GPUアクセラレーター　現行製品仕様一覧

（随時更新されます）

HPC/AI/ディープラーニングなど取り組まれる方に、最適なGPUアクセラレーター製品を搭載した計算環境をご提供します。
主要なGPU（NVIDIA® Tesla®、NVIDIA® Quadro®、NVIDIA® TITAN®、NVIDIA GeForce™）のラインナップをまとめます。

・GPGPU利用を想定し、グラフィックカードとしての機能は記載していません。
・実際の製品は、複数のメーカーからリリースされている場合があり、仕様が異なる場合があります。
・表中の表記：N/A (not applicable, no answer)、TBC (to be confirmed)

NVIDIA® Tesla®　サーバー、データ センター向け GPGPU

ECC（メモリのエラー検出・訂正の機能）対応

製品名	T4 PCIe	A100 SXM2	A100 PCIe	V100 SXM2	V100 PCIe	V100S PCIe
GPU Core名称	TU104	GA100	←	GV100	GV100	GV100
CUDA コア数	2560	6912	←	5120	←	←
Tensorコア数　※1	320	640	←	640	←	←
SM数	40	108	←	80	←	←
GPU ベースクロック	660MHz	1410MHz	←	1297 MHz	1230 MHz	1245 MHz
GPU ブーストクロック	1582MHz	TBC	←	1530 MHz	1380 MHz	1597 MHz
メモリ インターフェイス	256-bit	5120-bit	←	4096-bit	←	←
メモリ容量	16GB	40GB	←	16GB /32GB	16GB /32GB	32GB
メモリータイプ	GDDR6	HBM2E	←	HBM2	←	←
相互接続帯域幅(双方向)	32GB/sec	600GB/sec	TBC	300GB/sec	32GB/sec	32GB/sec
メモリー帯域幅	320GB/sec	1555GB/s	←	900GB/sec	900GB/sec	1134GB/sec
接続バスコネクタ	PCIe Gen3 x16	NVLink600GB/s 12Link/GPU	PCIe Gen4 x16	NVLink 300GB/s 6Link/GPU	PCIe Gen3 x16	←
パフォーマンス (理論値)						
FP64(TFLOPS)	N/A	9.7	9.7	7.8	7.0	8.2
FP32	8.1	19.5	19.5	15.7	14.0	16.4
FP16	16.2	77.8	77.8	30.4	28.0	32.8
INT32(TOPS)	8.1	TBC	TBC	14.9	TBC	TBC
Tensorコア 混合精度(TOPS)※1 FP16/32 INT8 INT4	 65 130 260	 19.5 312/156 624 1248	 19.5 312/156 624 1248	 N/A N/A / 125 N/A N/A	 N/A N/A / 112 N/A N/A	 N/A N/A / 130 N/A N/A
熱設計電力 (TDP)　最大消費電力	70W	400W	260W	300W	250W	250W
FAN	Passive	Passive	Passive	Passive	Passive	Passive
電源コネクタ	PCIeバス給電	ー	PCIeバス給電	ー	8pin x1	8pin x1
フォームファクタ	シングルスロット ロープロファイル	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





NVIDIA® Quadro®

デスクトップ ワークステーション向けGPU

プロフェッショナル用途で、耐久性や信頼性が高く、大容量メモリ搭載
グラフィックカードのため、基本的に倍精度浮動小数点演算能力は低い

製品名	GV100	RTX 8000	RTX 8000 Passive	RTX 6000	RTX 6000 Passive	RTX 5000	RTX 4000
GPU Core名称	GV100	TU102	TU102	TU102	TU102	TU104	TU104
CUDA コア数	5120	4608	←	←	←	3072	2304
Tensorコア数	640	576	←	←	←	384	288
SM数	80	72	←	←	←	48	36
GPU ベースクロック	1132 MHz	1395 MHz	1230 MHz	1440 MHz	1305 MHz	1620 MHz	1005 MHz
GPU ブーストクロック	1627 MHz	1770 MHz	1620 MHz	1770 MHz	1560 MHz	1815 MHz	1545 MHz
メモリ インターフェイス	4096 bit	384 bit	←	←	←	256 bit	256 bit
メモリ容量	32GB	48GB	48GB	24 GB	24 GB	16GB	8GB
メモリータイプ	HBM2	GDDR6	←	←	←	GDDR6	GDDR6
メモリー帯域幅	868.4 GB/s	672.0 GB/s	←	←	←	448.0 GB/s	416.0 GB/s
接続バスコネクタ	PCIe Gen3 x16	←	←	←	←	←	←
パフォーマンス (理論値)							
FP64(TFLOPS)	8.3	0.51	0.47	0.51	0.45	0.35	0.22
FP32	16.7	16.3	14.9	16.3	14.4	11.2	7.1
FP16	33.3	32.6	29.9	32.6	28.8	22.3	14.2
INT32(TOPS)	TBC	←	←	←	←	←	←
Tensorコア 混合精度(TOPS)※1							
FP16/32	133	130	119	130	115	89	57
INT8	N/A	260	238	260	230	178	114
INT4	N/A	520	476	520	460	356	228
熱設計電力 (TDP)	280W	260 W	←	←	←	230W	160 W
FAN	Active	Active	Passive	Active	Passive	Active	Active
電源コネクタ	8pin x1、6pin x1	8pin x1、6pin x1	8pin x1、6pin x1	8pin x1、6pin x1	8pin x1、6pin x1	8pin x1、6pin x1	8pin 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	デュアルスロット	デュアルスロット	シングルスロット
その他	ECC対応						

NVIDIA GeForce®

NVIDIA TITAN®

コンシューマ向け高性能GPU

GeForceはQuadroやTeslaと比べ、クロックが高く1コアあたりの性能が高い
グラフィックカードのため、基本的に倍精度浮動小数点演算能力は低い
コンシューマ向けの製品で安価

製品名	TITAN RTX	RTX 2080 Ti	RTX 2080 Super	RTX 2070 Super	RTX 2060 Super
GPU Core名称	TU102	TU102	TU104	TU104	TU106
CUDA コア数	4608	4352	3072	2560	2176
Tensorコア数	567	544	384	320	272
SM数	72	68	48	40	34
GPU ベースクロック　(MHz)	1350 MHz	1350 MHz	1650 MHz	1605 MHz	1470 MHz
GPU ブーストクロック　(MHz)	1770 MHz	1545 MHz	1815 MHz	1770 MHz	1650 MHz
メモリ インターフェイス	384 bit	352 bit	256 bit	←	←
メモリ容量	24GB	11 GB	8 GB	←	←
メモリータイプ	GDDR6	GDDR6	GDDR6	←	←
メモリー帯域幅	672 GB/s	616.0 GB/s	495.9 GB/s	448.0 GB/s	←
接続バスコネクタ	PCIe Gen3 x16	←	←	←	←
パフォーマンス (理論値)					
FP64(TFLOPS)	0.51	0.42	0.35	0.28	0.22
FP32	16.3	13.5	11.2	9.1	7.2
FP16	32.6	26.9	22.3	18.1	14.4
INT32(TOPS)	TBC	←	←	←	←
Tensorコア 混合精度(TOPS)※1					
FP16/32	130	107	89	72	57
INT8	260	214	178	144	114
INT4	520	428	356	288	228
熱設計電力 (TDP)	280W	250 W	250 W	215 W	160 W
電源コネクタ	8pin x2	8pin x2	8pin x1、6pin x1	8pin x1、6pin x1	8pin x1
FAN	Active	Active	Active	Active	Active
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	デュアルスロット
その他					

※1
Tensorコアは4x4のマトリックスの積和算ユニットで、FP16で4行4列の行列積を4列並列に実行します。
1個コアあたり、64の乗算ユニットと16の加算ユニットをもち、1サイクルで64の乗算と和算(FP32精度に拡張）が可能です。
TuringアーキテクチャのGPUとCUDA10の組み合わせでINT8/INT4 の計算も高速化します。原理的にはINT8 は2倍、INT4 は4倍の演算能力（OPS)となります。
上記のスペックでは「複合精度（Tensorコア）」として、以下の式により論理的なピーク値を記載しています。
[SMの個数]×[SMあたりのTensor Core数]×128[行列演算の演算数:64個の乗算と加算]×[動作クロック]