Industry Solution Brief

Manufacturing

Intel® HPC+AI Technology Accelerates Digital Prototyping in Manufacturing

Intel technologies and industry collaboration help improve and accelerate designs and reduce TCO

Table of Contents

Executive Summary

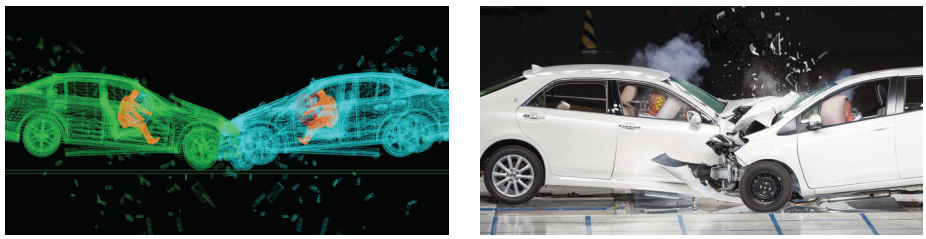
Digital prototyping through simulation and modeling on High Performance Computing (HPC) systems has changed large-scale manufacturing, such as in aerospace, automobiles, and robotics. It has also advanced small-scale manufacturing processes and designs. Companies work smarter when a design can be optimized through simulation and modeling rather than physical prototypes.

> *The less an aircraft company has to do in a wind tunnel with physical airplanes, the less they spend. HPC is expensive but building and testing real airplanes is much, much more expensive.*
>
> —Ken Lloyd, Intel

Simulation and modeling cover a variety of workloads, from Computational Fluid Dynamics (CFD) to Multiphysics, structural engineering, and crash simulation. The larger the object, the more computation required, but to stay competitive in an industry, the answers need to be delivered quickly. Thus, HPC systems—on-premises and in the cloud—are critical to infrastructure for large-scale digital prototyping and complex designs.

Additionally, both design and smart manufacturing embrace Artificial Intelligence (AI), machine learning (ML) deep learning (DL), and visualization. These advanced technologies are driving new methods of solving design problems and modernizing the manufacturing space, making it faster and more efficient.

Working with manufacturers, engineering domain experts, commercial independent software vendors (ISVs), and the open source community, Intel is helping solve the challenges in manufacturing. From enhancing silicon to optimizing software, Intel engineers, architects, and software scientists are helping to create and build safer products faster.



Using HPC to simulate crash testing verses using real automobiles saves manufacturers money and time.

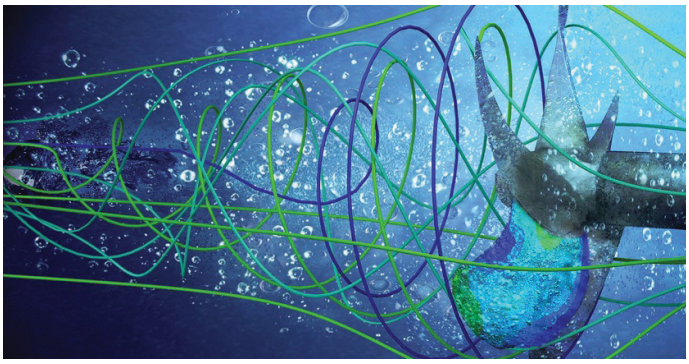
Challenges

Computational challenges begin with more data. The ability to more easily acquire large quantities of measurements from widespread IoT infrastructure, discrete sensors, and mobile devices gives engineers more data to process. The growing demand for deeper insight into a design requires higher resolution simulations, resulting in even more data. And autonomous systems require massive data for training.

In addition to general storage and throughput challenges with expanding datasets, each type of application makes its own demands on HPC infrastructure. Those demands must be addressed to optimize computing for manufacturing.

Memory Bandwidth—In aircraft and automobile design, Computational Fluid Dynamics (CFD) is one of the major HPC workloads engineers run. CFD is typically constrained by memory bandwidth.

With CFD problems, flow features scale matters. The rudder or wing of a Boeing 787 or Airbus A380 require greater detail and challenging flow regime (much larger dataset) and thus a bigger CFD problem than flow within a pipe for an industrial application. Some CFD problems, with billions of degrees of freedom, require Petascale and Exascale computing with large memory bandwidth to arrive at solutions in a timely manner.



ISVs like Cadence provide CFD software for a variety of disciplines including naval architects and marine engineers.

Memory Capacity—Building ships, airframes, buildings, automobiles, and smaller items requires solving structural engineering problems. These problems involve not just the design of a product, but also external forces, such as from earthquakes, waves, and other vehicles—all as simulated data. Again, the bigger the design and environment in which it is used, the more data to store and move around the infrastructure, especially when finer detail is desired.

Compute Capacity—With robotics and autonomous vehicles and systems, massive amounts of data gathering, analysis, labeling, and computing require large-scale HPC+AI infrastructure to accommodate model training. Additionally, as part of the Industry 4.0 vision, manufacturers are replacing manual operations with robots. They're also replacing sensor-based detection systems by training visual computing algorithms to detect faults in production lines using live cameras. To fully realize Industry 4.0, more computing is needed from the HPC+AI converged datacenter—both on-premises and in the cloud. At the edge, high-performance inference of streaming data requires low-power, compact higher-capacity computing.

Software Heterogeneity—Application code complexity and computational demand are as varied as the problems they address. Programs are a mixed bag of embarrassingly parallel to highly serial codes from commercial vendors to open source communities, written in various frameworks and languages. Optimizing these to deliver answers as fast as possible requires developers with domain expertise and technical knowledge and experience with the underlying hardware and software architecture.

Solutions

Intel scientists, engineers, and software developers collaborate with leading manufacturing enterprises,

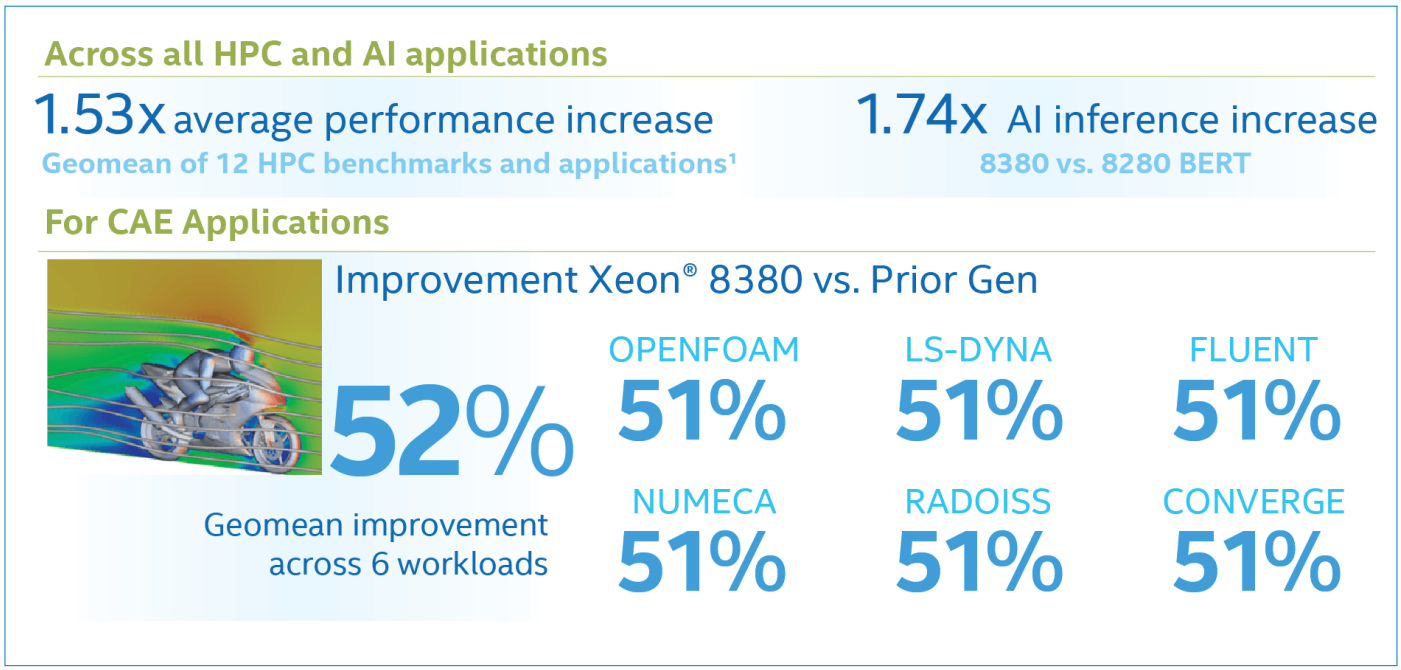


Figure 1. 3rd Generation Intel® Xeon® Scalable processors deliver significant performance gains over previous generation.[1]

commercial software companies, cloud service providers, and the open source community to solve manufacturing's biggest challenges. Working with these communities, Intel is helping industry innovate with new methodologies, next-generation technologies, and optimized software.

Memory Bandwidth/Capacity

HPC solutions built for CFD problems are designed to move data quickly through calculations. These solutions often involve high-core-count processors with large memory bandwidth and/or integrate specialized high-bandwidth memory (HBM). The Intel® Xeon® Scalable processor family offers a variety of CPU configurations with HBM to address today's CFD application needs.

Next-generation Intel® Xeon® Scalable Processors

In addition to HBM, next-generation Intel Xeon Scalable processors (codenamed Sapphire Rapids) will also support DDR5 DRAM and 3rd generation Intel® Optane™ persistent memory (codenamed Crow Pass). These options will provide flexible solutions that are both highly scalable and fast for bandwidth-bound workloads, accelerating calculations and speeding results.

3rd Generation Intel® Xeon® Scalable Processors

3rd Generation Intel® Xeon® Scalable processors are the latest generation of the datacenter processor family designed for HPC workloads. These processors deliver significant performance gains over previous generation Intel Xeon Scalable processors for many applications used in HPC, AI, and manufacturing (Figure 1).

Intel® Xeon® 9200 Processors

The Intel Xeon 9200 processor family was designed to deliver the highest memory bandwidth with greatest possible capacity and maximum number of cores in 2nd Generation Intel Xeon Scalable processors. This CPU family, offering up to 56 cores per CPU (112 cores in dual-socket servers), supports 12 DDR4 memory channels. Adding Intel Optane persistent memory, two-socket servers can be built with up to 12 TB when combined with DRAM.

Accelerating Data

The challenges in large computing begin with the scale of data. It extends into how the workload will use it—store it, move it, operate on it. For example, some large-scale problems that use massive datasets and in-memory type of computing require large memory capacity close to the processor. Intel Optane technologies help address data challenges across the hardware platform. Solutions gain greater advantage when combined with other advanced I/O technologies.

Intel® Optane™ SSDs are fast, high-capacity storage devices that combine attributes of memory (speed) and storage (density) to accelerate data movement from the storage tier to the memory tier.

Intel® Optane™ persistent memory (Intel® Optane™ PMem) is flexible memory technology that offers high capacity, affordability, and persistence of data on the memory bus. If DRAM alone is not enough, Intel Optane PMem enables up to 36 TB of memory capacity (with DRAM) in multi-socket servers. The larger amounts of memory keep calculations processing instead of moving between memory and storage.

Distributed Asynchronous Object Storage (DAOS) is the foundation of the Intel roadmap for Exascale HPC storage. It overcomes limitations of traditional distributed storage solutions. DAOS is an open source software-defined, scale-out object store that is designed to use low-latency, high-message-rate user space communications that bypass the operating system. When built with a combination of Intel Optane PMem, Intel Optane SSDs, and other 3D NAND storage, DAOS delivers high-bandwidth, low-latency, and high-input/output operations per second (IOPS) containers to HPC applications. The combination gives DAOS implementations the ability to massively scale storage in a cost-effective and operationally efficient manner.

High-bandwidth memory (HBM), to be introduced in versions of next-generation Intel® processors, reduces time to a solution by moving data faster during computation.

Additionally, the latest generations of Intel Xeon Scalable processors integrate the most recent and highest performing I/O technologies, such as PCIe 4.0. Faster Intel® Quick Path Interconnect keeps cores busy. And Intel® networking technologies enhance performance of data movement over the network.

Optimizing Workloads

Intel developers work directly with both ISVs and the open source community to optimize applications for the highest performance on Intel® architecture and technologies.

Intel Optimization Tools

Optimizations are built on the many Intel® libraries, primitives, and language distributions that take advantage of Intel® processor technologies. These technologies, such as Intel® Advanced Vector Extensions 512 (Intel® AVX-512) and others, accelerate performance for computation,

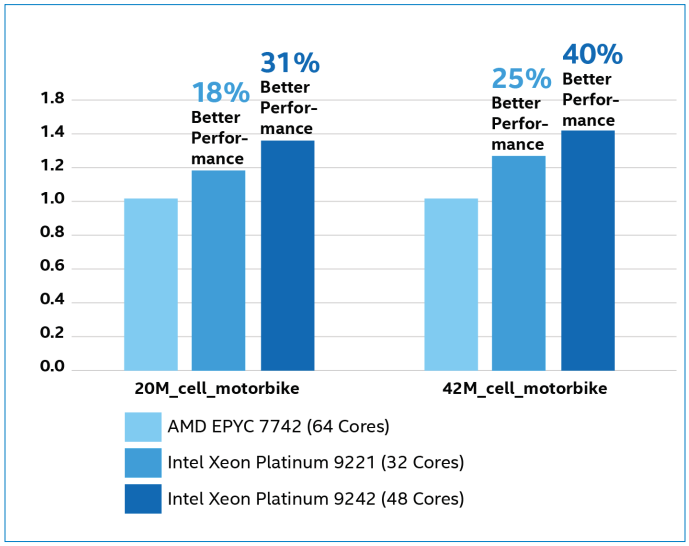


Figure 2. OpenFOAM performance on systems powered by 64-core AMD EPYC processors compared to systems powered by 32-core and 48-core Intel Xeon Scalable processors (these numbers are a geometric mean [geomean] of the testing results)

memory, and network in many applications when run on Intel architecture. Free Intel® oneAPI toolkits include the Intel libraries, optimizations, distributions, and analysis tools developers can leverage to accelerate their codes.

Applications written in Python and TensorFlow can achieve immediate performance gains with minimal code changes. Simply by employing Intel framework optimizations and language distributions, users are able to quickly see significant performance gains.

> *The lion's share of performance gains is through software optimizations available in Intel toolkits and libraries. Using these free toolkits and optimizations, developers working at and above the framework level—TensorFlow, PyTorch, and OpenVINO, for example—can enjoy performance gains quickly.*
>
> —Ed Groden, Intel

Some codes are already well tuned. Accelerating them further might require some innovation with help of Intel developers.

Open Source Contributions

Intel is a major contributor to the open source community, from the Linux operating system kernel through hundreds of applications. Intel developers have contributed significant enhancements to codes used in manufacturing that take advantage of Intel® technologies.

OpenFOAM is a key open source CFD code. OpenFOAM on Intel Xeon 9200 processor family, with 12 memory channels and using fewer cores, performs up to 40 percent better than on AMD EPYC[2] (Figure 2).

AI in Manufacturing

AI and visual computing are being increasingly employed in the industry, from large-scale operations to small product manufacturing.

AI in Automotive

Many Advanced Driver Assistance Systems (ADAS) are appearing in modern automobiles. Autonomous vehicles will be ubiquitous in the future. These self-driving solutions require large HPC+AI systems to analyze data, build algorithms, train models, and inference in the vehicle. Intel, along with its subsidiary vehicle automation company Mobileye, works closely with automobile manufacturers, such as VW, Ford, BMW, and Bentley. Collaborative efforts help the automotive industry develop technologies and algorithms and optimize processes for ADAS using HPC+AI.

AI for Smart Manufacturing

Industry 4.0 utilizes the power of AI to reduce costs while improving yield, quality, and efficiencies.

Asset Optimization—Optimizing equipment life on the manufacturing floor helps companies improve TCO. Organizations are using IoT monitoring systems and machine learning to gather and analyze equipment operations for predictive maintenance. Using AI, companies can see potential equipment malfunction before they occur.

Process Fault Detection—In manufacturing processes, defect inspection has been limited by the number of deployed sensors and specialized image analysis to detect problems. Today, manufacturing lines are employing high-definition cameras and visual computing to analyze entire processes. Visual computing is simpler to deploy and more scalable. Cameras are widely available, and algorithms can be quickly trained from data captured on the manufacturing line with coverage of much more than previous sensor-based systems. Visual computing can also be applied to safety and other compliance requirements across large-scale operations.

Safety Compliance—Improper use of or non-compliance of workers with personal protective equipment (PPE) and other types of safety hazards are key concerns in manufacturing. As with manufacturing line defect detection, strategically placed cameras and computer vision-based video analytics can reduce accidents and save lives. Real-time monitoring, analytics, and alerts/notifications allow company supervisors and managers to proactively respond to potential safety issues before they happen.

End-to-End Integration—Today's enterprise datacenters typically provide converged infrastructure for engineering and design, business operations, and manufacturing. With data and processes under a single umbrella, enterprise operations can take advantage of a wide variety of data to drive cost reductions and operational efficiencies.

For example, Bentley Motors sells one of the most expensive cars in the world—and mostly online, sight unseen. Bentley uses an end-to-end system, from the customer's design of the interior through manufacturing to ensure that the car ordered is the one built. Bentley's HPC infrastructure integrates highly accurate 3D visualization based on OSPRay rendering that is connected to their entire manufacturing system all in a single 3D pipeline.

Photorealistic rendering of a Bentley virtual showroom using OSPRay's path tracer. (Photo courtesy Bentley Motors)

Preventive Maintenance and more—Companies, such as Électricité de France (EDF), the second largest electric utility in the world, are exploring AI for predictive maintenance, consumption planning, cybersecurity, social media analysis, and more.

Intel® Technologies for Training, Inference, and Security

Intel continues to enhance its CPU families with technologies that accelerate AI operations, such as Intel® Deep Learning Boost (Intel® DL Boost) and Intel® Distribution of OpenVINO toolkit.

Intel® Deep Learning Boost (Intel® DL Boost) helps accelerate calculations that are used in training and inference. A suite of enhancements, Intel DL Boost components have been released in past generations of Intel CPUs. A new component, Intel® Advanced Matrix Extensions (Intel® AMX), will be integrated into next-generation Intel processors. Intel AMX enables GPU-like performance with a CPU on many matrix operations used in AI.

The Intel Distribution of OpenVINO™ toolkit accelerates visual computing performance by optimizing models trained with TensorFlow, PyTorch, and other frameworks. The toolkit includes a model optimizer and runtime and development tools. Built on oneAPI, the toolkit enables high-performance inference with a write once, deploy anywhere efficiency.

Where protecting data and functionality is paramount, Intel® Security technologies provide accelerated, end-to-end encryption from memory to storage. Intel® Software Guard Extensions allow use of secure enclaves that can be applied for greater protection of data and operations.

Manufacturing Case Studies

Across the manufacturing industry, Intel technologies have helped accelerate design solutions and improve operations. Intel technology-based solutions are deployed on-premises and in the cloud, offering access to accelerated HPC+AI resources for a variety of applications and business needs.

ST Microelectronics—High-frequency Simulation

The millimeter-scale, high-frequency V band (40-75 GHz) is enabling new close-proximity, point-to-point communications, especially in the unlicensed range of 57 to 71 GHz. STMicroelectronics developed its ST60 RF transceiver using Ansys HFSS 3D modeling software to study and optimize the device's design, especially for the highly sensitive antenna. Ansys works with Intel to optimize workload performance on Intel technologies, taking advantage of Intel AVX-512 and Intel® Math Kernel Library (Intel® MKL) to optimize efficiency and performance.

Intel software tools help engineers optimize their code to achieve high performance and code efficiency.

To complete the simulation and modeling of the ST60 transceiver, Ansys expanded its on-premises HPC infrastructure and added burstable capacity using the Microsoft Azure cloud. The engineering team accesses dedicated resources in the Azure cloud, using machines configured with the same Intel technologies and Ansys HFSS software as the on-premises systems.

> *With our Intel-based HPC and the performance increases of the last few years, we have been able to perform analyses that would not have been feasible before. An analysis of 512 cases that would have taken five days in June 2020, can now be done in two days. An analysis of 6,561 cases that previously took 11 days is now finished in less than a week.*[3]
>
> —Olivier Bayet, STMicroelectronics

Read the full case study.

OnScale—Multiphysics

Gaining insight across multiple physical domains for a design requires sophisticated modeling and simulation. Such high-level of modeling allows for accurate digital prototyping. For many designs, running Multiphysics studies on a workstation is not feasible, and for many companies, building large, on-premises HPC clusters are not possible. OnScale is a cloud company that offers powerful and highly scalable Multiphysics simulation services running on Google Cloud.

OnScale uses Intel software tools to compile and optimize their Multiphysics solvers for Intel architecture to achieve high performance and code efficiency. Their service is also intelligent, giving customers options on how they want to run their simulation.

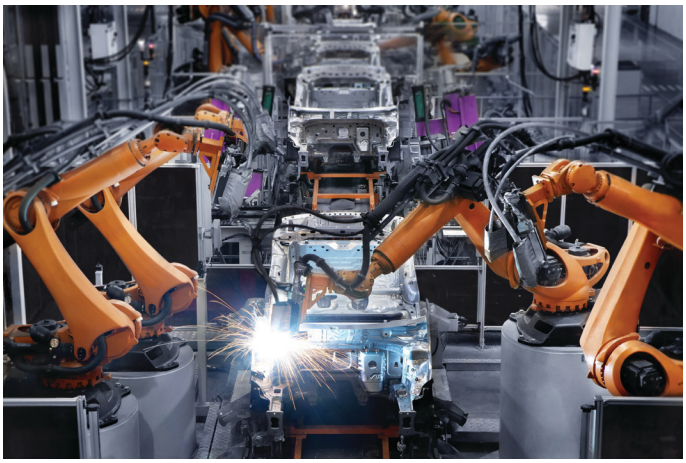
> *Our simulation platform mainly runs on compute-optimized C2 and memory-optimized M2 Google Cloud instance. We've created a machine learning engine and trained our models on one-half million simulations. That means when an engineer sets up a new simulation study, our AI can create the best cloud configuration based on his or her needs, optimizing for accuracy, cost, and runtime.*
>
> —Ian Campbell, OnScale CEO

OnScale customers have proven the accuracy of the simulation service on Intel architecture. It has been used by Polytec for non-destructive testing (NDT) analysis, and by University of Washington researchers for optical coherence elastography (OCE).

Read the full case study.

Midea and Flutura—AI-driven Industrial Inspection

Inspection/defect detection systems on manufacturing floors are full of challenges. The complex production environment, non-standardization of part features for recognition, and others result in long development cycles for inspection solutions. Midea leverages widely available and capturable data on a manufacturing floor to build computer vision solutions in its cloud-based service.



HPC enables the development and operation of robotics systems that are a key component of Industry 4.0 initiatives.

To optimize their platform, Intel worked with Midea to integrate Analytics Zoo, a unified analytics and AI platform, to build a fast and agile machine learning solution. A variety of Intel processors, including Intel Xeon Scalable processors, power the Midea computing platform.

Read the full case study.

Flutura, an Intel® Builders partner, delivers solutions to improve operations and reduce costs for the oil and gas, specialty chemicals, and OEM industries, such as PCB manufacturers. The company's solutions include computer vision-based defect detection and safety compliance monitoring and alert systems using high-definition cameras and video inferencing on trained models.

Flutura used the Intel Distribution of OpenVINO toolkit to optimize their model for PCB defect detection in manufacturing lines and for safety monitoring and compliance in processing facilities. For their PCB defect detection system, the use of an OpenVINO INT8 model improved inference time on 2nd Gen Intel Xeon processor-based systems by 2.25x with less than 0.1 percent accuracy loss.[4]

Read more: Video Analytics Safety Monitoring blog; PCB Defect Detection brief

Additional Examples

The above examples highlight recent work between Intel and the industry. Others include the following:

- Underwater robotics at University of Pisa
- Training algorithms for robots at Preferred Networks
- Redefining engineering at Royal Enfield Motorcycles
- Designing and managing nuclear power plants at Électricité de France (EDF)
- atNorth's HPC as a Service (HPCaaS) for CFD

Our Continuing Commitment to Technology Evolution

Powered by HPC+AI infrastructure, digital prototyping continues to drive product design and manufacturing. As engineering demands more computation capacity and capabilities, Intel has continued to collaborate with the industry across private and open source communities. This collaboration has helped address challenges in manufacturing and optimize solutions for HPC+AI in manufacturing.

Intel's technology continues to evolve, requiring ongoing commitments to infrastructure that can deliver next-generation technologies. New Intel engineering and manufacturing sites being built around the world will allow the company to continue to develop and create these technologies. Our goal is to address future manufacturing challenges with next-generation solutions.

intel.

[1] See [108, 123] at www.intel.com/3gen-xeon-config. Results may vary.

[2] Intel data. See OpenFOAM® on Intel® Xeon® Scalable Processors.

[3] Current server configuration is an HPE ProLiant based on the HP DL360 Generation 10 rack servers equipped with Intel Xeon Gold 6242 3.1 GHz processors, 40 cores with 292 machines available. Previous CPU setup included Intel Xeon Gold 6254, 36 cores with 210 machines available; there were multiple changes from the beginning to the end of this work that enabled this performance, including changing tool versions, hardware setup, tool setup and more. STMicroelectronics notes that when submitting a job on the compute farm, it can start on any type of machines which is why jobs are often using both types at the same time. Data provided by STMicroelectronics.

[4] BASELINE: Test by Flutura as of Jan/2021. 1-node, 2x Intel® Xeon® Gold 6252 CPU @ 2.10GHz Processor, 24 cores HT On Turbo ON Total Memory 192 GB (12 slots/16GB/ 2666 MHz), BIOS: SE 5C620.86B.02.01.0011.032620200659 (ucode: 0x4003003), Ubuntu 18.04.5 LTS, 4.15.0-130-generic, gcc 7.5.0 compiler, Inference Framework: OpenVINO (2020.1.023), Intel® MKL-DNN, PCB Defect Detection: Yolo V4, customer data, 1 instance/2 socket, Datatype: FP32

NEW: Test by Flutura as of Jan/2021. 1-node, 2x Intel® Xeon® Gold 6252 CPU @ 2.10GHz Processor, 24 cores HT On Turbo ON Total Memory 192 GB (12 slots/16GB/ 2666 MHz), BIOS: SE5C620 .86B.02.01.0011.032620200659 (ucode: 0x4003003), Ubuntu 18.04.5 LTS, 4.15.0-130-generic, gcc 7.5.0 compiler, Inference Framework: OpenVINO (2020.1.023), Intel® MKL-DNN, PCB Defect Detection: Yolo V4, customer data, 1 instance/2 socket, Datatype: INT8

04/07/2022/RJMJ/CP/PDF ♻ Please Recycle 350525-001US