Update on Ara

08/03/2023

Matteo Perotti
Matheus Cavalcante

Professor Luca Benini
Integrated Systems Laboratory
ETH Zürich

Energy Efficiency

- **Scaling**
  - #Lanes
  - Data width
  - Data type

`fmatmul, 128x128x128`

Lanes	Efficiency (GOPS/W)
2	30.09
4	**34.33**
8	**32.47**

Energy Efficiency

- **Scaling**
  - #Lanes
  - Data width
  - Data type

`fmatmul, 128x128x128`

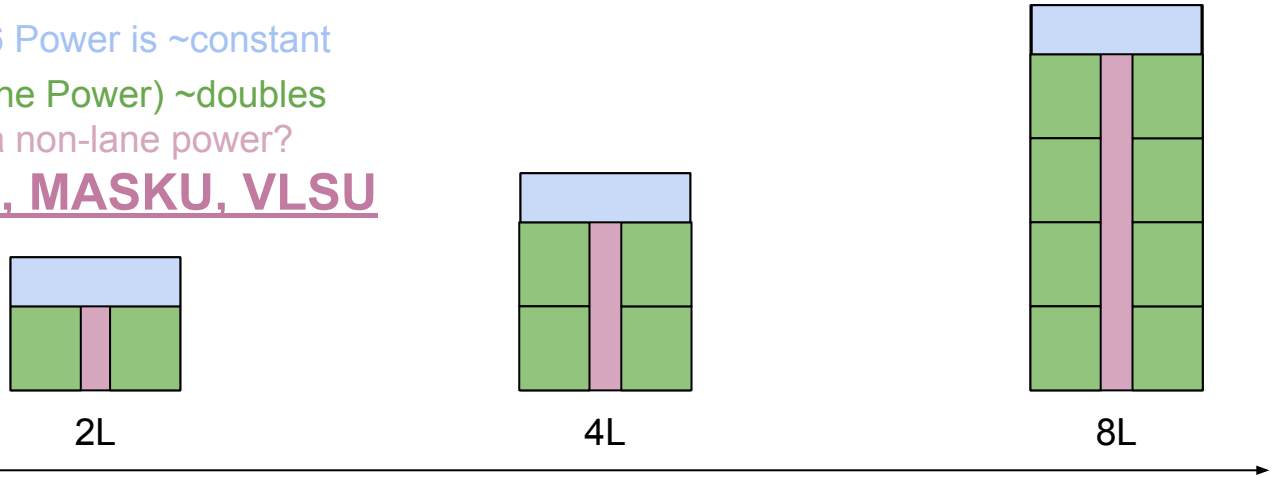
Lanes	Efficiency (GOPS/W)
2	30.09
4	**34.33**
8	**32.47**

CVA6 Power is ~constant

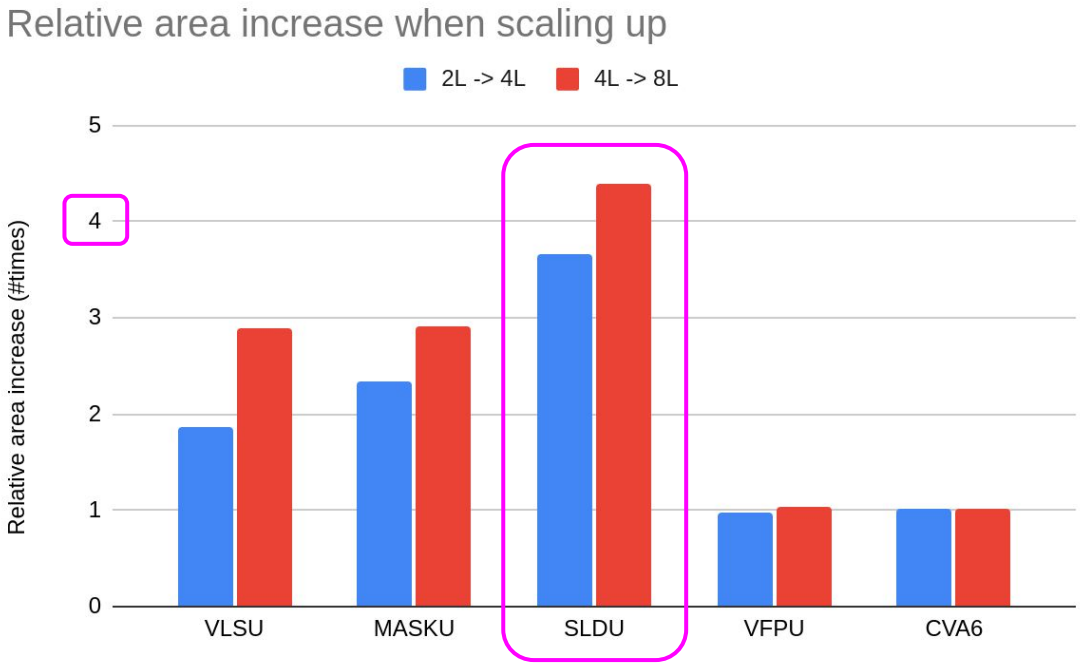
∑(Lane Power) ~doubles

Ara non-lane power?

SLDU, MASKU, VLSU

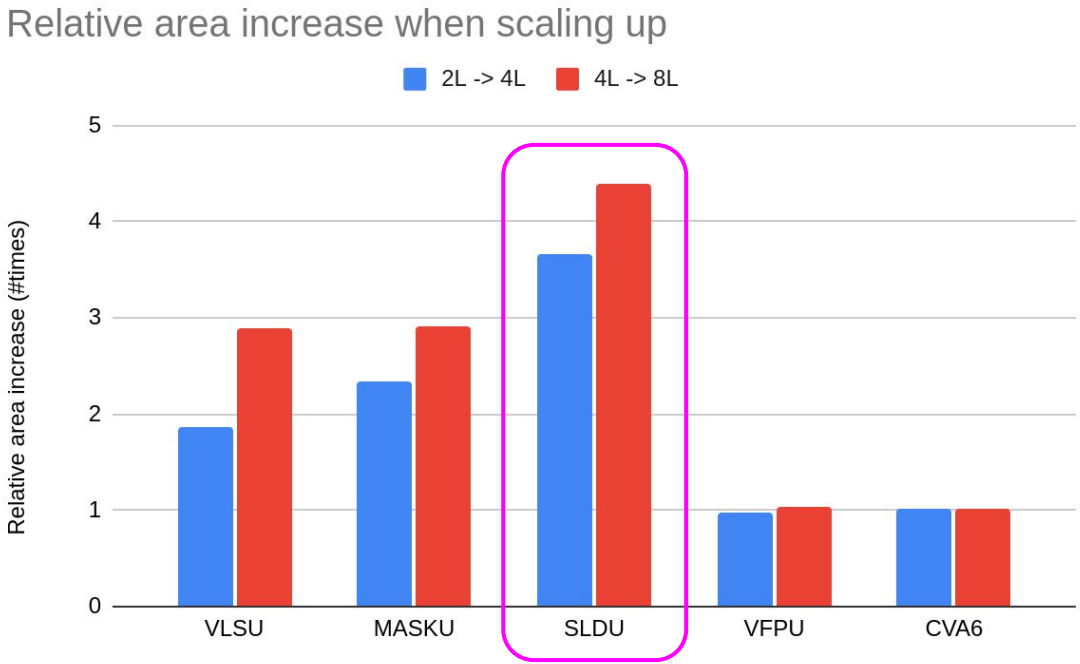


How do Ara's units scale in AREA?



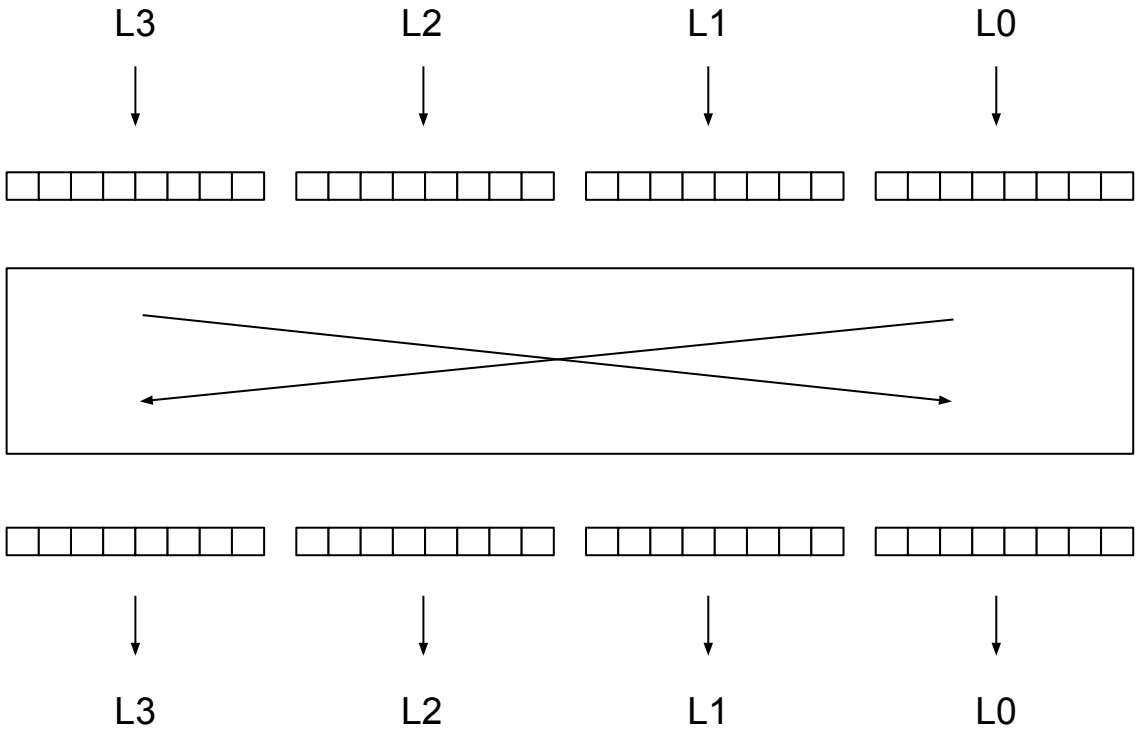
Largest growth $O(L^2)$

How do Ara's units scale in AREA?

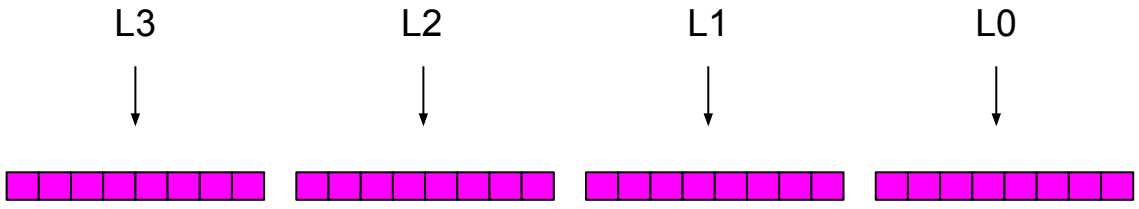


SLDU is also the largest unit
8L: SLDU area == 1.7x VLSU area

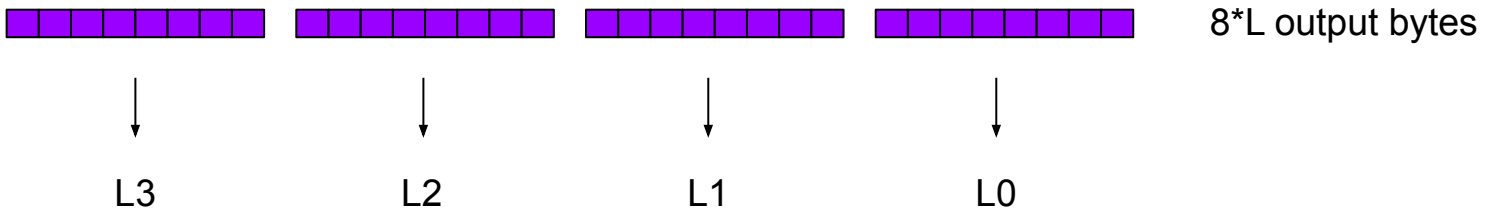
SLDU - Connections



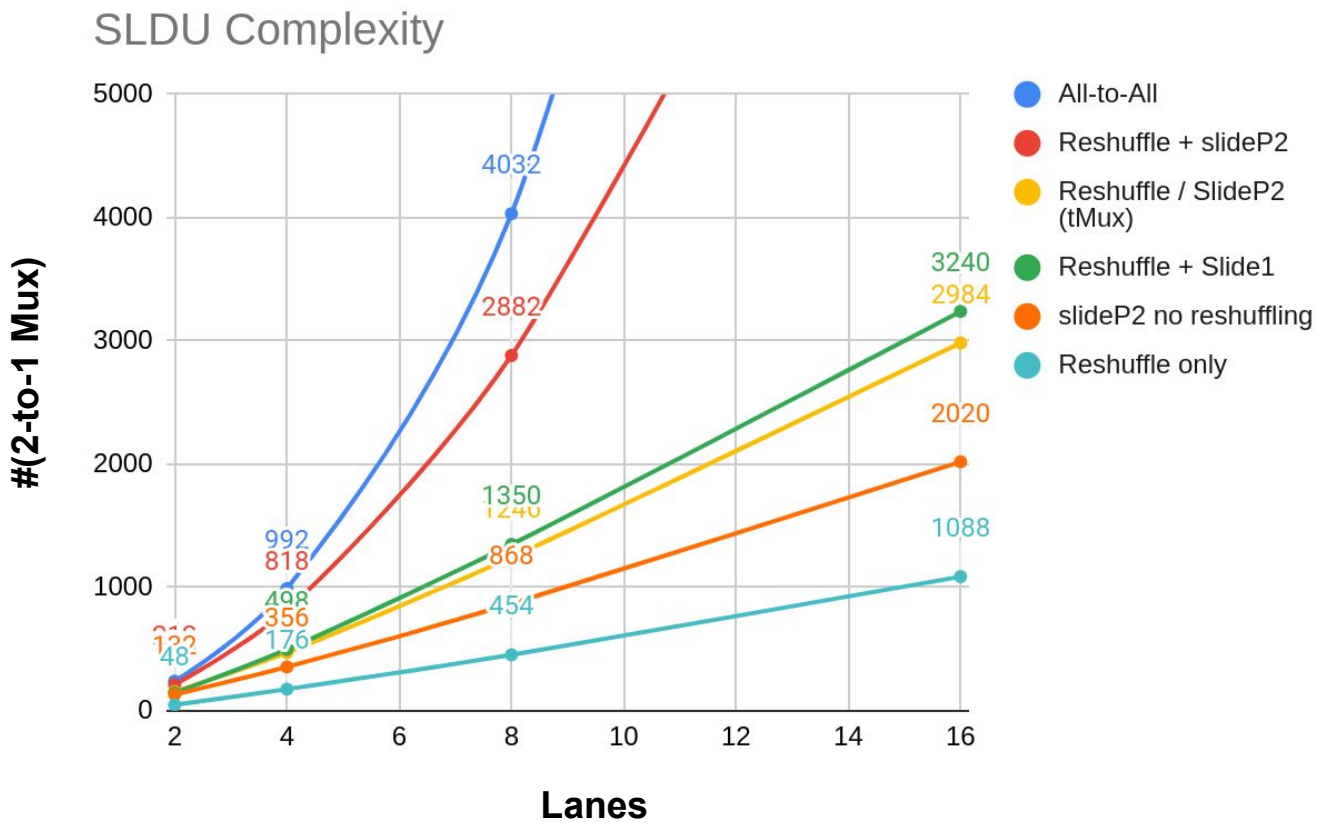
SLDU - Connections



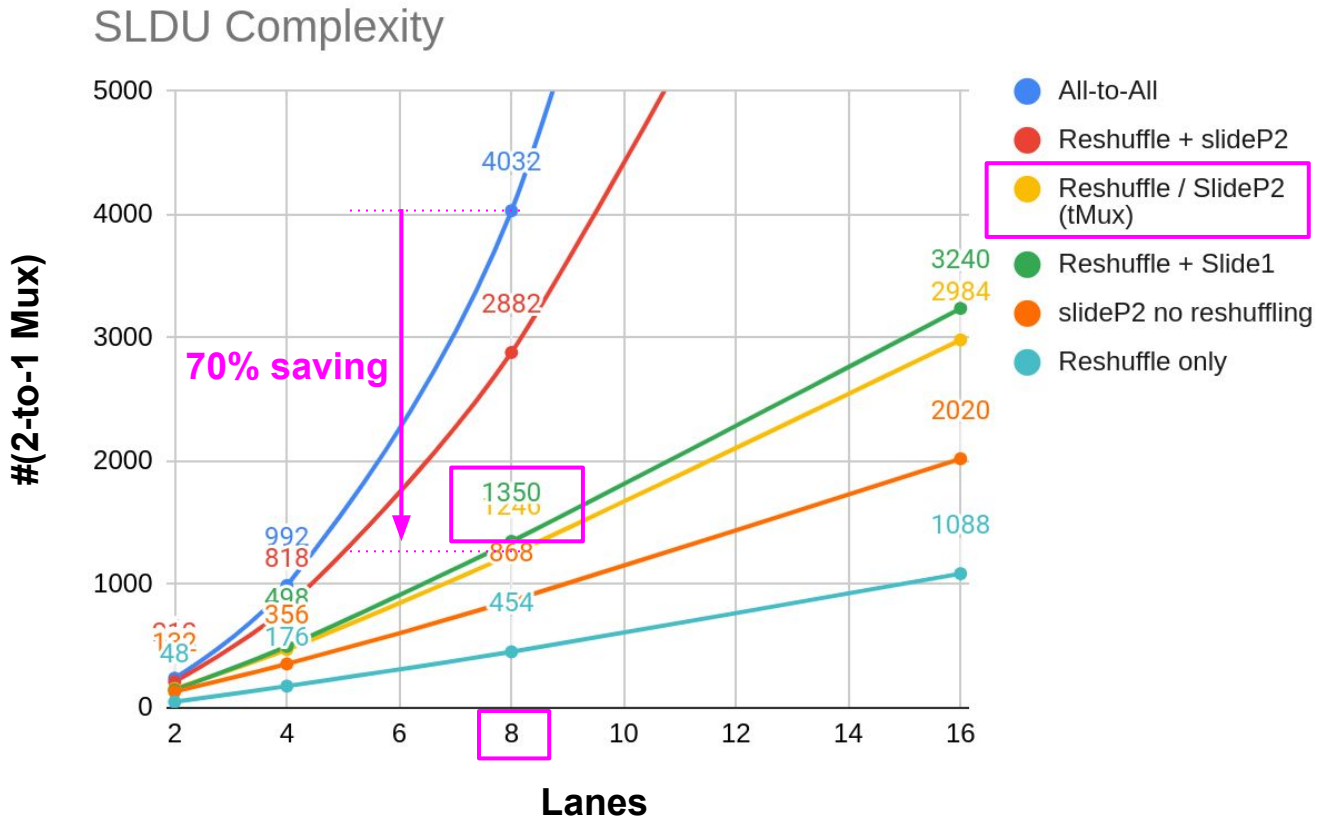
64*L² connections



SLDU - A New Hope



SLDU - A New Hope

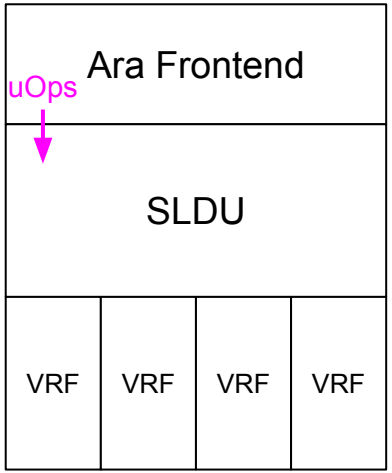


SLDU - A New Datapath

- Support slides by power-of-two strides only
- Either we slide or we re-encode
- Some difficulties to support undisturbed policy

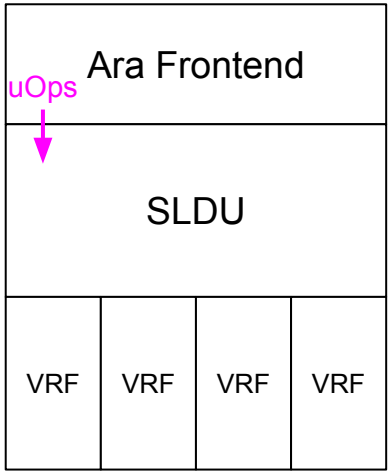
Non-power-of-two Strides?

- uOps injection from the frontend?
- Slide by 5?
  - ➢ Slide by 4 + Slide by 1

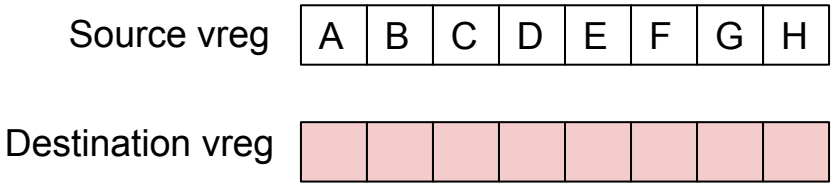


Non-power-of-two Strides?

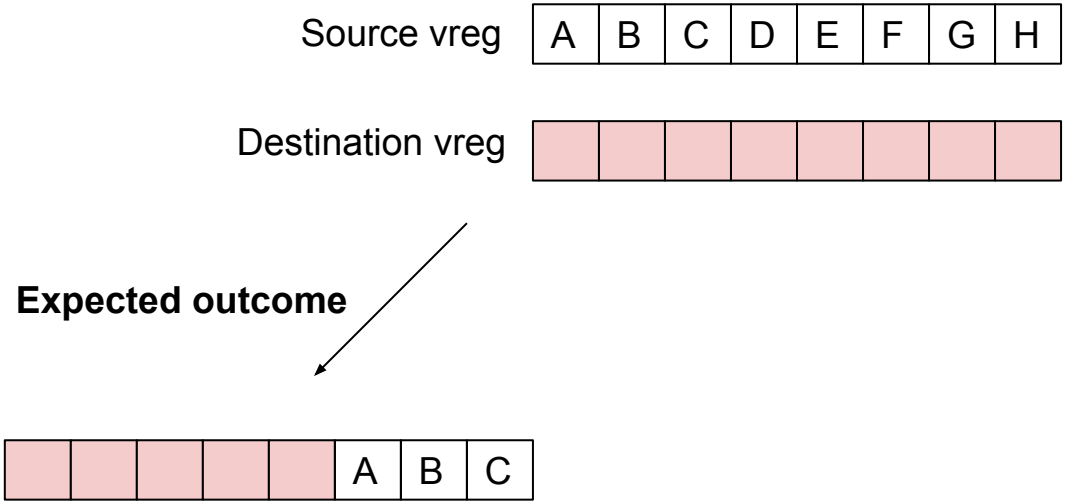
- uOps injection from the frontend?
- Slide by 5?
  - ➢ Slide by 4 + Slide by 1
- After the Slide by 4, store the intermediate result in the VRF
- Undisturbed policy? Masked elements?



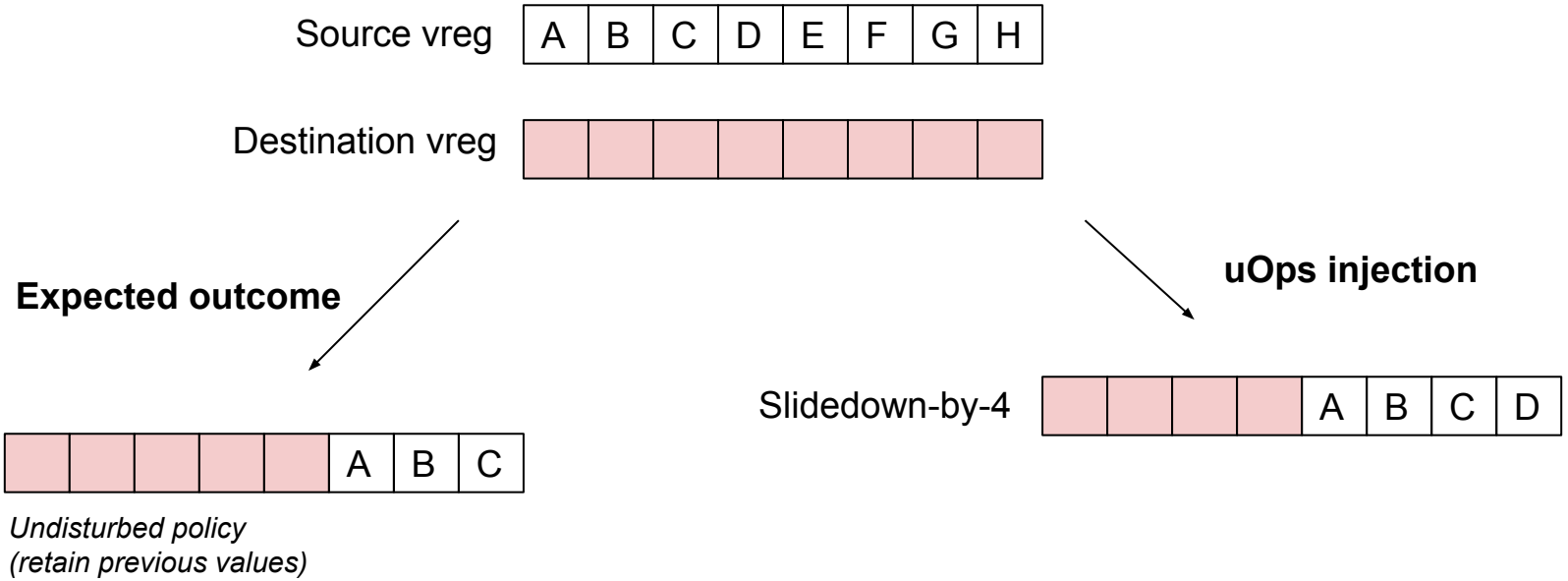
Slidedown-by-5



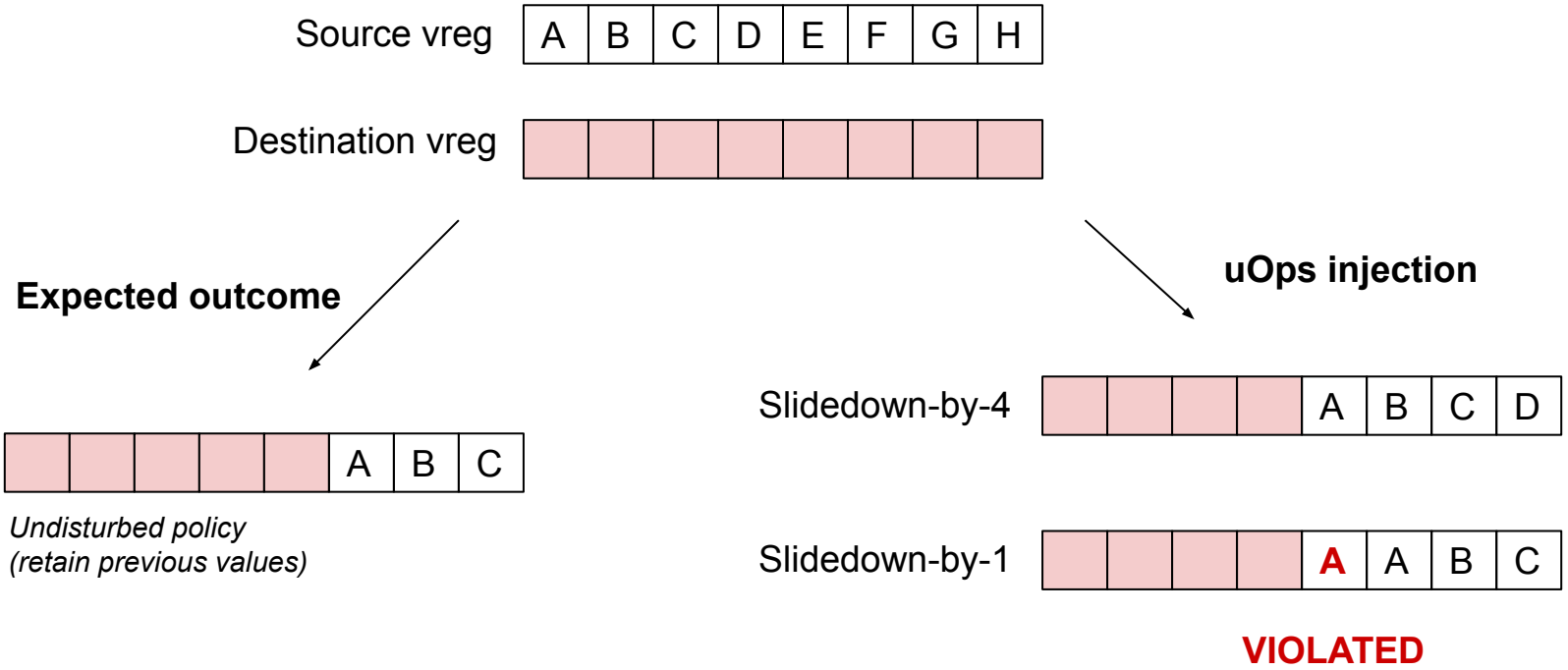
Slidedown-by-5



Slidedown-by-5

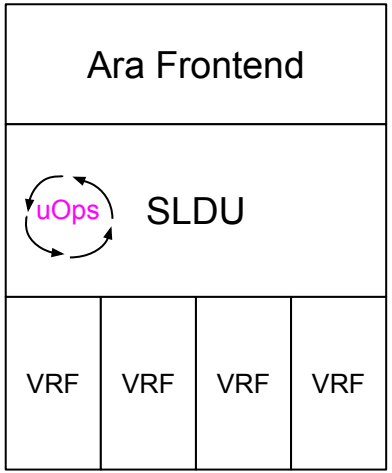


Slidedown-by-5

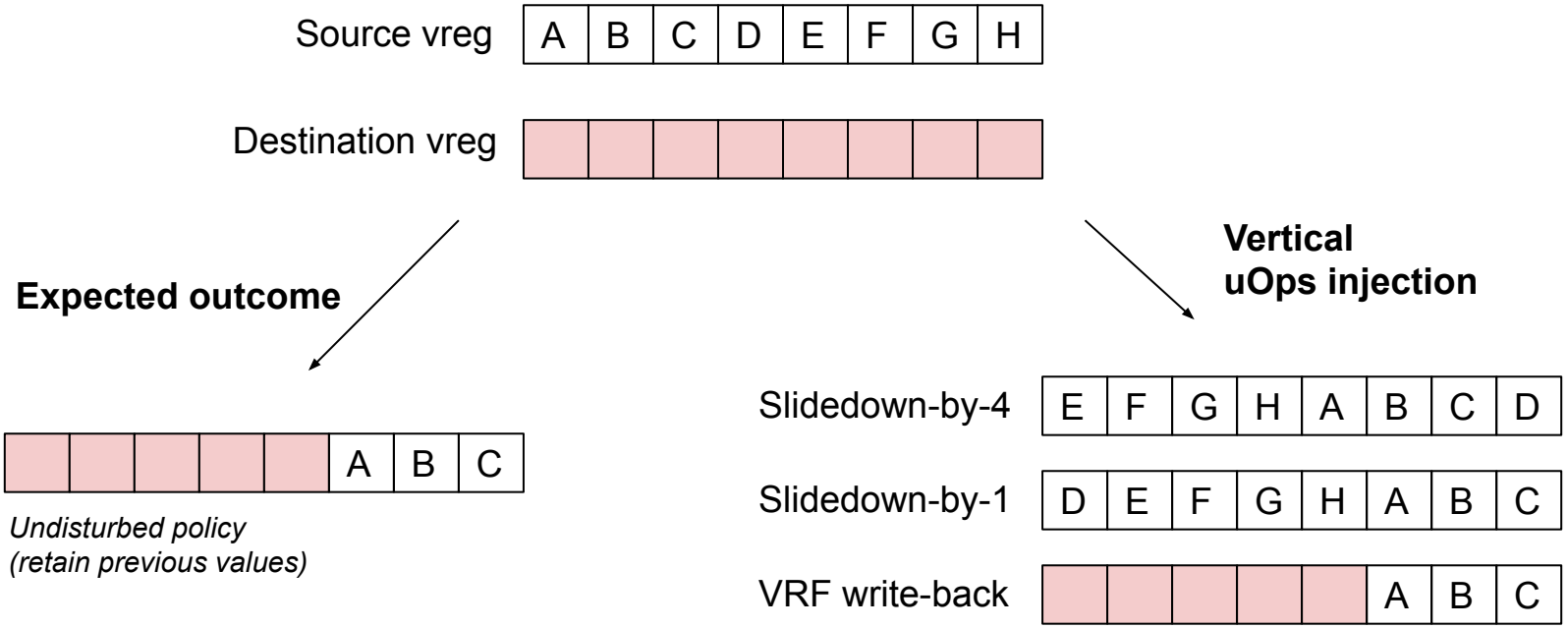


Solution - Vertical uOps

- uOps in the Slide Unit
- Slide by 5?
  - ➢ Slide by 4 + Slide by 1
- Buffer each chunk of vector in the SLDU
- Slide by 4 + Slide by 1 each chunk



Slidedown-by-5



Results

- 4-lane design
  - 26’000 µm² → 9’800 µm² (**-63%**)
- 8-lane design
  - 122’000 µm² → 17’000 µm² (**-86%**)

SLDU is the smallest all-to-all unit now!

Efficiency

- New **results** soon:
  - **Optimized SLDU**
  - **Clock-gate** the **macros** (D$, I$, VRF banks)
    - -8% power!
  - Different **data types**