ARTICLE IN PRESS

Microelectronic Engineering xxx (2009) xxx–xxx

Contents lists available at ScienceDirect

Microelectronic Engineering

journal homepage: www.elsevier.com/locate/mee

Roadmap for 22 nm and beyond

H. Iwai *

Frontierl Research Center, Tokyo Institute of Technology, 4259-J2-68, Nagatsuta, Midori-ku, Yokohama 226-8502, Japan

ARTICLE INFO

Article history:
Received 25 March 2009
Accepted 25 March 2009
Available online xxxx

Keywords:
CMOS logic
Roadmap
22 nm
Si
Nanowire

ABSTRACT

The down-scaling is still the most important and effective way for achieving the high-performance logic CMOS operation with low power, regardless of its concern for the technological difficulties, and thus, the past shrinking trend of the gate-length has been very aggressive. In this paper, logic CMOS technology roadmap for '22 nm and beyond' is described with ITRS (International Technology Roadmap for Semiconductor) as a reference. In the ITRS 2008 Update published just recently, there has been some significant change in the trend of the gate length. The future gate-length shrinking trend predicted in the past several versions of the ITRS has been too aggressive even for the most advanced semiconductor companies to catch up, and thus, the predicted trend has been amended to be less aggressive from the ITRS 2008 Update, resulting in the delay in the gate-length shrinkage for 3 years in the short term and 5 years in the long term from those predicted in ITRS 2007. Corresponding to this, the pace of the introduction of new technologies becomes slower. For the long term, the limit of the downsizing is a big concern. The limit is expected to be at the gate length of around 5 nm because of the too huge off-leakage current in the entire chip. Until that we will have probably six more generations or 'technology nodes', considering that we are now in the so-called 45 nm generation. It would take probably 20–30 years until we reach the final limit, because the duration between the generations will become longer when approaching the limit. In order to suppress the off-leakage current, double gate (DG) or fin-FET type MOSFETs are the most promising. Then, it is a natural extension for DG FETs to evolve to Si-nanowire MOSFETs as the ultimate structure of transistors for CMOS circuit applications. Si-nanowire FETs are more attractive than the conventional DG FETs because of higher on-current conduction due to their quantum nature and also because of their adoptability for high-density integration including that of 3D. Then, what will come next after reaching the final limit of the downsizing? The answer is new algorithm. In the latter half of this century, the application of algorithm used for the natural bio system such as the brains of insects and even human will make the integrated circuits operation tremendously high efficiency. Much higher performance with ultimately low power consumption will be realized.

© 2009 Elsevier B.V. All rights reserved.

1. Introduction

The progress of MOS LSIs has been accomplished by the down-scaling of their components such as MOSFETs, since their beginning of early 1970s. The down-scaling is still the most important and effective way or 'royal road' for achieving the high performance and low power consumption, and thus, the past shrinking trend of the gate-length has been very aggressive. In fact, the ideal scaling method proposed by Dennard [1] shown in Table 1 [2] increases the performance and integration beautifully without any increase in power consumption as long as the chip area is kept constant. In other words, the power consumption and production cost per performance or per transistor decrease significantly with the down-scaling. However, the real scaling trend for the 30 years from 1970 was more aggressive. While the gate length, junction depth, and gate oxide thickness decreased with 100×, the supply voltage decreased only 10× and the chip area did increase 10×. The result was that, while we enjoyed the increase of clock frequency with 1000× (only 10× is expected from the ideal scaling), we suffered from the increase of power consumption generation with 100,000× (no increase is expected under ideal scaling), as shown in Fig. 1 [2]. Thus, power consumption is now the limiting factor, and clock frequency and chip area have not very much increased recently. Now, the concern for the difficulty of the downsizing is stronger than the past, facing the tremendous cost increase in lithography, difficulties in developing new technologies, and expected large variations of electrical characteristics of smaller geometry MOSFETs. However, still, the downsizing is the 'royal road' and the effort of downsizing will continue by all means towards the limit until several more generations or 20–30 years, even though the duration between the generations would become longer. This paper describes a roadmap for high-performance logic

* Tel.: +81 45 924 5871; fax: +81 45 924 5584.
E-mail address: iwai.h.aa@m.titech.ac.jp.

0167-9317/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.mee.2009.03.129

Please cite this article in press as: H. Iwai, Microelectron. Eng. (2009), doi:10.1016/j.mee.2009.03.129

SAMSUNG EXHIBIT 1021

Table 1
Ideal down-scaling scheme.

Geometry and supply voltage	L_g, W_g, T_{ox}, V_{dd}	K	Scaling K: $K = 0.7$ for example
Drive current in saturation	I_d	K	$I_d = V_{sat} W_g C_o (V_g - V_{th})$ C_o: gateCper unit area $\rightarrow W_g (t_{ox}^{-1})(V_g - V_{th}) = W_g t_{ox}^{-1}(V_g - V_{th}) = KK^{-1}K = K$
I_d per unit W_g	$I_d/\mu m$	1	I_d per unit $W_g = I_d/W_g = 1$
Gate capacitance	C_g	K	$C_g = \varepsilon_o \varepsilon_{ox} L_g W_g / t_{ox} \rightarrow KK/K = K$
Switching speed	τ	K	$\tau = C_g V_{dd}/I_d \rightarrow KK/K = K$
Clock frequency	f	$1/K$	$f = 1/\tau = 1/K$
Chip area	A_{chip}	α	α: Scaling factor → In the past, $\alpha > 1$ for most cases
Integration (# of Tr)	N	α/K^2	$N \rightarrow \alpha/K^2 = 1/K^2$, when $\alpha = 1$
Power per chip	P	α	$fNCV^2/2 \rightarrow K^{-1}(\alpha K^{-2})K(K^1)^2 = \alpha = 1$, when $\alpha = 1$

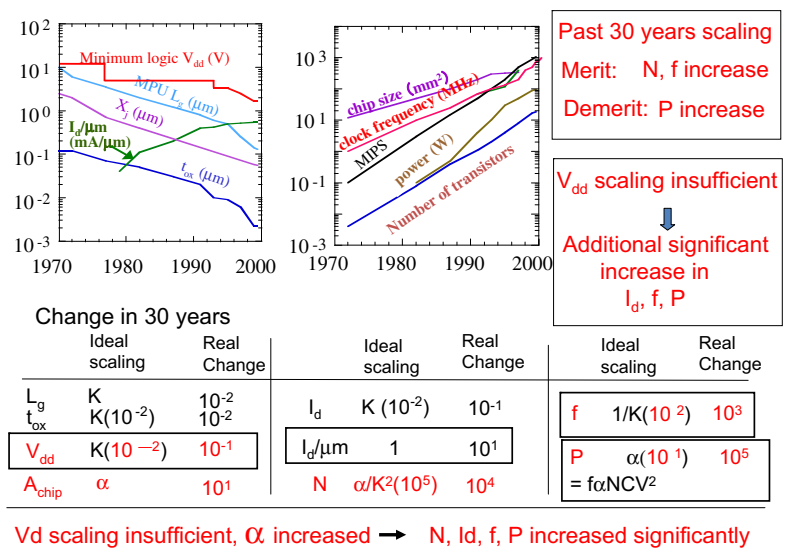

Fig. 1. Actual trend of down-scaling from 1970 to 2000.

CMOS technology for '22 nm and beyond' with ITRS 2008 Update [3] as a reference.

2. Roadmap for 22 nm

Before describing about the 22 nm CMOS logic technology, it would worth to explain what the '22 nm' means, or more generally, what the '*xx* nm' in *xx* nm CMOS technology means in the community. Table 2 [4] shows the expected staring years of the '*xx* nm logic CMOS' products and corresponding parameter values specified in the ITRS 2008 Update. It should be noted that there is neither '45 nm' nor '32 nm' numerical value appearing in the corresponding ITRS parameter sets. However, in the '22 and 16 nm technologies', '22 nm' and '16 nm' happen to appear as the physical gate lengths. Fig. 2 shows typical technology generations for logic LSIs since early 1970s towards future expected limit. Historically, '*xx* μm' or '*xx* nm' represented the lithography resolution which was the half pitch of the lines, the minimum gate length, and the metal line width. '*xx*' Value has decreased approximately with 0.7 in every 3 years in average. Since 0.35 μm generation, NTRS (National Technology Roadmap for Semiconductors in USA: the predecessor of ITRS before it became international) was established and NTRS and ITRS eventually laid down the '*xx* nm' values for all the future generations. However, from the '0.35 μm' or '350 nm' generation, logic CMOS introduced smaller physical gate length than the lithographical resolution by introducing resist-pattern-thinning technique, or resist ashing technique using oxygen plasma [5]. Since then, the LSI makers used '*xx* nm', which was ahead of few generations, for the name of their logic CMOS technologies, claiming that the physical gate length was much smaller. The '*xx* nm' used to have no relation with the physical gate length and just a commercial name as seen in the 45 and 32 nm nodes (Table 2). However, '*xx* nm' incidentally becomes close to the physical gate length from the 22 nm node as described above (Table 2), and this tendency would continue.

So far, the physical gate length of the logic CMOS has been much smaller than the half pitch of the lithography, however, the trends of the physical gate-length shrinkage predicted by recent versions of the ITRS have been even further aggressive for the most advanced semiconductor companies to catch up. Thus, the future trend has to be adjusted to be less aggressive in the ITRS 2008 Update, resulting in the delay in the gate-length shrinkage for 3 years in near future and even 5 years in the middle term as shown in Fig. 3 [4]. Due to the recent serious economical depression which started last year, all the semiconductor companies except Intel reduced the investment for R&D significantly. Thus, there is a high possibility that the gate-length shrinkage trend delays further. Corresponding to the delay in the gate length

Please cite this article in press as: H. Iwai, Microelectron. Eng. (2009), doi:10.1016/j.mee.2009.03.129

Table 2
Relation between '*xx* nm' and ITRS parameter values for '*xx* nm' logic CMOS.

'*xx* nm' CMOS technology			ITRS (2008 Update)		
Technology node (nm)	Starting Year		Year	Half pitch (1st metal) (nm)	Physical gate length (nm)
Commercial logic CMOS products for high-performance logic					
45	2007	↔	2007	68	32
			2008	59	29
32	2009?	↔	2009	52	27
			2010	45	24
22	2011?–2012?	↔	2011	40	22
			2012	36	20
16	2013?–2014?	↔	2013	32	18
			2014	29	16

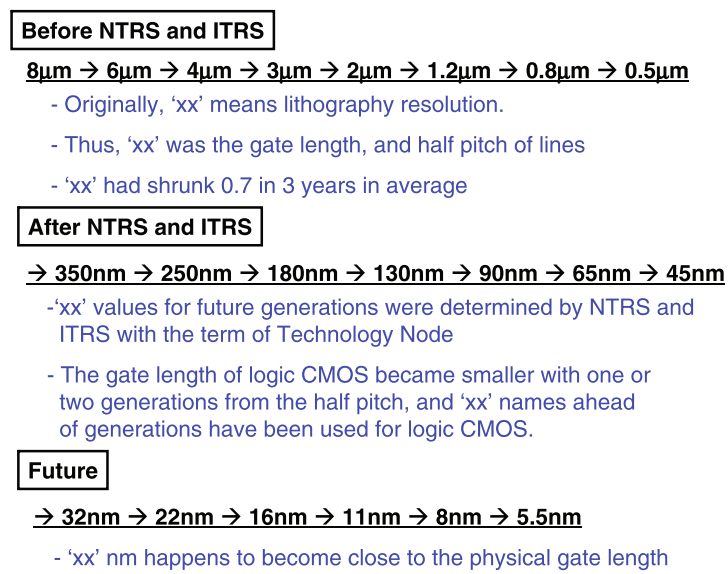

Fig. 2. Past and future technology nodes for MOSLSIs.

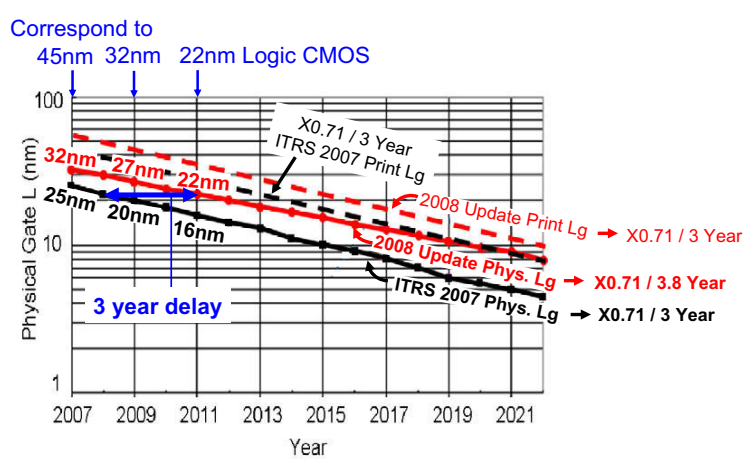

Fig. 3. Comparison of ITRS 2007 and 2008 Update for the trends of printed (resist) and physical gate lengths.

downsizing, the downsizing trend of EOT (Equivalent Oxide Thickness) of the gate insulator and junction depth will delay with the same pace. The critical dimension control or variation control of the gate length for 22 nm node becomes easier because of the gate length increase and red background of the column for the 22 nm node – which means no solution for the gate dimension control – turned to white and yellow. Also, the pace of the introduction of new technologies becomes slower. For example, introduction

Please cite this article in press as: H. Iwai, Microelectron. Eng. (2009), doi:10.1016/j.mee.2009.03.129

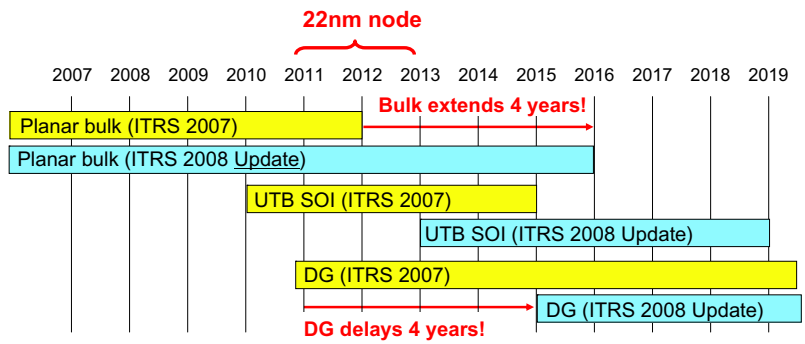

Fig. 4. Structure and technology innovation for MOSFETs.

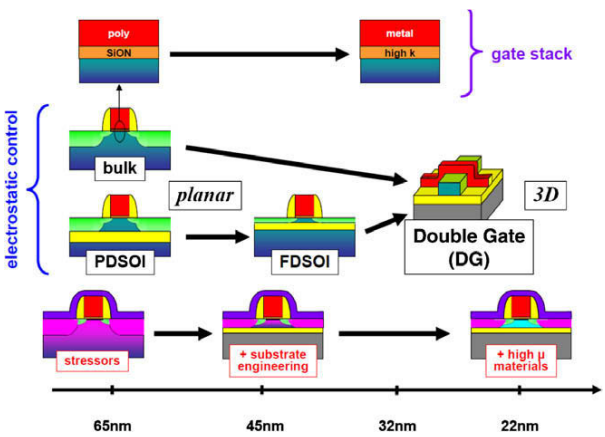

Fig. 5. Past trend of clock frequency for microprocessors.

of DG or fin-gate structure will delayed with 4 years, and 22 nm logic CMOS – which is expected to start production in 2011–12 –, can be made with the planer bulk CMOS, of course, as shown in Fig. 4 [3,4]. In other words, planar bulk CMOS will have a much longer life than expected by ITRS 2007. In ITRS 2007, it was expected that bulk planar structure is replaced by DG structure during 32 nm node and that the silicon channel is replaced by high μ (mobility) materials such as Ge and GaAs from 22 nm node as shown in Fig. 5 [3,4]. However, now, introduction of those structures and materials are thought to delay significantly.

Fig. 6 shows the past trend of the clock frequency of microprocessors [4,6]. The clock frequency had kept increase until it reached 3 GHz. However, recent trend is that it even decreased slightly down to 1–2 GHz when introducing the multi-core scheme. Too high clock frequency too much increases the power consumption and resulted heat generation. This is not a wise way, and ITRS predicts only a small increase of the clock frequency as the entire chip operation. However, local on chip clock frequency, or the core clock frequency is expected to keep increase in the ITRS 2007 as shown in Fig. 7 [3,4]. Already SRAM operation at 6 GHz was confirmed experimentally [3,4,7] and it is forecasted that the local clock frequency keep to increase with the same rate as before in the figure and even though 8% increase per year in IRTS 2008 Update with 6.3 GHz in 2011. It is not sure if the clock frequency can keep such an increase even it is a core frequency for a medium and long term.

Regarding the other issues for 22 nm technology node, 450 mm wafer is predicted to be introduced still in the 22 nm node from 2012 in ITRS 2008 Update. However, most of the people do not think it can be introduced such a near future. For low-k, there is a slight retardation in k value with 0.1–0.3 in ITRS 2008 Update.

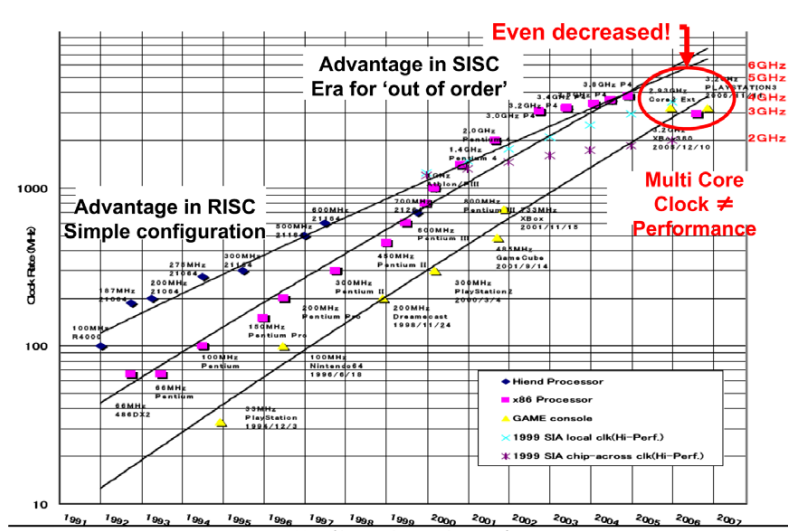

Fig. 6. Past trends of Clock frequency trend of microprocessors.

Please cite this article in press as: H. Iwai, Microelectron. Eng. (2009), doi:10.1016/j.mee.2009.03.129

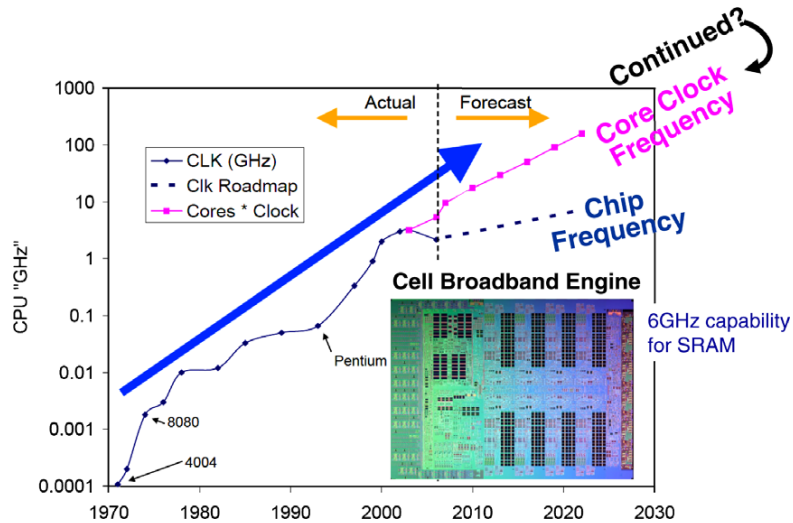

Fig. 7. Clock frequency trend in ITRS 2007 for local on-chip (Core) and entire chip.

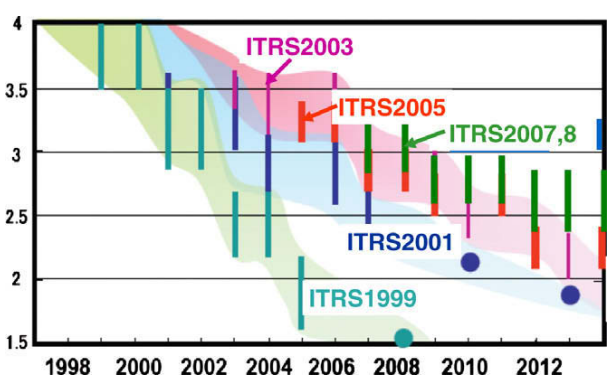

Fig. 8. Predicted low-*k* trends for various version of ITRS.

It is really difficult to realize the low-*k* value predicted by ITRS in production. In fact, *k* value kept retardation in almost every new version of the ITRS as shown in Fig. 8 [3,4].

3. Low supply voltage and off-leakage current

The increasing power consumption is the limiting factor of the logic CMOS, and lowering the supply voltage is the most effective way to decrease the dynamic power consumption. However, in order to decrease the supply voltage, the threshold voltage has to be reduced. This results in the significant increase in the 'off-leakage' current because of the significant increase of the subthreshold leakage current with low threshold voltage, as shown in Fig. 9 [4]. Thus, the threshold and hence, the supply voltages cannot be scaled-down easily. Their values are supposed to stay above 0.1 and 0.9 V, respectively, for next 10 years in ITRS 2008 Update as

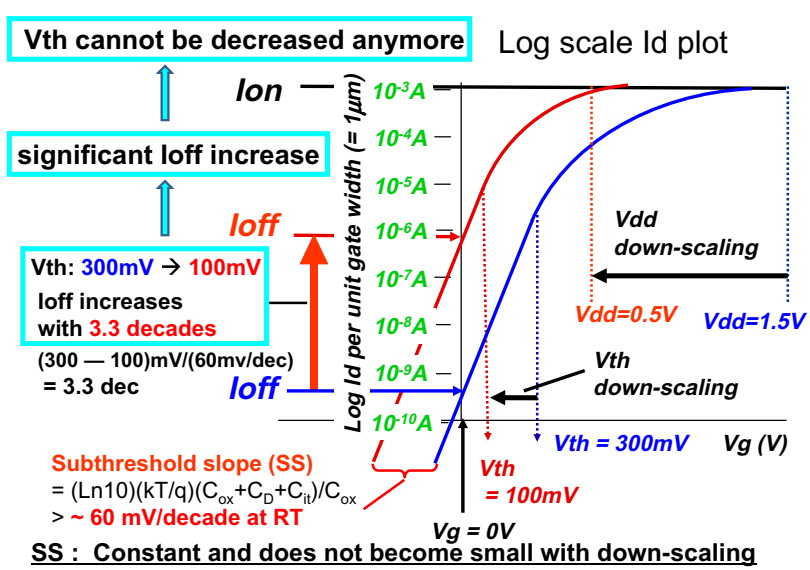

Fig. 9. Increase in off-leakage current ($V_g = 0$ V) with supply and threshold voltage reduction.

Please cite this article in press as: H. Iwai, Microelectron. Eng. (2009), doi:10.1016/j.mee.2009.03.129

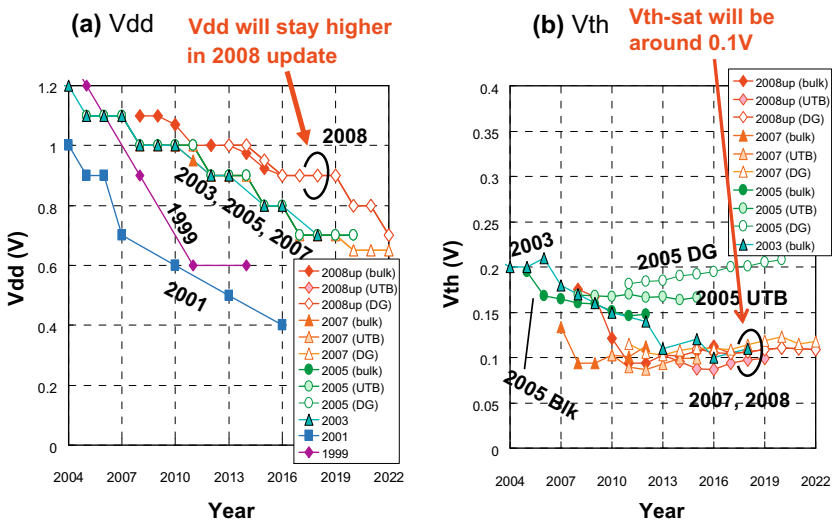

Fig. 10. Trend of: (a) supply voltage and (b) threshold voltage for various versions of ITRS.

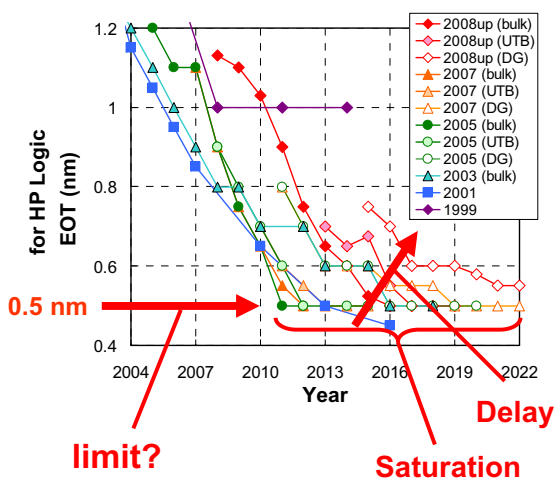

Fig. 11. Trend of EOT of the gate insulator for various versions of ITRS.

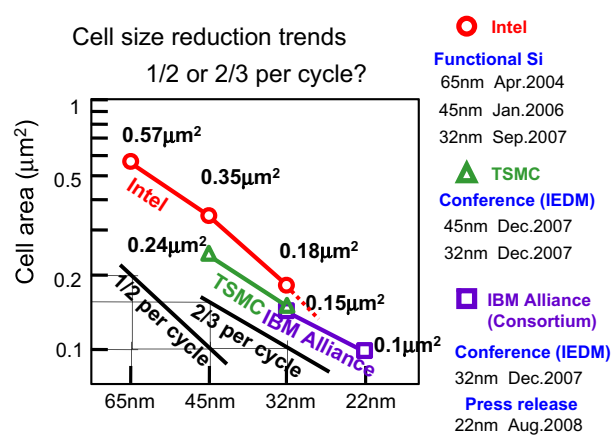

Fig. 12. Trend of experimentally fabricated SRAM cell size.

shown in Fig. 10 [3]. This kind of improper down-scaling, with keeping higher supply voltage and larger gate oxide thickness, is the solution for the downsizing for the moment. However, the improper scaling enhances the short channel effects, resulting in the larger off-leakage current and larger variation of the threshold voltage. The ITRS trends for the EOT of the gate insulator saturate at 0.5 nm as shown in Fig. 11 [3]. This will cause the increase in the off-leakage current and threshold variation in a future small geometry MOSFETs. Regarding the future possibility for the EOT below 0.5 nm, we have already experimentally confirmed a good operation of MOSFETs with EOT of 0.37 nm using the La_2O_3 gate insulator [8]. Thus, in future the EOT value predicted in the roadmap will decreases a little further to solve the problems, and then, the supply voltage would decrease further in order to suppress the short channel effects.

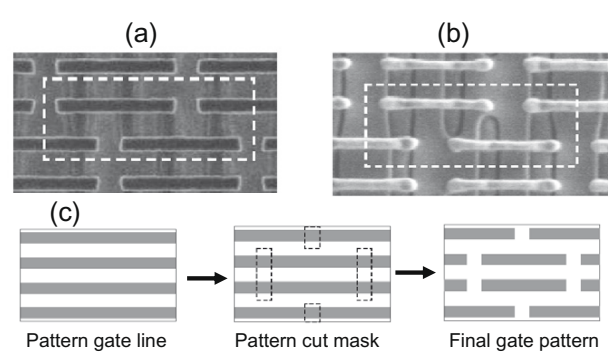

Fig. 13. Square endcap of the lines used recent SRAM cells; (a) non-square endcap, (b) square endcap, (c) double lithography use to realize the square endcap.

4. SRAM scaling

SRAM composes a significantly important part of logic devices as cache memories and its occupying area is quite large. Even a small off-leakage current of a single MOSFETs in a SRAM cell makes a large off-leakage in the entire chip, hence, it is especially difficult to decrease the gate length and supply voltage of the SRAM cell. Thus, the gate length and supply voltage used in the SRAM cell are often designed to be larger than those used in the logic part

Please cite this article in press as: H. Iwai, Microelectron. Eng. (2009), doi:10.1016/j.mee.2009.03.129

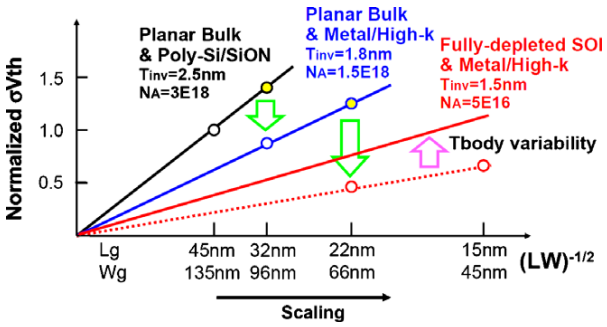

Fig. 14. Suppression of threshold voltage variation by proper down-scaling by the introduction of high-*k*/metal gate stack.

of the chip. Nevertheless, the experimental fabrications of the SRAM cell shows the same reduction trend – with the shrink rate from 1/2 to 2/3 for every generation – until the 22 nm node as shown in Fig. 12 [4]. In order to realize the 32 nm and 22 nm SRAM cells, new techniques are introduced. One is a double lithography to realize square endcap of the gate pattern as shown in Fig. 13 [9,10] and another is high-k/metal gate stack in order to suppress the threshold voltage variation with reducing the EOT as shown in Fig. 14 [3,4,9].

It should be noted that the reduction of the supply voltage in SRAM cell degrades the data retention of the cell. In order to improve the data retention, it is necessary to improve both the read and write voltage margin of the cell. However, it is difficult to optimize the read and write voltage margins at the same time, because the optimum gate width ratio of MOSFETs between the SRAM latch and transfer gate parts are opposite for read and write margins. In order to solve this, in the design of Intel's new multi-core microprocessor, Nehalem, they use 8T cell for L1 and L2 cache in a core (Fig. 15) [11], separating the read and write bit lines (as shown in Fig. 16 [12]) to chose optimum gate width ratio for the read and write, paying a penalty of the cell area increase of about 30%. In this way, the supply voltage for the cores can be decreased with keeping high data retention. The high density L3 cache commonly used in the entire chip level still uses the 6T cell in order to suppress the increase of the chip area, paying a penalty of higher supply voltage than the core. In some future further beyond the 22 nm node, introduction of new cell structure such as DG-FET cell or D-RAM capacitor cell will keep the cell size reduction rate.

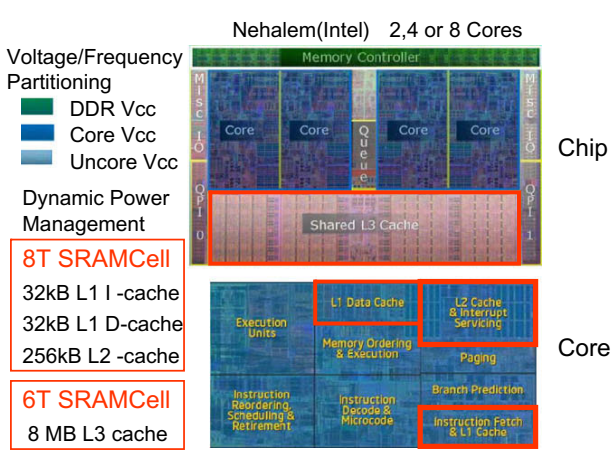

Fig. 15. Three kinds of supply voltages and 6- and 8-Tr. cells for Intel's microprocessor, Nahalem.

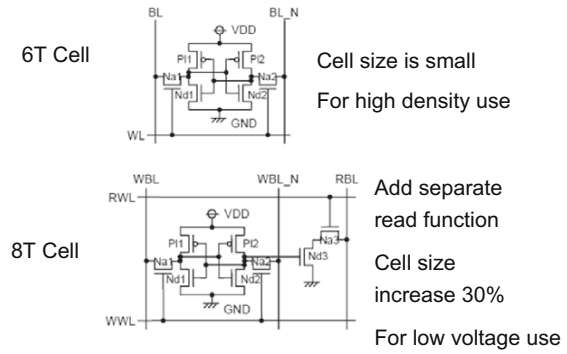

Fig. 16. Example of circuit design for 6- and 8-Tr. SRAM cells.

5. Roadmap for further future

The logic CMOS will encounter its downsizing limit sometime in 2020–2030 around the gate length of 5 nm [13], presumably due to the huge off-leakage current in the entire chip. Thus, probably we will have 6 more generations until then. Two types of FET's have been recently recognized as the emerging devices which could replace current planer bulk CMOS (Fig. 17) [3]. They are the Si-nanowire FET and the alternative channel (such as GaAs and Ge) FET. They are quite different from the current planar type Si CMOS devices, in terms of structure and material, respectively. Considering the compatibility with current Si CMOS process

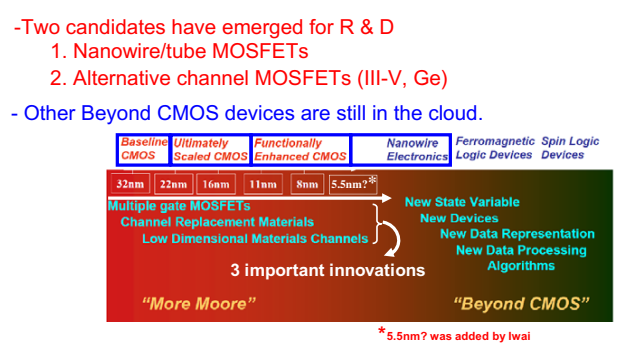

Fig. 17. Roadmap toward the down-scale limit.

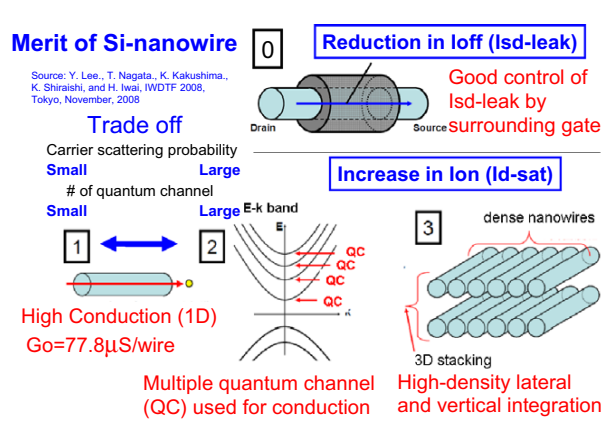

Fig. 18. Merit of Si-nanowire MOFETs.

Please cite this article in press as: H. Iwai, Microelectron. Eng. (2009), doi:10.1016/j.mee.2009.03.129

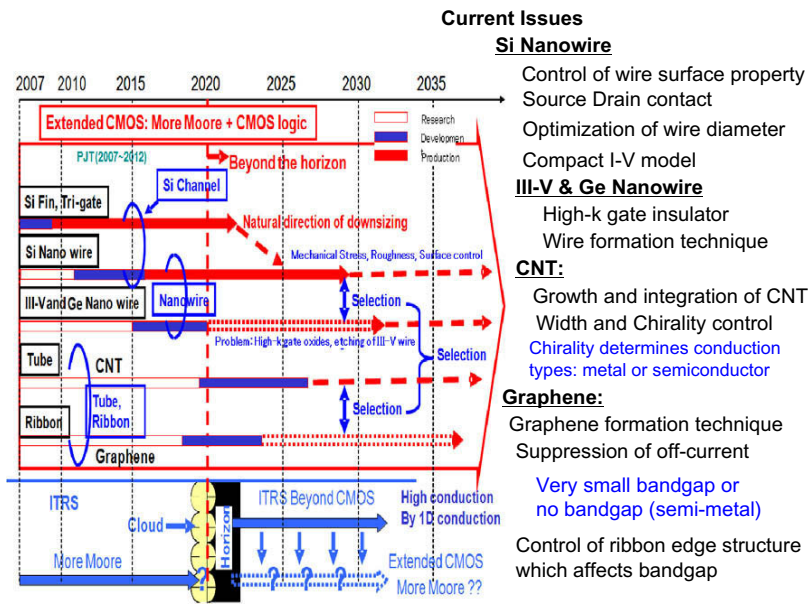

Fig. 19. Long range roadmap for logic CMOS transistors for next 30 years.

technologies, Si-nanowire FETs would be easier for production and more promising. Even if the alternative channel FET would become the main stream, the channel shape should be a wire type, because of the strong demands for the suppression of the off-leakage current.

The Si-nanowire FET has higher on-current conduction due to their quantum nature and also because of their adoptability for high-density integration including that of 3 dimensional stacked layer structure, as shown in Fig. 18 [14]. Because the nanowire pattern itself is simple, nano-inprint technology will be used for future high-density lithography with extremely small pitch. If the ideal one-dimensional ballistic conduction is realized for the nanowire, the nanowire itself has basically a high quantum conduction with 77.8 μS per wire regardless of the wire diameter and the channel length. In addition, the channel current is multiplied with the number of the quantum channel available for the conduction. However the increase of the quantum channel degrade the conduction because of the carrier scattering between the conduction bands, and there is a trade off relation between the one-dimensional ballistic conduction and the number of quantum channel in terms of the nanowire width. Smaller wire diameter is desirable for one-dimensional ballistic conduction and larger diameter is desirable for the larger number of quantum channel.

Fig. 19 [13] shows a long range roadmap including the period which ITRS does not covers. The Si-nanowire FETs is the most promising candidate as explained. III–V and Ge-nanowire FETs are the second candidate. However, technical barrier for the fabrication process is much higher compared with the Si nanowire. In

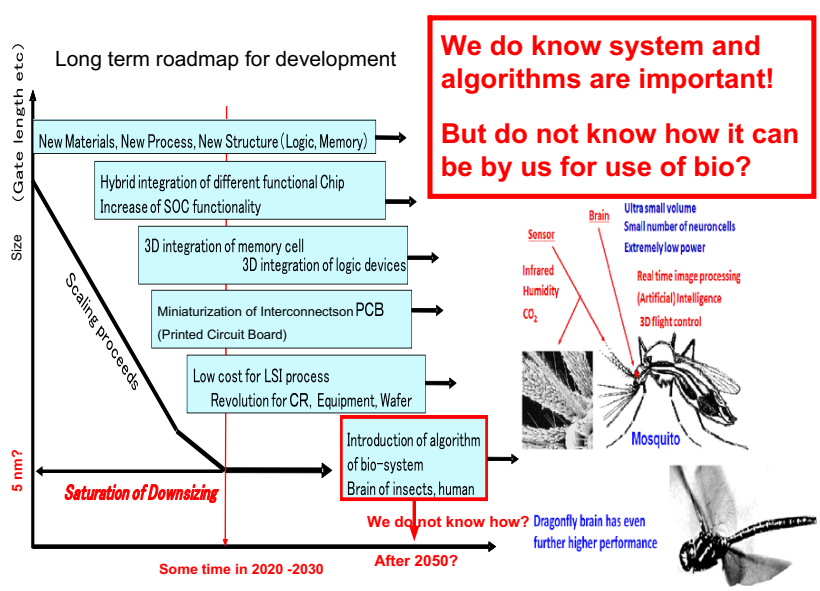

Fig. 20. Further long range roadmap in this century.

Please cite this article in press as: H. Iwai, Microelectron. Eng. (2009), doi:10.1016/j.mee.2009.03.129

further future, CNTs (Carbon Nanowire Transistor) and graphene ribbon FETs could be candidates to replace the Si-nanowire FETs. However, they are still too far at this moment because of no substantial idea for the integration method of so huge number transistors in a chip and bandgap control for high on/off ratio.

What will come next, after reaching the final limit of the downsizing? Probably, there will be the innovation or revolution in the production method of LSIs, and the LSIs will be produced with much cheaper cost. Then, the next step is to use new algorithm. In the latter half of this century, the application of algorithm used for the natural bio system such as the brains of insects as shown in Fig. 20 [15,16] and even human will make the integrated circuits operation tremendously high efficiency. Just for example, brain of the mosquito make the real time 3D flight control with image processing equipped with many sensors such as infrared and CO_2 with extremely small brain volume and extremely small energy consumption. The performance of dragonfly's brain is much higher. Today's performance and energy consumption of the microprocessor are not comparable to those of insect brains, at all. Introduction of the algorithm of the bio system will be the ultimate method in the roadmap.

6. Summary and Conclusion

A roadmap for high-performance logic CMOS technology for '22 nm and beyond' was explained with ITRS 2008 Update as a reference. The predicted trend of gate-length reduction in the past version of the ITRS was too aggressive for the industry to catch up and thus, the pace of the reduction in the gate length became less aggressive from the ITRS 2008 update. Corresponding this, introduction of the new technologies, structures, and materials will delay and 22 nm logic CMOS will be made with planar bulk MOSFETs. The supply voltage reduction is a very difficult item for the next 10 years, because of the difficulty in reducing the threshold voltage any more, and the supply voltage stays at 0.9 V even in 2019 in the ITRS 2008 Update. 22 nm SRAM cell for cache application can be made with planar MOSFETs with introduction of new technologies.

In the long term, Si-nanowire FETs are the most promising candidate because of process compatibility with the current planar CMOS LSIs, and also because of its small off-leakage current and high on-current. In further future, introduction of the algorithm of bio system will be the key for further improvement of the performance and energy consumption.

Acknowledgements

This study was partially supported by the 'Innovation Research Project on Nanoelectronics Materials and Structures' sponsored by Ministry of Economy, Trade and Industry, Japan. This paper is based on a lecture given at the IEDM Short Course given in San Francisco, US on December, 2009. The author would like to thank Drs. H. Ishiuchi of Toshiba M. Saito, Y. Urakawa, and T. Yabe of Toshiba, K. De Meyer of IMEC, Dr. Bohr, T. Ghani, and P. Gargini of Intel, B.S. Haran of IBM, Profs. K. Natori and K. Shiraishi of Tsukuba University, Profs. K. Yamada and K. Ohmori of Waseda University, and Profs. K. Kaukshima and P. Ahmet, and Messrs. T. Kawanago, S. Sato, and Y. Lee of Tokyo Institute of Technology for the useful information and discussion for the preparation of materials.

References

[1] R.H. Dennard, F.H. Gaensslen, H.-N. Yu, V.L. Rideout, E. Bassours, A.R. LeBlanc, J. Solid-State Circuits SC-9 (1974) 256.
[2] H. Iwai, S. Ohmi, Microelectron. Reliab. 42 (2002) 1251.
[3] <http://www.itrs.net/reports.html>.
[4] H. Iwai, SC, in: IEDM, 2008.
[5] M. Ono, M. Saito, T. Yoshitomi, C. Fiegna, T. Ohguro, H.S. Momose, H. Iwai, IEEE Trans. Electron. Dev. 42 (1995) 1510.
[6] M. Saito, private communication.
[7] J. Pille, C. Adams, T. Christensen, S. Cottier, S. Ehrenreich, F. Kono, D. Nelson, O. Takahashi, S. Tokito, O. Torreiter, O. Wagner, D. Wendel, in: ISSCC Dig. Tech., 2007, pp.322–323.
[8] K. Kakushima, K. Okamoto, K. Ttachi, P. Ahmet, K. Tsutsui, N. Sugii, T. hattori, H. Iwai, in: IWDTF, 2008, p. 9.
[9] K.J. Kuhn, in: IEDM Tech. Dig., 2007, p. 471.
[10] M. Bohr, in: Proc. ICSICT, 2008, p. 13.
[11] Intel Developer Forum, 2008, <http://www.intel.com/idf/index.htm>.
[12] Y. Morita, H. Fujiwara, H. Noguchi, Y. Iguchi, K. Nii, H. Kawaguchi, M. Yoshimoto, in: Symp. on VLSI Circ., 2007, p. 256.
[13] H. Iwai, in: IWJT, 2008, p. 1.
[14] Y. Lee, T. Nagata, K. Kakushima, K. Shiraishi, H. Iwai, in: IWDTF, 2008, p. 83.
[15] H. Iwai, H.S. Momose, M. Saito, M. Ono, Y. Katsumata, in: INFOS, 1995, p. 147.
[16] H. Iwai, in: IPFA, 2006, p. 1.

Please cite this article in press as: H. Iwai, Microelectron. Eng. (2009), doi:10.1016/j.mee.2009.03.129