INTEGRATION the VLSI journal 58 (2017) 74–81

Contents lists available at ScienceDirect

INTEGRATION, the VLSI journal

journal homepage: www.elsevier.com/locate/vlsi

Scaling equations for the accurate prediction of CMOS device performance from 180 nm to 7 nm

Aaron Stillmaker[a,b,*], Bevan Baas[a]

[a] Department of Electrical and Computer Engineering, University of California, Davis, One Shields Ave., Davis, CA 95616, USA
[b] Department of Electrical and Computer Engineering, California State University, Fresno, 2320 E. San Ramon Ave., Fresno, CA 93740, USA

ARTICLE INFO

Keywords:
Transistor scaling
Deep submicron performance
VLSI design
CMOS device

ABSTRACT

Classical scaling equations which estimate parameters such as circuit delay and energy per operation across technology generations have been extremely useful for predicting performance metrics as well as for comparing designs across fabrication technologies. Unfortunately in the CMOS deep-submicron era, the classical scaling equations are becoming increasingly less accurate and new practical scaling methods are needed. We curve fit second and third-order polynomials to circuit delay, energy, and power dissipation results based on HSpice simulations utilizing the Predictive Technology Model (PTM) and International Technology Roadmap for Semiconductors (ITRS) models. While the classical scaling equations give differences as much as 83×from the predictions of PTM and ITRS models, our predictive polynomial models with table-based coefficients yield a coefficient of determination, or R^2, value of greater than 0.95.

1. Introduction

The observation known as Moore's law [1] states that the number of devices per chip doubles roughly every two years and has held true for decades [2–4]. Until deep-submicron effects became more pronounced, for the most part, transistor characteristics scaled predictably with respect to transistor dimensions and supply voltage. These CMOS scaling metrics that have been in wide use for decades were first proposed by Dennard et al. in 1974 [5] and quantified into generalized scaling equations that took short-channel effects into consideration [5,6]. Performance gains were generated by simply using smaller transistors, with few outside factors [7]. These scaling factors are found in the literature [8,9], and are shown in Table 1 where scaling factor S is the ratio of the transistor dimensions between two transistor sizes and U is the ratio between two voltages. With both S and U, it is expected that all geometry and voltages scale together.

Unfortunately due to features of deep-submicron CMOS technologies such as a variety of short-channel effects and multiple-gate devices, these classical scaling equations have been increasingly inaccurate predictors in recent generations of CMOS technologies [8,9,7]. It is, however, still desirable to compare CMOS circuit performance results between circuits that are fabricated using different transistor sizes and supply voltages, so a new method is presented.

As transistors get smaller, however, short-channel effects and other issues such as process variation start playing a larger role, making the traditional scaling equations inaccurate [4,7,10,11]. Leakage current is affected greatly by gate length, oxide thickness, and threshold voltage, so it is becoming a large issue with deep submicron processes, where these values are small, and getting smaller. With these issues affecting transistor operation, designers started looking to optimize between technology nodes other than simple geometric scaling. Width, length, and oxide thickness are not scaling together, and neither is supply voltage, V_{DD}, and threshold voltage, V_T, which means that scaling factors S and U shown in Table 1 can not be determined. The above mentioned non-regularities were especially noticeable when the industry switched largely to using high-k dielectrics and metal gates with technology nodes at 45 nm and smaller [12] and again when the industry switched to multi-gate (double gate/FinFET or tri-gate) transistors at 20 nm and smaller [7,11,13,14]. This will of course be further complicated in the not so distant future beyond CMOS, when it becomes commercially viable to use different devices for continued performance gains, such as nano-electro-mechanical (NEM) devices [15], carbon nanotube transistors [16], or nanowire transistors [11,17]. The presented modeling method could potentially be used to characterize scaling to these devices, but is not covered in this paper. Equations have been proposed to describe how different performance characteristics scale based on specific aspects, such as length and thickness, while still using the same specific device [18–20], however none of these predict performance scaling from different device types.

This paper presents a method for quickly and accurately determin-

* Corresponding author at: Department of Electrical and Computer Engineering, California State University, Fresno, CA 93740, USA.
E-mail addresses: astillmaker@mail.fresnostate.edu (A. Stillmaker), bbaas@ucdavis.edu (B. Baas).

http://dx.doi.org/10.1016/j.vlsi.2017.02.002
Received 28 May 2016; Received in revised form 10 January 2017; Accepted 9 February 2017
Available online 13 February 2017
0167-9260/ © 2017 Elsevier B.V. All rights reserved.

Table 1
Traditional scaling equations for short-channel devices.
Source: Adapted from Rabaey [8].

Parameter	Relation	Full Scaling	General Scaling	Fixed V Scaling
W, L, t_{ox}		$1/S$	$1/S$	$1/S$
V_{DD}, V_T		$1/S$	$1/U$	1
$Area/Device$	WL	$1/S^2$	$1/S$	$1/S^2$
$Power$	$I_{sat}V$	$1/S^2$	$1/U^2$	1
$IntrinsicDelay$	$R_{on}C_G$	$1/S$	$1/S$	$1/S$
$Energy$	Pt	$1/S^3$	$1/SU^2$	$1/S$

ing a scaling factor of CMOS device performance between different technology nodes, characterized both by different transistor sizes and different device types, without needing to model the entire design using different Spice libraries. To our knowledge there is no current method to accomplish this in the literature.

1.1. Method for accurate scaling in deep submicron technologies

The physics of transistor operations in the submicron region become far more complicated than those in the micron region, with leakage and other above mentioned issues becoming large factors in energy consumption and delay. The most accurate way to get scaling factors in submicron processes is to use a Spice simulation tool, such as HSpice with a model that specifies the characteristics of the particular technology. Simulating a whole design in Spice with modified technology size and voltages would result in the most accurate comparison [8]. While this is an accurate method, it is impossible without the complete extracted design, which makes this an unviable option by which to compare multiple competitive designs. This work presents factors for estimated performance scaling between technology nodes. There is a lack of applicable methods in the literature to predict CMOS circuit design performance in deep submicron technology as it scales between different technology nodes to present and near future technologies without extracted netlists from the target CMOS circuit design.

This work expands upon a preliminary report [21] by using Spice simulation results to model CMOS device performance from different technology nodes to create scaling factors of energy, delay, power, and area between nodes and supply voltages. One of the large motivations for this project was to create the ability to compare different simple digital hardware implementations using a fair metric. A good performance approximation of a device in a certain technology can be achieved using inverters in a chain, with 4 inverters attached to each output, this is known as Fan Out 4, or FO4. A circuit that has a delay and consumption of X number of F04 inverter chains in a certain technology size should have roughly the same X number of FO4 inverter chains in a different technology size [22]. With this in mind, this work sets out to take simulated measurements of FO4 models in a range of different sizes and voltages to obtain approximated scaling factors for power, energy, and delay. These factors can be used to scale performance measurements when comparing digital designs with different fabrication technologies and supply voltages.

2. Background

2.1. International Technology Roadmap for Semiconductors (ITRS)

The International Technology Roadmap for Semiconductors (ITRS) [23] creates reports that predict where semiconductor technology is headed in the next 15 years. These reports are formed by a collaboration of many companies and research institutions. In this work, these reports were used to obtain industry standard technology sizes, and voltages commonly used, as well as general knowledge about transistor changes over the years. Area is also of interest in digital design, so this work evaluates minimum feature sizes, 1 half Metal 1 pitches, and 4 transistor (4 T) logic gate sizes as scaling factors. In this report, when technology process node sizes are referred to, i.e. 180 nm, 45 nm, etc., it is referring to the minimum feature size. Process sizes were generally identified by their smallest feature size, and for a long time, DRAM 1/2 pitch sizes were the smallest, and were therefore used to identify technologies. With new fabrications, this has not been the case, and ITRS discontinued identifying technologies by their minimum feature size. To try to stop confusion, they started to differentiate by using the first year of production. However, in their 2013 report, they started giving “node name” labels to more easily correlate to industry terminology [23]. As minimum feature technology nodes have continued to be the generally accepted term, in this work technologies are identified by both the production year and technology node, as shown in Table 2.

2.2. Predictive technology model (PTM)

The Predictive Technology Models [14,24–27], or PTMs, were used to simulate different performance characteristics as technology size and voltage scaled. The models were developed for designers who do not have access to proprietary transistor characteristics to test designs with future technology nodes. The PTMs are the most accurate models available, as semiconductor companies do not readily provide characteristics of their specific technologies. This lack of specificity of PTM

Table 2
Characteristics of different technology nodes [23]. The modeled measurements are for a single inverter in an FO4 chain. The energy value is the average energy required for a single inverter transition from low to high, or high to low.

				SimulatedPerformance of Inverter		
Production Year	Technology Node (nm)	Technology Type	V_{DD} (V)	Delay (ps)	Energy (fJ)	Power (μW)
1999	180	Bulk	1.8	77.2	27.5	105
2001	130	Bulk	1.2	34.7	5.20	26.1
2004	90	Bulk	1.1	26.5	2.62	13.0
2007	65	Bulk	1.1	19.8	1.72	8.58
2008	45	High-k	1.1	10.9	1.05	5.19
2010	32	High-k	0.97	9.8	0.51	2.47
2012	20	Multi-Gate	0.9	9.66	0.198	1.51
2013	16[a]	Multi-Gate	0.86	6.12	0.179	1.28
2013	14[a]	Multi-Gate	0.86	4.02	0.144	0.995
2015	10	Multi-Gate	0.83	3.24	0.122	0.866
2017	7	Multi-Gate	0.8	2.47	0.111	0.789

[a] The 2013 ITRS report labels a single "16/14" node.

has the added benefit of generality for our purpose of comparing designs across fabrication technologies and manufacturers.

3. HSpice device modeling

HSpice was used to model the scaling results. A fan out four, FO4, inverter chain was used. FO4 delay has been shown to be proportional to CV/I (intrinsic capacitance, voltage, and drive current of a device) [22]. Intrinsic capacitance and drive current are both proportional to device size, so they scale with 1/S and as previously mentioned, voltage scales at factor U, thus, using traditional scaling methods, delay should scale with U/S^2.

The inverters in the model were designed as a 4×minimum size CMOS inverter for the technology node, with a PMOS to NMOS ratio of $\beta = 2$ to keep the rise and fall time roughly balanced. In a multi-gate transistor, the effective channel width is equal to two times the height of the fin plus the width of the fin, or $W_{eff} = 2 \times h_{fin} + W_{fin}$ [14]. After the effective channel width of a single fin was determined, the number of fins in the transistor was modified to achieve a 4×minimum size CMOS inverter with a PMOS/NMOS ratio of $\beta = 2$.

The modeled inverter chain starts with one inverter with the output connected to 4 identical inverters, with that output connected to 16 inverters, and so on until the circuit ends with 64 inverters, creating a total of 4 FO4 stages, as shown in Fig. 1. A square wave was modeled as the input to the inverter chain. The set of 16 inverters in the middle of the chain, were used for the sampling. The delay between when the input signal to the set of 16 inverters crossed the midpoint and the output crossed the midpoint was measured. The voltage was measured along with the current, and calculations were made by Eqs. (1)– (4) where t_0 to t_1 is the transition time as the signal transitions from 10% V_{DD} to 90% V_{DD}, and t_2 to t_3 is the transition time from 90% V_{DD} to 10% V_{DD}. Eq. (1) calculates the average power by integrating the product of the current and voltage from time 0 to time T then dividing by T. Eq. (2) computes the energy for a low (0) to high (1) transition by integrating the product of the current and voltage from when the output voltage was 10% V_{DD} to 90% V_{DD}. Eq. (3) calculates the energy for a high (1) to low (0) transition by integrating the product of the current and voltage from when the output voltage was 90% V_{DD} to 10% V_{DD}. The average energy consumption of a transition is computed in Eq. (4) by summing Eqs. (2) and (3) and dividing the sum by 2.

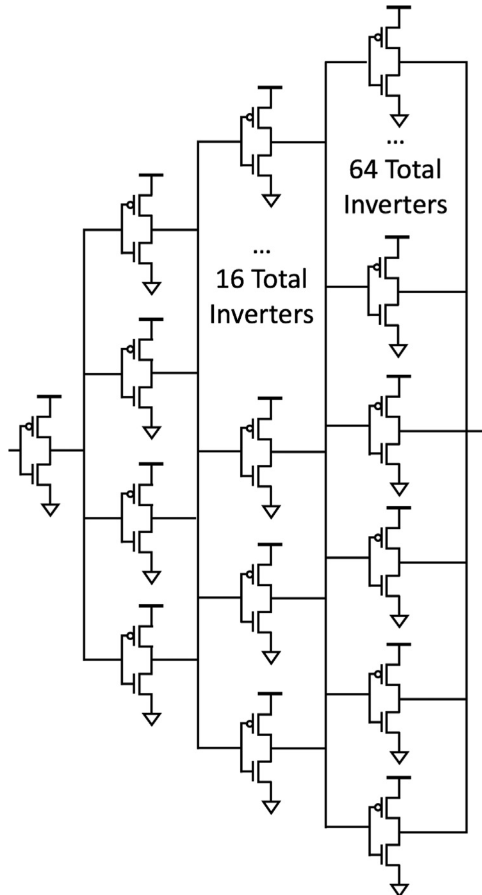

Fig. 1. Inverters connected into an FO4 chain were used to measure delay, energy, and power.

$$P_{ave} = \frac{1}{T}\int_0^T I(t)\cdot Vdt \quad (1)$$

$$E_{0\to 1} = \int_{t_0}^{t_1} I(t)\cdot Vdt \quad (2)$$

$$E_{1\to 0} = \int_{t_2}^{t_3} I(t)\cdot Vdt \quad (3)$$

$$E_{ave} = \frac{E_{0\to 1} + E_{1\to 0}}{2} \quad (4)$$

3.1. Simulation

The simulations were run on the following technology sizes: 180 nm, 90 nm, 65 nm, 45 nm, 32 nm, 20 nm, 16 nm, 14 nm, 10 nm, and 7 nm with supply voltages varying from 1.8V to 0.5 V in 0.05 V increments. Technology nodes are not designed to handle voltages much higher or lower than their target voltages, so even though HSpice gave results for the technology nodes operating at non-expected voltages, they were removed from the results as the PTM characteristics are not expected to hold for these values.

With the industry standard of high-k dielectric transistors at 45 nm and below, high-k PTM models, both for high performance (HP) and low power (LP), are used for the 45 nm and 32 nm nodes. High performance transistors are generally designed with lower threshold voltages, which allows for faster switching times, at the expense of leakage power. Low power transistors are generally the opposite, with high threshold voltages, which gives lower power consumption, especially while in standby, with low leakage. Also, as the industry standard for 20 nm and below is multi-gate transistors, the PTM models for both HP and low standby power (LSTP) are evaluated for these devices between 20 nm and 7 nm. Low standby power multi-gate transistors, similar to low power transistors, target lower power, at the cost of propagation delay.

As transistors become smaller, interconnect parasitics: resistance, capacitance, and inductance, are making a larger impact on total device performance [8,28]. Transistors are able to switch faster, while wires are getting smaller and closer together, which slows down the propagation of signals across wires [29]. However, the magnitude of these affects are largely determined by fabrication or design specific factors, such as how long a signal must travel on a wire, the number of contact vias between metal layers and their size, how close wires are together, and wire dimensions. If one wishes to determine the wire parasitic effects on their design they would need to extract the resistance, capacitance, and inductance values of their specific design using a design kit post layout and simulate their effect. Therefore, it was determined to be impractical to include a factor that can have so much undeterminable variance, so wire loads were not included. For larger technologies, and smaller designs, this will affect the factor less but it should be considered when using these scaling factors. As HSpice models are created in a simulated environment there are other effects, such as process variation, voltage fluctuation, and temperature effects, which are not taken into account.

Table 3
Geometric sizes of different technology nodes which affect area from ITRS reports [23].

Minimum FeatureSize (nm)	Metal I HalfPitch (nm)	(4 T) Logic GateSize (μm^2)
180	230	57
130	150	10.4
90	90	5.2
65	68	2.6
45	59	2.1
32	45	0.71
20	32	0.35
16/14	40	0.248
10	31.8	0.157
7	25.3	0.099

4. Simulation results and scaling factors

Table 2 shows the standard values labeled by ITRS at each technology node investigated, as well as the delay, energy, and power simulated using the inverter chain in HSpice as described in Section 3. The V_{DD} is taken from the ITRS tables for high-performance.

4.1. Area

To determine a factor for scaling area between technologies, minimum feature sizes, Metal 1 half pitches, and 4 T logic gate of MPU (High-volume Microprocessor) were taken from the ITRS reports, with details given in Table 3. The geometric characteristics from Table 3 with a single length dimension, minimum feature size and metal 1 pitch, were squared to get an area value. To combine all of these factors, two with units of mm^2 and one of mm^2, with equal weight given to each figure of merit so as to attain a useable scaling factor, the geometric mean was computed. Table 4 displays the scaling factors using the geometric mean of the area factors presented in Table 3. To scale area, determine the scaling factor by finding the intersection of the starting technology node and desired technology node row. Multiply that number by the starting chip's area to determine an equivalent area in the desired technology node. The closest scaling to the traditional scaling factors in Table 1 would be for a scaling factor S using the minimum feature size. An exact scaling would be dependent on the design, but using either of the aforementioned values should give a good estimate, especially for simple designs. So as to plot the different area figures of merit on a single graph, each of the values were normalized to their 7 nm node size. This relative scaling of area data is shown in Fig. 2.

4.2. Delay, energy, and power scaling factors

The results from the HSpice simulation of the inverter chain are given in Figs. 3–5. Fig. 3 plots the average propagation time of a single inverter in the middle of an FO4 inverter chain. Fig. 4 plots the average energy required for a state change of this single inverter. Fig. 5 plots the average power consumption of the single inverter over an entire 1000 ps clock period. This takes into account the increased leakage of the smaller technology nodes. Figs. 3–5 show the dichotomy between the different fabrication technologies. While inside of a specific transistor technology type, such as HP MG, one can see the sizing scales predictively, however there is a large non-linear relationship when comparing across different types, such as HP MG compared to HP High-k. This shows why a simple scaling equation is not a viable option to compare performance data from differing modern CMOS designs.

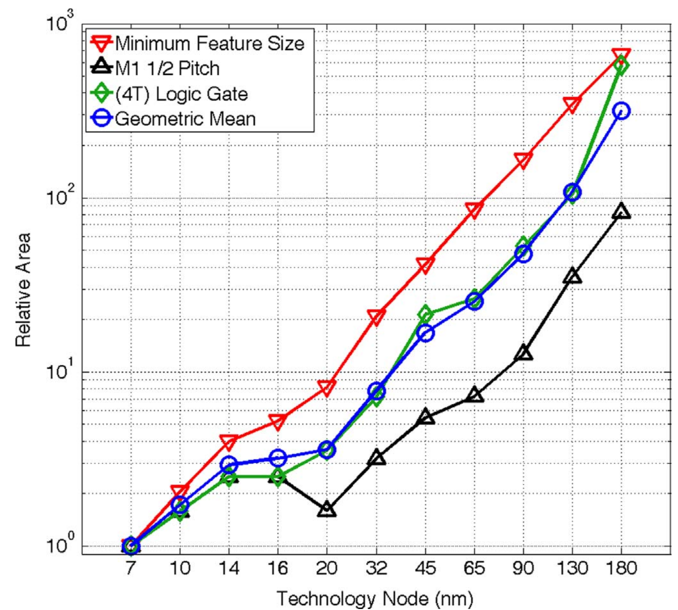

Fig. 2. Relative area scaling of different area sizes over different technology nodes, and the geometric mean of all three of the sizes from Table 3. The minimum feature size and metal 1 pitch values were squared to get an area number, and each area value was normalized to the 7 nm node.

To compare against traditional scaling methods, Figs. 6–8 use the nominal supply voltage values given in Table 2 to plot the modeled data of the major technology nodes. Using the traditional scaling equations in Table 1, the traditional scaling methods are used both scaling from the 180 nm node and the 7 nm node to the other technology nodes. The traditional methods fail by a considerable margin in all but scaling power values from 7 nm, as shown in Fig. 8.

As printing a data table containing all of the data points from the multitude of HSpice simulations would be prohibitive, polynomial approximations for each of the performance factors, i.e. the modeled

Table 4
Area scaling factors using geometric mean of area values given by the three sizes from Table 3.

		Starting Node										
		180 nm	130 nm	90 nm	65 nm	45 nm	32 nm	20 nm	16 nm	14 nm	10 nm	7 nm
Desired Node	180 nm	1	0.34	0.15	0.08	0.053	0.025	0.011	0.01	0.0093	0.0055	0.0032
	130 nm	2.9	1	0.44	0.23	0.16	0.072	0.033	0.03	0.027	0.016	0.0092
	90 nm	6.6	2.3	1	0.53	0.35	0.16	0.075	0.067	0.061	0.036	0.021
	65 nm	12	4.3	1.9	1	0.66	0.31	0.14	0.13	0.12	0.068	0.039
	45 nm	19	6.4	2.8	1.5	1	0.46	0.21	0.19	0.17	0.1	0.059
	32 nm	40	14	6.1	3.3	2.2	1	0.46	0.41	0.38	0.22	0.13
	20 nm	88	30	13	7.1	4.7	2.2	1	0.89	0.82	0.48	0.28
	16 nm	99	34	15	7.9	5.3	2.4	1.1	1	0.91	0.54	0.31
	14 nm	110	37	16	8.7	5.8	2.7	1.2	1.1	1	0.59	0.34
	10 nm	180	63	28	15	9.8	4.5	2.1	1.9	1.7	1	0.58
	7 nm	320	110	48	25	17	7.8	3.6	3.2	2.9	1.7	1

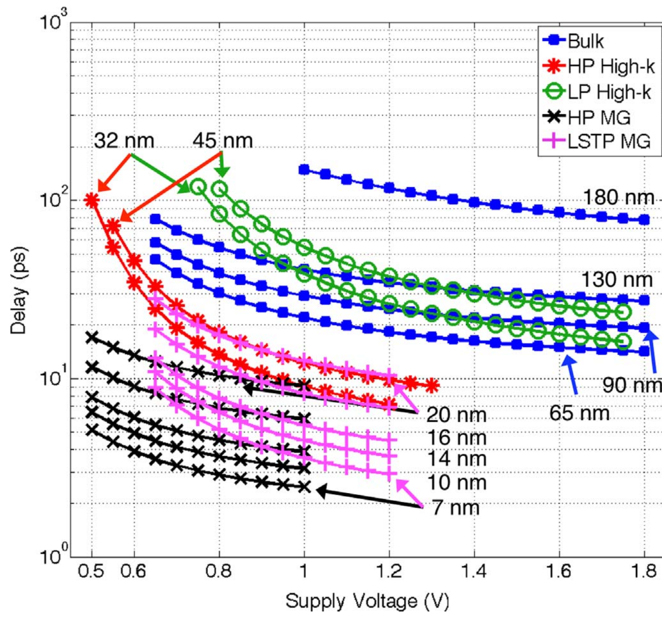

Fig. 3. Delay for a signal to propagate through one inverter in the middle of the FO4 inverter chain for different technologies: bulk, high performance (HP) high-k, low power (LP) high-k, high performance multi-gate (HP MG), and low standby power multi-gate (LSTP MG) nodes with scaling voltage. '*Interactive plot Delay1.csv and Delay2.csv here*'.

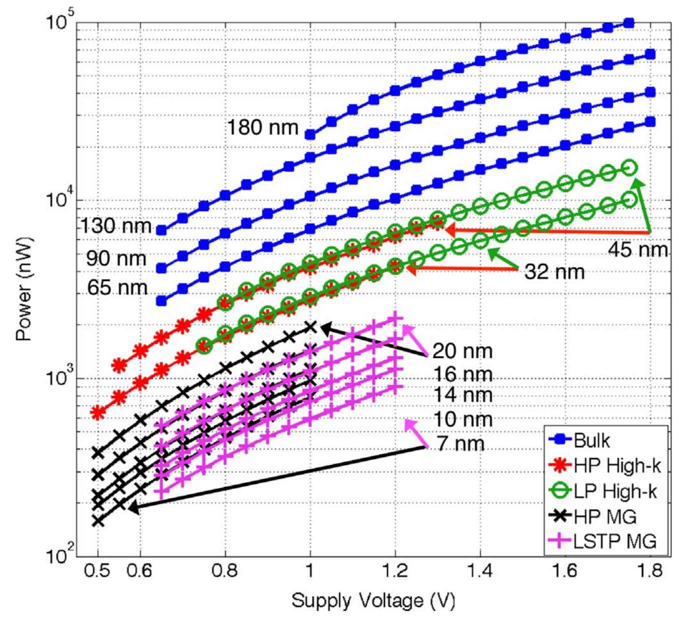

Fig. 5. Average power signal for a clock cycle in which a signal is propagated through one inverter in the middle of the FO4 inverter chain for different technologies: bulk, high performance (HP) high-k, low power (LP) high-k, high performance multi-gate (HP MG), and low standby power multi-gate (LSTP MG) nodes with scaling voltage with a clock frequency of 1 MHz. '*Interactive plot Power1.csv and Power2.csv here*'.

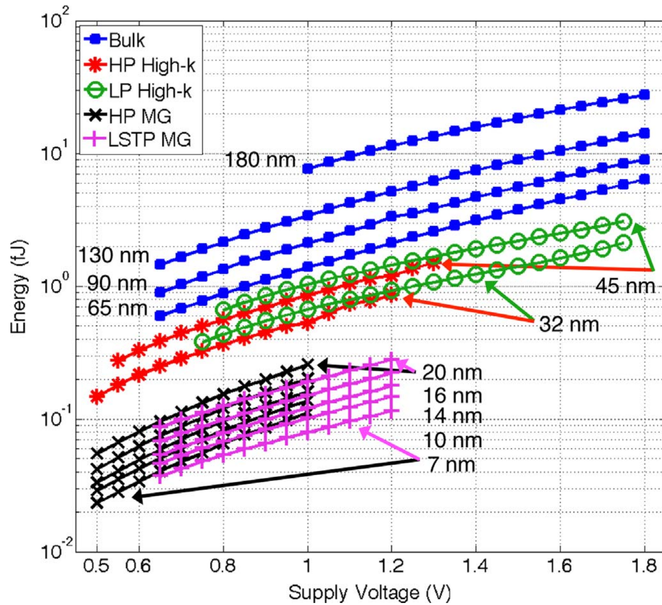

Fig. 4. Energy required for one inverter in the middle of the FO4 inverter chain to toggle for different technologies: bulk, high performance (HP) high-k, low power (LP) high-k, high performance multi-gate (HP MG), and low standby power multi-gate (LSTP MG) nodes with scaling voltage. '*Interactive plot Energy1.csv and Energy2.csv here*'.

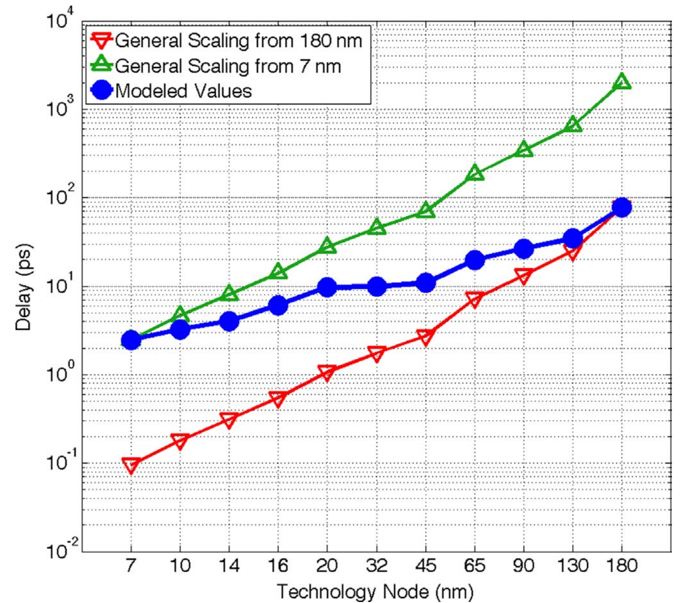

Fig. 6. Delay simulated for a signal to propagate through one inverter in the middle of the FO4 inverter chain using Table 2 values and scaled using Table 1 equations.

delay, energy, and power associated with a particular technology node and supply voltage, are generated for ease of use, without loss of accuracy. The polynomial approximations were created using a script that iteratively increased the order of the polynomial until a coefficient of determination, or R^2, value of greater than 0.95 was attained. This resulted in a third-order polynomial for the delay factor approximations, and second-order polynomials for energy and power factor approximations. The values attained using the polynomial approximations are indeed so close to the original, if plotted on Figs. 3, 4 and 5 they would completely cover the measured data from HSpice, as they are visually indistinguishable. Similarly, if the polynomial approximated values were plotted on Figs. 6–8 they would completely cover the "Modeled Values" plot lines.

Eqs. (5)–(7) are used to determine a *DelayFactor*, *EnergyFactor*, and *PowerFactor*, respectively, for a specific technology node and voltage.

$$DelayFactor = a_{d3}V^3 + a_{d2}V^2 + a_{d1}V + a_{d0} \quad (5)$$

$$EnergyFactor = a_{e2}V^2 + a_{e1}V + a_{e0} \quad (6)$$

$$PowerFactor = a_{p2}V^2 + a_{p1}V + a_{p0} \quad (7)$$

The coefficients for the above equations corresponding to each technology node can be found in Table 5. For simplicity, all coefficients were rounded to four significant figures, which did not significantly effect the coefficient of determination. This level of accuracy is more

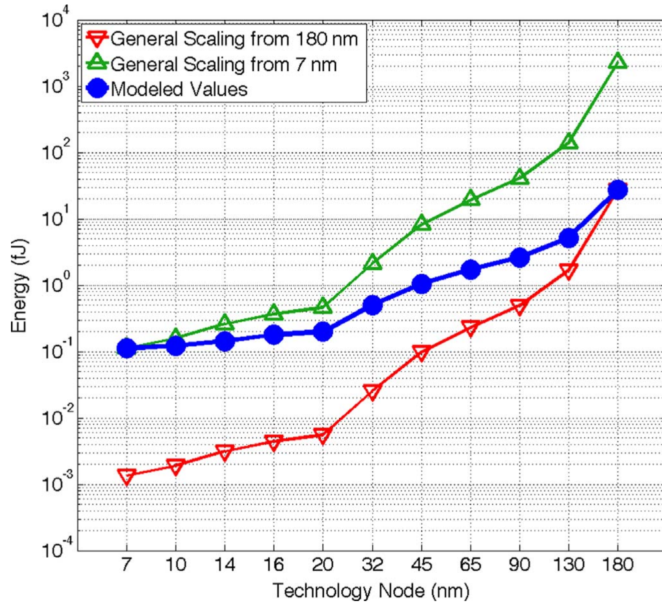

Fig. 7. Energy used to toggle one inverter in the middle of the FO4 inverter chain simulated using Table 2 values and scaled using Table 1 equations.

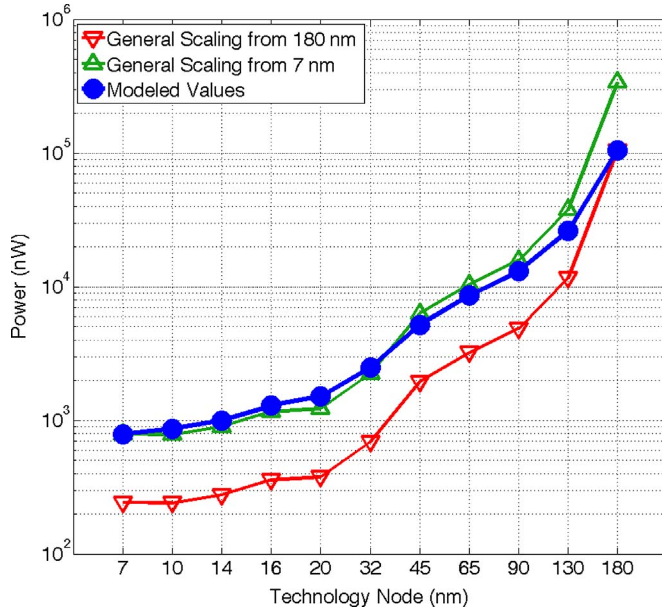

Fig. 8. Average power for a clock cycle simulated for a signal to propagate through one inverter in the middle of the FO4 inverter chain with a clock frequency of 1 MHz, using Table 2 values and scaled using Table 1 equations.

than sufficient given the imprecision inherent in technology scaling without full design knowledge. It is not recommended to use these factors for supply voltages without a corresponding data point in Figs. 3–5 for a particular technology node, as those voltages are outside of the normally operating voltages of that particular technology node.

Eqs. (8)–(10) may be used to scale delay, energy, and power, respectively, between technology nodes.

$$D_x = \frac{DelayFactor_x}{DelayFactor_y} \cdot D_y \tag{8}$$

$$E_x = \frac{EnergyFactor_x}{EnergyFactor_y} \cdot E_y \tag{9}$$

$$P_x = \frac{PowerFactor_x}{PowerFactor_y} \cdot P_y \tag{10}$$

As the modeled power values are an average over a 1000 ps clock period, it largely displays standby power for each node. If one wishes to scale both the operating frequency (governed by delay) and the technology node, they should use Eq. (11), which takes both changing values into account.

$$P_x = \frac{EnergyFactor_x \cdot DelayFactor_y}{EnergyFactor_y \cdot DelayFactor_x} \cdot P_y \tag{11}$$

In Eqs. (8)–(11), subscript x refers to the desired node, while y refers to the starting node. The factors *DelayFactor*, *EnergyFactor*, and *PowerFactor* are obtained from Eqs. (5)–(7), respectively. If this paper is viewed online, one can alternatively use the interactive plot to attain values for *DelayFactor*, *EnergyFactor*, and *PowerFactor* to be used in Eqs. (8)– (11).

4.2.1. Scaling example

The following example is given to illustrate the scaling procedure. To scale an example energy value of 1 pJ/Op from 1.3 V in 65 nm to 0.9 V in HP 32 nm, Eqs. (6) and (9) are used, as shown in Eqs. (12)–(14).

$$EnergyFactor_x = 0.5654(0.9)^2 - 0.2962(0.9) + 0.1148 \tag{12}$$

(Eq. 6 and Table 5)

$$EnergyFactor_x = 0.3062$$

$$EnergyFactor_y = 2.441(1.3)^2 - 2.831(1.3) + 1.276 \tag{13}$$

(Eq. 6 and Table 5)

$$EnergyFactor_y = 1.721$$

$$E_x = \frac{0.3062}{1.721} \cdot 1 \text{ pJ/Op} \tag{14}$$

(Eq. 9, 12 and 13)

$$E_x = \mathbf{0.1779 pJ/Op}$$

The resulting E_x is the new energy value, scaled to an approximation of its performance, if the example chip had been fabricated in 0.9 V in HP 32 nm.

5. Conclusion

This work presents a method and data sets from simulation that can be used to scale transistors to different technology nodes in a fair method. The data presented shows that traditional scaling methods do not hold into these submicron transistors, especially with the advent of radically changed devices. The general trend is similar, but does not make an accurate comparison, as illustrated by our gathered data that shows up to an 83× difference from measured values. Thus the method of using the modeled simulation data presented in this work is a more accurate estimation that can be used to compare two devices from different technologies and supply voltages. As models of more advanced technology nodes become available, this presented method could be used to add scaling data to and from these new nodes by simulating performance data in the same fashion presented and recreating a polynomial curve to make the scaling factors easily attainable.

Acknowledgements

The authors acknowledge Zhibin Xiao and Jon Pimentel and gratefully acknowledge support from ST Microelectronics, C2S2 Grant 2047.002.014, NSF Grant 0430090 and CAREER Award

Table 5
The polynomial coefficient values to be used with Eqs. (5)–(7) to attain the factors to be used to generate the scaling factor between two technology nodes and voltages.

Type		Node	Delay Coefficients (Eq. (5))				Energy Coefficients (Eq. (6))			Power Coefficients (Eq. (7))		
			a_{d3}	a_{d2}	a_{d1}	a_{d0}	a_{e2}	a_{e1}	a_{e0}	a_{p2}	a_{p1}	a_{p0}
Bulk		180 nm	–	97.09	−356.7	406.5	–	24.64	−17.98	–	101000	−79720
		130 nm	−76.65	334.9	−493.4	275.8	7.171	−6.709	2.904	27020	−15450	5630
		90 nm	−60.34	262.5	−384.2	210.9	4.762	−4.781	2.092	17320	−11230	4328
		65 nm	−53.3	230.4	−333.9	178.6	3.755	−4.398	1.975	12890	−10510	4362
High-k	HP	45 nm	−501.6	1567	−1619	566.1	1.018	−0.3107	0.1539	5462	−1760	522.4
		32 nm	−1047	2982	−2797	873.5	0.8367	−0.4341	0.1701	4001	−1733	533.6
	LP	45 nm	−285.7	1239	−1795	898.8	1.103	−0.362	0.2767	6297	−3009	1124
		32 nm	−325.9	1374	−1922	913.2	0.9559	−0.7823	0.471	4557	−3037	1323
Multi-Gate	HP	20 nm	–	34.63	−66.37	41.15	0.373	−0.1582	0.04104	2922	−1286	299.9
		16 nm	–	24.8	−47.52	28.87	0.2958	−0.1241	0.03024	2133	−882.6	197.7
		14 nm	−40.66	109.2	−100.6	35.92	0.2363	−0.09675	0.02239	1675	−711	159
		10 nm	−34.95	93.65	−85.99	30.4	0.2068	−0.09311	0.02375	1456	−621.6	143.8
		7 nm	−28.58	76.6	−70.26	24.69	0.1776	−0.09097	0.02447	1179	−515.7	123.4
	LSTP	20 nm	−160.5	514.1	−558.6	217.5	0.2632	−0.14	0.06841	2096	−962.4	287.1
		16 nm	−114.6	366.7	−397.4	153.6	0.2139	−0.1187	0.05639	1609	−715.5	205.7
		14 nm	−85.37	271.6	−292.2	111.4	0.1556	−0.06472	0.03066	1259	−554.1	152.3
		10 nm	−71.76	228.6	−246.3	93.91	0.1261	−0.0518	0.02769	1046	−422.7	118.9
		7 nm	−61.79	196.1	−210.3	79.55	0.09365	−0.03409	0.02043	815.2	−307.3	87.54

0546907, SRC GRC Grant 1598 and CSR Grant 1659, Intel, UC Micro, Intellasys, SEM, and a UCD Faculty Research Grant.

Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.vlsi.2017.02.002.

References

[1] G.E. Moore, Cramming more components onto integrated circuits, Electronics 38 (1965) 114–117. http://dx.doi.org/10.1109/JPROC.1998.658762.

[2] M. Horowitz, Computing's energy problem (and what we can do about it), in: Proceedings of the International Solid-State Circuits Conference, 2014, pp. 10–14. http://dx.doi.org/10.1109/ISSCC.2014.6757323.

[3] S.G. Narendra, L.C. Fujino, K.C. Smith, Through the looking glass: the 2015 edition: trends in solid-state circuits from isscc, in: IEEE Solid-State Circuits Magazine, Vol. 7, 2015, pp. 14–24. http://dx.doi.org/10.1109/MSSC.2014.2375071.

[4] S. Thompson, R. Chau, T. Ghani, K. Mistry, S. Tyagi, M. Bohr, In search of "forever," continued transistor scaling one new material at a time, IEEE Trans. Semicond. Manuf. 18 (1) (2005) 26–36. http://dx.doi.org/10.1109/TSM.2004.841816.

[5] R. Dennard, V. Rideout, E. Bassous, A. LeBlanc, Design of ion-implanted MOSFET's with very small physical dimensions, IEEE J. Solid-State Circuits 9 (5) (1974) 256–268. http://dx.doi.org/10.1109/JSSC.1974.1050511.

[6] G. Baccarani, M. Wordeman, R. Dennard, Generalized scaling theory and its application to a 1/4 micrometer MOSFET design, IEEE Trans. Electron Devices 31 (4) (1984) 452–462. http://dx.doi.org/10.1109/T-ED.1984.21550.

[7] M. Bohr, The evolution of scaling from the homogeneous era to the heterogeneous era, in: Proceedings of the Electron Devices Meeting (IEDM), IEEE International, 2011, pp. 1.1.1–1.1.6. http://dx.doi.org/10.1109/IEDM.2011.6131469.

[8] J.M. Rabaey, A.P. Chandrakasan, B. Nikolic, Digital Integrated Circuits: A Design Perspective, 2nd edition, Pearson Education, Upper Saddle River, NJ, 2003.

[9] J.P. Uyemura, Introduction to VLSI Circuits and Systems, 1st edition, John Wiley & Sons, Inc., Hoboken, NJ, 2002.

[10] K. Kuhn, CMOS transistor scaling past 32 nm and implications on variation, in: Proceedings of the Advanced Semiconductor Manufacturing Conference (ASMC), IEEE/SEMI, 2010, pp. 241–246. http://dx.doi.org/10.1109/ASMC.2010.5551461.

[11] K. Kuhn, Considerations for ultimate CMOS scaling, IEEE Trans. Electron Devices 59 (7) (2012) 1813–1828. http://dx.doi.org/10.1109/TED.2012.2193129.

[12] K. Mistry, et al., A 45 nm logic technology with high-k+metal gate transistors, strained silicon, 9 cu interconnect layers, 193 nm dry patterning, and 100 pb-free packaging, in: Proceedings of the Electron Devices Meeting, IEDM, IEEE International, 2007, pp. 247–250. http://dx.doi.org/10.1109/IEDM.2007.4418914.

[13] M. Jurczak, N. Collaert, A. Veloso, T. Hoffmann, S. Biesemans, Review of FinFET technology, in: Proceedings of the SOI Conference, IEEE International, 2009, pp. 1–4. http://dx.doi.org/10.1109/SOI.2009.5318794.

[14] S. Sinha, G. Yeric, V. Chandra, B. Cline, Y. Cao, Exploring sub-20 nm FinFET design with predictive technology models, in: Proceedings of the 49th Annual Design Automation Conference, DAC '12, ACM, New York, NY, USA, 2012, pp. 283–288.

[15] F. Chen, H. Kam, D. Markovic, T.-J. K. Liu, V. Stojanovic, E. Alon, Integrated circuit design with NEM relays, in: Proceedings of the IEEE/ACM International Conference on Computer-Aided Design, ICCAD, 2008, pp. 750–757.http://dx.doi.org/10.1109/ICCAD.2008.4681660.

[16] A.D. Franklin, M. Luisier, S.-J. Han, G. Tulevski, C.M. Breslin, L. Gignac, M.S. Lundstrom, W. Haensch, Sub-10 nm carbon nanotube transistor, Nano Lett. 12 (2) (2012) 758–762. http://dx.doi.org/10.1021/nl203701g.

[17] N. Singh, A. Agarwal, L. Bera, T. Liow, R. Yang, S. Rustagi, C. Tung, R. Kumar, G. Lo, N. Balasubramanian, D.-L. Kwong, High-performance fully depleted silicon nanowire (diameter ≤5 nm) gate-all-around CMOS devices, IEEE Electron Device Lett. 27 (5) (2006) 383–386. http://dx.doi.org/10.1109/LED.2006.873381.

[18] S.-H. Oh, D. Monroe, J.M. Hergenrother, Analytic description of short-channel effects in fully-depleted double-gate and cylindrical, surrounding-gate MOSFETs, IEEE Electron Device Lett. 21 (9) (2000) 445–447. http://dx.doi.org/10.1109/55.863106.

[19] T.K. Chiang, A novel scaling theory for fully depleted, multiple-gate MOSFET, including effective number of gates (ENGs), IEEE Trans. Electron Devices 61 (2) (2014) 631–633. http://dx.doi.org/10.1109/TED.2013.2294192.

[20] T.K. Chiang, A new subthreshold current model for functionless trigate MOSFETs to examine interface-trapped charge effects, IEEE Trans. Electron Devices 62 (9) (2015) 2745–2750. http://dx.doi.org/10.1109/TED.2015.2456040.

[21] A. Stillmaker, Z. Xiao, B. Baas, Toward more accurate scaling estimates of CMOS circuits from 180 nm to 22 nm, Tech. Rep. ECE-VCL-2011-4, VLSI Computation Lab, University of California, Davis, Dec. 2011.

[22] FO4 writeup: International technology roadmap for semiconductors 2003 edition, Tech. rep., ITRS, 2002.

[23] International technology roadmap for semiconductors, [Online]. Avaliable: http://www.itrs.net/, Oct. 2015.

[24] A. Balijepalli, S. Sinha, Y. Cao, Compact modeling of carbon nanotube transistor for early stage process-design exploration, in: Proceedings of the ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED), 2007, pp. 2–7.

[25] W. Zhao, Y. Cao, New generation of predictive technology model for sub-45 nm early design exploration, IEEE Trans. Electron Devices 53 (11) (2006) 2816–2823. http://dx.doi.org/10.1109/ISQED.2006.91.

[26] Y. Cao, T. Sato, D. Sylvester, M. Orshansky, C. Hu, New paradigm of predictive MOSFET and interconnect modeling for early circuit design, in: CICC, 2000, pp. 201–204.

[27] Predictive Technology Model, [Online]. Avaliable: http://ptm.asu.edu/ , Oct. 2015.

[28] J. Warnock, Circuit design challenges at the 14 nm technology node,

in: Proceedings of the 48th Design Automation Conference, DAC, ACM, New York, NY, USA, 2011, pp. 464–467.http://dx.doi.org/10.1145/2024724.2024833.
[29] N.H.E. Weste, D.M. Harris, CMOS VLSI Design: A Circuits and Systems Perspective, 4th edition, Addison-Wesley, 2011.

Aaron Stillmaker received the B.S. degree in computer engineering from the California State University, Fresno in 2008, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of California, Davis in 2013 and 2015, respectively. From 2008 to 2015 he was a Graduate Student Researcher in the VLSI Computation Laboratory at the University of California, Davis. In 2013 he interned with the Circuit Research Lab, Intel Labs in Hillsboro, OR. In 2017 he became an Assistant Professor in the Electrical and Computer Engineering Department at California State University, Fresno. His research interests include many-core processor architecture and physical design, many-core algorithms, and digital VLSI design.

Bevan Baas received the B.S. degree in electronic engineering from California Polytechnic State University, San Luis Obispo, in 1987, and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, in 1990 and 1999, respectively. From 1987 to 1989, he was with Hewlett-Packard, Cupertino, CA. In 1999, he joined Atheros Communications, Santa Clara, CA. In 2003 he joined the Department of Electrical and Computer Engineering at the University of California, Davis, where he is currently a Professor. He leads projects in architecture, hardware, software tools, and applications for VLSI computation with an emphasis on DSP workloads.