# Integrated Matrix Extension (IME)

TG Kick-off Meeting

**Guido Araujo**
**Jose Moreira**

02/12/24

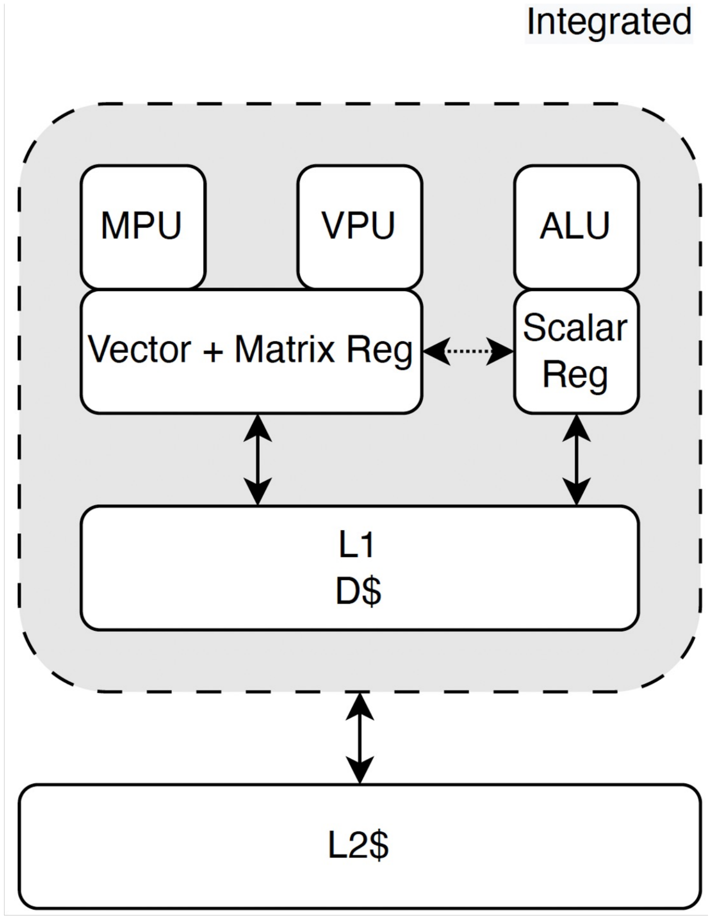
# Agenda

- IME preliminary proposals
- Qualitative vs Quantitative approaches
- Metrics and workloads

# Agenda

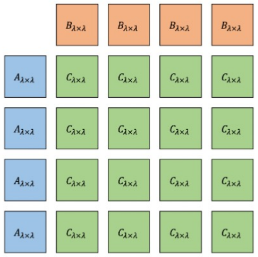
- IME preliminary proposals
- Qualitative vs Quantitative approaches
- Metrics and workloads

# IME - Integrated

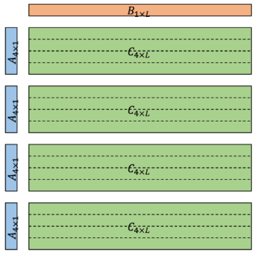


# IME preliminary proposals

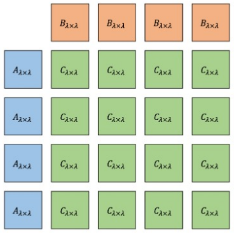
- Option A
  - One matrix per register vector


$B_{\lambda\times\lambda}$
$B_{\lambda\times\lambda}$
$B_{\lambda\times\lambda}$
$B_{\lambda\times\lambda}$
$A_{\lambda\times\lambda}$
$C_{\lambda\times\lambda}$
$C_{\lambda\times\lambda}$
$C_{\lambda\times\lambda}$
$C_{\lambda\times\lambda}$
$A_{\lambda\times\lambda}$
$C_{\lambda\times\lambda}$
$C_{\lambda\times\lambda}$
$C_{\lambda\times\lambda}$
$C_{\lambda\times\lambda}$
$A_{\lambda\times\lambda}$
$C_{\lambda\times\lambda}$
$C_{\lambda\times\lambda}$
$C_{\lambda\times\lambda}$
$C_{\lambda\times\lambda}$
$A_{\lambda\times\lambda}$
$C_{\lambda\times\lambda}$
$C_{\lambda\times\lambda}$
$C_{\lambda\times\lambda}$
$C_{\lambda\times\lambda}$

- Option B
  - One matrix in multiple register vectors

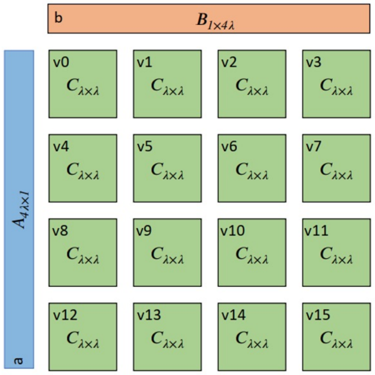


- Option C
  - Multiple matrices in one register vector

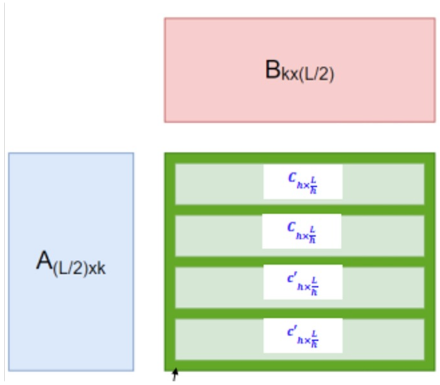

$B_{\lambda\times\lambda}$
$B_{\lambda\times\lambda}$
$B_{\lambda\times\lambda}$
$B_{\lambda\times\lambda}$
$A_{\lambda\times\lambda}$
$C_{\lambda\times\lambda}$
$C_{\lambda\times\lambda}$
$C_{\lambda\times\lambda}$
$C_{\lambda\times\lambda}$
$A_{\lambda\times\lambda}$
$C_{\lambda\times\lambda}$
$C_{\lambda\times\lambda}$
$C_{\lambda\times\lambda}$
$C_{\lambda\times\lambda}$
$A_{\lambda\times\lambda}$
$C_{\lambda\times\lambda}$
$C_{\lambda\times\lambda}$
$C_{\lambda\times\lambda}$
$C_{\lambda\times\lambda}$
$A_{\lambda\times\lambda}$
$C_{\lambda\times\lambda}$
$C_{\lambda\times\lambda}$
$C_{\lambda\times\lambda}$
$C_{\lambda\times\lambda}$

# IME preliminary proposals (cont.)

- Option D
  - Streaming buffers for the input matrices


b
$B_{1\times4\lambda}$
a
$A_{4\lambda\times l}$
v0 $C_{\lambda\times\lambda}$
v1 $C_{\lambda\times\lambda}$
v2 $C_{\lambda\times\lambda}$
v3 $C_{\lambda\times\lambda}$
v4 $C_{\lambda\times\lambda}$
v5 $C_{\lambda\times\lambda}$
v6 $C_{\lambda\times\lambda}$
v7 $C_{\lambda\times\lambda}$
v8 $C_{\lambda\times\lambda}$
v9 $C_{\lambda\times\lambda}$
v10 $C_{\lambda\times\lambda}$
v11 $C_{\lambda\times\lambda}$
v12 $C_{\lambda\times\lambda}$
v13 $C_{\lambda\times\lambda}$
v14 $C_{\lambda\times\lambda}$
v15 $C_{\lambda\times\lambda}$

- Option E
  - Variable matrix representation



# Agenda

- IME preliminary proposals
- Qualitative vs Quantitative approaches
- Metrics and workloads

# Qualitative vs. Quantitative approaches

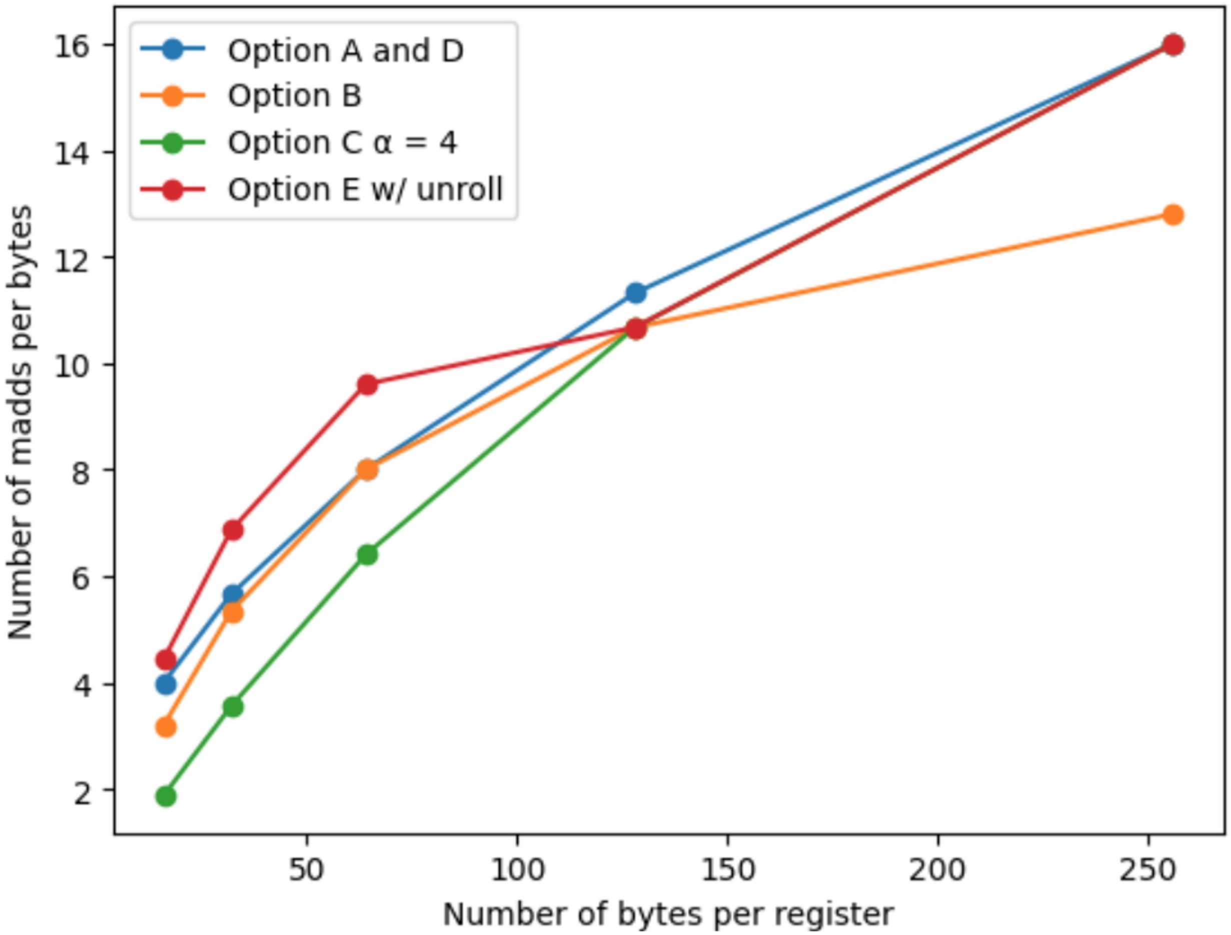
- Qualitative
  - ISA analytical evaluation
  - Analytical back-of-envelope estimates for metrics
  - Technical discussions on the merits of each approach
  - Faster progress, but ad-hoc

- Quantitative
  - Generate workloads
  - Run QEMU model, provided by proponent (open)
  - Run traces on simulator, provided by proponent (closed)
  - Measure metrics and evaluate
  - Slower progress, but precise

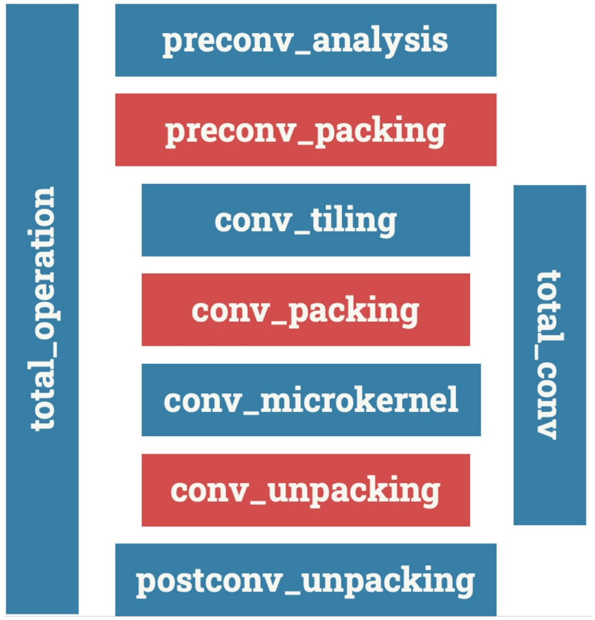
# Quantitative Approach

| | Polybench<br>Convbench<br>Mobile | QEMU | Options' Simulators | Metrics |
|---|---|---|---|---|
| • Option A | | | | |
| • Option B | | | | |
| • Option C | | | | |
| • Option D | | | | |
| • Option E | | | | |

# Qualitative approach



# Quantitative approach



total_operation

preconv_analysis

preconv_packing

conv_tiling

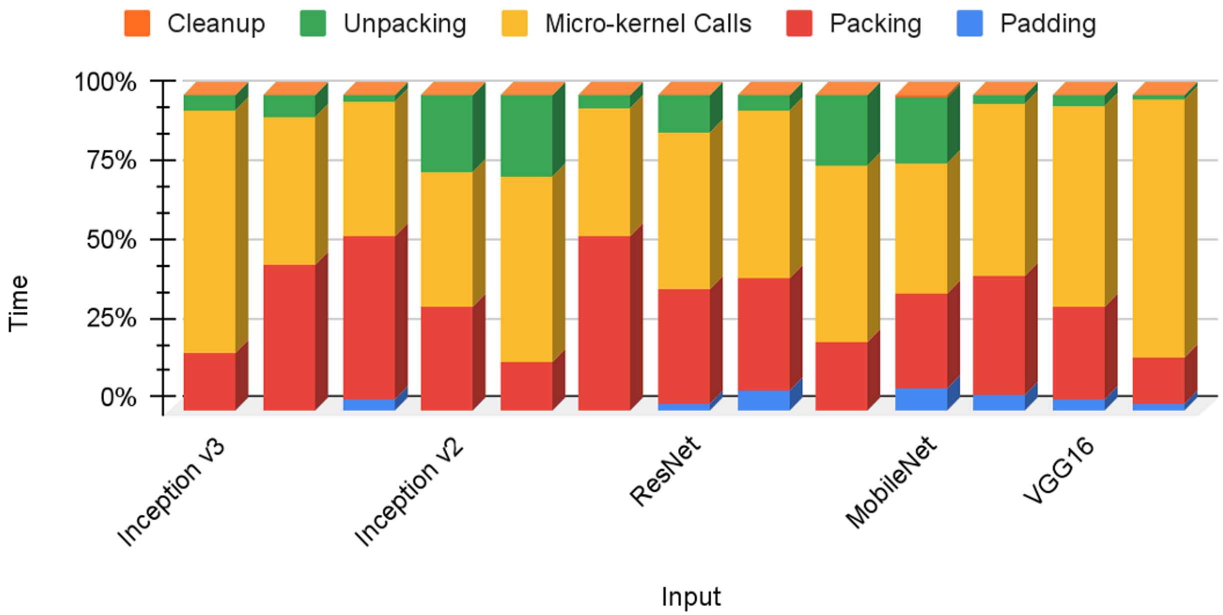
conv_packing

conv_microkernel

conv_unpacking

postconv_unpacking

total_conv



# Agenda

- IME preliminary proposals
- Qualitative vs Quantitative approaches
- Metrics and workloads

# Performance metrics

- Computational intensity
- GEMM execution time
- CONV execution time
- ISA integration with RVV
- Cache misses
- Your favorite metric: TBD

# Domain workloads

- ML: Convbench
- HPC: Polybench
- Mobile: TBD
- Your favorite domain: TBD

# Thanks!!

**Guido Araujo**
**Jose Moreira**

02/12/24