ONNX Community Day - June 24

Event details

This event was hosted on June 24, 2022 in-person at the brand-new Microsoft Silicon Valley Campus, as well as online on Teams and a live YouTube stream.

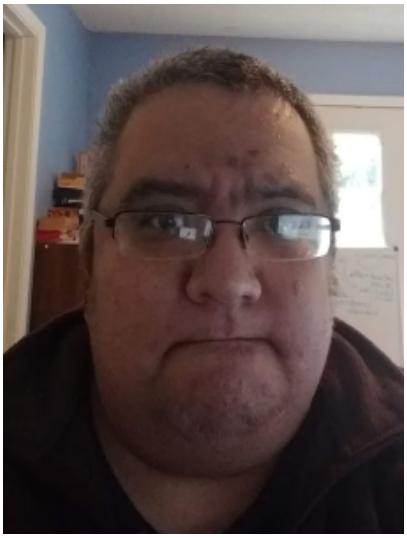
The main conference track included lightning talks showcasing the innovative ways ONNX is being used across the ecosystem. Additionally, the in-person event included a roundtables portion to continue the conversation, as well as an ONNX Runtime-hosted happy hour to relax and connect with the community.

Schedule

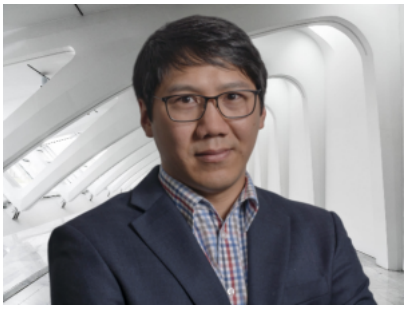
Talk	Presenter	Content
ONNX Steering Committee Update	Prasanth Pulavarthi, Microsoft Alexandre Eichenberger, IBM Rajeev Nalawadi, Intel Mayank Kaushik, NVIDIA Andreas Fehlner, TRUMPF Laser GmbH	Slides, Recording
ONNX SIG: Arch & Infra	Liqun Fu, Microsoft	Slides, Recording
ONNX SIG: ONNX Operators	Ganesan “Rama” Ramalingam, Microsoft	Slides, Recording
ONNX SIG: Converters	Kevin Chen, NVIDIA	Slides, Recording

ONNX SIG: Model & Tutorials	Jacky Chen, Microsoft	Slides, Recording
ONNX WG: Pre-processing	Joaquin Anton, NVIDIA	Slides, Recording
Designed to be Optimized ONNX Runtime allows us to export our models for different platforms and systems. Optimization, however, starts with design. Leaving it when everything seems ready to go a production could make effective optimization impossible. We will see the frequent mistakes and how to avoid them.	Mauro Bennici, GhostWriter.AI  *Mauro Bennici is the CTO and co-founder of "You Are My Guide" - "GhostWriterAI". He is active in the R&D area of Natural Language Processing (NLP) and Natural Language Generation (NLG). He is a Data Scientist, Professional Scrum Master (PSM), Microsoft Certified Trainer (MCT), and .NET foundation member. Member of the SME Focus Group on Artificial Intelligence, European Community. The main work area is to understand the text in any form to anticipate the author's intentions: this can be applied to create articles, advertising campaigns, customer care, speech analysis, churn prevention, etc. He is the author of research papers on understanding the text in Italian and English. He founded the Torino.NET meetup, and he is a speaker for Codemotion. Mentor for Techstars, facilitator for IamRemarkable.*	Slides, Recording

INT8 Inference of Quantization-Aware trained models using ONNX-TensorRT Accelerating Deep Neural Networks (DNN) inference is an important step in realizing latency-critical deployment of real-world applications such as image classification, image segmentation, natural language processing, etc. The need to improve DNN's inference latency has sparked interest in running those models in lower precisions, such as FP16 and INT8. In particular, running DNNs in INT8 precision can offer faster inference and a much lower memory footprint than its floating-point counterpart. NVIDIA TensorRT supports Quantization-Aware Training (QAT) techniques to convert floating-point DNN models to INT8 precision. In this talk, we shall demonstrate end-end workflow of converting Tensorflow QAT models into ONNX, which is a standard intermediate representation to deploy using TensorRT. We use TF2ONNX package to convert a quantized Tensorflow model into ONNX. ONNX format makes it easier to visualize graphs via netron which can provide users information about placement of quantized nodes.	Dheeraj Peri, NVIDIA  *Dheeraj Peri works as a deep learning software engineer at NVIDIA. Before that, he was a graduate student at Rochester Institute of Technology in New York, working on deep learning-based approaches for content retrieval and handwriting recognition tasks. Dheeraj's research interests include information retrieval, image generation, and adversarial machine learning. He received a bachelor's degree from Birla Institute of Technology and Sciences, Pilani, India.*	Slides, Recording
QONNX: A proposal for representing arbitrary-precision quantized NNs in ONNX We present extensions to the Open Neural Network Exchange (ONNX) intermediate representation format to represent arbitrary-precision quantized neural networks. We first introduce support for low precision quantization in existing ONNX-based quantization formats by leveraging integer clipping, resulting in two new backward-compatible variants: the quantized operator format with clipping and quantize-clip-dequantize (QCDQ) format. We then introduce a novel higher-level ONNX format called quantized ONNX (QONNX) that introduces three new operators —Quant, BipolarQuant, and Trunc— in order to represent uniform quantization. By keeping the QONNX IR high-level and flexible, we enable targeting a wider variety of platforms. We also present utilities for working with QONNX, as well as examples of its usage in the FINN and hls4ml toolchains. Finally, we introduce the QONNX model zoo to share low precision quantized neural networks.	Alessandro Pappalardo, AMD  *Alessandro is a Member of the Technical Staff at AMD AECG Research. His work focuses on fast inference algorithms, neural network HW/SW co-design and quantization across CPUs, GPUs, FPGAs and AI engines.*	Slides, Recording

How to reconcile AI and privacy AI is revolutionising many fields from healthcare to biometrics these recent years. However due to security and privacy concerns, data is still being siloed and not shared enough due to the fear of data exposure and IP leakage. Confidential Computing is a recent technology that enables end-to-end encryption when analysing sensitive data. By leveraging Confidential Computing, data owners can share their data to AI companies, for instance to train or consume an AI model, without ever risking their data being stolen, leaked or used for any other purpose, as data remains protected even when shared to third parties. This talk aims to introduce the high level principles of Confidential Computing and how it can be used to deploy privacy friendly AI models. We will present BlindAI (https://github.com/mithril-security/blindai), an AI deployment solution, serving ONNX models with privacy guarantees, and see how it can be used to unlock confidential medical document analysis in the Cloud, or facial recognition with privacy guarantees.	Daniel Huynh, Mithril Security  *Daniel Huynh is the CEO of Mithril Security. He is a graduate from Ecole Polytechnique with a specialisation in AI and data science. He worked at Microsoft on Privacy Enhancing Technologies under the office of the CTO of Microsoft France.* *He has written articles on Homomorphic Encryptions with the CKKS explained series (https://blog.openmined.org/ckks-explained-part-1-simple-encoding-and-decoding/). He is now focusing on Confidential Computing at Mithril Security and has written extensive articles on the topic: https://blog.mithrilsecurity.io/.*	Slides, Recording
Responsible AI @ ONNX: Metadata, Model Cards, and Provenance The space of AI is growing rapidly. At this pace, it can be challenging for key AI stakeholders to identify and address social and regulatory concerns with AI, motivating the need for tools and methods to approach AI ethics challenges. A popular approach in the responsible AI space is using metadata to encode a “model card,” a versatile report detailing the configuration, ethical considerations, limitations, and quantitative analysis of an AI model. This approach can be used to enable transparency and fairness of the use case, filtering of high-quality AI models, pain point identification in AI pipelines, and help with establishing compliance and lineage. In this session, we will present our proposal and end-to-end proof of concept for metadata fields and model cards incorporated in ONNX to capture aspects of the model such as provenance & mixed precision representation.	Rodolfo Gabe Esteves, Bhargavi Karumanchi, Ria Cheruvu, Rajeev Nalawadi, Intel  *Rodolfo (Gabe) Esteves has worked for Intel for over ten years, mostly showcasing hardware capabilities to programming languages and developer technologies. In the past few years, this has encompassed Machine Learning technologies, including ONNX. Gabe got his PhD in Computer Science from the University of Waterloo, ON, Canada.*  *Bhargavi is a Software Engineer at Intel working on enabling and optimizing AI workloads on Intel hardware. She holds a Master's degree in Computer Science from Cornell University, she has been with Intel for past 7 years.*	Slides, Recording

	 *Ria Cheruvu is AI Ethics Lead Architect at the Intel Network and Edge engineering group where she leads a team responsible for the productization of trustworthy and explainable AI technologies. She is an emerging speaker in the industry and delivered technical talks for TedX, DEFCON IoT Village, and Women in Data Science communities. Ria has a master's degree in data science from Harvard University, and her pathfinding domains include solutions for AI security, privacy, and fairness, and explainable and responsible AI systems.*	
ONNX and the JVM Integrating machine learning into enterprises requires building and deploying ML models in the environments enterprises build their software in. Frequently this is in Java, or another language running on the JVM. In this talk we'll cover some of our recent work bringing the ONNX ecosystem to Java. We'll discuss uses of ONNX Runtime from Java, and also our work writing model converters from our Java ML library into ONNX format.	Adam Pocock, Oracle  *Adam is an ML researcher in Oracle Labs' Machine Learning Research Group. He's worked on feature selection, scaling up Bayesian inference with GPUs and more recently NLP. He's the lead developer of the Tribuo ML library, maintains the Java API for ONNX Runtime, and co-leads the TensorFlow-Java project.*	Slides, Recording

Build your high-performance model inference solution with DJL and ONNX Runtime In many companies, Java is the primary language for the teams to build up services. To have ONNX model onboard and integration, developers faced several technical challenges on the resource allocation and performance tuning. In this talk, we will walk you through the inference solution built by DJL, a ML library in Java. In the meantime, we will share some customer success stories with model hosting using ONNXRuntime and DJL.	Qing Lan, AWS  *Qing is a Software Development Engineer in AWS. He has been working on several challenging products in Amazon, including high performance ML inference solutions and high performance logging system. Qing's team successfully launched the first Billion-parameter model in Amazon Ads with very low latency required. Qing has in-depth knowledge on the infrastructure optimization and Deep Learning acceleration. Qing is also a PPMC of Apache MXNet.*	Slides, Recording
Billions of NLP Inferences on the JVM using ONNX and DJL This session outlines the recently rolled out Hypefactors' MLOps infrastructure designed for billions NLP inferences a day. The workload serves media intelligence and OSINT use cases. The infrastructure is designed with a Java Virtual Machine-first approach that is enabled by ONNX interop and AWS' Deep Java Library. On top of that, we show how quantization drives further performance optimizations.	Viet Yen Nguyen, Hypefactors  *Viet Yen Nguyen is the CTO of Hypefactors.*	Slides, Recording -
What's New in ONNX Runtime This talk will share highlights of the ONNX Runtime 1.10-1.12 releases, including details on notable performance improvements, features, and platforms including mobile and web.	Ryan Hill, Microsoft  *Ryan Hill has been with the AI Frameworks team for the past 4 years, where he has mostly worked on operator kernels, C APIs, and dynamically loading execution providers. Prior to this he worked on the Office PowerPoint team, where his most widely seen work is many of the slideshow slide transitions. For fun he likes trying to use the latest C++ features and hitting internal compiler errors.*	Slides, Recording

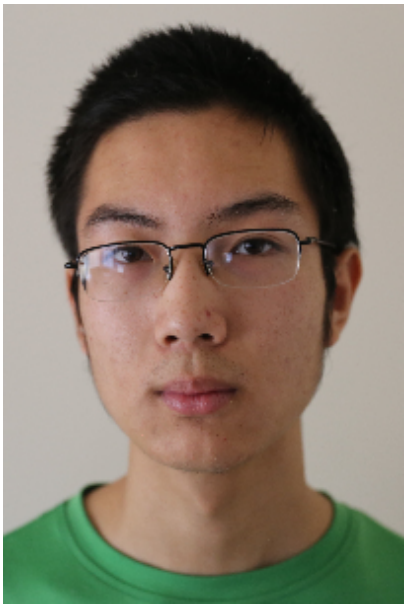
Accelerating Machine Learning with ONNX Runtime and Hugging Face Hugging Face has democratized state of the art machine learning with Transformers and the Hugging Face Hub, but deploying these large and complex models into production with good performance remains a challenge for most organizations. In this talk, Jeff Boudier will talk you through the latest solutions from Hugging Face to deploy models at scale with great performance leveraging ONNX and ONNX Runtime.	Jeff Boudier, Hugging Face  *Jeff Boudier builds products at Hugging Face, creator of Hugging Face Transformers, the leading open-source ML library. Previously Jeff was a co-founder of Stupeflix, acquired by GoPro, where he served as director of Product Management, Product Marketing, Business Development and Corporate Development.*	Slides, Recording
PyTorch-ONNX Converter This session will present an overview of the PyTorch-ONNX converter, its implementation, and recent improvements to support a wide range of models.	Bowen Bao, Microsoft  *Bowen is a software engineer working on the PyTorch-ONNX converter. He's contributed over 400 pull requests to PyTorch since 2018. He's also contributed to ONNX and ONNX Runtime. https://github.com/BowenBao*	Slides, Recording
Detect Safety Zone Violation in Manufacturing with SAS Event Stream Processing and ONNX models This session will present an in-production solution that takes advantage of SAS Event Stream Processing and ONNX runtime to support the detection of safety zone violations using computer vision pre-trained ONNX Model and involving multiple cameras. This solution was deployed at the factory edge with an architecture that, using Kubernetes and Kafka, ensures a reliable and stable environment for productionized computer vision solutions complemented with a cloud-centralized infrastructure to monitor, manage and collect information from multiple factories.	Allen Langlois, Saurabh Mishra, Daniele Cazzari, SAS 	Slides, Recording

Daniele Cazzari is a Global Lead in IoT, Edge and Cloud Analytics Solutions. Daniele brings over 10 years of experience on Internet of Things Edge to Cloud architecture supporting Automotive, Manufacturing, and Insurance customers. He is currently supporting the new SAS-Microsoft partnership as technical lead of cloud-native SAS IoT Solutions aiming to simplify data processing and analysis from edge devices. Before joining SAS, he worked as a Manager in Accenture's Industry X. 0 Capability where he was responsible for Connected Vehicle and Autonomous Driving project delivery. Daniele holds a Master's Degree in Industrial Engineering and Management from Polytechnic of Turin. In his free time, he enjoys sailing, swimming, skiing, and good food!

Allen Langlois is a Principal Software Developer. Allen has worked in IoT from the Edge to the Cloud. He started out on the Edge, configuring small devices and adding SAS Event Stream Processing, demonstrating the power of bringing the compute close to the data. Since then he's migrated his efforts to the Azure Cloud. He's currently focused on DevOps. Before joining SAS, Allen worked as a contractor on real time data acquisition and processing at NASA-Langley and on cloaking Naval ships from magnetic mines as an employee of the US Navy. Allen earned an MS in Electrical Engineering.

Saurabh Mishra

Improving the online shopping experience with ONNX Building and deploying AI solutions to the cloud at scale is complex. With massive datasets and performance considerations - finding a harmonious balance is crucial. This session will outline key learnings from deploying a Serverless application running inference on a sci-kit learn model using ONNX Runtime, and will share how to utilise the capabilities of ONNX runtime to improve the online shopping experience for shoppers and global brands.	Matthew Leyburn, Bazaarvoice  *Matthew Leyburn is a software engineer at Bazaarvoice in Belfast, Northern Ireland. After graduating from Queen's University with a BSc in Computer Science, he joined Bazaarvoice where he has focused on improving the online shopping experience through the use of AI. Matthew is involved in delivering e-commerce machine learning solutions and optimising cloud performance at scale. Matthew is passionate about harnessing the capabilities of innovative technologies to solve real-world problems.*	Slides, Recording
High-Performance Inference for Video and Audio ORT provides the foundations for inference for Adobe's audio and video products (Premiere Pro, After Effects, Character Animator) on both Mac and Windows. In this talk, we'll discuss how ORT with the DML backend is essential in enabling high-throughput inference for audio and video workflows on Windows, and how we use ORT to enable speech to text on Mac. Video workflows are unique because of the sheer amount of data they process; our customers frequently ingest high resolution video >= 4k@60fps of which each frame may need to be passed through our models. Likewise, video workflows are inherently resource limited: the GPU is also being used for hardware decode and render at the same time. ORT gives us the tools to build complex frameworks and workflows on top of so that we can deliver ML-based features while ensuring that we're able to provide the best experience for our customers.	Nikhil Kalra, Adobe  *Nikhil Kalra is a Sr. Computer Scientist at Adobe and is currently the engineering lead and architect for the Digital Video and Audio applied machine learning team.*	Slides, Recording

Deploying on desktop with ONNX Topaz Labs develops deep learning based image quality software for professional and hobbyist photographers, which means running on the user's desktop or laptop. ONNX is an essential part of our solution to producing consistent results while making the most of a variety of consumer hardware. This type of deployment poses unique challenges and opportunities. Some experiences in this task have driven us to adopt certain useful strategies, tools, and techniques. Others remain interesting avenues for future improvement.	Alexander Zhang, Topaz Labs  *Alexander Zhang is a software developer at Topaz Labs, primarily responsible for Gigapixel AI and the inference pipeline for image models.*	Slides, Recording
ONNX Tools: Polygraphy and ONNX-GraphSurgeon Over the years, NVIDIA's TensorRT team has developed tooling that makes it easy to generate, transform, and debug ONNX models. Among other things, this includes a sanitizer that can simplify your models, and an automated bisector for debugging ('git bisect' for ONNX!). In this talk, I'll cover some of these tools and how you can effectively leverage them in your workflow.	Pranav Marathe, NVIDIA  *Pranav has worked as part of the TensorRT team at NVIDIA since 2018, developing, among other things, ONNX tooling like Polygraphy and ONNX-GraphSurgeon.*	Slides, Recording

Using ONNX with Qualcomm powered devices from smartphones to the cloud edge and everything in between Whenever our clients target high performant AI cloud inferencing servers, create new and exciting AI based experiences on mobile phones or improve our lives by adding more and more AI features into cars, many of them use ONNX models as an interchange format. Qualcomm helps to deploy and accelerate natural language processing, computer vision, classification, segmentation, and transformer based models in various verticals: Mobile, IoT, XR, Compute and Automotive. We created a link between ONNX and Qualcomm AI Engine direct that allows us to run the same model not only on various backends such as CPU, GPU, Hexagon processor or Low Power AI subsystem of the same SoC, and migrate it to run on range of the devices due to the portability that ONNX provides. In addition to the above, we would briefly cover in this session the work we are doing with Microsoft on collaboration for ONNX RT Execution Provider for a range of our AI accelerators.	Felix Baum, Qualcomm  *Felix Baum is responsible for AI software products at Qualcomm Technologies Inc. (QTI). Felix has spent 20+ years in the embedded industry, both as an embedded developer and as a product manager. He previously led QTI product management for Hexagon software, supporting DSPs with scalar, vector and tensor accelerators for camera, video, machine learning and audio verticals. Prior to that, he led marketing and product management efforts for various real-time operating system technologies. His career began at NASA's Jet Propulsion Laboratory at the California Institute of Technology, designing flight software for various spacecrafts. Felix holds a Master's degree in CS from the Cal State Northridge and an MBA from the UCLA.*	Slides, Recording
Onnx-mlir: an MLIR-based Compiler for ONNX Models - The Latest Status Onnx-mlir is an open source compiler implemented using the Multi-Level Intermediate Representation (MLIR) infrastructure recently integrated in the LLVM project. It compiles ONNX models into native code for CPUs as well as specialized accelerators. It is able to compile models for many platforms including x86 (Linux/Windows/macOS), Power (Linux) and z/Architecture (Linux and z/OS). Onnx-mlir is a subproject inside the ONNX ecosystem and has attracted many contributions from IBM, Microsoft, Facebook, Arm and Universities since its incubation in 2019. In this talk, we will show the latest status of the project by providing the project overview as well as the latest features.	Tung D. Le, IBM  *Tung D. Le is a researcher at IBM Research - Tokyo. He got Ph. D. from National Institute of Informatics, Japan in 2016 with major of systematic program transformation. His interest includes systematic methods to program transformation, high performance computing and compilers for AI. He is an ACM Senior Member.*	Slides, Recording
PFVM - A Neural Network Compiler that uses ONNX as its intermediate representation PFVM is a neural network compiler developed by Preferred Networks, which relies on ONNX as the Intermediate Representation format. PFVM is used in production environments to deploy models to various devices such as GPUs, multiple edge computing architectures, and PFN's own accelerator, MN-Core. PFVM's most salient features are; automatic checkpointing, operator fusion, and graph simplification that can be applied even when models have dynamic axes or unknown shapes. ONNX Shape inference becomes a critical element for all these optimizations, and the importance of bringing up more advanced shape inference mechanisms to address complex optimization scenarios is discussed in this talk.	Zijian Xu, Preferred Networks  *Zijian is a Neural network compiler engineer at Preferred Networks and an ONNX SIG-archinfra member.*	Slides, Recording

Bring the power of ONNX to Spark as it never happened before

Both data processing platforms and deep learning frameworks are evolving in their own fields. Usually, Spark is used for offline data processing, and then various deep learning frameworks are used for data inference. A simplified API for DL Inferencing is very important as a bridge.

What does an ideal data and deep learning inference pipeline look like? We'll discuss how to build your AI application using Spark and ONNX, the current status and initial idea of Spark commiunty to improve this pipeline, and also make full use of the capabilities of Ascend Hardware Platform.

This topic help you know the latest progress of Ascend Hardware Platform integration in ONNX, as well as the initial idea of the inference pipeline improvement in the Spark community.

Xiyuan Wang, Yikun Jiang, Zhipeng Huang, Huawei

Xiyuan Wang and Yikun Liang are Senior OpenSource Engineers at Huawei.

Zhipeng Huang is the Director of AI Open Source at Huawei.

Slides, Recording