Interconnect Delay Model for Wide Supply Voltage Range Repeater Insertion in Sub-22 nm FinFET Technologies

Alexander E. Shapiro* and Eby G. Friedman
Department of Electrical Engineering, University of Rochester, Rochester, NY 14627, USA
(Received: 24 January 2017; Accepted: 23 March 2017)

Energy efficiency has been a primary focus over the past decade. Energy saving techniques such as dynamic voltage and frequency scaling, power gating, and many-core systems-on-chip, have been extensively studied. These techniques, however, struggle to deliver desired energy efficiencies while failing to exploit wide supply voltage ranges in mobile sub-22 nm microprocessors. Additionally, scaling exacerbates these problems, resulting in significant parasitic interconnect resistances. A major factor that limits dynamic wide voltage ranges and frequency scaling is the inability to optimize interconnect to support a wide voltage range from nominal to subthreshold voltages. The primary contribution of this work is the analysis and optimization of repeater insertion for wide supply voltage range applications. A closed-form delay model supporting wide voltage ranges is developed to enable this analysis. The model supports an ultra-wide voltage range from nominal voltages to deep subthreshold voltages, and repeater size multipliers up to 640 times the minimal size. The model is validated with SPICE using a commercial 14 nm FinFET transistor model for interconnect resistances and capacitances up to, respectively, 2 kΩ and 2 pF. The model exhibits good accuracy across the entire parameter space, with the worst case error ranging from −3.5% to 7.3% (−12% to 13% in the subthreshold region) for single stage delays, and from −17% to 9% (−15% to 17% in the subthreshold region) for long interconnect lines with repeaters. Challenges to repeater insertion are also evaluated based on the proposed model.

Keywords: Repeater Insertion, FinFET, DVFS.

1. INTRODUCTION

During the past decade, the microelectronics industry has shifted focus to mobile platforms as the personal computing standard. These battery powered devices emphasize energy efficiency in modern sub-22 nm CMOS technologies. Techniques to reduce power consumption include reducing the supply voltage due to the quadratic dependence on power consumption. In those cases where high speed operation is not required, the supply voltage is reduced to near the threshold voltage. Circuits that operate near the threshold voltage benefit from higher energy efficiency due to a balance between dynamic and static power consumption. In extreme cases where minimum power consumption is required, the supply voltage is reduced into the subthreshold region. In this region of operation, the dynamic power is sufficiently low that the static energy consumption is the major contributor to the total energy consumption.

The primary disadvantage of reduced supply voltages is slower circuit speed. In many cases, circuits with variable workloads operate at different processing speeds. These circuits are energy efficient when optimized to operate with a single voltage source. For these applications, dynamic voltage and frequency scaling (DVFS) is used to monitor the workload and adjust the supply voltage and speed to optimize the energy efficiency of the system. The DVFS technique is widely used in energy efficient microprocessors, providing significant savings in power. These energy savings are, however, constrained by the availability of different voltage/frequency performance states of the DVFS controller. Providing a DVFS controller with voltage/frequency performance states encompassing a wide voltage range is key to maximizing the energy efficiency of the DVFS-based system. Providing multiple voltage/frequency performance states is however a challenge since the circuit needs to be optimized for each performance state over a wide voltage range. A recent approach to overcome this ultra-wide voltage range optimization challenge is to split the problem into...
two (or multiple) separate parts.\textsuperscript{6} Optimize two (or more) separate cores for different regions of operation. One core is configured to work at nominal speeds using intermediate supply voltages while a second core is configured to operate at low voltages. Only one relevant core is active at a time to perform scheduled tasks with higher energy efficiency, while the other core is power gated.\textsuperscript{7} This solution, however, requires significant area overhead and complex synchronization between the cores.

These techniques, DVFS, and many-core systems share a common underlying difficulty. The challenge is to exploit the energy efficiency potential of these techniques while operating over a wide supply voltage range. An analytic model that supports a wide voltage range while providing accurate delay estimates of the critical paths with a variable number and size of inserted repeaters is needed.

Although repeater insertion has been extensively studied in the past,\textsuperscript{8–10} these single supply voltage solutions neglect the effects of a wide voltage range on the repeater insertion process. Additionally, these results predate FinFET technology which diminishes the relevance of these planar bulk CMOS $I–V$ models to modern applications.

The objective of this work is to address these limitations, enabling an analytic evaluation of the critical path delay with a variable number and size of inserted repeaters across a wide range of supply voltages. The proposed delay models and repeater insertion technique are validated by comparison to SPICE simulation with industrial 14 nm FinFET models. The simulation is performed with an input transition time of 1 ps, output load capacitance of 3 fF, temperature of 25 °C, and a wire resistance ranging from 200 \(\Omega\) to 2 k\(\Omega\) and capacitance from 200 fF to 2 pF.

The rest of the paper is structured as follows. An overview to the repeater insertion process and FinFET short-channel transistor models is provided in Section 2. The single stage RC delay model is described in Section 3. In Section 4, the single stage model is extended to a complete interconnect delay model with inserted repeaters. In Section 5, the proposed interconnect model is evaluated to address different challenges of repeater insertion for wide supply voltage range applications. Finally, the paper is concluded in Section 6.

2. EXISTING FinFET TRANSISTOR AND INTERCONNECT DELAY MODELS

The interconnect resistance in deeply scaled sub-22 nm technologies is increasing with each technology node due to longer distances and smaller cross sectional areas. To optimize the propagation delay through a resistive line, a repeater insertion technique is required that breaks the large resistance into sections. The repeater insertion technique proposed here is based on an interconnect delay model with repeaters that considers a wide supply voltage range. This wide supply voltage range delay model is based on the $RC$ delay expression and

\[ td(v_r) = 0.1 \ln \left( \frac{1}{1 - v_r} \right) (R_T C_T + R_T + C_T + 0.4) \]  
\[ R_T = r_i / R \]  
\[ C_T = c_i / C \]  
\[ v_r = V(out) / V_{dd} \]

In (1), $r_i$ and $c_i$ are, respectively, the resistance of the driving transistor and the input capacitance of the next stage (the gate capacitance of an inverter). The ratio of the output voltage at time $t_v$ over the supply voltage ($V_{dd}$) is $v_r$, and $R$ and $C$ are, respectively, the resistance and capacitance of the interconnect. This $RC$ delay model is widely used and provides high accuracy.\textsuperscript{11}

Although a number of analytic current models of FinFET transistors are available,\textsuperscript{12–16} these models either do not provide a closed-form expression or consider only long channel effects. The equivalent resistance ($r_i$) of the driving transistor is based on the short-channel FinFET transistor model of (5)\textsuperscript{16} which considers inversion sheet charge densities ($Q_{sa}$ and $Q_{sd}$),

\[ I_D = 4\mu_n W \frac{2}{L} V_{th}^2 \left[ (Q_{sa} - Q_{sd}) + 1/2 (Q_{sa}^2 - Q_{sd}^2) \right] \]  

This short-channel FinFET model provides sufficient accuracy for deeply scaled sub-22 nm technologies commonly used in industry.
3. SINGLE STAGE DELAY IN WIDE SUPPLY VOLTAGE RANGE APPLICATIONS

A single stage delay model of a CMOS inverter driving an RC load is described in this section. The single stage consists of an inverter driving a parasitic interconnect impedance and the input capacitance of the next stage. The model considers the discharge time through the NMOS FinFET transistor, and leakage current through the complementary PMOS transistor.

The single stage delay across a wide supply voltage range is

\[
 t_{\text{single-stage}}(v_t) = 0.1RC - \ln(1 - v_t)(R_T C_T + R_T + C_T + 0.4)RC + F_w[R, \text{mul}, V_{dd}] \tag{6}
\]

This model consists of two major parts. The first term of (6) is the delay as a function of the nominal voltage as described by Sakurai in (1). The second term, \( F_w \), as described in this section, provides the wide supply voltage range dependent component by relating the interconnect resistance \( R \) and repeater size multiplier \( \text{mul} \) across a wide range of \( V_{dd} \). This additional term is required since the \( R_T \) ratio in the first term describes the delay assuming a constant resistance of the driving transistor as opposed to a dynamically changing resistance over the transition duration. Additionally, the first term describes the delay at a nominal operating voltage as opposed to over a wide range of supply voltages. The wide voltage term, however, is not an explicit function of interconnect or load capacitance since these quantities remain constant with changes in the supply voltage and are considered within the first term of (6).

The equivalent transistor resistance \( r \), in the single stage delay model is based on the FinFET \( I-V \) model of (5) with \( V_{in} \approx 380 \text{ mV} \). The maximum resistance over the entire range of \( V_{DS} \) for a target gate voltage is

\[
 r(V_{GS}) = \max\left\{ \frac{V_{DS}}{I_D(V_{GS}, V_{DS})}, V_{DS} = 0, \ldots, V_{dd} \right\} \tag{7}
\]

The maximum resistance for each \( V_{GS} \) provides greater accuracy over a wide range of supply voltages as compared to a minimum, average, or weighted average resistance.\(^{11} \)

The input gate capacitance is

\[
 c_r(n, \text{fin}, W_{eff}) = 4 \times \text{mul} \times \text{fin} \times C_{ox} \times L \times W_{eff} \tag{8}
\]

and includes both of the PMOS and NMOS transistors of an inverter with a size multiplier \( \text{mul} \), number of fins \( \text{fin} \), effective width \( W_{eff} = H_{fin} + W_{fin}/2 \), and length \( L \).

The wide supply voltage range component of (6) is

\[
 F_w[R, \text{mul}, V_{dd}] = \alpha R^2 + \beta R + \gamma \tag{9}
\]

where \( \alpha, \beta, \) and \( \gamma \) are fitting coefficients that characterize the disparity between the simulated delay and the delay provided in (1). Note that the expression is a quadratic function of \( R \) and has units of time. This term supports a wide supply voltage range, and reduces the error, assuming a constant equivalent resistance of the driving transistor. The coefficients \( \alpha(C/\Omega), \beta(C), \) and \( \gamma \), respectively, (10), (11), and (12), compensate for the nonlinearity due to the resistance of the transistor as a function of the transistor size multiplier \( \text{mul} \) and the supply voltage.

\[
 \alpha = -3.8624e^{-17} \tag{10}
\]

\[
 \beta = -1.892e^{-13} \times V_{dd}^{3.739} \times \text{mul}^8 \tag{11}
\]

\[
 \gamma = 6.26e^{-14} \times [V_{dd}^{2.083} \ln(\text{mul}) - 12.46V_{dd}^{1.743}] \tag{12}
\]

\[
 \delta = -1.629V_{dd}^2 + 1.737V_{dd} - 1.468 \tag{13}
\]

\( F_w[R, \text{mul}, V_{dd}] \) supports a supply voltage range, ranging from 0.8 volts to 0.4 volts. The voltage range, however, can be further expanded to a wide supply voltage range including the near threshold and subthreshold regions from 0.4 volts to 0.2 volts. In this voltage range, the coefficients \( \beta, \gamma, \) and \( \delta \) are, respectively,

\[
 \beta = -2.348e^{-14} \times [V_{dd}^{5.603} \times \text{mul}^8 + 1] \tag{14}
\]

\( F_w[R, \text{mul}, V_{dd}] \) supports a supply voltage range, ranging from 0.8 volts to 0.4 volts. The voltage range, however, can be further expanded to a wide supply voltage range including the near threshold and subthreshold regions from 0.4 volts to 0.2 volts. In this voltage range, the coefficients \( \beta, \gamma, \) and \( \delta \) are, respectively,

\[
 \beta = -2.348e^{-14} \times [V_{dd}^{5.603} \times \text{mul}^8 + 1] \tag{14}
\]

\( F_w[R, \text{mul}, V_{dd}] \) supports a supply voltage range, ranging from 0.8 volts to 0.4 volts. The voltage range, however, can be further expanded to a wide supply voltage range including the near threshold and subthreshold regions from 0.4 volts to 0.2 volts. In this voltage range, the coefficients \( \beta, \gamma, \) and \( \delta \) are, respectively,
\[ \gamma = 2.194E - 15 \times [V_{dd}^{-5.5047} \ln(mul) - 22.62V_{dd}^{-4.742}] \]  
\[ \delta = 2.43V_{dd}^2 - 1.58V_{dd} - 0.7874 \]

The single stage delay model is validated across an extended range of repeater sizes, interconnect resistances and capacitances, and supply voltages. The error of the analytic delay as compared to simulation ranges between 7.3\% to 3.5\% for nominal to near threshold voltages. For an extended voltage range that includes subthreshold voltages, the error ranges from 13\% to 12\%. The error as a function of interconnect resistance is shown in Figure 2.

4. INTERCONNECT DELAY MODEL

An analytic model of the interconnect delay considering a wide supply voltage range is described in this section. The proposed interconnect delay model is an extension of the single stage model described in Section 3 for a number of stages. This delay model assumes that a stage starts to transition once the input passes the 50\% voltage level. \( v_i \) is larger than 50\% to compensate for the instantaneous input transition in (1).

The contribution of the first stage \( t_{\text{first}} = t_{\text{single-stage}}(0.6) \), intermediate stage \( t_{\text{intermediate}} = t_{\text{single-stage}}(0.68) \), and last stage \( t_{\text{last}} = t_{\text{single-stage}}(0.6) \) yields the total delay of the interconnect with \( N \) inserted repeaters,

\[ T_{d,\text{total}} = t_{\text{first}} + (N - 2) \times t_{\text{intermediate}} + t_{\text{last}} \]  
\[ (17) \]

Note that the rise and fall delays are not separated since the same single stage model provides the delay for rising and falling transitions, as described in Section 3.

This model is validated against SPICE, as depicted in Figure 3. Although only a subset of the results is provided in the figure, the model exhibits good accuracy across the entire parameter space. For nominal to near threshold voltages, the proposed model exhibits an error between −17\% to 9\%. For near threshold to subthreshold voltages, the proposed model exhibits an error between −15\% to 17\%.

5. REPEATER INSERTION FOR WIDE SUPPLY VOLTAGE RANGE APPLICATIONS

In this section, challenges of the repeater insertion technique for wide supply voltages are discussed. The optimal number of repeaters as a function of supply voltage is described in Section 5.1. The effect on the delay for varying supply voltages for a specific number of repeaters is discussed in Section 5.2. The maximum range of supply voltages considering delay constraints is described in Section 5.3.

5.1. Optimal Number of Repeaters Across a Range of Supply Voltages

The primary challenge of repeater insertion considering a wide supply voltage range is the conflicting number and size of the inserted inverters needed for different voltage levels. The interconnect resistance is not a function of supply voltage, and therefore remains constant. The resistance of a transistor is, however, a strong function of the supply voltage and can range from tens of ohms at nominal voltages for large transistors to mega-ohms for small transistors operating in the subthreshold voltage region. This issue is examined by evaluating (17) over a range of inverter sizes and supply voltages. The optimal number and size of the inserted repeaters enabling minimum total delay are illustrated in Figure 4. A disparity between the optimal number and size of the inserted repeaters is noted. The model provides an accurate approximation of the optimal number of inserted repeaters with a worst case error of two additional repeaters as compared to SPICE. The size of the repeaters provided by the model is also consistent with the results from SPICE, as shown in Figure 4(b).

5.2. Effect on Delay of a Fixed Number of Repeaters

The disparity in the optimal number and size of the repeaters for each supply voltage is demonstrated in Section 5.1. This disparity illustrates a significant design

![Fig. 3. Interconnect delay as a function of interconnect resistance and capacitance as compared to SPICE for (a) nominal to near threshold voltage range, and (b) near threshold to subthreshold voltage range.](image-url)
constraint due to the fixed number of inserted repeaters that cannot change as a function of voltage. When dealing with dynamically changing supply voltages, the interconnect being optimized for a specific operating point exhibits different delay overheads. A significant challenge is to determine the performance state that exhibits the lower delay overhead. The proposed interconnect model of (17) characterizes the delay penalty when optimizing the interconnect for a single supply voltage. From this model, the optimal supply voltage that minimizes the delay overhead can be determined, providing design guidelines for optimizing interconnect operating over a wide range of supply voltages.

As shown in Figure 5, two primary observations can be drawn from this analysis. The minimum delay overhead occurs at the 0.4 volt performance state, with three repeaters with a size multiplier of 140. With this configuration, the average delay overhead across the entire voltage range is 22.7%. Optimizing the low supply voltage provides a smaller delay overhead at high voltages.

5.3. Maximum Supply Voltage Range with Delay Constraint

The resistive and capacitive interconnect should be optimized at lower supply voltages to enable higher frequency of operation across a wide range of supply voltages, as described in Section 5.2. This condition does not hold when an external delay constraint is imposed. The delay expression in Section 5.1 is presented as a contour in Figure 6. In this figure, the relation between the minimum delay optimized at a specific supply voltage and the operating supply voltage is shown. For example, for a delay constraint $T_{\text{delay}} = 0.266$ ns, the corresponding contour line is highlighted in Figure 6. Within the highlighted region, the maximum operating voltage ranges from 0.5 volts to 0.8 volts. This range is enabled by repeater insertion optimized for an operating voltage of 0.5 volts.
6. SUMMARY

A closed-form model of interconnect delay with inserted repeaters supporting a wide supply voltage range in sub-22 nm FinFET technologies is provided in this paper. Utilizing this delay model, design issues relating to repeater insertion over a wide range of supply voltages are also addressed. The conflicting number and size of inserted repeaters required at different supply voltages is examined. To overcome this issue, a method for optimizing the number and size of inserted repeaters across a wide supply voltage range is proposed. The maximum attainable supply voltage range is also provided considering delay constraints.

The proposed delay model is validated against a 14 nm commercial SPICE model. The single stage delay model exhibits an accuracy ranging from $-3.5\%$ to $7.3\%$ for nominal to near threshold voltages. For an extended voltage range that includes subthreshold voltages, the error increases to within $-12\%$ to $15\%$. The interconnect delay model with inserted repeaters exhibits an error ranging between $-17\%$ to $9\%$ for nominal to near threshold voltages. For extended voltage range to subthreshold voltages, the error increases to $-15\%$ to $17\%$. The proposed delay model is suitable to address energy efficiency challenges in modern sub-22 nm FinFET technologies.

Acknowledgments: This research is supported in part by the Binational Science Foundation under Grant No. 2012139, the National Science Foundation under Grant Nos. CCF-1329374, CCF-1526466, and CNS-1548078, IARPA under Grant No. W911NF-14-C-0089, and by grants from Cisco Systems and Intel.

References

2. A. Shapiro and E. G. Friedman, Power efficient level shifter for 16 nm FinFET near threshold circuits. IEEE Transactions on Very Large Scale Integration (VLSI) Systems 24, 774 (2016).

Alexander E. Shapiro

Alexander E. Shapiro was born in Moscow. He received the Bachelor of Science degree in computer engineering from the Technion-Israel Institute of Technology, Haifa, Israel, in 2010, the Master of Science degree in electrical engineering from the University of Rochester, Rochester, New York, in 2012, and the Ph.D. degree in electrical engineering from the University of Rochester, Rochester, New York, in 2016. He is currently with Intel Hillsboro in Oregon. Between 2008 and 2011, he held a variety of software and hardware R&D positions at IBM and Intel in Israel. Alexander was employed as intern in circuits design group at Qualcomm, North Carolina in summer 2013 and in memory IP group at Intel, Oregon in summer 2015. His current research interests include the analysis and design of high performance integrated circuits, low power techniques, and near threshold circuits.
Eby G. Friedman

Eby G. Friedman received the B.S. degree from Lafayette College in 1979, and the M.S. and Ph.D. degrees from the University of California, Irvine, in 1981 and 1989, respectively, all in electrical engineering. From 1979 to 1991, he was with Hughes Aircraft Company, rising to the position of manager of the Signal Processing Design and Test Department, responsible for the design and test of high performance digital and analog IC’s. He has been with the Department of Electrical and Computer Engineering at the University of Rochester since 1991, where he is a Distinguished Professor, and the Director of the High Performance VLSI/IC Design and Analysis Laboratory. He is also a Visiting Professor at the Technion-Israel Institute of Technology. His current research and teaching interests are in high performance synchronous digital and mixed-signal microelectronic design and analysis with application to high speed portable processors and low power wireless communications. He is the author of over 400 papers and book chapters, 12 patents, and the author or editor of 16 books in the fields of high speed and low power CMOS design techniques, 3-D design methodologies, high speed interconnect, and the theory and application of synchronous clock and power distribution networks. Dr. Friedman is the Editor-in-Chief of the Microelectronics Journal, a Member of the editorial boards of the Analog Integrated Circuits and Signal Processing, Journal of Low Power Electronics, and Journal of Low Power Electronics and Applications, Chair of the IEEE Transactions on Very Large Scale Integration (VLSI) Systems steering committee, and a Member of the technical program committee of numerous conferences. He previously was the Editor-in-Chief of the IEEE Transactions on Very Large Scale Integration (VLSI) Systems, the Regional Editor of the Journal of Circuits, Systems and Computers, a Member of the editorial board of the Proceedings of the IEEE, IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing, IEEE Journal on Emerging and Selected Topics in Circuits and Systems, and Journal of Signal Processing Systems, a Member of the Circuits and Systems (CAS) Society Board of Governors, Program and Technical chair of several IEEE conferences, and a recipient of the IEEE Circuits and Systems 2013 Charles A. Desoer Technical Achievement Award, a University of Rochester Graduate Teaching Award, and a College of Engineering Teaching Excellence Award. Dr. Friedman is a Senior Fulbright Fellow and an IEEE Fellow.