Cache Design Under Spatio-Temporal Variability

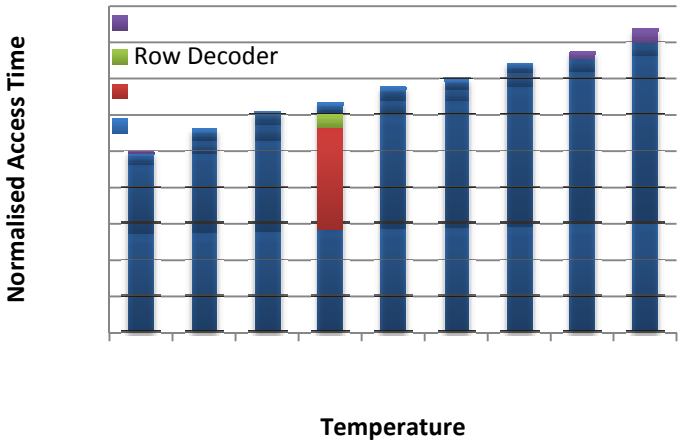
Shrikanth Ganapathy[1], **Ramon Canal**[1], **Antonio Rubio**[1], **Antonio Gonzalez** [1,2]

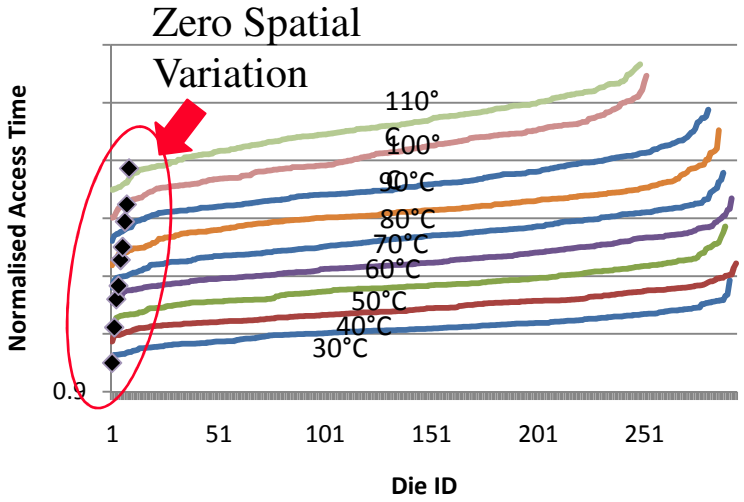
[1] *Universitat Politecnica de Catalunya, Spain*

[2] *Intel Barcelona Research Center, Spain*

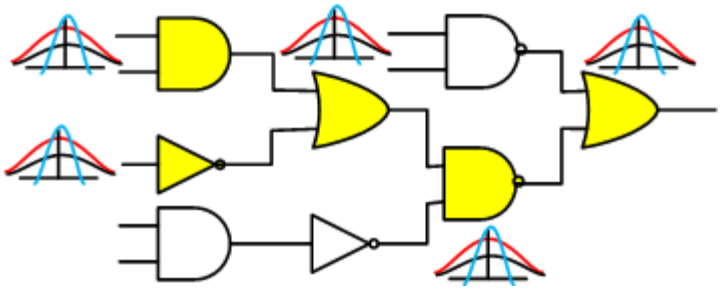
Motivation

- Manufacturing process induce variation in device parameters (Spatial).
- Adverse operating conditions make reliable operation tougher (Temporal).
- With reducing feature sizes, Memory designed from minimum geometry transistors will suffer the most from Intrinsic variations.
- Corner-case estimation of energy/delay imperative at design time.
- We propose to use a combination of Simulation and multivariate regression based curve fitting for analysis.
- Such idea also provides platform for simultaneous co-exploration of circuit-centric optimizations.





Delay Estimation



- **High dependence of delay on temperature translates to multiple PDFs as against single PDF suggested by SSTA techniques.**

$$D_i = \begin{vmatrix} \delta_{l_{eff}}^{p(1)} & \delta_{v_{th}}^{p(1)} & \delta_{l_{eff}}^{n(1)} & \delta_{v_{th}}^{n(1)} \\ \delta_{l_{eff}}^{p(2)} & \delta_{v_{th}}^{p(2)} & \delta_{l_{eff}}^{n(2)} & \delta_{v_{th}}^{n(2)} \\ \vdots & \vdots & \vdots & \vdots \\ \delta_{l_{eff}}^{p(j)} & \delta_{v_{th}}^{p(j)} & \delta_{l_{eff}}^{n(j)} & \delta_{v_{th}}^{n(j)} \end{vmatrix} \mathrm{X} \begin{vmatrix} m_{leff}^{pmos} \\ m_{vth}^{pmos} \\ m_{leff}^{nmos} \\ m_{vth}^{nmos} \end{vmatrix}$$

$$+ j * (D_{\text{nominal}}^{pmos} + D_{\text{nominal}}^{nmos}) + m_{temp} * \delta_{temp}$$

- **As delay is linearly dependent on threshold, effective length, we begin with a first order polynomial and fit to the best curve.**

Energy Estimation

- **The static and dynamic energy of every component in the array sub-block is estimated at different instants of time.**
- **Energy is estimated as a function of the integral of current through the supply (non-capacitive).**

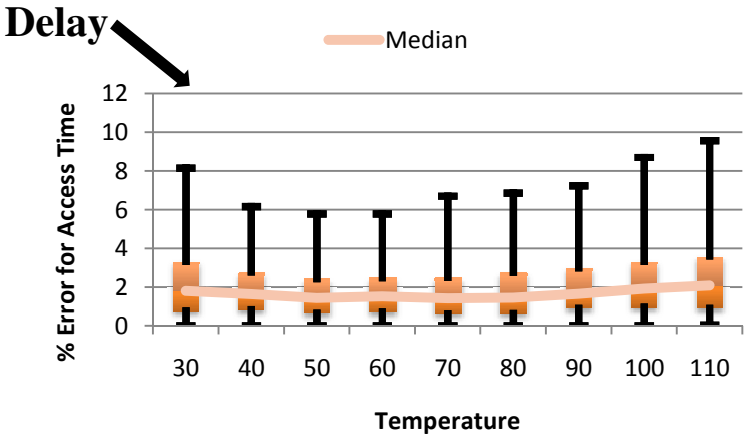
$$Cache_{Energy} = \begin{bmatrix} (E_{precharge} + E_{column-mux} + E_{Driver} + E_{Decoder}) + p * E_{(active/cell)} \\ + (n-p)E_{(wline-active/cell)} + (m-1)E_{(bline-active/cell)} + E_{control} \\ + (m-1)(n-p)E_{(inactive-cells/cell)} \end{bmatrix}$$

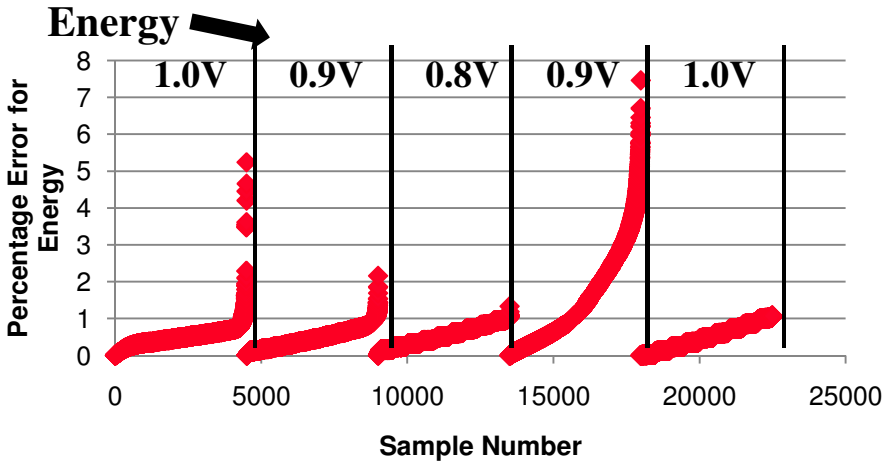
$$+ f(m,n) * E_{inactive-block}$$

- **In order to reduce the dimensions of the equations resulting from fitting using a first-order curve, we use *main-effect analysis.***

$$\delta_{Energy} \cong [f(\vec{x}_0, V_{dd-nom}, T) - f(\vec{x}_0, V_{dd-nom}, T_{nom})]$$
$$+ [f(\vec{x}_0, V_{dd}, T_{nom}) - f(\vec{x}_0, V_{dd-nom}, T_{nom})]$$
$$+ [f(\vec{x}, V_{dd-nom}, T_{nom}) - f(\vec{x}_0, V_{dd-nom}, T_{nom})]$$

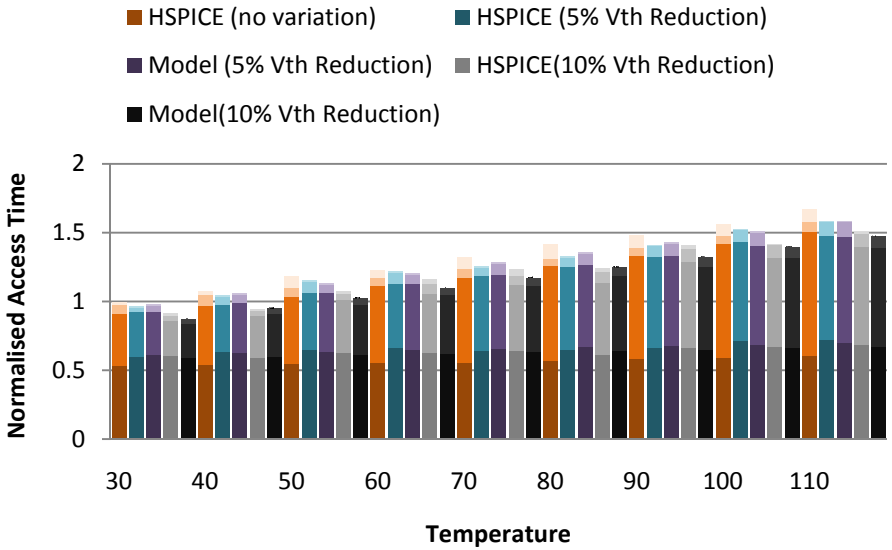
Error between Simulation & Model

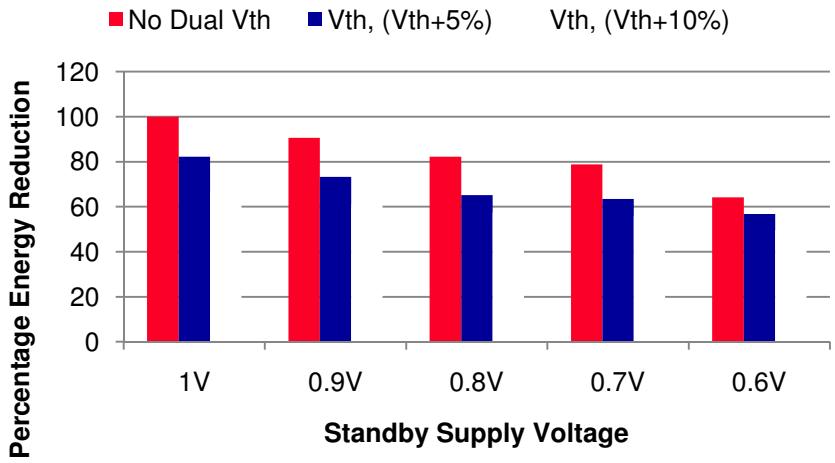




- **The computed error is independent of temperature.**
- **Model performance is degraded by structures driving large loads (address decoder, sense amplifier).**
- **At higher temperatures, access time failures are high.**
- **Higher order splines can be used to eliminate the non-linearity observed in 0.7V range.**

Usability in Circuit Optimizations





- V_{th} variability was exploited to assign dual-V_{th}.
- Lower threshold was assigned to delay critical paths.
- Delay was reduced by nearly 18% at high temperatures with minimal increase in leakage.
- Similarly, standby supply voltage of unused array sub-blocks was reduced with simultaneous dual-threshold assignment yielding energy reduction of around 50% for a single access.