TURING REINVENTS GRAPHICS WITH RAY TRACING & AI

- Breakthrough Architecture
- 9x Peak FLOPs
- 114 TFLOPS Tensor Core
- Up to 10 Giga Rays

Diagram showing comparison of Pascal TITAN Xp, Turing 2080 Ti performance in FP32 and INT32 with Tensor Core (TC).
ANNOUNCING
JUSTICE — FIRST CHINA RTX GAME
TURING TENSOR CORE ENABLES DLSS

[Image: Comparison of DLSS on and off effects in a video game environment, highlighting improved image quality with DLSS enabled.]
3.5X PERFORMANCE WITH RAY TRACING + DLSS
ANNOUNCING
JX3 — FIRST CHINA DLSS-ACCELERATED GAME
Coming in 2019 Q1
1.8X PERFORMANCE WITH DLSS
NVIDIA ACCELERATES WORLD’S TOP500

127 NVIDIA GPU Systems
#1 World, US — ORNL Summit | 144 PF
#1 Europe — CSCS Piz Daint | 21 PF
#1 Japan — AIST ABCI | 20 PF
22 of Top 25 Energy-efficient Supercomputers

41 New NVIDIA Systems
+48% Y-Y

Fastest Computer in the World
Summit at ORNL
27,648 NVIDIA V100 | 144 PF | 3.3 EF Tensor Core

Timeline:
- 2010
- 2012
- 2014
- 2016
- 2018

Graph: Volume of NVIDIA Systems over time
AI IS AUTOMATING THE WORLD

$31 BILLION
sales in China

$31 BILLION
sales in China

8 BILLION
doctor visits in China

2 TRILLION
miles per year driven in China

$5 TRILLION
goods made per year in China

$8 BILLION
at China’s box office

INTERNET / E-TAIL

HEALTHCARE

TRANSPORTATION

MANUFACTURING

ENTERTAINMENT
AI TRANSFORMING COMPUTING

HPC

PAST
Scientific Computing
Few Users — Big Jobs
10's of Thousands of “Strong” Nodes

PRESENT
+ ML, DL, Big Data Analytics
+ AI Developers, Data Scientists
+ CSP & Industry

HYPERSCALE

PAST
Search & Recommenders
Many Users — Small Jobs
Millions of “Weak” Nodes

PRESENT
+ ML, DL, Big Data Analytics
+ Cloud Computing
+ HPC & Industry
AI TRANSFORMING COMPUTING

HPC

NVIDIA HGX-2
AI & Data Analytics for HPC

HYPERSCALE

NVIDIA T4
AI & Data Analytics for Hyperscale
ANNOUNCING NVIDIA HGX-2 BROADLY ADOPTED BY TECH LEADERS
WIDESPREAD ADOPTION OF NVIDIA INFEERENCE PLATFORM
NEW NVIDIA T4 CLOUD GPU ACCELERATES HYPERSONE

NVIDIA T4 Multi-precision Tensor Core

Support for Tensor Core

TensorRT 5.0

DNN Models

TensorRT Inference Server

RESNET-50 INFERENCE (I/S)

P4 P4 T4

5.5 22 65 260

Maxwell FP32 Kepler FP32 Pascal FP32 Volta FP16 Turing INT8 Turing INT4

Turing INT8

NV DL SDK

NV Docker

Kubernetes

P4 T4

65 130 260

FLOAT INT8 INT4

FLOAT INT8 INT4
Images Per Sec:
Demand: 17000
Delivered: 16990
ANNOUNCING NVIDIA T4 CLOUD GPU CUSTOMERS

Every Major Computer Maker
50+ Server Designs

Cloud Computing Leaders
HPC AND HYPERSCALE — ONE ARCHITECTURE

NVIDIA HGX-2

NVIDIA T4
NEW — NVIDIA ACCELERATES MACHINE LEARNING

RAPIDS Open-Source Library
CUDA-Accelerated Data Science Workflow
cuDF is “Pandas-like”
cuML is “ScikitLearn-like”
BROAD ECOSYSTEM ADOPTION OF NVIDIA ACCELERATED DATA SCIENCE
GPU-ACCELERATED DATA SCIENCE FROM HOURS TO MINUTES

Jupyter Notebook on NGC

ETL

- 20 CPU Nodes
- 50 CPU Nodes
- 100 CPU Nodes

DGX-2

- 2 Hours
- 1 Hour

ML

- 20 CPU Nodes
- 50 CPU Nodes
- 100 CPU Nodes

DGX-2
TRADITIONAL DATA SCIENCE CLUSTER

300 Servers
$3M
180 kW
GPU-ACCELERATED MACHINE LEARNING
DGX-2 AND RAPIDS FOR PREDICTIVE ANALYTICS

1 DGX-2
10 kW

1/8 the Cost  1/15 the Space  1/18 the Power
RAPIDS SUCCESS STORIES ACROSS INDUSTRY

- Wireless Network Optimization
- Epidemic Disease Prediction
- Cancer Immunotherapy
CUDA ACCELERATION STACKS ON NVIDIA GPU CLOUD
CLOUD REGISTRY OF ACCELERATION CONTAINERS

NEW Acceleration Containers
NEW Multi-node Accelerated Stacks
NEW Singularity Container Support
Runs on “NGC-Ready” Workstation, Cluster, Cloud
ANNOUNCING
NGC-READY
SYSTEM PARTNERS

NGC-READY SYSTEM PARTNERS

- SCIENCE
- DATA ANALYTICS
- DEEP LEARNING
- MACHINE LEARNING
- HYPERSCALCE INFERENCE
- RENDERING & VIZ

- HUAWEI
- inspur
- Lenovo
- Sugon
THE NEW HPC IS SCIENTIFIC COMPUTING, DATA SCIENCE, AND AI
THE NEW HPC IS SCIENTIFIC COMPUTING, DATA SCIENCE, AND AI
THE NEW HPC IS
SCIENTIFIC COMPUTING, DATA SCIENCE, AND AI

$37B

ENTERPRISE

HYPERSCALE

SCIENTIFIC COMPUTING

INTERNET SERVICES
E-COMMERCE
CLOUD COMPUTING
MEDIA

RETAIL
HEALTHCARE
FINANCIAL SERVICES
TELECOM

SCIENTIFIC RESEARCH
ENERGY
HIGHER ED

NVIDIA V100 HGX-2
+ TENSOR CORE GPU
THE NEW HPC IS SCIENTIFIC COMPUTING, DATA SCIENCE, AND AI

$37B

ENTERPRISE

SCIENTIFIC COMPUTING

HYPERSONE

INTERNET SERVICES ELECTRONIC COMMERCE CLOUD COMPUTING MEDIA

- SCIENTIFIC RESEARCH ENERGY HIGHER ED

- RETAIL HEALTHCARE FINANCIAL SERVICES TELECOM

NVIDIA V100 HGX-2

+ TENSOR CORE GPU

+ NVIDIA T4

- TELECOM

- HEALTHCARE

- FINANCIAL SERVICES

- ELECTRONIC COMMERCE

- CLOUD COMPUTING

- MEDIA
THE NEW HPC IS
SCIENTIFIC COMPUTING, DATA SCIENCE, AND AI

$37B

科学计算
数据科学
人工智能

企业
超大规模
科学计算
研究
能源
高等教育
互联网服务
电子商务
零售
医疗
金融
电信

RAPIDS
NVIDIA T4
NVIDIA V100 HGX-2
+ TENSOR CORE GPU
+ NVIDIA T4
+ RAPIDS
AI IS AUTOMATING THE WORLD
XAVIER
WORLD’S FIRST AI COMPUTING PROCESSOR
NEW — XAVIER AUTO-SOC AND FULL PRODUCTION

Most Complex SOC Ever Made
9 Billion Transistors, 350mm², 12nFFN
~8,000 Engineering Years

Volta Tensor Core GPU
512 CUDA Tensor Cores
2,8 CUDA TFLOPS (FP16)
22.6 Tensor Core DL TOPS

ISP
2.4 GPIX/s
Native Full-range HDR
Tile-based Processing

Vision Accelerator
1.7 TOPS
Stereo & Optical Flow Engine
2x 3.1 TOPS

Multimedia Engines
1.2 GPY/s Encode
1.4 GPY/s Decode
4 GPY/s Video Image Compositor

Carmel ARM64 CPU
8 Cores
10-wide Superscalar
21 SpecInt2K6

Industry Standard High-Speed IO
PCIe Gen4 Root and Endpoint
USB 3.1 Gen2 Host and Device
UFS 2.1 Embedded Storage

16 Lane CSI
109 Gbps CPHY 5.1
1Gb Ethernet

120b PCIe LPDDR4X
137 GB/s

DLA
5.7 TFLOPS FP16
11.4 TOPS INT8

109 Gbps CPHY 5.1
1Gb Ethernet
XAVIER
DESIGNED FOR AI AND ROBOTICS

SENSOR

PERCEPTION

PLANNING

ACTION

CUDA
VDEC
ISP
PVA

TENSOR CORE
CUDA

TENSOR CORE
CUDA

TENSOR CORE
CPU

TENSOR CORE
CUDA

GPU
CPU
VENC
NVIDIA AGX FAMILY
SCALABLE AI COMPUTERS

High-speed SerDes — 109 Gbps + 320 Gbps I/O
Up to 320 TOPS Tensor Ops
Up to 25 TFLOPS FP32
Up to 16 GIGA Rays
Starting from 15W
NVIDIA AI COMPUTING PLATFORMS

- CLARA MEDICAL IMAGING
- METROPOLIS STREAMING ANALYTICS
- ISAAC ROBOTICS
- DRIVE AUTONOMOUS VEHICLES
NVIDIA CLARA

CLARA
- Kubernetes
- Docker
- Render Server
- TRT Inference Server

ACCELERATION LIBRARIES
- mGPU & vGPU
- OptiX RTX
- TensorRT
- OpenCV Computer Vision
- NVIDIA Imaging & Signal Processing
- cuBLAS

OS
- Linux
- CUDA
- OGL / ES
- Vulkan
- NVMedia

Computer Vision
- cuDNN

NVIDIA CLARA
### NVIDIA METROPOLIS

<table>
<thead>
<tr>
<th>DEEPMETRIC FRAMEWORK</th>
<th>ACCELERATION LIBRARIES</th>
<th>OS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Container</td>
<td>CUDA</td>
<td>LINUX</td>
</tr>
<tr>
<td>Sensors</td>
<td>CUDA</td>
<td>CUDA</td>
</tr>
<tr>
<td>TensorFlow</td>
<td>NPP Imaging</td>
<td>OGL / ES</td>
</tr>
<tr>
<td>Computer Vision</td>
<td>OpenCV Imaging</td>
<td>Vulkan</td>
</tr>
<tr>
<td>&amp; Signal Processing</td>
<td>cuDAS</td>
<td>NVMedia</td>
</tr>
</tbody>
</table>

---

**Diagram:**

- **Deepstream Framework:**
  - CUDA
  - DEEPMETRIC FRAMEWORK
  - Container
  - Sensors
  - TensorFlow
  - Computer Vision
  - & Signal Processing
  - cuDAS

- **Acceleration Libraries:**
  - CUDA
  - TRT
  - NPP Imaging
  - OpenCV Imaging
  - cuDAS

- **OS:**
  - LINUX
  - CUDA
  - OGL / ES
  - Vulkan
  - NVMedia

---

**Diagram Elements:**

- Video Sink (VID SINK)
-_deck
  - Video Source (ENC)
  - Video Source (RTSP)
  - Video Source (CUDA)
  - Video Source (CUDA)
  - Video Source (CUDA)
ANNOUNCING E-COMMERCE LEADERS SELECT JETSON AGX XAVIER
NVIDIA DRIVE – OPEN AUTONOMOUS VEHICLE PLATFORM

DRIVE AV
- Perception
- Localization & Mapping
- Planning

DRIVWORKS ACCELERATION LIBRARIES
- Sensor API
- TensorRT
- OpenCV
- NPP Imaging
- cuBLAS

DRIVE OS
- QNX
- CUDA
- OGL / ES
- Vulkan
- NVMedia

Signal Processing
NVIDIA DRIVE AVAILABLE NOW

DRIVE AGX XAVIER DEVELOPER KIT
Available Now

DRIVE IX
Available Now

DRIVE AV
Available Now
ANNOUNCING VOLVO CARS SELECTS NVIDIA DRIVE AGX

DRIVE AGX Xavier to Pilot Next-Generation Production Cars
ANNOUNCING NEV COMPANIES SELECT NVIDIA DRIVE AGX

XPENG MOTORS
SINGULATO MOTORS
SF MOTORS
ANNOUNCING MOBILITY SERVICES SELECT NVIDIA DRIVE AGX

WERIDE
MOMENTA
TUSIMPLE
AUTOX
ROADSTAR.AI
ANNOUNCING
NVIDIA SELECTED FOR
AUTONOMOUS TRUCK
HAILING SERVICE

Production in 2021
NVIDIA DRIVE GROWING TRANSPORTATION ECOSYSTEM

CARS

MOBILITY SERVICES

TRUCKS

TIER ONES

MAPPING
NEW NVIDIA PLATFORMS

TURING
Reinventing Graphics and AI Computing

NVIDIA ACCELERATED COMPUTING
HGX-2 Accelerates HPC | T4 Accelerates Hyperscale | NGC and NGC-Ready Systems

NVIDIA AGX
XAVIER In Full Production