# Memory Hierarchies and Matrix-Matrix Multiplication

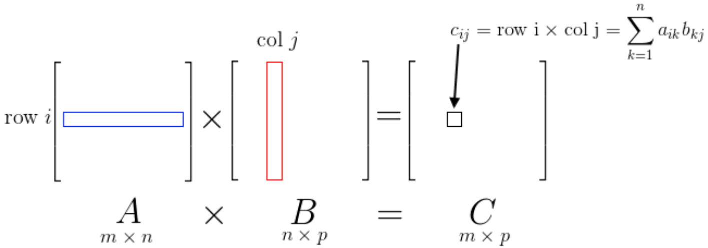
Pratyuksh Bansal

Università della Svizzera italiana

September 22, 2021

# What is Matrix multiplication?

If $A$ is $m \times n$ matrix and $B$ is $n \times p$ matrix, then $C = AB$ is $m \times p$ matrix.


col $j$
$c_{ij} = \text{row i} \times \text{col j} = \sum_{k=1}^{n} a_{ik} b_{kj}$
row $i$
$\times$
$=$
$\underset{m \times n}{A} \times \underset{n \times p}{B} = \underset{m \times p}{C}$

Also known as **GEMM** - General matrix multiplication.

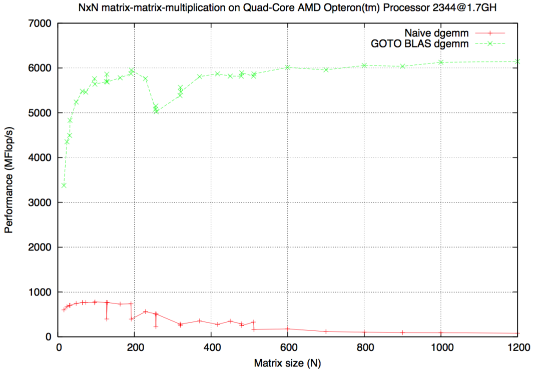
# Naive algorithm for GEMM

**Algorithm 1** Matrix multiplication

```
for $i = 1$ to $m$ do
    for $j = 1$ to $p$ do
        for $k = 1$ to $n$ do
            $C(i,j) = \sum_{k=1}^{n} A(i,k) * B(k,j)$
        end for
    end for
end for
```

Question: Is this the most optimal way to do it?

# Performance



Theoretical Peak Performance (TPP) :
$(1.7 \times 10^9 cycles/sec) \times (4 flops/cycle) = 6.8$ GFlops/sec.

More than $50\times$ difference in performance!!

# Memory hierarchy

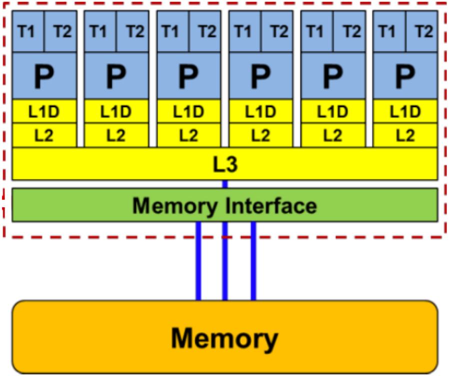


Figure: Memory hierarchy of a multi-core architecture

# Why is it important?

- Basic linear algebra operation, appears in several applications in physics, engineering, etc.
- Benchmark to compare the performance of processors.

# Why is it important?

- Basic linear algebra operation, appears in several applications in physics, engineering, etc.
- Benchmark to compare the performance of processors.

Learning objective :

Fundamental concepts and ideas used to optimize GEMM.

Thank you