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
<table>
<thead>
<tr>
<th>Time</th>
<th>Agenda Item</th>
<th>Speaker(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>7:00</td>
<td>Welcome</td>
<td>Sheng Zha (Amazon)</td>
</tr>
<tr>
<td>7:05</td>
<td>ONNX SC Updates</td>
<td>Jacky Chen (Microsoft)</td>
</tr>
<tr>
<td>7:25</td>
<td>Community Updates</td>
<td>Prasanth Pulavarthi (Microsoft)</td>
</tr>
<tr>
<td>9:05</td>
<td>Break</td>
<td>Harry Kim (Intel)</td>
</tr>
<tr>
<td>9:15</td>
<td>SIG Updates</td>
<td></td>
</tr>
<tr>
<td>9:55</td>
<td>Wrap Up</td>
<td></td>
</tr>
</tbody>
</table>
## Agenda

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
<th>Presenter</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>7:00</td>
<td>Welcome</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7:05</td>
<td>ONNX SC Updates</td>
<td>Patrick St-Amant (Zetane)</td>
<td>Extract the Maximum Benefits of ONNX to Shorten Your Development Cycle Time and Reduce Guesswork</td>
</tr>
<tr>
<td>7:25</td>
<td>Community Updates</td>
<td>Jianhao Zhang (OneFlow)</td>
<td>ONNX at OneFlow</td>
</tr>
<tr>
<td>9:05</td>
<td>Break</td>
<td>Morgan Funtowicz (Hugging Face)</td>
<td>Efficient Inference of Transformers Models: Collaboration Highlights Between Hugging Face &amp; ONNX Runtime</td>
</tr>
<tr>
<td>9:15</td>
<td>SIG Updates</td>
<td>Danilo Pau (ST Micro)</td>
<td>Flows and Tools to Map ONNX Neural Networks on Micro-controllers</td>
</tr>
<tr>
<td>9:55</td>
<td>Wrap Up</td>
<td>Fabian Bause (Beckhoff Automation)</td>
<td>Neural Automation: Fusion of Automation and Data Science</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Faith Xu (Microsoft)</td>
<td>ONNX Runtime Updates: Mobile, Quantization, Training, and More</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Jason Knight (OctoML)</td>
<td>Apache TVM and ONNX, What Can ONNX Do for DL Compilers (and vice versa)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Alexandre Eichenberger (IBM Research)</td>
<td>ONNX Support in the MLIR Compiler: Approach and Status</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Matteo Interlandi (Microsoft)</td>
<td>Hummingbird</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Neta Zmora (NVIDIA)</td>
<td>D/DQ is All You Need</td>
</tr>
<tr>
<td>Time</td>
<td>Session</td>
<td>Presenters</td>
<td>SIGs</td>
</tr>
<tr>
<td>-------</td>
<td>------------------</td>
<td>-----------------------------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>7:00</td>
<td>Welcome</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7:05</td>
<td>ONNX SC Updates</td>
<td>Ashwini Khade (Microsoft) &amp; Ke Zhang (Alibaba)</td>
<td>Architecture/Infrastructure SIG</td>
</tr>
<tr>
<td>7:25</td>
<td>Community Updates</td>
<td>Michał Karzyński (Intel) &amp; Emad Barsoum (Microsoft)</td>
<td>Operators SIG</td>
</tr>
<tr>
<td>9:05</td>
<td>Break</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9:15</td>
<td>SIG Updates</td>
<td>Chin Huang (IBM) &amp; Guenther Schmuelling (Microsoft)</td>
<td>Converters SIG</td>
</tr>
<tr>
<td>9:55</td>
<td>Wrap Up</td>
<td>Wenbing Li (Microsoft) &amp; Vinitra Swamy (EPFL)</td>
<td>Model Zoo/Tutorials SIG</td>
</tr>
</tbody>
</table>
ONNX | State of the state
Engagement & usage (compared to 4/9/20)

- Pull requests: 1734 (15%↑)
- Contributors: 174 (12%↑)
- GitHub stars: 9.1k (12%↑)
- GitHub forks: 1.6k (14%↑)
- Published Papers: 180 (41%↑)
- Models in Zoo: 36 (16%↑)
Adding new growth measure...

1773

Used By

New
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 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
Roadmap
(harry)
ONNX roadmap discussions

- onnx.ai/roadmap
- Feedback from community
- onnx.ai/impact
- Impact analysis
- Cost analysis

6 weekly community discussions
Suggested features & their rated impact

<table>
<thead>
<tr>
<th>Operator</th>
<th>Reduce # of ops</th>
<th>Simplify function definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>PyData alignment (numpy op definitions)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Introduce format (interface and coding style) for op reference implementation</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model Zoo</th>
<th>Expand model test to all models on Model Zoo</th>
<th>Improve tutorials on Model Zoo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Include quantized models in Zoo</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Arch/Infra</th>
<th>Improve support for large models</th>
<th>Improve model checker &amp; protobuf loading to prevent sudden termination</th>
</tr>
</thead>
<tbody>
<tr>
<td>Improved error handling / exception free</td>
<td>Shape inference (detect error via model checker)</td>
<td>Shape inference (reorg for easier debugging &amp; testing)</td>
</tr>
</tbody>
</table>
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/](https://onnx.ai/)
  - GitHub: [https://github.com/onnx](https://github.com/onnx)
  - Slack: (join [https://slack.lfai.foundation](https://slack.lfai.foundation) - email, password, then find #onnx-general)
  - Calendar: [https://onnx.ai/calendar](https://onnx.ai/calendar)
  - Mailing List: [https://lists.lfai.foundation/g/onnx-announce](https://lists.lfai.foundation/g/onnx-announce)