УДК 004.31:004.032.34 DOI 10.31114/2078-7707-2018-2-170-177

Нечувствительный к задержкам блок умножения-сложения-вычитания с плавающей точкой

И.А. Соколов, Ю.В. Рождественский, Ю.Г. Дьяченко, Ю.А. Степченков, Н.В. Морозов, Д.Ю. Степченков, Д.Ю. Дьяченко

Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук (ФИЦ ИУ РАН)

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

***Аннотация* — Представлено устройство совмещенного умножения-сложения-вычитания, независящее от задержек в элементах и проводниках. Оно полностью соответствует стандарту IEEE 754 и реализует одновременно операции сложения и вычитания третьего операнда из произведения первых двух. Каждый 64-разрядный операнд содержит либо одно число двойной точности, либо два числа одинарной точности. Для увеличения быстродействия умножитель, реализующий модифицированный алгоритм Бута, разбит на две ступени конвейера с ускоренным переключением в спейсер. Схема кодера Бута интегрирована во входное FIFO. Выполнение сложения и вычитания в троичном избыточном коде обеспечивает сокращение аппаратных затрат всего блока. С целью сокращения энергопотребления, блок построен как одноканальное устройство. Блок разработан на базе объемной КМОП технологии с проектными нормами 65 нм с использованием библиотеки стандартных элементов, дополненной самосинхронными элементами, и обеспечивает производительность на уровне 3 гигафлопс.**

***Ключевые слова*— избыточное кодирование, троичный сумматор, "дерево" Уоллеса, эквихронная зона, FIFO.**

I. Введение

Аппаратное совмещение умножения двух входных операндов и последующего сложения-вычитания с третьим входным операндом в одном устройстве может выполняться с промежуточным округлением произведения (типично для сигнальных процессоров первых поколений) или с одним округлением общего результата. Версия с одним округлением получила название fused multiply-add (FMA). Она стала де-факто стандартной операцией современных центральных процессоров, так как обеспечивает более высокую точность вычислений по сравнению с вариантом с промежуточным округлением.

Известные в настоящее время реализации данного устройства в подавляющем большинстве являются синхронными [1]-[3]. Асинхронные решения, претендующие на принадлежность к классу самосинхронных устройств[4]-[5], не отслеживают в полной мере завершение переключения во всех элементах схемы перед переходом в следующую фазу работы. Поэтому они не могут считаться устройствами, правильное функционирование которых не зависит от задержек в элементах и проводниках, т.е. нечувствительных к задержкам (НЧЗ, Delay-Insensitive) при любых условиях эксплуатации.

Сохранение работоспособности НЧЗ схем при сверхмалых значениях питающих напряжений открывает широкие перспективы для применения в портативных изделиях с аккумуляторным питанием и создания бортовых комплексов, не требовательных к объему и стабильности энергоресурсов. Устойчивая работа в экстремальных условиях достигается за счет аппаратной избыточности и дополнительных временных затрат на индикацию и фазу «спейсера» в работе НЧЗ схем. Однако грамотное проектирование НЧЗ схем позволяет существенно снизить эту избыточность, а в ряде случаев, например, в отказоустойчивых устройствах [6], получить результаты лучше, чем в синхронных аналогах.

Ранее авторами уже предпринимались попытки разработки устройства FMA гигафлопсного класса, независимого от задержек в элементах, – SIFMA [7]-[8] и SIFPC [9]-[10]. Однако для достижения предельного быстродействия в SIFMA использовался принцип спекулятивной индикации, не обеспечивавший его стопроцентной самопроверяемости, а SIFPC был построен как двухканальное устройство с двухступенчатым конвейером с общим входом и выходом и адаптивной индикацией, не учитывавшей реальных размеров эквихронной зоны (ЭЗ) [11].

В данной статье излагаются результаты проектирования 64-разрядного НЧЗ вычислительного устройства, выполняющего операции умножения со сложением и умножения с вычитанием с плавающей точкой в соответствии со стандартом IEEE 754 (Delay Insensitive Fused Multiply-Add-Subtract, DIFMAS). Оценки сложности реализации и временные характеристики обсуждаются в разделах II и III. В качестве прототипа математической модели вычислений, включая избыточное представление сомножителей, была выбрана синхронная реализация блока FMA [12], поскольку она обеспечивает наилучшие характеристики при использовании самосинхронного схемотехнического базиса. Методологические аспекты построения самосинхронного устройства FMA были подробно рассмотрены в [7].

II. Особенности DIFMAS

Как и в предшествующих разработках, каждый из трех обрабатываемых операндов содержит либо одно число двойной точности, либо два числа одинарной точности. В последнем случае выполняются две независимые операции: FMA и FMS, – над двумя тройками операндов одинарной точности.

A. Структурная схема DIFMAS

Разработка современных вычислительных средств ведется в направлении обеспечения минимального энергопотребления при достаточно высокой производительности. Это определяется тенденцией к использованию сравнительно невысокой тактовой частоты и большого числа вычислительных узлов на одну СБИС для высокопроизводительных компьютеров.

Наличие двух фаз в работе любой НЧЗ схемы (активной – рабочей и паузы – спейсерной) наталкивает на идею использовать два параллельных канала, фазы работы которых чередуются. Именно такая реализация была представлена в работах [9]-[10]. Она позволила достичь среднестатистической производительности на уровне 3,15 Гфлопс при работе с синхронным окружением и 3,90 Гфлопс при отсутствии непроизводительного ожидания отклика от синхронного окружения об успешном считывании результата с выхода FMA. Однако это вылилось в большие аппаратурные затраты и относительно большое энергопотребление устройства.

В связи с этим была предложена новая структурная схема устройства, вычисляющего сумму (FMA) и разность (FMS) произведения двух первых операндов и третьего операнда, – НЧЗ сопроцессора с плавающей точкой (DIFMAS), показанная на рис. 1.

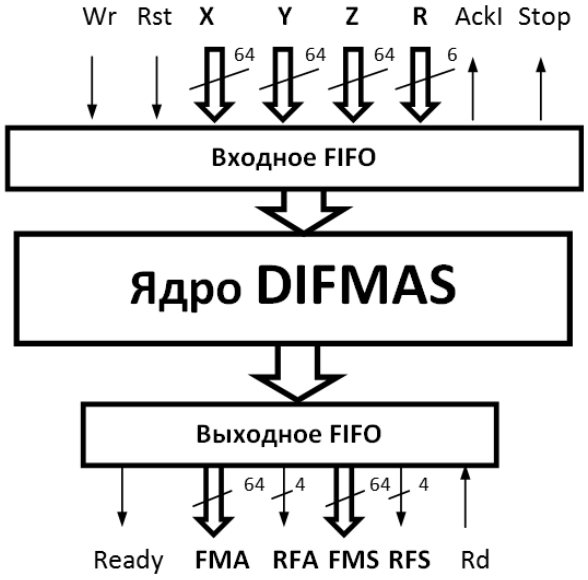

Рис. 1. Структурная схема DIFMAS

Помимо ядра DIFMAS, она содержит входное и выходное FIFO, которые повышают быстродействие DIFMAS при работе с синхронным окружением [7] за счет буферизации потока данных. Входное и выходное FIFO реализованы как регистровая самосинхронная память емкостью 4 слова.

Схема ядра DIFMAS представлена на рис. 2. В отличие от предыдущих разработок устройства FMA [7]-[10], в ней нет дублирующихся каналов или блоков умножителя. Тем не менее, она обладает лучшим, по сравнению с предшествующими реализациями, соотношением "быстродействие / аппаратурные затраты" за счет инновационной индикации блока умножителя и более широкого использования избыточного самосинхронного кодирования.

Рис. 2. Структурная схема ядра DIFMAS

B. Блок умножения DIFMAS

Блок умножения является наиболее ресурсоёмким (55–60% от всего устройства), энергопотребляющим и времязатратным блоком среди всех функциональных блоков DIFMAS. В данной работе за основу выбран базовый вариант блока умножения, использующийся в подавляющем большинстве современных вычислительных устройств. Это чисто комбинационная схема, состоящая из кодера Бута (Radix-2) и "дерева" Уоллеса на сумматорах с избыточным кодированием (redundant binary representation, RBR) [12]-[13].

Описанные в работах [7]-[10] варианты самосинхронного блока FMA включали в себя 2 блока умножения для достижения требуемой производительности. В качестве самосинхронного кода использовались парафазный для кодера Бута и избыточный (троичный) для "дерева" Уоллеса. Оба кода имеют двухфазную дисциплину кодирования, состоящую из рабочей и спейсерной фаз. Оптимизация самосинхронной дисциплины взаимодействия между ступенями конвейера FMA позволила обеспечить латентность вычислений согласно выражению:

$$T_L = N^* T_W + T_S,$$

где: T_L – латентное время работы конвейера FMA; T_W – длительность рабочей фазы ступени конвейера; T_S – длительность спейсерной фазы ступени конвейера; N – количество ступеней конвейера.

Однако длительность цикла работы одной ступени конвейера осталась по-прежнему равной суммарной длительности двух фаз и это определило производительность самосинхронного FMA. При реализации блока умножения в виде одной ступени конвейера в 65-нм КМОП технологии длительность среднестатистического цикла составила 1,5 нс без учёта топологической реализации. Использование двух параллельных блоков умножения в SIFMA и двух параллельных каналов обработки троек операндов в SIFPC позволило существенно ускорить работу блока самосинхронного FMA за счет увеличения аппаратурных затрат.

В то же время, простое разбиение блока умножения на две ступени конвейера позволяет уменьшить длительность каждой ступени до 1 наносекунды без учёта трассировки, что является явно недостаточным. Дальнейшее увеличение числа ступеней конвейера в блоке умножения будет чрезвычайно затратным и не приведёт к заметному сокращению длительности цикла работы ступени конвейера. Основной причиной этого являются существенные затраты времени на индикацию завершения всех процессов переключений элементов на каждой ступени конвейера при высокой разрядности обрабатываемых операндов. Дополнительные затраты привносятся и за счет организация самосинхронных регистров для хранения промежуточных результатов.

Для индикации только одного выходного регистра одной ступени конвейера блока умножения требуется пятислойное индикаторное "дерево" на трехвходовых гистерезисных триггерах (Г-триггерах, [11]). Суммарная задержка формирования индикаторных сигналов в двух фазах для 106 выходных разрядов составляет в 65-нм КМОП технологии 300-350 пс. Индикация регистров привносит ещё 100-150 пс. В результате суммарная длительность рабочей и спейсерной фаз приближается к 1 нс, в то время как максимальная длительность цикла работы ступени конвейера для обеспечения гигафлопсной производительности без учёта топологии не должна превышать 800-850 пс.

В разработанном варианте НЧЗ блока умножения удалось решить эту задачу с помощью совместного использования целого ряда специальных методов:

- принудительного ускорения спейсерной фазы самосинхронного цикла;
- ввода дополнительного регистра временного хранения промежуточных данных во второй стадии "дерева" Уоллеса;
- перехода от последовательной индикации, в пределах библиотечных элементов, к общей параллельной индикации внутренних переменных с последующей оптимизацией временных параметров;
- применения более быстрых троичных сумматоров в "дереве" Уоллеса.

Рис. 3 демонстрирует схему одноразрядного троичного сумматора из "дерева" Уоллеса блока DIFMAS. Схема сумматора на рис. 3 более сложная (на 18%), но обладает существенно лучшим быстродействием (на 27%) в составе конвейера в сравнении со схемой одноразрядного троичного сумматора-прототипа из предшествующих вариантов блока FMA [8].

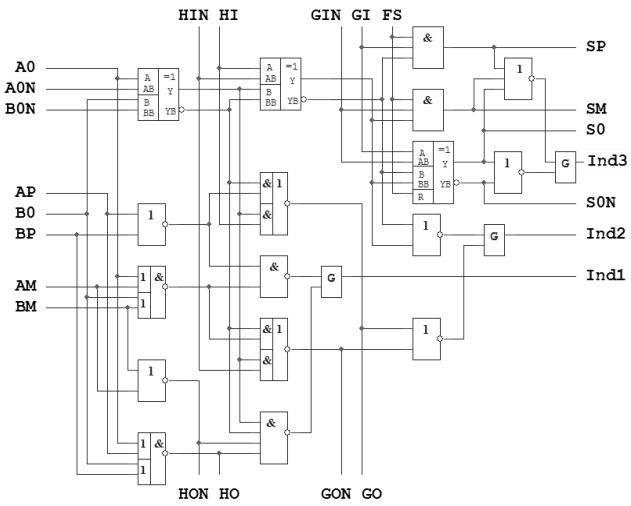

Рис. 3. Троичный НЧЗ сумматор

Вход *FS* на рис. 3 обеспечивает ускоренное переключение одноразрядного сумматора и всего "дерева" Уоллеса в спейсер. Три индикаторных выхода *Ind1* – *Ind3* объединяются в один индикаторный выход "дерева" Уоллеса распределенной индикаторной подсхемой, учитывающей соотношение задержек их формирования в разных разрядах и каскадах "дерева" Уоллеса. Реализация этого подхода позволила дополнительно повысить быстродействие умножителя на 12%.

Разработанный с использованием предложенных методов блок умножения DIFMAS реализуется в виде двух ступеней конвейера и обеспечивает суммарную длительность рабочей и спейсерной фазы 850-900 пс без учёта топологической реализации. Это в 1,7 раза меньше исходного классического варианта, что позволяет получить DIFMAS гигафлопсного класса с использованием одного умножителя. Дополнительные аппаратные затраты, обеспечивающие повышение бы-

стродействия умножителя, составляют 12-13%. Однако суммарные затраты на умножитель в составе DIFMAS в результате снижаются на 40-45% в сравнении с SIFMA и SIFPC за счет использования одного умножителя вместо двух. Пропорционально снизились энергетические затраты на операцию умножения и площадь топологии блока умножения DIFMAS.

При построении схемы умножителя использовались следующие принципы формирования парафазных входов комбинационной схемы (ступени конвейера):

1. Входы могут переключаться в рабочую фазу, если схема завершила переключение в спейсер и предыдущая и последующая ступени конвейера разрешили ей переключение в рабочую фазу; входы могут переключаться в спейсер, если схема завершила переключение в рабочую фазу и предыдущая и последующая ступени конвейера разрешили ей переключение в спейсер. Вход разрешения переключения в противоположную фазу работы формируется индикаторами данной, предыдущей и следующей ступеней конвейера.

2. Вход разрешения переключения в противоположную фазу работы является аналогом локального синхросигнала и может быть реализован в виде древовидной структуры типа "дерева тактового сигнала".

3. Формирователем входов является регистр на выходе предшествующей ступени конвейера. Разряд регистра реализуется схемой из двух Г-триггеров и индикаторного элемента (2И-НЕ или 2ИЛИ-НЕ в зависимости от типа спейсера входов).

Выходной регистр всей НЧЗ схемы реализуется двухтактными RS-триггерами с парафазным информационным входом, входом разрешения записи и бифазным выходом для обеспечения интерфейса с синхронным окружением.

C. Входное и выходное FIFO

Входное и выходное FIFO, использованные в предшествующих реализациях блока FMA [7]-[10], обладали одним существенным недостатком – относительно большим энергопотреблением. Это обусловлено их схемотехнической реализацией на основе полуплотного регистра сдвига [11, рис. 11.9]: слово данных, записанное во входную головку регистра, автоматически продвигается к выходной головке регистра до ближайшей не занятой ячейки.

С одной стороны, это обеспечивает строгую последовательность слов данных, записываемых в FIFO и считываемых из него, без дополнительных аппаратурных затрат на механизм адресации текущей выходной ячейки FIFO. С другой стороны, на пути к выходной головке FIFO слово данных вынужденно проходит через все промежуточные ячейки FIFO, вызывая перезаряд их паразитных емкостей, что приводит к дополнительному энергопотреблению.

С целью сокращения энергопотребления блока DIFMAS была предложена новая реализация входного и выходного FIFO, основанная на регистровом файле и указателях активного регистра для операций записи и чтения FIFO. Разряды регистров FIFO реализованы на самосинхронном однотактном триггере с унарным информационным входом, схема которого показана на рис. 4, что отвечает требованиям как синхронного, так и самосинхронного интерфейсов DIFMAS с окружением и упрощает управление FIFO. Здесь *D* – информационный унарный вход; *E* – вход разрешения записи; *I* – индикаторный выход.

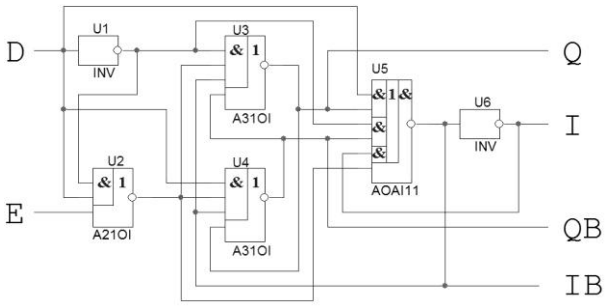

Рис. 4. Схема разряда регистра FIFO

Структурная схема FIFO показана на рис. 5. Сигналы *WrEn* (разрешение записи), *Wr* (инициация записи) и *Full* (признак заполнения FIFO) обеспечивают взаимодействие с синхронным устройством, формирующим входные операнды (*Op1*, *Op2*, *Op3*) для блока DIFMAS. Выходы *OK* (готовность операндов *X*, *Y*, *Z* на выходе FIFO) и *RdReq* (запрос на чтение следующей тройки операндов из FIFO со стороны ядра DIFMAS) реализуют запрос-ответное взаимодействие FIFO с остальной частью DIFMAS.

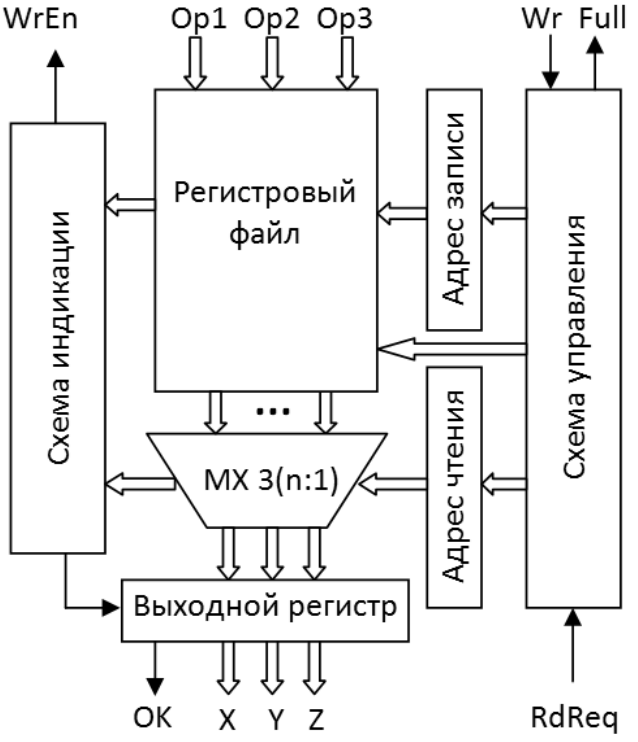

Рис. 5. Структурная схема FIFO

Регистровый файл рассчитан на хранение n троек операндов и сопровождающих их признаков операций. Блоки адреса чтения и записи реализованы на НЧЗ счетчиках. Схема управления выполняет арбитраж операций записи и чтения и предупреждает формирователь входных операндов о необходимости приостановки накачки FIFO (сигнал *Full*). Мультиплексор MX выбирает из регистрового файла тройки операндов в порядке их записи в FIFO с помощью счетчика адреса чтения.

Предложенная реализация FIFO в сравнении с реализацией на основе самосинхронного полуплотного регистра [8] обладает в 2,6 раза меньшим энергопотреблением при практически одинаковых аппаратурных затратах и быстродействии.

D. Индикация DIFMAS

DIFMAS является устройством, обрабатывающим многоразрядные данные и занимающим достаточно большую площадь на кристалле СБИС. В нем реализованы принципы оптимальной индикации на локальном уровне, отработанные в предшествующем варианте блока FMA (SIFPC) [10]. Однако на уровне крупных функциональных блоков, входящих в состав DIFMAS, индикация реализована как минимально необходимая с точки зрения принципов индикации самосинхронных схем и не учитывает ограничений, накладываемых на подсхему индикации размером ЭЗ [11].

Под ЭЗ понимаются фрагменты схемы, компоненты которых функционируют в "одном времени" [11, стр. 10] и разница в задержках сигналов в цепях межсоединений после разветвления не превышает минимальной задержки переключения произвольного элемента библиотеки стандартных элементов, использованной для реализации данной СБИС.

На рис. 6 показан пример разветвления сигнала для случая трех приемников сигнала, формируемого некоторым элементом U0. Задержки распространения сигнала с выхода элемента U0 до входов элементов U1, U2 и U3 ($t_{зд1}$, $t_{зд2}$ и $t_{зд3}$ соответственно) определяются технологией изготовления микросхемы и ее топологической реализацией: длинами отрезков цепи сигнала с выхода элемента U0 до элементов U1, U2 и U3 после точки разветвления сигнала *A* и их физической реализацией (на каких слоях трассировки проведены соответствующие отрезки трассы межсоединения). Если выполняются одновременно соотношения:

$$\begin{cases} |t_{зд1} - t_{зд2}| < t_{мин.зд.}, \\ |t_{зд2} - t_{зд3}| < t_{мин.зд.}, \\ |t_{зд1} - t_{зд3}| < t_{мин.зд.}, \end{cases} \quad (1)$$

где $t_{мин.зд.}$– минимальная задержка переключения выхода любого из элементов U1, U2, U3 относительно соответствующего входа, то схема на рис. 6 считается расположенной целиком в ЭЗ, и индицировать сигнал *A* можно на входе любого из элементов U1 – U3. В противном случае индицировать сигнал *A* необходимо на конце отрезка цепи с максимальной задержкой.

Поскольку реальные задержки выходного сигнала элемента U0 после точки разветвления сигнала *A* определяются конкретной топологической реализацией (взаимным расположением элементов, подключенных к одной цепи, и использованными технологическими слоями трассировки), достоверный анализ соотношений (1), во-первых, возможен только после топологической реализации схемы и, во-вторых, необходим после каждой коррекции топологии схемы.

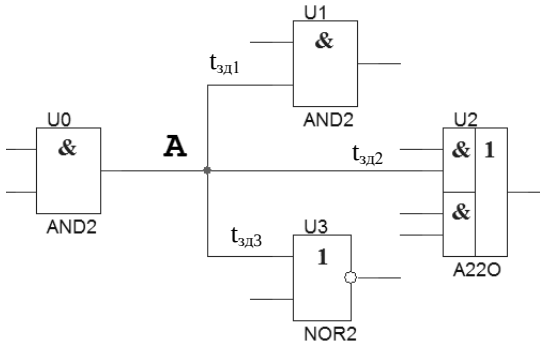

Рис. 6. Пример разветвления сигнала для случая трех приемников

Анализ влияния паразитных емкостей и сопротивлений в стандартной 65-нм КМОП технологии показывает, что использование для проведения трасс межсоединений слоев металла второго и третьего уровней в наихудшем случае приводит к появлению паразитных емкостей с погонным значением 202 фФ/мм, а при проведении той же трассы в слоях металла четвертого и пятого уровней – 198 фФ/мм. С учетом того, что типовая емкость входа стандартного библиотечного элемента не превышает 1,5 фФ, получается, что основной вклад в задержку распространения сигнала в микросхеме, изготовленной по 65-нм КМОП технологии, вносят трассы межсоединений.

Результаты моделирования схемы, показанной на рис. 6, с учетом паразитных параметров, восстановленных из ее топологической реализации, показывают следующее:

1. Для элементов с одинарной нагрузочной способностью разность задержек распространения сигнала по трассам, отличающимся по длине на 60 мкм, составляет около 5 пс, что соответствует задержке переключения одинарного инвертора для данной технологии ($t_{мин.зд.}$ = 5 пс). При этом разность задержек не зависит от типа элемента и сложности выполняемой им функции.

2. Для элементов с повышенной нагрузочной способностью разность задержек распространения сигнала по таким же трассам оказывается даже больше, чем для элементов с одинарной нагрузочной способностью (например, для инвертора с 40-кратной нагрузочной способностью сами задержки сокращаются, но разность между ними оказывается порядка 6 пс).

Таким образом, размер ЭЗ для КМОП технологии с проектными нормами 65 нм не превышает 60 мкм для

элементов с одинарной нагрузочной способностью и 50 мкм для элементов с большой нагрузочной способностью. Следует, однако, учесть, что размер ЭЗ ассоциируется с радиусом окружности, центр которой располагается в точке разветвления трассы.

Размер ЭЗ – понятие довольно условное. Элемент-приемник может находиться достаточно близко от элемента-передатчика, но трасса к нему после точки разветвления может "плутать" по топологии, из-за чего ее длина и соответствующая ей паразитная емкость, определяющая задержку сигнала, будут относительно большими. Поэтому целесообразно говорить об *эквихронных трассах* (ЭТ), имея в виду их длину (и задержку распространения сигналов по ним) после точки разветвления.

Для обеспечения свойства НЧЗ необходимо соблюдение следующих условий:

1. Парафазные сигналы должны индицироваться на входе их приемника, на конце самой длинной (задержанной) трассы.

2. При наличии трасс, выходящих за пределы подмножества ЭТ данного сигнала, индицироваться должны и сигналы на концах всех этих трасс.

Проверка описанных условий выполняется при анализе схемы на самосинхронность с помощью программы анализа [14] с учетом реальных паразитных параметров трасс, извлеченных из топологической реализации анализируемой схемы.

III. Параметры DIFMAS

DIFMAS был спроектирован в стандартной 65 нм КМОП объемной технологии с 6 слоями металлизации. Параметры DIFMAS в сравнении с синхронным аналогом близкой производительности [15] приведены в таблице. Временные и энергетические параметры получены на основе моделирования без учета паразитных параметров топологической реализации для статистически достоверного набора комбинаций входных операндов двойной и одинарной точности.

Таблица

Параметры DIFMAS

Наименование параметра	**Аналог**	**DIFMAS**
Частота работы, ГГц	1,03	1,02
Площадь топологии, мм^2	0,312	0,468
Латентность, нс	10,8	2,94
Производительность, Гфлопс	2,06	3,06
Эффективность площади, мм^2/Гфлопс	0,151	0,153
Диапазон работоспособности по напряжению питания$V_{пит}$	$V_{пит}$±10%	$V_{пор}$...$V_{проб}$
Обнаружение константных неисправностей	–	+

Быстродействие определялось для типовых условий эксплуатации (1,0 В напряжения питания, 25^{o}C), так как производительность НЧЗ схем всегда соответствует текущим условиям эксплуатации, а сами НЧЗ схемы не требуют учета наихудшего случая для обеспечения работоспособности схемы во всем гарантированном диапазоне изменений напряжения питания и температуры окружающей среды.

Следует отметить, что DIFMAS обладает большей функциональностью по сравнению с аналогом: за один цикл он способен обработать одну тройку операндов двойной точности или две тройки операндов одинарной точности, вычисляя при этом не только сумму, но и разность произведения первых двух операндов и третьего операнда (это учтено в показателе производительности). Кроме того, он имеет намного более широкий диапазон работоспособности, ограниченный лишь пороговыми напряжениями КМОП транзисторов ($V_{пор}$) и напряжением пробоя полупроводниковых структур ($V_{проб}$), и прекращает работу при обнаружении константных неисправностей [11]. Платой за эти преимущества является б*о*льшая сложность реализации и, в связи с этим, большее энергопотребление. Энергопотребление может быть снижено до требуемой величины за счет уменьшения питающего напряжения при соответствующем снижении производительности. За счет меньшего числа ступеней конвейера латентность DIFMAS в 3,7 раз меньше, чем у синхронного аналога.

Таким образом, представленный вариант DIFMAS обеспечивает производительность на уровне 3,06 Гфлопс. Он реализует современный тренд в построении вычислительных средств высокой производительности: использование большего количества процессоров с относительно низкой производительностью.

IV. Заключение

DIFMAS с одним блоком умножителя, разработанный по КМОП технологии с проектными нормами 65 нм, демонстрирует высокую среднюю производительность (3,06 Гфлопс при типовых условиях) и хорошую латентность (менее 3 нс).

Использование избыточного самосинхронного кодирования, двухступенчатой реализации умножителя и ускорения переключения умножителя в спейсер обеспечило разработку 64-разрядного вычислителя, реализующего FMA и FMS операции, соответствующего по производительности современным образцам синхронных аналогов и обладающего всеми преимуществами НЧЗ устройств: полной самопроверяемостью относительно константных неисправностей, сохранением работоспособности при сверхмалых значениях питающих напряжений.

Направлением дальнейших исследований является изучение возможности сокращения длительности спейсерной фазы всех ступеней конвейера DIFMAS за счёт разработки локализованной поразрядной системы ускорения перехода элементов блока в спейсерное состояние и рабочую фазу за счёт разработки более быстрого регистра дополнительной внутренней памяти.

Поддержка

Исследование выполнено при частичной поддержке Программы фундаментальных исследований Прези-

диума РАН № 14 "Исследование инновационных методов автоматизации проектирования СБИС и систем на кристалле на 2018-2020 годы" (проект 0063-2018-0004) в Институте проблем информатики ФИЦ ИУ РАН.

ЛИТЕРАТУРА

[1] R.V.K. Pillai, S.Y.A. Shah, A.J. Al-Khalili, and D. Al-Khalili, Low power floating point MAFs – A comparative study / Sixth International Symposium on Signal Processing and its Applications, Kuala Lumpur, 2001, V. 1.P. 284-287.

[2] P.-M. Seidel, Multiple path IEEE floating-point Fused Multiply-Add / Proc. 46th IEEE International Midwest Symposium on Circuits and Systems, Cairo, Egypt, 2003.P. 1359–1362.

[3] T. M. Bruintjes. Design of a Fused Multiply-Add Floating-Point and Integer Datapath. Master's thesis, University of Twente, Enschede, the Netherlands,2011. 154 p.

[4] J.R. Noche, and J.C. Araneta, An asynchronous IEEE floating-point arithmetic unit / Science Diliman, Philippines. 2007. V.19. No.2.P. 12–22.

[5] R. Manohar, and B.R. Sheikh, Operand-optimized asynchronous floating-point units and method of use therefor, US patent, № 20130124592. May 2013.

[6] Y. Stepchenkov, Y. Diachenko, V. Zakharov, Y. Rogdestvenski, N. Morozov, and D. Stepchenkov, Self-Timed Computing Device for High-Reliable Applications / Proc. International Workshop on power and timing modeling, optimization and simulation (PATMOS'2009), Delft, Netherlands, 2009.P. 276–285.

[7] Соколов И.А., Степченков Ю.А., Рождественский Ю.В., Дьяченко Ю.Г. Самосинхронное устройство умножения-сложения гигафлопсного класса: методологические аспекты // Проблемы разработки перспективных микро- и наноэлектронных систем - 2014. Сб. трудов / под общ. ред. академика РАН А.Л. Стемпковского. М.: ИППМ РАН, 2014. Ч. IV. С. 51-56.

[8] Степченков Ю.А., Рождественский Ю.В., Дьяченко Ю.Г., Морозов Н.В., Степченков Д.Ю., Сурков А.В. Самосинхронное устройство умножения-сложения гигафлопсного класса: варианты реализации // Проблемы разработки перспективных микро- и наноэлектронных систем - 2014. Сб. трудов / под общ. ред. академика РАН А.Л. Стемпковского. М.: ИППМРАН, 2014. Ч. IV. С. 57-60.

[9] Yuri Stepchenkov, Victor Zakharov, Yuri Rogdestvenski, Yuri Diachenko, Nikolai Morozov and Dmitri Stepchenkov. Speed-Independent Fused Multiply Add and Subtract Unit // Proceedings of IEEE EastWest Design & Test Symposium (EWDTS'2016), Yerevan, October, 14 - 17, 2016. P. 150-153.

[10] Ю.А. Степченков, Ю.В. Рождественский, Ю.Г. Дьяченко, Н.В. Морозов, Д.Ю. Степченков, Б.А. Степанов, Д.Ю. Дьяченко, А.В. Рождественскене. Самосинхронное устройство умножения-сложения с плавающей точкой // Проблемы разработки перспективных микро- и наноэлектронных систем - 2016. Сб. трудов / под общ. ред. академика РАН А.Л. Стемпковского. М.: ИППМРАН, 2016. Часть 3. С. 149-156.

[11] Варшавский В.И. и др. Автоматное управление асинхронными процессами в ЭВМ и дискретных системах. М.: Наука, 1986. 400 с.

[12] H. Makino, Y. Nakase, H. Suzuki, H. Morinaka, H. Shinohara, and K. Mashiko, "An 8.8-ns 54x54-bit multiplier with high speed redundant binary architecture" // IEEE Journal of Solid-State Circuits.1996. V. 31. No. 6, pp. 773-783.

[13] Stepchenkov Y.A., Zakharov V.N., Rogdestvenski Y.V., Diachenko Y.G., Morozov N.V., Stepchenkov D.Y. Speed-Independent Floating Point Coprocessor / IEEE Eeast-West Design and Test Symposium, Batumi, Georgia, September 26-29, 2015. P. 111- 114.

[14] Рождественский Ю.В., Морозов Н.В., Рождественскене А.В. Подсистема событийного анализа самосинхронных схем АСПЕКТ // Проблемы разработки перспективных микро- и наноэлектронных систем - 2010. Сб. трудов / под общ. ред. академика А.Л.Стемпковского. М.:ИППМ РАН, 2010. С. 26-31.

[15] S. Galal, and M. Horowitz, Energy-Efficient Floating-Point Unit Design // IEEE Transactions on computers. 2011. V. 60. No.7. P. 913–922.

Delay-Insensitive Floating Point Multiply-Add-Subtract Unit

I.A. Sokolov, Y.V. Rogdestvenski, Y.G. Diachenko, Y.A. Stepchenkov, N.V. Morozov, D.Y. Stepchenkov, D.Y. Diachenko

Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences (IPI FRC CSC RAS), IPI RAS

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

***Abstract* — The subject of this paper is a floating point unit implementing fused multiply-add-subtract operation. It belongs to the delay-insensitive self-timed circuits which do not depend on delays both in cells and on wires. It is fully compliant with IEEE 754 Standard and processes both a sum and difference between product of first two operands and third operand. Each 64-bit input operand contains either one double precision number, or two single precision numbers. Thus presented unit calculates either one operation with double precision numbers, or two simultaneous operations with single precision numbers. Multiplier utilizes modified Booth algorithm. In order to increase its performance, it is divided into two pipeline stages with accelerated forced switching to a spacer phase. Booth encoder circuit is integrated into an input FIFO. FIFO is implemented as a register file with an output multiplexer and read/write address counters. Using ternary redundant self-timed code for multiplying, adding and subtracting provides a reduction of unit's**

complexity. Indication subcircuit considers the constrains imposed by an equichronous zone for chosen fabrication technology. For decreasing energy consumption, the fused multiply-add-subtract unit implements one-channel pipeline. The unit is designed for 65-nm CMOS bulk technology using an industrial standard cell library supplemented by self-timed cells. It provides 3 Gflops performance and 2.9-ns latency.

***Keywords* — redundant coding, ternary adder, Wallace tree, equichronous zone, FIFO.**

REFERENCES

[1] R.V.K. Pillai, S.Y.A. Shah, A.J. Al-Khalili, and D. Al-Khalili, Low power floating point MAFs – A comparative study / Sixth International Symposium on Signal Processing and its Applications, Kuala Lumpur, 2001, V. 1. P. 284-287.

[2] P.-M. Seidel, Multiple path IEEE floating-point Fused Multiply-Add / Proc. 46th IEEE International Midwest Symposium on Circuits and Systems, Cairo, Egypt, 2003. P. 1359–1362.

[3] T. M. Bruintjes. Design of a Fused Multiply-Add Floating-Point and Integer Datapath. Master's thesis, University of Twente, Enschede, the Netherlands,2011. 154 p.

[4] J.R. Noche, and J.C. Araneta, An asynchronous IEEE floating-point arithmetic unit / Science Diliman, Philippines. 2007. V.19. No.2.P. 12–22.

[5] R. Manohar, and B.R. Sheikh, Operand-optimized asynchronous floating-point units and method of use therefor, US patent, № 20130124592. May 2013.

[6] Y. Stepchenkov, Y. Diachenko, V. Zakharov, Y. Rogdestvenski, N. Morozov, and D. Stepchenkov, Self-Timed Computing Device for High-Reliable Applications / Proc. International Workshop on power and timing modeling, optimization and simulation (PATMOS'2009), Delft, Netherlands, 2009.P. 276–285.

[7] Sokolov I.A., Stepchenkov Yu.A., Rozhdestvenskij Yu.V., Diachenko Yu.G. Samosinhronnoe ustroystvo umnojeniya-slojeniya gigaflopsnogo klassa: metodologicheskie aspektyi (Speed-Independent Fused Multiply-Add Unit of Gigaflops Rating: Methodological Aspects) // Sb. trudov "Problemyi razrabotki perspektivnyih mikro- i nanoelektronnyih sistem". M.: IPPM RAN, 2014. Ch. IV. S. 51-56.

[8] Stepchenkov Yu.A., Rozhdestvenskij Yu.V., Diachenko Yu.G., Morozov N.V., Stepchenkov D.Yu., Surkov A.V. Samosinhronnoe ustroystvo umnojeniya-slojeniya gigaflopsnogo klassa: variantyi realizatsii (Speed-Independent Fused Multiply-Add Unit of Gigaflops Rating: Implementation Variants) // Sb. trudov "Problemyi razrabotki perspektivnyih mikro- i nanoelektronnyih sistem". M.: IPPM RAN, 2014. Ch. IV. S. 57-60.

[9] Yuri Stepchenkov, Victor Zakharov, Yuri Rogdestvenski, Yuri Diachenko, Nikolai Morozov and Dmitri Stepchenkov. Speed-Independent Fused Multiply Add and Subtract Unit // Proceedings of IEEE EastWest Design & Test Symposium (EWDTS'2016), Yerevan, October, 14 - 17, 2016. P. 150-153.

[10] Stepchenkov Yu.A., Rozhdestvenskij Yu.V., DiachenkoYu.G., Morozov N.V., Stepchenkov D.Yu., Stepanov B.A., Diachenko D.Y., Rozhdestvenskene A.V. Samosinhronnoe ustroystvo umnojeniya-slojeniya s plavayuschey tochkoy (Self-Timed Floating Point Multiply-Add Unit) // Sb. trudov "Problemyi razrabotki perspektivnyih mikro- i nanoelektronnyih sistem". M.: IPPM RAN, 2016. Ch 3. S. 149-156.

[11] Varshvskij V.I. i dr. Avtomatnoe upravlenie asinhronnyimi protsessami v EVM i diskretnyih sistemah (Automatic control of the asynchronic processes in the computers and discrete systems). M.: Nauka, 1986. 400 s.

[12] H. Makino, Y. Nakase, H. Suzuki, H. Morinaka, H. Shinohara, and K. Mashiko, "An 8.8-ns 54x54-bit multiplier with high speed redundant binary architecture" // IEEE Journal of Solid-State Circuits.1996. V. 31. No. 6, pp. 773-783.

[13] Stepchenkov Y.A., Zakharov V.N., Rogdestvenski Y.V., Diachenko Y.G., Morozov N.V., Stepchenkov D.Y. Speed-Independent Floating Point Coprocessor / IEEE Eeast-West Design and Test Symposium, Batumi, Georgia, September 26-29, 2015. P. 111- 114.

[14] Rozhdestvenskij Yu.V., Morozov N.V., Rozhdestvenskene A.V. Podsistema sobyitiynogo analiza samosinhronnyih shem ASPEKT (ASPECT – a Subsystem of Event Analysis of Self-Timed Circuits) // Sb. trudov "Problemyi razrabotki perspektivnyih mikro- i nanoelektronnyih sistem". M., IPPM RAN, 2010. S. 26-31.

[15] S. Galal, and M. Horowitz, Energy-Efficient Floating-Point Unit Design // IEEE Transactions on computers. 2011. V. 60. No.7. P. 913–922.