A CPL-based Dual Supply 32-bit ALU for Sub 180nm CMOS Technologies

Bhaskar Chatterjee
Dept. of Electrical and Computer Engineering
University of Waterloo
Waterloo, ON, Canada
bhaskar@vlsi.uwaterloo.ca

Manoj Sachdev
Dept. of Electrical and Computer Engineering
University of Waterloo
Waterloo, ON, Canada
msachdev@vlsi.uwaterloo.ca

Ram Krishnamurthy
Circuits Research Lab,
Intel Corp.
Hillsboro, OR, USA
ram.krishnamurthy@intel.com

ABSTRACT

In this paper we present the design of a high performance 32-bit ALU for low power applications. We use dual power supply scheme and CPL logic for non-critical units of the ALU. In addition, latches with only n-MOS clocked transistors are used to interface logic operating at different power supplies and achieve static power free operation. Our simulation results indicate that, for the 180nm-65nm CMOS technologies it is possible to reduce the ALU total energy by 18%-24%, with minimal delay degradation. In addition, there is up to 22%-32% reduction in leakage power in the standby mode.

Categories and Subject Descriptors

B.7.1 [Integrated Circuits]: Types and Design Styles—*advanced technologies, VLSI (very large scale integration)*

General Terms: Performance, Design, and Theory

Keywords:DSM leakage control and scaling trends, dual supply ALU design, low power techniques.

1. INTRODUCTION

Modern microprocessors operate at clock frequencies more than 3GHz and have close to 100 million transistors on die. The need for improved performance and more functionality has resulted in aggressive technology scaling. As this trend continues into the future, it is expected that the device geometry, transistor threshold (V_{TH}) and supply (V_{DD}) voltages will be scaled further. This will lead to degraded short channel effects (*SCE*) and increased transistor OFF-state (I_{OFF}) current.

In addition, higher operating frequency, leakage currents and on-die transistor count will result in an increase in total power. Some of these scaling trends are shown in Figure 1 using data from ITRS reports (2001-2002) [1]. The ITRS long-term projections indicate that, by the year 2016, on-die clock frequency of high-end microprocessors might reach 29GHz while the total power consumption would be about 288W. This will offset the savings in switching energy obtained from technology scaling and result in lower battery life for mobile devices. This will increase the possibility of thermal hot spots and run away during stress testing (burn-in). Consequently, the long-term reliability of high end digital ICs may be compromised in deep submicron (*DSM*) technologies.

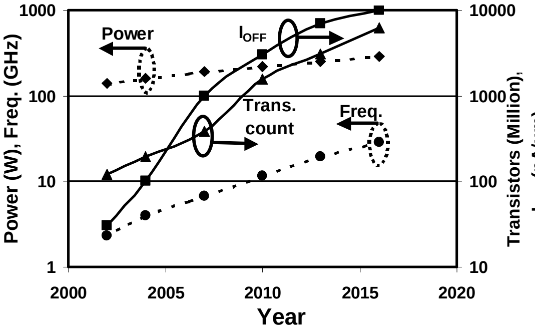

Figure 1: ITRS roadmap near and long-term projections

Several design techniques have been proposed to minimize transistor leakage and system level power consumption in high performance ICs [2-7]. These include transistor level leakage control techniques such as: dual V_{TH} techniques [2], multi-oxide or non-minimum channel length transistors [4], reverse body bias (*RBB*), and stack effect. In this paper, we present the design of a CPL-based dual supply, 32-bit ALU and demonstrate the scaling trends of its delay, total energy and leakage power for the 180nm-65nm bulk CMOS technologies. Our 180nm results correspond to a TSMC process while the 130nm-65nm technology results pertain to the Berkeley Predictive Technology Models [8, 9]. The rest of this paper is organized as follows: In section 2, we discuss the impact of supply voltage reduction on transistor leakage currents. In section 3 we present circuit level techniques used in the ALU design to achieve low power operation. We discuss the energy-delay tradeoffs and ALU scaling trends in section 4, while section 5 is for conclusions.

2. Supply Scaling and Transistor Currents

The impact of supply voltage scaling on transistor leakage and ON-state saturation currents is discussed in this section. A simplified expression for the transistor OFF state current is given by [2]:

$$I_{OFF} \approx Ae^{\left(\frac{V_{GS}-V_{TH0}-\gamma V_{SB}+\eta V_{DS}}{nv_T}\right)} \qquad (1)$$

where, $A = \mu_0 C_{ox} W/L_{eff}\ v_T^2 e^{1.8}$. Supply scaling reduces the transistor drain-source voltage and helps to minimize the *DIBL* current. The reduction in the transistor OFF-state current due to supply scaling can be expressed as [12]:

ISLPED'04, August 9–11, 2004, Newport Beach, California, USA.
Copyright 2004 ACM 1-58113-929-2/04/0008…$5.00.

$$\left.\frac{\Delta I_{OFF}}{I_{OFF}}\right|^{V_{DD}} = 1 - e^{-\frac{\eta \Delta V_{DS}}{n v_T}} \qquad (2)$$

The data in Figure 2 indicates that a 30% reduction in supply voltage results in up to 32% reduction in the I_{OFF} current while lowering by the I_{GATE} component by 84%. However, this also results in lower gate overdrive voltage and reduces the I_{DSAT} by ~48%. This may cause performance degradation in DSM logic circuits. In subsequent sections, we will demonstrate the selective usage of a dual supply scheme to reduce ALU total energy consumption, while maintaining the performance degradation within acceptable limits.

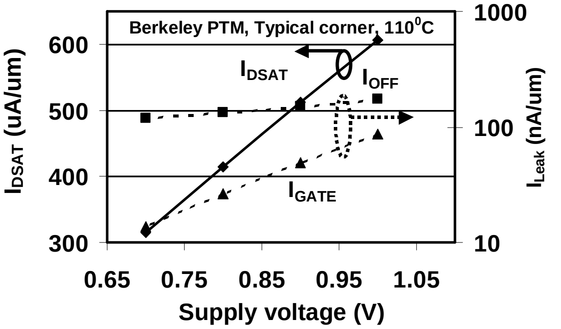

Figure 2: Supply scaling and 65nm transistor currents

3. Low Power Circuit Techniques for ALU Design

In this section, we focus on some of the circuit strategies adopted in this design to achieve low power ALU operation:

1. Using reduced swing clocking scheme for the non-critical path latches and flip-flops [11],
2. Using CPL logic and C^2MOS MUX-es to design the logic/shifter units to minimize overall switching capacitance and data buffer/driver sizes,
3. Designing the 32-bit adder PG unit (Propgate-generate unit) using shared clock footers and reducing buffer sizes.

3.1 Latch Design for Dual Supply Clocking

Traditionally, static latches, and master-slave FFs use transmission gate (TG) based designs that have both n-MOS and p-MOS clocked transistors. In order to maintain high performance while saving total energy, our goal was to keep the datapath circuitry at nominal supply voltage while lowering the clock swing of the latch/flip-flop clock transistors. However, under such a scheme, the TG p-MOS transistors do not turn OFF fully, resulting in static current (power) consumption. This problem is further aggravated for high performance datapath designs that normally operate at high junction temperatures (high switching frequency) and lower V_{TH} and thus exponentially higher I_{OFF}/µm currents. Figure 3(a-b) shows the circuitry and energy-delay tradeoffs of a latch that uses only n-MOS clocked transistors allowing static power-free dual supply ALU operation [11]. The master-slave flip-flops used in this design were obtained by cascading 2 of the latches shown in Figure 3(a). Simulation results indicate that as the clock swing is reduced (V_{DDL}) while maintaining nominal supply voltage (V_{DDH}) on the data transistors, total energy is reduced, while D→Q delay and T_{setup} are degraded.

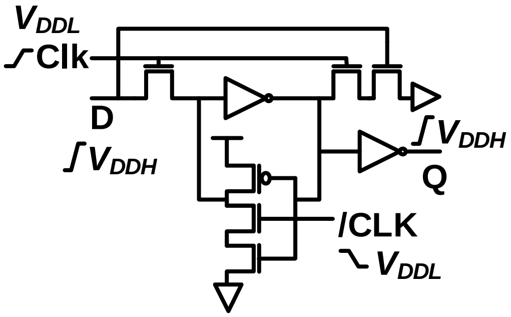

Figure 3(a): Latch supporting reduced swing ALU clocking

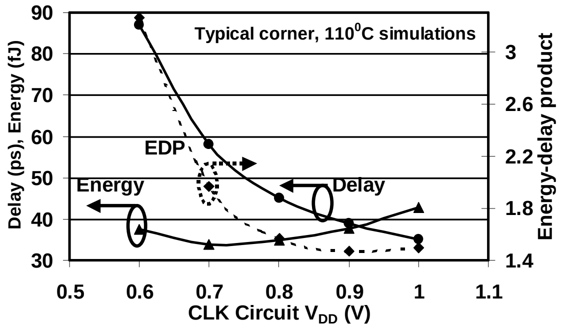

Figure 3(b): Dual supply latch 65nm energy-delay tradeoffs

For the 65nm technology (Berkeley PTM), there is a 21% reduction in total energy when $V_{DDL} = 0.7V_{DDH}$ while there is a 66% (90%) increase in the D→Q delay (T_{setup}). This corresponds to a D→Q delay increase of 23ps (T_{setup} increases 36ps). We restrict the usage of such a dual-supply latch (flip-flop) scheme to only the non-critical units and absorb the additional delay penalty in the timing slack. It should be noted that the region of minimum energy operation of the latch occurs when $V_{DDL}= (0.70\sim0.75)V_{DDH}$, while the region of minimum energy-delay product (EDP) operation is $V_{DDL}= (0.85\sim0.9)V_{DDH}$.

3.2 Swing Restored CPL-Based Logic Unit

In this paper, we use swing restored CPL logic to design the non-critical units of the ALU. Complementary pass transistor logic allows us to eliminate the p-MOS network required to implement a logic function when using the static CMOS style.

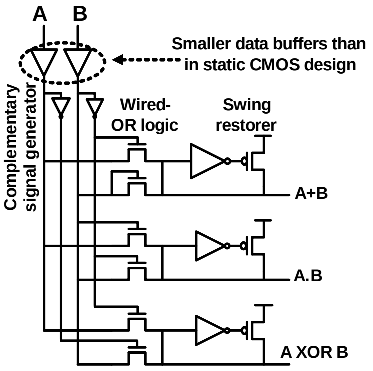
A
B
Smaller data buffers than in static CMOS design
Complementary signal generator
Wired-OR logic
Swing restorer
A+B
A.B
A XOR B

Figure 4: Logic unit bit-slice using swing restored CPL logic

This results in lower switching capacitance, smaller data buffer sizes and area for the logic unit. However, CPL logic using n-MOS pass transistors result in “weak” 1 and in our design we used output keepers to restore the CPL gate output signal to full swing. Figure 4 shows the usage of the CPL style to implement the logic unit gates for a single bit-slice of the 32-bit ALU.

3.3 PG Unit with Clock Footer Sharing

The 32-bit adder forms the performance critical core of the ALU and is implemented using compound domino logic (CDL) [10]. The PG unit of the adder outputs propagate (A+B) and generate (A.B) signals using dynamic gates with clocked footer transistors. In this design, we shared the 2 explicit clock transistors and use one common transistor as shown in Figure 5.

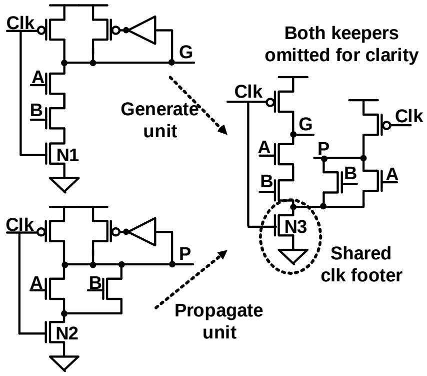

Figure 5: PG unit with shared clock footer for 32-bit ALU adder

It should be noted that by using this design strategy for the higher order adder bit-slices only, (shared footer used for bit slices 1 to 31, leaving bit slice 0 unchanged) it is possible to absorb the delay penalty in the existing slack. This allows us to obtain energy savings with minimal performance degradation for the worst-case delay vector.

Table 1: Clock footer sharing and domino energy-delay tradeoffs

N3/(N2+N1)	Ref =1	0.93	0.86	0.79	0.71
Delay (ps) (P/G)	10/20	11/21	12/22	13/23	14/24
Energy (fJ)	69	68	65	58	57

Table 1 indicates that, this sharing allows energy savings at the expense of performance. For example, when the effective n-MOS clock transistor width is reduced by ~29%, the P and G signal worst-case delays increase by ~4ps while allowing 16% energy savings for a data activity (α) of 0.1 (8% energy savings when α=1).

4. ALU Architecture and Design Overview

We now present an overview of the architecture of a 32-bit ALU and demonstrate the impact of the circuit techniques discussed earlier in ensuring low power operation. We also discuss the scaling trends and energy-delay tradeoffs associated with the circuit techniques for the 180nm-65nm CMOS technologies. The basic ALU architecture is shown in Figure 6, and is similar to that reported in [10]. This full-custom ALU design consists of approximately 11.5k transistors and has an operating frequency of 4.2GHz under worst-case conditions for 65nm CMOS (Berkeley PTM) technology. The decoder unit in Figure 6 determines the actual instruction (arithmetic, logical, shift) that is executed by the ALU. Both the decoder and logic/shift units are non-critical in terms of performance and have relaxed timings.

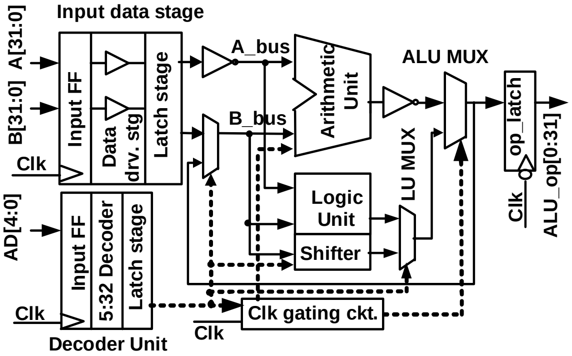

Figure 6: Conceptual block diagram of 32-bit ALU

Therefore, the decoder is realized using static CMOS logic, while the logic unit and shifter are implemented using swing-restored, complementary pass transistor logic (CPL). The MUX-es at the output of the logic unit are realized using C^2MOS logic (instead of transmission gates) to avoid the usage of cascaded pass transistors. The ALU critical path comprises of the arithmetic unit (adder front-end MUX + 32-bit adder), and the output MUX-es. In our design, these units were implemented using compound domino logic (CDL).

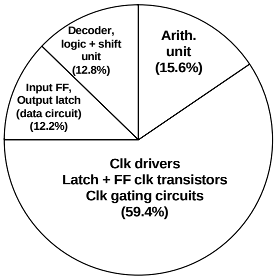

Figure 7: ALU (Design 1) energy break-up for180nm technology

Figure 7 shows the energy break-up for the ALU, averaged over 10 cycles of operation (includes logic, arithmetic and shift operations). The 180nm technology simulations indicate that the entire clock network contributes to 59.4% of the ALU total energy, while the arithmetic unit consumes 15.6% energy (worst case switching vector). In addition, the instruction decoder, logic unit and shifter contribute up to 12.8% of energy while the input stage flip-flops and ALU output latch data energy contribute 12.2%. It should be noted that the results in Figure 7 pertain to the baseline ALU design, with none of the design techniques discussed in section 3 incorporated in it. Henceforth we refer to this design as Design 1.

4.1 Energy-Delay Tradeoffs and Scaling Trends

Based on the energy break-up in Figure 7, we reduced the power supply for the latch and flip-flop units at the input-output boundaries of non-critical units like the decoder, and input data stage. The entire data network and the rest of the clock supply/drivers for the adder, output MUX-es and latch stages were maintained at a higher supply (V_{DDH}). Figure 8 demonstrates the ALU total energy savings associated with the different circuit techniques discussed in section 3.

We show the normalized energy plots for three different cases: Design 1 (Ref. ALU), Design 2 (all high V_{DDH}) and Design 2 (dual supply). Design 1 operates entirely at V_{DDH}, and uses static CMOS gates for its logic unit, and does not use clock sharing for the PG unit of its 32-bit adder. Design 2 (all high V_{DDH}) on the other hand demonstrates the energy savings obtained using CPL based logic unit and clock sharing in the PG unit while still operating at V_{DDH}. Our results indicate that, this results in 7%-10% savings in ALU total energy. However, when the power supply voltage is selectively lowered by ~30% (Design 2, dual supply case) there is 18%-24% savings in total energy for the 180nm-65nm technologies.

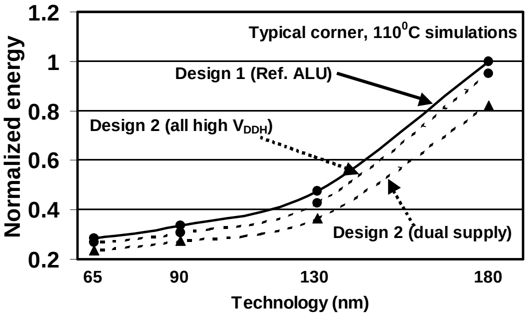

Figure 8: ALU energy savings and scaling trends

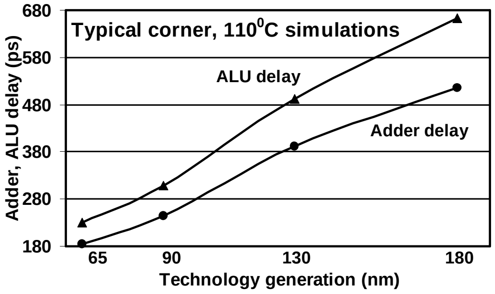

Figure 9: ALU and adder delay scaling trends

It should be noted that, for Design 2, the energy reduction was obtained while maintaining the worst-case ALU delay the same as in Design 1. This is possible because of the dual supply assignment strategy followed in this design, whereby all the critical unit signals were maintained at V_{DDH}. The scaling trends for the worst-case delays of both ALU and adder are shown in Figure 9.

4.2 ALU Leakage Power Demand

The different circuit techniques discussed in this paper, allow us to lower the ALU leakage power consumption as indicated in Figure 10. When Design 1 (Ref.) is scaled from 130nm to 65nm technology, there is a 27x increase in the standby mode leakage power (~30% gate leakage).

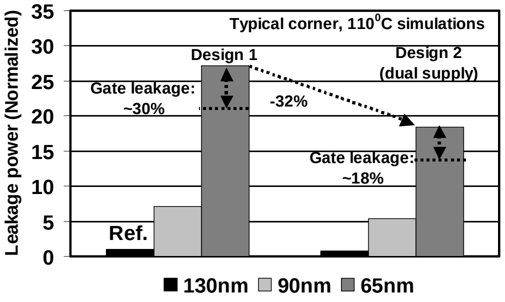

Figure 10: ALU standby mode leakage power plots

For Design 2 with dual supply, the total leakage power reduces by 22% (32%) for the 130nm (65nm) generation. The gate leakage reduces significantly (~40%) when the power supply is lowered for Design 2, and contributes to ~18% of the total ALU leakage power for the 65nm generation ($V_{DDL} = 0.72V_{DDH}$).

5. CONCLUSION

In this paper, we presented a high performance 32-bit ALU design and adopted a dual supply strategy to minimize total energy consumption. We discussed the impact of sharing the footer clock transistors (PG unit) and CPL logic in minimizing clock and data energy. We demonstrated the scaling trends for the 180nm-65nm CMOS technologies showing reductions in ALU total energy (18%-24%) and leakage power (22%-32%) demand.

6. ACKNOWLEDGMENTS

Authors would like to acknowledge O. Semenov, S. Naraghi and C. Kwong from the University of Waterloo, and S. Hsu and S. Borkar from Intel Corp. for encouragement and support.

7. REFERENCES

[1] Semiconductor Industry Associations, “International Technology Roadmap for Semiconductors, 2001Edition”, 2001.

[2] A. Chandrakasan, W.J. Bowhill, and F. Fox, *Design of High Performance Microprocessor Circuits*. IEEE Press, Piscataway, N.J., 2000.

[3] M. R. Stan, “Optimal Voltages and Sizing for Low Power,” *12th IEEE Intl. Conf. on VLSI Design*, pp. 428-433, 1999.

[4] N. Sirisantana, L. Wei, and K. Roy, “High-Performance Low-Power CMOS Circuits Using Multiple Channel Length and Multiple Oxide Thickness,” *Proc. ICCD*, pp. 227-232, 2000.

[5] T. Kuroda, “CMOS Design Challenges to Power Wall,” *Intl Conf. on Microprocessors and Nanotechnology*, pp. 6-7, 2001.

[6] T. Sakurai, H. Kawaguchi, and T. Kuroda, “Low-power CMOS design through V_{TH} control and low-swing circuits,” *ISLPED*, pp. 1-6, 1997.

[7] T. Kuroda, et. al, “A 0.9V, 150-MHz, 10-mW, 4mm^2, 2-D Discrete Cosine Transform Core Processor with Variable Threshold-Voltage (VT) Scheme,” *IEEE JSSC*, vol. 31, no. 11, pp. 1770-1779, Nov. 96.

[8] *http://www-device.eecs.berkeley.edu/~ptm*: BSIM3 files

[9] Y. Cao, T. Sato, D. Sylvester, M. Orshansky, and C. Hu, "New paradigm of predictive MOSFET and interconnect modeling for early circuit design," *Proc. of IEEE CICC*, pp. 201-204, Jun. 2000.

[10] S. Matthew, R. Krishnamurthy, M. Anders, R. Rios, K. Mistry, and K. Soumyanath, “Sub-500ps 64-b ALUs in 0.18μm SOI/Bulk CMOS: Design and Scaling Trends”, *IEEE JSSC*, vol. 36, no.-11, pp. 1636-1646, Nov. 2001.

[11] R. Krishnamurthy, S. Hsu, M. Anders, B. Bloechel, B. Chatterjee, M. Sachdev, and S. Borkar, “Dual supply voltage clocking for 5GHz 130nm integer execution core”, *Symp. on VLSI Circuits*, pp. 128-129, 2002.

[12] B. Chatterjee, M. Sachdev, S. Hsu, R. Krishnamurthy and S. Borkar, “Effectiveness and Scaling Trends of Leakage Control Techniques for Sub-130nm CMOS Technologies,” *Proc. of ISLPED*, pp. 122-127, 2003.