Quotas

cloud.google.com/functions/quotas

This document describes the quota limits for Google Cloud Functions.

To increase quotas above the defaults listed here, go to the Cloud Functions Quotas Page, select the quota(s) you want to modify, click **EDIT QUOTAS**, supply your user information if prompted, and enter the new quota limit for each quota you selected. Quotas for Google Cloud Functions encompass 3 areas:

- Resource Limits

  These affect the total amount of resources your functions can consume.

- Time Limits

  These affect how long things can run.

- Rate Limits

  These affect the rate at which you can call the Cloud Functions API and/or the rate at which resources can be used. You can think of rate quotas as "resources over time."

The different types of limits are described in more detail below.

Resource Limits

Resource limits affect the total amount of resources your functions can consume. The regional scope is per project, and each project maintains its own limits.

Quota	Description	Limit	Can be increased	Scope
Number of functions	The total number of functions that can be deployed per region	1,000	No	per region
Max deployment size	The maximum size of a single function deployment	100MB (compressed) for sources. 500MB (uncompressed) for sources plus modules.	No	per function

Max uncompressed HTTP request size	Data sent to HTTP Functions in an HTTP request	10MB	No	per invocation
Max uncompressed HTTP response size	Data sent from HTTP functions in an HTTP response	10MB	No	per invocation
Max event size for background functions	Data sent in events to background functions	10MB	No	per event
Max function memory	Amount of memory each function instance can use	8192MiB	No	per function

Note: If you are triggering a function using Pub/Sub, either via event-driven functions or as the HTTP target of a push subscription, be aware that Pub/Sub messages are base64-encoded. A 10 MB Pub/Sub message - the maximum size supported - is larger than 10 MB once it is encoded, and can thus exceed the Cloud Functions max size limit.

Time Limits

Quota	Description	Limit	Can be increased	Scope
Max function duration	The maximum amount of time a function can run before it's forcibly terminated	540 seconds	No	per invocation

Rate Limits

Quota	Description	Limit	Can be increased	Scope
API calls (READ)	Calls to describe or list functions via the Cloud Functions API	5000 per 100 seconds	Yes	per project
API calls (WRITE)	Calls to deploy or delete functions via the Cloud Functions API	80 per 100 seconds	No [1]	per project
API calls (CALL)	Calls to the "call" API	16 per 100 seconds	No [2]	per project

[1] You cannot increase the WRITE quota. Insufficient quota generally occurs due to one of the following:

- Use of a CI/CD system that deploys many functions concurrently or sequentially at a high rate.

- Use of the Firebase CLI to deploy multiple functions simultaneously.

In each case, you can avoid hitting this quota by changing the rate of deployments. For example, if you are deploying using the Firebase CLI, use the `--only` flag to deploy individual functions.

[2] You cannot increase the CALL quota. Insufficient quota generally occurs if you mistakenly use this API to invoke your functions in production. Please keep in mind that this API is meant for testing via Cloud Console or `gcloud functions call` CLI, and it cannot handle heavy traffic.

Scalability

Cloud Functions invoked by HTTP scale up quickly to handle incoming traffic, while background functions scale more gradually. A function's ability to scale up is dictated by a few factors, including:

- The amount of time it takes for a function's execution to complete (short-running functions can generally scale up to handle more concurrent requests).
- The amount of time it takes for a function to initialize on cold start.
- Rate limits, as described above.
- Your function's error rate.
- Transient factors, such as regional load and data center capacity.

Background functions have additional limits, as explained below. These limits do not apply to HTTP functions.

Additional quotas for background functions

Quota	Description	Limit	Can be increased	Scope
Max concurrent invocations	The maximum concurrent invocations of a single function **Example:** if handling each event takes 100 seconds, the invocation rate will be limited to 30 per second on average	3,000	No	per function
Max invocation rate	The maximum rate of events being handled by a single function **Example:** if handling an event takes 100ms, the invocation rate will be limited to 1000 per second even if only 100 requests, on average, are handled in parallel	1000 per second	No	per function

Max concurrent event data size	The maximum total size of incoming events to concurrent invocations of a single function **Example:** if events are of size 1MB and processing them takes 10 seconds, the average rate will be 1 event per second, because the 11th event will not be processed until processing one of the first 10 events finishes	10MB	No	per function
Max throughput of incoming events	The maximum throughput of incoming events to a single function **Example:** if events are of size 1MB, then the invocation rate can be maximum 10 per second, even if functions finish within 100ms	10MB per second	No	per function

When you reach a quota limit

When a function consumes all of an allocated resource, the resource becomes unavailable until the quota is refreshed or increased. This may mean that your function and all other functions in the same project will not work until then. A function returns an HTTP 500 error code when one of the resources is over quota and the function cannot execute.

To increase quotas above the defaults listed here, go to the Cloud Functions Quotas Page, select the quotas you want to modify, click **EDIT QUOTAS**, supply your user information if prompted, and enter the new quota limit for each quota you selected.