A Low-power 16-bit Multiplier-Accumulator using Series-regulated Mixed Swing Techniques

Ram K. Krishnamurthy, Herman Schmit, L. Richard Carley

Department of Electrical and Computer Engineering
Carnegie Mellon University, Pittsburgh, PA 15213

Abstract

*This paper describes an on-chip series-regulated mixed swing methodology with sleep-mode control for lowering the power consumption of high-performance DSP multiplier-accumulator (MAC) circuits. A 16*16+36-bit Overlapped bit-pair Booth recoded, Wallace tree MAC is fabricated in a commercial 0.5μm CMOS process in the proposed series-regulated methodology and conventional static CMOS. Up to 2.55X reduction in energy/operation is measured over static CMOS, while offering a simultaneous 1.8X improvement in low-voltage manufacturability. At the maximum clock frequency of 67MHz, the proposed approach consumes a total MAC power of 16.6mW in active mode and 152.5nW in standby mode. Measured peak-peak power/ground bounce is under 8% of the regulated low-swing voltage. Experimental results from comparisons in three additional (0.35μm, 0.25μm, 0.16μm) CMOS and fully-depleted SOI processes are also presented to demonstrate improved savings over static CMOS with process scaling.*

1.0 Introduction

With the increasingly stringent demands on battery space and weight in portable multimedia devices, there exists a strong necessity to investigate techniques for lowering the power dissipation of Digital Signal Processing (DSP) circuits. A majority of DSP circuits are signed, fixed-point, short bit-width (8 - 24 bit) datapath operators, specifically multipliers and/or multiplier-accumulators (MACs), and therefore, substantial attention has been devoted to lowering their power consumption.

Multiple voltage techniques have been reported earlier for lowering the power dissipation by operating non-critical path gates at reduced voltages [1], [2]. These techniques employ multiple voltages while retaining the static CMOS based logic gate structure unaltered. A four-power-rail methodology called *Mixed Swing QuadRail* has been proposed previously to construct standard digital logic gates using multiple voltages at the gate level [3], [4]. This approach performs logic by intermixing high- and low-swing signals while driving the load capacitance at the gate outputs at reduced swings. These multiple voltage techniques use explicit high- and low-voltage supplies to offer a nearly quadratic reduction in dynamic power since there exists no DC path between the supplies. However, they have three limitations: (i) the additional off-chip supply and its associated pin requirements add to the total system cost, making the techniques economically unattractive, (ii) low-voltage off-chip supplies are prone to significant inefficiencies, particularly if the drive-current requirements are high (e.g., if the supply delivers the drive-currents of many on-chip low-swing circuits). This degrades overall system power efficiency, (iii) due to the lack of any on-chip regulation (the separation between the high- and low-voltage supplies remains fixed), these techniques suffer from substantial low-voltage dispersion in delay and power across worst-case process and temperature corners, contributing significantly to parametric yield loss [5]. These are increasingly important concerns in future deep-submicron processes.

In this paper we describe a self-contained, on-chip series-regulated *Mixed Swing QuadRail* methodology with sleep-mode control. This technique locally generates the low-swing supply rails from the regular, high-swing supply rails, mitigating the above concerns. The next section describes details of the series-regulated Mixed Swing QuadRail technique. Section 3.0 describes the first reported implementation of a large-scale DSP datapath operator using the proposed series-regulated approach: a 16*16+36-bit DSP MAC fabricated in a commercial 0.5μm CMOS process. The same MAC is also fabricated in the off-chip regulated QuadRail approach (with explicit off-chip high- and low-voltage supplies) as well as conventional static CMOS to study the respective power-delay trade-offs. The remainder sections describe measured results, power-delay comparisons across three additional (0.35μm, 0.25μm, 0.16μm) CMOS and fully-depleted SOI (FDSOI) processes, and manufacturability analyses of the MACs.

2.0 Series-regulated Mixed Swing QuadRail

Fig. 1(a) shows the Mixed Swing QuadRail gate topology for a (3,2) counter, consisting of a logic stage operating between the high-swing power rails (Vd1-Vs1 = V_{logic}) and a driver/buffer stage operating between the low-swing power rails (Vd2-Vs2 = V_{buffer}). V_{logic} and V_{buffer} are approximately centered to maximize high and low noise margins and to equalize rising and falling delays in either stage. PMOS devices in each stage have independent N-wells for minimal body-effect on the buffer stage PMOS devices. Since our target process is single-threshold and N-well based, NMOS devices reside in the native P-substrate. Fig. 1(b) shows the proposed series-regulator circuit for local generation of the low-swing power rails (Vd2 and Vs2) from the regular, off-chip high-swing power rails (Vd1 and Vs1). The low-swing voltage is servoed to maintain a fixed ratio of off- to average on-drive current (I_{off}/I_{on}) in the QuadRail circuit in order to balance static and dynamic power. This achieves the same goal of minimizing total power as [6] but without mandating any process modifications. The transistor pairs (M3:M4) and (M7:M8) are ratioed Nx:1x, where 1x is the minimum-width transistor and N is the target I_{on}/I_{off}. The PMOS devices are appropriately ratioed wider than the NMOS devices to equalize their respective drive capabilities. The current mirror devices (M1:M2)

This work was supported by DARPA (Order A564), NSF (Grant MIP9408457), and SRC (Contract 068.007).

Reprinted from *Proceedings of the 1998 IEEE Custom Integrated Circuits Conference,* Santa Clara, CA, May 1998

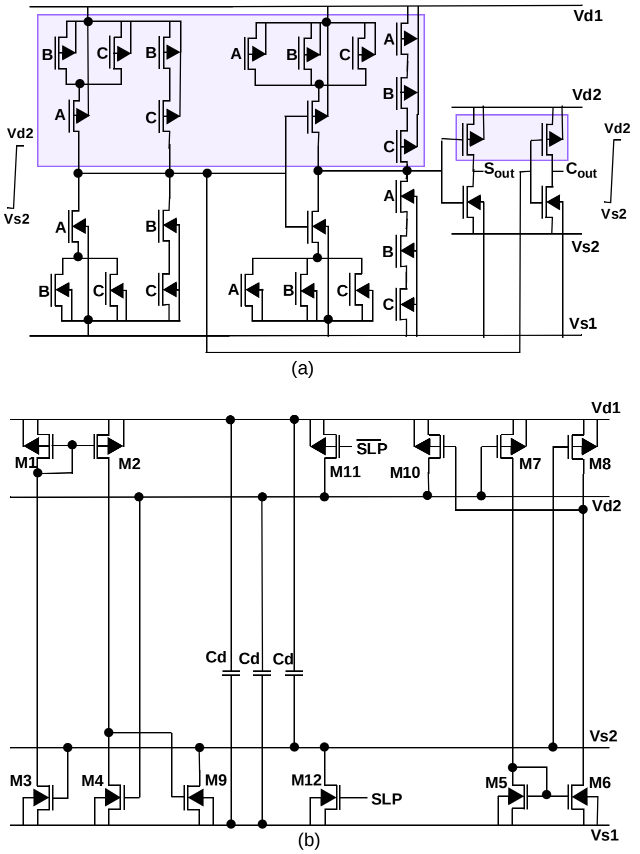

Figure 1: (a) Mixed Swing QuadRail (3,2) counter (b) I_{off}/I_{on} ratio based series regulator circuit.

and (M5:M6) are ratioed 1:1. M9 and M10 provide the DC series path between the power rails and are sized to be able to source/sink the QuadRail circuit's peak on-drive current requirement. Three local inter-rail decoupling capacitors (C_d) are inserted to reduce rippling on the low-swing power rails due to simultaneous switching noise on the high- and low-swing power rails. M11 and M12 are sleep-mode enable devices that are disabled (SLP=Vs1) during normal operation. During power-down mode (SLP=Vd1), the low-swing rails are shorted to the high-swing rails, eliminating the DC path power consumption that exists during active mode.

3.0 16*16+36-bit MAC Power-Delay Comparisons

A 2's complement, fixed-point 16*16+36-bit MAC is fabricated in a commercial 0.5μm CMOS process. The MAC comprises of an Overlapped bit-pair Booth-recoded, (3,2) counter-based Wallace tree 16*16-bit multiplier [7] and a 36-bit Block Carry Lookahead final accumulator [8], with a single pipeline stage between the multiplier and accumulator for enhanced throughput (Fig. 2(a)). The power distribution measured on a static CMOS implementation of the MAC is shown in Fig. 2(b). The Wallace tree multiplier is the most power-critical MAC component, consuming 75% of total power. This is due to the substantial interconnect capacitances driven by the 28-transistor-based (3,2) counters [9] within the Wallace tree. In order to lower the multiplier power, three versions of the MAC are fabricated with the multiplier constructed in series-regulated QuadRail, off-chip regulated QuadRail, and conventional static CMOS to study the relative power-delay tradeoffs. The final accumulator, due to its higher logic depth than the multiplier, is the most time-critical MAC component and hence sets the maximum clock frequency. It is therefore implemented in full-swing static CMOS in all MAC versions to retain a fixed, high throughput. All three MACs have CMOS-level I/Os to enable interfacing with external CMOS circuitry without level conversion.

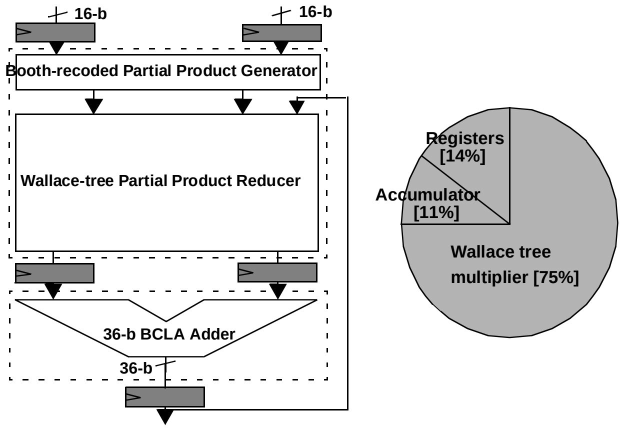

Figure 2: 16*16+36-bit MAC (a) architecture and (b) measured power distribution on a 0.5μm static CMOS implementation.

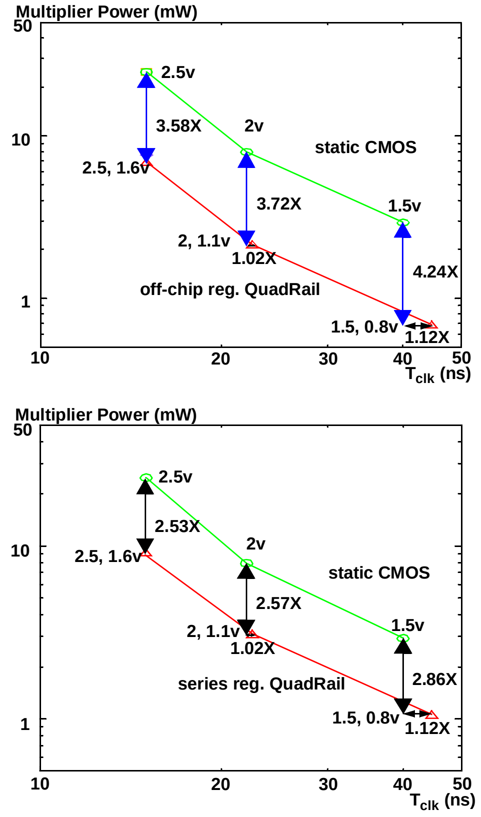

Figure 3: Static CMOS vs. QuadRail power-delay comparison measurements.

Fig. 3(a)-(b) show the measured Wallace tree multiplier power-delay comparisons for static CMOS vs. the QuadRail

Reprinted from *Proceedings of the 1998 IEEE Custom Integrated Circuits Conference*, Santa Clara, CA, May 1998

methodologies over a range of operating voltages (2.5-1.5V), i.e., V_{dd} for CMOS and V_{logic} for QuadRail. QuadRail's corresponding buffer voltages are selected to maintain an I_{off}/I_{on} ratio of 1:150, which balances static and dynamic power within the QuadRail multiplier while meeting the target delay constraints set by the CMOS MAC. Fig. 4 shows the low-swing rail waveforms from the series-regulated QuadRail MAC at Vd1=2V, Vs1=0V. Measured peak-to-peak power/ground bounce on the low-swing power rails is confined to within 8% of the low-swing voltage with 4pF on-chip inter-rail decoupling capacitors.

Power and delay are measured across 500 pseudo-random input vectors. The off-chip regulated QuadRail approach shows energy/operation savings ranging up to 3.79X over static CMOS, with the savings increasing with voltage scaling. The savings are attributed to the following:

- Average point-to-point net capacitance (due to both interconnect and fanout gate loading) extracted from the Wallace tree multiplier layout is 58fF. This, coupled with the inherently high switching activities of Wallace trees makes the effective switched capacitance per cycle substantial. A full quadratic reduction in buffer stage dynamic power is achieved due to the lowered output swing across this capacitance.

- 28% of the dynamic power within the multiplier is due to short-circuit power dissipation, despite the multipliers being optimally sized to maintain steep input rise/fall times. This is comparable to the short-circuit power reported for a similar multiplier in [10]. Thus, the reduced buffer stage swing offers a nearly cubic reduction in its short-circuit power component as well, contributing to the additional energy/operation savings.

Series-regulated QuadRail offers relatively lower energy/operation savings than off-chip regulated QuadRail, due to the DC series path between the power supplies. Therefore, the buffer stage dynamic power reduction factor drops from quadratic to linear. However, the nearly cubic reduction in buffer stage short-circuit power is still retained, contributing to an energy/operation savings slightly larger than linear. The savings range up to 2.55X, i.e., up to a 35% loss in savings compared to off-chip regulated QuadRail. At 67MHz/23MHz (maximum/minimum measured clock speed), the total series-regulated QuadRail MAC power (i.e., multiplier, accumulator, and registers) is 16.6mW/2.06mW. Series-regulated QuadRail's DC power disadvantage is offset by the following advantages:

- Standby power (152.5nW) is nearly three orders of magnitude lower than off-chip regulated QuadRail's standby power (143.8µW), because of the absence of the Vd1-Vs1 totempole current path during sleep mode. Further, transition between sleep and active modes is accomplished in a single clock cycle. Since transitioning to sleep mode essentially transforms QuadRail into conventional static CMOS, circuit state is still retained during standby. Thus, transitioning between sleep and active modes eliminates the need for any explicit state data transferring schemes similar to [11].

- Since the additional low-voltage supply is not required, series-regulated QuadRail is a self-contained methodology that can replace static CMOS operating from a regular, high-swing supply without mandating any system-level modifications.

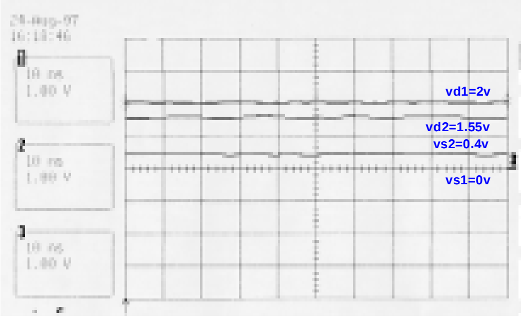

Figure 4: 0.5µm Series-regulated QuadRail MAC measured power-rail waveforms.

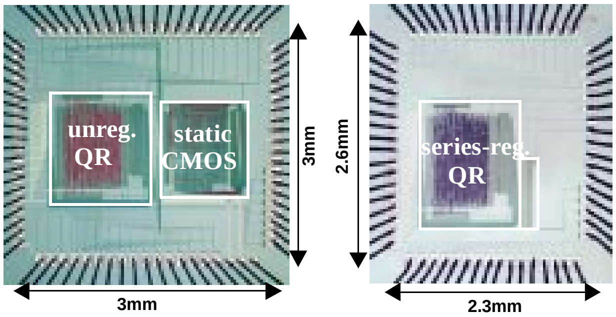

Figure 5: Static CMOS, QuadRail MAC die microphotographs.

Fig. 5 shows the static CMOS and QuadRail MAC die microphotographs. The off-chip regulated QuadRail MAC occupies about 10% larger layout area due to intrinsic cell-layout area penalty incurred by its dual-well requirement. Series-regulated QuadRail MAC incurs an additional 8% area penalty due to the on-chip decoupling capacitors.

The power-delay comparisons are extended over three additional commercial single-threshold processes: 0.35µm CMOS, 0.25µm FDSOI, and 0.16µm CMOS, to study the impact of process scaling on energy/operation savings (Fig. 6). Series-regulated QuadRail energy/operation savings increase with process scaling: up to 3.2X in 0.35µm, 3.45X in 0.25µm, and 3.8X in 0.16µm processes. The 0.25µm implementation's lowest energy/operation (at V_{logic} = 0.75V, V_{buffer} = 0.35V) is 6pJ. This is nearly 3.3X lower than one of the lowest reported energy/operation implementations in literature in a comparable multi-threshold 0.25µm process [10]. Since interconnect capacitance scales slower than gate capacitance with process scaling, the Wallace tree multiplier, because of its interconnect-dominated point-to-point net capacitances, becomes more and more power-critical. This, coupled with the increasing ratios of logic to buffer swings with process scaling, makes driving the multiplier's load capacitances at

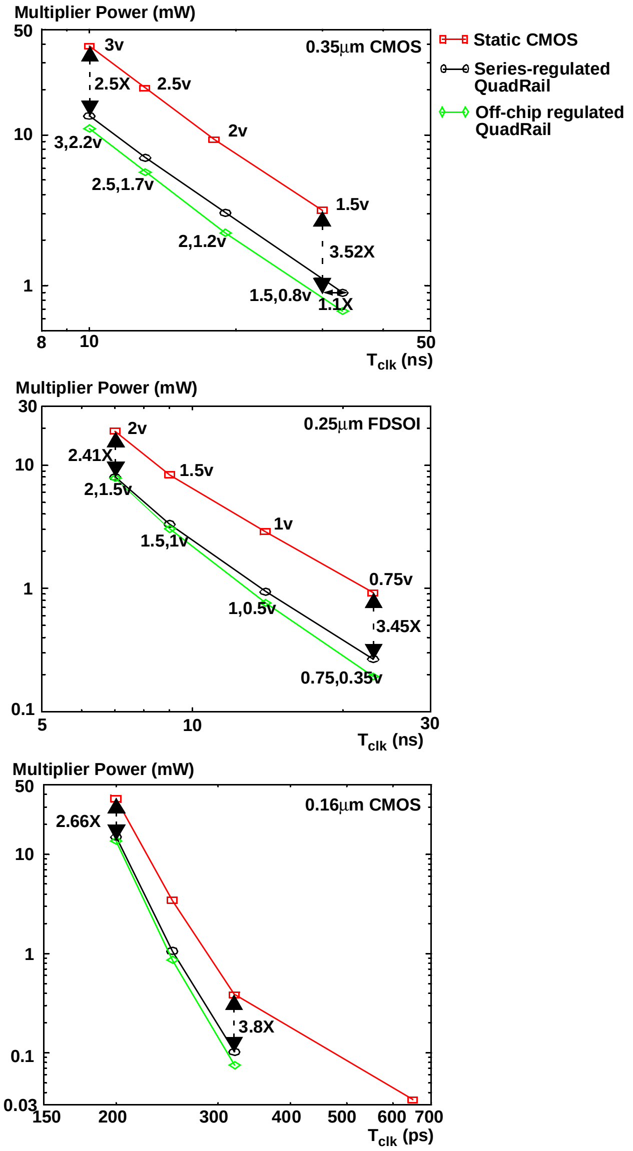

Figure 6: Static CMOS vs. QuadRail power-delay comparisons in 0.35µm CMOS, 0.25µm FDSOI, and 0.16µm CMOS process.

lower swings to offer improved energy/operation savings. The savings increase even further with process scaling beyond our range of analysis.

4.0 16*16+36-bit MAC Manufacturability

To study the impact of series-regulated QuadRail on manufacturability, worst-case process and temperature corner analysis is performed across industrial Slow-NMOS-Slow-PMOS and Fast-NMOS-Fast-PMOS corners on the CMOS and QuadRail multipliers in the 0.5µm process (Fig. 7). QuadRail demonstrates similar power*delay dispersions as CMOS at high voltages. With voltage scaling, the dispersion remains well controlled and at V_{logic}=1.5V, V_{buffer}=0.8V, the power*delay dispersion is 1.8X lower than CMOS, demonstrating improved low-voltage parametric yield. This is attributed to (i) the low-swing rails being dynamically offset across corners to maintain the target I_{off}/I_{on} ratio, thereby significantly compensating for the manufacturing variations, and (ii) the reduced output swings of QuadRail gates causing the power and delay sensitivities to worst-case corners to be relatively lower than in static CMOS. Further electronic variations control for both QuadRail and CMOS may be achieved through substrate/well back-biasing schemes [12].

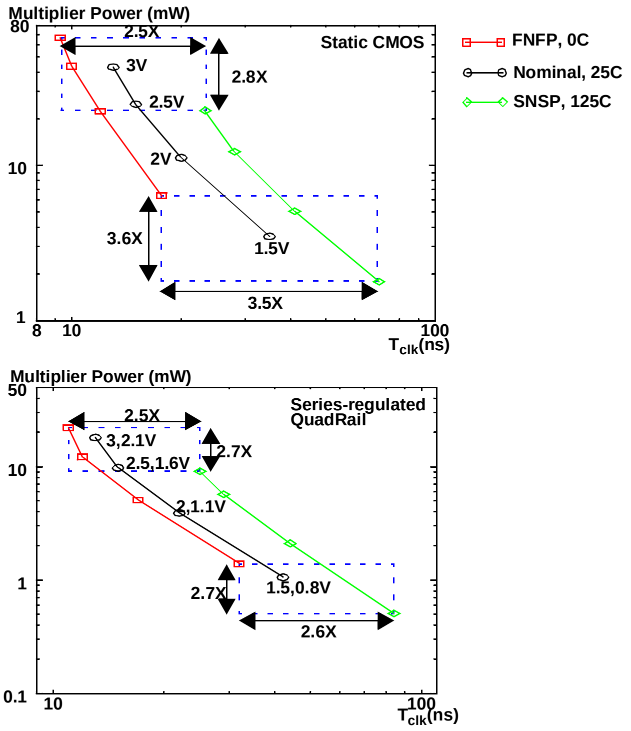

Figure 7: Static CMOS vs. series-regulated QuadRail power*delay dispersion analysis in 0.5µm process.

5.0 Conclusions

We have proposed a self-contained, on-chip series-regulated Mixed Swing QuadRail methodology with sleep-mode control. We have also described measured results from a 16*16+36-bit MAC fabricated in this methodology and static CMOS in a 0.5µm process. Up to 2.55X energy/operation savings was measured over static CMOS, while offering a simultaneous 1.8X low-voltage manufacturability improvement, without requiring any process or system-level modifications. Experimental results from three additional processes were also presented to show increased savings over static CMOS with process scaling.

References

[1] K.Usami, M.Horowitz, *Proc. ISLPD*, April 1995, pp. 3-8.
[2] K.Usami et al, *Proc. CICC*, May 1997, pp. 131-134.
[3] R.K.Krishnamurthy, I.Lys, L.R.Carley, *Proc. ISLPED*, August 1996, pp.381-386.
[4] R.K.Krishnamurthy, L.R.Carley, *IEEE Trans. VLSI Systems*, December 1997.
[5] A.J.Strojwas et al, *Proc. ISLPED*, August 1996, pp. 225-232.
[6] J.B.Burr, A.M.Peterson, Proc. *ICCD*, October 1991, pp. 593-600.
[7] J.F.Ardekani, *IEEE Trans. VLSI Systems*, June 1993, pp.120-125.
[8] J.F.Cavanagh, *Digital Computer Arithmetic*, McGraw Hill, 1984.
[9] R.K.Montoye et al, *Proc. ISSCC*, February 1990, pp. 336-337.
[10] M.Izumikawa et al, *IEEE JSSC*, January 1997, pp.52-60.
[11] S.Shigematsu et al, *IEEE JSSC*, June 1997, pp.861-869.
[12] T.Kobayashi, T.Sakurai, *Proc. CICC*, May 1994, pp.271-274.