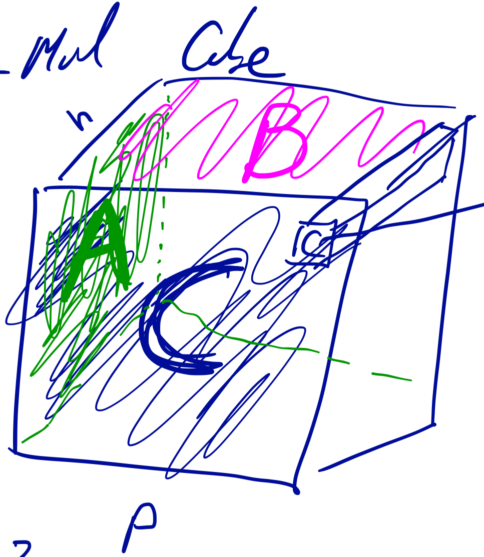
Lecture 15

The MatMul Cube

$A: m \times n$

$B: n \times p$



$c_{ij} = \sum_k a_{ik} b_{kj}$

Why the cube?

It elucidates memory locality

Why does this matter?

#1. Memory locality
+ Memory Hierarchy

#2. Vector Units
on CPUs

Vector Units
are specialized
processing units on
CPUs to do
multiple float ops
fast!

A common vector
op:

$$\begin{bmatrix}a_1\\a_2\\a_3\\a_4\end{bmatrix} + \begin{bmatrix}b_1\\b_2\\b_3\\b_4\end{bmatrix}$$

a vector
unit can
do all
4 @ once!

In Julia

```
@simd for ...

end
```

it will try and
use the vector unit.

```
function dot(a,b)
@simd for i = 1 : length(a)
        c += a[i]*b[i]
End
```

Another View on MatMul

SxS block

$$A = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix} \qquad B = \begin{bmatrix} \; & \; & \; \\ \; & \; & \; \\ \; & \; & \; \end{bmatrix}$$

"blocked Matrix partition"

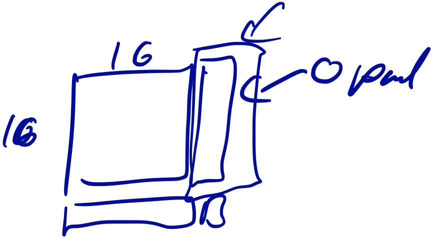
Then $C_{21} \leftarrow$ SxS block of C

$$= \underbrace{\boxed{A_{21} B_{11}}}_{\text{SxS - MatMul}} + A_{22} B_{21}$$

(16x16)
256

$$\underset{\uparrow \atop \text{SxS Matrix}}{C_{ij}} = \sum_{k=1}^{\text{"p"}} \underbrace{A_{ik} B_{kj}}_{\text{SxS MatMul}}$$

$+A_{23}B_{31}$

This is called the Blocked Approach