ВЫСОКОПРОИЗВОДИТЕЛЬНОЕ ПРОЦЕССОРНОЕ ЯДРО NMC4 ДЛЯ ОБРАБОТКИ ВЕКТОРНЫХ ДАННЫХ В ФОРМАТАХ С ПЛАВАЮЩЕЙ И ФИКСИРОВАННОЙ ТОЧКАМИ

HIGH-PERFORMANCE NMC4 VECTOR PROCESSOR CORE FOR FIXED AND FLOATING POINT CALCULATIONS

УДК 004.383

ЧЕРНИКОВ АЛЕКСАНДР ВЛАДИМИРОВИЧ
achernikov@module.ru

ЧЕРНИКОВ ВЛАДИМИР МИХАЙЛОВИЧ

ВИКСНЕ ПАВЕЛ ЕВГЕНЬЕВИЧ

ШЕЛУХИН АЛЕКСАНДР МИХАЙЛОВИЧ

ЗАО НТЦ «Модуль»
125190, г. Москва, а/я 166
Тел.: +7 (495) 531-30-80
http://www.module.ru

CHERNIKOV ALEXANDER V.
achernikov@module.ru

CHERNIKOV VLADIMIR M.

VIXNE PAVEL E.

SHELUKHIN ALEXANDER M.

RC Module JSC
P.O. Box 166, Moscow, Russia, 125190
Tel.: +7 (495) 531-30-80
http://www.module.ru

В докладе описана архитектура высокопроизводительного процессорного ядра NMC4 с архитектурой NeuroMatrix, позволяющая к управляющему RISC-процессору подключать сопроцессоры обработки векторных данных с фиксированной и плавающей точками. RISC-процессор выполняет функции управления, осуществляет обработку скалярных данных с фиксированной точкой и формирует адреса векторных данных. Векторный сопроцессор обработки данных с фиксированной точкой заимствован из процессорного ядра предыдущего поколения NMC3. Векторный сопроцессор плавающей арифметики содержит узел переупаковки данных и до восьми вычислительных ячеек, каждая из которых состоит из 8 векторных регистров 32×64 бита и вычислителя, выполняющего за один процессорный такт до 8 операций плавающей арифметики в формате одинарной точности или до 2 операций в формате двойной точности. Процессорное ядро NMC4 предназначено для построения высокопроизводительных систем на кристалле, применяемых в нейронных сетях, в ускорителях супер-ЭВМ, в радиолокации, при обработке больших потоков видеосигналов. На базе ядра NMC4 спроектирована и изготовлена двухпроцессорная СнК 1879ВМ6Я по КМОП технологии с нормами 65 нм. Одно процессорное ядро выполняет операции с плавающей точкой, другое — с фиксированной точкой. Микросхема работает на частоте 500 МГц и обеспечивает производительность до 16 GFLOP/s и до 112 GMAC/s. Микросхема содержит 16 Мбит внутренней памяти, 32-разрядный интерфейс с памятью DDR2, два байтовых коммуникационных порта с пропускной способностью 120 Мбайт каждый, интерфейсы USB, SPI, GPIO, таймеры, контроллеры прерываний и ПДП.
Ключевые слова: процессорное ядро с архитектурой NeuroMatrix; векторный сопроцессор; динамический VLIW; нейронные сети.
The article presents an architecture of high-performance NMC4 processor core that allows connecting vector coprocessors for fixed or floating point calculations to a single controlling RISC-processor. The RISC-core implements control functions, fixed-point scalar data processing and address generation for vector operations. Vector coprocessor for fixed-point calculations is inherited from previous generation of NeuroMatrix processor NMC3 core. The vector coprocessor for floating-point calculations contains data repack unit and up to 8 arithmetic cells. An arithmetic cell contains 8 vector registers (32×64 bits) and calculation unit capable of performing up to 8 single precision operations per cycle or up to 2 double precision operations per cycle. NMC4 processor core is designed for high-performance systems-on-chip that can be used as accelerators for supercomputers for neural network applications, radiolocation, video processing. NMC4 core is used inside dual-core 1879BM6Я SoC manufactured on 65nm CMOS technology node. One processor core performs floating-point calculations the other — fixed-point calculations. SoC is running on 500MHz and performs up to 16GFLOP/s and up to 112GMAC/s. SoC contain 16 Mbits of internal memory, 32-bit DDR2 interface, 2 communication links with up to 120MB/s throughput each, USB interface, SPI interface, GPIO interface, timers, interrupt controllers and DMA.
Keywords: NeuroMatrix processor core; vector processor; dynamic VLIW; neural networks; floating-point.

ВВЕДЕНИЕ

В настоящее время круг задач, требующих для своего решения применения мощных вычислительных ресурсов, все время расширяется [1], а высокая потребляемая мощность современных вычислительных систем выдвигает жесткие требования к энергоэффективности элементной базы, на которой реализованы данные системы [2]. В то же время, создание специализированной элементной базы, способной эффективно решать лишь узкий класс задач экономически нецелесообразно. Поэтому процессорные узлы современных вычислительных систем должны

быть достаточно универсальными, чтобы обеспечивать эффективное решение широкого класса задач. А наращивание производительности вычислительных систем без серьезной переработки программного обеспечения возможно, только если архитектура процессорных узлов будет хорошо масштабироваться. Таким образом, архитектура процессорного узла, применяемого в современных ускорителях высокопроизводительных вычислительных систем, должна обладать следующими свойствами: энергоэффективность, масштабируемость, универсальность.

Для решения этих проблем в ЗАО НТЦ «Модуль» была разработана архитектура высокопроизводительного ядра NMC4, позволяющая к управляющему RISC-процессору подключать вычислительные сопроцессоры различного типа. За счет использования сопроцессоров различного типа достигается высокая универсальность разрабатываемых микропроцессорных узлов. Масштабируемость и энергоэффективность систем достигается за счет использования сопроцессоров, оптимизированных для решения задач, в которых большую часть вычислений можно распараллелить. Данная архитектура способна эффективно решать задачи цифровой обработки широкополосных сигналов в радиолокации, навигации и связи, задачи, требующие обработки большого объема параллельных данных с применением векторно-матричных вычислений, использующие алгоритмы линейной алгебры и нейронных сетей.

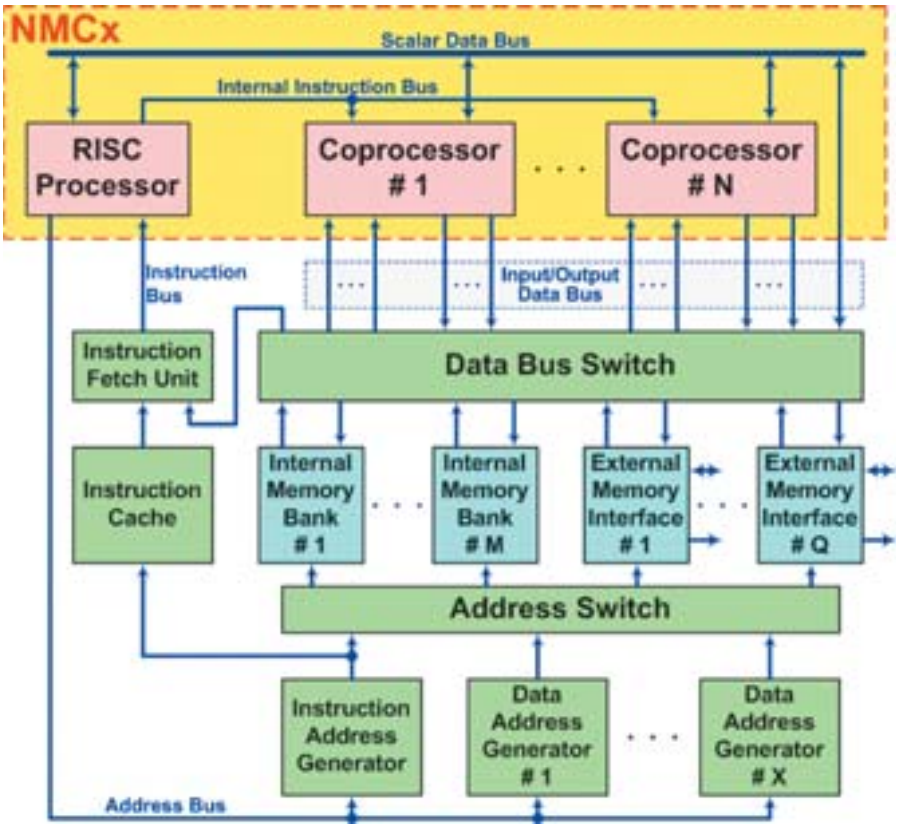

Рис. 1. Структурная схема процессорной системы на базе ядра NMC4

В докладе описана общая структура процессорной системы, построенной на базе ядра NMC4, описан векторный сопроцессор для обработки данных, представленных в формате с плавающей точкой, а также описана микросхема NM6407, реализованная на базе процессорного ядра NMC4 по КМОП технологии с топологическими нормами 65 нм.

ОБЩАЯ СТРУКТУРНАЯ СХЕМА ПРОЦЕССОРНОЙ СИСТЕМЫ NMC4

Требование обработки потоков данных в реальном времени с производительностью, сравнимой с производительностью универсальных процессоров, и потреблением, характерным для встроенных систем, привело к созданию в ЗАО НТЦ «Модуль» семейства архитектур NeuroMatrix [3]. Для расширения круга задач, эффективно решаемых с использованием процессоров семейства NeuroMatrix, была разработана архитектура процессорного ядра NMC4, обладающая следующими основными особенностями:

- векторно-конвейерный принцип выполнения операций, позволяющий одним потоком команд задавать большое количество параллельно выполняемых операций (динамический VLIW);
- наличие одного или нескольких векторных сопроцессоров, работающих под управлением одного RISC-процессора и имеющих свои шины ввода/вывода данных;
- использование внешних адресных генераторов, что обеспечивает эффективное использование адресных регистров ядра при адресации векторных данных;
- конвейер с очередью команд на ступени выборки операндов из памяти, обеспечивающий эффективную работу с банками внутренней и внешней памяти, имеющими различную глубину конвейера.

Обобщенная структура процессорной системы цифровой обработки сигналов на базе ядра NMC4 (NeuroMatrix Core 4), обладающая всеми вышеперечисленными свойствами, приведена на рис. 1.

В состав процессорной системы на базе ядра NMC4 входят следующие основные блоки:

NMC4 — процессорное ядро NeuroMatrix 4, содержащее управляющий RISC-процессор и несколько сопроцессоров, которые выполняют цифровую обработку данных, предварительно загруженных в память системы.

RISC-процессор с минимальными изменениями был заимствован из предыдущего поколения процессорных ядер NeuroMatrix [4] и решает следующие основные задачи:

- декодирует и выполняет команды, считываемые из памяти по шине команд (Instruction Bus);
- осуществляет через шину скалярных данных (Scalar Data Bus) конфигурирование сопроцессоров, настраивая их на обработку данных;
- выставляет на внутреннюю шину инструкций (Internal Instruction Bus) команды оперативного управления сопроцессорами;
- формирует и выставляет на адресную шину (Address Bus) адреса команд, а также скалярных и векторных данных.

Internal Memory Banks — банки внутренней памяти.

External Memory Interfaces — интерфейсы с внешней памятью, подключенной к соответствующим внешним шинам системы.

Instruction Address Generators — генератор последовательных адресов команд, содержимое которого может быть изменено произвольным образом при выполнении различных команд перехода.

Data Address Generator — генераторы адресов векторных данных, загружаемые через адресную шину начальным адресом, смещением адреса и количеством повторов операции ввода/вывода.

Address Switch — коммутатор адресов, который осуществляет пересылку адресов, формируемых адресными генераторами, и адресов, поступающих от внешних каналов прямого доступа к памяти, в банки внутренней памяти и в интерфейсы с внешней памятью.

Data Bus Switch — коммутатор шин данных, обеспечивающий обмен данными между процессорным ядром и памятью системы, а также обмен данными между каналами DMA и банками памяти.

Instruction Fetch Unit — блок выборки команд, который выстраивает в единую очередь команды, считываемые из внутренней или внешней памяти системы.

Количество и тип сопроцессоров, количество банков внутренней памяти и внешних шин системы, а также количество генераторов адресов векторных данных определяются требуемой производительностью системы, что позволяет легко наращивать производительность вычислительных систем в зависимости от класса решаемых задач с сохранением программной совместимости внутри семейства процессоров.

В зависимости от типа сопроцессора изменяется область применения микросхем, разработанных на базе процессорной системы NMC4. Эффективно решаемые задачи включают в себя, но не ограничиваются, такими задачами как: цифровая обработка сигналов в радиолокации, навигации и связи, вычисление преобразования Фурье, Адамара, цифровая фильтрация, цифровая коммутация, распознавание образов, обработка изображений, нейронные сети глубокого обучения.

СОПРОЦЕССОР АРИФМЕТИКИ С ПЛАВАЮЩЕЙ ТОЧКОЙ

Для ускорения выполнения арифметических операций над данными, представленными в формате с плавающей точкой, используется векторно-матричный сопроцессор арифметики с плавающей точкой. Сопроцессор выполняет операции в соответствии со стандартом IEEE 754-2008 [5] и работает с данными одинарной точности (32 разряда) и двойной точности (64 разряда).

Структурная схема сопроцессора представлена на рис. 2.

Основными узлами матрично-векторного сопроцессора являются:

CENTRAL CONTROL UNIT — центральный блок управления сопроцессором. Он принимает команды от управляющего RISC-ядра по внутренней шине команд (INTERNAL INSTRUCTION BUS), дешифрирует и проверяет их на правильность. Если команда правильная, то она запускается на выполнение при условии, что все требуемые ей ресурсы свободны. В случае ошибочности команды она не выполняется, и при этом формируется соответствующее прерывание от сопроцессора. 32-разрядная скалярная шина данных (SCALAR DATA BUS) используется для чтения/записи программно-доступных скалярных регистров блока управления.

REPACK UNIT — блок упаковки и распаковки данных. Этот блок предназначен для различных преобразований данных в формате с плавающей точкой в целые числа и обратно, а также 64-разрядных данных в 32-разрядные и наоборот.

SWITCH 9 –> 11 — коммутатор 9 в 11, обеспечивающий обмен данными между функциональными узлами сопроцессора с плавающей точкой и памятью процессорной системы.

FP PROCESSING CELL 0, ..., FP PROCESSING CELL 3–4 одинаковые процессорные ячейки, каждая из которых осуществляет арифметические операции над данными в формате с плавающей точкой как одинарной (32 разряда), так и двойной точности (64 разряда). Каждая ячейка имеет следующие 64-разрядные шины: две входных — IDB0, IDB1 и одну выходную — ODB, что позволяет за один такт осуществить до двух операций чтения и одной операции записи. 32-разрядная скалярная шина данных (SCALAR DATA BUS) используется для чтения/записи программно-доступных скалярных регистров процессорных ячеек. Процессорные ячейки являются основным вычислительным элементом сопроцессора арифметики с плавающей точкой. Структурная схема процессорной ячейки приведена на рис. 3.

Процессорная ячейка включает в себя следующие основные узлы:

VR0–VR7 — векторные регистры общего назначения, которые используются как в операциях ввода/вывода, так и в арифметических операциях над данными с плавающей точкой. Их максимальная емкость 32 64-разрядных слова упакованных данных. В некоторых типах операций векторные регистры образуют

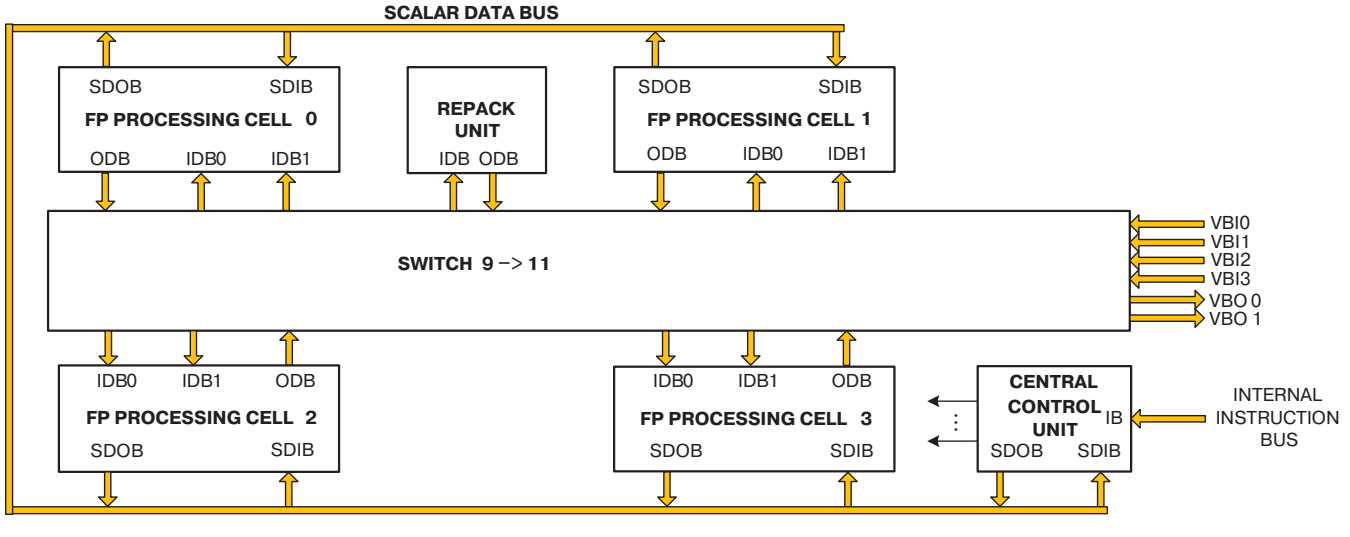

Рис. 2. Структурная схема сопроцессора арифметики с плавающей точкой

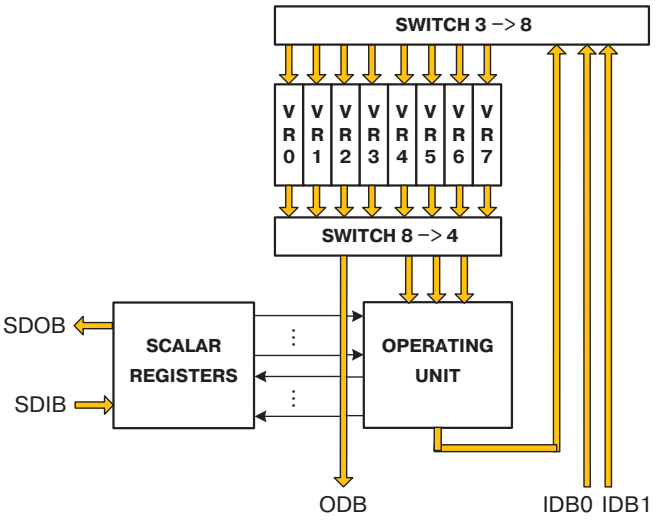

Рис. 3. Структурная схема процессорной ячейки

регистровые пары, формируя вектора максимальным размером 32 элемента по 128 разрядов.

SWITCH 3 –> 8 — коммутатор 3 в 8. Он позволяет одновременно записать результат арифметической операции и данные, считанные по шинам IDB0, IDB1 из памяти, а также из других ячеек или блока упаковки и распаковки, в три векторных регистра соответственно.

SWITCH 8 –> 4 — коммутатор 8 в 4. Он предназначен для выбора до 3-х источников для арифметических операций, а также одного источника для записи в память, в другие ячейки или в блок упаковки и распаковки по выходной шине ODB. Причем в качестве этих источников может использоваться любой из восьми векторных регистров.

SCALAR REGISTERS — программно-доступные скалярные регистры, входящие в состав каждой процессорной ячейки. Скалярные регистры хранят значение признаков, формируемых при выполнении арифметических операций, а также с помощью данных регистров осуществляется управление маскированием выполнения векторных операций.

OPERATING UNIT — операционное устройство для выполнения векторных и матричных операций над данными в формате с плавающей точкой. Данное устройство может работать в одном из четырех режимов:

- Операции над данными в формате с плавающей точкой двойной точности — см. рис. 4. В этом режиме все входные операнды A, B и C и результат D представляют собой 64-разрядные числа в формате с плавающей точкой двойной точности. Основная операция режима — умножение двух чисел и сложение с третьим.
- Операции над комплексными данными в формате с плавающей точкой одинарной точности — см. рис. 5. Данный режим характеризуется тем, что все входные операнды: A1 и A0, B1 и B0, C1 и C0, а также результат D1 и D0 представляют собой комплексные 64-разрядные числа, причем старшие 32 разряда содержат действительную часть (A1, B1, C1 и D1), а младшие 32 разряда (A0, B0, C0 и D0) — мнимую часть. Основная операция режима — умножение двух комплексных чисел и сложение с третьим.
- Операции над векторными данными в формате с плавающей точкой одинарной точности — см. рис. 6. В этом режиме все входные операнды A1 и A0, B1 и B0, C1 и C0, а также результат D1 и D0 образуют 64-разрядные вектора из двух 32-разрядных элементов. Основная операция режима — умножение двух векторов по два элемента и сложение с третьим двухэлементным вектором.
- Операции над матричными данными в формате с плавающей точкой одинарной точности — см. рис. 7. Данный режим характеризуется тем, что все входные операнды A1 и A0, B1 и B0, C1 и C0, а также результат D1 и D0 образуют 64-разрядные вектора из двух 32-разрядных элементов, как и в предыдущем случае. Основная операция режима — умножение вектора из двух элементов [A1 A0] на матрицу 2×2 элемента $\begin{bmatrix} B3 & B1 \\ B2 & B0 \end{bmatrix}$ и сложение с третьим двухэлементным вектором [D1 D0]. Особенностью данной операции является то, что матрица считывается сразу из пары векторных регистров: из того, что указан в команде (он обязан иметь четный номер), считывается столбец $\begin{bmatrix} B1 \\ B0 \end{bmatrix}$, а из регистра с номером на единицу больше — столбец $\begin{bmatrix} B3 \\ B2 \end{bmatrix}$. В остальных случаях операнд в виде одного 64-разрядного данного или вектора из двух 32-разрядных данных считывается только из одного векторного регистра. Результат также пишется только в один векторный регистр.

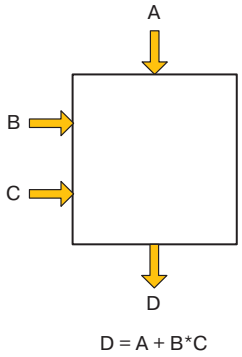

Рис. 4. Операции над данными в формате с плавающей точкой двойной точности

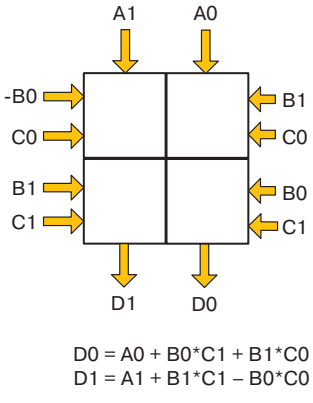

Рис. 5. Операции над комплексными данными в формате с плавающей точкой одинарной точности

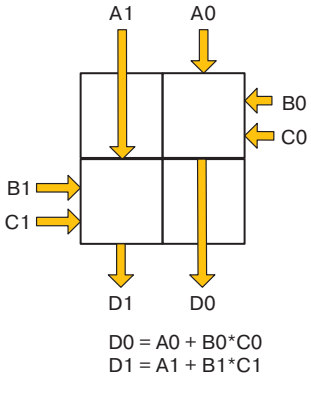

Рис. 6. Векторные операции над данными в формате с плавающей точкой одинарной точности

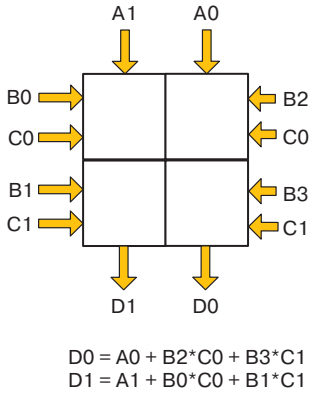

Рис. 7. Матричные операции над данными в формате с плавающей точкой одинарной точности

МИКРОСХЕМА NM6407

На базе процессорного ядра NMC4 в ЗАО НТЦ «Модуль» была разработана микросхема NM6407 [6], предназначенная для использования в качестве основного или дополнительного процессорного узла в вычислительных системах, интенсивно применяющих цифровую обработку сигналов и арифметику с плавающей точкой. Микросхема может быть использована в качестве базового элемента при построении многопроцессорных параллельных вычислительных систем.

Основные характеристики микросхемы NM6407:

- Наличие двух независимых RISC-процессорных ядер, одно из которых работает с матрично-векторным сопроцессором для обработки данных, представленных в формате с фиксированной точкой и программируемой разрядностью, а другое управляет матрично-векторным сопроцессором арифметики с плавающей точкой.
- Тактовая частота — до 500 МГц.
- Аппаратная поддержка операций умножения вектора на матрицу и матрицы на матрицу, как для данных в формате с фиксированной точкой, так и в формате с плавающей точкой.
- Формат обрабатываемых данных — 32-разрядные скалярные данные, а также:
  - вектора данных в формате с фиксированной точкой программируемой разрядности от 1 до 64 разрядов, упакованные в 64-разрядные слова;
  - вектора 32-разрядных данных в формате с плавающей точкой, упакованные в 64-разрядные слова (одинарная точность);
  - 64-разрядные данные в формате с плавающей точкой (двойная точность).
- Общий объем внутренней статической памяти 16 Мбит.
- Адресуемое пространство — 4Г 32-разрядных слов (16 Гбайт).
- Многотактовые векторные команды (возможность одновременного выполнения до 32-х векторных команд).
- Производительность для 32-разрядных данных в формате с плавающей точкой — 32 FLOP (операций с плавающей точкой) за такт.
- Производительность для 64-разрядных данных в формате с плавающей точкой двойной точности — 8 FLOP за такт.
- Встроенные системные интеграторы, обеспечивающие доступ процессорных ядер к внутренней и внешней памяти.
- Контроллеры прямого доступа к памяти для эффективного обмена периферийных устройств с памятью, а также для аппаратной поддержки пересылок типа память — память.
- Контроллер внутренних и внешних прерываний.
- Параллельный 32-разрядный интерфейс с внешней памятью типа DDR2 400 МГц.
- Четыре высокоскоростных байтовых коммуникационных порта с пропускной способностью не менее 125 Мбайт/с каждый.
- Интерфейс USB (device).
- Интерфейс SPI с четырьмя сигналами выборки кристалла.
- Восемь последовательных портов ввода/вывода общего назначения.
- JTAG интерфейс (IEEE-1149.1).
- Технология изготовления — 65 нм КМОП.

ЗАКЛЮЧЕНИЕ

Разработана архитектура высокопроизводительного ядра NMC4, позволяющая к управляющему RISC-процессору подключать вычислительные сопроцессоры различного типа, отвечающая современным требованиям энергоэффективности, масштабируемости и универсальности. В зависимости от типа сопроцессора изменяется область применения микросхем, разработанных на базе процессорной системы NMC4. Эффективно решаемые задачи включают в себя, но не ограничиваются, такими областями как: цифровая обработка сигналов в радиолокации, навигации и связи, вычисление преобразования Фурье, Адамара, цифровая фильтрация, цифровая коммутация, распознавание образов, обработка изображений, нейронные сети глубокого обучения.

Разработан векторно-матричный сопроцессор арифметики с плавающей точкой, выполняющий операции в соответствии со стандартом IEEE 754-2008 и работающий с данными одинарной и двойной точности. Сопроцессор имеет аппаратную поддержку преобразований данных в формате с плавающей точкой в целые числа и обратно, а также данных одинарной точности в данные двойной точности и наоборот. Арифметический узел позволяет выполнять операции над данными двойной точности, а также комплексными, векторными и матричными данными одинарной точности.

Разработана микросхема NM6407 на базе двух независимых процессорных ядер NMC4, одно из которых работает с матрично-векторным сопроцессором для обработки данных, представленных в формате с фиксированной точкой и программируемой разрядностью, а другое управляет матрично-векторным сопроцессором арифметики с плавающей точкой. Процессорные ядра работают на частоте 500 МГц, что обеспечивает производительность 16 GFLOP/s при работе с данными одинарной точности, 4 GFLOP/s для двойной точности и до 224 GMAC/s для данных с фиксированной точкой.

Микросхема NM6407 предназначена для использования в качестве основного или дополнительного процессорного узла в вычислительных системах, интенсивно применяющих цифровую обработку сигналов и арифметику с плавающей точкой, а также может быть использована в качестве базового элемента при построении многопроцессорных параллельных вычислительных систем.

ЛИТЕРАТУРА

1. Баденко В. Л. Высокопроизводительные вычисления: учебное пособие. — СПб.: Изд-во Политехн. ун-та. — 2010.
2. Энергоэффективность центров обработки данных. URL: https://www.bytemag.ru/articles/detail.php?ID=9167 (дата обращения 18.08.2017).
3. Черников В. М., Виксне П. Е., Шелухин А. М., Панфилов А. П. Отечественные высокопроизводительные процессоры цифровой обработки сигналов векторно-матричной архитектуры, перспективы развития // Материалы конференции «Перспективы развития высокопроизводительных архитектур. История, современность и будущее отечественного компьютеростроения». Сборник научных трудов ИТМиВТ. — М.: ИТМиВТ им. С. А. Лебедева РАН. — 2008. — Вып. 1. — С. 52–59.
4. Черников В. М., Виксне П. Е., Шелухин А. М., Шевченко П. А., Панфилов А. П., Косоруков Д. Е., Черников А. В. Семейство процессоров обработки сигналов с векторно-матричной архитектурой NeuroMatrix // Электронные компоненты. — 2006. — № 6. — С. 79–84.
5. IEEE, 754-2008 — IEEE Standard for Floating-Point Arithmetic, ©Copyright IEEE, 2008.
6. Черников В. М., Виксне П. Е., Шелухин А. М., Черников А. В. Новое ядро процессора обработки сигналов NMC4 семейства NeuroMatrix // Программа и тезисы докладов Шестого Московского суперкомпьютерного форума. — 2015. — С. 12–13.