# LF AI & Data Technical Advisory Council

Biweekly call - May 2, 2024

# Antitrust Policy

- Linux Foundation meetings involve participation by industry competitors, and it is the intention of the Linux Foundation to conduct all of its activities in accordance with applicable antitrust and competition laws. It is therefore extremely important that attendees adhere to meeting agendas, and be aware of, and not participate in, any activities that are prohibited under applicable US state, federal or foreign antitrust and competition laws.

- Examples of types of actions that are prohibited at Linux Foundation meetings and in connection with Linux Foundation activities are described in the Linux Foundation Antitrust Policy available at http://www.linuxfoundation.org/antitrust-policy. If you have questions about these matters, please contact your company counsel, or if you are a member of the Linux Foundation, feel free to contact Andrew Updegrove of the firm of Gesmer Undergone LLP, which provides legal counsel to the Linux Foundation.

# Recording of Calls

TAC calls are recorded and available for viewing on the TAC Wiki

# LF AI & Data Useful Links

- Web site: lfaidata.foundation
- Wiki: wiki.lfaidata.foundation
- GitHub: github.com/lfai
- Landscape: https://landscape.lfaidata.foundation
- Mail Lists: https://lists.lfaidata.foundation
- Slack: https://slack.lfaidata.foundation
- Youtube: https://www.youtube.com/channel/UCfasaeqXJBCAJMNO9HcHfbA
- LF AI Logos: https://github.com/lfai/artwork
- PPT Template: https://drive.google.com/file/d/1eiDNJvXCqSZHT4Zk_-czASlz2GTBRZk2/view
- Events: https://lfaidata.foundation/events/
- Events Calendar https://wiki.lfaidata.foundation/pages/viewpage.action?pageId=12091544
- Event Wiki https://wiki.lfaidata.foundation/pages/viewpage.action?pageId=10518553

# Agenda

- Roll Call  (1 mins)
- Approval of Minutes from previous meeting  (2 mins)
- IREE Requesting Sandbox (AMD, Google) (35 mins)
- Q&A  (10 mins)
- Voting  (5 mins)
- Next TAC Meetings (2 mins)
- Open Discussion

# TAC Voting Members

Note: we still need a few designated backups specified on [wiki]

| Company or Graduated Project | Level or Project Level | Eligibility | | | Representative Alternates |
|---|---|---|---|---|---|
| 4paradigm | Premier | Voting Member | China | Zhongyi Tan | |
| Microsoft | Premier | Voting Member | USA | Ali Dalloul | |
| Amazon Web Services | Premier | Voting Member | USA | Brian Granger | Mark Atwood |
| Ericsson | Premier | Voting Member | Sweden | Rani Yadav-Ranjan | |
| Huawei | Premier | Voting Member | China | Howard (Huang Zhipeng) | Charlotte (Xiaoman Hu), Leon (Hui Wang) |
| IBM | Premier | Voting Member | USA | Susan Malaika | Beat Buesser, Alexandre Eichenberger |
| OPPO | Premier | Voting Member | China | Jimmy (Hongmin Xu) | |
| Bytedance | General | Voting Member | USA | Vini Jaiswal* | |
| Sas | Premier | Voting Member | USA | Ruth Akintunde | Liz McIntosh |
| ZTE | Premier | Voting Member | China | Wei Meng | Liya Yuan |
| Adversarial Robustness Toolbox Project | Graduated Technical Project | Voting Member | USA | Beat Buesser | Kevin Eykholt |
| Angel Project | Graduated Technical Project | Voting Member | China | Jun Yao | |
| Egeria Project | Graduated Technical Project | Voting Member | UK | Mandy Chessell | Nigel Jones, David Radley, Maryna Strelchuk, Ljupcho Palashevski, Chris Grote |
| Flyte Project | Graduated Technical Project | Voting Member | USA | Ketan Umare | |
| Horovod Project | Graduated Technical Project | Voting Member | USA | Travis Addair | |
| Milvus Project | Graduated Technical Project | Voting Member | China | Xiaofan Luan | Jun Gu |
| ONNX Project | Graduated Technical Project | Voting Member | USA | Alexandre Eichenberger | Andreas Fehlner, Prasanth Pulavarthi, Jim Spohrer |
| Pyro Project | Graduated Technical Project | Voting Member | USA | Fritz Obermeyer | |
| Open Lineage Project | Graduated Technical Project | Voting Member | USA | Julien Le Dem | Michael Robinson, Mandy Chessell |
| Marquez Project | Graduated Technical Project | Voting Member | USA | Willy Lulciuc | TBD |

# Minutes approval

# Approval of April 4, 2024 Minutes

Draft minutes from the April 4, 2024, TAC call were previously distributed to the TAC members via the mailing list and available on the TAC wiki.

**Proposed Resolution:**

- That the minutes of the April 4, 2024, meeting of the Technical Advisory Council of the LF AI & Data Foundation are hereby approved.

# IREE

https://github.com/iree-org/iree/

Proposal to add in LF AI & Data at Sandbox level

# LF AI & Data proposal: IREE

Jacques Pienaar (Google) and Stella Laurenzo (AMD)
on behalf of IREE community

## vision | why we're here

The potential of ML remains unfathomed but can be discovered by providing a means to realize compelling performance on the continuously-evolving corpus of models expressed on its ever-mutating set of input forms on nascent, dynamic and heterogeneous hardware targets.

## mission | what we do

Build a suite of extensible, composable, re-targetable OSS tools for efficiently deploying popular and emerging ML model representations to popular and emerging targets ranging from embedded systems to datacenter. Biased for world class versatility with very good performance.

# Goals

- Unlock access to the hardware ecosystem from any framework
- Deliver industry-leading out-of-the-box performance and reduced TCO
- Becomes de facto toolkit for heterogeneous compute (IBM PC / LLVM analogy)
- Build an enduring compiler platform for the ML community to use and extend
- Increase the speed of ML innovation for all ML practitioners

# IREE in a nutshell

IREE is a toolkit that's designed to simplify the deployment of ML programs to a range of architectures and power regimes.

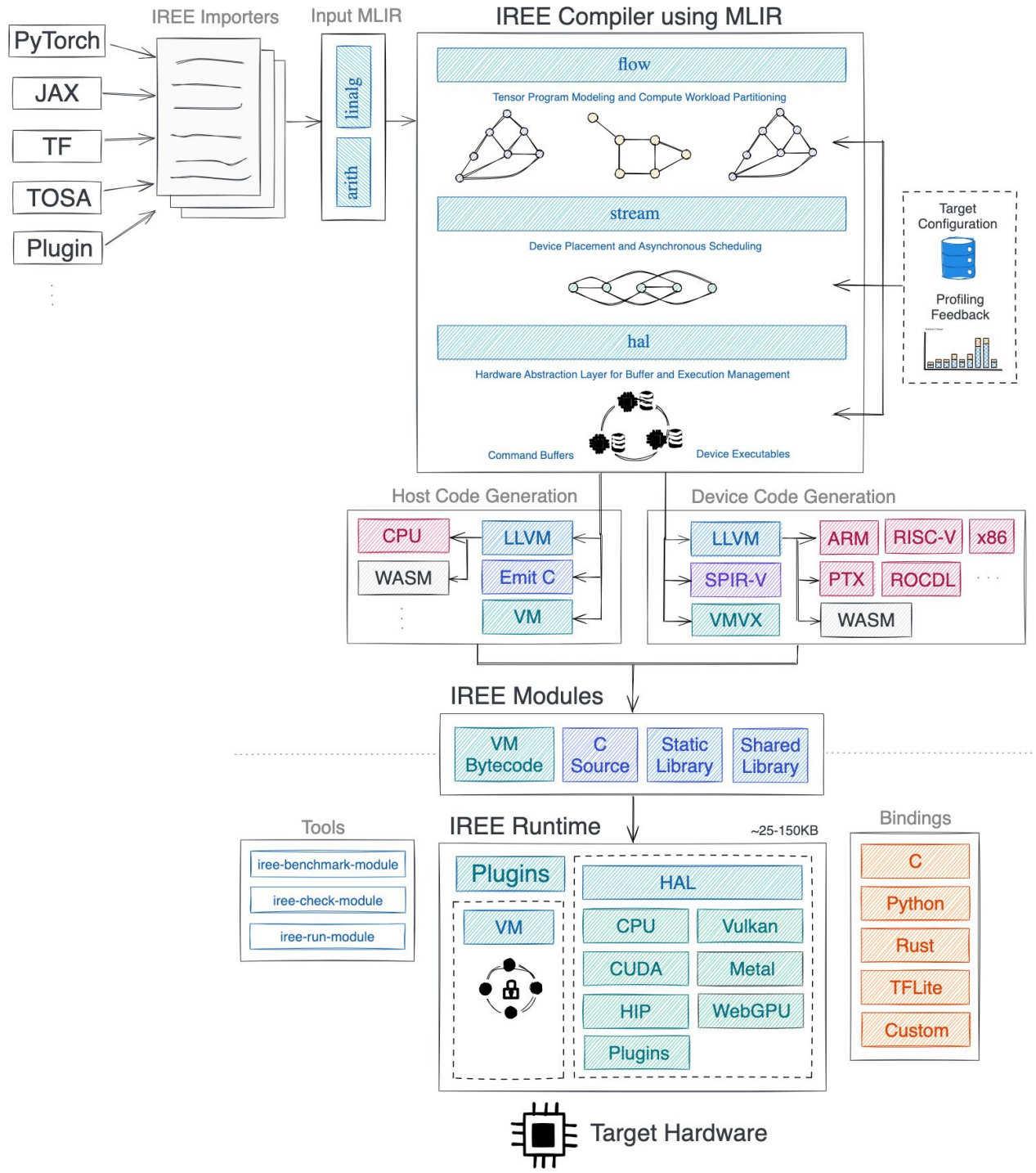
It compiles ML programs into deployable artifacts comprised of:

- Device-specific code representing the "kernels" of the program (as expressed in x86/Arm machine code, SPIR-V, PTX, etc.).
- Device-agnostic code that expresses the dataflow between these kernels.

A lightweight runtime executes these artifacts, providing flexible deployment and optimal execution.

# Architecture

- Extensible architecture with plugins for
  - Input/importers
  - Compiler backends
  - Devices (behind HAL)
- Build on top of efficient runtime
- Custom language bindings



# IREE Workflow

← Frameworks/integrations

## Compiler → Artifacts → Runtime

- Native integration with PyTorch via Dynamo/FX.
- Accepts StableHLO but also upstream MLIR dialects using tensors like arith, math, linalg, and TOSA.
- Importers for TensorFlow SavedModels, TFLite Flatbuffers, XLA protos, etc
- Partitions programs into host- and device-side logic and schedules execution
- Bakes out artifacts (IREE VM FlatBuffers, EmitC .c files, etc)

- Shared library-like modules
- Link against IREE built-in modules like the HAL or other user modules
- Contains the partitioned host program and embedded device programs ("executables") in various formats, like a fat binary
- Logic for device selection, branching paths for different device types, etc - *it's just code*

- Dynamic linker for loading modules into isolated contexts
- Optional bytecode VM for running host programs
- HAL (Hardware Abstraction Layer) API and backends (CPU, Vulkan, CUDA, etc)
- Utility APIs for ease of use (session-style invocation)
- Low-level; designed to be wrapped in bindings for languages/environments

# Target Platforms supported

Primary Platforms with committed investment:

- AMD HIP
- AMD XDNA (via in-progress plugin)
- CPUs generally, with a focus on x86_64 and aarch64
- NVIDIA CUDA
- Vulkan/SPIR-V for GPUs generally

Experimental or community projects:

- Apple Metal (via SPIR-V interop)
- RISC-V CPUs and ecosystem
- WebAssembly
- WebGPU

# Framework Integration

Full Featured:

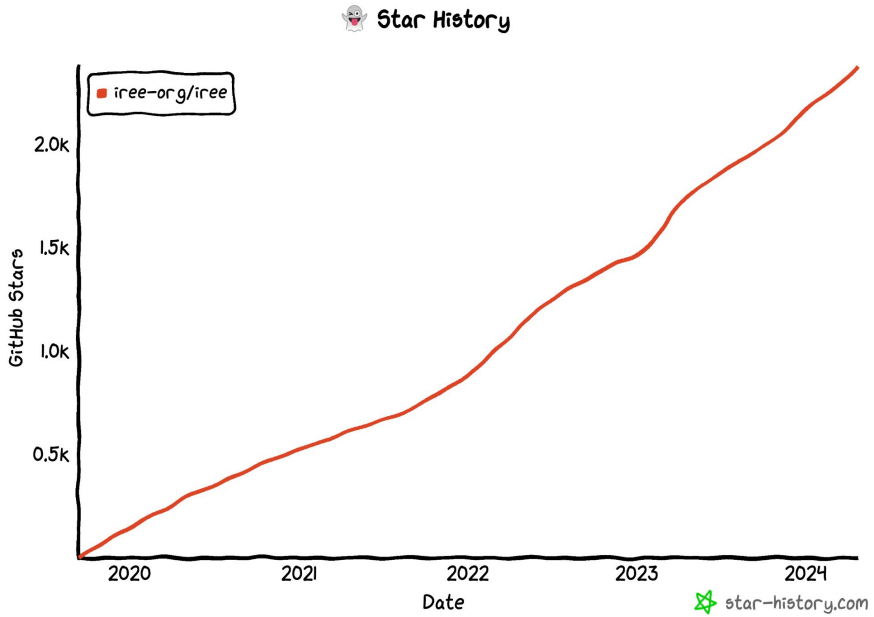
- Torch
  - IR compatibility via co-investment in torch-mlir
  - Full framework integration via IREE Turbine (née SHARK Turbine – repo migration in progress)
- ONNX
  - IR compatibility via co-investment in torch-mlir ONNX compatibility layer
  - Full tooling and EP prototyped and under development

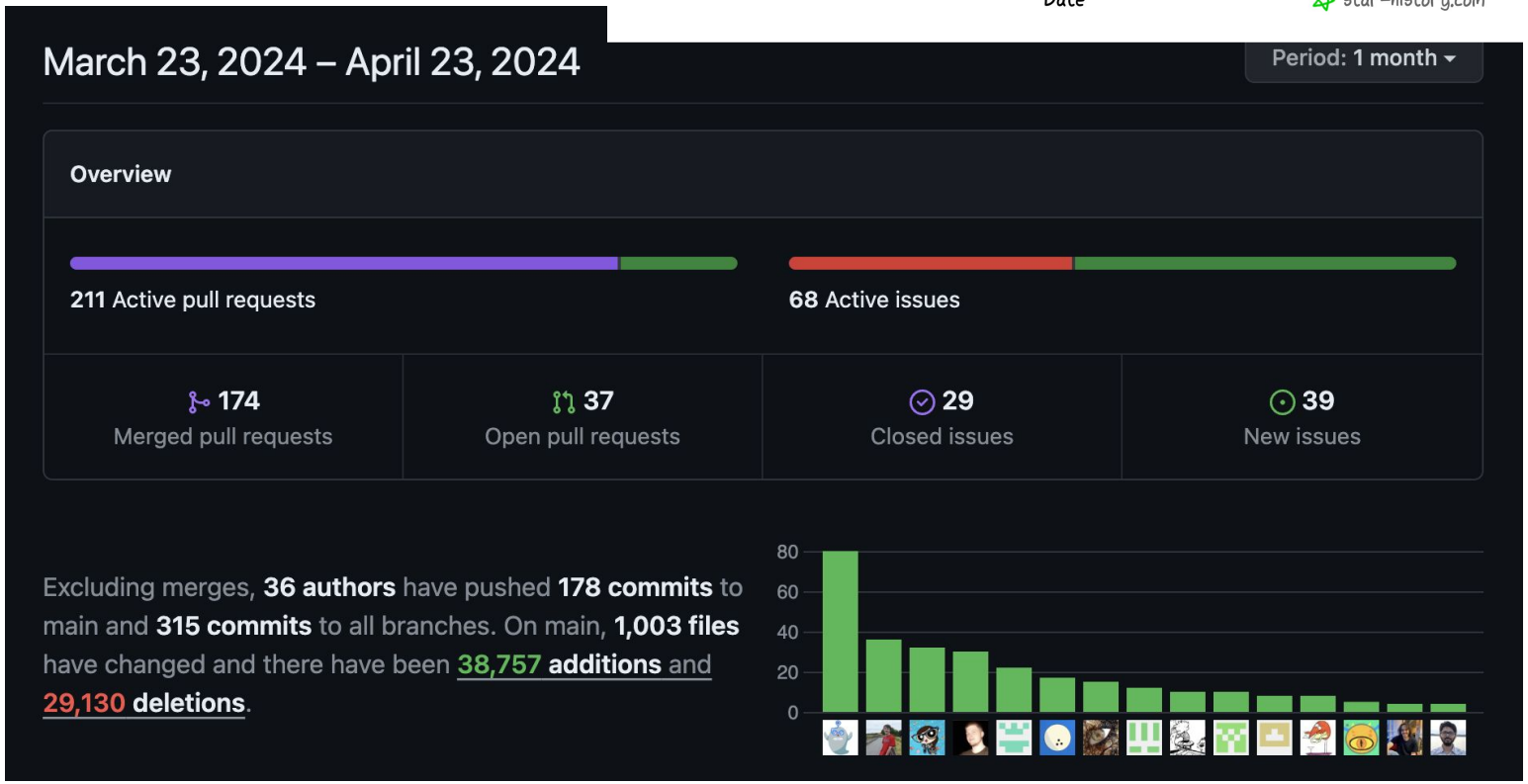
Compiler IR Compatibility:

- JAX (via StableHLO and PJRT)
- TFLite (via TOSA export)
- TensorFlow (via StableHLO)

# Community

- Open source since Sep 18, 2019
- 2.4K Stars
- 190 Contributors
- Well into 3rd-4th generation contributors that have spanned companies over the years
- Active community on Discord, mailing list, etc



# Why donate IREE?

Neutral holding ground

- Vendor-neutral, equal collaboration

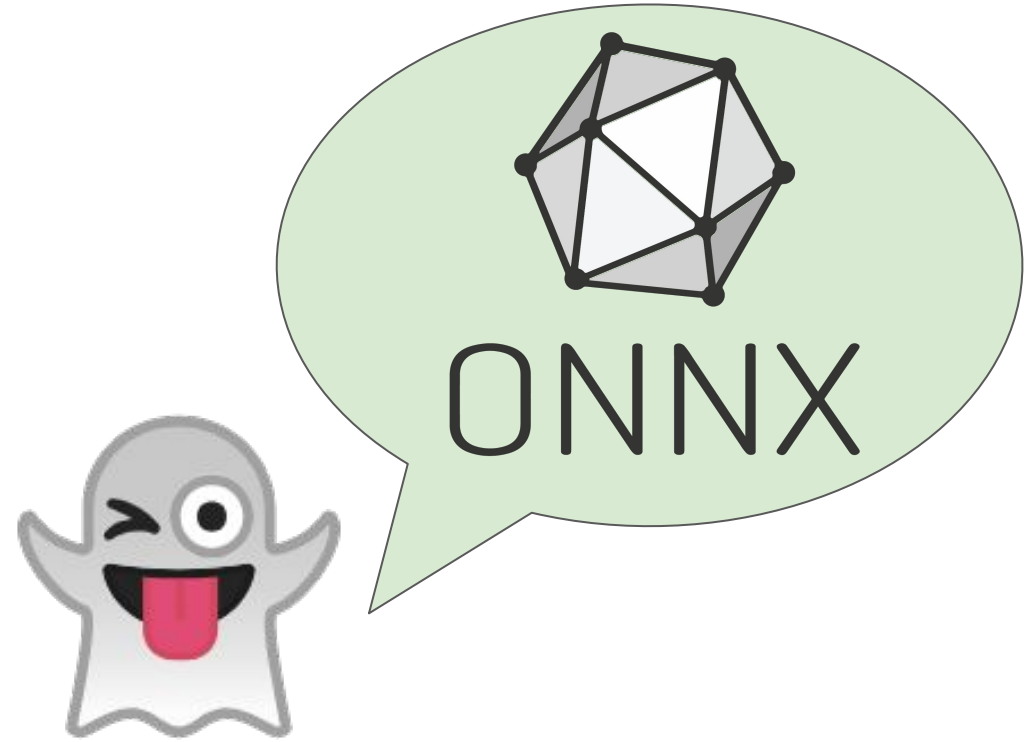
Growing community

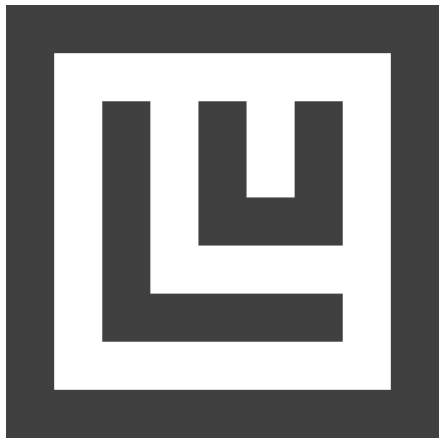
- Increase contributors by converting new & existing users
- Opportunities to collaborate with other hosted projects
- Increase users by broader outreach through the foundation

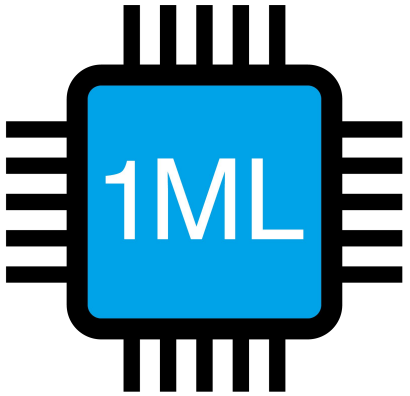
Open Governance model

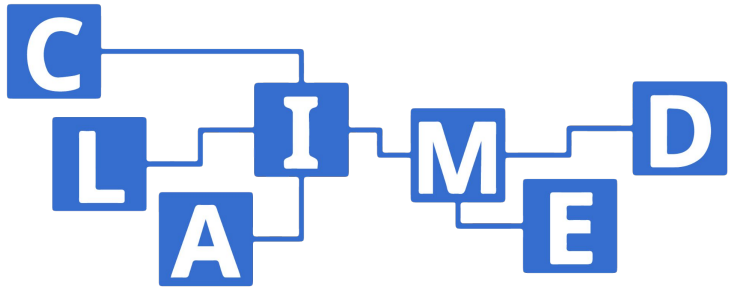
- Open governance + open source license
- Distills trust in the running & management of the project
- Neutral management of projects' assets by the foundation

# Possible collaborations with existing projects






ShaderNN

# Summary

- IREE focuses on scalable ML model deployment
- Connectivity bits across multiple input frameworks to multiple backends
- Toolbox to enable exploiting from application specialized deployment to hardware specialization
- Open community, using & driving standards to target many platforms

# Request to TAC

We'd like to contribute IREE to Linux Foundation's AI & Data foundation as sandbox project

We would like to work with you to address any concerns

# Thanks!

# Questions / Discussion on IREE

# Proposed Resolution

**Proposed Resolution:**

- That the Technical Advisory Council of LF AI & Data Foundation approves the IREE project at the Sandbox level

# Next steps

- The LF AI & Data team will follow up with the AMD on onboarding the project and integrating it with our services.
- Once completed, an announcement will follow.

# Open Discussion

# Call for Action

- **Nominate representatives** to our re-launched Outreach Committee (OC)
  - 1 appointed voting representative from each Premier Member
  - 1 appointed non-voting representative from each General Member

LF AI & DATA

OUTREACH COMMITTEE

# Upcoming TAC Meetings

| | |
|---|---|
| 05/16 | Delta Lake requesting addition to LF AI and Data |
| 05/30 | TRULENS requesting incubation |
| 06/13 | *ML Spec (Tentative)* |
| TBD | Invited talks: Harvard + Horovod project + Open Voice |

If you have a topic idea or agenda item, please send agenda topic requests to tac-general@lists.lfaidata.foundation

# TAC Meeting Details

TAC Biweekly Meeting LF AI & Data

Ways to join meeting:

1. Join from PC, Mac, iPad, or Android

https://zoom-lfx.platform.linuxfoundation.org/meeting/95332329356?password=c708f2ee-fb78-4a12-91a3-47daa19b708f

2. Join via audio

One tap mobile:

US: +12532158782,,95332329356# or +13462487799,,95332329356

Or dial:

US: +1 253 215 8782 or +1 346 248 7799 or +1 669 900 6833 or +1 301 715 8592 or +1 312 626 6799 or +1 646 374 8656 or 877 369 0926 (Toll Free) or 855 880 1246 (Toll Free)

Canada: +1 647 374 4685 or +1 647 558 0588 or +1 778 907 2071 or +1 204 272 7920 or +1 438 809 7799 or +1 587 328 1099 or 855 703 8985 (Toll Free)

Meeting ID: 95332329356

Meeting Passcode: 040721

# Legal Notice

- The Linux Foundation, The Linux Foundation logos, and other marks that may be used herein are owned by The Linux Foundation or its affiliated entities, and are subject to The Linux Foundation's Trademark Usage Policy at https://www.linuxfoundation.org/trademark-usage, as may be modified from time to time.
- Linux is a registered trademark of Linus Torvalds. Please see the Linux Mark Institute's trademark usage page at https://lmi.linuxfoundation.org for details regarding use of this trademark.
- Some marks that may be used herein are owned by projects operating as separately incorporated entities managed by The Linux Foundation, and have their own trademarks, policies and usage guidelines.
- TWITTER, TWEET, RETWEET and the Twitter logo are trademarks of Twitter, Inc. or its affiliates.
- Facebook and the "f" logo are trademarks of Facebook or its affiliates.
- LinkedIn, the LinkedIn logo, the IN logo and InMail are registered trademarks or trademarks of LinkedIn Corporation and its affiliates in the United States and/or other countries.
- YouTube and the YouTube icon are trademarks of YouTube or its affiliates.
- All other trademarks are the property of their respective owners. Use of such marks herein does not represent affiliation with or authorization, sponsorship or approval by such owners unless otherwise expressly specified.
- The Linux Foundation is subject to other policies, including without limitation its Privacy Policy at https://www.linuxfoundation.org/privacy and its Antitrust Policy at https://www.linuxfoundation.org/antitrust-policy. each as may be modified from time to time. More information about The Linux Foundation's policies is available at https://www.linuxfoundation.org.
- Please email legal@linuxfoundation.org with any questions about The Linux Foundation's policies or the notices set forth on this slide.

LF AI & DATA