

Semiconductors 101, unit 5

Masks and tape-out: how a design becomes a mask set

Handing a checked design to the fab is called tape-out, and changing it after that means new masks. The design can't be printed as it is either: it first has to be corrected and made into a set of photomasks. A mask is the "negative" for exposure, and one printable spot or one small error on it prints in every

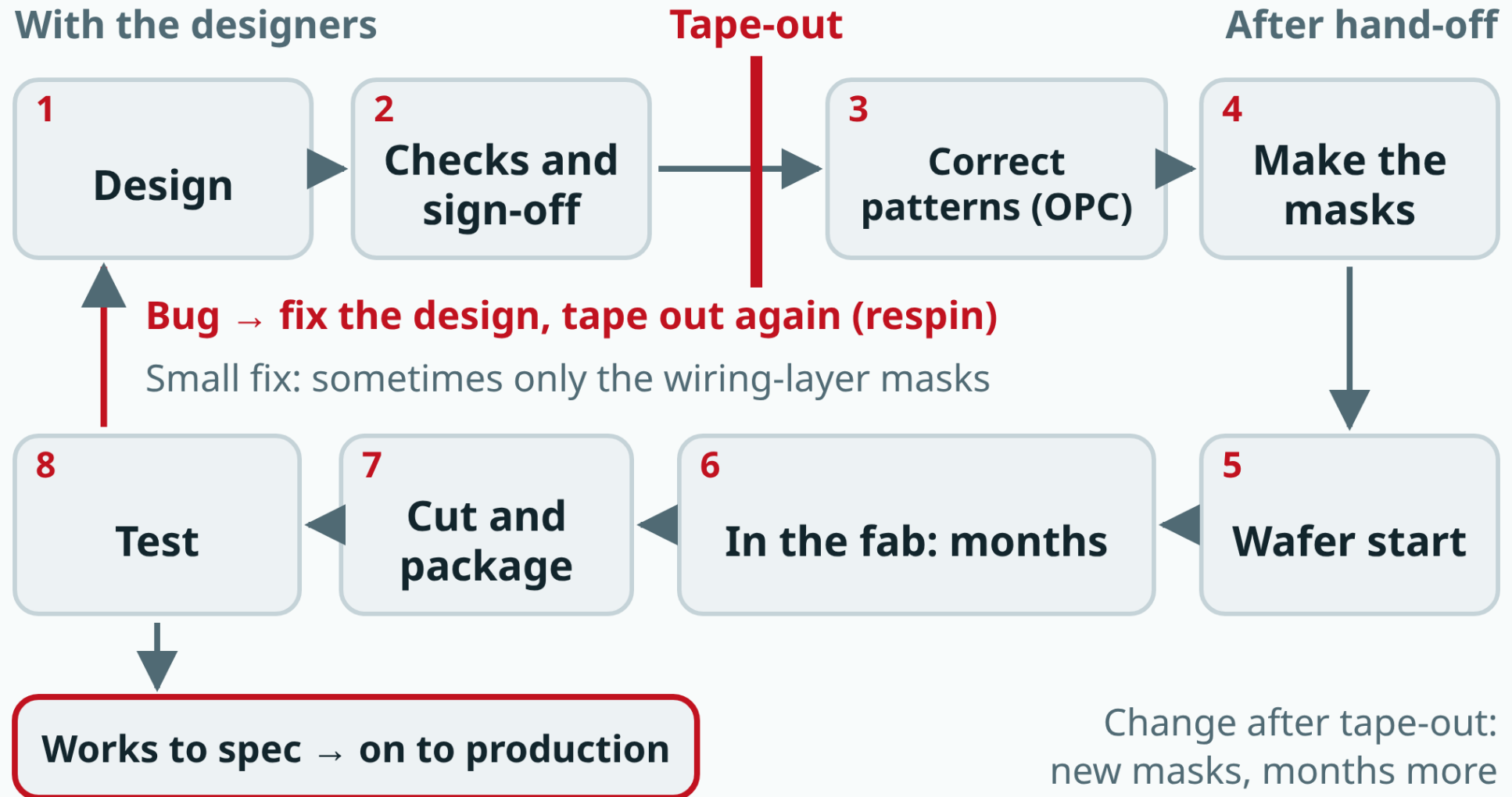
field of every wafer, so it has to be drawn extremely accurately and checked extremely closely.

Semiconductors 101: a free course following the life of a chip

Free to use in class and for non-commercial teaching. Please credit fab24hr. Do not sell it or present it as your own work.

Goals

- 1 Say what is checked one last time before tape-out, what is handed over, and why changing the design after tape-out is costly and slow.
- 2 Say how a mask prints its pattern onto the wafer, why one chip design needs dozens of masks, and why the pattern is deliberately changed first with OPC, with computers working out how much.
- 3 Know how a mask is written, inspected and repaired, why a set costs so much, and why it is checked again once it reaches the fab.



Tape-out is a dividing line: before it, the design is still in the designers' hands and can be changed; after it, the patterns are corrected by computer and then become masks and wafers, and a change means new masks and months more. A sketch; time is not to scale.

Step by step

1 Tape out: hand over the files with every layer's shapes; the fab checks them again itself when they arrive.

Why: The masks are made from these files, and every pattern later printed on the wafers comes from them.

If not: If a layer is left out or the wrong version goes out, the masks won't match the design that was checked.

2 For the finest layers, change the patterns on purpose by computer first (OPC), then check the whole layer by calculating it again.

Why: The lines are narrower than the wavelength of the light used to print them; printed as drawn, corners come out round and line ends come out short.

If not: What prints won't match the design, and things that should connect won't.

3 At the mask shop, write the pattern piece by piece with an electron beam or a laser onto a quartz plate coated with chromium and resist.

Why: A mask is the very first negative, with no other negative to print it from, and the finest patterns need a finely focused electron beam.

If not: If the machine drifts even slightly while writing, that layer won't line up with the one below once it's printed on the wafer.

Step by step

4 Develop it, etch away the uncovered chromium, then remove the leftover resist.

Why: It's the chromium that actually blocks the light; the resist is only a stencil along the way.

If not: Chromium not fully etched still blocks light where it should pass; etch too far, and the lines get thinner.

5 Measure line widths and positions, scan the whole plate for extra or missing bits, and repair them with an electron or ion beam.

Why: One printable defect on a mask prints in every field of every wafer.

If not: One missed spot can ruin many chips at once; a mask that can't be repaired has to be made again.

6 Clean it, inspect it once more, and mount the pellicle, a film floating a few millimeters above the pattern.

Why: Dust that lands on the film is too far from the pattern to be in focus, so it usually doesn't print.

If not: Dust that lands right on the pattern prints sharply in every field.

Step by step

7 Ship it to the fab in a dedicated mask pod; on arrival it is scanned again and its ID and report are checked.

Why: It was checked before shipping, but things can still go wrong on the way: a particle that gets under the pellicle onto the pattern prints just the same.

If not: Put on the scanner unchecked, one spot gets printed lot after lot, and many chips fail together.

Tape-out: the moment the design is final and goes out to be made into masks

Tape-out is the moment a finished, checked chip design is handed off as its final files to be made into photomasks. Before it goes, design software goes over the whole design once more: does any shape break the fab's rules, is the drawn layout the same circuit as the original, do signals arrive in time? The function has to be verified too, and only then do the people responsible approve the release; this gate is called sign-off. What goes out is the shapes of every layer, the whole stack of which is called the layout, and the fab checks them again itself when they arrive. Any change after that means new masks: paying again, and the wafers starting over from the first station, months again. So many designs scatter some spare little circuits across the chip beforehand; if a small bug turns up after tape-out, only the wiring above the transistors is changed to hook a spare circuit in, and only those few wiring-layer masks are remade. But only a limited range of bugs can be fixed this way; with bigger changes, the whole set is made again.

A mask works like a stamp: at least one per layer, dozens per set

A photomask is the "negative" used for exposure: a very flat quartz plate coated with a metal film drawn in the shape of the circuit (usually chromium). Where there's film, light is blocked; where there isn't, light passes. The scanner (the tool that prints patterns onto wafers) shines light through it and uses lenses to shrink the pattern (the pattern on the mask is usually drawn four times the size it will be on the chip), printing one small square at a time, called an exposure field; after each field the wafer shifts a little and the next one is printed, like covering a sheet of paper with the same stamp. So one extra spot on the mask means one extra spot in the same place in every field. A chip is built up layer by layer, and every layer has a different shape and needs at least one mask; the finest layers are sometimes split into two or three separate exposures, so a full set often has dozens.

Drawn as designed, it prints out of shape, so OPC changes it on purpose

The mask can't simply be drawn like the design: the deep ultraviolet light commonly used in scanners today has a wavelength of 193 nanometers, while a chip's finest lines are only tens of nanometers wide. Light through a very narrow gap spreads out (diffraction); the lens misses the widest-spreading light, which is exactly where the finest details are, so the print looks like a fine drawing traced with a thick marker: right angles come out round and line ends pull back, and one that pulls back too far misses its connection, breaking the circuit. Light spreading from nearby shapes also overlaps, so equal lines print at different widths in a crowd and alone: that is the "proximity" in the name. So before a mask is made, a computer deliberately changes the patterns to be printed (most of all on the finest layers) away from the design. This is OPC (optical proximity correction), like an archer who aims a little left because the wind blows right: small squares on corners, line ends drawn longer, crowded and lone lines drawn at different widths. How much is calculated: every edge is cut into short pieces, the computer works out how each will print, and pushes it out where the print falls short or pulls it in where it bulges; moving one piece changes its neighbors, so it takes several rounds. Finally the whole layer is calculated again, also with the focus and the amount of light a little off, to find lines that will merge or nearly break, because scanners drift a little every day.

The mask shop writes it stroke by stroke and can't leave a single defect, so masks are expensive

The corrected pattern goes to the mask shop. A mask is the very first negative, with no other negative to print it from, so it has to be written piece by piece: by a finely focused electron beam for the finest layers, or by a laser for less demanding ones, onto a quartz plate coated with chromium and resist (a film that changes where the beam or light hits it); then it is developed and the uncovered chromium is etched away. If the machine drifts even slightly while writing, that layer won't line up with the one below once it's printed on the wafer. The finished mask is measured and scanned all over for extra or missing bits, which are fixed with a very fine electron or ion beam; one that can't be fixed is made again. Last comes the pellicle, a film floating a few millimeters above the pattern: dust that lands on it is out of focus and usually doesn't print. It has to be extremely accurate with not one defect left, more OPC corners usually take longer to write, and a set of dozens serves only this one chip design, so masks are expensive; a design error found after tape-out means paying again, in both money and time, for new masks.

Common misconceptions

People say: An electron beam can draw the finest lines, so it could draw straight on the wafer and there would be no need for a mask.

In fact: It's too slow. A mask only has to be drawn stroke by stroke once; after that the scanner just flashes each field, so even a wafer with dozens to a hundred or more fields is printed quickly, and the same mask can print a great many wafers. Drawing every wafer directly with an electron beam couldn't keep up with volume production, so electron beams are used mostly for making masks, for research, and for things made in very small numbers.

People say: A mask is checked before it leaves the mask shop, so once it's in the fab it can be used indefinitely.

In fact: Checked before shipping doesn't mean clean on arrival: a particle can get in on the way, so the fab inspects each mask again when it arrives. And clean on arrival doesn't mean clean forever: with use, traces of chemicals left on the surface react with the exposure light and moisture and slowly grow fine crystals that look like fog, which is called haze; and every move, every trip onto and off the scanner, can pick up particles. That's why a fab has a mask engineer who inspects masks on arrival and takes them down for another scan after a while. If the inspection after developing finds the same spot in the same place in every field (a repeating defect), it is almost always the mask.

Play

Play a shift as the mask engineer: when a mask has haze, a new spot or a scratch, you decide whether it can still go on the scanner

Scan with a phone to play from here

fab24hr.com/en/?job=mask&utm_source=slides



Think

- 1 The finer the OPC, the bigger the mask data and the longer and more expensive the mask; the more conditions it's checked under, the longer the computers have to run. But a pattern that is only right under ideal conditions can fail in production. If the tape-out date were tight, which layers and which places would you spend the time on, and why?
- 2 One spot on a mask prints in every field, so masks are taken down and inspected after a while. Check too often and the mask spends more time out of use; check too rarely and a problem may only be found once it has grown. If you set the schedule, how would you balance it, and what would you look at?

Read

What is tape-out, and what happens after a chip design is handed off?

fab24hr.com/en/blog/what-is-tapeout/

What is a photomask? How is it made, and why does a set cost so much?

fab24hr.com/en/blog/what-is-photomask/

What is OPC, and why doesn't the photomask look like the chip?

fab24hr.com/en/blog/what-is-opc/

Further

What does a mask engineer do in a chip fab?

Semiconductors 101: a free course following the life of a chip

<https://fab24hr.com/en/learn/tapeout/>

Written from public textbooks, industry standards, university course notes and papers; the companies, people, machine numbers and figures in the game are fictional.

Free to use in class and for non-commercial teaching. Please credit fab24hr. Do not sell it or present it as your own work.

Generic illustration made with AI help from public sources; may contain mistakes; not any company's tool