Integrated Matrix Extension (IME)

Task Group Meeting

Guido Araujo
Jose Moreira

07/22/24

Agenda

- Update on working groups and TG schedule
- [Guido] Introduction on IME Bounds
- [Earl] Architecture proposal
- [Jose] Updates on Option C*

Agenda

- Update on working groups and TG schedule
- [Guido] Introduction on IME Bounds
- [Earl] Architecture proposal
- [Jose] Updates on Option C*

Working groups

Group	Coordinator	Members
Option A and A*	Marc Casas	Huayue Liang, Erich Focht
~~Option B~~		
Option C and C*	Jose Moreira	
Option D	Abel Bernabeu	
Option E	Jim (CN.Ke)	Yi-Xuan.Huang
Workloads and benchmarking	Guido Araujo	

Roadmap

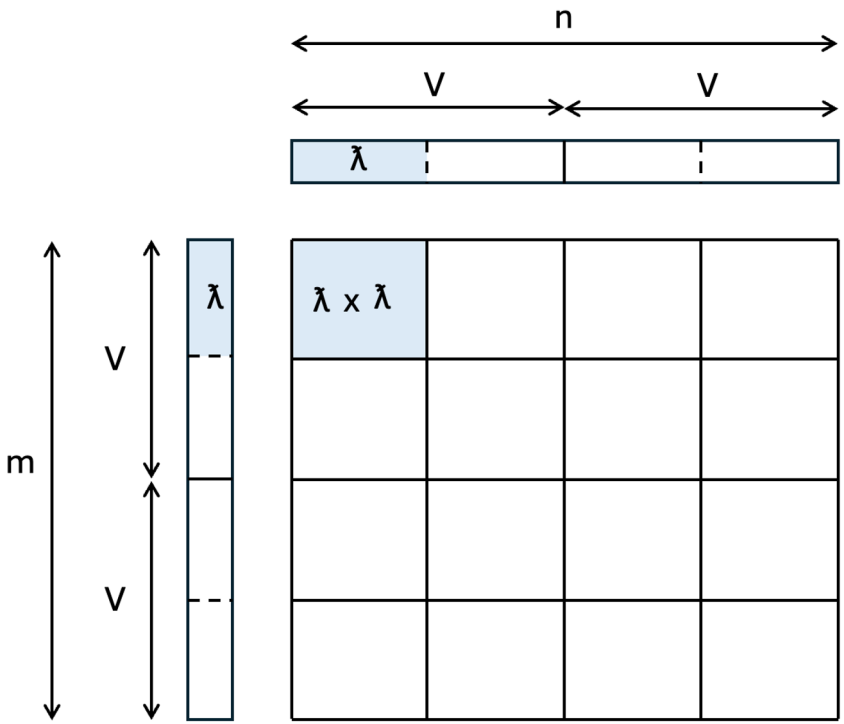
Task	Del.	Task Description	Meetings											
			1	2	3	4	5	6	7	8	9	10	11	12
1		Architectural features												
2		a. uArch: Overall analysis												
3		b. uArch: Memory access analysis												
4		c. ISA: Matrix data encoding												
5		d. ISA: Register usage and mapping												
6		e. ISA: Data type and geometry configuration												
7		f. ISA: Binary compatibility												
8		g. ISA: Computation operations definition												
9		h. ISA: Instruction encoding												
10		Workloads and bechmarking												
11		a. ML: T-Head profiling and ConvBench												
12		b. HPC: Polybench												
13		c. ML: POWER10 MMA transfers												
14		d. Workload analysis												
15		Quantitative analysis												
16		a. QEMU modelling												
17		b. Performace evaluation												
18		Definition of the final architecture												
19		a. RVM ISA v0												
20		b. RVM ISA v1												
21		RVM Spec writing												

Delivery date
Still discussing
Not touched
N Past meeting
Delayed

Agenda

- Update on working groups and TG schedule
- [Guido] Introduction on IME Bounds
- [Earl] Architecture proposal
- [Jose] Updates on Option C*

Notation and Intuition



$$CI = \frac{m.n}{m + n}$$

# of accumulator vector registers

source operand vector registers

Propositions

Lemma L

Computational intensity is maximized when m = n, and thus CI = n/2.

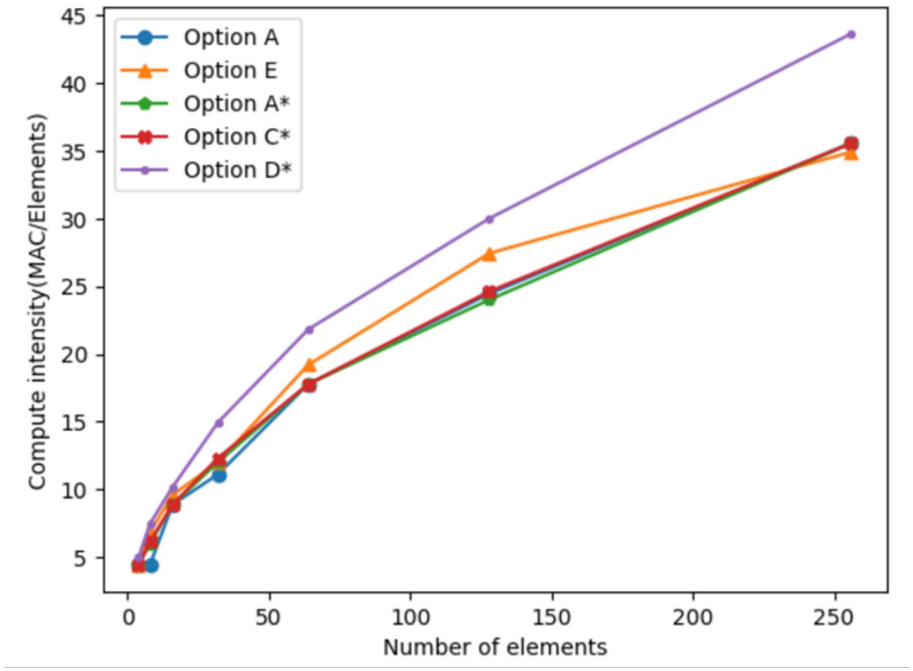
Corollary C

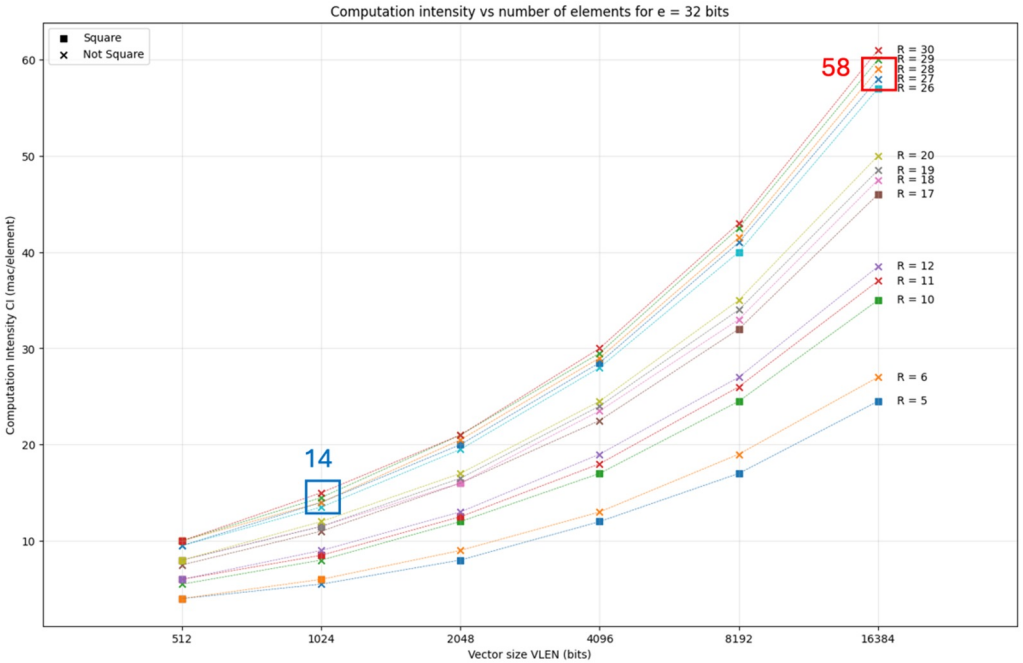
For a matrix multiplication engine E with R vector registers available for the operation and vectors with V elements, the dimension n that maximizes the computational intensity is:

$$n = \sqrt{RV + 1} - 1$$

Moreover, E requires just 2 source operand vector registers.

Results


Option A
Option E
Option A*
Option C*
Option D*
Compute intensity(MAC/Elements)
Number of elements
0
50
100
150
200
250
5
10
15
20
25
30
35
40
45


Computation intensity vs number of elements for e = 32 bits
Square
Not Square
Computation Intensity CI (mac/element)
Vector size VLEN (bits)
512
1024
2048
4096
8192
16384
10
20
30
40
50
60
14
58
R = 30
R = 29
R = 28
R = 27
R = 26
R = 20
R = 19
R = 18
R = 17
R = 12
R = 11
R = 10
R = 6
R = 5

Agenda

- Update on working groups and TG schedule
- [Guido] Introduction on IME Bounds
- [Earl] Architecture proposal
- [Jose] Updates on Option C*

Agenda

- Update on working groups and TG schedule
- [Guido] Introduction on IME Bounds
- [Earl] Architecture proposal
- [Jose] Updates on Option C*