Using Your Quota Efficiently

- Introduction
- Monitor Your Usage
    - Using the Dashboard
    - Directly on Your Instance
        - Monitoring CPU Usage
        - Monitoring Memory Usage
- Tune Your Application
- Use Smaller Flavors
- Large Datasets
- Use Containers
- Ask for a Quota Increase

Introduction

The Rapid Access Cloud provides a default quota of 8 instances, 8 virtual cores, and 8 gigabytes of memory. You can use this quota to create many different combinations of instances:

- 1 instance with 8 cores and 8gb of memory
- 2 instances with 4 cores and 4gb of memory
- 4 instances with 2 cores and 2gb of memory
- A mixture of the above with different sized instances

These combinations present the ability to use the Rapid Access Cloud for a wide array of scenarios, all within the standard quota limits.

While requesting an increase of your quota is possible, we ask that you first read through this document to determine if changes can be made to your current environment that would allow you to continue using the Rapid Access Cloud with your default quota.

Monitor Your Usage

Using the Dashboard

You can view detailed statistics of your usage on the Rapid Access Cloud Dashboard here.

This dashboard page gives you the ability to see usage aggregated from all instances across all regions as well as per-instance statistics.

If you see the CPU usage fairly low, this means your instance is idle and can safely run more processes.

> A note about memory: The memory statistics can be misleading. Once your instance has consumed all available memory, the statistics will report all memory being in use from that point forward -- the memory usage will never go down. This is a reporting limitation of the Rapid Access Cloud's hypervisor.

Directly on Your Instance

You can also monitor resource usage directly on your instance.

Monitoring CPU Usage

You can use this article to learn about several ways to monitor CPU usage in your instance.

In addition to the tools mentioned in the article, we also recommend `htop`. `htop` is similar to the `top` command but has the ability to show per-core usage. This is useful in determining if all cores in your virtual machine are in use. If you use `htop` and see that only one core is in use, this means the program you are running is "single-threaded". In these cases, you should run more than one instance of your program to make better efficent use of your instance.

To install `htop` do the following on Ubuntu:

```
$ sudo apt-get install htop
```

And the following on CentOS:

```
$ sudo yum install htop
```

Monitoring Memory Usage

Memory usage is a little more complicated than CPU usage. This is because memory usage can come in many different forms: free, cached, buffered, and in-use.

This article describes a few different tools you can use to see details about your instance's memory usage. In addition, we also recommend the site Linux Ate My RAM! as it explains in detail why all of your memory is not in use even though some tools might report it as such.

Tune Your Application

Depending on the application you're using, you might be able to tune it so it consumes a smaller amount of resources. For example, Java-applications are notorious for consuming large amounts of memory, however, this is only if the Java-based application has not been configured with a memory limit.

This article has details on how you can configure your Java application to use a specified amount of memory.

If you're using a web-based application which is based on Apache, you can also configure it to use a lower amount of memory. See this article for more details.

Use Smaller Flavors

The Rapid Access Cloud is meant for development and testing purposes. Most use-cases shouldn't require a lot of resources. For example, developing a web application only requires a few gigabytes of memory, and that includes the full web application stack.

If your use-cases requires multiple instances, you should try using the smallest possible flavor before moving up to a larger flavor. For example, given a scenario where you wanted to develop a web application, you could first try installing everything on a single m1.small instance. If you find your application is unable to effectively run in this flavor, you can do one of two things:

1. Create a larger instance.
2. Create another m1.small instance and move a portion of the web application to that new instance. For example, move the database service to a dedicated instance.



We highly recommend using a DevOps methodology when using the Rapid Access Cloud. This includes using tools such as Terraform and Ansible. These tools will allow you to codify your configurations and provide the ability to automatically rebuild your deployments.

Large Datasets

The Rapid Access Cloud is not designed for large High Performance Computing scenarios, so using datasets which are several gigabytes, or even terabytes, in size can become problematic.

It's possible that the application reading your dataset can efficiently read it without exhausting all available memory. For example, if your dataset is 32 gigabytes in size but your instance only has 8 gigabytes of memory, the application might recognize this and automatically read only a small portion of data at a time.

However, if your application doesn't do this, we recommend breaking your dataset up into smaller pieces, sized appropriately for the amount of memory available to your instance.

If your dataset cannot be broken up into smaller pieces, one other option is to add more swap space to your instance. Instructions on how to do this can be found here.

Use Containers

Containers can be thought of as a form of virtualization. You can install a container management service on your instance which will allow you to run "contained" applications or even full Linux distributions within it. Containers can be very resource efficient since all containers share the same underlying kernel (as opposed to instances which are discrete, isolated virtual machines all having a different operating system kernel).

There are two forms of containers you can use in your instance:

1. Docker
2. LXD

Ask for a Quota Increase

If you've read everything above and still find that your use-case requires access to more resources, please contact us. We are more than happy to provide users with a temporary quota increase to help resolve any difficulties.

We do request you let us know how much of each resource you require and for how long you wish to have the additional resources for.