

GPU Datasheet

datasheet

Brand	Model	GPU Memory	Memory ECC	Memory Bandwidth	Max Power Consumption	Host Interface (BW)	GPU-to-GPU Peer Interconnect (BW)	Slot Width	GPU Height/Length	Auxiliary Cable	Workload
AMD	MI300X OAM	192 GB HBM3	Y	5.3 TB/sec	750W	PCIe Gen5x16 (128 GB/sec)	Infinity Fabric Link (4th Gen-896 GB/sec)	N/A	N/A	N/A	AI / HPC
AMD	MI210	64 GB HBM2e	Y	1638 GB/sec	300W	PCIe Gen4x16 (64 GB/sec)	Infinity Fabric Link bridge8	DW	FHFL	CPU 8 pin	HPC/Machine learning training
Intel	Gaudi3 PCIe	128GB HBM3	Y	3.7 TB/s	600W	PCIe Gen5x16 (128 GB/sec)	3x200G RoCE	DW	FHFL		AI / HPC
Intel	Gaudi3 OAM	128GB HBM3	Y	3.6TB/s	850W	PCIe Gen5x16 (128 GB/sec)	3x200G RoCE	N/A	N/A	N/A	AI / HPC
Intel	Max1550 OAM	128 GB HBM2e	Y	3276.8 GB/sec	600W	PCIe Gen5x16 (128 GB/sec)	Intel XeLink	N/A	N/A	N/A	AI / HPC
Intel	Max1100	48 GB HBM2e	Y	1228.8 GB/sec	300W	PCIe Gen5x16 (128 GB/sec)	XeLink bridge8 (600GB/s)	DW	FHFL	PCIe 16 pin	AI / HPC
Intel	Flex 140	12 GB GDDR6	Y	336 GB/Sec	75W	PCIe Gen4 x8 (32 GB/sec)	XeLink bridge8 (336 GB/sec)	SW	HHHL/FHHL	N/A	Inferencing/Edge
Nvidia	H200 NVL	141GB HBM3e	Y	4.8 TB/s	600W	PCIe Gen 5×16 (128 GB/sec)	NVLink (4th Gen - 900 GB/s)	DW	FHFL	PCIe 16 pin	AI / HPC
Nvidia	H100 NVL	94 GB HBM3	Y	3.9 TB/s	350W- 400W	PCIe Gen 5×16 (128 GB/sec)	NVLink (4th Gen - 600 GB/s)	DW	FHFL	PCIe 16 pin	AI / HPC
Nvidia	B200 DLC	180 GB	Y			PCIe Gen 5×16 (128 GB/sec)	NVLink (4th Gen - 600 GB/s)				
Nvidia	H200 SXM5 (x8) DLC	141GB HBM3e	Y	4.8 TB/s	700W	PCIe Gen 5×16 (128 GB/sec)	NVLink (4th Gen - 900 GB/s)	DW			AI / HPC
Nvidia	H200 SXM5 (x8)	141GB HBM3e	Y	4.8 TB/s	700W	PCIe Gen 5×16 (128 GB/sec)	NVLink (4th Gen - 900 GB/s)	N/A	N/A	N/A	AI / HPC
Nvidia	H100 SXM5 (x8)	80 GB HBM3	Y	3 TB/sec	700W	PCIe Gen 5×16 (128 GB/sec)	NVLink (4th Gen - 900 GB/s)	N/A	N/A	N/A	AI / HPC
Nvidia	H100 SXM5 (x4)	80 GB HBM3	Y	3 TB/sec	700W	PCIe Gen 5×16 (128 GB/sec)	NVLink (4th Gen - 900 GB/s)	N/A	N/A	N/A	AI / HPC
Nvidia	A100 SXM4 (x8)	80 GB HBM2	Y	2039 GB/sec	500W	PCIe Gen 5×16 (128 GB/sec)	NVLink (3rd Gen - 600 GB/s)	N/A	N/A	N/A	HPC/AI/Database Analytics
Nvidia	A100 SXM4 (x4)	80 GB HBM2	Y	2039 GB/sec	500W	PCIe Gen 5×16 (128 GB/sec)	NVLink (3rd Gen - 600 GB/s)	N/A	N/A	N/A	HPC/AI/Database Analytics
Nvidia	A100 SXM4 (x4)	40 GB HBM2	Y	1555 GB/sec	400W	PCIe Gen 5×16 (128 GB/sec)	NVLink (3rd Gen - 600 GB/s)	N/A	N/A	N/A	HPC/AI/Database Analytics
Nvidia	H100	80 GB HBM2e	Y	2000 GB/sec	300-350W	PCIe Gen 5×16 (128 GB/sec)	NVLink8 (4th Gen - 600 GB/s)	DW	FHFL	PCIe 16 pin	HPC/AI/Database Analytics
Nvidia	A100	80 GB HBM2e	Y	1935 GB/sec	300W	PCIe Gen4 x16 (64 GB/sec)	NVLink8 (3rd Gen - 600 GB/s)	DW	FHFL	CPU 8 pin	HPC/AI/Database Analytics
Nvidia	L40S	48 GB GDDR6	Y	864 GB/sec	350W	PCIe Gen4 x16 (64 GB/sec)	N/A	DW	FHFL	PCIe 16 pin	AI/Performance graphics/VDI

Brand	Model	GPU Memory	Memory ECC	Memory Bandwidth	Max Power Consumption	Host Interface (BW)	GPU-to-GPU Peer Interconnect (BW)	Slot Width	GPU Height/Length	Auxiliary Cable	Workload
Nvidia	A30	24 GB HBM2	Y	933 GB/sec	165W	PCIe Gen4 x16 (64 GB/sec)	NVLink8 (3rd Gen - 200 GB/s)	DW	FHFL	CPU 8 pin	mainstream AI
Nvidia	L40	48 GB GDDR6	Y	864 GB/sec	300W	PCIe Gen4 x16 (64 GB/sec)	N/A	DW	FHFL	PCIe 16 pin	Performance graphics/VDI
Nvidia	A40	48 GB GDDR6	Y	696 GB/sec	300W	PCIe Gen4 x16 (64 GB/sec)	NVLink bridge8 (3rd Gen-112.5 GB/s)	DW	FHFL	CPU 8 pin	Performance graphics/VDI
Nvidia	A16	64 GB GDDR6	Y	800 GB/sec	250W	PCIe Gen4 x16 (64GB/sec)	N/A	DW	FHFL	CPU 8 pin	VDI
Nvidia	L4	24 GB GDDR6	Y	300 GB/s	72W	PCIe Gen4 x16 (64GB/sec)	N/A	SW	HHHL	N/A	Inferencing/Edge/VDI
Nvidia	L4	24 GB GDDR6	Y	300 GB/s	72W	PCIe Gen4 x16 (64GB/sec)	N/A	SW	FHHL	N/A	Inferencing/Edge/VDI
Nvidia	A2	16 GB GDDR6	Y	200 GB/sec	60W	PCIe Gen 4x8	N/A	SW	HHHL	N/A	Inferencing/Edge/VDI
Nvidia	A2 (v2)	16 GB GDDR6	Y	200 GB/sec	60W	PCIe Gen4 x8 (32 GB/sec)	N/A	SW	HHHL	N/A	Inferencing/Edge/VDI
Nvidia	A2	16 GB GDDR6	Y	200 GB/sec	60W	PCIe Gen4 x8 (32 GB/sec)	N/A	SW	FHHL	N/A	Inferencing/Edge/VDI
Nvidia	A2 (v2)	16 GB GDDR6	Y	200 GB/sec	60W	PCIe Gen4 x8 (32 GB/sec)	N/A	SW	FHHL	N/A	Inferencing/Edge/VDI
Nvidia	A10	24 GB GDDR6	Y	600 GB/sec	150W	PCIe Gen4 x16 (64GB/sec)	N/A	SW	FHFL	PCIe 8 pin	mainstream graphics/VDI
Nvidia	T4	16 GB GDDR6	Y	300 GB/sec	70W	PCIe Gen3 x16 (32GB/sec)	N/A	SW	HHHL	N/A	Inferencing/Edge/VDI
Nvidia	T4	16 GB GDDR6	Y	300 GB/sec	70W	PCIe Gen3 x16 (32GB/sec)	N/A	SW	FHHL	N/A	Inferencing/Edge/VDI
AMD	MI100	32 GB HBM2	Y	1228 GB/sec	300W	PCIe Gen4 x16	64 GB/sec (PCIe 4.0)	DW	FHFL	PCIe 8 pin	HPC/Machine learning training
Nvidia	M10	32 GB GDDR5	N	332 GB/sec	225W	PCIe Gen3x16	32 GB/sec (PCIe 3.0)	DW	FHFL	PCIe 8 pin	VDI
Nvidia	A100	40 GB HBM2	Y	1555 GB/sec	250W	PCIe Gen4x16/ NVLink bridge8	64 GB/sec5 (PCIe 4.0)	DW	FHFL	CPU 8 pin	HPC/AI/Database Analytics
Nvidia	V100S	32 GB HBM2	Y	1134 GB/sec	250W	PCIe Gen3x16	32 GB/sec (PCIe 3.0)	DW	FHFL	CPU 8 pin	HPC/AI/Database Analytics
Nvidia	V100	32 GB HBM2	Y	900 GB/sec	250W	PCIe Gen3x16	32 GB/sec (PCIe 3.0)	DW	FHFL	CPU 8 pin	HPC/AI/Database Analytics
Nvidia	V100	16 GB HBM2	Y	900 GB/sec	250W	PCIe Gen3x16	32 GB/sec (PCIe 3.0)	DW	FHFL	CPU 8 pin	HPC/AI/Database Analytics
Nvidia	V100	32 GB HBM2	Y	900 GB/sec	300W	NVIDIA NVLink	300 GB/sec (2nd Gen NVLink)	N/A	N/A	N/A	HPC/AI/Database Analytics
Nvidia	V100	16 GB HBM2	Y	900 GB/sec	300W	NVIDIA NVLink	300 GB/sec (2nd Gen NVLink)	N/A	N/A	N/A	HPC/AI/Database Analytics
Nvidia	RTX6000	24 GB GDDR6	Y	624 GB/sec	250W	PCIe Gen3x16/ NVLink bridge3	32 GB/sec3 (PCIe 3.0)	DW	FHFL	CPU 8 pin	VDI/ Performance Graphics
Nvidia	RTX8000	48 GB GDDR6	Y	624 GB/sec	250W	PCIe Gen3x16/ NVLink bridge3	32 GB/sec3 (PCIe 3.0)	DW	FHFL	CPU 8 pin	VDI/ Performance Graphics
Nvidia	P100	16 GB HBM2	Y	732 GB/sec	300W	NVIDIA NVlink	160 GB/sec (1st Gen NVLink)	N/A	N/A	N/A	HPC/AI/Database Analytics

Brand	Model	GPU Memory	Memory ECC	Memory Bandwidth	Max Power Consumption	Host Interface (BW)	GPU-to-GPU Peer Interconnect (BW)	Slot Width	GPU Height/Length	Auxiliary Cable	Workload
Nvidia	P100	16 GB HBM	Y	732 GB/sec	250W	PCIe Gen3x16	32 GB/sec (PCIe 3.0)	DW	FHFL	CPU 8 pin	HPC/AI/Database Analytics
Nvidia	P100	12 GB HBM2	Y	549 GB/sec	250W	PCIe Gen3x16	32 GB/sec (PCIe 3.0)	DW	FHFL	CPU 8 pin	HPC/AI/Database Analytics
Nvidia	P40	24 GB DDR5	N	346 GB/sec	250W	PCIe Gen3x16	32 GB/sec (PCIe 3.0)	DW	FHFL	CPU 8 pin	HPC/AI/Database Analytics

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

		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
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
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

enterprise:nvidia:description:gpu

From:

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

Permanent link:

https://micronode.ru/wiki/hardware/gpu_list

Last update:

2025/04/17 14:21

