Reflect, Rejoice, Envision

Navigating this year's oneAPI Journey

Sanjiv Shah
Vice President, Intel Software & Advanced Technology Group
General Manager, Developer Software Engineering

Joe Curley
Vice President, Intel Software & Advanced Technology Group
General Manager, Software Products & Ecosystem

oneAPI

intel

Welcome!

Our computing world
over the past 4+ decades:

run your code anywhere.

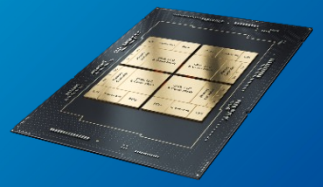
Why should the world of accelerators be any different?

Developer Challenges

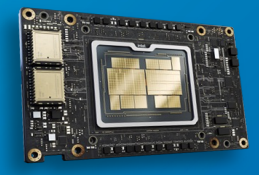
Unprecedented capabilities offered by accelerators

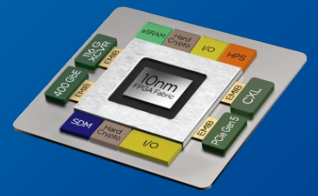
Developer Dilemma: Restrict where to run or bear the porting expense

CPU



Accelerators





oneAPI

We want to make accelerator programming portable and open

intel

oneAPI

Middleware and Frameworks

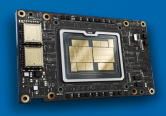
PyTorch · TensorFlow · scikit-learn · OpenVINO · NumPy · XGBoost

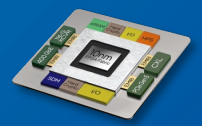
oneAPI Specification

Direct Programming	API-Based Programming		
SYCL (C++)	Math oneMKL	Threading oneTBB	Parallel STL oneDPL
	Analytics/ ML oneDAL	DNN oneDNN	AI Comm oneCCL

Low-Level Hardware Interface (oneAPI Level Zero)


CPU


GPU


FPGA


Other Accelerators

Freedom · **Performance** · **Productivity**

intel

Visit oneapi.io for more details

oneAPI Plug-ins for Nvidia* & AMD*

Codeplay Support for Nvidia & AMD GPUs to Intel® oneAPI Base Toolkit

oneAPI for NVIDIA & AMD GPUs

- Binary plugins to Intel® oneAPI DPC++/C++ Compiler
- Quarterly updates in-sync with SYCL 2020 conformance & performance

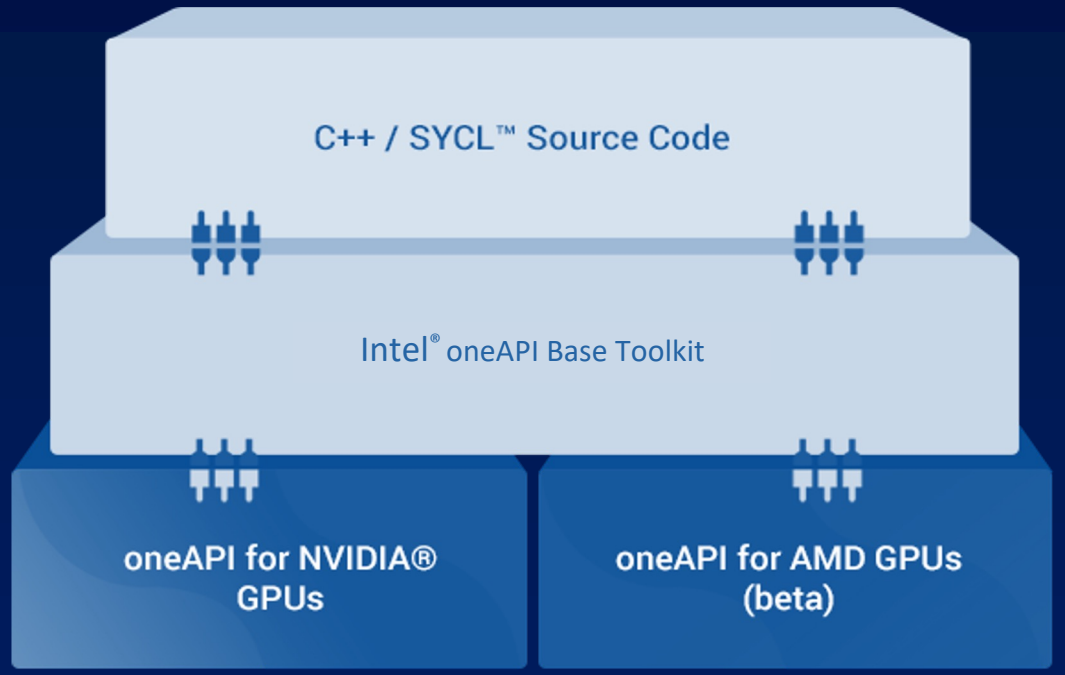


Image courtesy of Codeplay Software Ltd.

Nvidia GPU plug-in

AMD GPU plug-in

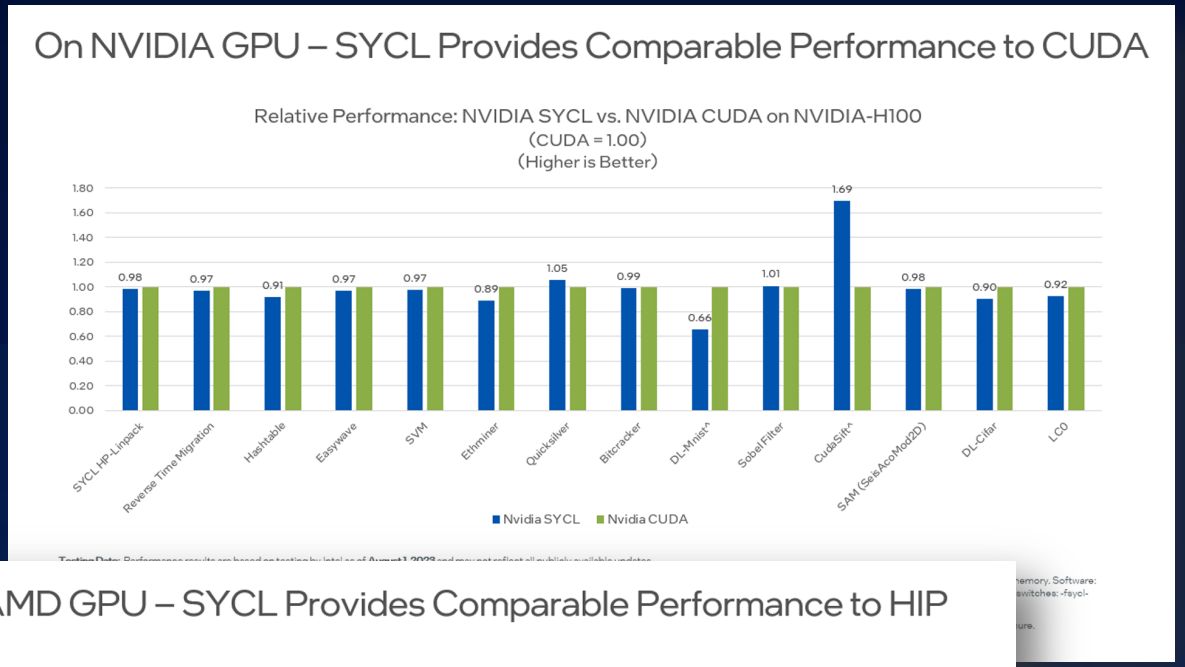
Codeplay blog

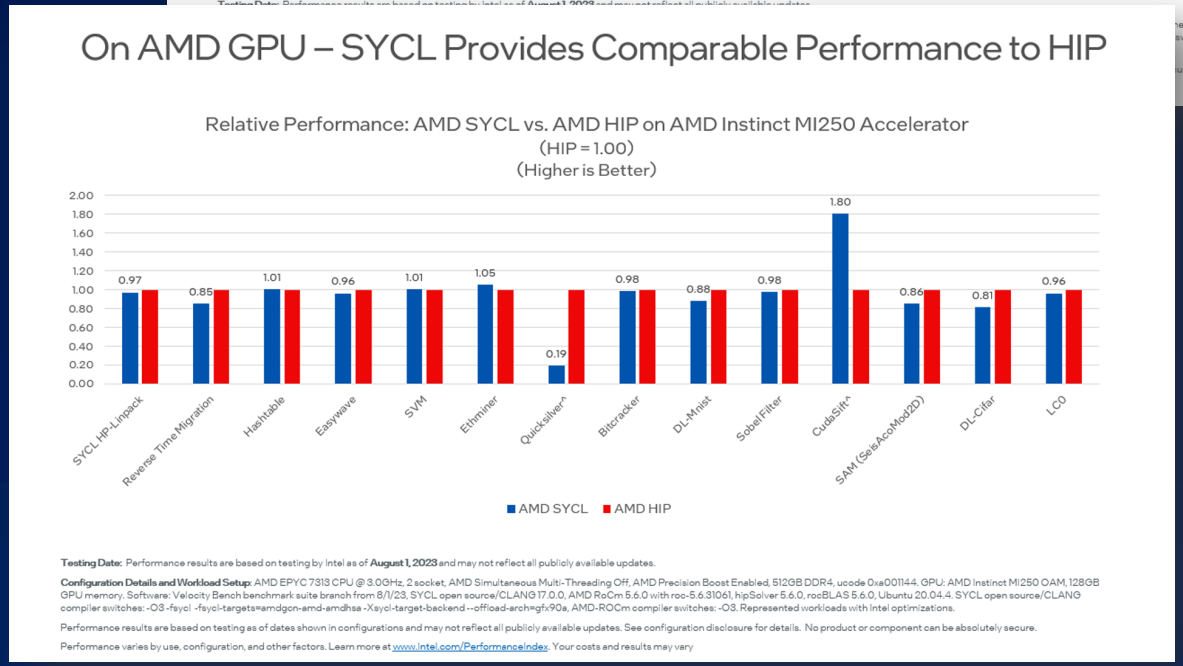
Codeplay press release

Accelerating Choice with SYCL* on NVIDIA and AMD

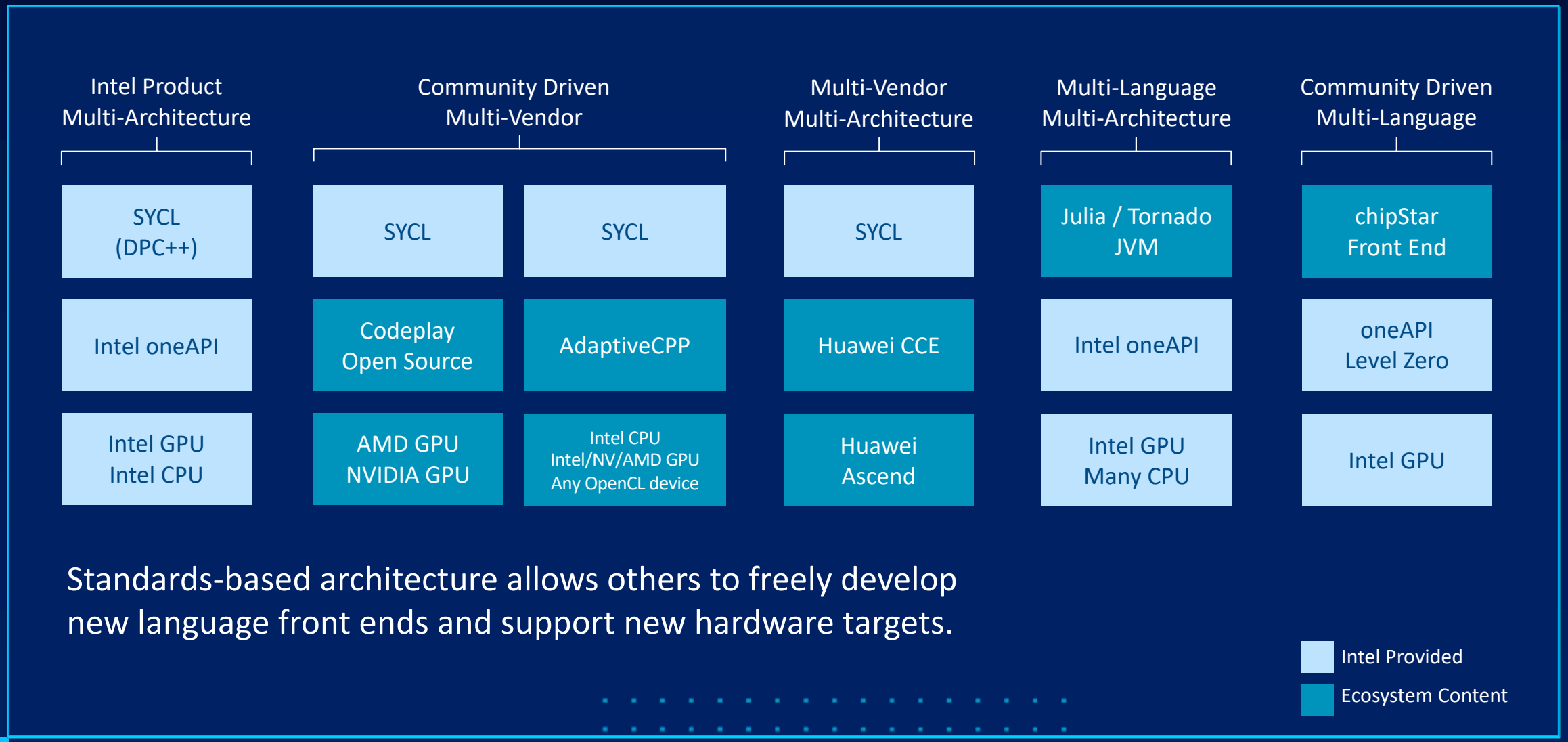
Khronos Group Standard

- Open, standards-based
- Multi-architecture performance
- Comparable performance to native:
  - CUDA on Nvidia GPUs
  - HIP on AMD GPUs
- Extension of widely used C++ language





Ecosystem Driven Innovation



Standards-based architecture allows others to freely develop new language front ends and support new hardware targets.

Open, unified standard accelerator programming model that delivers cross-platform performance and productivity

Now governed by the Linux Foundation

intel

Commitment to Open and Scalable Acceleration

Unified Acceleration (UXL) Foundation

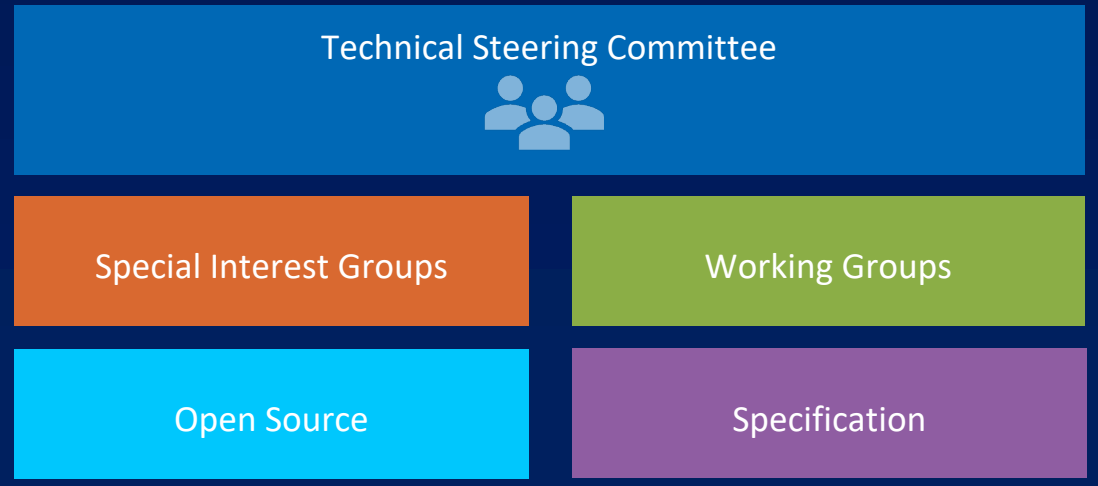
- Governance: Linux Foundation's Joint Development Foundation
- Mission: Unify the heterogeneous compute ecosystem around open standards
- Starting point: oneAPI Specification (oneAPI.io)
- Goal: broad-based industry participation and contributions
- SIGs: AI, Hardware, Language, Math
- Join Us: Participate in SIGs
  - www.UXLFoundation.org

UXL Unified Acceleration Foundation

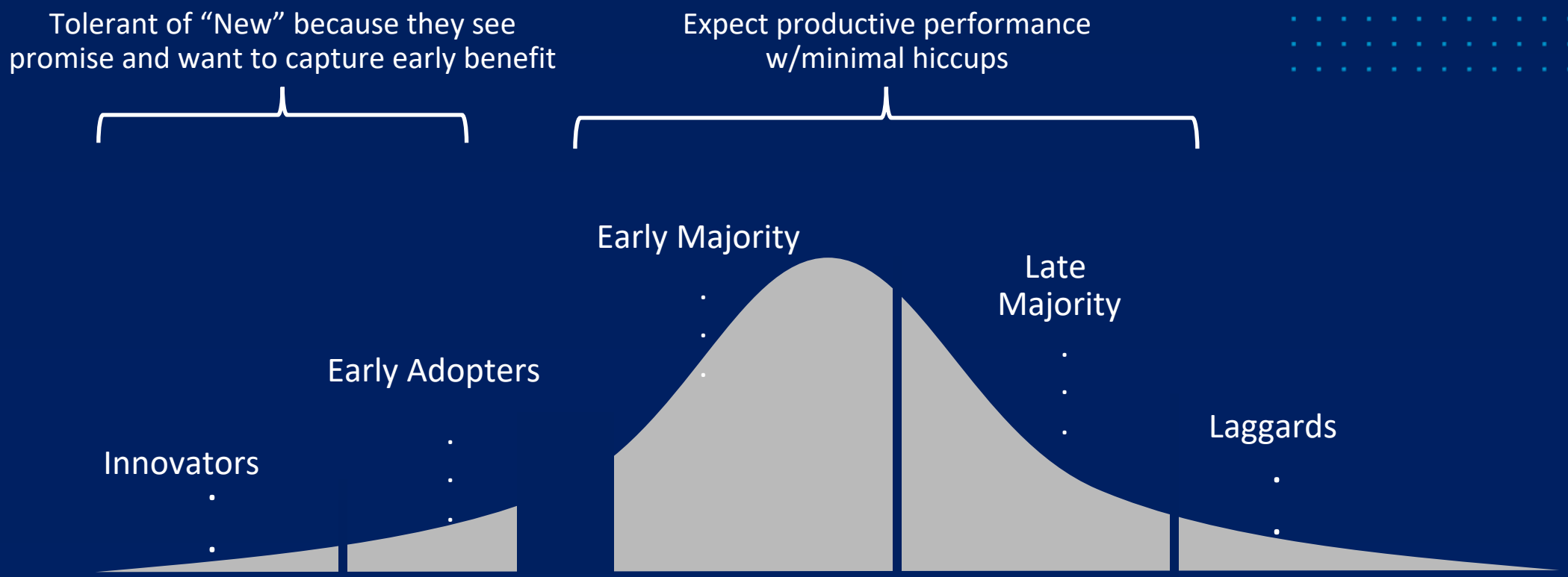
Steering Members

arm FUJITSU Google Cloud Imagination intel Qualcomm SAMSUNG vmware

UXL Foundation Structure



New Technology Adoption



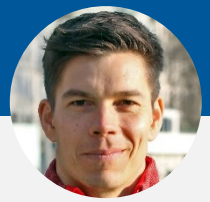
Crossing the Chasm: Marketing and Selling Disruptive Products to Mainstream Customers, By Geoffrey A Moore, 1991
Harper Business Essentials ISBN 0-060051712-3

oneAPI Centers of Excellence

Super Computing ‘23 Success Stories from oneAPI Innovators

Dr. Mohammad Zubair, ODU and Dr. Eric Nelson, NASA

“Optimization of Ported CFD Kernels on Intel Data Center GPU Max 1550 Using oneAPI ESIMD”



Dr. Hartwig Anzt and team, UTK

"Hands-On HPC Application Development Using C++ and SYCL" and “Porting Batched Iterative Solvers onto Intel GPUs with SYCL"

Dr. Joseph Insley, NIU

“Argonne and Intel Advancing Scientific Research and Visualization at Exascale —Rising to the Challenge"

Paper

Dr. Valerio Pascucci, UoU

“Scalable and Portable Blending of Massive Image Mosaics Using Intel® oneAPI Tools”

Panel Speaker

Dr. Andreas Goetz, UCSD

“Powering Amber Molecular Dynamics Simulations with oneAPI”

Success Stories in SYCL

THE UNIVERSITY OF TENNESSEE KNOXVILLE

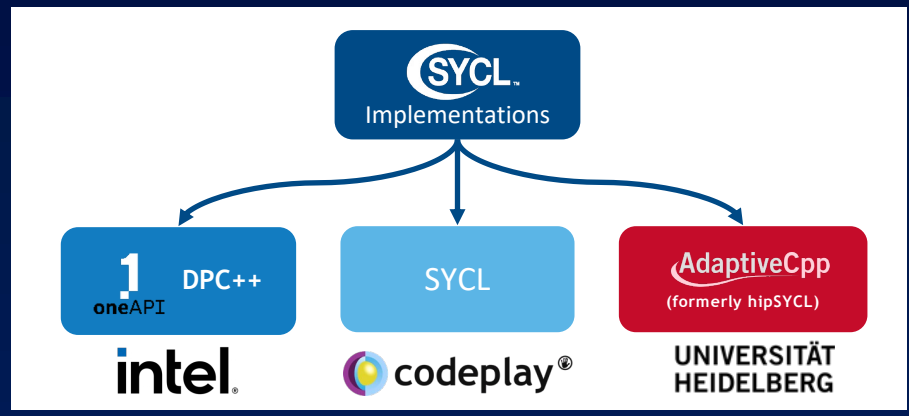
Full port of Ginkgo (popular linear algebra library) to SYCL

Arcvideo

Migrated CUDA → SYCL to create "single, portable code" with competitive perf & lower power

Argonne NATIONAL LABORATORY

#2 system in the world running oneAPI



OLD DOMINION UNIVERSITY

"SYCL version gave us portability, and we could run the code on Intel® GPUs and CPUs, and NVIDIA* GPUs"

SAMSUNG MEDISON

".. one source code for performance acceleration on different kinds of hardware"

KTH VETENSKAP OCH KONST · Stockholm University · FAST. FLEXIBLE. FREE. GROMACS

One of the most widely used HPC codes uses SYCL, calling it "a revolution for portability"

IST TÉCNICO LISBOA

"By using the Intel® DPC++ Compatibility Tool, over 90% of our hand-tuned CUDA* code was migrated"

THE UNIVERSITY OF UTAH

For the first time, deployed the Massive Image Blending capability on any cloud resource

intel

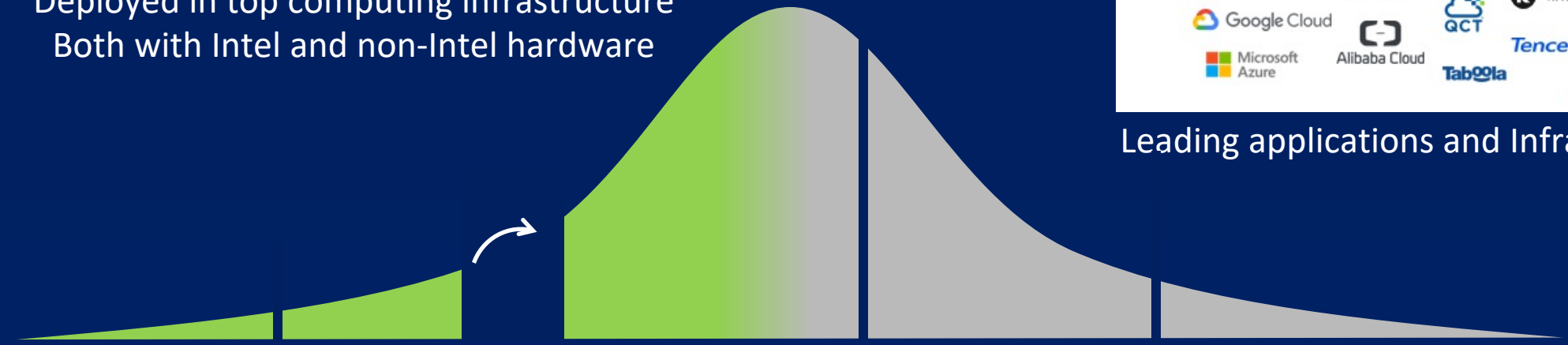
*Other names and brands may be claimed as the property of others. SYCL is a trademark of the Khronos Group Inc.
Source: https://www.intel.com/content/www/us/en/developer/tools/oneapi/ecosystem-support.html

From Concept to Deployment

Top Computing Infrastructures, leading applications

Deployed in top computing infrastructure
Both with Intel and non-Intel hardware

Leading applications and Infrastructure

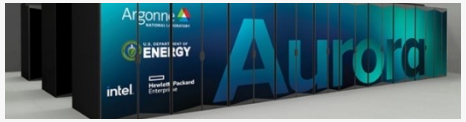


intel

Intel Software Developer Tools Use Cases

Multiarchitecture Performance & Productivity Value for Customers

HPC



Argonne rolled out Aurora performance using Intel® Data Center GPU Max Series

TACC's Frontera Supercomputer uses oneAPI to accelerate exascale scientific computing

Univ.of Cambridge strives for zettascale using oneAPI

Accelerating Google Cloud for HPC
Video | Podcast

AI/ML/DL

Red Hat optimizing Data Science Workflows

Scaling HuggingFace Transformer and Optimum Performance with Intel AI

Optimize performance of IBM Watson with NLP & NLU

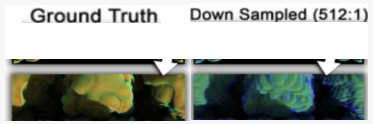
PyTorch Foundation

Advance PyTorch through Intel Optimizations

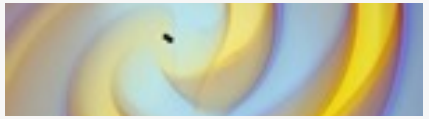
Rendering

THE UNIVERSITY OF TENNESSEE KNOXVILLE

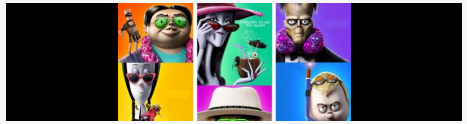
Univ. of Tennessee used oneAPI to enable a cloud-based Rendering-as-a-Service (RaaS) environment



Univ. of Calif. at Davis increased performance by 3x & delivered 100x data compression for scientific rendering[1]

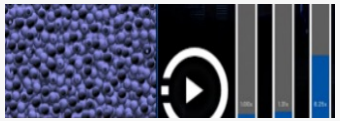


Stephen Hawking Centre for Cosmology Visualizes Cosmos Physics

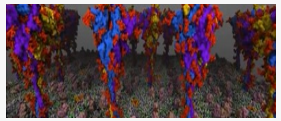


The Addams Family 2/Cinesite achieved up to 25% efficiency in rendering[1]

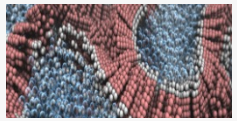
CUDA* Code Migration to SYCL*



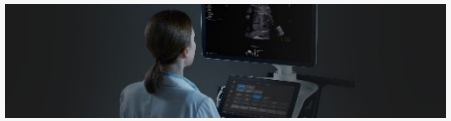
Preparing NAMD molecular dynamics for Aurora Supercomputer



LAMMPS: Speeding Exascale Material Discovery



University of Stockholm GROMACS 2022



Samsung Medison Uses oneAPI to Power Obstetric Ultrasound Systems

THE UNIVERSITY OF UTAH®

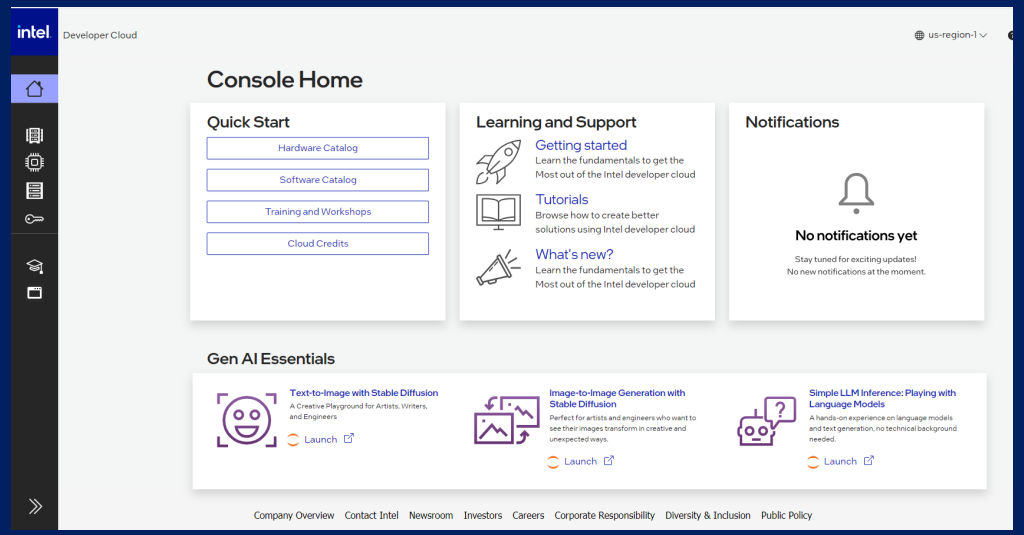
University of Utah
Using Massive Image Dataset Binding Using SYCL

intel

1. See Notices & Disclaimers for configuration details. Refer to software.intel.com/articles/optimization-notice for more information regarding performance & optimization choices in Intel software products. For workloads and configurations visit www.Intel.com/PerformanceIndex. Results may vary. Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy. *Other names and brands may be claimed as the property of others. SYCL is a trademark of the Khronos Group Inc.
2. Source for CUDA to SYCL migration: https://www.intel.com/content/www/us/en/developer/tools/oneapi/ecosystem-support.html

Try It Today – Intel Developer Cloud

For AI and Accelerated Computing



Easy To Use

Service *Use case*	Developer Services *Deploy AI workloads on Intel platforms*
Access	SSH / CLI, API
Software	Ubuntu OS Latest Intel kernel drivers Intel (optimized) AI frameworks
Runtime Environment(s)	Dedicated host – Linux OS VM k8s
Hardware	Intel Xeon Intel Data Center GPU Flex Series Intel Data Center GPU Max Series Intel Habana

Access to The Latest Hardware & Software

Intel Optimized Frameworks

PyTorch

TensorFlow

Toolkits and Programmability

Intel oneAPI

OpenVINO

ubuntu

kubernetes

Planned Q1 '24

Kernel drivers

Standards-based Developer Stack

intel

Or Download oneAPI Developer Tools

Accelerating Multiarchitecture Compute on Intel CPUs and GPUs – Available Now!

Hardware Choice

- Expand AI & HPC capabilities on Intel CPUs & GPUs with broader standards coverage including near complete SYCL* 2020 implementation
- Get faster performance on Python numeric workloads with new GPU support
- Enjoy scalable real-time rendering with expanded GPU support

Performance

- Maximize performance on upcoming 5th gen Intel® Xeon® Scalable & Intel® Core™ Ultra processors
- Accelerate AI performance, efficiency & innovation with improved deep learning framework CPU and GPU optimizations and faster Modin data tasks with new GPU support

Productivity

- Get AI frameworks & tools faster with a flexible, streamlined process for individual downloads & pre-set bundles
- Rapid elimination of memory leaks, uninitialized memory, thread data races, deadlocks, and undefined behavior on Intel CPUs with popular LLVM sanitizers

In Preview

Run C++ standard parallel algorithms (PSTL) on both CPU and GPU

Efficient scheduling to heterogenous compute resources

Launch multiple GPU operations via SYCL Graph

Native bindless image support via SYCL Graph

Learn, Develop & Contribute with the Community

For Developers

Intel® Innovators

For Professors

Educator Program

For Students

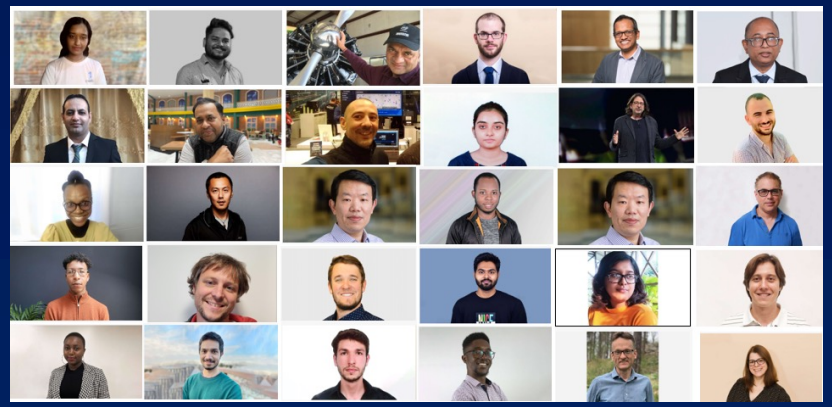
Student Ambassador

For Researchers

Centers of Excellence

For Startups

Intel Liftoff



Visit intel.com/oneapi

intel

In Closing

- Productively develop and deploy portable software
- Choose your hardware.
  Don’t let your software choose it for you.
- Make accelerated computing open and portable –
  Like things have been for the 4 prior decades
- Join us in this mission

Thank you!

intel


intel

Notices & Disclaimers – 1 of 2

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex. Results may vary.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details.
No product or component can be absolutely secure.

Texas Advanced Computing Center (TACC) Frontera references
Article: *HPCWire: Visualization & Filesystem Use Cases Show Value of Large Memory Fat Notes on Frontera*.
www.intel.com/content/dam/support/us/en/documents/memory-and-storage/data-center-persistent-mem/Intel-Optane-DC-Persistent-Memory-Quick-Start-Guide.pdf
software.intel.com/content/www/us/en/develop/articles/introduction-to-programming-with-persistent-memory-from-intel.html
wreda.github.io/papers/assise-osdi20.pdf

KFBIO
KFBIO m. tuberculosis screening detectron2 model throughput performance on 2nd Intel® Xeon® Gold 6252 processor: NEW: Test 1 (single instance with PyTorch 1.6: Tested by Intel as of 5/22/2020. 2-socket 2nd Gen Intel® Xeon® Gold 6252 Processor, 24 cores, HT On, Turbo ON, Total Memory 192 GB (12 slots/16 GB/2666 MHz), BIOS: SSE5C620.86B.02.01.0008.031920191559 (ucode: 0x500002c), Ubuntu 18.04.4 LTS, kernel 5.3.0-51-generic, mitigated Test 2 (24 instances with PyTorch 1.6: Tested by Intel as of 5/22/2020. 2-socket 2nd Gen Intel Xeon Gold 6252 Processor, 24 cores, HT On, Turbo ON, Total Memory 192 GB (12 slots/16 GB/2666 MHz), BIOS: SSE5C620.86B.02.01.0008.031920191559 (ucode: 0x500002c), Ubuntu 18.04.4 LTS, kernel 5.3.0-51-generic, mitigated BASELINE: (single instance with PyTorch 1.4): Tested by Intel as of 5/22/2020. 2-socket 2nd Gen Intel Xeon Gold 6252 Processor, 24 cores, HT On, Turbo ON, Total Memory 192 GB (12 slots/16 GB/2666 MHz), BIOS: SSE5C620.86B.02.01.0008.031920191559 (ucode: 0x500002c), Ubuntu 18.04.4 LTS, kernel 5.3.0-51-generic, mitigated.

Tangent Studios
Configurations for Render Times with Intel® Embree, testing conducted by Tangent Animation Labs. Render farm: 8x Intel® Core™ processors +hyperthread*2 + 128gig. In-office workstations: Intel® Xeon® processors HP blade c7000 chassis, with HP460 gen8 blades - 2x Intel Xeon E5-2650 V2, Eight Core 2.6GHz-128GB. Software: Blender 2.78 with custom build using Intel® Embree. For more information on Tangent's work with Embree, watch this video:
www.youtube.com/watch?time_continue=251&v=_2Ia4h8q3xs&feature=emb_logo
Recreation of the performance numbers can be recreated using Agent327, Blender and Embree.

Chaos Group - Up to 90% Memory Reduction for Displacement
Testing conducted by Chaos Group with Intel® Embree 2020. Software Corona Renderer 5 with Intel Embree. Up to 90% memory reduction calculated using Corona Renderer 5 with regular displacement grids per triangle of 154 bytes versus Corona Renderer 5 with Intel Embree, which has a displacement capability grid of 12 bytes per grid triangle. (12/154 = 7.8% usage or >90% memory reduction.) Recreation of the performance numbers can be accomplished using Corona Renderer 5 and Embree. For more information, visit the Corona Renderer Blog: blog.corona-renderer.com/corona-renderer-5-for-3ds-max-released/

The Addams Family 2 - **Gained a 10% to 20%—and sometimes 25%—efficiency in rendering, saving thousands of hours in rendering production time**.
Testing Date: Results are based on data conducted by Cinesite 2020-21. 10% to up to 25% rendering efficiency/thousands of hours saved in rendering production time/15 hrs per frame per shot to 12-13 hrs.
Cinesite Configuration: 18-core Intel® Xeon® Scalable processors (W-2295) used in render farm, 2nd gen Intel Xeon processor-based workstations (W-2135 and -2195) used. Rendering tools: Gaffer, Arnold, along with optimizations by Intel® Open Image Denoise.

Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

Notices & Disclaimers – 2 of 2

HippoScreen increased AI performance by 2.4x – Test configurations Intel® Xeon® Gold 6330 CPU @ 2.00 GHz, 2 sockets, 28 cores per socket, note: www.intel.com /performanceindex

Univ. of Calif. at Davis achieved 3x performance increase & >100x data compression for scientific rendering – VIDI researchers achieved compression rates of several hundred times by implementing a new compression mechanism that uses a combination of positional encoding (Figure 2) and multilayer perceptrons (Figure 3) to learn a mapping from sampling positions to volume densities. Figure 1. A comparison of training results between the proposed method and two state-of-the-art techniques: fV-SRN, which was adjusted to match the models' compression ratios, and tthresh, which was adjusted to match our models' PSNRs after 20k steps of training. Configurations: Models were trained on Windows* machine with RTX 3090, while fV-SRN models were trained on a Linux* machine with a faster RTX 3090Ti and Intel® Xeon® Scalable processor (E5-2699) due to operating system compatibility issues. The tthresh experiments were run on an 88-core Linux server with 256 GB of memory because tthresh is a CPU-based algorithm. The table indicates experiments that ran out of memory (as OOM). The method outperforms state-of-the-art techniques, with the best and worst results within each category highlighted in red and gray, respectively. Figure 13. Relative speed-up times when compared to the baseline ray-marching renderer. An Intel® Xeon® Scalable processor (E5-2699) with 256 gigabytes of memory were the 88-core/176-thread workhorses used to render datasets and train the machine learning models that powered these projects. More details are in *Instant Neural Representation for Interactive Volume Rendering*

The Addams Family 2/Cinesite achieving up to 25% efficiency in rendering - Testing Date: Results are based on data conducted by Cinesite 2020-21. 10% to up to 25% rendering efficiency/thousands of hours saved in rendering production time/15 hrs per frame per shot to 12-13 hrs. Cinesite Configuration: 18-core Intel® Xeon® Scalable processors (W-2295) used in render farm, 2nd gen Intel Xeon processor-based workstations (W-2135 and -2195) used. Rendering tools: Gaffer, Arnold, along with optimizations by Intel® Open Image Denoise.

University of Utah -: the performance for growing problem sizes (from 10'000x10'000 pixels to 40'000x40'000 mosaic) using Intel CPU (Intel® Xeon® Platinum 8480+) and GPU (Intel® Data Center GPU Max 1100) available on the Intel® Developer Cloud and using the original code on an NVIDIA A6000 on a workstation.

Ali- DP **Testing Date:** Performance results are based on **testing by Alibaba as of August 21, 2022**. Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

Configuration Details and Workload Setup:. 3rd Gen Intel® Xeon® Scalable processors 8369B CPU @ 2.70GHz, 32v CPU(s), 64G memory, 40G ESSD. Lammps configuration file: Lammps default configuration file. Release: 23 Jun 2022, Iteration Count: 2M, Number of test processes and threads: 32P1T, 16P2T, 8P4T. Comparing compilers: GCC-10.2, Intel(R) oneAPI DPC++/C++ Compiler 2022.0.0 (2022.0.0.20211123). Performance evaluation indicators: the exectuion time. Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See configuration disclosure for details. No product or component can be absolutely secure. Performance varies by use, configuration, and other factors. Learn more at www.Intel.com/PerformanceIndex. Your costs and results may vary

Performance varies by use, configuration and other factors. Learn

Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.