Process and Circuit Optimization for Power Reduction using DDC Transistors

The Deeply Depleted Channel (DDC) transistor architecture offers 2 to 3 times improvement in body coefficient and 60 percent improvement in local mismatch in 55-nm technology, extending design techniques such as body biasing with voltage scaling to more recent technology nodes. This article presents a body bias architecture for adaptive correction of manufacturing variation, with less than 0.5 percent area penalty for the bias generators.

Adaptive body bias has been proposed as a method for optimizing system power at a given performance level by correcting for manufacturing window variation. Furthermore, body bias and supply voltage can be optimized together for a given workload to achieve lower power at the required operating frequency. Higher supply voltage with more reverse bias reduces leakage at the expense of active power for low-frequency and low-activity factor designs, whereas reduced supply voltage and less reverse bias (or forward bias) might be optimal for workloads with high active power due to higher activity factors or higher frequency.

Although conventional planar transistors have seen the body effect shrinking at smaller geometries, SuVolta’s Deeply Depleted Channel (DDC) technology extends and improves the effectiveness of these techniques. DDC technology enables a transistor with a body coefficient 2 to 3 times higher than a conventional planar transistor, and a demonstrated 40 to 60 percent improvement in threshold voltage matching at Fujitsu Semiconductor’s 65-nm/55-nm DDC technology node. The improved threshold mismatch results in reduced minimum operation voltages ($V_{DD\min}$) for static RAM (SRAM), allowing lower supply voltages to be used in voltage-scaled designs without requiring an extra power supply for SRAM macros.

The CS250S 55-nm manufacturing process implemented by Fujitsu Semiconductor incorporates DDC technology. The process was optimized to achieve roughly 50 percent power savings on both leakage power and active power using both body-bias and voltage-scaling techniques. The strong body factor available in the DDC device allows effective leakage control by compensating for manufacturing variation with body bias, while matching performance at a reduced supply voltage. The improved threshold voltage mismatch lets the SRAM operate with good performance at the same lowered supply voltage. Because leakage and active power...
are improved by roughly the same amount, a predictable power savings versus a baseline design is achieved across both active-dominated and leakage-dominated designs.

This article expands on my Hot Chips 25 presentation, focusing on the methodology used to enable power reduction on SuVolta test chips manufactured at Fujitsu Semiconductor using DDC transistors, including a body bias architecture with bias generators and process monitors. I also present results on an ARM Cortex-M0 CPU with body bias.

The DDC transistor

The DDC transistor is a device architecture implemented in planar bulk CMOS. The device is planar, the layout rules for the transistors in the CS250S process are identical to those for conventional devices, and manufacturing is mask neutral.

A key difference between a DDC transistor and a conventional bulk transistor is the channel’s structure. A conventional planar-transistor channel is a single region, with the channel doped opposite to the source and drain regions. Modern planar devices use ion implantation through the gate corners (halo implants) to set the device’s threshold voltage and to control short-channel effects caused by drain electric fields and depletion region encroachment into the channel. The high concentration of dopants in the halo and doped channel is a primary source of threshold voltage variation in planar FETs. This variation, known as random dopant fluctuation (RDF), is due to the uneven placement of dopants in the channel.

The DDC transistor depicted in Figure 1 reduces threshold voltage variation by using an undoped or lightly doped channel. The dopants for setting the threshold voltage are placed at a distance below the gate rather than in the channel, thereby reducing RDF. A highly doped layer controls short-channel effects, increases effective drive current, and improves the body factor by 2 to 3 times over conventional devices. Additionally, this layer increases the device output impedance, \( R_{\text{out}} \), enabling a 4× gain increase in analog amplifiers.

In typical SoC designs, operating frequency is determined by slow-corner performance, and leakage power is determined by the fast corner. Worst-case speed and power define the specifications for the part, thus qualifying a part for any socket. Figure 2 shows three clouds of transistor data from skew wafers, representing a typical manufacturing distribution with slow (SS), typical (TT), and fast (FF) DDC transistors. The fast-corner cloud is leakier with higher drive current than the typical and slow-corner clouds. All the clouds are on the same \( I_{\text{off}} = I_{\text{on}} \) line, called the universal curve. This line represents a shift in \( V_T \) from high (at SS) to low (at FF). When a fixed body bias is applied to each corner cloud (0 V for SS, −0.3 V for TT, and −0.6 V for FF), the \( V_T \) for each cloud is shifted toward the slow corner, and the clouds overlap, indicating that the skewed corners can be pulled into a single corner. If a die-specific body bias were applied, the clouds could be collapsed nearly to a point. This is important for controlling leakage at the fast corner.

Figure 3 shows the strategy used to lower both leakage and active power in this work, using DDC transistors with body bias. We start with a comparison of the DDC transistor and the baseline (conventional) transistor at the same voltage, with matched threshold

Figure 1. Example of a Deeply Depleted Channel (DDC) transistor. A typical DDC transistor features a substantially undoped channel (layer 1) to reduce threshold voltage variability, a \( V_T \) setting layer (layer 2), and a screen layer (layer 3) that provides better short-channel effects and a stronger body coefficient compared to conventional devices.
voltage. The DDC transistor has superior performance given its higher effective drive current, $I_{\text{eff}}$. The DDC universal curve is to the left of the baseline transistor curve, indicating lower delay at matched leakage power (matched $V_{T,x}$).

To reduce active power, the operating voltage is reduced from 1.2 V to 0.9 V, a reduction of 25 percent. Delay is $t_{\text{delay}} = CV/I$, where $I$ is the drain current of the transistor that's charging a load capacitor ($C$), and $V$ is the voltage across the capacitor. Current $I$ is proportional to gate overdrive, $(V_{GS} - V_T)$. Because voltage is reduced by 25 percent, current can be reduced by 25 percent as well, while constant delay is maintained. However, with a threshold voltage of about 450 mV and a supply voltage of 1.2 V, overdrive for the baseline is roughly 750 mV, and overdrive at 0.9 V is only 450 mV at matched $V_T$, a reduction of 40 percent. This is more than can be recovered by the DDC transistor’s enhanced effective drive current at matched $V_T$. The DDC curve slides to the right, and is now slower than the baseline. To recover performance, we lower the DDC device’s $V_T$. Leakage is exponentially dependent on threshold voltage; therefore, leakage increases faster than delay improves when threshold voltage is reduced, resulting in higher leakage power at the fast corner if matched performance is achieved at the slow corner. Body bias is applied to reduce the device’s leakage at the corners, achieving a reduction of both active and leakage power while matching performance of the baseline device.

The DDC process implemented in the CS250S technology has been designed such that the device’s $V_T$ is low enough to meet performance requirements at the slow corner. This means that only reverse bias must be supplied, which simplifies the circuits. To create a reverse bias, $n$-channel MOS (NMOS) devices require a negative voltage, whereas $p$-channel MOS (PMOS) devices require a positive voltage that is higher than the supply voltage. Negative voltages can be supplied by a charge pump, whereas a positive voltage can be supplied by a charge pump or a regulator supplied by an I/O voltage, such as a 3.3-V supply.

**Body bias architecture**

The body bias system chosen by the test chip design team uses a two-stage approach for bias generation (Figure 4). Each device (PMOS and NMOS) has a global bias generator that supplies a constant bias voltage. Local digitally programmable low-dropout (LDO) regulators provide independently controllable bias voltages to separate bias regions on the die. The global NMOS bias generator is a charge pump that provides a $-1.2$-V global bias rail, whereas the global PMOS bias generator is an LDO regulator.

---

**Figure 2. Corner pull-in using body bias.** The low-threshold-voltage transistor (LVT) $n$-channel MOS (NMOS) DDC transistors form a distribution from slow to fast in manufacturing. A body bias setting applied per wafer pulls the wafer clouds on top of each other, shrinking the process corners dramatically.
Figure 3. Power reduction methodology used in this work. First, voltage is lowered to save active power, then $V_T$ is lowered to match performance. Finally, body bias is used to save leakage power.

Figure 4. Example of a bias generator architecture. Global bias generators provide unregulated voltages to local bias generator low-dropout (LDO) regulators. Bias generators are powered by the I/O voltage to drive the wells to a known voltage on start-up before the core voltage is brought up.
that provides a 2.2-V bias rail to the local bias generators. The local NMOS LDO regulator reduces ripple on the bias voltage generated by the charge pump, which is unregulated to reduce area. All of the bias generators are supplied by the 3.3-V supply voltage, which is present before the core voltage is brought up during the start-up power sequence. This lets the bias generators drive the wells on the chip to a reverse bias voltage at power-up, which prevents the wells from floating to a forward bias condition when the core voltage is applied.

The NMOS bias generators are sized to drive 1 mA of well current, which is enough to drive 13.7 mm² of logic circuits with −700-mV bias at 125°C. The local PMOS generator is sized to drive 200 μA of well current, which is enough to drive 25.6 mm² of logic circuits, whereas the global LDO regulator can supply 1.6 mA. The total area of the global bias generators is 0.037 mm², and a pair of local generators is 0.016 mm². The bias generators cost about 0.5 percent for a logic area of 10 mm² or larger. Unloaded power is 33 μW for the global LDO, 69 μW for the local LDO regulators, and 3 mW for the charge pump.

SRAM well leakage per unit area for NMOS is about 20 percent higher than logic well leakage at the maximum bias, owing to the density of the SRAM cells as well as the bias conditions of the transistors in an SRAM cell. The maximum bias for SRAM cells is lower than the logic bias because long-channel devices respond more strongly to bias. Bias generators supplying SRAM arrays cost about 0.5 percent for 10 mm² or more of SRAM array area. The local bias generators to supply the SRAM are separate from the standard cell bias generators because of the different bias requirements.

Each local bias region contains a process monitor and controller, as well as the local LDO regulators.

To determine the appropriate body bias for each part, a process monitor must be developed. Ono and Miyazaki suggest a critical-path replica to target delay, paired with a circuit to correct the ratio between the PMOS and NMOS threshold voltages.9 This is an indirect measurement of performance for the NMOS and PMOS devices. Ghosh et al. propose directly measuring the performance of the NMOS and PMOS devices employing slew measurements of an inverter in a ring oscillator using high-speed analog circuitry.10 The slew from each device is converted to a pulse, which is integrated to generate a voltage that powers a voltage-controlled oscillator (VCO). The oscillator’s frequency is indicative of the measured device’s speed.

SuVolta’s process monitor designs improve on the Ghosh approach, directly measuring the performance characteristics for NMOS and PMOS transistors, allowing independent correction of each device. In general, the process monitors are designed to allow a reference device to charge a capacitive load. This models the characteristics of a CMOS gate and allows a prediction of performance for a given device. The monitors are designed to provide a digital output value that is indicative of process corner and speed.

Lee et al. give an example of one of the process monitors developed.11 This monitor uses multiple reference transistor fingers to charge a load capacitor, creating a voltage ramp as the capacitor charges. The ramp is converted to a pulse using comparators that detect when the ramp crosses 20 percent and 80 percent thresholds, and the pulse is extended by a factor of 100 using an analog pulse extender. We measure the output pulse by counting the system clock pulses when the generated pulse is high. The width of the pulse is inversely proportional to the current supplied by the reference device, and therefore represents where the reference device lies in the process window. Compared to the Ghosh design, the analog circuits are simpler in that they are designed to respond to a slew rate that is four orders of magnitude slower than the ring oscillator slew rate. Also, the pulse extender is designed to be less sensitive to process variation than a VCO.

The approach discussed in this article improves the analog pulse-extender process monitor over Lee’s design by using a current mirror to reduce the current that charges the load capacitor. The reference transistor uses multiple fingers sized to match the standard cell library transistors. Multiple fingers are used to reduce the effects of random threshold voltage variation, in order to make
a true average process corner measurement. However, a reference made of a large number of fingers requires a large load capacitor to generate the target slew rate (about 40 ns). A slow slew rate is required in order to simplify the analog circuit design and to generate a final output pulse that is easily measurable. To reduce the required capacitor size, and thus the process monitor’s size, a current mirror samples the reference current and delivers a current into the capacitor that is multiplied by a factor of \(1/n\); this lets the capacitor be reduced by the same multiple. In this case, the original layout was 0.050 mm\(^2\) for the analog components of the NMOS and PMOS process monitors, whereas the improved version is 0.020 mm\(^2\) for the analog components, and 0.025 mm\(^2\) when control circuits and configuration registers are included.

Future iterations of slew-based process monitors might use cascaded current mirrors to further reduce the slew generation circuitry. Additionally, a digital pulse extender has been developed that has the following advantages over the described analog pulse extender: it is mostly synthesizable, except for the slew generator and a reference oscillator; it is smaller in area (0.009 mm\(^2\) versus 0.025 mm\(^2\)) due to the reduced use of analog components; and the pulse magnification factor is precise due to the digital nature of the calculation, whereas the analog pulse extender’s magnification is sensitive to process variation.

Figure 5 illustrates power and performance correction using a SuVolta process monitor, measured in 55-nm silicon. Inverter, NAND, and NOR ring oscillators manufactured on slow, fast, and typical wafers are used as representative circuits, to be adjusted using body bias. Several dies from each wafer are measured at zero bias to evaluate the uncorrected distribution. Each die contains a process monitor that is used to determine the process corner of the NMOS and PMOS transistors. The bias is set per die on the basis of the digital readout from the on-die process monitor. After applying the bias determined from the monitor, we remeasure the ring oscillators. As the figure shows, the performance monitor can decrease worst-case leakage power by 90 percent, as well as decrease the performance spread by 57 percent for the measured dies.

The third component of the bias system is the controller. The controller interprets the results of the process monitor and directs the bias generators to the appropriate voltage. The controller could be a software algorithm that implements a look-up table that correlates process monitor readout to the bias voltage; it could be a software algorithm that sweeps bias voltages at test time to determine NMOS and PMOS bias voltages that meet the performance requirement; or it could be implemented as a hardware control loop. The software look-up table solution has the advantage that it requires no silicon area for the control loop, but it has the disadvantage that it does not respond to process drift, and is inadequate for correcting both artificially skewed corner lots and production lots because the body bias response characteristic may be different between the artificial corners and corners that occur through natural manufacturing variation. The test-time bias sweep addresses the process drift concern, but it requires multiple test operations, probably involving scan shifts. SuVolta has chosen to implement a hardware control loop using a simplified proportional-
integral-derivative controller. The controller iteratively changes bias codes on the programmable LDO bias generators until the NMOS and PMOS performance targets are met. This can be performed at test, with the bias codes programmed into fuses for operation; it can be done at startup in a built-in self-initialization approach; or, the controller can be used in a continuous mode, which allows the system to change bias settings on the basis of performance variation due to device temperature changes.

The active power of the process monitors and controller is small and can be neglected in a fixed-at-test or set-at-power-up operation mode because the monitor is off when the chip is operating.

Results

Figure 6 shows the final result of applying the described methodology to the Cortex-M0 processor. The baseline operates at 1.2 V, and the DDC processor operates at 0.9 V. A reduction in power of about 50 percent is achieved at the fast corner, at matched 350 MHz slow-corner performance, with appropriate per-part body bias. In this evaluation, the same layout-identical design was manufactured in both the baseline and DDC processes.

Body bias pulls in the DDC fast and slow corners almost to a point, whereas the unbiased baseline has a 17 percent spread between fast and slow. Application of the same bias voltage to the baseline would not move the fast-corner leakage all the way to the slow corner, but could recover only approximately 7 percent of the total power. The DDC implementation has a substantial advantage in the reduction in active power achieved by lowering the supply voltage, while controlling leakage using the strong body coefficient. In this measurement, the DDC part uses only reverse bias, and a maximum reverse bias of $-600$ mV is applied at the fast corner to reduce the leakage power resulting from the reduced threshold voltage.

If the supply voltage to the DDC CPU is increased, matched total power is achieved at 1.1 V and 475 MHz, a performance increase of 35 percent. If the DDC CPU supply voltage is matched to the baseline 1.2 V, the DDC CPU performance gain is 55 percent. The DDC transistor in this work is qualified to 1.2 V, with matched reliability to the baseline transistor at the same voltage.

Fujitsu reported on 4 September 2013 that the first commercial product using DDC technology, its seventh-generation Milbeaut still-camera image processor, has entered mass production in its 55-nm CS250S process. This product achieves a power reduction of 30 percent with twice the performance of the previous 65-nm generation part.12 Fujitsu’s design trades off power...
for functionality, adding larger-scale circuits to increase performance and feature set while maintaining a power benefit.

Supply-voltage scaling for power reduction in products such as the Milbeaut camera chip are limited by $V_T$ variability, which increases SRAM minimum operating voltage ($V_{DDmin}$). The DDC transistor’s superior $V_T$ matching lowers the minimum operating voltage by 150 mV at the slow-NMOS, fast-PMOS corner at $125^\circ C$, allowing 0.9-V operation in SoCs with significant SRAM, as well as logic. $V_{DDmin}$ is dominated by mismatch between the transistors in the SRAM bit cells. Mismatch between the PMOS load transistor and the NMOS access transistor results in a write failure if the PMOS is too strong or the NMOS is too weak. In this condition, the access transistor cannot overwrite the load transistor. Mismatch between the NMOS pull-down and the NMOS access transistor results in a read failure if the access transistor is too strong or the pull-down is too weak. In this condition, the access transistor operates too strong of a high level, which turns on the opposite NMOS and flips the cell (read disturb). The DDC transistor improves mismatch in SRAMs by up to 60 percent on the NMOS devices, resulting in a $V_{DDmin}$ of 575 mV at $125^\circ C$, well below the level needed for 0.9-V operation.

Fujitsu Semiconductor Limited is currently in production with its first DDC product, a seventh-generation Milbeaut digital-camera processor that performs at twice the performance of the previous generation with 30 percent less power. At SuVolta, our future work includes building on the presented body bias architecture, including further optimization of the process monitor design and consolidation of the body bias generators, process monitors, and controllers into one drop-in block, which will make a body-bias-enabled design simpler.

In addition to circuit design and body bias architectures, we are continuing to improve and expand our DDC transistor technology offerings by working with commercial foundries and integrated device manufacturers at technology nodes from 65 nm to 20 nm. Applications include SRAM bitcell $V_{DDmin}$ improvement, sense amplifier offset reduction in SRAM and DRAM designs, and low-voltage operation for digital logic.

References

11. S. Lee et al., “A Slew-Rate Based Process Monitor and Bi-directional Body Bias Circuit for Adaptive Body Biasing in SoC.


David Kidd is the senior director of digital design at SuVolta, where he is responsible for SoC design, CAD, SRAM design, and IP development. His research interests include full-custom and semi-custom microprocessor design and design methodologies. Kidd has an MS in electrical engineering from the University of Michigan.

Direct questions and comments about this article to David Kidd at SuVolta, 130 D Knowles Drive, Los Gatos, CA 95032; dkidd@suvolta.com.