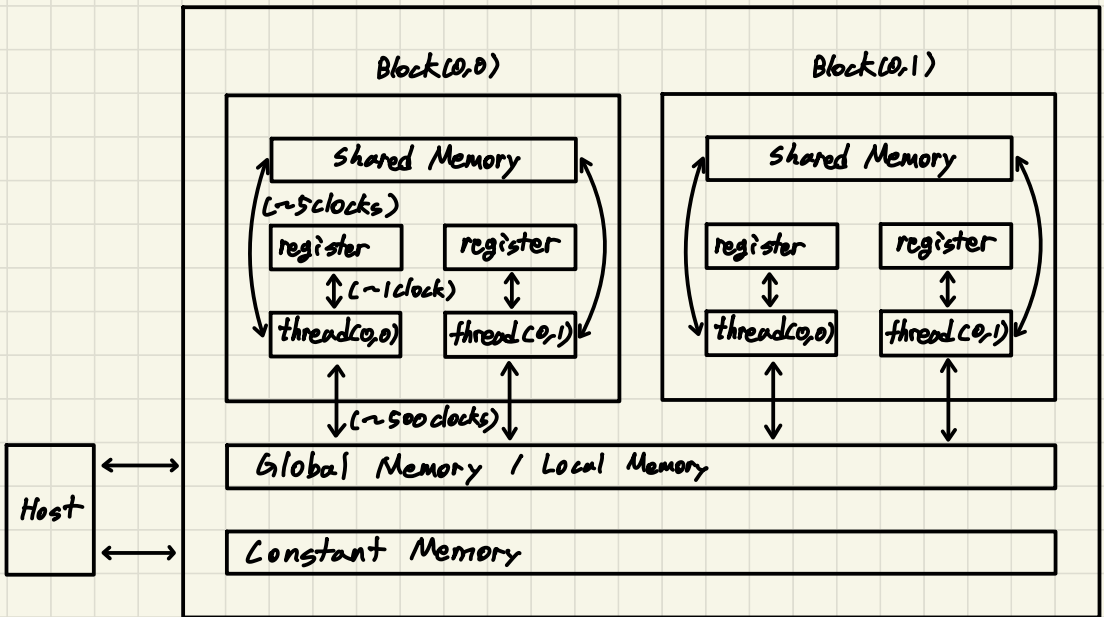


CUDA memory hierarchy

Grid



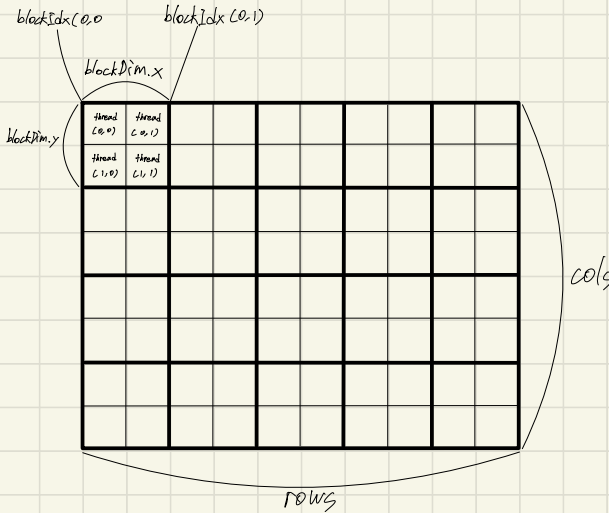
TILED MATRIX DOT PRODUCT

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)								



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;
    
```