UCLA

PROCEED: Pareto Optimization-based Circuit-level Evaluation Methodology for Emerging Devices

ASP-DAC 2014

Shaodi Wang, Andrew Pan, Chi-On Chui and Puneet Gupta

Department of Electrical Engineering, University of California, Los Angeles

Need for Device Evaluation

- Traditional CMOS technologies are reaching physical limits
- Many alternative emerging devices under investigation: TFET, CNT, Heterogeneous CMOS, etc. → need to be able quickly compare them to guide technology development
- How should we compare emerging devices ?
  - Comprehensive, systematic and automated comparison in *context of how they are going to be used*
    - *Account for various design types and circuit-level optimizations*
  - Fast and flexible evaluation framework
  - Cover the wide performance range (KHz to GHz)

Prior Work

- Three classes of works
  - Devices level[1,6]: I_{on}/I_{off}, Subthreshold Slope (SS), CV/I, CV^2
  - Canonical circuit level: Simple circuits + Analytical model[4] based power-delay tradeoff
  - Full design flow: Library generation, Synthesis, Placement and Routing
- Existing evaluation benchmarks neglect how modern circuits really use devices → <u>These can dramatically change the conclusions.</u>
  - Circuit topology dependence (e.g., logic depth)
  - Design-time power optimization (Multi-V_{th} and multiple gate sizes)
  - Runtime adaptive power management (DVFS, Gating)

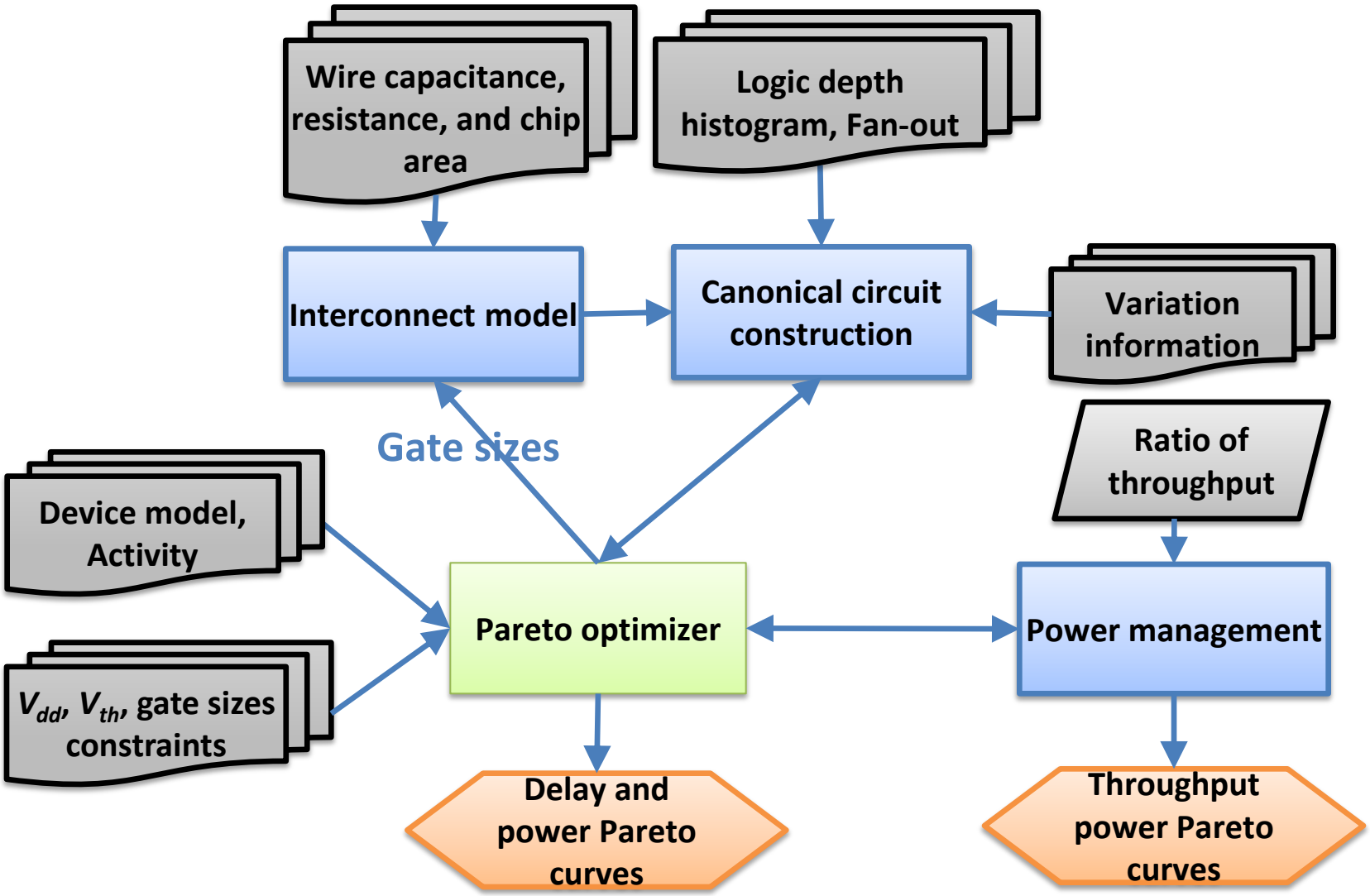
[1]L. Wei, et. al., IEDM, 2010
[4]D. J. Frank, et. al., IBM J, 2006
[6]M. Luisier, et. al., IEDM, 2011

Outline

PROCEED Methodology

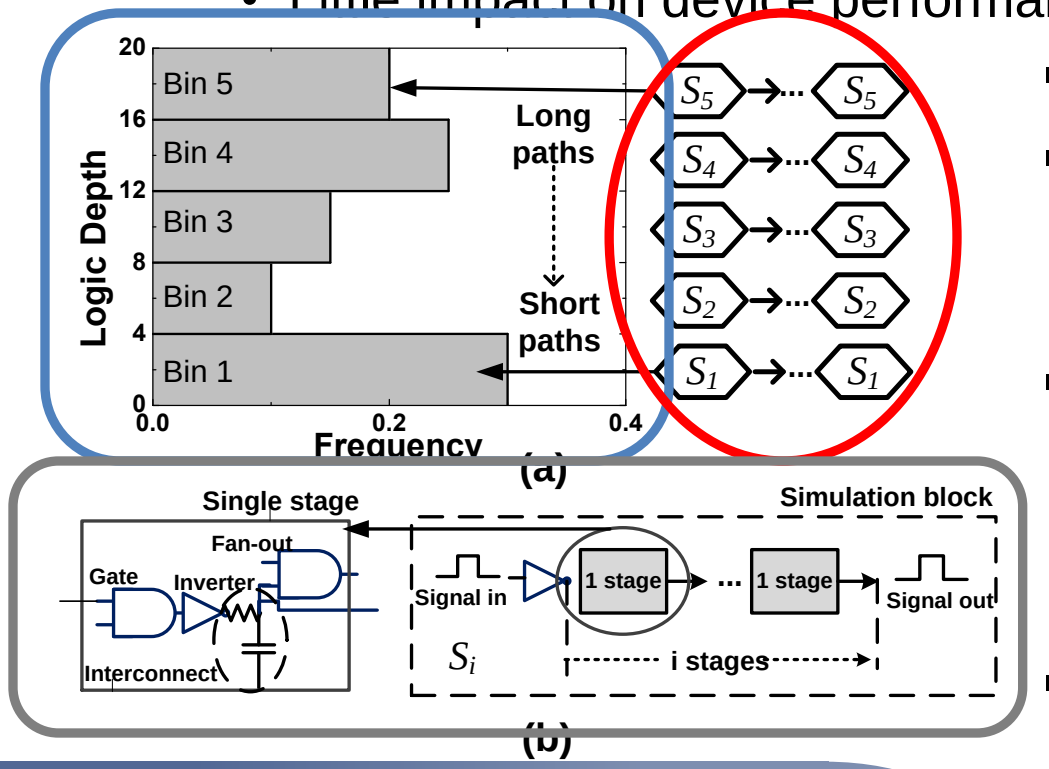
Example Experimental Results

Proposed Framework: PROCEED



Canonical Circuit Construction

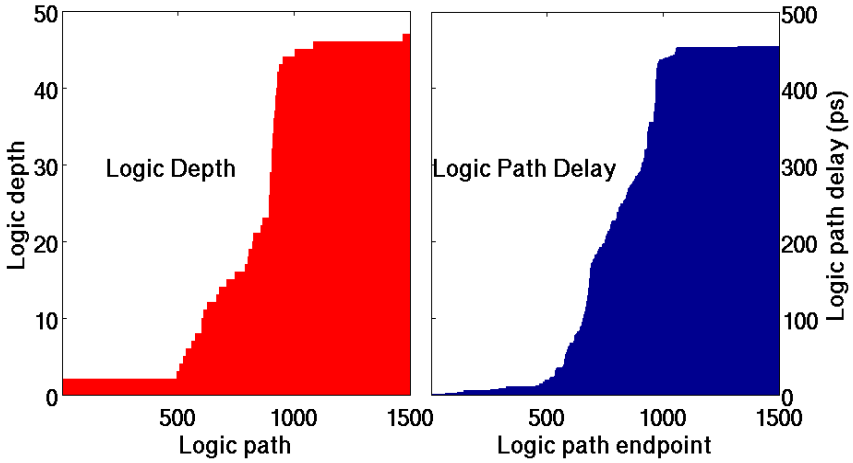
- Utilize essential design information
  - Logic depth histogram, average number of transistors per gate, average fan-out, average interconnect load and chip area
  - Ignore detailed circuit design
    - Little impact on device performance evaluation



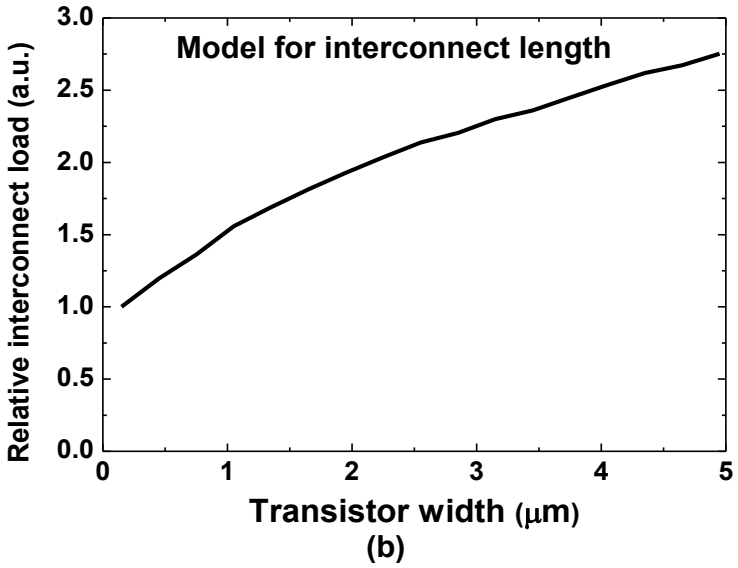
- **Logic depth histogram**
- **Simulation blocks (S_i)**
  - Construct logic paths in corresponding bins
- **Single stage**
  - Gate (Nand, XOR, etc.)
  - Buffer, interconnect, fan-out load
- **Tuning parameters**
  - Gate sizes, V_{dd}, V_{th}

Canonical Circuit Construction

- **Logic depth histogram is estimated using slack histogram**

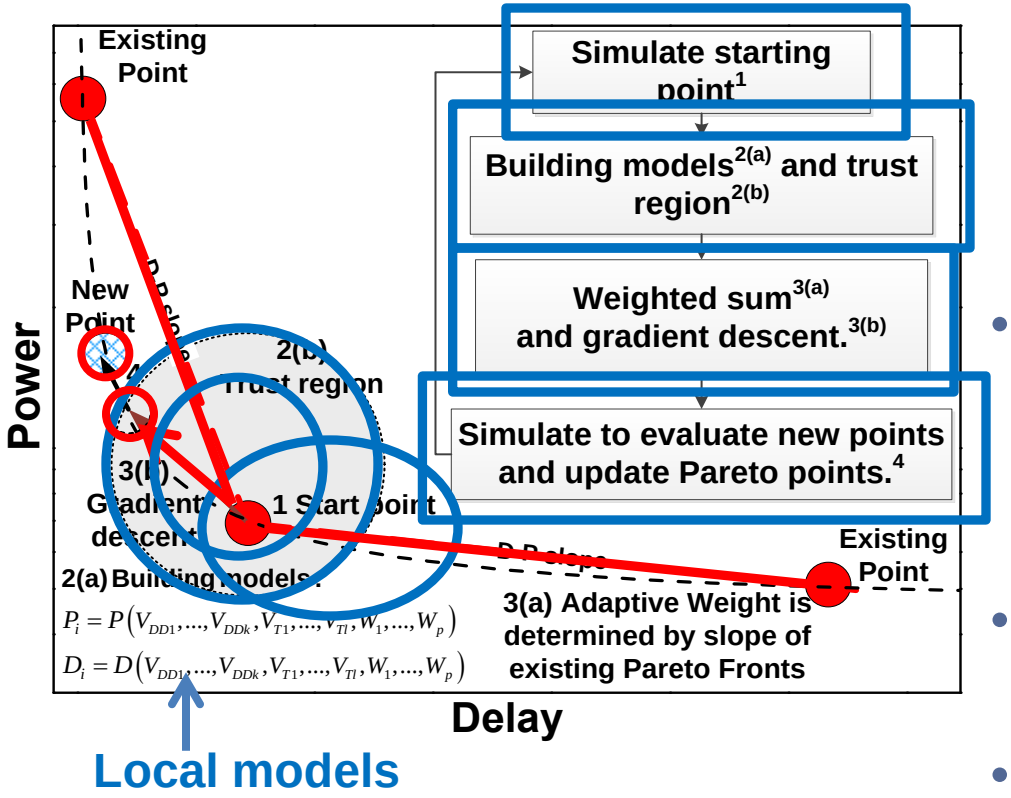


- **Interconnect**
  - Interconnect is proportional to the square-root of chip area[1]
  - Chip area is assumed to be linear to cell area
  - Cell area is modeled as a function of transistor width from DRE[2]



[1] J. A. Davis, et. al.,, *IEEE TED*, 1998.
[2] R. S. Ghaida and P. Gupta, *IEEE Trans. CAD*, 2012.

Pareto Optimization: Overview



- Objective functions are weighted sum of delay and power
  - Non-convex problem
  - Gradient descent used
- Delay and power are approximated with second order functions in trust region
- Trust region shrinks during optimization
- Logarithmic barrier is incorporated to confine parameter range

Pareto Optimization: Modeling

- **Build simulation-block-level delay and power models by utilizing circuit simulations results.**

$$D_{Si}\left(\boldsymbol{y_{i,0}} + \boldsymbol{\Delta y_i}\right) = D_{Si,0} + \boldsymbol{G_{Di}^T \Delta y_i} + \frac{1}{2}\boldsymbol{\Delta y_i^T H_{Di} \Delta y_i}$$

$$P_{Si}\left(\boldsymbol{y_{i,0}} + \boldsymbol{\Delta y_i}\right) = P_{Si,0} + \boldsymbol{G_{Pi}^T \Delta y_i} + \frac{1}{2}\boldsymbol{\Delta y_i^T H_{Pi} \Delta y_i}$$

- **Objective delay is the longest delay of all logic paths (constructed by simulation blocks)**
  - Using high order norm to estimate max-delay function

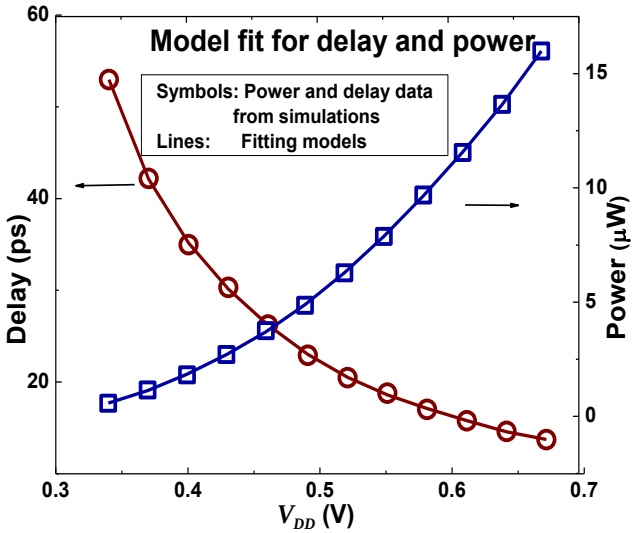
    → This can make the objective function continuous for gradient calculation

$$D(\boldsymbol{X}) = W_D \cdot \max\left(\left(D_{S1}(\boldsymbol{y_1}), D_{S2}(\boldsymbol{y_2}), ..., D_{Sn}(\boldsymbol{y_n})\right)\right) \approx W_D \cdot \left\|\left(D_{S1}(\boldsymbol{y_1}), D_{S2}(\boldsymbol{y_2}), ..., D_{Sn}(\boldsymbol{y_n})\right)\right\|_K$$

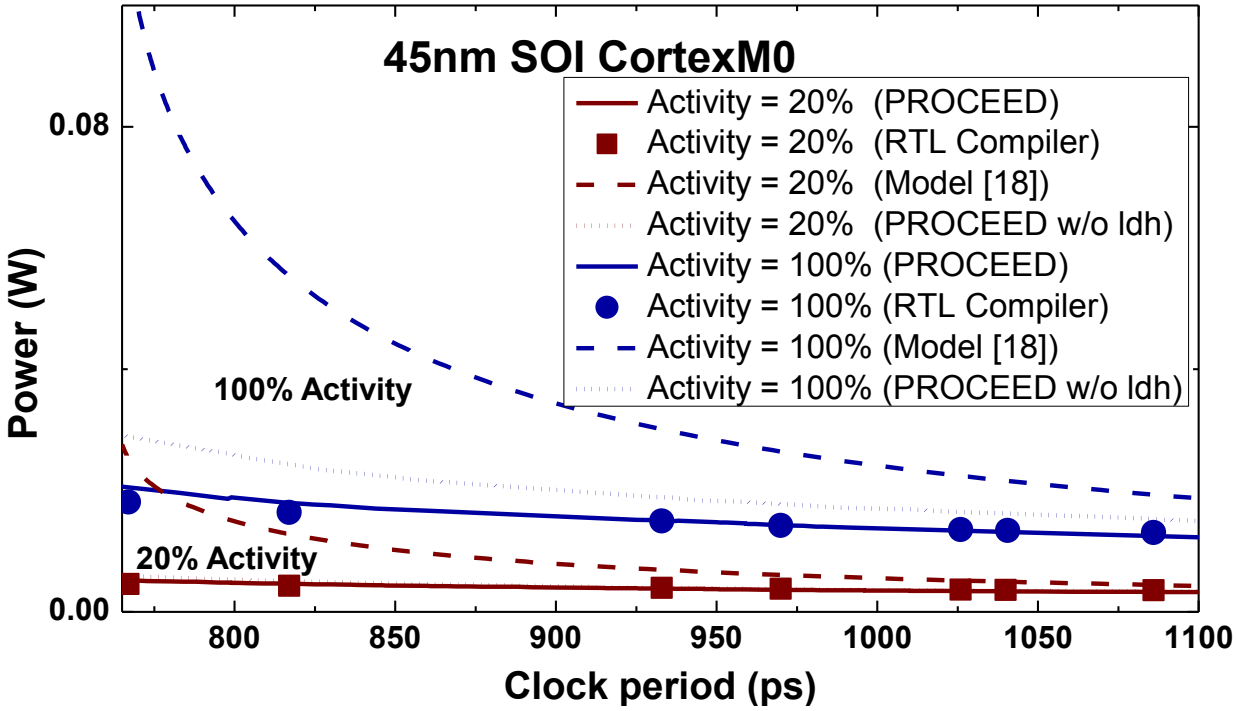
- **Objective power is the weighted total power** $P = \sum_{i=1}^{n} W_i \cdot P_{Si}$

Power Management: Overview

- **Modern circuits allow devices to operate in three modes: normal, power saving and sleep mode.**
  - A 2nd lower V_{dd} is applied to devices in power saving mode
  - Device is turned off in sleeping mode
- **Pick best 2nd V_{dd} and optimally divide time spent on each mode**
  - Need input of the ratio of average throughput to peak throughput
  - Use polynomial models for delay and power of simulation block as a function of V_{dd}
- **Minimizing average power consumption: $f_1P_1 + f_2P_2$**



Validation of PROCEED



- **Model[3] is frequently used for device evaluation**
  - Ignoring logic depth histogram and using analytical delay and power models is inaccurate
- **Proposed methodology is 21X (average) more accurate**
  - Efficient to evaluate devices with performance range from MHz to GHz

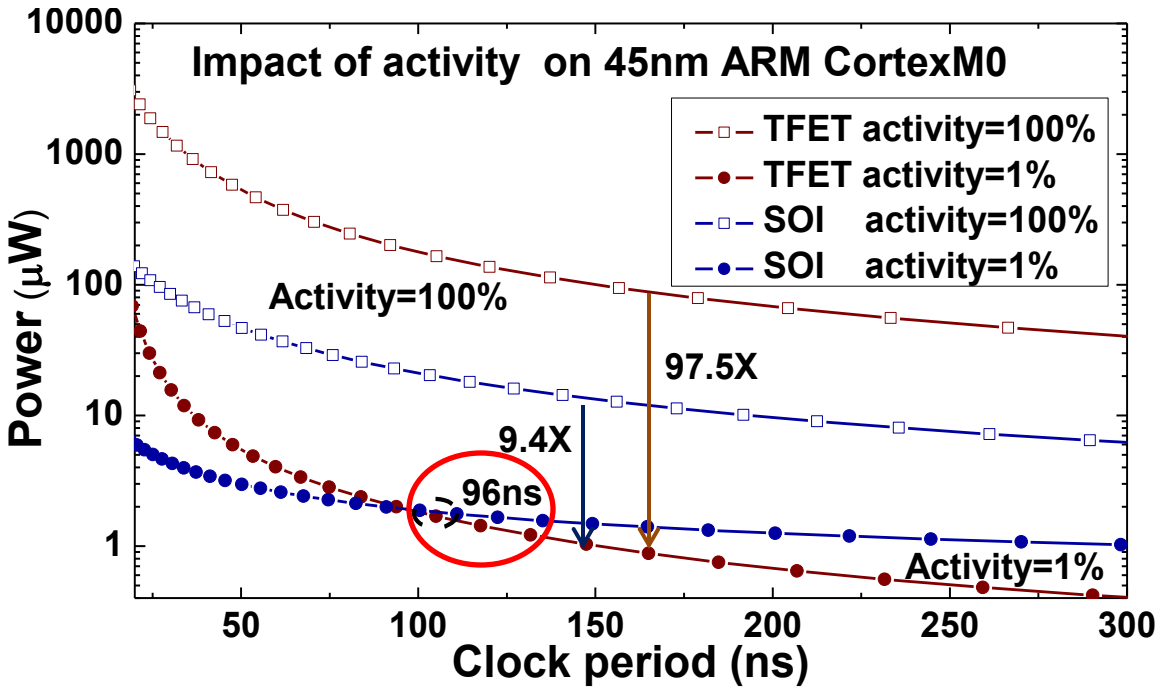
[3]D. J. Frank, et. al., *IBM J*, 2006

Outline

PROCEED Methodology

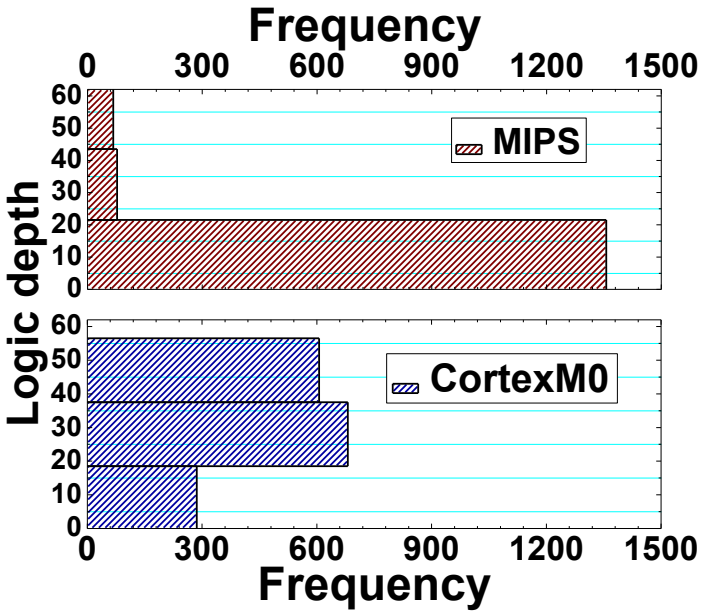
Example Experimental Results

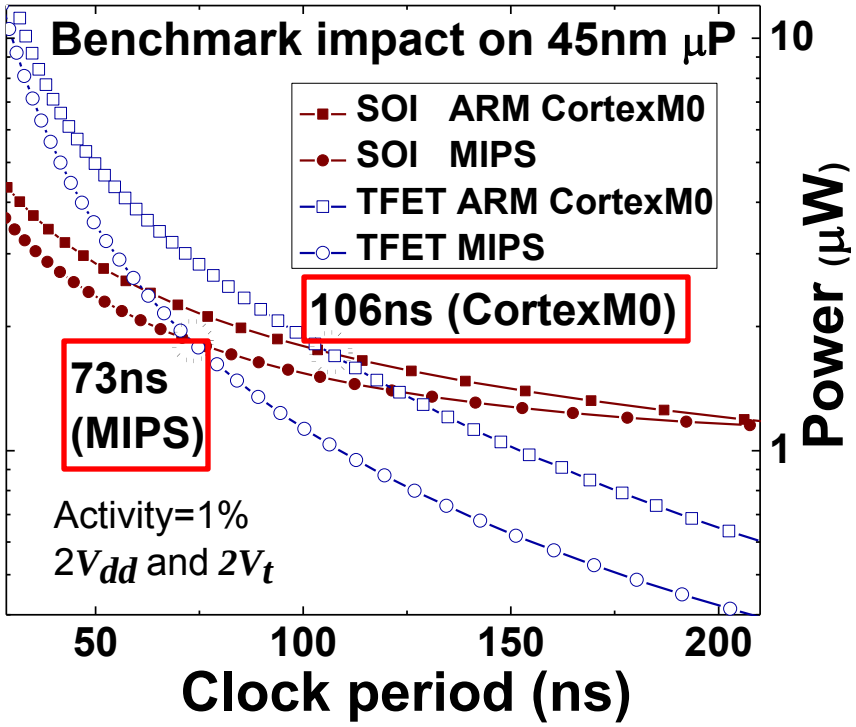
Impact of Activity



- **Low activity circuit benefits low leakage devices**
  - TFET better than SOI MOSFET for low-activity, low-performance circuits

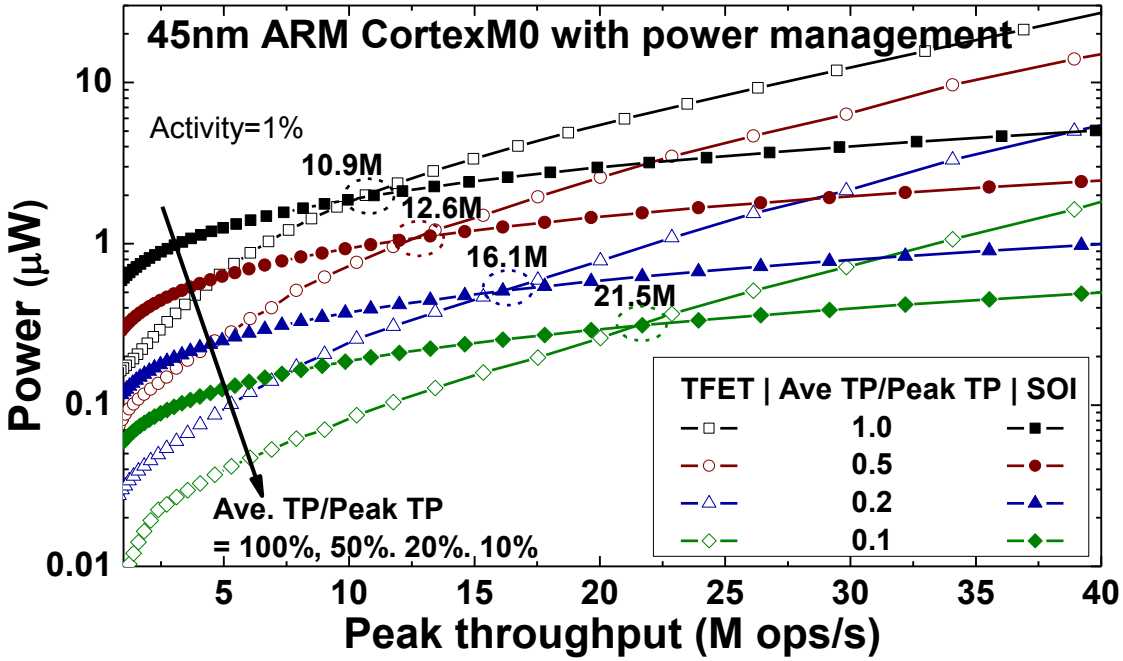
Impact of Circuit Topology





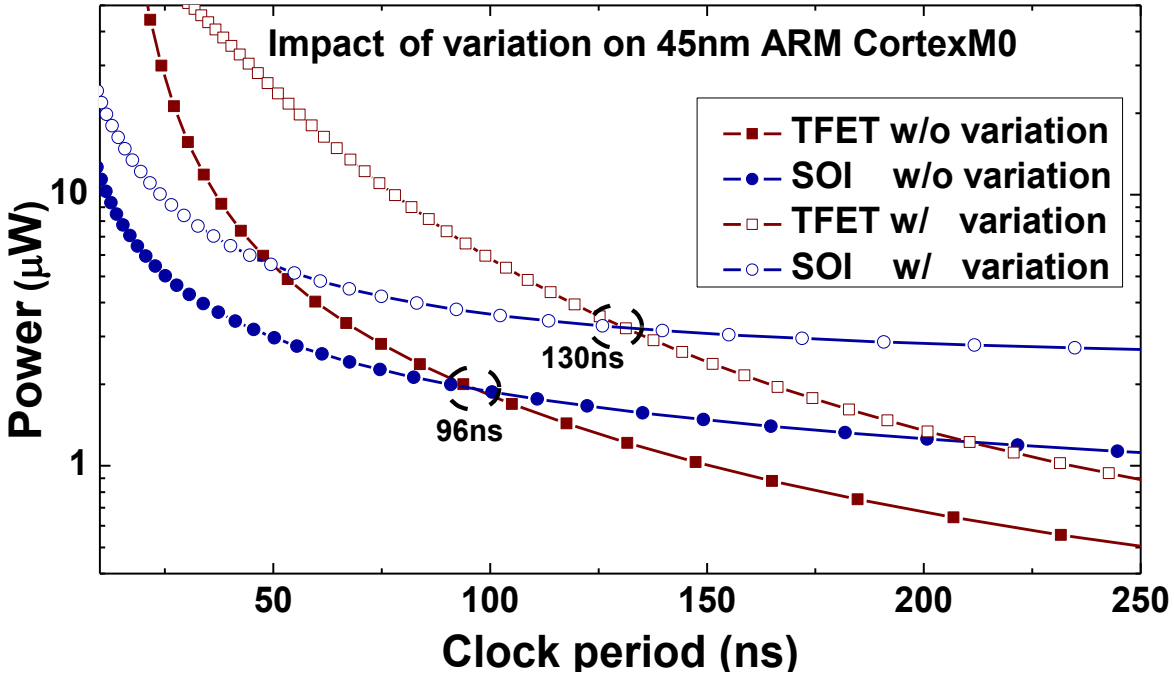
- **CortexM0 is more evenly distributed in LDH**
- **Power consumption in MIPS is dominated by short logic paths**
  - More accommodating to low power devices (TFETs)

Impact of Power Management



- **Peak throughput cross-points shift higher as ratio of average to peak throughput decreases**
  - This indicates TFET may be a better device for applications with wide dynamic range in performance needs

Impact of Variation



- **Variation evaluation indicates that TFET suffers more from effective voltage drop**
  - TFET and SOI are assumed to have peak 10% Vdd and 50mV Vth worst case variation
  - High Sub-threshold Slope (SS)devices are more sensitive to voltage drop

Conclusion

- **Proposed new methodology for evaluating emerging devices accounts for circuit topology, adaptivity, variability, and use context**
- **Proposed methodology is efficient and accurate for device evaluation over broad operating range.**
  - Effective Pareto -based optimization heuristic
  - Accurate circuit simulation and device compact model
- **Example comparison of TFET and SOI devices**
  - TFET is better for low activity, low logic depth and high dynamic performance design
  - TFET is more sensitive to voltage drop and threshold voltage shifting
- ***Entire PROCEED source-code (MATLAB, C++) will be made available openly.***

Q&A

Thanks!