Leakage-Tolerant Design Techniques for High Performance Processors (Invited Paper)

Vivek De
Microprocessor Research, Intel Labs, Hillsboro, OR, USA

In sub-100nm technology generation, transistor subthreshold leakage is 100-1000nA/μm for high performance microprocessor logic technology. As gate oxide thickness approaches sub-10Å regime, gate oxide leakage escalates to 10-100A/cm^2. Junction leakages also become significant as doping levels around the junction approach 5X10^{18} cm^{-3}. These excessive leakage currents contribute to large leakage power dissipation during (1) active operation, (2) standby or idle mode and (3) burn-in. In addition, excessive subthreshold leakage degrades noise margin or robustness of performance-critical circuits such as wide-OR domino gates, register files and cache. Large gate oxide leakage also limits circuit fanout. Therefore, high performance and low power processor designs must employ leakage power control techniques to alleviate active power dissipation and delivery challenges, extend battery life and prevent thermal runaway during burn-in. In addition, leakage-tolerant high performance circuits must be used to provide adequate circuit robustness.

The most effective design technique for reducing leakage power is dual-V_T design. Performance-critical transistors are made low-V_T to provide the required performance and high-V_T transistors are used everywhere else to minimize leakage power without impacting processor frequency. Optimal dual-V_T designs can provide the same frequency as a design with single low-V_T, while limiting low-V_T usage to 30%. As a result, leakage power during active, standby and burn-in is reduced by 3X without any performance impact. Of course, process complexity is slightly higher since extra critical masking steps are needed to provide additional transistor V_T's. EDA tools for optimal V_T allocation during all phases of the design flow are critical for successful dual-V_T designs. V_T allocations can be performed at logic gate level or transistor level. While transistor level allocation is most effective for leakage power reduction, it is also the most complex. 75-100mV difference between the high-V_T and low-V_T values is found to be most optimal for typical microprocessor designs.

Another technique for leakage power control is body bias. Traditionally, reverse body bias (RBB) is applied to increase V_T and thus, reduce leakage power during idle mode. However, V_T modulation achievable by RBB reduces at shorter channel lengths and lower V_T values because of worsening short-channel effects and weaker body effect. Therefore, RBB becomes less effective with technology scaling as both V_T and channel length are scaled down. The amount of useful reverse bias is limited to 500mV by increasing junction leakage and drain-body junction breakdown during burn-in. Alternately, forward body bias (FBB) can be used during active mode to lower V_T and provide the desired performance at low voltages. FBB is withdrawn during idle mode to reduce standby leakage power. Since FBB improves short-channel effects, it provides better V_T modulation capability with technology scaling. A 1.1V, 1GHz communication router chip in a 150nm logic technology with FBB demonstrates 3.5X standby leakage power reduction, when compared to lowering V_T by process technology. Full chip area and power overheads of on-chip body bias generators and bias grids are only 2% and 1%, respectively.

Leakage power of a chip is the sum total of individual transistor leakages. Within-die critical dimension (CD) variations cause lengths of many transistors to be below the target value. Since transistor leakages increase exponentially at smaller lengths, leakages of these devices are the dominant contributors to full-chip leakage. Leakage power estimation tools must, therefore, account for within-die CD variations accurately. Both die-to-die and within-die variations in device parameters dictate the frequency and leakage power distributions of microprocessors in volume manufacturing. Only those dies that meet both minimum frequency and maximum power constraints are acceptable. Bidirectional adaptive body bias (ABB) is effective for compensating for these variations. FBB is applied to speed up dies that are too slow. RBB is used to bring dies that are too leaky within the power envelope. The die acceptance rate and number of dies in the highest frequency bin can be improved significantly by ABB, as demonstrated by measurements on a testchip in 150nm technology.

Leakage current through a stack of two or more "off" transistors is an order of magnitude smaller than a single device leakage. This so-called "stack effect" becomes stronger with technology scaling as DIBL worsens. Many circuit blocks in a microprocessor already contain a significant number of transistor stacks in complex logic gates. Thus, leakage power depends strongly on the primary input vector to the block. These "natural stacks" can be exploited for standby leakage power reduction by activating the "minimum leakage" input vector during idle mode. 2X reduction in standby leakage power is achievable for a 32-bit adder with 3-80μs minimum time required in "standby" so that the switching energy consumed for entry into and exit from idle mode is less than 10% of the leakage power saved. In addition, transistors that are not performance-critical can be converted into stacks to reduce leakage without impacting overall processor performance. Thus, "stack forcing" allows one to emulate behavior of high-V_T devices not available from the process technology. Using a single-V_T process in conjunction with "stack forcing" can reduce leakage power of a 32-bit instruction decoder block by 3X without any performance degradation.

Finally, noise margin degradation of wide-OR domino gates and register files due to excessive leakage requires keeper transistors to be upsized and static stages to be deskewed. The resulting performance loss is severe and unacceptable. Conditional keepers can be used to provide the desired robustness with minimal performance loss. A pseudo-static local bitline scheme for register files also reduces bitline leakage significantly, allowing target performance to be achieved in the presence of excessive leakage.

ISPD'02, April 7-10, 2002, San Diego, California, USA.
ACM 1-58113-460-6/02/0004.