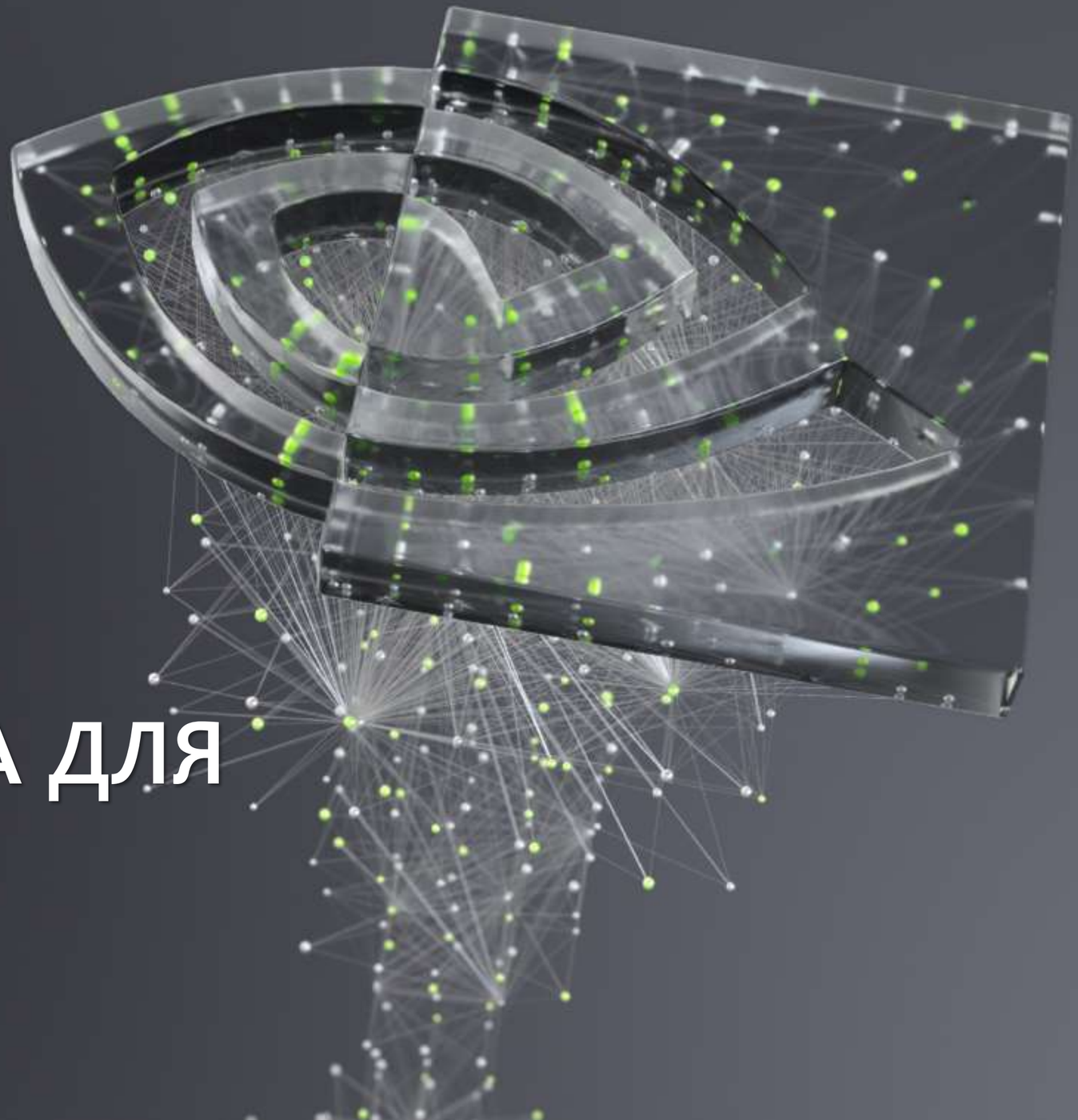
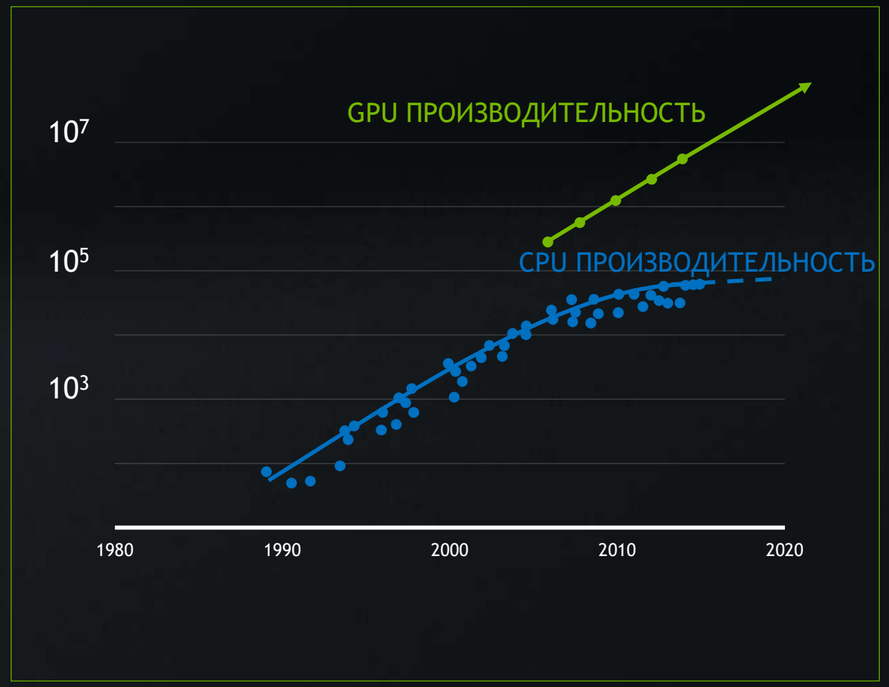


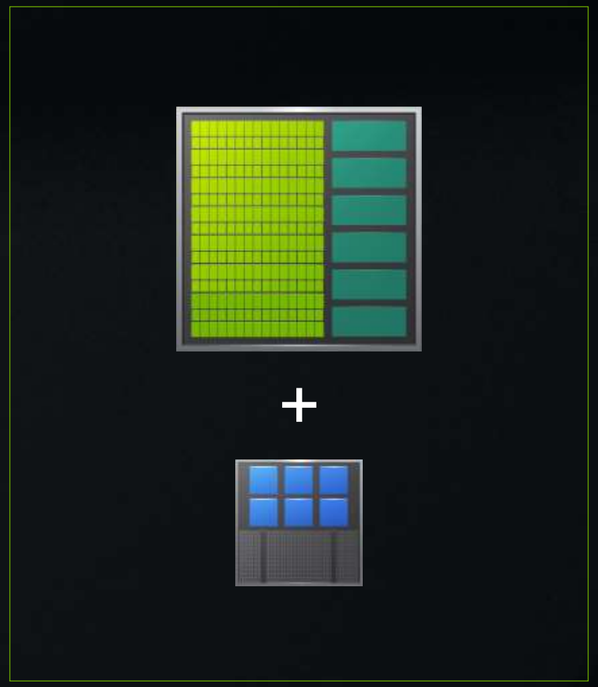
ПЛАТФОРМА NVIDIA ДЛЯ HPC И AI

Антон Джораев | adzhoraev@nvidia.com

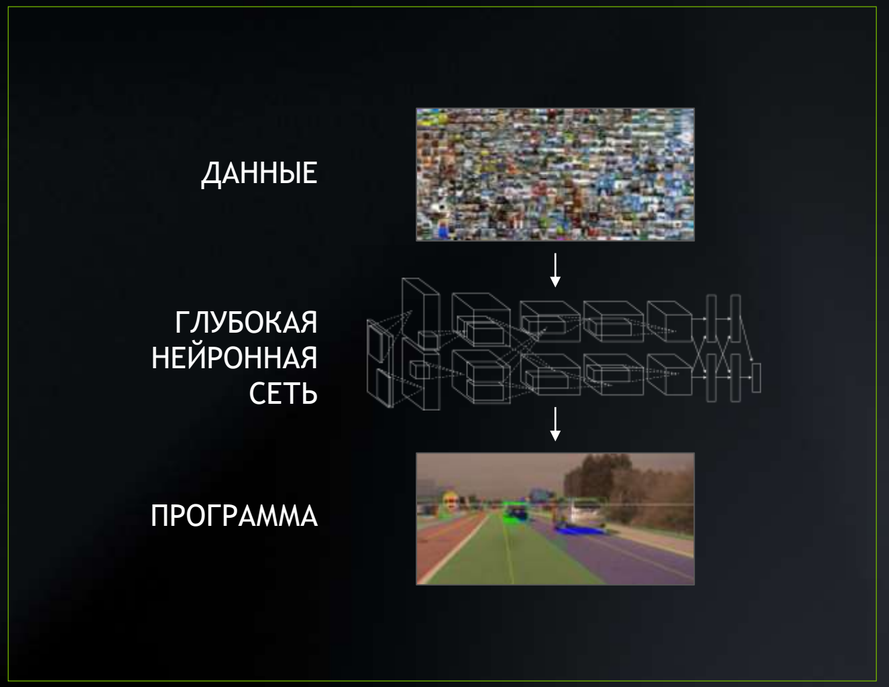
ВЫЧИСЛИТЕЛЬНЫЙ ЛАНДШАФТ СЕГОДНЯ



ЖИЗНЬ ПОСЛЕ ЗАКОНА МУРА



УСКОРЕННЫЕ ВЫЧИСЛЕНИЯ



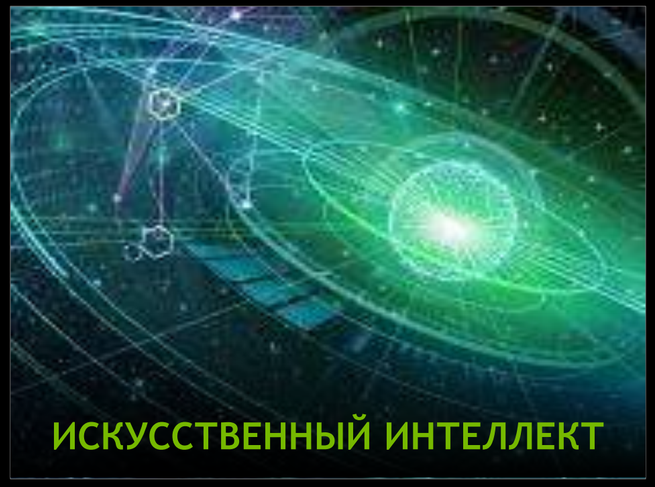
КОМПЬЮТЕРЫ СОЗДАЮТ ПРИЛОЖЕНИЯ

ПЛАТФОРМА NVIDIA ДЛЯ УСКОРЕННЫХ ВЫЧИСЛЕНИЙ

Быстрое решение актуальных задач на GPU

SPARK3.0 | RAPIDS | и другие

cuDF | cuML | cuGRAPH



TensorFlow | PyTorch | и другие

cuDNN| TensorRT | NCCL

NAMD | GROMACS | +700 приложений

cuBLAS| cuFFT | cuSOLVER

Разработка на рабочей станции

Размещение в ЦОД

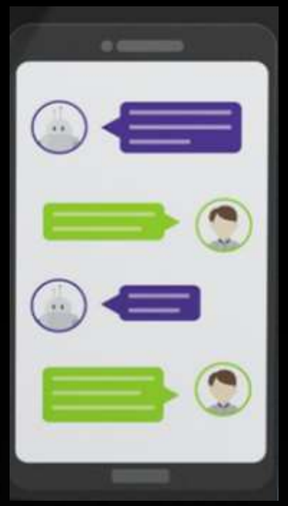
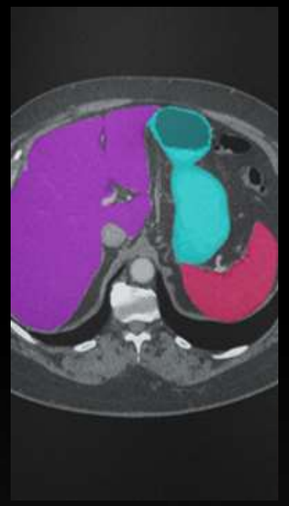
Ускоренные edge-вычисления

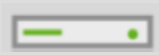
Суперкомпьютеры

GPU-ускоренные облака

КРАТЧАЙШИЙ ПУТЬ К ГОТОВОМУ ПРОДУКТУ НА БАЗЕ ИИ

Фреймворки для быстрого создания ваших собственных решений

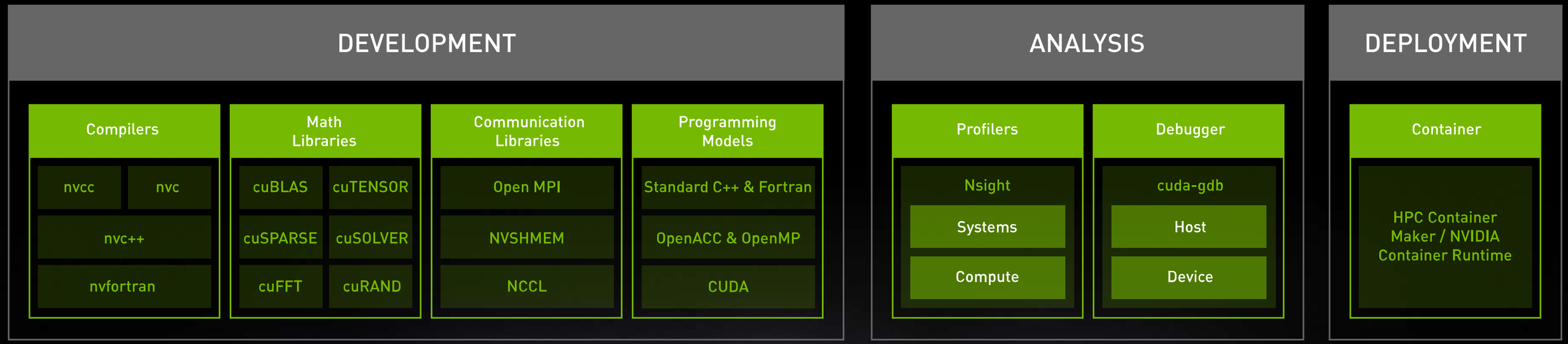
Разговорный ИИ	Рекомендательные системы	Умные города	Здравоохранение	Робототехника	Автономные автомобили	Телекоммуникации
						
Jarvis	Merlin	Metropolis	Clara	Isaac	Drive	Aerial

Разработка на рабочей станции	Размещение в ЦОД	Ускоренные edge-вычисления	Суперкомпьютеры	GPU-ускоренные облака
				

ФРЕЙМВОРК ДЛЯ ВЫСОКОПРОИЗВОДИТЕЛЬНЫХ ВЫЧИСЛЕНИЙ НА GPU

доступен свободно на developer.nvidia.com/hpc-sdk

NVIDIA HPC SDK



NVIDIA HPC Platform: GPU, CPU and Interconnect

HPC Libraries | GPU Accelerated C++ and Fortran | Directives | CUDA

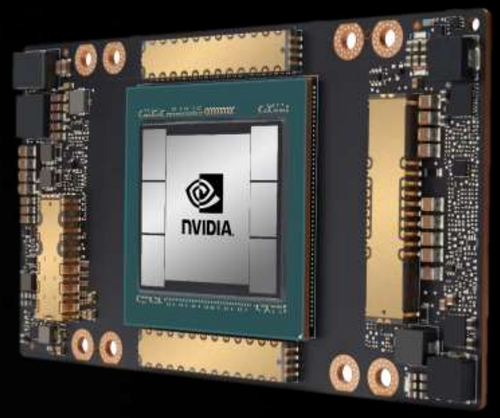
Compatible with 99% of Top500

NVIDIA A100

Большой скачок вперед: до 20 раз выше производительность в сравнении с Volta

	Peak		Vs Volta
FP32 TRAINING	312	TFLOPS	20X
INT8 INFERENCE	1,248	TOPS	20X
FP64 HPC	19.5	TFLOPS	2.5X
MULTI INSTANCE GPU			7X GPUs

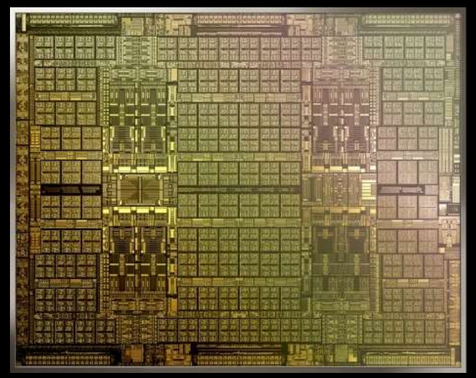
A100 PCIe



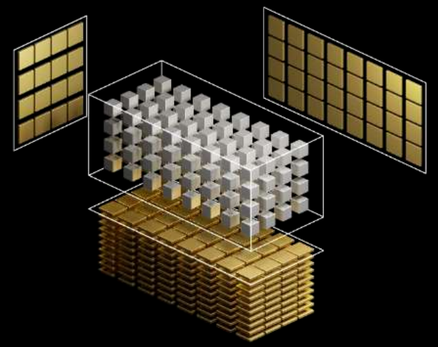
A100 SXM

54B XTOR | 826mm2 | TSMC 7N | 40GB Samsung HBM2

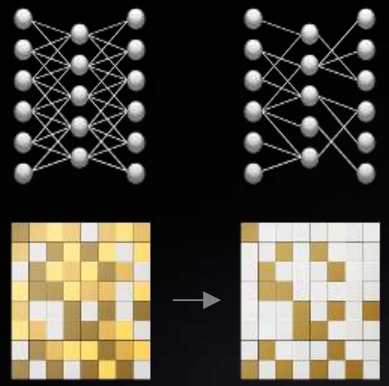
5 ОСНОВНЫХ ТЕХНОЛОГИЙ A100



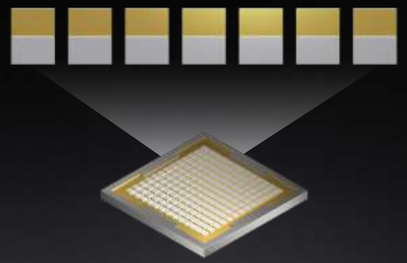
Ampere
World's Largest 7nm chip
54B XTORS, HBM2



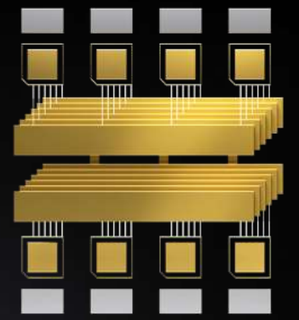
3rd Gen Tensor Cores
Faster, Flexible, Easier to use
20x AI Perf with TF32



New Sparsity Acceleration
Harness Sparsity in AI Models
2x AI Performance



New Multi-Instance GPU
Optimal utilization with right sized GPU
7x Simultaneous Instances per GPU



3rd Gen NVLINK and NVSWITCH
Efficient Scaling to Enable Super GPU
2X More Bandwidth

A100 ДОСТУПНЫ В ВАРИАНТАХ PCI-E И SXM4

A100 PCIe

Для обычных серверов

Энергопотребление 250 Вт

1-8 GPU на сервер, опциональный NVLink мост между парами GPU

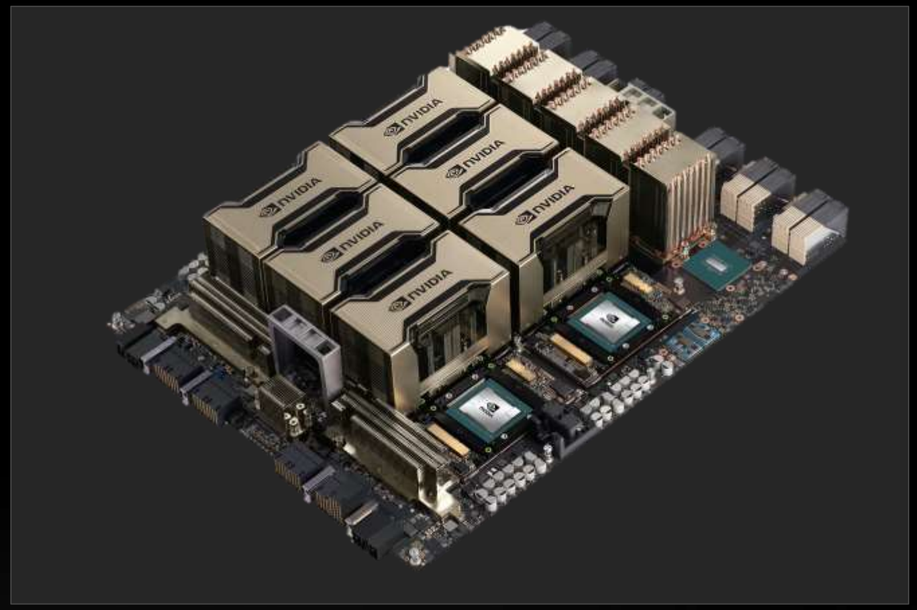
A100 SXM4 [4-GPU]

Масштабные вычисления, смесь AI и HPC нагрузок

Энергопотребление 400 Вт

4 процессора A100, объединенные с помощью NVLink

A100 SXM4 [8-GPU]

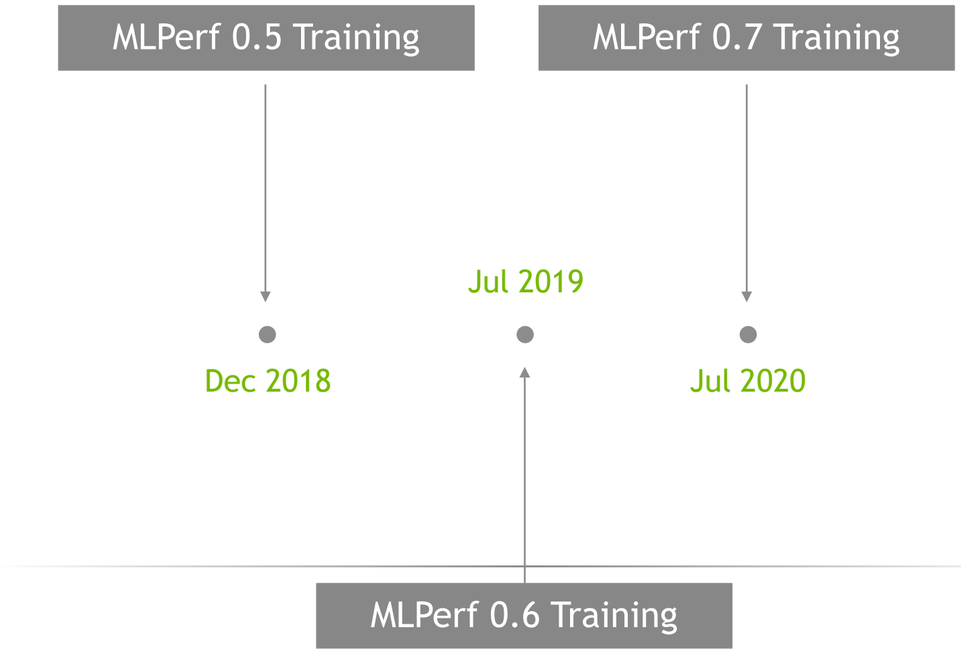


Масштабные вычисления, максимальная скорость обучения ИИ

Энергопотребление 400 Вт

8 процессоров A100, полная производительность шины NVLink между всеми GPU благодаря NVSwitch

MLPerf - отраслевой бенчмарк производительности ИИ-систем



MLPerf

Fair and useful benchmarks for measuring training and inference performance of ML hardware, software, and services.

Компании - контрибьюторы

Исследовательские организации - контрибьюторы

https://mlperf.org/

MLPERF – ОТРАСЛЕВОЙ БЕНЧМАРК ИИ СИСТЕМ

Три новых теста MLPerf v0.7 покрывают все актуальные направления ИИ: рекомендательные системы, языковые модели, обучение с подкреплением

Новые тесты

New	Recommendation Systems	DLRM
New	NLP	BERT
Enhanced	Reinforcement Learning	MiniGo Full – 19x19
	Translation (non-recurrent)	Transformer
	Translation (recurrent)	GNMT
	Object Detection (Heavy Weight)	Mask R-CNN
	Object Detection (Light Weight)	Single-Shot Detector with ResNet-34
	Image Classification	ResNet-50 v1.5

Множество конфигураций

От ЦОД до облака, от сервера до суперкомпьютера

Единая метрика

Время обучения модели до заданной точности

https://mlperf.org/

NVIDIA: ВСЕ 16 РЕКОРДОВ ПРОИЗВОДИТЕЛЬНОСТИ ДЛЯ ИИ

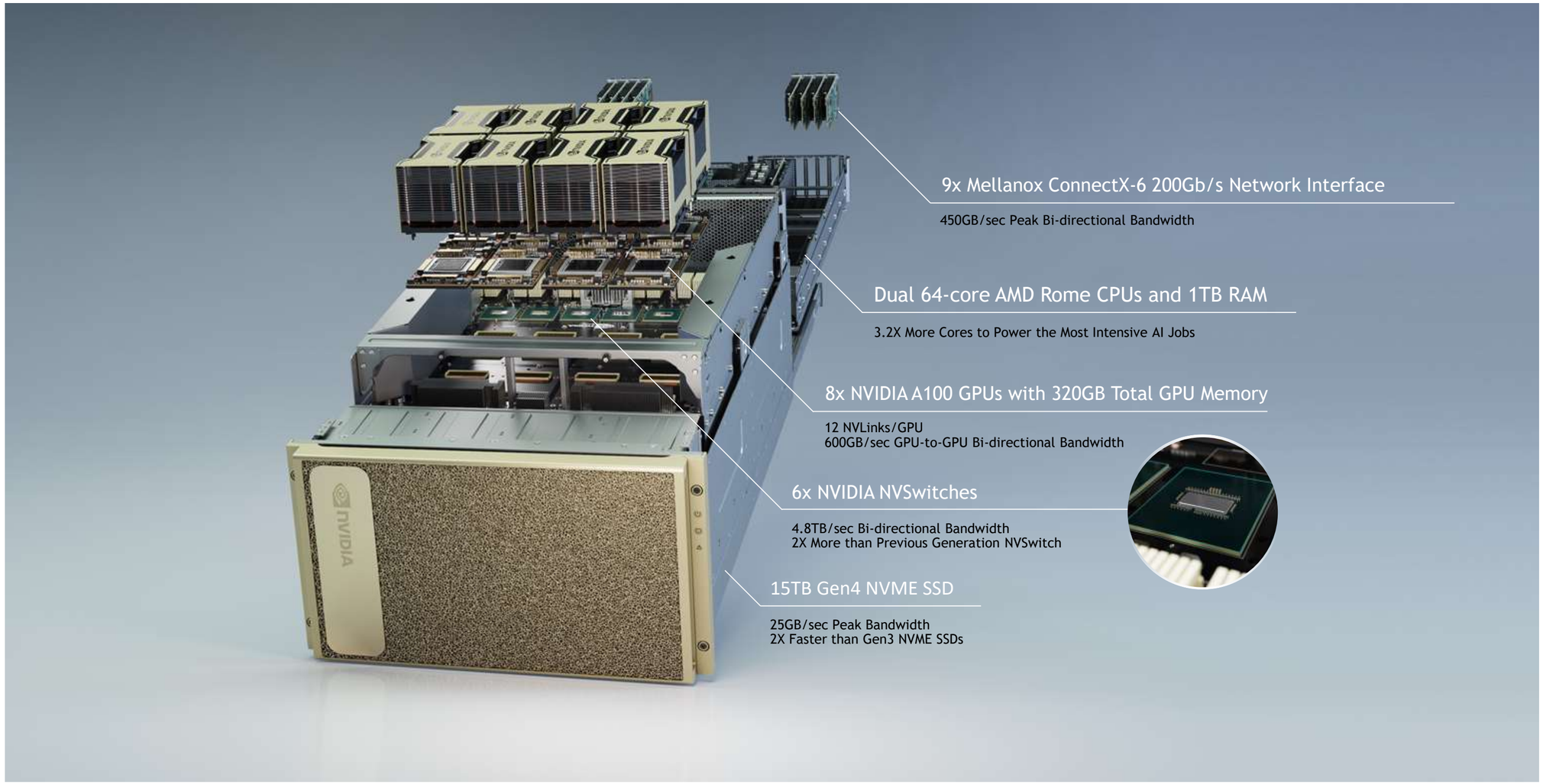
Среди коммерчески доступных продуктов

Тест	Масштабные вычисления (DGX SuperPOD, кластер DGX A100)	Расчет на одном процессоре (NVIDIA A100)
Recommendation (DLRM)	3.33 Min	0.44 Hrs
NLP (BERT)	0.81 Min	6.53 Hrs
Reinforcement Learning (MiniGo)	17.07 Min	39.96 Hrs
Translation (Non-recurrent) Transformer	0.62 Min	1.05 Hrs
Translation (Recurrent) GNMT	0.71 Min	1.04 Hrs
Object Detection (Heavy Weight) Mask R-CNN	10.46 Min	10.95 Hrs
Object Detection (Light Weight) SSD	0.82 Min	1.36 Hrs
Image Classification (ResNet-50 v1.5)	0.76 Min	5.30 Hrs

Per Chip Performance arrived at by comparing performance at the same scale when possible. Per Accelerator comparison using reported performance for MLPerf 0.7 NVIDIA A100 (8 A100s). MLPerf ID DLRM: 0.7-17, ResNet50 v1.5: 0.7-18, 0.7-15 BERT, GNMT, Mask R-CNN, SSD, Transformer: 07-19, MiniGo: 0.7-20. Max Scale: All results from MLPerf v0.7 using NVIDIA DGX A100 (8xA100s)\. MLPerf ID Max Scale: ResNet50 v1.5: 0.7-37, Mask R-CNN: 0.7-28, SSD: 0.7-33, GNMT: 0.7-34, Transformer: 0.7-30, MiniGo: 0.7-36, BERT: 0.7-38, DLRM: 0.7-17.

MLPerf name and logo are trademarks. See www.mlperf.org for more information.

DGX A100: ПРОИЗВОДИТЕЛЬНОСТЬ И УДОБСТВО



https://www.youtube.com/watch?v=TJcKYUTaBtg

СПЕЦИФИКАЦИЯ NVIDIA DGX A100

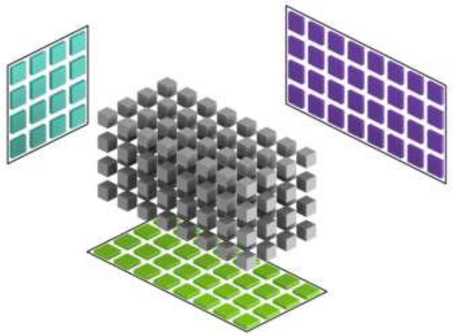
	Ключевые компоненты
GPU	8x NVIDIA A100 Tensor Core GPUs
GPU память	320GB Total
NVIDIA NVSwitch	6
Производительность	5 petaFLOPS AI 10 petaOPS, INT8
CPU	Dual AMD Rome, 128 cores total, 2.25 GHz (base), 3.4 GHz (max boost)
Системная память	1TB
Сеть	9x Mellanox ConnectX-6 VPI HDR InfiniBand/200GigE 10th Dual-port ConnectX-6 optional
Накопитель	OS: 2x 1.92TB M.2 NVME drives Internal Storage: 15TB (4x 3.84TB) U.2 NVME drives

	Энергопотребление и габариты
Энергопотребление	6.5 кВт
Вес	123 кг
Габариты	6U Высота: 264.0 мм Ширина: 482.3 мм Глубина: 897.1 мм
Рабочая температура	5°C ... 30°C
Охлаждение	Воздушное

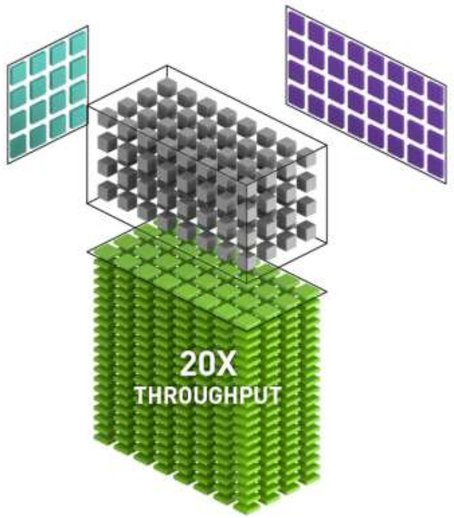
НОВЫЕ TF32 ТЕНЗОРНЫЕ ЯДРА A100

В 20 раз выше производительность для задач ИИ, неизменный код приложения

NVIDIA V100 FP32



NVIDIA A100 Tensor Core TF32 with Sparsity



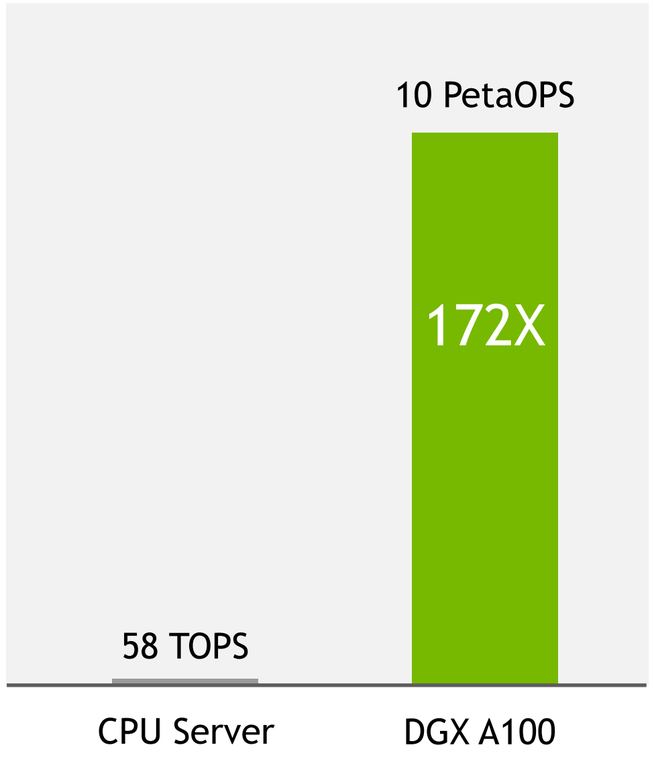
В 20 раз быстрее, чем Volta FP32 | Работает как FP32 для задач ИИ, с диапазоном FP32 и точностью FP16
Отсутствие необходимости менять код | Поддержка в контейнерах NGC с фреймворками PyTorch, TensorFlow и MXNet

https://blogs.nvidia.com/blog/2020/05/14/tensorfloat-32-precision-format/ https://blogs.nvidia.com/blog/2020/05/14/sparsity-ai-inference/

ПРОИЗВОДИТЕЛЬНОСТЬ DGX A100

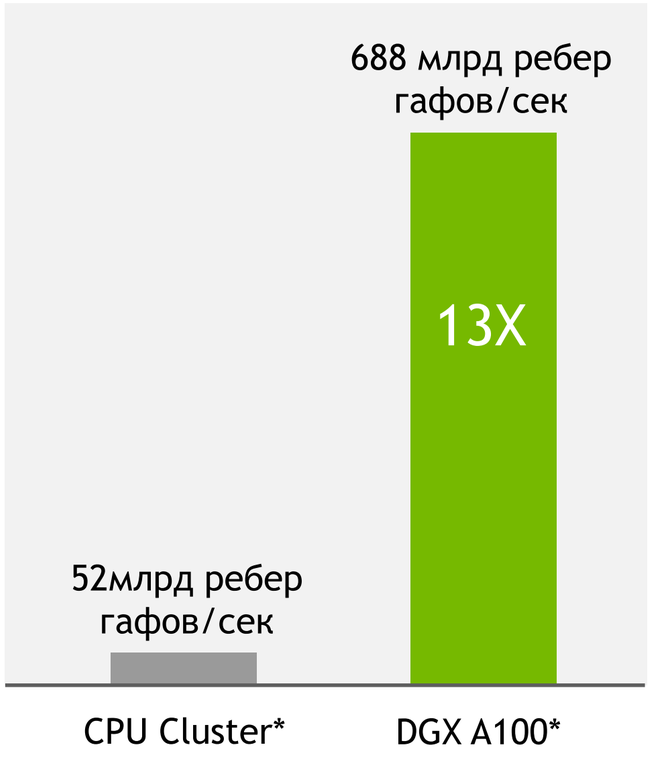
Обучение
NLP: BERT-Large

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 Server with 8x V100 using FP32 precision
DGX A100: DGX A100 with 8x A100 using TF32 precision



Инференс
Peak Compute

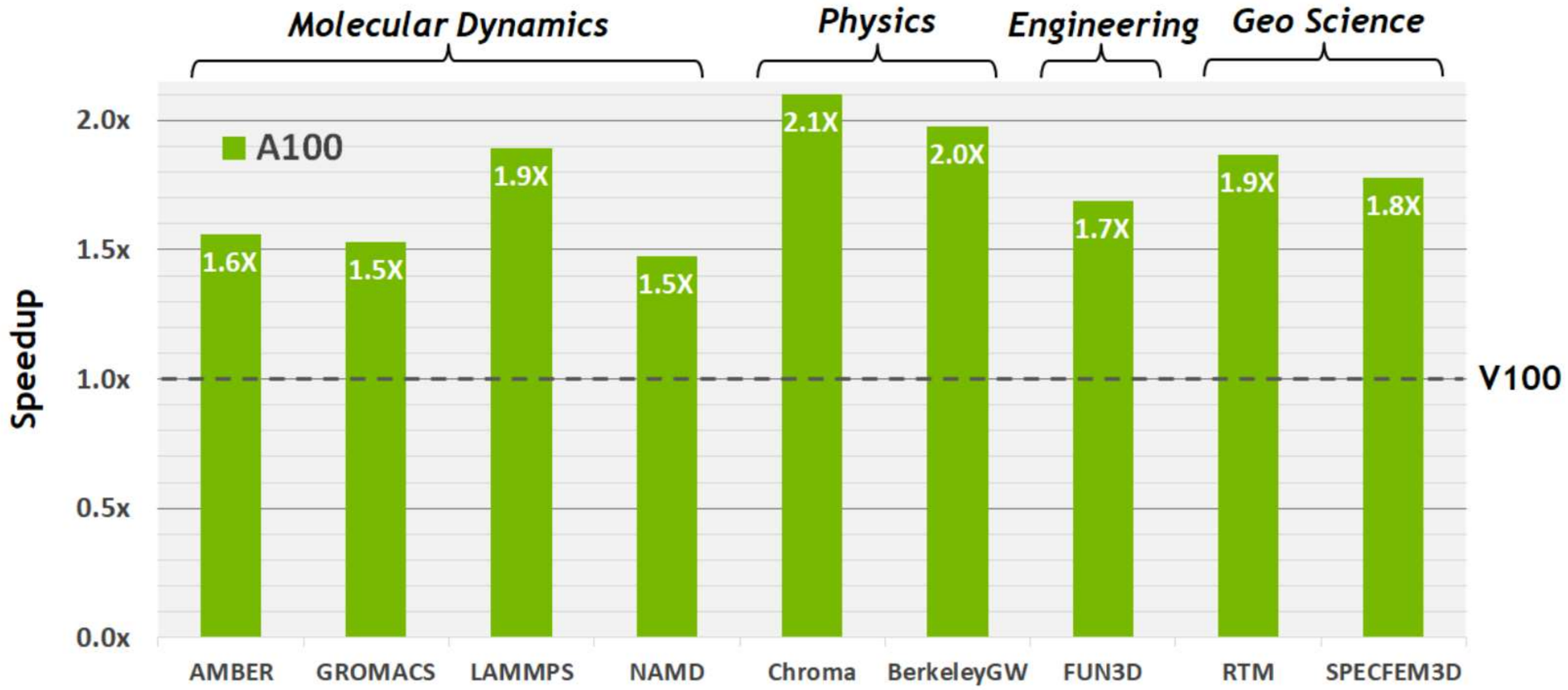
CPU Server: 2x Intel Platinum 8280 using INT8
DGX A100: DGX A100 with 8x A100 using INT8 with Structural Sparsity



Аналитика
PageRank

3000x CPU Servers vs. 4x DGX A100
Published Common Crawl Data Set: 128B Edges, 2.6TB Graph

УСКОРЕНИЕ HPC-ПРИЛОЖЕНИЙ



All results are measured
Except BerkeleyGW, V100 used is single V100 SXM2. A100 used is single A100 SXM4
More apps detail: AMBER based on PME-Cellulose, GROMACS with STMV (h-bond), LAMMPS with Atomic Fluid LJ-2.5, NAMD with v3.0a1 STMV_NVE
Chroma with szscl21_24_128, FUN3D with dpw, RTM with Isotropic Radius 4 1024^3, SPECFEM3D with Cartesian four material model
BerkeleyGW based on Chi Sum and uses 8xV100 in DGX-1, vs 8xA100 in DGX A100

Figure 3. A100 GPU HPC application speedups compared to NVIDIA Tesla V100.
https://devblogs.nvidia.com/nvidia-ampere-architecture-in-depth/

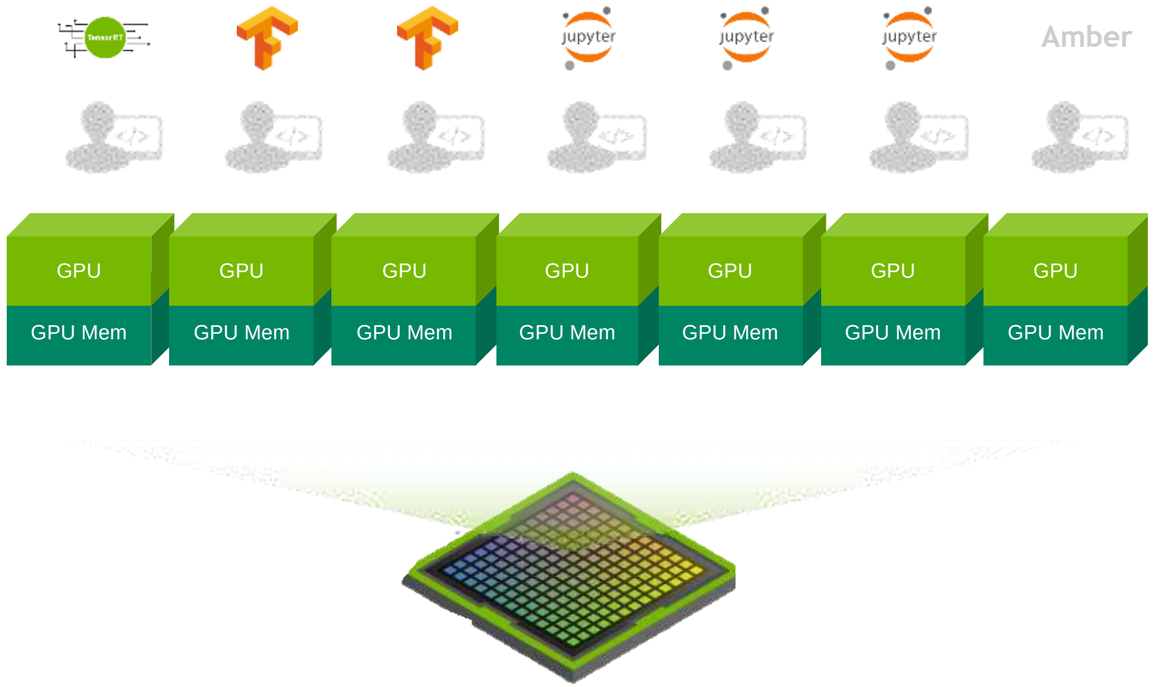
ПОКАЗАТЕЛИ ПРОИЗВОДИТЕЛЬНОСТИ A100

Peak FP64[1]	9.7 TFLOPS
Peak FP64 Tensor Core[1]	19.5 TFLOPS
Peak FP32[1]	19.5 TFLOPS
Peak FP16[1]	78 TFLOPS
Peak BF16[1]	39 TFLOPS
Peak TF32 Tensor Core[1]	156 TFLOPS | 312 TFLOPS[2]
Peak FP16 Tensor Core[1]	312 TFLOPS | 624 TFLOPS[2]
Peak BF16 Tensor Core[1]	312 TFLOPS | 624 TFLOPS[2]
Peak INT8 Tensor Core[1]	624 TOPS | 1,248 TOPS[2]
Peak INT4 Tensor Core[1]	1,248 TOPS | 2,496 TOPS[2]

Table 1. A100 Tensor Core GPU performance specs.
1) Peak rates are based on the GPU boost clock.
2) Effective TFLOPS / TOPS using the new Sparsity feature.
https://devblogs.nvidia.com/nvidia-ampere-architecture-in-depth/

САМАЯ ГИБКАЯ ПЛАТФОРМА ДЛЯ ИИ С MULTI-INSTANCE GPU (MIG)

Повышение утилизации GPU, гарантированный доступ к ресурсам большему числу пользователей



- До 7 GPU сущностей на одном A100
- Одновременное гарантированное исполнение задач Все MIG сущности работают параллельно с предсказуемой производительностью и задержкой
- Гибкость запуска любого типа задач на MIG сущностях
- Выделение оптимального кванта GPU ресурсов Различный размер MIG сущностей в зависимости от задач

САМЫЙ МОЩНЫЙ ИНСТРУМЕНТ ДЛЯ DS-КОМАНД

DGX A100 с MIG для необходимой производительности всем пользователям

Один DGX A100 предлагает:

- 5 petaFLOPS для обучения ИИ, или
- 10 petaOPS для инференса
- MIG позволяет команде из 25 разработчиков работать на одном DGX A100

Каждый разработчик получает:

- Более 180 teraFLOPS для обучения
  = 2 выделенные сущности с V100 в публичном облаке

или

- Более 357 teraOPS для инференса
  = 6 выделенных двухсокетных 28-ядерных CPU серверов

NVIDIA DGX SUPERPOD НА БАЗЕ DGX A100

Непревзойденная масштабируемость ЦОД и внедрение менее чем за 3 недели

Передовая ИИ инфраструктура

- Эталон производительности и масштабируемости ИИ на базе DGX A100
- Вся сила ИИ экспертизы от NVIDIA
- Для ранее нерешаемых задач
- Конфигурация блоками по 20 узлов

NVIDIA DGX SuperPOD уже в SATURNV

- 1,120 A100 GPU
- 140 DGX A100 систем
- 170 Mellanox 200G HDR свичей
- 4 PB высокопроизводительная СХД
- 700 PFLOPS производительности для ранее невиданных задач

https://www.nvidia.com/en-us/data-center/resources/nvidia-dgx-superpod-reference-architecture/

DGX A100 – САМОЕ УДОБНОЕ И МОЩНОЕ РЕШЕНИЕ ДЛЯ ИИ

Полный набор прикладного ПО и оптимизированные DL-фреймворки

Доступ к экспертизе NVIDIA в ИИ

Высочайшая производительность в ИИ и в HPC

Потрясающая масштабируемость ЦОД

ОСНОВНЫЕ МЫСЛИ

- фреймворки NVIDIA позволяют быстро получить результат с помощью ИИ
- вычисления на GPU стали еще проще с NVIDIA HPC SDK
- NVIDIA продолжает оставаться лидером в области решений для ИИ
- NVIDIA DGX A100 – самый удобный и мощный комплекс для ИИ

NVIDIA GTC

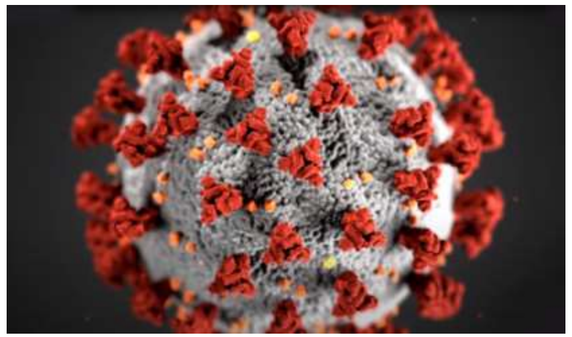
ИННОВАЦИИ НЕ ЖДУТ

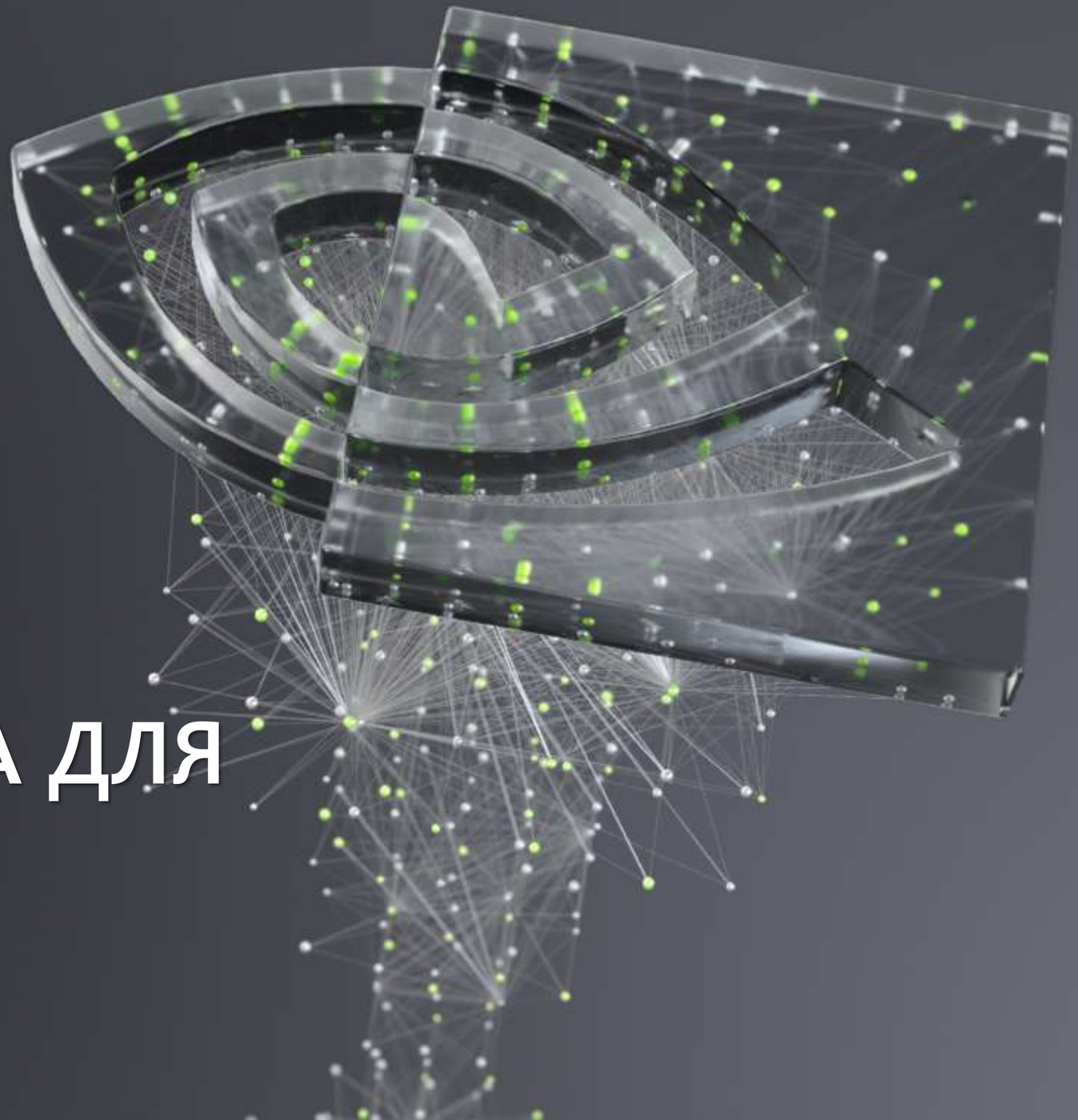
УЗНАЙТЕ ВСЕ НОВОСТИ НА NVIDIA GTC

На GTC вы узнаете о самых инновационных разработках в области AI, HPC и профессиональной графики, сможете пообщаться с экспертами и пройти обучение от NVIDIA Deep Learning Institute(DLI). Конференция пройдет в течение пяти дней с учетом семи часовых поясов.

Присоединяйтесь к трансляциям в реальном времени или изучите каталог сессий, чтобы ознакомиться с материалами в удобное вам время.

Онлайн 5 – 9 октября, 2020
Участие для ВУЗов и НИИ – бесплатное.
Регистрация www.nvidia.com/GTC





ПЛАТФОРМА NVIDIA ДЛЯ HPC И AI

Антон Джораев | adzhoraev@nvidia.com