

AI

2D

1. Dot 9
- 2.
3. L1/L2/L0 reshape swizzel
4. layout NHWC BPI BPK FF
- 5.
6. L0 L1 Fusion
7. mapping fusion
8. feature*weight vs weightT*featureT

2D Dot

1. 1D 2D fusion
1. L0 broadcast L1
2. load/store
3. 2D
4. / /
2. cuda simt warp thread
1. thread
2. 2D
- 3.
- 4.
5. sm subcore block group
3. GCU4.0 thread footprint order thread
1. L1load load L0 subcore
4. 2D-1D-2D Fusion
- 1.
1. 8bit int8 64 512bit 32bit 16
- 2.
3. layout NCHW NHWC
4. stride 1
5. 2D 1D 2D
6. L1 latency L1 L1_base x latency L0
7. swizzle renaming L0 bank conflict renaming



1. SRAM
 1.
 2. 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 Tensor memory ☐ 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

Revision #5

Created 11 January 2025 09:46:28 by Colin

Updated 24 January 2025 13:37:22 by Colin