Compute Engine Fundamentals 22 In corporate data centers, applications are deployed to physical servers Where do you deploy applications in the cloud? Rent virtual servers Virtual Machines - Virtual servers in GCP Google Compute Engine (GCE) - Provision & Manage Virtual Machines Google Compute Engine (GCE) 23 Compute Engine - Features Create and manage lifecycle of Virtual Machine (VM) instances Load balancing and auto scaling for multiple VM instances Attach storage (& network storage) to your VM instances Manage network connectivity and configuration for your VM instances Our Goal: Setup VM instances as HTTP (Web) Server Distribute load with Load Balancers 24 Let's create a few VM instances and play with them Let's check out the lifecycle of VM instances Let's use SSH to connect to VM instances Compute Engine Hands-on 25 What type of hardware do you want to run your workloads on? Different Machine Families for Different Workloads: General Purpose (E2, N2, N2D, N1) : Best price-performance ratio Web and application servers, Small-medium databases, Dev environments Memory Optimized (M2, M1): Ultra high memory workloads Large in-memory databases and In-memory analytics Compute Optimized (C2): Compute intensive workloads Gaming applications Compute Engine Machine Family 26 Compute Engine Machine Types How much CPU, memory or disk do you want? Variety of machine types are available for each machine family Let's take an example : e2-standard-2: e2 - Machine Type Family standard - Type of workload 2 - Number of CPUs Memory, disk and networking capabilities increase along with vCPUs 27 Image What operating system and what software do you want on the instance? Type of Images: Public Images: Provided & maintained by Google or Open source communities or third party vendors Custom Images: Created by you for your projects 28 Compute Engine Hands-on : Setting up a HTTP server Commands: sudo su - execute commands as a root user apt update - Update package index - pull the latest changes from the APT repositories apt -y install apache2 - Install apache 2 web server sudo service apache2 start - Start apache 2 web server echo "Hello World" > /var/www/html/index.html - Write to index.html $(hostname) - Get host name $(hostname -I) - Get host internal IP address #! /bin/bash sudo su apt update apt -y install apache2 sudo service apache2 start sudo update-rc.d apache2 enable echo "Hello World" > /var/www/html/index.html echo "Hello world from $(hostname) $(hostname -I)" > /var/www/html/index.html 29 External (Public) IP addresses are Internet addressable. Internal (Private) IP addresses are internal to a corporate network You CANNOT have two resources with same public (External) IP address. HOWEVER, two different corporate networks CAN have resources with same Internal (private) IP address All VM instances are assigned at least one Internal IP address Creation of External IP addresses can be enabled for VM instances (Remember) When you stop an VM instance, External IP address is lost DEMO: VM instances - Internal and External IPs Internal and External IP Addresses 30 Scenario : How do you get a constant External IP address for a VM instance? Quick and dirty way is to assign an Static IP Address to the VM! DEMO: Using Static IP Address with an VM instance Static IP Addresses 31 Static IP can be switched to another VM instance in same project Static IP remains attached even if you stop the instance. You have to manually detach it. Remember : You are billed for an Static IP when you are NOT using it! Make sure that you explicitly release an Static IP when you are not using it. Static IP Addresses - Remember 32 How do we reduce the number of steps in creating an VM instance and setting up a HTTP Server? Let's explore a few options: Startup script Instance Template Custom Image Simplify VM HTTP server setup 33 Bootstrapping with Startup script Bootstrapping: Install OS patches or software when an VM instance is launched. In VM, you can configure Startup script to bootstrap DEMO - Using Startup script #!/bin/bash apt update apt -y install apache2 echo "Hello world from $(hostname) $(hostname -I)" > /var/www/html/index.html 34 Why do you need to specify all the VM instance details (Image, instance type etc) every time you launch an instance? How about creating a Instance template? Define machine type, image, labels, startup script and other properties Used to create VM instances and managed instance groups Provides a convenient way to create similar instances CANNOT be updated To make a change, copy an existing template and modify it (Optional) Image family can be specified (example - debian-9): Latest non-deprecated version of the family is used DEMO - Launch VM instances using Instance templates Instance templates 35 Installing OS patches and software at launch of VM instances increases boot up time How about creating a custom image with OS patches and software pre-installed? Can be created from an instance, a persistent disk, a snapshot, another image, or a file in Cloud Storage Can be shared across projects (Recommendation) Deprecate old images (& specify replacement image) (Recommendation) Hardening an Image - Customize images to your corporate security standards Prefer using Custom Image to Startup script DEMO : Create a Custom Image and using it in an Instance Template Reducing Launch Time with Custom Image 36 Source: Automatic discounts for running VM instances for significant portion of the billing month Example: If you use N1, N2 machine types for more than 25% of a month, you get a 20% to 50% discount on every incremental minute. Discount increases with usage (graph) No action required on your part! Applicable for instances created by Google Kubernetes Engine and Compute Engine RESTRICTION: Does NOT apply on certain machine types (example: E2 and A2) RESTRICTION: Does NOT apply to VMs created by App Engine flexible and Dataflow Sustained use discounts https://cloud.google.com 37 For workloads with predictable resource needs Commit for 1 year or 3 years Up to 70% discount based on machine type and GPUs Applicable for instances created by Google Kubernetes Engine and Compute Engine RESTRICTION: Does NOT apply to VMs created by App Engine flexible and Dataflow Committed use discounts 38 Short-lived cheaper (upto 80%) compute instances Can be stopped by GCP any time (preempted) within 24 hours Instances get 30 second warning (to save anything they want to save) Use Preempt VM's if: Your applications are fault tolerant You are very cost sensitive Your workload is NOT immediate Example: Non immediate batch processing jobs RESTRICTIONS: NOT always available NO SLA and CANNOT be migrated to regular VMs NO Automatic Restarts Free Tier credits not applicable Preemptible VM 39 You are billed by the second (after a minimum of 1 minute) You are NOT billed for compute when a compute instance is stopped However, you will be billed for any storage attached with it! (RECOMMENDATION) Always create Budget alerts and make use of Budget exports to stay on top of billing! What are the ways you can save money? Choose the right machine type and image for your workload Be aware of the discounts available: Sustained use discounts Committed use discounts Discounts for preemptible VM instances Google Compute Engine - Billing 40 Compute Engine : Live Migration & Availability Policy How do you keep your VM instances running when a host system needs to be updated (a software or a hardware update needs to be performed)? Live Migration Your running instance is migrated to another host in the same zone Does NOT change any attributes or properties of the VM SUPPORTED for instances with local SSDs NOT SUPPORTED for GPUs and preemptible instances Important Configuration - Availability Policy: On host maintenance: What should happen during periodic infrastructure maintenance? Migrate (default): Migrate VM instance to other hardware Terminate: Stop the VM instance Automatic restart - Restart VM instances if they are terminated due to non-user-initiated reasons (maintenance event, hardware failure etc.) 41 What do you do when predefined VM options are NOT appropriate for your workload? Create a machine type customized to your needs (a Custom Machine Type) Custom Machine Type: Adjust vCPUs, memory and GPUs Choose between E2, N2, or N1 machine types Supports a wide variety of Operating Systems: CentOS, CoreOS, Debian, Red Hat, Ubuntu, Windows etc Billed per vCPUs, memory provisioned to each instance Example Hourly Price: $0.033174 / vCPU + $0.004446 / GB Compute Engine Features: Custom Machine Types 42 How do you accelerate math intensive and graphics-intensive workloads for AI/ML etc? Add a GPU to your virtual machine: High performance for math intensive and graphics-intensive workloads Higher Cost (REMEMBER) Use images with GPU libraries (Deep Learning) installed OTHERWISE, GPU will not be used GPU restrictions: NOT supported on all machine types (For example, not supported on shared-core or memory-optimized machine types) On host maintenance can only have the value "Terminate VM instance" Recommended Availability policy for GPUs Automatic restart - on Compute Engine Features: GPUs 43 Associated with a project Machine type availability can vary from region to regions You can only change the machine type (adjust the number of vCPUs and memory) of a stopped instance You CANNOT change the machine type of a running instance VM's can be filtered by various properties Name, Zone, Machine Type, Internal/External IP, Network, Labels etc Instances are Zonal (Run in a specific zone (in a specific region)) Images are global (You can provide access to other projects - if needed) Instance templates are global (Unless you use zonal resources in your templates) Automatic Basic Monitoring is enabled Default Metrics: CPU utilization, Network Bytes (in/out), Disk Throughput/IOPS For Memory Utilization & Disk Space Utilization - Cloud Monitoring agent is needed Virtual Machine - Remember 44 Choose Zone and Region based on: Cost, Regulations, Availability Needs, Latency and Specific Hardware needs Distribute instances in multiple zones and regions for high availability Choose right machine type for you needs: Play with them to find out the right machine type Use GPUs for Math and Graphic intensive applications Reserve for "committed use discounts" for constant workloads Use preemptible instances for fault-tolerant, NON time critical workloads Use labels to indicate environment, team, business unit etc Virtual Machine - Best Practices 45 Compute Engine Scenarios Scenario Solution What are the pre-requisites to be able to create a VM instance? 1. Project 2. Billing Account 3. Compute Engines APIs should be enabled You want dedicated hardware for your compliance, licensing, and management needs Sole-tenant nodes I have 1000s of VM and I want to automate OS patch management, OS inventory management and OS configuration management (manage software installed) Use "VM Manager" You want to login to your VM instance to install software You can SSH into it You do not want to expose a VM to internet Do NOT assign an external IP Address You want to allow HTTP traffic to your VM Configure Firewall Rules 46 Quick Review Image What operating system and what software do you want on the VM instance? Reduce boot time and improve security by creating custom hardened Images. You can share an Image with other projects Machine Types Optimized combination of compute(CPU, GPU), memory, disk (storage) and networking for specific workloads. You can create your own Custom Machine Types when existing ones don't fit your needs 47 Quick Review Static IP Addresses: Get a constant IP addresses for VM instances Instance Templates: Pre-configured templates simplifying the creation of VM instances Sustained use discounts: Automatic discounts for running VM instances for significant portion of the billing month Committed use discounts: 1 year or 3 year reservations for workloads with predictable resource needs Preemptible VM: Short-lived cheaper (upto 80%) compute instances for nontime-critical fault-tolerant workloads