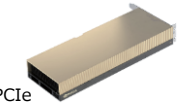
GPUアクセラレータボードシリーズ　HPCやデータセンターでの利用を想定した製品

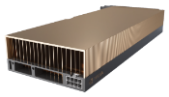
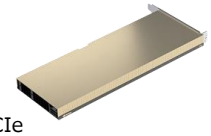
		H100		A100		A30	V100		V100S	T4
		 SXM5	 PCIe	 SXM4	 PCIe	 PCIe	 SXM2	 PCIe	 PCIe	 PCIe
アーキテクチャ		Hopper	Hopper	Ampere	Ampere	Ampere	Volta	Volta	Volta	Turing
GPU Core名称		GH100	GH100	GA100	GA100	GA100	GV100	GV100	GV100	TU104
Shadingコア数		8,448	7,296	6,912	6,912	8,192	5120	5120	5120	2560
Tensorコア数		第4世代 528	第4世代 456	第３世代 432	第３世代 432	第３世代 512	第2世代 640	第2世代 640	第2世代 640	320
SM数		132	114	108	108	128	80	80	80	40
GPU ベースクロック		1095 MHz	1125 MHz	765 MHz	765 MHz	930 MHz	1297 MHz	1230 MHz	1245 MHz	660MHz
GPU ブーストクロック		1755 MHz	1650 MHz	1410 MHz	1410 MHz	1440 MHz	1530 MHz	1380 MHz	1597 MHz	1582MHz
メモリ インターフェイス		5120-bit	5120-bit	5120-bit	5120-bit	3072 bit	4096-bit	4096-bit	4096-bit	256-bit
メモリ容量		80GB	80GB	40GB/80GB	40GB/80GB	24 GB	16GB /32GB	16GB /32GB	32GB	16GB
メモリータイプ		HBM3	HBM3	HBM2	HBM2	HBM2	HBM2	HBM2	HBM2	GDDR6
メモリー帯域幅		3000GB/s	2000GB/s	1555/2039GB/s	1555/1953GB/s	933.1 GB/s	900GB/s	900GB/s	1134GB/s	320GB/s
接続バスコネクタ		NVLink 900GB/s 12Link/GPU	PCIe Gen5 x16	NVLink 600GB/s 12Link/GPU	PCIe Gen4 x16	PCIe Gen4 x16	NVLink 300GB/s 6Link/GPU	PCIe Gen3 x16	PCIe Gen3 x16	PCIe Gen3 x16
パフォーマンス（理論値）	FP64	30	24	9.7	9.7	5.2	7.8	7.0	8.2	0.3
	FP32	60	48	19.5	19.5	10.3	15.7	14.0	16.4	8.1
	FP16	120	96	77.8	77.8	10.3	30.4	28.0	32.8	16.2
熱設計電力 (TDP)		700W	350W	400W	250/300W	165W	300W	250W	250W	70W
FAN		Passive	Passive	Passive	Passive	Passive	Passive	Passive	Passive	Passive
電源コネクタ		-	8pin x1 EPS	-	8pin x1 EPS	8pin x1 EPS	-	8pin x1	8pin x1	PCIeバス給電
フォームファクタ		SXM4	デュアルスロット	SXM4	デュアルスロット	デュアルスロット	SXM2	デュアルスロット	デュアルスロット	シングルスロット ロープロファイル
おすすめ用途 *		HPC/AI	HPC/AI	HPC/AI	HPC/AI	HPC/AI	HPC/AI	HPC/AI	HPC/AI	AI
その他　グラフィック出力		「Tenserコアによる複合精度演算」でFP64サポート					-	-	-	-

* [HPC] FP64演算を必要とする用途　[AI] FP32/16 演算で十分な用途

※性能は、実際のカードメーカの製品により異なる場合があります。

※Volta以降で搭載された「Tenserコアによる複合精度演算」は世代により効果は異なりますが、アプリケーションが対応していれば、FP16/32/FP64演算に於いて、桁違いの演算能力を発揮することができます。
例 H100の　GPU性能　ー>　Tensorコア性能
FP64　30　→　60TF
FP32　60 →　500TF(疎行列要素の場合 1000TF)
FP16　120 →　1000TF(疎行列要素の場合 2000TF)

GPUアクセラレータボードシリーズ　HPCやデータセンターでの利用を想定した製品

		GV100	A40	A16	A10	A2	RTX 8000 Passive	RTX 6000 Passive
		 PCIe	 PCIe	 PCIe	 PCIe	 PCIe	 PCIe	 PCIe
アーキテクチャ		Volta	Ampere	Ampere	Ampere	Ampere	Turing	Turing
GPU Core名称		GV100	GA102	GA107	GA102	GA107	TU102	TU102
Shadingコア数		5120	10,752	2,560	10,752	1,280	4608	4608
Tensorコア数		640	336	80	288	40	288	288
SM数		80	84	20	72	10	72	72
GPU ベースクロック		1132 MHz	1305MHz	885 MHz	885 MHz	1440 MHz	1230 MHz	1305 MHz
GPU ブーストクロック		1627 MHz	1740MHz	1695 MHz	1695 MHz	1770 MHz	1620 MHz	1560 MHz
メモリ インターフェイス		4096 bit	384-bit	128 bit	384 bit	128 bit	384bit	384 bit
メモリ容量		32GB	48GB	16 GB x4	24 GB	16 GB	48GB	24 GB
メモリータイプ		HBM2	GDDR6	GDDR6	GDDR6	GDDR6	GDDR6	GDDR6
メモリー帯域幅		868.4 GB/s	696.0GB/s	231.9 GB/s	600.2 GB/s	200.1 GB/s	672.0 GB/s	672.0 GB/s
接続バスコネクタ		PCIe Gen3 x16	PCIe Gen4 x16	PCIe Gen4 x16	PCIe Gen4 x16	PCIe Gen4 x8	PCIe Gen3 x16	PCIe Gen3 x16
パフォーマンス（理論値TF）	FP64	8.3	1.2	0.3	1.0	0.1	0.5	0.5
	FP32	16.7	37.4	8.7	31.2	4.5	14.9	14.4
	FP16	33.3	37.4	8.7	31.2	4.5	29.9	28.8
熱設計電力 (TDP)		280W	300W	250W	150W	250W	260W	260W
FAN		Active	Passive	Passive	Passive	Passive	Passive	Passive
電源コネクタ		8pin x1、6pin x1	8pin x1 EPS	8pin x1 EPS	8pin x1 EPS	8pin x1 EPS	8pin x1、6pin x1	8pin x1、6pin x1
フォームファクタ		デュアルスロット	デュアルスロット	デュアルスロット	シングルスロット	デュアルスロット	デュアルスロット	デュアルスロット
おすすめ用途 *		HPC/AI	AI・vGPU	AI・vGPU	AI・vGPU	AI	AI	AI
その他　グラフィック出力		-	DP（P1.4a）×3 デフォルトは非アクティブ NVIDIA vGPU対応 （DPアクティブ時は非対応） MIG（マルチGPUインスタンス）は非サポート	4GPU搭載 NVIDIA vGPU対応	NVIDIA vGPU対応	グラフィック出力ポートなし	グラフィック出力ポートなし	グラフィック出力ポートなし

* [HPC] FP64演算を必要とする用途　[AI] FP32/16 演算で十分な用途

※性能は、実際のカードメーカの製品により異なる場合があります。

NVIDIA RTX SERIES

プロフェッショナル グラフィックスカード製品

VISUAL TECHNOLOGY

		RTX A6000	RTX A5500	RTX A5000	RTX A4500	RTX A4000	RTX A2000	RTX 8000	RTX 6000	RTX 5000	RTX 4000
											
アーキテクチャ		Ampere	Ampere	Ampere	Ampere	Ampere	Ampere	Turing	Turing	Turing	Turing
GPU Core名称		GA102	GA102	GA102	GA102	GA104	GA104	TU102	TU102	TU104	TU104
CUDA コア数		10752	10240	8,192	7,168	6144	3228	4608	4608	3072	2304
Tensorコア数		336	320	256	224	192	104	576	576	384	288
SM数		84	80	64	56	48	26	72	72	48	36
GPU ベースクロック		1455 MHz	1170 MHz	1170 MHz	1050 MHz	735 MHz	562 MHz	1395 MHz	1440 MHz	1620 MHz	1005 MHz
GPU ブーストクロック		1860 MHz	1695 MHz	1695 MHz	1650 MHz	1560 MHz	1200 MHz	1770 MHz	1770 MHz	1815 MHz	1545 MHz
メモリ インターフェイス		384bit	384bit	384bit	320 bit	256 bit	192bit	384 bit	384 bit	256 bit	256 bit
メモリ容量		48GB	24 GB	24 GB	20GB	16GB	6GB/12GB	48GB	24GB	16GB	8GB
メモリータイプ		GDDR6	GDDR6	GDDR6	GDDR6	GDDR6	GDDR6	GDDR6	GDDR6	GDDR6	GDDR6
メモリー帯域幅		768.0 GB/s	768.0 GB/s	768.0 GB/s	640.0 GB/s	448.0 GB/s	288.0 GB/s	672.0 GB/s	672.0 GB/s	448.0 GB/s	416.0 GB/s
接続バスコネクタ		PCIe 4.0 x16	PCIe 4.0 x16	PCIe 4.0 x16	PCIe 4.0 x16	PCIe 4.0 x16	PCIe 4.0 x16	PCIe 3.0 x16	PCIe 3.0 x16	PCIe 3.0 x16	PCIe 3.0 x16
パフォーマンス (理論値)	FP64	1.3	1.1	0.9	0.7	0.6	0.6	0.5	0.5	0.4	0.2
	FP32	40.0	34.7	27.8	23.7	19.2	19.2	16.3	16.3	11.2	7.1
	FP16	40.0	34.7	27.8	23.7	19.2	19.2	32.6	32.6	22.3	14.2
熱設計電力 (TDP)		300W	230W	230W	200W	140W	70W	260W	260W	230W	160W
FAN		Active	Active	Active	Active	Active	Active	Active	Active	Active	Active
電源コネクタ		8-pin EPS	8pin x1	8pin x1	8pin x1	6pin x1	6pin x1	8pin x1、6pin x1	8pin x1、6pin x1	8pin x2、6pin x1	6pin x1
フォームファクタ		デュアルスロット	デュアルスロット	デュアルスロット	デュアルスロット	シングルスロット	デュアルスロットLP	デュアルスロット	デュアルスロット	デュアルスロット	シングルスロット
おすすめ用途 *		グラフィックス・AI	グラフィックス・AI	グラフィックス・AI	グラフィックス・AI	グラフィックス・AI	グラフィックス・AI	グラフィックス・AI	グラフィックス・AI	グラフィックス・AI	グラフィックス・AI
その他　グラフィック出力		DP 1.4a ×4	DP 1.4a ×4	DP 1.4a ×4	DP 1.4a ×4	DP 1.4a ×4	DP 1.4a ×4	DP 1.4 ×4、TypeC	DP 1.4 ×4、TypeC	DP 1.4 ×4、TypeC	4x DisplayPort

* [HPC] FP64演算を必要とする用途　[AI] FP32/16 演算で十分な用途

※性能は、実際のカードメーカの製品により異なる場合があります。

NVIDIA GeForce RTX™ 30 コンシューマ グラフィックスカード製品

		GEFORCE RTX 3090	GEFORCE RTX 3080Ti	GEFORCE RTX 3080	GEFORCE RTX 3070Ti	GEFORCE RTX 3070	GEFORCE RTX 3060 Ti	GEFORCE RTX 3060
								
アーキテクチャ		Ampere	Ampere	Ampere	Ampere	Ampere	Ampere	Ampere
GPU Core名称		GA102	GA102	GA102	GA104	GA104	GA104	GA106
CUDA コア数		10496	10240	8704	6144	5888	4864	3584
Tensorコア数		328	320	272	192	184	152	112
SM数		82	80	68	48	46	38	28
GPU ベースクロック		1400MHz	1365 MHz	1440MHz	1575 MHz	1500MHz	1410MHz	1320MHz
GPU ブーストクロック		1700MHz	1665 MHz	1710MHz	1730MHz	1730MHz	1670MHz	1780MHz
メモリ インターフェイス		384 bit	324 bit	324 bit	256 bit	256 bit	256 bit	192 bit
メモリ容量		24GB	12GB	10GB	8GB	8GB	8GB	12GB
メモリータイプ		GDDR6X	GDDR6X	GDDR6X	GDDR6X	GDDR6	GDDR6	GDDR6
メモリー帯域幅		936GB/s	912.4 GB/s	760GB/s	608.3 GB/s	448GB/s	448GB/s	360GB/s
接続バスコネクタ		PCIe 4.0 x16	PCIe 4.0 x16	PCIe 4.0 x16	PCIe 4.0 x16	PCIe 4.0 x16	PCIe 4.0 x16	PCIe 4.0 x16
パフォーマンス (理論値)	FP64	0.6	0.5	0.5	0.3	0.3	0.3	0.2
	FP32	35.6	34.1	29.8	21.8	20.3	16.2	12.7
	FP16	35.6	34.1	29.8	21.8	20.3	16.2	12.7
熱設計電力 (TDP)		350W	350W	320W	290W	220W	200W	170W
電源コネクタ		8Pin x2	8Pin x2	8Pin x2	8Pin x2	8Pin x1	8Pin x1	8Pin x1
FAN		Active	Active	Active	Active	Active	Active	Active
フォームファクタ		トリプルスロット	デュアルスロット	デュアルスロット	デュアルスロット	デュアルスロット	デュアルスロット	デュアルスロット
おすすめ用途 *		グラフィックス・AI	グラフィックス・AI	グラフィックス・AI	グラフィックス・AI	グラフィックス・AI	グラフィックス・AI	グラフィックス・AI
その他　グラフィック出力		HDMI 2.1,DP 1.4a x3	HDMI 2.1,DP 1.4a x3	HDMI 2.1,DP 1.4a x3	HDMI 2.1,DP 1.4a x3	HDMI 2.1,DP 1.4a x3	HDMI 2.1,DP 1.4a x3	HDMI 2.1,DP 1.4a x3

* [HPC] FP64演算を必要とする用途　[AI] FP32/16 演算で十分な用途

※性能は、実際のカードメーカの製品により異なる場合があります。