Update on Ara

22/03/2023

Matteo Perotti
Matheus Cavalcante

Professor Luca Benini
Integrated Systems Laboratory
ETH Zürich

Writing and Gathering Data…

Energy Efficiency - `fmatmul, 128x128x128`

- New SLDU
- Macro CLK-Gating

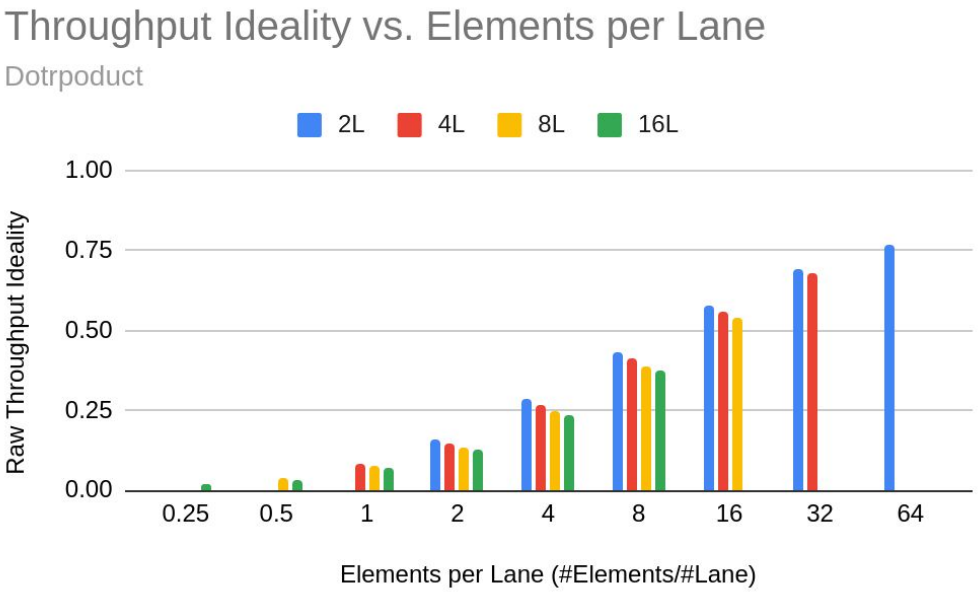
Lanes	Efficiency (DP-GFLOPS/W)
2	30.1
4	**34.3**
8	**32.5**

Still not the trend of AraV1, but important EE gains: +12%

`fmatmul, 128x128x128`

Lanes	Efficiency (DP-GFLOPS/W)
2	34.2
4	**38.7**
8	**35.8**

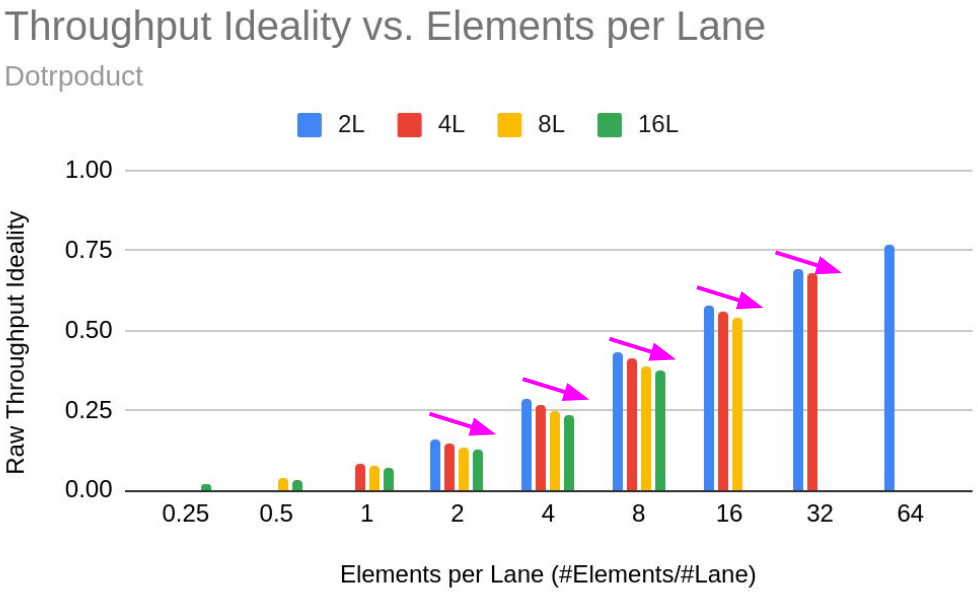
Performance is proportional to Bytes/Lane



	Lanes			
	2L	4L	8L	16L
4	25%	14%	7%	3%
8	55%	35%	19%	9%
16	82%	67%	41%	22%
32	93%	87%	70%	43%
64	96%	96%	89%	72%
128	99%	99%	96%	83%

Same Elements/Lane Ratio - Similar performance

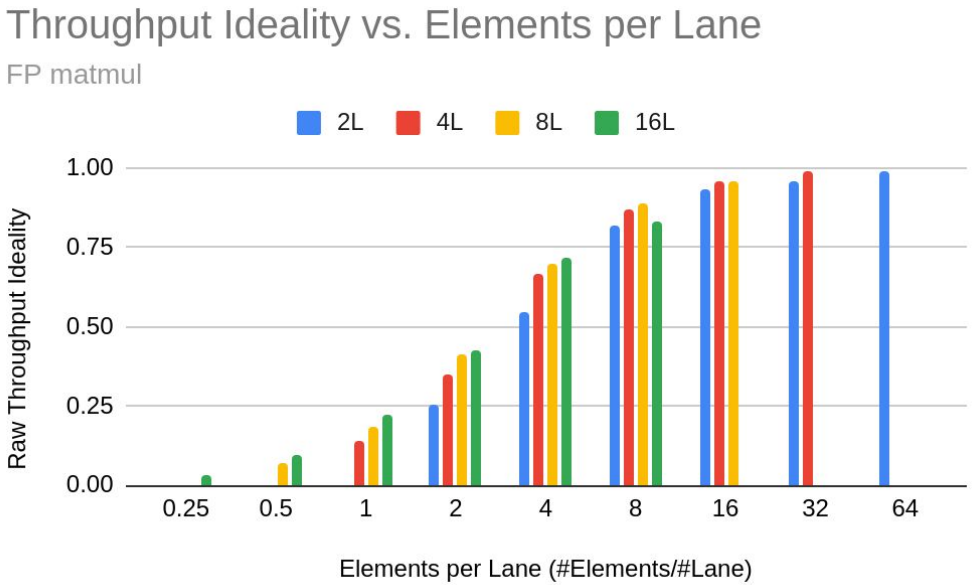
Performance is proportional to Bytes/Lane



	Lanes			
	2L	**4L**	**8L**	**16L**
4	25%	14%	7%	3%
8	55%	35%	19%	9%
16	82%	67%	41%	22%
32	93%	87%	70%	43%
64	96%	96%	89%	72%
128	99%	99%	96%	83%

Inter-Lane Reduction Phase
Latency depends on #Lanes!

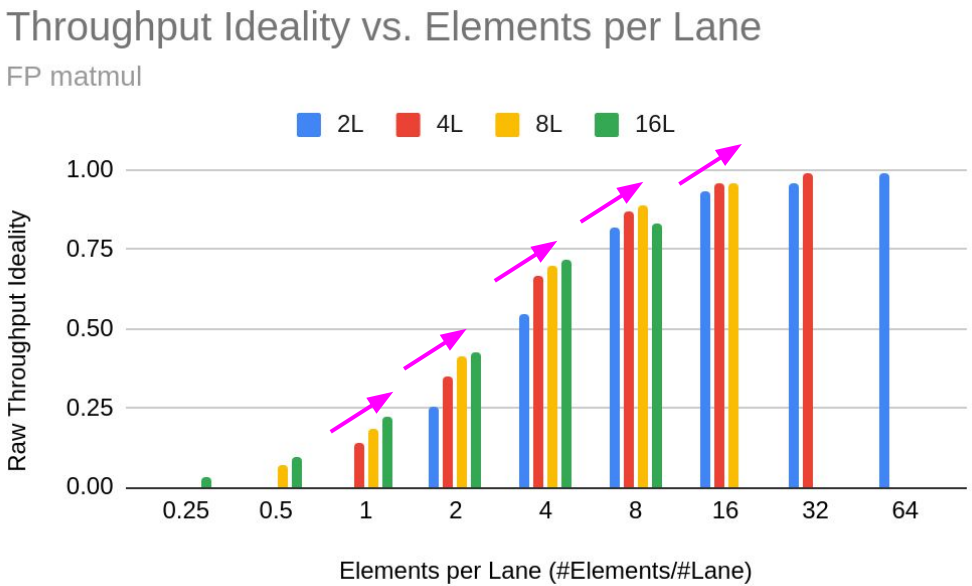
Performance is proportional to Bytes/Lane



	Lanes			
	2L	4L	8L	16L
4	16%	9%	4%	2%
8	29%	15%	8%	4%
16	43%	27%	14%	7%
32	58%	41%	25%	13%
64	70%	56%	39%	24%
128	77%	68%	54%	37%

Same Elements/Lane Ratio - Similar performance

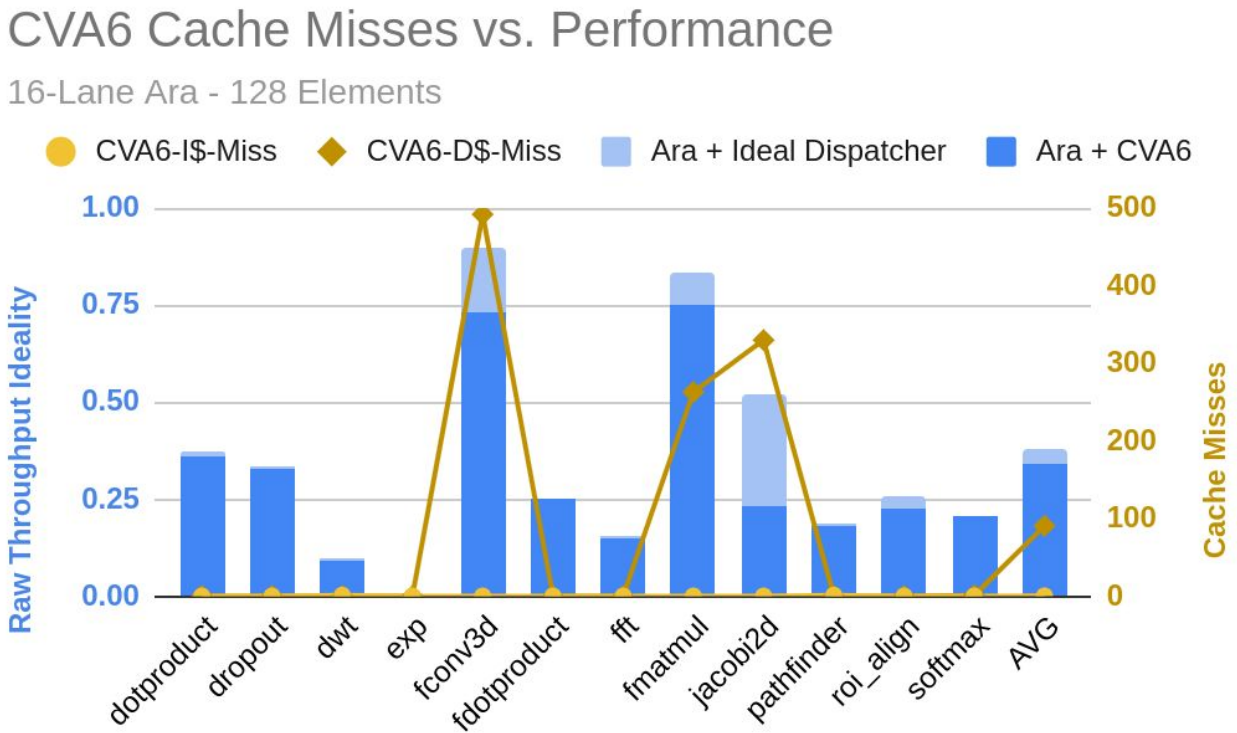
Performance is proportional to Bytes/Lane



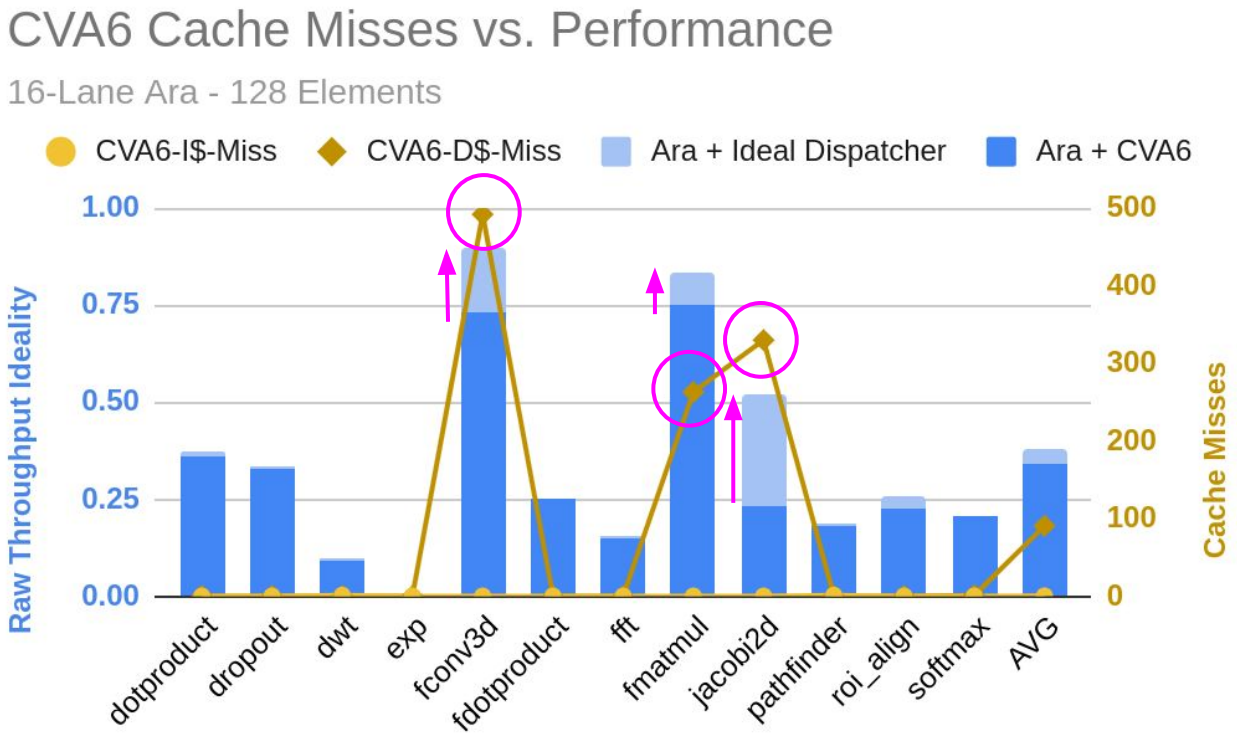
	Lanes			
	2L	4L	8L	16L
4	16%	9%	4%	2%
8	29%	15%	8%	4%
16	43%	27%	14%	7%
32	58%	41%	25%	13%
64	70%	56%	39%	24%
128	77%	68%	54%	37%

`fmatmul` on square matrices
More Elements - Higher arithmetic intensity!

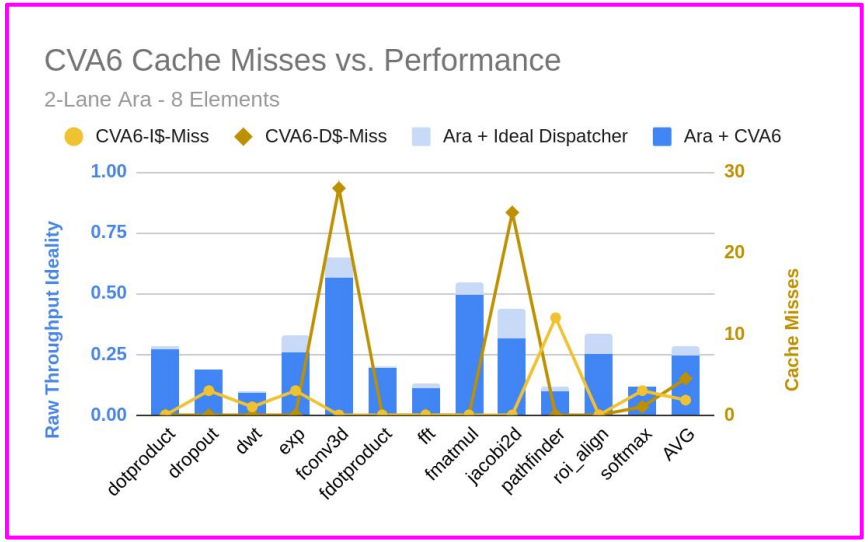
Correlation between Performance and CVA6's $-Misses

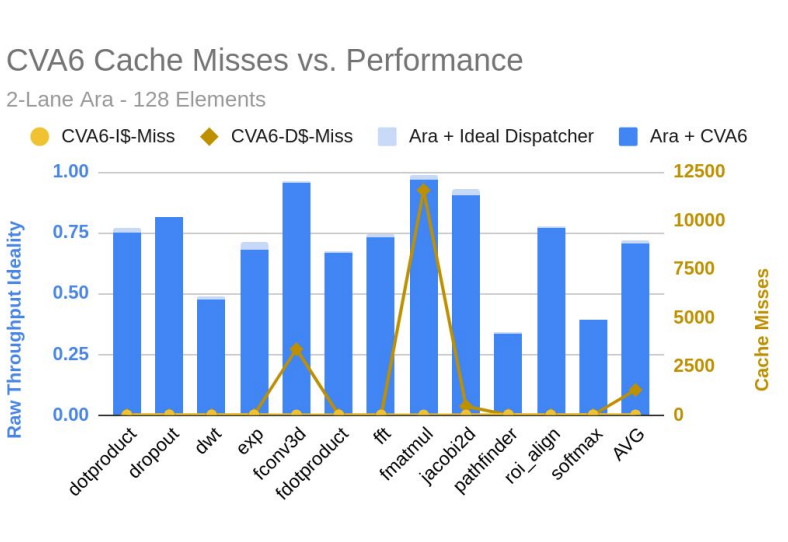


Correlation between Performance and CVA6's $-Misses

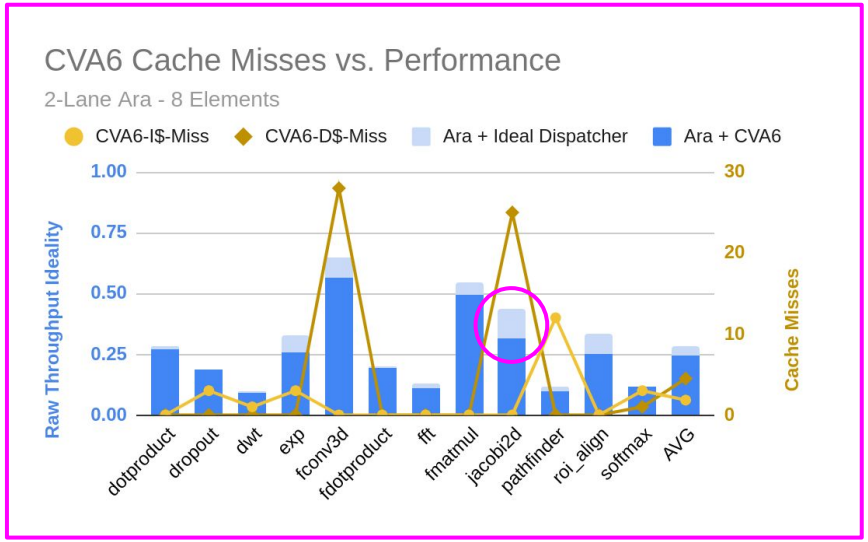


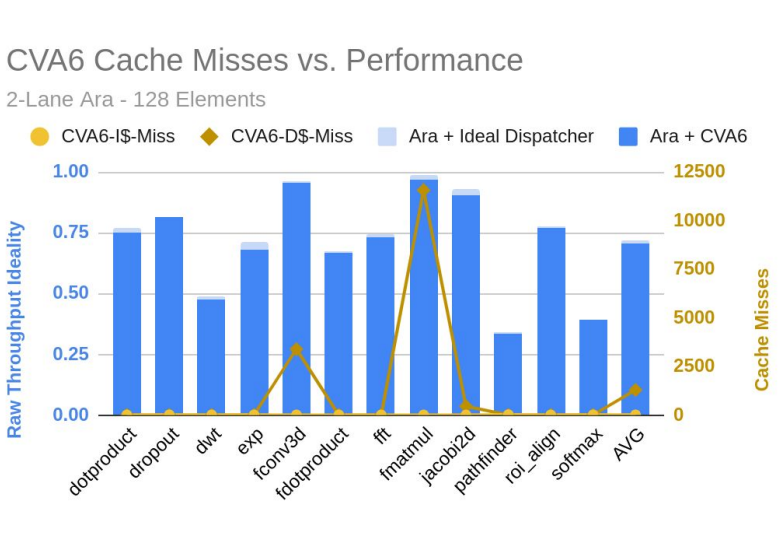
Edge cases - Don’t Blame CVA6





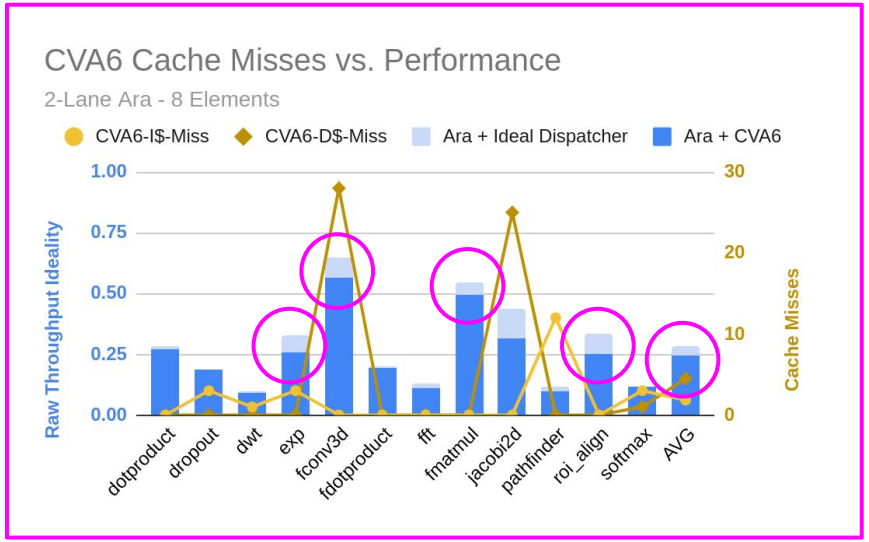
Edge cases - Don’t Blame CVA6

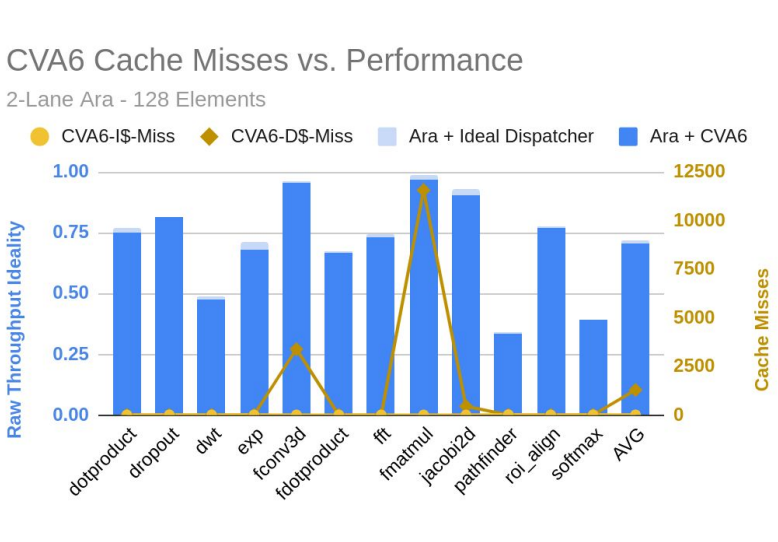




Jacobi2d - We can optimize it…

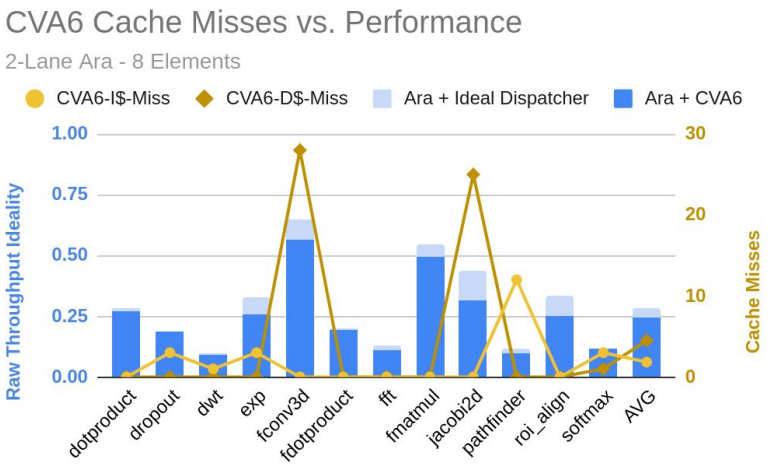
Edge cases - Don’t Blame CVA6

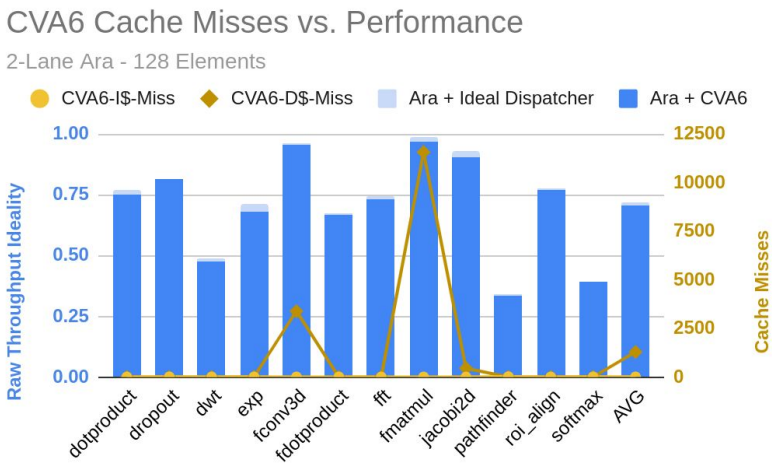




Average gain with ideal dispatcher is still small…

Edge cases - Don’t Blame CVA6





No gain when the #Byte/#Lanes ratio is high! (long vector)
Performance is peaking already