ВЫСОКОПРОИЗВОДИТЕЛЬНЫЕ ДИСКРЕТНЫЕ ПРЕОБРАЗОВАНИЯ НА ПРОЦЕССОРАХ NEUROMATRIX С ЯДРОМ NMC3

***Сергей Викторович Мушкаев**, ведущий инженер-программист*
Тел. +7(495)531-30-80 (225), e-mail: mushkaev@module.ru
ЗАО НТЦ «Модуль»
http://www.module.ru

В данной статье рассматриваются вопросы оценки производительности векторных вычислений в процессорах NeuroMatrix на примере дискретного Фурье, Уолша-Адамара, косинусного и вейвлет преобразования. Разбирается принцип построение оптимальных алгоритмов. Демонстрируется эффективность векторных вычислений на DSP процессорах NeuroMatrix.

Ключевые слова: быстрое преобразование, Фурье, Уолша-Адамара, вейвлет, косинусное, NMC3, NeuroMatrix, NM6406, 1879ВМ5, К1879ХК1Я, К1879ХБ1Я, МВ7707

Введение.

Платформа NeuroMatrix®

В задачах проектирования сложных аппаратно-программных систем, связанных с цифровой обработкой сигналов и изображений, важным аспектом является выбор аппаратной платформы. Решающим фактором, как правило, является скорость наиболее критичных функций. При этом не всегда весь требуемой набор функций присутствует в составе библиотек поддержки. На примере широко известных дискретных преобразований в данной статье рассматривается достаточно простой принцип расчета производительности любых векторизуемых алгоритмов для ядра NMC3.


С.В. Мушкаев

Фирмой ЗАО НТЦ Модуль разработаны следующие процессоры с ядром NMC3 (NeuroMatrix Core)[1]: NM6405, NM6406 (1879ВМ5), СБИС К1879ХК1Я и К1879ХБ1Я. Главным вычислительным узлом NMC3 ядра является матрично-векторный сопроцессор, осуществляющий арифметические и логические операции над 64-разрядными векторами данных произвольной разрядности. Элементарной 1-тактовой операцией для векторного сопроцессора может являться матричное преобразование входного 64-р слова. Это преобразование в зависимости от конфигурации разрядностей упакованных данных в векторе может включать от 1 до 2048 умножений с накоплением. Данный факт заставляет принципиально пересматривать стандартные быстрые алгоритмы для достижения максимальной производительности.

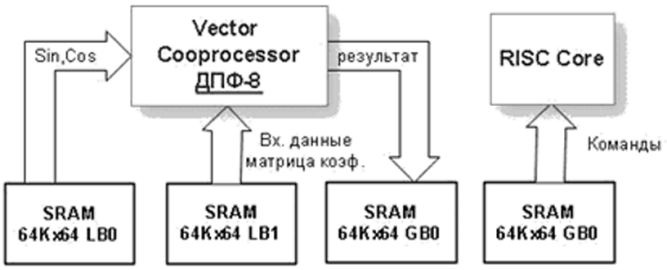

Рис. 1 Организации параллельных потоков команд и данных при вычислении ДПФ-8

Основной отличительной особенностью архитектуры NMC3 от архитектуры предыдущего поколения NMC является ускоренная загрузка весов в матричный умножитель, наличие 4 банков внутренней памяти и 6 внутренних шин ядра, благодаря которым возможно осуществить до шести операций ввода-вывода данных за один такт. В пересмотре на задачу с БПФ, где базисной единицей будет уже не «бабочка 2x2», а ДПФ-8 , это означает, что поток входных данных сигнала (веса для векторного умножителя), поток комплексных коэффициентов (sin и cos), выходной поток и поток команд будут направляться по независимым шинам и при использовании разных банков памяти :LB0,LB1,GB0,GB1 (см. рисунок 1) будут работать параллельно.

Т.к. данные на проходе обрабатываются векторным сопроцессором с темпом в один такт на одно 64-разрядное слово, то фактическое время вычислений будет определяться временем прохождения самого большого из этих потоков.

Быстрое преобразование Фурье-256

Так как базовой операцией сопроцессора является умножение вектора на матрицу, то базисной операцией в декомпозиции быстрого алгоритма БПФ может служить дискретное преобразование Фурье (ДПФ) с произвольным основанием (2,4,8,16…). Рассмотрим поиск оптимального основания для БПФ-256.

Так как суть ДПФ - комплексное умножение матрицы на вектор, то преобразование ДПФ-8 бьется на две стадии: умножение вектора входных данных X на матрицу коэффициентов W(Sin и cos) (см. рРис. 2) и последующую стадию сложения частичных произведений с учетом инверсии знака (см. рРис. 3). Более детально ознакомиться с алгоритмом БПФ-256 для NeuroMatrix можно в [2].

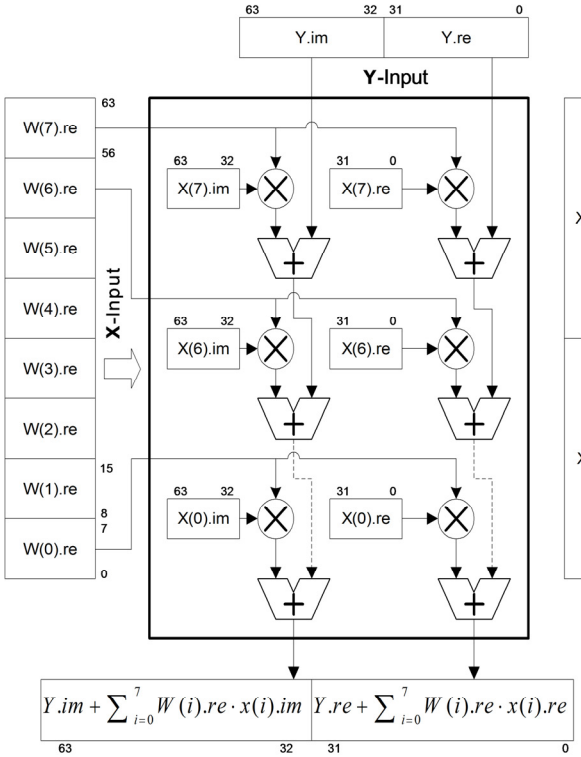

Рис. 2 Эквивалентная схема векторного умножителя при разбиении матрицы весовых коэффициентов - 8x2

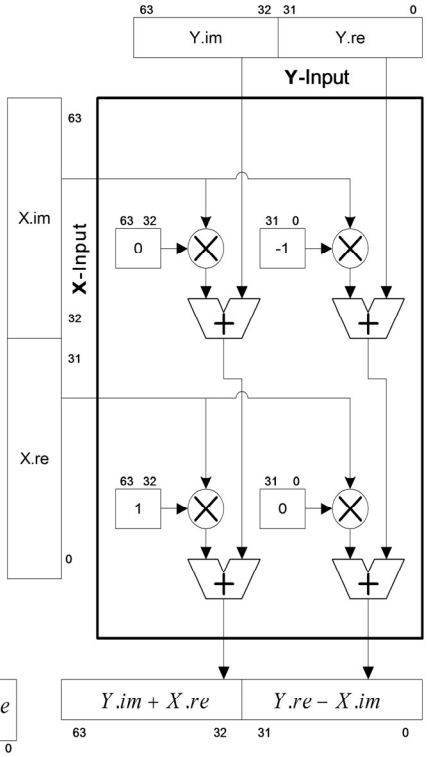

Рис. 3 Эквивалентная схема векторного умножителя при разбиении матрицы весовых коэффициентов - 2x2

Оценим время работы ДПФ-8. Разрядность входных и выходных данных возьмем 32, разрядность коэффициентов W.re и W.im (sin и cos) – 8 бит. При таких разрядностях имеем разбиение матрицы векторного умножителя на 2 столбца (Re и Im) и 8 строк (см. рРис. 2). Для вычисления ДПФ-8 через матрицу необходимо пропустить две (мнимую и действительную часть) 8x8 байтовых матрицы W(sin, cos). Так как данные передаются по 64-разрядной шине, то на это потребуется 2x8 тактов. Также 8 тактов потребуется, чтобы просуммировать частичные произведения от предыдущего шага (рРис. 3) и еще 8 тактов - на сдвиг результатов вправо для нормализации результата, так как вычисления происходят в целочисленной арифметике с фиксированной точкой. Т.е. для ДПФ-8 требуется 16+8+8=32 такта. Для других оснований-N аналогичный расчет дает общую формулу: $4*N^2/16+2*N/2+2*N/2 = N^2/4+2N$ тактов, где 16,2,2 в знаменателях – коэффициенты параллельности (число одновременных умножений в матрице) при соответствующих разбиениях матрицы умножителя. В зависимости от выбранного основания для базисных ДПФ, в схеме БПФ будет содержаться различное число слоев. Подста-

новка различных оснований в формулу дает зависимость времени вычисления одного слоя БПФ-256 (см. Таблица 1).

Таблица 1

Зависимость времени вычисления одного слоя БПФ-256 из отдельных ДПФ

Основание	T, тактов на DFT(N): $N^2/4+2N$	K, кол-во DFT(N) в слое для FFT(256)	Тактов на слой, T*K
DFT(2)	5	128	640
DFT'(2)	2	128	256'
DFT(4)	12	64	768
DFT(8)	32	32	1024
DFT(16)	96	16	1536
DFT(32)	320	8	2560
DFT(256)	16896	1	16896

DFT'(2)- вариант «бабочки» без умножения, выполненный только на сложениях и вычитаниях, где коэффициенты sin и cos = +1 и -1 .

Исходя из времени вычисления каждого слоя в Таблица 2 приведено время вычисления всего БПФ-256 для разных схем декомпозиции.

Таблица 2

Зависимость времени вычисления БПФ-256 от схемы декомпозиции

Схема FFT-256	**Слоев**	**Кол-во тактов**	**Тактов без 1 слоя нормализации**	**Тактов без 2 слоев нормализации**
2-2-2-2-2-2-2-2	8	8*640=5120	--	-- Теор./практ.
4-4-4-4	4	4*768=3072	2816	2560
4-8-8	3	768+1024+1024=**2816**	**2560**	2048
2'-8-16	3	256+1024+1536=**2816**	**2560**	**2304/2742 (+19%)**
16-16	2	1536+1536=3072	2816	2560/3008 (+17%)
8-32	2	1024+2560=3584	3584	3328
256	1	16896	--	--

Как видно из Таблица 2 оптимальной является схема 2'-8-16. Практическая реализация на процессоре этих схем отличается чуть меньше чем на 20%, это связанно с дополнительными расходами на инициализацию регистров, работу циклов и особенностями работы конвейера команд.

Двумерное дискретное косинусное преобразование (ДКП)

Принцип вычисления двумерного ДКП 8x8 очень схож с ДПФ-8 и имеет такую же структуру разбиения [3]. Разница только в том, что при первом проходе ДКП-8 (по горизонтали) 32р. коэффициенты косинусов загружаются в рабочую матрицу умножителя, а входные байтовые данные изображения в виде упакованных 64-р. слов подаются на вход умножителя. На втором проходе ДКП-8, наоборот – 32-р. результаты преобразований загружаются в матрицу умножителя, а на вход умножителя подаются байтовые вектора косинусов. И для первого, и для второго прохода требуется 4 раза сменить рабочую матрицу коэффициентов и на вход каждой подать по 8 64-р. слов. Таким образом, двумерное ДКП 8x8 требует 2*4*8=64 такта. На практике эта цифра составляет 78 тактов (1.2 такта/пиксель), что также на 20% выше расчетного.

Wavelet преобразование

В настоящее время алгоритмы кодирования на основе вейвлет-преобразований (ВП) играют большую роль в области сжатия изображений. Рассмотрим реализацию фильтра биортогонального ВП Коэна-Добеши-Фово 5/3 на платформе NeuroMatrix.

Как известно, ВП преобразование заключается в фильтрации исходного изображения по вертикали и горизонтали высокочастотным и низкочастотным фильтром с последующим прореживанием через один.

С учетом матричной структуры умножителя коэффициенты фильтров расположатся лесенкой как показано на рисунках Рис**. 4** и Рис**. 5**. Как видно из Рис. 4 для формирования 4 выходных отсчетов (как для высокочастотного, так и для низкочастотного фильтра) необходимо иметь две матрицы коэффициентов и соответственно две команды векторного умножения. Т.е. в среднем на обработку одного пикселя двумя фильтрами потребуется один такт. При вертикальной фильтрации полученные результаты загружаются в матрицу весовых коэффициентов умножителя, а коэффициенты фильтров в виде 4-разрядных чисел, упакованных в 64 разрядные слова подаются на умножающий вход (см. Рис. 5). На каждом такте генерируется 4 результирующих отсчета. Итоговая расчетная производительность одной ступени вейвлет преобразования составляет: 1+0.25= 1.25 такта на точку.

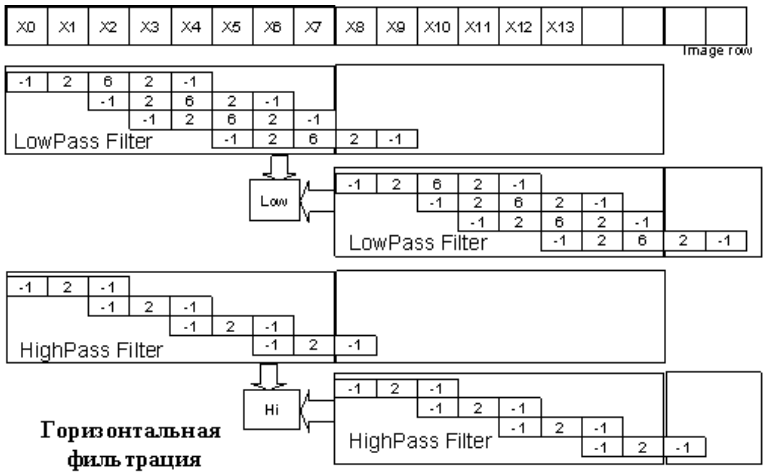

Рис. 4. Лестничная структура весовых коэффициентов векторного умножителя при прохождении низкочастотного и высокочастотного фильтра в горизонтальном направлении

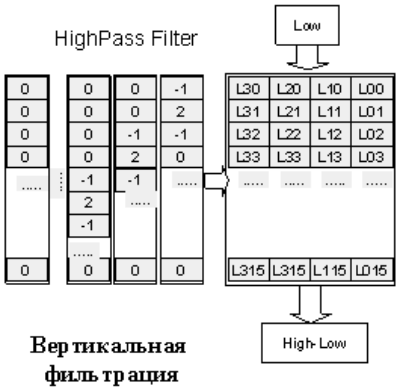

Рис. 5 Лестничная структура весовых коэффициентов векторного умножителя при прохождении высокочастотного фильтра в вертикальном направлении

Преобразование Уолша-Адамара

Наряду с ДПФ в системах цифровой связи получило широкое распространение дискретное преобразование Уолша-Адамара (DHT). В виду единичных коэффициентов DHT: -1 и +1 его можно считать самым быстродействующим на платформе NeuroMatrix. Высокая эффективность достигается за счет возможности работы с упакованными 2-разрядными данными. Так, если мы возьмем исходный вектор с 64-разрядными элементами длины n, то для вычисления преобразования DHT(n) нам понадобится $n^2/n=n$ тактов, где n^2 –общее кол-во операций умножения с накоплением, а n – коэффициент параллельности (при n<=32), полученный от разбиения матрицы на n строк. Схема быстрого преобразования FHT(N) создается через такую же декомпозицию до базисных DHT(n) в несколько слоев, как и в случае БПФ.

Кол-во дискретных преобразований DHT(n) в одном слое быстрого FHT(N) составляет S=N/n.Время вычисления одного слоя составляет S*n=(N/n)*n=N. Т.е. оно постоянно и не зависит от основания n (при n<=32). Однако, общее время вычисление FHT(N) зависит от кол-ва слоев k и составляет k*N , где $k=\log_{32}N$.

Таким образом:

при N≤32 требуется один слой, общее время составит N тактов.

при N≤32*32=1024 требуется два слоя, общее время составит 2*N тактов.

при N≤32*32*32= 32768 требуется три слоя, общее время составит 3*N тактов и т.д.

В итоге мы имеем гораздо меньшую вычислительную сложность - $O(N\log_{32}N)$, по сравнению со стандартной быстрой схемой $O(N\log_2N)$. Очевидно, что наибольшая эффективность достигается при длинах равных степеням 32-х.

Заключение

Реальные замеры времени на данных примерах демонстрируют, что развитая архитектура внутренних и внешних шин ядра NMC3 и наличие 4 внутренних банков памяти позволяет обеспечить вычисления в несколько параллельных потоков входных и выходных данных. Таким образом, исключается из расчета производительности временные задержи, связанные с пересылкой данных.

По сравнению с предыдущей архитектурой NMC в задачах DCT8x8 и FFT новое ядро NMC3 показало прирост производительности от 20% до 40% в тактах.

На примере стандартных дискретных преобразований показана относительная простота теоретической оценки их производительности, которая в чистом виде соответствует числу операций с учетом фактора параллельности. Это дает возможность для оценки времени исполнения векторизуемых задач без их предварительного программирования. Поправочный коэффициент реального времени исполнения составляет не более 20% от расчетного значения.

Представленные в Таблица 3 данные по производительности подтверждают высокую эффективность архитектуры NeuroMatrix в задачах цифровой обработки сигналов, в целом, и в вычислениях дискретных преобразований, в частности.

Таблица 3

Расчетная и практическая производительность дискретных преобразований на процессорах с ядром NMC3 320МГц

Преобразование-размер данных	Разрядность. вх.-вых. данных	Кол-во тактов (расчетное)	Кол-во тактов (практическое)	Время, мкс	Тактов/точку (пиксель)
FFT-256	32-32	3100	2742	8,5-9,5*	11
FFT-512	32-32	5376	6500	20-21*	13
FFT-1024	32-32	15360	18400	54-57*	18
DHT-1024	64-64	2048	2252	7	2,2
4xDHT-1024**, 4x1024	16-16	2048	2252	7	0,54
DCT-8x8, Кадр 128x128	8-32	16384	20215	63	1,2 (643fps для SD кадра)***
Wavelet,1 ступень Кадр 128x128	8-16	20480	26200	81	1,6 (480fps для SD кадра)***

*- разброс во временах обусловлен различными вариантами нормализации результатов [2]

** 4xDHT-1024 – означает одновременное вычисление четырех DHT над матрицей из 4 строк сигналов длиной 1024

*** данные по фреймрейту для SD кадра (720x576) являются пиковыми и получены путем пересчета из блоков 128x128, умещающихся во внутренней памяти. Реальная цифра будет зависеть от внешней памяти и работы контроллеров ПДП.

Литература

1. Семейство процессоров обработки сигналов с векторно-матричной архитектурой NEURO-MATRIX® / Черников В., Виксне П., Шелухин А. [и др.] // Электроные компоненты. 2006. № 6. С. 79-84.

2. Кашкаров В., Мушкаев С. Организация параллельных вычислений в алгоритмах БПФ на процессоре NM6403 // Цифровая обработка сигналов. 2001. № 1. С. 53-58.

3. *Мушкаев С.В., Ландышев С.В.* Применение процессора NM6403 (Л1879ВМ1) для сжатия изображений // Цифровая обработка сигналов. 2002. № 1. С. 12-18.

High-performance discrete transformations on NeuroMatrix processors with NMC3 core

***Sergey Viktorovich Mushkaev**, Leading software engineer*
Scientific and Technical Center «Modul»

The author discusses the performance evaluation of vector computing on NeuroMatrix processors . The examples of discrete Fourier, cosine, Walsh-Hadamard and wavelet transformations are analysed. The principle of optimal algorithms design is dwelled on. The effectiveness of vector calculations on NeuroMatrix processors is demonstrated.

Keywords: fast transformation, DCT, FFT, Wavelet, Walsh-Hadamard, NMC, NMC3, NeuroMatrix, NM6406, MB7707.