Annual Meeting

oneAPI Community Forum

Agenda

oneAPI Community Forum Update	Get a roundup of some of the highlights from 2022 as we close off the year and find out about some changes to the organization	Rod Burns, Codeplay Software
Breakout Room Discussions	We will break out into rooms on a range of topics for discussion.	•oneAPI 2023 and beyond •AI •Hardware Abstraction •Language
Stories from 2022	Hear from some of our members on what they have contributed to oneAPI this year.	Kevin Harms, Gordon Brown, Kentaro Kawakami and Robert Cohn
The future - a cross vendor, industry standard programming	Find out what the future can look like for heterogeneous computing	Andrew Richards, Codeplay CEO

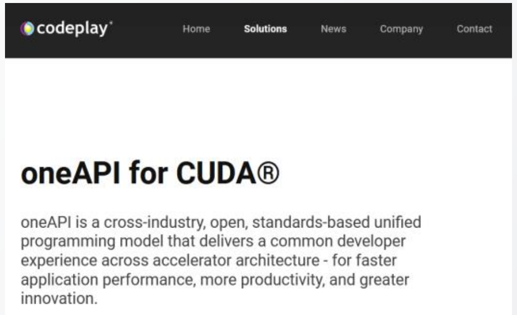
2022 oneAPI Roundup

Intel oneAPI GPU Rendering Appears Ready For Blender 3.3

Written by Michael Larabel in Intel on 13 July 2022 at 05:00 AM EDT. 10 Comments

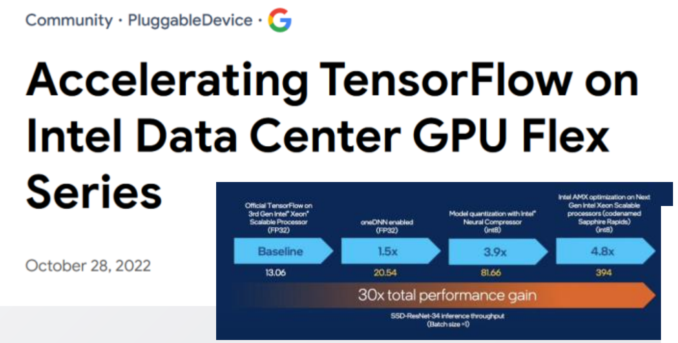
Intel's effort to add oneAPI/SYCL support to Blender for GPU acceleration with forthcoming Arc Graphics hardware appears all buttoned up for the upcoming Blender 3.3 release.

As I've written about before, for months Intel engineers with the Blender community have been working on oneAPI GPU rendering support for the Cycles engine to complement the NVIDIA CUDA/OptiX and AMD HIP targets. Since Blender 3.0 dropped OpenCL, this oneAPI target will be the only option for GPU acceleration with Intel discrete GPUs for this 3D modeling software.



oneAPI for CUDA®

oneAPI is a cross-industry, open, standards-based unified programming model that delivers a common developer experience across accelerator architecture - for faster application performance, more productivity, and greater innovation.

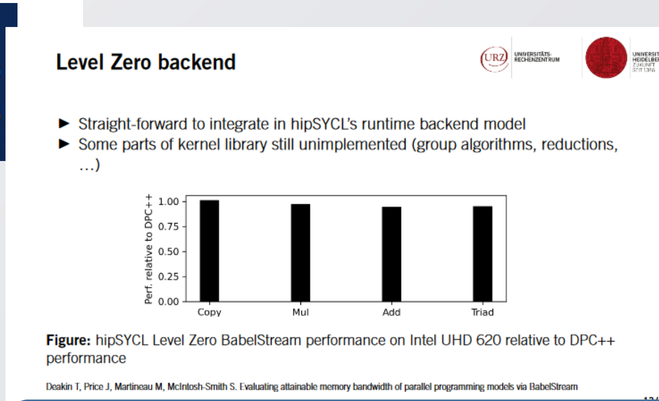


Community · PluggableDevice ·

Accelerating TensorFlow on Intel Data Center GPU Flex Series

October 28, 2022

Official TensorFlow on 3rd Gen Intel® Xeon® Scalable Processor (FP32)
Baseline
13.06
oneDNN enabled (FP32)
1.5x
20.54
Model quantization with Intel® Neural Compressor (int8)
3.9x
81.66
Intel AMX optimization on Next Gen Intel Xeon Scalable processors (codenamed Sapphire Rapids) (int8)
4.8x
394
30x total performance gain
SSD-ResNet-34 inference throughput (Batch size =1)



Level Zero backend

- Straight-forward to integrate in hipSYCL's runtime backend model
- Some parts of kernel library still unimplemented (group algorithms, reductions, ...)

Figure: hipSYCL Level Zero BabelStream performance on Intel UHD 620 relative to DPC++ performance

Deakin T, Price J, Martineau M, McIntosh-Smith S. Evaluating attainable memory bandwidth of parallel programming models via BabelStream

hipSYCL support for Level Zero and oneMKL

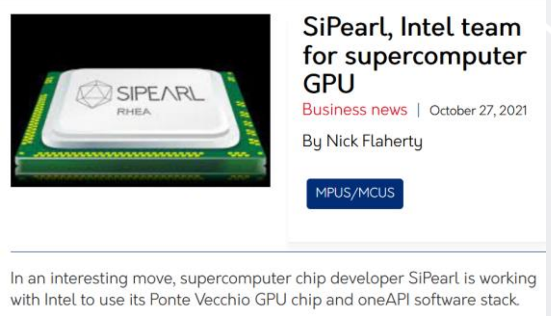
Tom Deakin
@tjdeakin

Busy day @comp_sci_durham at the SYCL Practitioners Hackathon hosted by @hpcsoftware_ - two codes and three different systems!

12:30 PM · Nov 23, 2022

Scientific Computing at Durham

SYCL Practitioners Hackathon 2022 (user group meeting)



SiPearl, Intel team for supercomputer GPU

Business news | October 27, 2021

By Nick Flaherty

MPUS/MCUS

In an interesting move, supercomputer chip developer SiPearl is working with Intel to use its Ponte Vecchio GPU chip and oneAPI software stack.

Huawei Extends DPC++ with Support for its Ascend AI Chipset

OCTOBER 28, 2021

Applying the oneAPI Industry Initiative to Real-World AI Use Cases

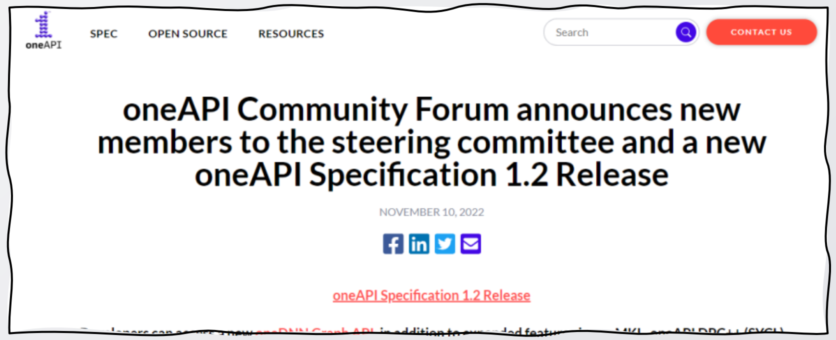
HUAWEI CONNECT
毕昇C++：SYCL在鲲鹏+昇腾上的落地实现
C++的优势
SYCL异构框架
https://www.khronos.org/sycl/
HUAWEI

oneAPI Initiative to Community Forum

- Intel managed oneAPI initiative → Moved to an open forum
- Codeplay is running the open community forum
- New Spec 1.2 released

- New Steering Committee Members:

Penporn Konantakool (Google)
Kevin Harms (Argonne National Lab),
Antonio Pena (Barcelona Supercomputer Centre)



oneAPI Community Forum announces new members to the steering committee and a new oneAPI Specification 1.2 Release

NOVEMBER 10, 2022

oneAPI Specification 1.2 Release

oneAPI Specification 1.2 released Nov. 10, 2022

oneAPI Deep Neural Network Library (oneDNN): graph API has been added, which compiles and executes a deep learning computation graph, identifying opportunities for fusing operators and other target-specific optimizations, working closely with industry partners who develop the major frameworks.

DPC++ (oneAPI's open source SYCL implementation): better management of contexts, queues, and memory management, and enhanced support for images.

oneMKL: enhancements for the BLAS libraries with new routines, including support for half/bfloat16, including dense matrix copy and transpose routines as well as updates for BLAS GEMM and GEMM batch.

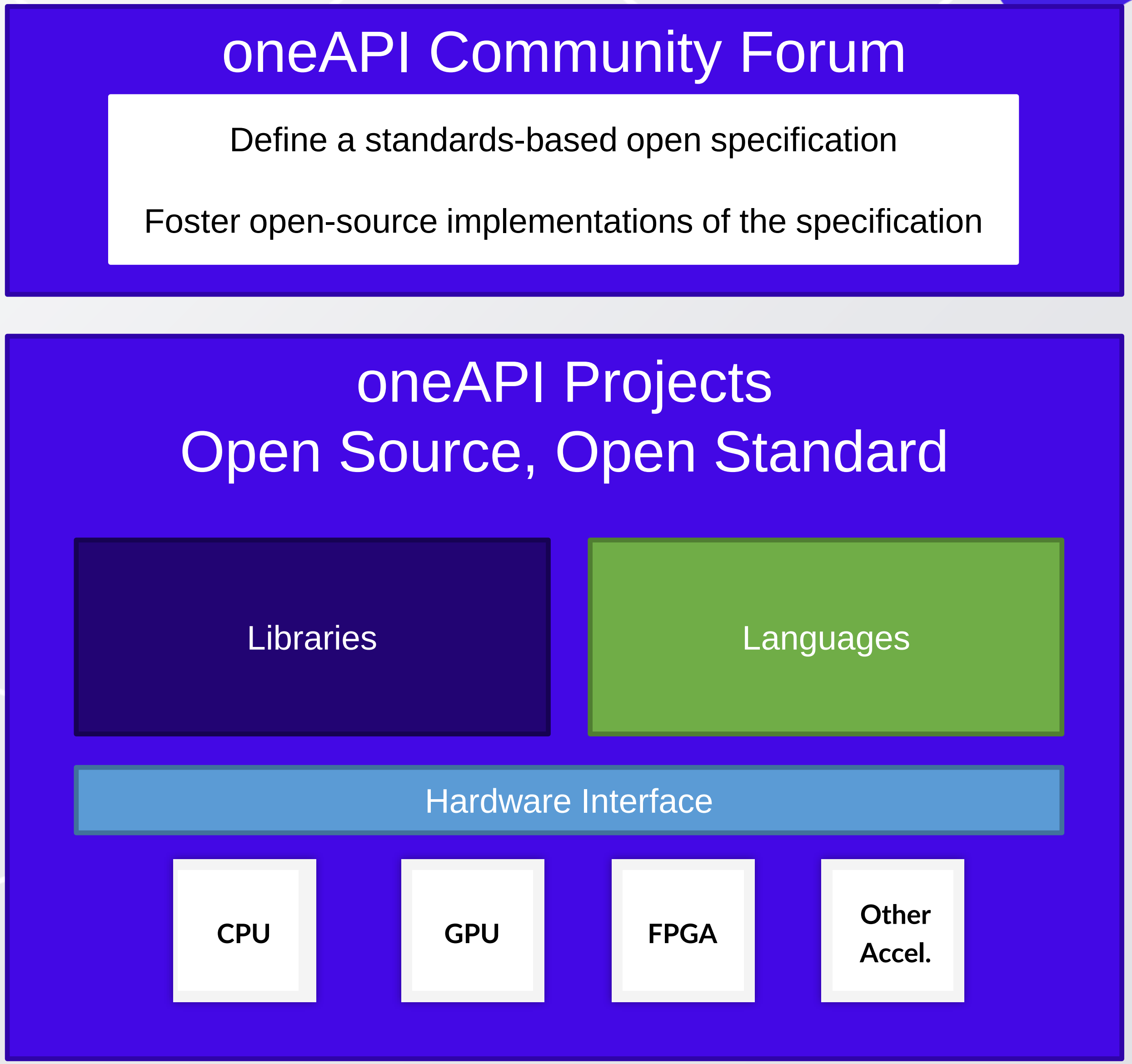
oneVPL added a new API for processing camera RAW data and support for more video color formats.

Level Zero added a fabric topology discovery API, support for sRGB, support for image copy with pitch as well as clarifications on existing API.

SC22 oneAPI Community Forum Meetup

oneAPI Mission

- Industry defined, open standard-based APIs for accelerated devices
- Bring open source implementations for wide adoption
- Gather industry leaders and contributors to support the mission
- Enable diverse processor designs



oneAPI Community Forum

Define a standards-based open specification

Foster open-source implementations of the specification

oneAPI Projects
Open Source, Open Standard

Libraries

Languages

Hardware Interface

CPU

GPU

FPGA

Other Accel.

Evolving TABs

What is the oneAPI Community Forum?

1

A cross industry group of hardware and software experts

2

Defines standard interfaces for accelerator computing

3

Multiple specialist technical working groups

4

Drives the future of open-standard accelerator computing

oneAPI Community Forum Steering Committee Formation

- Individuals are being appointed to roles on the Steering Committee
- This will bring on board members of the hardware and software community who have been making contributions to oneAPI

Steering Committee Responsibilities

- Agree and track annual goals
- Vote on proposals for creation of Working Groups
- Vote on proposals for creation of SIGs
- Ratify new versions of the specification
- Approve the Marketing plans

Rod Burns (Codeplay) – Chairperson

Members

Penporn Koanantakool (Google)
Kevin Harms (Argonne National Lab)
Antonio Pena (BSC)
Robert Cohn (Intel)

Others TBA

Technical Advisory Boards

- Highly engaged
- Great technical discussion
- Providing vital feedback on spec
- Constructive discussions on definitions of the spec

oneMKL

Level Zero

AI

oneIPL

Language

Governance Goals

- Open membership for groups
- Members vote on specification changes
- Members can propose new groups
- Feedback loop for implementations

These require broader scope and formal specification work alongside more general implementation discussion

New Technical Group Structure

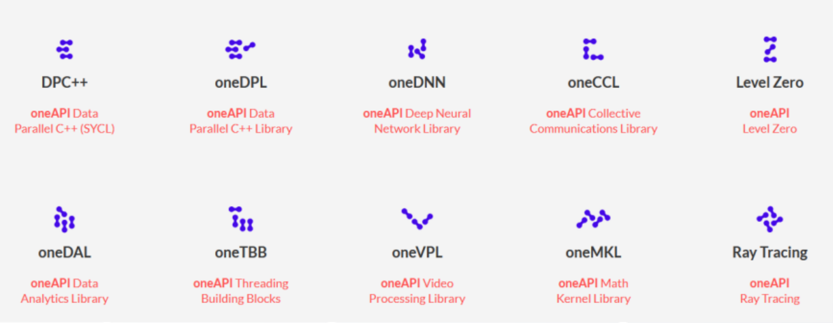
Special Interest Groups

- Form from existing TABs
- Facilitate technical discussions
- Can define their own scope
- May or may not feed into a Working Group

Working Groups

- Deal with specification proposals
- Vote on changes to the specification
- Proposals must be fully formed and draft

oneAPI Feedback Loop



DPC++
oneAPI Data Parallel C++ (SYCL)

oneDPL
oneAPI Data Parallel C++ Library

oneDNN
oneAPI Deep Neural Network Library

oneCCL
oneAPI Collective Communications Library

Level Zero
oneAPI Level Zero

oneDAL
oneAPI Data Analytics Library

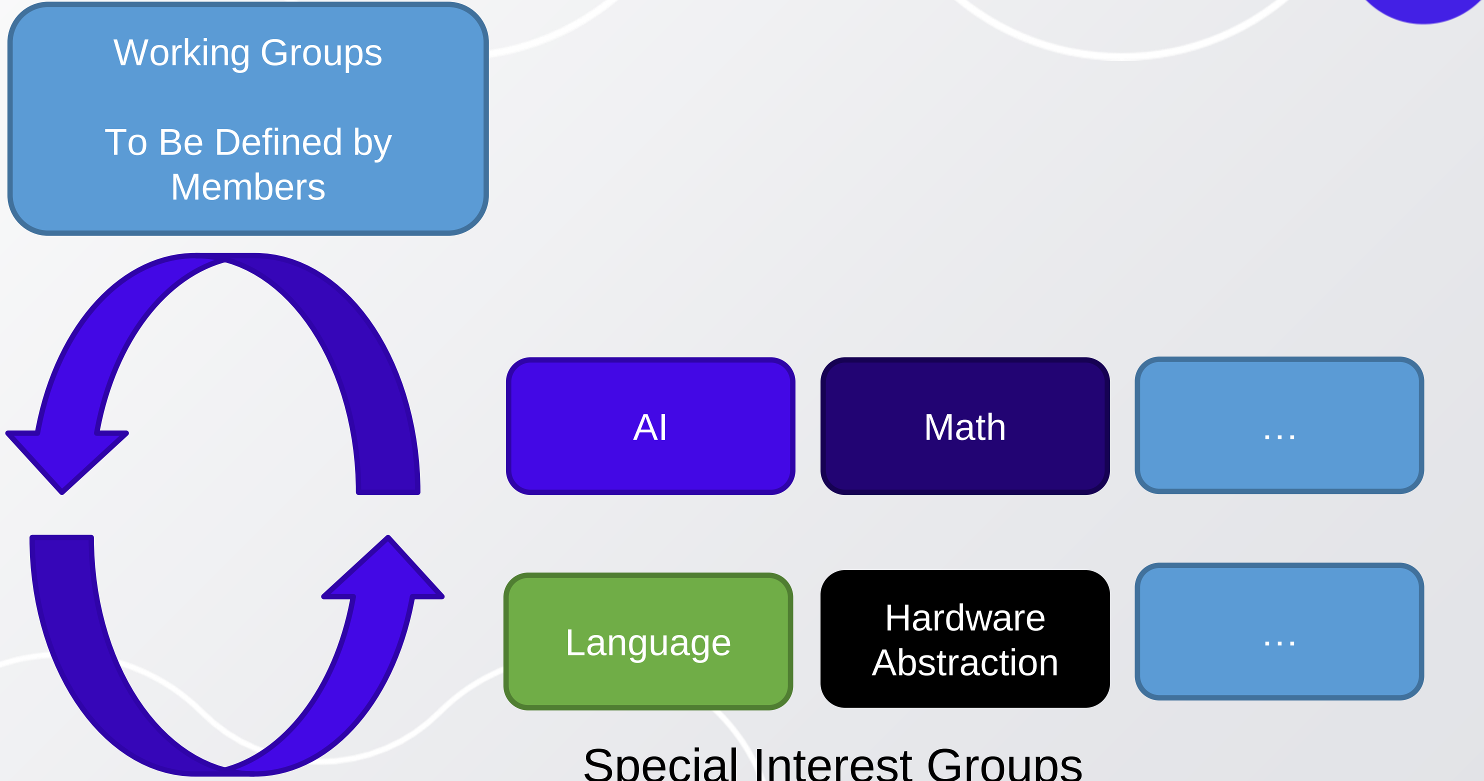
oneTBB
oneAPI Threading Building Blocks

oneVPL
oneAPI Video Processing Library

oneMKL
oneAPI Math Kernel Library

Ray Tracing
oneAPI Ray Tracing

Specification



Working Groups

To Be Defined by Members

AI

Math

…

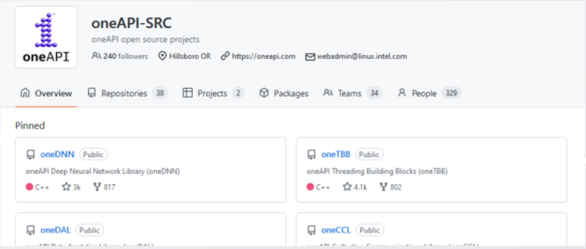
Language

Hardware Abstraction

…

Special Interest Groups

Implementations



oneAPI-SRC

oneAPI open source projects

240 followers · Hillsboro OR · https://oneapi.com · webadmin@linux.intel.com

Pinned

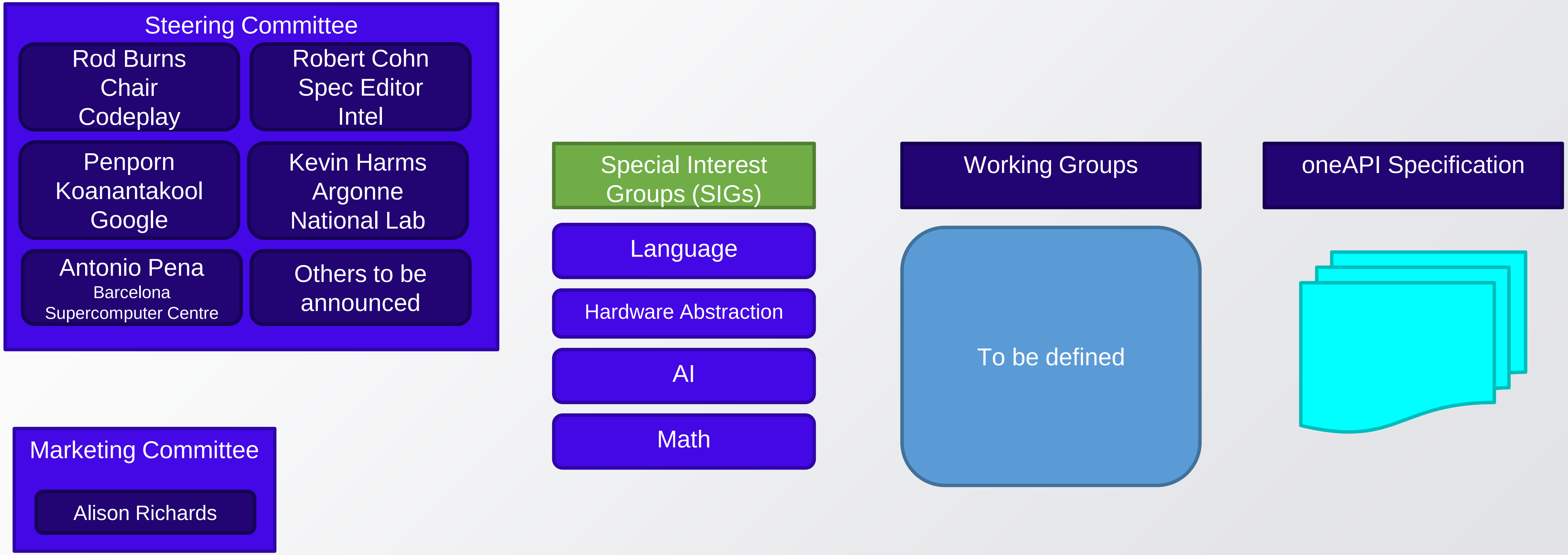
oneDNN Public
oneAPI Deep Neural Network Library (oneDNN)
C++ 3k 817

oneTBB Public
oneAPI Threading Building Blocks (oneTBB)
C++ 4.1k 802

oneDAL Public

oneCCL Public

oneAPI Community Forum Organization



Steering Committee

- Rod Burns
Chair
Codeplay
- Robert Cohn
Spec Editor
Intel
- Penporn Koanantakool
Google
- Kevin Harms
Argonne National Lab
- Antonio Pena
Barcelona Supercomputer Centre
- Others to be announced

Marketing Committee

- Alison Richards

Special Interest Groups (SIGs)

- Language
- Hardware Abstraction
- AI
- Math

Working Groups

- To be defined

oneAPI Specification

Next Steps

- TABs will become SIGs
- oneMKL TAB becomes **Math**
- Level Zero TAB becomes **Hardware Abstraction**
- Scope of SIGs will be agreed

- The GitHub project has been updated with governance and other information
  - https://github.com/oneapi-src/oneAPI-tab
- Steering Committee will meet to set goals and agree new group processes

AI	oneAPI Future
Language	Hardware Abstraction
You	Decide…

Contribute to the oneAPI Forum

- Join and lead SIGs and Working Groups
- Submit proposals for features and changes
- Vote on proposals

https://www.oneapi.io/community

Talk to me about the changes and give me your feedback rod@codeplay.com

Breakout Rooms

oneAPI Future

Rod Burns (Codeplay)

AI

Penporn Koanantakool (Google)

Language

Robert Cohn (Intel)

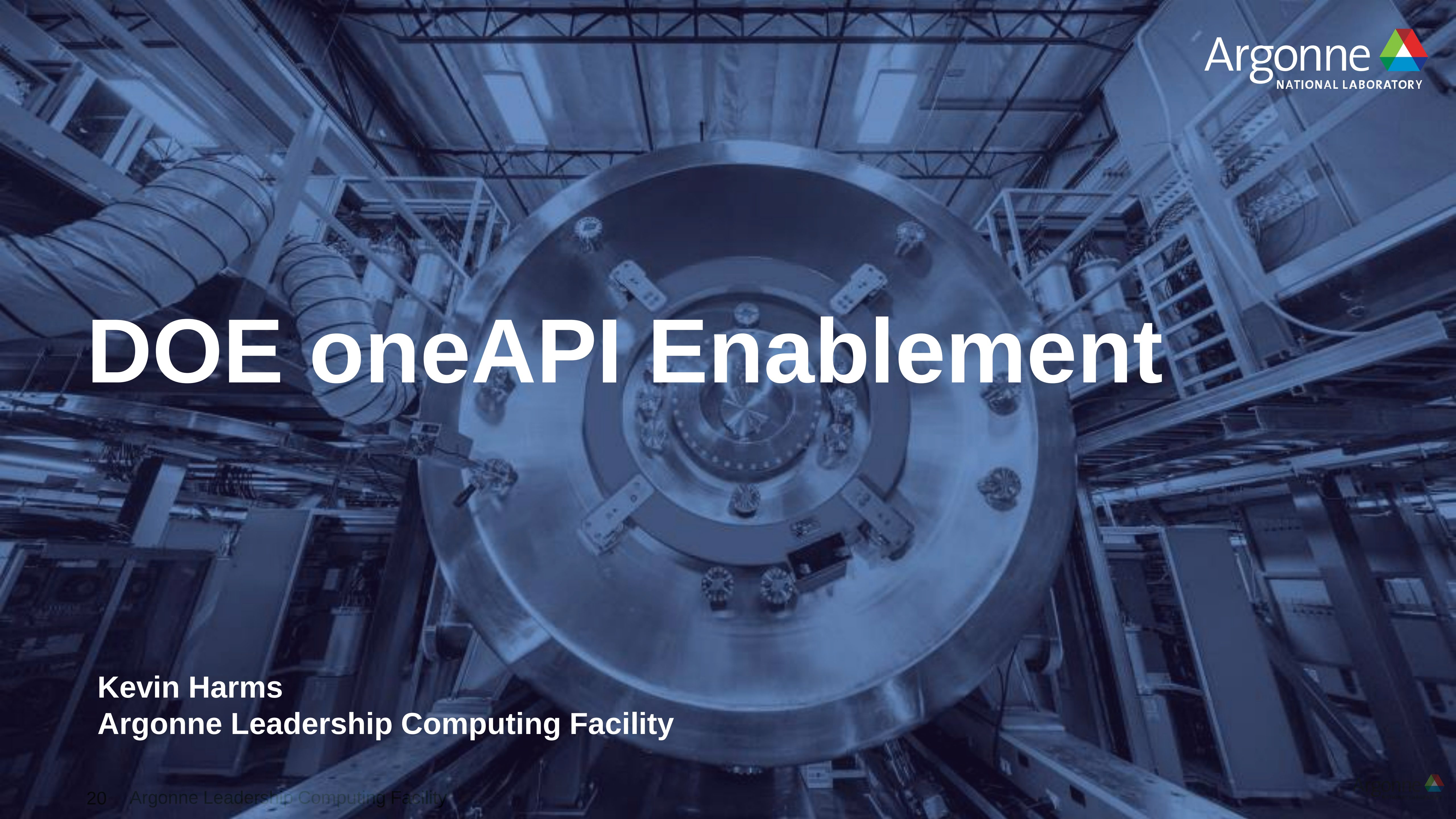
Hardware Abstraction

Kevin Harms (ANL)

oneAPI Future

- What areas are you most interested in?
  - **Introduce yourself and your interest in oneAPI**
- What is your perception of “oneAPI”?
- How “open” do you think oneAPI is?
- What can bring wide adoption to oneAPI

Stories from 2022



DOE oneAPI Enablement

Kevin Harms
Argonne Leadership Computing Facility

oneAPI Enablement – Aurora

- Support ALCF’s Aurora supercomputer
- Collaboration between ALCF and Intel
  - Funding via Non-Recurring Engineering (NRE) to support HPC focused improvements
- DPC++
  - Support for Intel PVC GPU
- Level Zero
  - Reviews and comments to the specification
  - Identificaton for multiple GPU nodes
- oneMKL
  - Batched interfaces
- oneDNN
  - Support for TF and pyTorch + optimizatoins
- oneDAL
  - Prioritize list of algorithms to be ported/optimized to GPU

Compute Node
2 Xeon Intel® Xeon® CPU Max processors
6 Intel® Data Center GPU Max
Node Unified Memory Architecture
8 fabric endpoints

GPU Architecture
Intel XeHPC architecture
High Bandwidth Memory Stacks

Node Performance
>130 TF

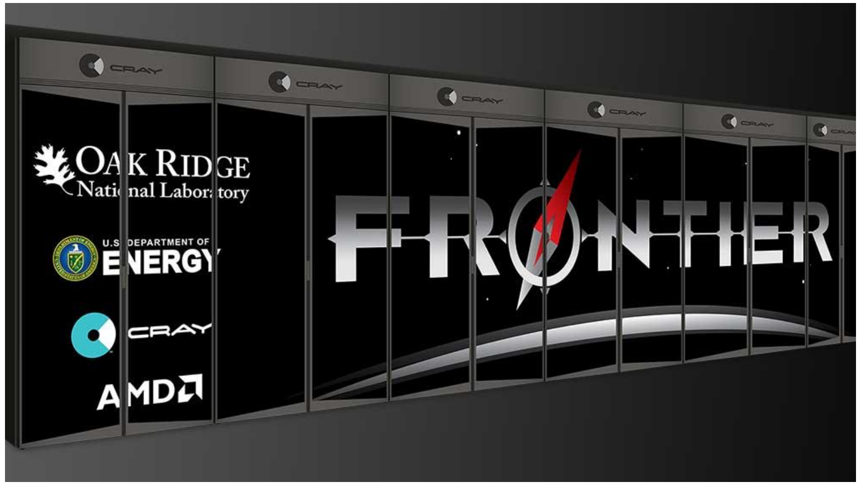
System Size
>9,000 nodes

oneAPI Enablement – Perlmutter / Polaris

- Support for NERSC’s Perlmutter supercomputer
- Collaboration between NERSC, ALCF and Codeplay
  - Focus on support of Nvidia A100 and SM_80 architecture
- Tensor Core support
- Atomics support
- Support for std::complex
  - https://github.com/argonne-lcf/SyclCPLX
- Interoperability of SYCL and OpenMP
- Support for multi-device contexts and peer-to-peer copy
- All work done against intel-llvm (Data Parallel C++)

oneAPI Enablement – Frontier

- Initial support for OLCF’s Frontier system
- Collaboration between OLCF, ALCF and Codeplay
  - Focus on initial implementation supporting AMD MI-50/MI-100 GPUs
  - Frontier uses AMD MI-250x
- Port SYCL CUDA backend to new HIP backend (PI_HIP)
- Support for four specific benchmarks
  - LULESH
  - BabelStream
  - SYCLDslash
  - RSbench
- Approximately 98% of performance for HIP version
- Performance comparisons across Nvidia and AMD hardware



Acknowledgements

This research used resources of the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357.

oneAPI TAB/ annual community forum meeting

Kenaro Kawakami

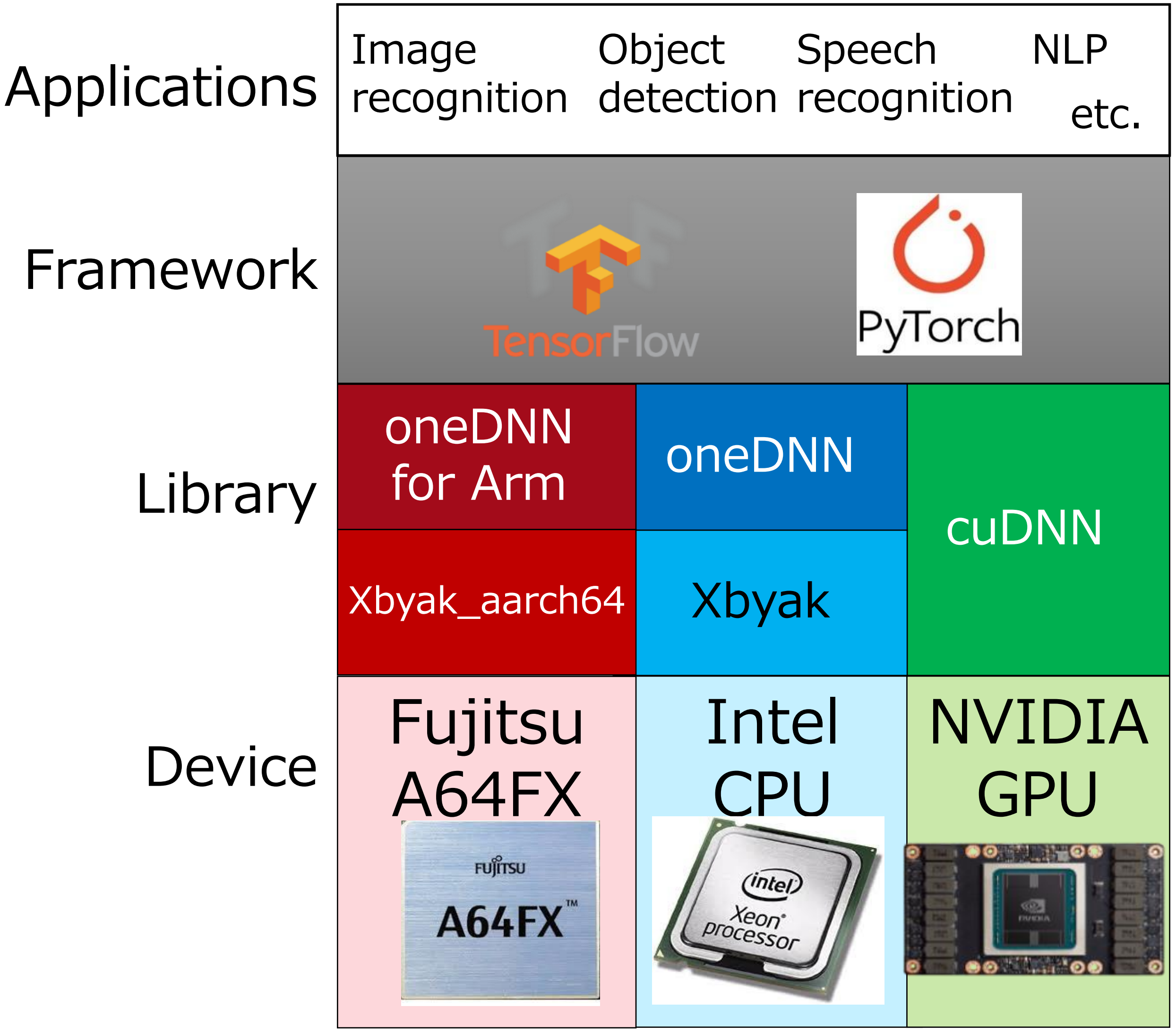
- kawakami.k@fujitsu.com
- https://github.com/kawakami-k

Computing Laboratory,

Fujitsu Research, Fujitsu Ltd.

oneDNN for Arm

- oneDNN has established an essential position in DL S/W stack.
  - The oneDNN allows users to run DL applications without being aware of device differences.
  - Implementations for Arm have already been merged into oneDNN and freely used by everyone.

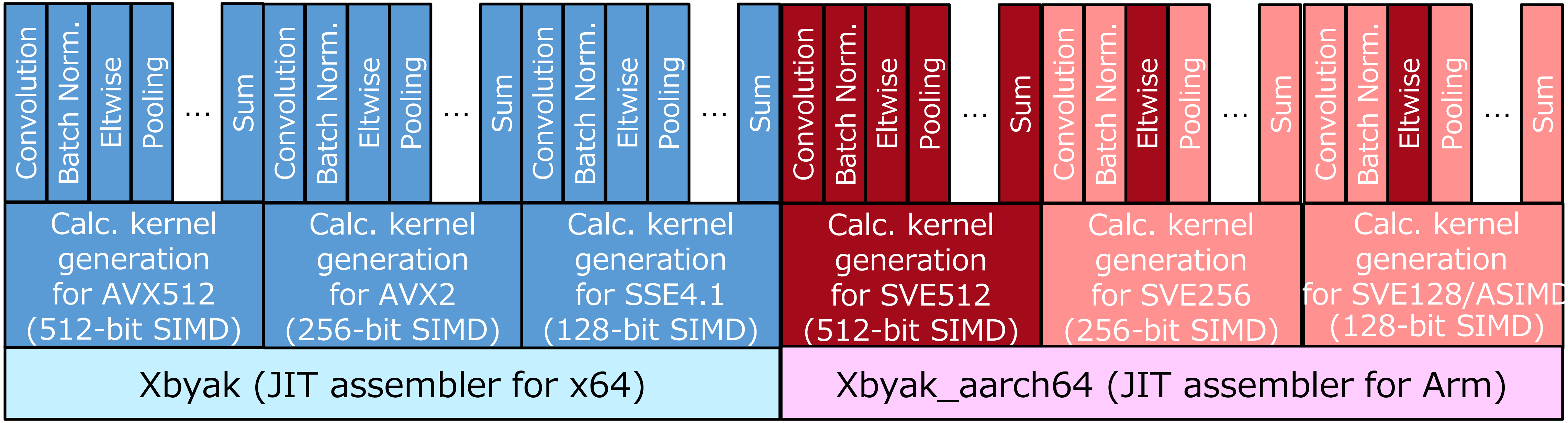


Arm v8.2a+SVE compliant

Recent activity of oneDNN for Arm

- Fujitsu developers, including myself, has been working on SVE 512(512-bit SIMD) of Armv8/9 instruction set support for oneDNN.
  - JIT-based implementation with Xbyak_aarch64
    - This allows efficient use of CPU resources to achieve high performance, just like the implementation for x64 with Xbyak.
  - The computational kernel required for CNN-based DL is almost ready.
    - Kernel type: convolution, batch-norm, relu and its variant, pooling, sum, reorder.
    - Data type: fp32, int8, uint8, int32.
  - Support for v3.0 has also been completed.
    - Release note https://github.com/oneapi-src/oneDNN/releases
    - API changes for v3.0 https://github.com/oneapi-src/oneDNN/blob/rfcs/rfcs/20220815-v3.0-API-cleanup/README.md

Kernel support status



Convolution | Batch Norm. | Eltwise | Pooling | … | Sum

Calc. kernel generation for AVX512 (512-bit SIMD)

Convolution | Batch Norm. | Eltwise | Pooling | … | Sum

Calc. kernel generation for AVX2 (256-bit SIMD)

Convolution | Batch Norm. | Eltwise | Pooling | … | Sum

Calc. kernel generation for SSE4.1 (128-bit SIMD)

Convolution | Batch Norm. | Eltwise | Pooling | … | Sum

Calc. kernel generation for SVE512 (512-bit SIMD)

Convolution | Batch Norm. | Eltwise | Pooling | … | Sum

Calc. kernel generation for SVE256 (256-bit SIMD)

Convolution | Batch Norm. | Eltwise | Pooling | … | Sum

Calc. kernel generation for SVE128/ASIMD (128-bit SIMD)

Xbyak (JIT assembler for x64)

Xbyak_aarch64 (JIT assembler for Arm)

AWS Graviton3

https://www.apple.com/jp/newsroom/2022/06/apple-unveils-m2-with-breakthrough-performance-and-capabilities/

Fully supported

Fully supported †

Partially supported ‡

Let's work together to extend the implementation for Arm devices!

† Data type of fp32/int8/uint8/int32 are supported.
‡ The figure shows the support status of JIT-based implementations.
oneDNN also includes Arm-compute-library-based implementation for Arm devices.

Thank you

codeplay®

THE HETEROGENEOUS SYSTEMS EXPERTS

Codeplay’s Journey to a Common Hardware Interface

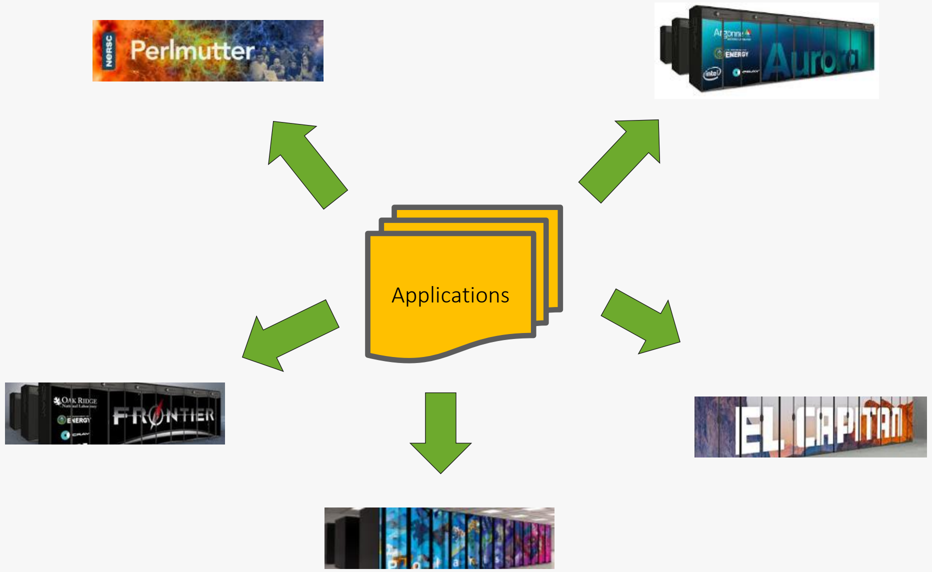
@codeplaysoft

/codeplaysoft

codeplay.com

Gordon Brown, Principal Product Owner, oneAPI

Extend oneAPI to Support Nvidia & AMD

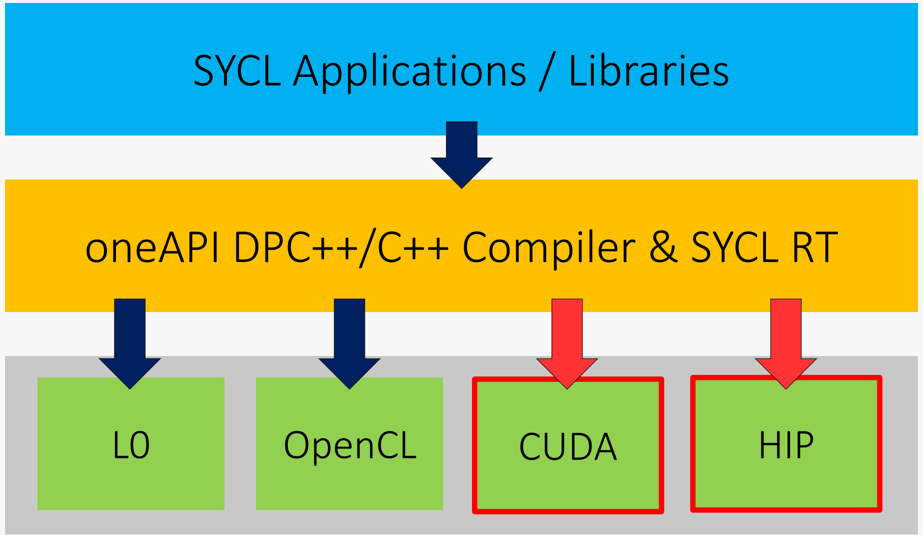


- Performance portability & long term stability in programming models are more important than ever
- Codeplay have been working with Intel and the US DoE to extend oneAPI to support Nvidia & AMD GPUs

Current Status

oneAPI Component	Nvidia GPUs	AMD GPUs
DPC++/C++ Compiler	~95% supported	~50% supported
oneDNN	Yes	In progress
oneMKL	Yes	In progress

First binary release of oneAPI for Nvidia/AMD GPUs coming soon!



SYCL Applications / Libraries
oneAPI DPC++/C++ Compiler & SYCL RT
L0
OpenCL
CUDA
HIP

- DPC++ has full support for Nvidia GPUs and partial support AMD GPUs
- oneMKL & oneDNN supports Nvidia GPUs and support for AMD GPUs is in progress
- Extensions have been introduced to support CUDA capabilities: tensor cores, cooperative groups, extended atomics, etc

What’s to Come in 2023

- Proposition of DPC++ extensions to SYCL Next
- Continued alignment with SYCL and ISO C++
- Continued maintenance of the DPC++ CUDA and HIP backends
- Further support for the DPC++ HIP backend
- Further performance optimizations for the CUDA and HIP backends
- Support for additional oneAPI libraries such as oneDPL and oneCCL

Presentation from Andrew Richards, Codeplay CEO

codeplay®

Enabling AI & HPC to be open, safe and accessible to all

oneAPI: Building the Future

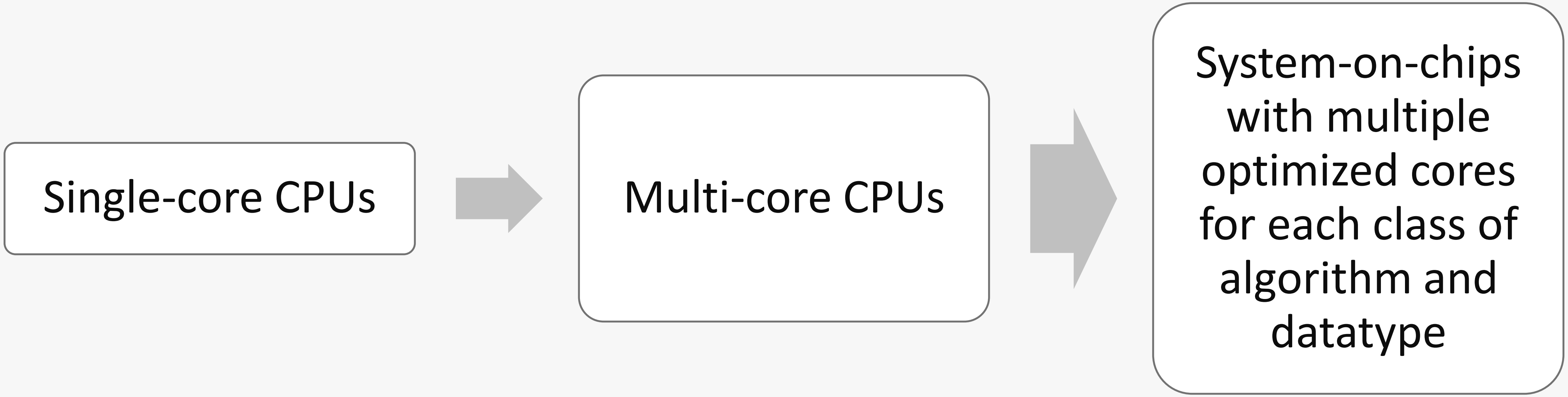
December 22

Andrew Richards

codeplay®

Enabling AI & HPC to be open, safe and accessible to all

Our Brave New World



Single-core CPUs

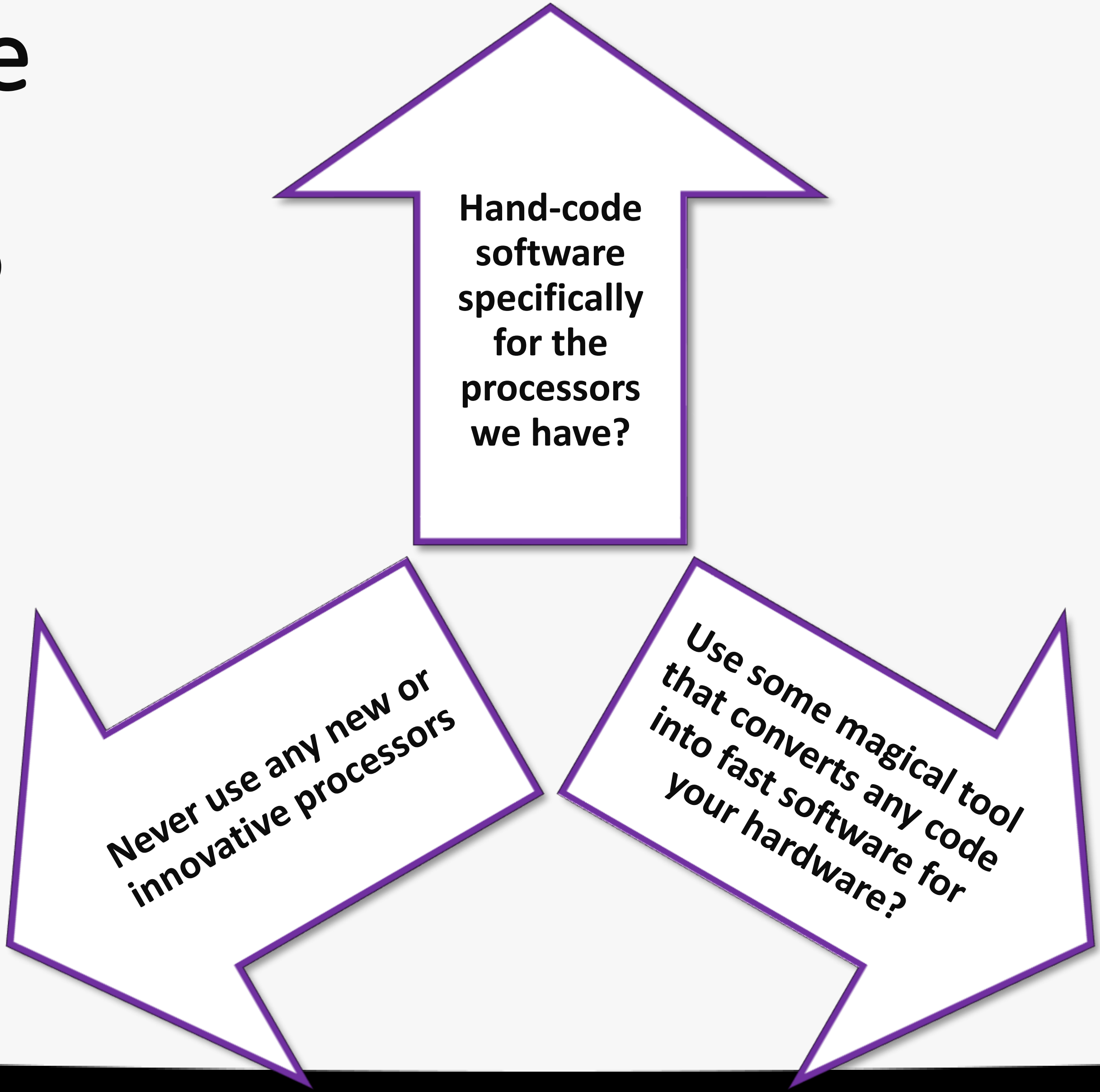
Multi-core CPUs

System-on-chips with multiple optimized cores for each class of algorithm and datatype

Great for processor architects, but how do we write the software?

How do we write fast software?

(Your hardware will be obsolete by the time you have optimized it)



Hand-code software specifically for the processors we have?

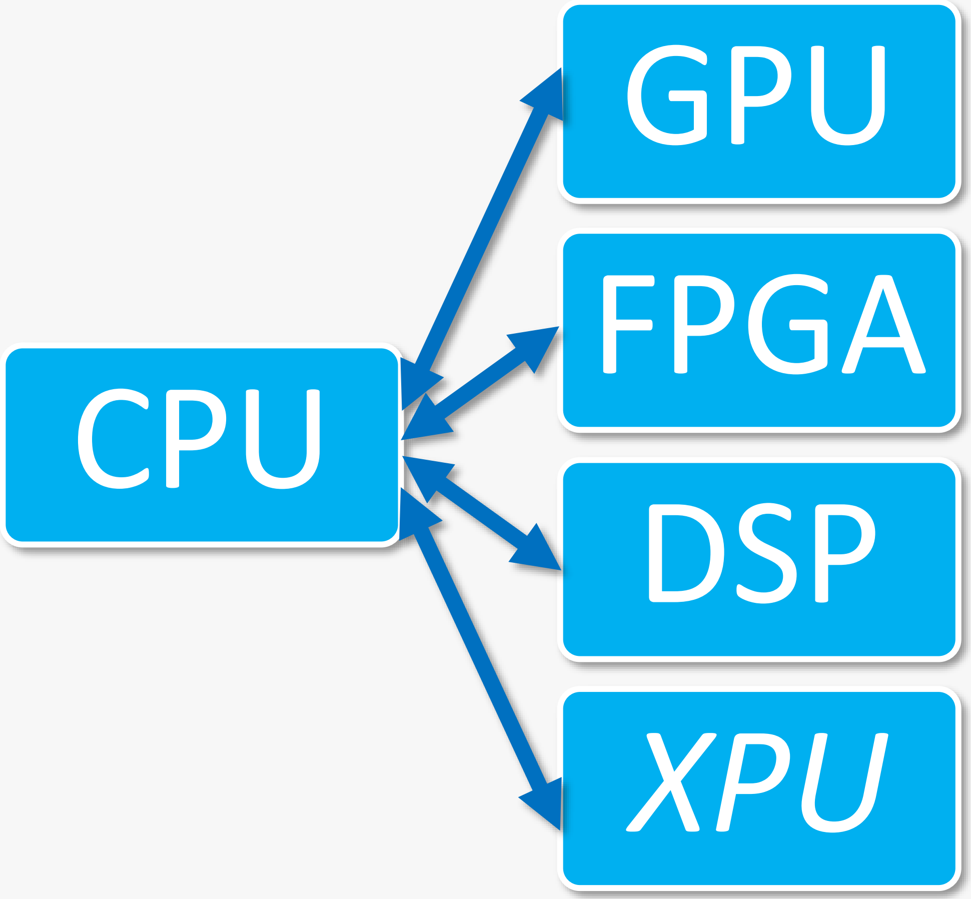
Never use any new or innovative processors

Use some magical tool that converts any code into fast software for your hardware?

(Those days are over)

(Only works if your magical tool has been pre-programmed to understand the software you just invented)

Our high performance future



CPU

GPU

FPGA

DSP

XPU

Moving software from the CPU to the whole system

What are we doing about it?

oneAPI

1. **Everything** you need to build software from CPUs to XPUs
2. **Open**: mix of open-source, open-standard, open-governance
3. **High performance**: optimized libraries for different processors

Components of oneAPI

DPC++

- SYCL compiler
- For C++ programmers who want performance across CPU, GPU, FPGA, *XPUs*

oneDPL

- ISO C++ Standard Parallelism library
- For C++ programmers who want performance easily

oneDNN

- AI graph compiler
- For people who want high performance deep learning

oneDAL

- Data analytics library
- For C++ programmers doing data analytics

oneTBB

- CPU-only parallelism library
- For C++ developers who want high performance CPU code

oneCCL

- Library for distributed processing across multiple hardware devices
- For big systems

SPIR-V

- Virtual instruction set for accelerators
- Enables different compilers & languages

Level Zero

- Low-level hardware interface
- Enables more languages to be accelerated with SPIR-V

oneVPL

- Accelerated video codec
- For people processing video: encodes, decodes & processes video

oneMKL

- Optimized math library
- Linear algebra (sparse+ dense), FFT/DFT, random numbers, LAPACK

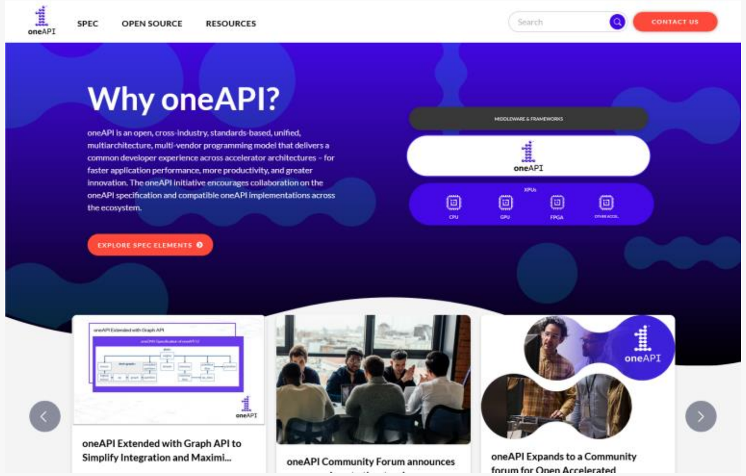
Ray Tracing

- Accelerated ray tracing
- For graphics programmers

Your Idea Here?

- It's an open project, so you can add your own concepts

The oneAPI ecosystem


oneAPI SPEC OPEN SOURCE RESOURCES Search CONTACT US

Why oneAPI?

oneAPI is an open, cross-industry, standards-based, unified, multiarchitecture, multi-vendor programming model that delivers a common developer experience across accelerator architectures – for faster application performance, more productivity, and greater innovation. The oneAPI initiative encourages collaboration on the oneAPI specification and compatible oneAPI implementations across the ecosystem.

EXPLORE SPEC ELEMENTS

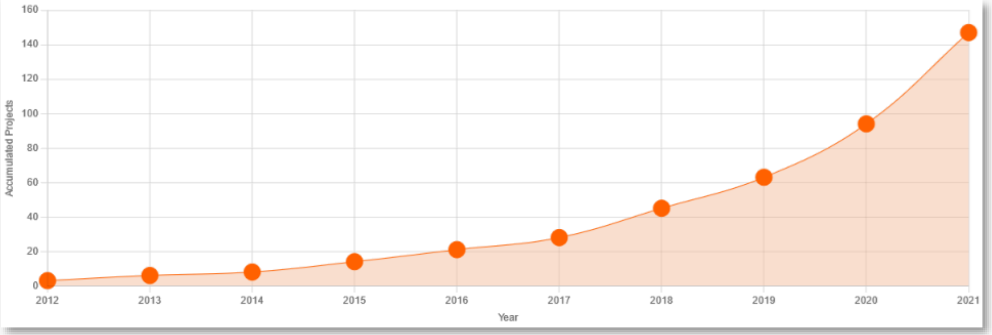
oneAPI Extended with Graph API to Simplify Integration and Maximi...

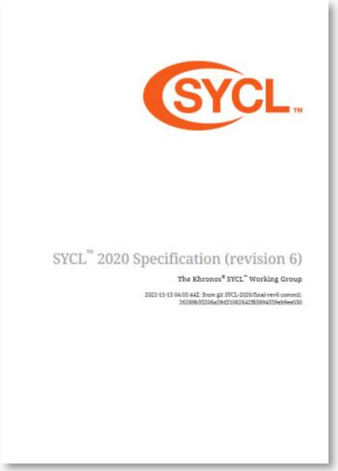
oneAPI Community Forum announces

oneAPI Expands to a Community forum for Open Accelerated

oneapi.io
Holds all the specifications and tracks everything going on in the ecosystem

Rapidly-growing range of projects using SYCL and oneAPI

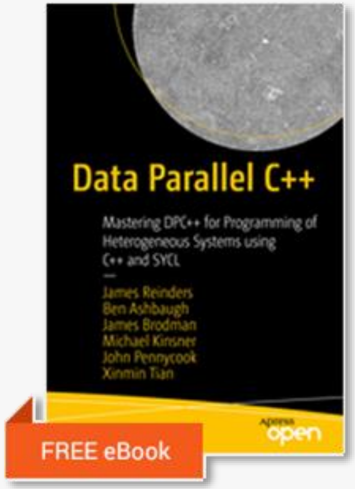



SYCL™

SYCL™ 2020 Specification (revision 6)

The Khronos® SYCL™ Working Group

2022-11-13 04:05:44Z: from git SYCL-2020/final-rev6 commit: 26269b35206a59d51062642f83894359eb9ee030


Data Parallel C++

Mastering DPC++ for Programming of Heterogeneous Systems using C++ and SYCL

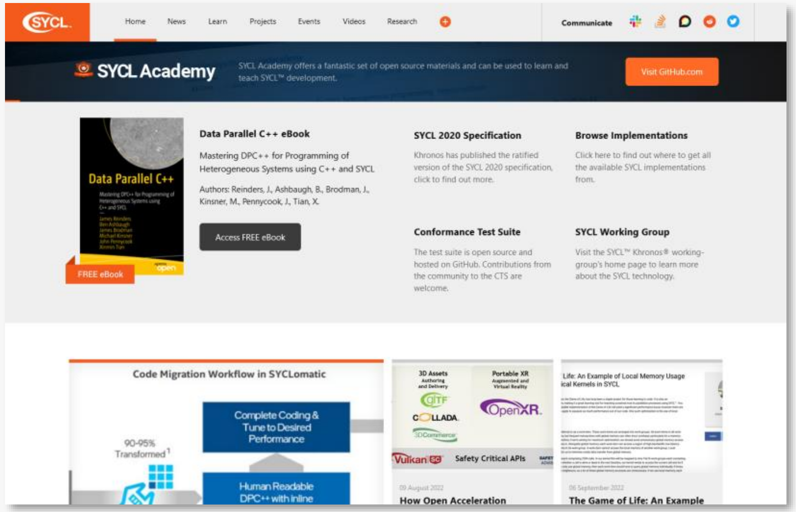
James Reinders
Ben Ashbaugh
James Brodman
Michael Kinsner
John Pennycook
Xinmin Tian

Apress open

FREE eBook

The SYCL specification and Data Parallel C++ book document the programming model of oneAPI

sycl.tech tracks everything going on in the world of SYCL



SYCL Academy

SYCL Academy offers a fantastic set of open source materials and can be used to learn and teach SYCL™ development.

Visit GitHub.com

Data Parallel C++ eBook

Mastering DPC++ for Programming of Heterogeneous Systems using C++ and SYCL

Authors: Reinders, J., Ashbaugh, B., Brodman, J., Kinsner, M., Pennycook, J., Tian, X.

Access FREE eBook

SYCL 2020 Specification

Khronos has published the ratified version of the SYCL 2020 specification, click to find out more.

Browse Implementations

Click here to find out where to get all the available SYCL implementations from.

Conformance Test Suite

The test suite is open source and hosted on GitHub. Contributions from the community to the CTS are welcome.

SYCL Working Group

Visit the SYCL™ Khronos® working-group's home page to learn more about the SYCL technology.

Code Migration Workflow in SYCLomatic

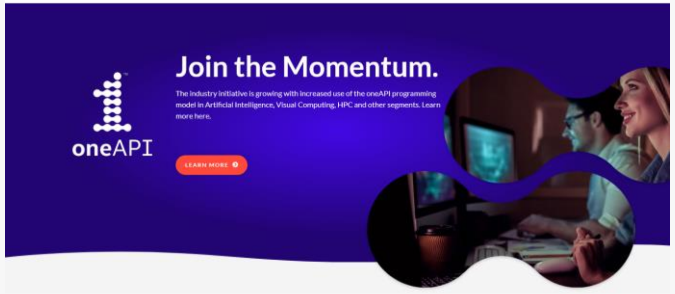
90-95% Transformed 1
Complete Coding & Tune to Desired Performance
Human Readable DPC++ with inline
3D Assets Authoring and Delivery
Portable XR Augmented and Virtual Reality
glTF
COLLADA
3DCommerce
OpenXR
Vulkan SC
Safety Critical APIs

09 August 2022
How Open Acceleration

Life: An Example of Local Memory Usage ical Kernels in SYCL

06 September 2022
The Game of Life: An Example

The future



Join the Momentum.

The industry initiative is growing with increased use of the oneAPI programming model in Artificial Intelligence, Visual Computing, HPC and other segments. Learn more here.

- Come and join us!
  - You can join both the oneAPI Community Forum and add your input to drive the project forwards

- Build out the performance-portable software ecosystem
  - There’s a huge opportunity to build the performance-portable frameworks of the future using oneAPI & SYCL, including ISO C++

- Bring your own hardware to developers
  - You can design your own chip for oneAPI, so you can accelerate all this software with your own hardware innovations

What interests me

Safety

- We're increasingly seeing AI being used to do things requiring safety such as driving cars or operating medical systems
- There's huge benefits to using AI in these ways
- But: there are huge dangers
- We're working on solving some of these challenges

Performance portability

- In 2023, we'll have a range of hardware supporting oneAPI and SYCL
- SYCL doesn't magically give you performance everywhere
- But: you can build performance frameworks using SYCL
- Let's build performance portable frameworks in 2023

Software-first-hardware

- When we build software with oneAPI we can target a variety of hardware
- It's very hard to design processors for very complex software
- Instead of designing software-for-hardware, let's design hardware-for-software
- We're particularly working with RISC-V to enable this

codeplay®

Enabling AI & HPC to be open, safe and accessible to all

What Happens Next?

- SIG meetings for 2023 will be scheduled in January
- We invite your proposals for new SIGs during 2023
- Processes and mechanisms will be updated
- The Steering Group will meet to set the goals for 2023

Tell us what you want to see in 2023 for the oneAPI Community Forum
oneapi@codeplay.com rod@codeplay.com