**RC24008 (W0607-060) July 18, 2006**
**Physics**

# IBM Research Report

## Simulation Study of a Metal / High-k Gate Stack for Low-Power Applications

**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**

# Simulation Study of a Metal / High-κ Gate Stack for Low-Power Applications

Arvind Kumar and Paul M. Solomon
IBM Semiconductor Research and Development Center (SRDC), T.J. Watson Research Center,
P.O. Box 218, Yorktown Heights, NY 10598; email: arvkumar@us.ibm.com, phone 914-945-3786, FAX: 914-945-2141

*Abstract*— **The performance of a near-midgap metal/high-κ gate stack for low-power applications is compared to that of a polysilicon/oxynitride gate stack using mixed-mode simulations. Realistic gate insulator stacks having leakages which are a small fraction of the source-drain leakage are used. In the first part of the study, the gate lengths are chosen based on fixed DIBL, and the performance of metal gate stacks is found to significantly exceed that of polysilicon gate stacks for low-power applications, but with poorer rolloff due to lower halo doping. In the second part of the study, the metal gate stack is allowed to have a longer gate length in order to match the rolloff of the polysilicon gate stack. In this case the impact of the longer gate length on performance is found to be surprisingly weak. The metal gate stack is found to have a performance advantage even when the effect of mobility degradation due to the high-κ stack is included. Finally, an ultralow power case is considered, where the lower halo doping of the metal gate stack leads to lower junction leakage.**

## I. INTRODUCTION

Conventional polysilicon gates (PG) suffer from carrier depletion which limits the effective scaling of the gate insulator stack. As a result, metal gates (MG) are currently of great interest as a means to continue device scaling through reduced effective gate insulator thickness [1-6]. Their competitiveness relative to PG, however, depends strongly on the MG work function as well as the off-state leakage target ($I_{soff}$) of the application desired [7,8]. If a near-midgap gate is used for a high-performance application, the required reduction in channel doping to attain the same $I_{soff}$ leads to a buried channel device due to the weak confining field at the surface [7-9]. On the other hand, the intrinsically high threshold voltage ($V_t$ ) of near-midgap gates makes them well-suited for low-power applications [8]. Fig. 1 illustrates this basic concept by showing how a low-power application ($I_{soff}$ =300 pA/μm) with near-midgap MG requires channel doping comparable to that of a high-performance one ($I_{soff}$ =300 nA/μm) with PG, thereby restoring the confinement needed to avoid a buried-channel device.

The purpose of this work is to study the competitiveness of a near-midgap metal/high-κ gate stack for a low-power application using mixed-mode simulations of inverter delay chains. We address key design issues including (1) choice of gate stack based on leakage requirements; (2) matching of rolloff characteristics between PG and MG; (3) effect of potential mobility degradation due the high-κ gate stack; and (4) junction leakage for ultralow power requirements.

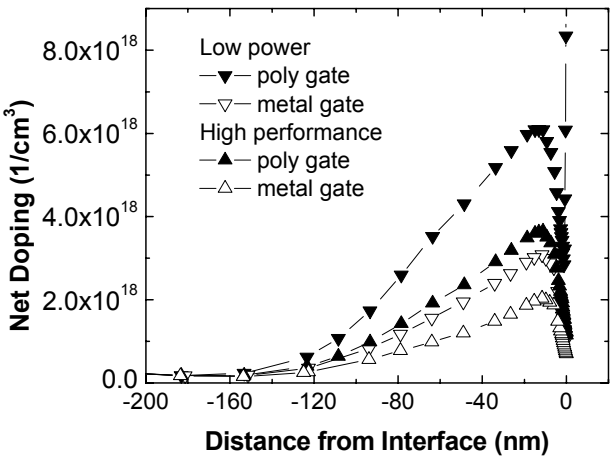


**Figure 1:** Comparison of channel doping for polysilicon and near-midgap metal gates required to meet high-performance and low-power off-state leakage targets. Note that the doping for a metal gate for a low-power application is comparable to that for polysilicon gate for a high-performance application so that the confinement is restored at the lower off-current target.

## II. SIMULATION RESULTS

Table 1 shows the three gate stacks on bulk Si studied in this work. We consider a low-power technology with an off-state leakage target of $I_{soff}$=300 pA/μm for both nFET and pFET, at $V_{dd}$=1.0 V and 27C. The gate stacks have been chosen to each have a gate leakage of ~12% of $I_{soff}$, or ~0.1 A/cm$^2$ based on Ref. [6]. Cases M1,M2 have workfunctions shifted 0.2 eV from midgap towards the conduction (valence) band edge for nFET (pFET). The gate lengths of cases P (L(nom)=47 nm) and M1 (L(nom)=42 nm) were chosen using the following method [8]. The gate length L is increased from a starting guess for L which is too small and the DIBL is monitored. The gate length is then successively increased in 1 nm increments until a DIBL target of 145 mV/V is achieved, with the halo dose adjusted in each case to achieve the target $I_{soff}$. The 5 nm difference in L reflects the thinner effective gate insulator of case M1. As discussed later, such a procedure results in different rolloff behaviors of the threshold voltage with L. Case M2 is intended to study the effect of increasing the gate length to match the rolloff of case P.

| case | Gate electrode | Gate dielectric | Equiv. $SiO_2$ thickness (nm) | L(-6σ) (nm) | L(-3σ) (nm) | L(nom) (nm) | L(+3σ) (nm) |
|---|---|---|---|---|---|---|---|
| P | poly | 2.3 nm SiON | 1.6 | 37 | 42 | 47 | 52 |
| M1 | metal | 0.6 nm $SiO_2$/2.5 nm $HfO_2$ | 1.1 | 32 | 37 | 42 | 47 |
| M2 | metal | 0.6 nm $SiO_2$/2.5 nm $HfO_2$ | 1.1 | 45 | 50 | 55 | 60 |

**Table 1:** Gate dielectric stacks used in this work. The stacks are chosen to have gate leakage ~0.1 A/cm$^2$, about 12% of the off-current leakage target. Also shown are the gate lengths simulated for each stack, chosen using method in Ref. [8] for P and M1. Gate length for case M2 is chosen to match rolloff of case P.

| case | Eff. cap. $C_{eff}$ (fF/μm) | Eff. drive current $I_{eff}$ (μA/μm) | nFET Vtlin (V) | nFET DIBL (mV/V) | nFET SSsat (mV/dec) | nFET Gmlin (μS/μm) | nFET Gmsat (μS/μm) | nFET Idsat (μA/μm) |
|---|---|---|---|---|---|---|---|---|
| P | 1.71 | 146 | 0.538 | 142 | 89 | 248 | 841 | 417 |
| M1 | 1.78 | 207 | 0.561 | 144 | 92 | 374 | 1240 | 568 |
| M2 | 1.98 | 220 | 0.490 | 113 | 85 | 531 | 1230 | 598 |

**Table 2:** 5-stage inverter delay chains with nFET and pFET widths of 45 nm and 90 nm, respectively, were simulated. Delay chains are either unloaded ($\tau_U$) or loaded ($\tau_L$) with capacitance $C_L$=1.2 fF/μm. Effective inverter capacitance is obtained from $C_{eff}=C_L\tau_U/(\tau_L-\tau_U)$ and effective inverter drive current is $I_{eff}= C_{eff} V_{dd} / \tau_U$ . Basic device properties of the nFETs at gate length L(-3σ) are also compared.

To compare performance, 5-stage inverter delay chains (β-ratio of 2) were simulated using mixed-mode FIELDAY with quantum-mechanical corrections [10] to accurately model carrier confinement. Fig. 2 shows the main result of this work, plotting quiescent leakage current ($I_{ddq}$) as a function of unloaded delay per stage, for the three cases. As compared to case P, case M1 offers a delay reduction of about 25% over the gate length range between L(-3σ) and L(nom).

Although case M1 offers substantial speedup compared to case P, it does so with markedly poorer rolloff which arises from the lower halo doping. Case M2 offers one possible solution, in which the gate length is increased by 13 nm, until its rolloff in $I_{ddq}$ matches that of case P. *Somewhat surprisingly, there is only a slight (~1%) performance penalty for this increase in gate length.*

Table 2 shows the effective inverter capacitance ($C_{eff}$) and effective inverter drive current ($I_{eff}$) extracted from unloaded vs. loaded delay. The speedup is primarily due to the high increase in drive current. Case M1 has a $C_{eff}$ only slightly higher than case P because of its higher $V_t$ as well as its lower junction capacitance which offsets the gate capacitance increase [8]. The transfer curves of the nFET shown in Fig. 3, with key parameters in Table 2, confirm the drive current gains seen for MG. Furthermore, the reduced DIBL of case M2 compared to M1 leads to a higher FET on-current ($I_{dsat}$), consistent with the load capacitance analysis which finds higher $I_{eff}$ (offset by higher $C_{eff}$) for case M2.

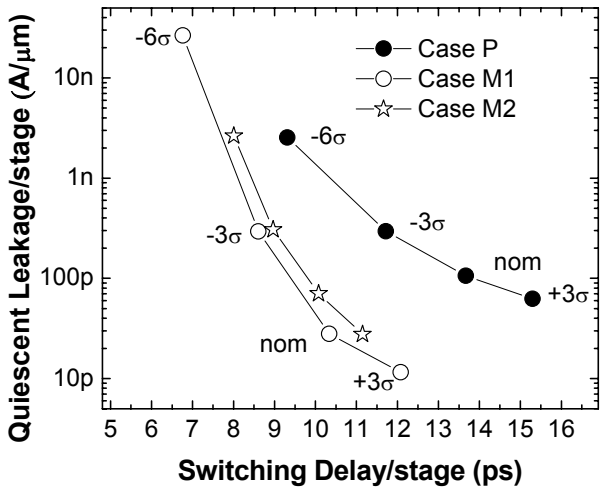


Figure 2: Comparison of quiescent leakage current vs. delay characteristic, showing significant advantage for metal gate cases. When the metal gate length is increased with appropriate adjustment in halo dose, only minor degradation in performance is observed.

High-κ gate stacks are thought to be more susceptible to soft optical phonon scattering than oxynitride ones, possibly leading to an intrinsic mobility degradation [11]. To model this effect, we repeated our simulations of case M2, reducing both the phonon component of the mobility and the saturation velocity by the same percentage. Figure 4

shows the effect of mobility degradation by comparing case P to case M2. Mobility reductions of 10% and 20% result in delay increases of 8% and 19%, respectively. However, significant speedup relative to case P is still observed.

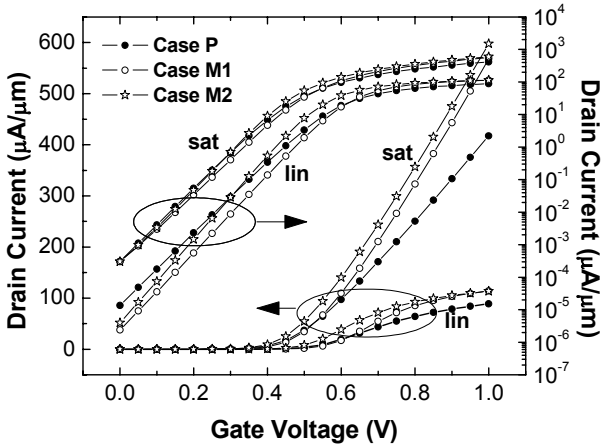


**Figure 3:** Comparison of nFET transfer characteristics using devices with gate length L(-3σ). Key parameters are summarized in Table 2.

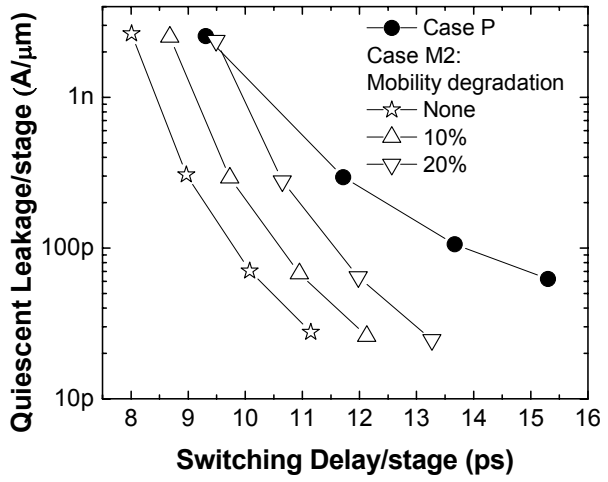


**Figure 4:** Effect of mobility degradation on quiescent leakage current vs. delay characteristic. Even with mobility degradation, near-midgap metal gate with high-κ dielectric is seen to outperform polysilicon gate with oxynitride dielectric.

The advantages of a metal/high-κ gate stack have been demonstrated using a relatively modest off-state leakage target of $I_{soff}$=300 pA/μm. For ultralow power applications, the junction leakage current ($I_j$) becomes a limiting factor. To understand its impact, we repeated our study using $I_{soff}$=10 pA/μm and $V_{dd}$=1.2 V, increasing the EOT of each gate stack by 0.1 nm to further reduce gate leakage by approximately a factor 3. Figure 5 shows the quiescent drain current leakage (including the junction leakage) as a function of switching delay. The lower doping level of the MG cases results in a significant reduction in junction leakage current. The performance advantage of the MG compared to PG exceeds that of the higher $I_{soff}$ target: case M2 has a delay reduction of about 40% for gate lengths between L(-3σ) and L(nom), compared to case P.

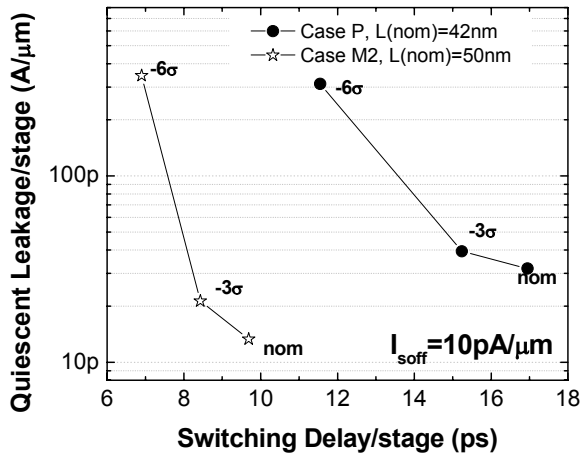


**Figure 5:** Comparison of quiescent leakage current vs. delay characteristic for the case of $I_{soff}$=10 pA/μm, $V_{dd}$=1.2 V. The metal gate cases offer not only significant performance advantage but also lower junction leakage.

## III. DISCUSSION AND CONCLUSION

We now discuss the important point that in-gap work functions [1-6] not only offer an advantage over polysilicon gates but are in fact well-suited for low-power applications. Metal gates with sub-band-edge work functions offer a performance advantage only when a buried-channel device with poor gate-channel coupling is avoided. We can thus gain insight by estimating the work function corresponding to zero transverse field, marking the transition from a surface channel to a buried channel device, as a function of the off-state leakage target. Figure 6 shows a band diagram for injection over a barrier from the source contact in the off-condition. The subthreshold current at temperature $T$ depends exponentially on the injection barrier as $\exp(-\Phi_{inj}/k_BT)$ for both diffusive transport and thermionic emission [12]. Using either mechanism, we find approximately $\Phi_{inj}=\Delta\Phi\sim0.4$ eV for $I_{soff}$ =300 pA/μm and $\Phi_{inj}=\Delta\Phi\sim0.48$ eV for $I_{soff}$ =10 pA/μm, corresponding to work functions shifted by 0.15 eV and 0.07 eV, respectively, from midgap.

The optimal work function range seeks an intermediate confinement which has a strong gate-channel coupling but a

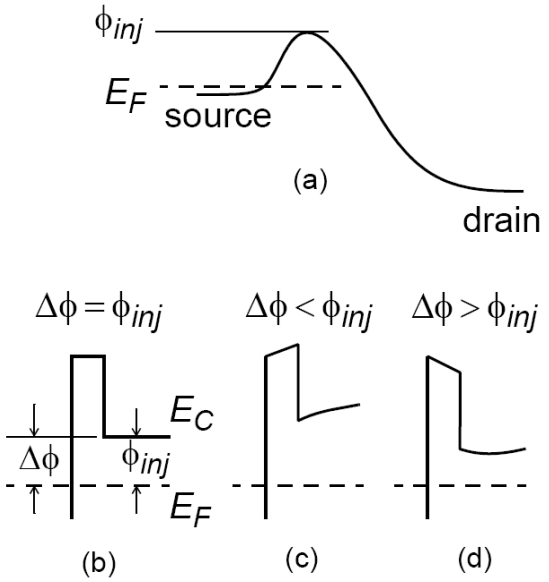


**Figure 6:** (a) Energy band diagram for nFET from source (with Fermi energy $E_F$) to drain showing energy barrier $\Phi_{inj}$ in the off-condition. (b-d) Energy band diagrams in transverse direction illustrating notation such that the work function shift of the gate electrode from a band-edge work function is denoted as $\Delta\Phi$ and the barrier seen by carriers injected from the source contact is denoted as $\Phi_{inj}$. (b) The case $\Delta\Phi=\Phi_{inj}$ results in no confining field in the off-state. (c) The case $\Delta\Phi<\Phi_{inj}$ results in a surface channel in the off-state. (d) The case $\Delta\Phi>\Phi_{inj}$ results in a buried channel in the off-state.

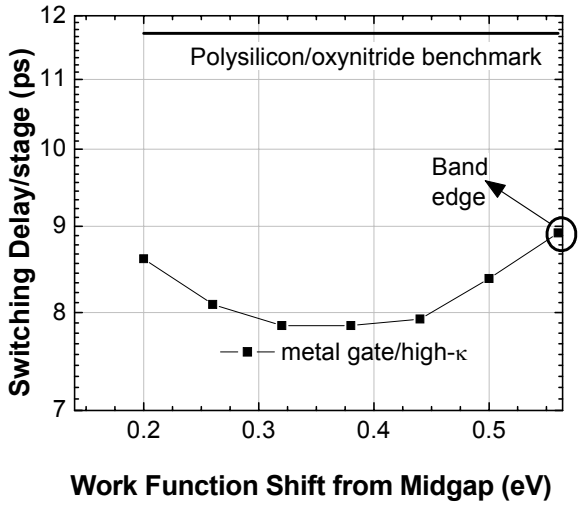


**Figure 7:** Dependence of switching delay on metal work function (shown relative to midgap) for low-power application with $I_{soff}$=300 pA/μm.

sufficiently weak transverse field that surface roughness and phonon scattering do not limit the mobility. Figure 7 shows for $I_{soff}$ =300 pA/μm the dependence of switching delay on work function shift from midgap for gate length L(-3σ), compared to a benchmark case of polysilicon/oxynitride gate stack (both nFET and pFET are shifted symmetrically). The performance is best and reasonably flat for work functions shifted from midgap by 0.25-0.45 eV, but is degraded outside this range consistent with the argument above. For a lower $I_{soff}$ target, we would expect a shift of this range towards midgap.

In conclusion, we have shown that a near-midgap metal/high-κ gate stack demonstrates strong advantages for low-power applications. Moreover, the poorer rolloff arising from lower halo doping in the metal gate case can be overcome by increasing the gate length with only minor performance penalty. The advantage is sustained even when mobility degradation is accounted for. Lowering the channel doping also benefits junction leakage, which should become a more severe limitation with increased scaling as band-to-band tunneling begins to dominate. In addition, lower channel doping reduces Coulomb scattering and helps to suppress $V_t$ fluctuations due to statistical dopant effects. Since most metal-gate stacks realized to date have in-gap work functions, we believe that low-power applications represent an excellent path to introducing metal gates into integration.

**Acknowledgment**: The authors thank X. Wang, P. Oldiges. S. Fischer, and S. Furkay for FIELDAY support; M. Khare, T. Hook, and W. Haensch for discussions; S. Fang for process conditions; and R. Miller for a critical reading of the manuscript.

**References:**

[1] B. Tavel et al., IEDM Tech. Dig. 2001, p. 825.
[2] W.P. Maszara et al., IEDM Tech. Dig. 2002, p. 367.
[3] Z. Krivokapic et al., 2003 Symp. VLSI Tech. Dig. Papers, p. 131.
[4] J. Kedzierski et al., IEDM Tech. Dig. 2003, p. 315.
[5] K.G. Anil et al., 2004 Symp. VLSI Tech. Dig. Papers, p. 190.
[6] E. Gusev et al., IEDM Tech. Dig. 2004, p. 79.
[7] Y. Abe et al., IEEE Elec. Dev. Lett., vol. 20, p. 632, 1999.
[8] A. Kumar and P.M. Solomon, IEEE Trans. Elec. Dev., vol. 53, p. 1208, 2006.
[9] A. Kumar and R.H. Dennard, unpublished.
[10] M. Ieong et al., Proc. SISPAD 1998, p. 129.
[11] M.V. Fischetti et al., J. Appl. Phys., vol. 90, p. 4587, 2001.
[12] S.M. Sze, *Physics of Semiconductor Devices*, 2nd edition, J. Wiley and Sons, New York, pp. 254-259, 1981.