RC23801 (W0511-115) November 15, 2005
Electrical Engineering

IBM Research Report

Suitability of Metal Gate Stacks for Low-Power and High-Performance Applications: Impact of Carrier Confinement

Arvind Kumar, Paul M. Solomon
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Suitability of Metal Gate Stacks for Low-Power and High-Performance Applications: Impact of Carrier Confinement

Arvind Kumar, *Member*, and Paul M. Solomon, *Fellow, IEEE*

Manuscript received Dec. 1, 2005. The authors are with the IBM Semiconductor Research and Development Center, Thomas J. Watson Research Center, Yorktown Heights, NY 10598. USA (914-945-3786; fax: 914-945-2141; e-mail: arvkumar@us.ibm.com.)

***Abstract*—A simulation study is carried out to assess the competitiveness of metal gate stacks for low-power and high-performance technologies using realistic oxynitride and high-permittivity gate dielectric stacks having insulator leakages appropriate for each application. In the first part of this work, the metal gate work function is fixed at a value near midgap. For this value of work function, the performance (obtained from mixed-mode simulations of inverter delay chains) of metal gate stacks is found to exceed that of polysilicon gate stacks for low-power applications, but to be uncompetitive for high-performance applications. Both of these observations are explained by understanding the role of carrier confinement determined by the channel doping required for each application. In the second part of this work, the metal gate work function is allowed to vary in order to obtain the optimal work function ranges for each application. Metal gate stacks are shown to be especially suitable for low-power applications over a wide range of possible work functions, with optimal performance away from the band edges. For high-performance applications, work functions near the band edges yield the best performance, but significant gains compared to polysilicon-gated devices are found only when additional scaling is achieved through use of a high-permittivity gate insulator.**

***Index Terms*—Semiconductor device modeling, work function, MOS devices**

I. Introduction

Reduction of the effective gate insulator thickness is a key component of MOSFET device scaling [1]. Through elimination of the polysilicon depletion effect, metal gates are currently being extensively studied as a means to reduce effective gate insulator thickness while incurring little additional gate leakage [2-9]. Unlike degenerately doped polysilicon gate stacks, however, metal gates typically have work functions that lie within the Si band gap rather than near the band edges, with values of 4.5-4.6 eV reported for fully silicided (FUSI) gates such as $CoSi_2$ [2] and NiSi [3] and near midgap (4.7 eV) for TaSiN [6] and TiN [7]. Although Ref. [5] has offered the intriguing possibility of work function adjustment through impurity segregation, the smallest work function achieved is still shifted 0.32 eV from a conduction band-edge work function.

Since the threshold voltage is set by the active net impurity concentration in the channel region, metal gate stacks require less channel doping than polysilicon ones to achieve a given target for the leakage current in the off-state, defined to be the source-drain current under zero gate-source bias (V_{gs}=0) and drain-source bias equal to the supply voltage (V_{ds}=V_{dd}). An immediate consequence of the reduced channel doping required for in-gap metal gates is a weaker potential to confine carriers near the surface, giving rise to weaker gate control of the channel and hence to poorer short-channel effects, potentially offsetting the performance advantages of the thinner effective gate insulator. Ref. [10] has shown how a midgap work function leads to a buried channel device with poor short channel effects, and Ref. [11] has studied degraded carrier confinement in metal gate stacks using fully quantum-mechanical simulations. Although these papers have explained fundamentally how degraded carrier confinement in metal-gated devices impacts their DC device characteristics, a study comparing their AC circuit performance to that of polysilicon-gated devices – and thus investigating the tradeoff of lower effective insulator thickness with poorer doping-induced carrier confinement -- has, to our knowledge, not been carried out.

The off-state leakage current target, usually set for a device whose gate length is significantly shorter than that of the nominal gate length of a given technology, must be chosen based on the power consumption requirements of the desired application. Fig. 1 illustrates for the case of an nFET why we might intuitively expect metal gates with in-gap work functions to be particularly well-suited for low-power applications having a low off-current target. In the off-state, a barrier ϕ_{inj} is set up limiting the off-current. The magnitude of ϕ_{inj} is a function of $\Delta\phi$, the offset of the metal work function from the conduction band edge in the source contact, but for a given off-current, ϕ_{inj} is approximately a constant and is larger the smaller the current. For $\Delta\phi = \phi_{inj}$ (Fig. 1(b)) the channel transverse field is zero (neglecting two-dimensional effects) so that for larger ϕ_{inj} (smaller off-current) than this (Fig. 1(c)) a positive confinement is obtained. Thus, the low-power case lends itself to the use of a higher work function gate metal. Furthermore, the reduction of the transverse field concomitant with the larger work function should result in higher channel mobility and improved performance for the low-power case.

In this work we investigate the competitiveness of metal-gated devices using mixed-mode simulations of inverter delay chains for both low-power and high-performance applications. Realistic oxynitride and high-permittivity (high-κ) gate insulator stacks are chosen from Ref. [9] to insure that the gate leakage is a small fraction of the source-drain leakage for each application. In the first part of this work, we fix the metal gate work function at a value typical of the widely studied fully silicided (FUSI) gates [2-5,8,9] in order to compare the impact of confinement degradation for a low-power and a high-performance technology. In the second part of this work, we sweep the range of metal gate work functions from band edge to nearly midgap in order to determine the work function ranges best suited for both low-power and high-performance applications. The twofold objectives of this work are, therefore, to elucidate what applications are well-suited for currently available in-gap metal gates and to provide valuable guidance to the industry regarding the range of work functions that would maximize performance for each application.

II. Problem Setup

Table 1 summarizes the gate insulator stacks studied in this work, consisting of a polysilicon gate (PG) or a metal gate (MG) combined with an oxynitride or a high-permittivity (high-κ) dielectric atop a bulk Si substrate. For a low-power technology, we have chosen an off-current target of I_{off}=300 pA/μm at room temperature for both nFET and pFET, with supply voltage taken to be V_{dd}=1 V. Gate stacks M1,P1 with a high-κ dielectric and M2,P2 with an oxynitride dielectric each have the same gate insulator leakage of approximately 0.1 A/cm^2, based on Ref. [9]. This gate insulator leakage corresponds to approximately 12% of our chosen source-drain leakage target. For a high-performance technology, we have increased the source-drain leakage target by a factor 1000, *i.e.*, I_{off}=300 nA/μm for both nFET and pFET, with V_{dd}=1 V. To realize insulator leakage of approximately 100 A/cm^2 at 1 V gate bias, we utilize gate stacks M3,P3 with a high-κ dielectric and M4,P4 with SiO_2 [9]. Gate stacks M3,P3 are based on extrapolations of the trend lines from lower leakage levels in Ref. [9], and other thickness combinations of SiO_2 and HfO_2 yielding the same equivalent insulator thickness t_{eq} may offer lower leakage. Note that gate stacks M4,P4, realized with SiO_2, and stacks M1,P1, realized with a high-κ dielectric stack, allow an easy comparison between the low-power and high-performance cases for the same t_{eq}. Gate stacks P1 and P3, consisting of polysilicon combined with a high-κ dielectric, are not realistic options at present due to Fermi-level pinning observed in HfO_2 and other high-κ dielectrics paired with a p-type polysilicon gate electrode [8], and are intended only to allow direct comparison to gate stacks M1 and M3, respectively.

Having determined the possible gate insulator stacks by the source-drain leakage targets of the application desired, we find the gate length using the following methodology. To set the short-channel effects, we choose a target value for the drain-induced barrier lowering (DIBL) of 145 mV/V. Using only the polysilicon gate stacks P1,P2,P3,P4, we then find the shortest gate length L_{off} such that the DIBL target can be achieved for both pFET and nFET, with the channel doping adjusted in each case through the halo implant dose such that the off-current target I_{off} is met. This procedure involves starting from an initial guess for the gate length that is too small to meet the DIBL target, and then incrementing L_{off} in successive 1 nm increments until the DIBL criterion is satisfied, recalibrating the halo implants to meet the I_{off} target at each step. The same gate length is then used for the metal gate counterparts M1,M2,M3,M4 corresponding to the polysilicon gate stacks P1,P2,P3,P4, respectively. As shown in Table 1, the use of a high-κ gate stack in place of an oxynitride one allows continued device scaling through smaller effective gate insulator thickness and corresponding gate length reduction of 5 nm.

Note that the above procedure, while convenient, does not necessarily optimize the performance of either the PG or MG case, but we did verify separately that the PG performance was close to optimum. The MG performance may not be optimum, so to this extent our results will be conservative. In addition, we expect the rolloff behavior of the PG and MG devices to be different when the same gate length is used to set the off-current target. Since we wish to focus on the tradeoff in short-channel effect degradation with smaller effective insulator thickness, we choose in this work to follow the above methodology, first comparing the performance at the same gate length L_{off}, and then to examine the implications of the different rolloff characteristics.

To evaluate performance, mixed-mode simulations of 5-stage inverter delay chains were carried out using FIELDAY with quantum-mechanical corrections [12] to accurately model electron confinement. The β-ratio (of pFET-to-nFET width) was 2:1, and the source and drain contacts were 0.30 μm long. The equivalent SiO_2 stacks shown in Table 1 (column 4) were used; however, as a check that two-dimensional effects in the gate insulator are not significant [13], case M1-κ, in which the physical high-κ stack was substituted, was found to differ insignificantly from case M1, using the same t_{eq}. Gate insulator leakage was not included, but is too low to affect significantly the AC performance studied here.

III. Near Midgap Work Function

In this Section we fix the nFET and pFET metal gate work functions using dual values of +/- 0.20 eV from midgap, which can be realized using a NiSi FUSI process [3]. We denote the work function as $\Delta\Phi$=0.36 eV, corresponding to 0.36 eV higher than a conduction band-edge work function for an nFET (as illustrated in Fig. 1); the complementary pFET has a work function 0.36 eV lower than a valence band-edge work function. Using gate stacks M1,P1 and M4,P4 which have the same equivalent insulator thickness, we show simulated potential and electron concentration profiles in Fig. 2 for off-current targets in the high-performance (I_{off} =300 nA/μm) and low-power (I_{off} =300 pA/μm) regimes. The potential profiles are in qualitative agreement with Fig. 1 and further illustrate the advantage of the in-gap metal for the low-power case. For a high-performance technology with low threshold voltage, carrier confinement is so severely degraded that the natural advantages of MG are effectively negated. In contrast, for a low-power technology, the increase in channel doping to achieve low off-current restores enough confinement to the MG device that its performance becomes competitive with the PG device. We focus first on a detailed study of this low-power regime followed by an analysis of the high-performance regime.

A. Low-power Regime

For the low-power regime, we fix the source-drain leakage target to be I_{off} =300 pA/μm and utilize gate stacks M1,P1,M2,P2 with insulator leakage 0.1 A/cm^2. As shown in Table 1, a 13% reduction in unloaded switching delay τ_U is seen for case M1 compared to P1, while case M2 shows hardly any improvement over case P2. The higher channel doping required for the shorter gate length of case M1 restores some of the degraded confinement, leading to its favorable comparison to its polysilicon counterpart, case P1. However, case P1 is not a realistic option due to Fermi level pinning. Thus, comparing case M1 (the best MG option) to case P2 (the best PG option), we obtain a delay reduction of 26%, corresponding to an improvement in speed (delay reciprocal) of 36%.

An effective inverter capacitance C_{eff}=$C_L\tau_U/(\tau_L- \tau_U)$ can be obtained by computing the switching delay τ_L in the presence of a load capacitance C_L (here taken to be C_L=1.2 fF/μm) at the output of each inverter stage. As shown in Table 1, the effective capacitance is comparable for PG and MG despite the effectively thinner gate insulator for MG. To gain further insight into these results, Fig. 3(a) summarizes the individual nFET and pFET small-signal capacitance components using the conventional measure of gate capacitance, C_{gg}, defined as gate capacitance $C_g(V_{gs})$ at gate voltage V_{gs}=V_{dd}=1 V. The much higher on-state gate capacitance C_{gg} of the MG is partially offset by a reduction in junction capacitance C_j which results from the lower required halo doping (overlap capacitance C_{ov} is comparable). Furthermore, as shown in Fig. 3(b), C_{gg} overstates the true capacitance penalty of the MG; due to its higher threshold voltage, $C_g(V_{gs})$ averaged over V_{gs} is only 9% higher for MG than PG. The net change in capacitance is small and slightly in favor of the metal gate. The near equality of the MG and PG C_{eff} values also means that the relative enhancements calculated here for unloaded delays are essentially unchanged in the presence of a load capacitance. If we define an effective inverter drive current using the relation $I_{eff}=C_{eff}V_{dd}/\tau_U$, we find that case M1 offers a 13% improvement over case P1. Thus, the intrinsic gain of MG over PG, based on comparing case M1 to case P1, stems almost exclusively from drive-current enhancement. Additional enhancement, based on comparing case M1 to case P2, arises from scaling of the effective insulator thickness and the gate length.

Fig. 4(a-b) compares transfer characteristics of PG (case P1) and MG (case M1) nFETs and pFETs, with parameter summary in Table 2. Although the MG devices have poorer DIBL and higher threshold voltage than the PG devices, their much higher transconductance leads to significantly higher on-currents and hence superior circuit performance. The transconductance improvement, in both linear and saturation regimes, is due not only to the higher inversion capacitance but also to higher channel mobility in the MG case. The reduced channel doping of the MG device leads to lower transverse field and hence to less surface roughness and phonon scattering [14,15]. Fig. 5 shows the greatly reduced transverse field and higher effective low-field mobility (each density-averaged in the transverse direction) at V_{gs}=1 V and V_{ds}=0.05 V for case M1 compared to case P1. The transverse field is lowered sufficiently to be removed from a surface-roughness-limited regime [14,15], resulting in a mobility boost of approximately 50%.

Up to now, we have compared MG and PG performance using the gate length L_{off} at which the off-current target was set. In reality, the total quiescent power consumption will be determined by a range of gate lengths. To compare the rolloff characteristics of MG and PG devices, we assumed a Gaussian distribution of gate lengths and also simulated, using the same halo implant conditions, devices having gate lengths of $L_{-6\sigma}=L_{off}-3\sigma$, $L_{nom}=L_{off}+3\sigma$, and $L_{+3\sigma}=L_{off}+6\sigma$, with 3σ taken to be 5 nm. As shown in Fig. 6(a-b) for nFETs, MG devices have greater variation of off-current and saturation threshold voltage with gate length than their PG counterparts, consistent with our initial expectations based on the lower halo implants of the MG devices. The leakage currents of devices with lengths between $L_{-3\sigma}$ and $L_{+3\sigma}$ are lower for MG than for PG. The higher leakage currents for gate lengths less than $L_{-3\sigma}$ are unlikely to be detrimental to circuit functionality, however, even for the statistically rare worst-case gate length of $L_{-6\sigma}$ having off-current ~30 nA/μm, and the number of these devices is small enough not to significantly impact overall chip power. To assess the performance impact, Fig. 6(c) shows the quiescent leakage current (obtained by summing the nFET and pFET leakages of a single inverter stage) plotted against the switching delay. Comparing case M1 to case P2, we see that the delay reduction of 26% for devices of length $L_{-3\sigma}$ is decreased only slightly (to 24%) for devices of length L_{nom} , with approximately one-third the leakage, and to 20% for devices of length $L_{+3\sigma}$. We therefore do not consider the poorer rolloff characteristics of the MG devices to be sufficient reason to modify our original methodology of keeping the gate lengths of the PG and MG devices equal.

B. High-performance Regime

We now turn to a high-performance technology with off-current target I_{off}=300 nA/μm, considering gate stacks M3,P3 with a high-κ gate stack and M4,P4 with a SiO_2 stack, so that the insulator leakage is approximately 100 A/cm^2 [9]. From Table 1, it is seen that the MG is slower than the PG for all cases. Contrary to the low-power case, this statement is true even when case M3, which benefits from gate length scaling, is compared to case P4. The fundamental reason for this poorer MG performance can be seen in Fig. 7, which compares PG and MG nFET transfer characteristics (the pFET case is similar). Because of the loss of confinement stemming from the lower doping [11], the DIBL increases from 143 mV/V to 193 mV/V, the saturation subthreshold swing increases from 90 mV/dec to 218 mV/dec, and drive current drops from 930 μA/μm to 847 μA/μm. Returning to Figs. 1 and 2, we see that this case results in a buried channel in the off-state [10], so that the reduction in effective insulator thickness cannot compensate for the pronounced deterioration in short-channel effects.

IV. Work Function Optimization

In this Section we vary the MG work function shift ΔΦ in order to determine the best work function for both low-power and high-performance applications. Fig. 8 shows the unloaded switching delay at gate length L_{off} as a function of work function shift using the four gate stacks in Table 1. The horizontal lines show the delay of polysilicon gate stacks P2 and P4 as benchmarks. For the low-power case, any MG with work function shift ΔΦ=0-0.36 eV is faster than PG using stack P2. The enhancement is markedly stronger for the high-κ gate stack which benefits from gate length and insulator scaling. The optimum performance for the low-power case occurs for ΔΦ=0.12-0.3 eV, where the channel doping is sufficient for adequate confinement but the transverse field is significantly lowered compared to that in a device with a band-edge metal gate. Compared to the case ΔΦ=0.36 eV studied above, the MG switching delay reduction of case M1 relative to case P2 increases from 26% to 33%, and the speed improvement increases from 36% to 49%.

For a high-performance technology, we find that the shortest delay is achieved for nearly band-edge MG, where strong confinement comparable to that of the PG case is restored. However, we note importantly that a pure band-edge work function is not required for optimal performance. Because of the competing effects of better confinement and higher transverse field as the band edge is approached, the switching delay is insensitive to work function over the range ΔΦ=0-0.18 eV. Only over this range does the performance of MG exceed that of PG for an oxide-based dielectric. For a high-κ dielectric, the additional scaling afforded by the thinner effective gate insulator allows modest gains of MG compared to PG over the range ΔΦ=0-0.3 eV. However, since the switching delay in Fig. 8 is plotted on a logarithmic scale, it is apparent that the relative gains achieved by using MG in place of PG are significantly lower for high-performance applications than for low-power ones.

V. Discussion and Conclusion

The choice of off-current target determines the extent to which short-channel effects caused by reduced confinement dominate over the beneficial effect of thinner effective gate insulator, restricting the range of acceptable metal gate work functions. As the work function moves towards midgap, the confining potential degrades until the transverse electric field at the surface reverses polarity, leading to the formation of a buried channel device with poor gate control [10]. We can thus gain insight into the results obtained above by estimating the work function that results in zero transverse field, marking the transition from a surface channel to a buried channel device. Returning to Fig. 1, we can estimate the Φ_{inj} required for each of our off-current targets by noting that the subthreshold current at temperature T depends exponentially on the injection barrier as exp(-Φ_{inj}/k_BT) for both diffusive transport and thermionic emission [16]. Using either mechanism, we find approximately Φ_{inj}~0.22 eV for I_{off} =300 nA/μm and Φ_{inj}~0.4 eV for I_{off} =300 pA/μm. Bearing in mind that $\Delta\Phi$ must be somewhat offset from these flatband conditions to have some confinement, we find that these estimates are consistent with the ranges found above for superior performance of the metal gates.

In this work we have focused on bulk devices which are well-suited for low-power applications. Use of silicon-on-insulator substrates could modify our findings in two ways. First, the reduction in junction capacitance for MG which counters the increased gate capacitance will be less pronounced. Second, when the silicon body is thinner than approximately 10 nm, the confinement induced by the buried oxide can compensate for the confinement lost from reduced doping [17]. However, any improvement in MG performance may be overshadowed by mobility degradation due to thin-silicon effects [18,19], and significant challenges remain before ultrathin silicon devices become manufacturable [19]. Finally, we note that we have not considered the possibility of mobility degradation caused by a high-κ gate dielectric stack [20] although recent work [21,22] suggests that this effect may not be too severe.

In conclusion we have found that sufficient carrier confinement is a necessary condition for in-gap metal gates to show a performance advantage over polysilicon gates. For low-power applications demanding a high threshold voltage, this condition is easily satisfied and significant performance enhancement can be achieved even for work functions far from the band edges. Additional benefits from scaling occur when a high-permittivity gate insulator is used, representing an option not available with polysilicon gate electrodes. Contrary to our initial expectations, the advantage of metal-gated devices is almost entirely due to its higher drive current with little or no penalty in effective capacitance. For high-performance applications requiring low threshold voltages, the condition of sufficient carrier confinement through doping is much harder to achieve with metal gates. Even in this case, we can obtain performance enhancements relative to polysilicon-gated devices by exploiting the scaling benefits of using a high-permittivity gate insulator. We also find, in this case, that the technologically severe requirement for a band-edge metal may be relaxed by ~0.2 eV without significant penalty.

.

ACKNOWLEDGMENT

The authors thank X. Wang and P. Oldiges for FIELDAY support, W. Haensch for discussions, and S. Fang for process conditions.

REFERENCES

[1] Y. Taur, “CMOS design near the limit of scaling,” *IBM J. Res. Develop.,* vol. 46, pp. 213-221, Mar./May 2002.

[2] B. Tavel, T. Skotnicki, G. Pares, N. Carriere, M. Rivoire, F. Leverd, C. Julien, J. Torres, and R. Pantel, “Totally silicided ($CoSi_2$) polysilicon: a novel approach to very low-resistive gate without metal CMP nor etching,” *IEDM Tech. Dig*., 2001, pp. 825-828.

[3] W.P. Maszara, Z. Krivokapic, P. King, J.-S. Goo, and M.-R. Lin, “Transistors with dual work function metal gates by single full silicidation (FUSI) of polysilicon gates,” *IEDM Tech. Dig.*, 2002, pp. 367-370.

[4] Z. Krivokapic, W. Maszara, K. Achutan, P. King, J. Gray, M. Sidorow, E. Zhao, J. Zhang, J. Chan, A. Marathe, and M-R. Lin, “Nickel silicide metal gate FDSOI devices with improved gate oxide leakage,” *IEDM Tech. Dig*., 2002, pp. 271-274.

[5] J. Kedzierski, D. Boyd, P. Ronsheim, S. Zafar, J. Newbury, J. Ott, C. Cabral, M. Ieong, and W. Haensch, “Threshold voltage control in NiSi-gated MOSFETs through silicidation induced impurity segregation (SIIS),” *IEDM Tech. Dig*., 2003*,* pp. 315-318.

[6] A. Vandooren, A.V.Y. Thean, Y. Du, I. To, J. Hughes, T. Stephens, M. Huang, S. Egley, M. Zavala, K. Sphabmixay, A. Barr, T. White, S. Samavedam, L. Mathew, J. Schaeffer, D. Triyoso, M. Rossow, D. Roan, D. Pham, R. Bai, B.-Y. Nguyen, B. White, M. Orlowski, A. Duvallet, T. Dao, and J. Mogab, “Mixed-signal performance of sub-100 nm fully-depleted SOI devices with metal gate, high K (HfO_2) dielectric and elevated source/drain extensions,” *IEDM Tech. Dig*., 2003*,* pp. 975-977.

[7] S. Datta, G. Dewey, M. Doczy, B.S. Doyle, B. Jin, J. Kavalieros, R. Kotlyar, M. Metz, N. Zelick, and R. Chau, "High mobility Si/SiGe strained channel MOS transistors with HfO_2/TiN Gate Stack," *IEDM Tech. Dig*., 2003, pp. 653-656.

[8] K.G. Anil, A. Veloso, S. Kubicek, T. Schram, E. Augendre, J-F. de Marneffe, K. Devriendt, A. Lauwers, S. Brus, K. Henson, and S. Biesmans, "Demonstration of fully Ni-silicided metal gates on HfO_2 based high-k gate dielectrics as a candidate for low power applications," *2004 Symp. VLSI Tech. Dig. Papers*, pp. 190-191.

[9] E. Gusev, C. Cabral, B.P. Linder, Y-H. Kim, K. Maitra, E. Cartier, H. Nayfeh, R. Amos, G. Biery, N. Bojarczuk, A. Callegari, R. Carruthers, S.A. Cohen, M. Copel, S. Fang, M. Frank, S. Guha, M. Gribelyuk, P. Jamison, R. Jammy, M. Ieong, J. Kedzierski, P. Kozlowski, V. Ku, D. Lacey, D. LaTulipe, V. Narayanan, H. Ng, P. Nguyen, J. Newbury, V. Paruchuri, R. Rengarajan, G. Shahidi, A. Steegen, M. Steen, S. Zafar, and Y. Zhang, "Advanced gate stacks with fully silicided (FUSI) gates and high-κ dielectrics: Enhanced performance at reduced gate leakage," *IEDM Tech. Dig*., 2004, pp. 79-82.

[10] Y. Abe, T. Oishi, K. Shiozawa, Y. Tokuda, and S. Satoh, "Simulation study on comparison between metal gate and polysilicon gate for sub-quarter-micron MOSFETs," *IEEE Elec. Dev. Lett*., vol. 20, pp. 632-634, Dec. 1999.

[11] A. Kumar and R.H. Dennard, "Effect of Metal Gate Work Function on Quantum Confinement in UTSOI Devices," *IEEE Elec. Dev. Lett.,* submitted for publication.

[12] M. Ieong, J. Johnson, S. Furkay, and P. Cottrell, "Efficient quantum correction model for multi-dimensional CMOS simulations," *Proc. SISPAD*, 1998, pp. 129-132.

[13] D. Frank, Y. Taur, and H.-S. P. Wong, "Generalized scale length for two-dimensional effects in MOSFETs," *IEEE Elec. Dev. Lett*., vol. 19 pp. 385-387, Oct. 1998.

[14] S. Takagi, A. Toriumi, M. Iwase, and H. Tango, "On the universality of inversion layer mobility in Si MOSFETs: Part I – Effects of substrate impurity concentration," *IEEE Trans. Elec. Dev*., vol. 41, pp. 2357-2362, Dec. 1994.

[15] S. Villa, "A physically-based model of the effective mobility in heavily-doped n- MOSFETs," *IEEE Trans. Elec. Dev*., vol. 45, pp. 110-115, Jan. 1998.

[16] S.M. Sze, *Physics of Semiconductor Devices*, 2nd edition, J. Wiley and Sons, New York, pp. 254-259, 1981.

[17] V.P. Trivedi and J.G. Fossum, "Scaling fully depleted SOI CMOS," *IEEE Trans. Electron Devices*, vol. 50, no. 10, pp. 2095-2103, Oct. 2003.

[18] D. Esseni, M. Mastrapasqua, G.K. Celler, F.H. Baumann, C. Fiegna, L. Selmi, and E. Sangiorgi, "Low field mobility of ultra-thin SOI N- and P-MOSFETs: Measurements and implications on the performance of ultra-short MOSFETs," *IEDM Tech. Dig*., 2000, pp. 671-674.

[19] B. Doris, M. Ieong, H. Zhu, Y. Zhang, M. Steen, W. Natzle, S. Callegari, V. Narayanan, J. Cai, S.H. Ku, P. Jamison, Y. Li, Z. Ren, V. Ku, D. Boyd, T. Kanarsky, C. D'Emic, M. Newport, D. Dobuzinsky, S. Deshpande, J. Petrus, R. Jammy, W. Haensch, "Device design considerations for ultra-thin SOI MOSFETs," *IEDM Tech. Dig.*, 2003, pp. 631-634.

[20] M.V. Fischetti, D.A. Neumayer, and E.A. Cartier, "Effective electron mobility in Si inversion layers in metal-oxide-semiconductor systems with a high-κ insulator: The role of remote phonon scattering," *J. Appl. Phys.*, vol. 90, pp. 4587-4608, Nov. 2001.

[21] R. Chau, S. Datta, M. Doczy, B. Doyle, J. Kavalieros, and M. Metz, "High-κ/metal-gate stack and its MOSFET characteristics," *IEEE Elec. Dev. Lett.*, vol. 25, pp. 408-410, Jun. 2004.

[22] V. Narayanan, E. Cartier, K. Maitra, B.P. Linder, V.K. Paruchuri, E.P. Gusev, P. Jamison, M.L. Steen, D. La Tulipe, J. Arnold, K. L. Lee, R. Carruthers, D.L Lacey, A. Callegari, and R. Jammy, "Process optimization for high electron mobility in aggressively scaled Metal/HfO_2 stacks," submitted to *IEEE Elec. Dev. Lett.*

Case	Gate electrode	Gate dielectric	Equiv. SiO_2 thickness (nm)	Insulator leakage (A/cm^2)	Gate length L_{off} (nm)	Unloaded delay τ_u (ps)	Eff. cap. C_{eff} (fF/μm)	Eff. drive current I_{eff} (μA/μm)
M1	metal	0.6 nm SiO_2/2.5 nm HfO_2	1.1	0.1	37	8.61	1.78	207
P1	poly	0.6 nm SiO_2/2.5 nm HfO_2	1.1	0.1	37	9.86	1.80	183
M2	metal	2.4 nm SiON	1.7	0.1	42	11.44	1.60	140
P2	poly	2.3 nm SiON	1.6	0.1	42	11.72	1.71	146
M1-κ	metal	same as M1	high-κ stack	0.1	37	8.51	1.74	204
M3	metal	0.6 nm SiO_2/1.0 nm HfO_2	0.8	100	32	5.49	1.87	340
P3	poly	0.6 nm SiO_2/1.0 nm HfO_2	0.8	100	32	4.72	1.87	397
M4	metal	1.1 nm SiO_2	1.1	100	37	6.27	1.79	285
P4	poly	1.1 nm SiO_2	1.1	100	37	5.21	1.83	351

Table 1: Gate dielectric stacks used in this work. Metal gate cases (M1,M2,M3,M4) have D F = 0.36 eV. Gate length L_{off} was chosen to keep DIBL of polysilicon-gated devices delow 145 mV/V while maintaining off-current requirement. Switching delay for unloaded 5-state inverter delay chains (t_U) is shown along with effective inverter capacitance $C_{eff} = C_L t_U/(t_L - t_U)$ and effective inverter drive current $I_{eff} = C_{eff} V_{dd} / t_U$ (t_L is the delay with a load capacitance, here taken to be 1.2 fF/1m).

Case	V_{tlin} (V)	DIBL (mV/V)	SS_{sat} (mV/dec)	G_{mlin} (μS/μm)	G_{msat} (μS/μm)	I_{dsat} (μA/μm)
M1, nFET	0.56	144	92	374	1240	568
P1, nFET	0.52	134	85	276	1030	522
M1, pFET	-0.50	125	94	133	613	283
P1, pFET	-0.41	93	82	87	444	239

Table 2: Linear threshold voltage (V_{tlin}), DIBL, saturation subthreshold swing (SS $_{sat}$), linear (G_{mlin}) and saturation (G_{msat}) transconductance, and on-current (I_{dsat}) for nFETs and pFETs for cases M1 and P1.

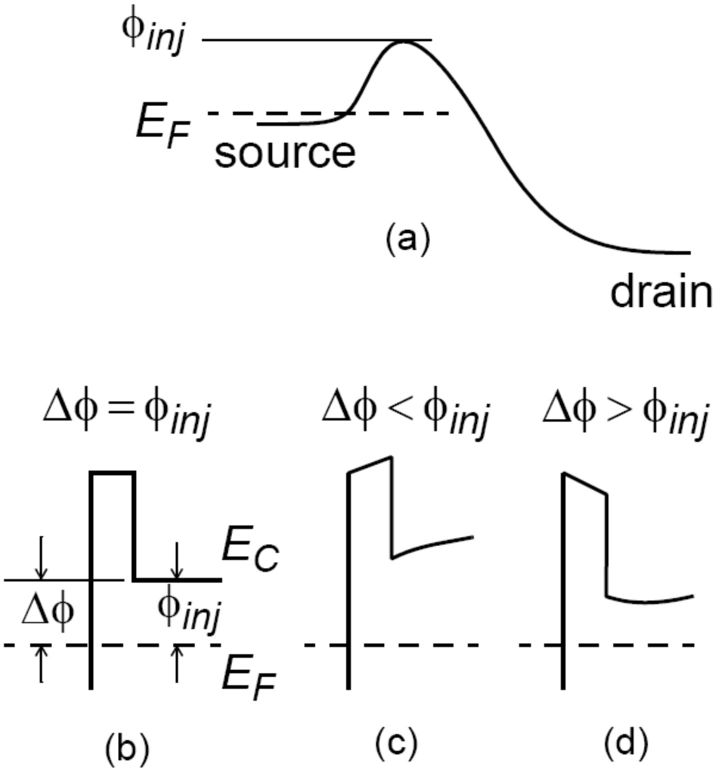
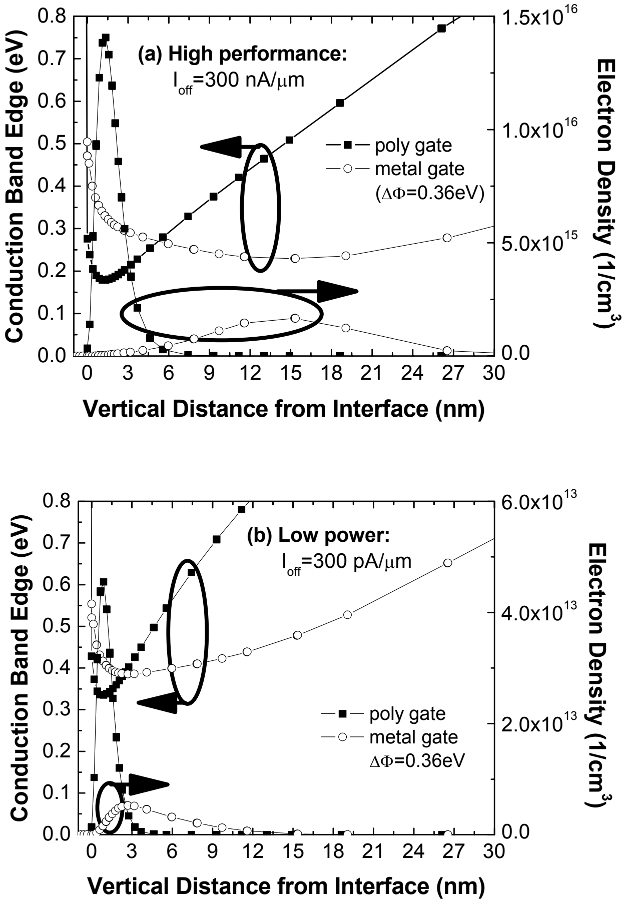

ϕ_{inj}
E_F
source
(a)
drain
$\Delta\phi = \phi_{inj}$
$\Delta\phi < \phi_{inj}$
$\Delta\phi > \phi_{inj}$
E_C
$\Delta\phi$
ϕ_{inj}
E_F
(b)
(c)
(d)

Fig. 1. (a) Energy band diagram for nFET from source (with Fermi energy E_F) to drain showing energy barrier v$_{inj}$ in the off-condition. (b) Notation used in this work such that the work function shift of the gate electrode from a band-edge work funtion is denoted as D F and the barrier seen by carriers injected from the source contact is denoted as v$_{inj}$. The case D F = v$_{inj}$ results in no confining field in the off-state (c) The case D F < v$_{inj}$ results in a surface channel in the off-state. (d) The case D F > v$_{inj}$ results in a buried channel in the off-state.



Comparison of transverse electron confinement in poly-and metal-gated nFETs for an off-current target in the (a) high-performance and (b) low-power regime. Cuts are taken in the off-condition at the peak of the potential barrier along the channel.

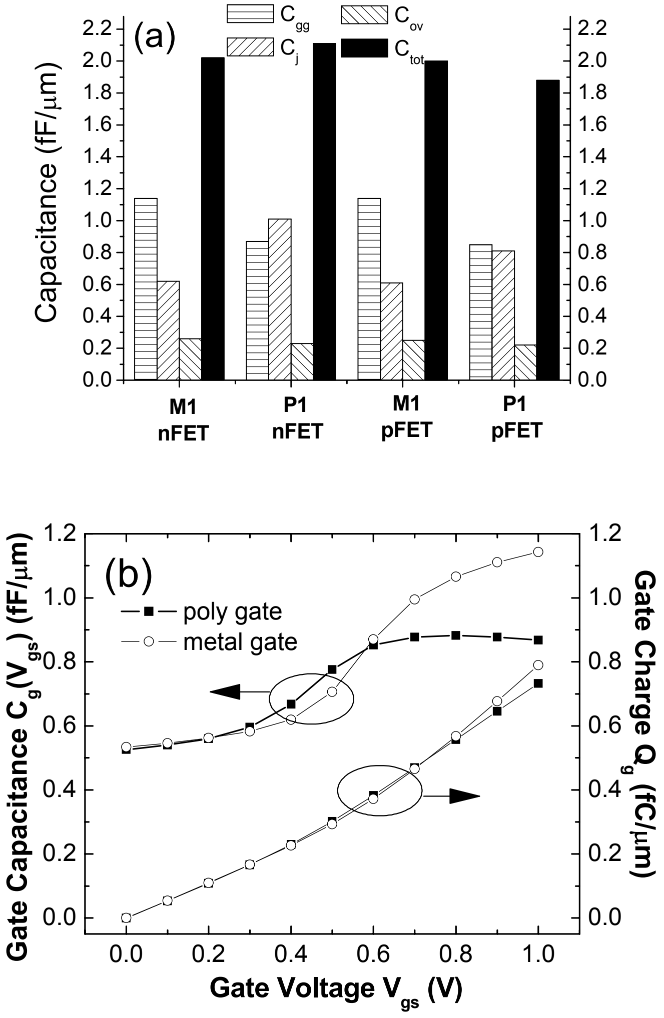


Fig. 3: (a) On-state gate capacitance (C_{gg}), overlap capacitance (C_{ov}), junction capacitance (Cj), and total capacitance C_{tot}=C_{gg} + C_{ov} + C_j for nFETs and pFETs for cases M1 andP1. (b) Gate capacitance and gate charge as a function of gate voltage for poly- and metal-gated nFETs. Although metal-gate C_{gg} (C_g and V_{gs} = 1V) is 32% highr than poly-gate C_{gg}, integrated capacitance (charge) is only 9% higher.

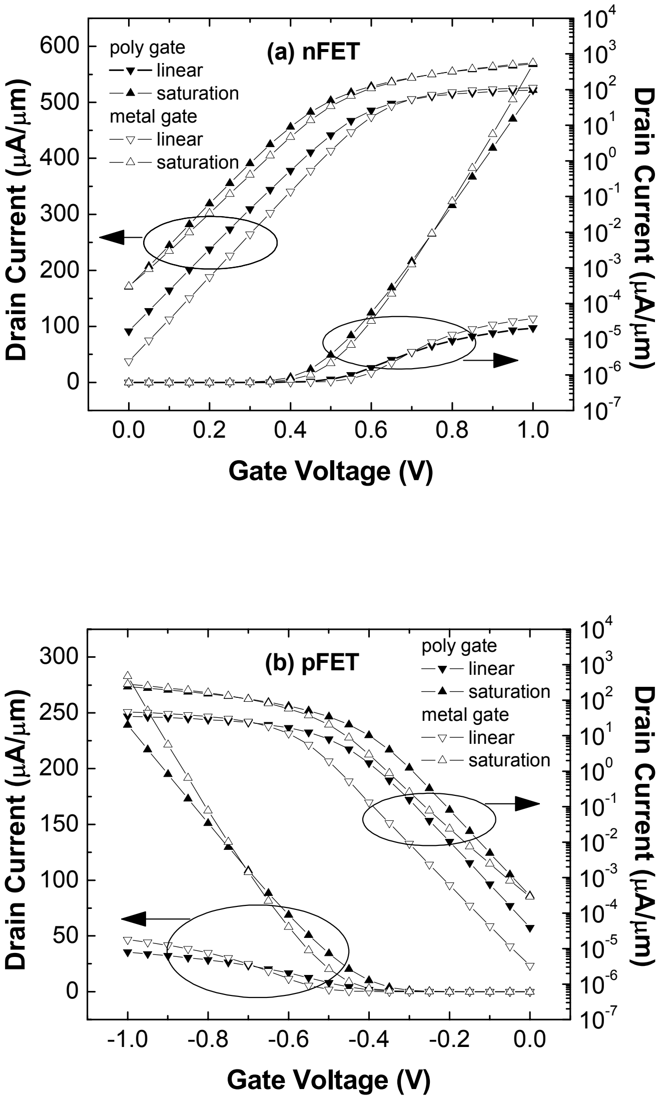


Fig. 4: Linear (V_{ds} = 0.05 V) and saturation (V_{ds} = 1.0 V) transfer characteristics for (a) nFETs and (b) pFETs with gate stacks M1 and P1 using low-power off-current target.

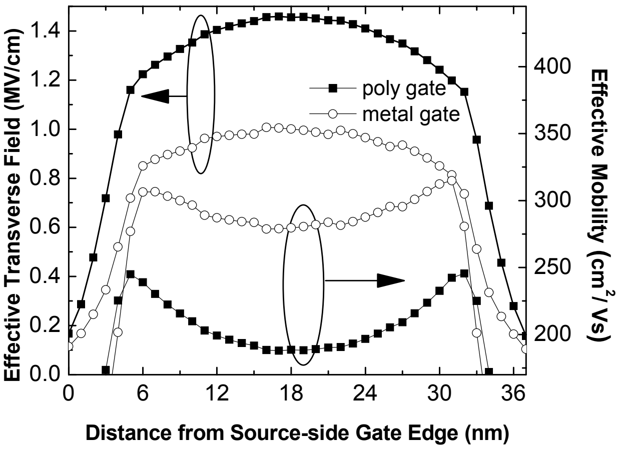


Fig. 5: Effective transverse field and effective low-field mobility (density-averaged in the transverse direction) as a function of position along the channel, at gate voltage $V_{gs} = V_{dd}$ = 1V and drain bias V_{ds} = 0.05 V. Transverse field is reduced, and low-field mobility correspondingly increased, for metal-Gated nFET compared to polysilicon-gated nFET.

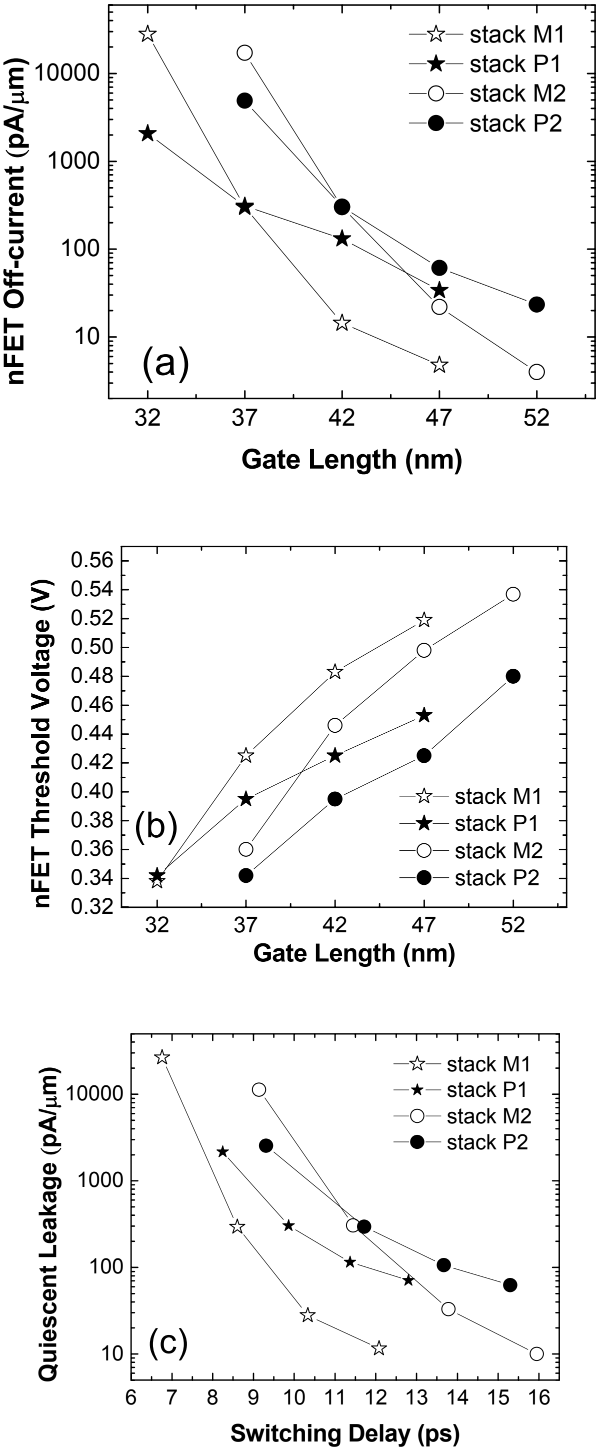


Fig. 6: Comparison of gate length dependence of (a) off-current and (b) saturation threshold voltage for polysilicon- and metal-gateed devices. (c) Comparison of quiescent leakage current vs. switching delay characteristic for inverters with polysilicon- and metal-gated devices. From left to right, points correspond to gate lengths $L_{-6\mathrm{r}} = L_{-3\mathrm{r}} - 3\mathrm{r}$, $L_{-3\mathrm{r}}$, $L_{nom} = L_{-3\mathrm{r}} + 3\mathrm{r}$, and $L_{+3\mathrm{r}} = L_{-3\mathrm{r}} + 6\mathrm{r}$ in a Gaussian distribution centered about L_{nom} with standard deviation r and 3r taken to be 5 nm.

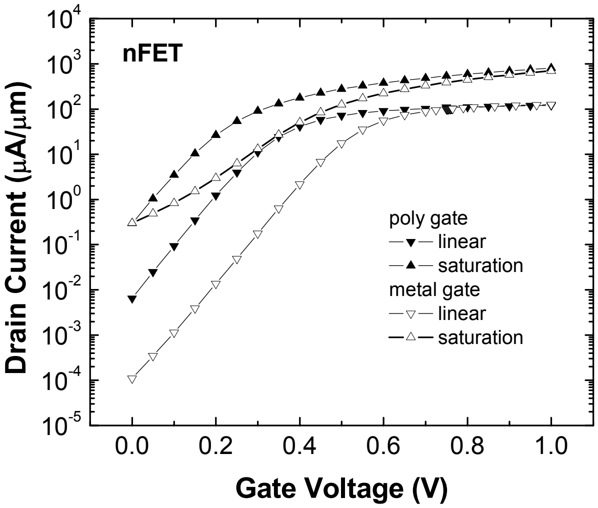


Fig. 7: Linear (V_{ds} = 0.05 V) and saturation (V_{ds} = 1.0 V) transfer characteristics for nFETs with gate stacks M4 and P4 using high-performance off-current target.

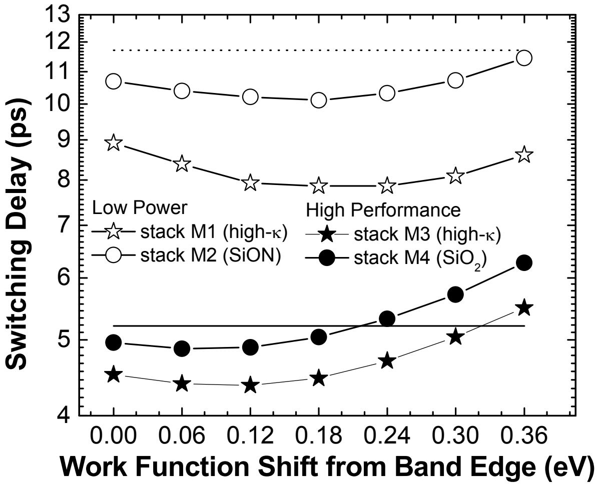


Fig. 8: Switching delay as a function of work funciton shift for low-power and high-performance cases. The optimum work function range is different due to the competing effects of enhanced confinement and mobility degradation at high transverse field. The dashed and solid lines are benchmarks for the low-power and high-performance cases using polysilicon-gate cases P2 and P4, respectively.