The current mirror

The CMOS inverter works as an amplifier only if its input is held at 0.712 V, where it sits on a slope with a gain of 116. This page is about what holds a bias point like that in place, and the answer is a **current** rather than a voltage.

It's the last new idea in AD103, and it's the one that turns the previous pages into a circuit you could build. You have already met the idea behind it in AD102, applied to resistors.

Why not distribute a voltage

Every amplifier on a chip needs to be biased at the right point. The obvious plan is to generate one good bias voltage and wire it to every amplifier. Before you do, count what that voltage would have to survive, using numbers from earlier pages:

It has to survive	measured
threshold moving with temperature	**−0.9025 mV/°C**, 148.9 mV across the automotive range
threshold moving with channel length	**171.44 mV** from L = 4 µm to L = 0.15 µm
threshold moving with source-to-bulk voltage	**232.791 mV** at V_{SB} = 1.2 V
process corners	Part 1 of AD102 Lab 04

A gate voltage that is 100 mV wrong, on a device with g_m = 334 µS, gives a current that is **33 µA** wrong, so the plan doesn't survive a real die.

AD102′s Lab 04 reached a conclusion that points the way out: **absolute values wander, but ratios hold.** Two devices side by side on the same die are at the same temperature and in the same corner, and they have the same threshold, so whatever moved one has moved the other too.

So instead of distributing a voltage, you distribute a *current*, using a circuit whose output is a **ratio** of its input.

The circuit

Take a transistor and tie its gate to its drain. It is now a two-terminal device, and it is in saturation automatically: $V_{DS} = V_{GS}$, so V_{DS} always exceeds $V_{GS} - V_{TH}$. This is called a **diode-connected** transistor.

If you push a known current into it, it settles at whatever gate voltage carries that current. In effect the device solves its own equation for you, which is useful because you don't know $\mu_n C_{ox}$ and it does.

Now connect a second transistor's gate to the same node. The second device has the same V_{GS}, is made of the same silicon and sits at the same temperature, so it carries the same current. That pair is a **current mirror**.

The capstone package's `spice/mirror.spice` forces 50 µA into the reference branch and prints the gate voltage it settles at:

```
--- part 1: the reference branch sets a gate voltage ---
v(ng) = 8.672678e-01
@m.xmref.msky130_fd_pr__nfet_01v8[vdsat] = 2.336319e-01
```

The diode-connected device chose **0.8672678 V** to carry 50 µA. Nobody calculated that number; the transistor found it for itself. If the die is hot or slow, or the threshold has moved, the number moves with it, which is what you want, because the copy has shifted in the same way, so the gate voltage that is right for the reference is right for the copy too.

With the copy's output held at 0.9 V, the deck measures the copied current:

```
i(vo2) = -5.00878e-05
```

That's **50.0878 µA against a 50 µA reference, a 0.176 % error**, from a circuit with no resistors, no feedback and no trimming.

Limit 1: output headroom

A saturated transistor is very nearly a current source, but if you drop it out of saturation it becomes a resistor. Sweeping the output node shows where that happens:

V_{out}	I_{out}	error vs 50 µA
0.1 V	28.4076 µA	**−43.18 %**
0.2 V	42.7022 µA	−14.60 %
0.4 V	48.1163 µA	−3.77 %
0.6 V	49.1572 µA	−1.69 %
0.9 V	50.0878 µA	+0.18 %
1.2 V	50.8174 µA	+1.63 %
1.8 V	52.0797 µA	+4.16 %

The error column contains two separate effects, and you need to be able to tell them apart.

Below about 0.25 V the copy collapses. The reference branch already printed the reason: `vdsat = 2.336319e-01`. With less than **0.2336 V** on its drain, the output device is in triode rather than saturation, and a triode device's current depends on its drain voltage. Every current mirror has a minimum output voltage like this, called its **compliance**, and for this one it's V_{DSAT}, about 234 mV. That's why a 1.8 V supply feels tight: every current source in a stack uses up about a quarter of a volt just to work.

Above that, the copy drifts upward by about 1.6 % per 0.3 V. That drift is r_o, the finite slope of saturation, and you can estimate it from the 0.6 V and 1.8 V rows, which are both clear of the collapse:

$$r_o \approx \frac{1.8 - 0.6}{(52.0797 - 49.1572)\ \mu\ \text{A}} = 410.6\ \text{k}\ \Omega$$

So this "current source" has 410 kΩ in parallel with it. That finite resistance is the reason the next circuit in every analog textbook is the **cascode**, which stacks a second transistor on top to shield the first from the output swing. It buys back a factor of $g_m r_o$ in output resistance, at the cost of another V_{DSAT} of compliance.

> **Before you call a mirror broken, ask what V_{out} was.** When half the current is missing, the cause is almost always a device in triode rather than a wiring error.

Limit 2: ratio error

If you widen the output device, the copy scales with it, so a mirror is also a *multiplier*. The deck draws 1:1, 1:2 and 1:4 mirrors by width, all from one gate voltage and with the outputs at 0.9 V, so widths of W = 5, 10 and 20 µm should give 1×, 2× and 4×.

drawn	W	I_{out}	actual ratio	error
1×	5 µm	50.0878 µA	1.0000	—
2×	10 µm	112.313 µA	**2.2423**	**+12.12 %**
4×	20 µm	239.347 µA	**4.7785**	**+19.46 %**

The 2× branch is twelve percent high and the 4× branch nearly twenty, which is a lot for a circuit whose whole selling point is exact ratios. The deck prints the cause on the next lines:

```
  and why the ratio is not what you drew:
@m.xmw1.msky130_fd_pr__nfet_01v8[vth] = 5.894596e-01
@m.xmw2.msky130_fd_pr__nfet_01v8[vth] = 5.766246e-01
@m.xmw1.msky130_fd_pr__nfet_01v8[gm] = 3.340839e-04
@m.xmw2.msky130_fd_pr__nfet_01v8[gm] = 7.187672e-04
```

The W = 10 µm device has a threshold **12.835 mV lower** than the W = 5 µm one. This course's Lab 04 measured the same shift independently (`vth(W=10) - vth(W=5) = -12.8354 mV`), so two different decks agree to five figures. The cause is the **narrow-width effect**: the edges of a channel behave differently from its middle, and in a wider device the edges are a smaller fraction of the whole channel.

Does 12.8 mV explain 12 %? The wide device's g_m is 718.7672 µS, so a threshold 12.835 mV lower is worth about

$$\Delta I \approx g_m \Delta V_{TH} = 718.7672\ \mu\ \text{S} \times 12.835\ \text{mV} = 9.23\ \mu\ \text{A}$$

on a branch that should carry $2 \times 50.0878 = 100.1756$ µA. The measured excess is **12.14 µA**, which is the same size and the same sign. The straight-line estimate reads low because at g_m/I_D = 6.40 /V (718.7672 µS on 112.313 µA) this device is in moderate inversion, where current rises faster than linearly with gate voltage. So the measurement confirms the mechanism even though the linear estimate is loose, and it's worth keeping those two claims separate: you've shown what causes the error, not that the straight-line formula predicts its size.

The fix: unit devices

Matching and Pelgrom's law gave you a rule for resistors:

> *If you need a ratio, build both sides from repeated copies of the same unit device, never from one big device and one small one.*

A 4:1 resistor ratio is built from four identical units and one unit, never from a long strip beside a short one, because the ends of a strip don't scale with its length. Transistors follow the same rule for the same reason: **the edges don't scale.**

So don't draw one W = 10 µm device. Draw **two W = 5 µm devices in parallel**, which have the same total width and the same total area as the single device, and in theory carry the same current.

```
--- part 4: one W=10 device vs two W=5 devices in parallel ---
i(vo3) = -1.12313e-04
i(vo5) = -1.00176e-04
```

output branch	I_{out}	ideal 2 ×	error
one W = 10 µm device	112.313 µA	100.1756 µA	**+12.12 %**
two W = 5 µm devices in parallel	**100.176 µA**	100.1756 µA	**+0.0004 %**

The parallel pair matches the ideal to six figures. Two copies of the unit device carry exactly twice the unit device's current because they *are* two copies of it, with the same width, the same edges and the same threshold, so there's nothing left to differ.

This is why real analog layout is full of arrays of identical transistors with their sources and drains strapped together, and why a schematic that says `m=4` means "four of these" rather than "one four times as wide". It's the rule AD102 showed on resistors (where it made a 3:1 divider land at 1.350000 V instead of 1.309327 V), reached here from a different direction.

> **When a schematic asks for a ratio of two devices,** check whether both sides are built from the same unit. If they aren't, the ratio has a systematic error that no amount of care will remove.

You'll run `spice/mirror.spice` yourself in the capstone lab.

After AD103

AD104 takes the inverter you draw in the capstone and has you build it as geometry. The capstone lab's last section shows what that looks like and which ideas from this course come back there.

Check yourself

1. Why is a diode-connected transistor always in saturation?
2. A mirror's output sits at 0.15 V. Which limit is it hitting, and what would you check first?
3. Using r_o = 410.6 kΩ, how much would the copied current change if V_{out} moved by 0.5 V?
4. Why do two W = 5 µm devices in parallel copy a current more accurately than one W = 10 µm device?

Next: Capstone: The CMOS inverter.