Compiler-based vs. Hardware-based Power Gating Techniques for Functional Units

Yen-Hsiang Fang  Yuan-Shin Hwang
Dept. of Computer Science & Engineering
National Taiwan Ocean University
Keelung 20224
Taiwan

Yi-Ping You  Jenq-Kuen Lee
Department of Computer Science
National Tsing Hua University
Hsinchu 30013
Taiwan

Abstract—Reducing leakage power of embedded systems is essential as it constitutes an increasing fraction of the total power consumption in modern embedded processors. Power gating of functional units has been proved to be an effective technique to reduce leakage, and its various implementations can be categorized into compiler-based and hardware-based approaches. Hardware-only designs rely on specific circuits and microarchitectural designs to monitor instruction executions to determine when to power-gate functional units, whereas compiler-based methods attempt to exploit global information of programs and let compilers embed special instructions to turn on and off functional units. This paper compares the efficiencies of hardware and software techniques for power gating of functional units. Experimental results of the DSPstone benchmarks on Wattch show that the hardware-only approach is generally effective in reducing leakage, while the compiler-based approach occasionally performs better as the global knowledge of programs gathered by compilers would avoid incurring excessive power-gating on/off activities. This outcome suggests a better scheme: a hardware-based technique is deployed as the default power gating mechanism, and a compiler would intervene only when its analysis indicates the default method is inferior for certain application programs.

I. INTRODUCTION

Minimizing power dissipation is critical for embedded systems, and it can be achieved by techniques designed at the algorithmic, architectural, logic, and circuit levels [5]. Various hardware and software techniques have been proposed to reduce dynamic for power dissipation with architecture designs and/or software arrangement at instruction level [1], [6], [11], [19], [20], [24], [25], [26]. For example, several types of code rearrangement have been used to reduce the dynamic power, such as utilizing the value locality of registers [6], swapping operands for Booth multipliers [20], scheduling VLIW instructions to reduce the power consumption on the instruction bus [19], gating the clock to reduce workloads [11], [25], [26], utilizing cache subbanking mechanism [24], and buffering instructions nested within loops in a minicache [1]. Dynamic energy consumption is the main concern of these methods since it is the dominant form of power dissipation then.

As the minimum feature size gets smaller and more transistors are packed densely onto processors, static power dissipation due to leakage takes an increasing fraction of total power in processors. Static power dissipation increases about 5 times each generation since the total leakage current increases about 7.5 times [2], [4]. Consequently, it is estimated that leakage power will be the dominant form of power...
dissipation soon [8], [13], [15], [17], [18], [23]. In order to minimize the impact, power gating could be used to reduce leakage power [4], [12], [14]. Specifically, a functional unit should be shut down every time it enters a sufficiently long idle period and be turned back on before it is needed. Therefore, the key issue here is how to identify the onsets and conclusions of long idle periods and then perform power gating to turn off and on functional units without incurring significant performance penalties.

Power gating techniques can be categorized into hardware-based and compiler-based approaches. Hardware-based methods rely on special circuits and microarchitectural designs to monitor instruction executions in order to determine when to turn off and on functional units [12], [14]. The advantage of this approach is that programs can be run without any modifications, but its disadvantage is that it cannot exploit global information of programs. By contrast, compiler-based methods attempt to integrate architecture and compiler power-gating mechanisms [9], [21], [27], [29], [30], [31]. This approach involves compilers inserting specific instructions into programs to shut down and wake up components based on data-flow analysis or profiling information. The benefit of this approach is that power gating will be performed based on the insight knowledge of programs gathered by compilers, while its drawback is that additional instructions must be implemented by the hardware to power-gate on/off functional units and they must be explicitly embedded into programs. Furthermore, extra efforts must be taken to carefully merge power-gating instructions to avoid code size explosion if multiple functional units are guarded by power-gating circuits.

This paper compares the efficiencies of hardware- and compiler-based techniques for power gating of functional units. A straightforward implementation of power gating based on a hardware counter [12] is compared with a system that runs programs with compiler-embedded on/off instructions [29]. In addition, the Sink-N-Hoist compiler framework is used to merge several power-gating instructions into a single compound instruction, and hence the code size explosion issue is minimized [28]. Experimental results of the DSP-stone benchmarks on Wattch [3] show that the hardware-only approach generally outperforms the compiler-based method, but there are still a couple of programs that compiler-assisted approach works better. This outcome reveals that hardware-only techniques will reduce more leakage power when idle and active phases of programs are distinctive and last relatively long time intervals. However, programs with short active and idle periods would generate excessive on/off activities and hence incur significant overheads for the hardware-only approach. In contrast, global knowledge of programs would help compiler-assisted techniques to avoid such a pitfall.

This observation suggests a better solution. A hardware-based technique can be deployed as the default power gating mechanism, since generally it is very efficient in leakage reduction. However, a compiler would intervene when its analysis identifies that the default method might not be effective for certain application programs. It could either reorganize the code or adjust the parameters of the default hardware mechanism. This study indicates that better leakage reduction can be achieved by the cooperation of hardware-based and compiler-based approaches, and hence further compiler research will be needed in order for a compiler to determine when and how to intervene.

The rest of this paper is organized as follows. Section II portrays the architecture of the target platform. Section III briefly describes the compiler technique to embed and merge power-gating instructions and Section IV outlines the simple hardware-only implementation. Experimental results will be presented in Section V, and Section VI summarizes this paper and discusses the future work.
II. MACHINE ARCHITECTURE

The instruction set architecture targeted by compiler-assisted techniques must support power-gating control at the component level. This paper focuses on reducing the power consumption of certain components by invoking power-gating technology. Power gating is analogous to clock gating, except that devices are powered off by switching off their supply voltage rather than the clock. This can be implemented by forcing transistors to be off or using MTCMOS (multi-threshold voltage CMOS technology) to increase the threshold voltage [4], [12], [14], [22].

Figure 1 illustrates an example of the target machine architecture based on a DEC Alpha 21264 processor with an instruction fetch, issue, and retire unit (Ibox), a block of integer functional units (Ebox), a block of floating-point functional units (Fbox), a memory reference unit (Mbox), and an external cache and system interface unit (Cbox) [7].

In the adapted DEC Alpha 21264 architecture model, the Ebox and Fbox were equipped with power-gated functions. The power state of each unit is controlled by the 64-bit integer power-gating control register (PGCR). In this case, one bit is used for the integer multiplier unit and three bits are used for the floating-point functional units. Setting the power-gating bit true powers on the corresponding module, and clearing the bit to zero powers off the corresponding module immediately in the following clock cycle. A new instruction was implemented to control units with the power-gated function by moving the appropriate value from a general-purpose register to the PGCR. The integer ALU unit is always powered on since it takes the responsibility for moving data to the PGCR.

III. COMPILER-ASSISTED POWER-GATING CONTROL PLACEMENT

This section reviews a compiler approach that statically analyzes the activities of power-gating candidates of the input programs and inserts power-gating instructions at appropriate positions with the consideration of code size issues [29], [30], [28]. The proposed framework, called Leakage-Power-Reduction Framework, is operated with three major phases:

1. Component-Activity Data-Flow Analysis (CADFA), which estimates the activities of the power-gating candidates within a given program,
2. Power-Gating Instruction Scheduling, which determines whether, where, and when power-gating controls should be employed so as to reduce energy dissipation, and
3. Sink-N-Hoist Analysis, which attempts to sink (postpone) power-off operations and hoist (advance) power-on operations for increasing the opportunity to merge power-gating instructions into compound instructions and thus reduce program code size.

Figure 2 sketches the process scenarios, corresponding to the above three phases, of a motivating example in the view of the compiler approach with the assumption that three floating-point units (an arithmetic logic unit, a multiplier, and a divider) are considered as the power-gating candidates. Each plot of Figure 2
shows the activities of the power-gating candidates, represented as boxes, in the timeline and also the placement of power-gating instructions. The leftmost plot shows the activity information produced by CADFA, where a shaded box represents a unit which is in use at that time, and is simply the case without power-gating controls; the middle plot shows the case when Power-Gating Instruction Scheduling is applied; and the rightmost one shows the case when Sink-N-Hoist Analysis is involved.

Basically, the Leakage-Power-Reduction Framework is performed with a set of data-flow equations and the control-flow graph of the input program. In CADFA, component-activities, the activities (active or inactive) of components, are propagated with union as the meet operation. A component-activity is generated at a block if a component is required for processing and it is killed if the component is released from the process. Once the activity information of components has been obtained, power-gating instructions can be inserted into programs at the appropriate points (i.e., the beginning and end of an inactive block) to power off and on unused components so as to reduce the leakage power. However, both shutdown and wake-up procedures are associated with an additional penalty, especially the latter due to peak voltage requirements. Power-Gating-Instruction Scheduling is then performed and takes account of the influence of conditional branches in programs — the time required to instigate power-gating controls on components is related to the number and complexity of program branches. The process seems to be done after the phase of Power-Gating-Instruction Scheduling. However, there are concerns about the amount of power-control instructions being added to programs with the increasing amount of power-gating candidates in a system-on-a-chip (SoC) design platform for embedded systems. Therefore, Sink-N-Hoist Analysis was proposed to generate balanced scheduling of power-gating instructions.

The main idea of Sink-N-Hoist analysis is to reduce the problem of too many instructions being added with code-motion techniques. The approach attempts to merge several power-gating instructions into one compound instruction by ‘sinking’ power-off instructions and ‘hoisting’ power-on instructions; that is, postponing the issuing of power-off instructions and bringing forward the issuing of power-on instructions. For instance, a power-off instruction can be postponed for several cycles to be merged with adjacent power-off instructions. This will result mainly in improvements to the code size, but also in performance and energy via grouping effects. Similarly to CADFA, Sink-N-Hoist Analysis is based on a set of data-flow equations to collect the information for the code motion of power-gating instructions, such as the information of possible positions to issue for each power-gating instruction and the information which power-gating instructions should be merged.

IV. Hardware-Only Approach

This section recaps the time-based power gating technique that is commonly deployed by the hardware-only approach [12]. Instead of executing explicit power-gating instructions, hardware-only techniques rely on logic circuits to detect the onsets and conclusions of sufficiently long idle periods and then to power-gate off/on functional units. The easiest way is to turn off a functional unit after it is idle over a certain number of cycles. Similar techniques
have been used for reducing leakage of caches [10], [16].

![State Diagram](image)

**Fig. 3. State Diagram (Adapted from Figure 5 in [12])**

In order to model this tactic, a state diagram shown in Figure 3 can be associated with every functional unit. The *idle_detect* state is the normal, active state, when the functional unit is ready for execution. A functional unit will be power-gated off and enter the *uncompensated* state after being idle for more than $T_{idledetect}$ cycles. If it is waked up during this state, it will miss the break-even point as the overhead of power-gating is greater than the leakage reduction. By contrast, power-gating will save leakage once the idle period lasts longer than $T_{breakeven}$, and the functional unit will enter the *compensated* state. The longer the functional unit stays in the *compensated* state, the more leakage power it will save. If an instruction is ready for execution when its corresponding functional unit is in the *compensated* or *uncompensated* state, the unit will move to the *wakeup* state and stay there for $T_{wakeup}$ cycles before entering the active *idle_detect* state. $T_{wakeup}$ is the latency of powering up functional units.

According to the state diagram, the efficiency of the hardware-based power-gating technique is determined by the three parameters, $T_{idledetect}$, $T_{breakeven}$, and $T_{wakeup}$. The first parameter, $T_{idledetect}$, determines how aggressive the power-gating mechanism would be. It is the only parameter among these three that can be dynamically adjusted, since the other two $T_{breakeven}$ and $T_{wakeup}$ are fixed once the physical circuit design technique is chosen. A small $T_{idledetect}$ would identify more idle periods but it might cause performance degradation if functional units have to be frequently turned back on before the break-even point. On the contrary, a large $T_{idledetect}$ might miss many opportunities to power-gate functional units.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clock</td>
<td>600 MHz</td>
</tr>
<tr>
<td>Process parameters</td>
<td>0.10 µm, 1.9 V</td>
</tr>
<tr>
<td>Instruction issuing</td>
<td>In-order</td>
</tr>
<tr>
<td>Decode width</td>
<td>8 instructions/cycle</td>
</tr>
<tr>
<td>Issue width</td>
<td>8 instructions/cycle</td>
</tr>
<tr>
<td>Commit width</td>
<td>8 instructions/cycle</td>
</tr>
<tr>
<td>RUU size</td>
<td>128</td>
</tr>
<tr>
<td>LSQ size</td>
<td>64</td>
</tr>
<tr>
<td>Functional units</td>
<td>4 integer ALUs</td>
</tr>
<tr>
<td></td>
<td>1 integer multiply/divide unit</td>
</tr>
<tr>
<td></td>
<td>4 FP ALUs</td>
</tr>
<tr>
<td></td>
<td>1 FP multiply/divide unit</td>
</tr>
<tr>
<td>Register files</td>
<td>32 64-bit integer registers</td>
</tr>
<tr>
<td></td>
<td>32 64-bit FP registers</td>
</tr>
<tr>
<td></td>
<td>1 64-bit power-gating control register (PGCR)</td>
</tr>
</tbody>
</table>

**TABLE I**

**BASELINE PROCESSOR CONFIGURATION**

V. EXPERIMENTAL RESULTS

A. Setup

The target platform is a DEC-Alpha-compatible architecture with the power-gating controls and instruction shown in Figure 1, and experiments are conducted on the Wattch simulator with 0.10-µm process parameters and 1.9V $V_{DD}$ [3]. Table I summaries the baseline configuration of the simulator. By default, the simulator performed out-of-order executions. The ‘-issue:inorder’ option is used in the configuration so that instructions would be executed in order to ensure the correctness of power-gating controls. Nevertheless, the software-assisted method can also be applied to out-of-order issue machines if the additional hardware supports proposed in [30] are employed. The benchmarks used in this paper are taken from the floating-point version of the
DSPstone benchmark suite [32]. The average IPC (instructions per cycle) of the benchmarks is 0.36 with the configuration in Table I.

Since Wattch does not model leakage at the component level per se, this paper assumes that leakage power makes up 10% of total power consumption. Furthermore, each wake-up operation is assumed to have a 20-cycle latency (i.e. $T_{\text{wakeup}} = 20$ cycles) and to dissipate ten times of the leakage power. Similarly, every turn-off instruction consumes twice of normal leakage energy. The energy consumption of fetching and decoding a power-gating instruction was assumed to be two times the leakage power. In addition, normalized leakage of the target platform will be computed relative to the leakage measured on Wattch $cc3$ with only clock-gating mechanism.

**B. Performance Evaluation**

This section will evaluate the performance of four hardware-only power-gating configurations, i.e. $T_{\text{idledetect}} = 16, 48, 64, \text{ and } 96$ cycles, and two software-assisted power-gating designs, namely CADFA and Sink-N-Hoist.

**Leakage**

Figure 4 illustrates that the normalized leakage of power-gated functional units for the DSPstone benchmarks under the above various software and hardware power-gating configurations. Basically the hardware-only configurations reduce much more leakage than the software-assisted designs. On average the hardware-only configurations lower the leakage power down to about 10%, while the software-assisted designs cut the leakage down to around 30%. However, CADFA and Sink-N-Hoist still manage to outperform the hardware-only configurations for the two benchmarks $f\text{ir2dim}$ and $m\text{atrix1}$, and only dissipate only about half of leakage. The main reason is that in some occasions hardware-only configurations might be too eager or too lukewarm due to the lack of global information of the programs, whereas the compiler-based techniques tend to know when to power-gate on/off functional units.

Power gating functional units incur overheads since turning off and on circuits do take energy. The red portions of bars in the figure denote such energy overheads, which are reasonably small for all hardware and software configurations. The hardware-only configurations with small $T_{\text{idledetect}}$ cycles introduce smaller overheads than the hardware-only configurations with large $T_{\text{idledetect}}$ numbers. The compiler-based approach usually suffers more overheads than the hardware-only method for it has to execute instructions explicitly to power-gate functional units. However, such overheads can be reduced by merging the power-gating instructions, as shown by the result that Sink-N-Hoist incurs much less penalty than CADFA.

**Run Time Impact**

Turning off/on functional units to reduce leakage power will definitely incur performance penalties since several to tens of cycles are needed to power on/off circuits. Figure 5 shows the performance impact of the various power-gating configurations. Hardware-only implementations commonly suffer higher performance degradation than compiler-based techniques, especially for the configurations with small $T_{\text{idledetect}}$ numbers due to excessive power-gating on/off activities. On average they might slow down the execution of the DSPstone benchmarks by 10% to 20%. In contrast, compiler-based designs incur only negligible performance degradation, roughly 2% on average.

Figure 4 and Figure 5 reveal that hardware-only configurations with reasonably large $T_{\text{idledetect}}$ cycles seem to be a favorable choice as they can significantly reduce leakage without incur considerable performance penalties.

**Wakeup Latency**

The hardware-only approach will be more effective if the wakeup latency $T_{\text{wakeup}}$ can
Fig. 4. Normalized Leakage Power
Fig. 5. Average Run Time Impact
Fig. 6. Normalized Leakage Power ($T_{\text{wakeup}} = 10$ Cycles)

Fig. 7. Average Run Time Impact ($T_{\text{wakeup}} = 10$ Cycles)
be reduced. Figure 6 and Figure 7 show the effects of cutting the wakeup latency by half to 10 cycles. Leakage and runtime impact can be further reduced as functional units sit idle for fewer cycles waiting for their circuits to be powered back on.

VI. SUMMARIES AND FUTURE WORK

This paper compares the efficiencies of hardware- and compiler-based techniques for power gating of functional units. A straightforward implementation of hardware power gating mechanism is compared with a system that runs programs with compiler-embedded on/off instructions. Experimental results of the DSPstone benchmarks on Wattch show that the hardware-only approach generally performs better than the compiler-based method, but compiler-based approach still manages to outperform in some occasions. This outcome reveals that hardware-only techniques might generate excessive on/off activities when the program executions do not fit the pattern, while global knowledge of programs would help compiler-based techniques to avoid such a pitfall. This observation suggests a combination of hardware-based and compiler-based approaches: a hardware-based technique can be deployed as the default power gating mechanism, and a compiler would intervene only when its analysis indicates the default method is inferior for certain application programs.

A compiler now has to take the responsibility of determining if the default hardware mechanism might inappropriately turn on and off functional units. If could adjust the parameters of the default mechanism, or even reorganize the instructions when adjusting parameters alone would not be enough. There is an on-going project here that studies how to perform analysis on the binary executable codes of application programs in order to decide if the compiler should intervene. In addition, this project also investigates how the object codes can be reorganized to fully exploit the default power gating mechanism.

REFERENCES


