DOI: 10.22184/1992-4178.2017.170.9.104.111

РЕШЕНИЕ ЗАДАЧ МАШИННОГО ЗРЕНИЯ НА БАЗЕ ГЕТЕРОГЕННОЙ ПЛАТФОРМЫ ГРИФОН

П.Галаган[1], Л.Кузьминский, к.ф-м.н.[2], А.Сорокин[3]

УДК 004.93'1, 004.272.43
ВАК 05.13.00

Визуальный канал получения информации – один из наиболее информативных источников данных о характеристиках изучаемого объекта, идет ли речь о тактической обстановке применительно к боевой информационно-управляющей системе (БИУС) той или иной единицы военной техники или о контроле качества выполнения поверхностного монтажа при изготовлении радиоэлектронных изделий. Поэтому сегодня в самых разных сферах практической деятельности наблюдается все более широкое применение систем машинного зрения.

В статье приводятся материалы по организации эффективной системы машинного зрения с применением параллельно-конвейерной обработки данных на примере системы обработки видео высокого разрешения в режиме реального времени на платформе ГРИФОН.

ВВЕДЕНИЕ

Машинное зрение (machine vision) – это обширный прикладной раздел междисциплинарной теории компьютерного зрения (computer vision). Как инженерная дисциплина оно развивается на стыке нескольких областей, таких как компьютерное зрение, встраиваемые системы, базы данных, машинное обучение. Среди разработок, в которых технологии машинного зрения получили наибольшее распространение, следует назвать системы визуального контроля и управления, системы безопасности, системы виртуальной и дополненной реальности, системы управления технических средств высокой степени автономности – от пилотажно-навигационных комплексов беспилотных летательных аппаратов до роботизированных технологических установок производственного и иного назначения.

Для встраиваемых систем реального времени, использующих машинное зрение для распознавания объектов, особое значение приобретают производительность и скорость реакции. Производительность системы может быть оценена по количеству обрабатываемых в единицу времени видеокадров, скорость реакции – по временной задержке между поступлением на приемник видеокадра и моментом принятия решения по данным с него. Комплекс возможностей, которыми потенциально располагают системы машинного зрения, стимулирует разработчиков к возложению на них все более сложных функций, что влечет за собой повышение требований к параметрам таких систем. Соответственно, сохраняется актуальность задачи совершенствования аппаратно-программных средств для работы с высокоинтенсивными потоками видеоинформации.

[1] Компания ДОЛОМАНТ, руководитель департамента ключевых проектов.
[2] Компания ДОЛОМАНТ, ведущий инженер-программист.
[3] Компания ДОЛОМАНТ, заместитель генерального директора по планово-организационной работе.

ПАРАЛЛЕЛЬНАЯ ГЕТЕРОГЕННАЯ СИСТЕМА

Задачи компьютерного зрения предоставляют разработчикам большой простор для распараллеливания. Например, входящие в состав вычислителей графические модули могут параллельно обрабатывать данные из нескольких видеопотоков, накладывать на один и тот же кадр различные фильтры, искать в кадре независимо друг от друга объекты различных типов и др.

Структура потока данных в системе может существенно меняться на различных этапах обработки, от объемных структурно-разнородных данных в разнообразных нестандартных форматах (видеопотоков от камер высокого разрешения) до небольших пакетов данных (сжатых на видеокарте кадров). Выбрать наиболее эффективную архитектуру для обработки каждого типа потока данных позволяет гетерогенная система, состоящая из существенно различных функциональных узлов. Например, для реализации ряда специальных прикладных алгоритмов или предварительной обработки нестандартных данных целесообразно использовать вычислитель на базе ПЛИС, для стандартной обработки видеопотоков – вычислители на базе графических процессоров, для решения задач контроля и принятия решений – вычислитель с центральным процессором.

Таблица 1. Вычислительные модули гетерогенной платформы ГРИФОН

Наименование	Описание	Производитель	Внешний вид
CPC510	Модуль центрального процессора (Intel i7-3555LE 2,5 ГГц, 8 Гбайт ОЗУ DDR3L)	ЗАО "НПФ "ДОЛОМАНТ"	
FPU500	Модуль реконфигурируемого процессора на базе ПЛИС Xilinx Virtex-6 с ОЗУ емкостью 4 Гбайт	ЗАО "НПФ "ДОЛОМАНТ"	
VIM556-01	Модуль графического процессора (графическая карта NVIDIA Quadro K2100M, 2 Гбайт ОЗУ)	ЗАО "НПФ "ДОЛОМАНТ"	
KIC550	Модуль-носитель HDD-накопителя	ЗАО "НПФ "ДОЛОМАНТ"	
TB-FMCH-3GSDI2A	Мезонинный модуль ввода	Texas Instruments	
Компактная трансляционная камера Full-HD	Marshall CV360-CGB (Full HD 1 920×1 080p)	Marshall	

Разработанная в компании ЗАО "НПФ "ДОЛОМАНТ" высокопроизводительная гетерогенная вычислительная платформа (ВГВП) ГРИФОН [1] предназначена для решения задач с высокими требованиями к вычислительной мощности и большими объемами анализируемой информации. Она позволяет создавать высокопроизводительные БИУС, в том числе многоканальные системы обработки видео. В состав системы могут входить процессорные модули, графические ускорители, ускорители на основе ПЛИС, объединенные межмодульной шиной PCI Express. Для некоторых ресурсоемких задач такое аппаратное решение может оказаться наилучшим с точки зрения производительности, стоимости и гибкости [2].

Платформу ГРИФОН выгодно отличает от аналогов возможность построения на ее базе параллельно-конвейерной системы за счет поддержки между вычислителями соединений типа "точка – точка" через PCI Express-коммутатор. Богатый аппаратный состав платформы и гетерогенность ее вычислительной среды позволяют достаточно эффективно и быстро организовать параллельно-конвейерную обработку. Идея использования гетерогенных вычислительных конвейеров заключается в выстраивании процесса обработки данных в цепочку, на каждом этапе которой (участке конвейера) с данными работает вычислитель с оптимальной для этого этапа аппаратной архитектурой. Своевременная загрузка конвейера новыми данными без накладных расходов на их пересылку позволяет организовать одновременную и слаженную работу всех вычислительных модулей.

Механизм параллельно-конвейерной обработки является признанным классическим методом повышения быстродействия систем обработки данных, и если структура данных и алгоритм позволяют распараллеливать задачу, это почти всегда повышает эффективность такой обработки.

РЕШЕНИЕ ЗАДАЧИ КОМПЬЮТЕРНОГО ЗРЕНИЯ

Постановка задачи

Рассмотрим возможность организации параллельно-конвейерной обработки данных на платформе ГРИФОН на примере системы обработки видео высокого разрешения. Для решения этой задачи требуется:

- в режиме реального времени принимать данные от двух камер разрешением 1920×1080;
- провести предварительную обработку кадров при приеме;
- применить к видеопотокам алгоритмы фильтрации и компьютерного зрения (поиск лиц, детектор движения, фильтр Собеля);
- отобразить полученный результат на мониторах;
- сжать видео кодеком MPEG-4;
- записать в режиме реального времени сжатое видео на жесткий диск.

Состав вычислителя

Для решения поставленной задачи в состав гетерогенного вычислителя нами были включены (табл.1):

- модуль центрального процессора CPC510, работающий под управлением Linux Ubuntu 14.04;
- модуль ПЛИС FPU500 с мезонинным модулем ввода TB-FMCH-3GSDI2A;
- модуль графического процессора VIM556;
- модуль-носитель HDD-накопителя KIC550.

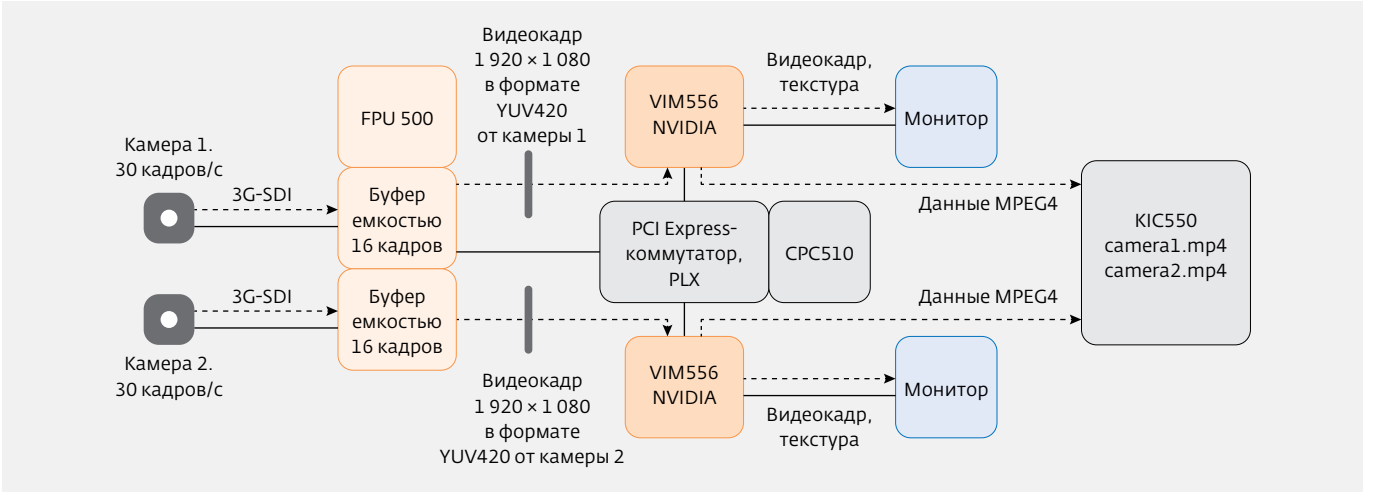

Рис.1. Общая схема системы обработки видео высокого разрешения на базе ГРИФОН. На схеме: 3G-SDI – цифровой видеоинтерфейс для передачи телевидения высокой четкости с прогрессивной разверткой потоком до 2970 Мбит/с посредством одного коаксиального кабеля; FPU500 – модуль реконфигурируемого процессора на базе ПЛИС Xilinx Virtex; VIM556 – модуль графического процессора; KIC550 – модуль-носитель HDD-накопителя

Организация взаимодействия между модулями вычислителя

Последовательность операций, которые требуется провести над видеопотоками, организована в виде независимо работающего конвейера; два видеопотока обрабатываются в отдельных, независимо работающих параллельных конвейерах (рис.1). Обработкой данных при этом занимаются модули FPU500 на базе ПЛИС и VIM556 на базе графического процессора. Задачей модуля центрального процессора CPC510 является только генерация управляющих команд, непосредственно в обработке данных он не задействован, что существенно снижает его загрузку, высвобождая ресурсы для выполнения других функций.

Каждый конвейер включает в себя:

- блок управления входными данными, реализованный на модуле ПЛИС FPU500;
- графическую видеокарту VIM556;
- набор управляющих программных потоков, выполняющихся на процессорном модуле CPC510.

Блок управления входными данными на ПЛИС написан на языке VHDL. В нем можно выделить следующие основные части: блок приема данных по протоколу 3G-SDI и их преобразования из формата YUV422 в формат YUV420; блок контроля и управления кольцевым буфером кадров; блок записи кадров в DDR-память модуля FPU500.

Реализацией алгоритмов компьютерного зрения в каждом видеопотоке занимаются вычислители VIM556, по одному на каждый поток. В их задачи входит проведение одной операции из списка: поиск лиц, детектирование движения, фильтрация Собеля. Результаты обработки видеоизображений вычислители сразу отображают на подключенных к ним мониторах, одновременно подготавливая кадр к отправке на жесткий диск путем его сжатия встроенным в видеокарту аппаратным видеокодеком H.264.

Управление конвейерами осуществляется приложением, выполняющимся на процессорном модуле CPC510. На обслуживание каждого конвейера в приложении выделено по два программных потока (нити), ответственных за контроль передачи данных и своевременное отображение кадров на графическом ускорителе.

Располагающийся на CPC510 коммутатор шины PCI Express Gen2 Switch PLX8624 и входящий в комплект поставки платформы ГРИФОН специальный драйвер обеспечивают устойчивую связь между всеми модулями системы.

В данном примере механизмы прямого межмодульного взаимодействия в режиме "каждый с каждым"

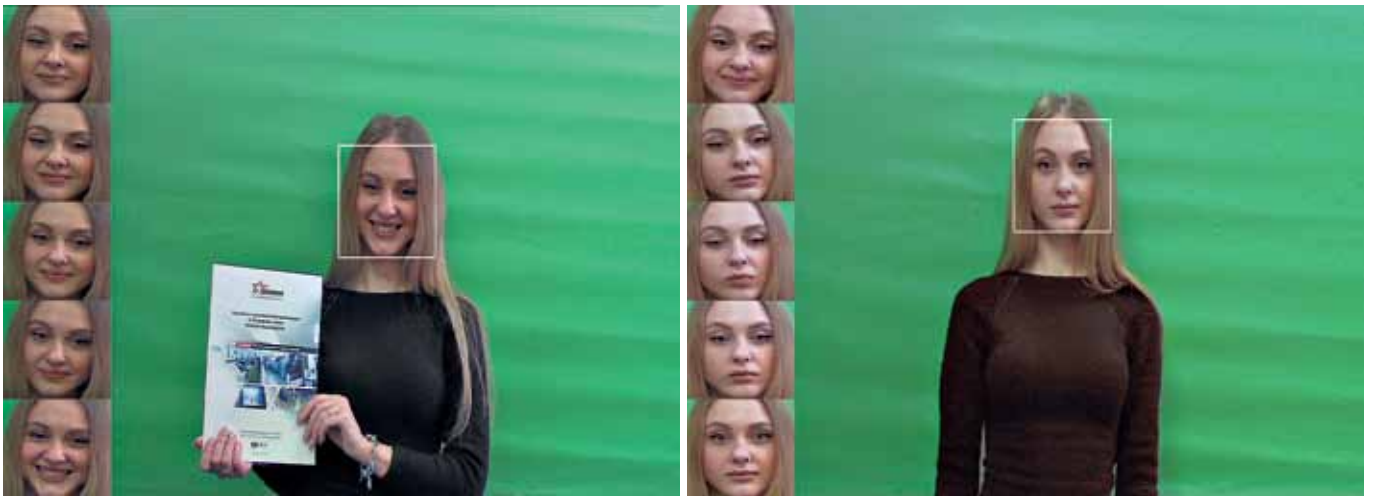

Рис.2. Поиск лиц (кадры из транслируемого видеопотока)

позволяют высвободить ресурсы центрального процессора и снизить нагрузку на основной транспортный интерконнект по шине PCIe, что на практике дает возможность минимизировать время обработки кадра всем конвейером.

Рассмотрим подробнее последовательности операций на основных этапах каждого конвейера.

Входной кадр разрешением 1920×1080 поступает через мезонин TB-FMCH-3GSDI2A на вход блока приема данных ПЛИС. В блоке приема изображение преобразуется "на лету" из формата YUV422 в более легковесный YUV420 и размещается в выделенной области DDR-памяти модуля FPU500, организованной в виде кольцевого буфера емкостью 16 кадров по 3 Мбайт. DDR-память модуля FPU500 доступна для чтения и записи через PCI Express всем вычислителям системы. Данные поступают в кольцевые буферы со скоростью 30 кадров в секунду. Отметим, что производительность системы такова, что кадры вычитываются из кольцевых буферов быстрее, чем они поступают в систему, и в каждом кольцевом буфере в произвольный момент времени находится не более одного кадра.

Рис.3. Детектирование движения, кадр из транслируемого видеопотока. Движущиеся области изображения детектируются видеокартой, на них накладываются квадраты

Записав кадр размером 3 Мбайт в DDR, FPU500 генерирует прерывание на шине, после чего переходит к ожиданию новых видеоданных. Весь алгоритм первичной обработки занимает не более 16 мкс.

Прерывание, полученное по PCI Express от FPU500, обрабатывается на CPC510 управляющим программным потоком, который выдает команду на копирование кадра из DDR-памяти FPU500 напрямую на VIM556 через коммутатор PLX8624. Получив новое изображение, видеокарта производит на нем одну из следующих операций на выбор: поиск лиц (рис.2), детектирование движения (рис.3) или фильтрацию Собеля (рис.4).

Обработка изображений выполнена на платформе параллельных вычислений CUDA с использованием функциональности библиотеки компьютерного зрения OpenCV: координаты лиц определяются методом Виолы – Джонса на основе каскадов Хаара [3, 4], при детектировании движения используются результаты выполнения алгоритма выделения фонового изображения с помощью распределений Гаусса [5], алгоритм выделения границ основывается на результатах применения к изображению оператора Собеля.

Результат обработки сразу отображается на подключенном к видеокарте мониторе и подвергается сжатию с помощью встроенного в VIM556 кодека H.264. Результат сжатия записывается в видеофайл в формате MPEG-4 на жестком диске модуля KIC550.

Несмотря на широкие возможности, которые предлагает библиотека OpenCV, для вывода кадров

Рис.4. Фильтрация Собеля, пример транслируемого видеопотока

с видеокарты сразу на дисплей применяются библиотеки OpenGL, GLEW и XLib. Кадры размещаются в областях памяти видеокарты типа "текстура", затем отрисовываются шейдерами на дисплее. Попытки использовать функции OpenCV для отображения приводили к излишним пересылкам кадров от VIM556 к CPC510 и обратно, что самым негативным образом сказывалось на производительности системы. По той же причине на CUDA пришлось реализовать функции рисования некоторых графических примитивов (прямоугольников). Контроль передаваемого по шине PCI Express трафика удобно производить с помощью PLX SDK, наглядно показывающего количество переданных и полученных байтов каждым устройством сети, а также скорости обмена.

Для сжатия видео встроенным в видеокарту кодеком используется NVIDIA Hardware Encoder SDK. Работа с кодеком построена таким образом, что его входные буферы, предназначенные для загрузки кадров, располагаются в локальной оперативной памяти VIM556 (рис.5). Любая излишняя пересылка данных по PCI Express, нарушающая принцип работы построенного конвейера, сразу приводила к простаиванию его элементов и резкому увеличению времени обработки кадра всей системой.

ПРОИЗВОДИТЕЛЬНОСТЬ

Оценим основные характеристики построенных конвейеров: конвейерную задержку, пропускную способность, уровень загрузки ЦП.

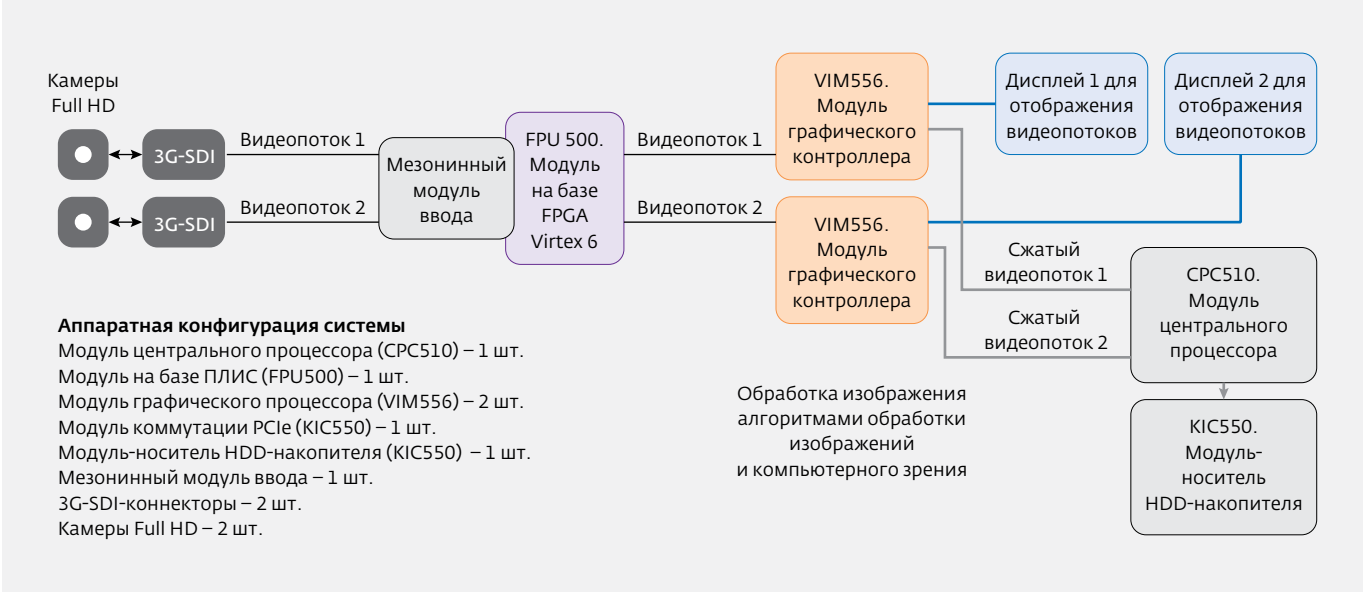

Рис.5. Параллельно-конвейерная обработка данных в системе обработки видео высокого разрешения в режиме реального времени, построенной на базе ВГВП ГРИФОН

Таблица 2. Продолжительность основных этапов цикла обработки кадра

Отображение и сжатие кадра с механизмом P2P	Передача кадра от FPU500 к VIM556	12 мс	16 мс
	Сжатие и сохранение кадра видеокодеком	4 мс	
Отображение и сжатие кадра без механизма P2P	Передача кадра от FPU500 к CPC510	12 мс	28 мс
	Передача кадра от CPC510 к VIM556	12 мс	
	Сжатие и сохранение кадра видеокодеком	4 мс	

Оценка конвейерной задержки

В табл.2 показаны длительности основных этапов цикла обработки кадра с использованием механизма "точка – точка" (P2P) и без его использования. Оценки были получены путем измерения длительности выполнения операций в управляющих потоках на процессорном модуле CPC510. Из приведенных данных видно, что реализованный в ГРИФОН механизм межмодульного взаимодействия позволяет значительно сократить величину конвейерной задержки. Действительно, при прямом обмене данными отпадает необходимость использовать процессорный модуль в качестве промежуточного звена передачи. В реальности выигрыш от применяемого механизма "точка – точка" еще более значителен, так как приведенные в таблице данные для режима "без PCIe P2P" не учитывают дополнительные временные затраты на пробуждение нитей на ЦП.

Малая величина задержки между моментом получения кадра 1920×1080 и его отображением на мониторе – менее 20 мс – подтверждает возможность построения на основе ГРИФОН систем видеотрансляции реального времени.

Оценка пропускной способности

Для оценки загруженности внутренней шины PCI Express использовался программный инструмент PLX SDK, показывающий потоки данных, проходящих через коммутатор PLX8624. Результаты мониторинга полностью соответствуют расчетным: из табл.3 видно, что исходящие от FPU500 видеопотоки объемом 89 Мбайт/с каждый поступают на соответствующие им графические модули VIM556. Размер видеопотока согласуется с размером кадров (3 Мбайт) и скоростью их выдачи (30 кадров/с). После сжатия кадры направляются на ЦП, что подтверждается наличием небольших потоков данных от графических ускорителей к ЦП.

Для сравнения в табл.4 приведены объемы потоков данных при работе ВГВП без механизма "точка – точка". При отсутствии возможности прямого межмодульного обмена видеокадры сначала попадают на процессорный модуль и лишь затем перенаправляются на графические ускорители.

Общая загрузка шины PCI Express не превышает 10% от максимально возможного значения.

Загрузка центрального процессора

При решении задачи обработки видео с помощью построенного конвейера центральному процессору необходимо только координировать работу входящих в состав ГРИФОН элементов – непосредственной обработкой данных CPC510 не занимается. В его функции входят выдача управляющих команд модулям на прием/передачу данных, управление кодеком NVIDIA, управление выводом изображения на мониторы видеокарт, а также общий контроль работоспособности системы.

Таблица 3. Объемы потоков данных, циркулирующих между вычислителями платформы ГРИФОН при использовании механизма "точка – точка"

Модуль	Входящий поток данных, Мбайт/с	Исходящий поток данных, Мбайт/с
FPU500	–	178
VIM556 N1	89	1
VIM556 N2	89	1
CPC510	2	0,7

Таблица 4. Объемы потоков данных, циркулирующих между вычислителями платформы ГРИФОН без использования механизма "точка – точка"

Модуль	Входящий поток данных, Мбайт/с	Исходящий поток данных, Мбайт/с
FPU500	–	178
VIM556 N1	89	1
VIM556 N2	89	1
CPC510	180	178,7

Таблица 5. Загрузка центрального процессора

Режим работы системы	Загрузка процессорной платы CPC510, %
Трансляция и сжатие видео при наличии в системе только одного видеопотока	4,5
Трансляция и сжатие видео при наличии в системе двух видеопотоков	12,5
Трансляция, поиск лиц и сжатие видео в обоих видеопотоках	25

Оценки загрузки центрального процессора в различных режимах мы проводили с помощью приложения htop, результаты измерений показаны в табл.5.

ЗАКЛЮЧЕНИЕ

Преимущества использования гетерогенных конфигураций для решения ряда ресурсоемких прикладных задач неоспоримы, и расширение их применения является сегодня одним из трендов развития вычислительных систем. При этом оценка характеристик производительности систем с гетерогенной вычислительной средой является пока нетривиальной задачей ввиду отсутствия готовых универсальных нагрузочных тестов и разнообразия способов решения прикладной задачи в гетерогенной вычислительной системе.

Продемонстрированный пример позволяет оценить наиболее критичные с точки зрения аспектов быстродействия и производительности характеристики гетерогенной системы при организации параллельно-конвейерной обработки данных в условиях высокой нагрузки. Так, разработанное для гетерогенной платформы ГРИФОН тестовое программное обеспечение позволило оценить ряд ключевых характеристик: конвейерную задержку, пропускную способность и загрузку центрального процессора в условиях достаточно серьезной нагрузки.

Полученные результаты решения задачи обработки потокового видео высокого разрешения подтверждают на практике эффективность реализованных в платформе ГРИФОН подходов к построению параллельно-конвейерной обработки в гетерогенной среде и наглядно демонстрируют ее основные преимущества:

- задействование каждого из вычислителей на своем участке конвейера, где он обрабатывает только те данные, для которых его архитектура оптимальна;
- параллельная работа различных звеньев цепи вычислительного конвейера;
- минимизация конвейерной задержки за счет межмодульного взаимодействия в режиме "каждый с каждым", или "точка – точка";
- разгрузка основного транспортного интерконнекта;
- существенное снижение нагрузки на центральный процессор и экономия его ресурсов для решения других задач.

Следует отметить, что выстроенные конвейерные цепочки поддерживают прямое масштабирование задачи – при необходимости обработки дополнительных видеопотоков к системе подключаются дополнительные звенья вычислительного конвейера – вычислители FPU500 и VIM556. При этом полученные конвейеры остаются не связанными между собой и работают параллельно, что определяет независимость значения конвейерной задержки системы для каждого потока. Уровень загруженности центрального процессора с увеличением числа видеопотоков и соответствующим ростом суммарного объема данных, обрабатываемых системой, возрастает линейно.

Разработанная в ЗАО "НПФ "ДОЛОМАНТ" высокопроизводительная гетерогенная вычислительная платформа ГРИФОН позволяет строить и эффективно применять гетерогенные вычислительные конфигурации не только для систем машинного зрения, но и для самого широкого спектра прикладных задач, в том числе для создания подсистем БИУС, вне зависимости от предъявляемых требований к надежности и производительности.

ЛИТЕРАТУРА

1. **Галаган П.** Платформа ГРИФОН для решения задач встраиваемых систем специального назначения // Современные технологии автоматизации. 2015. № 4.
2. **Alawieh M., Kasparek M., Franke N., Hupfer J.** A High Perfomance FPGA-GPU-CPU Platform for a Real-Time Locating System // 23rd European Signal Processing Conference (EUSIPCO). Fraunhofer Institute for Integrated Circuits IIS, Germany, 2015.
3. **Viola P, Jones M.J.** Rapid Object Detection using a Boosted Cascade of Simple Features // Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR2001). 2001.
4. **Viola P, Jones M.J.** Robust real-time face detection // International Journal of Computer Vision. 2004. Vol. 57. No. 2.
5. **KaewTraKulPong P., Bowden R.** An Improved Adaptive Background Mixture Model for Real-time Tracking with Shadow Detection // In Proc. 2nd European Workshop on Advanced Video Based Surveillance Systems. Sept 2001.