GPU Virtualization with MIG

Multi-Tenant Isolation for AI Inference Workloads

Yash Panchal

SDET III @ Percona

Overview

1. GPU sharing Methods
2. MiG Overview
3. MiG Slice Hierarchy, Overhead
4. MiG Partition Creation
5. Partition Combinations (**Wasted Potential cases**)
6. Execution on MiG Instances
7. Workload Performance for Video Generation Task
8. Monitoring MiG Instances
9. Conclusion (Workload suggestions)

GPU Sharing

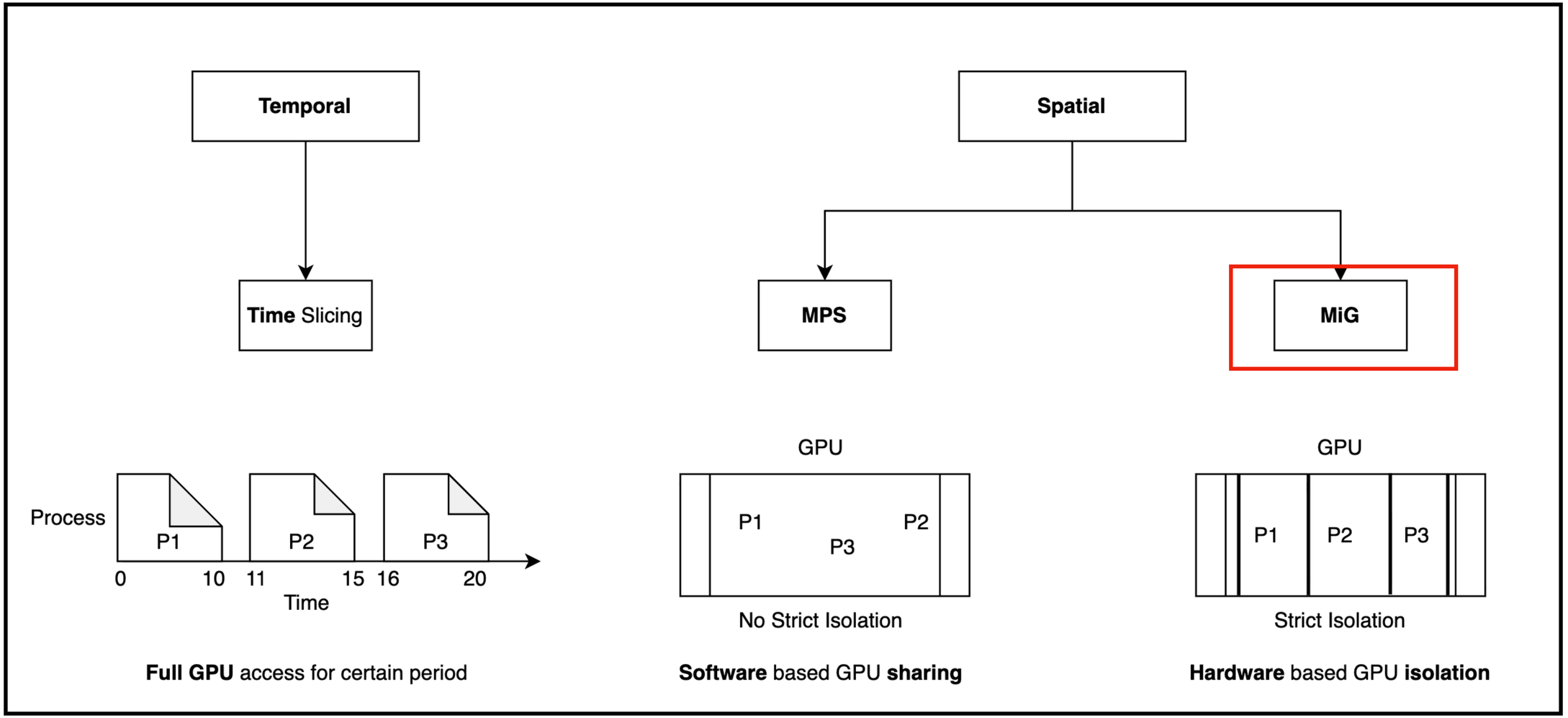

Figure A: GPU Sharing diagram

MiG Overview

Figure C: GPU instances

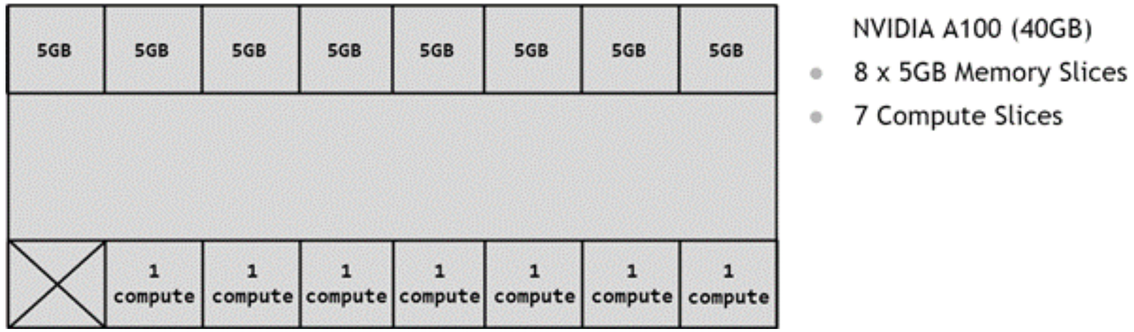

Figure D: GPU slices

- GPU Instance = GPU slices + GPU engines (CE, DEC, JPEG, ENC, OFA)
- GPU slice = GPU Memory slice + GPU Compute slice
- 1 GPU memory slice ~ 1/8 Total GPU Memory
- 1 GPU compute slice ~ 1/7 Total number of compute (SMs)

MiG Slice hierarchy

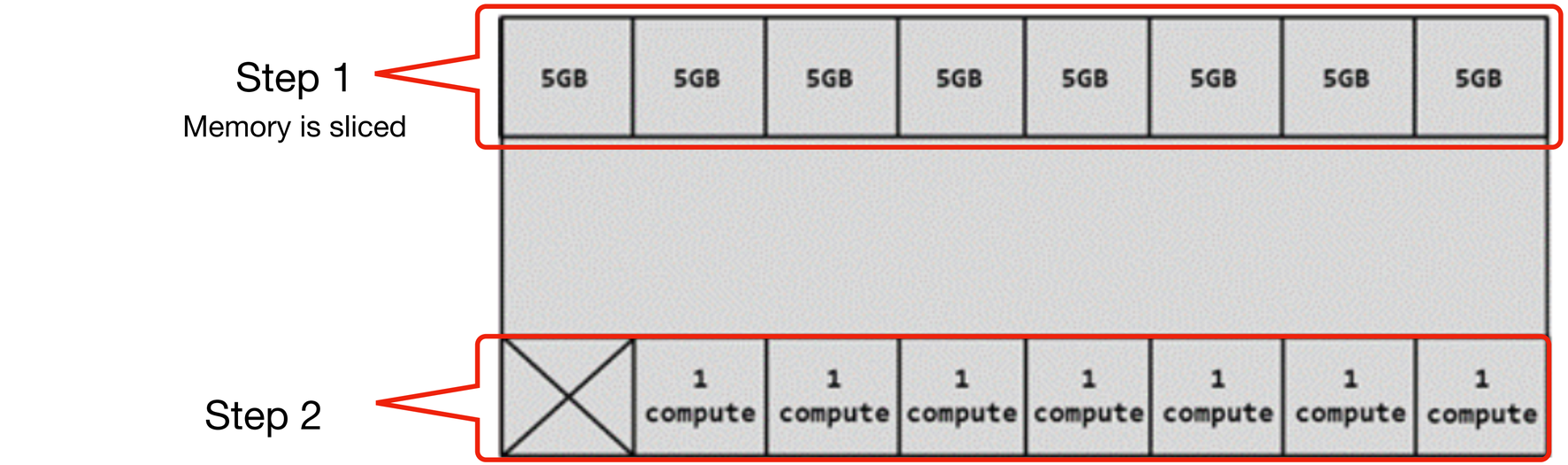

Figure E: 2 Level Slicing A100

- MiG instance creation has has two level hierarchy
- You cannot create compute slice first and then assign memory slice to it.

MiG Partitioning Combinations

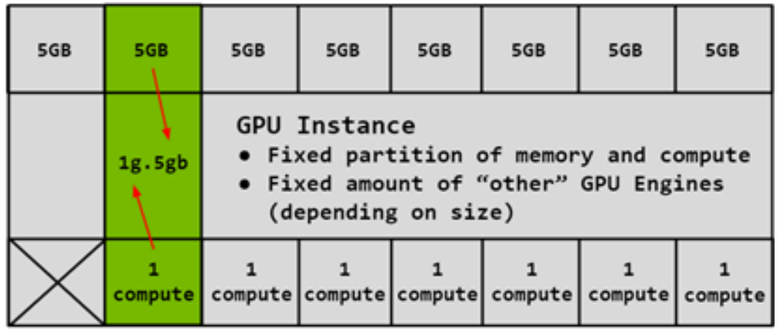

Figure G: Smallest Instance

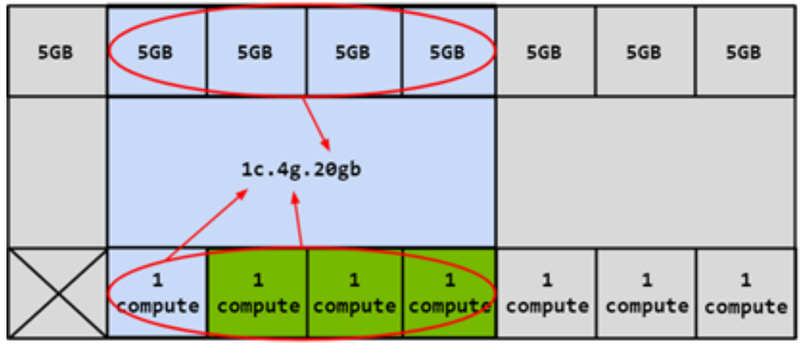

Figure H: isolated compute with shared Memory

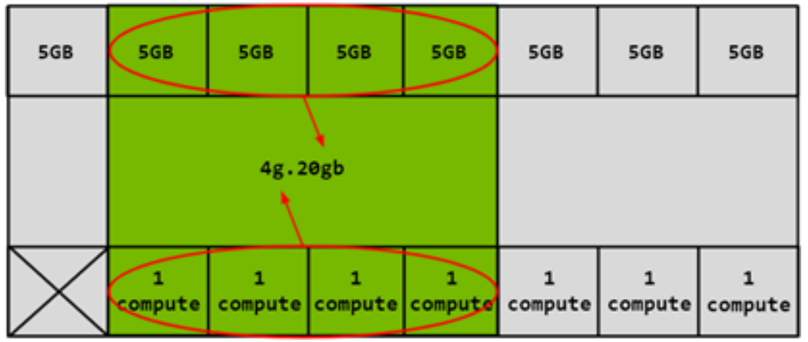

Figure I: Multiple compute with Large Memory

Single isolated compute instance

- 1 isolated 1g.5gb instance
- Full isolation of compute and memory
- Size might be an issue

Multiple isolated compute instances

- Single 4g.20gb split into 4 isolated 1c.4g.20gb instances
- The 20gb memory is shared by 4 instances
- Memory issues (potential OOM)

Single isolated large instance

- 1 isolated 4g.20gb instance
- The 20gb memory is shared by single instance
- Potential of unused compute (Idle compute)

MIG Overhead

```
root@h100-server1:~# nvidia-smi mig -i 0 -lgip
+-----------------------------------------------------------------------------+
| GPU instance profiles:                                                      |
| GPU   Name               ID    Instances   Memory     P2P    SM    DEC   ENC  |
|                               Free/Total   GiB              CE    JPEG  OFA  |
|=============================================================================|
|   0  MIG 1g.10gb         19     7/7         9.75       No     16     1     0   |
|                                                                  1     1     0   |
+-----------------------------------------------------------------------------+
|   0  MIG 1g.10gb+me      20     1/1         9.75       No     16     1     0   |
|                                                                  1     1     1   |
+-----------------------------------------------------------------------------+
|   0  MIG 1g.20gb         15     4/4         19.62      No     26     1     0   |
|                                                                  1     1     0   |
+-----------------------------------------------------------------------------+
|   0  MIG 2g.20gb         14     3/3         19.62      No     32     2     0   |
|                                                                  2     2     0   |
+-----------------------------------------------------------------------------+
|   0  MIG 3g.40gb          9     2/2         39.50      No     60     3     0   |
|                                                                  3     3     0   |
+-----------------------------------------------------------------------------+
|   0  MIG 4g.40gb          5     1/1         39.50      No     64     4     0   |
|                                                                  4     4     0   |
+-----------------------------------------------------------------------------+
|   0  MIG 7g.80gb          0     1/1         79.25      No     132    7     0   |
|                                                                  8     7     1   |
+-----------------------------------------------------------------------------+
```

Figure J: H100 MiG available profiles

- No Free Lunch: There will always be some overhead.
- You will compromise for utilizing MiG feature
- SM CE and Memory GiB are not exactly partitioned
- 1g = 16 SMs (base)
- 3g = 60 SMs (3 x base + few additional SMs)
- 7g = 132 SMs (all SMs)

MiG Partition Creation

```
root@h100-server1:~# nvidia-smi -L
GPU 0: NVIDIA H100 80GB HBM3 (UUID: GPU-f0adccc6-9d17-0af6-ba0b-82402cce84b5)
```

Figure K : List available GPU devices (Only one device)

```
root@h100-server1:~# nvidia-smi mig -i 0 -lgip
+-----------------------------------------------------------------------------+
| GPU instance profiles:                                                      |
| GPU   Name               ID    Instances   Memory     P2P    SM    DEC   ENC  |
|                               Free/Total   GiB              CE    JPEG  OFA  |
|=============================================================================|
|   0  MIG 1g.10gb         19     0/7         9.75       No     16     1     0  |
|                                                                1     1     0  |
+-----------------------------------------------------------------------------+
|   0  MIG 1g.10gb+me      20     0/1         9.75       No     16     1     0  |
|                                                                1     1     1  |
+-----------------------------------------------------------------------------+
|   0  MIG 1g.20gb         15     0/4         19.62      No     26     1     0  |
|                                                                1     1     0  |
+-----------------------------------------------------------------------------+
|   0  MIG 2g.20gb         14     0/3         19.62      No     32     2     0  |
|                                                                2     2     0  |
+-----------------------------------------------------------------------------+
|   0  MIG 3g.40gb          9     0/2         39.50      No     60     3     0  |
|                                                                3     3     0  |
+-----------------------------------------------------------------------------+
|   0  MIG 4g.40gb          5     0/1         39.50      No     64     4     0  |
|                                                                4     4     0  |
+-----------------------------------------------------------------------------+
|   0  MIG 7g.80gb          0     0/1         79.25      No     132    7     0  |
|                                                                8     7     1  |
+-----------------------------------------------------------------------------+
```

Figure L : List available MIG gpu instance profiles on a H100 GPU

Steps to create MIG Profiles

1. nvidia-smi -i 0 -mig 1 (Enable MiG Mode for GPU 0)
2. nvidia-smi mig -i 0 -cgi 9,9 (GPU instance creation)
3. nvidia-smi mig -i 0 -cci (Compute instance assignment)

nvidia-smi mig -i 0 -gi GI -cci 0,0 (2 separate compute instance within single gi)

```
root@h100-server1:~# nvidia-smi mig -i 0 -lgi
+-------------------------------------------------------+
| GPU instances:                                        |
| GPU   Name               Profile  Instance   Placement  |
|                            ID       ID       Start:Size |
|=======================================================|
|   0  MIG 3g.40gb           9        1          0:4     |
+-------------------------------------------------------+
|   0  MIG 3g.40gb           9        2          4:4     |
+-------------------------------------------------------+
```

Figure M: List created GPU Instances

```
root@h100-server1:~# nvidia-smi mig -i 0 -lci
+-------------------------------------------------------------------+
| Compute instances:                                                |
| GPU     GPU       Name             Profile   Instance   Placement  |
|       Instance                       ID        ID       Start:Size |
|         ID                                                        |
|===================================================================|
|   0       1       MIG 3g.40gb          2         0          0:4     |
+-------------------------------------------------------------------+
|   0       2       MIG 3g.40gb          2         0          0:4     |
+-------------------------------------------------------------------+
```

Figure N: List created Compute Instances

MiG Partitions post creation

```
root@h100-server:~# nvidia-smi -L
GPU 0: NVIDIA H100 80GB HBM3 (UUID: GPU-f0adccc6-9d17-0af6-ba0b-82402cce84b5)
  MIG 3g.40gb     Device  0: (UUID: MIG-cbd0b0c1-1449-5ed6-876a-dd5b483b65ca)
  MIG 3g.40gb     Device  1: (UUID: MIG-51c30d73-526a-5ca2-8914-900dfd3ceb1e)
```

Figure K : List available GPU devices (only one main GPU with 2 MiG instance)

```
root@h100-server1:~# nvidia-smi mig -i 0 -lgip
+-----------------------------------------------------------------------------+
| GPU instance profiles:                                                      |
| GPU   Name               ID    Instances   Memory     P2P    SM    DEC   ENC  |
|                              Free/Total    GiB              CE    JPEG  OFA  |
|=============================================================================|
|   0  MIG 1g.10gb         19     0/7        9.75       No     16     1     0   |
|                                                               1     1     0   |
+-----------------------------------------------------------------------------+
|   0  MIG 1g.10gb+me      20     0/1        9.75       No     16     1     0   |
|                                                               1     1     1   |
+-----------------------------------------------------------------------------+
|   0  MIG 1g.20gb         15     0/4        19.62      No     26     1     0   |
|                                                               1     1     0   |
+-----------------------------------------------------------------------------+
|   0  MIG 2g.20gb         14     0/3        19.62      No     32     2     0   |
|                                                               2     2     0   |
+-----------------------------------------------------------------------------+
|   0  MIG 3g.40gb          9     0/2        39.50      No     60     3     0   |
|                                                               3     3     0   |
+-----------------------------------------------------------------------------+
|   0  MIG 4g.40gb          5     0/1        39.50      No     64     4     0   |
|                                                               4     4     0   |
+-----------------------------------------------------------------------------+
|   0  MIG 7g.80gb          0     0/1        79.25      No     132    7     0   |
|                                                               8     7     1   |
+-----------------------------------------------------------------------------+
```

Figure L : List available MIG gpu instance profiles on a H100 GPU

```
root@h100-server1:~# nvidia-smi mig -i 0 -lgi
+-------------------------------------------------------+
| GPU instances:                                        |
| GPU   Name               Profile  Instance   Placement  |
|                            ID       ID       Start:Size |
|=======================================================|
|   0  MIG 3g.40gb           9        1          0:4     |
+-------------------------------------------------------+
|   0  MIG 3g.40gb           9        2          4:4     |
+-------------------------------------------------------+
```

Figure M: List created GPU Instances

```
root@h100-server1:~# nvidia-smi mig -i 0 -lci
+-----------------------------------------------------------------------+
| Compute instances:                                                    |
| GPU     GPU       Name             Profile   Instance   Placement  |
|       Instance                       ID        ID       Start:Size |
|         ID                                                          |
|=======================================================================|
|   0       1       MIG 3g.40gb         2         0          0:4     |
+-----------------------------------------------------------------------+
|   0       2       MIG 3g.40gb         2         0          0:4     |
+-----------------------------------------------------------------------+
```

Figure N: List created Compute Instances

MiG Partitioning combinations

- Combinations that waste Compute Slice
- Combinations that waste Memory Slice
- Combinations that use all available Compute and Memory Slices
- Overhead for using MiG

MiG combos that waste Compute !

A: 3g.40gb partition

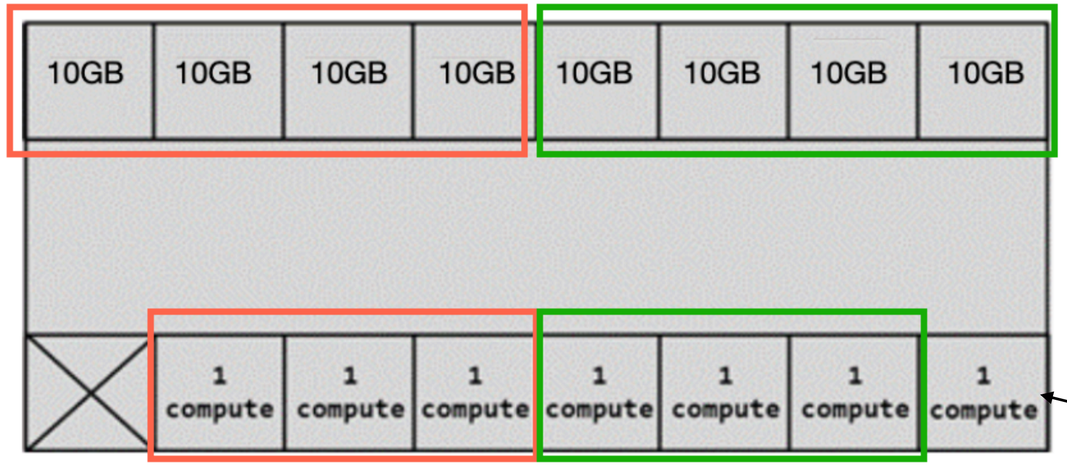

Figure E1: H100 overview

B: 3g.40gb partition

Wasted !

- GPU Memory equal sliced but 1 SM got wasted !
- **Combination of 6 compute slice is prone to this.**
- **3g.40gb sounds great but** 6/7 SMs are being used ,

  meaning 1/7 SM is WASTED (14% of your GPU compute is wasted)

H100 (80GB)

8 X 10GB Memory slices

7 X Compute slices

MiG combos that waste Memory

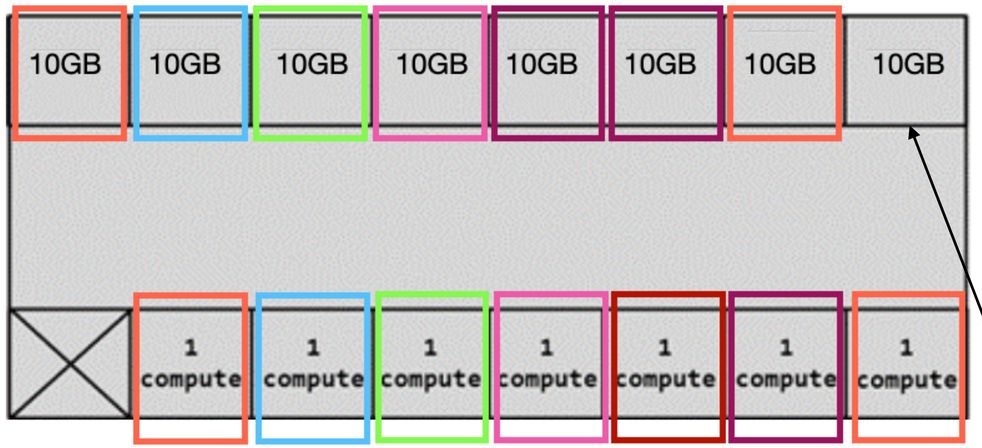

Figure E1: H100 overview

- 7 x 1c.1.g.10gb
- SMs fully used 7/7
- Memory 7/8 = 70 Gb used and 10 GB wasted

Overall Compute Profiles

Config	GPC Slice #0	GPC Slice #1	GPC Slice #2	GPC Slice #3	GPC Slice #4	GPC Slice #5	GPC Slice #6	OFA	NVDEC	NVJPG	P2P	GPU Direct RDMA
1	7							1	7	7	No	Supported MemBW proportional to size of the instance
2	4				3			0	4+3	4+3	No	
3	4				2		1	0	4+2+1	4+2+1	No	
4	4				1	1	1	0	4+1+1+1	4+1+1+1	No	
5	3			3				0	3+3	3+3	No	
6	3			2		1		0	3+2+1	3+2+1	No	
7	3			1	1	1		0	3+1+1+1	3+1+1+1	No	
8	2		2		3			0	2+2+3	2+2+3	No	
9	2		1	1	3			0	2+1+1+3	2+1+1+3	No	
10	1	1	2		3			0	1+1+2+3	1+1+2+3	No	
11	1	1	1	1	3			0	1+1+1+1+3	1+1+1+1+3	No	
12	2		2		2		1	0	2+2+2+2+1	2+2+2+2+1	No	
13	2		1	1	2		1	0	2+1+1+2+1	2+1+1+2+1	No	
14	1	1	2		2		1	0	1+1+2+2+1	1+1+2+2+1	No	
15	2		1	1	1	1	1	0	2+1+1+1+1+1	2+1+1+1+1	No	
16	1	1	2		1	1	1	0	1+1+2+1+1+1	1+1+2+1+1+1	No	
17	1	1	1	1	2		1	0	1+1+1+1+2+1	1+1+1+1+2+1	No	
18	1	1	1	1	1	2		0	1+1+1+1+1+2	1+1+1+1+1+2	No	
19	1	1	1	1	1	1	1	0	1+1+1+1+1+1+1	1+1+1+1+1+1+1	No	

Figure 12: Profiles on H100

Which Combinations leave no slice?

Configuration	instances	Compute Slice	Memory
1 x 7g.80gb	1	7	80 GB
1× 4g.40gb + 1× 3g.40gb	2	4 + 3 = 7	80 GB
1× 3g.40gb + 2× 2g.20gb, **1× 4g.40gb + 1× 2g.20gb + 1× 1g.20gb**	3	3 + 2 + 2 = 7 , 4 + 2 + 1 = 7	80 GB
3× 2g.20gb + 1× 1g.20gb , **1× 3g.40gb + 1× 2g.20gb + 2× 1g.10gb**	4	6 + 1 = 7 , 3 + 2 + 2 = 7	80 GB
1× 3g.40gb + 4× 1g.10gb. , **1× 3g.40gb + 1× 2g.20gb + 2× 1g.10gb**	5	3 + 4 = 7 , 4 + 2 +1 = 7	80 GB
6× 1g.10gb + 1× 1g.20gb	7	6 + 1 = 7	80 GB

Figure O: MiG Slices for an H100 GPU 80GB

Overhead for using MiG

```
root@h100-server1:~# nvidia-smi mig -i 0 -lgip
+-----------------------------------------------------------------------------+
| GPU instance profiles:                                                      |
| GPU   Name               ID    Instances   Memory     P2P    SM    DEC   ENC  |
|                              Free/Total    GiB              CE    JPEG  OFA  |
|=============================================================================|
|   0  MIG 1g.10gb         19     7/7         9.75       No     16     1     0  |
|                                                                 1     1     0  |
+-----------------------------------------------------------------------------+
|   0  MIG 1g.10gb+me      20     1/1         9.75       No     16     1     0  |
|                                                                 1     1     1  |
+-----------------------------------------------------------------------------+
|   0  MIG 1g.20gb         15     4/4         19.62      No     26     1     0  |
|                                                                 1     1     0  |
+-----------------------------------------------------------------------------+
|   0  MIG 2g.20gb         14     3/3         19.62      No     32     2     0  |
|                                                                 2     2     0  |
+-----------------------------------------------------------------------------+
|   0  MIG 3g.40gb          9     2/2         39.50      No     60     3     0  |
|                                                                 3     3     0  |
+-----------------------------------------------------------------------------+
|   0  MIG 4g.40gb          5     1/1         39.50      No     64     4     0  |
|                                                                 4     4     0  |
+-----------------------------------------------------------------------------+
|   0  MIG 7g.80gb          0     1/1         79.25      No     132    7     0  |
|                                                                 8     7     1  |
+-----------------------------------------------------------------------------+
```

- No Free Lunch: There will always be some overhead.
- You will compromise for utilizing MiG feature
- SM CE and Memory GiB are not exactly partitioned as name suggests
- 1g = 16 SMs (base)
- 3g = 60 SMs (3 x base + few additional SMs)
- 7g = 132 SMs (all SMs)

Execution on MiG Instance

1. Create MiG Instance
2. Find the MIG device ID

```
root@h100-server:~# nvidia-smi -L
GPU 0: NVIDIA H100 80GB HBM3 (UUID: GPU-f0adccc6-9d17-0af6-ba0b-82402cce84b5)
  MIG 3g.40gb     Device  0: (UUID: MIG-cbd0b0c1-1449-5ed6-876a-dd5b483b65ca)
  MIG 3g.40gb     Device  1: (UUID: MIG-51c30d73-526a-5ca2-8914-900dfd3ceb1e)
```

Figure K : List available GPU devices (only one main GPU with 2 MiG instance)

3. Assign MIG Device UUID to **CUDA_VISIBLE_DEVICES** env var

Example: export **CUDA_VISIBLE_DEVICES**=MIG-cbd0b0c1-1449-5ed6-876a-dd5b483b65ca (Using Device 0)

4. Run your workload in that environment :)

B200 GPU MiG Profiles

```
root@b200-server:~# nvidia-smi mig -i 0 -lgip
+-----------------------------------------------------------------------------+
| GPU instance profiles:                                                      |
| GPU   Name               ID    Instances   Memory     P2P    SM    DEC   ENC  |
|                               Free/Total   GiB              CE    JPEG  OFA  |
|=============================================================================|
|   0  MIG 1g.23gb         19     7/7         20.50      No     18     1     0  |
|                                                                 2     1     0  |
+-----------------------------------------------------------------------------+
|   0  MIG 1g.23gb+me      20     1/1         20.50      No     18     1     0  |
|                                                                 2     1     1  |
+-----------------------------------------------------------------------------+
|   0  MIG 1g.45gb         15     4/4         44.25      No     30     1     0  |
|                                                                 2     1     0  |
+-----------------------------------------------------------------------------+
|   0  MIG 2g.45gb         14     3/3         44.25      No     36     2     0  |
|                                                                 4     2     0  |
+-----------------------------------------------------------------------------+
|   0  MIG 3g.90gb          9     2/2         89.00      No     70     3     0  |
|                                                                 6     3     0  |
+-----------------------------------------------------------------------------+
|   0  MIG 4g.90gb          5     1/1         89.00      No     72     4     0  |
|                                                                 8     4     0  |
+-----------------------------------------------------------------------------+
|   0  MIG 7g.180gb         0     1/1         178.50     No     148    7     0  |
|                                                                 16    7     1  |
+-----------------------------------------------------------------------------+
root@b200-server:~#
```

Figure U: MiG Partitions options on B200

- B200 has 180GB VRAM
- Wan2.2 parallel video generation
  - TI2V 5B:
    - Max 3 relevant MiG Partitions
  - S2V 14B:
    - Max 2 relevant MiG Partitions

Video Generation

Wan2.2 S2V 14B

- Does not support batching like LLMs
- Large Model 12B Parameter
- prompt + image + audio to generate Video
- 480P
  - ~ 58 GB VRAM (Peak)
  - ~ 55 GB VRAM (Constant)

Wan2.2 TI2V 5B

- Does not support batching like LLMs
- Medium Model 5B Parameter
- prompt to generate Video
- 720P
  - ~ 33 GB VRAM (Peak)
  - ~ 21 GB VRAM (Constant)

Video Generation: Performance Tests

Test	Method	Memory Usage	Result	Time/Video
Baseline Wan2.2 S2V 14B	Full GPU	~ 55 GB Constan ~ 58 GB Peak	Works	4m 5s
Baseline Wan2.2 TI2V 5B	Full GPU	~ 21 GB Constant ~ 33 GB Peak	Works	4m 2s
2 x Wan2.2 s2v 14B	MPS 2 process	OOM	1 OOM 1 Survived	8m 50s
Wan2.2 S2V 14B + Wan2.2 TI2V 5B	MPS 2 process	~ 80 GB	Worked The 14B task finished early	6m 50s for 14B 7m 30s for 5B
3 x Wan2.2 TI2V 5B	MPS 3 process	OOM	1 OOM 2 Survived	5min 45s
2 x Wan2.2 TI2V 5B	**MIG** 2 Instances	~ 80 GB	Works	6m 50s
3 x Wan2.2 TI2V 5B 1m start delay	MPS 2 -> 3 -> 2 -> 1 Processes	~ 79 GB	Works	3m 20s
2 x Wan2.2 s2v 14B	**MIG** 2 Large instance	Around 60 GB IDLE as no other MIG combination possible	Not possible Model too large for H100 MiG Works for B200 4 x 1g.45gb and 3 x 2g.45gb	Use B200 :) 180GB VRAM 3 Videos using MPS 12 mins B200: 4m / video

Figure P: Running two Large Wan2.2 S2V 14B Model processes for 480P Video Generation on H100 having 80GB VRAM using MPS

H100 vs B200 for Wan2.2

Max Videos that can be generated in parallel	TI2V 5B	S2V 14B	Isolation	Allows other work ?
H100 using MPS	2 to 3(staggered)	1	Not Guaranteed	Yes
B200 using MPS	5	3	Not Guaranteed	Yes
H100 using MiG	2	1	Guaranteed	No Leftover GPU Resources go waste
B 200 using MiG	3 (faster) to 4 4 x 1g.45gb and 3 x 2g.45gb	1 to 2 2 x 3g.90gb and 1 x 4g.90gb	Guaranteed	No Leftover GPU Resources go waste

Figure W: Max no of video generation using MPS and MiG for two Wan2.2 models

Possible Failures

- Multiple CI within a Single GI OOM occurs
- Large variation workload a fixed partition

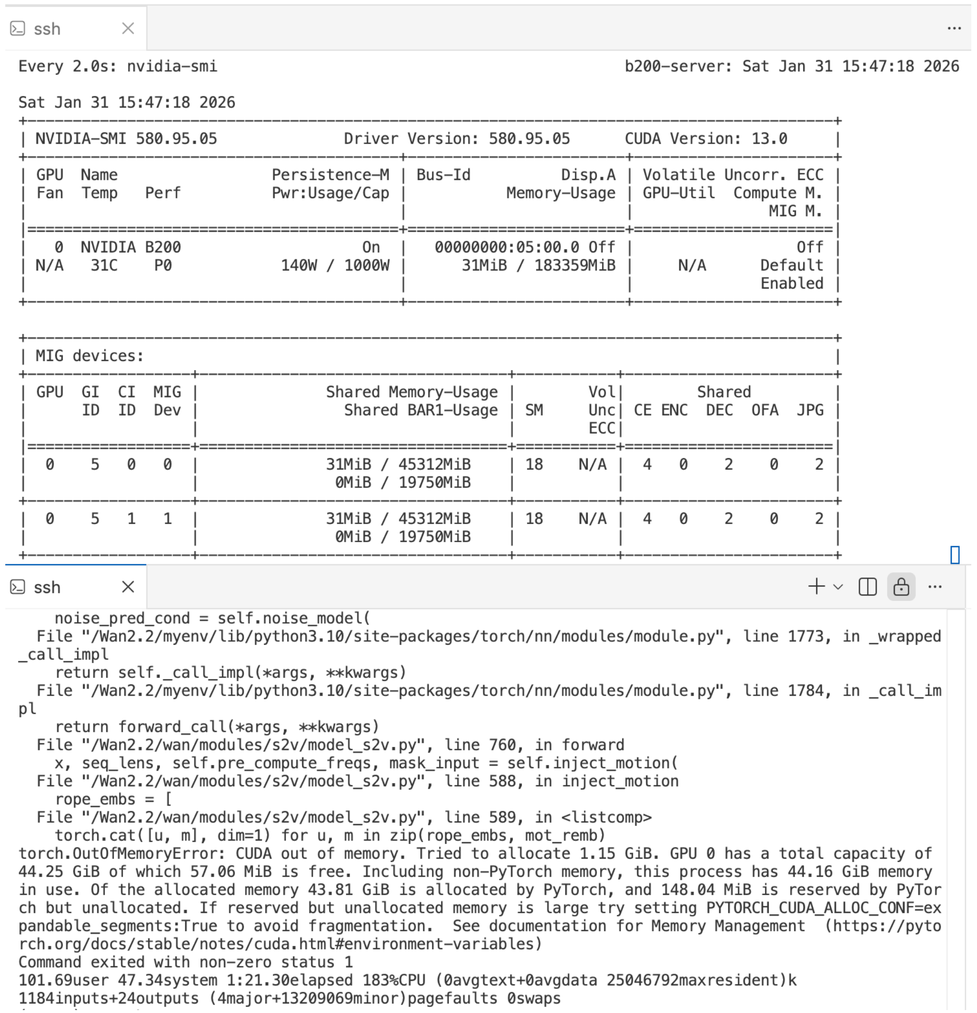

Figure O1: Trying to run high variable workload results in OOM

MiG: Monitoring

1. nvidia-smi (cli based quick view)
2. dcgm-exporter
   - An exporter binary that collects GPU metrics
   - Can be combined with Prometheus and Grafana for visualization
   - https://github.com/NVIDIA/dcgm-exporter.git

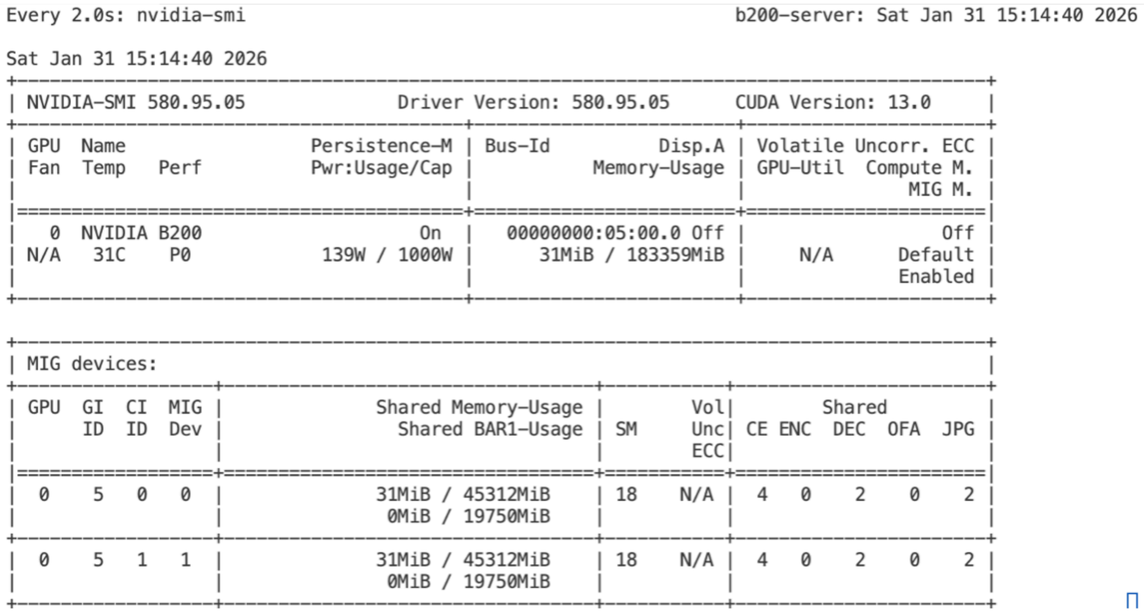

Figure P: MiG Slices two compute 1c.2g.45gb

Combining MiG and MPS

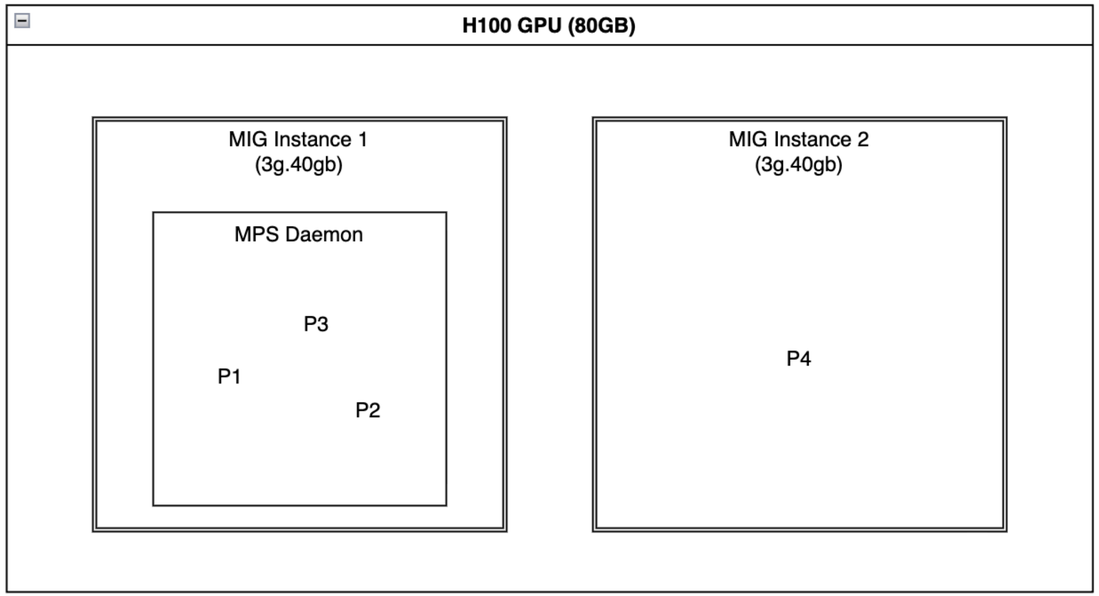

Figure O: H100 with 2 MIG and 1 MIG with MPS

MIG Instance can have MPS within it

MIG instances cannot be created if MPS is already enabled earlier.

(stop the MPS daemon and then MIG instances can be created)

Conclusion

- High Variance workload Provision profiles with buffers to avoid OOM
- For stable workloads MiG works well
- If too high variance use MPS instead of MIG as it will avoid waste of resources
- For large video generation models like wan 2.2 S2V 14B MiG use B200 instead of H100

Thank You