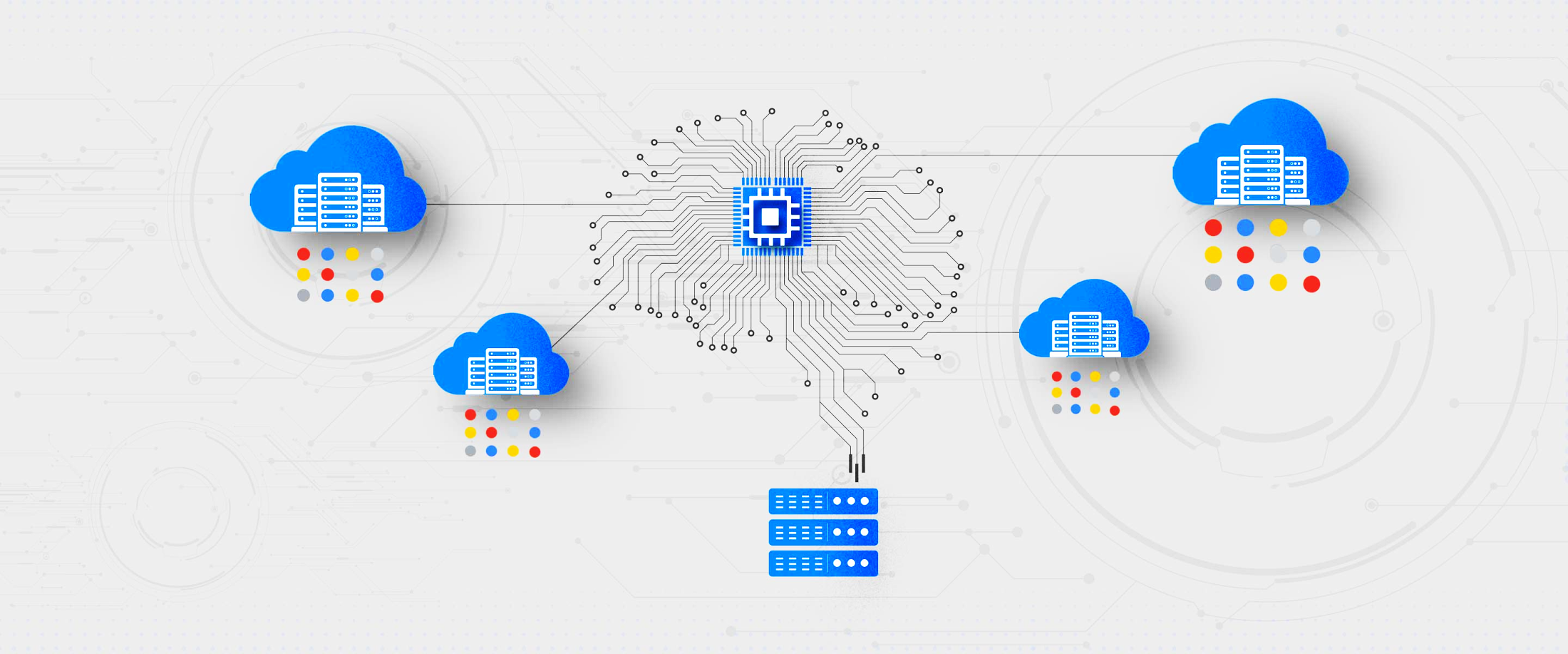
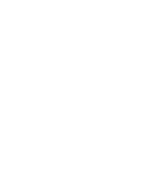
****

**High Performance Computing on Google Cloud Platform**

**January 2021**



C:\Users\sanish.raju\Desktop\1503-1413062_Brand update template creation\13-May-2015\Letter pieces\EY Approach_2lines LETTER_proposal.wmf

**Table of Contents –**

iNTRODUCTION 3

HPC COMPUTING 4

ON PREMISES HPC HAS ITS LIMITS 5

CLOUD A BETTER WAY TO HPC 7

GCP DELIVER A COMPLETE HPC SOLUTION 9

DRIVING INNOVATION ACROSS INDUSTRIES 12

CONCLUSION 15

|  |
| --- |
|  |

|  |
| --- |
| Apractice to build up Computing Power in a way that delivers much Higher Performance than one could get from a typical desktop computer or workstation in order to solve large problems in field of science, engineering or business.It is also referred to as Super-Computing for this reason. It enables the simulation or analysis of huge volumes of data that will otherwise not be possible to do with standard computer.High Performance Computing (HPC) has been key to solving the most complex problems in every industry, and changing the way we work and live. |

|  |
| --- |
| From weather modeling to genome mapping to the search for extraterrestrial intelligence, HPC is helping to push the boundaries of what’s possible with advanced computing technologies. Once confined to government labs, large enterprises, and select academic organizations, today it is found across a wide range of industries.In this paper we will discuss how cloud services put the world’s most advanced computing capabilities within reach for more organizations, helping them to innovate faster and gain a competitive edge. We will discuss the advantages of running HPC workloadsOn Google Cloud Platform(GCP), compared to traditional on-premises architectures. We will also illustrate these benefits in actual deployments across a variety of industries. |

**INTRODUCTION**

|  |
| --- |
| Although HPC applications share some common building blocks, they are not all similar. HPC applications are often based on complex algorithms that rely on high performing infrastructure for efficient execution.  These applications need hardware that includes high performance processors, memory, and communication subsystems. For many applications and workloads, the performance of compute elements must be complemented by comparably high-performance storage and networking elements. Some may demand high levels of parallel processing, but not necessarily fast storage or high-performance interconnect. Other applications are interconnect-sensitive, requiring low latency and high throughput networking. Similarly, there are many I/O-sensitive applications that, without a very fast I/O subsystem, will run slowly because of storage bottlenecks. And still other applications such as game streaming, video encoding, and 3D application streaming, need performance acceleration using GPUs. Today, many large enterprises and research institutions procure and maintain their own HPC |

|  |
| --- |
| infrastructure. This HPC infrastructure is shared across many applications and groups within the organization to maximize utilization of this significant capital investment.  Cloud-based services have opened up a new frontier for HPC. Moving HPC workloads to the cloud can provide near instant access to virtually unlimited computing resources for a wider community of users, and can support completely new types of applications. Today, organizations of all sizes are looking to the cloud to support their most advanced computing applications.  For smaller enterprises, cloud is a great starting point, enabling fast, agile deployment without the need for heavy capital expenditure. For large enterprises, cloud provides an easier way to tailor HPC infrastructure to changing business needs and to gain access to the latest technologies, without having to worry about upfront investments in new infrastructure, or ongoing operational expenses. When compared to traditional on-premises HPC infrastructures, cloud offers significant advantages in terms of scalability, flexibility, and cost. |

****

**HPC COMPUTING**

****

**HPC HAS Its LIMITS**

|  |
| --- |
| Today, on-premises HPC infrastructure handles most of the HPC workloads that enterprises and research institutions employ. Most HPC system administrators maintain and operate this infrastructure at varying levels of utilization. However, business is always competitive, so efficiency needs to be coupled with the flexibility and opportunity to innovate continuously. Some of the challenges with on-premises HPC are well known. These include long procurement cycles, high initial capital investment, and the need for mid-cycle technology refreshes. For most organizations, planning for and procuring an HPC system is a long and arduous process that involves detailed capacity forecasting and system evaluation |

|  |
| --- |
| cycles. Often, the significant up-front capital investment required is a limiting factor for the amount of capacity that can be procured. Maintaining the infrastructure over its lifecycle is an expensive proposition, as well. Previously, technology refreshes every three years was enough to stay current with the compute technology and incremental demands from HPC workloads. However, to take advantage of the faster pace of innovation, HPC customers are needing to refresh their infrastructure more often than before. And it is worth the effort. IDC reports that for every $1 spent on HPC, businesses see $463 in incremental revenues and $44 in incremental profit, so delaying incremental investments in HPC – and thus delaying the innovations it brings – has large downstream effects on the business. |

## There are other limitations of on-premises HPC infrastructure that are less visible and so are often overlooked leading to misplaced optimization efforts.

**Stifled Innovation:**

* Often the constraints of on-premises infrastructure mean that use cases or applications that did not meet the capabilities of the hardware were not considered. When engineers and researchers are forced to limit their imagination to what can be tried out with limited access to infrastructure, the opportunity to think outside the box and tinker with new ideas gets lost.

****

**Reduced Productivity:**

* On-premises systems often have long queues and wait times that decrease productivity. They are managed to maximize utilization – often resulting in very intricate scheduling policies for jobs. However, even if a job requires only a couple of hours to run, it may be stuck in a prioritized queue for weeks or months – decreasing overall productivity and limiting innovation. In contrast, with virtually unlimited capacity, the cloud can free users to get the same job done, but much faster, without having to stand in line behind others who are just as eager to make progress.

****

**Limited Scalability and Flexibility:**

* ****HPC workloads and their demands are constantly changing, and legacy HPC architectures cannot always keep pace with evolving requirements. For example, infrastructure elements like GPUs, containers, and serverless technologies are not readily available in an on-premises environment. Integrating new OS or container capabilities – or even upgrading libraries and applications – is a major system-wide undertaking. And when an on-premises HPC system is designed for a specific application or workload, it’s difficult and expensive to take on new HPC applications, as well as forecast and scale for future (frequently unknown) requirements.

**Lost Opportunities:**

* On-premises HPC can sometimes limit an organization’s opportunities to take full advantage of the latest technologies. For example, as organizations adopt leading-edge technologies like artificial intelligence/ machine learning technologies (AI/ML) and visualization, the complexity and volume of data is pushing on- premises infrastructure to its limits. Furthermore, most AI/ML algorithms are cloud-native. These algorithms will deliver superior performance on large data sets when running in the cloud, especially with workloads that involve transient data that does not need to be stored long term.

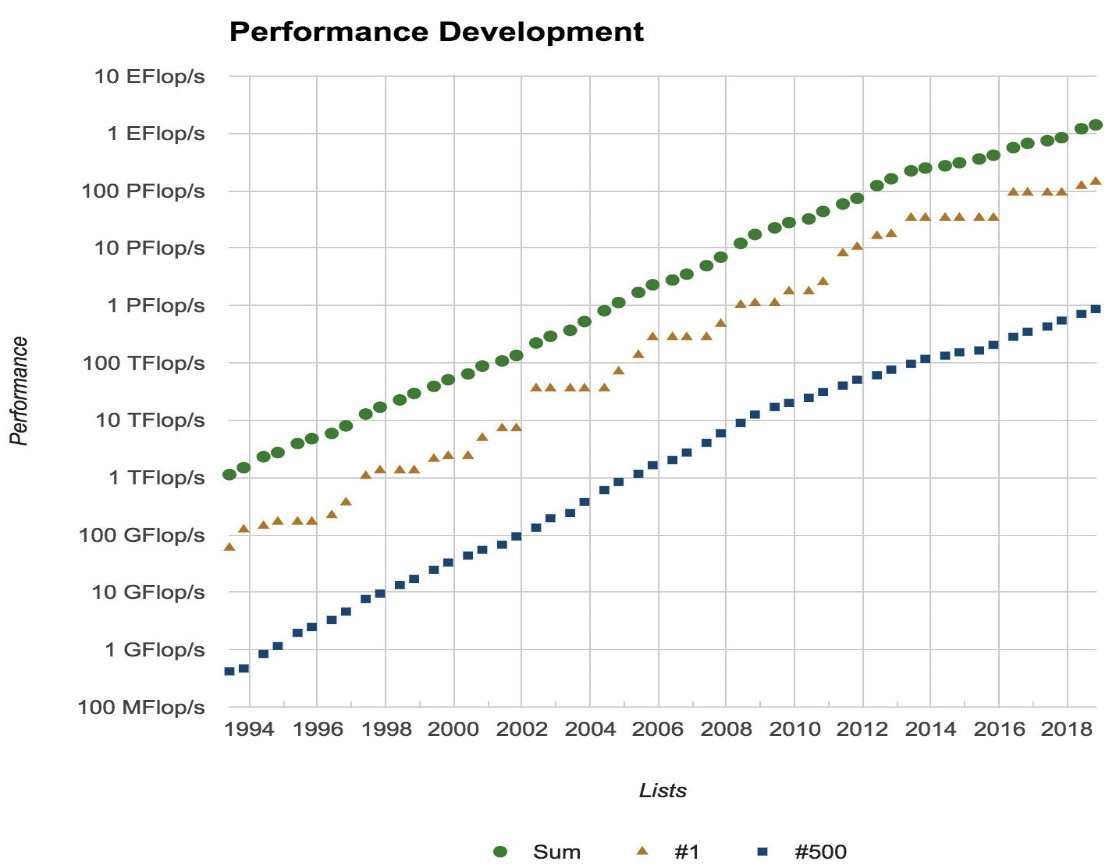
******CLOUD A BETTER WAY TO HPC**

To move beyond the limits of on-premises HPC, many organizations are leveraging cloud services to support their most advanced computing applications. Flexible and agile, the cloud offers strong advantages compared to traditional on-premises HPC approaches.

HPC on GCP, deliver significant leaps in compute performance, memory capacity, and bandwidth and I/O scalability. The highly customizable computing platform and robust partner community enable your staff to imagine new approaches so they can fail forward faster, delivering more answers to more questions without the need for costly, on-premises upgrades. In short, GCP frees you to rethink your approach to every HPC and big data analysis initiative and invites your team to ask questions and seek answers as often as possible**.**

**Growth Chart of the top 500 Supercomputers with the sum of the list now exceeding 1-exaflop:**





**Innovate Faster with a Highly Scalable Infrastructure:**

* Moving HPC workloads to the cloud can bring down barriers to innovation by opening up access to virtually unlimited capacity and scale. And one of the best features of working in a cloud environment is that when you solve a problem, it stays solved. You’re not revisiting it every time you do a major system-wide software upgrade or a bi-annual hardware refresh. Limits on scale and capacity with on-premises infrastructure, usually led to organizations being reluctant to consider new use cases or applications that exceeded their capabilities. Running HPC in the cloud enables asking the business-critical questions they couldn’t address before, and that means a fresh look at project ideas that were shelved due to infrastructure constraints.

**Increase Collaboration with Secure Access to Clusters Worldwide:**

* Running HPC workloads on the cloud enables a new way for globally distributed teams to collaborate securely. With globally-accessible shared data, engineers and researchers can work together or in parallel to get results faster. For example, the use of the cloud for collaboration and visualization allows a remote design team to view and interact with a simulation model in near real time, without the need to duplicate and proliferate sensitive design data. Using the cloud as a collaboration platform also makes it easier to ensure compliance with ever-changing industry regulations.
* The Google cloud is compliant with the latest revisions of GDPR, HIPAA, FISMA, FedRAMP, PCI, ISO 27001, SOC 1, and other regulations. Encryption and granular permission features guard sensitive data without interfering with the ability to share data across approved users, and detailed audit trails for virtually every API call or cloud orchestration action means environments can be designed to address specific governance needs and submit to continuous monitoring and surveillance. With a broad global presence and the wide availability of this technology-powered Google EC2 instances, HPC on GCP enables engineers and researchers to share and collaborate efficiently with team members across the globe without compromising on security.

**Optimize Cost with Flexible Resource Selection:**

* Running HPC in the cloud enables organizations to select and deploy an optimal set of services for their unique applications, and to pay only for what they use. Individuals and teams can rapidly scale up or scale down resources as needed, commissioning or decommissioning HPC clusters in minutes, instead of days or weeks. With HPC in the cloud, scientists, researchers, and commercial HPC users can gain rapid access to resources they need without a burdensome procurement process. Running HPC in the cloud also minimizes the need for job queues. Traditional HPC systems require researchers and analysts to submit their projects to open source or commercial cluster and job management tools, which can be time consuming and vulnerable to submission errors.
* GCP HPC solutions remove the traditional challenges associated with on-premises clusters: fixed infrastructure capacity, technology obsolescence, and high capital expenditures. GCP gives you access to virtually unlimited HPC capacity, built from the latest technologies. You can quickly migrate to newer, more powerful processor-based EC2 instances as soon as they are made available on GCP This removes the risk of on-premises CPU clusters becoming obsolete or poorly utilized as your needs change over time. As a result, your teams can trust that their workloads are running optimally at every stage.

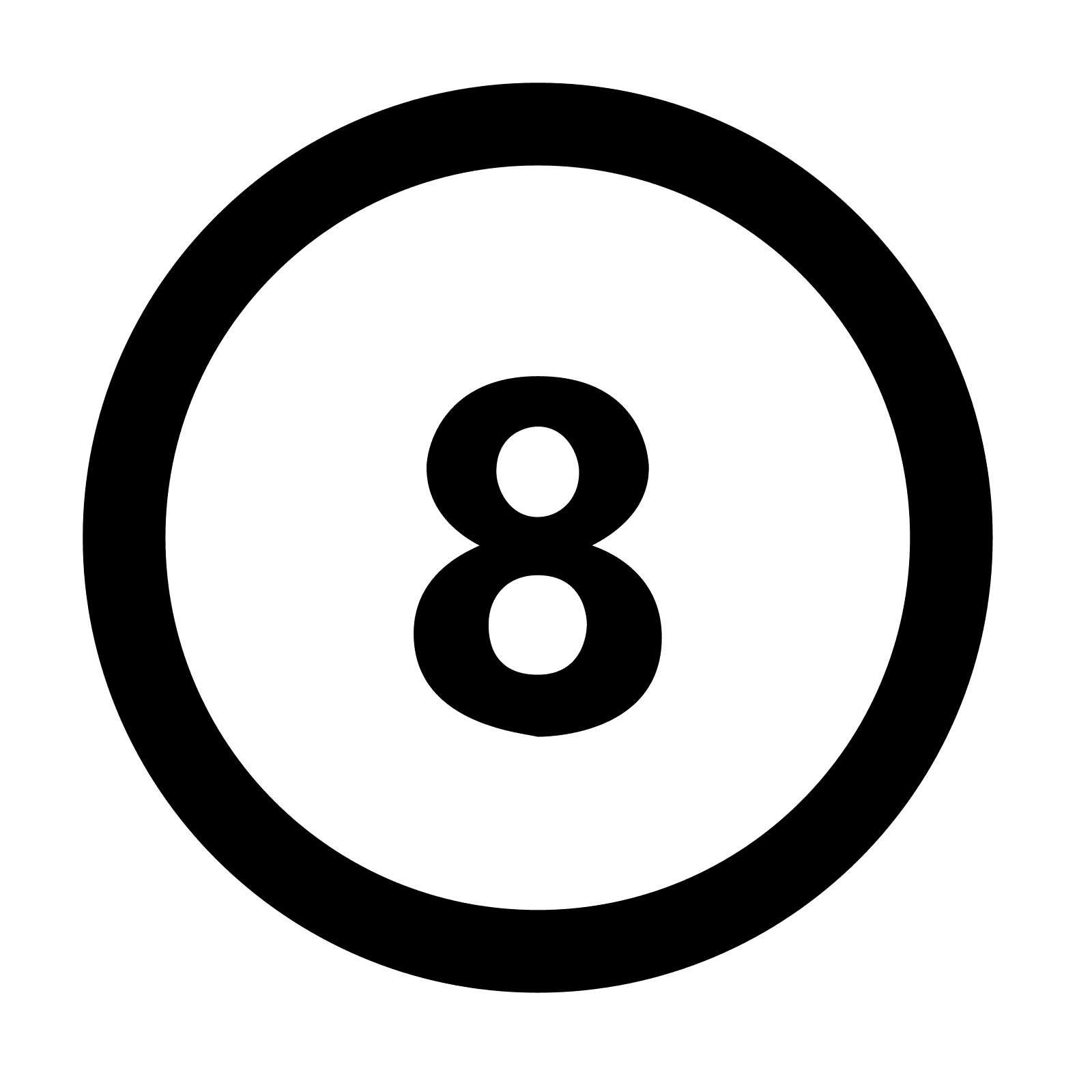
**GCP DELIVER A COMPLETE HPC SOLUTION**

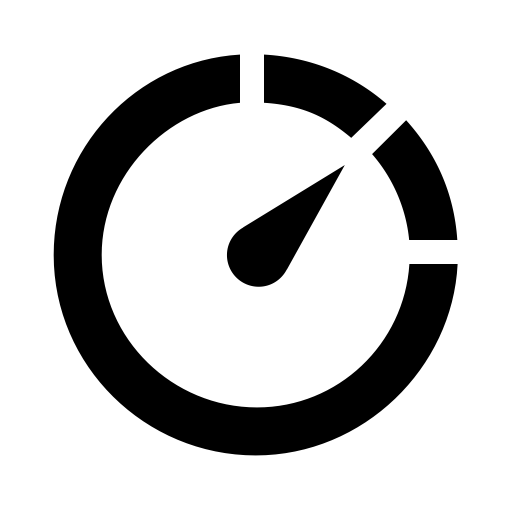
**Compute:**

The GCP HPC solution lets you choose from a variety of compute instance types that can be configured to suit your needs, including the latest processor- powered CPU instances, GPU-based instances, and field programmable gate array (FPGA)-powered instances.

The latest Google EC2 instances include the C5n, C5d and Z1d instances. C5n instances feature the processor with a sustained all core Turbo CPU clock speed of up to 3.5 GHz. C5n instances provide up to 100 Gbps of network bandwidth and up to 14 Gbps of dedicated bandwidth to Google EBS. C5n instances also feature 33% higher memory footprint compared to C5 instances. For workloads that require access to high-speed, ultra-low latency local storage, GCP offers C5d instances equipped with local NVMe-based SSDs. Google EC2 z1d instances offer both high compute capacity and a high memory footprint. High frequency z1d instances deliver a sustained all core frequency of up to 4.0 GHz, the fastest of any cloud instance. For HPC codes that can benefit from GPU acceleration, the Google EC2 P3dn instances feature 100 Gbps network bandwidth (up to 4x the bandwidth of previous P3 instances), local NVMe storage, the latest NVIDIA V100 Tensor Core GPUs with 32 GB of GPU memory, NVIDIA NVLink for faster GPU-to-GPU communication, processors running at 3.1 GHz sustained all-core Turbo. GCP Auto Scaling monitors your applications and automatically adjusts capacity to maintain steady, predictable performance at the lowest possible cost. Using GCP Auto Scaling, it’s easy to setup application scaling for multiple resources across multiple services in minutes.

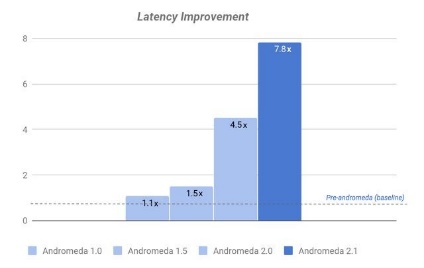
1. 

**Attach up to eight GPUs per instance**

**Attached directly to the VM via PCIe x16 to achieve near bare-metal performance**

1. **Per-second billing, Sustained Use Discounts Preemptible support (~70% off)**
2. **NVIDIA Tesla K80 NVIDIA Tesla P100 NVIDIA Tesla P4 NVIDIA Tesla V100 NVIDIA Tesla T4**

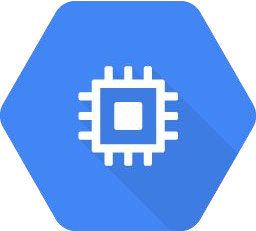
**Networking:**

1. Google EC2 instances support enhanced networking that allow EC2 instances to achieve higher bandwidth and lower inter-instance latency compared to traditional virtualization methods. Elastic Fabric Adapter (EFA) is a network interface for Google EC2 instances that enables you to run HPC applications requiring high levels of inter-node communications at scale on GCP. Its custom-built operating system (OS) bypass hardware interface enhances the performance of inter-instance communications, which is critical to scaling HPC applications. GCP also offers placement groups for tightly-coupled HPC applications that require low latency networking. Google Virtual Private Cloud (VPC) provides IP connectivity between compute instances and storage components.
2.  **Network/Bandwidth:**
3. 

**Storage:**

Storage options and storage costs are critical factors when considering an HPC solution. GCP offers flexible object, block, or file storage for your transient and permanent storage requirements. Google Elastic Block Store (Google EBS) provides persistent block storage volumes for use with Google EC2. Google Simple Storage Service (S3) is designed to store and access any type of data over the Internet and can be used to store the HPC input and output data long term and without ever having to do a data migration project again Google FSx for Lustre works natively with Google S3, making it easy for you to process cloud data sets with high performance file systems. When linked to an S3 bucket, an FSx for Lustre file system transparently presents S3 objects as files and allows you to write results back to S3. You can also use FSx for Lustre as a standalone high-performance file system to burst your workloads from on-premises to the cloud. By copying on-premises data to an FSx for Lustre file system, you can make that data available for fast processing by compute instances running on GCP. Google Elastic File System (Google EFS) provides simple, scalable file storage for use with Google EC2 instances in the GCP Cloud.



Google Cloud Storage

Exabyte-scale, feature-rich object storage, automatically scaling throughput

1. 

Persistent Disk

SSD/HDD Persistent Disk, High-performance, replicated block storage

1. 

Local Storage

Local SSD (NVMe) for scratch and fast access Physically attached to node via PCI

Cloud Filestore

Highly available, durable, POSIX-compliant shared storage across tens of thousands of nodes

Partner, hybrid, and open-source

Storage solutions for NetApp, Elastifile, DDN, Lustre, and more Move petabytes to GCS with the Data Transfer Appliance

**Security and Compliance:**

Security management and regulatory compliance are other important aspects of running HPC in the cloud. GCP offers multiple security related services and quick-launch templates to simplify the process of creating a HPC cluster and implementing best practices in data security and regulatory compliance. The GCP infrastructure puts strong safeguards in place to help protect customer privacy. All data is stored in highly secure GCP data centers. GCP Identity and Access Management (IAM) provides a robust solution for managing users, roles, and groups that have rights to access specific data sources. Organizations can issue users and systems individual identities and credentials, or provision them with temporary access credentials using the Google Security Token Service (Google STS). GCP manages dozens of compliance programs in its infrastructure. This means that segments of your compliance have already been completed. GCP infrastructure is compliant with many relevant industry regulations such as HIPAA, FISMA, FedRAMP, PCI, ISO 27001, SOC 1, and others.

1. **With HPC on GCP, organizations can flexibly tune and scale their infrastructure, as workloads dictate, instead of the other way around.**
2. ** DRIVING INNOVATION ACROSS INDUSTRIES**

**Life Sciences and Healthcare:**

1. Running HPC workloads on GCP lets healthcare and life sciences professionals easily and securely scale genomic analysis and precision medicine applications. For GCP users, the scalability is built-in, bolstered by an ecosystem of partners for tools and datasets designed for sensitive data and workloads. They can efficiently, dynamically store and compute their data, collaborate with peers, and integrate findings into clinical practice—while conforming with security and compliance requirements.
2. For example, Bristol-Myers Squibb (BMS), a global biopharmaceutical company, used GCP to build a secure, self-provisioning portal for hosting research. The solution lets scientists run clinical trial simulations on-demand and enables BMS to set up rules that keep compute costs low. Compute-intensive clinical trial simulations that previously took 60 hours are finished in only 1.2 hours on the GCP Cloud. Running simulations 98% faster has led to more efficient, less costly clinical trials—and better conditions for patients.

**Financial Services:**

1. Insurers and capital markets have long been utilizing grid computing to power actuarial calculations, determine capital requirements, model risk scenarios, price products, and handle other key tasks. Taking these compute-intensive workloads out of the data center and moving them to GCP helps them boost speed, scale better, and save money.

For example, MAPRE, the largest insurance company in Spain, needed fast, flexible environments in which to develop sales management insurance policy applications. The firm was looking for a cost-effective technology platform that could deliver rapid analysis and enable quick deployment of development environments in remote installations sites.

**Design and Engineering:**

Using simulations on GCP HPC infrastructure lets manufacturers and designers reduce costs by replacing expensive development of physical models with virtual ones during product development. The result? Improved product quality, shorter time to market, and reduced product development costs.

TLG Aerospace in Seattle, Washington put these capabilities to work to perform aerodynamic simulations on aircraft and predict the pressure and temperature surrounding airframes. Its existing cloud provider was expensive and could not scale to handle more performance-intensive applications. TLG turned to Google EC2 Spot instances, which provide a way to use unused EC2 computing capacity at a discounted price. The solution dramatically decreased simulation costs and can scale easily to take on new jobs as needed.



**Energy and Geo Sciences**

Reducing run-times for compute-intensive applications like seismic analysis and reservoir simulation is just one of the many ways the energy and geosciences industry has been utilizing HPC applications in the cloud. By moving HPC applications to the cloud, organizations reduce job submission time, track runtime, and efficiently manage the large data-sets associated with daily workloads.

For example, using GCP on-demand computing resources, Zenotech, a simulation service provider, can power simulations that help energy companies support advanced reservoir models.

1. 

**Media and Entertainment**

The movie and entertainment industries are shifting content production and post- production to cloud-based HPC to take advantage of highly scalable, elastic and secure cloud services, to accelerate content production and reduce capital infrastructure investment. Content production and post-production companies are leveraging the cloud to accelerate and streamline production, editing, and rendering workloads with highly scalable cloud computing and storage.

One design and visual effects (VFX) company, Fin Design + Effects, needed the ability to access vast amounts of compute capacity when big deadlines came around. Its on-premises render servers had a finite capacity and were difficult and expensive to scale. Fin started by using GCP Direct Connect to scale its rendering capabilities by establishing a dedicated Gigabit network connection from the Fin data center to GCP. Fin is also taking advantage of Google EC2 Spot instances. Fin now has the agility to add compute resources on the fly to meet last-minute project demands.

**AI/ML and Autonomous Vehicles**

The AI revolution which started with the rapid increase in accuracy brought by deep learning methods, has the potential to revolutionize a variety of industries. Autonomous driving is a particularly popular use case for AI/ML. Developing and deploying autonomous vehicles requires the ability to collect, store and manage massive amounts of data, high performance computing capacity and advanced deep learning frameworks, along with the capability to do real-time processing of local rules and events in the vehicle. GCP’s virtually unlimited storage and compute capacity and support for popular deep learning frameworks help accelerate algorithm training and testing and drive faster time to market.

**GCP cloud provides a broad range of scalable, flexible infrastructure solutions that organizations can select to match their workloads and tasks. This gives HPC users the ability to choose the most appropriate mix of resources for their specific applications.**

**CONCLUSION**

# Technology continues to change rapidly, and it’s clear that HPC has a critical role to play in enabling organizations to innovate faster, and enable them to adopt other leading-edge technologies like AI/ML and IoT. GCP puts the advanced capabilities of High-Performance Computing in reach for more people and organizations, while simplifying processes like management, deployment, and scaling. Accessible, flexible, and cost effective, it lets organizations unleash the creativity of their engineers, analysts, and researchers from the limitations of on-premises infrastructures.

# Unlike traditional on-premise HPC systems, GCP offers virtually unlimited capacity to scale out HPC infrastructure. It also provides the flexibility for organizations to adapt their HPC infrastructure to changing business priorities. With flexible deployment and pricing models, it lets organizations of all sizes and industries take advantage of the most advanced computing capabilities available. HPC on GCP lets you take a fresh approach to innovation to solve the world’s most complex problems.

**Hyderabad**

THE SKYVIEW 10

18th Floor, "Zone A"

Survey No 83/1, Raidurgam

Hyderabad – 500032

Tel: + 91 40 6736 2000

**Jamshedpur**

1st Floor, Shantiniketan Building

Holding No. 1, SB Shop Area

Bistupur, Jamshedpur – 831 001

Tel: + 91 657 663 1000

**Kochi**

9th Floor, ABAD Nucleus

NH-49, Maradu PO

Kochi - 682 304

Tel: + 91 484 433 4000

**Kolkata**

22 Camac Street

3rd Floor, Block ‘C’

Kolkata - 700 016

Tel: + 91 33 6615 3400

**Mumbai**

14th Floor, The Ruby

29 Senapati Bapat Marg

Dadar (W), Mumbai - 400 028

Tel: + 91 22 6192 0000

5th Floor, Block B-2

Nirlon Knowledge Park

Off. Western Express Highway

Goregaon (E)

Mumbai - 400 063

Tel: + 91 22 6192 0000

**Pune**

C-401, 4th floor

Panchshil Tech Park

Yerwada

(Near Don Bosco School)

Pune - 411 006

Tel: + 91 20 4912 6000

**Ahmedabad**

22nd Floor, B Wing, Privilon,

Ambli BRT Road, Behind Iskcon   
Temple, Off SG Highway,   
Ahmedabad - 380 015

Tel: + 91 79 6608 3800

**Bengaluru**

6th, 12th & 13th floor

“UB City”, Canberra Block

No.24 Vittal Mallya Road

Bengaluru - 560 001

Tel: + 91 80 6727 5000

Ground Floor, ‘A’ wing

Divyasree Chambers

# 11, O’Shaughnessy Road

Langford Gardens

Bengaluru - 560 025

Tel: + 91 80 6727 5000

**Chandigarh**

1st Floor, SCO: 166-167

Sector 9-C, Madhya Marg

Chandigarh - 160 009

Tel: + 91 172 331 7800

**Chennai**

Tidel Park, 6th & 7th Floor

A Block, No.4, Rajiv Gandhi Salai

Taramani, Chennai - 600 113

Tel: + 91 44 6654 8100

**Delhi NCR**

Golf View Corporate Tower B

Sector 42, Sector Road

Gurgaon - 122 002

Tel: + 91 124 443 4000

3rd & 6th Floor, Worldmark-1

IGI Airport Hospitality District

Aerocity, New Delhi - 110 037

Tel: + 91 11 4731 8000

4th & 5th Floor, Plot No 2B

Tower 2, Sector 126

NOIDA - 201 304

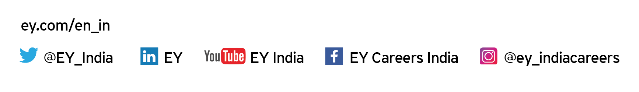
Gautam Budh Nagar, U.P.

Tel: + 91 120 671 7000

**Our offices**

Ernst & Young LLP

**EY** | Assurance | Tax | Strategy and Transactions | Consulting



About EY

EY is a global leader in assurance, tax, strategy, transaction and consulting services. The insights and quality services we deliver help build trust and confidence in the capital markets and in economies the world over. We develop outstanding leaders who team to deliver on our promises to all of our stakeholders. In so doing, we play a critical role in building a better working world for our people, for our clients and for our communities.

EY refers to the global organization, and may refer to one or more, of the member firms of Ernst & Young Global Limited, each of which is a separate legal entity. Ernst & Young Global Limited, a UK company limited by guarantee, does not provide services to clients. Information about how EY collects and uses personal data and a description of the rights individuals have under data protection legislation are available via ey.com/privacy. For more information about our organization, please visit ey.com.

Ernst & Young LLP is one of the Indian client serving member firms of EYGM Limited. For more information about our organization, please visit www.ey.com/en\_in.

Ernst & Young LLP is a Limited Liability Partnership, registered under the Limited Liability Partnership Act, 2008 in India, having its registered office at 22 Camac Street, 3rd Floor, Block C, Kolkata – 700016

© 2020 Ernst & Young LLP. Published in India.   
All Rights Reserved.

ED None

This publication contains information in summary form and is therefore intended for general guidance only. It is not intended to be a substitute for detailed research or the exercise of professional judgment. Neither EYGM Limited nor any other member of the global Ernst & Young organization can accept any responsibility for loss occasioned to any person acting or refraining from action as a result of any material in this publication. On any specific matter, reference should be made to the appropriate advisor.