Выбор платформы для встраиваемых систем захвата и обработки изображений. Часть 2.

В части I статьи, опубликованной в журнале "Системы безопасности" № 2/2018, речь шла об эволюции систем захвата и обработки изображений от простых аналоговых камер и видеомагнитофонов до современных распределенных систем. Были приведены примеры на базе гетерогенной архитектуры Qualcomm® Snapdragon™ и новинки от Intel для решения амбициозных задач, поставленных руководством компании, по кратному увеличению производительности своих устройств (не ограничиваясь x86) в области искусственного интеллекта. В части II мы продолжим рассказ о гибридных платформах с использованием ПЛИС на примере Xilinx® ZynQ® и графическом ускорителе для встраиваемых систем NVIDIA Jetson TX1/TX2.

В отличие от описанных ранее платформ, где основа бизнеса производителей – общеупотребительные процессоры, дополненные специализированными модулями (в частности, для работы с изображениями и видео), компании Xilinx® и NVIDIA разрабатывают и производят специализированные микросхемы на базе собственной архитектуры, лишь относительно недавно дополненные общеупотребительными процессорами. Однако вне зависимости от отправной точки, с разных направлений, все участники нашего обзора пришли к единому пониманию оптимального решения в виде гибридной архитектуры с участием обычных CPU и специализированных процессоров (GPU, VPU, ISP, DSP, FPGA и др.).

Внедрение архитектуры ARM

Эффективные в решении специализированных задач (в нашем случае – обработки изображений) Xilinx FPGA и NVIDIA GPU в силу своей природы мало приспособлены для универсальных применений. Это обстоятельство во многом сдерживало их широкое распространение в конечных устройствах. На помощь пришла архитектура ARM, получившая мощный толчок в развитии благодаря мировому рынку интеллектуальных телефонов и планшетов. Не вдаваясь в технические подробности, хочется отметить особенность бизнес-модели ARM-консорциума, которая как минимум способствовала выбору этой платформы большинством создателей разнообразных гаджетов: разработчики ARM-платформы не занимаются производством микросхем, в отличие, например, от Intel или AMD. Независимые производители процессоров приобретают у них лицензии на необходимые им ARM-платформы и компоненты и организуют их изготовление самостоятельно. Такой подход позволил разработчикам специализированных микросхем (GPU, FPGA, DSP) относительно легко интегрировать готовую архитектуру процессора широкого назначения с собственными устройствами, открыв им дорогу на широкий рынок встраиваемых конечных устройств.

Аппаратная оптимизация ПЛИС

Секрет эффективности технологии ПЛИС (FPGA) по сути своей прост: микросхема, архитектура которой оптимизирована для выполнения заданной программы, всегда выигрывает в производительности и потреблении в сравнении с универсальными процессорами, где оптимизация возможна лишь в программном коде, а аппаратная избыточность – неизбежная плата за универсальность. Очевидное решение – разработка логической схемы под конкретную задачу, такой подход давно известен: ASIC, Application-Specific Integrated Circuit, интегральная схема специального назначения. Это процесс небыстрый и очень дорогостоящий. Обычно он применяется для оптимизированной реализации несложных, стандартных функций, так как всегда существует опасность того что, появление новых технологий сделает такую разработку устаревшей еще до момента вывода "отлитого в камне" изделия на рынок. Другим вариантом решения может быть некий "волшебный" инструмент, позволяющий при создании продукта в разумных пределах задавать аппаратную конфигурацию готового кристалла и вносить в дальнейшем изменения, по мере совершенствования алгоритмов и технологий. Именно так работает ПЛИС.

Рис. 1. Архитектура Xilinx® ZynQ® UltraScale + MPSoC

Конкретная логическая структура ПЛИС (в частности, FPGA) описывается на специальном языке (VHDL) на этапе конфигурации полупроводника, что позволяет получить эффективный аппарат для решения конкретной вычислительной задачи и, при необходимости, его быстро изменить. Важно отметить, что параллельная обработка данных – важнейший элемент любой системы обработки изображений – есть базовое свойство любого ПЛИС.

Процессор реального времени ARM Cortex R5

Для встраиваемой электроники известный производитель FPGA-логики предлагает собственную экосистему на базе платформы Xilinx® ZynQ® (рис. 2). Помимо стандартных для гибридной архитектуры элементов (многоядерный ARM-процессор, управление памятью, графический сопроцессор Mali, аппаратные кодеки H.265/HEVC и программируемой логики FPGA), стоит обратить внимание на наличие процессора реального времени ARM Cortex R5, редко встречающегося в других решениях на базе ARM.

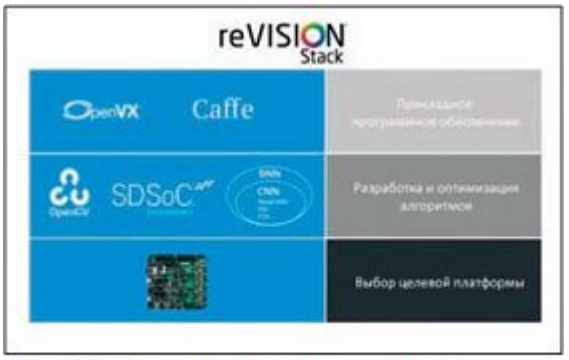


Рис. 2. Программная экосистема для обработки и распознавания reVision

ARM Cortex R5 обеспечивает быструю и детерминированную реакцию системы на внешнее воздействие в критической ситуации, содержит специальные механизмы внутреннего контроля работоспособности, что позволяет сертифицировать и использовать устройства на базе Xilinx®ZynQ® в задачах обеспечения безопасности человека и оборудования.

Секрет эффективности технологии ПЛИС (FPGA) по сути своей прост: микросхема, архитектура которой оптимизирована для выполнения заданной программы, всегда выигрывает в производительности и потреблении в сравнении с универсальными процессорами, где оптимизация возможна лишь в программном коде, а аппаратная избыточность – неизбежная плата за универсальность

Однако вернемся к нашей основной теме – обработке изображений. Как мы уже упоминали ранее, без эффективных и удобных инструментов программирования широкое применение даже самых замечательных решений вряд ли возможно. В экосистеме Xilinx® эта роль отводится набору библиотек и инструментов reVision Stack, открывающему пути к внедрению эффективных технологий FPGA в популярные среды разработки на различных этапах: создание оптимизированной аппаратной конфигурации, разработка алгоритмов и, наконец, конечного приложения в большинстве случаев без необходимости использования достаточно сложного в применении языка VHDL.

Принцип оптимизации вычислений

Выглядит работа программиста следующим образом (рис. 3). В привычной среде, например на базе OpenCV, создается программный код. С помощью отладчика определяются наиболее затратные по времени функции.

Рис. 3. Визуализация результатов работы алгоритма Dense Optical Flow

Затем они заменяются на аналогичные из библиотеки Xilinx, но выполняемые с использованием возможностей оптимизированной архитектуры и параллельных вычислений на базе FPGA. Производится компиляция кода с подключением фирменных библиотек Xilinx. По оценкам компании-производителя, это по сравнению с популярными компактными платформами на базе GPU дает выигрыш:

- в машинном обучении – в шесть раз (изображений/секунд/ватт);
- при обработке изображений – в 42 раза (кадров/секунд/ватт);
- сокращает время отклика в пять раз.

Так, широко применяемый алгоритм определения траектории движения объектов Dense Optical Flow для изображения 4K60 с MIPI-камеры выполняется на платформе разработчика ZCU102 без пропуска кадров (60 кадр/с при качестве 4K), при этом плата потребляет менее 5 Вт и задержка составляет менее 17 мс при загрузке процессора в 15%. А как же столь популярные сегодня нейронные сети? Стандартные подходы из конца 80-х были основаны на вычислениях с 32-разрядной точностью, что представляет определенные трудности для FPGA. Последние исследования в области оптимизации нейронных сетей показывают незначительное падение точности их работы при снижении разрядности весовых коэффициентов до восьми и даже трех разрядов, что позволяет эффективно использовать высочайшую скорость целочисленных вычислений FPGA в оконечных устройствах, реализующих в том числе оптимизированные модели нейронных сетей. Так же, как и на других гибридных платформах, особенно эффективным видится совместное применение алгоритмов обработки изображений и нейронных сетей в их аппаратной оптимизации на базе ПЛИС.

Таблица. Сравнительная таблица компактного воплощения NVIDIA CUDA Jetson TX1/TX2

Fastvideo SDK / NVIDIA GPU	*GeForce 1080*	*Quadro P6000*	*Tegra X1*	*Tegra X2*
JPEG Encoder				
2K gray (8-bit, q=90%)	0.216	0.11	1.5	1.2
2K (8-bit, q=90%, 4:2:0)	0.36	0.17	2.4	2.0
2K (8-bit, q=90%, 4:4:4)	0.40	0.21	3.7	3.1
4K gray (8-bit, q=90%)	0.55	0.35	6.3	5.0
4K (8-bit, q=90%, 4:2:0)	0.78	0.51	9.4	7.9
4K (8-bit, q=90%, 4:4:4)	1.12	0.74	14.9	12.5
4K gray (12-bit, q=90%)	0.83	0.54	11.2	8.5
4K (12-bit, q=90%, 4:2:0)	1.22	0.82	17.9	13.4
4K (12-bit, q=90%, 4:4:4)	1.90	1.32	28.6	22.4

Стоимость платформы разработчика на сайте производителя – от 895 до 1995 долларов в зависимости от емкости (количества логических элементов, памяти и др.) и производительности FPGA. В комплекте поставляются все необходимые аппаратные и программные компоненты для начала работы.

Графический акселератор – это не игрушка

Компания NVIDIA приобрела популярность в мире компьютерных технологий в первую очередь благодаря своим графическим акселераторам GeForce и Titan – неотъемлемым элементом любого высокопроизводительного игрового ПК, а в последнее время еще и, благодаря выдающимся вычислительным способностям, базовым компонентам "фермы" для майнинга (добычи) различных криптовалют. Многие современные суперкомпьютеры также используют акселераторы на основе CUDA-процессоров для обработки огромных объемов информации в коммерческих, научных целях, в задачах обеспечения безопасности и обороны.

Платформы Jetson и технология CUDA

Коммерческий успех в традиционных приложениях позволяет производителю инвестировать значительные средства в перспективные технологии для новых рынков. Предвосхищая широкое распространение компактных интеллектуальных устройств, компания NVIDIA много лет тому назад начала портацию своих технологий на компактные платформы в виде SoM с интегрированными ARM-процессорами.

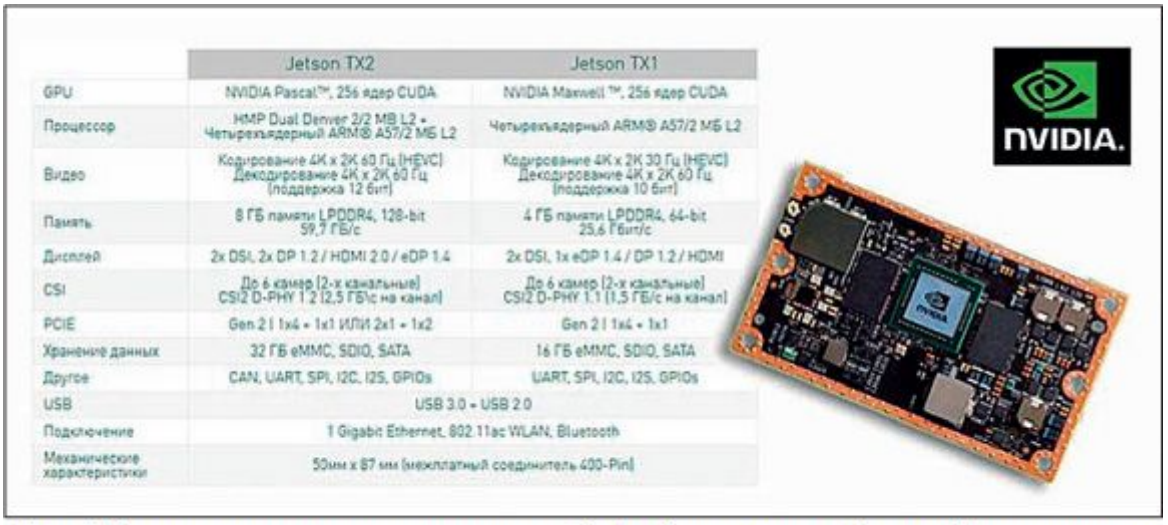
Первым заметным явлением была платформа Jetson TK1. Следом – Jetson TX1, в 2017 г. дополненная более производительным аналогом – Jetson TX2. Основа всех решений – фирменная технология параллельных вычислений CUDA – эффективная аппаратная масштабируемая архитектура, представляющая собой массив индивидуальных процессоров для параллельной обработки и комплект программных средств разработки для самых разных приложений, в частности для обучения и внедрения нейронных сетей и обработки видеопотока.

Результаты тестов

Оценить производительность процессоров NVIDIA в задачах обработки изображений можно по результатам тестов собственных алгоритмов (JPEG-сжатия и других), проведенных российским разработчиком программного обеспечения на базе технологий NVIDIA – компанией Fastvideo (Москва).

Вотличие от описанных ранее платформ, где основа бизнеса производителей – общеупотребительные процессоры, дополненные специализированными модулями (в частности, для работы с изображениями и видео), для Xilinx® и NVIDIA основа бизнеса – это специализированные микросхемы на базе собственной архитектуры, дополненные общеупотребительными процессорами.

На рис. 4 приведено время выполнения операции JPEG-сжатия кадра указанного разрешения с высоким качеством 90% в миллисекундах на целевых платформах без учета затрат на операции ввода изображения.



	Jetson TX2	Jetson TX1
GPU	NVIDIA Pascal™, 256 ядер CUDA	NVIDIA Maxwell ™, 256 ядер CUDA
Процессор	HMP Dual Denver 2/2 MB L2 + Четырехъядерный ARM® A57/2 MB L2	Четырехъядерный ARM® A57/2 MB L2
Видео	Кодирование 4K x 2K 60 Гц (HEVC) Декодирование 4K x 2K 60 Гц (поддержка 12 бит)	Кодирование 4K x 2K 30 Гц (HEVC) Декодирование 4K x 2K 60 Гц (поддержка 10 бит)
Память	8 ГБ памяти LPDDR4, 128-bit 59,7 ГБ/с	4 ГБ памяти LPDDR4, 64-bit 25,6 Гбит/с
Дисплей	2x DSI, 2x DP 1.2 / HDMI 2.0 / eDP 1.4	2x DSI, 1x eDP 1.4 / DP 1.2 / HDMI
CSI	До 6 камер (2-х канальные) CSI2 D-PHY 1.2 (2,5 ГБ\с на канал)	До 6 камер (2-х канальные) CSI2 D-PHY 1.1 (1,5 ГБ/с на канал)
PCIE	Gen 2	1x4 + 1x1 ИЛИ 2x1 + 1x2	Gen 2	1x4 + 1x1
Хранение данных	32 ГБ eMMC, SDIO, SATA	16 ГБ eMMC, SDIO, SATA
Другое	CAN, UART, SPI, I2C, I2S, GPIOs	UART, SPI, I2C, I2S, GPIOs
USB	USB 3.0 + USB 2.0	
Подключение	1 Gigabit Ethernet, 802.11ac WLAN, Bluetooth	
Механические характеристики	50мм x 87 мм (межплатный соединитель 400-Pin)	

Рис. 4. Оценка скорости выполнения операций обработки видеоизображений на различных платформах NVIDIA

Нужно отметить, что JPEG-сжатие позволяет сохранить на порядок больше информации, чем популярные алгоритмы сжатия семейства H.264/H.265. Это в значительной степени оказывает влияние не только на визуальное восприятие картинки человеком, но (что самое важное при машинной обработке) и на точность и достоверность алгоритмов распознавания. Подобное качество изображения требуется далеко не всегда, но в определенных задачах, где детализация изображения является основой получаемого результата, оно просто необходимо.

Для принятия решения о применении той или иной платформы требуется тщательное тестирование прототипа, для этого все производители предлагают комплекты разработчика и исчерпывающий набор средств

разработки программ.

Компания предлагает комплекты разработчика в России по цене от 41 тыс. рублей за Jetson TX2 с платой-носителем, где представлены основные интерфейсы. Доступны и комплекты предыдущих поколений по более низкой цене.

Трудности выбора

Для принятия решения о применении той или иной платформы требуется тщательное тестирование прототипа. Сегодня доступны комплекты разработчика и исчерпывающий набор средств создания программ от множества производителей. Для предварительной оценки, на наш взгляд, можно отталкиваться от следующих общих соображений:

- если речь идет о носимых, мобильных устройствах, где ключевыми факторами являются компактность и наличие современных беспроводных коммуникационных интерфейсов GSM, LT E и Wi-Fi или необходимость организации пользовательского интерфейса под Android, то очевидным выбором будет Qualcomm® Snapdragon™;
- для суперкомпактного воплощения оптимизированных нейронных сетей можно рассмотреть решения Movidus от Intel;
- более серьезные вычислительные возможности для обработки изображений и внедрения оптимизированных нейронных сетей для встраиваемых приложений предлагают Xilinx и NVIDIA;
- FPGA кажется предпочтительным с точки зрения энергетической эффективности и минимальной задержки в обработке данных, в том числе в критических задачах реального времени, например в автовождении или медицинских приложениях. Нужно отметить, что несмотря на значительные усилия производителей FPGA по упрощению создания прикладного ПО, в целом данное решение потребует более высокой квалификации разработчика. Следует обратить внимание также на более высокую стоимость стартового комплекта;
- NVIDIA отличается более развитой программной поддержкой и возможностью масштабирования CUDA-приложений на крупных платформах, вплоть до суперкомпьютеров. Стоимость стартового комплекта примерно в два раза ниже, чем у Xilinx® ZynQ®, но производительность и потребление конечного устройства, скорее всего, окажутся хуже, чем у FPGA;
- наконец, не стоит забывать и про старый добрый x86. Компании Intel и AMD уделяют большое внимание развитию возможностей собственных процессоров в плане работы с видеопотоками и обработки графики, что (принимая во внимание распространенность программных инструментов и широкие возможности интеграции на базе ПК других устройств, тех же GPU и FPGA) позволяет x86 успешно конкурировать с другими решениями во встроенных приложениях.

В любом случае, на наш взгляд, начинать выбор платформы стоит с изучения возможностей средств разработки программного обеспечения, которые сегодня во многом определяют время и стоимость реализации проектов и, безусловно, являются ключевым фактором успеха.

Автор

Максим Сорока
Генеральный директор ООО "ВиТэк"