Instance Groups 61 How do you create a group of VM instances? Instance Group - Group of VM instances managed as a single entity Manage group of similar VMs having similar lifecycle as ONE UNIT Two Types of Instance Groups: Managed : Identical VMs created using a template: Features: Auto scaling, auto healing and managed releases Unmanaged : Different configuration for VMs in same group: Does NOT offer auto scaling, auto healing & other services NOT Recommended unless you need different kinds of VMs Location can be Zonal or Regional Regional gives you higher availability (RECOMMENDED) Instance Groups 62 Managed Instance Group - Identical VMs created using an instance template Important Features: Maintain certain number of instances If an instance crashes, MIG launches another instance Detect application failures using health checks (Self Healing) Increase and decrease instances based on load (Auto Scaling) Add Load Balancer to distribute load Create instances in multiple zones (regional MIGs) Regional MIGs provide higher availability compared to zonal MIGs Release new application versions without downtime Rolling updates: Release new version step by step (gradually). Update a percentage of instances to the new version at a time. Canary Deployment: Test new version with a group of instances before releasing it across all instances. Managed Instance Groups (MIG) 63 Instance template is mandatory Configure auto-scaling to automatically adjust number of instances based on load: Minimum number of instances Maximum number of instances Autoscaling metrics: CPU Utilization target or Load Balancer Utilization target or Any other metric from Stack Driver Cool-down period: How long to wait before looking at auto scaling metrics again? Scale In Controls: Prevent a sudden drop in no of VM instances Example: Don't scale in by more than 10% or 3 instances in 5 minutes Autohealing: Configure a Health check with Initial delay (How long should you wait for your app to initialize before running a health check?) Time for a Demo Creating Managed Instance Group (MIG) 64 Rolling update - Gradual update of instances in an instance group to the new instance template Specify new template: (OPTIONAL) Specify a template for canary testing Specify how you want the update to be done: When should the update happen? Start the update immediately (Proactive) or when instance group is resized later(Opportunistic) How should the update happen? Maximum surge: How many instances are added at any point in time? Maximum unavailable: How many instances can be offline during the update? Rolling Restart/replace: Gradual restart or replace of all instances in the group No change in template BUT replace/restart existing VMs Configure Maximum surge, Maximum unavailable and What you want to do? (Restart/Replace) Updating a Managed Instance Group (MIG) 65 gcloud compute instance-groups managed Create instance group: create gcloud compute instance-groups managed create my-mig --zone us-central1-a --template myinstance-template --size 1 --health-check=HEALTH\_CHECK: How do you decide if an instance is healthy? --initial-delay: How much time should you give to an instance to start? Other similar commands - gcloud compute instance-groups managed delete/describe/list Setup Autoscaling: set-autoscaling/stop-autoscaling gcloud compute instance-groups managed set-autoscaling my-mig --max-num-replicas=10 --cool-down-period (default - 60s): How much time should Auto Scaler wait after initiating an autoscaling action? --scale-based-on-cpu --target-cpu-utilization --scale-based-on-load-balancing --target-load-balancing-utilization --min-num-replicas --mode (off/on(default)/only-scale-out) gcloud compute instance-groups managed stop-autoscaling my-mig Update existing MIG policies (ex: auto healing policies): gcloud compute instance-groups managed update my-mig --initial-delay: How much time should you give to the instance to start before marking it as unhealthy? --health-check: How do you decide if an instance is healthy? Playing with Managed Instance Groups - Command Line 66 Managed Instance Group - Command Line - Making Updates Resize the group: gcloud compute instance-groups managed resize my-mig --size=5 Recreate one or more instances (delete and recreate instances): gcloud compute instance-groups managed recreate-instances my-mig --instances=myinstance-1,my-instance-2 Update specific instances: gcloud compute instance-groups managed update-instances my-mig --instances=my-instance3,my-instance-4 (Update specific instances from the group) --minimal-action=none(default)/refresh/replace/restart --most-disruptive-allowed-action=none(default)/refresh/replace/restart Update instance template: gcloud compute instance-groups managed set-instance-template my-mig --template=v2- template After updating instance template, you can trigger roll out of the new template using update-instances, recreateinstances or rolling-action start-update commands 67 Managed Instance Groups - Command Line - Rolling Actions Scenario: You want to manage your new release - v1 to v2 - without downtime gcloud compute instance-groups managed rolling-action Restart(stop & start)- gcloud compute instance-groups managed rolling-action restart mymig --max-surge=5 or 10% (Max no of instances updated at a time) Replace(delete & recreate)- gcloud compute instance-groups managed rolling-action replace my-mig --max-surge=5 or 10% (Max no of instances updated at a time) --max-unavailable=5 or 10% (Max no of instances that can be down for the update) --replacement-method=recreate/substitute (substitute (default) creates instances with new names. recreate reuses names) Updates instances to a new template: Basic Version (Update all instances slowly step by step) - gcloud compute instance-groups managed rollingaction start-update my-mig --version=template=v1-template Canary Version (Update a subset of instances to v2) - gcloud compute instance-groups managed rolling-action start-update my-mig --version=template=v1-template --canary-version=template=v2-template,target-size=10% Options: --max-surge, --max-unavailable, --replacement-method 68 I want to ensure that I have one healthy instance running all the time: gcloud compute instance-groups managed set-autoscaling my-group --max-numreplicas=1 --min-num-replicas=1 I want to make a new release with no reduction in available number of instances. I want to update one instance at a time: gcloud compute instance-groups managed rolling-action start-update my-group -- version=template=my-v1-template --max-surge 1 --max-unavailable 0 Playing with Managed Instance Groups - Scenarios 69 Instance Group Scenarios Scenario Solution You want MIG managed application to survive Zonal Failures Create multiple zone MIG (or regional MIG) You want to create VMs of different configurations in the same group Create Un-managed Instance Group You want to preserve VM state in an MIG Stateful MIG - Preserve VM state (Instance name, attached Persistent disks and Metadata). Recommended for stateful workloads (database, data processing apps) You want high availability in an MIG even when there are hardware/software updates Use an instance template with availability policy automatic restart: enabled & on-host maintenance: migrate Ensures live migration and automatic restarts You want unhealthy instances to be automatically replaced Configure health check on the MIG (self healing) Avoid frequent scale up & downs Cool-down period/Initial delay