ABSTRACT

Sleep transistor insertion is one of today’s most promising and widely adopted solutions for controlling stand-by leakage power in nanometer circuits. Although single-cycle power mode transition reduces wake-up latency, it originates large discharge current spikes, thereby causing IR-drop and inductive ground bounce for the surrounding circuit blocks. We propose a new reactivation solution which helps in controlling power supply fluctuations and in achieving minimum reactivation times. Our structure limits the turn-on current below a given threshold through sequential activation of the sleep transistors, which are connected in parallel and are sized using a novel optimal sizing algorithm. The proposed methodology is validated using HSPICE simulations of several benchmark circuits, which have been synthesized onto a commercial 65nm CMOS technology library.

1. Introduction

Nanometer CMOS technologies are characterized by significant sub-threshold conductivity in MOS devices. This causes an upsurge of leakage power dissipation, which has become one of the major design challenges in modern VLSI circuits. Many different methodologies and circuit structures have been proposed to address this challenge.

Power-gating has emerged as a very effective leakage reduction technique. Its basic implementation is shown in Figure 1. A switch device (i.e., a sleep transistor) connects the logic circuit to the ground line, thus creating an intermediate virtual rail (i.e., virtual gnd). The sleep transistor is driven by a control signal (i.e., the sleep signal), which defines the operation mode of the system. During the idle state, the sleep signal is set to the HIGH logic value and the sleep transistor is turned-off. Since the ground path is cut by the sleep transistor, the logic gates are connected to GND only via a high-resistance path; therefore, the sub-threshold leakage is strongly reduced. In this period, due to leakage currents, the virtual-ground rail and all the internal nodes are slowly charged until VDD. When the sleep signal toggles to the LOW logic value, the sleep transistor is switched-on and the circuit steps into the active mode, where normal functionality is guaranteed. Needless to say, this ideal view is complicated by the resistance of the switched-on sleep transistor, which slows down the cells it is connected to. Several embodiments of power-gating do exist. Main differentiation factors include: granularity at which the sleep devices are inserted; methods for sleep transistor sizing under timing/area constraints; physical device implementation (see, for instance, [1]-[4]). One important concern in practical power-gating is the control of the transient currents associated to the transition between the stand-by and the active state (i.e., reactivation or turn-on or wake-up). When a low-to-high transition occurs at the gate of a sleep transistor, all the pre-loaded nodes discharge to GND and the circuit moves from the idle to the active mode. In this phase, both the reactivation time and the power plane integrity are key concerns. The smaller the reactivation time, the higher the performance. Unfortunately, during reactivation the virtual-ground voltage is much larger than in regular operation, and then the amount of discharge current injected into GND (i.e., the reactivation current) can be extremely high. An increase in the discharge current can be a major source of noise on the power distribution networks for the neighboring circuits [5]. Both IR-drop and inductive ground bounce effects can produce signal integrity problems and bit flips in memory elements. Thus, it is evident that maximum reactivation current and wake-up latency are strongly correlated. The critical issue is to find an optimum trade-off point between these two contrasting requirements.

In this paper, we address this issue. In particular, we propose a new power-gating architecture that reduces the turn-on current through sequential activation of an optimally-sized set of sleep transistors. Our structure reduces power consumption and global routing congestion by minimizing the number of control signals used during reactivation, while it minimizes area occupation thanks to a modular physical implementation. We present a novel sizing algorithm which takes, as input, the maximum allowed current, and computes the optimal size of each parallel sleep transistor. This new methodology can be applied at any level of granularity at which the sleep transistors are inserted. The proposed methodology is validated using accurate HSPICE simulations of netlists which include post-layout back-annotated parasitics; the achieved results demonstrate the superiority of our approach w.r.t. existing power-gating solutions.

The remainder of the paper is organized as follows. Section 2 provides an overview of the related work. Section 3 describes the reactivation time model, while Section 4 gives details on the design of the reactivation cell and a sketch of the sizing algorithm. The paper closes with the experimental section and some conclusions.
2. Previous Work

A conventional power-gating structure, shown in Figure 2, is characterized by a set of parallel sleep transistors driven by the same control signal. During reactivation, a large instantaneous current is injected into the ground network, generating high voltage fluctuations.

Figure 2: Single-ST power-gating structure.

In [5], the authors address this issue by proposing two structures in which the ratio of the turned-on transistor increases gradually. This results in a resistance modulation of the ground path and limits the peak current.

The first solution, shown in Figure 3, is a parallel approach (i.e., parallel-ST), in which the sleep transistor is split into a set of transistors with gradually increasing width and connected in parallel. By delaying the control signal, the sleep transistors are activated one at a time, from the smallest to the largest. The delays are generated using a shift register. One of the main drawbacks of this architecture is clearly the need of a dedicated clock signal, which increases power consumption and routing congestion. In addition, no information is given in the paper concerning the approach followed for sizing the transistors. Details on sizing are provided in [6], but unfortunately only a case study is discussed in this paper, and no systematic approach is considered.

Figure 3: Parallel-ST power-gating structure.

The second structure presented in [5], shown in Figure 4, is a gate voltage controlled architecture (i.e., staircase-ST). It modulates the discharge current controlled by the gate-to-source voltage. The sleep transistor operates in weak-inversion until the virtual ground voltage approaches zero, thereby limiting the current. After that, the gate-to-source voltage rapidly reaches VDD, thus forcing the complete turn-on of the sleep transistor. At turn-on, an under-sized pMOS slowly increases the gate voltage of the sleep transistor, while maintaining the device in weak-inversion. After a predefined delay, introduced by means of a shift register, the complete reactivation is initiated. One problem with this solution stands in the fact that, as the threshold-voltage is subject to strong process and thermal variations, an optimum step voltage is hard to find. In order to overcome this limitation, Quasi-Continuous overdrive increment was proposed in [7].

A switched capacitor circuit was used to increase the gate-to-source voltage of the switch device in arbitrarily small steps. Based on a charge sharing principle between pump and gate capacitance, this technique is self-regulating and insensitive to parameter variations. However, it requires additional clocking and circuits. Moreover, switched solutions are well known as noise sources, and could influence the neighboring circuits which are in active state.

Figure 4: Staircase-ST power-gating structure.

At last, in [8], the authors propose a parallel-ST structure with added bulk controller (i.e., PGB-ST). They use a coupling capacitor with sleep signal to generate a voltage boost at the bulk terminal of the sleep transistors. The physical implementation of this method is problematic in custom design, because of the difficulties in supporting multiple bulk polarizations.

An alternative sleep transistor partitioning solution is proposed in [9]. This method is based on the simple observation that a gate driven by another gate which is turning-on absorbs more short-circuit current. To avoid this behavior the proposal is to separate the cells in a number of logic subnetworks and to assign to each one a dedicated sleep transistor. Finally, by scheduling the activation signal of each switch device, it is possible to limit the amount of charge injected to GND. The main shortcoming of this technique is related to its physical implementation. In fact, apart from the large number of sleep signals which are required, it could be that two distant cells have to share the same sleep transistor, thereby causing heavy routing congestion.

All the contributions summarized above do not provide a methodology to control the turn-on current within a selectable bound. They may result in an excessively conservative peak current limitation, which causes heavy degradation on wake-up latency, and then on circuit performance. In contrast, this paper provides a method to automatically size the reactivation circuit for a user-settable maximum acceptable current, thus resulting in a well-controllable reactivation time.

3. Reactivation Time Modeling

A fine-grained MTCMOS design is shown in Figure 5. The power-gated block contains the cells that are connected to the virtual ground of a sleep transistor. For all the connections coming from the power-gated cells and directed to the external blocks, voltage control circuits (e.g., state retention, pull-ups) are inserted. This prevents invalid output voltages during the idle state to be fed to the inputs of non-power-gated cells, avoiding undesirable side-effects such as short-circuit currents and invalid latching. Alternatively, a pull-up pMOS can be inserted on the virtual ground [4], as shown in Figure 7. After the sleep transistor is turned-off, the pull-up rapidly charges the virtual ground rail to VDD, thus forcing a HIGH logic value to the output of all the connected cells. Besides limiting the presence of undefined logic values, this solution creates a stronger stacking effect in the power-gated cells, and thus a higher leakage reduction.
Figure 5: Typical MTCMOS circuit.

We conservatively assume that the virtual ground line has reached a VDD asymptotic value during the inactive period. This condition is the worst-case for the reactivation current, as the large capacitance associated to the virtual ground line is fully charged and can therefore inject the largest possible current peak. Circuit reactivation implies a discharging of the virtual ground capacitance (i.e., \( C_{vgnd} \)) from the initial voltage VDD to the maximum voltage acceptable during active mode (i.e., \( V_{on,MAX} \)).

The continuous line in Figure 6 shows the virtual ground current in the case of a single sleep transistor. Since the peak discharge current is larger than the maximum acceptable current on the power grid, this solution is unacceptable. The fastest way to discharge a capacitance, while meeting a maximum current constraint is to use an ideal current supply. Since the amount of charge is the same in both cases, we can define the ideal minimum reactivation time (i.e., \( T_{react,ideal} \)) as:

\[
T_{react,ideal} = \frac{Q_{tot}}{I_{max}} = \frac{\int I_{vgnd}(t) dt}{I_{max}} \tag{1}
\]

where \( I_{vgnd}(t) \) is the wake-up current of the cluster using a single sleep transistor and \( I_{max} \) is the maximum current.

Figure 6: Single-ST and ideal turn-on currents.

We observe that the value of \( T_{react,ideal} \) is a lower bound and no practical solution can achieve it. However, although an ideal current supply cannot be implemented, we can modulate the width of the sleep transistor in order to maintain its drain current as constant as possible, and as close as possible to the maximum allowed current during virtual ground discharge. This is the basic idea behind our approach.

4. Methodology for Reactivation Time Minimization

In this section, we describe a new design methodology to minimize the reactivation time while constraining the maximum discharge current. A minimum wake-up latency guarantees maximum performance, while the control of the discharge current addresses noise issues on power distribution.

As outlined in the previous section, we propose to modulate the sleep transistor size in order to maintain the discharge current as constant as possible. While the virtual ground rail discharges, the drain-to-source voltage across the sleep transistor decreases, causing a progressive current reduction and thus latency increase. By exploiting a parallel-ST structure, we modulate the effective width of the sleep transistor as a function of the virtual rail voltage.

The two key steps of the methodology, described in details in the next sections, consist of the determination of the optimal delay between the reactivation of two adjacent sleep transistors and the calculation of the size of each sleep transistor.

4.1 Cell Structure

The structure of our reactivation cell is shown in Figure 7.

Figure 7: Reactivation cell architecture.

The sleep signal is connected to the input of the buffer chain, which is properly designed to provide the desired delays. The low-to-high transition on the input activation signal propagates along the chain and the sleep transistors are turned-on with phase shifts given by the buffer delays. The done terminal signals the end of the wake-up transient. The pull-up pMOS eliminates the virtual ground rail fluctuation during stand-by (please refer to Section 3 and [4] for more details). The logic OR between sleep and done (which is the sleep signal translated in time) generates the control signal for the pMOS. This structure avoids the overlapping between nMOS and pMOS active regions, thus eliminating the short circuit current. Number and sizes of the sleep transistors change from design to design. Given the size of the power-gated circuit, the maximum turn-on current and the delay length it is possible to determine an optimal sizing. Differently from existing parallel-ST structures, our solution uses standard library buffers as delay elements. This choice eliminates the need of an external clock signal, thereby reducing power consumption and routing congestion. Moreover, it simplifies the integration in row-based standard cell designs, as it makes it possible to build self-contained, modular sleep transistor cells that include pull-up and reactivation circuits. To substantiate this point, Figure 8 shows the layout of a reactivation cell. White rectangles highlight the buffers, while the green ones correspond to the sleep transistors. The output of each buffer is connected to the gate terminal of one sleep transistor and to the input of the next buffer. The red line in the middle is the virtual-ground rail, which is shared by all the sleep transistors. The output of the last buffer generates the done signal, which is also connected to the input of the OR cell (yellow rectangle on the right). The OR cell drives the pull-up pMOS (cyan rectangle), which is placed over the last sleep transistor.
To guarantee the full compatibility between the standard cell design approach and the full custom design, the sleep transistors are made up of smaller parallel sleep transistors, which do not violate the geometry constraints of row-based design.

![Figure 8: Reactivation cell layout.](image)

The key issue for parallel-ST circuit design is to minimize the reactivation delay. Since we want to guarantee a constant current, the finer the sleep transistor width modulation, the better the current control. But a finer modulation means shorter delay periods. Obviously, while reducing the delay value we increase the length of the buffer chain and the area and power overhead.

Figure 9 shows the normalized reactivation time and area as functions of the delay for a generic power-gated cluster of cells. Sleep transistors are, by far, the elements with the highest area occupation. Hence, when the delay decreases, the cell area is not significantly influenced. In fact, using the minimum delay allowed by the technology (i.e., the delay of the minimum sized inverter with a fan-out of 1), the area overhead is just 3% with respect to the minimum one. In contrast, the reactivation time is highly sensitive to the buffer delay. For instance, when the delay is six times larger than the minimum one, the turn-on latency increases by around 20%. This analysis confirms that, by using the minimum delay, we obtain the minimum reactivation time without sensible area penalty.

![Figure 9: Reactivation time and area vs. delay.](image)

Concerning the power consumption, it has been observed that, if compared to that of the block controlled by the sleep transistors, the contribution of the buffer chain tends to be negligible.

4.2 Optimal Sizing Algorithm

As discussed in the previous sections, the size of each sleep transistor is strongly correlated to the maximum allowed discharge current, the virtual ground capacitance and the delay cells. We present an algorithm that performs optimal sleep transistor sizing under current constraint. The algorithm evaluates the virtual ground voltage every delay period and calculates the optimal sleep transistor width to maintain a constant discharge current.

Algorithm 1 provides the pseudo-code of our approach. The inputs of the algorithm are the power-gated cluster netlist (i.e., $PGnet$), the maximum allowed discharge current (i.e., $Ion_{max}$) and the total sleep transistor width (i.e., $Wtot$). The latter can be obtained by any method, as this does not impact the algorithm. Upon completion, the algorithm returns a list containing the width of each parallel sleep transistor (i.e., $Wlst$).

![Algorithm 1: Optimal Sleep Transistor Sizing.](image)

The core of the algorithm is function $STsize$ which, for a given virtual ground voltage, calculates the optimal width of the sleep transistor to ensure a desired current. The functionality of this method is based on transistor characterization, which is performed off-line and whose results are stored in tabular form. We calculated the optimal transistor size to obtain a unit of current ($Ids=1mA$) with a drain potential which varies from $Vdd$ to a few mVolt. The result of this analysis is shown in Figure 10.

![Figure 10: Sleep transistor size as a function of drain voltage for a constant current of 1mA.](image)
In the saturation region (VDSAT < Vx < VDD), the sleep transistor size is almost constant; this is because of its natural likeness to a current supply. Conversely, in the linear region (VTH < Vx < VDSAT), the current is strongly dependent on Vx and then transistor size. The drain current of a transistor is linearly proportional to its width, this table scales with the desired current. For instance, if we need the size for a current which is 10mA, then the optimal sleep transistor width will be 10 times bigger than the value stored in the table.

The sizing algorithm starts with the virtual ground capacitance measurement (line 4). Procedure CvngdMeasure receives the back-annotated netlist of the power-gated cluster and calculates the Cgnd as the sum of the GND capacitance of each standard cell and all metal line capacitances associated to the Vgnd distribution\(^1\). Lines 5 and 6 initialize two important variables: Delay and VgndL. Delay contains the value of the delay cell, in our case the minimum allowed by the reference library. VgndL represents the running virtual ground voltage, which is VDD at the beginning (turn-on).

In Line 7, the function STsize returns the size of the first sleep transistor. This value is stored as the first element of the output list (i.e., Wlst) and it is assigned to variable W (Line 8), which is the sleep transistor width reached until now. Line 9 sets the number of parallel sleep transistors to 1. The algorithm then enters the while loop of Line 10, in which the sizes of the various sleep transistors to be connected in parallel are calculated. In particular, at each iteration, the size of the next sleep transistor is determined. This is done by first evaluating the virtual ground voltage after the delay period. This task is performed by procedure VgndMeasure in Line 11. The procedure takes the starting virtual ground voltage VgndL, the sleep transistor width W and the delay value Delay, and it calculates the amount of dropped voltage after a delay period. In order to exploit an accurate transistor model, we decided to use a SPICE engine, but it is possible obtain very similar results by implementing mathematical models. Knowing the new virtual ground voltage degradation, we evaluate the size of the next parallel sleep transistor and store it as new element of the list Wlst (Line 12). Finally, we increment the current sleep transistor size and the number of parallel sleep transistors (Lines 13 and 14). The while loop repeats until the total sleep transistor Wlst size is reached (exit condition of Line 10).

It should be noticed that, differently from other approaches, our methodology takes the current constraint as a user-settable parameter. It is then possible to select several current values to meet the design constraints, thus obtaining different cell configurations. This provides our methodology with a high flexibility and a wide applicability.

5. Experimental Results

We have applied the proposed methodology to a set of benchmark circuits belonging to the ISCAS85 suite (only the largest examples were chosen). Each benchmark was synthesized using Synopsys PhysicalCompiler onto a 65nm technology library from STMicroelectronics. All the experimental data were obtained using HSPICE and the technology variables provided by the silicon vendor.

\(^1\)We performed off-line HSPICE simulation to determine the GND capacitance of each standard gate

The methodology presented in [10] was adopted to determine the maximum active current IonMAX of each benchmark. The obtained value was then used as upper bound for the maximum acceptable turn-on current and for the total sleep transistor size. The width of the sleep transistor is the one which guarantees a 5% maximum virtual ground voltage degradation when IonMAX is flowing. By constraining the wake-up current to the value of the active current IonMAX, we generated the reactivation cell netlist using our methodology and we determined the turn-on current and the reactivation time for each of the benchmarks. The reactivation time is defined as the time that elapses between the arrival time of the sleep signal and the time at which the virtual ground voltage reaches 5% of VDD. All the experimental data are collected in Tables 1 (turn-on current) and 2 (reactivation time).

In particular, Table 1 reports, in columns named Cgnd and IonMAX, the virtual ground capacitance and the current constraint used for performing the experiments. Next, the table shows (column Ionideal) the turn-on current (expressed as a percentage of the IonMAX) for the ideal case of a constant current supply discharging. Obviously, this is fixed to 100% for all benchmarks (meaning that the maximum quantitative of allowed discharge current is employed). Column Icali provides the turn-on current obtained using the design methodology presented in this paper. Finally, the last three columns report the information about the turn-on current for some of the techniques that we have surveyed in Section 2. In particular, column Icali refers to the single transistor architecture, column Ionkmi to the realization with four fixed size parallel sleep transistors and column ICali to the case of a staircase solution. On the other hand, Table 2 reports the information regarding the reactivation times.

The first observation that can be made about the results in the two tables is that only the single sleep transistor technique can originate a reactivation time shorter than that achievable in the ideal case (i.e., when 100% of the available discharge current is used). This comes at the price of a significant increase of the amount of current injected into GND (4.7 times larger than the maximum allowed, on average). The single transistor technique can thus obtain maximum performance, but the neighboring circuits could suffer from non-negligible ground-bounce effect and signal integrity problems. In contrast, all the other techniques do not violate the maximum current constraint, and can then be used without reliability loss.

Among the solutions that do not introduce a maximum current violation, we observe that the cali approach (that is, the methodology of this paper) outperforms the other two (i.e., km and sst) by a fair amount. In fact, while Icali is, on average, around 80% of the maximum acceptable current, Ionkm and Ionst are at a low 30% and 26% of IonMAX, respectively. This implies significantly larger reactivation times w.r.t. to the cali approach (i.e., Tcali is 5.7 times larger than Tkm and Tst is 27.9 larger than Tcali). In absolute terms, Tcali is only 2.5 times higher than the ideal case.

As an additional piece of information, it was found that all the reactivation times achieved by using the technique presented in this paper are shorter than the critical paths of the corresponding circuits. Therefore, the entire circuits could be reactivated in a single clock cycle. This clearly constitutes a significant added value of the technique, as it increases its practical applicability.
As shown in Table 1, due to the combinational nature of the benchmarks, virtual ground capacitance and maximum active current are linearly proportional. For this, even if circuits increase in size, the reactivation times are almost the same. In other kinds of circuits the maximum active current may increase at a lower pace than the size of the cluster.

To validate our methodology under this scenario, we evaluated the reactivation time for an increasing value of virtual ground capacitance, maintaining constant the peak current. Figure 11 plots the results. All the techniques guarantee reduced reactivation times decrease. The st technique (top curve) introduces a large wake-up latency, 2.5 times larger than the minimum one with a circuit which is only 5 times larger. The ideal case (bottom curve) remains the best one, while cali (proportionality factor of 0.3) outperforms kim (proportionality factor of 0.47).

**Table 2: Results: Reactivation Time.**

<table>
<thead>
<tr>
<th>Circuit</th>
<th>Tideal</th>
<th>Tcali</th>
<th>Tst</th>
<th>Tkim</th>
<th>Tideal</th>
<th>Tcali</th>
<th>Tst</th>
<th>Tkim</th>
</tr>
</thead>
<tbody>
<tr>
<td>c1908</td>
<td>183.10</td>
<td>47.20</td>
<td>64.99</td>
<td>595.00</td>
<td>183.10</td>
<td>47.20</td>
<td>64.99</td>
<td>595.00</td>
</tr>
<tr>
<td>c1335</td>
<td>207.80</td>
<td>48.30</td>
<td>69.93</td>
<td>601.00</td>
<td>207.80</td>
<td>48.30</td>
<td>69.93</td>
<td>601.00</td>
</tr>
<tr>
<td>c2670</td>
<td>182.50</td>
<td>43.00</td>
<td>60.54</td>
<td>560.00</td>
<td>182.50</td>
<td>43.00</td>
<td>60.54</td>
<td>560.00</td>
</tr>
<tr>
<td>c3540</td>
<td>150.80</td>
<td>418.20</td>
<td>57.38</td>
<td>474.00</td>
<td>150.80</td>
<td>418.20</td>
<td>57.38</td>
<td>474.00</td>
</tr>
<tr>
<td>c2315</td>
<td>196.70</td>
<td>493.70</td>
<td>68.92</td>
<td>544.00</td>
<td>196.70</td>
<td>493.70</td>
<td>68.92</td>
<td>544.00</td>
</tr>
<tr>
<td>c7352</td>
<td>163.40</td>
<td>478.50</td>
<td>58.20</td>
<td>453.00</td>
<td>163.40</td>
<td>478.50</td>
<td>58.20</td>
<td>453.00</td>
</tr>
<tr>
<td>c6288</td>
<td>181.40</td>
<td>560.60</td>
<td>60.29</td>
<td>505.00</td>
<td>181.40</td>
<td>560.60</td>
<td>60.29</td>
<td>505.00</td>
</tr>
<tr>
<td>Average</td>
<td>180.84</td>
<td>459.06</td>
<td>62.44</td>
<td>504.00</td>
<td>180.84</td>
<td>459.06</td>
<td>62.44</td>
<td>504.00</td>
</tr>
</tbody>
</table>

**Table 1: Results: Turn-on Current.**

**Table 3: Results: Power and Energy Consumption.**

As shown in Table 1, due to the combinational nature of the benchmarks, virtual ground capacitance and maximum active current are linearly proportional. For this, even if circuits increase in size, the reactivation times are almost the same. In other kinds of circuits the maximum active current may increase at a lower pace than the size of the cluster.

To validate our methodology under this scenario, we evaluated the reactivation time for an increasing value of virtual ground capacitance, maintaining constant the peak current. Figure 11 plots the results. All the techniques guarantee reduced reactivation times decrease. The st technique (top curve) introduces a large wake-up latency, 2.5 times larger than the minimum one with a circuit which is only 5 times larger. The ideal case (bottom curve) remains the best one, while cali (proportionality factor of 0.3) outperforms kim (proportionality factor of 0.47).

**Table 2: Results: Reactivation Time.**

<table>
<thead>
<tr>
<th>Circuit</th>
<th>Tideal</th>
<th>Tcali</th>
<th>Tst</th>
<th>Tkim</th>
<th>Tideal</th>
<th>Tcali</th>
<th>Tst</th>
<th>Tkim</th>
</tr>
</thead>
<tbody>
<tr>
<td>c1908</td>
<td>183.10</td>
<td>47.20</td>
<td>64.99</td>
<td>595.00</td>
<td>183.10</td>
<td>47.20</td>
<td>64.99</td>
<td>595.00</td>
</tr>
<tr>
<td>c1335</td>
<td>207.80</td>
<td>48.30</td>
<td>69.93</td>
<td>601.00</td>
<td>207.80</td>
<td>48.30</td>
<td>69.93</td>
<td>601.00</td>
</tr>
<tr>
<td>c2670</td>
<td>182.50</td>
<td>43.00</td>
<td>60.54</td>
<td>560.00</td>
<td>182.50</td>
<td>43.00</td>
<td>60.54</td>
<td>560.00</td>
</tr>
<tr>
<td>c3540</td>
<td>150.80</td>
<td>418.20</td>
<td>57.38</td>
<td>474.00</td>
<td>150.80</td>
<td>418.20</td>
<td>57.38</td>
<td>474.00</td>
</tr>
<tr>
<td>c2315</td>
<td>196.70</td>
<td>493.70</td>
<td>68.92</td>
<td>544.00</td>
<td>196.70</td>
<td>493.70</td>
<td>68.92</td>
<td>544.00</td>
</tr>
<tr>
<td>c7352</td>
<td>163.40</td>
<td>478.50</td>
<td>58.20</td>
<td>453.00</td>
<td>163.40</td>
<td>478.50</td>
<td>58.20</td>
<td>453.00</td>
</tr>
<tr>
<td>c6288</td>
<td>181.40</td>
<td>560.60</td>
<td>60.29</td>
<td>505.00</td>
<td>181.40</td>
<td>560.60</td>
<td>60.29</td>
<td>505.00</td>
</tr>
<tr>
<td>Average</td>
<td>180.84</td>
<td>459.06</td>
<td>62.44</td>
<td>504.00</td>
<td>180.84</td>
<td>459.06</td>
<td>62.44</td>
<td>504.00</td>
</tr>
</tbody>
</table>

**Figure 11: Reactivation Time VS. Cluster Size**

One interesting feature of the cali approach is its capability of distributing the total energy in time, thus ensuring a reduced power dissipation. We support this claim by comparing, in Table 3, power and energy consumption obtained using the technique of this paper to those achieved with the st architecture. For all the benchmarks, data include power (column P) and energy (column E) figures for both loaded and unloaded structures. From the results we can conclude that, while energy is almost the same for the two techniques, the cali solution halves power dissipation.

**6. Conclusions**

In this paper we have presented a novel parallel-ST based architecture, which can reduce power supply noise during the wake-up transient of power-gated circuit. We have also proposed an optimal sizing algorithm, which determines the sizes of the parallel sleep transistors as a function of the noise constraint. The proposed methodology was validated on benchmark circuits synthesized onto a commercial 65nm CMOS technology library. The experimental results have shown that our methodology is the one which exploits the maximum amount of available current, resulting in significant reactivation times decrease.

**7. REFERENCES**