

Update on Ara

05/04/2023

Matteo Perotti

Matheus Cavalcante

Professor Luca Benini

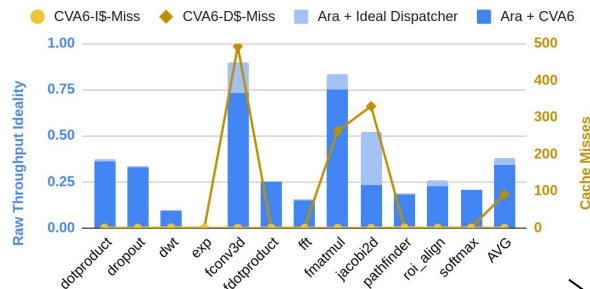
Integrated Systems Laboratory

ETH Zürich

Correlation between Performance and CVA6's \$-Misses

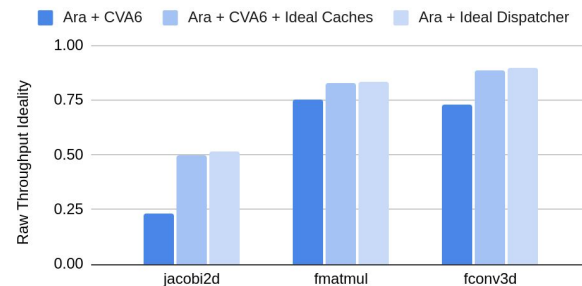
CVA6 Cache Misses vs. Performance

16-Lane Ara - 128 Elements



Performance Impact of the Scalar Memory System

16-lane Ara, 8 Elements/Lane

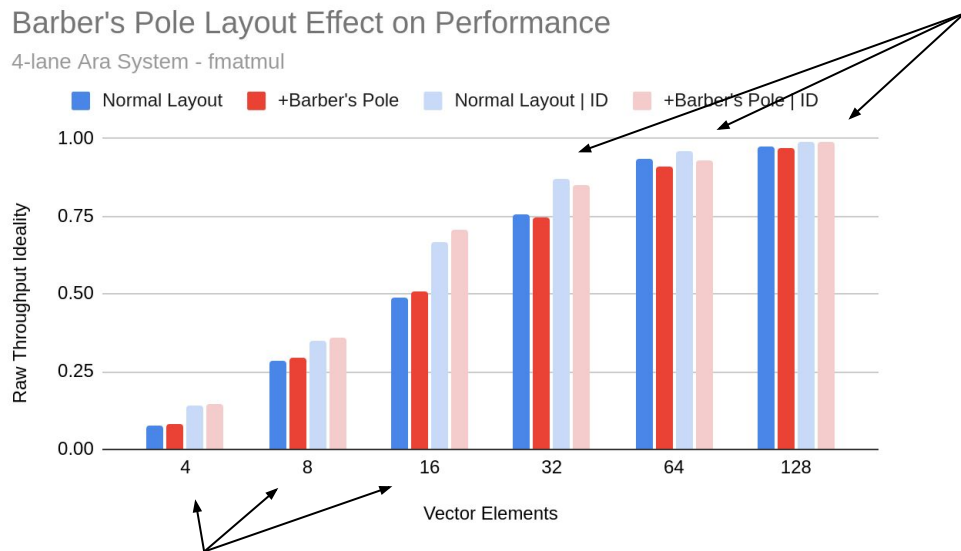


Performance drop mainly due to scalar cache size!

Barber's Pole Layout can be detrimental for medium vectors

Ara has 8 banks per lane

Perturbate VRF access pattern

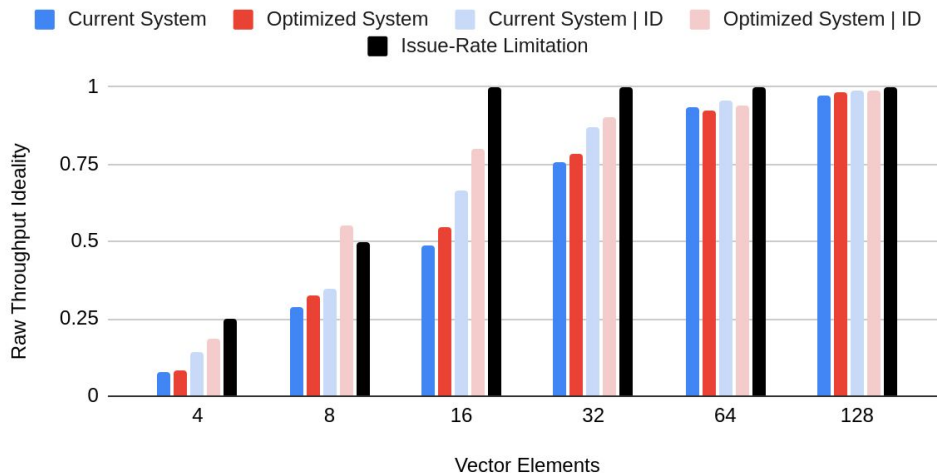


Increase the effective number of banks for short vector applications

Further Ara-optimizations

Additional Optimizations - Effect on Performance

4-lane Ara System - fmatmul



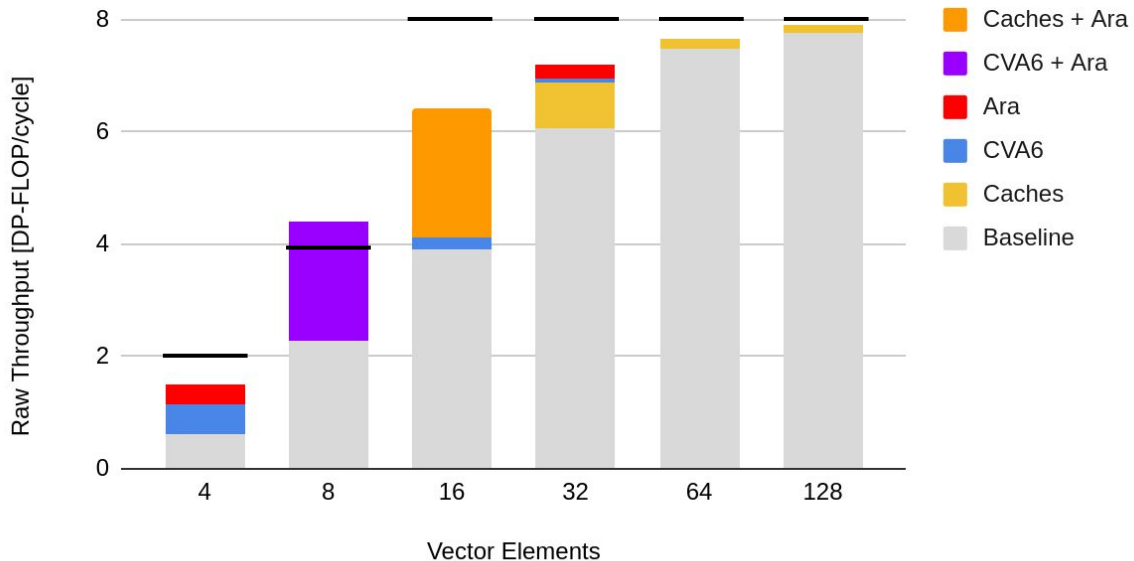
Further optimizing Ara does not boost performance much if not coupled with an improved CVA6 + Caches

Main Performance Drivers

Main Performance Drivers

4-lane Ara - fmatmul

Issue-Rate limitation



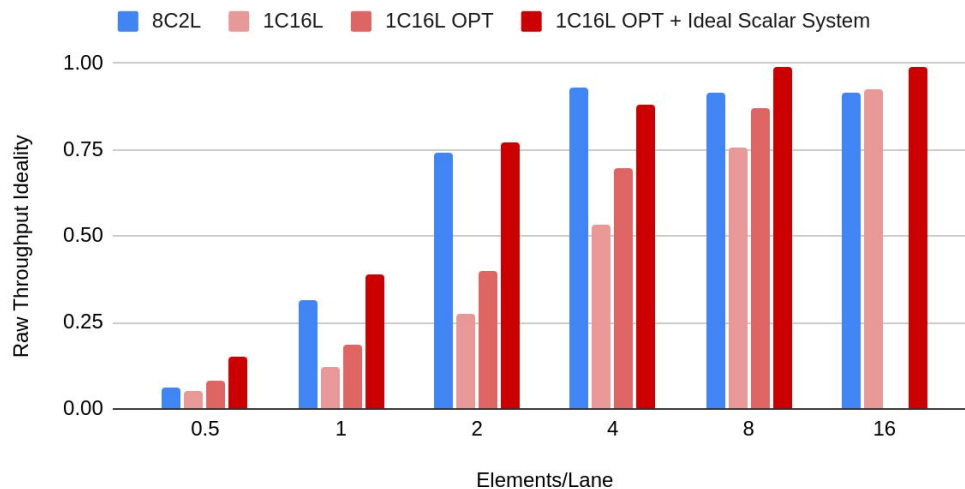
OptAra + CVA6 + Ideal \$
OptAra + Ideal Dispatcher
OptAra + CVA6 + \$
Ara + Ideal Dispatcher
Ara + CVA6 + Ideal \$
Ara + CVA6 + \$

CVA6 limits short vectors' performance, while scalar caches limit medium vectors'

16-FPU - Multi-Core System can overtake Single-Core ID-System

8C2L Multi-Core vs. 1C16L Single-Core Comparison

16-FPU System



Multi-Core can overcome the issue-rate limitation

Overcome the Issue Rate Limitation

fmatmul - 16 FPUs

