

Halbleitertechnologie von A bis Z

Zuverlässigkeit

ESD-Schutzstrukturen	3
<i>ESD-Modelle und Schadensmechanismen</i>	3
<i>Schutzdioden und Snapback-Verhalten</i>	3
<i>Grounded-Gate-NMOS und Power-Clamps</i>	4
Elektromigration, TDDB und NBTI/PBTI	5
<i>Elektromigration</i>	5
<i>Time-Dependent Dielectric Breakdown (TDDB)</i>	5
<i>Bias Temperature Instability (NBTI/PBTI)</i>	6
Burn-in und HTOL-Tests	6
<i>Die Badewannenkurve und der Zweck des Burn-in</i>	6
<i>HTOL-Testbedingungen</i>	7
<i>Das Arrhenius-Modell zur Lebensdauerabschätzung</i>	7

ESD-Schutzstrukturen

ESD-Modelle und Schadensmechanismen

Elektrostatische Entladung (ESD, Electrostatic Discharge) beschreibt den schlagartigen Ausgleich einer elektrostatischen Aufladung zwischen zwei Objekten unterschiedlichen Potentials. Berührt ein aufgeladener menschlicher Körper, ein Bauteil oder eine Maschine einen integrierten Schaltkreis, fließt für wenige Nanosekunden bis Mikrosekunden ein Strom von mehreren Ampere über die Anschlusspins in das Bauelement. Diese kurzzeitige, aber hohe Energiedichte kann Gate-Oxide durchschlagen, p-n-Übergänge thermisch zerstören oder Metallisierungsbahnen aufschmelzen.

Zur Bewertung der ESD-Festigkeit haben sich drei Modelle etabliert, die unterschiedliche Entladungsquellen nachbilden und sich in Pulsform, Anstiegszeit und Spitzenstrom deutlich unterscheiden:

Modell	Ersatzschaltung	Anstiegszeit	Pulsdauer	Typ. Spitzenstrom
HBM (Human Body Model)	100 pF über 1,5 k Ω	2-10 ns	ca. 150 ns	ca. 1,3 A bei 2 kV
MM (Machine Model)	200 pF, kein Serienwiderstand	<1 ns (oszillierend)	ca. 150-300 ns	mehrere A bei 200 V
CDM (Charged Device Model)	Bauteil-Eigenkapazität (wenige pF)	<1 ns	<1 ns	bis 10-15 A bei 500 V

Für die Qualifikation eines Bauelements werden üblicherweise Mindestfestigkeiten gefordert, etwa 2 kV nach dem HBM-Modell und 500 V bis 1 kV nach dem CDM-Modell für konsumentennahe Anwendungen; Automotive-Bauelemente verlangen häufig deutlich höhere Werte. Da beim CDM die gesamte im Bauteil gespeicherte Ladung innerhalb von unter einer Nanosekunde abfließt, entstehen hier trotz geringerer Spannung die mit Abstand höchsten Stromspitzen und die anspruchsvollsten Schutzanforderungen an die Reaktionsgeschwindigkeit der Schutzstruktur.

Schutzdioden und Snapback-Verhalten

Die einfachste ESD-Schutzstruktur besteht aus zwei Klemmdioden an jedem Eingangs- oder Ausgangspad: Eine sperrt gegen die Versorgungsspannung V_{DD} ,

die andere gegen Masse V_{SS} . Bei einem positiven ESD-Puls leitet die obere Diode und leitet den Strom in die V_{DD} -Schiene ab, bei einem negativen Puls übernimmt die untere Diode den Ableitpfad nach Masse. Die Klemmspannung der Dioden begrenzt dabei die am Gate-Oxid der nachfolgenden Kernschaltung anliegende Spannung auf einen unkritischen Wert – bei einer typischen Siliciumdiode liegt die Flussspannung bei etwa 0,7 V, so dass die Spannung am geschützten Knoten auf etwa $V_{DD} + 0,7$ V beziehungsweise $V_{SS} - 0,7$ V begrenzt wird.

Für höhere Stromfestigkeit werden anstelle einfacher Dioden häufig Snapback-Strukturen eingesetzt, deren Kennlinie einem parasitären Thyristor (Silicon Controlled Rectifier, SCR) entspricht. Überschreitet die Spannung am Bauelement die Durchbruchspannung – typischerweise im Bereich von 8 bis 15 V bei modernen CMOS-Technologien –, zündet ein parasitärer npn-Bipolartransistor, dessen Basisstrom wiederum einen parasitären pnp-Transistor aktiviert – die Struktur rastet in einen niederohmigen Zustand mit einer deutlich reduzierten Haltespannung von oft nur 1 bis 2 V ein. Dieses Snapback-Verhalten ermöglicht das Ableiten hoher ESD-Ströme bei geringer Verlustleistung, erfordert aber eine sorgfältige Abstimmung der Haltespannung oberhalb der Versorgungsspannung, um ungewolltes Latchup im Normalbetrieb zu vermeiden.

Grounded-Gate-NMOS und Power-Clamps

Eine weit verbreitete Snapback-Schutzstruktur ist der Grounded-Gate-NMOS (GGNMOS), bei dem Gate und Source eines NMOS-Transistors fest mit Masse verbunden sind, während der Drain-Anschluss am zu schützenden Pad liegt. Ein ESD-Puls treibt den Drain-Bulk-Übergang in den Avalanche-Durchbruch; die dabei erzeugten Löcher fließen über das Bulk-Gebiet ab und heben dessen Potential an, bis der parasitäre npn-Bipolartransistor zwischen Source, Bulk und Drain zündet und den ESD-Strom niederohmig ableitet. Die maximale Stromtragfähigkeit einer GGNMOS-Struktur vor thermischer Zerstörung (Failure Current I_{t2}) skaliert näherungsweise mit der Fingerweite des Transistors und liegt bei typischen Technologien im Bereich weniger Milliampere pro Mikrometer Gateweite.

Zum Schutz der internen Versorgungsschienen wird zusätzlich ein Power-Clamp zwischen V_{DD} und V_{SS} eingesetzt, der im Normalbetrieb hochohmig ist und erst bei einem schnellen ESD-Spannungsanstieg – erkannt über ein RC-Glied mit einer typischen Zeitkonstante von 0,1 bis 1 μ s, abgestimmt auf die deutlich langsamere Anstiegszeit regulärer Power-up-Vorgänge im Millisekundenbereich – für die Dauer des Pulses niederohmig durchschaltet. Beim Design aller ESD-Strukturen steht ein Zielkonflikt im Vordergrund: Größere Schutzstrukturen erhöhen die Stromfestigkeit, vergrößern aber die parasitäre Kapazität am Pad

und damit die Signallaufzeit, was besonders bei Hochfrequenz- und Hochgeschwindigkeitsschnittstellen kritisch abgewogen werden muss.

Elektromigration, TDDB und NBTI/PBTI

Elektromigration

Elektromigration bezeichnet den Materialtransport in einem Metalleiter, der durch den Impulsübertrag driftender Leitungselektronen auf die Metallionen des Kristallgitters verursacht wird (Elektronenwind). Bei den hohen Stromdichten moderner Verdrahtungsebenen von mehreren MA/cm² wandern Metallatome bevorzugt entlang von Korngrenzen in Stromrichtung, wodurch sich auf der Anodenseite Material anhäuft (Hillocks) und auf der Kathodenseite Leerstellen zu Hohlräumen (Voids) koaleszieren, die den Leiterquerschnitt bis zur vollständigen Unterbrechung verengen können.

Die mittlere Lebensdauer einer Leiterbahn bis zum elektromigrationsbedingten Ausfall wird durch die empirische Black-Gleichung $MTTF = A \cdot J^{-n} \cdot e^{E_a/(k \cdot T)}$ beschrieben, wobei J die Stromdichte, T die Temperatur, E_a die Aktivierungsenergie, k die Boltzmann-Konstante und der Stromdichteexponent n typischerweise nahe 2 liegt. Für Aluminiumverdrahtung liegt die zulässige Designstromdichte meist im Bereich von 1 bis 2 MA/cm² bei einer Aktivierungsenergie von etwa 0,5 bis 0,7 eV entlang der Korngrenzen; Kupfer erreicht dank einer Aktivierungsenergie von rund 0,7 bis 1,1 eV und einer geringeren Selbstdiffusion eine um ein Vielfaches höhere Elektromigrationsfestigkeit bei gleicher Stromdichte. Diese Verbesserung war einer der wesentlichen Treiber für die Einführung der Kupfertechnologie in den 1990er-Jahren, da sie höhere Stromdichten bei gleichzeitig sinkenden Leiterbahnquerschnitten überhaupt erst zuverlässig ermöglichte.

Time-Dependent Dielectric Breakdown (TDDB)

Time-Dependent Dielectric Breakdown (TDDB) beschreibt den zeitverzögerten elektrischen Durchbruch eines Gate-Dielektrikums unter einer Dauerbelastung, die deutlich unterhalb der intrinsischen Durchbruchspannung liegt. Ursache ist die kontinuierliche Erzeugung von Defekten im Oxid durch injizierte Ladungsträger, die sich mit der Zeit zu einem durchgehenden Pfad geringen Widerstands zwischen Gate und Kanal verketteten (Perkolationsmodell).

Bei modernen CMOS-Technologien liegt die effektive Gate-Oxiddicke (Equivalent Oxide Thickness, EOT) nur noch bei etwa 1 bis 2 nm, was im

Normalbetrieb bereits zu elektrischen Feldstärken von 8 bis 12 MV/cm im Oxid führt – ein Bereich, in dem TDDB zum dominanten Lebensdauer-limitierenden Mechanismus wird. Da die für einen Durchbruch nötige Defektdichte mit sinkender Oxiddicke ebenfalls abnimmt, ist TDDB einer der zentralen Gründe, warum die maximal zulässige Betriebsspannung moderner Technologien nicht beliebig erhöht werden kann. In der Qualifikation wird die Lebensdauer typischerweise durch beschleunigte Tests bei Feldstärken von 15 MV/cm und mehr sowie erhöhter Temperatur ermittelt und über ein Potenzgesetz auf die Nennbetriebsbedingungen extrapoliert, wobei geforderte Lebensdauern von zehn Jahren bei Feldern deutlich unterhalb der Testbedingungen erreicht werden müssen.

Bias Temperature Instability (NBTI/PBTI)

Bias Temperature Instability (BTI) bezeichnet die zeitabhängige Verschiebung der Schwellenspannung eines MOS-Transistors unter dauerhafter Gate-Spannungsbelastung bei erhöhter Temperatur. Bei PMOS-Transistoren tritt der Effekt als Negative BTI (NBTI) unter negativer Gate-Source-Spannung auf, bei NMOS-Transistoren als Positive BTI (PBTI) unter positiver Gate-Source-Spannung; beide äußern sich in einer betragsmäßigen Zunahme der Schwellenspannung und damit einer Reduktion des Drain-Stroms im Laufe der Betriebszeit.

Als Ursache gelten die Erzeugung von Grenzflächenzuständen an der Si-SiO₂-Grenzfläche sowie das Einfangen von Ladungsträgern in Traps innerhalb des Gate-Dielektrikums, wobei sich ein Teil der Verschiebung bei Entlastung der Gate-Spannung wieder zurückbildet (Recovery). Typische Schwellenspannungsverschiebungen liegen nach zehn Jahren Betrieb bei etwa 30 bis 80 mV, abhängig von Technologie, Betriebstemperatur und Versorgungsspannung; bei High-k-Metal-Gate-Technologien fällt NBTI in der Regel stärker ins Gewicht als PBTI. Da BTI die Schaltgeschwindigkeit von Logikpfaden über die Lebensdauer eines Chips kontinuierlich verringert, wird der Effekt beim Timing-Signoff durch einen Alterungs-Guardband von oft mehreren Prozent auf die Schwellenspannung beziehungsweise Gatterverzögerung berücksichtigt.

Burn-in und HTOL-Tests

Die Badewannenkurve und der Zweck des Burn-in

Die Ausfallrate elektronischer Bauelemente über der Betriebszeit folgt typischerweise der sogenannten Badewannenkurve: Zu Beginn der Nutzungsdauer treten gehäuft Frühausfälle durch Fertigungsdefekte auf (fallende Ausfallrate), es folgt eine lange Phase konstant niedriger Zufallsausfälle – häufig angegeben in FIT (Failures In Time, Ausfälle pro 10^9 Bauteilstunden) und typischerweise im Bereich weniger bis einiger hundert FIT für qualifizierte Halbleiterbauelemente –, bevor gegen Ende der Lebensdauer Verschleißausfälle durch Alterungsmechanismen wie Elektromigration, TDDB oder BTI die Ausfallrate wieder ansteigen lassen.

Burn-in bezeichnet einen Stresstest, bei dem Bauelemente für einen definierten Zeitraum, häufig 24 bis 168 Stunden, unter erhöhter Spannung und Temperatur betrieben werden, um die Frühausfallphase der Badewannenkurve vorwegzunehmen und defektbehaftete Bauteile vor der Auslieferung an den Kunden auszusortieren. Da Burn-in Kosten und Testzeit verursacht, wird er meist gezielt für Anwendungen mit hohen Zuverlässigkeitsanforderungen wie Automotive- oder Industrieelektronik eingesetzt, während er bei Consumer-Produkten mit ausgereiften, gut beherrschten Prozessen häufig durch statistische Stichprobenverfahren ersetzt wird.

HTOL-Testbedingungen

Highly Accelerated Temperature and Operating Life (HTOL) ist ein standardisierter Qualifikationstest, bei dem eine Stichprobe von Bauelementen – nach JEDEC JESD22-A108 häufig 77 Bauteile aus mindestens drei unterschiedlichen Fertigungslosen – über einen Zeitraum von typischerweise 1000 Stunden bei erhöhter Sperrschichttemperatur (häufig 125 °C, bei manchen Technologien auch 150 °C) und erhöhter Betriebsspannung im aktiven Schaltbetrieb belastet wird. Die Prüflinge werden vor, während und nach dem Test elektrisch charakterisiert, um Parameterdrifts oder Funktionsausfälle zu erkennen.

Für ein Bestehen des Tests wird üblicherweise gefordert, dass innerhalb der Stichprobe kein Bauteil ausfällt (Zero-Fail-Kriterium), was bei einer Stichprobe von 77 Einheiten statistisch einer Ausfallrate von unter etwa 0,3 % bei 60 % Konfidenzniveau entspricht. HTOL deckt in erster Linie diffusionsgetriebene und feldabhängige Alterungsmechanismen wie Elektromigration, TDDB und BTI ab und ergänzt damit ESD- und Latchup-Qualifikation um die zeitabhängige Betriebszuverlässigkeit.

Das Arrhenius-Modell zur Lebensdauerabschätzung

Da ein Test unter Nennbedingungen für die geforderten Feldlebensdauern von zehn Jahren und mehr nicht praktikabel wäre, werden HTOL-Tests bei erhöhter Temperatur durchgeführt und die Ergebnisse mithilfe des Arrhenius-Modells auf die Betriebstemperatur extrapoliert. Das Modell beschreibt die Geschwindigkeit eines thermisch aktivierten Alterungsmechanismus als Exponentialfunktion $k(T) = A \cdot e^{-E_a/(k \cdot T)}$ des Kehrwerts der absoluten Temperatur, gewichtet mit einer für den jeweiligen Ausfallmechanismus charakteristischen Aktivierungsenergie E_a .

Aus dem Verhältnis der Ausfallzeiten bei Testtemperatur und Betriebstemperatur ergibt sich ein Beschleunigungsfaktor $AF = e^{(E_a/k) \cdot (1/T_{use} - 1/T_{test})}$. Für einen typischen TDDB-Mechanismus mit $E_a \approx 0,7$ eV, einer Testtemperatur von 125 °C (398 K) und einer angenommenen Betriebstemperatur von 55 °C (328 K) ergibt sich ein Beschleunigungsfaktor von etwa 40 bis 50 – 1000 Teststunden entsprechen unter diesen Annahmen also einer äquivalenten Feldebetriebszeit von rund 4 bis 5 Jahren. Da unterschiedliche Ausfallmechanismen unterschiedliche Aktivierungsenergien besitzen – TDDB und Elektromigration liegen typischerweise im Bereich von 0,6 bis 1 eV –, muss für eine belastbare Lebensdauervorhersage der jeweils dominante Mechanismus bekannt sein.