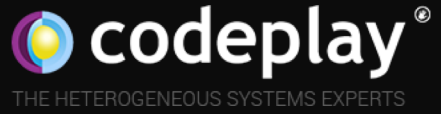


TensorFlow acceleration for neural network inference using SYCL ecosystem

May 2018, Embedded Vision Summit

Andrew Richards

TensorFlow on SYCL & OpenCL

- This is open-source TensorFlow, ported to support SYCL
- You can build TensorFlow & run on any accelerator that supports the right OpenCL features
  - But you may need to adapt some operators to achieve performance
- This is the same source base as the CUDA version
- This work supports training & inference and tools like TensorBoard
- This is separate from projects that extract TensorFlow graphs and execute them on a different inference engine

Codeplay provides the standard tools for transforming AI prototypes into volume products

TensorFlow

Prototype to Product

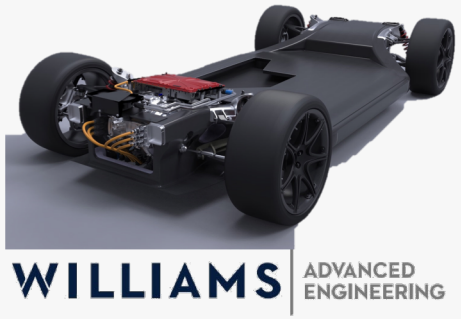
- Reduce power
- Reduce cost
- Standard programming models
- Safety qualification

Codeplay works in partnership to build industry standard solutions

Funding

Jez San: his company, Argonaut, created first videogames GPU (SuperFX)

Foresight group



WILLIAMS ADVANCED ENGINEERING

Standards

KHRONOS GROUP

HSA FOUNDATION

C++

MISRA

Processors

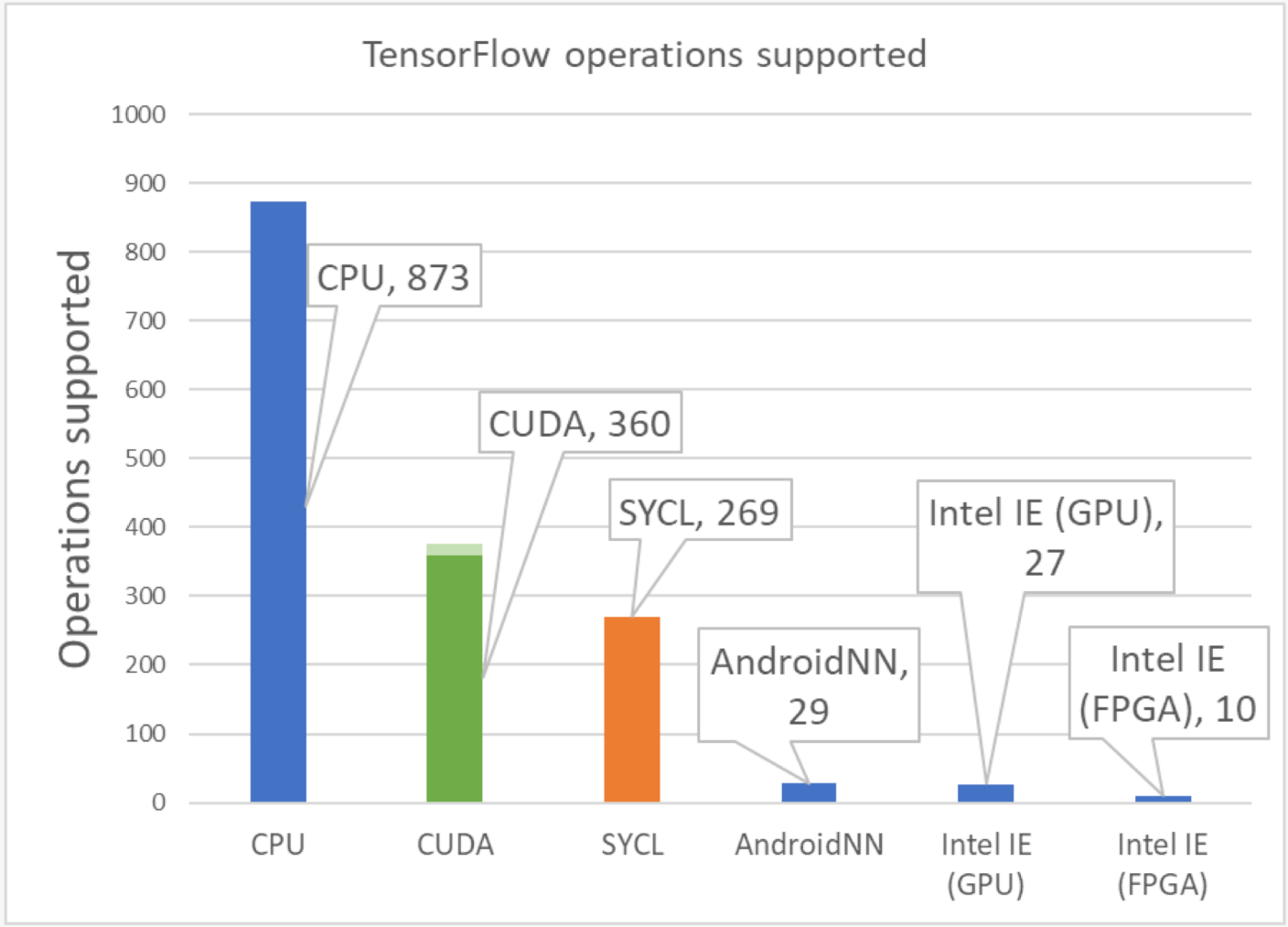
ARM Mali

Renesas R-Car

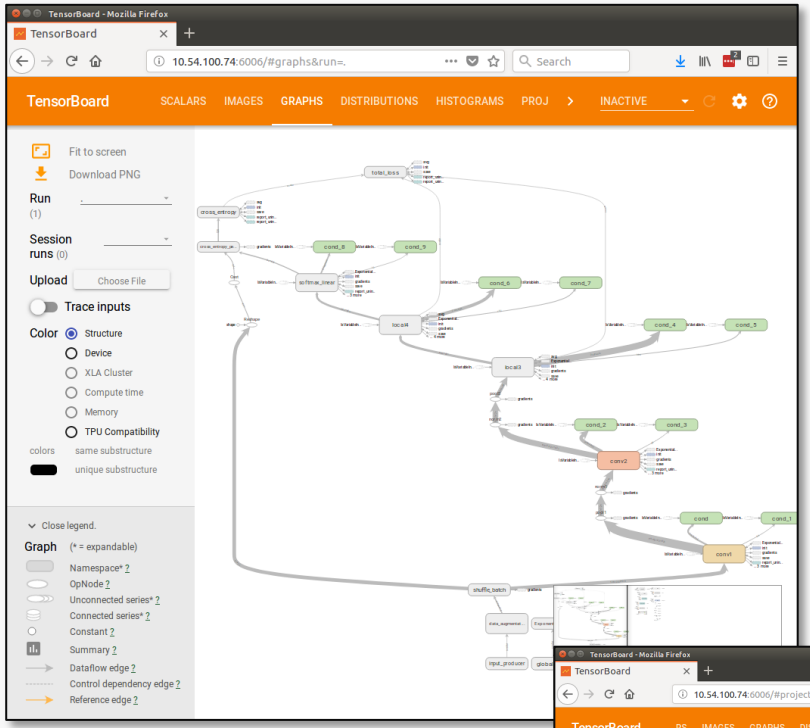
+ others unannounced

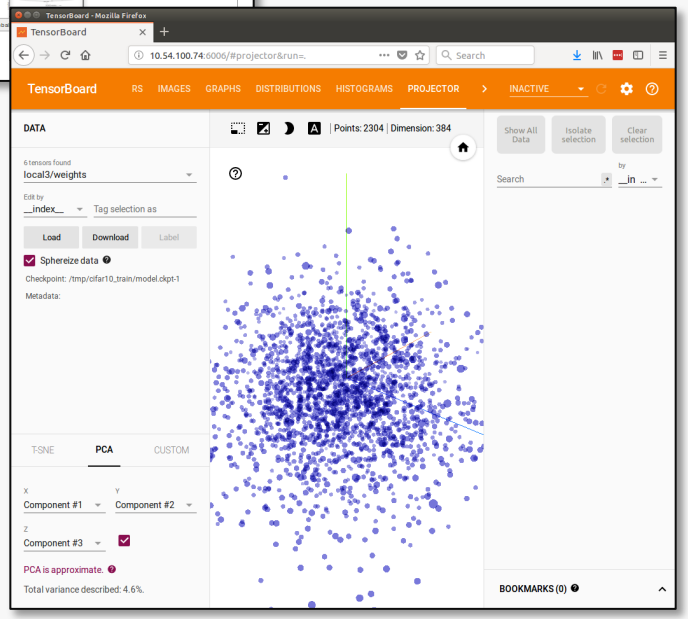
SYCL-TensorFlow

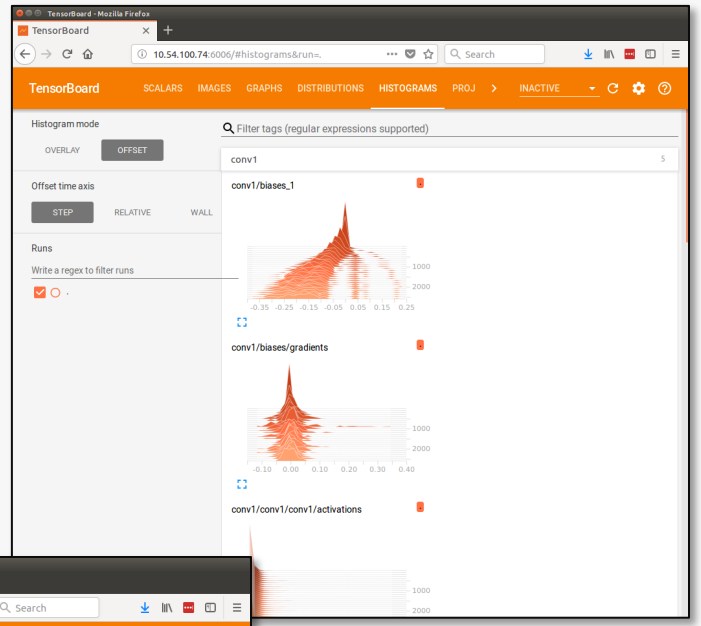
- ✓ Port of whole of TensorFlow to support SYCL and OpenCL
- ✓ Can take existing Python TensorFlow scripts and just run them as-is
- ✓ Training & Inference
- ✓ Profiling/tracing
- ✓ Custom operations, as they are added
- ✓ Open-source: can check out code from TensorFlow repository
- ✓ Python, C and C++ interfaces
- ✓ Continuous integration testing
- ✓ 37% of TensorFlow core operations, compared with CUDA's 54% and 24% with XLA
- ✓ (XLA falls back on non-XLA code for many ops)

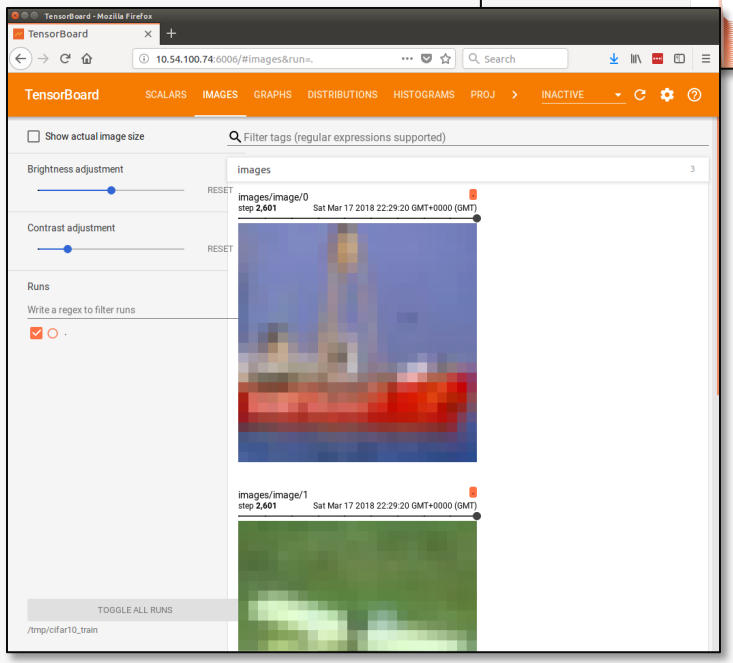


TensorBoard

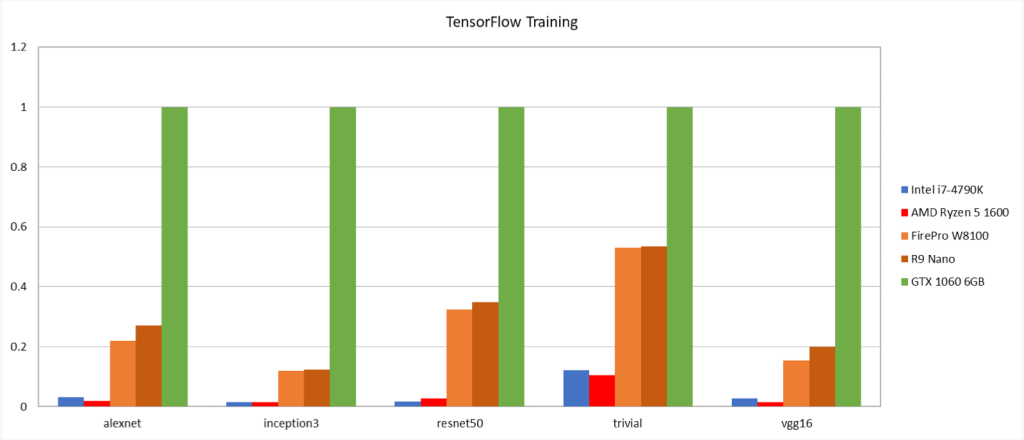


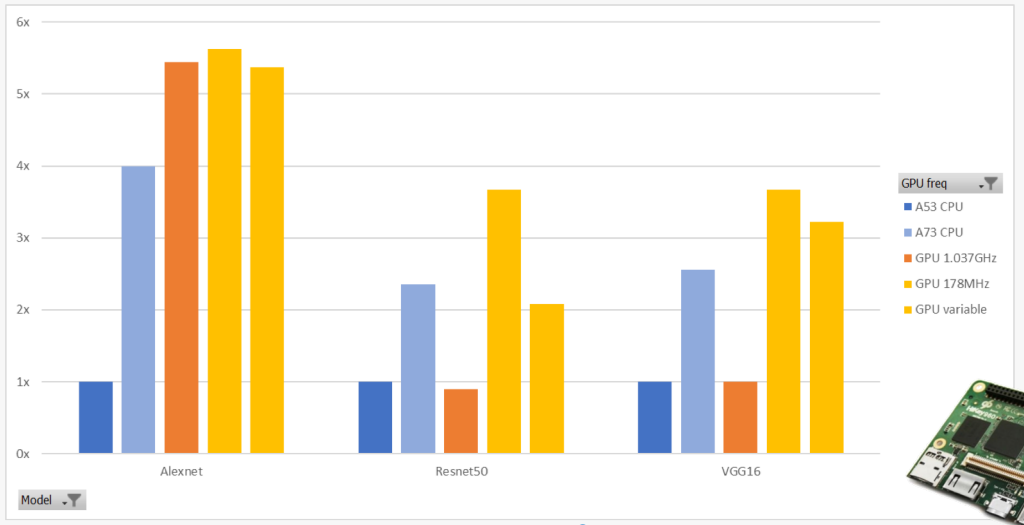




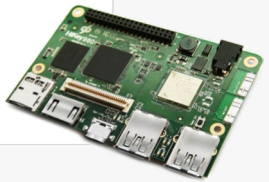


TensorFlow/Eigen performance



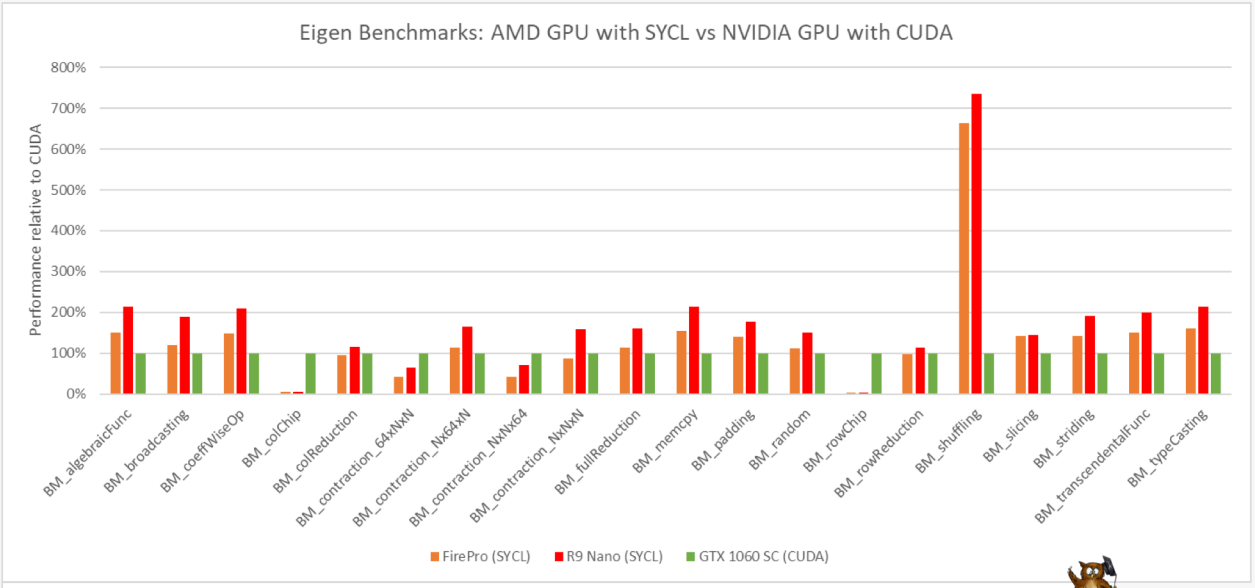


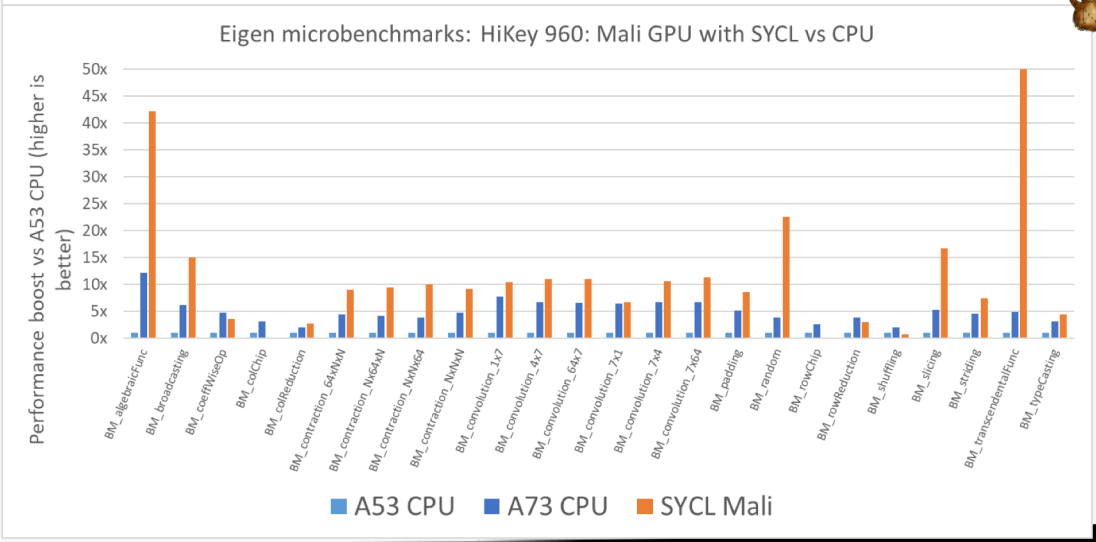
TensorFlow



Benchmarks taken from:

https://github.com/tensorflow/benchmarks

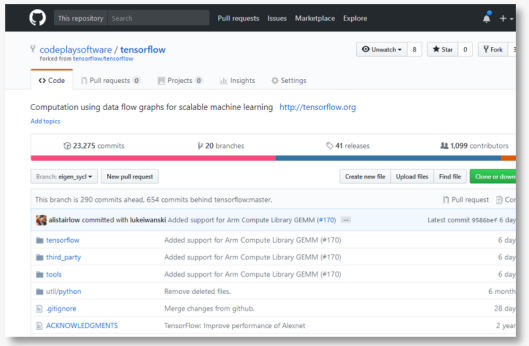


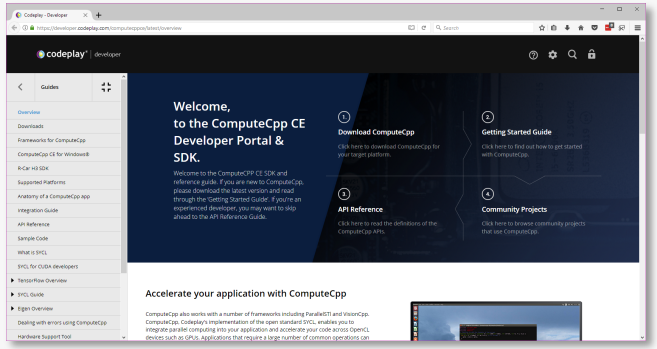


Getting TensorFlow using OpenCL/SYCL

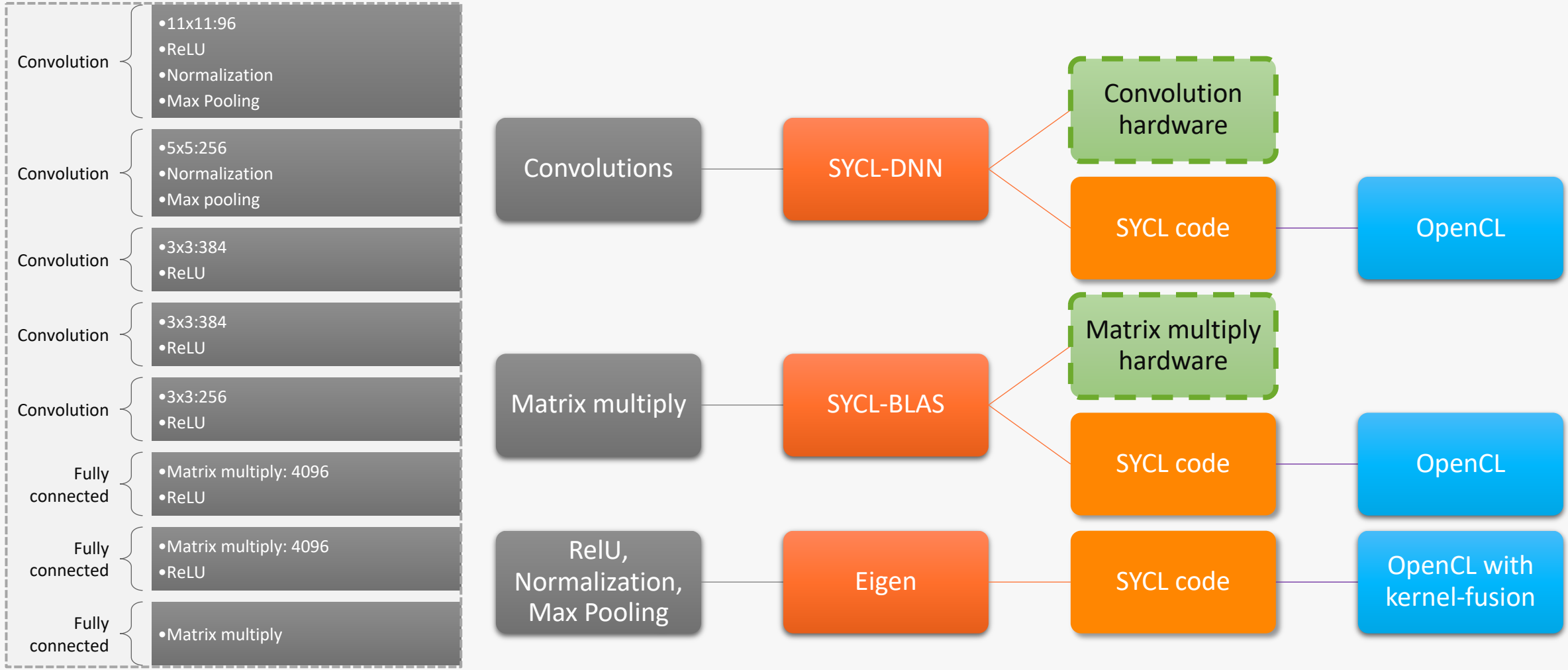
You need to rebuild from source

- Checkout the TensorFlow source from github (currently Mali support is only in Codeplay's github):
  - https://github.com/codeplaysoftware/tensorflow/tree/eigen_sycl
- Download the ComputeCpp for Arm package from Codeplay website:
  - http://www.codeplay.com/computecpp
- Download OpenCL implementation (with SPIR-V support) for your GPU or AI accelerator
- Compile and run:
  - Follow normal TensorFlow build options (using bazel)
  - Select OpenCL support during configuration
  - Need to provide the ComputeCpp directory

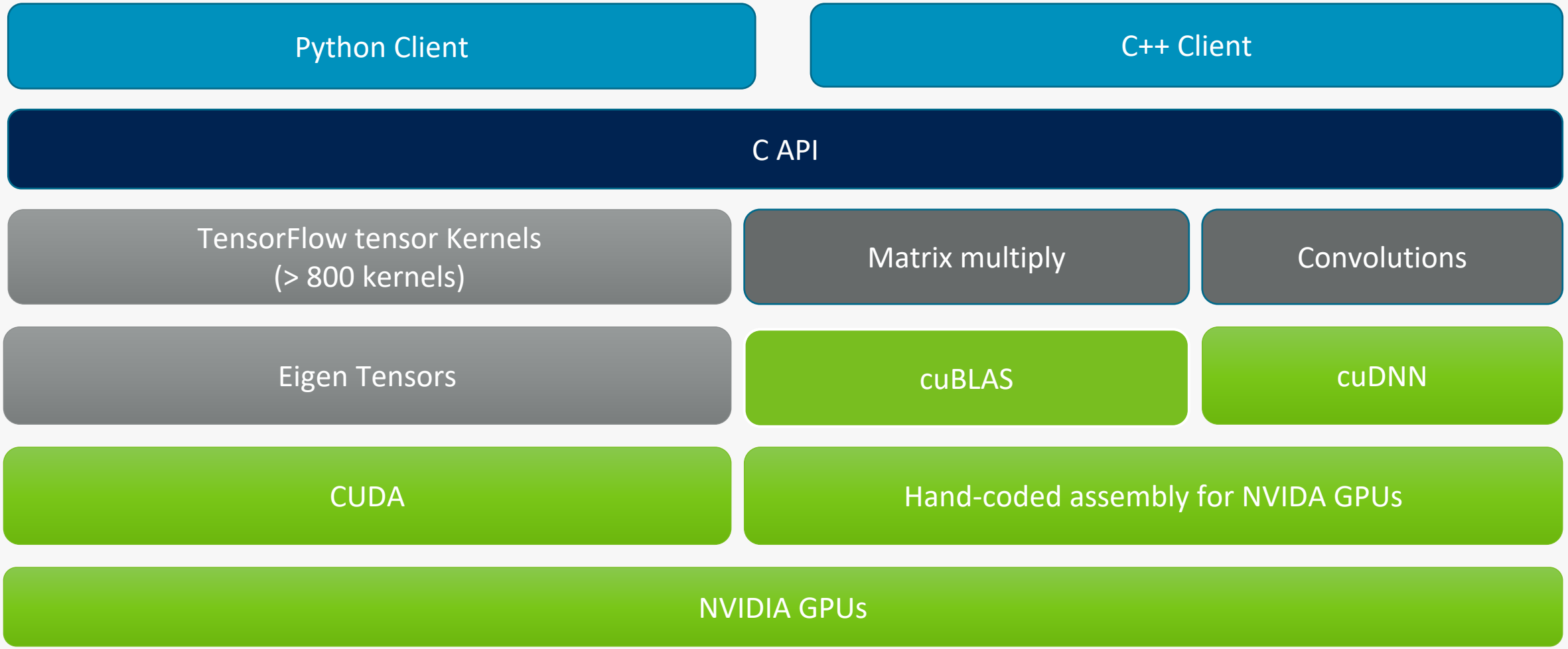




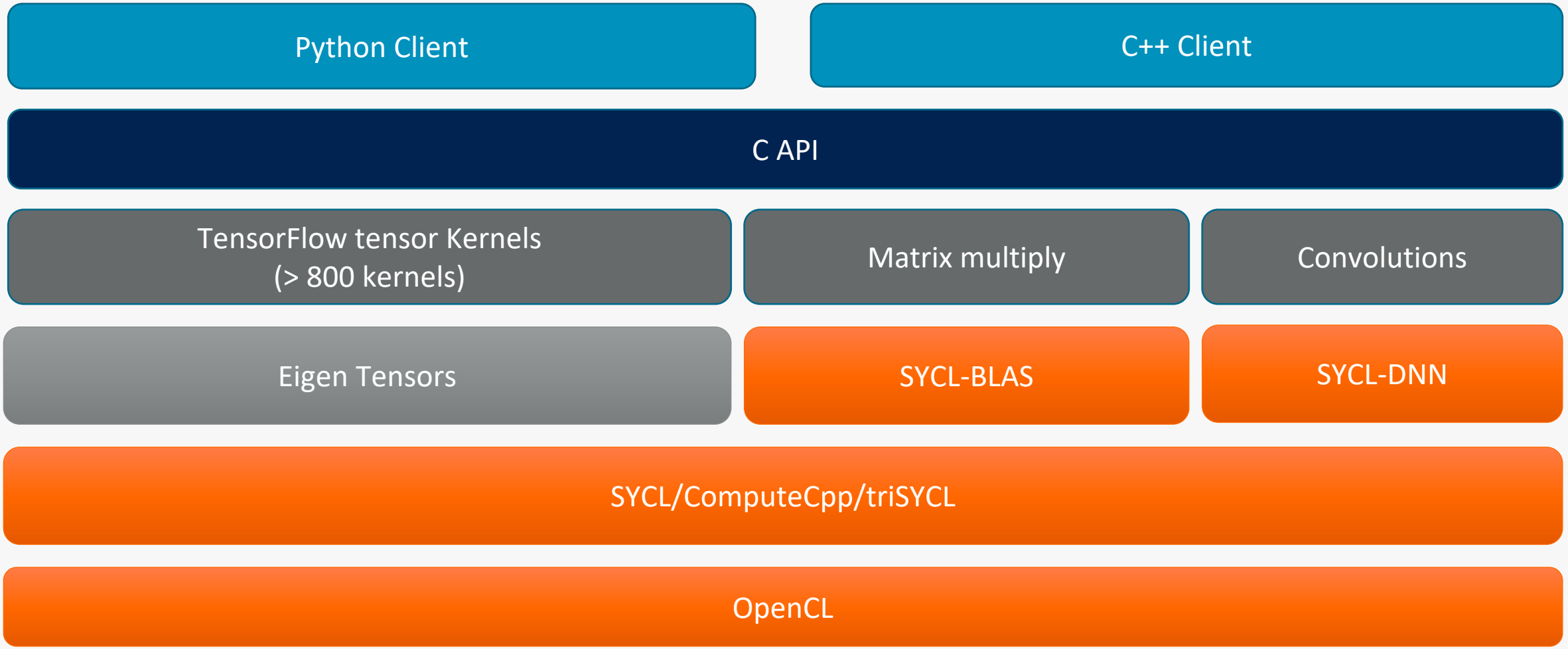
Mapping a neural network to SYCL



TensorFlow on CUDA



TensorFlow on SYCL / OpenCL



Optimizing TensorFlow

SYCL-DNN

- Highest impact on performance
- Algorithms need to be adapted to each new processor architecture
- Or, alternatively, make use of hand-coded assembly
- Or, make use of convolution hardware
- Some convolution operations make use of matrix multiplies (either SYCL-BLAS, or Eigen tensor contractions)

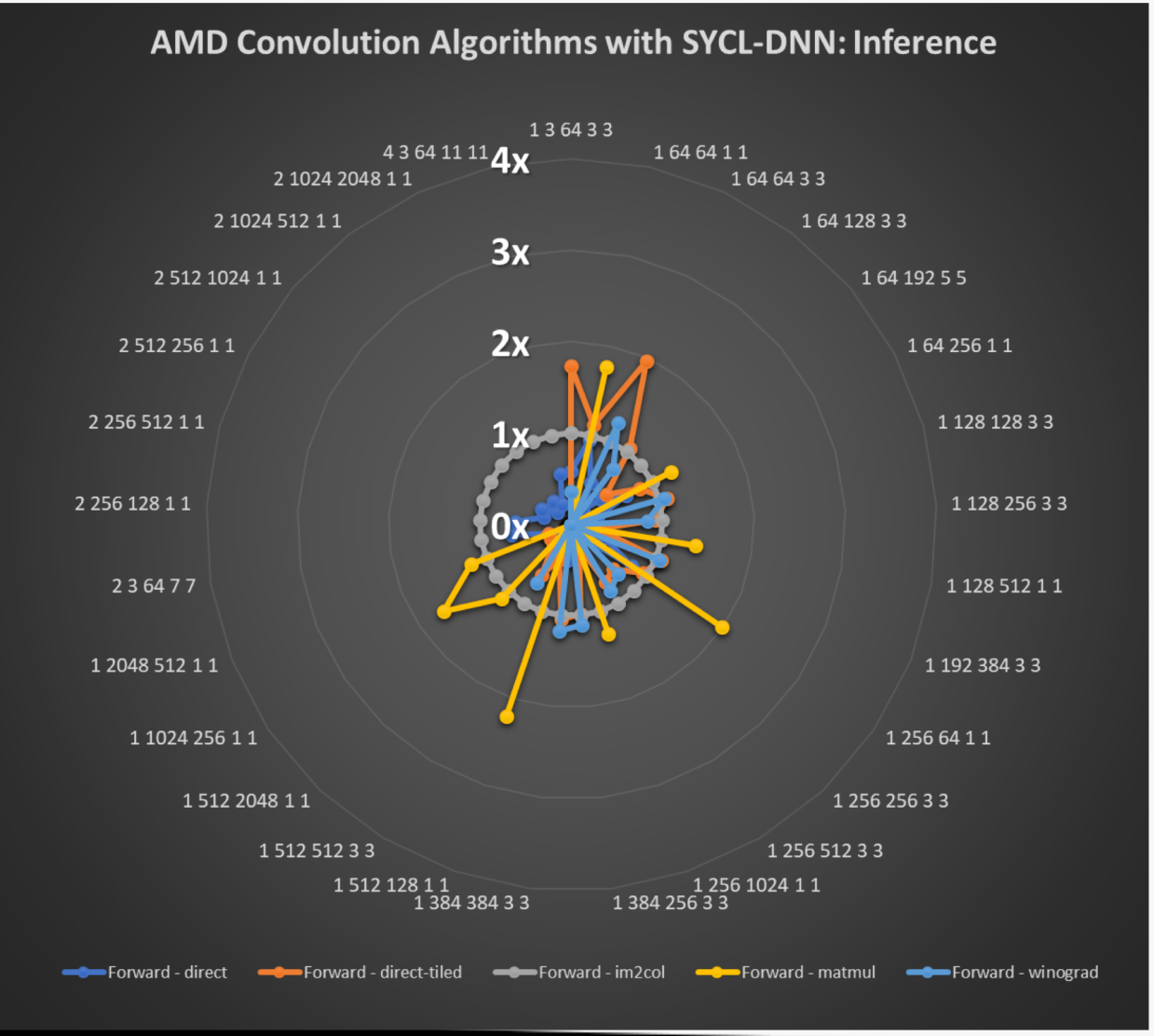
SYCL-BLAS

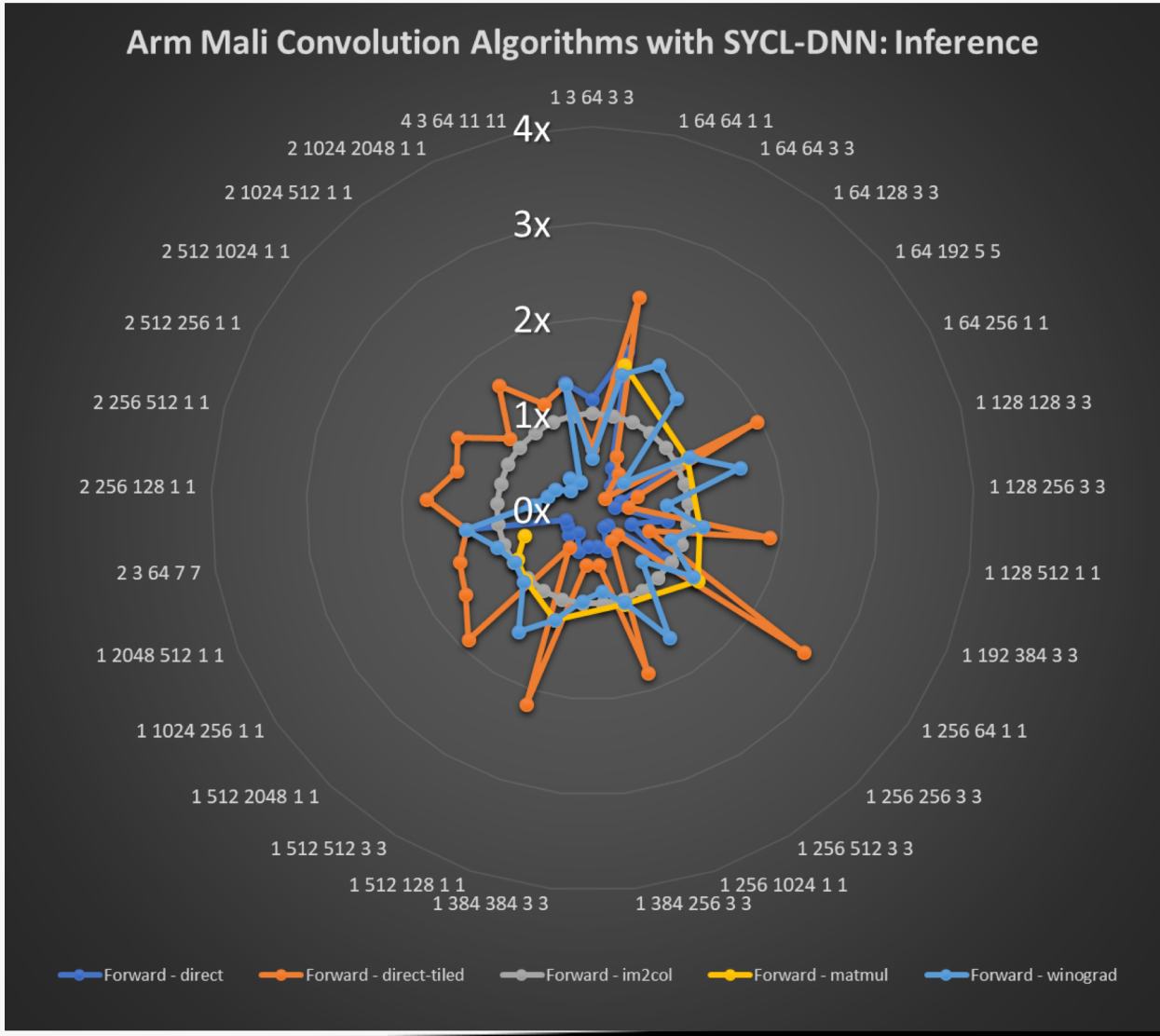
- Currently, we actually use Eigen tensor-contract
- The plan is to switch to SYCL-BLAS, making code-sharing easier
- Handles matrix multiplies
- Algorithms need to be adapted to each new processor architecture
- Or, alternatively, make use of hand-coded assembly
- Or, make use of matrix multiply hardware
- Necessary to have performance for a very wide range of matrix shapes & sizes

Eigen

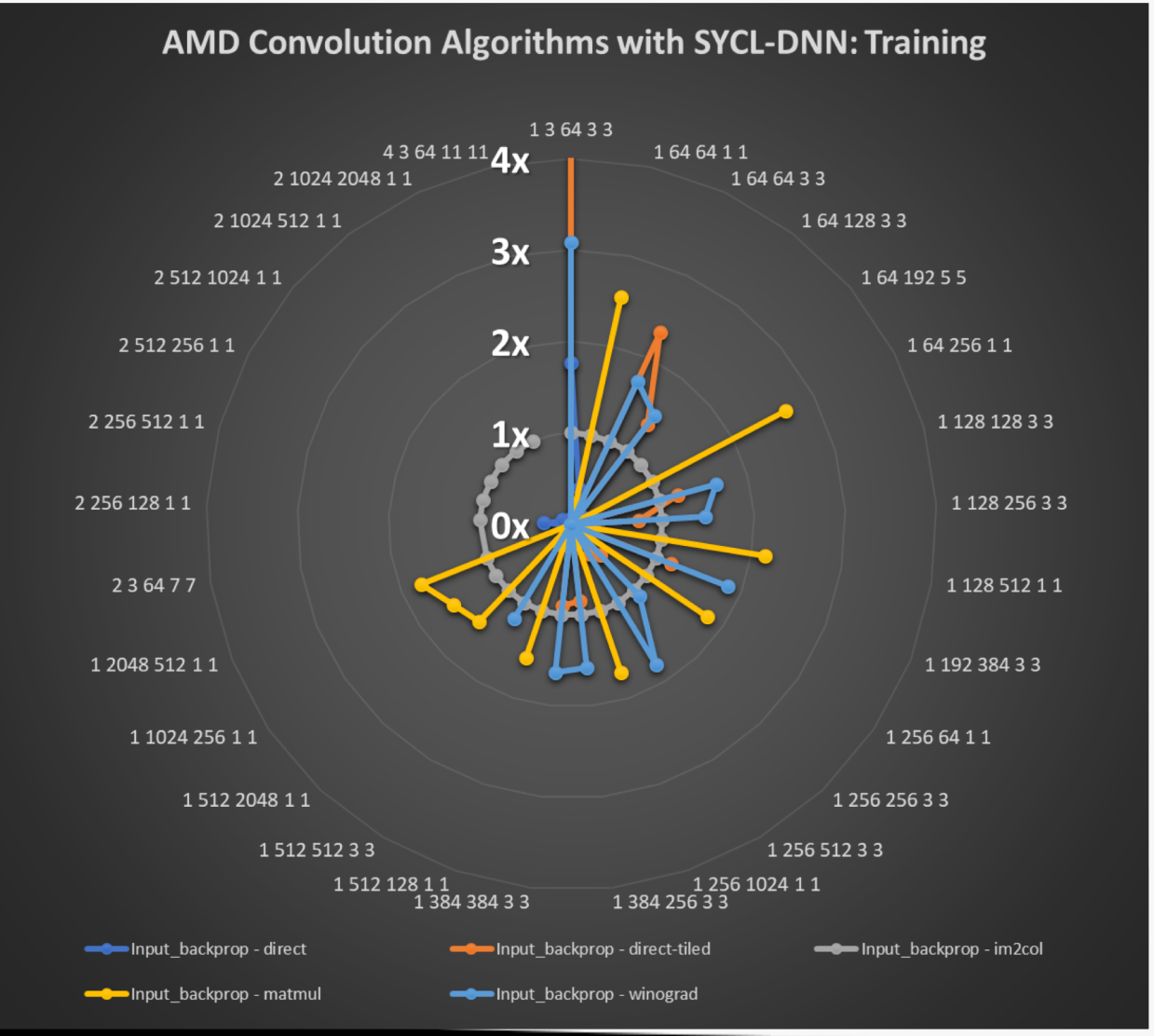
- The generic algorithms must be adapted for each architecture:
  - Component-wise operations
  - Reduction
  - Partial reduction
  - Shuffles and chipping
  - Tensor contractions
- All open-source
- Source code is shared between CPU, CUDA and SYCL
- Algorithms are written as C++ templates and can be configured per-device

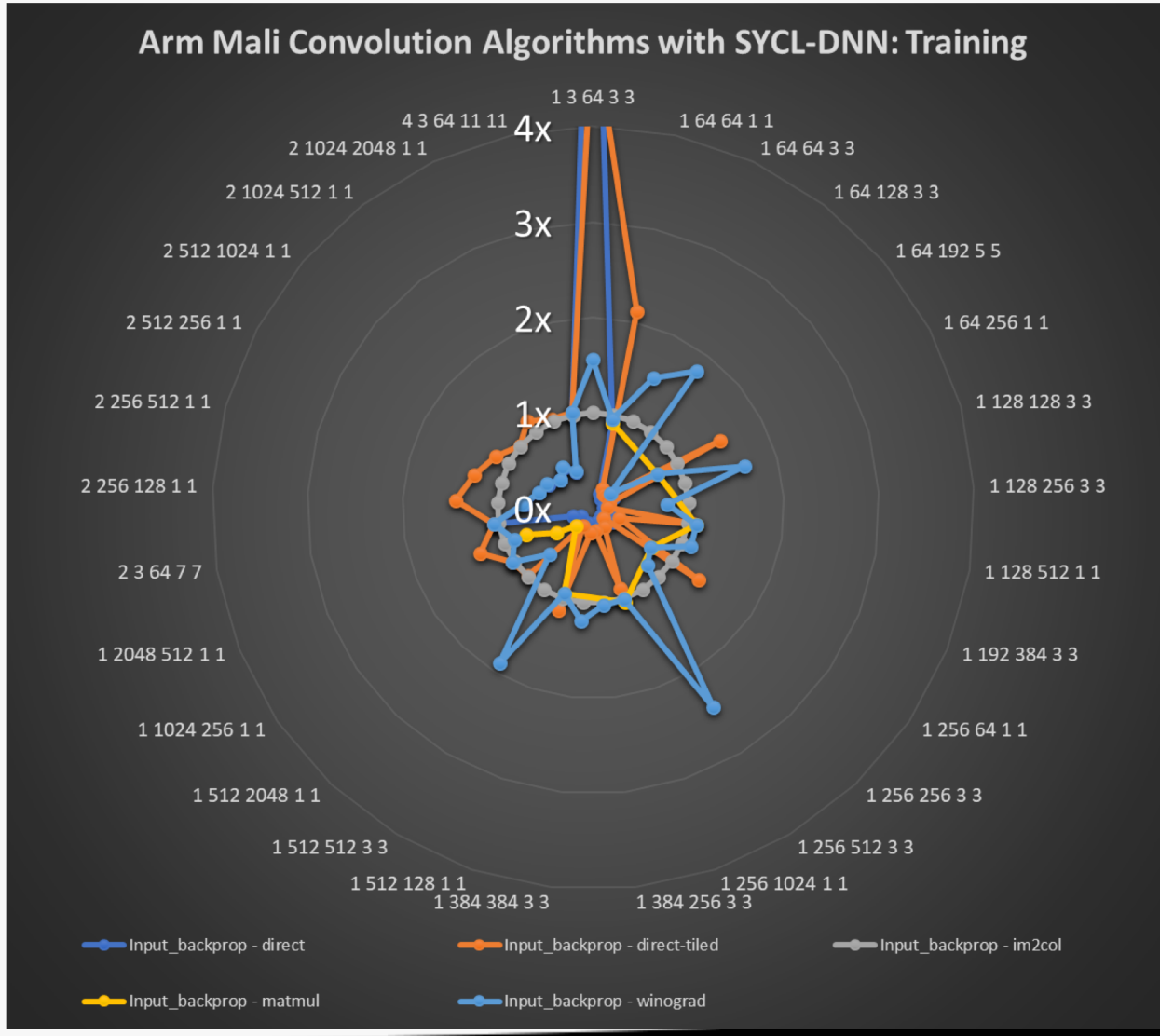
SYCL-DNN: algorithms performance (inference)





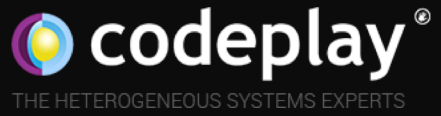
SYCL-DNN: algorithms performance (training)





What's next?

- Some further up-streaming required
  - Codeplay's branch has the latest support
- Open-source and integrate SYCL-DNN
- Integrate SYCL-BLAS
- More accelerator processor support
- Setup more continuous integration testing
- It's all open-source, open-standard, so easy to customize



Q&A