VISUAL TECHNOLOGY

2020/5/19

Product	A100	V100	P100
Architecture	Ampere	Volta	pascal
Code name	GA100	GV100	GP100
Process (nm)	7	12	16
Transistor	54 billion	21.1billion	15.3 billion
GPU Board Form Factor	SXM4	SXM2	SXM
SMs	108	80	56
FP32Cores/SM	64	64	64
FP64Cores/SM	32	32	32
INT32Cores/SM	64	64	-
TensorCores/SM	4	8	-
TPCs	56	40	28
Memory Capacity (GB)	40	32	16
Memory Bus Width (bit)	5120	4096	4096
Memory Clock (GHz)	2.4	1.75	1.4
Memory Bandwidth (GB/sec)	1600	900	720
NVLink Bandwidth (GB/sec)	600	300	160
(PCIe x16 model) (GB/sec)	(Gen4 63)	(Gen3 31.5)	(Gen3 31.5)
CUDA Cores (32 bit)	6912	5120	3584
CUDA Cores (64 bit)	3456	2560	1792
Tensor Cores	432	640	-
FP16 (TFLOPS)	78.0	31.4	21.2
FP32 (TFLOPS)	19.5	15.7	10.6
FP64 (TFLOPS)	9.7	7.8	5.3
INT32 (TOPS)	19.5	15.7	-
TC-INT4 (TOPS)	1248/2496※	-	-
TC-INT8(TOPS)	624/1248※	-	-
TC-FP16 (TFLOPS)	312/624※	125※※	-
TC-BF16 (TFLOPS)	312/624※	-	-
TC-TF32 (TFLOPS)	156/312※	-	-
TC-FP64 (TFLOPS)	19.5	-	-
TDP (W)	400	300	300

Tensor Core対応データ型の拡

※「Structural sparsity」:ハードウェアによるプルーニング機能を有効化した場合

※※TC-FP16/32。「Automatic Mixed Precision (AMP)」:混合精度演算を使用した場合

TC = Tensor Core
BF = Brain Floating Point
TF = Tensor Float

FP16	Sign 1-bit、Exponent 5-bit、Mantissa 10-bit	IEEE 754形式
FP32	Sign 1-bit、Exponent 8-bit、Mantissa 23-bit	IEEE 754形式
FP64	Sign 1-bit、Exponent 11-bit、Mantissa 52-bit	IEEE 754形式
BF16	Sign 1-bit、Exponent 8-bit、Mantissa 7-bit	
TF32	Sign 1-bit、Exponent 8-bit、Mantissa 10-bit	