Accounting for Very Deep Sub-Micron Effects in Silicon Models

By Farid Najm, Jay Abraham, Silicon Metrics Corp., EEdesign
Jan 9, 2001 (10:48 AM)
URL: http://www.eedesign.com/story/OEG20010109S1228

With feature sizes ranging from 180 to 100 nm, today's very deep submicron (VDSM) semiconductor technologies pose new modeling challenges to design closure and timing sign-off. For example, inadequate modeling of nonlinear waveforms and variations in input pin thresholds cause inaccurate interconnect delays and curious modeling anomalies like negative cell delays. In addition, instance-specific operating points are required to model the effects of on-chip voltage variations (IR drops) and significant temperature gradients accurately.

As the size of VDSM designs continues to increase, designers need efficient analysis at the physical level to maintain their productivity, let alone improve it. To quickly achieve timing closure and sign-off, Spice-level accuracy at higher levels of abstraction (starting with cell level) is required. However, problems with inconsistent use and inaccuracies of current cell library formats make this difficult. Because delay calculation is responsible for timing closure and sign-off throughout the design flow, it is important to consider the impact of library model accuracy and consistency on delay calculation.

Traditional submicron (>350 nm) flows partition path delay into two discrete components: cell delay and interconnect delay. Although the past few years have seen tremendous focus on the need for accurate interconnect models, archaic cell-delay models based on simple transition waveform descriptions are still in use. This methodology can no longer be applied in modern designs because VDSM effects cause significant nonlinear cell driver characteristics and transition (slew) waveform shape. To solve this dilemma, new sophisticated modeling capabilities must be developed for state of the art cell libraries. The following sections explain the problems and deficiencies of current cell models.

Waveform Shapes Signal delay through a logic cell traditionally is represented by two components, propagation delay and signal slope (called slew rate, signal slew or simply slew). Each of these is normally represented as a function of input slew and output load. This functional dependence is captured either in a table or with a simple linear or polynomial equation. This model has worked well for so long that many take it for granted today. However, the assumption that waveforms can be closely approximated with a simple linear model is breaking down with modern technologies, mainly because of the increased importance of interconnect delay.

Simply using accurate interconnect delay models does not enable accurate timing analysis. Interconnect models are significantly influenced by cell models, and the driver model (specifically nonlinear driver impedance) and the interconnect model are interdependent. Cell models also must be characterized accurately to account for their impact on interconnect delay.

Because of technology scaling, delays in logic cells have been reduced continuously so they are now

in the picosecond range. However, interconnect delays have not scaled down in the same way because reduced wire cross-sections lead to increased resistance and larger RC parasitic delays. As a result, interconnect delays have become a larger fraction of overall delay and are expected to become larger than cell delays for technologies smaller than 250 nm. The impact of this on waveform shape at the output of a gate leads to the conclusion that a linear ramp is no longer good enough. The waveform is significantly different from a straight line during a transition and includes a waveform tail, mainly because of the increased line resistance. An example of this behavior is shown in Figure 1, which shows a rising input, V1, to an inverter driving an interconnect load. V2 represents the waveform at the output of the inverter, and V3 represents the waveform at the output of the interconnect load. The far end (V3) waveform has a distinctive tail.

Figure 1: Non-Linear Output Waveform of an Inverter

As depicted in the figure, inaccuracies can result when approximating a nonlinear waveform with a linear waveform (shown in heavy gray). In the example, there are 50 picoseconds in slew variation when one incorrectly uses global 80 percent to 20 percent slew thresholds. Clearly, it is more accurate to choose path-unique 80 percent to 40 percent thresholds to accurately model the signal being generated at the output of this interconnect.

Negative Delay

Another problem with today's models is in the area of negative delay. When the input to a logic gate is slow (perhaps due to increased RC delay in its fan-in net), the output of the (very fast, i.e., low switching threshold) gate can make a full logic transition before its input has finished its logic transition. A linear ramp forced on the input signals (as part of the traditional delay model of propagation delay and slew rate) often leads to a situation where the propagation delay of the gate (measured as the separation between the 50 percent rise/fall points on the linear ramps) will be negative. Thus, the causality relationships between signals can be lost, which sometimes leads to strange situations that show unreal circuit errors. Figure 2 shows an example of this behavior for an inverter.

Figure 2: Negative Delay

One way to maintain causality is to force any negative delay to become zero or some small positive value, but this translates to a pretense that the gates are slower than they actually are and does not seem to make sense as a way of validating high-performance circuits.

Circuit-modeling techniques that do not account for varying switching thresholds cause negative delays. Specifying identical switching threshold values to all cells in a technology library does not model these cells accurately. Rather, accurate threshold-modeling techniques that are unique functions of cell types, pins, voltage, temperature and process are needed. Additionally, the use of a waveform model that is more sophisticated than a simple ramp also enables better modeling of these anomalies.

Supply Voltage Variations

To reduce the power dissipation, power supply voltages have been reduced over the years, from 5 V, to 3.3 V, to 1.2 V and even to less in the future. Because the MOSFET current is proportional to (Vgs-Vt), the threshold voltage has been reduced to maintain circuit speed. This leads to at least two problems: noise margin and leakage current. It also reduces the headroom available for supply voltage variations. At 1-V supply, a 200-mV variation is suddenly 20 percent of the power supply and cannot be tolerated. Circuit designers typically budget for a 5-to-10 percent variation in power supply. A large IR drop slows down a cellís performance significantly, potentially causing a circuit to fail. Supply voltage drop also can translate to clock jitter, which is problematic. Thus, supply voltage variations are suddenly a performance problem. Figure 3 highlights this concern.

Figure 3: Non-Linear Change of Slew With Respect to Voltage Variation.

SPICE analysis of 2-input NAND gate on a 180 nanometer process at 36*C.

The figure shows a 5 percent variation in voltage can result in a 15 percent change in slew. Additionally, the slew change is nonlinear with respect to voltage variations. However, most library models in use today account for voltage variations using simplistic linear approximations for changes in delay and slew. As the dashed linear reference shows, there can be as much as 10 percent error. If the slew is computed incorrectly, erroneous path delay values will follow because voltage variations were not accounted for.

The power supply voltage available on the power supply network inside a chip varies temporally and spatially (both across one chip and across different chips). Certainly, the variations in the external power supply are a factor; also important are the process variations (both intra- and interchip) in the layout of the power supply and ground network. These variations are well understood and usually can be bounded up front. They are typically accounted for by doing worst-case/best-case analysis to make sure that the circuit is designed to work under both supply voltage extremes.

But there is another kind of supply voltage variation that is very hard to bound upfront. It is caused by the voltage drop (also called IR drop) in the supply/ground network due to the power supply current. This creates a temporal and spatial variation in the power supply that traditionally was small. However, this variation now is increasingly becoming a significant effect as Vdd is scaled down, metal lines are becoming thinner (hence more resistive) and power supply currents are becoming larger. Figure 4 shows an example voltage map, in which a worst-case voltage drop of 160 mV is found (near the top right corner of the chip) at a certain time point in a dynamic simulation of the power grid.

Figure 4: Voltage (IR Drop) Map

This effect is hard to bind up front because either it is not clear what the worst case is or it is too pessimistic to apply the worst case everywhere. Therefor, checking whether the design is safe from supply voltage variations becomes part of the verification activity during physical design. The effect of supply voltage variations on circuit delay must be predicted. Given the size of the supply grid (100 million branches in large designs) this is a formidable task. The circuit first must be simulated to compute the supply currents drawn by the gates; then the power/ground network must be simulated to measure the voltages everywhere. Because the supply voltage affects gate delay, it might take two or three iterations for the results to become meaningful. To speed up the process, the circuit must be simulated at the cell level, instead of the transistor level. This creates a need for cell models that provide the power supply current as a function of supply voltages. No such modeling capability exists today.

Temperature Variations

Chip temperature is proportional to the power dissipation. With the increased power dissipation in recent years, chip temperatures also are increasing. Higher temperature causes a circuit to slow down because of reduced electron mobility, higher thresholds and increased interconnect resistance. To account for this, the traditional solution is to assume certain best-case and worst-case temperatures and to verify the circuit timing under both extremes. However, recent trends necessitate new solutions to this problem. Mainly, the increase in overall power dissipation and the use of low-power design techniques in which parts of the chip are selectively shut down to reduce power have led to big differences in temperature across the chip surface. For instance, it has been reported that temperatures across the surface of a large microprocessor can vary by as much as 30°C, as in the example shown in Figure 5. This translates directly to significant differences in delay between different parts of the circuit, which introduces skew between signals and can affect functionality.

Figure 5: Temperature Map

Table 1 lists typical slew values for a two-input NAND gate in a 180-nm technology. The table data shows that typical temperature variations on a die can result in a difference of more than 7 percent variation in slew of one simple cell. The variation in slew is even larger in complex cells because of their larger surface area.

Table 1: Temperature Dependence of Slew for a Simple Inverter (2.0 volts)

Thus, the answer lies in not treating temperature as a global fixed variable, but making it a local variable for each cell and making it variable over time. In other words, instance-specific temperature must be a variable in the cell-simulation model. Coupled with a temperature contour plot from a physical analysis of the layout, this leads to detailed analysis of the impact of temperature on the design.

Conclusion

It is clear that the advent of sub-150-nm VDSM technologies will bring to bear a host of previously ignored electrical and physical artifacts. These artifacts can no longer be ignored or approximated for high-performance IC designs where reducing design guard bands is the name of the game. Figure 6 highlights the interdependencies of slew, delay (as a dependent component of slew), voltage and temperature.

Figure 6: Inter-dependencies of Slew, Voltage, and Temperature

All of these items are linked so as to have nonlinear causal effects. Change in one can cause a change in another and thus cause a change in yet another and so forth. The interconnect delay component of path delay is heavily dependent on the input slew. Therefore, an error in slew leads to an error in the whole path delay. Design methodologies must evolve to incorporate these aspects of cell models into mainstream flows. Because these model characteristics are nonlinear, the use of file formats (like the industry de facto standard .LIB format) will not suffice. A new modeling methodology that can self-compute for given environmental condition (i.e., voltage, temperature, process and RLC load) is needed. Binding algorithms with the data permits this sort of self-evaluation for the models. Standards organizations and industry consortiums have attempted to address this by proposing API-based executable cell models. This standard, known as the IEEE 1481 Delay and Power Calculation System, is rapidly gaining momentum. This is a step in the right direction, because programmatic API-based models can evaluate delay values for any given unique environment.

However, as with all standards, adoption takes time. Given this, the failure to account for these complex cell model characteristics either results in designs with large guard bands that impair the technology providerís ability to showcase the performance of its silicon or, worst case, results in designs that simply do not work. Many specialists have focused intensely on the development of high-speed semiconductor devices with small feature sizes. However, equal consideration must be given to advanced modeling techniques--and those techniques, in turn, must drive accurate representation of whole path delays, so the design community can maximize the use of these advanced process technologies in the most efficient and productive way.