ONNX | Workshop 10/14/2020

ONNX

Welcome!

Disclaimer

All workshop presentations, SIG/WG sessions will be recorded and made available publicly afterwards.

Logistics

- Host of Zoom Meeting will share the slides on screen and record all presentations.
- All participants will be muted except when presenting.
- Questions should be posted in the Slack “onnx-general”
- Please “raise hand” (Zoom feature) if you would like to speak and engage in the discussion.

Goals for the Workshop

- Get the latest updates on ONNX - Processes, Roadmap Releases, and SIGs/WGs
- Learn from the community and how ONNX is being used
- Share feedback on what is working (and what isn’t)
- Learn how to get more involved with ONNX Steering Committee, SIGs and Working Groups

Agenda -1

7:00	Welcome
7:05	ONNX SC Updates
7:25	Community Updates
9:05	Break
9:15	SIG Updates
9:55	Wrap Up

Sheng Zha (Amazon)	Welcome Logistics Goals Agenda State of the State: ONNX Growth
Jacky Chen (Microsoft)	Release 1.8
Prasanth Pulavarthi (Microsoft)	Governance
Harry Kim (Intel)	Roadmap

Agenda - 2

7:00	Welcome
7:05	ONNX SC Updates
7:25	Community Updates
9:05	Break
9:15	SIG Updates
9:55	Wrap Up

Patrick St-Amant (Zetane)	Extract the Maximum Benefits of ONNX to Shorten Your Development Cycle Time and Reduce Guesswork
Jianhao Zhang (OneFlow)	ONNX at OneFlow
Morgan Funtowicz (Hugging Face)	Efficient Inference of Transformers Models: Collaboration Highlights Between Hugging Face & ONNX Runtime
Danilo Pau (ST Micro)	Flows and Tools to Map ONNX Neural Networks on Micro-controllers
Fabian Bause (Beckhoff Automation)	Neural Automation: Fusion of Automation and Data Science
Faith Xu (Microsoft)	ONNX Runtime Updates: Mobile, Quantization, Training, and More
Jason Knight (OctoML)	Apache TVM and ONNX, What Can ONNX Do for DL Compilers (and vice versa)
Alexandre Eichenberger (IBM Research)	ONNX Support in the MLIR Compiler: Approach and Status
Matteo Interlandi (Microsoft)	Hummingbird
Neta Zmora (NVIDIA)	D/DQ is All You Need

Agenda - 3

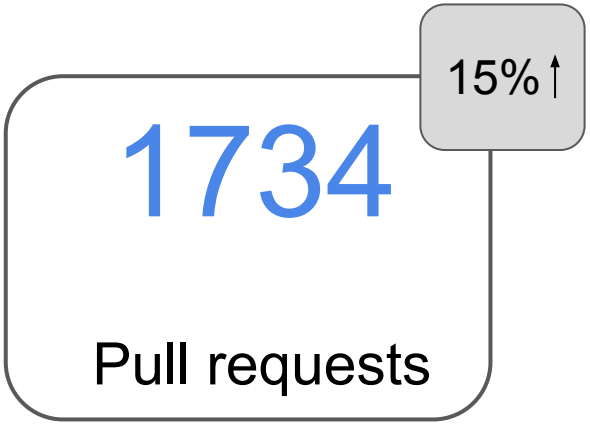
7:00	Welcome
7:05	ONNX SC Updates
7:25	Community Updates
9:05	Break
9:15	SIG Updates
9:55	Wrap Up

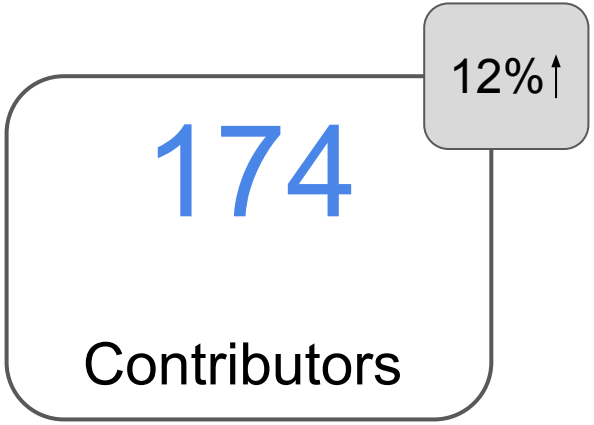
Ashwini Khade (Microsoft) & Ke Zhang (Alibaba)	Architecture/Infrastructure SIG
Michał Karzyński (Intel) & Emad Barsoum (Microsoft)	Operators SIG
Chin Huang (IBM) & Guenther Schmuelling (Microsoft)	Converters SIG
Wenbing Li (Microsoft) & Vinitra Swamy (EPFL)	Model Zoo/Tutorials SIG

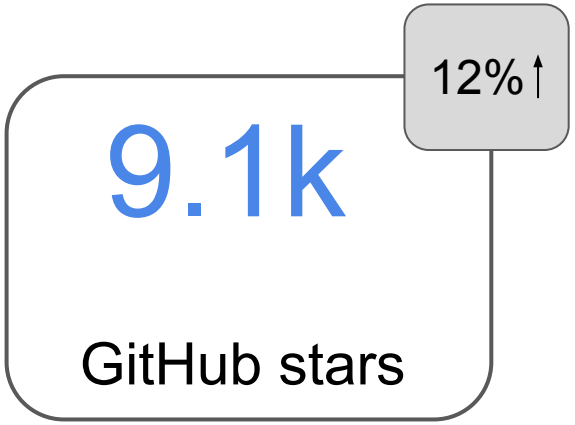
ONNX

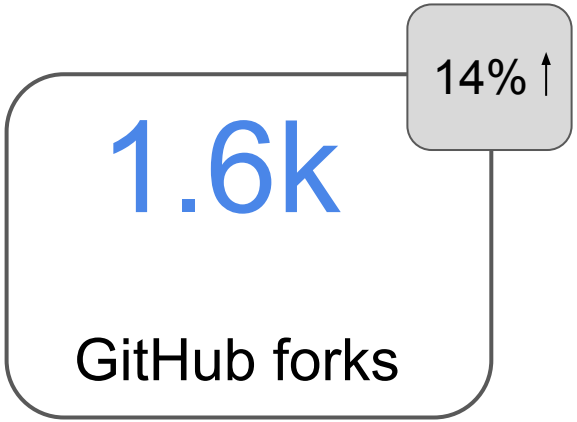
State of the state

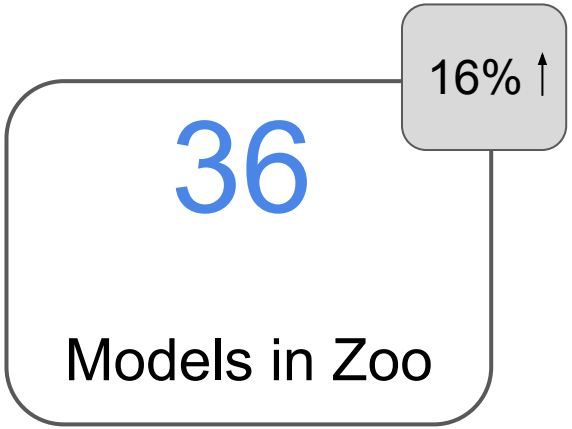
Engagement & usage (compared to 4/9/20)



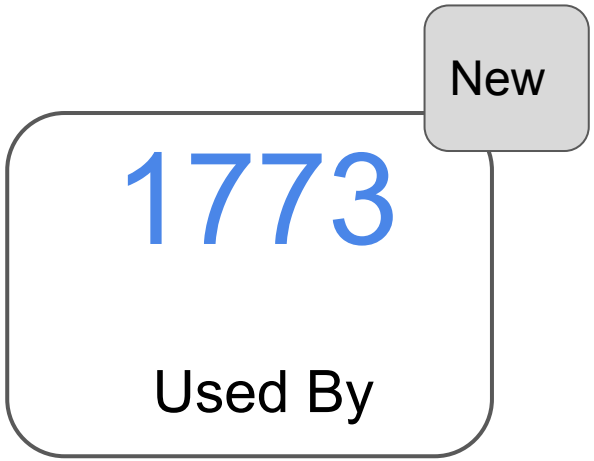








Adding new growth measure...



Support


NEW
Creation/
Manipulation
Caffe2
Chainer
Cognitive Toolkit
MATLAB
mxnet
ML.NET
PyTorch
PaddlePaddle
Keras
scikit learn
dmlc XGBoost
DeepCore
LibSVM
Neural Networ Librarie
sas
[M]s
SIEMENS
Yandex CatBoost
NeoML
Tencent
TENGINE
SINGA
NEW
Run/
Compile
intel AI
ONNX RUNTIME
Qualcomm
SYNOPSYS
NNOIR
Tencent
NVIDIA
MACE
Mobile AI Compute Engine
SOPHON
Windows
vespa
habana
Rockchip
CEVA
nGraph
skymizer
tvm
GRAPHCORE
cadence
tensilica
TwinCAT 3
Windows ML
Acumos
BITMAIN
Visualization/
Test Tools
NETRON
Visual DL
Sionnx

Coming soon: ONNX 1.8 (Jacky)

ONNX v1.8 comes with exciting new and enhanced features!

- Windows conda package will be available for the upcoming 1.8 Release (last for v1.1.1)
- Adding Differentiable tags to make Gradient operator better defined
- Remove GraphCall and eliminate the need to implement GraphCall
- Large model (>2GB model) support added for checker and shape_inference
- Graph level shape inference fixes to patch the IR gap introduced since IR version 4
- Node level shape inference fixes for operators
- More operators are supported by version converter
- Add serialization for inputs and outputs of Sequence and Map data types
- Opset 13
    - Extend ControlFlow to allow Sequence type for inputs and outputs
    - Support per-axis scaling for quantizing and dequantizing of tensors
    - Add bfloat16 support

Thank you everyone for your countless hours of work!

ONNX 1.8 Release Schedule

1. Week of Validation (10/13~)
    a. Cut ONNX Release branch
    b. ONNX Release candidate published in PyPI test
    c. Validation in ONNXRuntime
    d. Community validation
2. Week of Release (10/22~): Ready for ONNX 1.8 Release

ONNX

Governance (prasanth)

ONNX open governance update

Steering Committee
https://github.com/onnx/steering-committee

Prasanth Pulavarthi (MS)
Harry Kim (Intel)
Jim Spohrer (IBM)
Sheng Zha (AWS)
Joohoon Lee (Nvidia)

Special Interest Groups (SIGs)
https://github.com/onnx/sigs

Architecture & Infra: Ashwini Khade, Ke Zhang

Operators: Michał Karzyński, Emad Barsoum

Converters: Chin Huang, Guenther Schmuelling

Model Zoo & Tutorials: Wenbing Li

Working Groups (WGs)
https://github.com/onnx/working-groups

Training: Svetlana Levitan

ONNX open governance changes

Updated licensing: All code repos under ONNX will be Apache 2. Prior contributions will be reclassified with contributing organization sign-off. Document repos remain CCL.

CLA –> DCO:
DCO bot already enabled on all repos under ONNX. Will be made required by 10/19 (already required on main onnx repo). CLA will be turned off once license files updates.

To pass DCO bot, all commits in PRs need to be signed.
Easy to sign: if using command line, git commit **-s**
If using web UI or other tools, include "**Signed-off-by**: Humpty Dumpty <humpty.dumpty@example.com>" in the commit message (for each commit, not for the PR). Make sure email matches the account you are submitting with.

CONTRIBUTING.md will be updated with tips

ONNX Community Forums

Gitter - ONNX rooms will be deprecated by 10/19.
Please switch to GitHub Discussions and LF AI Slack

GitHub Discussions - new GitHub feature now enabled on onnx/onnx repo, will be enabled on other repos soon.
Good for technical questions and discussions that don't work well as Issues.
Issues can be converted to Discussions, but not vice versa.

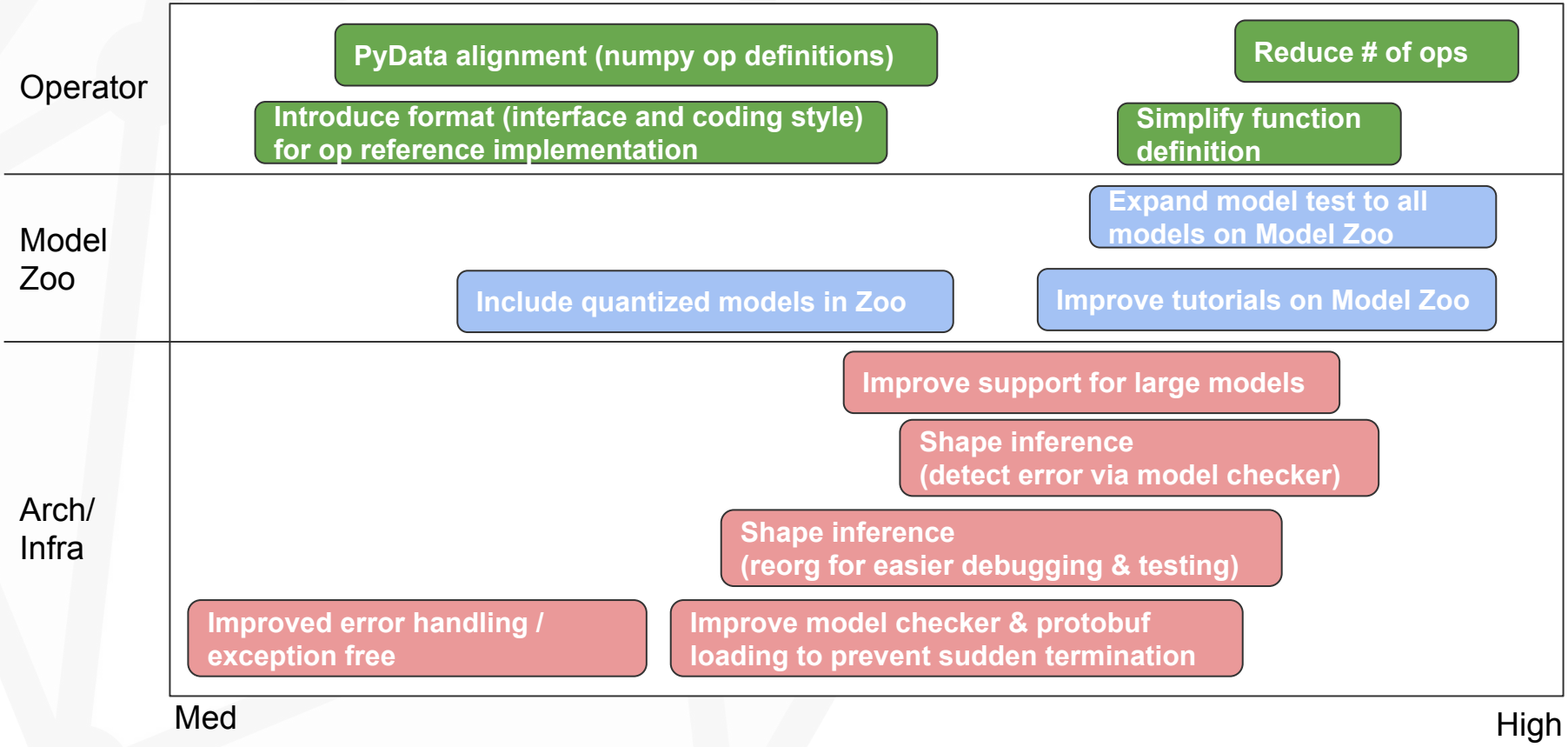
Slack - ONNX channels in LF AI Slack. Channels exist for each SIG and WG
Sign up for LF AI Slack and then join the ONNX channels

ONNX

Roadmap (harry)

ONNX roadmap discussions

Suggested features & their rated impact



Questions?

ONNX

Wrap up!

Thank you ...

- Recording of today's workshop and other applicable content will be shared via ONNX-Announce mailing list when available.
- Please stay engaged and continue to contribute to ONNX and ONNX 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