NVIDIA – from gaming to ChatGPT: The rise of accelerated computing ...or... Interesting things to do with GPUs

Paul Graham, Senior Solutions Architect, NVIDIA

pgraham@nvidia.com

NVIDIA

GPU Computing

Computer Graphics

Artificial Intelligence

Million-X Speedup for Innovation and Discovery

Simulation + AI

CLIMATE CHANGE
FourcastNet

DIGITAL BIOLOGY
Orbnet

Universe of GPU Computing

This material is released by NVIDIA Corporation under the Creative Commons Attribution 4.0 International (CC BY 4.0)

Generative AI

Multi-modal intelligence

Domain specific SDKs and Frameworks to accelerate AI and Data Science research

Fast track your AI with NVIDIA Pretrained models

RAPIDS

GPU-ACCELERATED DATA SCIENCE WORKFLOW

RAPIDS.ai

DATA PREPARATION - ETL

cuDF: Python drop-in **pandas** replacement built on CUDA.

GPU-accelerated Spark

MODEL TRAINING

cuML: GPU-acceleration of popular ML algorithms e.g. **XGBoost**

Easy-to-adopt, **scikit-learn** like interface

VISUALIZATION

Effortless exploration of datasets, billions of records in milliseconds

Dynamic interaction with data = faster ML model development

NVIDIA Clara for Healthcare and Life Sciences

World's Largest Data Industry | 36% CAGR by 2025

NVIDIA Modulus

Framework for Developing Physics Machine Learning Neural Network Models

TRAINING NEURAL NETWORKS USING BOTH DATA AND THE GOVERNING EQUATIONS

ADVANCING SCIENTIFIC DISCOVERY WITH MODULUS

RENEWABLE ENERGY
Siemens Gamesa: Up to 4000X Speedup of Wind Turbine Wake Optimization

CLIMATE CHANGE
45,000X Speedup of Extreme weather Prediction with FourCastNet

INDUSTRIAL HPC
NETL: 10,000X Faster Build Of high-fidelity surrogate models

HEALTHCARE
Achieve high-fidelity results faster for blood flow in inter-cranial aneurysm

DIGITAL TWINS
Kinetic Vision: Design Optimization Using parameterized models

MODULUS: Optimising heat sink design

FourCastNet
Full model emulation of ERA5

Scope	Global, Medium-Range Weather
Model Type	Full-Atmosphere AI Surrogate
Architecture	Fourier Neural Operator + Transformer
Resolution	25km
Training Data	ERA5 Reanalysis (1979 – Present)
Initial Condition	ERA5 / GFS / UFS
Training Time	O(1000) GPU-hours (NVIDIA A100)
Inference Time	0.5 sec (2-week forecast)
Calibration	IC + Bayesian model uncertainty
Speedup vs NWP	O(10,000 – 100,000)
Power Savings	O(10,000)
Max Stable Rollout	250+ days
Project Type	Open-source

FCN is fully data-driven.
Atmospheric dynamics were learned from data alone.

- **Skill** now matching European Center for individual,14-day forecasts.
  - Beating IFS on 60% of tracked metrics, 5-10% less error than IFS on 3-7 day timescale.

Before: IFS is more accurate

Now: SFNO becoming more accurate, a milestone.

FCN skill now matches or exceeds that of IFS (as measured by ACC and RMSE)

Digital twins – NVIDIA Omniverse for virtual worlds to simulate the largest challenges in science and engineering

OMNIVERSE

VIRTUAL WORLDS ARE ESSENTIAL FOR THE NEXT ERA OF AI

NVIDIA Omniverse Foundational Technology

Key Attributes of a Live, Full Fidelity Digital Twin

Enabling Accurate And Trustworthy Virtual Testing, Experimentation, And Optimization

GROUND TRUTH
Single Source of Truth for Virtual Datasets

PHYSICALLY ACCURATE REPLICA
Obeys the Laws of Physics

AI-ENABLING, AI-ENABLED
Optimized by AI & Enables Training of AI

PERFECTLY SYNCHRONIZED
Precision Timed, Perfectly Synced to the Real World

NVIDIA PhysX
Rigid & Soft Body Dynamics, Destruction, Fluid & Fire

Import Physics Instance
Offline from External 3rd Party Application

NVIDIA Modulus
Accelerated with Physics-ML Platform

Physics Simulation in Omniverse

Multiple Ways to Visualize and Simulate Physics in Omniverse

Reinforcement Learning – Isaac Gym

Digital twin – BMW Factory

Digital Twin Platform for Scientific Computing

Powered by Modulus, Omniverse

INDUSTRIAL DIGITAL TWIN FOR ENERGY GENERATION

SIEMENS energy

10,000x Faster Prediction

SCIENTIFIC DIGITAL TWIN OF EARTH

45,000X Speedup with FourCastNet

Maximizing Wind Energy Production Using Wake Optimization

NVIDIA Modulus and Omniverse

UKAEA using Omniverse To Develop Fusion Reactor Digital Twin

Video: https://www.youtube.com/watch?v=tJgR1TSBD0k

TORNADO VISUALIZATION

NanoVDB, IndeX, GPUDirect Storage, Iray, Omniverse

Generative AI – the “iPhone moment of AI”: implications and opportunities for science

Large Language Models

New opportunities

Content Generation
Marketing Copy, Blog Generator, Code Generator

Summarization
Catchy headlines, Article Summarization

Translation
Human language to code, Language to Language Translation

Chatbots
Question & Answering, Dialogue Generation

Virtual Avatars
Sentiment Analysis, Intelligent Responses

Healthcare
Protein Structure Prediction, Protein Discovery, Drug Discovery

Edify Generative AI

Foundation Models Create Images, Videos and 3D from Text

NVIDIA

GenAI For Science Research and Discoveries

The Race for Foundation Models for Science is on

Biology: AlphaFold

2021

Genomics: GenSLM

Nov 2022

Trillion Parameter Consortium

Summer 2023

Materials : MatterGen

Jan 2024

NVIDIA BioNeMo Framework

AI-powered drug discovery

AI Framework for BioMolecular Large Language Models

BioNeMo Framework

ESM-1, ProtT5, MMB

Framework - On-Prem or Cloud

Domain Specific

Drug discovery researchers need an LLM framework that speaks the language of biology and chemistry.

Optimized For Scale

LLMs are massive requiring model parallelism and supercomputing resources to train.

Pre-Trained Models

Training LLMs takes enormous compute resources and is very time consuming. Pre-trained models are ready for production.

Cloud Native

Use pre-trained models, create powerful customized workflows and deploy in the cloud.

Retrieval Augmented Generation (RAG)

Foundation/Custom Model Response

"What was the pressure of the tank at 11 p.m. last night?"

"I was trained 2 months ago and do not have the current data "

Information Retrieval Model Response

"What was the pressure of the tank at 11 p.m. last night?"

"The pressure at 11 p.m. last night was 345 psi"

Dataset

Example: UI automation

Using ChatGPT and GPT-4 to Generate 3D Content in Omniverse

Expert, Natural Q&A with NVIDIA Omniverse Avatar for Project Tokkio

NVIDIA Tokkio Showcase and Omniverse Avatar Cloud Engine (ACE)

Resources

NGC: GPU-Optimized Software Hub

Simplifying DL, ML and HPC Workflows

50+ Containers

DL, ML, HPC

Model Training Scripts

NLP, Image Classification, Object Detection and more

60 Pre-trained Models

NLP, Image Classification, Object Detection and more

Workflows

Medical Imaging, Intelligent Video Analytics

TensorFlow | PyTorch | more

RAPIDS | H2O | more

NAMD | GROMACS | more

ParaView | IndeX | more

NVIDIA

NVIDIA LaunchPad

Instantly experience end-to-end workflows for AI, data science, 3D design collaboration, and more

Get Started at nvidia.com/launchpad

RESOURCES AVAILABLE TO ACADEMICS

DLI Teaching Kits:

Material for incorporation into student courses, includes free access to online training for students, requested by a lecturer/professor. Edge AI and Robotics, Data Science, Accelerated Computing …

Academic Workshops:

The NVIDIA website lists free academic workshops that our Ambassadors are giving around the world that you can go and attend

Bootcamps:

~2 day tailored training events or workshops, typically for a target group e.g. N-ways to GPU, AI for Science

Hackathons:

Bring-your-code events, in-depth porting/optimization with access to expert mentors

Final thought

Driving Breakthrough HPC Application Performance

Center Panel: Geometric mean of application speedups vs. Broadwell in 2016 | benchmark applications | Amber [PME-Cellulose_NVE], Chroma [HMC], GROMACS [ADH Dodec], MILC [Apex Medium], NAMD [stmv_nve_cuda], PyTorch (BERT Large Fine Tuner], Quantum Espresso [AUSURF112-jR]; TensorFlow [ResNet-50], VASP 6 [Si Huge]; Random Forest make_blobs (160000 x 64 : 10).

Thankyou!

Paul Graham, Senior Solutions Architect

pgraham@nvidia.com