Google Cloud

NVIDIA Tesla V100: Known for its high performance and versatility, the Tesla V100 is widely used in deep learning and AI workloads. It offers high computational power and memory bandwidth, making it suitable for demanding tasks.

NVIDIA Tesla P100: Another powerful GPU suitable for deep learning tasks, the Tesla P100 offers excellent performance and is commonly used in cloud environments for machine learning workloads.

NVIDIA Tesla T4: Designed for inference workloads, the Tesla T4 offers a good balance of performance and cost-effectiveness. It's often used for deploying machine learning models in production environments where real-time inference is required.

NVIDIA A100 Tensor Core GPU: This is NVIDIA's latest GPU architecture optimized for AI and machine learning workloads. It offers significant performance improvements over previous generations and is available in various cloud platforms.

AMD Radeon Instinct MI100: While NVIDIA dominates the market, AMD's Radeon Instinct MI100 is also a powerful GPU option for machine learning tasks. It offers high performance and is used in some cloud environments.

Here's the information presented in a tabular format:

| Zones | Location | GPU platforms | NVIDIA RTX Virtual Workstations (vWS) |

|-----------------------------------------------|-------------------------------------------|--------------------------------------------------|---------------------------------------|

| africa-south1-a | Johannesburg, South Africa | | |

| africa-south1-b | | | |

| africa-south1-c | | | |

| asia-east1-a | Changhua County, Taiwan, APAC | L4, T4, P100, K80 | L4, T4, P100 |

| asia-east1-b | | L4, K80 | L4 |

| asia-east1-c | | L4, T4, V100, P100 | L4, T4, P100 |

| asia-east2-a | Hong Kong, APAC | T4 | T4 |

| asia-east2-b | | | |

| asia-east2-c | | T4 | T4 |

| asia-northeast1-a | Tokyo, Japan, APAC | A100 40GB, L4, T4 | L4, T4 |

| asia-northeast1-b | | | |

| asia-northeast1-c | | A100 40GB, L4, T4 | L4, T4 |

| asia-northeast2-a | Osaka, Japan, APAC | | |

| asia-northeast2-b | | | |

| asia-northeast2-c | | | |

| asia-northeast3-a | Seoul, South Korea, APAC | A100 40GB, L4 | L4 |

| asia-northeast3-b | Seoul, South Korea, APAC | A100 40GB, L4, T4 | L4, T4 |

| asia-northeast3-c | Seoul, South Korea, APAC | T4 | T4 |

| asia-south1-a | Mumbai, India, APAC | L4, T4 | L4, T4 |

| asia-south1-b | Mumbai, India, APAC | L4, T4 | L4, T4 |

| asia-south1-c | Mumbai, India, APAC | | |

| asia-southeast1-a | Jurong West, Singapore, APAC | L4, T4 | L4, T4 |

| asia-southeast1-b | Jurong West, Singapore, APAC | H100 80GB, A100 40GB, L4, T4, P4 | L4, T4, P4 |

| asia-southeast1-c | Jurong West, Singapore, APAC | H100 80GB, A100 80GB, A100 40GB, L4, T4, P4 | L4, T4, P4 |

| asia-southeast2-a | Jakarta, Indonesia, APAC | T4 | T4 |

| asia-southeast2-b | Jakarta, Indonesia, APAC | | |

| asia-southeast2-c | Jakarta, Indonesia, APAC | T4 | T4 |

| australia-southeast1-a | Sydney, Australia, APAC | T4, P4 | T4, P4 |

| australia-southeast1-b | Sydney, Australia, APAC | P4 | P4 |

| australia-southeast1-c | Sydney, Australia, APAC | T4, P100 | T4, P100 |

| europe-north1-a | Hamina, Finland, Europe | | |

| europe-north1-b | | | |

| europe-north1-c | | | |

| europe-central2-a | Warsaw, Poland, Europe | | |

| europe-central2-b | Warsaw, Poland, Europe | T4 | T4 |

| europe-central2-c | Warsaw, Poland, Europe | T4 | T4 |

| europe-west1-b | St. Ghislain, Belgium, Europe | L4, T4, P100, K80 | L4, T4, P100 |

| europe-west1-c | St. Ghislain, Belgium, Europe | L4, T4 | L4, T4 |

| europe-west1-d | St. Ghislain, Belgium, Europe | P100, K80, T4 | P100, T4 |

| europe-west2-a | London, England, Europe | L4, T4 | L4, T4 |

| europe-west2-b | London, England, Europe | L4, T4 | L4, T4 |

| europe-west2-c | London, England, Europe | | |

| europe-west3-a | Frankfurt, Germany, Europe | | |

| europe-west3-b | Frankfurt, Germany, Europe | L4, T4 | L4, T4 |

| europe-west3-c | Frankfurt, Germany, Europe | | |

| europe-west4-a | Eemshaven, Netherlands, Europe | A100 80GB, A100 40GB, L4, T4, V100, P100 | L4, T4, P100 |

| europe-west4-b | Eemshaven, Netherlands, Europe | H100 80GB, A100 40GB, L4, T4, P4, V100 | L4, T4, P4 |

| europe-west4-c | Eemshaven, Netherlands, Europe | H100 80GB, L4, T4, P4, V100 | L4, T4, P4 |

| europe-west6-a | Zurich, Switzerland, Europe | L4 | L4 |

| europe-west6-b | Zurich, Switzerland, Europe | L4 | L4 |

| europe-west6-c | Zurich, Switzerland, Europe | | |

| europe-west8-a | Milan, Italy, Europe | | |

| europe-west8-b | Milan, Italy, Europe | | |

| europe-west8-c | Milan, Italy, Europe | | |

| europe-west9-a | Paris, France, Europe | | |

| europe-west9-b | Paris, France, Europe | | |

| europe-west9-c | Paris, France, Europe | | |

| europe-west10-a | Berlin, Germany, Europe |

**Limitations**

For A100 40GB GPUs, the a2-megagpu-16g machine type is only available in the following regions and zones:

* Iowa, North America: us-central1-a,b,c,f
* Netherlands, Europe: europe-west4-a,b
* Singapore, APAC: asia-southeast1-c

**Local SSD availability by GPU regions and zones**

Local SSD disks are available as follows for each GPU type:

* NVIDIA H100 80GB: Local SSD disks are supported in all the available H100 80GB regions and zones.
* NVIDIA L4: Local SSD disks are supported in all the available L4 regions and zones.
* NVIDIA A100 80GB: Local SSD disks are bundled with the machine series in all available A100 80GB regions and zones.
* NVIDIA A100 40GB: Local SSD disks are supported in all the available A100 40GB regions and zones.
* NVIDIA T4: Local SSD disks are supported in all the available T4 regions and zones.
* NVIDIA V100: Local SSD disks are supported in most of the available V100 regions and zones with the exception of the following:
  + us-east1-c
* NVIDIA P100: Local SSD disks are supported in all the available P100 regions and zones.
* NVIDIA P4: Local SSD disks are only supported in zones us-central1-c and northamerica-northeast1-b.
* NVIDIA K80: Local SSD disks are supported in all the available K80 regions and zones.

CPU AVAILABILITY:

Here is the data presented in tabular format without truncation:

| Zones | Location | Machine types | CPUs | Resources | CO2 emissions |

|---------------------|-----------------------------|--------------------------------------------|----------------------------------------------------------------------------|-----------------------------------------------------|---------------|

| asia-east1-a | Changhua County, Taiwan, APAC | E2, N2, N2D, T2D, N1, M1, C2, C2D, G2 | Intel Ivy Bridge, Sandy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | GPUs | |

| asia-east1-b | Changhua County, Taiwan, APAC | E2, N2, N2D, T2D, N1, M1, M3, C2, C2D, G2 | Intel Ivy Bridge, Sandy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | GPUs | |

| asia-east1-c | Changhua County, Taiwan, APAC | E2, N2, N2D, T2D, N1, M1, M3, C2, C2D, G2 | Intel Ivy Bridge, Sandy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | GPUs | |

| asia-east2-a | Hong Kong, APAC | E2, N2, N2D, T2D, N1, C2 | Intel Ivy Bridge, Sandy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, AMD EPYC Rome, AMD EPYC Milan | GPUs | |

| asia-east2-b | Hong Kong, APAC | E2, N2, N2D, T2D, N1, C2 | Intel Ivy Bridge, Sandy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | | |

| asia-east2-c | Hong Kong, APAC | E2, N2, N2D, T2D, N1, C2 | Intel Ivy Bridge, Sandy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | GPUs | |

| asia-northeast1-a | Tokyo, Japan, APAC | E2, N2, N2D, T2D, N1, M1, M2, C2, A2, G2 | Intel Ivy Bridge, Sandy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | GPUs | |

| asia-northeast1-b | Tokyo, Japan, APAC | E2, N2, N2D, T2D, N1, M1, M2, C2 | Intel Ivy Bridge, Sandy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | | |

| asia-northeast1-c | Tokyo, Japan, APAC | E2, N2, N2D, T2D, N1, M1, C2, A2, G2 | Intel Ivy Bridge, Sandy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | GPUs | |

| asia-northeast2-a | Osaka, Japan, APAC | E2, N2, N2D, T2D, N1, M1, C2 | Intel Ivy Bridge, Sandy Bridge, Broadwell, Skylake, Cascade Lake, AMD EPYC Rome, AMD EPYC Milan | | |

| asia-northeast2-b | Osaka, Japan, APAC | E2, N2, N2D, T2D, N1, M1, M2, C2 | Intel Ivy Bridge, Sandy Bridge, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | | |

| asia-northeast2-c | Osaka, Japan, APAC | E2, N2, N2D, T2D, N1, M1, M2, C2 | Intel Ivy Bridge, Sandy Bridge, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | | |

| asia-northeast3-a | Seoul, South Korea, APAC | E2, N2, N2D, N1, M1, M2, C2, A2, G2 | Intel Ivy Bridge, Sandy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | GPUs | |

| asia-northeast3-b | Seoul, South Korea, APAC | E2, N2, N2D, N1, M1, M2, C2, A2, G2 | Intel Ivy Bridge, Sandy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | GPUs | |

| asia-northeast3-c | Seoul, South Korea, APAC | E2, N2, N2D, N1, C2 | Intel Ivy Bridge, Sandy Bridge, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | GPUs | |

| asia-south1-a | Mumbai, India, APAC | E2, N2, N2D, T2D, N1, M1, M2, C2, C2D, G2 | Intel Ivy Bridge, Sandy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | GPUs | |

| asia-south1-b | Mumbai, India, APAC | E2, N2, N2D, T2D, N1, M1, M2, C2, C2D, G2 | Intel Ivy Bridge, Sandy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | GPUs | |

| asia-south1-c | Mumbai, India, APAC | E2, N2, N2D, T2D, N1, M1, C2, C2D | Intel Ivy Bridge, Sandy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | | |

| asia-south2-a | Delhi, India, APAC | E2, N2, N2D, T2D, N

| northamerica-northeast1-a | Montréal, Québec, North America | E2, N2, N2D, T2D, N1, C2 | Intel Ivy Bridge, Sandy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | GPUs | leaf icon Low CO2 |

| northamerica-northeast1-b | Montréal, Québec, North America | E2, N2, N2D, T2D, N1, M1, M2, M3, C2 | Intel Ivy Bridge, Sandy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | GPUs | leaf icon Low CO2 |

| northamerica-northeast1-c | Montréal, Québec, North America | E2, N2, N2D, T2D, N1, M1, M2, M3, C2 | Intel Ivy Bridge, Sandy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | GPUs | leaf icon Low CO2 |

| northamerica-northeast2-a | Toronto, Ontario, North America | E2, N2, N2D, T2D, N1, M1, M2, M3 | Intel Ivy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | | leaf icon Low CO2 |

| northamerica-northeast2-b | Toronto, Ontario, North America | E2, N2, N2D, T2D, N1, M1, M2, M3 | Intel Ivy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | | leaf icon Low CO2 |

| northamerica-northeast2-c | Toronto, Ontario, North America | E2, N2, N2D, T2D, N1 | Intel Ivy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | | leaf icon Low CO2 |

| us-central1-a | Council Bluffs, Iowa, North America | E2, N2, N2D, C3, C3D, T2D, T2A, N1, Z3 (Preview), M1, M2, M3, H3, C2, C2D, A3, A2, G2 | Intel Ivy Bridge, Sandy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, Sapphire Rapids, AMD EPYC Rome, AMD EPYC Milan, AMD EPYC Genoa, Ampere Altra Arm | GPUs | leaf icon Low CO2 |

| us-central1-b | Council Bluffs, Iowa, North America | E2, N2, N2D, C3, C3D, T2D, T2A, N1, M1, M2, M3, C2, C2D, A2, G2 | Intel Ivy Bridge, Sandy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan, AMD EPYC Genoa, Sapphire Rapids, Ampere Altra Arm | GPUs | leaf icon Low CO2 |

| us-central1-c | Council Bluffs, Iowa, North America | E2, N2, N2D, C3, C3D, T2D, N1, Z3 (Preview), M1, M2, C2, C2D, A3, A2, G2 | Intel Ivy Bridge, Sandy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, Sapphire Rapids, AMD EPYC Rome, AMD EPYC Milan, AMD EPYC Genoa | GPUs | leaf icon Low CO2 |

| us-central1-f | Council Bluffs, Iowa, North America | E2, N2, N2D, T2D, T2A, N1, M1, M3, C2, C2D, A2 | Intel Ivy Bridge, Sandy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan, Ampere Altra Arm | GPUs | leaf icon Low CO2 |

| us-east1-b | Moncks Corner, South Carolina, North America | E2, N2, N2D, C3, C3D, T2D, N1, M1, M3, C2, C2D, A2, G2 | Intel Ivy Bridge, Sandy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, Sapphire Rapids, AMD EPYC Rome, AMD EPYC Milan, AMD EPYC Genoa | GPUs | |

| us-east1-c | Moncks Corner, South Carolina, North America | E2, N2, N2D, C3, C3D, T2D, N1, M1, C2, C2D, G2 | Intel Ivy Bridge, Sandy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, Sapphire Rapids, AMD EPYC Rome, AMD EPYC Milan, AMD EPYC Genoa | GPUs | |

| us-east1-d | Moncks Corner, South Carolina, North America | E2, N2, N2D, C3, C3D, T2D, N1, M1, C2, C2D, G2 | Intel Ivy Bridge, Sandy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, Sapphire Rapids, AMD EPYC Rome, AMD EPYC Milan, AMD EPYC Genoa | GPUs | |

| us-east4-a | Ashburn, Virginia, North America | E2, N2, N2D, C3, C3D, T2D, N1, M1, M2, M3, C2, C2D, A3, G2 | Intel Ivy Bridge, Sandy Bridge, Haswell,

Broadwell, Skylake, Cascade Lake, Ice Lake, Sapphire Rapids, AMD EPYC Rome, AMD EPYC Milan, AMD EPYC Genoa | GPUs | |

| us-east4-b | Ashburn, Virginia, North America | E2, N2, N2D, C3, C3D, T2D, N1, M1, M2, M3, C2, C2D | Intel Ivy Bridge, Sandy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, Sapphire Rapids, AMD EPYC Rome, AMD EPYC Milan, AMD EPYC Genoa | GPUs | |

| us-east4-c | Ashburn, Virginia, North America | E2, N2, N2D, C3, C3D, T2D, N1, M1, C2, C2D, A2, A3 | Intel Ivy Bridge, Sandy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, Sapphire Rapids, AMD EPYC Rome, AMD EPYC Milan, AMD EPYC Genoa | GPUs | |

| us-east5-a | Columbus, Ohio, North America | E2, N2, N2D, T2D, C2, C3, A3 | Intel Broadwell, Haswell, Skylake, Cascade Lake, Ice Lake, Sapphire Rapids, AMD EPYC Rome, AMD EPYC Milan | GPUs | |

| us-east5-b | Columbus, Ohio, North America | E2, N2, N2D, T2D, C2, C3 | Intel Broadwell, Haswell, Skylake, Cascade Lake, Sapphire Rapids, AMD EPYC Rome, AMD EPYC Milan | GPUs | |

| us-east5-c | Columbus, Ohio, North America | E2, N2, N2D, T2D, C2 | Intel Broadwell, Haswell, Skylake, Cascade Lake, AMD EPYC Rome, AMD EPYC Milan | | |

| us-west1-a | The Dalles, Oregon, North America | E2, N2, N2D, T2D, N1, M1, C2, C2D, G2 | Intel Ivy Bridge, Sandy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | GPUs | leaf icon Low CO2 |

| us-west1-b | The Dalles, Oregon, North America | E2, N2, N2D, T2D, N1, C2, C2D, M1, G2 | Intel Ivy Bridge, Sandy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | GPUs | leaf icon Low CO2 |

| us-west1-c | The Dalles, Oregon, North America | E2, N2, N2D, T2D, N1, C2, C2D, G2 | Intel Ivy Bridge, Sandy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | GPUs | leaf icon Low CO2 |

| us-west2-a | Los Angeles, California, North America | E2, N2, N2D, T2D, N1, C2 | Intel Ivy Bridge, Sandy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | | |

| us-west2-b | Los Angeles, California, North America | E2, N2, N2D, T2D, N1, M1, C2 | Intel Ivy Bridge, Sandy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | GPUs | |

| us-west2-c | Los Angeles, California, North America | E2, N2, N2D, N1, M1, C2 | Intel Ivy Bridge, Sandy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | GPUs | |

| us-west3-a | Salt Lake City, Utah, North America | E2, N2, N2D, T2D, N1, C2 | Intel Ivy Bridge, Sandy Bridge, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | | |

| us-west3-b | Salt Lake City, Utah, North America | E2, N2, N2D, T2D, N1, C2, A2 | Intel Ivy Bridge, Sandy Bridge, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | GPUs | |

| us-west3-c | Salt Lake City, Utah, North America | E2, N2, N2D, T2D, N1, C2 | Intel Ivy Bridge, Sandy Bridge, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | | |

| us-west4-a | Las Vegas, Nevada, North America | E2, N2, N2D, T2D, N1, C2, M1, M2, M3, G2 | Intel Ivy Bridge, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | GPUs | |

| us-west4-b | Las Vegas, Nevada, North America | E2, N2, N2D, N1, C2, T2D, M1, M2, M3, A2 | Intel Ivy Bridge, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | GPUs | |

| us-west4-c | Las Vegas, Nevada, North America | E2, N2, N2D, T2D, N1, C2, G2 | Intel Ivy Bridge, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | GPUs | |

| us-south1-a | Dallas, Texas, North America | E2, N2 | Intel Broadwell, Haswell, Skylake, Cascade Lake | | |

| us-south1-b | Dallas, Texas, North America | E2, N2 | Intel Broadwell, Haswell, Skylake, Cascade Lake | | |

| us-south1-c | Dallas, Texas, North America | E2, N2 | Intel Broadwell, Haswell, Skylake, Cascade Lake | | |

| southamerica-east1-a | Osasco, São Paulo, Brazil, South America | E2, N2, N2D, T2D, N1, C2 | Intel Ivy Bridge, Sandy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | GPUs | leaf icon Low CO2 |

| southamerica-east1-b | Osasco, São Paulo, Brazil, South America | E2, N2, N2D, T2D, N1, M1, M2, C2 | Intel Ivy Bridge, Sandy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | | leaf icon Low CO2 |

| southamerica-east1-c | Osasco, São Paulo, Brazil, South America | E2, N2, N2D, T2D, N1, M1, M2, C2 | Intel Ivy Bridge, Sandy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | GPUs | leaf icon Low CO2 |

| southamerica-west1-a | Santiago, Chile, South America | E2, N2, C2 | Intel Broadwell, Haswell, Skylake, Cascade Lake | | leaf icon Low CO2 |

| southamerica-west1-b | Santiago, Chile, South America | E2, N2, M1, M2, C2 | Intel Broadwell, Haswell, Skylake, Cascade Lake | | leaf icon Low CO2 |

| southamerica-west1-c | Santiago, Chile, South America | E2, N2, M1, M2, C2 | Intel Broadwell, Haswell, Skylake, Cascade Lake | | leaf icon Low CO2 |

| europe-north1-a | Hamina, Finland, Europe | E2, N2, N2D, T2D, N1, C2 | Intel Ivy Bridge, Sandy Bridge, Broadwell, Skylake, Cascade Lake, AMD EPYC Rome, AMD EPYC Milan | | leaf icon Low CO2 |

| europe-north1-b | Hamina, Finland, Europe | E2, N2, N2D, T2D, N1, C2 | Intel Ivy Bridge, Sandy Bridge, Broadwell, Skylake, Cascade Lake, AMD EPYC Rome, AMD EPYC Milan | | leaf icon Low CO2 |

| europe-north1-c | Hamina, Finland, Europe | E2, N2, N2D, T2D, N1, C2 | Intel Ivy Bridge, Sandy Bridge, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | | leaf icon Low CO2 |

| europe-central2-a | Warsaw, Poland, Europe | E2, N2, N2D, T2D, N1, M1 | Intel Ivy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | | |

| europe-central2-b | Warsaw, Poland, Europe | E2, N2, N2D, T2D, N1, M1 | Intel Ivy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | GPUs | |

| europe-central2-c | Warsaw, Poland, Europe | E2, N2, N2D, T2D, N1 | Intel Ivy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | GPUs | |

| europe-southwest1-a | Madrid, Spain, Europe | E2, N2, N2D, T2D, M1, M3 | Intel Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | | |

| europe-southwest1-b | Madrid, Spain, Europe | E2, N2, N2D, T2D | Intel Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | | |

| europe-southwest1-c | Madrid, Spain, Europe | E2, N2, N2D, T2D, M1, M3 | Intel Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | | |

| europe-west1-b | St. Ghislain, Belgium, Europe | E2, N2, N2D, C3, C3D, T2D, N1, M1, M2, M3, C2, C2D, G2 | Intel Ivy Bridge, Sandy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, Sapphire Rapids, AMD EPYC Rome, AMD EPYC Milan, AMD EPYC Genoa | GPUs | leaf icon Low CO2 |

| europe-west1-c | St. Ghislain, Belgium, Europe | E2, N2, N2D, C3, C3D, T2D, N1, M2, C2, C2D, G2 | Intel Ivy Bridge, Sandy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, Sapphire Rapids, AMD EPYC Rome, AMD EPYC Milan, AMD EPYC Genoa | GPUs | leaf icon Low CO2 |

| europe-west1-d | St. Ghislain, Belgium, Europe | E2, N2, N2D, C3, C3D, T2D, N1, M1, M2, M3, C2, C2D | Intel Ivy Bridge, Sandy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, Sapphire Rapids, AMD EPYC Rome, AMD EPYC Milan, AMD EPYC Genoa | GPUs | leaf icon Low CO2 |

| europe-west2-a | London, England, Europe | E2, N2, N2D, T2D, N1, M1, M2, C2, G2 | Intel Ivy Bridge, Sandy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | GPUs | leaf icon Low CO2 |

| europe-west2-b | London, England, Europe | E2, N2, N2D, T2D, N1, M1, M2, C2, C2D, G2 | Intel Ivy Bridge, Sandy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | GPUs | leaf icon Low CO2 |

| europe-west2-c | London, England, Europe | E2, N2, N2D, T2D, N1, M1, C2, C2D | Intel Ivy Bridge, Sandy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | | leaf icon Low CO2 |

| europe-west3-a | Frankfurt, Germany, Europe | E2, N2, N2D, T2D, N1, M1, M2, M3, C2, C2D | Intel Ivy Bridge, Sandy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | | leaf icon Low CO2 |

| europe-west3-b | Frankfurt, Germany, Europe | E2, N2, N2D, T2D, N1, M1, M2, M3, C2, C2D, G2 | Intel Ivy Bridge, Sandy Bridge, Haswell, Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | GPUs | leaf icon Low CO2 |

| europe-west3-c | Frankfurt, Germany, Europe | E2, N2, N2D, T2D, N1, M1, C2, C

| me-central1-a | Doha, Qatar, Middle East | E2, N2 | Intel Broadwell, Skylake, Cascade Lake, Ice Lake | | |

| me-central1-b | Doha, Qatar, Middle East | E2, N2, N2D, T2D, M3 | Intel Broadwell, Ivy Bridge, Skylake, Cascade Lake, Ice Lake, AMD Rome, AMD Milan | | |

| me-central1-c | Doha, Qatar, Middle East | E2, N2, N2D, T2D, M3 | Intel Broadwell, Ivy Bridge, Skylake, Cascade Lake, Ice Lake, AMD Rome, AMD Milan | | |

| me-central2-a | Dammam, Saudi Arabia, Middle East | E2, N2, N2D, T2D, M3 | Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | | |

| me-central2-b | Dammam, Saudi Arabia, Middle East | E2, N2, N2D, T2D | Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | | |

| me-central2-c | Dammam, Saudi Arabia, Middle East | E2, N2, N2D, T2D, M3 | Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | | |

| me-west1-a | Tel Aviv, Israel, Middle East | E2, N2, N2D, T2D, M1, C2 | Intel Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | | |

| me-west1-b | Tel Aviv, Israel, Middle East | E2, N2, N2D, T2D, C2, A2 | Intel Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | GPUs | |

| me-west1-c | Tel Aviv, Israel, Middle East | E2, N2, N2D, T2D, M1, C2, A2 | Intel Broadwell, Skylake, Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | GPUs | |

| Zones | Location | Machine types | CPUs | Resources | CO2 emissions |

| africa-south1-a | Johannesburg, South Africa | E2, N2, N2D, T2D | Intel Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | | |

| africa-south1-b | Johannesburg, South Africa | E2, N2, N2D, T2D | Intel Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | | |

| africa-south1-c | Johannesburg, South Africa | E2, N2, N2D, T2D | Intel Cascade Lake, Ice Lake, AMD EPYC Rome, AMD EPYC Milan | | |

**NVIDIA GPUs for compute workloads**

For compute workloads, GPU models are available in the following stages:

* NVIDIA H100 80GB: nvidia-h100-80gb: **Generally Available**
* NVIDIA L4: nvidia-l4: **Generally Available**
* NVIDIA A100
  + NVIDIA A100 40GB: nvidia-tesla-a100: **Generally Available**
  + NVIDIA A100 80GB: nvidia-a100-80gb: **Generally Available**
* NVIDIA T4: nvidia-tesla-t4: **Generally Available**
* NVIDIA V100: nvidia-tesla-v100: **Generally Available**
* NVIDIA P100: nvidia-tesla-p100: **Generally Available**
* NVIDIA P4: nvidia-tesla-p4: **Generally Available**
* NVIDIA K80: nvidia-tesla-k80: **Generally Available**. See [NVIDIA K80 end of support](https://cloud.google.com/compute/docs/eol/k80-eol).

**NVIDIA H100 GPUs**

To run NVIDIA H100 80GB GPUs, you must use an [A3 accelerator-optimized](https://cloud.google.com/compute/docs/accelerator-optimized-machines#a3-vms) machine type.

| **GPU model** | [**Machine type**](https://cloud.google.com/compute/docs/gpus/about-gpus#gpu-machine-types) | **GPUs** | **GPU memory\*** | **Available vCPUs** | **Available memory** | **Local SSD supported** |
| --- | --- | --- | --- | --- | --- | --- |
| [NVIDIA H100](https://www.nvidia.com/en-us/data-center/h100/) | a3-highgpu-8g | 8 GPUs | 640 GB HBM3 | 208 vCPUs | 1872 GB | Bundled (6000 GB) |
|  |  |  |  |  |  |

\*GPU memory is the memory that is available on a GPU device that can be used for temporary storage of data. It is separate from the VM's memory and is specifically designed to handle the higher bandwidth demands of your graphics-intensive workloads.

**NVIDIA L4 GPUs**

To run NVIDIA L4 GPUs, you must use a [G2 accelerator-optimized](https://cloud.google.com/compute/docs/accelerator-optimized-machines#g2-vms) machine type.

Each G2 machine type has a fixed number of [NVIDIA L4 GPUs](https://cloud.google.com/compute/docs/gpus#l4-gpus) and vCPUs attached. Each G2 machine type also has a default memory and a custom memory range. The custom memory range defines the amount of memory that you can allocate to your VM for each machine type. You can specify your custom memory during VM creation.

| **GPU model** | [**Machine type**](https://cloud.google.com/compute/docs/gpus/about-gpus#gpu-machine-types) | **GPUs** | **GPU memory\*** | **vCPUs** | **Default memory** | **Custom memory range** | **Max local SSD supported** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| [NVIDIA L4](https://www.nvidia.com/en-us/data-center/l4/) | g2-standard-4 | 1 GPU | 24 GB GDDR6 | 4 vCPUs | 16 GB | 16 - 32 GB | 375 GB |
| g2-standard-8 | 1 GPU | 24 GB GDDR6 | 8 vCPUs | 32 GB | 32 - 54 GB | 375 GB |
| g2-standard-12 | 1 GPU | 24 GB GDDR6 | 12 vCPUs | 48 GB | 48 - 54 GB | 375 GB |
| g2-standard-16 | 1 GPU | 24 GB GDDR6 | 16 vCPUs | 64 GB | 54 - 64 GB | 375 GB |
| g2-standard-24 | 2 GPUs | 48 GB GDDR6 | 24 vCPUs | 96 GB | 96 - 108 GB | 750 GB |
| g2-standard-32 | 1 GPU | 24 GB GDDR6 | 32 vCPUs | 128 GB | 96 - 128 GB | 375 GB |
| g2-standard-48 | 4 GPUs | 96 GB GDDR6 | 48 vCPUs | 192 GB | 192 - 216 GB | 1500 GB |
| g2-standard-96 | 8 GPUs | 192 GB GDDR6 | 96 vCPUs | 384 GB | 384 - 432 GB | 3000 GB |

\*GPU memory is the memory that is available on a GPU device that can be used for temporary storage of data. It is separate from the VM's memory and is specifically designed to handle the higher bandwidth demands of your graphics-intensive workloads.

**NVIDIA A100 GPUs**

To run NVIDIA A100 GPUs, you must use the [A2 accelerator-optimized](https://cloud.google.com/compute/docs/accelerator-optimized-machines#a2_vms) machine type.

Each A2 machine type has a fixed GPU count, vCPU count, and memory size.

[A100 40GB](https://cloud.google.com/compute/docs/gpus#a100-40gb)[A100 80GB](https://cloud.google.com/compute/docs/gpus#a100-80gb)

| **GPU model** | [**Machine type**](https://cloud.google.com/compute/docs/gpus/about-gpus#gpu-machine-types) | **GPUs** | **GPU memory\*** | **Available vCPUs** | **Available memory** | **Local SSD supported** |
| --- | --- | --- | --- | --- | --- | --- |
| [NVIDIA A100 40GB](https://www.nvidia.com/en-us/data-center/a100/) | a2-highgpu-1g | 1 GPU | 40 GB HBM2 | 12 vCPUs | 85 GB | Yes |
| a2-highgpu-2g | 2 GPUs | 80 GB HBM2 | 24 vCPUs | 170 GB | Yes |
| a2-highgpu-4g | 4 GPUs | 160 GB HBM2 | 48 vCPUs | 340 GB | Yes |
| a2-highgpu-8g | 8 GPUs | 320 GB HBM2 | 96 vCPUs | 680 GB | Yes |
| a2-megagpu-16g | 16 GPUs | 640 GB HBM2 | 96 vCPUs | 1360 GB | Yes |

\*GPU memory is the memory that is available on a GPU device that can be used for temporary storage of data. It is separate from the VM's memory and is specifically designed to handle the higher bandwidth demands of your graphics-intensive workloads.

**NVIDIA T4 GPUs**

VMs with lower numbers of GPUs are limited to a maximum number of vCPUs. In general, a higher number of GPUs lets you create instances with a higher number of vCPUs and memory.

| **GPU model** | [**Machine type**](https://cloud.google.com/compute/docs/gpus/about-gpus#gpu-machine-types) | **GPUs** | **GPU memory\*** | **Available vCPUs** | **Available memory** | **Local SSD supported** |
| --- | --- | --- | --- | --- | --- | --- |
| [NVIDIA T4](https://www.nvidia.com/en-us/data-center/tesla-t4/) | [N1 machine series](https://cloud.google.com/compute/docs/general-purpose-machines#n1_machines) except N1 shared-core | 1 GPU | 16 GB GDDR6 | 1 - 48 vCPUs | 1 - 312 GB | Yes |
| 2 GPUs | 32 GB GDDR6 | 1 - 48 vCPUs | 1 - 312 GB | Yes |
| 4 GPUs | 64 GB GDDR6 | 1 - 96 vCPUs | 1 - 624 GB | Yes |

\*GPU memory is the memory that is available on a GPU device that can be used for temporary storage of data. It is separate from the VM's memory and is specifically designed to handle the higher bandwidth demands of your graphics-intensive workloads.

**NVIDIA P4 GPUs**

For P4 GPUs, local SSD is only supported in select regions, see [Local SSD availability by GPU regions and zone](https://cloud.google.com/compute/docs/gpus/gpu-regions-zones#local-ssd-gpu).

| **GPU model** | [**Machine type**](https://cloud.google.com/compute/docs/gpus/about-gpus#gpu-machine-types) | **GPUs** | **GPU memory\*** | **Available vCPUs** | **Available memory** | **Local SSD supported** |
| --- | --- | --- | --- | --- | --- | --- |
| [NVIDIA P4](https://www.nvidia.com/en-us/deep-learning-ai/inference-platform/hpc/) | [N1 machine series](https://cloud.google.com/compute/docs/general-purpose-machines#n1_machines) except N1 shared-core | 1 GPU | 8 GB GDDR5 | 1 - 24 vCPUs | 1 - 156 GB | Yes |
| 2 GPUs | 16 GB GDDR5 | 1 - 48 vCPUs | 1 - 312 GB | Yes |
| 4 GPUs | 32 GB GDDR5 | 1 - 96 vCPUs | 1 - 624 GB | Yes |

\*GPU memory is the memory that is available on a GPU device that can be used for temporary storage of data. It is separate from the VM's memory and is specifically designed to handle the higher bandwidth demands of your graphics-intensive workloads.

**NVIDIA V100 GPUs**

For V100 GPUs, local SSD is only supported in select regions, see [Local SSD availability by GPU regions and zone](https://cloud.google.com/compute/docs/gpus/gpu-regions-zones#local-ssd-gpu).

| **GPU model** | [**Machine type**](https://cloud.google.com/compute/docs/gpus/about-gpus#gpu-machine-types) | **GPUs** | **GPU memory\*** | **Available vCPUs** | **Available memory** | **Local SSD supported** |
| --- | --- | --- | --- | --- | --- | --- |
| [NVIDIA V100](https://www.nvidia.com/en-us/data-center/tesla-v100/) | [N1 machine series](https://cloud.google.com/compute/docs/general-purpose-machines#n1_machines) except N1 shared-core | 1 GPU | 16 GB HBM2 | 1 - 12 vCPUs | 1 - 78 GB | Yes |
| 2 GPUs | 32 GB HBM2 | 1 - 24 vCPUs | 1 - 156 GB | Yes |
| 4 GPUs | 64 GB HBM2 | 1 - 48 vCPUs | 1 - 312 GB | Yes |
| 8 GPUs | 128 GB HBM2 | 1 - 96 vCPUs | 1 - 624 GB | Yes |

\*GPU memory is the memory that is available on a GPU device that can be used for temporary storage of data. It is separate from the VM's memory and is specifically designed to handle the higher bandwidth demands of your graphics-intensive workloads.

**NVIDIA P100 GPUs**

For some P100 GPUs, the maximum CPU and memory that is available for some configurations is dependent on the zone in which the GPU resource is running.

| **GPU model** | [**Machine type**](https://cloud.google.com/compute/docs/gpus/about-gpus#gpu-machine-types) | **GPUs** | **GPU memory\*** | **Available vCPUs** | **Available memory** | **Local SSD supported** |
| --- | --- | --- | --- | --- | --- | --- |
| [NVIDIA P100](http://www.nvidia.com/object/tesla-p100.html) | [N1 machine series](https://cloud.google.com/compute/docs/general-purpose-machines#n1_machines) except N1 shared-core | 1 GPU | 16 GB HBM2 | 1 - 16 vCPUs | 1 - 104 GB | Yes |
| 2 GPUs | 32 GB HBM2 | 1 - 32 vCPUs | 1 - 208 GB | Yes |
| 4 GPUs | 64 GB HBM2 | 1 - 64 vCPUs (us-east1-c, europe-west1-d, europe-west1-b)  1 - 96 vCPUs (all P100 zones) | 1 - 208 GB (us-east1-c, europe-west1-d, europe-west1-b)  1 - 624 GB (all P100 zones) | Yes |

\*GPU memory is the memory that is available on a GPU device that can be used for temporary storage of data. It is separate from the VM's memory and is specifically designed to handle the higher bandwidth demands of your graphics-intensive workloads.

**NVIDIA K80 GPUs**

**Caution:** NVIDIA K80 GPUs will reach end of life on May 1, 2024. For more information, see [NVIDIA K80 EOL](https://cloud.google.com/compute/docs/eol/k80-eol).

NVIDIA K80 boards contain two GPUs each. The pricing for K80 GPUs is by individual GPU, not by the board.

| **GPU model** | [**Machine type**](https://cloud.google.com/compute/docs/gpus/about-gpus#gpu-machine-types) | **GPUs** | **GPU memory\*** | **Available vCPUs** | **Available memory** | **Local SSD supported** |
| --- | --- | --- | --- | --- | --- | --- |
| [NVIDIA K80](https://www.nvidia.com/en-gb/data-center/tesla-k80/) | [N1 machine series](https://cloud.google.com/compute/docs/general-purpose-machines#n1_machines) except N1 shared-core | 1 GPU | 12 GB GDDR5 | 1 - 8 vCPUs | 1 - 52 GB | Yes |
| 2 GPUs | 24 GB GDDR5 | 1 - 16 vCPUs | 1 - 104 GB | Yes |
| 4 GPUs | 48 GB GDDR5 | 1 - 32 vCPUs | 1 - 208 GB | Yes |
| 8 GPUs | 96 GB GDDR5 | 1 - 64 vCPUs | 1 - 416 GB (asia-east1-a and us-east1-d)  1 - 208 GB (all K80 zones) | Yes |

\*GPU memory is the memory that is available on a GPU device that can be used for temporary storage of data. It is separate from the VM's memory and is specifically designed to handle the higher bandwidth demands of your graphics-intensive workloads.

**NVIDIA RTX Virtual Workstations (vWS) for graphics workloads**

If you have graphics-intensive workloads, such as 3D visualization, you can create virtual workstations that use [NVIDIA RTX Virtual Workstations (vWS)](https://www.nvidia.com/en-us/design-visualization/virtual-workstation/) (formerly known as NVIDIA GRID). When you create a virtual workstation, an NVIDIA RTX Virtual Workstation (vWS) license is automatically added to your VM.

For information about pricing for virtual workstations, see [GPU pricing page](https://cloud.google.com/compute/gpus-pricing).

For graphics workloads, NVIDIA RTX virtual workstation (vWS) models are available in the following stages:

* NVIDIA L4 Virtual Workstations: nvidia-l4-vws: **Generally Available**
* NVIDIA T4 Virtual Workstations: nvidia-tesla-t4-vws: **Generally Available**
* NVIDIA P100 Virtual Workstations: nvidia-tesla-p100-vws: **Generally Available**
* NVIDIA P4 Virtual Workstations: nvidia-tesla-p4-vws: **Generally Available**

**NVIDIA L4 vWS GPUS**

| **GPU model** | [**Machine type**](https://cloud.google.com/compute/docs/gpus/about-gpus#gpu-machine-types) | **GPUs** | **GPU memory** | **vCPUs** | **Default memory** | **Custom memory range** | **Max local SSD supported** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| [NVIDIA L4 Virtual Workstation](https://www.nvidia.com/en-us/data-center/l4/) | g2-standard-4 | 1 GPU | 24 GB GDDR6 | 4 vCPUs | 16 GB | 16 - 32 GB | 375 GB |
| g2-standard-8 | 1 GPU | 24 GB GDDR6 | 8 vCPUs | 32 GB | 32 - 54 GB | 375 GB |
| g2-standard-12 | 1 GPU | 24 GB GDDR6 | 12 vCPUs | 48 GB | 48 - 54 GB | 375 GB |
| g2-standard-16 | 1 GPU | 24 GB GDDR6 | 16 vCPUs | 64 GB | 54 - 64 GB | 375 GB |
| g2-standard-24 | 2 GPUs | 48 GB GDDR6 | 24 vCPUs | 96 GB | 96 - 108 GB | 750 GB |
| g2-standard-32 | 1 GPU | 24 GB GDDR6 | 32 vCPUs | 128 GB | 96 - 128 GB | 375 GB |
| g2-standard-48 | 4 GPUs | 96 GB GDDR6 | 48 vCPUs | 192 GB | 192 - 216 GB | 1500 GB |
| g2-standard-96 | 8 GPUs | 192 GB GDDR6 | 96 vCPUs | 384 GB | 384 - 432 GB | 3000 GB |

\*GPU memory is the memory that is available on a GPU device that can be used for temporary storage of data. It is separate from the VM's memory and is specifically designed to handle the higher bandwidth demands of your graphics-intensive workloads.

**NVIDIA T4 vWS GPUs**

| **GPU model** | [**Machine type**](https://cloud.google.com/compute/docs/gpus/about-gpus#gpu-machine-types) | **GPUs** | **GPU memory\*** | **Available vCPUs** | **Available memory** | **Local SSD supported** |
| --- | --- | --- | --- | --- | --- | --- |
| [NVIDIA T4 Virtual Workstation](https://www.nvidia.com/en-us/data-center/tesla-t4/) | [N1 machine series](https://cloud.google.com/compute/docs/general-purpose-machines#n1_machines) except N1 shared-core | 1 GPU | 16 GB GDDR6 | 1 - 48 vCPUs | 1 - 312 GB | Yes |
| 2 GPUs | 32 GB GDDR6 | 1 - 48 vCPUs | 1 - 312 GB | Yes |
| 4 GPUs | 64 GB GDDR6 | 1 - 96 vCPUs | 1 - 624 GB | Yes |

\*GPU memory is the memory that is available on a GPU device that can be used for temporary storage of data. It is separate from the VM's memory and is specifically designed to handle the higher bandwidth demands of your graphics-intensive workloads.

**NVIDIA P4 vWS GPUs**

For P4 GPUs, local SSD is only supported in select regions, see [Local SSD availability by GPU regions and zone](https://cloud.google.com/compute/docs/gpus/gpu-regions-zones#local-ssd-gpu).

| **GPU model** | [**Machine type**](https://cloud.google.com/compute/docs/gpus/about-gpus#gpu-machine-types) | **GPUs** | **GPU memory\*** | **Available vCPUs** | **Available memory** | **Local SSD supported** |
| --- | --- | --- | --- | --- | --- | --- |
| [NVIDIA P4 Virtual Workstation](https://www.nvidia.com/en-us/design-visualization/technologies/virtual-gpu/) | [N1 machine series](https://cloud.google.com/compute/docs/general-purpose-machines#n1_machines) except N1 shared-core | 1 GPU | 8 GB GDDR5 | 1 - 16 vCPUs | 1 - 156 GB | Yes |
| 2 GPUs | 16 GB GDDR5 | 1 - 48 vCPUs | 1 - 312 GB | Yes |
| 4 GPUs | 32 GB GDDR5 | 1 - 96 vCPUs | 1 - 624 GB | Yes |

\*GPU memory is the memory that is available on a GPU device that can be used for temporary storage of data. It is separate from the VM's memory and is specifically designed to handle the higher bandwidth demands of your graphics-intensive workloads.

**NVIDIA P100 vWS GPUs**

| **GPU model** | **Machine type** | **GPUs** | **GPU memory\*** | **Available vCPUs** | **Available memory** | **Local SSD supported** |
| --- | --- | --- | --- | --- | --- | --- |
| [NVIDIA P100 Virtual Workstation](https://www.nvidia.com/en-us/design-visualization/technologies/virtual-gpu/) | [N1 machine series](https://cloud.google.com/compute/docs/general-purpose-machines#n1_machines) except N1 shared-core | 1 GPU | 16 GB HBM2 | 1 - 16 vCPUs | 1 - 104 GB | Yes |
| 2 GPUs | 32 GB HBM2 | 1 - 32 vCPUs | 1 - 208 GB | Yes |
| 4 GPUs | 64 GB HBM2 | 1 - 64 vCPUs (us-east1-c, europe-west1-d, europe-west1-b)  1 - 96 vCPUs (all P100 zones) | 1 - 208 GB (us-east1-c, europe-west1-d, europe-west1-b)  1 - 624 GB (all P100 zones) | Yes |

\*GPU memory is the memory that is available on a GPU device that can be used for temporary storage of data. It is separate from the VM's memory and is specifically designed to handle the higher bandwidth demands of your graphics-intensive workloads.

**General comparison chart**

The following table describes the GPU memory size, feature availability, and ideal workload types of different GPU models that are available on Compute Engine.

| **GPU model** | **Memory** | **Interconnect** | **NVIDIA RTX Virtual Workstation (vWS) support** | **Best used for** |
| --- | --- | --- | --- | --- |
| **H100 80GB** | 80 GB HBM3 @ 3.35 TBps | NVLink Full Mesh @ 900 GBps |  | Large models with massive data tables for ML Training, Inference, HPC, BERT, DLRM |
| **A100 80GB** | 80 GB HBM2e @ 1.9 TBps | NVLink Full Mesh @ 600 GBps |  | Large models with massive data tables for ML Training, Inference, HPC, BERT, DLRM |
| **A100 40GB** | 40 GB HBM2 @ 1.6 TBps | NVLink Full Mesh @ 600 GBps |  | ML Training, Inference, HPC |
| **L4** | 24 GB GDDR6 @ 300 GBps | N/A |  | ML Inference, Training, Remote Visualization Workstations, Video Transcoding, HPC |
| **T4** | 16 GB GDDR6 @ 320 GBps | N/A |  | ML Inference, Training, Remote Visualization Workstations, Video Transcoding |
| **V100** | 16 GB HBM2 @ 900 GBps | NVLink Ring @ 300 GBps |  | ML Training, Inference, HPC |
| **P4** | 8 GB GDDR5 @ 192 GBps | N/A |  | Remote Visualization Workstations, ML Inference, and Video Transcoding |
| **P100** | 16 GB HBM2 @ 732 GBps | N/A |  | ML Training, Inference, HPC, Remote Visualization Workstations |
| **K80**EOL | 12 GB GDDR5 @ 240 GBps | N/A |  | ML Inference, Training, HPC |

To compare GPU pricing for the different GPU models and regions that are available on Compute Engine, see [GPU pricing](https://cloud.google.com/compute/gpus-pricing).

**Performance comparison chart**

The following table describes the performance specifications of different GPU models that are available on Compute Engine.

**Compute performance**

| **GPU model** | FP64 | FP32 | FP16 | INT8 |
| --- | --- | --- | --- | --- |
| **H100 80GB** | 34 TFLOPS | 67 TFLOPS |  |  |
| **A100 80GB** | 9.7 TFLOPS | 19.5 TFLOPS |  |  |
| **A100 40GB** | 9.7 TFLOPS | 19.5 TFLOPS |  |  |
| **L4** | 0.5 TFLOPS\* | 30.3 TFLOPS |  |  |
| **T4** | 0.25 TFLOPS\* | 8.1 TFLOPS |  |  |
| **V100** | 7.8 TFLOPS | 15.7 TFLOPS |  |  |
| **P4** | 0.2 TFLOPS\* | 5.5 TFLOPS |  | 22 TOPS† |
| **P100** | 4.7 TFLOPS | 9.3 TFLOPS | 18.7 TFLOPS |  |
| **K80**EOL | 1.46 TFLOPS | 4.37 TFLOPS |  |  |

\*To allow FP64 code to work correctly, a small number of FP64 hardware units are included in the T4, L4, and P4 GPU architecture.

†TeraOperations per Second.

**Tensor core performance**

| **GPU model** | FP64 | TF32 | Mixed-precision FP16/FP32 | INT8 | INT4 | FP8 |
| --- | --- | --- | --- | --- | --- | --- |
| **H100 80GB** | 67 TFLOPS | 989 TFLOPS† | 1,979 TFLOPS\*, † | 3,958 TOPS† |  | 3,958 TFLOPS† |
| **A100 80GB** | 19.5 TFLOPS | 156 TFLOPS | 312 TFLOPS\* | 624 TOPS | 1248 TOPS |  |
| **A100 40GB** | 19.5 TFLOPS | 156 TFLOPS | 312 TFLOPS\* | 624 TOPS | 1248 TOPS |  |
|  |  |  |  |  |  |  |
| **L4** |  | 120 TFLOPS† | 242 TFLOPS\*, † | 485 TOPS† |  | 485 TFLOPS† |
| **T4** |  |  | 65 TFLOPS | 130 TOPS | 260 TOPS |  |
| **V100** |  |  | 125 TFLOPS |  |  |  |
| **P4** |  |  |  |  |  |  |
| **P100** |  |  |  |  |  |  |
| **K80**EOL |  |  |  |  |  |  |

\*For mixed precision training, NVIDIA H100, A100, and L4 GPUs also support the bfloat16 data type.

†For H100 and L4 GPUs, structural sparsity is supported which you can use to double the performance value. The values shown are with sparsity. Specifications are one-half lower without sparsity

When you create a virtual machine (VM) on Compute Engine, you specify a machine series and a machine type for the VM. Each machine series is associated with one or more CPU platforms. If there are multiple CPU platforms available for a machine type, you can select a [minimum CPU](https://cloud.google.com/compute/docs/instances/specify-min-cpu-platform) platform for the VM.

A CPU platform offers multiple physical processors, and each of these processors are referred to as a core. For all processors available on Compute Engine, a single CPU core can run as multiple hardware multithreads through [Simultaneous multithreading (SMT)](https://wikipedia.org/wiki/Simultaneous_multithreading), which is known on Intel processors as [Intel Hyper-Threading Technology](https://www.intel.com/content/www/us/en/architecture-and-technology/hyper-threading/hyper-threading-technology.html). On Compute Engine, each hardware multithread is called a virtual CPU (vCPU). When vCPUs are reported to the VM as occupying different virtual cores, Compute Engine ensures that these vCPUs never share the same physical core.

The [machine type](https://cloud.google.com/compute/docs/machine-resource) of your VM specifies its number of vCPUs, and you can infer its number of physical CPU cores using the default vCPU per core ratio for that machine series:

* For the Tau T2D, Tau T2A, and H3 machine series, VMs always have one vCPU per core.
* For all other machine series, VMs have two vCPUs per core by default.

You can optionally [set a VM to have one vCPU per core instead of two vCPUs per core](https://cloud.google.com/compute/docs/instances/set-threads-per-core), which might benefit some workloads. Importantly, when you do this, the machine type of your VM no longer reflects the correct number of vCPUs. Instead, the [pricing](https://cloud.google.com/compute/docs/instances/set-threads-per-core#pricing) and number of physical CPU cores remains the same as it would be for the default two vCPUs per core ratio, and the number of vCPUs is half of the value indicated by the machine type.

**Arm processors**

For Arm processors, Compute Engine uses one thread per core. Each vCPU maps to a physical core with no SMT.

The following table describes the Arm processors that are available for Compute Engine VMs.

| **CPU processor** | **Processor SKU** | **Supported machine series and types** | **All-core sustained frequency (GHz)** |
| --- | --- | --- | --- |
| **Ampere Altra** | Q64-30 | * [Tau T2A](https://cloud.google.com/compute/docs/general-purpose-machines#t2a_machines) | 3.0 |

**x86 processors**

For most x86 processors, each vCPU is implemented as a single hardware thread. The Tau T2D machine series is the exception, with one vCPU representing one physical core.

**Intel processors**

On Intel Xeon processors, [Intel Hyper-Threading Technology](https://www.intel.com/content/www/us/en/architecture-and-technology/hyper-threading/hyper-threading-technology.html) supports multiple threads running concurrently on each core. The specific size and shape of your [VM instance](https://cloud.google.com/compute/docs/machine-types) determines the number of its vCPUs.

| **CPU processor** | **Processor SKU** | **Supported machine series and types** | **Base frequency (GHz)** | **All-core turbo frequency (GHz)** | **Single-core max turbo frequency (GHz)** |
| --- | --- | --- | --- | --- | --- |
| **Intel Xeon Scalable Processor (Sapphire Rapids) 4th generation** | Intel® Xeon® Platinum 8481C Processor | * [C3](https://cloud.google.com/compute/docs/general-purpose-machines#c3_series) * [Z3](https://cloud.google.com/compute/docs/storage-optimized-machines#z3_series) ([Preview](https://cloud.google.com/products#product-launch-stages)) * [H3](https://cloud.google.com/compute/docs/compute-optimized-machines#h3_series) * [A3](https://cloud.google.com/compute/docs/accelerator-optimized-machines#a3-vms) | 1.9 | 3.0 | 3.3 |
| **Intel Xeon Scalable Processor (Ice Lake) 3rd Generation** | Intel® Xeon® Platinum 8373C Processor | * [N2\*](https://cloud.google.com/compute/docs/general-purpose-machines#n2_machines) * [M3](https://cloud.google.com/compute/docs/memory-optimized-machines#m3_machine_types) | 2.6 | 3.4 | 3.5 |
| **Intel Xeon Scalable Processor (Cascade Lake) 2nd Generation** |  |  |  |  |  |
| Intel® Xeon® Gold 6268CL Processor | * [N2\*](https://cloud.google.com/compute/docs/general-purpose-machines#n2_series) | 2.8 | 3.4 | 3.9 |
| Intel® Xeon® Gold 6253CL Processor | * [C2](https://cloud.google.com/compute/docs/compute-optimized-machines#c2_machine_types) | 3.1 | 3.8 | 3.9 |
| Intel® Xeon® Platinum 8280L Processor | * [M2](https://cloud.google.com/compute/docs/memory-optimized-machines#m2_machine_types) | 2.5 | 3.4 | 4.0 |
| Intel® Xeon® Platinum 8273CL Processor | * [A2](https://cloud.google.com/compute/docs/accelerator-optimized-machines#a2-vms) * [G2](https://cloud.google.com/compute/docs/accelerator-optimized-machines#g2-vms) | 2.2 | 2.9 | 3.7 |
| **Intel Xeon Scalable Processor (Skylake) 1st Generation** | Intel® Xeon® Scalable Platinum 8173M Processor | * [E2](https://cloud.google.com/compute/docs/general-purpose-machines#e2_machine_types) * [m1-megamem memory-optimized machine types](https://cloud.google.com/compute/docs/memory-optimized-machines#m1_machine_types) * [N1](https://cloud.google.com/compute/docs/general-purpose-machines#n1_machines) | 2.0 | 2.7 | 3.5 |
| **Intel Xeon E7 (Broadwell E7)** | Intel® Xeon® E7-8880V4 Processor | * [m1-ultramem memory-optimized machine types](https://cloud.google.com/compute/docs/memory-optimized-machines#m1_machine_types) | 2.2 | 2.6 | 3.3 |
| **Intel Xeon E5 v4 (Broadwell E5)** | Intel® Xeon® E5-2696V4 Processor | * [E2](https://cloud.google.com/compute/docs/general-purpose-machines#e2_machine_types) * [N1](https://cloud.google.com/compute/docs/general-purpose-machines#n1_machines) | 2.2 | 2.8 | 3.7 |
| **Intel Xeon E5 v3 (Haswell)** | Intel® Xeon® E5-2696V3 Processor | * [N1](https://cloud.google.com/compute/docs/general-purpose-machines#n1_machines) | 2.3 | 2.8 | 3.8 |
| **Intel Xeon E5 v2 (Ivy Bridge)** | Intel® Xeon® E5-2696V2 Processor | * [N1](https://cloud.google.com/compute/docs/general-purpose-machines#n1_machines) | 2.5 | 3.1 | 3.5 |
| **Intel Xeon E5 (Sandy Bridge)** | Intel® Xeon® E5-2689 Processor | * [N1](https://cloud.google.com/compute/docs/general-purpose-machines#n1_machines) | 2.6 | 3.2 | 3.6 |

\*N2 machine types that have 96 or more vCPUs require the Intel Ice Lake CPU.

**AMD processors**

AMD processors provide optimized performance and scalability using SMT. In almost all cases, Compute Engine uses two threads per core, and each vCPU is one thread. Tau T2D is the exception where Compute Engine uses one thread per core and each vCPU maps to a physical core. The specific size and shape of your [VM instance](https://cloud.google.com/compute/docs/machine-resource) determines the number of its vCPUs.

| **CPU processor** | **Processor SKU** | **Supported machine series** | **Base frequency (GHz)** | **Effective frequency (GHz)** | **Max boost frequency (GHz)** |
| --- | --- | --- | --- | --- | --- |
| **AMD EPYC Genoa 4th Generation** | AMD EPYC™ 9B14 | * [C3D](https://cloud.google.com/compute/docs/general-purpose-machines#c3d_series) | 2.6 | 3.3 | 3.7 |
| **AMD EPYC Milan 3rd Generation** | AMD EPYC™ 7B13 | * [E2](https://cloud.google.com/compute/docs/general-purpose-machines#e2_machine_types) * [Tau T2D](https://cloud.google.com/compute/docs/general-purpose-machines#t2d_machines) * [N2D](https://cloud.google.com/compute/docs/general-purpose-machines#n2d_machines) * [C2D](https://cloud.google.com/compute/docs/compute-optimized-machines#c2d_machine_types) | 2.45 | 2.8 | 3.5 |
| **AMD EPYC Rome 2nd Generation** | AMD EPYC™ 7B12 | * [E2](https://cloud.google.com/compute/docs/general-purpose-machines#e2_machine_types) * [N2D](https://cloud.google.com/compute/docs/general-purpose-machines#n2d_machines) | 2.25 | 2.7 | 3.3 |

**Frequency behavior**

The previous tables describe the hardware specifications of the CPUs that are available with Compute Engine, but keep the following points in mind:

* **Frequency**: A PC's frequency, or clock speed, measures the number of cycles the CPU executes per second, measured in GHz (gigahertz). Generally, higher frequencies indicate better performance. However, different CPU designs handle instructions differently, so an older CPU with a higher clock speed can be outperformed by a newer CPU with a lower clock speed because the newer architecture deals with instructions more efficiently.

For more information about CPU clock cycles and performance, see [Clock rates and system performance](https://cloud.google.com/architecture/resource-mappings-from-on-premises-hardware-to-gcp#clock_rates_and_system_performance).

* **Base frequency**: The frequency at which the CPU runs when the system is idle or under light load. When running at its base frequency, the CPU draws less power and produces less heat.

A VM's guest environment reflects the base frequency, regardless of what frequency the VM is actually running at.

* **All-core turbo frequency**: The frequency at which each CPU typically runs when all cores in the socket are not idle at the same time. Different workloads place different demands on a system's CPU. Boost technologies address this difference and help processes adapt to the workload demands by increasing the CPU's frequency.
  + Most VMs get the all-core turbo frequency, even if only the base frequency is advertised to the guest environment.
  + Ampere Altra Arm processors can provide more predictable performance because the frequency for Arm processors is always the all-core turbo frequency.
* **Max turbo frequency**: The frequency a CPU targets when stressed by a demanding application like a video game or design modeling application. It's the maximum single-core frequency that a CPU achieves without overclocking.
* **Processor power management technologies**: Intel processors support multiple technologies to optimize the power consumption. These technologies are divided into two categories, or states:
  + C-states are states when the CPU has reduced or turned off selected functions.
  + P-states provide a way to scale the frequency and voltage at which the processor runs so as to reduce the power consumption of the CPU.

Certain C2 (30, 60 vCPUs), C2D (56, 112 vCPUs) and M2 (208, 416 vCPUs) machine types support instance-provided C-state hints by way of the MWAIT instruction.

Google Cloud VMs don't provide any facilities for customer control of P-states.

**CPU features**

Chip manufacturers add advanced technologies for computations, graphics, virtualization, and memory management to the CPUs they produce. Google Cloud supports the use of some of these advanced features with Compute Engine.

**Advanced Matrix Extensions (AMX)**

[Intel AMX](https://www.intel.com/content/www/us/en/products/docs/accelerator-engines/advanced-matrix-extensions/overview.html) is a new instruction set architecture (ISA) extension designed to accelerate artificial intelligence (AI) and machine learning (ML) workloads. AMX introduces new instructions that can be used to perform matrix multiplication and convolution operations, which are two of the most common operations in AI and ML.

AMX is supported on Intel Xeon 4th generation processors (code named Sapphire Rapids), which powers the A3 accelerator-optimized and C3 general-purpose VM series. All [C3 VM machine types](https://cloud.google.com/compute/docs/general-purpose-machines#c3_series) support AMX instruction sets.

AMX introduces 2-dimensional registers called *tiles* upon which accelerators can perform operations. AMX is intended as an extensible architecture. The first accelerator implemented is called tile matrix multiply unit (TMUL). Each CPU core of the Sapphire Rapids processor has an independent AMX TMUL unit.

More technical details about Intel AMX can be found at [Intel AMX support in 5.16](https://lwn.net/Articles/874846/). Intel offers a tutorial on AMX at [Code Sample: Intel® Advanced Matrix Extensions (Intel® AMX) - Intrinsics Functions](https://www.intel.com/content/www/us/en/developer/articles/code-sample/advanced-matrix-extensions-intrinsics-functions.html#gs.u3cqw2).

**Requirements for using AMX**

Intel AMX instructions have certain minimum software requirements such as:

* For custom images, AMX is supported with Linux kernel version 5.16 or later.
* Google Cloud offers support for AMX in the following [public images](https://cloud.google.com/compute/docs/images/os-details):
  + CentOS Stream 8 or later
  + Container-Optimized OS 109 LTS or later
  + RHEL 8 (latest build) or later
  + Rocky Linux 8 (latest build) or later
  + Ubuntu 22.04 or later
  + Windows Server 2022 or later
* [Tensorflow](https://blog.tensorflow.org/2023/01/optimizing-tensorflow-for-4th-gen-intel-xeon-processors.html) 2.9.1 or greater
* Intel extension for [Intel® Optimization for PyTorch](https://www.intel.com/content/www/us/en/developer/articles/technical/intel-for-pytorch-using-google-cloud-platform.html)

For regional availability of C3 VMs, see [Available regions and zones](https://cloud.google.com/compute/docs/regions-zones#available) and filter the table to show only C3 machine types.

**Confidential Computing**

To protect your data while it's in use, AMD EPYC 3rd generation CPUs (code named Milan) can be used in Confidential VM instances. They support the following attestation and memory encryption technologies:

* [AMD Secure Encrypted Virtualization (SEV)](https://www.amd.com/content/dam/amd/en/documents/epyc-business-docs/white-papers/memory-encryption-white-paper.pdf)
* [AMD Secure Encrypted Virtualization-Secure Nested Paging (SEV-SNP)](https://www.amd.com/content/dam/amd/en/documents/epyc-business-docs/white-papers/SEV-SNP-strengthening-vm-isolation-with-integrity-protection-and-more.pdf) ([Preview](https://cloud.google.com/products#product-launch-stages))