APOLLO:

AN AUTOMATED POWER MODELING FRAMEWORK FOR RUNTIME POWER INTROSPECTION IN HIGH VOLUME COMMERCIAL MICROPROCESSORS

Professor:
Prof .Vireshwar Kumar

Presented by:
Kashish jain
K.Laxman
Varshitha Reddy

INTRODUCTION:

Accurate power modeling is crucial for energy-efficient CPU design and runtime management.

An ideal power modeling framework needs to be accurate yet fast, achieve high temporal resolution (ideally cycle-accurate) yet with low runtime computational overheads.

Simultaneously satisfying such conflicting objectives is challenging and largely unattained despite significant prior research.

PROBLEM STATEMENT

- **PROBLEM 1**:**Design time CPU Power introspection :** Gain in IPC and FMAX
  And power resources are not keeping pace with CPU Power Demands

- **PROBLEM 2**:**Run time Introspection:** Peak power mitigation
  And Trigger abrupt changes in CPU current demand leading voltage-droop due to di/dt and as modern CPUs have complex underlying power

- Need software/algorithm that can predict and stop these undesirable extremes from happening, computer engineers can protect their hardware and increase its performance.

RELATED WORK:

Methods (hardware overhead %)	Demonstrated Application	Model Type	Temporal Resolution	PC / Proxy Selection	Cost or Overhead
300 %	Design-time software model	Analytical	>1K cycles	N/A	Low
		Proxies	>1K cycles	Automatic or no selection	High
					Medium
			Per-cycle		High
					Medium
16%	Design-time FPGA emulation	Proxies	Per-cycle	Automatic	High
			~100s cycles		Medium
4-10%			Per-cycle	Hybrid manual/auto	
	Runtime monitor	Event Counters	>1K cycles	Manual	Low
			~100s cycles		
		Proxies	>1K cycles	Automatic	Medium
APOLLO (0.2% overhead)			~100s cycles		
	Design-time model Runtime monitor	Proxies	Per-cycle	Automatic	Low

Table: Comparison among various power modeling approaches. The percentage numbers are area overheads

PROPOSED SCHEME:

APOLLO: An ideal Power estimation algorithm that is both accurate and fast and can easily be built into a processing core at a low power cost.

AI is used to detect and monitor a few signals to forecast chip performance in real time. Its learning process is autonomous and data-driven, so it can be implemented on any computer processor architecture.

Key Objectives :

Automated Power proxy Extraction:

- Uses ML(MCP) techniques to identify power correlated contributing events in design

Fast yet accurate on chip metering :

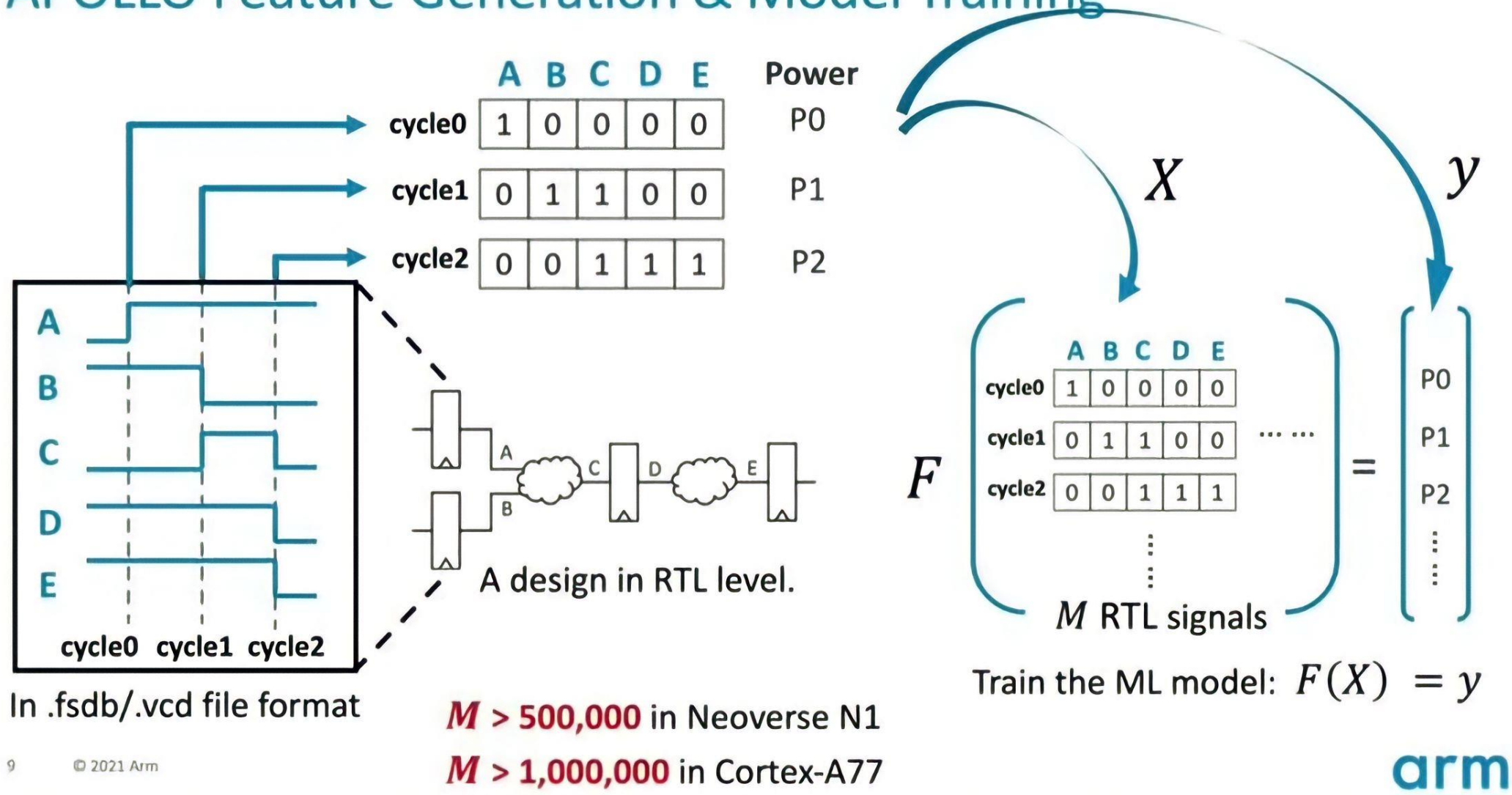
-proven on commercial CPUS with ~95% accuracy over different microprocessor

-0.2% area overhead over Neoverse N1 core Arm processor

Extensible to higher abstraction simulation :

-trade-off accuracy for pre-identified events

APOLLO Feature Generation & Model Training



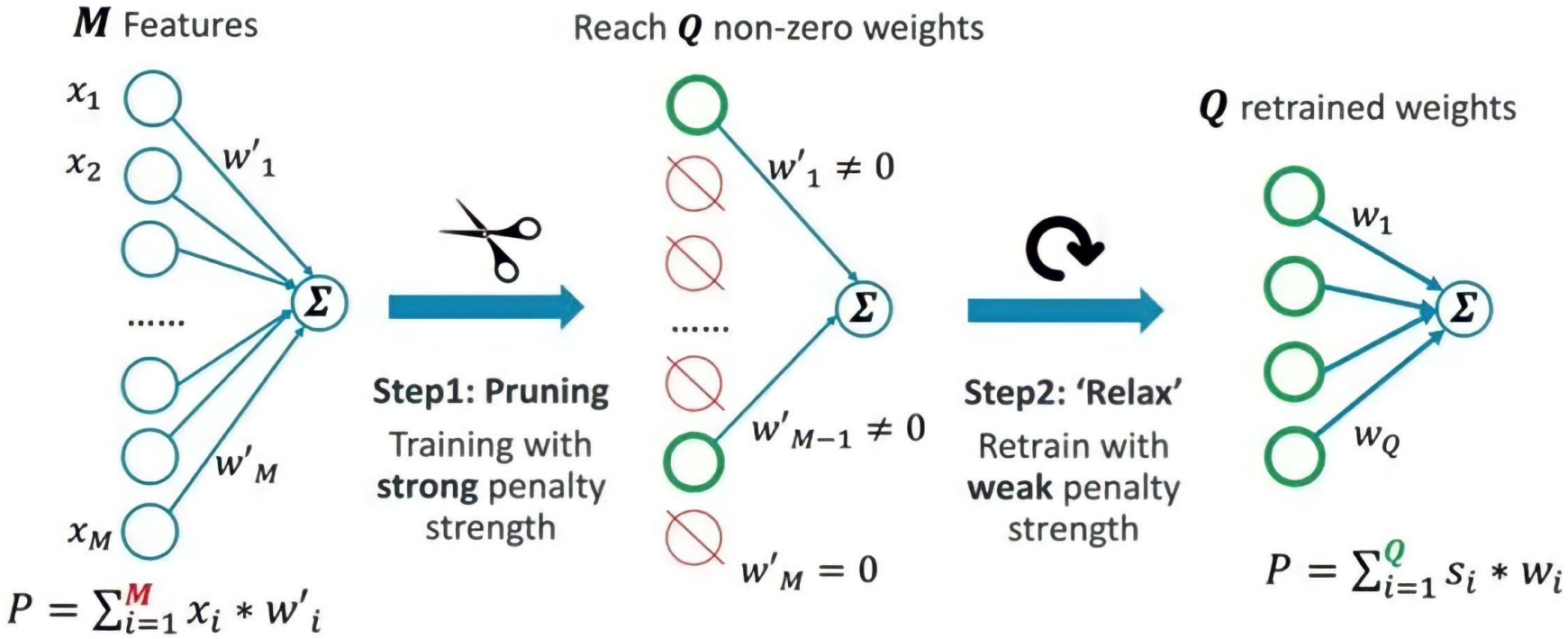
arm

Shows the procedure to construct features from the RTL- simulation traces and labels from power simulation results

ML-Based Power Proxies Selection

Model construction in two steps

Please check our paper for detailed discussion on MCP method



Minimax concave penalty (MCP) for pruning

arm

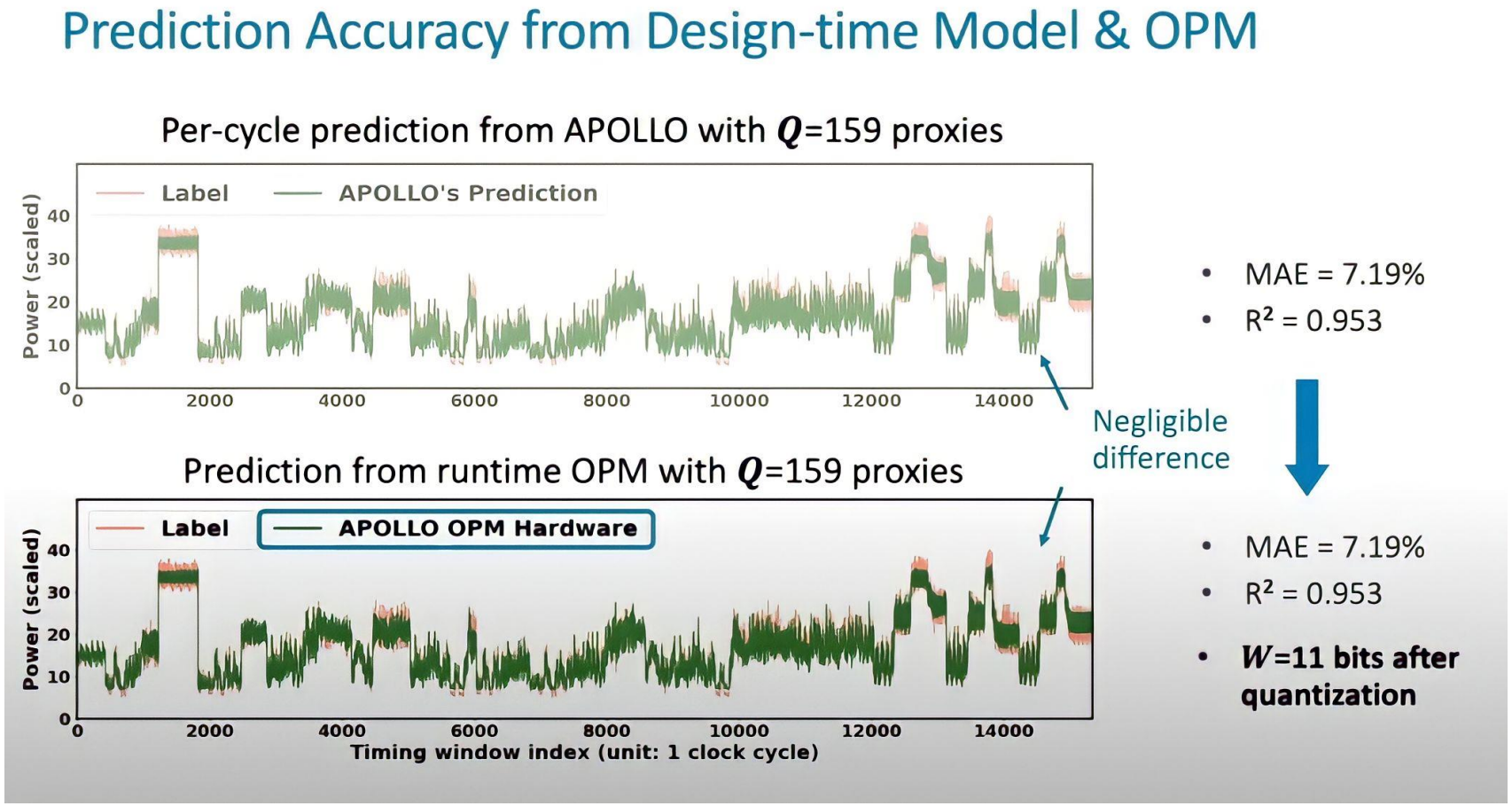
Above model is minimising the prediction error during training and also simultaneously shrinking all weights so that majority of weights eventually become zero. Non zero weights are selected as power proxies this method is called pruning.

• EXPERIMENTAL RESULTS

- **TRAINING DATA** :- From more than 1000 random micro-benchmarks 300 micro benchmark are selected
- **TESTING DATA** :- 12 representative micro-benchmarks handcrafted by CPU designers corresponding to various use cases
- Implemented the MCP algorithm and the coordinate descent algorithm with NumPy .
- MCP regressor converges within 200 iterations, with the threshold of unpenalized weights set to $\gamma = 10$.
- Experiments are performed on
  - --> Neoverse N1 - trace lengths N for training and testing are 30,000 and 15,000 cycles
  - --> Cortex-A77 - trace lengths N for training and testing are 5,000 and 2,000 cycles

The numbers of RTL signals M are $> 5 \times 10^5$ and $> 1 \times 10^6$ for Neoverse N1 and Cortex-A77, respectively.

• ACCURACY OF APOLLO



- Detailed evaluation of the APOLLO model with Q = 159, which obtains NRMSE = 9.4% and R^2 = 0.95 using 15, 000-cycle testing dataset, covering all 12 handcrafted microbenchmark.

CONCLUSION:

- Fast Power modeling has a material impact in how we design and deploy CPUs

- Here demonstrated that by monitoring < 0.05% RTL signals, the OPM achieves R^2 > 0.95 with <1% area/power overhead when integrated with Neoverse N1.

- Micro-architecture agnostic methodology is automated and can scale to multiple compute-solutions-CPUs, GPUs, NPUs, and even for sub blocks

- In modern computer processors, cycles of computations are made on the order of 3 trillion times per second. Keeping track of the power consumed by such intensely fast transitions is important to maintain the entire chip's performance and efficiency

- Ultimately, the APOLLO capability can enable the development of new mechanisms for smarter power and thermal management in future SoCs.