

NVIDIA GPU Cheatsheet

Blackwell

	H100	H200	GB200 Superchip	NVL72	NVL36
Configuration	Hopper GPU	Hopper GPU	2x Blackwell GPU, 1x Grace CPU	36 Grace CPUs : 72 Blackwell GPUs	18 Grace CPUs : 36 Blackwell GPUs
FP4 Tensor Dense/Sparse	N/A	N/A	20/40 petaflops	720/1440 PFLOPS	360/720 PFLOPS
FP6/FP8 Tensor Dense/Sparse	2/4 petaflops	2/4 petaflops	10/20 petaflops	360/720 PFLOPS	180/360 PFLOPS
INT8 Tensor Dense/Sparse	2/4 petaflops	2/4 petaflops	10/20 petaflops	360/720 PFLOPS	180/360 PFLOPS
FP16/BF16 Tensor Dense/Sparse	1/2 petaflops	1/2 petaflops	5/10 petaflops	180/360 PFLOPS	90/180 PFLOPS
TF32 Tensor Dense/Sparse	0.5/1 petaflops	0.5/1 petaflops	2.5/5 petaflops	90/180 PFLOPS	45/90 PFLOPS
FP32	67 teraflops	67 teraflops	180 teraflops	6480 teraflops	6480 TFLOPS
FP64 Tensor Core	34/67 teraflops	34/67 teraflops	90 teraflops	3240 TFLOPS	1620 TFLOPS
Memory Type	HBM3	HBM3e	HBM3e	HBM3e	HBM3e
Memory	80GB(5x16GB)	141GB(6x24GB)	up to 384GB (2x8x24GB)	up to 13.5TB HBM3e	up to 6.75TB HBM3e
Memory Bandwidth	3.35TB/s	4.8TB/s	16 TB/s	576 TB/s	288 TB/s
NVLink Bandwidth	900GB/s	900GB/s	2x 1.8 TB/s	130TB/s	65TB/s
Power	700W	700W	Up to 2700W	Up to 123.6kW	Up to 67kW

Hopper

		A100 40GB PCIe	A100 80GB PCIe	A100 40GB SXM	A100 80GB SXM	H100 80GB SXM	H100 80GB PCIe	H100 94GB NVL	H200 141GB NVL	H200 141GB SXM
GPU memory	GB	40	80	40	80	80	80	94	141	141
FP64	TFLOPS	9,7	9,7	9,7	9,7	34	26	30	34	34
FP64 Tensor Core	TFLOPS	19,5	19,5	19,5	19,5	67	51	60	67	67
FP32	TFLOPS	19,5	19,5	19,5	19,5	67	51	60	67	67
TF32 Tensor Core	TFLOPS	312	312	312	312	989	756	835	989	989
BFLOAT16 Tensor Core	TFLOPS	624	624	624	624	1979	1513	1671	1979	1979
FP16 Tensor Core	TFLOPS	624	624	624	624	1979	1513	1671	1979	1979
FP8 Tensor Core	TFLOPS					3958	3026	3341	3958	3958
INT8 Tensor Core	TFLOPS	1248	1248	1248	1248	3958	3026	3341	3958	3958
GPU memory bandwidth	TB/s	1,55	1,935	1,55	1,935	3,35	2	3,9	4,8	4,8

		A100 40GB PCIe	A100 80GB PCIe	A100 40GB SXM	A100 80GB SXM	H100 80GB SXM	H100 80GB PCIe	H100 94GB NVL	H200 141GB NVL	H200 141GB SXM
Decoders						7 x NVDEC 7 x JPEG	7 x NVDEC 7 x JPEG	7 x NVDEC 7 x JPEG	7 x NVDEC 7 x JPEG	7 x NVDEC 7 x JPEG
Max thermal design power (TDP)	W	250	300	400	400	700	350	400	700	700
Multi-Instance GPUs		7	7	7	7	7	7	7	7	7
Form factor		PCIe	PCIe	SXM	SXM	SXM	PCIe	PCIe	PCIe	SXM
NVLink	GB/s	600	600	600	600	900	600	600	900	900
PCIe		Gen4	Gen4	Gen4	Gen4	Gen5	Gen5	Gen5	Gen5	Gen5

From:

<https://micronode.ru/> - **micronode.ru**

Permanent link:

<https://micronode.ru/enterprise/nvidia/description/gpu>

Last update:

2025/01/20 13:58

