Update on Ara

22/02/2023

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	?

Energy Efficiency

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

`fmatmul, 128x128x128`

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

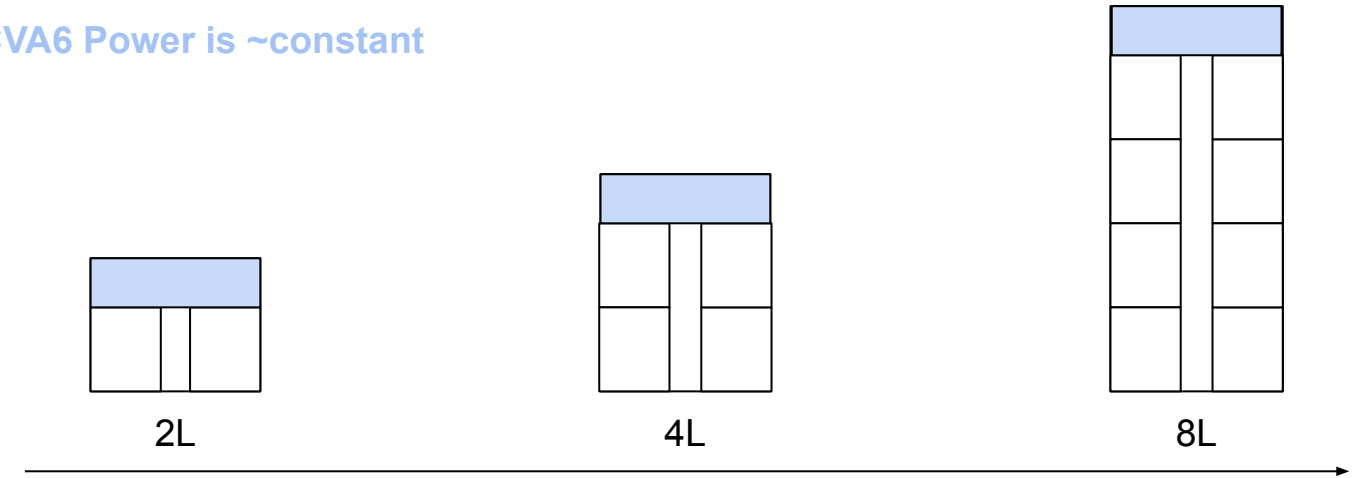
Energy Efficiency

- **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



Energy Efficiency

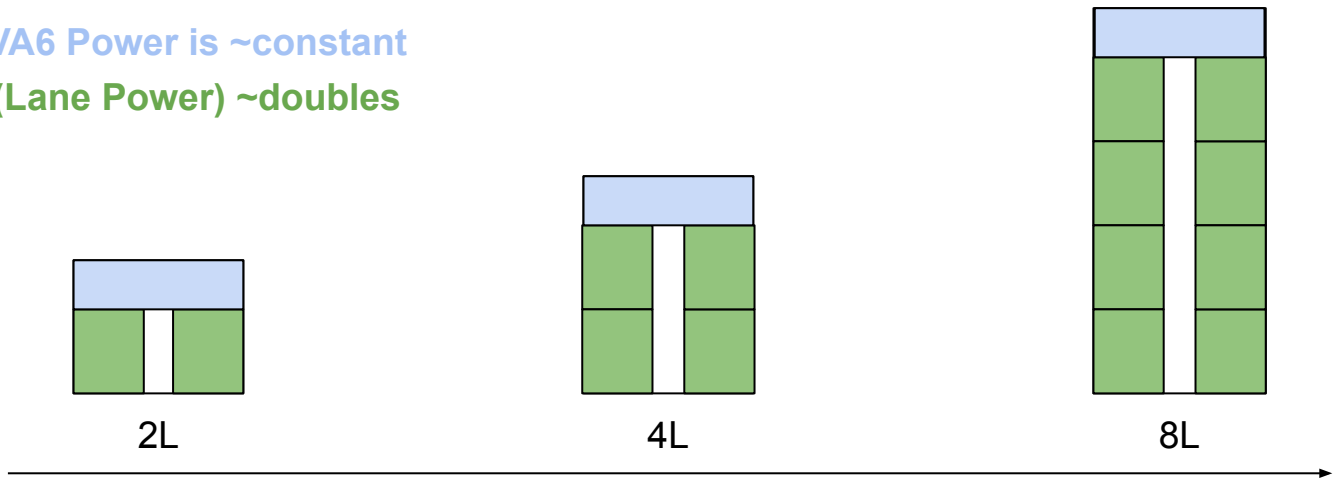
- **Scaling**
  - #Lanes
  - Data width
  - Data type

`fmatmul, 128x128x128`

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

CVA6 Power is ~constant

∑(Lane Power) ~doubles



Energy Efficiency

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

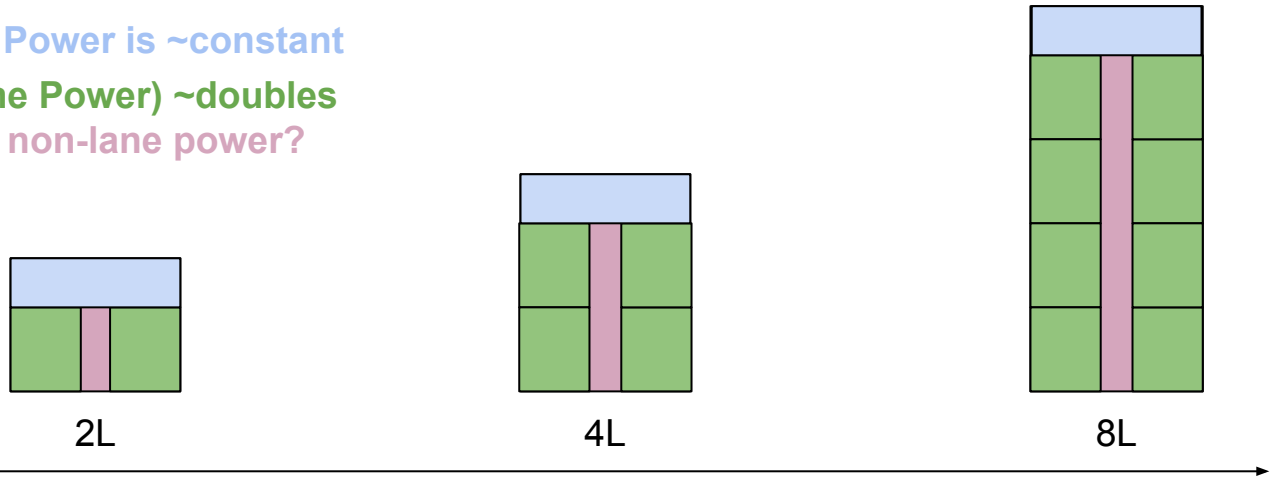
`fmatmul, 128x128x128`

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

CVA6 Power is ~constant

∑(Lane Power) ~doubles

Ara non-lane power?



Energy Efficiency

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

`fmatmul, 128x128x128`

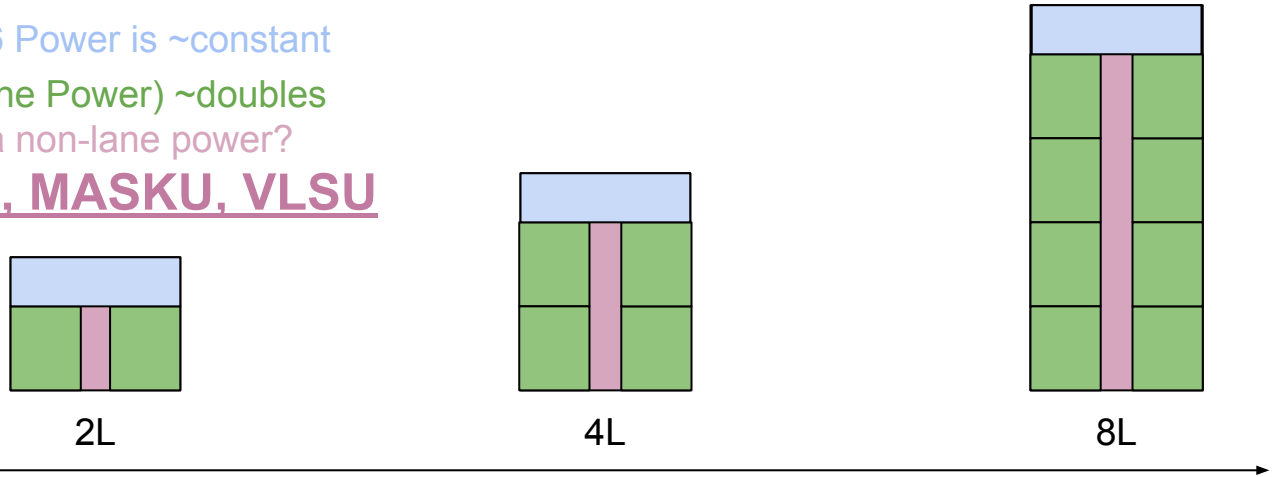
Lanes	Efficiency (GOPS/W)
2	30.09
4	**34.33**
8	**32.47**

CVA6 Power is ~constant

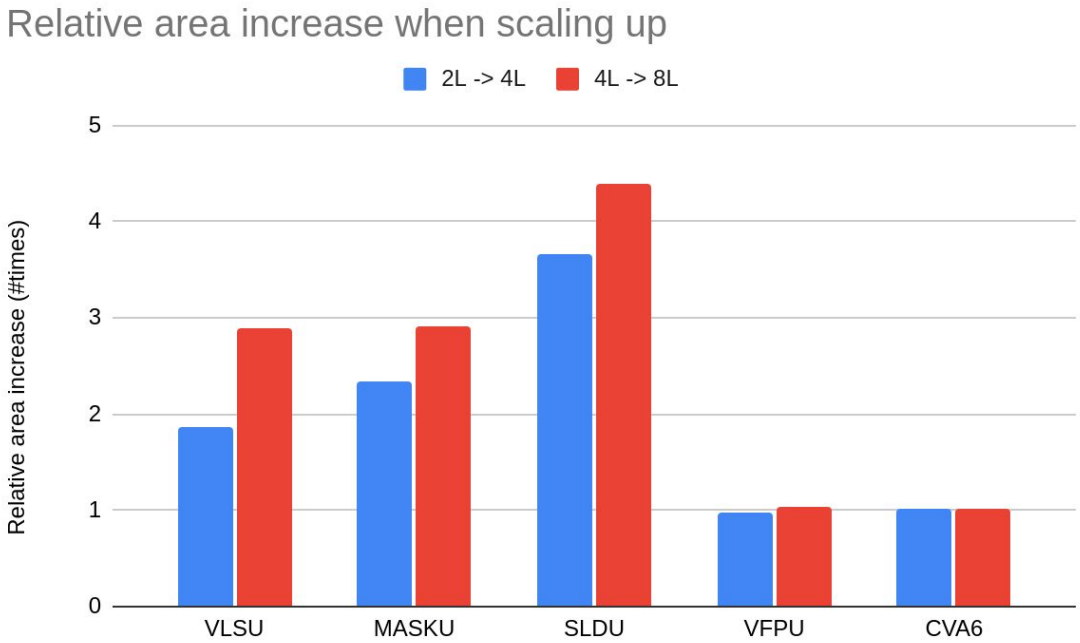
∑(Lane Power) ~doubles

Ara non-lane power?

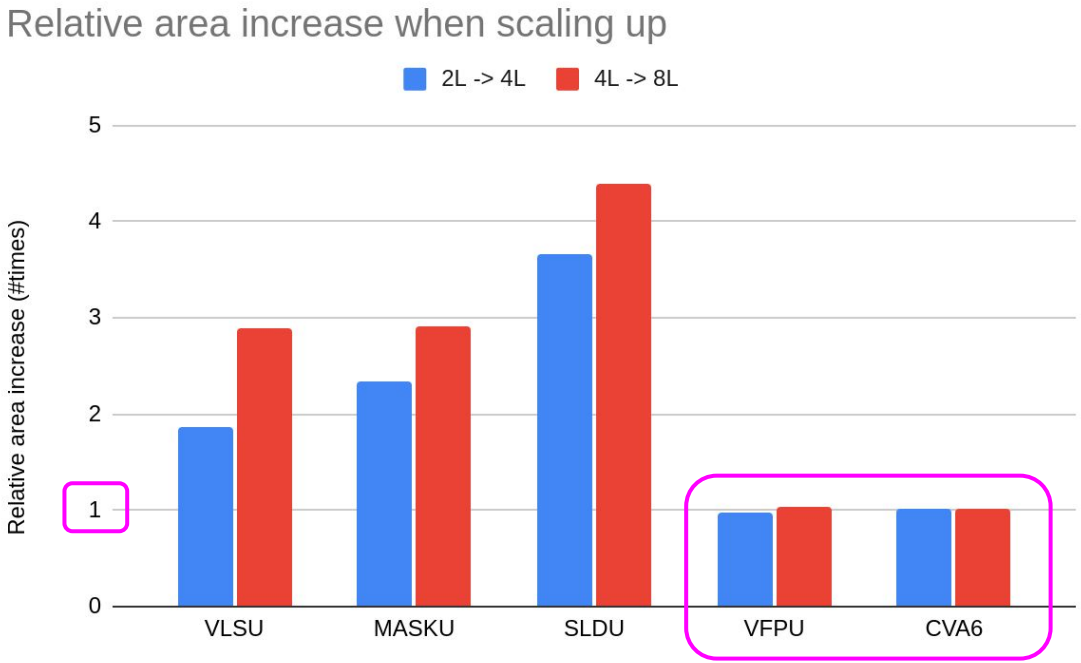
SLDU, MASKU, VLSU



How do Ara's units scale in AREA?

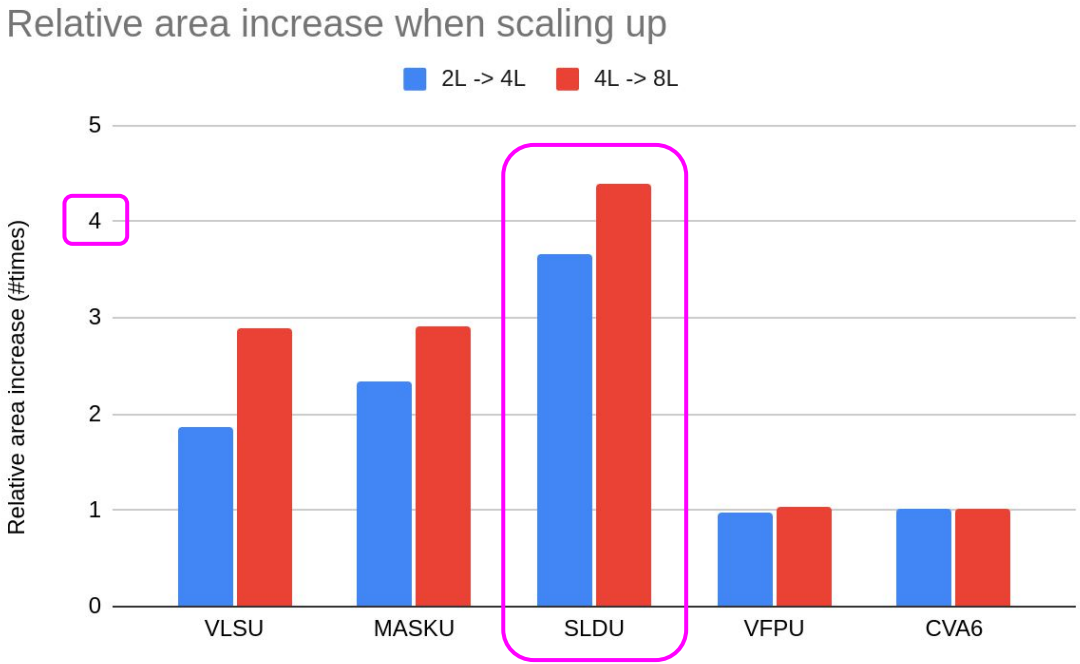


How do Ara's units scale in AREA?



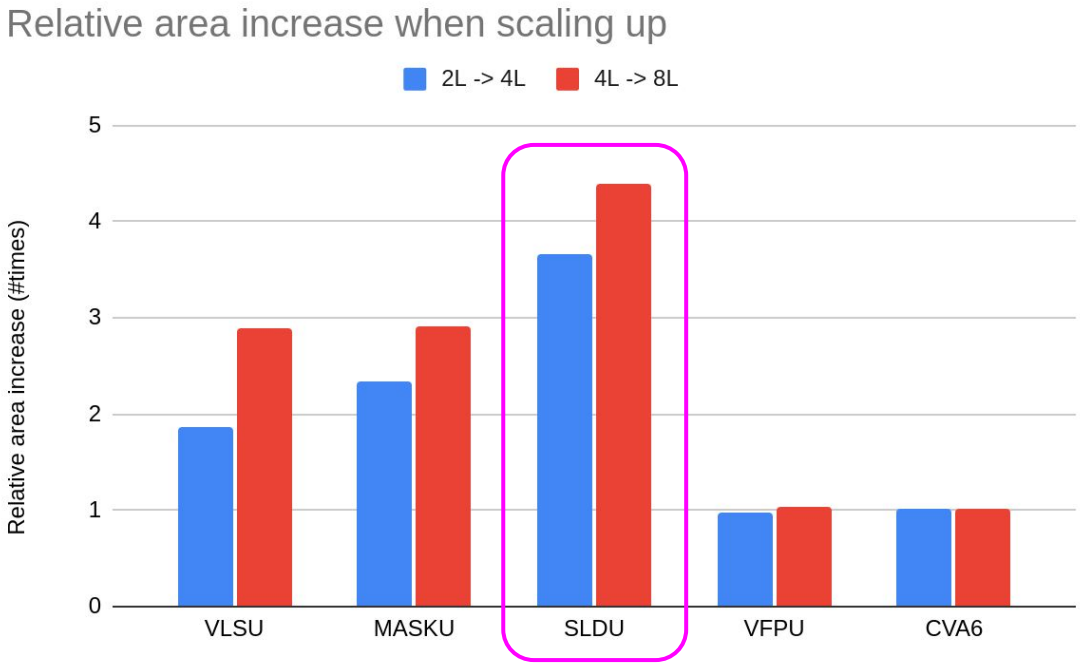
Same Area, as expected

How do Ara's units scale in AREA?



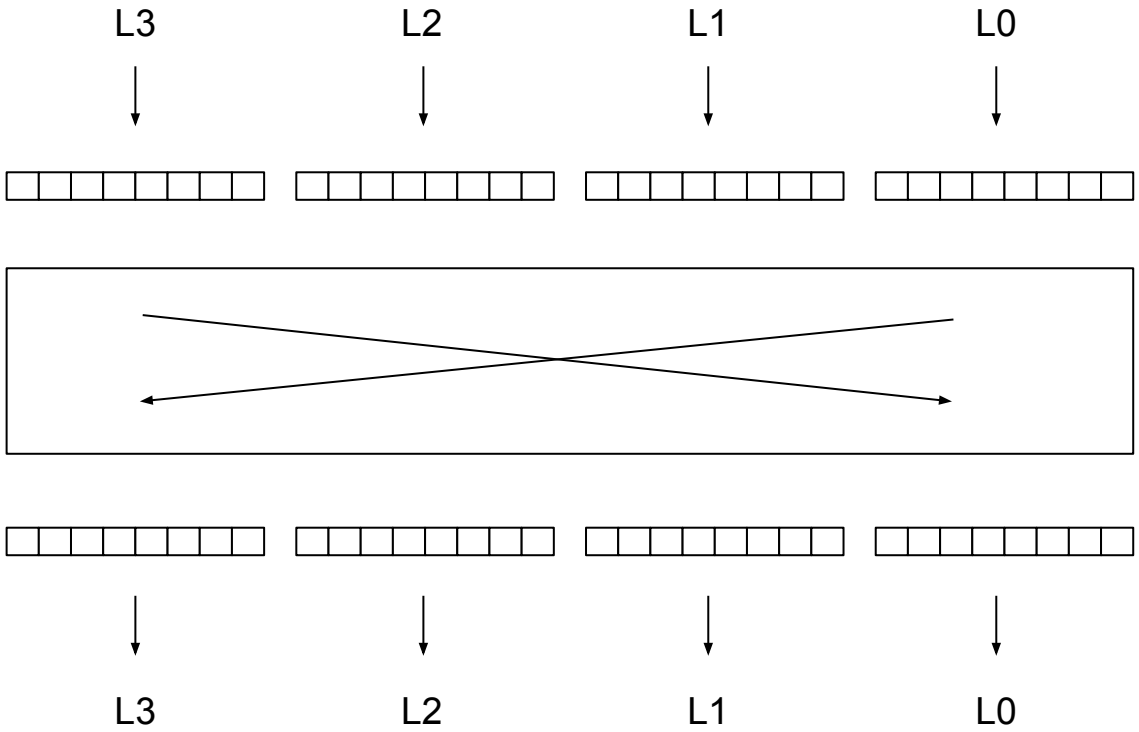
Largest growth $O(L^2)$

How do Ara's units scale in AREA?

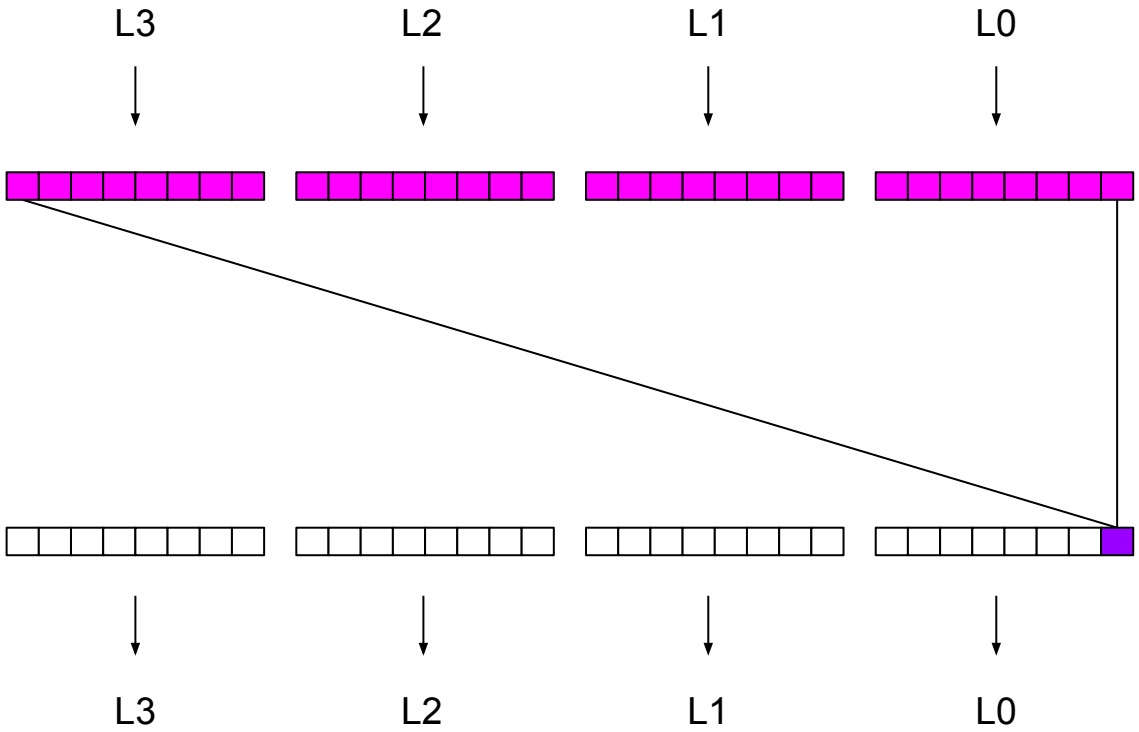


SLDU is also the largest unit
8L: SLDU area == 1.7x VLSU area

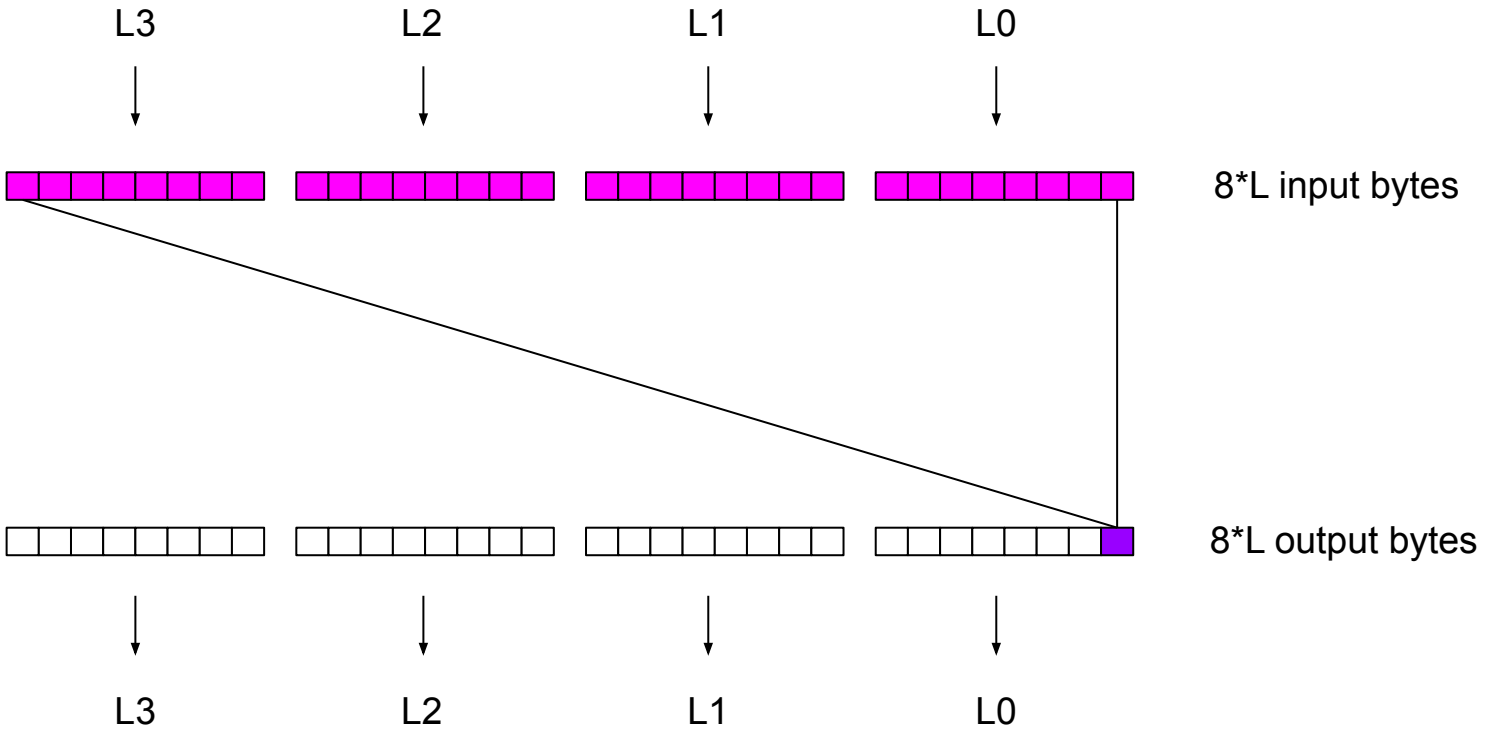
SLDU - Connections



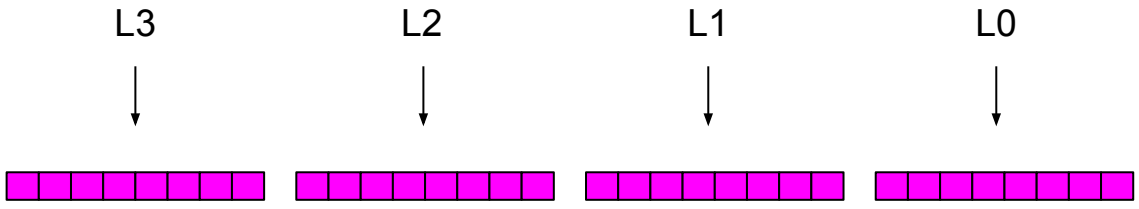
SLDU - Connections



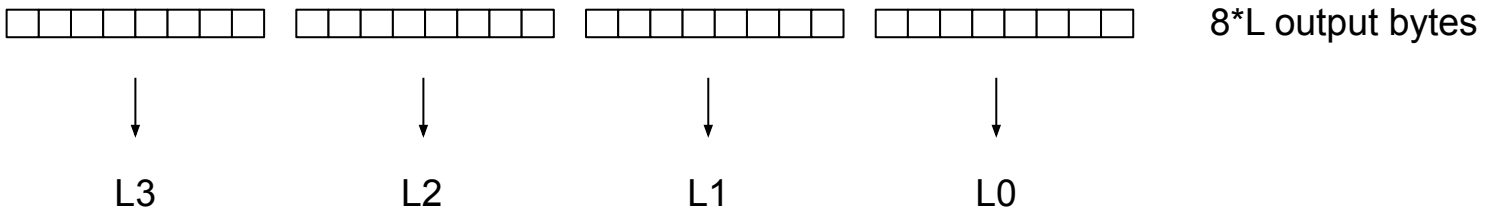
SLDU - Connections



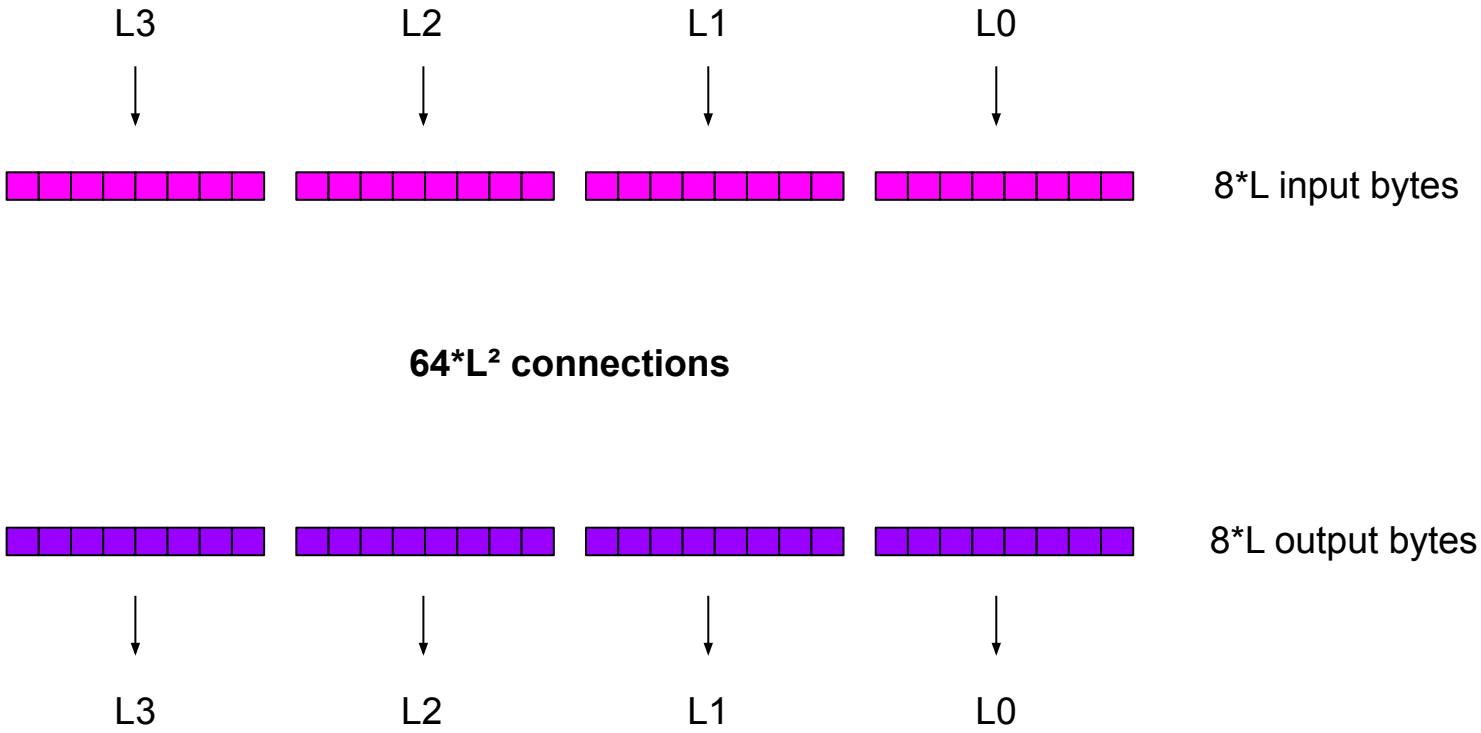
SLDU - Connections



All-to-All to support **arbitrary slides** with **8-bit elements** (and **re-encoding**)



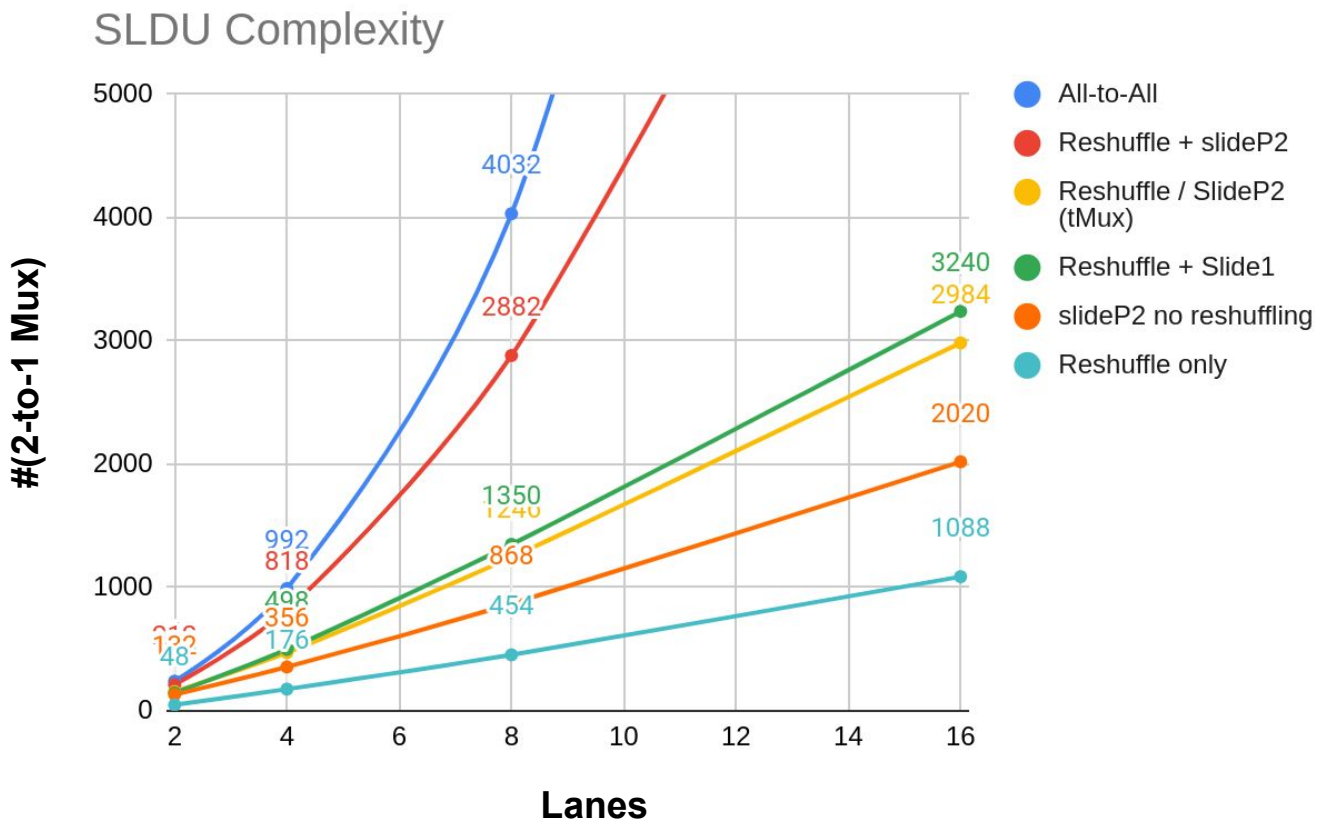
SLDU - Connections



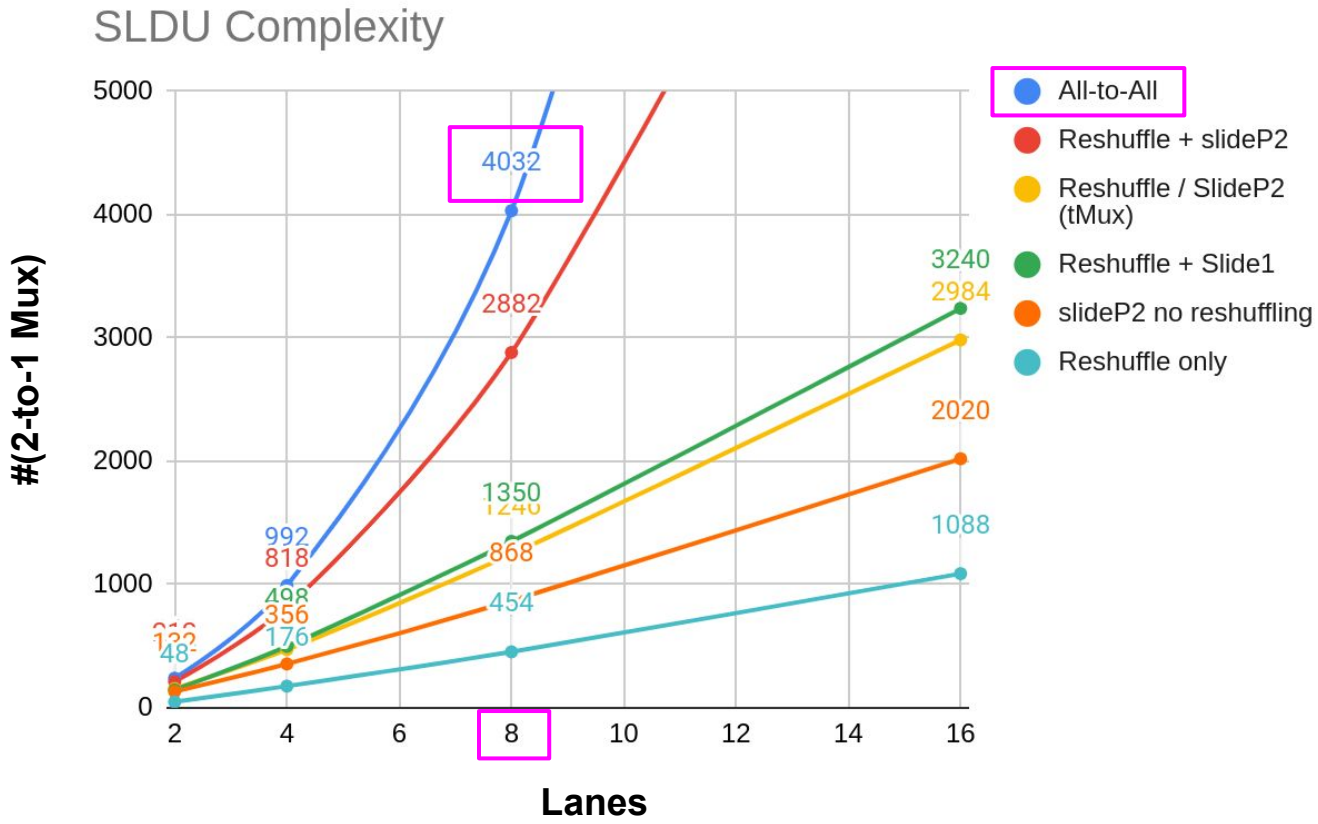
SLDU - A New Hope

- How to **simplify** the **SLDU?**
  - Only **slides** by **1?**
  - Only **slides** by **powers of 2?**
  - **Time Multiplex slides** and **encoding?**

SLDU - A New Hope

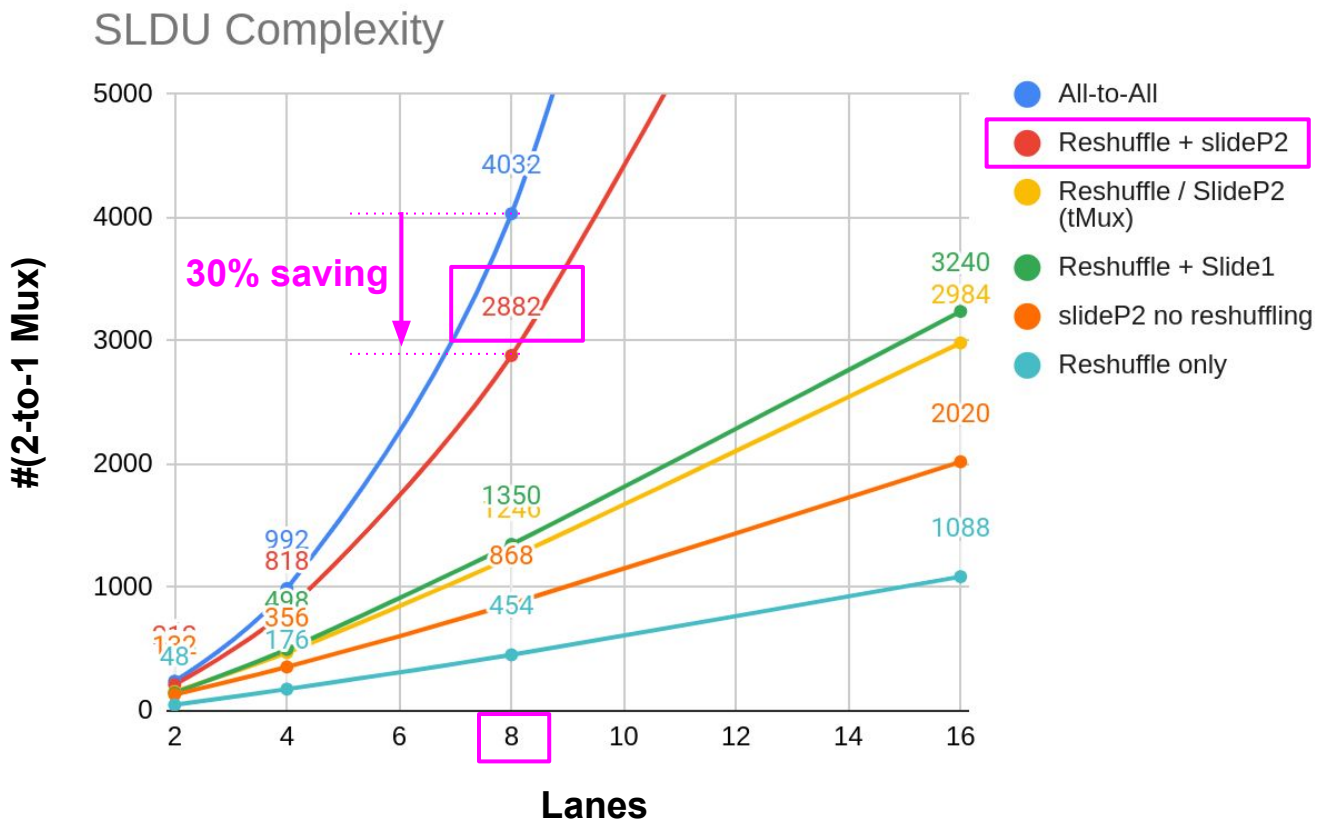


SLDU - A New Hope

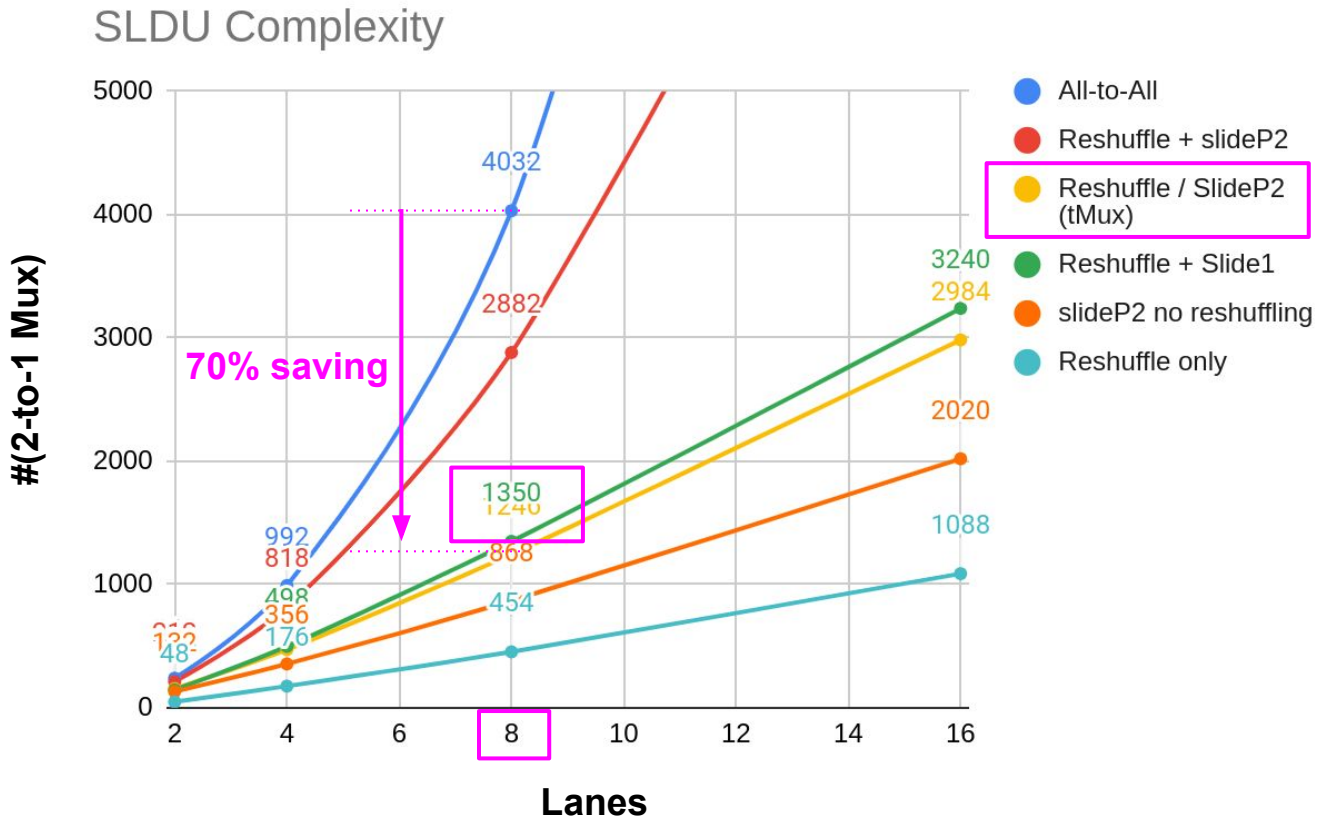


Original SLDU

SLDU - A New Hope



SLDU - A New Hope



SLDU - A New Datapath

- Time-multiplex p2 slides and re-encoding
- Some difficulties to support undisturbed policy
- In parallel:
  - Energy efficiency of 4L and 8L without SLDU