УДК 004.272.50

ПАРАЛЛЕЛЬНО-КОНВЕЙЕРНАЯ ОБРАБОТКА ИНФОРМАЦИИ В ГЕТЕРОГЕННОЙ ВЫСОКОПРОИЗВОДИТЕЛЬНОЙ ВЫЧИСЛИТЕЛЬНОЙ ПЛАТФОРМЕ (ВГВП)

A PARALLEL-PIPELINED INFORMATION IN A HETEROGENEOUS HPEC PLATFORMS (VGVP)

П.В. Галаган, С.М. Чудинов
P.V. Galagan, S.M. Chudinov

АО «НИИВК им. М.А. Карцева», Россия, 117437, Москва, ул. Профсоюзная, 108

M.A. Kartsev Scientific and Research Institute of Computing Systems, 108 Profsoyuznaya St, Moscow, 117437, Russia

E-mail: galagan@fastwel.ru, chud35@ yandex.ru

Аннотация: В статье приводятся материалы по эффективному применению вычислительных возможностей, организации параллельно-конвейерной обработки информации ВГВП на примере системы обработки видео высокого разрешения в режиме реального времени. Рассмотрены полученные экспериментально значения параметров оценки конвейерной задержки, при обработки информации, и оценка пропускной способности ВГВП.

Resume: This article contains material on the effective use of computing power, the organization of parallel-pipelined data VGVP video processing system an example of a high-resolution real-time. Examined experimentally obtained values conveyor delay estimation parameters for information processing and evaluation capacity VGVP.

Ключевые слова: параллельно-конвейерная обработка данных, гетерогенность вычислительной среды отечественная высокопроизводительная гетерогенная вычислительная платформа, компьютерное зрение, машинное зрение, результаты экспериментальных значений параметров обработки.

Keywords: parallel-pipelined data processing, heterogeneous computing environments domestic high-performance heterogeneous computing platform, computer vision, machine vision, the results of the experimental values of the processing parameters.

В последнее время в России все большее значение уделяется проблеме импортозамещения в области информационных технологий и наукоемкой продукции в виде вычислительной техники. В связи с этим особое внимание направлено на создание отечественных образцов вычислительной техники, не уступающих по характеристикам зарубежным аналогам.Один из вариантов решения проблем импортозамещения лежит в разработке отечественной высокопроизводительных гетерогенных реконфигурируемых вычислительных платформ, в которых в составе одного блока можно использовать модули с разными архитектурами в различных конфигурациях. В такие вычислительные платформы могут входить микропроцессоры общего назначения (x86, Эльбрус, Байкал), графические процессоры, вычислительные модули на базе программируемых логических интегральных схем (ПЛИС). Создание проблемно-ориентированной конфигурации на базе такой платформы достигается за счет выбора и установки в вычислительную платформу необходимого набора модулей, исходя из максимальной эффективности выполнения алгоритмов. Применение модулей на базе отечественных и зарубежных процессоров в рамках одной вычислительной платформы определяет концепцию постепенного импортозамещения, следование которой позволит не только создавать аппаратуру современного уровня уже сейчас, но и стимулирует разработку отечественной элементной базы, аналоги которой на данный момент отсутствуют. В рамках подхода постепенного импортозамещения в АО «НИИВК им. М.А.

Карцева»проводятся работы по созданию многопроцессорной гетерогенной вычислительной платформы (далее - МВП) с разнородной архитектурой, направленной на обработку больших объемов информации в том числе, видео высокого разрешения в режиме реального времени [Чудинов С.М., 2016, Галаган П.В., 2016].

Следует отметить, что машинное зрение (machinevision) – это обширный прикладной раздел междисциплинарной теории компьютерного зрения (computervision), представляющий существенный потенциал для встраиваемых систем. Машинное зрения как инженерная дисциплина находится на стыке нескольких областей, таких как компьютерное зрение, встраиваемые системы, базы данных, машинное обучение. Среди многочисленных направлений применения наиболее обширные внедрения наблюдаются в области промышленных и военных применений по следующим направлениям: системы визуального контроля и управления; системы безопасности; системы виртуальной и дополненной реальности; технические средства высокой степени автономности - от пилотажно-навигационных подсистем БИУС и до полностью автономных роботизированных технических средств [Головкин Б.А., 1980, Головастов А., 2010].

Для подобных систем характерно наличие нескольких потоков структурно-разнородных данных (в первую очередь это видеопотоки от камеры высокого разрешения), необходимость приема данных в нестандартных форматах, необходимость максимизации быстродействия для отработки сценариев по предназначению системы в режиме реального времени.

Для обработки каждогоиз потоков данных целесообразно использовать ту архитектуру, которая будет эффективнее при обработке каждого из потоков данных. Например, для реализации ряда специальных прикладных алгоритмов или предварительной обработки нестандартных данных целесообразно использовать вычислитель на базе ПЛИС, для обработки интенсивных потоков видео – вычислители на базе графических процессоров, для решения задач контроля и принятия решений – вычислитель центрального процессора, и т.д.

Отечественная высокопроизводительная гетерогенная вычислительная платформа (ВГВП) позволяет строить и эффективно применять гетерогенные конфигурации. Выбор конкретной гетерогенной конфигурации обусловлен комплексом исходных технических требований, типом данных и режимов их обработки.

На базе ВГВП представляется возможным осуществлять конвейерную обработку данных с применением гетерогенной архитектуры. Идея использования гетерогенных вычислительных конвейеров заключается в том, чтобы на каждом этапе последовательной обработки (участке конвейера)обработчик на базе оптимальной для работы с конкретным типом данных архитектурой, выполнив свою работу, передавал бы результат для дальнейшей обработки на следующий участок конвейера для обработки вычислителем – обработчиком другой архитектуры, одновременно принимая новый объем входных данных для следующей итерации цикла конвейерной обработки [Головкин Б.А., 1980].

При этом большинство задач машинного зрения хорошо поддаются распараллеливанию при обработке данных. Например, каждая видеокамера передает один видеопоток, если таких камер несколько, то для повышения общего быстродействия весьма эффективно разделить конвейер на участки параллельной обработки, где это возможно, получив прирост производительности.

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

Так, гетерогенность, архитектурные решения и программные механизмы взаимодействия модулей различной архитектуры позволяют эффективно применять ВГВП для гетерогенной параллельно-конвейерной обработки данных.

Рассмотрим возможности ВГВП для организации параллельно-конвейерной обработки данных на примере системы обработки видео высокого разрешения в режиме реального времени.

Постановку задачи можно кратко сформулировать следующим образом: требуется в режиме реального времени принять данные от четырех камер высокого разрешения,провести предварительную обработку,передать данные на отдельныйобработчик для отработки прикладных алгоритмов компьютерного зрения с дальнейшей передачей результата для принятия решения центральным процессором.

Исходя из постановки данной задачи был сконфигурирован аппаратный состав базового вычислительного блока – гетерогенного вычислителя на базе ВГВП – Табл. 1., а дополнительные аппаратные средства представлены в Табл. 2.

Таблица 1
Table 1

Аппаратный состав гетерогенного вычислителя обработки видео высокого разрешения на базе ВГВП
The composition of the heterogeneous hardware calculator processing high-definition video on the basis of VGVP

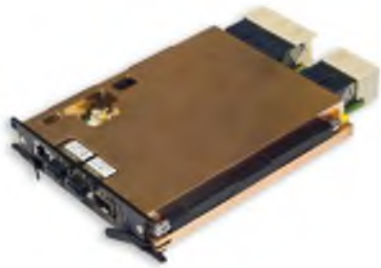
Наименование	Описание	Внешний вид	Количество
CPC512	Модуль центрального процессора (может использоваться в микропроцессорах с архитектурой Эльбрус)		1 шт.
FPU500	Модуль ПЛИС		1 шт.
VIM556	Модуль графического процессора		4 шт.
KIC551	Модуль коммутации PCIe		1 шт.
KIC550	Модуль-носитель HDD-накопителя		1 шт.
MIC2003	Мезонинный модуль ввода		1 шт.

Таблица 2
Table 2

Дополнительные аппаратные средства
Additional Hardware

Наименование	Количество
Камеры full-hd	4 шт.
3G-SDI-коннекторы	4 шт.
Мониторы	4 шт.

На рис. 1 представлена схема параллельно-конвейерной обработки данных на базе ВГВП.

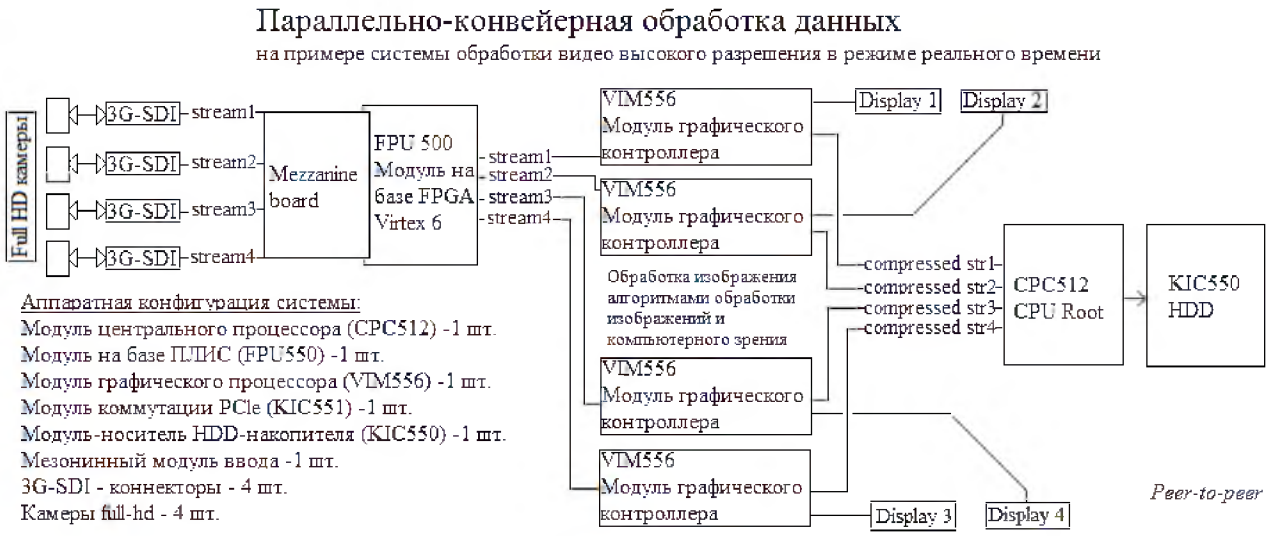


Рис. 1. Параллельно-конвейерная обработка данных на примере системы обработки видео высокого разрешения в режиме реального времени, построенной на базе ВГВП
Fig. 1. A parallel-pipelined data processing of high resolution on the example of the video processing system in real time, based on the constructed VGVP

В статье рассмотрены этапы работы конвейера на конкретном примере.

Для ввода данных в вычислительный контур сразу от нескольких камер по стандарту 3G-SDI используется мезонинный субмодуль MIC2003, монтированный на вычислительный модуль FPU500, что позволяет, во-первых осуществить ввод данных через нестандартные интерфейсы, а во-вторых осуществить ввод«напрямую» (без транзита по общей транспортной шине PCIe) на модуль FPU500 для дальнейшей обработки.

Поступающие на модульFPU500 кадры видеоизображения разрешением 1920x1080 в формате 3G-SDI, далее декодируются и сохраняются в памяти модуля в формате YUV420, организованной в виде кольцевого буфера емкостью 16 кадров для каждой камеры. При очередной записи кадра модуль генерирует прерывание на шине PCIe, по которому управляющая программа на модуле центрального процессора CPC512 выдает команду на копирование кадра из памяти FPU500 в память модуля графического процессора VIM556 по линиям шины PCIe.Один модуль FPU500 может одновременно обслуживать видеопотоки от 4-х видеокамер.

На модуле графического процессора VIM556 в режиме реального времени средствами CUDA и компонентами библиотеки OpenCV отрабатываются нужные прикладные алгоритмы: поиск лиц(рис. 2), детектирование движения(рис. 3), дополнительная фильтрация (рис 4). Далее средствами библиотек OpenGL и XLib прошедший обработку на VIM556 кадр без передачи по PCIев режиме реального времени отображается на подключенном к модулю VIM556 мониторе.

Рис. 2. Поиск лиц. Пример выведенного на монитор кадра из транслируемого видеопотока
Fig. 2. Search for individuals. Example outputted to the monitor frame of broadcast video

Пояснение 1: Поиск лиц в кадре производится на видеокарте с помощью объекта класса cv::cuda::CascadeClassifier библиотеки OpenCV. Функция поиска лиц в OpenCV – синхронная операция, занимающая порядка 20 мс, поэтому она запускается в отдельном потоке CPU, чтобы не замедлять отображение кадров. Обнаружив объект, программа выделит его местоположение в кадре белым прямоугольником и плавно выдвинет найденное изображение в левую часть экрана. Для снижения времени поиска кадр сжимается в 4 раза.

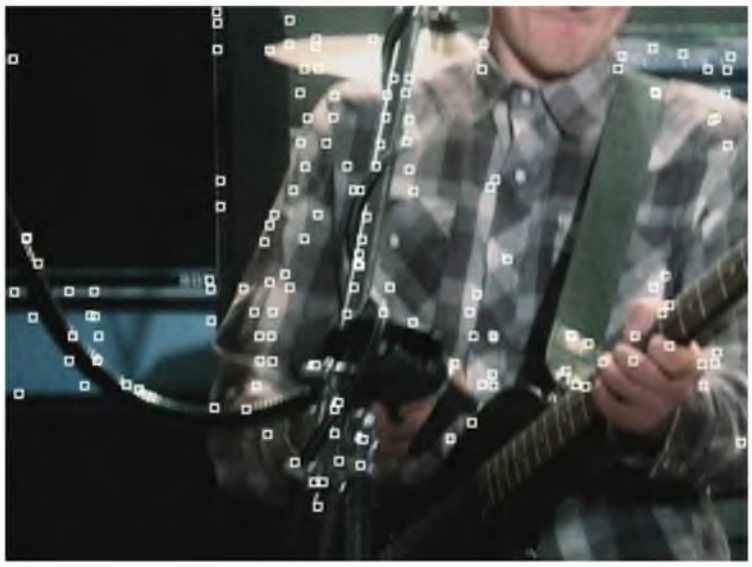


Рис. 3. Детектирование движения. Пример выведенного на монитор кадра из транслируемого видеопотока
Fig. 3. Motion Detection. Example outputted to the monitor frame of broadcast video

Пояснение 2: В основе процедуры поиска движения лежит объект класса cv::cuda::BackgroundSubtractorMOG библиотеки OpenCV, который работает с памятью видеокарты и вычисляет "опорное" фоновое изображение по последним полученным N кадрам. Вычитая фон из каждого нового кадра можно получить маску движения. Полученная маска разбивается примерно на 500 частей, в каждой из которых с помощью CUDA проводится фильтрация крупных движущихся участков. Используя найденные координаты движущихся объектов на оригинальное изображение накладываются белые квадратики.

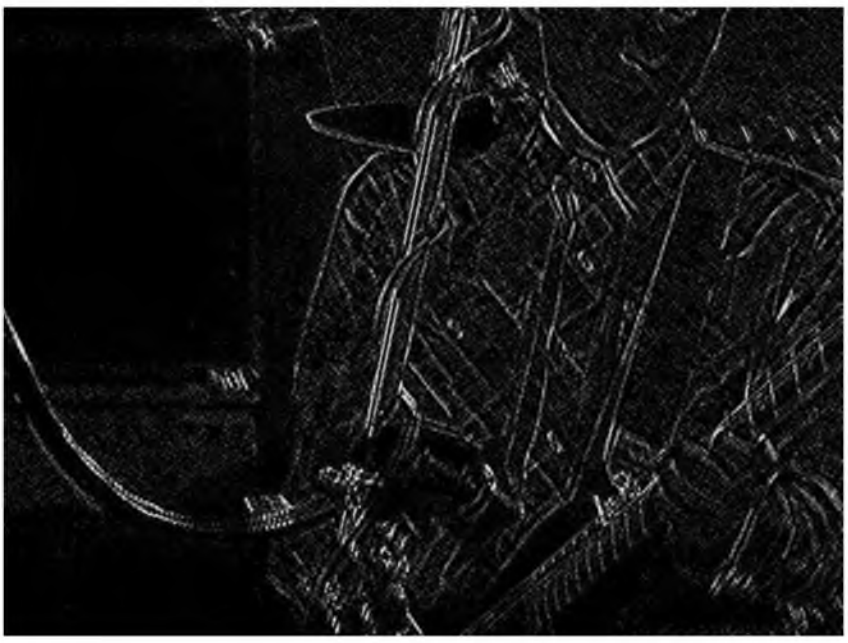


Рис. 4. Фильтрация Собеля. Пример транслируемого видеопотока
Fig. 4. Filter Sobel. An example of the broadcast video stream

Пояснения 3: Фильтрация Собеля выполняется с помощью объекта cv::cuda::SobelFilter библиотеки OpenCV.

Фильтр выделяет белым цветом границы областей различной яркости.

Процесс такой обработки идет по 4 параллельным гетерогенным конвейерам по количеству входных потоков данных – в данном примере задействованы 4 камеры. При этом основная нагрузка делегируется для выполнения средствами модуля на базе ПЛИС FPU500 и модулей графического процессора VIM556. Модуль центрального процессора CPC512 не задействован непосредственно в обработке данных, а выдает только управляющие команды, что существенно снижает его загрузку, высвобождая ресурсы для выполнения другого функционала.

Действительно, следует особо отметить, что одним из важных преимуществ ВГВП является поддержка режима «каждый с каждым» (peer-to-peer/P2P) при межмодульном взаимодействии по высокоскоростной шине PCIe. Это позволяет осуществлять пересылку данных от одного вычислительного модуля другому без участия центрального процессора.

В данном примере механизмы прямого межмодульного взаимодействия в режиме «каждый с каждым» позволяют высвободить ресурсы центрального процессора и снизить нагрузку на основной транспортный интерконнект по шине PCIe, что на практике позволяет минимизировать время обработки кадра по конвейеру.

Важным параметром ВГВП при разработке является производительность

К основным характеристикам производительности конвейера можно отнести следующие параметры:

- конвейерная задержка;
- пропускная способность;
- уровень загрузки ЦП.

В статье рассмотрены полученные экспериментально значения этих параметров на базе представленной системы.

1. Оценка конвейерной задержки

В таблице 3 показаны длительности основных этапов цикла обработки кадра как вместе, так и без механизма “каждый с каждым”. Из приведенных данных видно, что реализованный в “Грифон” механизм межмодульного взаимодействия позволяет значительно сократить величину конвейерной задержки. На самом деле выигрыш от применяемого механизма «каждый с каждым» еще более значителен, так как приведенные в таблице данные для режима “без PCIeP2P” не учитывают дополнительные временные затраты на пробуждение нити на CPU.

Таблица 3
Table 3

Длительность основных этапов цикла обработки кадра
Frame duration basic stages of the processing cycle

Отображение и сжатие кадра с PCIeP2P	Передача кадра от FPU500 к VIM556	12 мс	16 мс
	Конвейер видеокодека NVIDIA	4 мс	
Отображение и сжатие кадра без PCIeP2P	Передача кадра от FPU500 к VIM556	12 мс	28 мс
	Передача кадра от CPC512 к VIM556	12 мс	
	Конвейер видеокодека NVIDIA	4 мс	

2. Оценка пропускной способности

В представленном примере модульFPU500 готовит кадры объемом 3110400 байт для VIM556 от нескольких камер, например, 2-х камер, по 30 кадров в секунду. Общий объем видеоданных, поступающих в систему по PCI-Express, составляет 178 MB/s. На каждую видеокарту поступает половина от указанного объема. С каждой из 2-х видеокарт сжатые кадры отправляются на CPU в объеме 1 MB/s (таблица 4).

Таблица 4
Table 4

Объем видео данных
Volume of video data

Модуль	Входящий поток данных, MB/s	Исходящий поток данных, MB/s
FPU500		178
VIM556 N1	89	
VIM556 N2	89	
CPC512	2	

Для сравнения в таблице 5 приведены объемы потоков данных при работе стенда без механизма “каждый с каждым”.

Таблица 5
Table 5

Объемы потоков данных
Volumes of data flows

Модуль	Входящий поток данных, MB/s	Исходящий поток данных, MB/s
FPU500		178
VIM556 N1	89	
VIM556 N2	89	
CPC512	180	178

3. Загрузка центрального процессора

В задачи центрального процессора (ЦП) входят выдача управляющих команд модулям на прием/передачу данных, управление кодеком NVIDIA при сжатии видео в формат MPEG4 на видеокарте, управление выводом изображения на мониторы видеокарт.

Результаты оценки загрузки ЦП в различных режимах проведены с помощью приложения htopи показаны в таблице 6.

Таблица 6
Table 6

Результаты загрузки ЦП в различных режимах
Results of CPU usage in different modes

Режим работы стенда	Загрузка процессорной платы CPC512
Трансляция и сжатие видео от 1-й видеокамеры	Одно из 4-х ядер загружено на 36%
Трансляция и сжатие видео от 2-х видеокамер	Одно из 4-х ядер загружено на 50%
Трансляция, поиск лиц и сжатие видео от 2-х видеокамер	Одно из 4-х ядер загружено на 100%

Показано, что основное преимущество организации такой параллельно-конвейерной обработки в гетерогенной среде заключается в том, что:

во-первых, каждый вычислитель задействован на своем участке конвейера, где он обрабатывает те данные, для которых его архитектура оптимальна

во-вторых, организация межмодульного взаимодействия по принципу каждый с каждым, позволяет минимизировать конвейерную задержку – задержку при отработке одного полного цикла конвейера в момент времени.

в-третьих, позволяет разгрузить основной транспортный интерконнект

в-четвертых, позволяет существенно снизить нагрузку на центральный процессор и сэкономить его ресурсы для других задач.

Существенное развитие математического аппарата, методов и алгоритмов, применяемых в теории компьютерного зрения, все чаще находят практическое применение в различных прикладных областях раздела компьютерного зрения – машинного зрения, в том числе в системах реального времени. Как правило, задачи машинного зрения достаточно ресурсоемки, поэтому одной из важных задач эффективного практического применения этих теоретических результатов компьютерного зрения является поиск путей минимизации потребляемых вычислительных ресурсов при достижении требуемого быстродействия работы системы. Благодаря архитектурным возможностям ВГВП представляется возможным достигать оптимального результата при решении задач компьютерного зрения.

Список литературы:
References

Головастов А. 2010. Машинное зрение и цифровая обработка изображений. СТА Современные технологи автоматизации № 4: 8-18.

Golovastov A. 2010. Machine vision and digital image processing. the CTA of modern technology automation of number 4: 8-18.

Головкин Б.А. 1980. Параллельные вычислительные системы. Москва.«НАУКА» Главная редакция физико-математической литературы.

Golovkin B.A. 1980. Parallel Computing System. Moscow. "SCIENCE" Home edition of Physical and mathematical literature.

Галаган П.В., Тумакин Д.А. 2016. Высокопроизводительная гетерогенная вычислительная платформа для построения встраиваемых систем. Вопросы радиоэлектроники Серия Радиолокационная техника (РЛТ) № 10. ВЫПУСК 2: 21-31.

Galagan PV Tumakin D.A. 2016. High-performance heterogeneous computing platform for building embedded systems. Questions electronics. Series radar technology (RLT) number 10. Release 2: 21-31.

Чудинов С.М., Парфенов А.В. 2016. Тенденции развития технологии вычислительной техники. Научные ведомости БелГУ. Сер. Экономика. Информатика. 16 (237). 98-106.

Chudinov S.M., Parfenov A.V. 2016. Trends in the development of computing technology. Belgorod State University Scientific Bulletin. Economics Information technologies. 16(237). 98-106.