

Нейронные сети и их практическое применение.

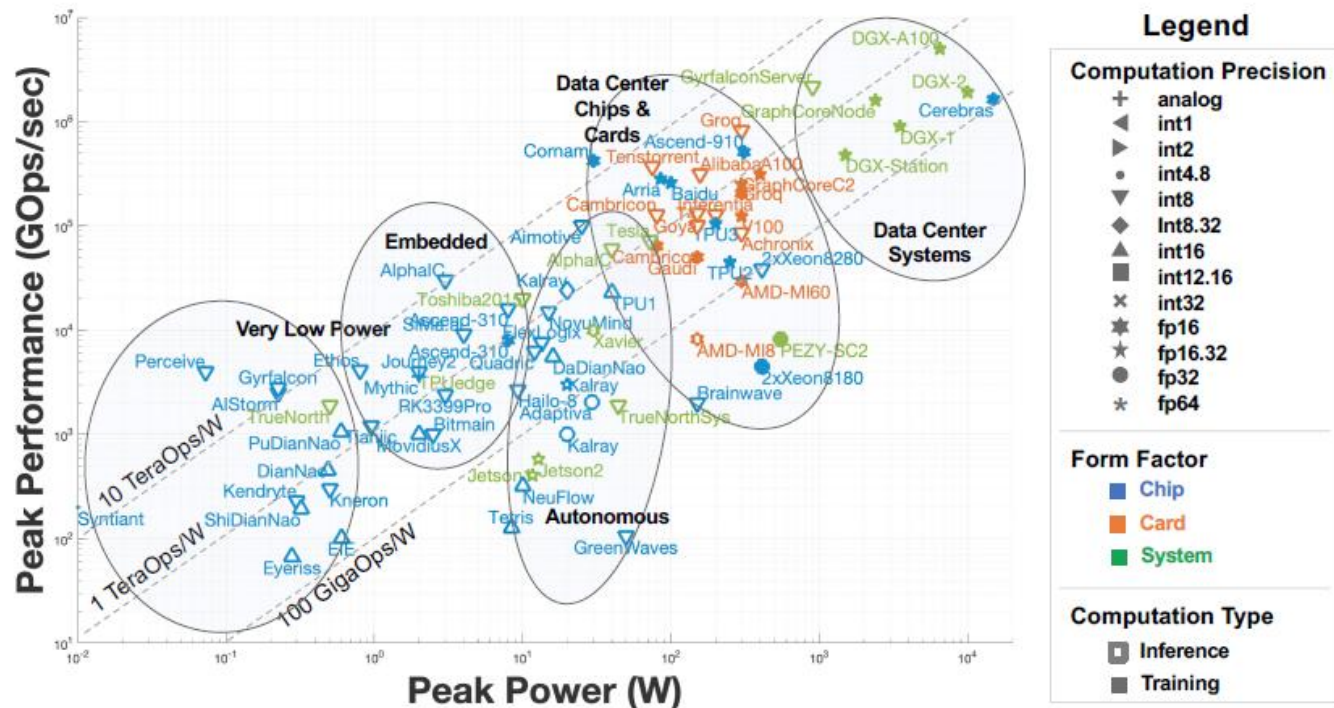
Лекция 10. Вычислительные платформы для реализации НС.

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 ИС



Зависимость пиковой производительности от потребляемой мощности ИС 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 $> 1\text{TeraOps/W}$
- ❑ 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?

Чипы с низким энергопотреблением

❑ Предназначены только для применения НС.

❑ Syntiant

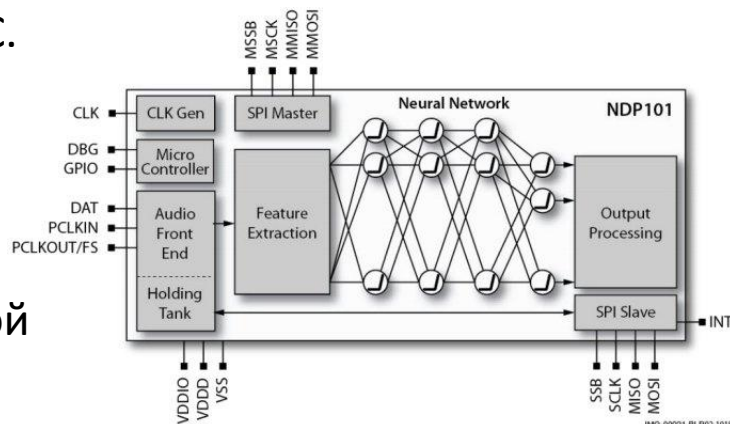
- Processor-in-memory
- веса (int4), активации (int8)
- < 200мВт
- Amazon Alexa – распознавание ключевой фразы.

❑ Mythic Intelligent Processing Unit

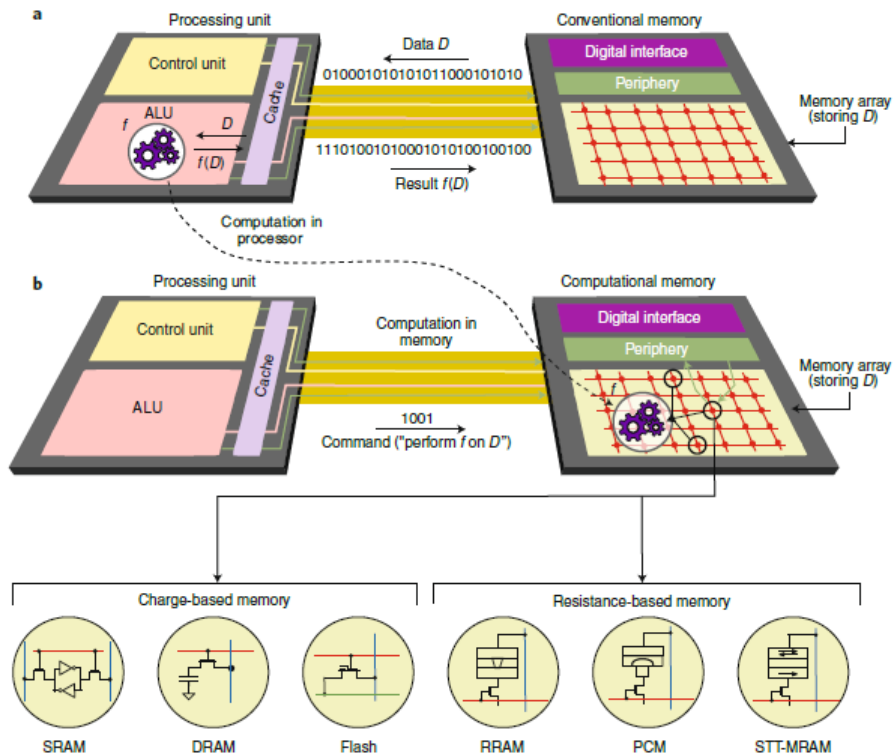
- аналоговая схема для реализации матричного умножения (~Syntiant)
- цифровой управляющий процессор RISK-V
- встроенные системы и центры обработки данных.

❑ AIStorm

- вычисления выполняются «на» сенсоре в аналоговом виде (~Syntiant)
- обработка сигналов биометрических сенсоров.



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.

Чипы для встроенных решений

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

❑ ARM, процессоры Etos

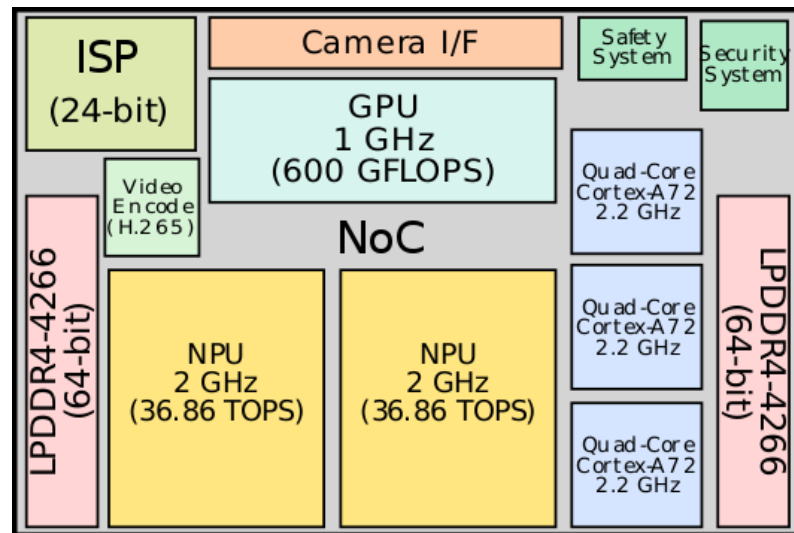
- Применение MAC Compute Engines (MCEs)
- Производительность MCE: 1TOP/s на 1GHz.

❑ Gyr Falcon, Lightspeed

- processor-in-memory
- обработка видео
- применение в серверах обработки данных (128 чипов).

Автономные системы

- ❑ Системы помощи водителю, автопилот, роботы.
- ❑ Huawei HiSilicon Ascend 310
 - CPU + ускоритель для вычислений AI
 - Высокая производительность, как с int8, так и с fp16.
- ❑ Tesla Full Self-Driving (FSD) Computer chip
 - 3 quad Cortex-A72 + Mali G71 + 2 Neural Processing Unit
 - применение в беспилотных системах (уровни автономности 4 и 5).



Автономные системы. Nvidia Jetson

❑ Встраиваемые системы

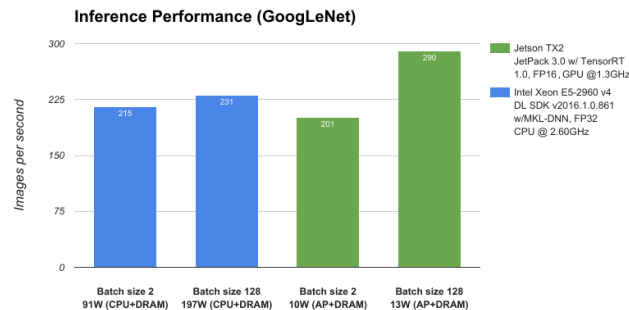
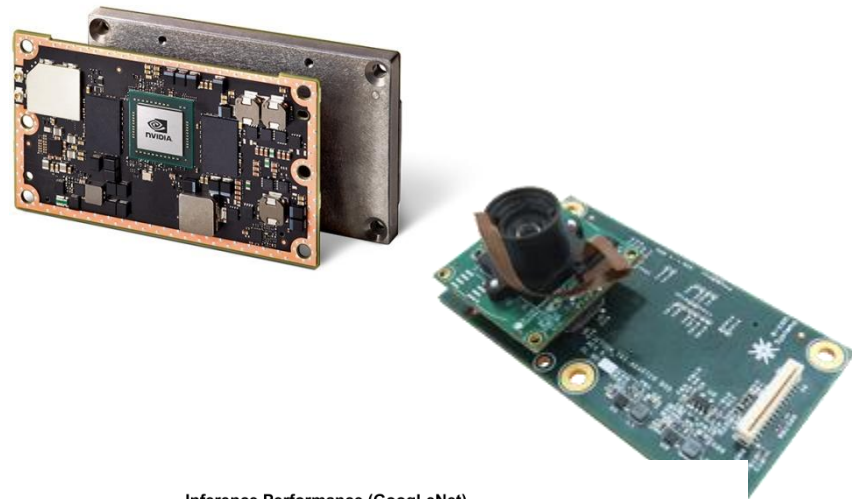
- небольшой размер;
- низкий уровень энергопотребления;
- высокопроизводительные вычисления.

❑ Модули NVIDIA Jetson Tx2

- GPU: 256 ядер CUDA (Pascal);
- CPU: ARM 6 ядер;
- Память: 8Гб.

❑ 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, с применением GPU.
- ❑ Intel Xeon.
- ❑ Preferred Networks, PFN-MN-3
 - 4 чипа + 32Гб RAM
 - Каждый чип имеет 4 матрицы из 2048 вычислительных элементов и 512 блоков матричных вычислений.
 - Поддержка вычислений fp16;
 - Применяются в одном из японских суперкомпьютеров (exascale).
- ❑ Intel Arria
 - Intel Xeon + Altera Arria FPGA
 - Загрузка FPGA конфигурации, HC вычисляется на FPGA

Чипы для центров обработки данных. GPU карты

❑ NVIDIA (Ampere, Turing, HGX)

▪ Тензорные ядра

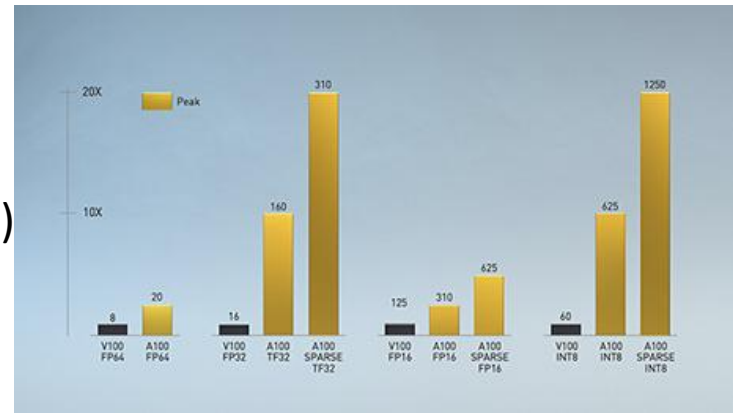
- . Предназначены для умножения матриц за такт;
- . Volta, 4x4, fp16 (64GEMM) → Ampere (256GEMM)
- . Обработка разреженных матриц: Sparse Tensor Cores;

▪ Turing – эффективное вычисление HC;

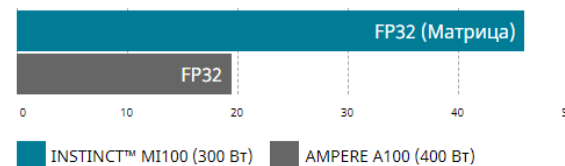
▪ HGX – до 16 A100 на одной карте

❑ AMD (Instinct)

- MI100;
- MI250 (128Гб, 2 GPU чипа, 560Вт, 95.7 TFLOPS матричная производительность).



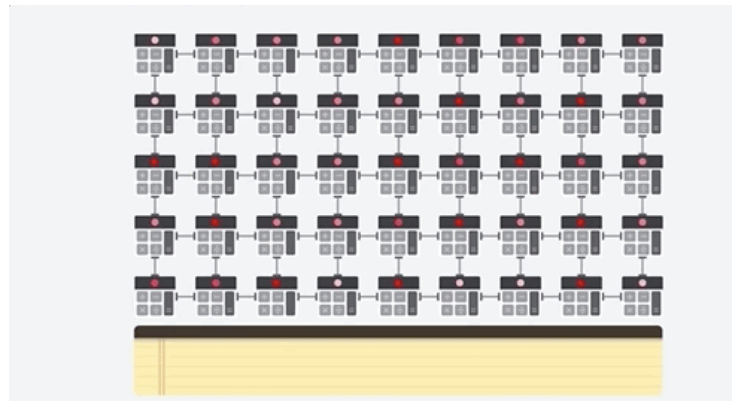
(Пиковая производительность в терафлопсах) в вычислениях для чисел смешанной точности¹



Сравнение MI100 и A100 на задачах машинного обучения
(<https://www.amd.com/ru/products/server-accelerators/instinct-mi100>)

Специальные чипы для центров обработки данных.

- ❑ Разработаны специально для обучения и применения НС
 - Последовательность вычислений в НС детерминирована;
 - Проектирование HW с учетом вычислений и обращений к памяти в НС
 - Применение в облачных платформах
- ❑ Amazon Web Services: Inferentia chip
- ❑ Baidu: Kunlun
- ❑ Google: TPU (tensor processing units)
 - Появление 2015г.
 - Архитектура конвейерного массива – уменьшить обращение к памяти при умножении матриц;
 - Меньше энергопотребление и размер чипа.



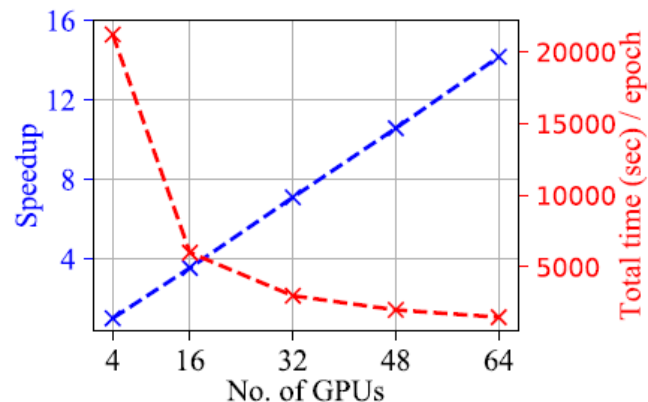
HPC системы для центров обработки данных

- ❑ Объединение нескольких GPU на базе серверного решения.
- ❑ Набор узлов: CPU, несколько GPU, высокоскоростной интерфейс (например, NVLink).
- ❑ Соединение между узлами: Infiniband/Ethernet
- ❑ ПО для реализации распределенных вычислений

- ❑ Области применения
 - искусственный интеллект
 - облачные центры обработки данных
 - корпоративные вычисления
 - высокопроизводительные вычисления

Зачем HPC для ИС

- ❑ Уменьшение времени на обучение ИС.
- ❑ 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)

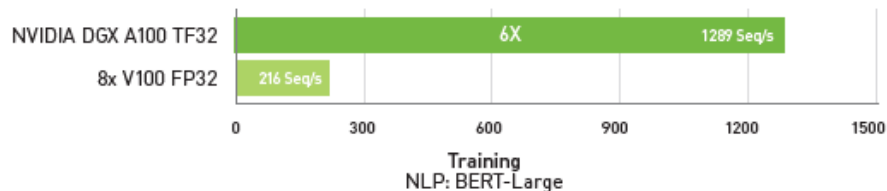
Семейство DGX

❑ DGX-1, DGX-2, DGX A100.

❑ Параметры DGX A100

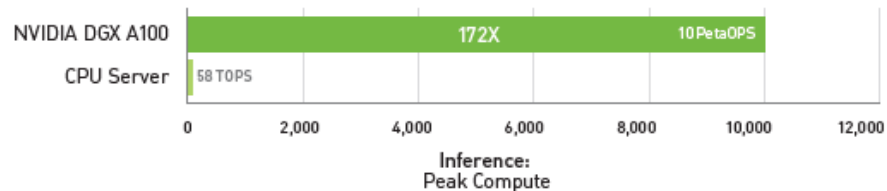
- 5 Pflops на задачах AI;
- 8xNVIDIA A100 Tensor Cores GPU;
- Размер общей памяти GPU: 320Gb;
- CPU: Dual AMD Rome 7742 (128 cores), 2.25GHz;
- Обмен GPU-GPU: 600GB/s (12 NVLinks/GPU);
- Между узлами: HDR Infiniband, 200GB/s;
- Общий объем памяти: 1TB.

DGX A100 Delivers 6 Times The Training Performance



BERT Pre-Training Throughput using PyTorch including [2/3]Phase 1 and [1/3]Phase 2 | Phase 1 Seq Len = 128, Phase 2 Seq Len = 512 | V100: DGX-1 with 8x V100 using FP32 precision | DGX A100: DGX A100 with 8x A100 using TF32 precision

DGX A100 Delivers 172 Times The Inference Performance



CPU Server: 2x Intel Platinum 8280 using INT8 | DGX A100: DGX A100 with 8x A100 using INT8 with Structural Sparsity

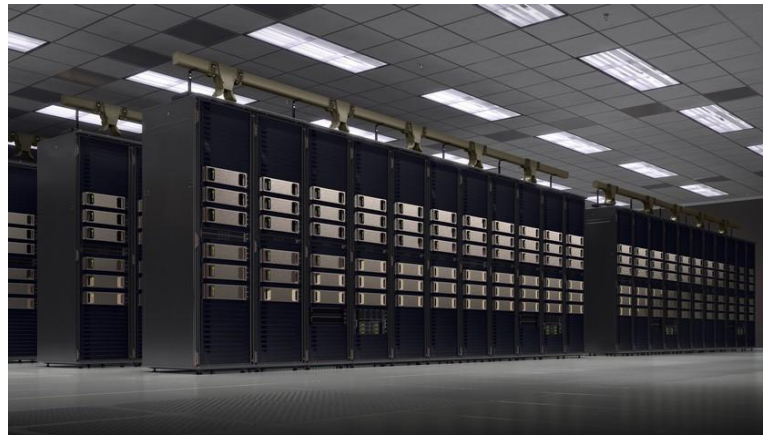
NVIDIA DGX Pod

❑ Масштабируемая архитектура ЦОД для задач ИИ

- Стойки серверов DGX;
- Сетевое оборудование;
- Устройства хранения данных;
- Электропитание, охлаждение;
- Полный программный стек для ИИ.

❑ SATURNV – пример реализации на основе DGX Pod (116 в Top500, 11.2021)

❑ NVIDIA DGX SuperPod – второе поколение.



Вычислительный центр NVIDIA SATURNV

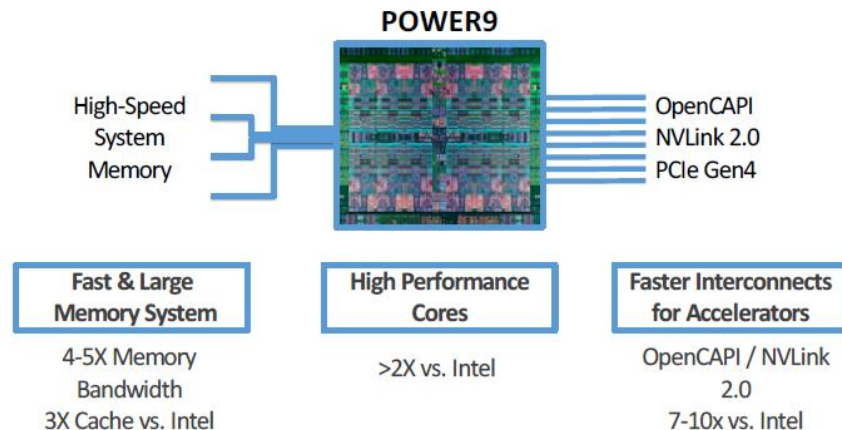
Платформа IBM Power

❑ Базовая платформа для HPC серверов

- Анализ больших данных, высокопроизводительные вычисления, ИИ;
- Масштабируемость: сервер → суперкомпьютер

❑ IBM Power9

- коммутаторные межсоединения;
- аппаратная поддержка ускорителей;
- расширенная система команд (ИИ, когнитивные вычисления);
- высокопроизводительные вычисления.



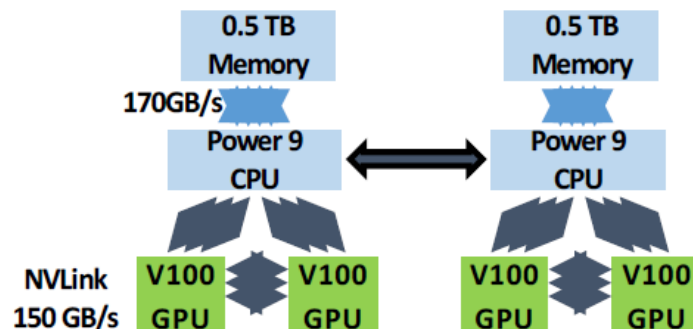
❑ IBM Power10

- Вычислительная эффективность до 20 раз выше чем Power9;
- Энергоэффективность до 3 раз выше, чем Power9;
- Серверные решения – вторая половина 2021г.

НРС на базе IBM Power9

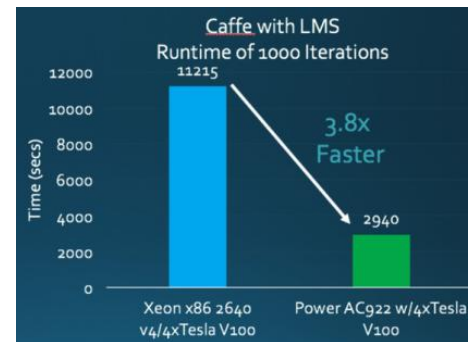
❑ IBM AC922 Power System

- Технология объединения памяти CPU и GPU;
- Поддержка больших моделей, больших наборов данных;
- Ускорение 95% от линейного на задачах глубокого обучения.



❑ Сравнение с Xeon 2640

- Ускорение передачи данных CPU-GPU ~5 раз;
- Увеличение скорости вычислений в 3.8 раз.



IBM PowerAI

❑ PowerAI – набор дистрибутивов популярных библиотек для глубокого обучения, оптимизированных для серверных решений IBM;

- TensorFlow, Caffe, Torch, ...;

❑ Distributed Deep Learning (DDL) – коммуникационная библиотека (на базе MPI) для эффективной реализации распределенных вычислений;

- Близкое к линейному масштабирование (до 256 GPU);

❑ Large Model Support (LMS) – совместное использование основной памяти и памяти GPU при обучении больших моделей.

