What moves the threshold

In Lab 03 you extracted $V_{TH} = 0.6016$ V for the W = 5 µm, L = 1 µm transistor. A second method gave 0.5159 V, and BSIM's own value was 0.5895 V, so the number you call $\boldsymbol{V_{TH}}$ **depends on the method you used to extract it**.

This page covers the other half of the story. Even with the method fixed, the *same transistor* has a different threshold depending on three things you have been holding still so far: the voltage between its source and bulk, the length you drew, and the temperature.

The fourth terminal

Look at any device line in this course:

```
XM1 d g 0 0 sky130_fd_pr__nfet_01v8 L=1 W=5
```

It names four nodes, in the order **drain, gate, source, bulk**. On every deck since Lab 01 you have wired the last two to the same place, node 0, so the voltage between them has always been zero.

The **bulk** (also called the *body* or *substrate*) is the silicon the transistor is built in. For an NMOS in this process that's the p-type wafer itself, which every other NMOS on the chip shares. You don't usually get to choose its voltage, because it's tied to ground.

What you *can* change, sometimes without meaning to, is the **source** voltage. If you stack two NMOS transistors in series, as in the bottom half of a two-input NAND gate, the upper one's source sits on the lower one's drain, somewhere above ground. Its bulk is still at 0, so for that device $V_{SB} > 0$.

Body effect

The Lab 03 deck `spice/vth_body.spice` uses the same W = 5 µm, L = 1 µm device with the source lifted onto its own node. V_{GS} and V_{DS} are held fixed *relative to the source* (V_{GS} = 1.8 V, V_{DS} = 0.05 V), so the only thing that changes is V_{SB}:

V_{SB}	V_{TH}	shift	I_D
0 V	0.5894596 V	—	66.3374 µA
0.3 V	0.6607482 V	**+71.289 mV**	61.6210 µA
0.6 V	0.7213725 V	**+131.913 mV**	57.7533 µA
0.9 V	0.7745817 V	**+185.122 mV**	54.4634 µA
1.2 V	0.8222507 V	**+232.791 mV**	51.5967 µA

Lifting the source 1.2 V above the bulk moved the threshold by **232.791 mV**, which is nearly a quarter of a volt on a supply of only 1.8 V, and it cost **22.2 %** of the drain current at the same gate drive, even though nothing was rewired and no dimension changed.

This is the **body effect**. It's the reason a two-input NAND gate is slower than an inverter by more than the extra transistor's capacitance accounts for. Because the upper device in the stack doesn't have its source at ground, it works against a threshold a hundred-odd millivolts higher than the one on its datasheet.

The body-effect formula

The standard formula, which your device-physics course will give you, is

$$V_{TH}(V_{SB}) = V_{TH0} + \gamma\left(\sqrt{2\phi_F + V_{SB}} - \sqrt{2\phi_F}\right)$$

Here γ is the **body-effect coefficient**, in $\sqrt{\text{V}}$, and $2\phi_F$ is roughly twice the Fermi potential of the substrate. Both are set by how heavily the silicon is doped, and neither is printed anywhere easy to read, so instead you can fit them to the five measured points, which also tells you whether the formula is any good.

Take $V_{TH0} = 0.5894596$ V from the $V_{SB} = 0$ row, then do a least-squares fit of the other four points for γ and $2\phi_F$:

$$\gamma = 0.4108\ \sqrt{\text{V}} \qquad 2\phi_F = 0.600\ \text{V}$$

V_{SB}	measured	formula	difference
0 V	0.5895 V	0.5895 V	—
0.3 V	0.6607 V	0.6610 V	**−0.23 mV**
0.6 V	0.7214 V	0.7213 V	**+0.11 mV**
0.9 V	0.7746 V	0.7744 V	**+0.20 mV**
1.2 V	0.8223 V	0.8224 V	**−0.15 mV**

With two parameters fitted to five points, the worst error is **0.23 mV**. That's four significant figures of agreement between a formula from the 1960s and a BSIM4 model with several hundred parameters.

Compare that with the last two pages, where the square law missed the saturation knee by 35 % and the $\sqrt{I_D}$ extraction missed the threshold by 12 %. Here a formula from the same era lands within a quarter of a millivolt. The difference is that the square law relies on physics that short channels changed, while the body-effect formula rests on the depletion charge under the channel, which doesn't depend on how long the channel is. That's why the square-root law survived where the square law didn't.

> **When a hand formula disagrees with a simulator,** ask which assumption of the formula the device violates. If you can't name one, suspect your arithmetic.

Threshold and drawn length

Shortening a device raises its current, and it also moves its threshold, in a direction few people predict. The Lab 03 deck `spice/vth_vs_l.spice` has nine copies of the same W = 5 µm transistor at nine lengths, all at V_{DS} = 50 mV so that the drain doesn't influence the answer:

```
L        model vth
0.15 um    7.078517e-01 V
0.2 um     6.872240e-01 V
0.3 um     6.553824e-01 V
0.5 um     6.300670e-01 V
0.8 um     6.003750e-01 V
1 um       5.894596e-01 V
1.5 um     5.627056e-01 V
2 um       5.495175e-01 V
4 um       5.364102e-01 V
```

Most textbooks introduce **threshold roll-off**: as a channel gets shorter, the source and drain depletion regions take over some of the charge the gate would otherwise have to support, the gate's job gets easier, and V_{TH} **falls**. These nine devices go the other way. Their threshold rises monotonically as the channel gets shorter, from **0.5364102 V** at L = 4 µm to **0.7078517 V** at L = 0.15 µm, a span of **171.44 mV**, so the shortest device has the *highest* threshold.

This is the **reverse short-channel effect**, and it comes from a manufacturing decision rather than from unavoidable physics. Processes at this node implant extra p-type dopant in pockets right beside the source and drain, called **halo** or **pocket** implants, specifically to fight roll-off, because a short device whose threshold collapses also leaks. In SKY130 the halo over-corrects: the average doping under a short channel ends up *higher* than under a long one, so the gate has more charge to support rather than less.

This has two practical consequences. The first is that you can't shorten a transistor and keep its bias point, because 171 mV of threshold is most of an overdrive voltage. The second is that analog designers who need two transistors to behave identically always give them **the same length**, even when the equations say only W/L matters. In Lab 04 you'll see five devices with identical W/L carry five different currents, and this is one of the reasons.

Threshold and temperature

The same `vth_body.spice` run ends by simulating the device at $V_{SB} = 0$ at three temperatures, spanning the automotive range that real parts are specified over:

T	V_{TH}	I_D
−40 °C	0.6499250 V	93.8168 µA
27 °C	0.5894596 V	66.3374 µA
125 °C	0.5010177 V	45.0650 µA

V_{TH} falls by **148.9 mV** across the range, a slope of **−0.9025 mV/°C**, which is close enough to the −1 mV/°C rule of thumb that you can simply remember the rule of thumb.

Before you read the current column, predict which way it goes. A lower threshold means more overdrive at the same gate voltage, so you might expect the hot device to carry *more* current.

In fact the hot device carries **32.1 % less** current than at 27 °C, and the cold one carries **41.4 % more**.

That's because temperature moves more than the threshold. Carrier **mobility** falls as the lattice heats up and scatters carriers more, roughly as $T^{-1.5}$, and at this overdrive the mobility effect wins comfortably. The two effects have opposite signs, and which one dominates depends on the bias. At very low overdrive, near threshold, the threshold term wins and the current goes *up* with temperature. The gate voltage at which the two cancel is called the **zero-temperature-coefficient point**, and circuits that must not drift are sometimes deliberately biased there.

> **"Does this get better or worse when it's hot?"** has no general answer for a MOSFET, so simulate both ends. This is the corner discipline AD102 drilled into you with resistors, and it matters more here, because a resistor's temperature coefficient doesn't change sign on you.

You'll run `vth_body.spice` and `vth_vs_l.spice` yourself in Lab 04, Part 2.

Check yourself

1. In a stack of two NMOS transistors, which one sees $V_{SB} > 0$, and does that make it stronger or weaker?
2. Using γ = 0.4108 $\sqrt{\text{V}}$ and $2\phi_F$ = 0.600 V, estimate the threshold shift at V_{SB} = 0.6 V and compare it with the table.
3. Why does the halo implant raise the threshold of a short device more than a long one?
4. At this overdrive, the drain current falls when the device heats up even though V_{TH} falls. Which effect wins, and would the answer be the same just above threshold?

Next: Transconductance and output resistance.