Analysis and Optimization of Sleep modes in Subthreshold Circuit Design

Mingoo Seok, Scott Hanson, Dennis Sylvester, David Blaauw
{mgseok, hansons, dennis, blaauw}@umich.edu, University of Michigan, Ann Arbor, MI

ABSTRACT
Subthreshold operation is a promising method for reducing power consumption in ultra-low power applications, such as active RFIDx and sensor networks. It was shown that operating at the so-called $V_{min}$ supply voltage results in optimal energy operation, where $V_{min}$ typically falls below the threshold voltage. However, all previous subthreshold analyses ignore the leakage current in standby mode. Hence, for applications where operation at $V_{min}$ results in completion of the task well ahead of the required deadline, the energy consumption can be significantly under estimated. In this paper, we investigate the effect of the non-zero standby energy on the optimal energy consumption in subthreshold operation. We first analyze energy consumption both with and without a cutoff technique in standby mode. Two parameters are proposed to capture the cutoff structure's effect on the energy consumption. Second, a methodology to minimize the total energy consumption is addressed. The right selection of the cutoff structure is examined by comparing three different structures. Then, a co-optimization method to optimize the size of the cutoff structure concurrently with the supply voltage, is proposed. This approach reduces energy by 99.2% compared to standby energy unaware optimization.

1. INTRODUCTION
Ultra low power operation has become a key concern in VLSI design. Traditionally, voltage scaling has been used as a method to reduce energy-per-operation due to the quadratic dependence of dynamic energy on supply voltage. This has lead to the use of so-called just-in-time computation where the supply voltage is lowered to the point where a task is completed at exactly the time of its deadline with significant energy savings.

![Figure 1 Vmin/Emin Curve with Zero Sleep energy](image)

Figure 1 $V_{min}$/Emin Curve with Zero Sleep energy

As the supply voltage is lowered, the circuit delay increases and hence the ration of leakage energy over dynamic energy increases as devices spend more time leaking. Particularly as the supply voltage falls below the threshold voltage of the devices, the delay increases exponentially leading to a rapid increase in the leakage energy. In [1], it was shown that there exists a supply voltage call $V_{min}$ below which the reduction in dynamic energy with supply voltage cannot overcome the increase in leakage energy with supply voltage. Hence, the energy-per-operation reaches its optimum $E_{min}$ at the supply voltage $V_{min}$ as shown in Figure 1 and lowering the supply voltage below $V_{min}$ only increases the energy consumption.

Just-in-time operation is only energy efficient if the supply voltage falls above $V_{min}$. For typical designs, $V_{min}$ was shown to fall around 300mV resulting in a frequency of operation in the hundreds of kHz range [3][4]. However, numerous ultra-low power applications, such as active RFIDx, sensor networks and implantable medical devices require substantial lower operating frequencies, ranging from 100’s to 1000’s of Hz or even less [5]. Since these applications are also extremely power constrained, operation at $V_{min}$ is considered a viable option for them. Hence, determination of $V_{min}$ and reliable operation at $V_{min}$ has received significant research in recent years [1][2][8][9].

We refer to the completion time of a task when the processor is operating at $V_{min}$ as $T_{min}$.

![Figure 2 Illustration of Task Scheduling at Different Deadlines](image)

Figure 2 Illustration of Task Scheduling at Different Deadlines

For many sensor applications, operating at the energy optimal voltage $V_{min}$, $T_{min}$ will fall well in advance of the required deadline $T_{deadline}$. For the time that remains between $T_{min}$ and $T_{deadline}$, they therefore enter a standby or sleep mode where the clock has been disabled. However, all previous minimum energy analyses were performed with the assumption that the leakage energy in this standby mode is zero, as shown in Figure 2(a) [1][2]. While the energy consumption in sleep mode can be significantly reduced, some leakage current will always remain, as shown in Figure 2(b). If the device spends relatively little time in standby mode, the impact of this standby energy is negligible. However, if the device spends the vast majority of time in standby mode, as is the case for many sensor applications, the standby energy can be many time that of the energy consumed in active mode. The omission of the standby energy consumption in previous minimum energy analyses was first noted in [6]. However, no detailed analysis of its impact on $V_{min}$ and the cutoff structures was presented.

In this paper, we therefore present a new analysis of the minimum energy operation for applications that have performance requirements that fall well below the performance obtained at $V_{min}$. We first examine operation in the absence of any cutoff structures, where only the clock is gated in standby mode. In this case, the energy optimal supply voltage can scale far below the traditional $V_{min}$, as $T_{deadline}$ is increased beyond $T_{min}$. We also show that ignoring the standby energy results in highly suboptimal energy consumption.
We then examine three different cutoff structures, namely MTCMOS, DTCMOS, and stack forcing. We show that all cutoff structures present a trade-off between leakage reduction and operating frequency degradation. In super-threshold operation, the operating frequency loss due to cutoff structures, such as MTCMOS, is typically limited to 10% or less [7]. However, in subthreshold operation the circuit delay is exponentially dependent on the supply voltage and hence has a much higher performance impact. Since the leakage power is substantial in active mode, the performance penalty introduces a significant energy penalty. Hence, we find that the sizing of cutoff structures is non-trivial and represents a trade-off between standby mode leakage energy and active mode leakage energy.

Since leakage current is weakly dependent on supply voltage, finding the optimal operation for very large values of \( T_{\text{deadline}} \) involves a co-optimization of the sizing of the cutoff structure and the operating voltage. In other words, by increasing \( V_{dd} \) above \( V_{min} \), we can achieve the small size of the cutoff structure which reduces the standby mode leakage energy. Also, the higher \( V_{dd} \) allows the operation to move into the regime of less active mode leakage energy. Though the active switching energy increases due to the increased \( V_{dd} \), however the gain from the leakage energy from both standby and active modes affords this approach. We present results showing that optimization of only the sizing of the cutoff structure can improve the energy consumption by 98% for a \( T_{\text{deadline}} \) that is 1000x that of \( T_{\text{min}} \). By performing a co-optimization of both the sizing and the supply voltage, this energy reduction increases to 99.2%, along with 93% area saving for the cutoff structure, demonstrating the importance of accurate accounting of standby energy in subthreshold design.

The remainder of this paper is organized as follows: In Section 2, we extend the existing analytical energy optimization analysis to include standby energy. In Section 3, we examine the common cutoff structures and present the proposed optimization methodology. We also present comparison between when optimizing their sizes with a fixed supply voltage, \( V_{min} \), and when co-optimizing of sizes and supply voltage together. In Section 4 we present our conclusions.

### 2. Impact of non-zero standby energy on \( E_{\text{min}} \)

The leakage, switch and total energy consumption with voltage scaling is shown in Figure 1 and demonstrates the existence of an energy minimum, \( E_{\text{min}} \), at supply voltage \( V_{min} \). The expression of the minimum energy consumption, as derived in EQ1 is also shown. However, this and other previous analyses overlooked the possible standby leakage. In order to examine how non-zero standby energy affects the \( E_{\text{min}}/V_{min} \) curve, we first investigate the case in which a circuit should finish a task at \( T_{\text{deadline}} \) and define the ratio of \( T_{\text{deadline}} \) over \( T_{\text{min}} \) as \( K_{\text{duty}} \). \( K_{\text{duty}} = T_{\text{deadline}} / T_{\text{min}} \). In other words, \( K_{\text{duty}} \) expresses the delay decrease compared to the completion time of the task at \( V_{min} \). Hence, for \( K_{\text{duty}} > 1 \), there is non-zero standby time unless the supply voltage is lowered below \( V_{min} \).

Initially, we assume that there is no cutoff structure present and then examine different cutoff structures in subsequent sections. Without cutoff structures, the total energy can be expressed as:

\[
E_{\text{Total}} = E_{\text{switch}} + E_{\text{delay}} + E_{\text{sleep}}
\]

[EQ2]

As shown in Figure 3, using no cutoff structure significantly increases the total energy consumption for large values of \( K_{\text{duty}} \). Therefore, cutoff circuitry such as MTCMOS is commonly introduced to reduce the energy consumption in standby mode. Note that there are two eminent effects of the cutoff circuitry on the main circuit. The major effect of the circuitry is to reduce the leakage current during the standby mode, which is the main purpose of the circuitry. However the cutoff circuitry can degrade the performance of the main circuit as a side effect. This is caused by the reduced swing between virtual supply rails. The reduced swing induces a smaller gate-source voltage, thus degrading current driving capacity of the gate, which in turn degrades the performance [7].

Both two effects should be considered to understand the impact of the cutoff structure on \( E_{\text{min}}/V_{min} \). We therefore propose two parameters in EQ3 to capture these effects. The first parameter, denoted by \( K_{\text{leak}} \), leakage reduction factor, indicates how much the leakage in the sleep mode reduces compared to the leakage current without any cutoff structure. The second parameter, the delay degradation factor, denoted by \( 1/K_{\text{delay}} \), indicates the ratio of the delay increase caused by the cutoff circuitry.

\[
\frac{1}{K_{\text{delay}}} = \frac{\int_{V_{min}}^{V_{min}/K_{\text{leak}}}}{\int_{V_{min}/K_{\text{leak}}}}
\]

[EQ3]

As can be seen, both parameters can only take on values between 0 and 1. We first examine the behavior of the \( E_{\text{min}} \) with an imaginary cutoff structure. EQ4 shows the total energy when a cutoff structure with characteristics \( K_{\text{leak}} \) and \( 1/K_{\text{delay}} \) is used. The parameter \( T_{\text{min}} \) denotes the main circuit delay at \( V_{min} \) without the cutoff structure and \( P_{\text{leak}} \) denotes the leakage power without the cutoff structure.
cutoff structure.

\[
E_{\text{total}} = E_{\text{switch}} + E_{\text{delay}} + E_{\text{sleep}} \quad \text{[EQ4]}
\]

where \( E_{\text{switch}} \) denotes the switch energy, which, theoretically, is also affected by the cutoff structure since the cutoff structure limits the voltage swing. However, this was found to be a secondary effect, which can be ignored without significant loss in accuracy.

We now examine \( E_{\text{min}} \) for different values of \( K_{\text{leak}} \) and \( 1/K_{\text{delay}} \). In Figure 4(a), \( K_{\text{leak}} \) is swept from 0 to 0.01. \( K_{\text{leak}} = 1 \) implies that there is no leakage reduction. Large values of \( K_{\text{leak}} \) increase the total energy as expected from the equation. Furthermore, for large values of \( K_{\text{leak}} \), \( V_{\text{min}} \) should be decreased to reduce the increase of the \( E_{\text{sleep}} \) as shown in the Figure 4(b).

Figure 4(c) shows the effect of \( 1/K_{\text{delay}} \) on \( E_{\text{leak}} \) and \( V_{\text{min}} \). Note that \( E_{\text{leak}} \) is small at superthreshold voltages because of the small \( t_{\text{delay}} \). Therefore the effect of the \( 1/K_{\text{delay}} \) is generally negligible at the superthreshold operation. On contrary, the effect of \( 1/K_{\text{delay}} \) is pronounced at the low voltage at which \( t_{\text{delay}} \) is large and \( E_{\text{leak}} \) has a substantial portion of total energy. As a result, \( V_{\text{min}} \) increases as \( 1/K_{\text{delay}} \) increases, as shown in the Figure 4(d).

Figure 4 shows that \( E_{\text{min}} \) and \( V_{\text{min}} \) have a strong dependence on the characteristics of the cutoff structure in subthreshold operation. The value of \( 1/K_{\text{delay}} \) primarily affects \( E_{\text{leak}} \), while the value of \( K_{\text{leak}} \) affects \( E_{\text{sleep}} \). Hence, as the values of \( 1/K_{\text{delay}} \) and \( K_{\text{leak}} \) change for different cutoff structures, either \( E_{\text{leak}} \) or \( E_{\text{sleep}} \) can become the more dominant portion of the total energy. In response, the optimal operating voltage will try to minimize this dominant energy component. Hence, as \( 1/K_{\text{delay}} \) increases \( E_{\text{leak}} \), \( V_{\text{min}} \) will increase to reduce the circuit delay and thereby \( E_{\text{leak}} \). On the other hand, as \( K_{\text{leak}} \) increases \( E_{\text{sleep}} \), it becomes advantageous to increase circuit delay and reduce \( V_{\text{min}} \). Therefore, the two factors \( 1/K_{\text{delay}} \) and \( K_{\text{leak}} \) present conflicting influences on \( V_{\text{min}} \) and require detailed analysis for optimal energy operation.

**2.2 MTCMOS cutoff structures**

For an actual MTCMOS cutoff structure, as shown in Figure 5 the constants \( 1/K_{\text{delay}} \) and \( K_{\text{leak}} \) are not independent, but are related to each other through the value of \( V_{\text{dd}} \) and the MTCMOS footer width.

First, we derive \( 1/K_{\text{delay}} \) for an MTCMOS structure in EQ5 where \( V_{\text{swing}} \) is the voltage across virtual rails. The expression for \( K_{\text{leak}} \) is given in EQ6. In this equation, it is assumed that the voltage across the footer is \( V_{\text{dd}} \) in the standby time, due to the high resistance of the footer compared to the general circuitry. Accordingly, it is shown that \( K_{\text{leak}} \) is a linear function of the footer width.

![Image of MTCMOS Circuit](image)

**Figure 5 MTCMOS Circuit**

\[
1/K_{\text{delay}} = \frac{t_{\text{delay,}}}{t_{\text{delay,}} + t_{\text{min,}}}
\]

\[
K_{\text{leak}} = \frac{1}{t_{\text{leak,}}}
\]

\[
E_{\text{leak}} = CV_{\text{leak}} \mu C_{\text{leak}} W + \mu C_{\text{leak}} W V_{\text{leak}}^2 (1-m) e^{-\frac{V_{\text{leak}}}{V_{\text{leak}}}} (1-\exp(-\frac{V_{\text{leak}}}{V_{\text{leak}}})) \quad \text{[EQ5]}
\]

\[
E_{\text{delay}} = \frac{CV_{\text{leak}} W}{\mu C_{\text{leak}} W} V_{\text{leak}}^2 (1-m) e^{-\frac{V_{\text{leak}}}{V_{\text{leak}}}} + (1-\exp(-\frac{V_{\text{leak}}}{V_{\text{leak}}})) \quad \text{[EQ6]}
\]

![Image of MTCMOS Energy Equations](image)

**Figure 6 (a) and (b)** show the dependency of \( 1/K_{\text{delay}} \) and \( K_{\text{leak}} \) with the footer width and the supply voltage, respectively. The value of \( 1/K_{\text{delay}} \) can easily approach 1 by increasing the footer width at the high voltage, while it more slowly increases at the low voltage. On the other hands, the \( K_{\text{leak}} \) is simply linear function of the footer width across a wide voltage range. The plots in Figure 6 (a) and (b), compare the accuracy of the proposed model against spice simulation, demonstrating acceptable accuracy. The final plot in Figure 6 (c) shows the interdependence between \( K_{\text{leak}} \) and \( 1/K_{\text{delay}} \) as the width of the MTCMOS device is swept. The ideal cutoff structure point lies at \( K_{\text{leak}} = 0 \) and \( K_{\text{deley}} = 1 \). This plot provides a means to compare the efficacy of different cutoff structures for subthreshold design, as discussed further in Section3.

Using \( 1/K_{\text{delay}} \) and \( K_{\text{leak}} \), we now derive the energy equation, EQ7, by plugging EQ5 and EQ6 into EQ4. \( E_{\text{switch}} \) and \( E_{\text{leak}} \) can be substituted from EQ1. Note that the energy equation is a function of \( t_{\text{leak}} \), \( K_{\text{leak}} \), \( K_{\text{delay}} \) and technology constants. In Figure 7, we plot the expression in EQ7 to illustrate how total energy changes with the different footer width and the \( V_{\text{dd}} \) in subthreshold operation. First, for superthreshold supply voltage (> 450mV), \( E_{\text{leak}} \) is relatively small compared to \( E_{\text{leak}} \). Therefore the effect of the \( 1/K_{\text{delay}} \) on \( E_{\text{leak}} \) is negligible for the total energy. Hence, \( K_{\text{leak}} \) can be reduced linearly with the footer width while \( E_{\text{leak}} \) is also reduced slightly with smaller footer width due to the reduced \( V_{\text{leak}} \). Therefore reduced footer width tends to minimize the total energy at superthreshold supply voltages.
On the other hands, at low voltage, $E_{\text{leak}}$ must be taken into consideration. In this case, increasing footer width helps to reduce $1/K_{\text{delay}}$ as a large width increases on-conductance and increases $V_{\text{swing}}$. However $E_{\text{sleep}}$ is proportional with the footer width. Therefore the energy initially decreases with increasing width and after a point starts to increase again due to increased $E_{\text{sleep}}$. Hence, a large value of $K_{\text{duty}}$ makes $E_{\text{sleep}}$ more dominant in the energy equation reducing the optimal footer width.

3. Energy Minimization

In the Section 2, the change of the $V_{\text{min}}/E_{\text{min}}$ curve with non-zero sleep energy condition was analyzed. As shown, the total energy in subthreshold operation is a strong function of the size and type of cutoff structure. In this section, we therefore first compare the $K_{\text{leak}}$ and $K_{\text{delay}}$ of different cutoff structure quantitatively. In doing so, it can be determined how effective each cutoff technique can be, compared to an ideal cutoff technique.

Secondly, we propose a co-optimization method in this section, which optimize the footer size of the cutoff structure as well as the supply voltage. Since EQ7 is a function of $K_{\text{duty}}$, this co-optimization also depends on $K_{\text{duty}}$.

3.1 Comparison of cutoff structures

Values of $K_{\text{leak}}=0$ and $K_{\text{delay}}=1$ for a cutoff structure imply that the structure can eliminate the leakage completely during the standby time along with no delay penalty in the active time.

3.1.1 Comparison of cutoff structures

Values of $K_{\text{leak}}=0$ and $K_{\text{delay}}=1$ for a cutoff structure imply that the structure can eliminate the leakage completely during the standby time along with no delay penalty in the active time.

3.1.2 Comparison of cutoff structures

Values of $K_{\text{leak}}=0$ and $K_{\text{delay}}=1$ for a cutoff structure imply that the structure can eliminate the leakage completely during the standby time along with no delay penalty in the active time.

3.1.3 Comparison of cutoff structures

Values of $K_{\text{leak}}=0$ and $K_{\text{delay}}=1$ for a cutoff structure imply that the structure can eliminate the leakage completely during the standby time along with no delay penalty in the active time.

3.1.4 Comparison of cutoff structures

Values of $K_{\text{leak}}=0$ and $K_{\text{delay}}=1$ for a cutoff structure imply that the structure can eliminate the leakage completely during the standby time along with no delay penalty in the active time.

3.1.5 Comparison of cutoff structures

Values of $K_{\text{leak}}=0$ and $K_{\text{delay}}=1$ for a cutoff structure imply that the structure can eliminate the leakage completely during the standby time along with no delay penalty in the active time.

3.1.6 Comparison of cutoff structures

Values of $K_{\text{leak}}=0$ and $K_{\text{delay}}=1$ for a cutoff structure imply that the structure can eliminate the leakage completely during the standby time along with no delay penalty in the active time.
3.2 Co-Optimization of Vdd and footer width

In the conventional optimum voltage analysis in the subthreshold operation\cite{1}, \(V_{\text{min}}\) was selected without considering the sleep energy. We have shown, however, that the sleep energy can significantly affect the \(V_{\text{min}}/E_{\text{min}}\) values. In addition, we also showed that the cutoff structure impact the \(V_{\text{min}}/E_{\text{min}}\) values as well. Given a fixed supply voltage, the footer width can change the total energy consumption. Therefore, it is necessary to optimize not only the supply voltage but also the footer width to fully minimize the total energy consumption under non-zero sleep energy.

For small values of \(K_{\text{duty}}\), we examine how to optimize \(V_{\text{min}}\) and the footer width together. If \(K_{\text{duty}}\) is unity, the largest energy saving can be achieved by supplying the conventional \(V_{\text{min}}\) without any cutoff structure. Adding cutoff structures induce extra delay, which increase \(E_{\text{leak}}\). Because there is no standby time, i.e. \(K_{\text{duty}} = 1\), the sleep leakage reduction is of no use in this case.

For small values of \(K_{\text{duty}} > 1\), the optimum \(V_{\text{dd}}\) will be similar to the conventional \(V_{\text{min}}\) without the cutoff structure and the footer width will be large. A small \(K_{\text{duty}}\) implies that the \(E_{\text{sleep}}\) is small; therefore the increase of the \(E_{\text{sleep}}\) due to increased width is negligible to the total energy. Rather, the increased footer allows nearly same active mode leakage energy with small performance penalty at around \(V_{\text{min}}\) which minimizes the total energy.

In the Figure 10 (a), each curve represents the change of the total energy with the footer width for small \(K_{\text{duty}}\). Note that standby energy is negligible due to the small \(K_{\text{duty}}\). For high \(V_{\text{dd}}\) such as 0.5V and 0.4V, the energy curve is increased with the width due to dominance of \(E_{\text{active}}\). However, for small \(V_{\text{dd}}\) such as 0.3V and 0.2V, small width exponentially increases the delay, eventually resulting in large energy. However, note that as the width increases, the energy at the low \(V_{\text{dd}}\) decreases and becomes less than the energy consumption at high voltage. For example, in Figure 10(a), the energy curve of 0.2V is larger at small width, but becomes smaller at large width than the energy curve of 0.3V. Therefore for the small \(K_{\text{duty}}\), it is better to use large width and small \(V_{\text{dd}}\) together to minimize the total energy consumption.

Figure 10(b) shows the same energy curve with large values of \(K_{\text{duty}}\). Because of the large value of \(K_{\text{duty}}\), \(E_{\text{sleep}}\) is no longer negligible. Therefore, increasing width may increase the total energy consumption. Because of the increased \(E_{\text{sleep}}\), the energy consumption at 0.2V with large footer cannot become smaller than the energy curve at 0.3V as in Figure 10(a). Instead, the 0.3V curve has a minimum point in at reduced the footer width. Therefore \(V_{\text{min}}\) is increased while the footer width is reduced for the energy minimization for the large \(K_{\text{duty}}\).
3.3 Comparison of the optimization methods

In this final section, three strategies to minimize the energy consumption with non-zero standby energy are compared. The first approach is to use no cutoff technique. Only supply voltage can be optimized to minimize the energy consumption. In the second approach, referred as fixed-$V_{min}$-optimization, a cutoff structure is used and the footer width is optimized while a conventional fixed $V_{min}$ is used. The final approach is to optimize both the footer width and the supply voltage, referred to co-optimization.

Figure 12 (a) shows the change of $V_{min}$ at each strategy. In the first strategy, the $V_{min}$ is reduced to the functional limit $V_{dd}$ as long as the task can be completed in $K_{duty}$. The fixed-$V_{min}$-optimization uses the conventional $V_{min}$. In case of the co-optimization, the $V_{min}$ increases with $K_{duty}$, as explained in the Section 3.2. Figure 12(b) indicates the behavior of the optimum footer width. In the fixed-$V_{min}$-optimization, the footer width is reduced less compared to the co-optimization, due to the low supply voltage as expected.

Finally, the total energy consumption of each strategy is shown in the Figure 12 (c). Even at relatively small $K_{duty}$, the first strategy induces significantly large energy consumption. In addition, the difference of the energy between fixed-$V_{min}$ - optimization starts to increase at larger $K_{duty}$. Therefore the co-optimization is required to minimize the energy consumption, in particular large $K_{duty}$ that exceeds 1000. Note however, that for many sensor applications, values of $K_{duty}$ can easily exceed 10,000, resulting in a 99% energy loss without use of the proposed co-optimization.

Figure 12 Comparison of three optimization Strategies.

4. Conclusion

In this paper, the interaction of the optimal energy, supply voltage and cutoff structures are investigated for subthreshold design. We show that ignoring standby leakage current in subthreshold can significantly impact the energy efficiency of the design. Furthermore, we show that counter to intuition, applications that spend a large portion of time in standby mode require an increased supply voltage compared to the traditional $V_{min}$ combined with a much smaller cutoff transistor width. We also proposed two metrics to compare the effectiveness of cutoff structures in subthreshold operation. Finally, we show that by using the proposed co-optimization of voltage and cutoff width, more that 99.2% energy reduction can be obtained for applications with large values of $K_{duty}$.

References