Cuda Tensor Core

□□□□□□□□□□□□□□□□

□□□□□□ GPU□□□□□□□□□□□□□□□□□□□□□□□□□□□□

- WGMMA□□□□□□□□□□□□□□□□□□□□
- □□□□□□□□□□□□□□□□□□□□□□□□□□□□
- □□□□□□□□□□
- □□□□□□□□□□□□□□□□□□□□□□□□□□□□

□□ GPU□□□□□□□□□□ wmma.mma.sync□ mma.sync□□□ SM□□□□□□□□ 32□□□□□□ warp □□□□□□□□□□□□□□□□□□□□

wgmma.mma_async□□□□□□□□□□ 128□□□□□□ SM □□□□□□□□□□□□□□□□□□□□

BF16	-	-	√	√
FP16	√	√	√	√
FP8	-	-	-	√
INT8/UINT8	-	√	√	√
INT4/UINT4	-	√	√	-
INT1	-	√	√	-

Volta	Turing	Ampere	Hopper	
SM/TPC	2	2	2	2
process blocks/SM	4	4	4	4
FP64/PB	8	-	8	16
FP32/PB	16	16	16	32
INT32/PB	16	16	16	16
tensorCore/PB	2	2	1	1
LSU/PB	8	4	8	8
register file	64KB(16384*32bit)	64KB(16384*32bit)	64KB(16384*32bit)	64KB(16384*32bit)
L0 ICache/PB	1	1	1	1
L1/SHM/SM	128KB	96KB	192KB	256KB
warp scheduler/PB	1(32thread/clk)	1(32thread/clk)	1(32thread/clk)	1(32thread/clk)
dispatch uint/PB	1(32thread/clk)	1(32thread/clk)	1(32thread/clk)	1(32thread/clk)

US AI Semiconductor Controls

GPU	Memory Capacity (GB)	Memory Bandwidth (Tbps)	TeraFLOPs	Bitlength	TPP (TeraFLOPS x Bitlength)	Die size (mm^2)	Performance density (TPP / Die size)	Rule 3A090.a	Rule 3A090.b
H100 SXM	**80**	**3.4**	1,979	8	15,832	814	19.4	APPLIES	DOESN'T APPLY
H20 SXM	**96**	**4.0**	296	8	2,368	814	2.9	DOESN'T APPLY	DOESN'T APPLY
L40S	48	0.9	733	8	5,864	608	9.6	APPLIES	DOESN'T APPLY
L40	48	0.9	362	8	2,896	608	4.8	DOESN'T APPLY	APPLIES
L20	48	0.9	239	8	1,912	608	3.1	DOESN'T APPLY	DOESN'T APPLY
L4	24	0.3	242	8	1,936	295	6.6	APPLIES	DOESN'T APPLY
L2	24	0.3	193	8	1,544	295	5.2	DOESN'T APPLY	DOESN'T APPLY
A100 SXM	40	1.6	312	16	4,992	826	6.0	APPLIES	DOESN'T APPLY
V100 SXM	16	0.9	125	16	2,000	815	2.5	DOESN'T APPLY	DOESN'T APPLY
RTX 4090[1]	24	1.0	661	8	5,285	609	8.7	APPLIES	DOESN'T APPLY
RTX 4080[1]	16	0.7	320	8	2,560	379	6.8	APPLIES	DOESN'T APPLY
AMD MI210	64	1.6	181	16	2,896	770	3.8	DOESN'T APPLY	APPLIES
AMD MI250X	128	3.2	383	16	6,128	1,540	4.0	APPLIES	DOESN'T APPLY
AMD MI300X[2]	192	5.6	2,400	8	19,200	2,381	8.1	APPLIES	DOESN'T APPLY
Intel Gaudi2[2]	96	2.5	700	8	5,600	826	6.8	APPLIES	DOESN'T APPLY

1. Not "designed" for datacenter
2. No official specs, estimated

tensorCore□□□□□□ SM□□ thread□ warp□□□□□□□□□ thread □□□□□□□□□□□□□□□□□□□□□□□□□□□□□□ SM□□□ thread□ tensorCore□□□□□□□

Revision #1
Created 11 January 2025 09:46:27 by Colin
Updated 12 January 2025 06:34:03 by Colin