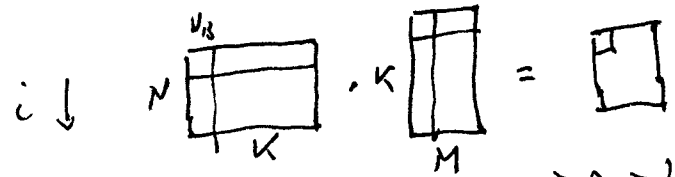
Model-based ATLAS (Yotov et al.)

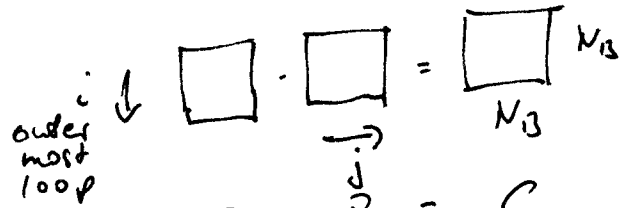
parameters used include

C_1	L1 cache size
B_1	cache line size
C_I	I-cache size
L_x	Fl.p. latency

1.) Estimating N_{13} (assuming L1-cache fully associative)
refine model step by step



blocked into mini-MMM's:



$A \cdot B = C$

a.) working set $3N_{13}^2$

$\Rightarrow 3N_{13}^2 \leq C_1$

b.) closer analysis

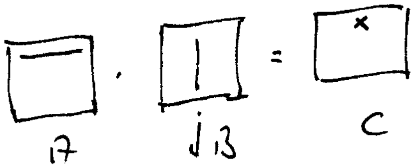
$N_{13}^2 + N_{13} + 1 \leq C_1$

↑ B; ↑ row of A; ↑ element of C

c.) take into account cache line size
units are cache line

$$\left\lceil \frac{N_{13}^2}{B_1} \right\rceil + \left\lceil \frac{N_{13}}{B_1} \right\rceil + 1 \leq \frac{C_1}{B_1}$$

d.) take into account LRU replacement



touches

$A_{00}, B_{0j}, A_{01}, B_{1j}, \ldots, A_{0 N_B-1}\, B_{N_B-1\, j}, C_{0j}$

⟶ more recent

for all j: (computing 1. row of C)
(i = 0)

← want to keep for i = 1

$A_{00}\, B_{00}, \ldots, A_{0N_B-1}\, B_{N_B-1\, 0}, C_{00},$

$A_{00}\, B_{01}, \ldots, A_{0N_B-1}\, B_{N_B-1\, 1}, C_{01},$

$\ldots$

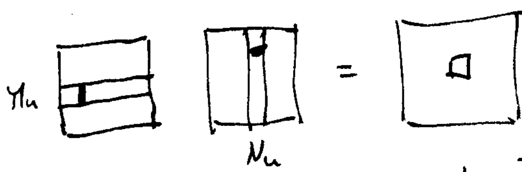
$A_{00}\, B_{0N_B-1}, \ldots, A_{0N_B-1}\, B_{N_B-1\, N_B-1}, C_{0N_B-1}$

in the next iteration (i = 1) we want to reuse B

$\Rightarrow$ for i = 1 also $\underbrace{A_{0*}}_{1.\text{row } A}$ and $\underbrace{C_{0*}}_{1.\text{row } C}$ have to fit (i.e., entire B, 2 rows of A, 1 row of C, 1 elem. of C)

$$\Rightarrow \left\lceil \frac{N_B^2}{B_1} \right\rceil + 3\left\lceil \frac{N_B}{B_1} \right\rceil + 1 \leq \frac{C_1}{B_1}$$

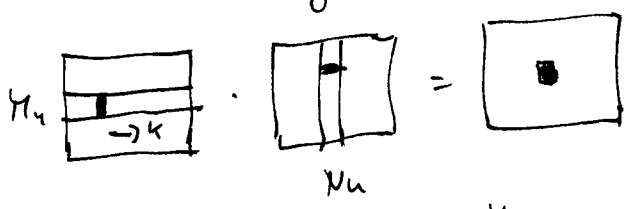
e.) take into account register tiling



$$\Rightarrow \left\lceil \frac{N_B^2}{B_1} \right\rceil + 3\left\lceil \frac{N_B \cdot M_u}{B_1} \right\rceil + \left\lceil \frac{M_u \cdot N_u}{B_1} \right\rceil \leq C_1$$

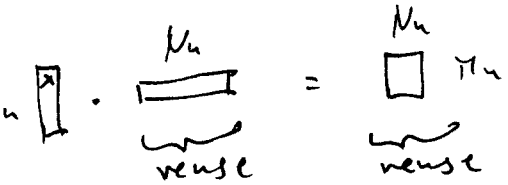
f.) make sure $M_u, N_u \mid N_B$ (no clean-up code)

2.) Estimating M_u and N_u



(kji - order in micro-MMM)

a.) micro MMM



$$\Rightarrow M_u N_u + N_u + 1 \leq N_{12}$$

b.) scalar replacement: loads entire [] and ═

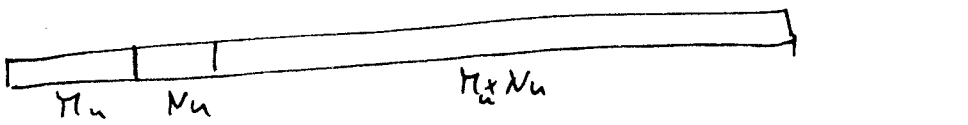
$$\Rightarrow M_u N_u + M_u + N_u \leq N_{12}$$

c.) skewing

$L_s = 1$: mul$_1$, add$_1$, mul$_2$, add$_2$, ⋮

L_s: mul$_1$, ⋮, mul$_{L_s}$ } extra $L_s - 1$ tmp. variables; add$_1$, mul$_{L_s+1}$

$$\Rightarrow M_u N_u + M_u + N_u + L_s \leq N_{12}$$



3.) Estimating K_u

- choose such that fits into I-cache
- K_u / N_{13} (avoid clean-up code)

4.) Estimating L_s (assuming one FP unit)

$$\Rightarrow 2L_s - 1 \geq L_x$$

$$\Rightarrow L_s = \left\lceil \frac{L_x + 1}{2} \right\rceil$$

mul$_1$, ⋮, mul$_{L_s}$, add$_1$, mul$_{L_s+1}$, add$_2$, ⋮ — $2L_s - 1$ instructions in between