

NEW TECHNOLOGIES FOR IMPROVED COMPUTING WITH INTEL PROCESSORS

MARIUS CORNEA, INTEL CORPORATION
ARITH-26

JUN 10, 2019

Intel Turned 50 on July 18th, 2018!

Timeline Highlights

1968: Founded on July 18, by Robert Noyce and Gordon Moore - NM Electronics, renamed Intel Corporation

1969: Intel releases the [3101 static random access memory](#) (SRAM) in April, its first product

1969: Intel launches the [1101](#), the first metal oxide semiconductor (MOS) static RAM

1971: Intel releases the 4-bit 4004, the first microprocessor

1971: Intel introduces [erasable programmable read-only memory](#) (EPROM)

1972: Intel opens first international manufacturing facility in Penang, Malaysia

1973: Intel opens wafer fabrication facility in Livermore, California – its first outside Silicon Valley

1974: Intel releases the 8-bit [8080 microprocessor](#)

1976: Intel debuts the [MCS-48 microcontroller family](#)

1978: Intel releases the [8086 processor](#), the first 16-bit processor & first based on the x86 architecture

1981: IBM selects Intel's 8088 microprocessor for the IBM PC

1982: Intel launches the first 286 processor, the 16-bit 80286

1985: Intel introduces the 386 processor, a 32-bit chip that runs multiple software programs at once

1991: The Intel Inside marketing campaign begins

1993: Intel introduces the Pentium processor

1995: Intel collaborates on Universal Serial Bus (USB) specification, which provides a standard peripheral-to-PC connection

2003: Intel releases the Centrino processor, integrating a mobile processor, related chipsets and 802.11 wireless network functions

2007: Intel produces processors that use [45-nanometer transistors](#)

2011: Intel announces the [Ultrabook laptop specification](#), its first new device category specification for PC manufacturers

2016: Intel restructures from a PC-centric company to a data-focused company

2017: Intel develops chips that use 10-nm transistors



New Technology Focus Areas

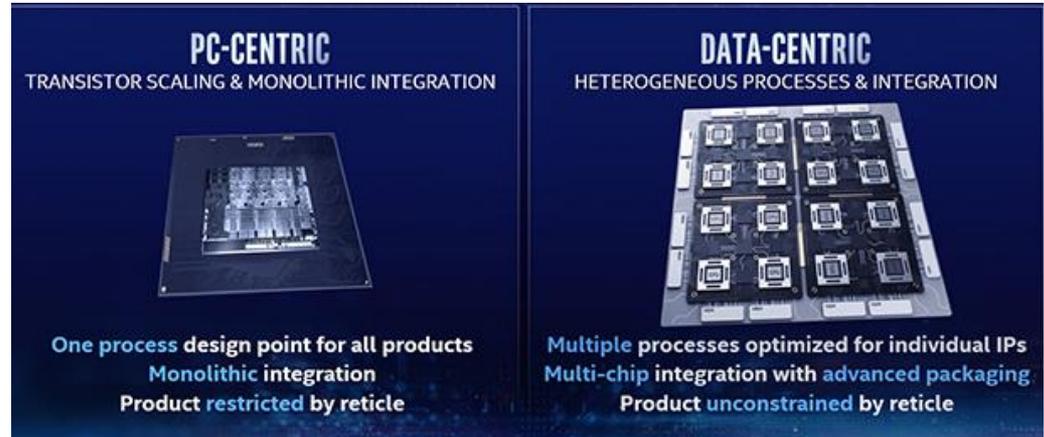
For the past 50 years, Intel focused on process and architecture

Now, Intel is transforming from a PC-centric to a data-centric company

- CPUs are still very important, but we emphasize workload-optimized platforms and effortless customer & developer innovation

Current focus areas:

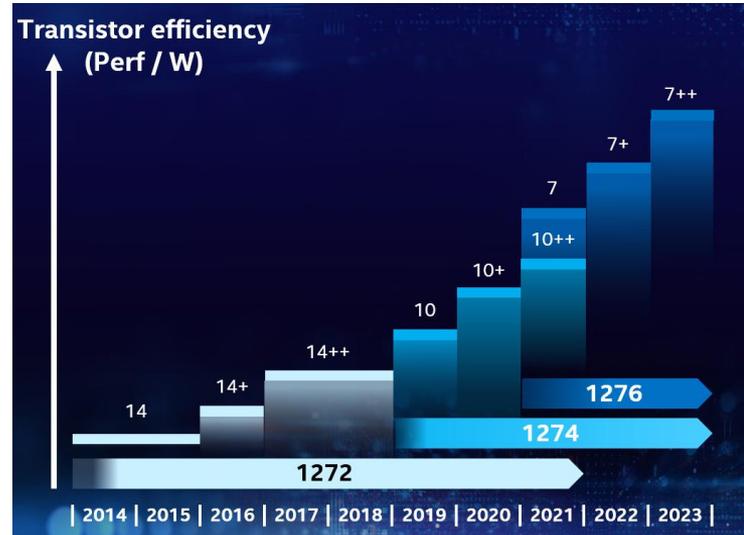
- Manufacturing Processes
- Processor and System Architecture
- Memory
- Interconnect
- Security
- Software, and New Compute Instructions (includes much Computer Arithmetic!)



Intel's goal is to supply a mix of scalar, vector, matrix, and spatial architectures

Manufacturing Processes

- Transistor efficiency is increasing over time
- Intel 10nm CPUs began to ship in June 2019
 - 1st volume 10nm platform is “Ice Lake”
 - Other 10 nm products will follow in 2020-21
- 7nm planned for 2021
 - 2 times scaling, approximately 20 percent increase in performance per watt, and 4 times reduction in design rule complexity versus 10 nm
 - Intel’s first commercial use of extreme ultraviolet (EUV) lithography - will help drive scaling for multiple node generations
 - The lead 7nm product is expected to be an Intel X^e architecture-based, general-purpose discrete GPU for data center AI and high-performance computing



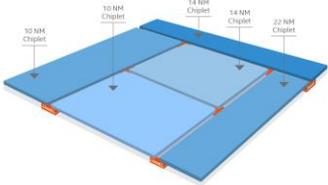
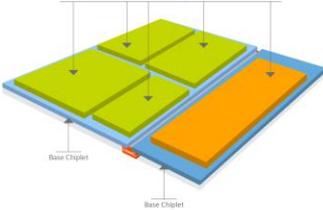
Manufacturing Processes

- “Foveros”: a new 3D packaging technology – uses 3D stacking to enable logic-on-logic integration
- Follows the Embedded Multi-die Interconnect Bridge (EMIB) 2D packaging technology of 2018
- Devices and systems can combine high-performance, high-density and low-power silicon process technologies
- Provides flexibility, to “mix and match” technology IP blocks w/ various memory & I/O elements in new device form factors; products can be broken up into smaller, stacked “chiplets”
- Planning to launch a range of products using Foveros beginning in the second half of 2019

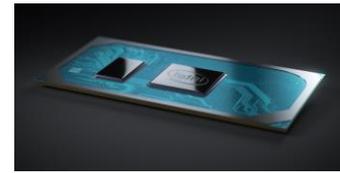


2D AND 3D PACKAGING DRIVE NEW DESIGN FLEXIBILITY

The combination of advanced 2D and 3D packaging technologies allows Intel to flexibly combine smaller chiplets of IP to meet the demands of a huge range of applications, power envelopes, and form factors. Intel® embedded multi-die interconnect bridge (EMIB) and Foveros are advanced 2D and 3D packaging technologies, delivering high performance at low cost.

MONOLITHIC	2D INTEGRATION	3D INTEGRATION
Integrate functions on a single die for high performance on a single silicon technology	Combine IPs built with separate processes into a single package with Intel EMIB, helping improve yield, cost, time-to-market, and total capability	All the benefits of 2D integration plus a new level of density thanks to Foveros, allowing for a radical re-architecture of systems-on-chips
		

Processor and System Architecture



- “Cascade Lake” (now shipping)
 - Intel Xeon Scalable processor, in 14nm technology
 - Introduced Intel Optane DC persistent memory and a set of new AI features (Intel DL Boost – AVX512-VNNI speeds up DL inference); this embedded AI accelerator will speed deep learning inference.
- “Coffee Lake” – 9th Gen Intel Core i9-9900KS special edition desktop processor – the first to feature all 8 cores running at a turbo frequency of 5.0 GHz
- “Ice Lake”: 10 nm process (now shipping) also adds architecture innovations
 - Expected to deliver approximately 3x faster wireless speeds, 2x faster video transcode speeds, 2x faster graphics performance, and 2.5x to 3x faster AI performance over previous generation products
 - New integrated graphics unit, “Gen 11”, with at least 1 TF for 32-bit floating-point operations, and more than twice the performance-per-clock of its predecessor – at least 64 EUs versus 24 EUs in “Gen 9”



Processor and System Architecture

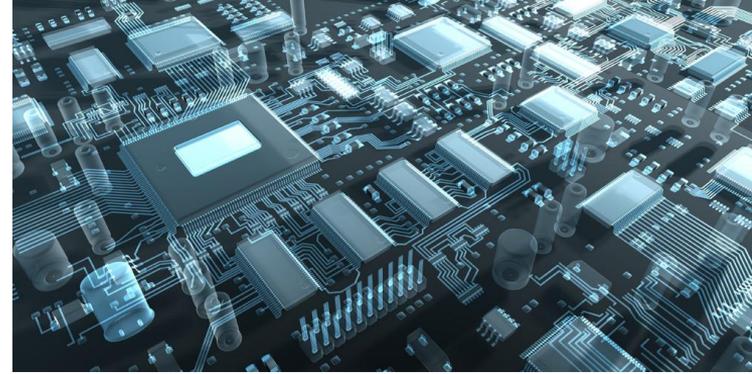
- “Lakefield”
 - New client platform, code-named “Lakefield”, with a hybrid CPU
 - Featuring the first iteration of the Foveros 3D packaging technology - combines different pieces of IP that might have previously been discrete into a single product with a smaller motherboard footprint
 - Expected to be in production before the end of 2019
- “Cooper Lake”: upcoming 14nm server CPU, expected in 2020
 - Performance improvements, new hardware-enhanced security features, new I/O features, new Intel® DL Boost capabilities (bfloat16) for AI/DL training performance, and additional Intel Optane DC innovations
- “Tiger Lake” – future big core CPU
- “Tremont” – future Atom CPU

Memory



- Memory hierarchy – optimized by capacity, latency, bandwidth, cost, and other features
- Persistent memory and high-bandwidth memory were introduced relatively recently in Intel systems, influencing the way the memory system is structured
- Intel Optane memory H10 with solid state storage combines the responsiveness of the Intel Optane memory technology, with the storage capacity of Intel® Quad Level Cell (QLC) 3D NAND technology in a space-saver form factor
 - High-speed and large SSD storage capacity on a single drive
 - Faster access to frequently used applications and files; better responsiveness
- Intel® Optane™ M10/M15 - a new generation of Intel Optane memory, available by ~Sep 2019, with higher performance and lower power consumption compared to the previous generation
 - Leads to shorter boot times, fast application launches, and fast gaming and browsing

Interconnect



- A consortium was founded, to develop Compute Express Link Technology (CXL) – will improve performance and remove bottlenecks in computation-intensive workloads
- Processing data in emerging applications requires a diverse mix of scalar, vector, matrix and spatial architectures deployed in CPU, GPU, FPGA, networking and other accelerators
- In March, Intel contributed the Thunderbolt protocol to the USB Promoter Group, enabling other chipmakers to build Thunderbolt-compatible silicon devices under the new USB4 specification

Security

- Ice Lake will support several new instructions
 - Galois Field New Instructions (GFNI for SSE, AVX and AVX512)
 - 'EVEX' and 'VEX' versions of AES and PCLMULQDQ
 - Intel Secure Guard Extensions (SGX)
 - PCONFIG (Platform Configuration, for MKTME – multiple key total memory encryption)
- Tiger Lake will support
 - Control-flow Enforcement Technology (CET)
- Tremont will support
 - Galois Field New Instructions (GFNI for SSE)
 - Intel Secure Guard Extensions (SGX)

Software

- “One API” software project
 - Intel announced the “One API” project to simplify the programming of diverse computing engines across CPU, GPU, FPGA, AI and other accelerators; it includes a comprehensive and unified portfolio of developer tools for mapping software to the hardware that can best accelerate the code; a public project release is expected to be available in 2019.
- Intel® OpenVINO™ Toolkit - Open Visual Inference & Neural Network Optimization Toolkit
 - It extends workloads across Intel® hardware (including accelerators) and maximizes performance.
 - Enables CNN-based deep learning inference at the edge
 - Supports heterogeneous execution across computer vision accelerators—CPU, GPU, Intel® Movidius™ Neural Compute Stick, and FPGA—using a common API

Software

- Open-Source Deep Learning Reference Stack (Open Source)
 - An integrated, highly-performant open source stack optimized for Intel® Xeon® Scalable Processors
 - Includes Intel® Deep Learning Boost (Intel DL Boost), and is designed to accelerate AI use cases such as image recognition, object detection, speech recognition and language translation
- Data Analytics Reference Stack (Open Source)
 - Developed to help enterprises analyze, classify, recognize and process large amounts of data built on Intel® Xeon® Scalable platforms using Apache Hadoop and Apache Spark™

New Compute Instructions in Intel® Architecture Processors

- Cooper Lake (CPX)
 - AVX512_VNNI: Vector Neural Network Instructions
 - AVX512_BF16CVT: Conversions of Single Precision Floating-Point Values to BF16
 - AVX512_VDPBF16PS: Dot Product of BF16 Pairs, Accumulated into Packed Single Precision
- Ice Lake
 - AVX512_BITALG and AVX512_VPOPCNTDQ : Vector Bit Algebra Instructions – Byte and Word Pop Count, and Shuffle; Double and Quad Word Pop Count
 - AVX512_VBMI2: Vector Bit Manipulation Instructions – Compress, Expand, Logical Shift Left, Logical Shift Right

Exascale Computing



- Intel and Cray Inc. (sub-contractor) plan to deliver the Aurora supercomputer, first US exascale system, to the Argonne National Laboratory in 2021
- ExaFLOP performance: 10^{18} , a “quintillion” floating-point operations per second
- For both traditional HPC and AI applications, it will include
 - A future generation of the Intel® Xeon® Scalable processor
 - Intel’s X^e compute architecture
 - A future generation of Intel® Optane™ DC Persistent Memory
 - The Intel One API software
 - Cray’s next-generation supercomputer system, code-named “Shasta,” w/ more than 200 cabinets; Cray’s Slingshot™ high-performance scalable interconnect; the Shasta software stack optimized for Intel architecture

Other Technology Highlights

- Intel® Movidius™ Myriad™ X VPU for high-performance, low-power AI workloads – powers the Intel® Neural Compute Stick 2 (Intel® NCS 2)
- Intel® Nervana™ NNP-I (AI inference processor), a general-purpose GPU; joins Intel® Nervana™ NNP-L for training
- FPGAs: Intel® Agilex™ family of FPGAs, introduced in April 2019 – a family of field programmable gate arrays for data-centric business applications for embedded, network and data center areas
- 5G-ready network system-on-chip (SoC)
- Mobileye - Autonomous Driving
- Quantum Computing

References

- Intel® 64 and IA-32 Architectures Software Developer Manuals: <https://software.intel.com/en-us/articles/intel-sdm>
- Intel® Architecture ISA Extensions: <https://software.intel.com/en-us/isa-extensions>
- Intel® Newsroom: <https://newsroom.intel.com/>

NOTICES AND DISCLAIMERS

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration.

No computer system can be absolutely secure.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. For more complete information about performance and benchmark results, visit <http://www.intel.com/benchmarks>.

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit <http://www.intel.com/benchmarks>.

Intel® Advanced Vector Extensions (Intel® AVX)* provides higher throughput to certain processor operations. Due to varying processor power characteristics, utilizing AVX instructions may cause a) some parts to operate at less than the rated frequency and b) some parts with Intel® Turbo Boost Technology 2.0 to not achieve any or maximum turbo frequencies. Performance varies depending on hardware, software, and system configuration and you can learn more at <http://www.intel.com/go/turbo>.

Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

Cost reduction scenarios described are intended as examples of how a given Intel-based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate. Intel, the Intel logo, and Intel Xeon are trademarks of Intel Corporation in the U.S. and/or other countries.

*Other names and brands may be claimed as property of others.

© 2019 Intel Corporation.

