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

<table>
<thead>
<tr>
<th>EXPERTISE</th>
<th>INSTALLATION</th>
<th>OPTIMIZATION</th>
<th>PRODUCTIVITY</th>
<th>MAINTAINENCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Building AI-centric solutions requires expertise</td>
<td>Complex, time consuming, and error-prone</td>
<td>Requires expertise to optimize framework performance</td>
<td>Users limited to older features and lower performance</td>
<td>IT can’t keep up with frequent software upgrades</td>
</tr>
</tbody>
</table>
NGC - SIMPLIFYING AI & HPC WORKFLOWS

**EMBEDDING EXPERTISE**
Deliver greater value, faster

**FASTER DEPLOYMENTS**
Eliminates installations. Simply Pull & Run the app

**OPTIMIZED SOFTWARE**
Key DL frameworks updated monthly for perf optimization

**HIGHER PRODUCTIVITY**
Better Insights and faster time-to-solution

**ZERO MAINTENANCE**
Empowers users to deploy the latest versions with IT support
ANNOUNCING NEW NGC CAPABILITIES

Workflow/Value Chain

Industry Solutions
- Smart Cities | Medical Imaging
  - Training SDK
  - Deployment SDK
  - KubeFlow Pipelines

Model Scripts
- Classification | Object Detection | NLP/Translation
- Text to Speech | Recommender

Containers

Proficiency of Skills
- Advanced ML/DL Practitioner
- Developers & Data Scientists
THE NEW NGC
GPU-optimized Software Hub. Simplifying DL, ML and HPC Workflows

- **50+ Containers**
  - DL, ML, HPC

- **Pre-trained Models**
  - NLP, Classification, Object Detection & more

- **Industry Workflows**
  - Medical Imaging, Intelligent Video Analytics

- **Model Training Scripts**
  - NLP, Image Classification, Object Detection & more

- **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

DEEP LEARNING
TensorFlow | PyTorch | more

INFERENC
TensorRT | DeepStream | more

MACHINE LEARNING
RAPIDS | H2O | more

HPC
NAMD | GROMACS | more

VISUALIZATION
ParaView | IndeX | more

GENOMICS
Parabricks
DALI
Eliminating CPU Bottleneck for DL Workflows

CPU Bottleneck Waste GPU Cycles

- 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
Mixed Precision Accelerator - Enabled by AMP

FP32
Higher Precision
Range: +/- 3.402823x10^{38}

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

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

4x4 Product and Accumulate
FP32 = FP16 x FP16 + FP32

8X hardware throughput of FP32
2-4X gains seen in DL Training

D = A*B+C

TENSOR CORES BUILT FOR AI AND HPC
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

HPC Applications

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
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

<table>
<thead>
<tr>
<th>Publisher</th>
<th>NVIDIA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Application</td>
<td>Classification</td>
</tr>
<tr>
<td>Version</td>
<td>4.0.4</td>
</tr>
<tr>
<td>Modified</td>
<td>11/21/2018 03:27 PM</td>
</tr>
<tr>
<td>Size</td>
<td>91.02 GB</td>
</tr>
</tbody>
</table>

### Labels
- Classification
- fp32
- gpu-optimized
- image classification
- resnet-50
- tensorflow
- trtplan
- v100

## Related Models

<table>
<thead>
<tr>
<th>Model</th>
<th>Publisher</th>
<th>Version</th>
<th>Format</th>
<th>Precision</th>
<th>GPU Model</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet-50 for Classification</td>
<td>NVIDIA</td>
<td>4.0.4</td>
<td>Caffe</td>
<td>FP12</td>
<td></td>
<td>4.0.4</td>
</tr>
<tr>
<td>ResNet-50 for Classification</td>
<td>TensorFlow</td>
<td>7.1.4</td>
<td>TensorFlow</td>
<td>FP16</td>
<td></td>
<td>18.5.2</td>
</tr>
<tr>
<td>ResNet-50 for Classification</td>
<td>NVIDIA</td>
<td>4.0.4</td>
<td>NVcaffe</td>
<td>FP16</td>
<td></td>
<td>4.0.4</td>
</tr>
<tr>
<td>ResNet-50 for Classification</td>
<td>PyTorch</td>
<td>20.05</td>
<td>PyTorch</td>
<td>INT8</td>
<td></td>
<td>05/26/2018</td>
</tr>
</tbody>
</table>
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

Transfer Learning Toolkit
- SAMPLE TRAINING PIPELINES
- TRAIN
- PRUNE
- TUNE
- PRE-TRAINED MODELS

DeepStream SDK
- Processing
- Decoding
- REFERENCE APPLICATIONS
  - PERCEPTION GRAPH
    - Calibration
    - AI Inference
    - Tracking

Data Converters

QUERY

VIDEO FEED
Frames detecting new class of objects

RETRAIN WITH NEW DATA
NVIDIA CLARA AI PLATFORM
Organ Segmentation for Medical Imaging

Clara Train SDK
- TRAINING PIPELINES
- DICOM 2 NIFTI
- AI-ASSISTED ANNOTATION
- TRANSFER LEARNING
- PRE-TRAINED MODELS
- TUNE

Clara Deploy SDK
- DEPLOYMENT PIPELINES
- PIPELINE MANAGER
- TRT INFERENCE SERVER
- STREAMING RENDER
- WEBUI

CT SCANS OF PATIENT’S LIVER
RETRAIN WITH NEW DATA
SEGMENTED LIVER

Clara Train SDK
Clara Deploy SDK
NGC-READY SYSTEMS & SUPPORT SERVICES
NGC-READY SYSTEMS

VALIDATED FOR FUNCTIONALITY & PERFORMANCE OF NGC SOFTWARE

T4 & V100-ACCELERATED

- Atos
- Cisco
- Cray
- Dell EMC
- Fujitsu
- HPE
- Huawei
- Inspur
- Lenovo
- Sugon
- Supermicro
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

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

• Live phone support during local biz hours
• 24/7 phone, portal, email to create support cases
THE NEW NGC
GPU-optimized Software Hub. Simplifying DL, ML and HPC Workflows

50+ Containers
DL, ML, HPC

Pre-trained Models
NLP, Classification, Object Detection & more

Model Training Scripts
NLP, Image Classification, Object Detection & more

Industry Workflows
Medical Imaging, Intelligent Video Analytics

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