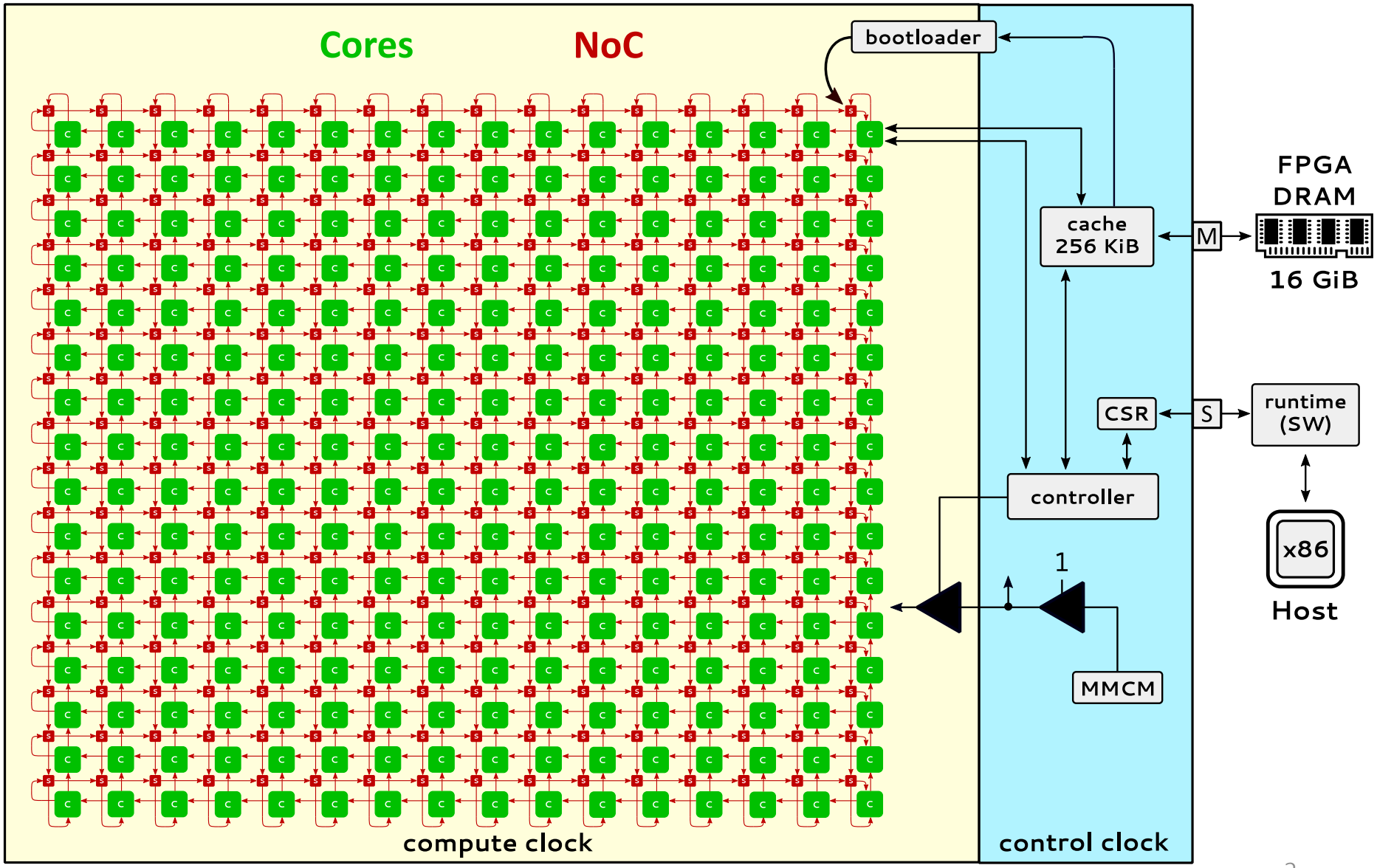
Manticore:
A 475 MHz Manycore FPGA Accelerator for RTL Simulation

Sahand Kashani[1], Mahyar Emami,
Keisuke Kamahori, Sepehr Pourghannad, Ritik Raj,
James R. Larus

[1]Now at MangoBoost

A Manycore RTL Simulation Accelerator

- 256 cores
  - Custom arch.
  - Custom ISA
- 2D torus NoC
  - Uni-directional
- No stall logic!
  - Static schedule
- 2 clock domains
  - Compute (**clock-gated**)
  - Control (always running)
- BSP programming model

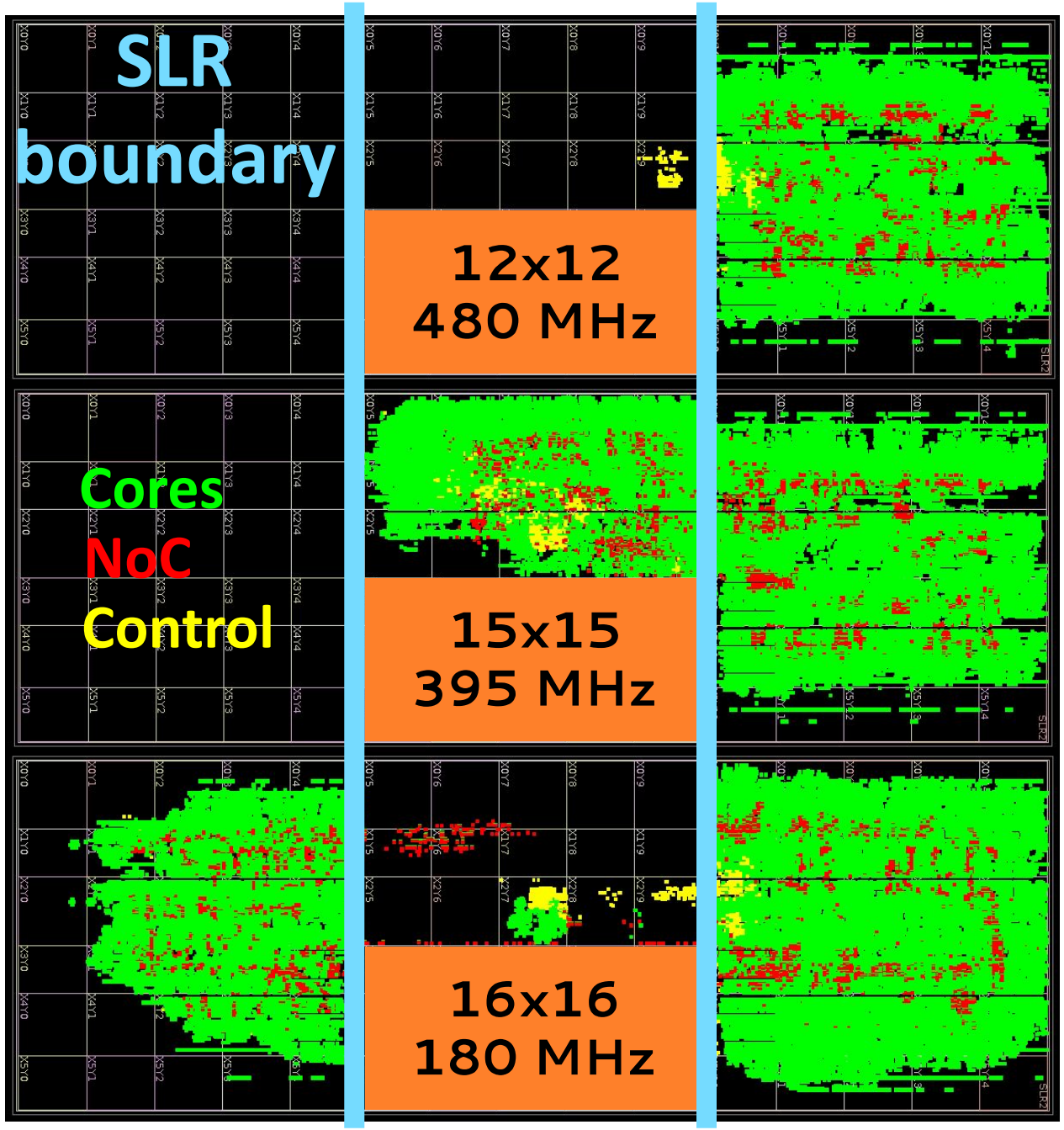


Unconstrained P&R

Alveo U200

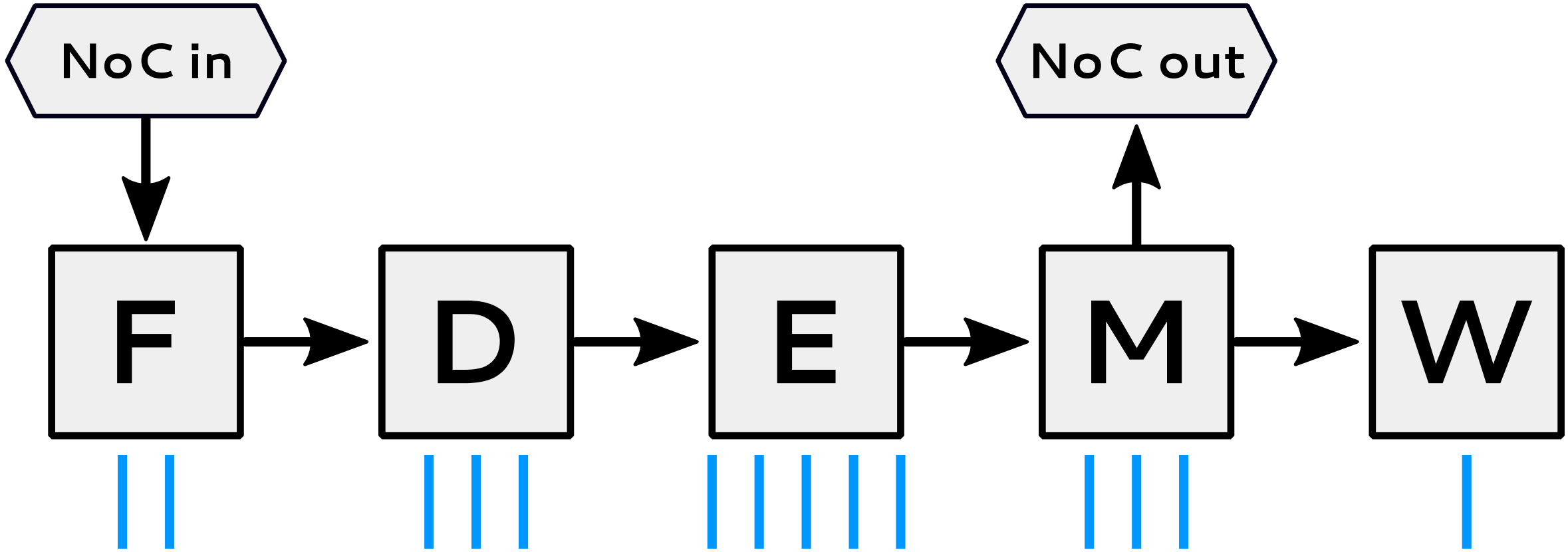
- Goal: High freq. & dense impl.

- QoR
  - 1 core @ 500 MHz
  - 256 cores @ 180 MHz
    → 2.7x degradation



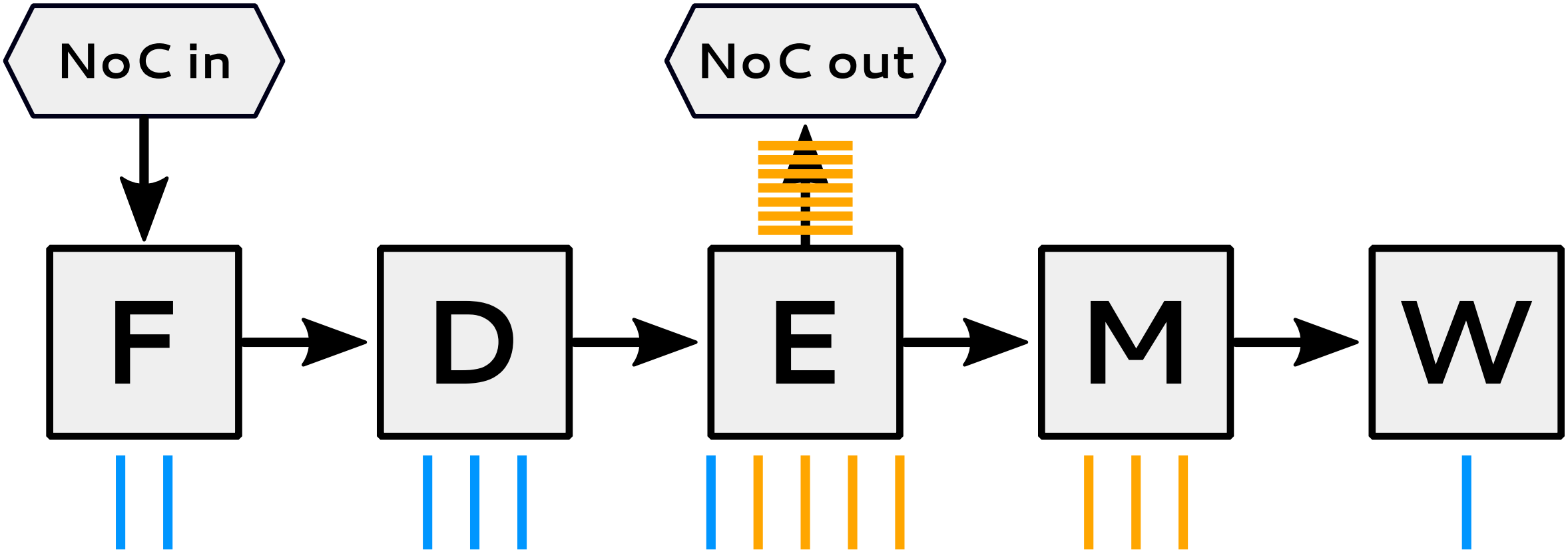
Decoupling cores from their NoC switches

Original 500 MHz core: Core placement tightly coupled with switch



Decoupling cores from their NoC switches

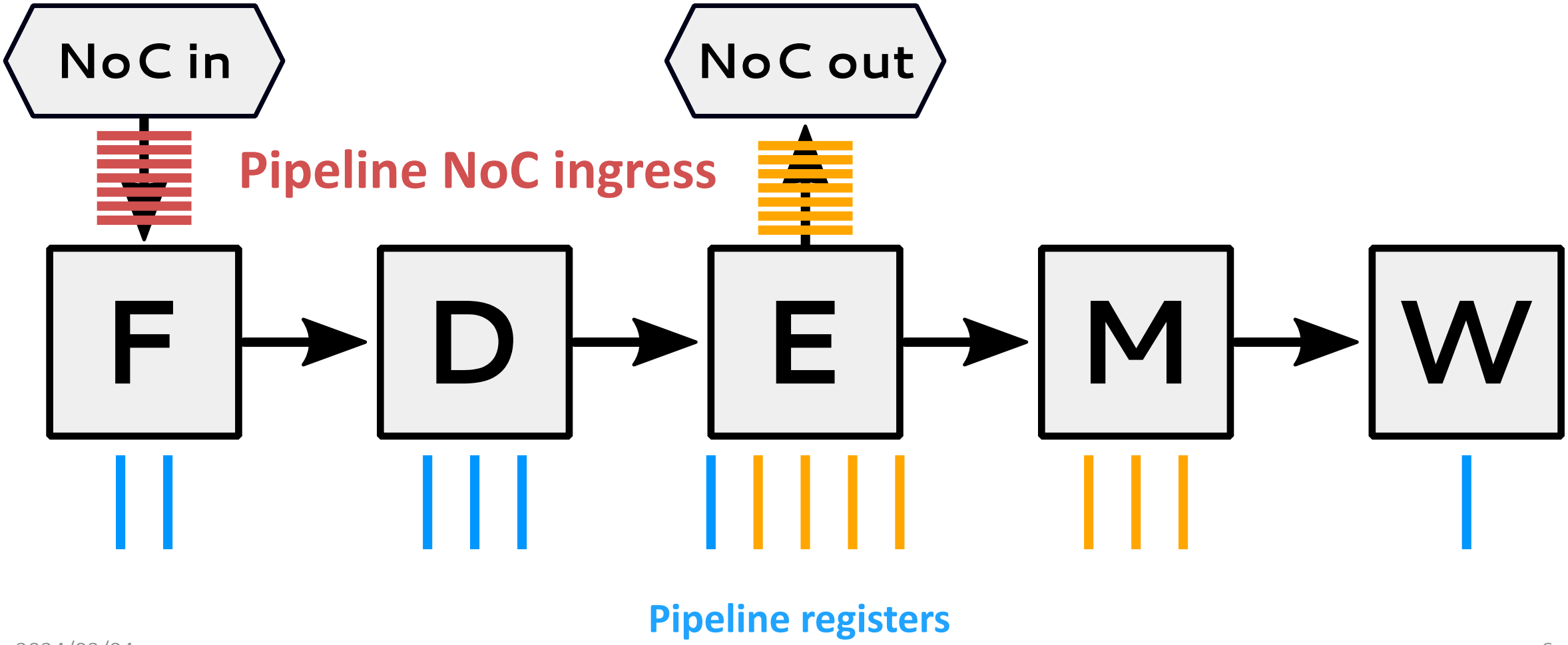
Move extra pipelining slack from inside the core to outside the core



Pipeline registers

Decoupling cores from their NoC switches

Move extra pipelining slack from inside the core to outside the core



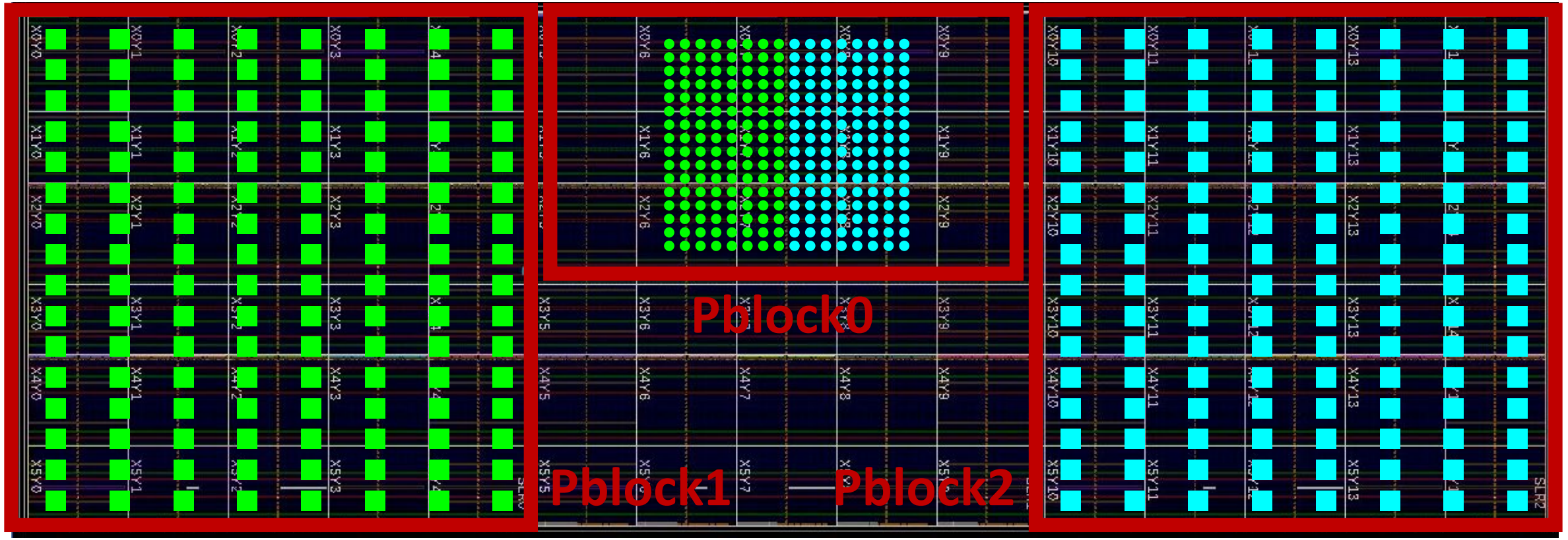
Split core-switch floorplan

cores (left)

switches (left)

switches (right)

cores (right)



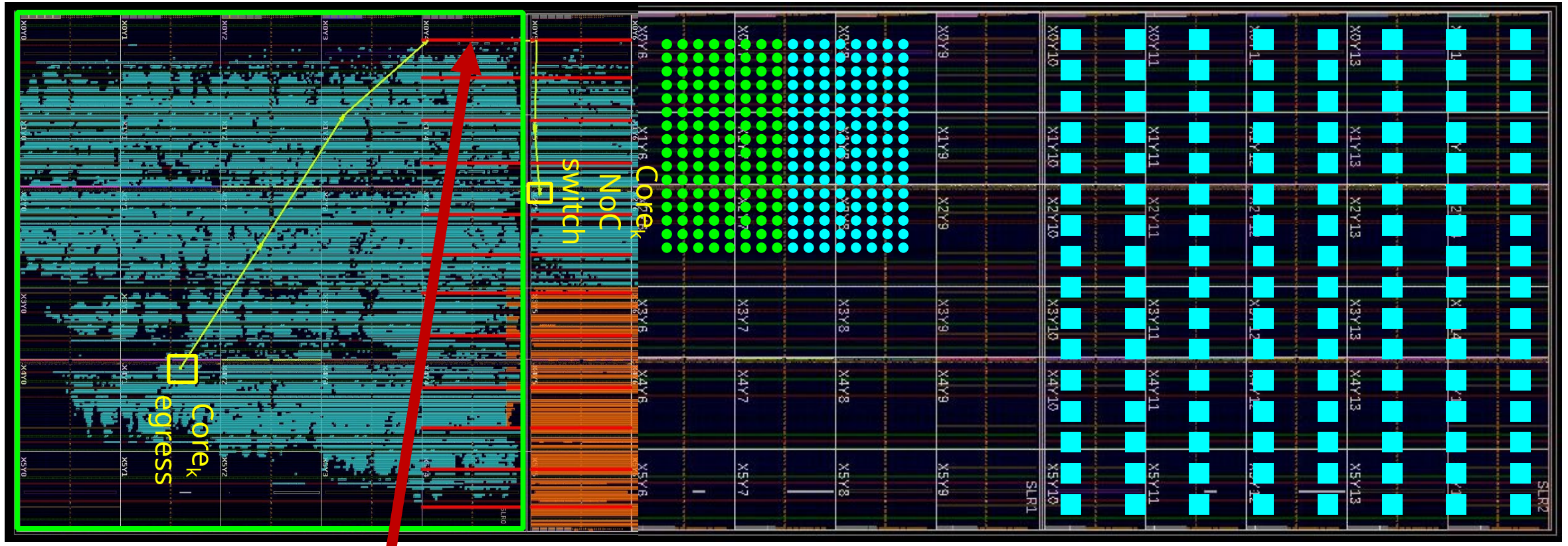
Split core-switch floorplan

cores (left)

switches (left)

switches (right)

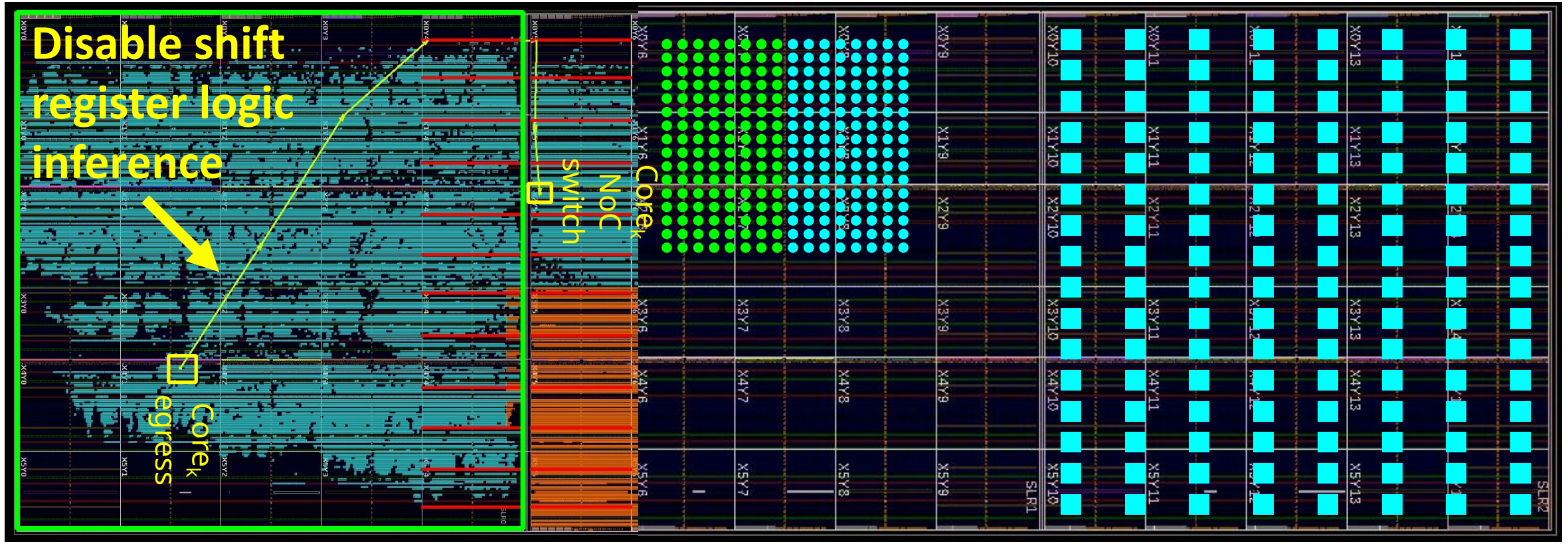
cores (right)



Hard inter-SLR pipeline registers

Split core-switch floorplan

cores (left) | switches (left) | switches (right) | cores (right)



Additional timing optimizations

- Minimizing clock skew between compute and control domains (Clock Root placement)
- Relatively-Placed Macros (RPM) for core register file
- Reset tree SLR boundary pipelining

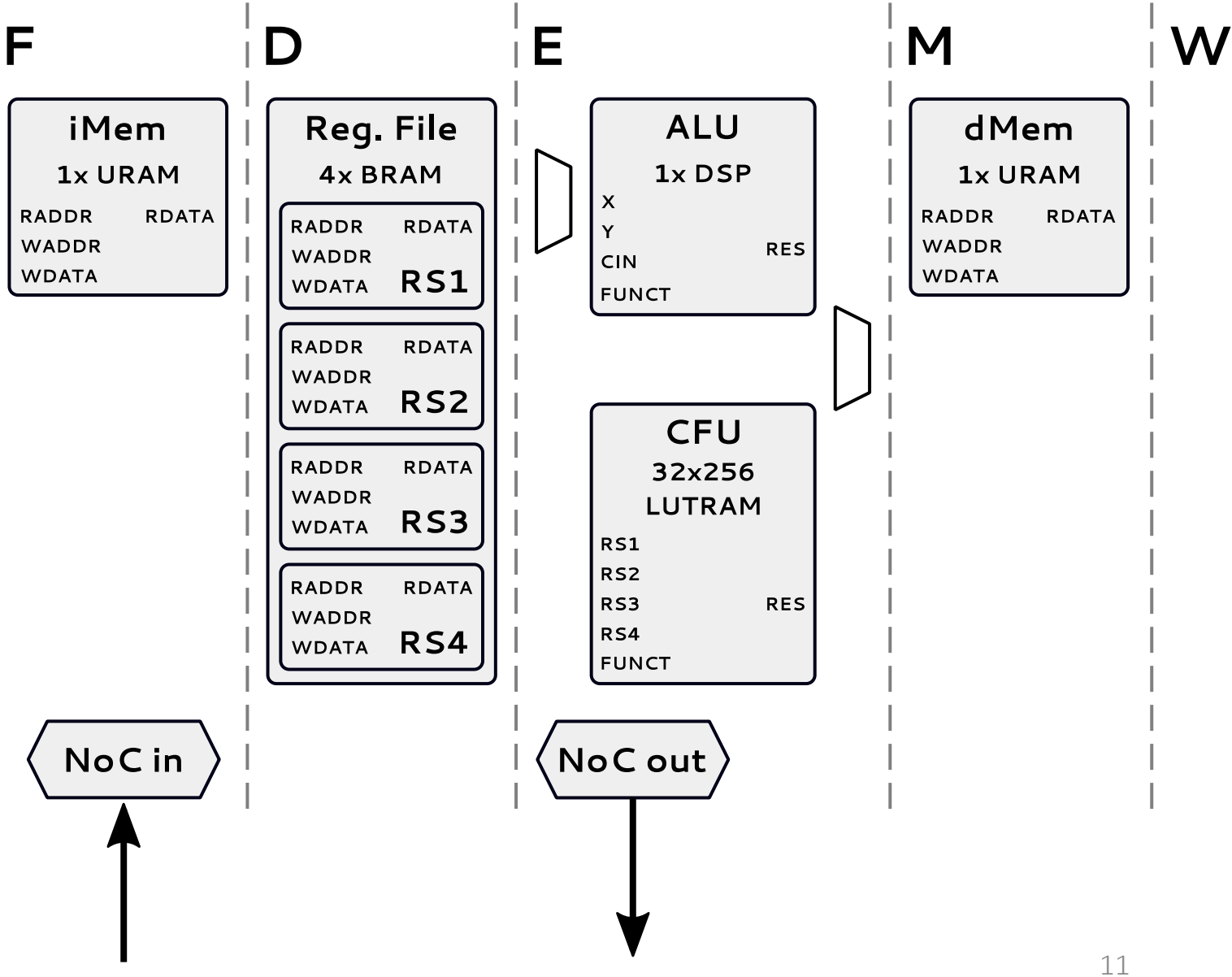
More details in the paper…

Reducing memory footprint

- Cores are memory-heavy
  - 4 BRAMs (large register file)
  - 2 URAMs (iMem + dMem)

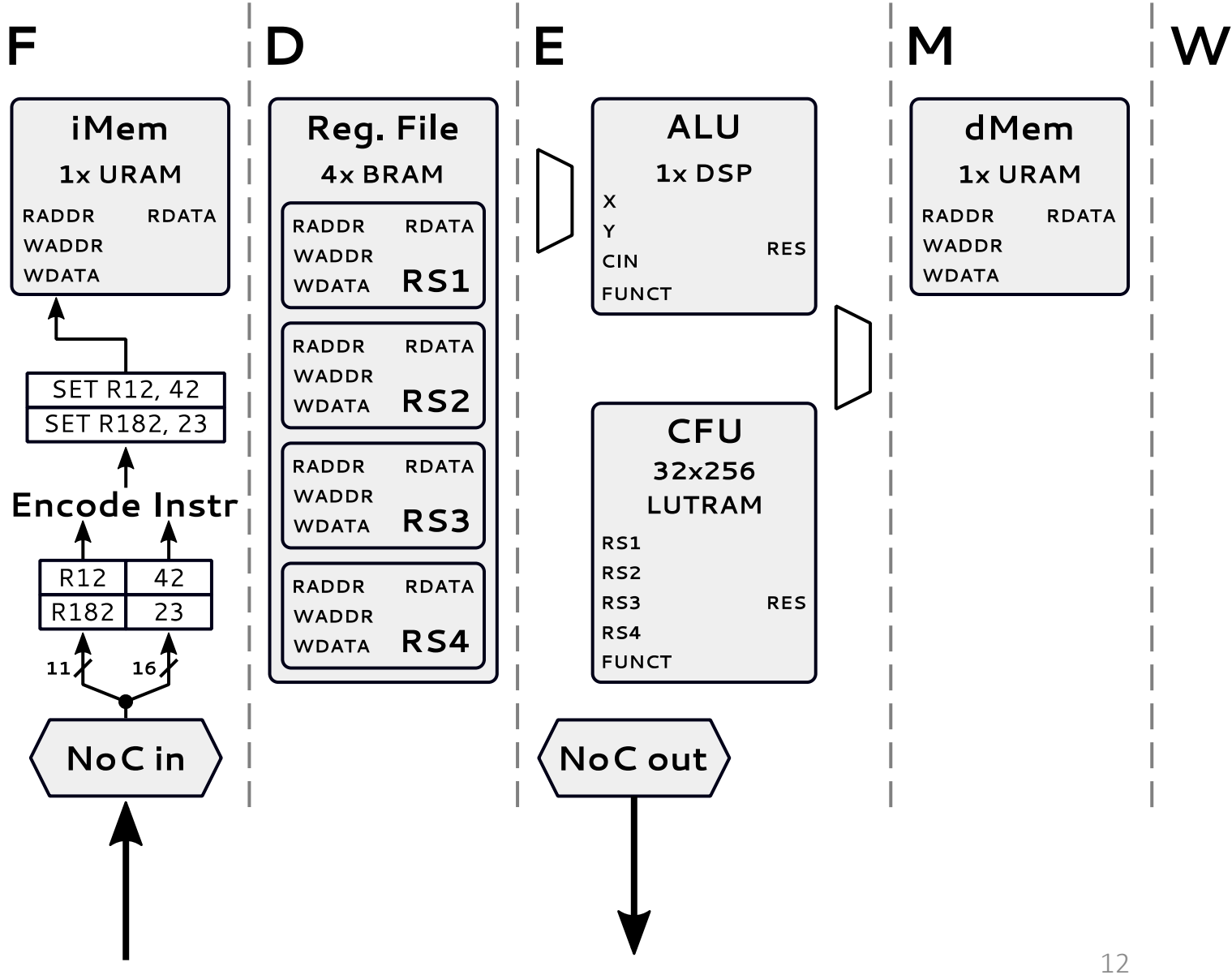
- Need queue for ingress writes received from other cores
  - Additional BRAM per core

  → Reduces total core count!

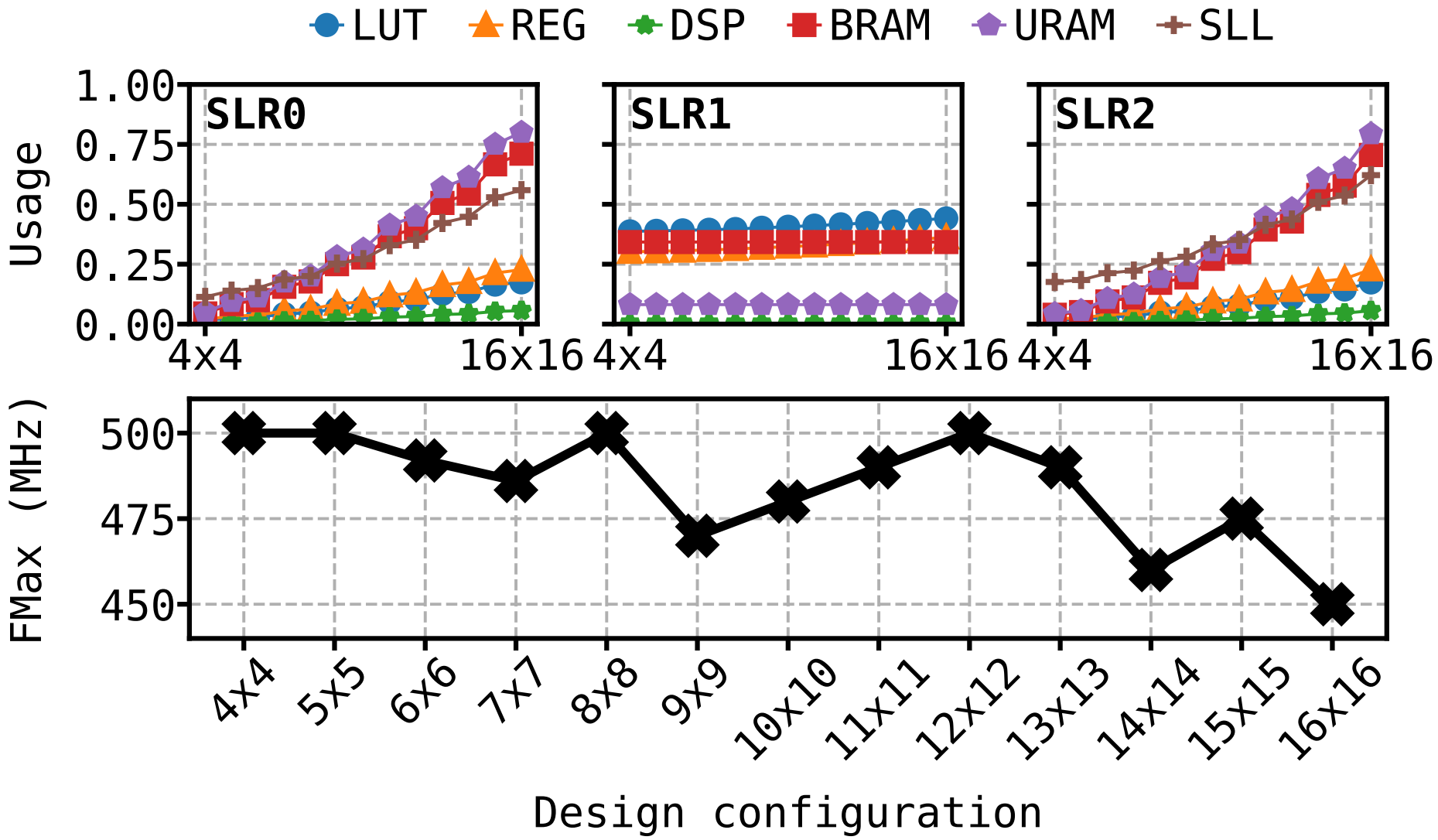


Reducing memory footprint

- Encode incoming writes as “set-immediate” instr.
- Push instr. to tail of iMem
  - Write port is unused after bootloading
- Core executes remote write when PC reaches it

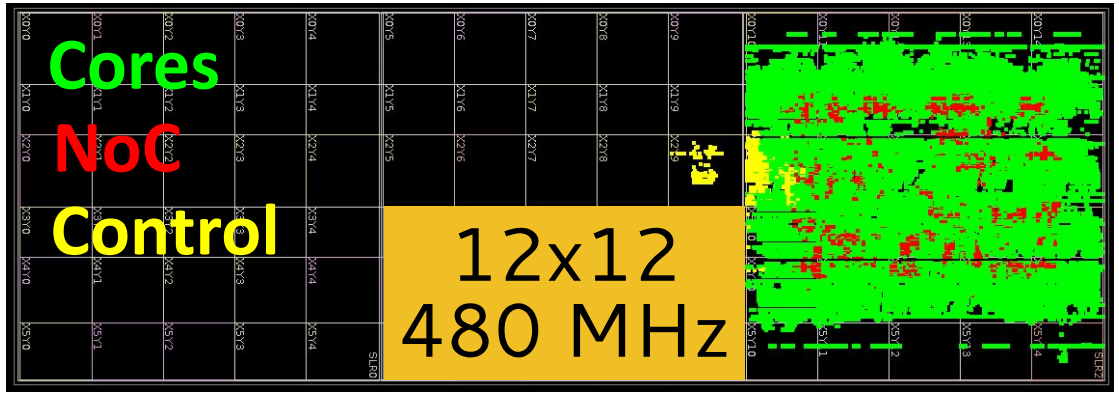


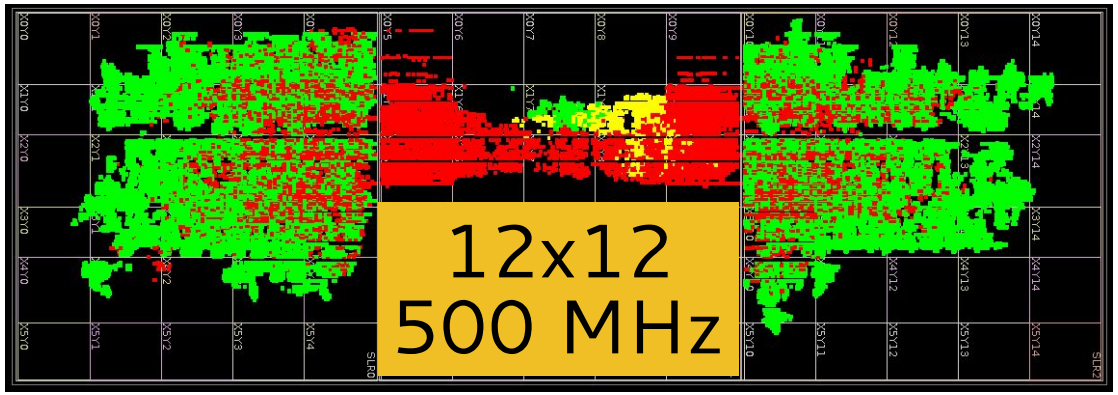
Quality of Results

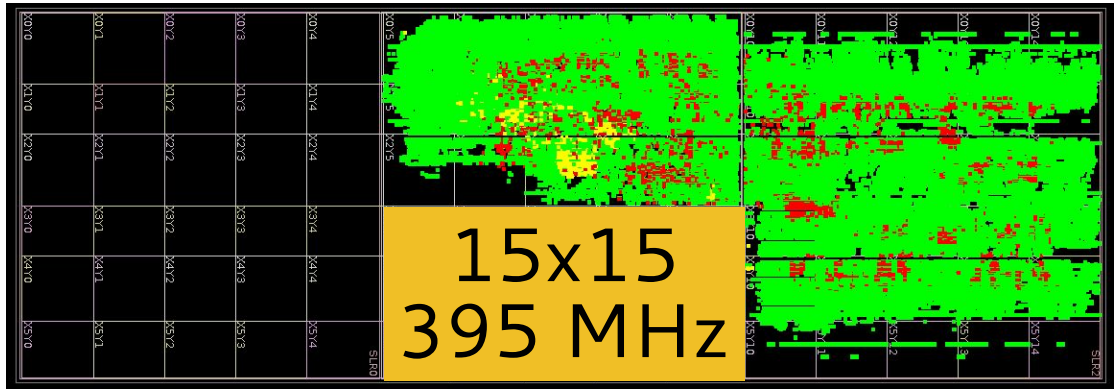


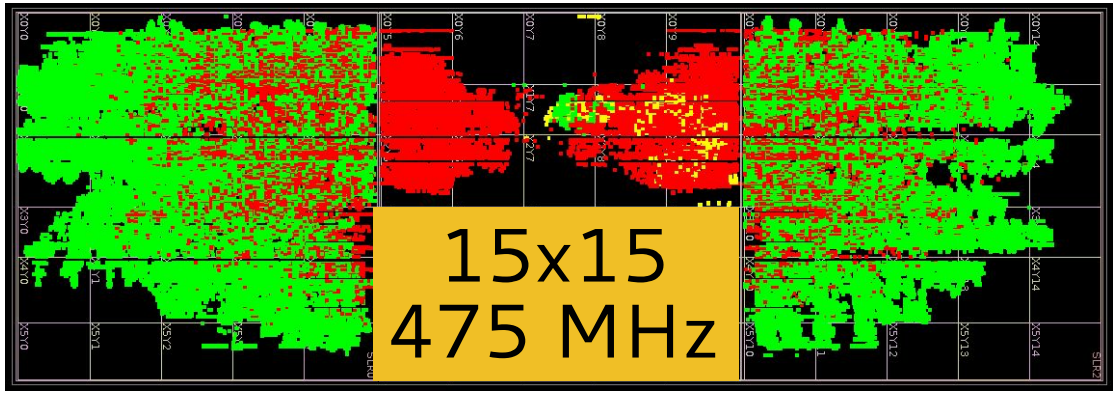
Auto P&R

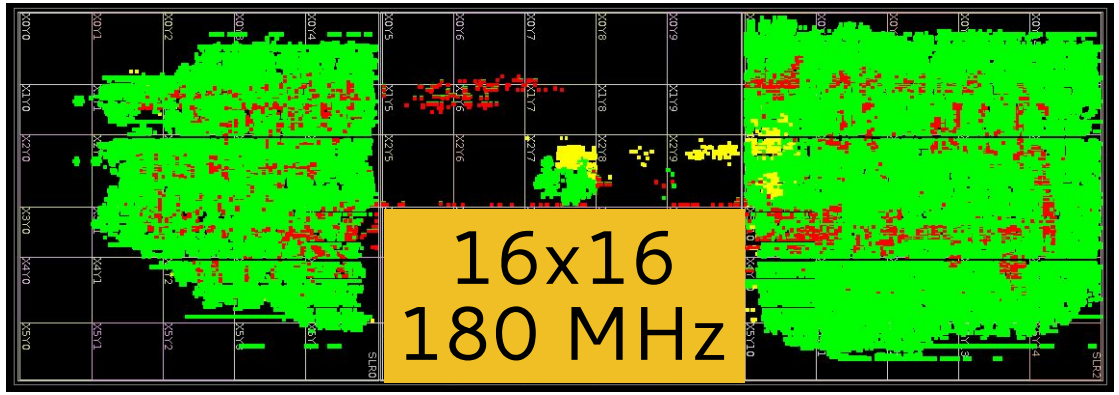
Split design

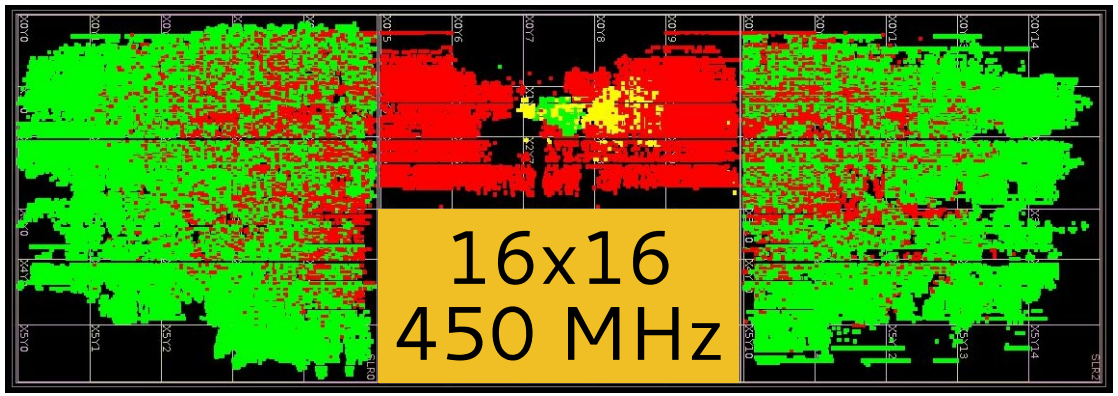












Summary

- Manticore is a manycore FPGA accelerator for RTL simulation
  - 225 cores @ 475 MHz (U200)
  - 256 cores @ 450 MHz (U200)
- Key contribution (physical implementation)
  - Decoupling core/switch placement
- Irregular placement can lead to good QoR

And there's more in the paper…

Architecture and application paper

RTL, compiler, runtime & pre-built bitstreams