Preface

Lin Chao
Editor
Intel Technology Journal

This Q3'98 issue of the *Intel Technology Journal (ITJ)* marks our first anniversary of publishing the *ITJ* on the Internet. From your questions and feedback so far, we know that the *ITJ* is being read all over the globe. Thank you for all your feedback and questions. Please keep them coming.

This Q3'98 issue describes one of Intel's most important technologies: our semiconductor process technology, the prize of Intel's technology jewels. The magic of silicon is basically that in every generation the dimension gets smaller, and processors get smaller, faster, and cheaper to build. This issue of the *ITJ* describes the challenges inherent in making this silicon magic happen.

The first two papers describe our 0.25 micron process technology used to manufacture the Intel® Celeron™ and Pentium® II processors. The 0.25 micron refers to the line-width dimensions etched into the silicon wafers. To illustrate how small 0.25 microns actually is, think about the fact that a typical pollen microspore measures between 10 and 100 microns. Using the 0.25 micron process technology, you could place between 40 and 400 transistors in the width of a pollen spore. Intel is mass producing the 0.25 micron process technology. Next year, in 1999, a 0.18 micron technology will be in production, and we are already working in the lab on a 0.13 micron technology.

The third paper describes the challenges faced when shrinking transistors below the 0.13 - 0.10 micron range. Intel is now addressing these challenges in preparation for the turn of the millennium.

And finally, the fourth paper describes the future of lithography used to imprint very small patterns onto silicon wafers. Optical projection lithography is used today, but will not be able to imprint the ever finer patterns needed in the future. Over the next several years, a new lithographic technology needs to be developed that can print lines of 50 nanometers and smaller. Extreme Ultraviolet Lithography (EUVL) is one of the technologies being evaluated.

Intel's 0.25 Micron, 2.0Volts Logic Process Technology

A. Brand, A. Haranahalli, N. Hsieh, Y.C. Lin, G. Sery, N. Stenton, B.J. Woo
California Technology and Manufacturing Group, Intel Corp.

S Ahmed, M. Bohr, S. Thompson, S. Yang
Portland Technology Development Group, Intel Corp.

Index words: CMOS, shrink, interconnect

Abstract

Process 856 is a 0.25µm-generation logic technology currently in volume manufacturing, which has been optimized for high performance, yield, and density. This process is being used to manufacture high performance products including the Intel® Celeron™ and Pentium® II microprocessors. The process has a high equipment re-use rate to reduce cost. Using the older equipment has increased the challenge of scaling to smaller pitch, particularly in the interconnect process. Transistor optimization allows volume production of Pentium II microprocessors at 450 MHz. High yield has also been achieved, both before and after a 5% linear shrink of the initial 0.25µm design rules.

Introduction

Process 856 (P856) is Intel's quarter micron (0.25µm) logic technology. In developing P856, the important goals were to achieve low cost through high equipment re-use, deliver a gate delay improvement of 30%, and deliver high yield. An equipment re-use goal of 70% was set: the actual level achieved was 85% [1]. A performance goal of 30% transistor delay improvement was set: this was exceeded by 18%. The yield improvement curve for the P856 is the fastest of any Intel process so far.

Each generation of high-performance, low-power microprocessor products requires progressively faster transistors with lower operating voltage, produced with higher density. Historically the rate of improvement in gate delay has been 30% per generation. Normally it takes two to three years to develop a new technology, and each technology generation is progressively more expensive. Through scaling and the introduction of key architectural features such as halo NMOS, P856 delivered a better than 30% delay improvement at certification, the key checkpoint for volume manufacturing.

A second post-certification technology enhancement project delivered a 5% linear shrink with an additional 18% delay improvement, using the same equipment set. This represents nearly a half technology generation improvement in performance and yield, and it was delivered at very low cost. The post-certification improvement was achieved through control improvement and further transistor scaling, including a reduction of gate oxide thickness, enhanced halo processing, and general optimization of transistor implant conditions. This transistor enhancement has been critical in achieving good binsplit for Pentium II processors at 450 Mhz.

In this paper, we describe the important architectural features in P856 that enabled scaling of the interconnect process and transistor enhancement. The transistor improvements made in the pre- and post-certification stages are described. We discuss some of the important issues for interconnect processing with quarter micron features. We also describe the approach used to achieve a 5% shrink of the initial design rules.

Transistor Integration

P856 Architectural Enhancement

A fundamental constraint for short channel length transistors is that as the channel length is reduced to improve drive current, the barrier to off-state leakage is decreased. Throughout the development of P856, the transistor was optimized to achieve the best Idsat at a given margin to leakage, while also striving for low capacitance. High transistor performance in P856 was achieved through aggressive scaling to 40.8A electrical

gate oxide and sub-quarter micron poly dimensions, and through the addition of the following architectural enhancements, to be described in detail:

- Silicon pre-amorphization implants
- NMOS and PMOS halo implants
- Junction compensation implants

Like the previous generation P854 (0.35µm) CMOS process, the P856 process flow uses 200mm P-/P+ epi wafers and begins with shallow trench isolation followed by implantation of N and P wells. The gate oxide thickness is scaled from 60A on P854 to 40.8A on P856. Complimentary doped polysilicon is used to obtain matched Vt in N- and P-MOS devices. Nitride spacers are used to separate the deep source drain regions from the shallow source drain extensions. TiSi2 is selectively formed on polysilicon and source drain regions, obtaining a worst-case sheet resistance of 5 Ω/sq. The transistor structure is illustrated in Figure 1.

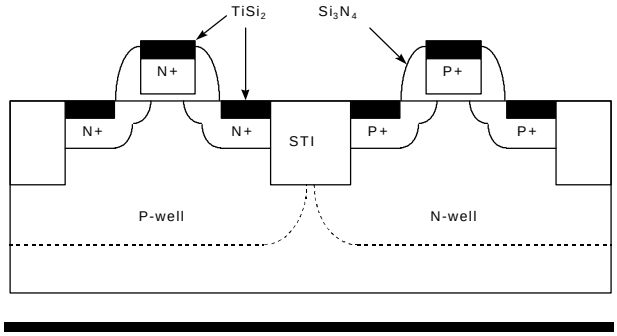


Figure 1: Schematic cross section of transistors

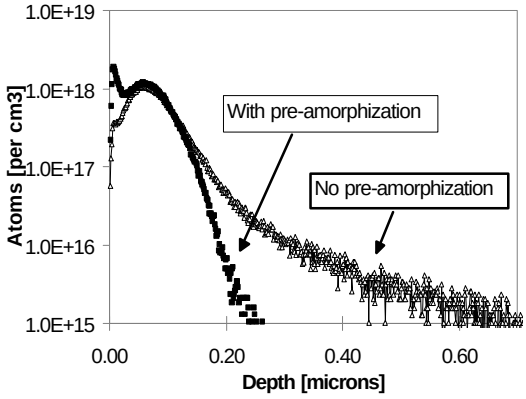


Figure 2: SIMS depth profile shows reduction in As implant tail due to Si pre-amorphization

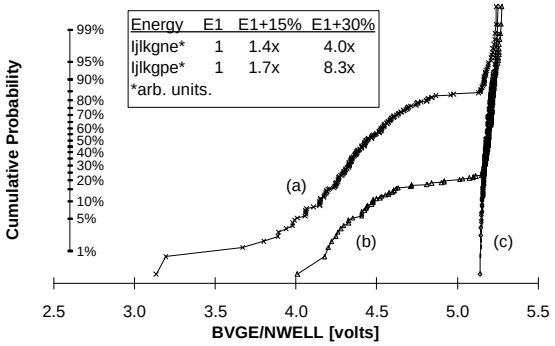


Figure 3: Increased Si pre-amorphization dose reduces the gate breakdown voltage. BVG failure rate is shown vs. a) 2X PA dose, b) 1.5X PA dose, and c) nominal PA dose.

Silicon Pre-Amorphization Implants

A silicon implant is introduced in P856 after poly gate definition. It is used to create an amorphous layer in the polysilicon gate and source/drain regions of both the N and P devices. The amorphous layer reduces the channeling tails of subsequent implant steps resulting in abrupt implant profiles (see Figure 2). Reducing the lateral implant tails under the poly gate region is key to controlling the sub-threshold leakage in short channel devices. The dose and energy of the Si implant need to be high enough to amorphize the underlying region without degrading the gate oxide. Figure 3 shows that gate oxide leakage increases for higher energy implant, and that gate oxide failure, as measured by lower breakdown voltage (BVG), can occur when the dose is too high. The table inset in Figure 3 shows the impact on gate leakage.

NMOS and PMOS Halo Implants

The short channel behavior of both NMOS and PMOS transistors was further enhanced by the introduction of halo implants. The halo implant is a high-angle implant introduced after Si pre-amorphization in the same lithography step used to dope the source/drain extension regions. Since the halo implant uses a high angle it must be done in four 90-degree rotations in the implant tool to ensure both sides of the channel are doped and that transistors oriented in both X and Y directions get doped. The halo implant uses the same implant type as the original well dopant (for example, N type dopant for the Nwell of the PMOS device).

The halo implant, together with the well implant, sets the threshold voltage of the transistor. By reducing the initial well implant dose and introducing the halo implant after

gate patterning, a non-uniform channel doping profile is achieved. Due to the angled implant, short channel devices receive a higher dopant concentration than do longer channel devices. There are several benefits when these implants are optimized. The halo implant reduces the Vt roll-off in short channel devices as shown in Figure 4. Since the same Vt is achieved with lower average channel concentration, the Vt with substrate bias is reduced as shown in Figure 5. Most important, higher Idsat at target is achieved because with a given Vt, the halo device has a more abrupt drain-channel junction and higher channel mobility than a non-halo device.

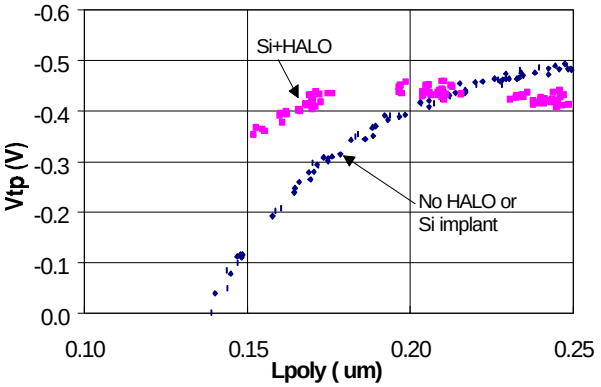


Figure 4: Reduction in PMOS threshold voltage roll-off with Si pre-amorphization and halo implant

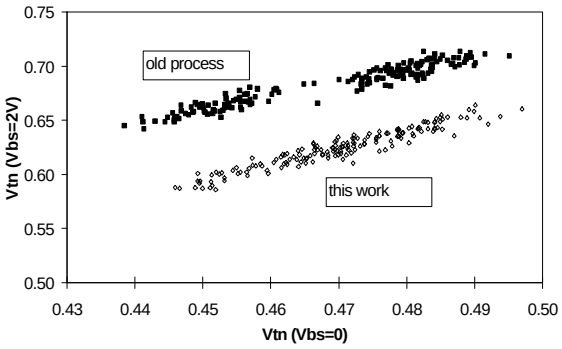


Figure 5: Reduction in substrate bias Vt effect

Junction Compensation Implants

The third major transistor modification on P856 is the use of compensation implants to reduce junction capacitance. AC parameters play an increasingly important role in overall transistor performance, and junction capacitance was a high leverage parameter contributing to the performance of P856. A compensation implant is introduced in both N and PMOS devices during the same lithography sequence used for source and drain (S/D) implants. This implant uses the same type species as the S/D implant but with a lower dose and higher energy to give a more graded implant profile at the junction (see Figure **6**). The compensation implant conditions were chosen to give approximately a 20-30% reduction in junction capacitance (see Table 1) with no degradation of the isolation performance or the implant penetration of the gate oxide.

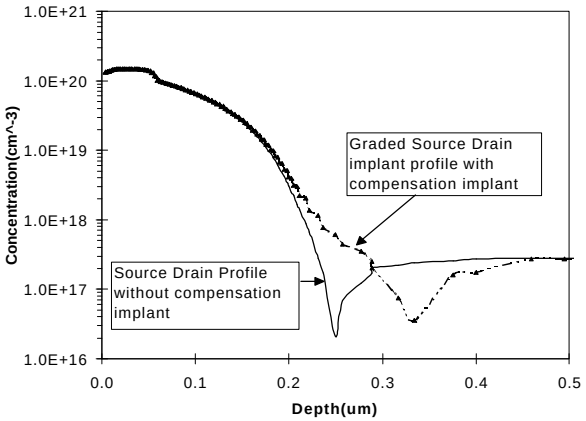


Figure 6: Junction doping profile with the addition of a compensation implant to reduce junction capacitance

Type	Before	With	Change
N	1.0 fF/µm2	0.7 fF/µm2	-30%
P	1.25 fF/µm2	1.0 fF/µm2	-20%

Table 1: Junction capacitance area component reduction attributed to compensation implants

Transistor Performance Results

P856 was certified in Q3 1997 using halo implants, Si pre-amorphization implants, and n+ junction compensation [3]. Based on the common industry metric of 1nA/µm worst-case device leakage, the Idsat target of 0.585mA/µm for NMOS and 0.250mA/µm for PMOS was achieved. A simulated transistor delay metric known as FEM95 showed that the performance goal of a 30% delay improvement over P854 had been achieved.

Time	NMOS Idsat	PMOS Idsat	FEM95 vs P854	FEM95
Certification	0.585	0.250	-33.2%	ref
Cert+2Q	0.670	0.295	-45.8	-18.8%
Cert+4Q	0.700	0.310	-49.9	-23.6

Table 2: Idsat target and FEM95 benchmark results as a function of time (in quarters) from certification (the FEM95 reference is P854)

To rapidly deliver significant additional performance, two process revisions were developed and implemented within a year of certification. The enhancement involved further thinning of the gate oxide to 40.8A, scaling of the poly target due to improved poly control, implementation of a p+ junction compensation implant, and re-optimization of the NMOS and PMOS halo, well, and S/D implants.

The halo implant re-optimization allowed a reduction in the N and P well surface implant, favoring an increase in the halo implant. The resulting transistors have well behaved sub-threshold characteristics (see Figure 7). As shown in Figure 8, we achieved Idsat at 1nA/μm of 0.755mA/μm for NMOS and 0.350mA/μm for PMOS. Accounting for the channel length control margin, we achieved industry pace-setting Idsat at target of 0.700mA/μm for NMOS and 0.310mA/μm for PMOS [4],[5]. These results are summarized in Table 2.

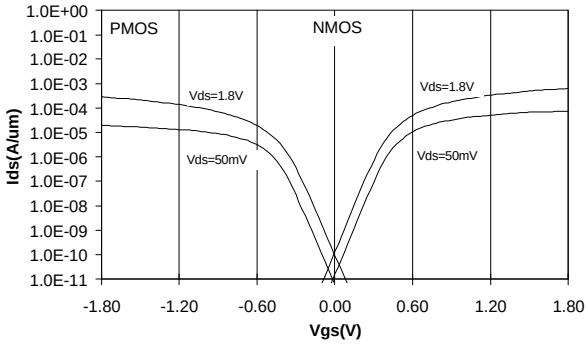


Figure 7: IV sub-threshold characteristics for NMOS and PMOS devices for target devices

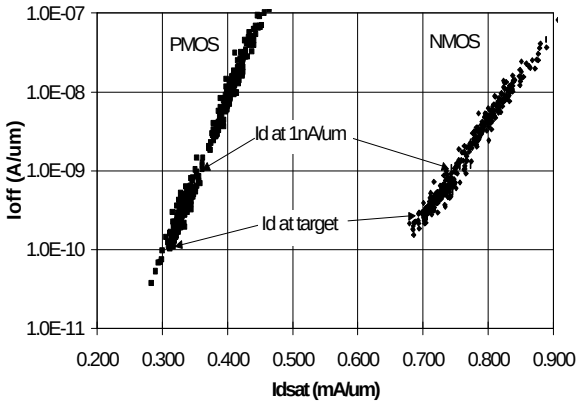


Figure 8: NMOS and PMOS drive current vs. leakage (the reference leakage current is 1nA/μm)

The improvement in performance has been demonstrated using the Pentium II microprocessor. Maximum speed measurements made at low-voltage and low-temperature conditions primarily show the improvement made in transistor performance. Under these conditions there is little influence from interconnect RC delay, because the interconnect sheet rho is reduced at low temperature. Figure 9 shows the progression in microprocessor path delay (period) as a function of time from certification. (In this figure, the data is smoothed for clarity, and the same stepping and test program is used in all cases.) A net 18.1% delay improvement has been observed on the same stepping of the Pentium II microprocessor. While there is dilution of the transistor improvement due to RC limited paths, with this enhanced process, better than 50% Fmax improvement has been achieved in microprocessor speed compared to the prior 0.35μm technology [6].

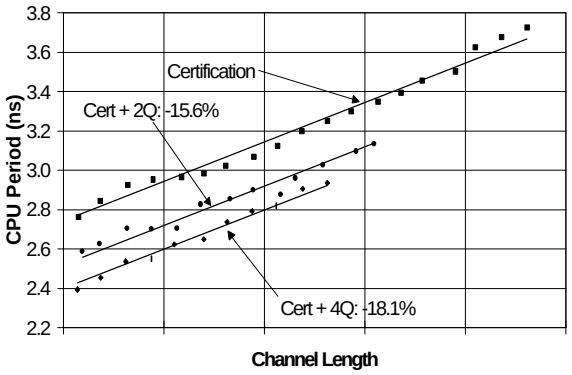


Figure 9: Microprocessor low voltage/low temperature delay improvement from post-certification process enhancement

All of the benchmarks discussed in this section are based on 1.8V transistor test conditions, and the P854 reference assumes the P854 and P856 run under nominal 2.5V and 1.8V conditions. To enable further performance enhancement, the reliability characterization of P856 was converted to a 2.0V nominal criteria. On products that can tolerate higher power consumption due to increased supply voltage, the 2.0V operation improves performance. Microprocessor characterization shows that there is an additional 9-10% frequency enhancement at 2.0V compared to 1.8V. At certification, P856 met the reliability goals for 2.0V operation.

Interconnect Integration

P856 uses five metal layers that are optimized for microprocessor performance and density. Table 3 shows the intended functions for each layer. Intel's technologies for logic are optimized for high aspect ratio to provide the most competitive RC performance at the

best density. The M1 to M3 layers use tight pitch, which is necessary for good SRAM and logic cell routing density. The M4 and M5 layers use wide pitch and high thickness, resulting in the low sheet rho needed for power distribution and cross die interconnect.

As with previous Intel processes, the metal stack is Ti/Al-Cu/Ti/TiN, which provides low line and via resistance while meeting electromigration requirements. Also, as before, the first inter-layer dielectric (ILD) above poly is Boro-Phosphosilicate-Glass (BPSG). The BPSG is planarized using chemical-mechanical polishing (CMP). The remaining ILD layers are PTEOS oxide that use a deposition followed by an etch-back process followed by CMP planarization. The CMP steps improve layer planarity, which is necessary for the uniform lithographic and etch processing of multi-layer interconnects. Contacts and vias are all filled with tungsten plugs formed by blanket tungsten deposition followed by CMP.

Layer	Pitch	Thick ness	AR	Purpose
M1	608 nm	480 nm	1.6	local connections
M2	882	900	2.0	intermediate length RC
M3	882	900	2.0	intermediate length RC
M4	1520	1325	1.7	power / long RC
M5	2432	1900	1.6	power / long RC

Table 3: Metal layer pitch, aspect ratio, and intended applications

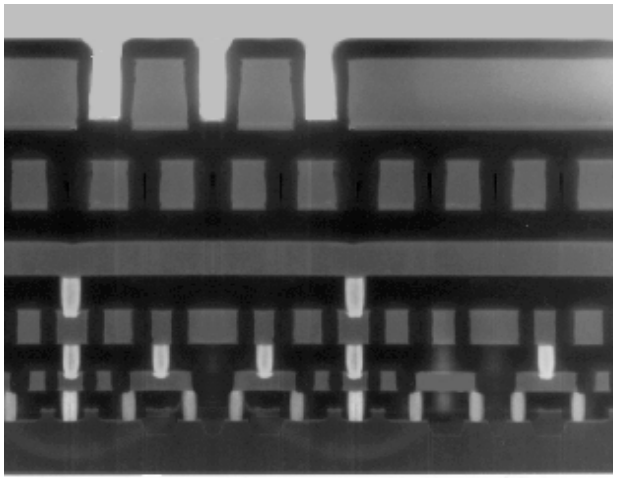


Figure 10: Five-layer metal interconnect cross section

To achieve cost savings, most of the metal-processing tools used in P856 were used in P854. The same stepper, metal deposition, contact etcher, metal etcher, and planarization equipment are used. A key challenge in the P856 interconnect has come from optimizing the lithographic and etch processes to work with the 20% smaller pitch of P856.

Just as Poly stretches the line width capability of DUV tools, Metal 1 patterning challenges the DUV lithography for space-limited capability, as the minimum space required is beyond the wavelength limits. This tight pitch (608 nm) demands thin photoresist for resolution, which in turn degrades the margin for metal etch due to resist erosion. The resist erosion results in poor metal line profile (shelving) and poor metal line critical dimension (CD) control.

Stringent control in depth of focus is also needed to ensure the integrity of the lithographic patterning. In order to achieve a planar surface for metal lithography, CMP is used prior to metal deposition for both ILD0 and contact plug steps. However, density variation causes local ILD erosion during CMP, which can result in severe variation in topography. For example, a depression as deep as 180 nm has been seen on the surface near a boundary between a dense memory array area and a loose periphery area. This depression causes a local area to be printed out of focus and results in a distorted metal line, as shown in Figure 11. Improved oxide and tungsten polishes that reduce the topographical step have been developed to ensure enough depth of focus on the surface.

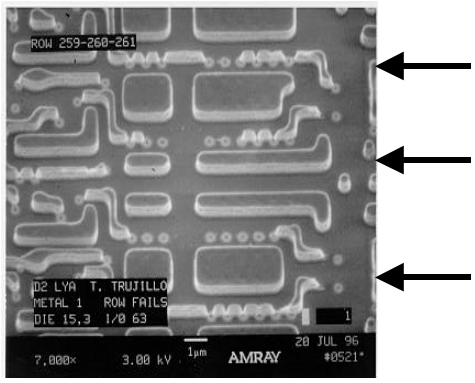


Figure 11: Metal 1 line distortion caused by ILD erosion induced out of focus lithography

Another limitation of lithographic capability is evident in the pullback at the end of a metal line. This pullback can cause a reliability problem when it is so severe that the metal line does not adequately cover a contact at the end of the metal line. Figure 12 shows a Metal 1 void bake

failure due to Metal 1 pullback and improper contact coverage. Twenty to 160 nm pullback has been detected in Metal 1 lines, depending on whether the structure is nested or isolated and on the location on the wafer. One solution to this pullback problem is to use optical proximity correction (OPC). These features compensate for the lithographic pullback effect at the end of the metal line. Contact coverage better than 75% has been achieved with OPC improvement, and the failure has essentially been eliminated.

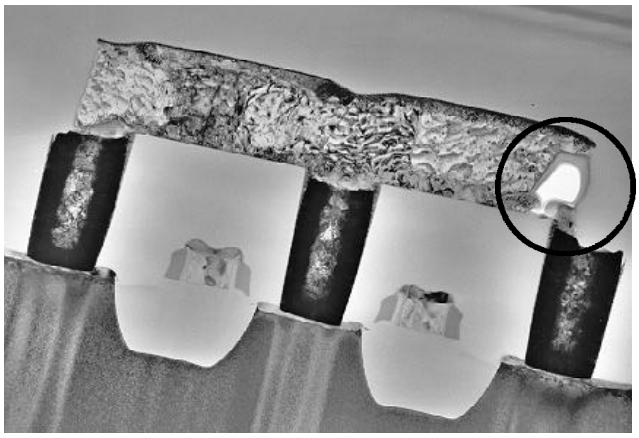


Figure 12: Example of a Metal 1 void failure due to pullback, before process optimization

The P856 technology also places stringent demands on the metal etch control. Magnatron current and RF power are optimized to reduce the erosion of photoresist during etch and to provide enough sidewall passivation to protect the metal profile. A vertical Metal 1 profile without undercut and shelving was achieved while providing good metal CD control (Figure 13). In the high aspect ratio M1 process, incomplete etching due to cross wafer thickness and CD non-uniformity can result in metal stringer defects. This is addressed by limiting the M1 sputter deposition target lifetime, controlling the M1 grain size through minimum deposition chamber heating, improving the uniformity of metal thickness, reducing the metal electrical CD, and slightly increasing the over-etch time.

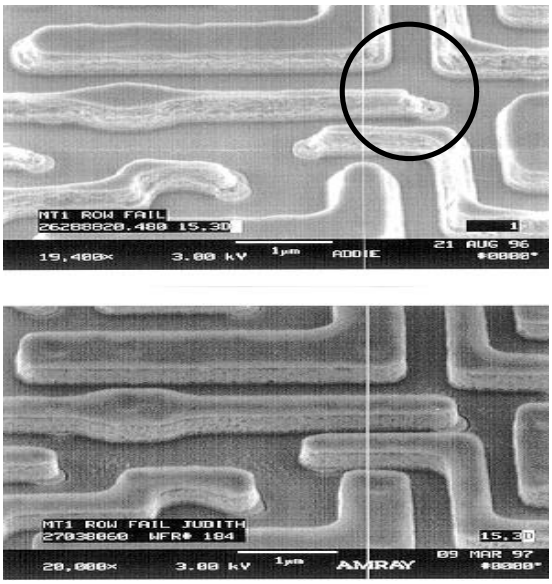


Figure 13: Metal 1 profile before optimization (top) showing shelving and M1 pullback, and after optimization (bottom) with good vertical profile & good contact coverage

Just as Metal 1 challenges the DUV limits, the Metal 2 and Metal 3 patterning with very thick films stress the I-line lithography limits. Both metals required significant improvements on processing issues, such as shelving of the metal profile, pullback at the end of metal lines, and bridging between narrow spaces. Optimized operating conditions have been determined for individual I-line lithographic tools to provide the best focus and exposure window. Together with an optimized metal-etching recipe, shelving is eliminated in the metal profile. Optimized reticle sizing for narrow spacing is used to provide adequate margin for the metal bridging. OPC is also used in Metal 2 and Metal 3 reticles to reduce pullback. The combination of these enhancements has successfully provided needed process capability in a production environment.

5% Shrink

A 5% linear shrink, known as P856.5, was applied to P856.0 in order to reduce die cost. A 5% technology shrink has been used in Intel in many generations as a standard means of cost reduction. Due to high equipment re-use in P856, the margin for shrink initially appeared tighter than in previous technologies. This required increased optimization of individual layers. Process margins and design rule margins were examined closely in order to achieve a "smart shrink" for minimum margin loss on the tightest part of the technology.

The smart shrink strategy uses optimum sizing for all critical layers and optimum targeting for critical

dimensions. For example, the high performance interconnect has high aspect ratio metal spacing as well as lines. The lithographic and etch margins are more critical for spacing than for lines in manufacturing. Therefore, a strategy of avoiding or minimizing shrinking metal spacing was adopted. Whenever possible, the metal line width rather than the space is shrunk. This helps ensure no degradation in speed due to the increased cross-talk if spacing were shrunk.

The reduced metal line width in the shrink technology does reduce certain design rule margins such as metal overlap of underlying via and metal enclosure of top via. To overcome this difficulty, OPC techniques were used creatively to systematically prevent the degradation of the design rule margin. For the contact layer, the proximity effect from clustered contacts became much worse after shrink, and a special selective sizing method was used on the reticle to restore the process margin. Due to very tight constraints, many layers required re-characterization. Table 4 shows the approaches used for critical DUV and I-line layers.

Layer	Shrink Strategy	Re-Characterized	Mask Fix/OPC
Isolation	Line / Space	Yes	No
Poly Gate	Space	Yes	OPC
Contact	Space	Yes	Selective sizing
Metal 1	Mostly line	Yes	Improved OPC
Via 1	Line /Space	Yes	No
Metal 2	Mostly line	Yes	Improved OPC
Via 2	Space	No	No
Metal 3	Mostly line	Yes	Improved OPC

Table 4: Shrink strategy

The shrink reduces SRAM cell size from 10.26μm^2 to 9.26μm^2. With the 5% shrink, there is a 15% increase approximately in sorted good die due to smaller die size. The shrink technology went to production four months after product tape-out thereby setting a new benchmark. All quality and reliability requirements were met, and products were synchronized just in time for the volume production ramp.

Conclusion

At certification, P856 met its principal performance, yield, and density goals, while achieving an 85% equipment re-use rate. Within one year of certification, and with only low-cost changes, a further 5% shrink was implemented. With the same equipment set, re-optimization of the transistors combined with control enhancement has allowed an 18% improvement in gate delay, more than a half technology step.

Acknowledgments

Many people contributed to the results discussed in this paper. Key people include Max Wei, Brian Johnson, Maurice DeCourcy, Domenic Pipitone, Karen Lubic, Brett Huff, Haiping Dun, Sam Hu, Bill Kavanaugh, Yung-Huei Lee, Wallace Lin, Steven Soss, Andrew Stack, John Mardinly, Li-Jia Ma, K.C. Patel, Tom Castro, Nevine Malek, Mike Maxim, Melinda Hoppe, and Ajay Chatterjee. In addition to those mentioned, we acknowledge the contributions of many others from the CTM, PTD, and virtual factory module and integration groups.

References

[1] M. Bohr, Y. El-Mansy, "Technology for Advanced High-Performance Microprocessors," *IEEE Transactions on Electron Devices*, March 1998, pp. 620-625.

[2] S. Wolf, *Silicon Processing for the VLSI Era, Volume 3: The Submicron MOSFET.*

[3] M. Bohr, S.S. Ahmed, S.U. Ahmed, M. Bost, T. Ghani, J. Greason, R. Hainsey, C. Jan, P. Packan, S. Sivakumar, S. Thompson, J. Tsai, and S. Yang, "A High Performance 0.25um Logic Technology Optimized for 1.8V Operation," *IEDM Technical Digest*, 1996, pp. 847-850.

[4] S. Venkatesan, A. Gelatos, B. Smith, R. Islam, et.al., "A High Performance 1.8V, 0.20um CMOS Technology with Copper Metallization," *IEDM Technical Digest*, 1997, pp. 769-772.

[5] M. Chang, J. Ting, J. Shy, L. Chen, "A Highly Manufacturable 0.25 um Multiple-Vt Dual Gate Oxide CMOS Process for Logic/Embeded IC Foundry Technology." *1998 Symposium on VLSI Technology Digest*, pp. 150-151.

[6] J, Schutz, R. Wallace, "A 450MHz IA32 P6 Family Microprocessor," *ISSCC Technical Digest*, 1998, p. 236-237.

Authors' Biographies

Adam Brand received his BSEE and his MSEE from the Massachusetts Institute of Technology in 1991. He joined Intel in 1991 and is currently working in the California Technology and Manufacturing 0.25μm Device Group. His interests include transistor

performance optimization, high voltage device development, and circuit modeling. His email address is adam.d.brand@intel.com .

Aravinda Haranahalli received an MS in Physics in 1976 and a Ph.D in Materials Engineering 1980 from the University of Florida. He joined Intel in 1984 and has held various management positions in technology, manufacturing, and business development. He currently manages interconnect technology development for 0.2µm. Before joining Intel he held technology positions at Texas Instruments and Fairchild. His current interests include technology, manufacturing, and business management. His email address is aravinda.r.haranahalli@intel.com .

Ning Hsieh received a Ph.D. in Materials Science from Northwestern University in 1979. He worked for various semiconductor companies including IBM, Fairchild, and DEC. He joined Intel in 1993 and has worked in CTM Technology Development since then. His work experience is mostly in process integration. He has published six external papers and has six patents. His email address is ning.hsieh@intel.com .

Yi-Ching Lin graduated from the University of California, Berkeley with a Ph.D. in EECS in 1981. Prior to joining Intel in 1987, he was with Texas Instruments and Monolithic Memories, Inc. He has been working in the area of process integration for microprocessor, Flash and EPROM memories. He had also worked on technology transfer from D2 to foreign foundries, including those located in Taiwan and Japan. His email address is yi-ching.lin@intel.com .

George E. Sery is an Intel Fellow and director of Device Technology Optimization in Intel's California Technology and Manufacturing group. Mr. Sery is currently responsible for directing process characterization, performance improvement, and capability enhancement for Intel's 0.25 micron CMOS logic technology. He received a B.S. and M.S. in electrical engineering from the University of Minnesota in 1976 and 1978 respectively. He joined Intel in 1978 as part of the SRAM Technology Development group. He has been involved with the development of NMOS and CMOS technologies for logic, SRAM, and Flash memory applications. For each technology, he has led the device physics team responsible for device development and process characterization. His email address is george.sery@intel.com .

Nicky Stenton received a M.S. in Materials Engineering from Lehigh University in 1982. She joined Intel in 1982 and most recently has been working on transistor process development in the California Technology and Manufacturing P856 Integration group. Her email address is nicky.stenton@intel.com .

Been-Jon Woo received a B.S. in Chemical Engineering from the National Taiwan University in 1975 and a Ph.D. from USC in 1979. She joined Intel in 1984 after working at Fairchild. She has worked in EPROM, Flash, and logic technology integration in the California Technology and Manufacturing group. She is currently the 0.25µm transistor integration manager. Her email address is been-jon.k.woo@intel.com .

Shahriar Ahmed joined Intel in 1985, initially as a interconnect device engineer working on Process 448. He subsequently was part of the team that developed P648 and coordinated the final transfer to high-volume manufacturing. Shahriar then moved on to be the device engineer for Intel's first bi-CMOS process. His next project was P856, which he developed together with a team from California Technology and Manufacturing. Currently he is in working on 0.18µm process development. His email address is shahriar.ahmed@intel.com

Mark T. Bohr joined Intel in 1978 after receiving a MSEE from the University of Illinois. He has been a member of the Portland Technology Development group since 1978 and has been responsible for process integration and device design on a variety of DRAM, SRAM, and logic technologies, including recently 0.35µm and 0.25µm logic technologies. He is an Intel Fellow and director of process architecture and integration. He is currently directing development activities on 0.18µm and 0.13µm logic technologies. His email address is mark.bohr@intel.com .

Scott Thompson joined Intel in 1992 after completing his Ph.D. under Professor C. T. Sah at the University of Florida on thin gate oxides. He has worked on transistor design and front-end process integration on Intel's 0.35, 0.25, and 0.18µm silicon process technology design for the Pentium® and the Pentium® II microprocessors. Scott is currently managing the development of Intel's 0.13µm transistor design. His email address is scott.thompson@intel.com .

Simon Yang received his B.S. in Electrical Engineering from the Shanghai University of Science and Technology (Shanghai, PRC). He then received his M.S. in Physics and a Ph.D. in Materials Engineering from the Rensselar Polytechnic Institute in New York. He joined Intel after graduating in 1987 and is currently leading transistor and yield improvement for Intel's 0.18µm logic technology. His email address is shi-ning.yang@intel.com.

The Quality and Reliability of Intel's Quarter Micron Process

Krishna Seshan, Technology and Manufacturing Group, Intel Corp.
Timothy J. Maloney, Design Technology, Intel Corp.
Kenneth J. Wu, Technology and Manufacturing Group, Intel Corp.

Index words: quality, reliability, ESD protection, electromigration, mechanical stress

Abstract

This paper describes how the quality and reliability of Intel's products are designed, measured, modeled, and maintained. Four main reliability topics: ESD protection, electromigration, gate oxide wearout, and the modeling and management of mechanical stresses are discussed. Based on an analysis of the reliability implications of device scaling (the process of a planned reduction of dimensions and operating parameters), we show how these four topics are of prime importance to component reliability. We conclude with a brief discussion of the future challenges of energy scaling.

Introduction

The maintenance of quality and reliability is an important aspect of Intel's product goals. Intel's goal for reliability is to strive to reach failures-in-time (FITs) to less than the hundred range by the end of the century. FITs are defined as the number of device failures in 1.0E9 or billion device hours. In order to reach this goal, defects have to be reduced to less than 100 ppm. For more details refer to Intel's *Component Quality and Reliability Handbook* [1]. Intel's reliability goals are shown in Figure 1.

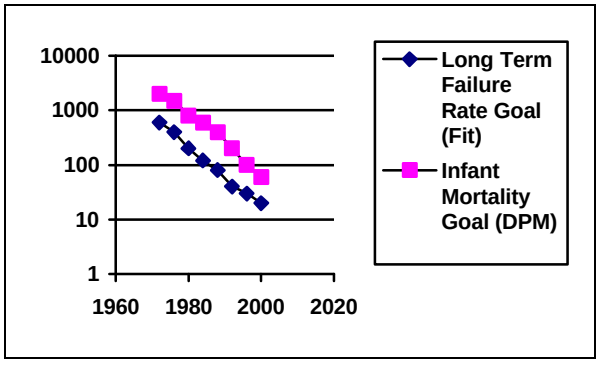


Figure 1: Failure rate (FIT) & defect rate (DPM) goals (the top curve represents infant mortality goals, which can only be achieved by reducing defects)

In this paper we discuss four of the main topics pertaining to the maintenance of reliability for the 0.25µm process also known as P856 so that Intel meets its product goals. The topics are as follows:

1. electrostatic discharge (ESD) protection
2. electromigration failures resulting from increased current densities
3. gate oxide wearout failures resulting from decreasing gate oxide thickness
4. modeling and management of the effects of mechanical stress resulting from silicon-package interactions

There are two major challenges to maintaining quality and reliability. The first is the continued increase in die size. Even though transistor density increases, new features and functionality are added to the microprocessor causing die size to grow. This is depicted in Figure 2, which shows the size growth of Intel's products. Some of the microprocessors using the 0.25µm process generation are as large as 800 mils on the side and have in excess of seven million transistors.

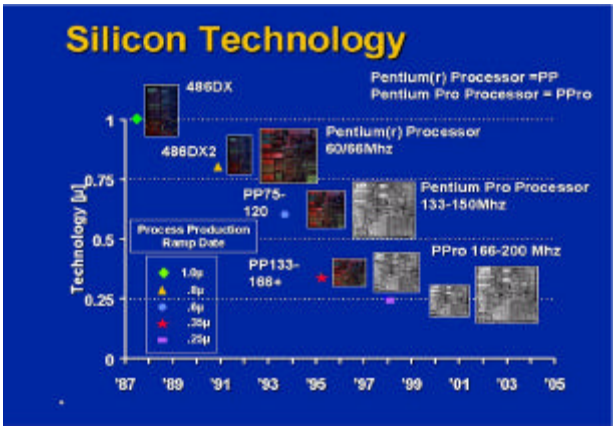


Figure 2: The continued growth of the microprocessor despite the increase in transistor density

The die sizes and the area per transistor for the products shown in Figure 2 are plotted generationally in Figure 3.

The graph shows the decrease in area per transistor (mil*2/transistor) that has enabled the three-fold compaction per decade. Also plotted is the die size in Mil -Sq. This is the square root of the area. Note that die size increases generationally, and that die sizes as large as 800 mils-square are allowed by the reliability envelope.

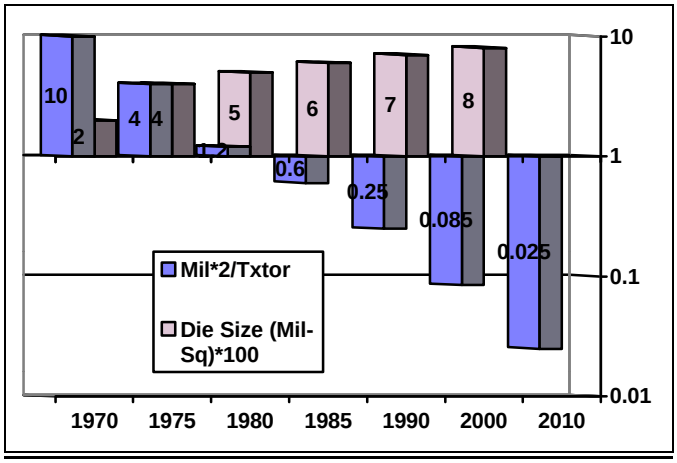


Figure 3: Generational graph of area per transistor and die size trends (increase in functionality contributes to the increase in die size)

As microprocessors grow in complexity, Intel's customers have come to expect improved reliability. The reliability of devices and packaged products is measured by subjecting devices to various reliability tests aimed at accelerating failures. The results of these tests are then displayed in a graph such as is shown in Figure 1.

Reliability Implications of Scaling

Before going into detail on the four reliability topics mentioned, we briefly discuss scaling and its implications on reliability.

Scaling is the process by which device dimensions are reduced or "scaled" from one process technology to the next. Continued scaling of transistors to improve speed results in increased frequency and this in turn requires an increase of current density in metal lines and vias. This increase accelerates failures by electromigration. As metal line dimensions are decreased, so is gate oxide thickness. The resulting thinner gate requires carefully designed protection against electrostatic discharge events. The thinner gates also suffer from the effects of wearout caused by the hot electron bombardment of the gate oxide. In order to provide increased protection, the operating voltage was decreased or "scaled" to 2.0 volts. This then leads to a decrease in current density and offsets the increase in frequency. A second level of protection is obtained by using design rules based on an understanding of the electromigration failure mechanisms.

Various aspects of quality and reliability constitute the so-called "Reliability Envelope." As device length L scales, various parts of this envelope scale as "K," where K is the scaling factor and is > 1. Therefore, the channel length would scale by L* (1/K). Table 1 shows several parameters of this envelope that will scale "ideally." Table 2 shows how "ideal scaling" applies to various aspects of reliability of the integrated circuit.

Scaling factor K>1	Ideal Scaling	Reliability Implications
Channel Length L	1/K	Latchup
and shallow junctions		Hot-electron effects
Gate Oxide Thickness	1/K	Oxide wearout and ESD protection.
		Process Charging
Metal line width	1/K	Electromigration

Table 1: Reliability impacts on ESD, electromigration, etc. caused by ideal scaling (note that this table only deals with "ideal " scaling on device dimensions)

The most important consequence of the data from Table 1 is that in order to maintain a constant "E_Field" and preserve gate oxide reliability, (that is, maintain the electric field across the gate) *operating voltage must be scaled.* This leads to so-called "supply voltage scaling" that is shown in Table 2.

Table 2 shows the main implications to component reliability from scaling: ESD, gate oxide wearout, electromigration, and stress.

Electrical Parameter	**Ideal Scaling with Scaled Supply Voltage**	**Reliability Implications of Scaling**
Operating Voltage	Vcc * 1/K	Hot e and gate oxide reliability are rendered equivalent in the scaled voltage scheme.

Device Current	1/sqrtK - 1/K	
Metal Current Den	K**1.5	EM, and self heating increases.
Die Size	Does not scale	Package stress effects on metal lines and dielectric layers.
Mechanical stress from die-package interactions	Does not seem to scale	Stress effects on Devices and interconnections.
Power dissipation per gate	1/k**1.5	Total power does not scale with Vcc This creates challenges for cooling.
Gate Delay τd	1/k**1.5	Main contributor to performance enhancements.
Delay x Power	1/k**3	Even though power delay per gate scales, total power does not.

Table 2: Impact of dimensional scaling on device electrical parameters with a scaled supply voltage

As can be seen from Table 2, some parameters do not scale at all, notably current and size. This has a significant impact on the amount of scaling that can occur. Reliability is affected by scaling because scaling gives rise to larger current densities, higher chip temperatures, and higher electric fields during device operation. However, if Vcc and process are both scaled, then electric field (E) can be maintained invariant. This then begs the question of how low Vcc can be scaled and how much power dissipation can be lowered? This question is dealt with in the Discussion section at the end of this paper. Other parameters such as stress and power that do not scale even with the scaled supply voltage are also discussed in the Discussion section.

We now return to the discussion of the four main topics of component reliability: ESD protection, electromigration, gate oxide wearout, and modeling and management of mechanical stress. These four topics are presented below with an introduction in the beginning aimed at the general reader in each topic.

Aspects of Electrostatic Discharge (ESD) Protection

In recent years, CMOS FET scaling, power supply voltage scaling, and FET engineering for performance have caused a continued need for ESD protection methods that can be easily applied to inputs and outputs, without interfering with process development. Despite the scaling of devices to sub-micron dimensions, where oxides break down at, say, 5V, and junctions and wells are shallower, the ESD test goals are the same (e.g. 2000V human body model). How have designers been able to achieve the same ESD performance with the new devices? The ideal low-cost ESD design exploits devices that are available "for free" as part of the process, and which do not need to be engineered for ESD performance and then made compatible with other goals. We will discuss how these ambitious design goals are met for the 0.25µm process.

To understand these protection methods, we define smooth ESD current paths through the chip for the possible ESD events in stress testing and in actual handling. The natural diodes to power and ground are used, and current paths are linked together with the help of power supply clamps. The latest designs for power supply clamps in the 0.25µm process technology take full advantage of device scaling, which only in recent years has made it possible to dissipate ESD-scale currents (on the order of amperes, but only for nanoseconds) within small amounts of chip area (bond pad size) by using MOS FET conduction. In earlier days, some kind of avalanche breakdown event had to be used, but sensitivity to process and triggering events made these methods very difficult to execute. With dual diodes for basic inputs and outputs, and special PMOS FET circuits for power supply linkage, smooth ESD current paths can be defined for nearly all varieties of chip interface with the outside world. PMOS FET clamping methods for ESD have become important for all of Intel's low voltage deep sub-micron CMOS products. They also help to solve the ESD protection problems for mixed voltage products, where compatibility with signals from earlier technologies is desired.

ESD Protection Issues for 0.25µm Process

The scaling of power supply voltages below 5V in recent years has meant that components need to be backward compatible, to some extent, with chips running on higher voltage supplies. Table 3 summarizes the situation with four typical CMOS integrated circuits processes of the past few years, with Proc1 being the last of the processes allowing a continuous 5 volts across the gate oxide.

Proc4 is the 0.25µm process, under discussion here.

Many designers use (what might be called) the dual diode principle as much as possible in their chips. For example, a typical CMOS input/output (I/O) pad such as in Figure 4 has driver devices T1 and T2, which have parasitic diodes to power and ground resembling the dual diodes of an input-only. Even though the NMOS T1 FET might have its source on a separate Vssp supply as shown in Figure 4, its diode to Vss (substrate) is a particularly good one if the CMOS process is on epitaxial silicon with a conducting p+ substrate (as used on Intel's 0.25µm process. This diverts most of the current in one polarity of ESD pulse. The other polarity is steered toward T2's inherent diode to Vccp, which can be optimized (or even augmented) through obvious layout methods. Thus not much current is being handled by the breakdown mode of the NMOS T1 device. In recent years, the NMOS device has become weaker and weaker in ESD due to self-aligned silicide (salicide) on the drain and source, and also because of lightly doped drain (LDD) structures. Even when salicide is blocked between drain and gate with a section of n-well [2], it is best to use dual diode current steering and avoid much breakdown current flowing through the T1 transistor during ESD. For that reason, dual diode methods are commonly used on outputs as well as inputs.

The final link in the ESD protection scheme is that between one power supply and another. Much work on the use of diodes for cross-linking similar power supplies has been done by Intel [3]. Less obvious is how to clamp dissimilar power supplies, such as Vcc to Vss. These stand-alone power supply clamps also can solve the problem of powerup sequencing (as when "similar" Vccx power supplies are powered up and may overstress their crosslinking diodes) and they have become increasingly popular as a result of their success.

5.0V	**3.3V**	**2.5V**	**1.8-2.0V**
Proc1	Proc1 low		
Proc2 hi	Proc2		
Proc3 hi+	Proc3 hi	Proc3	
	Proc4 hi+	Proc4hi	Proc4

Table 3: Compatibility of sub-micron CMOS technology: Proc1 is the last 5V process, Proc4 is the 0.25µm process

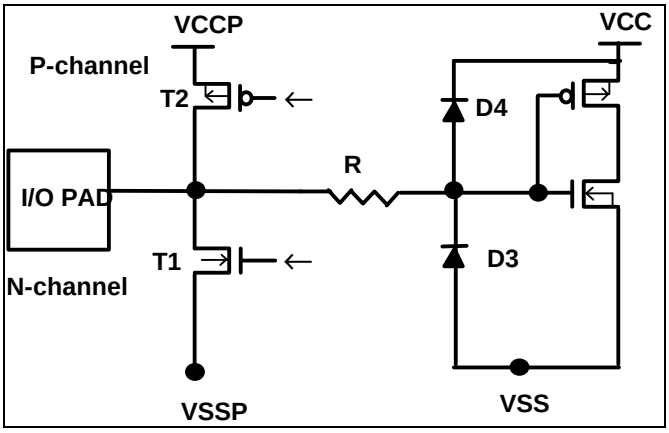


Figure 4: CMOS Input/Output buffer protection (T1 and T2 have built-in parasitic dual diodes that can be enhanced through layout)

Work at Intel in the 1992-95 time frame pointed toward the need for power supply clamping in ESD protection, and to the need for a design that would yield equivalent or better performance with each process generation. Our rigorous qualification standards required universal application of power clamp cells, meaning that each clamp should pass all standard ESD tests with some margin (>4-8kV HBM, >1.2kV CDM) in stand-alone mode, so that an arbitrarily small power supply would be protected. In addition, pulsed I-V behavior must be consistent with sinking at least a 2kV HBM peak current (1.33 amps) below the known danger Vcc voltage for all ordinary circuits in the process, even vulnerable ones (it was expected that every supply would have at least two clamps). These criteria, and the cost-driven desire not to add masks or tamper with the performance-engineered FET process, drove us away from NMOS FET clamps [4,5] because the salicided devices, even large ones, failed miserably on all ESD tests. Unsalicided NMOS devices resembling Worley's [5] had the same problems with size and CDM behavior.

However, the properties of power clamps made from PMOS FETs [6] were quite favorable. Dimensions at or near minimum could be used, so the disadvantage of PMOS current drive per unit gate width over NMOS was hardly noticeable. The PMOS devices (pmosclamps) in this driven-gate mode were very rugged in all the aforementioned ESD and pulsed I-V tests, sometimes almost impossible to destroy. We did not have to intervene in the performance-oriented process development cycle with wafer splits and such; we just

evaluated the process changes as they happened to confirm continued good performance. As the PMOS FET is free of the positive feedback and negative differential resistance effects of npn snapback [8], it appears to have no difficulty conducting uniformly over a large area, even in the high-voltage breakdown regime. The results discussed here are from devices fabricated on 0.35µm and 0.25µm processes, now in manufacturing. Some details of the processes have been released publicly [8,9].

The basic pmosclamp (Figure 5) is built around a large (around 3000µm) p-channel transistor (T1) of near-minimum gate length. Its gate is driven temporarily to ground in two ways. First, a MOS capacitor (C1) helps to overcome the capacitive coupling of the large gate to Vcc. But more important, the inverter driving the T1 gate is heavily weighted toward the NMOS device, T2, pulling the T1 gate low with considerable strength. The RC timer formed from T3 (long channel) and C2 sets the time constant (microseconds), while the first inverter trip point is set midway between ground and Vcc for high noise immunity.

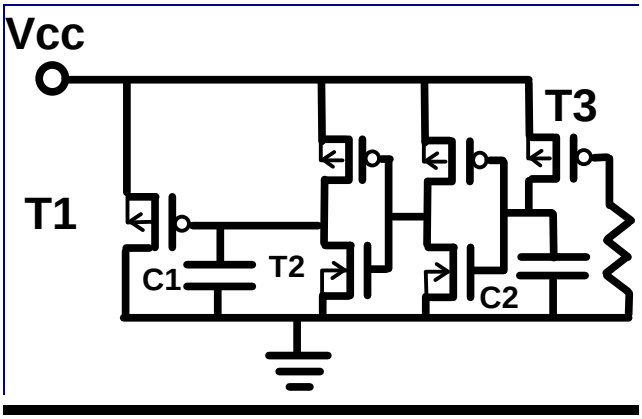


Figure 5: RC-timed circuit for PMOS FET power clamp (pmosclamp); T2 and C1 enhance gate drive

Figure 6 shows pulsed I-V curves for a pmosclamp protection circuit as in Figure 5, occupying about 7700 µ m^2 in a 0.25µm process. The "idealized" curve is from a test pattern with the T1 gate artificially hard-wired to Vss, and it shows how close we come to the desired grounding of the gate during the pulse. The I-V of a pmosclamp without the optimized trigger circuit including C1 and T2 (data not shown) shows clearly degraded characteristics as the gate does not fully turn on. The gate length used in Figure 3 matched for the two examples shown and happened to be well above the process minimum; the pmosclamp now routinely used in products has about a 10% higher pulsed current than shown, and its gate length is still substantially above the process minimum. The sub-threshold leakage of these pmosclamps is not an issue; it is below 1 µA until considerably above 100 C. The clamps were also shown to be robust against power supply noise, which was simulated on test chips with a high-frequency signal applied to the power supply node. There have been no reliability problems with the clamps on recent products.

Simulations of these circuits (using the standard process MOSFET model) match the pulsed I-V curves almost perfectly to the device model's voltage limit of 4-5V. Note how the pmosclamp continues to conduct (without destruction) up to 9-10V, far beyond the observed dc punchthrough voltage, around 5-6V. Thus the HBM self-protection of these clamps was measured at 8 kV, and CDM did not fail to the limit of the 2kV Keytek socketed tester. This is a hopeful sign for CDM protection of products as well. The empirical product results are very good so far.

The equivalent pmosclamp for the 0.35µm process has roughly the same I-V curve as in Figure 6, and it is in a still-reasonable 12000μm^2, but this uses over 50% more area than the 0.25µm process. The trend should continue until such MOS conduction of pulsed currents runs into thermal limits. All of this is because, with shorter FET channels, we can achieve more pulsed and dc current sinking per unit area as processes scale. In the days of process feature size of 0.8µm and above, the same PMOS FETs for sinking ESD currents would have been absurdly large. However, just in the past few years, it has become possible to sink more than an ampere of pulsed current through ordinary MOS conduction in a production PMOS FET less than the size of a bond pad. Moreover, while devices have scaled dramatically due to Moore's Law, ESD events have not—the human being, source of the HBM, has not scaled noticeably (!), and while electronic packages, source of the CDM, have proliferated into a variety of sizes and shapes, the CDM event is roughly the same as always. Thus device scaling once again teaches us to be on the lookout for opportunities as well as drawbacks.

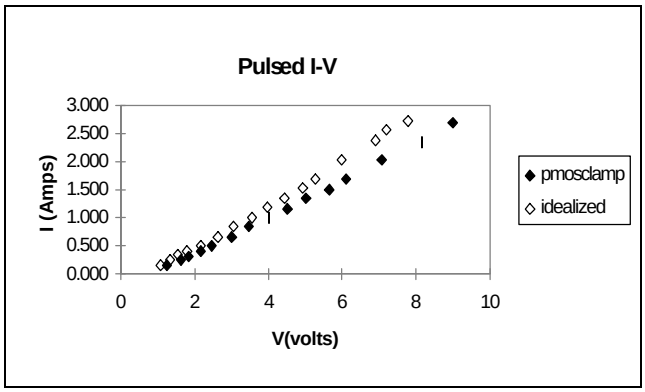


Figure 6: Pulsed I-V behavior of pmosclamp in 0.25µm process; idealized curve has T1 gate artificially grounded

For compatibility with signals from chips with earlier-generation power supply voltage, we want to enable an on-chip power supply Vccx, greater than the voltage that can be safely applied for long-term reliability to a gate oxide in the process. A stand-alone solution is often desired, allowing Vccx to be on when Vcc is off. This is allowed with the stacked pmosclamp (vtolclamp) as shown in Figure 7. There are two large (about 4000µm in the 0.25µm process) p-channel devices in the same n-well, with no required contact to the common node, thus allowing tight layout. The midpoint voltage of approximately Vccx/2 is set by long channel devices T4 and T5. This reference voltage allows only Vccx/2 to be dropped across any of the gates in the circuit. The trigger circuits were modeled after those in the pmosclamp, where the capacitors and NMOS FETs pull the gates as low as possible, and RC circuits time them out.

The ESD and TLP (Figure 8) performance of the vtolclamp was on par with the pmosclamp for both the 0.35µm and 0.25µm processes, with device sizes scaled similar to the pmosclamp as described earlier. About twice as much area was used for the vtolclamp due to conservative layout and circuit design. Prospects are good for compaction of the layout and for use of more aggressive circuits, improving the current per unit area of the vtolclamp in the future by perhaps 30-50%.

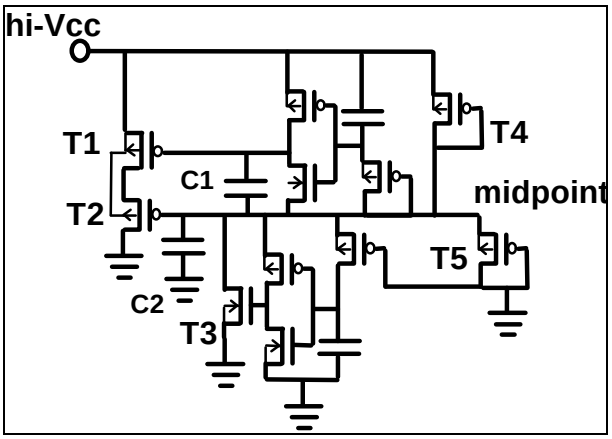


Figure 7: Circuit for stacked-gate high-voltage tolerant PMOS clamp (vtolclamp). T1 and T2 are large FETs built in the same n-well; circuitry drives their gates low temporarily. T4 and T5 bias the midpoint, rendering dc gate oxide voltages safe.

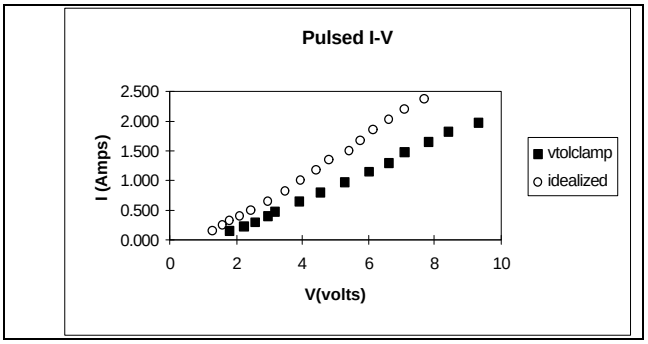


Figure 8: Pulsed I-V of vtolclamp in 0.25µm process; idealized curve has T1 and T2 gates artificially grounded

Electromigration

Another reliability concern is electromigration. Electromigration failures result from increased current densities. The current generation of highly integrated microprocessors, requiring dense interconnects and large amounts of current, has highlighted the concern for metal interconnect reliability. Formation of metal voids induced by electromigration during normal microprocessor operation will cause an interconnect open or high resistance resulting in malfunction or speed degradation.

The continued scaling of transistors for speed improvement in 0.25µm process technology achieves gate delays for n-channel and p-channel transistors of 3.5 and 7.8 psec (CV/I) [8], respectively, which is half that of the previous 0.35µm technology [9]. Although transistor drive current is about the same as in the previous

technology, this gate delay improvement increases the current density in metal lines and vias for high performance microprocessors. Five metal layers are developed to provide low metal line/via resistance and good electromigration performance. Metal interconnect pitch and thickness are summarized in Table 4 along with those for the 0.35µm technology [8-9].

	0.25um Technology		0.35um Technology	
Layer	Pitch (um)	Thickness (um)	Pitch (um)	Thickness (um)
Metal 1	0.64	0.48	0.88	0.60
Metal 2	0.93	0.90	1.16	0.80
Metal 3	0.93	0.90	1.16	0.80
Metal 4	1.60	1.33	1.70	1.70
Metal 5	2.56	1.90	N/A	N/A

Table 4: Metal layer pitches and thickness

Without major architectural changes in metallization to improve electromigration resistance, the thickness of M2 and M3 lines (used for intermediate interconnect) was increased from 0.80µm to 0.90µm. The inter-level dielectric process was optimized to support aggressive metal aspect ratios. However, the M1 line (used for local interconnect) thickness was decreased from 0.60µm to 0.48µm for narrow pitch planarity improvement. M1 current density increases significantly as compared to the other layers, and effort has been focused on process improvement, electromigration design rule characterization and implementation.

The thin Ti shunt layer used in Ti/Al-Cu/Ti/TiN metal stack forms a $TiAl_3$ compound at the end of silicon processing. The quality and thickness uniformity of the shunt layer has been found to be key to M1 line electromigration resistance. In addition, the top TiN ARC process has also been optimized to become a reliable shunt layer. However, metal width and length dependence of electromigration performance was not considered in the previous 0.35µm process technology. Therefore, during the 0.25µm process development, attention was paid to characterization of the electromigration of narrow and short metal lines.

M1 electromigration structures with different metal width and length were designed in the SRAM test chip; constant temperature and current density stresses were used in the characterization. Figure 9 shows M1 electromigration performance vs. line width. It is clear that minimum metal lines with 0.4µm drawn (pre-shrink) improves performance ~50% over the wide line structures, which are used for process monitoring. In microprocessors, the majority of M1 lines are used for local interconnect with minimum width for density improvement. Designers can use this narrow metal width electromigration advantage to support enhancement of transistor drive current density.

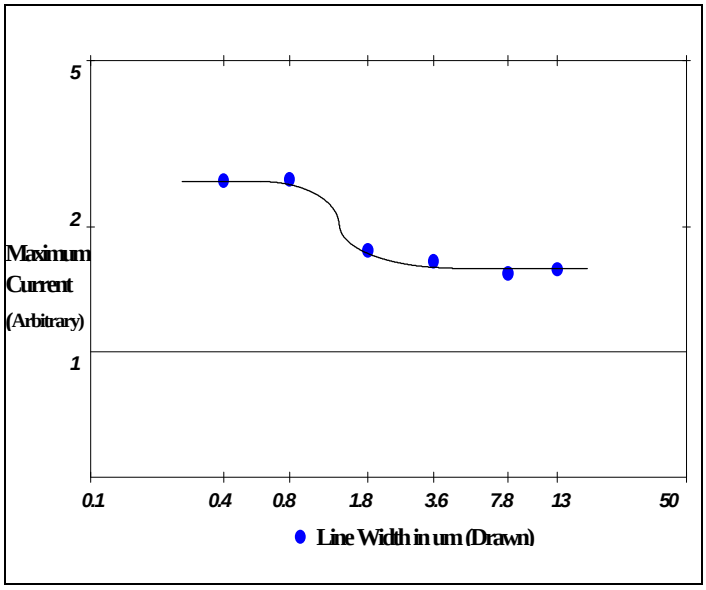


Figure 9: M1 electromigration performance vs. line width

It has been reported that short metal lines with tungsten plugs significantly improve electromigration due to vacancy back pressure effects [10]. Different stress current densities were applied to various metal length structures, and resistance changes vs. stress time were recorded to characterize the void formation. It is interesting to find that when metal line length is reduced to a certain value, void formation is saturated especially under relative low current density stress, indicating that electromigration depletion and back diffusion reach an equilibrium. The maximum percentage of line resistance increase is well below 30%, which is the electromigration failure criterion. Therefore, a short metal line electromigration design rule is conservatively implemented to support short local transistor interconnect.

Electromigration occurs during unidirectional current stress but not during AC current stress. A design rule is developed for AC signal lines, based on a maximum

allowable amount of resistive heating in the interconnects. Heat transfer through metal lines and inter-layer dielectric was simulated using a two-dimensional model. The design rule was derived based on a reasonable local temperature rise; and experimental data were taken on the test structures to calibrate the model. Electromigration requirements were built into the development of standard library cells, and design-rule checks were developed at the Function Unit Block (FUB) and Full Chip stages.

Gate Oxide Reliability

Gate oxide integrity is another one of the reliability concerns for high-density, high-performance microprocessors. Transistor and capacitor leakage current will be degraded under voltage and temperature stresses leading to function or speed failures. To ensure the product Defect Per Million (DPM) and Failure In Time (FIT) rate meet Intel's reliability goals, a high-quality gate oxide process is required for the ultra thin oxide technology.

The 0.25µm process technology gate-delay improvement comes from both transistor architect and gate oxide thickness reduction from 60 nm to 42 nm. Power supply voltage was reduced from 2.8V to 2V to keep the oxide electric field unchanged while maintaining acceptable hot electron reliability and reducing power consumption in high performance microprocessors. Although the electrical field across the gate oxide increases slightly on 0.25µm process technology, thin gate oxide reliability in terms of initial gate leakage, latent defect, and intrinsic integrity was well characterized during technology development. Besides the appropriate surface clean prior to the gate oxide growth and poly silicon gate deposition to improve oxide quality, process charging damage elimination and antenna layout rules are also implemented to ensure a low product field failure rate due to gate oxide breakdown.

Breakdown Voltage of Gate (BVG), Constant/Ramp Current Density Stress (JT), Ig Gate Current Measurement, and Time Depend Dielectric Breakdown (TDDB) test methodologies were used to characterize process charging induced gate oxide damage [11]. High Voltage Extent Life Test (HVELT) was also used on the test chip and on products to calculate the field product failure rate. Figure 9 shows the HVELT Time-To-Fail (TTF) distributions of 0.35µm and 0.25µm Test Chip and products after normalizing to the same electrical field and temperature. Product A in the 0.35µm process technology has the highest gate oxide failure rate though it still meets Intel's reliability goal of less than 0.1% failures in 10 years of product life. Detailed fault isolation and failure analysis unveiled gate oxide damage; and circuit layout analysis discovered a huge metal antenna ratio (to the gate area) was the culprit for oxide breakdown. The gate oxide failure rate of Product B without metal antenna violations is improved by 3.8X over Product A.

With this knowledge, the 0.25µm technology process charging induced gate oxide damage was extensively characterized on Inter-Layer-Dielectric (ILD) deposition/etching and metal etch processes. Appropriate test structures were designed in the Test Chip such that reliable metal antenna design rules could be derived. Reliability validation tools to check antenna layout rules were also developed to catch and fix any design rule violations before products are taped out.

Gate oxide failure rate in 0.25µm pre-mature process was measured on the Test Chip. The result is quite similar to that of the 0.35µm Product B shown in Figure 10. Subsequent processes resulted in a reduction in the product failure rate. TTF (without area normalization) distributions of the Test Chip, Product C, and Product D in mature 0.25µm technology are plotted in Figure 2. Taking a conservative approach, when data were fitted with a -1 sigma distribution as shown by the solid line in Figure 10, the 0.25µm product failure rate improves 7.2X over that of the 0.35µm product failure rate. This improvement has opened the way to additional reductions in gate-oxide thickness, improving process speed.

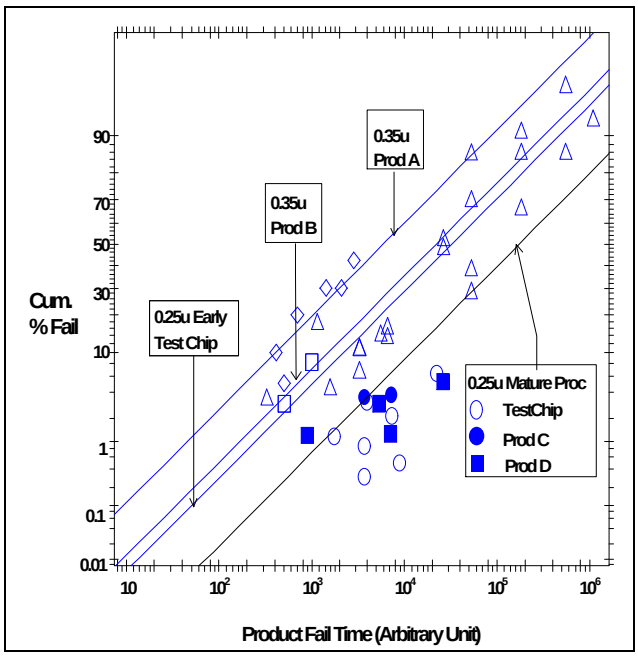


Figure 10: Comparison of 0.35µm and 0.25µm gate oxide breakdown TTF distributions

Aspects of Mechanical Stress

The last reliability topic we discuss is modeling of mechanical stress effects. These effects are the result of large die sizes and the use of new and novel package technologies such as Intel's plastic-mounted flip-chip technologies. As both the die size and the number of back-end layers increase, mechanical interactions between the package and the silicon die, metallization, and device become concerns of both reliability and failures. We now describe the approach taken to both model and mitigate such failures.

There are two parts to mechanical stress: the intrinsic part (σi) and the externally applied part (σe). The total stress is the sum of the two as shown in Eq(1).

$$\sigma \text{ total} = \sigma \text{ intrinsic} + \sigma \text{ applied} \quad (1)$$

We have used finite element models to calculate the magnitude of the strains resulting from σ applied, the externally applied stress. The basis of this model is shown in Figure 11 where the die and the package are treated as two beams. (Note that the "fliped-chip" is being modeled in this figure). We use this dual-beam approach to estimate stresses in various parts of the final packaged component.

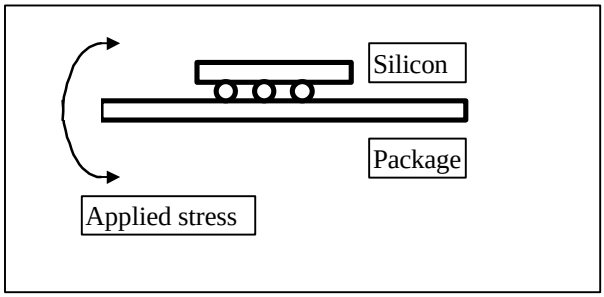


Figure 11: Externally applied stresses on silicon with the chip and package viewed as two independent beams

Using this kind of modeling, an informed choice can be made when selecting materials for various parts of the complete package. Materials are selected on the basis of compatible coefficients of thermal expansion, elastic moduli, and strength in order to maximize reliability performance.

A second use of this model is to examine in greater detail some of the spatial stress relationships. In order to perform these calculations, we start by making a "3-D finite-element mesh" of the package die. An example of this is shown in Figure 12.

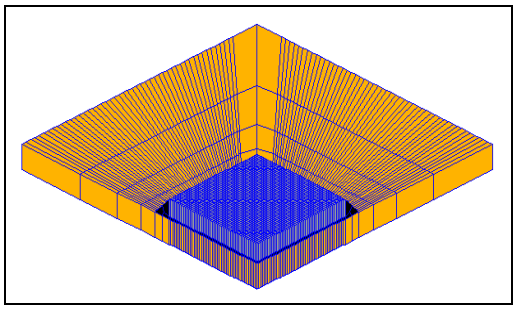


Figure12a: Global model showing a quarter-slice of a die (in gray) mounted on a plastic package (yellow)

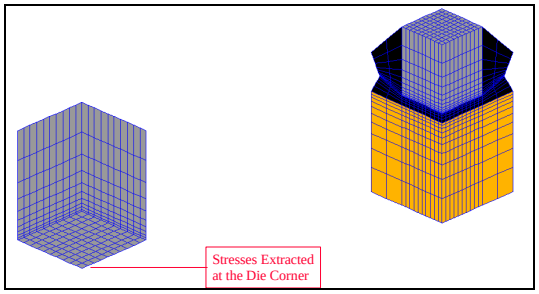


Figure 12b: Stress extracted at the die corner

Figures 12a and 12b show how global package models are meshed and how local stresses are determined. Figure 12a is the "global" model showing a quarter-slice of a die mounted on a plastic package. (Only the half plane is shown and the other half follows by symmetry. The mesh is provided courtesy of Drs. George Raiser and Nancy Fang, both at Intel.) When the material's properties (modulus, coefficient of expansion, etc.) are put in, the model will provide stress in various layers.

The global model in Figures 12a and 12b is then taken and put into a detailed die-level model. The die-level model is shown in Figure 13a.

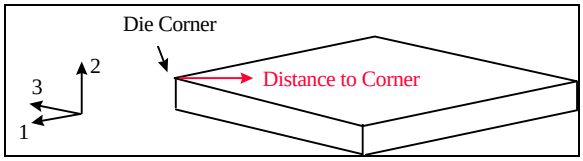


Figure 13a: Stress components

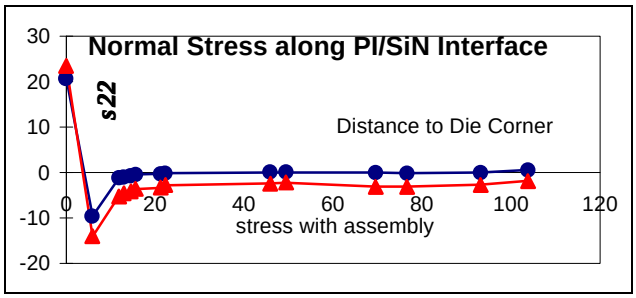


Figure 13b: Result of calculations showing that the stresses are altered after assembly

The results of this analysis (Figure 13b) help us both in selecting materials as well as in defining the layout design rules to mitigate failures. These models allow detailed analysis of corner areas that are subject to the most stress.

We have also studied the effects of stress on transistors, and our results show that stress does not play any part in the degradation of transistor stability.

Discussion—Limits to Energy Scaling

From Table 2 it may occur to the discerning reader that a reduction in the total power consumed by microprocessors will lead to an enhancement in the lifetime of a device. In fact, Von Neuman stated that the process of manipulating 0s and 1s could be accomplished without the expenditure of entropy and hence energy. We refer to this as the vN computer. Using this as an absolute reference you may ask what "practical" power dissipation is possible. Both Keyes [13] and Meindl [12] have shown that the "ideal" switching process in an ersatz but "practical" quantum mechanical computer could switch with power as low as 10e-41 Joules per switching even. Real computers take much more energy—about 10 e-11 Joules. It can be seen from this that present day computers expend vastly more energy per switching event than the ersatz computer (in fact by a factor of e+31 in the example above).

It can therefore be argued that from a reliability scaling point of view, enhanced reliability could be achieved if the energy per switching event could be reduced. Simple scaling of voltage could continue to about ~10kT, but we may approach other materials' limits before ever reaching energy limits, Meindl [12]. However, significant effort has to be made to reduce the power consumption of future microprocessors, and this effort will also contribute to their extended reliability.

Before reaching the lowest possible operating voltage, it is likely that other limits like materials' limited RC delays will set in. This is likely to call for new materials with lower RC constants (such as, copper with low-k dielectrics), and these new materials will undoubtedly bring fresh reliability challenges. One may therefore expect to see both new combinations of materials and new reliability phenomena in the coming generations.

Conclusion

In this paper, we show that for Intel's 0.25µm process technology based products, electrostatic discharge protection of gates, electromigration in metal lines, gate oxide reliability, and mechanical reliability have been modeled, measured, studied, and characterized; and that our design methodology ensures that the quality of our products is equal to that of previous generations.

We note that the channel length, gate thickness, and voltage undergo a scaling process with operating voltage and are internally consistent with a "constant E-filed" scaling scheme. However, power, current, and size of the integrated and multi-functional microprocessors—and the stress effects on them when mounted in complex packages—are not scaling in a systematic manner. We believe that both these tendencies will constitute the challenges of the future.

Acknowledgments

We acknowledge useful technical discussions with Neal Mielke and Paco Leon and leaders of the Protection, EM, and Stress working groups. We also thank our immediate Intel management for support and encouragement in writing this paper.

References

1. *Component Quality and Reliability*, Intel Technical Publication, Literature Center, POB 7641, Mt. Prospect Il 60056-7641.
2. G. Notermans, "On the Use of N-Well Resistors for Uniform Triggering of ESD Protection Elements," *1997 EOS/ESD Symposium Proceedings*, pp. 221-229.
3. T.J. Maloney and S. Dabral, "Novel Clamp Circuits for IC Power Supply Protection," *1995 EOS/ESD Symposium Proceedings*, pp. 1-12. Revised version published in *IEEE Trans. on Components, Packaging, and Manufacturing Technology, Part C, 19*, pp. 150-161, July 1996.
4. R. Merrill and E. Issaq, "ESD Design Methodology," *1993 EOS/ESD Symposium Proceedings*, pp. 223-237.
5. E.R. Worley, et. al., "Sub-micron Chip ESD Protection Schemes Which Avoid Avalanching Junctions," *1995 EOS/ESD Symposium Proceedings*, pp. 13-20.
6. T.J. Maloney and T.M. Eiles, "MOSFET-Based Power Supply Clamps for Electrostatic Discharge

Protection of Integrated Circuits," US Patent application, filed 3/25/97.

7. T. Toyabe, et. al., "A Numerical Model of Avalanche Breakdown in MOSFETs," *IEEE Trans. Electron Devices, ED-25*, 825-832 (1978).
8. M. Bohr, et. al., "A High Performance 0.35µm Logic Technology for 3.3V and 2.5V Operation," *1994 Proceedings of the IEEE International Electron Devices Meeting*, pp. 273-276.
9. M. Bohr, et. al., "A High Performance 0.25µm Logic Technology Optimized for 1.8V Operation," *1996 Proceedings of the IEEE International Electron Devices Meeting*, pp. 847-850.
10. R. Filippi et al., *Journal Of Applied Physics*, 1995, pp. 3756 – 3768.
11. YH Lee, et al., *P2ID Technical Digest*, 1998, pp. 38 – 41.
12. J. D. Meindl "Low Power Microelectronics: Retrospect and Prospect" *Proceedings of IEEE v.83* (4), pp. 619-635, 1995.
13. R.W. Keyes "Physical Limits in Digital Electronics," *Proceedings IEEE v.63*, pp. 740-766. May 1975.

Authors' Biographies

Krishna Seshan received a M.Sc in Low Temperature Physics from the University of Lancaster, UK, and a Ph.D. in Materials Science EE from the University of California at Berkeley in 1975. He is involved in aspects of mechanical stress management and the interaction of stress with all device levels. He works as a technical staff member in Intel's 0.25 µm Process Integration team. His email address is krishna.seshan@intel.com.

Timothy J. Maloney received degrees in physics from the Massachusetts Institute of Technology and Cornell University and ended with a Ph.D. (1976) in electrical engineering and postdoctoral studies at Cornell University. He was employed in semiconductor research at Varian Associates, Palo Alto, CA, from 1977 until he joined Intel in 1984. Since then, he has been concerned with integrated circuit ESD protection, CMOS latchup testing, fab process reliability, and design and testing of standard IC layouts. In 1994, he received the Intel Achievement Award for his patented ESD protection devices, which have achieved breakthrough ESD performance enhancements for a wide variety of Intel products. In 1996, he became a Principal Engineer at Intel. He is a Senior Member of the IEEE. His email address is timothy.j.maloney@intel.com.

Kenneth J. Wu received his B.S. (E.E.) from the National Taiwan University in 1975, his M.S. (E.E.) from Northwestern University in 1978, and his Ph.D. (E.E.) from Princeton University in 1982. He joined Intel Corporation in 1982 working on process technology development in SRAM, Microprocessor, and Non-Volatile Memory technologies. He is interested in the areas of dielectric, gate charging, hot carrier, charge retention, electromigration, and assembly/package related reliability issues. Currently, he is Intel's 0.25µm Microprocessor Technology Reliability Program Manager. His email address is kenneth.j.wu@intel.com.

MOS Scaling: Transistor Challenges for the 21st Century

Scott Thompson, Portland Technology Development, Intel Corp.
Paul Packan, Technology Computer Aided Design, Intel Corp.
Mark Bohr, Portland Technology Development, Intel Corp.

Index words: SDE, transistor, scaling

Abstract

Conventional scaling of gate oxide thickness, source/drain extension (SDE), junction depths, and gate lengths have enabled MOS gate dimensions to be reduced from 10μm in the 1970's to a present day size of 0.1μm. To enable transistor scaling into the 21st century, new solutions such as high dielectric constant materials for gate insulation and shallow, ultra low resistivity junctions need to be developed. In this paper, for the first time, key scaling limits are quantified for MOS transistors (see Table 1). We show that traditional SiO_2 gate dielectrics will reach fundamental leakage limits, due to tunneling, for an effective electrical thickness below 2.3 nm. Experimental data and simulations are used to show that although conventional scaling of junction depths is still possible, increased resistance for junction depths below 30 nm results in performance degradation. Because of these limits, it will not be possible to further improve short channel effects. This will result in either unacceptable off-state leakage currents or strongly degraded device performance for gate lengths below 0.10μm. MOS transistor limits will be reached for 0.13μm process technologies in production during 2002. Because of these problems, new solutions will need to be developed for continued transistor scaling. We discuss some of the proposed solutions including high dielectric constant gate materials and alternate device architectures.

FEATURE	LIMIT	REASON
Oxide Thickness	2.3 nm	Leakage (I_{GATE})
Junction Depth	30 nm	Resistance (R_{SDE})
Channel Doping	V_T=0.25 V	Leakage (I_{OFF})
SDE Under Diffusion	15 nm	Resistance (R_{INV})
Channel Length	0.06μm	Leakage (I_{OFF})
Gate Length	0.10μm	Leakage (I_{OFF})

Table 1: Fundamental scaling limits for conventional MOS devices

Introduction

For more than 30 years, MOS device technologies have been improving at a dramatic rate [1,2]. A large part of the success of the MOS transistor is due to the fact that it can be scaled to increasingly smaller dimensions, which results in higher performance. The ability to improve performance consistently while decreasing power consumption has made CMOS architecture the dominant technology for integrated circuits. The scaling of the CMOS transistor has been the primary factor driving improvements in microprocessor performance. Transistor delay times have decreased by more than 30% per technology generation resulting in a doubling of microprocessor performance every two years. In order to maintain this rapid rate of improvement, aggressive engineering of the source/drain and well regions is required. In this paper, key methods for improving device performance are discussed. Creating shallow source/drain extension (SDE) profiles for improved short channel effects, the use of retrograde and halo well profiles to improve leakage characteristics, and the effect of scaling the gate oxide thickness are discussed in detail. Fundamental tradeoffs and scaling trends in engineering these effects are analyzed through experimental data and computer simulations. The impact of these trends associated with circuit requirements including power supply, threshold voltage, and off-state leakage on transistor design is also explored. We show that the scaling trends of the last ten years will be extremely difficult if not impossible to maintain unless new methods for device improvement are found. In addition to the conventional MOS transistor, several alternate device architectures are analyzed to understand the potential gains and tradeoffs associated with each device. The ability to overcome current physical technology limits such as gate oxide thickness and shallow junction formation as well as tradeoffs in circuit design will

determine if MOS transistors can be scaled into the next century.

Oxide Scaling

Gate oxide thickness scaling has been instrumental in controlling short channel effects as MOS gate dimensions have been reduced from 10µm to 0.1µm. Gate oxide thickness must be approximately linearly scaled with channel length to maintain the same amount of gate control over the channel to ensure good short channel behavior. Figure 1 plots the electrical channel length divided by gate oxide thickness for Intel's process technologies over the past 20 years. Each data point represents a process technology, developed approximately every three years, which was used to fabricate Intel's leading-edge microprocessors.

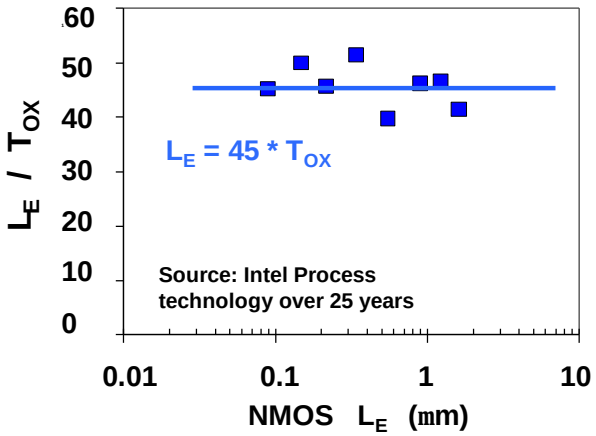


Figure 1: Channel length divided by gate oxide thickness versus channel length

From Figure 1, a simple relationship between oxide thickness and the minimum channel length set by short channel effects is observed:

$$L_E = 45 * T_{OX} \qquad \text{(Eq. 1)}$$

This relationship exists because the channel depletion layer is engineered to become smaller as the gate oxide thickness is decreased. In addition, short channel behavior is governed by the ratio of channel depletion layer thickness to channel length. The channel depletion layer is inversely proportional to the square root of the channel doping concentration. During device optimization, channel doping is increased as the oxide is scaled to maintain approximately the same device threshold voltage. Figure 2 illustrates this point. In Figure 2, the thickness of the channel depletion layer for two devices with different oxide thicknesses is shown. Figure 2a shows the depletion layer for a device with an oxide thickness of 4.5 nm while Figure 2b shows a device with an oxide thickness of 3.2 nm.

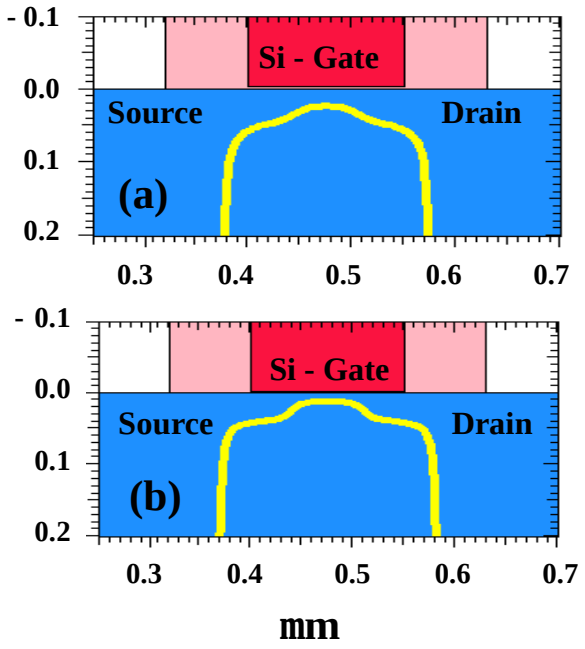


Figure 2a and 2b: Device simulations showing channel depletion layer thickness for devices with two oxide thicknesses: (a) 4.5 nm, (b) 3.2 nm

Both devices have the same off-state leakage. The device with the thinner oxide has a smaller channel depletion layer and hence improved short channel characteristics. The improved short channel effects can be taken advantage of by targeting a smaller channel length. Thus, for continued MOS channel length scaling, the gate dielectric thickness must continue to be scaled. Figure 3 shows the Semiconductor Industry Association's (SIA) road map for gate dielectric thickness. This roadmap predicts that continued gate dielectric scaling will be required with a new gate dielectric material needed for the 2002-2005 time frame.

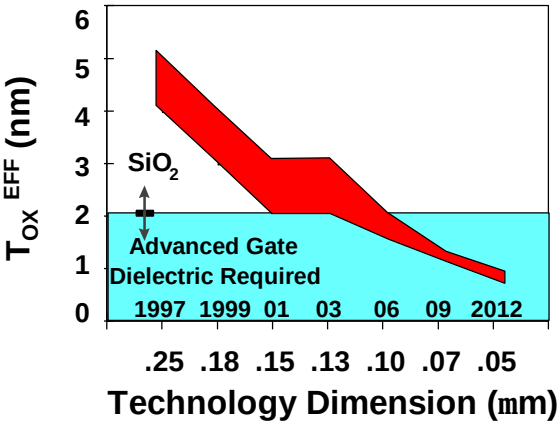


Figure 3: SIA road map for junction depth

Scaling Limit for SiO_2

SiO_2 or nitrided SiO_2 has been the gate dielectric used by the semiconductor industry for over 30 years. The thickness limit is the same for both materials and is not limited by manufacturing control. Today, it is technically feasible to manufacture 1.5 nm and thinner oxides on 200 mm wafers [3]. The thickness limit for SiO_2 is set instead by gate-to-channel tunneling leakage. Figure 4 schematically shows the tunneling leakage process for an NMOS device biased in inversion.

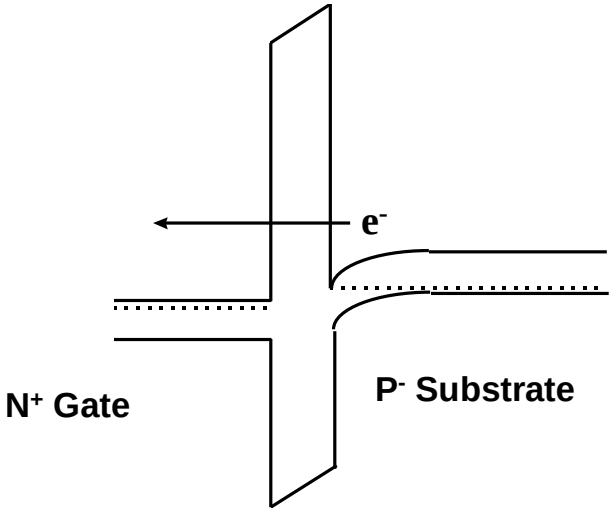


Figure 4: Direct tunneling leakage mechanism for thin SiO_2

As the thickness of the dielectric material decreases, direct tunneling of carriers through the potential barrier can occur. Because of the differences in height of barriers for electrons and holes, and because holes have a much lower tunneling probability in oxide than electrons, the tunneling leakage limit will be reached earlier for NMOS than PMOS devices. The SiO_2 thickness limit will be reached approximately when the gate to channel tunneling current becomes equal to the off-state source to drain sub-threshold leakage (currently ~1nA/µm). Figure 5 shows the area component of gate leakage current in A/cm^2 versus gate voltage. If we assume the gate leakage limit occurs for devices with 0.1µm gate length designed for 1.0V operation, the SiO_2 thickness limit occurs at ~1.6 nm.

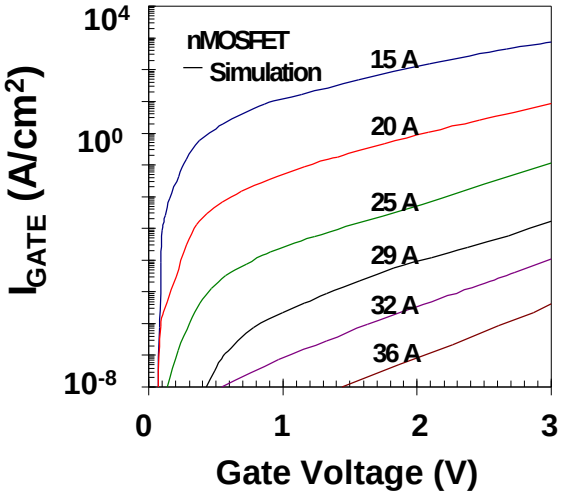


Figure 5: Gate leakage versus gate voltage for various oxide thicknesses [5]

We now have established that the thickness limit for SiO_2 is ~1.6 nm. However, due to quantum mechanical and poly-Si gate depletion effects, both the gate charge and inversion layer charge will be located at a finite distance from the SiO_2/Si interfaces with the charge location being a strong function of the bias applied to the gate. Figure 6 shows the location of the inversion layer charge in the silicon substrate for a transistor with a typical bias when quantum mechanical effects are taken into account [4]. The centroid for the inversion charge is ~1.0 nm from the SiO_2/Si interface. This increases the effective SiO_2 thickness (T_{OX}^{EFF}) by ~0.3 nm. By taking into account the charge distribution on both sides of the gate, the minimum effective oxide thickness for a MOS device bias in inversion (at voltages used in our 0.25 or 0.18µm technologies) is increased by approximately 0.7 nm. Thus, the 1.6 nm oxide tunneling limit results in an effective oxide thickness of approximately 2.3 nm.

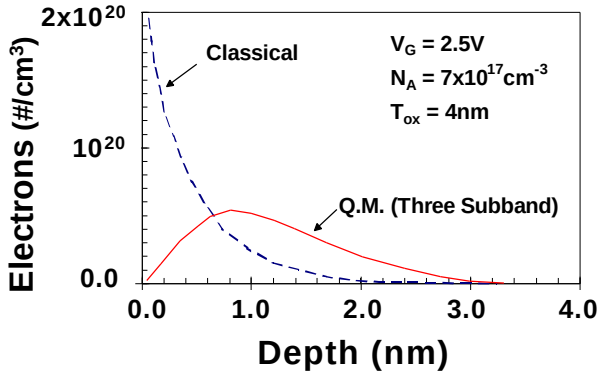


Figure 6: Position of inversion channel charge versus depth

Based on the previous arguments for controlling short channel effects, a limit for SiO_2 thickness will set a limit on the gate and channel length of MOS devices. Figure 7 plots gate and channel length versus effective oxide thickness. From this figure, we see that the limit for gate and channel length for an SiO_2 gate dielectric MOSFET is 0.1µm and 0.06µm, respectively. Since in leading-edge logic technologies, the gate dimension is printed smaller than the technology features, the SiO_2 thickness limit and the gate length limit will be reached for ~0.13µm technologies.

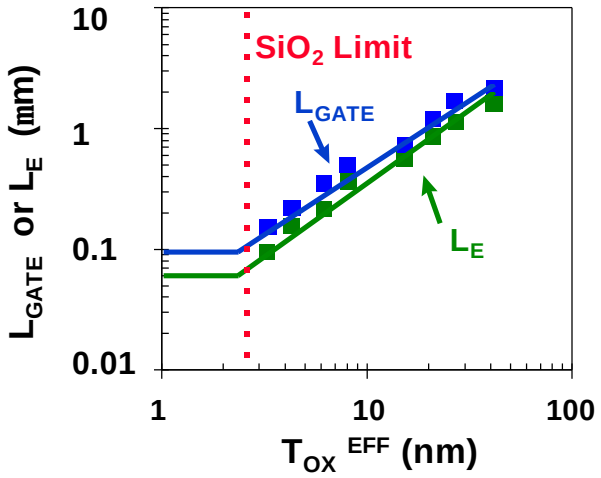


Figure 7: Gate and channel length versus effective oxide thickness

Alternative High Dielectric Constant Materials

Alternative high dielectric constant materials will be the key to continued MOSFET scaling past 0.1µm gate dimensions. With these materials, thicker dielectric layers can be used yet the same inversion layer characteristics can be maintained. These thicker layers result in less carrier tunneling, and they permit further scaling of the effective oxide thickness. Table 2 lists the leading alternative dielectrics and their status.

OPTION	ISSUES / STATUS
Si_3N_4 /	Small advantage especially with buffer layer
nitride	Close to being ready (G. Lucovsky, T. P. Ma)
Ta_2O_5	Need SiO_2 buffer/ no poly-silicon gate
	Very early stages (S. Kamiyama)
TiO_2	Need SiO_2 buffer/ no poly-silicon gate
	Very early stages (S. A. Campbell)
BST	Deep states/ buffer layer/ no poly-silicon gate
	Early stages FET (large DRAM interest)

Table 2: Alternate high dielectric constant materials [6-9]

All these materials, with the possible exception of Si_3N_4, need an SiO_2 buffer layer between the high dielectric constant materials and the silicon substrate in order to obtain an interface with low interface states. They also need a metal electrode to eliminate a reaction between the alternate dielectric and the poly-Si that usually forms SiO_2. This is extremely unfortunate since it can be shown that if an SiO_2 buffer layer is needed, and since quantum mechanical effects and poly-Si gate depletion cannot be eliminated, an Si_3N_4 gate dielectric with a buffer layer can only improve the effective oxide thickness by 0.3 nm before it reaches its tunneling thickness limit [10]. The problem with using a metal gate electrode with an alternative dielectric material is that the metal gate is not compatible with deep sub-micron complementary CMOS devices. A metal gate with a work function equal to intrinsic silicon such as tungsten would produce complementary CMOS devices. However, a mid-bandgap gate metal is not compatible with deep sub-micron devices because of degraded short channel behavior. Figure 8 shows the depletion layer obtained from a device simulator for two NMOS devices with the same threshold voltage but with different gate electrodes: (a) with an N+ poly-Si gate and (b) with a tungsten gate. As can be seen from this figure, the device with the tungsten gate has a significantly larger depletion layer and hence degraded short channel behavior.

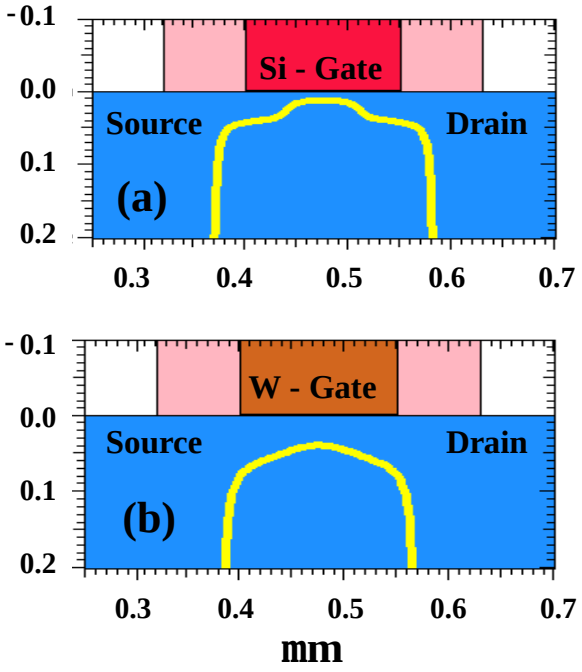


Figure 8: Device simulation of two devices showing depletion layers: a) N+ poly-Si and b) tungsten gate

Source/Drain Engineering

In this section, we investigate the scaling of source/drain extension (SDE) depth and gate overlap for MOSFETs of 0.1µm and below. For the purposes of this discussion, the SDE is the shallow diffusion that connects the channel with the deep source and drain. Junction depth always refers to the SDE junction depth. The deep source/drain junction depth is held constant. Overlap is defined as the distance the SDE extends under the gate. The metallurgical spacing (L_{MET}) is the distance between the source and drain SDE (see Figure 9).

We show that a minimum SDE to gate overlap of 15-20 nm is needed to prevent degradation of drive current (I_{DSAT}). We also show that scaling SDE vertical depths below 30-40 nm results in little to no performance benefit for 0.1µm devices and beyond. This is because any improvement in short channel effects due to reduced charge sharing is offset by a large increase in external resistance and too small an overlap between the SDE and gate.

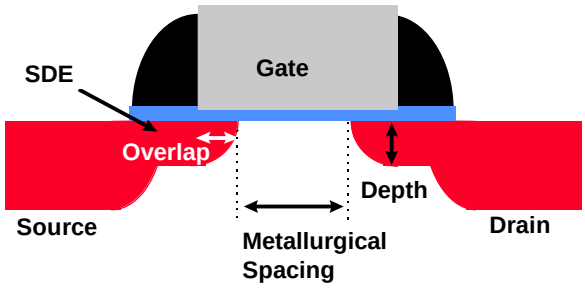


Figure 9: Terminology used in this discussion

Shallow Junction Formation

Very short gate length transistors with shallow SDE junctions and small gate overlap have been reported [11,12]. Many of these transistors have lower than expected drive currents given their extremely short channel lengths. We propose that these low drive currents are the result of an SDE that is too shallow and therefore leads to a high external resistance and too small of an overlap between the SDE and gate. Junction depths are currently 50-100 nm for 0.25µm process technologies and are predicted to be as low as 10 nm for future deep sub-micron devices (see Figure 10). The fabrication of these shallow junctions is less of an issue than whether or not the shallow junctions offer any device benefit. Shallow junctions can be fabricated by carefully controlling transient enhance diffusion (TED) [13-17]. Methods for reducing TED include lowering implant energies, amorphization followed by solid phase expitaxial regrowth and high temperature, and short time rapid thermal anneal cycles. Figure 11 shows an example of a shallow 35 nm junction formed by a low energy implant and a rapid thermal anneal. Alternate architectures such as removable spacer process flows can also be used to minimize SDE depths. In this architecture, an initial disposable spacer is used. High temperature cycles for forming the S/D and doping the poly-Si gate are used before the introduction of the SDE structure. These cycles permit the use of extremely low temperature anneal cycles engineered to minimize SDE junction depths and maximize dopant concentrations.

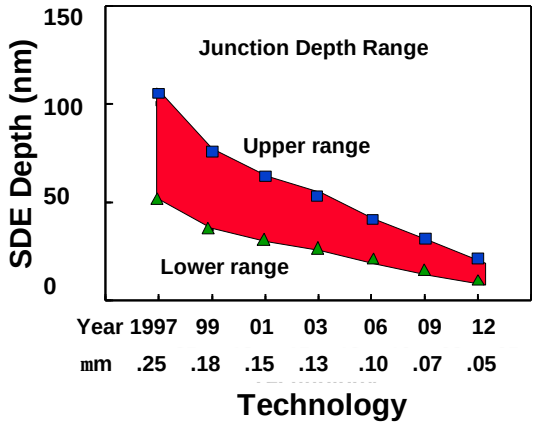


Figure 10: SIA roadmap for junction depth

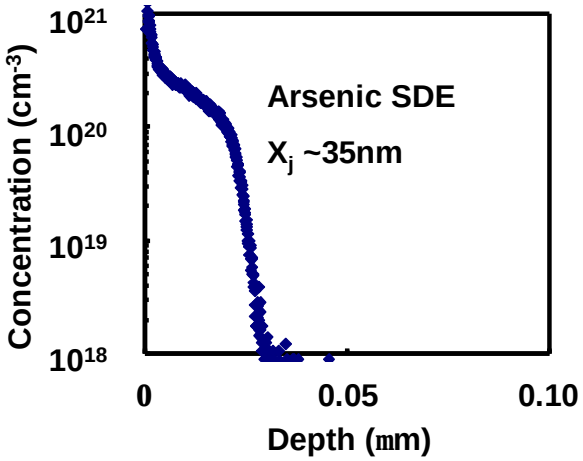


Figure 11: Shallow 30.0 nm SDE formed by a low energy implant and rapid thermal anneal

SDE Junction Scaling

Reducing SDE junction depths will improve device short channel characteristics by reducing the amount of channel charge controlled by the drain. This may not,

however, lead to improved device performance. Figures 12a and 12b show the potential contours for two devices with junction depths of 30 and 150 nm, respectively, biased in the off-state condition. In this figure, the potential contours extend much further into the channel for the device with the deep junction.

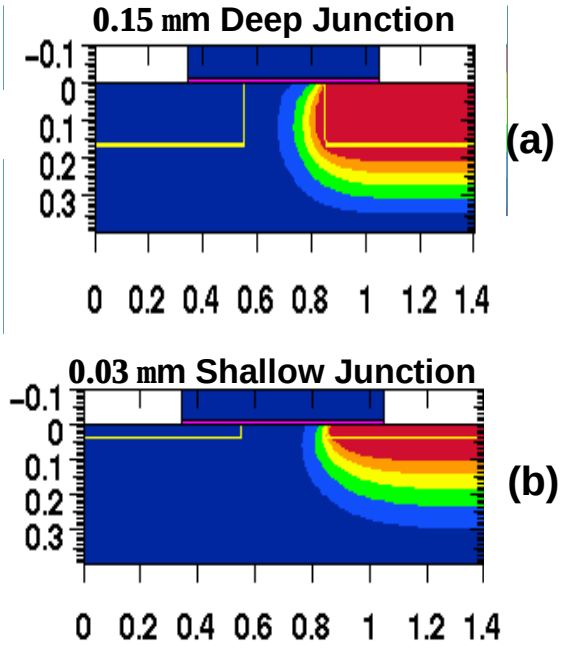


Figure 12: Potential contours for two devices biased in an off-state condition (a) 30 nm shallow junction and (b) 150 nm deep junction

Thus, transistors with deeper junctions will have worse short channel characteristics. Unfortunately, shallow SDE junctions can increase the external resistance of the device. Figure 13 shows the various components of external resistance for a MOS device. Current flows from the channel inversion layer into the SDE accumulation region ($R_{ACCUMULATION}$). The current then spreads out into the SDE ($R_{SPREADING}$) region and through the bulk SDE area (R_{SHUNT}). The final component of resistance is associated with the deep source/drain and salicide ($R_{CONTACT}$). In deep sub-micron devices, particularly NMOS, the SDE accumulation and spreading components are the dominant components of external resistance. The components associated with the SDE region become a greater problem as the transistor feature size is scaled (channel length and SDE depth reduced) since the scaling reduces channel resistance while increasing the components of SDE resistance.

A second scaling limit is the minimum SDE-to-gate overlap for a device. Reducing this overlap causes the current to spread out into a lower doping location of the SDE. This can strongly increase accumulation and spreading resistance and increase the total external resistance. For example, if the overlap is zero, the current flow would spread out at the gate edge where the SDE doping concentration would be zero. In the next section, we investigate scaling limits for SDE to junction depth and gate overlap.

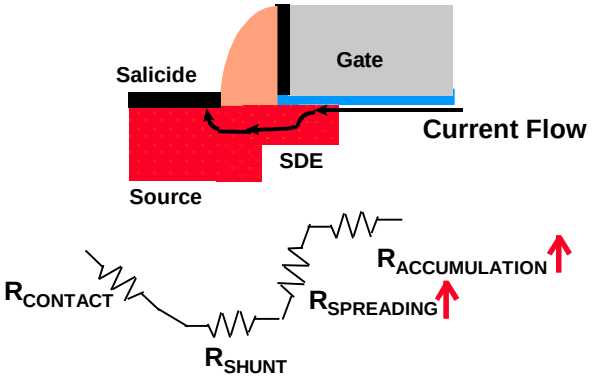


Figure 13: Components of external resistance

Minimum SDE-to-Gate Overlap

The test structure shown in Figure 14 is used to evaluate the effect of SDE-to-gate overlap on I_{DSAT}. In this test structure, the SDE implant is performed after the formation of a thin offset spacer. By varying the thickness of the offset spacer, the SDE-to-gate overlap and vertical junction depth can be independently varied. The transistor data presented are measured on devices with a process flow similar to our 0.25µm technology [2].

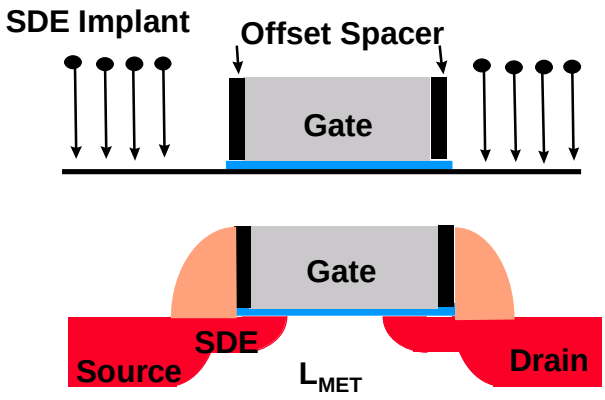


Figure 14: Test structure to evaluate minimum SDE-to-gate overlap

Also included is data on transistors with gate length, gate oxide, and power supply scaled by 0.7 and $(0.7)^2$ from our 0.25µm technology. All transistors have controlled sub-threshold slopes of less than 85mV/decade, 1nA/µm off-state leakage, and electrical channel lengths (L_E) between 0.06 and 0.14µm.

With the above test structure fabricated for a range of poly-Si gate lengths, the transistor saturation drive current versus the SDE overlap for both fixed vertical SDE depth and fixed SDE metallurgical spacing was measured. The SDE metallurgical spacing is kept constant by adjusting the poly-Si gate length to maintain 1nA/µm off-state leakage. Figure 15 shows the vertical SIMS profile of an SDE junction used in the experiment (1.0e15cm^{-2}, 5keV arsenic implant RTA annealed). Figure 16 shows the effect of spacer offset on overlap capacitance and I_{DSAT}. For spacer offsets greater than 40 nm, there is a flattening in overlap capacitance implying minimal SDE-to-gate overlap. A degradation in I_{DSAT} is also clearly observed for offset spacer widths greater than 20 nm.

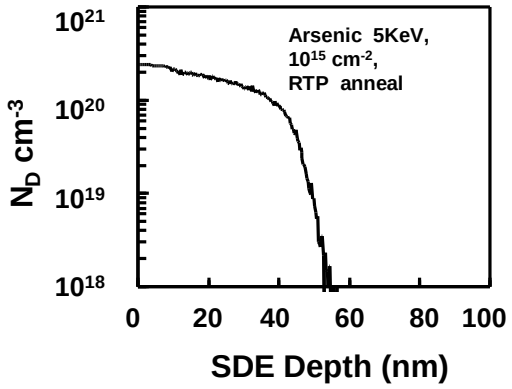


Figure 15: Vertical SIMS profile of Arsenic SDE

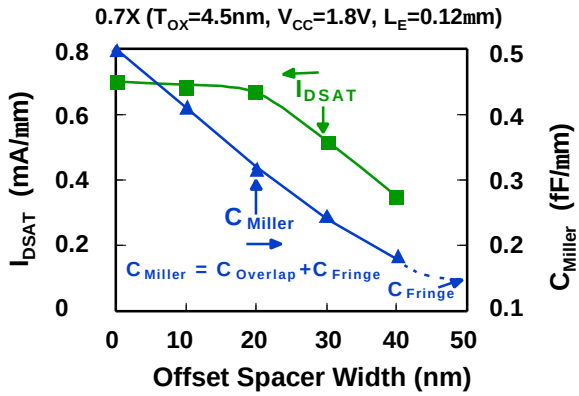


Figure 16: I_{DSAT} and C_{MILLER} versus spacer offset

The lateral diffusion of the SDE junction under the gate edge is estimated to be 0.6 - 0.7 times the vertical depth minus the offset spacer width. This estimate is obtained from process simulations and junction-staining measurements. Experimentally, the offset spacer width is varied from 0 to 40 nm and is used to modulate the SDE-to-gate overlap from approximately 40 to 0 nm. Figures 17 and 18 show I_{DSAT} versus SDE overlap for both NMOS and PMOS 0.25µm devices as well as the 0.7 scaled devices. These figures also show that, independent of the feature size of the process technology, a degradation in I_{DSAT} is observed if the overlap is less than 15-20 nm.

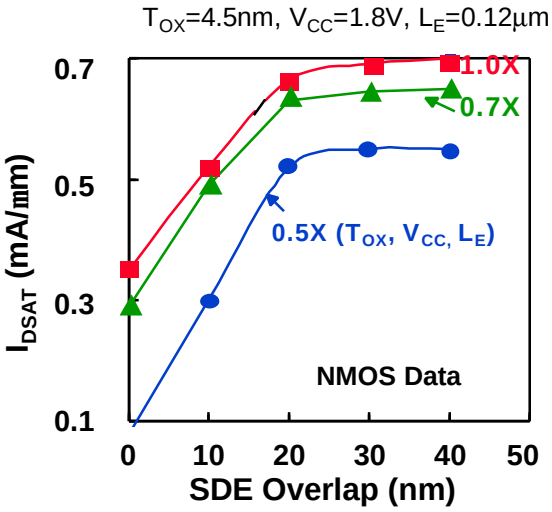


Figure 17: I_{DSAT} versus SDE overlap (NMOS)

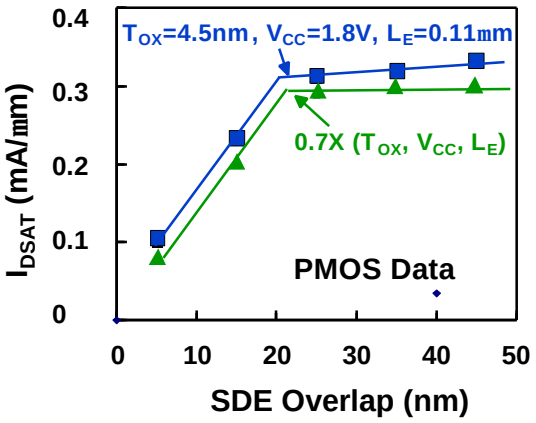


Figure 18: I_{DSAT} versus SDE overlap (PMOS)

Minimum SDE Junction Depth

The optimal SDE vertical depth is now investigated. For this set of experiments, both the conventional and removable spacer flows were used. Figure 19 shows NMOS and PMOS drive current versus SDE depth for devices with 1nA/µm of off-state leakage. The SDE depths were adjusted by varying the implant energy (500eV - 40KeV) and the RTA temperature. In Figure 19, we see that a maximum in I_{DSAT} occurs when the vertical junction depth is 35-40 nm. With an SDE deeper than 35-40 nm, short channel effects degrade due to increased charge sharing. This necessitates a larger channel length to meet the off-state criteria and a loss in I_{DSAT}. SDE depths shallower than 35-40 nm result in degraded I_{DSAT} due to increased external resistance and an overlap between the SDE and gate that is too small.

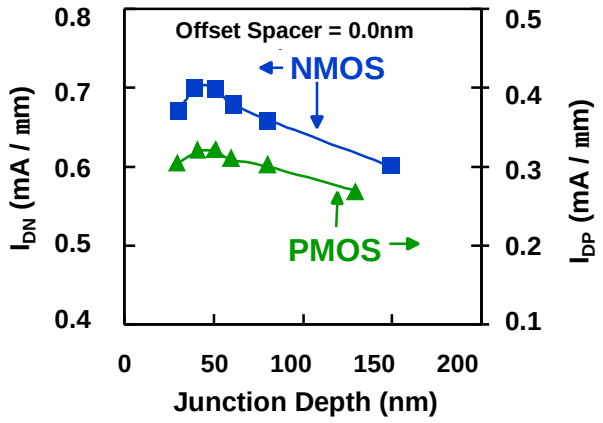


Figure 19: I_{DSAT} versus SDE depth

Simulation results for the above experiment are shown in Figure 20. In this figure, external resistance and short channel behavior (defined by source-to-drain distance at 1nA/µm off-state leakage) versus SDE junction depth are quantified.

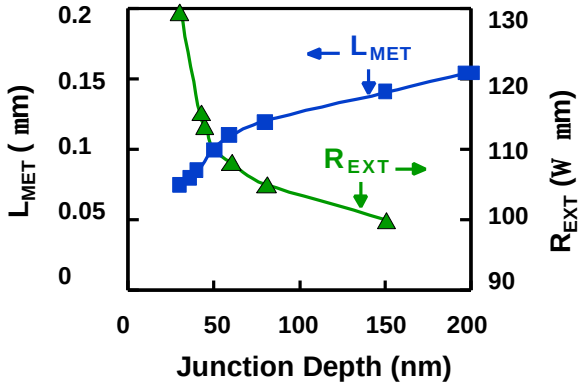


Figure 20: Simulation data quantifying R_{EXT} and L_{MET} versus junction depth

These results support the conclusion that the observed drive current maximum at a 35-40 nm junction depth results from tradeoffs in short channel effects, external resistance, and SDE-to-gate coupling. Note that these conclusions implicitly assume that the maximum SDE concentration is solid solubility limited for these devices.

Channel Engineering

Up to now we have shown how gate oxide thickness and junction scaling has enabled channel length scaling by improving short channel characteristics. We have also quantified scaling limits for these two techniques. The third and final technique to improve short channel characteristics is well engineering. By changing the doping profile in the channel region, the distribution of the electric field and potential contours can be changed. The goal is to optimize the channel profile to minimize the off-state leakage while maximizing the linear and saturated drive currents. Super Steep Retrograde Wells (SSRW) and halo implants have been used as a means to scale the channel length and increase the transistor drive current without causing an increase in the off-state leakage current [18-23]. Figure 21 is a schematic representation of the transistor regions that are affected by the different types of well engineering. Retrograde well engineering changes the 1D characteristics of the well profile by creating a retrograde profile toward the Si/SiO_2 surface. The halo architecture creates a localized 2D dopant distribution near the S/D extension regions. The use of these two techniques to increase device performance is discussed in the following sections. We show that channel doping optimization can improve

circuit gate delay by ~10% for a given technology. However, we also show that well doping engineering cannot provide the generation after generation channel length scaling that gate oxide and SDE junction depth scaling has provided.

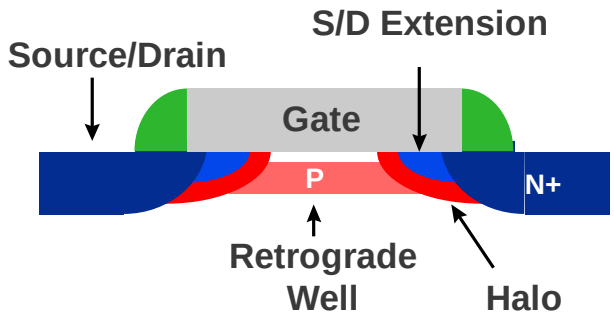


Figure 21: Schematic representation of different aspects of well engineering

Retrograde Well Engineering

The use of retrograde well profiles to improve device performance has been reported [18,21]. The retrograde profile is typically created by using a slow diffusing dopant species such as arsenic or antimony for PMOS devices and indium for NMOS devices. It has been established that SSRW can improve short channel effects, increase surface mobility, and can lead to either an increase or a decrease in saturated drive current depending on a variety of technology issues [18-20]. Although retrograde wells do not appreciably improve saturated drive currents, we will show that for today's deep sub-micron technologies, they do improve linear drive currents and lead to improved circuit performance. Unfortunately, as S/D junction depths continue to decrease, this gain in linear drive current is further diminished.

The process flow used for the devices in this study has been reported [1]. In this study, aggressive SSRW wells created by indium (NMOS) and arsenic (PMOS) implants are compared to uniform wells formed by boron (NMOS) and phosphorus (PMOS). Figure 22 shows the vertical doping profile for an SSRW formed by an arsenic implant and by a conventional flat phosphorus well. As can be seen, the well doping profile formed by the arsenic implant is clearly retrograde to the surface. Although the SSRW profile has a lower surface concentration, the profile was engineered to give the same threshold voltage as the flat well case to ensure an accurate comparison.

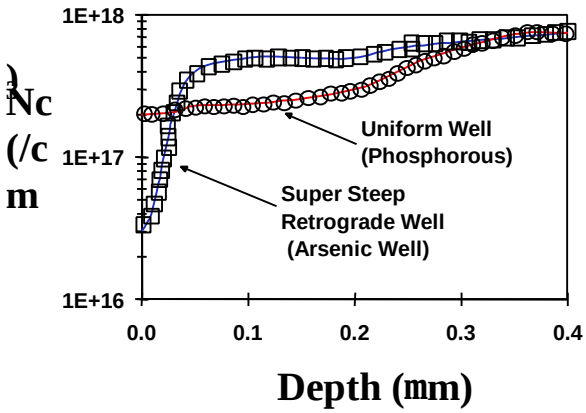


Figure 22: Vertical concentration doping profile for SSRW and conventional well doping profiles

Figure 23 shows the minimum channel length that can be supported for an off-state leakage current of 1nA/µm for a range of threshold voltages for both SSRW and uniform well transistors. As expected, higher threshold voltages support smaller gate lengths due to the increase in channel doping. This figure shows that the SSRW architecture supports smaller channel lengths compared to the uniform well case for all threshold voltages. Similar results are seen for antimony (PMOS) and indium (NMOS). For the purposes of this paper, only PMOS data will be shown. Figures 24 and 25 compare I_{OFF} and I_{DSAT} versus electrical channel length for SSRW and uniform well transistors. Figure 24 shows improved source-to-drain leakage for the SSRW device for sub-0.25µm channel lengths implying improved short channel effects. However, Figure 25 shows a decrease in saturated drive current for the same SSRW device. Figure 26 shows families of curves for drain current versus drain voltage for SSRW and uniform well devices. The devices have a channel length of 0.15µm. For devices with the same channel length, the linear drive current is approximately equal, indicating no change in mobility for SSRWs. However, the current does saturate at a lower drain bias.

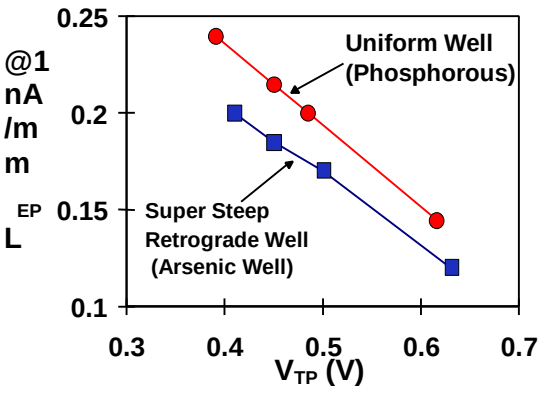


Figure 23: Channel length at which 1nA/µm of off-state leakage current occurs as a function of threshold voltage for SSRW and uniform well profiles

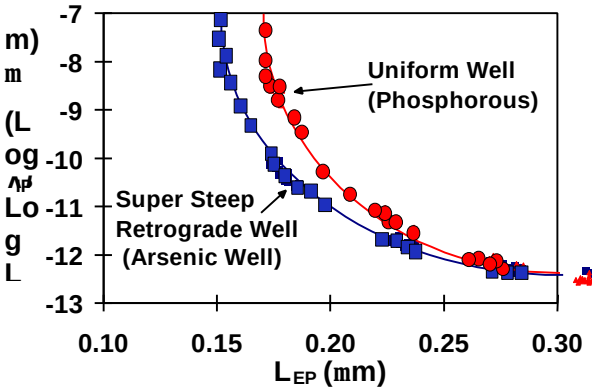


Figure 24: Leakage current as a function of channel length for SSRW and uniform well transistors with the same threshold voltage

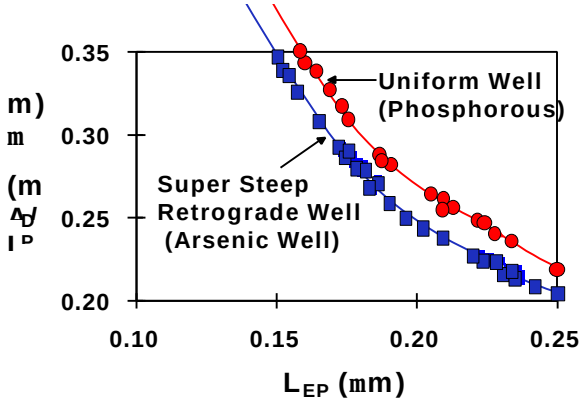


Figure 25: Saturated drive current (IDSAT) versus channel length for SSRW and uniform well transistors

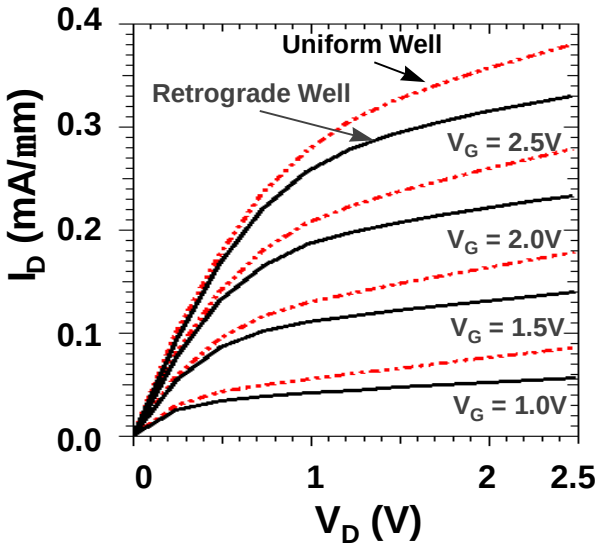


Figure 26: $I_D V_D$ characteristics for SSRW and uniform well devices as a function of gate voltage

In the next section, device simulations are used to understand this decrease in V_{DSAT}. Figure 27 shows the IV characteristics for SSRW and uniform well devices in which both devices have the same value of I_{OFF} (1nA/µm). Even though the SSRW device can support smaller channel lengths due to improved short channel effects, only a slight gain in I_{DSAT} is seen. The linear drive current, however, is clearly increased. For logic gate delays with fast input rise times and large loads, drive current in the linear mode is at least as important as drive current in saturation. Measured circuits showed that the increase in linear drive current improved inverter switching delays by up to 10%.

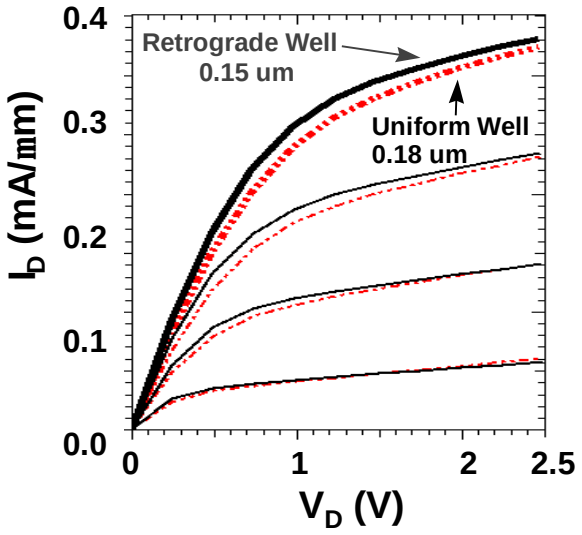


Figure 27: I_DV_D characteristics for SSWR and uniform well devices both having the same I_{OFF} criteria

Fundamental Operation of SSRW

In the classical derivation of the NMOS transistor, the drive current is calculated by integrating the inversion charge along the channel [24]:

$$I_D = \frac{W}{L}\int_{VS}^{VD} m_n\, Q_n(V)\, dV \qquad \text{Eq. 1}$$

It is typically assumed that the depletion charge and V_T are constant along the channel for this calculation. As shown schematically in Figure 28, the depletion charge and V_T actually increase along the channel from source to drain due to the body effect. This is true for both the SSRW and uniform well device. However, the increase in depletion charge and consequently V_T is larger for the SSRW device because of the higher doping in the substrate (see Figure 22) resulting in a larger body effect. The larger V_T for the SSRW device at high drain bias lowers the saturation voltage ($V_{DSAT}=V_G-V_{T(Drain)}$). This causes the reduction in I_{DSAT} for the SSRW device shown in Figure 26. The improvement in transistor performance due to SSRW strongly depends on the ability to scale the channel length due to improved short channel effects. Figure 29 shows the net change in performance due to SSRW versus junction depth. As S/D junction depths are scaled, the improvement in short channel effects from the use of SSRW decreases.

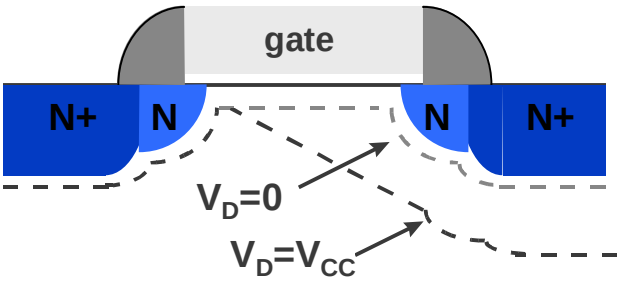


Figure 28: Schematic representation of the depletion layer for low and high drain bias

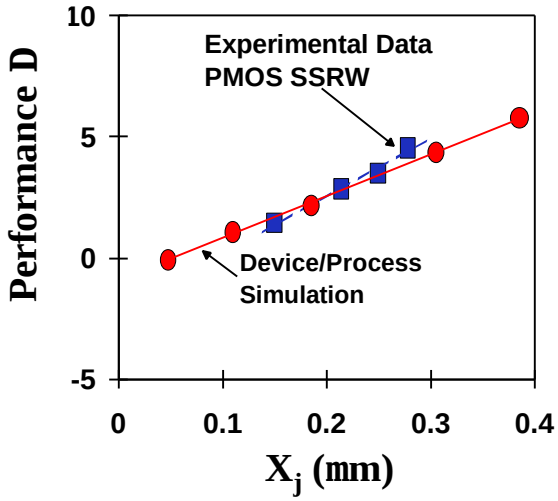


Figure 29: Improvement in device performance for SSRW over uniform well devices versus S/D depth

Halo Engineering

The addition of well implants to create a non-uniform well profile to improve short channel effects has been reported [25-27]. These implants may be vertical or angled and are typically done after gate patterning. They add additional well dopant around the source and drain regions providing an increased source-to-drain barrier for current flow. For long channel devices, the additional halo dopants only modestly change the threshold voltage. For short channel devices, however, a large increase in threshold voltage is seen. In order to maintain a constant threshold voltage for the target devices, the nominal threshold implant must be lowered for the halo devices (see Figure 30). This results in a lower long channel threshold voltage, and it can create a curvature reversal in the threshold voltage versus channel length curve. It will be shown in the following sections that although the use of halos can improve performance by compensating

for manufacturing variability, halos do not fundamentally improve device performance.

The process flow for the devices reported here has been presented previously [1,2]. Figure 30 shows a lateral surface cut of the doping profile for both a conventional and halo device. For the halo device, there is a lateral decay of the well doping profile toward the center of the channel. As the gate length of the halo device is decreased, the average well concentration increases resulting in a higher V_T.

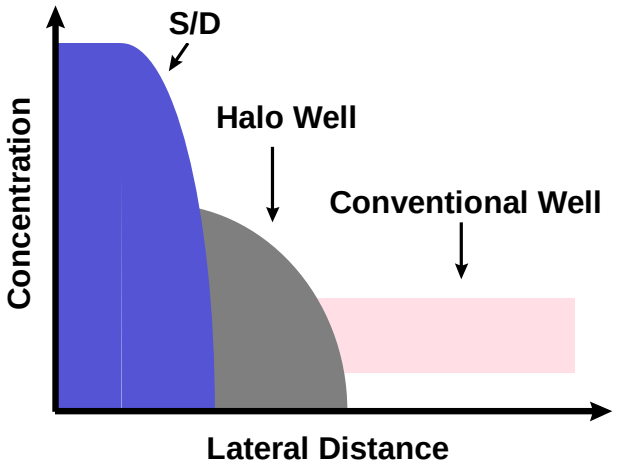


Figure 30: Schematic showing a lateral surface cut of the well doping near the Si/SiO_2 interface

Figures 31 and 32 show the threshold and off-state leakage characteristics versus channel length for conventional and halo devices. It should be noted that the change in well doping as a function of size makes extraction of effective channel length a strong function of extraction methodology for halo devices and often becomes much less meaningful. Because of this, it is often clearer to use I_{DSAT} versus I_{OFF} when comparing device performance for halo devices. Figure 33 shows I_{DSAT} versus I_{OFF} characteristics for a halo and non-halo device. As seen, there is very little improvement in I_{DSAT} at the targeted I_{OFF} for the halo device (Figure 33).

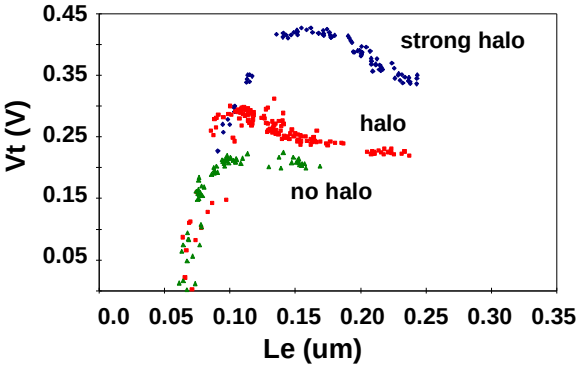


Figure 31: Threshold voltage as a function of channel length for a no halo, moderate halo, and strong halo device

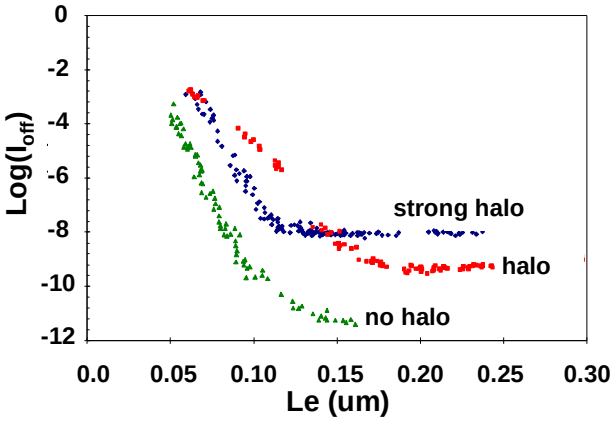


Figure 32: Off-state leakage current as a function of channel length for a no halo, moderate halo, and strong halo device.

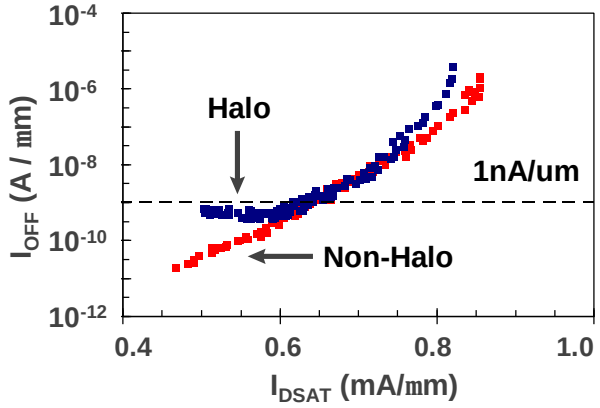


Figure 33: I_{OFF} versus I_{DSAT} for a halo and a conventional device (little to no gain in I_{DSAT} is seen for a given I_{OFF}.)

Fundamental Operation of Halo Well Profiles

Halo profiles are created by implanting extra dopants into the wells immediately after tip implantation. The implant is typically performed at an angle and energy high enough to ensure the implant dose is outside the final SDE profile. After spacer processing and S/D anneal, the resulting profile diffuses due to TED effects, resulting in a relatively flat profile over the dimensions of current device sizes. Figure 34 shows experimental results for the as implanted and final doping profile for a typical boron halo implant. The data includes the effect from damage generated by the SDE and S/D implants. As can be seen, the profile is quite flat over the characteristic channel length dimensions for today's 0.25µm and 0.18µm technologies. However, even though the halo profile is relatively flat, it still causes an increase in well doping as the gate length is decreased. This is because the same halo implant dose is confined in a smaller area. For flat well devices, I_{OFF} quickly decreases as the channel length is increased. This is due to the exponential relationship between the current and the potential barrier in the sub-threshold region. For the halo cases, the leakage current does not decrease as quickly with size. In fact, for extremely strong halos, an increase in I_{OFF} with increasing size can be seen. This can be explained by the change in the source-to-drain potential barrier for different size devices in the case of the halo well. For the strong halo devices, the threshold voltage is rapidly decreasing as the device size increases.

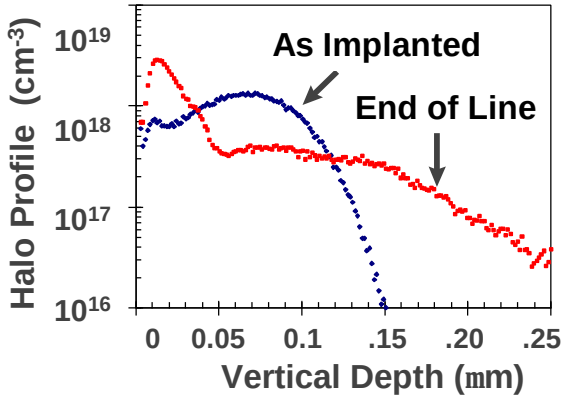


Figure 34: As implanted and end of line vertical halo profile (due to TED effects, a large amount of diffusion is seen)

This decrease compensates for the reduction in the electric field due to the increased channel length that results in less change in I_{OFF}. The strength of the halo depends not only on the halo doping concentration, but also on the lateral confinement of the halo. Figure 35 shows the simulation results on the effect of halo confinement for I_{OFF} versus device size. In this figure, I_{OFF} is plotted versus L_E for several values of **s** where **s** is defined as the characteristic lateral decay length of a gaussian halo doping profile, which begins at the transistor gate edge. Increasing the halo confinement increases the localization of the halo effect. A comparison of simulation and experimental results (Figures 32 and 35) shows that a relatively non-localized halo profile matches the experimental data. This is in agreement with the SIMS data of Figure 34. Therefore, for a single device size, both the halo and conventional device will have close to the same doping profile for the same off-state leakage criteria. However, there will be a large difference in the well doping level and threshold voltage for the device variations around this device. For the halo device, the threshold voltage will be lower for larger device sizes. Due to manufacturing variation, the target device will be necessarily larger than the worst-case device defined by maximum tolerable I_{OFF}. The gate drive (V_{CC}-V_T) for the target device is increased for the halo device resulting in an increase in I_{DSAT}. A halo can cause a greater than 10% increase in I_{DSAT} for the target device, relative to a non-halo process.

In order to scale deep sub-micron devices, halo implants must be used to improve the performance of target devices. Current technologies have used halo architectures to increase performance by up to 10%. Due to strong TED effects, halo profiles are not well confined in the technology now being used. A complicated interaction between halo dopant profiles, short channel effects, off-state leakage currents, and threshold voltages determines the final device performance gain.

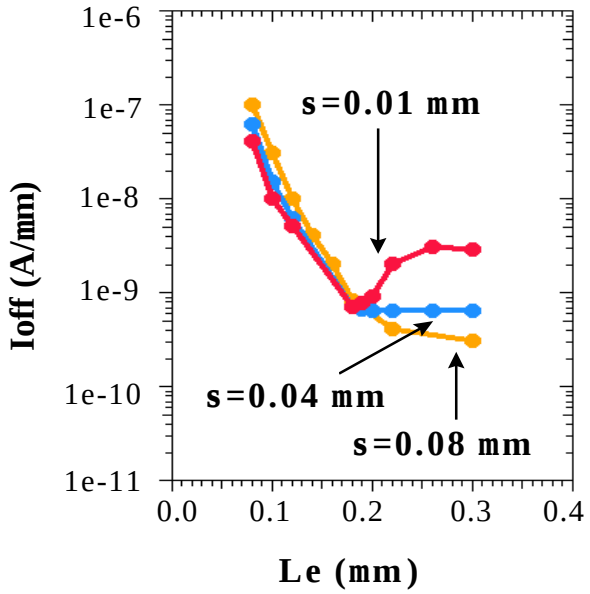


Figure 35: Simulation results showing the effect of halo confinement on I_{OFF} where σ is defined as the characteristic lateral decay length of a gaussian halo profile and is in units of µm.

The halo architecture does not improve device performance for the worst-case device, but instead provides a subtle benefit by improving the performance for the target devices. The smaller the difference between the worst case and target device (smaller device variability), the smaller the device improvement for halo well architecture.

Circuit and Device Interactions

The choice of power supply (V_{CC}) and threshold voltage (V_T) will be critical in determining whether the performance of 0.1µm transistors can continue to be scaled. These parameters strongly affect chip active power, chip standby power, and transistor performance.

In this section, we review the power supply and threshold voltage scaling trends. We show that the loss in gate over drive (V_{CC}-V_T) is becoming so severe that this trend cannot continue without substantial loss in device performance. One possible solution that has been proposed is the use of dual threshold voltage transistors. It will be shown, however, that this will only extend the scaling trend by one technology generation at most.

V_{CC} and V_T Scaling

Figure 36 shows power supply and threshold voltage trends for Intel's microprocessor process technologies. As seen, the power supply is decreasing much more rapidly than threshold voltage. This has severe implications for device performance. Transistor drive current and therefore circuit performance is proportional to gate over drive (V_{CC}-V_T) raised to the power **n** where **n** is between 1 and 2 ($(V_{CC}-V_T)^n$).

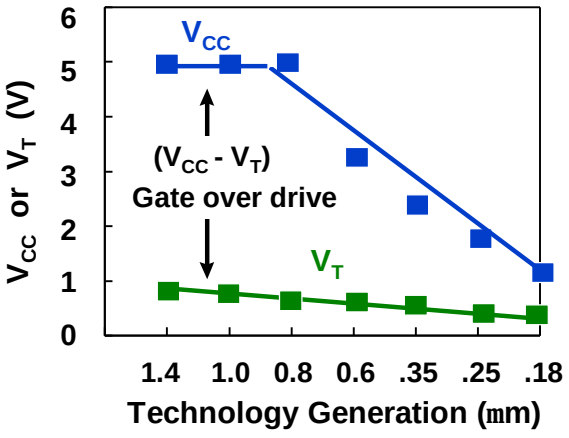


Figure 36: Power supply and threshold voltage scaling trend

In Figure 36, the gate over drive is shown to be rapidly decreasing for deep sub-micron devices, thereby strongly degrading device performance. As discussed previously, aggressive oxide, SDE, and well engineering are used to overcome the loss in gate drive and maintain the historical rate of transistor improvement.

To understand why these power supply and threshold voltages are being chosen, we need to understand chip active and standby power trends. Active power is set by circuit switching and is defined as $P = C_{LOAD} V_{CC}^2 f$ where **f** is the operating frequency and $\mathbf{C_{LOAD}}$ is the switching capacitance of the gate and wire load. Chip active power and frequency trends are shown for Intel's process technologies in Figure 37. Standby power results from junction and transistor sub-threshold source-to-drain leakage. For 0.1µm transistors, the sub-threshold leakage is the dominant contributor to standby power.

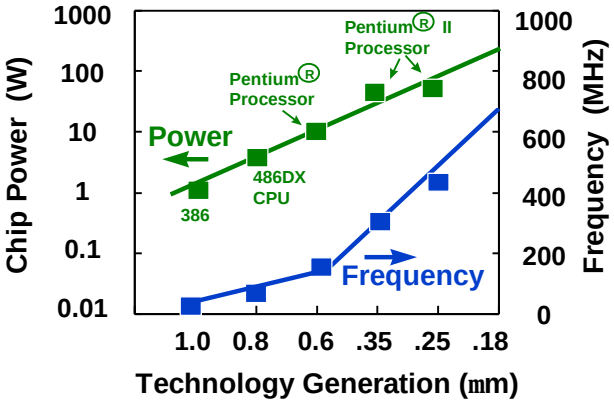


Figure 37: Chip power and frequency trends for Intel's process technologies

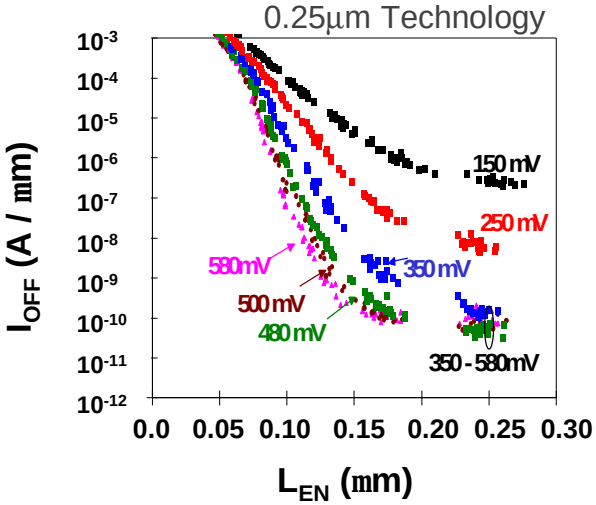


Figure 38: Off-state leakage versus channel length for 0.25µm transistors with different threshold voltage

Sub-threshold leakage is fundamental to silicon MOSFET operation and is set by the device threshold voltage. Sub-threshold off-state leakage versus channel length characteristics is shown in Figure 38. The active and standby power trends for Intel's process technologies are shown in Figure 39. In this figure, several interesting points can be observed. First, as microprocessor complexity increases, chip power is increasing to ~10-20W. Second, standby power for 1µm technology was .01% of active power, but is approaching 10% of active power in 0.1µm technologies. In order to limit the increase of standby power, threshold voltages need to increase. However, this increase strongly affects device performance because of reduced gate over drive. To maintain acceptable leakage values, the V_T's of transistors will need to increase by >0.25 V.

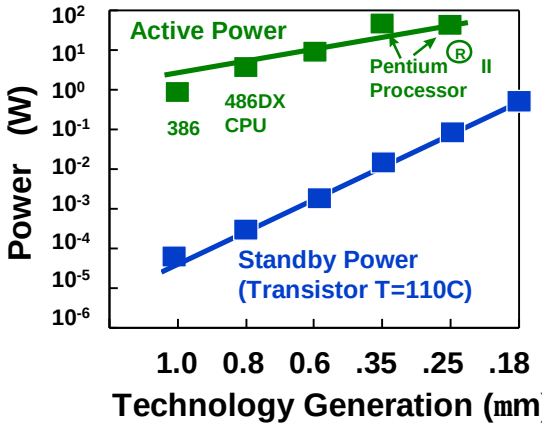


Figure 39: Active and standby power trends for Intel's technologies

Dual V_T Architecture

If power supply and threshold voltage scaling continues at the current trend, further reduction in gate overdrive will occur. A general rule for high performance transistor design is to maintain a V_{CC}/V_T ratio of at least four. A ratio of four provides a gate swing of one V_T to turn the device off and three V_T to drive the device. Figure 40 plots the V_{CC}/V_T ratio for Intel's previous technologies as well as the current projected trend. The projected scaling trend shows that beyond the 0.25µm technology, the ratio of V_{CC}/V_T will drop below 4. One technique to improve the gate drive and standby power trend is to offer circuit designers dual threshold voltage devices. This would consist of designing a high-performance, high-leakage, low-threshold voltage device and a low-performance, low-leakage, high-threshold voltage device. A chip would be designed such that only the critical paths would use the high-performance/high-leakage devices.

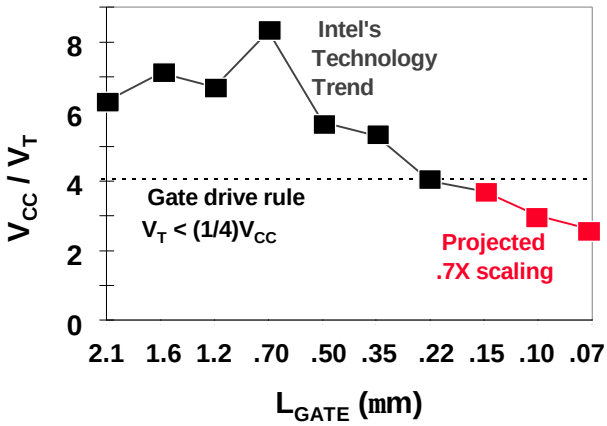


Figure 40: V_{CC}/V_T trend for Intel's process technologies

Figure 41 shows the performance and leakage current tradeoff for 0.25µm technology, lower threshold voltage devices. A 100x increase in leakage current would be required to extend the present performance trend by one generation. Whether or not a 100x increase in leakage could be tolerated would depend heavily on the circuit architecture and power constraints of the chip.

Alternate Device Options

Many designers have proposed new device architectures to improve device and circuit performance. In this section, we evaluate three of the most widely explored options and discuss the potential advantages and disadvantages of each.

SOI Device

One technique proposed to improve CMOS performance is to fabricate the devices on a silicon on insulator (SOI) substrate. SOI devices are classified into two types depending on the extent of the channel depletion layer (partially depleted or fully depleted) compared to the silicon thickness (T_{Si}). Fully depleted devices are not practical for deep sub-micron devices since the silicon thickness needs to be ~10.0 nm to control short channel effects. This silicon thickness is extremely difficult to manufacture and causes large device external resistance due to shallow SDE depths. Partially depleted devices are more suitable for deep sub-micron devices. However, since the channel region of the silicon layer is not entirely channel depleted, a partially depleted device offers no advantage for short channel effects or channel length scaling.

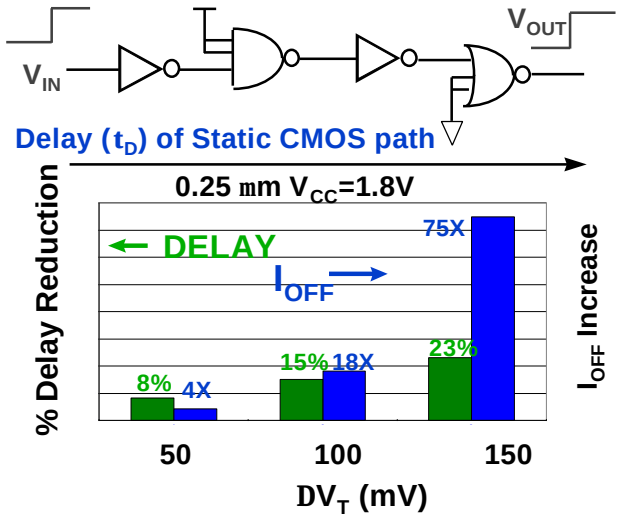


Figure 41: Performance and leakage current tradeoff for lower threshold voltage devices

Actually the partially depleted floating body can degrade short channel effects because of an uncontrolled lowering of V_T that is caused by impact ionization [28]. If the floating body can be controlled, partially depleted devices offer improvements in junction area capacitance, device body effect, and a gate-to-body coupling, which potentially results in a slightly larger drive current during switching.

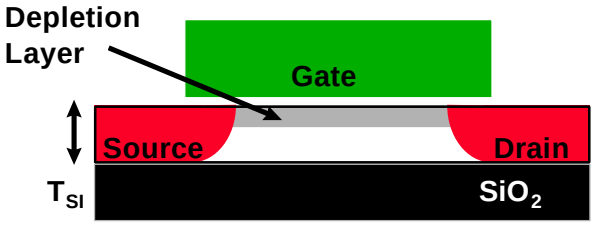


Figure 42: Cross section of an SOI device

Parameter	Best Case Gain
Junction Capacitance	12%
Body Factor	3%
Gate-to-Body Coupling	3%
Channel Length	0%
Total	18%

Table 3: Estimated improvement in circuit speed by device feature for a SOI device with unconstrained I_{OFF}

The best case estimated impact of these parameters on current generation circuit's speed improvements is shown in Table 3. We call it best case, since to date, no literature paper has demonstrated these device parasitic improvements without increasing the transistor off-state leakage. Studies done at Intel indicate that NMOS SOI devices require a somewhat higher threshold voltage than bulk devices to maintain an equivalent off-state leakage due to the floating body effect[28]. This higher threshold voltage offsets some of the other potential performance advantages of SOI. Also, in future high performance microprocessors where interconnect capacitances are becoming more dominant, the junction capacitance advantage of SOI will become less important. In summary, the performance gain going to the SOI architecture is less than one generation and will pose serious complications for circuit design due to floating body effects.

$Si_{1-x}Ge_x$ Channel Device

Another technique to improve transistor performance is to fabricate the device in a $Si_{1-x}Ge_x$ channel (see Figure 43). The $Si_{1-x}Ge_x$ channel region has been shown to increase hole mobility [29]. There are two reasons for the mobility gain: $Si_{1-x}Ge_x$ under compressive strain has improved mobility over Si; and the valence band offset between Si and $Si_{1-x}Ge_x$ localizes the hole inversion charge away from the SiO_2/Si interface, which reduces the effects of surface roughness scattering. Unfortunately, improving mobility becomes less important as the transistor is scaled into the deep sub-micron regime. This is due to the high lateral electric fields that cause the carrier velocity to saturate.

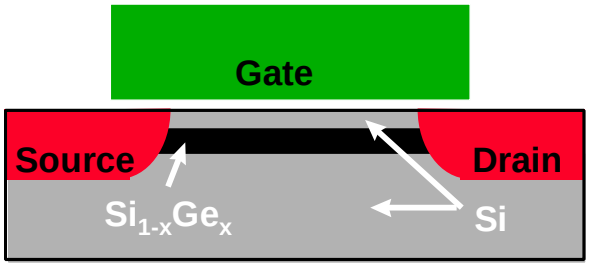


Figure 43: Cross section of a transistor fabricated with a $Si_{1-x}Ge_x$ channel

In Figure 44, the ratio of saturated drive current to mobility change is plotted for different device sizes. For long channel device lengths, the improvement in drive current is equal to the improvement in mobility. However, for deep sub-micron devices with channel lengths of ~0.1µm, a 4% improvement in mobility improves drive current by only 1%. If a $Si_{1-x}Ge_x$ channel improved electron or hole saturation velocity, there would be an improvement in drive current.

Unfortunately, electron and hole saturation velocities are similar if not slightly lower in SiGe than they are in silicon.

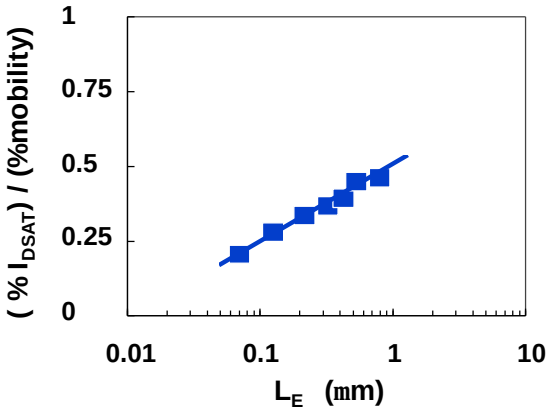


Figure 44: Ratio of I_{DSAT} change to mobility change versus channel length (for smaller devices, high electric fields cause velocity saturation)

Dynamic V_T Device

For low supply voltage operation (<0.6 V), a dynamic threshold voltage MOS device (DTMOS) has been proposed [30,31]. A DTMOS is formed by connecting the gate to the well as shown in Figure 45. This connection causes the threshold voltage of the device to be lowered during switching thereby increasing the transistor drive current. This technique is limited to supply voltages less than 0.6V to prevent the forward bias well-to-source junction from conducting large forward bias diode currents. The DTMOS technique has been proposed for devices fabricated on either bulk silicon or SOI substrates. Fabrication of these devices on SOI substrates is easier due to the electrical isolation of both n- and p-wells.

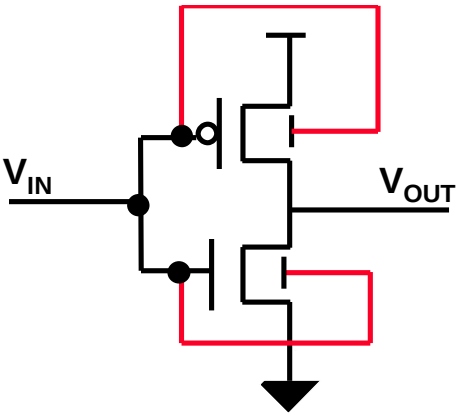


Figure 45: Circuit schematic of a dynamic threshold voltage MOS inverter

This technique can increase transistor drive current by over 20% through improved gate over drive (V_G-V_T). However, this technique offers little to no net gain over high performance, optimized, static V_T CMOS when differences in chip area are considered. When DTMOS is implemented on bulk silicon substrate (see Figure 46), there is a large performance degradation due to the increase in the switching load capacitance that is comprised of junction (C_J) and depletion (C_D) capacitance.

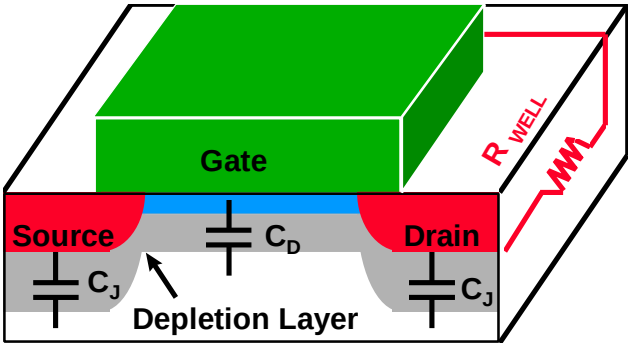


Figure 46: Transistor cross schematic of a dynamic threshold voltage MOS inverter

The performance degradation from the junction and depletion capacitance can be significantly reduced for DTMOS fabricated on an SOI substrate. However, for DTMOS on SOI, the RC time constant associated with the well resistance (R_{WELL}) and depletion capacitance (C_D) is not compatible with high frequency microprocessor applications. The R_{WELL}*C_D time constant can be ~1ns, which would consume half of the clock period for today's 500 MHz microprocessors. To minimize the RC delay associated with the poly-Si gate, companies have added metals to reduce the resistance to 2-3Ω/sq. By comparison, a DTMOS device in SOI can easily have a resistance component (R_{WELL}) on the order of 10^4-10^5 Ω/sq. or greater.

Although each of these alternate device structures has certain advantages, the overall device improvement is relatively small. In addition, manufacturing costs and circuit issues make it extremely difficult to justify the adoption of any of these device architectures.

Conclusions

Current performance scaling trends will not continue past the 0.13 - 0.10µm device technologies by using traditional scaling methods. Fundamental limits in SiO_2 scaling due to tunneling currents, in SDE junction depths due to large increases in external resistance, and in well engineering due to leakage constraints are currently being reached. At present, there is no clear alternate device architecture that has shown the potential for continuing the performance trends seen in the last 20 years. Aggressive exploration of high dielectric constant materials as well as developing a way to decrease SDE resistance offer the best hope for device and circuit improvements into the next century. These should be strongly supported.

Acknowledgments

The authors would like to acknowledge the collaborative efforts of our colleagues in the Portland Technology Development and Technology Computer Aided Design groups: T. Ghani, R. Rios, M. Stettler, M. Alavi, I. Post, S. Tyagi, R. Chau, M. Taylor, R. Nagisetty, J. Sandford, S. Ahmed, and S. Yang. The management support from R. Gasser, J. Garcia, S. Yang, L. Yau, and Y. El-Mansy is greatly appreciated.

References

[1] M. Bohr, S.U. Ahmed, L. Brigham, R. Chau, R. Gasser, R. Green, W. Hargrove, E. Lee, R. Natter, S. Thompson, K. Weldon and S. Yang, *IEDM Technical Digest*, 1994, p. 273.

[2] M. Bohr, S.S. Ahmed, S.U. Ahmed, M. Bost, T. Ghani, J. Greason, R. Hainsey, C. Jan, P. Packan, S. Sivakumar, S. Thompson, J. Tsai, and S. Yang, *IEDM Technical Digest*, 1996, p. 847.

[3] H.S. Momose, M. Ono, T. Yoshitomi, T. Ohguro, S-I. Nakamura, M. Saito, and H. Iwai, *IEDM Technical Digest*, 1994, p. 593.

[4] S.A. Hareland, S. Krishnamurthy, S. Jallepalli, C.-F. Yeap, K. Hasnat, A.F. Tasch, and C.M. Maziar, *IEDM Technical Digest*, 1995, p. 933.

[5] S.-H.Lo, D.A. Buchanan, Y. Taur, and W. Wang, *IEEE Electron Device Letter*, 1997, p. 209.

[6] S. A. Campbell, D.C. Gilmer, X.-C. Wang, M.-T. Hsieh, H.-S. Kim, W.L. Gladfelter, and J.Yan, *IEEE Electron Device*, 1997, p. 104.

[7] S. Kamiyama and T. Saeki, *IEDM Technical Digest*, 1991, p. 827.

[8] C.G. Parker, G. Lucovsky, and J.R. Hauser, to be published, 1997.

[9] H.-H. Tseng, P.G.Y. Tsui, P.J. Tobin, J. Mogab, M. Khare, X.W. Wang, T.P. Ma, R. Hegde, C.Hobbs, J.Veteran, M. Hartig, G. Kenig, V. Wang, R. Blumenthal, R. Cotton, V. Kaushik, T. Tamagawa, B.L. Halpern, G. J. Cui, and J. J. Schmitt, *IEDM Technical Digest*, 1997, p. 647.

[10] S. Thompson, *VLSI Symposium Technology Short Course*, 1998.

[11] M. Ono, M. Saito, T. Yoshitomi, C. Fiegna, T. Ohguro, and H. Iwai, *IEDM Technical Digest*, 1993, p. 119.

[12] A. Hori, H. Nakaoka, H. Umimoto, K. Yamashita, M. Takase, N. Shimizu, B. Mizuno and S. Odanaka, *IEDM Technical Digest*, 1994, p. 485.

[13] P.A. Packan and J.D. Plummer, *Applied Physics Letter*, 1990, p. 1787.

[14] D.F. Downey, C.M. Osburn, S.D. Marcus, *Solid State Technology*, 1997, p. 71.

[15] D.J. Eaglesham, P.A. Stolk, H.-J. Gossmann, and J.M. Poate, *Applied Physics Letter*, 1994, p. 2305.

[16] A.D. Lilak, S.K. Earles, K.S. Jones, M.E. Law, *IEDM Technical Digest*, 1997, p. 493.

[17] M.E. Law and K.S. Jones, *Proceedings of the Process Physics Symposium of the Electrochemical Society*, 1996, p. 374.

[18] S.E. Thompson, P.A. Packan, and M.T. Bohr, *VLSI Symposium Digest*, 1996, p. 154.

[19] S. Venkatesan, J.W. Lutze, C. Lage and W.J. Taylor, *IEDM Technical Digest*, 1995, p. 419.

[20] M. Rodder, S. Aur, And I.-C. Chen, *IEDM Technical Digest*, 1995, p. 415.

[21] J.B. Jacobs and D. Antoniadis, *IEEE Transactions Electron Devices*, 1995, p. 870.

[22] G.G. Shahidi, J.D. Warnock, J. Comfort, S. Fischer, P.A. McFarland, A. Acovic, T.I. Chappell, B.A. Chappell, T.H. Ning, C.J. Anderson, R.H. Dennard, J.Y.C. Sun, M.R. Polcari, and B. Davari, *IBM Journal Research Development*, 1995, p. 229.

[23] M. Cao, P. Griffin, P. Vande Voorde, C. Diaz, and W. Greene, *VLSI Symposium Digest*, 1997, p. 85.

[24] C.T. Sah, *Fundamentals of Solid-State Electronics*, 1991, p. 553.

[25] C.F. Codella and S. Ogura, *IEDM Technical Digest*, 1985, p. 230.

[26] T. Hori, *IEDM Technical Digest*, 1994, p. 75.

[27] Y. Taur and E.J. Nowak, *IEDM Technical Digest*, 1997, p. 215.

[28] R. Chau, R. Arghavani, M. Alavi, D. Douglas, J. Greason, R. Green, S. Tyagi, J. Xu, P. Packan, S. Yu, and C. Liang, *IEDM Technical Digest*, 1997, p. 591.

[29] K. Ismail, J.O. Chu, and B. S. Meyerson, *Applied Physics Letter, vol. 64*, 1994, p. 3124.

[30] F. Assaderaghi, D. Sinitsky, S. Parke, J. Bokor, P.K. Ko, and C. Hu, *IEDM Technical Digest*, 1994, p. 809.

[31] A. Shibata, T. Matsuoka, S. Kakimoto, H. Kotaki, M. Nakano, K. Adachi, K. Ohta, and N. Hashizume, *VLSI Symposium Digest*, 1998, p. 76.

Authors' Biographies

Scott Thompson joined Intel in 1992 after completing his Ph.D. under Professor C. T. Sah at the University of Florida on thin gate oxides. He has worked on transistor design and front-end process integration on Intel's 0.35, 0.25, and 0.18µm silicon process technology design for the Pentium® and the Pentium® II microprocessors. Scott is currently managing the development of Intel's 0.13µm transistor design. His email address is scott.thompson@intel.com.

Paul Packan received his Ph.D. degree in Electrical Engineering in 1991 from Stanford University. He joined Siemens AG in Munich Germany in 1991 working in the area of high speed bipolar transistor architecture. In 1992 he joined Intel Corp. working in the field of process and device simulation for MOS devices. He worked on the development of the 0.35, 0.25 and 0.18 um technologies and is currently managing the process and device modeling group. His email address is paul.a.packan@intel.com.

Mark T. Bohr joined Intel in 1978 after receiving a MSEE from the University of Illinois. He has been a member of the Portland Technology Development group since 1978 and has been responsible for process integration and device design on a variety of DRAM, SRAM, and logic technologies, including recently 0.35µm and 0.25µm logic technologies. He is an Intel Fellow and director of process architecture and integration. He is currently directing development activities on 0.18µm and 0.13µm logic technologies. His email address is mark.bohr@intel.com.

EUV Lithography—The Successor to Optical Lithography?

John E. Bjorkholm
Advanced Lithography Department, Technology and Manufacturing Group, Santa Clara, CA.
Intel Corporation

Index words: EUV lithography, lithography, microlithography

Abstract

This paper discusses the basic concepts and current state of development of EUV lithography (EUVL), a relatively new form of lithography that uses extreme ultraviolet (EUV) radiation with a wavelength in the range of 10 to 14 nanometer (nm) to carry out projection imaging. Currently, and for the last several decades, optical projection lithography has been the lithographic technique used in the high-volume manufacture of integrated circuits. It is widely anticipated that improvements in this technology will allow it to remain the semiconductor industry's workhorse through the 100 nm generation of devices. However, some time around the year 2005, so-called Next-Generation Lithographies will be required. EUVL is one such technology vying to become the successor to optical lithography. This paper provides an overview of the capabilities of EUVL, and explains how EUVL might be implemented. The challenges that must be overcome in order for EUVL to qualify for high-volume manufacture are also discussed.

Introduction

Optical projection lithography is the technology used to print the intricate patterns that define integrated circuits onto semiconductor wafers. Typically, a pattern on a mask is imaged, with a reduction of 4:1, by a highly accurate camera onto a silicon wafer coated with photoresist. Continued improvements in optical projection lithography have enabled the printing of ever finer features, the smallest feature size decreasing by about 30% every two years. This, in turn, has allowed the integrated circuit industry to produce ever more powerful and cost-effective semiconductor devices. On average, the number of transistors in a state-of-the-art integrated circuit has doubled every 18 months.

Currently, the most advanced lithographic tools used in high-volume manufacture employ deep-ultraviolet (DUV) radiation with a wavelength of 248 nm to print features that have line widths as small as 200 nm. It is believed that new DUV tools, presently in advanced development, that employ radiation that has a wavelength of 193 nm, will enable optical lithography to print features as small as 100 nm, but only with very great difficulty for high-volume manufacture. Over the next several years it will be necessary for the semiconductor industry to identify a new lithographic technology that will carry it into the future, eventually enabling the printing of lines as small as 30 nm. Potential successors to optical projection lithography are being aggressively developed. These are known as "Next-Generation Lithographies" (NGL's). EUV lithography (EUVL) is one of the leading NGL technologies; others include X-Ray lithography, ion-beam projection lithography, and electron-beam projection lithography. [1]

In many respects, EUVL may be viewed as a natural extension of optical projection lithography since it uses short wavelength radiation (light) to carry out projection imaging. In spite of this similarity, there are major differences between the two technologies. Most of these differences occur because the properties of materials in the EUV portion of the electromagnetic spectrum are very different from those in the visible and UV wavelength ranges. The purpose of this paper is to explain what EUVL is and why it is of interest, to describe the current status of its development, and to provide the reader with an understanding of the challenges that must be overcome if EUVL is to fulfill its promise in high-volume manufacture.

Why EUVL?

In order to keep pace with the demand for the printing of ever smaller features, lithography tool manufacturers have found it necessary to gradually reduce the wavelength of the light used for imaging and to design imaging systems with ever larger numerical apertures. The reasons for these changes can be understood from the following equations that describe two of the most fundamental characteristics of an imaging system: its

resolution (RES) and depth of focus (DOF). These equations are usually expressed as

$$RES = k_1 \lambda / NA \quad (1a)$$

and

$$DOF = k_2 \lambda / (NA)^2, \quad (1b)$$

where λ is the wavelength of the radiation used to carry out the imaging, and NA is the numerical aperture of the imaging system (or camera). These equations show that better resolution can be achieved by reducing λ and increasing NA. The penalty for doing this, however, is that the DOF is decreased. Until recently, the DOF used in manufacturing exceeded 0.5 µm, which provided for sufficient process control.

The case $k_1 = k_2 = ½$ corresponds to the usual definition of diffraction-limited imaging. In practice, however, the acceptable values for k_1 and k_2 are determined experimentally and are those values which yield the desired control of critical dimensions (CD's) within a tolerable process window. Camera performance has a major impact on determining these values; other factors that have nothing to do with the camera also play a role. Such factors include the contrast of the resist being used and the characteristics of any etching processes used. Historically, values for k_1 and k_2 greater than 0.6 have been used comfortably in high-volume manufacture. Recently, however, it has been necessary to extend imaging technologies to ever better resolution by using smaller values for k_1 and k_2 and by accepting the need for tighter process control. This scenario is schematically diagrammed in Figure 1, where the values for k_1 and DOF associated with lithography using light at 248 nm and 193 nm to print past, present, and future CD's ranging from 350 nm to 100 nm are shown. The "Comfort Zone for Manufacture" corresponds to the region for which $k_1 > 0.6$ and DOF > 0.5 µm. Also shown are the k_1 and DOF values currently associated with the EUVL printing of 100 nm features, which will be explained later. As shown in the figure, in the very near future it will be necessary to utilize k_1 values that are considerably less than 0.5. Problems associated with small k_1 values include a large iso/dense bias (different conditions needed for the proper printing of isolated and dense features), poor CD control, nonlinear printing (different conditions needed for the proper printing of large and small features), and magnification of mask CD errors. Figure 1 also shows that the DOF values associated with future lithography will be uncomfortably small. Of course, resolution enhancement techniques such as phase-shift masks, modified illumination schemes, and optical proximity correction can be used to enhance resolution while increasing the effective DOF. However, these techniques are not generally applicable to all feature geometries and are difficult to implement in manufacturing. The degree to which these techniques can be employed in manufacturing will determine how far optical lithography can be extended before an NGL is needed.

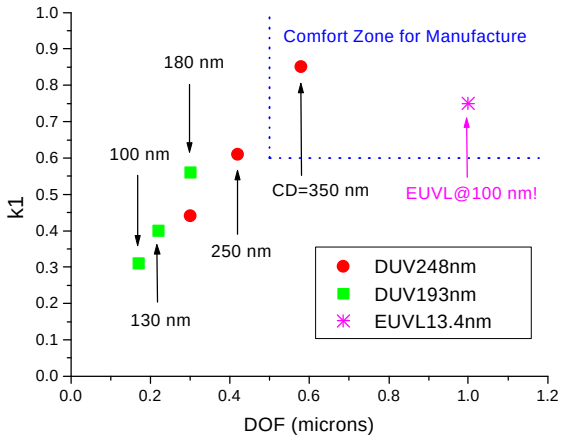


Figure 1: The k_1 and DOF values associated with 248 nm and 193 nm lithographies for the printing of CD values ranging from 350 nm down to 100nm assuming that $k_2 = k_1$ and NA = 0.6

EUVL alleviates the foregoing problems by drastically decreasing the wavelength used to carry out imaging. Consider Figure 2. The dashed black line shows the locus of points corresponding to a resolution of 100 nm; the region to the right of the line corresponds to even better resolution.

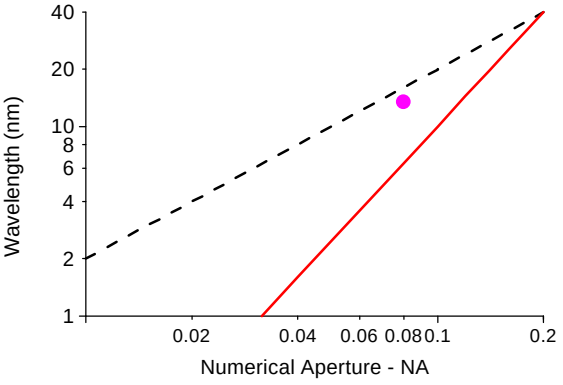


Figure 2: The region between the lines shows the wavelength and numerical aperture of cameras simultaneously having a resolution of 100 nm or better and a DOF of 0.5 µm or better

The solid red line shows the locus of points for which the DOF is 0.5 μm; in the region to the left of that line the DOF values are larger. Points in the region between the two lines correspond to situations in which the resolution is 100 nm or better, and the DOF is 0.5 μm or longer. As shown, to be in this favorable region, the wavelength of the light used for imaging must be less than 40 nm, and the NA of the imaging system must be less than 0.2. The solid circle shows the parameters used in current imaging experiments. Light having wavelengths in the spectral region from 40 nm to 1 nm is variously referred to as extreme uv, vacuum uv, or soft x-ray radiation. Projection lithography carried out with light in this region has come to be known as EUV lithography (EUVL). Early in the development of EUVL, the technology was called soft x-ray projection lithography (SXPL), but that name was dropped in order to avoid confusion with x-ray lithography, which is a 1:1, near-contact printing technology.

As explained above, EUVL is capable of printing features of 100 nm and smaller while achieving a DOF of 0.5 μm and larger. Currently, most EUVL work is carried out in a wavelength region around 13 nm using cameras that have an NA of about 0.1, which places the technology well within the "Comfort Zone for Manufacture" as shown in Figure 1 by the data point farthest to the right.

EUVL Technology

In many respects, EUVL retains the look and feel of optical lithography as practiced today. For example, the basic optical design tools that are used for EUV imaging system design and for EUV image simulations are also used today for optical projection lithography. Nonetheless, in other respects EUVL technology is very different from what the industry is familiar with. Most of these differences arise because the properties of materials in the EUV are very different from their properties in the visible and UV ranges.

Foremost among those differences is the fact that EUV radiation is strongly absorbed in virtually all materials, even gases. EUV imaging must be carried out in a near vacuum. Absorption also rules out the use of refractive optical elements, such as lenses and transmission masks. Thus EUVL imaging systems are entirely reflective. Ironically, the EUV reflectivity of individual materials at near-normal incidence is very low. In order to achieve reasonable reflectivities near normal incidence, surfaces must be coated with multilayer, thin-film coatings known as distributed Bragg reflectors. The best of these function in the region between 11 and 14 nm. EUV absorption in standard optical photoresists is very high, and new resist and processing techniques will be required for application in EUVL.

Because EUVL utilizes short wavelength radiation for imaging, the mirrors that comprise the camera will be required to exhibit an unprecedented degree of perfection in surface figure and surface finish in order to achieve diffraction-limited imaging. Fabrication of mirrors exhibiting such perfection will require new and more accurate polishing and metrology techniques.

Clearly, then, there are a number of new technology problems that arise specifically because of the use of EUV radiation. Intel has formed a consortium called the EUV, LLC (the LLC), which currently also includes AMD and Motorola, to support development of these EUV-specific technologies. The bulk of this development work is carried out by three national laboratories functioning as a single entity called the Virtual National Laboratory (VNL). Participants in the VNL are Lawrence Livermore National Laboratory, Sandia National Laboratories, and Lawrence Berkeley National Laboratory. Development work is also carried out by LLC members, primarily on mask fabrication and photoresist development. Recently, additional support for some of this work has come from Sematech. The work described in the following sections was carried out within this program, primarily by workers within the VNL.

Multilayer Reflectors

In order to achieve reasonable reflectivities, the reflecting surfaces in EUVL imaging systems are coated with multilayer thin films (ML's). These coatings consist of a large number of alternating layers of materials having dissimilar EUV optical constants, and they provide a resonant reflectivity when the period of the layers is approximately $\lambda/2$. Without such reflectors, EUVL would not be possible. On the other hand, the resonant behavior of ML's complicates the design, analysis, and fabrication of EUV cameras. The most developed and best understood EUV multilayers are made of alternating layers of Mo and Si, and they function best for wavelengths of about 13 nm. Figure 3 shows the reflectivity and phase change upon reflection for an Mo:Si ML that has been optimized for peak reflectivity at 13.4 nm at normal incidence; similar resonance behavior is seen as a function of angle of incidence for a fixed wavelength. While the curve shown is theoretical, peak reflectivites of 68% can now be routinely attained for Mo:Si ML's deposited by magnetron sputtering.

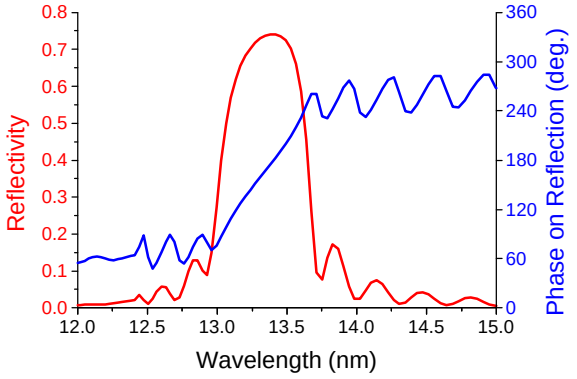


Figure 3: Curve showing the normal incidence reflectivity and phase upon reflection of an Mo:Si ML as a function of wavelength; the coating was designed to have peak reflectivity at 13.4 nm

This resonance behavior has important implications for EUVL. A typical EUVL camera is composed of at least four mirrors, and light falls onto the various mirrors over different angular ranges. As a consequence, the periods of the ML's applied to the various mirrors must be different so that all the mirrors are tuned to reflect the same wavelength. Proper matching of the peak wavelengths is crucial for achieving high radiation throughput and good imaging performance. The range of angles of incidence over a single mirror surface must also be considered. For some optical designs, the angular ranges are small enough that ML's with a uniform period over the surface can be used. In other designs, the angular ranges are so large that the ML period must be accurately varied over the surface in order to achieve uniform reflectivity. There are optical designs in which the angular ranges are so large that ML reflectors can not be utilized.

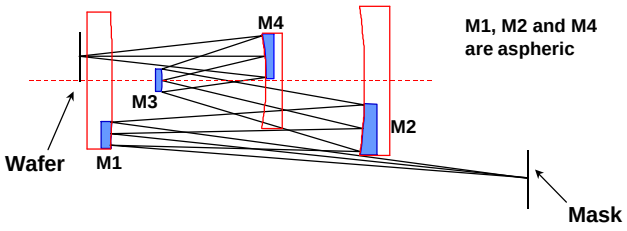
The effects on imaging performance due to the variations of ML reflectivity and phase with wavelength and angle have been extensively modeled. The effects have been shown to be minimal for cameras of interest to us. The primary perturbations of the wavefront transmitted by the camera are described as a simple tilt and defocus.

In our work we are fabricating two types of EUV cameras. The first is a small field, microstepper-like design that utilizes two mirrors and that images with a reduction factor of 10. We call it the "10X camera." This camera has been used extensively in our early investigations of EUV imaging. One of the mirrors in this camera requires a strongly graded ML coating. Three of these cameras have been fabricated and have been shown to perform well. (Examples of the imaging performance of these cameras are shown later in this paper.) The second camera, currently being fabricated, is a prototype lithography camera with a ring field of 26 mm X 1.5 mm. This camera was designed so that it will perform well with uniform ML coatings. The VNL has demonstrated the ability to achieve the ML matching, uniformity, and grading requirements of EUVL cameras currently of interest.

EUV Cameras

Designing an all-reflective camera that achieves lithographic-quality imaging is more difficult than designing a refractive imaging system because mirrors have fewer degrees of freedom to vary than do lenses. As a result, most of the mirrors in an EUVL camera will have aspheric surfaces. The detailed reasoning that leads to this conclusion was first discussed in 1990. [2]

A schematic of a four-mirror camera that the VNL is in the process of fabricating is shown in Figure 4. The mirror segments shown in blue are the pieces actually being fabricated, while the full, on-axis "parent" mirrors are shown in red. This camera will become part of an



"engineering test stand," so it is called the ETS camera.

Figure 4: Schematic diagram of the 4-mirror ETS camera

It has an NA = 0.1 and is designed to be used with Mo:Si ML's at a wavelength of 13.4 nm. Mirror 3 is spherical, and the other three mirrors are aspheres. Some of the most important features of this camera are as follows:

- Its resolution is better than 100 nm over a 26 mm x 1.5 mm, ring-shaped field.
- It images with a reduction factor of 4.
- The departures of the aspheres from a best-fit sphere are less than 10 µm.

The camera is intended for use in a step-and-scan lithography system. In actual operation, the mask and wafer are simultaneously scanned in opposite directions, with the mask moving four times faster than the wafer, as

done in current DUV step-and-scan systems. The design of this camera has been optimized so that the effective distortion when scanning (about 1 nm) is considerably less than the distortion obtained for static printing (15 nm).

Because short wavelength radiation is used to carry out the imaging, the surfaces of the mirrors are required to exhibit unprecedented perfection. In order to achieve diffraction-limited imaging at 13.4 nm, the root-mean-square (rms) wavefront error of the camera must be less than 1 nm. Assuming that the surface errors on the mirrors are randomly distributed, this means that the surface figure (basic shape) of each mirror must be accurate to 0.25 nm (2.5 angstroms!) rms, or better. Until recently, achieving this kind of surface figure accuracy was out of the question, even for spheres. Furthermore, aspheres are much more difficult to fabricate than are spheres. We have been working closely with optics fabricators to address this issue, and dramatic progress has been made over the last 18 months.

The figure of a surface refers to its basic shape. Stringent requirements must also be placed on the roughness of the surfaces. For our purposes, we define surface figure errors as those errors that have a spatial wavelength scale of 1 mm or longer; such errors are typically measured deterministically using instruments such as interferometers. We define surface roughness as surface errors with a spatial wavelength scale shorter than 1 mm. Typically such surface errors are described and measured statistically. We define roughness with wavelengths in the range of 1 mm through 1 μm as mid-spatial frequency roughness (MSFR). Roughness in this frequency range causes small-angle scattering of light off the mirror surfaces. This scattering causes a reduction in the contrast of images because it scatters light from bright regions of the image plane onto regions intended to be dark. This scattering is often called flare. Because the effects of scatter scale as $1/\lambda^2$, the deleterious effects of flare are becoming more evident as the wavelengths used for lithography continue to be reduced. For a given surface roughness, the amount of scattering at 13.4 nm is approximately 340 times larger than that at 248 nm. In order to keep flare to manageable levels in EUVL, the MSFR must be 0.2 nm rms, or less. Until recently, even the best surfaces exhibited MSFR of 0.7 nm rms. Roughness with spatial wavelengths less than 1 μm is called high-spatial-frequency roughness (HSFR), and it causes large angle scattering off the mirrors. Light scattered at such angles is typically scattered out of the image field and represents a loss mechanism for light. We require HSFR to be less than 0.1 nm rms. Optical fabricators have for some time been able to use "super-polishing" techniques to produce surfaces with HSFR even better than this. A well-polished silicon wafer also exhibits such HSFR.

The challenge for a fabricator of optics for EUVL is to achieve the desired levels of figure accuracy and surface roughness simultaneously. The manufacturer we have been working with has made exceptional progress in this regard. As a measure of the progress that has been made, the first copy of Mirror 3 has been completed, and its surface has been measured and found to have the following characteristics:

- Surface figure: 0.44 nm rms
- MSFR: 0.31 nm rms
- HSFR: 0.14 nm rms

This result demonstrates excellent progress towards the surface specifications that we need to achieve.

Metrology

The progress made in optics fabrication described above could not have been achieved without access to appropriate metrology tools. Some of the required tools were recently developed by workers within the VNL.

Two very significant advances have been made in the measurement of figure. Previous to these advances, no tools existed that could measure figure to the accuracy we require. The first of these innovations is the Sommargren interferometer, which uses visible light to achieve unprecedented accuracy. [3] In this version of a "point-diffraction interferometer," the wavefront to be measured is compared with a highly accurate spherical wave generated by an optical fiber or by an accurate, small pinhole. Interferogram stitching algorithms have been developed that allow aspheric surfaces to be measured without the need for null optics, which are typically the weak link in such measurements. An accuracy of 0.25 nm rms has already been demonstrated, and an engineering path exists for improvements down to one half that value. Four versions of the interferometer have been supplied to our optics manufacturer for use in the fabrication of the four individual mirrors of the ETS camera. The interferometer can also be configured to measure the wavefront quality of an assembled camera. However, visible light does not interact with ML reflectors in the same manner as EUV light. Thus it is of great importance to be able to characterize an EUV camera using light at the wavelength of intended operation. To this end, an EUV interferometer has been developed which will be used to characterize the wavefront quality of assembled EUV cameras and to

guide final adjustments of the camera alignment. [4] This system has been shown to have an innate rms accuracy of better than 0.003 waves at the EUV wavelength! Its accuracy is far better than needed to qualify an EUV camera as diffraction-limited.

Several commercial instruments have been used to measure surface roughness. An interference microscope was used to measure MSFR, and an atomic force microscope (AFM) was used to measure HSFR. The relevance of these measurements was verified by making detailed precision measurements of the magnitude and angular dependence of EUV scattering off of surfaces characterized with the other instruments. Excellent agreement has been obtained between the direct scattering measurements and the predictions based on the measurements of MSFR and HSFR.

Masks

EUVL masks are reflective, not transmissive. They consist of a patterned absorber of EUV radiation placed on top of an ML reflector deposited on a robust and solid substrate, such as a silicon wafer. Membrane masks are not required. The reflectance spectrum of the mask must be matched to that of the ML-coated mirrors in the camera. It is anticipated that EUVL masks will be fabricated using processing techniques that are standard in semiconductor production. Because a 4:1 reduction is used in the imaging, the size and placement accuracy of the features on the mask are achieved relatively easily.

Nonetheless, there are a number of serious concerns about mask development. The foremost is the fact that there is no known method for repairing defects in an ML coating. Since masks must be free of defects, a technique must be developed for depositing defect-free ML reflectors. The defect densities in ML coatings produced by magnetron sputtering have been found to be adequate for camera mirrors, but far too high for mask blanks. As a result, a much cleaner deposition system that uses ion-beam sputtering has been constructed. A reduction of about 1000 in the density of defects larger than 130 nm, to a level of better than $0.1/cm^2$, has been obtained with this system, but further improvement will certainly be required. Present defect detection techniques use visible light, and it is all but certain that the density of defects printable with EUV light is higher. Defects can take the form of amplitude or phase perturbations, and the proper tools for detecting EUV-printable defects are currently being developed. Initially it will be necessary to inspect the mask blanks using EUV radiation. In the long run, it is hoped that experience will show that adequate inspection can be carried out with commercially available visible-light and e-beam inspection tools.

Finally, in current practice, pellicles are used to protect masks from contamination. The use of pellicles in EUVL will not be possible because of the undesirable absorption that would be encountered. Other methods for protecting EUV masks are under development.

Sources of EUV Radiation

A number of sources of EUV radiation have been used to date in the development of EUVL. Radiation has been obtained from a variety of laser-produced plasmas and from the bending magnets and the undulators associated with synchrotrons. Our work has used a succession of continually improved laser-produced plasma sources. Work is also being done on the development of discharge sources that might be able to provide adequate power in the desired wavelength range. Eventually a source will be required that reliably provides sufficient power to yield adequate wafer throughput in a manufacturing tool.

Resists

The main problem to be confronted in developing a satisfactory photoresist for EUVL is the strong absorption of EUV radiation by all materials. The absorption depth in standard organic resists used today is less than 100 nm. EUV resists will most likely be structured so that printing occurs in a very thin imaging layer at the surface of the resist. Resist types being actively worked on include silylated single-layer resists, refractory bi-layer resists, and tri-layer resists. A resist acceptable for high volume manufacture must exhibit high contrast for printing in combination with a sensitivity that will yield an acceptable throughput. A resist sensitivity of 10 mJ/cm^2 is our goal since it represents a good compromise between the need for high throughput and the desire to minimize the statistical fluctuations due to photon shot noise. Of course, a successful resist must also possess excellent etch resistance. As the features printed in resist have continued to shrink, the roughness at the edges of resist lines has begun to be a serious problem for all lithographies. While not strictly an EUVL problem, a successful EUV resist will be required to solve the line-edge roughness (LER) problem.

Experimental Results

Our imaging experiments to date have been carried out using the 10X EUVL microstepper. These experiments have allowed us to evaluate the EUV imaging performance of the camera and to relate it to the measured surface figure and surface roughness of its mirrors. The imaging performance also correlated well with the camera wavefront as measured directly with the EUV interferometer. Additionally, these experiments have been used to investigate various resists and masks

and to help us understand a number of system issues. Three cameras have been built for this system, all of which image with a 10X reduction. The camera itself is a simple Schwarzschild design and is comprised of two spherical mirrors. A schematic diagram of this camera is shown in Figure 5. As shown in the lower part of the figure, we used off-axis portions of the full mirrors to avoid obscuration of the light by the mirrors; the NA used was 0.07 or 0.08.

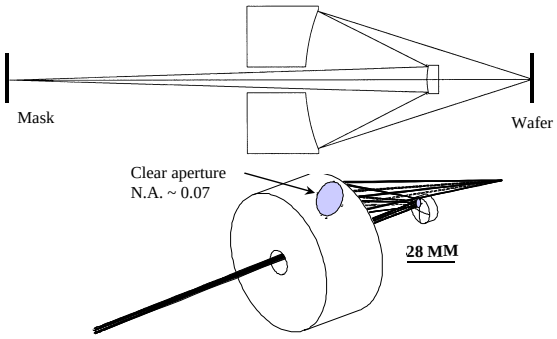


Figure 5: Schematic of the 10X EUVL camera

The cameras were originally aligned using visible interferometry. Subsequent EUV interferometry revealed that the at-wavelength measurements yielded nearly identical results. Not all camera designs allow for alignment with visible light.

Figure 6 shows the cross-sectioned profiles of dense lines and spaces printed in resist with the 10X camera. The figure shows resist profiles of lines and spaces with widths of 200 nm, 150 nm, and 100 nm. As can be seen, the resist profiles are well defined. From a series of measurements like this it is possible to demonstrate the excellent linearity of the printing.

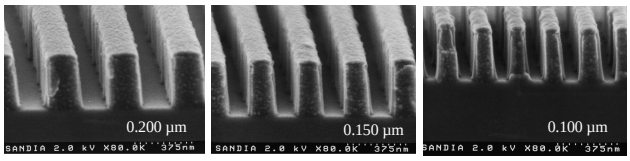


Figure 6: Resist profiles of line and space patterns imaged by the 10X camera for line and space widths of 200 nm, 150 nm, and 100 nm

That is, the width of the resist image is equal to the intended size as written on the mask. Figure 7 demonstrates excellent linearity for dense lines and spaces from a line width of 250 nm down to 80 nm.

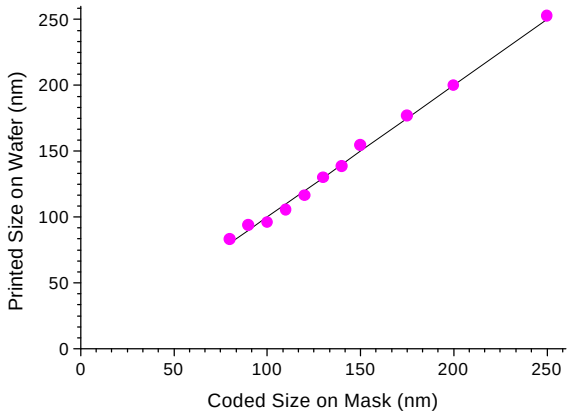


Figure 7: Linearity of printing by the 10X camera in resist for line and space patterns with linewidths from 200 nm down to 80 nm

Exposures such as the above can also be used to demonstrate the large DOF inherent in EUVL. Figure 8 presents the data from such a series of exposures: it shows how the line width of a 130 nm line (the remaining resist) varies as the camera image is defocused on the wafer. As seen, the line width only changes by about 5% as the wafer is moved from best focus to a position 2 µm away from best focus. This observation is in reasonable agreement with the behavior predicted by Equation 1. In manufacturing of high-performance IC's, it is desired to control the critical line widths to +/- 10% or better.

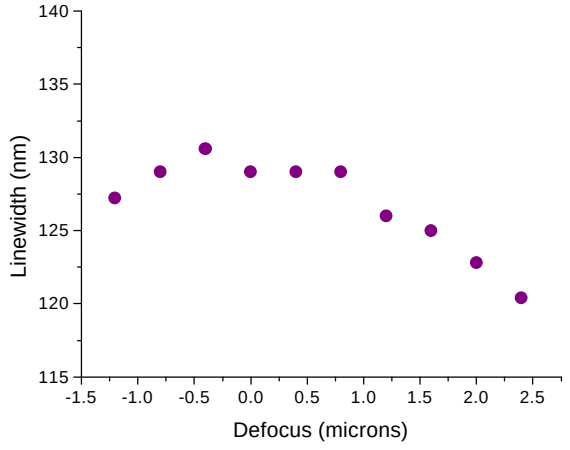


Figure 8: Variation in the size of 130 nm dense lines as a function of defocus; the feature size varies by only 5% as the wafer is defocused by 2 µm

Finally, in Figure 9, we show cross-sectioned resist images of 80 nm lines and spaces (with a line space ratio of 1:2). This demonstrates the resolving power of the

10X camera and our ability to print such fine features in resist.

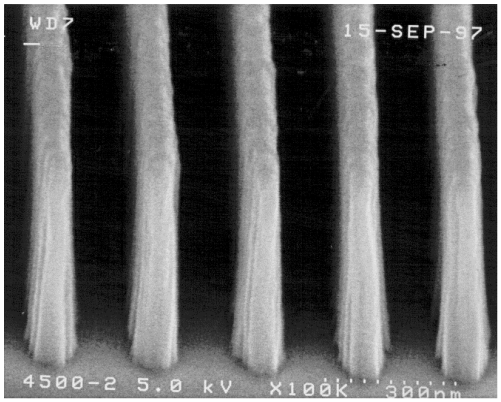


Figure 9: Printing of 80 nm lines and spaces (with a 1:2 pitch) by the 10X camera

While the 10X camera has been of great use in our program, we look forward to the completion of the ETS camera so that we can explore EUV imaging with a camera of the kind needed for production-type lithography.

Conclusion

Successful implementation of EUVL would enable projection photolithography to remain the semiconductor industry's patterning technology of choice for years to come. However, much work remains to be done in order to determine whether or not EUVL will ever be ready for the production line. Furthermore, the time scale during which EUVL, and in fact any NGL technology, has to prove itself is somewhat uncertain. Several years ago, it was assumed that an NGL would be needed by around 2005 in order to implement the 0.1 μm generation of chips. Currently, industry consensus is that 193 nm lithography will have to do the job, even though it will be difficult to do so. There has recently emerged talk of using light at 157 nm to push the current optical technology even further, which would further postpone the entry point for an NGL technology. It thus becomes crucial for any potential NGL to be able to address the printing of feature sizes of 50 nm and smaller! EUVL does have that capability.

The battle to develop the technology that will become the successor to 193 nm lithography is heating up, and it should be interesting to watch!

Acknowledgments

The work described in this paper was done by a large number of people, too numerous to mention individually, within the VNL and the LLC. I am indebted to all of them for allowing me to play a small role in the overall effort and to summarize their work here.

References

[1] For readers interested in digging deeper, I recommend the following sources:

For a compilation of papers on EUVL see "OSA Trends in Optics and Photonics Vol. 4," *Extreme Ultraviolet Lithography*, G.D. Kubiak and D.R. Kania, eds. (Optical Society of America, Washington, DC 1996).

For recent papers on the various NGL's and on optical lithography see J. Vac. Sci. Technol. **B15**, Nov./Dec. 1997.

[2] T.E. Jewell, J.M. Rodgers, and K.P. Thompson, J. Vac. Sci. Technol. **B8**, 1509 (1990).

[3] G.E. Sommargren, Laser Focus World **32**, 61 (1996).

[4] E. Tejnil, et al., J. Vac. Sci. Technol. **B15**, 2455 (1997).

Author's Biography

John E. Bjorkholm has been a Principal Scientist at Intel Corporation in Santa Clara, CA since 1996. In 1994, after 28 years at Bell Laboratories, he retired as a Distinguished Member of the Technical Staff. He was a key player there in the work on laser cooling and trapping of atoms. He received a BSE with highest honors in EE-Physics from Princeton University in 1961, and in 1966 he received a Ph.D in Applied Physics from Stanford University. He is a Fellow of the OSA and the APS. He served as an OSA Director-at-Large (1988-90) and as the OSA Treasurer (1992-96). He also served as a Trustee of Princeton University (1991-95). His email address is John.Bjorkholm@intel.com.