■ NVIDIA データセンターシリーズ

モデル	H100 for SXM5	H100 for PCIe	L40	L40S	L4	A100 for SXM 80GB	A100 for PCIe 80GB	NVIDIA A800 40GB Active	A40	A30	A16	A10	A2
GPUアーキテクチャ	Hopper	Hopper	Ada Lovelace	Ada Lovelace	Ada Lovelace	Ampere	Ampere	Ampere	Ampere	Ampere	Ampere	Ampere	Ampere
製造プロセス(nm)	TSMC 4nm	TSMC 4nm	TSMC 4nm	TSMC 5nm	TSMC 4nm	TSMC 7nm	TSMC 7nm	TSMC 7nm	Samsung 8nm	TSMC 7nm		Samsung 8nm	Samsung 8nm
トランジスタ数(billion)	80.0	80.0	76.0	76.0	35.8	54.2	54.2	54.2	28.3	54.2		28.3	
チップのコードネーム	GH100	GH100	AD102	AD102	AD104	GA100	GA100	GA100	GA102	GA100	GA107 x4	GA102	GA107
Compute Capabilty	9.0	9.0	8.9	8.9	8.9	8.0	8.0	8.0	8.6	8.0	8.6	8.6	8.6
ベースクロック(MHz)	-	1125 MHz	735 MHz	1065 MHz	795 MHz	1275 MHz	1065 MHz	765 MHz	1305 MHz	930 MHz	1312 MHz	885 MHz	1440 MHz
GPU Boost クロック(MHz)	-	1755 MHz	2490 MHz	2520 MHz	2040 MHz	1410 MHz	1410 MHz	1410 MHz	1740 MHz	1440 MHz	1755 MHz	1695 MHz	1770 MHz
CUDAコア数	16,896	14,592	18,176	18,176	7,424	6,912	6,912	6,912	10,752	3,804	1280 x4	9,216	1,280
Tensorコア数	528	456	568	568	232	432	432	432	336	224	40 x4	288	40
RTコア数	0	0	142	142	58	0	0	-	84		10 x4	72	10
SM数	132	114	142	-	58	108	108	108				72	
メモリ容量	80GB HBM3	80GB HBM2e	48GB GDDR6	48GB GDDR6	24GB GDDR6	80GB HBM2e	80GB HBM2e	40GB HBM2	48GB GDDR6	24GB HBM2	64GB GDDR6 (16GB per GPU)	24GB GDDR6	16GB GDDR6
メモリバス	5120 bit	5120 bit	384 bit	384 bit	192 bit	5120 bit	5120 bit	5120 bit	384 bit	3072 bit	128 bit	384 bit	128 bit
メモリ帯域幅	3350 GB/s	2048 GB/s	864 GB/s	864 GB/s	300 GB/s	2039 GB/s	1935 GB/s	1555.2 GB/s	696 GB/s	933 GB/s	200 GB/s x4	600 GB/s	200 GB/s
ECC機能サポート	○	○	○	○	○	○	○	○	○	○	○	○	○
バスインターフェース	SXM5	PCI-Express 5.0 x16	PCI-Express 4.0 x16	PCI-Express 4.0 x16	PCI-Express 4.0 x16	SXM4	PCI-Express 4.0 x16	PCI-Express 4.0 x16	PCI-Express 4.0 x16	PCI-Express 4.0 x16	PCI-Express 4.0 x16	PCI-Express 4.0 x16	PCI-Express 4.0 x8
TDP(W)	700 W	350 W	300 W	350 W	72 W	400 W	300 W	240W	300 W	165 W	250 W	150 W	40 - 60 W
放熱機構	Passive	Passive	Passive	Passive	Passive	Passive	Passive	Active	Passive	Passive	Passive	Passive	Passive
NVLink Bridge 対応	NVLink Switch	○	×	×	×	NVLink Switch	○	○	○	○	×	×	×
補助電源	不要	PCIe CEM5 16-pin x1	PCIe CEM5 16-pin x1	PCIe CEM5 16-pin x1	不要	不要	CPU(EPS) 8pin x1	PCIe CEM5 16-pin x1	CPU(EPS) 8pin	CPU(EPS) 8pin	CPU(EPS) 8pin	PCIe 8pin	不要
占有PCIe SLOT数	N/A (SXM module)	2	2	2	1	N/A (SXM module)	2	2	2	2	2	1	1
倍精度性能 FP64 (TFLOPS)	34.0	26.0	-	-	-	9.7	9.7	9.7	-	5.2	-	-	
倍精度性能 FP64 Tensor Core (TFLOPS)	67.0	51.0			-	19.5	19.5		-	10.3	-	-	
単精度性能 FP32 (TFLOPS)	67.0	51.0	90.5	91.6	30.3	19.5	19.5	19.5	37.4	10.3	4.5 x4	31.2	4.5
単精度性能 FP32 Tensor Core (TFLOPS)		756.0			-						-	-	-
TensorFloat-32性能 TF32		378.0	90.5	183.0	60.0	156.0	156.0		74.8	82.0	9 x4	62.5	9.0
TensorFloat-32性能 TF32 with sparsity	989.0	756.0	181.0	366.0	120.0	312.0	312.0	311.8	149.6	165.0	18 x4	125.0	18.0
BFLOAT16 Tensor Core (TFLOPS)		-	181.1	362.1	121.0	312.0	312.0	312.0	149.7	165.0	-	125.0	18.0
BFLOAT16 Tensor Core (TFLOPS) with sparsity	1979.0	1513.0	362.1	733.0	242.0	624.0	624.0	624.0	299.4	330.0	-	250.0	36.0
FP16 Tensor Core (TFLOPS)		756.0	181.1	362.1	121.0	312.0	312.0	312.0	149.7	165.0	17.9 x4	125.0	18.0
FP16 Tensor Core (TFLOPS) with sparsity	1979.0	1513.0	362.1	733.0	242.0	624.0	624.0	624.0	299.4	330.0	35.9 x4	250.0	36.0
FP8 Tensor Core (TFLOPS)		1513.0		733.0	242.5								
FP8 Tensor Core (TFLOPS) with sparsity	3958.0	3026.0		1466.0	485.0								
INT8 Tensor Core (TOPS)		1513.0	362.0	733.0	242.5	624.0	624.0		299.3	330.0	35.9 x4	250.0	36.0
INT8 Tensor Core (TOPS) with sparsity	3958.0	3026.0	724.0	1466.0	485.0	1248.0	1248.0	1247.4	598.6	661.0	71.8 x4	500.0	72.0
INT4 Tensor Core (TOPS)	-	-	724.0	733.0	-	-	-	1248.0	598.7	661.0		500.0	72.0
INT4 Tensor Core (TOPS) with sparsity	-	-	1448.0	1466.0	-	-	-	2496.0	1197.4	1321.0		1000.0	144.0
RTコア性能 (TFLOPS)			209.0	212.0					73.1				
ピーク Tensor TFLOPS				733.0	485.0			623.8					
ディスプレイポート	なし	DP 1.4a x4	DP 1.4 x4	DP 1.4a x4	なし	なし	なし	なし	DP 1.4 x4	なし	なし	なし	なし
vGPU対応	○	○	○	○	○	○	○		○	○	○	○	○
Multi-Instance GPU (MIG)対応	○	○	×	×	×	○	○	○	×	○	×	×	