NVIDIA GPU Comparison (Tesla)

	GP100	GM200	GK110	GF110
	P100	M40	K40	M2090
Architecture	Pascal	Maxwell	Kepler	Fermi
Year	2016	2015	2013	2011
FP32 Cuda Cores (FMAD)	3584	3072	2880	512
SMs	56	24	15	16
Clock Base (Boost)	1328MHz (1480MHz)	948MHz (1114MHz)	745MHz (875MHz)	1300MHz (base 650MHz)
GFLOPS (FP32) @base clock	9519 (10609) GFLOPS	5824 (6844) GFLOPS	4291 (5040) GFLOPS	1331 GFLOPS
FP64 FMAD	1792	96	960	256
GFLOPS (FP64) @base clock	4760 (5304) GFLOPS	182 (214) GFLOPS	1430 (1680) GFLOPS	666 GFLOPS
FP16 FMAD	7168	3072	2880	512
GFLOPS (FP16) @base clock	19038 (21217) GFLOPS	5824 (6844) GFLOPS	4291 (5040) GFLOPS	1331 GFLOPS
Register Files	14336 KB	6144 KB	3840 KB	2048 KB
Shared Memory	3584 KB	2304 KB	960 KB	768 KB
L2 Cache	4096 KB	3072 KB	1536 KB	768 KB
Warp In-Flight	3584	1536	960	768
Memory Bandwidth	720 GB/sec	288 GB/sec	288 GB/sec	178 GB/sec
TDP (Thermal Design Power)	300W	250W	235W	225W
FP32 GFLOPS /TDP	31.7	23.3	18.3	5.9
Transistors	15.3B	8B	7.08B	3B
Die Size	610mm2	601mm2	561mm2	520mm2
Process Technology	16 nm	28 nm	28 nm	40nm
Byte/ FLOPS(FP32)	0.075	0.049	0.067	0.133