



# ONNX

Community Meetup  
2023-06-28

## **ONNX Steering Committee**

Alexander Eichenberger (IBM)

Andreas Fehlner (Trumpf)

Mayank Kaushik (NVIDIA)

Prasanth Pulavarthi (MSFT)

Saurabh Tangri (Intel)



ONNX

Welcome!

# Agenda for Steering committee presentation

- ONNX Open Governance
- ONNX “State of the State”
- Companies / Tool Chains supporting ONNX
- Summary of ONNX Requests/Proposals for 2023
- ONNX Releases
- Learn how to get more involved with ONNX Steering Committee, SIGs and Working Groups

# Disclaimer

**All community meetup presentations will be recorded and made available publicly afterwards**

# Open Neural Network Exchange

The open standard for machine learning interoperability

GET STARTED

ONNX is an open format built to represent machine learning models. ONNX defines a common set of operators - the building blocks of machine learning and deep learning models - and a common file format to enable AI developers to use models with a variety of frameworks, tools, runtimes, and compilers. [LEARN MORE](#) >

## KEY BENEFITS



### Interoperability

Develop in your preferred framework without worrying about downstream inferencing implications. ONNX enables you to use your preferred framework with your chosen inference engine.

[SUPPORTED FRAMEWORKS](#) >



### Hardware Access

ONNX makes it easier to access hardware optimizations. Use ONNX-compatible runtimes and libraries designed to maximize performance across hardware.

[SUPPORTED ACCELERATORS](#) >



ONNX

Governance

# ONNX is a Community Project

## Steering Committee

<https://github.com/onnx/steering-committee>

Alexander Eichenberger (IBM)  
Andreas Fehlner (Trumpf)  
Mayank Kaushik (NVIDIA)  
Prasanth Pulavarthi (MSFT)  
Saurabh Tangri (Intel)

## Special Interest Groups (SIGs)

<https://github.com/onnx/sigs>

**Architecture & Infra:** Liqun Fu, Ke Zhang

**Operators:** Michał Karzyński, Ganesan Ramalingam

**Converters:** Thiago Crepaldi, Kevin Chen

**Model Zoo & Tutorials:** Jacky Chen

**Pre-processing:** Joaquin Anton

**Compilers:** Alexander Eichenberger, Philip Lassen



ONNX

State of the state

# Global Community



And more...

# Usage and Engagement

## Dependent Repos

15.8k

+130%

since June 2022

## Monthly Downloads

2.8M

+149%

since June 2022

## Pull Requests

9.4k

+21%

since June 2022

## Forks

3.6k

+29%

since June 2022

## Contributors

6.5k

+18%

since June 2022



ONNX

Progress Report  
on Requests

# ONNX 1.14 Released

[Release v1.14.0 onnx/onnx \(github.com\)](https://github.com/onnx/onnx/releases/tag/v1.14.0)

ONNX v1.11.0 comes with following updates:

- Opset Version 19, Protobuf v21
- New operators (DeformConv)
- Operator extensions (Equal, AveragePool, Pad, Resize)
- Introduces four new types for quantization / computation to speed up deep learning for GPUs and specialized accelerators.
  - FLOATE8E4M3FN, FLOATE8E5M2, FLOATE8E4M3FNUZ, FLOATE8E5M2FNUZ
  - Operator support for Cast, CastLike, QuantizeLinear, DequantizeLinear
- **Python 3.7 will be deprecated**

Visit ONNX.AI website to learn more

Thank you everyone for your countless hours of work!

# ONNX 1.13 Released

[Release v1.13.0 onnx/onnx \(github.com\)](https://github.com/onnx/onnx/releases/tag/v1.13.0)

ONNX v1.13.0 comes with following updates:

- New operators (Col2Im, BitwiseNot , BitwiseAnd, BitwiseOr and BitwiseXor)
- Operator extensions (Resize, Pad, OptionalHasElement, OptionalHasElement, OptionalGetElement, ScatterElement,ScatterND, Split , LpPool)
- Function Updates
  - CenterCropPad, mish,GroupNormalization
- ONNXIFI: has been deprecated.
- ONNX 1.13.0 supports Python 3.11
- Support for Apple M1/M2 ARM processors

Visit ONNX.AI website to learn more

Thank you everyone for your countless hours of work!

# 2023 Proposals

- 1) **Support for Hybrid FP8 and MHA operator.** Arnab Raha / Intel
- 2) **Support for FP8 data formats.** Naveen Mellempudi / Intel
- 3) **Float16 support in quantizelinear and dequantizelinear.** Murali Ambati / Intel
- 4) **Adding bit parameter to QuantizeLinear and DequantizeLinear operator for improved hardware compatibility.** Lucas Fischer
- 5) **Adopting a wider set of quantized ops into ONNX.** Peter van Beek, Aleksandar Sutic, Thomas Gardos / Intel
- 6) **Distributed Inference and Communication Collectives.** Ganesan Ramalingam / Microsoft
- 7) **Add a tokenization op.** W. Tambellini / RWS
- 8) **Tokenizer / Support for Transformers.** Bourhan Dernayka / IBM
- 9) **Overhauling Training Info Proto for more comprehensive training information.** Taka Shinagawa / Microsoft
- 10) **Web AI GPU support for JavaScript.** Alexander Visheratin
- 11) **Double calculation on Tree Ensemble (Regressor & Classifier) ML operators.** Stefan Acin / Aizon
- 12) **Additional beam search support and Dot Product Attention Op.** William Tambellini / RWS/LanguageWeaver
- 13) **Additional OP DCNv2 in ONNX format.** Wei Wen / Intel
- 14) **Op support and usability improvements to ONNX for GNNs.** Ramakrishnan Sivakumar / Groq
- 15) **Multi-Head attention operation on ONNX.** Alessandro Palla / Intel
- 16) **Sparse Tensor Support.** Jacob Renn / AI Squared
- 17) **Define operator attributes and add data-driven post-training sparsification capabilities.** Manuj Sabharwal, Ken Koyanagi / Intel
- 18) **Datetime parsing.** Christian Bourjau / QuantCo

# Thank you ...

- Please stay engaged and continue your contributions to ONNX and its related projects.
- Remember to use the following ONNX resources:
  - Website: <https://onnx.ai/>
  - GitHub: <https://github.com/onnx>
  - Slack: (join <https://slack.lfai.foundation> - email, password, then find onnx-general)
  - Calendar: <https://onnx.ai/calendar>
  - Mailing List: <https://lists.lfai.foundation/g/onnx-announce>



Questions?