특집

Future Challenges and Opportunities of the Semiconductor Industry

Changhyun Kim(Senior Vice President, Memory Division, Samsung Electronics)

Ⅰ. Introduction

The still ongoing digital revolution of the late 20th century has changed the daily life of a great number of people in a way second only to the industrial revolution. The explosive growth of the semiconductor industry, which reached 16% per year or 4.4 times per decade in the past (Fig. 1), was based on two pillars, the incessant demand for information-related goods, which continuously created new markets, and the advance in device integration technology according to the predictions of Moore & Hwang[1, 2], which satisfied the need for enhancing system performance and reducing production

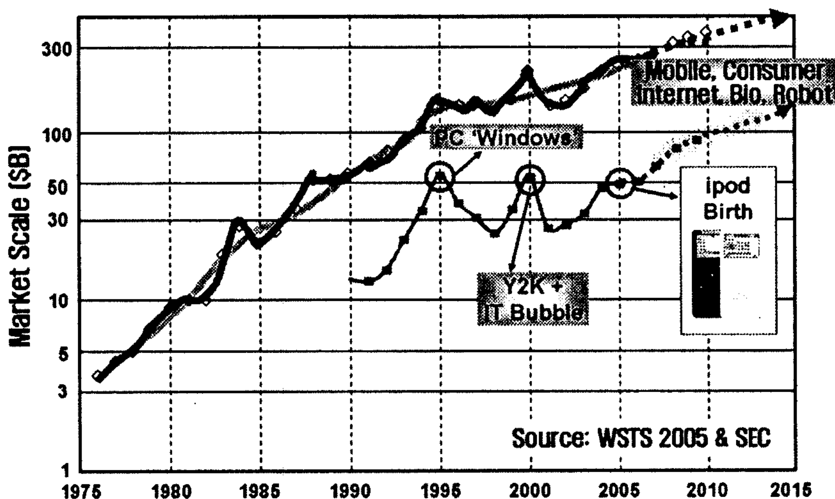

Fig.1: Semiconductor Market Development.

costs. Despite this success story, it is generally accepted that the semiconductor industry will meet unprecedented challenges in the near future. On the technology side, it seems that linear scale-down of device sizes will soon approach economic limits even before feature sizes approach fundamental physical barriers. On the market side, the growth rate of the semiconductor industry recently shows slight saturation, despite the steady growth of established markets. The future path will thus be less clear-cut and straight than it has been in the past decades. In this paper challenges of the semiconductor industry together with current and future possible solutions will be outlined. A general prospect of the future semiconductor business will also be given.

II. Development and Current Status of Semiconductor Markets and Technologies

The historical development of the semiconductor industry (Fig. 1) has shown that the driving force behind new technologies has been the successive creation of new application fields demanding higher performance on the one side, and the continuous pressure for reducing product costs on the other. The high-end server, home computer and video game market of the 1970's and 80's marked a first boost of mass consumer demand for CPU and DRAM chips. The first generation of computer and/or game enthusiasts was fascinated by the possibilities of the new computer technology. The consolidation of the home computer market, leaving IBM compatible PCs as an industrial quasi-standard, was paralleled by the extension of PC usage since the late 1980's and 90's as a universal tool not only for "freaks" and "computer kids" b ut nearly all kinds of people. Contrary to the semiconductor CPU and DRAM market, high density storage for PC and servers by this time was completely satisfied by hard disk technology whereas mobile consumer products were nearly non-existent before the late 1990's. Thus high density nonvolatile semiconductor memories did not enter consumer markets of high volume at that time, despite continuous progress in NAND flash memory technology. Beginning in the late 90's, the growth of the well-established EDP market began to stagnate on a high-level. However, new diversified consumer markets for entertainment (game consoles) and mobile applications appeared (cell phones, MP3, digital cameras, smart cards, etc.). This latter development opened new opportunities for processor and DRAM. Restricted battery life drove the industry to develop new, low power consumption

chips such as Mobile DRAM. The graphics and game digital consumer sector required special high-speed I/O memories (Rambus, XDR with currently 1~3 Gbps bandwidth). Apart from the specialization of the existing main memory and CPU market, the increased demand for mobile multimedia contents also led to the explosive demand for low cost, nonvolatile mass-storage solutions, which had to be mechanically robust and compact in size at the same time. This demand, which could not easily be met by the traditional hard disk technology, gave birth to the still on-going boom of the NAND flash memory market.

On the technology side, the need for reducing production costs and increasing performance has led to the continuous increase in speed and density, a development which is common to the current three main high-volume semiconductor product groups CPU, DRAM and NAND. This has been made possible by the rapid evolution of optical lithography, proceeding to ever smaller device features and the effects of transistor scaling theory, which allowed scaling without dramatic changes in transistor structure. Until recently the doubling rate was once every two years for CPU transistors, once every 18 months for DRAM bit density and even once every year NAND flash bit density. Feature sizes approach 60 nm for DRAM and 40 nm for CPU and NAND flash. The growth rate of NAND integration density (predicted by Hwang[2]) exceeding the other semiconductor

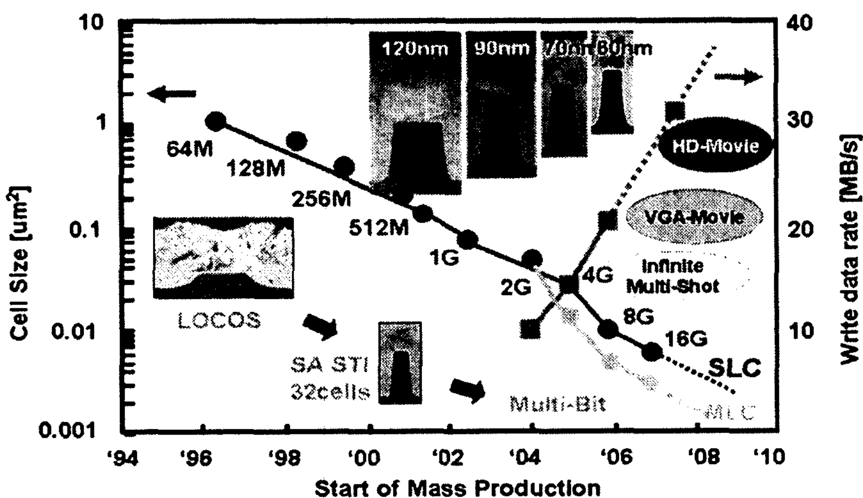

Fig. 2: Evolution of NAND density, cell size, cell structure and write data rate.

product groups was made possible by innovative MLC technolgy (Fig. 2), which allowed the effective cell density to exceed the real physical density by a factor of two. Bit density has now reached 6~10 Gbits/cm2. Additionally, the decrease of transistor delay which occurs with the downscaling of transistor length was the main factor in increasing device speed.(CPU 200 GFLOP, DRAM bandwidth 1 Gbps in 2007).

However, with ongoing scaling of device sizes increasing difficulties specific to each semiconductor product occurred. For DRAM, the maintenance of data retention time has always been a great challenge. Sufficient amount of capacitance in each memory cell and extremely low level of leakage current from the storage node are getting more difficult with scaling. The first requirement led to the evolution of storage capacitor architecture from planar to 3-D. 3-D capacitors, beginning from simple stack shape, evolved to combined stack and Hemispherical Grain(HSG) technology, cylindrical shape to finally combined cylindrical shape and HSG technology. The DRAM cell transistor on the other hand had to be optimized between low sub-threshold leakage and junction leakage current on the one hand and sufficient on-current for high speed operation. Among other things, DRAM access transistors had to be designed to have immunity against the off-current increase due to short-channel effects. As a solution the RCAT(Recessed Channel Array Transistor) has been developed, thereby increasing the effective gate length. Furthermore, some companies adopted an open bit-line structure for its DRAM products to shrink cell size beyond mere shrink of feature size, leading to a reduction from 8F2 to 6F2. The resulting increase in bit density could otherwise have been achieved only by an advance in design rule by one generation. NAND flash also has undergone several changes in cell structure. Cells manufactured with LOCOS have early been replaced by self-aligned STI technology, which has later advanced to a technology where substrate, floating gate and control gate are all patterned in a self-aligned process(see insets in Fig. 2).

III. Challenges and Limitations in the Near Future

The future market requirements on density, speed and power consumption are fundamentally similar to the past. These requirements are expected to get increasingly difficult to realize technologically, as the scaling of device sizes faces limits from various sides. Common to all semiconductor product

groups are the limits of photolithography and process variability. The current use of photolithography for pattern sizes below 50 nm has exceeded the early pessimistic predictions that it would hardly get beyond the micrometer barrier. However, the transition from current DUV (193/157 nm) to EUV (13.5 nm) technology is a major challenge. Although the current status of EUV tools suggest commercial availability and high volume use by 2009[3], the cost-effectiveness as well as the optimal time of transition of is still a matter of debate. Recent efforts therefore show the trend to extend the existing DUV photolithographic technologies before investing in next generation lithography facilities. The recent scaling in feature size down to 50 nm and less, which is far beyond the optical wavelength of the used DUV light sources, is achieved by various "tricks" such as phase shift masks or immersion lithography. High-index immersion lithography is expected to be the latest development on this line which might push the lithographic limit to 30 nm. A combination of Double-Exposure and current immersion lithography may even allow existing DUV facilities to extend to the 20nm range. Litho-friendly layout at the design stage and advanced OPC are efforts to postpone the mentioned limits. Another common issue besides lithography is process variability which is one of the main sources of parametric or design-limited yield loss in the deep nanometer range. As dimensions approach the atomic scale, the atomic "graininess" such as intrinsic line edge roughness or dopant fluctuations becomes inevitable and is therefore necessary to be solved at the design level.

Other scaling issues are more product-specific. CPU's which are moving towards the 45 nm node in the near future face the problem of increasing transistor subthreshold and gate leakage leading to excessive power consumption and heat. The gate oxide thickness of currently 1.2 nm is already at the limit below which power dissipation due to direct gate oxide tunneling becomes intolerable. This means, that the overall performance gain, which until recently relied 80% on shrink induced frequency scaling, approaches a fundamental physical barrier. However, it has to be added that a great part of the slowing down of speed gain is related to memory latency and throughput ("memory wall" and "von-Neumann bottleneck"). While the demand for processor performance is expected to increase at the same rate as in the past, constraints on power consumption will remain or even get tighter, forcing future technical solutions to focus on both aspects (see multicore architecture in next section)

Looking at the scaling barriers of NAND flash, tunneling oxide and interpoly-dielectric layer thicknesses have already reached their lower limits, below which data loss due to direct tunneling becomes critical. This means that contrary to the lateral dimensions the stack height of a NAND flash cell cannot be reduced significantly, adding considerable difficulties to processing. Another problem is, that as the lateral dimensions shrink and cells get closer to each other, the effective coupling ratio decreases due to parasitic coupling with neighbouring cells resulting in high programming and erase voltage. The most fundamental limitations to current NAND technology is thought to be the decreasing amount of charge within the floating gate cell. Even if the charge loss rate of current devices is maintained, future NAND cells will have much worse data retention properties. The scaling problems of DRAM outlined in the previous chapter will also get more severe in the future. Beyond the 50 nm node, solutions for improving increased transistor GIDL current and sensitivity to the bulk bias are crucial. At even smaller device sizes fluctuation of subthreshold leakage due to inevitable process variation is expected to be eventually a fundamental limit to DRAM-type memories.

IV. Possible Technical Solutions

The issues in the preceding chapters make it clear that device scaling can only be maintained with accompanying process technologies which attenuate and thus postpone the mentioned limitations. For CPU, the development of device technology points into the direction of Tri-gate FinFET combined with high-k gate dielectrics and metal gate to reduce off-current and gate leakage. Solutions for the inherent problems of this technology, Fermi-level pinning and phonon-scattering, are expected to near the manufacurable stage at the 40 nm node[4]. High-k dielectrics also become increasingly crucial for NAND, in order to increase the coupling ratio between inter-poly and floating gate. Charge trap cell structure(e.g. TANOS) is a possible alternative technology to solve the fundamental scaling problems of FG cells and prevent interference between neighbouring cells. However, economic factors will decide whether CT NAND is going to be adopted or the current FG technology is maintained despite its data retention issues[5]. DRAM cell transistors will possibly maintain the basic RCAT concept even beyond 50nm, if GIDL current and back-bias sensitivity can be solved reliably with minor modifications. For even smaller sizes, body-tied FinFET (Fin field effect transistor) is a candidate for providing

superior short channel properties, higher speed, and lower sub-threshold leakage [6].

Despite these ongoing efforts of process technology, the inherent difficulties in the nanometer region raise the question of whether there exist alternative means, which may enable us to postpone the technological barriers imposed by device scaling but get similar effective gains in performance or density. Two concepts already applied for DRAM and NAND, respectively, may be extended in the future. The past transition of DRAM cell structure, which led to the reduction in size from 8F2 to 6F2 and thus temporarily relaxed the pressure on design rule scaling may be repeated to 4~5F2 cell size, which however will require a completely new concept of cell structure. Research on multi-gate transistor structures (FinFET) or vertical pillar structures are going on to make use of the vertical dimension. With this architecture it may be possible to maintain a reasonable value of leakage current by keeping transistor length while reducing the lateral dimensions for scaling down the occupied layout area. This will however raise technological challenges of their own and still have to prove to be economic in the end. As device scaling of current main memories is getting more difficult, new memory types like PRAM, MRAM and other new RAM-types may provide alternative ways to better shrinkability and reliability. Recently, PRAM with its superior endurance, random access time and cell size is traded as a possible NOR flash replacement, though not yet as a candidate for mass data storage or main memory. The MLC concept of NAND, which doubles the effective cell density may be broadened to physically 3-D stacked memory cells, thereby increasing the effective cell density without scaling of design rule or increasing layout area. Stacked resistive memory-types and other alternative stacked memory technologies are currently being researched at many companies.

Alternatively to scaling technology, but still offering a boost in performance and density, intelligent system level MCP/SIP-based Fusion Technology is gaining in importance. OneNAND achieves both cost-effectiveness and performance by combining NAND cells and embedded SRAM. Maximum read bandwidths of more than 100 Megabyte/s, four times faster than conventional NAND Flash memory, and write bandwidth of up to 17MByte/s, 30 times faster than multi-level-cell NOR Flash memory are achieved in currently available products. Similar recent examples include OneDRAM or FlexOneNAND. Contributions for higher performance and lower power consumption are also expected to come from stacked package, where inter-chip connections beyond the conventional wire

bonding such as Through-Silicon Vias(TSV) technology(Fig. 3) using new laser and RIE processes may serve as a means for drastically reducing parasitic line resistances and capacitances, while significantly increasing functionality and device packing density. In order to reach low parasitic capacitances, developing low-k materials is crucial up to the extreme point of realizing interconnections suspended in air or vacuum.

For maximizing the device performance with the same process technology, design technology plays a key role for the concentration of various functions within a chip. For CPU, attempts to circumvent the fundamental problems of frequency and device size scaling (leakage current, memory wall, von-Neumann bottlenecks) come from organizing the previously monolithic cores into many processing units and thus achieving massive parallelization of operation (chip multi-processing, CMP). This is expected to enhance performance while maintaining power consumption and heat dissipation on a reasonable level by intelligent control of processing resources like powering down idle cores. Finally, as yield loss due to reliability and PVT variability is expected to undo a great part of the gain attained by scaling, tools and methods which consider process issues already at the design stage also become become a key technology[7], as well as the parallel development of high level synthesis CAD

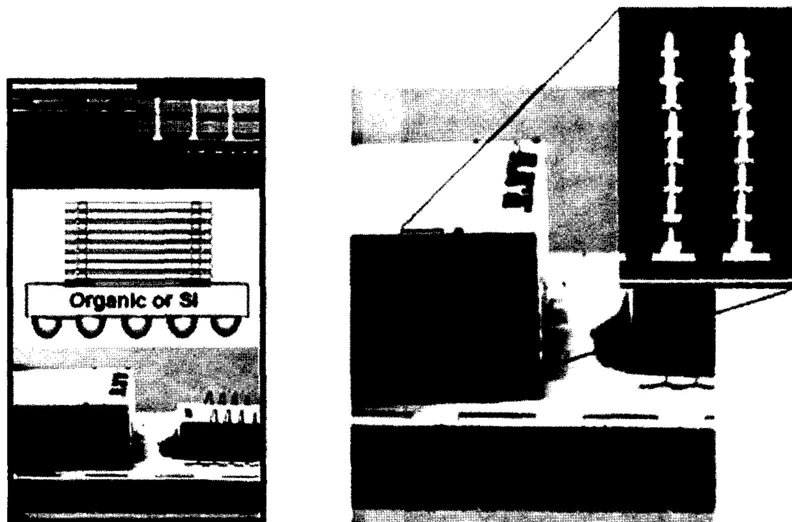

Fig.3: Stack package using Through-hole vias.

tools with standardized design rules and design automation.

A few general remarks are given to summarize the rough overview of this section and to emphasize the impact of the manufacturing cost factor on future device integration. Undoubtedly, device scaling technology hasn't by far reached the fundamental physical limit, and developments will eventually converge to one of the many nanotechnologies of molecular sized devices (CNT, molecular memory, etc.), once some fundamental problems have been solved, including molecular-sized memory core technology or aligning and interfacing new materials and conventional CMOS. However, concerning the near future, maintaining the rate of device scaling is not so much about whether there exist physically realizable solutions (of course they exist!), but rather points at whether device scaling solutions if accompanied by increased immediate manufacturing costs and low yield due to PVT variation will make economically sense [1]. Thus the question is how to bridge the transition area illustrated in Fig. 4 with the best economic solutions, before radically new technologies become available. Recent efforts concerning alternative solutions in lithography or chip design, while trying to keep established device structures and extending existing manufacturing facilities have to be viewed in this light.

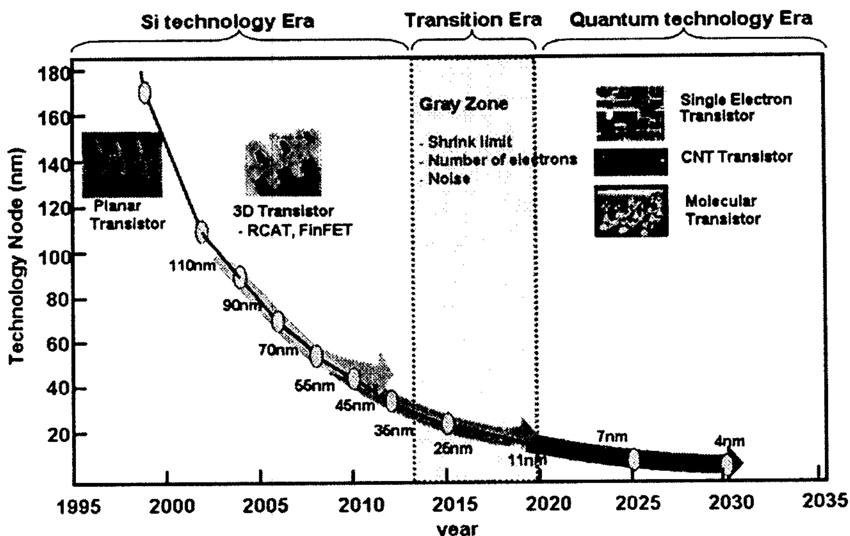

Fig. 4: Long-term prospects of Si shrink technology illustrating the critical transition era between conventional silicon technology and the quantum domain of molecular sized devices.

V. An Optimistic View on Future Semiconductor Industry

As the desire for nonmaterial, information-related audio-visual contents and information is an age-old intrinsic feature of the human mind, it is unquestionable that the demand for ever higher performance semiconductor products which can support these contents will remain unbroken in the future. The ubiquitous demand for digital contents will drive the markets to present new killer applications which in turn require high-performance platforms to run on. Past and ongoing examples in the PC market show that every new generation of Operating Systems (Windows XP, Vista) and applications require a jump in system resources to enhance functionality and user-friendliness. For the mobile market, portable media players, portable digital assistants and mobile device which send large amount of data like images and movie contents between them show that real time networking, data processing and storage will have to get much faster in the future but at the same time have to maintain low power consumption. The entertainment sector sees each new generation of console hardware requiring rapid development of processing and memory technology. Newer consoles compete to output a greater range of natural colors, and use of graphical technologies such as scaling, and vector graphics. This shows the demand for realistic virtual output environments, in answer to which high-end game consoles will continue to pose unprecedented technical challenges on the performance of CPU and data processing engine as well as memory. A particularly interesting example which may point to a characteristic feature of future developments is the wireless controller of Nintendo' s Wii, which can be used as a handheld pointing device and can detect motion and rotation in three dimensions. It is a hint that the future man-machine interfaces will increasingly directly adress natural human senses in 3-D instead of using machine-like knobs and analog sticks.

Looking a bit farther into the future, while the directly consumer-related multimedia and entertainment part of the semiconductor market and technology will still be governed by the strict cost-effectiveness requirements of the past, the co-evolution of other industries which require completely new functionalities will create a highly value-added sector where the price argument of semiconductor chips will play a comparatively minor role compared to performance and reliability. Already high-performance and/or high-reliability data processing is hidden in numerous applications which are not

readily associated with the semiconductor business, a trend that will continue to a much greater extent in the future. Automobiles, aerospace and robotics are fields with still high potential. Extrapolating the current technology trend, the growing capacitiy and performance of semiconductor chips which will eventually approach the complexity of the the human brain will increasingly be able to take over complex decision-making and organizing tasks. A current example, where artificial intelligence already vastly increases functionality is Driver Assistance Systems in automobiles and more advanced technologies in aerospace, where real-time data processing increasingly takes over critical security-related tasks for which humans have proved to be inherently error-prone. Future uses will also be in robotics, humanoid as well as task-specific, where near-intelligent machines will take over many tasks of everyday life by virtue of advanced sensor and actuator technology. Demand for higher processing and storage capability will not be restricted to consumer products for individual use, as the demand for massive data processing will also continue to exist in many fields of basic and applied sciences, where massive data gathering, processing or simulation is needed (climate simulations, earthquakes/tsunamis, DNA analysis, computational neuroscience [8], etc.). Industrial supercomputing, which will require sophisticated simulation, will also profit heavily from enhanced data processing capability. Finally, the explosive growth of the biotechnology sector, which is related to increased life spans and quality of life, opens unforeseen opportunities with advances in nanobiology such as living system-semiconductor interface technologies or health-related nano-sensor technology (early recognition of diseases, screening of harming environment factors). Future fusion technology will enable us to stack multi-functional electronics such as RF, image sensors and bio-sensors (nano lab-on-a-chip) and semiconductor logic and memory. We can thus imagine a future, where the enhanced performance and diverse functionality of the semiconductor chips and the fusion of hitherto separate technologies (IT+BT+NT) will have a great impact on our everyday life (Fig. 5).

VI. Conclusion

Semiconductor industry has the potential of continued growth despite the expected difficulties of shrink technology. This potential will be realized by discovering and stimulating future customer needs, creation of new application fields and the use of fusion technology, through which semiconductor

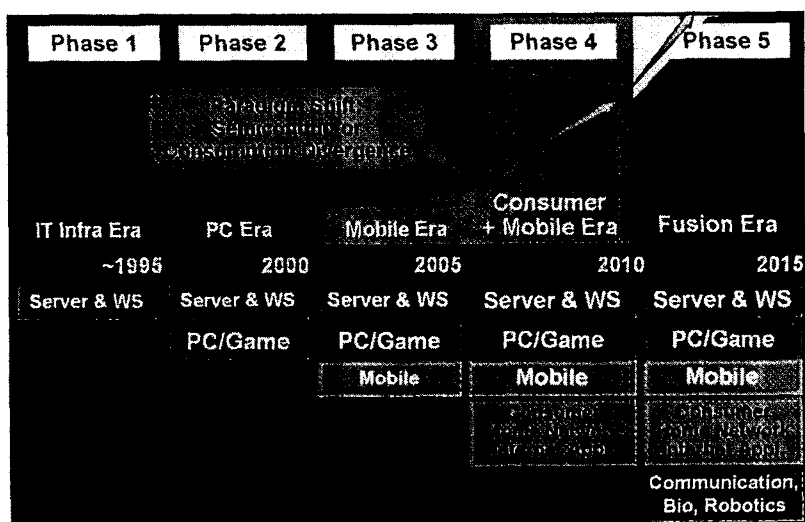

Fig. 5: Past and future paradigms of the semiconductor industry.

industry will combine with other disciplines (IT+BT+NT). Despite the economic and technological challenges reviewed in this paper semiconductor technology will play an increasingly important role in future industries and societies.

References

[1] G. Moore, "Cramming more components onto integrated circuits", Electronics Magazine 19 April (1965)

[2] C-G. Hwang, "Semiconductor memories for IT era", (plenary invited), , Digest of technical papers of 2002 ISSCC, (2002); C-G. Hwang, "Nanotechnology enables a new memory growth model", Proceedings of the IEEE, 91 pp. 1765-1771, (2003)

[3] IMEC, Scientific Report 2006, www.imec.be

[4] R. Chau et al., "Tri-Gate Transistor Architecture with High-k Gate Dielectrics, Metal Gates and Strain Engineering", VLSI Technical Digest, pp. 50-51 (2006)

[5] K. Kim et al., "Future outlook of NAND Flash technology for 40 nm node and beyond", (invited), pp. 9-11, Technical digest of 21st NVSMW, 2006

[6] C. H. Lee et al., "Novel body tied FinFET cell array transistor DRAM with negative word line operation for sub 60 nm technology and beyond",

VLSI Technical Digest, pp.130-131, (2004)

[7] J.-T. Kong, "CAD for Nanometer Silicon Design", IEEE Transactions on VLSI Systems, vol. 12, No. 11, 2004

[8] R. Brette, "Simulation of networks of spiking neurons: a review of tools and strategies", J. of Computational Neuroscience, in press 2007

저자소개

김 창 현

1982년 2월 서울대학교 공과대학 학사
1984년 2월 서울대학교 공과대학 석사
1994년 5월 Univ. of Michigan 박사
1984년 3월-1989년 8월 삼성 반도체 입사, DRAM설계
1995년 2월-현재 삼성전자 반도체 총괄 메모리 사업부 근무

주관심 분야 : High Performance Memory 설계와 연구