This webinar will begin shortly at
19:00 CET | 13:00 EST | 10:00 PDT

Accelerating Machine Learning with OpenCL

May 11, 2022
How to Participate

Speaker Questions
During the presentations, please submit speaker questions using the Zoom Q&A button (not the chat button). At the end of the talk, our moderator will put as many questions as possible to the speakers.

General Questions and Comments
Please use the Zoom Chat feature for logistical questions or if you are having issues with Zoom.

Recording
We are recording this webinar and will publicly post the link at the Forum Home Page.

Survey
To help us design future Khronos ML Forum events, we appreciate you completing the short survey form that we will distribute after the session.
Accelerating Machine Learning with OpenCL

Khronos ML Forum
Neil Trevett, Khronos

Qualcomm Extensions for Advancing Machine Learning Acceleration
Balaji Calidas, Qualcomm

A Case Study on OpenCL vs GPU Assembly for Machine Learning Performance
Roy Oursler, Intel

Q&A Panel Moderator
Kevin Petit, Arm
Introduction to the Khronos Machine Learning Forum

Neil Trevett
Khronos President
Khronos Open Standards Mission

Open, royalty-free interoperability standards to harness the power of GPU, XR and multiprocessor hardware

3D graphics, augmented and virtual reality, parallel programming, inferencing and vision acceleration

Non-profit, member-driven standards organization, open to any company

Proven multi-company governance and Intellectual Property Framework

Established in 2000

~200 Members ~ 40% US, 30% Europe, 30% Asia

Over 180 members worldwide

Any organization is welcome to join
Increasing industry interest in parallel compute acceleration to combat the ‘End of Moore’s Law’
## OpenCL and Machine Learning

### Machine Learning Compilers

<table>
<thead>
<tr>
<th>Import Formats</th>
<th>Caffe, Keras, MXNet, ONNX</th>
<th>TensorFlow Graph, MXNet, PaddlePaddle, Keras, ONNX</th>
<th>PyTorch, ONNX</th>
<th>TensorFlow Graph, PyTorch, ONNX</th>
</tr>
</thead>
<tbody>
<tr>
<td>Front-end / IR</td>
<td>NNVM / Relay IR</td>
<td>nGraph / Stripe IR</td>
<td>Glow Core / Glow IR</td>
<td>XLA HLO</td>
</tr>
<tr>
<td>Output</td>
<td>OpenCL, LLVM, CUDA, Metal</td>
<td>OpenCL, LLVM, CUDA</td>
<td>OpenCL, LLVM</td>
<td>OpenCL, LLVM, XLA IR, TensorFlow Lite / NNAPI (inc. HW accel)</td>
</tr>
</tbody>
</table>

### Common Steps
1. Import Trained Network Description
2. Graph-level optimizations e.g., node fusion, node lowering and memory tiling
3. Decompose to primitive instructions and emit programs for accelerated run-times

### Machine Learning Compilers and Frameworks using OpenCL Acceleration

**Inferencing Libraries and Frameworks**
- Alibaba MNN
- Arm Compute Library
- Baidu PaddlePaddle/Paddle-Lite
- Berkeley Caffe
- Intel cLDNN and OpenVINO

**Google TensorFlow and NNAPI**
- SYCL-DNN
- Synopsis MetaWare EV
- Texas Instruments DL Library (TIDL)
- VeriSilicon Acuity
- Xiaomi Mace

**Embedded NN Compilers**
- CEVA Deep Neural Network (CDNN)
- Cadence Xtensa
- Neural Network Compiler (XNNC)

**Inferencing Libraries and Frameworks**
- Alibaba MNN
- Arm Compute Library
- Baidu PaddlePaddle/Paddle-Lite
- Berkeley Caffe
- Intel cLDNN and OpenVINO
# OpenCL Extension Pipeline

A significant percentage of OpenCL extensions are used for Machine Learning acceleration

### New OpenCL KHR extensions shipped since IWOCL 2021

<table>
<thead>
<tr>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subgroup rotate extension for efficient data exchange among work-items</td>
</tr>
<tr>
<td>Workgroup Uniform Arithmetic for new work-group scan and reduction operators</td>
</tr>
<tr>
<td>Command Buffers Record and Replay (Provisional)</td>
</tr>
<tr>
<td>Asynchronous DMA</td>
</tr>
<tr>
<td>Expect Assume Hints</td>
</tr>
<tr>
<td>Integer Dot Product</td>
</tr>
<tr>
<td>External memory objects and semaphores (Provisional)</td>
</tr>
</tbody>
</table>

### EXT and Vendor Extension Pipeline

<table>
<thead>
<tr>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generalized Image from buffer (EXT)</td>
</tr>
<tr>
<td>Unified Shared Memory (EXT)</td>
</tr>
<tr>
<td>Floating Point Atomics (EXT)</td>
</tr>
<tr>
<td>Cooperative Matrices (EXT)</td>
</tr>
<tr>
<td>Machine Learning Operations (Qualcomm)</td>
</tr>
<tr>
<td>...</td>
</tr>
</tbody>
</table>

This work is licensed under a Creative Commons Attribution 4.0 International License
Khronos Machine Learning Forum

Productive ongoing communication and cooperation on ML Acceleration
... between Machine Learning hardware and software communities

Forum is free to join, no NDA or IP commitments
Dedicated meetings, email and slack channels for group communication

Use cases, requirements, updates and presentations

Machine Learning Community

Khronos Hardware and Software Members

API guidelines, updates and roadmaps
Machine Learning Forum Meeting Series

Public Meetings

- **Khronos ML Summit**
  - October 2021
  - Session #1 Video
  - Session #2 Video

- **ML Summit Response Session**
  - January 2022

- **Vulkan ML Webinar**
  - 5th May 2022
  - Slides and videos

- **OpenCL ML Webinar**
  - You are here!

- **SYCL and OpenVX Webinars**
  - Being scheduled

**Forum Member Meetings** will start in July 2022

Input and requests for specific topics welcome!

All the information you need to join!

[https://www.khronos.org/machine-learning](https://www.khronos.org/machine-learning)

This work is licensed under a Creative Commons Attribution 4.0 International License

© The Khronos® Group Inc. 2022 - Page 10
Accelerating Machine Learning with OpenCL

Khronos ML Forum
Neil Trevett, Khronos

Qualcomm Extensions for Advancing Machine Learning Acceleration
Balaji Calidas, Qualcomm

A Case Study on OpenCL vs GPU Assembly for Machine Learning Performance
Roy Oursler, Intel

Q&A Panel Moderator
Kevin Petit, Arm
Ask the Experts

Use the Zoom Q&A feature to ask your questions
Thank you!

https://www.khronos.org/events/accelerating-machine-learning-with-opencl