



The University of Texas at Austin  
WHAT STARTS HERE CHANGES THE WORLD

# Koios: A Deep Learning Benchmark Suite for FPGA Architecture and CAD Research

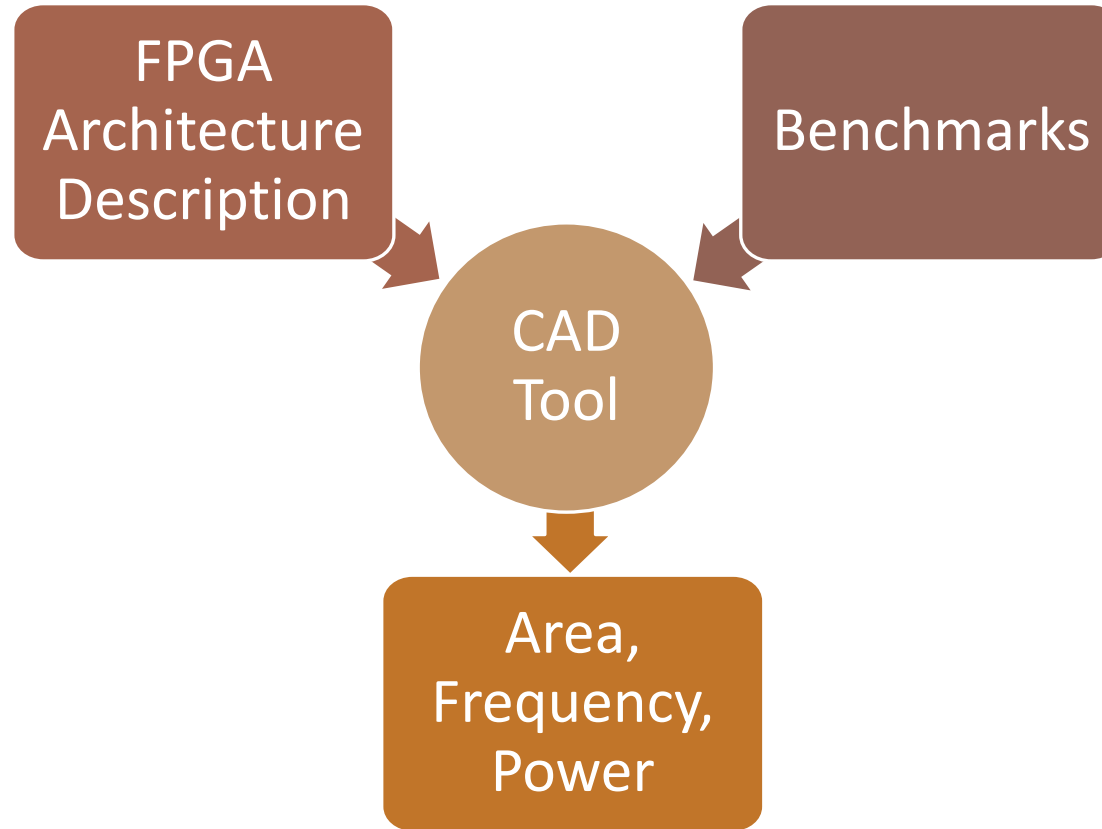
Aman Arora and **Lizy K. John**

*Open-Source Computer Architecture Research (OSCAR)*

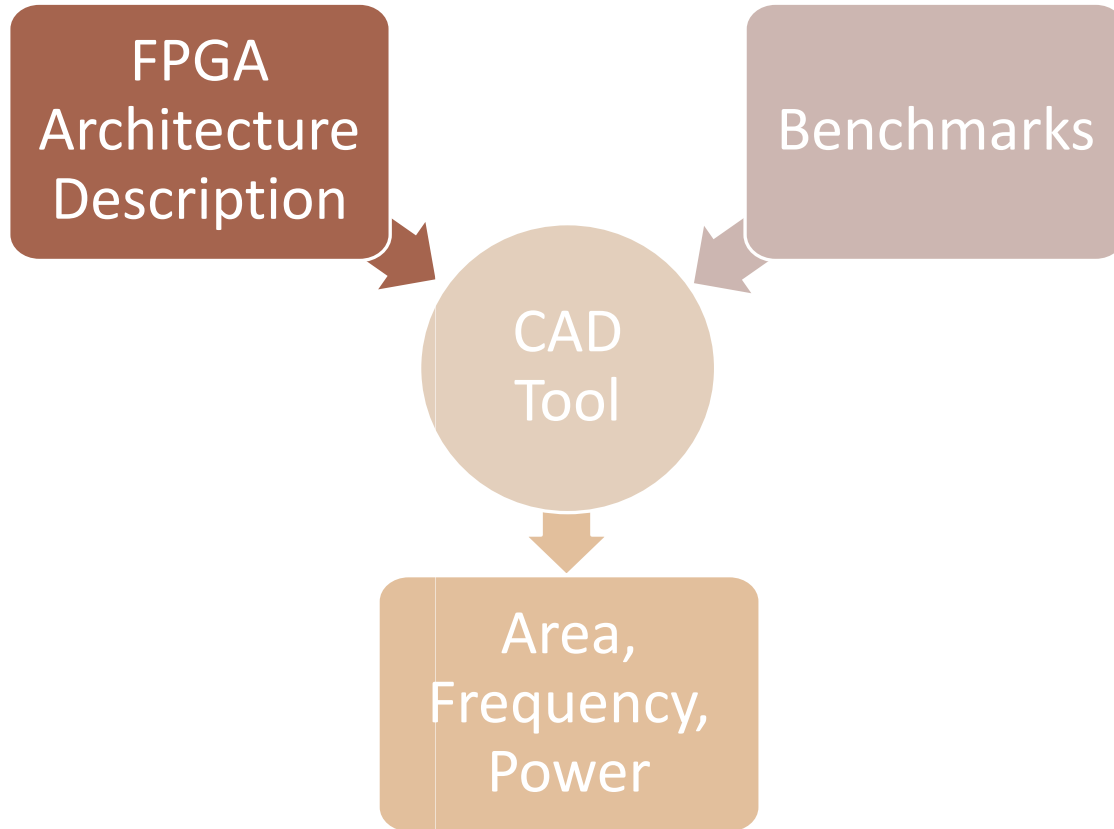
*June 18, 2023*

# FPGA Architecture and CAD Research

---



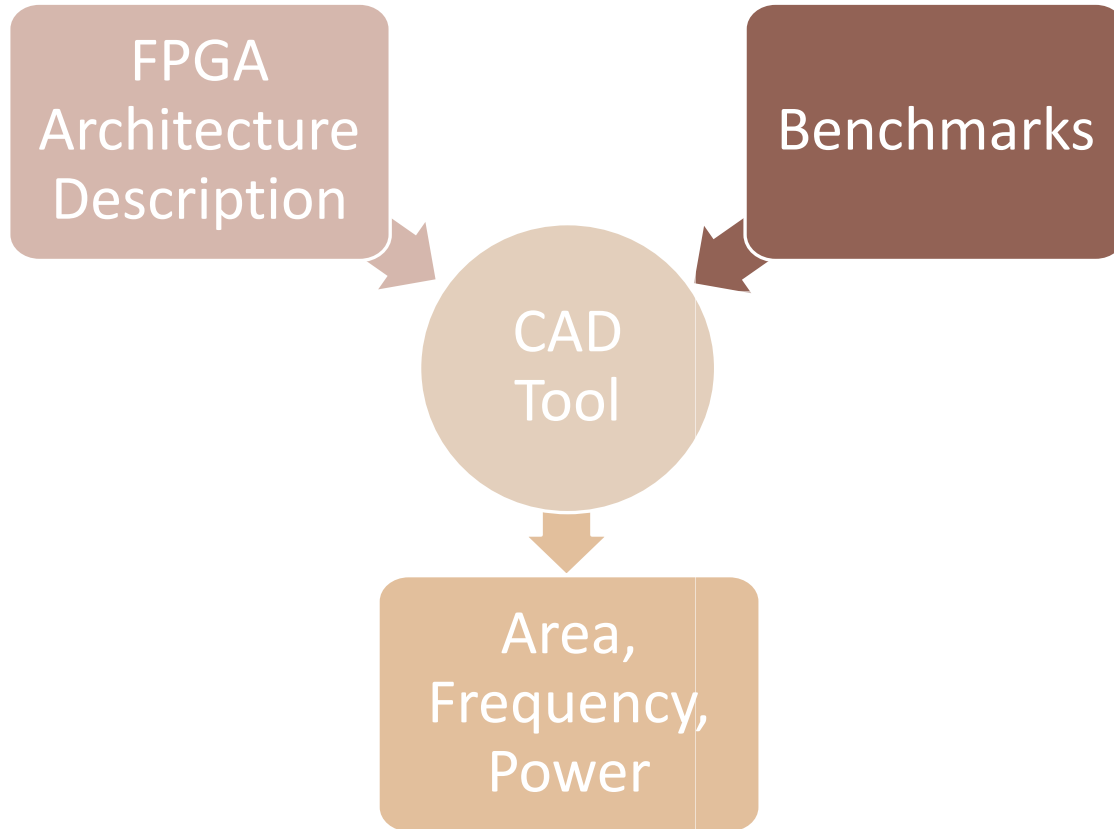
# FPGA Architecture and CAD Research



## DL-optimized FPGAs

SHADOW MULTIPLIER IN LOGIC BLOCKS  
SPECIALIZED OVERLAYS  
INTEL TENSOR BLOCK  
IN-BRAM COMPUTE  
ACHRONIX ML PROCESSOR BLOCKS  
TENSOR TILES  
LOW-PRECISION DSPS  
XILINX AI ENGINES  
REGISTER FILE IN DSPS  
FLEXLOGIC NNMAX TILES  
**TENSOR SLICES**

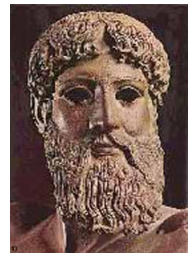
# FPGA Architecture and CAD Research



## Existing FPGA Benchmark Suites

Benchmark Suite	Medium-Large	Heterogenous	Open-source CAD	DL-specific
MCNC20	x	x	✓	x
UMass RCG	✓	-	x	x
Groundhog	-	✓	-	x
ERCBench	-	✓	x	x
VTR	x	✓	✓	x
Titan	✓	✓	x	x
Koios	✓	✓	✓	✓

# Koios – The Titan of Intelligence



A DL-specific benchmark suite for FPGA research

40 benchmarks that cover a diverse representative space

Includes some proxy/synthetic benchmarks

Contains original designs, and designs re-created from prior works

Suitable for DL-specific FPGA architecture exploration and CAD research

# Contributors



Name	Affiliation	Role
Aman Arora	University of Texas at Austin	Graduate Student (Ph.D.)
Andrew Boutros	University of Toronto	Graduate Student (Ph.D.)
Seyed Alireza Damghani	University of New Brunswick	Graduate Student (Master's)
Daniel Rauch	University of Texas at Austin	Graduate Student (Master's)
Aishwarya Rajen	University of Texas at Austin	Graduate Student (Master's)
Mohamed Elgammal	University of Toronto	Graduate Student (Ph.D.)
Karan Mathur	University of Texas at Austin/BITS Pilani	Research Intern (Bachelor's Student)
Vedant Mohanty	University of Texas at Austin/BITS Pilani Hyd	Research Intern (Bachelor's Student)
Tanmay Anand	University of Texas at Austin/BITS Pilani	Research Intern (Bachelor's Student)
Samidh Mehta	University of Texas at Austin/BITS Pilani Goa	Research Intern (Bachelor's Student)
Pragnesh Patel	University of Texas at Austin/BITS Pilani Goa	Research Intern (Bachelor's Student)

# Open Sourced with VTR

Currently at version 2.0

verilog-to-routing / vtr-verilog-to-routing Public

Edit Pins Watch 69 Fork 308 Starred 744

Code Issues 322 Pull requests 41 Discussions Actions Security Insights

master vtr-verilog-to-routing / vtr\_flow / benchmarks / verilog / koios /

Go to file Add file ...

aman26kbm Adding li

README.md

attention\_layer.v

bnn.v

bwave\_like.fixed.large.v

bwave\_like.fixed.small.v

bwave\_like.float.large.v

bwave\_like.float.small.v

clstm\_like.large.v

## README.md

### Koios 2.0 Benchmarks

#### Introduction

Koios benchmarks are a set of Deep Learning (DL) benchmarks for FPGA architecture and CAD research. They are suitable for DL related architecture and CAD research. There are 40 designs that include several medium-sized benchmarks and some large benchmarks. The designs target different network types (CNNs, RNNs, MLPs, RL) and layer types (fully-connected, convolution, activation, softmax, reduction, eltwise). Some of the designs are generated from HLS tools as well. These designs use many precisions including binary, different fixed point types int8/16/32, brain floating point (bfloat16), and IEEE half-precision floating point (fp16).

#### Documentation

A brief documentation of Koios benchmarks is available [here](#).

#### How to Use

Koios benchmarks are fully compatible with the full VTR flow. They can be used using the standard VTR flow described [here](#).

Koios benchmarks use advanced DSP features that are available in only a few FPGA architectures provided with VTR. These benchmarks



<https://tinyurl.com/vtrkoios>

# Agenda



Introduction



Koios



Results



Conclusion



# The Koios Benchmark Suite

---

Benchmark	Description
dla_like (S/M/L)	Intel-DLA-like accelerator
clstm_like (S/M/L)	CLSTM-like accelerator
deepfreeze	ARM FixyNN design
tdarknet_like (S/L)	Accelerator for Tiny Darknet
brainwave_like	Microsoft-Brainwave-like design
lstm	LSTM engine
bnn	4-layer binary neural network
lenet	Accelerator for LeNet-5
tpu_like.ws (S/L)	Google-TPU-v1-like accelerator
tpu_like.os (S/L)	Google-TPU-v1-like accelerator

Benchmark	Description
dnnweaver	DNNWeaver accelerator
gemm_layer	Matrix multiplication engine
attention_layer	Transformer self-attention layer
conv_layer	GEMM based convolution
robot_rl	Robot+maze application
reduction_layer	Add/max/min reduction tree
spmv	Sparse matrix vector multiplication
eltwise_layer	Matrix elementwise add/sub/mult
softmax	Softmax classification layer
conv_layer_hls	Sliding window convolution

... and 8 proxy benchmarks (more on this later)

# The Koios Benchmark Suite

---

Design Size

Implementation  
Style

Target Neural  
Network

Acceleration  
Paradigm

Numerical  
Precisions

Circuit Properties

# The Koios Benchmark Suite

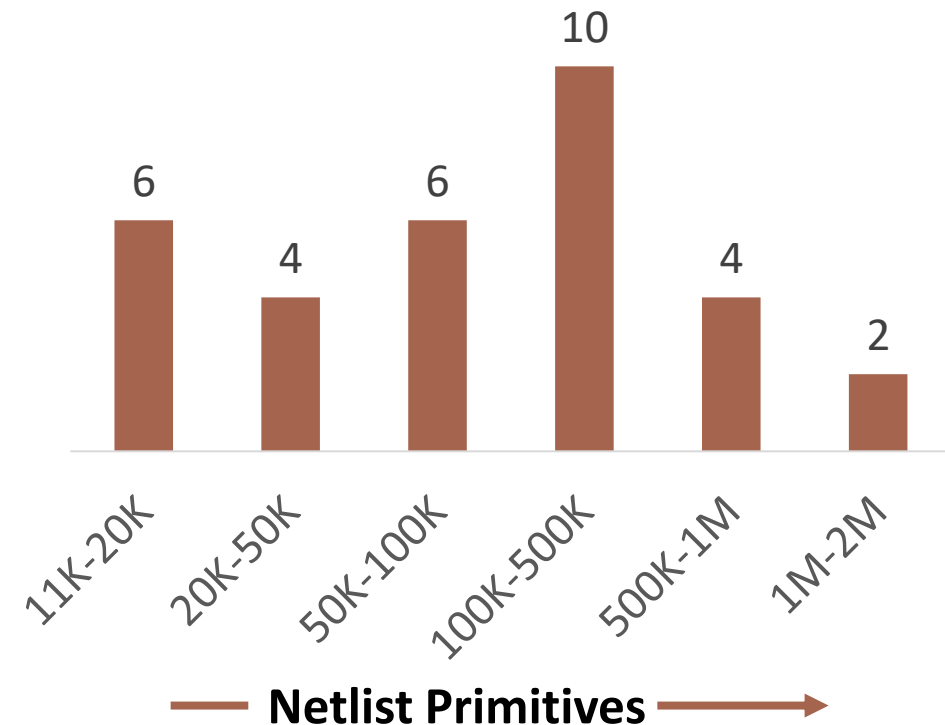
## Design Size

From ~11K to ~2M netlist primitives

Some have multiple variants (L/M/S)

Large → Challenging for CAD tools

Small → For early-stage experiments



# The Koios Benchmark Suite

---

## Implementation Style

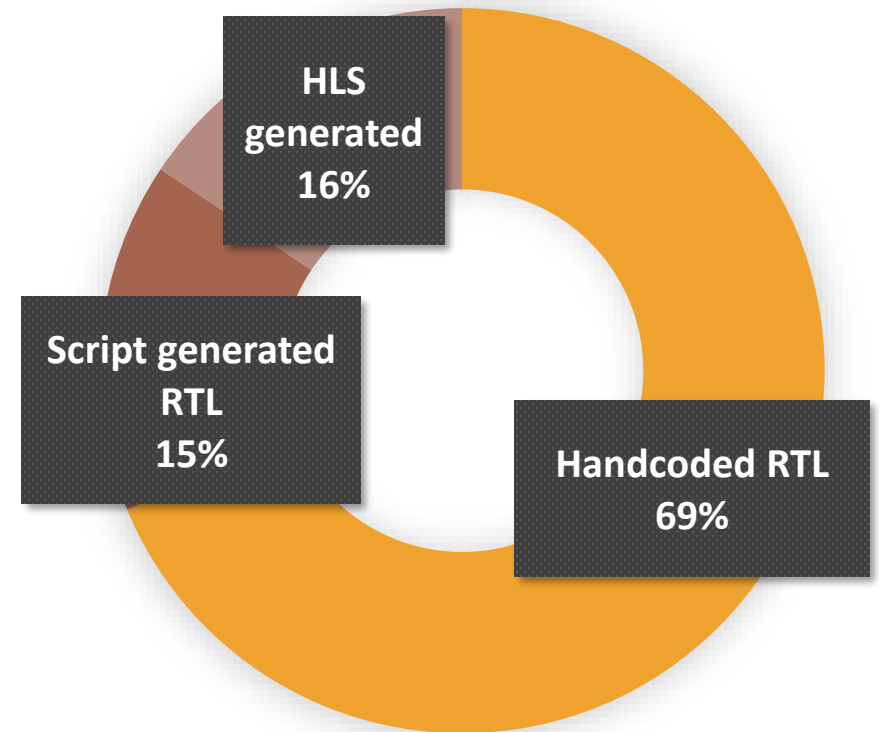
Benchmarks are in Verilog

Some implemented in RTL directly

Some implemented using script based  
RTL generators

Some generated using HLS tools

- Widely distributed control signals
- Complex state machines



# The Koios Benchmark Suite

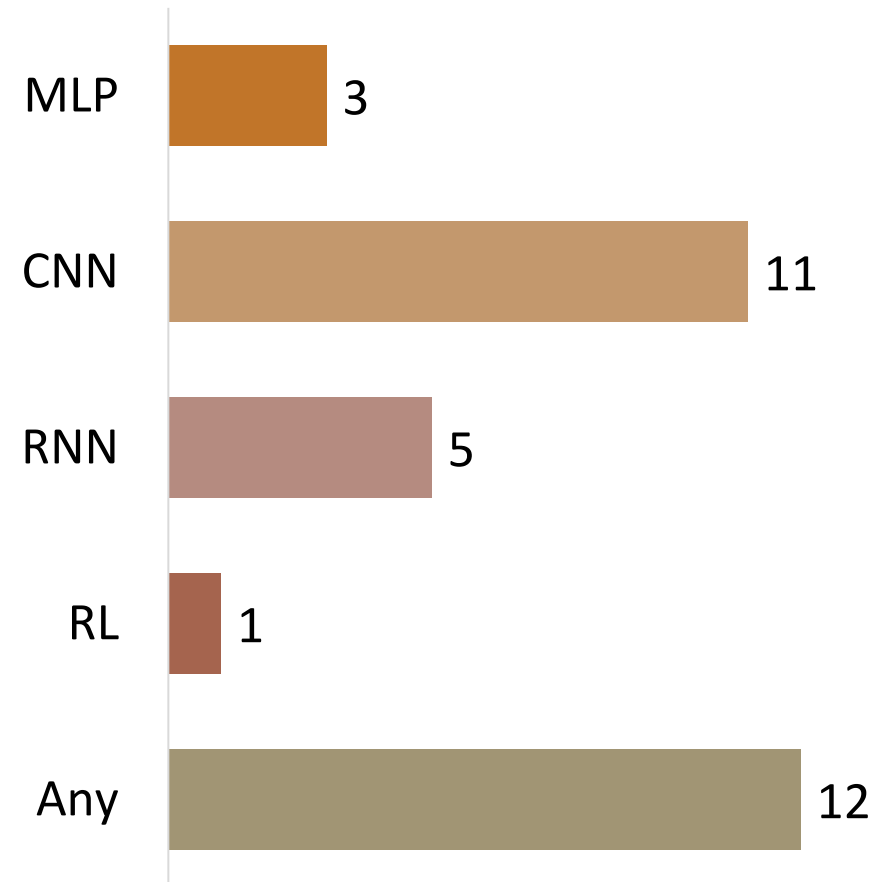
---

Major classes of neural networks

Different compute and memory requirements

- Reflects in resource breakdown

## Target Neural Network



# The Koios Benchmark Suite

---

# The Koios Benchmark Suite

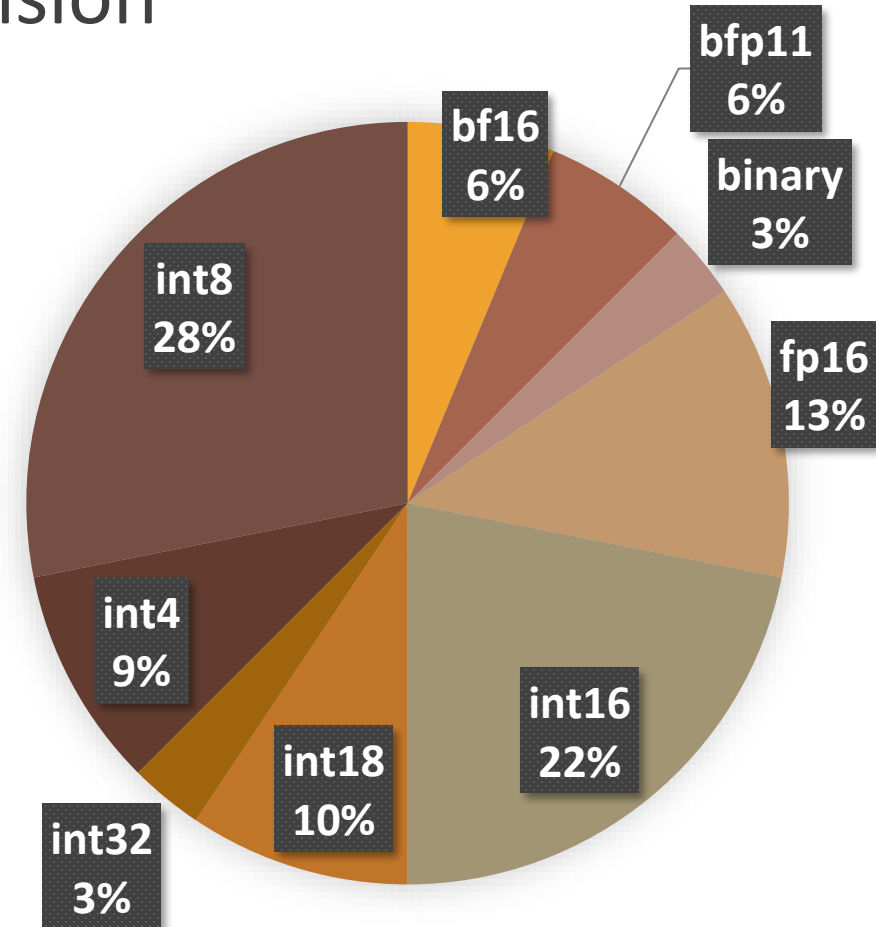
## Numerical Precision

Custom numerical precisions is key for DL

Common precisions used:

- Binary
- INT 4/8/16/18/32
- BFloat16, IEEE half-precision (FP16)
- Block floating point (BFP11)

Explore new DSP and BRAM architectures

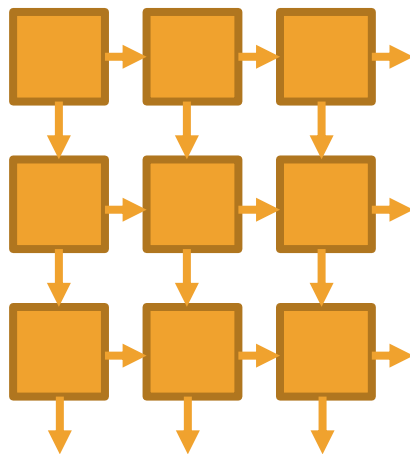


# The Koios Benchmark Suite

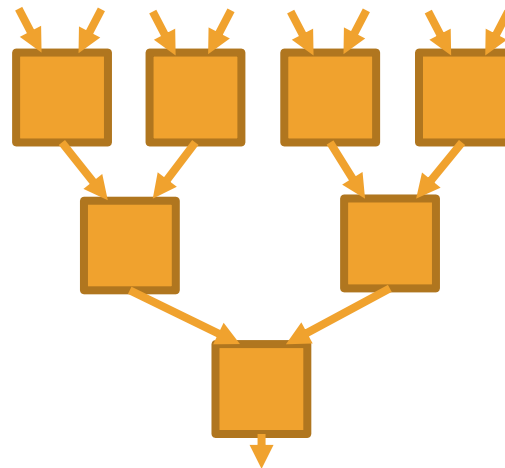
## Circuit Properties

Different circuit styles exercise CAD tools in different ways

Regular structures like systolic arrays



Large reduction trees

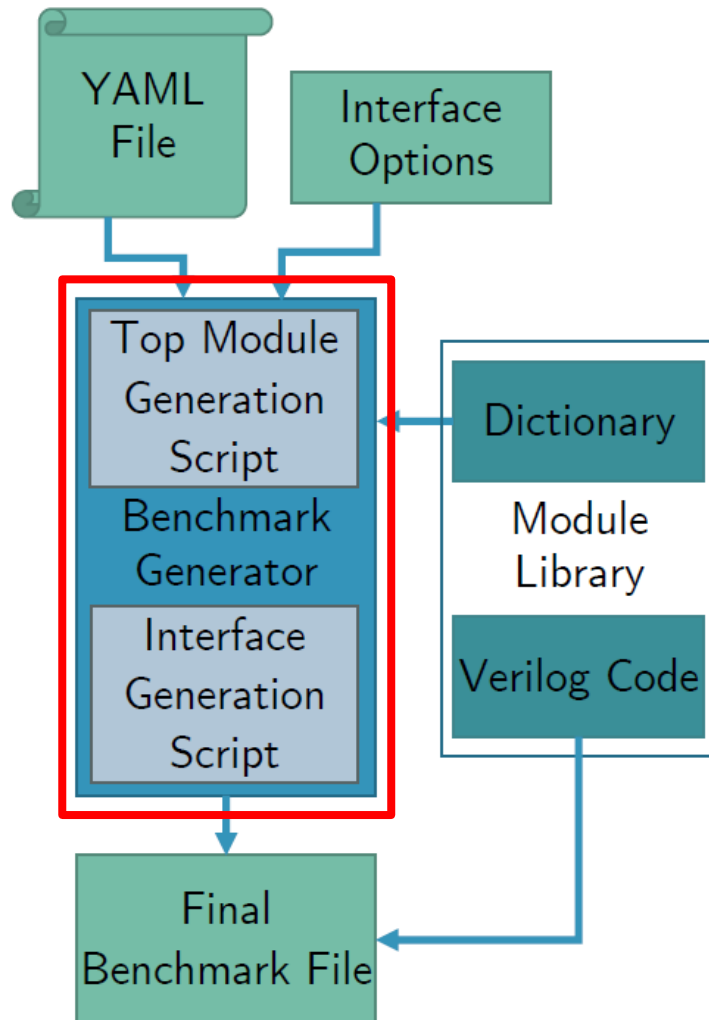


Long cascades of hard blocks





# Proxy Benchmarks in Koios



Heart of the generator are 2 scripts written in Python:

- One generates the top-level of the benchmark using component module descriptions
- Another generates code for interfaces between the component modules

# Agenda



Introduction



Koios

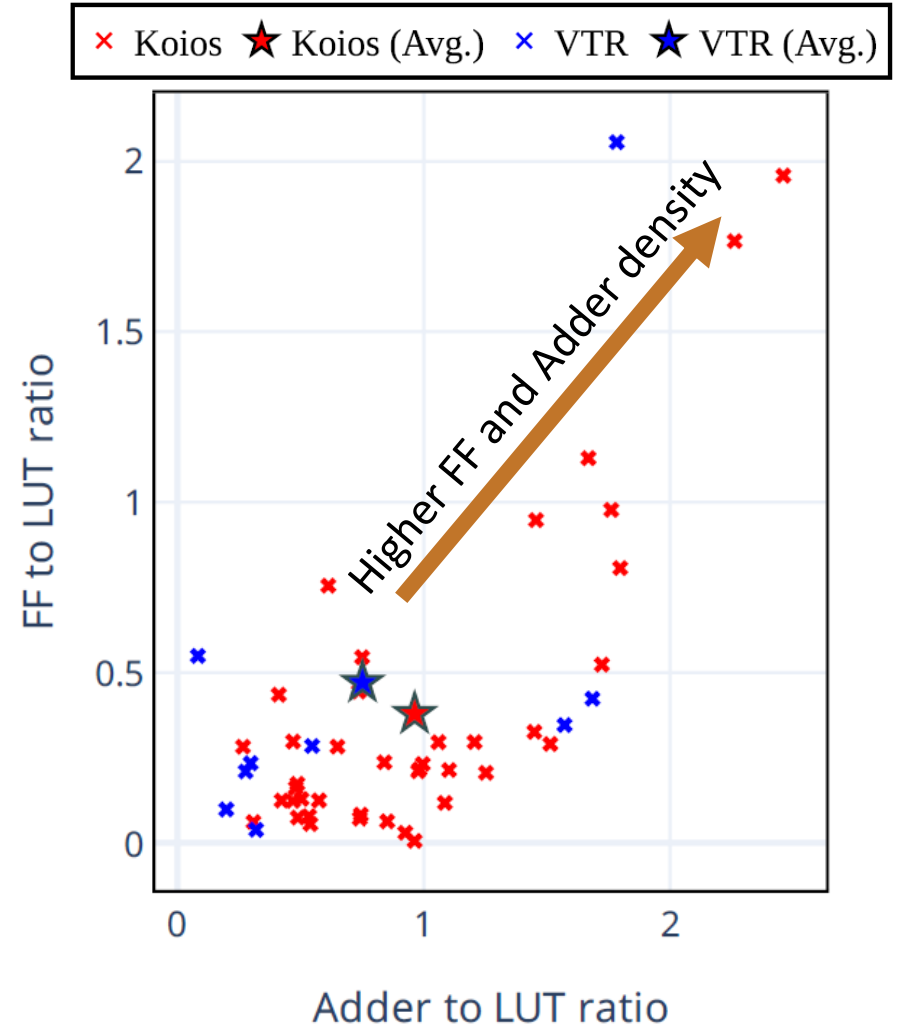
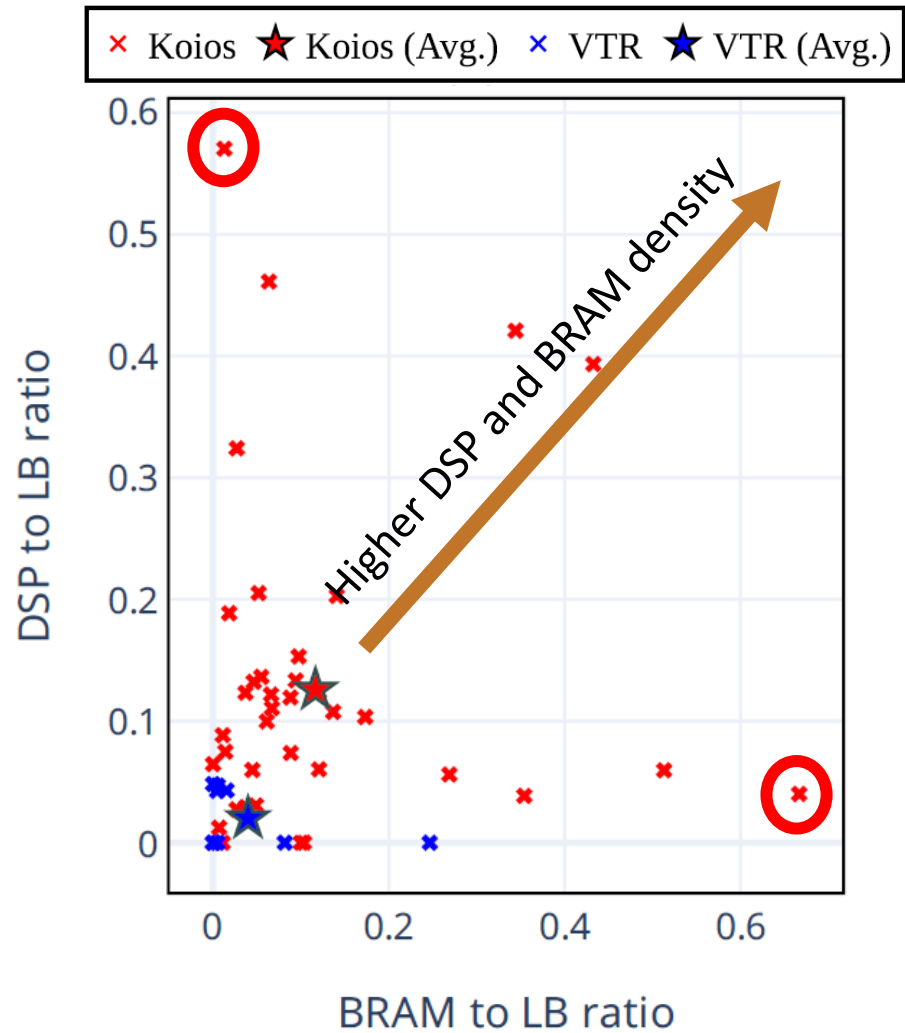


Results

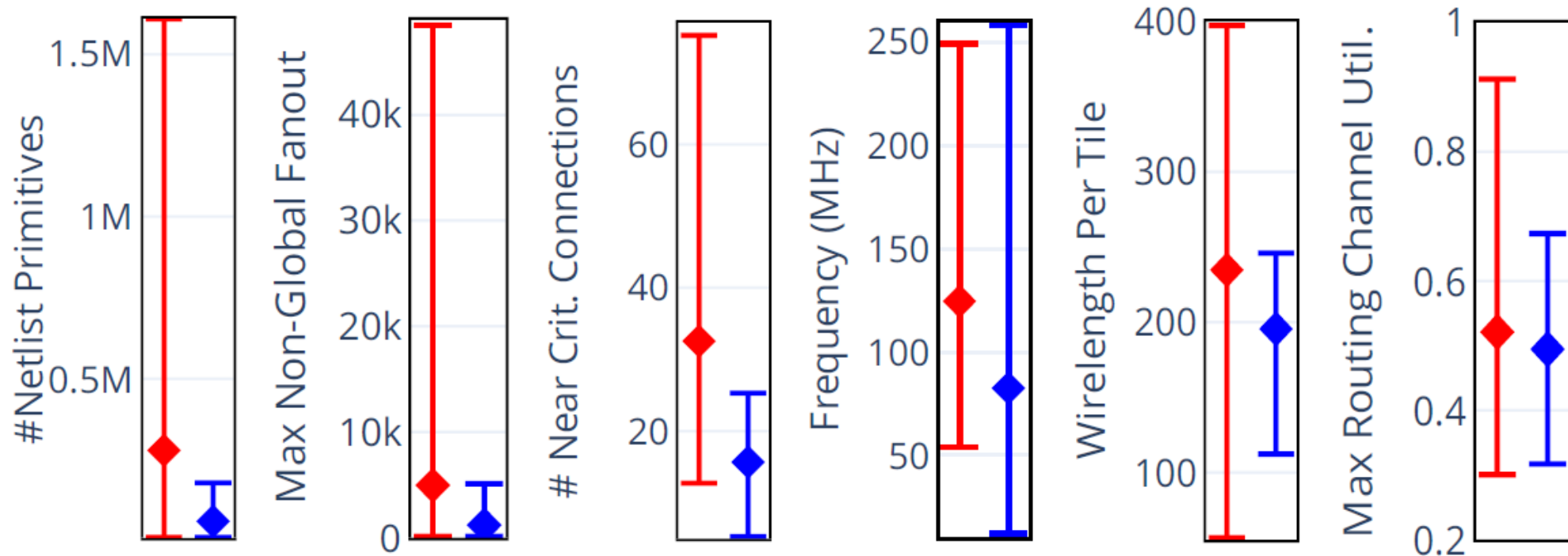


Conclusion

# Comparison with VTR Benchmarks

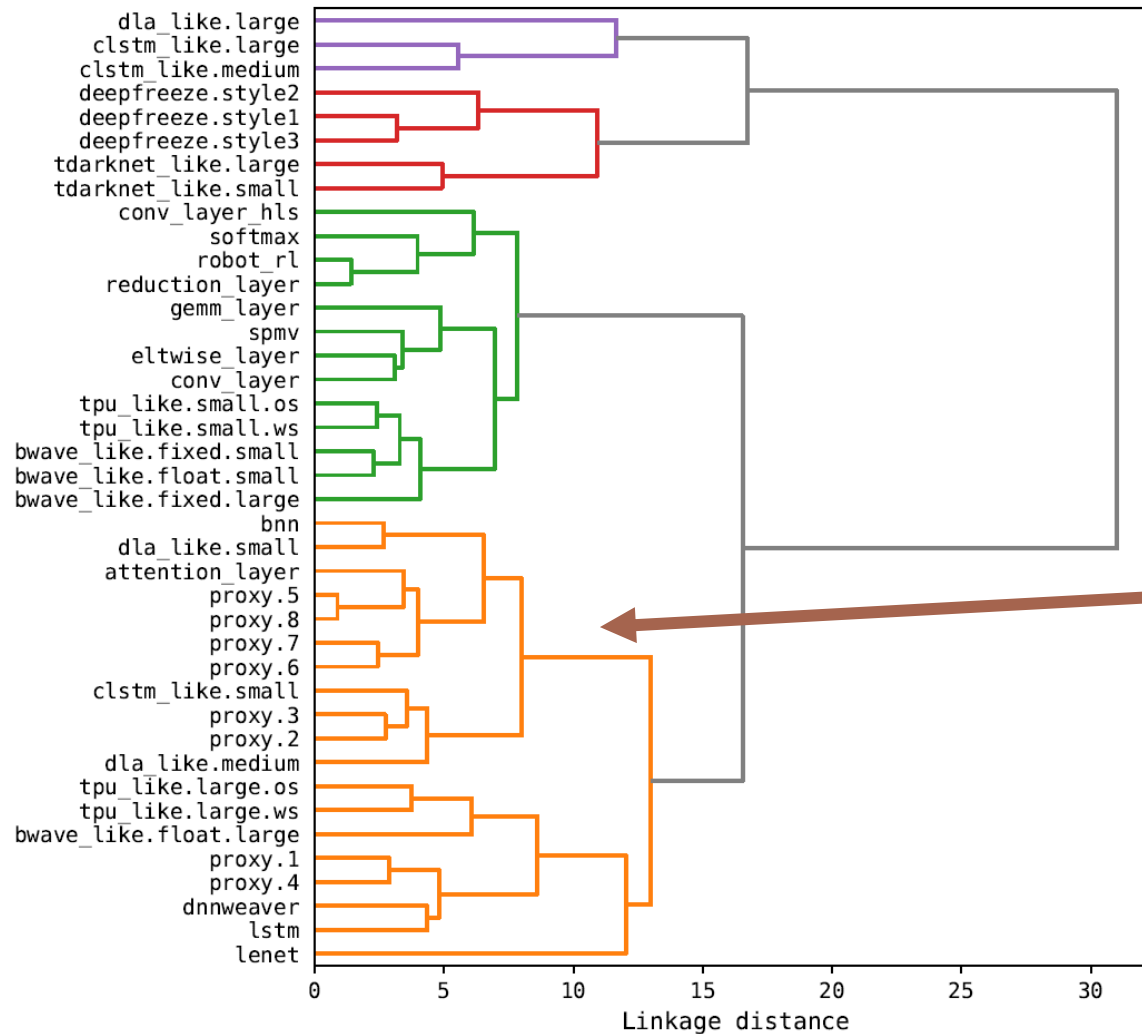


# Comparison with VTR Benchmarks



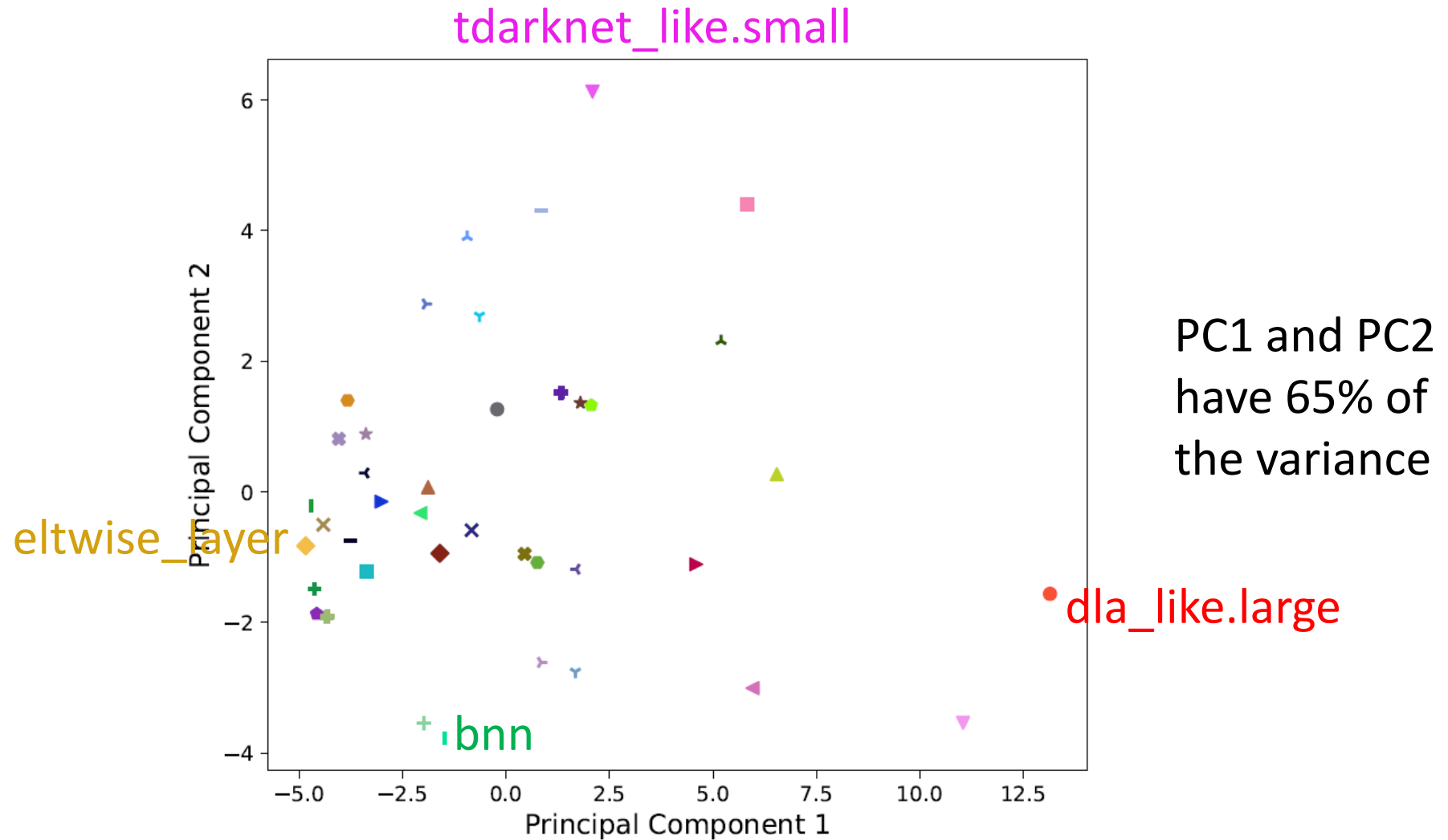
Red = Koios, Blue = VTR

# Dendrogram



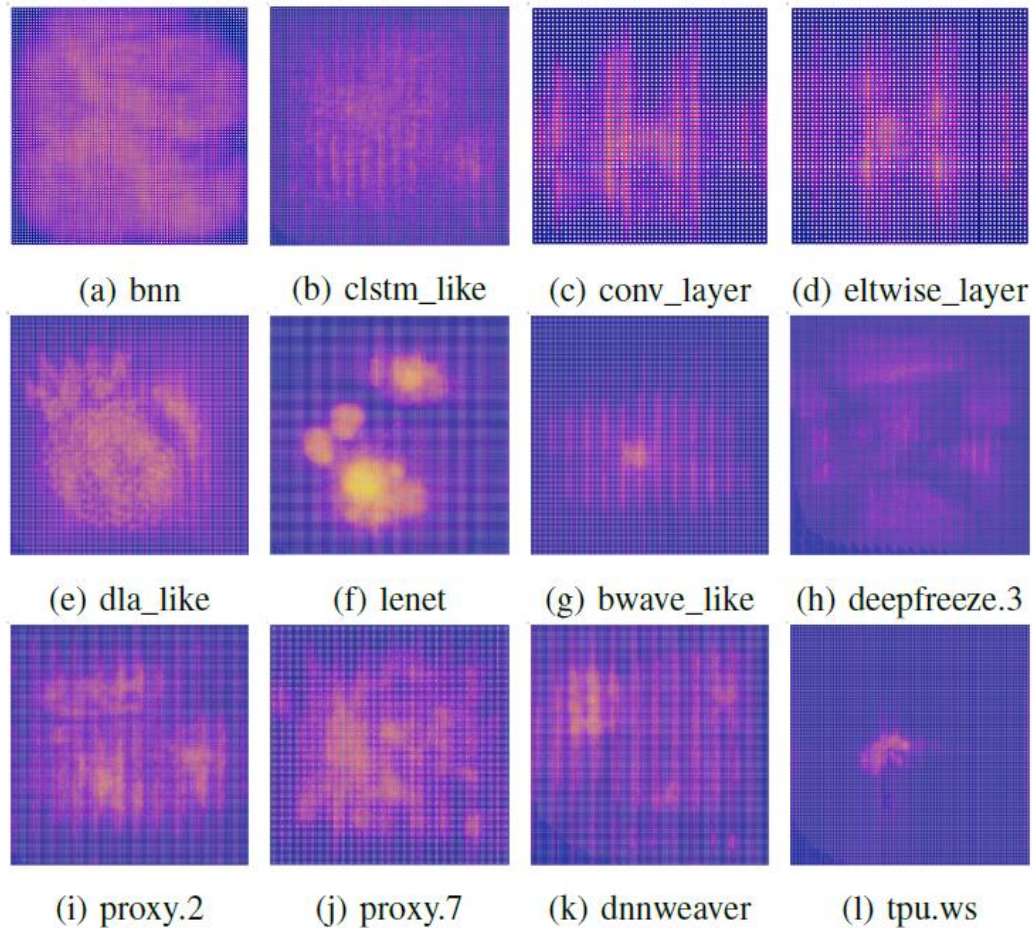
Future work:  
Generate more  
diverse proxy  
benchmarks

# PCA Analysis



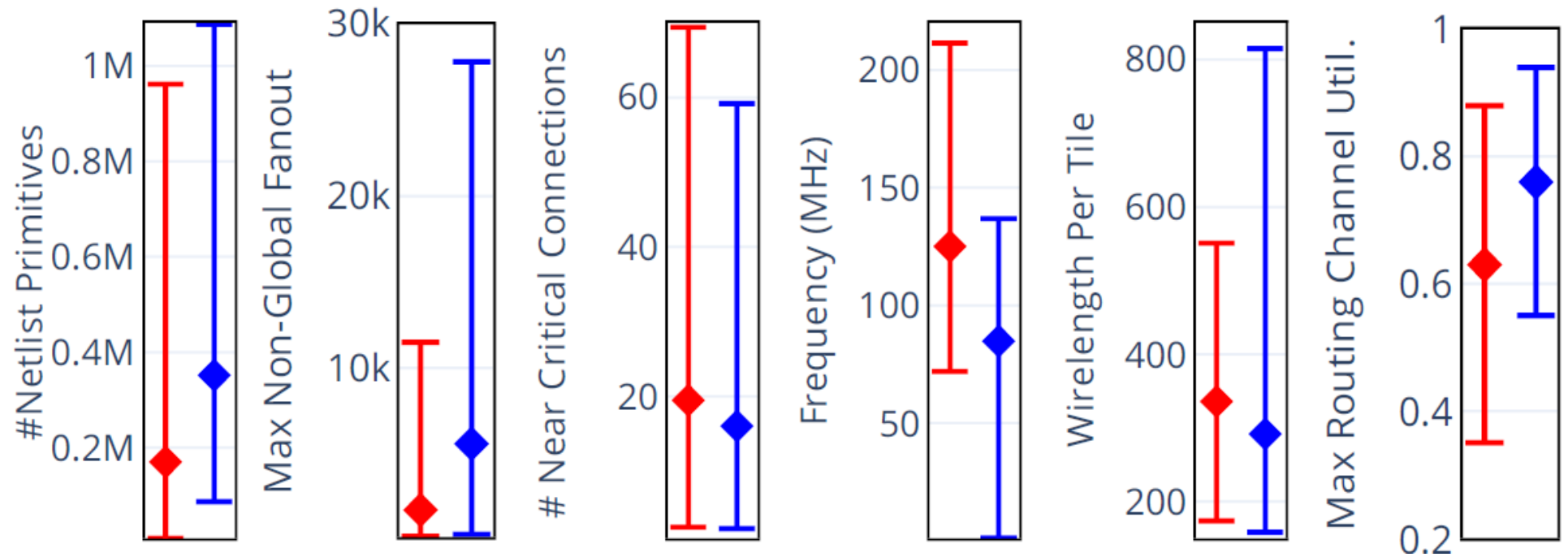
# Variation in routing utilization

---



- Lighter color means higher routing congestion
- Diversity in routing requirements and patterns in the benchmarks
- Exercise FPGA CAD tools (for placement and routing) in different ways

# Comparison with Titan Benchmarks

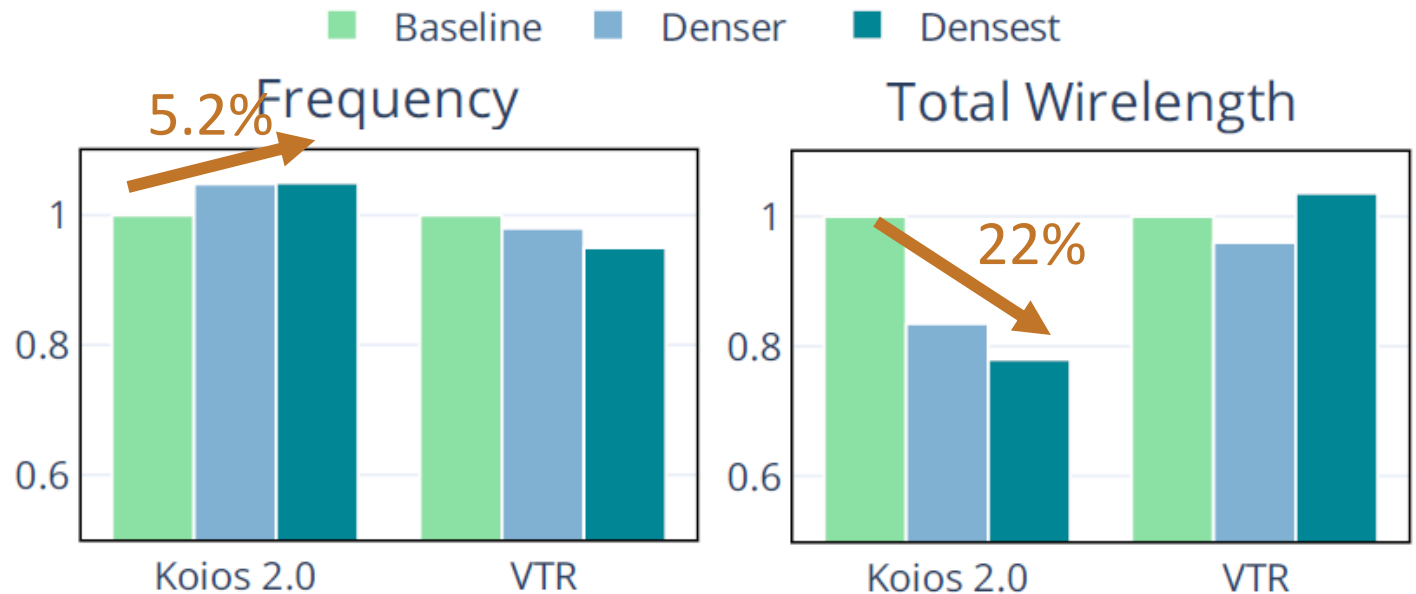
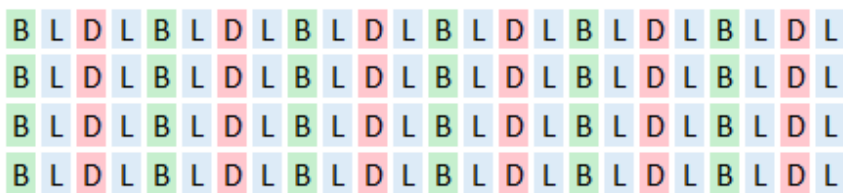
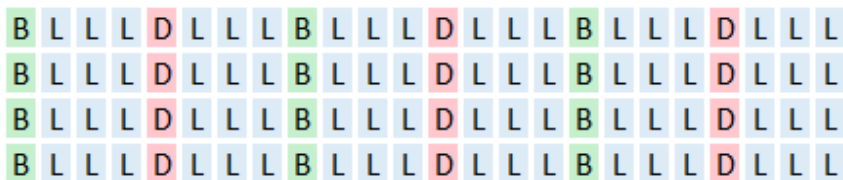


Red = Koios, Blue = Titan



# Arch Exploration Case Study

## Hard Block to Soft Logic Ratio



**Takeaway:**  
Right benchmarks, Right conclusions

# Agenda



Introduction



Koios



Results



Conclusion

# Conclusion

---



First DL benchmark suite for FPGA architecture and CAD research



Open-source and compatible with VTR (the most popular FPGA research framework)



Call for action: Use and contribute

# Find more at...

---

Koios: A Deep Learning Benchmark Suite for FPGA Architecture and CAD Research

- **IEEE International Conference on Field-Programmable Logic and Applications (FPL) 2021**

Koios 2.0: Open-Source Deep Learning Benchmarks for FPGA Architecture and CAD Research

- **IEEE Transactions on Computer Aided Design of Integrated Circuits & Systems (TCAD) 2023**



<https://tinyurl.com/vtrkoios>

Thanks