Semiconductor Technology from A to Z

Reliability

ESD Protection Structures

ESD Models and Damage Mechanisms

Electrostatic discharge (ESD) describes the abrupt equalization of an electrostatic charge between two objects at different potentials. When a charged human body, a component, or a machine touches an integrated circuit, a current of several amperes flows through the connection pins into the device for a few nanoseconds to microseconds. This brief but high energy density can break down gate oxides, thermally destroy p-n junctions, or melt metallization traces.

Three models have become established for evaluating ESD robustness, each reproducing a different discharge source, and they differ markedly in pulse shape, rise time, and peak current:

Model	Equivalent circuit	Rise time	Pulse duration	Typical peak current
HBM (Human Body Model)	100 pF through 1.5 kΩ	2–10 ns	approx. 150 ns	approx. 1.3 A at 2 kV
MM (Machine Model)	200 pF, no series resistance	<1 ns (ringing)	approx. 150–300 ns	several A at 200 V
CDM (Charged Device Model)	device self-capacitance (a few pF)	<1 ns	<1 ns	up to 10–15 A at 500 V

Qualification of a device typically requires minimum robustness levels, such as 2 kV under the HBM model and 500 V to 1 kV under the CDM model for consumer-oriented applications; automotive devices often demand considerably higher levels. Because the entire charge stored in the component discharges within under a nanosecond under CDM, this model produces by far the highest current peaks despite the lower voltage, and the most demanding requirements on the response speed of the protection structure.

Protection Diodes and Snapback Behavior

The simplest ESD protection structure consists of two clamping diodes at every input or output pad: one blocks toward the supply voltage V_{DD}, the other toward ground V_{SS}. During a positive ESD pulse, the upper diode conducts and diverts the current into the V_{DD} rail; during a negative pulse, the lower diode takes over the discharge path to ground. The clamping voltage of the diodes limits the voltage that reaches the gate oxide of the downstream core circuit to a non-critical level – for a typical silicon diode the forward voltage is about 0.7 V, so the voltage at the protected node is clamped to roughly V_{DD} + 0.7 V or V_{SS} –

0.7 V.

For higher current robustness, snapback structures are frequently used instead of simple diodes. Their characteristic corresponds to a parasitic thyristor (Silicon Controlled Rectifier, SCR). Once the voltage across the device exceeds the breakdown voltage – typically in the range of 8 to 15 V in modern CMOS technologies –, a parasitic npn bipolar transistor turns on, whose base current in turn activates a parasitic pnp transistor – the structure latches into a low-impedance state with a significantly reduced holding voltage of often only 1 to 2 V. This snapback behavior allows large ESD currents to be discharged with low power dissipation, but requires careful tuning of the holding voltage above the supply voltage to avoid unintended latchup during normal operation.

Grounded-Gate NMOS and Power Clamps

A widely used snapback protection structure is the grounded-gate NMOS (GGNMOS), in which the gate and source of an NMOS transistor are permanently tied to ground while the drain is connected to the pad to be protected. An ESD pulse drives the drain-bulk junction into avalanche breakdown; the resulting holes flow through the bulk region and raise its potential until the parasitic npn bipolar transistor between source, bulk, and drain turns on and discharges the ESD current at low impedance. The maximum current-carrying capability of a GGNMOS structure before thermal failure (failure current I_{t2}) scales roughly with the finger width of the transistor and is typically on the order of a few milliamperes per micrometer of gate width.

To protect the internal supply rails, a power clamp is additionally placed between V_{DD} and V_{SS}. It remains high-impedance during normal operation and switches to a low-impedance state only for the duration of a fast ESD voltage rise, detected through an RC network with a typical time constant of 0.1 to 1 µs, tuned to be much shorter than the millisecond-scale rise time of a regular power-up event. The design of every ESD structure involves a fundamental trade-off: larger protection structures increase current robustness but also increase the parasitic capacitance at the pad and therefore the signal delay, which must be weighed carefully for high-frequency and high-speed interfaces.

Electromigration, TDDB and NBTI/PBTI

Electromigration

Electromigration refers to the transport of material within a metal conductor

caused by the momentum transfer from drifting conduction electrons to the metal ions of the crystal lattice, often called the electron wind. At the high current densities of modern interconnect levels, on the order of several MA/cm², metal atoms preferentially migrate along grain boundaries in the direction of current flow. Material accumulates on the anode side, forming hillocks, while vacancies coalesce into voids on the cathode side, which can narrow the conductor cross-section until it fails completely.

The mean time to failure of an interconnect due to electromigration is described by the empirical Black equation $MTTF = A \cdot J^{-n} \cdot e^{E_a/(k \cdot T)}$, where J is the current density, T the temperature, E_a the activation energy, k the Boltzmann constant, and the current density exponent n is typically close to 2. For aluminum interconnect, the allowable design current density is usually in the range of 1 to 2 MA/cm² with an activation energy of about 0.5 to 0.7 eV along grain boundaries; copper, with an activation energy of roughly 0.7 to 1.1 eV and lower self-diffusion, achieves a much higher electromigration resistance at the same current density. This improvement was one of the key drivers behind the introduction of copper interconnect technology in the 1990s, since it was what first made higher current densities reliably possible as wire cross-sections continued to shrink.

Time-Dependent Dielectric Breakdown (TDDB)

Time-dependent dielectric breakdown (TDDB) describes the delayed electrical breakdown of a gate dielectric under continuous stress well below its intrinsic breakdown voltage. The cause is the continuous generation of defects in the oxide by injected charge carriers, which over time link up into a continuous low-resistance path between gate and channel, as described by the percolation model.

In modern CMOS technologies, the effective gate oxide thickness (equivalent oxide thickness, EOT) is only about 1 to 2 nm, which already produces electric field strengths of 8 to 12 MV/cm in the oxide during normal operation – a regime in which TDDB becomes the dominant lifetime-limiting mechanism. Because the defect density required to trigger breakdown also decreases with oxide thickness, TDDB is one of the central reasons why the maximum allowable operating voltage of modern technologies cannot simply be increased. In qualification testing, lifetime is typically determined through accelerated tests at field strengths of 15 MV/cm and above combined with elevated temperature, and extrapolated to nominal operating conditions using a power law, with required lifetimes of ten years having to be achieved at fields well below the test conditions.

Bias Temperature Instability (NBTI/PBTI)

Bias temperature instability (BTI) refers to the time-dependent shift of a MOS transistor threshold voltage under sustained gate bias at elevated temperature. In PMOS transistors, the effect appears as negative BTI (NBTI) under negative gate-source voltage, while in NMOS transistors it appears as positive BTI (PBTI) under positive gate-source voltage; both manifest as an increase in the magnitude of the threshold voltage and a corresponding reduction in drain current over the operating lifetime.

The generation of interface states at the Si–SiO_2 interface and the trapping of charge carriers within the gate dielectric are considered the main causes, and part of the shift recovers once the gate bias is removed. Typical threshold voltage shifts after ten years of operation are on the order of 30 to 80 mV, depending on technology, operating temperature, and supply voltage; in high-k metal-gate technologies, NBTI usually dominates over PBTI. Because BTI continuously reduces the switching speed of logic paths over the lifetime of a chip, the effect is accounted for during timing signoff through an aging guardband, often several percent, applied to the threshold voltage or gate delay.

Burn-in and HTOL Testing

The Bathtub Curve and the Purpose of Burn-in

The failure rate of electronic components over operating time typically follows the so-called bathtub curve: early in the service life, infant mortality failures caused by manufacturing defects dominate and the failure rate decreases; this is followed by a long period of low, roughly constant random failures – often expressed in FIT (Failures In Time, failures per 10^9 device-hours) and typically in the range of a few to a few hundred FIT for qualified semiconductor devices –, before wear-out failures driven by aging mechanisms such as electromigration, TDDB, or BTI cause the failure rate to rise again toward the end of the useful life.

Burn-in refers to a stress test in which components are operated for a defined period, often 24 to 168 hours, under elevated voltage and temperature in order to accelerate through the infant mortality phase of the bathtub curve and screen out defective units before they are shipped to the customer. Because burn-in adds cost and test time, it is mainly applied to applications with high reliability requirements, such as automotive or industrial electronics, while mature consumer products with well-controlled processes are often covered by

statistical sampling instead.

HTOL Test Conditions

Highly Accelerated Temperature and Operating Life (HTOL) testing is a standardized qualification test in which a sample of components – per JEDEC JESD22-A108 often 77 units drawn from at least three different production lots – is stressed for typically 1000 hours at an elevated junction temperature (often 125 °C, or 150 °C for some technologies) and elevated operating voltage under active switching conditions. The units under test are electrically characterized before, during, and after the test to detect parameter drift or functional failures.

To pass the test, it is usually required that no unit fails within the sample (a zero-fail criterion), which for a sample of 77 units statistically corresponds to a failure rate of less than about 0.3 % at 60 % confidence. HTOL primarily covers diffusion-driven and field-dependent aging mechanisms such as electromigration, TDDB, and BTI, complementing ESD and latchup qualification with time-dependent operating reliability.

The Arrhenius Model for Lifetime Prediction

Because testing under nominal conditions for the required field lifetimes of ten years or more is not practical, HTOL tests are run at elevated temperature and the results are extrapolated to the operating temperature using the Arrhenius model. The model describes the rate of a thermally activated aging mechanism as an exponential function $k(T) = A \cdot e^{-E_a/(k \cdot T)}$ of the inverse absolute temperature, weighted by an activation energy E_a characteristic of the respective failure mechanism.

The ratio of failure times at test temperature and operating temperature yields an acceleration factor $AF = e^{(E_a/k) \cdot (1/T_{use} - 1/T_{test})}$. For a typical TDDB mechanism with $E_a \approx 0.7$ eV, a test temperature of 125 °C (398 K), and an assumed operating temperature of 55 °C (328 K), the acceleration factor works out to roughly 40 to 50 – under these assumptions, 1000 test hours therefore correspond to an equivalent field operating time of about 4 to 5 years. Because different failure mechanisms have different activation energies – TDDB and electromigration are typically in the range of 0.6 to 1 eV –, a reliable lifetime prediction requires knowledge of the dominant mechanism.