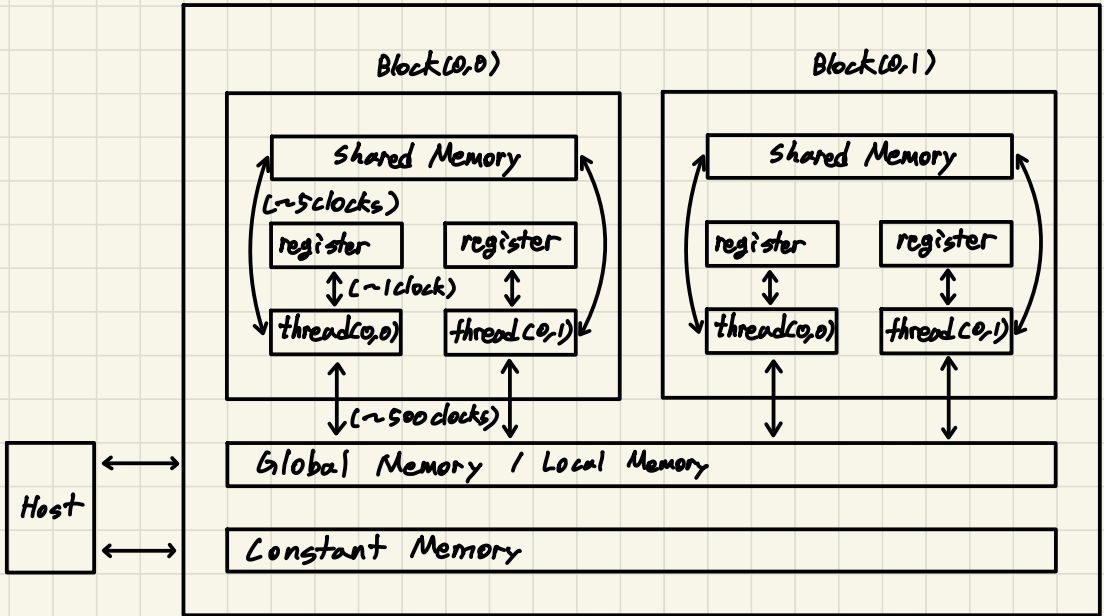


CUDA memory hierarchy

Grid



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

```
--global-- void kernelFunc(float* dst, const float* src) {  
    float p = 0; // register per thread  
    float arr[10]; // local per thread  
    --shared-- float partial_sum[1024]; // shared Memory per Block  
}
```

TILED MATRIX MULTIPLICATION Basic

rhs matrix

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)						

lhs matrix

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)

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

blockDim.x

result matrix

blockDim.y

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

cols

rows

pseudo code

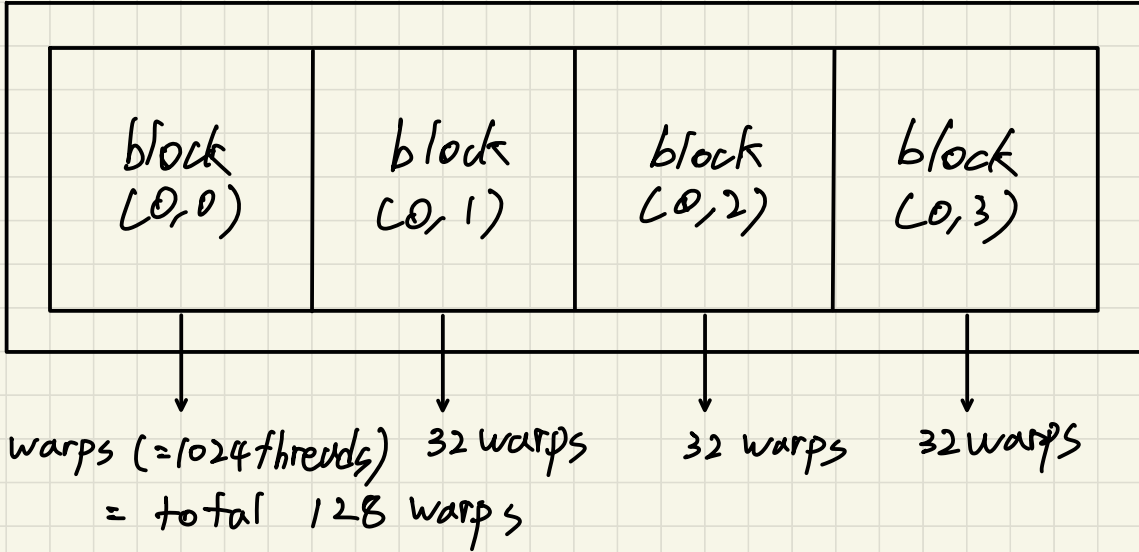
$\text{int } x = \text{blockIdx}.x \times \text{blockDim}.x + \text{threadIdx}.x$

$\text{int } y = \text{blockIdx}.y \times \text{blockDim}.y + \text{threadIdx}.y$

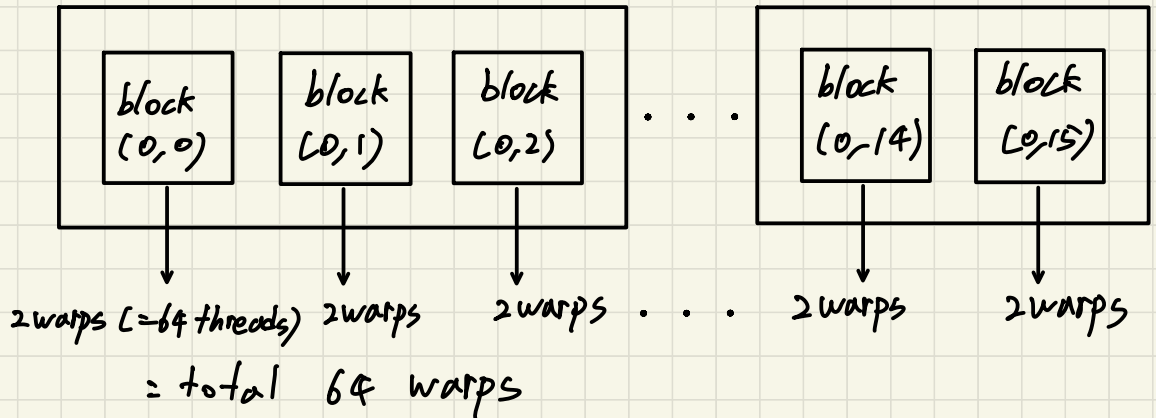
$\text{result}[y, x] = \text{sum} [\text{lhs}[y, :] \times \text{rhs}[:, x]]$

Calculate the number of warp
(result matrix dimension 256 x 1)

TILE WIDTH = 32



TILE WIDTH = 8



TILED MATRIX MULTIPLICATION using shared memory (TILE WIDTH = 2)

lhs_mat (16x32)

M_{00}	M_{01}	M_{02}	M_{03}	M_{04}	M_{05}	M_{06}	M_{07}
M_{10}	M_{11}	M_{12}	M_{13}	M_{14}	M_{15}	M_{16}	M_{17}

step 1
 copy to shared memory
 (using - sync threads)

M_{00}	M_{01}	N_{00}	N_{01}
M_{10}	M_{11}	N_{10}	N_{11}

calculation

$$\begin{aligned}
 P_{00} &\pm M_{00} \times N_{00} + M_{01} \times N_{10} \\
 P_{01} &\pm M_{00} \times N_{01} + M_{01} \times N_{11} \\
 P_{10} &\pm M_{10} \times N_{00} + M_{11} \times N_{10} \\
 P_{11} &\pm M_{10} \times N_{01} + M_{11} \times N_{11}
 \end{aligned}$$

N_{00}	N_{01}			
N_{10}	N_{11}			
N_{20}	N_{21}			
N_{30}	N_{31}			
N_{40}	N_{41}			
N_{50}	N_{51}			
N_{60}	N_{61}			
N_{70}	N_{71}			

rhs_mat (32x32)

P_{00}	P_{01}			
P_{10}	P_{11}			

TILE WIDTH

result mat (16x32)

step 4

copy to shared memory
 (using - sync threads)

M_{06}	M_{07}	N_{10}	N_{11}
M_{16}	M_{17}	N_{10}	N_{11}

calculation

$$\begin{aligned}
 P_{00} &\pm M_{06} \times N_{60} + M_{07} \times N_{70} \\
 P_{01} &\pm M_{06} \times N_{61} + M_{07} \times N_{71} \\
 P_{10} &\pm M_{16} \times N_{60} + M_{17} \times N_{70} \\
 P_{11} &\pm M_{16} \times N_{61} + M_{17} \times N_{71}
 \end{aligned}$$