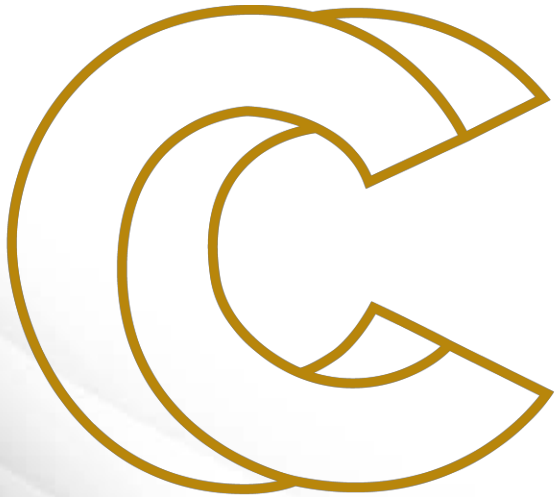
TÜBİTAK

Sabancı Üniversitesi



EURO^2

Performance Engineering on CPUs and GPUs:

- GPUs: Things to be Careful for Performance -

Kamer Kaya, Sabancı University

ncc@ulakbim.gov.tr

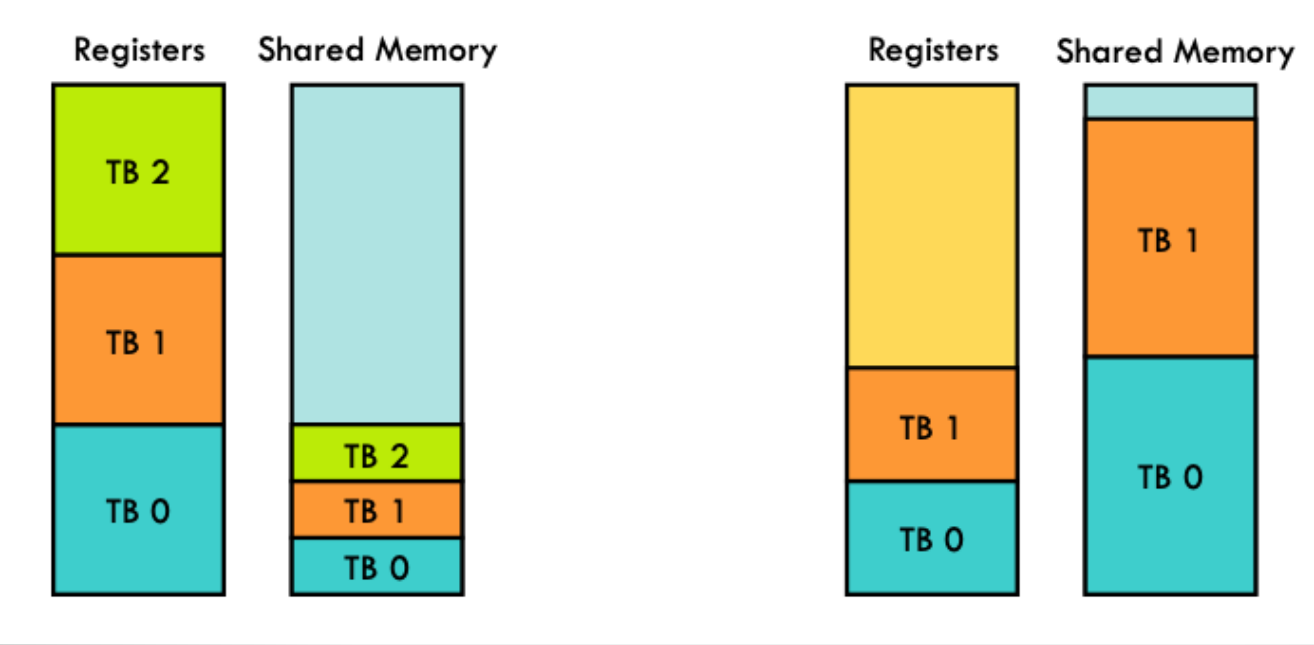
Things to be Careful for Performance: (4) Occupancy

EURO²

- What determines occupancy?
- Limited resources
  - Register usage per thread.
  - Shared memory per thread block.

Things to be Careful for Performance: (4) Occupancy

```
Total amount of shared memory per block:       49152 bytes
Total shared memory per multiprocessor:        98304 bytes
Total number of registers available per block: 65536
Warp size:                                     32
Maximum number of threads per multiprocessor:  2048
Maximum number of threads per block:           1024
```

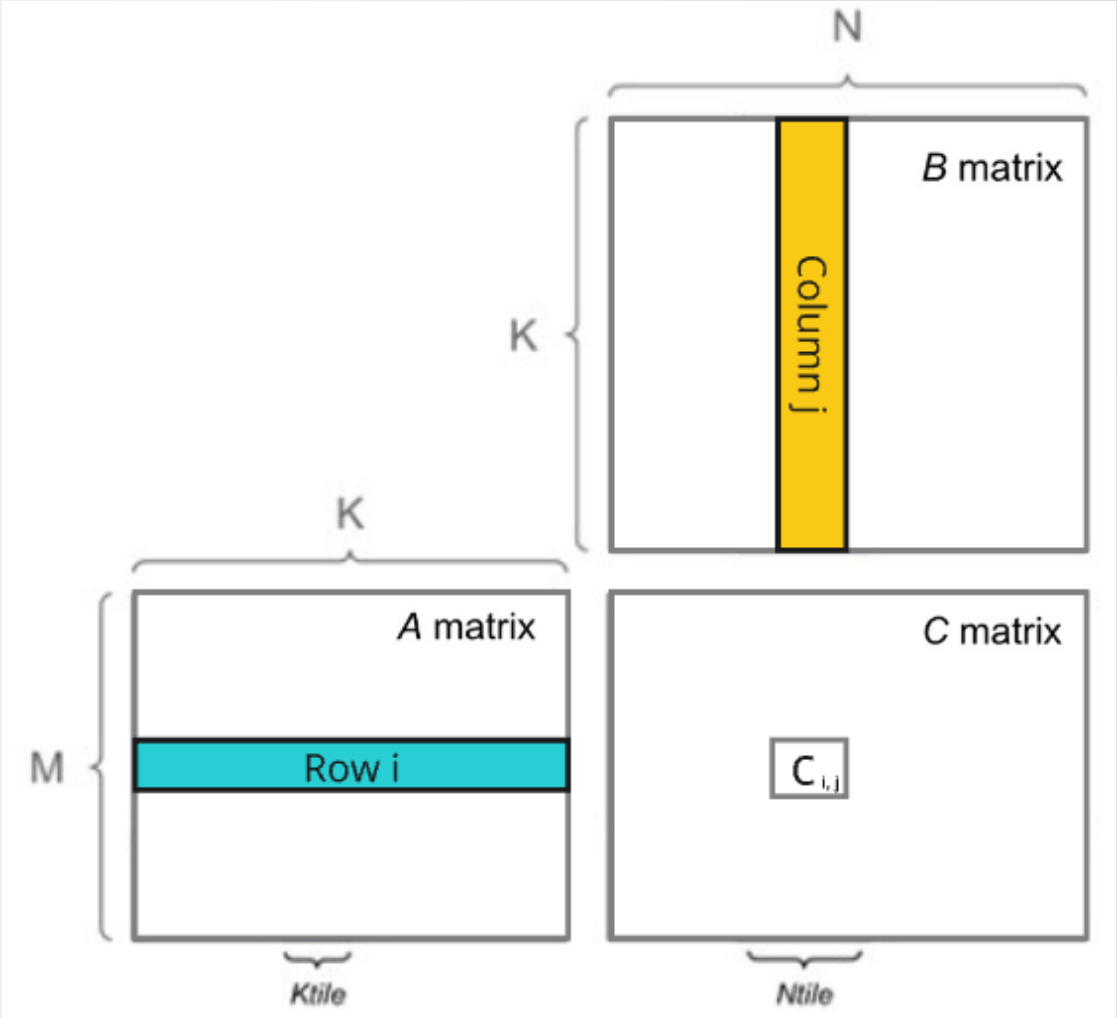


Summary: Matrix Multiplication

EURO²

$$C = A \times B$$

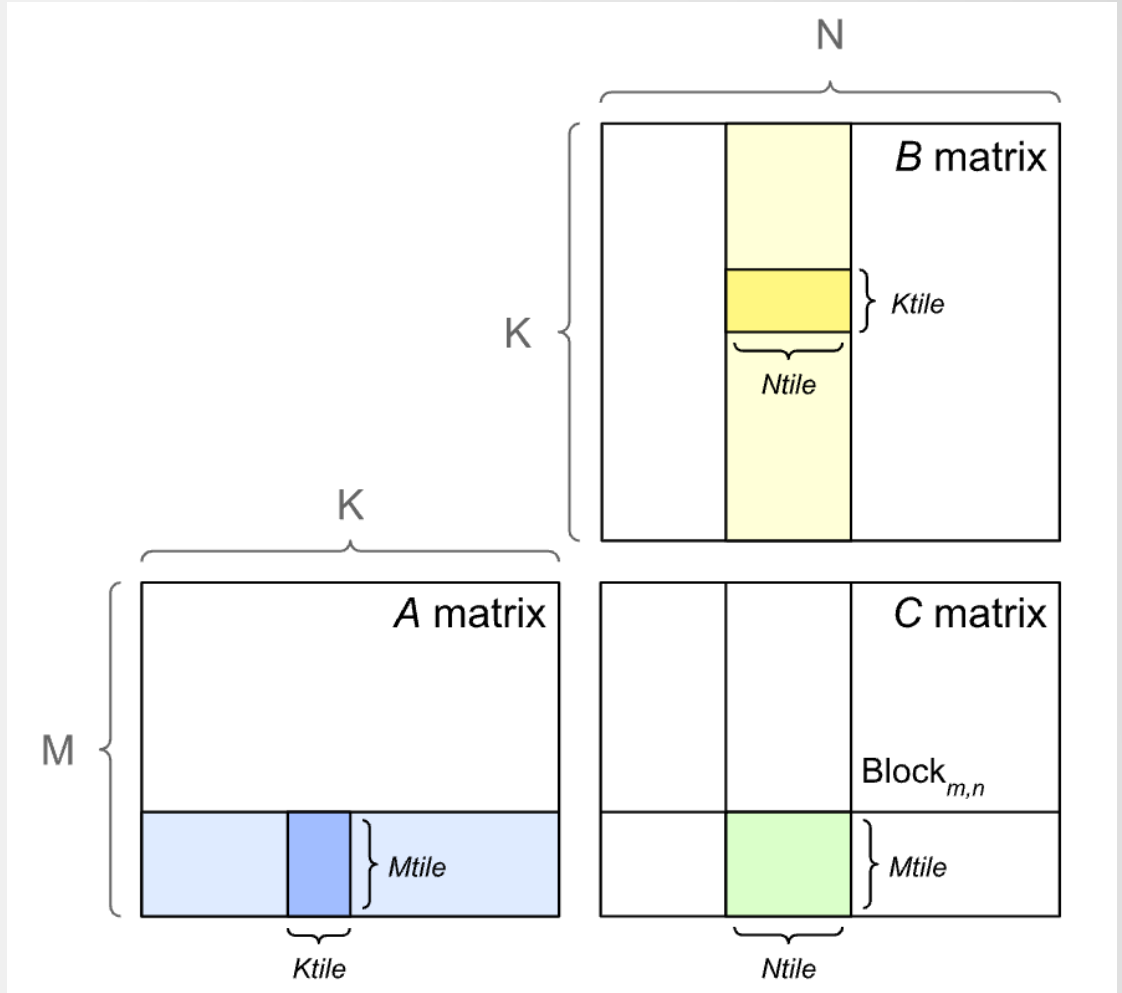
Trivial approach: Assign each entry to a single thread. Each thread will perform a dot product.



Summary: Matrix Multiplication

Another approach with shared memory use: Assign a **C** tile to a single block.

- The **A** and **B** tiles are brought to shared memory
- Computation will be performed with these tiles.
- The block will process the next tiles from **A** and **B**.



Summary: Matrix Multiplication

Let’s check the code.

Thanks

This project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 101101903. The JU receives support from the Digital Europe Programme and Germany, Bulgaria, Austria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, Greece, Hungary, Ireland, Italy, Lithuania, Latvia, Poland, Portugal, Romania, Slovenia, Spain, Sweden, France, Netherlands, Belgium, Luxembourg, Slovakia, Norway, Türkiye, Republic of North Macedonia, Iceland, Montenegro, Serbia