$$\text{Layer Norm}(x) = \frac{x - \mu}{\sigma} \cdot \gamma + \beta$$

$\mu:$ 均值

$\sigma:$ 方差

$$\sigma = \sqrt{\frac{\sum\limits_{1}^{n}(x_i - \mu)^2}{n}}$$

(-128, 127)

$x:$ int32?

$\mu:$ int32?

$\sigma:$ float?

γ ?

β ?

(0.01, 0.1)

$1e^{-3}$

error_rate: < 10%

DDR

width: 32
stream1

depth: 32

cols: 512
ram2
rows: 32

LOAD

DECODE

128
ram_idx

32
ram_count

MUL

STORE

outputs

LOAD:

DDR
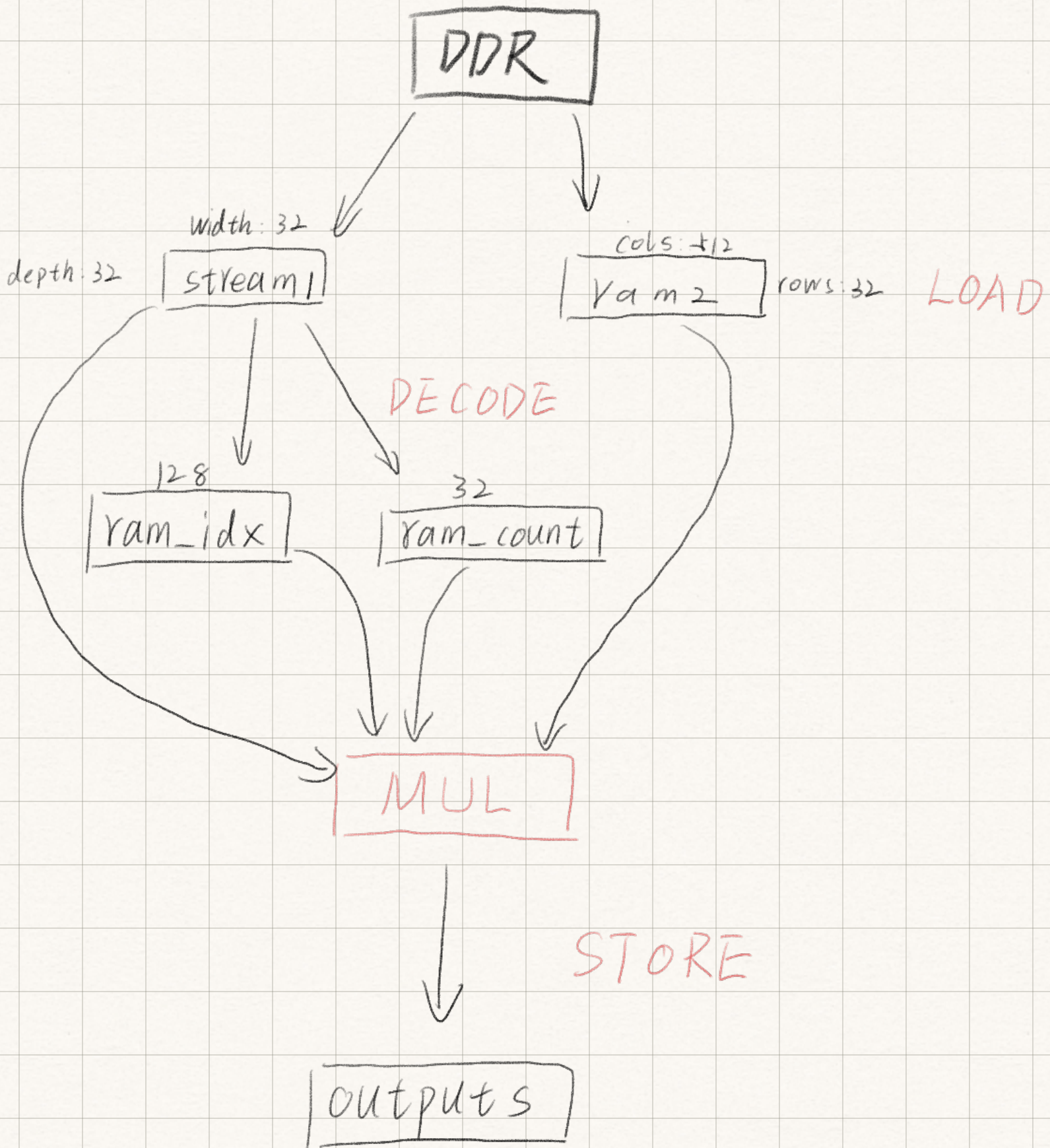
$addr_1$ → stream1

for times
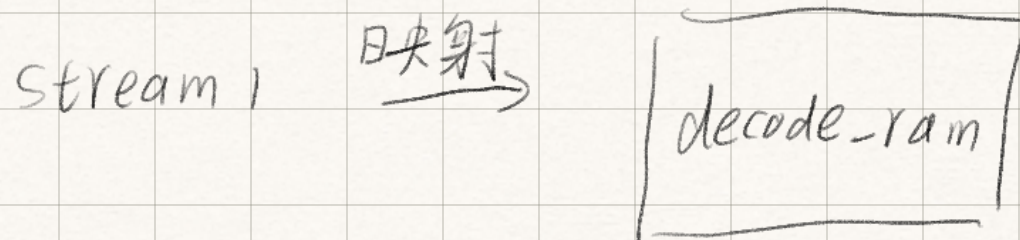memcpy(data, inputs)
stream1.write
(data)

$addr_2$ → stream2

memcpy (stream2,
inputs,
ROWS * COLS *
sizeof (DataType_IN)
/ sizeof (DataType_AXI)

DECODE

$$decode\_ram\,[ROWS][COLS]$$

memcpy (&decode_ram[0], & stream1[0],

　　　　ROWS * COLS * sizeof (DataType_IN));

stream1　　映射→　　| decode_ram |

idx_stream　　widpth: 8　　depth: 128

count_stream　　widpth: 8　　depth: 32

for ROWS
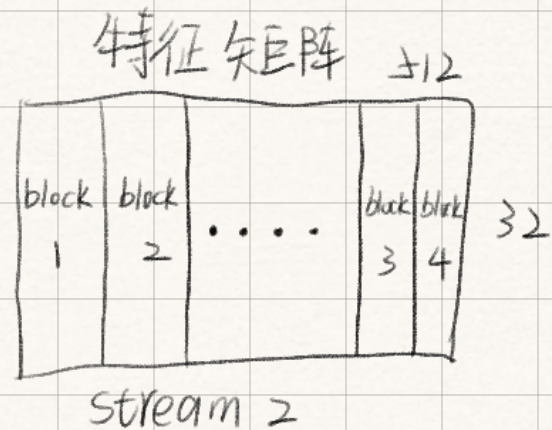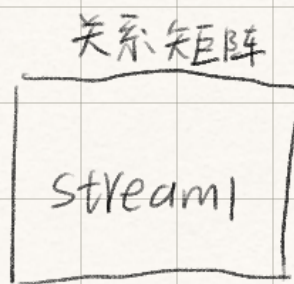
　　　　int8 idx ;
　　for COLS
　　　　decode_ram [row][col] ==1 ?
　　　　count ++
　　　　idx [count] = col;

　　idx_stream. write (idx);

　　count_stream. write (count);

MUL

关系矩阵

特征矩阵 512

stream1

| block 1 | block 2 | . . . . | block 3 | block 4 | 32

stream 2

for COLs ↗32

   for block_nums ↗ 512/32

   count_stream.read();

   idx = idx_stream.read();

     for count {

    memcpy(data, ram2[idx * 512], PE * DataType);

       for PE

          sum += data;

    data_out_stream.write(sum);

    }