Accelerator Integration Working Group Meeting Notes

2025-12-04 Next Meeting

Meeting: 2025-12-04 10:00 AM (UTC+8)(Convert to your timezone)
Join: Click here /Meeting Number: 827 3042 5910 / Meeting pass: Sa2XYA

2025-11-20 Meeting

Meeting: 2025-11-20 10:00 AM (UTC+8)(Convert to your timezone)
Join: Click here /Meeting Number: 827 3042 5910 / Meeting pass: Sa2XYA
Attendees

- Dmitry Rogozhkin (@dvrogozh, Intel)
- Rabi
- Hunter Carlisle
- Joshua Rosenkranz
- Maen Suleiman (AWS)
- Jiawei LI (@fffrog, Huawei)
- Zesheng Zong (@zeshengzong, Huawei)
- Jiahao Chen (@can-gaa-hou, Huawei)

Agenda

[Jiahao Chen] OpenReg Current Progress

Introducing current implementation and progress of OpenReg, more features will be gradually expanded and tracked here.

Discuss:

- How does autoload implement in OpenReg?
  - Here's a PR to show how to enable autoload.
- Why is OpenReg placed in the cpp_extension folder?
  - The module was initially planned to be placed in this directory and has not been adjusted. Currently, OpenReg has not only a C++ backend but also a Python frontend.
- Would OpenReg test other backends like CUDA?
  - OpenReg used to test the functionality of backend integration via PrivateUse1 DispatchKey.

[Dmitry Rogozhkin] PyTorch Compute Platform Quality Criteria Proposal

See: RFC, PR

Update:

- Added “Documentation update guideline” section to the PR with the proposal on how to add brand new compute platform and promote/denote existing platform
- Related to PR comment. **Proposal** is to:
    - 1) Clarify that sw/hw version changes correspond to different “flavors” of *the same* platform unless there is a major breaking changes somewhere
    - 2) Drop OS requirement items (platform tested on Windows +1, Linux +1, etc.) from the criteria scoring. Motivation is that we might wish to consider OS as a differentiating factor for compute platforms due to potential significant differences in driver stack implementations and different user experiences due to that.
- Filed RFC (XPU) with the request to add XPU to the Getting Started page. This RFC contains assessment for XPU following proposed criteria
- Category naming proposals:

Category 1 name	Category 2 name	Category 3 name	Comments
Stable	Unstalbe	Engineering	That’s *current* naming
General Availability (GA)	Preview	Development	shauheen
Stable	Preview	Development	shauheen
Stage 1	Stage 2	Stage 3	
Stable	Beta	Alpha	

Discuss:

- Need to clarify requirements and standards for the full life cycle.
    - [Dmitry Rogozhkin] Already has a **Documentation update guideline** section in PR, and needs more review opinion.
- Defining the support matrix and sorting out standards(Fully support, partial support, not support).
    - [Dmitry Rogozhkin] Marked as a following task after the criteria of classification is clear.

Open Discussion

- Call for roadmap of Accelerator Integration WG
    - [Zesheng Zong] A roadmap for the first quarter of the working group will be initiated. Ideas will be collected via a form, and several meetings will be held to finalize it.
- Migrating meeting calendar to LFX calendar
    - [Zesheng Zong] The meeting will be moved to the Linux calendar, and the link will be sent out before the next meeting.

2025-11-06 Meeting

Meeting: 2025-11-06 10:00 AM (UTC+8)(Convert to your timezone)
Join: Click here /Meeting Number: 827 3042 5910 / Meeting pass: Sa2XYA
Attendees

- Cyril Bortolato (cborto@amazon.com, AWS Neuron)
- Ashok Emani (Intel)
- Dmitry Rogozhkin (@dvrogozh, Intel)
- Jerome Anand (@jeromean, Intel)
- Fan Mo (@fmo-mt)
- Zesheng Zong (@zeshengzong, Huawei)
- Jiahao Chen (@can-gaa-hou, Huawei)
- Zhi Chen

Agenda

PyTorch Compute Platform Quality Criteria Proposal

[Dmitry Rogozhkin] Introduced the proposal to establish quality criteria for compute platforms. (link)

To discuss:

1. Items to close to finalize current version of proposal (i.e. to close doc and switch to edit/comment PR):
    a. Category naming and standard naming (guide standard or criteria)
        i. Naming should balance professionalism and user-friendliness to avoid negative impacts.
        ii. List all options and launch a vote and submit voting results to TAC for final recommendations.
2. Process to submit evidence of passing criteria for the platform?
    a. Hardware Availability: Vendors are not required to provide hardware to the community, sharing a testing report is sufficient.
    b. CI test reports must be publicly accessible. Can use torchbench to track performance regressions and vendors should submit performance/regression test reports.
3. Process to add/exclude platforms which match/don't match the criteria? (Not discussed)
4. Which organization should be in charge of judging results? (Not discussed)

Next steps:

1. [Dmitry Rogozhkin] Create RFC issue for Compute Platform Quality Criteria proposal and migrate proposal content to github. (RFC)
2. [Dmitry Rogozhkin] Create PR to add Compute Platform Quality Criteria to PyTorch documentation. (PR)
3. [Zesheng Zong] Organize a draft run with several vendors, sync the result to TAC and request review.
4. [All] Follow up on Slack to resolve questions and coordinate next steps, especially regarding RFC drafting and test data collection.

2025-10-16 Meeting

Meeting: 2025-10-16 10:00 AM (UTC+8)(Convert to your timezone)
Join: Click here /Meeting Number: 827 3042 5910 / Meeting pass: Sa2XYA
Attendees

- Zesheng Zong (@zeshengzong, Huawei)
- Jiawei LI (@fffrog, Huawei)
- Han Qi (@qihqi, Google)

Agenda

Python Level Registration for PyTorch

[Han Qi] Using JAX as backend for PyTorch, also met another developer who wants to use Modular as backend as well.

- Initiate a RFC this week.
- need a place to practice and discuss how Python level register API looks like
    - **[Zesheng Zong]** Create sub-module in pytorch-fdn/accelerator-integration-wg repo.

[Jiawei Li] There's a gap using pure Python level registration, need to combine cpp and Python, also introduced usage and implementation of device guard in PyTorch.

2025-10-02 Meeting is cancelled due to the holiday

2025-09-25 Temporary Meeting

Attendees

- Zesheng Zong (@zeshengzong, Huawei)
- Jiawei LI (@fffrog, Huawei)
- Han Qi (@qihqi, Google)
- Fan Mo (Moore Threads)

Agenda

The design of torch at the Layout level is more inclined towards devices with a CUDA-like architecture, such as in the handling of aspects like stride.This design is not sufficiently memory-access friendly for DSA (Domain-Specific Architecture) devices, and the framework lacks the ability to perceive the memory access characteristics of underlying hardware.

Conclusions:

- The Layout design of Torch needs to be optimized to better support the memory access requirements of DSA devices.
- The backend extension capabilities of PyTorch should be strengthened to support more types of hardware and computing libraries.

2025-09-18 Meeting

Meeting: 2025-09-18 10:00 AM (UTC+8)(Convert to your timezone)
Join: Click here /Meeting Number: 827 3042 5910 / Meeting pass: Sa2XYA

Attendees

- Zesheng Zong (@zeshengzong, Huawei)
- Jiawei LI (@fffrog, Huawei)
- Jerome Anand(@jeromean, Intel)

Agenda

1. [zeshengzong] CI workflow updates

- Integrated torchtune with CI workflow
- Working on Integrating torchtitan, but need ability to skip tests for function not supported yet

2. [fffrog] OpenReg updates

- Runtime support for device management, streams, events, and DeviceGuard.
- Operator registration in different use cases.
- AMP and Autoload are both supported.
- Developer guide completed for operator registration, AMP, and Autoload.

3. PyTorch backend discussion (Skip)

- PR #157859

4. [zeshengzong, jeromean] PyTorch Conference Speech

- Discuss speech syllabus
- Assign tasks owners and time arrangement

Open Questions

2025-09-04 Meeting is cancelled

2025-08-21 Meeting

Meeting: 2025-08-21 10:00 AM (UTC+8)(Convert to your timezone)
Join: Click here /Meeting Number: 827 3042 5910 / Meeting pass: Sa2XYA

Attendees

- Zesheng Zong (@zeshengzong, Huawei)
- Jiawei LI (@fffrog, Huawei)
- Lei Wang (FlagGems, BAAI)

Agenda

1. CI workflow

- Introduce FlagGems current workflows and implementation
- Introduce PyTorch-fdn/accelerator-integration-wg workflows
- Discuss integrating FlagGems E2E CI test into working group repo.

Open Questions

2025-08-07 Meeting

Meeting: 2025-08-07 10:00 AM (UTC+8)(Convert to your timezone)
Join: Click here /Meeting Number: 827 3042 5910 / Meeting pass: Sa2XYA

Slides: Accelerator Integration Working Group Meeting 20250807

Attendees

- Zesheng Zong (@zeshengzong, Huawei)
- Jiawei LI (@fffrog, Huawei)
- Chunlei Men (FlagGems, BAAI)
- Jinjie Liu (FlagGems, BAAI)

Agenda

1. CI workflow

- Progress
  - Integration torchtune in CI workflow for E2E test
  - Draft an outline of Accelerator Integration Quality Guidance

2. OpenReg

- Progress
  - Submit RFC of enhancing Accelerator Integration
  - Create Roadmap of optimize OpenReg module as reference for testing and document
  - Draft first version of developer guide , OpenReg OSX/Windows support and refactor (#158644, #159441, #159640)

3. FlagGems

- Introduce
  - Consider switching their test-suit to torchbench aline test and verification with community, can also combine with working group upstream CI workflow.
  - FlagGems wrapper level code has device specific logic that can be improved, which may benefit from generic device API (torch.accelerator) in PyTorch core.

Open Discussion

1. Do we need more meetings? Current meeting schedule: Thu. of the first week in a month.

- Change to Bi-week meeting, next on 21st, Aug.

2025-07-03 Meeting

Meeting: 2025-07-03 10:00 AM (UTC+8)(Convert to your timezone)
Join: Click here /Meeting Number: 827 3042 5910 / Meeting pass: Sa2XYA

Slides: Accelerator Integration Working Group Meeting 20250703

Attendees

- Zesheng Zong (@zeshengzong, Huawei)
- Jiawei LI (@fffrog, Huawei)
- Jerome Anand(@jeromean, Intel)

Agenda

1. WG Status Updates

- From “OOTA Exploratory Group” Convert to “Accelerator Integration Working Group”
- Setup repo rule for PR (need at least 1 approve, not mandatory CI pass for now)
- Setup group project panel pytorch-fdn/Projects/accelerator-integration-wg

2. CI workflow

- Current Status
  - Pause Gaudi workflow for data center migration
  - Fix Ascend workflow for host upgrading
- Discuss and setup high priority tasks in the project
  - Define Quality Standards of Accelerator Integration Process (#28)
  - Rewrite readme content to show working group guidance (#22)
  - Support displaying multiple Test Results in dashboard (#21)

3. OpenReg

- Introduce the background and current advantages and disadvantages of OpenReg
- Proposal a new design by C++ with reference to torch_xla and Ascend NPU
  - Keeping consistent with the real integration way can more effectively ensure stable quality
  - Provide E2E docs and tutorials matching the above code

4. Device Generic Improvements

- Inspired by PyTorch Roadmap 2025H1, migrate device specific code in repo to torch.accelerator API, like torchtune, torchao etc., reduce workload of new accelerator integration.

Open Discussion

1. Do we need more meetings? Current meeting schedule: Thu. of the first week in a month.

- Keep it once a month for now, adjust frequency based on tasks progress.

2. Other valuable tasks need to be tracked by WG?

- Acquire more opinions and suggestions about group working content from others.