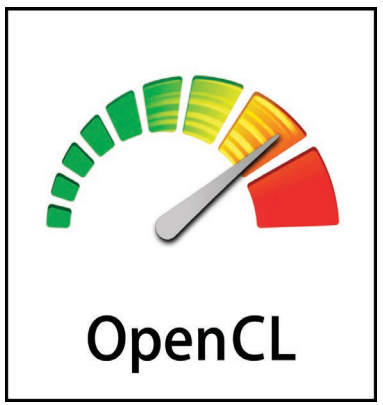
KHRONOS GROUP



Khronos Update

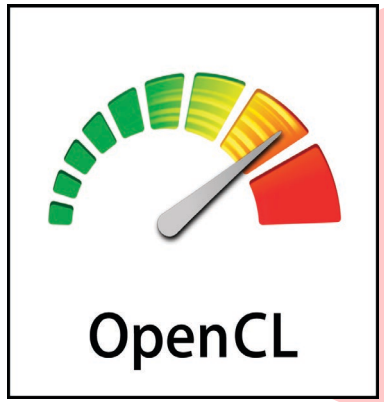
OpenCL, SYCL and SPIR - The Next Steps

Neil Trevett | Khronos President
NVIDIA Vice President Developer Ecosystem
OpenCL Working Group Chair
ntrevett@nvidia.com | @neilt3d

Boston, May 2019

OpenCL Update from the Khronos Perspective

OpenCL is the only cross-platform industry standard for low-level heterogeneous compute



1. Widening support and industry usage

Heterogenous compute is the 'new Moore's Law'
Critical to new-generation mobile/embedded systems

2. Strengthening the OpenCL Ecosystem

Increasing community engagement
Leveraging the power of open source resources

3. Deploying New Functionality

Extensions and core specs
Processor deployment flexibility
Community-based kernel language tooling

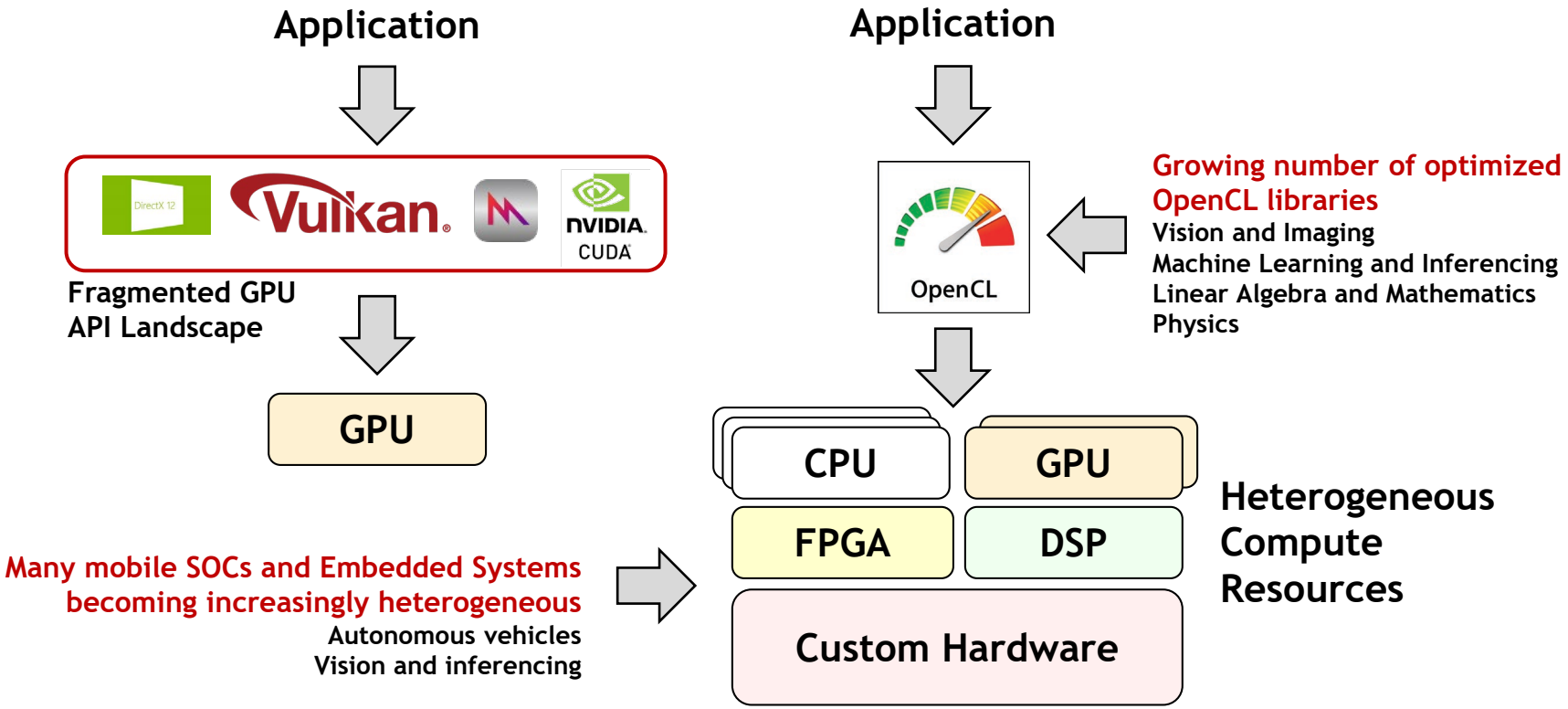
4. Platform Deployment Flexibility

Enabling OpenCL with no native drivers

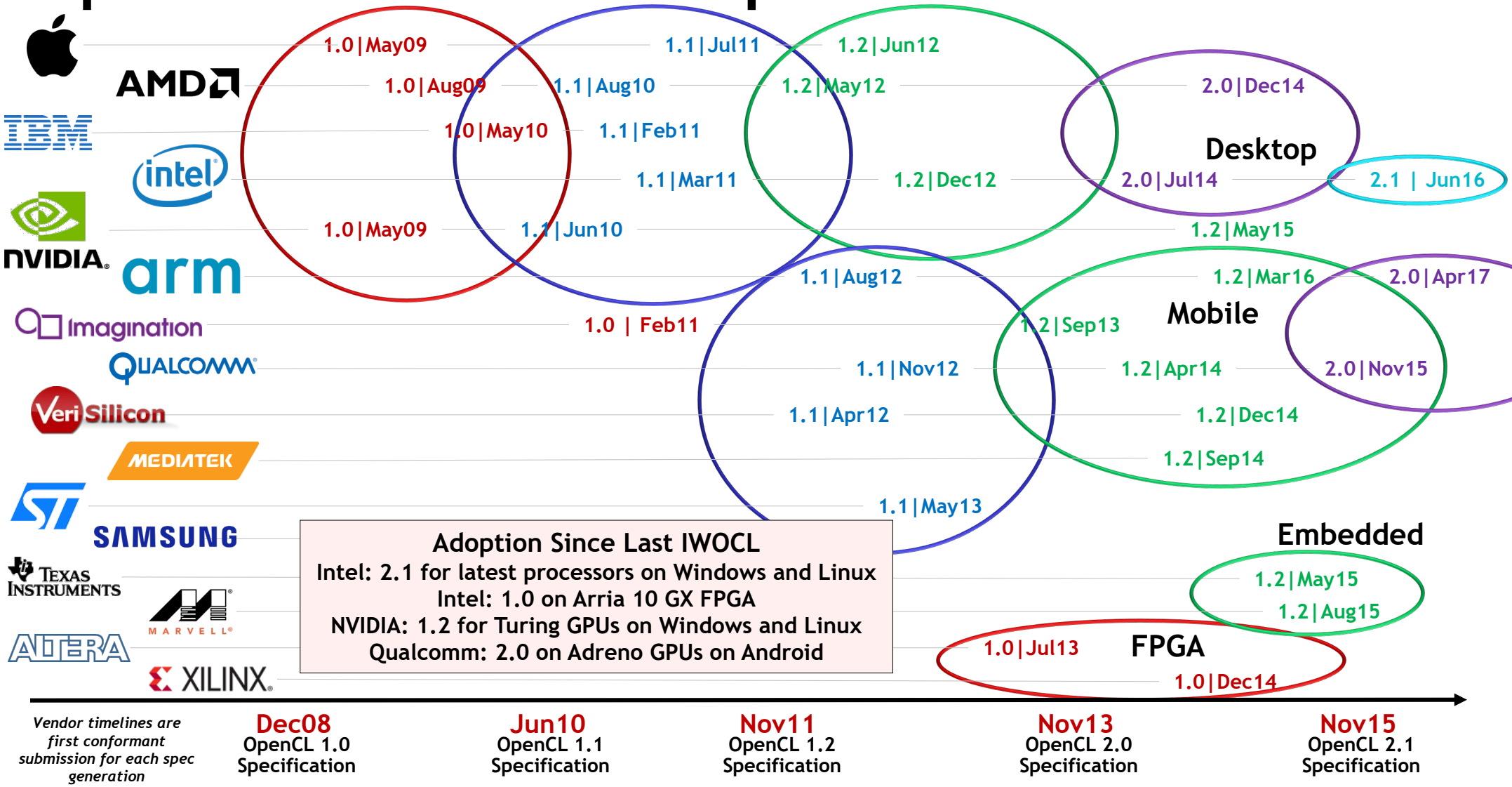
KHRONOS GROUP

OpenCL Heterogeneous Computing

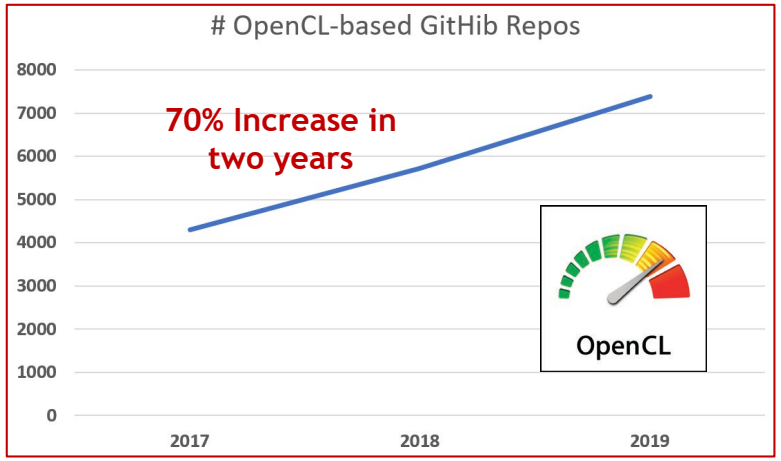
A programming and runtime framework for heterogeneous compute resources
Low-level control over memory allocation and parallel task execution
Simpler and relatively lightweight compared to GPU APIs



OpenCL Conformant Implementations



OpenCL User Adoption



OpenCL is Pervasive!

Desktop Creative Apps

Parallel Computation Languages

Linear Algebra Libraries

Vision and Imaging Libraries

Math and Physics Libraries

Machine Learning Libraries

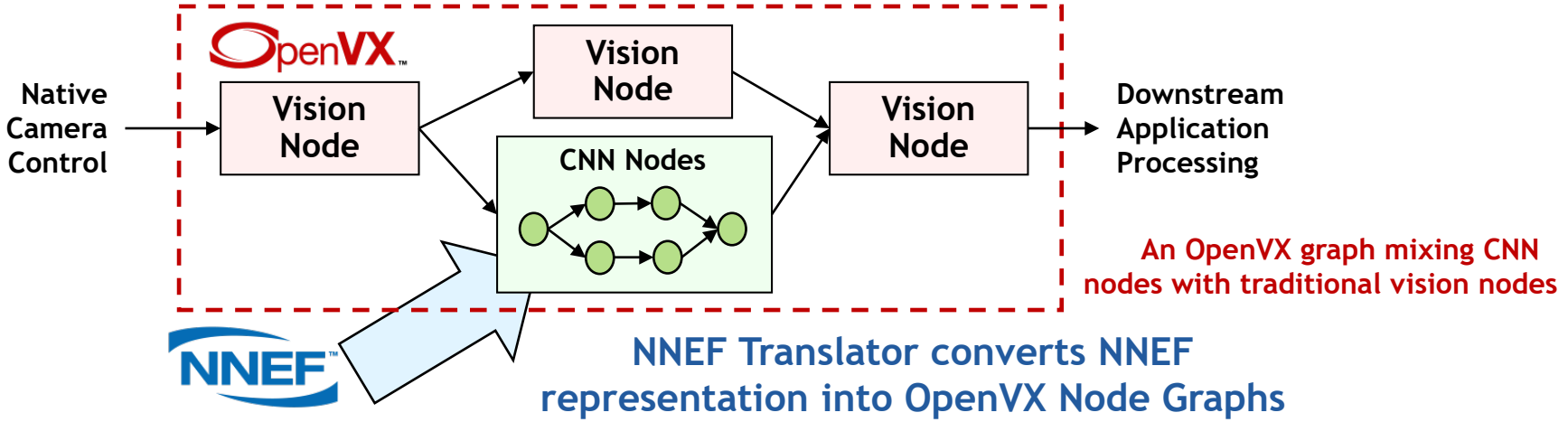
Machine Learning Inferencing Compilers

https://www.khronos.org/opencl/resources/opencl-applications-using-opencl
https://www.iwocl.org/resources/opencl-libraries-and-toolkits/

Khronos OpenVX and NNEF for Inferencing

OpenVX

A high-level graph-based abstraction for Portable, Efficient Vision Processing
Can be implemented on almost any hardware or processor



NNEF (Neural Network Exchange Format)

For transferring trained Neural Networks into inferencing accelerators
Provides stability needed by hardware vendors through true multicompany governance

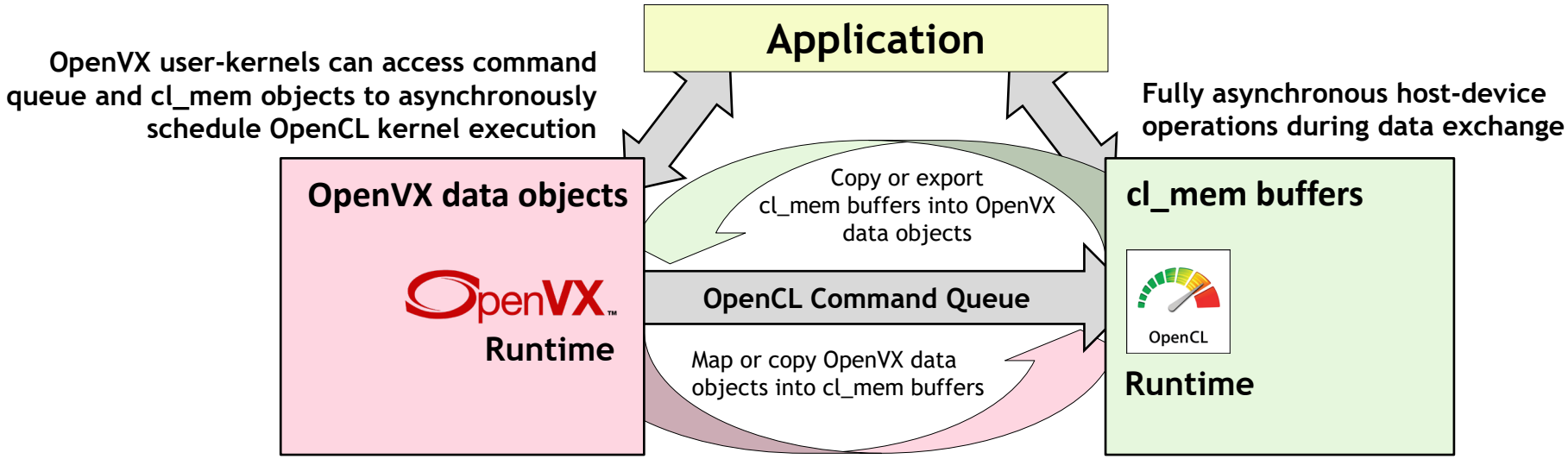
Extending OpenVX for Custom Nodes

OpenVX/OpenCL Interop

- Provisional Extension
- Enables custom OpenCL acceleration to be invoked from OpenVX User Kernels
- Memory objects can be mapped or copied

Kernel/Graph Import

- Provisional Extension
- Defines container for executable or IR code
- Enables arbitrary code to be inserted as a OpenVX Node in a graph

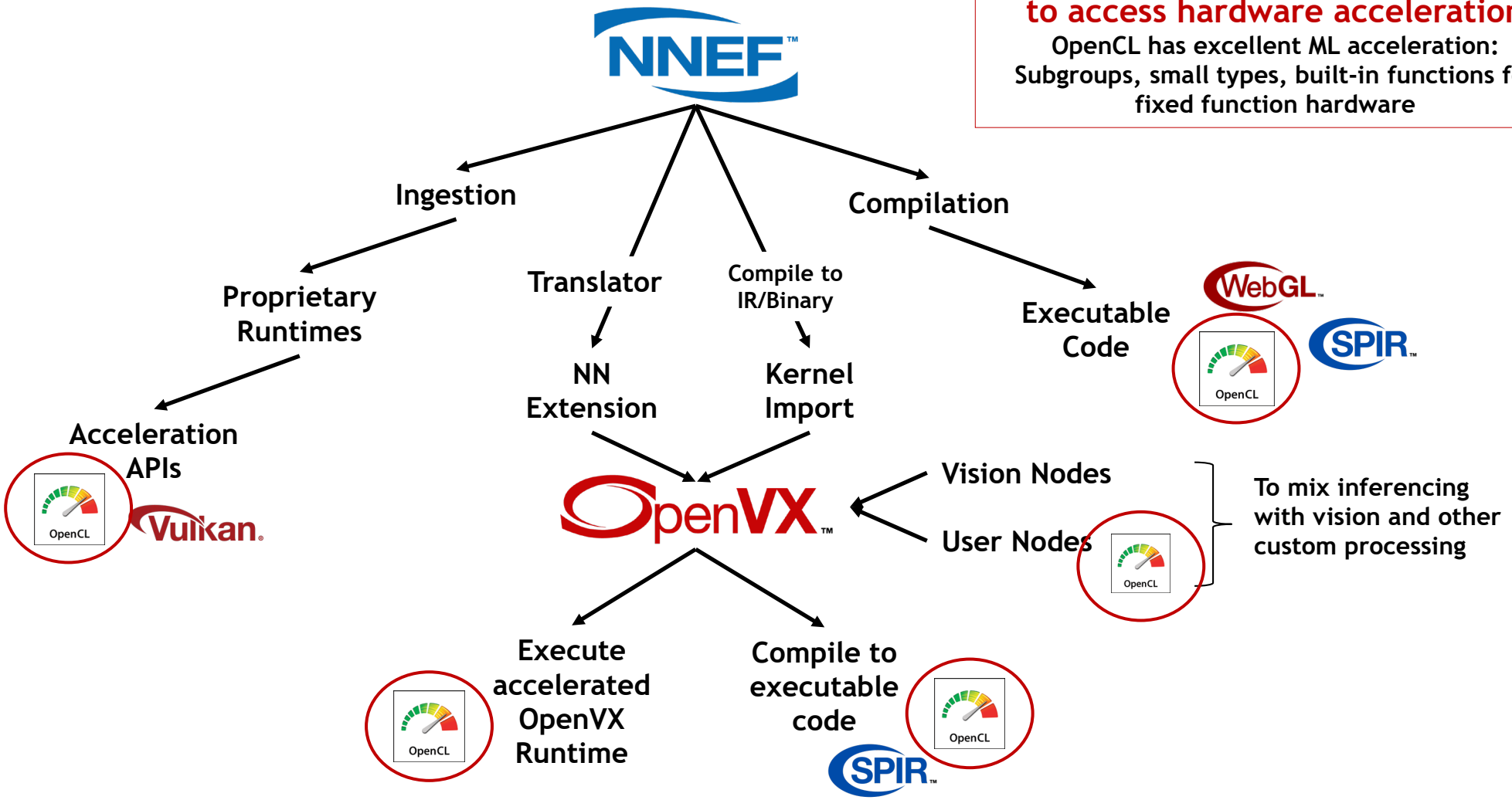


OpenVX/OpenCL Interop

Inferencing Acceleration

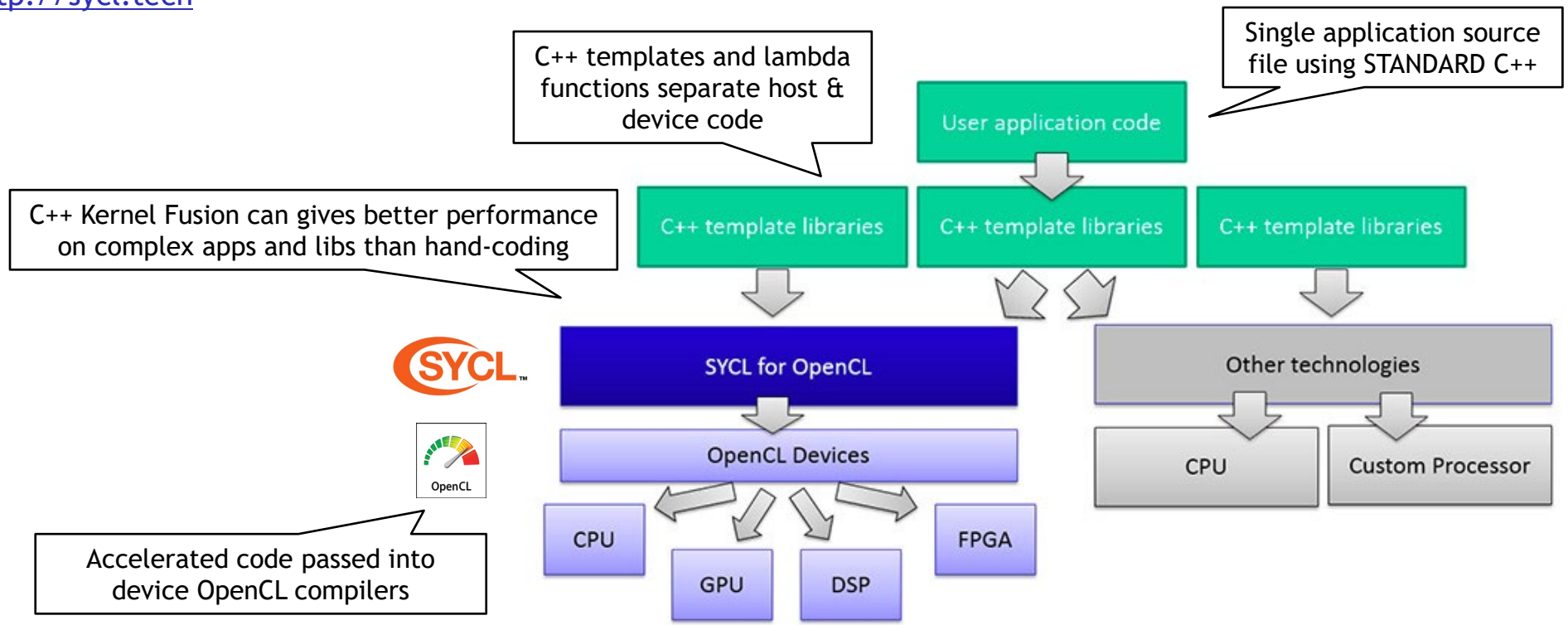
Many inferencing stacks use OpenCL to access hardware acceleration

OpenCL has excellent ML acceleration: Subgroups, small types, built-in functions for fixed function hardware

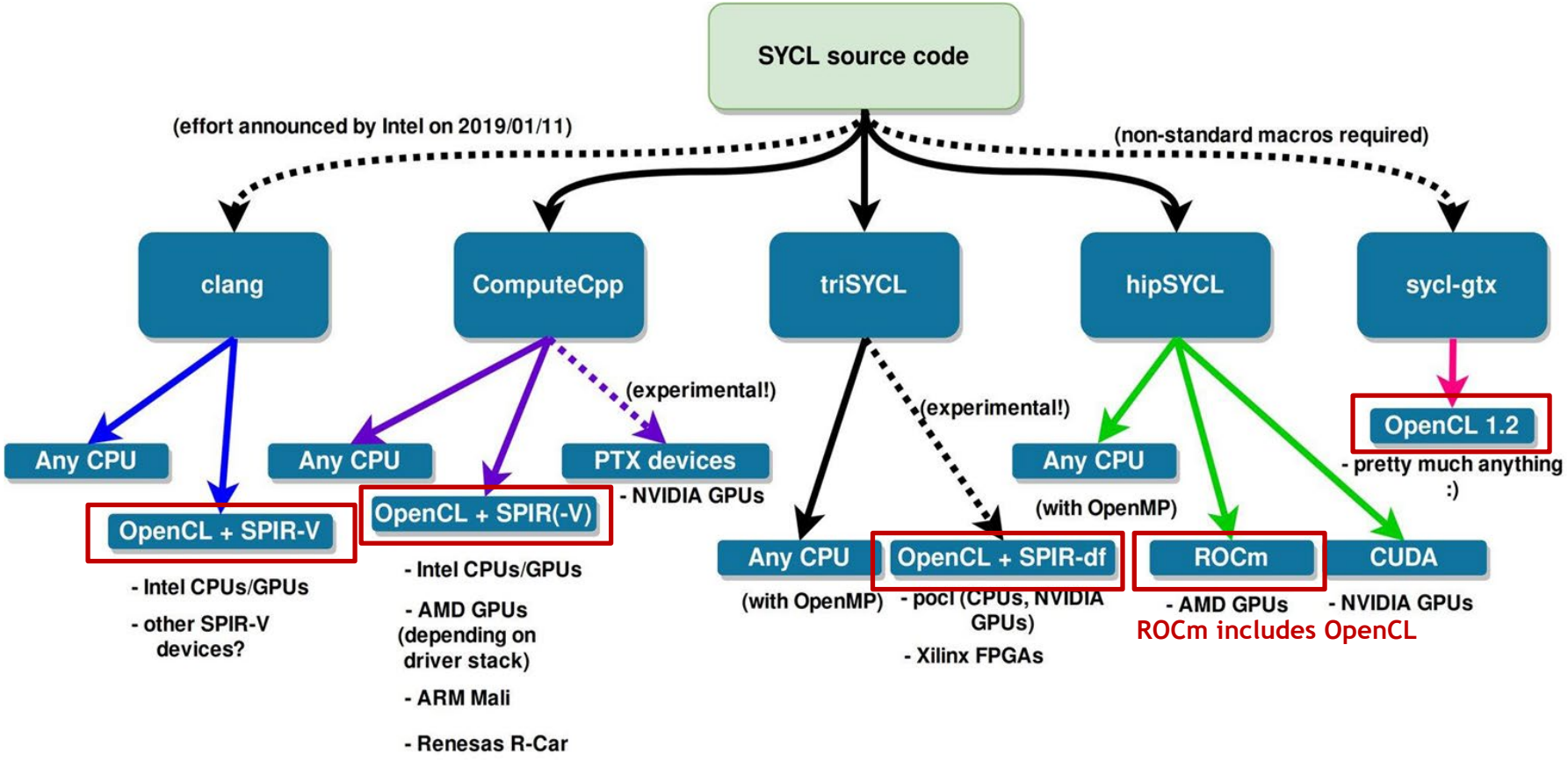


SYCL Single Source C++ Parallel Programming

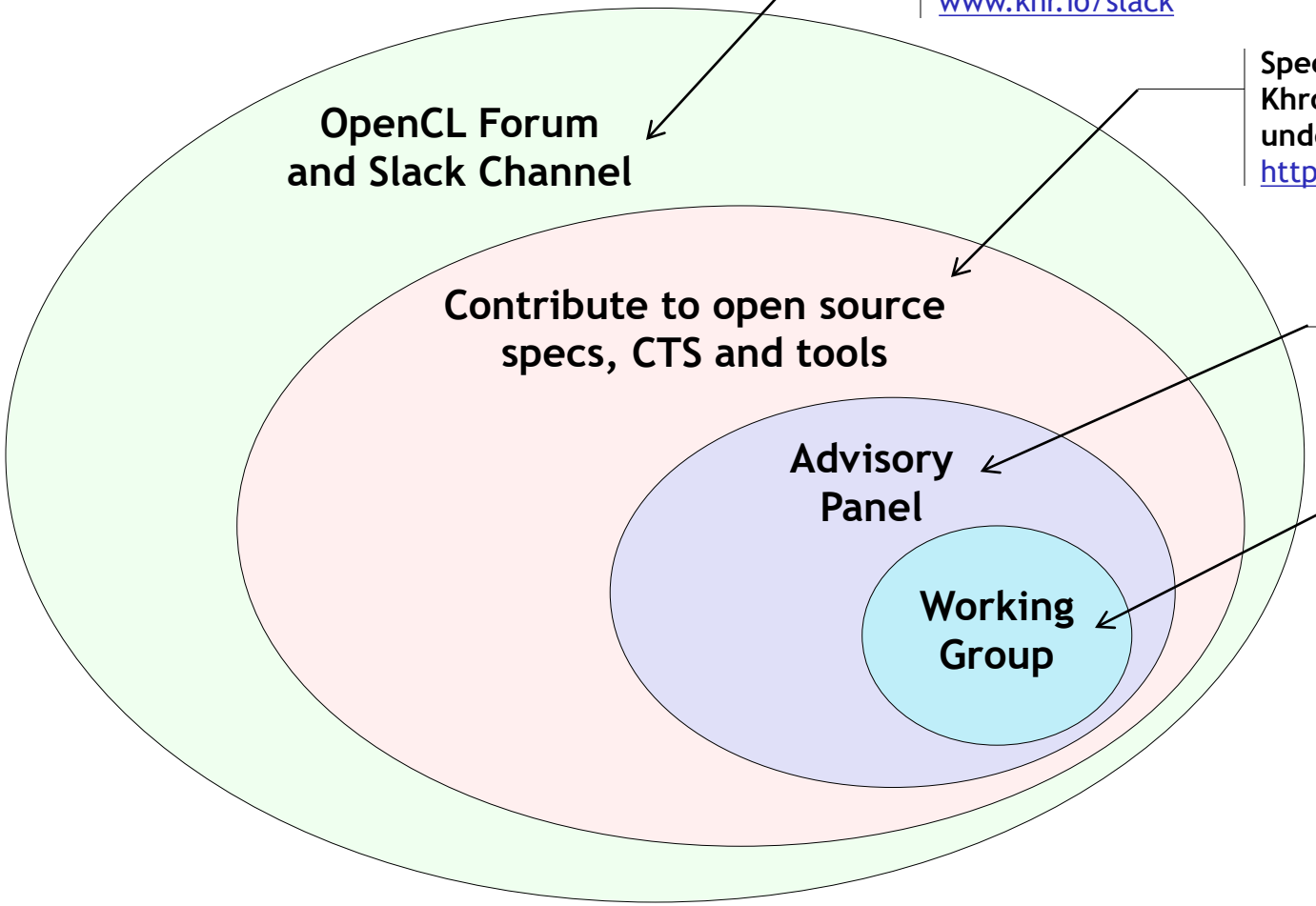
- **SYCL 1.2.1 Adopters Program released in July 2018 with open source conformance tests soon**
  - https://www.khronos.org/news/press/khronos-releases-conformance-test-suite-for-sycl-1.2.1
- **Multiple SYCL libraries for vision and inferencing**
  - SYCL-BLAS, SYCL-DNN, SYCL-Eigen
- **Multiple Implémentations shipping: triSYCL, ComputeCpp, HipSYCL**
  - http://sycl.tech



SYCL Implementations



Ecosystem Engagement



Open to all!
https://community.khronos.org/c/opencl
www.khr.io/slack

Spec fixes and suggestions made under the Khronos IP Framework. Open source contributions under repo's CLA - typically Apache 2.0
https://github.com/KhronosGroup

Advisors under the Khronos IP Framework can comment and contribute to requirements and draft specifications

Khronos members under IP Framework and NDA can participate and vote in working group meetings

OpenCL Open Source Specs and Tests

Khronos has open sourced OpenCL Specifications and Conformance Tests

Merge requests welcome from the community (subject to review by OpenCL working group)
File bugs for specifications, headers etc.
Mix your own documentation!

Source Materials for Specifications and Reference Documentation CONTRIBUTED Under Khronos IP Framework (you won't assert patents against conformant implementations, and license copyright for Khronos use)

Spec and Ref Language Source

Redistribution under CC-BY 4.0

Contributions and Distribution under Apache 2.0

Spec Build System and Scripts

Contributions and Distribution under Apache 2.0

Conformance Test Suite Source

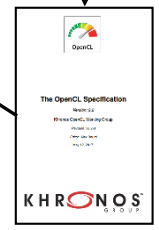
Khronos builds and Ratifies Canonical Specification under Khronos IP Framework. No changes or re-hosting allowed

Spec and Ref Language Source and derivative materials. Re-mixable under CC-BY by the industry and community

Khronos Adopters Program

Anyone can test any implementation at any time

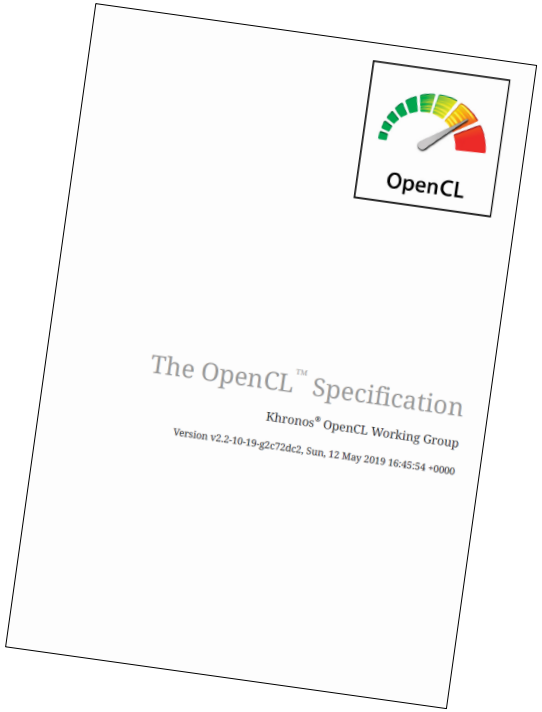
Conformant Implementations can use trademark and are covered by Khronos IP Framework



Community built documentation and tools

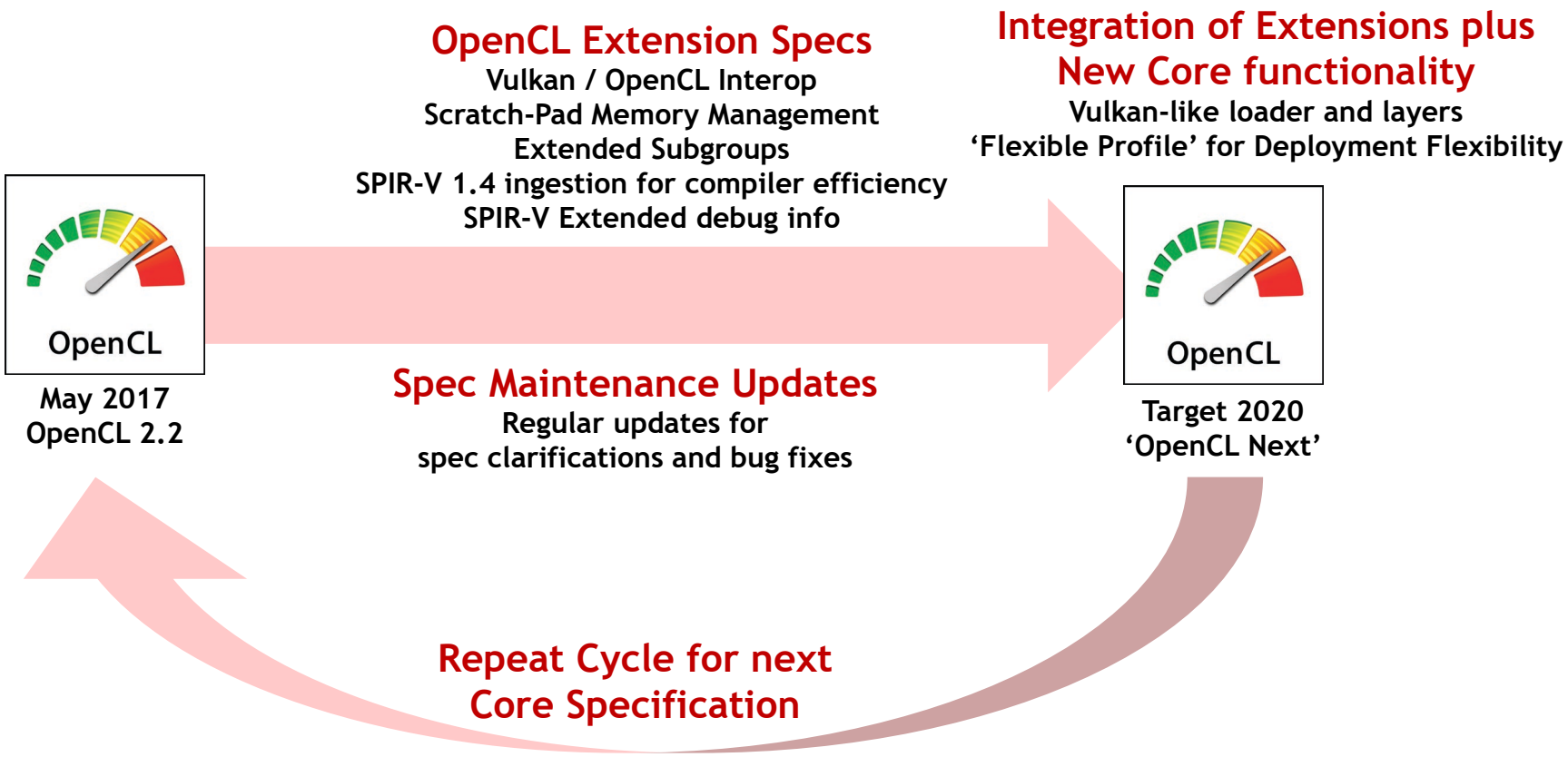
Underway - Unified OpenCL Specification

- **Unified OpenCL API specification will describe the API for all versions of OpenCL**
  - Rather than having a separate specification per version
- **OpenCL SPIR-V environment, extension and SPIR-V specs are already unified**
  - Working well – good developer feedback
- **Easier for developers to navigate**
  - And to consistently apply specification fixes and clarifications
- **Working Group has started prototyping the spec work**
  - Short introductory section describing the unified aspects
  - "missing before X.Y" and "deprecated by X.Y" language
  - Roughly as SPIR-V specification
  - https://github.com/KhronosGroup/OpenCL-Docs/issues/77
- **DRAFT unified spec is already uploaded!**
  - https://github.com/KhronosGroup/OpenCL-Docs/files/3170333/OpenCL_API.pdf
  - Feedback welcome!



Eases opportunity to coherently include deprecation and version evolution rationale in specification - as requested in yesterday's BOF

OpenCL Evolution



Embedded Processors & OpenCL Conformance

- **The embedded market is a new frontier needing advanced heterogenous compute**
  - E.g. Vision and inferencing using a wide range of processor architectures
- **BUT OpenCL is currently monolithic – and arguably desktop/HPC-centric**
  - E.g. a processor without 32-bit IEEE floating point cannot realistically be conformant
  - Vendors and developers do not want software emulation of higher precisions
- **Many functionality requirements change between different markets and processors**

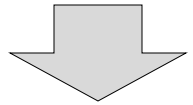
OpenCL is disenfranchising one of its most important emerging market opportunities!

Supported Precisions	DSP A	DSP B	DSP C
8-bit integer	✓	✓	✓
16-bit integer	✓	✓	✓
32-bit integer	✓	✓	✓
64-bit integer	✗	✓	✗
16-bit float	✗	✓	✓
32-bit float	✗	✗	✓
64-bit float	✗	✗	✗
Possible to be OpenCL Compliant?	No	No	Yes

OpenCL Next 'Flexible Profile'

Goals

Enable Conformant OpenCL implementations on more diverse processors and platforms
Enable vendors to ship functionality precisely targeting their customers and markets
A conformant OpenCL can expose precisely what is available in the hardware
Enable incremental feature adoption



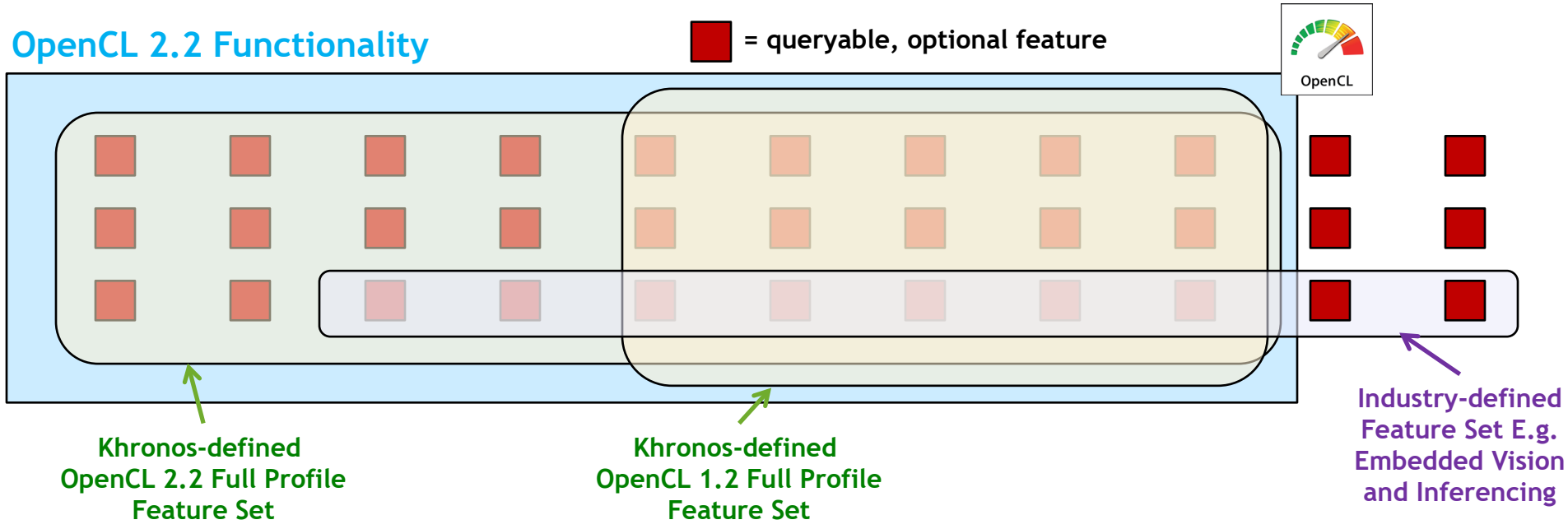
Design Philosophy

More OpenCL features become optional for enhanced deployment flexibility
Optionality includes both API and language features e.g. floating point precisions
Enhanced query mechanisms - precisely which features are supported by a device?

OpenCL Next aims to be a flexible run-time framework that can be pervasively and cost-effectively deployed across a wider range of heterogenous devices

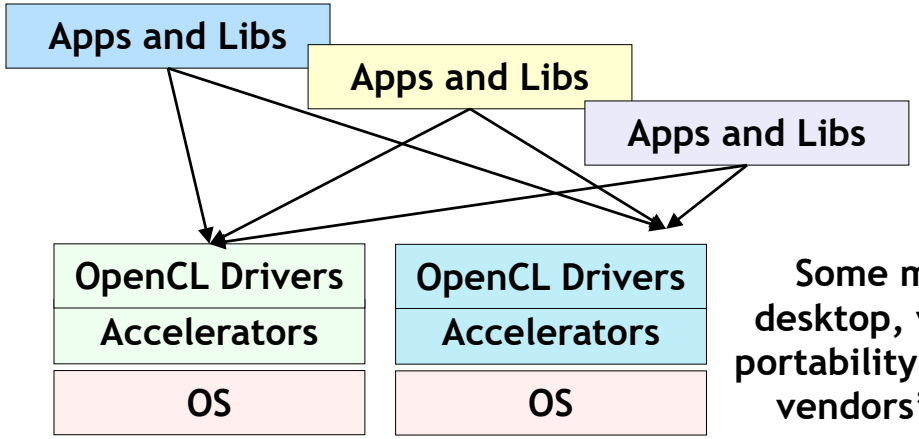
Flexible Profile and Feature Sets

- **In OpenCL Next Flexible Profile features become optional for enhanced deployment flexibility**
  - API and language features e.g. floating point precisions
- **Feature Sets reduce danger of fragmentation**
  - Defined to suit specific markets – e.g. desktop, embedded vision and inferencing
- **Implementations are conformant if fully support feature set functionality**
  - Supporting Feature Sets will help drive sales – encouraging consistent functionality per market
  - An implementation may support multiple Feature Sets



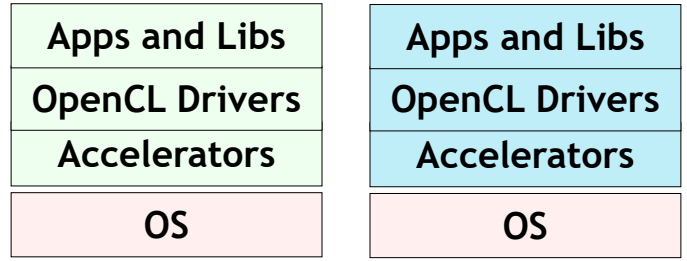
OpenCL Next Feature Set *Discussions*

- **OpenCL Next Flexible Profile will leverage the new Unified OpenCL Specification**
  - Feature Sets can easily select from any previous functionality
- **What could be useful Feature Sets?**
  - Feature sets for previous spec versions e.g. OpenCL 1.2?
  - 'Desktop' Feature Set to raise the universally available baseline above OpenCL 1.2?
  - Set of OpenCL functionality that runs efficiently over Vulkan?
  - Vertical market focused – e.g. inferencing, vision processing
- **Some vertically integrated markets don't care about cross-vendor app portability**
  - But still want to use industry standard programming framework – reduces costs for engineering, tooling, training etc.
  - Allow conformance for *any* combination of features – no Feature Sets
  - Enables minimal footprint OpenCL per system – ideal for Safety Certification



Some markets, e.g. desktop, value software portability across multiple vendors' accelerators

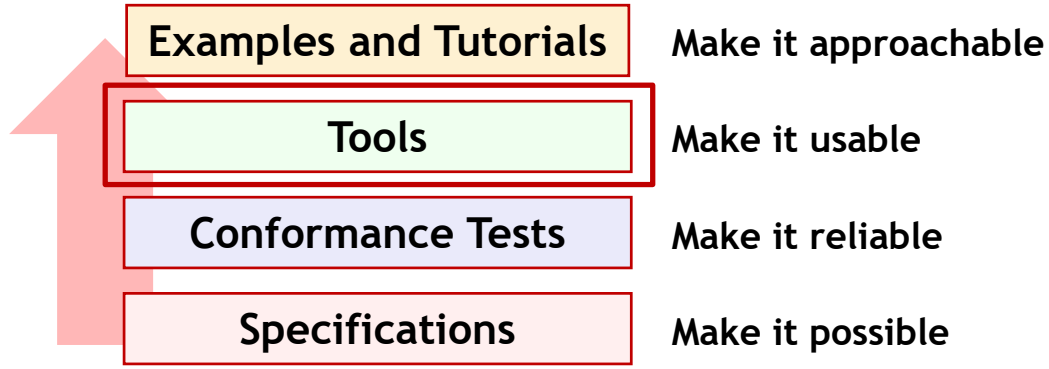
In some vertically integrated embedded markets, application portability is not so important



OpenCL Tooling Ecosystem Subgroup

- **Coordinating collaborative opportunities between SPIR-V and LLVM ecosystems**
  - Encouraging joint development of new features and tooling integration
- **Active open source projects – making SPIR-V a first-class LLVM citizen**
  - Extending SPIR-V<->LLVM Translation for OpenCL - release 8.0 is out!
  - https://github.com/KhronosGroup/SPIRV-LLVM-Translator
  - Libclc: implementation of the OpenCL C 1.1 library for use with Clang
  - https://libclc.llvm.org/
  - Upstream SPIR-V backend translation from Clang/LLVM – in discussion
  - https://clang.llvm.org/
  - Front-end support for all OpenCL C language versions in Clang
- **C++ for OpenCL in Clang**
  - Experimental support for C++ in OpenCL

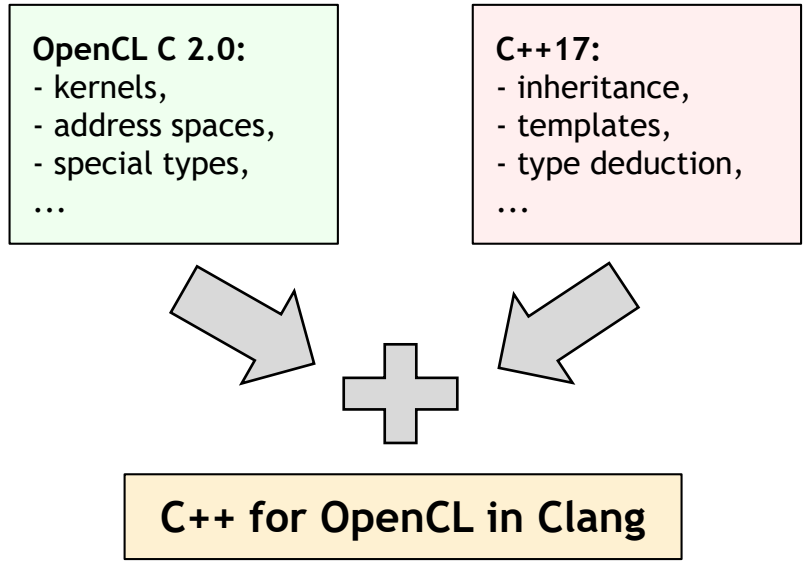
Khronos increasing efforts around developing, coordinating and releasing open source tooling



C++ for OpenCL in Clang Project

- **Front end with OpenCL C 2.0 and C++17 capabilities**
  - Experimental support in Clang 9.0
  - Expect Alpha in September 2019
- **Existing OpenCL C code is valid and fully compatible**
  - Enables gradual transition to C++ for existing apps
- **Offline compilation into SPIR-V or device binary**
  - Generates SPIR-V 1.0 for most features
  - Uses SPIR-V 1.2 where necessary
- **Works with any OpenCL 2.0 driver**
  - Possible future driver updates may take advantage of enhanced language capabilities
- **Check it out in Compiler Explorer**
  - https://godbolt.org/z/nGvxAC

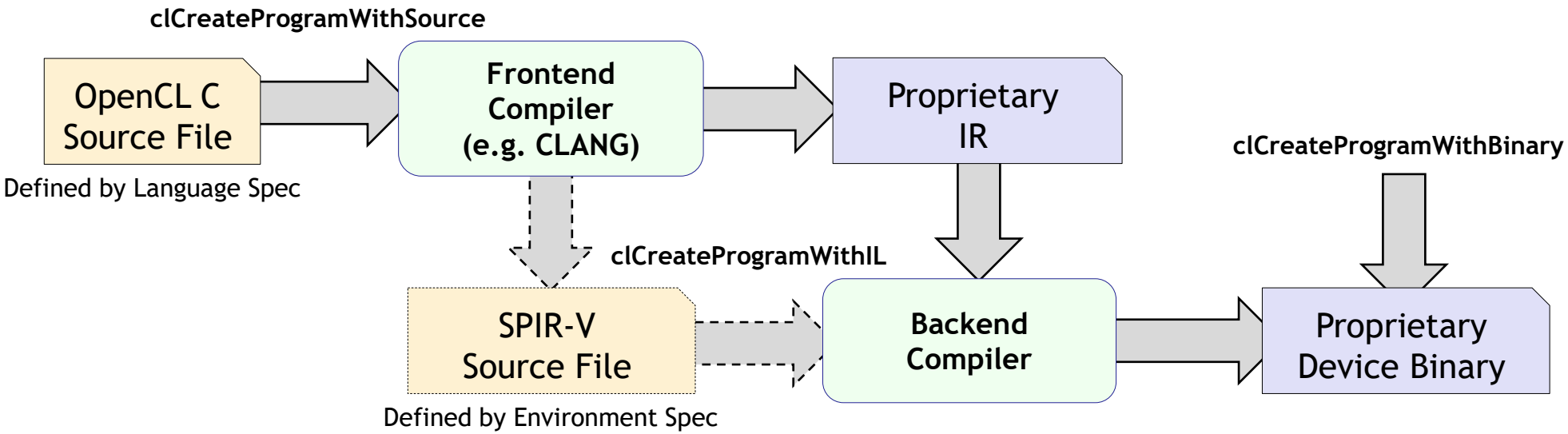
Good example of the power and flexibility of offline compilation using SPIR-V - how can we embrace and support this class of language project?



clang -std=c++ test.cl

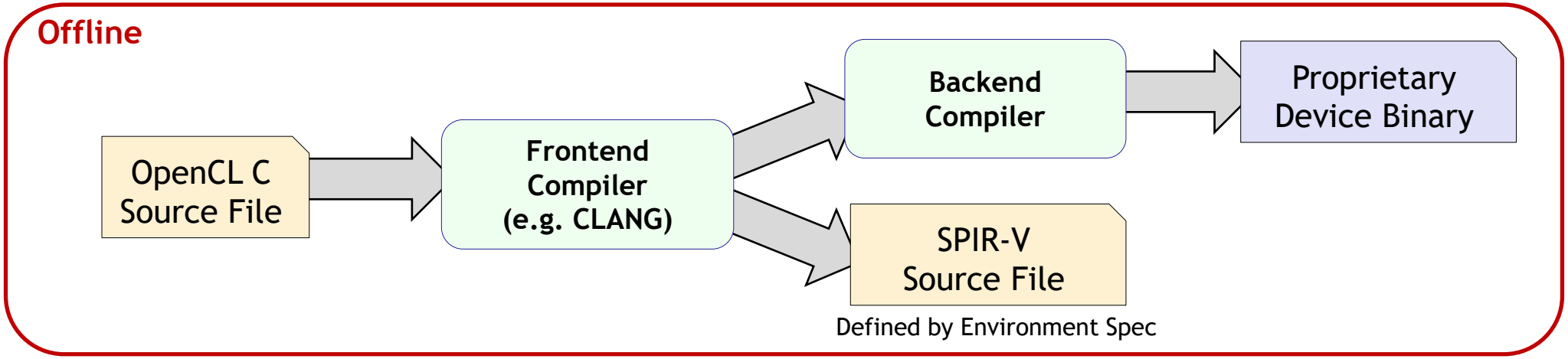
```
template<class T> T add( T x, T y )
{
  return x + y;
}

__kernel void test( __global float* a, __global float* b)
{
  auto index = get_global_id(0);
  a[index] = add(b[index], b[index+1]);
}
```

Current OpenCL Compilation Flow

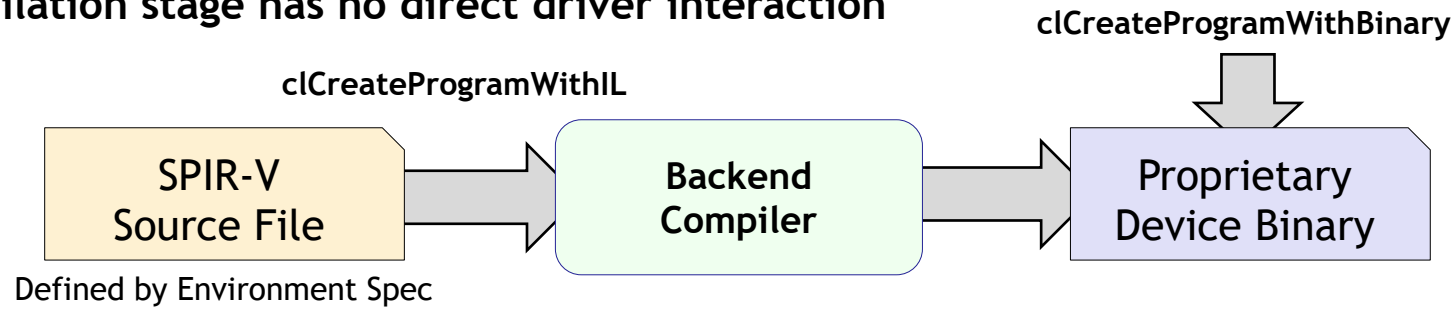


Frontend and backend compilers ship with OpenCL driver
Adding new language capabilities → new driver version

Offline Compilation with SPIR-V



Offline compilation stage has no direct driver interaction

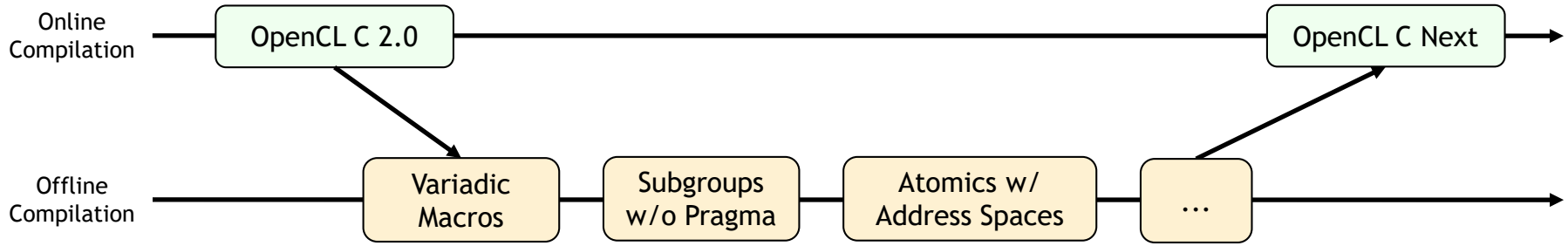


If offline compiler generates valid SPIR-V, front-end can add language capabilities without requiring a driver update

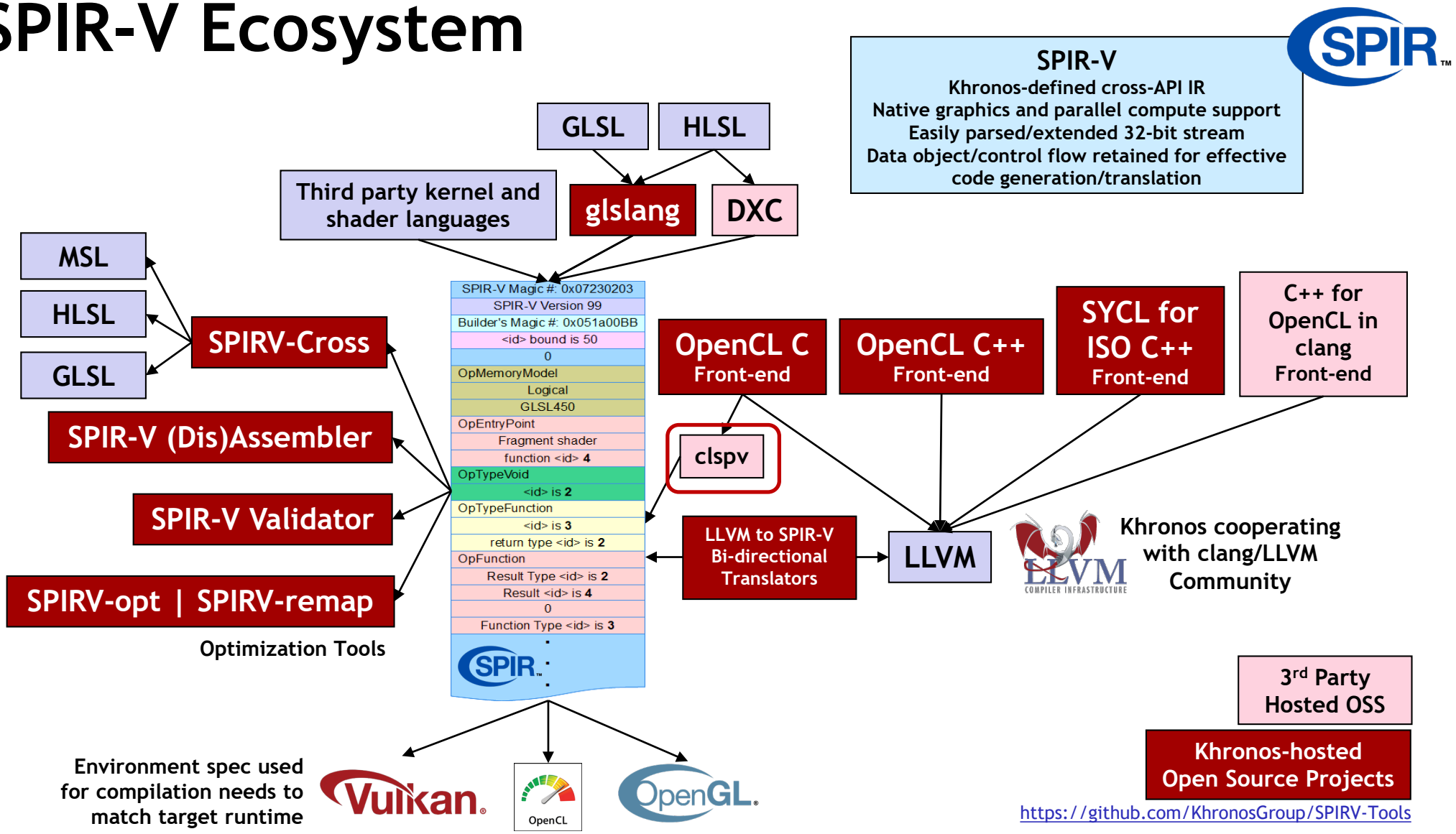
Examples: Variadic Macros, Atomic Functions w/ Address Spaces, Templates are possible with no SPIR-V changes

Community OSS-Driven Language Evolution

- **Language capabilities can move faster with offline compilation**
  - Benefits Khronos: easier to explore and iterate on new features
  - Benefits implementers: no driver updates required for new language features
  - Benefits developers: one consistent tool enables all implementations
- **Proposal -> accelerate towards community-based language ecosystem**
  - Host non-ratified offline language documentation at Khronos – agile updates
  - Add Compiler Capabilities extension to core OpenCL – using preprocessor #defines
  - Code can choose to interrogate extension for enhanced compiler capabilities
- **N.B. NOT proposing to remove OpenCL C online compilation!**
  - Key to many use-cases – and can also absorb new features over time
  - https://github.com/KhronosGroup/OpenCL-Docs/issues/65

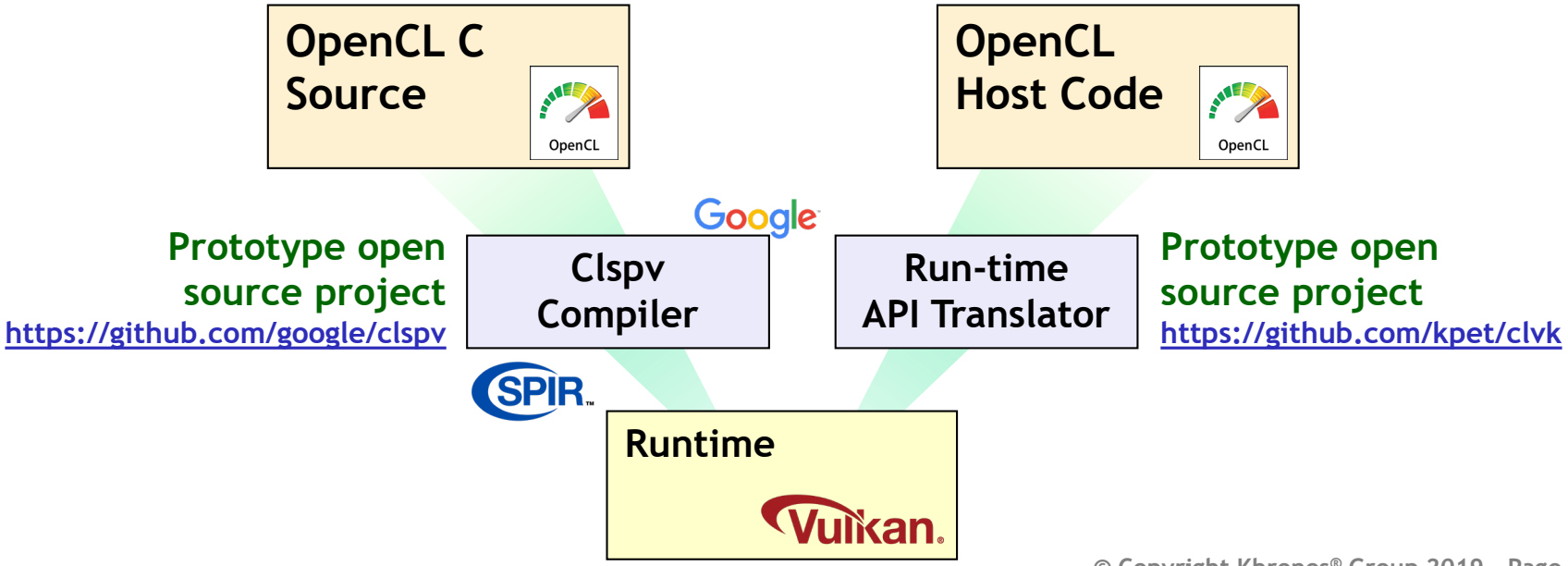


SPIR-V Ecosystem

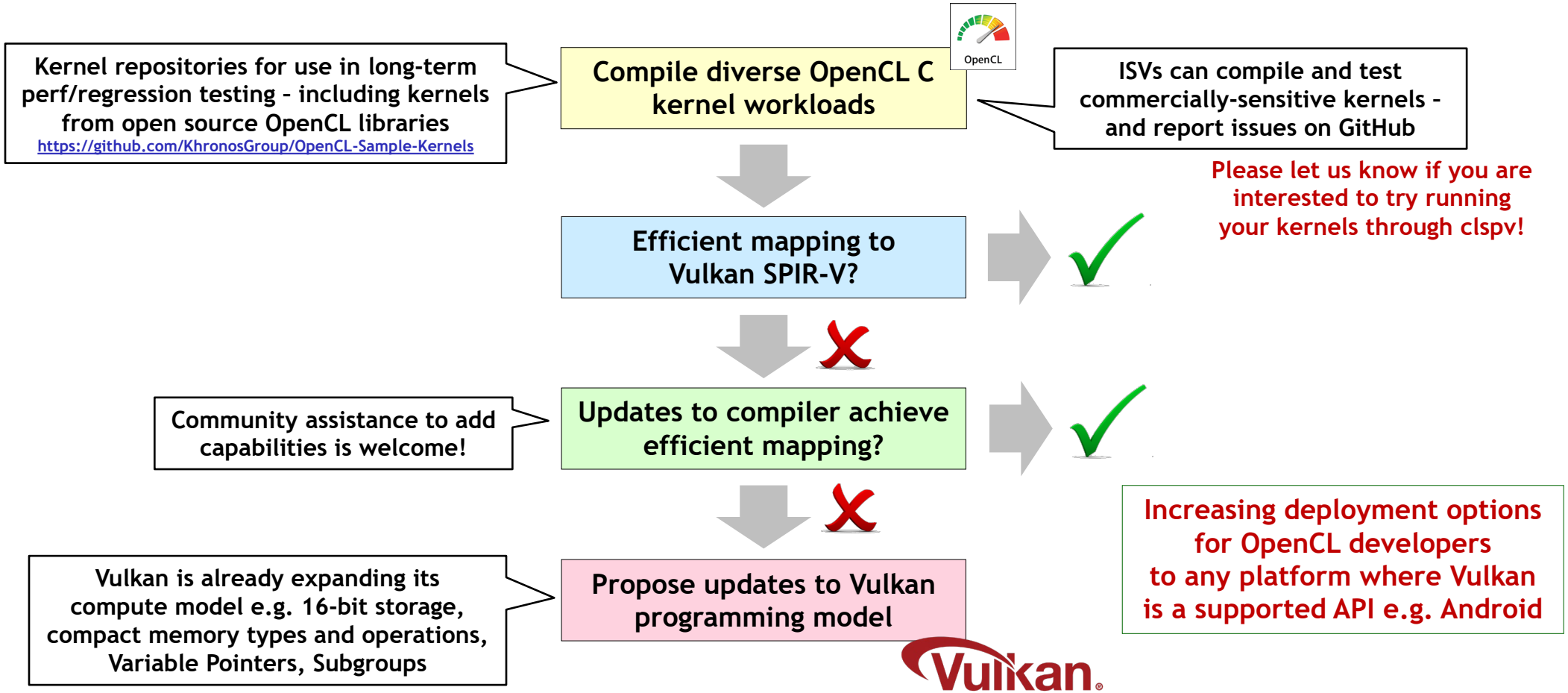


OpenCL Platform Deployment Flexibility

- **Clspv - Google's experimental compiler for OpenCL C to Vulkan SPIR-V**
  - Open source - tracks top-of-tree LLVM and clang, not a fork
  - Originally tested on over 200K lines of Adobe OpenCL C production code
  - Sony is now working with Google to compile their production kernels
    - 355 kernels in 75 files - 40 files compiled successfully at first pass
- **Clvk - experimental OpenCL to Vulkan API shim by Kevin Petit**
  - Early days – but Halide's OpenCL back-end successfully running over Vulkan



Refining clspv with Diverse Workloads



Vulkan Portability Initiative on Apple

Almost all mandatory Vulkan 1.0 functionality is supported:
No Triangle Fans
No separate stencil reference masks
Events are not supported

Selected Optional Features and Extensions are added as required - driven by industry input and feedback
Robust buffer access
BC texture compressed formats
Fragment shader atomics

https://github.com/KhronosGroup/MoltenVK

Open source SDK to build, run, and debug applications on macOS - including validation layer support
https://vulkan.lunarg.com/

LUNARG

SPIRV-Cross
Convert SPIR-V shaders to Metal Shaders

KHRONOS GROUP

macOS / iOS Run-time
Maps Vulkan to Metal API

MoltenVK supports macOS 10.11 / iOS 9.0 and up

gfx-rs

Open source beta release for macOS

Molten VK

OPEN SOURCE.
Free to use - no fees or royalties - including commercial applications

KHRONOS GROUP

Apps Shipping On Apple with Vulkan Backend

Forsaken Remastered was just updated with **Vulkan** support! If you're on Linux, you're probably hitting 60fps with the existing OpenGL renderer, but it's good to be future proof. If you're on a Mac, though, you *definitely* want to switch. On my MacBook, the framerate goes from around 15 to a solid 60!

Initial Vulkan Performance On macOS With Dota 2 Is Looking Very Good

Written by Michael Larabel in Valve on 1 June 2018 at 05:37 PM EDT. 34 Comments

Yesterday Valve released Vulkan support for Dota 2 on macOS. Indeed, this first major game relying upon MoltenVK for mapping Vulkan over the Apple Metal drivers is delivering performance gains.

Valve Releases Artifact As Its Cross-Platform, Vulkan-Powered Digital Card Game

Written by Michael Larabel in Valve on 28 November 2018 at 04:16 PM EST. 29 Comments

Valve managed to ship their latest game today as planned and without any major delays.

Artifact is now available with launch-day support for Linux, macOS, and Windows. Artifact is a competitive digital card game; game is targeting Dota 2 players as well as card gaming enthusiasts. Valve still plans to evolve Artifact and its gameplay and

DOTA 2

Production Dota 2 on Mac Ships - up to 50% more perf than Apple's OpenGL

First iOS Apps using MoltenVK ship through app store

Qt

Qt Running on Mac through MoltenVK

Multiple iOS and macOS apps organically ported - support through MoltenVK website e.g. Forsaken Remastered on Mac

WINE

Initial ports of Wine games in progress using Vulkan on Mac

Filament

Google Filament PBR Renderer on Mac

ARTIFACT

Artifact from Steam ships on MoltenVK on macOS - first Vulkan-only Valve app on Mac

RPCS3

RPCS3 PlayStation 3 Emulator on Mac

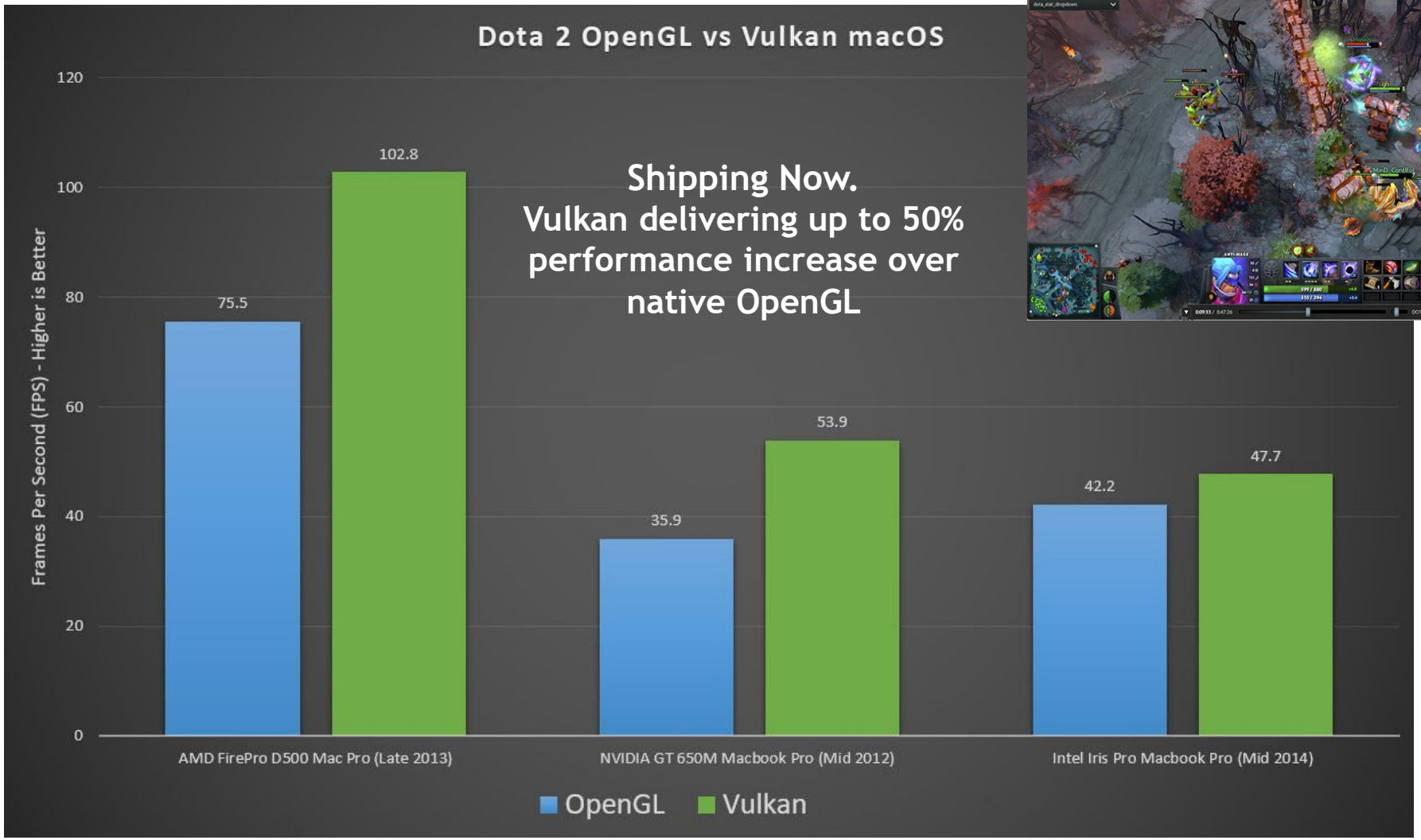
Dolphin GameCube and Wii Emulator working on MacOS

Diligent Graphics

Diligent Engine runs on MacOS

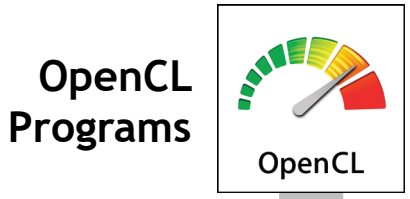
June 2018 | **September 2018** | **November 2018** | **January 2019**

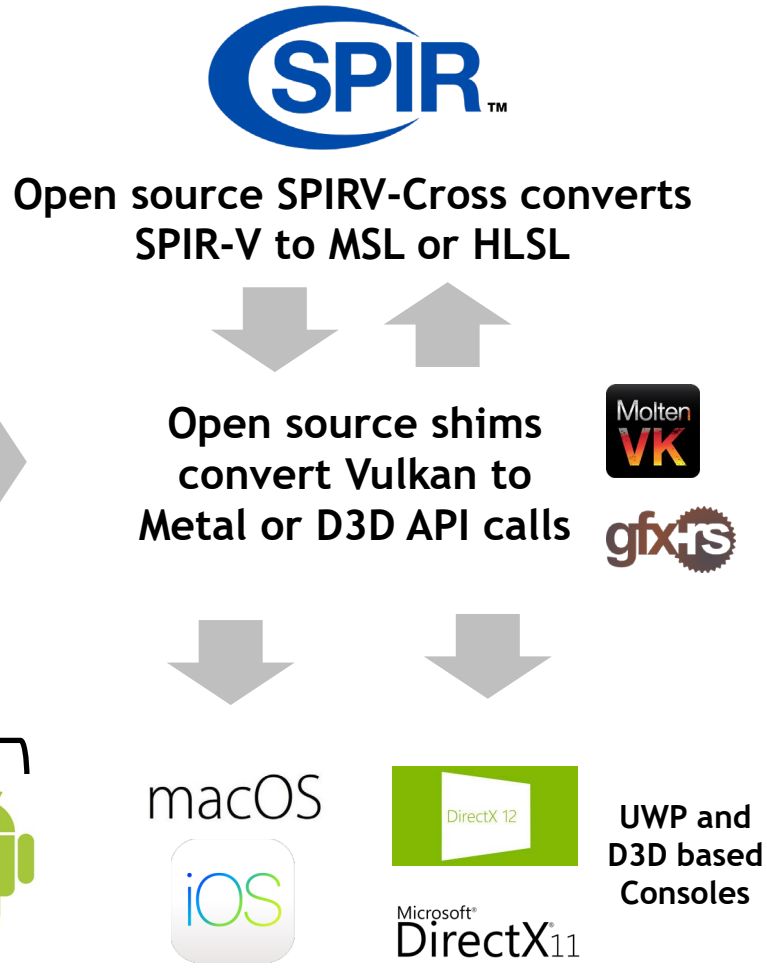
Valve - Vulkan Dota 2 on macOS



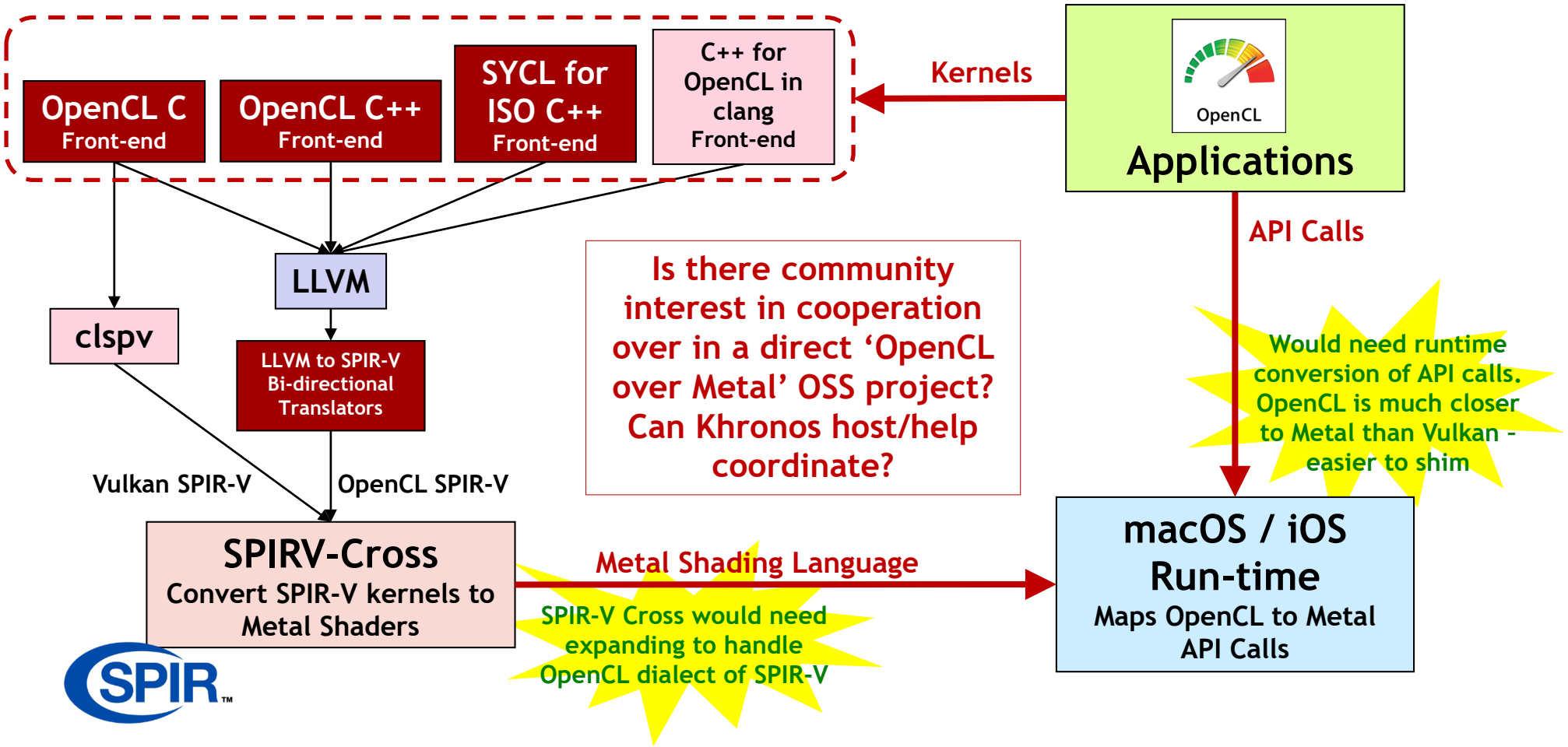
Universal Deployment Flexibility?

Open source tools enable OpenCL and Vulkan apps to be increasingly deployed on many platforms





OpenCL Portability Initiative on Apple??



Get Involved!

- **OpenCL is driving to new levels of usability and deployment flexibility**
  - We want to know what *you* need from OpenCL!
- **OpenCL Next and Feature Sets**
  - Let us know what you think!
- **Multiple ways to engage and help OpenCL evolve**
  - Join Khronos for a voice and a vote in any of these standards
  - Or ask about an invite to the OpenCL Advisory Panel
  - Or consider getting involved in OpenCL OSS projects
    https://github.com/KhronosGroup
  - Or talk to us on Slack and the forums
  - https://community.khronos.org/c/opencl
  - www.khr.io/slack
- **Neil Trevett**
  - ntrevett@nvidia.com
  - @neilt3d
  - www.khronos.org

If you need OpenCL let your hardware vendors know! Your voice counts!