ICCD 2026
The 44th IEEE International Conference on Computer Design
Nov. 16-18, 2026 Hong Kong SAR, China

HCI: A Flexible and Scalable Interconnect for Tightly-Coupled Heterogeneous Accelerator Clusters

Sergio Mazzola (smazzola@iis.ee.ethz.ch),
Arpan Suravi Prasad, Davide Rossi, Luca Benini, Francesco Conti

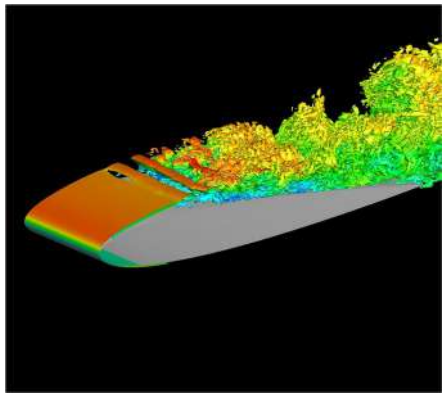


PULP Platform
Open-Source Hardware, the way it should be!

ETH zürich
ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

The workforce of next-generation computing

- Shared-memory, tightly-coupled **heterogeneous cluster**
  - End-to-end machine learning
  - From edge to high-performance



Suzuki Laboratory, University of Tokyo



Project Aria [Meta]
https://www.projectaria.com/glasses/

The workforce of next-generation computing

- Shared-memory, tightly-coupled **heterogeneous cl**
  - End-to-end machine learnin
  - From edge to high-perform

Tens to hundreds of general-purpose **control cores**

Word-granular, irregular access to TCDM **32/64-bit interfaces**

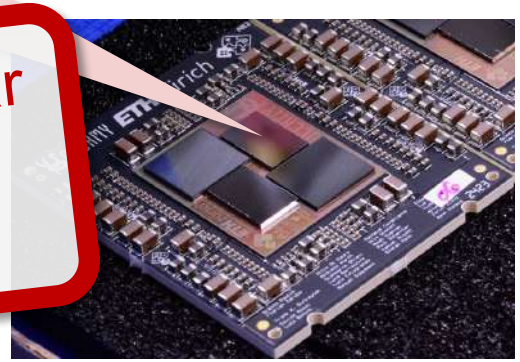
Large local TCDM, up to **few MiB**

Neural Processing Units

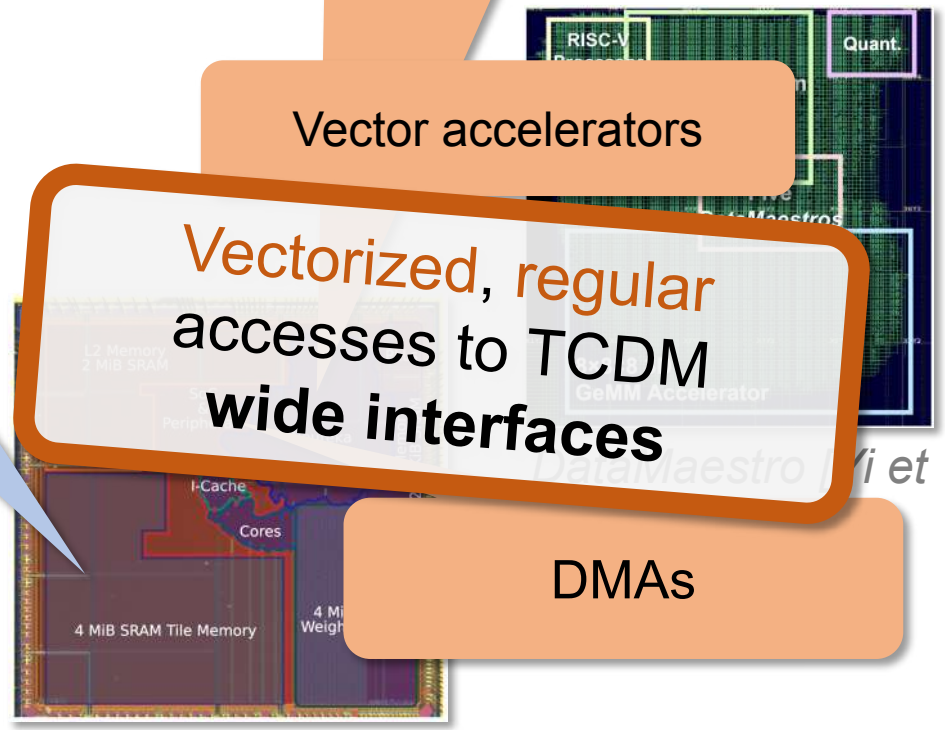
Vector accelerators

Vectorized, regular accesses to TCDM **wide interfaces**

DMAs



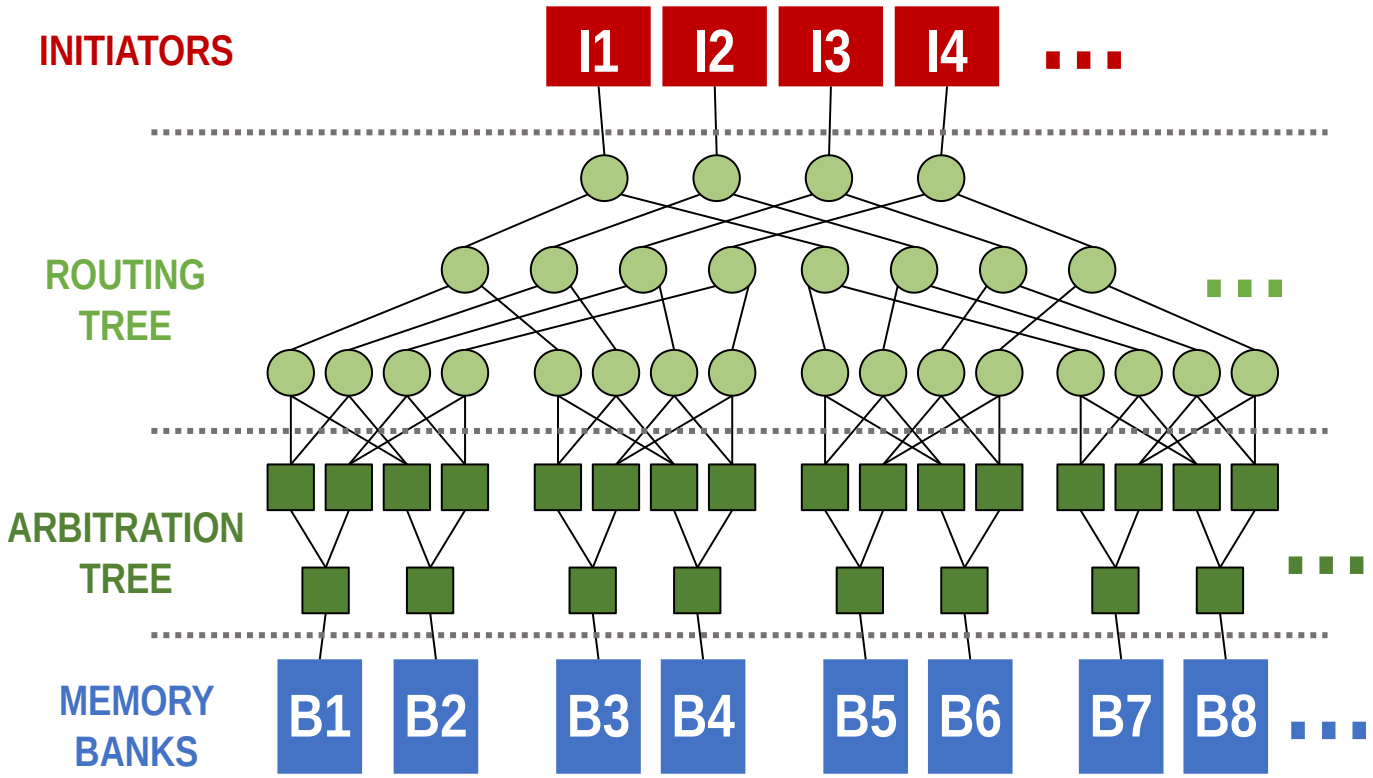
Occamy [Paulin et al., VLSI Symp 2024]



Siracusa 16nm [Prasad et al, JSSC 2024]

The hidden cost of sharing memory: **interconnect**

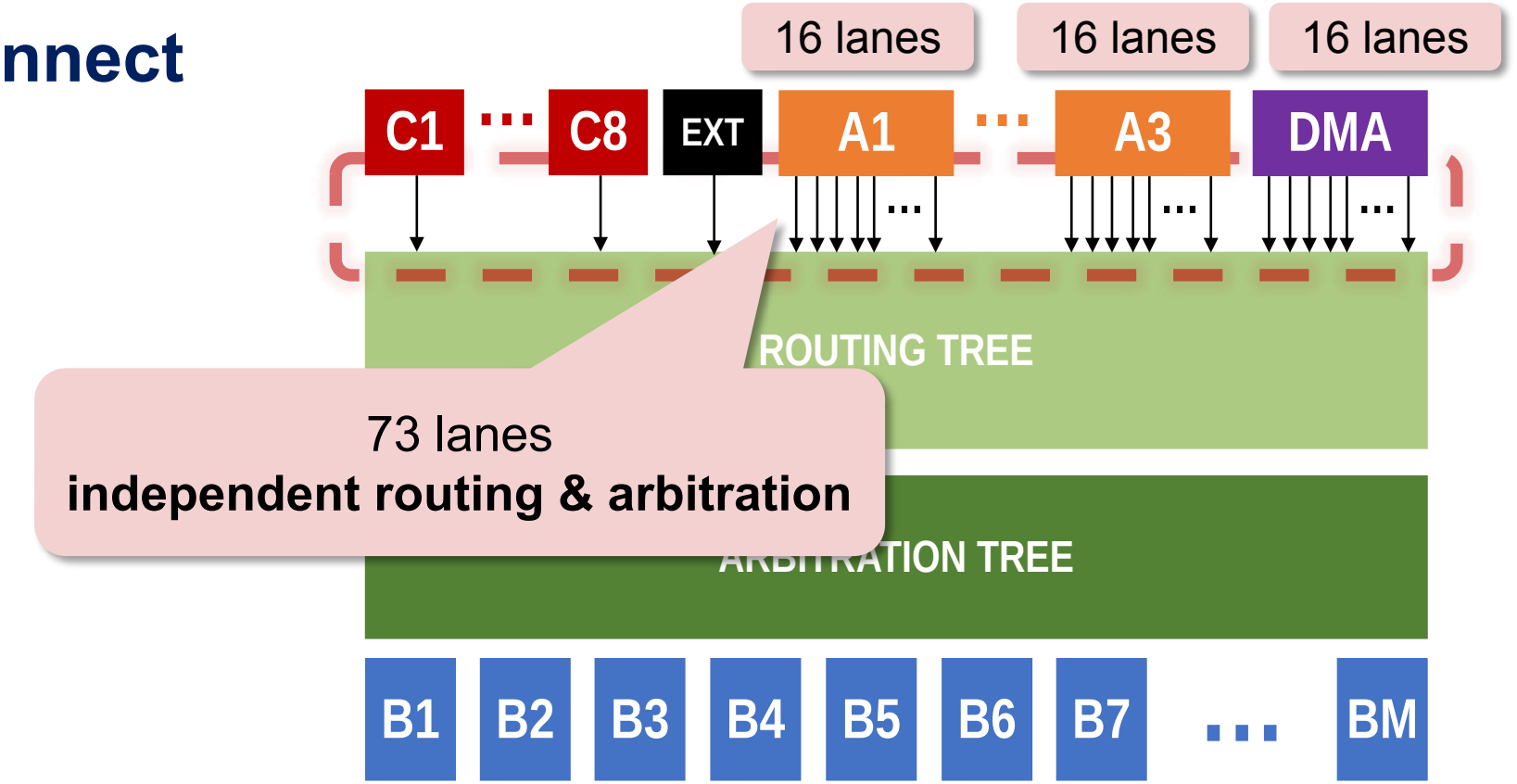
- **Logarithmic interconnect**
  - Full flexibility
  - Few-cycle latency



The hidden cost of sharing memory: **interconnect**

- **Logarithmic interconnect**
  - Full flexibility
  - Few-cycle latency

8x 32-bit CORES
3x 512-bit ACCELERATORS
1x 512-bit DMA
1x 32-bit EXTERNAL PORT

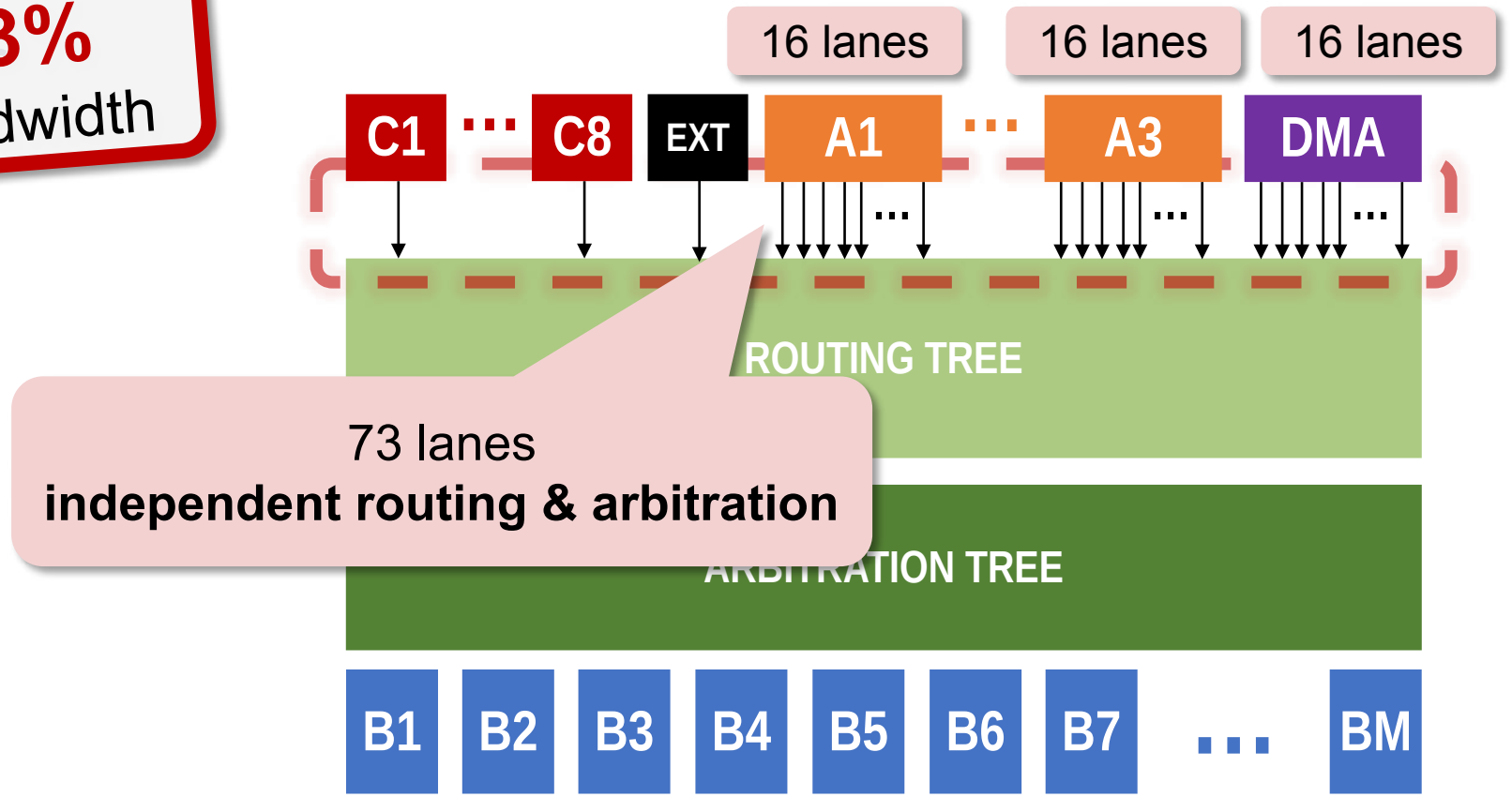


The hidden cost of sharing memory: **interconnect**

1. **Performance** under contention
2. Physical **scalability**

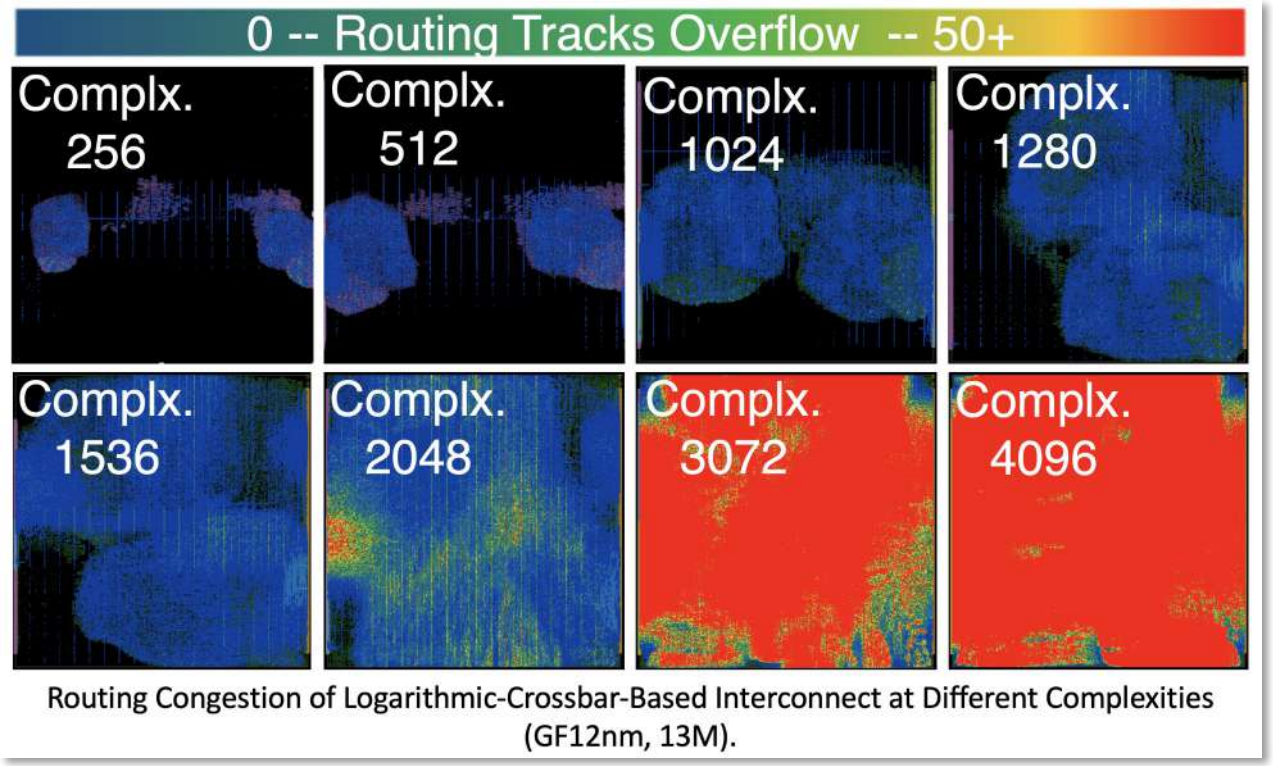
23% bandwidth

8x 32-bit CORES
3x 512-bit ACCELERATORS
1x 512-bit DMA
1x 32-bit EXTERNAL PORT



The hidden cost of sharing memory: **interconnect**

1. **Performance** under contention
   - **23%** bandwidth
2. Physical **scalability**
   - **13%** f_{MAX}
   - **32%** area



Routing Congestion of Logarithmic-Crossbar-Based Interconnect at Different Complexities (GF12nm, 13M).

Physical Design Is Not Push Button
[Yichao Zhang, EFCL Winter School 2026]

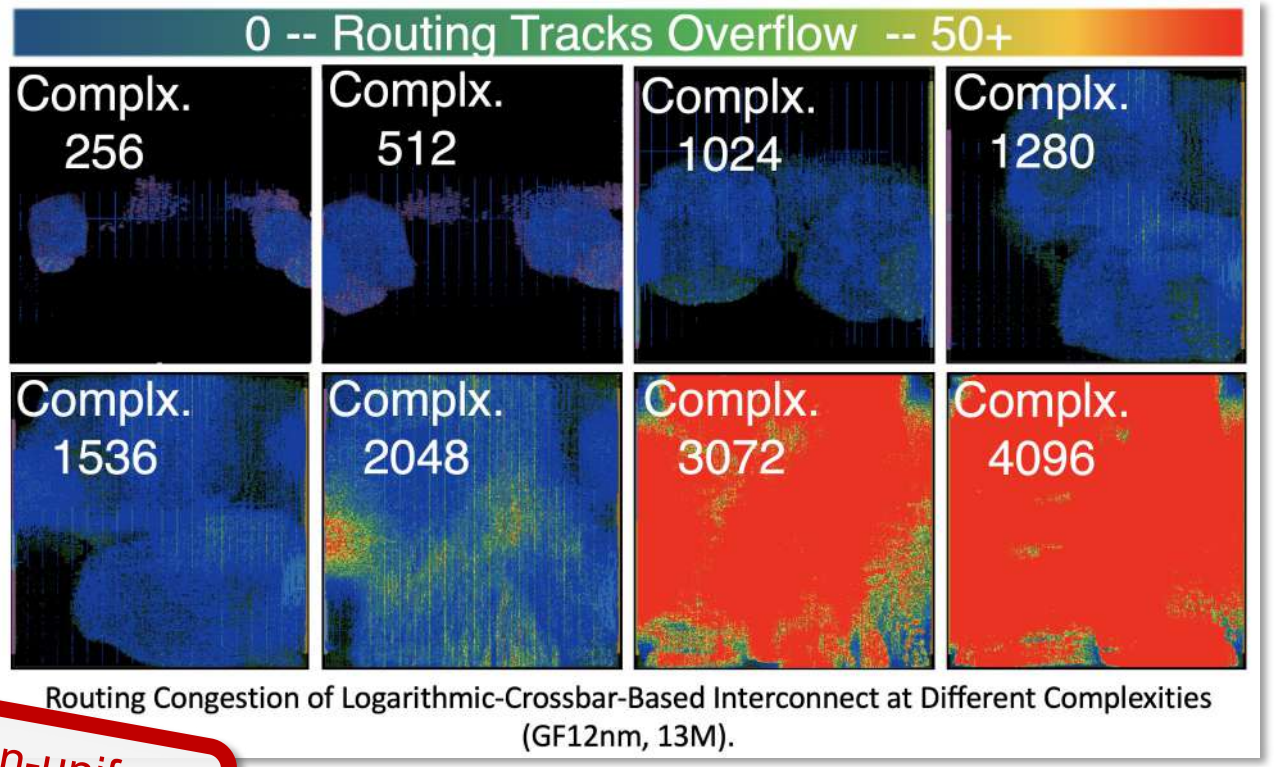
The hidden cost of sharing memory: **interconnect**

1. **Performance** under contention
2. Physical **scalability**

23% bandwidth

13% f_{MAX}

32% area



Routing Congestion of Logarithmic-Crossbar-Based Interconnect at Different Complexities (GF12nm, 13M).

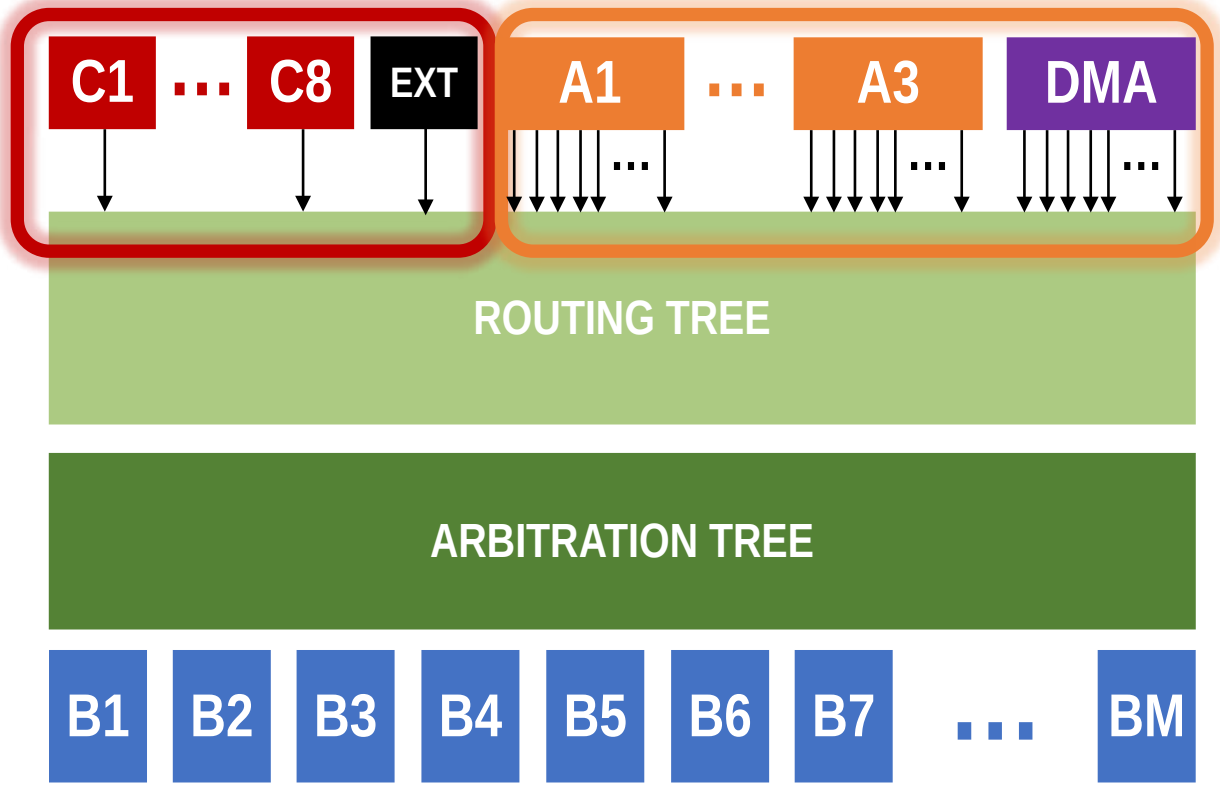
- More scalable solutions
  - NoC
  - Hierarchical crossbars
  - come with their own cons…

Longer, non-uniform latency

Physical Design Is Not Push Button
[Yichao Zhang, EFCL Winter School 2026]

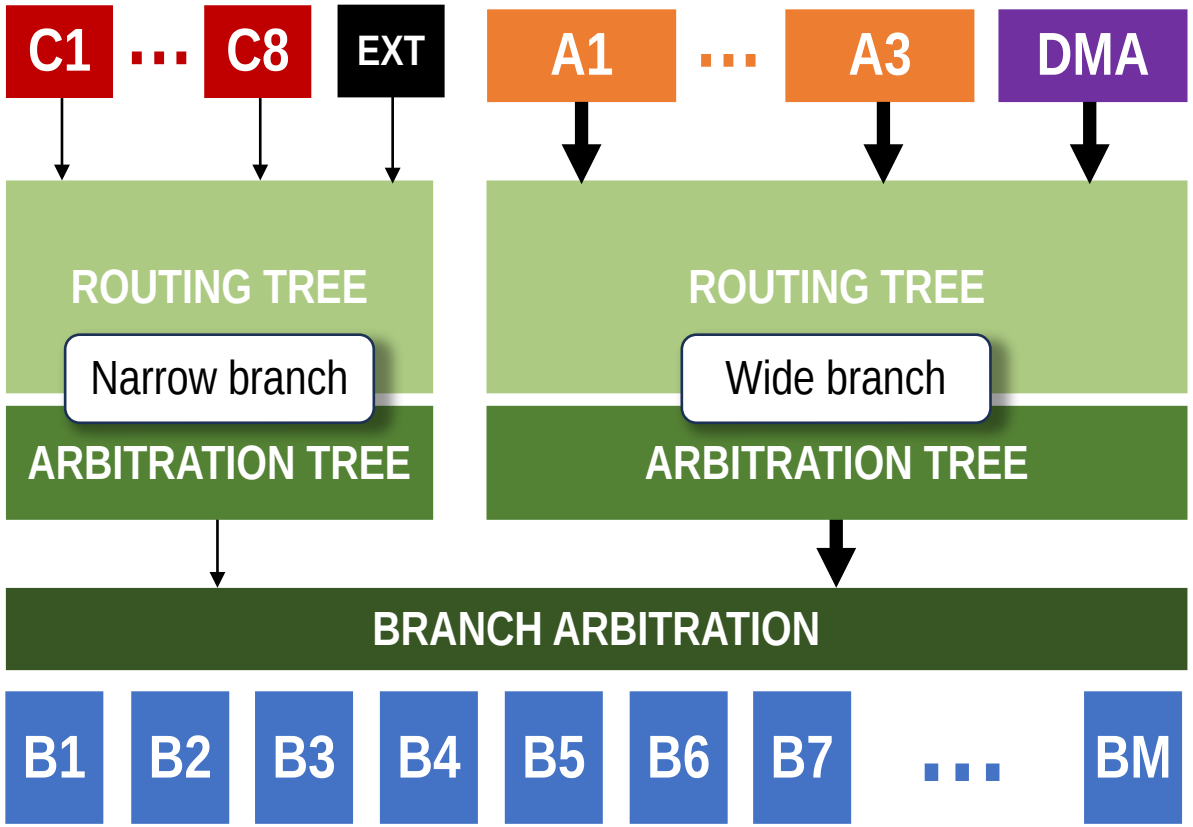
HCI: Heterogeneous Cluster Interconnect

- Word-granular access only where needed



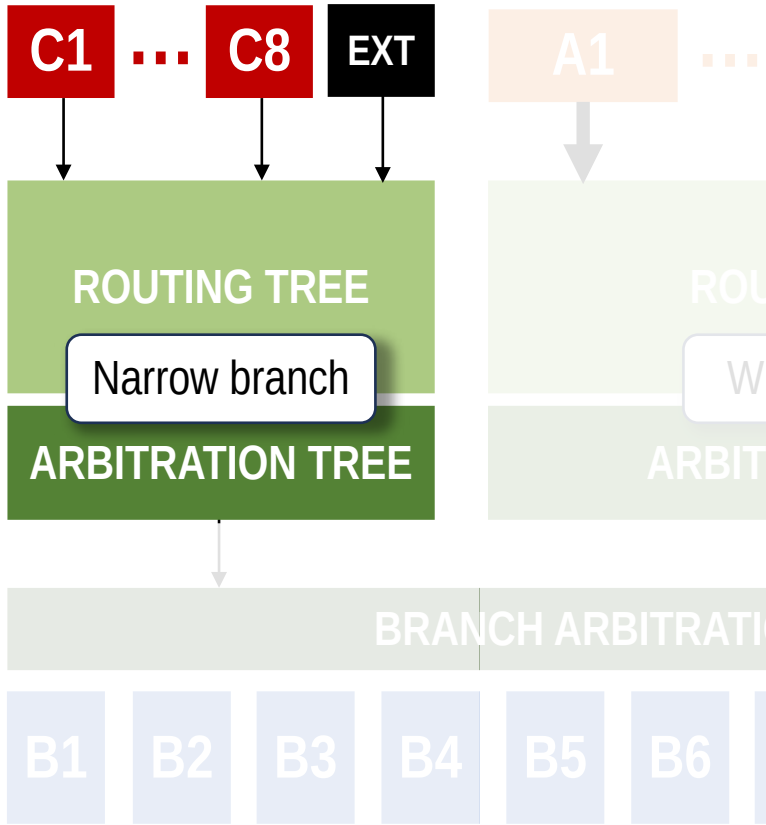
HCI: Heterogeneous Cluster Interconnect

- Word-granular access only where needed
- A **hybrid** interconnect
  - Separate initiators by **memory access pattern**



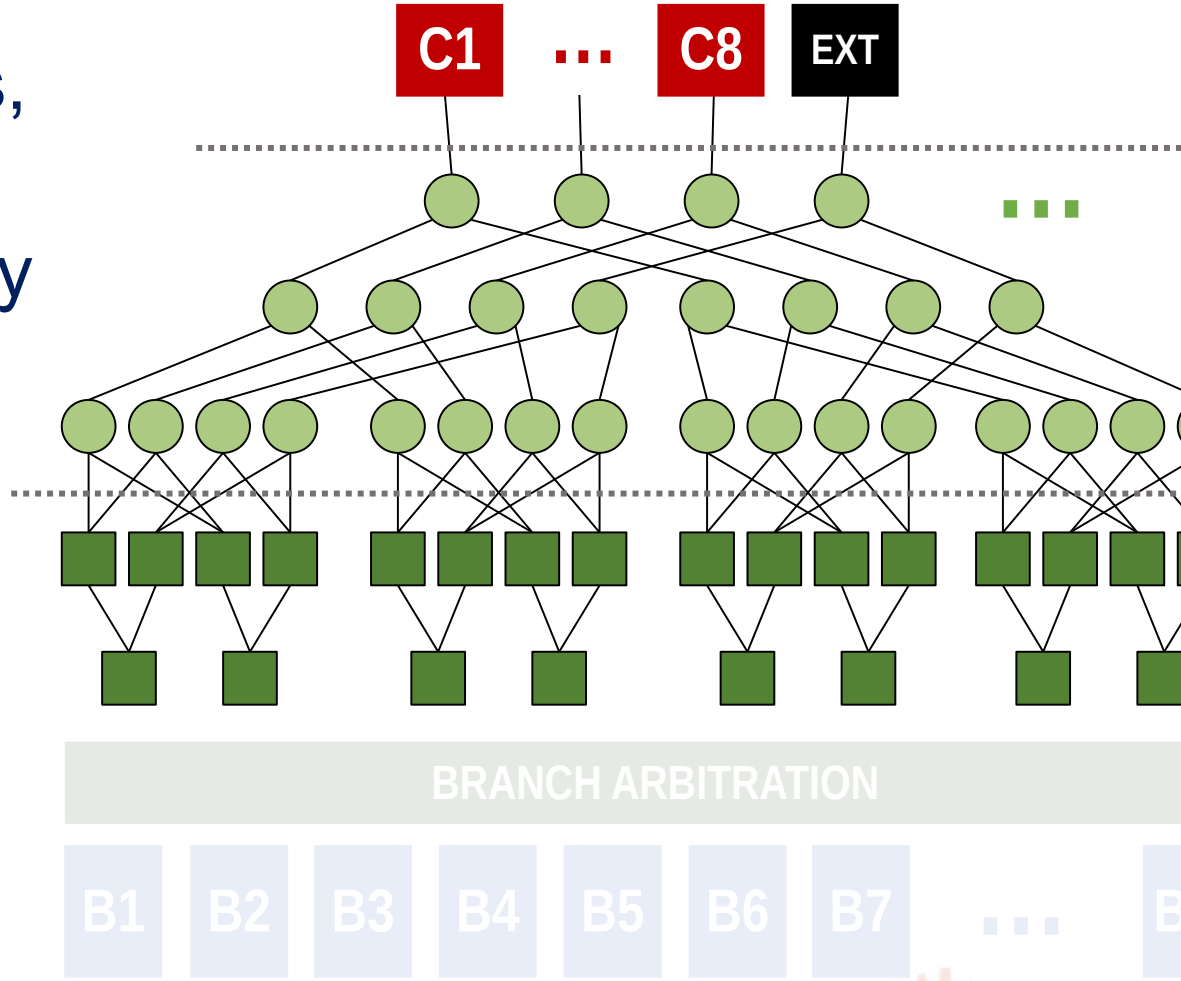
github.com/pulp-platform/hci

Architecture: **narrow HCI branch** for word-level access

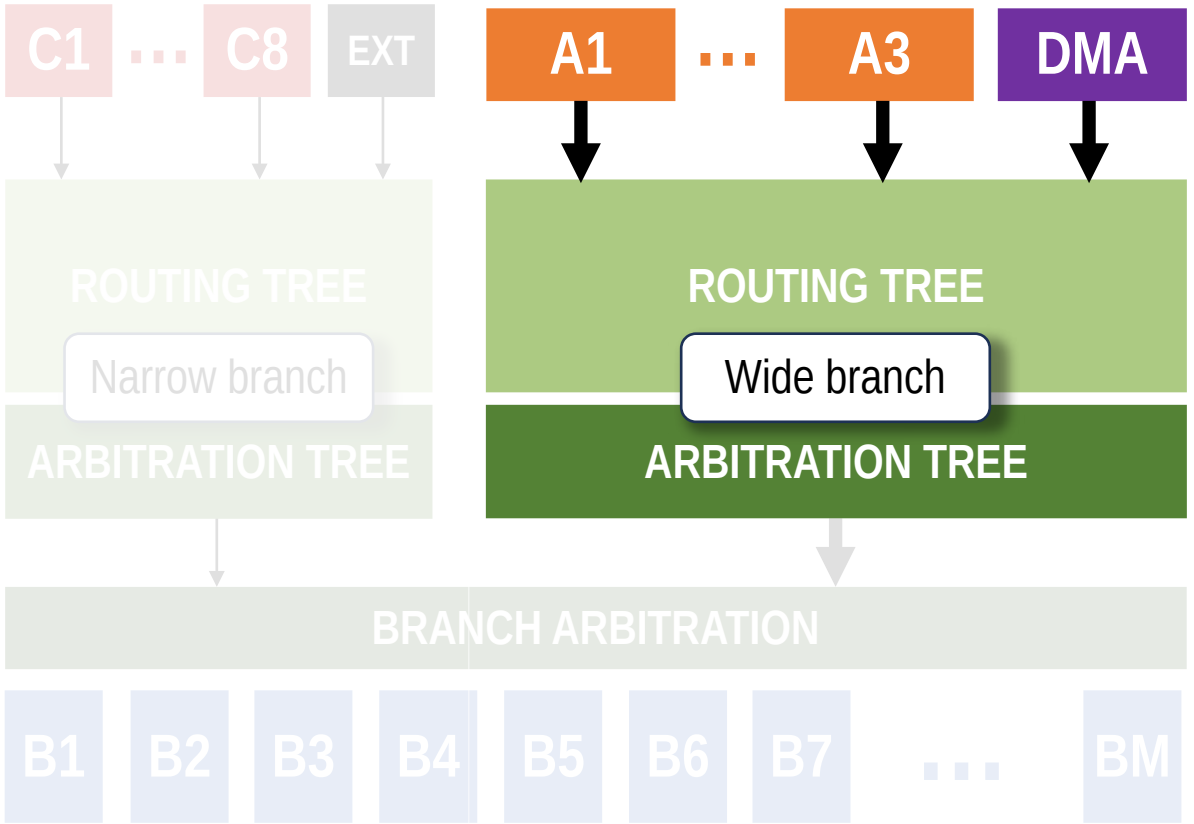


Architecture: **narrow HCI branch** for word-level access

- **Word-granular**, individual accesses, max flexibility
- No assumption on address contiguity
- Flat, fully-connected log xbar

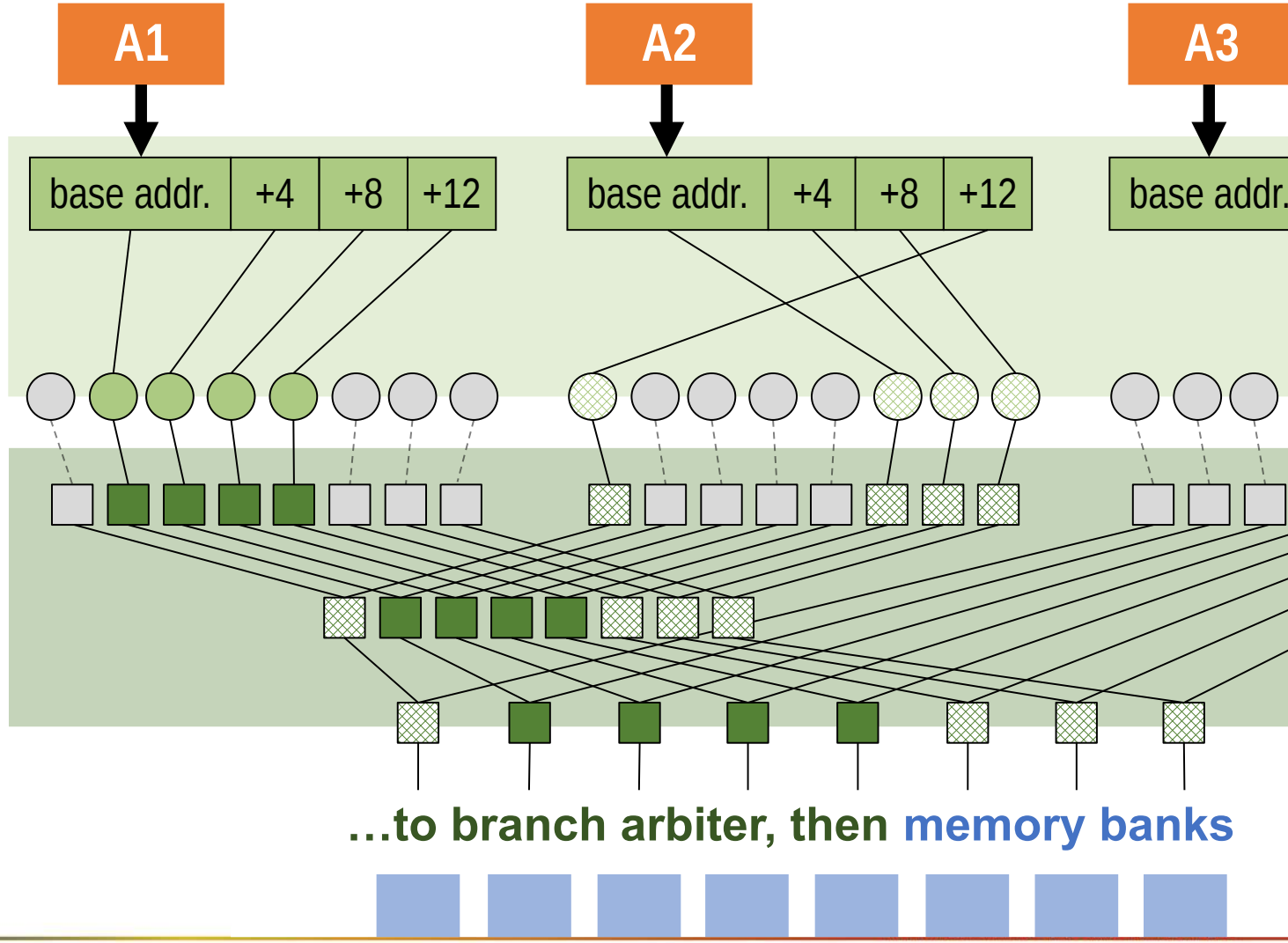


Architecture: **wide HCI branch** for vectorized accesses

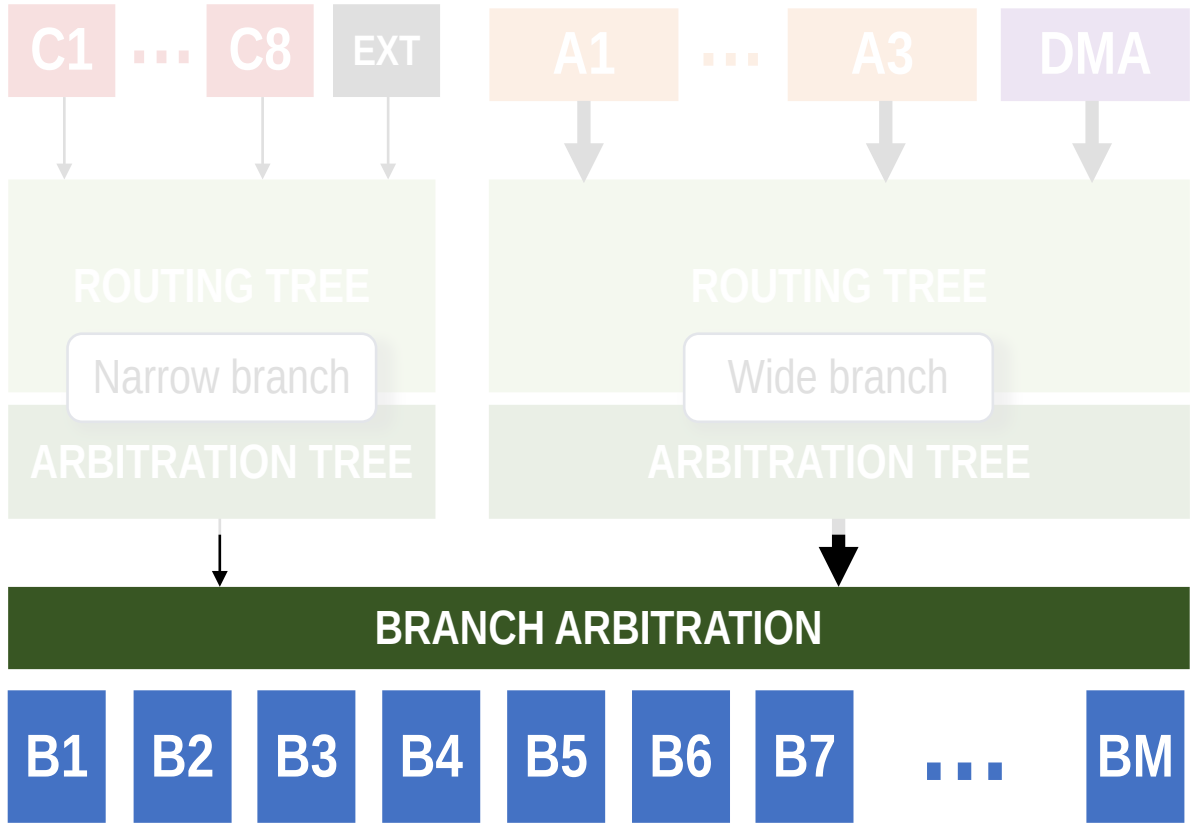


Architecture: **wide HCI branch** for vectorized accesses

- Wide access
  - **Multiple adjacent banks**
  - Simplified routing
- Shared req/gnt
  - by all sub-lanes
- Atomic grant
  - No need to track partial completion
- Binary arbiter tree

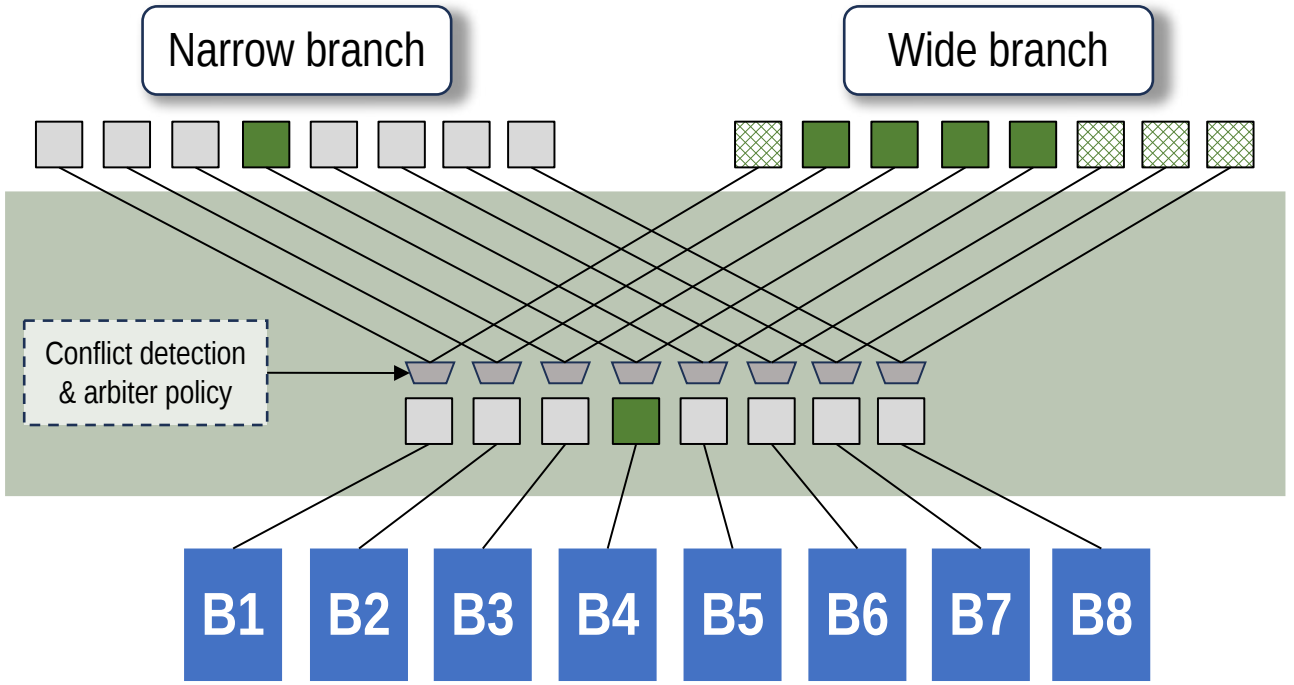


Architecture: **branch arbitration**



Architecture: **branch arbitration**

- Arbitrate conflicts between **narrow** and **wide** branch
- Runtime-configurable arbiter policy
  - Allocate desired bandwidth per branch



Physical scalability analysis

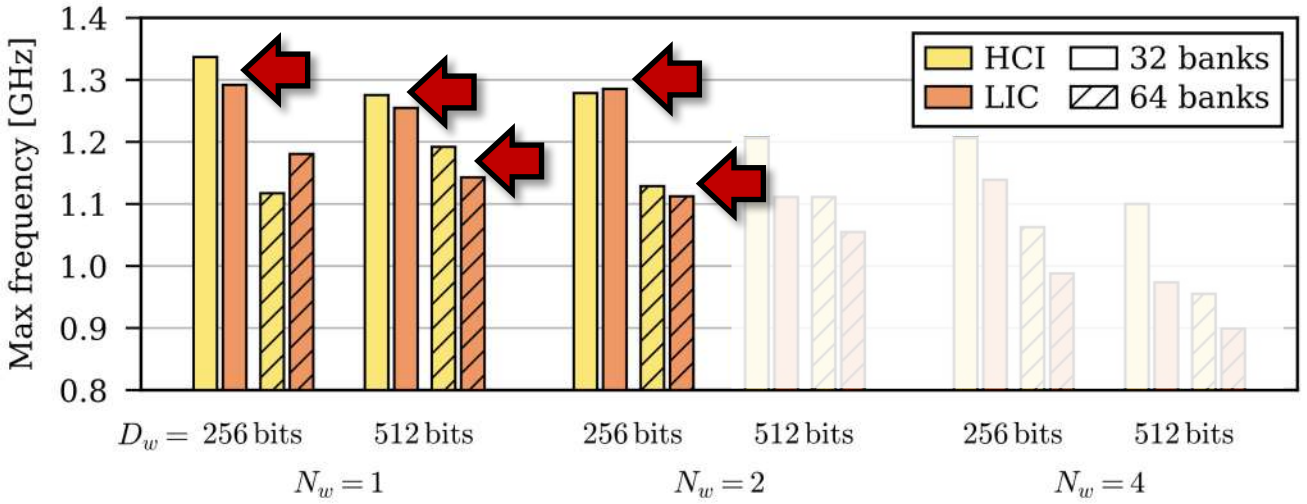
- Technology: GF 22FDX
- 24 interconnect configs
  - **9** 32-bit initiators
  - **1-4** wide initiators
    - **256/512-bit**
  - **32-64** TCDM banks
    - 128-256 KiB
  - **LIC** vs **HCI**
- 500 MHz – 1.5 GHz
- SS, 0.72V, - 40°C



HCI improves **max frequency** & **timing pressure**

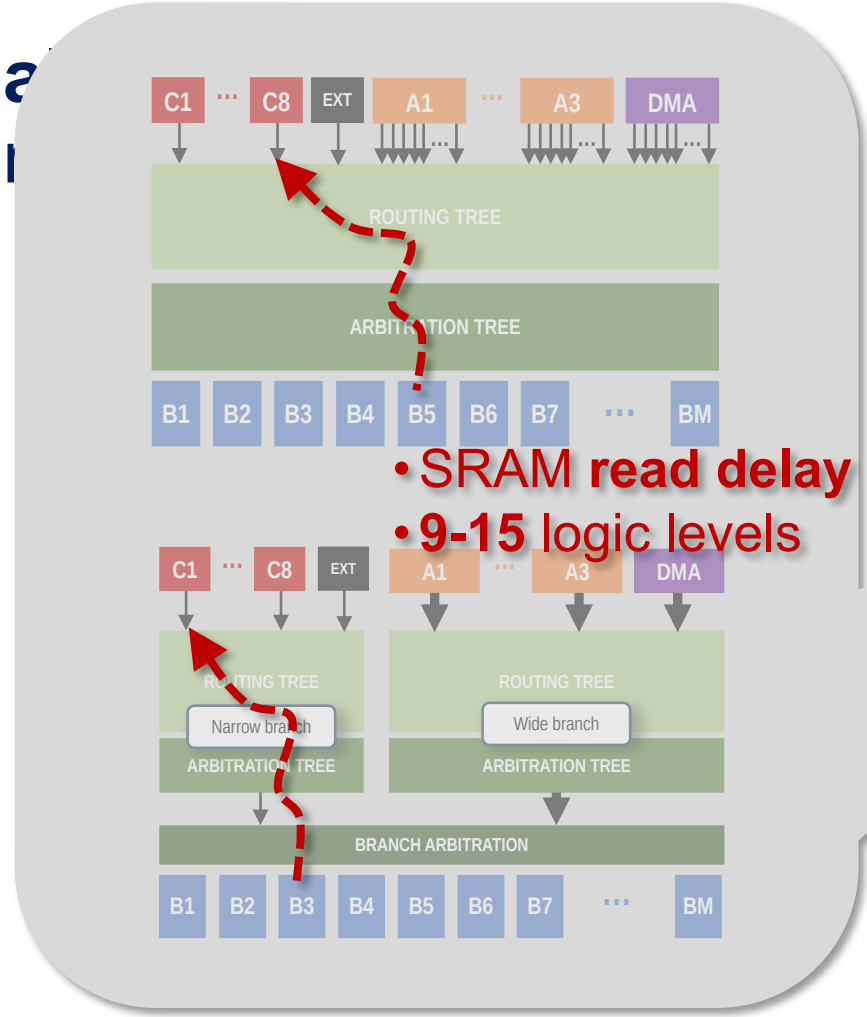
- **Small** configs have similar timing

Max frequency of all configs

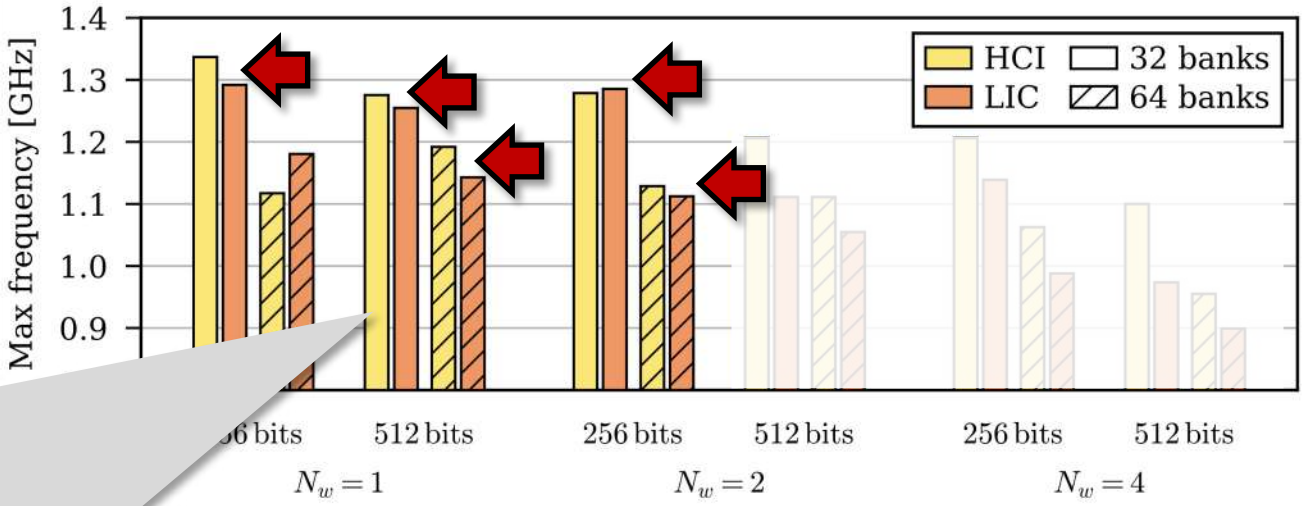


HCI improves **max frequency** & **timing pressure**

- **Sma**
  timi



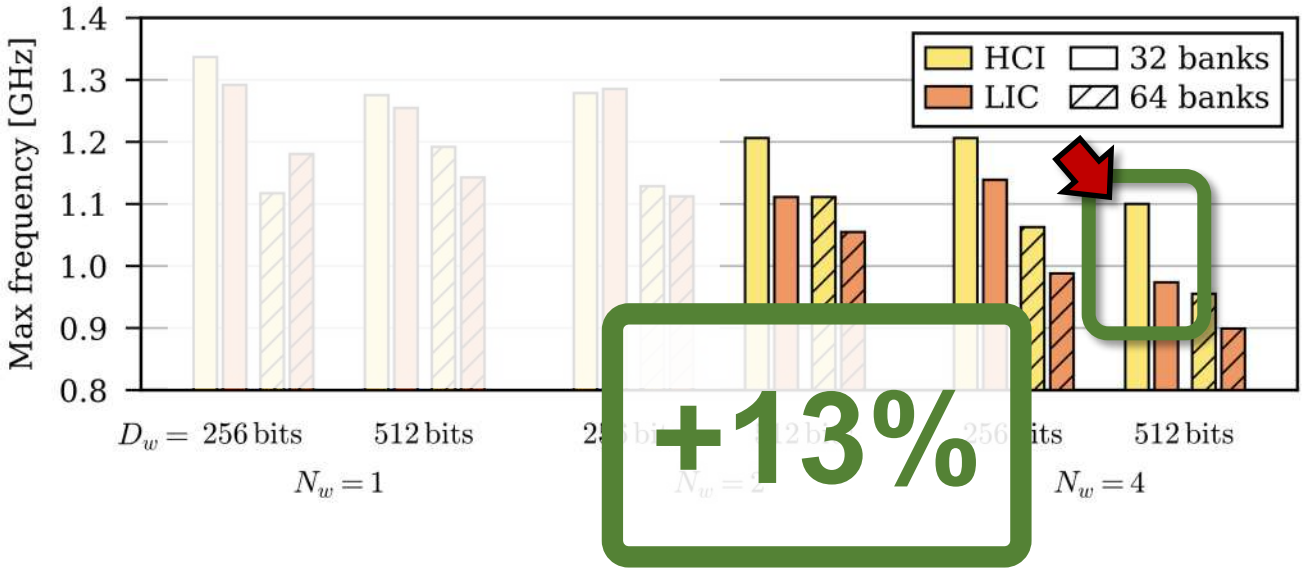
Max frequency of all configs



HCI improves **max frequency** & **timing pressure**

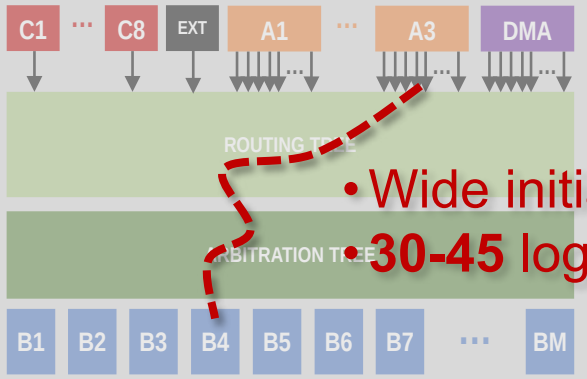
- **Small** configs have similar timing
- The **more complex**, the better HCI **max frequency**

Max frequency of all configs

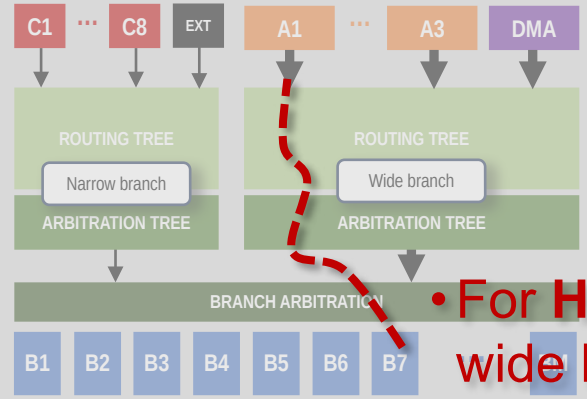


What alleviates timing pressure?

- **Sma**
  timi
- The
  bett

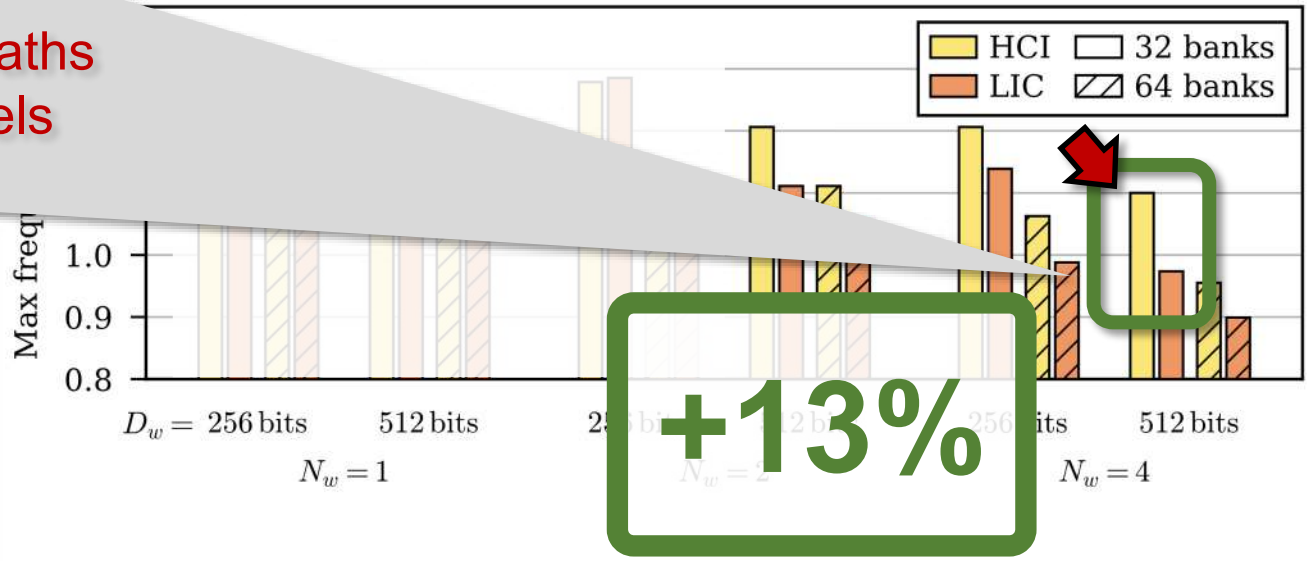


- Wide initiator paths
- **30-45** logic levels



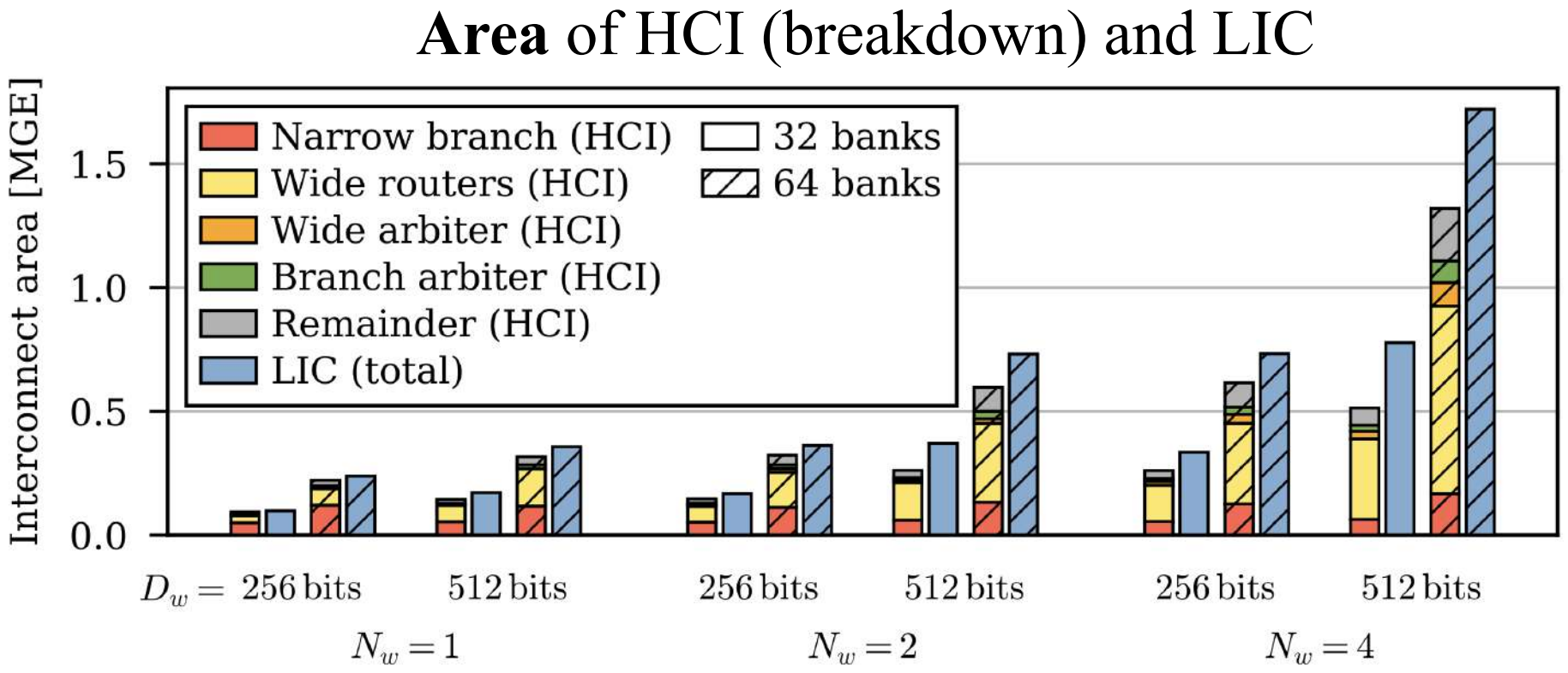
- For **HCI**: through wide branch
- **20-40** logic levels

Max frequency of all configs



Lower timing pressure means better **area scalability**...

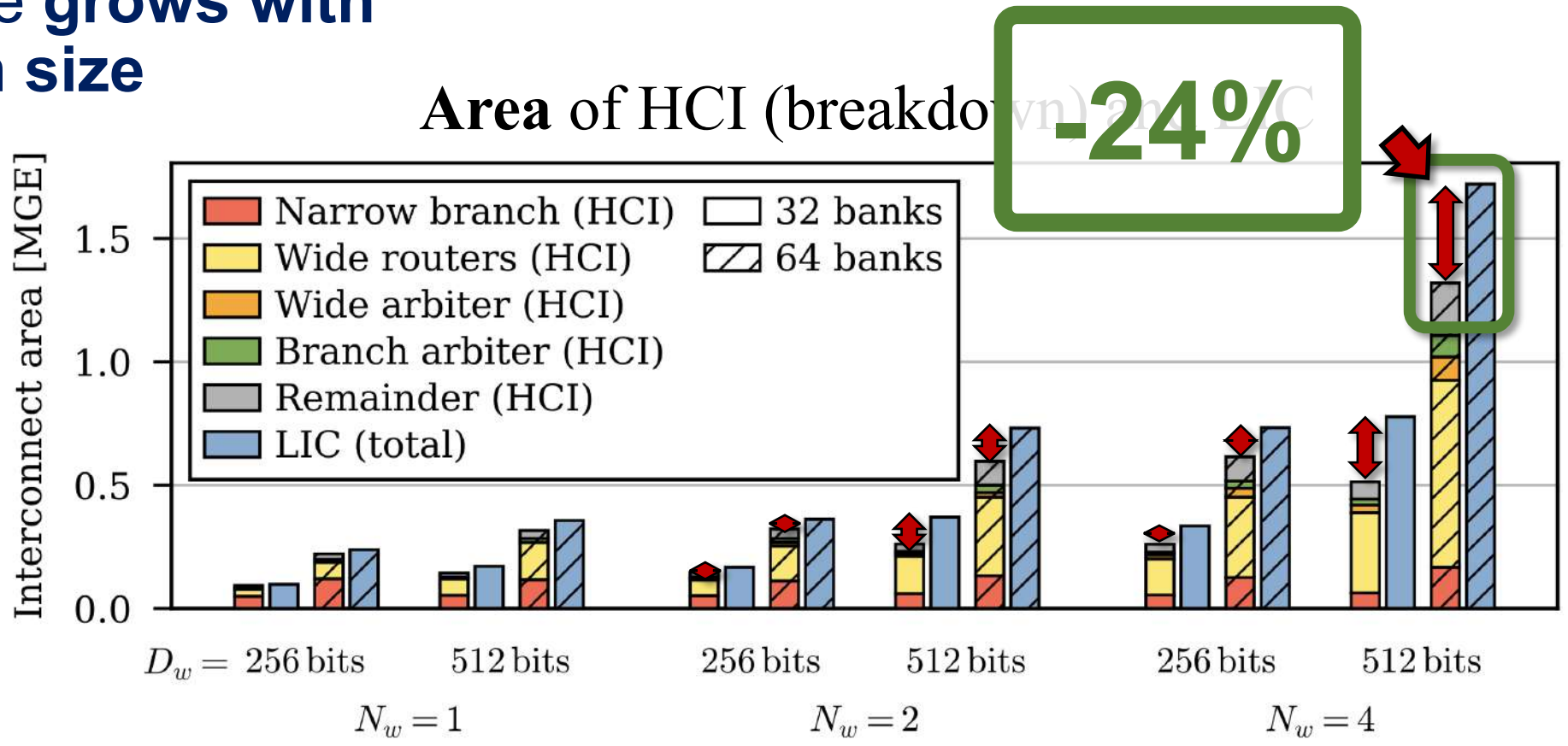
- Results for 1 GHz target



Lower timing pressure means better **area scalability**...

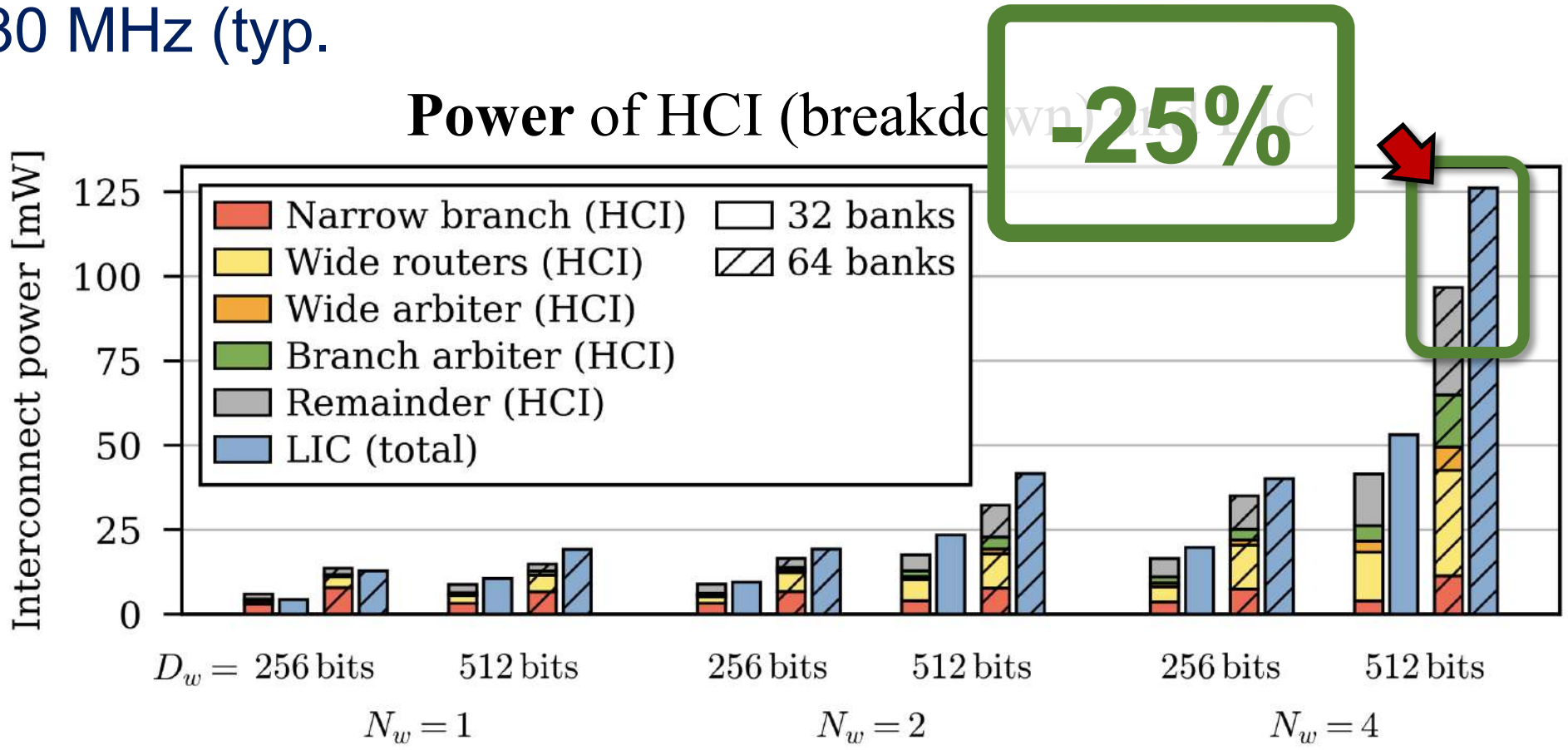
- Results for 1 GHz target
- HCI advantage **grows with configuration size**

Area of HCI (breakdown) and LIC



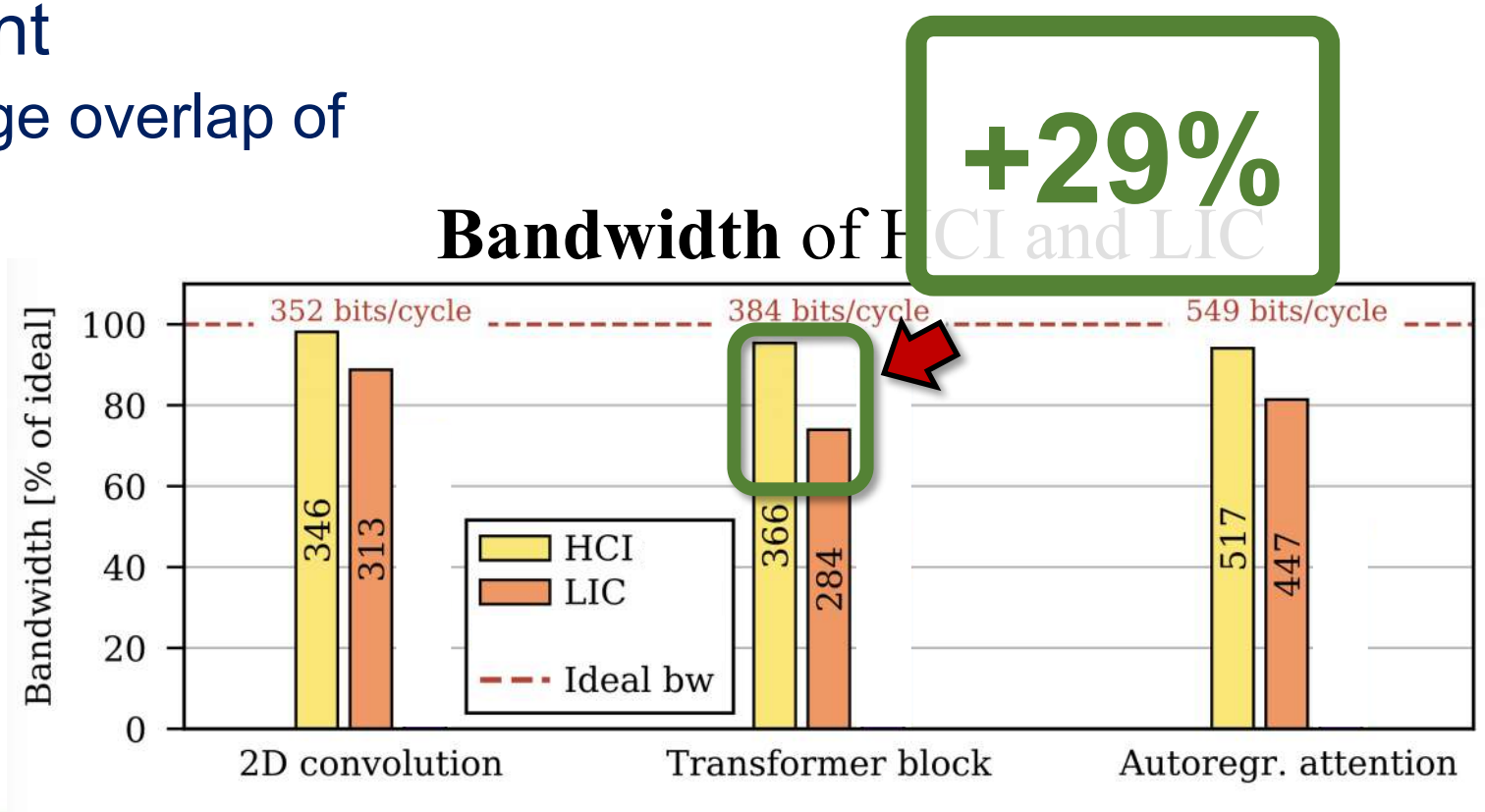
...which means also **lower power**!

- Double-buffered GEMM
- Running at 830 MHz (typ. conditions)



What about performance?

- **8** 32-bit cores, **3** 256-bit initiators
  - 64-bank 256 KiB TCDM
- Highest HCI improvement
  - High arith. intensity + large overlap of narrow/wide traffic



Conclusions

- The hierarchical architecture reduces **interconnect complexity**
  - Relaxes timing pressure
  - Reduces area
- **HCI** improves bandwidth over LIC
  - but not with a higher power cost (actually, smaller!)

+13% max freq

-24% leakage

up to **-24%**

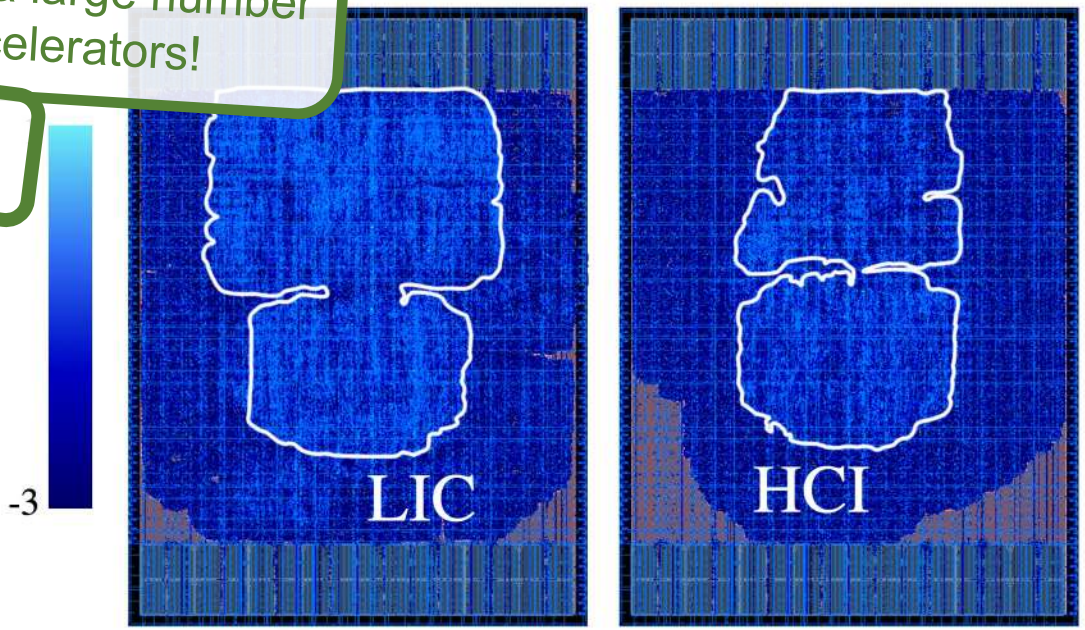
Low, uniform latency
Even with a large number of accelerators!

up to **+29%**

up to **-25%**

3D Routing congestion

9 32-bit, 4 512-bit initiators, 64 banks
@ 1.5 GHz target



github.com/pulp-platform/hci

ICCD 2026

ETH zürich

ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

The 44th IEEE International Conference on Computer Design

Nov. 16-18, 2026 Hong Kong SAR, China

Thank you!

pulp-platform.org

@pulp_platform

company/pulp-platform

youtube.com/pulp_platform

Sergio Mazzola smazzola@iis.ee.ethz.ch
Arpan Suravi Prasad
Davide Rossi
Luca Benini
Francesco Conti

Integrated Systems Laboratory (IIS), ETH Zürich, Switzerland
Department of Electrical, Electronic, and Information Engineering (DEI), University of Bologna, Italy

PULP Platform
Open-Source Hardware,
the way it should be!

