CABOT PARTNERS
OPTIMIZING BUSINESS VALUE

Intel Innovation 2023: Intel Expanding Its Product Roadmap Around AI Growth

Jean S. Bozman, President,
Cloud Architects Advisors LLC
M. R. Pamidi, Ph.D., Principal Analyst

September 2023

Cabot Partners Group, Inc.
100 Woodcrest Lane, Danbury, CT 06810.
www.cabotpartners.com
info@cabotpartners.com

Intel Expanding Its Product Roadmap Around AI Growth

Intel's Innovation 2023 conference provided a new view of the company's roadmap for hardware and software as platforms and enabling software are being transformed to help customers navigate that new landscape. This time, Artificial Intelligence (AI) was presented as a focal point for Intel's product roadmap on the hardware and software sides of AI deployments.
Because the Intel Innovation event is first and foremost a software and hardware developers' conference for next-generation products and services, it showcased software tools, use-cases, and AI-enabled customer solutions in retail, healthcare, manufacturing, and finance, among other sectors—all with the *Bringing AI Everywhere* mantra. The ecosystem around the rapidly expanding AI marketplace is also of great interest to Intel as it looks to grow revenues and profits commensurate with the AI growth that is happening throughout the world.

A Bit of Recent History

The AI revolution, which caught the public's attention following the emergence of ChatGPT in November 2022, is impacting the full spectrum of AI software – from desktops to the largest supercomputers. Many elements of this rapidly expanding AI ecosystem already exist, but hardware and software providers are shipping many more products and services into the marketplace. Intel's strong position in CPUs and its plan to grow its position in GPUs is a leverage point for growing the AI use cases for its vendor partners, cloud providers, and the worldwide customer base for Intel-based systems.

Intel is showing how its roadmap for next-generation products is changing – especially in markets where AI is expected to drive rapid growth in developer tools and AI workloads. It is blending its roadmap and AI messages to show that its processors and silicon (e.g., FPGAs, chiplets) will be building blocks for AI and Machine Learning (ML) systems. Key characteristics of AI hardware include low power/cooling costs, high-density compute capability, and the ability to scale up AI hardware resources as AI/ML tasks demand.

The Intel Development Cloud

The Intel Developer Cloud, in beta-test since last year (2022), debuted at this Innovation conference – providing a platform and work environment for developers to build, try, test, and iterate new AI applications. At the same time, Intel presented a multi-layered view of its software and hardware that would, when used together, host customers' applications and data for the Core, the Cloud, and the Edge – all of them jumping-off points for customers' AI deployments.

Conference Announcements

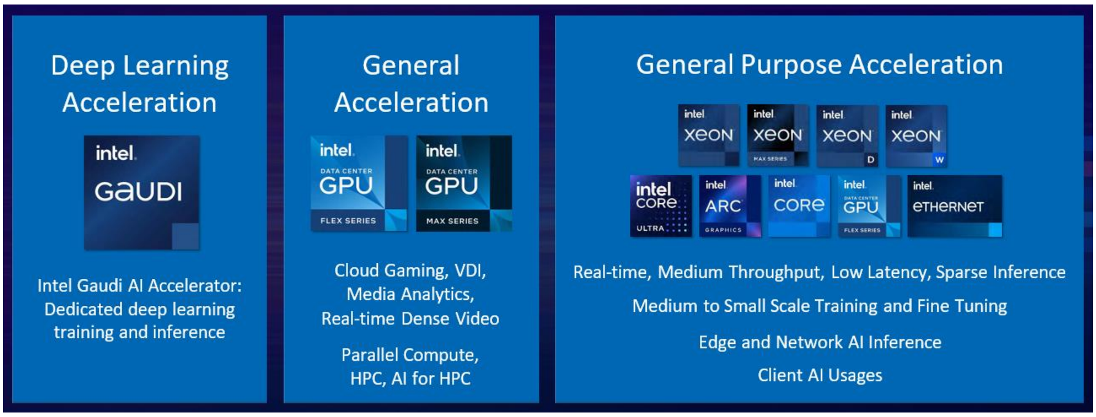
Intel's reach into the PC, server, storage, and networking/communications markets is broad, and all of them are being transformed by AI and AI-enabled workloads.

At the conference, Intel CEO Pat Gelsinger's keynote introduced a new word, *Siliconomy* -- combining the words *silicon* and *economy.* Here's why: *the worldwide silicon marketplace saw a* 4-fold increase in connected devices over the past four years and a forecasted 15-fold growth over the next ten years.

According to industry estimates, Silicon is already a $574 billion industry worldwide, and the global tech economy is worth $8 trillion worldwide. That is why Intel CEO Pat Gelsinger and Intel CTO Greg Lavender announced and placed into context these critical pieces of AI-enabled infrastructure. Highlights of these announcements included the following topics.

Processors

Two keynotes were delivered – one by CEO Pat Gelsinger and the other by CTO Greg Levander. Focusing on AI's importance in the marketplace, these keynotes provided a glimpse of current and upcoming Intel® processors from the Cloud to Edge to Client—covering the entire spectrum of computing (Figure 1).



Deep Learning Acceleration

Intel Gaudi AI Accelerator: Dedicated deep learning training and inference

General Acceleration

Cloud Gaming, VDI, Media Analytics, Real-time Dense Video

Parallel Compute, HPC, AI for HPC

General Purpose Acceleration

Real-time, Medium Throughput, Low Latency, Sparse Inference

Medium to Small Scale Training and Fine Tuning

Edge and Network AI Inference

Client AI Usages

Figure 1: Intel AI: From Cloud to Edge to Client

At a high level, highlights of the Intel Innovation announcements include:

- Intel Core Ultra (code-named Meteor Lake), a new client processor built on the Intel 4 process node using Intel's 3D high-performance hybrid architecture, is designed to deliver performance, power efficiency, and AI-enabled features at scale. The product is scheduled to launch in December 2023.
- **Intel Core processor** roadmap – showing that next-generation processors code-named Arrow Lake, Lunar Lake, and Panther Lake —will reach silicon readiness by 2025, all with AI capabilities (Figure 2).
- **Intel Gaudi 2 processor** for AI workloads, made by Intel's Habana Labs (Intel acquired Habana, a maker of programmable deep-learning accelerators, in 2019). Intel's Gaudi is commercially available and has been submitted to MLPerf for benchmarking inference and training. Gaudi uses a cluster of programmable Tensor Processing Cores (TPCs) that work together with software tools and software "libraries" to tackle AI workflows. It is not a GPU but will work on AI/ML tasks, including at-scale "dedicated training" of LLMs and ML tracing.
- **Intel Sierra Forest.** As noted at the conference, a 288-core version of the efficient core (E-Core) Intel Xeon processor family, code-named Sierra Forest, will become generally available in 2024. It will be based on the Intel 3 process. Sierra Forest is designed to be a fast and energy-efficient chip for hyper-scalers and cloud service providers (CSPs) that must balance performance requirements with power efficiency in their data centers. It is expected to deliver 2.5 times better rack density and 2.4 times higher performance/watt than Intel 4th-generation processors.

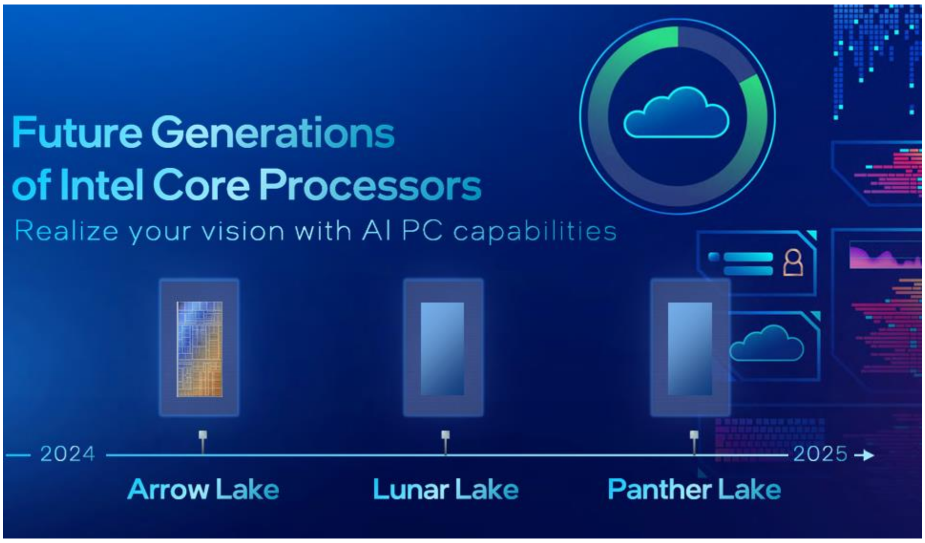

Future Generations of Intel Core Processors
Realize your vision with AI PC capabilities
2024
2025
Arrow Lake
Lunar Lake
Panther Lake

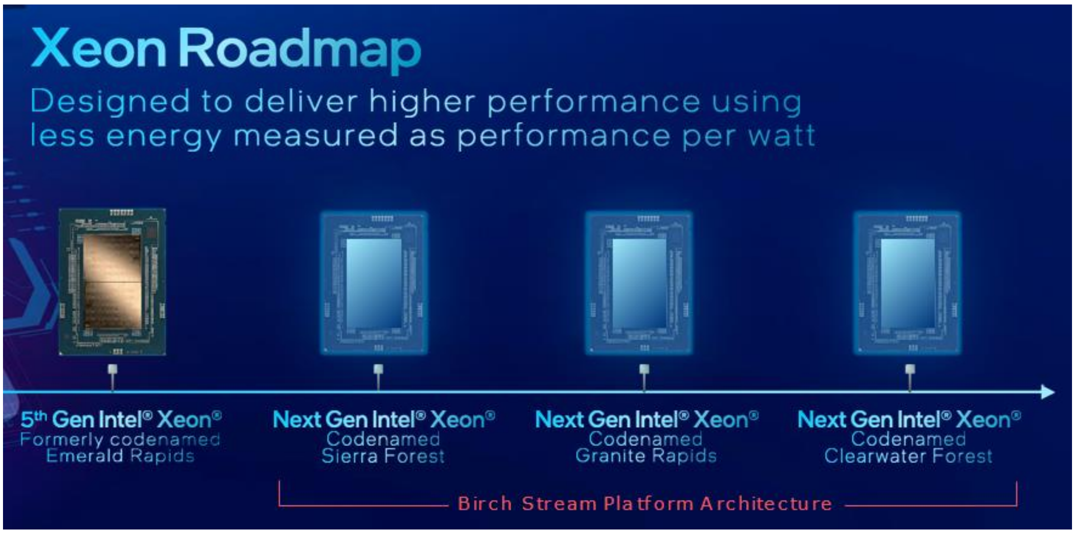
Figure 2: Intel Core Processors Roadmap

In 2024 and 2025, a generation of AI PCs based on Intel processors will ship into the marketplace, running AI tasks (e.g., training models, inference, visualization, and recommendations). For this PC use case, AI processing can be done locally rather than in the cloud, reducing latency and dependency on high bandwidth, especially for compute- and storage-intensive applications. Overall, these attributes will improve performance for customers of Intel-based client systems.

The Intel Xeon roadmap will include more features for high performance and low energy (measured as performance/watt) (Figure 3), both attributes good for AI workloads, the company said.

The roadmap includes:

- **5th Generation Intel Xeon® Scalable Processor**. It was formerly codenamed Emerald Rapids and is scheduled to be launched on December 14, 2023. Intel executives view this launch as an inflection point, signaling Intel's return to more competitive CPU processor speeds. Intel plans to retain its market share of over 80% of CPUs worldwide and grow it with AI-enabled systems.
- **Next-Generation Processor Roadmap.** The new processor roadmap will fulfill Intel's promise of delivering "five nodes in four years" – starting with the Intel 7 "node" in high-volume production; the Intel 4 node, now ready for manufacturing, and the Intel 3 node, which is due to debut by late 2023. These are all Intel designations for the node types on the roadmap, as shown in on-stage keynotes and presentations at the Innovation Conference.
- **Intel Xeon Sierra Forest.** With 288 cores (see the discussion above), this processor will be launched in 2024, followed by Xeon *Granite Rapids* and Xeon *Clearwater Forest,* expected in 2024 and 2025, respectively.
- **Intel Data Center Max Series GPU.** At the conference, Intel discussed the Intel Data Center Max Series GPU, which packs over 100 billion transistors into a package and contains up to 128 X^{e} cores, Intel's foundational GPU compute building block, designed for high performance in AI and HPC applications. As announced at the Intel Innovation conference, the Aurora Supercomputer, a powerful supercomputer built upon Max Series GPUs, is installed at the Argonne National Laboratory in Illinois. Aurora will showcase the power of pairing Max 1550 Series GPUs and CPUs in a single system with over 10,000 blades, each containing six Max Series GPUs and two Xeon Max CPUs. With 63,744 Max 1550 Series GPUs, Intel described the Aurora Supercomputer as one of the largest GPU clusters in the world.



Xeon Roadmap

Designed to deliver higher performance using less energy measured as performance per watt

5th Gen Intel® Xeon® Formerly codenamed Emerald Rapids

Next Gen Intel® Xeon® Codenamed Sierra Forest

Next Gen Intel® Xeon® Codenamed Granite Rapids

Next Gen Intel® Xeon® Codenamed Clearwater Forest

Birch Stream Platform Architecture

Figure 3: Intel Xeon Roadmap

Products for Developers/Programmers

Most of the Intel software tools for application developers are based on open-source code, making it accessible to a broader spectrum of the world's software developers/programmers. The Intel Developer Cloud's toolset allows developers to invoke AI logic and programming algorithms on open hardware platforms conforming to open-systems APIs. This will pave the way to broader accessibility to Intel's technology for AI software and systems-level OEM products.

Developers' Environment

Intel CTO Greg Levander provided a more granular look at how Intel's product line is changing – both in hardware and software. Significant software tools for developing AI, based on deep-learning processes such as Intel Geti – a software platform that allows subjects- to build AI models in a fraction of the time, using less data – will enable applications to be deployed for distributed, parallelized processors using Intel's OpenVINO software for developers and programmers.

We see OpenVINO as a key enabler for Intel's accelerated roadmap for supporting AI workloads. OpenVINO (supporting programming for deployments) will be paired with Intel Geti (supporting the development and testing of AI tasks). Examples at the conference demos were AI for factory processes and AI for modeling real-world dynamic applications, such as simulations of racecar pit stops.

OpenVINO is an acronym for Open Visual Inference and Neural Network Optimization. It is an open-source toolkit designed to help developers optimize and deploy deep learning models on various hardware platforms, such as CPUs, FPGAs, GPUs, and NPUs (Neural Processing Units) (see the workflows, as shown in Figure 4).

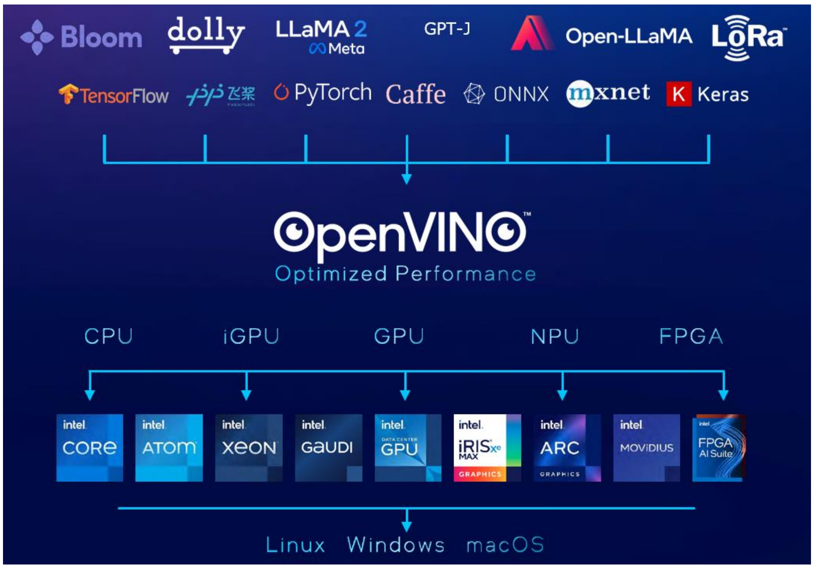


Figure 4: AI Workflows on Intel Processors

OpenVINO's key capabilities include:

- **Cross-Platform Support**: Designed to be cross-platform, it allows developers to target a wide range of operating systems and hardware configurations, making it versatile for various deployment scenarios.
- **Custom Layer Support**: Allows developers to implement custom layers and plugins for their specific hardware configurations, allowing them to optimize their deep learning models further.
- **Edge Computing**: Well-suited for edge computing scenarios, where processing happens locally on devices with limited computational resources. This enables real-time inference on edge devices, vital for AI applications involving autonomous vehicles, robotics, and surveillance systems, to name three examples.
- **Runtime:** Runtime allows developers to deploy optimized models for real-time inference on Intel-based devices, taking advantage of hardware acceleration, such as Intel's CPUs, FPGAs, GPUs, and NPUs, to achieve high inference performance.
- **Integration with computer-vision libraries:** Integrates with popular computer-vision libraries like OpenCV, making it easier to work with computer-vision tasks in combination with deep-learning models.
- **Model Conversion API:** Provides tools for optimizing and converting deep learning models trained in popular frameworks like TensorFlow, PyTorch, and ONNX (Open Neural Network Exchange) into a format that can be efficiently executed on cross-platform hardware.
- **Model Zoo: The Model Zoo** provides pre-trained deep-learning models for various tasks, including object detection, image segmentation, text recognition, and other functionalities that can be used and fine-tuned for specific applications.

Leveraging oneAPI Open-Source for Developers

Intel promotes oneAPI as a powerful tool to help developers write portable, performant, and productive code for various hardware platforms. Intel's oneAPI is an open, cross-architecture programming model and set of tools for developing high-performance, data-centric applications across different hardware platforms. Importantly, this includes CPUs, GPUs, and FPGAs, providing many system choices for Intel's OEMs. It allows developers to tap the latest hardware innovations without rewriting their code for each platform. This makes it easier for developers to port their applications to new hardware platforms.

Developers' benefits include performance, portability, and developer productivity. Intel's oneAPI code can be executed on a variety of hardware platforms, including CPUs, GPUs, and FPGAs, without having to be rewritten. This makes it easier for developers to port their applications to new hardware platforms and to take advantage of the latest hardware innovations.

In the age of multi-cloud applications, this is an important attribute – because most end-to-end multi-cloud deployments run on a mix of processor types. This is also true of CSP infrastructure, comprising AMD, ARM, Intel, NVIDIA, and other processor types across a widely scaled-out, distributed deployment supported in expansive CSP data centers located around the world.

Global Strategy

Intel CEO Gelsinger cited the company's existing or planned offices and/or fabs around the world. His comments illuminated a broad plan to develop parallel supply chains across the globe, increasing overall fab production and improving fab-to-OEM delivery times across all major geographic regions of the world.

This plan will enable a "follow the sun" strategy, allowing Intel to produce multiple processors at two or more locations, depending on product type and supply chain.

The Intel method of a proven "copy exact" principle will allow Intel to quickly transfer processes between two or more fabs, speeding production for new products—worldwide. This strategy will apply to Intel fabs in Oregon, Arizona, New Mexico, Ohio (under construction), and Costa Rica in the Americas; Ireland, Germany, Czech Republic, and Poland in Europe; and China, Taiwan, Malaysia, Vietnam, and India (planned) in Asia/Pacific. Intel is one of the few fab owners that can boast of building such a large worldwide presence. Across the industry, taking a global approach to manufacturing is expected to impact the geographic mix of semiconductor products worldwide.

Summary

Intel can make the case that it is laying the groundwork for a more stable production (and sourcing) of processors, FPGAs, and other silicon-based products than is seen today, with nearly 90% of all of the world's semiconductor products made in one geographic region – Asia/Pacific. Diversification is driving more fab-building and gaining support for governments in the U.S. and EMEA to help fund the fab-building construction process.

Why is this happening? Intel, with more than 80% market share in the CPU market, knows that it has to accelerate its move to new, smaller processor dimensions (e.g., 5nm and below) as it continues to compete with AMD in CPUs, and to both compete with and partner with its GPU competitor NVIDIA. NVIDIA is the leader in the worldwide GPU market, with an 80.2 % worldwide market share, followed by AMD and Intel.

In summary, Intel has always been a leading global provider of semiconductor products, and it is entering a new wave of diversified and differentiated product innovation and manufacturing. Intel plans to ride the wave of fast-growing AI deployments to increase its market share in a range of markets, including CPUs, GPUs, PCs, and communications/networking devices, on a worldwide basis.

CABOT PARTNERS
OPTIMIZING BUSINESS VALUE

Cabot Partners Group, Inc.

100 Woodcrest Lane, Danbury, CT 06810

www.cabotpartners.com | info@cabotpartners.com