ONNX | Workshop 3/24/2021

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 1.9

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)	ONNX on microcontrollers
Krishna Gade (FiddlerAI-USA)	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

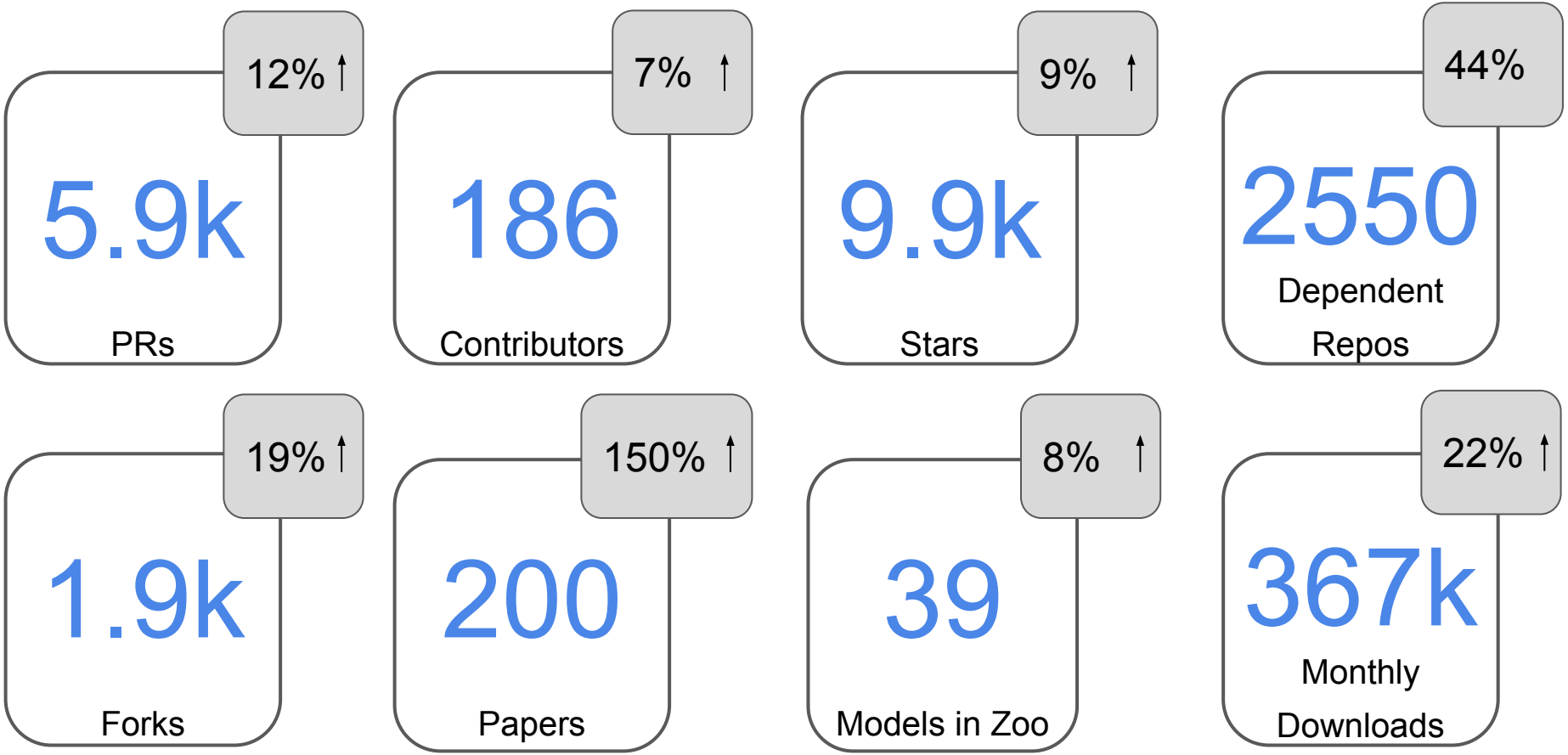
8:00/ 5:00	Welcome
8:05/ 5:05	ONNX SC Updates
8:25/ 5:25	Community Updates
10:05/ 7:05	Break
10:15/ 7:15	SIG Updates
10:55/ 7:55	Wrap Up

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) and Kevin Chen (NVIDIA)	Converters SIG
Wenbing Li, Microsoft	Model Zoo/Tutorials SIG

ONNX

State of the state (Sheng)

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



Support

Creation/ Manipulation

Caffe2, Chainer, Cognitive Toolkit, LibSVM, mxnet, NeoML, ML.NET, PyTorch, Keras, Tencent, [M]s, Yandex CatBoost, scikit learn, dmlc XGBoost, PaddlePaddle, SAS, MyCaffe, SINGA, DeepCore, MATLAB, Neural Network Libraries, SIEMENS, Tengine

NEW

DataAPIs
https://data-apis.org/

Run/ Compile

ONNX RUNTIME, intel AI, Qualcomm, SYNOPSYS CEVA, TwinCAT 3, Tencent, Windows, NVIDIA, habana, Acumos, nGraph, skymizer, SOPHON, MACE Mobile AI Compute Engine, NNOIR, GRAPHCORE, Rockchip, vespa, tvm, cadence tensilica

NEW

HAILO

Visualization/ Test Tools

NETRON, Visual DL, Sionnx, ZETANE

ONNX

Governance

ONNX open governance update

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

Prasanth Pulavarthi (MS)
Harry Kim (FB)
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, Ganesan Ramalingam

Converters: Guenther Schmuelling, Chin Huang

Model Zoo & Tutorials: Wenbing Li

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

Training: Svetlana Levitan

Recently closed WGs: ***Release***

ONNX open governance changes

Updated licensing: All code repos under ONNX are now under Apache License v2.0.

CLA -> DCO:
DCO bot enabled on all repos under ONNX and are required, replacing CLA.

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.

See CONTRIBUTING.md for more 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.

ONNX open governance election process

Now accepting applications until 4/19 (Mon)**: onnx.ai/sc-apply**

Eligibility

Candidate: Self-nominated and not required to be a Contributor
Voter: Contributors to ONNX project

Methodology

- 1 vote per Member Company.
- Condorcet voting with Schultz method (Ranked preference)
- Contributor votes roll up to associated company
- Final result based on Member Company votes
- All votes published for transparency

Timeline

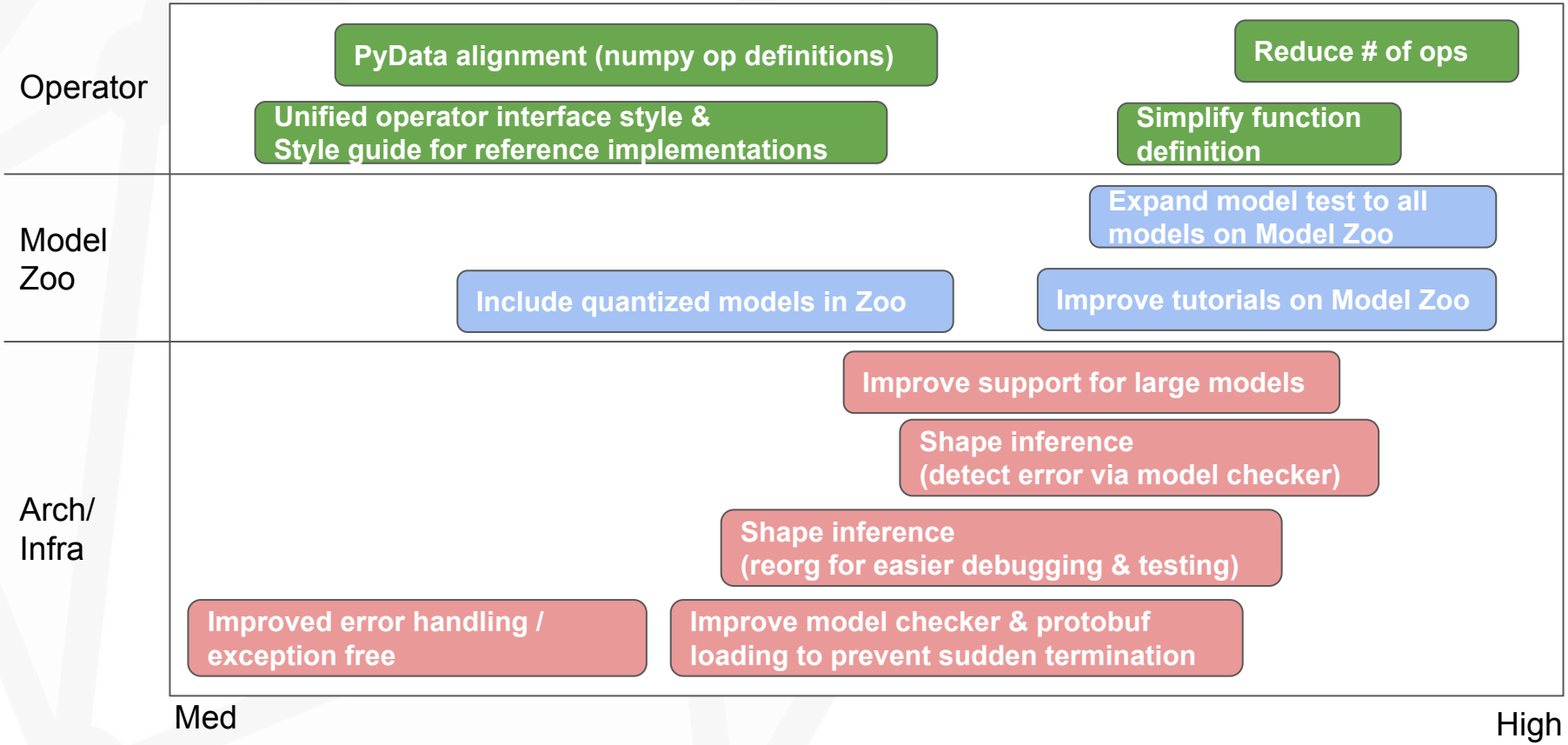
- April: Nomination and candidate campaigns
- May: Election and transition
- June: New Steering Committee

ONNX

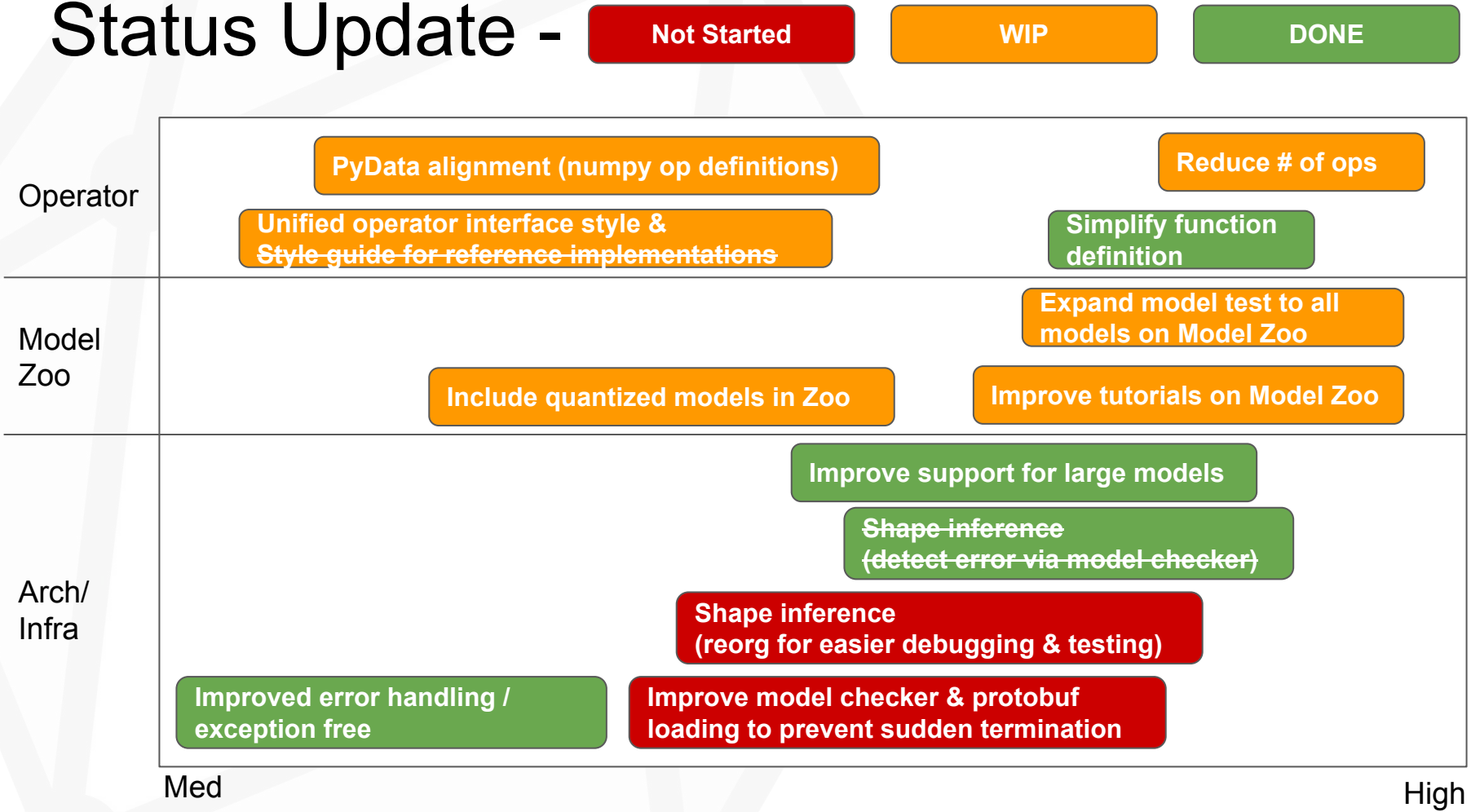
Roadmap (Joohoon)

ONNX roadmap discussions

Suggested features & their rated impact



Status Update -



Coming soon: ONNX 1.9 (Release mgr: Michal Karzynski)

ONNX v1.9 comes with exciting new and enhanced features!

- Symbolic shape inference
- Removing optimizers from onnx packages
- Updates to external data helpers
- Selective load of ONNX schema by specific opset_version
- ONNX Parser
  - Text-based syntax for ONNX models
  - Simplify function definitions & test cases
- Opset 14
  - Updated - Cumsum, Relu, Reshape, GRU, LSTM, RNN, BatchNorm
  - New - Trilu, HardSwish

Thank you everyone for your countless hours of work!

ONNX 1.9 Release Schedule

1. Cut ONNX 1.9 release branch (3/31)
2. Week of validation (4/2~)
    a. ONNX release candidate published in PyPI test
    b. Validation in ONNX Runtime
    c. Community validation
    d. Converters validation
3. Target release date (Week of 4/12~)
    a. 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)
My experience implementing ONNX import for GAP processors Contact: Martin Croome (Greenwaves Tech)	deck / recording
How we are making it insanely easy to deploy ml/ai models from jetson nano to Azure with ONNX Contact: Mahesh Yadav (Microsoft)	deck / prerecording
Visualizing ONNX models' internal data: Key things to look for? Contact: Mina Amiri (Zetane)	deck/ recording
Deploying 3rd Party Models in PaddlePaddle via X2Paddle Converter Contact: JiaJun Jiang (Baidu)	deck / prerecording

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