проблем, даже несмотря на то что аппаратные ускорители могут напрямую обращаться к памяти хоста ЦП, расположенного на одном кристалле.

Перечисленные выше системы, по сути, представляют собой стандартные кластеры, поддерживаемые MPI, дополненные высокопроизводительными ресурсами FPGA, которые обрабатывают отдельные трудоемкие участки алгоритмов. Управление исполнением алгоритма возлагается на универсальные процессоры. Вычисления выполняются, согласно принципам «control flow». При данной организации потоки программы можно представить в виде графа, где узлы соответствуют инструкциям программы, а рёбра – переходам между ними. Стандартные интерфейсы не способны обеспечить высокоскоростной обмен большими массивами данных и пакетов команд.

Ко второму типу вычислительных систем относятся РВС, использующие FPGA в качестве основного вычислительного ресурса. FPGA собираются в мощные вычислительные поля, в рамках которых пользователь может создавать различные проблемно-ориентированные вычислительные структуры, адекватные алгоритму решаемой задачи, с использованием универсальных процессоров для вспомогательных функций. Чем больше будет такое поле FPGA, тем ниже будут неоправданные накладные расходы на организацию вычислительного процесса и тем выше будет ее реальная производительность.

Такой аппаратной базой для задач по сжатию данных в темпе их поступления может стать перспективный реконфигурируемый вычислительный блок «Арктур». РВБ «Арктур» содержит 16 вертикально расположенных на кросс-плате вычислительных модулей (ВМ) по шесть FPGA XCVU37P фирмы Xilinx в каждом. В блоках используется новый тип оперативной памяти HBM – высокопроизводительная динамическая оперативная память DRAM с многослойной компоновкой кристаллов.

Особенность ВМ «Арктур» - расширенные возможности информационного обмена. Связь между FPGA обеспечивается по дифференциальным линиям с помощью встроенных в FPGA мульти-гигабитных трансиверов (MGT), вспомогательными связями являются дифференциальные LVDS линии, подключенные к HP-банкам FPGA с применением топологии 2D-тор. Между парой FPGA реализованы 24 дифференциальные линии со скоростью передачи данных до 24 Гбит/с в пределах ВМ и 24 дифференциальные линии со скоростью передачи данных до 16 Гбит/с между ВМ. Общая пропускная способность каналов связи ВМ – 15,6 Тбит/с, в том числе между ВМ – 9 Тбит/с. Возможна организация информационного взаимодействия между РВБ «Арктур» при построении вычислительных комплексов, которое осуществляется через оптические каналы. Такая архитектура позволяет эффективно решать вычислительно трудоемкие сильносвязанные задачи различных проблемных областей, в которых количество пересылок данных между функциональными устройствами сравнимо с числом вычислительных операций, обеспечивая линейный рост производительности при наращивании аппаратного ресурса [7]. Ключевые технические решения позволяют достичь реальной производительности решения на РВБ прикладных задач до

200 Тфлопс (single precision), а при построении суперкомпьютеров достигнуть уровня реальной производительности сотен петафлопс.

3. Заключение

Анализ существующих архитектур РВС на базе FPGA показывает, что однокристальные РВС используются как ускорители вычислений и в большей степени пригодны только для реализации небольших задач. Многокристальные гибридные РВС, где FPGA также отводится роль ускорителя, не лишены проблем однокристальных РВС, из которых зачастую и формируются. В основу данных вычислительных систем заложен классический принцип процедурных вычислений. Система управления в подобных системах является иерархической и возлагается на CPU, который распределяет пакеты команд и данных для обработки по поддоменам. Подобная организация с интенсивным обменом данными через стандартные интерфейсы приводит к задержкам связи между блоком управления, памятью и реконфигурируемым ускорителем, ограничению доступа к FPGA для разных CPU, снижению пропускной способности и общей производительности системы. Отдельной проблемой стоит синхронизация потоков данных.

К РВС второго типа относятся системы, построенные по принципу «dataflow», где вычислительная структура обрабатывает потоки поступающих операндов параллельно-конвейерным образом. FPGA выступают основным вычислительным ресурсом и формируют единое вычислительное поле, на базе которого можно создать вычислительную структуру, соответствующую информационному графу задачи. Управление вычислениями заложено в структуру и не требует дополнительной синхронизации потоков входных и выходных данных между операционными вершинами графа задачи. Широкие аппаратные архитектурные связи между кристаллами подобных РВС обеспечивают обмен данными с огромной пропускной способностью и позволяют решать вычислительно трудоёмкие информационно сильносвязанные задачи с высокой степенью информационных разрывов, такие как задачи сжатия данных в темпе их поступления. При наращивании аппаратного ресурса в подобных РВС гарантируется близкий к линейному рост производительности при решении потоковых задач, требующих обработки больших массивов данных в темпе их поступления за оперативно приемлемое время.

Список литературы

1. Abel F., Weerasinghe J., Hagleitner C., et. al. An FPGA Platform for Hyperscalers // IEEE 25th Annual Symposium on High-Performance Interconnects (HOTI), 2017. P. 29-32.
2. Caulfield A.M., Chung E.S., Putnam A., et. al. A cloud-scale acceleration architecture // 2016. P. 1-13.
3. Tarafdar N., Eskandari N., Sharma V., Lo C., Chow P. Galapagos: A Full Stack Approach to FPGA Integration in the Cloud // IEEE Micro 38, 2018. Vol. 6. P. 18-24.
4. George A.D., Herbordt M.C., et. al. Novo-G#: Large-scale reconfigurable computing with direct and programmable interconnects // IEEE High Performance Extreme Computing Conference (HPEC), 2016. P. 1-7.
5. Sano K., Koshiba A., Miyajima T., Ueno T. ESSPER: Elastic and Scalable FPGA-Cluster System for High-Performance Reconfigurable Computing with Supercomputer Fugaku // HPC Asia '23: Proceedings of the International Conference on High Performance Computing in Asia-Pacific Region, 2023. P. 140-150.
6. Biagioni A., Cretaro P., Frezza O., Lo Cicero F., et. al. EuroEXA Custom Switch: an innovative FPGA-based system for extreme scale computing in Europe // The European Physical Journal Conferences, 2020.
7. Каляев И.А., Левин И.И. Реконфигурируемые вычислительные системы на основе ПЛИС. Ростов-на-Дону: ЮФУ, 2021. 458 с.