Intelligent Circuit Design and Implementation with Machine Learning in EDA

Yiran Chen

Dept. Electrical & Computer Engineering

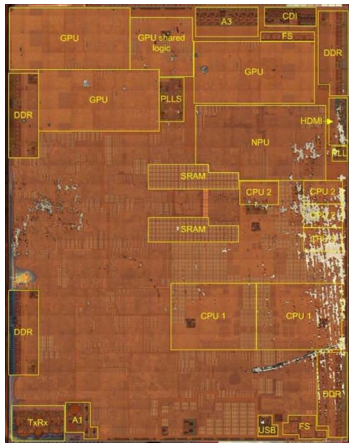
Duke University

Electronic Devices are Everywhere

*These images are found in public domain

Designers Try to Deliver Generational Gains

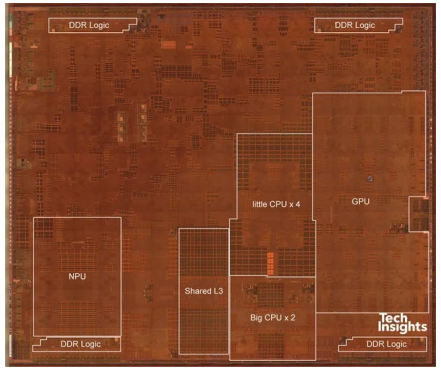
iPhone 8, X



Apple A11

10nm
4.3 B trasistors

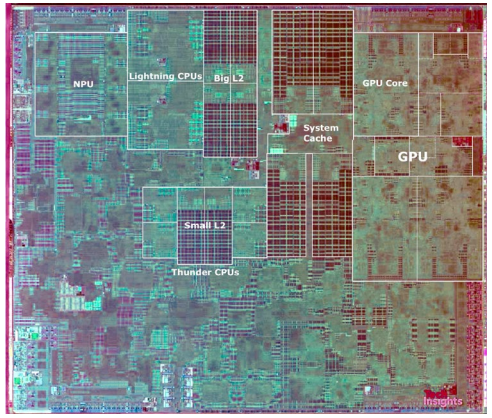
iPhone XS, XR



Apple A12

7nm
6.9 B trasistors

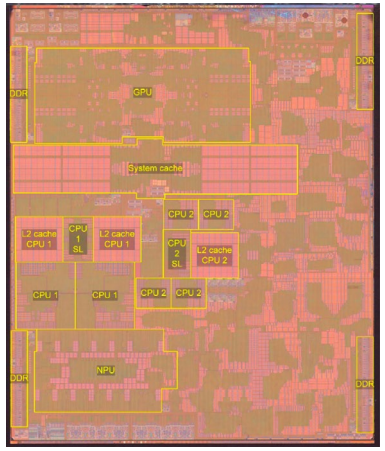
iPhone 11



Apple A13

7nm
8.5 B trasistors

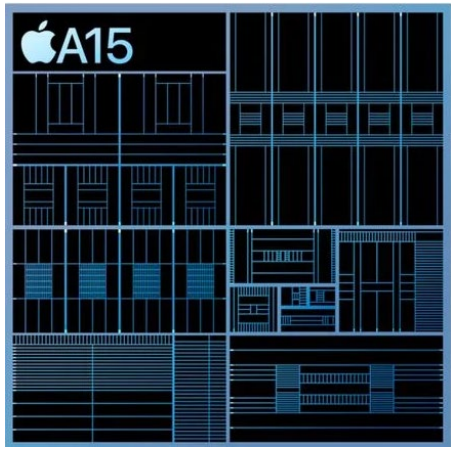
iPhone 12



Apple A14

5nm
11.8 B trasistors

iPhone 13



Apple A15

5nm
15 B trasistors

Benefit from increased **integration** and **architecture** improvements

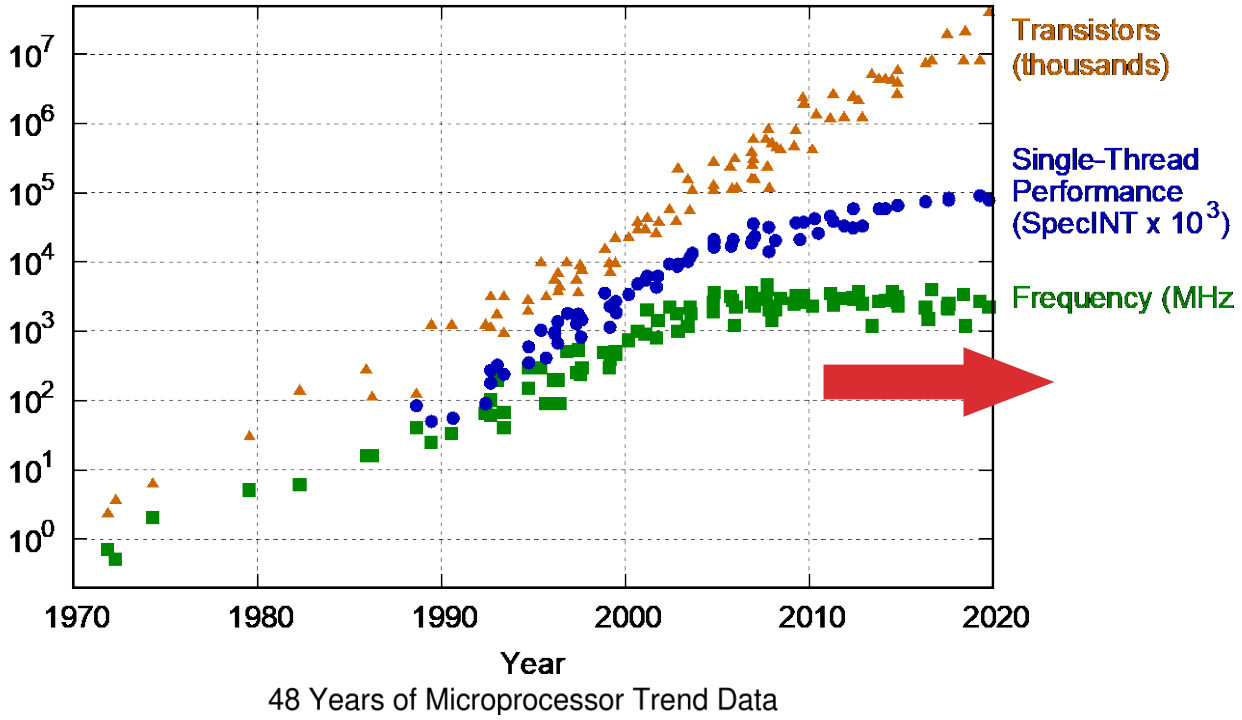
*Source: TechInsights Inc.

Duke UNIVERSITY

Chip Design Challenges

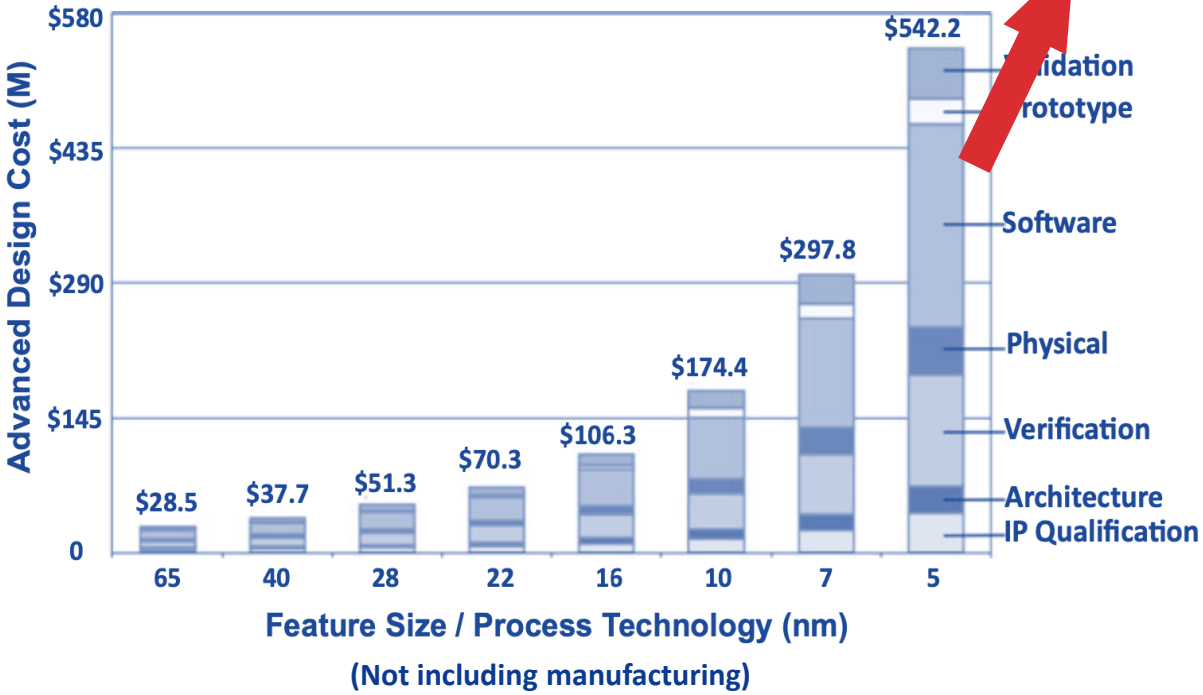
Diminishing performance gain and increasing design cost

Per-Core Performance Gain is Diminishing



Partially collected by M. Horowitz et al. Plotted by Karl Rupp, 2020

IBS Design Cost is Skyrocketing



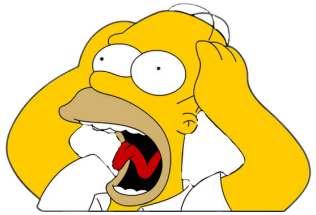
International Business Strategies, 2020

Duke UNIVERSITY

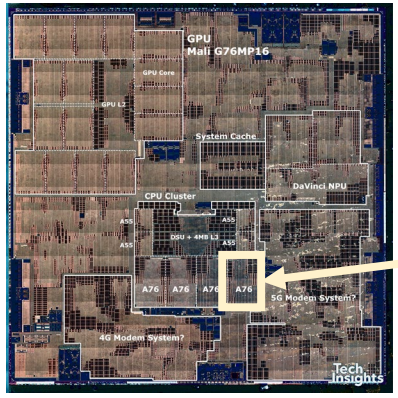
This is Real Problem!

Challenges at advanced node

- **Pressure** from IPC and frequency
- Peak power is **increasing**
- Power fluctuation is **more abrupt**
- Power delivery technique is **limited**
- **Increasing** design rules to meet
- **Increasing** wire parasitics, causing wire delay and noise
- ……



Inefficient chip design methodologies



For one Arm CPU core with ~3 million gates

- Accurate power simulation takes **~2 weeks**
- One iteration in physical design take **~1 week**
- Solutions **repeatedly** constructed from scratch
- Solutions rely on designer **intuition**
- ……

Source: The Kirin 990 SoC. TechInsights Inc.

Our Work: Intelligent Circuit Design & Implementation

PPA

Power

Power & Power Delivery Challenges

Xie et al. [ICCAD'20], Xie et al. [ASPDAC'20],

Xie et al. [MICRO'21] (**Best Paper Nomination**)

Performance

Timing & Interconnect Challenges

Barboza et al. [DAC'19], Liang et al. [ICCAD'20],

Xie et al. [ASPDAC'21], Xie et al. [TCAD] (in review)

Area

Routability Challenges

Xie et al. [ICCAD'18], Huang et al. [DATE'18],

Chang et al. [ICCAD'21]

Overall Flow Tuning

Xie et al. [ASPDAC'20]

Covered in this talk

Case Study 1:

Power & Power Delivery Challenges

Problem 1 – Design-time CPU Power Introspection

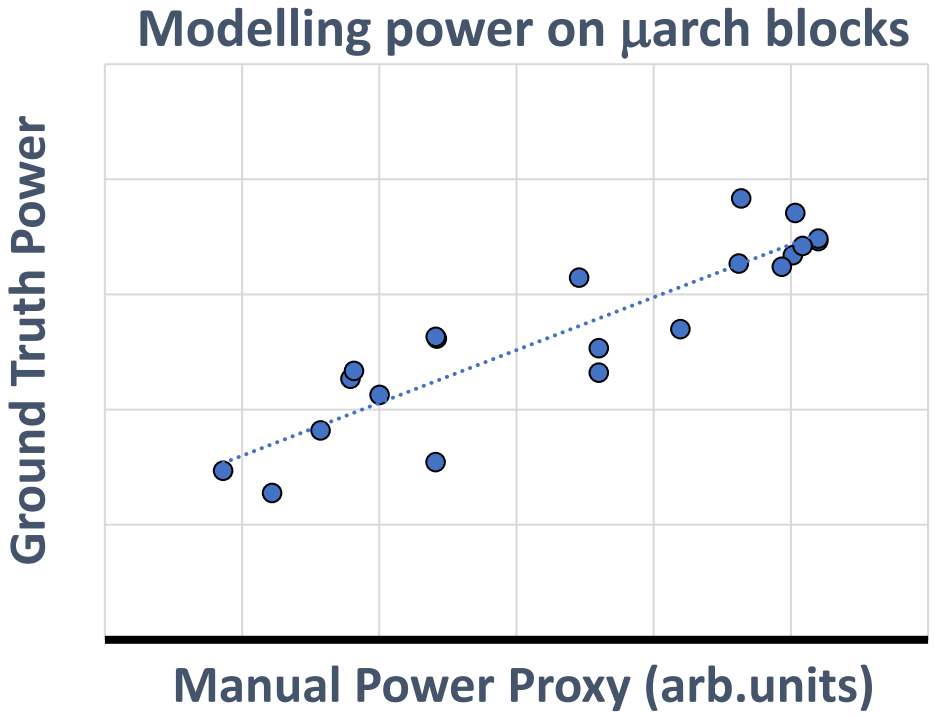
- **Delivering generational gains in IPC and F_{MAX} adversely impacts CPU power**
  - Diminishing returns with speculation, wide-issue and vectored execution

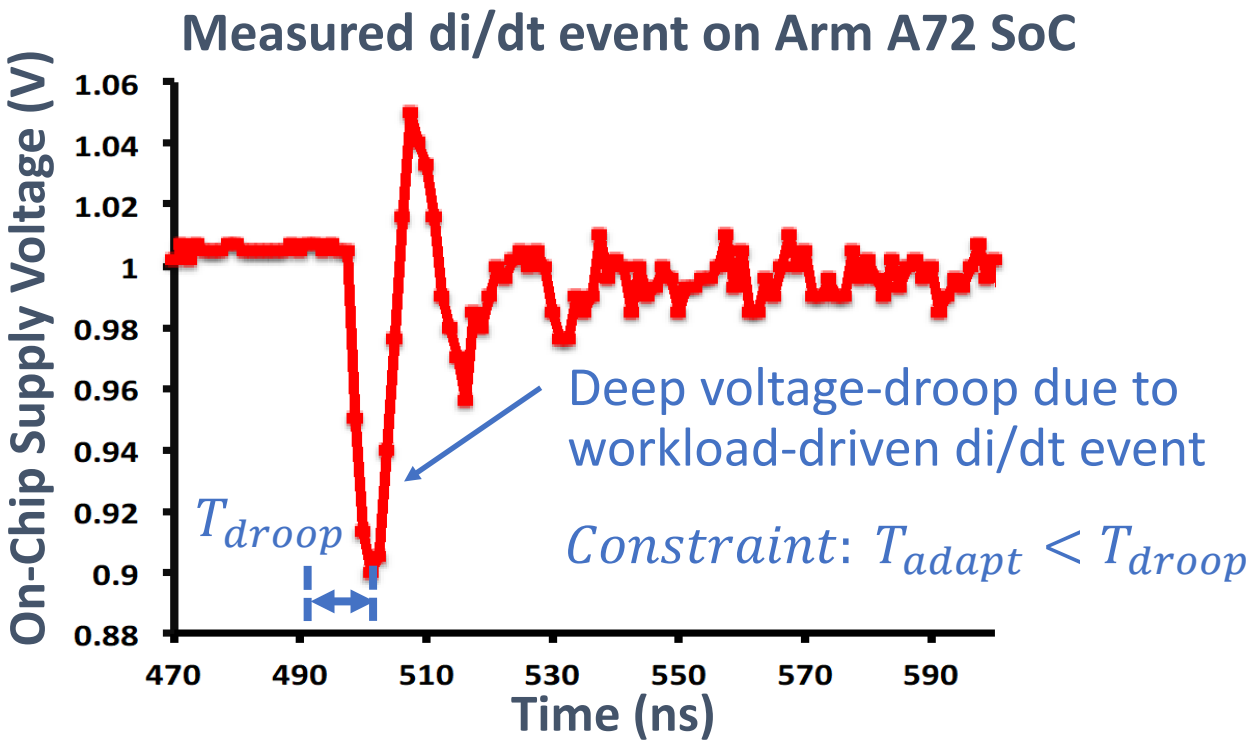
- **Power consumption is adversely impacted and trends upwards**
  - Efficiency gains through Moore's law scaling has effectively stalled
  - Parallel execution and greater transistor integration => increased switching activities

- **Power-delivery resources not keeping pace with CPU power demands**
  - Resistive interconnects at scaled technology nodes => greater sensitivity to peak-power
  - Packaging technology unable to sustain di/dt demands

- **Increasing power-sensitivity drives the need for design-time introspection**

Problem 2 – Run-time Introspection



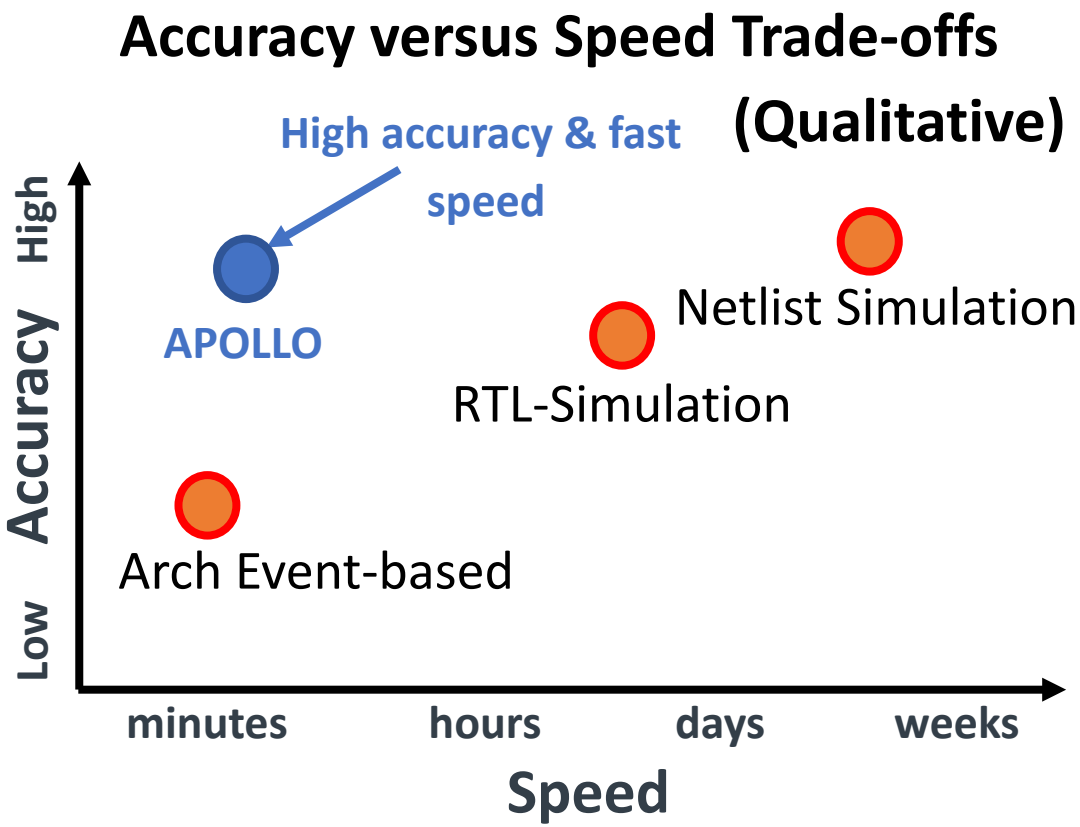


- **Peak-Power mitigation requires accurate power-estimation to drive throttling decisions**
  - Manually inferring proxies is difficult, particularly in modern CPUs with complex underlying µarch
- **Micro-architectural interactions (branch-mispredicts, ROB issue, hit-after-miss) trigger abrupt changes in CPU current-demand leading voltage-droop due to di/dt events**

Duke UNIVERSITY

APOLLO – Key Objectives and Attributes

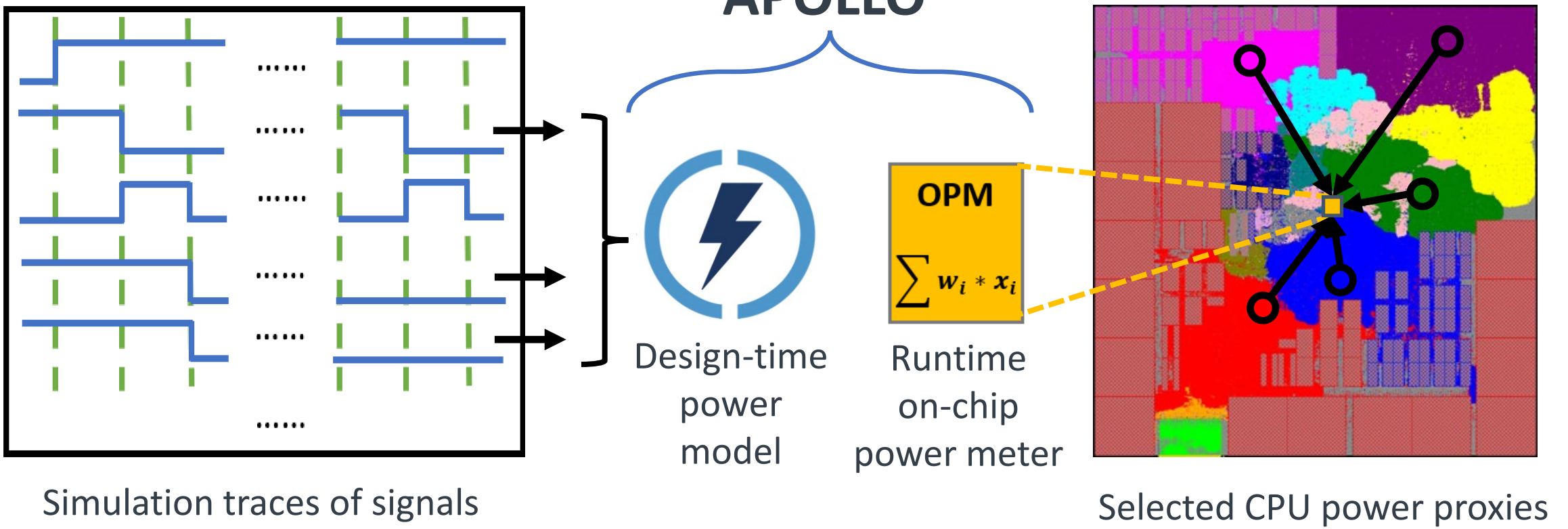
Problem: Prior art suffers from stark trade-offs between accuracy and speed



- **Automated Power-Proxy Extraction**
  - Use ML techniques to identify correlated events
- **Fast, yet accurate on-chip metering**
  - Proven on commercial CPUs with >95% accuracy
  - 0.2% area overhead over Neoverse N1 core
- **Per-cycle temporal resolution**
  - Unify simulation, Ldi/dt mitigation, emulator-tracing within the same framework
- **Extensible to higher abstraction simulation**
  - Trade-off accuracy for pre-identified events

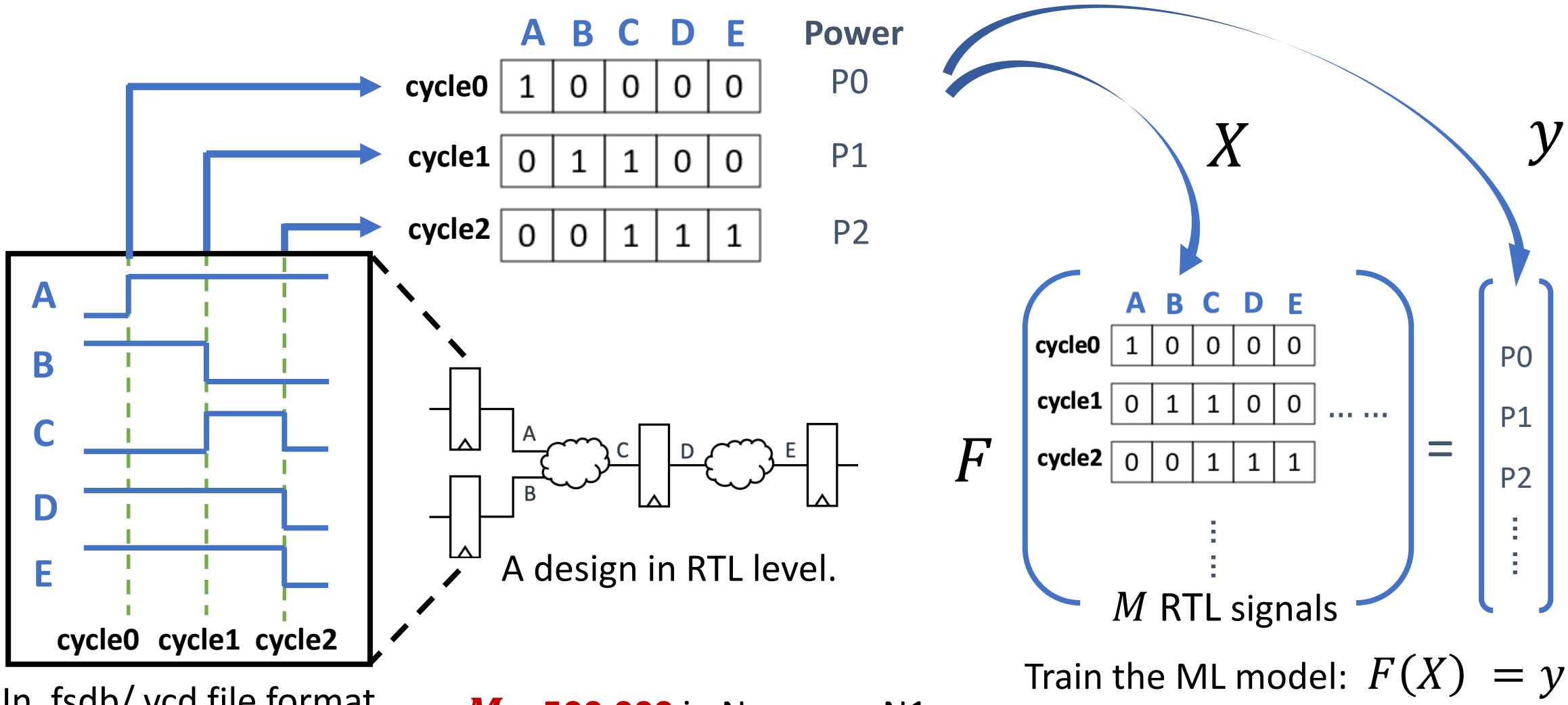
Duke UNIVERSITY

APOLLO Includes Design-time Model and Runtime OPM



Duke UNIVERSITY

APOLLO Feature Generation & Model Training



M > 500,000 in Neoverse N1

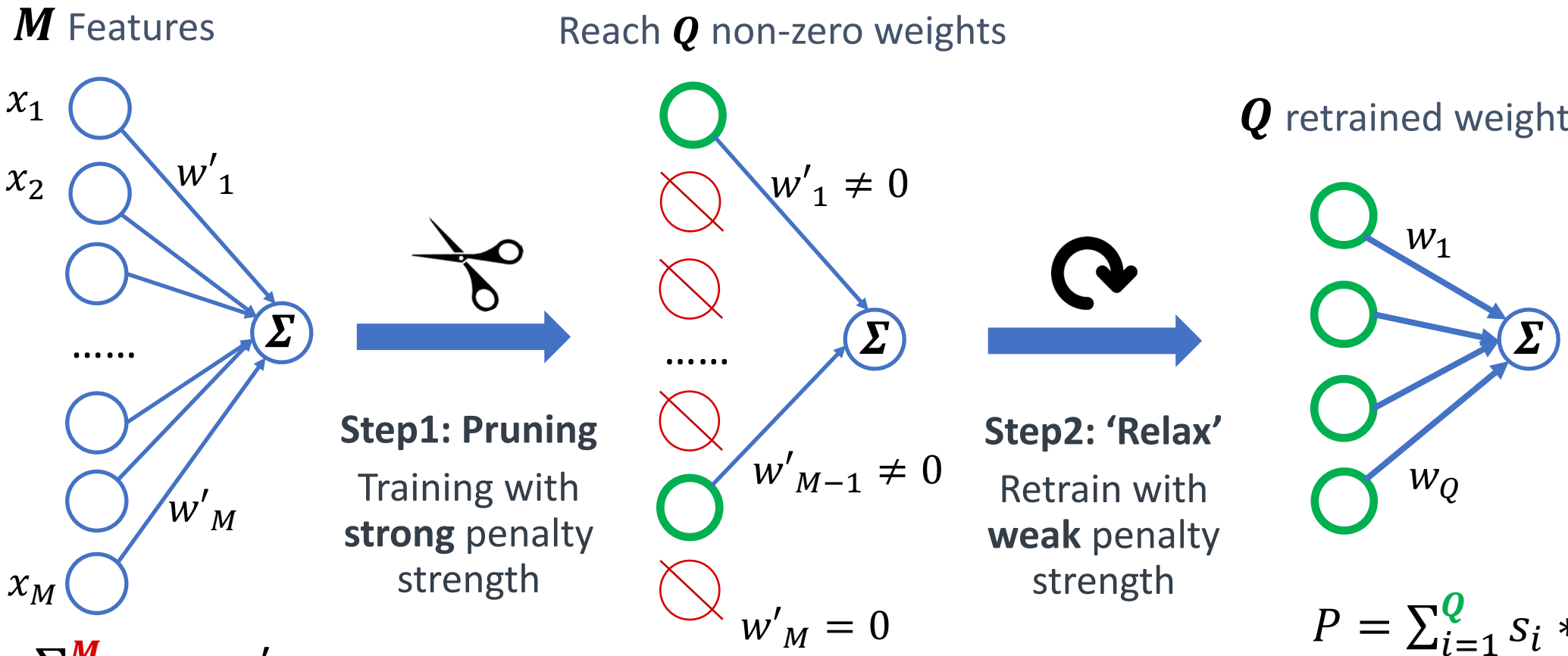
M > 1,000,000 in Cortex-A77

Duke UNIVERSITY

ML-Based Power Proxies Selection

Model construction in two steps

Please check our **paper** [MICRO'21] for detailed discussion on MCP method



Minimax concave penalty (**MCP**) for pruning

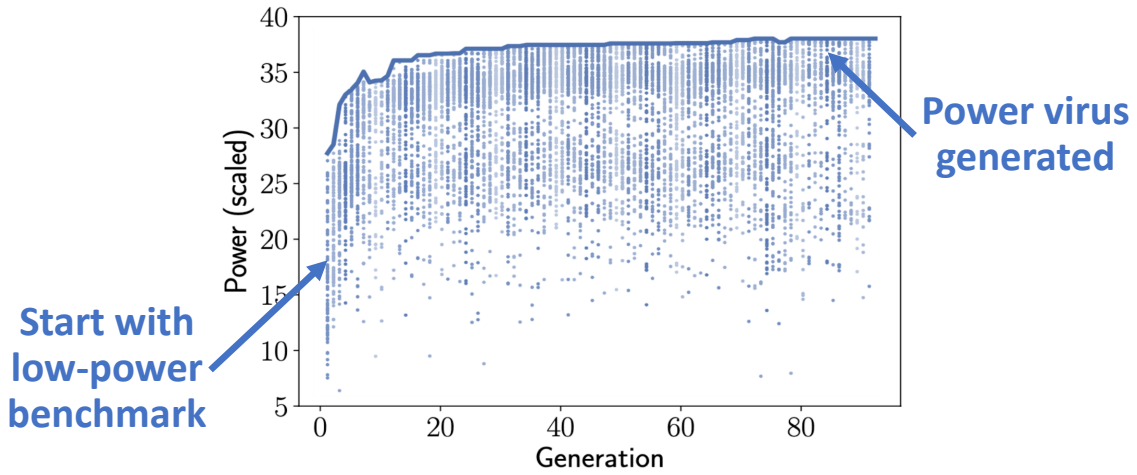
Duke UNIVERSITY

Our Proposed Power Modeling Approach

A “diverse” set of random (micro-)benchmarks is critical

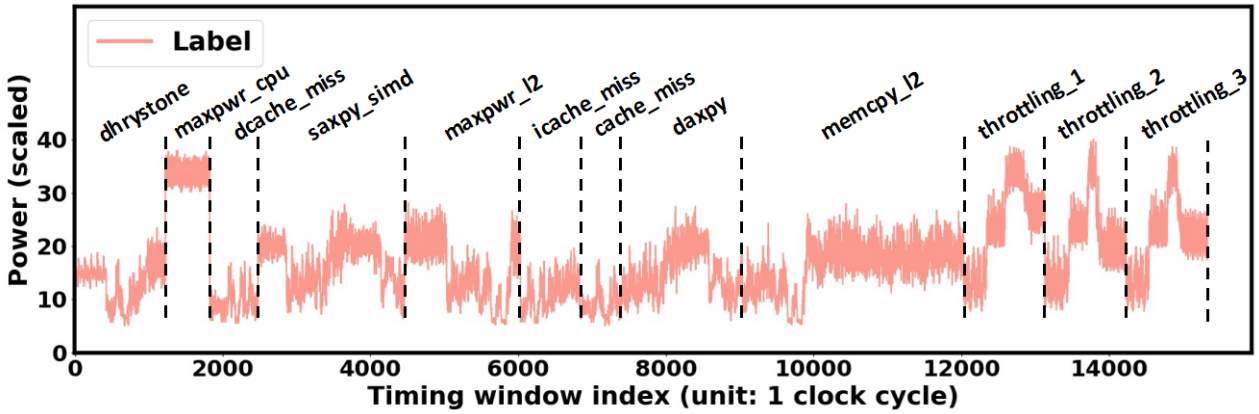
Training data automatically generated

- Micro-architecture agnostic **genetic algorithm** to automatically generate max-power virus
- A “diverse” set is generated: lower-power in early generations and higher-power in later generations



Model testing

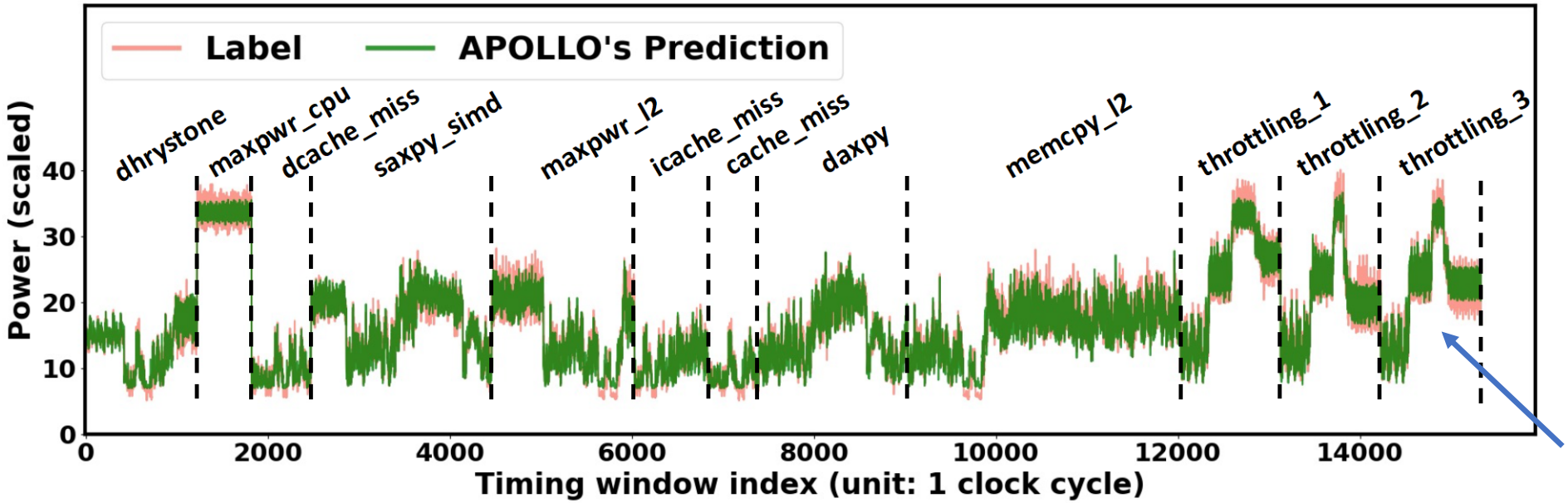
- Experiments on 3GHz 7nm microprocessors **Neoverse N1** and **Cortex A77**
- Testing on Arm power-indicative workloads
  - Steady-state, transient, and throttling regions
  - High- and low-power-consumption regions



Duke UNIVERSITY

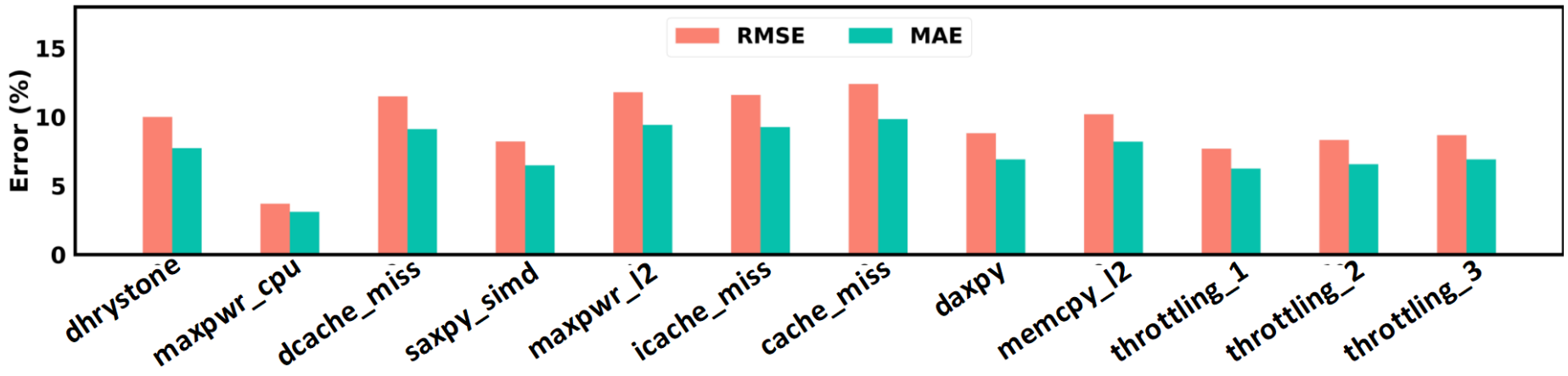
Prediction Accuracy as Design-Time Power Model

Per-cycle prediction from APOLLO with $\boldsymbol{Q}$=159 proxies



- MAE = 7.19%
- RMSE = 9.13%
- R^2 = 0.953

Prediction trace shows great agreement with ground-truth

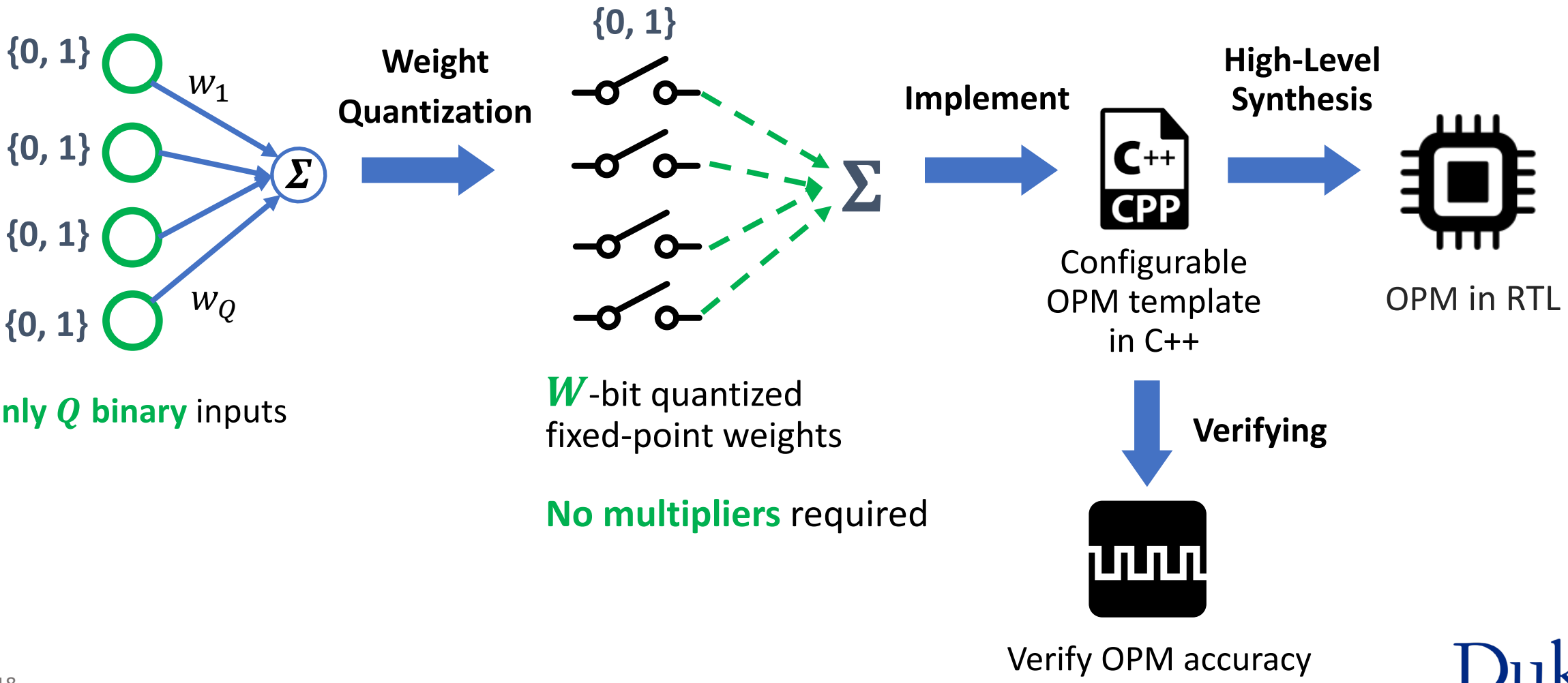


- MAE < 10% for all workloads

Duke UNIVERSITY

Automated Low-Cost Runtime OPM Implementation

APOLLO is designed to be hardware-friendly



Only Q binary inputs

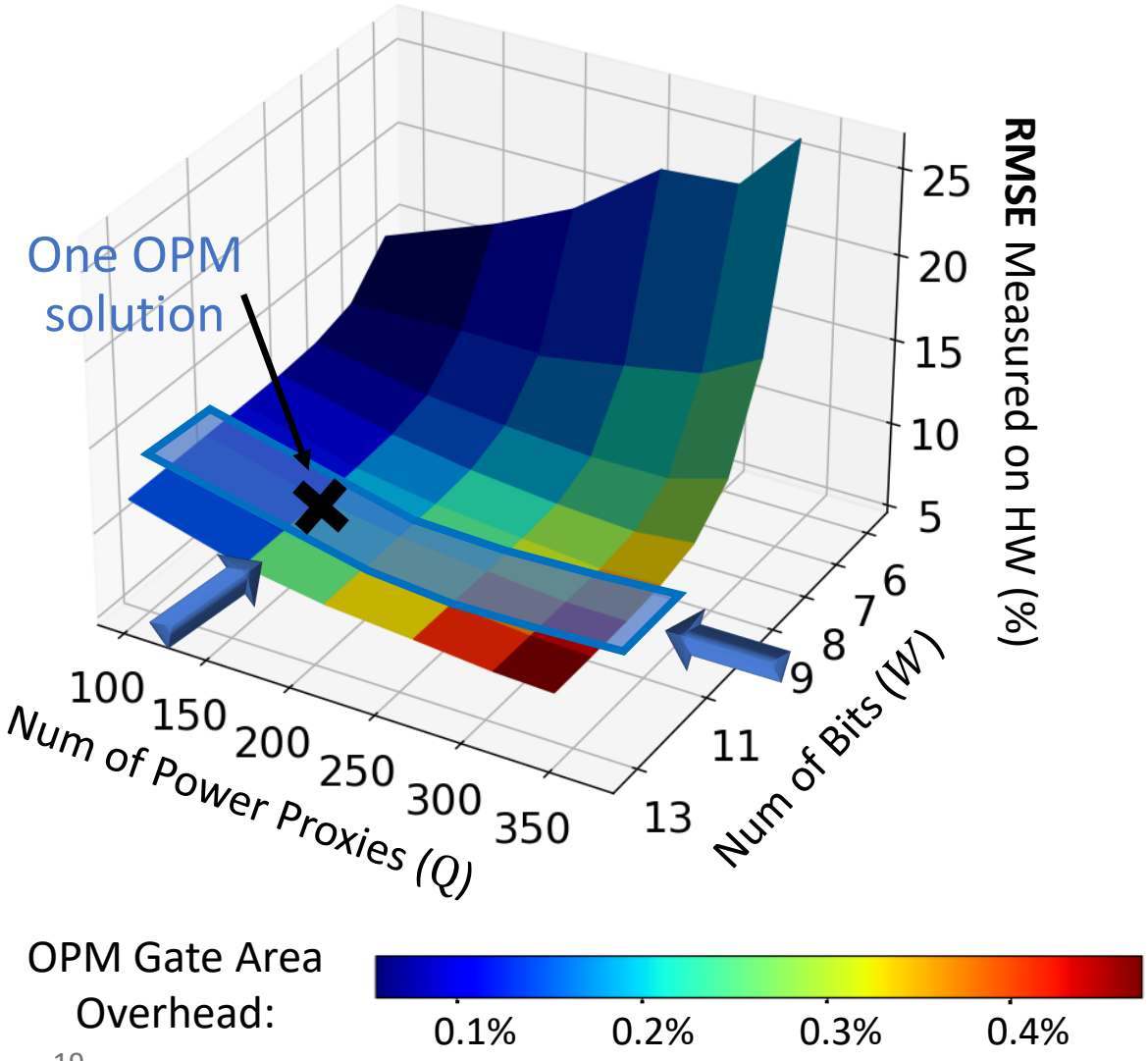
W-bit quantized fixed-point weights

No multipliers required

Duke UNIVERSITY

Accuracy vs. Hardware Cost (Area Overhead) of the OPM

Runtime OPM implementation on Neoverse N1



- Trade-off accuracy and hardware cost
- Sweep proxy num $\boldsymbol{Q}$ and quantization bits $\boldsymbol{W}$
- **Strategy**
  - Keep $\boldsymbol{W}$ = 10 to 12
  - Vary $\boldsymbol{Q}$ for different solutions
- **For an OPM with $\boldsymbol{Q}$=159, $\boldsymbol{W}$=11**
  - **< 0.2%** area overhead of Neoverse N1
  - **< 10%** in the error (RMSE)

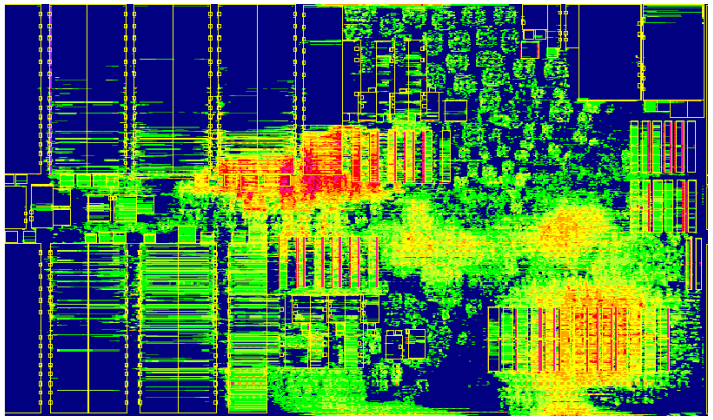
Duke UNIVERSITY

Case Study 2:

Routability Challenges

Routability Background

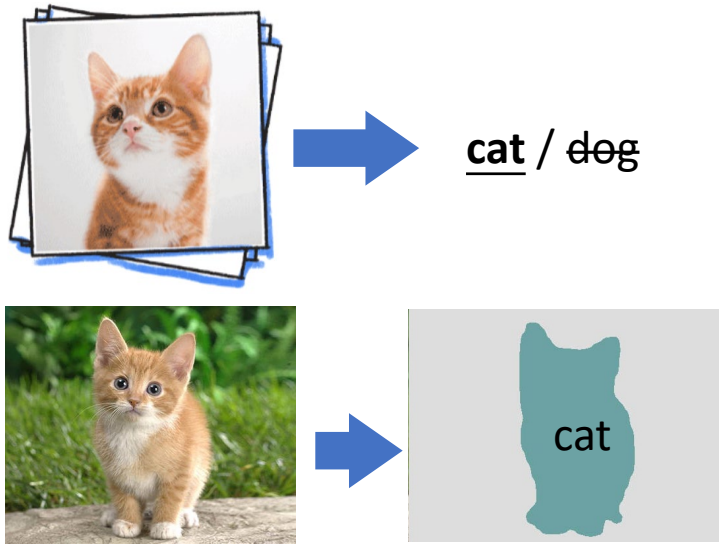
- Routability: post-routing design rule checking (DRC) violations
- Early routability prediction enables early mitigations of DRC violations
- Previous methods:
  - Analytical techniques: very fast but not enough fidelity
  - Trial routing: acceptable fidelity but not fast enough
  - Traditional machine learning (ML) models like logistic regression (LR), support vector machine (SVM): global information of the whole layout not captured

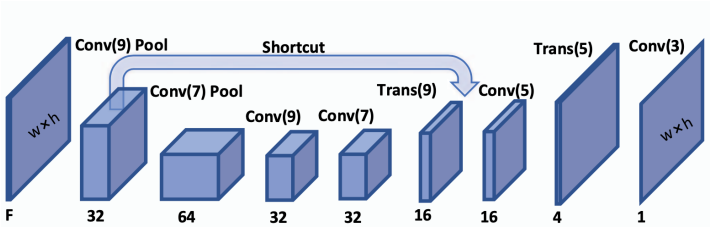


Routing congestion

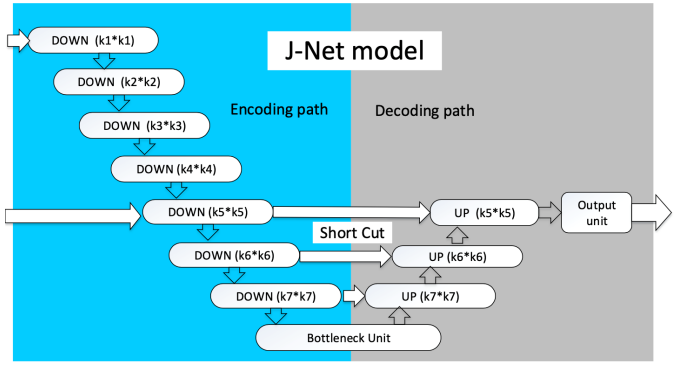
Our Works for Routability Prediction

- Input placements can be view as 2D images
  - CNN is naturally a good fit to predicting #DRV
- DRC hotspot is also a 2D prediction
  - Pinpointing DRC locations viewed as semantic segmentation
- CNN/FCN are suitable for routability prediction

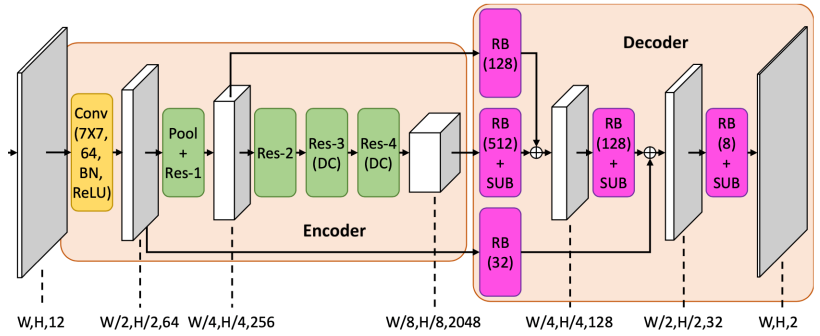




RouteNet [Xie, et al., ICCAD'18]



J-Net [Liang, et al., ISPD'20]

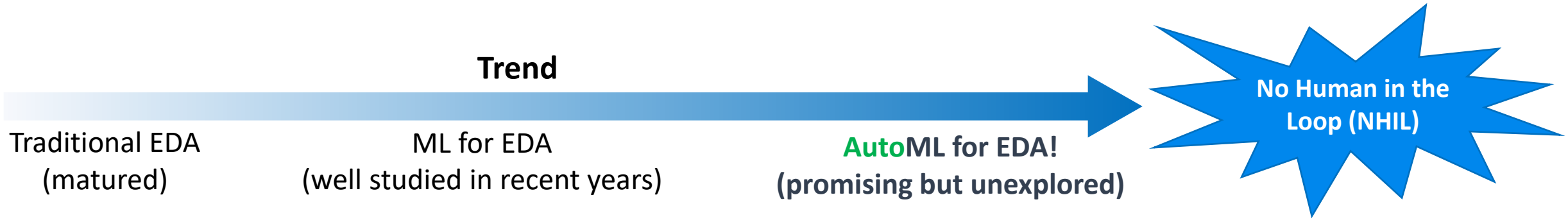


PROS [Chen, et al., ICCAD'20]

Duke UNIVERSITY

AutoML for EDA

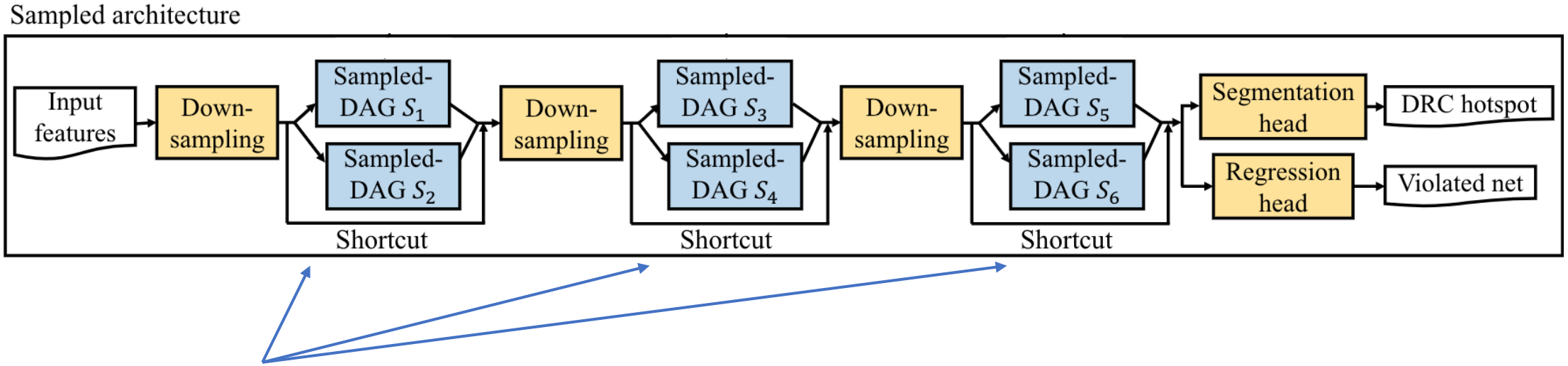
- **AutoML for EDA**: a higher level of automation
  - Truly no human in the loop



- Automate ML model designs: Neural Architecture Search

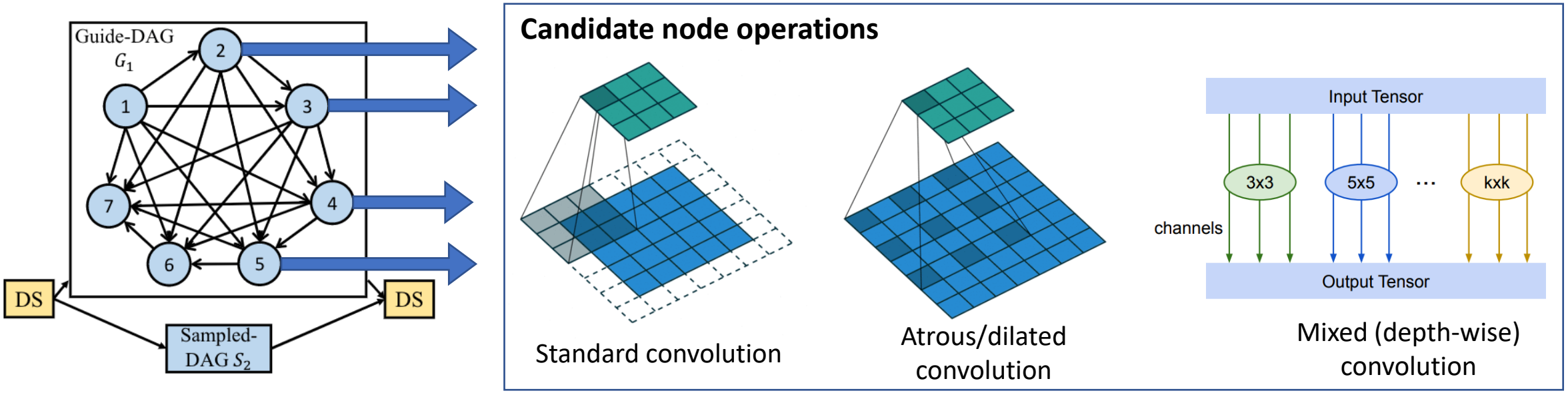
23

Methods – The Model Architecture



- Three layers, each with parallel and shortcut structures.
- Six **blue blocks** are searchable parts, **yellow blocks** are fixed parts.
- Two different heads for different tasks.
  - Regression head for **violated net count prediction**, providing one scalar number prediction
  - Segmentation head for **DRC hotspots detection,** providing two-dimensional prediction

Methods – The Model Architecture



- Three components in NAS: 1. search space; 2.evaluate strategy; 3.search strategy
- Sample from a completely-ordered graph (guide-DAG G_i) to form a sampled-DAG (S_i)
- After training, we get the evaluation metric of the sampled model
- Update the sampling probability (weight) by the metric accordingly

Experimental Results – Accuracy

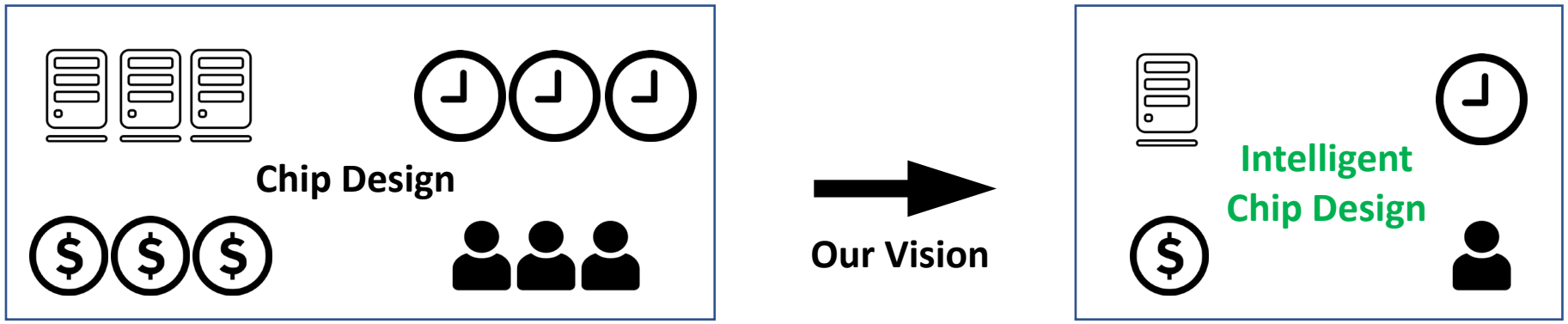
Comparison of the violated net count prediction

Models	Kendall's τ on designs (#nets)				Kendall's τ on all 74 designs	Pearson's correlation on all 74 designs
	s349 (270)	mem_ctrl (9.3k)	b17 (33.8k)	DSP (73.1k)		
RouteNet [3]	0.3620	0.1547	0.1779	0.4414	0.5264	0.7224
NAS-crafted model	**0.6369**	**0.4657**	**0.2683**	**0.7302**	**0.5572**	**0.7930**

Comparison of the DRC hotspot detection

Models	ROC-AUC on designs (#nets)				ROC-AUC on all 74 designs
	s349 (270)	mem_ctrl (9.3k)	b17 (33.8k)	DSP (73.1k)	
RouteNet [3]	0.829	0.844	0.902	0.866	0.847
PROS [6]	0.487	0.483	0.478	0.489	0.676
cGAN [4]	0.516	0.515	0.521	0.517	0.510
NAS-crafted model	**0.865**	**0.891**	**0.911**	**0.884**	**0.865**

Summary and Takeaway



- **Problem**: Increasing Challenges in Chip Design
  - Cost, Time-to-Market, Reliance on Designers, Diminishing Performance Return, ……
- **Our Work**: Intelligent Circuit Design and Implementation
  - Develop Customized ML Methods: Pruning, Linear Model, CNN, GNN, ……
  - Tackle Key Design Objectives: Power, Timing, Area, ……
  - Benefit: Less Simulation Time, Better Solution Quality, Less Human Designer, ……

References

- Zhiyao Xie, Yu-Hung Huang, Guan-Qi Fang, Haoxing Ren, Shao-Yun Fang, Yiran Chen and Jiang Hu, "RouteNet: Routability Prediction for Mixed-Size Designs Using Convolutional Neural Network," ICCAD, 2018
- Yu-Hung Huang, Zhiyao Xie, Guan-Qi Fang, Tao-Chun Yu, Haoxing Ren, Shao-Yun Fang, Yiran Chen and Jiang Hu, "Routability-Driven Macro Placement with Embedded CNN-Based Prediction Model," DATE, 2019
- Erick Carvajal Barboza, Nishchal Shukla, Yiran Chen, Jiang Hu, “Machine Learning-Based Pre-Routing Timing Prediction with Reduced Pessimism”, DAC, 2019.
- Zhiyao Xie, Guan-Qi Fang, Yu-Hung Huang, Haoxing Ren, Yanqing Zhang, Brucek Khailany, Shao-Yun Fang, Jiang Hu, Yiran Chen and Erick Carvajal Barboza, "FIST: A Feature-Importance Sampling and Tree-Based Method for Automatic Design Flow Parameter Tuning," ASPDAC, 2020
- Zhiyao Xie, Haoxing Ren, Brucek Khailany, Ye Sheng, Santosh Santosh, Yiran Chen, Hu Jiang, "PowerNet: Transferable Dynamic IR Drop Estimation via Maximum Convolutional Neural Network," ASPDAC, 2020
- Rongjian Liang, Zhiyao Xie, Jinwook Jung, Vishnavi Chauha, Yiran Chen, Jiang Hu, Hua Xiang, and Gi-Joon Nam. “Routing-Free Crosstalk Prediction.” ICCAD, 2020
- Zhiyao Xie, Hai Li, Xiaoqing Xu, Jiang Hu, Yiran Chen, “Fast IR Drop Estimation with Machine Learning” (invited) ICCAD, 2020
- Zhiyao Xie, Rongjian Liang, Xiaoqing Xu, Yixiao Duan, Jiang Hu and Yiran Chen, "Net2: A Graph Neural Network Method Customized for Pre-Layout Wirelength Estimation," ASPDAC, 2021
- Chen-Chia Chang, Jingyu Pan, Tunhou Zhang, Zhiyao Xie, Jiang Hu, Weiyi Qi, Chung-Wei Lin, Rongjian Liang, Joydeep Mitra, Elias Fallon, Yiran Chen, “Automatic Routability Predictor Development Using Neural Architecture Search”, ICCAD, 2021
- Zhiyao Xie, Xiaoqing Xu, Matt Walker, Joshua Knebel, Kumaraguru Palaniswamy, Jiang Hu, Huanrui Yang, Yiran Chen, Shidhartha Das, “APOLLO: An Automated Power Modeling Framework for Runtime Power Introspection in High-Volume Commercial Microprocessors”, MICRO, 2021
- Zhiyao Xie, Rongjian Liang, Xiaoqing Xu, Jiang Hu, Chen-Chia Chang, Jingyu Pan, Yiran Chen, Pre-Placement Net Length and Timing Estimation by Customized Graph Neural Network. TCAD. (Under Review)

Thanks! Questions?

Duke

- Title: Efficient Digital Design and Implementation with Machine Learning in EDA
-
- Abstract:
- EDA technology has achieved remarkable progress over the past decades. However, chip design is not completely automatic yet in general. For example, automation of EDA flow is still largely restricted to individual tools with little interplay across tools and design steps, and tools in early steps cannot efficiently judge if their solutions eventually lead to satisfactory designs. In addition, solutions are constructed from scratch even if similar optimizations have already been performed repeatedly. We believe such limitations can be largely addressed by knowledge reuse with machine learning, whose major strength is to explore highly complex correlations between design stages based on prior data. In this talk, I will share our recent research about customized ML algorithms in EDA. They cover a wide range of design stages from the RTL level to post-routing, solving primary chip-design problems including power, timing, interconnect, IR drop, routability, and design flow tuning. After introducing these research efforts systematically, I will present two latest progresses with more details. They are power estimation and monitoring implemented at the RTL level, and efficient routability prediction performed during layout. Finally, I will share our experience and vision in enabling efficient digital design and implementation with machine learning in EDA.
-

Duke UNIVERSITY