

# SLEEP TRANSISTOR SIZING USING TIMING CRITICALITY AND TEMPORAL CURRENTS\*

**Anand Ramalingam<sup>1</sup>, Bin Zhang<sup>1</sup>, Anirudh Devgan<sup>2</sup>, and David Z. Pan<sup>1</sup>**

<sup>1</sup>Department of Electrical and Computer Engineering,  
The University of Texas, Austin, TX 78712

{anandram, bzhang, dpan}@ece.utexas.edu

<sup>2</sup>Austin Research Laboratory, IBM Research Division, Austin, TX 78758

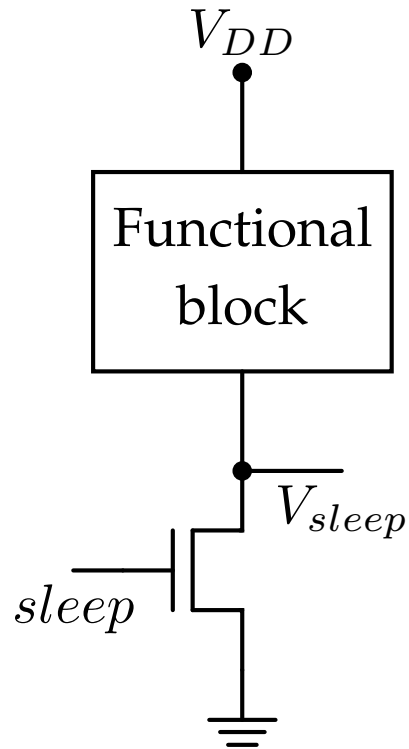
devgan@us.ibm.com

\* This work is partially sponsored by IBM Faculty Award. We used computers donated by Intel Corporation.

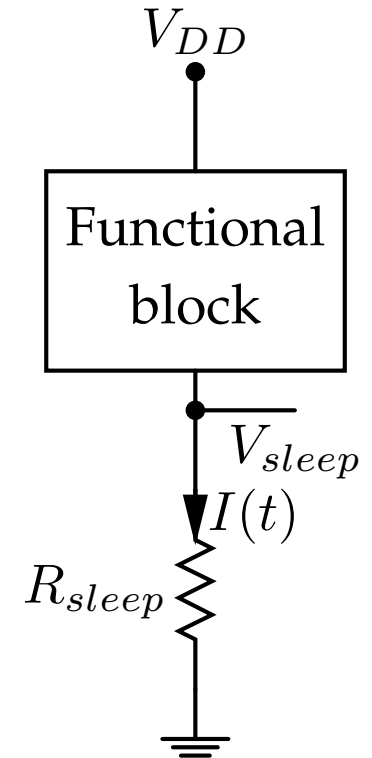
## MOTIVATION

- As technology scales, *dynamic power increases*
  - To reduce dynamic power, scale down  $V_{DD}$
- Scaling down  $V_{DD}$  results in *performance loss*
  - To maintain performance, scale down  $V_T$
- Scaling down  $V_T$  results in *exponential increase* of subthreshold current
  - Subthreshold leakage accounts for 42% of total power in 90nm technology [Kao et al, ICCAD 2002]
- **Power gating** is a technique used to reduce the subthreshold leakage
  - The circuit is gated from the power supply by **sleep transistor**
  - A major challenge in power gating is *sizing* the sleep transistor

## SLEEP TRANSISTOR



- Sleep transistors are *high*  $V_T$  transistors
  - Switched off when the circuit is *idle*
- Sleep transistor can be modeled as resistor
  - Results in reduction of gate overdrive to  $V_{GS} - V_{sleep}$
  - This results in *performance loss* when the circuit is *active*



**PROBLEM** Given a *combinational* circuit and a specified *delay penalty*, find the size of the sleep transistor for the entire circuit efficiently

## LITERATURE REVIEW

1. Module based design [Kao et al, DAC 1998]
  - Single sleep transistor to gate the entire circuit
2. Cluster based design [Anis et al, DAC 2002]
  - Circuit is divided into different clusters to *minimize*  $I_{peak}$
  - Each cluster has an individual sleep transistor
3. Distributed sleep transistor network [Long-He, DAC 2003]
  - The sleep transistors in the cluster based design are connected
  - The current is shared by the network reducing the sleep size
  - All the above methodologies base their sizing on  $I_{peak}$

**OBSERVATION** The sizing can be *improved* if we can estimate the switching current  $I(t)$  efficiently

## EXPECTED CURRENT OF A GATE $I_{exp}$

- The  $I_{exp}$  of the gates are needed to estimate the switching current  $I(t)$

FIND-EXPECTED-CURRENT(*gate*)

- 1 Find  $I_{peak}$  for each *gate* in the library using HSPICE
  - ▷  $\alpha_s$  is the switching factor
- 2  $I_{exp} = \alpha_s \times I_{peak}$
- 3 **return**  $I_{exp}$

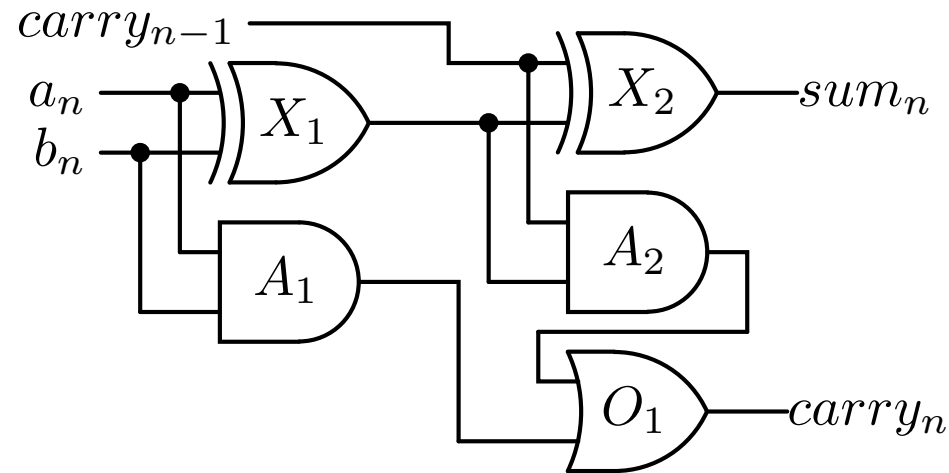
- The  $I_{exp}$  of NOR2 is illustrated
- The switching factor  $\alpha_s = P\{Y = 1 \rightarrow 0 | Y = 1\} \times P\{Y = 1\}$ 
  - For NOR2,  $\alpha_s = \frac{1}{4} \times \frac{3}{4} = \frac{3}{16}$
- The expected current  $I_{exp} = \alpha_s \times I_{peak}$ 
  - For NOR2,  $I_{exp} = \alpha_s \times I_{peak} = \frac{3}{16} \times 0.72ma = 0.12ma$

## PSEUDOCODE FOR ESTIMATING THE SWITCHING CURRENT $I(t)$

ESTIMATE-SWITCHING-CURRENT(*circuit*)

- 1 Run *PrimeTime* on the *circuit* to get timing windows
- 2  $I(t) \leftarrow 0$
- 3 **for every** *gate* in the *circuit*
- 4     **do**  $I_{exp} \leftarrow \text{GET-EXPECTED-CURRENT}(gate)$
- 5          $I_{gate}(t) \leftarrow$  timing windows bounded by  $I_{exp}$
- 6          $I(t) \leftarrow I(t) + I_{gate}(t)$
- 7 **return**  $I(t)$

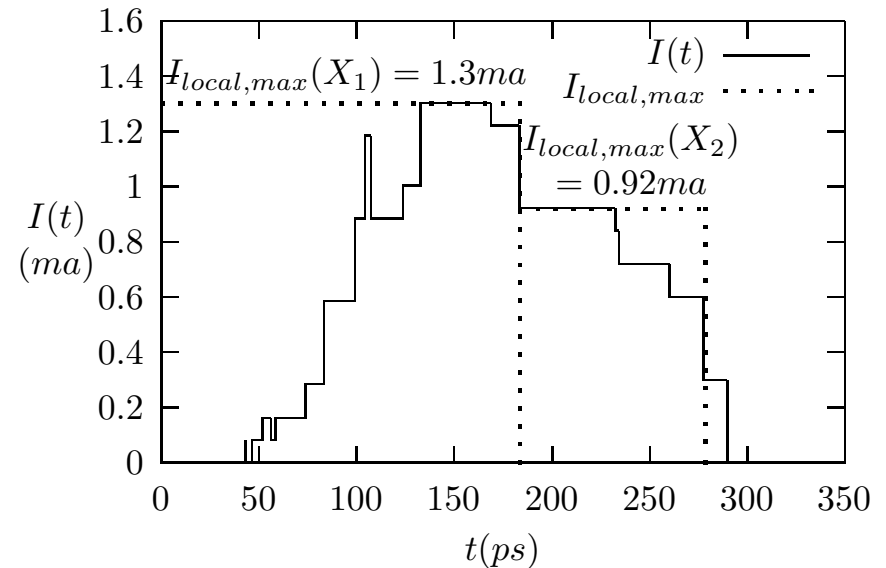
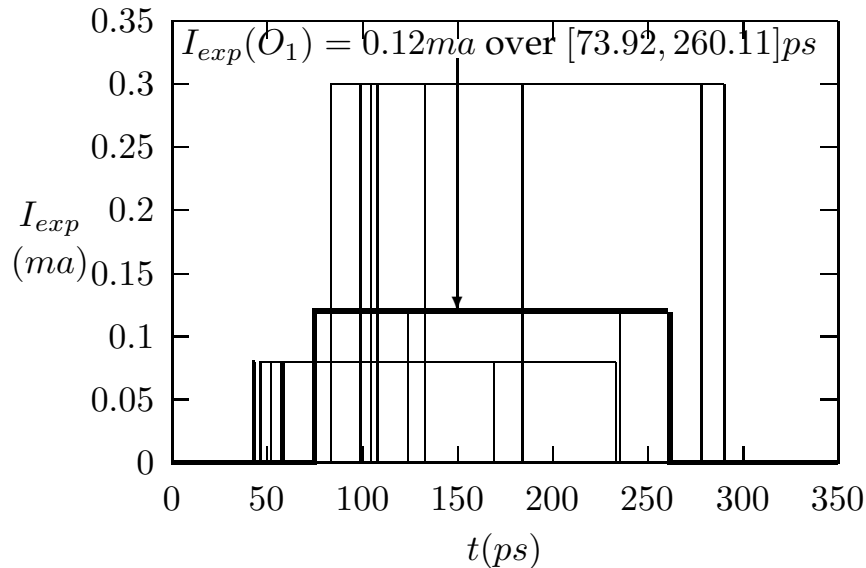
## TIMING WINDOWS OF A 1-BIT CLA



- Timing windows after running *PrimeTime* on 1-bit CLA (time in *ps*)

| Gate  | $rise_{min}$ | $rise_{max}$ | $fall_{min}$ | $fall_{max}$ |
|-------|--------------|--------------|--------------|--------------|
| $X_1$ | 98.90        | 107.52       | 104.24       | 183.28       |
| $A_1$ | 51.52        | 56.33        | 42.82        | 43.01        |
| $X_2$ | 83.22        | 290.10       | 132.75       | 277.42       |
| $A_2$ | 58.31        | 168.80       | 46.43        | 232.62       |
| $O_1$ | 123.60       | 234.09       | 73.92        | 260.11       |

## SWITCHING CURRENT OF 1-BIT CLA



- On the *left*,  $I_{exp}$  bounds the timing windows of each gate
- On the *right*, the estimated current discharge  $I(t)$ 
  - Got by summing up all the currents on the *left*
  - Also shown is the *local maximum* current of the gates  $X_1$  and  $X_2$

## TIMING CRITICALITY BASED SIZING

- Sleep transistor sizing based on paths has two *drawbacks*
  - Worst case paths are *not* the same in CMOS and MTCMOS
  - The number of paths is *exponential*
- **HEURISTIC** Take top  $K$  paths to size the sleep transistor

SIZE-SLEEP-TRANSISTOR(*circuit*, *penalty*,  $K$ )

- 1 Run *PrimeTime* on the *circuit* to get critical paths
- 2  $R_{sleep} \leftarrow \infty$
- 3 **for**  $path \leftarrow 1$  **to**  $K$   $\triangleright$  Size using top  $K$  critical paths
- 4     **do**  $R_{path} \leftarrow$  Max resistance tolerated by a *path* for given a *penalty*
- 5          $R_{sleep} \leftarrow \text{MIN}(R_{sleep}, R_{path})$
- 6      $(\frac{W}{L})_{sleep} \leftarrow$  Size using  $R_{sleep}$
- 7 **return**  $(\frac{W}{L})_{sleep}$

## PATH BASED SIZING IN A 1-BIT CLA

- Consider a path through the gates  $X_1$  and  $X_2$
- The gate and path delays after running *PrimeTime*

| Gate          | $\tau_d(ps)$ | $\tau_d^{path}(ps)$ | fall/rise |
|---------------|--------------|---------------------|-----------|
| $X_1$         | 183.28       | 183.28              | fall      |
| $X_2$         | 94.13        | 277.42              | fall      |
| $t_{arrival}$ |              | 277.42              |           |

- $R_{path}$  is the maximum resistance tolerated by a *path* for a given *penalty*
  - Let the penalty be 5% of the delay ( $0.05 \times t_{arrival}$ )

$$\begin{aligned}
 R_{path} &= \frac{(V_{DD} - V_{TL}) \tau_{penalty}^{path}}{\sum_{gate \in path} I_{local,max} \tau_d} \\
 &= \frac{(3.3 - 0.7) (0.05 \times 277.42ps)}{183.28ps \times 1.3ma + 94.13ps \times 0.92ma} \\
 &= 111\Omega
 \end{aligned}$$

## PATH BASED SIZING IN A 1-BIT CLA

- $R_{path}$  is calculated for top  $K$  paths
  - The *minimum* of  $R_{path}$  is the resistance of the sleep transistor  $R_{sleep}$
- For the purpose of illustration, let  $R_{sleep} = 111\Omega$

$$\begin{aligned}
 \left(\frac{W}{L}\right)_{sleep} &= \frac{1}{\mu_n C_{ox} (V_{DD} - V_{TL}) R_{sleep}} \\
 &= \frac{1}{1.25 \times 10^{-4} (3.3 - 0.7) 111} \\
 &= 27.77\lambda
 \end{aligned}$$

- Let  $L_{sleep} = 2\lambda$ , where  $\lambda = 0.1\mu m$

$$W_{sleep} = 55.55\lambda \approx 56\lambda$$

**MODULE BASED DESIGN**

| Circuit        | Module ( $\lambda$ ) |          |
|----------------|----------------------|----------|
|                | Kao                  | Proposed |
| CLA4           | 825                  | 125      |
| Parity checker | 960                  | 235      |
| Wallace tree   | 1365                 | 427      |
| c432           | 3438                 | 475      |
| c499           | 3840                 | 1171     |

- Comparison with [Kao et al, DAC 1998] based on  $W_{sleep}$ 
  - The unit of  $W_{sleep}$  is  $\lambda = 0.1\mu m$
  - The performance penalty is 5%
- The proposed approach yields an sleep area reduction of 80%

## CLUSTER BASED DESIGN

| Circuit        | Cluster ( $\lambda$ ) |          |
|----------------|-----------------------|----------|
|                | Anis                  | Proposed |
| CLA4           | 204                   | 127      |
| Parity checker | 369                   | 284      |
| Wallace tree   | 1201                  | 698      |
| c432           | 1272                  | 385      |
| c499           | 2094                  | 1351     |

- Comparison with [Anis et al, DAC 2002] based on  $W_{sleep}$ 
  - The unit of  $W_{sleep}$  is  $\lambda = 0.1\mu m$
  - The performance penalty is 5%
- The proposed approach yields an sleep area reduction of 49%

## COMPARISON OF PROPOSED METHODS

| Circuit | Proposed ( $\lambda$ ) |         |
|---------|------------------------|---------|
|         | Module                 | Cluster |
| c880    | 638                    | 509     |
| c1908   | 479                    | 457     |
| c3540   | 1979                   | 1933    |
| c7552   | 12955                  | 8325    |

- Comparison of our proposed module and cluster based designs
- For bigger circuits clustering wins comfortably
  - If the critical path is in one cluster, then we have slack to exploit in other clusters
  - This leads to smaller sizes in cluster based design

## CONCLUSIONS

- A new *path based* methodology to size sleep transistors using temporal currents and timing windows
  - Area reduction of sleep transistors by 80% and 49% compared to module based design and cluster based design respectively
  - Area reduction is mainly due to the usage of *local maximum* current instead of *global maximum* current
- An efficient method to estimate the temporal switching current  $I(t)$  of the circuit.