H200 Partition

The H200 partition is comprised of 6 nodes with 54 H200 GPUs. Access to this partition is limited and is only available by direct request from a faculty PI.

General use of an H200 GPU

We have enabled MIG on some of the GPUs within the H200ea partition.

Slurm settings:

```
Account: h200ea
Partition: h200ea
--gres=gpu:h200
```

Fractionalized GPUs

Two GPUs on dcc-h200-gpu-05 (0 and 1) have been fractionalized (one into 7 parts and one into 2 parts). Fractional GPUs can be requested with:

```
--gres=gpu: h200_1g.18gb:1
--gres= gpu:h200_4g.71gb:1
--gres= gpu:h200_3g.71gb:1
```

Submission

All h200ea jobs will need to specify the "gres" with the "h200" GPU type, e.g.

```
#SBATCH --gres=gpu:h200:1
```

This is to accommodate the MIG settings on dcc-h200-gpu-05.

Due to the increased usage of the h200ea partition, we ask that jobs that will require more than 72 hours to complete be limited to a single H200 GPU (7-day partition limit). If the jobs will require less than 72 hours, 2-GPU jobs can be submitted with the line

```
#SBATCH --time=72:00:00
```

added to the job script. There is currently a per user max GPU limit of 12 H200s allocated at one time.