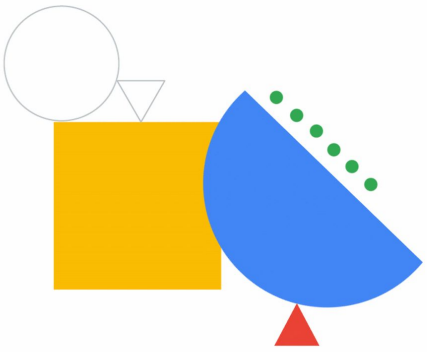
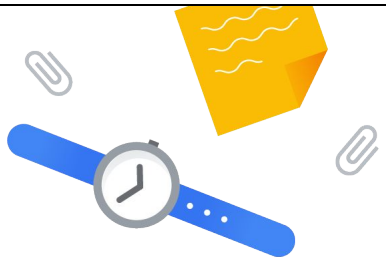
Google Cloud

Fundamentals of Cloud Run



In this module, we discuss the fundamentals of developing and running applications on the Cloud Run platform.



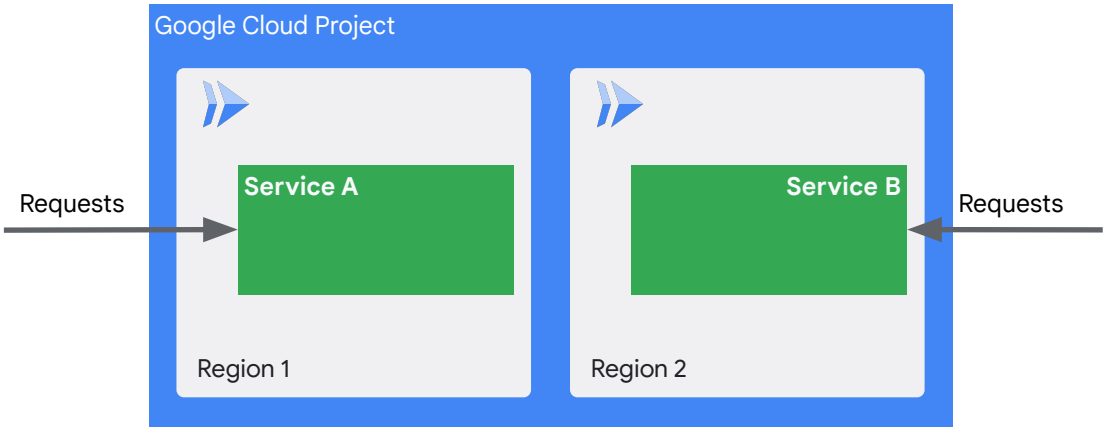
01 Overview

02 Resource model

Agenda

Let’s now discuss the Cloud Run resource model.

Cloud Run services



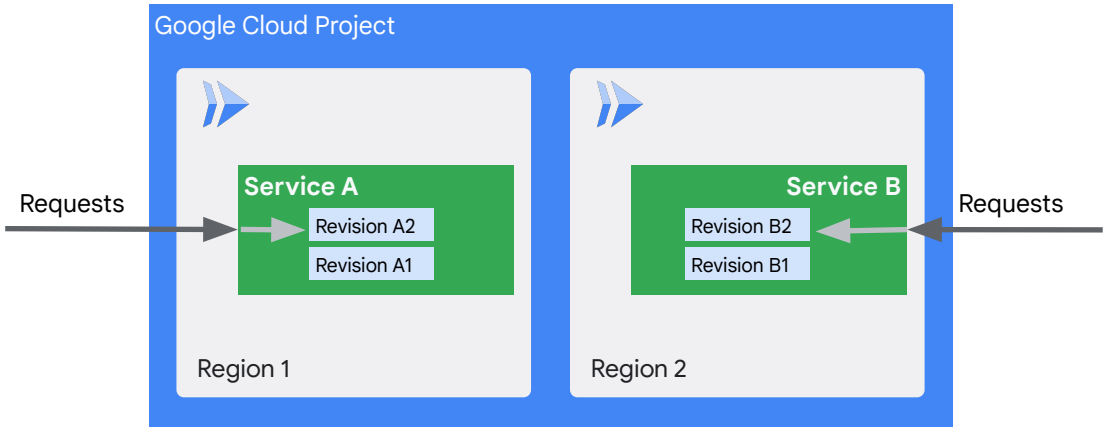
The service is the main resource of Cloud Run. Each service is located in a specific Google Cloud region where Cloud Run is available.

Services are a regional resource and container instances of a service can start in any zone in the region. For redundancy, services with high traffic and many container instances are spread out over multiple zones in the region. This means that if Cloud Run is experiencing issues in one zone, your service will continue to serve requests.

A given Google Cloud project can run many services in different regions.

Each service exposes a unique endpoint and automatically scales the underlying infrastructure to handle incoming requests.

Cloud Run revisions

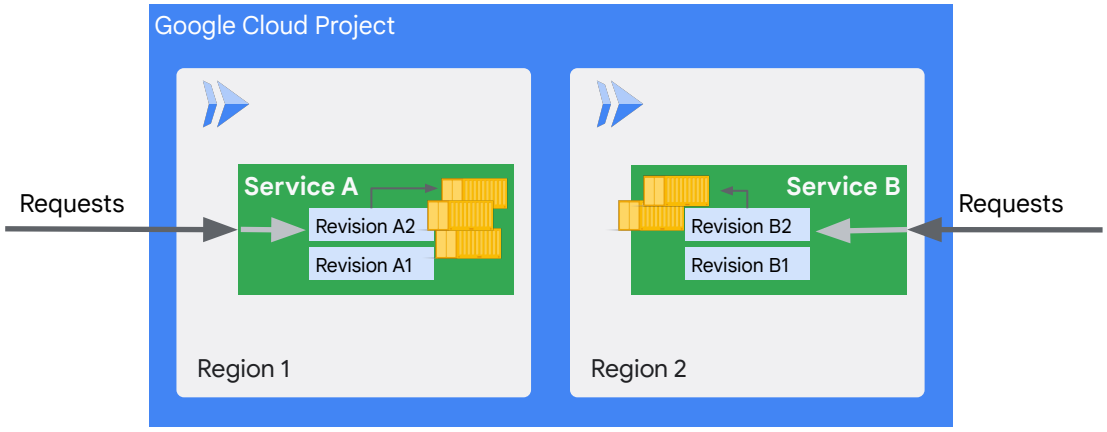


Each deployment of your application container image to Cloud Run creates a service revision. A revision consists of a specific container image, along with environment settings such as environment variables, memory limits, or concurrency value.

Revisions are immutable. Once a revision has been created, it cannot be modified. For example, when you deploy a container image to a new Cloud Run service, the first revision is created. If you then modify your application code and deploy a different container image to that same service, a second revision is created.

Requests to your application are automatically routed as soon as possible to the latest healthy service revision.

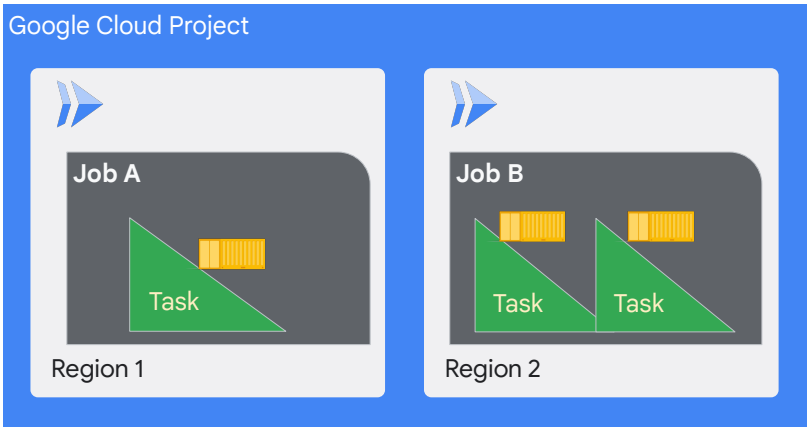
Cloud Run revisions



Each service revision that receives requests is automatically scaled with the number of container instances needed to handle all these requests.

A container instance can receive many requests at the same time. With the concurrency setting, you can set the maximum number of requests that can be sent in parallel to a given container instance.

Cloud Run jobs



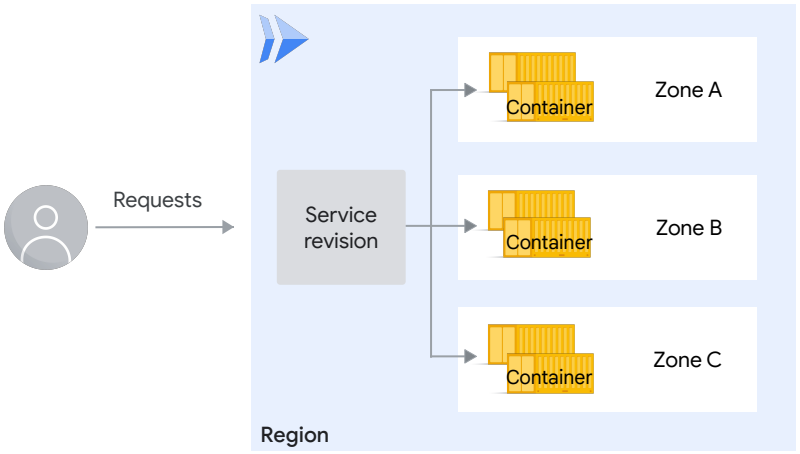
Each job is located in a specific Google Cloud region.

A job consists of one or multiple independent tasks that are executed in parallel in a given job execution. Each task runs one container instance.

When a job is executed, a job execution is created in which all job tasks are started. All tasks in a job execution must complete successfully for the job execution to be successful.

To handle task failures, you can set timeouts on tasks and specify the number of retries.

Regions and zones



A **region** is a geographic location where cloud resources are hosted. (Iowa, North America)

A region has three or more **zones.** A zone is a deployment area for cloud resources within a region.

Cloud Run is a regional service that lets you choose a region where your containers are deployed. A region is a specific geographical location where your Google Cloud resources are hosted.

A region consists of three or more zones. Zones and regions are logical abstractions of underlying physical resources that are provided in one or more data centers. An example of a region is *us-central1* in Iowa, North America.

A zone is a deployment area for cloud resources within a region. Zones are considered to single failure domains within a region.

For high availability, Cloud Run distributes your containers over multiple zones in a region, making your application resilient against the failure of a zone.

Remember

1. The main resource in Cloud Run is a service.
2. A service consists of one or more revisions. Service revisions are immutable.
3. A service revision consists of a specific container image, and configuration such as environment variables, or concurrency value.
4. A container instance handles requests to a service revision.
5. On Cloud Run, you can also run your code as a job that performs work and quits when done.

In summary:

- The main resource in Cloud Run is a service. Each service is located in a specific Google Cloud region where Cloud Run is available.
- A service consists of one or more revisions. Each deployment of your application container image to Cloud Run creates a service revision. Revisions are immutable.
- A revision consists of a specific container image, along with environment settings such as environment variables, memory limits, or concurrency value.
- A container instance handles requests to a service revision.
- On Cloud Run, you can also run your code as a job. Jobs are used to run code that performs work and quits when the work is done.