NGC

Adel El Hallak – Director of Product Management
Phil Rogers – Chief Software Architect

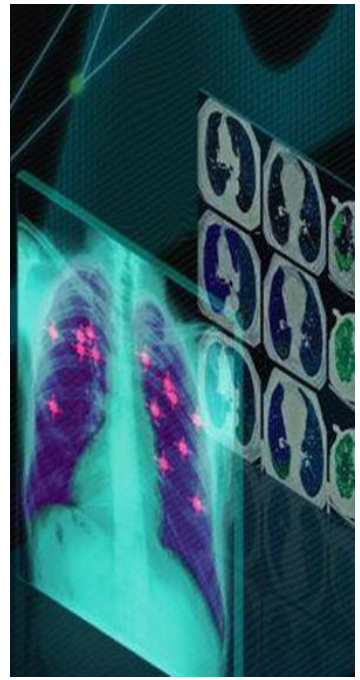
March 2019

GRAND CHALLENGES REQUIRE MASSIVE COMPUTING

AUTONOMOUS DRIVING

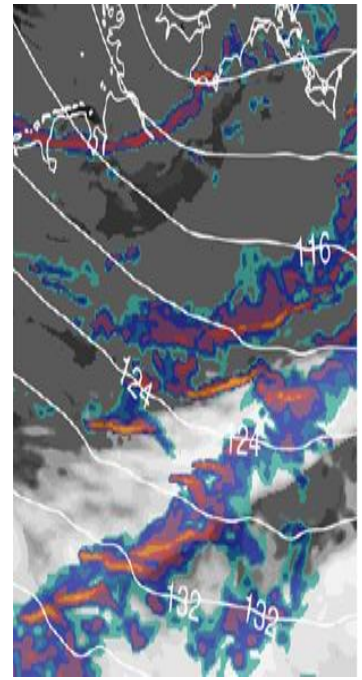
ASTROPHYSICS

GENOMICS



MEDICAL IMAGING

NUCLEAR FUSION



WEATHER

DIFFERENT ROLES. SAME GOALS.

Driving Productivity and Faster Time-to-Solutions

Data Scientists and Researchers

Eliminate mundane tasks, focus on science and research

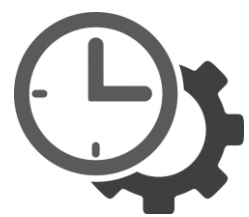
Developers

Speed up development with existing building blocks

Sysadmins

Deploy to production immediately

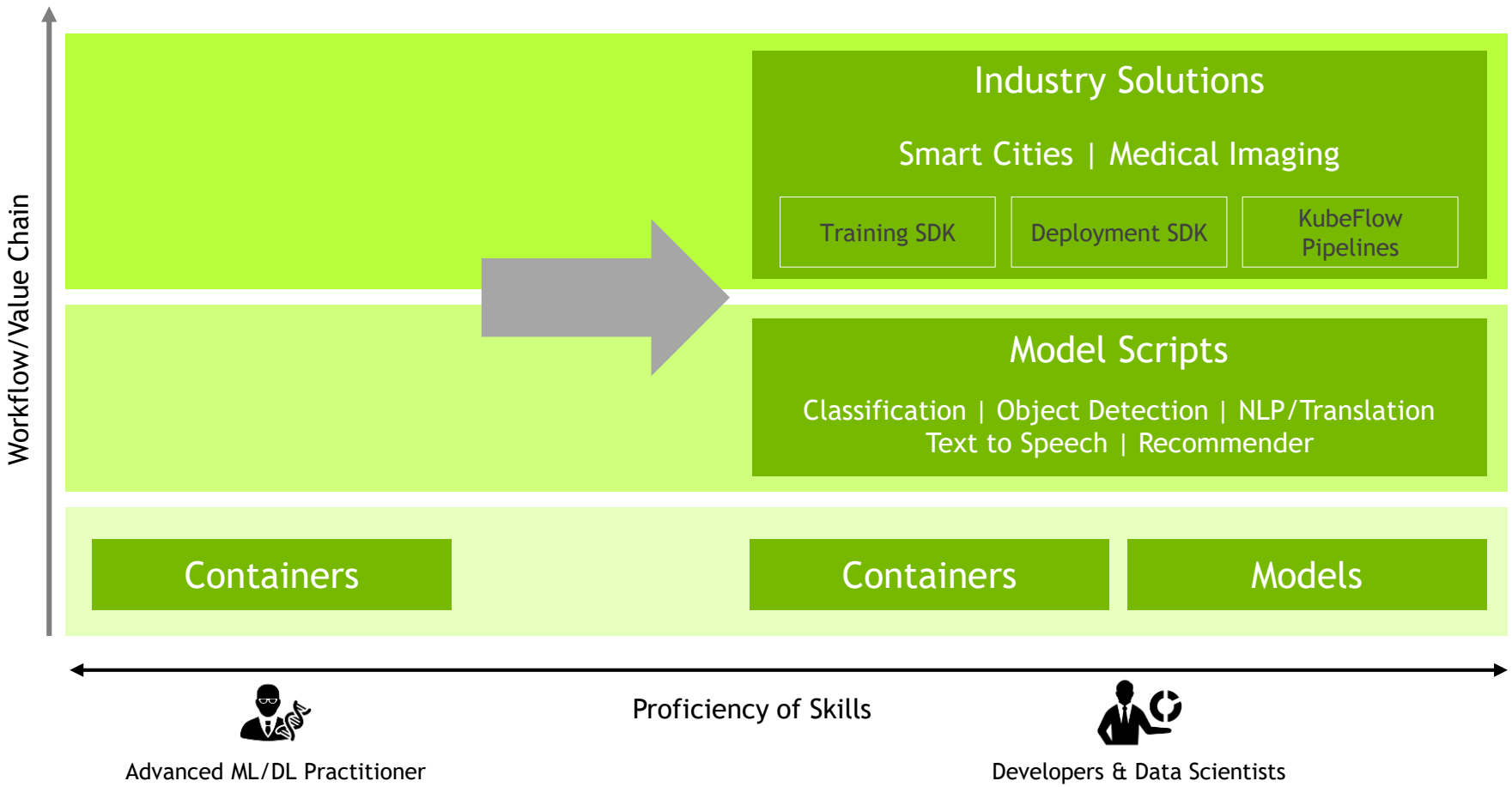
CHALLENGES UTILIZING AI & HPC SOFTWARE

EXPERTISE	INSTALLATION	OPTIMIZATION	PRODUCTIVITY	MAINTAINENCE
				
Building AI-centric solutions requires expertise	Complex, time consuming, and error-prone	Requires expertise to optimize framework performance	Users limited to older features and lower performance	IT can’t keep up with frequent software upgrades

NGC – SIMPLIFYING AI & HPC WORKFLOWS

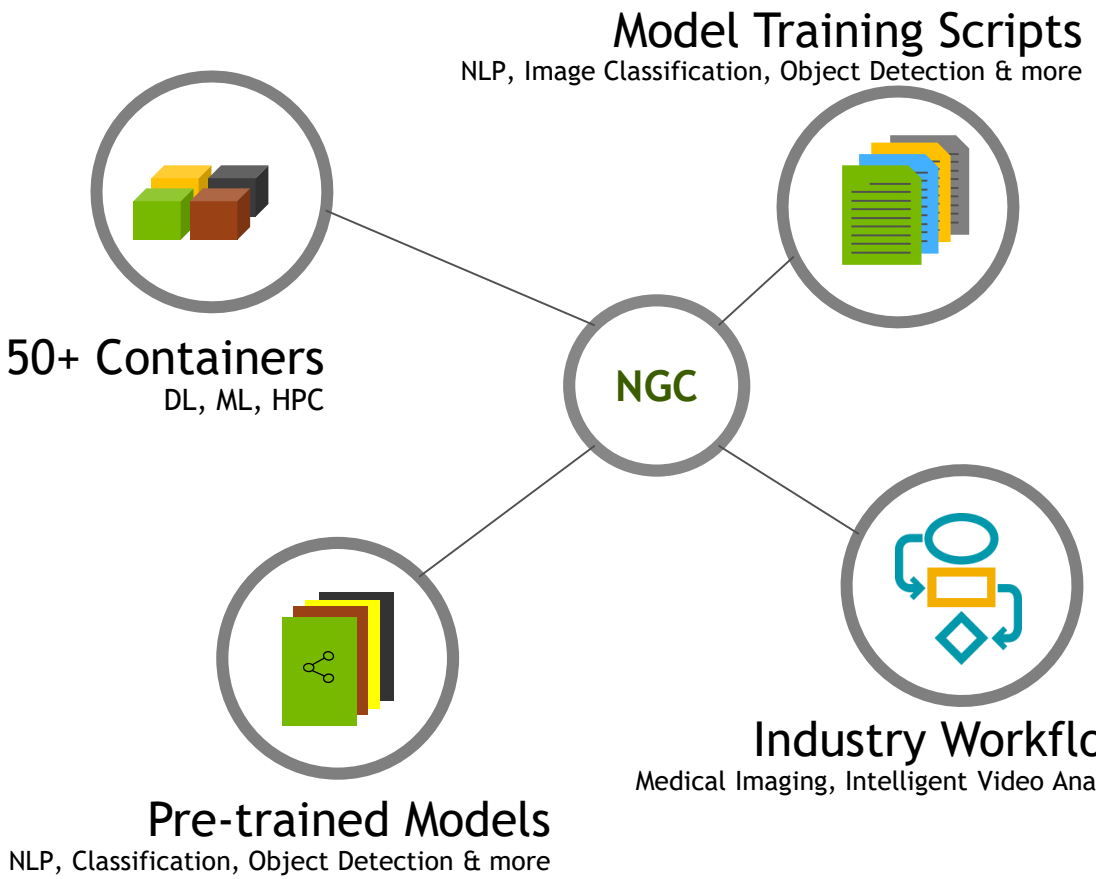
EMBEDDING EXPERTISE	FASTER DEPLOYMENTS	OPTIMIZED SOFTWARE	HIGHER PRODUCTIVITY	ZERO MAINTENANCE
				
Deliver greater value, faster	Eliminates installations. Simply Pull & Run the app	Key DL frameworks updated monthly for perf optimization	Better Insights and faster time-to-solution	Empowers users to deploy the latest versions with IT support

ANNOUNCING NEW NGC CAPABILITIES



THE NEW NGC

GPU-optimized Software Hub. Simplifying DL, ML and HPC Workflows



Simplify Deployments

Innovate Faster

Deploy Anywhere

CONTAINERS

CONTAINERS: SIMPLIFYING WORKFLOWS

WHY CONTAINERS

Simplifies Deployments

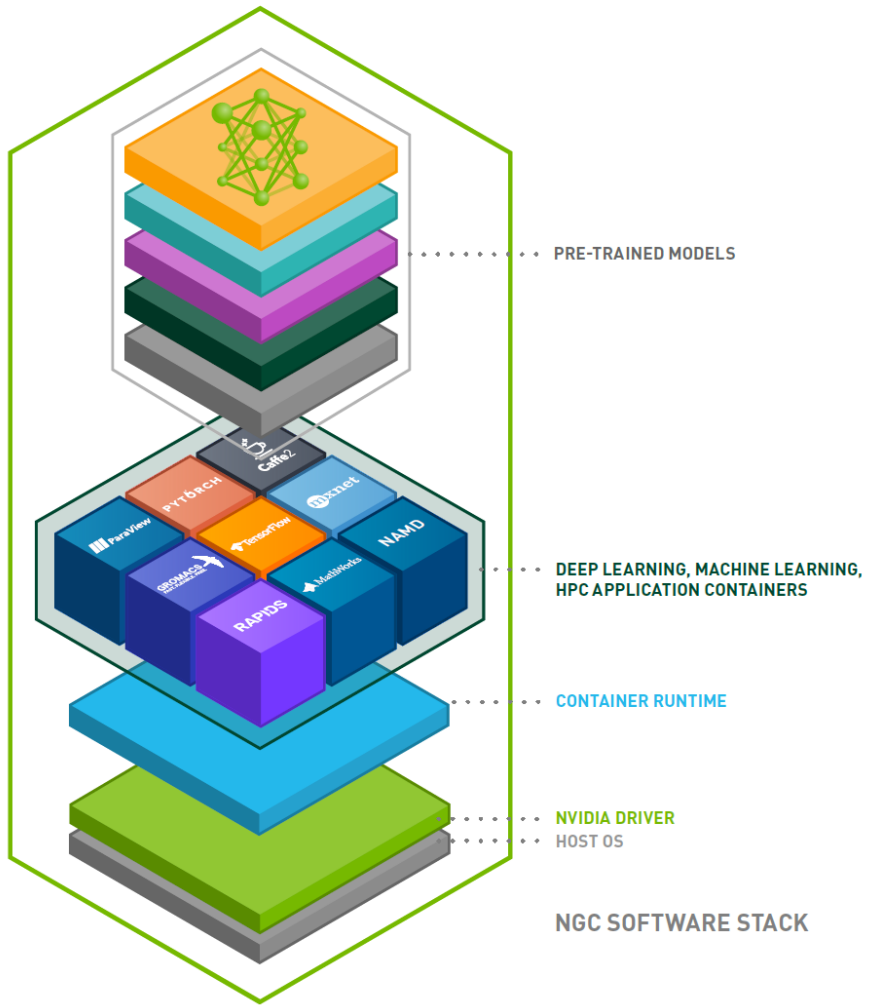
- Eliminates complex, time-consuming builds and installs

Get started in minutes

- Simply Pull & Run the app

Portable

- Deploy across various environments, from test to production with minimal changes



NGC CONTAINERS: ACCELERATING WORKFLOWS

WHY CONTAINERS

Simplifies Deployments

- Eliminates complex, time-consuming builds and installs

Get started in minutes

- Simply Pull & Run the app

Portable

- Deploy across various environments, from test to production with minimal changes

WHY NGC CONTAINERS

Optimized for Performance

- Monthly DL container releases offer latest features and superior performance on NVIDIA GPUs

Scalable Performance

- Supports multi-GPU & multi-node systems for scale-up & scale-out environments

Designed for Enterprise & HPC environments

- Supports Docker & Singularity runtimes

Run Anywhere

- Pascal/Volta/Turing-powered NVIDIA DGX, PCs, workstations, servers and top cloud platforms

GPU-OPTIMIZED SOFTWARE CONTAINERS

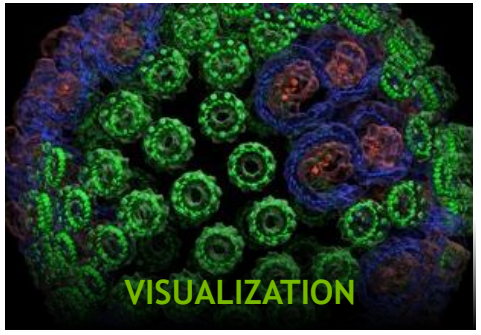
Over 50 Containers on NGC

TensorFlow | PyTorch | more

TensorRT | DeepStream | more

RAPIDS | H2O | more

NAMD | GROMACS | more



ParaView | IndeX | more

Parabricks

DALI

Eliminating CPU Bottleneck for DL Workflows

CPU Bottleneck Waste GPU Cycles

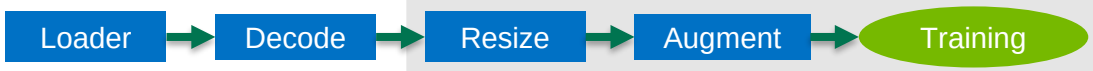


Image classification (ResNet-50)

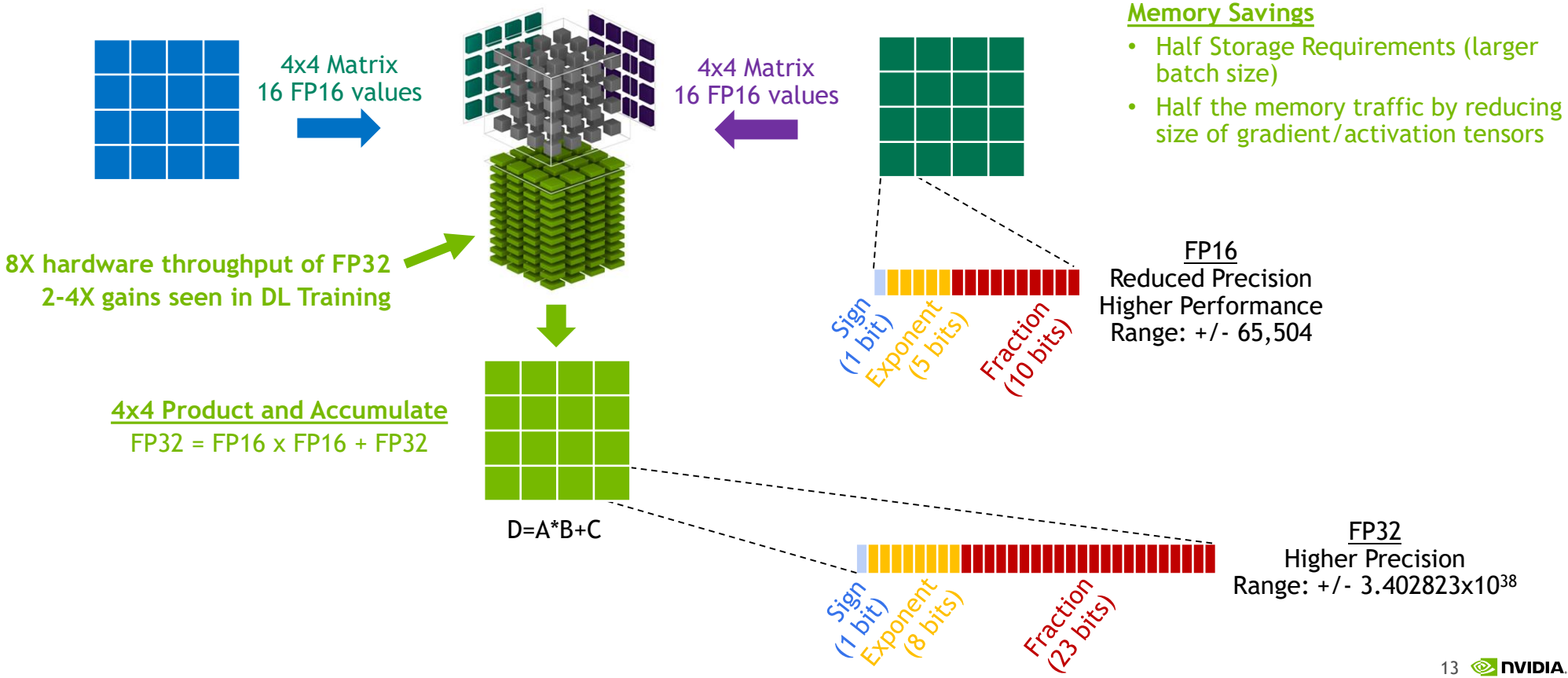
- Complex I/O pipelines
- Multi-pipeline frameworks
- Decreasing CPU:GPU ratio

DALI Shifts Workloads to GPUs

- Full input pipeline acceleration including data loading and augmentation
- Integrated in PyTorch, TF, MxNET
- Supports Resnet50 & SSD

TENSOR CORES BUILT FOR AI AND HPC

Mixed Precision Accelerator – Enabled by AMP



Memory Savings

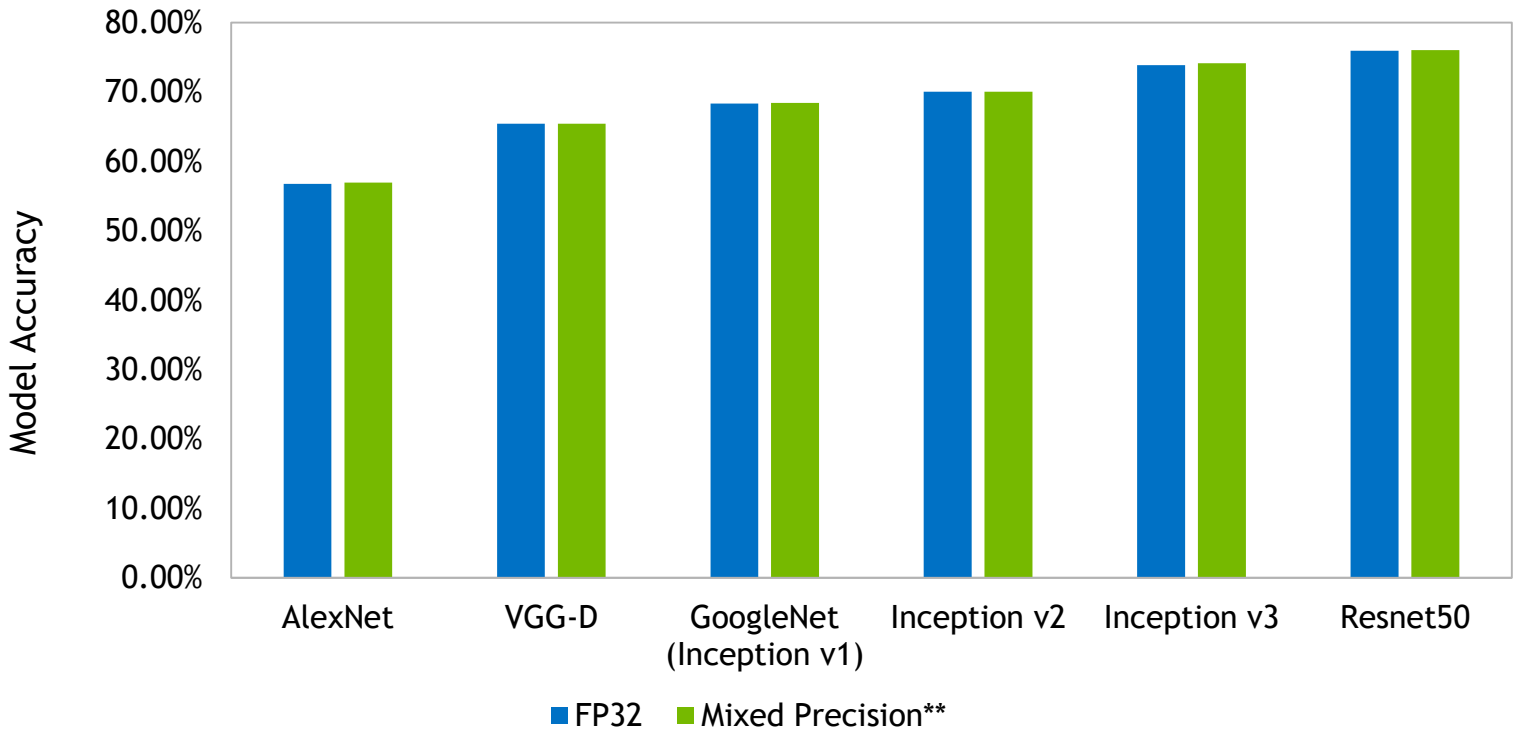
- Half Storage Requirements (larger batch size)
- Half the memory traffic by reducing size of gradient/activation tensors

FP16
Reduced Precision
Higher Performance
Range: +/- 65,504

FP32
Higher Precision
Range: +/- 3.402823×10^{38}

MIXED PRECISION MAINTAINS ACCURACY

Benefit From Higher Throughput Without Compromise

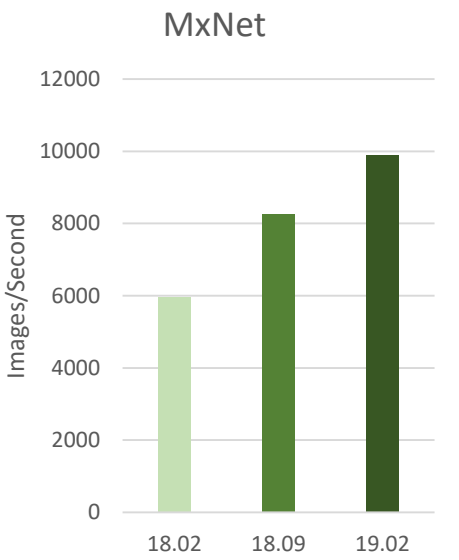


ILSVRC12 classification top-1 accuracy.
(Sharan Narang, Paulius Micikevicius *et al.*, "Mixed Precision Training", ICLR 2018)
**Same hyperparameters and learning rate schedule as FP32.

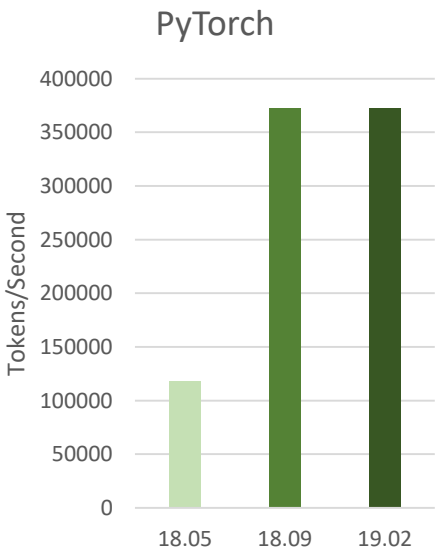
CONTINUOUS PERFORMANCE IMPROVEMENT

Developers' Software Optimizations Deliver Better Performance on the Same Hardware

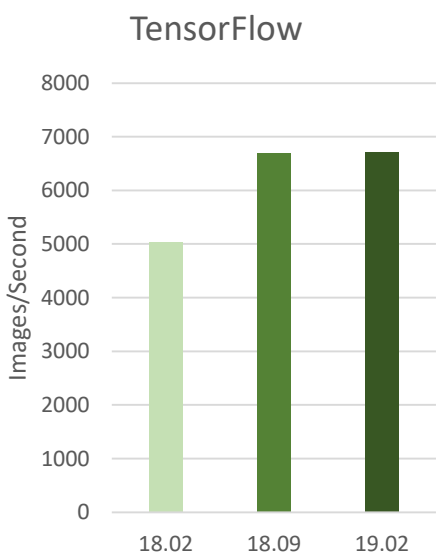
Monthly DL Framework Updates & HPC Software Stack Optimizations Drive Performance



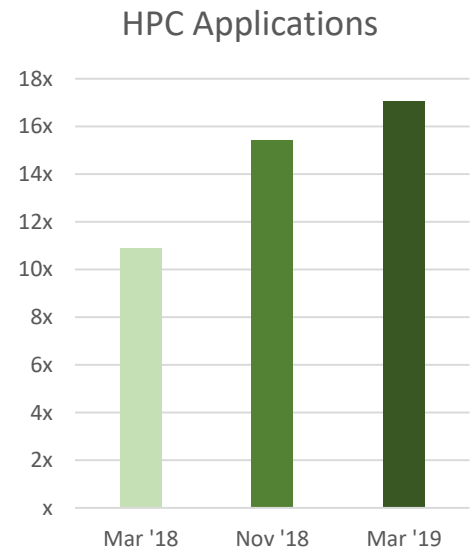
Mixed Precision | 128 Batch Size | ResNet-50 Training | 8x V100



Mixed Precision | 128 Batch Size | GNMT | 8x V100



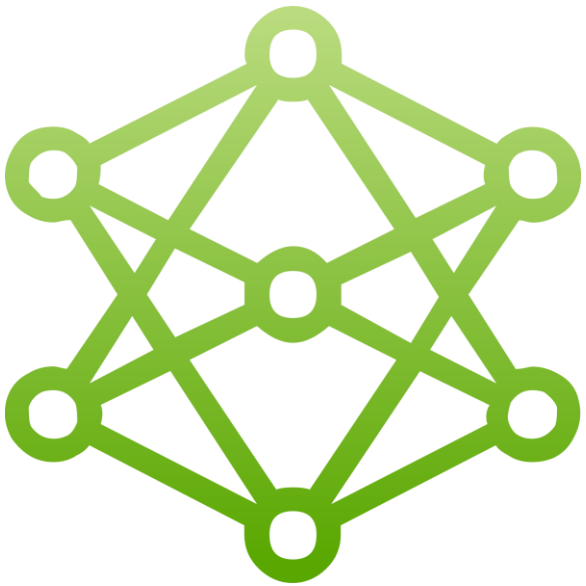
Mixed Precision | 256 Batch Size | ResNet-50 Training | 8x V100



Speedup across Chroma, GROMACS, LAMMPS, QE, MILC, VASP, SPECFEM3D, NAMD, AMBER, GTC, RTM | 4x V100 v. Dual-Skylake | CUDA 9 for Mar '18 & Nov '18, CUDA 10 for Mar '19

MODEL REGISTRY & MODEL SCRIPTS

ANNOUNCING THE NGC MODEL REGISTRY



Repository of Popular AI Models

- Starting point to retrain, prototype or benchmark against your own models
- Use As-Is or easily customize
- Private hosted registry for NGC Enterprise accounts to upload, share and version

DOMAIN SPECIFIC | INFERENCE-READY

PRE-TRAINED MODELS

- Domain specific for video analytics and medical imaging
- Use transfer learning and your own data to quickly create accurate AI
- Available models: Organ & tumor segmentation, x-ray classification, classification and object detection for video analytics

TENSORRT MODELS

- Ready for inference with Tensor Cores
- Precision: INT8, FP16, FP32
- Optimized for multiple GPU architectures
- Available Models: ResNet50, VGG16, InceptionV1, Mobilenet

ResNet-50 for Classification

Remove from Favorites | Download Latest Model

Publisher	Application	Version	Modified	Size
NVIDIA	Classification	4.0.4	11/21/2018 03:27 PM	93.02 GB

Training Framework	Inference Framework	Model Format	Precision	GPU Model
TensorFlow	TensorRT	TRTPlan	INT8	v100

Description

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nam viverra volutpat ipsum dignissim ve egestas. Maecenas egestas vestibulum erat, eu dapibus purus tempus. Fusce ornare molestie tortor, sed eleifend nisi vulputate vel. Sed semper ornare lacinia.

Labels

classification | fp32 | gpu-optimized | image classification | resnet-50 | tensorflow | tensorrt | trtplan | v100

Overview | Version History | File Browser | Release Notes | Related Model Scripts

Classification with ResNet-50
Caffe | FP16, INT8
4.0.4
built by NVIDIA 06/25/2018

Classification with ResNet-50
TensorFlow | FP16, INT8
4.0.4
built by NVIDIA 06/25/2018

LEARN | BUILD | OPTIMIZE | DEPLOY

MODEL SCRIPTS

- Best practices for training models
- Faster Performance with Optimized Libraries and Tensor Cores
- State-of-the-Art Accuracy
- Scripts for Classification, Detection, Recommendation, NLP, Segmentation, Speech Synthesis, Translation

Classification with ResNet-50

Add to Favorites | Download Latest Version

Publisher	Application	Version	Modified	Size
NVIDIA	Classification	4.0.4	11/21/2018 03:27 PM	93.02 GB

Training Framework	Model Format	Precision	GPU Model
TensorFlow	TRTPlan	FP16, INT8	v100

Description

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nam viverra volutpat ipsum dignissim ve egestas. Maecenas egestas vestibulum erat, eu dapibus purus tempus. Fusce ornare molestie tortor, sed eleifend nisi vulputate vel. Sed semper ornare lacinia.

Labels

classification | fp32 | gpu-optimized | image classification | resnet-50 | tensorflow | tensorrt | trtplan | v100

Overview Setup Quick Start Guide Performance Version History File Browser Release Notes Related Models

ResNet-50 for Classification
Caffe | FP32
4.0.4
built by NVIDIA 06/25/2018

ResNet-50 for Classification
TensorFlow | FP16
7.1.4
built by NVIDIA 06/25/2018

ResNet-50 for Classification
NVCaffe | FP16
18.5.2
built by NVIDIA 05/18/2018

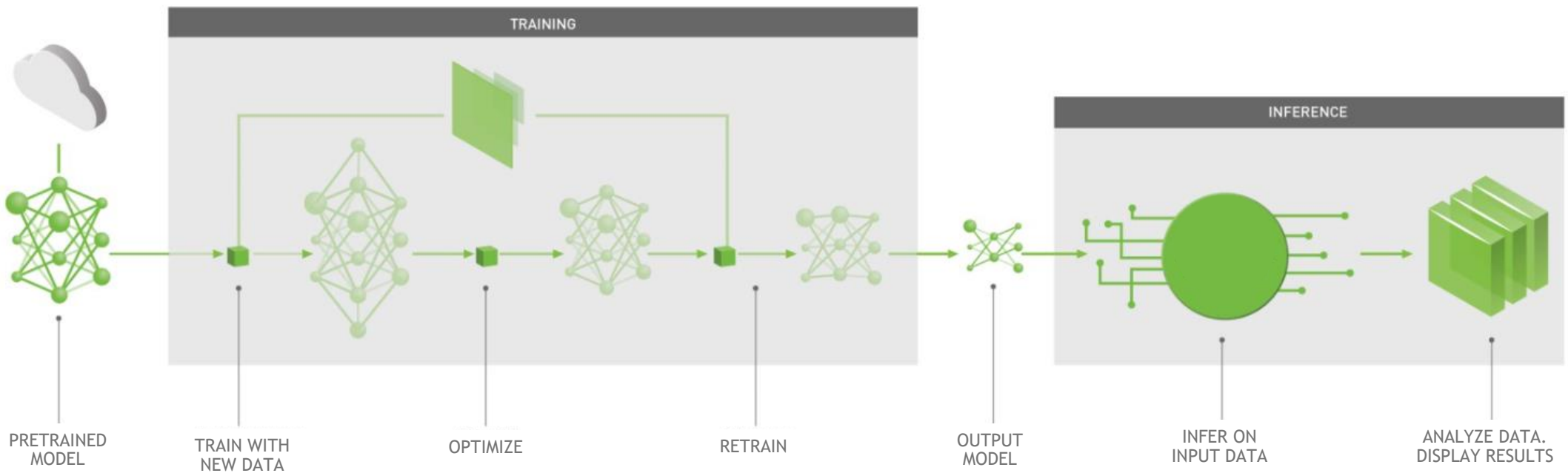
ResNet-50 for Classification
TensorFlow | INT8
4.0.4
built by NVIDIA 06/25/2018

ResNet-50 for Classification
PyTorch | FP16
20.05
built by NVIDIA 05/20/2018

INDUSTRY SOLUTIONS

END-TO-END DEEP LEARNING WORKFLOW

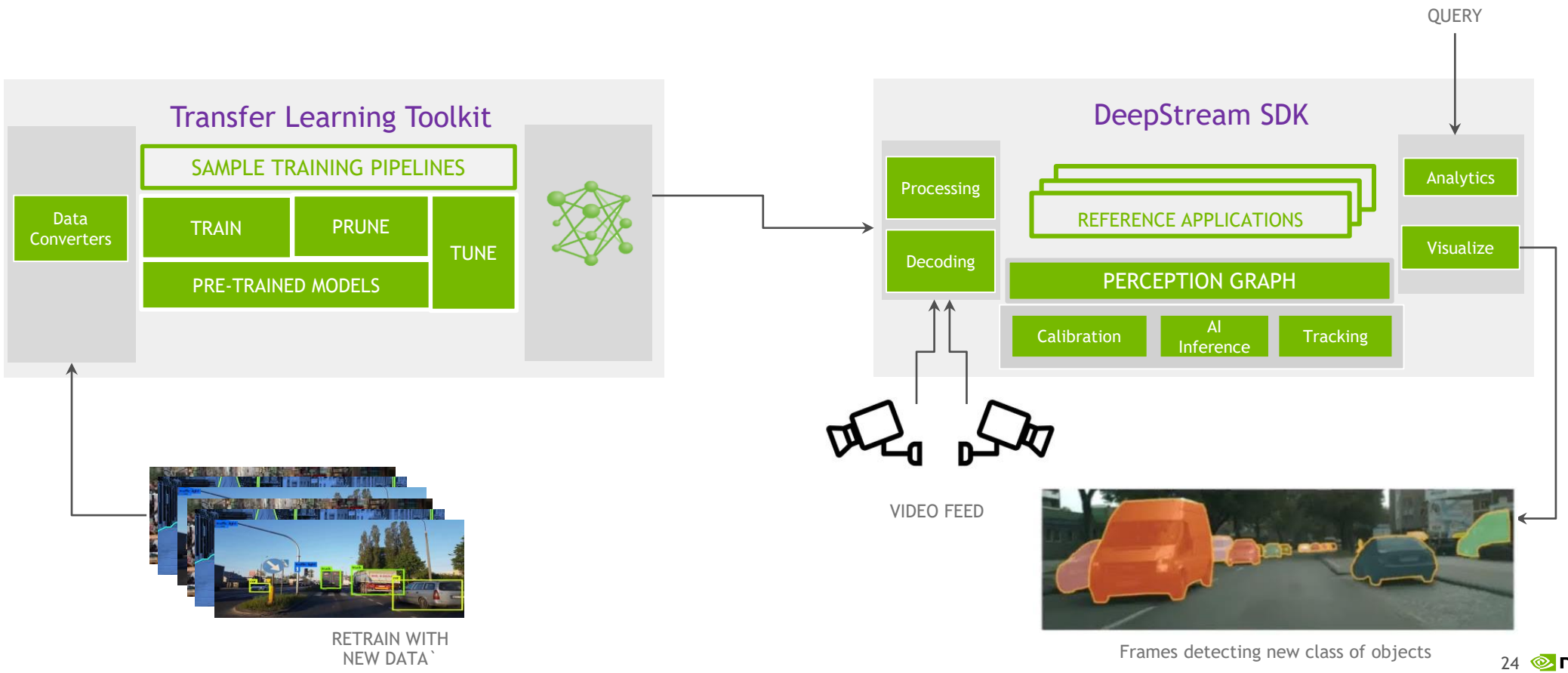
Pre-Trained Models | Training & Adaptation | Ready to Integrate



Accelerate time to market

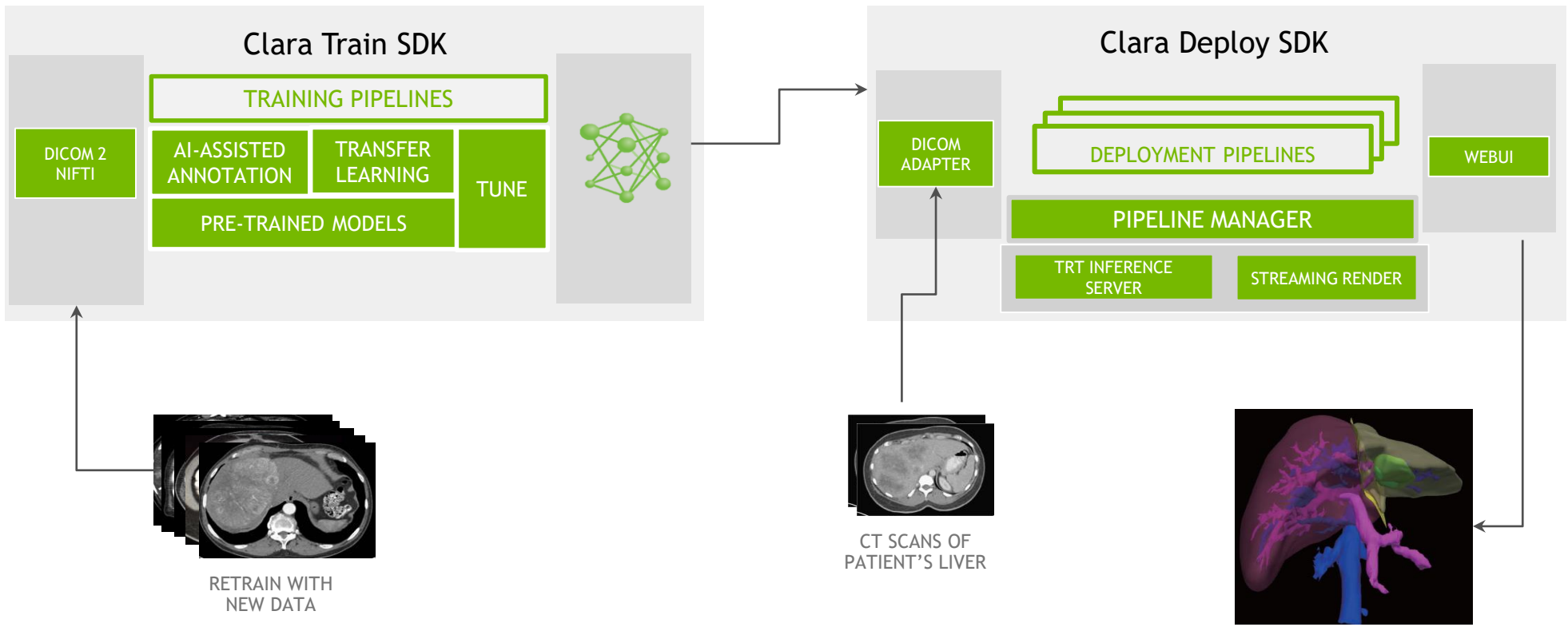
NVIDIA METROPOLIS

Intelligent Video Analytics for Smart Cities



NVIDIA CLARA AI PLATFORM

Organ Segmentation for Medical Imaging



NGC-READY SYSTEMS & SUPPORT SERVICES

NGC-READY SYSTEMS

VALIDATED FOR FUNCTIONALITY & PERFORMANCE OF NGC SOFTWARE

T4 & V100-ACCELERATED

NVIDIA NGC SUPPORT SERVICES

Minimize Downtime And Maximize System Utilization

Support Coverage

- NGC DL & ML containers
- NVIDIA drivers
- Kubernetes Device Plug-In
- NVIDIA Container Runtime
- CUDA

L1-L3 Support by NVIDIA's subject matter expert

- Live phone support during local biz hours
- 24/7 phone, portal, email to create support cases

Availability

- Exclusively for V100 & T4 NGC-Ready systems
- Availability
  - Now: Cisco
  - Q2: Dell, HPE, Lenovo
- Agreement between NVIDIA & end-customer
- Purchase from OEM

RUN ANYWHERE

Workflow / Value Chain

NGC

INDUSTRY SOLUTIONS

Smart Cities | Medical Imaging

TRANSFER LEARNING TOOLKITS | DEPLOYMENT SDKs | KUBEFLOW PIPELINES

MODEL SCRIPTS

Containers | Models

CLI

UI

NGC-Ready Systems

Atos, Cisco, Cray, Dell EMC, Fujitsu, Hewlett Packard Enterprise, inspur, Lenovo, Sugon, Supermicro

NVIDIA DGX Systems

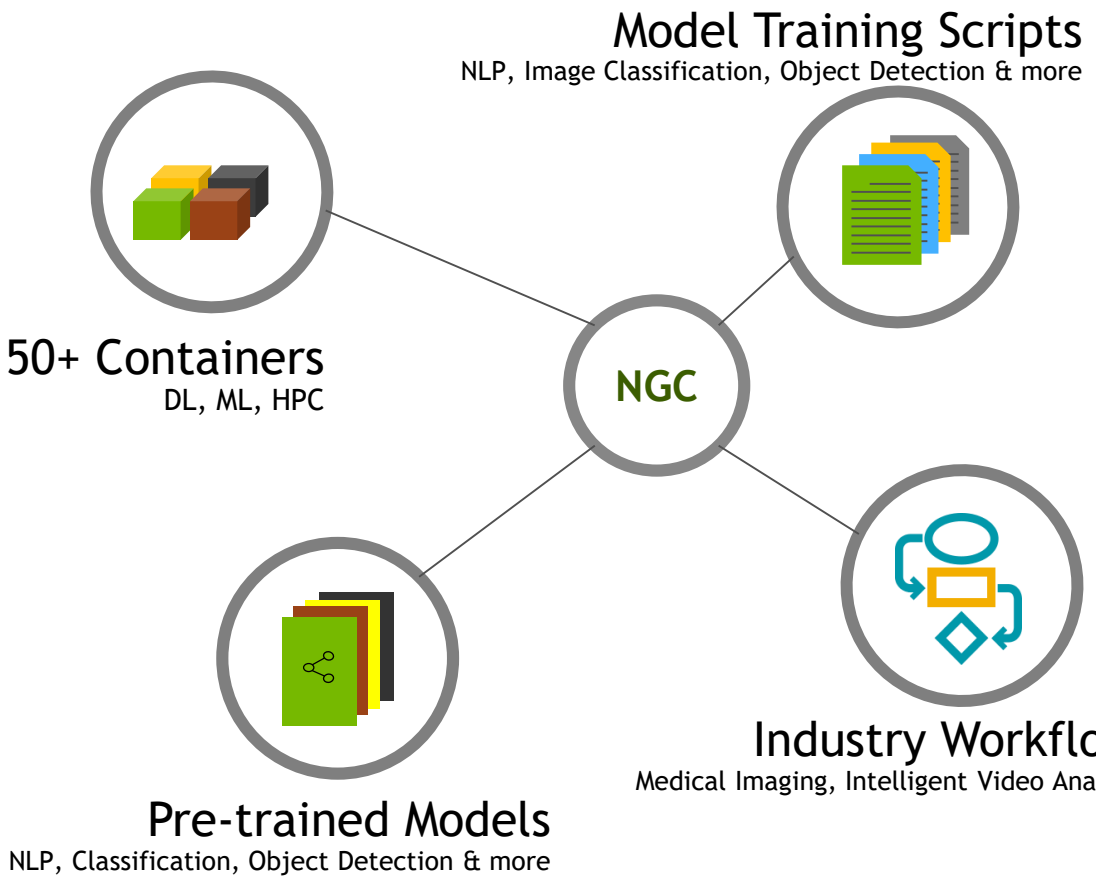
NVIDIA Support Services

All Major CSPs

Alibaba Cloud, aws, Azure, Baidu Cloud, Google Cloud, Oracle Cloud Platform, OVH Innovation is Freedom

THE NEW NGC

GPU-optimized Software Hub. Simplifying DL, ML and HPC Workflows



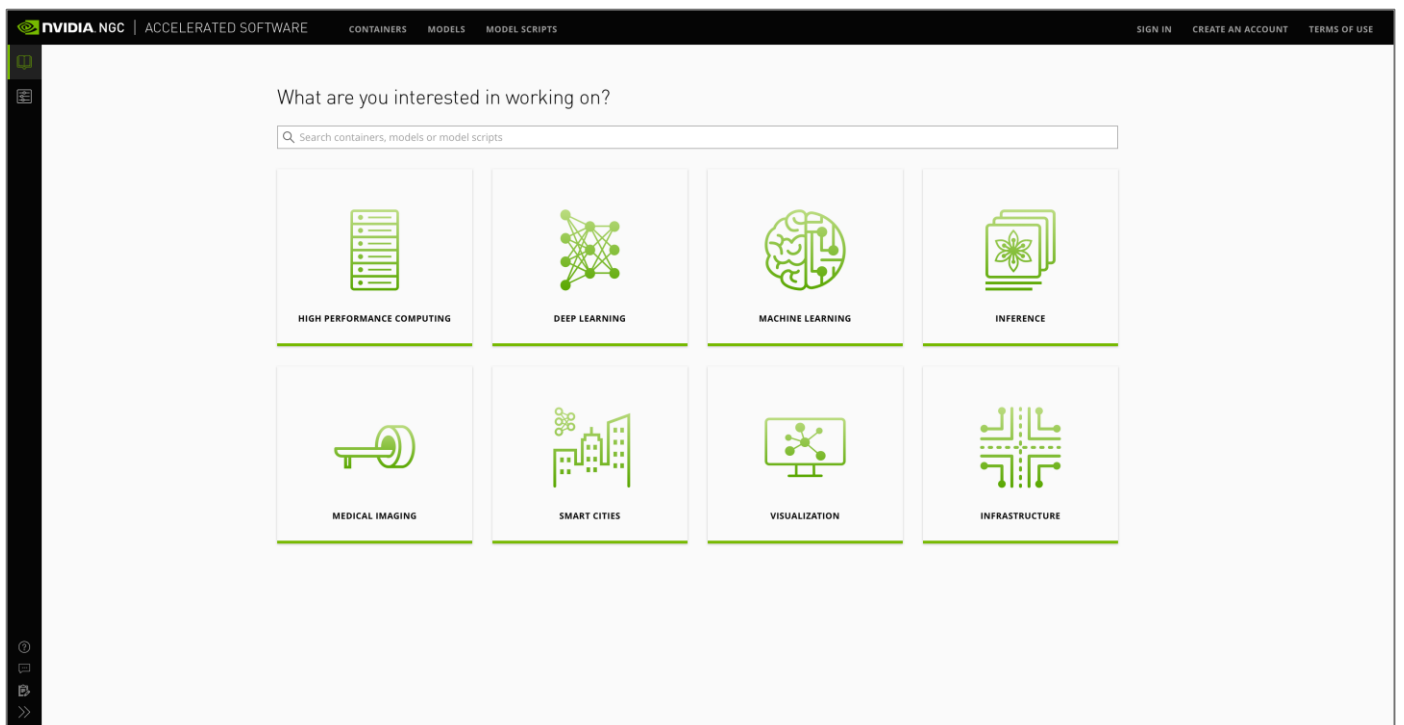
Simplify Deployments

Innovate Faster

Deploy Anywhere

GET STARTED WITH NGC

Explore the NGC Registry for DL, ML & HPC



Deploy containers:
ngc.nvidia.com

Learn more about NGC offering:
nvidia.com/ngc

Technical information:
developer.nvidia.com

GTC TALKS & RESOURCES

L9128 - High Performance Computing Using Containers WORKSHOP TU 10-12

S9525 - Containers Democratize HPC TU 1-2

S9500 - Latest Deep Learning Framework Container Optimizations W 9-10

SE285481 - NGC User Meetup W 7-9

Connect With the Experts

- NGC W 1-2
- NVIDIA Transfer Learning Toolkit for Industry Specific Solutions TU 1-2 & W 2-3
- DL Developer Tool for Network Optimization W 5-6