Fleet pricing

When you run a fleet, charges apply only for the CPU, memory resources, and potentially GPU resources, that are consumed while the fleet runs.

For each fleet you run, you can choose to allow Code Engine to deploy worker nodes to meet the fleet resource requirements, or you can choose to deploy a specific worker node profile that you specify. Review the cost considerations for both options below.

If you let CE automatically provision worker nodes

When you run your fleet, you specify the amount of resources required to run an instance of your code to complete a task, and the maximum number of instances to run concurrently. Code Engine deploys worker nodes of potentially various profiles to meet these resource requirements most efficiently. In this scenario, the cost is based on the worker nodes deployed, but can be approximated using the instance resource requirements you specify, the number of tasks, and the average instance runtime. The formula to approximate the cost of a fleet based on these values is:

```
[ (total cost of vCPU seconds) + (total cost of GB seconds) ] x (# of tasks) x (average runtime of each task in seconds)
```

For example, if an instance of your code requires 2 vCPU and 4 GB, you run 100 tasks, and the average runtime of each task is 0.5 seconds, the formula to approximate the total cost is:

```
[ 2 x (cost of 1 vCPU second) + 4 x (cost of 1 GB second) ] x (100) x (0.5)
```

The total cost to run the fleet is the accumulation of costs for each worker node utilized during the runtime of the fleet. Additionally, unsuccessful instances can have their runtimes increased by retries, which can add to the fleet cost. You can configure retry settings when you create the fleet.

> **Note:** Code Engine does not automatically deploy GPUs for fleets. To deploy GPUs for your fleet, you must choose to deploy a specific worker profile or GPU family.

If you choose a specific worker profile

If you choose to deploy a specific worker node profile or family for your fleet, then only that type of worker node is deployed for your fleet. The total cost to run the fleet is the accumulated cost to run each worker node that is deployed. You can also choose worker profiles with GPUs.

> **Note:** This option is currently available in the CLI only.

Keep in mind that this may not be the most efficient way to run your fleet and might result in a higher cost if worker resources deployed exceed resources required.

Non-GPU worker profiles

If you do not use GPUs, then the cost to run a worker can be approximated with the following formula:

```
[ (total cost of worker VPU seconds) + (total cost of worker GB seconds)  ] x  (average worker runtime in seconds)
```

For example, if you choose a worker profile of 16 vCPU and 64 GB, and your fleet runs for 10 seconds, the formula to approximate the total cost is:

```
[ 16 x (cost of 1 vCPU second) + 64 x (cost of 1 GB second) ] x (10)
```

The total cost to run the fleet is the accumulated cost of running each worker node that deploys. For example, if `2` worker nodes are deployed, you would multiple the above formulas by `2` to approximate the total cost. Keep in mind that the number of worker nodes depends on the required instances resources and the maximum number of concurrent instances.

GPU worker profiles

Each GPU worker incurs an additional charge for GPU seconds. You can approximate the cost of running a GPU worker with the following formula:

```
[ (total cost of  GPU-seconds) + (Total cost of vCPU seconds) + (total cost of 1 GB second) ] x (average worker runtime in seconds)
```

For example, consider the worker node profile `gx3-24x120x2l40s`. You can use the values in the worker node profile to approximate the cost of using the worker. The profile is composed of these parts:

- The first value `gx3` is the worker node category.
- The second value, `24` is the vCPU
- The third value `120` is the GB of memory
- The fourth value `2 L40s`, represents the number of L40s GPU cores.

The approximate cost per second to use a worker node with this profile is:

```
[ 24 x (cost of 1 vCPU second) + 120 x (cost of 1 GB second) + 2 x (cost of 1 L40 GPU-second) ] x (average worker runtime in seconds)
```

The total cost to run the fleet is the accumulated cost of running each worker node that deploys. For example, if `2` GPU workers are deployed, you would multiple the above formulas by `2` to approximate the total cost. Keep in mind that the number of worker nodes depends on the required instance resources and the maximum number of concurrent instances.