On-Line Adjustable Buffering for Runtime Power Reduction

Andrew B. Kahng
CSE and ECE Departments
UC San Diego
abk@ucsd.edu

Sherief Reda
Division of Engineering
Brown University
sherief_reda@brown.edu

Puneet Sharma
ECE Department
UC San Diego
sharma@ucsd.edu

Abstract

We present a novel technique to exploit the power-performance tradeoff. The technique can be used stand-alone or in conjunction with dynamic voltage scaling, the mainstream technique to exploit the tradeoff. Physical design, specifically repeater insertion and sizing, is naturally signed-off at the highest performance mode. We observe that through simple modifications to the repeaters (buffers and inverters), it is possible to dynamically customize the repeater driving capacity of the design. This customization opens the door to a novel opportunity for on-line power-performance tradeoff: customizable repeaters can trade away performance for reductions in power, or vice versa. We describe a simple customization of repeaters to have an additional adjustable low-power operation mode besides their regular operational mode. Using selective logic remapping, we demonstrate how to use the new customized repeaters in a design flow that does not impact the high-performance signoff, yet attains considerable power reductions in low-performance mode. With industrial tools and real-world benchmarks at the 90nm node, we observe an average of 8.34% reduction in total system power in lower performance modes, while ensuring no sacrifice to high-performance modes. We estimate the overhead of our approach to be a tolerable 2.89% in the total device area and 3.41% in the total routing requirements, which is likely easily accommodated in the whitespace of a design.

1. Introduction

In today's designs, power is a major design concern and has become a bottleneck for CMOS scaling into the 90nm and 65nm nodes. High power consumption shortens battery life, increases packaging costs, and reduces circuit reliability. CMOS power is composed of dynamic and static (leakage) components. Dynamic power is due to charging and discharging of capacitors (*switching power*) and intermittent short-circuiting during CMOS switching (short-circuit power). Leakage power is composed of three main components: (1) subthreshold leakage, (2) gate leakage, and (3) reverse biased drain substrate and source-substrate junction band-to-band tunneling leakage [2]. With scaling, dynamic power increases as a result of increasing clock frequencies, relatively poor interconnect scaling, and increasing design complexity. Leakage, on the other hand, increases due to lowered threshold voltage (V_{th}), shrinking channel length and gate oxide thickness, and increasing complexity.

Dynamic voltage scaling (DVS) is a popular technique to adjust the power-performance of a circuit in an on-line fashion. Unfortunately, the feasibility of this technique is now being challenged by issues such as signal integrity and reliability. Dynamic power increases tremendously with supply voltage and consequently circuits are operated at the minimum possible voltage that still allows them to meet timing constraints. On the other hand, supply voltage must be kept high enough to ensure sufficient noise margins to cope with crosstalk- and process- induced failures. Soft errors are also known to increase substantially when supply voltage is reduced. As a consequence supply voltage is no longer allowed to scale at runtime or is scaled over a very small range. In this paper we propose an alternative technique to exploit the power-performance tradeoff in an on-line fashion. We evaluate the technique both as a replacement of DVS and as a complementary technique that reduces power beyond the savings from DVS. Our technique reduces dynamic and leakage power. In DVS, due to source voltage reduction, dynamic and leakage power decrease quadratically. However, to avoid high delay penalty of reduced supply voltage, threshold voltage has to be reduced which causes a near-exponential increase in leakage power. Therefore, DVS is primarily a dynamic power reduction technique, and its benefit in terms of leakage reduction is questionable.

Poor interconnect scaling has created the buffering problem [12]. An ever-increasing number of repeaters must be inserted into a design, and an explosion in the number of repeaters is underway as the electronic industry is moving to 65nm and beyond. Consequently, power consumption of repeaters accounts for an increasing proportion of total power consumption. Repeater insertion and sizing, or physical design optimizations, are mostly signed-off at the highest performance mode. For example, repeater sizes are determined to meet the timing requirements of the highest performance mode. However, the physically-designed high-performance repeaters are not necessarily needed in lower-performance operating modes – e.g., it might be possible to use a 2× repeater instead of a 4× repeater and still meet the timing requirements of low-performance modes.

In this paper, we propose an *on-line, adjustable buffering methodology* for ASICs. Our approach dynamically trades off total system power and performance to take advantage of periods (modes) of low utilization that occur during the operation of many chips. Depending on the mode of operation, our methodology dynamically adjusts the buffering capacity of the system: in the highest performance mode, our customizations leave the performance untouched, but in lower performance modes, we take advantage of the slack in performance to reduce repeater driving capacity and hence power consumption. We propose a number of techniques to attain the timing of the highest-performance mode, and to minimize the area overhead of our customization structures. We empirically validate our approach using a state-of-the-art commercial physical design flow and 90nm technology and libraries, as well as realistic industrial benchmarks. Our approach gives a 8.34% reduction in total system power of low-performance modes, with no impact to the high-performance modes. Our area overhead analysis shows an estimated increase of 2.89% in the examples studied.

Proceedings of the 8th International Symposium on Quality Electronic Design (ISQED'07)
0-7695-2795-7/07 $20.00 © 2007 IEEE

Motivation
Repeaters – whether buffers or inverters – are inserted to meet timing and slew constraints, and account for a significantly increased portion of current designs. There are two reasons for this trend: (1) devices are getting smaller in size with every technology node, limiting their driving capability; and (2) poor interconnect length scaling [12], together with increased interconnect resistance due to shrinking dimensions, is further straining the drivers. Not only do inserted repeaters require significant die area, but they also significantly contribute to total power.

During runtime, a circuit may dynamically adjust its V_{dd} or operating frequency in order to reduce the total power consumption. This typically occurs when there is low performance load on the chip, or when the chip temperature becomes critical. Reducing the frequency leads to an interesting natural phenomenon: the positive slack of critical paths starts increasing, since the high-performance frequency was originally designed, with some tolerance, to be the reciprocal of the critical path delay. The increased positive slack presents us with an opportunity: if somehow the repeaters can be adjusted, say, by converting 24× repeaters to 12× repeaters, then we can reduce the repeater power consumption at the tradeoff of reducing the amount of "free" positive slack. This reduction of positive slack must be achieved without bringing the positive slack down to zero, or violating the performance and slew requirements.

To exploit this opportunity, we must overcome the following key challenges.

1. How can we design *customized repeaters* that can adjust their driving capacity in an on-line fashion, depending on the operating mode of the circuit?
2. How can we minimize the area overhead of such customized repeaters?
3. How can we physically design a chip with customized repeaters, so that there is no impact on performance at the highest-frequency modes, but yet we achieve power reductions at lower performance modes and at all operating supply voltages?

Contributions of This Work
The key contributions of this work are as follows:

- Design of customized repeaters that offer a tradeoff between driving capacity and power. Characterization of such repeaters to evaluate their power and performance at different operating modes.
- Methodology to utilize the customized repeaters in place of traditional repeaters to: (1) ensure no sacrifice in performance at high performance mode, (2) reduce power at lower performance modes, and (3) minimize area overhead.
- Validation of the approach on real-world testcases to evaluate the power reduction and area overhead.

The organization of this paper is as follows. Section 2 gives the motivation for our work. Section 3 presents our on-line customization methodology and analyzes its impact on performance, power, and device area. Experimental results are given in Section 4. Finally, Section 5 reviews the highlights of our approach and outlines future work.

2. An Adjustable Repeater Methodology

In this section, we address the key challenges mentioned in the previous section. Our approach starts with a customized repeater design. An inverter is chosen for such customizations; similar modifications are realizable for buffers.

2.1 Adjustable Repeater Design

In standard cell libraries, devices of very large width cannot be laid out due to cell-row height restrictions. Therefore, in most repeaters multiple devices are connected in parallel to effectively act as larger-width devices and provide the required drive strength. This is known as device fingering. For example, in the Artisan TSMC 90nm library available to us, 11 out of the 14 inverters have two or more devices connected in parallel to provide the required drive strength. Our customization adds the ability to disable some of the parallel-connected devices to reduce the repeater's drive strength and power. To create a customized inverter, we alter an inverter cell to include two *control gates* as shown in Figure 1. The figure illustrates an inverter cell with two fingers each of NMOS and PMOS; *one* PMOS (NMOS) finger is connected to V_{dd} (V_{ss}) through a control gate. If the control signal (LPM, which stands for low-performance mode) is turned *off* then the inverter functions as a regular 2× inverter. If the LPM signal is turned *on* then half of the inverter is shut down, and the inverter works as 1× inverter. In general, for larger inverters and buffers, half of the PMOS fingers are controlled by one PMOS control gate and half of the NMOS fingers by one NMOS control gate. The approach is similar to the widely-used power gating (also known as MTCMOS) technique [10], except that instead of introducing a control gate between all MOS devices in a cell and power/ground, we introduce the control gate only between half of the MOS devices and power/ground. Comparing a customized inverter with a regular inverter, we expect that:

- *At regular operating mode (LPM=0)*, the customized inverter works as a regular inverter, with the exception that it has a slightly higher output resistance than the traditional design. The output resistance can be readily brought down by properly upsizing the control gate size. Upsizing the control gate, however, increases the cell area and also the gate leakage of the control gate. We discuss this tradeoff later.
- *At low-performance mode (LPM=1)*, the customized inverter has only two operating transistors versus four in the traditional 2× inverter, devices of which cannot be disabled. At low-performance mode, one of the two parallel-connected devices is connected to power or ground through an OFF control gate and is effectively disabled due to the stacking effect. This provides a higher resistance drain to source path for subthreshold and short-circuit currents in comparison to that provided by two parallel-connected devices in traditional inverters. Consequently subthreshold leakage and short-circuit power decrease. Dynamic power of a cell (i.e., internal power), which is composed of short-circuit power and charging/discharging power of internal devices, is almost entirely due to short-circuit power [1]. In our test circuits we consistently observe internal power between 65% and 70%. The remainder of the dynamic power is primarily contributed by the interconnects. The customized inverter is slower, but as stated in the motivation, this is allowable in low-performance modes of operations.

We validate the above-mentioned hypotheses by characterizing repeaters for their performance and power using SPICE simulations. Table 1 presents the input capacitance, delay, leakage, and internal power for INVX8 and BUFX8, and their customized counterparts in regular (LPM=0) and low-performance (LPM=1) modes. With respect to traditional cells, we observe the corresponding customized repeaters to have: (1) marginally higher delay in regular mode, (2) nearly unchanged capacitance in both modes, (3) over 20% smaller leakage in low performance mode, and (4) around 30% smaller internal power in low performance mode.

With these customized repeaters, the entire design requires just one additional signal, LPM, to choose the op-

Cell	Input Capacitance (fF)			Rise Delay (ns)			Fall Delay (ns)		
	trad-	adjustable		trad-	adjustable		trad-	adjustable	
	itional	LPM=0	LPM=1	itional	LPM=0	LPM=1	itional	LPM=0	LPM=1
INVX8	6.663	6.664	6.732	0.039	0.042	0.059	0.018	0.019	0.035
BUFX8	2.577	2.582	2.590	0.060	0.064	0.088	0.079	0.084	0.117
Cell	Leakage (nW)			Rise Short-Circuit Energy (pJ)			Fall Short-Circuit Energy (pJ)		
	trad-	adjustable		trad-	adjustable		trad-	adjustable	
	itional	LPM=0	LPM=1	itional	LPM=0	LPM=1	itional	LPM=0	LPM=1
INVX8	84.628	90.957	62.837	0.019	0.018	0.013	0.010	0.010	0.006
BUFX8	125.316	135.076	93.211	0.014	0.013	0.011	0.019	0.019	0.015

Table 1: Input capacitance, delay, leakage, and short-circuit energy for two cells and their customized counterparts in regular and low-performance modes.

eration mode. The LPM signal has to be routed to all customized repeaters, and the $\overline{LPM}$ signal is locally generated as described later.

A few points are noteworthy:

- The control gates switch only when modes are changed and hence do not contribute to the dynamic power. The change in input capacitance of the switching devices is negligible as expected and verified by SPICE simulations (Table 1). Therefore, dynamic power does not increase due to this customization. Gate leakage is contributed by the control gates. However, in our technology it is negligible especially at operating temperatures and due to sharing of control gates as described later.
- At low-performance mode, due to the reduced drive strength, output slews increase which can lead to potential slew constraint violations. To ensure that no slew constraint violations are introduced, we convert custom repeaters back to traditional ones at the outputs of which slew constraints occur (described later). Also, we do not expect the susceptibility to noise or its magnitude to increase. In fact, due to increased slew, noise is likely to decrease. We note that separate signoffs must be performed for the regular and low-performance modes. This requirement, however, is true whenever multiple operating modes are present (e.g., in DVS).
- We do not expect the output voltage at low-performance mode to deteriorate because of the disabled devices. The disabled devices do not provide a low impedance path to power/ground and we have verified this expectation with SPICE simulations.

As noted earlier, the delay of a customized repeater depends on the size of its control gates. Figure 2 shows the delay overhead of a customized repeater in comparison to a traditional repeater as the width of the control gates is varied. W_{CG} is the width of the PMOS (NMOS) control gate and W_L is the total width of all PMOS (NMOS) gates being controlled by it. In all our experiments, we use a ratio of 4 for $W_{CG} : W_L$ since it offers a good tradeoff between delay penalty and area overhead. In our technology, gate leakage of the control gate is extremely small. If the gate leakage is considerable, then it can also be used to determine the $W_{CG} : W_L$ ratio. With a ratio of 4, an inverter with half its devices controlled has a customization overhead (i.e., area of the control gates) of 200%! In the next subsection we present a methodology to share the control gates among multiple repeaters to reduce the area overhead.

2.2 Area Overhead Reduction

The addition of control gates for each customized repeater can intolerably increase the total device area overhead. To reduce this overhead, we suggest that control gates be shared among a number of repeaters. Figure 3 illustrates a modified design, where the two inverters share the same control gates. In the new design, each customized inverter will have two new nets for two virtual power V'_{dd} and ground V'_{ss} signals driven from the common header (PMOS) and footer(NMOS) control gates.

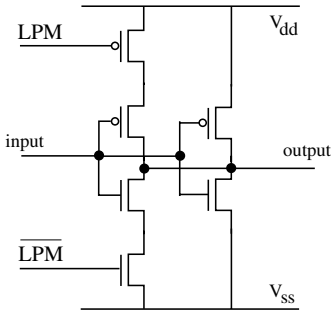

Figure 1: An adjustable inverter that can operate as either a $1\times$ inverter or a $2\times$ inverter.

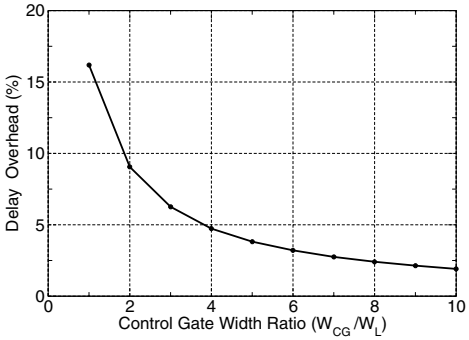

Figure 2: Delay overhead of the adjustable repeater vs. its control gate area.

Not all repeaters switch in all clock cycles and not all of the repeaters that switch in a clock cycle have overlapping time windows in which they switch. Clearly, the repeaters that do not have overlapping time windows can share a control gate without incurring a noticeable performance penalty. The number of repeaters that can share a common control header and footer is strictly controlled by the *simultaneous switch-on rate (SSR)* [10], which gives the maximum fraction of repeaters that have overlapping switching time windows. A high SSR indicates a large number of transistors switching at the same time, which increases the control gate size requirement to keep the voltage drop across the control gate small. We calculate the SSR of each design as follows: (1) the switching interval is calculated for each and every repeater during a clock period; (2) an interval graph is constructed; (3) the largest clique in the graph is computed; and (4) the SSR is the largest clique size divided by the total number of repeaters. Note that calculating the largest clique in interval graphs requires $O(R \log R)$ time, where R is the total number of repeaters. In our SSR calculation, we assume all repeaters to switch in every clock cycle (i.e., our calculation is pessimistic and we overestimate the control gate requirement). If switching activity is available (e.g., in the form of value change dump (VCD) or switching activity interchange format (SAIF)), less pessimistic SSR calcula-

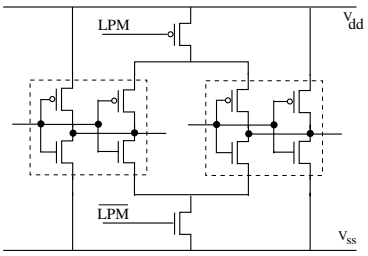

Figure 3: Reducing the control gate overhead. Transistors are not drawn in proportion.

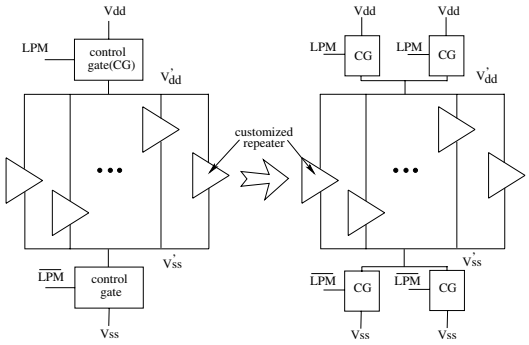

Figure 4: Distributing the control overhead.

tion, which will reduce the number of contol gates, can be performed.

Given the SSR and the total design repeater width W_d, the size of the shared control gates is readily calculated as follows. From the previous subsection, we know that the area overhead of a single repeater is twice the repeater's total width. Thus, the size of the shared control header and footer is equal to $2 \times W_d \times SSR$. The left part of Figure 4 gives a schematic of a control gate that distributes V'_{ss} and V'_{dd} to all customized repeaters. Note that the shared control header and footer transistors can be split into manageable-size ones, and all connected to the common virtual ground and supply nets as shown in the right part of Figure 4 (similar to [10]). In Figure 4, the smaller control gates are connected in parallel and form a "pool" to satisfy the current requirements of simultaneous switching repeaters.

2.3 Performance in Regular Operating Mode

Even though the adjustable repeaters are marginally slower than traditional ones, we ensure that performance is not compromised at high-performance modes by not using them on the timing-critical paths. In addition, we also ensure that the slew constraints are not violated in any mode of operation. In our flow, the repeaters are first characterized to estimate their power, input capacitances, delays, and slews at various input slew and load capacitance points. Then a synthesis tool is used to remap traditional repeaters to adjustable ones such that: (1) the critical path delay is not altered at the high-performance mode, and (2) slew constraints are never violated. In our experimental results, we have observed that almost all traditional repeaters get remapped to adjustable ones. This is due to the small difference in delay and input capacitance between traditional and adjustable repeaters, and to small slew times.

3. Experimental Validation

In our experiments we compare power reduction between the following two flows: (1) the proposed approach is used without DVS, and (2) the proposed approach is used with DVS to reduce power after voltage can no longer be scaled down. The nominal operating voltage is 1.1V and we assume that it can be scaled down to 0.9V.

circuit	cells	nets	speed (MHz)	leakage (mW)	dynamic (mW)
s38417	8890	8997	709	0.977	1.840
AES	15272	15887	445	1.797	2.400
OpenRisc	46732	53165	374	7.971	2.957

Table 2: Details of our testcases. Leakage gives the leakage power at 1.1V. Dynamic gives the dynamic power at the maximum frequency and 1.1V.

Experimental Setup

We use the following library, benchmarks, and tools:

- We use *Artisan TSMC* 90nm technology with *Synopsys Design Compiler W-2004.12-SP3* for logic synthesis.
- Delay and power characterization of standard cells (including the customized repeaters) is performed with *Cadence SignalStorm v4.1* and *Synopsys HSPICE U-2003.09*. We use $90nm$ BSIM4 SPICE models from a leading foundry that model both subthreshold and gate leakage.
- The three designs used in our experiments are summarized in Table 2. We assume an activity factor of 0.01 at primary inputs for dynamic power computation. We observe high leakage in Circuit *OpenRisc* primarily due to the presence of high-performance RAM.
- Timing-driven placement, clock tree generation, timing-driven routing, and parasitic extraction is performed with *Cadence Encounter v04.10*.
- We calculate circuit delay, dynamic power, and leakage with *Design Compiler*.

Experimental Flow

Our experimental flow is as follows.

1. Cell selection for characterization: Our library comprises 50 combinational cells, 12 sequential cells, and one latch. We create customized versions of four buffers (BUFX4, BUFX8, BUFX16, BUFX20) and four inverters (INVX4, INVX8, INVX16, INVX20).

2. Cell Characterization: We characterize all the selected cells, as well as the newly created customized cells, at worst delay conditions (slow process corner, 125C temperature) for both 1.1V and 0.9V. For the characterization of the customized repeaters, we assume PMOS and NMOS control gates to be 4× the size of PMOS and NMOS gates controlled by them. This is primarily due to the inability of characterization, which is performed per-cell, to handle control-gate sharing. We note that this characterization gives a pessimistic estimate of the delays because, as described in Section 2.2, we ensure that the total control-gate area is no less than 4× the area of simultaneously switching devices. Additionally, the gate leakage of control gates is overestimated because control gates are assumed to be dedicated for each customized repeater and not shared.

3. Netlist synthesis and physical design: We synthesize all designs under tight delay constraints using iterative synthesis. The slew constraint was set to 1ns in all our designs. Placement is performed with 60% utilization in timing-driven mode, followed by clock-tree generation with reasonable clock skew constraints, and finally timing-driven routing and parasitic extraction.

4. Logic remapping to adjustable repeaters: We remap the synthesized netlists, that have been back-annotated with parasitic delays, to use the adjustable repeaters where possible. This crucial step avoids introducing timing and slew constraint violations by using the adjustable repeaters only on the non-critical paths. With this selective remapping, we ensure that the circuit speed is not deteriorated in high-performance mode due to the use customized repeaters. The slew constraints are met at all operating modes.

Circuit	freq (MHz)	Traditional Buffering				Adjustable Buffering							
		slack (ns)	leakage (mW)	dyn (mW)	total	mode	slack (ns)	leakage (mW)	leakage Δ(%)	dyn (mW)	dyn Δ(%)	total (mW)	total Δ(%)
s38417	709	0.000	0.977	1.840	2.817	LPM=0	0.000	0.999	+2.25%	1.836	-0.22%	2.835	+0.64%
	683	0.055	0.977	1.771	2.748	LPM=0	0.055	0.999	+2.25%	1.768	-0.17%	2.767	+0.69%
	656	0.114	0.977	1.702	2.679	LPM=1	0.000	0.846	-13.41%	1.672	-1.76%	2.518	-6.01%
	616	0.214	0.977	1.597	2.574	LPM=1	0.100	0.846	-13.41%	1.569	-1.75%	2.415	-6.18%
AES	445	0.000	1.797	2.400	4.197	LPM=0	0.000	1.821	+1.33%	2.368	-1.33%	4.189	-0.19%
	438	0.035	1.797	2.364	4.161	LPM=0	0.037	1.821	+1.34%	2.329	-1.48%	4.150	-0.26%
	432	0.069	1.797	2.328	4.125	LPM=1	0.000	1.681	-6.46%	1.984	-14.78%	3.665	-11.15%
	389	0.325	1.797	2.096	3.893	LPM=1	0.271	1.681	-6.46%	1.776	-15.27%	3.457	-11.20%
OpenRisc	192	0.000	7.971	2.957	10.928	LPM=0	0.000	8.109	+1.73%	2.947	-0.34%	11.056	1.17%
	187	0.160	7.971	2.868	10.839	LPM=0	0.160	8.109	+1.73%	2.859	-0.31%	10.968	1.19%
	181	0.338	7.971	2.776	10.747	LPM=1	0.000	7.226	-9.35%	2.676	-3.60%	9.902	-7.86%
	173	0.593	7.971	2.653	10.624	LPM=1	0.282	7.226	-9.35%	2.558	-3.58%	9.784	-7.91%

Table 3: Performance and power results for all benchmarks when the technique is used without DVS.

V_{th}	Delay	Leakage	Switching Energy
SVT	+63%	-26%	-35%
HVT	+62%	-22%	-22%
LVT	+75%	-28%	-43%

Table 5: Delay, leakage and switching energy difference between a traditional INVX8 and a customized INVX8 in low-performance mode for three threshold voltages.

Results and Discussion

1. Performance and power calculation: We analyze the original and remapped netlists for power at various operating frequencies. We assess the power reduction at different operating frequencies when the proposed technique is used (a) as stand-alone and (b) used in conjunction with DVS.

(a) Adjustable buffering without DVS: We gradually reduce the speed of the circuit and compare the power when traditional repeaters are used and when customized repeaters are used. For the circuit with customized repeaters, low-performance mode is turned ON (LPM=1) and power is observed to reduce. Because of selective mapping the circuit with customized repeaters is able to operate at the peak circuit speed. Table 3 summarizes our results. We observe that both leakage and dynamic power reduce by $6-11\%$ when the LPM mode is turned on. We, however, see a marginal increase in leakage power at regular mode. We attribute this to the gate leakage of the added control gates. We note that this increase is an over-estimate because characterization, which estimates the leakage at the cell level and forms the basis of circuit leakage estimation, assumes each customized repeater to be controlled by a separate control gate. In reality, however, due to sharing of control gates by customized repeaters, the ratio between control gate area and logic area is much smaller, as discussed in Section 2.2. We also observe that the dynamic power decreases slightly in comparison to the traditional case even when the LPM mode is turned off. This is attributed to the slight drop in V_{dd} (V_{ss}) across the PMOS (NMOS) control gate to half of the PMOS (NMOS) devices.

From the results it is clear that adjustable buffering can reduce power when used with or without DVS. The testcase *s38417* has small circuit delay and slew rates leading to a large ratio between switching and short-circuit power. We also note that for the design *OpenRisc*, power reductions are less than the other benchmarks because a large percentage of the design consists of RAMs that we do not touch.

(b) Adjustable buffering with DVS: We assume that the voltage is scaled down to 0.9V from 1.1V and cannot be scaled down any further due to electrical, noise, or reliability issues. We successively reduce frequency and compare power when traditional repeaters are used and when customized repeaters are used. When customized repeaters are used, low-performance mode is turned ON when the performance requirements allow. Table 4 presents the results. We observe that in low-performance mode, power reduces by $5-7\%$. Leakage power and dynamic power marginally increase and decrease for the same reasons as when applying adjustable buffering without DVS.

In all our experiments standard threshold voltage (SVT) devices are used. Multi-V_{th} is a mainstream leakage reduction technique in which devices of multiple threshold voltages are used to tradeoff leakage and performance. Therefore it is important for adjustable buffering to work well for devices of other threshold voltages. Table 5 presents the delay-power tradeoff associated with low-performance mode for a customized INVX8 for three threshold voltages. We see similar tradeoff for low threshold voltage (LVT) and high threshold voltage (HVT) as SVT. Therefore, results similar to Tables 3 and 4 can be expected for circuits optimized with multi-V_{th}.

2. Area and Routing Overhead: As described in Section 2.2, we compute the SSR for each of our testcases and use it to estimate the area overhead due to the control gates. Table 6 summarizes our area overhead results. We observe that area overhead due to insertion of control gates is small and ranges between 0.91% to 5.57%. However, a large number, if not all, of the added control gates may be easily placeable in the whitespace since they connect to global interconnects only. Therefore, in practice we expect the area overhead to be negligible. As for the routing area overhead, we estimate the wirelength required to route V'_{dd}, V'_{ss} to the customized repeaters as minimum Steiner trees. We find the wirelength overhead is on the average 3.41% of the total routed wirelength. Note that the LPM signal is directly connected to the control gate, and does not need any routing to the customized repeaters. Since the number of control gates is relatively small due to their sharing by customized repeaters, we expect the routing overhead of the LPM signal to be small. The $\overline{LPM}$ signal can either be routed with LPM signal or generated locally from LPM to feed to a group of customized repeaters. If the technique is used with a runtime failure detection methodology such as [4], that generates the LPM signal automatically, the LPM signal routing overhead can be further reduced.

4. Conclusions

In this paper we presented a new method for total runtime power reduction. We add control gates to operate the customized repeaters either in regular mode or in low-power mode. At high frequencies, the repeater driving capacity is similar to a traditional repeater of same size. At lower frequency, however, it is possible to reduce the repeater driving capacity, which in turns reduces both leakage and dynamic power. We have proposed efficient customization

Circuit	freq (MHz)	Traditional Buffering				Adjustable Buffering							
		slack (ns)	leakage (mW)	dyn (mW)	total	mode	slack (ns)	leakage (mW)	leakage Δ(%)	dyn (mW)	dyn Δ(%)	total (mW)	total Δ(%)
s38417	575	0.000	0.645	1.153	1.798	LPM=0	0.000	0.650	+0.78%	1.153	0.00%	1.803	+0.28%
	543	0.103	0.645	1.089	1.734	LPM=0	0.103	0.650	+0.78%	1.089	0.00%	1.739	+0.29%
	511	0.219	0.645	1.024	1.669	LPM=1	0.000	0.570	-11.63%	1.019	-0.49%	1.589	-4.79%
	479	0.350	0.645	0.960	1.605	LPM=1	0.131	0.570	-11.63%	0.955	-0.52%	1.525	-4.98%
AES	354	0.000	1.149	1.059	2.208	LPM=0	0.000	1.155	+0.52%	1.053	-0.57%	2.208	0.00%
	349	0.041	1.149	1.044	2.193	LPM=0	0.041	1.155	+0.52%	1.038	-0.57%	2.193	0.00%
	343	0.091	1.149	1.026	2.175	LPM=1	0.000	1.082	-5.83%	0.946	-7.80%	2.028	-6.76%
	337	0.142	1.149	1.008	2.157	LPM=1	0.056	1.082	-5.83%	0.929	-7.84%	2.011	-6.77%
OpenRisc	164	0.000	5.764	2.163	7.929	LPM=0	0.000	5.799	+0.61%	2.163	0.00%	7.962	+0.42%
	159	0.201	5.764	2.095	7.859	LPM=0	0.201	5.799	+0.61%	2.095	0.00%	7.894	+0.45%
	154	0.405	5.764	2.029	7.775	LPM=1	0.000	5.334	-7.64%	2.019	-0.49%	7.353	-5.45%
	149	0.623	5.764	1.963	7.727	LPM=1	0.253	5.334	-7.64%	1.943	-1.02%	7.277	-5.82%

Table 4: Performance and power results for all benchmarks when the technique is used with DVS at lowered supply voltage. Results for high supply voltage with DVS are the same as in Table 3.

circuit	total area (μ^2)	total repeater area (μ^2)	SSR	customization overhead	total new area (μ^2)	area Δ(%)	routing overhead (%)
s38417	77294	15929	13.5%	4304	81598	5.57%	3.46%
AES	112749	15577	7.95%	2480	115229	2.20%	3.37%
OpenRisc	931963	91056	9.77%	8505	940468	0.91%	3.41%

Table 6: Adjustable repeater area overhead. Total area is total original device area including the repeater area. SSR is the calculated maximum simultaneous switching-on ratio of the circuit. Customization overhead is the total area of LPM control gates. Total new area is the new customized total device area. Area Δ is the percentage change in total device area. Routing overhead gives the routing requirements in comparison to the total wirelength.

strategies and carefully analyzed the area and routing overhead required by this customization. We have also examined how to customize the entire design while meeting the high-performance requirements. This has been achieved by careful selection of the control gate sizes, as well as selective customization of non-critical repeaters. We assessed the power reductions due to the proposed approach when used with and without DVS. With state-of-the-art design tools and real-world benchmarks, customized buffering gives an average reduction of 8.34% in total system power, while not impacting the performance at the highest-frequency modes. The average area overhead for our approach is 2.94% and we expect the additional control gates to fit in whitespace areas in today's designs. Assuming V'_{dd} and V'_{ss} to be routed as steiner trees, we estimate the routing overhead to be 3.41%. In our evaluation, such a tradeoff is acceptable for many, if not all, low-power, battery-powered designs.

Our **future work** is in two directions: (1) rigorous area and routing layout implementation, and (2) enhancements for further power reduction. In the first direction, we plan to create layouts of the customized repeaters and control gates using a layout editor, e.g., Cadence Virtuoso. Next, using the extracted layout parasitics, we will characterize the customized repeaters more accurately. Finally, the customized repeaters will be placed in the whitespace, and the design rerouted and extracted. Such a flow enables a more accurate estimation of power and performance, and area overhead.

In the second direction, we plan to customize clock tree repeaters and other logic cells such as NANDs, NORs, ANDs and ORs. This enhancement would open room for more significant power reduction since a large percentage of dynamic power is consumed by the clock repeaters due to their high activity, and in logic cells due to their larger numbers.

5. References

[1] Synopsys. Power Compiler User Guide. 2003.

[2] A. Agarwal, C. H. Kim, S. Mukhopadhyay and K. Roy. Leakage in Nano-Scale Technologies: Mechanisms, Impact and Design Considerations. In *Proc. ACM/IEEE Design Automation Conference*, pages 6–11, 2004.

[3] T. D. Burd, T. A. Pering, A. J. Stratakos and R. W. Brodersen. A Dynamic Voltage Scaled Microprocessor System. *IEEE Intl. Solid-State Circuits Conference Digest of Technical Papers*, pages 294–295, 2000.

[4] D. Ernst, S. Das, S. Lee, D. Blaauw, T. Austin, T. Mudge, N. S. Kim and K. Flautner. Razor: Circuit-level correction of timing errors for low-power operation. *IEEE Micro*, 24(6):10–20, November 2004.

[5] P. Gupta, A. B. Kahng, P. Sharma and D. Sylvester. Selective Gate-Length Biasing for Cost-Effective Runtime Leakage Control. In *Proc. ACM/IEEE Design Automation Conference*, pages 327–330, 2004.

[6] J. Halter and F. Najm. A Gate-level Leakage Power Reduction Method for Ultra Low Power CMOS Circuits. In *IEEE Custom Integrated Circuits Conference*, pages 475–478, 1997.

[7] M. Horiguchi, T. Sakata and K. Itoh. Switched-Source-Impedance CMOS Circuit for Low Standby Sub-Threshold Current Giga-Scale LSI's. *IEEE Journal of Solid-State Circuits*, 28(11):1131–1135, 1993.

[8] I. Hyunsik, T. Inukai, H. Gomyo, T. Hiramoto and T. Sakurai. VTCMOS Characteristics and its Optimum Conditions Predicted by a Compact Analytical Model. In *Intl. Symp. on Low Power Electronics and Design*, pages 123–128, 2001.

[9] J. Kao, S. Narendra and A. Chandrakasan. MTCMOS Hierarchical Sizing Based on Mutual Exclusive Discharge Patterns. In *Proc. ACM/IEEE Design Automation Conference*, pages 495–500, 1998.

[10] S. Mutoh, T. Douseki, Y. Matsuya, T. Aoki, S. Shigematsu and J. Yamada. 1-V Power Supply High-Speed Digital Circuit Technology with Multithreshold-Voltage CMOS. *IEEE Journal of Solid-State Circuits*, 30(8):847–854, 1995.

[11] S. Mutoh, S. Shigematsu, Y. Matsuya, H. Fukada, T. Kaneko and J. Yamada. 1V Multithreshold-Voltage CMOS Digital Signal Processor for Mobile Phone Application. *IEEE Journal of Solid-State Circuits*, 31(11):1795–1802, 1996.

[12] P. Saxena, N. Menezes, P. Cocchini and D. Kirkpatrick. Repeater Scaling and its Impact on CAD. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 23(4):451–463, 2004.

[13] S. Shigematsu, S. Mutoh, Y. Matsuya, Y. Tabae and J. Yamada. A 1-V High-Speed MTCMOS Circuit Scheme for Power-Down Application Circuits. *IEEE Journal of Solid-State Circuits*, 32(6):861–869, 1997.

[14] L. Wei, Z. Chen, M. Johnson, K. Roy and V. De. Design and Optimization of Low Voltage High Performance Dual Threshold CMOS Circuits. In *Proc. ACM/IEEE Design Automation Conference*, pages 489–494, 1998.

[15] Q. Wu, M. Pedram and X. Wu. Clock-gating and its application to low power design of sequential circuits. *IEEE Trans. Circuits and Systems I: Fundamental Theory and Applications*, 47(3):415–420, 2000.

[16] Y. Ye, S. Borkar and V. De. A New Technique for Standby Leakage Reduction in High-Performance Circuits. In *Proc. Symp. on VLSI Circuits*, pages 40–41, 1998.