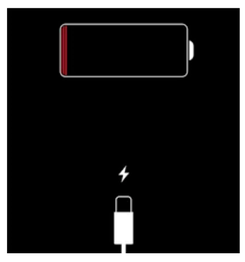
Simmani: Runtime Power Modeling with Automatic Signal Selection

Donggyu Kim, Jerry Zhao, Jonathan Bachrach, Krste Asanovic

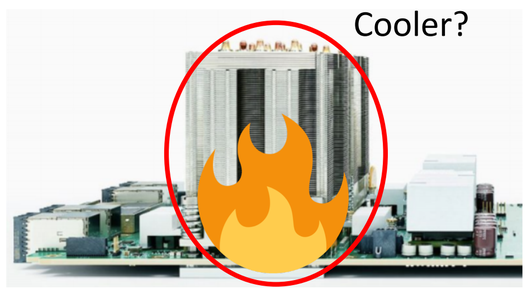
Why do we care about power & energy efficiency?

- All Day Battery Life

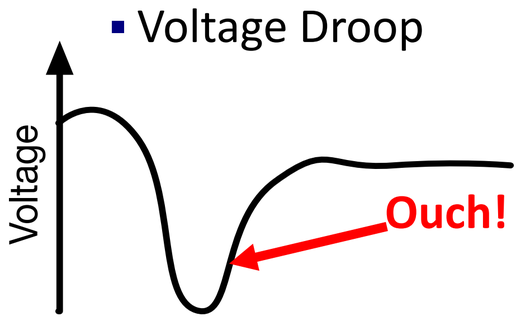


- Electricity Bill

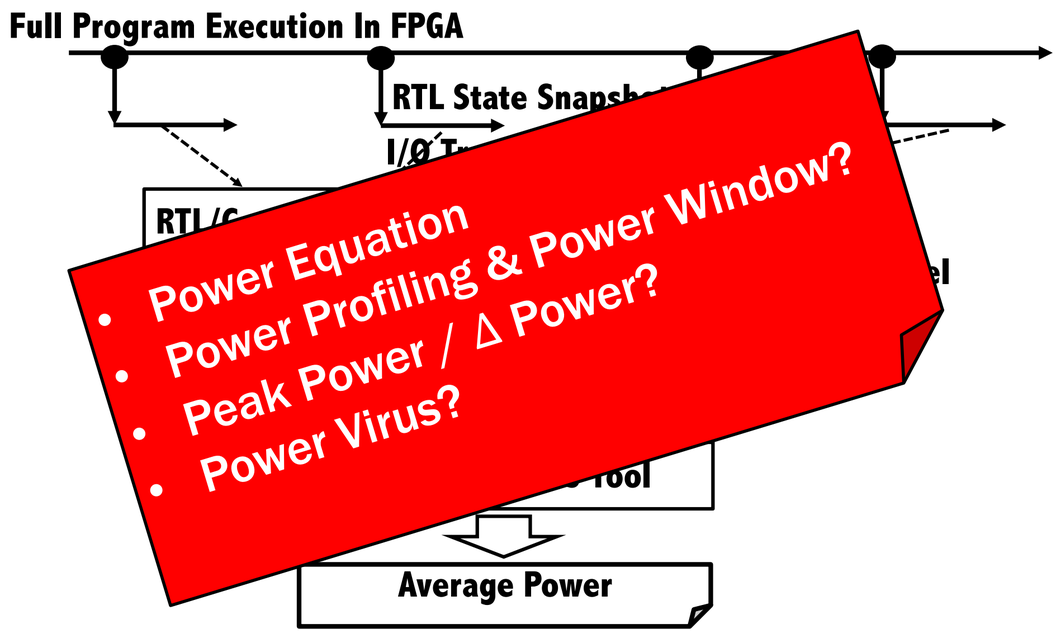
- Heat



- Voltage Droop



Strober Power/Energy Modeling [ISCA `16]

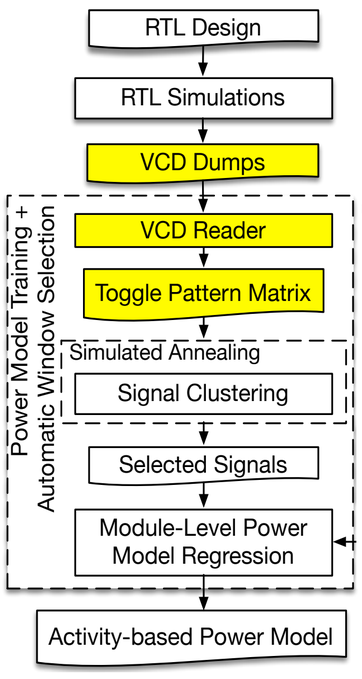


[ISCA `16] Kim et al. "Strober: Fast and Accurate Sample-Based Energy Simulation for Arbitrary RTL"

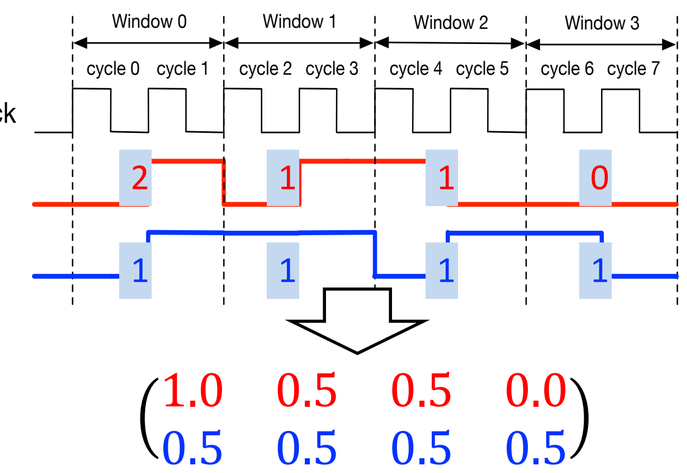
Our Approach

- Goal
  - Find *key signals for power dissipations from any RTL*
- Observation
  - Signals showing similar toggle patterns
    ➔ Similar effect on dynamic power dissipation
- Our Approach
  - Construct ***toggle pattern matrix*** from VCD dumps
  - Select key signals with ***signal clustering***
  - ***Module-level power model*** regression against power traces from CAD tools
  - ***Automatic counter instrumentation*** for runtime power estimation with FPGAs

Constructing Toggle Pattern Matrix

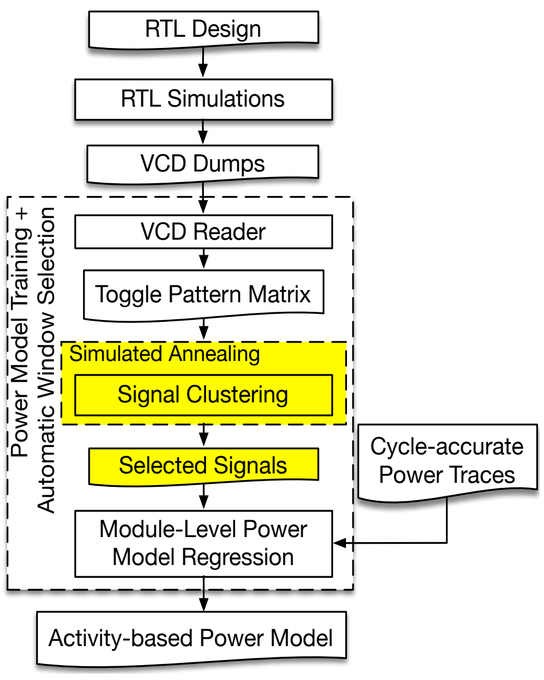


- Each element: toggle densities of signal in window



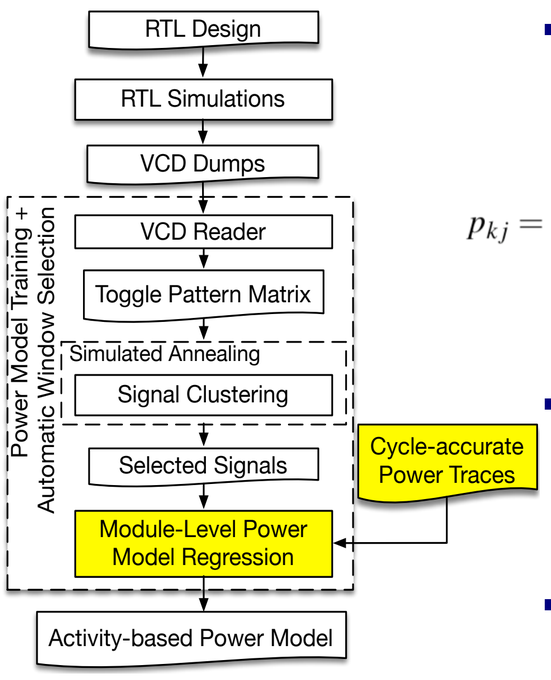
- Matrix is big, but signal activities are sparse
  ➔ Compressed sparse row (CSR) format

Selecting Key Signals



- Signal clustering
  - Project high-dimensional data into low-dimensional space
  - Clustering (k-means) on the projected points
  - Select signals closed to the cluster centers
- Optimal number of signals
  - Model selection (# clusters) with Bayesian Information Criterion (BIC)
  - Simulated annealing for global optimum

Power Model Regression



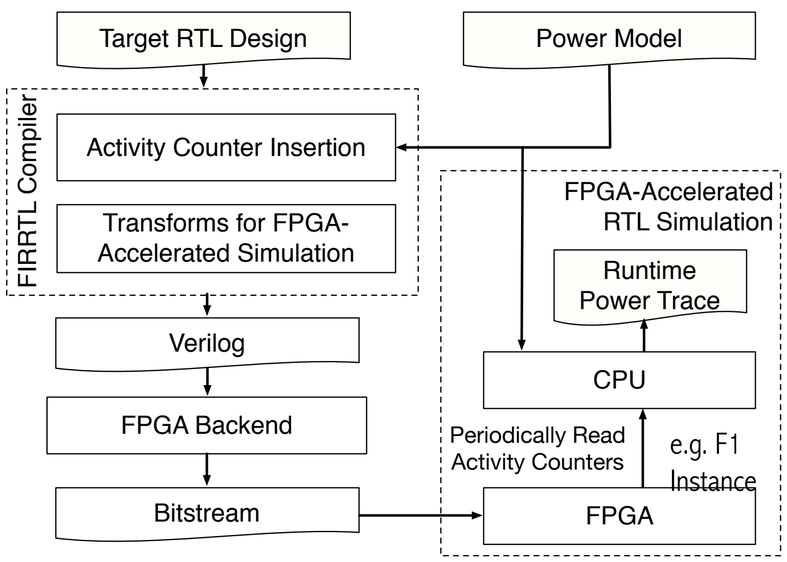
- Linear model with polynomial features → simple, intuitive, low training & inference overhead

$$p_{kj} = \alpha + \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots \beta_n x_{nj} + \beta_{11} x_{1j}^2 + \beta_{22} x_{2j}^2 + \cdots + \beta_{12} x_{1j} x_{2j} + \cdots + \beta_{123} x_{1j} x_{2j} x_{3j} + \cdots$$

- Regularization & variable selection with the elastic net → prediction accuracy + interpretability
- Constrain coefficients >= 0 (except uncore)

Power Model Instrumentation

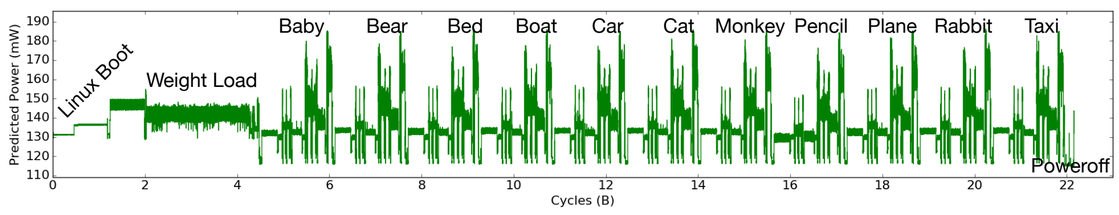
- Automatically insert activity counters for selected signals
- Read counters periodically from FPGA-accelerated simulation



Power Trace of SqueezeNet on Hwacha

- Inference on 11 images (~22B cycles)
- Counters are samples every 100K cycles from the FPGA
- Errors against Strober <= 10%

SqueezeNet Baseline



SqueezeNet Quantization+Compression