Update on Ara

01/12/2021

Matteo Perotti
Matheus Cavalcante
Nils Wistoff

Professor Luca Benini
Integrated Systems Laboratory
ETH Zürich

Summary

- New features:
  - Indexed memory operations
  - Integer reductions
  - Parametric D-cache AXI data-width
- Back-end trials:
  - Performance, Area
  - Analysis of the critical path
- New benchmark
  - Integer reductions

Reductions

- Cycle count for reduction benchmark
  - Varying vector length in Byte

	Cycle Count (#)		
	64 B	512 B	4096 B
2 Lanes	25 / 23	55 / 51	279 / 275
16 Lanes	33 / 32	36 / 32	64 / 60

Longer vectors == Higher Efficiency!!

Reductions

- Cycle count for reduction benchmark
  - Varying vector length in Byte
  - Varying vector element size: *8-bit / 64-bit*

	Cycle Count (#)		
	64 B	512 B	4096 B
2 Lanes	25 / 23	55 / 51	279 / 275
16 Lanes	33 / 32	36 / 32	64 / 60

Cycle count almost independent on element size
8-bit elements == ~8x the throughput of 64-bit elements

Reductions

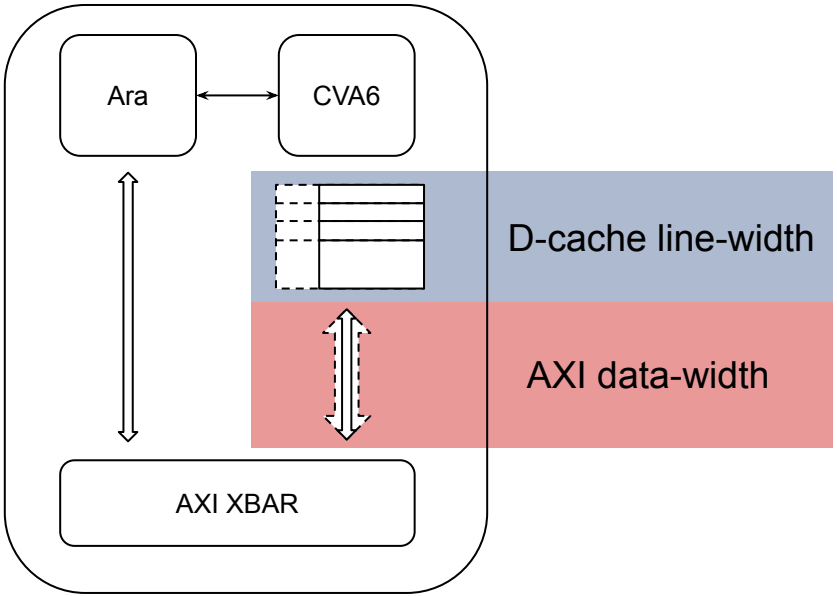
- Cycle count for reduction benchmark
  - Varying vector length in Byte
  - Varying vector element size: *8-bit / 64-bit*
  - Varying number of lanes

	Cycle Count (#)		
	64 B	512 B	4096 B
2 Lanes	25 / 23	55 / 51	279 / 275
16 Lanes	33 / 32	36 / 32	64 / 60

High overhead by inter-lanes reduction phase for shorter vectors or more lanes

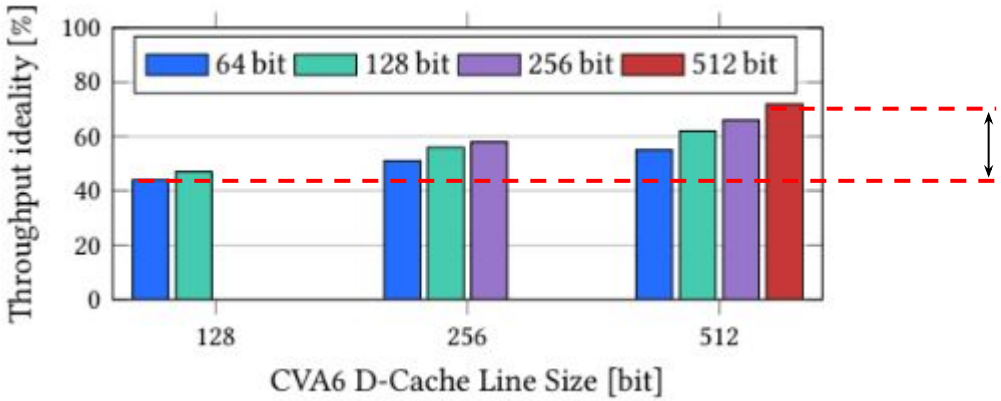
Parametric scalar D-cache

- CVA6's D-cache line-width and AXI data-width are parametric
- Vary AXI data-width and line-width
  - Modify miss penalty
  - Modify miss rate
- **Impact on the throughput!**
  - Especially with short vectors



Parametric scalar D-cache

- Ara 16 lanes: throughput ideality - `fmatmul` with 16x16 matrices
- Ideality w.r.t. system with “ideal” CVA6 and scalar memory system



From 42% to 72%!

Post layout - Ara 4 Lanes

- Configuration:
  - -75% of the original vector register file
- Performance:
  - Comparable or better IPC than Ara V1
  - Better typical corner frequency: 1.36 GHz in typical corner
- Area:
  - 10% smaller die
  - Cell area of one lane: +7% with enhanced features

Post layout - Ara 4 Lanes

- Improving the worst case frequency
  - Critical path between scalar core and D-cache
- Ongoing:
  - Move the D-cache banks