August 19, 2021

Intel Editorial: Intel Advances Architecture for Data Center, HPC-AI and Client Computing

Intel unveils new architectures for two x86 CPU cores, two data center SoCs, two discrete GPUs and a revolutionary multicore performance hybrid architecture for client.

SANTA CLARA, Calif.--(BUSINESS WIRE)-- *The following is an opinion editorial from Raja Koduri, senior vice president and general manager of the Accelerated Computing Systems and Graphics Group at Intel Corporation.*

This press release features multimedia. View the full release here: https://www.businesswire.com/news/home/20210819005219/en/

Raja Koduri, Intel senior vice president and general manager of the Accelerated Computing Systems and Graphics Group, displays a wafer with Intel Arc high-performance discrete graphics hardware as part of a presentation during Intel Architecture Day 2021. The virtual event was held in August 2021. (Credit: Intel Corporation)

Architecture is alchemy of hardware and software. It blends the best transistors for a given engine, connects them through advanced packaging, integrates high-bandwidth, low-power caches, and equips them with high-capacity, high-bandwidth memories and low-latency scalable interconnects for hybrid computing clusters in a package, while also ensuring that all software accelerates seamlessly. Disclosing the architectural innovations that they have been working on for products that are imminent is something Intel's architects look forward to each year at Intel Architecture Day, and this year's event – our third – was the most exciting yet.

Today, we unveiled our biggest shifts in Intel® architectures in a generation. This includes the first in-depth look at Alder Lake, our first performance hybrid architecture with two new generations of x86 cores and the intelligent Intel® Thread Director workload scheduler; Sapphire Rapids, Intel's new standard-setting data center architecture with our new Performance-core and various accelerator engines; our new discrete gaming graphics processing unit (GPU) architecture; new infrastructure processing units (IPUs); and Ponte Vecchio, our tour-de-force data center GPU architecture with Intel's highest ever compute density.

These architectural breakthroughs set the stage for our next era of leadership products, starting soon with Alder Lake. The breakthroughs we disclosed today also demonstrate how architecture will satisfy the crushing demand for more compute performance as workloads from the desktop to the data center become larger, more complex and more diverse than ever.

More: Intel Architecture Day 2021 (Press Kit) | Expanding Intel's Foundry Partnerships: A Critical Piece of IDM 2.0 (Stuart Pann Editorial) | Intel Unveils Biggest Architectural Shifts in a Generation for CPUs, GPUs and IPUs (Architecture Day Fact Sheet)

Our architects are working hard, combining Intel's unique and rich selection of scalar, vector, matrix and spatial compute engines, to create hybrid computing architectures that deliver non-linear gains on our customers' most demanding workloads.

You should visit our Newsroom press kit to watch the presentations, but let me walk you through the highlights:

Efficient-core: A highly scalable x86 microarchitecture for addressing compute requirements across the entire spectrum of our customers' needs, from low-power mobile applications to many-core microservices. Compared with Skylake, Intel's most prolific CPU microarchitecture, the Efficient-core delivers 40% more single-threaded performance at the same power, or the same performance while consuming less than 40% of the power.[1] For throughput performance, four Efficient-cores deliver 80% more performance while still consuming less power than two Skylake cores running four threads or the same throughput performance while consuming 80% less power.[1]

Performance-core: This x86 core is not only the highest performing CPU core Intel has ever built, but it also delivers a step function in CPU architecture performance that will drive the next decade of compute. It was designed as a wider, deeper and smarter architecture to expose more parallelism, increase execution parallelism, reduce latency and increase general purpose performance. It also helps support large data and large code footprint applications. Performance-core provides a Geomean improvement of about 19%, across a wide range of workloads over our current 11th Gen Intel® Core™ architecture (Cypress Cove core) at the same frequency.[1]

Targeted for data center processors and for the evolving trends in machine learning, Performance-core brings dedicated hardware, including Intel's new Advanced Matrix Extensions (AMX), to perform matrix multiplication operations for an order of magnitude performance – a nearly 8x increase in artificial intelligence acceleration.[1] This is architected for software ease of use, leveraging the x86 programming model.

Intel Thread Director: Intel's unique approach to scheduling was developed to ensure Efficient-cores and Performance-cores work seamlessly together, dynamically and intelligently assigning workloads from the start and optimizing the system for maximum real-world performance and efficiency. With intelligence built directly into the core, Intel Thread Director works seamlessly with the operating system to place the right thread on the right core at the right time.

Alder Lake: Reinventing the multicore architecture, Alder Lake will be Intel's first performance hybrid architecture with the new Intel Thread Director. This is Intel's most intelligent client system-on-chip (SoC) architecture, featuring a combination of Efficient-cores and Performance-cores, scaling from ultra-mobile to desktop, and leading the industry transition with multiple industry leading I/O and memory. Products based on Alder Lake will begin shipping this year.

X^e HPG and Alchemist SoC: A new discrete graphics microarchitecture is designed to scale to enthusiast-class performance for gaming and creation workloads. The X^e HPG microarchitecture features a new X^e-core, a compute-focused programmable and scalable element, and full support for DirectX 12 Ultimate. New matrix engines inside the X^e-cores (referred to as X^e Matrix eXtensions, XMX) accelerate artificial intelligence workloads such as X^eSS, a novel upscaling technology that enables high-performance and high-fidelity gaming. X^e HPG-based Alchemist SoCs (formerly code-named DG2) will be coming to market in the first quarter of 2022 under the new Intel® Arc™ brand.

Sapphire Rapids: Combining Intel's Performance-cores with new accelerator engines, Sapphire Rapids sets the standard for next-generation data center processors.At the heart of Sapphire Rapids is a tiled, modular SoC architecture that delivers significant scalability while still maintaining the benefits of a monolithic CPU interface thanks to Intel's EMIB multi-die interconnect packaging technology and advanced mesh architecture.

Infrastructure Processing Unit: Mount Evans is Intel's first dedicated ASIC-based IPU, along with a new FPGA-based IPU reference platform, Oak Springs Canyon. With an Intel IPU-based architecture, cloud service providers (CSPs) can maximize data center revenue by offloading infrastructure tasks from CPUs to IPUs. Offloading infrastructure tasks to the IPU allows CSPs to rent 100 percent of their server CPUs to customers.

X^e HPC, Ponte Vecchio: The most complex SoC Intel has ever built and a great example of our IDM 2.0 strategy come to life, Ponte Vecchio takes advantage of several advanced semiconductor processes, our revolutionary EMIB technology, and our Foveros 3D packaging. With this product, we are bringing to life our moon-shot project, the 100 billion-transistor device that delivers industry-leading FLOPs and compute density to accelerate artificial intelligence, high performance computing and advanced analytics workloads. At Architecture Day, we showed that our early Ponte Vecchio silicon is already demonstrating leadership performance, setting an industry-record in both inference and training throughput on a popular AI benchmark.[1] Our A0 silicon is already providing greater than 45 TFLOPS FP32 throughput, greater than 5 TBps Memory Fabric bandwidth and greater than 2 TBps connectivity bandwidth. Ponte Vecchio, as with our X^e architectures, will be enabled by oneAPI, our open, standards-based, cross-architecture and cross-vendor unified software stack.

Looking back at just the past year, technology was at the heart of how we all communicated, worked, played and coped through the pandemic. Enormous computing power proved crucial. Looking ahead, we face a massive demand for compute – potentially a 1,000x need by 2025. That 1,000-times boost in four years is Moore's Law to the power of five.

As our CEO, Pat Gelsinger, also an architect, stated at Architecture Day: "We face daunting compute challenges that can only be solved through revolutionary architectures and platforms … Our talented architects and engineers made possible all this technology magic."

The world is counting on architects and engineers to solve the most difficult computational problems, to enrich people's lives. Our strategy and execution are accelerating to meet these demands – at a torrid pace.

About Intel

Intel (Nasdaq: INTC) is an industry leader, creating world-changing technology that enables global progress and enriches lives. Inspired by Moore's Law, we continuously work to advance the design and manufacturing of semiconductors to help address our customers' greatest challenges. By embedding intelligence in the cloud, network, edge and every kind of computing device, we unleash the potential of data to transform business and society for the better. To learn more about Intel's innovations, go to newsroom.intel.com and intel.com.

[1] For workloads and configurations visit www.intel.com/ArchDay21claims.

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.

View source version on businesswire.com:
https://www.businesswire.com/news/home/20210819005219/en/

Bruce Fienberg
1-510-220-6461
bruce.fienberg@intel.com

Source: Intel Corporation