

# 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\*featueT

## 2D Dot

1. 1D 2D fusion  
  1. L0 broadcast L1  
2D
  2. load/store
  3. 2D
  4. / /
2. cuda simt warp thread  
2D PPA  
  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  
L0\_base L1\_base
7. swizzle renaming L0 bank conflict renaming



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