Industry Solution Brief

Intel® HPC+AI Technologies
Healthcare and Life Sciences

Intel HPC+AI Technology Accelerates Time to Discovery in Healthcare and Life Sciences

Intel commitment, collaboration, and contributions to the industry help advance insight in digital biology, understanding of disease, and therapeutic development

Table of Contents

Executive Summary

Advancements in technologies and computational methods are revolutionizing healthcare and life sciences, enabling clearer insights into digital biology, digital chemistry, and biophysical research. Breakthroughs using advanced computation are leading to personalized medicine, faster and more accurate diagnoses, accelerated drug and vaccine development, better understanding of epidemiology, and development of new therapies.

Intel engineers, solution architects, and scientists collaborate with healthcare and life sciences experts to understand their ongoing challenges, to help enable solutions that accelerate discovery. From enhanced silicon to optimized, open workflows, Intel supports the researchers and scientists seeking answers to our most critical health issues.

HPC advances the effectiveness of cryo-electron microscopy, an essential technology for studying the architectures of cells, viruses, and protein strutures at molecular resolution.

Challenges

The work in healthcare and life sciences crosses multiple scientific domains, including omics, modeling/visualization (Cryo-EM), quantum mechanics, and molecular dynamics. Together, they address the challenges that researchers, clinicians, and developers face with data expansion, regulatory compliance, and software and computing capabilities. Key challenges include the following:

- **Domain scale**
  The biophysical and pharmacological molecule domains are massive. Computationally traversing these domains efficiently to find and create molecules not available in nature that can address disease requires new methods and ever-greater compute capabilities.
- **Data expansion**
  Data expansion within life sciences is exponential. From DNA sequencing to single-cell RNA sequencing and high-resolution imaging, these rich data resources give scientists massive repositories for collaborative research and potential for discovery. But big data presents significant challenges that impact storage, access latency, security, and computational throughput.
- **Complexity**
  Computational workflows involve multiple and various software pipelines that have evolved using a variety of programming languages and frameworks. For example, the Genomics Analysis Toolkit (GATK) contains 24 individual steps, of which 18 are single-threaded and six are multi-threaded. Hardware manufacturers and software vendors offer technology-specific optimization libraries for acceleration. The confluence of all these creates complexity across the entire toolchain.
- **Compute heterogeneity**
  Computational problems drive the workloads and technologies needed to run them efficiently, quickly, and cost effectively—irrespective of whether deployed in the cloud or on-premises (Table 1). Different steps in a pipeline might benefit most from specific types of computing, such as CPUs, GPUs, FPGAs, accelerators, distributed computing clusters, large memory nodes, and cloud or on-premises deployments. The need for different silicon architectures and software development to address these challenges and achieve fastest results must be balanced against optimizing TCO for enterprise operations.

Every time humans get to see better (e.g., telescope, microscope), we revolutionize scientific insight and transform the world forever. The unprecedented resolution into biological processes today translates to the rise of data and the need for tools that help us make sense of that data and "see" better. Like any other end to end problem, it starts with data and the challenges associated with data. Processing of the massive scale of biological data has significant implications for computing as a paradigm. We will need novel platform architecture and breakthrough algorithms in this space, followed by optimized software stacks to deal with massive data.

—Bharat Kaul, Director, Parallel Computing Lab, Intel Labs

Domain	Workload Profile	Example Pipeline Components
Genomics	• Compute-intensive • Memory-bandwidth-intensive (e.g., BWA), scales well to ~40 cores • Datasets require small footprint (e.g., GATK)	Burrows-Wheeler Aligner (BWA-MEM and BWA-MEM2) Minimap2 HaplotypeCaller
Cryo-EM	• Compute-intensive • Datasets require large memory footprints: 256+ GB RAM • Highly scalable codes	RELION 3.x
Quantum Mechanics	• Compute-intensive • Increased memory and storage for batch size management	VASP NWChem
Molecular Dynamics	Compute-intensive Highly scalable codes	NAMD GROMACS LAMMPS

Table 1. Sample HLS workloads and pipelines

Supercomputing for COVID-19

Within healthcare, pharmacological therapies are some of the most time-consuming products to deliver to market. Vaccines typically take 10–15 years to develop.[1] The first COVID-19 vaccines were developed in an unprecedented time of less than 15 months. Before COVID-19, the Mumps vaccine was developed the quickest (in four years)[2] followed by the Ebola vaccine, which took five years.[3]

The speed of the COVID-19 vaccines development was supported by the power of supercomputing around the world. In the history of vaccine development, the amount of supercomputing for COVID-19 is unprecedented.[4] According to the COVID-19 HPC Consortium, institutions thus far have dedicated 603 petaFLOPS from 6.4 million CPU cores and 49,000 GPUs across 141,000 nodes.[5] Much of this work has been done in the cloud and on the world's fastest supercomputers. These resources have been and continue to be dedicated to understanding the SARS-CoV-2 genome, discovering the right molecule to inhibit its spike protein, and developing the drugs.

The insights we have gained from fighting this pandemic will help us address future ones more effectively. The discoveries and breakthroughs in healthcare and life sciences are due to the dedicated and brilliant scientists across multiple domains and the computational capabilities available to them.

Solutions

Intel software and engineering teams are addressing challenges with technology advancements that optimize and accelerate computational tools across different—possibly unique—workflows.

> *Components of an entire workflow have different computational demands. We need to think about the entire toolchain, from sequencing to variant calling and analysis, to modeling and manipulation of Cryo-EM-generated data, to quantum mechanics, molecular dynamics, and simulation during drug design, discovery, and screening.*
>
> —Michael J. McManus, PhD, Director, Precision Medicine & Principal Engineer, Intel

Optimizing Workloads

Intel has done extensive work to optimize several important workloads across the domains to deliver performance, TCO, and a better user/developer experience on Intel architecture.

Intel optimizations for specific languages and operations take advantage of integrated Intel silicon capabilities. These optimizations allow open source and commercial codes to run faster and more efficiently on less hardware—deployed on-premises and in the cloud. For example, Intel® Optimizations for TensorFlow*, Intel® Distribution for Python*, and Intel® oneAPI toolkits help accelerate workloads with none to few code changes. For example, RELION-3, with vector acceleration and tuned for Intel architecture, runs 2.26 times faster on a 40-core Intel® Xeon® 8380 processor-based two-socket server versus a 24-core Intel Xeon 8268 processor-based two-socket server.[6]

Intel has collaborated with the community to optimize many important workloads on Intel® architecture.

Evolving AI

The convergence of HPC+AI offers opportunities to researchers to accelerate understanding and discovery. For example, combining outputs from Cryo-EM data and RELION-3 with AI algorithms can potentially accelerate finding dockable sites around an epitope and identifying manufacturable vaccine molecules that can attach to it to address mitigating disease.

> *"With AI, it is like we just discovered the wheel with its many possibilities yet to be realized. AI will bring a new horizon of possibilities in digital biology."*
>
> —Bharat Kaul, Director, Parallel Computing Lab, Intel Labs

Computer science is in the early stages of AI, machine learning, and deep learning. Technologies continue to evolve across the industry to enhance and accelerate these methodologies. With each new generation of Intel processor architecture, Intel introduces new AI-focused capabilities in their silicon. For example, Intel® Deep Learning Boost (Intel® DL Boost) and Intel® Advanced Vector Instructions 512 (Intel® AVX512) are parts of Intel AI enhancements released in past generations of Intel CPUs. Advanced Matrix Extensions (Intel® AMX), that will be integrated into 4th Generation Intel Xeon Scalable processors, enables GPU-like performance with a CPU on many matrix operations used in AI. Other Intel technologies that can assist AI-based research include:

- Security-enhancements, including Intel® Software Guard Extensions (Intel® SGX), full disk encryption, and others, allow building software tools that protect data. Securing data enables users to build better trained models by being able to federate their deep learning across multiple datasets.
- Intel® Xᵉ-HPC architecture is designed to further accelerate model training with large datasets found in life sciences.
- OpenVINO™ and the Intel® Distribution of OpenVINO™ toolkit accelerate computer vision solutions development and inference performance. The toolkit is also part of Intel oneAPI toolkits.

Open Omics Acceleration Framework for Optimizing Genomics Workloads

To help accelerate digital biology research, Intel is developing the Open Omics Acceleration Framework. Open Omics is an open-sourced high throughput framework that brings together key digital biology pipelines, biological compute motifs, AI, and data management. It efficiently utilizes the underlying hardware architecture to enable productive performance. The Open Omics framework is:

Community driven—being built with extensive discussions with thought leaders to understand the requirements of the user community.

Modular—the developer community can use modules to achieve faster performance for existing and new software tools without any change in accuracy.

Open source—anyone can customize it for variations in use-cases.

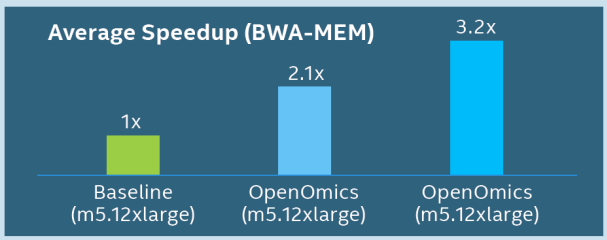
Hardware accelerated—data science experiments can be done quickly and efficiently to help reduce computing costs.

Supporting the full application stack—including application layer (e.g., genomics, single-cell analysis, drug discovery); middleware, with efficient and scalable implementations of key building blocks; and architecture (processor, memory, storage, and interconnect).

Recent benchmarking of Open Omics on Amazon Web Services (AWS)—the older c5 and m5 and the newer c6i and m6i instances of three key genomics applications and their well-optimized versions—demonstrated significant speedup. These applications are sequence mapping software tools (BWA-MEM and minimap2) and single cell ATAC-Seq data analysis. Compared to the baseline implementations running on the single or m5 instances, the Open Omics versions of these applications running on single c6i or m6i instances achieve speedups of up to 3.5x for BWA-MEM, 2.5x for minimap2, and 11.9x (for ATAC-Seq data analysis.[7] For ATAC-Seq data analysis, the Open Omics version gets further speedup using 16 c6i instances achieving nearly 145x speedup compared to baseline running on a c5 instance.[7] AWS c5 and m5 instances run on 2nd Gen Intel® Xeon® Scalable processors. AWS c6i and m6i instances use 3rd Gen Intel Xeon Scalable processors.

The following figures chart the results of these benchmarks for both performance and price-performance. In all tests, the Open Omics versions perform better and cost less to complete than the baseline, non-Open Omics workloads.

Read about the Intel-AWS collaboration.



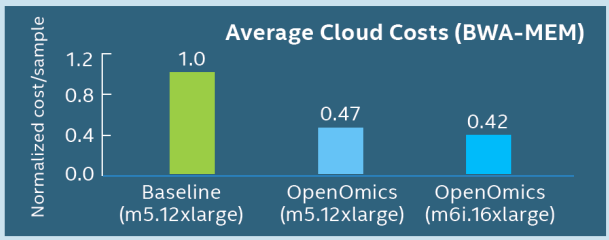
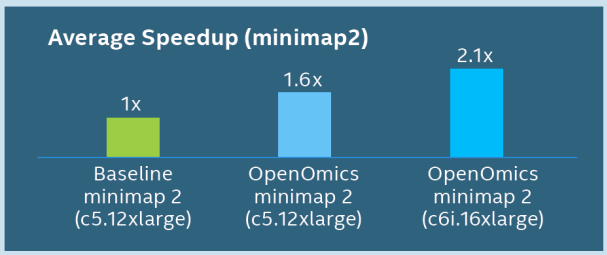


Figure 2. Performance (left) and price-performance (right) charts of BWA-MEM on AWS m5 and m6i instances using the OpenOmics framework.



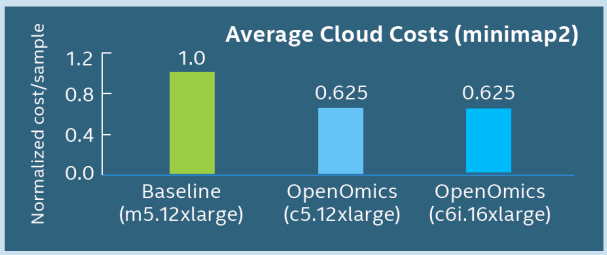
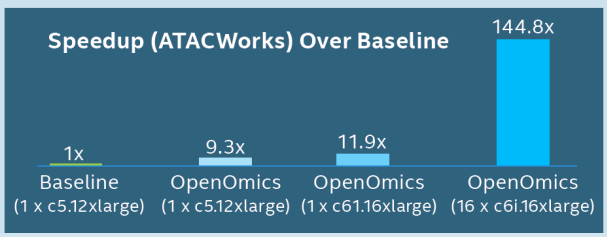


Figure 3. Performance (left) and price-performance (right) charts of minimap2 on AWS c5 and c6i instances using the OpenOmics framework.



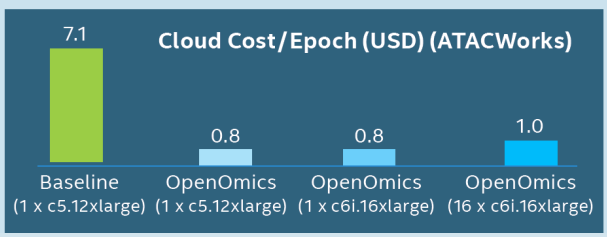
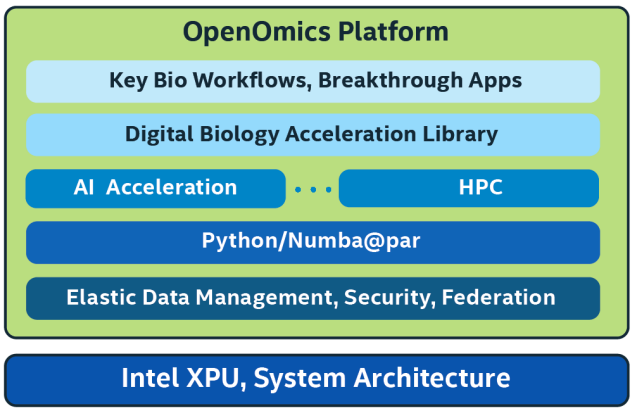


Figure 4. Performance (left) and price-performance (right) charts of ATACWorks on AWS c5 and c6i instances using the Open Omics framework.



The Open Omics Acceleration Framework utilizes the underlying hardware architecture to enable productive performance.

Scientists in Intel Labs are working closely with researchers, clinicians, and the commercial community to help develop AI algorithms and optimize Intel architecture to help accelerate discovery and disease diagnoses.

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. Intel® Optane™ technologies help address data challenges across the hardware platform.

- Intel® Optane™ persistent memory (Intel® Optane™ PMem) offers high capacity, affordability, and persistence of data on the memory bus. It puts larger amounts (multiple terabytes) of data closer to the processor for lower latencies and enhanced performance. New designs around Intel Optane PMem, such as the MemVerge Memory Machine, offer accelerated computation with big memory platforms.[8]
- Intel® Optane™ SSDs combine attributes of memory (speed) and storage (density) to accelerate data movement.
- High-bandwidth memory (HBM), to be introduced in 4th Gen Intel Xeon scalable 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.

Simplifying Development

Different computational problems demand different silicon solutions. But algorithm and software developers should be enabled to move quickly to deliver their codes, irrespective of the underlying technology. The oneAPI standard and Intel oneAPI toolkits help unify development across different technologies—both Intel and other manufacturers—to simplify code development. Research and solutions with oneAPI continue. For example, Oakridge National Laboratory established a oneAPI Center for Excellence to evaluate single-model parallel codes, many of which are used in sciences.

And GROMACS 2022 is now accelerated by oneAPI open programming and multiarchtecture tools running on Intel data center GPUs.

Examples

Across the HLS industry, Intel technologies have helped scientists and researchers discover and deliver new AI-powered diagnostic tools, pharmacological therapies, and treatments for our most critical diseases. From genomics analytics to simulation and modeling of molecules and multiphysics, Intel HPC has provided HPC solutions that enable discovery.

The Broad Institute—Building an optimized pipeline for Google Cloud on Intel architecture

The Broad Institute of MIT and Harvard (Broad Institute) is a world-renowned organization dedicated to transforming medicine. It is a major center of genomics-based research, not only as a provider of analytics tools and computational resources, but also as a sequencing service provider. Their genome sequencing facility sequences a whole human genome every three to five minutes, 24 hours a day. Each genome produces about 350 gigabytes of data. They currently manage about 30 petabytes of genomic data.

In collaboration with Google Cloud and Intel, the Broad Institute optimized their genomics workloads for fast, cost-effective execution on Google Cloud N1 and N2 instances. The Broad Institute cloud vision involved workload migration and building a new collaboration platform called Terra.bio. The cloud delivers scalable computing resources as well as data management facilities and tools the Broad Institute needed for growth. The collaboration resulted in 85 percent reduction in cost of data processing after optimization compared to the initial deployment of workloads on Google Cloud.[9]

> *Terra provides a user-friendly environment that enables researchers to access the datasets they need, and apply the tools they want, securely and at scale. The platform also makes it easy to share their work at any stage, either privately with their collaborators or publicly with the world, in a form that makes their analysis completely reproducible and extensible.*
>
> —Geraldine Van der Auwera, The Broad Institute

By migrating to the cloud and optimizing their workloads for Google instances, the Broad Institute solved their storage capacity and computational capability challenges in a scalable, forward-looking way. Building the Terra platform in partnership with Microsoft and Verily further enabled the Broad Institute to empower not only their own research teams, but life scientists around the world. Terra allows researchers to take advantage of these optimized tools and pipelines and to participate in a federated data ecosystem that opens many exciting new possibilities for biomedical research. Read the entire case study here.

TRON—SARS-CoV-2 and the CoVigator

TRON
Science in Translation

TRON gGmbH (for Translational Oncology) is a nonprofit research organization established as an independent spin-off of the University Medical Center of the Johannes Gutenberg University Mainz (Germany). Researchers at TRON have been studying the immuno-biology of cancers—and genetic biomarkers of disease—for more than ten years.

A genetic biomarker can indicate early presence of disease before it shows clinical symptoms. Molecular biomarkers can also suggest a particular therapy for an individual patient. Biomarkers are important to understanding the evolution of a disease and how individualized medicine can affect the disease in different patients.

TRON scientists in the Biomarker Development Center applied their expertise and knowledge in genome analytics to study how the SARS-CoV-2 virus spike-glycoprotein variants attack host cells. To conduct necessary gene sequence analysis, TRON needed to extend its computational capacity. They acquired Intel® Server System nodes to run their Coronavirus Navigator NGS (CoVigator NGS) genome alignment and analytical pipelines. As documented in the case study, the new cluster allowed them to complete a study of nearly 2 million virus genomes and over 30,000 virus genome sequencing datasets, discovering many variants of the spike protein. The research was released as a bioRxiv pre-print. The CoVigator is now a publicly available database research tool that is updated as more genomes are analyzed.

> *As a project, without the new servers, we would have only been able to complete the initial study. But we would not be able to provide an ongoing study and service that analyzes millions of samples and identifies spike protein variants. With this platform, we are able to keep the work going.*
>
> —Thomas Bukur, TRON

U Buffalo—Digital Biology and More

University at Buffalo
Center for Computational Research

Innovative medical technology companies in Western New York, such as Marion Surgical whose VR/AR technology is featured above, take advantage of the supercomputing capabilities at the University at Buffalo's Center for Computational Research (image courtesy Marion Surgical)

The University at Buffalo's (UB) Center for Computational Research (CCR) offers unique opportunities to Western New York's many businesses. The CCR provides dedicated converged high-performance computing (HPC) and Artificial Intelligence (AI) capabilities with an HPC+AI compute cluster (the CCR cluster). The new system enables a large community of customers to develop innovative solutions through simulation, modeling, and machine learning.

Virtual Reality—powered Surgery Rehearsal—Marion Surgical uses virtual reality (VR) and a specialized haptic feedback robot to realistically simulate complex surgical procedures. The simulations provide an interactive environment with the actual feel of surgical devices as they penetrate different tissues: skin, fat, muscle, and different layers of kidney. Surgeons can learn or rehearse a procedure before they enter the surgical theater.

Marion Surgical engineers use the CCR cluster to build their simulations and integrate patient-specific imagery. They create a 3D model from patient CT scans and integrate the model into the rehearsal. The surgeon then sees physical pathways as the surgical instrument moves through the model, while the robot provides feedback. Surgeons can gain an immersive experience before entering the surgical suite.

Simulating Treatment for Orthopedic Implants—Garwood Medical Devices is developing BioPrax*, a treatment that attacks bacteria on surgical implants. Biofilm bacteria are

Intel® Select Solutions for Genomics Analytics

Intel Select Solutions are verified hardware and software stacks that are workload-optimized across compute, storage, and networking resources. Built on 2nd and 3rd Gen Intel Xeon Scalable processors, Intel Select Solutions for GATK help ensure enterprises get the scalability and performance they require.

The Intel-Broad Center for Genomic Data Engineering brings together science and technology to optimize genomics analytics codes and workflows and to define an optimized infrastructure. The result is the Intel Select Solutions for Genomics Analytics for running GATK workloads. The Intel Select Solutions for Genomics Analytics enable faster analysis and quicker times to deploy hardware solutions that are customized for genetics analysis. The solutions demonstrated the following benefits:

- 5x overall performance improvement running GATK 4.0 compared to previous versions of the genomics software[11]
- Reduced setup time for deploying an infrastructure to accelerate genomics workflows[11]
- 75 percent performance speedup for BWA using Intel® SSDs[11]

highly resistant to antibiotics and difficult to eliminate. Infection is a major problem that can end in implant replacement or may result in amputation and death.[10]

The BioPrax treatment injects current into an implant, which creates electrolysis around it and forms an environment that kills bacteria. Researchers at the University at Buffalo working with Garwood Medical use the CCR to model and simulate the treatment with Multiphysics software. The modeling and simulation integrate the physical environment of the patient, size, geometry, and metallic makeup of the implant, and application of electrical current to the body. The validated model matches well experimentally. The FDA has accepted the model and brought the treatment into the FDA's Breakthrough Devices Program.

Across the healthcare and life sciences domain, Intel's role and commitment of silicon, hardware, and software optimization have and will continue to support the advancements of digital biology, digital chemistry, and development of new therapeutic solutions. Find out more on the Intel Healthcare and Life Sciences Technology Solutions web site.

[1] https://www.historyofvaccines.org/content/articles/vaccine-development-testing-and-regulation

[2] https://www.history.com/news/mumps-vaccine-world-war-ii

[3] https://www.nature.com/articles/s41541-020-0204-7

[4] https://newsroom.ibm.com/IBM-helps-bring-supercomputers-into-the-global-fight-against-COVID-19

[5] This does not include work done by affiliates, such as EU PRACE COVID-19 Initiative, NCI Australia and Pawsey Supercomputing Centre facilities, and others, including South Africa Center for High Performance Computing.

[6] Testing performed by Intel 7/13/2020 (RELION 3.0.4), 1/19/2021 (8268), 5/26/2021 (8358, 8380),
BASELINE: INTEL® XEON® PLATINUM 8268 PROCESSOR: Dual-Socket Intel® Xeon® Platinum 8268 processor, 2.9GHz, 24 Cores/Socket, 48 cores total, turbo and HT on, BIOS SE5C620.8 6B.02.01.0011.032620200659, 192GB total memory, 24 slots (12 populated) / 16 GB / 2933 MT/s / DDR4 RDIMM, 1 x 800GB Intel® SSD SC2BA80, CentOS* 7.8.2003 kernel 3.10.0-1127.13.1.el7. crt1.x86_64, microcode microcode_ctl-2.1-61.10.el7_8.x86_64/0x500002c.
NEW: INTEL® XEON® PLATINUM 8380 PROCESSOR: Dual-Socket Intel® Xeon® Platinum 8380 processor, 2.3GHz, 40 Cores/Socket, 80 cores total, turbo and HT on, BIOS SE5C6200.86B.0020. P23.2103261309, 256GB total memory, 32 slots (16 populated) / 16 GB / 3200 MT/s / DDR4 RDIMM, 1 x 800GB Intel® SSD SC2BA80, CentOS* 8.3.2011 kernel 4.18.0-240.15.1.el8_3.crt1.x86_64, microcode microcode_ctl-20200609-2.20210216.1.el8_3.x86_64/0xd000270.

[7] https://aws.amazon.com/blogs/hpc/accelerating-genomics-pipelines-using-intel-open-omics-on-aws/

[8] For example, https://phoenixnap.com/company/press/memverge-memory-virtualization-in-bare-metal-cloud

[9] Results reported by the Broad Institute.

[10] Ercan B, Kummer KM, Tarquinio KM, Webster TJ. Decreased Staphylococcus aureus biofilm growth on anodized nanotubular titanium and the effect of electrical stimulation. Acta Biomater. 2011;7:3003–12. doi: 10.1016/j.actbio.2011.04.002.

[11] https://www.intel.com/content/www/us/en/products/docs/select-solutions/select-solutions-for-genomics-analytics-brief-v3.html

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 backup for configuration details. No product or component can be absolutely secure.

For workloads and configurations visit www.Intel.com/PerformanceIndex. Results may vary.

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

Your costs and results may vary.

Intel technologies may require enabled hardware, software, or service activation.

0822/RJMJ/CP/PDF ♻ Please Recycle 351388-001US