nvidia-smi GPU monitoring

CONCEPTS

Run with no arguments first: the table shows driver and CUDA versions, temperature, power, memory and processes per GPU.

nvidia-smi summary table
↓ `-q` / `--query-gpu`
Detail or CSV pick fields
↓ `-l` / `dmon`
Live monitor loop or stream
↓ `-pm` / `-pl`
Tuning needs root

Anything that changes state (persistence, power, clocks, reset) needs root and usually does not survive a reboot.

INSPECT

`nvidia-smi`
Show GPU summary table

`nvidia-smi -L`
List GPUs with UUIDs

`nvidia-smi -q`
Show all details for every GPU

`nvidia-smi -i` *`<id>`*
Limit output to one GPU

`nvidia-smi -q -d MEMORY,UTILIZATION`
Show only chosen sections

`nvidia-smi -q -x`
Print details as XML

`nvidia-smi --version`
Show tool and driver version

MONITOR

`nvidia-smi -l` *`<sec>`*
Repeat every N seconds

`nvidia-smi -lms` *`<ms>`*
Repeat every N milliseconds

`nvidia-smi dmon`
Stream power, temp, clocks, load

`nvidia-smi dmon -s um`
Choose metric groups to stream

`nvidia-smi pmon`
Show load per process

MONITOR (cont.)

`nvidia-smi pmon -s um`
Per-process utilization and memory

`watch -n 1 nvidia-smi`
Refresh the table in place

SCRIPTED QUERIES

`nvidia-smi --query-gpu=`*`<fields>`* **`--format=csv`**
Query chosen fields as CSV

`nvidia-smi --query-gpu=name,memory.used,memory.total --format=csv`
Memory use per GPU

`nvidia-smi --query-gpu=temperature.gpu --format=csv,noheader,nounits`
Bare numbers for scripts

`nvidia-smi --query-compute-apps=pid,process_name,used_memory --format=csv`
Compute processes and memory

`nvidia-smi --help-query-gpu`
List every queryable field

`nvidia-smi --query-gpu=`*`<fields>`* **`--format=csv -l 1`**
Log CSV every second

PERSISTENCE AND COMPUTE MODE

`sudo nvidia-smi -pm 1`
Enable persistence mode

`sudo nvidia-smi -c` *`<mode>`*
Set compute mode

POWER AND CLOCKS

`sudo nvidia-smi -pl` *`<watts>`*
Set power limit

`nvidia-smi -q -d CLOCK`
Show current and max clocks

`nvidia-smi -q -d PERFORMANCE`
Show throttle reasons

TOPOLOGY AND OTHERS

`nvidia-smi topo -m`
Show GPU interconnect matrix