Device Scaling of High Performance MOSFET with Metal Gate High-K at 32nm Technology Node and Beyond

Xinlin Wang, *Ghavam Shahidi, Phil Oldiges and Mukesh Khare
IBM Semiconductor Research and Development Center
Systems and Technology Group, Hopewell Junction, NY 12533, Research Division, *IBM T.J. Watson Research Center

Abstract — **In this work, two different methodologies are used to quantitatively evaluate devices with metal high-k gate dielectrics for their scaling benefits over conventional polysilicon gate devices. For each method, device characteristics and ring oscillator delay calculations are performed. Our results show that aggressive channel length scaling continually provides transistor performance gain with the use of metal gate high-k technology. A band edge work function for the metal gate offers potential benefits for device scaling over conventional polysilicon gates for high performance (HP) application at the 32nm CMOS technology node and beyond.**

Keywords-device scaling, metal high-k gate, channel length scaling, high performance CMOS

I. INTRODUCTION

Scaling the transistor gate length (L_{gate}), which is one of the key parameters driving MOSFET scaling, has significant performance impact at the 32nm node and beyond [1,2]. Because the gate oxide cannot be further scaled down due to large gate tunneling currents, channel length scaling without gate dielectric scaling beyond the 45nm node actually degrades transistor drive current and performance. With a high-k material as gate dielectric, effective oxide thickness (EOT) can be further scaled down without increasing gate tunneling leakage. Also, using metal as a gate electrode (MG), the polysilicon gate (PG) depletion effect is eliminated, which allows designers to push the scaling limit through a reduction of the inversion layer thickness (T_{inv}) [3,4]. There are different ways to evaluate the electrostatic and performance advantage associated with metal gate high-k technology. In our previous study [5], the off-state leakage current (I_{off}) is constrained to a fixed value for minimum channel length (L_{min}) devices at V_{dd}= 1V, then we compared metal gate high-k ((MG/HK) devices to poly oxide gate (PG/OX) devices at nominal channel length (L_{nom}). However the sub-threshold leakage (I_{off}) does not match to that of PG/OX control devices at L_{nom}. In this work, we evaluate EOT and T_{inv} scaling benefits quantitatively by two different methods: a single point methodology and double point methodology. For the single point methodology, we compared devices at nominal channel length with a fixed I_{off} target. For the double point methodology, devices are compared at both nominal and minimum channel length with the fixed I_{off} targets. By using the double point method, for the first time we show a fair performance comparison between devices with metal gate high-k dielectric and polysilicon gate oxide dielectric at a fixed total leakage current of a chip.

II. SIMULATION METHOODOLOGY

Drift diffusion simulations were performed on PDSOI MOSFETs with either PG/OX or MG/HK gate stack as showed in Fig. 1, and ring oscillator delays are calculated by FIELDAY [6] mixed-mode simulations. In this study, we focus on a band-edge metal gate which is required for higher performance applications [5]. NFET devices are simulated with L_{nom} ranging from 23nm to 35nm, T_{inv} is 19A for PG/OX and 14A or12A for MG/HK as listed in the table 1. In order to de-couple the performance adder between MG NFETs and PFETs, all of the PFETs used in the ring oscillator simulations are PG/OX devices with T_{inv}=20A. Two methodologies are used in our scaling study. (I) fixed drain induced barrier lowering (DIBL) scaling at a single channel length (one point). In this method, we center the devices to satisfy the I_{off} constraint at different nominal channel lengths and different inversion layer thicknesses. Then we study how T_{inv} scaling enables channel length scaling for the fixed DIBL condition. (II) Double-point methodology, in which we assume 3σ L_{gate} variation in the process between L_{nom} and L_{min}, then we match I_{off} of the MG/HK devices to that of the PG/OX control devices at both L_{nom} and L_{min} by adjusting the doping profile, nominal channel length and T_{inv} of the MG/HK devices.

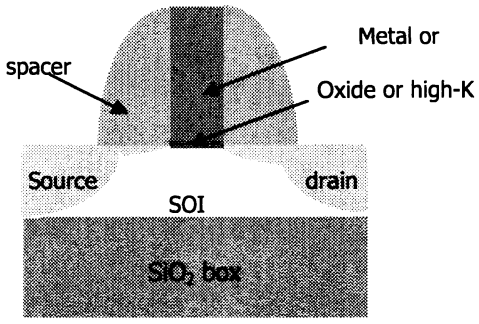

Fig. 1 Simulated PDSOI structures with different gate and dielectrics. Gate: metal or polysilicon; dielectrics: oxide or high-K.

978-1-4244-1753-7/08/$25.00 ©2008 IEEE

TABLE I. CHANNEL LENGTH AND T_{INV} FOR SIMULATED MOSFET.

	L_{nom} (nm)	T_{inv} (A)
Polysilicon/oxide (PG)	35	19
Metal/High-k (MG)	35, 31, 27, 23	14 or 12

III. SIMULATION RESULTS AND DISSCUSIONS

Short channel device properties of MG NFETs are compared to 35nm PG NFETs. I_{off} is matched between MG and PG at L_{nom} for Figs 1-6.

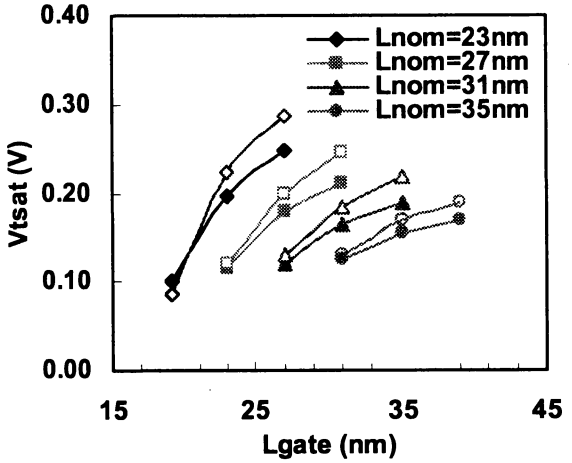

Fig. 2 *Threshold voltage (V_t) vs. L_{gate}* at different L_{nom} (V_{ds}=1.0V). At each nominal channel length, I_{off} is matched. Open symbol lines: PG (T_{inv}=19A), Solid symbol lines: MG (T_{inv} =14A). When nominal length is scaled down, Vt roll-off becomes worse. When MG and PG devices have the same EOT, MG devices provide better V_t roll-off compared to that of the PG devices at the same nominal channel length.

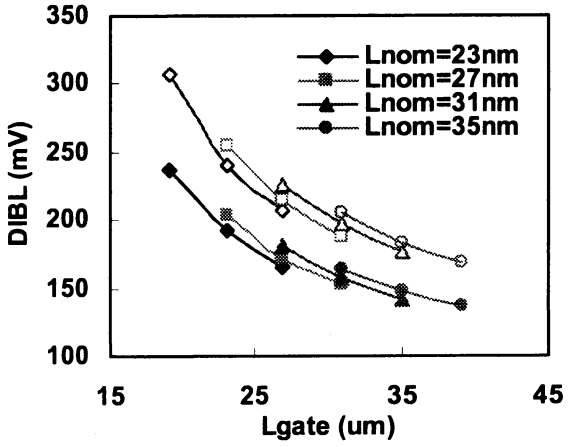

Fig. 3 *DIBL vs. L_{gate}* at different L_{nom} (V_{dd}=1.0V). Open symbol: PG; solid symbol: MG. When the nominal channel length is scaled down without EOT scaling, DIBL is degraded. Because there is different potential drop in the poly gate electrode at the sub-threshold conditions when V_{ds}=50mV or V_{dd}, MG devices show 30-50mV DIBL reduction compared to PGs at different L_{nom}. The DIBL benefit of MG over PG becomes larger when the channel length is scaled down.

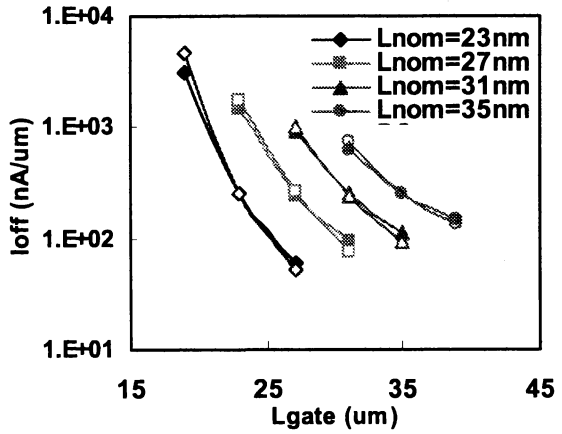

Fig. 4 *I_{off} vs. L_{gate}* at different L_{nom}. Open symbol lines: PG (T_{inv}=19A), solid symbol lines: MG (T_{inv} =14A). Compared to PG, there is little improvement for MG on the *I_{off} vs. L_{gate}* curves when EOT is not scaled. Because at the I_{off} condition (Vgs=0), there is little poly depletion at the gate electrode. So EOT has to be scaled down to improve *I_{off} vs. L_{gate}* when the channel length is scaled down.

In Figs. 2-4, T_{inv} =14A for MG and T_{inv} =19A for PG and effective oxide thickness (EOT) is the same for both MG and PG devices. During the channel length scaling without EOT scaling, V_t roll-off and DIBL become worse even though halo doses are increased to control I_{off} at the target L_{nom}. And I_{off} of minimum channel length device keeps increasing due to worse short channel effect (SCE) control at shorter channel length. From Figs. 2-3, we can see that at the same EOT, MG devices show better V_t roll-off and DIBL reduction compared to PG counterpart due to the removal of the potential drop in the poly gate electrode under the sub-threshold condition. However there is only small improvement in *I_{off} vs. L_g* for the MG over the PG when EOT is not scaled as shown in Fig. 4.

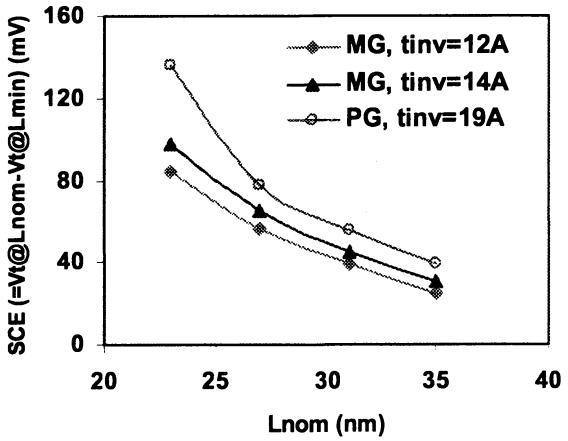

Fig. 5 SCE (=*V_t@L_{nom}-V_t@L_{min}*) *vs. L_{nom}*. Scaling channel length degrades SCE even though the halo doping concentration is increased to control I_{off} at L_{nom}. SCE degrades faster for the PG devices compared to the MG Devices at the Same EOT (compared MG 14A case to PG 19A case). SCE can be further improved by EOT scaling for MG devices (MG 12A case vs. MG 14A case).

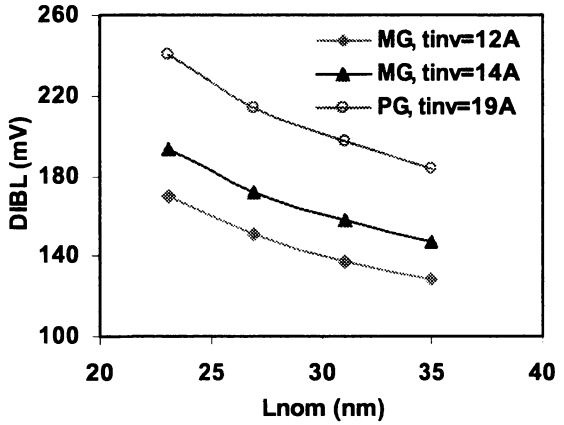

Fig. 6 *DIBL* comparison at different L_{nom}. Removal of poly depletion in the gate electrode improves DIBL at the same channel length. 1A EOT scaling for MG devices provides 3~4 nm channel length scaling for the fixed DIBL condition.

Fig. 5 shows that short channel effect (SCE) is degraded during the channel length scaling even though the halo dose increases to meet I_{off} target at shorter L_{gate}. The PG device SCE degrades faster than that of MG devices for shorter devices. Fig. 6 shows the DIBL comparison at different L_{nom} for scaled EOT (MG, Tinv =12A) and non-scaled EOT (MG, tinv=14A) cases compared to PG. For MG devices, 1A EOT scaling provides 3~4 nm channel length scaling for the fixed DIBL condition; and at the fixed L_{nom}, 1A EOT scaling provides 10mV DIBL improvement. AC performance is compared between PG and MG by running mix-mode ring oscillator simulation. Fig. 7 shows that for the loaded ring oscillator, there is ~13% performance improvement by replacing a 35nm PG NFET (T_{inv}=19A) with a 27nm band-edge MG/HK NFET (T_{inv} =12A); and 2A EOT scaling of MG/HK yields 4% AC performance benefit. If we apply MG to the PFET, the performance gain over PG will be doubled.

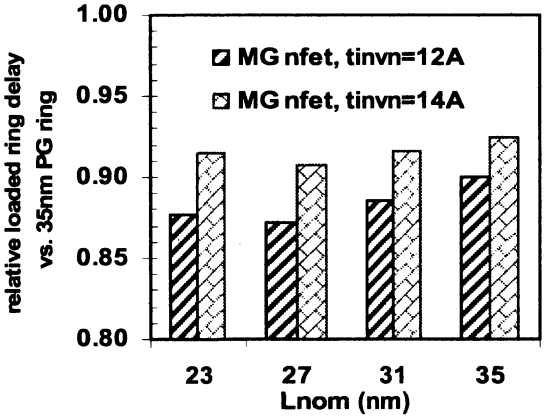

Fig. 7 Relative loaded ring oscillator delay for different L_{nom} at fixed I_{off} condition. At L_{nom}=35nm, replace PG with MG for NFET will provide 8% performance gain. Continually scaling channel length for MG devices will reduce higher effective capacitance loading penalty due to smaller Tinv for MG, so more performance benefit is obtained by L_g scaling down to 27nm for MG.

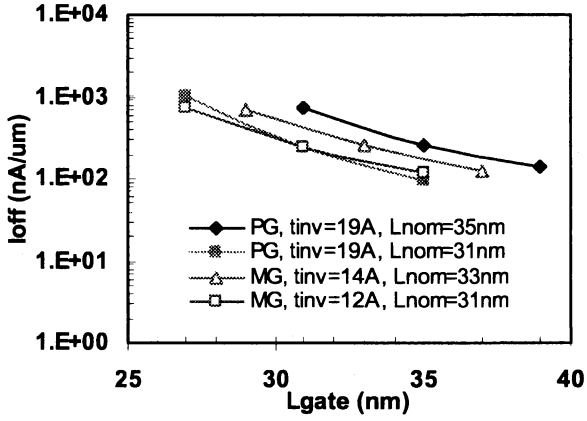

Fig. 8 I_{off} vs. L_{gate} comparison between MG and PG NFETs. MG devices provide 2nm Lg scaling with T_{inv}=14A and 4nm Lg scaling with T_{inv} =12A for the same I_{off} targets at both L_{nom} and L_{min} respectively to PG NFETs. For the PG devices, simply increasing halo dose for the shorter L_g without EOT scaling degrades I_{off} *vs.* L_g curve due to degraded SCE.

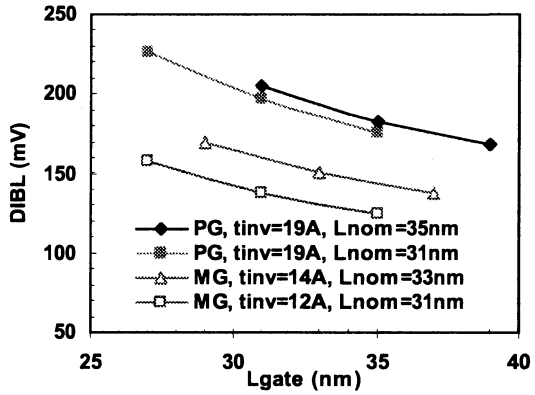

Fig. 9 DIBL *vs.* L_{gate} comparison between MG and PG NFETs. MG (T_{inv}=12A) provides more than 50mV DIBL benefit over PG (T_{inv} =19A) at 30nm channel length.

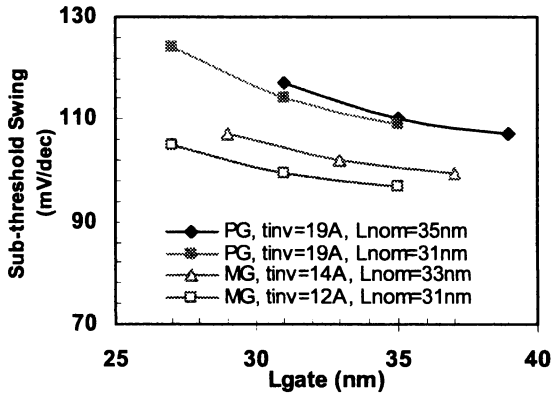

Fig. 10 Sub-threshold swing *vs.* L_{gate} comparison between MG and PG NFETs. MG (T_{inv}=12A) provides 15mV/dec sub-threshold swing (SS) benefit over PG (T_{inv} =19A) at 30nm channel length. For MG, with EOT scaling, SS is not degraded at both L_{nom} and L_{min}. However, due to lack of EOT scaling for PG, shrinking L_g degrades SS at both L_{nom} and L_{min}.

In Figs. 8-13, we used the double-point methodology to study the scaling benefit of MG over PG devices. Fig. 8 shows that scaling L_{nom} of PG device from 35nm to 31 without scaling T_{inv} doubles I_{off} at L_{min}. However, the MG device can match I_{off} at both L_{nom} and L_{min} respectively when the channel length is scaled down due to Tinv and EOT scaling. In other words, the performance is improved, but the total sub-threshold leakage of a whole chip does not increase during L_{gate} scaling for MG devices. Figs 9-11 are DIBL, sub-threshold swing and V_t roll-off curves for MG and PG. We can see that MG exhibits superior SCE control over PG. There is a 37% drive current (I_{on}) improvement for MG with T_{inv} =12A as shown in Fig. 12. However, EOT of PG can not be scaled down due to larger gate leakage, so there is no current improved when L_{gate} continually scaled.

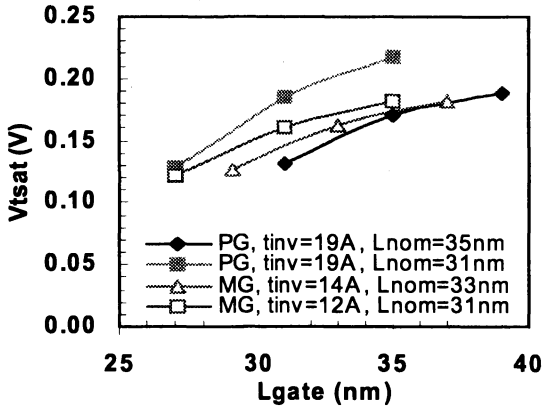

Fig. 11 V_{tsat} *vs.* L_{gate} comparison between MG and PG NFETs (V_{ds}=V_{dd}). There is Vt roll-off improvement during EOT scaling.

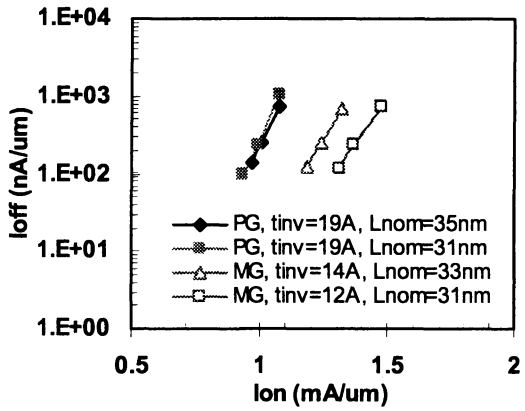

Fig. 12 I_{off} *vs.* I_{on} comparison between MG and PG NFETs. There is 24% and 37% I_{on} improvement for MG with T_{inv}=14A, L_g=33nm and T_{inv}=12A L_g=31nm respectively.

Fig. 13 shows that replacing a 35nm PG NFET (T_{inv}=19A) with a 31nm band-edge MG NFET (T_{inv} =12A), loaded ring delay is improved 12% without degrading the total Ioff of the chip.

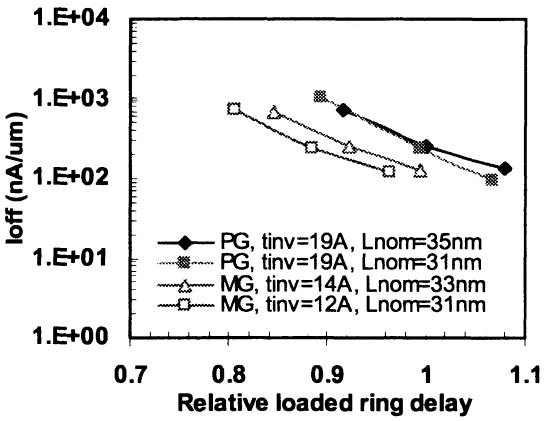

Fig. 13 I_{off} *vs.* loaded ring oscillator delay. Replacing 35nm PG NFET (tinv=19A) with 31nm BEMG/HK NFET (tinv=12A), loaded ring delay is improved 12% without degrading I_{off}. While channel length scaling on PG gate does not provide any performance gain.

IV. CONCLUSIONS

Two different methodologies are used in our simulations to quantitatively evaluate the benefits of channel length scaling together with T_{inv} and EOT scaling. Our results show that MG devices with scaled Tinv and EOT exhibit electrostatic and performance advantages over PG devices. MG/HK also provides additional channel length scaling without degrading the total I_{off} of the chip for 32nm high performance CMOS devices and beyond.

ACKNOWLEDGMENT

This work was performed by the Research Alliance Teams at various IBM Research and Development Facilities.

REFERENCE

[1] 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", IEDM Tech. Dig. 2006

[2] P. M. Zeitzoff, "Proceedings of the Custom Integrated Circuits Conference", 2004, p. 233-240.

[3] Y. Abe, T. Oishi, K. Shiozawa, Y. tokuda, and S. Satoh.," Simulation study on comparision between Metal gate and polysilicon gate for sub-quarter-micron MOSFET's", *IEEE Elec. Dev. Lett., vol. 20, No. 12,* p632-634. 1999.

[4] E. Gusev *et al.*, "Advanced Gate Stacks with Fully Silicided (FUSI) Gates and High-κ Dielectrics Enhanced Performance at Reduced Gate Leakage", IEDM Tech. Dig. 2004, p79-82.

[5] X. Wang, A. Bryant, P. Oldiges, S. Narasimha, R. Dennard and W. Haensch, "Simulation Study on Channel Length Scaling of High Performance Partially Depleted Metal Gate and Poly Gate SOI MOSFETs"SISPAD 2006, p283-286.

[6] E. M. Buturla, P. E. Cottrell, B. M. Grossman, and K. A. Salsburg, "Finite-element analysis of semiconductor device: the FIELDAY program," IBM J. Res. Develop., vol. 25, pp. 218, 1981.