MIG (Multi-Instance GPU) and new Service Units(SU) Charge Model

Configuration • Billing • Usage • Best Practices

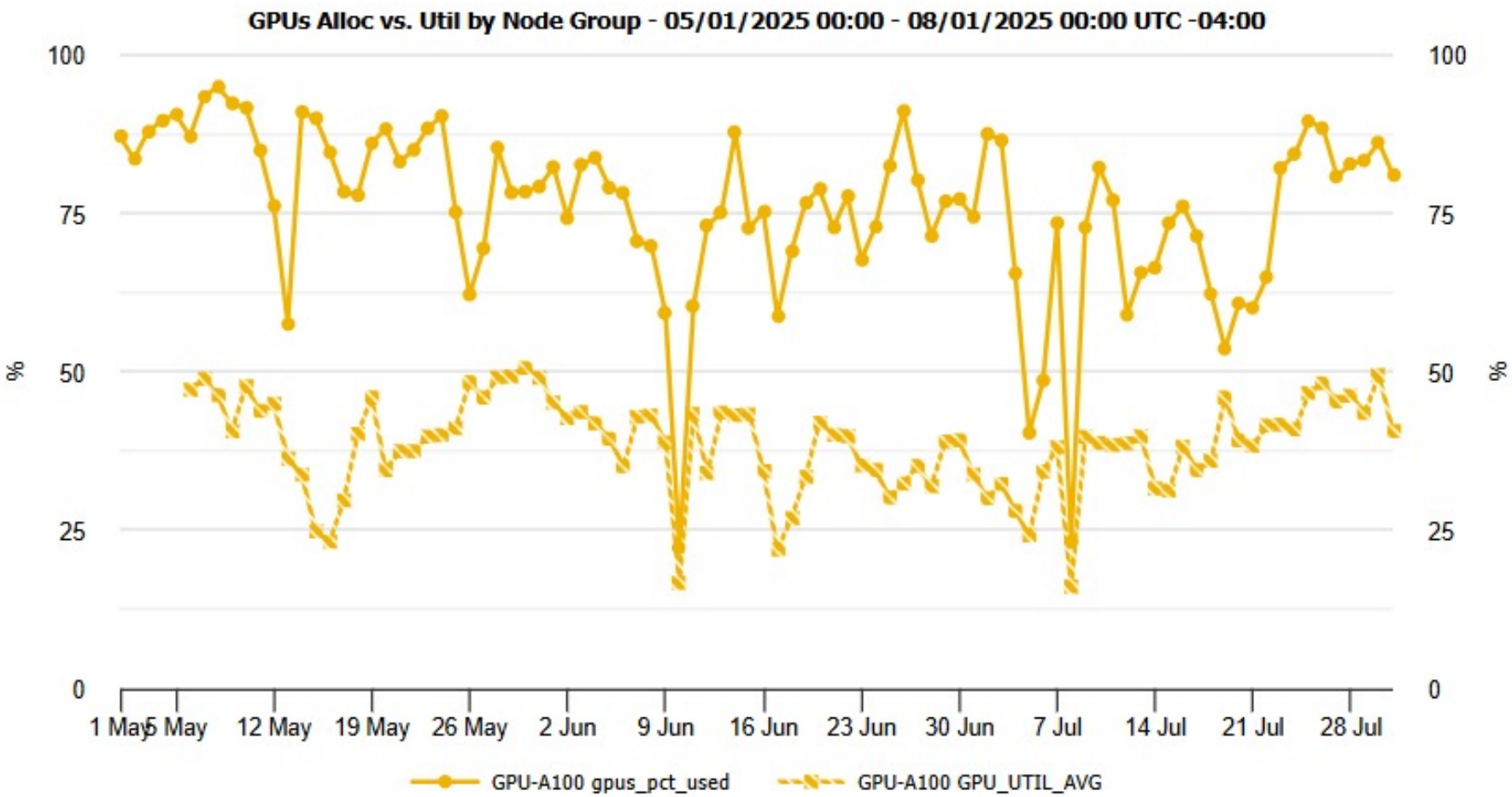
Agenda

- What is MIG?
- Why MIG on Wulver?
- MIG Configuration Example
- Submitting Jobs
- New Billing Model
- Summary & Q&A

What is MIG?

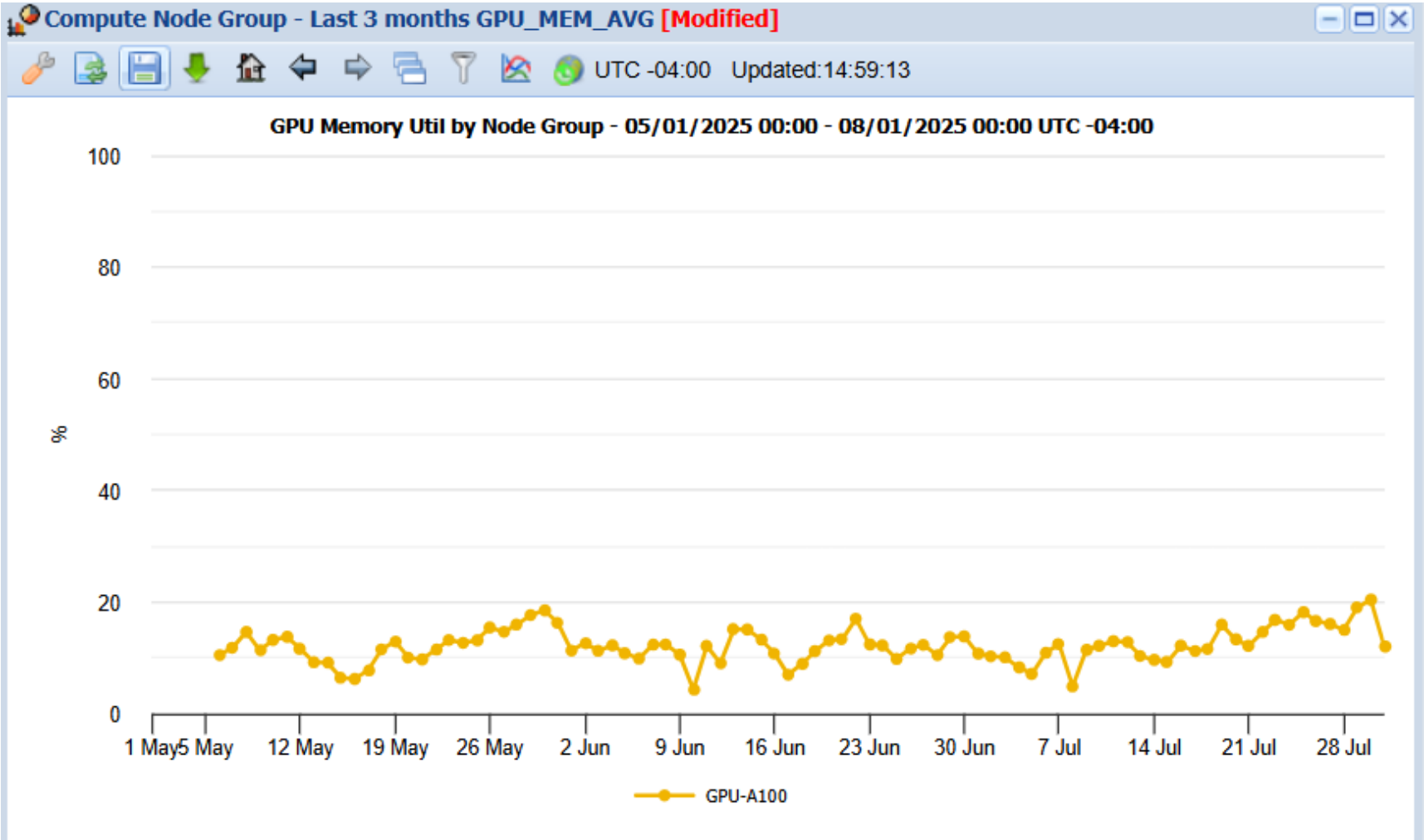
- NVIDIA Multi-Instance GPU partitions an A100 into isolated GPU instances.
- Each instance has dedicated GPU memory, compute cores, cache, and bandwidth.
- Appears as a separate GPU to CUDA and frameworks.
- Up to 7 GPU instances per A100 (depending on profiles).

GPU Allocated vs used



GPUs were reserved ~85–95% of the time, while average compute utilization was ~25–50%

GPU Mem Usage



GPUs were **heavily allocated** but **lightly utilized**

Why MIG on Wulver?

- Reduce queue times: smaller slices fit into schedule easier.
- Improve overall cluster utilization.
- Fair, predictable performance via hardware isolation.
- Lower SU cost for small/medium jobs compared to full A100.

Available Profiles & Naming

- Slurm GRES naming on Wulver
- MIG
  - a100_10g → ~10 GB slice
  - a100_20g → ~20 GB slice
  - a100_40g → ~40 GB slice
- full GPU
  - a100 (--gres=gpu:a100:1) (80 GB)
- Choose the smallest profile that fits your workload.

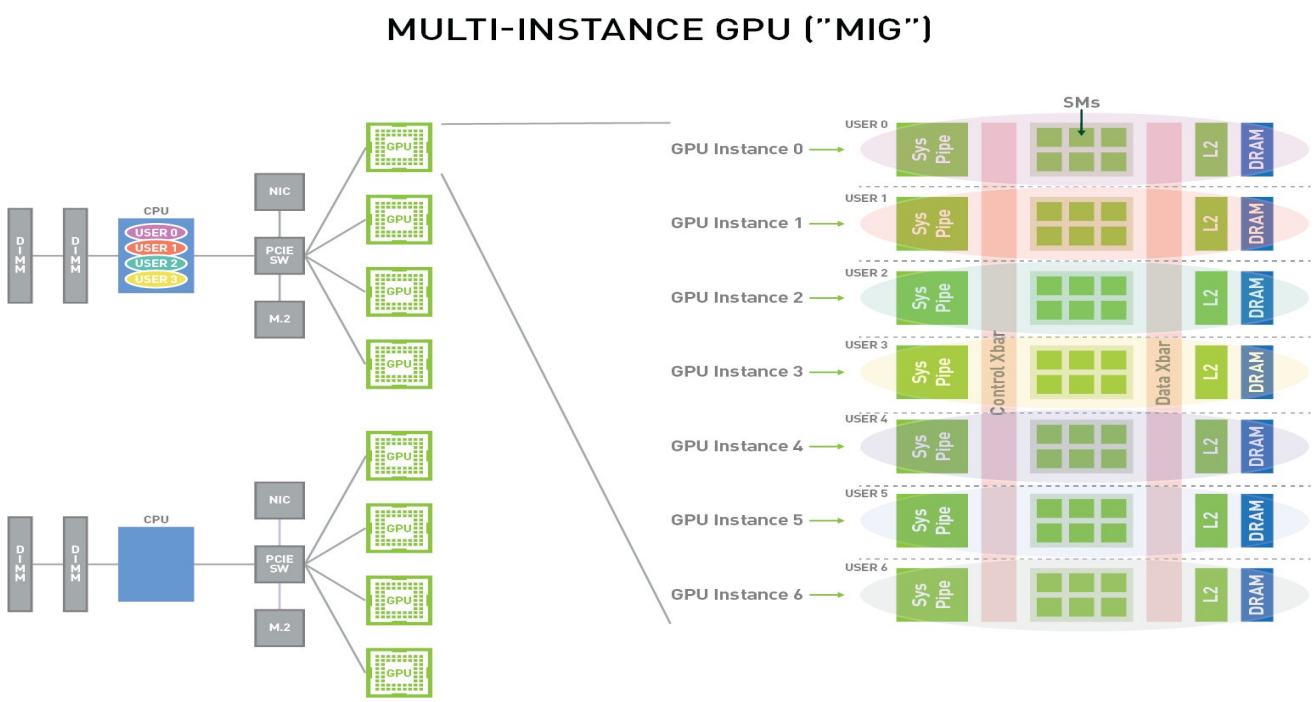
GPU MIG	Slurm Directive
10G MIG	`--gres=gpu:a100_10g:1`
20G MIG	`--gres=gpu:a100_20g:1`
40G MIG	`--gres=gpu:a100_40g:1`

MIG Profiles on A100 MIG

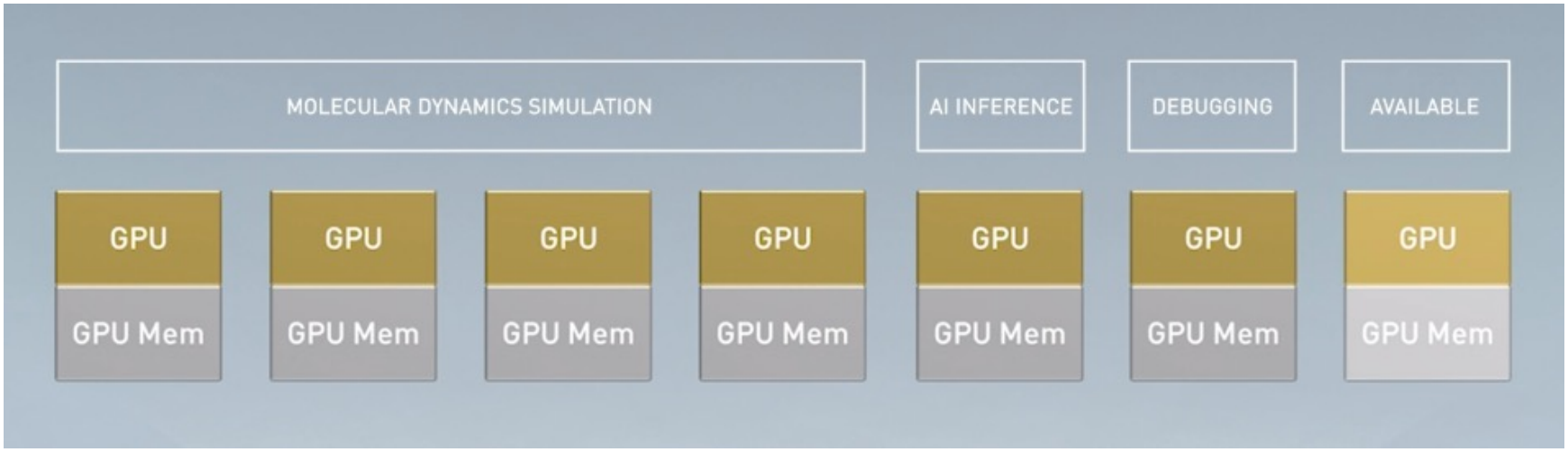
Config	GPC Slice #0	GPC Slice #1	GPC Slice #2	GPC Slice #3	GPC Slice #4	GPC Slice #5	GPC Slice #6	OFA	NVDEC	NVJPG	P2P	GPU Direct RDMA
1	7							1	5	1	No	Supported MemBW proportional to size of the instance
2	4				3			0	2+2	0	No	
3	4				2		1	0	2+1+0	0	No	
4	4				1	1	1	0	2+0+0+0	0	No	
5	3			3				0	2+2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

Reference: https://docs.nvidia.com/datacenter/tesla/mig-user-guide/index.html#a100-mig-profiles

MIG Configuration Example



MIG Configuration Example – Cont.



Source: https://developer.nvidia.com/techdemos/video/disc03

Learn More

- NVIDIA MIG Overview Video: https://developer.nvidia.com/techdemos/video/disc03
- NJIT MIG Documentation: https://hpc.njit.edu/MIG/

Submitting Jobs (Interactive & Batch)

```
$srun --partition=debug_gpu --account=$PI_ucid --qos=debug \
      --gres=gpu:a100_10g:1 --time=00:59:00 --pty bash
```

```
#!/bin/bash -l
#SBATCH --job-name=mig_test
#SBATCH --output=%x.%j.out
#SBATCH --error=%x.%j.err
#SBATCH --partition=debug_gpu
#SBATCH --qos=debug
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=1
#SBATCH --mem-per-cpu=4000M
#SBATCH --gres=gpu:a100_40g:1
#SBATCH --time=02:00:00

module load Miniforge3
conda activate torch-cuda
srun python torch_tensor.py
```

Testing Phase on Wulver

- Access via --partition = debug_gpu (no SU charges) --qos=debug.
- Max walltime: 8 hours.
- MIG slices available for all users.
- Encouraged: test workloads and provide feedback.

```
$ nvidia-smi -L
GPU 0: NVIDIA A100-SXM4-80GB (UUID: GPU-c21a578f-0d0a-8fe2-109a-99d2167fe8f9)
  MIG 3g.40gb     Device  0: (UUID: MIG-c84607d6-2922-525c-a640-b25faead8d90)
  MIG 2g.20gb     Device  1: (UUID: MIG-3fe3432d-1448-5eba-8f14-3bf62ef3f0cf)
  MIG 1g.10gb     Device  2: (UUID: MIG-551d4ef1-778e-5ca6-b86a-bc297fce85c3)
  MIG 1g.10gb     Device  3: (UUID: MIG-34d26a4c-a436-58fb-836d-16443b7e274e)
GPU 1: NVIDIA A100-SXM4-80GB (UUID: GPU-2e5b5c97-11f5-0dc1-abb5-7dd0b2e6d5a8)
  MIG 3g.40gb     Device  0: (UUID: MIG-36f31c31-a1c5-55e7-b02e-78db0ea6d4a1)
  MIG 2g.20gb     Device  1: (UUID: MIG-9175d779-eb2d-5fff-8e45-84c026553abb)
  MIG 1g.10gb     Device  2: (UUID: MIG-e871f386-1ae2-5f80-8793-eab8a04e24e1)
  MIG 1g.10gb     Device  3: (UUID: MIG-7200a2df-6354-5553-997a-668de032174a)
GPU 2: NVIDIA A100-SXM4-80GB (UUID: GPU-83fc584c-8c1c-8bbd-f5a8-6fb4b0b98786)
  MIG 3g.40gb     Device  0: (UUID: MIG-dc4f94f7-063d-5920-ae5a-8e4d075a79e3)
  MIG 2g.20gb     Device  1: (UUID: MIG-31965e78-8e28-56b1-bc71-817c8f7883d6)
  MIG 1g.10gb     Device  2: (UUID: MIG-af845557-7eb3-5163-8614-e808679ed183)
  MIG 1g.10gb     Device  3: (UUID: MIG-4c42fca0-4504-5e7e-94c4-e5bb806bbf38)
GPU 3: NVIDIA A100-SXM4-80GB (UUID: GPU-b72d2661-55fe-a7c1-da5e-d2173cd3658d)
  MIG 3g.40gb     Device  0: (UUID: MIG-6d7f639f-3156-5297-bfa3-02a7c0aaec8e)
  MIG 2g.20gb     Device  1: (UUID: MIG-ee404a6d-6555-5c8d-9afb-c403ed12b621)
  MIG 1g.10gb     Device  2: (UUID: MIG-cfb88930-ae16-511d-b3f0-f6692c6e36a7)
  MIG 1g.10gb     Device  3: (UUID: MIG-07816ffa-6f20-5299-b2d9-40e542061d5f)
```

Tips & Best Practices

- Avoid over-allocating CPU/RAM (affects SU).
- Use smallest MIG profile that fits your workload.
- Check GPU usage with nvidia-smi or PyTorch memory summary.
- Document job resource usage for reproducibility.

Service Unit (SU) Calculation

SU = MAX(#CPUs, Memory(in GB)/4) + 16 × (GPU Memory requested / 80GB)

- CPU/Memory term: Accounts for CPU cores and RAM usage
- GPU term: Scales with GPU memory requested
- Applies to both MIG slices and full A100 GPUs
- HPC Service Units resource page: https://hpc.njit.edu/Running_jobs/service-units/

How SU Charges Are Applied

- CPU and memory usage are always included in SU calculations
- Full A100 GPU (80 GB) → billed as 16 SU/hr
- MIG slices → billed in fractions, proportional to GPU memory share

Examples of SU Charges

- Example 1: 4 CPUs + full A100 GPU (80 GB)

  $SU = MAX(4, RAM/4) + 16 \times (80/80) = 20$ SU/hr

- Example 2: 2 CPUs + MIG slice (20 GB)

  $SU = MAX(2, RAM/4) + 16 \times (20/80) = 6$ SU/hr

Billing Examples (4 CPUs)

Profile	GPU Memory	CPU Alloc	SU/hr
a100_10g	10 GB	4	6
a100_20g	20 GB	4	8
a100_40g	40 GB	4	12
Full A100	80 GB	4	20

Bigmem Partition

- For --partition=bigmem jobs:

  $SU = MAX(1.5 \times \#CPUs, Memory / 16GB)$

- Different scaling due to high-memory nodes
- no GPU nodes, GPU term not included

Summary & Q&A

- MIG enables right-sizing:
  a100_10g / 20g / 40g / full GPU.

- SU charges are now calculated more precisely, incorporating full details for CPU, memory, and GPU requests.

- Debug partition available for exploration.

- Questions?

Upcoming Events

Intro to Wulver: Job Scheduler & Submitting Jobs	October 1 2:30pm - 3:30pm		Online	Abhishek Mukherjee
Intro to Wulver: Focus on Job Efficiency	October 8 2:30pm - 3:30pm		Online	Abhishek Mukherjee
Machine Learning and Big Data	October 14-15 11am - 5pm	TBA	Library Faculty Commons CAB 2032	Pittsburgh Supercomputing Center
HPC User Meeting - Cluster Tools & Monitoring	October 22 2:30pm - 3:30pm	TBA	CTR 235	Hui(Julia) Zhao
Conda for Shared Environments	November 5 2:30pm - 3:30pm	TBA	Online	Hui(Julia) Zhao
HPC User Meeting	December 3 2:30pm - 3:30pm	TBA	CTR 235	Hui(Julia) Zhao

https://hpc.njit.edu/HPC_Events_and_Workshops/