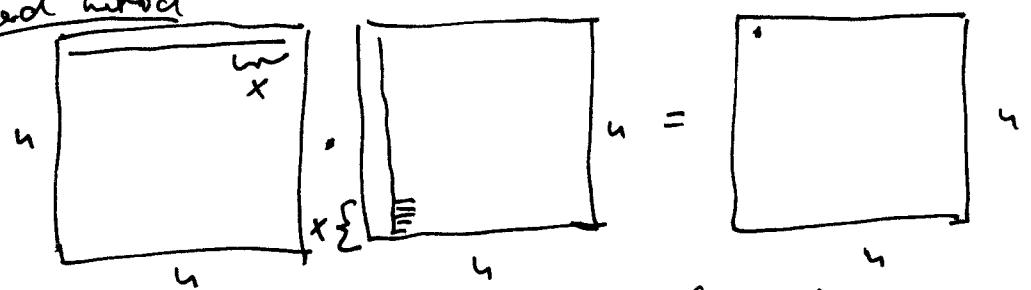
Why blocking? assume cache size = $C \ll n^2$, cache line = 8 doubles

1.) standard method



1. entry: $\frac{n}{8} + n$ CM (compulsory)

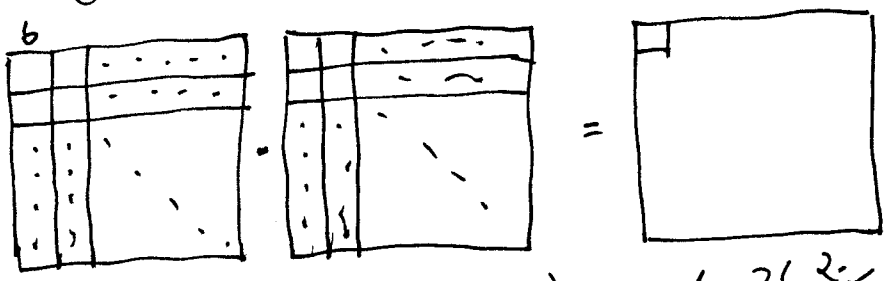
afterwards (x) in cache

2. entry: no reuse, so again $\frac{n}{8} + n$ CM

- - - - -

$\Rightarrow$ total $\approx n^2\left(\frac{n}{8} + n\right) = \frac{9}{8} n^3$ CM's

2.) blocking



choose: $b \geq 8$ (cache line size) and $8 \mid b$ and $3b^2 \leq C \Leftrightarrow b \leq \sqrt{\frac{C}{3}}$

(3 blocks fit into cache)

1. block: $\left(\frac{b^2}{8} + \frac{b^2}{8}\right)\frac{n}{b} = \frac{nb}{4}$

2. block: assume the same

$\Rightarrow$ total $\approx \left(\frac{n}{b}\right)^2 \cdot \frac{nb}{4} = \frac{n^3}{4b}$

choose max $b = \sqrt{\frac{C}{3}}$: $\approx \frac{\sqrt{3}}{4\sqrt{C}} \cdot n^3$ CM's

Blocking exploits locality of mat-mat-mult:

$O(n^2)$ data, $O(n^3)$ computations

$\Rightarrow O(n)$ comp. / data