

Freedom, Productivity, and Performance for Accelerated Computing

Next-Gen HPC + AI

Intel's Open, efficient, performant portfolio

Bob Gaines

HPC + AI Solutions Architect

Public Sector + Financial Services

April 9-10, 2024



Notices & Disclaimers

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

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.

Your costs and results may vary.

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

Results have been estimated or simulated.

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

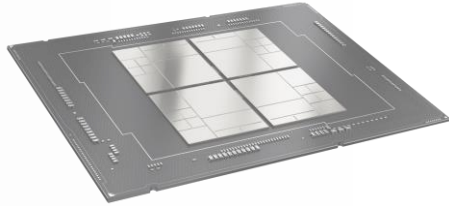
All product plans and roadmaps are subject to change without notice.

Statements in this document that refer to future plans or expectations are forward-looking statements. These statements are based on current expectations and involve many risks and uncertainties that could cause actual results to differ materially from those expressed or implied in such statements. For more information on the factors that could cause actual results to differ materially, see our most recent earnings release and SEC filings at www.intc.com.

Code names are used by Intel to identify products, technologies, or services that are in development and not publicly available. These are not "commercial" names and not intended to function as trademarks.

© Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.

Executing on Our Acceleration Roadmap

	2023	Today	2024 (closely following Sierra Forest)
CPU P-Core	 <p>4th Gen Intel® Xeon® Scalable processors</p>	 <p>5th Gen Intel® Xeon® codenamed Emerald Rapids</p>	 <p>Next-Gen Intel® Xeon® codenamed Granite Rapids</p>
Discreet Accelerators	 <p>Intel® Data Center GPU Flex Series Intel® Data Center GPU Max Series Intel® Gaudi® 2 AI accelerator</p>	 <p>Intel® Gaudi® 3 AI accelerator</p>	 <p>Next Gen GPU/AI Accelerator</p>

Dawn Phase 1, University of Cambridge, UK

- Fastest AI supercomputer deployed in the UK today
- Concept to HPL run in less than 6 months, installation in less than 4 weeks
- Dawn will drive advancements in healthcare, green fusion energy, climate modelling
- Co-designed by Intel, Dell Technologies, University of Cambridge and UKAEA
- Dell PowerEdge XE9640 with 4th Gen Intel Xeon and Intel Data Center GPU Max Series
- Direct liquid cooling to provide a more power-efficient and cost-effective solution

Compute			Fabric		
512 CPUs	1024 GPUs	256 Nodes	25.6 TB/s Peak Injection Bandwidth	12.8 TB/s Peak Bisection Bandwidth	
Memory				Storage	
256 TB DDR Capacity	157 TB/s Peak DDR BW	128 TB HBM Capacity	3.3 PB/s Peak HBM BW	3 PB Storage	2 TB/s Bandwidth



SuperMUC-NG Phase 2, LRZ, Germany

- Among the [fastest HPC systems in Germany](#) as of today
- Broad AI/HPC userbase to drive life & environmental sciences
- Software stack enabled by [oneAPI](#) for easy portability
- [Lenovo ThinkSystem SD650-I V3 Neptune DWC](#) servers
- Hot water cooling to increase efficiency and lower the TCO
- Assembled in Europe to reduce carbon footprint and shipping timelines

Compute			Fabric		
480 CPUs	960 GPUs	240 Nodes	12 TB/s Peak Injection Bandwidth	6 TB/s Peak Bisection Bandwidth	
Memory				Storage	
123 TB DDR Capacity	147 TB/s Peak DDR BW	123 TB HBM Capacity	3.1 PB/s Peak HBM BW	1 PB DAOS Capacity	750 GB/s 42 DAOS Bandwidth Nodes

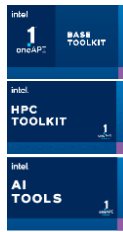


Leibniz Supercomputing Centre
of the Bavarian Academy of Sciences and Humanities



oneAPI

Specification and Open Source



Open industry initiative driving a vendor-neutral software ecosystem for multiarchitecture accelerated computing.

Now governed by the Linux Foundation.



Freedom to Make Your Best Choice

- An open alternative to single-vendor/proprietary lock-in enables easy architecture retargeting
- Open, standards-based programming (C++ with SYCL) so software investments continue to add value in future hardware generations

Performance – Realize All the Hardware Value

- Expose and exploit all the cutting-edge features and maximize performance across CPUs, GPUs, FPGAs, and other accelerators.
- Powerful libraries for acceleration of domain-specific functions

Productivity – Develop Performant Code Quickly

- One programming model for all – easy integration with existing code including migration of CUDA code to SYCL
- Based on familiar C++ – no need to learn a new language
- Interoperable with existing HPC standards including Fortran, C/C++, OpenMP, and MPI, as well as Python with a rich set of optimized Python libraries

Visit oneapi.io or <https://uxlfoundation.org/> for more details

*Other names and brands may be claimed as the property of others. SYCL is a trademark of the Khronos Group Inc.

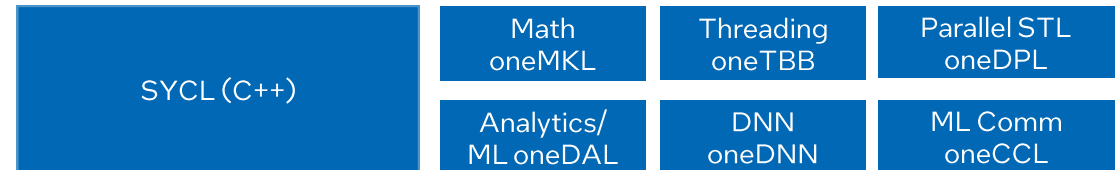
Middleware and Frameworks



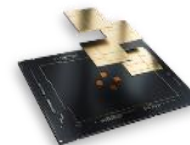
oneAPI Industry Specification

Direct Programming

API-Based Programming



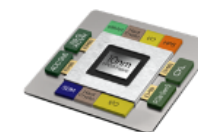
Low-Level Hardware Interface (oneAPI Level Zero)



CPU



GPU



FPGA



Other Accelerators

Intel® HPC + AI Portfolio

Open Software Environment



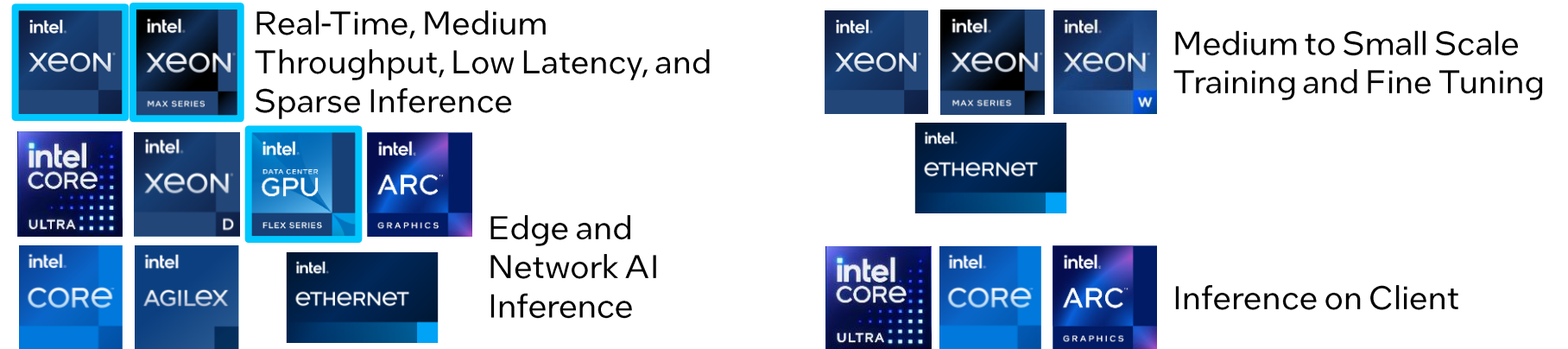
Deep Learning Acceleration



General Acceleration

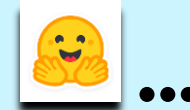
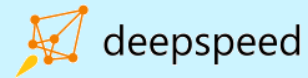


General Purpose



Seamless ...Transitioning

Performant AI Code with Minimal Changes



1
oneAPI

Unified Programming Model

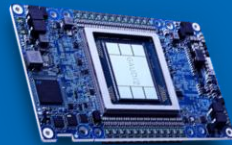
Intel® Gaudi® Software Suite

1
oneAPI  **UXL**
Unified Acceleration Foundation

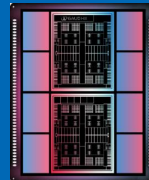
Transitioning Into a
Single Software Environment



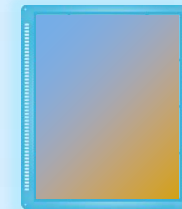
Intel® Data Center GPU Max Series



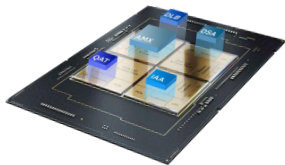
Intel® Gaudi® 2 AI accelerator



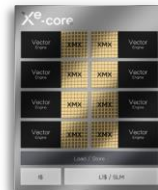
Intel® Gaudi® 3 AI accelerator



Next-Gen GPU



Intel® Accelerator Compatibility



Intel® Data Center Acceleration Series

Accelerate your HPC and AI workloads!

Highly Performant

Competitive performance across popular AI and HPC workloads

Versatile

Designed to handle a broad array of AI and HPC workloads

Open

Say goodbye to proprietary vendor lock-in

Available

Tens of thousands shipped since launch

