Energy-Efficient Low-Voltage Operation of Digital CMOS Circuits Through Charge-Recycling

Saravanan Rajapandian, Zheng Xu, and K. L. Shepard
Columbia Integrated Systems Lab, Department of Electrical Engineering
Columbia University, New York, NY 10027
{sara,zx2004,shepard}@cisl.columbia.edu

Abstract

This paper describes an energy-efficient means to achieve on-chip dc-dc conversion for digital CMOS circuits. The approach uses balanced voltage islands running at fractions of the off-chip supply voltage. Charge "discarded" by one domain is "recycled" to supply energy for another. When the domains are ideally balanced, all the energy dissipated by electrons in "dropping" to lower potentials is used for active computation. We describe the design and measurement of a prototype system in a 0.18 μm CMOS process that provides active on-chip voltage regulation and controlled dc-dc conversion with this technique.

Introduction

Dynamic (or adaptive) voltage scaling (DVS) offers the ability to trade-off power and performance through adjustment of the supply voltage, V_{DD} [1]–[3]. Significant energy savings can be realized by lowering the supply voltage until circuits can just meet specified performance requirements. Most DVS systems are based on the idea that multiple power grids are available to be "tapped into" to support multiple voltage operation, which comes at the cost of additional complexity and area [4], [5]. An alternative to a set of externally-generated fixed voltage supplies that are switched into on-chip voltage domains is to provide for dynamic dc-dc conversion for each island, which would allow for continuous scaling and negate the need for multiple global power grids. To accomplish this, what is needed is a very efficient adaptive power supply regulator, preferably one that is small and can be completely integrated on-chip.

Efficient on-chip dc-dc downconversion is also becoming a critical component in the design of deeply scaled digital CMOS ICs. Operating at supply voltages below 1 V, 90-nm (and below) technologes still demand in excess of 100 W of power in the largest chips. Delivering this power at the reduced supply voltage levels required by scaling results in high current requirements, exacerbating power supply integrity issues (i. e., forcing very low impedance requirements on the power distribution). Being able to bring the power onto the chip at higher voltage levels, which are then downconverted to the required supply voltage, significantly reduces the off-chip current requirements. The active regulation of the power supply that such an approach implies also reduces the on-chip decoupling capacitance requirement.

The most efficient dc-dc converters are buck-type regulators, which generate a reduced dc level by filtering a pulse-width modulated (PWM) signal through a simple LC filter. [6] By varying the frequency or duty-cycle of the PWM signal, different dc levels can be generated. While buck converters can operate at very high efficiencies (> 80%), they require off-chip filter components, which limits their usefulness. To deal with this limitation, two other types of dc-dc converters are possible which can be easily integrated on-chip: linear regulators and switched-capacitor power supplies.

This work was supported by the National Science Foundation under grant CCR-00-86007, by the MARCO C2S2 Center, by the SRC, and by a gift from the Intel Corporation.

A linear regulator is a power transistor (shown as a variable resistor in Fig. 1(a)) that is controlled by a feedback amplifier to keep the intermediate supply voltage V_{int} constant with changing load current demands. The efficiency of such a linear regulator is limited to V_{int}/V_{DD}, which provides for low efficiencies at small values of V_{int}.

Switched-capacitor (SC) supplies allow one to produce lower voltages at higher efficiencies than linear regulators. SC supplies are capacitance dividers, in which the capacitors are periodically "exchanged" as they are discharged by the load current. Overall efficiency in generating, for example, a $V_{DD}/2$ supply voltage is still quite poor even in carefully design systems [7] (about 60 - 65 %). Furthermore, both linear regulators and switched-capacitor supplies consume huge on-chip areas for both the power transistors (of the linear regulator) and the capacitors (of the switched-capacitor supply).

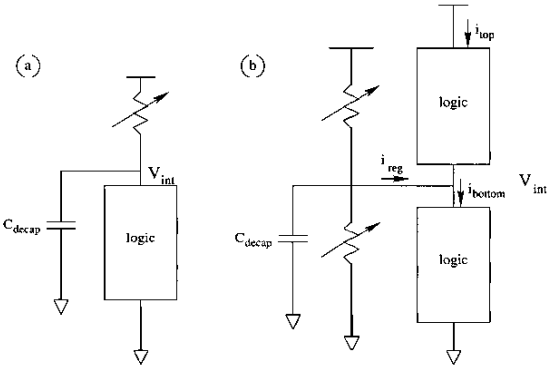

Fig. 1. (a) Linearly regulated reduced voltage requires a large power transistor; (b) this power transistor can be replaced with a logic and only a small push-pull regulator is needed to regulate the node V_{int}.

In this paper, we explore an alternative, *charge-recycling low-voltage regulation*, that allows for very energy-efficient on-chip generation of reduced supply voltages [8]. The idea borrows, in part, from charge-recycling ideas as applied in buses [9] or the highly capacitive lines of memories [10]. In this case, as shown in Fig. 1(b), we replace the power transistor of the linear regulator by *another* domain of logic. Now, the energy lost by electrons in dropping from potential V_{DD} to potential V_{int} is usefully employed in computation in the "top" logic domain. This charge is then "recycled" to be used in the logic computation of the "bottom" domain. If $V_{int} = V_{DD}/2$, then this top logic domain is operating between V_{DD} and $V_{DD}/2$ while the bottom domain is operating between $V_{DD}/2$ and ground. To achieve high efficiency, the charge demands of the top and bottom domain must be "balanced," so that the charge required by the bottom domain can be completely provided by the top domain. The system must also provide for the active regulation of V_{int}. We note that this approach is general in that voltages other than $V_{DD}/2$ may be regulated; for example, one domain may operate at $V_{DD}/3$,

 0-7803-8287-0/04/ $20.00 © 2004 IEEE

while another operates at $2V_{DD}/3$. Similarly, the scheme can be generalized to more than two domains and to regulate down from voltages above V_{DD}; for example, one could bring in an off-chip $3V_{DD}$ supply and recycle charge through three domains, one operating between $3V_{DD}$ and $2V_{DD}$, the second operating between $2V_{DD}$ and V_{DD} and the third operating between V_{DD} and ground.

In the next section, we explore general implementation issues for charge-recycling low-voltage regulation. A specific model system prototype is described in the following section. Measurement results on this system are presented next. The last section concludes.

Implementation issues for charge-recyling voltage domains

In using charge recycling for dc-dc conversion, charge imbalances will inevitably come about because of differences in the evaluation node capacitances of the two domains or because of differences in circuit activity in the two domains. Therefore, a full system must provide for the active regulation of the V_{int} node, which can be accomplished with three "levels" of regulation. To smooth out imbalances on short time scales, there must be adequate decoupling capacitance provided on the V_{int} node. For "medium" time constants, a linear regulator is used to add or subtract charge from the V_{int} supply. To compensate for large imbalances or imbalances that exist for extended periods of time, logic can be moved from one domain to the other.

To be able to add or subtract charge from the V_{int} supply node, a push-pull configuration for the linear regulator, as shown schematically in Fig. 1(b), is most effective. The power transistors (shown in Fig. 1(b) as two variable resistors) can be far smaller than the power transistor required for the linear regulator of Fig. 1(a) because they must only provide the supply-current *difference* between the two domains. Similarly, the decoupling capacitance requirements on the V_{int} node in Fig. 1(b) are significantly lower than the requirements in Fig. 1(a).[1]

The more current that must be sourced or sinked by the regulator, the lower the efficiency of the dc-dc conversion. The "instantaneous" efficiency is given by:

$$\eta(t_o) = \frac{V_{DD} i_{top}(t) + V_{int} i_{reg}(t)}{V_{DD} i_{V_{DD}}(t)} \quad (1)$$

where i_{top} and i_{reg} are the currents as shown in Fig. 1(b). $i_{V_{DD}}$ is the total current flowing out of the V_{DD} supply.

To provide a mechanism to rebalance domains on long time scales, each domain is divided into a set of switchable units, referred to as *granules*. A domain may consist of tens or hundreds of granules depending on the domain size. Granules can then be exchanged between domains to compensate for charge-consumption mismatch. To accomplish this, each granule shares a set of *granule multiplexer transistors* (as shown in Fig. 2) in both the pull-up and pull-down networks which determine the domain assignment of a particular granule; these switches also allow a domain to be configured for full-rail operation. The switches are similar to the "sleep transistors" which can be employed to control standby power due to subthreshold leakage and can also be used to perform this function if implemented with "high" V_T devices [11]. The drain nodes of these multiplexer transistor represent virtual supply and ground nodes. With the help of decoupling capacitance on these virtual nodes, the switches should be sized large enough to keep V_{DS} less than 5 % of the target supply voltage for the domain. At the system level, a given logic block can be easily configured to run at full-supply or, for example, half-supply. In the latter case, the block would be partitioned into two domains with one domain supplying charge for the other.

[1]The decoupling capacitance requirements on the chip supply are also reduced because of the reduced current demands.

Charge recycling dc-dc conversion does not work well in "normal" bulk CMOS because of body effect. Because their bodies are still tied to ground, nFETs in the upper voltage domain are heavily body affected. This problem is avoided in a triple-well process in which the nFETs are constructed in a p-well within an n-well. These p-wells are then tied to the virtual ground of the granule; similarly, the pFET n-wells are tied to the virtual supply of the granule. The junction capacitance of these wells adds intrinsic decoupling to the virtual supply and ground nodes, improving power supply integrity for a given switch width. Silicon-on-insulator (SOI) technology is also an attractive alternative for this technique, since the bodies float to the required voltage by action of the gate, source, and drain.

Special consideration must go into the logic that controls the switching of granules between domains to guarantee system stability and ensure (because of the power overhead associated with switching granules) that switching occurs only when the linear regulator is providing too much current for an extended period of time. Switching granules between domains dissipates energy because of the power required to switch the capacitance of the gates of the (large) granule multiplexer transistors. Furthermore, decoupling capacitance on the virtual supply and ground nodes (which for the triple-well implementation is provided by the intrinsic well capacitance) must be charged or discharged when a granule switches between domains. Fortunately, however, most of the device capacitances and interconnect coupling capacitances between wires of the same domain have the character of floating capacitors, simply translating in voltage as domains are switched. It is also possible for granules to switch domains while the digital logic is functioning without stalling or stopping execution.

Model system

To test the efficacy of charge-recycle low-voltage regulation, we have designed and fabricated a simple prototype system consisting of a 16 × 16 carry-save-array multiplier in a TSMC $0.18\mu m$ triple-well process. The die photo is shown in Fig. 3. Two on-chip SRAMs feed data into the multiplier and a third stores the result. The multiplier can be dynamically configured to run at 400 MHz at full supply or 200 MHz at half supply. Nominal full supply is 1.8 V, giving a fanout-of-four (FO4) delay of approximately 60 psec. At a supply of 900 mV, the FO4 delay is approximately 120 psec. The multiplier has 20 pipeline stages and is decomposed into 16 granules, each granule representing a total gate capacitance of approximately 8 pF. Because of the well capacitance, each granule has a decoupling capacitance of approximately 0.5 pF on the virtual supply and ground nodes. The nFET (pFET) multiplexer transistors have a total gate width of approximately 400 μm (800 μm) per granule. The complete system is as shown in Fig. 2.

The linear regulator design, consisting of two single stage differential amplifiers and a push-pull output stage (transistors M1 and M2), is shown in detail in Fig. 4. A simple switched-capacitor divider is used to generate the $V_{DD}/2$ reference (**half_vdd_ref**) for the linear regulator. The regulator has an open-loop gain of 38 dB and a unity gain bandwidth of 130 MHz with a phase margin of 70 degrees. The amplifier driving transistor M1 is biased with $200\mu A$, while the amplifier driving M2 is biased with $400\mu A$. The output stage has a quiescent current of $50\mu A$. (This quiescent current has an unfortunately strong dependence on mismatch, an issue that will be address in a future implementation.) The regulator can source or sink 30 mA of current before losing regulation.

Power transistors M1 and M2 have widths of 600 μm and 1.2 mm, respectively. Transistors M3 through M5 mirror out a current proportional to that flowing through transistor M1 for integration onto the capacitor C_{int}, which is approximately 400 fF. Similarly, transistors M6 through M8 mirror out a current proportional to that flowing through transistor M2, also integrated

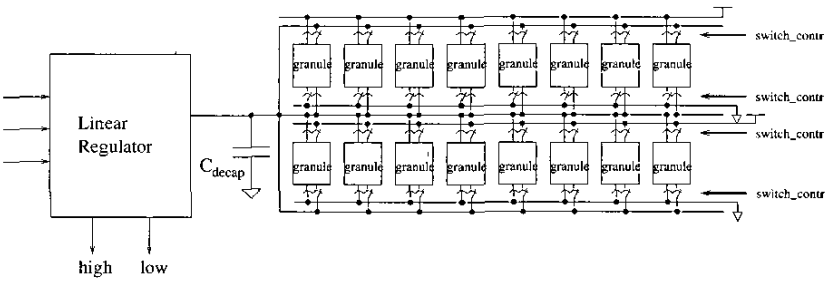

Fig. 2. Charge-recycling prototype system. A linear regulator controls the V_{int} node. When the linear regulator is sourcing or sinking a large amount of current, the **high** or **low** signals trigger the granule-switching control logic to switch granules between domains to remove the imbalance.

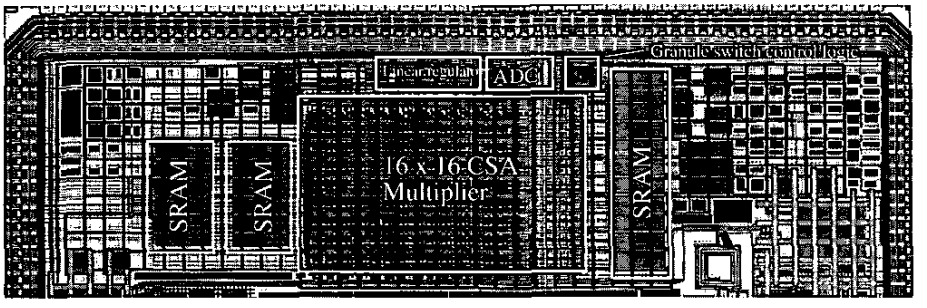

Fig. 3. Die photo of the prototype system.

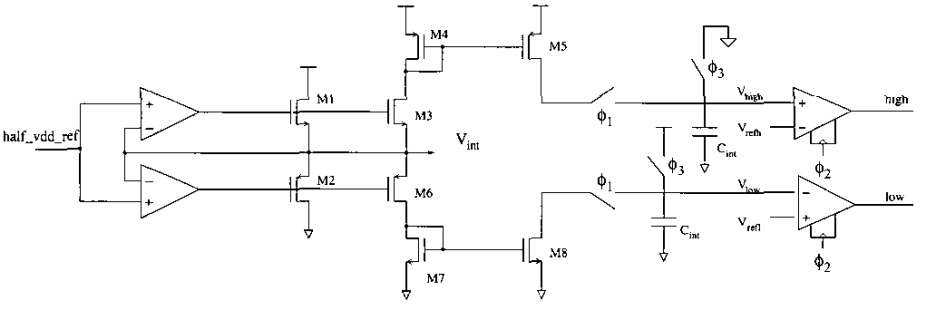

Fig. 4. Linear regulator design with domain charge mismatch monitoring.

onto a capacitor. Clock phases ϕ_1 and ϕ_3 are used to control the integration, establishing an integration time of approximately $t_{int} = 150nsec$. After the integration window, ϕ_3 clocks the comparators to compare the voltages V_{high} and V_{low} with the reference levels V_{refh} and V_{refl}, respectively, producing the signals **high** and **low** to the granule-switching control logic.

The use of source-follower transistors M1 and M2 provides better stability and the need for less decoupling capacitance than the more traditional common-source output stage. In many linear regulator applications, the common-source is preferred because of its lower dropout voltage. In this application, dropout voltage is not a concern since we are regulating far from the rails. Decoupling capacitance on V_{int} must ensure a low enough impedance beyond 100 MHz, where the linear regulator is ineffective.

$$C_{decap} \cong \frac{\Delta i_{reg}}{\Delta V_{int} 2\pi f} \quad (2)$$

For $\Delta i_{reg} \cong 5mA$ and $\Delta V_{int} \cong 90mV$, this yields C_{decap} of approximately 9 pF. 4pF of this is provided by explicit on-chip thin-oxide decoupling capacitance, while the remainder is provided by non-switching circuits and well capacitance.

When the average current sourced (sinked) by the linear regulator exceeds 3 mA over an interval of approximately 150 nsec, the **high (low)** signal is asserted to indicate that granules should be switched between domains to remove the imbalance. To switch logic between domains, the controller randomly chooses a granule by means of a linear feedback shift register (LFSR). At most one granule can be "switched" every t_{int}.

Granules for this implementation are chosen to be latch-bounded logic partitions, allowing the latches between granules to provide full-rail interfaces. Gate-isolated sense-amplifier-based flip-flops are used. Unfortunately, the switching energy of these full-fail interface latches does not scale when the logic is configured to run at reduced supply, reducing the overall achievable power scaling.

Measurement results

Table I summarizes the measured power characteristics in a balanced configuration (i. e., when the system is in a steady-state with relatively matched charge consumption of the top and

bottom domains). The multiplier functions correctly running at both 1.8-V and at a charge-recycled 0.9-V supply; functionality is not affected by dynamics of domain switching. A switching energy of approximately 108 pJ is associated with the full-rail boundary latches and associated full-rail components of the clock distribution that do not scale with supply voltage. This latch "overhead" can be expected to be inversely proportional to the granule size; it is relatively high in this model system because of the relatively small granule size chosen here (8 pF of total gate capacitance). This latch switching energy can also be reduced by approximately 30% by scaling the supply of the RS-latch in the boundary sense-amplifier flip-flop. We consider the efficiency of our system independent of this overhead because these components are not regulated at the $V_{DD}/2$ supply. The overhead should be reduced, if not eliminated, by the introduction of a true level-shifting approach in future implementations.

In Fig. 5, we show how the measured efficiency of the system degrades with mismatch between the top and bottom domains. In this case, the domain-switching is disabled and granule assignments are "hard-wired." The pairs of numbers in Fig. 5 denote the number of granules (out of a total of 16) assigned to the "top" and "bottom" domains. "8-8" is an approximately matched configuration, while "2-14" is heavily skewed. We note that the currents demand of each granule are not the same, although the assignments below start from a balanced configuration and achieve the skewed configurations by moving granules from one domain to the other. Efficiency exceeds 85% for the matched configuration and approaches 50% when the domains are out of balance (and the linear regulator must source or sink all of the current for the logic). Measurements are on-going to also measure the instantaneous efficiency during domain switching.

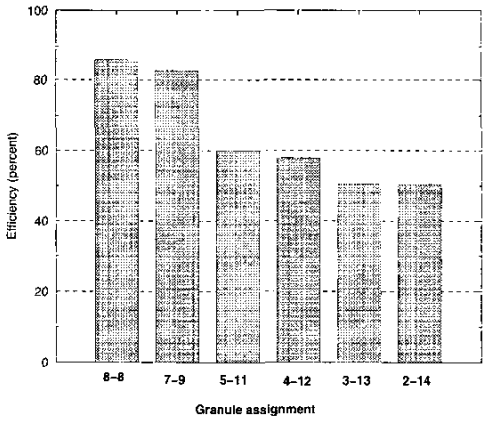

Fig. 5. Efficiency of the charge-recycling dc-dc conversion as a function of charge mismatch. The number of granules assigned to each domain is noted by the pair of numbers.

The chip also contains a 40-MHz, 6-bit flash analog-to-digital converter to monitor the regulation of the $V_{int} = V_{DD}/2$. Fig. 6 shows the output of this ADC, showing 10% regulation, for the multiplier with a random input pattern.

Table I also considers the area overhead associated with implementing this charge-recycling scheme. In our model system, the regulator and switch overhead add only approximately 3-4% to the area of the system.

Conclusions and acknowledgments

In this paper, we have described a new approach for on-chip dc-dc conversion using charge-balanced voltage islands running at fractions of the supply voltage. Charge "discarded" by one domain is "recycled" to supply energy to another. In a simple prototype system, we achieve greater than 85 % measured efficiency for $V_{DD}/2$ conversion.

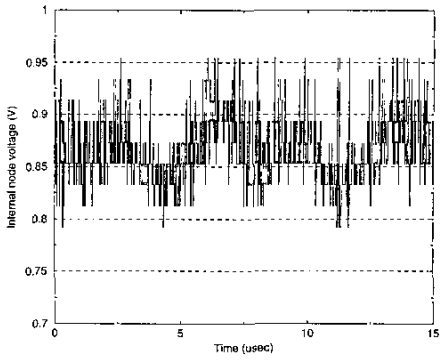

Fig. 6. Regulation of the V_{int} node as measured by the 6-bit on-chip flash ADC

TABLE I
SUMMARY OF MEASURED SYSTEM CHARACTERISTICS

Scaling power @ V_{DD} and f=400 MHz	158 mW
Scaling power @ $V_{DD}/2$ and f=200 MHz	20 mW
Full-rail latch and clock power at f=400 MHz	108 mW
Full-rail latch and clock power at f=200 MHz	54 mW
Multiplier area	1.4 mm^2
Linear regulator area	0.024 mm^2
Switch control logic area	0.0196 mm^2

The authors gratefully acknowledge T. Karnik and P. Hazucha of Intel for many helpful discussions.

References

[1] G. Qu and et al. Energy minimization of system pipelines using multiple voltages. In *Proc. IEEE International Symposium on Circuits and Systems*, 1999.

[2] G. Semeraro, G. Magklis, R. Balasubramonian, D. H. Albonesi, S. Dwaradasas, and M. L. Scott. Energy-efficient processor design using multiple clock domains with dynamic voltage and frequency scaling. In *Proc. Intl. Symp. on High Performance Computer Architecture*, 2002.

[3] T. D. Burd and R. W. Broderson. Design issues for dynamic voltage scaling. In *Intl. Symp. on Low Power Electronics and Design*, 2000.

[4] K. Usami and M. Horowitz. Clustered voltage scaling technique for low-power design. In *Proc. Workshop on Low Power Design*, 1995.

[5] D. E. Lackey, P. S. Zuchowski, T. R. Bednar, D. W. Stout, S. W. Gould, and J. M. Cohn. Managing power and performance for system-on-chip designs using voltage islands. In *Proceedings of the IEEE/ACM International Conference on Computer Aided-Design*, pages 195 – 202, 2002.

[6] G.-Y. Wei and M. Horowitz. A fully digital, energy-efficient adaptive power-supply regulator. *IEEE Journal Solid-State Circuits*, pages 520 – 528, April 2000.

[7] G. Patounakis, Y. W. Li, and K. L. Shepard. A fully integrated on-chip DC-DC conversion and power management system. *IEEE Journal Solid-State Circuits*, March 2004.

[8] S. Rajapandian, X. Zheng, and K. L. Shepard. Charge-recycling voltage domains for energy-efficient low-voltage operation of digital cmos circuits. In *Proceedings of the International Conference on Computer Design*, pages 98 – 102, 2003.

[9] H. Yamauchi, H. Akamatsu, and T. Fujita. An asymptotically zero power charge-recycling bus architecture for battery-operated ultrahigh data rate ULSI's. *IEEE Journal Solid-State Circuits*, pages 423 – 431, April 1995.

[10] Byung-Do Yang and Lee-Sup Kim. A low-power ROM using charge recycling and charge sharing techniques. *IEEE Journal Solid-State Circuits*, 38:641 – 653, April 2003.

[11] J. T. Kao and A. P. Chandrakasan. Dual-threshold voltage techniques for low-power digital circuits. *IEEE Journal Solid-State Circuits*, pages 1009 – 1018, July 2000.