УДК 004.27; 004.31; 004.382.2.

Ю. С. ЯКОВЛЕВ

Институт кибернетики имени В.М. Глушкова НАН Украины, г. Киев

ОСОБЕННОСТИ АРХИТЕКТУРЫ И СТРУКТУРЫ РЕКОНФИГУРИРУЕМЫХ ГРАФИЧЕСКИХ УСКОРИТЕЛЕЙ

Анотація. розглянуто особливості побудови графічних прискорювачів з реконфігурацією шляхом використання: комутаційних схем для вибору оптимальних ресурсів прискорювача; масштабування системи за рахунок використання модифікованої кільцевої шини; застосування ПЛИС для конфігурації прискорювача під тип вирішуваної задачі; архітектури типу "процесор-в-пам'яті" із застосуванням запропонованого методу розподілу графічної задачі по процесорах системи. При цьому запропоновані архітектурно-структурні рішення захищені патентами України.

Ключові слова: реконфігурація, графічний прискорювач, "процессор-в-памяти", програмована логічна інтегральна схема (ПЛІС).

Аннотация. рассмотрены особенности построения графических ускорителей с реконфигурацией путем использования: коммутационных схем для выбора оптимальных ресурсов ускорителя; масштабирования системы за счет использования модифицированной кольцевой шины; применения ПЛИС для конфигурации ускорителя под тип решаемой задачи; архитектуры типа "процессор-в-памяти" с применением предложенного метода распределения графической задачи по процессорам системы. При этом предложенные архитектурно-структурные решения защищены патентами Украины.

Ключевые слова: реконфигурация, графический ускоритель, "процессор-в-памяти", программируемая логическая интегральная схема (ПЛИС)

Abstract. features of construction of accelerators with reconfiguration by usage are observed: diagrammes of connections for sampling of optimal resources of the accelerator; scalings of system at the expense of usage of the updated ring bus; applications the PLIS for an accelerator pattern under type of the solved task; architectures of type "processor-in-memory" with application of the offered method of allocation of the graphics task on system processors. Thus offered is architectural-structural solutions are protected by patents of Ukraine.

Keywords: reconfiguration, the accelerator, the "processor-in-memory", the programmed logical chip (PLC)

Введение

Обработка графических изображений используется в различных областях, в том числе в играх для отображений различных игровых сцен, в медицине, например, при обработке гистограмм, снимков и т.д., в военном деле, в системах автоматизированного проектирования (САПР) и т. д. При этом на каждом этапе развития средств вычислительной техники (ВТ) требования к качеству и скорости обработки изображений постоянно восрастают, что в принципе может быть достигнуто за счет расширения полосы пропускания информации по каналу процессор-память, совешенствования системы памяти, обеспечивающей малое время доступа к данным, применением более мощных процессоров и множества других факторов, которые в совокупности, как правило, представлены в существующем множестве графических ускорителей основных фирм – произволителей ATI (AMD) и NVIDIA [1, 2]. Практически во всех указанных случаях их применения заметный положительных эффект может быть достигнут, если алгоритмы подлежат хорошему распараллеливанию и содержат большое количество простых однотипных вычислений. С этой точки зрения, обработка изображения как раз и является той областью, где распараллеленная часть алгоритма, как правило, имеет большой процент по сравнению с другими целевыми прикладными программами.

Помимо графических ускорителей для повышения качества и скорости обработки изображений используются также различные средства, выполненные на микросхемах ASIC (например, устройства цифровой обработки сигналов – ЦОС), на микросхемах памяти ("Processor-in-memory" или PIM-системы), обширный класс которых относят к категории CIMA (Computing-in-Memory Architecture), а также средства, обладающие свойствами реконфигурации, в том числе – использующие программируемые логические интегральные схемы (ПЛИС)

Актуальность

Актуальность данной работы продиктована, с одной стороны, необхлдимостью улучшения качества отображения графической информации за счет совершенствования средств и способов обработки графики и, с другой стороны, – отсутствием систематизированного подхода к анализу и построению перспективных архитектурно-структурных решений средств обработки графики, обеспечивающих по сравнению с известными более высокое качество и скорость обработки изображений при меньших затратах ресурсов системы и улучшенных параметрах пользовательских характеристик (например, энергопотребления и тепловыделения).

Цель

В соответствие с этим, целью данной работы является анализ и отображение существующих и перспективных архитектурно - структурных решений, обладающих свойствами реконфигурации, выполненных на современной элементной базе и ориентированных на обработку графики, включая PIM-системы, а также программируемые логические интегральные схемы – ПЛИС и однокристальные

графические процессоры. При этом не рассматриваются, так называемые, графические ускорители фирм ATI (AMD) и NVIDIA, анализу и применению которых посвящены многочисленные публикации.

Задачи

Основными задачами являются:

1) Анализ особенностей архитектурно-структурной организации специализированных графических процессоров, выполненных на современной элементной базе и ориентированных на обработку графики (например, совершенствованные технологии VLSI, технологии цифровой обработки сигналов ЦОС и др.)

2) Анализ особенностей архитектурно-структурной организации и применения средств для обработки графических изображений, в том числе PIM-систем, обладающих свойствами реконфигурации.

3) Сравнение основных параметров реконфигурируемых графических систем, выполненных с применением различных архитектурно-структурных решений.

Решение задач

Все возрастающие требования к качеству и скорости обработки изображения приводит к выбору высокопроизводительного микропроцессора при построении соответствующих систем, для которых обычно предъявляются весьма строгие требования к таким параметрам, как габариты, стоимость, потребляемая мощность и время вывода нового изделия на рынок. Эти требования часто не могут быть удовлетворены только применением более мощного микропроцессора. Поэтому в промышленных изделиях функции обработки изображения осуществляются как с применением графических ускорителей фирм ATI (AMD) и NVIDIA, так и с применением специализированных процессоров таких, как процессоры цифровой обработки сигналов (ЦОС) или– Digital Signal Processors (DSPs), а также с помощью специфического программного обеспечения Application Specific Standard Products (ASSPs). Однако, при увеличения сложности обработки изображения для ЦОС необходимо использовать большое количество параллельных модулей. Такие скоростные ЦОС становятся дорогими, и их производительность имеет тенденцию отставать от соответствующих требований при обработке изображения. С другой стороны, ASSPs являются негибкими, дорогими и трудоемкими, отнимающими много времени при модификации, несмотря на то, что врожденный параллелизм при обработке изображения предлагает приложению высокую эффективность технологии вычислений (High Performance Computing - HPC).

Аппаратное ускорение – эффективный способ увеличить производительность при использовании специализированной аппаратной архитектуры, которая выполняет параллельную обработку. С появлением программируемой вентильной матрицы (Field Programmable Gate Array – FPGA), специализированная аппаратная архитектура может быть реализована с более низкой стоимостью, при этом мощность потребления может быть уменьшена, поскольку схема оптимизируется под приложение. Реконфигурируемое вычисление может также уменьшить размер канала “память-процессор” (и, следовательно, стоимость) и обеспечить дополнительную гибкость за счет динамической реконфигурации во время выполнения приложения. В результате производительность обработки изображения может быть улучшена по сравнению с мощным микропроцессором на несколько порядков, что соответствует требованиям для многих приложений.

Операции обработки изображения включают большое количество операции умножения-накопления (MAC). При этом микропроцессоры выполняют несколько последовательных шагов (выборка машинной команды, декодирование, выполнение и получение результата на запрос) для каждой операции MAC. Так как эти шаги выполняются последовательно, то трудно ускорить вычисление без архитектурных изменений. FPGAs может обеспечить высокоэффективное MAC – вычисление, используя специализированную архитектуру, которая в состоянии выполнить несколько операций MAC одновременно

Аппаратные сопроцессоры – ускорители базовых операций обработки графики. Ниже рассматриваются две разновидности архитектур, которые вычисляют базовые операции алгоритма обработки изображений: взаимную корреляцию в пространственных и спектральных областях. Другие операции, такие как фильтрация и свертка можно рассматривать как специфические случаи корреляционных [3]. Обе архитектуры могут быть реконфигурированы, чтобы соответствовать различным приложениям и размерам FPGA. На рис.1 представлена упрощенная архитектура аппаратного средства для обработки графики, которая отнесена к категории пространственной. Основой этой архитектуры является систолический массив модулей MAC, подключенных к памяти, где каждый модуль MAC выполняет операции и посылает результат к следующему модулю в той же самой строке. Последний MAC каждой строки выдает результат вычисления всей строки, который подается на вход MAC следующей строки и т.д. (рис.1). Результат вычисления матрицы получают путем суммирования результатов последовательных строк N при накоплении результата предыдущей строки к следующей.

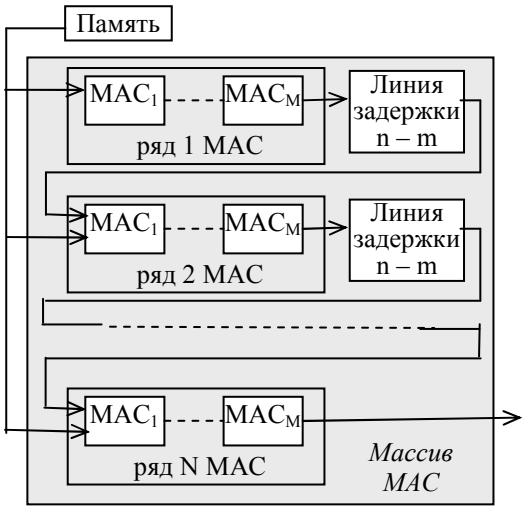

Рисунок 1 – Фрагмент пространственной архитектуры реконфигурируемого графического ускорителя

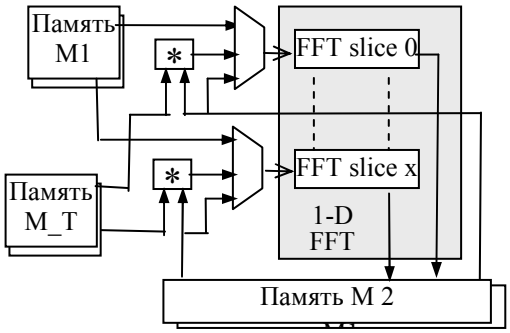

Рисунок 2 – Фрагмент архитектура для вычисления взаимной корреляция в спектральной области

Такая архитектура ориентирована на выполнение операций линейной фильтрации, свертки и методов корреляции, которые обычно используются во многих приложениях обработки изображения.

Строка задержки N - M. циклически повторяется и вставляется между выводом последнего MAC модуля в каждой i - й строке и вводом результата в первый модуль MAC в следующей (i+ 1) строке. При этом схемы задержки могут быть реализованы как сдвиговые регистры или как память типа FIFOs в зависимости от выбранной технологии и типа устройства. Для этого итерационного процесса требуется промежуточная память для хранения результата каждого уровня, а также мультиплексор между массивом MAC и памятью, чтобы выбрать соответствующие входные данные при каждой итерации.

Недостатком этой архитектуры является необходимость применения линий задержки и малая емкость памяти, размещенная на кристалле FPGA. Поэтому при организации на FPGA памяти приходится увеличивать её емкость, одновременно уменьшая тактовую частоту в два раза. При этом увеличение ширины шины памяти в два раза обеспечивают требуемую массивом MAC скорость за счет передачи данных одновременно для двух последовательных пикселов.

На рис. 2 представлена архитектура для вычисления взаимной корреляция (CC) в спектральной области [3]. В этой архитектуре 1-D FFT массив используется вместо массива MAC, приведенного на рис.1. Для вычисления взаимной корреляция в спектральной области изображение разбивается на строки или столбцы, чтобы выполнить параллельную обработку. Модули FFT (Fast Fourier Transform) вычисляют строки/столбцы FFTs и называются секторами FFT.

Максимальное число распределенных секторов FFT зависит от размера изображения и выбранного устройства для реализации. В этом процессе используются три различных блока памяти: блок памяти M1 (рис. 2), предназначенный для запоминания входного изображения и результата, блок памяти M_T для запоминания FFT(T) и блок памяти M2 для запоминания промежуточных результатов. У всех этих блоков памяти – одинаковая емкость, часть которой предназначена для хранения размера входного изображения, а другая – для хранения результатов вычисления FFT. Эти блоки памяти могут быть выполнены, используя внутренние блоки микросхемы FPGA. Блоки памяти M1 и M2 являются независимыми, чтобы обеспечить достаточное количество портов памяти для каждого сектора FFT. Чтобы избежать конфликтов при доступе к памяти, обращения к блокам памяти осуществляют по диагонали при сдвиге во времени секторов FFT друг относительно друга. Описанная архитектура полезна также для свертки и фильтрации спектральной области, и требуется только незначительные изменения: в частности при реализации свертки удаляют сопряженную операцию при выводе информации из блока памяти M_T.

Анализ экспериментальных результатов, приведенных в [3], показал, что при применении FPGA можно достичь ускорения больше, чем на два порядка величины относительно программных реализаций. Например, пространственная аппаратная реализация архитектуры на FPGA для шаблона 16x16 обеспечивает в 688 раз быстрее, чем эквивалентная программная реализация и в 244 раза быстрее, чем спектральная реализация на современном персональном компьютере без потери точности. При этом можно отметить, что у пространственной архитектуры – умеренное потребление ресурсов FPGA, которое увеличивается с увеличением размеров шаблона. Основное ограничение для пространственной

архитектуры - число секторов DSP, доступных в выбранном устройстве FPGA. С другой стороны, спектральная архитектура потребляет большинство ресурсов FPGA, включая логику, внедренные блоки памяти и DSP сектора, и только немного быстрее, чем пространственная реализация. Однако, спектральная архитектура может быть выгодной для больших размеров шаблона, поскольку продолжительность обработки не зависит от размера шаблона.

Реконфигурируемые однокристальные графические ускорители. Рассмотренные выше архитектуры ориентированы на выполнение отдельных (базовых) операций алгоритма обработки изображений. Чтобы управлять потоком данных и выполнить остальную часть алгоритма, необходима интеграция с другими элементами (микропроцессором или PC).

Эта проблема решается в пределах сценария “Система на Программируемом Чипе”(“System on Programmable Chip” – SoPC). Микропроцессор, размещенный на FPGA, подключен к аппаратному сопроцессору, который ускоряет самые сложные операции. Поскольку все компоненты SoPC располагаются в единственном чипе, то соответственно уменьшается размер, стоимость и потребляемая мощность. Для исследования реальной системы была выбрана микросхема Virtex 5 FPGA, XC5VSX50T фирмы Xilinx. Система предназначается для высокоэффективной обработки изображения и содержит внедренный микропроцессор, блоки памяти и коммуникационные шины между элементами, а также аппаратный сопроцессор, который ускоряет критические операции. Упрощенная структурная схема такой системы приведена на рис.3. При этом сопроцессор обладает свойствами реконфигурации и спроектирован с использованием пространственной архитектуры из-за её многосторонности (многофункциональности) и меньшим потреблением ресурсов. Архитектура сопроцессора показана на рис. 4.

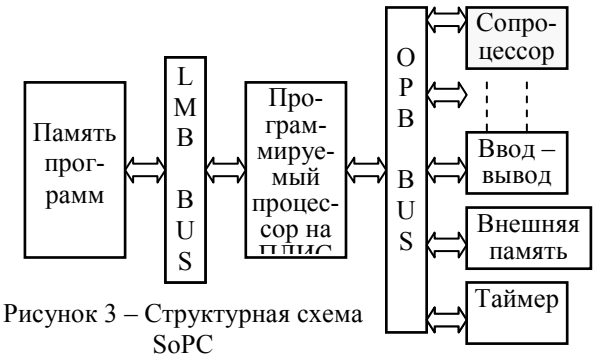

Рисунок 3 – Структурная схема SoPC

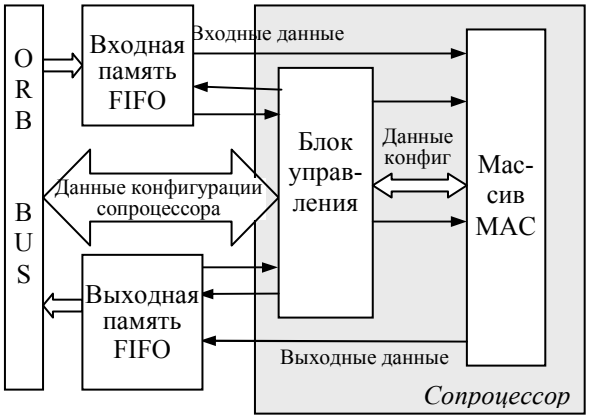

Рисунок 4 – Упрощенная архитектура сопро-цессора (рис.3)

Параметры конфигурации упаковываются в 4-х 32-разрядных регистрах так, чтобы полная конфигурация сопроцессора могла быть осуществлена для 4-х транзакций OPB. Как только сопроцессор запрограммирован, обработка данных начинается немедленно, пока данные доступны при вводе в порядке поступления. Сопроцессор работает параллельно с порядком поступления, и выходные данные памяти выводятся также в порядке поступления. Данные ввода-вывода передаются из/в оперативную память по каналу прямого доступа в память без вмешательство микропроцессора.

Как только результаты готовы, они сохраняются в порядке поступления и посылаются в микропроцессор через OPB. Реконфигурируемая производительность сопроцессора была измерена для предложенного алгоритма. Результаты были получены с и без сопроцессора для сравнения. Программные результаты были получены с Pentium IV PC, 3.2 ГГц и 2 Гбайта оперативной памяти. Потребление ресурса сопроцессора главным образом определяется размером массива MAC и длиной строк задержки. Потребление ресурса FPGA умеренно. Половина FPGA остается пустой, поэтому больше компонентов может быть добавлено в систему SoPC. Результаты SoPC были проанализированы без обнаружения потеря точности в алгоритмах. Операционная частота SoPC составляла 100 МГц, она является и частотой OPB. Производительность была измерена для предложенных алгоритмов. Эти измерения также взяты для SoPC без сопроцессора и PC. SoPC работает намного быстрее(до 49 раз быстрее) чем программируемый процессор на ПЛИС (Microblaze) и немного быстрее, чем PC. Для больших размеров изображений SoPC в 157 раз работает быстрее, чем Microblaze и приблизительно в 3 раза быстрее чем PC.

Таким образом, экспериментальные результаты демонстрируют, что все представленные решения обеспечивают ускорение до 3 порядков величины по отношению к эквивалентной программной реализации. Эта архитектура может быть реконфигурируема для многих приложений и к потребностям

многих систем обработки изображений. Архитектура является масштабируемой для любого семейства FPGA и приспосабливаемой к любому FPGA. Чтобы получить аппаратное ускорение при обработке изображения, реконфигурируемые аппаратные средства и сопроцессор интегрируется в SoPC. Эта система может обеспечить гибкость, микропроцессор, который размещен на FPGA, обеспечивая тем самым высокую производительность при выполнении сложных (критических) операций.

Реконфигурируемые графические ускорители с архитектурой "Processor-in-memory" (процессор-в-памяти). Большинство PIM – архитектур лучше всего функционирует как сопроцессоры компьютерных систем, выполняя функции, для которых они оптимизированы, в то время как другие прикладные программы, в которых основной является последовательная часть, реализуются главным (хост) процессором. Следует отметить, что PIM-система хорошо приспособлена под обработку изображения [4], поскольку обеспечивает возможность параллельной обработки с помощью множества процессоров (процессорных ядер – ПЯ) большого количества фрагментов алгоритма (несколько сотен и более), выполняя одновременные соответствующие вычисления для сотен и тысяч пикселов изображения . При этом, благодаря особенностям архитектурно-структурной организации PIM-систем, обеспечивается широкая полоса пропускания по каналу процессор-память, что существенно сказывается на повышении производительности графической системы в целом. Из-за малых размеров данных, связанных с большинством операций обработки графики (8 – 16 бит), вычислительная мощность современных 32-или 64-разрядных процессоров с возможностями обработки с плавающей запятой, не является необходимой. Вместо этого, матрица или вектор из более простых 8-разрядных элементарных процессоров (PEs), специализированных к определенным потребностям применений, может быть весьма эффективной

Начало обработки изображений с помощью средств типа PIM-систем положило простейшее устройство с архитектурой SIMD – Single-Instruction stream and Multiple-Data stream (один поток команд – поток множества данных), содержащее на одном кристалле множество одноразрядных средств обработки со схемами сдвига влево и вправо, каждый из которых подключен к одному столбцу данных, размещенных в памяти [5]. Это один из первых нового поколения CIMA (Computing-in-Memory Architecture) проектов, который предложен и построен как CRAM проект, разработанный Университетом Торонто в 1992. Благодаря распараллеливанию алгоритма обработки изображений и в итоге широкой эквивалентной полосе пропускания по каналу множество процессоров – множество столбцов данных, размещенных в памяти, это устройство, используемое как приставка-ускоритель к основному процессору, обеспечивает производительность порядка Tera (Тера) команд в секунду. Устройство подобного типа также предложено в [6], где часть вычислений выполняется в узле PIM, однако основные вычисления выполняются на суперскалярном процессоре MP. Тем самым PIM-система используется в качестве ускорителя МП. При этом, для увеличения производительности системы, увеличивают количество узлов памяти типа PIM. Подобный простой проект PIM - устройства Terasys, был объявлен в 1995. Этот проект содержит большую матрицу простых элементов вычисления (типично более чем 1.000) которые сформированы в матрице DRAM. Элементарные процессорные элементы объединены по выходному сигналу усилителей считывания и управляются единственным устройством управления. Поэтому эта архитектура работает как SIMD архитектура - в любом цикле обращения к памяти типа DRAM, каждый процессор в матрице выполняет одну и ту же команду над её собственными локальными данными. Несмотря на то, что с архитектурной точки зрения, этот проект – самый простой, тем не менее, он может теоретически обеспечить высокую производительность при реализации алгоритмов с хорошим распараллеливанием. Однако, эти типы массово – параллельных, однопроцессорных элементов SIMD-приборов в составе PIM-системы могут быть эффективны только на наборе прикладных программ, которые могут быть легко распараллелены. Для прикладных программ с существенным количеством последовательных вычислений, увеличение скорости за счет применения этой архитектуры – ограничено.

В целом можно отметить, что в настоящее время диапазон проектируемых устройств, которые подпадают под архитектурный стиль "Processor-in-memory", богат и различен [3, 4], и большинство из них могут успешно применяться для обработки графических изображений.

Дальнейшее развитие PIM-систем, ориентированных на обработку графики, шло одновременно с развитием интегральной технологии по пути совершенствования процессорных элементов, коммутационной среды и обеспечения возможностей реконфигурации с целью настройки архитектуры на особенности приложения и возможностей реконфигурируемой среды (микросхемы ПЛИС). С этой точки зрения особый интерес представляют архитектура PIM-систем типа VIRAM (vector-intelligent-random access memory) и DIVA (Data IntensiVe Architecture – интенсивная архитектура данных) [7]. При этом VIRAM, используя новую технологию PIM, содержит встроенную в кристалл динамическую оперативную память (DRAM) с векторным сопроцессором, обеспечивая тем самым эффективное использование полосы пропускания по каналу память-процессор. Особенностью архитектура чипа

VIRAM является то, что это – законченная система на чипе, содержащая элементарные процессоры и стандартную динамическую оперативную память (DRAM).

PIM-система типа DIVA содержит множество чипов PIM в качестве интеллектуальных сопроцессоров и использует широкие информационные каналы памяти большой емкости, применяя при этом мелкомодульную параллельность. Каждый чип DIVA PIM-системы включает ЗУ с аппаратными средствами вычисления и коммуникации и содержит два информационных канала, механизмы которых координируются единственным блоком управления: 32-разрядный скалярный информационный канал и 256-разрядный широкий информационный канал. Скалярный информационный канал - стандартная архитектура RISC-процессора с расширенными функциями DIVA-specific для того, чтобы координироваться с широким информационным каналом. Широкий информационный канал воздействует на составные объекты (суперслова) 256 бит, выполняя параллельные работы SIMD на установленных по размеру полях объекта (8, 16, и 32-разрядных полях). В дополнение к условным арифметическим действиям и логическим операциям, широкое арифметико-логическое устройство также поддерживает большой набор операций для того, чтобы манипулировать данными, включая перегруппировку данных в пределах широкого слова-операнда, передачи между широкими и скалярными регистрами, собирая в структуру и распаковывая операции. Кроме того, широкое арифметико-логическое устройство поддерживает выборочное выполнение команд на базисе информационного канала в зависимости от состояния кодов условия.

Графический ускоритель, построенный на принципах потоковой обработки. Другой подход к повышению производительности системы за счет эффективного использования ширины полосы пропускания памяти основан на принципах потоковой обработки. Например, программируемый потоковый микропроцессор Imagine [7], ориентирован на обработку интенсивных приложений, отличающихся высоким параллелизмом данных с небольшим глобальным многократным их использованием. Imagine – это 32-разрядная архитектура с поддержкой выполнения операций для 16-и 8-разрядных данных, обеспечивая в результате в два и в четыре раза повышение пиковой производительности соответственно.

В [8] предлагается архитектура PIM-системы, ориентированная на обработку компьютерной графики с массовым обращением к памяти. Чтобы достигнуть максимальной производительности, рабочая нагрузка должна быть равномерно распределена среди процессоров, а размещение данных и распределение задачи пользователя должны уменьшить коммуникацию между процессорами,

Предложения по повышению эффективности графических ускорителей, выполненных на PIM-системах.

Реконфигурируемая PIM-система с выбором оптимальных ресурсов средствами коммутации

Совершенствование PIM-системы, ориентированной на обработку графики, предлагается путем расширения её функциональных возможностей за счет динамической настройки архитектуры системы на класс решаемых задач, а также за счет оптимальной загрузки всех процессорных элементов полезной (вычислительной) работой. При этом за основу принимаются: максимальный учет особенностей построения и применения существующих PIM-систем и методов их организации и ориентация PIM-систем на обработку хорошо распараллеливаемых алгоритмов при массовом обращении к памяти за данными, а также возможность реализации PIM-чипа на современных и перспективных ПЛИС. Показано, что основные свойства ПЛИС как элементной базы соответствуют особенностям структурной организации PIM-систем, на основании чего определены факторы, определяющие пригодность PIM-системы для реализации на ПЛИС [9].

Для расширения функциональных возможностей на кристалле PIM-системы размещены (на рис. 5 закрашены серым цветом) [10]:

– набор (вместо одного) проблемно-ориентированных (или специализированных) ведущих процессоров (ВП) вместе с соответствующими блоками памяти типа КЭШ;

– блок анализа входного управляющего пакета (БОУП1) для выработки всей совокупности управляющих сигналов, необходимых для функционирования БИС;

– селектор С1 для выбора и активации ведущих процессоров;

– набор селекторов С2 и С3 для выбора процессорных ядер (ПЯ) и соответствующих банков памяти;

– устройство для загрузки программ и данных в массивы памяти (УЗг);

– устройство интерфейса диагностики и отладки (Ин.Д/О)

Селекторы (коммутаторы) С1 – С3 управляются блоком обработки управляющего пакета и формирования управляющих сигналов (БОУП1), необходимых для работы всей системы.

Управляющий пакет содержит поля и признаки, отражающие последовательность действий PIM-системы, обеспечивающих: настройку архитектурно-структурного образа системы перед запуском её на решение конкретной задачи, перестройку системы в процессе работы, а также различные режимы работы системы памяти, в том числе: в качестве обычной памяти с использованием емкости памяти,

размещенной только на кристалле (чипе), а также на кристалле и дополнительной внешней памяти, подключенной к чипу через интерфейс внешней памяти; в качестве "процессора-в-памяти" с использованием ресурсов обработки информации только собственного кристалла, а также с использованием собственных ресурсов обработки информации и дополнительных ресурсов за счет других чипов, подключенных через интерфейс ввода/вывода и интерфейс внешней памяти.

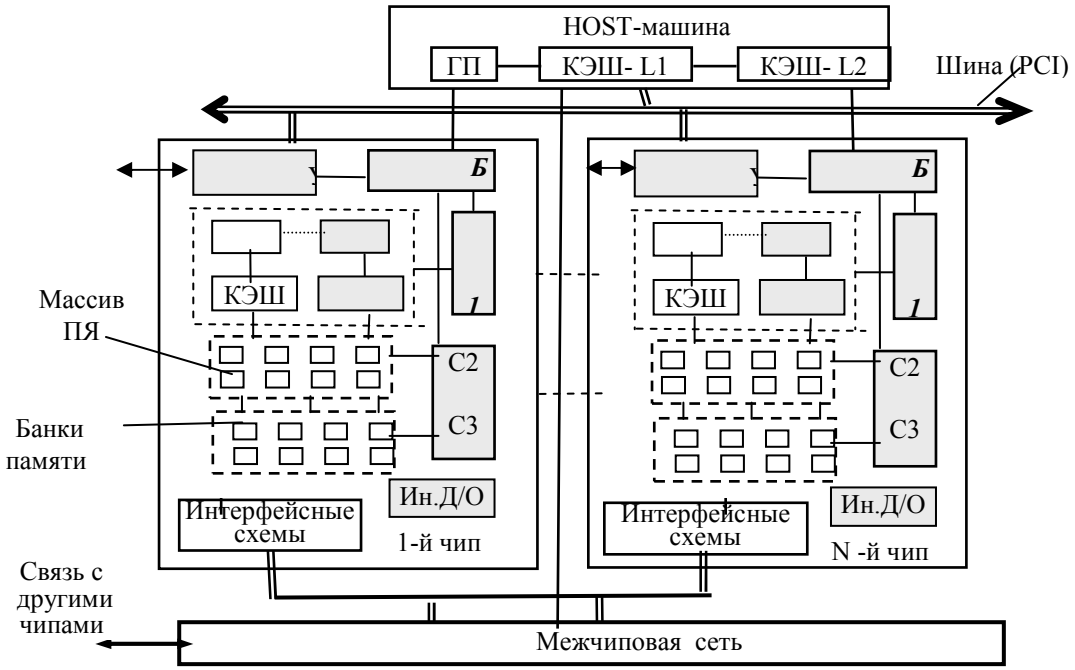

Рисунок 5 – Укрупненная структурная схема реконфигурируемой PIM-системы

Кроме того, управляющий пакет обеспечивает формирование управляющих сигналов и сигналов синхронизации для работы всех узлов системы памяти: управление работой КЭШ-памяти с учетом реализации механизмов когерентности; реализация режимов "Чтение/Запись" при обращении к памяти внутри чипа и за его пределами через интерфейс внешней памяти и др.

Преимущества предлагаемой PIM-системы по сравнению с известными PIM-системами прежде всего состоит в том, что в ней реализована возможность повышения производительности системы (от 2,5 до 10 раз) за счет исходной настройки с помощью селекторов С1 – С3 архитектуры на класс решаемых задач, а также её перестройки в процессе работы (в динамике). Это достигается путем использования оригинальных схемотехнических решений и соответствующих режимов работы БИС памяти, обеспечивающих: выбор оптимального ведущего процессора, выбор необходимого массива памяти для решения конкретной задачи с возможностью подключения дополнительной памяти, находящейся за пределами чипа; выбор оптимальной разрядности данных и количества обрабатываемых слов в пределах всей строки массива памяти, ориентируясь на соответствующие методы обработки и имеющиеся ресурсы для распараллеливания алгоритма решаемой задачи.

Кроме того, существенно снижена мощность, потребляемая одним чипом за счет исключения "холостых пробегов" процессорных ядер массива памяти, так как они загружены только полезной вычислительной работой. При этом сокращается количество одновременно передаваемых бит, так как передачи по линиям связи не всегда реализуются полной N-разрядной строкой (как в известных структурах подобного типа), а при необходимости произвольными кратными R-разрядными группами (например, частями строки разрядностью R, 2R,…, αR; $\alpha R \le N$). Это обеспечивает возможность более глубокого распараллеливания процесса обработки информации при тех же исходных ресурсах за счет повышения эффективности использования процессоров. Увеличивается также и серийность PIM-чипа и соответственно уменьшается стоимость, поскольку он из специализированного преобразован в чип широкого назначения за счет настройки его архитектуры и необходимых ресурсов на различные типы решаемых задач, в том числе – в части обработки графики.

Реконфигурируемая PIM-система с кольцевой модифицированной шиной. Для дальнейшего повышения эффективности применения PIM-систем в качестве графических ускорителей необходимо было решить следующие задачи: упростить коммутационную среду между компонентами системы,

сохранив при этом возможности распараллеливания алгоритма обработки графики и исключив ограничения по её масштабированию. Разместить на том же кристалле блок служебных функций по управлению памятью, сократив при этом время реализации наиболее важных его функций, в частности – время распределения реализуемого алгоритма по процессорам системы.

На основе анализа существующих PIM-систем и выявления их недостатков предложена оригинальная архитектурно-структурная организация системы такого класса с модифицированной кольцевой шиной, которая содержит на одном кристалле вместе со средствами обработки и хранения данных, набор блоков, которые реализуют служебные функции по управлению памятью и распределения программ обработки графики по процессорам системы. Блок-схема такой PIM-системы приведена на рис.6, которая защищена патентом Украины на изобретение [11].

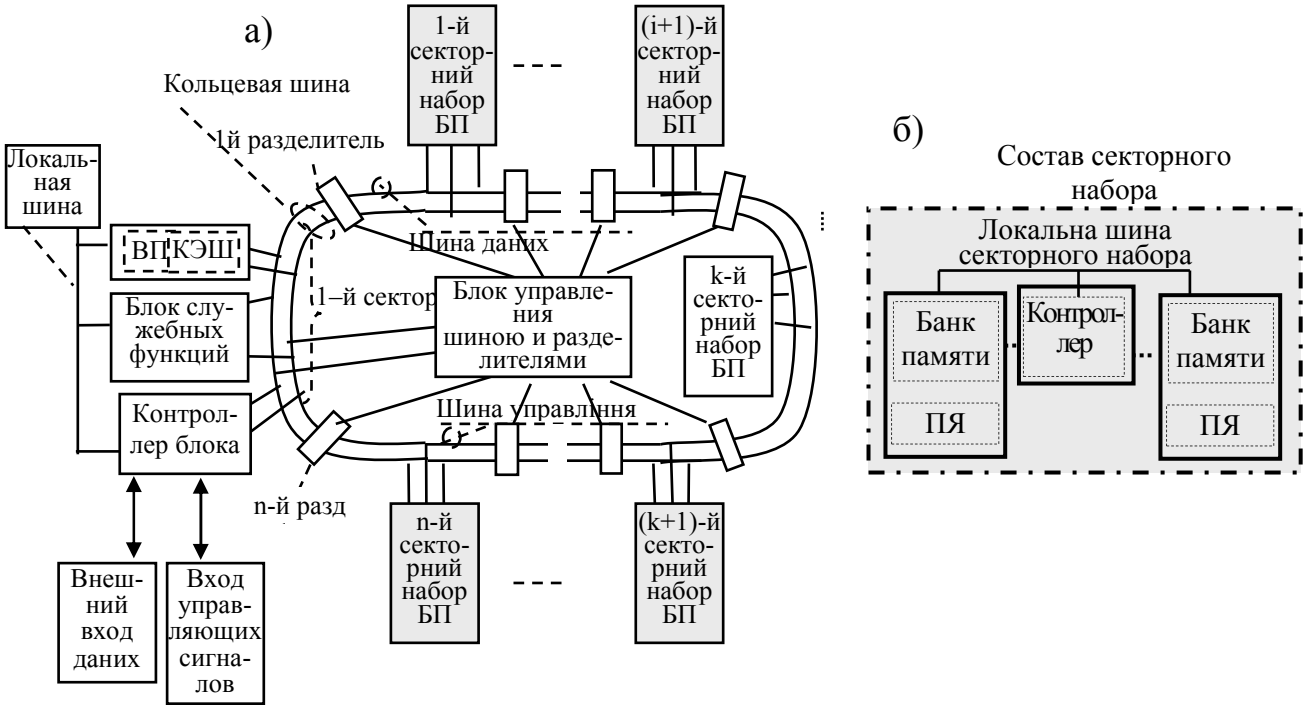

Рисунок 6 – PIM-система со средствами поддержки вычислительного процесса: а) общая структура PIM - системы з кольцевою шиною; б) состав секторного набора

Особенностью такой структуры является то, что в кольцевую шину введены секторные разделители, которые управляются специальным блоком. Два соседних секторных разделителя выделяют сектор кольцевой шины, длина которого меньше длины всей кольцевой шины в n раз, где n – количество секторов, при этом емкостная нагрузка на секторную часть кольцевой шины также уменьшается в n раз, и каждый банк памяти (БП) секторного набора (рис. 6,б) может общаться с секторными шинами быстрее. При широковещательной передаче сигналов по кольцевой шине емкостная нагрузка всей кольцевой шины также существенно уменьшена, поскольку каждый секторный разделитель при сквозном прохождении сигналов через него усиливает сигнал, возобновляя форму его импульса. Кроме того, возможность блокировки с помощью секторных разделителей каждого сектора кольцевой шины от влияния других секторов, позволяет реализовать параллельные вычисления в секторных наборах заблокированных секторов, что также существенно увеличивает производительность системы в целом и снимает ограничения по наращиванию параллельно работающих секторных наборов.

Блок служебных функций реализует предложенный модифицированный алгоритм обработки изображений, основанный на равномерном (с точки зрения загрузки процессоров) распределении распараллеленных фрагментов алгоритма по процессорам PIM-системы [12] с аппаратной его поддержкой, структурная схема которого защищена патентом Украины [13] и содержит: блок памяти микропрограмм, блок распределения фрагментов приложения, блок управления, блок памяти таблиц распределения, а также логический блок и блок буферной памяти вместе с интерфейсом с рабочей PIM-системой. Предложенная система уменьшает общее время распределения по процессорам алгоритма обработки изображения, а также расширяет сферу ее приложения для рабочих систем, выполненных на разных аппаратно-программных платформах.

Оценка эффективности применения реконфигурируемых графических ускорителей. Реконфигурируемые процессоры могут адаптироваться к изменению потребностей обработки графических изображений (например, задач обработки media-фильтров) для легко распараллеливаемых. алгоритмов. Интегрирование памяти большой емкости на том же самом кристалле, что и логические схемы, обеспечивает возможность получения более высокой ширины полосы памяти, чем была бы практически достигнута на традиционной архитектуре. Например, работая с 256-разрядной шиной от DRAM-банка (сегмента), можно достичь пропускную способность для каждого сегмента памяти до 4.8GB/s., размещенного на чипе. С архитектурной точки зрения, это устройство является весьма простым, однако теоретически может обеспечить высокую производительность (табл. 1) [14].

Таблица 1. Параметры производительности CIMA- архитектур для различных задач

Тип архитектуры	Время обработки, мс		
	Гистограмм	Фильтра	Пороговой величины
Традиционная RISC-архитектура	19	67	17
Процессор-в-памяти	71	0,6	0,1
Векторная DRAM-память (333МГц)	28	50	0,9
Одиночный мультипроцессорный чип с 4-мя процессорами (333МГц)	8	25	–
Merged CFD без быстрого переноса (16K LE's)	37	6	6,5
Merged CFD с быстрым переносом (16K LE's)	12	1,2	1,0

Для CIMA-системы с четырьмя банками памяти, общая ширина полосы обеспечивает пропускную способность до 19.2GB/s.. Это – больше, чем на порядок по сравнению с традиционными системами. Для сравнения, система оперативной памяти AlphaServer 8200 обеспечивает максимальную пропускную способность только 1.2GB/s. Таким образом, наиболее очевидное преимущество CIMA-архитектур по сравнению с другими типами архитектур при обработке изображения - увеличенная ширина полосы пропускания памяти и более низкая задержка сигнала при обращении к памяти, размещенной на чипе. Например, если на DRAM- чипе можно передавать одновременно 128 бит каждые 6 наносекунд, то данные полного буфера изображения емкостью 1МБ можно передавать через память меньше, чем за 1/2000 с. Конечно, это требует достаточной скорости процессора, чтобы не отставать от этой огромной ширины полосы. Оптимальная архитектура CIMA должна обеспечить высокий уровень вычислительного параллелизма, чтобы эксплуатировать полную ширину полосы пропускания памяти [7].

Технология реализации PIM – систем, как и других систем класса CIMA, позволяет основную оперативную память выполнить в непосредственной близости к элементарным процессорам, обеспечивая тем самым более низкую латентность памяти и значительно более широкий её интерфейс, чем в обычных микропроцессорах. Так разработанная за рубежом PIM-система VIRAM-1 имеет пиковую производительность 1.6 GFlop/s для данных разрядностью 32 бита и является чипом малой мощности, потребляя всего 2 ватта энергии.

В таблице 2 показаны различия в параметрах между PIM-системами типа VIRAM, DIVA, потоковым процессором Image, а также классической системой типа Power3, которая представляет собою 64-битовую реализацию процессора PowerPC с поддержкой выполнения восьми команд в цикле.

Таблица 2 – Различия в параметрах VIRAM, Image, DIVA и архитектуры Power3 [7].

Параметры	VIRAM	Imagine memory	DIVA (1 PIM)	Power 3
Пропускная способность по каналу процессор-память, ГВ/с	6,4	2,7	1,77	1,6
Пиковая производительность, Гфлопс/с	1,6	20	1,3	1,5
Пиковая производительность, Флопс/слово	1	30	1	3,75
Тактовая частота, МГц	200	600	166	375
Область (размеры) чипа	15x18 мм (270 мм^2)	12x12мм (144мм^2)	9,8x9,8 мм (96мм^2)	270мм^2
Ширина поддерживаемых данных, бит	64/32/16	42/16/8	64/32/16/8	64
Количество транзисторов, млн. шт.	130	21	55	16
Потребляемая мощность, Вт	2	4	1,6	33

Как большинство обычных суперскалярных архитектур, Power3 многократно использует КЭШ, чтобы уменьшить задержку основной памяти.

Согласно табл.2, Imagine имеет на порядок более высокую пиковую производительность, в то время как VIRAM – в два раза больше полосу пропускания памяти и потребляет половину мощности. При этом у VIRAM и DIVA имеется достаточная полоса пропускания, чтобы реализовать одну операцию при доступе к памяти, в то время как Image, требует 30 операций

Выводы

Таким образом, несмотря на совершенствования в технологии VLSI, которая обеспечила более высокие вычислительные возможности и большие блоки памяти, классическая архитектура, не в состоянии достаточно эффективно использовать эти достижения для эффективных параллельных вычислений из-за узких мест канала “память – процессор”, так как графическая обработка требует интенсивной коммуникации памяти процессора для передач данных, текстурирования, и буферизации. Архитектура PIM с малым временем доступа к памяти может увеличить производительность системы.

Применение реконфигурируемых архитектур существенно увеличивает производительность в десятки и сотни раз по сравнению с программной реализацией обработки графики и высокопроизводительным процессором за счет настройки архитектуры и структуры системы на размеры графических изображений и их сложность, а также за счет широкой шины (полосы пропускания) памяти. С этой точки зрения архитектурно-структурная организация КС с модифицированной кольцевой шиной, а также реконфигурируемой PIM-системы имеют обоснованную перспективу использования и дальнейшего развития вследствие особенностей их архитектурно-структурной организации. При этом в настоящее время все новые решения по созданию систем такого типа защищены патентами Украины.

Список литературы

1. Романюк О.Н. Класифікація графічних відеоадаптерів./ Романюк О.Н., Довгалюк Р.Ю.,.Олійник С.В.//Наукові праці ДонНТУ випуск 14(188). Серія “Інформатика, кібернетика та обчислювальна техніка”. С. 2011. – 215.

2..Додонов А.Г. Сравнительный анализ многопроцессорных систем формирования графических изображений / Додонов А.Г.,.Мельников А.Н, Резник Я.В.//The Sixth International Conference “INTERNET –EDUCATION - SCIENCE” Vinnytsia, Ukraine, October 7 –11, 2008. С. 369 – 373

3. Almudena Lindoso . Hardware Architectures for Image Processing Acceleration/ Almudena Lindoso and Luis Entrena// Электронный ресурс. Режим доступа: http://www.intechopen.com/books/image-processing/hardware-architectures-for-image-processing-acceleration. – Дата доступа: 17.10.2013.

4. Яковлев Ю.С. Однокристальные компьютерные системы высокой производительности. Особенности архитектурно- структурной организации и внутренних процессов: Монография / Ю.С. Яковлев. – Винница: ВНТУ, 2009. – 294 с.

5. Thinh M. Le. SIMD Processor Arrays for Image and Video Processing: A Review/ Thinh M. Le, Snelgrove W. M., and Panchanathan S..//Электронный ресурс. Режим доступа: http://www.thinhmle.com/imagesnus/SIMDProcessor ArraysforImageandVideoProcessingAReview.pdf. – Дата доступа: 17.10.2013.

6.. Tashev T. Design of Advanced Computer Architectures, Based on PIM - Processors in Memory/ Tashev T., Tashev S., Tasheva N.// Information technologies and control. – 2010. – №3. – С. 19 – 22

7. Oliker Leonid. Evaluation of Architectural Paradigms for Addressing the Processor-Memory Gap/ Oliker Leonid, Husbands Parry, Jacqueline Gorden Griem// Электронный ресурс. Режим доступа: http://www.escholarship.org/uc/item/5s59f02m#page-3. htm. – Время доступа: 18.10.2013

8. Jae Chul Cha. A PIM (Processor-In-Memory) for Computer Graphics: Data Partitioning and Placement Schemes/ Jae Chul Cha and Sandeep K. Gupta //World Academy of Science, Engineering and Technology 14 - 2008. С. –123 – 131. – Электронный ресурс. Режим доступа: http://www.waset.org/journals/waset/v14/v14-22.pdf.– Время доступа: 13.10.2013

9. Палагин А.В. Особенности подхода к выбору ПЛИС для проектирования PIM- систем / Палагин А.В. , Яковлев Ю.С. , Елисеева Е.В. // Математичні машини і системи. – 2012. – № 3. – С. 19–28.

10. Патент № 6259 Украина, МПК G06F13/00, G06F12/00. Система пам’яті з інтеграцією функцій зберегання та обробки інформації на одному кристалі /Сергієнко І. В., Кривонос Ю. Г., Палагін О. В., Коваль В. М., Яковлєв Ю. С., Тихонов Б. М.; Інститут кібернетики Імені В.М.Глушкова НАН України; 15.04.2005. Бюл. № 4. – 24с.

11. Пат. на винахід 99164 Україна, МПК G06F 15/16, G06F 13/42. Інтелектуальна розподілена система пам'яті з кільцевою шиною/ Палагін О.В., Яковлєв Ю. С., Тихонов Б. М., Єлісєєва О. В.; Інститут кібернетики ім. В.М. Глушкова НАН України; заявл. 16.07.2010; опубл. 25.07.2012, Бюл. № 14.– 21с.

12. Пат. 71719 Україна, МПК G06F 9/44, G06F 9/45. Спосіб розподілу програми користувача для комп'ютерної системи / Сергієнко І. В. Палагін О. В., Боюн В. П., Яковлєв Ю. С., Єлісєєва О. В.; Інститут кібернетики ім. В.М. Глушкова НАН України; заявл. 03.01.2012; опубл. 25.07.2012, Бюл. № 14.– 15с.

13 Пат. 73424 Україна, МПК G06F 9/44 (2006.01), G06F 9/45(2006.01). Система для розподілу програми користувача / Сергієнко І. В. Палагін О. В., Боюн В. П., Яковлєв Ю. С., Єлісєєва О. В.; Інститут кібернетики ім. В.М. Глушкова НАН України; заявл. 27.02.2012; опубл. 25.09.2012, Бюл. № 18.

14. Landis David. Evaluation of Computing in Memory Architectures for Digital Image Processing Applications / Landis David, Hulina Paul and Deno Scott, Roth Luke and Coraor Lee// Электронный ресурс. Режим доступа: http://design.ecs.psu.edu/SmartDIMM/papers/ICCD_CIMA.pdf. – Дата доступа: 16.10.2013

Стаття надійшла: 10.10.13.

Сведения об авторе

Яковлев Юрий Сергеевич – докт. техн. наук, зав. отделом, Институт кибернетики имени В.М. Глушкова Национальной академии наук Украины.