

1. SRAM
1. bank
2. cycle by cycle
3. latency

NV Blackwell

PTX SASS AI

<https://docs.nvidia.com/cuda/parallel-thread-execution/index.html#tensorcore-5th-generation-family-instructions> <https://docs.nvidia.com/cuda/cuda-binary-utilities/index.html#blackwell-instruction-set>

1. L0 memory
 1. tensor memory 2D/1D NPU/DSA
 2. Tensor memory to memory Tensor
 3. 1D load/store memory
 4. L0 2D/1D L0 fusion code Fusion
2. 128/256 2D
3. TPC 2 SM 2D 2 core
4. OCP-MX micro-scaling
5. thread issue Tensor SIMT style thread
6. weight-stationary GEMM mask bit padding 0