Update on Ara

07/02/2023

Matteo Perotti
Matheus Cavalcante

Professor Luca Benini
Integrated Systems Laboratory
ETH Zürich

Summary

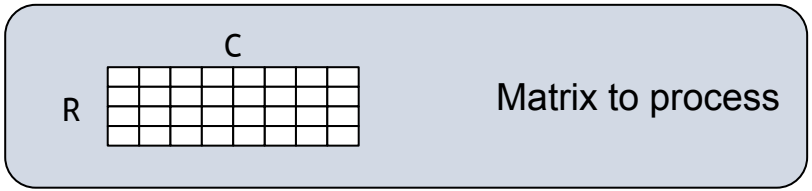
- **Multi-Core Experiment**
  - Concepts
  - 16 Lanes Experiment
  - 2, 4, 8, 16 Lanes Experiment

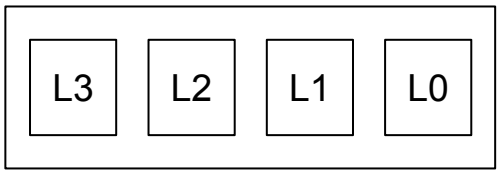
Multi-Core Experiment

- **Ara - Vector processor**
  - Parameter: #Lanes
  - 1 Lane → 1 FPU

Multi-Core Experiment

- **Ara - Vector processor**
  - Parameter: #Lanes
  - 1 Lane → 1 FPU
  - All the lanes work on a single vector!



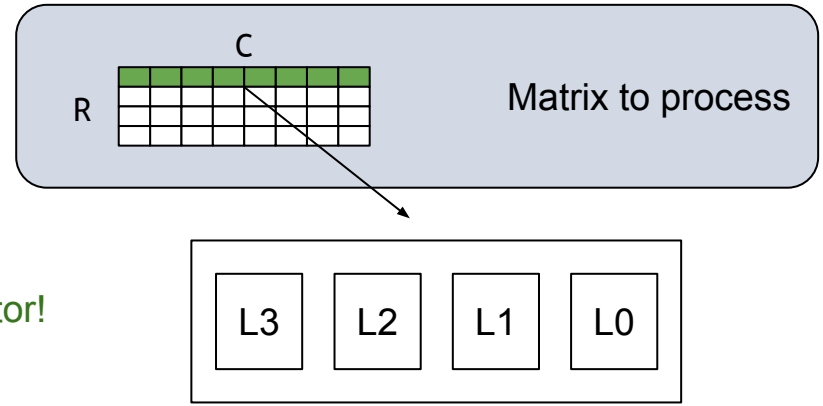


4 FPUs in total

Ara (4 Lanes)

Multi-Core Experiment

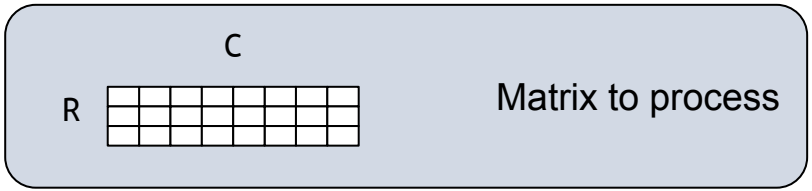
- **Ara - Vector processor**
  - Parameter: #Lanes
  - 1 Lane → 1 FPU
  - All the lanes work on a single vector!



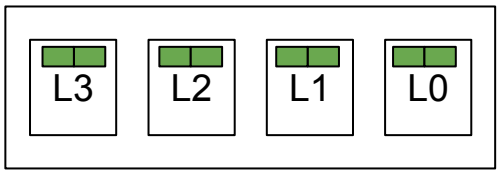
Longest vector!

Multi-Core Experiment

- **Ara - Vector processor**
  - Parameter: #Lanes
  - 1 Lane → 1 FPU
  - All the lanes work on a single vector!



Lanes are not filled



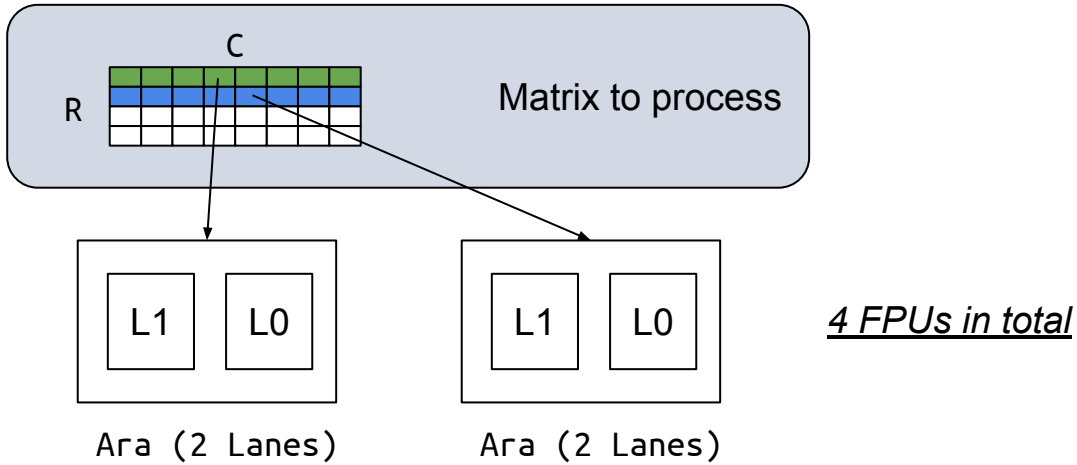
Ara (4 Lanes)

Multi-Core Experiment

- **Ara - Multi-Core**
  - Parameter: #Lanes
  - 1 Lane → 1 FPU

Multi-Core Experiment

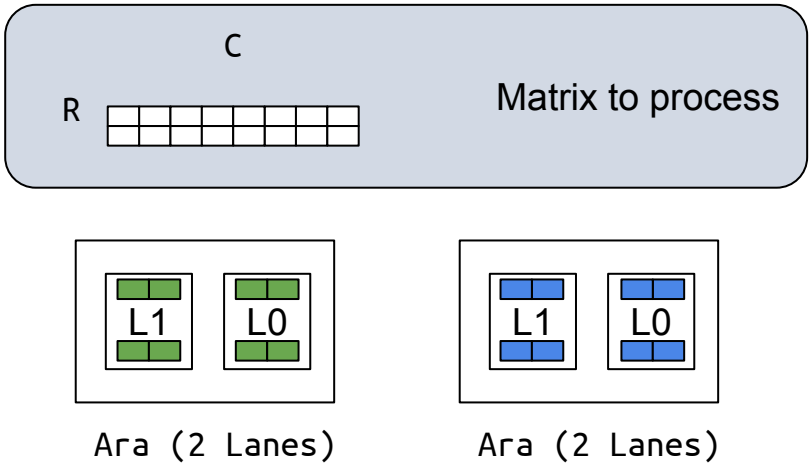
- **Ara - Multi-Core**
  - Parameter: #Lanes
  - 1 Lane → 1 FPU
  - Two Ara can work on two vectors!



4 FPUs in total

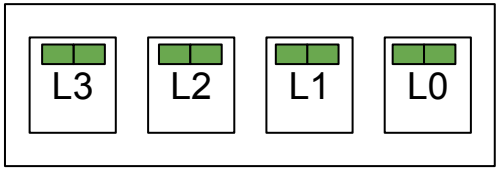
Multi-Core Experiment

- **Ara - Multi-Core**
  - Parameter: #Lanes
  - 1 Lane → 1 FPU
  - Each Ara works on a vector!



Multi-Core Experiment

1 Core, 4 Lanes

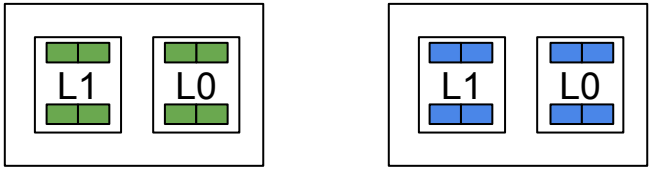


Ara (4 Lanes)

4 FPUs

2 Elements/Lane

2 Cores, 2 Lanes each



Ara (2 Lanes) Ara (2 Lanes)

4 FPUs

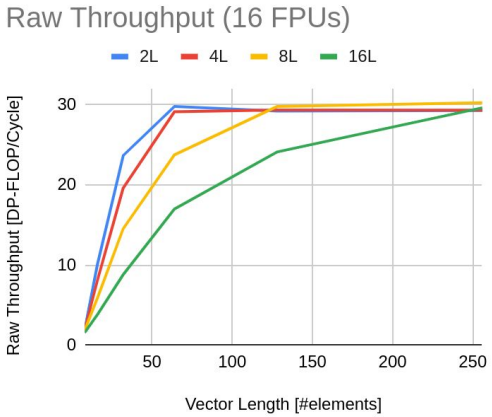
4 Elements/Lane

Multi-Core Experiment

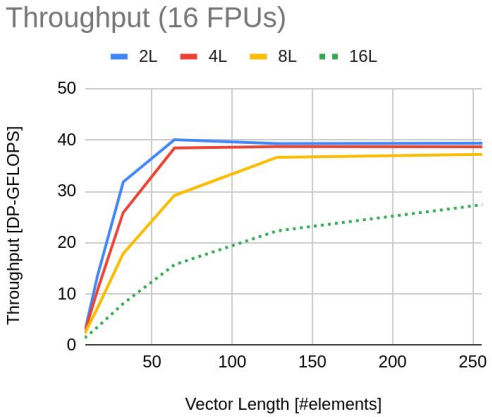
This happens when the vectors are short!

Multi-Core Experiment - Results

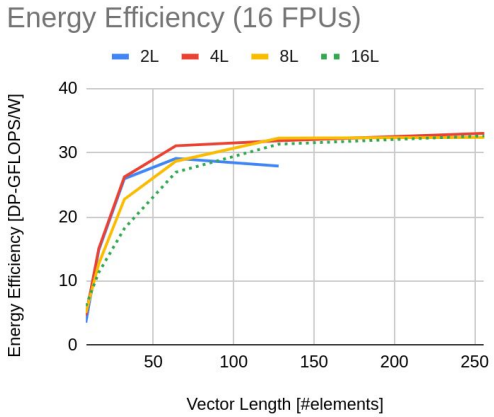
- **16 FPUs experiment**
  - FP-matmul
  - Elements from [8, 16, 32, 64, 128, 256]



FLOP/cycle



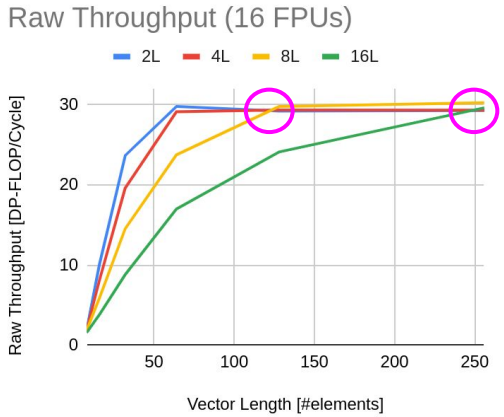
GFLOPS



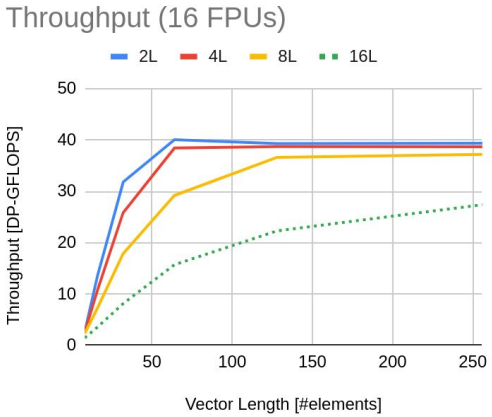
GFLOPS/W

Multi-Core Experiment - Results

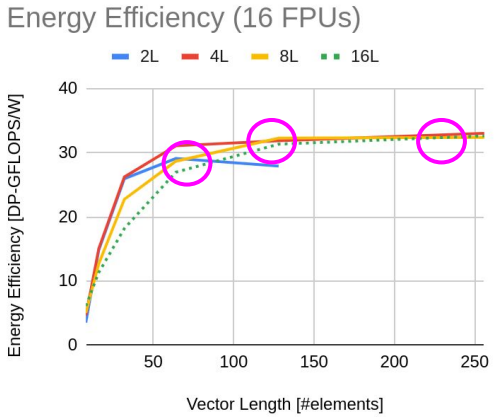
- **16 FPUs experiment**
  - FP-matmul
  - Elements from [8, 16, 32, 64, 128, 256]



FLOP/cycle



GFLOPS

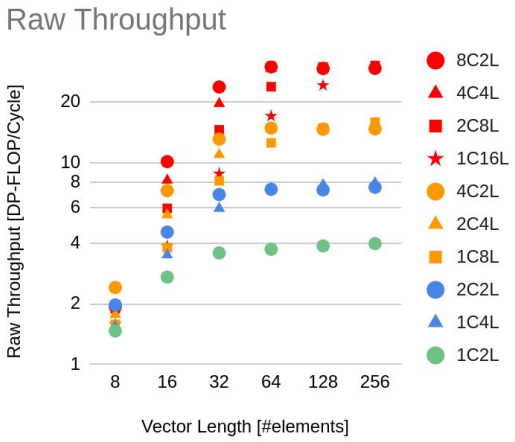


GFLOPS/W

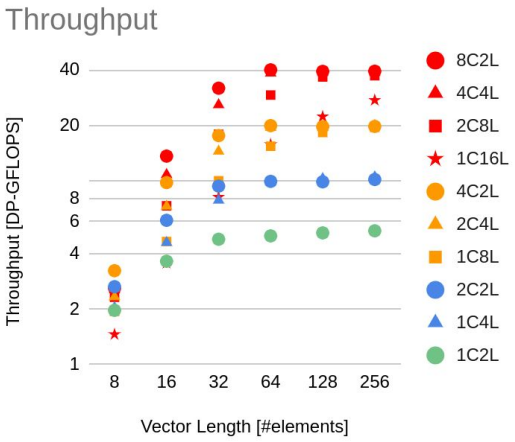
Multi-Core Experiment - Results

- **2, 4, 8, 16 FPUs experiments**
  - FP-matmul
  - Elements from [8, 16, 32, 64, 128, 256]

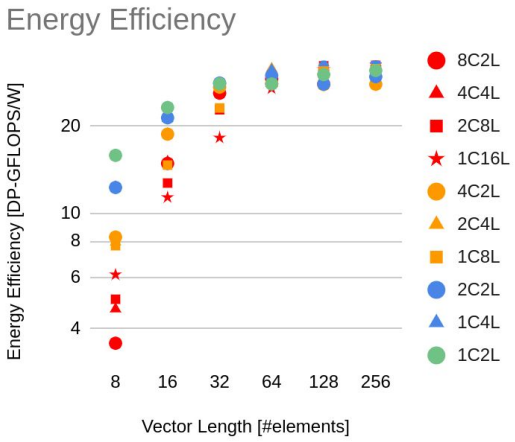
+Complex Shape
==
+Complex Ara core



FLOP/cycle



GFLOPS



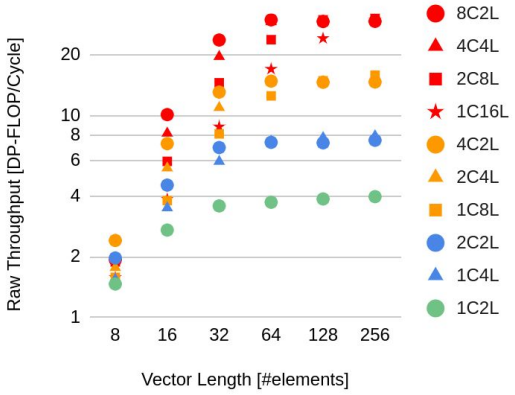
GFLOPS/W

Multi-Core Experiment - Results

- **2, 4, 8, 16 FPUs experiments**
  - FP-matmul
  - Elements from [8, 16, 32, 64, 128, 256]

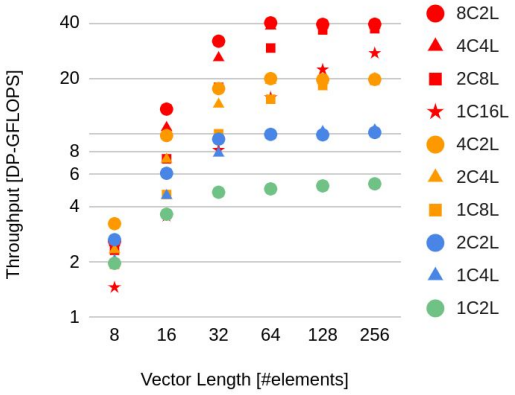
The complex shapes emerge with longer vector lengths

Raw Throughput



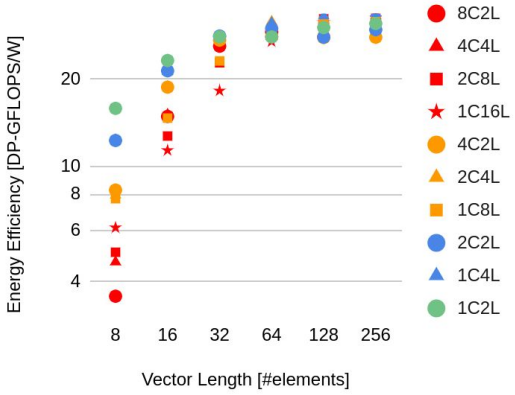
FLOP/cycle

Throughput



GFLOPS

Energy Efficiency



GFLOPS/W