Review in Low Power VLSI Design

S.H.Prasad
AEC,ECE, Surampalem
Hariprasad.satti@aec.edu.in

G.Rama Naidu
AEC,ECE, Surampalem
Ramanaidu.gangu@aec.edu.in

G.Ajay Shankar
AEC,ECE, Surampalem
Ajay.sankar1989@gmail.com

S.B.G.Tilak Babu
AEC,ECE, Surampalem
thilaksayila@gmail.com

Abstract: The recent trends in the developments and advancements in the area of low power VLSI Design are surveyed in this paper. Though Low Power is a well established domain, it has undergone lot of developments from transistor sizing, process shrinkage, voltage scaling, clock gating, etc., to adiabatic logic. This paper aims to elaborate on the recent trends in the low power design.

Key words: Multi threshold, dynamic voltage and frequency scaling, split level charge recovery logic, efficient charge recovery logic, positive feedback adiabatic logic, pre-resolve and sense adiabatic logic.

1. Introduction

1.1. Classification of Power Consumption

Though there are different types of power consumption, the major types that affect CMOS circuits are dynamic power and leakage power [1].

1.1.1. Dynamic power

Dynamic power [2] is the power that is consumed by a device when it is actively switching from one state to another [3]. Dynamic power consists of switching power consumed while charging and discharging the loads on a device, and internal power (also referred to as short circuit power), consumed internal to the device while it is changing state [4]. Fig. 1 shows the dynamic power dissipation that can occur in CMOS circuits.

![Dynamic power reduction](image-url)
1.1.2. Leakage power

Leakage power is the power consumed by a device not related to state changes [2]. Leakage power is actually consumed when a device is both static and switching, but generally the main concern with leakage power is when the device is in its inactive state, as all the power consumed in this state is considered “wasted” power [3].

![Fig. 2. Causes of leakage power.](image)

Different causes for the leakage power like reverse bias current, subthreshold channel leakage current, drain induced barrier lowering leakage, gate induced drain leakage, punch through, narrow width effect, gate oxide tunneling current, and hot carrier injection current [2] are depicted in Fig. 2.

Various techniques have been developed to reduce both dynamic and leakage power. CMOS circuit dynamic power consumption equation is

\[ P = A C V^2 F_{CLK} \]

Where:
- \( P \) is the power consumed,
- \( A \) is the activity factor, i.e., the fraction of the circuit that is switching,
- \( C \) is the switched capacitance, \( V \) is the supply voltage, and \( F \) is the clock frequency. If a capacitance of \( C \) is charged and discharged by a clock signal of frequency \( F \) and peak voltage \( V \), then the charge moved per cycle is \( CV \) and the charge moved per second is \( CVF \). Since the charge packet is delivered at voltage \( V \), the energy dissipated per cycle, or the power, is

\[ \text{Power} = \text{Capacitive load} \times \text{Voltage}^2 \times \text{Clock Frequency} \]

The data power for a clocked flip-flop, which can toggle at most once per cycle, will be half of the stated power. When capacitances are clock gated or when flip-flops do not toggle every cycle, their power consumption will be lower. Hence, a constant called the activity factor (0 ≤ \( A \) ≤ 1) is used to model the average switching activity in the circuit.

2. Traditional Power Reduction Techniques

To minimize this power, Technology scaling, voltage scaling, clock frequency scaling, reduction of switching activity, etc., were widely used.

The two most common traditional, mainstream techniques are:
Clock Gating:

Clock gating is a technique which is shown in Fig. 3 for power reduction, in which the clock is disconnected from a device it drives when the data going into the device is not changing. This technique is used to minimize dynamic power.

Clock gating is a mainstream low power design technique targeted at reducing dynamic power by disabling the clocks to inactive flip-flops.

To save more power, positive or negative latch can also be used as shown in Fig. 4 and Fig. 5. This saves power in such a way that even when target device's clock is 'ON', controlling device's clock is 'OFF'. Also when the target device's clock is 'OFF', then also controlling device's clock is 'OFF'. In this more power can be saved by avoiding unnecessary switching at clock net [6].

Multi-Vth optimization/ (Multi Threshold - MTCMOS):

MTCMOS is the replacement of faster Low-\(V_t\) (Low threshold voltage) cells, which consume more leakage
power, with slower High-$V_{th}$ (high threshold voltage) cells, which consume less leakage power [7]. Since the High-$V_{th}$ cells are slower, this swapping can only be done on timing paths that have positive slack and thus can be allowed to slow down. Hence multiple threshold voltage techniques use both Low $V_t$ and High $V_t$ cells [8]. It uses lower threshold gates on critical path while higher threshold gates off the critical path [9].

![Graph showing variation of threshold voltage with respect to delay and leakage current]

Fig. 6. Variation of threshold voltage with respect to the delay and leakage current.

Fig. 6 shows the variation of threshold voltage with respect to the delay and leakage current. As $V_t$ increases, delay increases along with a decrease in leakage current. As $V_t$ decreases, delay decreases along with an increase in leakage current. Thus an optimum value of $V_t$ should be selected according to the presence of the gates in the critical path. As technologies have shrunk, leakage power consumption has grown exponentially, thus requiring more aggressive power reduction techniques to be used.

Several advanced low power techniques have been developed to address these needs. The most commonly adopted techniques today are in below:

1) Dual $V_{DD}$

A Dual $V_{DD}$ Configuration Logic Block and a Dual $V_{DD}$ routing matrix is shown in Fig. 7.

![Dual $V_{DD}$ architecture diagram]

Fig. 7. Dual $V_{DD}$ architecture.

In Dual $V_{DD}$ architecture [10], the supply voltage of the logic and routing blocks are programmed to reduce the power consumption by assigning low-$V_{DD}$ to non-critical paths in the design, while assigning high-$V_{DD}$ to the timing critical paths in the design to meet timing constraints as shown in Fig. 8.

However, whenever two different supply voltages co-exist, static current flows at the interface of the $V_{DDL}$ part and the $V_{DDH}$ part. So level converters can be used to up convert a low $V_{DD}$ to a high $V_{DD}$. 

Available online: https://edupediapublications.org/journals/index.php/IJR/
Fig. 8. High $V_{DD}$ for critical paths and low $V_{DD}$ for non-critical paths.

2) Clustered Voltage Scaling (CVS)

This is a technique to reduce power without changing circuit performance by making use of two supply voltages [11]. Gates of the critical path are run at the lower supply to reduce power, as shown in Fig. 9. To minimize the number of interfacing level converters needed, the circuits which operate at reduced voltages are clustered leading to clustered voltage scaling.

Fig. 9. Gates of the critical paths are run at lower supply.

Here only one voltage transition is allowed along a path and level conversion takes place only at flipflops.

3) Multi-voltage (MV)

MV deals with the operation of different areas of a design at different voltage levels [9]. Only specific areas that require a higher voltage to meet performance targets are connected to the higher voltage supplies. Other portions of the design operate at a lower voltage, allowing for significant power savings. Multi-voltage is generally a technique used to reduce dynamic power, but the lower voltage values also cause leakage power to be reduced.

4) Dynamic Voltage and Frequency Scaling (DVFS)

Modifying the operating voltage and/or frequency at which a device operates, while it is operational, such that the minimum voltage and/or frequency needed for proper operation of a particular mode is used is termed as DVFS, Dynamic Voltage and Frequency Scaling [12].

5) Adaptive Voltage Scaling (AVS)

Adaptive Voltage Scaling (AVS) provides the lowest operation voltage for a given processing frequency by utilizing a closed loop approach [13]. The AVS loop regulates processor performance by automatically adjusting the output voltage of the power supply to compensate for process and temperature variation in the processor [14]. In addition, the AVS loop trims out power supply tolerance. When compared to open loop voltage scaling solutions like Dynamic Voltage Scaling (DVS), AVS uses up to 45% less energy as shown in Fig. 10.
AVS is a system level scheme that has components in both the processor and power supply. The Advanced Power Controller (APC) provides the AVS loop control and resides on the processor. The Slave Power Controller (SPC) resides on the power supply and interprets commands from the APC. The IP provided in the APC and SPC automatically handle the handshaking involved in frequency and voltage scaling, simplifying system integration in the application.

3. Adaptive Techniques

The power and the delay dependence on the threshold voltage at 0.5 $V_{DD}$ is shown in Fig. 11. From Fig. 11, it is inferred that to achieve high performance, $V_{th}$ has to be decreased. But decreasing $V_{th}$ could cause a significant increase in static leakage power component.

There are several approaches to reduce the stand by leakage current like MTCMOS (Multi Threshold CMOS) and VTCMOS (Variable Threshold CMOS) [11]. These schemes cannot suppress the active leakage power. Another approach is a dual threshold voltage approach, which is to partition a circuit into critical and non critical gates and use low $V_{th}$ transistors only in the critical gates. The drawback of this scheme is that the leakage current cannot be sufficiently suppressed since the large leakage current always flows through the low $V_{th}$ transistors.

1) $V_{th}$ Hopping

Dynamic threshold voltage hopping scheme solves these problems [15]. This scheme utilizes dynamic adjustment of frequency and $V_{th}$ through back gate bias control depending on the workload of the processor.
When the workload is decreased, less power would be consumed by increasing $V_{th}$ as depicted in Fig. 12. This approach is similar to the dynamic $V_{DD}$ scaling, DVS. In the DVS scheme, voltage and frequency are controlled dynamically based on the workload variation.

![Fig. 12. Power dependence on workload.](image)

DVS is effective when the dynamic power is dominant. On the other hand, $V_{th}$ hopping is effective in the low $V_{DD}$ designs, where $V_{th}$ is low and the active leakage component is dominant in the total power consumption.

2) Power gating is the complete shut off of supply nets to different areas of a design when they are not needed. Since the power has been completely removed from these shutdown areas, the power for these areas is reduced essentially to zero. This technique is used to reduce leakage power.

![Fig. 13. Schematic of a power gating methodology.](image)

Power gating uses high-$V_{th}$ “sleep transistors” (also referred to as power switches) to disconnect power supplies to higher-speed and higher-power logic when that logic is not being actively used as depicted in Fig. 13. Power can be gated using either header cells (which disconnect the $V_{dd}$) or footer cells (which disconnect the Ground). It is very common to see multi-voltage and power gating used together on the same design, whereby different regions operate at different voltages, and one or more of those regions can also be shutdown.

3) Multi-Corner, Multi-Mode (MCMM)

Multi-corner, multi-mode (also known as Multi-Scenario) considers optimization at multiple operating corners, and in multiple operational modes, concurrently, instead of using an iterative process that may never converge.

- **State Retention**

  It is the capability to retain the critical state of sequential elements within a block when the block is powered down. State retention generally requires saving the registers and possibly memory contents of the block.

- **Well Biasing**

  Separate voltage supplies can be used to connect to the NMOS and PMOS bulk regions in triple well CMOS technologies. Modification of these voltages with respect to the primary power and ground supplies is
called well-biasing. These supplies can be modulated to provide a back-bias voltage which causes an increase in device $V_{th}$, reducing the sub-threshold leakage. These supplies can also be modulated in the reverse direction to provide a forward-bias voltage which causes a decrease in device $V_{th}$ that increases the speed at which the transistors switch, at a cost of increased sub-threshold leakage. Thus, well-biasing can be used to directly adjust between high performance and low power consumption.

4) Reduction of Clock Frequency

Processors have been able to increase clock frequency to run faster as IC circuits have become smaller. A faster clock boosts performance, but unfortunately also increases power levels. So turning off the clock, or slowing down the clock whenever excess CPU time is available can be used to reduce power levels. Many processors have hardware support to vary the clock frequency or even turn off the clock (i.e., Sleep mode). Some reduced static power levels will still be required even with the clock off to save the values in registers and volatile RAM memory, so that the device can wake up without a full reboot. Interrupt hardware is used to wake a device from Sleep mode, so the hardware used for wakeup can’t be turned off.

The primary goal of using hardware accelerators to reduce power is to lower the clock frequency of the FPGA logic while maintaining acceptable performance levels. However, some applications require rapid response time to asynchronous events such as interrupts in addition to a particular level of data throughput.

Unfortunately, by lowering the clock frequency of the entire system, the clock frequency of the processor is also lowered, effectively slowing its response time to such events. Therefore, if an application requires a fast CPU response time to asynchronous events, the option of lowering the clock frequency of the processor will not be there. However, even when a design requires rapid response time, we can still attain significant power savings by adding hardware accelerators.

Two separate clock domains can be used: a slower domain for the hardware accelerators and a faster domain for the processor. By adding hardware accelerators running at a very low clock frequency, we can relieve the processor of heavy processing work that consumes more power, reducing overall system power consumption without having to reduce the processor clock frequency. Fig. 14 shows the variation of throughput with respect to clock frequency. From the Fig. 14, it is clear that as the clock frequency is reduced, energy per operation can be maintained constant at the cost of delivered throughput.

4. Low Power Buses

In Buses, power consumption takes place by the high capacitance lines and the high switching activities as shown:

$$P = \sum_i \frac{1}{2} \alpha_i f_i C_i V^2$$
Bus Coding-Frequent Value Encoding

This is a technique in which, power dissipation can be reduced by reducing the number of transitions [16]. To minimize the transitions in bus with large capacitance an encoder and a decoder are used as shown in Fig. 15.

![Bus Coding-Frequent Value Encoding](image)

Instead of sending the entire data, a coded data is sent for the frequent data, which reduces the switching activity. Otherwise the data is sent unchanged [17].

5. Non-Conventional Methods of Low Power Design

5.1 Adiabatic Switching

Adiabatic circuits are low power circuits which use "reversible logic" to conserve energy. Unlike traditional CMOS circuits, which dissipate energy during switching, adiabatic circuits reduce dissipation by following two key rules [18]:

1) Never turn on a transistor when there is a voltage potential between the source and drain.
2) Never turn off a transistor when current is flowing through it.

To meet today's power requirement, most research has focused on building adiabatic logic, which is a promising design for low power applications.

Adiabatic logic works with the concept of switching activities which reduces the power by giving stored energy back to the supply. Thus, the term adiabatic logic is used in low-power VLSI circuits which implements reversible logic. In this, the main design changes are focused in power clock which plays the vital role in the principle of operation. Each phase of the power clock gives user to achieve the two major design rules for the adiabatic circuit design [19]. During the recovery phase energy will be restored to the power clock, resulting in considerable energy saving.

These include only turning switches on when there is no potential difference across them, only turning switches off when no current is flowing through them, and using a power supply that is capable of recovering or recycling energy in the form of electric charge. To achieve this, in general, the power supplies of adiabatic logic circuits have used constant current charging (or an approximation thereto), in contrast to more traditional non-adiabatic systems that have generally used constant voltage charging from a fixed-voltage power supply. The power supplies of adiabatic logic circuits have also used circuit elements capable of storing energy. This is often done using inductors, which store the energy by converting it to magnetic flux. There are a number of synonyms that have been used by other authors to refer to adiabatic logic type systems, these include: “Charge recovery logic”, “Charge recycling logic”, “Clock-powered logic”, “Energy recovery logic” and “Energy recycling logic”.

Yet some complexities in adiabatic logic design perpetuate. Two such complexities, for instance, are circuit implementation for time-varying power sources needs to be done and computational implementation by low overhead circuit structures needs to be followed.
There are two big challenges of energy recovering circuits; first, slowness in terms of today's standards, second it requires ~50% of more area than conventional CMOS, and simple circuit designs get complicated.

- **Split level Charge Recovery Logic (SCRL)**
  
  Knight and Younis developed a family of adiabatic circuits known as Splitlevel Charge Recovery Logic or SCRL [20]. This circuit is very similar to a conventional NAND, which is shown in Fig. 16; however, one of the main differences is that the top and bottom rails are driven by trapezoidal clocks (Ø₁ and /Ø₁) rather than $V_{dd}$ and ground.

In the beginning the whole circuit is set at $V_{dd}/2$ except for P1 which is set to ground and /P1 which is set to $V_{dd}$ so that the transmission gate is off. In the next step, the transmission gate is turned on by gradually switching the value of P1 and /P1. Following, Ø₁ and /Ø₁ which were at $V_{dd}/2$ are split to $V_{dd}$ and ground respectively. At this point, the gate computes the NAND of A and B like a non-adiabatic gate would. Once the output is used by the next gate, the transmission gate can be turned back off gradually. Then Ø₁ and /Ø₁ are gradually returned to $V_{dd}/2$ and now the input can change and the next cycle can begin. It is important not to change the input until the rails are back to $V_{dd}/2$ so that a transistor is not turned on when there is a potential difference thus violating the first rule [18].

- **Two Level Adiabatic Logic or 2LAL**
  
  Another interesting adiabatic circuit family is the Two Level Adiabatic Logic or 2LAL developed by Frank. Like SCRL, this family can be fully pipelined at the gate level. Fig. 17 shows the basic building block of 2LAL, a pair of transmission gates which transmit signal A and A' respectively both of which are represented by the single “box” on the left. The fact that 2LAL only requires a basic switching device and is not dependent on CMOS makes it ideal for use with new technologies. Fig. 17 shows the basic buffer element of 2LAL which consist of two sets of transmission gates. Ø₁ and Ø₀ are both trapezoidal clocks but Ø₁ is a quarter cycle behind Ø₀ .

Initially all the nodes are at 0. As the input gradually raises to 1 (if it is 1) or stays at 0, Ø₀ transitions to 1.
On the next step, \( \phi_1 \) transitions to 1 which sets the output to 1 if the input was one and otherwise leaves it at 0 which itself reduces the power dissipation because no charge passes through the transistor. On the third step \( \phi_0 \) transitions back to 0 resetting the input to 0. Finally \( \phi_1 \) transitions back to 0 and the output is restored to 0 by the following gate in order to accommodate for full pipelining and thus the circuit is ready to process a new input. Another feature of 2LAL is that inverters can be easily created by simply crossing over the rails when going from one gate to the next.

**Quasi Adiabatic Buffers:** The term “Quasi-Adiabatic Logic” is used to describe logic that operates with a lower power than static CMOS logic, but which still has some theoretical non-adiabatic losses. Because high-Q inductors are not available in CMOS, inductors must be off-chip, so adiabatic switching with inductors is limited to designs which use only a few inductors. Quasi-adiabatic stepwise charging avoids inductors entirely by storing recovered energy in capacitors. Stepwise charging (SWC) can use on-chip capacitors.

- **Efficient charge recovery logic (ECRL)**

  Efficient charge recovery logic (ECRL) is proposed as a candidate for low-energy adiabatic logic circuit. Energy recovery process is explained with an inverter example of ECRL (Efficient Charge Recovery Logic inverter) as depicted in Fig. 18. Power supply \( PC \) is with trapezoidal pulses [9].

![ECRL inverter](image)

In the initial state holds \( a=1 \), and the \( M_{n1} \) is conducting \((Q=0)\). While \( PC \) rises from 0 to \( V_{dd} \) over conductive transistor \( M_{p2} \) the output \( Q \) follows the variation of \( PC \). When \( PC \) reaches the \( V_{dd} \) value, then it holds \( Q =1 \), and \( Q=0 \) and those conditions are valid logic states at inputs of next stage. During the fall of \( PC \) from \( V_{dd} \) to zero, the right capacitor \( CL \) discharges over the conductive \( M_{p2} \) and \( PC \), and therefore recovers accumulated energy to the \( PC \) supply [21].

Power comparison with other logic circuits is performed on an inverter chain and a carry look ahead adder (CLA) by Yong Moon and Deog-Kyoon Jeong [22]. ECRL CLA is designed as a pipelined structure for obtaining the same throughput as a conventional static CMOS CLA. Proposed logic shows four to six times power reduction with a practical loading and operation frequency range. An inductor-based supply clock generation circuit is proposed. Circuits are designed using 1.0-\( \mu \)m CMOS technology with a reduced threshold voltage of 0.2 V.

- **Positive Feedback Adiabatic Logic (PFAL)**

  The structure of Positive Feedback Adiabatic Logic PFAL logic [23] is shown in Fig. 19. Two n-trees realize the logic functions. This logic family also generates both positive and negative outputs. The two major differences with respect to ECRL are that the latch is made by two PMOSFETs and two NMOSFETs, rather than by only two PMOSFETs as in ECRL, and that the functional blocks are in parallel with the transmission PMOSFETs [24]. Thus the equivalent resistance is smaller when the capacitance needs to be charged. The
ratio between the energy needed in a cycle and the dissipated one can be seen in figure below.

During the recovery phase, the loaded capacitance gives back energy to the power supply and the supplied energy decreases. The partial energy recovery circuit structure so called Positive Feedback Adiabatic Logic (PFAL) has good robustness against technological parameter variations [25]. It is a dual rail circuit; the core of all the PFAL circuit is adiabatic amplifier, a latch made up by the two PMOS and two NMOS that avoids a logic level degradation on the output nodes. The two n-tree release the logic functions. The functional blocks are in parallel with P-MOSFETs and form a transmission gate.

![Fig. 19. PFAL logic circuit.](image1)

![Fig. 20. Two phase adiabatic static clocked logic.](image2)

- **Phase Adiabatic Static Clocked Logic (PASCL)**

  Fig. 20 shows the circuit diagram of Two Phase Adiabatic Static Clocked Logic (2PASCL) [25]. The logic families which include diode in charging path suffer from output amplitude degradation [26]. To deal with this problem, a new logic family was proposed, which was named as two phase clocked adiabatic static CMOS logic. This logic family like other families discussed have used MOSFET as diode by shorting gate and drain of MOSFET together, which does not include diode in charging path, so that output amplitude degradation does not occur. The 2 phase clocked adiabatic static CMOS logic uses a two phase clocking split level sinusoidal power supply. One is in phase while the other is inverted. The voltage level of \( V_{clk} \) exceeds that of \( V_{DD} \) by \( V_{DD}/2 \).

  By using these two split-level sinusoidal waveforms, which have peak to peak voltages of 0.9V, the voltage difference between the current carrying electrodes can be minimized and subsequently, power consumption can be suppressed. It uses two diodes-one diode is placed between output node and power clock, \( V_{clk} \) and the other diode is placed adjacent to nmos logic circuit and connected to other power clock \( V_{clk} \). Both the diodes are used to recycle the charge from output node and to improve the discharging speed of internal nodes.

- **Pre-resolve and Sense Adiabatic Logic (PSAL) Buffer**

  The novel pre-resolve and Sense Adiabatic Logic (PSAL) is a less complex quasi-adiabatic logic circuit usable for frequency range from 100 KHz to 500 MHz [27]. It employs a large height pre-resolved nMOS structured tree and a differential sensing logic. The logic realizes superior energy efficiency through reduced silicon area requirement, low circuit latency, glitch-free output and less switching transients. Significant reduction in switched capacitance realizes enhanced speed performance. Furthermore, evaluation of more than one level of gate (or a complex gate) in each phase makes use of less number of buffers possible, in the adiabatic pipeline. With circuit latency being a major impediment of four-phase adiabatic logic styles, PSAL achieves better throughput and reduced critical path length leading to improved frequency performance. The nMOS structured cascode tree and differential sensing logic help overcome the incomplete charge-recovery and the floating output node problems prevalent in adiabatic logic structures.
Full custom and modular flow is adopted in the circuit designs. • Energy Recovery Clocking

Energy recovery technique mainly works on the clock networks and input gate capacitances. Because of the slow falling/rising transition of energy recovery signals, applying energy recovery techniques to internal nodes could cause the short-circuit power. This sinusoidal clock technique depicted in Fig. 21 can reduce the power due to clock distribution by more than 90% compared to square-wave clocking.

![Fig. 21. Energy recovery clock.](image)

Sinusoidal clock waveforms provide synchronization for coarse-grain recovery, synchronization and power delivery for fine-grain recovery, provides Metal-only clock distribution without clock buffers. This also provides substantially reduced clock uncertainty due to elimination of buffers, substantially reduced clock jitter due to decrease in clock power replenishing requirements, reduced gate leakage due to lower average voltages across gates and increased reliability. Also sinusoidal clock reduces the Electro Magnetic Interference.

• Asynchrobatic Logic

Asynchrobatic Logic, introduced in 2004, is a CMOS logic family design style using internal stepwise charging that attempts to combine the low-power benefits of the seemingly contradictory ideas of "clock-powered logic" (adiabatic circuits) and "circuits without clocks" (asynchronous circuits).

**ENERGY DISSIPATION COMPARISON**

Fig. 22 shows the comparison of the power dissipation in picoJoules between energy recovery logic and different stages of static CMOS logic, for various operating voltages & frequency.

![Fig. 22. Energy comparison between different logics.](image)

It is clear that Energy Recovery logic dissipates the least power even at higher operating frequencies.

6. Conclusion

Thus Energy Recovery logic paves way for reusing the power in high speed power hungry circuits. This logic can be used in memories to save power to a greater extent.
References


[16] Goyal C., & sood, I. Low power data bus encoding & decoding schemes.


