Using Ray with GPUs

GPUs are critical for many machine learning applications. Ray enables remote functions and actors to specify their GPU requirements in the `ray.remote` decorator.

Starting Ray with GPUs

In order for remote functions and actors to use GPUs, Ray must know how many GPUs are available. If you are starting Ray on a single machine, you can specify the number of GPUs as follows.

```
ray.init(num_gpus=4)
```

If you don't pass in the `num_gpus` argument, Ray will assume that there are 0 GPUs on the machine.

If you are starting Ray with the `ray start` command, you can indicate the number of GPUs on the machine with the `--num-gpus` argument.

```
ray start --head --num-gpus=4
```

Note: There is nothing preventing you from passing in a larger value of `num_gpus` than the true number of GPUs on the machine. In this case, Ray will act as if the machine has the number of GPUs you specified for the purposes of scheduling tasks that require GPUs. Trouble will only occur if those tasks attempt to actually use GPUs that don't exist.

Using Remote Functions with GPUs

If a remote function requires GPUs, indicate the number of required GPUs in the remote decorator.

```
@ray.remote(num_gpus=1)
def gpu_method():
    return "This function is allowed to use GPUs {}.".format(ray.get_gpu_ids())
```

Inside of the remote function, a call to `ray.get_gpu_ids()` will return a list of integers indicating which GPUs the remote function is allowed to use.

Note: The function `gpu_method` defined above doesn't actually use any GPUs. Ray will schedule it on a machine which has at least one GPU, and will reserve one GPU for it while it is being executed, however it is up to the function to actually make use of the GPU. This is typically done through an external library like TensorFlow. Here is an example that actually uses GPUs. Note that for this example to work, you will need to install the GPU version of TensorFlow.

```
import os
import tensorflow as tf

@ray.remote(num_gpus=1)
def gpu_method():
    os.environ["CUDA_VISIBLE_DEVICES"] = ",".join(map(str, ray.get_gpu_ids()))
    # Create a TensorFlow session. TensorFlow will restrict itself to use the
    # GPUs specified by the CUDA_VISIBLE_DEVICES environment variable.
    tf.Session()
```

Note: It is certainly possible for the person implementing `gpu_method` to ignore `ray.get_gpu_ids` and to use all of the GPUs on the machine. Ray does not prevent this from happening, and this can lead to too many workers using the same GPU at the same time. For example, if the `CUDA_VISIBLE_DEVICES` environment variable is not set, then TensorFlow will attempt to use all of the GPUs on the machine.

Using Actors with GPUs

When defining an actor that uses GPUs, indicate the number of GPUs an actor instance requires in the `ray.remote` decorator.

```
@ray.remote(num_gpus=1)
class GPUActor(object):
    def __init__(self):
        return "This actor is allowed to use GPUs {}.".format(ray.get_gpu_ids())
```

When the actor is created, GPUs will be reserved for that actor for the lifetime of the actor.

Note that Ray must have been started with at least as many GPUs as the number of GPUs you pass into the `ray.remote` decorator. Otherwise, if you pass in a number greater than what was passed into `ray.init`, an exception will be thrown when instantiating the actor.

The following is an example of how to use GPUs in an actor through TensorFlow.

```
@ray.remote(num_gpus=1)
class GPUActor(object):
    def __init__(self):
        self.gpu_ids = ray.get_gpu_ids()
        os.environ["CUDA_VISIBLE_DEVICES"] = ",".join(map(str, self.gpu_ids))
        # The call to tf.Session() will restrict TensorFlow to use the GPUs
        # specified in the CUDA_VISIBLE_DEVICES environment variable.
        self.sess = tf.Session()
```

Troubleshooting

Note: Currently, when a worker executes a task that uses a GPU, the task may allocate memory on the GPU and may not release it when the task finishes executing. This can lead to problems. See this issue.