Parasitic Capacitances: Analytical Models and Impact on Circuit-Level Performance

Lan Wei, *Member, IEEE*, Frédéric Boeuf, *Member, IEEE*, Thomas Skotnicki, *Fellow, IEEE*, H.-S. Philip Wong, *Fellow, IEEE*

***Abstract*— Parasitic capacitances have become a main issue for advanced technology nodes. In this paper, we develop analytical models for parasitic capacitance components for several device structures, including bulk devices, fully depleted silicon-on-insulator (FDSOI) devices, and double gate (DG) devices. With these models, we analyze the impact of parasitic capacitances on the circuit-level performance for logic applications. Si CMOS roadmap projection is revisited beyond 32nm technology, with different device design scenarios examined.**

***Index Terms*— parasitic capacitances, CMOS, bulk, FDSOI, DG, circuit-level performance.**

I. Introduction

HISTORICALLY, the delay of CMOS devices is benchmarked by the intrinsic delay $\tau_{int}=C_{gc}V_{dd}/I_{dsat}$, where C_{gc} is the intrinsic gate-to-channel capacitance, V_{dd} is the supply voltage, and I_{dsat} is saturation current when the device is biased at $V_{gs}=V_{ds}=V_{dd}$. In the long channel regime, parasitic capacitances are negligible compared with C_{gc}; thus the device-level intrinsic delay is a good indicator for the circuit-level speed for CMOS logic applications. According to Dennard's scaling theory **[1]**, when all three dimension of the device and the supply voltage are simultaneously scaled by $1/\kappa$, τ_{int} is scaled by $1/\kappa$, which means the circuit built by scaled devices can run κ times faster. Dennard scaling has been the technology engine for many generations down to sub-100nm regime. Technology boosters such as strain **[2][3]**, high-k/metal gate**[4][5]** have helped the continuation of the historic performance trend down to the 45 nm technology node. As device physical gate length is reduced below 30 nm, gate length scaling becomes less effective, because the shorter gate lengths must be traded off against various leakage (subthreshold, gate, BTBT) currents **[6]-[11]**. Meanwhile, the increasing contribution of parasitic capacitances and series resistances has become a challenge since 90nm technology node. In particular, parasitic capacitances **[10]-[12]**, which are also charged/discharged during the device switching, are no longer negligible, and significantly affect the circuit-level speed.

In this paper, we first introduced simple analytical capacitance models for bulk CMOS, fully-depleted silicon-on-insulator (FDSOI) devices, and planar double gate (DG) devices in Section II. Then, we study which device-level capacitance components affect circuit-level delay performance in Section III and develop an expression of the actual average load capacitance for circuit-level delay estimation, including both intrinsic and parasitic capacitances. Furthermore, the capacitance models are implemented into a fast device and circuit simulator, the Model for Assessment of cmoS Technologies And Roadmaps (MASTAR) **[13]**. With MASTAR, the Si CMOS roadmap projection beyond 32nm technology with a variety of device design scenarios is revisited in Section IV.

II. Analytical Models for Capacitance Components

Device capacitances are split into different components according to their physical origins, including gate-to-channel capacitance (C_{gc}), gate-to-substrate capacitance (C_{gb}), overlap capacitance (C_{ov}), inner fringe capacitance (C_{if}), outer-fringe capacitance (C_{of}), gate-to-plug capacitance (C_{PCCA}), corner capacitance (C_{corner}), and junction capacitance (C_j). Simple analytical models of each component are developed for bulk CMOS device (Figure 1a), FDSOI devices (Figure 1b), and DG devices (Figure 1c), as described in the following paragraphs. Figure 1d shows the capacitance elements for the bulk device. Models of C_{ov}, C_{if}, C_{of}, C_{PCCA}, C_{corner}, and C_j are for one side (source or drain). L_g is the channel length. ΔL is the gate to source/drain overlap length in total. W is the gate width. W_{ext} is the width of gate-overhang per side. L_{pitch} is the contacted gate pitch. L_{PCCA} is the distance between contact plug to gate edge. t_{ox} and t_{ox_el} are the physical and electrical gate dielectric thickness, respectively. t_{gate} is the metal gate thickness. t_{si} is the silicon body thickness in

Manuscript received September 27, 2010. This work was partially supported by C2S2 Focus Center, one of six research centers funded under the Focus Center Research Program (FCRP), a Semiconductor Research Corporation subsidiary, and the member companies of the Stanford Initiative for Nanoscale Materials and Processes (INMP). L. Wei is additionally supported by the Stanford Graduate Fellowship.

L. Wei was with Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA. She is now with Microsystems Technology Laboratories, Massachusetts Institute of Technology, Cambridge 02139 USA (email: lanwei@mit.edu, phone: +1-617-253-0732).

Frédéric Boeuf and Thomas Skotnicki are with STMicroelectronics, Crolles 38960 France (email: frederic.boeuf@st.com, thomas.skotnicki@st.com).

H.-S. Philip Wong is with Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA (email: hspwong@stanford.edu).

FDSOI and DG devices. t_{box} is the thickness of body insulator layer in FDSOI devices. t_{rSD} is the thickness of raised-source/drain in FDSOI devices. x_j is the source/drain junction depth in bulk devices. ε_{ox}, ε_{cap}, ε_{Si} are the permittivity of gate dielectric, capping layer, and silicon body, respectively.

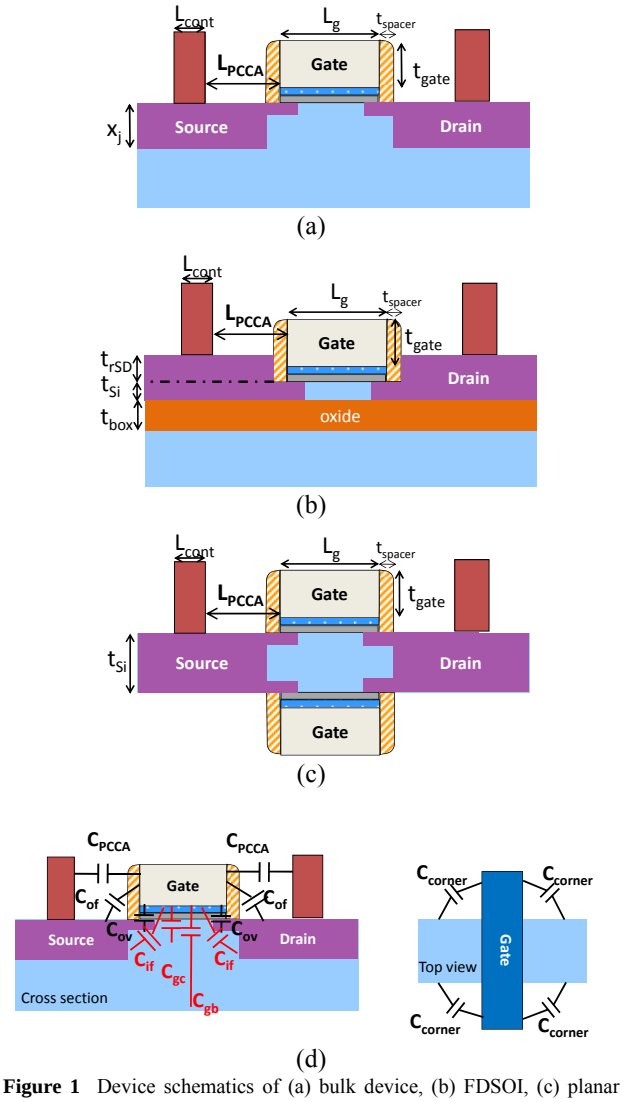


Figure 1 Device schematics of (a) bulk device, (b) FDSOI, (c) planar double gate, and (d) capacitance elements for a bulk device.

A. Gate-to-channel capacitance (C_{gc}), overlap capacitance (C_{ov}), and gate-to-substrate capacitance (C_{gb})

C_{gc}, C_{ov}, and C_{gb} are modeled as parallel plate capacitances. C_{gc} only appears when the device is turned on, when an inversion layer is formed and shield the substrate from the gate. Electrical gate dielectric thickness (t_{ox_el}) is adopted for C_{gc} to include the quantization effect. While the device is in off-state, the gate is electrically coupled to the substrate by C_{gb}. For double-gate devices, the middle of the silicon body is considered as a virtual ground. However, no actual charge goes into or out of the undoped silicon body during device switching. Thus, C_{gb_DG}=0.

$$C_{gc\_bulk} = C_{gc\_SOI} = 0.5C_{gc\_DG} = \frac{\varepsilon_{ox}}{t_{ox\_el}}\left(L_g - \Delta L\right)W. \quad (1)$$

$$C_{ov\_bulk} = C_{ov\_SOI} = 0.5C_{ov\_DG} = 0.5\frac{\varepsilon_{ox}}{t_{ox\_el}}\Delta L \cdot W \quad (2)$$

$$C_{gb\_bulk} = \left\{\left[\frac{\varepsilon_{ox}}{t_{ox}}(L-\Delta L)W\right]^{-1} + \left[\frac{\varepsilon_{si}}{t_{dep}}(L-\Delta L)W\right]^{-1}\right\}^{-1},$$

$$C_{gb\_SOI} = \left\{\begin{aligned}&\left[\frac{\varepsilon_{ox}}{t_{ox}}(L-\Delta L)W\right]^{-1} + \left[\frac{\varepsilon_{si}}{t_{dep}}(L-\Delta L)W\right]^{-1}\\ &+\left[\frac{\varepsilon_{box}}{t_{box}}(L-\Delta L)W\right]^{-1}\end{aligned}\right\}^{-1} \quad (3)$$

$$C_{gb\_DG} = 0.$$

B. Inner fringe capacitance (C_{if})

To the first order, inner and outer fringe capacitances are determined by the shorter one of the two electrodes. Analytical model for inner fringe capacitance (per side) in long channel devices has previously been developed [14], under the assumption that channel length is much longer than the junction depth in bulk devices or silicon thickness in FDSOI/DG devices. However, this assumption can be problematic for advanced technology nodes where devices are designed well into the short channel regime [15]. The following analytical models are developed, so that C_{if} is determined by the smaller one of $0.5L_g$ and t_{eff}. Here, t_{eff} is the effective body thickness as defined in (4). a_{if} is a fitting parameter around 0.05~0.1. Good agreement is achieved between the numerical simulation and our analytical models, as in Figure 2. The numerical simulation is done by Maxwell 2D [16], a 2D static electromagnetic field solver.

$$t_{eff} = x_j(bulk), t_{si}(SOI) \text{ or } 0.5t_{si}(DG)$$

$$C_{if\_bulk,SOI} = \frac{2\varepsilon_{si}}{\pi}W\ln\left(1+\frac{\min(0.5L_g - t_{ox}, t_{eff})}{a_{if}\cdot t_{ox}}\right),$$

$$C_{if\_DG} = \frac{4\varepsilon_{si}}{\pi}W\ln\left(1+\frac{\min(0.5L_g - t_{ox}, t_{eff})}{a_{if}\cdot t_{ox}}\right) \quad (4)$$

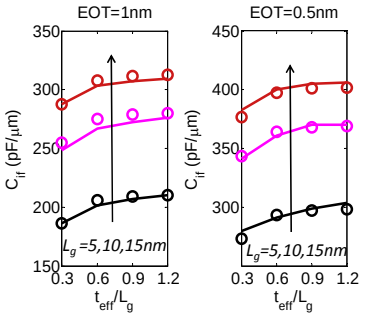


Figure 2 Comparison of analytical model (lines) and numerical simulation (circles) of C_{if}. t_{eff} is defined in (4)

C. Outer-fringe capacitance(C_{of})

Outer-fringe capacitance is modeled with conformal mapping methodology as in [17][18]. The usual assumption is that the length of the source/drain extension region is much larger than the gate height, and therefore the integration along electric field line is typically done along the gate electrode in the vertical direction. However, for advanced technology nodes, the device channel length scaling and hence the gate height scaling has slowed down, while the device pitch scaling is still about 0.7x per generation and continues to scaled down the source/drain extension length. As a result, this assumption is no longer valid. Following the conformal mapping methodology in [17]-[19], while integrating along the electrode of the source/drain extension in the horizontal direction, C_{of} can be modeled as in (5). To the first order, C_{of} is independent of gate height and contact plug height, because the distance between gate edge and contact plug is much smaller than the gate height and the contact plug height, according to Table 1. Figure 3 shows the comparison between numerical simulation and the analytical models. a_{of} is fitted to be 0.35.

$$C_{of\_bulk} = C_{of\_SOI} = \frac{\varepsilon_{cap} W}{\pi} \ln\left(\frac{L_{PCCA} + \sqrt{t_{ox}^2 + L_{PCCA}^2}}{t_{ox}}\right) + a_{of} \cdot \frac{\varepsilon_{cap} W}{\pi} \ln\left(\frac{\pi W}{t_{ox}}\right)$$

$$C_{of\_DG} = \frac{2\varepsilon_{cap} W}{\pi} \ln\left(\frac{L_{PCCA} + \sqrt{t_{ox}^2 + L_{PCCA}^2}}{t_{ox}}\right) + a_{of} \cdot \frac{2\varepsilon_{cap} W}{\pi} \ln\left(\frac{\pi W}{t_{ox}}\right) \quad (5)$$

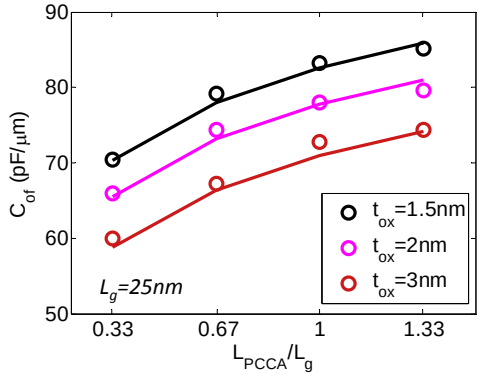


Figure 3 Comparison of analytical model (lines) and numerical simulation (circles) of C_{of}.

D. Gate-to-plug capacitance (C_{PCCA})

Gate-to-plug capacitance is split into two parts: C_{PCCA_nr} is associated with the normal field, while C_{PCCA_fr} is associated with the fringe field. In bulk devices, C_{PCCA_nr} is described by a one-piece parallel plate capacitance; while in FDSOI devices where raised-source/drain is present, C_{PCCA_nr} is described by two parallel plate capacitances in parallel. C_{PCCA_fr} is modeled following [20]. For planar DG structure as in Figure 1c, only the coupling between the top gate electrode to the contact plugs is included. The model agrees with numerical simulation as in Figure 4.

$$t_{rSD\_eff} = 0 (bulk, DG) \text{ or } t_{rSD} (SOI)$$

$$t_{gate\_eff} = t_{gate} (bulk, DG) \text{ or } (t_{gate} - t_{rSD}) (SOI)$$

$$C_{PCCA\_bulk} = C_{PCCA\_SOI} = C_{PCCA\_DG} = C_{PCCA\_fr} + C_{PCCA\_nr} = \frac{0.5\pi\varepsilon_{cap} W}{\ln\left[\frac{2\pi\left(L_{PCCA} + L_g\right)}{2L_g + \tau_{bk} \cdot t_{gate\_eff}}\right]} + \frac{\varepsilon_{cap} W t_{gate\_eff}}{L_{PCCA}} + \frac{\varepsilon_{spacer} W t_{rSD\_eff}}{t_{spacer}}$$

$$\tau_{bk} = \exp\left[2 - 2\sqrt{1 + \frac{2\left(t_{gate\_eff} + L_g\right)}{L_{PCCA}}}\right] \quad (6)$$

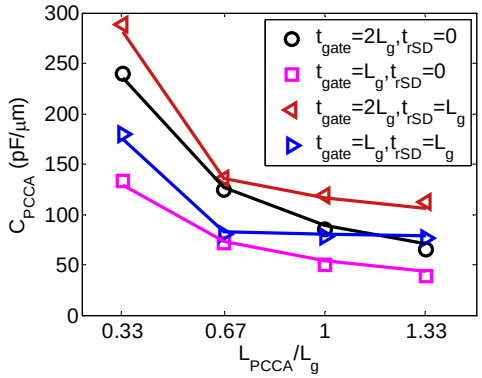


Figure 4 Comparison of analytical model (lines) and numerical simulation (circles) of C_{PCCA}.

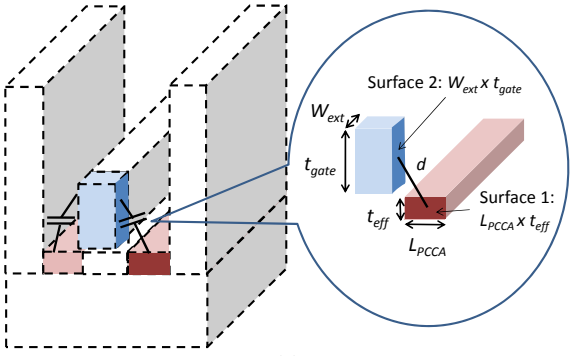
E. Corner capacitance (C_{corner})

The corner capacitance is the capacitance between the gate overhang to source/drain extension regions. The electric field associated with C_{corner} is 3-dimensional. C_{corner} is empirically modeled by (7). Here, d is the distance between the mid-points of the two shaded surfaces (surface 1 and 2) as in Figure 5a. A is the effective area, which is the geometric mean of the areas of the two shaded surfaces (surface 1 and 2). a_{corner} is fitted to be 5. Figure 5b shows the comparison between the analytical model and the numerical simulation by Maxwell 3D [21], a 3D electromagnetic simulator. The coupling between the gate overhang to the substrate is not included in this work because the distance between the gate and substrate is still relatively large in the assumed structure. It could potentially be important for ultra-thin body device when the gate and substrate are close to each other, and can be modeled in a similar way.

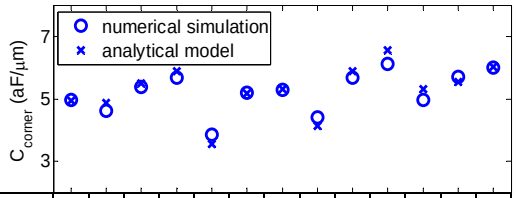
$$d = \sqrt{\left(0.5W_{ext}\right)^2 + \left(0.5t_{poly} + t_{ox} + 0.5t_{eff}\right)^2 + \left(0.5L_{PCCA}\right)^2}$$

$$A = \sqrt{W_{ext} \cdot t_{gate} \cdot L_{PCCA} \cdot t_{eff}}$$

$$C_{corner\_bulk} = C_{corner\_SOI} = 0.5C_{corner\_DG} = a_{corner} \cdot \frac{\varepsilon_{cap}}{d} \cdot A \quad (7)$$



(a)



t_{gate}/L_g	2	1	2	2	2	2	2	2	2	2	1	2	2
W_{ext}(nm)	24	24	30	36	24	24	24	16	24	24	24	30	36
L_{PCCA}/L_g	$\frac{2}{3}$	$\frac{2}{3}$	$\frac{2}{3}$	$\frac{2}{3}$	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{2}{3}$	$\frac{2}{3}$	1	$\frac{4}{3}$	$\frac{2}{3}$	$\frac{2}{3}$	$\frac{2}{3}$
t_{ox}(nm)	4	4	4	4	4	2	1	4	4	4	1	1	1

(b)

Figure 5 (a) 3D schematic of the device geometry that gives rise to C_{corner}. The gate overhang extends beyond the device width by W_{ext}. (b) comparison of analytical model (crosses) and numerical simulation (circles) of C_{corner}.

F. Junction Capacitance (C_j)

Source/drain junction capacitance is modeled as a conventional p-n junction capacitance in a bulk device, and as a parallel plate capacitance of the buried oxide between the source/drain and substrate in a FDSOI device.

$$C_{j\_bulk} = W\left[x_j + 0.5\left(L_{pitch} - L_g\right)\right]\sqrt{\frac{q\varepsilon_{si}}{2V_{bi}}\frac{N_a N_d}{N_a + N_d}} \Big/ \left(1 + \frac{V}{V_{bi}}\right)^{0.33}$$

$$C_{j\_SOI} = \frac{\varepsilon_{box}}{t_{box}} W \cdot 0.5(L_{pitch} - L_g)$$

$$C_{j\_DG} = 0 \quad (8)$$

TABLE 1 KEY PARAMETERS FOR NOMINAL DEVICES FOLLOWING THE ITRS 2009 [15].

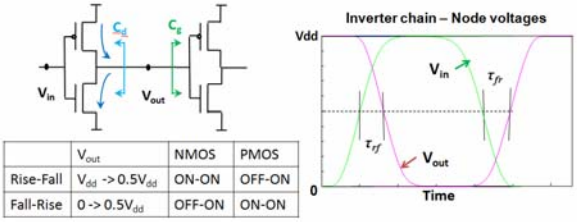
Parameter Definition	Unit	2010 32nm mBulk	2013 22nm FDSOI	2016 16nm FDSOI	2019 11nm DG
Metal 1 pitch (*M1*)	nm	90	64	44	32
Contacted gate pitch (L_{pitch})	nm	120	85	59	43
Channel length (L_g)	nm	27	20	15	12
Length of lightly doped source/drain (ΔL)	nm	3.9	2.9	2.1	1.7
Gate-to-contact length (L_{PCCA})	nm	31.5	22	14.5	10
Physical effective oxide thickness (*EOT*)	nm	0.95	0.7	0.57	0.62
Electrical effective oxide thickness ($EOT_{\_el}$)	nm	1.26	1.1	0.97	1.02
Gate height (t_{gate})	nm	54	40	30	24
Spacer thickness (t_{spacer})	nm	21	14.7	9.7	6.7
Raised source/drain thickness (t_{rSD})	nm	0	20	15	12
Contact size (L_{cont})	nm	30	21.3	14.7	10.7
Junction/Si body thickness (x_j or t_{si})	nm	12	7	5.1	3.9
Minimum NMOS width (W_{nmin})	nm	45	32	22	16
Minimum PMOS width (W_{pmin})	nm	90	64	44	32
Supply voltage (V_{dd})	V	0.97	0.87	0.78	0.71
Off-state current (I_{off})	nA/μm	100	100	100	100
NMOS Saturation current (I_{dsat})	μA/μm	1200	1470	1730	1970
N/P current ratio (I_{dsatN}/I_{dsatP})		1.3	1.25	1.2	1.18

III. ACTUAL DEVICE DELAY

The total capacitance (C_{tot}) charged/discharged at the output node is the capacitance that determines the circuit-level delay. C_{tot} includes three parts, the drain capacitance of the driving stage (C_d), the gate capacitance of the load stage (C_g), and the wiring capacitance between the two stages (C_{int}). We do not include C_{int} in this work, but it can be included in a straight forward manner as it is a constant. The drain capacitance of the driving stage consists of the drain-to-gate capacitances

(C_{dg}) and drain-to-substrate junction capacitance (C_j). Because the voltages applied to the two electrodes of C_{dg} change in a reverse direction during the switching, a multiplier M ($M=2$) is included to account for the Miller Effect. In bulk CMOS, the voltage across C_j is V_{out}. For rise-to-fall delay, V_{out} falls from V_{dd} to $0.5V_{dd}$; while for fall-to-rise delay, V_{out} rises from 0 to $0.5V_{dd}$. Thus, to the first order, C_j is calculated as the average of $C_j(V_{out}=0)$ and $C_j(V_{out}=V_{dd})$ to average τ_{rf} and τ_{fr}. C_g is also V_{out}-dependent. For MOSFET devices, when the device is off and no channel charge sheet is present, the gate is electrostaticly coupled to source/drain junction through the inner fringe capacitance (C_{if}) and to the substrate through the gate-to-substrate capacitance (C_{sub}). When the device is on and a conductive inversion layer is formed as the channel, both C_{if} and C_{sub} are shielded by the channel. Instead, the gate is coupled to the source/drain junction by the gate-to-channel capacitance (C_{gc}). For the rise-to-fall delay (τ_{rf}), V_{out} falls from V_{dd} to $0.5V_{dd}$. During this process, PMOS of the load stage is switched from the off-state to the on-state, while NMOS of the load stage stays in the on-state (Figure 6). Conversely, for the fall-to-rise delay (τ_{fr}), V_{out} rises from 0 to $0.5V_{dd}$. PMOS stays in the on-state, while NMOS is switched from the off-state to the on-state (Figure 6). Device delay (τ) is the average of τ_{rf} and τ_{fr}. To average the contribution of PMOS and NMOS in both τ_{rf} and τ_{fr}, one quarter of C_g in the off-state (C_{g_off}) and three quarters of C_g in the on-state (C_{g_on}) are included to estimate C_{tot}, to the first order, as in (9). M is the coefficient of Miller Effect, and FO is the electric fan-out. FO1 ($FO=1$) and FO3 ($FO=3$) delays of inverter chains are widely adopted figures of merit for circuit-level delay benchmarking [4].

$$\begin{aligned} C_{tot} &= C_d + \left(0.25C_{g\_off} + 0.75C_{g\_on}\right)\cdot FO \\ &= C_{dg}\cdot M + C_j + \left(0.25\cdot C_{g\_off} + 0.75\cdot C_{g\_on}\right)\cdot FO \\ &= \left(C_{ov} + C_{of} + C_{pcca} + C_{corner}\right)\cdot M + C_j \\ &\quad + \begin{bmatrix} 0.25\left(C_{gb\_off} + 2C_{ov} + 2C_{if} + 2C_{of} + 2C_{pcca} + 2C_{corner}\right) \\ +0.75\left(C_{gc} + 2C_{ov} + 2C_{of} + 2C_{pcca} + 2C_{corner}\right) \end{bmatrix}\cdot FO \end{aligned} \quad (9)$$



	V_{out}	NMOS	PMOS
Rise-Fall	V_{dd} -> $0.5V_{dd}$	ON-ON	OFF-ON
Fall-Rise	0 -> $0.5V_{dd}$	OFF-ON	ON-ON

Figure 6 Circuit and waveform of a 2-stage inverter chain. During the switching process in logic applications, an input signal (V_{in}) is applied at the gate of the driving stage. In CMOS circuits with full-swing signals, the delay is generally measured as the time duration from the 50% input level to the 50% output level. Depending on the switching directions, rise-to-fall delay (τ_{rf}) corresponds to the rising input to falling output process, while fall-to-rise delay (τ_{fr}) refers to the falling input to rising output progress. Actual device delay (τ) is the average of τ_{rf} and τ_{fr}.

IV. Parasitic Capacitance Engineering

The Model for Assessment of cmoS Technologies And Roadmaps (MASTAR) application [13]is a simple device and circuit simulator based on analytical models, used for ITRS Roadmap calculation and projection [15]. We first build nominal devices according to ITRS 2009 specifications [15], which produce the projected I-V characteristics as the ITRS 2009 projection for both n-type and p-type devices. The key parameters are listed in Table 1. Then, we calculate the average total load capacitances during a switching event for FO1 and FO3 inverters following (9), which takes into consideration the voltage dependent capacitive elements. The average load capacitances are used as the load capacitance for circuit simulation. With the full I-V characteristics and the load capacitance, we simulate 5-stage inverter chains with MASTAR, and generate the wave form similar as that shown in Figure 6. The delay is the average of τ_{rf} and τ_{fr}, which can be extracted from the waveform. By implementing (9) within the MASTAR program, the delay of an inverter chain is calculated. FO1 and FO3 delay with nominal devices including parasitic capacitances are illustrated in Figure 7. Figure 7 also include a curve showing the intrinsic CV/I projected by ITRS 2009. Though the metric of CV/I still predicts the trend of the device speed with technology scaling, it cannot be used as a metric to estimate the absolute numbers of delay. Parasitic capacitances have to be taken into considerations for both first order estimation and accurate simulation.

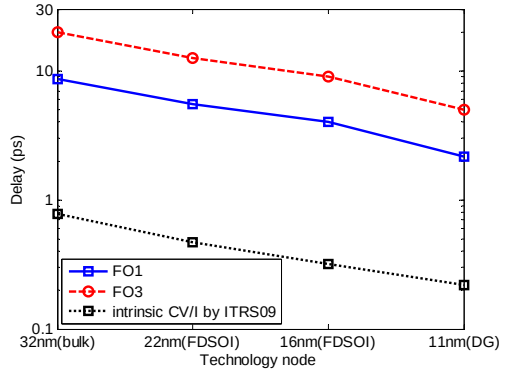
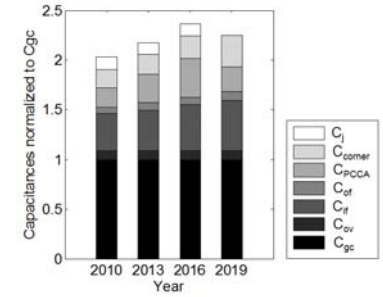


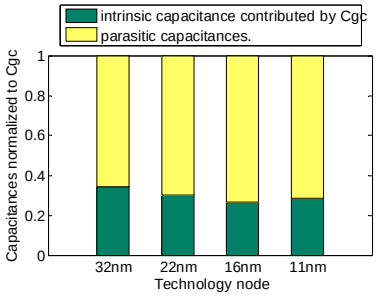
Figure 7 FO1 and FO3 delay of the nominal devices based on ITRS 2009. As projected by ITRS2009, the typical devices for Year 2010 (32nm node), Year 2013 (22nm node), Year 2016 (16nm node), and Year 2019 (11nm node) are bulk devices, FDSOI devices, FDSOI devices, and DG devices, respectively.

Figure 8a shows different capacitance components normalized to C_{gc} for nominal devices. Note that C_{ov}, C_{if}, C_{of}, C_{PCCA}, C_{corner}, and C_j are capacitances per side. Several parasitic capacitances, including C_{corner}, C_{PCCA}, and C_{if}, are comparable to C_{gc}. C_{PCCA}/C_{gc} increases when devices are scaled to more advanced technology. Taking Miller effect into consideration, C_{tot} of an FO1 inverter chain is calculated following (9). Figure 8b shows the partitions of intrinsic capacitance (contributed by C_{gc}) and parasitic capacitances in C_{tot} for nominal devices. The intrinsic capacitance only

account for 30%~40% of the total capacitance charged/discharged during the switching process. The weight of intrinsic capacitance decreases for more advanced technologies. From Year 2016 to 2019 (32nm technology node to 11nm technology node), by switching from single-channel device structures (bulk, FDSOI) to double-gate devices structure, the weight of the parasitic capacitances in the total capacitances decreases, because dual-channel structure doubles C_{gc}, but not all parasitic components are doubled. Different possible design scenarios are examined below.



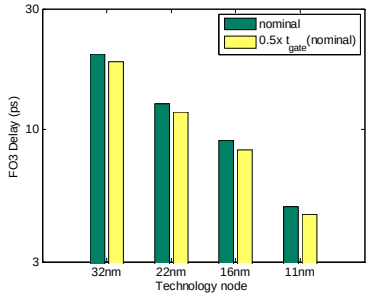
(a)



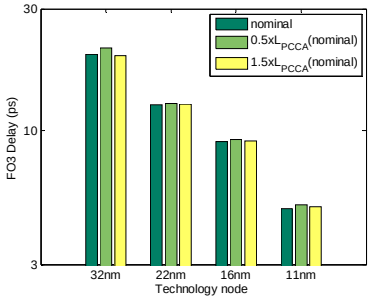
(b)

Figure 8 (a) Capacitance components normalized to C_{gc} for nominal devices based on ITRS2009, and (b) partition of intrinsic capacitance (due to C_{gc}) and parasitic capacitances in C_{tot} of a self-loaded FO1 inverter chain. Note that C_{ov}, C_{if}, C_{of}, C_{PCCA}, C_{corner}, and C_j are capacitances per side.

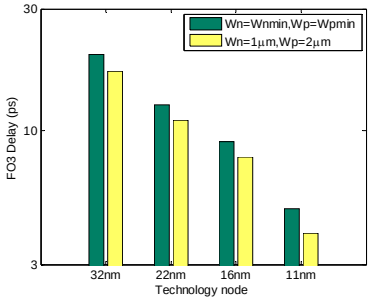
First, the dependence of C_{tot} on t_{gate} is examined (Figure 9a). When t_{gate} is reduced, C_{PCCA}, C_{of}, and C_{corner} decreases. Thus, devices with a lower gate height suffer less from parasitics. Lowering the gate height by half of the nominal case speeds up the devices by 7%, for all technologies from 32nm to 11nm (Year 2010 to 2019). This is in agreement with earlier modeling [22] and experimental studies [23]. Secondly, Figure 9b shows the delay dependence on L_{PCCA} at each technology node. At the device level, reducing L_{PCCA} reduces C_{of} and C_{corner}, with the penalty of increasing C_{PCCA}. At the circuit level, because a smaller device pitch enables shorter interconnect, thus a smaller interconnect RC delay [24] is achieved. Without taking interconnect into consideration, the delay of the inverter chain changes less than 5% when reducing L_{PCCA} by half or increasing L_{PCCA} to 1.5x. Thirdly, the C_{corner} becomes significant for small device width. Traditionally, devices can be divided into two main families according to the statistical distribution of the device widths: devices in logic cells with "large" width (~0.5μm) and devices in SRAM cells with small width (<0.1μm). As logic cell size is scaled down, the average "large" width is becoming smaller and is now close to 0.1μm. Most capacitance components are proportional to the device width to the first order, except for C_{corner}. As shown in Figure 9c, wide devices, where C_{corner} is negligible, is 13-20% faster than the one with minimum widths. Fourthly, enhancing PMOS current driving capability improves the circuit level performance. Conventionally, for static circuit design, the ratio of the width of PMOS over NMOS is about 2. Intense effort has been made to boost PMOS driving capability [25]-[27] (Figure 10a), which can potentially reduce the PMOS device width and thus reduce the total capacitance. P/N ratio is defined as the ratio of I_{dsat} of PMOS over that of NMOS. Figure 10b shows that 27% delay reduction is obtained by boosting PMOS driving capability by 2x, (i.e. same driving capability as NMOS). Boosting PMOS driving capability by 1.5x achieves 2/3 of the improvement of 2x case, i.e., 18%.



(a)



(b)



(c)

Figure 9 FO3 delay dependences on (a) gate height, (b) gate-to-contact distance, and (c) average device width.

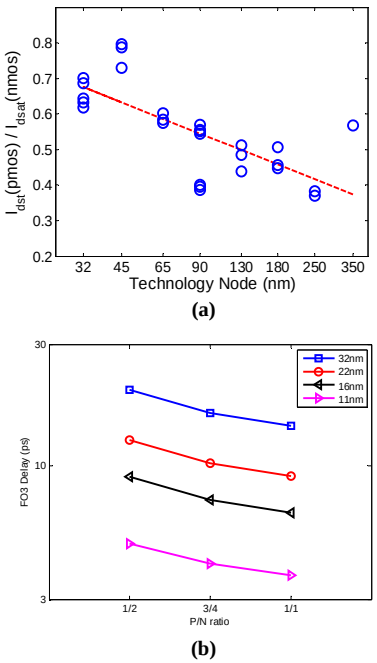


Figure 10 (a) Historical data of the ratios of saturation current (I_{dsat}) of PMOS and NMOS (P/N ratio) [3],[5],[6]-[9], [28]-[38], and (b) FO3 delay dependence on P/N ratio.

From the analysis above, lowering the gate height is an effective way to reduce the parasitic capacitances, because both C_{PCCA} and C_{of} are reduced. Reducing L_{PCCA} does not introduce a big delay penalty, because the increase of C_{PCCA} is offset by the reduction of C_{of}. When interconnect is taken into consideration, reducing L_{PCCA} can potentially improve the circuit-level speed because of the shorter interconnect. C_{corner} must be included for advanced technology nodes, where it is comparable to other capacitance components in narrow devices. Enhancing PMOS current driving capability can effectively reducing the circuit-level delay by reducing the total load capacitance. With parasitic engineering, we propose new scaling scenarios as shown in Figure 11. By introducing lower gate heights and better PMOS, the device delay performance is improved by ~30% compared with the nominal devices. Without sacrificing too much of device speed (<5%) at the device level, reducing L_{PCCA} shortens the interconnect length. A shorter interconnect length improves speed at the circuit level.

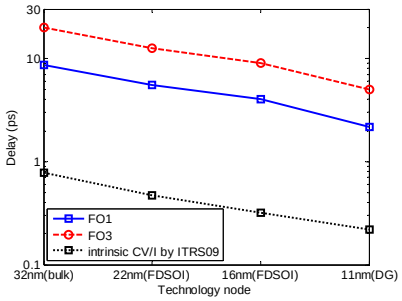


Figure 11 Proposed roadmaps beyond 32nm technology with several design scenarios to reduce parasitic capacitances.

V. Conclusions

For technologies beyond 32nm, the intrinsic delay $C_{gc}V/I$, which relies mainly on L_{gate} and does not capture the relevant design parameters necessary for MOSFET optimization, is inadequate for device performance evaluation. Parasitic capacitance has become a severe problem, which cannot be ignored for technology evaluation. We develop simple analytical models for parasitic capacitance components according to their physical origins. Starting from the ITRS 2009 edition, we use the MASTAR model to study the device delay including the parasitic capacitances. By selectively engineering the P/N ratio and the device dimensions, such as gate height, gate-to-contact distance, device width, and gate length, the device performance can be boosted by reducing parasitics and preserving electrostatic integrity. A possible improvement of the roadmap is proposed.

Acknowledgement

The authors would like to thank G. Bidal and K. Herve for useful discussion and programming help.

References

[1] R. Dennard *et al.*, "Design of Ion-Implanted MOSFET's with Very Small Physical Dimensions," *IEEE Journal of Solid-State Circuits,* vol. SC-9, no. 5, pp. 256 –268, Oct 1974.

[2] S. E. Thompson, *et al.*, "A logic nanotechnology featuring strained-silicon," *Electron Device Letters, IEEE,* vol. 25, pp. 191-193, 2004.

[3] V. Chan, *et al.*, "High speed 45nm gate length CMOSFETs integrated into a 90nm bulk technology incorporating strain engineering," *Electron Devices Meeting, 2003. IEDM 2003. IEEE International*, 2003, pp. 77-80, 2003.

[4] R. Chau, *et al.*, "High-κ/metal-gate stack and its MOSFET characteristics," *Electron Device Letters, IEEE,* vol. 25, pp. 408-410, 2004.

[5] K. Mistry et al., "A 45nm Logic Technology with High-k+Metal Gate Transistors, Strained Silicon, 9 Cu Interconnect Layers, 193nm Dry Patterning, and 100% Pb-free Packaging," *IEEE International Electron Devices Meeting (IEDM)*, pp. 247-250, Washington DC, December 10-12, 2007.

[6] S. Natarajan, *et al.*, "A 32nm logic technology featuring 2nd-generation high-k + metal-gate transistors, enhanced channel strain and 0.171 μm^2 SRAM cell size in a 291Mb array," *Electron Devices Meeting, 2008. IEDM 2008. IEEE International*, 2008, pp. 941-943.

[7] F. Arnaud, *et al.*, "32nm general purpose bulk CMOS technology for high performance applications at low voltage," *Electron Devices Meeting, 2008. IEDM 2008. IEEE International*, 2008, pp. 633-636.

[8] X. Chen, *et al.*, "A cost effective 32nm high-K/ metal gate CMOS technology for low power applications with single-metal/gate-first process," *VLSI Technology, 2008 Symposium on*, 2008, pp. 88-89.

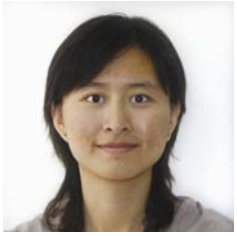
[9] C. H. Diaz, *et al.*, "32nm gate-first high-k/metal-gate technology for high performance low power applications," *Electron Devices Meeting, 2008. IEDM 2008. IEEE International*, 2008, pp. 629-632.

[10] S. E. Thompson and S. Parthasarathy, "Moore's law: the future of Si microelectronics," *Materials Today*, vol. 9, pp. 20-25, 2006.

[11] J. W. Sleight, I. Lauer, O. Dokumaci, D. M. Fried, D. Guo, B. Haran, S. Narasimha, C. Sheraw, D. Singh, M. Steigerwalt, X. Wang, P. Oldiges, D. Sadana, C.Y. Sung, W. Haensch, M. Khare, "Challenges and Opportunities for High Performance 32 nm CMOS Technology, " *IEEE International Electron Devices Meeting (IEDM)*, pp. 697-700, San Francisco, CA, December 11-13, 2006.

[12] J. Mueller, R. Thoma, E. Demircan, C. Bermicot, and A. Juge, "Modeling of Mosfet Parasitic Capacitances, and their Impact on Circuit

Performance," *Solid State Electronics* Vol 51, pp 1485- 1493, Nov-Dec 2007.
[13] MASTAR model, http://public.itrs.net/models.html
[14] R. Shrivastava and K. Fitzpatrick, "A simple model for the overlap capacitance of a VLSI MOS device," *Electron Devices, IEEE Transactions on,* vol. 29, pp. 1870-1875, 1982.
[15] ITRS Roadmap, http://itrs.net/reports.html
[16] Maxwell 2D®, Ansoft Corporation, PA.
[17] A. Bansal, B. C. Paul, and K. Roy, "Modeling and optimization of fringe capacitance of nanoscale DGMOS devices," *IEEE Transactions on Electron Devices*, vol. 52, No. 2, pp. 256–262, Feb. 2005.
[18] W. Wu and M. Chan, "Analysis of Geometry-Dependent Parasitics in Multifin Double-Gate FinFETs," *IEEE Transactions on Electron Devices*, vol. 52, No. 4, pp. 692- 698, Apr. 2007.
[19] R. Plonsey and R. Collin, Principles and Applications of Electromagnetic Fields, p146-163. New York: McGraw-Hill, 1961.
[20] J. Deng, H.-S. P. Wong, "Modeling and analysis of planar gate electrostatic capacitance for 1-D FET with multiple cylindrical conducting channels," *IEEE Trans. on Electron Devices*, vol. 54, No. 9, pp.2377-2385, 2007.
[21] Maxwell 3D®, Ansoft Corporation, PA.
[22] J. Deng, K. Kim, C.-T. Chuang, H. -S. P. Wong, "The Impact of Device Footprint Scaling on High-Performance CMOS Logic Technology," *IEEE Transactions on Electron Devices*, pp. 1148-1155, vol. 54, No.5, May 2007.
[23] Z. Ren, K. T. Schonenberg, V. Ontalus, I. Lauer, and S. A. Butt, "CMOS Gate Height Scaling," The 9th International Conference on Solid-State and Integrated-Circuit Technology (ICSICT2008), paper A1.6, Beijing, China, October 20-23, 2008.
[24] L. Wei, *et al.*, "Selective device structure scaling and parasitics engineering: a way to extend the technology roadmap," *IEEE Transactions on Electron Devices,* vol. 56, pp. 312-320, February 2009.
[25] B. F. Yang, *et al.*, "Stress dependence and poly-pitch scaling characteristics of (110) PMOS drive current," in *VLSI Technology, 2007 IEEE Symposium on*, 2007, pp. 126-127.
[26] K.-M. Tan, *et al.*, "A High-Stress Liner Comprising Diamond-Like Carbon (DLC) for Strained p-Channel MOSFET," *Electron Device Letters, IEEE,* vol. 29, pp. 192-194, 2008.
[27] H. C. H. Wang, *et al.*, "High-Performance PMOS Devices on (110)/<111'> Substrate/Channel with Multiple Stressors," in *Electron Devices Meeting, 2006. IEDM '06. International*, 2006, pp. 1-4.
[28] R. A. Chapman, J. W. Kuehne, P. S-H. Ying, W. F. Richardson, A. R. Peterson, A. P. Lane, I. -C. Chen, L. Velo, C. H. Blanton, M. M. Moslehi, and J. L. Paterson, "High performance sub-half micron CMOS using rapid thermal processing," *Electron Devices Meeting, 1991. IEDM 1991. IEEE International*, 1991, pp. 101-104.
[29] M. Rodder, Q. Z. Hong, M. Nandakumar, S. Aur, J. C. Hu, and I. C. Chen, "A sub-0.18μm gate length CMOS technology for high performance (1.5V) and low power (1.0V)," *Electron Devices Meeting, 1996. IEDM 1996. IEEE International*, 1996, pp. 563-566, 1996.
[30] P. Gilbert, *et al.*, "A high performance l.5V, 0.10μm gate length CMOS technology with scaled copper metalization," *Electron Devices Meeting, 1998. IEDM 1998. IEEE International*, 1998, pp. 1013-1016, 1998.
[31] K. K. Young, *et al.*, "A 0.13 μm CMOS technology with 193 nm lithography and Cu/low-k for high performance applications," *Electron Devices Meeting, 2000. IEDM 2000. IEEE International*, 2000, pp. 563-566, 2000.
[32] S. Tyagi, *et al.*, "A 130 nm generation logic technology featuring 70 nm transistors, dual Vt transistors and 6 layers of Cu interconnects," *Electron Devices Meeting, 2000. IEDM 2000. IEEE International*, 2000, pp. 567-570.
[33] W.-H. Lee, *et al.*, "High performance 65 nm SOI technology with enhanced transistor strain and advanced-low-K BEOL," *Electron Devices Meeting, 2005. IEDM 2005. IEEE International*, 2005, pp.56-59, 2005.
[34] S. Tyagi, *et al.*, "An advanced low power, high performance, strained channel 65nm technology," *Electron Devices Meeting, 2005. IEDM 2005. IEEE International*, 2005, pp. 1070-1072, 2005.
[35] P. Bai, *et al.*, "A 65nm logic technology featuring 35nm gate lengths, enhanced channel strain, 8 Cu interconnect layers, low-k ILD and 0.57 μm^2 SRAM cell," *Electron Devices Meeting, 2004. IEDM 2004. IEEE International*, 2004, pp. 657-660, 2004.
[36] C. Auth, *et al.*, "45nm High-k + metal gate strain-enhanced transistors," in *VLSI Technology, 2008 Symposium on*, 2008, pp. 128-129.
[37] K. Henson, *et al.*, "Gate length scaling and high drive currents enabled for high performance SOI technology using high-κ/metal gate," in *Electron Devices Meeting, 2008. IEDM 2008. IEEE International*, 2008, pp. 645-648.
[38] C. H. Jan, *et al.*, "A 45nm low power system-on-chip technology with dual gate (logic and I/O) high-k/metal gate strained silicon transistors," in *Electron Devices Meeting, 2008. IEDM 2008. IEEE International*, 2008, pp. 637-640.



Lan Wei (M'11) received her B. S. in microelectronics and economics from Peking University, Beijing, China in 2005, M. S. and Ph. D. in electrical engineering from Stanford University, Stanford, CA USA, in 2007 and 2010, respectively. She is currently a post-doctoral associate in Microsystems Technology Laboratories, Massachusetts Institute of Technology, Cambridge, USA. Her research interests include device/circuit interactive design and integrated bio-system.
She worked as a research intern at Intel (2006), IBM Research (2007), STMicroelectronics (2008), and Grenoble Institute of Technology (2008). She has contributed to the PIDS (Process Integration, Devices, and Structures) Chapter of ITRS (International Technology Roadmap for Semiconductors) 2009 Edition. Lan Wei was a recipient of a number of awards, including Stanford Graduate Fellowship (2005-2009).

Frédéric Boeuf, born 1972**,** obtained his M.Eng. and M.Sc. degree from Institut National Polytechnique de Grenoble in 1996 and Ph.D. from the University Joseph Fourier of Grenoble (France) in 2000. He joined STMicroelectronics in 2000, where he worked on the pre-development phase of 65nm and 45nm CMOS technologies. He actively participated to the development of the MASTAR model, used for the definition of the 2005 and 2007 editions of the "International Technology Roadmap for Semiconductors" roadmap to which he collaborated. He is currently managing the Advanced Devices Research Group working towards the 20/16nm CMOS thin-films technology and is industrial advisor of several Ph.D. theses in the field of device integration and modelling. His fields of expertise are semiconductor physics and CMOS device physics. He served to several European commission funded-projects as workpackage or sub-project leader. He also served as Chair of the "CMOS Devices" technical subcommittee and European Arrangement Chair of the International Electron Device Meeting conference (IEDM) in 2007 and 2008-2009 respectively. He is member of the "Device Physics" subcommittee of the "Solid State Devices and Materials" (SSDM) conference since 2006, and participated to the Technical Program Committee of the European Solid State Device Research Conference (ESSDERC) in 2004-2005. He authored and co-authored over 135 publications, including books chapters, invited talks, and presented papers at several conferences in the field. He submitted several patents and served as reviewer for the main journals in the domain of physics, devices, technology and microelectronics.

Thomas Skotnicki

H.-S. Philip Wong received the B.Sc. (Hons.) in 1982 from the University of Hong Kong, the M.S. in 1983 from the State University of New York at Stony Brook, and the Ph.D. in 1988 from Lehigh University, all in electrical engineering. He joined the IBM T. J. Watson Research Center, Yorktown Heights, New York, in 1988. In September, 2004, he joined

Stanford University as Professor of Electrical Engineering.

While at IBM, he worked on CCD and CMOS image sensors, double-gate/multi-gate MOSFET, device simulations for advanced/novel MOSFET, strained silicon, wafer bonding, ultra-thin body SOI, extremely short gate FET, germanium MOSFET, carbon nanotube FET, and phase change memory. He held various positions from Research Staff Member to Manager, and Senior Manager. While he was Senior Manager, he had the responsibility of shaping and executing IBM's strategy on nanoscale science and technology as well as exploratory silicon devices and semiconductor technology.

His research interests are in nanoscale science and technology, semiconductor technology, solid state devices, and electronic imaging. He is interested in exploring new materials, novel fabrication techniques, and novel device concepts for future nanoelectronics systems. Novel devices often enable new concepts in circuit and system designs. His research also includes explorations into circuits and systems that are device-driven. His present research covers a broad range of topics including carbon nanotubes, semiconductor nanowires, self-assembly, exploratory logic devices, nanoelectromechanical devices, novel memory devices, and biosensors.

He is a Fellow of the IEEE and served on the IEEE Electron Devices Society (EDS) as elected AdCom member from 2001 – 2006. He served on the IEDM committee from 1998 to 2007 and was the Technical Program Chair in 2006 and General Chair in 2007. He served on the ISSCC program committee from 1998 – 2004, and was the Chair of the Image Sensors, Displays, and MEMS subcommittee from 2003-2004. He serves on the Executive Committee of the Symposia of VLSI Technology and Circuits. He was the Editor-in-Chief of the IEEE Transactions on Nanotechnology in 2005 - 2006. He is a Distinguished Lecturer of the IEEE Electron Devices Society (since 1999) and Solid-State Circuit Society (2005 – 2007).