SKYWAVE™

A STERLING SOLUTION

Sterling SkyWave GPU Partition Plugin

Version: 0.1
Date: 07/22/24
https://sterling.com/skywave/

Overview

This plugin provides a blueprint to configure the ARC-OTA gNB GPU in a Supermicro GH200 server into multiple MIG partitions. This allows the end researcher to experiment with multiple simultaneous GPU accelerated workloads using a single server. de

MIG Partitions

The included configuration creates two MIG partitions.

Partition	SMs	Memory	Use Case
0	4	48 GB	Aerial cuBB Layer1
1	3	48 GB	Other CUDA Applications

Modifying the MIG Configuration

This release includes a configmap (skywave-service-management/files/custom-mig-parted-config.yaml) that defines a custom profile (mixed-3g4g.48gb). If the default configuration listed above does not meet the end researcher's needs, the configmap can be edited to create additional profiles. Customizing the configmap should edited prior to following the installation process below.

Installation Process

1. Create the MIG Partition configmap.

```
kubectl create configmap custom-mig-parted-config --from-file=config.yaml=/home/aerial/skywave-service-management/files/custom-mig-parted-config.yaml -n gpu-operator
```

2. Edit the cluster policy to set the new configmap as the active configuration.

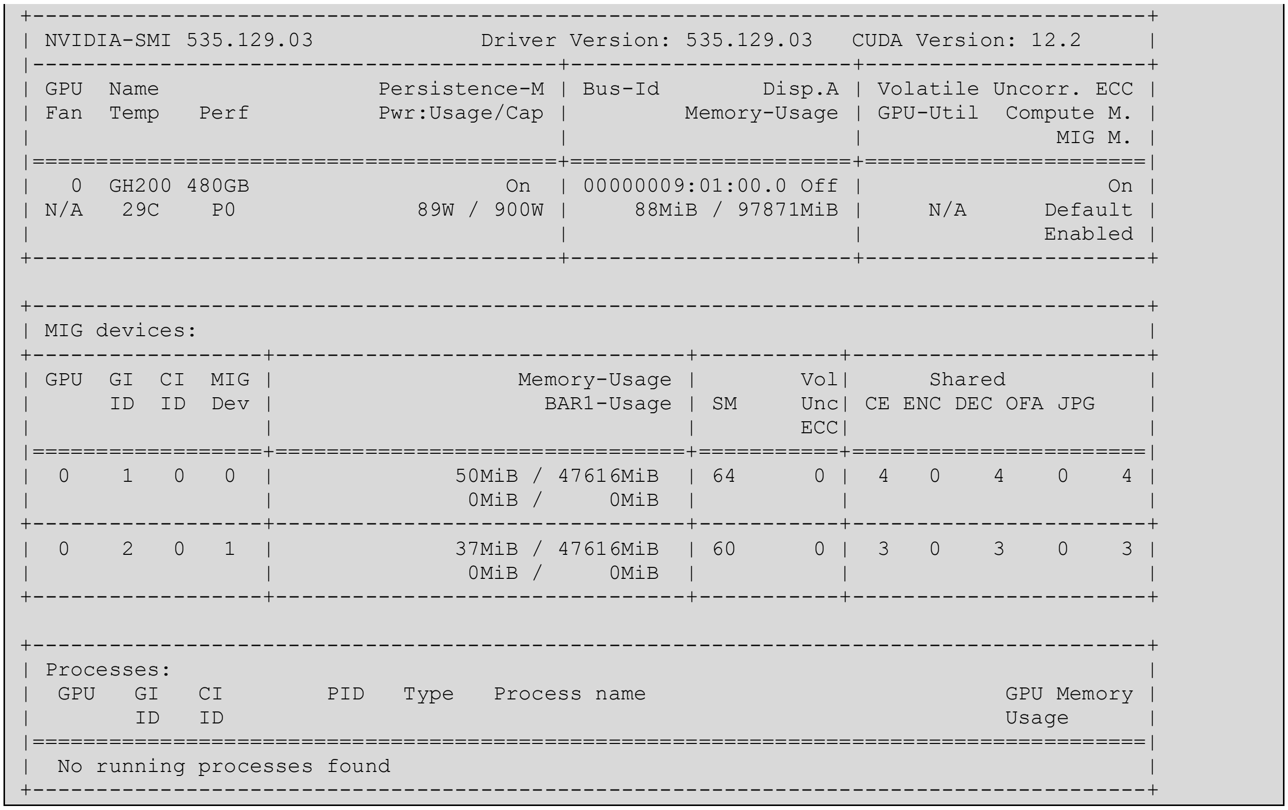
```
kubectl patch clusterpolicy -n gpu-operator cluster-policy --type json -p='[{"op": "replace", "path": "/spec/migManager/config/name", "value": "custom-mig-parted-config"}]'
```

3. Label the node to create the MIG partitions.

```
kubectl label nodes `hostname` nvidia.com/mig.config=mixed-3g4g.48gb –overwrite
```

4. Verify the MIG configuration.

```
kubectl exec -n gpu-operator `kubectl get pods -n gpu-operator | grep nvidia-driver-daemonset | cut -d' ' -f 1` -- bash -c "/usr/bin/nvidia-smi"
```



```
+---------------------------------------------------------------------------------------+
| NVIDIA-SMI 535.129.03             Driver Version: 535.129.03   CUDA Version: 12.2     |
|-----------------------------------------+----------------------+----------------------+
| GPU  Name                 Persistence-M | Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp   Perf          Pwr:Usage/Cap |         Memory-Usage | GPU-Util  Compute M. |
|                                         |                      |               MIG M. |
|=========================================+======================+======================|
|   0  GH200 480GB                    On  | 00000009:01:00.0 Off |                   On |
| N/A   29C    P0              89W / 900W |     88MiB / 97871MiB |     N/A      Default |
|                                         |                      |              Enabled |
+-----------------------------------------+----------------------+----------------------+

+---------------------------------------------------------------------------------------+
| MIG devices:                                                                          |
+------------------+--------------------------------+-----------+-----------------------+
| GPU  GI  CI  MIG |                   Memory-Usage |        Vol|      Shared           |
|      ID  ID  Dev |                     BAR1-Usage | SM     Unc| CE ENC DEC OFA JPG    |
|                  |                                |        ECC|                       |
|==================+================================+===========+=======================|
|  0    1   0   0  |              50MiB / 47616MiB  | 64      0 |  4   0    4    0    4 |
|                  |               0MiB /     0MiB  |           |                       |
+------------------+--------------------------------+-----------+-----------------------+
|  0    2   0   1  |              37MiB / 47616MiB  | 60      0 |  3   0    3    0    3 |
|                  |               0MiB /     0MiB  |           |                       |
+------------------+--------------------------------+-----------+-----------------------+

+---------------------------------------------------------------------------------------+
| Processes:                                                                            |
|  GPU   GI   CI        PID   Type   Process name                            GPU Memory |
|        ID   ID                                                             Usage      |
|=======================================================================================|
|  No running processes found                                                           |
+---------------------------------------------------------------------------------------+
```