

# Minimum-Power Retiming for Dual-Supply CMOS Circuits

Farhana Sheikh  
University of California  
Berkeley, CA 94720

farhana@eecs.berkeley.edu

Andreas Kuehlmann  
Cadence Berkeley Labs  
Berkeley, CA 94704

kuehl@cadence.com

Kurt Keutzer  
University of California  
Berkeley, CA 94720

keutzer@eecs.berkeley.edu

## ABSTRACT

The use of dual-supply voltages at the gate level is an effective technique to limit dynamic power consumption while preserving performance. However, its use in commercial circuit designs is limited primarily due to lack of CAD tool support. Very little work has been carried out to leverage multiple supply voltages for timing, area, and power trade-offs during logic synthesis. This paper describes an extension to the retiming framework which is leveraged to synthesize low-power CMOS circuits using dual-supply voltages. A mathematical formulation of the problem is presented with the central objective to minimize dynamic power while maintaining the target clock period.

## Categories and Subject Descriptors

B.2.5 [Register-Transfer-Level Implementation]: Design Aids—*Automatic synthesis, optimization*; B.3.6 [Logic Design]: Design Aids—*Automatic synthesis, optimization*

## General Terms

Retiming Theory, Low-Power Design

## Keywords

Low-power, dual-supply, synthesis, retiming theory

## 1. INTRODUCTION

The exponential growth in the number of devices per chip combined with another increase in their operating frequencies results in a dramatic growth of overall power consumption. For example, extrapolating the trends in power dissipation for the Intel processor family shows that by the year 2008, a processor would consume 18kW of power [1]. As a result, power management increasingly becomes the primary design objective for enabling higher chip integration levels.

The use of multiple supply voltages at the gate level is an effective technique to limit dynamic power consumption while preserving the required performance [11]. However, thus far only few

custom designs employ dual supplies at the gate level. This is primarily due to the lack of CAD tool support. The synthesis problem for dual-supply (dual- $V_{DD}$ ) designs is to generate a gate level netlist such that the assignment of high or low voltage to each gate results in a power-optimal circuit that meets all given timing constraints.

Little work has been carried out to utilize multiple supply voltages for timing, area, and power trade-offs during logic synthesis. As far as the authors are aware, only one paper has been published that introduces dual-supply voltages at the gate level during logic synthesis [9]. A simple heuristic is presented that assigns supply voltages to gates in combinational circuits. Another publication [10] describes a methodology based on gate sizing that is used to exploit dual-supply voltages post logic synthesis.

In this paper we show how an extension to the standard retiming framework can be used to formulate the synthesis problem for dual- $V_{DD}$  circuits as an integer linear program (ILP). The next section presents an introduction to multiple supply voltages. In Section 3 we briefly review the classical retiming formulation. Section 4 presents a new retiming formulation that minimizes the power consumption of a dual-supply circuit implementation for a given constraint for the clock period. In Section 5 an example is provided and Section 6 discusses conclusions and future work.

## 2. CIRCUIT DESIGN WITH MULTIPLE SUPPLY VOLTAGES

Power consumption in CMOS circuits is currently dominated by dynamic switching power which decreases quadratically by lowering the supply voltage as follows:

$$P = \alpha \cdot C_{load} \cdot f_{clk} \cdot V_{DD}^2 \quad (1)$$

Here,  $\alpha$ ,  $C_{load}$ , and  $f_{clk}$  denote the switching activity, load capacitance, and clock frequency, respectively [13]. However, lowering  $V_{DD}$  also increases the individual gate delays by:

$$t_d \sim \frac{C_{load} \cdot V_{DD}}{2} \left[ \frac{1}{K_n \cdot (V_{DD} - V_{Tn})^2} + \frac{1}{K_p \cdot (V_{DD} - |V_{Tp}|)^2} \right]$$

where all technology and gate topology parameters are lumped into constants  $K_n$  and  $K_p$ . Further,  $V_{Tn}$  and  $V_{Tp}$  denote the transistor threshold voltages [7].

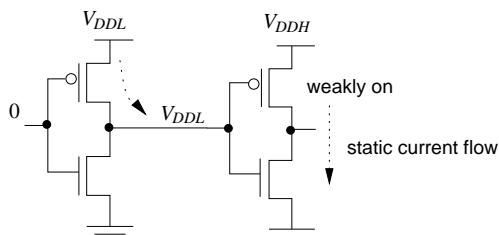
The increased delay results in a performance degradation only if the supply voltage is reduced for gates on the critical path. One option for compensating for this performance loss is to adjust the threshold voltage  $V_T$  of these gates [4]: however, decreasing  $V_T$  results in increased standby leakage and thus larger static power consumption. Another proposed compensation approach is based on parallel or pipelined architectures [2] but causes large area penalties.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

TAU'02 December 2–3, 2002, Monterey, California, USA.  
Copyright 2002 ACM 1-58113-526-2/02/0012 ...\$5.00.

A multi- $V_{DD}$  approach is presented in [11] whereby gates off the critical path are allowed to operate at low  $V_{DD}$  ( $V_{DDL}$ ) and gates on the critical path operate at high  $V_{DD}$  ( $V_{DDH}$ ). This methodology allows a significant power reduction without compromising the performance of the circuit. The selective adjustment of  $V_{DD}$  can be used at different levels of granularity. Applied at the block level, the supply voltage can only be reduced if the entire block does not contain any critical paths. In contrast, when applied at the gate level [12], the power saving potential is significantly larger because the supply for all non-critical gates can be reduced.

When using dual  $V_{DD}$  special attention must be paid at the boundaries of gates with different supply voltages.  $V_{DDH}$  gates can safely drive  $V_{DDL}$  gates. However,  $V_{DDL}$  gates cannot directly feed  $V_{DDH}$  gates without using special level converters. As illustrated in Figure 1 [12], the pull-up device of a high-voltage gate that is driven by a low-voltage gate will not completely shut off for a logical-1 input. Therefore, special circuit structures similar to sense amplifiers used in memories, must be inserted at supply boundaries from  $V_{DDL}$  to  $V_{DDH}$ . Such level converters consume additional power and area and add to the overall delay. A realistic model for power minimization must take this into account.



**Figure 1: Static leakage current for direct connections of  $V_{DDL}$  to  $V_{DDH}$  gates.**

Generally, there are two types of level converters [12, 13]. The asynchronous type simply adjusts the voltage level from  $V_{DDL}$  to  $V_{DDH}$  whereas synchronous converters are also clocked by the system clock and thus combine the function of a flip-flop with the actual level conversion (LCFF). The exclusive application of synchronous level converters is typically referred to as Clustered Voltage Scaling (CVS) [11]. Extended Clustered Voltage Scaling (ECVS) denotes its extension where a limited number of asynchronous converters are added. Asynchronous level converters consume more area and are inherently less stable than their synchronous counterpart. Furthermore, they are more sensitive to delay changes due to voltage variations. As a result circuit designers prefer using CVS. Our paper takes this into account by focusing on a retiming formulation for CVS, however, an extension to ECVS is possible.

It should be noted that if ECVS is employed, there is more freedom in assigning  $V_{DDH}$  or  $V_{DDL}$  to gates as asynchronous level converters maybe inserted anywhere in the circuit and are not limited to synchronous boundaries. This may result in significant overhead due to possible proliferation of level converters if no limit is imposed on their number.

Table 1 compares the parameters for a selected set of combinational gates, standard flip-flops designed for different supply levels, and flip-flops which are combined with synchronous level converters [3]. The data presented in Table 1 is based on a generic 130nm technology process. As shown, there are significant differences in the power and delay characteristics between gates at high and low supply voltage. In addition to modeling the power used by the flip-flops themselves, a correct formulation for power optimization

Gate Type	Setup [ps]	Hold [ps]	Clk-Q [ps]	Delay [ps]	Energy [fJ]
FF $V_{DDH}$	19.6	21.5	126.5	152	15.84
FF $V_{DDL}$	37.4	35.5	192.7	240	7.28
LCFF	61.1	63.9	214.8	287	9.13
INV $V_{DDH}$	-	-	-	36.8	0.91
INV $V_{DDL}$	-	-	-	54.7	0.41
NAND2 $V_{DDH}$	-	-	-	51.2	1.15
NAND2 $V_{DDL}$	-	-	-	79.7	0.53
NOR2 $V_{DDH}$	-	-	-	62.8	1.22
NOR2 $V_{DDL}$	-	-	-	95.1	0.57

**Table 1: Differences between flip-flop types and combinational gates at varying voltages in a generic 130nm technology.**

tion must also take into account the clocking circuitry. In practical designs, the clock distribution accounts for a large fraction of the overall power dissipation [13]. Consequently, for manual circuit optimization, designers attempt to assign as many flip-flops as possible (and thus their clocking circuitry) to  $V_{DDL}$ . In rare cases, it may be necessary to operate flip-flops at  $V_{DDH}$  where path delays violate the target clock period when all gates are assigned  $V_{DDH}$ . A corresponding scheme is described in [2] and [13].

Reported benchmark results [12, 2, 14, 13] demonstrate that a significant reduction in power consumption, ranging from 30 – 45%, can be achieved by dual-supply methods. In these results, the reported increase of area and average interconnection length is less than 10%.

### 3. CLASSICAL RETIMING

Classical retiming is a structural optimization technique that relocates the flip-flops in a circuit with the objective of minimizing their total count, maximizing the circuit performance, or achieving both goals simultaneously [5, 6, 8]. The original formulation presented by Leiserson and Saxe is based upon the following assumptions:

- The gate delays are modeled by non-negative constants and are assumed to be independent of their fanout.
- The circuit does not contain combinational loops.
- A single clocking scheme and edge-triggered flip-flops with identical skews are used. Later publications extend this to multi-phase clocking and level-sensitive flip-flops.
- Any reset state can be handled safely by the environment. Later publications provide more general solutions to the reset problem
- All flip-flop delays are assumed to be identical and handled by a corresponding adjustment of the target clock period.

#### 3.1 Min-area Retiming

Let  $C = (V, E)$  denote a circuit graph where  $V$  corresponds to the set of gates including a single vertex, the *host*, representing all primary inputs and outputs.  $E \subseteq V \times V$  is a set of edges connecting the gates and I/Os. Each edge  $(u, v) \in E$  is associated with a non-negative weight  $w(u, v) \in \mathbb{N}$  which is equal to the number of flip-flops along this edge. Furthermore,  $d(v)$  represents the delay of gate  $v \in V$ .

A retiming of  $C$  is defined as a gate labeling  $r : V \rightarrow \mathbb{N}$ , where  $r(u)$  is the *lag* of gate  $u$  denoting the number of flip-flops that are

moved backward through it. The new set of arc weights  $w_r$  of the retimed circuit  $C_r$  is computed as follows:

$$w_r(u, v) = w(u, v) + r(v) - r(u) \quad (3)$$

For a retiming to be valid the number of flip-flops at all edges must be non-negative, i.e.,  $w_r \geq 0$ . The min-area retiming objective is to minimize the total number of flip-flops, i.e.,  $\sum w_r \rightarrow \min$ . Using (3), this leads to the following optimization problem:

$$\sum_{v \in G} r(v) \cdot (|FI(v)| - |FO(v)|) \rightarrow \min \quad (4)$$

with the constraint

$$\forall (u, v) \in E : r(v) - r(u) \geq -w(u, v) \quad (5)$$

where  $FI(v)$  and  $FO(v)$  represent the set of fanin and fanout gates of gate  $v$ , respectively, and  $|S|$  denotes the cardinality of set  $S$ . For simplicity we omit in this and following formulations the special handling of shared flip-flops at multi-fanout gates. The sharing can be modeled by a simple extension of the circuit graph as described in detail in [6].

### 3.2 Clock-period-constrained Min-area Retiming

Individual circuit paths need to be analyzed to account for a target clock period during retiming. A circuit path  $p$  is defined as a sequence of connected vertices, i.e.,  $p = (v_1 \xrightarrow{e_1} v_2 \xrightarrow{e_2} \dots \xrightarrow{e_{n-1}} v_n); v_i \in V, e_i = (v_i, v_{i+1}) \in E$ . The weight  $w(p)$  and delay  $d(p)$  of a path  $p$  are defined as follows:

$$w(p) = \sum_{e_i \in p} w(e_i),$$

$$d(p) = \sum_{v_i \in p} d(v_i)$$

and used in the following definitions of the weight matrix  $W$  and delay matrix  $D$ :

$$W(u, v) = \min_p \{w(p) : u \xrightarrow{p} v\} \quad (6)$$

$$D(u, v) = \max_p \{d(p) : u \xrightarrow{p} v \wedge w(p) = W(u, v)\} \quad (7)$$

Every path with a delay larger than the target clock period  $\phi$  must include at least one flip-flop after retiming:

$$\forall D(u, v) > \phi : r(v) - r(u) \geq -W(u, v) + 1 \quad (8)$$

The clock-period-constrained min-area retiming problem can be formulated by the ILP composed by the objective function (4) and the constraints (5) and (8).

## 4. MIN-POWER RETIMING

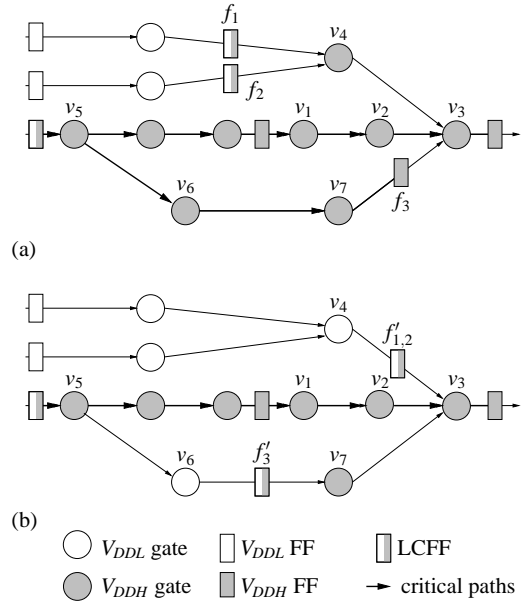
This section describes an extension of standard retiming theory to accommodate dual-supply circuits with the central objective to minimize dynamic power. We first give a motivation for using a retiming framework for dual- $V_{DD}$  synthesis which is then followed by a detailed description of the mathematical formulation.

### 4.1 Motivation

Recall that CVS is the most popular scheme for designing dual-supply circuits in which level conversion is only applicable in combination with flip-flops. To avoid the use of asynchronous level converters, connections from high-voltage gates to low-voltage gates are inhibited in this scheme. As a result, a simple heuristic that attempts to assign low-voltage gates to the part of a circuit that is not

timing critical will frequently be blocked by combinational paths that start at non-critical gates and lead to a critical sections of the circuit.

The example in Figure 2(a) illustrates this situation. Here the critical paths formed by  $V_{DDH}$  gates  $\{v_1, v_2, v_3\}$  forces the non-critical gate  $v_4$  to be assigned to  $V_{DDH}$ , otherwise an illegal configuration of a high-voltage gate followed by a low-voltage gate would occur. This requirement results in a significant limitation of possible power savings because a gate can only be assigned to low voltage if all gates of its combinational fanout structure are also assigned to  $V_{DDL}$ . Figure 2(b) shows how retiming can alleviate this limitation by moving flip-flops to the positions where level converters are required. Flip-flops  $f_1$  and  $f_2$  are retimed forward permitting  $v_4$  to be assigned to  $V_{DDL}$ . In addition, by retiming flip-flop  $f_3$  backward, the timing of the otherwise critical path  $\{v_5, v_6, v_7\}$  is relaxed. As a result gate  $v_6$  can be assigned to low voltage.



**Figure 2: Example for low-power retiming (gate delay = 1 unit): (a) original circuit using 9 gates at  $V_{DDH}$ , (b) retimed circuit using 7 gates at  $V_{DDH}$ .**

### 4.2 Delay Constraints

The presented min-power retiming formulation for dual-supply circuits is based on a number of additional assumptions which are outlined below. These assumptions reflect the typical scenario for using multiple supply voltages in practical circuit design.

- We focus our effort on clustered voltage scaling (CVS) as the most common form of applying multiple supply voltages. An extension of our formulation to ECVS is possible.
- The power consumption and delays of flip-flops at low-supply voltage are assumed to be identical to the values of flip-flops with level converters (LCFF). This simplification reflects the fact that the largest power savings comes from avoiding high-voltage flip-flops (see Table 1).
- The gates of the clocking tree for flip-flops at low-supply voltage and LCFFs are assigned to  $V_{DDL}$ . Similarly, the clock distribution of the flip-flops at high-supply voltage is driven

by  $V_{DDH}$ . The power consumption of the two clocking trees is assumed to grow linearly with the number of their respective flip-flops.

- The switching activity of the circuit and the pattern of glitches causing additional power consumption is not effected by retiming or the corresponding changes in power dissipation are negligible.

For modeling min-power retiming, the supply levels of the gates are encoded as a gate labeling  $x : V \mapsto \{0, 1\}$ , where  $x(v) = 0$  and  $x(v) = 1$  denote that gate  $v$  is assigned to  $V_{DDL}$  and  $V_{DDH}$ , respectively. Let  $d_L(v)$  and  $d_H(v)$  be the delay of the  $V_{DDL}$  and  $V_{DDH}$  version of gate  $v$ , respectively and  $p_L(v)$  and  $p_H(v)$  represent their respective power dissipation. Then the delay and power consumption of  $v$  are:

$$d(v) = x(v) \cdot d_H(v) + (1 - x(v)) \cdot d_L(v) \quad (9)$$

$$p(v) = x(v) \cdot p_H(v) + (1 - x(v)) \cdot p_L(v) \quad (10)$$

For the delay of a path  $p = (v_1 \xrightarrow{e_1} v_2 \xrightarrow{e_2} \dots \xrightarrow{e_{n-1}} v_n)$  we introduce an upper and a lower bound as follows:

$$d_L(p) = \sum_{v_i \in p} d_L(v_i)$$

$$d_H(p) = \sum_{v_i \in p} d_H(v_i)$$

Recall that in classical retiming the delay matrix  $D$  (7) is used to generate flip-flop constraints for paths between gates that would exceed the target clock period. In dual-supply retiming, we must distinguish between three cases: (1) The slowest path between the gates *does not violate* the clock period  $\phi$  if all its gates are assigned to  $V_{DDL}$ ; (2) The slowest path *does violate*  $\phi$  even with all its gates at  $V_{DDH}$ ; (3) The slowest path *would violate*  $\phi$  with its gates at  $V_{DDL}$  but can sufficiently be powered-up by assigning some of its gates to  $V_{DDH}$ .

We can apply lower and upper delay bounds to distinguish between these three cases. Using the classical definition (6) of the weight matrix  $W$ , the upper and lower bounds for the delay between two gates  $u$  and  $v$  are expressed by the following matrices:

$$D_L(u, v) = \max_p \{d_L(p) : u \xrightarrow{p} v \wedge w(p) = W(u, v)\}$$

$$D_H(u, v) = \max_p \{d_H(p) : u \xrightarrow{p} v \wedge w(p) = W(u, v)\}$$

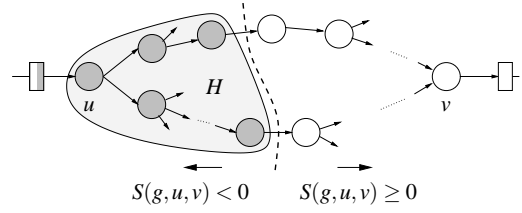
The third case of the above mentioned classification occurs when the target clock period  $\phi$  lies between  $D_L(u, v)$  and  $D_H(u, v)$ . For computing the set of gates that need to be assigned to  $V_{DDH}$ , we introduce the concept of *voltage slack*. Let  $t_A(g, p)$  and  $t_R(g, p)$  denote the *path fast arrival time* and *path slow required time*, respectively, of gate  $g$  along path  $p = u \xrightarrow{p_1} g \xrightarrow{p_2} v$ :

$$t_A(g, p) = \sum_{g' \in p_1} d_H(g')$$

$$t_R(g, p) = \phi - \sum_{g' \in p_2} d_L(g') - d_L(g)$$

The path fast arrival time is simply computed by summing up the delays of the  $V_{DDH}$  versions of the gates starting from  $u$  up to but not including  $g$ . The path slow required time starts from  $v$  and subtracts from the clock period all delays of the  $V_{DDL}$  versions.

Let  $T_A(g, u, v)$  and  $T_R(g, u, v)$  denote the arrival and required time



**Figure 3: Computation of the set of gates  $H$  that must be assigned to  $V_{DDH}$  if no flip-flop is placed between  $u$  and  $v$ .**

matrices of gate  $g$  with respect to gate pair  $(u, v)$ :

$$T_A(g, u, v) = \max_p \{t_A(g, p) : u \xrightarrow{p} v \wedge w(p) = W(u, v)\}$$

$$T_R(g, u, v) = \min_p \{t_R(g, p) : u \xrightarrow{p} v \wedge w(p) = W(u, v)\}$$

These values can be computed by a simple topological traversal similar to static timing analysis. Based on  $T_A(g, u, v)$  and  $T_R(g, u, v)$  we can now define the voltage slack matrix  $S$  as follows:

$$S(g, u, v) = T_R(g, u, v) - T_A(g, u, v)$$

Note that in contrast to the classical slack definition, the value of the voltage slack generally increases along any path from  $u$  to  $v$ . This is because the faster  $V_{DDH}$  gate versions are used for the fast arrival times whereas the slow required times are computed based on the slower  $V_{DDL}$  versions. The voltage slack can be used as a criteria for assigning supply levels to gates:

**Theorem:** In the absence of a flip-flop on a path between gate  $u$  and  $v$  of circuit  $C$ , if the voltage slack  $S(g, u, v)$  of gate  $g$  is negative, then  $g$  must be assigned to  $V_{DDH}$  in order for  $C$  to meet the clock period  $\phi$ .

For the generation of the clock-period constraints, let  $H(u, v)$  denote the set of gates on all paths between  $u$  and  $v$  with negative voltage slack values, i.e.:

$$H(u, v) = \{g \mid S(g, u, v) < 0\}$$

Figure 3 illustrates the situation in which the set of gates between  $u$  and  $v$  contained in  $H$  that must be assigned to  $V_{DDH}$  in order to realize the clock period.

In the following we discuss the setup of the clock-period constraints for the min-power retiming formulation based on the previously introduced concepts. As mentioned earlier, we need to distinguish between three different cases:

1. If the target clock period  $\phi$  is equal or greater than the upper bound of the delays of all paths between  $u$  and  $v$ , i.e.  $\phi \geq D_L(u, v)$ , then no additional flip-flops are needed between the two gates because  $V_{DDL}$  gates are sufficient to meet the timing requirements. Thus no constraint is generated.
2. If the target clock period is smaller than the lower delay bound then at least one flip-flop must be placed on all paths between the two gates, i.e.:

$$\forall D_H(u, v) > \phi : r(v) - r(u) \geq -W(u, v) + 1 \quad (11)$$

This is independent of the  $V_{DD}$  assignment for the gates.

3. In the remaining case the target clock period is equal or greater than the lower delay bound but less than the upper delay

bound ( $D_H(u, v) \leq \phi < D_L(u, v)$ ). Here, either (1) a flip-flop must be placed on all paths between the gates or (2) all gates between  $u$  and  $v$  with negative slack must be assigned to  $V_{DDH}$ . Formally stated:

$$\begin{aligned} \forall D_H(u, v) \leq \phi < D_L(u, v) : \\ (r(v) - r(u) \geq -W(u, v) + 1) \vee (\forall g \in H(u, v) : x(g) = 1) \end{aligned}$$

This condition can be expressed by the following linear constraint:

$$\begin{aligned} \forall D_H(u, v) \leq \phi < D_L(u, v) : \\ |H(u, v)| \cdot (r(v) - r(u) + W(u, v) - 1) + \sum_{g \in H(u, v)} x(g) \geq 0 \end{aligned} \quad (12)$$

Note that the expression  $r(v) - r(u) + W(u, v)$  is equal to the number of flip-flops on all paths  $p : u \xrightarrow{p} v$  with  $w(p) = W(u, v)$  and thus non-negative due to constraints (5).

The presented formulation does not consider the supply voltage assignment of the flip-flops at the outgoing edges of gate  $u$ . In cases that are extremely timing critical it may be necessary to assign, in addition to all gates between  $u$  and  $v$ , these flip-flops also to  $V_{DDH}$ . For simplicity, we did not include the corresponding components in our formulation. However, modeling the supply level of the flip-flops can easily be incorporated by including them into the computation and analysis of the voltage slack.

In addition to constraint (5), which forces non-negative edge weights and the previously described delay equations, the retiming solutions must also be restricted to CVS configurations, i.e., no  $V_{DDH}$  gate must follow a  $V_{DDL}$  gate without a flip-flop in-between:

$$(x(u) = 0 \wedge x(v) = 1) \Rightarrow (r(v) - r(u) + w(u, v) > 0)$$

converted into a linear constraint:

$$\forall (u, v) \in E : r(v) - r(u) + w(u, v) + x(u) - x(v) \geq 0 \quad (13)$$

### 4.3 Power Consumption

The power dissipation of a circuit includes mainly static power caused by leakage and other parasitic effects and the power dynamically consumed by the gates, flip-flops, and interconnections. Equation (10) gives the dynamic power  $p(v)$  of gate  $v$  for a given assignment of  $x(v)$ . This equation is based on the constants  $p_L(v)$  and  $p_H(v)$  that give the power dissipation of a gate for its  $V_{DDL}$  and  $V_{DDH}$  version, respectively. Based on our original modeling assumptions (Section 3 and Section 4.2) we do not take into account any changes to the interconnection structure caused by retiming. Therefore, we can simply include the power consumption of the interconnection into the constants  $p_L$  and  $p_H$ .

Even though our approach does not directly address power due to glitching, reducing supply voltage for as many gates as possible (weighted by the load) will reduce power due to glitching.

For the flip-flops we made the assumption that the size of the clocking tree and thus its power dissipation grows linearly with the number of flip-flops. We further assumed that all  $V_{DDL}$  flip-flops and LCFFs consume an identical amount of power. Therefore, the power consumption of all flip-flops can be modeled by a simple weighted summation.

In summary, for a given retiming  $r$  and supply level assignment  $x$  the total power consumption of a circuit  $C$  is computed as:

$$P = P_{const} + \sum_{v \in G} p(v) + p_{ff} \cdot \sum_{(u, v) \in E} w_r(u, v) \quad (14)$$

where,  $P_{const}$  and  $p_{ff}$  denote the static power dissipation of the circuit and the power consumption of a single  $V_{DDL}$  flip-flop, respectively. For a generalized retiming formulation that also considers  $V_{DDH}$  flip-flops, this equation can easily be extended by adding a term for them.

### 4.4 Complete Retiming Formulation

Summarizing the constraints and objective function described in the previous sections, the formulation for min-power dual-supply retiming is composed of four components:

1. The constraint from (5) ensures that there is a positive number of flip-flops on each edge:

$$\forall (u, v) \in E : r(v) - r(u) \geq -w(u, v)$$

2. Clock period constraints derived from (11) and (12):

$$\forall D_H(u, v) > \phi :$$

$$r(v) - r(u) \geq -W(u, v) + 1$$

$$\forall D_H(u, v) \leq \phi < D_L(u, v) :$$

$$|H(u, v)| \cdot (r(v) - r(u) + W(u, v) - 1) + \sum_{g \in H(u, v)} x(g) \geq 0$$

The first states that at least one flip-flop is present along all paths whose lower delay bound is greater than the target clock period. The second constraint relates to all paths whose lower delay bound is less than or equal to the clock period but whose upper delay bound exceeds the target clock period. At least one register must be present along these paths or a specific set of combinational gates on the path must be assigned to  $V_{DDH}$ .

3. Limitation to CVS configurations stated in (13):

$$\forall (u, v) \in E : r(v) - r(u) + w(u, v) + x(u) - x(v) \geq 0$$

4. Objective to minimize power given in (14):

$$\sum_{v \in G} p(v) + p_{ff} \cdot \sum_{v \in G} r(v) \cdot (|FI(v)| - |FO(v)|) \rightarrow \min$$

These equations present an Integer Linear Program with the integer variables  $r$  and  $x$  encoding the actual shifting of flip-flops and gate assignments to supply voltages, respectively. Unfortunately, in contrast to the original retiming formulation, this ILP cannot be directly translated into a network-flow problem for which efficient algorithms are available. Our current effort focuses on developing an efficient solver for the given problem by exploiting practical artifacts for further reducing the problem size.

## 5. EXAMPLE

This section presents a simple example based on a circuit that implements the following arithmetic equations:

$$z(n) = v(n) + t(n - 1)$$

$$t(n) = a \cdot z(n - 1) + b \cdot z(n - 2)$$

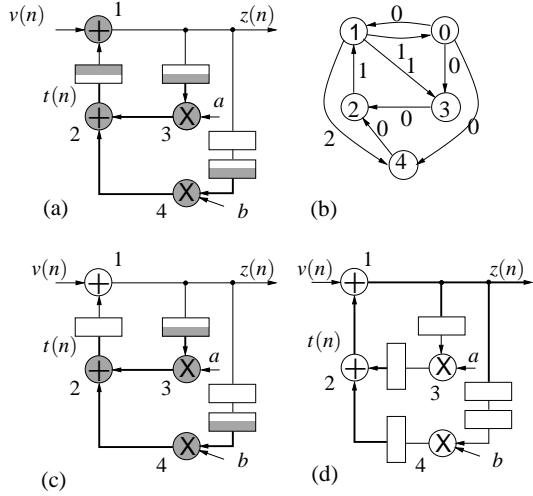
Figure 4 (a) shows the corresponding circuit structure, which uses four registers, two adders (gates 1 and 2) and two multipliers (gates 3 and 4). Part (b) of that figure displays the retiming graph  $C$  that includes four regular vertices and the host vertex 0 for the gates and input/outputs respectively. Note that a register must be inserted between node 0 and node 1 as a loop between nodes 0 and 1 form a cycle. According to classical retiming, a cycle in

the retiming graph must contain at least one register. This register cannot be moved however: therefore  $r(0) = 0$ .

Table 2 gives the delay and power consumption values for the circuit elements which results in a minimum circuit clock period of  $\phi = 3$  when all components are operating at high supply voltage.

Gate Type	Delay	Energy
Adder $V_{DDL}$	2	1
Adder $V_{DDH}$	1	3
Multiplier $V_{DDL}$	4	2
Multiplier $V_{DDH}$	2	5
Register $V_{DDL}$		1

**Table 2: Delay and power values of the gates in Figure 4.**



**Figure 4: Example: (a) original circuit  $\phi = 3$  (b) retiming graph (c) solution for target  $\phi = 3$  (d) solution for target  $\phi = 4$ .**

The extension to classical retiming theory for dual-supply CMOS circuits is illustrated below for the example circuit shown in Figure 4(a) where power is minimized given a target clock period  $\phi = 3$ .

1. Legal retiming constraints:

$$\begin{aligned} r(0) - r(1) &\leq 0 & r(1) - r(0) &\leq 1 & r(2) - r(1) &\leq 1 \\ r(0) - r(3) &\leq 0 & r(1) - r(3) &\leq 1 & r(3) - r(2) &\leq 0 \\ r(0) - r(4) &\leq 0 & r(1) - r(4) &\leq 2 & r(4) - r(2) &\leq 0 \end{aligned}$$

2. Clock period constraints:

The weight matrix  $W$  is given as:

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 2 & 1 & 0 & 2 & 2 \\ 2 & 1 & 0 & 0 & 2 \\ 2 & 1 & 0 & 2 & 0 \end{bmatrix}$$

The delay matrices  $D_L$  and  $D_H$  are:

$$\begin{bmatrix} 0 & 2 & \boxed{6} & \boxed{4} & \boxed{4} \\ 2 & 2 & 8 & \boxed{6} & \boxed{6} \\ \boxed{4} & \boxed{4} & 2 & 8 & 8 \\ 8 & 8 & \boxed{6} & \boxed{4} & 12 \\ 8 & 8 & \boxed{6} & 12 & \boxed{4} \end{bmatrix} \begin{bmatrix} 0 & 1 & 3 & 2 & 2 \\ 1 & 1 & \boxed{4} & 3 & 3 \\ 2 & 2 & 1 & \boxed{4} & \boxed{4} \\ \boxed{4} & \boxed{4} & 3 & 2 & \boxed{6} \\ \boxed{4} & \boxed{4} & 3 & \boxed{6} & 2 \end{bmatrix}$$

For an assumed target clock period of  $\phi = 3$  all values of the  $D_L$  matrix are highlighted that generate constraints from condition (12). Similarly, those values in the  $D_H$  matrix are marked that produce constraints from condition (11).

For all paths where the lower delay bound is greater than the target clock period, the following constraints are generated according to (11):

$$\begin{aligned} r(1) - r(2) &\leq 0 & r(2) - r(3) &\leq 1 & r(2) - r(4) &\leq 1 \\ r(3) - r(0) &\leq 1 & r(3) - r(1) &\leq 0 & r(3) - r(4) &\leq 1 \\ r(4) - r(0) &\leq 1 & r(4) - r(1) &\leq 0 & r(4) - r(3) &\leq 1 \end{aligned}$$

The path fast arrival and path slow required time matrices for each gate are given as follows for  $\phi = 3$ .

$T_A(0, u, v)$  and  $T_R(0, u, v)$ :

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & & & 1 & \\ 2 & & & 2 & \\ 4 & & & 4 & \\ 4 & & 4 & & \end{bmatrix} \begin{bmatrix} 3 & 1 & -3 & -1 & -1 \\ 3 & & & & -1 \\ 3 & & & & -1 \\ 3 & & & & -1 \\ 3 & & & & -1 \end{bmatrix}$$

$T_A(1, u, v)$  and  $T_R(1, u, v)$ :

$$\begin{bmatrix} 0 & & & & \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & & 1 & 1 \\ 3 & 3 & & 3 & \\ 3 & 3 & 3 & & \end{bmatrix} \begin{bmatrix} 1 & & & & \\ 1 & 1 & -5 & -3 & -3 \\ 1 & 1 & & -3 & -3 \\ 1 & 1 & & & -3 \\ 1 & 1 & & & -3 \end{bmatrix}$$

$T_A(2, u, v)$  and  $T_R(2, u, v)$ :

$$\begin{bmatrix} 2 & & & & \\ 3 & & & & \\ 0 & 0 & 0 & 0 & 0 \\ 2 & 2 & 2 & 2 & \\ 2 & 2 & 2 & 2 & \end{bmatrix} \begin{bmatrix} 1 & & & & \\ 1 & & & & \\ -1 & -1 & 1 & -5 & -5 \\ -1 & -1 & 1 & & -5 \\ -1 & -1 & 1 & -5 & \end{bmatrix}$$

$T_A(3, u, v)$  and  $T_R(3, u, v)$ :

$$\begin{bmatrix} 0 & 0 & & & \\ 1 & 1 & & & \\ 2 & & & & \\ 0 & 0 & 0 & 0 & 0 \\ 4 & & & & \end{bmatrix} \begin{bmatrix} -3 & -1 & & & \\ -3 & -1 & & & \\ & & -1 & & \\ -5 & -5 & -3 & -1 & -9 \\ & & & & -1 \end{bmatrix}$$

$T_A(4, u, v)$  and  $T_R(4, u, v)$ :

$$\begin{bmatrix} 1 & 0 & & & \\ 1 & 1 & & & \\ & & 2 & & \\ & & & 4 & \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} -3 & -1 & & & \\ -3 & -1 & & & \\ & & -1 & & \\ & & & -1 & \\ -5 & -5 & -3 & -9 & -1 \end{bmatrix}$$

Based on the two delay matrices and the slack matrix calculated from the above path fast arrival and path slow required time matrices, the resulting  $H$  matrix is:

$$H = \begin{bmatrix} & & \{0, 2, 3, 4\} & \{0, 3\} & \{0, 4\} \\ & & & \{1, 3\} & \{1, 4\} \\ \{2\} & \{2\} & & & \\ & & \{2, 3\} & \{3\} & \\ & & \{2, 4\} & & \{4\} \end{bmatrix}$$

The resulting constraints are:

$$\begin{aligned}
4(r(2) - r(0) - 1) + x(0) + x(2) + x(3) + x(4) &\geq 0 \\
2(r(3) - r(0) - 1) + x(0) + x(3) &\geq 0 \\
2(r(4) - r(0) - 1) + x(0) + x(4) &\geq 0 \\
2(r(3) - r(1)) + x(1) + x(3) &\geq 0 \\
2(r(4) - r(1)) + x(1) + x(4) &\geq 0 \\
(r(0) - r(2) + 1) + x(2) &\geq 0 \\
(r(1) - r(2)) + x(2) &\geq 0 \\
2(r(2) - r(3) - 1) + x(2) + x(3) &\geq 0 \\
-1 + x(3) &\geq 0 \\
2(r(2) - r(4) - 1) + x(2) + x(4) &\geq 0 \\
-1 + x(4) &\geq 0
\end{aligned}$$

3. Limitation to CVS configurations:

$$\begin{aligned}
r(0) - r(1) - x(0) + x(1) \leq 0 & \quad r(1) - r(0) - x(1) + x(0) \leq 1 \\
r(0) - r(3) - x(0) + x(3) \leq 0 & \quad r(1) - r(3) - x(1) + x(3) \leq 1 \\
r(0) - r(4) - x(0) + x(4) \leq 0 & \quad r(1) - r(4) - x(1) + x(4) \leq 2 \\
r(2) - r(1) - x(2) + x(1) \leq 1 & \quad r(3) - r(2) - x(3) + x(2) \leq 0 \\
r(4) - r(2) - x(4) + x(2) \leq 0 &
\end{aligned}$$

4. Objective to minimize power:

$$\begin{aligned}
2x(1) + 2x(2) + 3x(3) + 3x(4) + 6 + \\
1 \cdot [-2r(0) - r(1) + r(2) + r(3) + r(4)] \rightarrow \min
\end{aligned}$$

The corresponding solution to the minimum-power retiming problem results in a circuit structure given in Figure 4(c). As shown, the two multipliers and one of the adders are assigned to  $V_{DDH}$  whereas the other adder can be set to  $V_{DDL}$ . This results in a reduced power consumption by two units assuming each register consumes one unit of power (i.e.  $p_{ff} = 1$ ).

If the target clock period is relaxed by one unit to  $\phi = 4$ , all four gates in the retimed circuit shown in Figure 4(d) can operate at  $V_{DDL}$  given that the register between nodes 1 and 2 is retimed backwards through node 2. Here the additional power savings is 7 units.

## 6. CONCLUSIONS AND FUTURE WORK

In this paper we have presented early work that describes a mathematical framework for dual- $V_{DD}$  synthesis. For this we have extended the standard Leiserson and Saxe retiming formulation to optimize for minimal dynamic power by utilizing any available timing slack. The core of the extension includes the use of two delay matrices instead of one. These matrices are applied to differentiate between three cases for placing flip-flops between pairs of gates in order to meet the clock-period constraint, namely: (1) no flip-flop is needed, (2) a flip-flop must always be inserted, and (3) either a flip-flop is inserted or a specific set of combinational gates is assigned to  $V_{DDH}$  for fast performance.

Unfortunately, the new formulation cannot be solved directly by a standard network-flow approach as done for classical retiming. This has a significant impact on the practical runtime complexity. Our current research focuses on an efficient implementation of a corresponding ILP solver that makes this approach applicable for practical-sized designs. We further investigate generalizations of this work to also consider other power optimization techniques such as multiple threshold values and gate resizing.

The use of dual-supply voltages at the gate level impacts physical design for a CMOS circuit. For this we investigate alternative physical design methodologies and power network distribution schemes for dual-supply circuits.

## 7. ACKNOWLEDGMENTS

The authors would like to thank Borivoje Nikolić, Dave Chinery, and Michael Orshansky for providing valuable feedback during the development of the mathematical framework presented in this paper. Furthermore, we would like to thank Fujio Ishihara for providing the detailed simulation data for Table 1.

## 8. REFERENCES

- [1] S. Borkar. Design challenges of technology scaling. *IEEE Micro*, 19(4):23–29, July/August 1999.
- [2] M. Igarashi, K. Usami, K. Nogami, F. Minami, Y. Kawasaki, T. Aoki, M. Takano, S. Sonoda, M. Ichida, and N. Hatanaka. A low-power design method using multiple supply voltages. In *Proceedings of the International Symposium on Low Power Electronics and Design*, pages 36–41, 1997.
- [3] F. Ishihara. Level-converting flip-flop tradeoffs. Technical Report Unpublished research report, University of California, Berkeley, Spring 2002.
- [4] T. Kuroda and T. Sakurai. Low-power circuit design techniques for multimedia CMOS VLSIs. *Electronics and Communications in Japan, Part 3*, 81(9), 1998.
- [5] C. Leiserson and J. Saxe. Optimizing synchronous systems. *Journal of VLSI and Computer Systems*, 1(1):41–67, January 1983.
- [6] C. Leiserson and J. Saxe. Retiming synchronous circuitry. *Algorithmica*, 6:5–35, 1991.
- [7] J. M. Rabaey. *Digital Integrated Circuits: A Design Perspective*. Prentice-Hall, New Jersey, 1996.
- [8] N. Shenoy. Retiming: Theory and practice. *Integration, The VLSI Journal*, 22(1-2):1–21, August 1997.
- [9] V. Sundararajan and K. K. Parhi. Synthesis of low power CMOS VLSI circuits using dual supply voltages. In *Proceedings of the 36th ACM/IEEE Design Automation Conference*, pages 72–75, New Orleans, LA, June 1999.
- [10] Torsten Mahnke and Sebastian Panenka and Martin Embacher and Walter Stechele and Wolfgang Hoeld. Power optimization through dual supply voltage scaling using power compiler. In *SNUG Europe*, 2002.
- [11] K. Usami and M. Horowitz. Clustered voltage scaling technique for low-power design. In *Proceedings ISPLD*, pages 3–8, April 1995.
- [12] K. Usami and M. Igarashi. Low-power design methodology and applications utilizing dual supply voltages. In *Proceedings of the Asia and South Pacific Design Automation Conference*, pages 123–128, January 2000.
- [13] Usami, K. and Igarashi, M. and Minami, F. and Ishikawa, T. and Kanzawa, M. and Ichida, M. and Nogami, K. Automated low-power technique exploiting multiple supply voltages applied to media processor. *Journal of Solid-State Circuits*, 33(3):463–472, 1998.
- [14] C. Yeh, Y.-S. Kang, S.-J. Shieh, and J.-S. Wang. Layout techniques supporting the use of dual supply voltages for cell-based designs. In *Proceedings of the 36th ACM/IEEE Design Automation Conference*, pages 62–67, New Orleans, LA, June 1999.