Simulation Study on Channel Length Scaling of High Performance Partially Depleted Metal Gate and Poly Gate SOI MOSFETs

Xinlin Wang, Andres Bryant[†], Phil Oldiges, Shreesh Narasimha, Robert Dennard[*] and Wilfried Haensch[*]

IBM Semiconductor Research and Development Center
Systems and Technology Group, Hopewell Junction, NY 12533, [†]System and Technology Group, Essex Junction, VT, 05452
[*]Research Division, IBM T.J. Watson Research Center
Email: xinlinw@us.ibm.com Phone:(845)894-4740 Fax:(845)892-6483

Abstract— **In this work, two-dimensional numerical device simulations and 6-stage inverter chain delay calculations are done to examine whether aggressive channel length scaling continually provides transistor performance gain and whether metal gates (MG) offer potential for device scaling over poly gate (PG) for high performance (HP) applications. Our simulation show that for HP application (1) there is an optimized channel length, at which maximum performance gain is obtained both for MG and PG; (2) At short channel length regime (< 46nm), there is no performance gain of QG-MG relative to PG due to lack of carrier confinement, which result in severe sub-threshold slope degradation of QG-MG; (3) BE-MG stacks show 10% gain on a inverter delay over PG.**

Keywords: MOSFET, channel lengh scaling, poly gate depletion, metal gate, high performance, short channel effect.

I. Introduction

The transistor physical gate length (L_g) is a key parameter driving overall MOSFET scaling. For high performance (HP) logic, the scaling and the device design aim at maximizing the transistor speed. With rapid scaling of L_g, it is difficult to control short channel effect (SCE) and the mobility of the carriers in the inversion layer is much degraded due to high electric field and high doping concentration in the channel, so the transistor performance goal may not be met in the extremely scaled device. Another critical challenge as MOSFETs are scaled to deep sub-micron dimensions is polysilicon gate depletion. The preferred solution to the poly depletion is to use metal gate electrode, which is virtually no depletion and can push scaling limit through reduced the equivalent oxide thickness (EOT) [1-3]. However, generally the work function of metal gate is located near silicon mid-gap and in order to get the desired symmetric threshold voltage (Vt) for NFETs and PFETs, much less channel doping is needed compared to poly gate devices, which lead to poor SCEs due to much weaker carrier conferment [3,4]. In this work, we use simulation to investigate whether aggressive channel length scaling continually provides transistor performance gain and whether metal gates offer potential for device scaling over poly gate for HP applications.

II. Simulation Methodology

Drift-diffusion simulations were performed on Partially Depleted Silicon-On-Insulator (PDSOI) MOSFETs with either PG or MG electrode as shown in Fig. 1 and the first switch delays of a 6-stage inverter chain are calculated by mixed-mode FIELDAY device simulator with quantum-mechanical corrections to accurately model carrier confinement [5,6]. Three different nominal channel length (L_{nom}) designs are studied. Assuming 3σ L_g variation in the process, the off-state leakage current (I_{off}) is constrained to a fixed number for minimum channel length (L_{min}) devices at V_{dd}=1V as listed in table 1 by adjusting halo implant dose. Dual metal gate electrodes with two different work functions were chosen. One is quarter gap metal gate (QG-MG), whose work function is 250mV away from the silicon conduction band (E_c) or valance band (E_v) for NFETs or PFETs and the other one is band edge metal gate (BE-MG) as shown in Fig. 2.

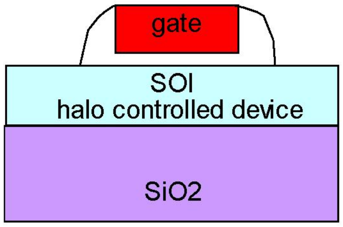

Fig.1 Simulated PDSOI structure.

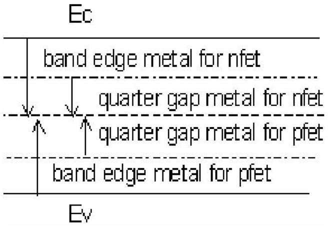

Fig. 2 work functions for metal gate NFETs and PFETs. Dual metal electrodes are used for CMOS.

Lmin(nm)	25	35	42
Lmon(nm)	28	38	46

Table. 1 L_{min} and L_{nom} for three studied design points.

Case I is referred to the widely studied "fully silicided" (FUSI) gate as reported recently [7-8], and case II is used to demonstrate the maximum advantage of MG over PG devices through EOT reduction.

1-4244-0404-5/06/$20.00 © 2006 IEEE

III. SIMULATION RESULTS AND DISCUSSION

Although both P and N devices were simulated, the DC characteristics reported here are for N devices. Fig. 3 shows the relative halo dose required to maintain I_{off} for the various design points. When MG work function moves away from band edge, halo dose has to be reduced. QG-MG device needs 30% less halo at the 25nm node and 60% less halo dose at 42nm node relative to PG case. When L_g is scaled from 42nm to 25nm, the halo dose increases 2.5X for all gate options in order to maintain reasonable SCE.

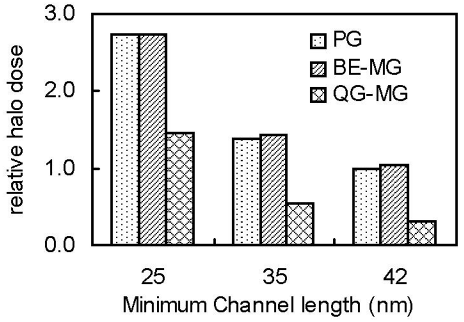

Fig. 3 Relative halo implant dose used in the NFETs for different channel length design.

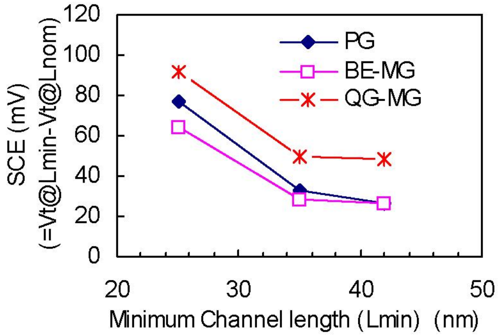

Fig. 4 SCE for different nominal channel length design points.

Fig. 4 shows that SCE becomes worse as the MG work function moves towards mid-gap. Compared to PG, because the channel doping is reduced, QG-MG FETs show worse SCE due to poor carrier confinement as shown in Fig. 5, while BE-MG devices that have channel doping profiles similar to PG, show improved SCE due to reduced EOT. When L_{nom} is scaled, SCE degrades rapidly for PG compared to MG devices, which means that using BE-MG can push the L_g scaling limit further. From Fig. 6(a) we can see that BE-MG shows sub-threshold swing (SS) benefits relative to PG, while Fig. 6 (b) shows severe SS degradation of QG-MG devices. 2-D carrier concentration profiles as shown in Fig. 7 explain why QG-MG FETs have very poor SS. In the PG case, carriers are constrained close to the top-gate. In contrast, carriers in the QG metal case are spread into the PDSOI body and away from the top-gate. Due to lack of carrier confinement resulting from near-zero vertical electric fields (VE_{field}) in body (VE_{field} =~5e5(V/cm) for PG and ~1e3(V/cm) for QG-MG at Vg=0V) QG-MG FETs have poor SS.

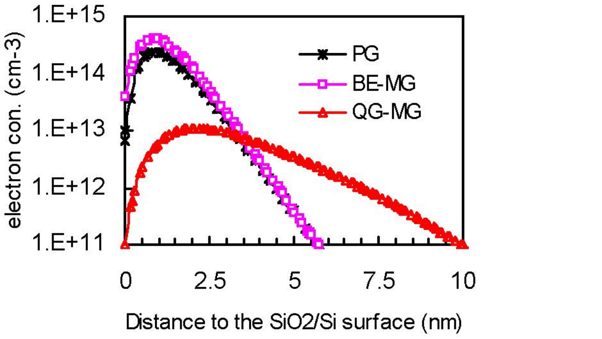

Fig. 5 Electron density at zero bias along a vertical cut in the PDSOI, Lg=25nm.

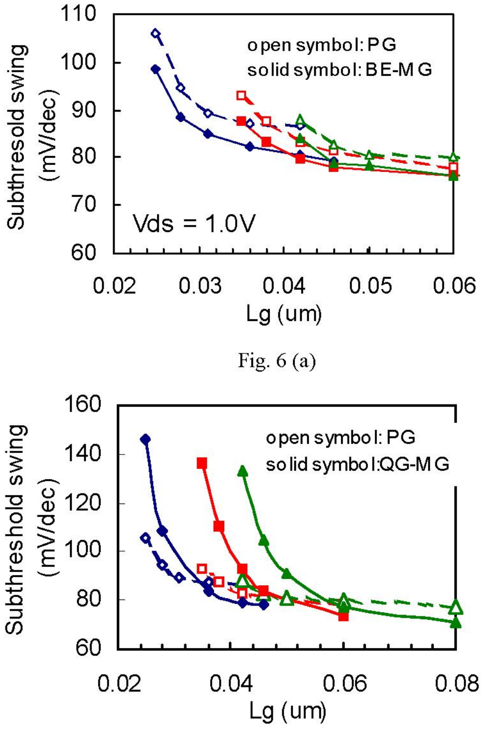

Fig. 6 (a)

Fig. 6 (b)

Fig. 6 sub-threshold swing vs. L_g for PG and MG stacks. (diamond: L_{nom}= 28nm, square: L_{nom}=38nm, triangle: L_{nom}=46nm)

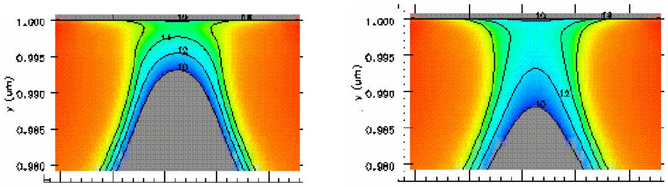

Fig. 7 (a) PG at L_g =25nm

Fig. 7 (b) QG-MG at L_g=25nm

Fig. 7 2D carrier concentration profiles in log-scale. PG shows better carrier conferment over QG-MG.

Figs. 8 and 9 show that when L_{nom} is scaled, V_t roll-off and drain induced barrier lowering (DIBL) become worse for all gate options. QG-MG FETs have high V_t, while BE-MG devices show lower V_t compared to PG cases. Drive current, I_{on}, is strongly dependent on the over-drive (V_{dd}- V_t).

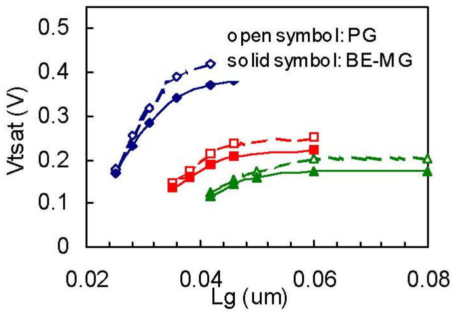

Fig 8 (a)

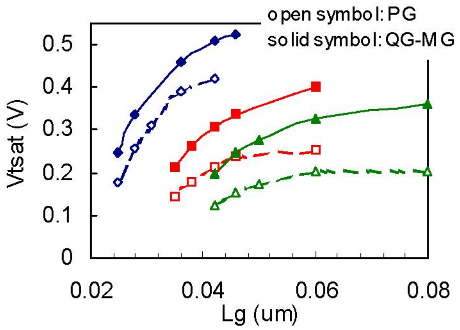

Fig 8 (b)

Fig. 8 V_t *vs.* L_g at V_{ds}=1.0V. (diamond: L_{nom} = 28nm, square: L_{nom}=38nm, triangle: L_{nom}=46nm)

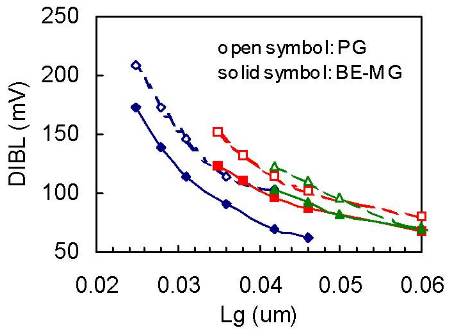

Fig. 9 (b)

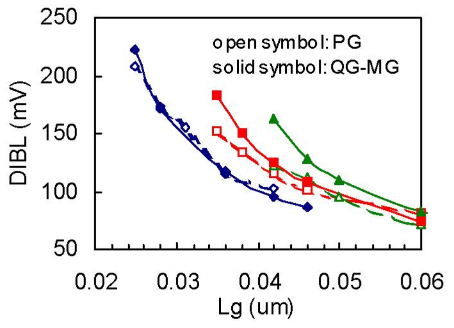

Fig. 9 (b)

Fig. 9 DIBL vs. L_g for PG and MG. (diamond: L_{nom} = 28nm, square: L_{nom}=38nm, triangle: L_{nom}=46nm)

Fig.10a shows a 27% I_{on} improvement at L_{nom} for BE-MG relative to PG. This is due to: (1) Better SS in BE-MG due to good carrier confinement and (2) smaller V_t yielding a larger over-drive. Fig. 10b shows that QG-MG I_{on} is degraded despite the elimination of gate depletion. This is due to: (1) smaller over-drive voltage due to higher V_t, (2) SS degradation limiting I_{on} gain.

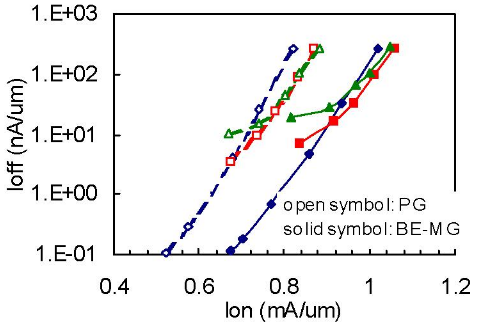

Fig. 10 (a)

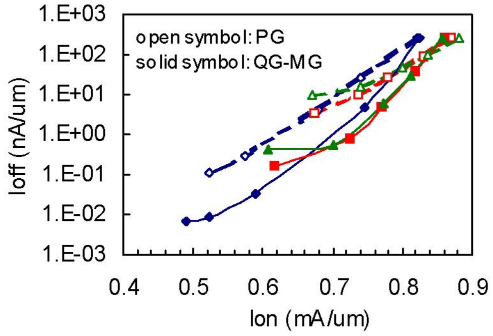

Fig. 10 (b)

Fig. 10 Ioff vs. Ion for PG and MG stacks. Compared to PG, BE-MG shows 27% Ion improvement at L_{nom} and 24% Ion improvement at L_{min} with fixed Ioff, while QG-MG does not Ion improvement at L_{min} and L_{nom}. (diamond: L_{nom} = 28nm, square: L_{nom}=38nm, triangle: L_{nom}=46nm)

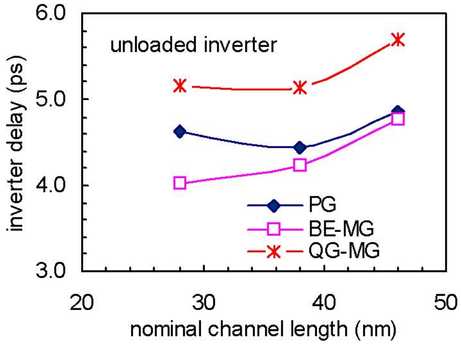

Fig. 11 First switch delay *vs.* L_{nom} for unloaded inverter chain. There is optimized L_{nom}, at which minimum inverter delay is obtained for PG and QG-MG; for BE-MG delay improvement starts to saturate when L_{nom} < 30nm

Next we focus on the circuit performance using comparably designed N and P devices. Fig 11 and 12 show the first switch inverter delays with different capacitance load. BE-MG speed gain relative to PG is 13% and 17% for unloaded and loaded inverters, while QG-MG shows more than 10% delay degradation because the effective drive current of inverter (I_{eff})

(see Fig. 13) does not increase sufficiently to overcome increased loading (see Fig. 14). There is a tradeoff between high current drive and high gate loading for MG devices. More importantly, we find that when L_{nom} is scaled below 30nm, there is NO performance gain for PG and QG-MG and the gain starts to saturate for BE-MG.

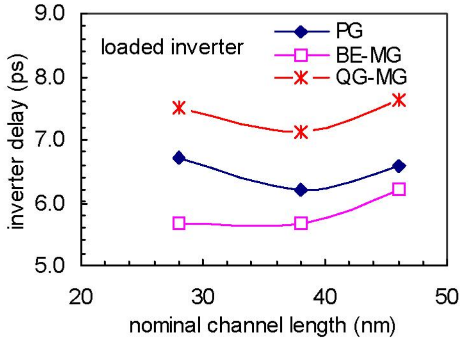

Fig. 12 First switch delay *vs.* L_{nom} for loaded inverter chain. There is optimized L_{nom} between 30 ~ 40nm, at which minimum inverter delay is obtained for PG and MG devices.

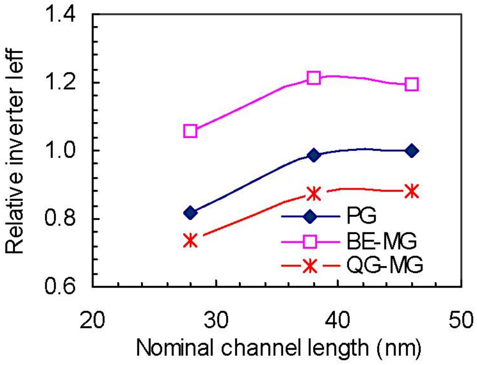

Fig. 13 Relative inverter I_{eff} of CMOS inverter. When channel length is scaled under 40nm, inverter I_{eff} start to decrease.

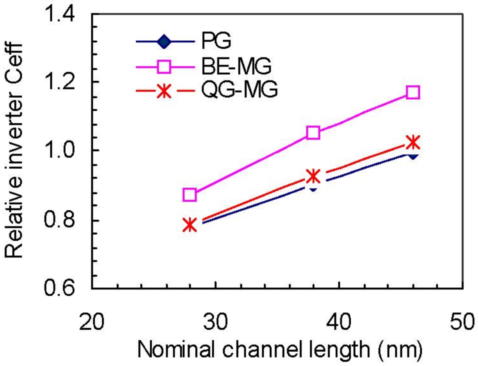

Fig. 14 Relative C_{eff} of inverters. Scaling channel length reduces C_{eff}. MG has larger C_{eff} compared to PG due to smaller inversion layer thickness.

There is an optimized channel length, at which maximum performance gain is obtained. The optimal channel length depends on capacitance load and gate electrode options. For example, from Fig. 13-14, we can see PG I_{eff} keeps constant when L_{nom} scales from 46nm to 38nm. At 28nm, I_{eff} degrades by 20%, while C_{eff} reduces 10% at 38nm and 20% at 28nm. Since delay $\tau = C_{eff}V_{dd}/I_{eff}$, changes on C_{eff} and I_{eff} are canceled at L_{nom}=28nm, so there is no performance gain at 28nm gate length. Fig 15. shows that at the 28nm design point, BE-MG is 10-13% faster than PG. There is a cross point of PG and QG-MG I_{off} *vs.* delay curves. If we compare performance at larger I_{off} criteria, QG-MG is 10% slower; while if we choose smaller I_{off} criteria QG becomes faster than PG.

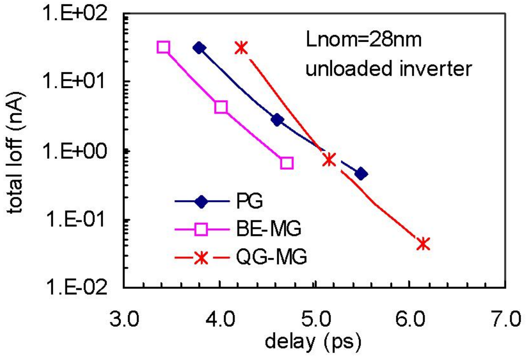

Fig. 15 total I_{off} *vs.* unloaded inverter delay for 28nm channel length design.

Conclusions

We examined the impact of channel length scaling on PDSOI performance with MG and PG stacks. Our simulations show that for HP application (1) there is an optimized channel length, at which maximum performance gain is obtained both for MG and PG; (2) In the short channel length regime (< 46nm), there is no performance gain of QG-MG relative to PG due to lack of carrier confinement, which results in severe sub-threshold slope degradation of QG-MG; (3) BE-MG stacks show 10% gain on a inverter delay over.

References

[1] P. M. Zeitzoff, "MOSFET scaling trends and challenges through the end of the roadmap ", Proceedings of the Custom Integrated Circuits Conference, p. 233-240, 2004.

[2] E. Gusev, *et al.*"Advanced gate stacks with fully silicided (FUSI) fates and high-κ dielectrics: Enhanced performance at reduced gate leakage" IEDM Tech. Dig. 2004, p79-82.

[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] A. Kumar and R. Dennard, "Carrier Confinement in UTSOI Devices: Impact of Metal Gate Work Function" unpublished.

[5] 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

[6] MeiKei Ieong, Ronald Logan, and James Slinkman, "Efficient quantum correction for multi-dimensional CMOS simulations", Proc. SISPAD, p. 129, 1998.

[7] P. Ranade, T. Ghani, K. Kuhn, K. Mistry, S. Pae, L. Shifren, M. Stettler, K. Tone, S. tyagi and M. Bohr, "High performance 35nm Lgate CMOS transistors featuring NiSi metal gate (FUSI), uniaxial strained silicon channels and 1.2nm gate oxide", IEDM Tech. Dig., p. 227-230, 2005.

[8] S. Su. *et al.*"45-nm node NiSi FUSI on nitrided oxide bulk CMOS fabricated by a novel integration process", IEDM Tech. Dig., p. 231-234, 2005.