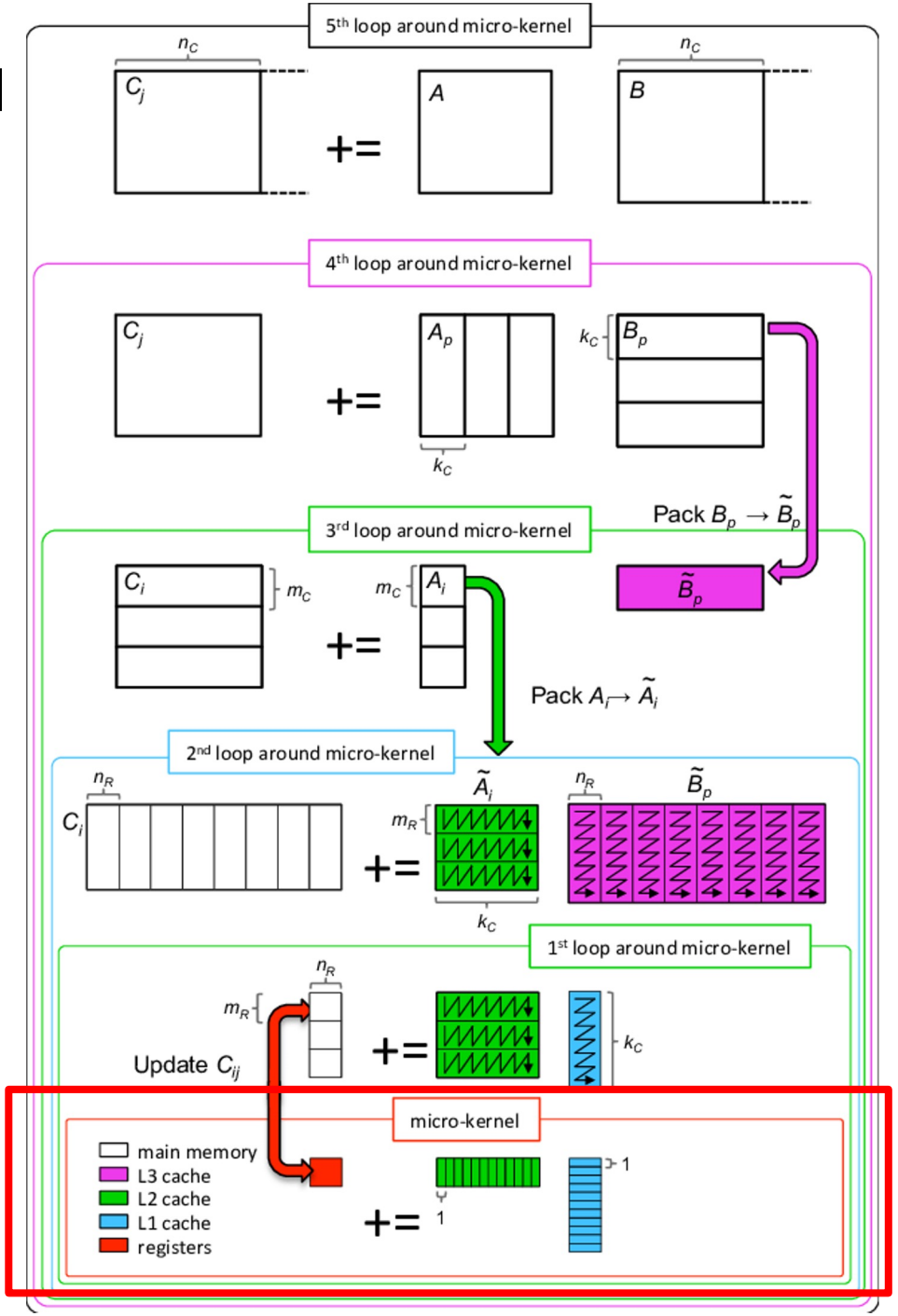
# Burst and Packing Analysis

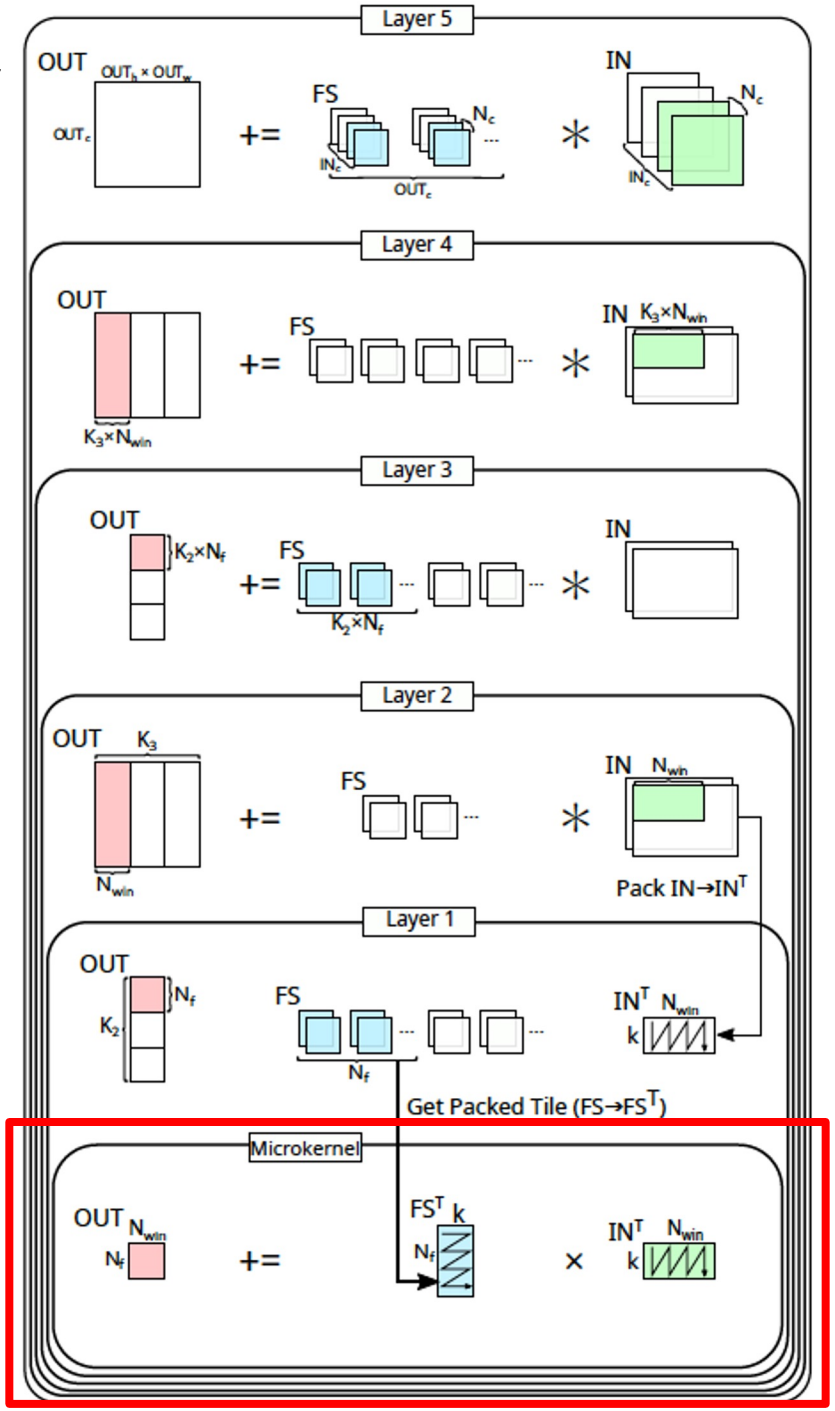
IME Task Group

**Guido Araujo**

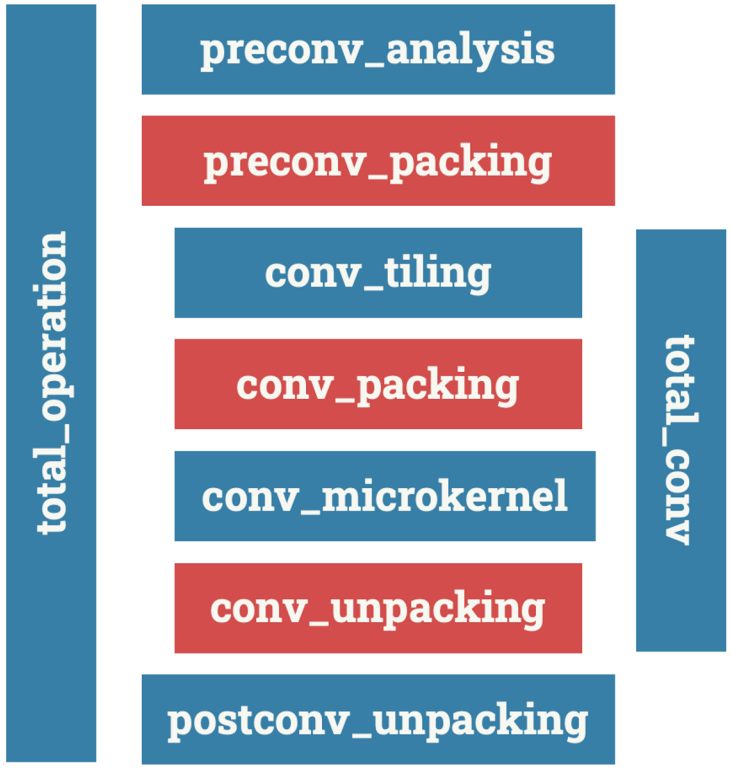
GEMM

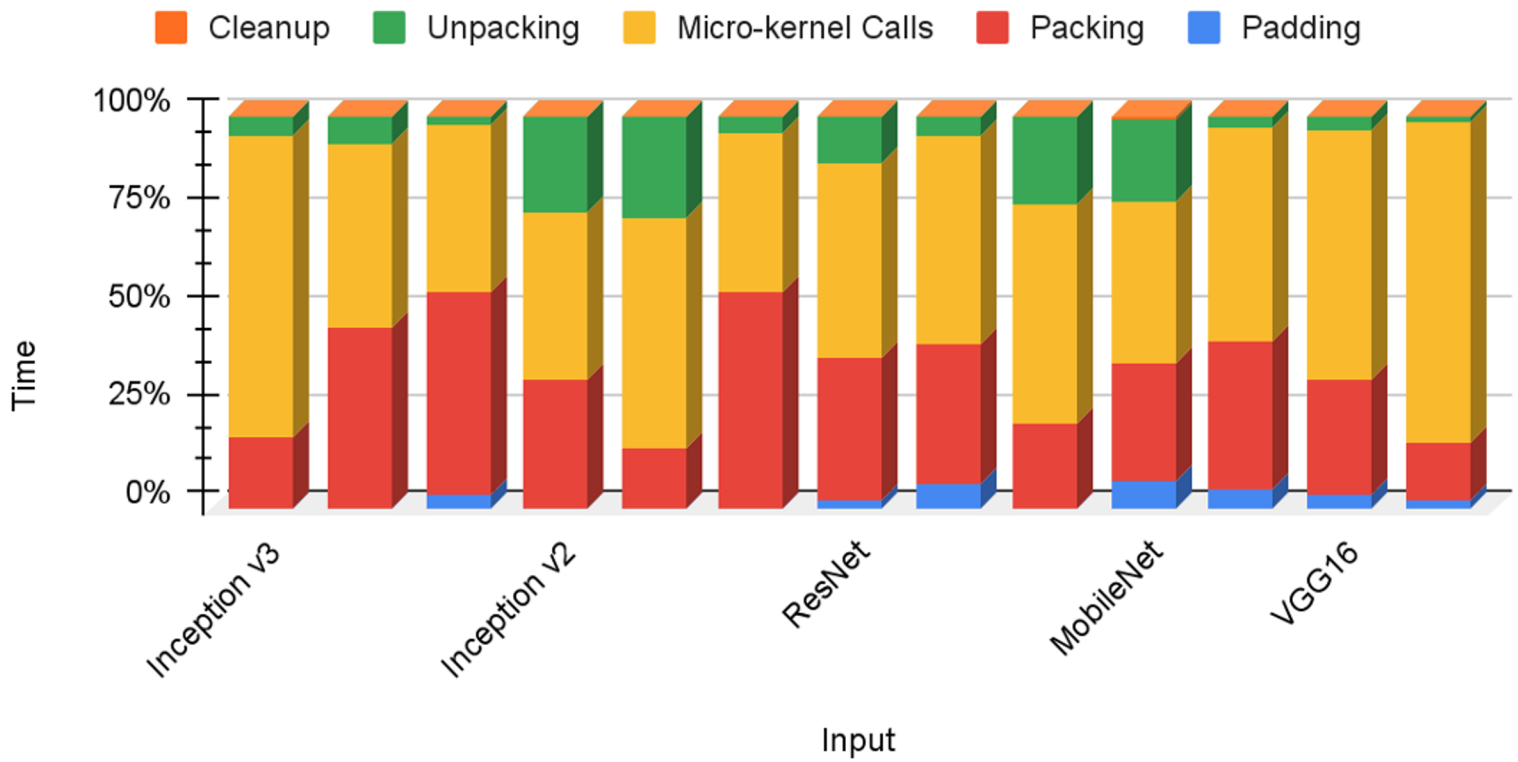

5th loop around micro-kernel
4th loop around micro-kernel
3rd loop around micro-kernel
2nd loop around micro-kernel
1st loop around micro-kernel
micro-kernel
Pack $B_p \rightarrow \tilde{B}_p$
Pack $A_i \rightarrow \tilde{A}_i$
Update $C_{ij}$
main memory
L3 cache
L2 cache
L1 cache
registers

CONV

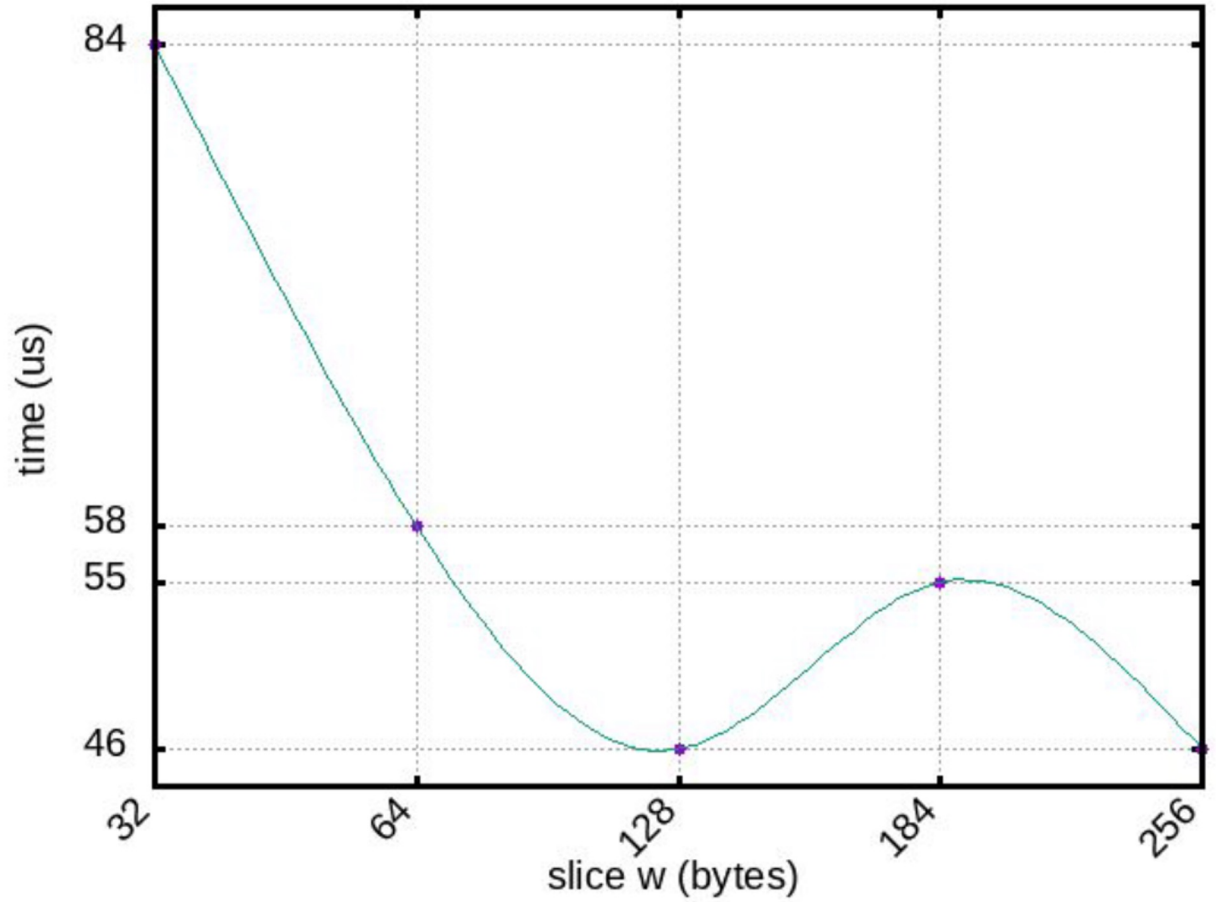

Layer 5
Layer 4
Layer 3
Layer 2
Layer 1
Pack IN→IN^T
Get Packed Tile (FS→FS^T)
Microkernel

# Locality is important for packing



# Tiling and memory burst

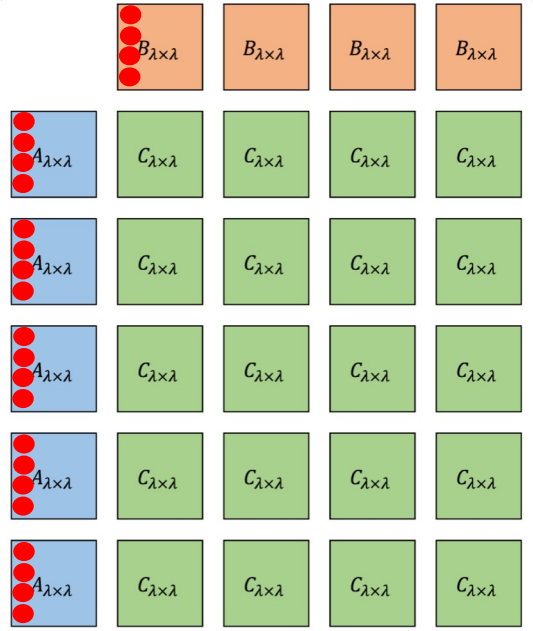


# Tiling and Memory Bursts Analysis

- Assumptions
  - Registers containing L words
    - $L = 16$ words (VLEN=512)
  - Last-level cache blocks of $w$ words
    - $w$ = 16 words (64 bytes)

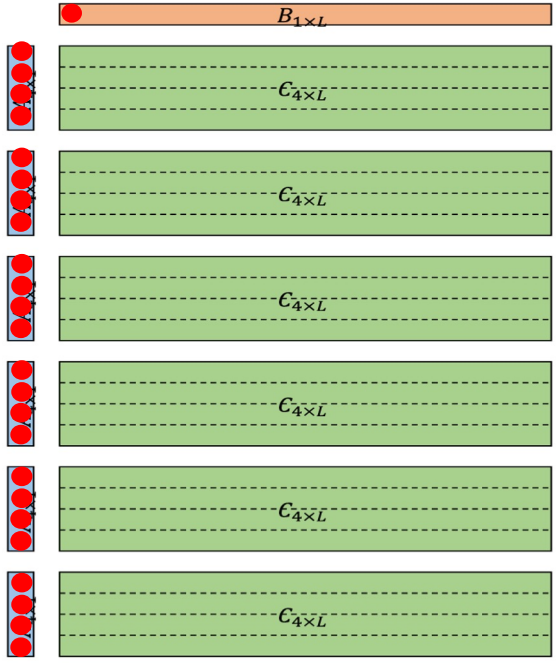
# Tiling and Memory Bursts: Option A

- Option A: One matrix per vector register
  - $\lambda = \sqrt{L}$
  - Configuration: $SRC_A = 5, SRC_B = 4,\ ACC_C = 20$
- Bursts
  - A: $SRC_A \cdot \lambda$
  - B: $\lambda$
  - For this configuration:
    - $6\lambda = 6\sqrt{L} =$ 24 bursts



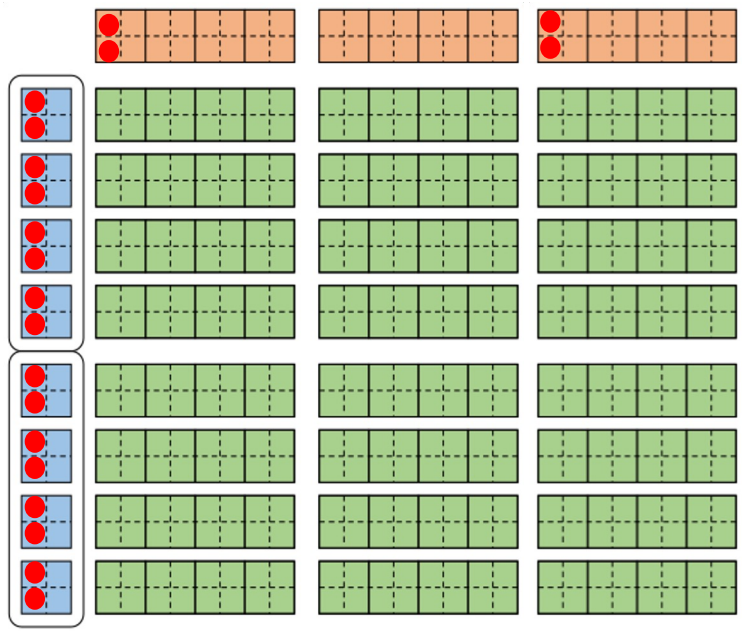
# Tiling and Memory Bursts: Option B

- Option B: One matrix in multiple register vectors
  - Configuration: $SRC_A = 6, SRC_B = 1, ACC_C = 24$
- Bursts
  - A: $ACC_C$
  - B: 1
  - For this configuration
    - 25 bursts



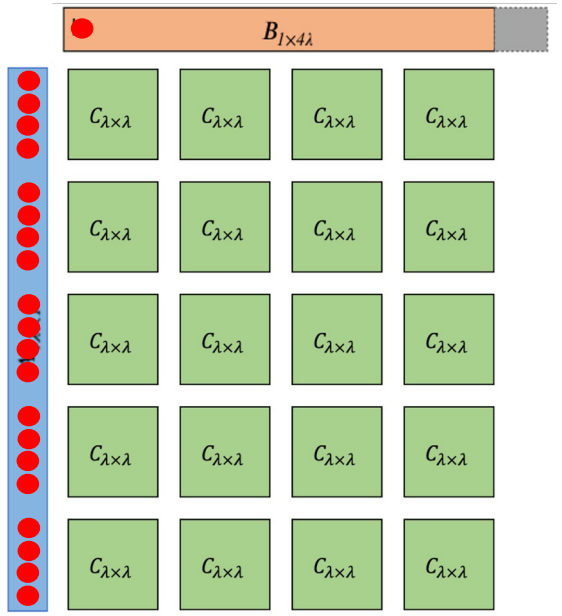
# Tiling and Memory Bursts: Option C

- Option C: Multiple matrices in one vector register
  - Configuration: $SRC_A = 2, SRC_B = 3, ACC_C = 24$
  - $\lambda = 2$
  - Issue: Rows of B may be longer than cache line
- Bursts
  - A: $\frac{L \cdot SRC_A}{\lambda}$
  - B: $\left\lceil \frac{SRC_B \cdot L}{\lambda \cdot w} \right\rceil \cdot \lambda$
  - For this configuration
    - 20 bursts



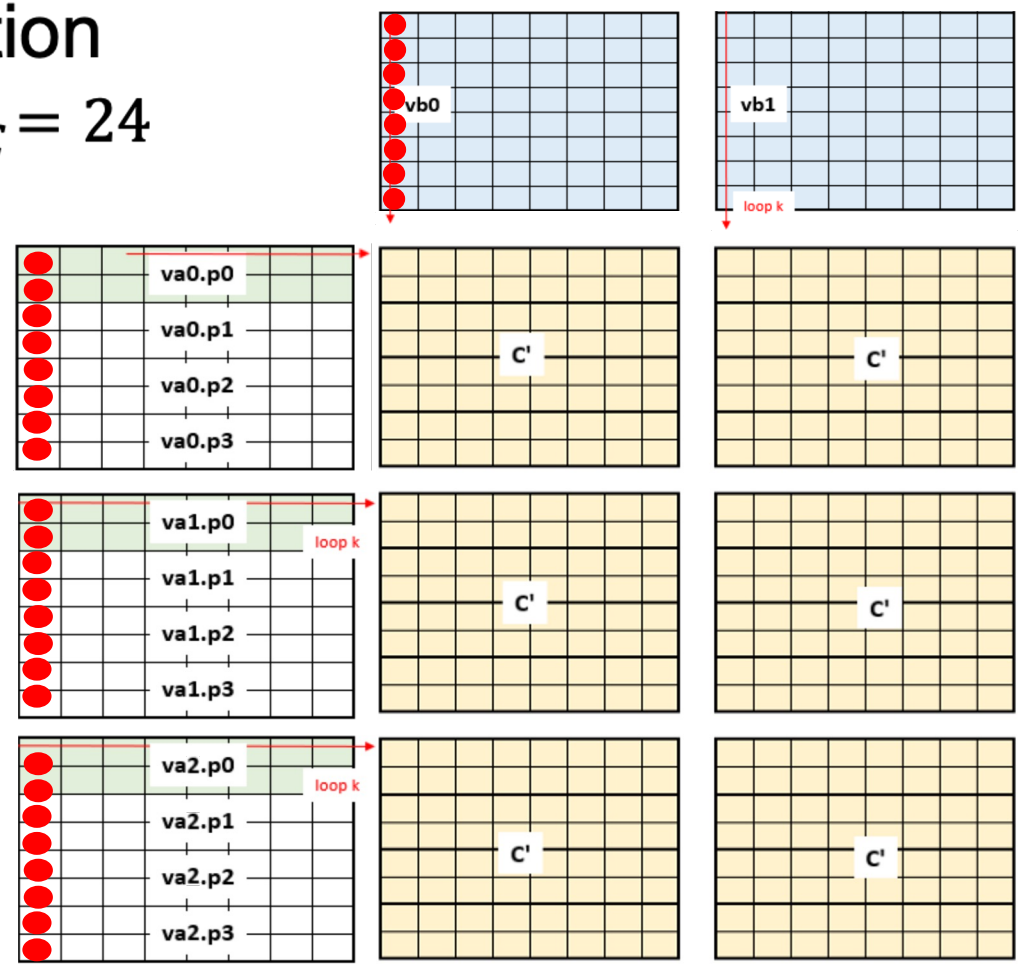
# Tiling and Memory Bursts: Option D

- Option D: Streaming buffers for the input matrices
  - $\lambda = \sqrt{L}$
  - Configuration: $SRC_A = 5, SRC_B = 2, ACC_C = 24$
- Bursts
  - A: $SRC_A\lambda$
  - B: 1
  - For this configuration
    - $5\lambda + 1 = 21$ bursts



# Tiling and Memory Bursts: Option E

- Option E: Variable matrix representation
  - Configuration: $SRC_A = 3, SRC_B = 2, ACC_C = 24$
  - $L_{min} = 4$
- Bursts
  - A: $SRC_A \cdot \frac{L}{\sqrt{L_{min}}}$
  - B: $\sqrt{L_{min}}$
  - For this configuration
    - 26 bursts



# Evaluating impact of bursts