УДК 681.325.57 DOI: 10.31114/2078-7707-2022-4-109-116

Быстродействующие умножители для аппаратной реализации искусственных нейронных сетей

С.Э. Миронов, О.И. Буренева, К.М. Зибарев

Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина), semironovspb@yandex.ru

***Аннотация* — В основе функционирования сверточных нейронных сетей (СНС) лежит операция умножения вектора на матрицу, в связи с чем для построения производительных СНС требуется разработка быстродействующих вычислителей. Один из путей проектирования таких устройств связан с аппаратной реализацией алгоритмов быстрого умножения, в частности алгоритмов умножения на группу разрядов (алгоритмы Бута, Мак-Сорли и др.). Полученные матричные структуры могут быть оптимизированы при разработке топологии с целью минимизации площади кристалла. В статье рассматриваются варианты ускорения работы умножителей с использованием методов умножения на группу разрядов, предложены варианты топологических реализаций рассмотренных решений, позволяющие достичь компромисса между быстродействием и площадью кристалла.**

***Ключевые слова*— элементы нейронных сетей, аппаратная реализация нейронных сетей, матричные умножители, умножение на группу разрядов, алгоритм Бута.**

I. Введение

Концепция построения и функционирования искусственных нейронных сетей (ИНС) является основой для решения разнообразных задач: классификации, аппроксимации, моделирования, распознавания, кластеризации и др.Чаще всего реализуются сверточные нейронные сети, в которых результат генерируется путем свертки входных матриц с использованием фильтров. Повышение производительности и точности сети обычно связано с увеличением ее сложности. В [1] показано, что классификация изображения размером 227 × 227 пикселей требует миллиардов арифметических операций умножения и сложения. Сеть с архитектурой MCN-MobileNet требует 0,58 миллиарда операций для классификации изображения, а модель VGG-19 большого размера требует около 20 миллиардов операций на классификацию. Решение проблемы повышения производительности ИНС связано с применением параллельных вычислительных структур, реализованных на специализированной аппаратной базе: графических процессорах, программируемых логических интегральных схемах, заказных БИС [2].Аппаратная реализация ИНС актуальна и при решении задач, которые требуют автономной работы устройств (камеры видеонаблюдения, автономное вождение, смартфоны и т.д.) с поддержкой методов искусственного интеллекта [3]-[4].

В основе функционирования сверточных нейронных сетей лежит операция умножения вектора на матрицу, реализуемая в каждом слое[5]. Полученный в результате умножения вектор используется при вычислении функции активации, в которой также в большинстве случаев требуется операция умножения. Таким образом, повышение производительности ИНС непосредственно связано с разработкой аппаратных быстродействующих умножителей, что может быть достигнуто за счет применения оригинальных аппаратных решений [6]-[8] или быстрых алгоритмов арифметических операций [9]-[12].

Быстродействующие (с малой задержкой) и высокопроизводительные (с высокой тактовой частотой) умножители могут быть построены на основе матричного подхода к реализации. Он позволяет значительно увеличить быстродействие за счет того, что каждая из элементарных операций реализуемого алгоритма выполняется на «персональном» блоке. Применительно к операции «умножение» такими блоками являются сумматоры, собранные в матрицу и используемые для сложения частичных произведений, каждое из которых является простым кратным множимого. Количество сумматоров в матрице пропорционально разрядности множителя. Очевидно, что такое решение отличается значительными аппаратными затратами, выражающимися в площади, занимаемой схемой на кристалле. Однако площадь такого матричного умножителя (МУ) может быть значительно снижена благодаря использованию методов группировки разрядов множителя.

В статье рассматриваются варианты аппаратной реализации быстрого умножения, основанного на умножении одновременно на группу разрядов. Эти устройства очень часто используются для реализации систем. Особая архитектура этих компонентов заставляет разработчика использовать методы синтеза, несколько отличающиеся от тех, которые применяются для ASICs (интегральных схем, ориентированных на конкретные приложения), для которых существуют стандартные библиотеки ячеек.

В данной статье приводятся результаты исследований авторов, посвященных вопросам оптимизации параметров устройств умножения, выполнявшихся в рамках проводимых в СПбГЭТУ «ЛЭТИ» работ в области аппаратных решений для искусственного интеллекта.

II. Методы ускорения умножения

Время умножения пропорционально числу сложений частичных произведений, поэтому одной из важных задач проектирования является уменьшение числа сложений с целью повышения быстродействия.

Это достигается двумя способами:

1) распараллеливанием процесса сложения;
2) уменьшением числа слагаемых.

A. *Распараллеливание процесса сложения*

Распараллеливание процесса сложения частичных произведений предполагает:

- группировку частичных произведений,
- их сложение в каждой группе,
- сложение результатов, полученных в группах.

На рис. 1 приведены фрагменты структурных схем умножителей, матрицы которых разбиты на две параллельно работающие подматрицы. В одном случае (рис. 1, *а*) они топологически совмещены (их строки чередуются, как слои в пироге), а во втором (рис. 1, *б*) топологически разделены. Квадратами на схеме изображаются сумматоры, а точками – элементы умножения (элементы «2И» или мультиплексоры).

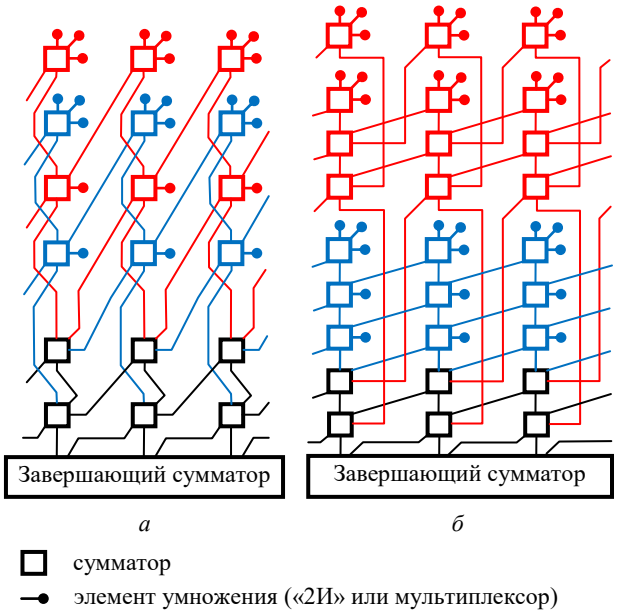


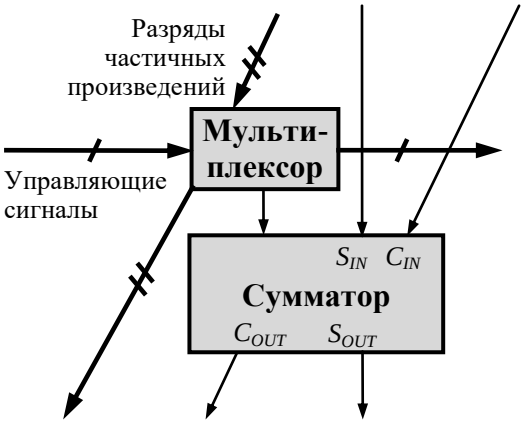
Рис. 1.Фрагменты матриц умножителей, с топологически совмещенными (а) и с топологически разделенными подматрицами (б)

Матрица может быть разделена и на большее число подматриц. Это уменьшит глубину схемы (количество строк сумматоров, через которые сигналы проходят от входа в матрицу до ее выхода). Однако при увеличении количества подматриц, увеличивается и число связей между ними и длина связей. Это приводит к увеличению площади, занимаемой схемой на кристалле.

B. *Уменьшение числа слагаемых*

Уменьшения числа слагаемых в процессе суммирования частичных произведений добиваются, умножая множимое *A* не на один разряд, а на группу разрядов множителя *B* одновременно.

В этом случае в базовой ячейке матрицы МУ, состоящей из сумматора и элемента умножения (рис. 2), последний будет не элементом «2И», а мультиплексором. В зависимости от значения управляющих сигналов (формируемых на основе группы разрядов множителя) он передает на вход сумматора разряд одного из кратных множимому частичных произведений.



S_{IN}иS_{OUT} – входной и выходной сигналы суммы

C_{IN}и C_{OUT} – входной и выходной сигналы переноса

Рис. 2. Базовая ячейка матричного умножителя с группировкой разрядов множителя

Увеличение сложности элемента умножения при группировке разрядов множителя (по сравнению с умножением на один разряд) с лихвой компенсируется кратным уменьшением числа строк устройства.

C. *Результаты анализа методов ускорения умножения*

На основании изложенных в этом разделе материалов можно сделать следующие выводы.

1) Распараллеливание процесса сложения частичных произведений приводит к существенному повышению быстродействия МУ. Однако при этом в связи с появлением протяженных межматричных шин существенно возрастают аппаратные затраты и снижается регулярность устройства.
2) Матричные умножители с группировкой разрядов множителя обладают большей однородностью и регулярностью, чем умножители с разбиением матриц на части, что упрощает их проектирование. Кроме того умножение на группу разрядов множителя позволяет существенно сократить число строк в матрице, что обеспечивает снижение аппаратных затрат.
3) Рассмотренные способы ускорения умножения могут применяться совместно для достижения большего быстродействия.

В заключение этого раздела необходимо отметить, что, если матрицы умножителей можно разбивать теоретически на любое число параллельно работающих частей, то число разрядов множителя, на которые осуществляется умножение, не превышает двух.

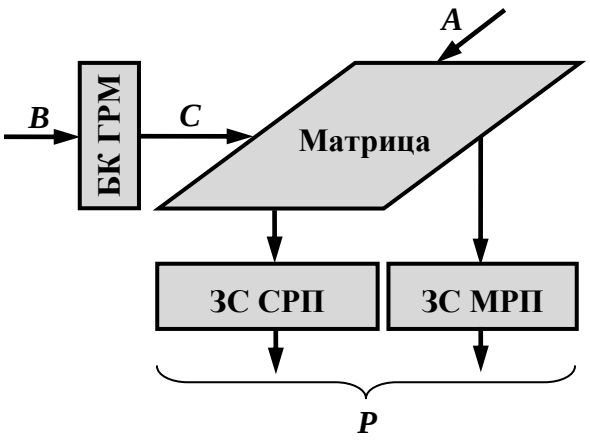
Однако в следующем разделе, посвященном детальному рассмотрению вопросов группировки разрядов множителя, будет предложена область применения этого подхода практически без ограничений на число разрядов в группе.

III. Умножение с группировкой разрядов множителя

A. *Структура матричного умножителя с группировкой разрядов множителя*

Из материалов предыдущего раздела следует, что использование алгоритмов умножения с группировкой разрядов множителя приводит к изменению структуры матричного умножителя по сравнению с традиционным его вариантом (умножение на один разряд множителя).

Поясним это с помощью структурной схемы матричного умножителя с группировкой разрядов множителя, приведенной на рис. 3.



A,***B***, ***P***– множимое, множитель, произведение

C– сигналы управления мультиплексорами

БК ГРМ – блоки кодирования групп разрядов множителя

ЗС СРП – завершающий сумматорстарших разрядов произведения

ЗС МРП – завершающий сумматормладших разрядов произведения

Рис. 3. Структура матричного умножителя с группировкой разрядов множителя

Очевидно, что при умножении сразу на несколько разрядов число строк в матрице умножителя сокращается пропорционально числу разрядов в группе.

При этом строки матрицы удлиняются на число ячеек, равное числу разрядов *K* в группе и оказываются сдвинутыми друг относительно друга уже не на один разряд, а на *K* разрядов.

Это в свою очередь приводит к тому, что на выходе правой части матрицы формируются не младшие разряды результата, а двухрядный код, представляющий собой множество выходных сигналов суммы и переноса, для сложения которых в схему вводят завершающий сумматор младших разрядов произведения.

В связи с необходимостью управления работой мультиплексоров в состав МУ с группировкой разрядов вводятся блоки кодирования групп разрядов множителя.

Существует целый ряд алгоритмов кодирования групп разрядов (Бута [10], [11], Мак-Сорли [12]), ускоряющих процесс умножения.

Пример топологии матричного умножителя Бута, соответствующей приведенной выше структурной схеме, представлен на рис. 4. Топология 16-разрядного матричного умножителя Бута [**7**] получена методом программной генерации с помощью разработанной на кафедре вычислительной техники СПбГЭТУ «ЛЭТИ» системы иерархического сжатия топологии «*Matching of cell*» [13].

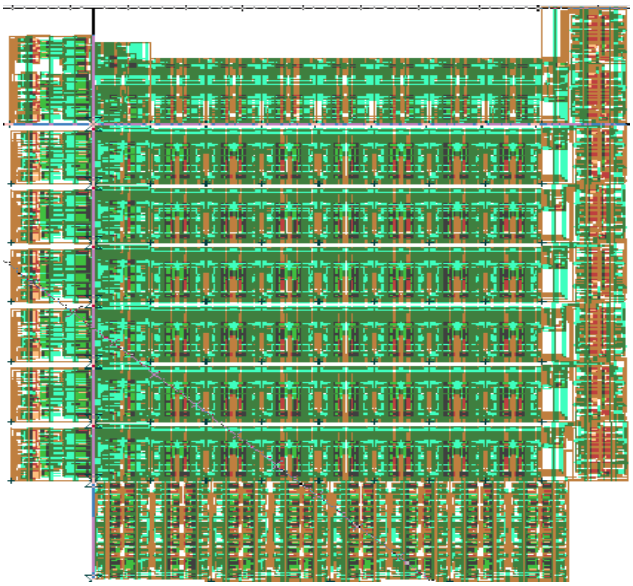


Рис. 4. Топология 16-разрядного матричного умножителя Бута

Принцип работы алгоритмов умножения с группировкой разрядов множителя поясним на примере наиболее известного и широко применяемого на практике метода группировки пар разрядов множителя в соответствии с алгоритмом **Бута** [**10**], [**11**].

B. *Алгоритм умножения Бута*

В алгоритме умножения Бута значения пары разрядов множителя *B* суммируются с переносом из соседней с ними справа (младшей) пары разрядов. В зависимости от полученной величины анализируемой паре ставятся в соответствие перенос в старшую по отношению к ней пару и так называемое «число со знаком», задающее коэффициент кратности частичного произведения множимому ***A***.

В табл. 1 представлено кодирование пары разрядов в соответствии с модифицированным вариантом алгоритма, в котором перенос из пары разрядов совпадает со старшим разрядом пары.

Таблица 1

Кодирование пары разрядов в алгоритме умножения Бута

B_n	B_{n+1}	C_{in}	$\sum$	C_{out}	Число со знаком
Вес					
2	1	1	1	4	1
0	0	0	0	0	0
0	0	1	1	0	+1
0	1	0	1	0	+1
0	1	1	2	0	+2
1	0	0	2	1	-2
1	0	1	3	1	-1
1	1	0	3	1	-1
1	1	1	4	1	0

Возможные варианты частичных произведений подаются на информационные входы мультиплексоров с прямых и инверсных выводов регистра множимого со сдвигом влево или без сдвига в зависимости от абсолютного значения чисел со знаком.

Однако при умножении на большее двух число разрядов ситуация принципиально изменяется в связи с тем, что числа со знаком перестают быть степенями двойки, и частичные произведения уже не могут быть сняты непосредственно с регистра множимого.

В табл. 2 это иллюстрируется на примере варианта группировки трех разрядов множителя.

Как видно из приведенного в табл. 2 примера кодирования, числа со знаком принимают целые значения из диапазона от «-2^{K-1}» до «$+2^{K-1}$», где K - число разрядов в группе.

Таким образом, при умножении одновременно более чем на два разряда, частичные произведения с коэффициентами кратности множимому, не равными степеням двойки, должны быть оперативно подготовлены с помощью быстродействующих многоразрядных сумматоров, которые будут вносить существенный вклад и в задержку, и в площадь устройства.

Такой алгоритм может найти применение в задачах фильтрации, где один из операндов неизменен (по крайней мере, в течение продолжительного времени).

Однако применение этого алгоритма возможно и в столь востребованных сейчас системах искусственного интеллекта при решении задачи распознавания.

Таблица 2

Кодирование трех разрядов в алгоритме умножения Бута

B_n	B_{n+1}	B_{n+2}	C_{in}	$\sum$	C_{out}	Число со знаком
Вес						
4	2	1	1	1	8	1
0	0	0	0	0	0	0
0	0	0	1	1	0	+1
0	0	1	0	1	0	+1
0	0	1	1	2	0	+2
0	1	0	0	2	0	+2
0	1	0	1	3	0	+3
0	1	1	0	3	0	+3
0	1	1	1	4	0	+4
1	0	0	0	4	1	-4
1	0	0	1	5	1	-3
1	0	1	0	5	1	-3
1	0	1	1	6	1	-2
1	1	0	0	6	1	-2
1	1	0	1	7	1	-1
1	1	1	0	7	1	-1
1	1	1	1	8	1	0

Более детальному рассмотрению вопросов построения быстродействующих матричных умножителей с группировкой разрядов будет посвящен следующий раздел.

IV. Аппаратная реализация алгоритма умножения с группировкой разрядов множителя

Как уже было сказано выше, основными компонентами базовых ячеек МУ с группировкой разрядов множителя (рис. 2) являются сумматоры и мультиплексоры.

В качестве примера в работе используется популярное схемотехническое решение полного одноразрядного сумматора (рис. 5), впервые описанное Хэмплом (*Hampel*) [14].

Очевидно, что внутренние связи ячеек следует реализовывать в топологических слоях, наиболее близких к транзисторам, а именно в слоях поликремния и нижнего металла.

В таком случае, чтобы не мешать внутренней разводке, шины «земли» и «питания» (с которых нулевой и единичный сигналы должны поступать не только на сами транзисторы, но и на расположенные между областями транзисторов разного типа контакты к карма-

нам и к подложке) должны организовываться в слое верхнего металла.

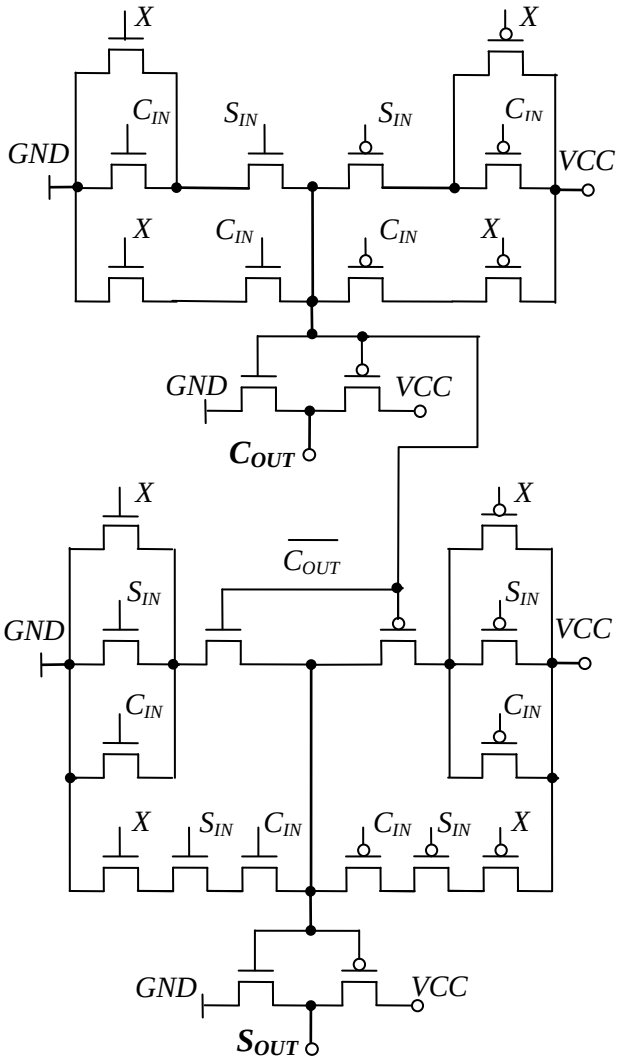


Рис. 5. Схема полного одноразрядного сумматора Хэмпла

По этой же причине в слое верхнего металла над столбцами ячеек должны проводиться и другие шины:

- шины разрядов кратных множимого, с которых данные поступают на информационные входы мультиплексоров;
- шины сигналов сумм и переносов, проходящие между сумматорами разных строк над мультиплексорами.

Поскольку перечисленные шины проводятся через матрицу вертикально, то вертикальными должны быть и шины «земли» и «питания».

В связи с этим горизонтальные шины выходных сигналов блоков кодирования множителя, управляющие работой мультиплексоров, должны идти через строки матрицы в слое нижнего металла.

Во избежание роста аппаратных затрат мультиплексоры следует строить на одно-транзисторных ключах. Это позволит сократить, как число транзисторов, так и число шин нижнего металла, подающих на *n*-транзисторные ключи мультиплексоров управляющие сигналы.

В качестве иллюстрации на рис. 6 приведена схема мультиплексора «(2^K+1) в 1» *I*-й ячейки строки МУ. Как видно из рисунка, для формирования *I*-го разряда частичного произведения на информационные входы мультиплексора подаются *K* разрядов в соответствии с принципами кодирования в алгоритме умножения Бута (иллюстрированными табл. 1 и 2):

- прямые и инверсные значения (по 2^K шинам) *I*-ых разрядов кратных множимого с коэффициентами кратности от 2^{K-1}до 1 ($[2^{K-1}\times A]_I$, $[(2^{K-1}-1)\times A]_I$, $[(2^{K-1}-2)\times A]_I$, ..., $[2\times A]_I$);
- напряжение питания *VCC*, после инвертирования преобразующееся в «0»– разряд числа $[0\times A]$.

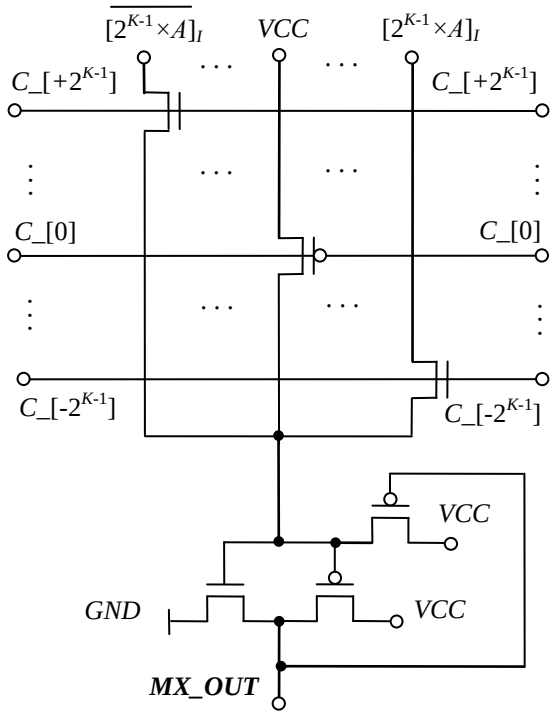


Рис. 6. Схема мультиплексора «(2^K+1) в 1»

На управляющие же входы мультиплексора подаются соответствующие числам со знаком (табл. 1 и 2) управляющие сигналы «$C\_[+2^{K-1}]$», «$C\_[+2^{K-1}-1]$», …, «$C\_[0]$»,…,«$C\_[-2^{K-1}+1]$»,«$C\_[-2^{K-1}]$».

Топология ячейки состоящей из сумматора и мультиплексора «9 в 1» для МУ Бута с группировкой 3-х разрядов множителя приведена на рис. 7. Она была получена методом программной генерации с помощью оригинальных средств проектирования, включающих в себя средства генерации топологии по электрической схеме [15] и систему иерархического сжатия топологии [13].

Высота этой ячейки превышает высоту сумматора с ускоренным манчестерским переносом. Это снимает

проблему быстрого сложения чисел на выходах младших K разрядов строк при формировании младших разрядов произведения, так как позволяет вписать сумматор с ускоренным переносом в вертикальный габарит строки умножителя.

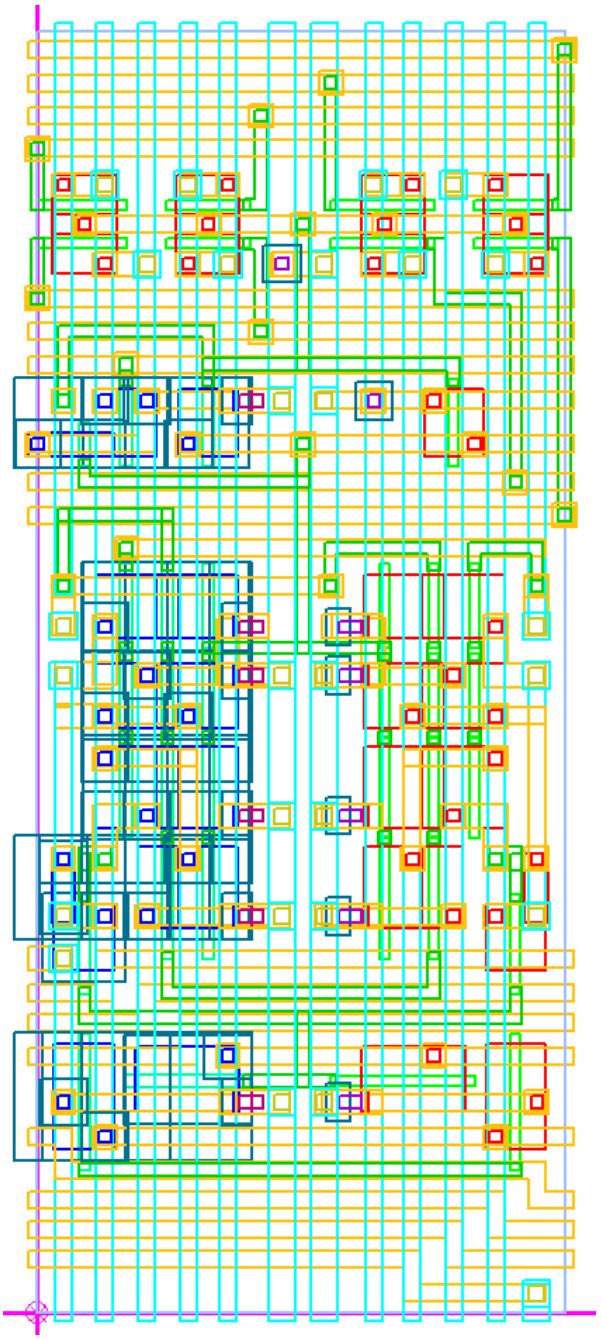


Рис. 7. Топология ячейки матричного умножителя Бута с группировкой 3-х разрядов множителя

Необходимо отметить, что увеличение разрядности групп K приведет к необходимости расширения ячейки, так как число проходящих через нее вертикальных шин второго металла, определяемое выражением (2^K+4)(2 шины земли и питания, 2 шины суммы и переноса, 2^K шин разрядов кратных множимого), даже при K=4 будет равно уже 20. То есть горизонтальный габарит ячейки возрастет в 1,5 раза (с 13 шагов координатной сетки до 21 шага).

Следовательно, во избежание возникновения в ячейке пустот, топология сумматора должна быть «перепланирована» – расширена до габаритов мультиплексора «(2^K+1) в 1». Но даже в этом случае площадь последнего в значительной степени (более чем на половину) будет занята горизонтальными шинами управляющих сигналов.

И хотя при этом не стоит забывать о существенном снижении площади всего умножителя вследствие сокращения числа его строк, тем не менее, уже при K=5 число вертикальных шин будет равно36. А это значит, что горизонтальный габарит ячейки (37шагов координатной сетки) превысит длину полосы из 28 транзисторов (14 n-типа и 14p-типа) сумматора Хэмпла.

Выходом из этой ситуации, позволяющим расширить диапазон значений разрядности групп K, может стать использование мультиплексора «(2^K+1) в 1» с управлением инвертированием. Принцип его схемотехнической реализации иллюстрируется рис. 8.

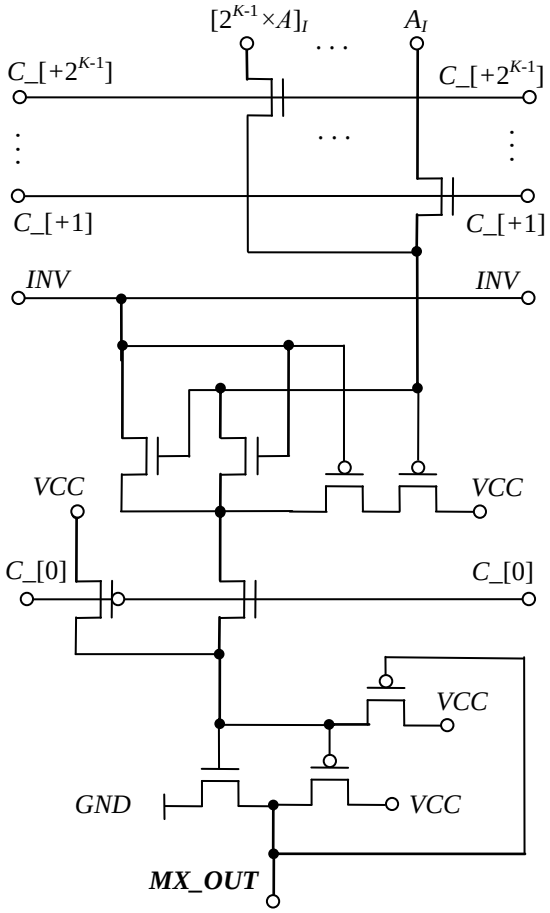


Рис. 8. Схема мультиплексора «(2K+1) в 1» с управлением инвертированием

Введение в схему управляемого инвертора (сумматора по модулю два, управляемого сигналом инвертирования «*INV*») позволяет уменьшить число шин кратных множимого (и соответственно число горизонталь-

ных шин управляющих сигналов мультиплексора) по сравнению с ранее рассмотренным вариантом в 2 раза.

V. Заключение

В ходе исследования проблемы аппаратной реализации искусственных нейронных сетей были рассмотрены варианты построения быстродействующих матричных умножителей. Учет особенностей их функционирования (в частности при решении задачи распознавания) позволил предложить эффективный вариант их построения на основе алгоритмов умножения с группировкой большего числа разрядов множителя, чем в традиционном варианте алгоритма Бута.

Разработанные варианты схемо-топологических решений (в том числе с применением управляемого инвертирования разрядов кратных множимого) позволяют за счет уменьшения числа строк матрицы снизить как аппаратные затраты, так и время задержки.

В настоящий момент авторами ведутся работы по созданию средств кремниевой компиляции матричных умножителей с использованием оригинальных средств иерархического сжатия топологии.

Литература

[1] RajuMachupalli, MasumHossain, MrinalMandal. Review of ASIC accelerators for deep neural network // Microprocessors and Microsystems. 2022. V. 89.doi: 10.1016/j.micpro.2022.104441

[2] NurvitadhiE., Jaewoong Sim, SheffieldD., MishraA., Krishnan S., MarrD. Accelerating recurrent neural networks in analytics servers: Comparison of FPGA, CPU, GPU, and ASIC // 26th International Conference on Field Programmable Logic and Applications (FPL).2016. P. 1-4. doi: 10.1109/FPL.2016.7577314

[3] Кустов А.Г., Соловьев Р.А., Стемпковский А.Л., Тельпухов Д.В. Аппаратная реализация нейронной сети для обнаружения объектов в базисе ПЛИС // Проблемы разработки перспективных микро- и наноэлектронных систем (МЭС). 2022. Вып. 1. С. 27-34. doi:10.31114/2078-7707-2022-1-27-34

[4] Вирясова А.Ю., Климов Д.И., Хромов О.Е., Губайдуллин И.Р., Орешко В.В. Анализ возможности применения сверточных нейронных сетей и их аппаратной реализации для задачи термо-видеотелеметрии // Радиотехника. 2021. № 9. С. 115–126.doi:10.18127/j00338486-202109-11

[5] Слюсар В.И. Модели нейросетей на основе тензорно-матричной теории // Проблемы разработки перспективных микро- и наноэлектронных систем (МЭС). 2021. Вып. 2. С. 23-28. doi:10.31114/2078-7707-2021-2-23-28

[6] Якунин А.Н., Аунг Мьо Сан. Повышение скорости работы многоразрядного двоичного умножителя // Проблемы разработки перспективных микро- и наноэлектронных систем. 2018. Вып. 2. С. 149-155. doi:10.31114/2078-7707-2018-2-149-155

[7] Mironov S.E., Bureneva O.I., Milakin A.D. Analysis of Multiplier Architectures for Neural Networks Hardware Implementation // III International Conference on Neural Networks and Neurotechnologies (NeuroNT). 2022. doi:10.1109/NeuroNT55429.2022.9805564

[8] Camus V., Mei L., Enz C., Verhelst M. Review and Benchmarking of Precision-Scalable Multiply-Accumulate Unit Architectures for Embedded Neural-Network Processing // IEEE Journal on Emerging and Selected Topics in Circuits and Systems.2019.V. 9.№ 4.P. 697-711. doi: 10.1109/JETCAS.2019.2950386.

[9] Пасынков С.В., Ильясов Р.Ф. Оценка использования систолических массивов при реализации алгоритмов умножения матриц на ПЛИС // Проблемы разработки перспективных микро- и наноэлектронных систем (МЭС). 2021. Вып. 3. С. 76-80. doi:10.31114/2078-7707-2021-3-76-80

[10] Rooban S., Nagesh M., Prasanna M. V. S. L., Rayudu K., Sai G. D. Implementation of 128-bit Radix-4 Booth Multiplier // International Conference on Computer Communication and Informatics (ICCCI). 2021.P. 1-7. doi: 10.1109/ICCCI50826.2021.9457004

[11] Jean-Pierre Deschamps, GeryJean Antoine Bioul, Gustavo D. Sutter.Synthesis of arithmetic circuits: FPGA, ASIC and embedded systems. John Wiley & Sons, Inc.. Hoboken. New Jersey.2006. 556 p.

[12] MacSorley O. L. High-Speed Arithmetic in Binary Computers // Proceedings of the IRE. 1961. V. 49.№ 1. P. 67-91. doi: 10.1109/JRPROC.1961.287779.

[13] Mironov S.E., Vasiliyev A.Yu., Safyannikov N.M. Means Of Automating The Hierarchical Design Of Complex Microelectronic Circuits With Uncertainty Of Design Rules // Problems of advanced micro- and nanoelectronic systems development (MES). SELECTED ARTICLES of the VIII All-Russia Science&Technology Conference MES-2018. Moscow.: – FSFIS Institute for Design Problems in Microelectronics RAS. 2019. P. 7-13. DOI: 10.31114/2078-7707-2019-1-7-13.

[14] 18. Hample D., McGuire K.E., Prost K.J. CMOS/SOS serial-parallel multiplier // IEEE journal of solid-state circuits, 1975. – V. SC-10. – № 5. – P. 307-314.

[15] Миронов С.Э., Андреев Л.Е., Зибарев К.М. Технология комплексной параметризации топологических проектов регулярных макроблоков СБИС // Проблемы разработки перспективных микро- и наноэлектронных систем (МЭС). 2020. Вып.3. С. 35-40. DOI: 10.31114/2078-7707-2020-3-35-40

Fast Multipliers for Hardware Implementation of Artificial Neural Networks

S.E. Mironov, O.I. Bureneva, K.M. Zibarev

Saint Petersburg Electrotechnical University «LETI», Saint Petersburg, semironovspb@yandex.ru

Abstract — The article is devoted to the structural, circuit and layout design of matrix multipliers, taking into account the peculiarities of their functioning in artificial neural networks. Taking into account the specifics of calculations performed by multipliers makes it possible to reduce their delay and area on a chip.

Purpose. Development and research of high-speed equipment for performing the multiplication operation for artificial neural networks.

Methods. Ensuring high performance and reducing hardware costs achieved by using the method of grouping multiplier digits. The layout was developed in a technologically invariant concept using original software tools based on combinatorial transistor placement methods and layout compaction methods.

Results. An analysis of the features of artificial neural networks functioning made it possible to propose the implementation of a matrix multiplier with grouping of more than two digits of the multiplier (in contrast to traditional multiplication options). The proposed implementation option can also be used in filtering problems.

Discussion. The structure and scheme of a matrix multiplier with grouping of the multiplier bits are described. The problem of the efficiency of multiplier cells layout designing with a grouping of a large number of multiplier bits is discussed.

***Keywords* — neural networks elements, neural networks hardware implementation, matrix multipliers, multiplication by a group of digits, Booth's algorithm.**

REFERENCES

[1] RajuMachupalli, MasumHossain, MrinalMandal. Review of ASIC accelerators for deep neural network // Microprocessors and Microsystems. 2022. Vol. 89.doi: 10.1016/j.micpro.2022.104441

[2] NurvitadhiE., Jaewoong Sim, SheffieldD., MishraA., Krishnan S., MarrD. Accelerating recurrent neural networks in analytics servers: Comparison of FPGA, CPU, GPU, and ASIC // 26th International Conference on Field Programmable Logic and Applications (FPL). 2016. P. 1-4. doi: 10.1109/FPL.2016.7577314

[3] Kustov A.G., Solovyev R.A., Stempkovsky A.L., Telpukhov D.V. Hardware Implementation of a Neural Network for Object Detection in FPGA // Problems of Perspective Micro- and Nanoelectronic Systems Development - 2022. Issue 1. P. 27-34. doi:10.31114/2078-7707-2022-1-27-34

[4] Viryasova A.YU., Klimov D.I., Hromov O.E., Gubajdullin I.R., Oreshko V.V. Analiz vozmozhnosti primeneniya svertochnyh nejronnyh setej i ih apparatnoj realizacii dlya zadachi termo-videotelemetrii (Rich feature hierarchies for accurate object detection and semantic segmentation) // Radiotekhnika. 2021. № 9. S. 115–126. doi:10.18127/j00338486-202109-11

[5] Slyusar V.I. Neural Networks Models based on the tensor-matrix theory // Problems of Perspective Micro- and Nanoelectronic Systems Development - 2021. Issue 2. P. 23-28. doi:10.31114/2078-7707-2021-2-23-28

[6] Yakunin A.N., Aung Myo San Increasing the Speed of a Multi-bit Binary Multiplier // Problems of Perspective Micro- and Nanoelectronic Systems Development - 2018. Issue 2. P. 149-155. doi:10.31114/2078-7707-2018-2-149-155

[7] MironovS.E., BurenevaO.I., MilakinA.D.Analysis of Multiplier Architectures for Neural Networks Hardware Implementation // III International Conference on Neural Networks and Neurotechnologies (NeuroNT). 2022. doi:10.1109/NeuroNT55429.2022.9805564

[8] Camus V., Mei L., Enz C., Verhelst M. Review and Benchmarking of Precision-Scalable Multiply-Accumulate Unit Architectures for Embedded Neural-Network Processing // IEEE Journal on Emerging and Selected Topics in Circuits and Systems. 2019,Vol. 9,no. 4,pp. 697-711. doi: 10.1109/JETCAS.2019.2950386.

[9] Pasynkov S.V., Iliasov R.F. Evaluation of the use of systolic arrays in the implementation of matrix multiplication algorithms on FPGAs // Problems of Perspective Micro- and Nanoelectronic Systems Development - 2021. Issue 3. P. 76-80. doi:10.31114/2078-7707-2021-3-76-80

[10] Rooban S., Nagesh M., Prasanna M. V. S. L., Rayudu K., Sai G. D. Implementation of 128-bit Radix-4 Booth Multiplier // International Conference on Computer Communication and Informatics (ICCCI). 2021.P. 1-7. doi: 10.1109/ICCCI50826.2021.9457004

[11] Jean-Pierre Deschamps, GeryJean Antoine Bioul, Gustavo D. Sutter.Synthesis of arithmetic circuits: FPGA, ASIC and embedded systems. John Wiley & Sons, Inc., Hoboken, New Jersey, 2006. 556 p.

[12] MacSorley O. L. High-Speed Arithmetic in Binary Computers // Proceedings of the IRE. 1961. Vol. 49, no. 1, pp. 67-91. doi: 10.1109/JRPROC.1961.287779.

[13] Mironov S.E., Vasiliyev A.Yu., Safyannikov N.M. Means Of Automating The Hierarchical Design Of Complex Microelectronic Circuits With Uncertainty Of Design Rules // Problems of advanced micro- and nanoelectronic systems development (MES). SELECTED ARTICLES of the VIII All-Russia Science&Technology Conference MES-2018. Moscow.: – FSFIS Institute for Design Problems in Microelectronics RAS. 2019. P. 7-13. DOI: 10.31114/2078-7707-2019-1-7-13.

[14] Hample D., McGuire K.E., Prost K.J. CMOS/SOS serial-parallel multiplier // IEEE journal of solid-state circuits, 1975. – V. SC-10. – № 5. – P. 307-314.

[15] Mironov S.E., Andreev L.E., Zibarev K.M. Complex parameterization technology for topological projects of regular VLSI macroblocks // Problems of Perspective Micro- and Nanoelectronic Systems Development - 2020. Issue 3. P. 35-40. doi:10.31114/2078-7707-2020-3-35-40