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

Configuring Memory Limits

This page describes how to set memory limits.

Understanding memory usage

Cloud Run container instances that exceed their allowed memory limit are terminated.

The following count towards the available memory of your container instance:

- running the application executable (as the executable must be loaded to memory)
- allocating memory in your application process
- writing files to the filesystem

The size of the deployed container image does not count towards the available memory.

Setting and updating memory limits

By default, the memory allocated to each container instance of a revision is 256MiB.

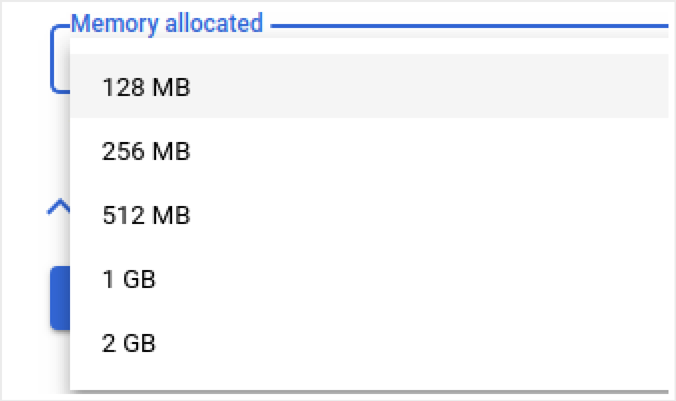
Like any configuration change, setting an memory limits leads to the creation of a new revision. Subsequent revisions will also automatically get this memory limit unless you make explicit updates to change it.

You can set memory limits using the Cloud Console or the gcloud command line when you create a new service (https://cloud.google.com/run/docs/deploying#service) or deploy a new revision (https://cloud.google.com/run/docs/deploying#revision):

CONSOLE COMMAND LINE YAML

1. **GO TO CLOUD RUN** (HTTPS://CONSOLE.CLOUD.GOOGLE.COM/RUN)

2. Click **CREATE SERVICE** if you are setting memory limits on a new service you are deploying to. If you are setting limits on an existing service, then click on the service, then click **DEPLOY NEW REVISION**.
3. Click **SHOW OPTIONAL SETTINGS**.



4. Select the desired memory size from the dropdown list.
5. Click **Create** or **Deploy**.

Maximum amount of memory

The maximum amount of memory you can configure depends on the Cloud Run platform you are deploying to:

- *Cloud Run (fully managed)*: 2 gibibyte (`2Gi`).
- *Cloud Run for Anthos*: limited by the configuration of your GKE cluster.

Optimizing memory

The peak memory requirement for a service can be found using the following: **(Standing Memory) + (Memory per Request) * (Service Concurrency)**

Accordingly,

- If you raise the concurrency of your service, you should also increase the memory limit to account for peak usage.
- If you lower the concurrency of your service, consider reducing the memory limit to save on memory usage costs.

For more guidance on minimizing per request memory usage read Development Tips on Global Variables (https://cloud.google.com/run/docs/tips#using_global_variables).

Last updated January 10, 2020.