DEPLOYING AI AT THE EDGE

JARED CONWAY - EDGE, ENTERPRISE AND CLOUD SOLUTIONS | NVIDIA
THE MOST POWERFUL TECHNOLOGY FORCE OF OUR TIME
DRIVERS FOR AI AT THE EDGE

IOT: IOT devices projected to grow to >150B by 2025, >1T by 2035

5G: 5G will deliver 1000X better bandwidth and 10X lower latency than 4G

AI: By 2025, AI at the edge has a potential total economic impact of up to $11T/year
THE CHALLENGE WITH AI AT THE EDGE

**CLOUD NATIVE**
Continuous improvement and deployment from cloud/data center to edge

**MANAGEMENT**
Software and model deployment, management and orchestration at scale

**SECURITY**
Enforce trust for every element of the stack
A new class of distributed GPU accelerated AI computing systems designed to gather and analyze continuous streams of data at the edge of the network

NVIDIA EGX is for applications that require:
- Low-latency interactions
- Reduced bandwidth to the cloud
- Data privacy or sovereignty

Supported by leading ecosystem partners

.5 TOPS  520 TOPS  10,000 TOPS
BIG DATA

NVIDIA DGX/HGX

NGC

NVIDIA AI

DEPLOY

MANAGE

CUDF
CUDIO
GPU DIRECT
RAPIDS
NVIDIA DGX/HGX

CUDNN
CUGRAPHC
CUML
RAPIDS

MELLANOX

CUDF
CUDIO
GPU DIRECT
RAPIDS
NVIDIA DGX/HGX

CUDNN
CUGRAPHC
CUML

MELLANOX

ANALYTICS
CLARA
METROPOLIS
JARVIS
ISAAC
ARIEL

NVIDIA EGX
Running on NVIDIA GPUs

IOT
- Helm charts in NGC for top AI applications
- Vertical Industry SDKs
- Cloud-Native EGX Stack w GPU Operator
- Scale from 2W to 2 Petaflops
- Commercially off the shelf (COTS) NGC-Ready for Edge HW
- Powered by NVIDIA GPU
ANNOUNCING NGC-READY FOR EDGE
Built for AI & ML Workloads

Validated for performance of AI stack
Additional testing for security and remote management
Pick preferred supplier

T4-ACCELERATED.

- acer
- Advantech
- ASRock Rack
- Atos
- DELL EMC
- Fujitsu
- GIGABYTE
- Hewlett Packard Enterprise
- Lenovo
- Mitac
- QCT
- Supermicro
- TYAN
EGX STACK GPU OPERATOR
Consistent Management Across GPU & CPU Nodes

• **Simple** | Single gold image for IT to deploy on all worker nodes
• **Easy** | One step install for GPU enablement
• **Productive** | Accelerates GPU-powered K8s clusters
• **Cloud Native** | Containerized GPU enablement
• **Accessible** | On NGC & Enterprise K8s services by Red Hat, VMWare, Nutanix, Cisco
AI FRAMEWORKS TARGET VERTICAL INDUSTRIES

METROPOLIS

AERIAL

JARVIS
HELM REGISTRY ON NGC

Automating Deployments on K8s
GPU-ACCELERATED AI APPLICATIONS

NGC REGISTRY OF CONTAINERS, MODELS AND HELM CHARTS

APPLICATION FRAMEWORKS

NVIDIA EGX STACK

NGC-READY FOR EDGE OEM EGX SYSTEMS

GLOBAL OEMs/ODMs

ENTERPRISE HYBRID CLOUD
The Industry’s Most Advanced AI Computing Platform

Largest domain of AI models
High throughput & low latency
One architecture scalable from device to cloud

NVIDIA EGX is AI-Optimized for Industry Use Cases

Hardware & software optimized for AI, storage, networking & security
Easy development & deployment of AI at edge
Optimized AI frameworks in NGC

NVIDIA has Deep AI Expertise

End-to-end, from development to deployment, from tools to experts