Sizing and Placement of Charge Recycling Transistors in MTCMOS Circuits

Ehsan Pakbaznia
Farzan Fallah*
Massoud Pedram

University of Southern California
* Fujitsu Laboratories of America

Outline

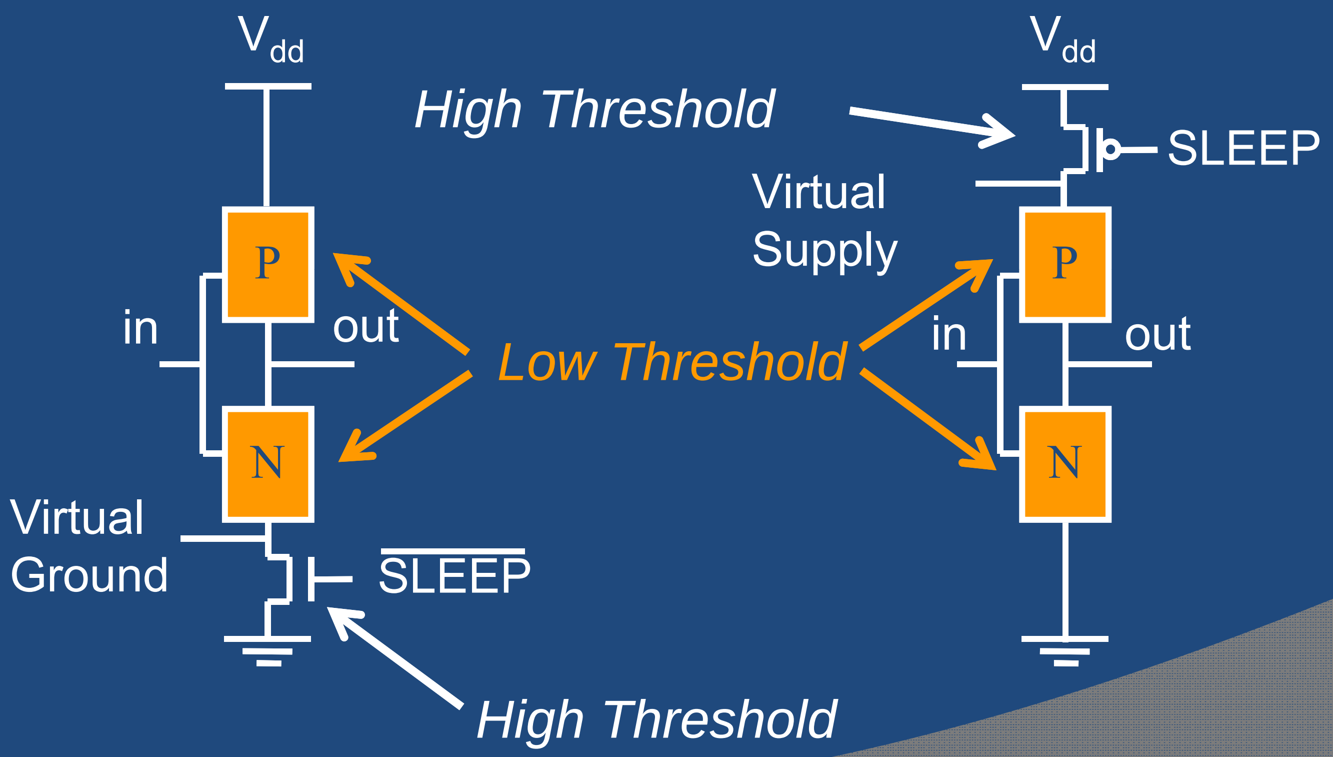
- Introduction
- Charge Recycling (CR) for Multi-Threshold CMOS (MTCMOS) Circuits
- Row-Based Layout Style for CR-MTCMOS
- Sizing and Placement of CR Transistors
- Experimental Results
- Conclusion

Leakage in CMOS Technology

- V_{dd} is reduced with CMOS technology scaling
- V_{th} must be lowered to recover the transistor switching speed
- The subthreshold leakage current increases exponentially with decreasing V_{th}
- A highly effective leakage control mechanism has proven to be the MTCMOS technique

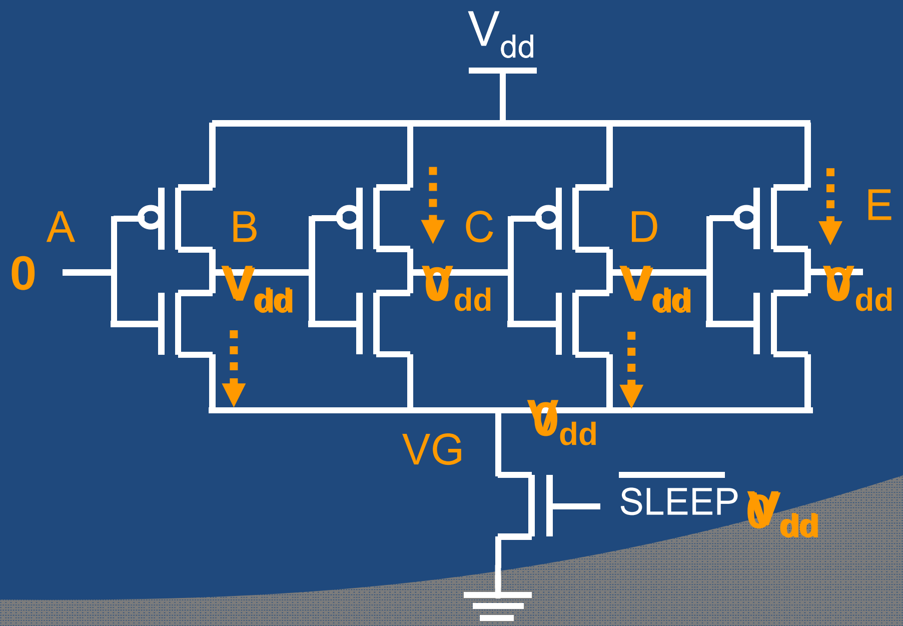
Overview of MTCMOS

- A high-V_{th} transistor is used to disconnect low-V_{th} transistors from the ground or the supply rails



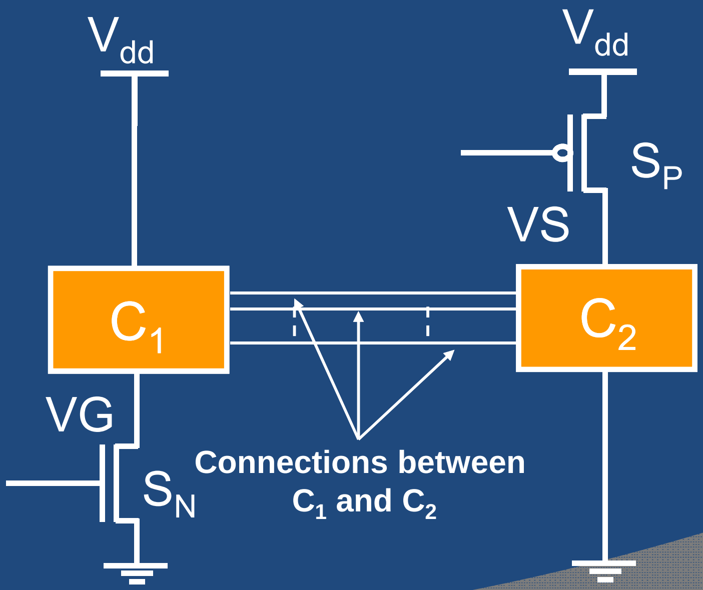
Some Drawbacks of MTCMOS

- State of internal nodes is corrupted, that is, with a footer sleep transistor, all internal nodes and the virtual ground (VGND) are charged up to a level near V_{dd}
- Energy is wasted when switching from the Sleep mode to the Active mode or vice versa
  - This means energy cannot be saved by the MTCMOS technique unless the sleep time is sufficiently long

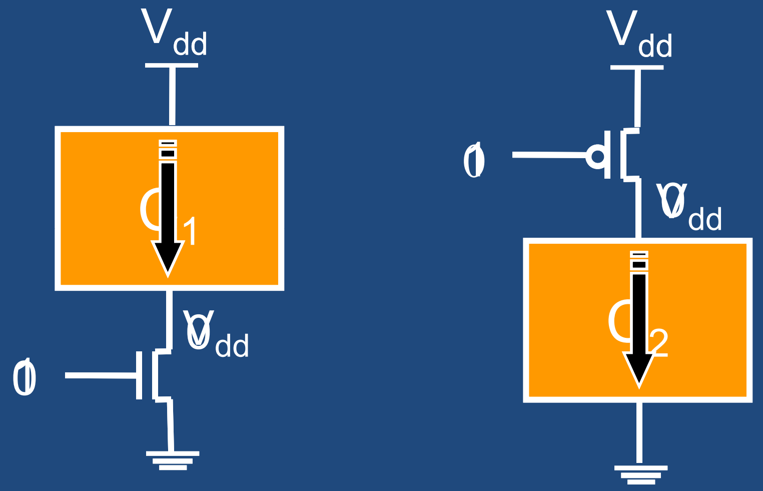


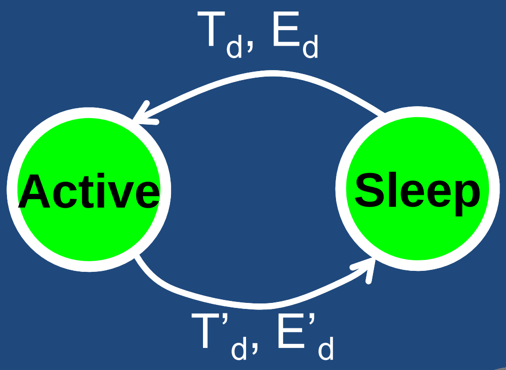
Charge Recycling (CR) MTCMOS

- The charge recycling technique uses both nMOS and pMOS sleep transistors
- Circuit C is divided into two sub-circuits:
  - Sub-circuit C_1 is connected to the nMOS sleep transistor, S_N
  - Sub-circuit C_2 is connected to the pMOS sleep transistor, S_P



Mode Transition in MTCMOS





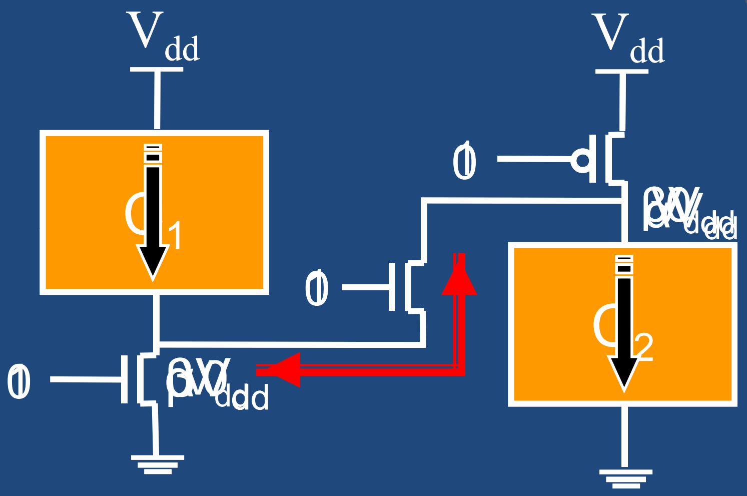
CR-MTCMOS

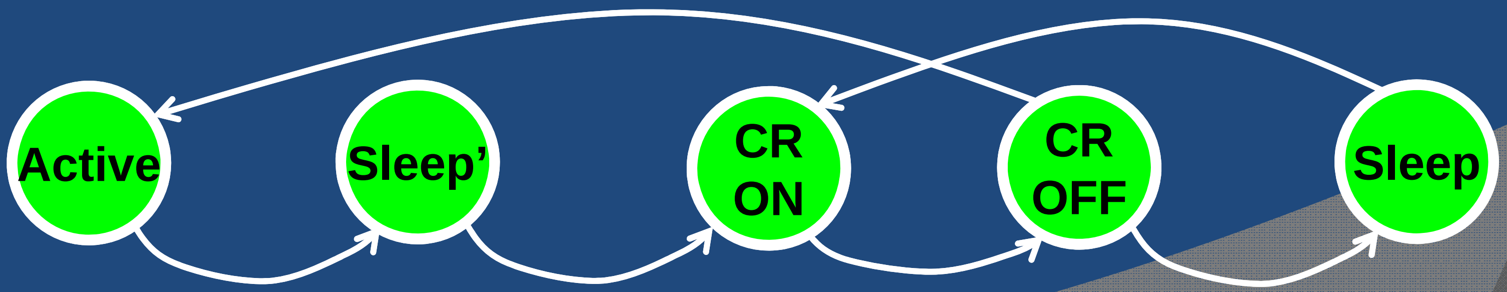
Energy Saving Ratio:

$$ESR(X) = \frac{E_{conv.} - E_{CR}}{E_{conv.}} = \frac{2X}{(1+X)^2}$$

$$ESR_{\max} = ESR(X=1) = 50\%$$

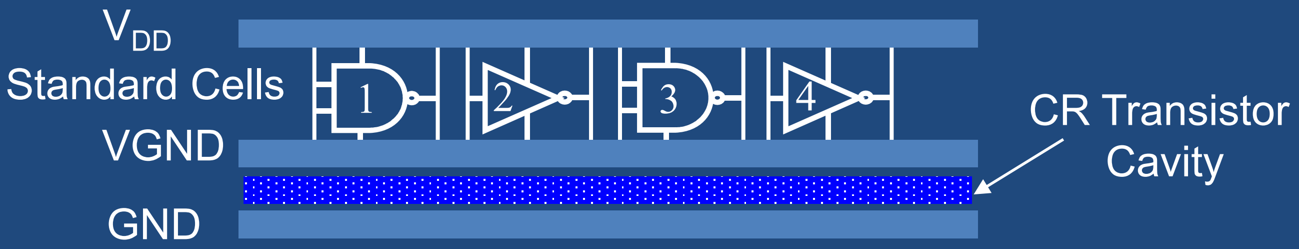
X : ratio of the VGND to VV$_{DD}$ capacitances



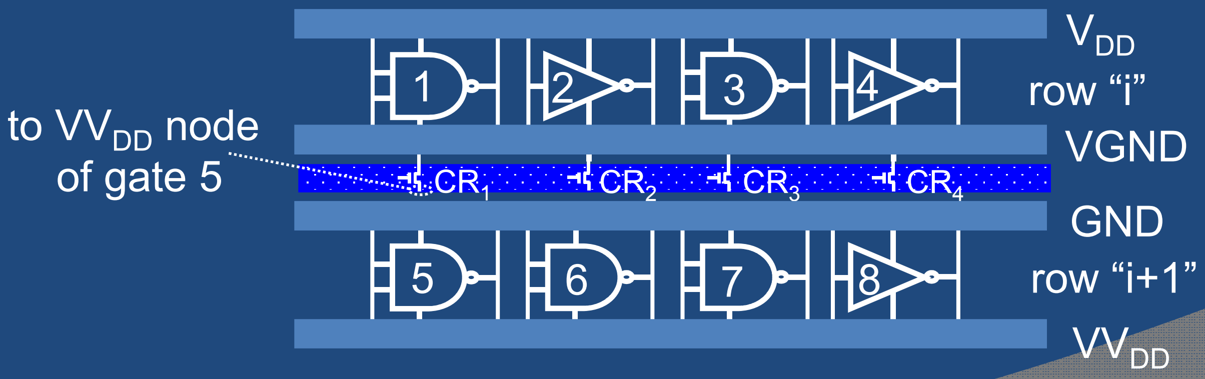


Row-Based Layout for CR-MTCMOS

- Layout style for a single cell row:



- Two adjacent rows use different types of sleep transistors e.g., nMOS for row *i*, and pMOS for row *i*+1



Problem Statement

- Objective:
  - Maximizing the ESR = Minimizing the CR overhead
- Constraint:
  - Maximum wakeup-time increase is limited to γ%
- Decision variables:
  - Widths of the CR transistors (which may also be set to zero)

$$\boxed{\begin{cases} Min\left(E_{CR-overhead}\right) \\ s.t. \\ \qquad t_w^{CR} \le \left(1+\gamma\right) \times t_w \end{cases}}$$

- $t_w{}^{CR}$: wakeup time of the CR-MTCMOS circuit
- γ: percentage increase in the wakeup time
- t_w: wakeup time of the original MTCMOS circuit

Power Overhead for CR-MTCMOS

- Dynamic power overhead
  - Due to switching ON and OFF the CR transistors
- Static power overhead
  - Due to extra sneak leakage path in CR-MTCMOS [Pakbaznia-DAC07]
- Total power dissipation overhead (dynamic + static):

$$\boxed{P_{CR-overhead} = \sum_{i=1}^{M} C_{g_i} f V_{DD}^2 + \sum_{i=1}^{M} I_{leak_i} V_{DD}}$$

  - *M* : CR transistor count in the row under consideration
  - *f* : mode transition frequency
  - $C_{g,i}$: input gate capacitance of the ith CR transistor
  - $I_{leak,i}$: sub-threshold leakage current of the ith CR transistor

Power Overhead (cont'd)

- It can be shown that the power dissipation overhead is proportional to the total width of CR transistors:

$$P_{CR-overhead} = \kappa \sum_{i=1}^{M} W_i$$

where κ is a constant coefficient which is calculated as:

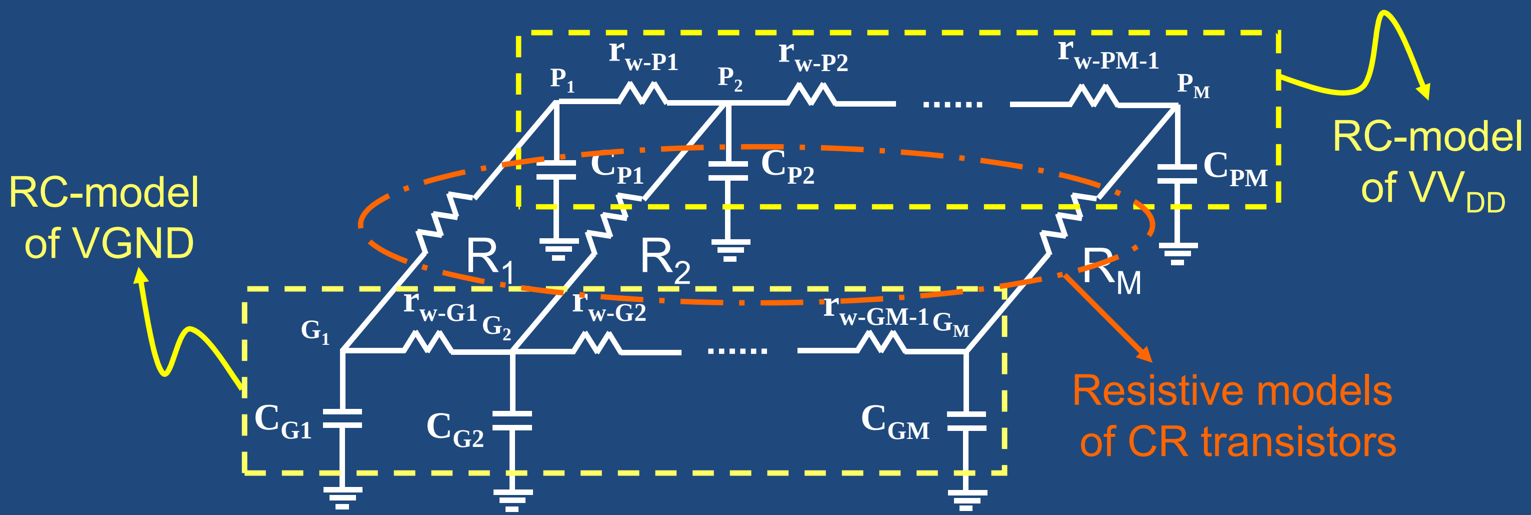
$$\kappa = L\, C_{ox}\, f\, V_{DD}^2 + \frac{\mu_0 \varepsilon_{ox}}{L t_{ox}} V_{DD}\, v_T^2\, e^{1.8} \exp\left(\frac{-V_{th}}{S\, v_T} \right)$$

- The new objective function is thus:

$$\boxed{Min\left(\sum_{i=1}^{M} W_i \right)}$$

RC Model for Charge Recycling Operation

- VGND and VV$_{DD}$ are replaced by equivalent RC models
- CR transistors are modeled as linear resistors



RC-model of VGND

RC-model of VV$_{DD}$

Resistive models of CR transistors

R_i : ON drain-source resistance of the i$^{\mathrm{th}}$ CR transistor, $\mathrm{R_i} = \eta/W_i$

C_{G_i}, C_{P_i} : Diffusion + interconnect capacitances at $\mathrm{G_i}$ and $\mathrm{P_i}$

$\mathrm{r_{w\text{-}G_i}}$: Wiring resistance between $\mathrm{G_i}$ and $\mathrm{G_{i+1}}$

r_{w-P_i} : Wiring resistance between $\mathrm{P_i}$ and $\mathrm{P_{i+1}}$

Wakeup Time Constraints

- The original wakeup-time constraint can be written as M separate constraints, one for each G_i node:

$$t_{w_i}^{CR} \le (1+\gamma) \times t_w \qquad \forall\ 1 \le i \le M$$

- $t_{w,i}^{CR}$ is the summation of two terms:
  - Charge-recycling delay
  - Delay due to discharging the remaining charge in the VGND rail

$$t_{w_i}^{CR} = d_i^{CR} + t_i^{rem} \qquad \forall\ 1 \le i \le M$$

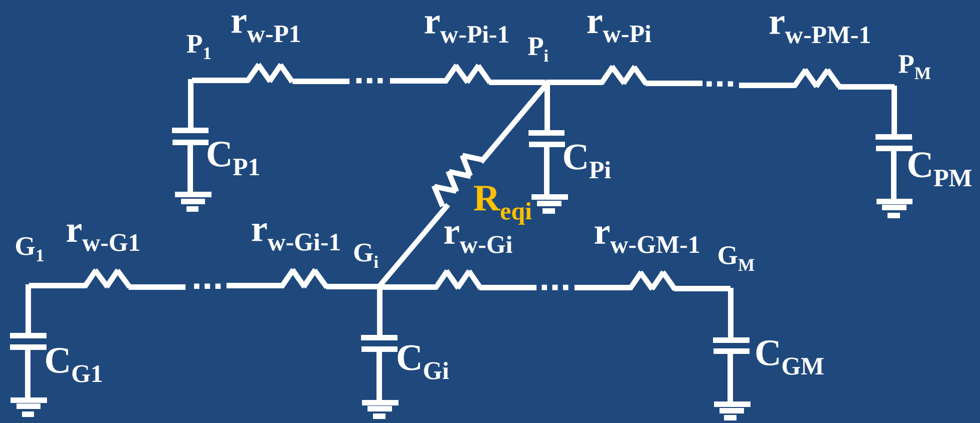
- New set of equivalent constraints:

$$\boxed{d_i^{CR} \le (1+\gamma) \times t_w - t_i^{rem} \qquad \forall\ 1 \le i \le M}$$

- t_w and t_i^{rem} are easily obtained from Hspice simulations
- d_i^{CR} must be calculated

Simplified RC Model

- A single equivalent transistor with width $W_{eq,i}$ (and resistance $R_{eq,i}$) is defined for each G_i ; P_i pair:

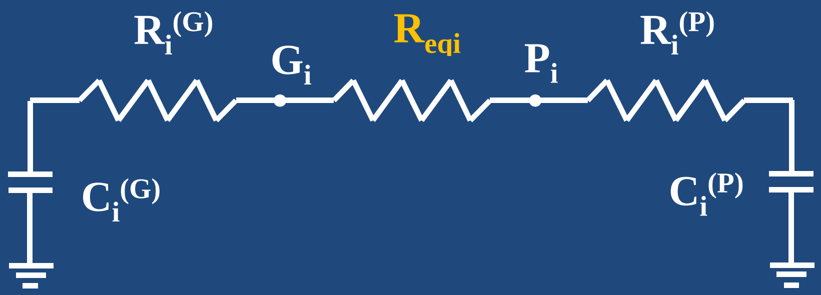


where

$$\boxed{R_{eqi} = \frac{\eta}{W_{eq_i}} = \frac{\eta}{\sum_{j=1}^{M}\left(1-\alpha\left|i-j\right|\right)W_j} \qquad 1\le i, j\le M}$$

$$\eta = \frac{L}{\mu C_{ox}\left(V_{DD}-V_{th}\right)} \qquad \text{and} \qquad \alpha = \frac{\sum_{i=1}^{M}\left(r_{w-G_i}+r_{w-P_i}\right)}{\sum_{i=1}^{M}R_i} << 1 \qquad \forall\,\text{each pass}$$

Replacing Virtual Rails with Their Effective RC Models



- $R_i^{(G)}$, $C_i^{(G)}$: RC-lumped model of the VGND rail at G_i
- $R_i^{(P)}$, $C_i^{(P)}$: RC-lumped model of the VV$_{DD}$ rail at P_i
  - For example, for G_i we have:

$$\boxed{C_i^{(G)} = Y_{G,1i} \text{ and } R_i^{(G)} = -\frac{Y_{G,2i}}{Y_{G,1i}^2}}$$

  - $Y_{G,1i}$ and $Y_{G,2i}$ are the first and second moments of the total admittance at G_i which may be recursively calculated as in [Kahng-VLSI Design99]

Charge-Recycling Delay

- The 0-δ% CR delay for node G_i is:

$$d_i^{CR} = \frac{1}{\ln(\delta)} \times \frac{\left(R_i^{(G)} + R_{eq_i} + R_i^{(P)}\right) C_i^{(G)} C_i^{(G)}}{\left(C_i^{(G)} + C_i^{(P)}\right)}$$

- Recall that $d_i^{CR} \leq (1+\gamma) \times t_w - t_{rem_i}$
- The set of the constraints can thus be re-written as:

$$\boxed{\sum_{j=1}^{M} b_{ij} W_j \geq \eta \left[\left[(1+\gamma) t_w - t_{rem_i} \right] \ln(\delta) \frac{\left(C_i^{(G)} + C_i^{(P)}\right)}{C_i^{(G)} C_i^{(P)}} - R_i^{(G)} - R_i^{(P)} \right]^{-1} \qquad 1 \leq i \leq M}$$

where: $\boxed{b_{ij} = 1 - \alpha \left| i - j \right|}$

Modified Problem Statement

$$\text{Minimize}\left(\sum_{i=1}^{M} W_i\right)$$

$$\text{s.t.:}\quad \sum_{j=1}^{M}\left(1-\alpha\left|i-j\right|\right)W_j \ge \eta\left[\left[\left(1+\gamma\right)t_w - t_{rem_i}\right]\ln\left(\delta\right)\frac{\left(C_i^{(G)}+C_i^{(P)}\right)}{C_i^{(G)}C_i^{(P)}} - R_i^{(G)} - R_i^{(P)}\right]^{-1}, \ \forall i \quad 1\le i\le M$$

$$W_i \ge 0 \quad , \quad \forall i \quad 1\le i\le M$$

- This is a linear Programming (LP) problem, which can be solved optimally in polynomial time

CR Transistors Placement

- The sizing problem is solved assuming there is one CR transistor between each G_i ; P_i pair
- CR transistors that have a width less than W_{min} will be removed; this is called the rounding step (W_{min} is the minimum acceptable transistor width)
- The sizing problem will be solved again for the remaining CR transistors
- Sizing + rounding operations will be repeated until the improvement in the total CR transistor width is negligible

Simulation Approach

- The proposed approach was compared with two other approaches
  - Single CR-MTCMOS: one CR transistor placed at the leftmost corner of each row
  - Uniform CR-MTCMOS: 3 uniformly-distributed CR transistors placed on each row
- CR transistor(s) in both approaches are sized such that the maximum wakeup delay increase is γ%

Experimental Results in 90nm Technology for ISCAS Benchmarks

Circuit	# of cells	# of rows	Total sleep tx width (λ)	Total CR TX width (λ)			Total CR TX width comparison (%)	
				SCR	UCR	DCR	DCR vs. SCR	DCR vs. UCR
9Sym	276	4	7152	1667	833	417	75	50
C432	204	2	4600	625	382	208	67	45
C880	432	6	9936	2326	1458	625	73	57
C1355	526	6	11320	2118	1597	625	71	61
C3540	1295	10	30656	6458	4792	1875	71	61
C5315	1727	10	38992	11042	6458	2292	79	65
average	-	-	-	4039	2587	1007	75	61

- MT = MTCMOS
- SCR = Single CR-MTCMOS
- UCR = Uniform CR-MTCMOS
- DCR = Distributed CR-MTCMOS (proposed)

Experimental Results (cont'd)

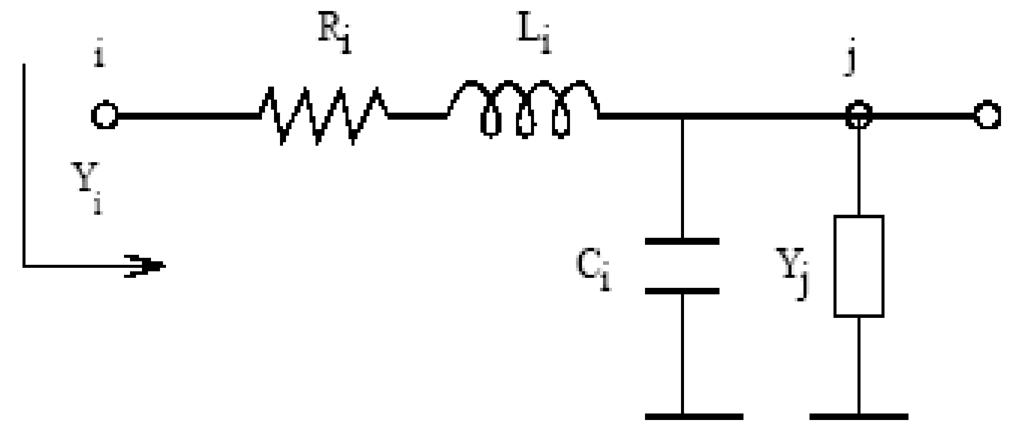
Circuit	# of cells	# of rows	Total sleep tx width	Switching energy in one complete active-sleep cycle (pJ)				DCR ESR (%)	ESR comparison (%)	
				MT	SCR	UCR	DCR		DCR vs. SCR	DCR vs. UCR
9Sym	276	4	7152	14.4	12	9.6	8.4	42	25	8
C432	204	2	4600	9.6	6.6	5.9	5.4	44	13	5
C880	432	6	9936	20.4	16.9	14.4	12	41	24	12
C1355	526	6	11320	25.2	18.7	17.2	14.4	43	17	11
C3540	1295	10	30656	90	63.6	58.8	50.4	44	15	9
C5315	1727	10	38992	147.6	105.6	92.4	80.4	46	17	8
average	-	-	-	51.1	37.2	33	28.4	44.4	18.5	8.8

- MT = MTCMOS
- SCR = Single CR-MTCMOS
- UCR = Uniform CR-MTCMOS
- DCR = Distributed CR-MTCMOS (proposed)

Conclusion

- CR-MTCMOS is the only known method for reducing energy consumed during transitions between Sleep and Active modes
- The placement and sizing problem of CR transistors can be formulated and solved as an LP problem
- The proposed concurrent sizing and placement technique allows us to employ CR-MTCMOS for row-based designs
- The technique achieves nearly the full potential of CR-MTCMOS in terms of saving switching energy during mode transitions (ideal 50%, in practice 44%)

Backup Slide: Recursive Admittance Calculation



$$
\begin{aligned}
Y_{1,i} &= Y_{1,j} + C_i \\
Y_{k,i} &= Y_{k,j} - R_i \sum_{l=1}^{k-1} Y_{l,i} Y_{k-l,j} - L_i \sum_{l=1}^{k-2} Y_{l,i} Y_{k-1-l,j} - R_i C_i Y_{k-1,i} \\
&\quad - L_i C_i Y_{k-2,i} \quad \text{for } k \geq 2
\end{aligned}
$$