ETH zürich

Update on Ara

23/11/2022

Matteo Perotti

Matheus Cavalcante

Professor Luca Benini

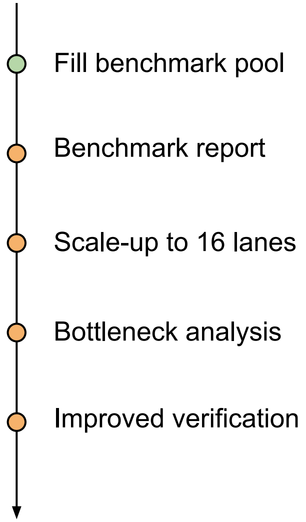
Integrated Systems Laboratory

ETH Zürich

Summary

- **Software**
  - New benchmarks upstream
  - Stall analysis

- **Hardware (RTL + Backend)**
  - 8-lanes trials
  - Mask instructions support
  - Multi-Core



Fill benchmark pool

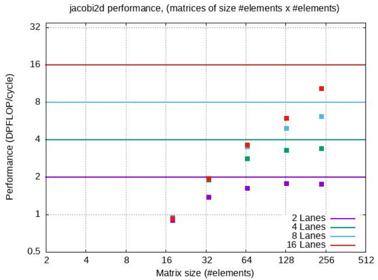
Benchmark report

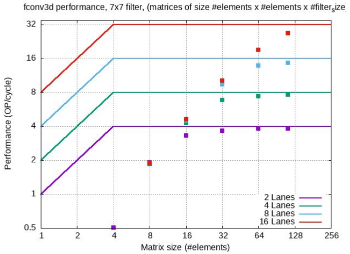
Scale-up to 16 lanes

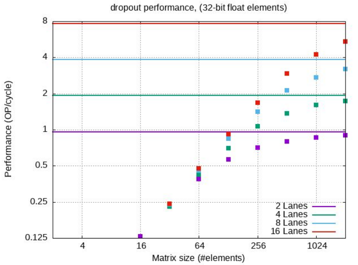
Bottleneck analysis

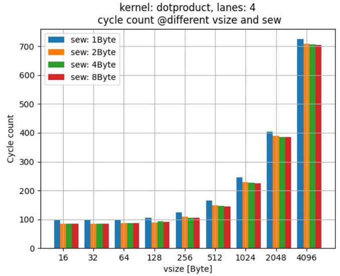
Improved verification

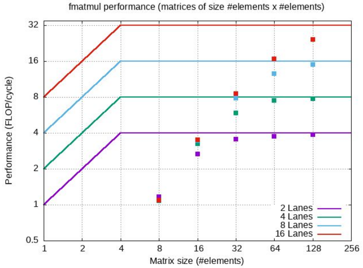
Longer vectors perform better

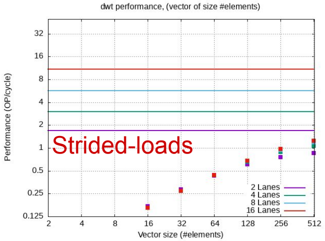




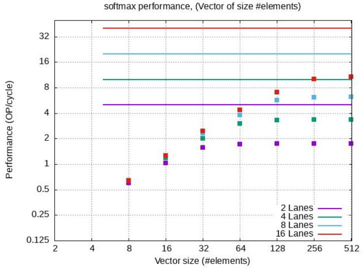


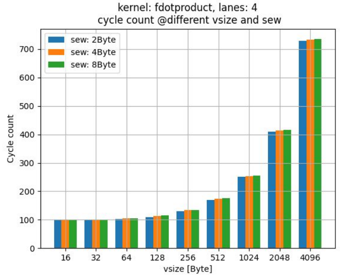


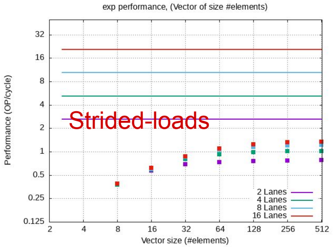




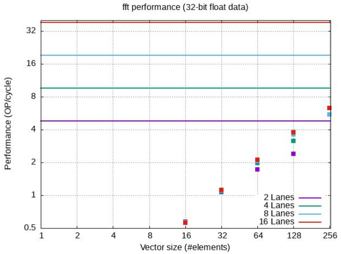
Strided-loads

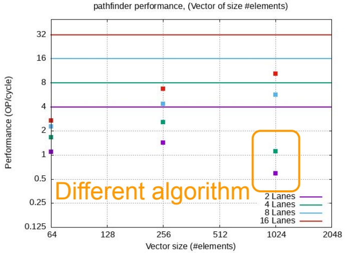




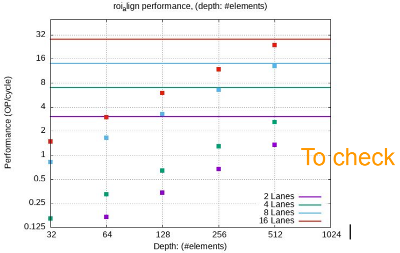


Strided-loads





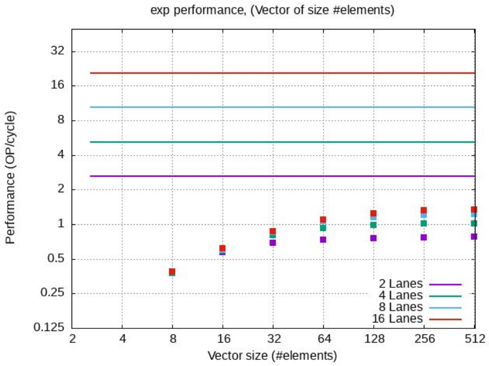
Different algorithm



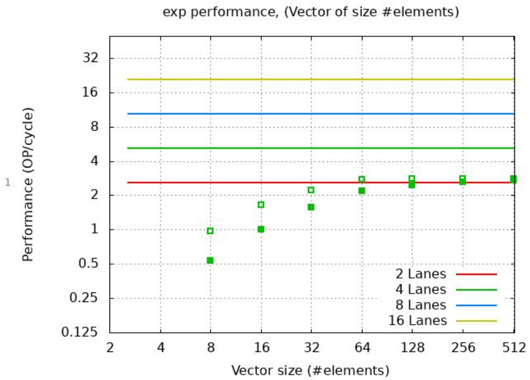
To check

Benchmarks

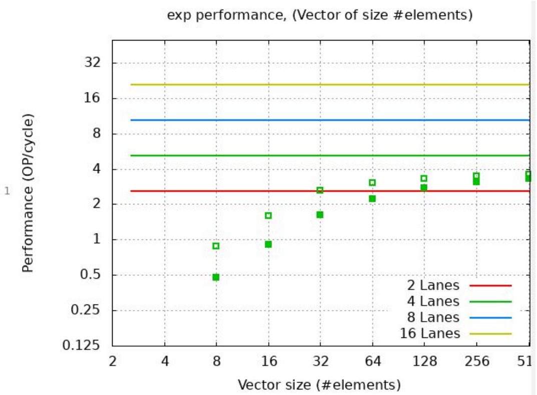
- Vector intrinsics
- How architectural choices reflect on programming model
- Pitfalls and coding guidelines



Naive intrinsics



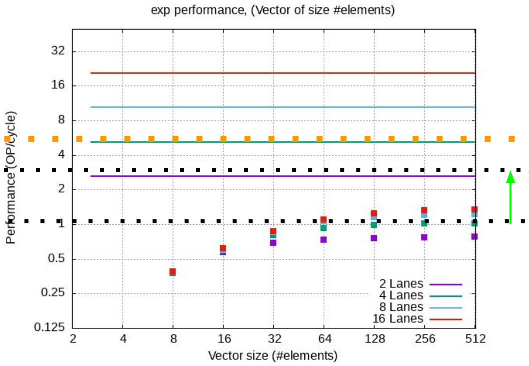
Optimized intrinsics



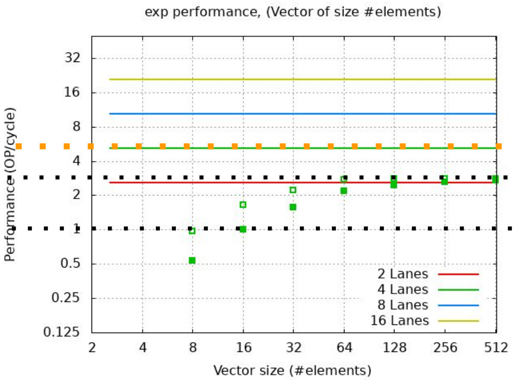
Optimized ASM

Benchmarks

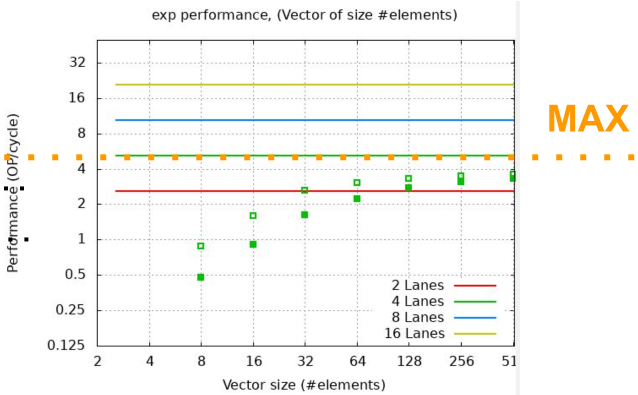
- Vector intrinsics
- How architectural choices reflect on programming model
- Pitfalls and coding guidelines



Naive intrinsics



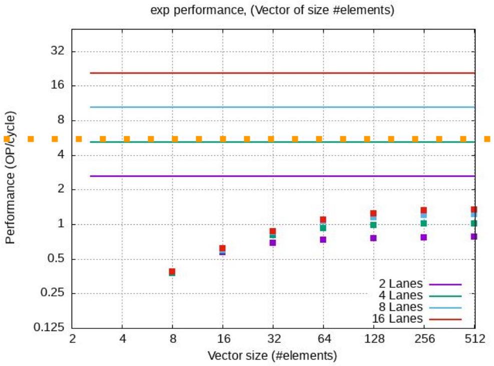
Optimized intrinsics



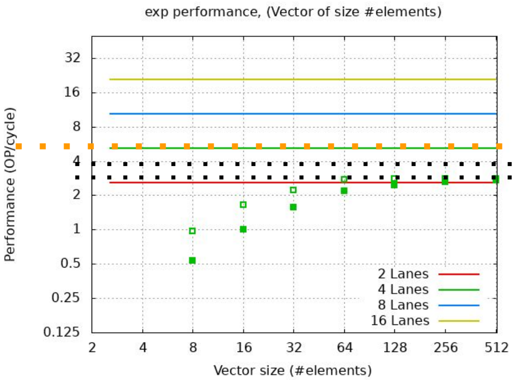
Optimized ASM

Benchmarks

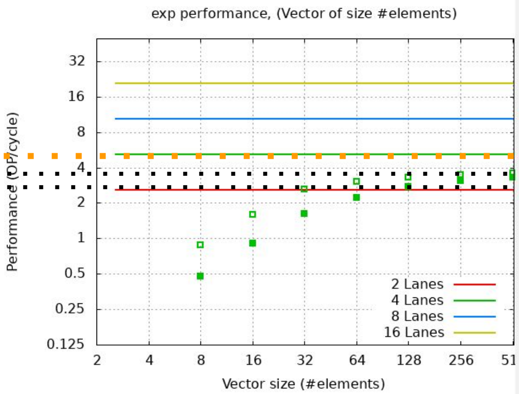
- Vector intrinsics
- How architectural choices reflect on programming model
- Pitfalls and coding guidelines



Naive intrinsics



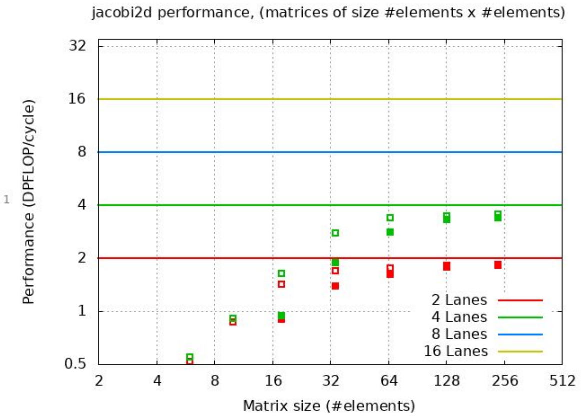
Optimized intrinsics

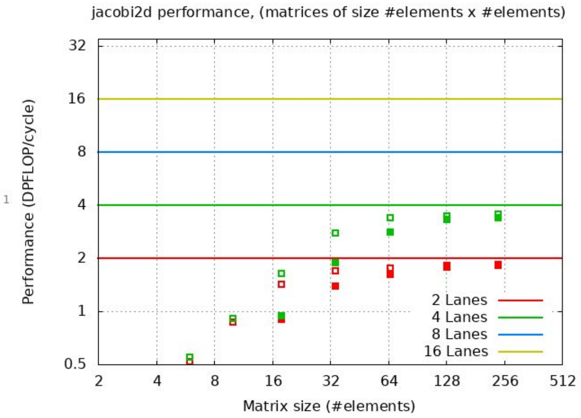


Optimized ASM

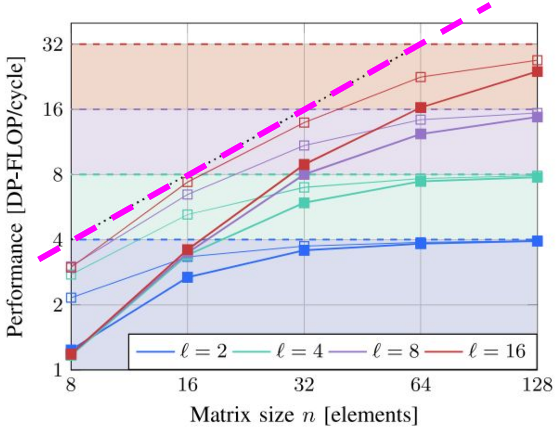
MAX

Scalar core problem?





Issue rate limitation only for real system!

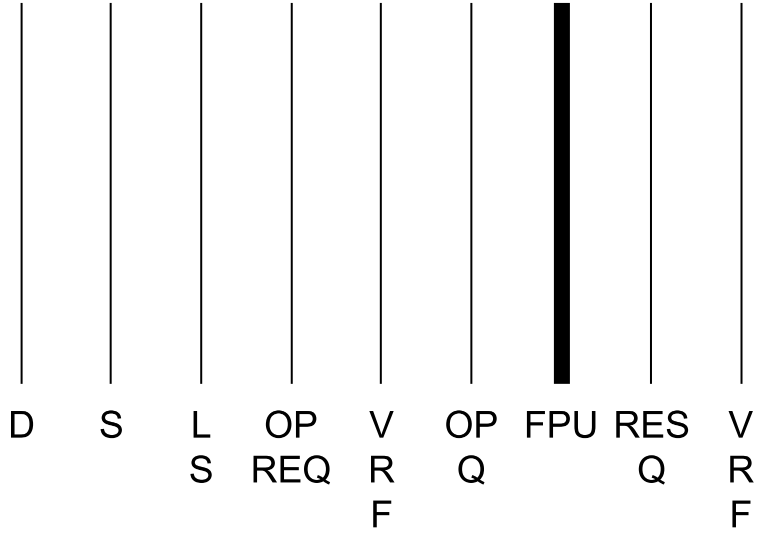


□ Ideal Dispatcher + Ara

■ CVA6 + Ara

- **Ideal dispatcher analysis + Tune miss rate/penalty**
- CVA6 + Scalar Mem Sys bottlenecks the gains
  - Many in-Ara improvements are hidden by CVA6
- **Better scalar core** does **not** fully **solve** the **problems**
- Should **repeat analysis after internal optimization** with ideal dispatcher (barber's pole, new hazard handling engine)

Ara problem?



D

S

LS

OP REQ

VRF

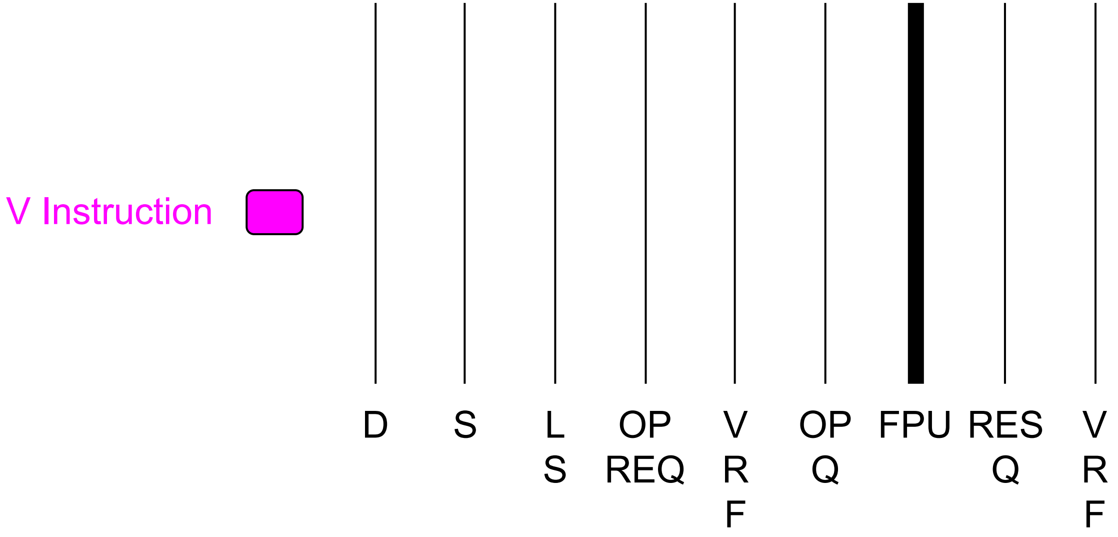
OP Q

FPU

RES Q

VRF

Ara problem?



V Instruction

D

S

L
S

OP
REQ

V
R
F

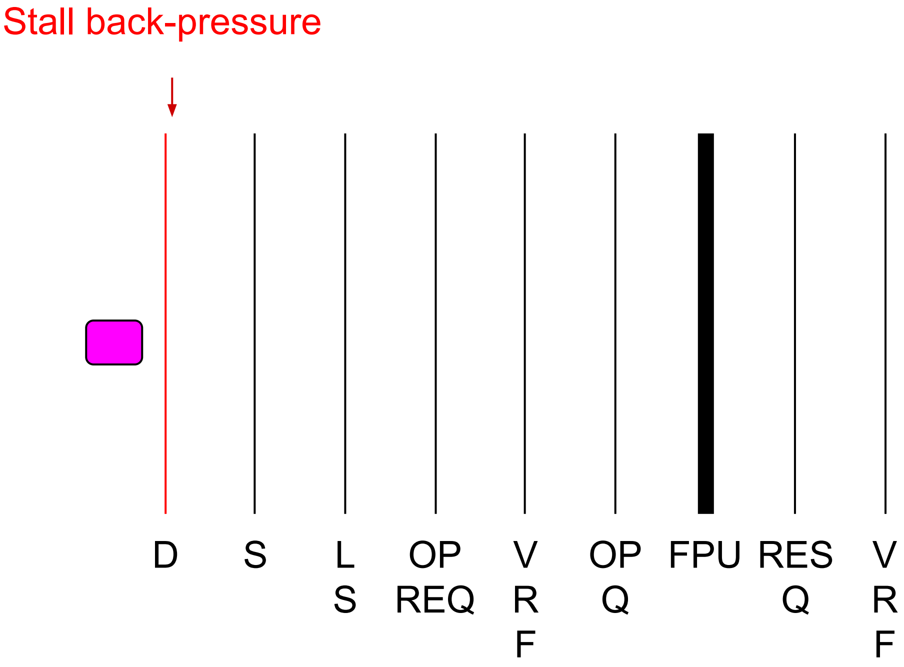
OP
Q

FPU

RES
Q

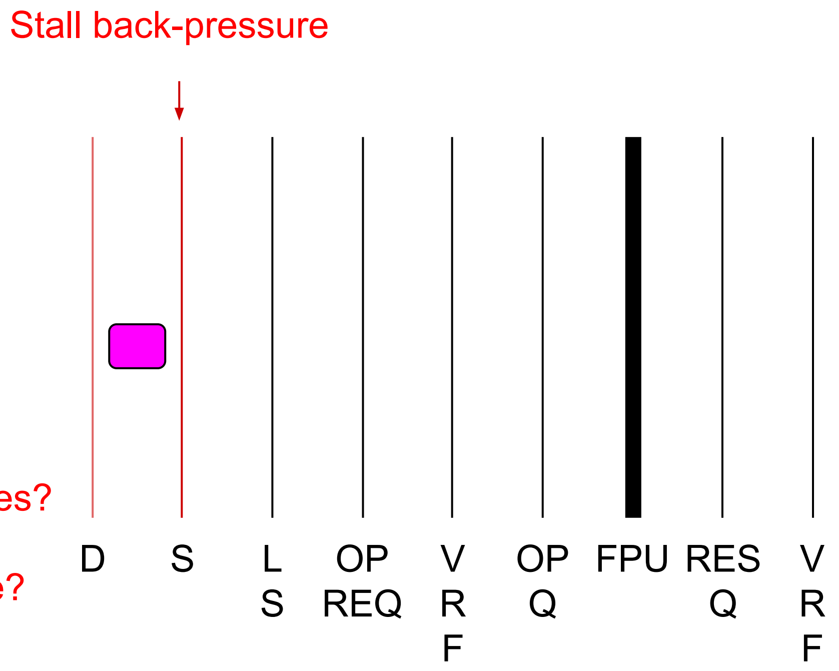
V
R
F

Ara problem?



1. Lower LMUL?

Ara problem?

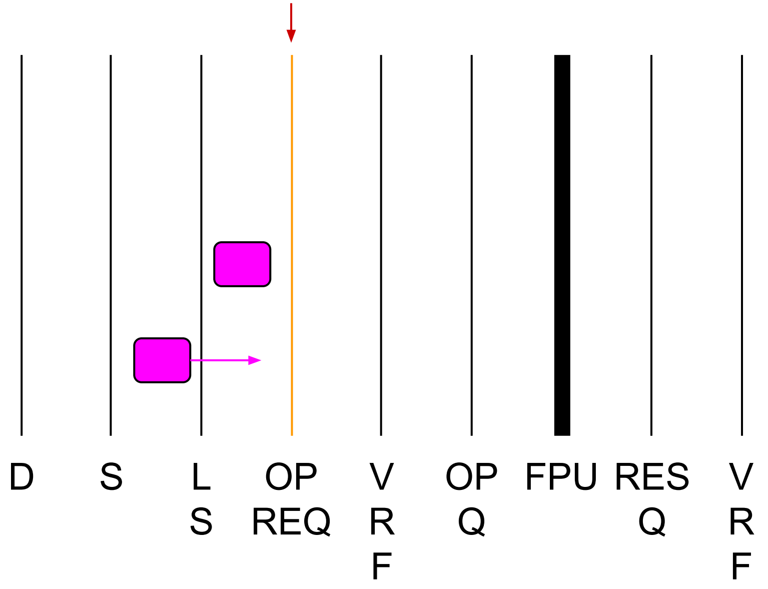


1. Downstream full queues?
2. Slide stall?
3. WAR, WAW no source?
4. Exception check?
5. Feedback answer?
6. Back pressure from next stage?

Crucial stalls for short vectors!

Ara problem?

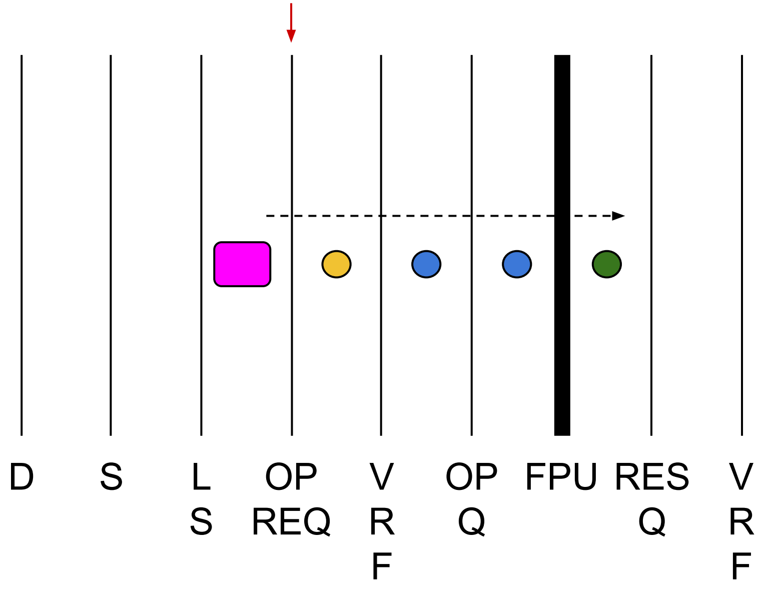
Stall **without** back-pressure for != units



1. Data hazards?
2. Arbitration?
3. Bank conflict?

Ara problem?

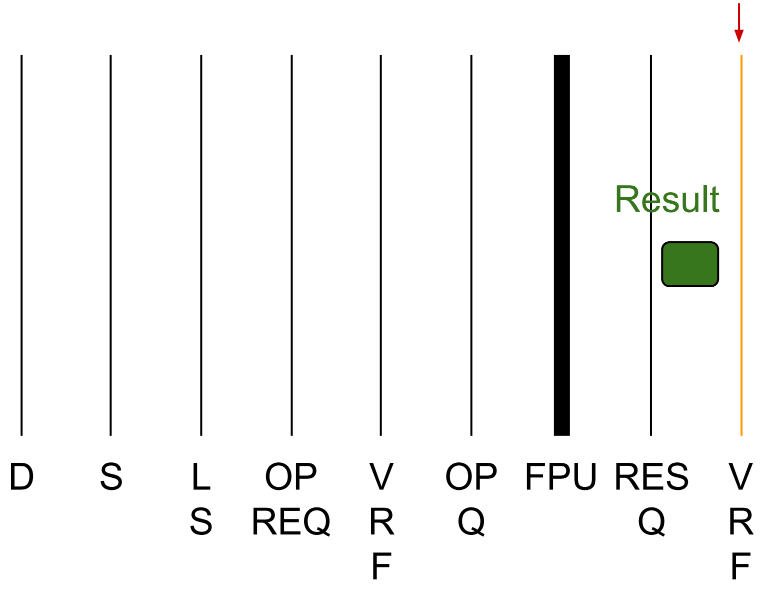
Stall **without** back-pressure for != units



From here on, the vector length help hide stalls from behind!

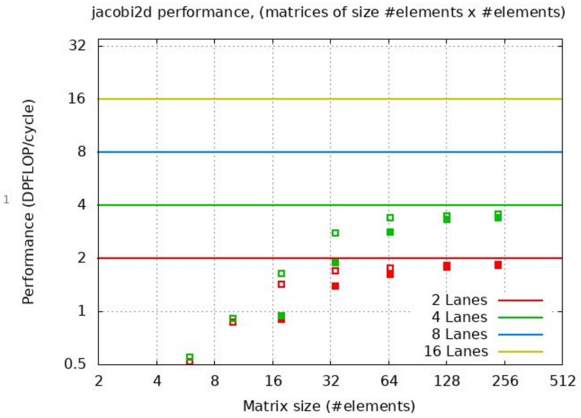
Ara problem?

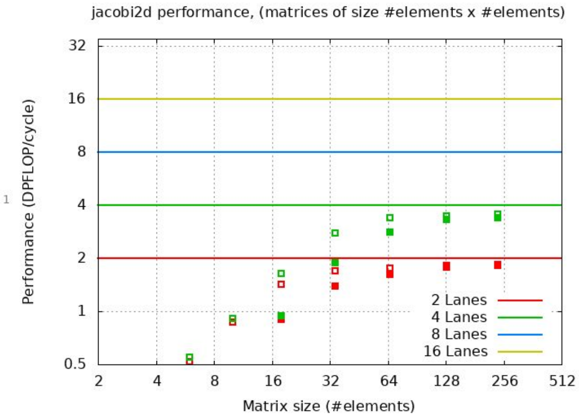
Stall **without** back-pressure for != units



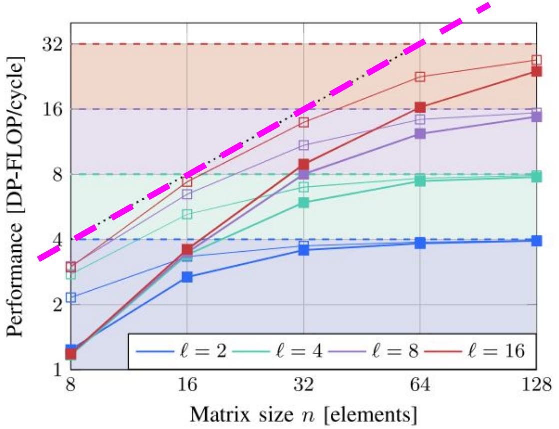
1. Bank conflict?

Ara problem?





Issue rate limitation only for real system!



□ Ideal Dispatcher + Ara

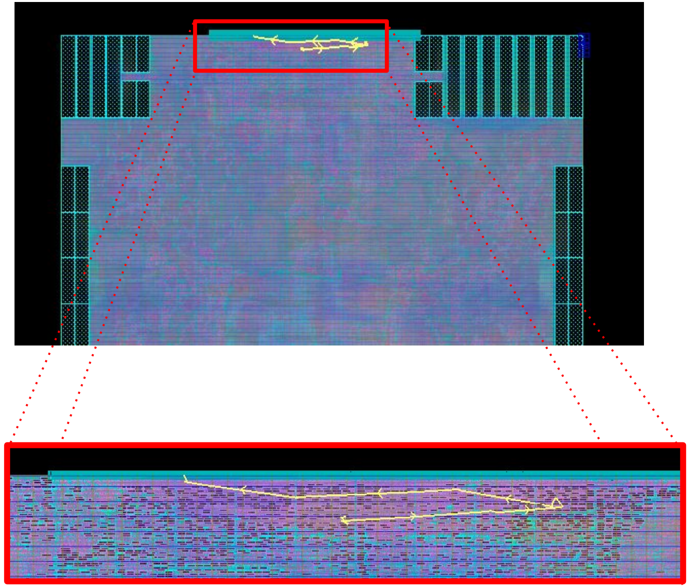
■ CVA6 + Ara

- CVA6 bottlenecks the gains
  - Many in-Ara improvements are hidden by CVA6
- **Do we need renaming? WAW-WAR analysis**
- **Why short-vectors are a problem? Stall analysis**
- **Diminishing returns** to improve short vector perf (larger buffers)
- Should **repeat analysis after internal optimization** with ideal dispatcher (barber’s pole, new hazard handling engine)

Scale-up problem

- **HW**:
  8 lanes WIP for full closure
  16 lanes is almost infeasable:
  - `With    SLDU - Infeasable`
  - `Without SLDU - WNS: -400 ps`

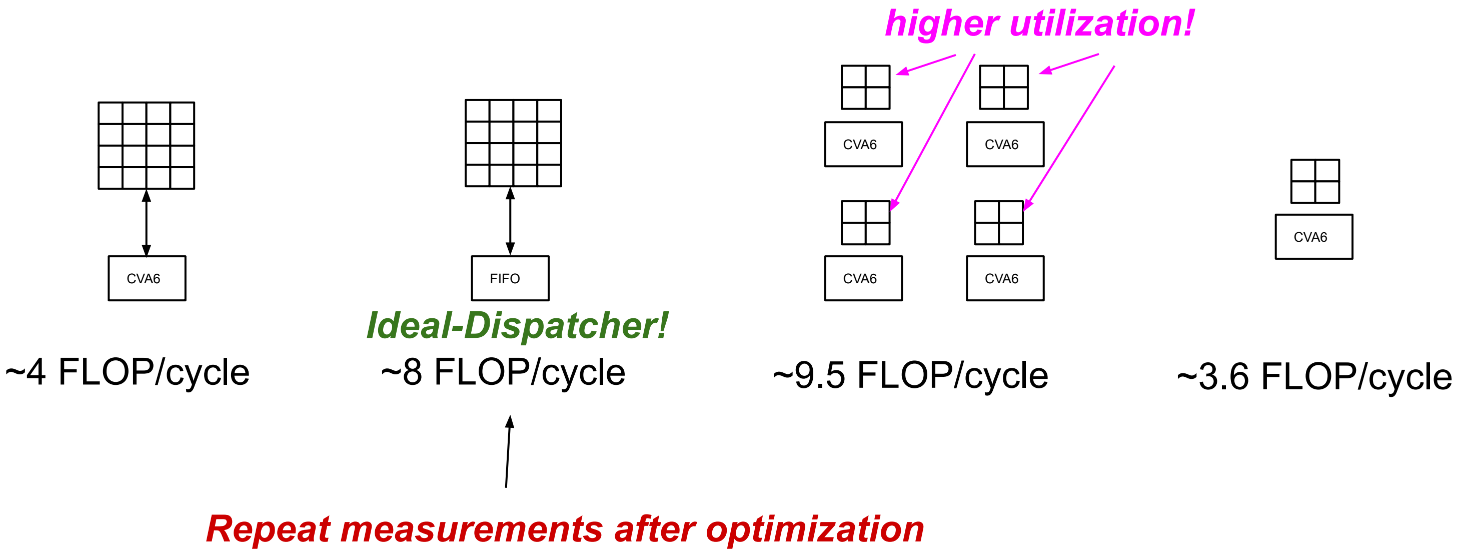
- **SW**:
  Efficient use of the resources:
  - Longer vectors
  - Hard to partition a problem!



Vector Multi-Core

- ✓ **Partition** the problem
- ✓ **Utilization is higher** since vectors are “longer”
- ✓ **Scale-up** in terms of **FPUs**
- ➢ **Should repeat the analysis after optimization**
- ✗ **Increased CVA6 traffic to upper level of memory**

Vector Multi-Core (8x8 `fmatmul`)



higher utilization!

CVA6

CVA6

CVA6

CVA6

CVA6

CVA6

FIFO

Ideal-Dispatcher!

~4 FLOP/cycle

~8 FLOP/cycle

~9.5 FLOP/cycle

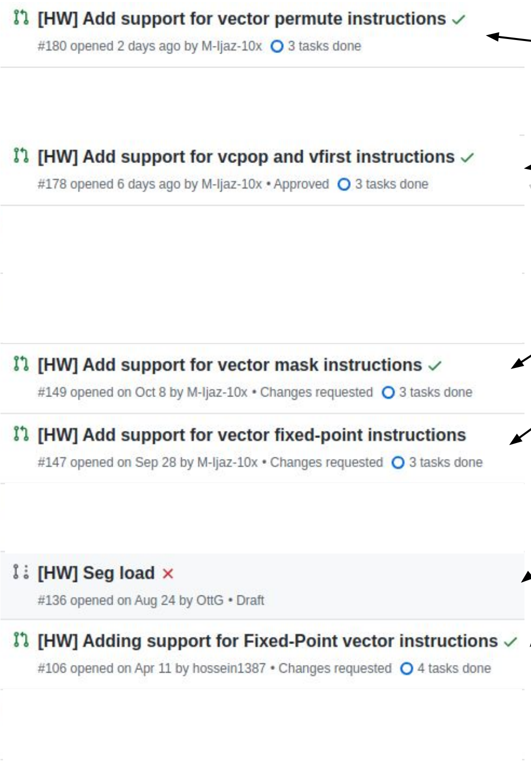
~3.6 FLOP/cycle

Repeat measurements after optimization

Power and efficiency measurements - Many Ara systems in an SoC wrapper

- Memory left outside
- Interconnect left outside

PRs - Compliance



[HW] Add support for vector permute instructions ✓
#180 opened 2 days ago by M-Ijaz-10x ○ 3 tasks done

[HW] Add support for vcpop and vfirst instructions ✓
#178 opened 6 days ago by M-Ijaz-10x • Approved ○ 3 tasks done

[HW] Add support for vector mask instructions ✓
#149 opened on Oct 8 by M-Ijaz-10x • Changes requested ○ 3 tasks done

[HW] Add support for vector fixed-point instructions
#147 opened on Sep 28 by M-Ijaz-10x • Changes requested ○ 3 tasks done

[HW] Seg load ✗
#136 opened on Aug 24 by OttG • Draft

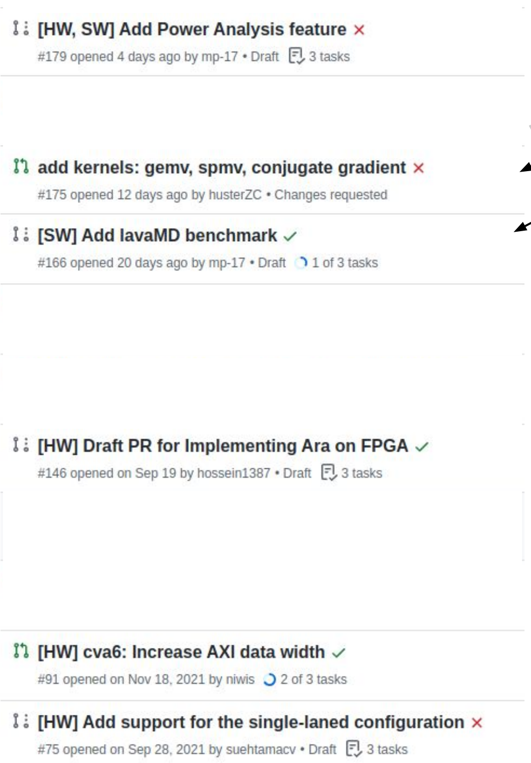
[HW] Adding support for Fixed-Point vector instructions ✓
#106 opened on Apr 11 by hossein1387 • Changes requested ○ 4 tasks done

➢	Vector complex shuffling	
✓	Mask-reg instructions	(1)
✓	Mask-reg instructions	(2)
✓	Fixed-Point support	(1)
✓	Fixed-Point support	(2)
✗	Segment memory ops	

✗ strncpy, strncmp

✓ AWB

PRs - Benchmarks



[HW, SW] Add Power Analysis feature ✗
#179 opened 4 days ago by mp-17 • Draft 3 tasks

add kernels: gemv, spmv, conjugate gradient ✗
#175 opened 12 days ago by husterZC • Changes requested

[SW] Add lavaMD benchmark ✓
#166 opened 20 days ago by mp-17 • Draft 1 of 3 tasks

[HW] Draft PR for Implementing Ara on FPGA ✓
#146 opened on Sep 19 by hossein1387 • Draft 3 tasks

[HW] cva6: Increase AXI data width ✓
#91 opened on Nov 18, 2021 by niwis 2 of 3 tasks

[HW] Add support for the single-laned configuration ✗
#75 opened on Sep 28, 2021 by suehtamacv • Draft 3 tasks

✓ gemv, spmv, conjugate gradient

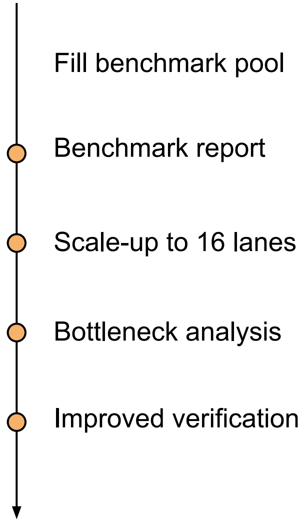
✓ lavaMD

✗ strncpy, strncmp

✓ AWB

Further

- **Software**
  - Stall analysis
  - WAW / WAR profiling
- **Hardware (RTL + Backend)**
  - Compliance + Verification
  - Scale up to 16 lanes



Fill benchmark pool

Benchmark report

Scale-up to 16 lanes

Bottleneck analysis

Improved verification