White Paper

Intel® Data-Centric Technologies
Cloud Services

Today's Top Clouds Are Powered by Intel

Large cloud service providers (CSPs) rely on Intel architecture in their hyperscale operations.

If you manage a data center, you know the challenges: scaling infrastructure to handle growing amounts of data, optimizing systems for demanding new workloads such as artificial intelligence (AI), guarding application and data privacy against evolving threats, and always working to accelerate performance at every level. Cloud service providers (CSPs) face these same challenges, operating at hyperscale, and they often build their solutions with collaboration from Intel.

Intel has been at the forefront of hyperscaling cloud services for decades. Through co-engineering and business relationships with top CSPs globally, Intel has delivered generations of custom silicon optimized and built for cloud scale. Many of the features originally designed for hyperscalers have since been engineered into Intel's data-centric technology portfolio for use in data centers of all sizes.

This paper describes some of the key Intel technologies in use at the biggest CSPs in the world. This information is intended to be valuable for you in two ways:

- If you are migrating workloads to a public cloud, this information can help you identify and understand the benefits of Intel technology optimizations that support the kinds of workloads you are migrating so you can seek out CSP offerings that best meet your needs.
- If you are continuing to run critical workloads on premises in your own data center, you should consider adopting some of the same technologies that the leading CSPs are using to tackle their hyperscale challenges. Choose technologies that have been hardened in the largest data centers on earth, and you create a level of infrastructure compatibility that makes it easier to move workloads between your data center and public clouds as needed.

While some technologies, such as Intel Turbo Boost Technology and Intel Hyper-Threading Technology (Intel HT Technology), might be integrated into almost all Intel-based instances at a CSP, other technologies might be available only in a specialized subset of those instances—which you'll need to identify. Intel is committed to delivering ubiquitous access to cloud services to enable you to determine the best business model for your company.

The Intel technologies discussed in this paper are grouped into the following categories:

- **Intelligent network connectivity**. CSPs rely on high-performance network connectivity that is flexible and adapts to changing demands and workloads to move data faster and maintain high compute performance.
- **Application and data privacy**. Companies like yours need to trust that workloads and data that move to the cloud remain private and protected from malicious threats—and from even the ability of the cloud provider itself to access the data.
- **AI and high-performance computing (HPC)**. CSPs are increasingly able to accommodate these highly compute-intensive workloads in specialized instances built on Intel technology.
- **Performance acceleration**. Organizations are continuously asked to do more with less by improving workload performance and density with innovations in processors, memory, and storage.

Navigating these times of rapid change requires agile, trusted, and scalable solutions to ensure business continuity and long-term success. Business needs are evolving fast, and IT infrastructure must be able to respond at equal speed. This might mean scaling quickly on-premises or optimizing

workload performance in the cloud to meet changing demands. Intel technologies are at the heart of both the enterprise data center and public CSP service offerings, delivering a trusted technology foundation for smooth transitions and migrations, in addition to performance tailored to meet the needs of today's and tomorrow's workloads, with improved operational efficiency to scale from the data center to the cloud to the edge.

What are the key Intel cloud technologies, and how do they work? In what ways are they being used by CSPs, and what business value do they provide to the CSPs and their customers? This paper provides a brief overview of each technology to help answer questions like these. Understanding how hyperscalers are using Intel technologies can be useful for infrastructure architects and other decision makers at digital enterprises building a cloud strategy. These decision makers need a clear understanding of the implications of their decisions about public and private cloud services and their underlying Intel technologies.

High-performance intelligent network connectivity

Hyperscalers choose intelligent, scalable, software-defined infrastructure over single-purpose fixed hardware and appliances. Two key trends are driving this choice: distribution of computing power to remote sites and the movement of data center functions closer to the edge. Compute performance and data processing and analysis can be limited by the network infrastructure not moving data fast enough to meet increased demands on the network from growth in traffic. Software-defined networking (SDN)—high-performance network connectivity that is flexible and adapts to changing demands and workloads—is essential for making the most of compute performance. From Ethernet network interface controllers (NICs) and Intel Silicon Photonics to switching products and technologies, Intel has a long history of innovation in networking products including hardware, software, and solutions, enabling a broad ecosystem that benefits hyperscalers.

Intel Ethernet Controllers and Adapters

As Ethernet bandwidth continues to expand from 25 gigabits per second (Gbps) to 50 Gbps to 100 Gbps, and multiple applications contend for network bandwidth, the processing overhead required to serve the network interface has become an issue. The problem is exacerbated by the increasing complexity of modern-day data center networks, which include support for performance-acceleration techniques, virtualization, and overlay networks. Data center operators need Ethernet adapters that are fast and intelligent to support the dynamic demands on the network. This can include accelerating some of the processing of infrastructure workloads on the NIC. CSPs use Intel Ethernet Network Adapters for their proven broad interoperability, critical performance optimizations, and agility.

Intel Ethernet Network Controllers and Adapters support speeds up to 100 Gbps. A few of the key features of Intel Ethernet Network Adapters that are designed to benefit hyperscalers and large data centers more generally, include:

- **Faster packet processing with improved performance for network functions virtualization (NFV)**. Intel Ethernet Network Adapters use a combination of hardware- and software-acceleration features, including Dynamic Device Personalization (DDP) to enable customizable packet filtering, enhanced Data Plane Development Kit (DPDK) support for advanced packet forwarding, and highly efficient packet processing for both cloud and NFV workloads.
- **Advanced traffic steering to better meet service-level agreements (SLAs).** The Intel Ethernet 800 Series with Application Device Queues (ADQ) provides dedicated queues and shapes traffic for the transfer of data over Ethernet for critical applications. It improves application response time predictability, lowers latency, and improves throughput for key applications such as databases, the web tier, and caching applications. This improves the ability of hyperscalers to meet SLAs.
- **Server virtualization with flexible and scalable input/output (I/O) virtualization**. Intel virtualization delivers I/O performance and reduces I/O bottlenecks through technologies like Virtual Machine Device Queues (VMDq) and Flexible Port Partitioning (FPP) using single root I/O virtualization (SR-IOV) for networking traffic per virtual machine (VM). This approach enables near-native

performance and VM scalability. With Intel Virtualization Technology (Intel VT), Intel Ethernet Network Adapters deliver outstanding I/O performance in virtualized server environments.

Smart network interface cards (SmartNICs)

Like other data center operators, CSPs want to optimize for critical workloads while being efficient in managing infrastructure tasks. When infrastructure tasks can be accelerated by intelligent network components, server CPUs can be freed up for critical workloads.

Hyperscalers use SmartNICs based on Intel field-programmable gate arrays (FPGAs) to reduce server overhead by accelerating infrastructure work usually performed by CPU cores by running them on the SmartNICs themselves. SmartNICs can deliver performance for infrastructure workloads, and they allow for changes in network technology through software updates in the field. CSPs can use SmartNICs to free up host CPU cores for running end-customer VMs, or they can use a SmartNIC to partition a server as a bare-metal server to support a hosting business.

SmartNICs using Intel FPGAs enable the following advantages for CSPs, who have made sizeable investments in tailoring and integrating SmartNICs into their infrastructures:

- Increased business efficiency and scaling of applications, because more host CPU cores become dedicated to VMs for customer workloads
- Accelerated infrastructure workloads, such as virtual network functions (VNF) like open vSwitch, security like IPsec and Transport Layer Security (TLS), and storage like NVM Express (NVMe) over Fabrics (NVMe-oF)
- The ability to future-proof solutions by making use of the flexibility and programmability of an Intel FPGA: organizations can add functionality or enhance the performance of workloads running on the FPGAs to adapt to changing requirements and standards
- Lower total cost of ownership (TCO), because the same FPGA SmartNIC can be provisioned as needed to run different workloads; for example, the same hardware can run security applications during peak times and be updated to run data analytics workloads when network traffic is lower

P4-programmable switching using Tofino

Hyperscale providers don't want black boxes from vendors without visibility into the underlying code and functionality. They need open systems that they can control, debug, and reprogram as necessary. That's why they can benefit from using P4-programmable switches in their massive SDNs for agility and scalability. Barefoot Networks was a pioneer in the development of the P4 programming language, and it has been instrumental in growing the P4 ecosystem to well over 100 members since P4.org was established.

Intel supports a range of open source network operating systems including SONiC. Originally created by Microsoft and in use at a number of hyperscalers,[2] SONiC is based on Linux, and it runs on switches from multiple vendors and application-specific integrated circuits (ASICs). Intel supports SONiC code for the 6.4 terabits per second (Tbps) Tofino and the 12.8 Tbps Tofino 2 switch ASICs. Hyperscalers can get their hands on Tofino and Tofino 2–based switches through a robust OEM and ODM ecosystem. In addition to the white box switches, OEMs like Arista and Cisco have introduced programmable switches built on Tofino and supporting SONiC.

Switches with Tofino programmability make Ethernet switches programmable at the data-plane level, allowing users to define the functionality of the hardware in a simple P4 program, compile it down to the ASIC, and run it at multi-Tbps speeds. For example, the box could be set up to run standard switching and routing protocols at a better scale and quality but also add on top functions like distributed denial of service (DDoS) screening, deep packet inspection, or network address translation (NAT) via quick software updates.

Powering remote work during COVID-19

Like many companies, Intel responded to the outbreak of COVID-19 by quickly enabling many employees to work from home. These newly remote workers needed immediate access to their work environments via virtual private networks (VPNs). The challenge was scaling to the necessary number of VPNs with low enough latency to support worker productivity in the new environment.

Intel IT responded with a two-pronged strategy, executed both on premises and in a public cloud. Both initiatives relied on Intel Ethernet Network Adapters with support for SR-IOV—a specification that allows a PCIe device to appear to be multiple separate physical PCIe devices—to scale large numbers of VPN appliances with predictably low latency.

The public cloud was able to scale most quickly and bore the brunt of the new load during the first two weeks. Intel chose instances carefully to include Intel Ethernet Controllers with SR-IOV support and Intel® Xeon® Scalable processors. These VMs proved able to support VPN users at about 90 percent CPU utilization, based on Intel IT records from the first two weeks of the response.

With Intel Ethernet Network Adapters optimized for SR-IOV, Intel IT was able to add 50 percent more users per VPN server with consistent latency.[1]

The on-premises response used servers running a variety of Intel processors, some of them older, using equipment that was immediately available for reuse. Some systems that did not have Intel Ethernet Controllers proved unable to scale to the necessary network/VPN capacity and delivered inconsistent latency. Intel IT replaced the NICs in these systems with Intel Ethernet Network Adapters with support for SR-IOV, which proved able to scale to meet the demand for large numbers of users experiencing consistent low latency.

Because both the public cloud instances and the on-premises servers were using SR-IOV, load balancing between the two was simplified, and more of the workload was moved back on premises as those systems came online.

Intel Silicon Photonics

Intel Silicon Photonics delivers high-speed, long-distance optical connectivity to support massive warehouse-sized hyperscale data centers, where moving data fast and over longer distances is critical to get data to servers where it can

be processed or analyzed quickly. Inside data centers, optical links connect switches through a complex network of fiber-optical cable and optical transceivers. Servers and storage are no longer required to sit physically together and can be scaled independently based on a variety of requirements such as business continuity and data protection. Intel Silicon Photonics Optical Transceiver products enable large data centers to deploy 100 Gbps solutions for connecting switches across hundreds of meters or several kilometers, rather than just a few meters. Intel's unique approach integrates hybrid silicon lasers as part of the photonic chip, allowing wafer-scale manufacturing and testing, and providing industry-leading quality at the high volumes hyperscalers require to support their growing demand for data. As bandwidth within the switching infrastructure in these massive data centers continues to increase from 3.2/6.4 Tbps to 12.8 Tbps and 25.6 Tbps switches, the need for higher bandwidth optical connectivity also grows. Intel is continuing to drive innovation by also producing 200 Gbps and 400 Gbps Intel Silicon Photonics optical modules. Large private data centers should consider following the lead of the hyperscalers and adopting this mature and cost-effective technology.

Security and privacy for data, applications, and collaboration

As infrastructure scales, so do security threats and business risks. Privacy is a major concern for companies using the public cloud, and therefore a top priority for CSPs. Confidential computing is an emerging industry initiative focused on helping secure data in use, enabling encrypted data to be processed in memory without exposing it to the rest of the system, reducing exposure to sensitive data, and providing greater control and transparency for users. Intel is a founding member of the Confidential Computing Consortium and a contributor of Intel technologies that enable CSPs—and anyone else—to help secure the privacy of applications and data in their data centers.

Intel Software Guard Extensions (Intel SGX)

Data-protection requirements grow ever more stringent. Even data that is well-protected by encryption at rest and in transit can be vulnerable when it is being processed. CSPs and other data centers use Intel Software Guard Extensions (Intel SGX) to help protect data during that critical moment of processing when the data is not encrypted. This helps improve security to support use cases such as the following:

- **Federated (machine) learning**. Federated learning is a distributed approach to ML that enables multiple organizations to collaborate on ML projects, but it requires sensitive data to be protected. Intel SGX is ideal for building trusted execution hardware environments in federated learning solutions in the cloud.
- **Confidential containers and VMs**. In multitenant cloud environments, customers worry that containers and VMs might be open to attack. Support for trusted execution through Intel SGX can be used to help protect container and VM processes from outside attacks.
- **Confidential databases**. Many organizations are moving databases to the public cloud. Intel SGX can be used to increase protection of these databases through isolation of sensitive data or of cryptographic keys.
- **Blockchain**. Intel SGX helps CSPs increase privacy and security for blockchain transaction processing, consensus, smart contracts, and key storage.

Intel SGX is a set of instructions for creating security-enabled enclaves—small, trusted environments within a CPU that can execute code in a way that is not accessible by the normal operating system. These enclaves are also remotely attestable, so that one party can cryptographically verify that an enclave running on another party's computer is running trusted, unmodified code.

Intel SGX enables an additional level of security in the cloud. Now CSPs are beginning to offer customers the ability to keep their sensitive data so private that even the CSP itself is unable to access the data.[3] Even in the event of undetected malware or a rogue administrator, Intel SGX helps protect customers' data from exposure.

Surveys show the top cloud security concern of cybersecurity professionals is data loss and leakage.[4] And 70 percent of business executives surveyed in a 2020 study felt security concerns had held back wider adoption of the public cloud.[5] CSPs see the higher level of security and privacy provided by Intel SGX as a major opportunity to ease their customers' concerns about protecting the confidentiality of data in the public cloud, which could open the floodgates to cloud migration at an even bigger scale.

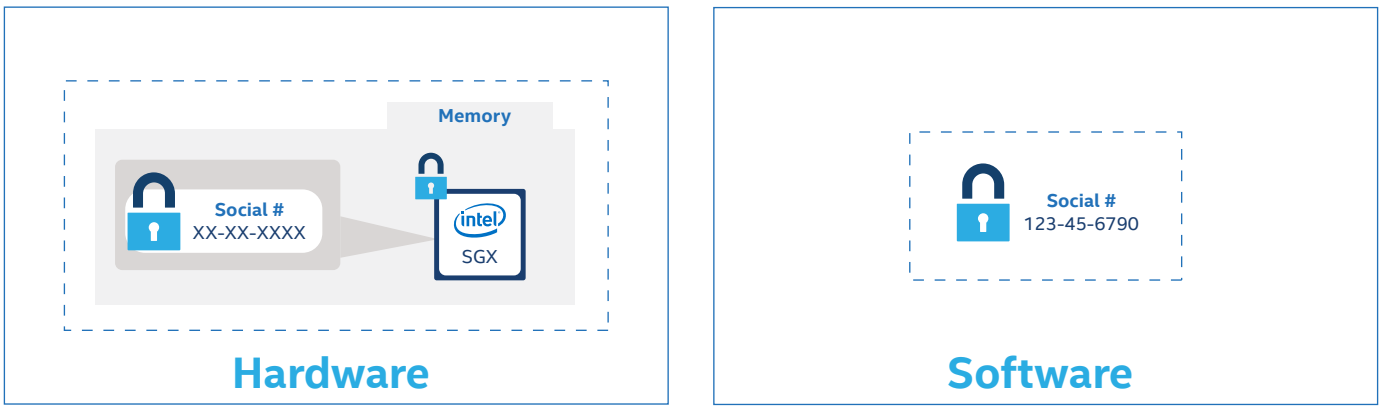


Figure 1. Intel SGX protects information in an enclave when the application is running, and in the hardware when it is not

Intel QuickAssist Technology (Intel QAT)

Intel QuickAssist Technology (Intel QAT) provides hardware acceleration for compute-intensive operations including ciphers, hashes, public-key cryptography, and compression. By offloading this work from CPU cores, performance gains up to 4.3x can be achieved using Intel QAT.[6]

Intel QAT also provides an important security feature called Intel Key Protection Technology (Intel KPT), which helps protect private cryptographic keys.[7] With Intel KPT, the private key is encrypted before it gets inside the VM that needs to use it. Only Intel QAT inside the chip can decrypt the private key, meaning the key is better protected at the hardware level.

CSPs benefit from the acceleration features of Intel QAT by increasing the number of VMs available to offer customers. They can use the Intel KPT feature to differentiate their security, fight the growing sophistication in software and hardware attacks, and make their clouds compliant with regulatory requirements.

AI and HPC

Compute-intensive workloads—AI, advanced analytics, and simulations—are increasingly well supported by cloud services. With cloud computing, organizations can increase or decrease computing resources based on application needs. Hyperscalers, in collaboration with Intel, deliver infrastructure and resources that can be ideally constructed to efficiently handle compute-intensive applications that required large amounts of compute power for extended periods of time. Intel's data-centric portfolio addresses the unique challenges of compute-intensive workloads by bringing together HPC, AI acceleration, and advanced data analytics into a single computing environment that CSPs can use to support their customers' work in scientific simulations, financial analytics, AI/deep learning (DL), and 3D modeling and analysis.[8]

The following sections describe some of the Intel technologies that make HPC and AI applications feasible in the cloud.

Intel Advanced Vector Extensions 512 (Intel AVX-512)

Vector processing performs an arithmetic operation on a large array of integers or floating-point numbers in parallel, which can be highly intensive in applications such as scientific simulations and 3D modeling, for example.

Intel Advanced Vector Extensions 512 (Intel AVX-512) is a set of CPU instructions that boost vector processing-intensive compute workloads for up to 1.6x performance gains.[9]

Intel AVX-512 impacts compute, storage, and network functions. The number 512 refers to the width, in bits, of the register file, which sets the parameters for how much data a set of instructions can operate upon at one time. Intel AVX-512 enables twice the number of floating-point operations per second (FLOPS) compared to its predecessor, Intel AVX2. This means Intel AVX-512 enables processing of twice the number of data elements that Intel AVX2 can process with a single instruction, and four times that of Streaming SIMD Extensions (SSE).

Intel works with hyperscalers to enable the acceleration of operations with no code modifications on their existing frameworks.

Intel AVX-512 shows the greatest performance benefit on workloads that require the same kind of vector/matrix operations to be performed on large amounts of data, such as DNA sequencing.[10] The ability to operate on more information at once means Intel AVX-512 can help handle computational tasks and accelerate performance for workloads and usages such as AI/DL, scientific simulations, financial analytics, and 3D modeling and analysis.

If you are looking to move these kinds of AI and HPC workloads to a public cloud, be sure to choose an instance from a CSP that offers Intel AVX-512 to get excellent performance per dollar.

Intel Deep Learning Boost (Intel DL Boost)

DL applications require HPC capabilities and low latency. Traditionally, graphics processing units (GPUs) have been used to perform ML and DL workloads, resulting in higher hardware costs to achieve the necessary performance for specific cloud instances. Advancements in CPU technology have created new opportunities for CSPs to extend their core infrastructure services to include built-in AI acceleration.

CSPs offer Intel Deep Learning Boost (Intel DL Boost) in high-performance instances to provide customers the optimal environment for inner convolutional neural network loops and some other compute-intensive workloads. Performance gains of up to 3.4x can be achieved using Intel DL Boost for these kinds of AI operations.[11]

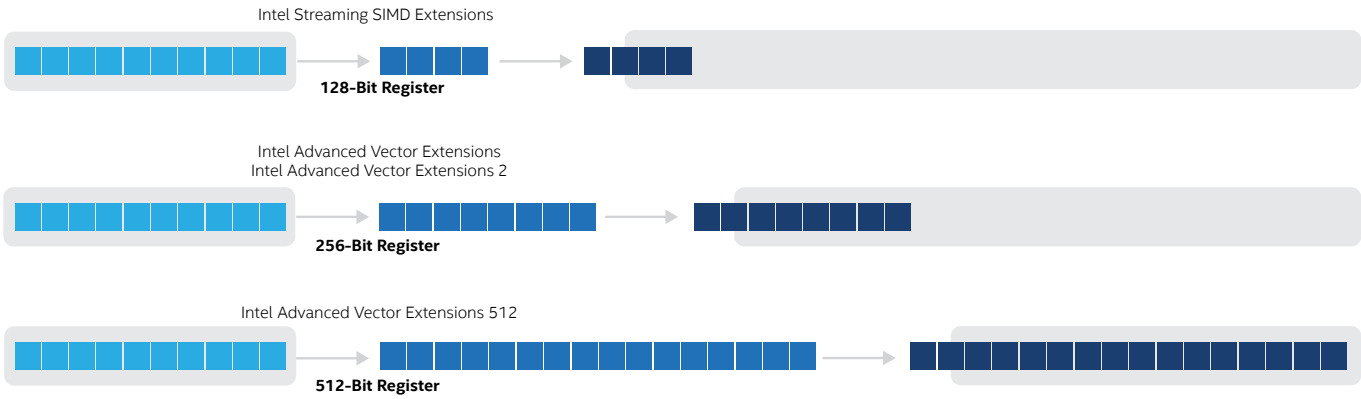


Figure 2. Intel AVX-512 processes more data with a single instruction

Intel DL Boost refers to a set of Intel AVX-512 instructions called Vector Neural Network Instructions (VNNI), which extends the Intel AVX-512 foundation by introducing four new instructions for accelerating inner convolutional neural network loops.[12] Intel DL Boost can result in dramatic performance improvements for this kind of ML, which is used in AI applications such as image recognition, video analysis, and natural language processing (NLP).[13]

Simply by activating Intel DL Boost, the same hardware platform that customers use for other HPC workloads can also be optimized for AI workloads.

Your AI/DL initiatives can benefit from Intel DL Boost, whether on premises on the same HPC infrastructure you use for other workloads or in the cloud using an instance from a hyperscaler that provides the same functionality.

Intel field-programmable gate arrays (FPGAs)

In addition to the SmartNIC use cases of Intel FPGAs noted earlier, Intel FPGAs are also used to accelerate many other workloads, including ML operations, by hardcoding operations into the hardware. Applications that benefit greatly from ML implementations on an FPGA include intelligent vision, scientific simulations, and life science and medical data analysis, among others.[14]

Because they can be reconfigured for different types of ML models, Intel FPGAs can be used to accelerate AI operations where major algorithmic changes come several times a year. Intel FPGAs make it possible to achieve low latency for real-time inference requests.[15] Implementations of Intel FPGA neural processing units don't require batching, and therefore the latency can be much lower than with CPU or GPU processors.

CSPs make Intel FPGA acceleration available to their customers both in the cloud and at the edge.[16] This added acceleration for real-time inference requests makes it much more feasible to migrate compute-intensive workloads to the cloud from your on-premises HPC environment.

Intel Math Kernel Library for Deep Neural Networks (Intel MKL-DNN)

Intel Math Kernel Library for Deep Neural Networks (Intel MKL-DNN) accelerates DL frameworks on Intel architecture. It includes highly vectorized and threaded building blocks for implementing convolutional neural networks with C and C++ interfaces. Intel works with hyperscalers to integrate its Intel MKL-DNN with the frameworks they use, including Apache MXNet and TensorFlow.

Intel® OpenVINO™ toolkit

The Intel distribution of the Open Visual Inference and Neural Network Optimization (OpenVINO) toolkit enables developers to quickly deploy applications and solutions that emulate human vision. It also accelerates other AI workloads such as audio, speech, language, and recommendation systems.

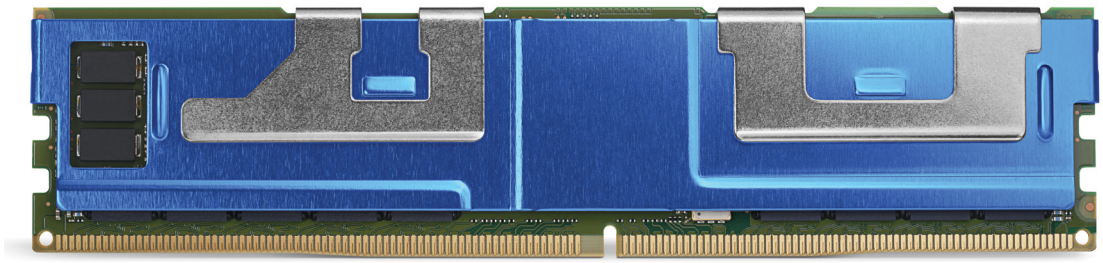
The Intel OpenVINO toolkit enables DL inference on the edge, across heterogeneous Intel hardware including:

- Intel CPUs
- Intel integrated graphics
- Intel FPGAs
- Intel® Movidius™ Neural Compute Stick (NCS) or NCS 2
- Intel Vision Accelerator Design with Intel Movidius Vision Processing Units (VPUs)

You can use the Intel OpenVINO toolkit to implement computer-vision inference on your edge devices using models trained on high-performing instances provided by CSPs.[17]

Workload performance acceleration

Applications are critical to competing in business today, and their performance depends on getting the right data center resources at the right time. Whether it's revolutionary and recent, like Intel® Optane™ persistent memory (PMem), or established and foundational, like Intel Turbo Boost Technology, CSPs and their customers rely on Intel to deliver performance acceleration at every level.



Intel Optane Persistent Memory

Big
High capacity for scalability

Persistent
Data persistence without disk input/output (I/O)

Affordable
Low cost

Robust
Less downtime

Figure 3. Advantages of Intel Optane PMem

Intel Optane persistent memory (Intel Optane PMem)

Intel Optane PMem represents an innovative class of memory and storage technology that allows CSPs to offer customers larger amounts of data closer to the processor, with consistent, low latencies and near-DRAM performance.

CSPs use Intel Optane PMem to cost-effectively expand the capacity of memory available to support higher quantities of "hot" data available for processing with demanding workloads. In-memory databases are a prime example of workloads that benefit from scaling up memory, because they are most efficient when the whole dataset can be processed in-memory at once. But the prohibitive cost of DRAM puts pressure on the cost effectiveness of scaling up memory.

SAP HANA is at the forefront of optimizing in-memory database management systems (DBMSs) for Intel Optane PMem—including the modifications needed to take full advantage of its memory-persistence capabilities. Consequently, CSPs have been working to roll out offerings with instances optimized to run SAP HANA in the cloud on Intel Xeon Scalable processors with large amounts of Intel Optane PMem.[18] Organizations running SAP solutions on premises can also benefit from adding Intel Optane PMem to their private clouds to reap the benefits of memory persistence, including order-of-magnitude faster restart times.[19]

CSPs are also finding Intel Optane PMem valuable for running their own enormous in-house database systems.[20]

Intel Optane technology fills the gap between more expensive DRAM memory and slower NAND storage. When deployed on the memory bus as a DIMM, persistent memory represents a cost-effective alternative to DRAM for scaling up memory—with additional benefits of memory persistence that will become increasingly widespread as more software vendors optimize for it.

Intel Optane Solid State Drives (SSDs)

The same Intel Optane technology that supports larger and more affordable memory is also used in Intel Optane Solid State Drives (SSDs) to provide high performance and high endurance in the storage tier. While Intel QLC 3D NAND SSDs meet the needs of CSPs for massive capacity storage, Intel Optane SSDs are appealing for scenarios where lower latency and greater endurance are required, such as in the caching tier.

CSPs see the value of Intel Optane SSDs for caching because of their ability to quickly read and write large quantities of data. As customer workloads migrate to the cloud with ever larger datasets, caching speeds can become a bottleneck, preventing complex analytics and AI applications, for example, from deriving insights from the data as rapidly as possible. Scaling up a faster caching tier can improve the performance of these complex workloads and better satisfy customer expectations.

Another feature of Intel Optane SSDs that's important to CSPs is their endurance. Like car tires, SSDs are rated for the "mileage" they will get before they wear out. In the case of SSDs, this endurance is measured in total petabytes written (PBW) over the lifecycle of the drive. Hyperscalers can easily push cache drives to the limit of their read/write speeds 24 hours a day, so endurance is a critical factor in the total cost of ownership (TCO) for those drives. One hyperscaler found that two Intel Optane SSDs in an array would last more than four times the PBW as an array of six NAND SSDs, while also delivering lower-latency caching.[21]

Intel Optane SSDs can help CSPs—and private clouds—to scale a fast caching tier that writes massive amounts of data without burning through drives so quickly.

Intel Hyper-Threading Technology (Intel HT Technology)

Some established technologies originally pioneered by Intel continue to be central to the ability of CSPs to provide services at a tremendous scale. Intel HT Technology is a good example.

Intel HT Technology is used to improve parallelization of computations to increase the number of independent instructions in the pipeline. With Intel HT Technology, one physical core appears as two processors to the operating system, allowing concurrent scheduling of two processes per core. CSPs typically use Intel HT Technology to offer customers double the number of virtual CPUs as there are physical cores, while never splitting one core's threads between two customers, to help avoid security risks from shared resources. CSPs also allow customers to turn off Intel HT Technology for applications that need high single-threaded performance from a core.

Intel Turbo Boost Technology

Intel Turbo Boost Technology is another established technology you might take for granted in on-premises environments. When you move workloads to the cloud, you should be sure to use instances based on Intel processors in order to achieve the high performance levels you are accustomed to.

Intel Turbo Boost Technology accelerates processor performance for peak loads, automatically allowing processor cores to run faster than the rated operating frequency if the processor is working within power, temperature, and specification limits. This enables CSPs to offer better performance to customers when their workloads require it, and it is one of the reasons CSPs can command premium pricing for Intel-based instances.

Great clouds are built on Intel technologies

Challenges faced by typical data centers are also faced by the top CSPs at hyperscale. Hyperscalers are the first to encounter new problems and new limits as they move to operations of unprecedented scale. And, as this paper has highlighted, CSPs trust Intel technologies to help them solve problems and overcome the limits of hyperscaling today and tomorrow.

You can benefit from the experience of the hyperscalers. As a consumer of cloud services, knowing about the underlying technologies that power cloud instances can help you choose the right service options to make sure you get the capabilities you need and ensure a seamless migration. If you manage a private cloud, knowing the technology that underpins the top CSPs gives you a best-practices reference to help choose the best available technologies for meeting the challenges of your own data center. By following the best

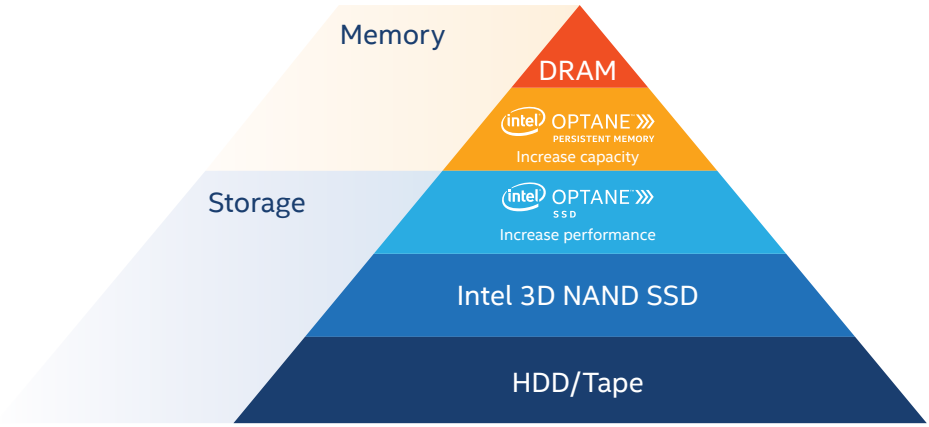


Figure 4. Intel Optane technology fills the gap between expensive memory and slow storage

practices of the hyperscalers, you can accelerate your time to market and scale your private cloud with confidence.

As organizations evolve business strategies and continue to adopt cloud computing services, knowing what's inside your cloud will help you mitigate risk and deliver performance, scale, and choice.

Building your cloud strategy around Intel architecture can also increase your agility and flexibility in choosing the right CSP and architecture for your business needs.

Workloads optimized for Intel architecture can help organizations shift data and applications more easily among different clouds that support the same key Intel technologies. This is especially valuable for the increasing number of enterprises who both operate on-premises data centers and make use of multiple public clouds. You can use the same best practices in your private cloud that the hyperscalers are using, and that means your workloads can run in multiple locations as determined by your particular business and security considerations.

Learn more

Amazon Web Services (AWS) solution brief: **intel.com/content/www/us/en/cloud-computing/amazon-ec2-cloud-powered-by-intel-brief.html**

AWS with Intel webpage: **intel.com/content/www/us/en/artificial-intelligence/aws.html**

Azure solution brief: **intel.com/content/www/us/en/cloud-computing/azure-stack-and-intel-drive-business-value-brief.html**

Azure with Intel webpage: **intel.com/content/www/us/en/big-data/partners/microsoft/overview.html**

Google Cloud solution brief: **intel.com/content/www/us/en/cloud-computing/google-cloud-enabled-by intel-brief.html**

Google Cloud with Intel webpage: **intel.com/content/www/us/en/artificial-intelligence/google-cloud-platform.html**

[1] Based on Intel IT records from the first two weeks of the COVID-19 response.

[2] According to Dave Maltz, distinguished engineer at Microsoft's Azure Networking division, more than 10 hyperscalers and cloud builders have adopted SONiC as their switch operating system, with Microsoft and Alibaba being the two big CSPs that are on the record. Source: The Next Platform. "Is Microsoft's SONiC Winning the War of the Noses?" May 2020. nextplatform.com/2020/05/12/is-microsofts-sonic-winning-the-war-of-the-noses/.

[3] Microsoft Azure was the first CSP to announce general availability of confidential computing based on Intel SGX in April 2020. Source: Microsoft. "DCsv2-series VM new generally available from Azure confidential computing." April 2020. https://azure.microsoft.com/en-us/blog/dcsv2series-vm-now-generally-available-from-azure-confidential-computing/. Used with permission from Microsoft.

[4] In a survey of 700 IT and security professionals, 81 percent of cloud users said they encountered significant security concerns, including concerns over risks of data losses and leakage (cited by 62 percent), closely followed by regulatory compliance concerns (57 percent). Source: AlgoSec. "Cloud Security Alliance Study Identifies New and Unique Security Challenges in Native Cloud, Hybrid and Multi-cloud Environments." May 2019. globenewswire.com/news-release/2019/05/21/1833639/0/en/Cloud-Security-Alliance-Study-Identifies-New-and-Unique-Security-Challenges-in-Native-Cloud-Hybrid-and-Multi-cloud-Environments.html. See also: Cybersecurity Insiders. "2019 Cloud Security Report (ISC)2." cybersecurity-insiders.com/portfolio/2019-cloud-security-report-isc2/.

[5] Barracuda Networks. "Future shock: the cloud is the new network." March 2020. https://lp.barracuda.com/BEU-AMER-WBN-20200304-FutureShockCloudReport_LP-Registration.html.

[6] 4.3x performance gain with Intel QAT as measured by Private Key Exchange TLS 1.2 RSA2K workload as of November 5, 2019: Intel Server Board S2600WFD, Intel Xeon Gold 6252N processor (2.30 GHz, 24 cores, 2 UPI links) run on 18 cores/36 threads with Intel Turbo Boost Technology, with Intel QuickAssist Adapter 8970, 12 x 32 GB DDR4-2,933, BIOS: SE5C620.86B.0X.02.0040.060420190144, microcode: 0x5000026, Ubuntu 19.04, 5.0.0-23-generic, GCC 8.3 compiler, Intel Ethernet Controller XXV710-DA2, NGINX 1.14.2, OpenSSL 1.1.0k, Intel QAT Engine v0.5.41, Intel QAT Driver L05000007.

[7] Intel. "Intel® QuickAssist Technology with Intel® Key Protection Technology in Intel Server Platforms Based on Intel® Xeon® Processor Scalable Family." 2017. intel.com/content/www/us/en/architecture-and-technology/key-protection-technology-white-paper.html.

[8] For more information on Intel's data-centric portfolio, see: Intel. "Intel's Data-Centric Portfolio Accelerates Convergence of High-Performance Computing and AI Workloads." June 2019. https://newsroom.intel.com/news/intels-data-centric-portfolio-accelerates-convergence-high-performance-computing-ai-workloads/.

[9] 1.6x performance average gain with Intel AVX-512 as measured by financial services kernels workload as of November 1, 2019: Intel Server Board S2600WF with 2-socket Intel Xeon Platinum 8268 processors (2.9 GHz, 24 cores, 2 UPI links), 12 x 16 GB DDR4-2,933, one SSD, BIOS: SE5C620.86B.02.01.0008.031920191559; microcode: 0x500001c, Red Hat Enterprise Linux 7.7, kernel 3.10.0-1062.1.1. FSI kernels v2.0: Geomean (three workloads: Binomial Options, Black Scholes, Monte Carlo), Intel AVX2 256 build versus Intel AVX-512 build, Intel Compiler 2019u5, Intel Math Kernel Library (Intel MKL) 2019u5, BIOS: binomial (Intel HT Technology on, Intel Turbo Boost Technology on, SNC off, three threads/core), Black Scholes (Intel HT Technology off, Intel Turbo Boost Technology on, SNC off, one thread/core), Monte Carlo (Intel HT Technology on, Intel Turbo Boost Technology on, SNC off, two threads/core).

[10] For example, see: DeepVariant Blog. "The Power of Building on an Accelerating Platform: How DeepVariant Uses Intel's AVX-512 Optimizations." April 2019. https://google.github.io/deepvariant/posts/2019-04-30-the-power-of-building-on-an-accelerating-platform-how-deepVariant-uses-intels-avx-512-optimizations/.

[11] 3.4x performance gain with Intel® DL Boost (VNNI) as measured by ResNet-50 inference throughput performance; tested by Intel on 12/25/2019: 1-node, 2 x Intel Xeon Gold 6258R processor on Intel reference platform with 384 GB (12 slots, 32 GB, 2,933) total memory, ucode 0x500002c, Intel HT Technology on, Intel Turbo Boost Technology on, with Ubuntu 19.10, 5.3.0-24-generic, AIXPRT image classification AIXPRT v1.01, Intel Distribution of OpenVINO toolkit version 2019 R3, ResNet50 v1, for INT8 with Intel DL Boost: BS=4, 56 instances, for FP32 BS=4, 56 instances.

[12] Intel. "Lower Numerical Precision Deep Learning Inference and Training." January 2018. intel.com/content/www/us/en/artificial-intelligence/solutions/lower-numerical-precision-deep-learning-inference-and-training.html.

[13] In tests measuring inference images/second per socket using PyTorch, integrating Intel MKL-DNN improved performance anywhere from 7.7x to over 105x in different scenarios at fp32 and int8 performance gains over baseline (fp32 without Intel MKL-DNN) for ResNet50, Faster R-CNN, and RetinaNet using batch size 1 on a single socket Intel Xeon Platinum 8280 (Cascade Lake) processor. Source: Intel. "Intel and Facebook collaborate to boost PyTorch CPU performance." April 2019. https://software.intel.com/content/www/us/en/develop/articles/intel-and-facebook-collaborate-to-boost-pytorch-cpu-performance.html.

[14] See, for example, Microsoft's ongoing project Brainwave. In 2018, Bing and Microsoft Azure deployed new multi-FPGA appliances into data centers, shifting the ratio of computing power between CPUs and FPGAs, with multiple Intel® Arria® 10 FPGAs in each server. Source: Microsoft. "Project Catapult." microsoft.com/en-us/research/project/project-catapult/. Used with permission from Microsoft.

[15] The Intel® Stratix® 10 NX FPGA, for example, enables up to 15x more INT8 throughput for AI applications. Based on internal Intel estimates. For more information, see: Intel. "Intel Stratix 10 NX FPGAs."intel.com/content/www/us/en/products/programmable/fpga/stratix-10/nx.html.

[16] Microsoft. "What are field-programmable gate arrays (FPGA) and how to deploy." March 2020. https://docs.microsoft.com/en-us/azure/machine-learning/how-to-deploy-fpga-web-service. Used with permission from Microsoft.

[17] For example, see: Intel. "Simplifying Cloud to Edge AI Deployments with the Intel® Distribution of OpenVINO™ Toolkit, Microsoft Azure, and ONNX Runtime." intel.com/content/www/us/en/artificial-intelligence/posts/microsoft-azure-openvino-toolkit.html.

[18] "For SAP HANA solutions, these new offerings help lower total cost of ownership (TCO), simplify the complex architectures for HA/DR and multi-tier data, and offer 22 times faster reload times." Source: Microsoft. "Next Generation SAP HANA Large Instances with Intel® Optane™ drive lower TCO." April 2020. https://azure.microsoft.com/en-us/blog/next-generation-sap-hana-large-instances-with-intel-optane-drive-lower-tco/. Used with permission from Microsoft.

[19] Restart time reduced from 50 minutes to 4 minutes. Based on testing as of May 30, 2018. SAP HANA simulated workload for SAP BW edition for SAP HANA Standard Application Benchmark Version 2 as of 30 May 2018. Baseline configuration with traditional DRAM: Lenovo ThinkSystem SR950 server with 8 x Intel Xeon Platinum 8176M processors (28 cores, 165 watt, 2.1 GHz). Total memory consists of 48 x 16 GB TruDDR4 2,666 MHz RDIMMS and 5 x ThinkSystem 2.5" PM1633a 3.84 TB capacity SAS 12 Gb hot-swap solid-state drives (SSDs) for SAP HANA storage. The operating system is SUSE Linux Enterprise Server 12 SP3 and uses SAP HANA 2.0 SPS 03 with a 6 TB dataset. Average start time for all data finished after table preload for 10 iterations: 50 minutes.

New configuration with a combination of DRAM and Intel Optane PMem: Lenovo ThinkSystem SR950 server with 8 x Intel Xeon Platinum 8176M processors (28 cores, 165 watt, 2.1 GHz). Total memory consists of 48 x 16 GB TruDDR4 2,666 MHz RDIMMs and 48 x 128 GB Intel Optane PMem modules, and 5 x ThinkSystem 2.5" PM1633a 3.84 TB capacity SAS 12 Gb hot-swap solid-state drives (SSDs) for SAP HANA storage. The operating system is SUSE Linux Enterprise Server 12 SP3 and uses SAP HANA 2.0 SPS 03 with a 6 TB dataset. Average start time for all data finished after table preload for 10 iterations: 4 minutes (12.5x improvement).

[20] Intel. "Baidu Feed Stream Services Restructures Its In-Memory Database with Intel® Optane™ Technology." https://newsroom.intel.com/wp-content/uploads/sites/11/2019/08/baidu-feed-case-study.pdf.

[21] Baidu needed caching disks to continuously deliver sequential data to tape drive libraries. Six NAND drives delivered 36.75 PBW, while two Intel Optane SSDs delivered 164 PBW. Source: Intel. "Improve tape-backup speeds by caching with Intel® Optane™ SSDs." July 2020. intel.com/content/www/us/en/products/docs/storage/baidu-improves-tape-backup-case-study.html.

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 www.intel.com/benchmarks.

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**.

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.

Your costs and results may vary.

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

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

Printed in USA 0920/SM/PRW/PDF Please Recycle 343925-001US