TWO FORCES DRIVING THE FUTURE OF COMPUTING

40 Years of CPU Trend Data

TWO FORCES DRIVING THE FUTURE OF COMPUTING

The Big Bang of Deep Learning

40 Years of CPU Trend Data

Original data up to year 2010 collected and analyzed by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten New plot and data collected for 2010-2015 by K. Rupp
RISE OF NVIDIA GPU COMPUTING

40 Years of CPU Trend Data

GPU-Computing perf
1.5X per year

1.1X per year

Single-threaded perf
1.5X per year

NVIDIA Volta GPU | 21B xtors | 120 TFLOPS
RISE OF NVIDIA GPU COMPUTING

- Global GTC Attendees: 22,000 in 2012, 645,000 in 2017 (10X in 5 Years)
- GPU Developers: 22,000 in 2012, 645,000 in 2017 (15X in 5 Years)
- CUDA Downloads: 1.8M in 2012, 645,000 in 2017 (5X in 5 Years)
NVIDIA GPU-ACCELERATED INDUSTRIES

HPC
- Weather
- Molecular Dynamics
- Materials Science

INTERNET SERVICES
- Photo Album
- Recommendations
- Social Media

TRANSPORTATION
- Ride Sharing
- Self-Driving Cars

MEDICAL IMAGING
- Radiology
- Volumetric Rendering

LOGISTICS
- AV Truck
- Route Management
- Warehouse Automation

DESIGN & SIMULATION
- Weather
- Molecular Dynamics
- Materials Science

PHOTO ALBUM
- Amazon
- Twitter

RADIOLOGY
- 3D CT

WEATHER
- Materials Science

MATERIALS SCIENCE
- Design & Simulation

SOCIALLY MEDIA
- Twitter

SOCIAL MEDIA
- Twitter
NVIDIA GPU ACCELERATES 2017 NOBEL PRIZES IN CHEMISTRY AND PHYSICS

Cryogenic Electron Microscopy
Jacques Dubochet, Joachim Frank, Richard Henderson

Detection of Gravitational Waves
Rainer Weiss, Barry Barish, Kip Thorne
ANNOUNCING NVIDIA HOLODECK
THE DESIGN LAB OF THE FUTURE

Photorealistic Models
Physically Simulated Interaction
Virtual Team Collaboration
GPU-accelerated AI
CREATE AND COLLABORATE IN NVIDIA HOLODECK

CATIA / Siemens NX, Creo / Alias

Maya / 3dsMAX

CREATE AND COLLABORATE IN NVIDIA HOLODECK
ANNouncing
NVIDIA HOLODECK
THE DESIGN LAB OF THE FUTURE
Photorealistic Models
Physically Simulated Interaction
Virtual Team Collaboration
GPU-accelerated AI
Early Access NOW
nvidia.com/holodeck
THE ERA OF AI

- Deep Learning Papers Published 13X in 3 Years
- AI Startup Funding 12X in 5 Years
- NIPS Registration
SOLVING THE UNSOLVABLE

NVIDIA
Interactive Ray Tracing
SOLVING THE UNSOLVABLE

NVIDIA / Remedy
Audio-driven Facial Animation

NVIDIA Interactive Ray Tracing
SOLVING THE UNSOLVABLE

NVIDIA / Remedy Audio-driven Facial Animation

WRNCH Pose Estimation
SOLVING THE UNSOLVABLE

NVIDIA Interactive Ray Tracing

NVIDIA / Remedy Audio-driven Facial Animation

WRINCH Pose Estimation

University of Edinburgh Character Animation
THE WORLD’S AI PLATFORM

Every Framework
- Caffe2
- Chainer
- MXNet
- PaddlePaddle
- PyTorch
- TensorFlow
- Theano

NVIDIA Inception: 1,900 DL Startups

Every Cloud and Data Center
- Alibaba Cloud
- Amazon Web Services
- Microsoft Azure
- NVIDIA
- VMware
- VMware
- Xilinx

NVIDIA AI PLATFORM
AI INFEERENCE IS THE NEXT GREAT CHALLENGE.

Training: Caffe2, Chainer, mxnet, PaddlePaddle, PyTorch, TensorFlow, Theano

DNN Model

Inferencing: Car, Drone, Robot, Security Camera, Sensor
EXPLOSION OF INTELLIGENT MACHINES

20M Inference Servers

100s of Millions of Autonomous Machines

Trillions of IoT Devices
EXPLOSION OF NETWORK DESIGN

Convolution Networks
- PReLU
- ReLU
- BatchNorm
- Concat
- Dropout
- Pooling

Recurrent Networks
- LSTM
- GRU
- Highway
- BiDirectional

Generative Adversarial Networks
- 3D-GAN
- Rank GAN
- Conditional GAN
- Coupled GAN
- Speech Enhancement GAN
- Latent space GAN

Reinforcement Learning
- A3C
- Dueling DQN
- DQN
- Environments
- Rewards
- States
NEW NVIDIA TENSORRT 3
PROGRAMMABLE INFERENCER ACCELERATOR

Compile and Optimize Neural Networks
Support for Every Framework
Optimize for Each Target Platform
NEW NVIDIA TENSORRT 3
PROGRAMMABLE INFERENCE ACCELERATOR

Weight & Activation Precision Calibration
Layer & Tensor Fusion
Kernel Auto-Tuning
Multi-Stream Execution
NEW NVIDIA TENSORRT 3
PROGRAMMABLE INFERENCE ACCELERATOR

40x Speed-up on ResNet-50
140x Speed-up on OpenNMT

Images/Sec (ResNet-50)

Sentences/Sec (OpenNMT)
NEW NVIDIA TENSORRT 3
PROGRAMMABLE INFERENCCE ACCELERATOR

40x Speed-up on ResNet-50
140x Speed-up on OpenNMT
Half the Latency
NVIDIA TENSORRT
10X BETTER DATA CENTER TCO

160 CPU servers
45,000 images / second
65 kWatts
NVIDIA TENSORRT
10X BETTER DATA CENTER TCO

1 NVIDIA HGX with 8 Tesla V100 GPUs
45,000 images / second
3 KWatts

1/6 the Cost | 1/20 the Power | 4 Racks in a Box
INFERENCE ON IMAGES
INFERENCE ON SPEECH
Listening...
VINCENT AI
THE AUTONOMOUS VEHICLE REVOLUTION
NVIDIA DRIVE
AV COMPUTING PLATFORM
Sensor Fusion: RADAR, LIDAR, Camera
Deep Learning, CV, Parallel Computing
Diversity of Algorithms
ASIL-D Functional Safety
Fully Integrated into NVIDIA BB8

DRIVE PX — AI CAR COMPUTER
“RIDE-HAILING INDUSTRY EXPECTED TO GROW EIGHTFOLD TO $285B BY 2030”

-Goldman Sachs
ROBOTAXI DEMANDS EXTREME COMPUTING

+ 10X camera resolution
+ Surround LIDAR point-cloud processing
+ Camera & LIDAR localization to HD map
+ Tracking all surrounding objects
+ New map generation
+ Sophisticated path planning & control
+ Algorithm diversity
+ Sensor & computing fail-operate
+ Excess computing capacity
STATE-OF-THE-ART DRIVERLESS VEHICLES
ANNOUNCING “PEGASUS”
ROBOTAXI DRIVE PX

320 TOPS CUDA TensorCore
16x GMSL | 4x 10G | 8x 1G | 16x 100M
Auto-grade | ASIL D
500W
Late Q1 Early Access Partners

Supercomputing Data Center in Your Trunk
ANNOUNCING “PEGASUS”
ROBOTAXI DRIVE PX

320 TOPS CUDA TensorCore
16x GMSL | 4x 10G | 8x 1G | 16x 100M
Auto-grade | ASIL D
500W
Late Q1 Early Access Partners

Supercomputing Data Center in Your Trunk
STATE-OF-THE-ART DRIVERLESS VEHICLES
ANNOUNCING “PEGASUS”

ROBOTAXI DRIVE PX

320 TOPS CUDA TensorCore
16x GMSL | 4x 10G | 8x 1G | 16x 100M
Auto-grade | ASIL D
500W
Late Q1 Early Access Partners

Supercomputing Data Center in Your Trunk
ANNOUNCING
NVIDIA DRIVE IX SDK
INTELLIGENT EXPERIENCE TOOLKIT
Sense Inside & Outside the Vehicle
Deep Learning Powered
Early Access Q4

Your Car is an AI
NVIDIA DRIVE
AI CAR PLATFORM
From Training to Driving
From Processor to AI to AV Stack
From AV to IX
From 3D Simulation to Super-Real-Time Testing
Open Platform for Partners
SUPER REAL-TIME SIMULATION WITH DGX AND NEW TENSORRT 3

With 8 NVIDIA DGX
Re-Sim 300,000 miles in 5 hours

Virtually Drive Every Paved Road in U.S. in 2 Days
145 AV STARTUPS ON NVIDIA DRIVE
THE ERA OF AUTONOMOUS MACHINES
XAVIER
WORLD’S FIRST AUTONOMOUS MACHINE PROCESSOR
Deep Learning, CV, Parallel Computing
Rich High-Speed Sensor IOs
Extreme Energy Efficiency
30 TOPS at 30W
A NEW COMPUTING ERA

NVIDIA Holodeck
Design Lab of the Future

NVIDIA
The World’s AI Computing Platform

NVIDIA DRIVE
Open AI Platform for the Transportation Industry

NVIDIA DRIVE PX
Pegasus Robotaxi AI Computer

Project Isaac
AI Robot Simulator