Brief

Intel® Data-Centric Technologies
Cloud Services

This Is Google Cloud Enabled by Intel

Google Cloud

Cloud computing is a fact of life for all industries. The cloud is a critical business platform involving and transforming people, processes, and applications. Google Cloud and Intel have developed a variety of resources and technologies for high-performance computing (HPC), big data, artificial intelligence (AI), and connecting with the edge to meet the needs of organizations today and tomorrow. 2nd Generation Intel® Xeon® Scalable processors provide the foundation of the cloud computing services deployed by Google Cloud. Google Cloud instances enabled by Intel processors have the largest breadth, global reach, and availability of compute instances across geographies (24 regions and 73 availability zones),[1] tailored to meet all your diverse computing needs. Whether deploying a hybrid cloud with Google Cloud Anthos, SAP HANA in an M2 instance, or HPC with a C2 instance, Google Cloud VMware Engine (GCVE), and Oracle bare metal, Intel architecture is at the heart of Google Cloud computing instances, providing strong data protection, fast processing of large data volumes, and service flexibility, without a hit to performance. Intel processors feature:

- **Intel Advanced Vector Extension 512 (Intel AVX-512)**, which offers accelerated application performance, 2x better than previous-generation technologies, enabling significant improvements in workload and data application speeds.[2]
- **Intel Deep Learning Boost (Intel DL Boost)** offers built-in AI acceleration, with up to 14x better inference performance on image classification in 2nd Generation Intel Xeon Scalable processors, compared to competing processors.[3]
- **Intel Turbo Boost Technology** accelerates processor and graphics performance for peak loads, automatically allowing processor cores to run faster than the rated operating frequency if they're operating below power, current, and temperature specification limits.
- **Intel AES New Instructions (Intel AES-NI)** improves upon the original Advanced Encryption Standard (AES) algorithm to provide faster data protection and greater security. All current-generation Google Cloud instances support this processor feature.

Broad choice of Google Cloud services, from edge to cloud, enabled by Intel

Google Cloud and Intel share a passion for customers and innovation. Together, they are continuously developing new service solution offerings to empower customers to move faster, adopt next-generation technologies, and scale with confidence. Google Cloud has a broad offering of cloud computing solutions, enabled by Intel Xeon processors, that are designed to manage costs and complexity, accelerate business outcomes, and scale to meet current and future computing requirements.

Google Cloud N2—general purpose

N2 instances are the latest generation of general-purpose computing instances powered by 2nd Generation Intel Xeon Scalable processors. They are ideal for applications that can take advantage of improved network throughput and packet-rate performance. These instances provide a balance of compute, memory, and network resources, and they are a good choice for many applications.

> The performance difference with Google Cloud Platform was amazing—superior to the others by a factor of four. Combined with security features in Google and better pricing, it was clearly the best strategic technology choice.
>
> — Andreas Pettersson, CEO, Arcules

Use cases

Small and mid-size databases, data processing tasks that require additional memory, caching fleets, and for running back-end servers for SAP, Microsoft SharePoint, cluster computing, and other enterprise applications. These instances are also good for web and application servers, small and mid-sized databases, cluster computing, gaming servers, and other enterprise applications.

Features

- N2 instances are powered by 2nd Generation Intel Xeon Scalable processors with up to 3.9 GHz turbo frequency
- 30 percent better price/performance compared to N1 instances[4]—the best price/performance for a wide range of applications
- Multiple shapes and options with 4 to 60 vCPUs and 16 to 240 GB (4:1 GB to vCPU ratio)
- Intel AVX-512
- Intel Turbo Boost Technology, delivering frequencies up to 3.1 GHz
- Support for Intel DL Boost Vector Neural Network Instructions (VNNI), which helps speed up typical machine learning (ML) operations like convolution and automatically improves inference performance over a wide range of deep learning (DL) workloads

Google Cloud C2—compute optimized

Compute-optimized instances are ideal for compute-bound applications that benefit from high-performance Intel Xeon Scalable processors. Instances belonging to this family are well suited for batch processing workloads, media transcoding, high-performance web servers, HPC, scientific modeling, dedicated gaming servers, ad server engines, ML inference, and other compute-intensive applications, delivering 40 percent more performance per vCPU than N1 instances.[5] C2 instances offer up to 3.4x better performance per dollar.[6]

"Blue Skies Ahead: ClimaCell Delivers Innovative Weather Prediction Solutions": intel.com/content/dam/www/public/us/en/ai/documents/climacell-case-study.pdf

Use cases

HPC, batch processing, ad serving, video encoding, gaming, scientific modelling, distributed analytics, and CPU-based ML inference.

Features

- C2 instances are powered by 2nd Generation Intel Xeon Scalable processors
- Highest performance/core on Google Cloud
- Multiple shares and options from 4 to 60 vCPUs and 16 to 240 GB (4:1 GB to vCPU ratio)
- Intel AVX-512
- Intel Turbo Boost Technology, delivering frequencies up to 3.9 GHz—the fastest per-core speed available on Google Cloud
- Support for Intel DL Boost VNNI, which helps speed up typical ML operations like convolution and automatically improves inference performance over a wide range of DL workloads

Google Cloud M2—memory optimized

M2 instances deliver 3x more memory capacity per virtual machine (VM) than M1 instances,[7] and they are ideal for memory-bound workloads including high-performance databases, distributed web-scale in-memory caches, mid-sized in-memory databases, real-time big data analytics, and other enterprise applications.

Use cases

M2 instances are well suited for memory-intensive applications such as high-performance databases, distributed web-scale in-memory caches, mid-size in-memory databases, real-time big data analytics, and other enterprise applications.

Features

- M2 instances are powered by 2nd Generation Intel Xeon Scalable processors with up to 3.8 GHz turbo frequency
- New, larger memory VMs, with 12 TB per VM—3x more than M1[7]
- Up to 416 and 12 TB

> Moving to a 12 TB virtualized environment with the help of Google Cloud is going to provide us with a better platform for growth as we look to optimize and scale. It's been a great partnership.
>
> — Duy Trinh, SAP Center of Excellence, Cardinal Health

- High GB per core, with up to 29 GB per vCPU
- Only Google Cloud instances certified for SAP[8]
- Intel AVX-512
- Intel Turbo Boost Technology, delivering frequencies up to 3.1 GHz
- Support for Intel DL Boost VNNI, which helps speed up typical ML operations like convolution and automatically improves inference performance over a wide range of DL workloads

Google Cloud Anthos (hybrid cloud)

With Google Cloud Anthos powered by Intel Xeon Scalable processors, you can extend Google Cloud services with N2, M2, and C2 instances on premises. You can quickly and easily deploy and manage apps and workloads using VMs or Kubernetes containers across your on-premises environment, on Google Cloud, or to your preferred cloud provider. Simplify deployment and configuration of Google Cloud Anthos with Intel Select Solutions for Google Cloud's Anthos—hyperconverged solutions built on a familiar VMware stack and tools and designed to integrate seamlessly with Google Cloud. Intel Select Solutions for Google Cloud's Anthos allow users to run concurrent database and analytics jobs without sacrificing application responsiveness and job throughput. The solutions include Google Kubernetes Engine (GKE) On-Prem on VMware vSphere using the VMware ESXi private cloud deployment model. They also make use of VMware vSAN storage, but they can support remote storage as an alternative to a hyperconverged infrastructure (HCI) deployment.

Intel Select Solutions for Google Cloud's Anthos are built on Intel hardware technologies that are part of the VMware hardware-compatibility list. In addition, the solutions are built using Intel-verified platforms and reference designs for vSAN, following tight specifications from Intel and VMware. Intel Select Solutions for Google Cloud's Anthos are also tested and verified by Intel for balanced and optimized performance—from hardware up through the firmware stack to the vSAN software and GKE On-Prem.

"Key Bank Modernizes Applications with Google Cloud and Cisco HyperFlex": youtube.com/watch?v=AXnpYt60rjg#action=share

Use cases

A unified management console across multi-cloud environments, blending on-premises and cloud infrastructure, in addition to VM migration to a public cloud.

Features

- Performance-optimized specifically for vSAN
- Reduce the time required to evaluate, select, and purchase the necessary hardware components
- Minimize the time required to deploy new infrastructure
- Deliver performance optimized to a specific threshold across compute, storage, and network on trusted Intel architecture
- Benefit from years of engineering partnership between Intel and Google

Intel has rigorously benchmark-tested configurations that are optimized specifically for the performance of Anthos. This means Google's top OEMs and system integrators (SIs) can accelerate their deployments of Anthos by verifying against minimum performance guarantees, as established by Intel. Partners such as Lenovo, World Wide Technology (WWT), and Redapt have Intel Select Solutions for Google's Cloud Anthos designs available to accelerate time to deployment, reduce risk, and improve the operational efficiency of any single migration.

Google Cloud Bare Metal Solution

Google Cloud Bare Metal Solution is designed for the performance and high-availability needs of mission-critical, enterprise-grade applications. It offers state-of-the-art dedicated servers, based on 2nd Generation Intel Xeon Scalable processors, that come in a variety of sizes. Depending on your needs, you can choose a bare-metal server with as few as 16 cores, or all the way up to 112 cores with 3 TB of DRAM, all to handle your most demanding workloads. These servers running 2nd Generation Intel Xeon Scalable processors are certified by a broad ecosystem of software companies. Google Cloud Bare Metal Solution deploys in a region extension with less than two millisecond latency to Google Cloud; in most cases, the latency is measured to be sub-millisecond.[9]

Bare-metal solutions

Google Cloud VMware Engine (GCVE)

GCVE, optimized on 2nd Generation Intel Xeon Scalable processors, provides everything you need to run your VMware environment natively in Google Cloud. The service delivers a fully managed VMware Cloud Foundation hybrid cloud platform, including VMware vSphere, VMware vCenter, VMware vSAN, VMware NSX-T, and VMware HCX—in a dedicated environment on the high-performance and reliable infrastructure of Google Cloud to support your

Google can reduce your VMware environment TCO by 25% or more[10]

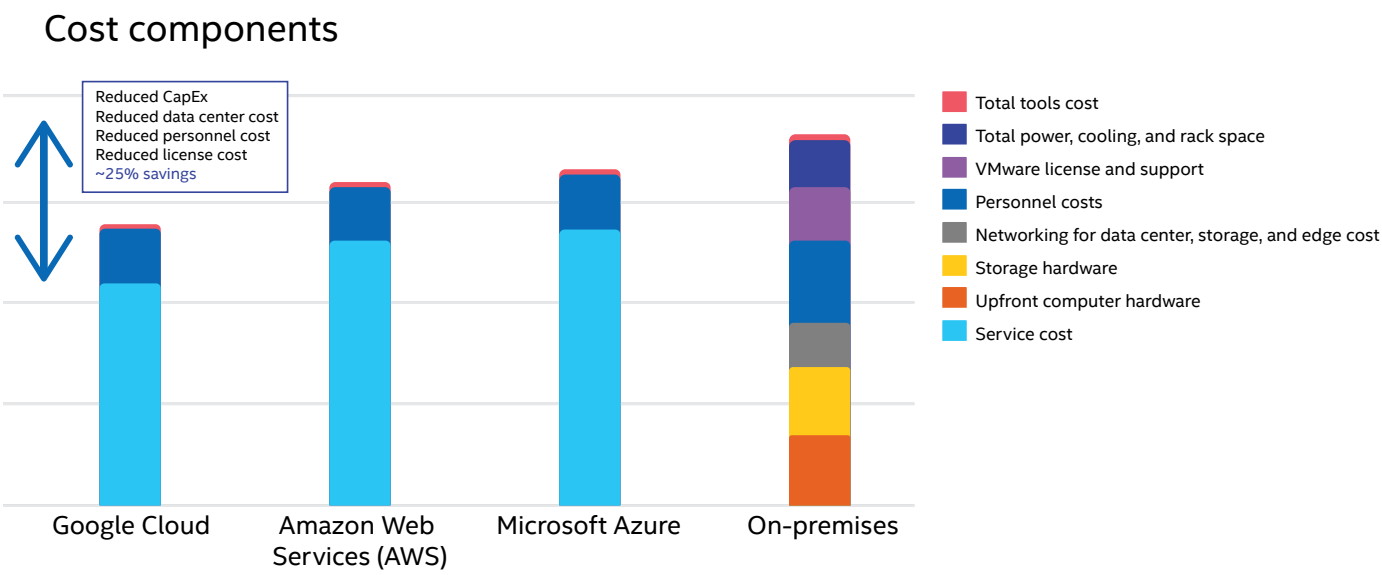

Figure 1. Google Cloud total cost of ownership (TCO) analysis[10]

enterprise production workloads. Seamlessly extend or bring your on-premises workloads to Google Cloud in minutes—and without changes—by connecting to a dedicated VMware environment enabled by Intel architecture.

Features

- Intel with GCVE enables the vast majority of customers running on-premises VMware workloads to seamlessly migrate to Google Cloud
- GCVE means reducing the TCO of a VMware environment by 25 percent or more[10]
- Larger machines with 72 Intel vCPUs and ample memory per solid state drive (SSD) enable lower TCO per VM
- Intel-based machines are verified and certified by VMware
- Available in 10 worldwide regions by the end of 2020

Oracle bare metal

Oracle bare metal can streamline the process for applications that require an Oracle database. It can also provide hardware to run specialized workloads with low latency on Google Cloud. With Oracle bare metal, you can reduce overhead, drive innovation, and increase agility with a wide variety of databases for your Oracle workloads. From bare-metal solutions to cloud-native databases, Google Cloud offers infrastructure to meet your needs.

- Oracle bare metal supports all Oracle Database options, including Oracle Real Application Clusters (RAC).
- Servers can be configured with different operating systems, such as Windows or Linux.
- Oracle bare metal can be used for consolidation of databases.
- Oracle VMs and Oracle Logical Volume Management (LVM) can be used in Oracle bare metal. Both options are known as hard partitioning within the Oracle licensing frame. This means that specific enterprise options can be restricted to specific VMs within the Oracle environment. This again means that, only for this restricted part, a higher license fee will be taken into account (rather than for the whole server).
- Older applications and operating systems need to run on VMs. More precisely, the following are not being supported by processors and need to run on VMs: all database version before 11204 or certified with Red Hat 6 need to run on VMs.
- The first migration to an Oracle bare metal offering with Google Cloud needs around 8–10 weeks.

Features

- Certified solution with Oracle
- Run old versions of Oracle databases, packaged apps, and the hypervisor of your choice
- Reduce Oracle cloud deployment risk by running on Oracle-certified Intel hardware
- Select from four different Intel-enabled compute machine types—small, medium, large, X-large—enabling customers to choose the best fit
- Fully managed data center, with enterprise-grade security and reliability
- Built-in Intel performance acceleration and optimizations for Oracle to support migration and accelerate time to deployment

Table 1. Product table (intel.com/content/www/us/en/artificial-intelligence/google-cloud-platform.html)

INSTANCE TYPE	C2 COMPUTE OPTIMIZED	N2 GENERAL PURPOSE	M2-ULTRAMEM AND MEGA-MEM MEMORY OPTIMIZED	N1 GENERAL PURPOSE	M1-ULTRAMEM MEMORY OPTIMIZED	M1-MEGAMEM MEMORY OPTIMIZED	GOOGLE CLOUD VMWARE ENGINE
TARGET WORKLOAD	Apps needing single-thread frequency (such as HPC, electronic design automation [EDA], and gaming)	Mainstream	High-memory database (such as SAP HANA)	Mainstream	High-memory database (such as SAP HANA)	High-memory database (such as SAP HANA)	VMware-based instances
CPU	2nd Gen Intel Xeon Scalable processor	2nd Gen Intel Xeon Scalable processor	2nd Gen Intel Xeon Scalable processor	Various Intel Xeon processors, including Intel Xeon Scalable processors	Intel Xeon processor E7 v4	Intel Xeon Scalable processor	2nd Gen Intel Xeon Scalable processor
VCPUS	Up to 60	Up to 80	208/416	Up to 96	40/80/160	96	72
MEMORY	Up to 240 GB	Up to 624 GB	6 TB/12 TB	Up to 642 GB	1 TB/2 TB/4 TB	1.4 TB	1 TB
ALL-CORE TURBO BOOST SPEED	3.8 GHz (sustained)	3.4 GHz	3.4 GHz	2.7 GHz (Intel Xeon Scalable processor)	2.6 GHz	2.7 GHz	3.9 GHz
BASELINE CLOUD SPEED	3.1 GHz	2.8 GHz	2.5 GHz	2.0 GHz (Intel Xeon Scalable processor)	2.2 GHz	2.0 GHz	2.6 GHz
INTEL AVX-512 (2X FLOATING PT)	Yes	Yes	Yes	Intel Xeon Scalable processor (using MinCPU platform)	No	Yes	Yes
INTEL DEEP LEARNING BOOST	Yes	Yes	Yes	No	No	No	Yes
AVAILABLE IN GCP REGIONS	12	13	5	24	17	10	2
CAN TAKE-IN AVX-512 AND VNNI INSTRUCTIONS	Yes	Yes		Yes			

Learn More

"Today's top clouds are powered by Intel" white paper: **intel.com/content/www/us/en/cloud-computing/top-clouds-powered-by-intel-paper.html**

Intel cloud computing: **intel.com/cloud**

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

Intel and Google Cloud webpage: **https://cloud.google.com/intel**

[1] No other architecture is available in more Google Cloud Platform geographies than the Intel architecture in the C2 instance. For details, see the geographies page at: https://cloud.google.com/compute/docs/regions-zones#available.

[2] FSI kernels: B**aseline: Intel Xeon Platinum 8268 processor configuration**: Intel "Wolf Pass" platform with 2-socket Intel Xeon Platinum 8268 processors (2.9 GHz, 24 cores), 12 x 16 GB DDR4-2,933, 1 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), AVX2_256 build, Intel Compiler 2019u5, Intel Math Kernel Library (Intel MKL) 2019u5, BIOS: Binomial (Intel Hyper-Threading Technology [Intel HT Technology] on, Intel Turbo Boost Technology on, SNC off, 2 threads/core), Black Scholes (Intel HT Technology off, Intel Turbo Boost Technology on, SNC off, 1 threads/core), Monte Carlo (Intel HT Technology on, Intel Turbo Boost Technology on, SNC off, 2 threads/core). Test by Intel as of 11/1/2019. **With AVX-512: Intel Xeon Platinum 8268 processor configuration**: Intel "Wolf Pass" platform with 2-socket Intel Xeon Platinum 8268 processors (2.9 GHz, 24 cores), 12 x 16 GB DDR4-2,933, 1 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), AVX-512 build, Intel Compiler 2019u5, Intel Math Kernel Library (Intel MKL) 2019u5, BIOS: BIOS: Binomial (Intel HT Technology on, Intel Turbo Boost Technology on, SNC off, 2 threads/core), Black Scholes (Intel HT Technology off, Intel Turbo Boost Technology on, SNC off, 1 threads/core), Monte Carlo (Intel HT Technology on, Intel Turbo Boost Technology on, SNC off, 2 threads/core). Test by Intel as of 11/1/2019.

[3] Up to 14x AI performance improvement with Intel DL Boost compared to Intel Xeon Platinum 8180 processor (July 2017). Tested by Intel as of 2/20/2019. 2 socket Intel Xeon Platinum 8280 processor, 28 cores, Intel HT Technology on, Intel Turbo Boost Technology on, 384 GB total memory (12 slots, 32 GB, 2,933 MHz), BIOS: SE5C620.86B.0D.01.0271.120720180605 (ucode: 0x200004d), Ubuntu 18.04.1 LTS, kernel4.15.0-45-generic, SSD 1 x sda INTEL SSDSC2BA80 745.2 GB, nvme1n1 INTEL SSDPE2KX040T7 3.7 TB, deep learning framework: Intel Optimization for Caffe version: 1.1.3 (commit hash: 7010334f159da247db3fe3a9d96a3116ca06b09a) , ICC version 18.0.1, Intel Math Kernel Library for Deep Neural Networks (Intel MKL-DNN) version: v0.17 (commit hash: 830a10059a018cd2634d94195140cf2d8790a75a, model: https://github.com/intel/caffe/blob/master/models/intel_optimized_models/int8/resnet50_int8_full_conv.prototxt, BS=64, DummyData, 4 instance/2 socket, datatype: INT8 vs. tested by Intel as of July 11, 2017: 2-socket Intel Xeon Platinum 8180 processor at 2.50 GHz (28 cores), Intel HT Technology disabled, Intel Turbo Boost Technology disabled, scaling governor set to "performance" via intel_pstate driver, 384 GB DDR4-2,666 ECC RAM. CentOS Linux release 7.3.1611 (core), Linux kernel 3.10.0-514.10.2.el7.x86_64. SSD: Intel SSD DC S3700 (800 GB, 2.5-in SATA 6 Gb/s, 25 nm, MLC). Performance measured with: environment variables: KMP_AFFINITY='granularity=fine, compact', OMP_NUM_THREADS=56, CPU frequency set with cpupower frequency-set -d 2.5G -u 3.8G -g performance. Caffe: (http://github.com/intel/caffe/), revision f96b759f71b2281835f690af267158b82b150b5c. Inference measured with "caffe time --forward_only" command, training measured with "caffe time" command. For "ConvNet" topologies, dummy dataset was used. For other topologies, data was stored on local storage and cached in memory before training. Topology specs from https://github.com/intel/caffe/tree/master/models/intel_optimized_models (ResNet-50), Intel C++ compiler ver. 17.0.2 20170213, Intel MKL small libraries version 2018.0.20170425. Caffe run with "numactl –l."

[4] N2 instances: For more information, see: Google. "Expanding virtual machine types to drive performance and efficiency." August 2019. https://cloud.google.com/blog/products/compute/expanding-virtual-machine-types-to-drive-performance-and-efficiency.

[5] "40 percent more performance per vCPU" claim: 2.8 GHz N2 vs. 2.0 GHz N1 base frequency. For more information, see: Google. "Expanding virtual machine types to drive performance and efficiency." August 2019. https://cloud.google.com/blog/products/compute/expanding-virtual-machine-types-to-drive-performance-and-efficiency.

[6] Compute-optimized Intel Xeon processor–based C2 instance provides 3.43x better performance/dollar compared to N2D instances based on competing processors for HPC applications. The highest instance sized compared: C2-standard-60 vs. N2D-standard-224 for LAMMPS Polyethelene (AIREBO) workload. Compute-optimized: C2 instances. For more information, see: Google. "Expanding virtual machine types to drive performance and efficiency." August 2019. https://cloud.google.com/blog/products/compute/expanding-virtual-machine-types-to-drive-performance-and-efficiency.

[7] Memory-optimized M2 instance: Comparing M1-UltraMem-160 to M2-UltraMem-416. For more information, see: Google. "Machine types." https://cloud.google.com/compute/docs/machine-types.

[8] SAP. "Find Certified IaaS Platforms." August 2020. sap.com/dmc/exp/2014-09-02-hana-hardware/enEN/iaas.html#categories=Google%20Cloud%20Platform.

[8] Google. "Bare Metal Solution: Coming to a Google Cloud data center near you." June 2020. https://cloud.google.com/blog/products/compute/bare-metal-solution-comes-to-five-new-google-cloud-regions.

[10] Example calculation based on: 200 VMs, 8 hosts, 2 full-time employees (FTE) needed to operate on-premises infrastructure.

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.**

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 1020/SM/PRW/PDF Please Recycle 344061-001US