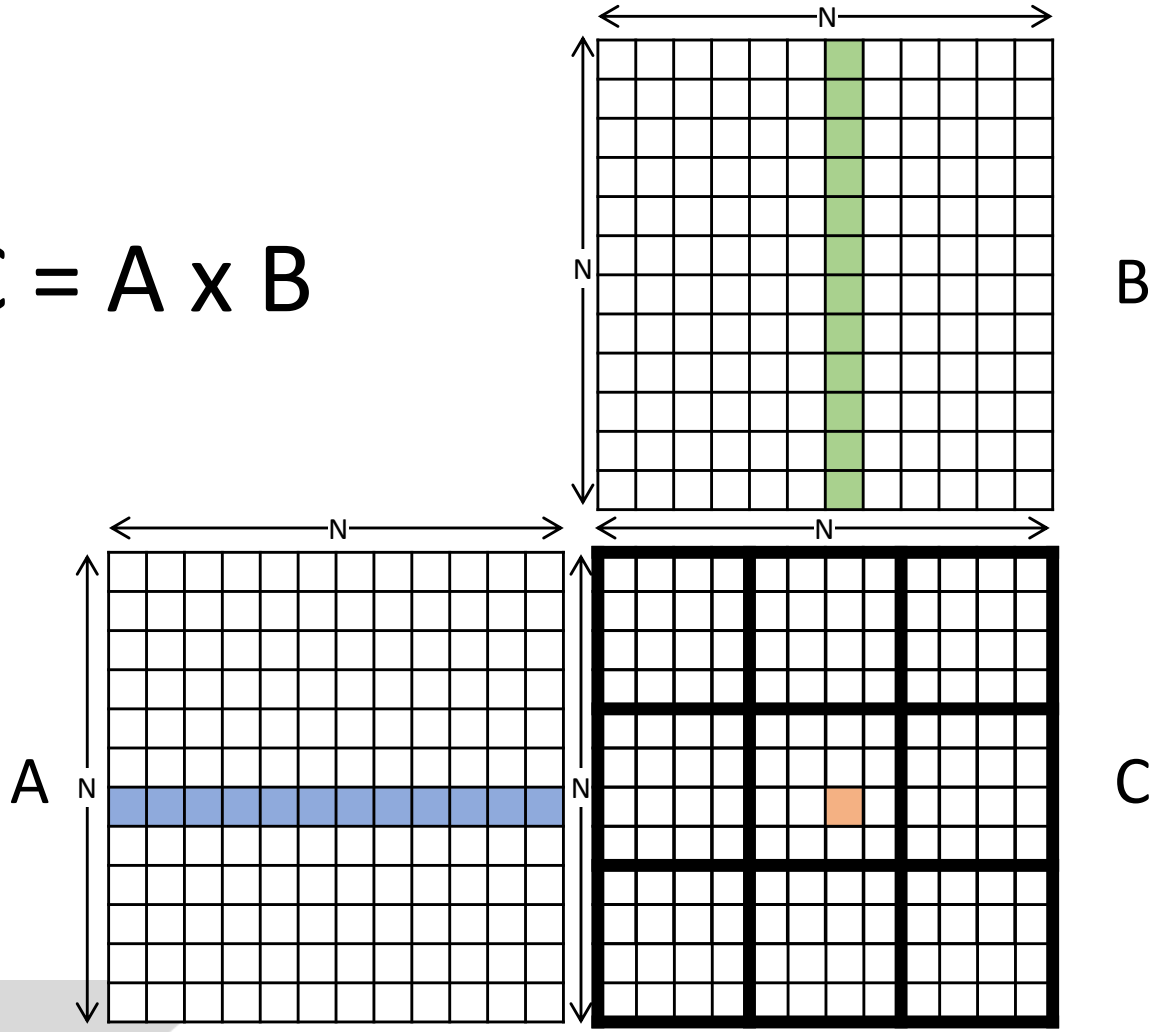
Example: Matrix-Matrix Multiplication

$C = A \times B$



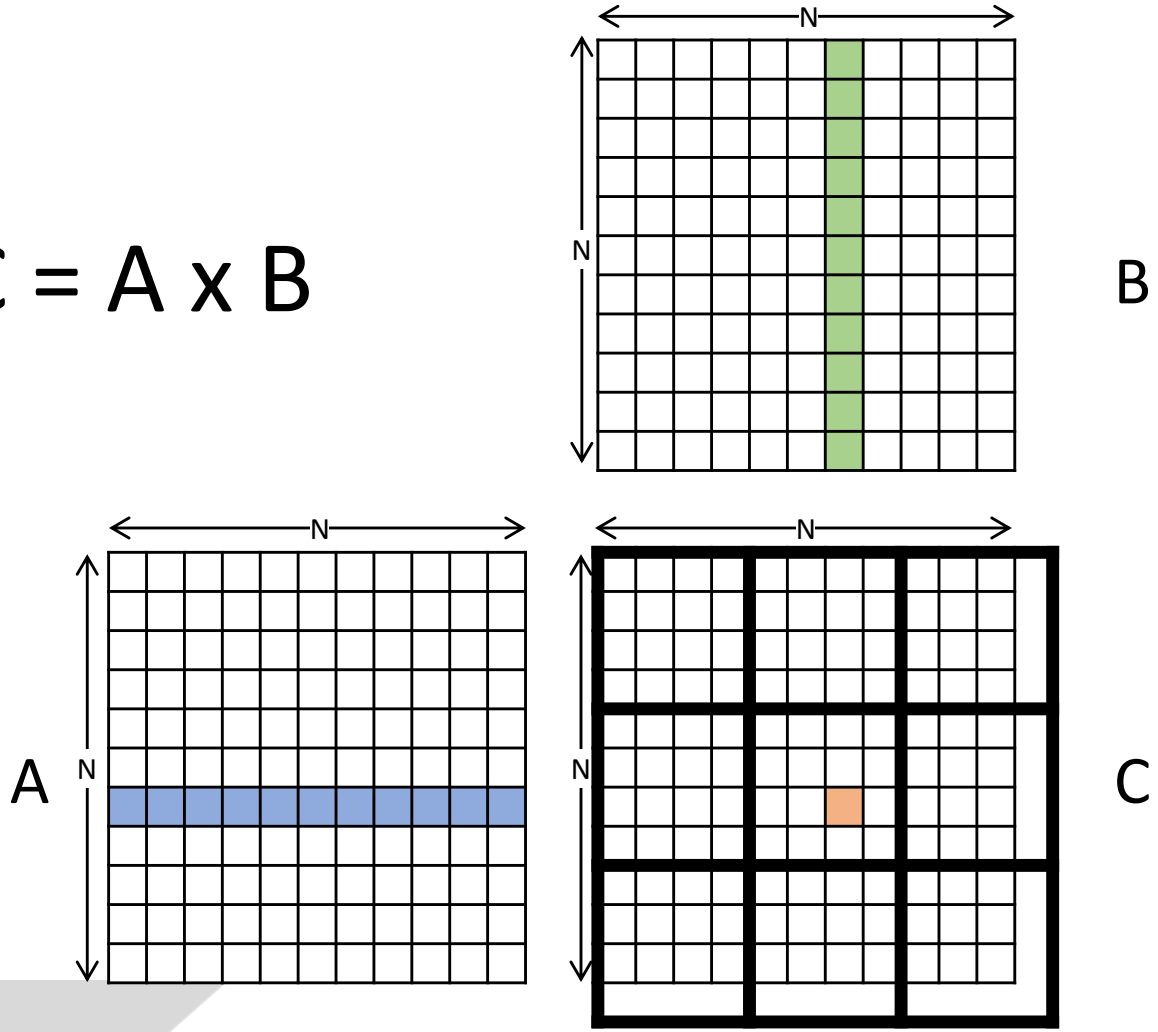
Parallelization approach: assign one thread to each element in the output matrix (C)

Example: Matrix-Matrix Multiplication

```
__global__ void mm_kernel(float* A, float* B, float* C, unsigned int N) {

    unsigned int row = blockIdx.y*blockDim.y + threadIdx.y;
    unsigned int col = blockIdx.x*blockDim.x + threadIdx.x;

    float sum = 0.0f;
    for(unsigned int i = 0; i < N; ++i) {
        sum += A[row*N + i]*B[i*N + col];
    }
    C[row*N + col] = sum;

}
```

Boundary Conditions

$$C = A \times B$$



Different Matrix Dimensions

$C = A \times B$