A Case for Open EDA Verticals

Zhiru Zhang
School of ECE, Cornell University

March 2023 “at” ISPD
Panel on EDA for Domain Specific Computing

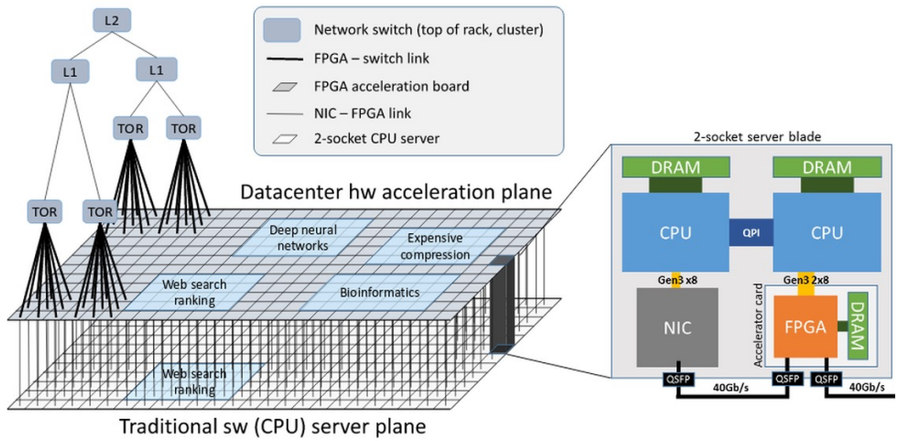
Cornell University

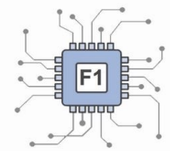
CSL

A Golden Age for Hardware Specialization

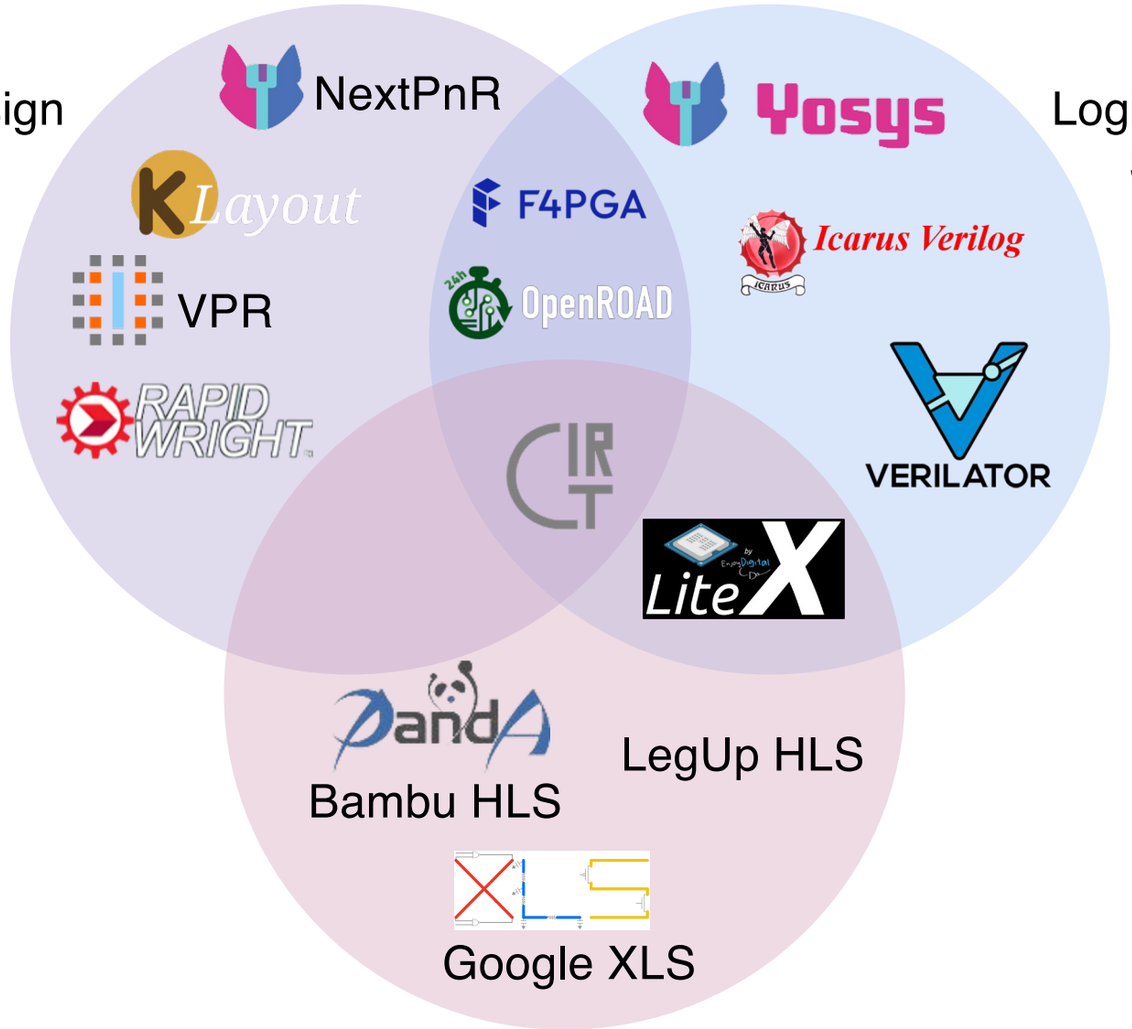
- **Higher demand:** emerging apps need to run faster with less energy use

- **Lower barrier:** open-source hardware & acceleration in cloud coming of age





Open-Source EDA Tools are Catching Up



Deployment of Open-Source HLS

CHIPS

Why YouTube decided to make its own video chip

With YouTube designing its own custom chips, the company joins a growing group of big tech companies seeking to offer something unique in the data center. Operating behind the scenes, Argos made YouTube's data centers much more efficient.

YouTube Argos chip, or video-coding unit (VCU)

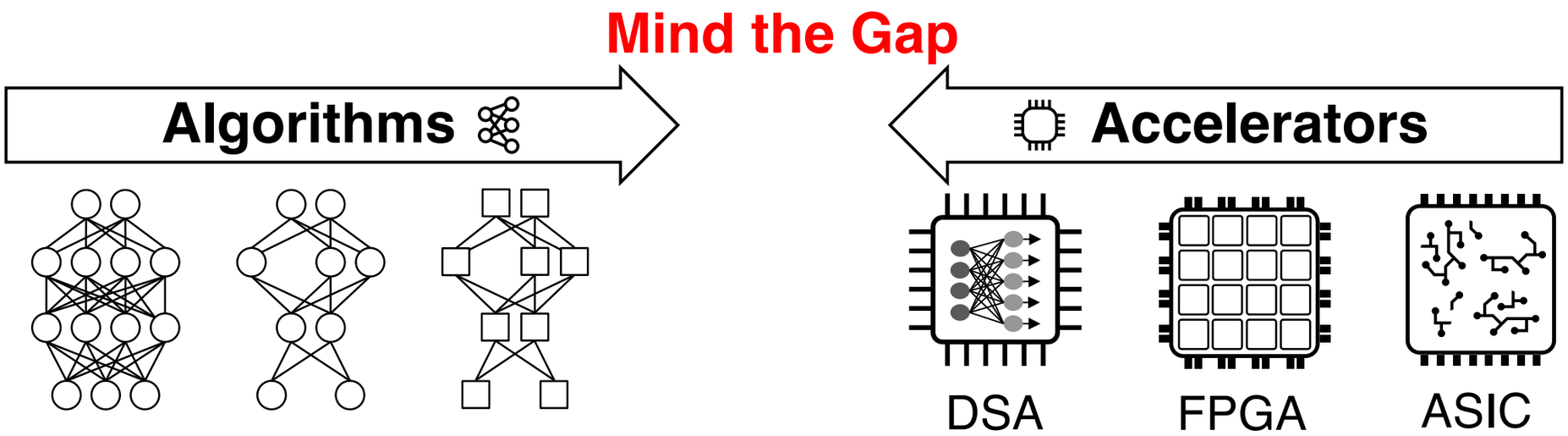
"… **Argos is a piece of hardware defined by software**, which meant that the engineers working on the chip could use what are called high-level synthesis techniques to iterate on the design much more quickly. **Google developed its own version of high-level synthesis software called Taffel that it used to help make the TPUs and the Argos processors** …"

https://www.protocol.com/enterprise/youtube-custom-chips-argos-asics (8/24/2022)

Caveat: currently only affordable by large enterprises for high-value apps

EDA for Pervasive Hardware Specialization

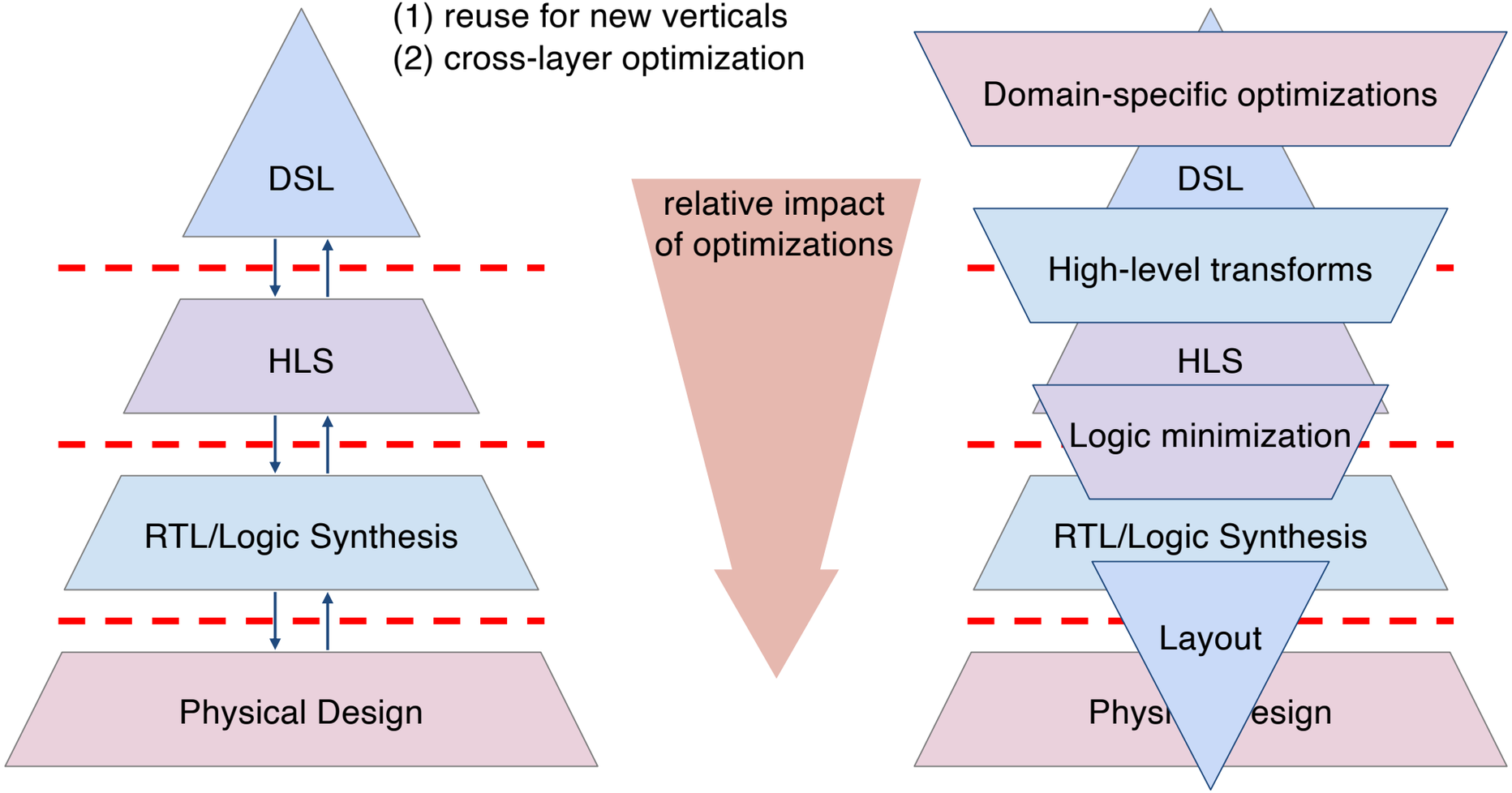
Can a small team of domain experts create and utilize accelerators for long-tail applications in a novel domain?



Opportunities: Open EDA Verticals

Open-standard interfaces facilitate:

(1) reuse for new verticals
(2) cross-layer optimization



Panel Statement

- We need IMPROVED design abstractions, not just higher-level ones
  - Develop systematic methods for exposing the search space, cost, and constraints
  - Provide open and interoperable interfaces to facilitate customization

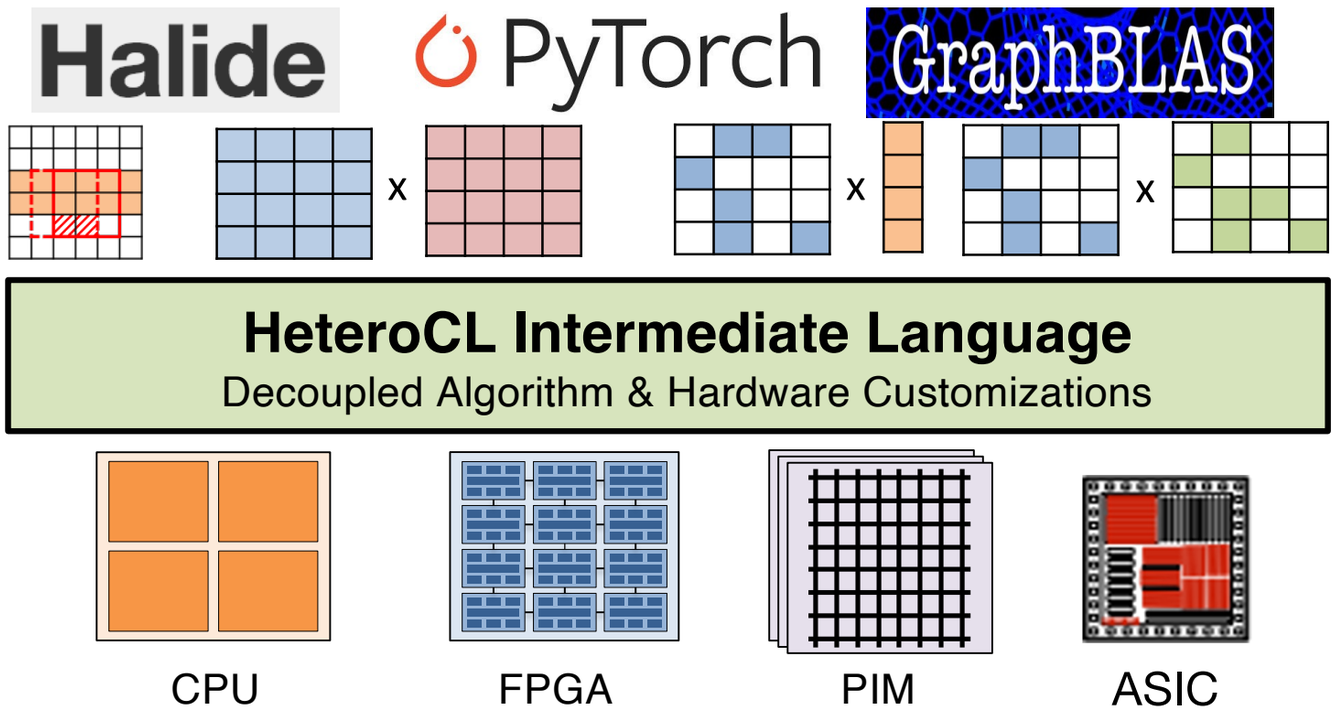
↳ Enable the construction of extensible DA infrastructures and reusable libraries

↳ Rapid customization of DA tool stack for a novel domain (i.e. the verticals)

↳ Productive end-to-end mapping from DSLs to accelerators

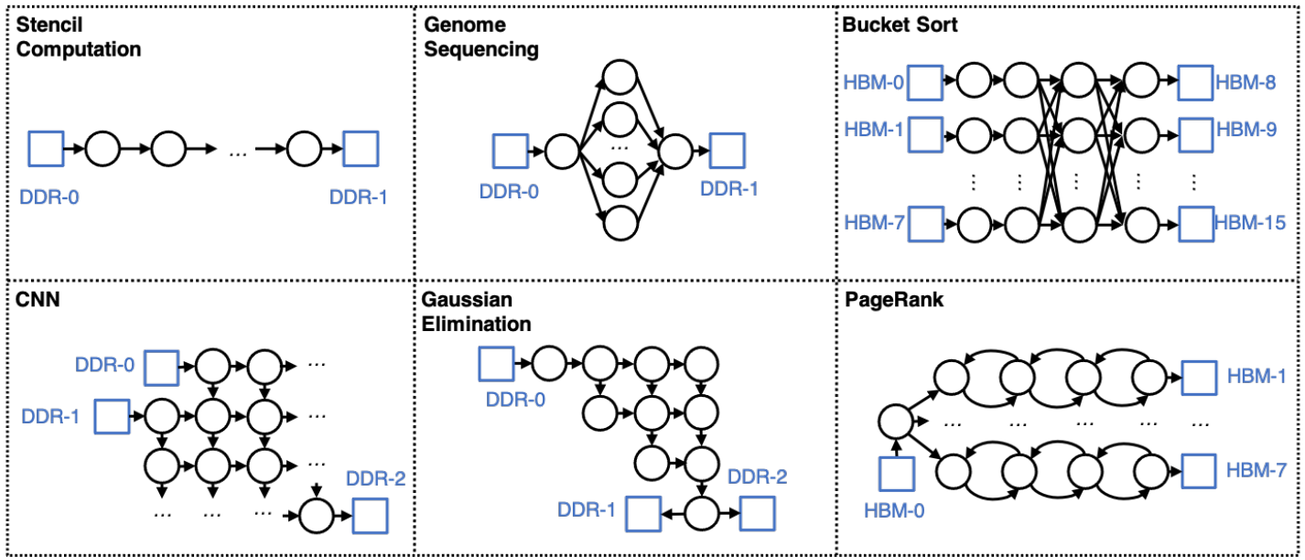
Example: HeteroCL Accelerator Design Framework

1. Productive, performant, and portable mapping to accelerators
2. Flexible support for dense & sparse workloads
3. Convenient interfaces to high-level DSLs



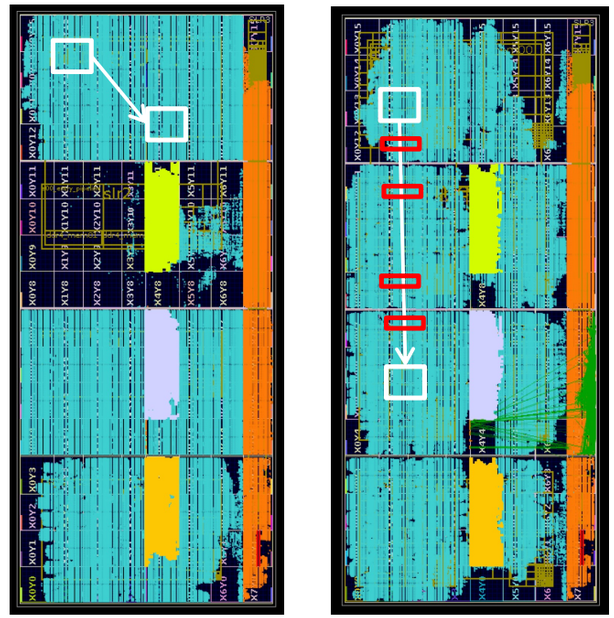
github.com/cornell-zhang/heterocl

AutoBridge: Coupling HLS and Floorplanning



A total of 43 design configurations

- 16 of them originally failed in routing with commercial tools
- AutoBridge improved frequency from 147 MHz to 297 MHz on average (2X increase)

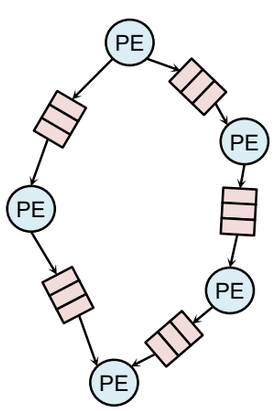


Default @ ~200 MHz (High local congestion)

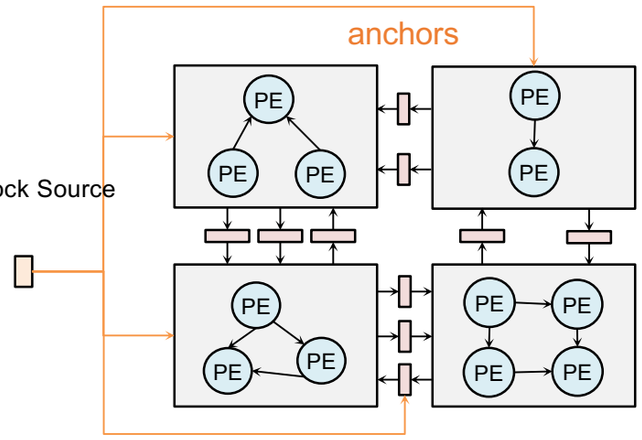
AutoBridge @ 300+ MHz (Long wire pipelined)

L. Guo, et al. AutoBridge: Coupling Coarse-Grained Floorplanning and Pipelining for High-Frequency HLS Design on Multi-Die FPGAs, FPGA'2021. (**Best Paper Award**)

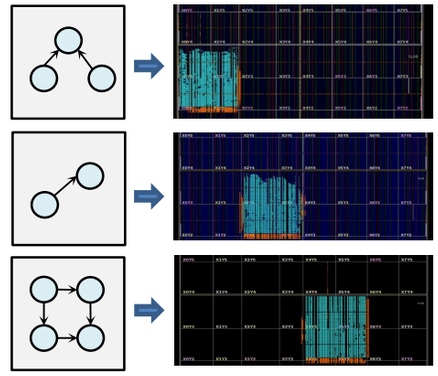
RapidStream: Parallel Split Compilation from HLS to Layout



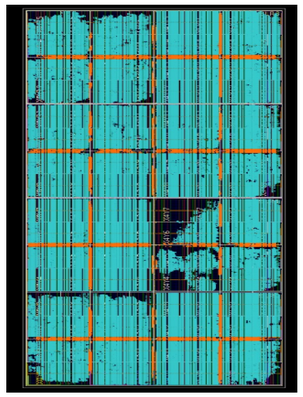
Input dataflow design
in C/C++



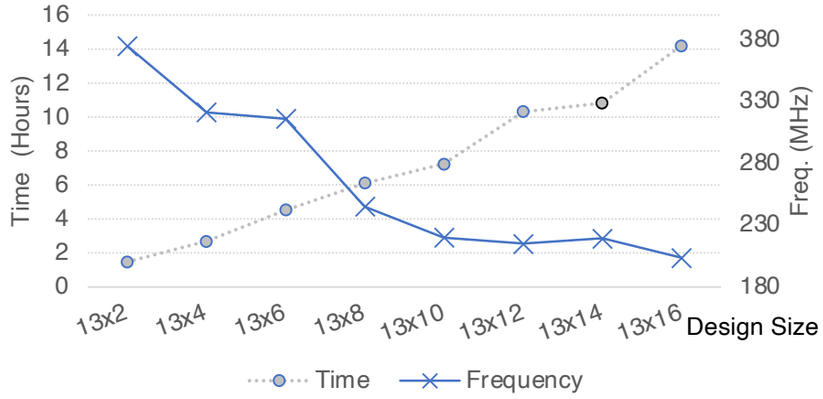
Partition the design into islands
Insert anchor registers

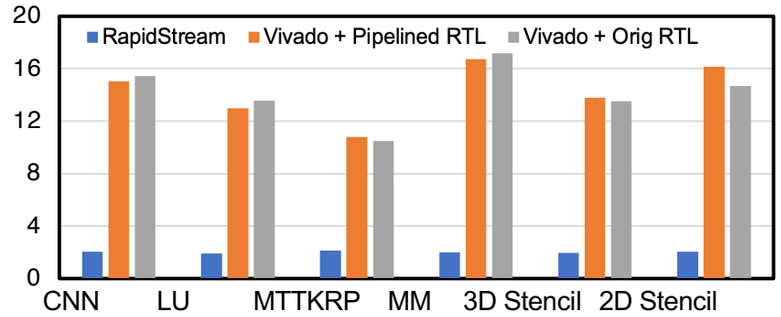


Islands and anchors
placed & routed in parallel



All islands stitched together
w/ inter-island routing





Compile time (hours)
5-7x reduction

L. Guo et al., RapidStream: Parallel Physical Implementation of FPGA HLS Designs, FPGA'22 **(Best Paper Award)**