BIG BANG OF MODERN AI

- Detection
- Auto Answer
- Sentiment Classification
- AlexNet
- Segmentation
- Translation
- Search
- Recommender
- Anomaly Detection

2012 2014 2016 2018

IMAGE
LANGUAGE
DECISION MAKING

DECISION MAKING
A DAY IN THE LIFE OF AI

EXTRACT

Telemetry Data
Usage Data
Inventory Data

TRANSFORM

PROCESS DATA
Increasing Data Sizes

LOAD

TRAIN AI
Rising Network Complexity

DEPLOY AI
Billions of Devices
DATA PROCESSING

- Python
- CUDA
- Apache Arrow
- CUDF, CUGRAPH
- RAPIDS
- CuDF
- CuML
- CuGraph
- CuDNN
- DASK
- PANDAS
- SGX-ML / XGBOOST
- Deep Learning Frameworks

25 MILLION CUSTOMERS IN 40+ STATES IN U.S.
Second Largest Cable Operator

DATA SCIENCE WORKSTATION
AI TRAINING

**Frameworks**
- CLOUD ML SERVICES
  - Amazon SageMaker
  - Google Cloud ML
  - Azure Machine Learning

**Deployment**
- Workstation CloudServer
- Neo

**AI Training**
- PROJECT MEGATRON
  - 8.3B Parameters, 8-way Model Parallel, 64-way Data Parallel, 24x Larger than BERT

**Number of Parameters by Network**
- ResNet-50: 24M
- Transformer: 340M
- OPT-1: 1.5B
- BERT-base: 340M
- GPT-1: 8M
- GPT-2: 153B
- GPT-2 8B

**CUDA-X AI**
- NLP - Generative Tasks
  - Chatbots
  - E-mail Auto-completion
  - Document Summarization

**DGX SUPERPOD**
- 1536 V100 GPUs, Mellanox Interconnect
- 9.4 Petaflops

**Number of Parameters by Network**
- BERTLARGE: 8.3B

**Number of Parameters by Network**
- Original: GPT-1TransformerResNet-50 GPT-2 GPT-2 8B

**NLP**
- Q&A
- Sentiment
- Translation

**IMAGE**
- Autonomous Vehicles
- Social Tagging
- Visual Search

**DGX SUPERPOD**
- 9.4 Petaflops

**CUDA-X AI**
- NLP - Generative Tasks
  - Chatbots
  - E-mail Auto-completion
  - Document Summarization

**DGX SUPERPOD**
- 1536 V100 GPUs, Mellanox Interconnect
- 9.4 Petaflops

**CUDA-X AI**
- NLP - Generative Tasks
  - Chatbots
  - E-mail Auto-completion
  - Document Summarization

**DGX SUPERPOD**
- 9.4 Petaflops

**CUDA-X AI**
- NLP - Generative Tasks
  - Chatbots
  - E-mail Auto-completion
  - Document Summarization

**DGX SUPERPOD**
- 9.4 Petaflops

**CUDA-X AI**
- NLP - Generative Tasks
  - Chatbots
  - E-mail Auto-completion
  - Document Summarization

**DGX SUPERPOD**
- 9.4 Petaflops

**CUDA-X AI**
- NLP - Generative Tasks
  - Chatbots
  - E-mail Auto-completion
  - Document Summarization

**DGX SUPERPOD**
- 9.4 Petaflops

**CUDA-X AI**
- NLP - Generative Tasks
  - Chatbots
  - E-mail Auto-completion
  - Document Summarization

**DGX SUPERPOD**
- 9.4 Petaflops

**CUDA-X AI**
- NLP - Generative Tasks
  - Chatbots
  - E-mail Auto-completion
  - Document Summarization

**DGX SUPERPOD**
- 9.4 Petaflops

**CUDA-X AI**
- NLP - Generative Tasks
  - Chatbots
  - E-mail Auto-completion
  - Document Summarization

**DGX SUPERPOD**
- 9.4 Petaflops

**CUDA-X AI**
- NLP - Generative Tasks
  - Chatbots
  - E-mail Auto-completion
  - Document Summarization

**DGX SUPERPOD**
- 9.4 Petaflops

**CUDA-X AI**
- NLP - Generative Tasks
  - Chatbots
  - E-mail Auto-completion
  - Document Summarization

**DGX SUPERPOD**
- 9.4 Petaflops

**CUDA-X AI**
- NLP - Generative Tasks
  - Chatbots
  - E-mail Auto-completion
  - Document Summarization

**DGX SUPERPOD**
- 9.4 Petaflops

**CUDA-X AI**
- NLP - Generative Tasks
  - Chatbots
  - E-mail Auto-completion
  - Document Summarization

**DGX SUPERPOD**
- 9.4 Petaflops

**CUDA-X AI**
- NLP - Generative Tasks
  - Chatbots
  - E-mail Auto-completion
  - Document Summarization

**DGX SUPERPOD**
- 9.4 Petaflops

**CUDA-X AI**
- NLP - Generative Tasks
  - Chatbots
  - E-mail Auto-completion
  - Document Summarization

**DGX SUPERPOD**
- 9.4 Petaflops

**CUDA-X AI**
- NLP - Generative Tasks
  - Chatbots
  - E-mail Auto-completion
  - Document Summarization

**DGX SUPERPOD**
- 9.4 Petaflops

**CUDA-X AI**
- NLP - Generative Tasks
  - Chatbots
  - E-mail Auto-completion
  - Document Summarization

**DGX SUPERPOD**
- 9.4 Petaflops

**CUDA-X AI**
- NLP - Generative Tasks
  - Chatbots
  - E-mail Auto-completion
  - Document Summarization

**DGX SUPERPOD**
- 9.4 Petaflops

**CUDA-X AI**
- NLP - Generative Tasks
  - Chatbots
  - E-mail Auto-completion
  - Document Summarization

**DGX SUPERPOD**
- 9.4 Petaflops

**CUDA-X AI**
- NLP - Generative Tasks
  - Chatbots
  - E-mail Auto-completion
  - Document Summarization

**DGX SUPERPOD**
- 9.4 Petaflops

**CUDA-X AI**
- NLP - Generative Tasks
  - Chatbots
  - E-mail Auto-completion
  - Document Summarization

**DGX SUPERPOD**
- 9.4 Petaflops

**CUDA-X AI**
- NLP - Generative Tasks
  - Chatbots
  - E-mail Auto-completion
  - Document Summarization

**DGX SUPERPOD**
- 9.4 Petaflops

**CUDA-X AI**
- NLP - Generative Tasks
  - Chatbots
  - E-mail Auto-completion
  - Document Summarization

**DGX SUPERPOD**
- 9.4 Petaflops

**CUDA-X AI**
- NLP - Generative Tasks
  - Chatbots
  - E-mail Auto-completion
  - Document Summarization

**DGX SUPERPOD**
- 9.4 Petaflops

**CUDA-X AI**
- NLP - Generative Tasks
  - Chatbots
  - E-mail Auto-completion
  - Document Summarization

**DGX SUPERPOD**
- 9.4 Petaflops

**CUDA-X AI**
- NLP - Generative Tasks
  - Chatbots
  - E-mail Auto-completion
  - Document Summarization

**DGX SUPERPOD**
- 9.4 Petaflops

**CUDA-X AI**
- NLP - Generative Tasks
  - Chatbots
  - E-mail Auto-completion
  - Document Summarization

**DGX SUPERPOD**
- 9.4 Petaflops

**CUDA-X AI**
- NLP - Generative Tasks
  - Chatbots
  - E-mail Auto-completion
  - Document Summarization

**DGX SUPERPOD**
- 9.4 Petaflops

**CUDA-X AI**
- NLP - Generative Tasks
  - Chatbots
  - E-mail Auto-completion
  - Document Summarization

**DGX SUPERPOD**
- 9.4 Petaflops

**CUDA-X AI**
- NLP - Generative Tasks
  - Chatbots
  - E-mail Auto-completion
  - Document Summarization

**DGX SUPERPOD**
- 9.4 Petaflops

**CUDA-X AI**
- NLP - Generative Tasks
  - Chatbots
  - E-mail Auto-completion
  - Document Summarization

**DGX SUPERPOD**
- 9.4 Petaflops

**CUDA-X AI**
- NLP - Generative Tasks
  - Chatbots
  - E-mail Auto-completion
  - Document Summarization

**DGX SUPERPOD**
- 9.4 Petaflops

**CUDA-X AI**
- NLP - Generative Tasks
  - Chatbots
  - E-mail Auto-completion
  - Document Summarization

**DGX SUPERPOD**
- 9.4 Petaflops

**CUDA-X AI**
- NLP - Generative Tasks
  - Chatbots
  - E-mail Auto-completion
  - Document Summarization

**DGX SUPERPOD**
- 9.4 Petaflops

**CUDA-X AI**
- NLP - Generative Tasks
  - Chatbots
  - E-mail Auto-completion
  - Document Summarization

**DGX SUPERPOD**
- 9.4 Petaflops

**CUDA-X AI**
- NLP - Generative Tasks
  - Chatbots
  - E-mail Auto-completion
  - Document Summarization

**DGX SUPERPOD**
- 9.4 Petaflops

**CUDA-X AI**
- NLP - Generative Tasks
  - Chatbots
  - E-mail Auto-completion
  - Document Summarization

**DGX SUPERPOD**
- 9.4 Petaflops
MODERN USE-CASES: REAL TIME AND MULTI-MODEL
Enabling AI Conversational with Performance and Versatility

What are different types of lighting for a living room?

Ambient, task and accent

20-30 Containers End-to-end | RNN, CNN, MLP in INT8, FP16, FP32 | Latency <300ms
“George Washington was the first president of the United States.”
HEALTHCARE IMAGING INDUSTRY CHALLENGES

10'S OF MODALITIES

8,000 IMAGES VIEWED PER DAY

3 PAPERS PER MINUTE

PubMed Papers per Month


0 25,000 50,000 75,000 100,000 125,000
CLARA SDK
Build, Manage and Deploy AI Applications for Imaging and Genomics

- AI-Assisted Annotation
  - 10x Less Training Data Needed
  - 13 World-Class Pre-Trained Models
  - Reference Training and Deployment Pipelines

https://developer.nvidia.com/clara
ACCELERATING AI IN IMAGING WITH TRANSFER LEARNING

ANNOTATED SEGMENTATION DATA

TRANSFER LEARNING PROCESS

INTEGRATION WITH KUBEFLOW
5G REVOLUTION

1000X MORE BANDWIDTH | ULTRA-RELIABLE | 1MS LATENCY | 1M DEVICES PER SQUARE KM
A NEW COMPUTING PLATFORM
IOT | AR/VR | CLOUD GAMING | ROBOTICS | TRILLIONS OF DEVICES | EDGE PROCESSING
ANNOUNCING NVIDIA EGX EDGE SUPERCOMPUTING PLATFORM

Powered by NVIDIA CUDA Tensor Core GPU
Secured Boot Root of Trust
Industrial-strength Cloud Native and AI Stack
New Aerial 5G SDK
ANNOUNCING
USPS ADOPTION OF NVIDIA AI
World’s Largest Delivery Service Invests in AI to Boost Sorting Efficiency

485M Mail Pieces per Day, 146B Annually
10x Faster, Higher Accuracy
200 Processing Facilities
END-TO-END WORKFLOW TO DEVELOP AN AUTONOMOUS VEHICLE

- Ingestion
- Curation
- Labeling
- Training
- Replay
- Simulation
AUTONOMOUS VEHICLE
AI-ENABLED ROBOTS

RETAIL

AGRICULTURE

DELIVERY
NVIDIA JETSON END-TO-END AUTONOMOUS ROBOTICS PLATFORM FOR RAPID DEPLOYMENT

<table>
<thead>
<tr>
<th>DEEPSTREAM SDK</th>
<th>ISAAC SDK</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SOFTWARE MODELS</strong></td>
<td></td>
</tr>
<tr>
<td>Depth Estimation</td>
<td>Object Detection</td>
</tr>
<tr>
<td>Pose Estimation</td>
<td>Gesture Recognition</td>
</tr>
<tr>
<td>Path Planning</td>
<td>Autonomous Navigation</td>
</tr>
</tbody>
</table>

**JETSON COMPUTER**

- **JETSON TX2 Series**
  - 7.5 – 15W*
  - 1.3 TFLOPS (FP16)

- **JETSON AGX XAVIER Series**
  - 10 – 30W
  - 5.5 - 11 TFLOPS (FP16)
  - 20 - 32 TOPS (INT8)

- **JETSON NANO**
  - 5 - 10W
  - 0.5 TFLOPS (FP16)

**AI at the Edge**

**Fully Autonomous Machines**

MULTIPLE DEVICES, SAME SOFTWARE
COLLABORATIVE ROBOTS

IFM
Warehouse inventory management
NATIONAL AI STRATEGIES TO MAINTAIN LEADERSHIP

AI STANDARDS
PUBLIC-PRIVATE PARTNERSHIPS
OPEN DATASETS
AI INFRASTRUCTURES
WORKFORCE TRAINING
NVIDIA DLI (DEEP LEARNING INSTITUTE) – GET YOUR WORKFORCE AI READY

ACCELERATED COMPUTING FUNDAMENTALS
DEEP LEARNING FUNDAMENTALS
DEEP LEARNING BY INDUSTRY

INDIVIDUAL
Online Learning

TEAMS
Instructor Led Workshops

UNIVERSITIES
Teaching Kits
PUTTING AI TO WORK

PROCESS DATA | TRAIN | DEPLOY

VERTICAL INDUSTRY FRAMEWORKS
ngc.nvidia.com

GETTING STARTED
Visit: nvidia.com/deploy-ai
FEDRAMP AND EC2 G4 INSTANCES
Amazon Web Services on Demand Computing
Scales With Your Needs
ANNOUNCING MODZY AI MARKETPLACE
Curated Models With a RESTful API
DELL AI READY SOLUTION & GPU BUILDING BLOCKS

NVIDIA GPU Cloud

Precision WS
Hyper-Converged
Dell PowerEdge
Dell DSSB440
DGX-1 Server

TensorFlow
Caffe
Spark
ML Rapids
mxnet
PyTorch

DELL EMC
CHALLENGES WITH HUMANITARIAN ASSISTANCE AND DISASTER RELIEF

REAL-TIME UNIFIED VIEW

DATA SOURCES

SOCIAL MEDIA

TRAFFIC/BODY CAMERAS

RADIOS

DRONES

LOCATING PERSONNEL

INFRASTRUCTURE STATUS

SURVIVOR STATUS

SYNCHRONIZING COLLECTION

USE CASES

LOCKHEED MARTIN