

Update on Ara

24/01/2023 Matteo Perotti

Matheus Cavalcante

Professor Luca Benini Integrated Systems Laboratory ETH Zürich

Summary

SW

Benchmark results

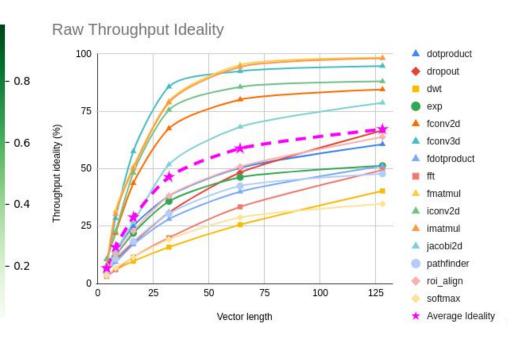
HW

- Popcount optimization
- Implementation: 2, 4, 8, 16 lanes
- Energy efficiency
 - fmatmul
 - imatmul



Relative kernel performance on maximum achievable

lotproduct -	0.082	0.14	0.25	0.38	0.5	0.61
dropout -	0.055	0.1	0.18	0.31	0.48	0.67
dwt -	0.039	0.061	0.097	0.16	0.26	0.4
exp -	0.068	0.12	0.22	0.36	0.46	0.51
fconv2d -	0.11	0.22	0.44	0.68	0.8	0.85
fconv3d -	0.076	0.28	0.58	0.86	0.92	0.95
otproduct -	0.062	0.093	0.17	0.28	0.4	0.51
fft -	0.031	0.06	0.11	0.2	0.33	0.49
fmatmul -	0.084	0.32	0.51	0.79	0.95	0.98
iconv2d -	0.11	0.23	0.48	0.76	0.86	0.88
imatmul -	0.074	0.3	0.5	0.79	0.94	0.98
jacobi2d -	0.077	0.14	0.27	0.52	0.68	0.79
pathfinder -	0.046	0.11	0.18	0.3	0.42	0.48
roi_align -	0.066	0.13	0.23	0.38	0.51	0.64
softmax -	0.034	0.068	0.11	0.19	0.29	0.35
	4	8	16	32	64	128
		Ved	tor Lengt	h [elemer	nts]	



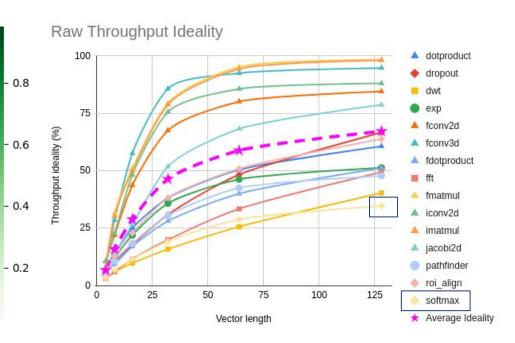


jump to function? + heavy function state preparation

softmax - ... vexp() ... vdiv ...

Relative kernel performance on maximum achievable

		o. p.o.				acilicianic
lotproduct -	0.082	0.14	0.25	0.38	0.5	0.61
dropout -	0.055	0.1	0.18	0.31	0.48	0.67
dwt -	0.039	0.061	0.097	0.16	0.26	0.4
exp -	0.068	0.12	0.22	0.36	0.46	0.51
fconv2d -	0.11	0.22	0.44	0.68	0.8	0.85
fconv3d -	0.076	0.28	0.58	0.86	0.92	0.95
otproduct -	0.062	0.093	0.17	0.28	0.4	0.51
fft -	0.031	0.06	0.11	0.2	0.33	0.49
fmatmul -	0.084	0.32	0.51	0.79	0.95	0.98
iconv2d -	0.11	0.23	0.48	0.76	0.86	0.88
imatmul -	0.074	0.3	0.5	0.79	0.94	0.98
jacobi2d -	0.077	0.14	0.27	0.52	0.68	0.79
pathfinder -	0.046	0.11	0.18	0.3	0.42	0.48
roi_align -	0.066	0.13	0.23	0.38	0.51	0.64
softmax -	0.034	0.068	0.11	0.19	0.29	0.35
	4	8 Ved	16 ctor Lengt	32 h [elemen	64 nts]	128



ETH Zürich | 4 |

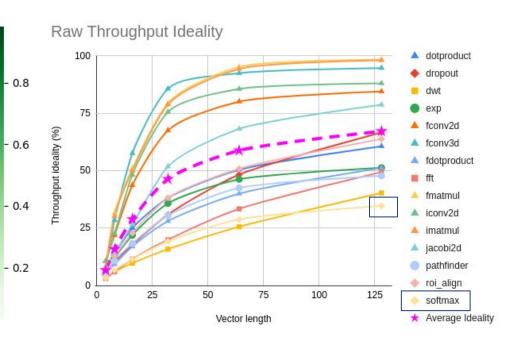


vdiv is way slower than other fp operations

softmax - ... vexp() . . vdiv ..

Relative kernel performance on maximum achievable

		c. p.c.			.,	acinevabi
lotproduct -	0.082	0.14	0.25	0.38	0.5	0.61
dropout -	0.055	0.1	0.18	0.31	0.48	0.67
dwt -	0.039	0.061	0.097	0.16	0.26	0.4
exp -	0.068	0.12	0.22	0.36	0.46	0.51
fconv2d -	0.11	0.22	0.44	0.68	0.8	0.85
fconv3d -	0.076	0.28	0.58	0.86	0.92	0.95
otproduct -	0.062	0.093	0.17	0.28	0.4	0.51
fft -	0.031	0.06	0.11	0.2	0.33	0.49
fmatmul -	0.084	0.32	0.51	0.79	0.95	0.98
iconv2d -	0.11	0.23	0.48	0.76	0.86	0.88
imatmul -	0.074	0.3	0.5	0.79	0.94	0.98
jacobi2d -	0.077	0.14	0.27	0.52	0.68	0.79
pathfinder -	0.046	0.11	0.18	0.3	0.42	0.48
roi_align -	0.066	0.13	0.23	0.38	0.51	0.64
softmax -	0.034	0.068	0.11	0.19	0.29	0.35
4 8 16 32 64 128 Vector Length [elements]						

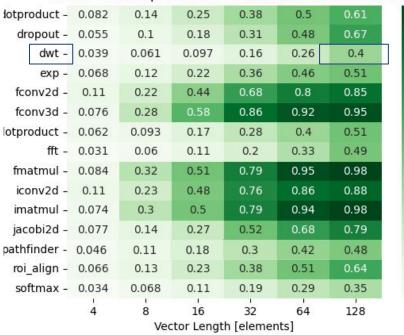


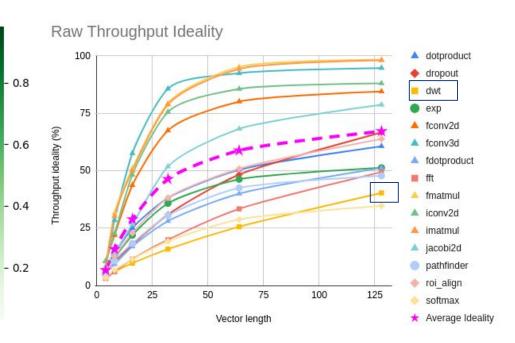


misaligned stride load - slow!

dwt - vsld (addr), vsld(addr + 1), compute

Relative kernel performance on maximum achievable

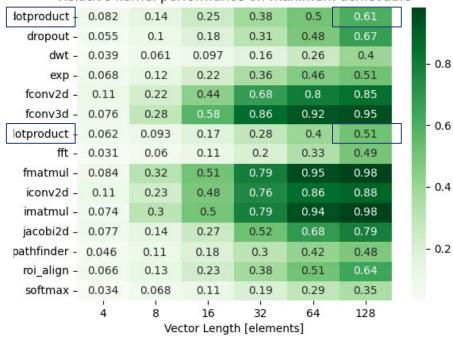


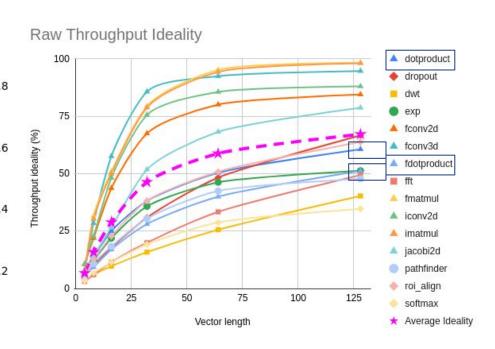




[f]dotproduct - vld, vld, vmul, vreduce

Relative kernel performance on maximum achievable



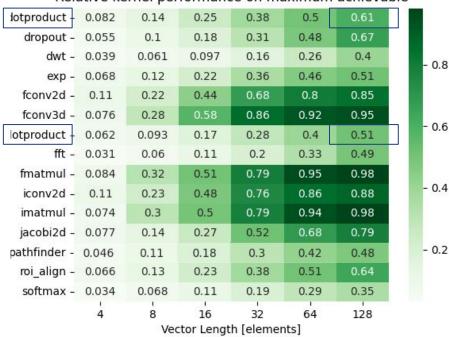


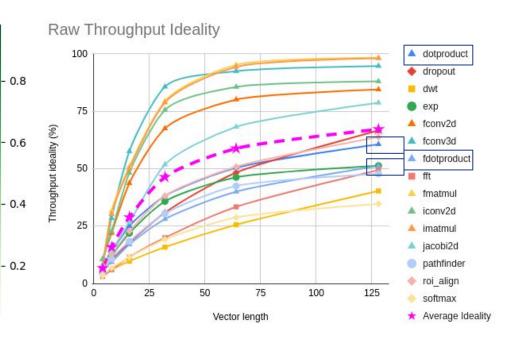


Cannot finish before the end of second vld

[f]dotproduct - vld, vld, vmul, vreduce

Relative kernel performance on maximum achievable



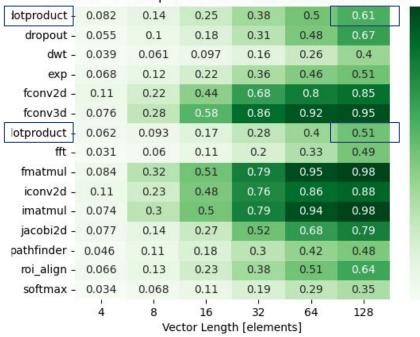


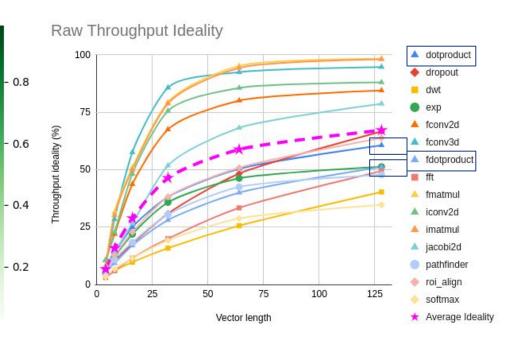


vreduce - structural hazard with vmul (fdotproduct)

[f]dotproduct - vld, vld, vmul, vreduce

Relative kernel performance on maximum achievable



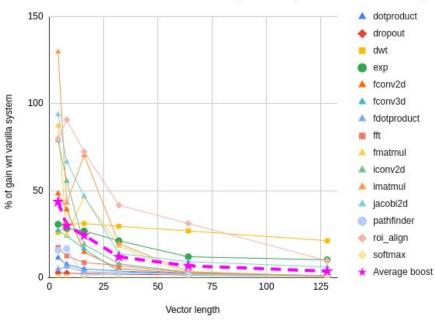


ETH Zürich | 9 |



Ideal Dispatcher boost

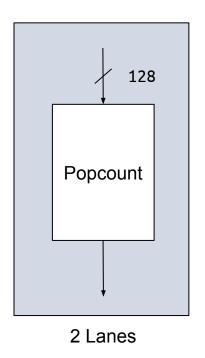


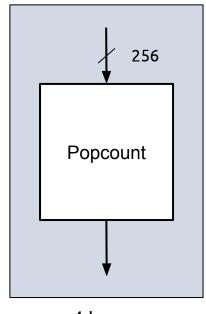


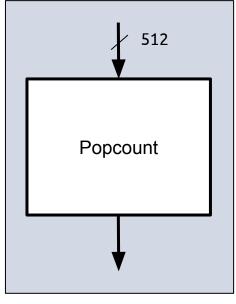
ETH Zürich | 10 |

Mask Unit - Popcount

Popcount in MASKU does not scale well







ETH Zürich

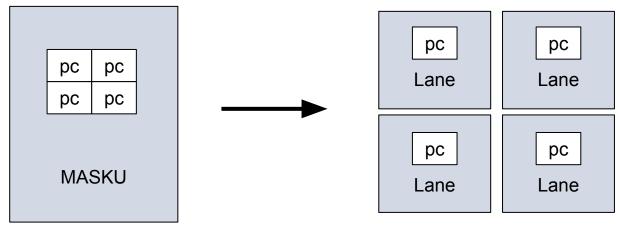
4 Lanes

8 Lanes

11

Mask Unit - Popcount

- Popcount in MASKU does not scale
- > Move popcount tree in the lanes?
 - × Huge HW modifications



Mask Unit - Popcount

- Popcount in MASKU does not scale
- Move popcount tree in the lanes?
 - Huge HW modifications
- ✓ Parametric implementation
 - In MASKU
 - Multi-cycle
 - Limited HW impact

Parameter: POPC BWIDTH = 64 bit 512 bit **Popcount**

ETH Zürich

12

HW - Compliance

RVV 1.0

- Recently implemented:
 - FP estimate reciprocal [sqrt]
 - FP rounding toward odd
- Missing instructions:
 - Three VRF shuffling instructions
 - Segment memory operations

HW - Implementation

Merged the last HW modifications

Implement 2, 4, 8, 16 lanes

Plug for Multi-Core runs (efficiency)

ETH Zürich | 15 |

HW - Implementation

fmatmul, 128x128x128

Lanes	Raw TP (OP/cycle)	TP (GOPS)	SS_f (MHz)	TT_f (GHz)	Power @TT_f (mW)	Efficiency (GOPS/W)
2	3.87	5.19	955.00	1.34	172.32	30.09
4	7.78	10.43	955.00	1.34	303.72	34.33
8	14.90	19.97	925.00	1.34	614.94	32.47

imatmul, 128x128x128

Lanes	Raw TP (OP/cycle)	TP (GOPS)	SS_f (MHz)	TT_f (GHz)
2	3.89	5.21	955.00	1.34
4	7.80	10.45	955.00	1.34
8	15.22	20.40	925.00	1.34

32b, 16b, 8b SIMD multipliers polluted the power results

HW - Implementation

fmatmul, 128x128x128

Lanes	Raw TP (OP/cycle)	TP (GOPS)	SS_f (MHz)	TT_f (GHz)	Power @TT_f (mW)	Efficiency (GOPS/W)
2	3.87	5.19	955.00	1.34	172.32	30.09
4	7.78	10.43	955.00	1.34	303.72	34.33
8	14.90	19.97	925.00	1.34	614.94	32.47

imatmul, 128x128x128

Lanes	Raw TP (OP/cycle)	TP (GOPS)	SS_f (MHz)	TT_f (GHz)	Power @TT_f (mW)	Efficiency (GOPS/W)
2	3.89	5.21	955.00	1.34	149.92	34.77
4	7.80	10.45	955.00	1.34	264.24	39.55
8	15.22	20.40	925.00	1.34	534.99	38.12

Extrapolated values!

Netlist with gated SIMD-Multipliers soon