Data Center Innovation Day

Transform your business with Artificial Intelligence and Machine Learning

Dennis Ang
Director, ASEAN & ANZ
ENTERPRISE
AI/DEEP LEARNING
GAMING
PRO VISUALIZATION
DATA CENTER/
GPU COMPUTING
AI/DEEP LEARNING

THE WORLD LEADER IN
VISUAL AND AI COMPUTING
GPU COMPUTING PLATFORM AND AI DEEP LEARNING
RISE OF GPU COMPUTING

APPLICATIONS
ALGORITHMS
SYSTEMS
CUDA
ARCHITECTURE

NVIDIA DATA CENTER ACCELERATION PLATFORM
Single Platform Drives Utilization and Productivity

CUSTOMER USE CASES
- Speech
- Translate
- Recommender
- Healthcare
- Manufacturing
- Finance

CONSUMER INTERNET & INDUSTRY APPLICATIONS

SCIENTIFIC APPLICATIONS
- Molecular Simulations
- Weather Forecasting
- Seismic Mapping

VIRTUAL GRAPHICS
- Creative & Technical
- Knowledge Workers

APPS & FRAMEWORKS
- python
- TensorFlow
- RAPIDS
- PyTorch
- mxnet
- Chainer
- ONNX

AMBER
NAMD
+600 Applications

VIRTUAL GRAPHICS
- CATIA
- 3DS MAX
- Adobe
- Windows

CUDA-X & NVIDIA SDKs

MACHINE LEARNING
- cuDF
- cuML
- cuGRAPH

DEEP LEARNING
- cuDNN
- CUTLASS
- TensorRT

HPC
- OpenACC
- cuFFT

VIRTUAL GPU
- vDWS
- vPC
- vAPPS

CUDA & CORE LIBRARIES - cuBLAS | NCCL

TESLA GPUs & SYSTEMS
- TESLA GPU
- NVIDIA DGX FAMILY
- NVIDIA HGX
- SYSTEM OEM
- CLOUD
NVIDIA DEEP LEARNING SDK WITH DIFFERENT FRAMEWORKS

High Performance GPU-Acceleration for Deep Learning

DEEP LEARNING FRAMEWORKS

VISION
- Image Classification
- Object Detection

SPEECH
- Voice Recognition
- Language Translation

BEHAVIOR
- Recommendation Engines
- Sentiment Analysis

DEEP LEARNING
- cuDNN

MATH LIBRARIES
- cuBLAS
- cuSPARSE
- cuFFT

MULTI-GPU
- NCCL
AI AND DEEP LEARNING
AI REVOLUTION

Big Data
- 350 million images uploaded per day
- Petabytes of customer data hourly
- 300 hours of video uploaded every minute

Better Algorithms

Better Compute

“The Three Breakthroughs that have Finally Unleashed A.I. on the World”
GPU DEEP LEARNING IS A NEW COMPUTING MODEL

Billions of Trillions of Operations
GPU train larger models, accelerate time to market

10s of billions of image, voice, video queries per day
GPU inference for fast response, maximize data center throughput

Device

Data center inference
HIGHER ACCURACY WITH DEEP LEARNING

Human / Vehicle Detection
Facial Analysis
Person & Vehicle Analysis
People Counting
License Plate Recognition
Vehicle Classification

DATA: Hikvision, Dahua, and Uniview
AI IS DRIVING CUSTOMER ENGAGEMENT

**IMAGES / VIDEO**
- 60X Better Latency for Real-Time User Experience

**SPEECH / VOICE**
- 20% Higher Accuracy For Better User Experience

**RECOMMENDATIONS**
- 50% Improvement in User Engagement
AI TRANSFORMING EVERY INDUSTRY

**HEALTHCARE**
>80% Accuracy & Immediate Alert to Radiologists

**INFRASTRUCTURE**
50% Reduction in Emergency Road Repair Costs

**IOT**
>$6M / Year Savings and Reduced Risk of Outage
PLATFORM BUILT FOR AI
Accelerating Every Framework And Fueling Innovation

All Use-cases

Speech
Video
Translation
Personalization

All Major Frameworks

Caffe2
Chainer
Microsoft Cognitive Toolkit
mxnet
PaddlePaddle
PYTORCH
TensorFlow

Tensor Cores
Volta Tensor Core, NVSwitch, NVLink

NVLink
NVSwitch
NVIDIA AI PLATFORM
From data center to machines

Simulation
Model Training
Tesla GPUs
Datacenter

Deployment
Jetpack
Jetson
Robot | AIOT
CISCO GPU ACCELERATED DATA CENTER

Virtualization

GPU-Accelerated VDI

Windows 10 Virtual Workstations

Artificial Intelligence

Deep Learning

Machine Learning & Analytics

Citrix
vmware Horizon View

Office Ai Ps Id

Windows 10

mxnet TensorFlow

Caffe

RAPIDS

Caffe

RapidM

Microsoft CNTK

OmniSci

Graphistry
Cisco UCS C480ML – Purpose Built Server for AI/ML/DL

**Simplified Management**
Cisco Intersight or UCSM/CIMC managed

**Prevents Operation Silos**
 Extends Existing UCS Environments with Consistent, Cloud-Based Management

**Industry leading Spec**
Fully Integrated Platform Designed to Accelerate Deep Learning

**Accelerate AI/ML Projects**
Validated with Popular Machine Learning Software

➢ 40,000+ CUDA Cores
NVIDIA & CISCO PORTFOLIO: VIRTUALIZATION & COMPUTE

Test & Dev and Model Training

- **C240**
  - 2 x V100

- **HyperFlex 240**
  - 2 x V100 Per Node

  Option of GPU Only Nodes

- **C480**
  - 6 x PCIe V100

- **C480 ML**
  - 8x SXM2 V100 with NVLink

Inferencing & VDI

- **C/HX 220**
  - C/HX 240

  - 2x M10
  - 4 x T4
  - 6 x T4

Converged

AVAILABLE TODAY

Better Together, Customer Choice, Cisco Validated Design with Eco-system
Thank you!