This webinar will begin shortly at

19:00 CET | 13:00 EST | 10:00 PDT

KHRONOS® GROUP

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

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

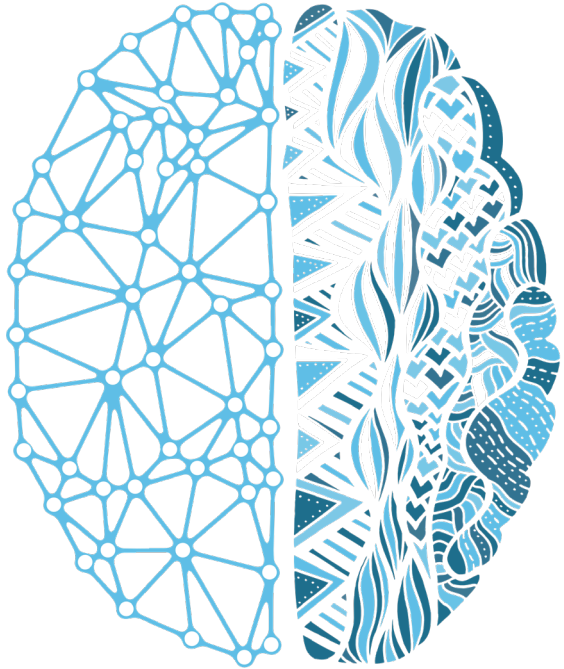
Qualcomm Extensions for Advancing Machine Learning Acceleration
Balaji Calidas, Qualcomm

Q&A Panel Moderator
Kevin Petit, Arm

KHRONOS GROUP

Introduction to the Khronos Machine Learning Forum

Neil Trevett
Khronos President



Vulkan SPIR OpenCL SYCL OpenVX NNEF

Khronos Open Standards Mission

Founded in 2000
~200 Members ~ 40% US, 30% Europe, 30% Asia

KHRONOS GROUP

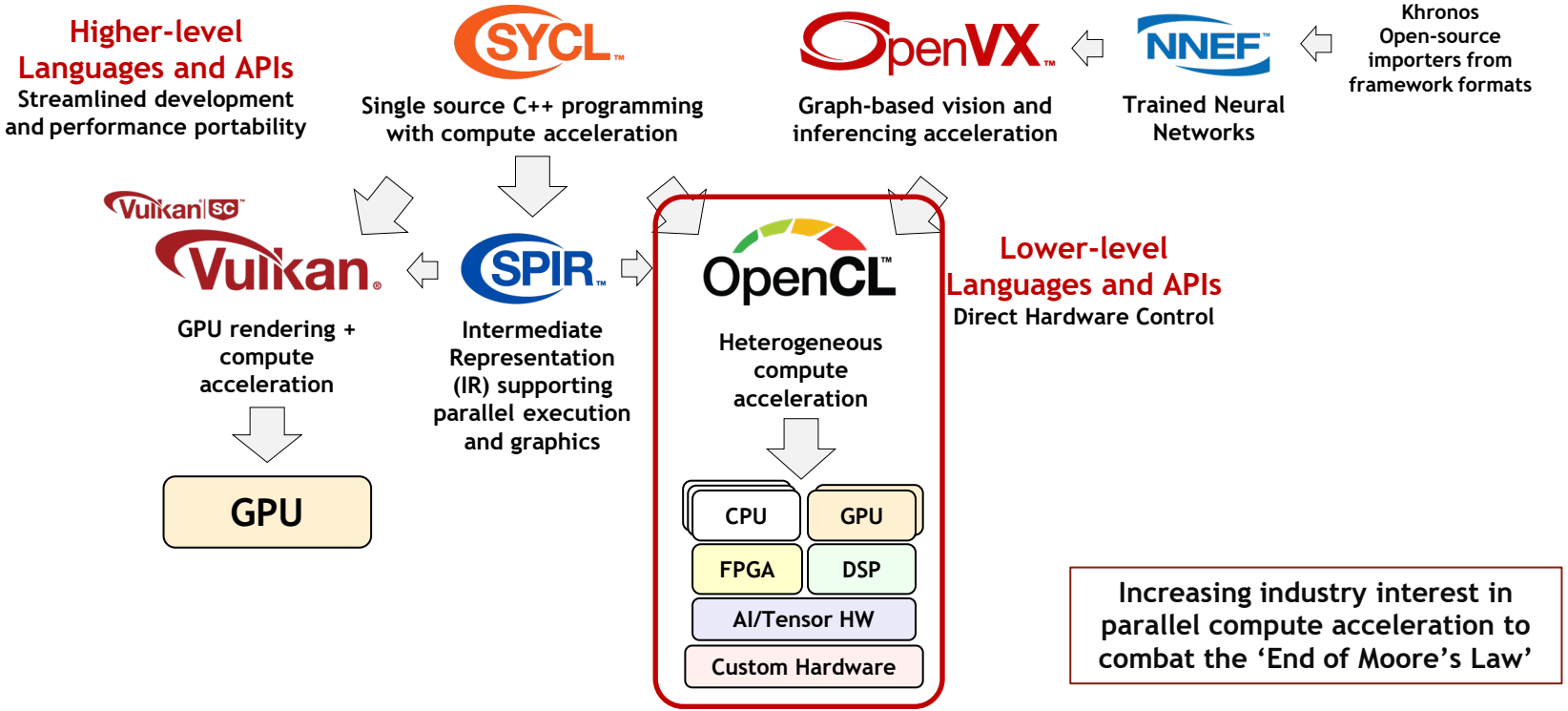
Open, royalty-free interoperability standards to harness the power of GPU, XR and multiprocessor hardware

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

Khronos Compute Acceleration Standards



Increasing industry interest in parallel compute acceleration to combat the 'End of Moore's Law'

OpenCL and Machine Learning

Machine Learning Compilers	tvm.ai	plaidML	GLOW	XLA
	amazon	intel		Google
Import Formats	Caffe, Keras, MXNet, ONNX	TensorFlow Graph, MXNet, PaddlePaddle, Keras, ONNX	PyTorch, ONNX	TensorFlow Graph, PyTorch, ONNX
Front-end / IR	NNVM / Relay IR	nGraph / Stripe IR	Glow Core / Glow IR	XLA HLO MLIR
Output	OpenCL, LLVM, CUDA, Metal	OpenCL, LLVM, CUDA	OpenCL LLVM	OpenCL LLVM, TPU IR, XLA IR TensorFlow Lite / NNAPI (inc. HW accel)
				SPIR Vulkan

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)

Acuity ML/AI Software Foundation · Caffe · MACE Mobile AI Compute Engine · MNN Mobile Neural Network · OPENVINO™ · 飞桨 PaddlePaddle

OpenCL Extension Pipeline

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

New OpenCL KHR extensions shipped since IWOCL 2021

- **Subgroup rotate extension for efficient data exchange among work-items**
- Workgroup Uniform Arithmetic for new work-group scan and reduction operators
- **Command Buffers Record and Replay (Provisional)**
- Asynchronous DMA
- Expect Assume Hints
- **Integer Dot Product**
- External memory objects and semaphores (Provisional)

EXT and Vendor Extension Pipeline

- **Generalized Image from buffer (EXT)**
- Unified Shared Memory (EXT)
- **Floating Point Atomics (EXT)**
- **Cooperative Matrices (EXT)**
- **Machine Learning Operations (Qualcomm)**

....

Khronos Machine Learning Forum

Productive ongoing communication and cooperation on ML Acceleration

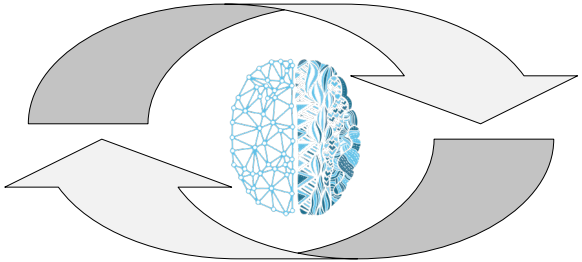
… between Machine Learning *hardware* and *software* communities

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

Presentations from leading ML practitioners on progress AND acceleration API pain points

Feedback and actions from Khronos working groups on Summit requests and ideas

Public Meetings

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

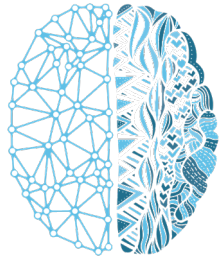
ML Summit Response Session
January 2022

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

Accelerating Machine Learning with OpenCL

Khronos ML Forum
Neil Trevett, Khronos

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

Qualcomm Extensions for Advancing Machine Learning Acceleration
Balaji Calidas, Qualcomm

Q&A Panel Moderator
Kevin Petit, Arm

Accelerating Machine Learning with OpenCL
11 May 2022

Ask the Experts

Use the Zoom Q&A feature to ask your questions

IWOCL & SYCLcon 2022

Accelerating Machine Learning with OpenCL
11 May 2022

Thank you!

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