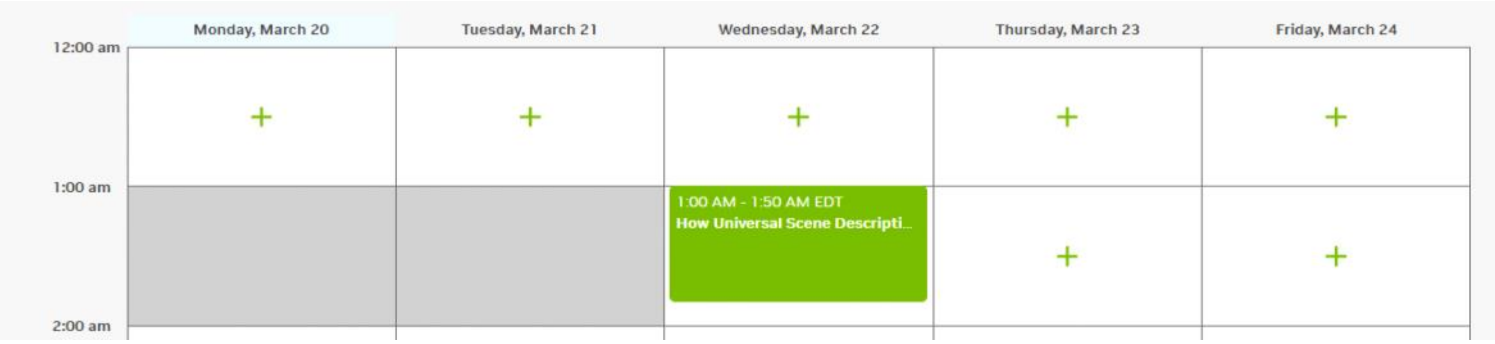
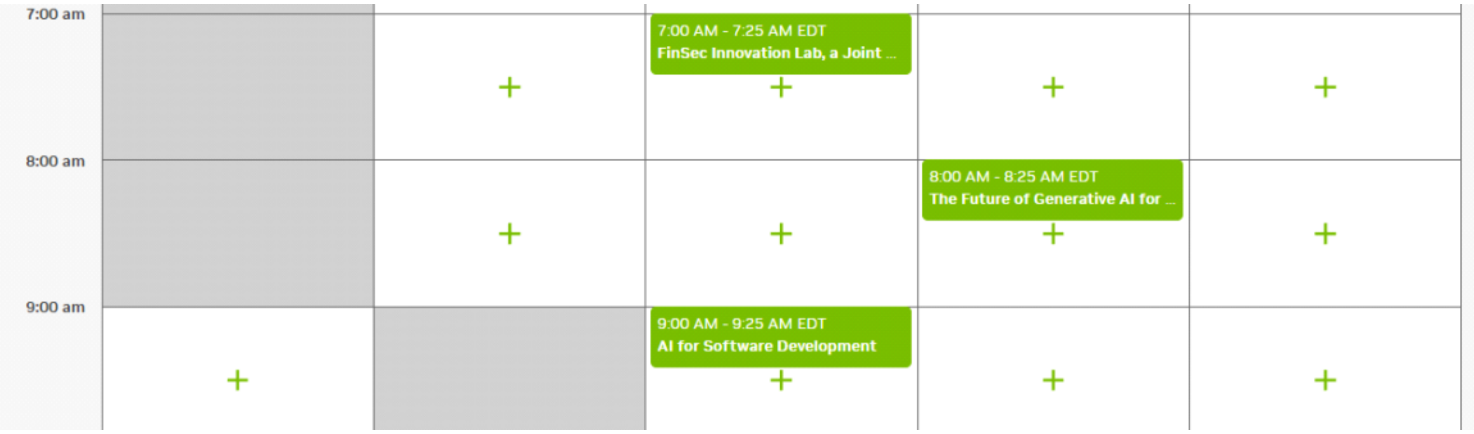
GTC 3/2023

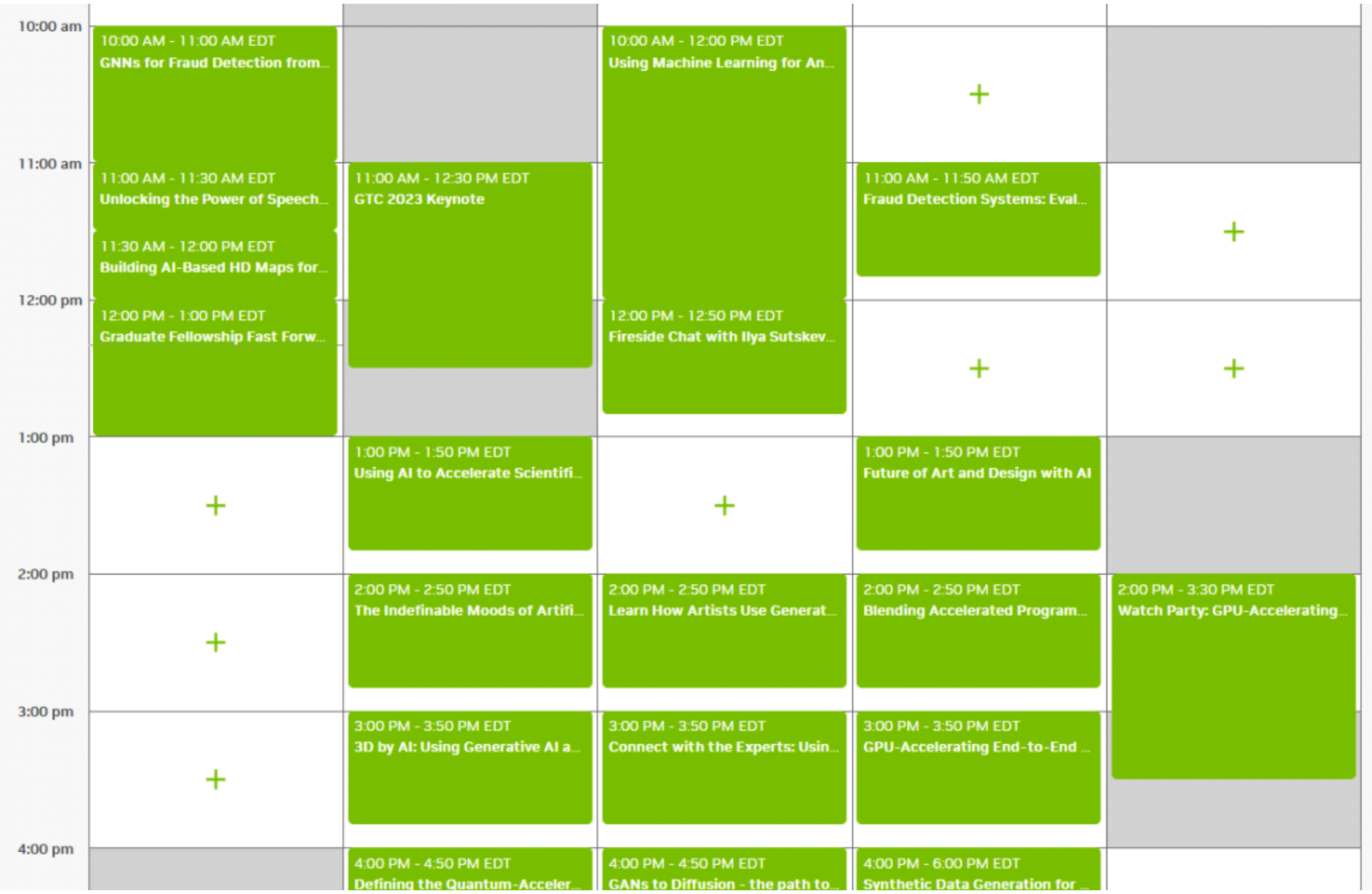
Monday, March 20, 2023 10:38 AM



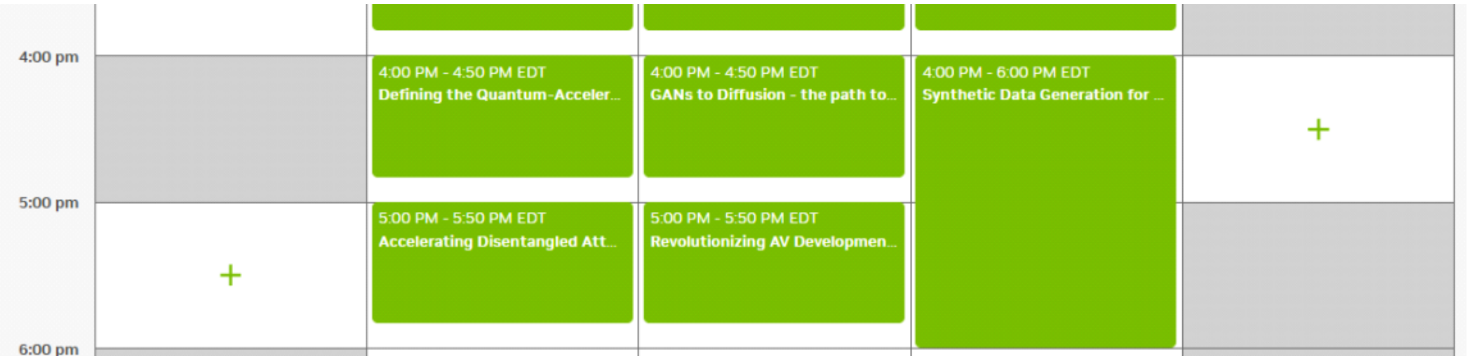
	Monday, March 20	Tuesday, March 21	Wednesday, March 22	Thursday, March 23	Friday, March 24
12:00 am	+	+	+	+	+
1:00 am			1:00 AM - 1:50 AM EDT How Universal Scene Descripti...	+	+
2:00 am					



7:00 am		+	7:00 AM - 7:25 AM EDT FinSec Innovation Lab, a Joint ... +	+	+
8:00 am		+	+	8:00 AM - 8:25 AM EDT The Future of Generative AI for ... +	+
9:00 am	+		9:00 AM - 9:25 AM EDT AI for Software Development +	+	+



10:00 am	10:00 AM - 11:00 AM EDT GNNs for Fraud Detection from...		10:00 AM - 12:00 PM EDT Using Machine Learning for An...	+	
11:00 am	11:00 AM - 11:30 AM EDT Unlocking the Power of Speech... 11:30 AM - 12:00 PM EDT Building AI-Based HD Maps for...	11:00 AM - 12:30 PM EDT GTC 2023 Keynote		11:00 AM - 11:50 AM EDT Fraud Detection Systems: Eval...	+
12:00 pm	12:00 PM - 1:00 PM EDT Graduate Fellowship Fast Forw...		12:00 PM - 12:50 PM EDT Fireside Chat with Ilya Sutskev...	+	+
1:00 pm	+	1:00 PM - 1:50 PM EDT Using AI to Accelerate Scientifi...	+	1:00 PM - 1:50 PM EDT Future of Art and Design with AI	
2:00 pm	+	2:00 PM - 2:50 PM EDT The Indefinable Moods of Artifi...	2:00 PM - 2:50 PM EDT Learn How Artists Use Generat...	2:00 PM - 2:50 PM EDT Blending Accelerated Program...	2:00 PM - 3:30 PM EDT Watch Party: GPU-Accelerating...
3:00 pm	+	3:00 PM - 3:50 PM EDT 3D by AI: Using Generative AI a...	3:00 PM - 3:50 PM EDT Connect with the Experts: Usin...	3:00 PM - 3:50 PM EDT GPU-Accelerating End-to-End ...	
4:00 pm		4:00 PM - 4:50 PM EDT Defining the Quantum-Acceler...	4:00 PM - 4:50 PM EDT GANs to Diffusion - the path to...	4:00 PM - 6:00 PM EDT Synthetic Data Generation for ...	



4:00 pm		4:00 PM - 4:50 PM EDT **Defining the Quantum-Acceler...**	4:00 PM - 4:50 PM EDT **GANs to Diffusion - the path to...**	4:00 PM - 6:00 PM EDT **Synthetic Data Generation for ...**	+
5:00 pm	+	5:00 PM - 5:50 PM EDT **Accelerating Disentangled Att...**	5:00 PM - 5:50 PM EDT **Revolutionizing AV Developmen...**		
6:00 pm					

GNN fraud

Monday, March 20, 2023 10:38 AM

[GNNs for Fraud Detection from Industry Leaders [S51704]](https://register.nvidia.com/flow/nvidia/gtcspring2023/attendeeportal/page/sessioncatalog/session/1666638995788001G5Zm)

From
<https://register.nvidia.com/flow/nvidia/gtcspring2023/attendeeportal/page/sessioncatalog/session/1666638995788001G5Zm>

Leaders from various industries will speak on how graph neural networks have been used in fraud detection use cases to unlock critical business benefits. Financial fraud, fake reviews, bot assaults, account takeovers, and spam are all examples of online fraud and harmful activity. In recent years, GNNs have gained traction for fraud detection problems, revealing suspicious nodes (in accounts and transactions, for example) by aggregating their neighborhood information through different relations — in other words, by checking whether a given account has sent a transaction to a suspicious account in the past. In the context of fraud detection, the ability of GNNs to aggregate information contained within the local neighborhood of a transaction enables them to identify larger patterns that may be missed by just looking at a single transaction. Hear from industry leaders like Mastercard, TigerGraph, Bunq, and BNY Mellon.

Nitendra Rajput, Senior Vice President and Head, AI Garage, Mastercard

Benika Hall, Senior Data Scientist and Solutions Architect, Financial Services Industry, NVIDIA (moderator)

Jay Yu, VP Product and Innovation, TigerGraph

Jason Rosado, Principal, Product Management – Payment Validation, BNY Mellon

Ali el Hassouni, Data Professional, Bunq

Industry: All Industries
Topic: Deep Learning - Frameworks

- **Scheduled**

Monday, Mar 2010:00 AM - 11:00 AM EDT

From
<https://register.nvidia.com/flow/nvidia/gtcspring2023/attendeeportal/page/sessioncatalog/session/1666638995788001G5Zm>

Billions of transactions, very large nodes and edges. Building graph data difficult. pipeline and GPU processing is difficult. Used on top of rule-based fraud detection, for recommendation of suspicious transactions.

BNY Mellon, not yet using GNN. 20% of all global assets. Banker's bank, not much retail activity, credit card activity.

Customers are nodes, but merchants and customer banks are also nodes. Heterogeneous nodes, different networks.

In production, response time has to be milliseconds.

BNY - GNN is component of a system. False positives very bad, biggest challenge. Need to be correct 99% of time. Can't hold up a multi-million dollar transaction for no reason. Use external other data sources to predict fraud.

Bunq - GNN can be whole system or a component. unlock additional value from looking at clustered network. Start as a component, then grow into the single solution.

Mastercard - graph can capture a lot of relationships. Table style rules that captures all the rich relationships in graphs is not there. But when orthogonal information (rule), the accuaracy improves. But I think traditional rule based algorithms can be surpasses with just GNN when all the rich information in it can be captured and used.

Jay - customers, it's a journey. Complicated process. Centrality - how important this node, clustering to fraud node, ... improved by 50%. Don't need full graph, but a few graph features does most of the job. GNN is just boosting the semantic information into ML models.

Ali - get the data right. importing right structure.
Jay - GNN is real but hard. Start evolving.
Jason - lots to learn with GNNs. growing trend among institutions on data sharing re fraud. GNN will help with more data from other institutions.

speech AI

Monday, March 20, 2023 10:55 AM

- **Unlocking the Power of Speech AI for Exceptional Communication and Collaboration [SE52349]**

  To maintain world-class platforms that meet customer expectations across the globe, progressive companies are adopting highly accurate and performant real-time speech AI – localized to support required languages -- and integrating it within the existing cloud, hybrid, or on-premises systems and applications. We'll see how Avaya is leveraging the capabilities of Riva speech AI technology to transform employee and customer experiences. We will also look at exciting future capabilities being explored "behind the curtain!" This talk is a must-attend for businesses looking to maximize the potential of AI Speech technology to accelerate their business goals.

  Stephen Brock, Avaya Marketing DIrector, Avaya

  Industry: Telecommunications

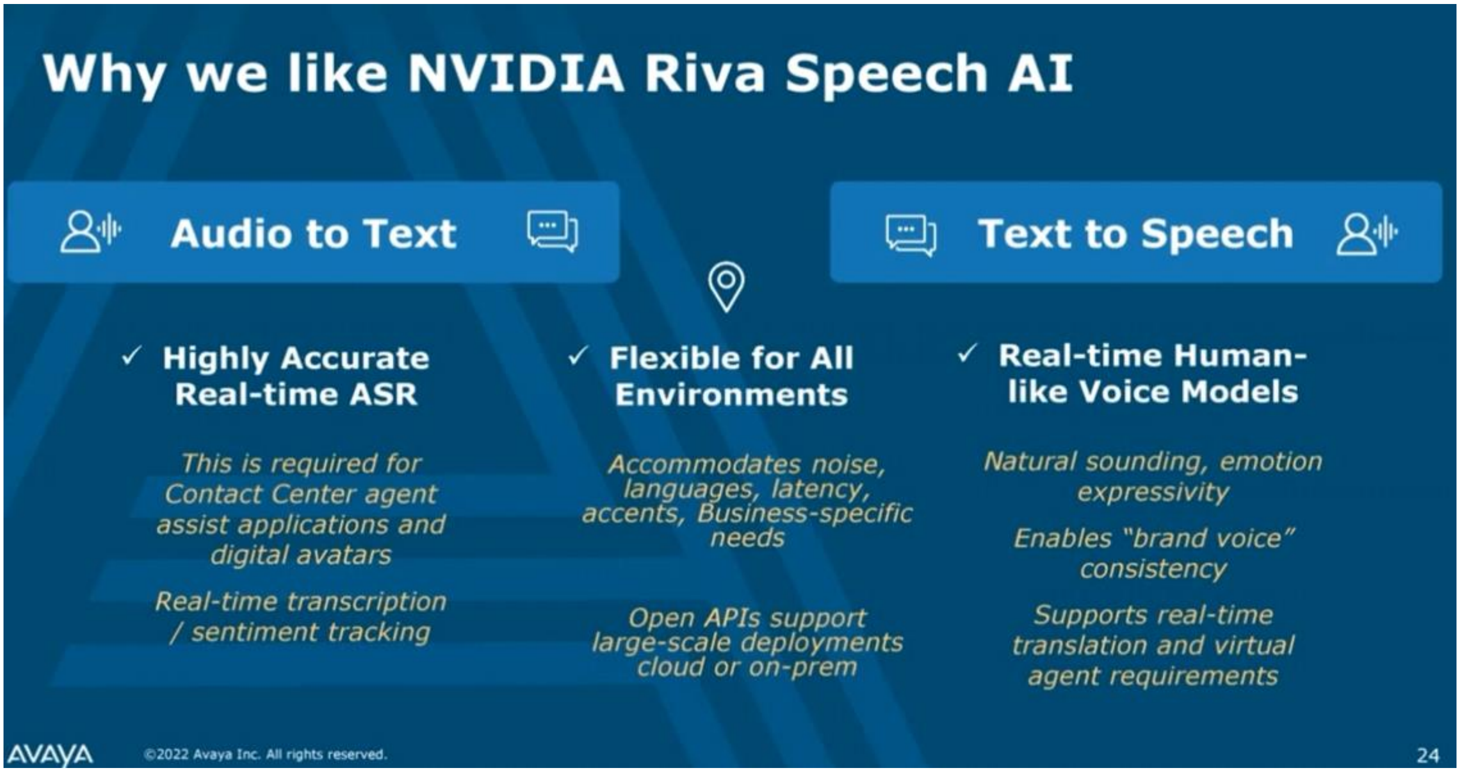
  Topic: Conversational AI / NLP

- **Scheduled**

  Monday, March 2011:00 am - 11:30 am EDT

From <https://register.nvidia.com/flow/nvidia/gtcspring2023/attendeeportal/page/schedule>

blog - Effortless, Steve Brock.
blog - Next 'Verse" to shape the customer experience? The YOUniverse, of course. Steve Brock.



Why we like NVIDIA Riva Speech AI

Audio to Text

Text to Speech

✓ **Highly Accurate Real-time ASR**

This is required for Contact Center agent assist applications and digital avatars

Real-time transcription / sentiment tracking

✓ **Flexible for All Environments**

Accommodates noise, languages, latency, accents, Business-specific needs

Open APIs support large-scale deployments cloud or on-prem

✓ **Real-time Human-like Voice Models**

Natural sounding, emotion expressivity

Enables "brand voice" consistency

Supports real-time translation and virtual agent requirements

AVAYA ©2022 Avaya Inc. All rights reserved. 24

Mostly marketing.
Building AI talkers for customer service, personal assistant (Siri, Alexa).
Change gender, accent, language AI model, remove background noise.
90% of users demand human customer rep option, to not go through AI speaker.

25% deman AI option??? Seems fake. Probably they don't want to wait extremely long time, if no human or AI is available. No one PREFERS to have AI option, at least until their ability to solver user's problem becomes higher than a Human's ability.

AI high-def maps (urban)

Monday, March 20, 2023 11:23 AM

- **Building AI-Based HD Maps for Autonomous Vehicles [SE50001]**
  As autonomous vehicles scale into urban areas, the need for HD map coverage dramatically increases. This session discusses how HD map companies can use AI software to accelerate the development and deployment of an automated map creation pipeline.
- Rambo Jacoby, Principal Engineer, NVIDIA
  Vijay Chintalapudi, Engineering Manager, NVIDIA
  Michael Boone, Product Marketing Manager, Autonomous Vehicles & Computer Vision, NVIDIA
- Industry: Automotive / Transportation
  Topic: Autonomous Vehicles
- **Scheduled**
  Monday, March 2011:30 am - 12:00 pm EDT

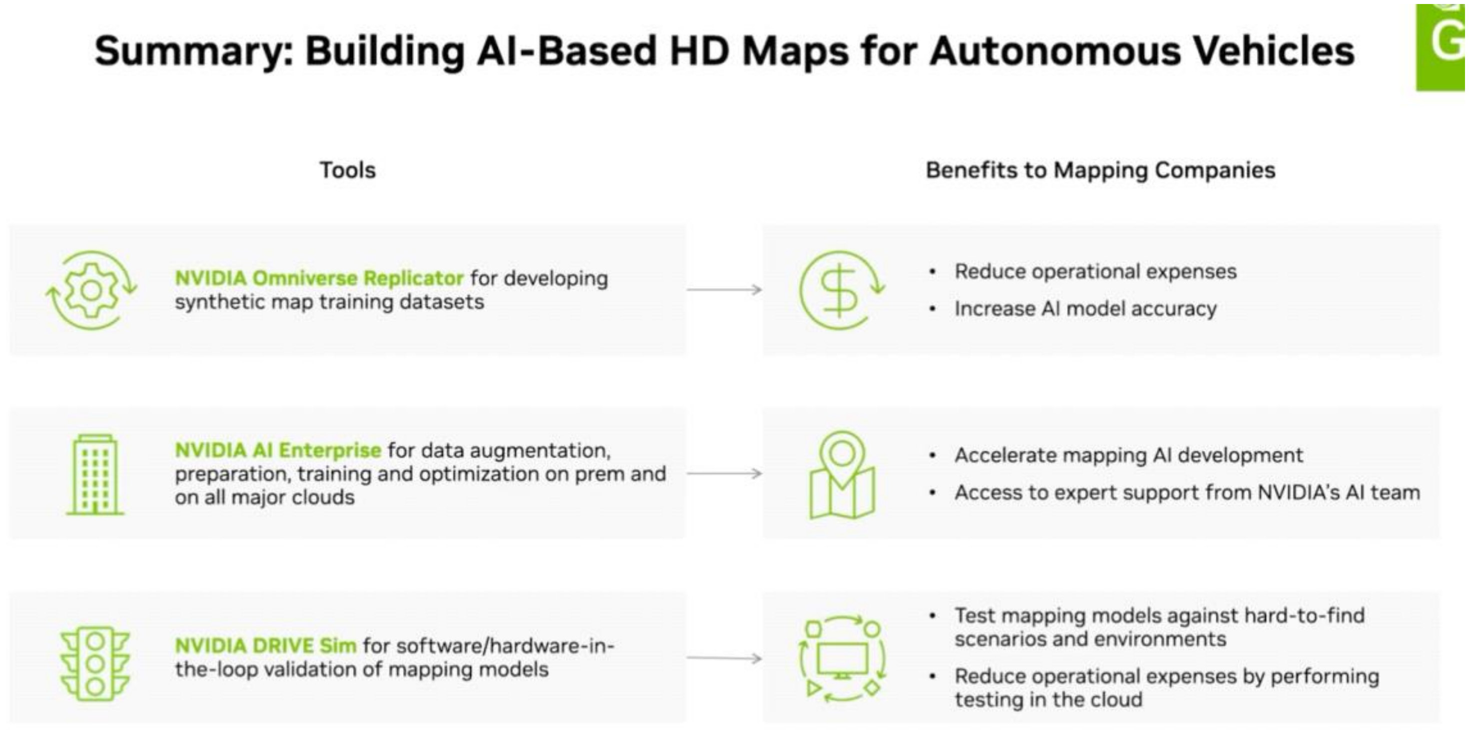
From <https://register.nvidia.com/flow/nvidia/gtcspring2023/attendeeportal/page/schedule?search.daytime=20230323t09>

* Training Data - ground truth maps. Distance between objects, signs, signal lights.

* 2023 - 10% scenes synthetic generated. Capture variations in weather, day/night, dusk/dawn, lighting.
  Improves training accuracy for edge cases, does much better for small sample conditions.
  Combine with drive path information, sign information, sensor information -- to lead to correct reading of driving conditions.

PyTorch, CV-CUDA
blur - personally identifiable info, faces, license plates.
Pad stack & resize (color channel, objects, image size)
Denoise - improve image quality in faint light. Pick up features lane marking, road signs.

Drive Sim an simulate 10,000 miles of driving in 8 hours.
Simulate edge cases that will be harmful to your customer.
Gather much more edge cases, and infinite variation in road conditions via augmented, photo realistic, driving simulation (video).



Summary: Building AI-Based HD Maps for Autonomous Vehicles

Tools	Benefits to Mapping Companies
NVIDIA Omniverse Replicator for developing synthetic map training datasets	• Reduce operational expenses • Increase AI model accuracy
NVIDIA AI Enterprise for data augmentation, preparation, training and optimization on prem and on all major clouds	• Accelerate mapping AI development • Access to expert support from NVIDIA's AI team
NVIDIA DRIVE Sim for software/hardware-in-the-loop validation of mapping models	• Test mapping models against hard-to-find scenarios and environments • Reduce operational expenses by performing testing in the cloud

graduates fast talks

Monday, March 20, 2023 11:55 AM

- **Graduate Fellowship Fast Forward Talks [S51990]**
  Join a special presentation from our 2022-23 Graduate Fellowship recipients to learn "what's next" from the world of research and academia. Sponsored projects involve a variety of technical challenges, including deep learning, robotics, computer vision, computer graphics, architecture, circuits, high performance computing, life sciences, and programming systems. We believe that these minds lead the future in our industry and we're proud to support the 2022-23 NVIDIA Graduate Fellows. We'll also announce the 2023-24 Graduate Fellows at this session. For more information on the NVIDIA Graduate Fellowship program, visit www.nvidia.com/en-us/research/graduate-fellowships
  Bill Dally, Chief Scientist and Senior Vice President of Research, NVIDIA
  Sylvia Chanak, NVIDIA

  - Davis Rempe, Ph.D. Student, Stanford University
  - Hao Chen, Ph.D. Student, University of Texas, Austin
  - Mohit Shridhar, Ph.D. Student, University of Washington
  - Sai Praveen Bangaru, PhD student, MIT
  - Shlomi Steinberg, PhD student , UCSB
  - Sneha Goenka, Ph.D. Student , Stanford University
  - Yufei Ye, PhD Student , CMU - hand grasping, hand size, simulation to objects, bowl, cup, phone, teddy bear.
  - Yuke Wang, PhD Candidate, UCSB - graph NN
  - Yuntian Deng, PhD Candidate, Harvard - Markup-to-image diffusion models.
  - Zekun Hao, Ph.D. Student , Cornell University - Implicit Neural Representations. image generation.

  Industry: All Industries

  Topic: Computer Vision - Research

- **Scheduled**
  Monday, March 2012:00 pm - 1:00 pm EDT

From <https://register.nvidia.com/flow/nvidia/gtcspring2023/attendeeportal/page/schedule>

Chris Newburn

Monday, March 20, 2023 12:02 PM

[Accelerating Data Movement Between GPUs and Storage or Memory [S51142]]

In the past few years, there's been great progress in expediting data movement and access from storage to GPUs. Customers have experienced significant performance improvements by using GPUDirect™ Storage technology in a variety of frameworks and applications focused on visualization, data analytics, AI, and machine learning across multiple industries including health & life sciences and oil & gas. There are also exciting new directions for "just getting data" whether it's in storage or memory, initiating storage transfers from the GPU, and for objects and keys, not just files. We'll outline a vision for transferring and accessing data and share proof-of-concept results that are engaging to framework users and developers. We feature early prototype work on initiating a large volume of small transfers from the GPU, and its applicability to graph machine learning. Da Zheng, leader of Amazon’s forthcoming GraphStorm offering, and CJ Newburn, distinguished engineer at NVIDIA, will co-present their usage requirements and discuss the applicability of NVIDIA’s prototype to their work.

CJ Newburn, Distinguished Engineer, NVIDIA

Harish Arora, Magnum IO Product Manager, NVIDIA

Da Zheng, Senior Applied Scientist, AWS

Industry: All Industries
Topic: Accelerated Computing & Dev Tools - Programming Languages / Compilers

- **Add to Schedule**
  Wednesday, Mar 22 4:00 PM - 4:50 PM EDT

From <https://register.nvidia.com/flow/nvidia/gtcspring2023/attendeeportal/page/sessioncatalog?tab.catalogallsessionstab=16566177511100015Kus&search=Newburn>

[Blending Accelerated Programming Models in the Face of Increasing Hardware Diversity [S51215]]

Choosing a programming model for accelerated computing applications depends on a wide range of factors, which weigh differently across application domains, institutions, and even countries. Why does one application use standard programming languages like C++, while another uses embedded programming models like Kokkos or directives such as OpenACC, and yet another directly programs in vendor-specific languages like CUDA or HIP? We'll compare the various choices and share hands-on experience from developers in different countries and fields of expertise. We’ll explore both technical and nontechnical reasons for how the various approaches are mixed. Join us for a fun and insightful session!

CJ Newburn, Distinguished Engineer, NVIDIA

Tom Deakin, Lecturer in Advanced Computer Systems, University of Bristol

Fernanda Foertter, Director, Voltron

Torsten Hoefler, Professor, ETH Zurich

Somnath Roy, Associate Professor Mechanical Engineering Centre for Computational and Data Sciences, IIT Kharagpur

Christian Trott, Principal Member of Technical Staff, Sandia National Labs

Industry: All Industries
Topic: Accelerated Computing & Dev Tools - Programming Languages / Compilers

- **Scheduled**
  Thursday, Mar 23 2:00 PM - 2:50 PM EDT

From <https://register.nvidia.com/flow/nvidia/gtcspring2023/attendeeportal/page/sessioncatalog?tab.catalogallsessionstab=16566177511100015Kus&search=Newburn>

- Status is reachable
  Chris (CJ) Newburn

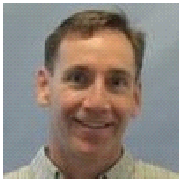
  1st degree connection· 1st
  Architect of NVIDIA HPC software strategy, roadmap, and developer engagement; Data center and zero trust architect
- DEC 28, 2010Jennifer Yoon sent the following messages at 12:54 AM
  View Jennifer's profile

Jennifer Yoon 12:54 AM
Hello Chris, Thanks for contacting me. Did you know that Jason turned his personality around 180 degrees when he turned 30 and quit computers and is into new age spiritual healing? Well, he still builds a 2-seater plane from his garage so he's not totally out of engineering, but... He is happy in Pt Reyes, CA. You might give him a shout out jasonjangho@yahoo.com. I am living in HHMI research facility for brain science. My husband Bill Katz is the scientist and computer programmer. I went to Chicago for MBA and am into finance & securities. Bill and I were in Silicon Valley working on startups, esp with Google App Engine and Ruby on Rails, maybe you've heard of them.. Came back to Virginia when we ran through all of our savings. Love it here. I'm still job hunting. Hope to find something in 2011. Dad & Mom retired. Goes line dancing and Dad is learning Chinese. He seems happy, to my surprise. I didn't think he will do retirement well. It was forced onto him due to major changes in immigration law after 9-11. I'll be happy to give a tour of Janelia Research Center whenever you're nearby Ashburn, VA. What's you home address? I'll send you a card. J

- View Jennifer's profile

Jennifer Yoon 1:14 AM
Dear Chris (CJ), I've written this recommendation of your work to share with other LinkedIn users. Details of the Recommendation: "Excellent colleague to work with, always cheerful and courteous, and a very smart computer scientist. I always wondered what a smart guy like him was doing at such a fly-by-night startup. The others were direct imports from Moscow and Beijing who were in need of a Green Card, so I could understand about them :-) Seriously, we had top talent working on an impossible machine. I hope he enjoyed the wild ride and many, many extremes in pushing the boundary of reality. I think we were 20 years ahead of the curve."(Edited)

- JAN 13, 2011Chris (CJ) Newburn sent the following messages at 10:27 PM
  View Chris (CJ)'s profile



Chris (CJ) Newburn 10:27 PM
Thanks for the news, Jennifer. Glad to hear that your family is doing well. My address is 8034 Renee Dr. South Beloit IL 61080. Haven't been to VA is many years. Best of luck with job search - that can get

difficult. Have a happy and prosperous 2011! CJ

From <https://www.linkedin.com/feed/>

GPU in Financial Apps

Monday, March 20, 2023 1:03 PM

- **Using NVIDIA GPUs in Financial Applications: Not Just for Machine Learning Applications [S52211]**

  Deploying GPUs to accelerate applications in the financial service industry has been widely accepted and the trend is growing rapidly, driven in large part by the increasing uptake of machine learning techniques. However, banks have been using NVIDIA GPUs for traditional risk calculations [large matrix correlation, sigma volatility calc, pricing pdf calc.] for much longer, and these workloads present some challenges due to their multi-tenancy requirements. We'll explore the use of multiple GPUs on virtualized servers leveraging NVIDIA AI Enterprise to accelerate an application that uses Monte Carlo techniques for risk/pricing application in a large international bank. We'll explore various combinations of the virtualized application on VMware to show how NVIDIA AI Enterprise software runs this application faster. We'll also discuss process scheduling on the GPUs and explain interesting performance comparisons using different VM configs. We'll also detail best practices for application deployments.

- Manvender Rawat, Senior Manager, Product Management, NVIDIA
  Justin Murray, Technical Marketing Architect, VMware
  Richard Hayden, Executive Director and Head of the QR Analytics Team, JP Morgan Chase
  Industry: Financial Services
  Topic: Data Center / Cloud Infrastructure - Technical

- **Scheduled**
  Monday, March 201:00 pm - 2:00 pm EDT

From <https://register.nvidia.com/flow/nvidia/gtcspring2023/attendeeportal/page/schedule>

VMWare vSphere, virtual GPU servers. On-Demand Kubernetes Clusters and Virtual Machines.
Intrinsic Security and Lifecycle Management
Live Migration and Load Balancing.

HPC Applications - atomic calcs, not dependent on other units, GPU capacity highly parallelized.

JP Morgan used GPUs for 10 years+ (2011) to calculate pricing, risk, volatility, interest rate products, FX, credit risk.
Speed ups 50X, 100X.
GPU used in quant library.
 * mostly C++, some Python, CUDA.
 * Monte Carlo
 * Partial Diff eq
 * Some calibration

A few varieties of GPUs
 * Hand-written CUDA kernels
 * Thrust (custom language, pricing inst)

* Auto-generated CUDA kernels.

GPUs used for speedup, massive parallelizm, net cost savings vs CPUs. Overnight run of risks SLA report (sensitivity loss report?)

Calculate greeks for an exotic instrument - GPU fractional uses (multi-tenancy).
Only 10% of calculation time is on GPU. Batch process 10 instruments so GPU can be used for 100% of time. Cost effective this way. Throughput time much faster.
Can't get 100% GPU usage, too much CUDA API/driver calls, memory calls, context switches etc.
Fractional multi-GPU usage.

Refactor code to combine many small memory operations into few large memory operations. Reduce context switching.
Next generation A100 Ampere GPU, even denser GPU, bottlenecks even worse. Need double precision floating point, can't go to single, so need A100. Only 1 and 2 GPUs provide benefit, throughput is worse with 3 and 4 GPUs.
Hosting costs makes it uneconomical if only 1 or 2 GPUs can be used per blade (server).
Initially, really bad performace, 42% loss on 4 GPUs.
Got much better performance with 2 VMs (virtual GPUs) with 2GPU each.
Even better with 4 VMs with 1GPU each. (CUDA mps scaling formula.)

Eventually will hit a limit where faster GPUs with more cores will translate into faster processing of risk engines/calculations. Not enough CPU cores to keep GPU cores busy.

Hassabis Deep Mind

Tuesday, March 21, 2023 1:08 PM

Using AI to Accelerate Scientific Discovery [S51831]

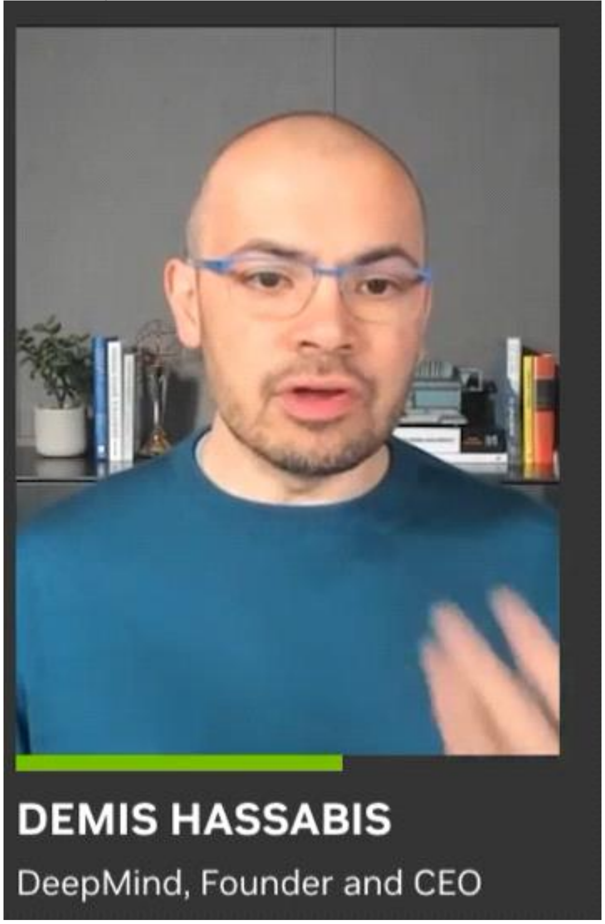
The past decade has seen incredible advances in artificial intelligence. DeepMind has been in the vanguard of many of these big breakthroughs, pioneering the development of self-learning systems like AlphaGo, the first program to beat the world champion at the complex game of Go. Games have proven to be a great training ground for developing and testing AI algorithms, but the aim at DeepMind has always been to build general learning systems ultimately capable of solving important problems in the real world. We're on the cusp of an exciting new era in science, with AI poised to be a powerful tool for accelerating scientific discovery itself. We recently demonstrated this potential with our AlphaFold system, a solution to the 50-year grand challenge of protein structure prediction, culminating in the release of the most accurate and complete picture of the human proteome and release of the predicted structures of over 200 million proteins — nearly all catalogued proteins known to science.

Demis Hassabis, Founder and CEO, DeepMind

Industry: Academia / Higher Education
Topic: Deep Learning - Frameworks

- **Scheduled**

Tuesday, Mar 211:00 PM - 1:55 PM EDT



From
<https://register.nvidia.com/flow/nvidia/gtcspring2023/attendeeportal/page/sessioncatalog/session/1666649457197001tw0E
>

Moods AI dreams

Tuesday, March 21, 2023 1:53 PM

- **The Indefinable Moods of Artificial Intelligence [S51838]**

  Learn about the direct correlation of human dreams to the evolution of drawing, painting, and animation, and how AI reflects the collective unconscious, driving new forms of art creation, structure, and narrative forms. We'll look at how AI is reflecting the dream world — our hopes, fears, desires, biases, and human frailty. How can we better use AI for good? How can it enhance all humans to return to creative practice and not separate art from everyday life? NVIDIA RTX GPUs are widely used at the University of Southern California's School of Cinematic Arts and the new Expanded Animation - XA Program focuses on animation for AI, virtual characters, and robotics. Through USC's XA Program we'll showcase how NVIDIA Omniverse, NVIDIA Canvas, GauGAN2, and NeRFs are being used to teach, experiment, and express dream imagery, inspiring new forms of image making and creative process.

  Kathy Smith, Artist and Professor, School of Cinematic Arts, University of Southern California

  Industry: Academia / Higher Education

  Topic: Graphics - AI Applications, Art

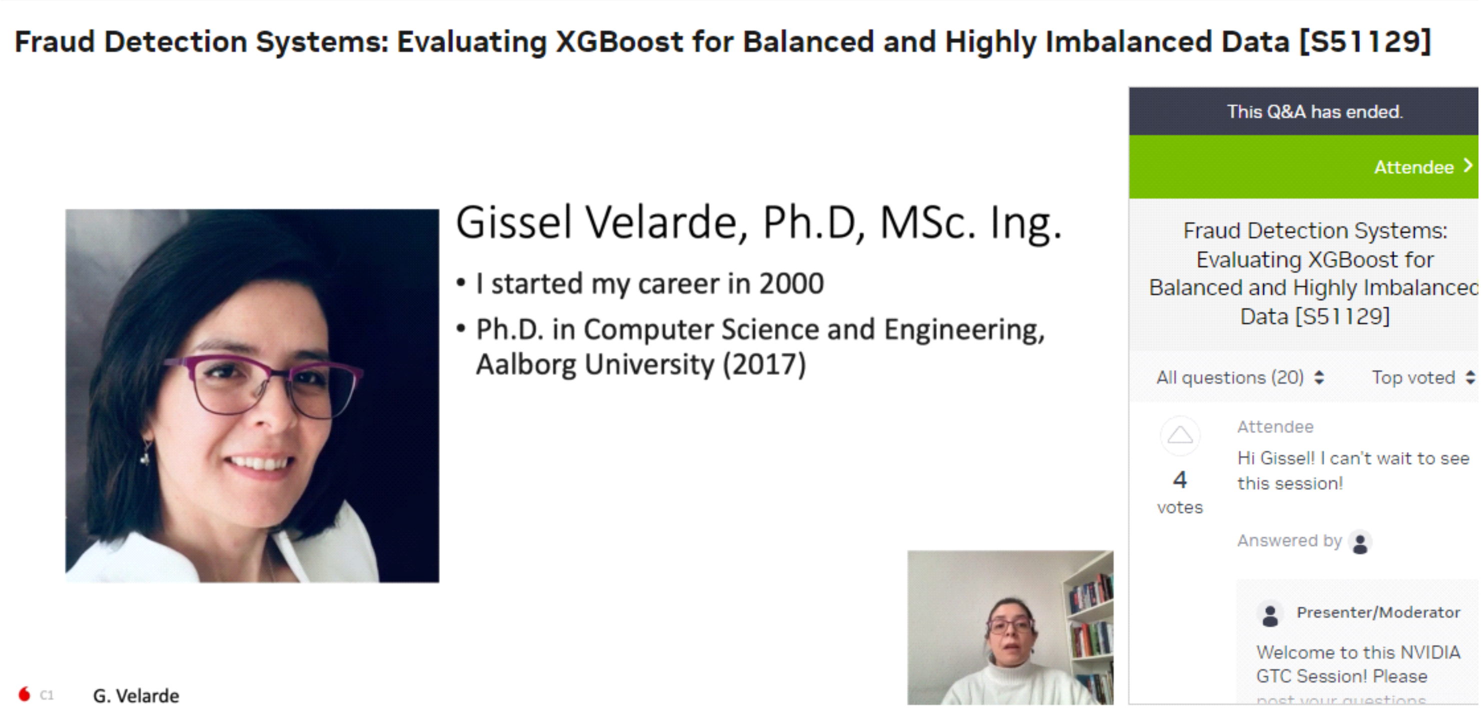
- **Scheduled**

  Tuesday, March 212:00 pm - 2:55 pm EDT

From <https://register.nvidia.com/flow/nvidia/gtcspring2023/attendeeportal/page/schedule>

Fraud Detection XGBoost

Thursday, March 23, 2023 12:53 PM



Fraud Detection Systems: Evaluating XGBoost for Balanced and Highly Imbalanced Data [S51129]

Gissel Velarde, Ph.D, MSc. Ing.

- I started my career in 2000
- Ph.D. in Computer Science and Engineering, Aalborg University (2017)

C1 G. Velarde

This Q&A has ended.

Attendee

Fraud Detection Systems: Evaluating XGBoost for Balanced and Highly Imbalanced Data [S51129]

All questions (20) Top voted

4 votes

Attendee

Hi Gissel! I can't wait to see this session!

Answered by

Presenter/Moderator

Welcome to this NVIDIA GTC Session! Please post your questions

Fraud Detection Systems: Evaluating XGBoost for Balanced and Highly Imbalanced Data [S51129]

From <https://register.nvidia.com/flow/nvidia/gtcspring2023/attendeeportal/page/sessioncatalog/session/1665741940906001a0N9>

Fraud detection is a challenging task since fraudsters continuously change their behavior, they may represent rare cases, and fraud patterns may even be unseen during training. Expert human inspectors develop an intuition of possible fraudulent cases. Still, they need an accurate and automated process to help them spot fraud out of thousands of daily samples. Fraud detection is treated as a binary classification problem where the positive class is of extreme interest and may be highly underrepresented in the data. In this talk, we evaluate XGBoost on balanced and imbalanced data, and present key points that helped us improve our models in terms of detection precision, recall, and training speed. We will explain the concepts behind our choices and present code examples in well-known python libraries. Finally, we will reflect on lessons learned from our experience using XGBoost for fraud detection, considering that the types of mistakes a binary classifier makes have different impacts on different applications.

Gissel Velarde, Senior Expert Data Scientist, Vodafone GmbH.

Industry: Telecommunications

Topic: Cybersecurity / Fraud Detection

- S51129 - Fraud Detection Systems_ Evaluating XGBoost for Balanced and Highly Imbalanced Data.pdf
- [White Paper] S51129 - Velarde et al 2023 Evaluating XGBoost for Balanced and Imbalanced Data.pdf

From <https://register.nvidia.com/flow/nvidia/gtcspring2023/attendeeportal/page/sessioncatalog/session/1665741940906001a0N9>

PDF

S51129 - Fraud Det...

PDF

[White Paper] S5...

[White Paper] S5...

geospatial GPU-Accel

Thursday, March 23, 2023 3:18 PM

- **GPU-Accelerating End-to-End Geospatial Workflows [S51796]**

  Both the federal community and the commercial marketplace have critical mission needs to rapidly geolocate imagery that has no associated geospatial information for a wide variety of computer vision applications, such as search and rescue, natural hazards detection, and environmental monitoring. These tasks are extremely difficult due to the computational and time requirements — i.e., being able to process a large landmass spanning hundreds of thousands of square miles in minutes, rather than hours or days. We'll explain how NVIDIA's AI data processing and GPU acceleration software libraries (Triton Inference Server, DALI, and RAPIDS) are incorporated in an end-to-end geospatial application workflow to implement an extremely fast geospatial search-and-retrieval application.

  Kevin Green (WFO), Senior Solutions Architect, NVIDIA

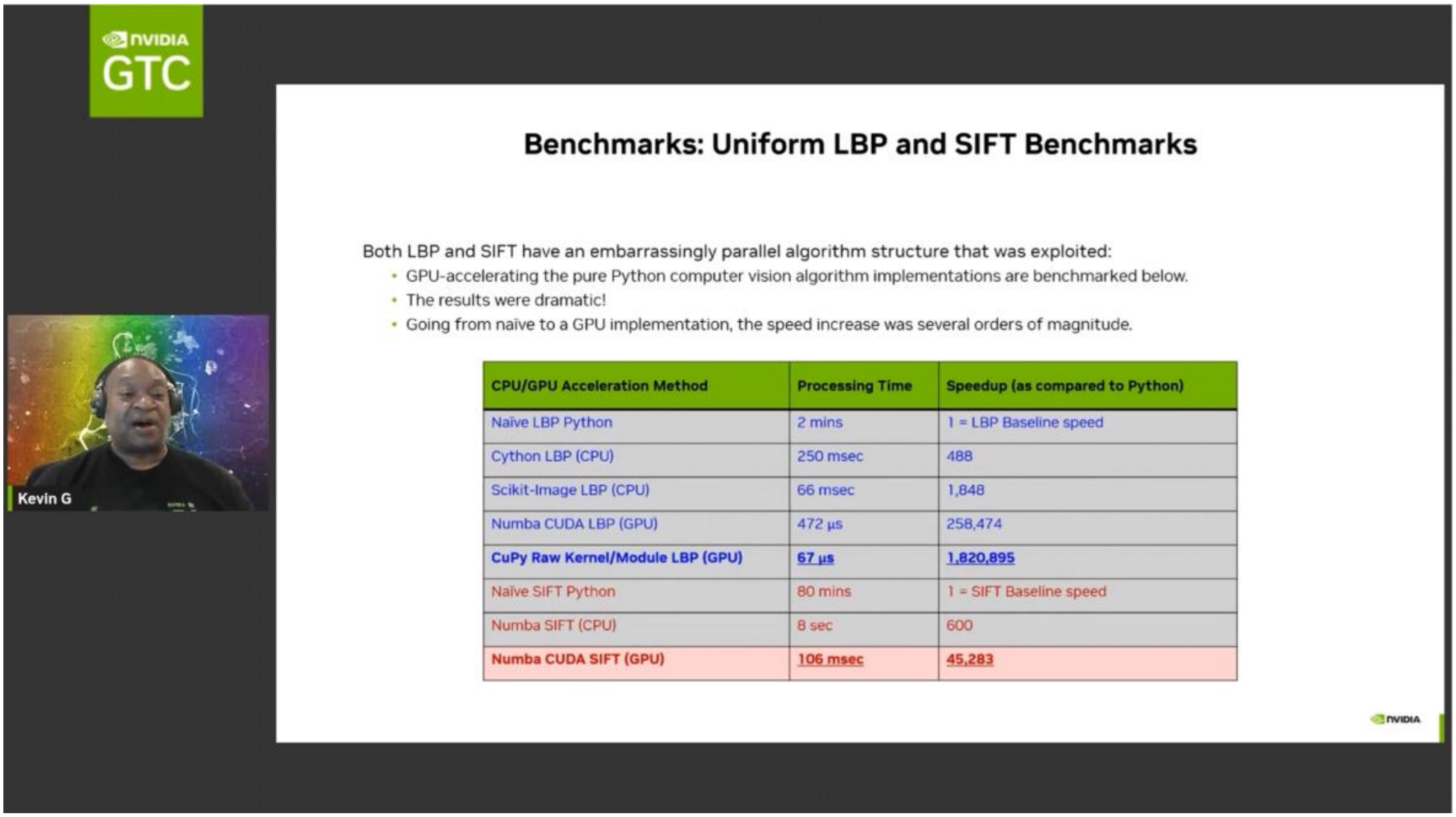
  Industry: Public Sector

  Topic: Data Science

- **Scheduled**

  Thursday, March 233:00 pm - 3:50 pm EDT

From <https://register.nvidia.com/flow/nvidia/gtcspring2023/attendeeportal/page/schedule>

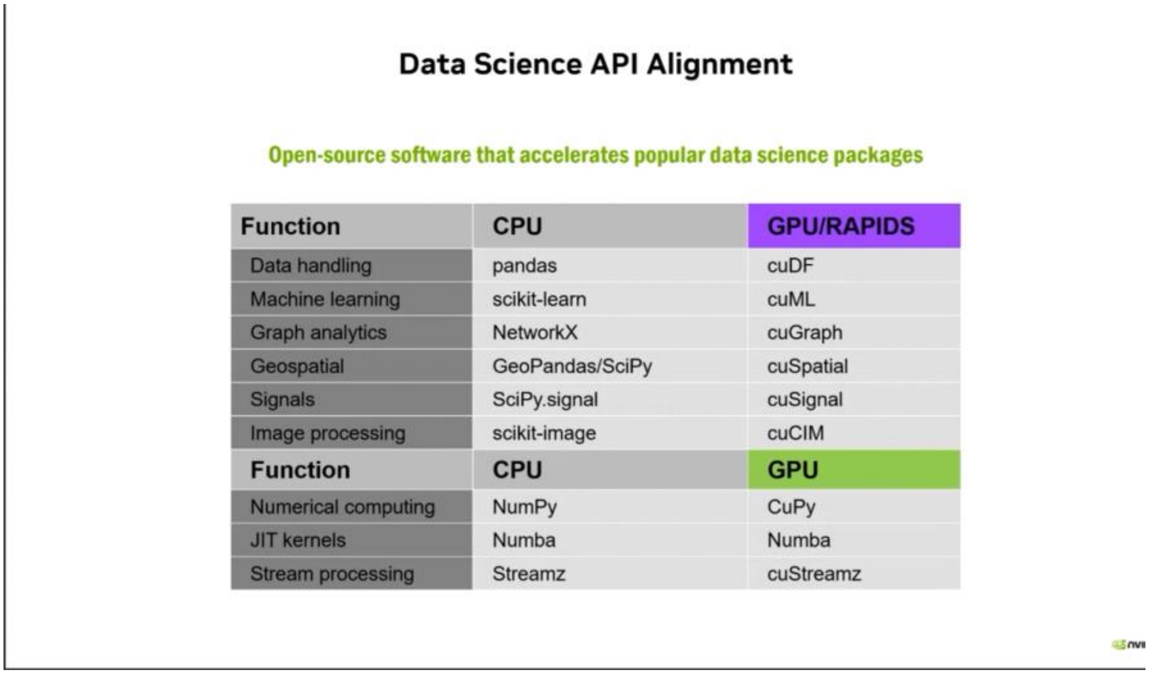


Benchmarks: Uniform LBP and SIFT Benchmarks

Both LBP and SIFT have an embarrassingly parallel algorithm structure that was exploited:

- GPU-accelerating the pure Python computer vision algorithm implementations are benchmarked below.
- The results were dramatic!
- Going from naïve to a GPU implementation, the speed increase was several orders of magnitude.

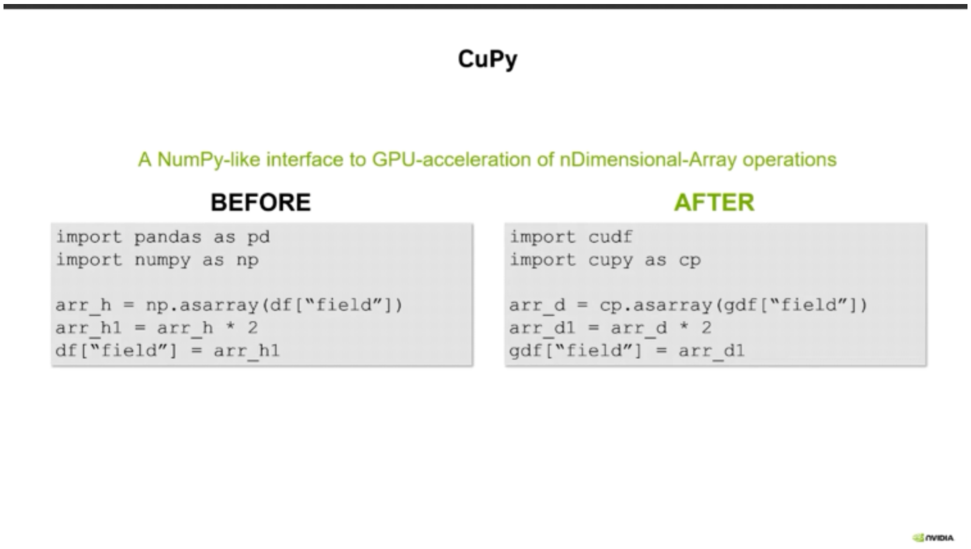
CPU/GPU Acceleration Method	Processing Time	Speedup (as compared to Python)
Naïve LBP Python	2 mins	1 = LBP Baseline speed
Cython LBP (CPU)	250 msec	488
Scikit-Image LBP (CPU)	66 msec	1,848
Numba CUDA LBP (GPU)	472 µs	258,474
CuPy Raw Kernel/Module LBP (GPU)	**67 µs**	**1,820,895**
Naïve SIFT Python	80 mins	1 = SIFT Baseline speed
Numba SIFT (CPU)	8 sec	600
Numba CUDA SIFT (GPU)	**106 msec**	**45,283**



Data Science API Alignment

Open-source software that accelerates popular data science packages

Function	CPU	GPU/RAPIDS
Data handling	pandas	cuDF
Machine learning	scikit-learn	cuML
Graph analytics	NetworkX	cuGraph
Geospatial	GeoPandas/SciPy	cuSpatial
Signals	SciPy.signal	cuSignal
Image processing	scikit-image	cuCIM
Function	**CPU**	**GPU**
Numerical computing	NumPy	CuPy
JIT kernels	Numba	Numba
Stream processing	Streamz	cuStreamz



CuPy

A NumPy-like interface to GPU-acceleration of nDimensional-Array operations

BEFORE

```
import pandas as pd
import numpy as np

arr_h = np.asarray(df[“field”])
arr_h1 = arr_h * 2
df[“field”] = arr_h1
```

AFTER

```
import cudf
import cupy as cp

arr_d = cp.asarray(gdf[“field”])
arr_d1 = arr_d * 2
gdf[“field”] = arr_d1
```

Very easy drop-in cuXX code for standad python library.
pandas as pd -> cudf
numpy as np -> cupy as cp