Update on Ara

13/10/2021

Matteo Perotti
Matheus Cavalcante
Nils Wistoff

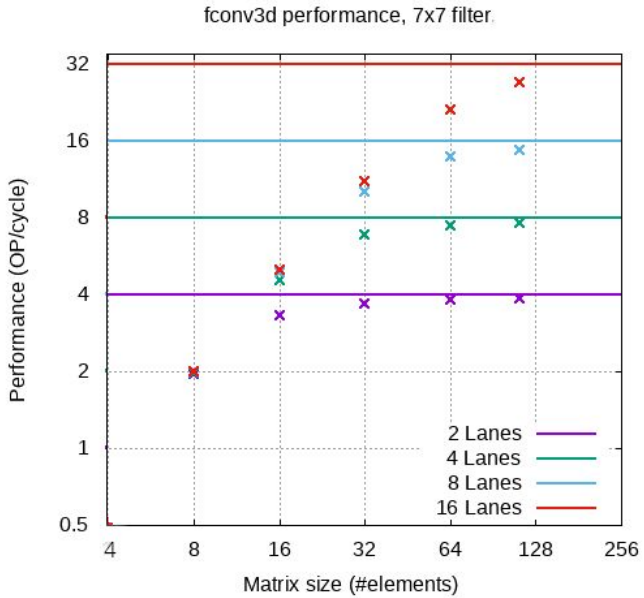
Professor Luca Benini
Integrated Systems Laboratory
ETH Zürich

Summary

- **Comparison** with **AraV1**
  - Kernels TP
  - PPA + Efficiency
- **SW** optimization
  - Convolutions on small matrices
- **HW** optimization
  - Timing

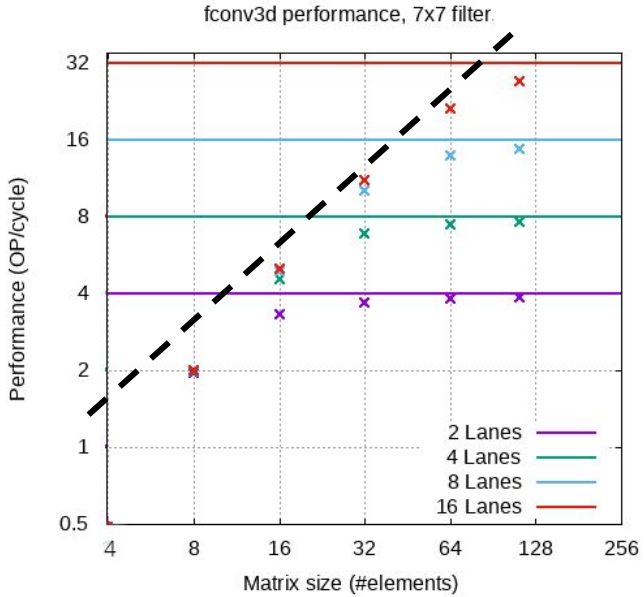
Convolutions on small matrices

- Convolution, 7x7 filter, 3 channels



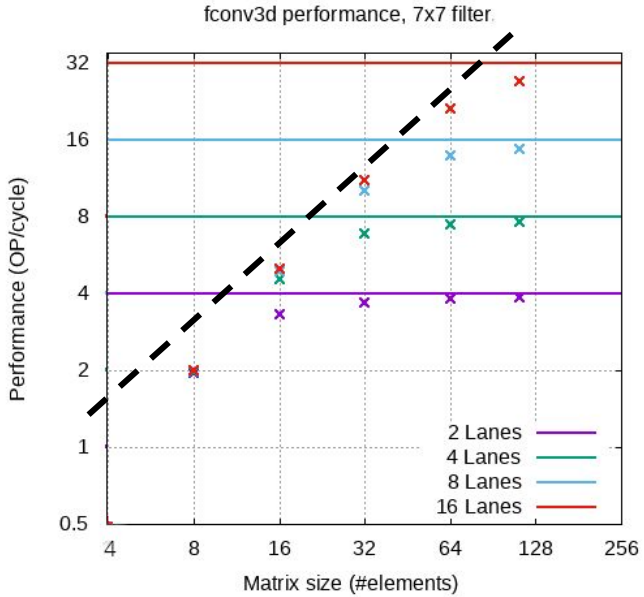
Convolutions on small matrices

- Convolution, 7x7 filter, 3 channels



Convolutions on small matrices

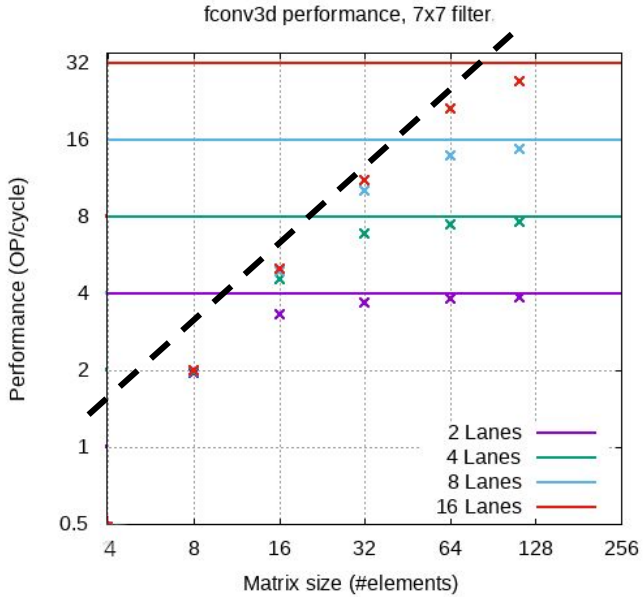
- Convolution, 7x7 filter, 3 channels
- Arithmetic intensity is always > 0.5 DPFLOP/B



Convolutions on small matrices

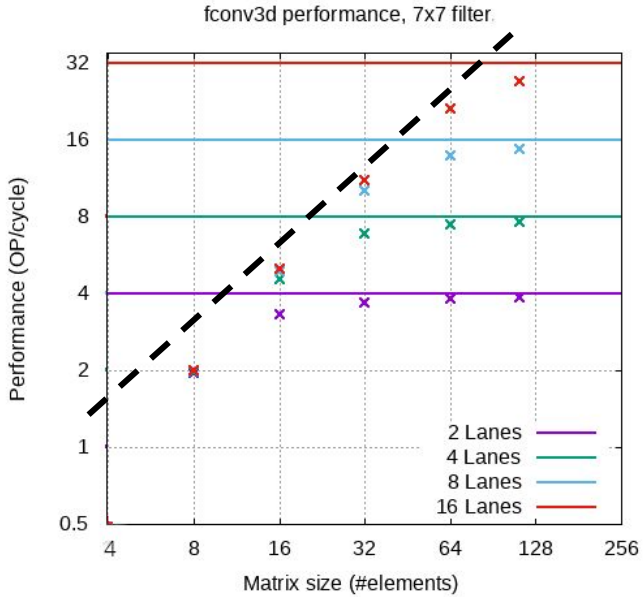
- Convolution, 7x7 filter, 3 channels
- Arithmetic intensity is always > 0.5 DPFLOP/B

Ara computation-bound region



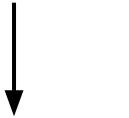
Convolutions on small matrices

- Convolution, 7x7 filter, 3 channels
- Arithmetic intensity is always > 0.5 DPFLOP/B
- Ara memory BW does not limit performance

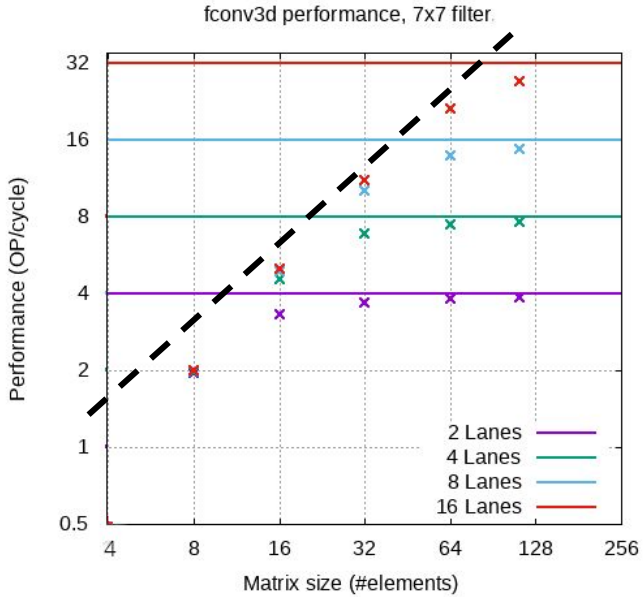


Convolutions on small matrices

- Convolution, 7x7 filter, 3 channels
- Arithmetic intensity is always > 0.5 DPFLOP/B
- Ara memory BW does not limit performance
- Issue rate problem?



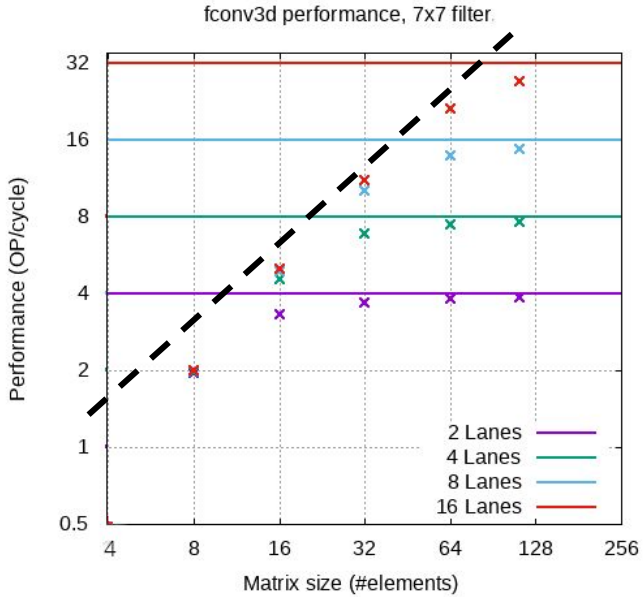
Investigation + optimization



Convolutions on small matrices

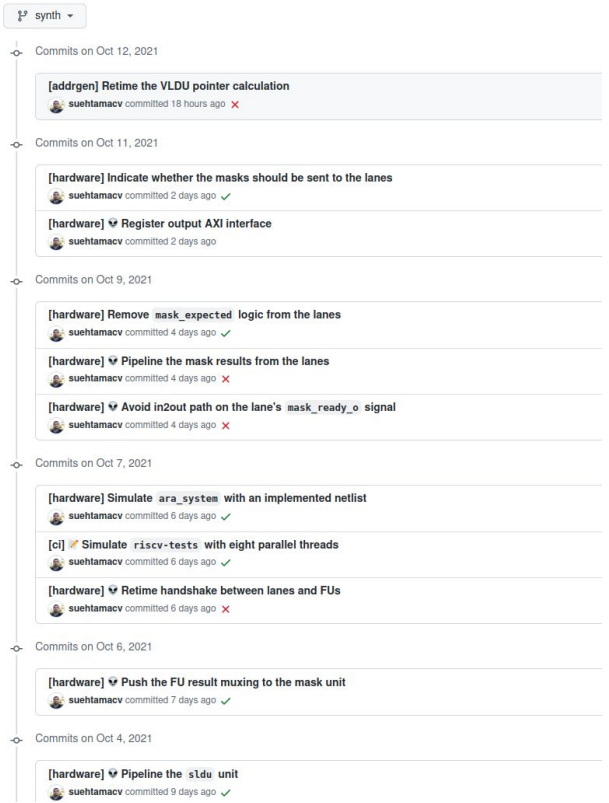
Information:

- Linear dependency w.r.t. the matrix size
- The performance limit due to the issue rate is linearly dependent on the matrix size
- This is a problem for short vectors, as the issue latency is not amortized
- A vector is *short* depending on the size of the vector and the number of lanes



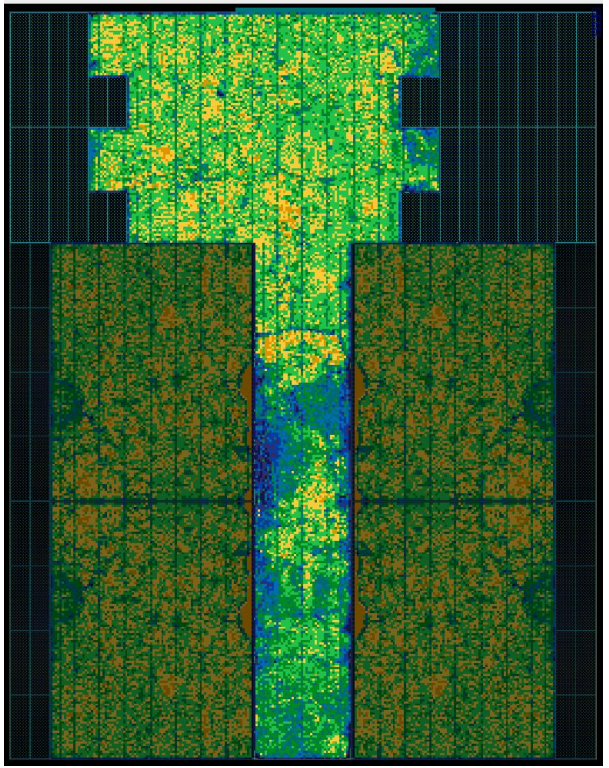
Ara's implementation in GF 22FDX

- Currently implementing **Ara's system** (Ara, Ariane, AXI mux) on 22FDX
  - Synopsys-based flow
- Lanes reach post-layout 1 GHz target without major issues
- Top-level suffers from *in2out* paths at the lane interface
  - About 10 paths that limit performance
  - Fixing them on GitHub



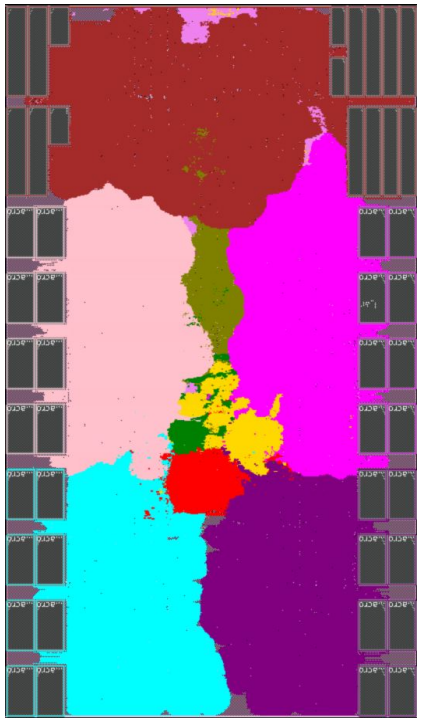
Some interesting details about the design

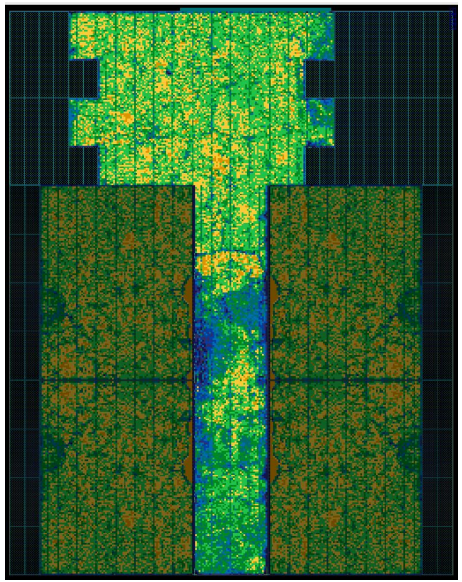
- We are targeting a high cell density
  - The lanes achieve a cell density of 75%
  - The top-level system achieves a cell density of 72%
  - No "empty regions" in the floorplan
- Current size: 0.83 x 1.06 mm
- The system can be clocked at 980 MHz (wc)
  - Potential to be pushed even further, by solving some critical paths on the Ara-Ariane interface



AraV2 is 10% smaller than the original AraV1 implementation

AraV1 area:
0.75 x 1.30 mm²





AraV2 area:
0.830 x 1.059 mm²