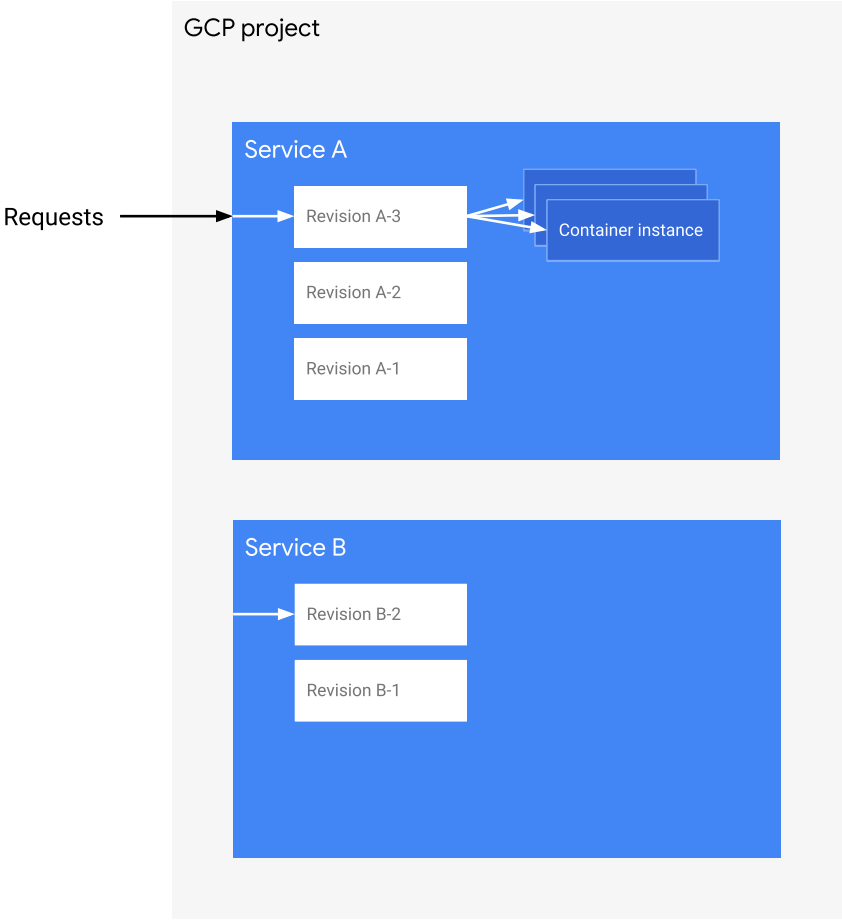
Serverless Computing (https://cloud.google.com/products/serverless/)
Cloud Run: Serverless Computing (https://cloud.google.com/run/)
Documentation (https://cloud.google.com/run/docs/) Guides

Resource model

The following diagram shows the Cloud Run resource model:



The diagram shows a Google Cloud project containing two Cloud Run services, `Service A` and `Service B`, each of which has several revisions.

In the diagram, `Service A` is receiving many requests, which results in the startup and running of several container instances. Note that `Service B` is not currently receiving requests, so no container instance is started yet.

Cloud Run services

The service is the main resource of Cloud Run. Each service is located in a specific GCP region (https://cloud.google.com/compute/docs/regions-zones/) (Cloud Run) or in a GKE cluster namespace (https://cloud.google.com/blog/products/gcp/kubernetes-best-practices-organizing-with-namespaces) (Cloud Run for Anthos on Google Cloud). For redundancy and failover, services are automatically replicated across multiple zones in the region they are in. A given GCP project can run many services in different regions or GKE clusters.

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

Cloud Run revisions

Each deployment to a service creates a 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 deploy a different container image to that same service, a second revision is created. If you subsequently set an environment variable, a third revision is created, and so on.

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

Cloud Run container instances

Each revision receiving requests is automatically scaled to the number of container instances needed to handle all these requests. Note that a container instance can receive many requests at the same time. With the concurrency setting (https://cloud.google.com/run/docs/configuring/concurrency), you can set the maximum number of requests that can be sent in parallel to a given container instance.

Last updated December 4, 2019.