Нейронные сети и их практическое применение.

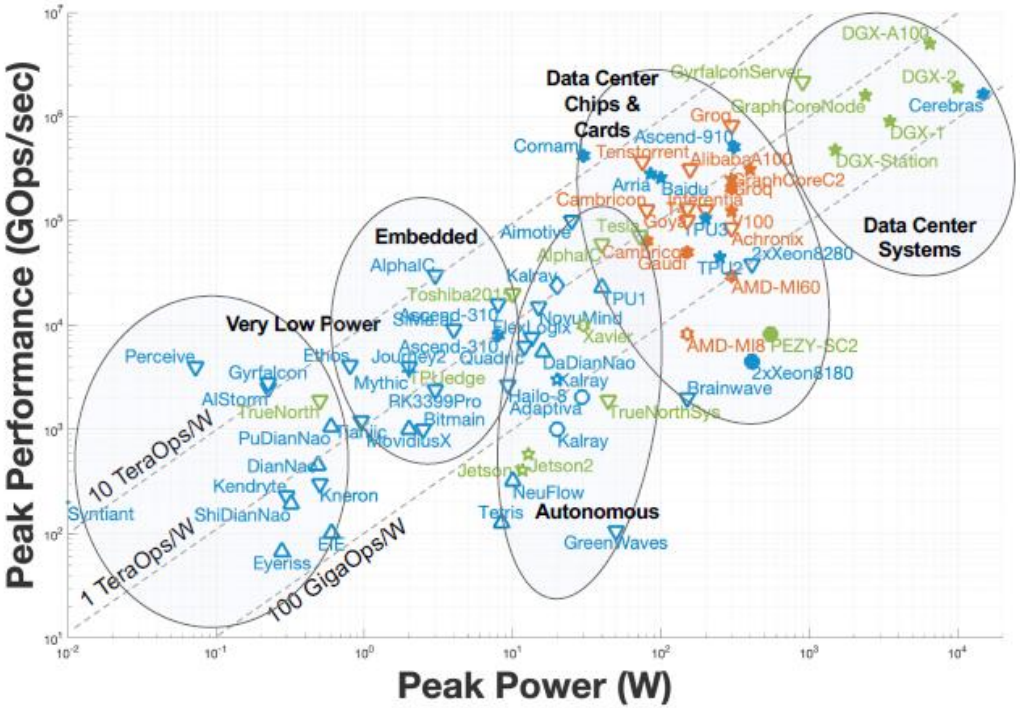
Лекция 10. Вычислительные платформы для реализации НС.

Дмитрий Буряк
к.ф.-м.н
dyb04@yandex.ru

Параметры вычислителей для НС

- Обучение и применение НС требует большой объем вычислений.
- Многие достижения в применении НС связаны с появлением новых вычислителей.
- Основные параметры для сравнения:
  - производительность
  - энергопотребление
- Факторы
  - разрядность: аналоговая, 1бит, …, 64бит.
  - форм-фактор – важна производительность одного чипа
  - обучение – вычисление (inference)

Ускорители для НС



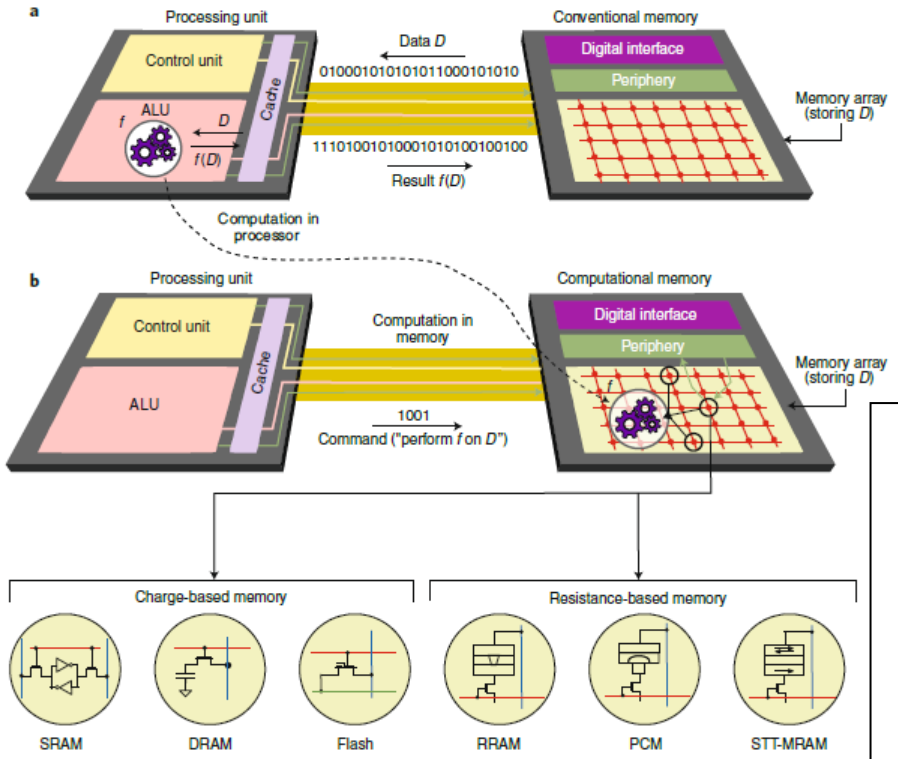
Зависимость пиковой производительности от потребляемой мощности НС ускорителей (A.Reuther, et al., Survey of Machine Learning Accelerators, 2020)

- 5 категория по типу применения
- низкое энергопотребление (обработка речи, маленькие сенсоры)
- встроенные (камеры, роботы)
- автономные (автопилоты в автомобилях)
- Чипы и карты для центров обработки данных
- Системы в центрах обработки данных

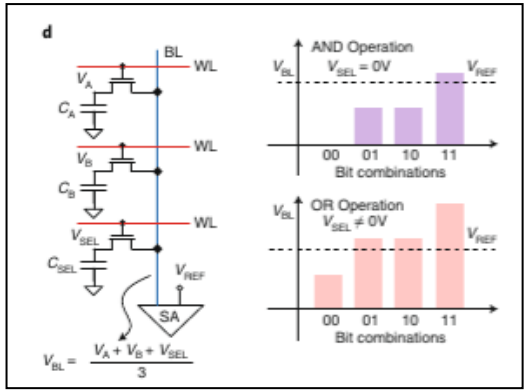
Тенденции

- ❑ Высокая плотность решений для всех областей применения
- ❑ Разнообразие архитектур и технологий
- ❑ Многие вычислители имеют эффективность > 1TeraOps/W
- ❑ Для реализации обучения требуется как минимум 100Вт
- ❑ Разнообразие решений по разрядности выполняемых вычислений
  - ▪ сложность сравнения вычислителей между собой
  - ▪ какая достаточная точность для выполнения вычислений в НС?

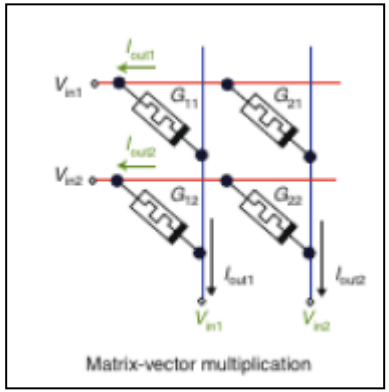
Processor-in-memory



- ❑ Затраты времени и энергии на передачу данных из памяти в процессор.
- ❑ Реализация вычислений в памяти.
- ❑ Веса кодируются в запоминающих устройствах: SRAM, Flash, RRAM и т.д.
- ❑ Умножение матриц реализовано через прохождение аналогового сигнала (тока) через запоминающие устройства.



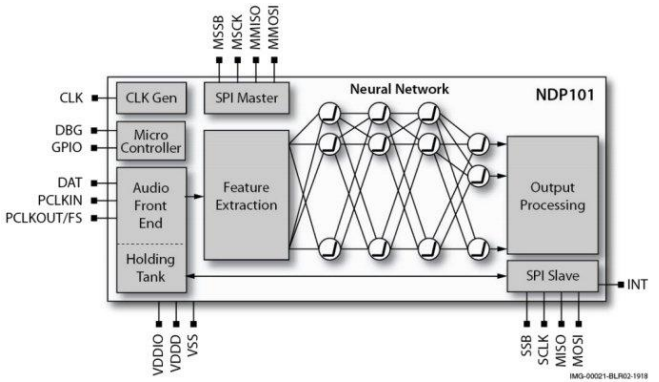
Пример реализации логических функций в ячейках памяти на основе заряда



Реализация умножения матрицы на вектор в резистивных ячейках

Чипы с низким энергопотреблением

- ❑ Предназначены только для применения НС.
- ❑ Syntiant
  - ▪ Processor-in-memory
  - ▪ NDP101
    - • НС с 4 слоями
    - • веса (int4), активации (int8)
    - • энергопотребление <200мкВт
    - • Задачи обработки 1D сигналов
    - • Amazon Alexa – распознавание ключевой фразы.
  - ▪ NDP200
    - • «Любые» НС
    - • Добавление DSP
    - • Задачи анализа изображений.

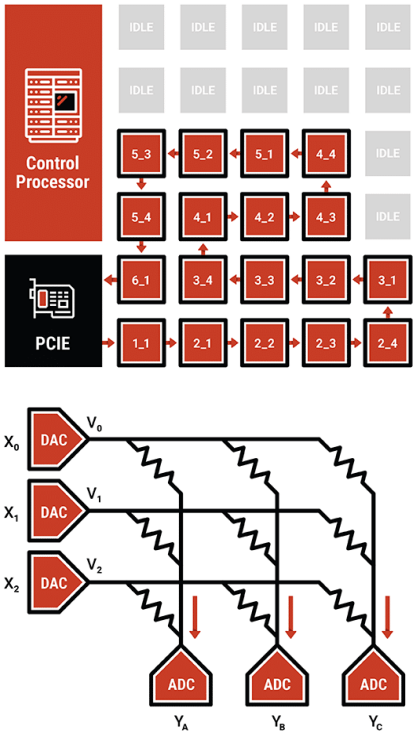


Чипы для встроенных решений

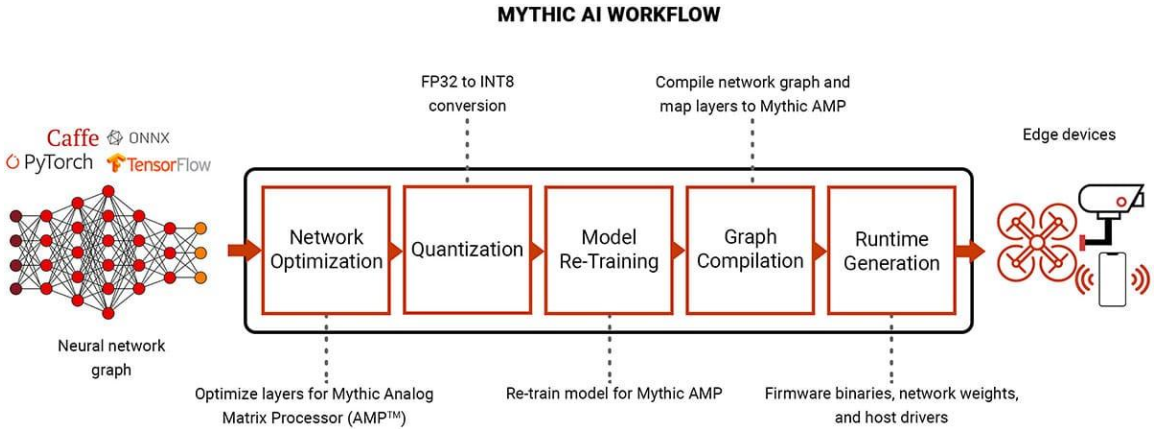
❑ Встроенные приложения с высокой производительностью при небольшом энергопотреблении и малом форм-факторе: обработка видео, небольшие БЛА и роботы.

❑ Mythic Intelligent Processing Unit (25 TOPs)- встроенные системы и центры обработки данных.

- Распараллеливание вычисления графа НС
- Архитектура потоков данных
- аналоговая схема для реализации матричного умножения
- цифровой управляющий процессор RISK-V
- декомпозиция исходных данных для параллельной обработки отдельными модулями (tiles)



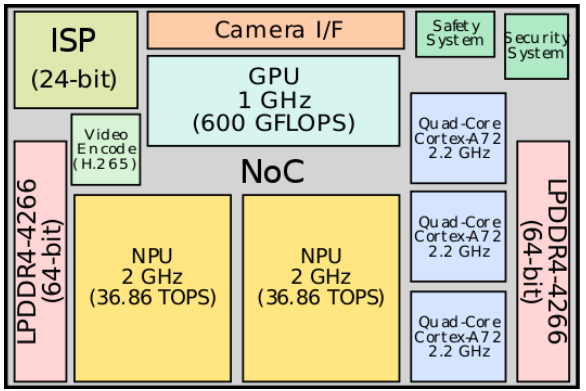
Разработка НС для энергоэффектиных чипов



- ❑ Разработка исходной НС (PyTorch, TensorFlow)
- ❑ Оптимизация, квантизация (Int8)
- ❑ Дообучение
- ❑ Генерация кода

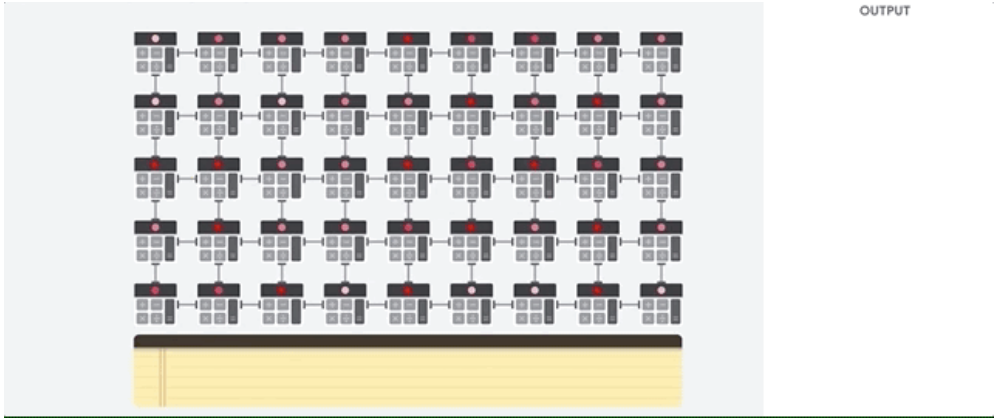
Автономные системы

- Системы помощи водителю, автопилот, роботы.
- Совместная обработка видео и сигналов других сенсоров
- aiMotive
  - программируемые FPGA
  - масштабируемость до 1024 TOPS
- Tesla Full Self-Driving (FSD) Computer chip
  - 3 quard Cortex-A72 + Mali G71 + 2 Neural Processing Unit
  - применение в беспилотных системах (уровни автономности 4 и 5)
  - Пиковая производительность 73TOPS @72w.
- Модули NVIDIA Xavier
  - GPU: 256 ядер CUDA (Volta) + 64 tensor cores
  - CPU: ARM 8 ядер;
  - Память: 32Гб;
  - 32 TOPs, 10Вт.



Чипы для центров обработки данных.

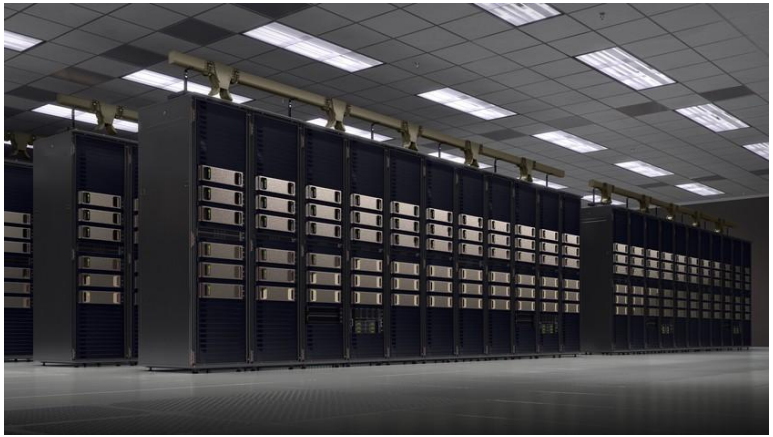
- Применение различных технологий: на основе CPU, программируемые FPGA.
- NVIDIA (Blackwell,Hopper)
  - Тензорные ядра
    . Умножения матриц за такт;
    . Ускорение операций с разреженными матрицами
- Специализированные для НС: Google TPU
  - Последовательность вычислений в НС детерминирована;
  - Проектирование HW с учетом вычислений и обращений к памяти в НС: уменьшить обращение к памяти при умножении матриц
  - GPU (общее назначение) →Google TPU (tensor processing unit) – матричный процессор
- Сравнение производительности на LLM
- Потребление – сотни ватт.



https://habr.com/ru/post/422317/

HPC системы для центров обработки данных

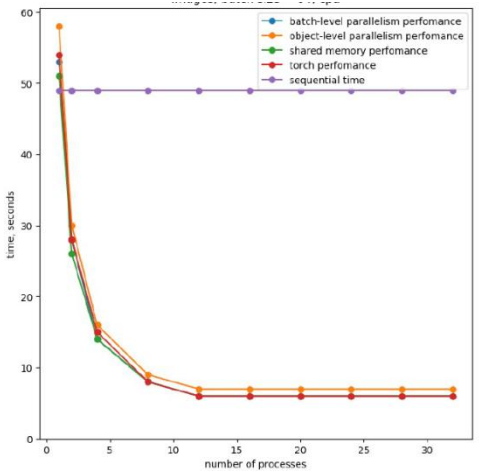
- Объединение нескольких GPU на базе серверного решения.
- Набор узлов: CPU, несколько GPU, высокоскоростной интерфейс
- ПО для реализации распределенных вычислений
- NVIDIA DGX SuperPod - масштабируемая архитектура ЦОД для задач ИИ
  - Стойки серверов DGX:
    - 36 узлов (Grace CPU+2xBlackwell GPU)
  - Сетевое оборудование;
  - Устройства хранения данных;
  - Электропитание, охлаждение;
  - Полный программный стек для ИИ.
- SATURNV – пример реализации на основе DGX Pod (116 в Top500, 11.2021)



Вычислительный центр NVIDIA SATURNV

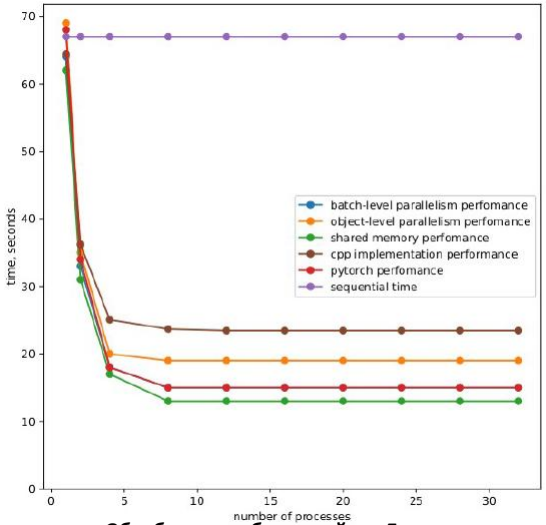
Вычислительный кластер МГУ

- ❑ МГУ-270. Характеристики узла:
  - Процессор: AMD EPYC 7742 64-Core Processor
  - Оперативная память: 2 TB
  - Видеокарты: 8xNvidia A100 80GB
  - Память (ПЗУ): 28 TB.
- ❑ Спроектирован для ИИ вычислений
- ❑ Пиковая производительность: 400 ПФлопс



Обработка изображений на МГУ-270

- ❑ Ломоносов-2. Характеристики узла:
  - Процессор: Intel Haswell-EP E5-2697v3, 14 cores
  - Оперативная память: 64 GB
  - Видеокарты: NVidia Tesla K40M
- ❑ Общее количество узлов: 1730
- ❑ Пиковая производительность: 5.5 ПФлопс

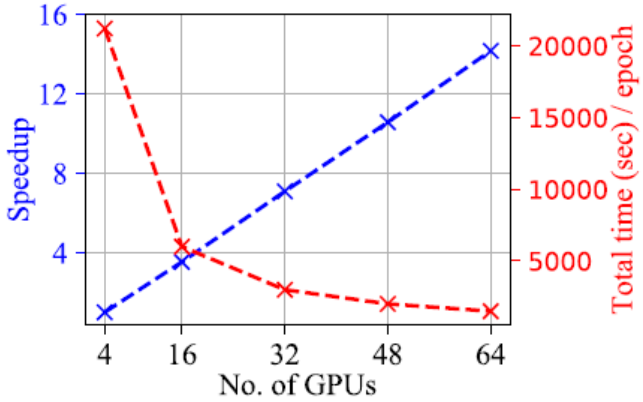


Обработка изображений на Ломоносов

Зачем HPC для НС

- Уменьшение времени на обучение НС.
- Возможность построения больших моделей
  - увеличенный размер пакета;
  - увеличенная размерность входных данных;
  - больше параметров в модели.

- Победитель ILSVRC -2017 – SENet
  - Реализация распределенного обучения на 64 GPU: размер пакета 2048 изображений.
- Обучение сети для обработки гравитационных волн
  - HAL кластер (64 GPU NVIDIA V100): сокращение времени обучения с 1 месяца до 12.4ч.



Время обучения сети для анализа гравитационных волн (E.A. Huerta, et al., Convergence of artificial intelligence and high Performance computing on NSF-supported cyberinfrastructure, 2020)