Setting a maximum number of container instances

This page describes how to set the maximum number of container instances that can be used for your Cloud Run service. Specifying maximum instances in Cloud Run allows you to limit the scaling of your service in response to incoming requests. Use this setting as a way to control your costs or to limit the number of connections to a backing service, such as to a database.

Note that to specify a maximum number of instances greater than 1000 for Cloud Run (fully managed), you must first request a quota increase (https://cloud.google.com/run/quotas#increase).

For more information on the way Cloud Run autoscales container instances, refer to Instance autoscaling (https://cloud.google.com/run/docs/about-instance-autoscaling).

Setting and updating maximum instances

Like any configuration change, setting a maximum number of container instances leads to the creation of a new revision. Subsequent revisions will also automatically get this maximum number of container instances unless you make explicit updates to change it.

By default, container instances can scale up to 1000 instances. You can change this default 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

1. GO TO CLOUD RUN (HTTPS://CONSOLE.CLOUD.GOOGLE.COM/RUN)
2. Click **CREATE SERVICE** if you are setting a maximum number of container instances on a new service you are deploying to. If you are setting a maximum number of container instances on an existing service, then click on the service, then click **DEPLOY NEW REVISION**.
3. Click **SHOW OPTIONAL SETTINGS**.

4. In the field labelled *Maximum number of instances*, specify the desired maximum number of container instances, using any integer value from `1` to `1000` or more if you requested a quota increase.
5. Click **Create** or **Deploy**.

Last updated December 9, 2019.