Low Power RT-Level Synthesis Techniques: A Tutorial

Massoud Pedram and Afshin Abdollahi

Dept. of Electrical Engineering

University of Southern California

Abstract – Power consumption and power-related issues have become a first-order concern for most designs and loom as fundamental barriers for many others. And, while the primary method used to date for reducing power has been supply voltage reduction, this technique begins to lose its effectiveness as voltages drop to sub-one volt range and further reductions in the supply voltage begin to create more problems than are solved. Under these circumstances, the process of design and the automation tools required to support that process become the critical success factors. In the last decade, huge effort has been invested to come up with a wide range of design solutions that help solve the power dissipation problem for different types of electronic devices, components and systems. These techniques range from RTL power management and multiple voltage assignment, to power-aware logic synthesis and physical design, to memory and bus interface design. This tutorial paper explains a number of representative low power design techniques from this large set. More precisely, we will describe basic techniques, applicable at RT-level and below, that have proven to hold good potential for power optimization in practical design environments.

1 Introduction

A dichotomy exists in the design of modern microelectronic systems: they must be low power and high performance, simultaneously. This dichotomy largely arises from the use of these systems in battery-operated portable (wearable) platforms. Accordingly, the goal of low-power design for battery-powered electronics is to extend the battery service life while meeting performance requirements. Unless optimizations are applied at different levels, the capabilities of future portable systems will be severely limited by the weight of the batteries required for an acceptable duration of service. In fixed, power-rich platforms, the packaging cost and power density/reliability issues associated with high power and high performance systems also force designers to look for ways to reduce power consumption. Thus, reducing power dissipation is a design goal even for non-portable devices since excessive power dissipation results in increased packaging and cooling costs as well as potential reliability problems. Meanwhile, following Moore's Law, integrated circuit densities and operating speeds have continued to go up in unabated fashion. The result is that chips are becoming larger, faster, and more complex and because of this, consuming increasing amounts of power.

These increases in power pose new and difficult challenges for integrated circuit designers. While the initial response to increasing levels of power consumption was to reduce the supply voltage, it quickly became apparent that this approach was insufficient. Designers subsequently began to focus on advanced design tools and methodologies to address the myriad of power issues. Complicating designers' attempts to deal with these issues are the complexities – logical, physical, and electrical – of contemporary IC designs and the design flows required to build them.

This article reviews a number of representative RT-level design automation techniques that focus on low power design. It should be of interest to designers of power efficient devices, IC design engineering managers, and EDA managers and engineers. More precisely, it covers techniques for, sequential logic synthesis, RT-level power management, multiple voltage design, and low power bus encoding techniques. Interested readers can find wide-ranging information on various aspects of low power design in [1]-[3]. Note that although, in many of today's designs, the leakage component of power consumption has become comparable to the dynamic component, this tutorial does not discuss the leakage issue. Interested readers may refer to any of the excellent references on leakage power, including those in the abovementioned edited books.

2 Multiple-Voltage Design

Using different voltages in different parts of a chip may reduce the global energy consumption of a design at a rather small cost in terms of algorithmic and/or architectural modifications. The key observation is that the minimum energy consumption in a circuit is achieved if all circuits paths are timing-critical (there is no positive slack in the circuit.) A common voltage scaling technique is thus to operate all the gates on non-critical timing paths of the circuit at a reduced supply voltage. Gates/modules that are part of the critical paths are powered at the maximum allowed voltage, thus, avoiding any delay increase; the power consumed by the modules that are not on the critical paths, on the other hand, is minimized because of the reduced supply voltage. Using different power supply voltages on the same chip of circuitry requires the use of level shifters at the boundaries of the various modules (a level converter is needed between the output of a gate powered by a low V_{DD} and the input of a gate powered by a high $V_{DD,}$ i.e., for a step-up change.) Figure 1 depicts a typical level converter design. Notice that if a gate that is supplied with $V_{DD,L}$ drives a fanout gate at $V_{DD,H}$, transistors N1 and N2 receive inputs at reduced supply and the cross-coupled PMOS transistors do the level conversion. Level converters are obviously not needed for a step-down change in voltage. Overhead of level converters can be mitigated by doing conversions at register boundaries and embedding the level conversion inside the flip flops (see [4] for details.)

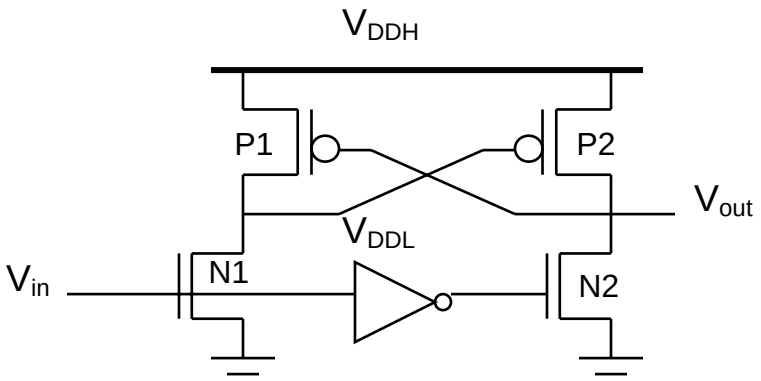


Figure 1: A typical level-converter design.

A polynomial time algorithm for multiple-voltage scheduling of performance-constrained non-pipelined designs is presented by Raje and Sarrafzadeh in [5]. The idea is to establish a supply voltage level for each of the operations in a data flow graph, thereby, fixing the latency of that operation. The goal is then to minimize the total power dissipation while satisfying the system timing constraints. Power minimization is in turn accomplished by ensuring that each operation will be executed using the minimum possible supply voltage. The proposed algorithm is composed of a loop where, in each iteration, slacks of nodes in the acyclic data flow graph are calculated. Then, nodes with the maximum slack are assigned to lower voltages in such a way that timing constraints are not violated. The algorithm stops when no positive slack exists in the data flow graph. Notice that this algorithm assumes that the Pareto-optimal voltage versus delay curve is identical for all computational elements in the data flow graph. Without this assumption, there is no guarantee that this algorithm will produce an optimal design.

In [6], the problem is addressed for combinational circuits, where only two supply voltages are allowed. A depth-first search is used to determine those computational elements, which can be operated at low supply voltage without violating the circuit timing constraints. A computational element is allowed to operate at $V_{DD,L}$ only is all its successors are operating at $V_{DD,L}$. For example, Figure 2(a) demonstrates a clustered voltage scaling (CVS) solution in which each circuit path starts with $V_{DD,H}$ and switches to $V_{DD,L}$ when delay slack is available. The timing-critical path is shown with thick line segments. Here gray-colored cells are running at $V_{DD,L}$. Level conversion (if necessary) is done in the flip flops at the end of the circuit paths. An extension to this approach is proposed in [7], which is based on the observation that by optimizing the insertion points of level converters, one can increase the number of gates using $V_{DD,L}$ without increasing the number of level converters. This leads to higher power savings. For example, in the CVS solution depicted in Figure 2(a), assume that the path delay from flip-flop FF3 to gate G2 is much longer than that of the path from FF1 to G2. In addition, assume that if we apply $V_{DD,L}$ to G2, then the path from FF3 to FF5 through G2 will miss its target combinational delay i.e., G2 must be assigned a supply level of $V_{DD,H}$. With the CVS approach, it immediately follows that G3 must be assigned $V_{DD,H}$ although a potentially large positive slack remains in the path from FF1 to G2. The situation is the same for G4 and G5. Consequently, the CVS approach can miss opportunities for applying $V_{DD,L}$ to

some gates in the circuit. If the insertion point of the level converter LC1 is allowed to move up to the interface between G3 and G2, the gates G3 through G5 can be assigned a supply of $V_{DD,L}$, as depicted in Figure 2(b). The structure shown there is one that can be obtained by the extended CVS (ECVS) algorithm. Both CVS and ECVS assign the appropriate power supply to the gates by traversing the circuit from the primary outputs to the primary inputs in a levelized order. ECVS allows a $V_{DD,L}$-driven gate to feed a $V_{DD,H}$ driven gate along with the insertion of a dedicated level converter.

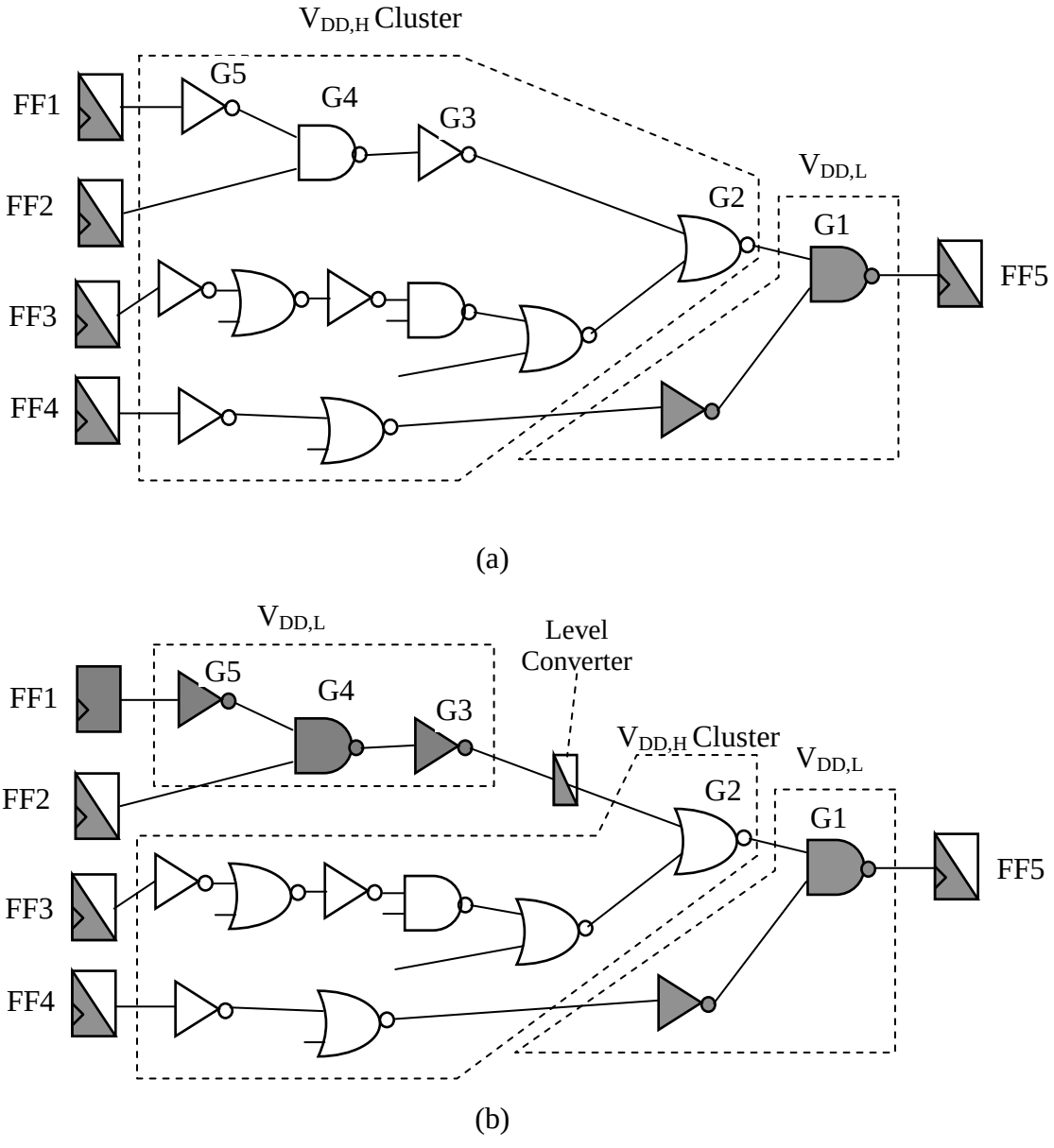


Figure 2: Examples of (a) CVS solution, (b) ECVS solution.

In [8], the authors propose an approach for voltage assignment in combinational logic circuits. First, a lower bound on dynamic power consumption is determined by exploiting

the available slacks and the value of the dual-supply voltages that may be used in solving the problem of minimizing dynamic power consumption of the circuit. Next, a heuristic algorithm is proposed for solving the voltage-assignment problem, where the values of the low and the high supply voltages are either specified by the user or fixed to the estimated ones.

In [9], the authors present resource and latency constrained scheduling algorithms to minimize power/energy consumption when the resources operate at multiple voltages. The proposed algorithms are based on efficient distribution of slack among the nodes in the data-flow graph. The distribution procedure tries to implement the minimum energy relation derived using the Lagrange multiplier method in an iterative fashion.

An important phase in the design flow of multiple-voltage systems is that of assigning the most convenient supply voltage, selected from a fixed number of values, to each operation in the control-date flow graph (CDFG). The problem is to assign the supply voltages and to schedule the tasks so as to minimize the power dissipation under throughput/resource constraints. An effective solution has been proposed by Chang and Pedram in [10]. The technique is based on dynamic programming and requires the availability of accurate timing and power models for the macro-modules in a RTL library. A preliminary characterization procedure must then be run to determine an energy-delay curve for each module in the library and for all possible supply-voltage assignments. The points on the curve represent various voltage assignment solutions with different tradeoffs between the performance and the energy consumption of the cell. Each set of curves is stored in the RTL library, ready to be invoked by the cost function that guides the multiple supply-voltage scheduling algorithm. We provide a brief description of the method for the simple case of control and data flow graphs (CDFG's) with a tree structure. The algorithm consists of two phases: first, a set of possible power-delay tradeoffs at the root of the tree is calculated; then, a specific macro-module is selected for each node in such a way that the scheduled CDFG meets the required timing constraints. To compute the set of possible solutions, a power-delay curve at each node of the tree (proceeding from the inputs to the output of the CDFG) is computed; such a curve represents the power-delay tradeoffs that can be obtained by selecting different instances of the macro-modules, and the necessary level shifters, within the subtree rooted at each specific node. The computation of the power-delay curves is carried out recursively, until the root of the CDFG is reached. Given the power-delay curve at the root node, that is, the set of tradeoffs the user can choose from, a recursive preorder traversal of the tree is performed, starting from the root node, with the purpose of selecting which module alternative should be used at each node of the CDFG. Upon completion, all the operations are fully scheduled; therefore, the CDFG is ready for the resource-allocation step.

More recently, a level-converter free approach is proposed in [11] where the authors try to eliminate the overhead imposed by level converters by suggesting a voltage scaling technique without utilizing level converters. The basic initiative is to impose some constraints on the voltage differences between adjacent gates with different supply voltages based on the observation that there will be no static current if the supply voltage of a driver gate is higher than the subtraction of the threshold voltage of a PMOS from

the supply voltage of a driven gate. In [12], the authors propose behavioral-level power optimization algorithms that use voltage and frequency scaling. In this work, the operators in a data flow graph are scheduled in the modules of the given architecture, by applying voltage and frequency scaling techniques to the modules of the architecture such that the power consumed by the modules is minimized. The global optimal selection of voltages and frequencies for the modules is determined through the use of an auction-theoretic model and a game theoretic solution. The authors present a resource constrained scheduling algorithm, which is based on applying the Nash equilibrium function to the game theoretic formulation.

3 RT-level Power Management

Digital circuits usually contain portions that are not performing useful computations at each clock cycle. Power reductions can then be achieved by shutting down the circuitry when it is idle.

3.1 Precomputation Logic

Precomputation logic, presented in [13], relies on the idea of duplicating part of the logic with the purpose of precomputing the circuit output values one clock cycle before they are required, and then uses these values to reduce the total amount of switching in the circuit during the next clock cycle. In fact, knowing the output values one clock cycle in advance allows the original logic to be turned off during the next time frame, thus eliminating any charging and discharging of the internal capacitances. Obviously, the size of the logic that pre-calculates the output values must be kept under control since its contribution to the total power balance may offset the savings achieved by blocking the switching inside the original circuit. Several variants to the basic architecture can then be devised to address this issue. In particular, sometimes it may be convenient to resort to partial, rather than global, shutdown, i.e., to select for power management only a (possibly small) subset of the circuit inputs.

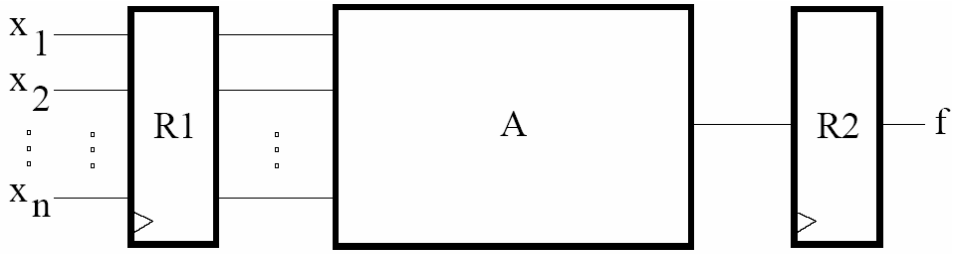


Figure 3: A pipeline stage of a data path.

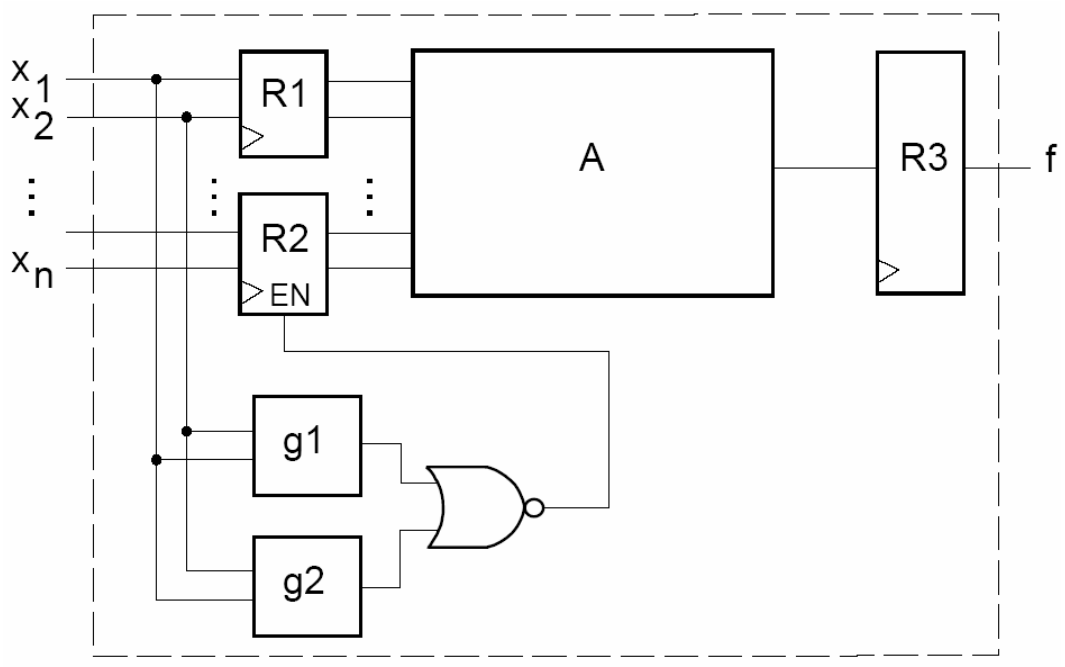


Figure 4: A precomputation logic realization of the pipeline stage (subset-input disabling architecture).

Figure 3 shows a combinational block A that implements an n-input, single-output Boolean function f, with registers R1 and R2 connected to its inputs and output pins, respectively. A precomputation architecture realization of this same logic block placed between register sets R1 and R2 is depicted in Figure 4. The key elements of the precomputation architecture are two n-input, single-output predictor functions g1 and g2, which satisfy the following constraints:

$$g1 = 1 \Rightarrow f = 1$$

$$g2 = 1 \Rightarrow f = 0$$

If, at the present clock cycle, g1 or g2 evaluate to one, then the load enable signal, LE, goes to zero, and the inputs to block A at the next clock cycle are forced to retain the current values. Hence, no gate output transitions inside block A occur, while the correct output value for the next time frame is provided by the two registers located on the outputs of g1 and g2. Note that the precomputation logic is a function of a subset of the input variables, hence, it is called a "subset input-disabling architecture."

The synthesis algorithm presented in [13] suffers from the limitation that if a logic function is dependent on the values of several inputs for a large fraction of the applied input combinations, then no reduction in switching activity can be obtained. In [14], the authors focus on a particular sequential precomputation architecture where the precomputation logic is a function of all of the input variables. The authors call this architecture the "complete input-disabling architecture." It is shown that the complete input disabling architecture can reduce power dissipation for a larger class of sequential circuits compared to the subset input-disabling architecture. The authors present an

algorithm to synthesize precomputation logic for the complete input-disabling architecture.

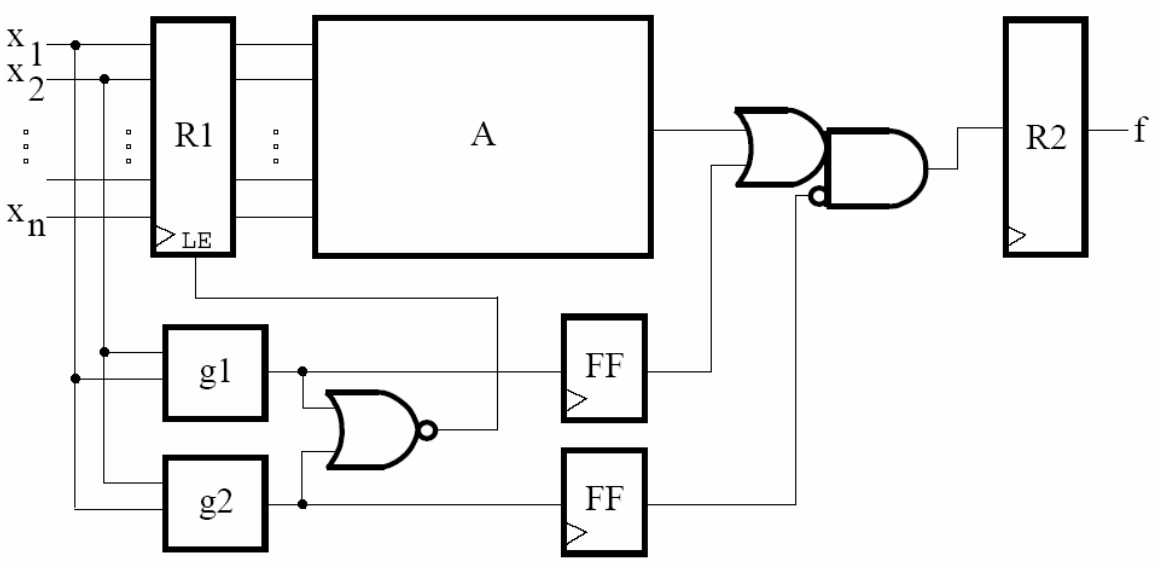


Figure 5: An example of a complete input-disabling precomputation architecture.

In Figure 5, a complete input-disabling precomputation architecture for a comparator circuit is shown. Functions g1 and g2 satisfy the conditions of (1) and (2) as before. During clock cycle t; if either g1 or g2 evaluates to a 1, the load enable signal of register R1 is set to be 0. This means that in clock cycle t + 1; none of the inputs to the combinational logic block A change. If g1 evaluates to 1 in clock cycle t, the input to register R2 is a 1 in clock cycle t+1, and if g2 evaluates to a 1, then the input to register R2 is a 0. Note that g1 and g2 cannot both be 1 during the same clock cycle, due to the conditions imposed by (1) and (2). The important difference between this architecture and the subset input-disabling architecture shown in Figure 4 is that the precomputation logic can be a function of all input variables, allowing us to precompute any input combination.

3.2 Clock Gating

Another approach to RT and gate-level dynamic power management, known as gated clocks [15]–[17], provides a way to selectively stop the clock, and thus, force the original circuit to make no transition, whenever the computation that is to be carried out at the next clock cycle is redundant. In other words, the clock signal is disabled according to the idle conditions of the logic network. For reactive circuits, the number of clock cycles in which the design is idle in some wait states is usually large. Therefore, avoiding the power waste corresponding to such states may be significant.

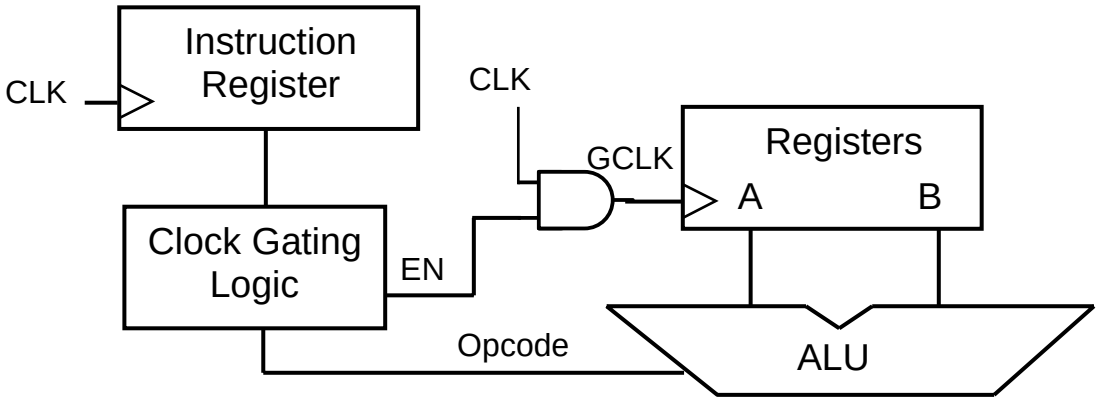


Figure 6: Clock gating logic for ALU in a typical processor microarchitecture with negative-edge triggered flip-flops.

The logic for the clock management is automatically synthesized from the Boolean function that represents the idle conditions of the circuit (cf. Figure 6.) It may well be the case that considering all such conditions results in additional circuitry that is too large and too power consuming. It may then be necessary to synthesize a simplified function, which dissipates the minimum possible power and stops the clock with maximum efficiency. The use of gated clocks has the drawback that the logic implementing the clock-gating mechanism is functionally redundant, and this may create major difficulties in testing and verification. The design of highly testable-gated clock circuits is discussed in [18].

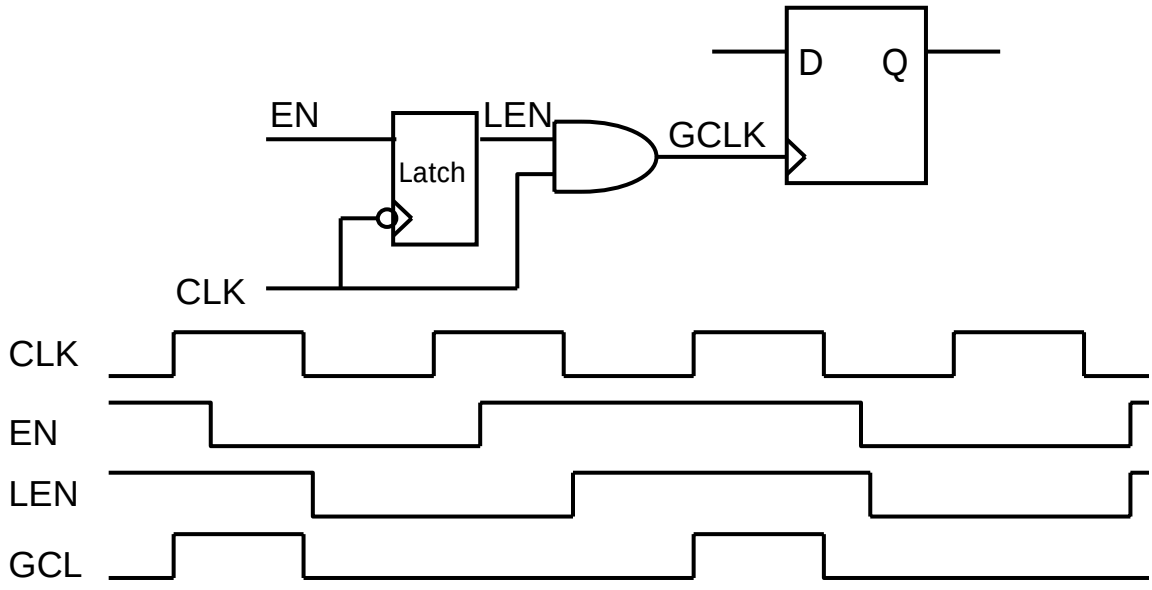


Figure 7: Clock is disabled when EN = 0; Furthermore, a hazard on EN will be stopped from reaching GCLK.

Another difficulty with clock gating is that one must stop hazards/glitches on EN signal from corrupting the clock signal to the register sets. This can be accomplished by

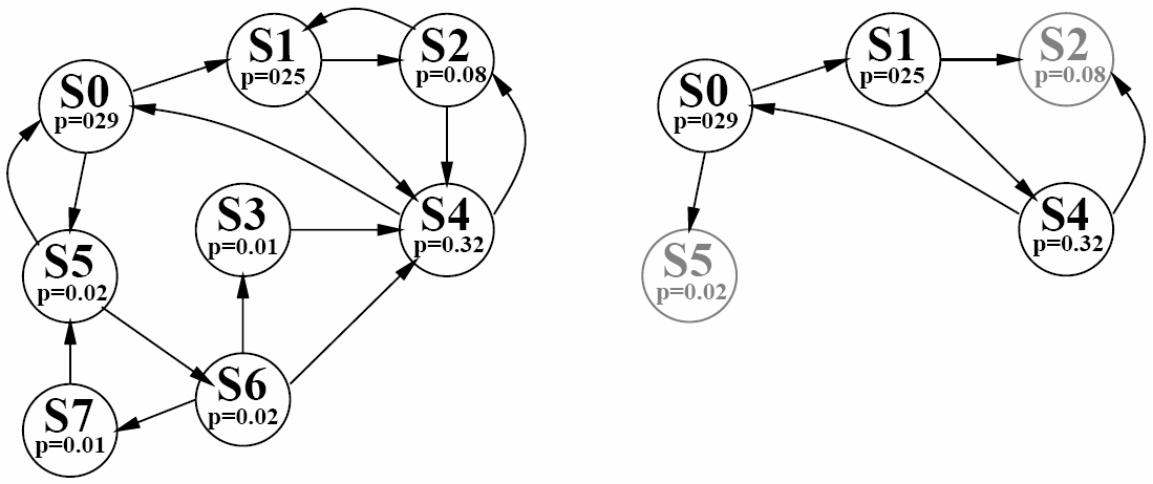
introducing a transparent negative latch between EN and the AND gate as shown in Figure 7.

3.3 Computational Kernels

Sequential circuits may have an extremely large number of reachable states, but during normal operation, these circuits tend to visit only a relatively small subset of the reachable states. A similar situation occurs at the primary outputs; while the circuit walks through the most probable states, only a few distinct patterns are generated at the combinational outputs of the circuit. Many researchers have proposed approaches for synthesizing a circuit that is fast and power-efficient under typical input stimuli, but continues to operate correctly even when uncommon input stimuli are applied to the circuit.

Reference [19] presents a power optimization technique by exploiting the concept of computational kernel of a sequential circuit, which is a highly simplified logic block that imitates the steady-state behavior of the original specification. This block is smaller, faster, and less power consuming than the circuit from which it is extracted and can replace the original network for a large fraction of the operation time.

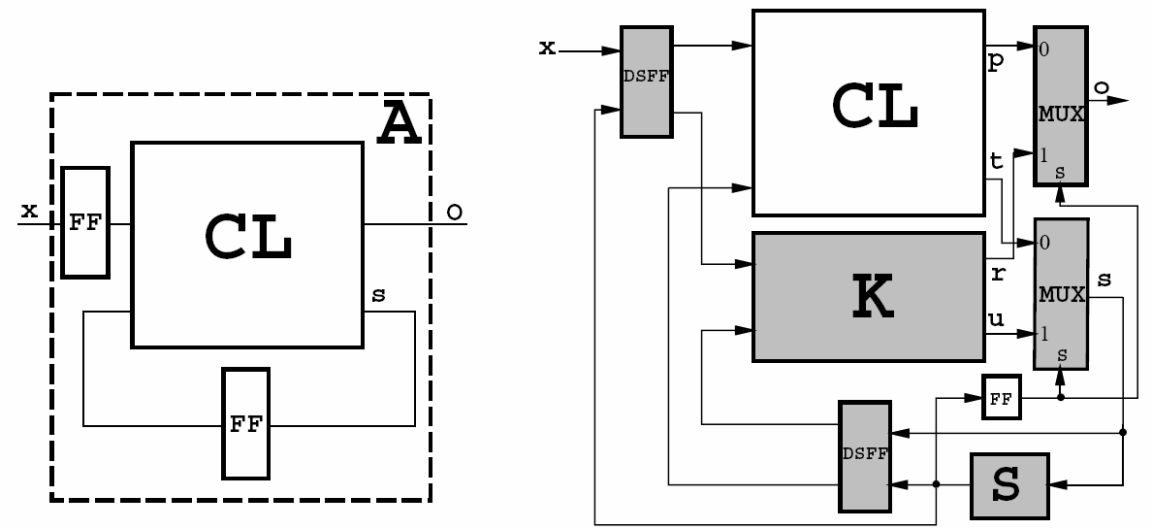
The *p*-order computational kernel of an FSM is defined with respect to a given probability threshold p and includes the subset of the states, S_P, of the original FSM whose steady-state occupation probabilities are larger than *p*. The combinational kernel also includes the subset of states, R_P, where for each state in R_p there is an edge from a state in S_p to that state. As an example, consider the simple FSM shown in Figure 8(a) in which the input and output values are omitted for the sake of simplicity and the states are annotated with the steady-state occupation probabilities calculated through Markovian analysis of the corresponding state transition graph (STG.) If we specify a probability threshold of p=0.25, then the computational kernel of the FSM is depicted in Figure 8(b). States in black represent set S_p, while states in grey represent R_p. The kernel probability is $\text{Prob}(S_p) = 0.29 + 0.25 + 0.32 = 0.86$.



(a) **(b)**

Figure 8: (a) Moore-type FSM and (b) its 0.25-order computational kernel.

Given a sequential circuit with the standard topology depicted in Figure 9(a), the paradigm for improving its quality with respect to a given cost function (e.g., power dissipation, latency) is based on the architecture shown in Figure 9(b).



(a) **(b)**

Figure 9: Kernel-based optimized architecture.

The basic elements of the architecture are: the combinational portion of the original circuit (block CL), the computational kernel (block K), the selector function (block S), the double state flip-flops (DSFF), and the output multiplexers (MUX.)

The computational kernel can be seen as a “dense" implementation of the circuit from which it has been extracted. In other terms, K implements the core functions of the original circuit, and because of its reduced complexity, it usually implements such functions in a faster and more efficient way. The purpose of selector function S is that of deciding what logic block, between CL and K, will provide the output value and the next-state in the following clock cycle. To take a decision, S examines the values of the next-state outputs at clock cycle n. If the output and next-state values in cycle n+1 can be computed by the kernel K, then S takes on the value 1. Otherwise, it takes on the value 0. The value of S is fed to a flip-flop, whose output is connected to the MUXes that select which block produces the output and the next-state. The optimized implementation is functionally equivalent to the original one. Computational kernels are a generalization of the precomputation architecture from combinational and pipelined sequential circuits to finite state machines. The authors in [19] proposed an algorithm for generating the computational kernel of a FSM by iterative simplification of the original network by redundancy removal.

In [20], the authors raise the level of abstraction at which the kernel-based optimization strategy can be exploited and show how RTL components for which only a functional specification is available can be optimized using the computational kernels. They present a technique for computational kernel extraction directly from the functional specification of a RTL module. Given the state transition graph (STG) specification, the proposed algorithm calculates the kernel exactly through symbolic procedures similar to those employed for FSM reachability analysis. The authors also provide approximate methods to deal with large STG's. More precisely, they propose two modifications to the basic procedure. The first one replaces the exact probabilistic analysis of the STG with an approximate analysis. In the second solution, symbolic state probability computation is bypassed and the set of states belonging to the kernel is determined directly from RTL simulation traces of a given (random or user-provided) stream.

3.4 State Machine Decomposition

Decomposition of finite state machines for low power has been proposed in [21]. The basic idea is to decompose the STG of a finite state machine (FSM) into two STGs that jointly produce the equivalent input-output behavior as the original machine. Power is saved because, except for transitions between the two sub-FSMs, only one of the sub-FSMs needs to be clocked. The technique follows a standard decomposition structure. The states are partitioned by searching for a small subset of states with high probability of transitions among these states and a low probability of transitions to and from other states. This subset of states will then constitute a small sub-FSM that is active most of the time. When the small sub-FSM is active, the other larger sub-FSM can be disabled. Consequently, power is saved because most of the time only the smaller, more power-efficient, sub-FSM is clocked.

In [22], the combinational logic block is partitioned (for example to CL1 and CL2) and the active part is decided based on the encoding of the present state. The states selected for one of the sub-FSMs (i.e., M1) are all encoded in such a way that the enable signal is always on for CL1 while it is off for CL2. Conversely, for all states in the other sub-FSM (i.e., M2), the enable signal is always off for CL1 while it is on for CL2. Consequently, for all transitions within M1, only CL1 will be active and vice-versa.

Consider as an example *dk27* FSM from the MCNC benchmark set, depicted in Figure 10. Assume that the input signal values, 0 and 1, occur with equal probabilities. The steady state probabilities which are shown next to the states in this figure have been computed accordingly. Suppose we partition the FSM into two sub-machines M1 and M2 along the dotted line. Then around 40% of the transitions occur in submachine M1, 40% of the transitions occur in submachine M2, and 20% of the transitions occur between sub-machines M1 and M2. Now suppose that the FSM is synthesized as two individual combinational circuits for sub-machines M1 and M2. Then we can turn off the combinational circuit for submachine M2 when transitions occur within submachine M1. Similarly, we can turn off the combinational circuit for submachine M1 when transitions occur within submachine M2. The states are partitioned such that the probability of transitions within any sub-FSM is maximized and the estimated overhead is minimized.

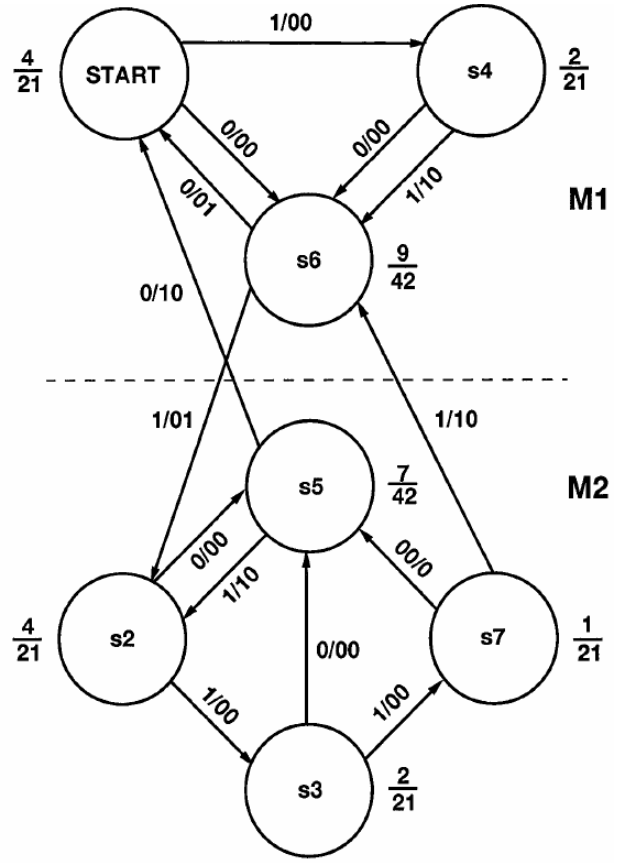


Figure 10: Example of an FSM (dk27) that may be decomposed into two sub-FSMs such that one sub-FSM can be shut off when the other is active and vice versa. .

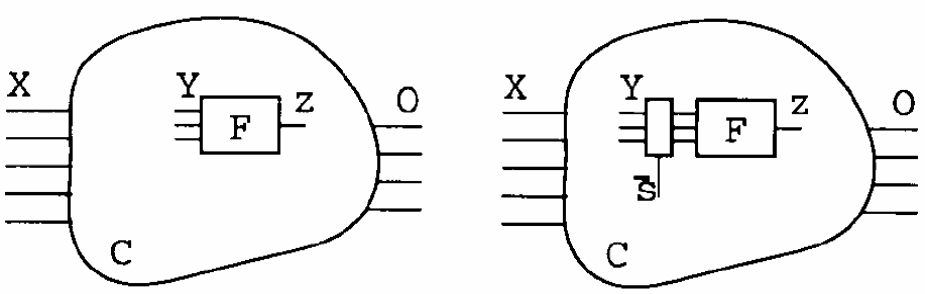
These methods for FSM decomposition can be considered as extensions of the gated-clock for FSM self-loops approach proposed in [23]. In FSM decomposition the cluster of states that are selected for one of the sub-FSMs can be considered as a "super-state" and then transitions between states in this cluster can be seen as self-loops on this "super-state".

3.5 Guarded Evaluation

Guarded evaluation [24] is the last RT and gate-level shutdown technique we review in this section. The distinctive feature of this solution is that, unlike precomputation and gated clocks, it does not require one to synthesize additional logic to implement the shutdown mechanism; instead, it exploits existing signals in the original circuit. The approach is based on placing some guard logic, consisting of transparent latches with an enable signal, at the inputs of each block of the circuit that needs to be power managed. When the block must execute some useful computation in a clock cycle, the enable signal makes the latches transparent. Otherwise, the latches retain their previous states, thus, blocking any transition within the logic block.

Guarded evaluation provides a systematic approach to identify where transparent latches must be placed within the circuit and by which signals they must be controlled. For

Example, Let C be a combinational logic block (cf. Figure 11(a)), X be the set of primary inputs to C, and z be a signal in C. Furthermore, let F be the portion of logic that drives z and Y be the set of inputs to F. Finally, let $D_Z(X)$ be the observability don't-care set for z (that is, the set of primary input assignments for which the value of z does not influence the outputs of C). Now consider a signal s in C which logically implies $D_Z(X)$, that is, $s \Rightarrow D_Z(X)$. Then, if s=1, then the value of z is not required to compute the outputs of C. If we call $t_e(Y)$ the earliest time at which any input to F can switch when s=1, and $t_l(s)$ as the latest time at which s settles to one, then signal s can be used as the guard signal for F (cf. Figure 11(b)) if $t_l(s) < t_e(Y)$. This is because z is not required to compute the outputs of C when s=1, and therefore, block F can be shut down. Notice that the condition $t_l(s) < t_e(Y)$ guarantees that the transparent latches in the guard logic are shut down before any of the inputs to F makes a transition.



(a) (b)

Figure 11: Example of guard logic insertion.

This technique, referred to as pure guarded evaluation, has the desirable property that when applied, no changes in the original combinational circuitry are needed. On the other hand, if some resynthesis and restructuring of the original logic is allowed, a larger number of logic shutdown opportunities may become available.

4 Sequential Logic Synthesis for Low Power

Power can be minimized by appropriate synthesis of logic. The goal in this case is to minimize the so-called switched capacitance of the circuit by low power driven logic minimization techniques.

4.1 State Assignment

State encoding/assignment, as a crucial step in the synthesis of the controller circuitry, has been extensively studied. Roy et al. was the first to address the problem of reducing switching activity of input state lines of the next state logic, during the state assignment, formulating it as a Minimum Weighted Hamming Distance problem [25]. Olson et al. used a linear combination of switching activity of the next state lines and the number of literals as the cost function [26]. Tsui et al. [27] used simulated annealing as a search strategy to find a low power state encoding that accounts for both the switching activity of the next state lines and switched capacitance of the next state and output logic.

For example, consider the state transition graph for a BCD to Excess-3 Converter depicted in Figure 12. Assume that the transition probabilities of the thicker edges in this figure are more than those of the thin edges. The key idea behind all of the low power state assignment techniques is to assign minimum Hamming distance codes to the states pairs that have large inter-state transition probabilities. For example the coding, S_0=000, S_1=001, S_2=011, S_3=010, S_4=100, S_5=101, S_6=111, S_7=110 fulfills this requirement.

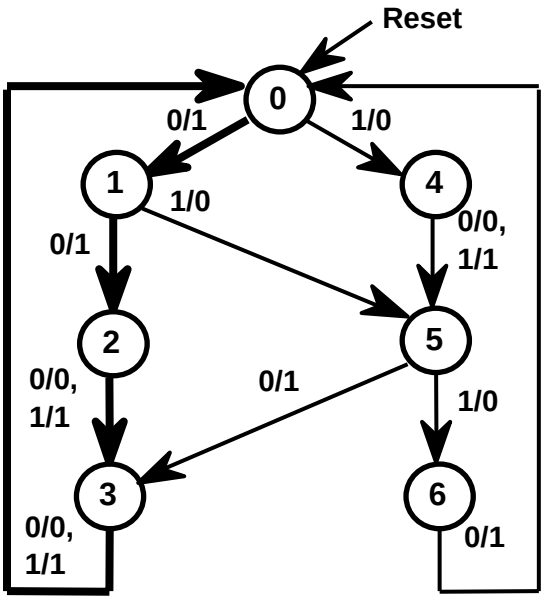


Figure 12: Excess-3 Converter state transition graph.

In [28], Wu et al. proposed the idea of realizing a low power FSM by using T flip-flops. The authors showed that use of T flip flops results in a natural clock gating and may result in reduced next state logic complexity. However, that work was mostly focused on BCD counters which have cyclic behavior. The cyclic behavior of counters results in a significant reduction of combinational logic complexity and, hence, lowers power consumption. Reference [29] introduces a mathematical framework for cycle representation of Markov processes and based on that, proposes solutions to the low power state assignment problem. The authors first identify the most probable cycles in the FSM and encode the states on these cycles with Gray codes. The objective function is to minimize the Weighted Hamming Distance. This reference also teaches how a combination of T and D flip-flops as state registers can be used to achieve a low power realization of a FSM.

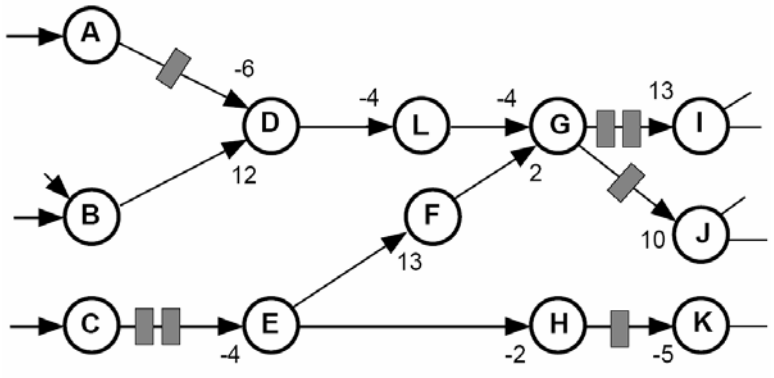
4.2 Retiming

Retiming is to reposition the registers in a design to improve the area and performance of the circuit without modifying its input-output behavior. The technique was initially proposed by Leiserson and Saxe [30]. This technique changes the location of registers in the design in order to achieve one of the following goals: 1) minimize the clock period; 2) minimizing the number of registers; or 3) minimize the number of registers for a target clock period.

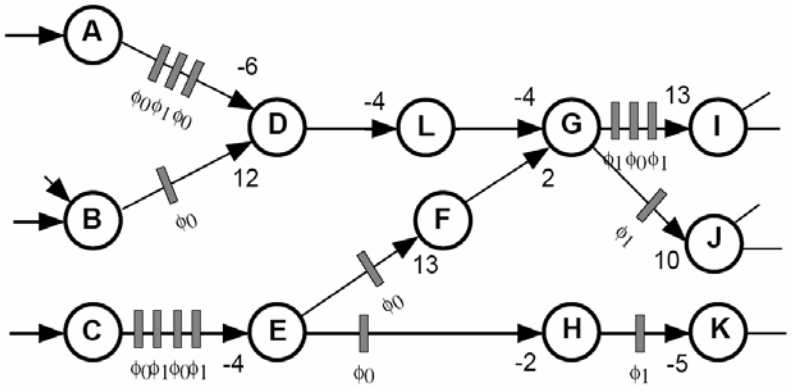
Minimizing dynamic power for synchronous sequential digital designs is addressed in the literature. In [31], Monteiro et al. presented heuristics to minimize the switching activity in a pipelined sequential circuit. Their approach is based on the fact that registers have to be positioned on the output edges of the computational elements that have high switching activity. The reason for power savings is that in this case the output of a register switches only at the arrival of the clock signal as opposed to potentially switching many times in the clock period. Consider the simple example of a logic gate belonging to a synchronous circuit and a capacitive load driven by the output gate. In CMOS technology, the power dissipated by gate is proportional to the product of the switching activity of the output node of the gate and the output load. At the output of gate some spurious transitions (i.e., glitches) may occur, which can result in a significant power waste. Suppose a register is inserted between the output of the gate and the capacitive load. In the new circuit, the output of the register can make, at most, one transition per clock cycle. In fact, the gate output may have many redundant transitions but they are all filtered out by the register; hence, these logic hazards do not propagate to the output load.

The heuristic retiming technique of [31] applies to a synchronous network with pipeline structure. The basic idea is to select a set of candidate gates in the circuit such that if registers are placed at their outputs, the total switching activity of the network gets minimized. The selection of the gates is driven by two factors: the amount of glitching that occurs at the output of each gate and the probability that such glitching propagates to the gates located in the transitive fanout. Registers are initially placed at the primary inputs of the circuit, and backward retiming (which consists of moving one register from all gate inputs to the output) is applied until all the candidate gates have received a register on their outputs. Then, registers that belong to paths not containing any of the candidate gates are repositioned, with the objective of minimizing both the delay and the total number of registers in the circuit. This last retiming phase does not affect the registers that have been already placed at the outputs of the previously selected gates. In [32], fixed-phase retiming is proposed to reduce dynamic power consumption. The edge-triggered circuit is first transformed to a two-phase level-clocked circuit, by replacing each edge-triggered flip-flop by two latches. Using the resulting level-clocked circuit, the latches of one phase are kept fixed, while the latches belonging to the other phase are moved onto wires with high switching activity and loading capacitance.

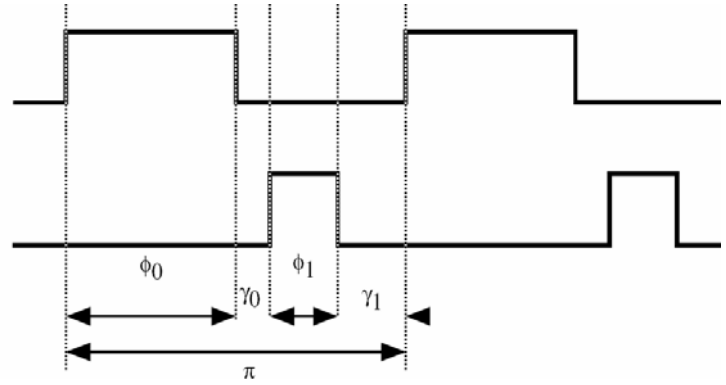
Fixed-phase retiming is best illustrated by the example shown below. Figure 13(a) shows a section of a pipelined circuit with edge-triggered flip-flops. The numbers on the edges represent the potential reduction in power dissipation when an edge-triggered flip-flop is present on that edge, assuming that the rest of the circuit remains unchanged. Negative values of power reduction indicate an increase in power dissipation when a flip-flop is placed on an edge. This reduction in power dissipation can be achieved if the edge has a high glitching-capacitance product [3]. After replacing each edge-triggered flip-flop by two back-to-back level-clocked latches, the resulting circuit is fixed-phase retimed to obtain the circuit in Figure 13(b).



(a)



(b)



(c)

Figure 13: Illustration of fixed-phase retiming. (a) Initial edge-triggered circuit. (b) Fixed-phase retimed circuit. (c) A two-phase clocking scheme $\pi = \langle \phi_0 = 4, \gamma_0 = 1, \phi_1 = 4, \gamma_1 = 1 \rangle$.

Assuming a non-overlapping two-phase clocking scheme $\pi = \langle \phi_0 = 4, \gamma_0 = 1, \phi_1 = 4, \gamma_1 = 1 \rangle$ such as the one shown in Figure 13(c), power dissipation can be reduced by 11.8 units. Specifically, the glitching on edges B→D, E→F and E→H is "masked" for 60% of the clock cycle which decreases power dissipation by 0.6×(12 + 13 -2) = 13.8 units of power. At the same time, the glitching on edges G→J and H→K is "exposed" for 40% of the clock cycle which increases power dissipation by 0.4×(10 – 5) = 2 power units. In order to simplify the computation of changes in power dissipation for this example, it is assumed that glitching is uniformly distributed over the entire clock period and that the relocation of latches does not change glitching significantly.

In [33], the authors propose a hybrid retiming and supply voltage scaling. They observe that critical paths are related to the position of registers in a design so they try not only to scale down the supply voltage of computational elements that are off the critical paths, but also to move registers to maximize the number of computational elements that are off the critical paths, thereby further minimizing the circuit power consumption. Registers have to be moved from their positions by the standard retiming technique. Instead of unifying basic retiming and supply voltages scaling, the authors propose to apply "guided retiming" followed by the application of voltage scaling on the retimed design. Polynomial time algorithms based on dynamic programming to realize the guided retiming as well as the supply voltage scaling on the retimed design are proposed.

5 Bus Encoding for Low Power

A lot of power is consumed in the on-chip and off-chip busses in a VLSI circuit. These buses, which connect various internal blocks of the circuit or connect the circuit to the external environment, have large capacitive loads and high transition counts. Power on these buses can be reduced by properly coding the data and/or address bus values so as to minimize the number of transitions that occur on the bus.

Musoll, et al. proposed the working zone method in [34]. Their method takes advantage of the fact that data accesses tend to remain in a small set of working zones. For the addresses that lie in each of these zones, a relatively high degree of locality is observed. Each working zone requires a dedicated register called zone register that is used to keep track of the accesses in that zone. When a new address arrives, the offset of the address is calculated with respect to all zone registers. The address is, thus, mapped to the working zone with the smallest offset. If the offset is sufficiently small, one-hot encoding is performed and the result is sent on the bus using transition signaling (by transition signaling we mean that instead of sending the code itself we XOR it with the previous value of the bus). Otherwise, the address itself is sent over the bus. The working zone method uses one extra line to show whether encoding has been done or the original value has been sent. It also uses additional lines to identify the working zone that was used to compute the offset. Based on this information, the decoder on the other side of the bus can uniquely decode the address.

The working zone method also has the ability to detect a stride in any of the working zones. A stride is a constant offset that occurs between multiple consecutive addresses repeatedly and if detected, can be used to completely eliminate the switching activity for such addresses. For instruction addresses, stride corresponds to the offset of sequential instructions. Stride is very important when instruction address encoding is tackled. In fact, the large number of sequential instructions with constant stride is the foundation of considerable transition savings that is usually seen in instruction address encoding techniques. For data addresses, stride can occur when, for example, a program is accessing elements of an array in the memory. Except for special cases, detecting and utilizing strides has a very small impact on decreasing the switching activity of data addresses.

Another encoding method that can be used for data addresses is the bus-invert method [35]. The bus-invert selects between the original and the inverted pattern in a way that minimizes the switching activity on the bus. The resulting patterns together with an extra bit (to notify whether the address or its complement has been sent) are transition signaled over the bus (cf. Table I, column 4.) This technique is quite effective for reducing the number of one's in addresses with random behavior, but it is ineffective when addresses exhibit some degree of locality. To make the bus-invert method more effective, the bus can be partitioned into a handful of bit-level groups and a bus-invert can be separately applied to each of these groups. However, this scheme will increase the number of surplus bits required for the encoding, which is absolutely undesirable.

TABLE I- EXAMPLE SHOWING THE T0, BI, AND T0-BI CODES

Address (Hex)	Source word	T0 Code word	BI Code word	T0-BI Code word
31	0011 0001	0 0011 0001	0 0011 0001	00 0011 0001
32	0011 0010	1 0011 0001	0 0011 0010	10 0011 0001
33	0011 0011	1 0011 0001	0 0011 0011	10 0011 0001
C2	1100 0010	0 1100 0010	1 0011 1101	01 0011 1101
C3	1100 0011	1 1100 0010	1 0011 1100	11 0011 1101
C4	1100 0100	1 1100 0010	1 0011 1011	11 0011 1101
C2	1100 0010	0 1100 0010	1 0011 1101	01 0011 1101
C3	1100 0011	1 1100 0010	1 0011 1100	11 0011 1101
C4	1100 0100	1 1100 0010	1 0011 1011	11 0011 1101
	Tr. Cnt = 19	Tr. Cnt = 11	Tr. Cnt = 16	Tr. Cnt = 9

In [36], Benini et al. proposed the T0 code, which exploits data sequentiality to reduce the switching activity on the address bus. The observation is that addresses are sequential except when control flow instructions are encountered or exceptions occur. T0 adds a redundant bus line, called INC. If the addresses are sequential, the sender freezes the value on the bus and sets the INC line. Otherwise, INC is de-asserted and the original address is sent (cf. Table I, column 3.) Several methods that are combinations of the Bus-Invert and T0 encodings were proposed in [37]. For instance, one of the introduced methods called T0-BI, adds two redundant bits, named INV and INC to the bus. If the addresses are sequential, T0 encoding is applied and the bus is frozen; otherwise, the new

address, which is not sequential, is encoded based on the Bus-Invert coding. INC and INV bits are used to correctly decode the bus value on the receiver side (cf. Table I, column 5.)

The major drawback of the encoding methods introduced in this work is that they introduce redundant bits. In T0 code, one extra bit was used to identify between these two cases in the receiver. Aghaghiri et al [38] improved on this technique by eliminating the redundant bit in T0-concise. The idea is to send previous source plus stride if the bus value is equal to the current non-sequential source word. This is the only thing that the receiver does not expect, therefore, it can correctly decode it as a jump back to the address that was frozen on the bus at the beginning of the current sequential access.

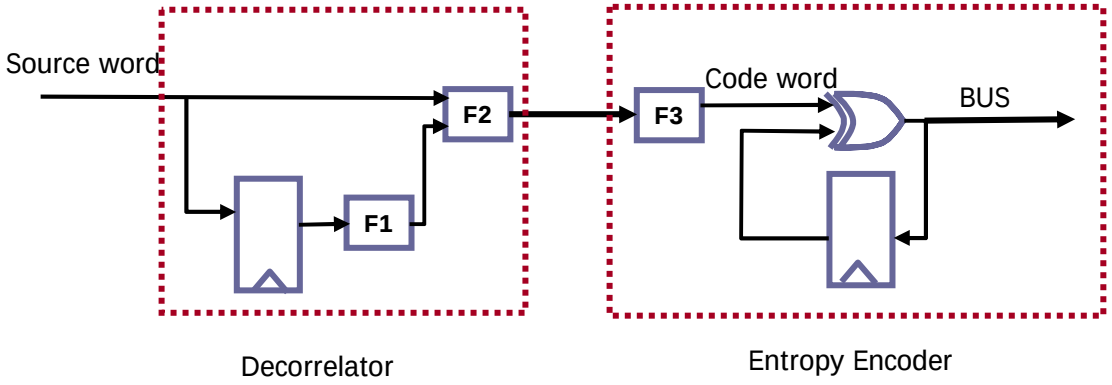


Figure 14: The block diagram of a generic low power encoder.

Reference [39] proposes a low-power coding framework for address and data buses. They describe the general architecture of a low power encoder (cf. Figure 14.) In this Figure, choices for function F1 include identity or increment transformations, choices for F2 include XOR operation, subtraction, or difference-based mapping, and choices for F3 are inversion or probability-based mapping. For example, the INC-XOR encoder, also known as T0-XOR, generates the new bus value as the XOR of the previous bus value and the new code word (this is known as the transition signaling over the bus.) The new code word is in turn obtained as the XOR of the new source word and the summation of the previous source word and the stride value. Obviously, when consecutive addresses grow by the stride, no transitions will occur on the bus. The Offset-XOR encoder also relied on transition signaling. However, the new code word is obtained as the new source word minus the summation of the previous source word and the stride value. In [40], Aghaghiri and Pedram presented the Offset-XOR-SM encoding whereby the new code word is again transition signaled over the bus. The new code word itself is generated as LSB-Invert of the Offset-XOR code word followed by a codebook-based mapping. The LSB-Invert function is a simple mapping function that reduces number of one's in the binary representation of small negative numbers (LSB-INV(X)=if(X>0) X; otherwise X $\oplus$ (2^{N-1} - 1). The codebook maps small offsets (say up to 10 bits) to K-limited codes in order to reduce the number of 1's in the new code word (recall that a 1 in the code word translates to a bit-level activity after transition signaling the code word over the bus.)

In [41], Mamidipaka, et al. proposed an encoding technique based on the notion of self-organizing lists. They use a list to create a one-to-one mapping between addresses and codes. The list is reorganized in every clock cycle to map the most frequently used addresses to codes with fewer one's. For multiplexed address buses, the authors used a combination of their method and INC-XOR. The size of the list in this method has a significant impact on the performance. To achieve satisfactory results, it is necessary to use a long list. However, the large hardware overhead associated with maintaining long lists makes this technique quite expensive. Furthermore, the encoder and the decoder hardware are practically complex and their power consumption appears to be quite large.

In [42], the authors introduced a class of irredundant low power techniques for encoding instruction or data source words before they are transmitted over buses. The key idea is to partition the source word space into a number of sectors with unique identifiers called *sector heads* (SH). These sectors can, for example, correspond to address spaces for the code, heap, and stack segments of one or more application programs. Each source word is then mapped to the appropriate sector and is encoded with respect to the sector head. Suppose X is an N-bit source word to be encoded. There are 2^k fixed sectors with 2^k sector heads, SH[0]…SH[2^k-1]. The code word is comprised of k most significant Sec-ID bits used for identifying the sector, and N-k least significant difference bits representing the XOR difference between the source word and the corresponding sector head. The encoder takes a source word X: (X_N…X_1), and assigns it to the corresponding sector by examining its Sec-ID bits. Next, it sets the N-k LSB's of the code word to the XOR difference between the N-k LSB's of the source word and the corresponding bits of the SH for the identified sector. The SH of the identified sector is set to X. Finally, the code word is transition signaled over the bus. As an example, consider a 5-bit space with 4 sector heads initialized at equal distances from each other i.e., {00000,01000,10000,11000}. Table II shows the results of the fixed four-sector encoder.

TABLE II- EXAMPLE SHOWING A FIXED 4-SECTOR ENCODER

X	SH[0]	SH[1]	SH[2]	SH[3]	Code(X)
01111	00 000	01 000	10 000	11 000	01111
00010	00 000	01 111	10 000	11 000	00010
00011	00 010	01 111	10 000	11 000	00001
01110	00 011	01 111	10 000	11 000	01001
10001	00 011	01 110	10 000	11 000	10001
00011	00 011	01 110	10 001	11 000	00000
01100	00 011	01 110	10 001	11 000	01010
Tr. Cnt=12	00 011	01 100	10 001	11 000	Tr. Cnt=8

Note that the sectors are fixed, but the sector heads are dynamically updated. In a generalization of this approach the sectors can dynamically be defined based on program behavior. This feature is very useful because the source word space is very large while the total working zone of a program is usually small. Therefore, it pays off to have dynamically defined sectors which can "zoom into" the working zone of a program. The sector-based encoding techniques are quite effective in reducing the number of inter-pattern transitions on the, bus while incurring rather small power and delay overheads.

In [43], the authors provide a modified bus-invert (MBI) technique which besides reducing delay and power also minimizes the crosstalk noise that results from inductive coupling between the bus lines. Their proposed approach is based on the observation that opposite skews can reduce the crosstalk noise. Therefore, the authors propose to minimize the number of transitions that are in the same direction by selectively inverting the data patterns. The method requires an extra line which carries the "invert signal" and is used by the decoder in order to restore the original data. In the encoder the bus lines are partitioned into pairs and each pair of adjacent lines as well as their values from the previous clock cycle drive the inputs of a logic cell which encodes the types of events that occur on the pair of bus lines. This cell generates 11 if the transitions occur in the same direction, 00 if both lines are idle, 01 if either only one line switches, or both lines switch in opposite directions. A majority voter takes the outputs of the logic cells and the previous invert signal and sets the invert signal to 0 when the count of 1's is less than n and to 1 otherwise.

6 Conclusion

Several key elements emerge as enablers for an effective low power design methodology. The first is the availability of accurate, comprehensive power models. The second is the existence of fast, easy to use high level estimation and design exploration tools for analysis and optimization during the design creation process, while the third is the existence of highly accurate, high capacity verification tools for tape-out power verification. As befitting a first-order concern, successfully managing the various power-related design issues will require that power be addressed at all phases and in all aspects of design, especially during the earliest design and planning activities. Advanced power tools will play central roles in these efforts.

This paper reviewed a number of techniques for low power design of VLSI circuits including RT level synthesis, bus encoding and voltage scaling. Emphasis was placed on runtime power management techniques and sequential circuit synthesis. A review of techniques for low power design of combinational logic circuits can be found in many references, including [44], [45].

References

[1] M. Pedram and J. Rabaey (editors), *Power Aware Design Methodologies*, Kluwer Academic Publishers, Boston, 2002.

[2] E. Macii (editor), *Ultra Low-Power Electronics and Design*, Kluwer Academic Publishers, Boston, 2004.

[3] C. Piguet (editor), *Low Power Electronics Design*, The CRC Press, 2004.

[4] M. Hamada, M. Takahashi, H. Arakida, A. Chiba, T. Terazawa, T. Ishikawa, M. Kanazawa, M. Igarashi, K. Usami, and T. Kuroda,"A top-down low power design technique using clustered voltage scaling with variable supply-voltage scheme," in *Proc. IEEE Custom Integrated Circuits Conference* (CICC'98), May 1998, pp. 495-498,.

[5] S . Raje and M. Sarrafzadeh, "Variable voltage scheduling," in *Proc. Int'l. Workshop Low Power Design*, Aug. 1995, pp. 9–14.

[6] K. Usami and M. Horowitz, “Clustered voltage scaling technique for low-power design,” in *Proc. Int’l. Workshop Low Power Design*, 1995, pp. 3–8.
[7] Usami, K. Igarashi, M. Minami, F. Ishikawa, T. Kanzawa, M. Ichida, M. Nogami, K. “Automated low-power technique exploiting multiple supply voltages applied to a media processor," *IEEE Journal of Solid-State Circuits*, vol. 33, no. 3, Mar. 1998, pp 463 – 472.
[8] C. Chen, A. Srivastava, and M. Sarrafzadeh, “On gate level power optimization using dual supply voltages,” *IEEE Trans. on VLSI Systems*, vol. 9, Oct. 2001, pp. 616–629.
[9] A. Manzak and C. Chakrabarti, “A Low Power Scheduling Scheme with Resources Operating at Multiple Voltages,” IEEE Trans. on VLSI Systems, vol. 10, no. 1, Feb. 2002, pp. 6-14.
[10] J. M. Chang and M. Pedram, “Energy minimization using multiple supply voltages,” *IEEE Trans. VLSI Systems*, vol. 5, no. 4, 1997, pp. 436–443.
[11] Y.-J. Yeh, S.-Y. Kuo, and J.-Y. Jou, “Converter-free multiple-voltage scaling techniques for low-power CMOS digital design,” *IEEE Trans. Computer-Aided Design*, vol. 20, Jan. 2001, pp. 172–176.
[12] A. K. Murugavel, N. Ranganathan, “Game Theoretic Modeling of Voltage and Frequency Scaling during Behavioral Synthesis,” in *Proc. of VLSI Design*, 2004, pp. 670-673.
[13] M. Alidina, J. Monteiro, S. Devadas, A. Ghosh, and M. Papaefthymiou, “Precomputation-based sequential logic optimization for low power,” *IEEE Trans. VLSI Systems*, vol. 2, no. 4, 1994, pp. 426–436.
[14] J. Monteiro, S. Devadas, A. Ghosh, “Sequential Logic Optimization For Low Power Using Input-disabling,” *IEEE Trans. on Computer-Aided Design*, vol. 17, no. 3, 1998, pp. 279–284.
[15] L. Benini, P. Siegel, and G. De Micheli, “Automatic synthesis of gated clocks for power reduction in sequential circuits,” *IEEE Design Test Computer Magazine*, vol. 11, no. 4, pp. 32–40, 1994.
[16] L. Benini and G. De Micheli, “Transformation and synthesis of FSM’s for low power gated clock implementation,” *IEEE Trans. on Computer-Aided Design*, vol. 15, no. 6, 1996, pp. 630–643.
[17] L. Benini, G. De Micheli, E. Macii, M. Poncino, and R. Scarsi, “Symbolic synthesis of clock-gating logic for power optimization of control-oriented synchronous networks,” in *Proc. European Design and Test Conf.*, Paris, France, Mar. 1997, pp. 514–520.
[18] L. Benini, M. Favalli, and G. De Micheli, “Design for testability of gated-clock FSM’s,” in *Proc. European Design and Test Conf.*, Paris, France, Mar. 1996, pp. 589–596.
[19] L. Benini, G. De Micheli, A. Lioy, E. Macii, G. Odasso, and M. Poncino, “Synthesis of Power-Managed Sequential Components Based on Computational Kernel Extraction,” *IEEE Trans. on Computer-Aided Design*, vol. 20, no. 9, September 2001, pp. 1118-1131.
[20] L Benini, G. De Micheli, E. Macii, G. Odasso, M. Poncino, “Kernel-Based Power Optimization of RTL Components: Exact and Approximate Extraction Algorithms,” in *Proc. of Design Automation Conf.*, 1999, pp.247-252.

[21] J. Monteiro and A. Oliveira. Finite State Machine Decomposition for Low Power. In *Proc. of Design Automation Conference*, June 1998, pages 758-763.
[22] S-H. Chow, Y-C. Ho, and T. Hwang. "Low Power Realization of Finite State Machines A Decomposition Approach," *ACM Trans. on Design Automation of Electronic Systems*, vol. 1 no. 3, July 1996, pp.315-340,.
[23] L. Benini, P. Siegel, and G. De Micheli. Automatic Synthesis of Low-Power Gated-Clock Finite-State Machines. *IEEE Trans. on Computer-Aided Design*, 15(6):630643, June 1996.
[24] V. Tiwari, S. Malik, and P. Ashar, "Guarded evaluation: Pushing power management to logic synthesis/design," in *Proc. ACM/IEEE Int'l. Symp. Low Power Design*, Dana Point, CA, Apr. 1995, pp. 221–226.
[25] K. Roy and S. Prasad, "Syclop: Synthesis of CMOS Logic for Low-Power Application," *Proc. of Int'l Conf. on Computer design*, pp. 464-467, Oct. 1992.
[26] E. Olson and S. M. Kang, "Low-Power State Assignment for Finite State Machines," in *Proc. of Int'l Workshop on Low Power Design*, pp. 63-68, April 1994.
[27] C. Y. Tsui, M. Pedram and A. M. Despain, "Low-Power State Assignment Targeting Two- and Multilevel Logic Implementation," *IEEE Trans. on Computer-Aided Design*, vol. 17, no. 12, Dec. 1998, pp. 1281-1291.
[28] X. Wu, J. Wei, Q. Wu, and M. Pedram, "Low-Power Design of Sequential Circuits Using a Quasi-Synchronous Derived Clock," *Int'l Journal of Electronics*, Taylor and Francis Publishing Group, vol. 88, no. 6, Jun. 2001, pp. 635-643.
[29] A. Iranli, P. Rezvani, and M. Pedram, "Low power synthesis of finite state machines with mixed D and T flip-flops," in *Proc. of Asia and South Pacific Design Automation Conference*, Jan. 2003 pp. 803-808.
[30] C. E. Leiserson and J. B. Saxe, "Optimizing synchronous systems," *Journal of VLSI Computer Systems*, vol. 1, no. 1, 1983, pp. 41–67.
[31] J. Monteiro, S. Devadas, and A. Ghosh, "Retiming sequential circuits for low power," in *Proc. Int'l. Conf. Computer-Aided Design*, Santa Clara, CA, Nov. 1993, pp. 398–402.
[32] K. N. Lalgudi and M. Papaefthymiou, "Fixed-phase retiming for low power," in *Proc. Int'l. Symp Low-Power Electronics and Design*, 1996, pp. 259–264.
[33] N. Chabini and W. Wolf, "Reducing Dynamic Power Consumption in Synchronous Sequential Digital Designs Using Retiming and Supply Voltage Scaling" *IEEE Trans. on VLSI Systems*, vol. 12, no. 6, June 2004, pp.573-589.
[34] E. Musoll, T. Lang, and J. Cortadella, "Exploiting the Locality of Memory References to Reduce the Address Bus Energy," in *Proc. Int'l Symp. on Low Power Electronics and Design*, 1997, pp. 202-207.
[35] M. R. Stan, W. P. Burleson, "Bus-Invert Coding for Low Power I/O," *IEEE Trans. on VLSI Systems*, vol. 3, no. 1, Mar. 1995, pp. 49-58.
[36] L. Benini, G. De Micheli, E. Macii, D. Sciuto, C. Silvano, "Asymptotic Zero-Transition Activity Encoding for Address Buses in Low-Power Microprocessor-Based Systems," *Proc. Seventh Great Lakes Symposium on VLSI*, pp. 77-82, Mar. 1997.

[37] 4. L. Benini, G. De Micheli, E. Macii, D. Sciuto, and C. Silvano, "Address Bus Encoding Techniques for System-Level Power Optimization," *Proc. Design Automation and Test in Europe*, pp. 861-866, 1998.
[38] Y. Aghaghiri, F. Fallah and M. Pedram, "A class of irredundant encoding techniques for reducing bus power," Special Issue on Low Power Design in *Journal of Circuits, Systems, and Computers*, World Scientific Publishers, Vol. 11, No. 5 (2002), pp. 445-457.
[39] S. Ramprasad, N. Shanbhag, I. Hajj, "A Coding Framework for Low Power Address and Data Busses," *IEEE Trans. on VLSI Systems*, vol. 7, no. 2, Jun. 1999, pp. 212-221.
[40] Y. Aghaghiri, F. Fallah, and M. Pedram, "Irredundant address bus encoding for low power," *Proc. of Symp. on Low Power Electronics and Design*, Aug. 2001, pp. 182-187.
[41] M. Mamidipaka, D. Hirschberg, N. Dutt, "Low Power Address Encoding Using Self-Organizing Lists," in *Proc. Intl. Symp. on Low Power Electronics and Design*, 2001, pp. 188-193.
[42] Y. Aghaghiri, F. Fallah, and M. Pedram, "Transition reduction in memory buses using sector-based encoding techniques," *IEEE Trans. on Computer-Aided Design*, Vol. 23, No. 8, Aug. 2004, pp. 1164-1174.
[43] M. Lampropoulos, B.M. Al-Hashimi, P.M. Rosinger, "Minimization of Crosstalk Noise, Delay and Power Using a Modified Bus Invert Technique," Proc. Design Automation and Test in Europe, 2004, pp. 1372-1373.
[44] J. Rabaey and M. Pedram (editors), *Low Power Design Methodologies*, Kluwer Academic Publishers, Boston, Oct. 1995.
[45] M. Pedram, "Power minimization in IC design: principles and applications," A*CM Trans. on Design Automation of Electronic Systems*, vol. 1, no. 1, 1996, pp. 3-56.