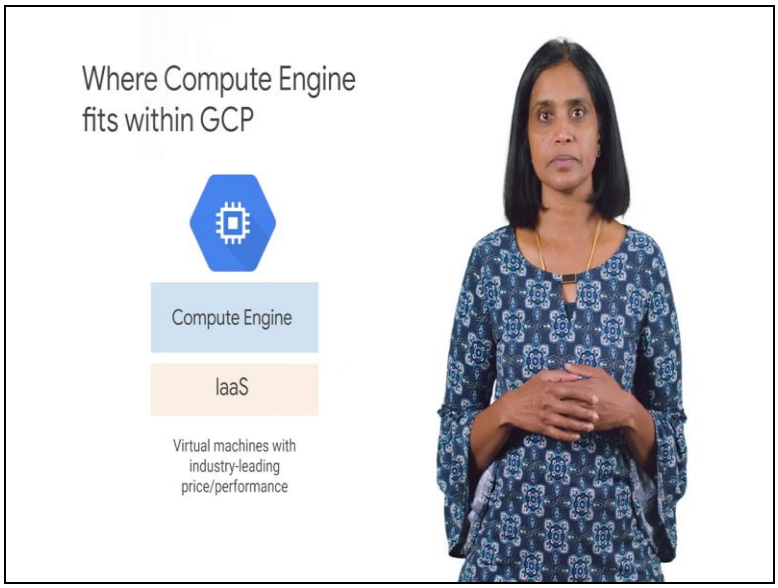
Google Cloud Computing Foundation Course
Sowmya Kannan
Google Cloud

Lecture-19
Exploring IaaS with Compute Engine

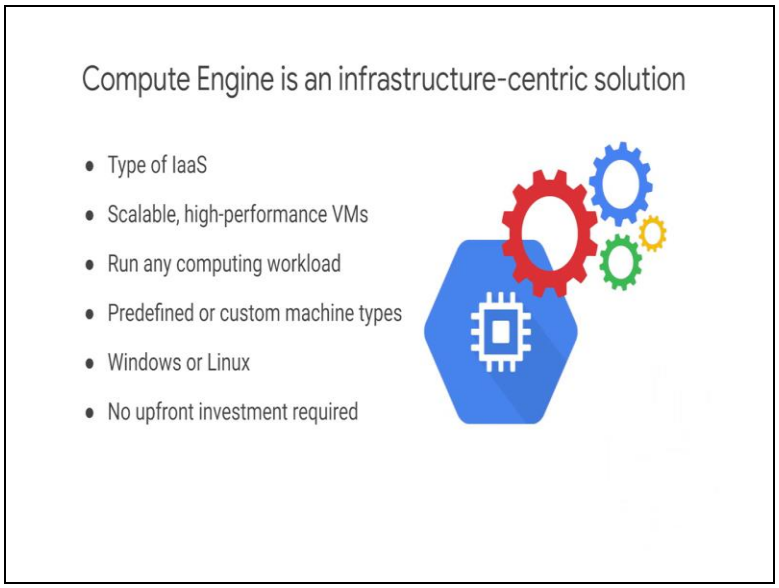
(Refer Slide Time: 00:07)



Next you will discover how to build and deploy applications with compute engine. Compute engine delivers virtual machines running in Google's innovative data centers and worldwide fiber network. Compute engine is ideal if you need complete control over the virtual machine infrastructure need to make changes to the kernel such as providing your own network or graphics drivers to squeeze out the last drop of performance.

Or if you need to run a software package that cannot easily be containerized or have existing VM images to move to the cloud.

(Refer Slide Time: 00:43)

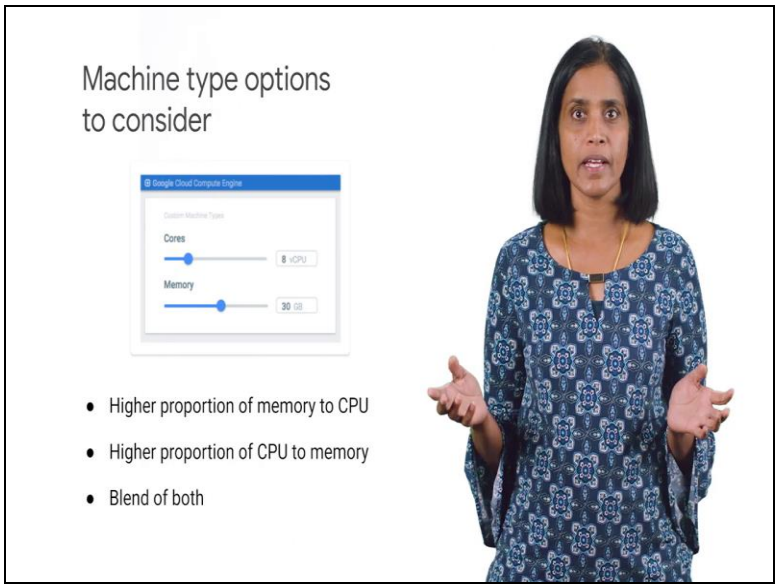


Compute engine is a type of infrastructure as a service. It delivers scaleable high performance virtual machines that run on Google's infrastructure. Compute engine VMs boot quickly come with persistent disk storage and deliver consistent performance. You can run any computing workload on compute engine such as web server hosting, application hosting and application backends. Virtual servers are available in many configurations including predefined sizes.

Alternatively there is the option to create custom machine types optimized for specific needs. Compute engine also allows users to run their choice of operating system. And while compute engine allows users to run thousands of virtual CPUs in a system that has been designed to be fast and offer strong performance consistency there is no upfront investment required. The purpose of custom virtual machines is to ensure you can create virtual services with just enough resources to work for your application.

For example you want to run your application on a virtual machine but none of the predefined versions will fit the resource footprint you require or your application needs to run on a specific CPU architecture or GPUs are required to run your application. Custom virtual machines allow for creating a perfect fit for your applications.

(Refer Slide Time: 02:28)



To meet your workload requirements there are different machine type options that you can consider. For example a higher proportion of memory to CPU a higher proportion of CPU to memory or a blend of both through Google's standard configuration. Compute engine offers predefined machine types that you can use when you create an instance. A predefined machine; type has a preset number of virtual CPUs or me CPUs and amount of memory and are charged at a set price.

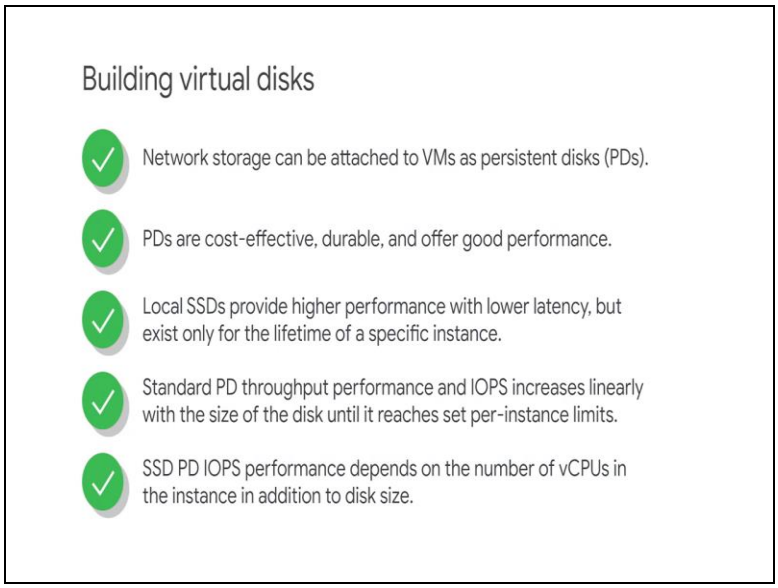
You can choose from the general-purpose machine types memory optimized machine types and compute optimized machine types. Predefined virtual machine configurations range from micro instances of 2v CPUs and 8 gigabytes of memory to memory optimized instances with up to 160 V CPUs and 3.75 terabytes of memory.

(Refer Slide Time: 03:42)

Compute engine also allows you to create virtual machines with the V CPU and memory that meet workload requirements. This has performance benefits and also reduces cost significantly. One option is to select from predefined configurations. A general-purpose configuration provides a balance between performance and memory or you can optimize for memory or for performance. You can create a machine type with as little as one v CPU and up to ATV CPUs or any even number of V CPUs in between.

You can configure up to 8 gigabytes of memory per V CPU. Alternatively if none of the predefined virtual machines fit your needs you have the option to create a custom virtual machine. When you create a custom virtual machine you can choose the number of CPUs the amount of memory required, the CPU architecture to leverage and the option of using GPUs.

(Refer Slide Time: 04:54)



Network storage up to 64 terabytes in size can be attached to VMs as persistent disks. Persistent disks are the most common storage option due to their price performance and durability and can be created in HDD or SSD formats. If a VM instance is terminated its persistent disk retains data and can be attached to another instance. You can also take snapshots of your persistent disk and create new persistent disks from that snapshot.

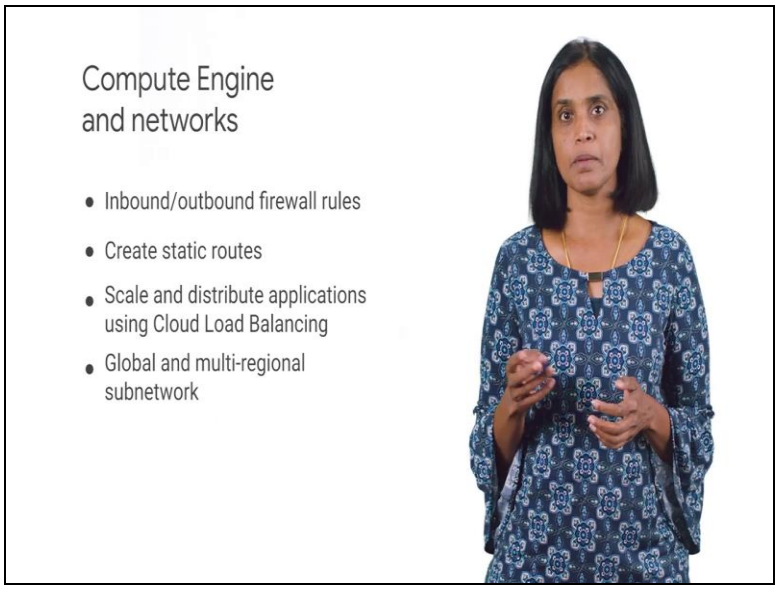
Compute engine offers always encrypted local SSD block storage unlike standard persistent disks local SSDs are physically attached to the server hosting the VM instance offering very high input output operations per second and very low latency compared to persistent disks. Predefined local SSD sizes up to 3 terabytes are available for any VM with at least one V CPU. By default most compute engine provided Linux images will automatically run an optimization script that configures the instance for peak local SSD performance.

Standard persistent disk performance scales linearly up to the VM performance limits. A V CPU counter 4 or more for your instance does not limit the performance of standard persistent disks. A V CPU count of less than 4 for an inn stance reduces the right limit for input output operations per second or IOPS because network egress limits are proportional to the V CPU count. The right limit also depends on the size of input outputs or iOS.

For example 16 kilobyte iOS consume more bandwidth than 8 kilobyte iOS at the same IOPS level. Standard persistent disk IOPS and throughput performance increase linearly with the size of the disk until it reaches set per instance limits. The IOPS performance of SSD persistent disks depends on the number of V CPUs in the instance in addition to disk size. Lower core VMs have lower right IOPS and throughput limits due to the network egress limitations on write throughput.

SSD persistent disk performance scales linearly until it reaches either the limits of the volume or the limits of each compute engine instance. SSD read bandwidth and IOPS consistency near the maximum limits largely depends on network ingress utilization. Some variability is to be expected especially for 16 kilobyte iOS near the maximum IOPS limits.

(Refer Slide Time: 08:18)

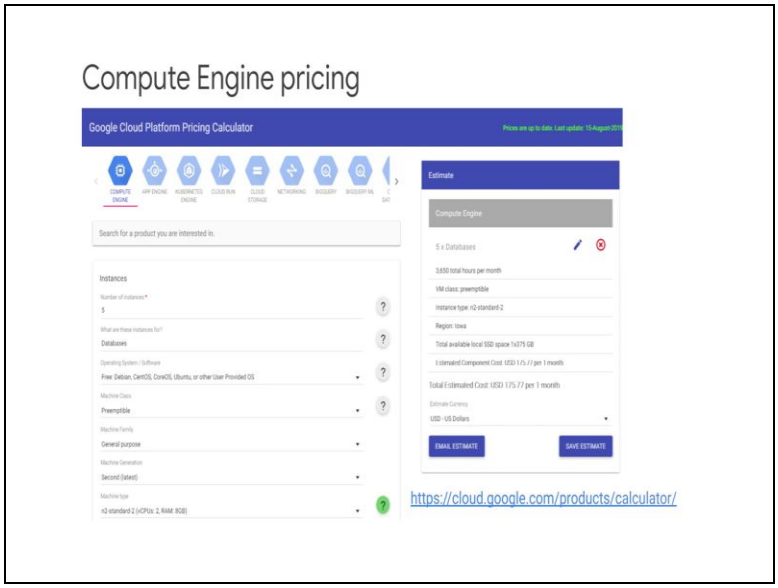


Networks connect compute engine instances to each other and to the internet. Networks in the cloud have a lot of similarities with physical networks. You can segment networks use firewall rules to restrict access to instances and create static routes to forward traffic to specific destinations. You can scale up applications on compute engine from 0 to full throttle with cloud load balancing. Distribute your load balance compute resources in single or multiple regions to users and to meet your high availability requirements.

Sub-networks segment your cloud network IP space sub network prefixes can be automatically allocated or you can create a custom topology sub networks and cloud load balancing are both discussed in the module it helps to network. When you build a compute engine instance you use a virtual network adapter which is part of the instance to connect a virtual machine to a network. Much in the same way you would connect a physical server to a network.

For compute engine you can have up to 8 virtual adapters sub networks and cloud load balancing are both discussed in the module it helps to network.

(Refer Slide Time: 09:49)



All virtual machines are charged for one minute at boot time which is the minimum charge for a beer after that per second pricing begins meaning that you only pay for the compute time used. Google offers sustained use discounts which automatically provide discounted prices for long-running workloads without the need for signup fees or any upfront commitment. Predefined machine types are discounted based on the percent of monthly use.

While custom machine types are discounted on a percent of total use. The GCP pricing calculator is a great way to see pricing estimates based on the different configuration options that are available and instances salt-and in notes, persistent disks load balancing and cloud tepees.