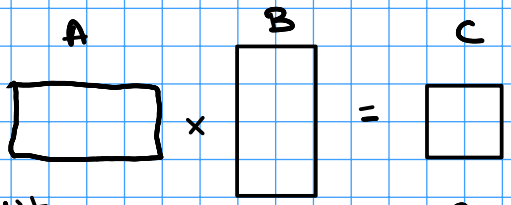
DISTRIBUTED MATMUL



1) Split computations among GPU

Take C
4 GPU

 → 4 blocks

2) Divide in threads = # GPU = # Blocks

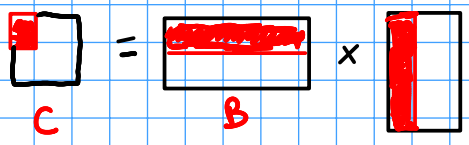
- Each thread will take care of 1 Block

FROM NOW PARALLEL

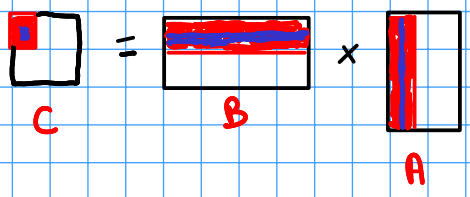
3) Check free memory in the GPU

Divide each block in CHUNKS

example: the RED PART of C to be computed needs the RED parts of B and A



If all the red part (FROM ABC) cannot fit all at once in the GPU
we have to split it, FOR THIS REASON → CHUNKS



In Blue a chunk is represented

In general a chunk does not have to contain Full rows of A and B

4) COMPUTE MATMUL FOR EACH CHUNK

for CHUNK in Block {

- send chunk to GPU
- COMPUTE
- Get the result
- FREE GPU MEM

}