Selective Device Structure Scaling and Parasitics Engineering: A Way to Extend the Technology Roadmap

Lan Wei, *Student Member, IEEE*, Jie Deng, *Member, IEEE*, Li-Wen Chang, Keunwoo Kim, *Senior Member, IEEE*, Ching-Te Chuang, *Fellow, IEEE*, and H.-S. Philip Wong, *Fellow, IEEE*

***Abstract*—We propose a path for extending the technology roadmap when currently considered technology boosters (e.g., strain, high-κ/metal gate) reach their limits and physical gate length can no longer be effectively scaled down. By judiciously engineering the device parasitic resistance and parasitic capacitance, and considering the impact of the interconnect wiring capacitance, we propose scenarios of selective device structure scaling that will enable technology scaling and contacted gate pitch scaling for several generations beyond the currently perceived limits.**

***Index Terms*—CMOS, contacted gate pitch, device geometry, device scaling, footprint, parasitic.**

I. Introduction

TECHNOLOGY boosters such as strain, high-κ/metal gate [1] have helped the continuation of the historic performance trend down to 45-nm node. As device physical gate length is reduced below 20 nm, gate length scaling becomes less effective because of the increasing contribution of parasitic capacitance [2]–[4]. Furthermore, the shorter gate lengths must be traded off against various leakage (subthreshold, gate, BTBT) currents. In [3], the role of device pitch was explored at the device level. Mueller *et al.* [4] discussed and modeled the layout dependence of the parasitic capacitances and their impact on the circuit performances. Deng *et al.* [5] explored the concept of selective scaling and proposed new scaling scenarios substantiated by simulation and intuitive explanations, showing that even if gate length scaling slows down, significant performance gains can be achieved through aggressive scaling of the device footprint selectively. In this paper, we propose new selective scaling scenarios which extend the study to examine the effect of detailed device features in both the lateral and vertical directions (contact sizes, overlay tolerances, gate heights, and plug heights) on both the parasitic resistances and capacitances. In particular, the impact of selective scaling of device structures on the tradeoff between the parasitic capacitances (outer-fringe capacitance, gate-to-plug capacitance, plug-to-plug capacitance) and parasitic resistance are quantified. Circuit-level simulations are performed to verify the benefits of selective device structure scaling. Guidelines for selective scaling are proposed, and a more comprehensive and efficient selective scaling scenario than was described in [5] is developed. A methodology for extending technology scaling roadmap is introduced to continue the historic performance trend for several generations even without scaling the gate length or sacrificing the contacted gate pitch (or device density).

II. Background of Contacted Gate Pitch Scaling

Contacted gate pitch (L_{pitch}) is the main driver for cost and performance. It has scaled along with general lithography from 1 μm through the 45-nm node [1], [6]. We introduce the concept of selectively scaled footprint [5], which is analogous to aggressive selective scaling of the gate length introduced in 0.35- to 0.25-μm era. Selective footprint scaling enables us to trade part of parasitic capacitance (C_{par}) with extension series resistance (R_{ext}) and interconnect wiring lengths. The speed and power efficiency are improved at the circuit level with a reduced wire length and chip area [5]. This paper examines how to optimize the selective scaling in device structures in general, in both the horizontal (contact sizes, overlay tolerances) and the vertical directions (gate heights, contact plug heights). We make the bold, yet plausible, proposal that contact sizes, overlay tolerances, and heights of gates and plugs should be aggressively reduced (faster than general lithography) through process innovations. For example, aggressive reduction of contact sizes and overlay tolerance can potentially be achieved by self-assembly patterning techniques augmented by conventional photolithography. Block copolymer [7] (an organic material similar to photoresist) can self-organize into sub-20-nm holes that are self-registered

Manuscript received May 29, 2008; revised October 28, 2008. Current version published January 28, 2009. This work was supported in part by the Focus Center Research Program (FCRP) Center for Circuit and System Solutions (C2S2), by Stanford INMP, by the SRC, by the NSF/NRI, and by the PERCS DARPA (NBCH3039004). The work of L. Wei was supported in part by the Stanford Graduate Fellowship (SGF). The review of this paper was arranged by Editor H. Jaouen.

L. Wei and H.-S. P. Wong are with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA (e-mail: lanw@stanford.edu).

J. Deng was with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA. He is now with the IBM SRDC, Hopewell Junction, NY 12533 USA.

L. W. Chang is with the Department of Materials Science and Engineering, Stanford University, Stanford, CA 94305 USA.

K. Kim is with the IBM T. J. Watson Research Center, Yorktown Heights, NY, 10598 USA.

C.-T. Chuang was with the IBM T. J. Watson Research Center, Yorktown Heights, NY, 10598 USA. He is now with the Department of Electronic Engineering, National Chiao Tung University, Hsinchu, Taiwan.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TED.2008.2010573

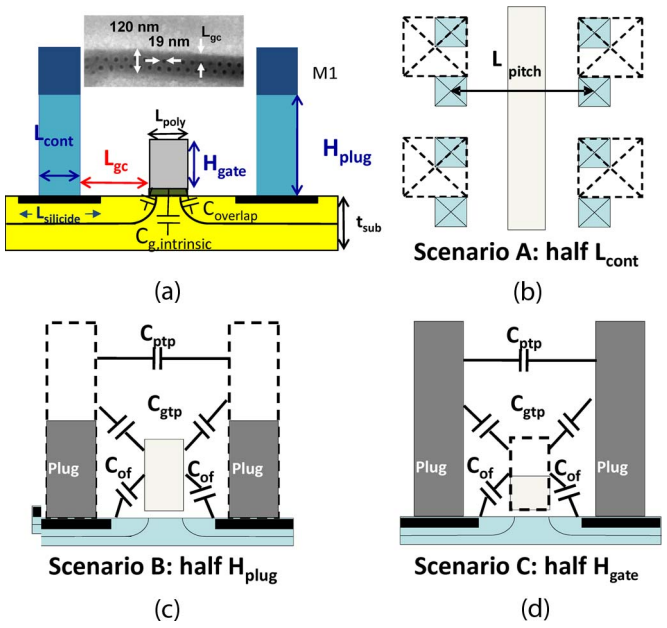


Fig. 1. (a) Schematics of planar device with related parasitics. Inset is the SEM photo of block copolymer self-assembled contact hole patterns. The holes are self-aligned to the edge of a topography 40 nm deep. Holes are 19 nm $\pm$ 1.8 nm with a pitch of 42 nm $\pm$ 2.7 nm. $L_{\rm gc}$ is the offset of the holes from the edge. (b–d) Schematics of the three proposed scaling approaches. (b) Scale contact size $(L_{\rm cont})$. (c) Scale contact plug height $(H_{\rm plug})$. (d) Scale gate height $(H_{\rm gate})$. Dashed lines denote the structures before scaling.

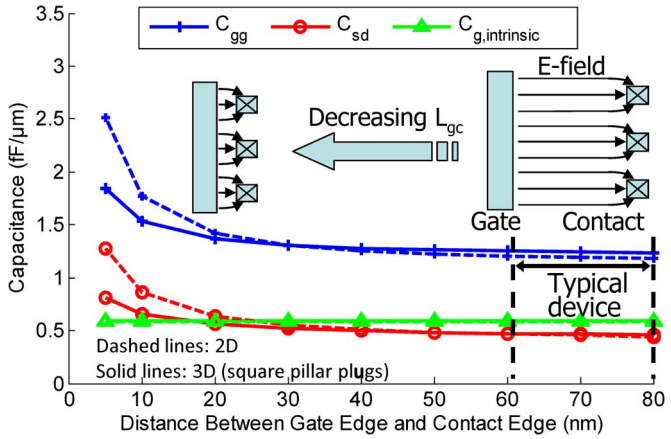


Fig. 2. Due to the increased portion of the elliptical shape of the E-field for the shorter $L_{\rm gc}(< 20$ nm), (solid lines) 3-D predicted capacitance values become much lower, compared with (dashed lines) 2-D counterparts.

to an existing 40-nm topography [inset of Fig. 1(a)] [8], [9]. In addition, the reduction of the gate heights and plug heights are enabled by metal gate technology [1].

III. Selective Device Structure Scaling: Effects on Parasitic Capacitances and Series Resistances

We first focus on the delay merit at the device level. The tradeoff between the series resistances and parasitic capacitances determines the device speed. We decompose possible selective scaling scenarios into two categories: (I) Reducing the distance between the gate edge to the inner edge of the contact plug $(L_{\rm gc})$; (II) reducing the lateral size of the contact hole $(L_{\rm cont})$, the contact plug height $(H_{\rm plug})$, and the gate height $(H_{\rm gate})$. For category (I), the series resistance is reduced because of the shorter source/drain extension region, obtained by sacrificing the parasitic capacitances; while, for category (II), the parasitic capacitances are effectively reduced, by trading off the contact series resistance. Furthermore, both (I) and (II) efficiently reduce the layout pitch, which directly shortens the interconnect length. Fig. 1(a) shows a schematic of the device structure with the geometric parameters labeled.

A. Reducing $L_{\rm gc}$

We start with a detailed analysis of the total gate capacitance $(C_{\rm gg})$ and S/D node capacitance $(C_{\rm sd})$ by both 3-D and 2-D simulations. A full 3-D simulation [10] accurately captures the 3-D fringing capacitance (Fig. 2) from the gate to contact plugs. The geometric parameters of the nominal case used for simulation are listed in Table II. The parameters are chosen based on 65-nm technology. The reason for using a 65-nm technology for analysis rather than 45-nm technology is explained in Section IV. Further discussions about more advanced technologies are given in Section VI with an extended roadmap. For the total gate capacitance $(C_{\rm gg})$ and the S/D node capacitance $(C_{\rm sd})$ which take into consideration the parasitic components, 2-D slot approximation of the plug matches the 3-D results quite well at long $L_{\rm gc}$, but overestimates the parasitic capacitances at short $L_{\rm gc}$ (Fig. 2). The reason is that for short $L_{\rm gc}$ $(< 15$ nm), the fringing effect is significant enough that the elliptical shape of the E-field cannot be ignored as in the 2-D case, as shown by the inset of Fig. 2. It is important to be clear about the range where 3-D simulation is necessary to minimize computational cost with an acceptable accuracy. 3-D simulations shows negligible difference between cylindrical plugs and square pillar plugs for $L_{\rm gc} > 15$ nm, and even for $L_{\rm gc} = 5$ nm, the cylindrical plugs give only 6% less $C_{\rm gg}$ and 7% less $C_{\rm sd}$ than square pillar plugs.

Fig. 3 shows the components of $C_{\rm gg}$ and $C_{\rm sd}$ as a function of $L_{\rm gc}$. The intrinsic gate capacitance is only $\sim$50% of $C_{\rm gg}$, the total gate capacitance. Parasitic capacitances are contributed by the outer-fringe capacitance $(C_{\rm outer-fringe})$, the gate-to-plug capacitance $(C_{\rm gate-plug})$, and the plug-to-plug capacitance $(C_{\rm plug-plug})$. The $C_{\rm outer-fringe}$ and $C_{\rm gate-plug}$ are responsible for $\sim$40%. At small $L_{\rm gc}$, the rapid increase of capacitance is due to $C_{\rm gate-plug}$. This is shown numerically in Fig. 3 and can also be explained by the analytical model described in [11]. In [11], $C_{\rm gate-plug}$ is decomposed into normal and fringing parts. The normal part is roughly inversely proportional to $L_{\rm gc}$, while the fringing part is inversely proportional to the logarithms of $L_{\rm gc}$. Both parts increase when $L_{\rm gc}$ decreases.

To first order, the increase in $C_{\rm gg}$ $(C_{\rm sd})$ will be less than 8% (15%) of the nominal values for "typical" devices (designed with standard design rules) if $L_{\rm gc}$ is larger than 0.4$\times$ gate height $(H_{\rm gate})$. For $L_{\rm gc} < 0.4\ H_{\rm gate}$, there is significant performance loss due to increasing parasitic capacitance.

Reducing $L_{\rm gc}$ reduces the series resistances, mainly due to shortening of the source/drain extension region. With the same supply voltage and channel structures, reducing the series

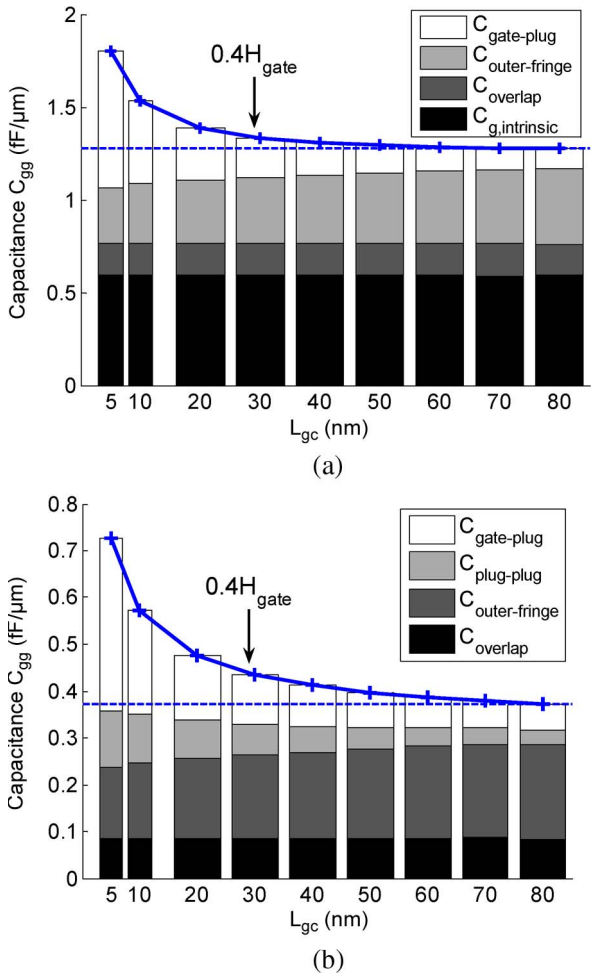


Fig. 3. (a) Gate capacitance, (b) S/D node capacitance, and their breakdown including Miller effect versus gate to plug edge distance $L_{\rm gc}$. The minimum $L_{\rm gc}$ is about $0.4H_{\rm gate}$ to keep $C_{\rm gg}(C_{\rm sd})$ within 8% (15%) of the nominal values for typical devices, respectively.

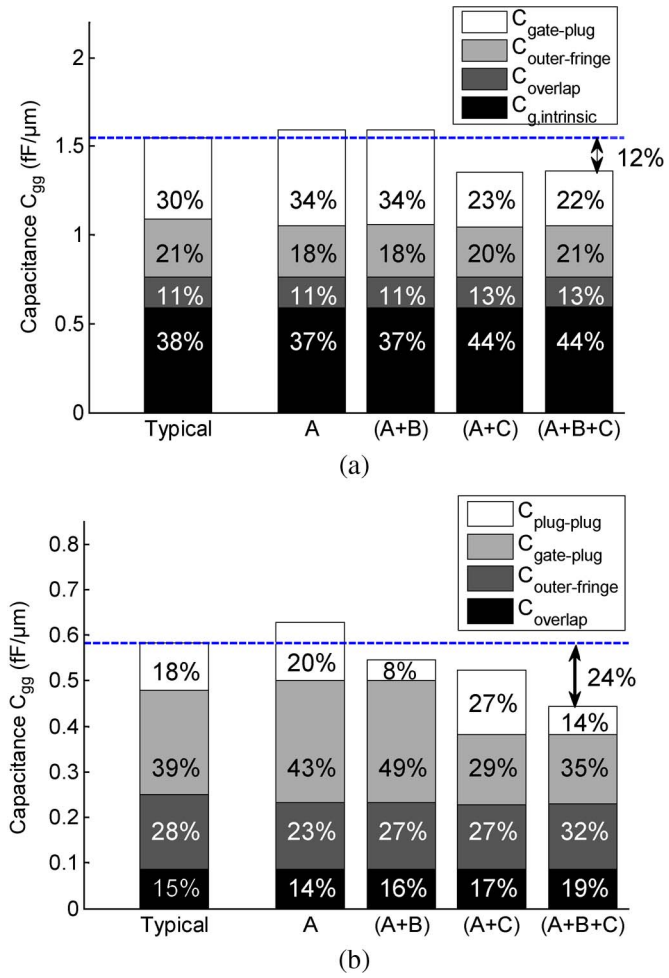


Fig. 4. (a) Gate capacitance, (b) S/D node capacitance, and their breakdown with different selective scaling scenarios. $L_{\rm gc}$ is 10 nm. Scaling down the gate height is the most effective way of reducing device parasitic capacitance.

resistances increases the on-current. Ultrathin Body Silicon-on-Insulator (UTBSOI) devices and bulk devices are built with Taurus Devices, following the conventional design parameters (Tables II and III). By reducing $L_{\rm gc}$ gradually without changing the other parameters, the on-current is improved by 10% for UTBSOI and 8% for bulk devices. The on-current improvement saturates for very small $L_{\rm gc}$ when the source/drain extension region is too short to contribute significantly to the total series resistance. For a very rough estimation, the resistance of the extension region can be approximated to be proportional to the length of the extension region and inversely proportional to the junction depth. For further studies, Kim *et al.* [12] has carefully modeled the series resistances.

B. Reducing Contact Size, Plug Height, and Gate Height

Selectively scaling the device structure (the contact size, plug height, and gate height) in all three directions can reduce the parasitic capacitance. To illustrate the concept, we studied three scaling scenarios [Fig. 2(b–d)] and their combinations: A) reducing the planar dimensions of S/D contacts by half; B) reducing the S/D contact plug height by half; and C) reducing the gate height by half. The rest of this paragraph gives a first-order qualitative analysis, followed by quantitative results in the next paragraph. Scenario "A" does not reduce either $C_{\rm gg}$ or $C_{\rm sd}$. However, it potentially reduces the interconnect length by shrinking the circuit area. Scenario "A" may increase the contact resistance sharply at very advanced technology nodes, due to limitations related to the transfer length as discussed later. Scenario "B" reduces $C_{\rm sd}$ by reducing $C_{\rm plug-plug}$, but does not reduce $C_{\rm gg}$. Reducing $H_{\rm gate}$ (Scenario "C") is effective in reducing both $C_{\rm gg}$ and $C_{\rm sd}$. The effect of any other structure scaling scenarios without changing the $L_{\rm gc}$ can be evaluated as a combination/superposition of the effects of these three basic scaling scenarios.

Three-dimensional simulation are performed with $L_{\rm gc} =$ 10 nm at the 65-nm node, to capture the impacts on the capacitances by reducing contact size, plug height, and gate height to half of the typical values as listed in Table II. Fig. 4 shows device gate capacitance [Fig. 4(a)], S/D node capacitance [Fig. 4(b)], and their components with different selective scaling scenarios. The height of each bar indicates the absolute values of the total gate capacitance $(C_{\rm gg})$ or the total S/D node capacitance $(C_{\rm sd})$, while the percentage numbers correspond to their contributions from different components. The effect of

TABLE I
EFFECTS OF SELECTIVE SCALING SCENARIOS ON PARASITICS PER UNIT GATE WIDTH

	$C_{overlap}$	$C_{outer\text{-}finge}$	$C_{gate\text{-}plug}$	$C_{plug\text{-}plug}$	R_c+R_{plug}	L_{pitch}
Scenario A	-	↓	↑	↑	↑	↓
Scenario B	-	-	↓	↓↓	↓	-
Scenario C	-	↓	↓↓↓	↑	-	-

scaling scenarios A, (A + B), (A + C), and (A + B + C) are shown in Fig. 4.

As verified by Fig. 4, to the first order, scenario "A" does not reduce either $C_{\rm gg}$ or $C_{\rm sd}$; scenario "B" reduces $C_{\rm sd}$ by reducing $C_{\rm plug-plug}$, but does not reduce $C_{\rm gg}$ to the first order; reducing $H_{\rm gate}$ (Scenario "C") is effective in reducing both $C_{\rm gg}$ and $C_{\rm sd}$. As a general rule, the most effective way to reduce the device parasitic capacitance is to reduce the height of the lowest components, e.g., the gate height for a planar bulk device, or the raised S/D and gate height for UTBSOI. That is confirmed by the simulation result that (A + C) is more effective than (A + B) in terms of reducing the parasitic capacitances.

The impacts of these selective scaling scenarios on the parasitics are summarized in Table I. At the device level, the effectiveness of reducing parasitic capacitances is, in descending order: (B + C) > (A + B + C) > C > (A + C) > B > (A + B) > Not Scaled > A. Since Scenario "A" effectively reduces the chip area, the interconnect length is reduced. Consequently, the reduction of interconnect capacitance on the critical path improves the circuit speed. When interconnect capacitance at the circuit level is also considered, this order becomes: (A + B + C) > (B + C) > (A + C) > C > B > (A + B) > A > Not Scaled, because a smaller device footprint helps reduce the interconnect capacitance. Reducing the contact size has some secondary effect on the capacitance, such as increasing $C_{\rm gate-plug}$ and $C_{\rm plug-plug}$.

Fig. 5 shows the on-current and gate capacitances for selectively scaled structures, normalized to those of the typical devices, for both SOI and bulk devices. A complete device is built using Taurus Device [14], with the geometric parameters as in Table II and the doping profiles as in Table III. The channel structure remains the same for different $L_{\rm pitch}$, thus the magnitude of the on-current reveals the difference in the parasitic resistances in a reverse way. Reducing the plug height and the gate height do not change the series resistance to first order.

In our simulation for the 65-nm technology (Fig. 5), the contact resistance is not very sensitive to the contact size for device structure scaling scenarios for current-generation technologies. However, for future technologies, it is quite possible that reducing the contact size can dramatically increase the series resistance. We use the transmission line model [13] to estimate this impact: $R_{\rm cont} = \sqrt{R_S\rho_C}\coth(L_{\rm silicide}\sqrt{R_S/\rho_C})$. Here, $R_{\rm cont}$ is the contact resistance between the diffusion layer and the silicide layer, in units of ohm. R_S is the sheet resistance per square of the underlying heavily doped silicon layer in units of ohm/□, ρ_C is the specific contact resistivity between the metal and the diffusion layer, in units of ohm square centimeter. A transfer length $l_t = \sqrt{\rho_C/R_S}$ is defined [13]. The calculated transfer length for our 65-nm technology devices is around 60 and 30 nm for NMOS and PMOS, respectively. Most current paths end within a length of l_t, which means the contact resistance only slightly depends on the length of the silicide region ($L_{\rm silicide}$), when $L_{\rm silicide}$ is greater than l_t. However, once $L_{\rm silicide}$ is less than l_t, a sharp increase of contact resistance is expected if $L_{\rm silicide}$ is further reduced. Typically, $L_{\rm silicide}$ is proportional to the technology feature size. To overcome this contact resistance issue, $L_{\rm silicide}$ has to be relaxed, since $L_{\rm silicide}$ is less than l_t in the advanced technologies.

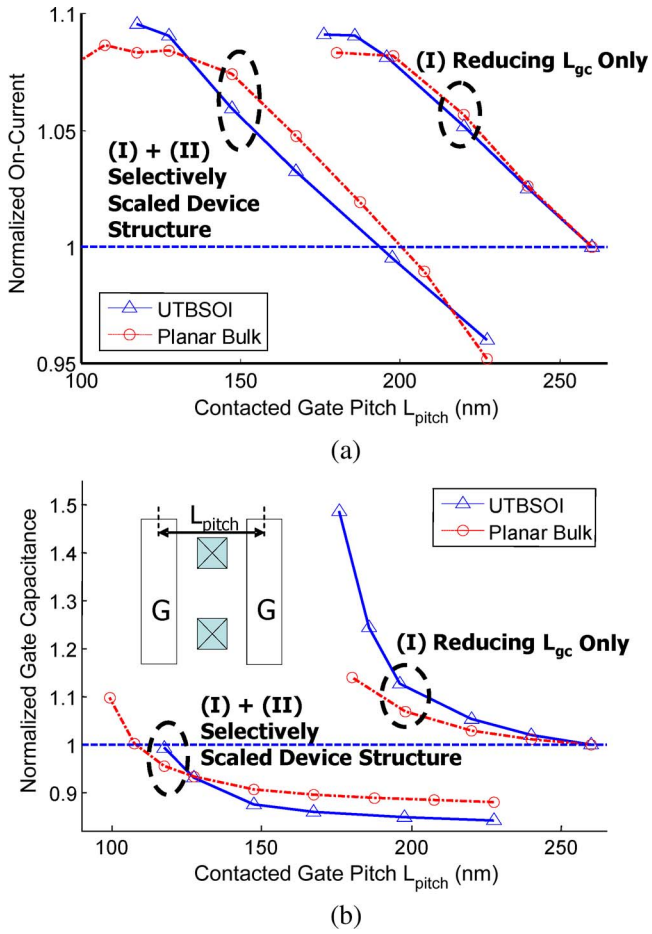


Fig. 5. (a) On-current and (b) gate capacitance versus $L_{\rm pitch}$ for both planar bulk CMOS and UTBSOI in 65-nm node with (I) reducing $L_{\rm gc}$ only and (I) + (II) selectively scaled device structure. The currents corresponding to (I) + (II) are of the same magnitudes as (I), only with a displacement in $L_{\rm pitch}$ axis.

Generally, the device structure should be selectively scaled in the following ways. 1) Reduce the vertical electrode heights, particularly the height of the lower electrode, i.e., $H_{\rm gate}$. 2) Moderately reduce the overlay tolerance ($L_{\rm gc}$). A rule of thumb for $L_{\rm gc}$ is about 0.4 $H_{\rm gate}$. 3) Reduce the lateral contact size down to the level of the transfer length. In the following sections, we show examples to illustrate the effectiveness of selective device structure scaling for improving device and circuit level performance.

TABLE II
GEOMETRIC PARAMETERS USED FOR THE TYPICAL CASE OF 65-nm TECHNOLOGY DEVICES

L_{poly}	L_{gc}	L_{cont}	$L_{silicide}$	H_{gate}	H_{plug}	t_{ox}	k_{ox}	t_{sub}
35nm	80nm	65nm	130nm	70nm	210nm	1.2nm	3.9	100nm

TABLE III
KEY PARAMETERS FOR DEVICE DOPING PROFILES

UTBSOI		bulk	
body thickness	15nm	S/D junction depth	35nm
p-channel doping	$5.4x10^{18}cm^{-3}$	S/D peak doping	$1x10^{20}cm^{-3}$
n-channel doping	$5.25x10^{18}cm^{-3}$	S/D extension junction depth	11nm
poly doping	$5x10^{19}cm^{-3}$	S/D extension peak doping	$2x10^{20}cm^{-3}$
S/D peak doping	$5x10^{19}cm^{-3}$	p-channel peak doping	$6x10^{18}cm^{-3}$
		n-channel peak doping	$1.8x10^{18}cm^{-3}$
		poly doping	$1x10^{20}cm^{-3}$

IV. INVERTER DELAY IMPROVEMENT

To validate the concept of selective device structure scaling, we use mixed-mode device/circuit simulations (Taurus [14]) to optimize FO4 inverter delay for both planar bulk CMOS and UTBSOI. The nominal devices are built with the key parameters listed in Tables II and III. For the group labeled "Reducing $L_{\rm gc}$ only," we gradually reduce $L_{\rm gc}$ with all the other parameters unchanged. For the group labeled "Selectively scaled device structure," we gradually reduce $L_{\rm gc}$, while halving $L_{\rm cont}$, $H_{\rm gate}$, and $H_{\rm plug}$.

First, we analyze the behavior of the saturation on-current and the device capacitances as a function of the distance $(L_{\rm sd})$ between the gate edge and the S/D contact stud for both planar bulk CMOS and UTBSOI with the scaling scenarios: (I) reducing $L_{\rm gc}$ only and (I) + (II) 3-D $(\rm A + B + C)$ selectively scaled device structure. We choose 65-nm technology rather than the up to date 45-nm technology mainly because reliable compact device model is available to us only up to 65-nm node. Moreover, 65- and 45-nm technologies are similar in the way that the contact size scaling is not expected to increase the series resistance significantly. Clearly, reducing the distance $L_{\rm gc}$ results in a tradeoff between higher on-current due to reduced series resistance [Fig. 5(a)] and higher parasitic capacitance [Fig. 5(b)]. The maximum on-current improvement of the scaled device over the device with the standard digital circuit design rule (identified as "typical device" in the following) is about 10%. When evaluating dynamic performance (circuit speed), capacitance effects are also significant. Three-dimensional device structure scaling $(\rm A + B + C)$ is able to reduce the gate capacitance by about 10% as compared to that of the typical device [Fig. 5(b)]. The general trend for bulk CMOS and UTBSOI are similar.

A four-stage inverter chain is used to examine the dynamic performance for both planar bulk CMOS and UTBSOI. Circuit simulations are performed using device/circuit mixed-mode numerical simulation in Taurus Device [14]. The device width ratio between pFET and nFET is designed at 1.5 to balance the pull-up and the pull-down delay. The inverter delay (τ) is evaluated between the 50% to 50% points. A tradeoff between the on-current and capacitances exists when sizing the dimensions of S/D region. Up to 5% higher speed can be obtained simply by pushing the gate to silicide distance $L_{\rm gs}$ ($L_{\rm gc} = L_{\rm gs} + \Delta$, Δ is the distance from the silicide edge to the plug edge.) smaller (Fig. 6). With 3-D selective structure scaling $(\rm A + B + C)$, the inverter speed is about 15% faster for both the planar bulk device and UTBSOI, compared with devices with standard layout rules ($L_{\rm sd} = 12\ \lambda$). At the design point with minimum FO4 delay, the on-current improvement over typical device is about 7%, and the total device length is about $L_{\rm sd0} = 6.6\lambda$ which is 45% smaller (isolated device) layout area than the typical device.

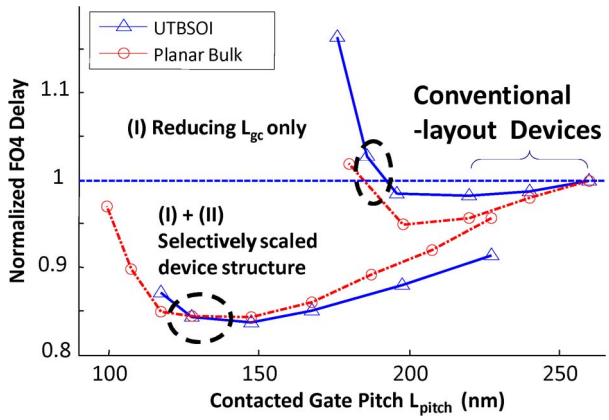


Fig. 6. FO4 delay versus $L_{\rm pitch}$ for both planar bulk CMOS and UTBSOI with (I) reducing $L_{\rm gc}$ only and (I) + (II) selectively scaled device structures.

The reduced junction capacitance for bulk device with selective structure scaling has trivial impact $(< 3\%)$ on speed because $C_{\rm plug-plug}$, $C_{\rm gate-plug}$, $C_{\rm outer-fringe}$, and $C_{\rm overlap}$ are the largest components of $C_{\rm sd}$. This also explains the similarity of the trend between bulk CMOS and UTBSOI. Mobility degradation due to reduced stress of the smaller active area ($\sim$50 MPa less stress by reducing $L_{\rm gc}$ by 65 nm) is smaller than 4.5% [15], [16] which corresponds to less than 3% on-current degradation. As a result, even though the stress-dependent mobility is ignored in our analysis, the general trend and the conclusions addressed in this paper remain valid. For more advanced technologies which depend heavily on strain-induced mobility enhancement, the reduction of drive-current with reduced stress must be further studied.

TABLE IV
SUMMARY OF PARAMETERS USED FOR ITRS TECHNOLOGY PROJECTION [2005]. THE RC DELAY IS EXPECTED TO SCALING ACCORDING TO ITRS ROADMAP. THE TOTAL RESISTANCE R IS THE SUM OF THE CHANNEL RESISTANCE ($R_{\rm chan}$), THE EXTENSION RESISTANCE ($R_{\rm ext}$), AND THE CONTACT RESISTANCE (R_c). THE $R_{\rm chan}$ IS ASSUMED TO BE ABOUT 70% OF THE TOTAL RESISTANCE IN A WELL-DESIGNED DEVICE UNDER THE TYPICAL LAYOUT RULES. THE EXTENSION RESISTANCE, CONTACT RESISTANCE, AND CAPACITANCE ARE COMPUTED BASED ON THE PARAMETERS LISTED HERE

Technology node	nm	90	65	45	32	22	16	11
Metal 1 half pitch (contacted)	nm	90	65	45	32	22	16	11
Supply voltage (V_{dd})	V	1.2	1	1	0.9	0.8	0.7	0.6
Contact junction depth (X_j)	nm	55	38.5	27.5	19.8	15.4	12.1	11
Silicide thickness ($t_{silicide}$)	nm	27.5	19.25	13.75	9.9	7.7	6.05	5.5
Physical gate length (L_{gate})	nm	50	35	25	18	14	11	10
Drain extension junction depth (X_{jext})	nm	17.5	12.25	8.75	6.3	4.9	3.85	3.5
Contact maximum resistivity (ρ_c)	ohm-cm^2	3e-8	3e-8	3e-8	3e-8	1e-8	1e-8	1e-8

V. CIRCUIT-LEVEL IMPROVEMENT

The performance improvement at the circuit macrolevel is larger than the inverter delay discussed above because a smaller device footprint results in a smaller layout area and reduces the interconnect capacitance. We verify this for UTBSOI by a full custom 53-bit multiplier using both the conventional layout and the optimized footprint ($L_{\rm sd} = 7\lambda$, the minimum point in Fig. 6), in a similar way as described in [5]. Compared with the multiplier made with typical devices ($L_{\rm sd} = 12\lambda$), the multiplier built with the selectively scaled devices ($L_{\rm sd} = 7\lambda$) occupies 30% less layout area, operates at 25% higher speed ($\sim$10% comes from the shorter interconnects), and consumes 10% less dynamic power due to the smaller interconnect capacitance of the smaller circuit layout area. The principles of selective structure scaling for bulk devices are the same as UTBSOI, and the amount of improvement in the device level and simple circuit level are very similar for UTBSOI and bulk devices. Since the major difference for a circuit macrolevel analysis is the reduction of wiring capacitances due to the tighter pitches, we expect the bulk devices have the similar improvement at the system level with the $L_{\rm sd} = 7\lambda$ footprint.

VI. EXTENDING THE TECHNOLOGY ROADMAP

Fig. 8 shows a scaling scenario in which aggressive $L_{\rm pitch}$ scaling compensates for the slower than 0.7$\times$ per node $L_{\rm gate}$ scaling. With selective device structure scaling in both the horizontal (reducing contact size and $L_{\rm gc}$) and vertical (reduced gate and plug height) directions, the technology roadmap can be extended to the 11-nm node with physical gate length no shorter than 10 nm. The parameters used for the projection are listed in Table IV. $L_{\rm pitch}$ scaling is bounded by parasitic capacitance and contact resistance. Scaling the length of the S/D silicide ($L_{\rm silicide}$) below the current transfer length causes rapid increase of contact resistance R_c below 45-nm technology node (Fig. 7). In order not to degrade the on-current for aggressively scaled devices, a relaxed silicide length ($L_{\rm silicide}$) in source/drain contact regions has to be used once $L_{\rm silicide}$ becomes comparable to or less than the transfer length. As shown in Fig. 7(b), the contact resistance with the silicide length reduced by half becomes much larger than that of the typical device beyond 45-nm technology. For the extended technology roadmap, we relax $L_{\rm silicide}$ in a way that the total series resistance after selective scaling is no larger than that for the conventional layout, as indicated by the solid line in Fig. 7(a). The main limiter of the transfer length is the specific

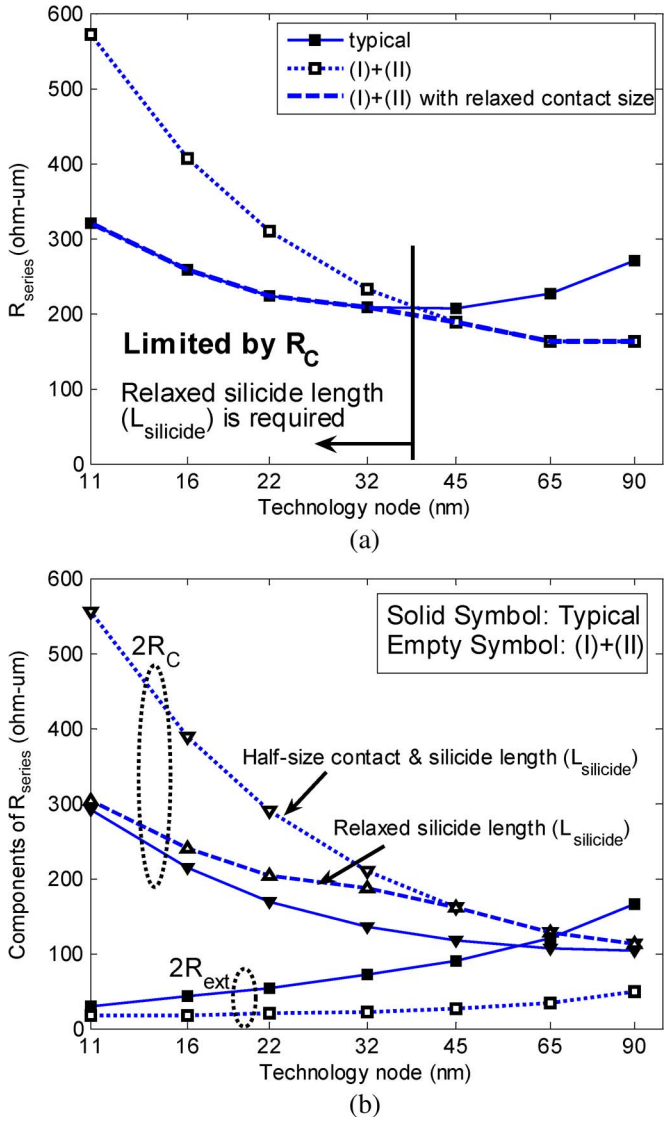


Fig. 7. (a) $R_{\rm series}$ and (b) its components as a function of technology nodes. R_c increase quickly with half-size contact scaling. A relaxed contact silicide length ($L_{\rm silicide}$) should be used for future technology nodes in order not to degrade on-current.

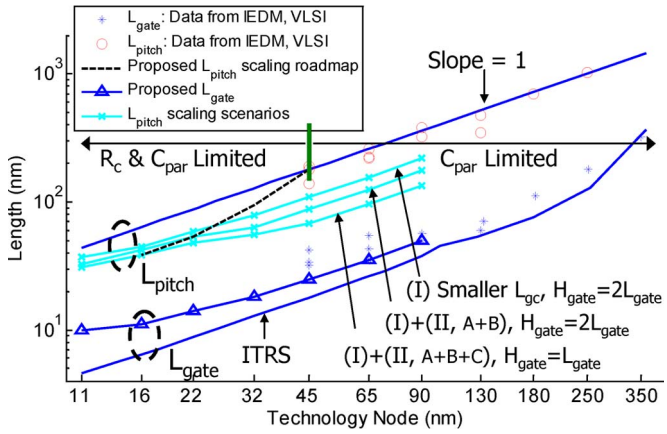


Fig. 8. Contacted gate pitch ($L_{\rm pitch}$) and physical gate length ($L_{\rm gate}$) versus technology node down to 11-nm node. The minimum $L_{\rm pitch}$ (lines with "×" symbol) is bounded by either parasitic capacitance ($C_{\rm par}$) or contact resistance (R_c) or both. By device selective structure scaling, the historic performance trend can continue for another two to three generations even without gate length scaling. The suggested technology scaling path is denoted by dashed curve.

contact resistivity, which is difficult to reduce with present techniques. We therefore assume that the transfer length does not scale significantly along with the technology [17], [18]. Potentially, the specific contact resistivity can be reduced by further increasing the doping concentration in the diffusion layer or lowering the barrier heights by choosing a different metal or silicide [17]–[19]. Metal source/drain with fermi-level depinning in the Schottky junctions, is a possible candidate to reduce source/drain resistance [20]–[23]. The $L_{\rm pitch}$ scaling is bounded by the relaxed $L_{\rm silicide}$ in more advanced technologies. Fig. 8 marks the $L_{\rm pitch}$ boundary within which the device on-current is greater than or equal to the values for "typical" devices with zero or trivial parasitic capacitance penalty. The extended scaling path requires tight pitch patterning, tight overlay tolerances, small contact holes, and a short gate height processes. They are all potential yield limiters. On the other hand, the benefits of selective device structure scaling are significant. Novel nanofabrication techniques are needed to realize the substantial benefits offered by $L_{\rm pitch}$ scaling and parasitics engineering. The potential candidates of such fabrication techniques include diblock copolymer for small contact sizes and overlay tolerance, and metal gate process for low gate height.

VII. Conclusion

In this paper, we propose a new device scaling scenario for sub-45-nm technology node high-performance CMOS technology. We postulate that even with the gate length remaining essentially the same, selectively scaling the device structure will provide significant circuit-level performance improvement from technology generation to technology generation. The benefit comes from optimizing the tradeoff between series resistances and parasitic capacitances and the reduction of the interconnect capacitances. By shrinking the lateral distance between the gate edge and the source/drain contact edge, the parasitic capacitance increases but both the series resistance and interconnect capacitance decreases. By vertically lowering the gate heights and plug heights and reducing the contact sizes, the parasitic capacitance and interconnect capacitance are reduced with the penalty of series resistance. For small benchmark circuits, such as inverter chains, and a fully custom designed 53-bit multiplier, the selectively scaled device with reduced footprint achieves smaller layout area, higher speed, and energy efficiency. The results are verified by 2-D and 3-D electrostatic simulation [10] (for capacitances), Taurus Device [14] and analytical calculation [12] (for series resistances), Taurus mix-mode simulation [14] (for inverter chains) and complex circuit simulation as in [5] (for 53-bit multiplier). This paper provides a strong incentive to develop innovative technologies, such as small contacts, tight tolerances, low-κ spacers [24], low gate and plug height integration schemes [25], and low specific contact resistivity source/drain technology [18], [19], such as metal S/D with unpinned Fermi level [20], [22], [23].

Appendix

The key parameters used for simulations are listed in Tables II–IV.

Acknowledgment

The authors would like to thank D. Frank, W. Haensch, M. Ieong, J. Sleight (IBM), K. Saraswat, M. Horowitz (Stanford), D. Antoniadis (MIT), and S. Thompson (U. Florida) for the discussions.

References

[1] 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 *IEDM Tech. Dig.*, Washington DC, Dec. 10–12, 2007, pp. 247–250.

[2] S. E. Thompson and S. Parthasarathy, "Moore's law: The future of Si microelectronics," *Mater. Today*, vol. 9, no. 6, pp. 20–25, Jun. 2006.

[3] 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, and M. Khare, "Challenges and opportunities for high performance 32 nm CMOS technology," in *IEDM Tech. Dig.*, San Francisco, CA, Dec. 11–13, 2006, pp. 697–700.

[4] J. Mueller, R. Thoma, E. Demircan, C. Bermicot, and A. Juge, "Modeling of MOSFET parasitic capacitances, and their impact on circuit performance," *Solid State Electron.*, vol. 51, no. 11/12, pp. 1485–1493, Nov./Dec. 2007.

[5] J. Deng, K. Kim, C.-T. Chuang, and H.-S. P. Wong, "The impact of device footprint scaling on high-performance CMOS logic technology," *IEEE Trans. Electron Devices*, vol. 54, no. 5, pp. 1148–1155, May 2007.

[6] ITRS Roadmap. [Online]. Available: http://itrs.net/reports.html

[7] C. T. Black, K. W. Guarini, R. Ruiz, E. M. Sikorski, I. V. Babich, R. L. Sandstrom, and Y. Zhang, "Polymer self assembly in semiconductor microelectronics," in *IEDM Tech. Dig.*, San Francisco, CA, Dec. 11–13, 2006, pp. 439–442.

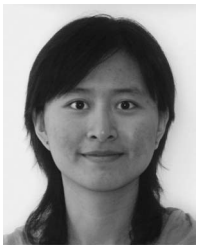
[8] L.-W. Chang and H.-S. P. Wong, "Diblock copolymer directed self-assembly for CMOS device fabrication," in *Proc. 31st SPIE Int. Symp. Microlithography, Des. Process Integr. Microelectron. Manuf. IV*, A. A. K. Wong and V. K. Singh, Eds, 2006, vol. 6150, pp. 329–334.

[9] J. Bang, S. H. Kim, E. Drockenmuller, M. J. Misner, T. P. Russell, and C. J. Hawker, "Defect-free nanoporous thin films from ABC triblock copolymers," *J. Amer. Chem. Soc.*, vol. 128, no. 23, pp. 7622–7629, May 2006.

[10] *Maxwell 3D*, Pittsburgh, PA: Ansoft Corp.

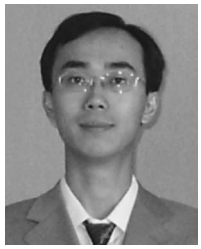
[11] J. Deng and H.-S. P. Wong, "Modeling and analysis of planar gate electrostatic capacitance for 1-D FET with multiple cylindrical conducting channels," *IEEE Trans. Electron Devices*, vol. 54, no. 9, pp. 2377–2385, Sep. 2007.

[12] S.-D. Kim, C.-M. Park, and J. C. S. Woo, "Advanced model and analysis of series resistance for CMOS scaling into nanometer regime. I. Theoretical derivation," *IEEE Trans. Electron Devices*, vol. 49, no. 3, pp. 457–466, Mar. 2002.
[13] H. H. Berger, "Contact resistance and contact resistivity," *J. Electrochem. Soc.*, vol. 119, no. 4, pp. 507–514, Apr. 1972.
[14] *Taurus-Device*, Santa Clara, CA: Synopsys Corp., Version 2005.10.
[15] S. Thompson, G. Sun, K. Wu, J. Lim, and T. Nishida, "Key differences for process-induced uniaxial vs. substrate-induced biaxial stressed Si and Ge channel MOSFETs," in *IEDM Tech. Dig.*, San Francisco, CA, Dec. 13–15, 2004, pp. 221–224.
[16] N. Shah, "Stress Modelling of Nanoscle MOSFET," M.S. thesis, Univ. Florida, Gainesville, FL, 2005.
[17] R. Shenoy and K. Saraswat, "Optimization of extrinsic source/drain resistance in ultrathin body double-gate FETs," *IEEE Trans. Nanotechnol.*, vol. 2, no. 4, pp. 265–270, Dec. 2003.
[18] A. Yagishita, T.-J. King, and J. Bokor, "Schottky barrier height reduction and drive current improvement in metal source/drain MOSFET with strained-Si channel," *Jpn. J. Appl. Phys.*, vol. 43, no. 4B, pp. 1713–1716, 2004.
[19] A. Kinoshita, C. Tanaka, K. Uchida, and J. Koga, "High-performance 50-nm-gate-length Schottky-source/drain MOSFETs with dopant-segregation junctions," in *VLSI Symp. Tech. Dig.*, Jun. 14–16, 2005, pp. 158–159.
[20] D. Connelly, C. Faulkner, D. Grupp, and J. Harris, "A new route to zero-barrier metal source/drain MOSFETs," *IEEE Trans. Nanotechnol.*, vol. 3, no. 1, pp. 98–104, Mar. 2004.
[21] T. Takahashi, T. Nishimura, L. Chen, S. Sakata, K. Kita, and A. Toriumi, "Proof of Ge-interfacing concepts for metal/high-k/Ge CMOS," in *IEDM Tech. Dig.*, Washington DC, Dec. 10–12, 2007, pp. 697–700.
[22] M. Kobayashi, A. Kinoshita, K. Saraswat, H.-S. P. Wong, and Y. Nishi, "Fermi-level depinning in metal/Ge Schottky junction and its application to metal source/drain Ge NMOSFET," in *Proc. VLSI Symp. Technol.*, Honolulu, HI, Jun. 17–20, 2008, pp. 54–55.
[23] J. Hu, D. Choi, J. S. Harris, K. Saraswat, and H.-S. P. Wong, "Fermi-level depinning of GaAs for Ohmic contacts," in *Proc. Device Res. Conf.*, Santa Barbara, CA, Jun. 23–25, 2008.
[24] J. Park and C. Hu, "Air spacer MOSFET technology for 20 nm node and beyond," presented at the The 9th Int. Conf. on Solid-State and Integrated-Circuit Technology ICSICT, Beijing, China, Oct. 20–23, 2008, Paper A1.10.
[25] Z. Ren, K. T. Schonenberg, V. Ontalus, I. Lauer, and S. A. Butt, "CMOS gate height scaling," in The 9th Int. Conf. on Solid-State and Integrated-Circuit Technology ICSICT, Beijing, China, Oct. 20–23, 2008.



Lan Wei (S'06) received the B.S. degree in microelectronics from Peking University, Beijing, China, in 2005, the M.S. degree in electrical engineering from Stanford University, Stanford, CA, in 2007, where she is currently working toward the Ph.D. degree.

Currently, she is working on device scaling and benchmarking from a perspective of circuit-level performance, particularly including parasitic capacitances. Her research interests focus on device and circuit areas.

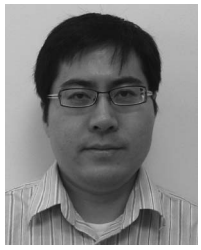


Jie Deng (S'05–M'08) received the B.S. degree in electrical engineering from the Beijing University, Beijing, China, in 2001 and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, in 2004 and 2007, respectively.

He was with IBM SRDC, Hopewell Junction, NY, in June 2007. He is working on Logic performance benchmarking and Iddq modeling. He is currently with the Department of Electrical Engineering, Stanford University.

Li-Wen Chang received the B.S. degree in material science and engineering from National Tsing Hua University, Hsinchu, Taiwan, in 2002 and the M.S. degree in material science and engineering from Stanford University, Stanford, CA, in 2006.

From 2002 to 2004, she was with UMC, a semiconductor foundry, as a Process Integration Engineer working on 0.35- and 0.25-μm processes. She is currently with the Department of Materials Science and Engineering, Stanford University. She is currently working on fabricating CMOS devices using diblock copolymer self-assembly. Her research interests are in the fabrication and characterization of novel devices.



Keunwoo Kim (S'98–M'01–SM'06) was born in Daegu, Korea, in 1968. He received the B.S. degree in physics from Sung-Kyun-Kwan University, Seoul, Korea, in 1993 and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Florida, Gainesville, in 1998 and 2001, respectively. His doctoral research was in the area of SOI and double-gate device design and modeling.

Since June 2001, he has been with the VLSI Design Department, IBM T. J. Watson Research Center, Yorktown Heights, NY, as a Research Staff Member. He has worked on the design of high-performance and low-power microprocessors, novel VLSI circuit techniques, scaled and exploratory CMOS technology performance/power evaluation, and physics/modeling for bulk-Si, SOI, strained-Si, SiGe, hybrid orientation/device, and double-gate technologies. He has published over 70 papers in technical journals and conference proceedings. He is the holder of ten U.S. patents with another five U.S. patents pending. His present work includes IBM's POWER7 processor design, the analysis and prediction of SRAM variability/yields, and system-level performance/power projections.

Dr. Kim has received five invention achievement awards from IBM. He has been a Reviewer of the journal publications for IEEE TRANSACTIONS ELECTRON DEVICES, IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN OF INTEGRATED CIRCUITS AND SYSTEMS, IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION SYSTEM, AND SOLID-STATE ELECTRONICS. He was listed in *Who's Who in America* in 2007 and 2008.

Ching-Te Chuang (S'78–M'82–SM'91–F'94) received the B.S.E.E. degree from National Taiwan University, Taipei, Taiwan, in 1975 and the Ph.D. degree in electrical engineering from the University of California, Berkeley, in 1982.

From 1977 to 1982, he was a Research Assistant with the Electronics Research Laboratory, University of California, Berkeley, working on bulk and surface acoustic wave devices. He was with the IBM T. J. Watson Research Center, Yorktown Heights, NY, in 1982. From 1982 to 1986, he worked on scaled bipolar devices, technology, and circuits. He studied the scaling properties of epitaxial Schottky barrier diodes, did pioneering works on the perimeter effects of advanced double-poly self-aligned bipolar transistors, and designed the first subnanosecond 5-kb bipolar ECL SRAM. From 1986 to 1988, he was Manager of the Bipolar VLSI Design Group, working on low-power bipolar circuits, high-speed high-density bipolar SRAMs, multigigabits per second fiber-optic data-link circuits, and scaling issues for bipolar/BiCMOS devices and circuits. Since 1988, he has managed the High Performance Circuit Group, investigating high-performance logic and memory circuits. Since 1993, his group has been primarily responsible for the circuit design of IBM's high-performance CMOS microprocessors for enterprise servers, PowerPC workstations, and game/media processors. Since 1996, he has been leading the efforts in evaluating and exploring scaled/emerging technologies, such as PD/SOI, UT/SOI, strained-Si devices, hybrid orientation technology, and multigate/FinFET devices, for high-performance logic and SRAM applications. Since 1998, he has been responsible for the Research VLSI Technology Circuit Codesign strategy and execution. His group has also been very active and visible in leakage/variation/degradation tolerant circuit and SRAM design techniques.

Dr. Chuang has received one Outstanding Technical Achievement Award, one Research Division Outstanding Contribution Award, five Research Division Awards, and 12 Invention Achievement Awards from IBM. He took early retirement from IBM to join National Chiao-Tung University, Hsinchu, Taiwan, as a Chair Professor with the Department of Electronic Engineering in February 2008. He has received the Outstanding Scholar Award from Taiwan's Foundation for the Advancement of Outstanding Scholarship for 2008 to 2013. He served on the Device Technology Program Committee for IEDM in 1986 and 1987, and the Program Committee for Symposium on VLSI Circuits from 1992 to 2006. He was the Publication/Publicity Chairman for Symposium on VLSI Technology and Symposium on VLSI Circuits in 1993 and 1994, and the Best Student Paper Award Sub-Committee Chairman for Symposium on VLSI Circuits from 2004 to 2006. He was elected an IEEE Fellow in 1994 "For contributions to high-performance bipolar devices, circuits, and technology." He has authored many invited papers in international journals such as *International J. of High Speed Electronics*, PROCEEDINGS OF IEEE, *IEEE Circuits and Devices Magazine*, and *Microelectronics Journal*. He has presented numerous plenary, invited or tutorial papers/talks at international conferences such as International SOI Conference, DAC, VLSI-TSA, ISSCC Microprocessor Design Workshop, VLSI Circuit Symposium Short Course, ISQED, ICCAD, APMC, VLSI-DAT, ISCAS, MTDT, WSEAS, VLSI Design/CAD Symposium, etc. He was the corecipient of the Best Paper Award at the 2000 IEEE International SOI Conference. He is the holder of 27 U.S. patents with another 14 pending. He has authored or coauthored over 270 papers.

H.-S. Philip Wong (S'81–M'82–SM'95–F'01) received the B.Sc. degree (Hons.) in electrical engineering from the University of Hong Kong, Kowloon, Hong Kong, in 1982, the M.S. degree in electrical engineering from the State University of New York at Stony Brook, in 1983, and the Ph.D. degree in electrical engineering from Lehigh University, Bethlehem, PA, in 1988.

He was with the IBM T. J. Watson Research Center, Yorktown Heights, NY, in 1988. Since September 2004, he has been with Stanford University, Stanford, CA as Professor of electrical engineering. While at IBM, he worked on CCD and CMOS image sensors, double-gate/multigate MOSFET, device simulations for advanced/novel MOSFET, strained silicon, wafer bonding, ultrathin 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, and novel memory devices.

Dr. Wong is a member of the Emerging Research Devices Working Group of the International Technology Roadmap for Semiconductors. He served on the IEEE Electron Devices Society as elected AdCom member from 2001 to 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 to 2004, and was the Chair of the Image Sensors, Displays, and MEMS subcommittee from 2003 to 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 and Solid-State Circuit Society. He has taught several short courses at the IEDM, ISSCC, Symp. VLSI Technology, SOI conference, ESSDERC, and SPIE conferences.