Semiconductor Technology from A to Z

Digital technology

Design Rules

Introduction & Motivation

Why Design Rules?

Every integrated circuit exists as a geometric layout: a stack of overlapping layers of diffusion, polysilicon, contacts, and multiple metal layers that together form transistors and their interconnects. For this layout to be manufactured correctly and reliably on the wafer, it must obey a set of geometric constraints known as design rules. They translate the physical limits of a fabrication process into concrete requirements for chip design.

These limits arise from several sources at once. Lithography can only image structures cleanly down to a certain resolution; below that, edges blur or lines break. Masks cannot be aligned to each other with perfect precision during exposure, so a safety margin for misalignment must be built in between successive layers. Etch and deposition processes vary from wafer to wafer and across the wafer surface, so minimum spacings must also guard against electrical shorts caused by process variation. On top of this come effects such as latchup in CMOS well structures or charge damage from the so-called antenna effect, which require additional, electrically motivated rules.

Design rules are therefore always a trade-off: generous spacings improve manufacturing yield and reliability but cost chip area and thus money. Tight, aggressive rules allow higher packing density but reduce yield and increase the risk of fabrication defects. Each foundry publishes a complete rule set for its specific process, which is checked automatically against the finished layout by a Design Rule Check (DRC) tool before a mask is allowed to be exposed.

Basic Rule Types

Basic Rule Types

Although individual rule sets differ substantially between foundries and technology nodes, nearly all design rules can be traced back to a small number of geometric primitives. These are usually defined either on a single layer (one-layer rule) or between two layers (two-layer rule):

Width specifies the minimum width a structure on a given layer may have — for example a polysilicon gate line or a metal trace. Below this width, lithographic and etch fidelity is no longer guaranteed; the structure could break or become

electrically too resistive.

Spacing describes the minimum gap between two structures on the same layer. It prevents neighboring traces from merging or shorting due to process variation, while also accounting for the optical resolution limit of the exposure.

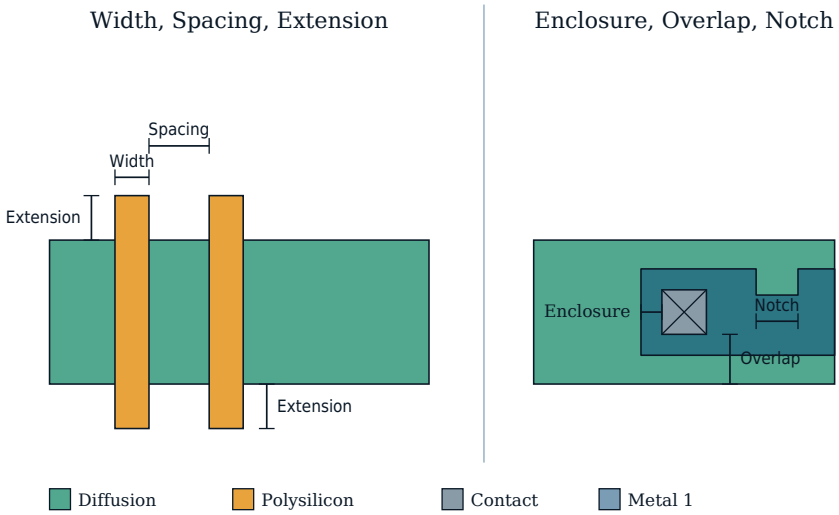
Enclosure requires one layer to surround another by a minimum margin — classically at a contact: the metal must extend past the underlying contact on all sides by a fixed amount, so that misalignment cannot leave an exposed, uncontacted edge.

Overlap is closely related to enclosure but is expressed from the lower layer's perspective: diffusion must overlap the contact so that it lands entirely on conductive material.

Extension requires a structure to project beyond the edge of another — for instance the polysilicon gate, which must extend past the diffusion edge on both sides so that source and drain remain reliably separated during ion implantation, even with slight misalignment.

Notch limits the minimum size of a re-entrant corner or narrow indentation in a structure. Very sharp or narrow notches cannot be imaged crisply by lithography and round off in the real process.

The diagram below illustrates these six basic concepts on a single, simplified layout excerpt: a polysilicon gate over a diffusion area with one contact and the overlying first metal layer.



Lambda-based vs. Absolute Rules

Scalable Rules: the Lambda Concept

In the early 1980s, as integrated circuits first began to be designed on a large scale outside individual semiconductor manufacturers, the Mead-Conway approach introduced an influential concept: instead of specifying every design rule in absolute units such as micrometers, all spacings and widths were expressed as multiples of a single reference quantity called lambda (λ). A typical rule might read, for example, "minimum poly width = 2λ" or "minimum diffusion-to-contact spacing = 3λ".

The advantage of this approach lay in portability: a layout designed entirely in lambda units could, in principle, be ported to a different fabrication process with finer or coarser resolution simply by redefining the numeric value of λ — without adjusting any geometry in the layout individually. For teaching and academic designs, where circuits were often fabricated at multiple foundries (for instance through multi-project-wafer programs), this offered a substantial practical benefit.

Why Modern Processes Use Absolute Rules

As technology nodes continued to scale, however, the lambda model ran into its limits. Not all structures in a modern process scale uniformly: transistor gates typically shrink faster than contact spacings, and metal layers at different levels often have entirely different minimum widths and spacings depending on their role in the interconnect hierarchy. A single scaling factor can no longer capture these asymmetries.

On top of this come physical effects that do not scale linearly with feature size — such as antenna ratios, latchup spacings, or density requirements for chemical-mechanical planarization (CMP). For these reasons, modern foundries publish their design rules almost universally in absolute units (micrometers or nanometers), each with its own rule set often comprising several hundred individual rules per process variant. The lambda concept remains valuable for teaching the underlying principle of design rules, but plays essentially no role in commercial chip development today.

Example Rule Set

A Typical, Simplified Rule Set

To put the concepts introduced in Section 2 into concrete numbers, the table

below shows a simplified, illustrative rule set typical of a generic single-poly, multi-metal sub-micron CMOS process. The values are modeled on common teaching design kits and are not the rules of any specific real foundry process — real rule sets comprise several hundred individual rules depending on the technology node and naturally differ from manufacturer to manufacturer.

Layer / Rule	Meaning	Minimum Value
Poly Width	min. width of a poly gate line	0.35 µm
Poly Spacing	min. spacing between two poly lines	0.35 µm
Poly Extension	gate overhang past diffusion edge	0.25 µm
Diffusion Width	min. width of a diffusion area	0.4 µm
Diffusion Spacing	min. spacing between two same-type diffusion areas	0.5 µm
Contact Size	fixed contact hole size	0.4 × 0.4 µm
Contact Spacing	min. spacing between two contact holes	0.4 µm
Diffusion Overlap of Contact	diffusion margin around contact	0.15 µm
Metal-1 Width	min. width of metal-1	0.5 µm
Metal-1 Spacing	min. spacing between two metal-1 traces	0.5 µm
Metal-1 Enclosure of Contact	metal-1 margin around contact	0.15 µm
Via-1 Size	fixed via size (metal-1 to metal-2)	0.4 × 0.4 µm
Metal-2 Width / Spacing	min. width / spacing of metal-2	0.5 / 0.5 µm

Even in this highly simplified excerpt, a typical pattern emerges: contacts and vias are usually square and fixed in size (not freely scalable), while enclosing layers such as diffusion and metal must maintain an additional safety margin. The enclosure values are deliberately smaller than the width and spacing values of their respective layer — they primarily compensate for mask-to-mask misalignment rather than the resolution limit of the lithography itself.

In practice, such a rule set is never looked up by hand; it is encoded into a machine-readable technology file (e.g. in LEF/DEF or a vendor-specific format), against which automated tools such as layout editors and the Design Rule Checker (see Section 8) continuously verify the layout.

Latchup and Well Rules

The Latchup Mechanism

In a CMOS process, NMOS and PMOS transistors are not electrically isolated from one another but share the same substrate: PMOS transistors sit inside an N-well, while NMOS transistors sit directly in the P-substrate. This arrangement

inevitably creates two parasitic bipolar transistors: a vertical PNP transistor (emitter = PMOS P^+ source/drain, base = N-well, collector = P-substrate) and a lateral NPN transistor (emitter = NMOS N^+ source/drain, base = P-substrate, collector = N-well).

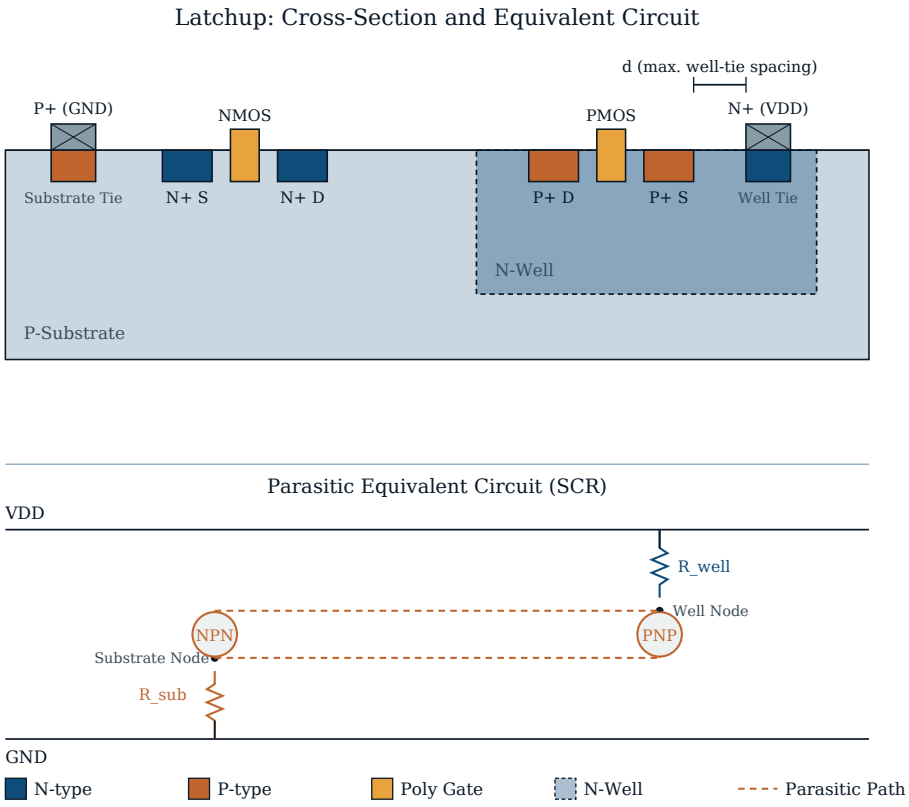
The collector of each parasitic transistor doubles as the base of the other — together with the lateral resistances of the well (R_{well}) and the substrate (R_{sub}), the PNP and NPN form a feedback four-layer structure electrically identical to a thyristor (SCR). If a voltage spike, a substrate current, or crosstalk momentarily turns on one of the two base-emitter junctions, this feedback can reinforce itself: the resulting low-resistance path between supply and ground — latchup — persists until the supply voltage is removed, and can destroy the chip thermally.

Well and Substrate Contacts as a Countermeasure

The current gain of the parasitic transistors can be effectively limited through layout: the lower the lateral resistances R_{well} and R_{sub}, the smaller the voltage drop needed to turn on the respective other transistor. Design rules therefore specify a maximum distance between every source/drain diffusion and the nearest well or substrate contact — usually set considerably tighter than pure fabrication tolerances alone would require.

In practice, these contacts are often implemented as a continuous guard ring: an N^+ ring around especially latchup-sensitive PMOS structures (tied to VDD), or a P^+ ring around NMOS structures (tied to GND), which diverts the base current along the shortest possible path before it can turn on the parasitic transistor. Such rings are particularly important around I/O drivers and other circuit blocks that carry high or fast-switching currents.

The diagram below shows the cross-section of a simple inverter with substrate and well contacts, along with the overlaid parasitic thyristor path.



Antenna Rules

The Plasma Charging Effect

Many etch and deposition steps in semiconductor fabrication — particularly reactive ion etching (RIE) of polysilicon and metal — operate using a plasma. During these steps, conductive structures on the wafer surface that are not yet electrically connected to anything else charge up, because the electron and ion currents arriving from the plasma do not hit all surfaces symmetrically. A long polysilicon or metal line that is still incompletely wired acts like an antenna: it collects charge from the plasma during etching, which — as long as the line terminates only at a transistor gate and is not yet connected to source/drain or a stabilizing lower layer — must discharge across the thin gate oxide.

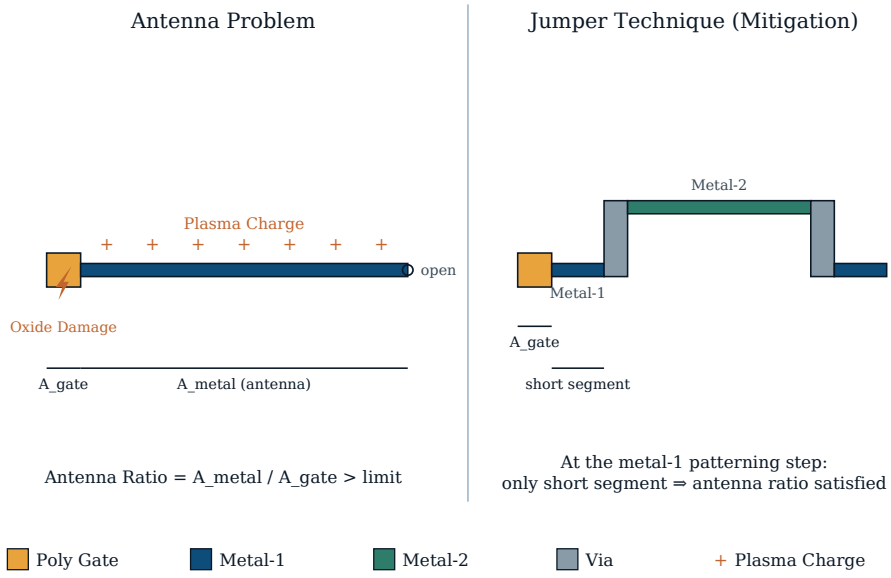
At only a few nanometers thick, the gate oxide is the most sensitive structure in the entire device. If the accumulated charge is enough to build up a sufficiently high electric field across the oxide, a Fowler-Nordheim tunneling current flows, damaging the oxide or, in extreme cases, causing breakdown. Such damage is often not immediately apparent as a failure but shows up later as premature device aging or elevated leakage currents, which makes debugging considerably harder.

Antenna Ratio as a Design Rule

To limit this risk, foundries define a maximum allowed antenna ratio for every conductive layer: the ratio of the area of all conductor segments electrically connected to a gate (poly, metal-1, metal-2, and so on, each counted up to the fabrication step at which it is patterned) to the area of the connected gate oxide itself. If a net connection exceeds this limit, the DRC reports an antenna violation.

Two common remedies exist in practice. The antenna diode adds an extra, reverse-biased diode to the substrate on the at-risk line, which bleeds off accumulated charge in a controlled way without affecting circuit function. The jumper technique instead breaks the long line on its original layer and routes it onward through a higher metal level: since that higher level is only patterned in a later, temporally separate fabrication step, the area connected to the gate at any given point in time is much smaller — keeping the antenna ratio within the allowed limit at every individual fabrication step.

The diagram below shows the antenna problem on the left, on a long line connected directly to the gate, and the mitigation via a metal jumper on the right.



Density Rules (CMP)

Why Pattern Density Is a Design Rule

Modern processes with multiple metal layers need a surface that is as flat as

possible between layers — only then can the next layer be exposed sharply, without focus deviations caused by height differences degrading the lithography. This planarization is handled by chemical-mechanical planarization (CMP), in which a rotating polishing pad wetted with slurry mechanically and chemically removes material from the wafer's oxide or metal surface.

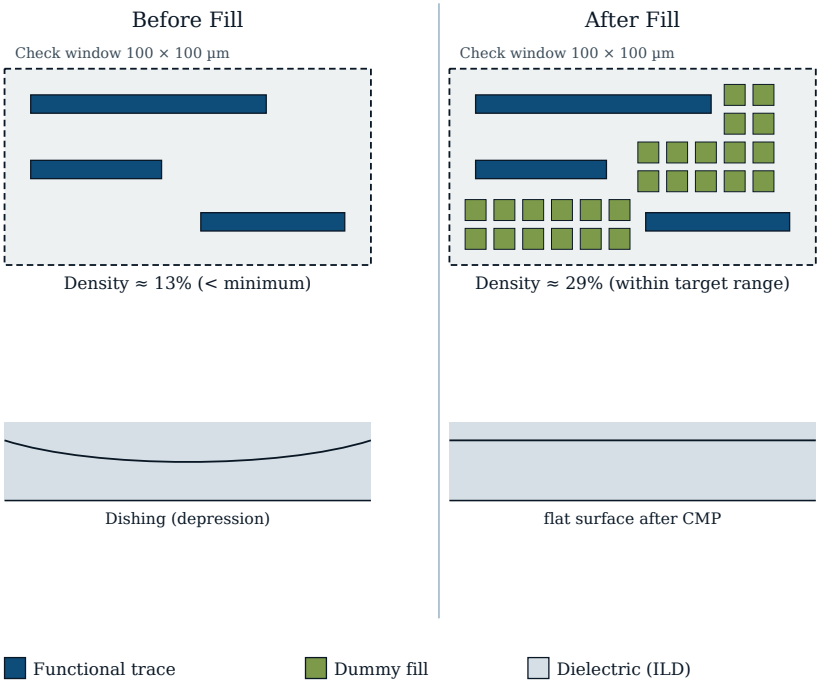
CMP, however, does not polish at the same rate everywhere: the removal rate depends on the local pattern density of the underlying layer. In large, mostly empty areas — for example above a wide, sparsely patterned trace — the pad polishes in deeper than intended, an effect known as dishing. Conversely, in densely packed areas with many narrow, closely spaced structures, erosion occurs, in which the surrounding dielectric is also removed more than intended. Both effects create height variations across the wafer surface that propagate into subsequent fabrication steps, where they can cause focus and yield problems.

Fill Structures as a Countermeasure

To prevent this, design rules specify a minimum and maximum local pattern density for each layer, typically measured over a sliding check window of fixed size (e.g. 100 µm × 100 µm) that is swept across the entire layout. If the density in a window falls below the minimum, automated fill generators insert additional, electrically non-functional dummy structures into the gaps — small, evenly distributed metal or polysilicon shapes that are not connected to the circuit's nets (floating fill) and exist solely to satisfy the density requirement.

These fill structures are themselves subject to design rules — for instance a minimum spacing to functional traces, to limit parasitic capacitance or unwanted coupling. Fill generation today is almost always fully automated and runs as the last step before tape-out, once the actual circuit layout has already been finalized.

The diagram below shows a check window before and after fill insertion, along with a schematic view of the resulting surface topography after the CMP step.



Design Rule Check (DRC)

Automated Rule Checking

A complete rule set is of little use unless it is consistently checked against every layout — for modern chips with billions of individual structures, manual inspection is out of the question from the start. This task is handled by the Design Rule Check (DRC): a software tool that checks the layout against the rules stored in a technology file (the so-called DRC deck) and reports every violation found as a flagged error.

At its core, a DRC tool works with Boolean and geometric operations on the individual layers: it computes intersections, unions, and spacing measurements between polygons on the same or different layers, and compares the results against the limits stored in the rule set. A width rule, for instance, is checked by measuring every structure on the relevant layer at its narrowest point; a spacing rule by determining the minimum distance between neighboring polygons.

Role in the Design Flow

In practice, DRC does not run just once at the end but repeatedly throughout the entire layout process: many layout editors offer real-time checking that flags violations with color highlighting as they are drawn, supplemented by full batch runs over the entire design ahead of major milestones. Every violation found

must be resolved before the next fabrication step — a layout is considered "DRC clean" only once a run over the complete design reports no remaining open violations. Only after that do further checks typically follow, such as layout-versus-schematic (LVS) verification, which confirms that the drawn layout actually implements the same circuit as the schematic.

The diagram below shows a typical spacing violation — two poly lines with insufficient spacing — compared to the corrected, rule-compliant geometry.