AI PRINCIPLE AND SDK - DL
DEVELOPMENT FRAMEWORK

Jay Chen | Solution Architect
jaych@nvidia.com
AGENDA

NGC - (10 minutes)
Accelerating AI workflows with NGC

SimNet - (3 minutes)
AI accelerated Simulation toolkit

AI for Weather - (3 minutes)
Trend of researches/applications
NGC: ACCELERATING AI WORKFLOWS WITH NGC
AI HAS REACHED THE TIPPING POINT

Over 50 Percent of Organizations Are Using AI in Production

AI Adoption In The Enterprise 2020 O’Reilly Survey
AI WORKFLOWS ARE COMPLEX

Data Engineer
- Extract
- Transform
- Load
- Feature Engineering

Data Scientist
- Setup
- Build Model
- Train
- Optimize

Developer
- Integrate
- Optimize
- Build App

DevOps
- Deploy & Monitor

Timeline
NGC CATALOG HELPS SIMPLIFY AND ACCELERATE AI WORKFLOWS

Timeline

Containers  Pre-Trained Models  SDKs  Helm Charts

Accelerate ETL  Simplify Training  Build Faster  Deploy Reliably
NGC CATALOG - GPU-OPTIMIZED SOFTWARE

Build AI Faster, Deploy Anywhere

NGC Catalog

CONTAINERS  TRAINED MODELS  INDUSTRY APP FRAMEWORKS  HELM CHARTS  COLLECTIONS

100+  30+  End-to-End AI Workflows  ML, Inference  Task based collections

x86  |  ARM  |  POWER

CLOUD  ON-PREM  HYBRID CLOUD  EDGE
NGC CONTAINERS ENABLE YOU TO FOCUS ON BUILDING AI

ENTERPRISE READY SOFTWARE
Scanned for CVEs, malware, crypto
Tested for reliability
Backed by Enterprise support

PERFORMANCE OPTIMIZED
Scalable
Updated monthly
Better performance on the same system

DEPLOY ANYWHERE
Docker | cri-o | containerd | Singularity
Bare metal, VMs, Kubernetes
Multi-cloud, on-prem, hybrid, edge
DO WHAT YOU DO BEST, FASTER

DO WHAT YOU DO BEST, FASTER

Scanned for CVEs, malware, crypto
Tested for reliability
Backed by Enterprise support

BERT Large
v20.05 (V100)

BERT Large
v20.07 (V100)

BERT Large
v20.07 (A100)

Scalable
Updated Monthly
Better performance on the same system

 ENTERPRISE READY SOFTWARE

PERFORMANCE OPTIMIZED

Deployed
Bare metal, VMs, Kubernetes
Multi-cloud, on-prem, hybrid, edge

DOCKER ANYWHERE

Scanned for CVEs, malware, crypto
Tested for reliability
Backed by Enterprise support

BERT Large and ResNet-50 v1.5 Training performance with TensorFlow on a single node BERT Large (32GB) & A100 (40GB) Mixed Precision. Batch size for BERT: 10 (V100), 24 (A100), ResNet: 512 (V100, v20.05), 256 (v20.07)

BERT Large and ResNet-50 v1.5 Training performance with PyTorch on 1x V100 & 1x A100 Mixed Precision. Batch size 32768

BERT Large and ResNet-50 v1.5 Training performance with PyTorch on 1x V100 & 1x A100 Mixed Precision. Batch size 32768

BERT Large and ResNet-50 v1.5 Training performance with PyTorch on 1x V100 & 1x A100 Mixed Precision. Batch size 32768
EASILY IDENTIFY THE RIGHT MODELS WITH CREDENTIALS

WIDE RANGE OF USE CASES
- ResNet-50, SSD, MobileNet, VGG16
- WaveGlow, BERT, NeMo
- Wide & Deep, DLRM & many more

PRE-TRAINED MODELS
- Faster training
- Higher accuracy
- Transparency through credentials

RESOURCES
- Get started with code samples
- Customize NGC models
- Reproduce with recipes
INDUSTRY APP FRAMEWORKS FOR END-TO-END AI WORKFLOWS

TRANSFER LEARNING TOOLKIT
Domain adaptability
Significantly reduce development time

TENSORRT
Optimizes for low latency and high-throughput
Integrated with major frameworks

TRITON
High performance inference
Supports multiple frameworks
CLOUD NATIVE APPROACH DELIVERS CONSISTENT DEPLOYMENTS

**KUBERNETES & HELM CHARTS**
- Consistent deployment across platforms
- Simplify multi-container deployments
- Auto scaling and self healing

**GPU OPERATOR**
- Automate GPU resources in Kubernetes
- Installs drivers and monitoring tools
- Single golden image for IT to manage

**NGC HELM REGISTRY**
- NVIDIA AI
- Industry application frameworks
- Partner applications
EVERYTHING YOU NEED TO BUILD YOUR AI IN ONE LOCATION

Search NGC for the app or use case

The Collection walks through all the required assets and how to use them together

Fine-tune and deploy

Deploy container

Start Jupyter Notebook instance

Download model

COLLECTIONS

Compatible assets grouped together, removes guesswork
Curated software by use cases
Detailed documentation further simplifies work for users

READY-TO-USE

Conversational AI
Computer Vision
NVIDIA AI App Frameworks
START BUILDING AI FASTER WITH THE NGC CATALOG
ngc.nvidia.com

Resources

- Technical blogs: Train | Optimize | Infer | Deploy - Conversational AI
- Webinars: Building AI solutions on: AWS Sagemaker | OpenShift
- Other blogs, webinars, success stories on Pathfactory
- Learn more at nvidia.com/ngc
SIMNET: AI ACCELERATED SIMULATION TOOLKIT
**SIMNET: AI BASED SIMULATION**

**Physical Prototyping**
- Physical Prototyping is iterative, time consuming, costly and not optimized for material and characteristics.

**Traditional Simulations**
- Traditional numerical solvers work on one problem at a time making design process time consuming, do not address real-time simulations, data assimilation, inverse problems.

**AI based Techniques**
- Data driven NN require data, are oblivious to physics laws, suffer from interpolation/extrapolation errors and are not generalizable.

**Past**
- Physical Prototyping

**Present**
- Virtual Prototype

**Future**
- Product
AI IN COMPUTATIONAL SCIENCES

Primary Driver: Data vs. Physics

Forward Solution  Inverse Solution  Data-Driven Solution
## DATA DRIVEN METHODS

<table>
<thead>
<tr>
<th><strong>Pros</strong></th>
<th><strong>Cons</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Not dependent on Physics</td>
<td>No physics awareness; Generalization ability may be limited</td>
</tr>
<tr>
<td></td>
<td>Need to generate a lot of simulations (accuracy dependent on the simulation code)</td>
</tr>
<tr>
<td></td>
<td>Not very efficient for complex 3D geometries/curved surfaces</td>
</tr>
<tr>
<td></td>
<td>Interpolation/extrapolation errors</td>
</tr>
</tbody>
</table>
FPGA HEAT SINK

Multi-Physics Application: Fluids + Heat Transfer

https://www.youtube.com/watch?v=Oq2Mpi5pF1w
Early Access

https://developer.nvidia.com/simnet

AI-Accelerated Simulation Toolkit

Simulations are pervasive in science and engineering. They are computationally expensive and don’t easily accommodate measured data coming from sources such as sensors or cameras. NVIDIA SimNet™ is a simulation toolkit, which addresses these challenges using AI and physics. Whether you’re looking to get started with AI-driven physics simulations or working on complex nonlinear physics problems, NVIDIA SimNet is your toolkit for solving forward, inverse, or data assimilation problems.
WEATHER AI: TREND OF RESEARCHES/APPLICATIONS
MACHINE LEARNING IS NEW & POWERFUL

AI STRATEGIES & WORKSHOPS

NOAA Artificial Intelligence Strategy

ECMWF-ESA Machine Learning Workshop

NOAA 2nd Workshop on Leveraging AI

ERAD 2020 AI Section

https://www.star.nesdis.noaa.gov/star/meeting_2020AIWorkshop.php

https://nrc.noaa.gov/LinkClick.aspx?fileticket=0Ip2p2Gu3rA%3D&tabid=91&portalid=0

https://www.ecmwf.int/en/learning/workshops/ecmwf-esa-workshop-machine-learning-earth-system-observation-and-prediction

https://www.ecmwf.int/en/learning/workshops/ecmwf-esa-workshop-machine-learning-earth-system-observation-and-prediction


https://www.star.nesdis.noaa.gov/star/meeting_2020AIWorkshop.php

https://www.ecmwf.int/en/learning/workshops/ecmwf-esa-workshop-machine-learning-earth-system-observation-and-prediction

IMPROVING CLIMATE PREDICTIONS

Improve physics-based models conditionally

FASTER SOLUTION
HIGHER RESOLUTION

MORE ACCURATE
PHYSICS

IMPROVED
DYNAMICS

DATA ASSIMILATION

DATA THINNING

DATA INFILLING
AN EXAMPLE: INPAINTING

Repair an image that has missing data

Image Inpainting for Irregular Holes Using Partial Convolutions
Guilin Liu Fitsum A. Reda Kevin J. Shih Ting-Chun Wang Andrew Tao Bryan Catanzaro, NVIDIA Corporation
https://arxiv.org/abs/1804.07723, ECCV 2018
RECONSTRUCT MISSING CLIMATE INFO

Train on Climate Model data (20CR, CMIP5) to repair Observational data (HadCRUT4)

Artificial intelligence reconstructs missing climate information

Christopher Kadow, David Matthew Hall and Uwe Ulbrich

HadCRUT4 1901-2012

HadCRUT4 by 20crAI 1901-2012

https://www.nature.com/articles/s41561-020-0582-5, Nature Geoscience 2020
# RECONSTRUCT MISSING CLIMATE INFO

Kriging / PCA / AI

<table>
<thead>
<tr>
<th>Method</th>
<th>20CR Test r (r.m.s.e.)</th>
<th>CMIP Test r (r.m.s.e.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Masked</td>
<td>0.9706 (0.1142)</td>
<td>0.9857 (0.0684)</td>
</tr>
<tr>
<td>Kriging</td>
<td><strong>0.9899 (0.0732)</strong></td>
<td><strong>0.9959 (0.0388)</strong></td>
</tr>
<tr>
<td>PCA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20crPCA</td>
<td>0.9879 (0.0887)</td>
<td>0.9895 (0.0918)</td>
</tr>
<tr>
<td>cmipPCA</td>
<td>0.9860 (0.0801)</td>
<td>0.9948 (0.0426)</td>
</tr>
<tr>
<td>AI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20crAI*</td>
<td>0.9941 (0.0547)</td>
<td>0.9984 (0.0239)</td>
</tr>
<tr>
<td>cmipAI*</td>
<td>0.9942 (0.0486)</td>
<td>0.9984 (0.0190)</td>
</tr>
</tbody>
</table>

Evaluation on model data (20CR, CMIP5)

El Niño, July 1877 (HadCRUT4)
GTC Fall 2020

Exploring the Frontiers of Deep Learning for the Earth System [A22147]


Exploring the Frontiers of Deep Learning for the Earth System [A22147], GTC Fall Oct 2020
KEY TAKEAWAYS

NGC

SimNet

AI + Weather
加入 NVIDIA 開發者計畫

請掃描左邊 QR code 加入 NVIDIA 開發者計畫，活動結束前完成註冊，即可換取精美小禮物。

請向 NVIDIA 工作人員出示主旨為「Your application for the program NVIDIA Developer Program is approved」的 email。

developer.nvidia.com/developer-program
Thank You