Нейропроцессоры и нейро-ЭВМ.

Агниашвили Д.В. ИВТ-41-15

Чебоксары 2017-2018

Нейрокомпьютеры - это системы, в которых алгоритм решения задачи представлен логической сетью элементов частного вида - нейронов с полным отказом от булевских элементов типа И, ИЛИ, НЕ. Как следствие этого введены специфические связи между элементами, которые являются предметом отдельного рассмотрения.
В отличие от классических методов решения задач нейрокомпьютеры реализуют алгоритмы решения задач, представленные в виде нейронных сетей. Это ограничение позволяет разрабатывать алгоритмы, потенциально более параллельные, чем любаядругая их физическая реализация.

Нейрокомпьютер - устройство переработки информации на основе принципов работы естественных нейронных систем, на практике - это вычислительная система с архитектурой MSIMD, в которой реализованы два принципиальных технических решения: упрощен до уровня нейрона процессорный элемент однородной структуры и резко усложнены связи между элементами; программирование вычислительной структуры перенесено на изменение весовых связей между процессорными элементами.

Общее определение нейрокомпьютера может быть представлено в следующем виде. Нейрокомпьютер - это вычислительная система с архитектурой аппаратного и программного обеспечения, адекватной выполнению алгоритмов, представленных в нейросетевом логическом базисе.

	Машина фон Неймана	<>Биологическая нейронная система
Процессор	Сложный	Простой
	Высокоскоростной	Низкоскоростной
	Один или несколько	Большое количество
Память	Отделена от процессора	Интегрирована в процессор
	Локализована	Распределенная
	Адресация не по содержанию	Адресация по содержанию
Вычисления	Централизованные	Распределенные
	Последовательные	Параллельные
	Хранимые программы	Самообучение
Надежность	Высокая уязвимость	Живучесть
Специализация	Численные и символьные операции	Проблемы восприятия
Среда функционирования	Строго определенная	Плохо определенная
	Строго ограниченная	Без ограничений

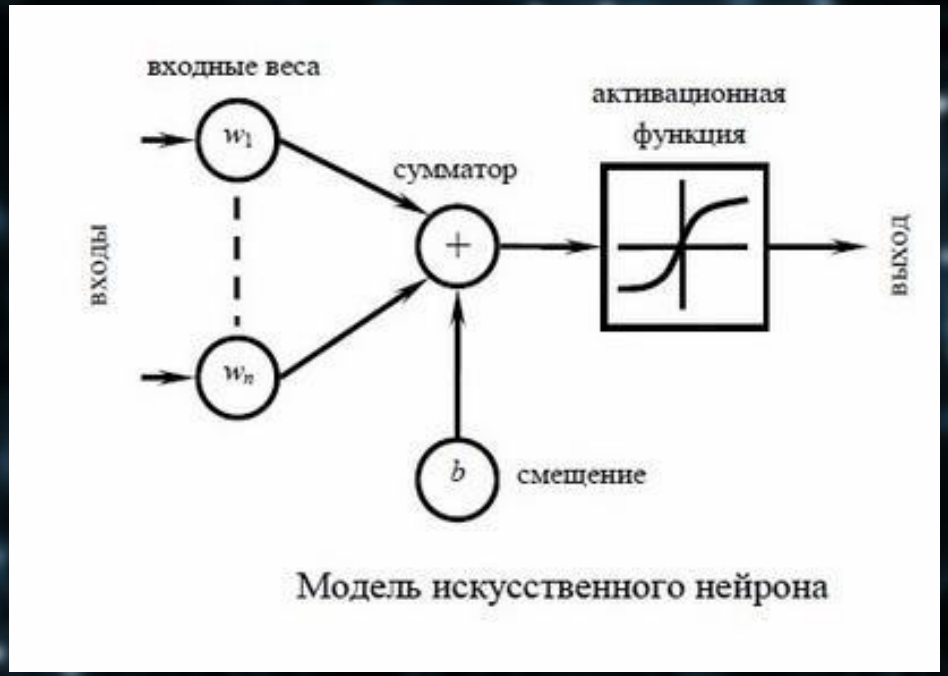
Машина фон Неймана по сравнению с биологической нейронной системой

Основные понятия

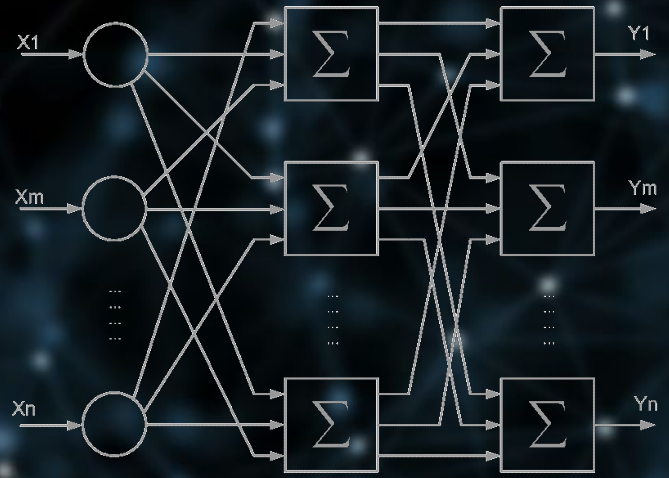
Модель технического нейрона

МакКаллок и Питтс предложили использовать бинарный пороговый элемент в качестве модели искусственного нейрона. Этот математический нейрон вычисляет взвешенную сумму n входных сигналов x_j, j = 1, 2... n, и формирует на выходе сигнал величины 1, если эта сумма превышает определенный порог u, и 0 - в противном случае.

Часто удобно рассматривать u как весовой коэффициент, связанный с постоянным входом x_0 = 1. Положительные веса соответствуют возбуждающим связям, а отрицательные - тормозным. МакКаллок и Питтс доказали, что при соответствующим образом подобранных весах совокупность параллельно функционирующих нейронов подобного типа способна выполнять универсальные вычисления. Здесь наблюдается определенная аналогия с биологическим нейроном: передачу сигнала и взаимосвязи имитируют аксоны и дендриты, веса связей соответствуют синапсам, а пороговая функция отражает активность сомы.

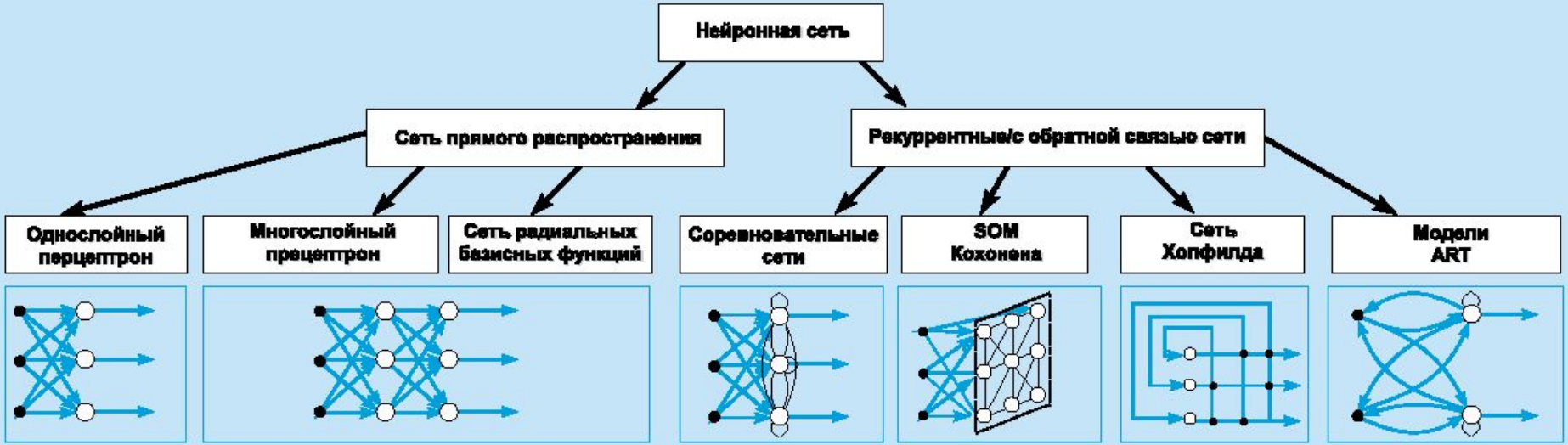


Модель искусственного нейрона



$F(x) = 1/(1 + e^{-x})$.

$$S = \sum_{i=1}^{n} X_i \cdot W_i + W_0$$



Систематизация архитектур сетей прямого распространения и рекуррентных (с обратной связью)

Области применения нейросетей:

- Прогнозирование;
- Распознавание образов;
- Классификация;
- Кластеризация и др.

Отличия нейросетей от традиционных вычислительных систем:

- **Высокая скорость обработки данных;**
- **Высокий уровень отказоустойчивости;**
- **Возможность обучения.**

Выделяют несколько (обычно три) основных типов нейронных сетей, отличающихся структурой и назначением:

1. Иерархические сети. Информация в таких сетях передается в процессе последовательного перехода от одного уровня иерархии к другому. Нейроны образуют два характерных типа соединений — конвергентные, когда большое число нейронов одного уровня контактирует с меньшим числом нейронов следующего уровня, и дивергентные, в которых контакты устанавливаются со все большим числом нейронов последующих слоев иерархии.
2. Локальные сети, формируемые нейронами с ограниченными сферами влияния. Нейроны локальных сетей производят переработку информации в пределах одного уровня иерархии. При этом функционально локальная сеть представляет собой относительно изолированную тормозящую или возбуждающую структуру.
3. Важную роль также играют так называемые дивергентные сети содним входом. Командный нейрон, находящийся в основании такой сети может оказывать влияние сразу на множество нейронов, и поэтому сети с одним входом выступают согласующим элементом в сложном сочетании нейросетевых систем всех типов.

В зависимости от используемой в НС выходной функции нейрона различают бинарные и аналоговые сети. Первые из них оперируют с двоичными сигналами, и выход каждого нейрона может принимать только два значения: логический ноль ("заторможенное" состояние) и логическая единица ("возбужденное" состояние). В аналоговых сетях выходные значения нейронов способны принимать непрерывные значения при замене ступенчатой (пороговой) функции сигмоидной.

Нейронные процессоры относятся к вычислительной технике и используются для аппаратного ускорения эмуляции работы нейронных сетей и цифровой обработки сигналов в режиме реального времени. Как правило нейропроцессор содержит регистры, блоки памяти магазинного типа, коммутатор и вычислительное устройство — содержащее матрицу умножения, дешифраторы, триггеры и мультиплексоры.

На современном этапе (по состоянию на 2017 год) к классу нейронных процессоров могут относится разные по устройству и специализации типы чипов, например:

- Нейроморфные процессоры — построенные по кластерной асинхронной архитектуре разработанной в Корнеллском университете (принципиально отличающейся от фон Неймановской и Гарвардской компьютерных архитектур, используемых последние 70 лет в IT-отрасли). В отличие от традиционных вычислительных архитектур, логика нейроморфных процессоров изначально узкоспециализированна для создания и разработки разных видов искусственных нейронных сетей. В устройстве используются обычные транзисторы из которых строятся вычислительные ядра (каждое ядро как правило содержит планировщик заданий, собственную память типа SRAM и маршрутизатор для связи с другими ядрами), каждое из ядер эмулирует работу нескольких сотен нейронов и таким образом, один чип состоящий из нескольких тысяч ядер алгоритмически может воссоздать массив из нескольких сотен тысяч нейронов и на порядок больше синапсов. Как правило такие процессоры применяются для алгоритмов глубокого машинного обучения.
- Тензорные процессоры — устройства как правило являющиеся сопроцессорами управляемыми центральным процессором, оперирующие тензорами — объектами, которые описывают преобразования элементов одного линейного пространства в другое и могут быть представлены как многомерные массивы чисел, обработка которых осуществляется с помощью таких программных библиотеках как например TensorFlow. Они как правило оснащаются собственной встроенной оперативной памятью и оперируют низко-разрядными (8-битными) числами, и узкоспециализированы для выполнения таких операций как матричное умножение и свёртка — используемая для эмуляции свёрточных нейронных сетей, которые используются для задач машинного обучения.
- Процессоры машинного зрения — во многом похожи на тензорные процессоры, но они узкоспециализированы для ускорения работы алгоритмов машинного зрения — в которых используются методы свёрточных нейронных сетей (CNN) и масштабно-инвариантная функция преобразования (SIFT). В них делается большой акцент на распараллеливание потока данных между множеством исполнительных ядер включая использование модели блокнотной памяти. — как в многоядерных цифровых сигнальных процессорах, и они также как тензорные процессоры используются для вычислений с низкой точностью принятой при обработке изображений.

ПРИНЦИПЫ НЕЙРОННОЙ ОБРАБОТКИ ИНФОРМАЦИИ

В отличие от цифровых систем, представляющих собой комбинации процессорных и запоминающих блоков, нейропроцессоры содержат память, распределённую в связях между очень простыми процессорами, которые часто могут быть описаны как формальные нейроны или блоки из однотипных формальных нейронов. Тем самым основная нагрузка на выполнение конкретных функций процессорами ложится на архитектуру системы, детали которой в свою очередь определяются межнейронными связями. Подход, основанный на представлении как памяти данных, так и алгоритмов системой связей (и их весами), называется коннекционизмом.

Три основных преимущества нейрокомпьютеров:

a) Все алгоритмы нейроинформатики высокопараллельны, а это уже залог высокого быстродействия.
b) Нейросистемы можно легко сделать очень устойчивыми к помехам и разрушениям.
c) Устойчивые и надёжные нейросистемы могут создаваться и из ненадёжных элементов, имеющих значительный разброс параметров.

Разработчики нейрокомпьютеров стремятся объединить устойчивость, быстродействие и параллелизм АВМ — аналоговых вычислительных машин — с универсальностью современных компьютеров.

Существующие процессоры

Процессоры машинного зрения:

1. Intel Movidius Myriad 2. — который является многоядерным ИИ-ускорителем основанном на VLIW-архитектуре, с дополненными узлами предназначенными для обработки видео.
2. Mobileye EyeQ — это специализированный процессор ускоряющий обработку алгоритмов машинного зрения для использования в беспилотном автомобиле.

Тензорные процессоры:

1. Google TPU (Tensor Processing Unit) — представлен как ускоритель для системы Google TensorFlow, которая широко применяется для свёрточных нейронных сетей. Сфокусирован на большом объеме арифметики 8-битной точности.
2. Intel Nervana NNP (Neural Network Processor) — это первый коммерчески доступный тензорный процессор предназначенный для постройки сетей глубокого обучения, компания Facebook была партнёром в процессе его проектирования.

Нейроморфные процессоры:

1. IBM TrueNorth — нейроморфный процессор построенный по принципу взаимодействия нейронов, а не традиционной арифметики. Частота импульсов представляет интенсивность сигнала. По состоянию на 2016 год среди исследователей ИИ нет консенсуса, является ли это правильным путем для продвижения, но некоторые результаты являются многообещающими, с продемонстрированной большой экономией энергии для задач машинного зрения.
2. Adapteva Epiphany — предназначен как сопроцессор, включает модель блокнотной памяти сети на кристалле , подходит к модели программирования потоком информации, которая должна подходить для многих задач машинного обучения.
3. KnuPath — процессор компании KnuEdge предназначен для работы в системах распознавания речи и прочих отраслях машинного обучения, он использует соединительную технологию LambdaFabric и позволяет объединять в единую систему до 512 тысячи процессоров.

Процессор NM6403

- 50 Mhz
- RISC ядро
  - 32-битные данные
  - 32-битные операции
  - 8 + 8 регистров
- Векторное устройство
  - Переменная разрядность
  - До 2048 параллельных умноженей

НТЦ «Модуль» учрежден в 1990 году известными предприятиями военно-промышленного комплекса - НПО «Вымпел» и НИИ Радиоприборостроения. За свою историю НТЦ «Модуль» прошел путь от прикладных исследований в области распознавания образов до разработки уникальных аппаратных средств цифровой обработки сигналов и изображений и построении на их основе функционально законченных вычислительных комплексов. В настоящее время НТЦ «Модуль» - лидирующая российская hi-tech-компания, работающая в области электроники. Деятельность НТЦ «Модуль» опровергает мнение о России как о сырьевом придатке развитого мира, стране, полностью утратившей научно-технический потенциал.

Процессор Л1879ВМ1 (NM6403) представляет собой высокопроизводительный микропроцессор с элементами VLIW и SIMD архитектур. В его состав входят устройства управления, вычисления адреса и обработки скаляров, а также узел для поддержки операций над векторами с элементами переменной разрядности. Кроме того, имеются два идентичных программируемых интерфейса для работы с внешней памятью различного типа, а также два коммуникационных порта, аппаратно совместимых с портами ЦПС TMS320C4x, для возможности построения многопроцессорных систем.

Нейропроцессор предназначен для обработки 32 - разрядных скалярных данных и данных программируемой разрядности, упакованных в 64 - разрядные слова.

RISC-ядро

- 5-ти ступенчатый 32-разрядный конвейер;
- 32- и 64-разрядные команды (обычно выполняется две операции в одной команде);
- Два адресных генератора, адресное пространство - 16 GB;
- Два 64-разрядных программируемых интерфейса с SRAM/DRAM-разделяемой памятью;
- Формат данных - 32-разрядные целые;
- Регистры:
- 8 32-разрядных регистров общего назначения;
- 8 32-разрядных адресных регистров;
- Специальные регистры управления и состояния;
- Два высокоскоростных коммуникационных порта ввода/вывода,
- Аппаратно совместимых с портами TMS320C4x.

VECTOR-сопроцессор

- Переменная 1-64-разрядная длина векторных операндов и результатов;
- Формат данных - целые числа, упакованные в 64-разрядные блоки, в форме слов переменной длины от 1 до 64 разрядов каждое;
- Поддержка векторно-матричных и матрично-матричных операций;
- Два типа функций насыщения на кристалле;
- Три внутренних 32x64-разрядных RAM-блока

Производительность

- Скалярные операции:
  - 50 MIPS;
  - 200 MOPS для 32-разрядных данных;
- Векторные операции:
  - от 50 до 50.000+ MMAC (миллионов умножений с накоплением в секунду);
- I/O и интерфейсы с памятью:
  - пропускная способность двух 64-разрядных интерфейсов с памятью - до 800 Мбайт/сек;
- I/O коммуникационные порты - до 20 Мбайт/сек кажд

Особенности NM64003 (1)

- Возможность работы с входными сигналами (синапсами) и весами переменной разрядности (от 1 до 64 бит), задаваемой программно, что обеспечивает уникальную способность нейропроцессора увеличивать производительность с уменьшением разрядности операндов;
- Быстрая подкачка новых весов на фоне вычислений;
- (24 операции умножения с накоплением за один такт при длине операндов 8 бит);
- V аппаратная поддержка эмуляции нейросетей большой размерности;
- Реализация функции активации в виде пороговой функции или функции ограничения;

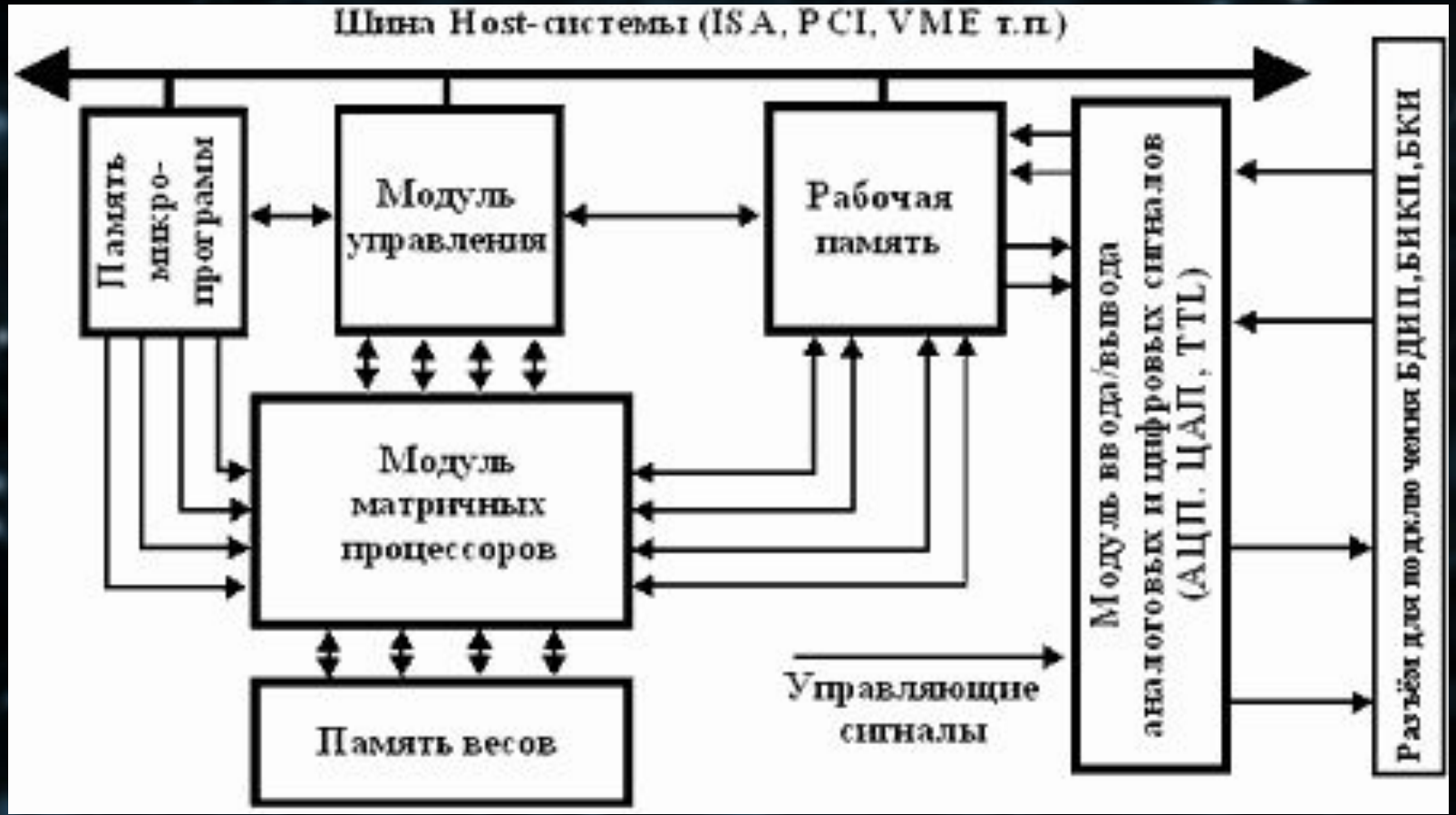
Особенности NM64003 (2)

- Наличие двух широких шин (по 64 разряда) для работы с внешней памятью любого типа: до 4Мб SRAM и до 16 Гб DRAM;
- Наличие двух байтовых коммуникационных портов ввода/вывода, аппаратно совместимых с коммуникационными портами TMS320C4x для реализации параллельных распределенных вычислительных систем большой производительности.
- Возможность работать с данными переменной разрядности по различным алгоритмам, реализуемым с помощью хранящихся во внешнем ОЗУ программ

Системы на NM 6403

- MC431 – однопроцессорная плата
- NM4 – четырехпроцессорная плата
- 6MCBO – 4 платы по 6 процессоров и платы для обработки входных сигналов

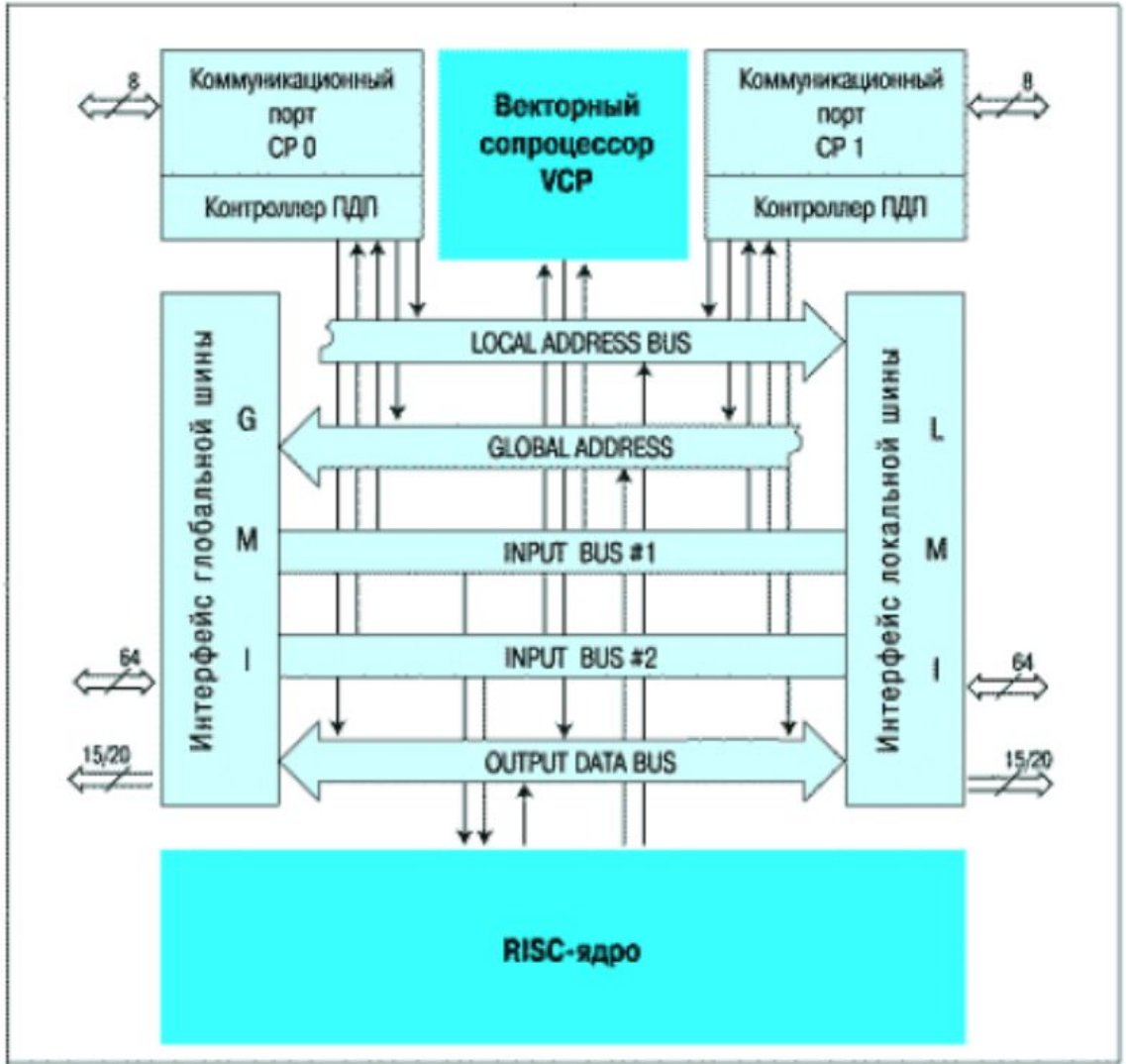
Схема нейровычислителя



Область применения

- Акселераторы для PC и рабочих станций
- Эмуляция нейронных сетей
- Сигнальная обработка
- Блок для построения болших суперпараллельных вычислительных систем и реализаиинейросетевых технологий
- Векторно-матричные вычислители

Архитектура процессора



Требования к аппаратуре

Компоненты БПО нейропроцессора предназначены для работы в среде Windows95. Все компоненты представляют собой консольные приложения Windows95. Интерфейс взаимодействия с пользователем представляет собой интерфейс командной строки.

Минимальные требования, необходимые для работы компонент:

1) объем оперативной памяти - 8 Мб (минимальное требование для работы Windows95),

2) наличие на компьютере установленной операционной системы Windows95,

3) не менее 20 Мб свободного дискового пространства

Редактор связей поддерживает различные варианты конфигурации памяти нейропроцессора.

Когда редактор связей обрабатывает входные объектные файлы, он выполняет следующие функции:

1) объединяет секции с одинаковыми именами и создает для них собственные таблицы перемещений, необходимые для перенастройки ссылок на конкретную конфигурацию памяти нейропроцессора,
2) в процессе построения исполняемых файлов с настройкой на конкретную конфигурацию нейропроцессора вычисляет адреса символов и секций, настраивает все ссылки, хранящиеся в таблицах перемещений,
3) объединяет загружаемые секции в программные сегменты для ускорения и упрощения загрузки программы в память нейропроцессора,
4) разрешает неопределенные внешние ссылки между входными файлами,
5) дает возможность удалять из выходного файла неиспользуемые программой секции и символы, а также отладочную информацию,
6) выдает информацию о найденных в процессе редактирования связей ошибках.

Несмотря на перечисленные преимущества эти устройства имеют ряд недостатков:

1. Они создаются специально для решения конкретных задач, связанных с нелинейной логикой и теорией самоорганизации. Решение таких задач на обычных компьютерах возможно только численными методами.
2. В силу своей уникальности эти устройства достаточно дорогостоящи.

ПРАКТИЧЕСКОЕ ПРИМЕНЕНИЕ НЕЙРОКОМПЬЮТЕРОВ

Несмотря на недостатки, нейрокомпьютеры могут быть успешно использованы в различных областях народного хозяйства.

— Управление в режиме реального времени: самолетами, ракетами и технологическими процессами непрерывного производства (металлургического, химического и др.);
— Распознавание образов: человеческих лиц, букв и иероглифов, сигналов радара и сонара, отпечатков пальцев в криминалистике, заболеваний по симптомам (в медицине) и местностей, где следует искать полезные ископаемые (в геологии, по косвенным признакам);
— Прогнозы: погоды, курса акций (и других финансовых показателей), исхода лечения, политических событий (в частности результатов выборов), поведения противников в военном конфликте и в экономической конкуренции;
— Оптимизация и поиск наилучших вариантов: при конструировании технических устройств, выборе экономической стратегии и при лечении больного.

Этот список можно продолжать, но и сказанного достаточно для того, чтобы понять, что нейрокомпьютеры могут занять достойное место в современном обществе.