CHALLENGES: ACCELERATING BIG AND SMALL

3000X Higher Compute Required to Train Largest Models Since Volta

Every AI Powered Interaction Needs Varying Amount of Compute

Source: OpenAI and NVIDIA Analysis
MODERN CLOUD DATA CENTER
Diverse Applications | Scale-Up & Scale-Out Workloads | Insatiable Demand
REIMAGINING THE GPU
Three Breakthroughs to Fuel the Next Era of Modern Accelerated Data Centers

20X
A GIANT PERFORMANCE LEAP

UNIFIED AI TRAINING AND INference ACCELERATION

1-50
SCALABILITY FOR THE ELASTIC DATACENTER
ANNOUNCING NVIDIA A100
GREATEST GENERATIONAL LEAP – 20X VOLTA

<table>
<thead>
<tr>
<th></th>
<th>Peak</th>
<th>Vs Volta</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>FP32 TRAINING</strong></td>
<td>312 TFLOPS</td>
<td>20X</td>
</tr>
<tr>
<td><strong>INT8 INFERENCE</strong></td>
<td>1,248 TOPS</td>
<td>20X</td>
</tr>
<tr>
<td><strong>FP64 HPC</strong></td>
<td>19.5 TFLOPS</td>
<td>2.5X</td>
</tr>
<tr>
<td><strong>MULTI INSTANCE GPU</strong></td>
<td>7X GPUs</td>
<td></td>
</tr>
</tbody>
</table>
ANNOUNCING NVIDIA A100
GREATEST GENERATIONAL LEAP – 20X VOLTA

- 54 BILLION XTORS
- 3RD GEN TENSOR CORES
- SPARSITY ACCELERATION
- MULTI INSTANCE GPU
- 3RD GEN NVLINK & NVSWITCH

54B xtors | 826mm² | TSMC 7N | 40GB Samsung HBM2 | 3rd gen Tensor Core GPU | 600 GB/s NVLink
### NEW TF32 TENSOR CORES

- **Range of FP32 and Precision of FP16** | **Input in FP32 and Accumulation in FP32** | **No Code Change Speed-up for Training**

<table>
<thead>
<tr>
<th>Format</th>
<th>FP32</th>
<th>Tensor Float32</th>
<th>FP16</th>
<th>BFloat16</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bits</strong></td>
<td>8</td>
<td>9</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td><strong>Precision</strong></td>
<td>23</td>
<td>10</td>
<td>10</td>
<td>7</td>
</tr>
</tbody>
</table>
NEW TENSOR CORE ACCELERATION FOR SPARSITY
Optimized For Sparse AI Tensor Ops | 2X Faster Execution | Supported on TF32, FP16, BFLOAT16, INT8 and INT4

Dense Matrix  →  Sparse Matrix  →  A100 Sparsity Optimized Tensor Core

2X Effective
NEW MULTI-INSTANCE GPU FOR ELASTIC GPU COMPUTING

7x Higher Throughput of V100 with Simultaneous Instances per GPU
UNIFIED AI TRAINING AND INFRINGEMENT ACCELERATION

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 8xV100 using FP32 precision | A100: DGX A100 Server with 8xA100 using TF32 precision

BERT Large Inference | T4, V100: TRT 7.1, Precision = FP16, Batch Size = 256 | A100 MIG: Pre-production TRT, Batch Size = 94, Precision = INT8 with Sparsity
ANNOUNCING NVIDIA DGX A100 3RD GENERATION INTEGRATED AI SYSTEM

5 PetaFLOPS of Performance in a Single Node

Unified System for End-to-End Data Science and AI
Fully Accelerated Stacks — Spark 3.0, RAPIDS, TensorFlow, PyTorch, Triton
Elastic Scale-Up or Scale-Out Computing
High Scalability with Mellanox Networking

INT8  10 PetaOPS Peak
FP16  5 PFLOPS Peak
TF32  2.5 PFLOPS Peak
FP64  156 TFLOPS Peak
ANNOUNCING NVIDIA A100 LIGHTHOUSE CUSTOMERS
Elastic Data Center Accelerator Choice of Industry Leaders
TODAY’S AI DATA CENTER

50 DGX-1 Systems for AI training
600 CPU Systems for AI Inference
$11M
25 Racks
630 kW

$11M 630 kW
DGX A100 AI

5 DGX A100 Systems for AI Training and Inference

$1M

1 Rack

28 kW

$1M  28 kW

1/10th COST

1/20th POWER
ANNOUNCING
NVIDIA DGX A100 SUPERPOD

140 DGX A100 Systems (1,120 A100)
170 Mellanox Quantum 200G InfiniBand Switches
280 TB/s Network Fabric - 15km of Optical Cable
4 PB of All-Flash Networked Storage
700 PFLOPS of AI Performance
Built in under 3 Weeks
NVIDIA EXPANDS SATURNV

Before Expansion
1,800 DGX Systems
1.8 ExaFLOPS

Adding 4 DGX SuperPODs
560 DGX A100 = 2.8 ExaFLOPS

4.6 ExaFLOPS Total Capacity
SMART EVERYTHING REVOLUTION

ALWAYS-ON | INSTANT SENSE-INFER-ACT | DISTANT | TRILLIONS
ANNOUNCING
NVIDIA EGX A100 WITH MELLANOX CX6 DX

NVIDIA Mellanox ConnectX-6 DX
- Dual 100 Gb/s Ethernet or InfiniBand
- Line-speed TLS/IPSec Crypto Engine
- Time Triggered Transmission Tech for Telco (5T for 5G)
- ASAP^2 SR-IOV and VirtualIO Offload

NVIDIA Ampere GPU
- 3rd generation Tensor Core
- New Security Engine for Confidential AI
- Secure, Authenticated Boot
GTC 2020 ANNOUNCEMENTS

Data Center-Scale Computing

Omniverse RTX Server

NVIDIA AI

DGX A100
Powered by A100

Merlin Recommender System

JARVIS Conversational AI

EGX A100
ISAAC, and BMW

Mellanox Technologies

RTX Server

NVIDIA AIOmniverse

DGX A100
Powered by A100

EGX A100
ISAAC, and BMW
ANNOUNCING
NVIDIA MERLIN — DEEP RECOMMENDER APPLICATION FRAMEWORK

1TB Ads Dataset

CPU
ETL 2 HR TRAINING 1 ½ Days

GPUs
ETL 3 Min TRAINING 16 Min

NVIDIA Merlin

NVTabular → HugeCTR

RAPIDS → cuDNN

Magnum IO → Magnum IO

TensorRT

Data Lake 100’s PB

User Query

0(10) Recommendation

CANDIDATES GENERATION

0(Billions) ITEMS EMBEDDING

0(1000) ITEMS EMBEDDING

0(1000) CANDIDATES GENERATION
ANNOUNCING
NVIDIA JARVIS — MULTIMODAL CONVERSATIONAL AI SERVICES FRAMEWORK

PRE-TRAINED MODEL
NVIDIA GPU CLOUD
NVIDIA AI TOOLKIT

RE-TRAIN
Transfer Learning
NuMo
Service Maker

NVIDIA JARVIS
Dialog Manager
Multi-Speaker transcription
Chatbot
Gesture Recognition
Look-to-Talk

TRITON INFERENCE SERVER
Speech
Language Model
Encoder
Acoustic Model
Feature Extraction
NLU & Recommenders
Speech Synthesis
Voice Encoder

JESSICA: What will you have ready for Wednesday?
DOUGLAS: I expect to have early designs of the packaging.
JESSICA: Great.

Join Early Access Program
developer.nvidia.com/nvidia-jarvis
ANNOUNCING
NVIDIA ACCELERATES SPARK 3.0