intel FACT SHEET

Intel Unveils Biggest Architectural Shifts in a Generation for CPUs, GPUs and IPUs

Intel powers the next era of computing for data center, edge and client for the workloads and computing challenges of tomorrow.

Aug. 19, 2021 – At Intel's Architecture Day 2021, Raja Koduri and Intel architects provided details on two new x86 core architectures; Intel's first performance hybrid architecture, code-named "Alder Lake," with the intelligent Intel® Thread Director workload scheduler; "Sapphire Rapids," the next-generation Intel® Xeon® Scalable processor for the data center; new infrastructure processing units; and upcoming graphics architectures, including the Xᵉ HPG and Xᵉ HPC microarchitectures, and Alchemist and Ponte Vecchio SoCs.

These new architectures will power upcoming high-performance products and establish the foundations for the next era of Intel innovation aimed at meeting the world's ever-growing demand for more computing power.

Raja Koduri addressed the importance of architectural advancement to meet this demand, saying: "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. … The breakthroughs we disclosed today 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."

x86 Cores

Efficient-core

Intel's new Efficient-core microarchitecture, previously code-named "Gracemont," is designed for throughput efficiency, enabling scalable multithreaded performance for modern multitasking. This is Intel's most efficient x86 microarchitecture with an aggressive silicon area target so that multicore workloads can scale out with the number of cores. It also delivers a wide frequency range. The microarchitecture and focused design effort allow Efficient-core to run at low voltage to reduce overall power consumption, while creating the power headroom to operate at higher frequencies. This allows Efficient-core to ramp up performance for more demanding workloads.

Efficient-core utilizes a variety of technical advancements to prioritize workloads without being wasteful with processing power and to directly enhance performance with features that improve instruction per cycle (IPC), including:

- 5,000 entry branch target cache that results in more accurate branch prediction
- 64 kilobyte instruction cache to keep useful instructions close without expending memory subsystem power
- Intel's first on-demand instruction length decoder that generates pre-decode information
- Intel's clustered out-of-order decoder that enables decoding up to six instructions per cycle while maintaining energy efficiency
- A wide back end with five-wide allocation and eight-wide retire, 256 entry out-of-order window and 17 execution ports
- Intel® Control-flow Enforcement Technology and Intel® Virtualization Technology Redirection Protection
- The implementation of the AVX ISA, along with new extensions to support integer artificial intelligence (AI) operations

Compared with the Skylake CPU core, Intel's most prolific central processing unit (CPU) microarchitecture, in single-thread performance, the Efficient-core achieves 40% more performance at the same power or delivers the same performance while consuming less than 40% of the power[1]. For throughput performance, four Efficient-cores offer 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

Intel's new Performance-core microarchitecture, previously code-named "Golden Cove," is designed for speed and pushes the limits of low latency and single-threaded application performance. Workloads are growing in their code footprint and demand more execution capabilities. Datasets are also massively growing along with data bandwidth requirements. Intel's new Performance-core microarchitecture provides a significant boost in general purpose performance and better support for large code footprint applications.

The Performance-core features a wider, deeper and smarter architecture:

- Wider: six decoders (up from four); eight-wide µop cache (up from six); six allocation (up from five); 12 execution ports (up from 10)
- Deeper: Bigger physical register files; deeper re-order buffer with 512 entry
- Smarter: Improved branch prediction accuracy; reduced effective L1 latency; full write predictive bandwidth optimizations in L2

The Performance-core is the highest performing CPU core Intel has ever built and pushes the limits of low latency and single-threaded application performance with:

- A Geomean improvement of ~19% across a wide range of workloads over current 11th Gen Intel® Core™ processor architecture (Cypress Cove) at ISO frequency for general purpose performance[1]
- Exposure for more parallelism and an increase in execution parallelism
- Intel® Advanced Matrix Extensions, the next-generation, built-in AI acceleration advancement, for deep learning inference and training performance. It includes dedicated hardware and new instruction set architecture to perform matrix multiplication operations significantly faster
- Reduced latency and increased support for large data and large code footprint applications

Client

Alder Lake Client SoC

Intel's next-generation client architecture, code-named Alder Lake, is Intel's first performance hybrid architecture, which for the first time integrates two core types – Performance-core and Efficient-core – for significant performance across all workload types. Alder Lake is built on the Intel 7 process and supports the latest memory and fastest I/O.

Alder Lake will deliver incredible performance that scales to support all client segments from ultra-portable laptops to enthusiast and commercial desktops by leveraging a single, highly scalable system-on-chip (SoC) architecture with three key design points:

- A maximum performance, two-chip, socketed desktop with leadership performance, power-efficiency, memory and I/O
- A high-performance mobile BGA package that adds imaging, larger X^e graphics and Thunderbolt 4 connectivity
- A thin, lower-power, high-density package with optimized I/O and power delivery

The challenge of building such a highly scalable architecture is meeting incredible bandwidth demands of the compute and I/O agents without compromising power. To solve this challenge, Intel has designed three independent fabrics, each with real-time, demand-based heuristics:

- The compute fabric can support up to 1,000 gigabytes per second (GBps), which is 100 GBps per core or per cluster and connects the cores and graphics through the last level cache to the memory
    - Features high dynamic frequency range and is capable of dynamically selecting the data path for latency versus bandwidth optimization based on actual fabric loads
    - Dynamically adjusts the last-level cache policy – inclusive or non-inclusive – based on utilization

- The I/O fabric supports up to 64 GBps, connecting the different types of I/Os as well as internal devices and can change speed seamlessly without interfering with a device's normal operation, selecting the fabric speed to match the required amount of data transfer
- The memory fabric can deliver up to 204 GBps of data and dynamically scale its bus width and speed to support multiple operating points for high bandwidth, low latency or low power

Intel Thread Director

In order for Performance-cores and Efficient-cores to work seamlessly with the operating system, Intel has developed an improved scheduling technology called Intel Thread Director. Built directly into the hardware, Thread Director provides low-level telemetry on the state of the core and the instruction mix of the thread, empowering the operating system to place the right thread on the right core at the right time. Thread Director is dynamic and adaptive – adjusting scheduling decisions to real-time compute needs – rather than a simple, static rules-based approach.

Traditionally, the operating system would make decisions based on limited available stats, such as foreground and background tasks. Thread Director adds a new dimension by:

- Using hardware telemetry to direct threads that require higher performance to the right Performance-core at that moment
- Monitoring instruction mix, state of the core and other relevant microarchitecture telemetry at a granular level, which helps the operating system make more intelligent scheduling decisions
- Optimizing Thread Director for the best performance on Windows 11 through collaboration with Microsoft
- Extending the PowerThrottling API, which allows developers to explicitly specify quality-of-service attributes for their threads
- Applying a new EcoQoS classification that informs the scheduler if the thread prefers power efficiency (such threads get scheduled on Efficient-cores)

X^{e} HPG Microarchitecture and Alchemist SoCs

X^{e} HPG is a new discrete graphics microarchitecture designed to scale to enthusiast-class performance for gaming and creation workloads. The X^{e} HPG microarchitecture powers the Alchemist family of SoCs, and the first related products are coming to market in the first quarter of 2022 under the Intel® Arc™ brand. The X^{e} HPG microarchitecture features a new X^{e}-core, a compute-focused, programmable and scalable element.

The client graphics roadmap includes Alchemist (previously known as DG2), Battlemage, Celestial and Druid SoCs. During the presentation, Intel provided microarchitectural details and shared demos running on a pre-production Alchemist SoCs, showing real gameplay, an Unreal Engine 5 health test and a new neural-based super sampling technology called X^{e}SS.

Alchemist SoCs, based on the Xᵉ HPG microarchitecture, are engineered to deliver great scalability and compute efficiency with key architectural features:

- Up to eight render slices with fixed function designed for DirectX 12 Ultimate
- New Xᵉ-cores with 16 vector engines and 16 matrix engines (referred to as XMX – Xᵉ Matrix eXtensions), cache and shared local memory
- New ray tracing units with support for DirectX Raytracing (DXR) and Vulkan Ray Tracing
- 1.5x frequency uplift and 1.5x performance/watt improvement compared with Xᵉ LP microarchitecture through a combination of architecture, logic design, circuit design, process technology and software optimizations[1]
- Manufactured on TSMC's N6 process node

Central to Intel's graphics efforts is a software-first approach:

- The Xᵉ architecture is being engineered in close collaboration with developers, driving alignment to industry standards
- Intel's first high-performance gaming graphics processing unit (GPU) prioritizes performance and quality through a driver design that covers integrated and discrete graphics products in one unified codebase
- Intel has completed a re-architecture of core graphics driver components, specifically the memory manager and compiler, resulting in improved throughput for CPU-bound titles by 15% (and as much as 80%) and improved game load times by 25%

XᵉSS

XᵉSS takes advantage of Alchemist's built in XMX AI acceleration to deliver a novel upscaling technology that enables high-performance and high-fidelity visuals. It uses deep learning to synthesize images that are close to the quality of native high-resolution rendering. With XᵉSS, games that would only be playable at lower quality settings or lower resolutions can run smoothly at higher quality settings and resolutions.

- XᵉSS works by reconstructing subpixel details from neighboring pixels, as well as motion-compensated previous frames
- Reconstruction is performed by a neural network trained to deliver high performance and great quality, with up to a 2x performance boost[1]
- XᵉSS delivers AI-based super sampling on a broad set of hardware, including integrated graphics, by leveraging the DP4a instruction set
- Several early game developers are engaged on XᵉSS, and the SDK for the initial XMX version will be available for ISVs this month, with the DP4a version available later this year

intel FACT SHEET

Data Center

Next-Generation Intel Xeon Scalable Processor (code-named "Sapphire Rapids")
Sapphire Rapids represents Intel's biggest data center platform advancement. The processor delivers substantial compute performance across dynamic and increasingly demanding data center usages and is workload-optimized to deliver high performance on elastic compute models like cloud, microservices and AI.

At the heart of Sapphire Rapids is a tiled, modular SoC architecture that leverages Intel's embedded multi-die interconnect bridge (EMIB) packaging technology to deliver significant scalability while maintaining the benefits of a monolithic CPU interface. Sapphire Rapids provides a single balanced unified memory access architecture, with every thread having full access to all resources on all tiles, including caches, memory and I/O. The result offers consistent low-latency and high cross-section bandwidth across the entire SoC.

Sapphire Rapids is built on Intel 7 process technology and features Intel's new Performance-core microarchitecture, which is designed for speed and pushes the limits of low-latency and single-threaded application performance.

Sapphire Rapids delivers the industry's broadest range of data center-relevant accelerators, including new instruction set architecture and integrated IP to increase performance across the broadest range of customer workloads and usages. The new built-in acceleration engines include:

- **Intel® Accelerator Interfacing Architecture (AIA)** – Supports efficient dispatch, synchronization and signaling to accelerators and devices
- **Intel® Advanced Matrix Extensions (AMX)** – A new workload acceleration engine introduced in Sapphire Rapids that delivers massive speed-up to the tensor processing at the heart of deep learning algorithms. It can provide an increase in computing capabilities with 2K INT8 and 1K BFP16 operations per cycle. Using early Sapphire Rapids silicon, optimized internal matrix-multiply micro benchmarks run over 7x faster using new Intel AMX instruction set extensions compared to a version of the same micro benchmark using Intel AVX-512 VNNI instructions, delivering substantial performance gains across AI workloads for both training and inference
- **Intel® Data Streaming Accelerator (DSA)** – Designed to offload the most common data movement tasks that cause the overhead seen in data center scale deployments. Intel DSA improves processing of these overhead tasks to deliver increased overall workload performance and can move data among CPU, memory and caches, as well as all attached memory, storage and network devices

These architectural advancements enable Sapphire Rapids to deliver great out-of-the-box performance for the broadest range of workloads and deployment models in the cloud, data center, network and intelligent

edge. The processor is built to drive industry technology transitions with advanced memory and next-generation I/O, including PCIe 5.0, CXL 1.1, DDR5 and HBM technologies.

Infrastructure Processing Unit (IPU)
The IPU is a programmable networking device designed to enable cloud and communication service providers to reduce overhead and free up performance for CPUs.

Intel's IPU-based architecture has several major advantages:

- The strong separation of infrastructure functions and tenant workload allows tenants to take full control of the CPU
- The cloud operator can offload infrastructure tasks to the IPU, maximizing CPU utilization and revenue
- IPUs can manage storage traffic, which reduces latency while efficiently using storage capacity via a diskless server architecture. With an IPU, customers can better utilize resources with a secure, programmable and stable solution that enables them to balance processing and storage

Recognizing "one-size-does-not-fit-all," Intel offered a deeper look at its IPU architecture and introduced the following new members of the IPU family – all designed to address the complexity of diverse and dispersed data centers.

Mount Evans is Intel's first ASIC IPU. Mount Evans has been architected and developed hand-in-hand with a top cloud service provider and integrates learnings from multiple generations of FPGA SmartNICs.

- Hyperscale-ready, it offers high-performance network and storage virtualization offload while maintaining a high degree of control
- Provides a best-in-class programmable packet processing engine enabling use cases like firewalls and virtual routing
- Implements a hardware accelerated NVMe storage interface scaled up from Intel Optane technology to emulate NVMe devices
- Deploys advanced crypto and compression acceleration, leveraging high-performance Intel® Quick Assist technology
- Can be programmed using existing, commonly deployed software environments, including DPDK, SPDK; and the pipeline can be configured utilizing P4 programming language pioneered by Intel's Barefoot Switch Division

Oak Springs Canyon is an IPU platform built with the Intel® Xeon D and the Intel® Agilex™ FPGA, the industry's leading FPGA in power, efficiency and performance, to:

- Offload network virtualization functions like open virtual switch (OVS) and storage functions like NVMe over fabric and RoCE v2, and provide a hardened crypto block providing a more secure, high-speed 2x 100 gigabit Ethernet network interface
- Enable Intel's partners and customers to customize their solutions with Intel Open FPGA Stack, a scalable, source-accessible software and hardware infrastructure
- Be programmed using existing, commonly deployed software environments, including DPDK and SPDK, which have been optimized on x86

The Intel N6000 Acceleration Development Platform, code-named "Arrow Creek," is a SmartNIC designed for use with Xeon-based servers. It features:

- Intel's Agilex FPGA, the industry's leading FPGA in power, efficiency and performance; Intel Ethernet 800 Series controller for high-performance 100 gigabit network acceleration
- Support for several infrastructure workloads enabling communication service providers (CoSPs) to offer flexible accelerated workloads like Juniper Contrail, OVS and SRv6, building upon the success of Intel's PAC-N3000, which is already deployed in some of the world's top CoSPs

X^e HPC and Ponte Vecchio

Ponte Vecchio, based on the X^e HPC microarchitecture, delivers industry-leading FLOPs and compute density to accelerate AI, high performance computing (HPC), and advanced analytics workloads. Intel disclosed IP block information of the X^e HPC microarchitecture; including eight Vector and Matrix engines (referred to as XMX – X^e Matrix eXtensions) per X^e-core; slice and stack information; and tile information including process nodes for the Compute, Base, and X^e Link tiles. At Architecture Day, Intel showed that early Ponte Vecchio silicon is demonstrating leadership performance, setting an industry-record in both inference and training throughput on a popular AI benchmark.[1] Intel's A0 silicon performance is providing greater than 45 TFLOPS FP32 throughput, greater than 5 TBps memory fabric bandwidth and greater than 2 TBps connectivity bandwidth. Intel also shared a demo showing ResNet inference performance of over 43,000 images per second and greater than 3,400 images per second with ResNet training, both of which are on track to deliver performance leadership.[1]

Ponte Vecchio is comprised of several complex designs that manifest in tiles, which are then assembled through an EMIB tile that enables a low-power, high-speed connection between the tiles. These are put together in Foveros packaging that creates the 3D stacking of active silicon for power and interconnect density. A high-speed MDFI interconnect allows scaling from one to two stacks.

Compute Tile is a dense package of X^e-cores and is the heart of Ponte Vecchio.

- One tile has eight X^e-cores with a total of 4MB L1 cache, its key to delivering power-efficient compute
- Built on TSMC's most advanced process technology, N5

- Intel has paved the way with the design infrastructure set-up and tools flows, and methodology to be able to test and verify tiles for this node
- The tile has an extremely tight 36-micron bump pitch for 3D stacking with Foveros

Base Tile is the connective tissue of Ponte Vecchio. It is a large die built on Intel 7 optimized for Foveros technology.

- The Base Tile is where all the complex I/O and high bandwidth components come together with the SoC infrastructure – PCIe Gen5, HBM2e memory, MDFI links to connect tile-to-tile and EMIB bridges
- Super-high-bandwidth 3D connect with high 2D interconnect and low latency makes this an infinite connectivity machine
- The Intel technology development team worked to match the requirements on bandwidth, bump pitch and signal integrity

Xe Link Tile provides the connectivity between GPUs supporting eight links per tile.

- Critical for scale-up for HPC and AI
- Targeting the fastest SerDes supported at Intel – up to 90G
- This tile was added to enable the scale-up solution for the Aurora exascale supercomputer

Ponte Vecchio is powered on, is in validation and has begun limited sampling to customers. Ponte Vecchio will be released in 2022 for HPC and AI markets.

oneAPI

The oneAPI industry initiative provides an open, standards-based unified software stack that is cross-architecture and cross-vendor, allowing developers to break free from proprietary languages and programming models. There are now Data Parallel C++ (DPC++) and oneAPI library implementations for Nvidia GPUs, AMD GPUs and Arm CPUs. oneAPI is being adopted broadly by independent software vendors (ISVs), operating system vendors, end users and academics. Key industry leaders are helping to evolve the specification to support additional use cases and architectures. Intel also has a commercial product offering that includes the foundational oneAPI base toolkit, which adds compilers, analyzers, debuggers and porting tools beyond the spec language and libraries.

oneAPI delivers compatibility across architectures, improving developer productivity and innovation:

- There are more than 200,000 unique installs of Intel's oneAPI toolkits
- More than 300 applications deployed in the market use oneAPI's unified programming model
- More than 80 HPC and AI applications are functional on the Xe HPC microarchitecture using Intel oneAPI Toolkits

- The provisional version 1.1 spec released in May adds new graph interfaces for deep learning workloads and advanced ray tracing libraries, and is expected to be finalized by the end of the year

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

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 www.Intel.com/ArchDay21claims for configuration details. No product or component can be absolutely secure.

All product plans and roadmaps are subject to change without notice. Results that are based on pre-production systems and components as well as results that have been estimated or simulated using an Intel Reference Platform (an internal example new system), internal Intel analysis or architecture simulation or modeling are provided to you for informational purposes only. Results may vary based on future changes to any systems, components, specifications, or configurations. Intel technologies may require enabled hardware, software or service activation.

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.

Statements that refer to future plans and expectations are forward-looking statements that involve a number of risks and uncertainties. Words such as "anticipates," "expects," "intends," "goals," "plans," "believes," "seeks," "estimates," "continues," "may," "will," "would," "should," "could," and variations of such words and similar expressions are intended to identify such forward-looking statements. Statements that refer to or are based on estimates, forecasts, projections, uncertain events or assumptions, including statements relating to future products and technology and the expected availability and benefits of such products and technology, market opportunity, and anticipated trends in our businesses or the markets relevant to them, also identify forward-looking statements. Such statements are based on management's current expectations and involve many risks and uncertainties that could cause actual results to differ materially from those expressed or implied in these forward-looking statements. Important factors that could cause actual results to differ materially from the company's expectations are set forth in Intel's reports filed or furnished with the Securities and Exchange Commission (SEC), including Intel's most recent reports on Form 10-K and Form 10-Q, available at Intel's investor relations website at www.intc.com and the SEC's website at www.sec.gov. Intel does not undertake, and expressly disclaims any duty, to update any statement made in this presentation, whether as a result of new information, new developments or otherwise, except to the extent that disclosure may be required by law.

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

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.