Retiming and Dual-supply Voltage Based Energy Optimization for DSP Applications

Teng Xu, Miodrag Potkonjak
Computer Science Department
University of California, Los Angeles
{xuteng, miodrag}@cs.ucla.edu

***Abstract*—Low energy, as one of the most important metrics in modern design, has also been considered as an important criterion to many DSP applications. To address the low energy issue of DSP applications, in this work, we have proposed a new energy minimization approach which optimally combines the dual-supply voltage (dual-V_{dd}) with the retiming technique. We have specifically focused on various hardware-implemented transforms and filters such as the fast Fourier transform (FFT) and the discrete cosine transform (DCT) which are widely used in DSP applications. Our key idea is to use a maximum-flow/min-cut strategy to reduce the required number of flip-flop logic and while maintaining the delay of the circuit in such a way that to enable efficient V_{dd} allocation in the subsequent stage. The effectiveness of our approach is demonstrated on both FFT and DCT applications using an industrial cell library. Given a target delay, a significant energy saving using our approach is observed.**

***Keywords*—Energy Optimization, Retiming, Dual Supply-Voltage, DSP Applications**

I. Introduction

Low energy has emerged to become one of the most important design metrics in the last 25 years. It is especially important for DSP applications due to their highly constrained resources in many scenarios. On the other hand, transforms and filters such as FFT and DCT are commonly used in the DSP applications such as video compression, audio signal processing, and image processing. In order to gain the benefit of fast processing speed and small area, these transforms and filters are often implemented in hardware. Therefore, the main goal of our work is to optimize the energy consumption of such hardware-implemented transforms and filters used in the DSP applications.

Two previous techniques have been widely applied to optimize the energy consumption of circuits. The first technique is the dual-V_{dd} optimization, where the key idea is to assign two supply voltages to a circuit instead of one. The major benefit is that the circuit path with long delay is significantly shorten by assigning the gates on that path to high voltage. However, it has its own constraints. The most important one is that only the gates with high voltage can drive the gates with low voltage. This significantly impedes the optimal voltage allocation of gates due to the fact that now the process of voltage allocation highly depends on the structure of the circuit. Another widely employed energy optimization technique is retiming. It is dedicated to change the structure of the circuit and is performed in such a way that the circuit functionality (e.g., relation between the inputs and outputs) is not altered. Similarly to dual-V_{dd} optimization, retiming has its own drawbacks. For instance, the process to minimize clock period may cause an unacceptable increase in the number of flip-flops. Also, retiming itself has structure constraints, e.g., the flip-flops whose inputs are primary inputs can not be retimed.

The objective goal of our work is to propose an approach to combine the above two technologies to enable an even more effective energy optimization scheme while trying to leverage as many constraints as possible. The intuition behind our objective is to take advantage of retiming to change the structure of the circuit in such a way that the new structure enables an effective usage of dual-V_{dd} assignment. For example, we initially assign all flip-flops at higher supply voltages and structure all combinational gates in such a way that they receive inputs from either flip-flops and/or other gates that operate at a higher or equal supply voltage. Therefore, it is important that the width of the circuit is small at the positions immediately after the flip-flops so that fewer gates can be placed at higher V_{dd} while still satisfying the delay constraints.

The transforms and filters circuit in DSP applications provide an ideal platform to apply our proposed technique. It is because such circuits usually employ a large number of flip-flops and a levered structure with long delay paths, which provide a large space to optimize. In the following paper, we first briefly review the previous literature on dual-V_{dd} optimization and retiming. Then we give a detailed algorithmic description of our flow of optimization. And lastly, we apply our approach on the FFT and DCT circuit to demonstrate its effectiveness.

II. Related Work

In this section, we briefly survey the most directly related work to retiming and dual-V_{dd} optimization.

A. Retiming

Leiserson and Saxe were the first to study retiming [1]. They presented polynomial time algorithms that minimize either delay or the number of sequential elements in use. Consequently, several research groups introduced approaches that combine the effectiveness of retiming with algebraic transformation technology. Berkeley's CAD group introduced a technique for optimizing a sequential network by moving registers to the boundary of the network using an extension of retiming [2]. Mishchenko and his co-authors integrated not only logic synthesis but also technology mapping and retiming

[3]. Soha and Ebeling used conceptual similarities between retiming and the pipeline to optimize latency-constrained circuits [4]. Cong and his group at UCLA presented techniques combining retiming with the physical planning [5].

B. Dual Supply Voltages

Kimiyoshi Usami was the first to propose the use of multiple supply voltages as a way to reduce energy [6][7]. Salil and Sarrafzadeh applied multiple supply voltages at the behavior level for energy minimization [8]. Ishihara proposed the level converter required for dual-V_{dd} systems [9]. Srivastava had introduced technology that minimizes both switching and static power using simultaneous supply and voltage assignment [10]. Chang and his group studied the use of dual-V_{dd} by considering the requirements for power-network planning [11]. He. proposed pre-defined dual-voltage assignment for interconnections on FPGA [12][13]. Finally, Agrawal demonstrated the effectiveness using dual sub-threshold supply voltages for energy minimization in CMOS circuits [14].

III. Technical Approach

In this section, we summarize our contribution on two algorithms, respectively retime for minimal flip-flops (RTMF) and dual-V_{dd} optimization. The first algorithm we propose is RTMF. The goal of the algorithm is to minimize the number of flip-flops in the circuit after retiming. We design our algorithm in such a way that the critical path delay achieved by the optimal retiming will be kept during the RTMF. The second algorithm we propose is dual-V_{dd} optimization. The objective is to find the best low and high V_{dd} pair for energy optimization as well as the corresponding cell V_{dd} assignment.

A. Retime for Minimal Flip-flops

We present a motivational example in Figure 1. Given the original circuit on the left side of (a), if we choose to assign all the flip-flops and all the gates directly connected to the flip-flops at high V_{dd} to achieve delay reduction, then 4 flip-flops and 4 gates need to be assigned. However, if we retime the circuit, the number of flip-flops decreases to only 1 and the number of gates that directly connect to flip-flops decreases to 1 as well. Consequently, in order to achieve the same critical path delay with dual-V_{dd} compared to the original circuit in (a), only 1 flip-flop and 1 gate are required to be assigned to high V_{dd}. Note that both the original circuit as well as the retimed circuit after applying dual-V_{dd} have 1 flip-flop and 4 gates on the critical path in which 1 gate is assigned to high V_{dd}. Therefore, the delay of the dual-V_{dd} circuit remains unaltered before and after retiming.

The pseudocode for the RTMF is illustrated in Algorithm 1. The core idea is to apply the minimum-cut retiming on the circuit so that fewer flip-flops and gates after flip-flops need to be assigned to high voltage. However, we don't want to increase the circuit delay during the process, therefore, we only use minimum-cut retiming on the flip-flops that will not influence the circuit delay. The way to find such flip-flops is to first fix the position of the flip-flops in the critical path of the original circuit and then apply minimum-cut. Afterwards, we check whether we have a new critical path. If the critical path has changed, we revert to the previous circuit state, add the flip-flops in the new critical path to our set of fixed flip-flops, then reapply the minimum-cut retiming procedure again. On the other hand, if the critical path does not change, we halt our algorithm and use the current configuration. Our algorithm guarantees that the circuit delay after RTMF is not influenced by the minimum-cut procedure. Our overall approach targets at retrieving the optimal minimum-cut on the circuit on the premise not to increase the circuit delay.

Algorithm 1 Retiming for minimum flip-flops (RTMF)

Input: C_0 - original circuit.
Input: CP_0 - critical path on C_0.
Input: FF_0 - flip-flops on C_0.
FF_{fix} is a vector that contains all the flip-flops that are fixed.

$FF_{fix} = \emptyset$
do
  for all ff_i in FF_0
    if ff_i is in CP_0
      FF_{fix}.append(ff_i)
    end if
  end for
  $C_{pre} = C_0$
  $(C_0, CP_0, FF_0) = Mincut((C_0, CP_0, FF_0 - FF_{fix}))$
while $CP_0 != CP_{pre}$
Output: C_0

Algorithm 2 Dual-vdd Optimization (DV)

Input: C - original circuit.
Input: Vdd_0 - original supply voltage on C_0.
Vec_{high} is a vector that contains the gates with high vdd.
Vec_{low} is a vector that contains the gates with low vdd.

Vec_{high}= $\emptyset$.
Vec_{low}= all gates in C_0.
$Vdd_{high} = Vdd_{low} = Vdd_0$
do
  $CP_0 = CriticalPath(C, Vdd_{low}, Vdd_{high}$
    $Vec_{low}, Vec_{high})$
  $POW_0 = Power(C, Vdd_{low}, Vdd_{high}, Vec_{low}, Vec_{high})$
  for all $Gate_i$ in CP_0
    Vec_{high}.append($Gate_i$ with smallest slack)
    Vec_{low}.erase($Gate_i$ with smallest slack)
  end for
  for all possible $Vdd_{newlow} \leq Vdd_0$
    binary search $Vdd_{newhigh}$
    until $CP_0 = CriticalPath(C, Vdd_{newlow},$
      $Vdd_{newhigh}, Vec_{low}, Vec_{high})$
    $POW_1 = Power(C, Vdd_{lownew},$
      $Vdd_{highnew}, Vec_{low}, Vec_{high})$
    if $POW_1 < POW_0$
      $Vdd_{low} = Vdd_{newlow}$
      $Vdd_{high} = Vdd_{newhigh}$
      $POW_0 = POW_1$
    end if
  end for
while Vec_{low}!= $\emptyset$
Output: C

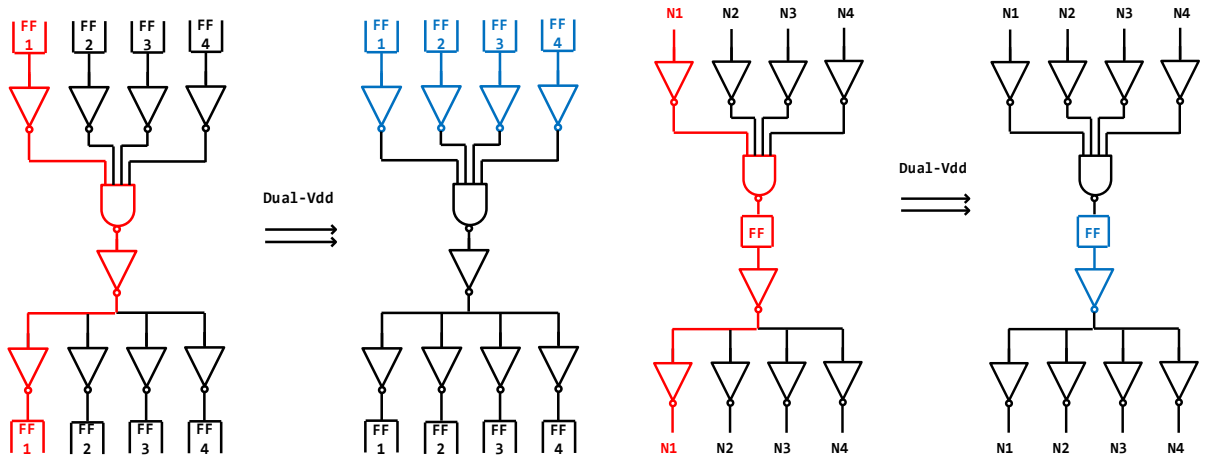

Fig. 1: An example of how our method of retiming influences the assignment of dual-V_{dd}. (a) Assign dual-V_{dd} to the original circuit: The red lines represents the critical path in the original circuit. The blue line represents the flip-flops and the gates to be put in high V_{dd}. (b) Assign dual-V_{dd} to the retimed circuit.

B. Dual Supply Voltage Optimization (DV)

When applying dual supply voltages to a circuit, two essential questions need to be answered. The first is what voltages should be used, the second is which part of the circuit should be assigned to high voltage (or low voltage). In answering these two questions, we arrived at the procedure in Algorithm 2 that heuristically approximates the best voltage pairs and the corresponding coverage. We assume that in our design, only the gates with high supply voltage can drive the gates with low voltage, thus we do not need to use level converters in our circuit. In each iteration, we choose one gate in the current critical path with the shortest arrival time and place it in the group of gates with high supply voltage. Subsequently, given the current circuit configuration, we use binary search to traverse the pairs of voltages that meet the given target delay and find the pair that achieves the smallest power consumption. We repeat these steps until all gates in the circuit have been placed in the high voltage group. From here, we choose the lowest point of energy consumption from all explored iterations. In practice, the minimal energy is normally achieved when only a small subset of gates are put in high voltage, therefore, the algorithm can be stopped when no more energy is reduced within some number of iterations. Our algorithm provides a heuristic method to approximate the best pair of supply voltages as well as the corresponding circuit configuration in order to achieve the minimal energy consumption configuration.

Figure 2 depicts an example of the performance of our algorithm. The tested circuit, DMA, has the following initial configuration: the total number of gates is 25301, the initial supply voltage is 0.7V, the critical path delay is 10737ps, and the power is 52337μw. According to Algorithm 2, we guarantee that the circuit always has the same target delay during the whole process while the high voltages and low voltages can be adjusted as long as the target delay is fixed. We observe in Figure 2 that the minimal power is achieved at iteration 3217, with high V_{dd} set to 0.75V and low V_{dd} set to 0.60V. Only the initial 4000 iterations are shown in the figure

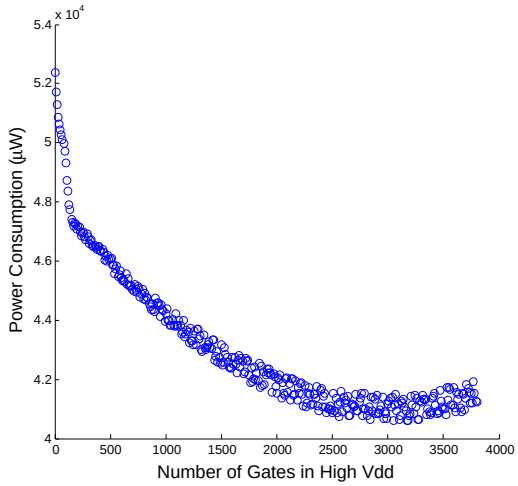

Fig. 2: An example of the performance of the dual-voltage optimization algorithm on the DMA circuit.

as the energy consumption of the rest iterations will gradually increase and return back to the starting point. A detailed data analysis indicates that the power (or energy) at iteration 3217 corresponds to a 22.69% reduction compared to the original circuit. The reason that the initial part of the iteration causes more energy reduction is that as more gates are assigned to high voltage, the circuit becomes balanced, thus the marginal effects on energy reduction using dual-voltage is reduced. Note that throughout the whole process, the delay of the circuit does not change, only the voltages are scaled to meet the delay constraints.

IV. Experiment Setup

We adopt the ISPD2012 standard cell library [15] and fit accordingly to Markovic's EKV formulation for enabling dual-V_{dd} optimization [16]. The nominal V_{dd} is set to 0.7V and V_{th} is set to 0.3V. For our dual-V_{dd} approach, we consider

Circuit	Target Delay (ns)	Initial Energy (mJ)	Supply Voltage(s) (V)			Energy Savings (% compared to initial)	
			Scaled (after RTMF)	(V_{min}, V_{max})		RTMF	RTMF+DV
FFT-64	47.35	38.75	0.57V	0.53V	0.64V	34.57 %	45.08%
FFT-128	47.35	153.15	0.55V	0.52V	0.61V	39.27 %	56.06%
DCT-8x8	59.83	117.3	0.56V	0.52V	0.63V	35.38 %	65.1%
DCT-16x16	55.39	2988.18	0.59V	0.56V	0.68V	28.18 %	39.58%
Average	-	-	-	-	-	34.35 %	51.46 %

TABLE I: Energy savings when using standard retiming for minimum flip-flops (RTMF), and dual-V_{dd} (DV). Energy consumption is relative to each method satisfying a target delay achieved by the original configuration operating at the initial supply voltage V_{dd} (0.7V). Also presented are the scaled V_{dd} after using RTMF in column 4 and the best dual-V_{dd} pair under RTMF+DV is in columns 5 and 6. The bottom row provides an average energy savings when applying RTMF, and RTMF+DV across all the tested circuits.

V_{dd} within the range of 0.35V to 0.70V. Two famous linear transforms, DCT and FFT are applied as the benchmarks. We evaluate and synthesize the netlist of each of them using Cadence Encounter. Each design was optimized in accordance to the ISPD2012 design contest suite in satisfying each slew and cell load restrictions.

We show our simulation workflow in Figure 3. We start from the characterization using the cell library, subsequently, we use the gate-level simulation to quantify the delay, switching power, and leakage power. Then the next step is to apply the circuit optimization with the procedures of RTMF and DV. During the optimization, we use the delay of the original circuit as our target delay, the energy consumption after each single step of optimization is calculated under the target delay.

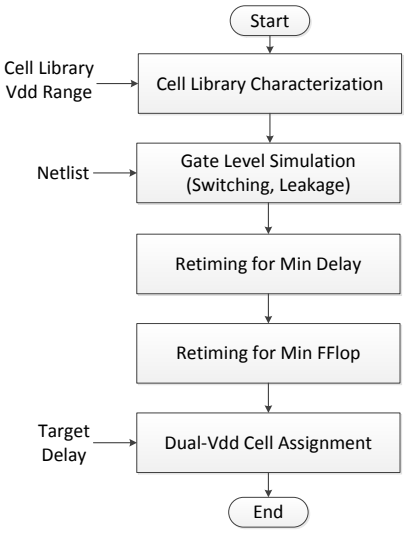

Fig. 3: Overall workflow of our simulation.

V. Results

We present results comparing the energy consumption of our retiming methods against the original circuit under identical timing constraints in Table I. All the optimization is conducted under the same target delay which is achieved through the initial circuit under original setting (without RTMF or DV, $V_{dd} = 0.7$). Energy savings are presented in an incremental fashion with respect to the achieved energy and delay of the original circuit operating at the nominal set supply voltage of 0.7V. The techniques we consider independently are minimum delay optimization through retiming and minimum flip-flop minimization (RTMF) with V_{dd} scaling such that the final delay after voltage scaling satisfies the delay target. Finally, we evaluate our approach which combines minimum retiming and minimal flip-flop with our dual-V_{dd} technique (RTMF+DV).

We show the results of our optimization in 2 steps. The first step of the result is achieved through initially retiming the original circuit and then conducting a maximum-flow/minimum-cut algorithm for flip-flop minimization. Retiming involves rearranging the sequential elements (e.g., flip-flops, latches, etc.) of the circuit to achieve the minimum delay while preserving the functionality of the design. We observe an energy reduction between 28.18% to 39.27% (34.35% avg.) is achieved when conducting RTMF alone. After the process of RTMF, since the target delay is fixed for each individual circuit, we reduce the supply voltage of the entire design uniformly across all cells to reclaim slack (increase delay back to target delay). The V_{dd} that satisfies the target delay under RTMF is recorded as the scaled V_{dd}.

Our RTMF+DV solution achieves between 39.58%-65.1% (51.46% avg.) energy reduction over the original design under the same delay target constraints. These savings translate to an additional 17.11% average energy reduction over RTMF, which indicates the potential benefits of utilizing dual-V_{dd}. The improvement is achieved through the flexibility in assigning non-critical cells at lower V_{dd} while simultaneously maintaining circuit delay on critical paths; RTMF is limited in this regard, since only a single V_{dd} is enabled and all cells, critical or non-critical, are set to this configuration.

VI. Conclusion

We have developed a new approach for energy minimization that employs retiming as an essential step for the consequent application of dual supply voltages. Our approach provides an ideal solution to enable the energy reduction of hardware implemented transforms and filters in DSP applications. The new retiming approach minimizes the number of sequential elements under the constraint that the retimed circuit has provably minimal delay achievable by any retiming. The minimization of the number of flip-flops is a heuristic measure that maximally reduces the number of gates that require placement on the high supply voltage. The approach is generic in a sense that can be easily used in standard synthesis flows and can be further extended to cover transformations such as unfolding and another degree of optimization freedom such as multiple thresholds. We have explored our techniques on the widely used FFT and DCT linear transforms using accurate gate-level models and cell sizing techniques. An average energy reduction by 51.46% is observed.

REFERENCES

[1] C. E. Leiserson and J. B. Saxe, "Retiming synchronous circuitry," *Algorithmica*, pp. 5–35, 1991.

[2] S. Malik, E. M. Sentovich, R. K. Brayton, and A. Sangiovanni-Vincentelli, "Retiming and resynthesis: Optimizing sequential networks with combinational techniques," *Computer-Aided Design of Integrated Circuits and Systems*, IEEE Transactions on 10, no. 1, pp. 74-84, 1991.

[3] A. Mishchenko, S. Chatterjee, and R. Brayton, "Integrating logic synthesis, technology mapping, and retiming," *In Proc. IWLS'05*. 2006.

[4] S. Hassoun, and C. Ebeling, "Architectural retiming: pipelining latency-constrained circuits," *In Design Automation Conference Proceedings 1996, 33rd*, pp. 708-713. IEEE, 1996.

[5] J. Cong, and S. Lim, "Physical planning with retiming," *In Proceedings of the 2000 IEEE/ACM international conference on Computer-aided design*, pp. 2-7. IEEE Press, 2000.

[6] K. Usami, and M. Horowitz, "Clustered voltage scaling technique for low-power design," *In Proceedings of the 1995 international symposium on Low power design*, pp. 3-8. ACM, 1995.

[7] M. Igarashi, K. Usami, K. Nogami, F. Minami, Y. Kawasaki, T. Aoki, M. Takano, et al, "A low-power design method using multiple supply voltages," *In Proceedings of the 1997 international symposium on Low power electronics and design*, pp. 36-41. ACM, 1997.

[8] S. Raje, and M. Sarrafzadeh, "Variable voltage scheduling," *In Proceedings of the 1995 international symposium on Low power design*, pp. 9-14. ACM, 1995.

[9] F. Ishihara, F. Sheikh, and B. Nikolic, "Level conversion for dual-supply systems," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on* 12, no. 2 (2004): 185-195.

[10] A. Srivastava, and D. Sylvester, "Minimizing total power by simultaneous Vdd/Vth assignment," *Computer-Aided Design of Integrated Circuits and Systems*, IEEE Transactions on 23, no. 5 (2004): 665-677.

[11] W. Lee, H. Liu, and Y. Chang, "An ILP algorithm for post-floorplanning voltage-island generation considering power-network planning," *In Proceedings of the 2007 IEEE/ACM international conference on Computer-aided design*, pp. 650-655. IEEE Press, 2007.

[12] F. Li, Y. Lin, L. He, and J. Cong, "Low-power FPGA using pre-defined dual-Vdd/dual-Vt fabrics," *In Proceedings of the 2004 ACM/SIGDA 12th international symposium on Field programmable gate arrays*, pp. 42-50, 2004.

[13] Y. Lin, and L. He, "Statistical dual-Vdd assignment for FPGA interconnect power reduction," *Design, Automation & Test in Europe Conference & Exhibition*, pp. 1-6, 2007.

[14] K. Kim, and V. D. Agrawal, "Minimum energy cmos design with dual subthreshold supply and multiple logic-level gates," *In Quality Electronic Design(ISQED), 2011 12th International Symposium on*, pp. 1-6. IEEE, 2011.

[15] M. M. Ozdal, C. Amin, A. Ayupov, S. Burns, G. Wilke, C. Zhuo, "The ISPD -2012 Discrete Cell Sizing Contest and Benchmark Suite", *Proc. ACM International Symposium on Physical Design*, pp. 161-164, 2012.

[16] D. Markovic, C. C. Wang, L. P Alarcon, L. Tsung-Te and J. M. Rabaey, "Ultralow-Power Design in Near-Threshold Region," *IEEE*, pp.237-252, 2010.