Integrated Matrix Extension (IME)

Task Group Meeting

Guido Araujo
Jose Moreira

03/11/24

Agenda

- Chair/Vice-chair selection update

- Moving forward on qualitative analysis
  - Computational Intensity
  - Locality evaluation

- Presentation
  - Matrix Tile Extension: Portable ISA For Vector-Integrated Matrix Unit
  - Erich Focht (NEC) and Marc Casas (BSC)

Agenda

- Chair/Vice-chair selection update

- Moving forward on qualitative analysis
  - Computational Intensity
  - Locality evaluation

- Presentation
  - Matrix Tile Extension: Portable ISA For Vector-Integrated Matrix Unit
  - Erich Focht (NEC) and Marc Casas (BSC)

Chair/Vice-chair selection update

- Call started: February 23rd.
- Call Closure: Friday, March 8th.
- Completion of Voting/Selection: Friday, March 29th

Agenda

- Chair/Vice-chair selection update

- Moving forward on qualitative analysis
  - Computational Intensity
  - Locality evaluation

- Presentation
  - Matrix Tile Extension: Portable ISA For Vector-Integrated Matrix Unit
  - Erich Focht (NEC) and Marc Casas (BSC)

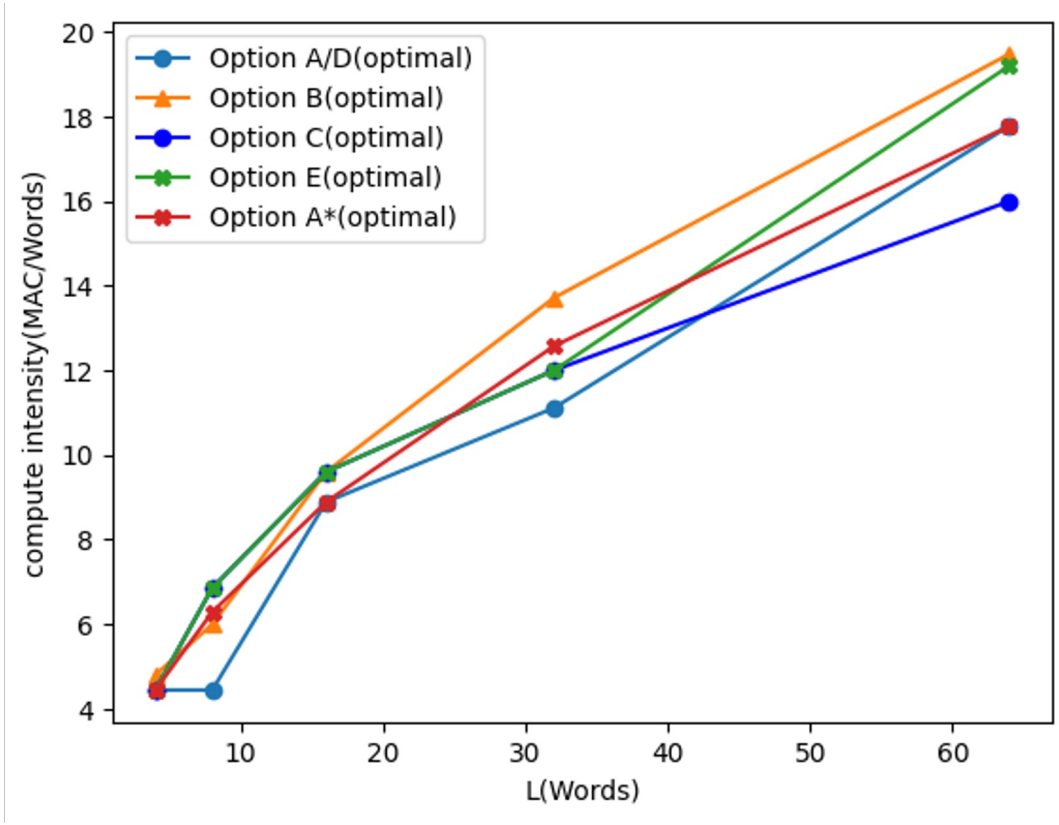
Agenda

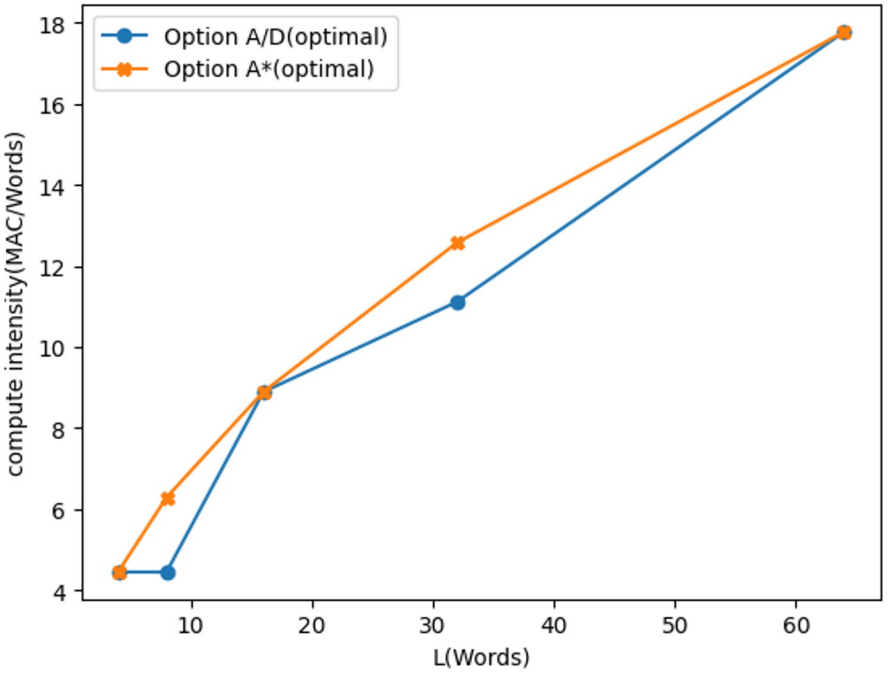
- Chair/Vice-chair selection update

- Moving forward on qualitative analysis
  - Computational Intensity
  - Locality evaluation

- Presentation
  - Matrix Tile Extension: Portable ISA For Vector-Integrated Matrix Unit
  - Erich Focht (NEC) and Marc Casas (BSC)

Computational Intensity (Adding A*)





Thanks

CN-Ke
BING

Charter Criteria Discussion

- Based on proposing Matrix-TG → 9 explicitly guide-lines
- Suggest to add Metrics for Performance/uArch Cost

No.	Guides	Option A	Option B	Option C	Option D	Option E
1	**VLEN agnostic** at binary level	Not Disclosed while L!=λ^2	To Be Discussed (seems support?)		Not Disclosed while L!=λ^2	(1) Source Compliant (2) Binary Compliant for vmul/fused ISA
2	**Deterministic Result** (FMAC rounding/ordering)	Shall support	Shall support	Shall support	Shall support	Support if finalized
3	**Re-producible result** with plain scalar/vector	Shall support	Shall support	Shall support	Shall support	Support (BIT TRUE test)
4	**Near peak (~90%) performance** for GEMM kernels is possible					GEMM kernels Near Peak U-rate
5	**Higher (~2X) performance** than vector					Over (>3X) enhancement than RVV
6	Maximization **computation intensity** for GEMM kernels					GEMM kernels Near Peak U-rate
7	Minimization additional **architecture state**	None new state	None new state	None new state	Not Support (New Streaming buffer for A/B)	None new state (not considering ZOB)
8	**Live-migration** with larger vector registers	Not Support for different lambda	To Be Discussed (seems support?)	Not Support for different vector element type	Not Support for different lambda	Under Working (AMM 2.0)
9	Proper Support for **packing/reformat data**	May Need Additional handling while L!=λ^2	Support (No additional interleaving/shuffle required)	Support (No additional interleaving/shuffle required)	May Need Additional handling while L!=λ^2	Support (No additional interleaving/shuffle required)
10	**Metrics for Performance/uArch cost (Suggest to consider)**	Feasible uArch cost(VRF R/W) for Specific MAC Performance	High uArch cost(VRF R/W) for MAC Performance†	feasible uArch cost(VRF R/W) for MAC Performance	Feasible uArch cost(VRF R/W) for MAC Performance	Feasible uArch cost(VRF R/W) for MAC Performance

†: refer details at

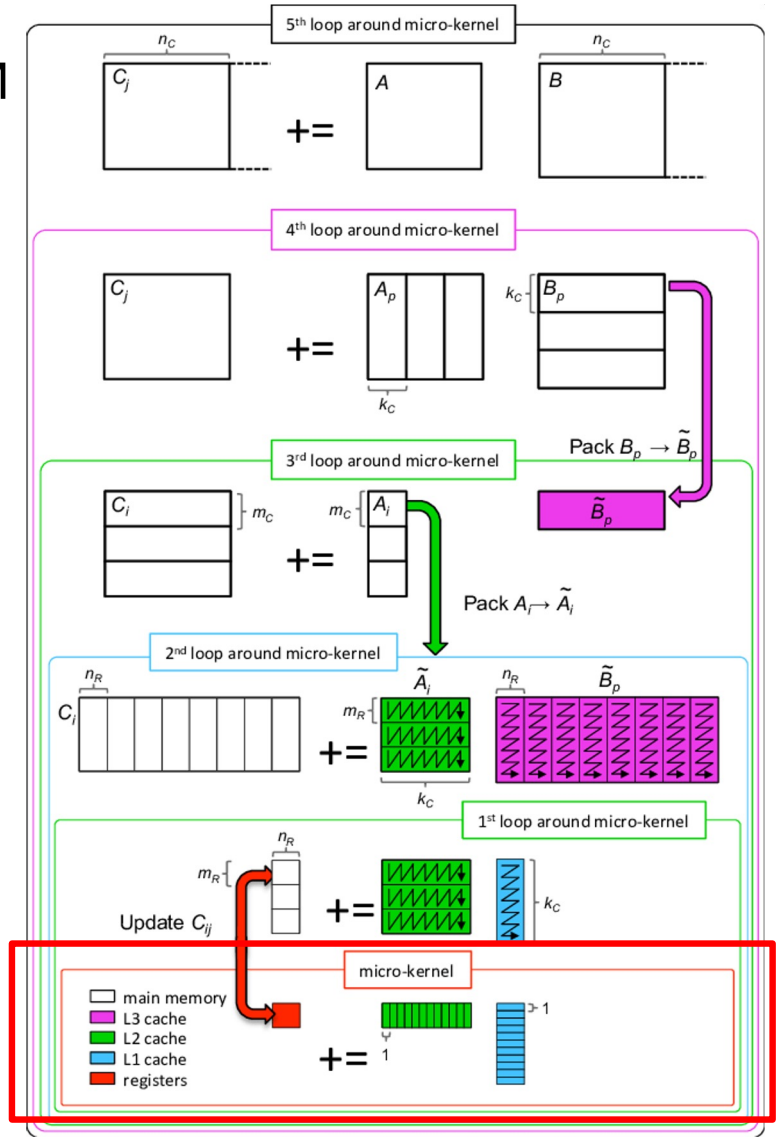
Agenda

- Chair/Vice-chair selection update

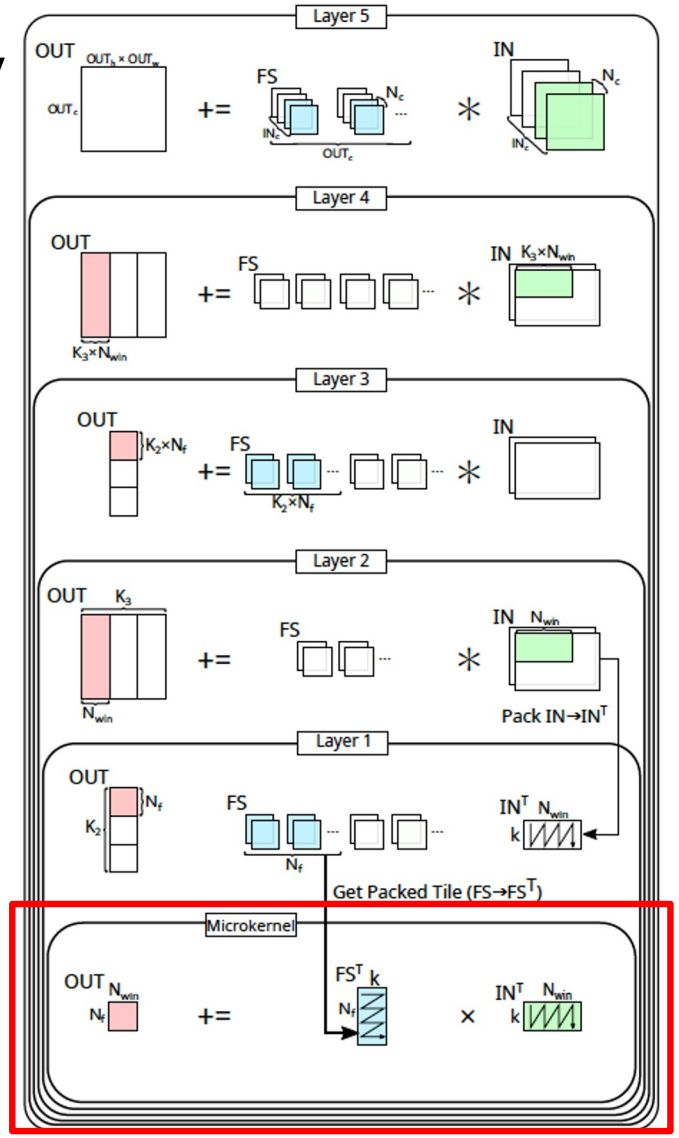
- Moving forward on qualitative analysis
  - Computational Intensity
  - Locality evaluation

- Presentation
  - Matrix Tile Extension: Portable ISA For Vector-Integrated Matrix Unit
  - Erich Focht (NEC) and Marc Casas (BSC)

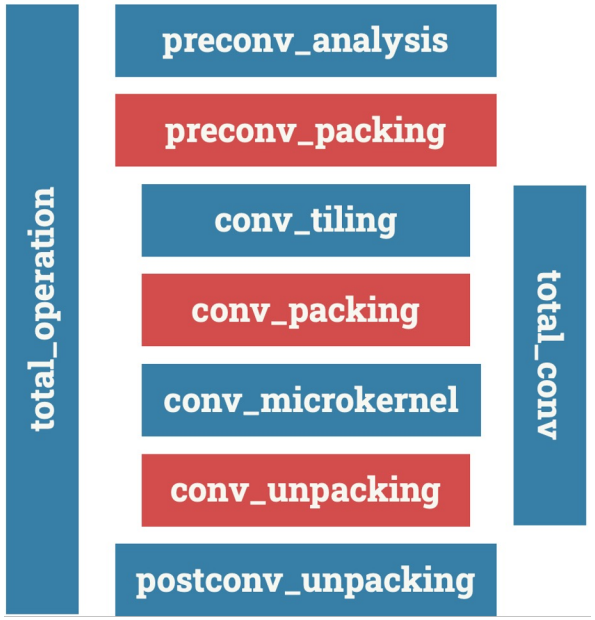
GEMM

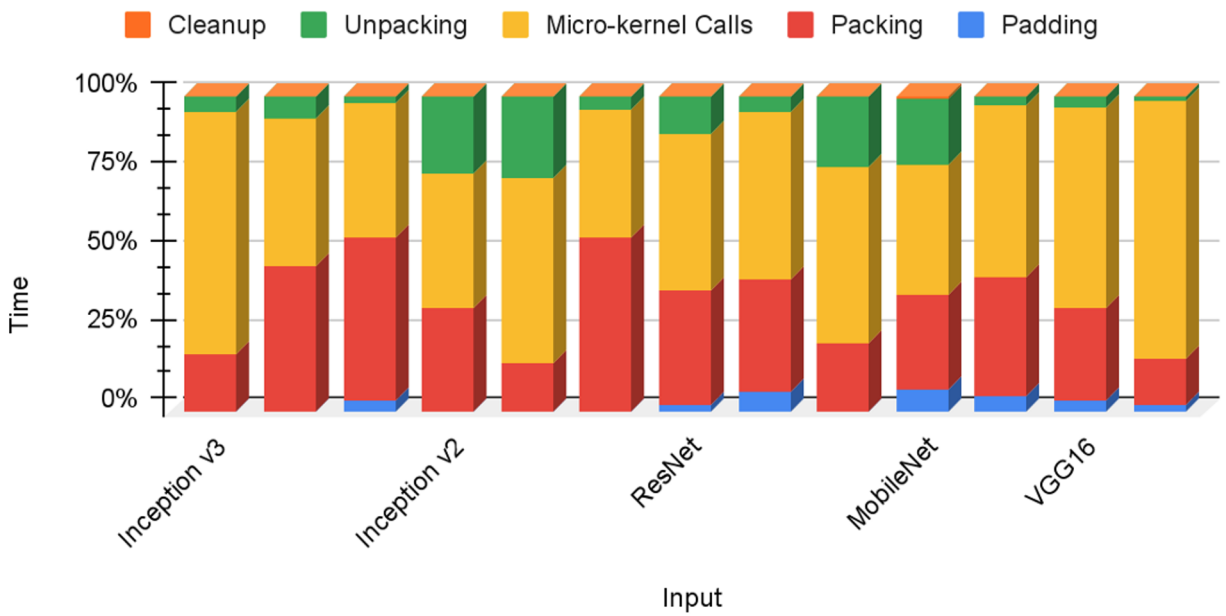

5th loop around micro-kernel
4th loop around micro-kernel
3rd loop around micro-kernel
2nd loop around micro-kernel
1st loop around micro-kernel
micro-kernel
Pack $B_p \rightarrow \tilde{B}_p$
Pack $A_i \rightarrow \tilde{A}_i$
Update C_{ij}
main memory
L3 cache
L2 cache
L1 cache
registers

CONV

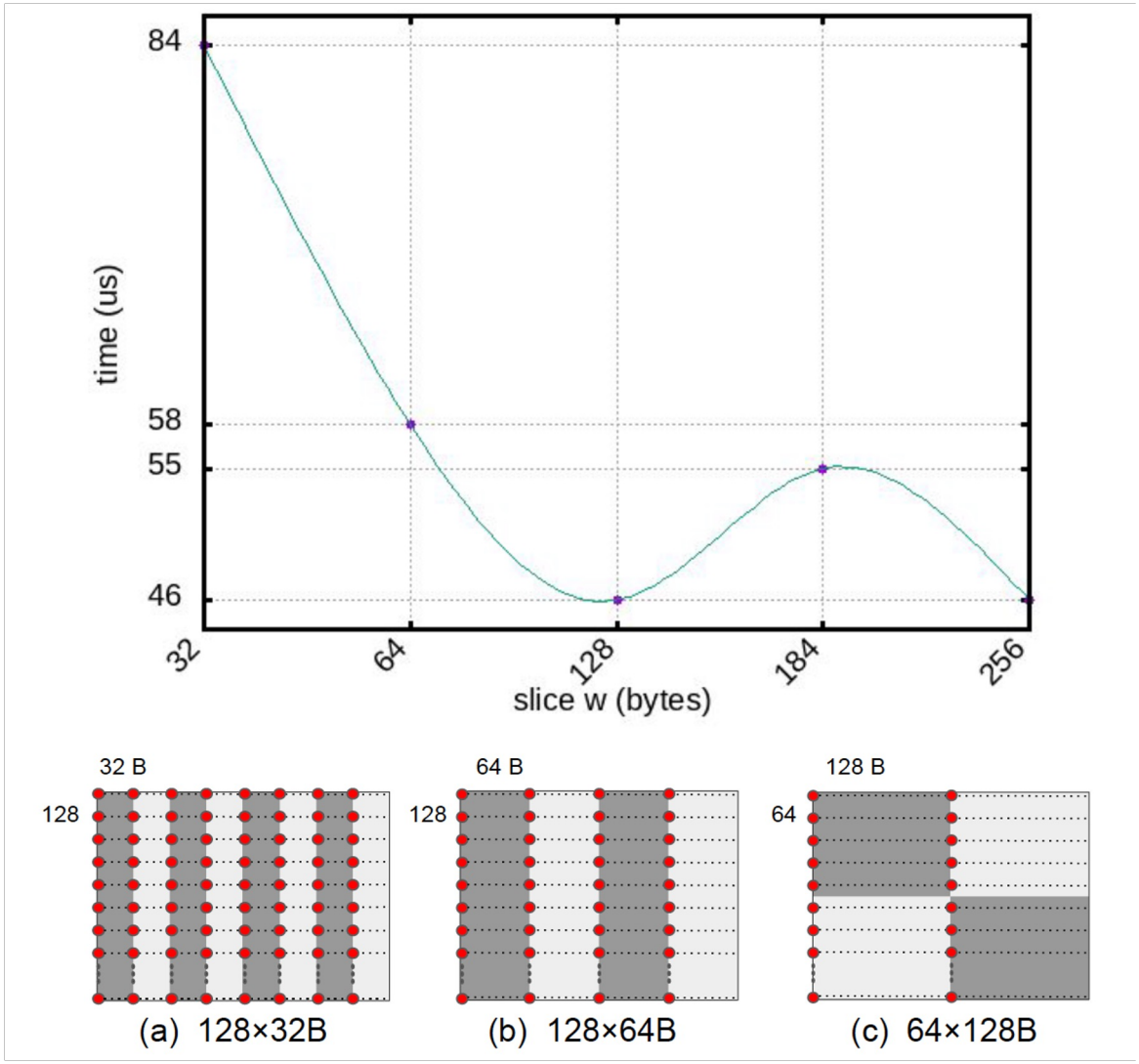

Layer 5
Layer 4
Layer 3
Layer 2
Layer 1
Microkernel
Pack IN→IN^T
Get Packed Tile (FS→FS^T)

Locality is important for packing





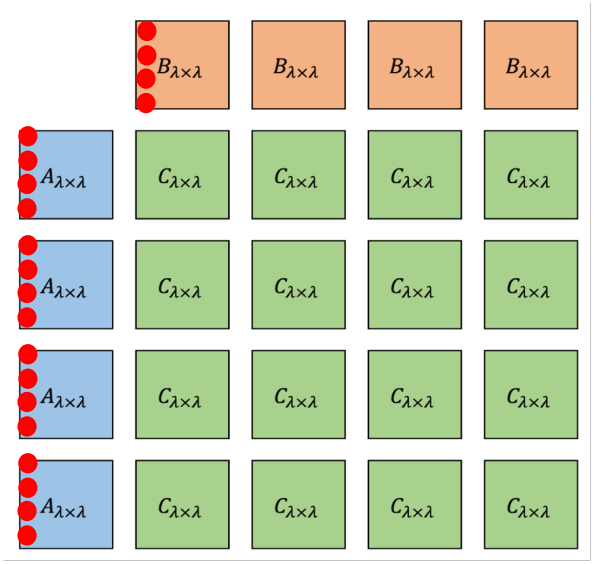
Tiling and memory burst



Evaluate impact on packing

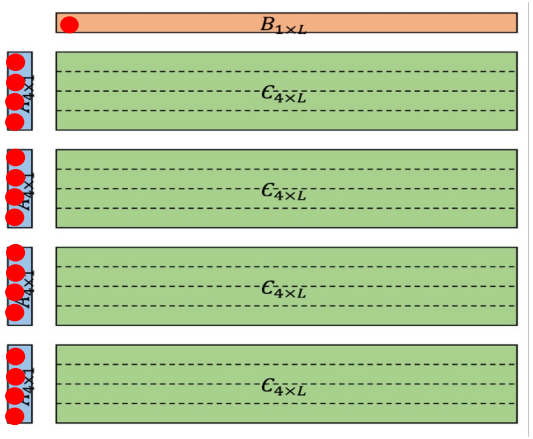
- Option A
  - One matrix per register vector

    5λ



- Option B
  - One matrix in multiple register vectors

    17



Agenda

- Chair/Vice-chair selection update

- Moving forward on qualitative analysis
  - Computational Intensity
  - Locality evaluation

- Presentation
  - Matrix Tile Extension: Portable ISA For Vector-Integrated Matrix Unit
  - Erich Focht (NEC) and Marc Casas (BSC)