



ONNX |

Workshop
3/24/2020



ONNX

Welcome!

Disclaimer

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

Welcome Message from Host - Ti Zhou and Baidu
PaddlePaddle Team in China (Picture of team
potentially)



Logistics

- Host of Zoom Meeting will share the slides on screen and record all presentations. Simul-cast of zoom on Bilibili link
- All participants will be muted except when presenting.
- Questions should be posted in the Slack “onnx-general” or zoom chat or Bilibili
- 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

8:00/ 5:00	Welcome
8:05/ 5:05	ONNX SC Updates
8:25/ 5:25	Community Updates
10:05/ 7:05	Break
10:15/ 7:15	SIG Updates
10:55/ 7:55	Wrap Up

Ti Zhou, Baidu

Welcome
Logistics
Goals
Agenda

Sheng Zha, Amazon

State of the State: ONNX Growth
Governance and Election

Joohoon Lee, NVIDIA

Roadmap and Release 1.8

Agenda - 2

8:00/ 5:00	Welcome
8:05/ 5:05	ONNX SC Updates
8:25/ 5:25	Community Updates
10:05/ 7:05	Break
10:15/ 7:15	SIG Updates
10:55/ 7:55	Wrap Up

Han Zhao (GraphCore-UK)	popONNX: Support ONNX on IPU
Yu Feng Wei (SenseTime-HongKong)	Spring Project: Multi Backend Neural Network Auto Quantization and Deploy over ONNX
Tom Wildenhain (Microsoft-USA)	ONNX Runtime for Mobile Scenarios: From model to on-device inferencing
Wranky Wang (Baidu-China)	Introduction to DL Framework PaddlePaddle and Paddle2ONNX Module
Rohit Sharma (AITechSystems-USA_CA)	ONNX on microcontrollers
Krishna Gade (FiddlerAI-USA_CA)	Monitoring and Explaining ONNX Models in Production
Philippe Dooze (Orange-France) (picture)	ONNX client for Acumos
Leon Wang (Huawei-China)	Deploy ONNX model seamlessly across the cloud, edge, and mobile devices using MindSpore
Peng Wang (Microsoft China)	ONNX Runtime Training
Haihao Shen (Intel - China) and Saurabh Tangri (Intel)	Quantization support for ONNX using LPOT (Low precision optimization tool)

Agenda - 3

8:00/ 5:00	Welcome
8:05/ 5:05	ONNX SC Updates
8:25/ 5:25	Community Updates
10:05/ 7:05	Break
10:15/ 7:15	SIG Updates
10:55/ 7:55	Wrap Up

Ashwini Khade (Microsoft) and Jacky Chen (Microsoft)	Architecture/Infrastructure SIG
Michał Karzyński (Intel) and Ganesan Ramalingen (Microsoft)	Operators SIG
Chin Huang (IBM) and Guenther Schmuelling (Microsoft)	Converters SIG
Wenbing Li, Microsoft	Model Zoo/Tutorials SIG

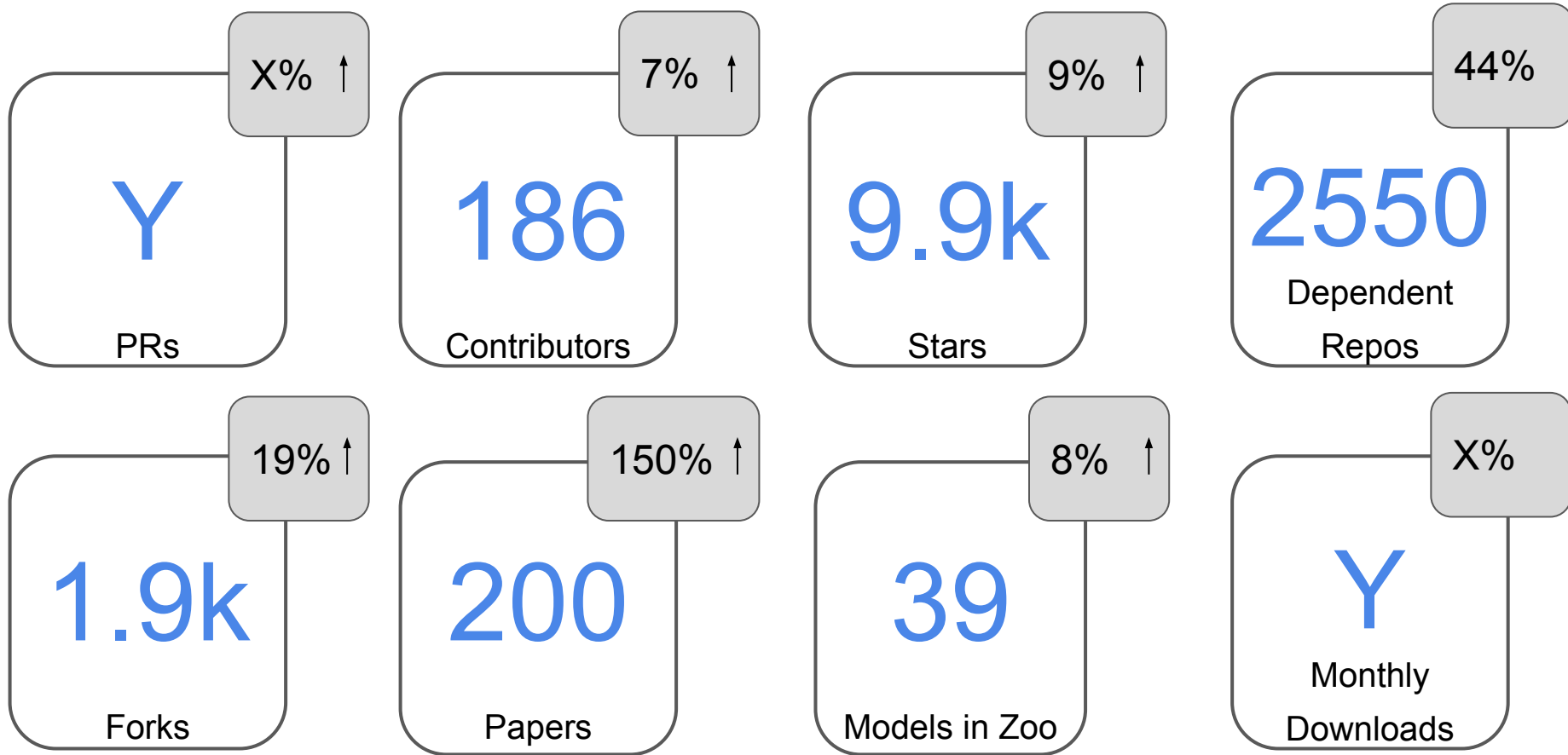


ONNX

State of the state

Engagement & usage (from 11/9/20 to 3/18/21)

(please check and update speaker notes too – will help for next time :-)



Support

Creation/ Manipulation



NEW

Run/ Compile



NEW

Visualization/ Test Tools





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, Jacky Chen

Operators: Michał Karzyński, Ganesan Ramalingen

Converters: Guenther Schmuelling, Kevin Chen, Chin Huang

Model Zoo & Tutorials: Wenbing Li

Working Groups (WGs)

<https://github.com/onnx/working-groups>

Release: TBD

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

Slack - ONNX channels in [LF AI & Data Slack](#). Channels exist for each SIG and WG
Sign up for LF AI & Data Slack and then join the ONNX channels

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.

Bi-Annual LF AI & Data Day Virtual Meetups - Looking for future host of virtual meetups, one each Fall and Spring. Planning starts 3 months before events - key gather 10 community talks.

Face-to-Face Workshops – TBD Post-Pandemic.

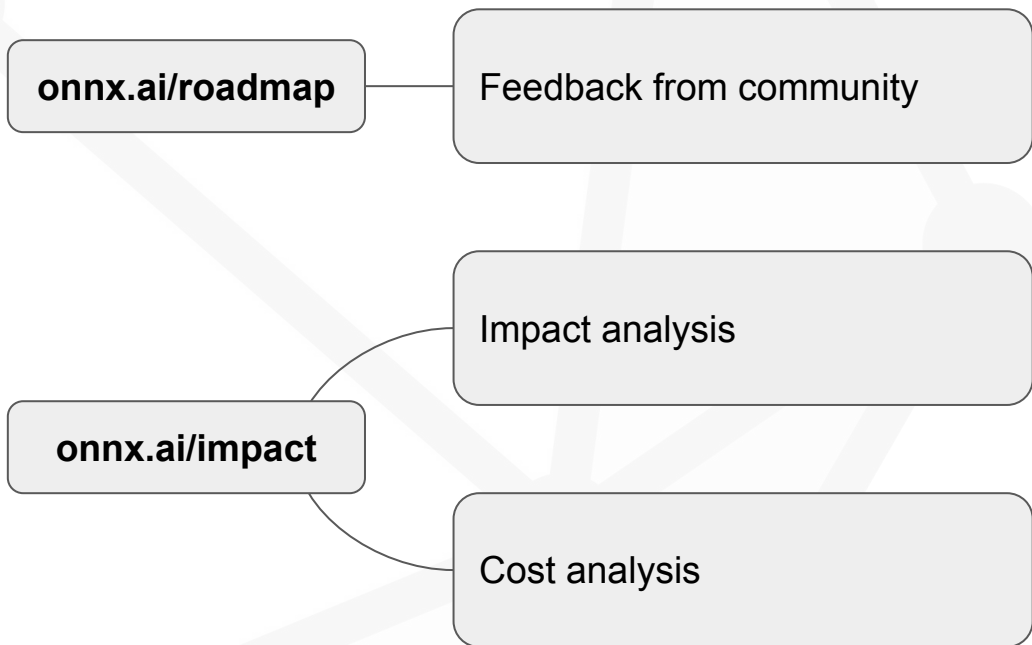
Upcoming SC Elections – Harry (slide)



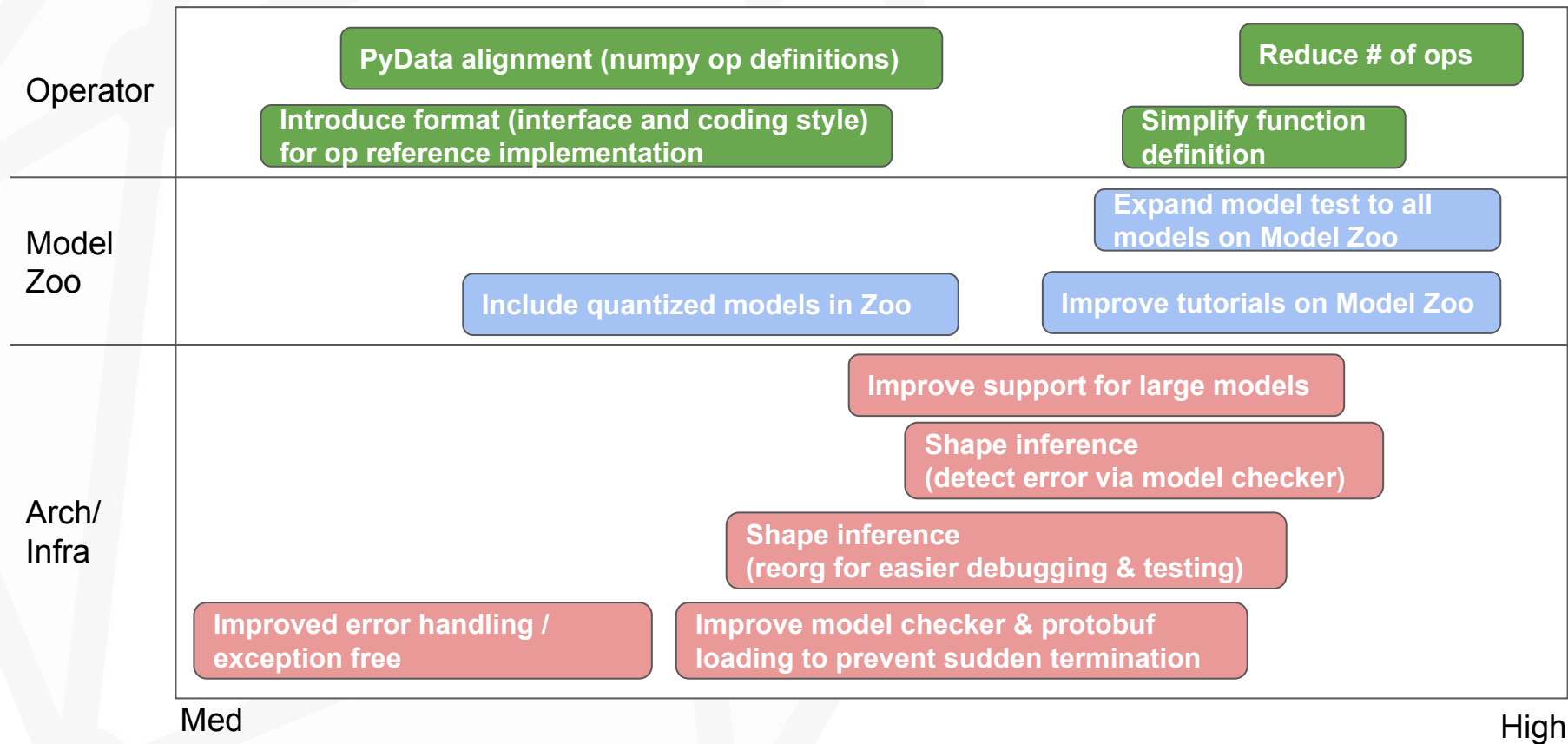
ONNX

Roadmap
(Harry)

ONNX roadmap discussions



Suggested features & their rated impact



Coming soon: ONNX 1.9 (Release mgr: TBD)

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 1.9 Release Schedule

1. Week of Validation (3/19~)
 - a. Cut ONNX Release branch
 - b. ONNX Release candidate published in PyPI test
 - c. Validation in ONNXRuntime
 - d. Community validation
2. Week of Release (4/5~): Ready for ONNX 1.9 Release



Questions?



ONNX

Community
Presentations

Agenda - 2

8:00/ 5:00	Welcome
8:05/ 5:05	ONNX SC Updates
8:25/ 5:25	Community Updates
10:05/ 7:05	Break
10:15/ 7:15	SIG Updates
10:55/ 7:55	Wrap Up

Han Zhao (GraphCore-UK)	popONNX: Support ONNX on IPU
Yu Feng Wei (SenseTime-HongKong)	Spring Project: Multi Backend Neural Network Auto Quantization and Deploy over ONNX
Tom Wildenhain (Microsoft-USA)	ONNX Runtime for Mobile Scenarios: From model to on-device inferencing
Wranky Wang (Baidu-China)	Introduction to DL Framework PaddlePaddle and Paddle2ONNX Module
Rohit Sharma (AITechSystems-USA_CA)	ONNX on microcontrollers
Krishna Gade (FiddlerAI-USA_CA)	Monitoring and Explaining ONNX Models in Production
Philippe Dooze (Orange-France) (picture)	ONNX client for Acumos
Leon Wang (Huawei-China)	Deploy ONNX model seamlessly across the cloud, edge, and mobile devices using MindSpore
Peng Wang (Microsoft China)	ONNX Runtime Training
Haihao Shen (Intel - China) and Saurabh Tangri (Intel)	Quantization support for ONNX using LPOT (Low precision optimization tool)



ONNX

Break

Resume at 9:15 PST



ONNX

SIG
Presentations

Agenda - 3

8:00/ 5:00	Welcome
8:05/ 5:05	ONNX SC Updates
8:25/ 5:25	Community Updates
10:05/ 7:05	Break
10:15/ 7:15	SIG Updates
10:55/ 7:55	Wrap Up

Ashwini Khade (Microsoft) and Jacky Chen (Microsoft)	Architecture/Infrastructure SIG
Michał Karzyński (Intel) and Ganesan Ramalingen (Microsoft)	Operators SIG
Chin Huang (IBM) and Guenther Schmuelling (Microsoft)	Converters SIG
Wenbing Li, Microsoft	Model Zoo/Tutorials SIG



ONNX

Wrap up!



ONNX

Wish we had more
time!

Session	Deck / Prerecording (Links)
<p>My experience implementing ONNX import for GAP processors Contact: Martin Croome (Greenwaves Tech)</p>	<p>deck / recording</p>
<p>How we are making it insanely easy to deploy ml/ai models from jetson nano to Azure with ONNX Contact: Mahesh Yadav (Microsoft)</p>	<p>deck / prerecording</p>
<p>Visualizing ONNX models' internal data: Key things to look for? Contact: Mina Amiri (Zetane)</p>	<p>deck/recording</p>
<p>Deploying 3rd Party Models in PaddlePaddle via X2Paddle Converter Contact: JiaJun Jiang (Baidu)</p>	<p>deck / prerecording</p>

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>