Deploying Accelerated ML & AI:
The Role of Khronos Open Standards

Neil Trevett

President

The Khronos Group

KHRONOS GROUP

Khronos Connects Software to Silicon

KHRONOS GROUP

Non-profit Standards Consortium creating open, royalty-free standards

Focused on runtime APIs and file formats for 3D, XR, AI, vision, parallel compute acceleration

Member-driven, open to any company

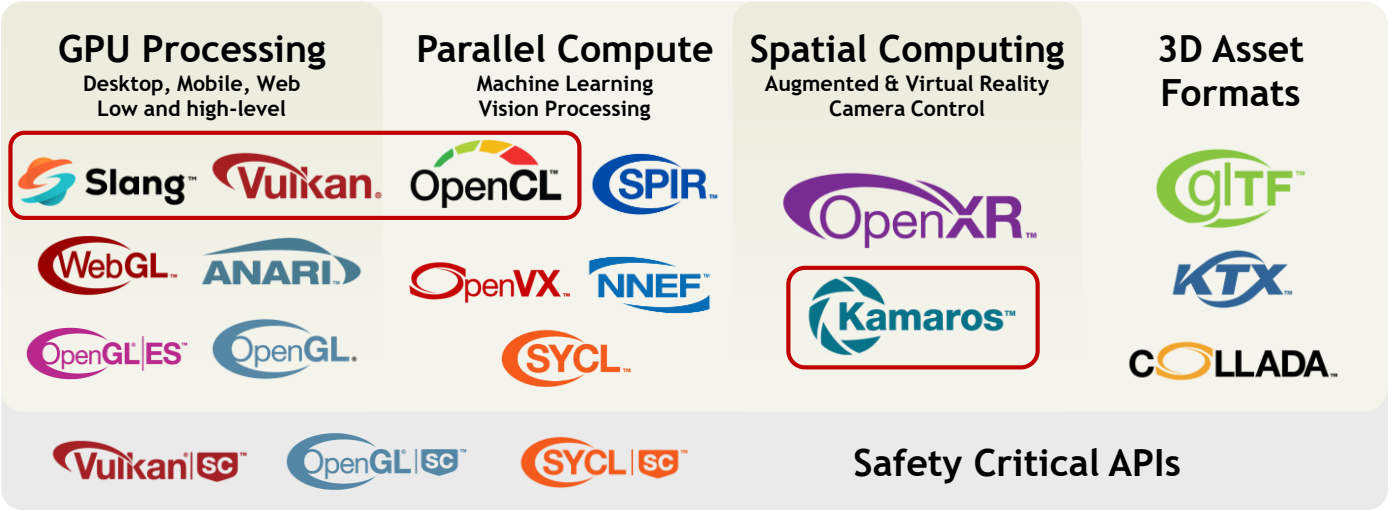
Founded in 2000

~ 160 Members | ~ 40% US, 30% Europe, 30% Asia

ISO/IEC JTC 1 PAS Submitter

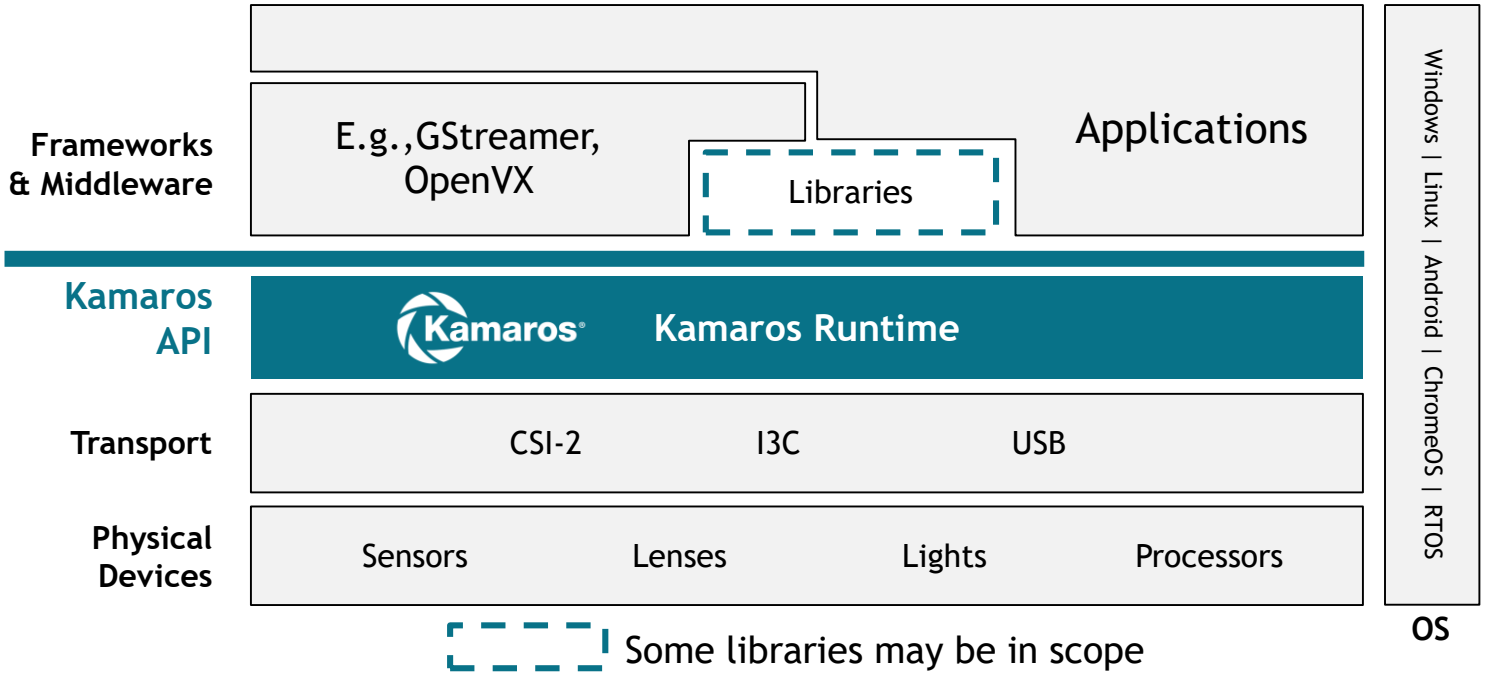
Topics

- Update on the Kamaros camera control API currently in development
- New Slang GPU shading language with machine learning capabilities
- Updates on OpenCL and Vulkan for machine learning acceleration
- Ongoing evolution of the ML acceleration ecosystem



Kamaros Camera Control API

Developer-facing API for detailed and performant access to embedded cameras

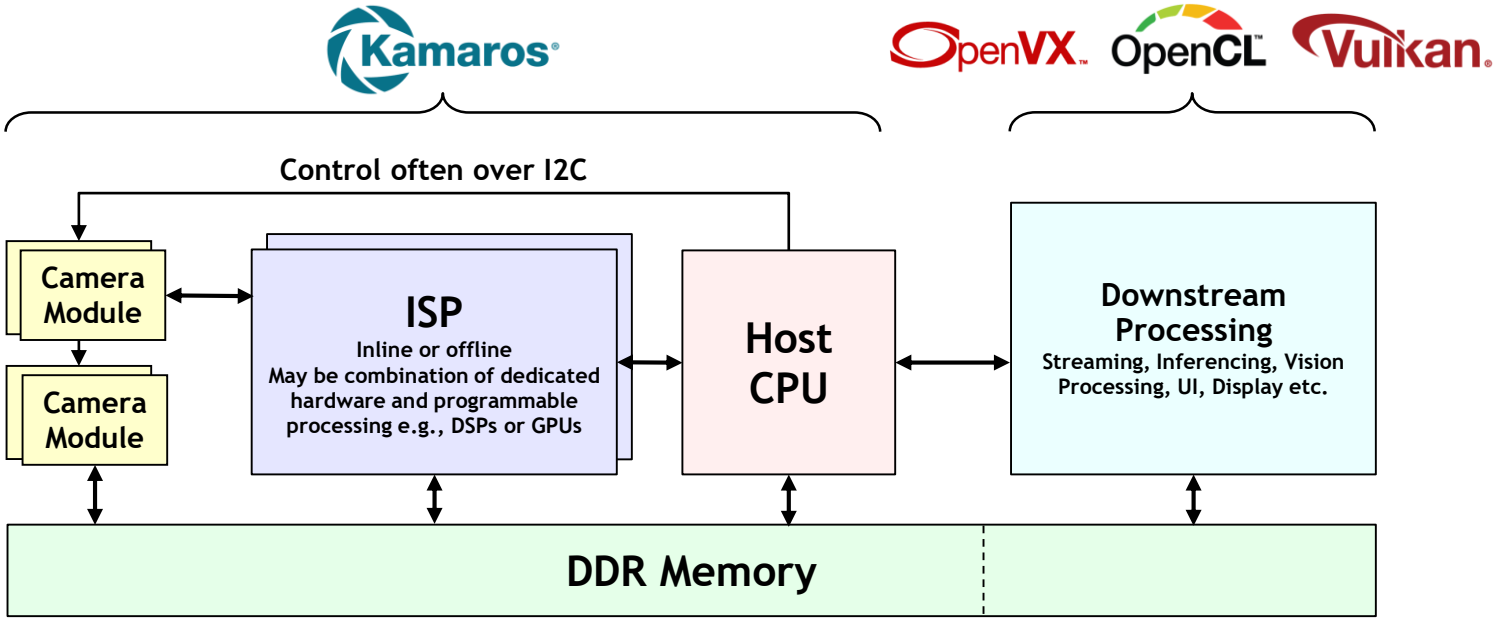


Names of transport layers, framework and operating systems are illustrative examples

Kamaros Scope

Kamaros API provides application facing controls for Camera Modules and close-to-sensor Image Signal Processing (ISP) hardware

Kamaros will be deployable as a standalone driver or integrated into a Vulkan runtime



Kamaros and Sensor Vendor Benefits

- Kamaros has a careful API abstraction: defines operation, not implementation
  - No need for implementers to expose proprietary technology
  - FASTER software ecosystem access to and utilization of sensor innovations
- Kamaros will be extensible using Khronos' proven extension process
  - To rapidly track and evolve to handle sensor innovations
  - Vendors can experiment with custom extensions before collaborative standardization
- Kamaros Pipeline Templating enable definition of tested, well-performing pipelines
  - Reduce risk of sensor mis-configurations
- Kamaros Working Group welcomes input and feedback from sensor vendors
  - Kamaros drivers will be often implemented and shipped by SOC vendors or system integrators
  - How can Kamaros best support sensor integration?
  - What sensor controls are needed beyond exposure time, gain, cropping and scaling?

Kamaros Timeline

Camera Exploratory Group

In response to industry requests EMVA and Khronos cooperate to explore industry interest and build consensus on use cases and requirements

Over 70 companies join and contribute to the discussions

Kamaros Working Group Created

Working Group formed under the Khronos membership and IP framework

Work starts on the detailed specification of the API

Kamaros Spec Drafting

Including cooperative prototyping using Linux on Raspberry Pi

Kamaros 1.0 Release

Including open-source sample implementation and Conformance Test Suite

Two ways to get involved:

1. Join Khronos to directly influence API design and evolution
2. Join Kamaros Advisory Panel to review pre-release spec drafts ($0)

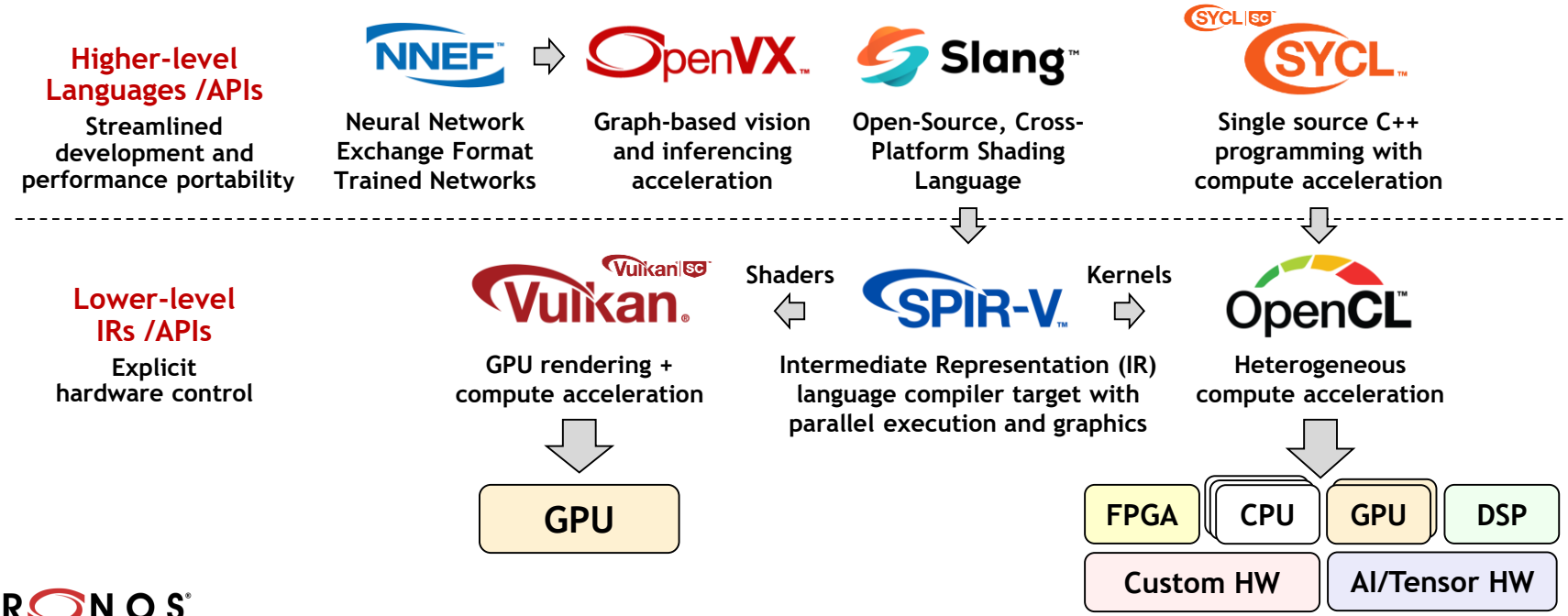
Email: memberservices@khronosgroup.org

Khronos Compute Acceleration Standards

EMBEDDED VISION SUMMIT

Choice of programming models to meet the needs of diverse developers

Higher-level applications, libraries, and languages and APIs often use lower-level standards for hardware access



KHRONOS GROUP

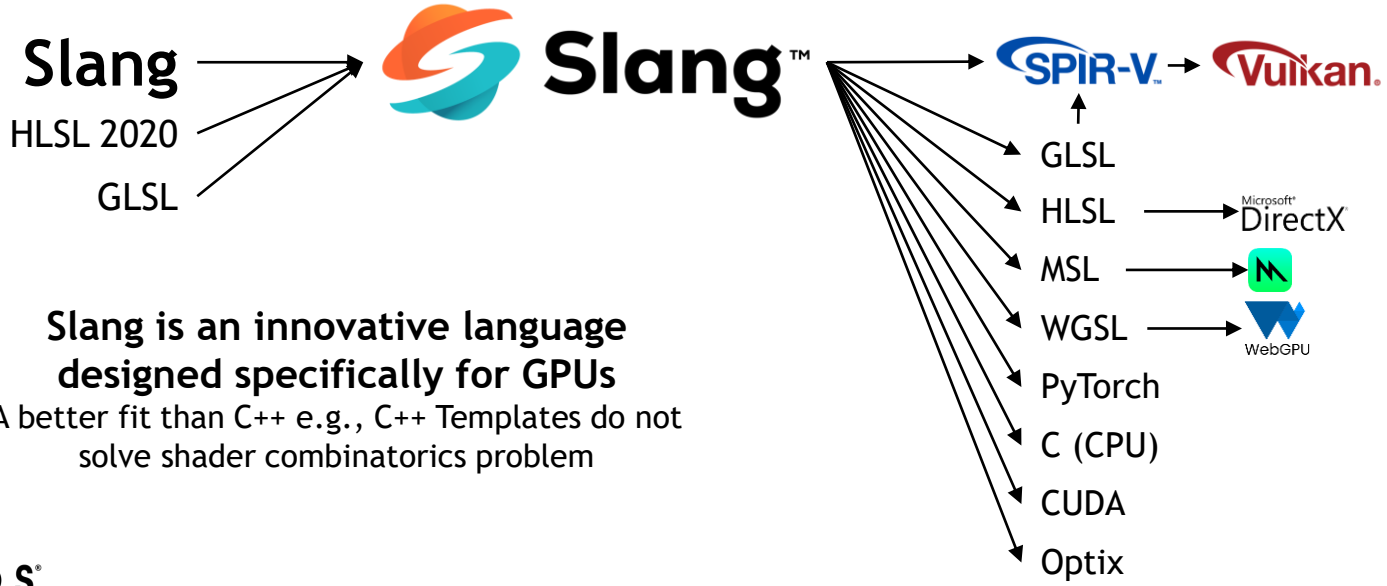
Slang Open-Source, Cross-Platform Compiler

Slang Developer Benefits

On-ramp existing code bases and incrementally transition to a modern language

Improved maintainability of large-scale code bases, language expressivity, machine learning

Write Once - Run 'Everywhere' with multiple, diverse compiler backends

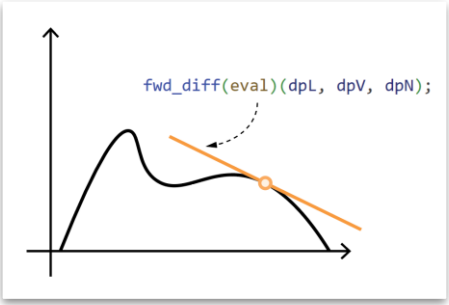


Slang is an innovative language designed specifically for GPUs

A better fit than C++ e.g., C++ Templates do not solve shader combinatorics problem

Slang and Machine Learning

- Slang Automatic Differentiation
  - Fine-grained, modular automatic differentiation directly within GPU shaders
  - Used in gradient descent solutions to train neural networks
  - E.g., custom convolution operations, optical flow, stereo matching
- SlangPy Python library simplifies integration of Slang GPU-accelerated code
  - Integrate and differentiate high-performance Slang kernels within PyTorch
  - Accelerate complex, non-standard ops: sparse convolutions, geometric transformations



python™ ⇄ Slang™

```python
import slangpy as sp
import numpy as np

module = sp.Module.load_from_file(
    device, "example.slang")
a = numpy_array_1
b = numpy_array_2
result = module.add(a, b)
```

```
float add(float a, float b)
{
    return a + b;
}
```

Khronos Slang Open-Source Shading Language

- Slang is the first Khronos initiative where the primary deliverable is open source
  - Not an open standard interoperability specification
- Leveraging 15 years of R&D and deployment experience
  - Originally hosted at NVIDIA from 2017
- Now hosted at Khronos hosting to foster industry-wide collaboration and innovation
  - Open governance provides all an equal chance to influence and decide Slang's evolution

Technical work 100% under streamlined open-source project

Welcome contributors from any engaged company

Community-driven project structure and best practices suited to a shading language

Working Group bit only for logistical and funding support

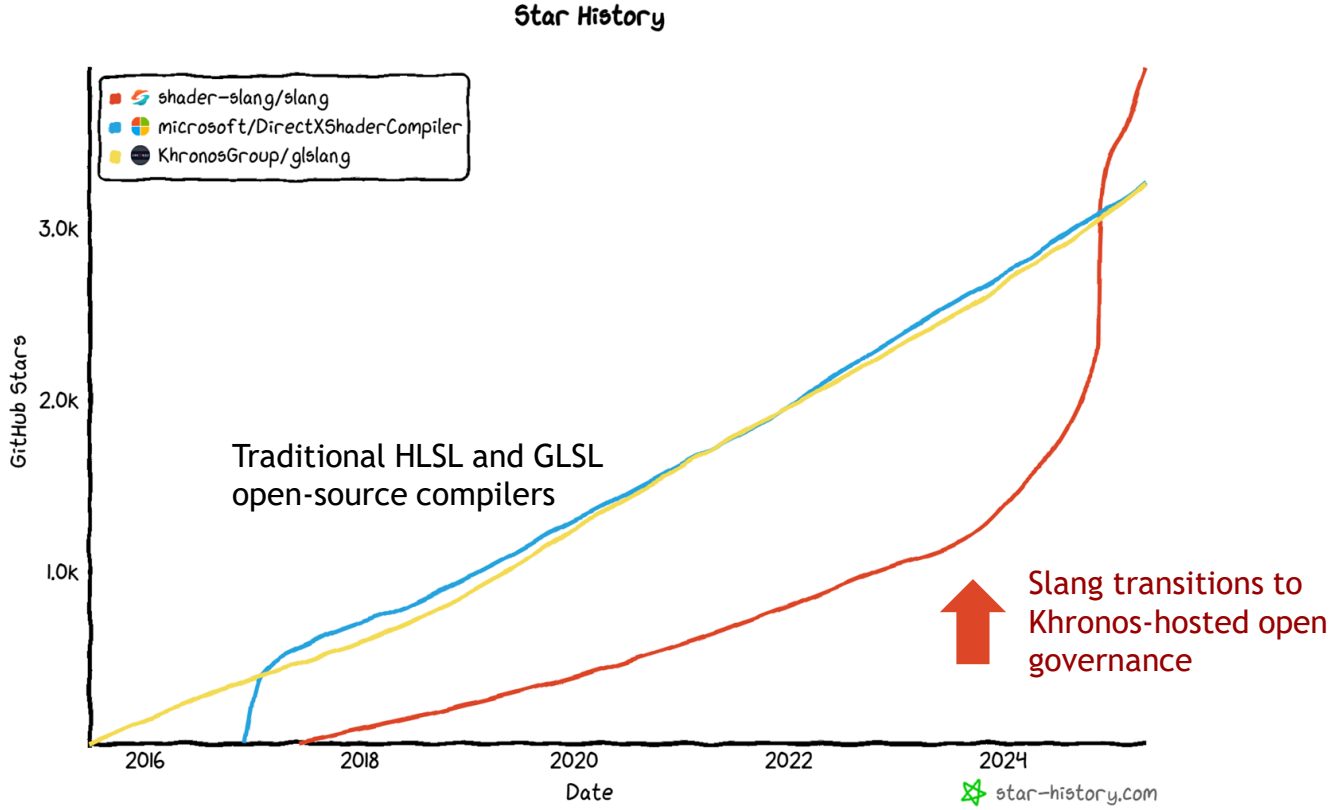
Enable the open-source project to focus on technical forward progress

Explore and leverage synergy between Slang, Vulkan and SPIR-V

Try Slang right now in your browser!

https://shader-slang.org/slang-playground/

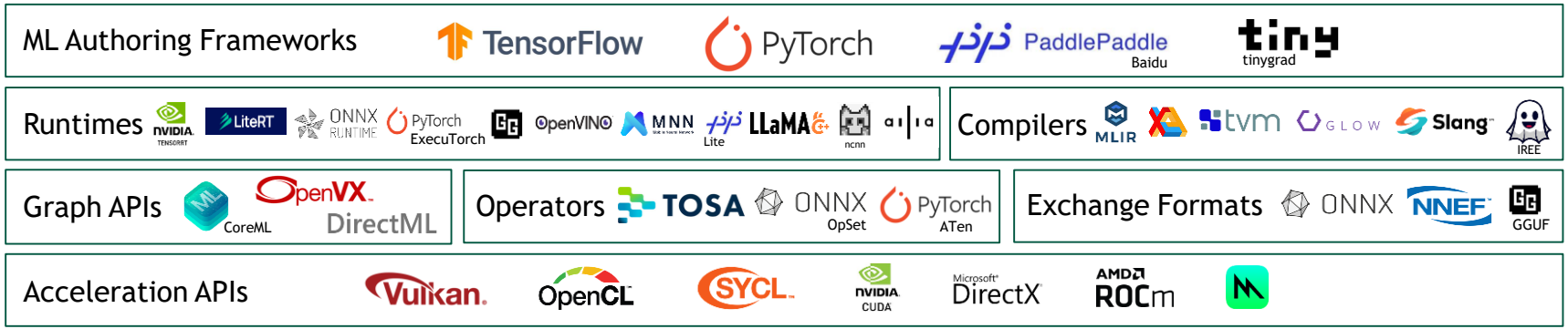
Slang Industry Interest



Accelerated ML Landscape

The native machine learning acceleration stack is complex

Similar discussions also happening in the JavaScript stack for machine learning in the Web



Acceleration APIs are the foundation of this stack

Khronos provides the industry's only non-proprietary accelerator APIs

OpenCL and Machine Learning

- Programming and runtime framework for parallel heterogeneous processors
  - Offload compute-intensive kernels onto diver hardware: CPUs, GPUs, DSPs etc.
- OpenCL is often used as a backend for ML compilers and inference engines
  - Especially in the embedded and mobile markets
- OpenCL has a robust future pipeline of AI-related extensions
  - Recordable command buffers, including mutable command buffers
  - 'Cooperative matrix' for standard access to dedicated matrix hardware
  - New AI data types, such as bfloat16 and fp8

Machine Learning frameworks, libraries and compilers using OpenCL for offload acceleration

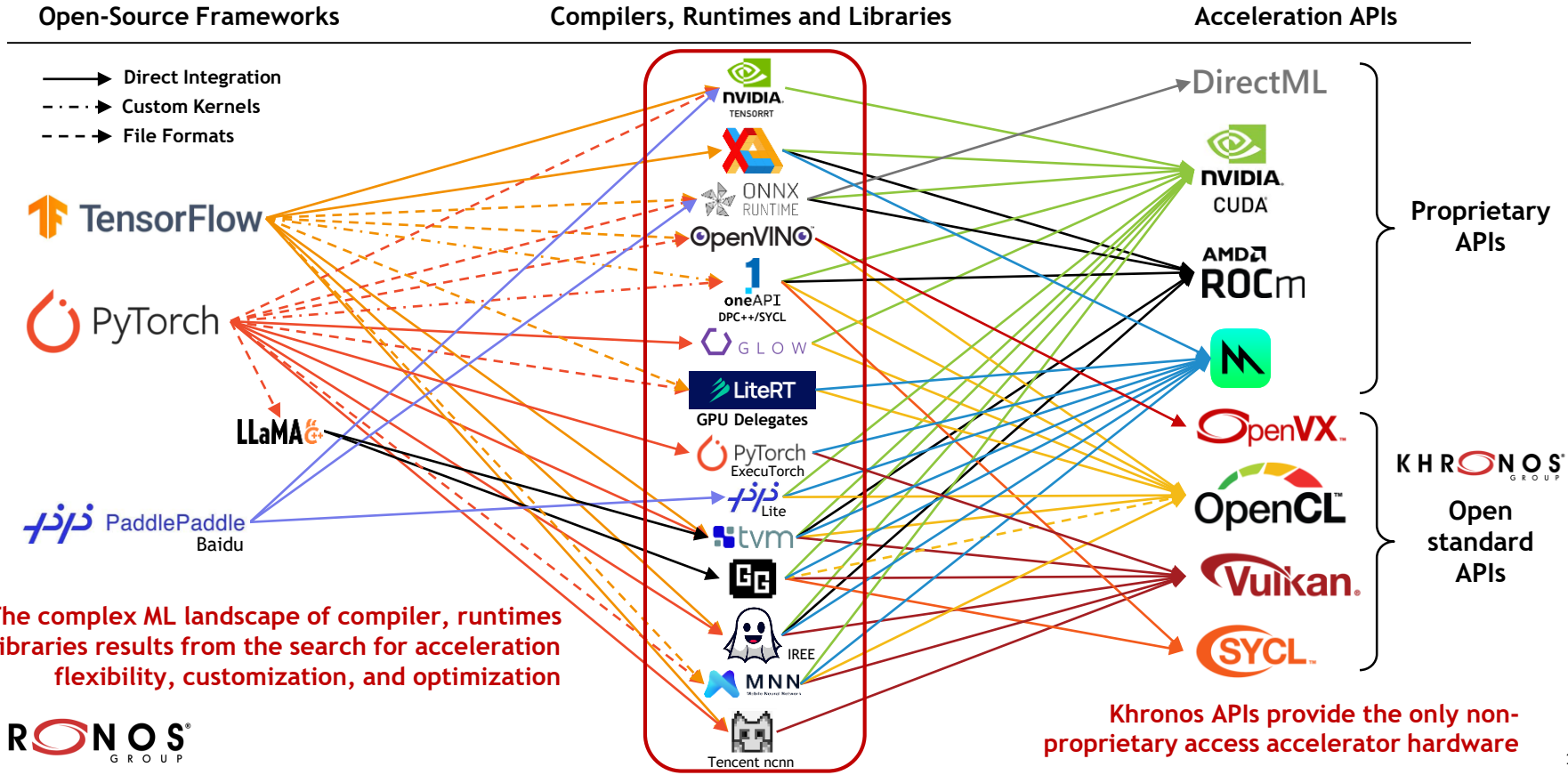
LiteRT (Google)
OpenVINO (Intel)
Arm Compute Library
CDNN (CEVA)
Acuity (VeriSilicon)
飞桨 PaddlePaddle (Baidu)
GLOW (Meta)
SYCL
MACE Mobile AI Compute Engine (Xiaomi)
Qualcomm Neural Processing SDK
TI DL Library (TIDL)
XNNC (Cadence)
tvm (Apache)
MetaWare EV (Synopsis)
MNN Mobile Neural Network (Alibaba)
OpenCV

Vulkan for ML Inferencing

- "Explicit" low-level graphics and compute API providing detailed GPU control
  - Widely adopted and deployed in mobile, embedded desktop and cloud platforms
  - Native drivers for Windows, Linux, Android and Nintendo Switch
- Extensions provide an increasing range of high-performance ML-focused primitives
  - VK_KHR_cooperative_matrix - matrix-matrix multiply in a subgroup
  - VK_NV_cooperative_matrix2 - matrix-matrix multiply in a workgroup scope
  - VK_NV_cooperative_vector - matrix-vector multiplies with fp8/fp16 & int8 support
  - Precisions: VK_KHR_shader_bfloat16 and VK_KHR_shader_float16_int8

Vulkan is being increasingly used to provide backend acceleration in inference frameworks

Machine Learning Acceleration Complexity



The complex ML landscape of compiler, runtimes and libraries results from the search for acceleration flexibility, customization, and optimization

Khronos APIs provide the only non-proprietary access accelerator hardware

Khronos ML Market Research

- Can open standard acceleration APIs reduce this market friction and complexity?
  - Faster performance, streamlined development, reduced porting and support costs
- What are the real-world industry needs that would have to be met?
  - Do we need a common ML hardware acceleration stack for CPUs, NPUs and GPUs?
  - What functionality should HAL APIs provide for effective tensor acceleration?
  - What is the most effective way to balance inferencing and other loads on a GPU?
- What are the most promising potential solutions to explore?
  - Will a graph abstraction enable optimal and portable performance?
  - Standardization of a tensor operator set such as Arm's TOSA?
  - Communicate optimization data between layers of the stack?

Khronos is initiating funded market research to try to begin answer these questions and more!

Seeking first round of input in this online survey

Please take few minutes - your input is appreciated!

Khronos Group AI Opportunity Survey

Call for Input and Participation

- Camera or sensor vendor?
  - Please consider participation in the Kamaros Working Group
  - Join Khronos or request to join the Advisory Panel for free
- Using GPUs for inferencing or image processing?
  - Try out the open-source Slang Shading Language for free on the Web!
- Seeking portable and effective inferencing acceleration?
  - Make your requirements known in the Khronos high-level ML Survey!
  - Let us know if you would like to provide more detailed input and feedback
- All welcome to join Khronos to directly influence standards design and evolution
  - www.khronos.org/members/
  - Email memberservices@khronosgroup.org
- More information on all Khronos APIs
  - https://www.khronos.org/