Highly Power Efficient, Uncompromised Performance Cache Design Using Dual-Edged Clock

Rajesh Kannan Megalingam, Nived Krishnan, Arjun Ashok V., Arunkumar M.
Dept. of Electronics and Communication, Amrita School of Engineering,
Amrita Vishwa Vidyapeetham, Amritapuri Campus,
Kollam-690525, Kerala
rajeshm@amritapuri.amrita.edu, nivedkrish@ieee.org,
arjunashok619@ieee.org, arunkumar.m@ieee.org,

Abstract—**Dynamic power dissipation is a major field where various researches have been held to reduce the energy consumption in present clocked systems. Here we discuss the various tiers of power management in design followed by a treatise on dynamic power dissipation in CMOS circuits and power efficient cache designs under research. We further propose a technique in reducing the clock frequency of circuits without compromising the performance. This is achieved by using a dual-edged clock for the operation, allowing the operating frequency to be halved, which immediately translates into a tremendous gain in power efficiency. The applications of this technique in processor caches are also studied. The designs proposed herein were implemented using Verilog HDL and the power requirements analyzed using Xilinx XPower 10.1. Experimental results and the conclusions drawn are included.**

Keywords-Dual-edged clock; CMOS; Dynamic power;

I. INTRODUCTION

Caches are normally the largest structures in a processor. Most of the existing processors have data caches whose size is in the range 16K to 1M. Caches may consume up to 50% of a microprocessor's total power including both static and dynamic. [2]

However, reducing cache power dissipation is a delicate process, because power conservation at the cost of performance usually precipitates increased power consumption. Current techniques being explored for reducing cache power consumption include bit-line segmentation, sub-banking, gray code addressing and cache partitioning.

This paper proposes a technique that holds promise for reducing power consumption in caches – dual-edged clock access.

II. STATIC AND DYNAMIC POWER

The average power consumption in CMOS digital can be expressed as the sum of three main components:[3]

1) *The dynamic (switching) power consumption,*
2) *The short-circuit power consumption,*
3) *The static power consumption.*

We'll limit our discussion to the conventional static and dynamic power dissipation.

Static dissipation is due to leakage current or other current drawn continuously from the power supply. Dynamic power is due to switching transient current and due to charging and discharging of load capacitances.

The dynamic power dissipation is of more importance to us, this occurs mainly when the node voltage of a CMOS logic gate makes a logic transition. Current also charges and discharges the output capacitive load and this capacitive charging and discharging current is dominant in dynamic dissipation. During the charge-up phase, the output node voltage typically makes a full transition from 0 to V_{DD} and one half of the energy drawn from the power supply is dissipated as heat in the pMOS transistors. As output voltage drops from VDD to 0, the energy stored in the output capacitance during charge-up phase is dissipated as heat in the conducting nMOS transistors.

The main cause of static dissipation due to leakage of current, leakage of current is due to reverse bias leakage between diffusion regions and substrate. The sub- threshold conduction also contributes to the static dissipation.

The static power dissipation is given by the product of device leakage current and the supply voltage. It usually occurs in pseudo nMOS gates, where there is direct path between power and ground. [10]

III. EXISTING POWER OPTIMIZATION TECHNIQUES

Different power optimization techniques in use are at system architecture level, micro-architecture level, RTL level and gate/physical implementation level. SW-HW partitioning, OS/firmware-level APLs for standby/sleep modes, single core vs. multi cores, bus and memory architecture and communication vs. computation tradeoffs are optimization techniques at the system architecture level. Frequency and voltage scaling, memory/register file banking, auto-inferencing of appropriate FIFOs and other communication channels form optimization techniques at the micro-architecture level. Combinational clock gating, sequential clock gating and power gating are techniques at the RTL level. Multi-V_{DD}, multi-V_{th} technology mapping, clock network optimization, high-k transistors and novel

978-1-4244-4520-2/09/$25.00 ©2009 IEEE

circuit structures/logic families are some techniques at the Gate/Physical implementation level.[11]

IV. DYNAMIC POWER DISSIPATION IN CMOS CIRCUITS

CMOS circuits consume very less power when their inputs trigger signals are static. In contrast, a significant portion of the total power consumption occurs when the inputs or other triggers are made to toggle. During the transition state, current is drawn directly from the supply, while the only current being drawn in a static state is contributed by the leakage current. This transition power is termed as dynamic power. The more the number of transitions a CMOS circuit or a memory experiences, the more is the power consumed. It is often seen that, clock signals have the highest toggling rates.

Dynamic power, P_D, is a factor of load capacitance, C_L, supply voltage, V_{DD} and clock frequency, f_C.

$$P_D = C_L V_{DD}^2 f_C \text{ [3]} \qquad (1)$$

Since P_D is proportional to f_C, reducing the clock frequency decreases the dynamic power consumption by the same factor. However, reduction in frequency usually translates to a corresponding hit in performance.

V. POWER EFFICIENT DATA CACHE DESIGNS

Some of the already proposed ideas for improving power efficiency in caches are discussed below. Leakage power can be reduced by powering off cache lines whose content is not expected to be reused. Sub- banking which saves bit line energy, multiple line buffers that save accesses of the data arrays but, at the same time, also save the access to the tag arrays and bit-line segmentation reduce dynamic power dissipation in superscalar processor caches. [7]

Other methods include vertical cache partitioning, horizontal cache partitioning, Gray code addressing to reduce dynamic power, different low power SRAM circuit design strategies like divided bit line, pulsed word line and isolated bit line. Study of dependence of lowering the supply and threshold voltages on the energy efficiency of CMOS circuits yields power efficient cache designs. [1]

Clock gating the energy recovery clock (sinusoidal clock) reduces the flip-flop power during idle periods [5]. Different cache bank predictors on clustered micro-architectures with a distributed L1 data cache and disambiguation hardware reduce the power requirements [6]. Energy usage is reduced by partitioning tag arrays into two parts and accessing in two phases [8].

VI. CONVENTIONAL CACHE ACCESS

Cache memories play an important role in today's processors in that they contribute to increased performance in a cost-effective manner. They work by bypassing high latency access to main memory and letting the processor access the low latency cache memory instead. It would certainly be uneconomical to replace the larger capacity main memory with high performance variants.

Since cache sizes are typically far lesser than main memory sizes, the data that is stored in cache needs to be extremely selective. The principles of spatial locality and temporal locality are used for deciding which locations populate the cache. Spatial locality dictates that if a memory location is accessed, the locations near to it are also likely to be accessed soon. Temporal locality dictates that if a memory location is accessed, the same is likely to be accessed again soon. These two principles allow the cache to be populated in such a way that it has highest probability of containing requested information.[9]

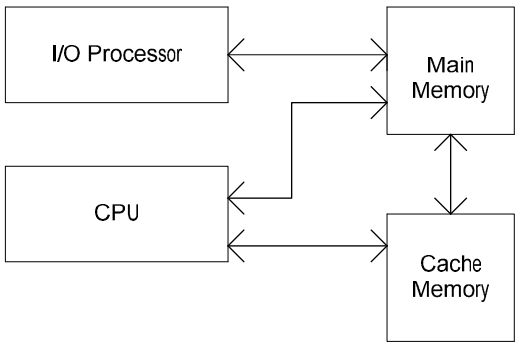

Figure.1. Memory organization in a conventional system

Studies of memory access behavior of processors have revealed that most of the access is confined to some few localized areas of memory [3].

Conventional caches are single-edge triggered devices. They access their content only during either the positive clock edge or the negative clock edge. Their frequency of operation is usually in the GHz range for the current generation processors. Thus they contribute to a significant amount of the dynamic power consumed by the chip.

VII. DUAL-EDGED CLOCK ACCESS

If the clock frequency, f_C, can be reduced without affecting cache performance in a significant way, it translates to a direct decrease in power needs. The proposed method makes use of a clock frequency that is halved compared to a conventional clock, with little performance impact.

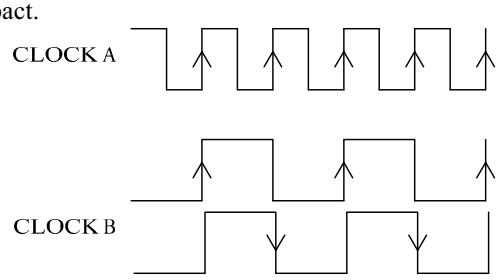

Figure. 2. Comparison of the conventional clock and postulated clock.

It is obvious from the figure 2 that the two clocks, one of frequency f and the other of frequency f/2, have edges at the same instants, albeit with a change in the sense. Therefore, if both the edges of clock B could be used for triggering, the net gain in power consumption would be huge, the performance remaining almost unaffected compared to triggering using clock A.

Consider the two 4-bit shift registers shown in figure 3. The first one in figure 3a., receives clock A and the second in i3b., receives clock B as specified figure 2. At the end of four cycles of clock A, or correspondingly, two cycles of clock B, the input gets shifted to the output of both the 4 – bit shifter registers shown in figure 3. This is an interesting result in that we are able to achieve exactly the same performance using clocks of two widely different frequencies. Note that the power consumption values of these two configurations are distinctly different. This is proved as follows.

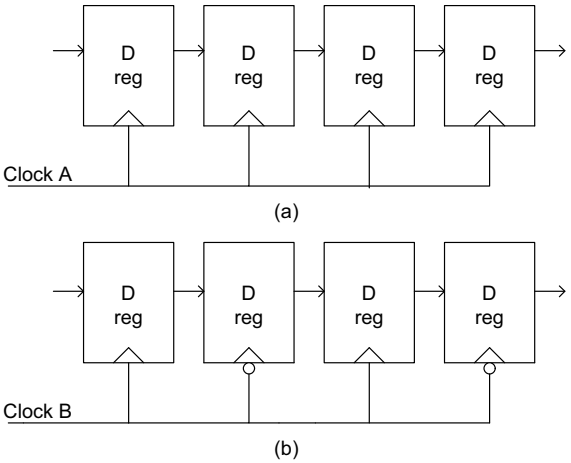

Figure.3. Logical organization of (a) traditional shift register versus (b) dual-edge triggered shift register.

If one D flip-flop consumes power P at operating frequency f, it consumes power P/2 at power at f/2, as in equation (1). In the first shift register of figure 3a., if the operating frequency is f, we have total power consumed as 4P, whereas in the second shift register, the power consumed is 4(P/2), or 2P, or half of the former.

If we are able to implement the same design scheme for large scale memories and other circuits, it would lead to tremendous gains in power efficiency. Here, we have applied this idea to cache memory. Therefore, even a slight reduction in the cache power requirement can provide significant increase in overall power efficiency.

To achieve this, assuming a direct-mapped cache, the entire cache memory block is conceptually divided into two banks, each containing only either odd or even locations as shown in figure 5. The design is such that even-numbered addresses are served during the positive edge of the clock, and odd-numbered addresses by the negative edge.

The downside is that if an even-numbered address is requested during a negative edge, it needs to be relegated to the subsequent positive edge, or vice-versa.

This translates into a loss of a single clock cycle with respect to the conventional clock. However, this is a small delay compared to that incurred by a cache miss, and with intelligent addressing, can be entirely avoided. Many programs especially of multimedia applications exhibit high spatial locality by operating nearly sequentially, that too most of the time in loops [4]. Since the clock frequency in dual-edged clock access can be half of that of conventional systems for equivalent performance, the power consumption is also theoretically reduced by the same factor.

In the case of a set-associative cache, the above model translates into alternate access for even and odd-numbered sets during positive and negative edges respectively. Obviously, the order of access can be interchanged to mean odd-positive/even-negative without affecting the model.

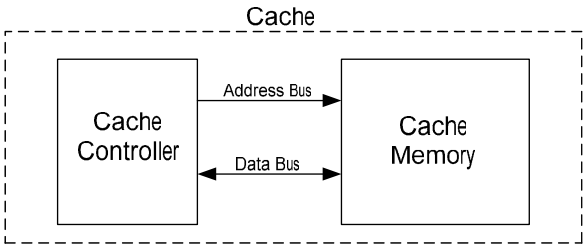

Figure.4. Cache internal organization.

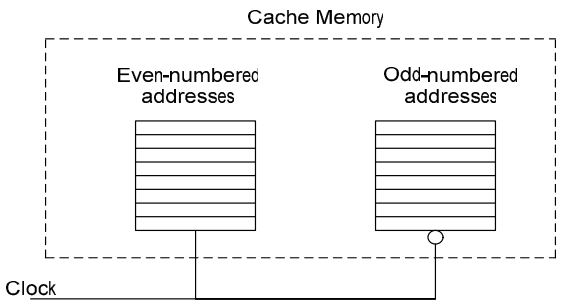

Figure.5. Proposed cache memory internal organization.

From equation (1), we have,

$$P_{D1} = C_L V_{DD}^2 f_A \text{ [3]}$$

$$P_{D2} = C_L V_{DD}^2 f_B \text{ [3]}$$

$$f_A = 2f_B$$

Therefore,

$$P_{D2} = P_{D1}/2 \quad (2)$$

VIII. EXPERIMENTAL OBSERVATIONS AND RESULTS

The experimental procedure followed in our research is diagrammatically represented in figure 6. The hypothetical cache with a 100% hit ratio and logical architectures true to the previously proposed idea was designed and coded in Verilog HDL.

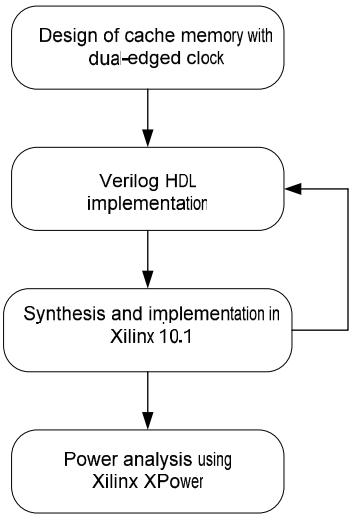

Figure.6. Experimental methodology.

A portion of the simulation output is given in figure 7. This simple proof of concept design was used for verification of the propounded design.

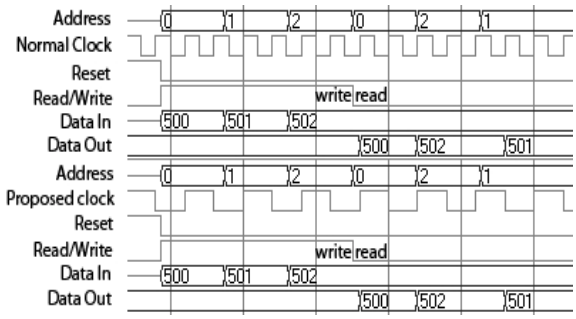

Figure.7. Timing diagram obtained from the simulation result.

Next, this design was imported into Xilinx 10.1 and implemented in Spartan2 xc2s200 device. Power analysis of the same was performed for various clock frequencies using Xilinx XPower. The same procedure was repeated for two different variants of the design viz.

1) *Clock dependant reads, without output buffers:*
In the verilog code designed, read operation is carried out at every clock edge when read-enable is set. The output lines were made to retain the previous driven values until the next address value from which the data to be read arrives. Power consumption values corresponding to clock frequencies 10MHz to 1280MHz were obtained. The collected data are given in table I. The observations obtained are plotted in figure 8. It is obvious from the figure that power reduction is not much significant at lower frequencies. At higher frequencies, power consumed has been reduced almost to 50% as predicted theoretically. Even for the same clock frequency, the proposed model consumes 5% lesser power than the conventional design. To achieve an operating frequency of 1280MHz, the clock of the conventional single edge clock should be running at 1280MHz, thereby consuming power worth 0.84W whereas to achieve the same speed of operation, dual-edge clock frequency has to be 640 MHz, producing power worth 0.41W. We readily observe that the reduction in power consumed by using the proposed design is 50%.

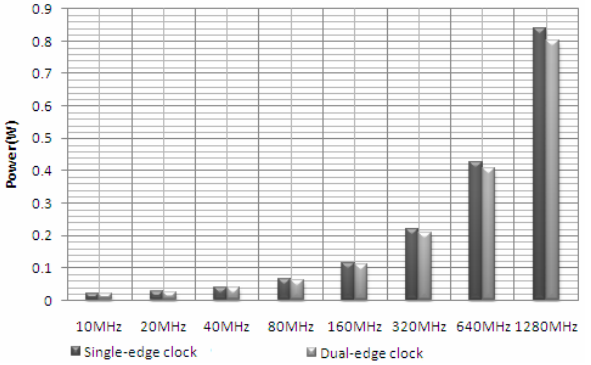

Figure.8. Output retaining previous value and clock dependant read

TABLE I POWER CONSUMPTION VALUES OBTAINED

Operating Frequency (MHz)	Conventional Cache		Proposed Cache	
	Clock Frequency (MHz)	Power(W)	Clock Frequency (MHz)	Power(W)
20	20	0.026	10	0.019
40	40	0.039	20	0.025
80	80	0.065	40	0.038
160	160	0.116	80	0.062
320	320	0.219	160	0.112
640	640	0.426	320	0.210
1280	1280	0.838	640	0.407

2) *Clock independent reads, without output buffers:*
In the verilog code designed, read operation is carried out whenever read enable is set. Read operation is carried out irrespective of clock edge. The output lines are turned off or driven to 'Z' once read operation is disabled. Here also the power consumption values corresponding to clock frequencies 10MHz to 1280MHz were collected. The collected data are given in table II. The observations obtained are plotted in figure 9. Similar to the previous design, this concept holds good at higher frequencies than lower frequencies. At higher frequencies, as expected power consumption has been brought down to almost 50% of the single-edged clock accessed conventional cache. At the same clock frequency, the propounded design has very less power gain, nearly 2%. To achieve an operating frequency

of 1280MHz, the clock of the conventional single edge clock should be running at 1280MHz, thereby consuming power worth 0.84W. The clock frequency of dual-edge clock to achieve the same speed of operation is just 640 MHz, producing power worth 0.42W. We readily observe that the reduction in power consumed by using the proposed design is 50%.

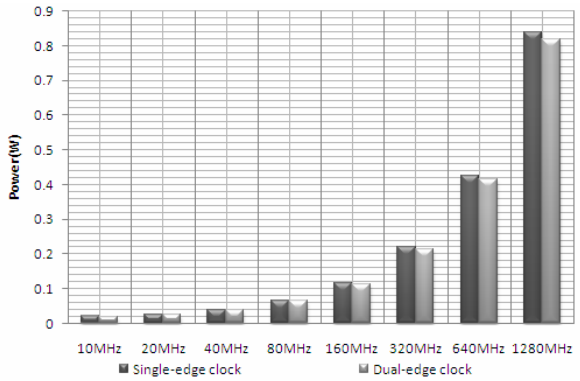

Figure.9. Output retaining previous value and clock independent read

TABLE II POWER CONSUMPTION VALUES OBTAINED

Operating Frequency (MHz)	Conventional Cache		Proposed Cache	
	Clock Frequency (MHz)	Power(W)	Clock Frequency (MHz)	Power(W)
20	20	0.026	10	0.019
40	40	0.039	20	0.026
80	80	0.065	40	0.038
160	160	0.116	80	0.064
320	320	0.219	160	0.114
640	640	0.426	320	0.214
1280	1280	0.838	640	0.416

As a control, the normal single-edge triggered cache design was also simultaneously implemented and analyzed.

The results indicate that implementations using output tri-state buffers consume more power than those using output registers. These results are contradictory and indicate one of the limitations of the experiment in that no bus logic was concurrently implemented. Results would no doubt agree with expectations if this were not the case.

As far as power consumption is concerned, in all the cases, the suggested cache design has lower power consumption than the control design at the same frequencies. Moreover, the underlying principle of the design is that it is capable of almost the same performance as a traditional design at half the operating frequency, or correspondingly almost twice the performance at the same operating frequency. It is readily apparent that a reasonably large saving in power is achieved. Furthermore, the difference in power consumption between f and f/2 is far greater at higher operating frequencies as compared to lower operating frequencies. This holds promise in a world where memories keep getting faster.

IX. CONCLUSION

Usage of dual-edged clock has proven to be an efficient method both theoretically and practically that it can be applied to the present cache systems to reduce the dynamic power dissipation. Theoretically, 50% reduction in power was expected. But as per the observations made using the Xilinx tool, reduction in power was a bit less than the expected though an appreciable reduction has been observed. At lower frequencies, the power dissipation was not much significant. Even for the same clock frequency, the suggested design has shown to be efficient with slightly lesser power consumption. Hence we conclude that this method is highly efficient in high frequency clocked circuits.

ACKNOWLEDGMENT

We gratefully acknowledge the Almighty GOD who gave us strength and health to successfully complete this venture. The authors wish to thank Amrita Vishwa Vidyapeetham, in particular the Digital library, for access to their research facilities.

REFERENCES

[1] Jaume Abella and Antonio González, "Power Efficient Data Cache Designs", Proceedings of the 21st International Conference on Computer Design (ICCD'03).

[2] Chuanjun Zhang, "An Efficient Direct Mapped Instruction Cache for Application-Specific Embedded Systems", Sept. 19–21, 2005, Jersey City, New Jersey, USA.

[3] Neil.H.Weste and Kamran Eshraghian, "Principles of CMOS VLSI design".

[4] Praveen Kalla and Xiaobo Sharon Hu, "Distance-Based Recent Use (DRU): An enhancement to Instruction Cache Replacement Policies for Transition Energy Reduction", IEEE transactions on very large scale integration (VLSI) systems, vol. 14, no. 1, january 2006.

[5] Vishwanadh Tirumalashetty and Hamid Mahmoodi, "Clock Gating and Negative Edge Triggering for Energy Recovery Clock",2007.

[6] Stefan Bieschewski, Joan-Manuel Parcerisa1, and Antonio González "Memory Bank Predictors", Proceedings of the 2005 International Conference on Computer Design.

[7] Kanad Ghose and Milind B. Kamble, "Reducing power in superscalar processor caches using sub-banking, multiple line buffers and bit-line segmentation", 1999.

[8] Rui Min, Wenben Jone and Yiming Hu, "Phased Tag Cache: An efficient Low Power Cache System", 2004.

[9] Moris Mano, "Computer System Architecture", Third Edition.

[10] Sung–Mo Kang and Yusuf Leblebici, "CMOS Digital Integrated Circuits", Third Edition, Tata McGraw–Hill edition,2003.

[11]Anmol Mathur, Qi Wang and Vani Dimri, "Power Reduction Techniques at the RTL and System Level", 22nd International Conference on VLSI Design, January 5–9, 2009.