УДК 004.31:004.032.34

# Самосинхронное устройство умножения-сложения гигафлопсного класса: варианты реализации

Ю.А. Степченков[1], Ю.В. Рождественский[1], Ю.Г. Дьяченко[1], Н.В. Морозов[1], Д.Ю. Степченков[1], А.В. Сурков[2]

[1]Учреждение Российской академии наук Институт проблем информатики РАН (ИПИ РАН),

{YStepchenkov, YRogdest, YDiachenko, NMorozov, DStepchenkov}@ipiran.ru

[2]Учреждение Российской академии наук Научно-исследовательский институт системных исследований РАН, surkov@cs.niisi.ras.ru

***Аннотация* — В докладе изложены результаты разработки вариантов независимого от задержек устройства умножения-сложения (SIFMA – Speed-Independed Fused Multiply-Add), соответствующего стандарту IEEE 754 и выполняющего либо одну операцию двойной точности, либо одновременно две операции одинарной точности над тремя операндами. Устройство разработано по стандартной КМОП технологии с проектными нормами 65 нм. Оно работает с синхронным или асинхронным окружением и обеспечивает среднюю производительность на уровне 1 гигафлопс при напряжении питания 1В и температуре 25$^0$С. Энергопотребление при этом не превышает 970 мДж/ГГц.**

***Ключевые слова* — самосинхронная схема, умножитель, сумматор, вычитатель, конвейер, индикация.**

## I. Введение

Операция "умножение двух операндов и сложение с третьим операндом" (Fused Multiply-Add, FMA) - одна из наиболее часто используемых в распределенных вычислениях. Наилучшее сочетание потребительских характеристик блока FMA обеспечивается применением самосинхронных схем, не зависящих от задержек элементов (Speed Independent, SI). Методологические аспекты проблемы реализации блока SIFMA, особенности его алгоритмического и аппаратного исполнения и обоснование целесообразности разработки двух вариантов реализации рассмотрены в докладе [1].

Цель данного доклада – разработка двух вариантов устройства умножения-сложения, соответствующего стандарту IEEE 754 [2], принадлежащего к классу SI-устройств [3], обладающего расширенными функциональными возможностями и имеющего сбалансированные характеристики. Первый вариант (далее по тексту – асинхронный) предназначен для работы с асинхронным окружением и позволяет в максимальной степени использовать преимущества самосинхронных схем. Второй вариант (синхронный) учитывает необходимость согласования входного и выходного интерфейса с синхронным окружением.

## II. Структурная схема SIFMA

Ядром обоих вариантов SIFMA является собственно вычислитель операции умножения-сложения, структурная схема которого показана на рис. 1. Его входной и выходной интерфейсы содержат все необходимые сигналы для организации взаимодействия SIFMA с асинхронным окружением.

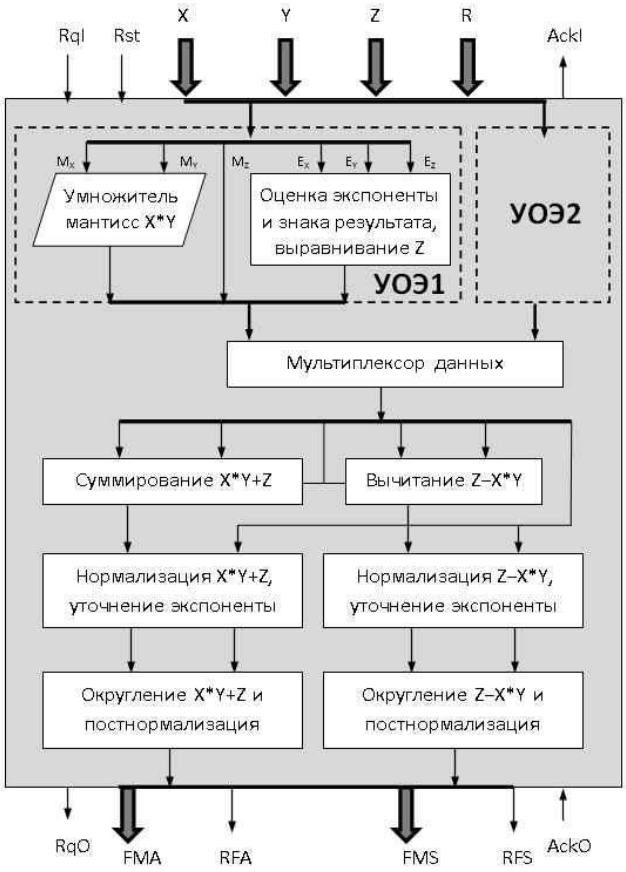


**Рис. 1. Структурная схема ядра SIFMA**

Входные операнды – обрабатываемые числа X, Y, Z и признаки операции R (тип округления, точность и характер операции) поступают на вход SIFMA асинхронно. Их готовность подтверждается сигналом RqI со стороны источника. SIFMA запоминает данные во входных регистрах блоков Умножителя и Обработки Экспонент (УОЭ1 и УОЭ2) и извещает об окончании этой процедуры сигналом AckI.

Умножитель 53×53 (старший разряд – подразумеваемая, но отсутствующая в формате представления чисел по стандарту IEEE 754, единица целых; младшие 52-разряда – мантисса числа) реализован на основе модифицированного алгоритма Бута (Booth) с основанием Radix2 и дерева Уоллеса (Wallace) [4]. Дерево Уоллеса обеспечивает быстрое получение произведения в избыточной форме.

Исследования показали [1], что дополнительное повышение быстродействия при практически таких же аппаратурных затратах достигается при использовании специального самосинхронного кодирования, основанного на троичном избыточном представлении обрабатываемых операндов (вместо парафазного кода). Это позволило сократить количество каскадов сумматоров дерева Уоллеса с 7 до 4 и за счет этого на 16% повысить быстродействие блока УОЭ.

Специальное форматирование входных данных для алгоритма Бута позволяет в рамках одного умножения получить либо один результат операции двойной точности, либо сразу два результата двух операций одинарной точности.

Предлагаемая реализация SIFMA включает два блока УОЭ, работающих параллельно. Это обеспечивает максимальное быстродействие и сбалансированность ступеней конвейера при разумных аппаратурных затратах. Каждый блок УОЭ выполняет умножение операндов X и Y, а также анализ и обработку экспонент всех трех операндов и выравнивание операнда Z.

Очередность предоставления новых операндов блокам УОЭ не зависит от времени вычисления каждого блока и определяется простым чередованием. Это позволяет отказаться от использования арбитража как на входе блоков УОЭ, так и на их выходе, и сохранить последовательность появления результата на выходе SIFMA соответствующей порядку задания операндов на входе. Таким образом, окружение SIFMA всегда знает, результат какой именно операции присутствует на его выходе: SIFMA представляет собой как бы одно FIFO со сложной функциональной начинкой.

Последующие блоки устройства SIFMA обрабатывают поступающие данные в самосинхронном режиме: по мере готовности данные передаются из текущего блока в следующий. Отметим, что SIFMA способен одновременно выполнить две операции: "Z + X*Y" и "Z – X*Y", благодаря наличию двух параллельных путей обработки произведения и третьего операнда.

На выходе SIFMA формируются сумма FMA и/или разность FMS и сопутствующие флаги результата RFA и RFS. Готовность результата индицируется сигналом RqO. Окончание чтения результата асинхронное окружение SIFMA подтверждает сигналом AckO.

Детерминированность процесса подачи операндов на вход SIFMA (на каждом такте системной частоты) при наличии синхронного окружения не позволяет в полной мере использовать тот факт, что время выполнения операции в SIFMA зависит от типа операции и значения операндов. Поэтому для обеспечения максимальной эффективности вычислительного процесса и достижения предельного быстродействия в синхронный вариант были введены устройства сопряжения с синхронным окружением – входное и выходное FIFO, также выполненные в стиле SI-устройств. Результирующая схема устройства показана на рис. 2.

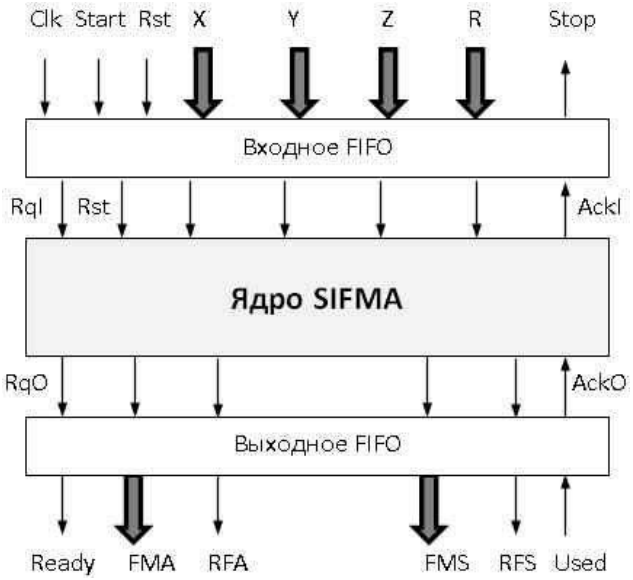


**Рис. 2. SIFMA в синхронном окружении**

Входные данные записываются в FIFO синхронно, по фронту тактового сигнала Clk. Синхронное окружение не ждет подтверждения приема данных от SIFMA, но следит за дополнительным сигналом Stop, свидетельствующим о заполнении FIFO. Выходные данные также считываются синхронно по фронту тактового сигнала Clk, но только при наличии активного уровня на выходе Ready. В свою очередь, синхронное окружение информирует выходное FIFO сигналом Used о том, что оно приняло текущий результат и больше не нуждается в нем. В качестве такого сигнала может использоваться синхросигнал в регистре синхронного окружения, фиксирующем результат операции.

В качестве FIFO использованы самосинхронные полуплотные регистры сдвига [5, рис. 11.9], емкостью четыре слова данных.

## III. Конвейер SIFMA

Конвейер SIFMA обеспечивает производительность на уровне 1 Гфлопс для среднестатистической комбинации входных операндов при типовом напряжении питания (Uип = 1 В) и температуре окружающей среды T = 25$^0$C. При обработке наихудшей, с точки зрения времени выполнения, комбинации входных операндов производительность может снижаться, но в среднем она будет не хуже 1 Гфлопс.

Конвейер самосинхронного устройства обычно строится на основе традиционного запрос-ответного взаимодействия между ступенями (рис. 3). Гистерезисные триггеры (Г-триггеры [5]) на основе индикаторных выходов предыдущей и последующей ступе-

ней конвейера формируют сигналы управления, разрешающие переключение соответствующей ступени из рабочей фазы в спейсер и обратно. Однако в ряде случаев удается оптимизировать взаимодействие ступеней конвейера, повысив его быстродействие путем разделения индикаторных сигналов, управляющих предыдущей и последующей ступенями конвейера. В данном случае это оказалось возможным благодаря структуре каждой ступени конвейера, изображенной на рис. 4. Выходы DGi индицируют поразрядные индикаторы и вместе с DPi составляют общую совокупность парафазных информационных выходов ступени.

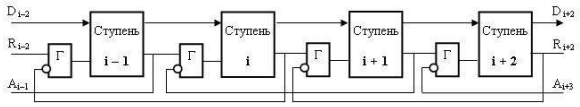


**Рис. 3. Традиционное взаимодействие ступеней конвейера**

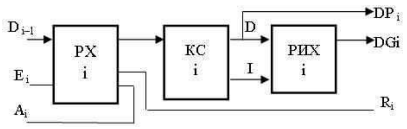


**Рис. 4. Структурная схема одной ступени конвейера**

Каждая ступень конвейера, за исключением входного и выходного FIFO, включает три блока:

1) входной регистр хранения (РХ);

2) комбинационную схему (КС), реализующую алгоритм обработки данных в текущей ступени;

3) выходной регистр индикации и хранения (РИХ).

Реализация РИХ (рис. 5а) и РХ (рис. 5б) на основе Г-триггера позволяет оптимизировать по быстродействию индикаторную подсхему SIFMA. Здесь X, XB – информационные парафазные входы, E – общий для всех разрядов РИХ сигнал управления, I – индикаторный выход соответствующего разряда РИХ, Y, YB – парафазные выходы с нулевым спейсером.

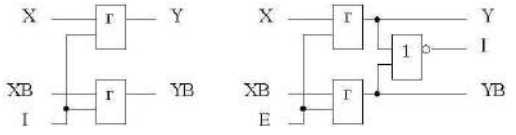


**Рис. 5. Схема одного разряда регистров РИХ (а) и РХ (б)**

Регистры хранят как рабочую фазу, так и спейсер входных сигналов X, XB. Это избавляет от необходимости использовать блок преобразования бифазного сигнала (выхода бистабильной ячейки триггера – разряда регистра хранения) в парафазный сигнал со спейсером, с которым работают комбинационные самосинхронные схемы. Кроме того, РИХ индицирует своими выходами поразрядные индикаторы I схемы КС, а РХ – некоторый сигнал управления E, формируемый индикаторными выходами текущей и следующей ступеней конвейера. Тем самым упрощается и ускоряется формирование общего индикатора многоразрядной КС. Результирующая схема взаимодействия ступеней конвейера показана на рис. 6. Парафазные информационные выходы DP и DG объединяются в одну шину информационных данных.

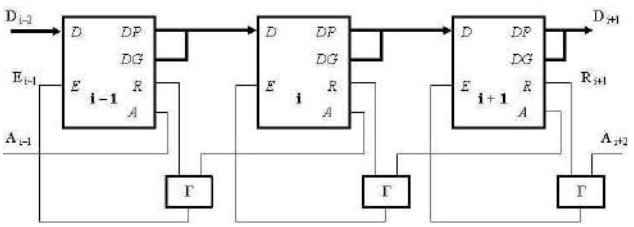


**Рис. 6. Оптимизированная схема взаимодействия ступеней конвейера SIFMA**

Каждая ступень конвейера имеет спекулятивный (упрощенный, на основе критического пути) и полный индикаторный выходы. Спекулятивный индикаторный выход ступени A используется для формирования сигнала E управления предыдущей ступенью конвейера, а полный индикаторный выход R – для формирования сигнала управления текущей ступенью.

Такая реализация позволила достичь средней производительности конвейера при типовых условиях 1 Гфлопс и обеспечить среднее время выполнения операции не более 6 нс.

Описанная организация запрос-ответного взаимодействия между ступенями конвейера была проанализирована программой анализа схемы на самосинхронность АСПЕКТ [6] и АСИАН [7], которые подтвердили принадлежность схемы конвейера к классу SI-устройств.

С учетом необходимости и достаточности индицирования в полном объеме только спейсерной фазы SIFMA индикаторная подсхема КС строится с использованием только комбинационных логических элементов, без гистерезисных триггеров, и реализует логическую функцию "И" (для единичного спейсера) или "ИЛИ" (для нулевого спейсера). Это сокращает аппаратурные затраты и ускоряет формирование индикаторных сигналов, участвующих в запрос-ответном взаимодействии ступеней конвейера.

Разбиение SIFMA на ступени конвейера выполнялось исходя из принципа обеспечения максимального быстродействия SIFMA с учетом приемлемых аппаратурных затрат и сбалансированности конвейера. В целом оно соответствует структурной схеме SIFMA на рис. 1. В варианте с синхронным окружением входное и выходное FIFO являются дополнительными ступенями конвейера.

## IV. Характеристики SIFMA

Характеристики любого устройства во многом определяются базисом его реализации. Базис реализации SI-устройств зависит от используемых принципов организации запрос-ответного взаимодействия между функциональными блоками и способа их индикации. В настоящее время известны два основных базиса реализации SI-схем:

1) избыточная NCL-логика [8], обеспечивающая независимость не только от задержек элементов, но и от задержек в цепях межсоединений;

2) неизбыточная КМОП-логика [5], обеспечивающая независимость от задержек элементов, а в пределах эквихронной зоны [9] – и от задержек в цепях межсоединений.

После сравнения преимуществ и недостатков обоих схемотехнических базисов [10] предпочтение было отдано неизбыточной КМОП-логике, так как она обеспечивает меньшие аппаратурные затраты (в 4,49 раза при реализации двоичного счетчика, в 1,13 раза при реализации умножителя 4x4, до двух раз при реализации более простых логических схем), бо́льшую производительность и меньшее энергопотребление по сравнению с NCL схемами.

Для проектирования SIFMA использовалась библиотека элементов, разработанная для стандартной КМОП технологии с проектными нормами 65-нм. Элементы, вошедшие в состав библиотеки, являются подмножеством элементов из библиотеки для проектирования самосинхронных схем [11]. Они были апробированы при разработке и изготовлении SI-вычислителя [3].

Характеристики вариантов SIFMA с асинхронным (A) и синхронным (C) окружением, выполненных по 65-нм КМОП технологии с шестью слоями металлизации, приведены в таблице.

Таблица

*Характеристики вариантов SIFMA*

| **Параметр** | **Интерфейс SIFMA** | |
|---|---|---|
| | **A** | **C** |
| Сложность реализации, транзисторы | 639000 | 724000 |
| Площадь, мм$^2$ | 0,78 | 0,96 |
| Производительность, Гфлопс | 1,0 | 1,0 |
| Время выполнения операции, нС | 5,95 | 6,90 |
| Энергопотребление, мДж/ГГц | 970 | 1140 |

Быстродействие определялось для типовых условий эксплуатации, так как производительность SI-схем всегда соответствует текущим условиям эксплуатации, а сами SI-схемы не требуют учета наихудшего случая.

## V. Заключение

Представленные варианты устройства, выполняющего операцию FMA в соответствии со стандартом IEEE 754, относятся к классу устройств, поведение которых не зависит от задержек элементов (SI-устройств). Они позволяют получить сумму и разность третьего операнда и произведения двух первых операндов для одной тройки чисел двойной точности или для двух упакованных троек чисел одинарной точности.

Вариант с синхронным окружением отличается от варианта с асинхронным окружением наличием входного и выходного FIFO емкостью в 4 слова данных.

Средняя производительность обоих вариантов устройства при типовых условиях равна 1,0 Гфлопс.

Энергопотребление SIFMA составляет 970 и 1140 мДж/ГГц для вариантов с асинхронным и синхронным окружением, соответственно.

В настоящее время варианты SIFMA запущены на изготовление в составе тестовой БИС.

## Поддержка

Исследование выполнено при финансовой поддержке РФФИ в рамках научных проектов №№ 13-07-12062 офи_м и 13-07-12068 офи_м, а также при частичной финансовой поддержке Программы фундаментальных исследований ОНИТ РАН за 2013 г. (проект 1.5).

## Литература

[1] Соколов И.А., Бобков С.Г., Степченков Ю.А., Рождественский Ю.В., Дьяченко Ю.Г. Самосинхронное устройство умножения-сложения гигафлопсного класса: методологические аспекты / В настоящем сборнике трудов.

[2] IEEE Computer Society. IEEE Standard for Floating-Point Arithmetic IEEE Std 754-2008. doi:10.1109/IEEESTD.2008.4610935.

[3] Степченков Ю.А., Дьяченко Ю.Г., Рождественский Ю.В., Морозов Н.В., Степченков Д.Ю. Разработка вычислителя, не зависящего от задержек элементов // Системы и средства информатики. М.: Торус Пресс, 2010. Т. 20. № 1. С. 5–23.

[4] Hensley J., Lastra A., Singh M. A scalable counterflow-pipelined asynchronous radix-4 booth multiplier // In Proceedings of the International Symposium on Asynchronous Circuits and Systems. 2005. P. 128-137.

[5] Автоматное управление асинхронными процессами в ЭВМ и дискретных системах / Под ред. В.И. Варшавского. М.: Наука, 1986. 400 с.

[6] Рождественский Ю.В., Морозов Н.В., Рождественскене А. АСПЕКТ: Подсистема событийного анализа самосинхронных схем // IV Всероссийская научно-техническая конференция "Проблемы разработки перспективных микро- и наноэлектронных систем-2010". Сборник научных трудов. М.: ИППМ РАН. 2010. С. 26–31.

[7] Рождественский Ю.В., Морозов Н.В., Степченков Ю.А., Рождественскене А.В. Универсальная подсистема анализа самосинхронных схем // Системы и средства информатики. М.: Наука. 2006. Т. 16. С. 463–475.

[8] Fant K.M. Logically determined design: clockless system design with NULL convention logic. N. Y.: John Wiley, 2005. 292 p.

[9] Varshavsky V., Kishinevsky M., Marakhovsky V., et al. Self-timed Control of Concurrent Processes. Dordrecht. The Netherlands: Kluwer Academic Publishers, 1990. 245 p.

[10] Соколов И.А., Степченков Ю.А., Бобков С.Г., Захаров В.Н., Дьяченко Ю.Г., Рождественский Ю.В., Сурков А.В. Базис реализации супер-ЭВМ эксафлопсного класса // Информатика и ее применения. 2014. Т. 8. № 1. 28 с.

[11] Морозов Н.В., Степченков Ю.А., Дьяченко Ю.Г., Степченков Д.Ю. Функциональная полузаказная библиотека самосинхронных элементов ML03 / Свидетельство о регистрации № 2010611908 от 12.03.10.