НОВАЯ ПРОГРАММИРУЕМАЯ ПЛАТФОРМА ДЛЯ УСКОРЕНИЯ ВЫЧИСЛЕНИЙ XILINX VERSAL

ИЛЬЯ ТАРАСОВ, д.т.н., Московский технологический университет МИРЭА

Состоявшийся в 2018 г. анонс новой платформы ПЛИС стал если не революционным изменением, то важным шагом в направлении развития элементной базы программируемой электроники. В статье дается обзор основных характеристик нового семейства ПЛИС компании Xilinx, отнесенных производителем к архитектуре ACAP (Adaptive Compute Acceleration Platform – адаптивная платформа ускорения вычислений). Отличие от предыдущих классов ПЛИС заключается, в первую очередь, в размещении на кристалле нового класса вычислительных ресурсов – матрицы процессорных ядер. Семейство Versal будет выпускаться с соблюдением технологических норм 7 нм на производственных мощностях компании TSMC.

ВВЕДЕНИЕ

В настоящее время, когда на массовом рынке широко представлены многоядерные процессоры, многие разработчики хорошо понимают, что повышение производительности вычислений путем повышения тактовой частоты практически исчерпало свой потенциал. Уменьшение технологических норм не приводит к существенному приросту тактовой частоты, причиной чего является множество технических факторов. Следовательно, основным способом увеличения производительности вычислительных систем является добавление новых ядер и переход к параллельным вычислениям. Отдельным классом являются массово-параллельные вычислительные устройства, имеющие сотни, тысячи и даже более вычислительных узлов на одной микросхеме.

Долгое время ИС программируемой логики, или ПЛИС, имеющие объективно меньшую тактовую частоту из-за программируемых соединений, устойчиво занимали свою позицию среди вычислительных платформ именно благодаря возможности реализации в матрице программируемых ячеек большого количества специализированных вычислительных ядер. В этом случае отставание по плотности компонентов и тактовой частоте компенсировалось тем, что на кристалле ПЛИС можно было разместить большое количество устройств, которые в сумме давали высокую производительность, а во многих случаях – еще и обеспечивали выигрыш по стоимости и потребляемой мощности относительно конкурирующих решений. Однако для такого эффекта было необходимо, чтобы решаемая с помощью ПЛИС задача не имела готового аппаратного решения на базе ЦП, GPU или сигнального процессора.

В процессе эволюции системной архитектуры ПЛИС производители постоянно добавляли в матрицу программируемых ячеек дополнительные компоненты для аппаратной реализации часто используемых в ПЛИС функций. Несмотря на то, что программируемые ячейки, представляя собой универсальный компонент цифровых систем, могут реализовать практически любую применимую на практике схему, многие полезные цифровые узлы при их реализации на базе универсальных ячеек оказываются достаточно «рыхлыми», используя программируемые ресурсы неэффективно. Например, реализация статической памяти на базе триггеров логических ячеек оставляет большинство ресурсов неиспользуемыми. По этой причине одним из первых компонентов, добавленных в состав ПЛИС, стали блоки статической памяти (Block RAM).

По мере выявления преимущественных областей применения ПЛИС в их составе появились и дополнительные устройства – например, аппаратные умножители независимых операндов, и практически сразу же – блоки «умножение с накоплением», являющиеся основой многих алгоритмов цифровой обработки сигналов. Основания для размещения такого аппаратного ресурса в целом те же, что и для статической памяти – реализация операции умножения только с помощью логических ячеек оставляет много ресурсов незадействованными, а тактовая частота такого узла невысока из-за большого числа коммутируемых соединений в умножителе. Еще одним примером аппаратного узла является transceiver MGT, или «мультигигабитный последовательный приемопередатчик».

В отличие от памяти или умножителя, такой компонент, использующий сигналы гигагерцового диапазона и смешанную аналого-цифровую схемотехнику, не может быть реализован на базе только цифровых компонентов. С другой стороны, задачи проводной и беспроводной связи, а также скоростные интерфейсы вычислительных систем (PCI Express, 10 G Ethernet, Interlaken и т. д.) могут быть эффективно решены на базе конфигурируемых микросхем, которым для этого не хватает именно приемопередатчиков. На протяжении долгого времени компания Xilinx выпускает серию Virtex, имеющую на кристалле десятки блоков MGT.

На начало 2010-х гг. облик микросхем ПЛИС с архитектурой FPGA практически сложился и представлял собой матрицу программируемых логических ячеек, внутри которой размещались блоки памяти, «умножение с накоплением», а на периферии кристалла – модули MGT. Часть аппаратных модулей могла быть и не задействована, что, тем не менее, делало ПЛИС достаточно эффективными даже в задачах, не тре-

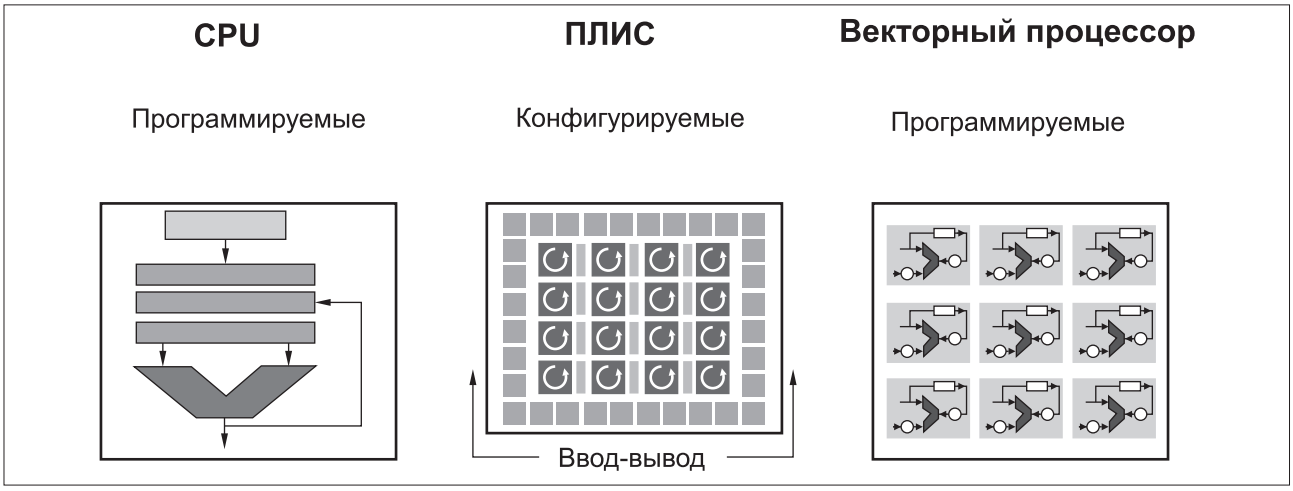

Рис. 1. **Классы микросхем для высокопроизводительных вычислительных устройств**

бующих всей функциональности микросхемы.

Анонс нового семейства Zynq-7000 в составе микросхем серии 7, выполненной с технологическими нормами 28 нм, сопровождался введением в практику нового обозначения подкласса ПЛИС – APSOC (All-Programmable System-On-Chip), или ППСнК (полностью программируемая система-на-кристалле). Основанием для такого шага со стороны компании Xilinx стало размещение в виде аппаратного компонента подсистемы процессора ARM Cortex-A с полным набором периферийных устройств. Матрица программируемых ячеек при этом выступала не столько программируемой периферией ARM, сколько аппаратной платформой для реализации высокопроизводительных параллельных вычислений, управляемых процессором ARM.

На рисунке 1 показаны современные классы микросхем, которые можно использовать для построения высокопроизводительных систем. Процессоры общего назначения (ЦП), преимущественно с архитектурой x86, являются универсальными устройствами, обеспечивающими высокую частоту при работе одного вычислительного потока. Однако число ядер такого процессора ограничено, и при решении задач с высокой степенью параллелизма эффективность процессора оказывается невысокой. Общеизвестно, что программная эмуляция трехмерной графики на процессоре общего назначения выполняется крайне медленно по сравнению со специализированным графическим процессором (GPU).

Рассматривая системы на базе GPU, можно несколько расширить этот класс, включив в него векторные процессоры как таковые. Они могут и не предназначаться для трехмерной графики, однако следует обратить внимание, что эти графические процессоры были на определенном этапе адаптированы для выполнения вычислений общего назначения. Таким образом, можно рассматривать эти микросхемы как матрицу процессоров, которые по отдельности слабее ЦП, однако предназначены для выполнения подкласса задач (например, как в GPU, для ускорения вычислений для трехмерной графики). Современные GPU компании Nvidia имеют несколько тысяч ядер CUDA, которые в сумме дают более высокую производительность по сравнению с ЦП.

В этой статье рассматривается новая архитектура ПЛИС, которая объединяет три основных компонента, показанных на рисунке 1. Если матрица программируемых логических ячеек является основой ПЛИС, а процессоры общего назначения представлены в ППСНК Zynq, то анонсированная архитектура ПЛИС Versal добавляет на кристалл новую подсистему – массив векторных процессоров.

АРХИТЕКТУРА ПЛИС VERSAL

В анонсе микросхем Versal [1] внимание обращается на главное отличие новых ПЛИС от предыдущих поколений этих микросхем. Размещенные на кристалле процессоры существенно отличаются и от программируемых ячеек, и от многоядерного ARM. Процессоры представляют собой RISC-ядра с векторными расширениями, оптимизированными для обработки сигналов в 5G-сетях и задач искусственного интеллекта (Artificial Intelligent, AI). Таким образом, в новых ПЛИС совмещаются три основных класса высокопроизводительных вычислительных архитектур:

- процессоры общего назначения в виде многоядерного ARM Cortex-A72 и двуядерного процессора реального времени ARM Cortex-R5;
- матрица программируемых ресурсов: логических ячеек, блоков памяти и цифровой обработки сигналов (Adaptable Hardware);
- матрица процессоров с векторными расширениями системы команд (AI engines).

Основные компоненты Versal показаны на рисунке 2.

Сводные характеристики ACAP Versal по состоянию на момент подготовки статьи приведены в таблице. Предполага-

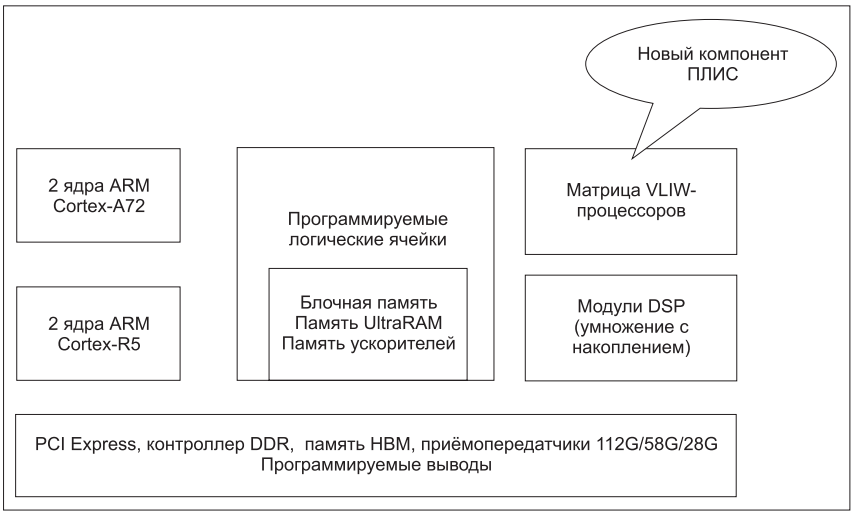

Рис. 2. **Основные компоненты ПЛИС Versal**

Таблица. **Сводные характеристики ACAP Versal**

Семейство	AI core	Prime
Сводная производительность при операциях над 8-разрядными целыми числами, Топ/с	49–147	3–27
Число логических ячеек, тыс.	540–1968	352–2154
Объем памяти, Мбит	68–191	40–324
Число ядер DSP	928–1968	472–3984
Число ядер AI	128–400	–
Число последовательных приемопередатчиков (NRZ, PAM4)	8–44	12–66
Пропускная способность последовательных приемопередатчиков (макс.), Тбит/с	2,9	4,2
Число программируемых выводов	346–692	238–778
Число контроллеров памяти	2–4	1–6

Рис. 3. **Процессорное ядро в ПЛИС Versal**

ется, что номенклатура семейств Versal расширится не только в сторону высокоинтегрированных решений, включающих накристальную динамическую память HBM, но и в сторону компактных, в т. ч. автомобильных микросхем.

Архитектура процессора, входящего в состав матрицы процессорных ядер, показана на рисунке 3. Процессор имеет VLIW-подобные расширения и способен выполнять до шести операций за такт.

ПРИМЕНЕНИЕ ПЛИС VERSAL

Наличие в составе ПЛИС Versal подсистем трех разных типов, предусматривающих разные подходы к решению задач, заставляет обращать большое внимание на методологию проектирования, и особенно на распределение задач между подсистемами. С появлением в составе ПЛИС процессоров ARM одним из распространенных неэффективных решений стало использование разработчиками ПЛИС Zynq в качестве процессоров ARM с очень гибко настраиваемой периферийной частью. Разумеется, при таком подходе соотношение цены и производительности при сравнении с другими микросхемами на базе ARM оказывалось далеко не в пользу Zynq. При этом разработчики-программисты преимущественно оставались в рамках концепции «мощное ядро процессора ARM – основное вычислительное устройство, выполняющее все операции, а задача программируемой части состоит в обеспечении доступа к данным».

Такой подход не позволял эффективно использовать десятки и сотни блоков DSP, способных на порядки превзойти ARM в производительности. С существенным дополнением – в задачах цифровой обработки сигналов: например, БПФ, КИХ и БИХ-фильтров, выделения промежуточной частоты с помощью цифрового гетеродина и т. д. Общим признаком задач, эффективно реализуемых на базе ускорителя, является параллельный характер вычислений при относительной простоте каждого отдельного потока. Таким образом, в системе ARM+ПЛИС оказывается важным разделить задачи на простые, но массово-параллельные (для ПЛИС) и относительно сложные алгоритмически, но не предъявляющие чрезмерных требований к производительности (для ARM).

Появление в Versal третьего компонента – массива процессорных ядер – в еще большей мере повышает важность этапа архитектурного проектирования. Чтобы оправдать использование достаточно дорогих ПЛИС Versal, необходимо выделить для массива процессорных ядер задачи, которые решаются с их помощью эффективнее, чем для ARM или ячеек ПЛИС.

С точки зрения Xilinx, к таким задачам могут относиться обработка данных в сетях пятого поколения (5G) и машинное обучение (Machine Learning). Именно эти направления представляются перспективными с точки зрения производителя. Такие задачи как анализ генома, анализ рисков на финансовых рынках, реализация нейросети GoogleNet ускоряются, по данным Xilinx [1], до 90 раз по сравнению с ЦП, а анализ изображений и обработка данных для сетей 5G – до пяти раз по сравнению с ПЛИС.

Предполагаемый маршрут проектирования для Versal показан на рисунке 4. Видно, что для битового представления конфигурации ПЛИС необходимо разработать три компонента – конфигурацию ПЛИС-части, программу для ARM и набор программ для массива процессоров. Эти три компонента будут

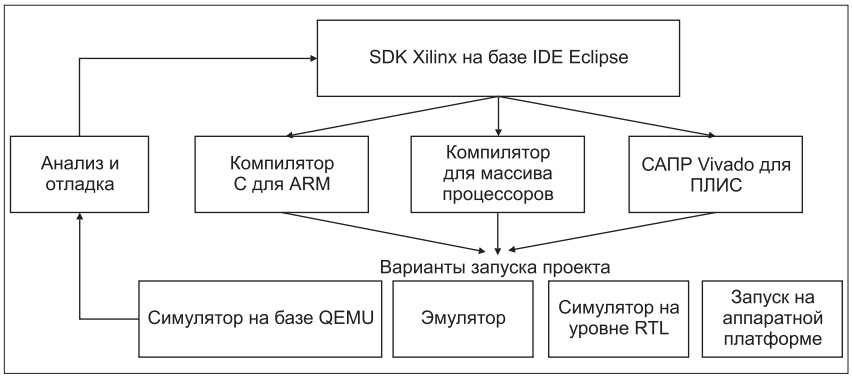

Рис. 4. **Маршрут проектирования для ПЛИС Versal**

объединены в среде проектирования на базе IDE Eclipse.

ВЫВОДЫ

Имеющаяся в настоящий момент информация и состояние нового семейства, даже с учетом того, что в течение 2019 г. едва ли стоит ожидать поставок таких микросхем, позволяет уже сейчас обратить внимание на потенциальные возможности новой платформы и запланировать ее освоение и применение в новых проектах. Эффективное использование подобной дорогостоящей и сложной элементной базы нельзя считать легко достижимым, однако при наличии подходящей задачи и квалифицированного коллектива разработчиков можно ожидать положительных результатов.

ЛИТЕРАТУРА

1. Versal: The First Adaptive Compute Acceleration Platform (ACAP) WP505 (v1.0). October 2. 2018//www.xilinx.com/support/documentation/white_papers/wp505-versal-acap.pdf.

2. Xilinx AI Engines and Their Applications WP506 (v1.0.2) October 3. 2018//www.xilinx.com/support/documentation/white_papers/wp506-ai-engine.pdf.

3. Versal Architecture and Product Data Sheet: Overview DS950 (v1.0) October 2. 2018. Advance Product Specification//www.xilinx.com/support/documentation/data_sheets/ds950-versal-overview.pdf.