New SOC (system-on-chip) process technology brings with it some process, voltage, and temperature effects, as well as IR drop, which all impact timing closure. Designers can model the on-chip variation of these parameters without guardbanding and needlessly wasting design margin in increased area, increased power, or reduced performance. As core voltage drops below 1V in designs using 90-nm and smaller processes, IR-drop effects become more prominent. If they are not adequately addressed in the design flow, these IR drops can lead to setup-and-hold failures than can render the system inoperable. The clock tree is one of the most sensitive parts of the design; thus, designers must carefully analyze it for on-chip-variation effects. Although you can quantify some of these effects using physical data, such as relative cell location and proximity to the power grid, you must statistically model other effects, such as transistor-threshold-voltage variation (Reference 1). Another issue that complicates timing analysis is RC-interconnect variation. Interconnect delay in deep-submicron processes can often dominate total delay; thus, you should model it as accurately as possible.

**EXPLORING THE PROBLEM**

A variety of wafer-processing parameters can affect the overall performance characteristics of an SOC. Historically, designers have analyzed digital chips at the various voltage and temperature corners of a process without modeling across-chip-variation effects. In the last few years, however, integrators of larger and more complex SOCs have started investigating on-chip-variation effects on overall performance—a factor that analog designers have always modeled through transistor-mismatch analysis. Designers have typically analyzed these effects by introducing a constant-variation variable. An alternative model can drastically reduce the pessimism of such constant-variation factors by quantifying the nonrandom effects of variation and using physical information for all gates in a design.

The variations that can directly affect timing are variations in transistor width and channel length, reduction in supply voltage at the cell from IR drop in the supply lines, and variations in interconnect resistance. Variations in critical dimensions of minimum-width lines can occur during mask generation and in the process stages for active and polysilicon layers. These variations cause changes in both the width of a transistor’s active area and the device’s effective channel length. The active area and polysilicon layers are independent and therefore can vary independently. Manufacturers’ libraries typically characterize these layers at the absolute worst-, typical-, and best-case corners. The final die most likely does not have both worst- and best-case-speed transistors on it, but it will see significant variation in both effective channel length and width (Reference 2).

Variations in operating voltage can be dramatic, as well. For example, assume a die in a 90-nm process with nominal power-supply voltage of 1V±10%. Further, assume that IR drop on the supply line is no more than 50 mV at any cell and that the standard cells have an increase in delay due to IR drop of 1.5% per 10 mV of drop in power-supply voltage. If you can control your power grid only within a 100-mV maximum IR drop, the result is: 100 mV×(1.5%/10 mV)=15% delay variation. Even if you limit the IR drop to 50 mV, the variation would be 50 mV×(1.5%/10 mV)=7.5% delay variation. Both cases are realistic and significant.

Variations in the width or height of interconnecting metal lines can also cause effects. For example, consider a signal route in copper interconnect for a deep-submicron process with nominal statistical-process-control limits of ±10% for width and ±15% for height. The worst-case scenario for RC extraction is generally the corner with maximum width and height, and the best case generally comprises minimum width and height. Calculating for a process results in a delta range from −14 to +24%. Looking at the absolute delta delay for a 100-micron-long line results in a nominal delay of approximately 116 psec, with a best- and worst-case-delay vari-

**Figure 1** IR-drop differences in clock-tree branches can induce skew, causing setup or hold errors.
atation of approximately 16 psec faster or 28 psec slower.

Typically, designers think of analyzing timing with worst- or best-case parasitic extraction, but a manufacturer could have processed two layers of interconnect on the same die at opposite extremes. For example, Metal 2 could be a best-case scenario, and Metal 3 could be a worst-case scenario. For longer nets that often switch layers, this effect will average out. Again, critical paths most likely contain many short lengths, and the variation may not have much impact. The noncritical paths may have long nets, which variations in R and C could affect, but these paths would also have greater margin to compensate for these effects.

**TIMING FLOW FOR ON-CHIP VARIATION**

Digital designers typically simulate circuits at extreme process corners for timing sign off. That analysis typically includes no mismatch in gate or interconnect performance at any corner. Designers assumed the cells to be at the worst- or best-case corner. Unfortunately, that assumption is no longer valid. Because of the complexity of deep-submicron processes, designers can no longer ignore the variation between devices and interconnect characteristics on the same die. This fact is most evident on the clock network in which speedup and slowdown in the clock latency to logically dependent flip-flops can lead to slower parts and to failure to hit performance targets. In the worst case, these issues can lead to setup-and-hold failures and, ultimately, inoperable devices (Figure 1).

A design flow can avoid such issues without excessive guard-banding. You can describe on-chip variation using a random and a deterministic component. The random component of critical parameter variation occurs from lot to lot, wafer to wafer, and die to die. Examples are variations in gate-oxide thickness, implant doses, and metal or dielectric thickness. The deterministic component comprises variations that you can predict from their location on the wafer or the nature of surrounding patterns. These variations relate to hot spots, density effects, and the relative distance of devices. Examples are variations in gate length or width and interconnect width. Another component of deterministic variation results from radial or linear gradients specific to each processing tool.

To model these effects, you can use timing-derating factors during static-timing analysis that allow you to specify a percentage of the timing that you want to adjust on a percentage basis the speeding up or slowing down of all or specific gates, nets, or both in the design. The total derating is a sum of the random and deterministic variations on a per-cell or-net basis.

**MODELING VOLTAGE AND IR DROP**

Designers have traditionally performed IR-drop analysis for power-grid design to ensure electromigration compliance and to meet a targeted overall IR-drop allowance. But, in advanced processes, IR drop can have a significant impact on transistor performance, and, thus, you should account for it in static-timing analysis. If you assume that the clock-tree buffers for the launch flip-flop in Figure 2 are on the higher end of the IR-drop budget and that the clock-tree buffers for the capture flip-flop are on the lower end of the budget, a significant difference will occur in the arrival times of these two clocks at their respective flip-flops. If the skew is too large, hold failures will result. To account for the timing impact of IR drop, you must establish a methodology that does not rely on k factors to scale timing arcs within the impacted cell (Figure 3).

To quantify the magnitude of the impact of IR drop on timing, designers can perform Spice simulations on a sample of the standard cells in a number of libraries, each of which includes clock buffers, inverters, and integrated clock-gating cells. When the IR drop is within 5 to 10% of the nominal power-supply voltage for each process, voltage, and temperature effect, the increase in cell delay is approximately linear with decreasing power-supply voltage for the 130- and 90-nm processes. Figure 4 shows the delay for a gated-clock cell at a worst-case corner with nominal power-supply voltage of 1.35V. Note that the delta delay is only approximately linear. Because the increase in delay for the region of interest is a linear function of the voltage, you can approximate the percentage increase in delay per a given number of millivolts’ IR drop for each cell of interest.

Unfortunately, the slope of the curve can differ for each standard cell in a library. If the slope of the delta delay were identical for each cell in a library, then global k factors, if available, would suffice. But data shows that, in the real world, they can be unreliable.

**Figure 2** You can use a map of IR drop across the die to adjust the delay values on individual cells in the netlist.

![Figure 3](image) Measured values of delay versus power-supply voltage show only an approximately linear relationship.
Once you have determined the derating factor for each type of cell, you can write a script to combine the output of the IR-drop-analysis tool with the Spice-derived derating factor, creating a timing derating for each instance, as the IR-drop tool’s output describes. Static-timing analysis can now more realistically account for the actual power-supply values at each cell.

INTERCONNECT VARIATION

Another area of on-chip variation is in interconnect height and width, resulting in variation in both resistance and capacitance. Because the delay from interconnect is becoming more dominant as geometries shrink, you should pay attention to accurate modeling of interconnect variations. Two potential sources of this variation are the CMP (chemical-mechanical-planarization) process and the proximity effects in the photolithography and etch processes. Variation in the CMP process results from the difference in hardness between the interconnect material and the dielectric. Ideally, after the designer has etched trenches into the dielectric below an interconnect layer and copper on the wafer, the CMP process removes the unwanted copper, leaving only lines and vias. The copper line is softer than the dielectric material, resulting in “dishing” and erosion, which cause uneven removal of the copper and dielectric. Dishing is a function of line width and density, and erosion is a function of line space and density (Figure 5).

Another source of variation in thickness due to CMP is a more random variation resulting in a gradient across the wafer. You can see this gradient in die-to-die variations and even across-die variations for large die. You would ideally model this random, nondeterministic variation statistically. However, if you can obtain process data to model this variation, then you can model it deterministically as a function of position on the wafer. In this scenario, you give an adder or subtracter, depending on the x,y position on the die, to the RC value.

Etch-proximity effects appear as “microloading,” which means that the etch process overetches isolated lines. A dual-damascene structure uses only a single metal-deposition step to simultaneously form the main metal lines and the metal in the vias. Photolithographic effects also cause problems. Diffraction and local scattering in photolithography may overexpose densely spaced lines and underexpose isolated lines (Figure 6). Tiling and metal slotting reduce the variation in feature density and mitigate these effects. Tiling algorithms give different results, but a general rule states that a less dense gradient yields smaller line-width variations on the die. Tiling can also result in small additional delay effects on timing. The final design may not meet the desired target frequency once you account for tiling. Whether it does depends largely on the design and the methods you use to meet the tiling requirements.

MODELING PROCESS VARIATIONS

The key parameters that control CMOS transistors’ drive current are width and length, including random and nonrandom effects, and threshold voltage and gate-oxide thickness, both including only random effects. Random effects are the day-to-day, lot-to-lot, or wafer-to-wafer variations. These include variations due to implant doses, oxide-growth rates, and varying stress levels in the gate oxide, across wafer-photo gradients, or across etch gradients. Transistor mismatch is proportional to the area.

Because each process and library has a standard channel length in standard-cell design, you minimize variation by choosing cells that are wider than minimum. By restricting the clock-tree cells to high-drive cells, you ensure smallest variation and thus smallest mismatch across the various subtrees. You may see
a large distribution of effective channel length and width across a die. That variation can differ, based on the fab and the process, but a worst-case variation should be one-half the overall process limit for length and width across the wafer.

Temperature variations can also cause differences in electrical behavior and, hence, timing. Fortunately, it is uncommon to find opposite temperature corners on the same die during operation. But nonuniform on-chip power distribution, interconnect heating, and thermal characteristics of the die and package materials can influence actual operating temperatures. Temperature profiles, for the most part, follow IR-drop maps but may differ slightly because of density, hard-macro placement, and other effects.

Timing analysis and closure in deep-submicron processes are becoming more challenging as process variation increases. Designers must more carefully analyze variation’s effects to achieve reasonable yields across process, voltage, and temperature effects. This variation is increasing, and designers must realistically model it. If they do not, designs incur a much greater risk of missing performance targets, having lower yields, or not functioning. If designers realistically model variation, they can realize higher yields.

REFERENCES

AUTHORS’ BIOGRAPHIES
Anis Jarrar is principal design engineer at Freescale Semiconductor (Austin, TX), where he has worked since 1993. He is currently developing PowerPC- and ColdFire-based chips, focusing on synthesis, timing, and optimization techniques for deep-submicron designs. He has bachelor’s and master’s degrees in electrical engineering from the Georgia Institute of Technology (Atlanta).

Kirk Taylor is a principal design engineer at Freescale Semiconductor’s Design Environment Organization (Austin, TX), where he has worked since 1992. He currently focuses on methodology development in synthesis, static-timing analysis, signal integrity, and on-chip variation in deep-submicron design. He has a bachelor’s degree in electrical engineering from Texas Tech University (Lubbock, TX). His personal interests include travel, food, and wine.