Concurrency

In Cloud Run, each revision (https://cloud.google.com/run/docs/resource-model#revisions) is automatically scaled to the number of container instances needed to handle all incoming requests.

When more container instances are processing requests, more CPU and memory will be used, resulting in higher costs. When new container instances need to be started, requests might take more time to be processed, decreasing the performances of your service.

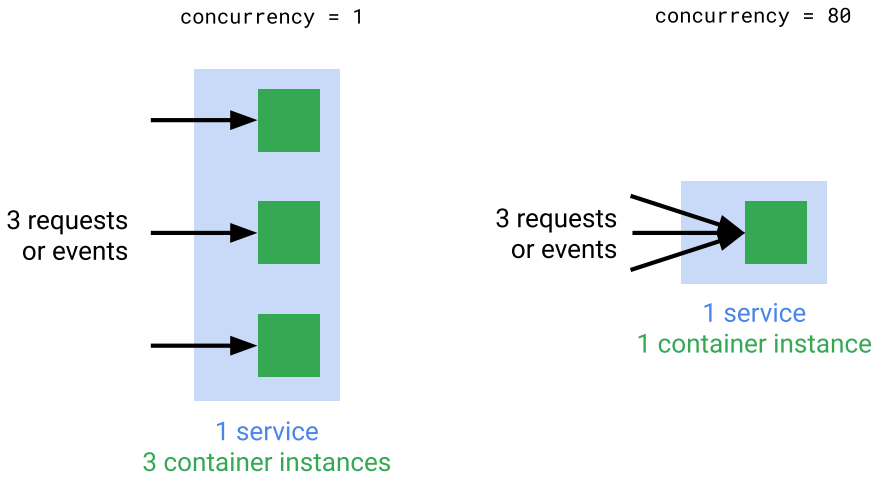
To give you more control, Cloud Run provides a *concurrency* setting that specifies the maximum number of requests that can be processed simultaneously by a given container instance.

Concurrency values

By default Cloud Run container instances can receive many requests at the same time (up to a maximum of 80). Note that in comparison, Functions-as-a-Service (FaaS) solutions like Cloud Functions have a fixed concurrency of 1.

If you want to change this, you can change the concurrency setting (https://cloud.google.com/run/docs/configuring/concurrency) at any time. The specified concurrency value is a maximum and Cloud Run might not send as many requests to a given container instance if the CPU of the instance is already highly utilized.

The following diagram shows how the concurrency setting affects the number of container instances needed to handle incoming concurrent requests:



When to limit concurrency to one request at a time.

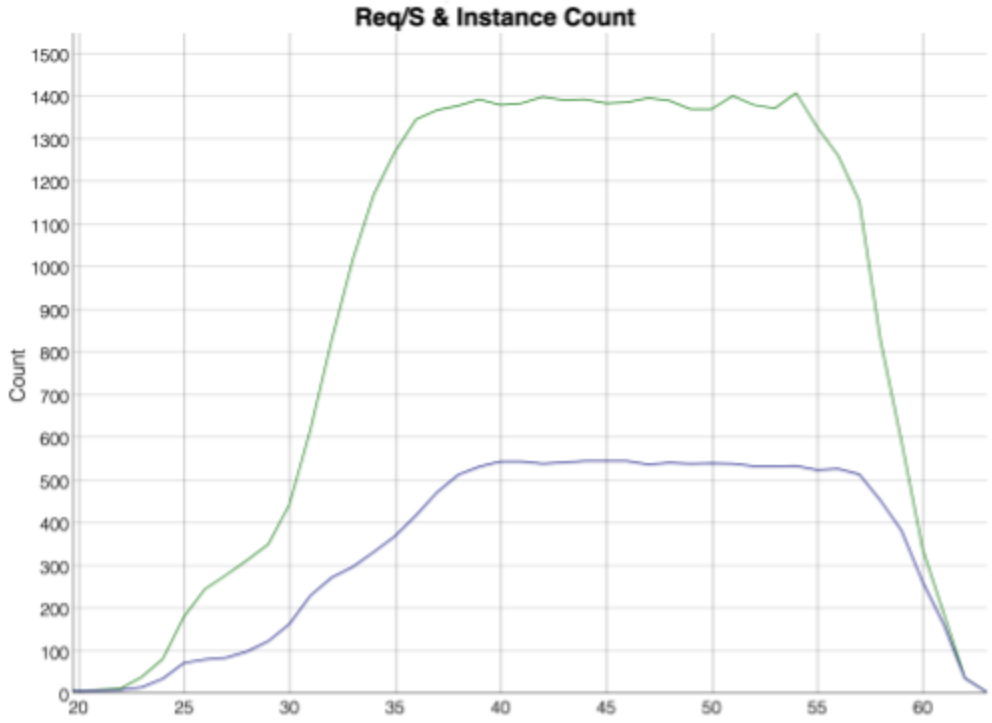
You can limit concurrency so that only one request at a time will be sent to each running container instance. You should consider doing this in cases where:

- Each request uses most of the available CPU or memory.
- Your container image is not designed for handling multiple requests at the same time, for example, if your container relies on global state that two requests cannot share.

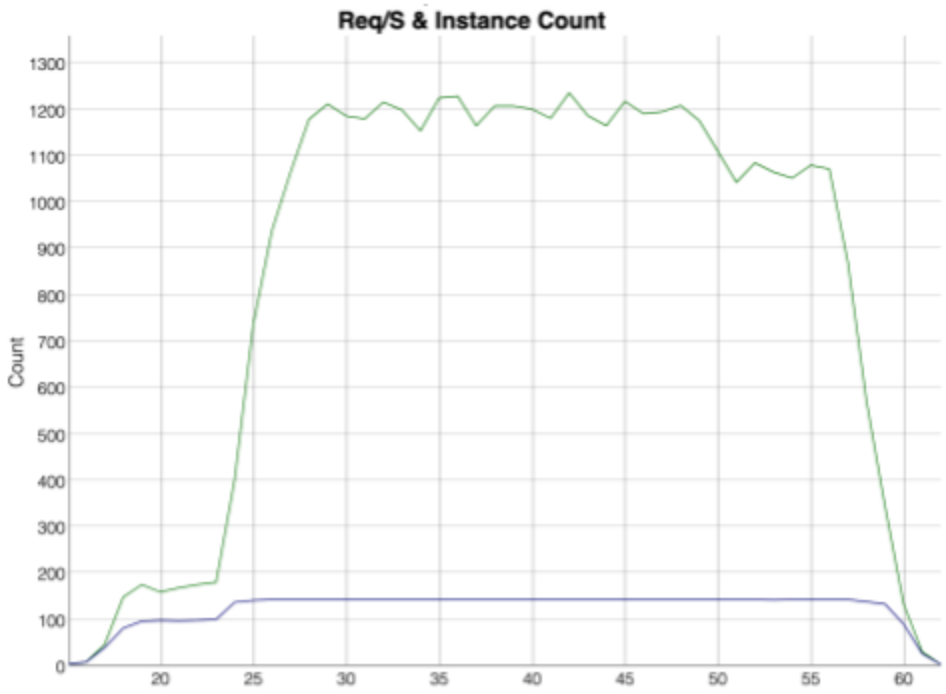
Note that a concurrency of `1` is likely to negatively affect scaling performance, because many container instances will have to start up to handle a spike in incoming requests.

Case study

The following metrics show a use case where 400 clients are making 3 requests per second to a Cloud Run service that is set to a maximum concurrency of 1. The green top line shows the requests over time, the bottom blue line shows the number of container instances started to handle the requests.



The following metrics show 400 clients making 3 requests per second to a Cloud Run service that is set to a maximum concurrency of 80. The green top line shows the requests over time, the bottom blue line shows the number of container instances started to handle the requests. Notice that far fewer instances are needed to handle the same request volume.



What's next

To manage the concurrency of your Cloud Run services, see Setting concurrency (https://cloud.google.com/run/docs/configuring/concurrency).

To optimize your concurrency setting, see development tips for tuning concurrency (https://cloud.google.com/run/docs/tips#tuning-concurrency).

Last updated November 14, 2019.