УДК 004.272.23

ПРОГНОЗИРОВАНИЕ ПРОИЗВОДИТЕЛЬНОСТИ ПРИ РЕАЛИЗАЦИИ АЛГОРИТМОВ НА ГИБРИДНЫХ АРХИТЕКТУРАХ С СОПРОЦЕССОРАМИ

Андреев А.Е., Силкин И.М., Шафран Ю.В.

ВолгГТУ «Волгоградский государственный технический университет» (400005, Россия, г. Волгоград, пр. Ленина, 28), andan2005@yandex.ru.

В статье рассматривается проблема прогнозирования производительности реализаций алгоритмов с использованием сопроцессоров: графических процессоров (GPU) и программируемых пользователем вентильных матриц (FPGA). Предлагается использовать существующую методику оценки производительности Reconfigurable Amenability Test (RAT), предназначенную для архитектуры FPGA. Оценивается применимость методики RAT для архитектуры GPU, предлагаются методы адаптации. Предлагаются варианты реализаций алгоритма ГОСТ 28147-89 с использованием GPU и FPGA, производится оценка производительности предложенных реализаций по методике RAT, прогнозируемые производительности реализаций сравниваются с экспериментальными результатами. Рассматривается вариант реализации алгоритма триангуляции матриц на основе QR-разложения методом вращений Гивенса, выполняемых процессорами CORDIC, на FPGA, производится оценка производительности данной реализации по методике RAT.

Ключевые слова: прогнозирование производительности, Reconfigurable Amenability Test (RAT), FPGA, GPGPU, ГОСТ 28147-89, криптографический алгоритм, QR-разложение, CORDIC.

PERFOMANCE PREDICTION OF ALGORITHMS IMPLEMENTATIONS ON HYBRID ARCHITECTURES WITH COPROCESSORS

Andreev A.E., Silkin I.M., Shafran Y.V.

VSTU «Volgograd State Technical University» (400005, Russia, Volgograd, Lenin avenue 28), andan2005@yandex.ru

The article deals with predicting of algorithms implementations perfomance when using co-processors: graphics processor (GPU) and field-programmable gate arrays (FPGA). Use of known Reconfigurable Amenability Test (RAT) methodology, designed to estimate performance of implementations using FPGA is considered. Authors assess the applicability and propose the methods to adapt RAT methodology for the GPU architecture. Implementations of GOST 28147-89 algorithm using GPU and FPGA are proposed. Authors make the perfomance forecast for these implementations using RAT methodology and compare it with the experimental results. Implementation of matrix triangulation algorithm based on QR-decomposition with Givens rotations perfomed by CORDIC processors on FPGA is estimated. Authors make the perfomance forecast for this implementation using RAT methodology.

Key words: the prediction of productivity, Reconfigurable Amenability Test (RAT), FPGA, GPGPU, GOST 28147-89, cryptographic algorithm, QR decomposition, CORDIC.

Введение

На современном этапе акцент при увеличении производительности приложений сместился в сторону активного использования параллельных систем и сопроцессоров. В качестве сопроцессоров к CPU (Central Processing Unit) обычно выступают GPU (Graphics Processing Unit), реже сопроцессоры на базе FPGA (Field Programmable Gateway Array). В первом случае повышение производительности достигается благодаря широкому использованию массивного векторного параллелизма, во втором – благодаря конвейеризации и специализации вычислителей, настраиваемых под конкретную задачу. Однако разработка приложений, использующих GPU или FPGA, требует дополнительных временных затрат и

специфических знаний. Наибольшую сложность представляет использование FPGA, так как в этом случае требуется разработка не только программной, но и аппаратной части (логической схемы, размещаемой в устройстве), что в свою очередь предполагает знания в области схемотехники. При этом именно FPGA в ряде случаев позволяют достигать ускорения вычислений там, где традиционные параллельные архитектуры не дают прироста производительности.

В виду сложности разработки для сопроцессоров важной задачей является предварительная оценка прироста производительности при реализации алгоритма с использованием сопроцессора, которая позволит удостовериться в целесообразности разработки. Существует несколько методик прогнозирования производительности алгоритмов для аппаратных архитектур. Большинство из них обладают теми или иными недостатками (сложность и длительность расчетов, по сути сводящие на нет смысл предварительной оценки; низкая точность прогноза; пренебрежение моментами, характерными именно для архитектуры FPGA). На наш взгляд, оптимальной методикой для быстрого, простого и эффективного исследования реализации алгоритмов на FPGA является методика Reconfigurable Amenability Test (RAT), предложенная B. Holland, K. Nagarajan и A. D. George в [7]. Несмотря на то что методика RAT рассчитана на FPGA, мы считаем, что она применима и для GPU. В качестве целевой архитектуры GPU мы рассматриваем видеопроцессоры NVIDIA. Для реализации алгоритмов мы используем технологию NVIDIA CUDA, потому что, как отмечается в [2] и [3], эта технология на данный момент позволяет достигать лучших по сравнению с OpenCL показателей производительности.

Методика Reconfigurable Amenability Test (RAT)

RAT использует три критерия для определения целесообразности реализации алгоритма на FPGA: пропускную способность, числовую точность и необходимые ресурсы. Данные факторы являются доминирующими при оценке эффективности реализации алгоритмов на FPGA. Методология RAT предполагает, что первоначально определяется необходимая точность приложения с учётом типа и количества доступных ресурсов, затем проводится тест пропускной способности для получения прогнозируемой производительности алгоритма.

Для GPU под тестом числовой точности мы подразумеваем выбор чисел с плавающей запятой одинарной (float) или двойной (double) точности, а также выбор математических функций для выполнения действий над ними. Под анализом необходимых ресурсов мы понимаем подбор оптимальной схемы размещения данных в различных видах памяти, что сказывается на быстродействии операций чтения и записи.

Прогнозируемое время выполнения алгоритма определяется исходя из двух показателей: времени обмена данными между CPU и сопроцессором и времени непосредственных вычислений. Время обмена данными учитывает объём данных и скорости чтения и записи, время непосредственных вычислений – опять же объём данных, количество операций над ними, частоту и количество одновременно выполняемых сопроцессором операций. Кроме того, при вычислении времени выполнения алгоритма предлагается учитывать использование одиночной или двойной буферизации. Двойная буферизация подразумевает одновременное осуществление передачи данных и непосредственных вычислений. Расчётные формулы приведены в [7].

Количество одновременно выполняемых операций на GPU в общем случае будет равно количеству ядер. Однако необходимо учитывать, что некоторые операции выполняются для ядер одного мультипроцессора последовательно.

Оценка производительности реализаций ГОСТ 28147-89

В качестве примера применения представленной методики проведём расчёт прогнозируемого ускорения криптографического алгоритма ГОСТ 28147-89 при шифровании 1 Гб данных в режиме простой замены. Предполагаемые схемы вычислений на GPU и FPGA приведены на рисунках 1 и 2 соответственно. Процедура шифрования одного 64-битного блока данных по алгоритму ГОСТ 28147-89 состоит из 32 раундов, описание которых вместе с расчётом количества операций над одним блоком данных приведено в таблице 1. Уже на этапе подсчёта количества операций видно, что задание собственной аппаратной логики на FPGA позволяет сильно снизить количество операций за счёт параллельности подстановки по таблице S-boxes и отсутствия обращений к памяти во время обработки блока.

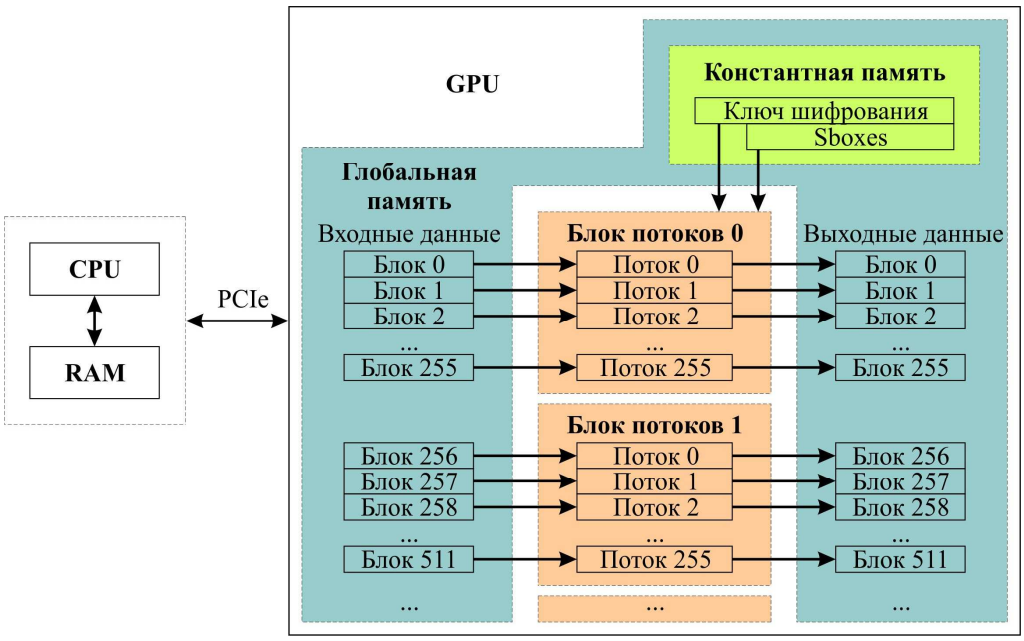

Рис. 1. Схема вычислений ГОСТ 28147-89 на GPU.

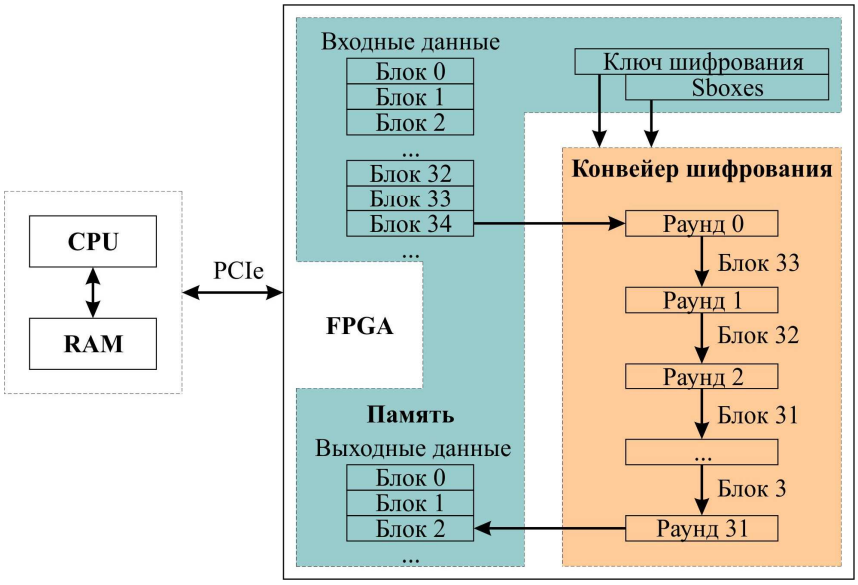

Рис. 2. Схема вычислений ГОСТ 28147-89 на FPGA.

Таблица 1 – Расчёт количества операций ГОСТ 28147-89

Операция	Повторов	GPU	FPGA
Чтение блока данных из памяти	1	402	1
Сложение полублока с ключом	32	2	1
Подстановка в полублок по таблице S-boxes	32	8 * 4	1
Циклический сдвиг полублока	32	2	0

Сложение полублоков по модулю 2	32	1	1
Перестановка блока	32	2	0
Запись преобразованного блока данных в память	1	402	1
Всего		2052	98

В таблице 2 приведена оценка времени выполнения шифрования 1 Гбайта данных согласно методике RAT для устройств: CPU Intel Core i3-2125 (3,30 ГГц), GPU NVIDIA GT 335M (72 ядра, 450 МГц), GPU NVIDIA GTX 260 (216 ядер, 576 МГц), GPU NVIDIA Tesla C1060 (240 ядер, 602 МГц), FPGA Altera Arria II GX EP2AGX125.

Таблица 2 – Оценка производительности реализаций ГОСТ 28147-89 согласно RAT

Устройство	**GPU**			**FPGA**
	NVIDIA GT 335M	**NVIDIA GTX 260**	**NVIDIA Tesla C1060**	**Altera Arria II GX EP2AGX125**
Количество элементов	134 217 728			
Величина элемента, байт	8			
Скорость чтения, Мбайт/с	1684,1	3005,1	4265,3	748,0
Скорость записи, Мбайт/с	1595,0	2932,7	4603,4	748,0
Время обмена данными, с	1,25	0,69	0,46	2,74
Количество операций над элементом данных	2052	2052	2052	98
Частота вычислителя, МГц	450	576	602	125
Количество операций вычислителя за такт	72	216	240	98
Время непосредственных вычислений, с	8,50	2,21	1,91	1,07
Время вычислений (одиночная буферизация), с	9,75	2,90	2,37	3,81
Время вычислений (двойная буферизация), с	8,50	2,21	1,91	2,74
Время вычислений (CPU Intel Core i3-2125), с	24,07			
Расчётное ускорение	2,83	10,89	12,60	8,78
Экспериментальное ускорение	3,05	10,29	12,10	8,36
Разница с оценкой, %	7,59	5,56	4,02	4,86

Экспериментальные результаты реализации алгоритма, разработанной с использованием технологии CUDA, оказались достаточно близкими к теоретической оценке,

что косвенно подтверждает применимость методики RAT для оценки производительности реализации алгоритмов на GPU.

Оценка производительности триангуляции матрицы методом Гивенса

В качестве второго примера проведём с помощью методики RAT прогнозирование реализации алгоритма QR разложения методом Гивенса для матриц большой размерности.

Оптимальная, на наш взгляд, схема использования GPU при реализации вращений Гивенса на GPU приведена в [5], однако, как показано в [5] и [6], ускорение мало и теряется при увеличении размера обрабатываемых матриц. Это связано с высокой долей взаимосвязанных вычислений, которые необходимо проводить последовательно, это значит, что большую долю операций приходится проводить в рамках одного блока потоков GPU. А одновременное выполнение только 8 операций (по числу ядер в мультипроцессоре) при простом пересчёте на частоту ядра даёт сопоставимую с CPU производительность. По изложенным причинам оценку QR разложения методом Гивенса по методике RAT для GPU мы не проводим.

Стандартная архитектура устройства, реализующего метод вращений Гивенса, приведённая в [4], представляет собой систолический массив спецпроцессоров CORDIC [1] в режиме вращения и векторном режиме. Анализ ресурсов показывает, что ёмкости современных FPGA недостаточно для того, чтобы вместить весь систолический массив для матриц большой размерности. Поэтому мы использовали представленную на рисунке 3 архитектуру, представляющую собой несколько CORDIC-устройств, предназначенных для обработки части строки матрицы.

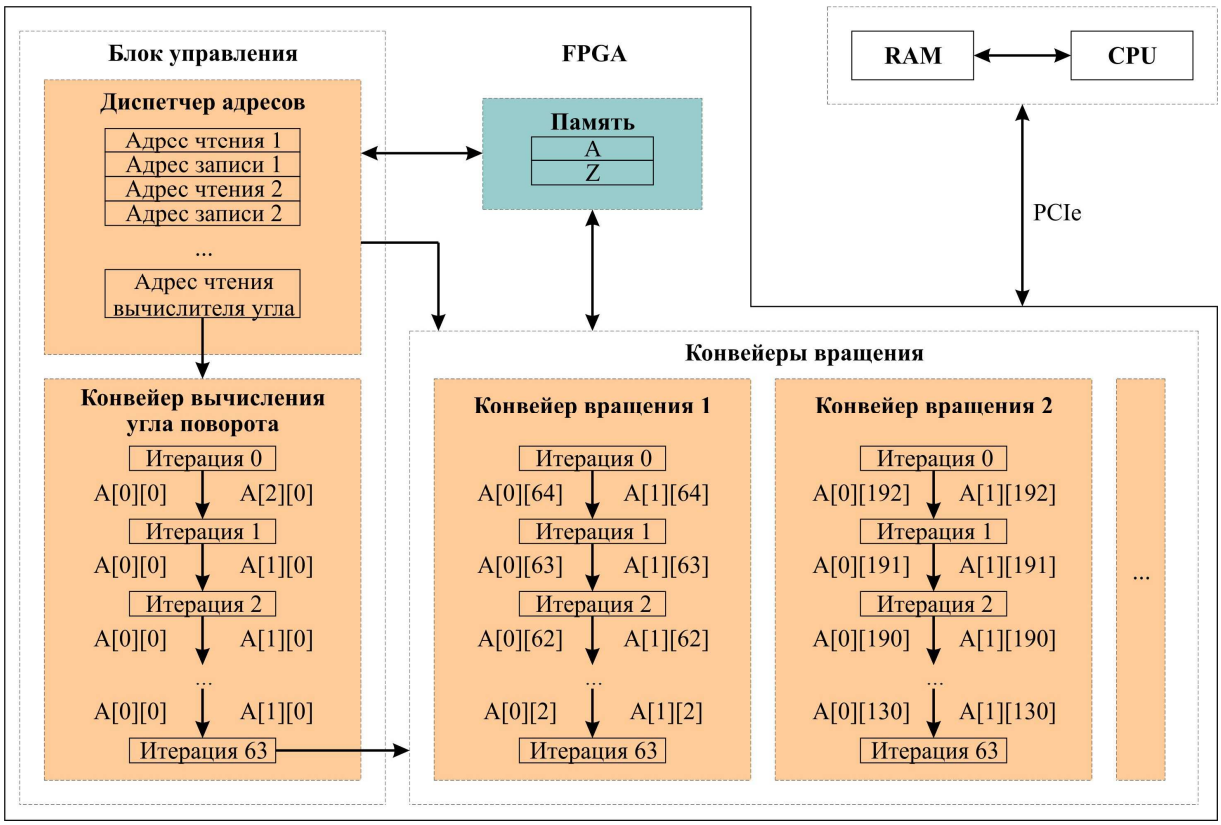

Рис. 3. Схема вычислений вращений Гивенса на FPGA.

Оценка по методике RAT для вращений Гивенса на FPGA приведена в таблице 4. За элемент мы принимаем всю обрабатываемую матрицу. Для обеспечения необходимой точности рассматриваются 64-разрядные конвейерные CORDIC-процессоры, которых, по нашим расчётам, на кристалле FPGA Altera Arria II GX EP2AGX125 в используемом нами сопроцессоре можно разместить не более семи.

Таблица 3 – Оценка производительности реализаций вращений Гивенса согласно RAT

Устройство	FPGA Altera Arria II GX EP2AGX125			
Линейный размер матрицы	512	1024	4096	8192
Количество конвейеров CORDIC	7			
Количество элементов	1			
Величина элемента, Мбайт	2	8	128	512
Скорость чтения, Мбайт/с	748,0	748,0	748,0	748,0
Скорость записи, Мбайт/с	748,0	748,0	748,0	748,0
Время обмена данными, с	0,0054	0,0214	0,3423	1,3692
Количество операций над элементом данных, 10^9	3	23	1 467	11 732
Частота вычислителя, МГц	125			

Количество операций вычислителя за такт	384			
Время непосредственных вычислений, с	0,0600	0,4786	30,5644	244,4254
Время вычислений (одиночная буферизация), с	0,0654	0,5000	30,9067	245,7945
Время вычислений (двойная буферизация), с	0,0600	0,4786	30,5644	244,4254
Время вычислений (CPU Intel Core i3-2125), с	0,07	0,56	34,93	274,62
Расчётное ускорение	1,2000	1,1659	1,1427	1,1235

По результатам прогнозирования получены незначительные показатели ускорения. Кроме того, обеспечение одновременной загрузки данных из памяти на конвейеры и записи выходных данных всех конвейеров в память представляется достаточно сложной задачей без использования памяти на чипе. Однако ресурсы чипа полностью отданы конвейерам. Ввиду изложенных обстоятельств реализация вращений Гивенса на имеющемся аппаратном обеспечении не представляется нам перспективной.

Заключение

Предварительное прогнозирование производительности до начала реализации алгоритмов на FPGA и GPU является одним из важных моментов создания приложений, потому что оно позволяет на начальном этапе отделить задачи, неперспективные с точки зрения повышения производительности на рассматриваемых архитектурах. Результаты, полученные для алгоритма ГОСТ 28147-89, показывают, что методику RAT можно применять не только для FPGA, но и для GPU.

Список литературы

1. Егунов В.А. Аппаратные методы решения задач линейной алгебры : монография / В.А. Егунов, В.С. Лукьянов // ВолгГТУ. – Волгоград, 2007. – 152 с.

2. Блохин О.Д. Исследование высокопроизводительного решения задачи N тел на базе платформы OpenCL / О.Д. Блохин, Д.К. Боголепов, М.М. Захаров, Д.П. Сопин. – 2010. [Электронный ресурс]. – Режим доступа : http://www.itlab.unn.ru/archive/MSConf10/msconf-2010_book.pdf.

3. Karimi K. A Performance Comparison of CUDA and OpenCL / Kamran Karimi, Neil G. Dickson, Firas Hamze // D-Wave Systems Inc. [Электронный ресурс]. – Режим доступа : http://arxiv.org/ftp/arxiv/papers/1005/1005.2581.pdf.

4. Misra M. Design of Systolic arrays for QR Decomposition / Manoj Misra, Rajat Moona. – 1994. [Электронный ресурс]. – Режим доступа : http://www.cse.iitk.ac.in/users/moona/papers/iccse94.pdf/

5. Kerr A. GPU Performance Assessment with the HPEC Challenge / Andrew Kerr, Dan Campbell, Mark Richards // HPEC 2008, Lexington, 23–25 September 2008 / Lincoln Laboratory, Massachusetts Institute of Technology. – Lexington, 2008. [Электронный ресурс]. – Режим доступа : http://www.ll.mit.edu/HPEC/agendas/proc08/Day3/58-Day3-Session6-Kerr-abstract.pdf.

6. Kerr A. QR Decomposition on GPUs / Andrew Kerr, Dan Campbell, Mark Richards // Georgia Institute of Technology, Georgia Tech Research Institute. – 2009. [Электронный ресурс]. – Режим доступа : http://www.akerr.net/andrew/publications/Kerr_Campbell_Richards_QRD_on_GPUs.pdf.

7. Holland B. RAT: RC Amenability Test for Rapid Performance Prediction / Brian Holland, Karthik Nagarajan, Alan D. George. Электронный ресурс]. – Режим доступа : http://www.hcs.ufl.edu/pubs/TRETS08_F3.pdf.

Рецензенты

Завьялов Д.В., д.физ.-мат.н., профессор, заместитель заведующего кафедрой общей физики, Волгоградский государственный социально-педагогический университет, г. Волгоград.

Муха Ю.П., д.техн.н., профессор, заведующий кафедрой «Вычислительная техника», Волгоградский государственный технический университет, г. Волгоград.