CUDA memory hierarchy

Grid

Block(0,0)

Block(0,1)

Shared Memory

(~5clocks)

register   register

↕(~1clock)

thread(0,0)   thread(0,1)

Shared Memory

register   register

thread(0,0)   thread(0,1)

(~500clocks)

Global Memory / Local Memory

Host

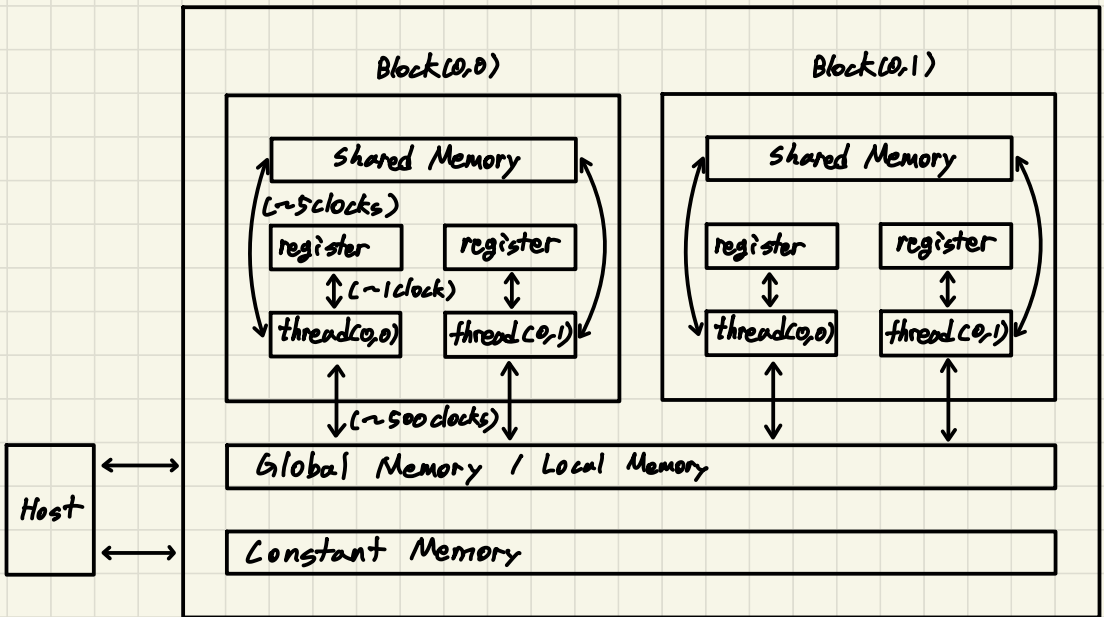Constant Memory

1. max 1024 threads per 1 block
2. 1 block per SM
3. (32 threads = 1 warp) per SM

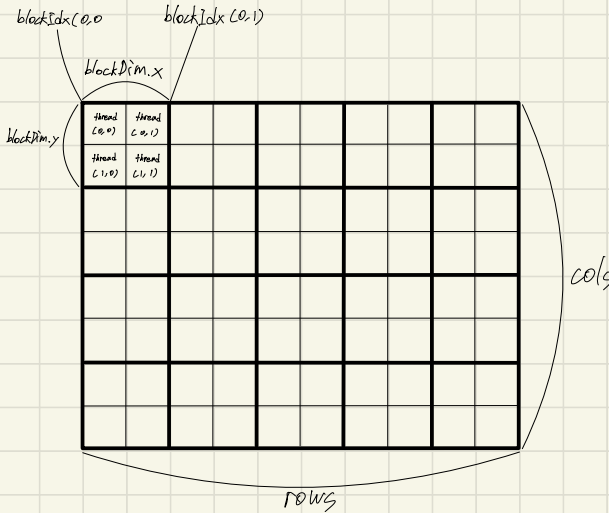# TILED MATRIX DOT PRODUCT Basic

lhs

thread (0,0)	thread (0,1)	thread (0,0)	thread (0,1)	thread (0,0)	thread (0,1)
thread (1,0)	thread (1,1)	thread (1,0)	thread (1,1)	thread (1,0)	thread (1,1)

rhs

thread (0,0)	thread (0,1)						
thread (1,0)	thread (1,1)						
thread (0,0)	thread (0,1)						
thread (1,0)	thread (1,1)						
thread (0,0)	thread (0,1)						
thread (1,0)	thread (1,1)						

blockIdx (0,0)   blockIdx (0,1)

blockDim.x

blockDim.y

thread (0,0)	thread (0,1)						
thread (1,0)	thread (1,1)						

cols

rows

```
pseudo code
int x = blockIdx.x × blockDim.x + threadIdx.x;
int y = blockIdx.y × blockDim.y + threadIdx.y;
double sum = 0.0;
for (int i=0;  i < lhs_cols; ++i)
{
     sum += lhs[y × lhs_cols + i] × rhs [i × rhs_cols + x];
}
result[y × rows + x] = sum;
```

Result matrix dimension 256 x 1
TILED MATRIX DOT PRODUCT

TILED WIDTH = 32

block (0,0)	block (0,1)	block (0,2)	block (0,3)

32 warps   32 warps   32 warps   32 warps

= total 128 warps