Update on Ara

24/01/2023

Matteo Perotti
Matheus Cavalcante

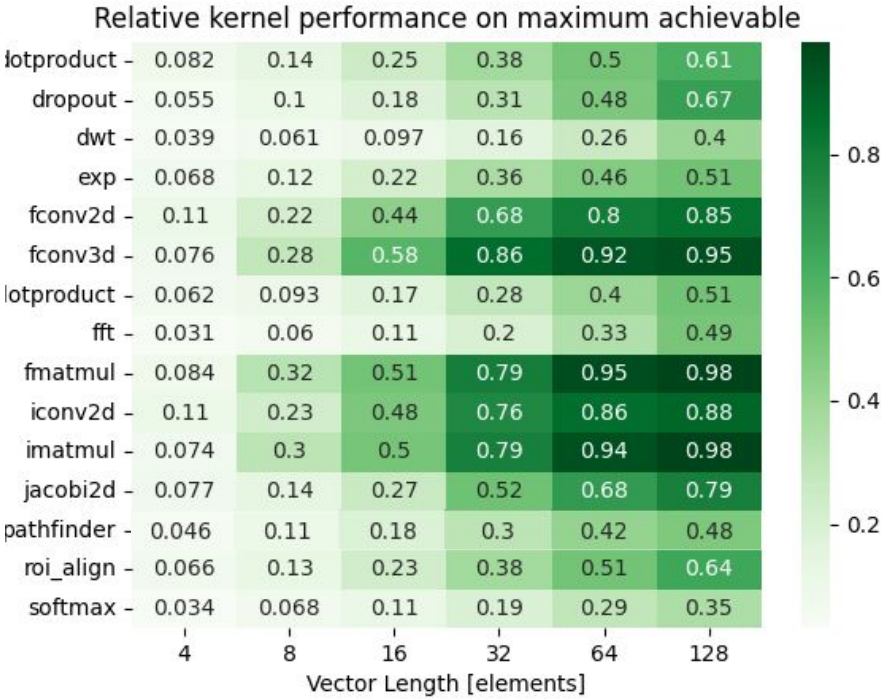
Professor Luca Benini
Integrated Systems Laboratory
ETH Zürich

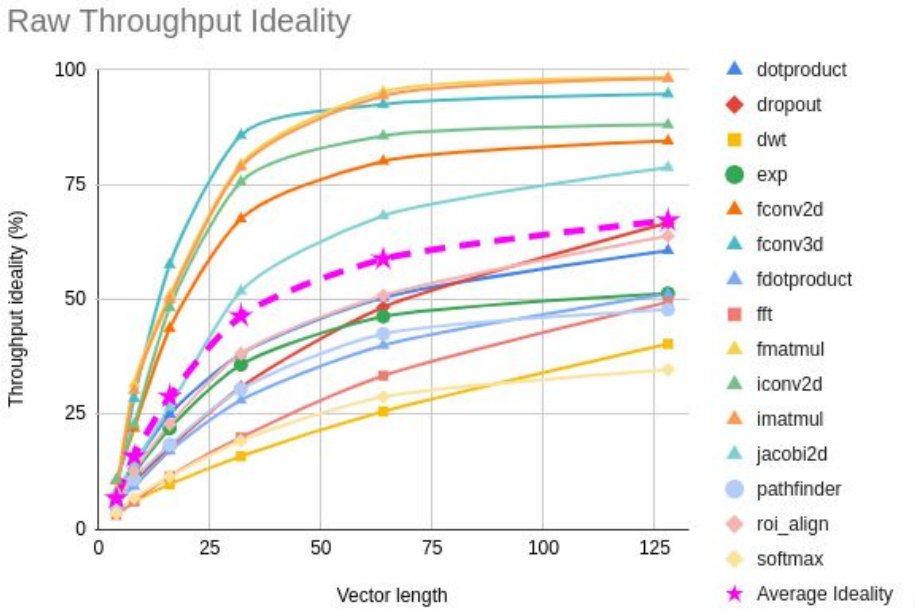
Summary

- **SW**
  - Benchmark results

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

SW performance

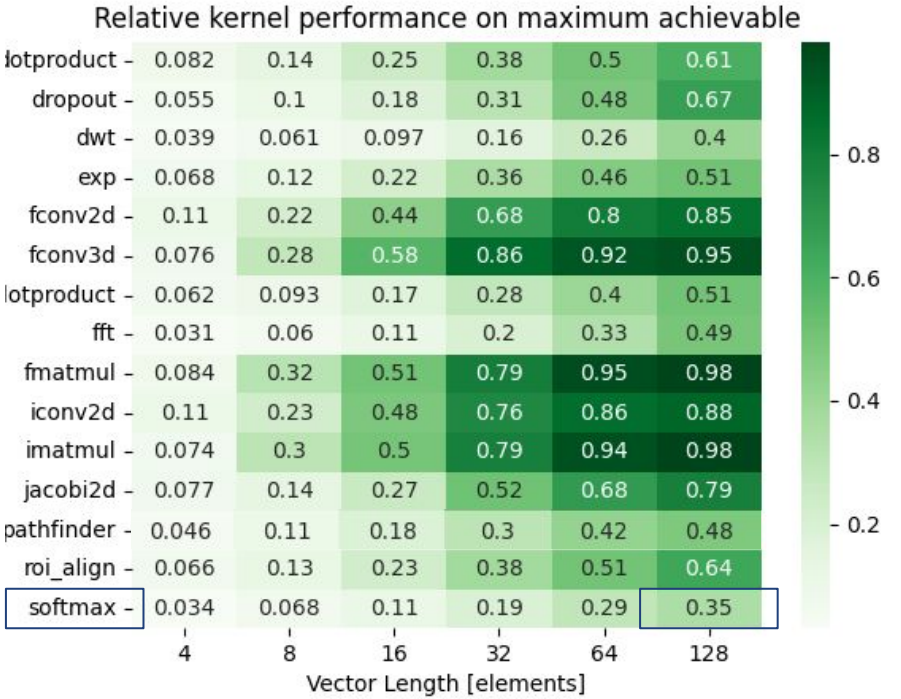


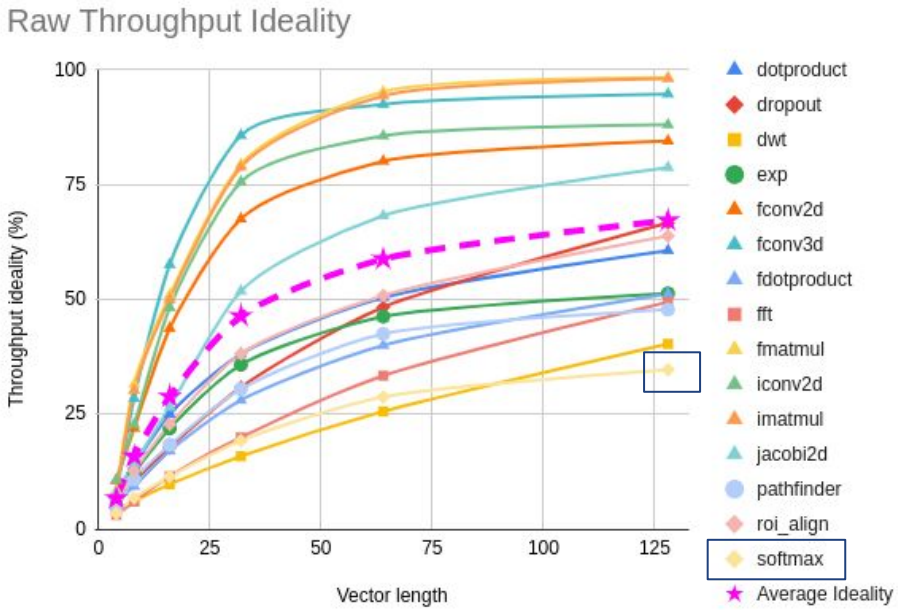


SW performance

jump to function? + heavy function state preparation

`softmax` - … vexp() … vdiv …

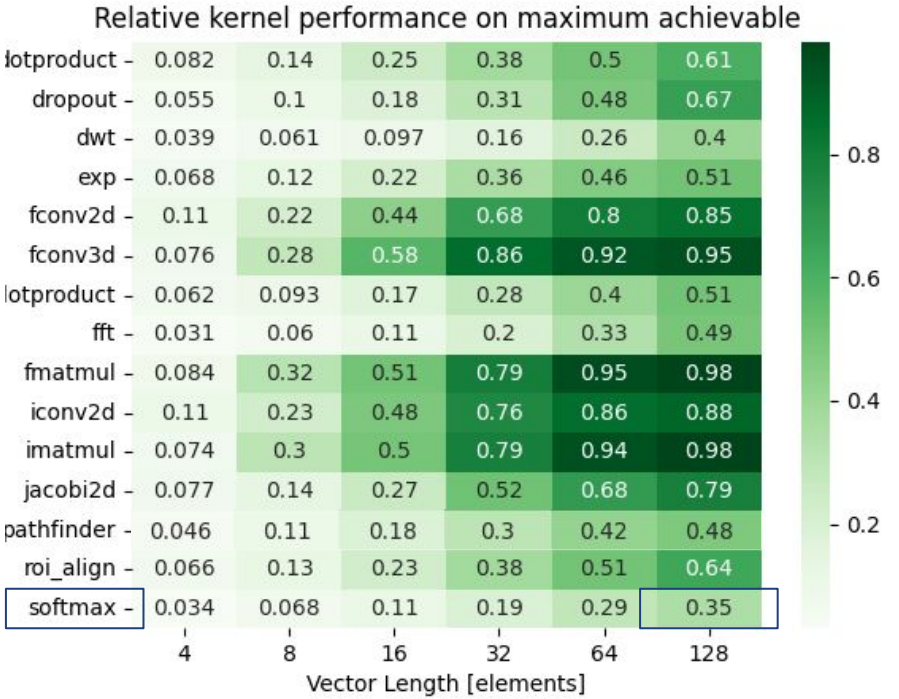


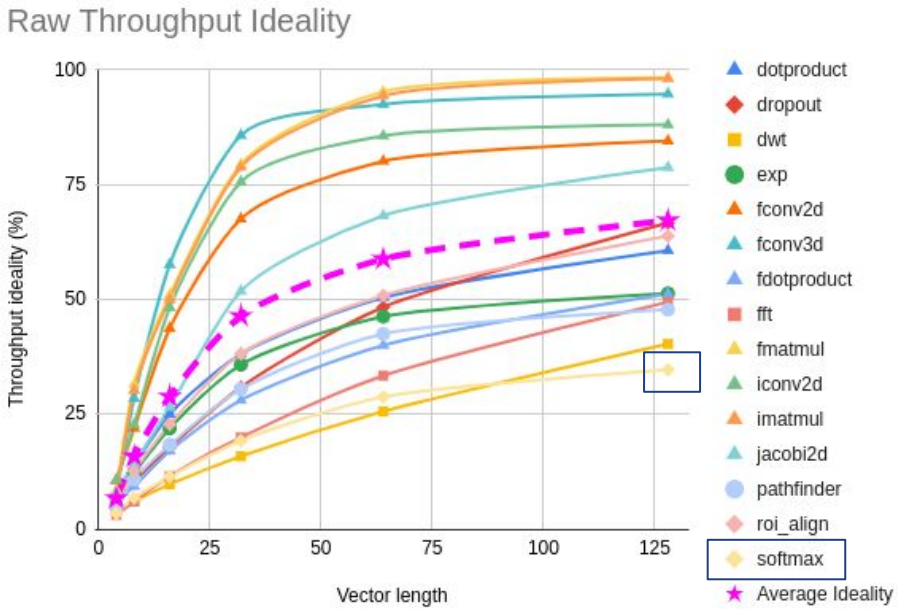


SW performance

vdiv is way slower than other fp operations

`softmax` - … vexp() . . vdiv ..

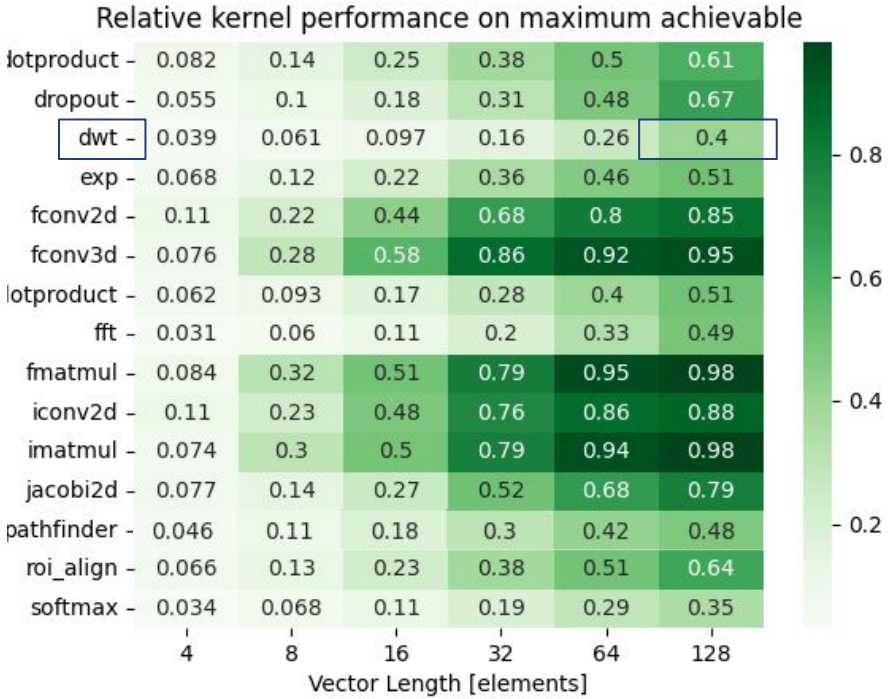


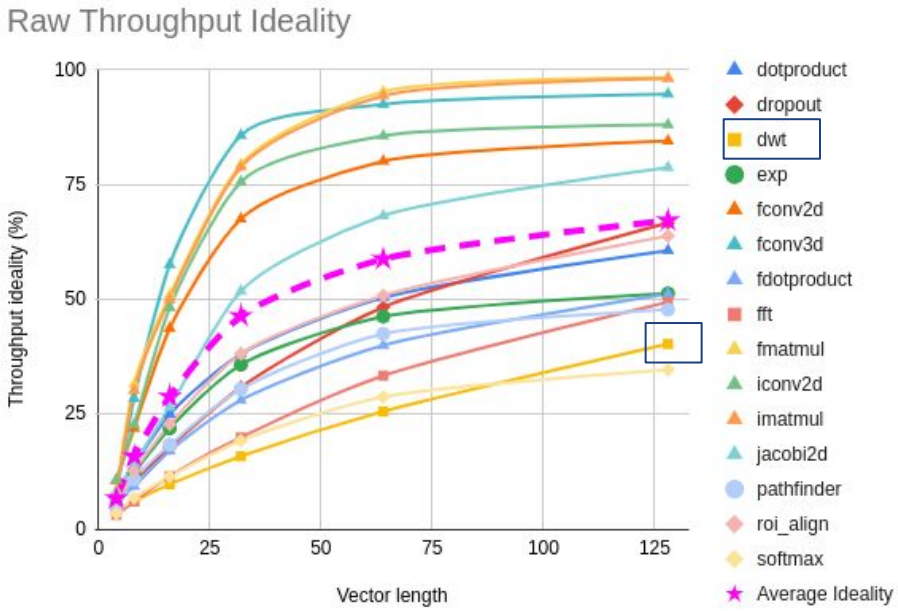


SW performance

misaligned stride load - slow!

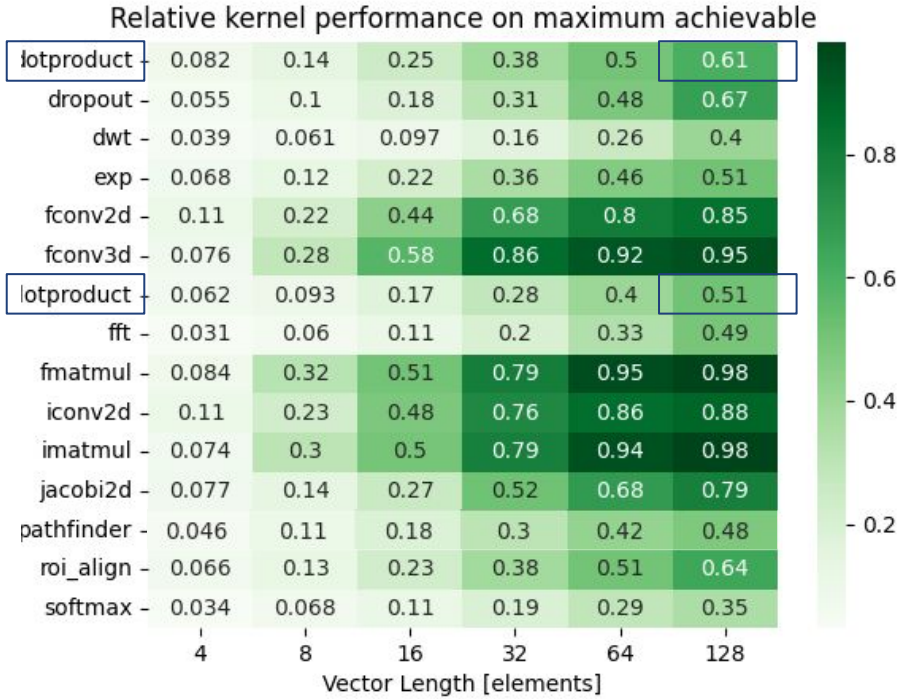
`dwt - vsld (addr), vsld(addr + 1), compute`

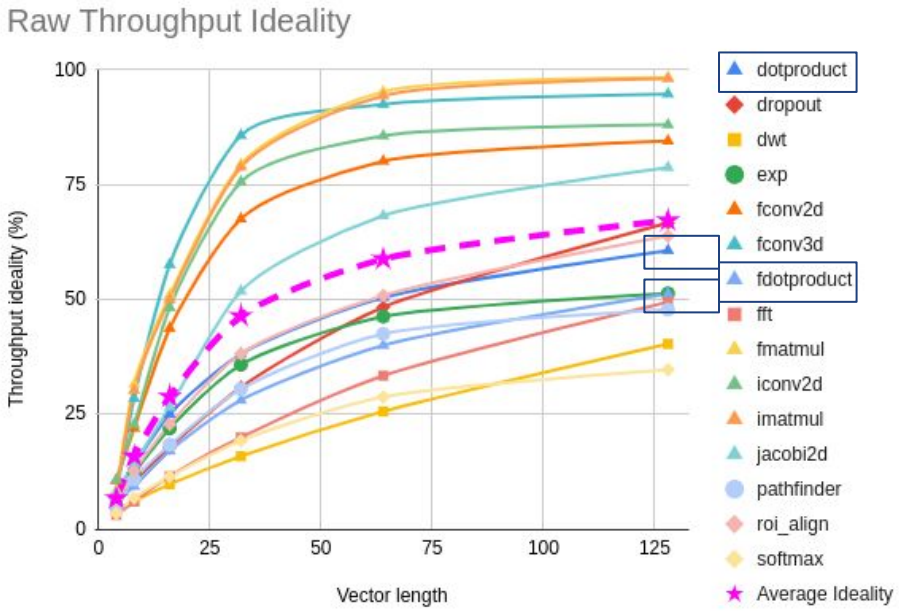




SW performance

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

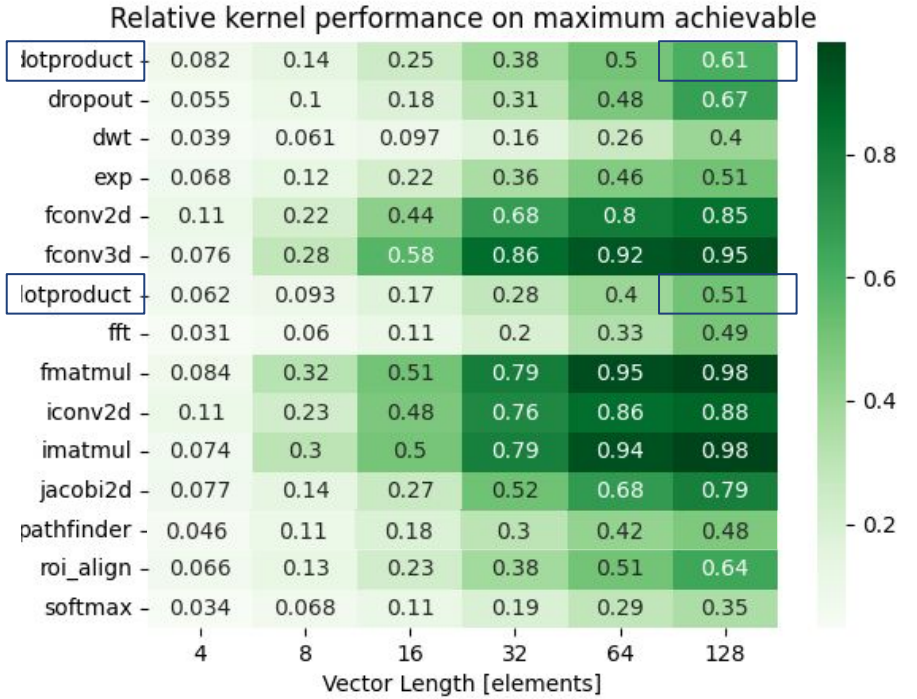


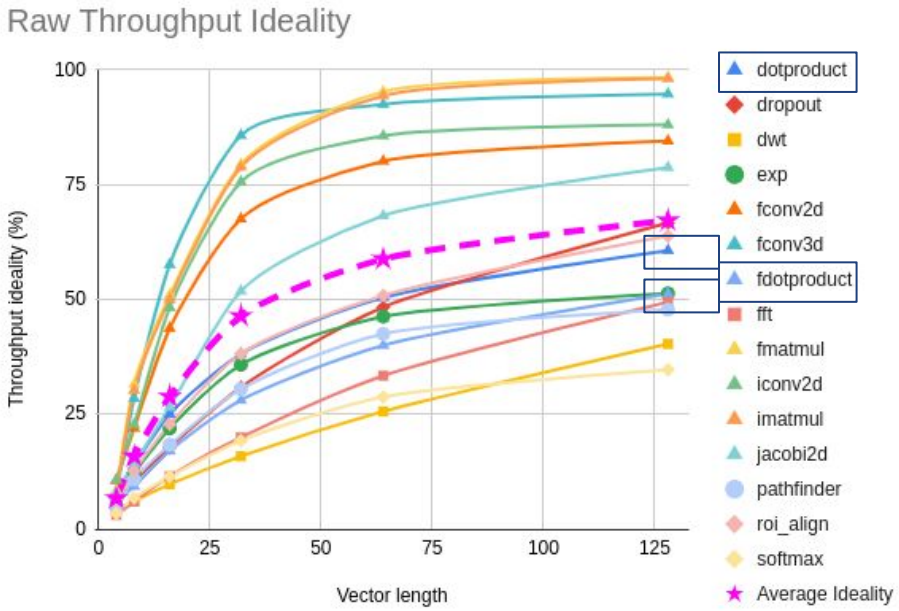


SW performance

Cannot finish before the end of second vld

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

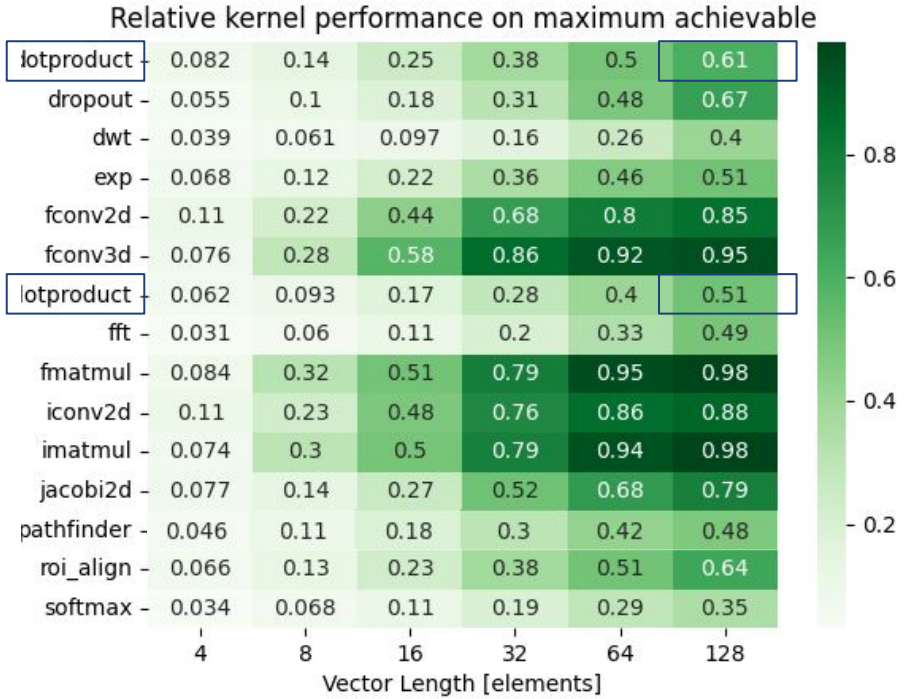


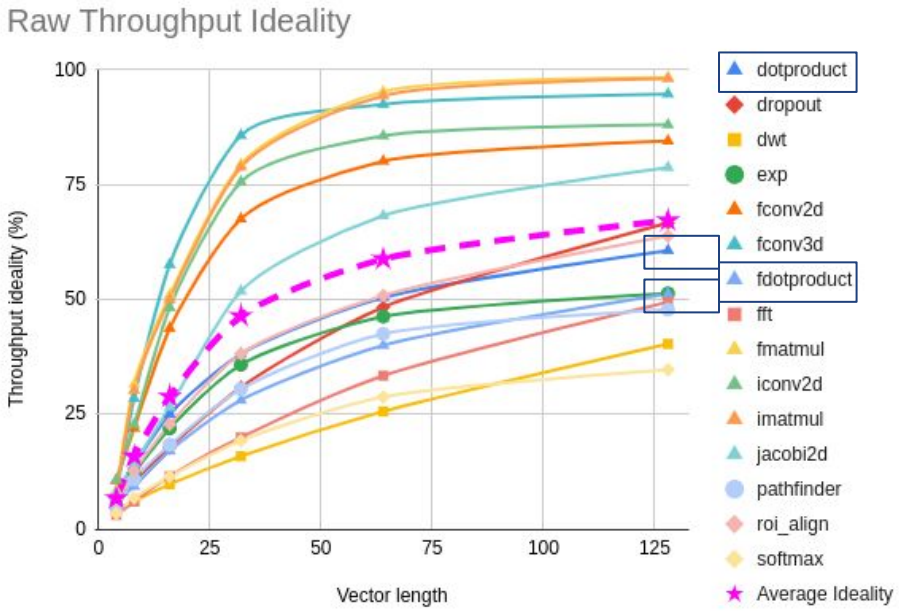


SW performance

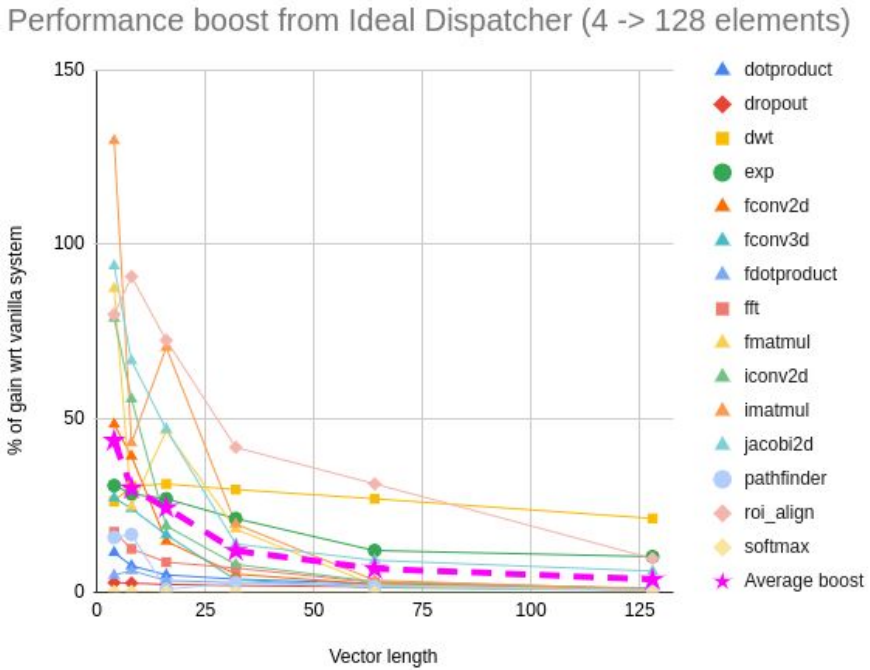
vreduce - structural hazard with vmul (fdotproduct)

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



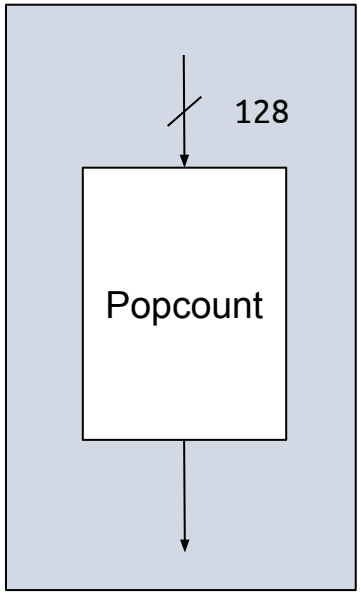


Ideal Dispatcher boost

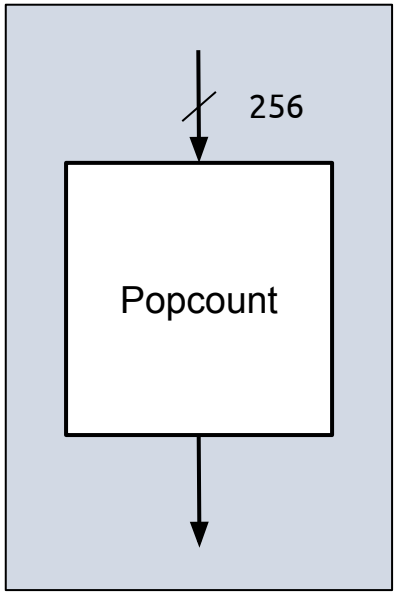


Mask Unit - Popcount

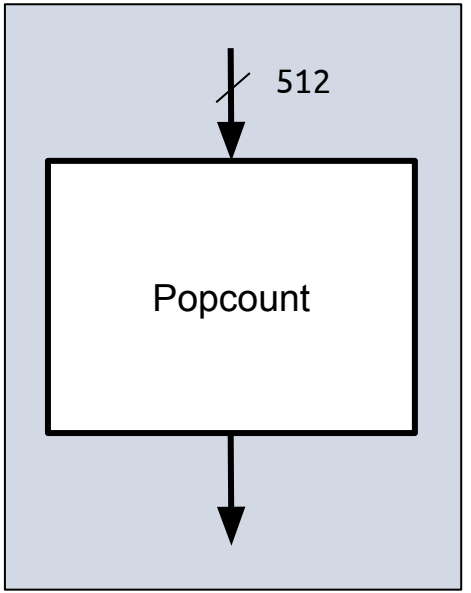
- **Popcount** in **MASKU does not scale well**



2 Lanes



4 Lanes



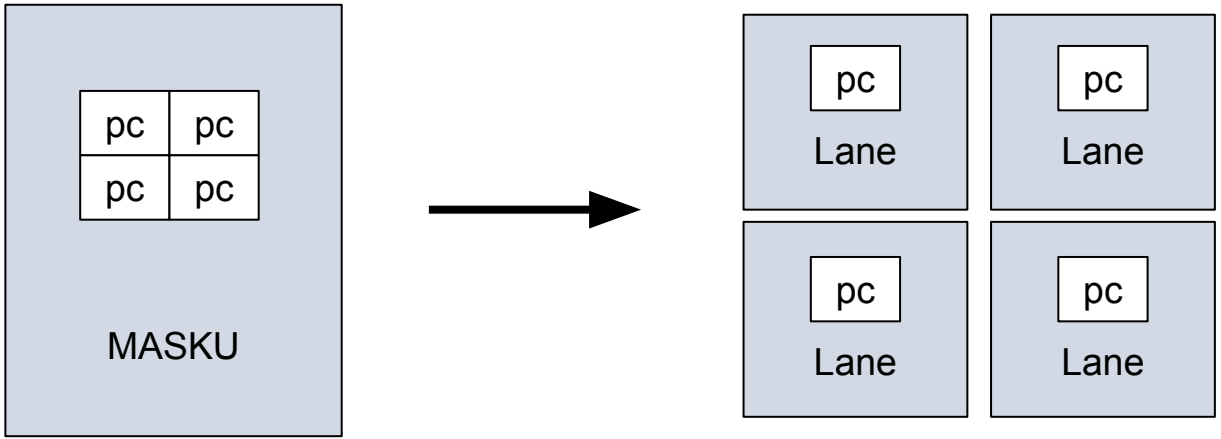
8 Lanes

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**

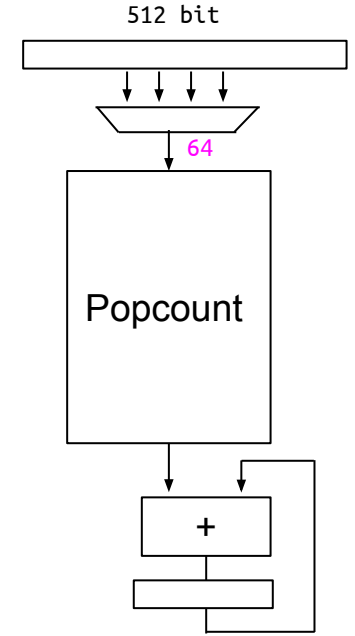
➢ **Move popcount** tree **in** the **lanes?**

✗ Huge HW modifications

✓ **Parametric** implementation

- **In MASKU**
- **Multi-cycle**
- **Limited HW impact**

Parameter:
POPC_BWIDTH = 64 bit



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)

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