

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.

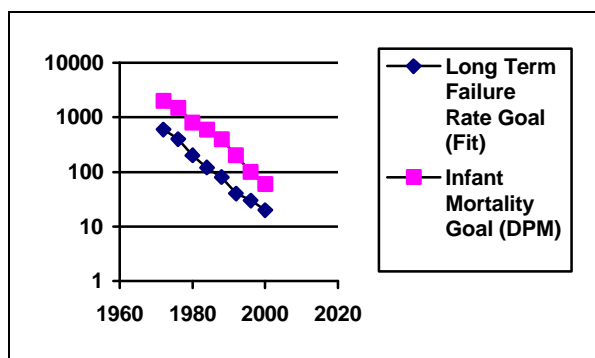


Figure 1: Failure rate (FIT) & defect rate (DPM) goals (the top curve represents infant mortality goals, which can only be achieved by reducing defects)

In this paper we discuss four of the main topics pertaining to the maintenance of reliability for the $0.25\mu\text{m}$ 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 $0.25\mu\text{m}$ process generation are as large as 800 mils on the side and have in excess of seven million transistors.

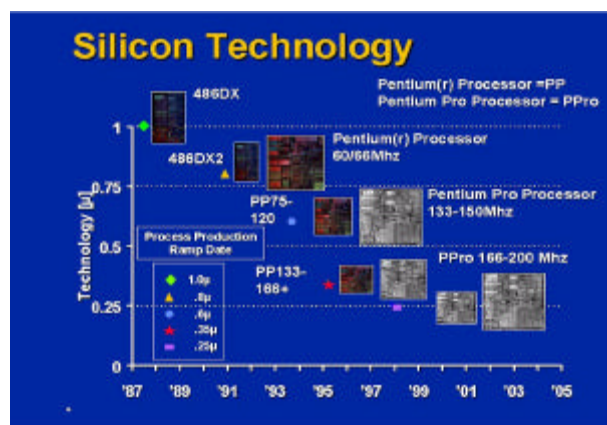


Figure 2: The continued growth of the microprocessor despite the increase in transistor density

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 (mil²/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.

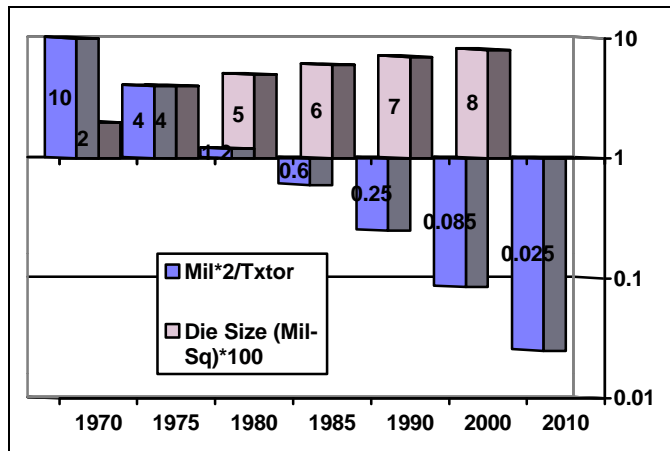


Figure 3: Generational graph of area per transistor and die size trends (increase in functionality contributes to the increase in die size)

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 <i>K</i> >1	Ideal Scaling	Reliability Implications
Channel Length <i>L</i> and shallow junctions	1/ <i>K</i>	Latchup Hot-electron effects
Gate Oxide Thickness	1/ <i>K</i>	Oxide wearout and ESD protection. Process Charging
Metal line width	1/ <i>K</i>	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	<i>V</i> _{cc} * 1/ <i>K</i>	Hot e and gate oxide reliability are rendered equivalent in the scaled voltage scheme.

Device Current	$1/\sqrt{K} - 1/K$	
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 τ_d	$1/k^{**1.5}$	Main contributor to performance enhancements.
Delay Power x	$1/k^{**3}$	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 0.25 μ m 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 0.25 μ m 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 0.25 μ m 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

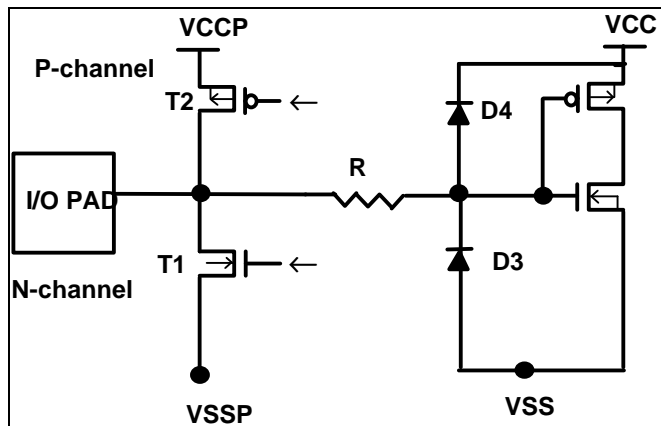


Figure 4: CMOS Input/Output buffer protection (T1 and T2 have built-in parasitic dual diodes that can be enhanced through layout)

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 0.35 μm and 0.25 μm 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 3000 μm) 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.

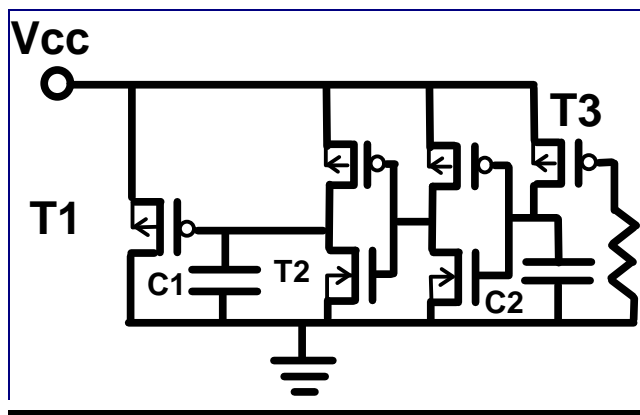


Figure 5: RC-timed circuit for PMOS FET power clamp (pmosclamp); T2 and C1 enhance gate drive

Figure 6 shows pulsed I-V curves for a pmosclamp protection circuit as in Figure 5, occupying about 7700 μm^2 in a 0.25 μm 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 0.35 μm process has roughly the same I-V curve as in Figure 6, and it is in a still-reasonable 12000 μm^2 , but this uses over 50% more area than the 0.25 μm 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 0.8 μm 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.

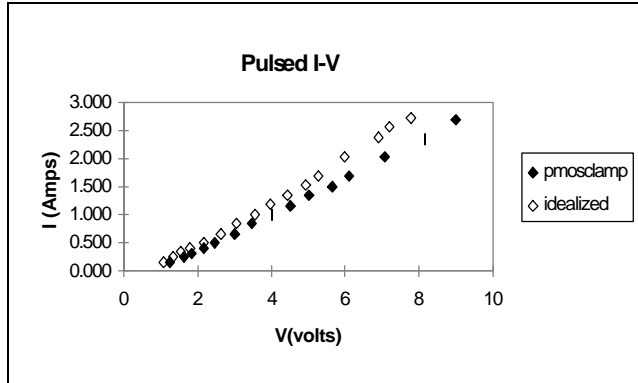


Figure 6: Pulsed I-V behavior of pmosclamp in 0.25µm process; idealized curve has T1 gate artificially grounded

For compatibility with signals from chips with earlier-generation power supply voltage, we want to enable an on-chip power supply V_{ccx} , 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 V_{ccx} to be on when V_{cc} 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 $V_{ccx}/2$ is set by long channel devices T4 and T5. This reference voltage allows only $V_{ccx}/2$ 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%.

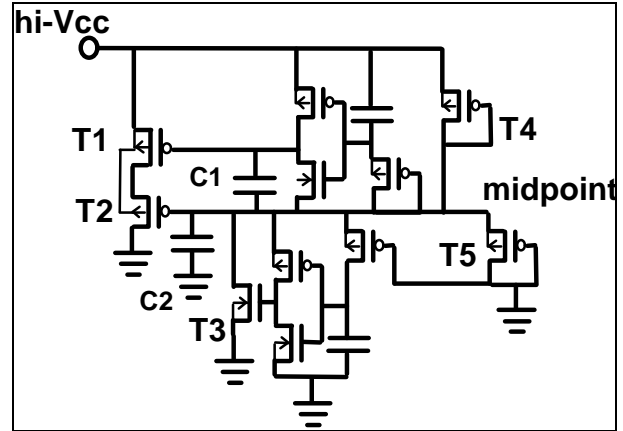


Figure 7: Circuit for stacked-gate high-voltage tolerant PMOS clamp (vtolclamp). T1 and T2 are large FETs built in the same n-well; circuitry drives their gates low temporarily. T4 and T5 bias the midpoint, rendering dc gate oxide voltages safe.

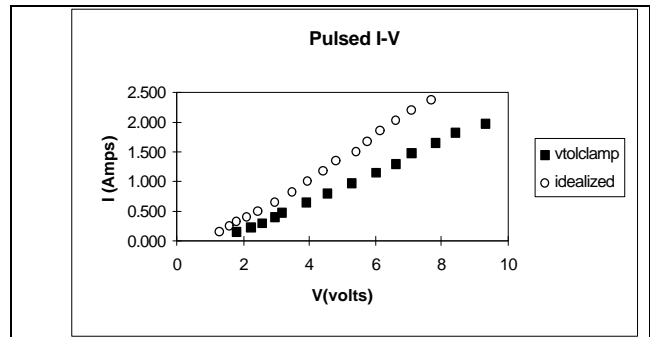


Figure 8: Pulsed I-V of vtolclamp in 0.25µm process; idealized curve has T1 and T2 gates artificially grounded

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 0.35 μm technology [8-9].

Layer	0.25 μm Technology		0.35 μm Technology	
	Pitch (μm)	Thickness (μm)	Pitch (μm)	Thickness (μm)
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 0.80 μm to 0.90 μm . 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 0.60 μm to 0.48 μm 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 0.35 μm process technology. Therefore, during the 0.25 μm 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 0.4 μm 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.

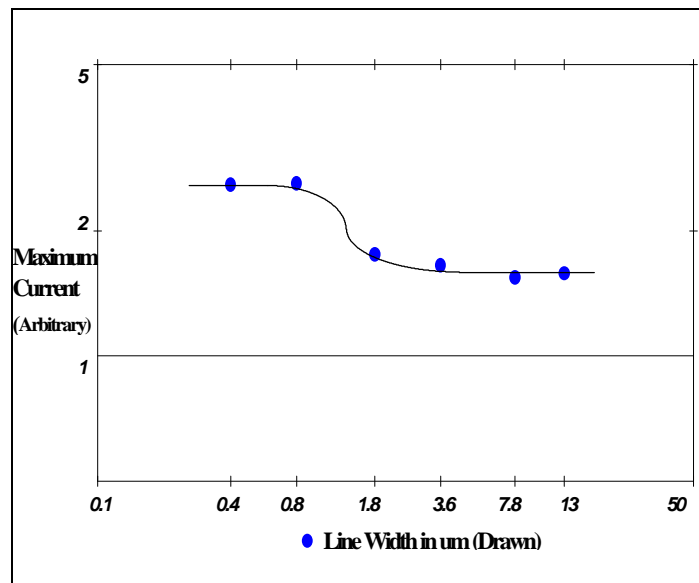


Figure 9: M1 electromigration performance vs. line width

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 0.25 μ m 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 0.25 μ m 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), I_g 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 0.35 μ m and 0.25 μ m Test Chip and products after normalizing to the same electrical field and temperature. Product A in the 0.35 μ m 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 0.25 μ m 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 0.25 μ m pre-mature process was measured on the Test Chip. The result is quite similar to that of the 0.35 μ m 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 0.25 μ m 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 0.25 μ m product failure rate improves 7.2X over that of the 0.35 μ m product failure rate. This improvement has opened the way to additional reductions in gate-oxide thickness, improving process speed.

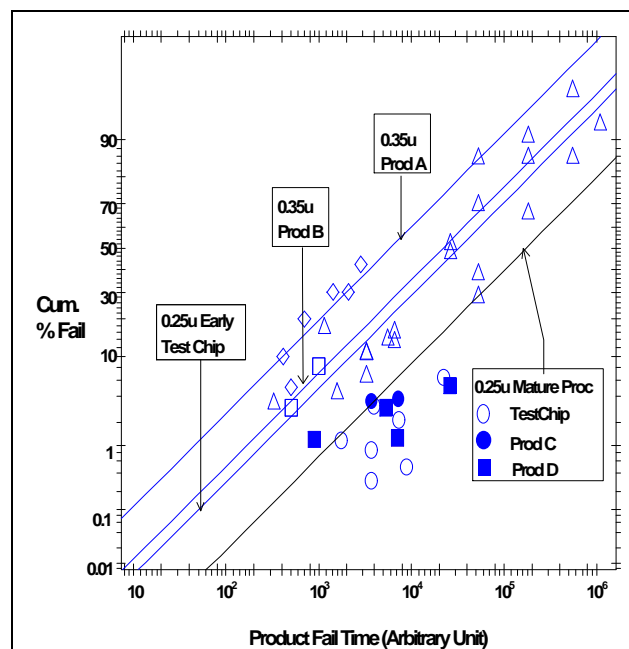


Figure 10: Comparison of 0.35µm and 0.25µm gate oxide breakdown TTF distributions

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 (σ_i) and the externally applied part (σ_e). The total stress is the sum of the two as shown in Eq(1).

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

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 “flipped-chip” is being modeled in this figure). We use this dual-beam approach to estimate stresses in various parts of the final packaged component.

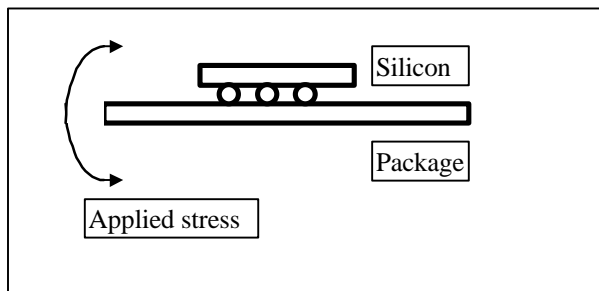


Figure 11: Externally applied stresses on silicon with the chip and package viewed as two independent beams

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.

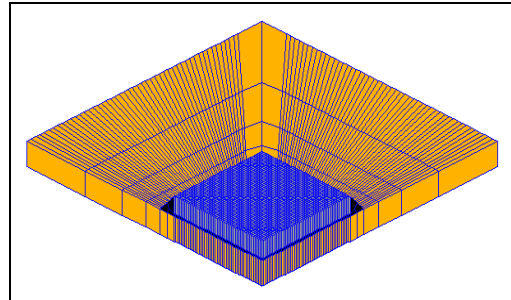


Figure12a: Global model showing a quarter-slice of a die (in gray) mounted on a plastic package (yellow)

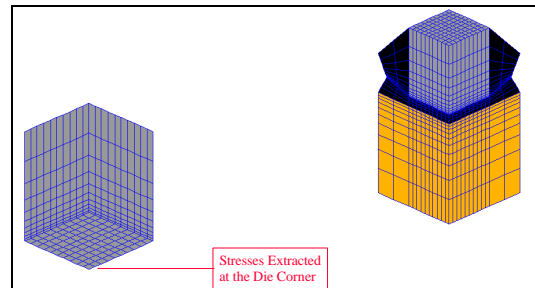


Figure 12b: Stress extracted at the die corner

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.

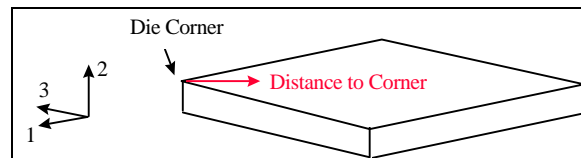


Figure 13a: Stress components

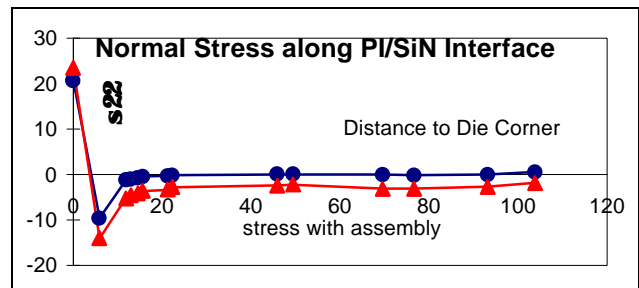


Figure 13b: Result of calculations showing that the stresses are altered after assembly

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 event. 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 $\sim 10kT$, 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 0.25 μ m 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 0.35 μ m 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 0.25 μ m 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 μ m 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 0.25 μ m Microprocessor Technology Reliability Program Manager. His email address is kenneth.j.wu@intel.com.