



The Design of a Low-Voltage Bandgap Reference

Most integrated circuits incorporate bandgap references (often simply called *bandgaps*) to define certain dc voltages or currents that serve various building blocks. In this article, we introduce a step-by-step procedure for the design of low-voltage bandgaps. As presented in Figure 1, a typical power-management environment employs a low-dropout (LDO) circuit that, from a global supply of 1.2 V, generates a moderately regulated voltage around 1 V. This voltage acts as a local supply for the bandgap circuit and some other building blocks. It is desirable for the bandgap to provide substantial supply rejection to minimize corruption in its output due to the electronic noise produced by the LDO and the transient perturbations caused by the switching activities within the other building blocks.

We target the following specifications:

- output voltage = 0.5 V
- output voltage variation < 5 mV from 0° C to 100° C
- supply rejection > 40 dB
- power consumption < 1 mW
- supply voltage = 1 V ± 5%.

We design the circuit in 28-nm CMOS technology. The reader is referred to [1]–[12] for background information.

Basic Operation

We wish to generate a voltage that is nominally independent of the

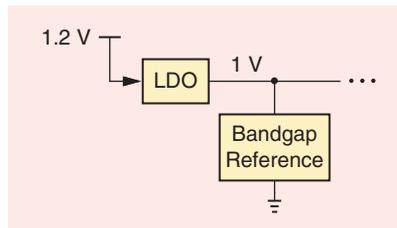


FIGURE 1: A typical power-management environment.

temperature. This can be accomplished by summing two voltages that have opposite temperature coefficients (TCs), as practiced in [1], [3], and [11]. Alternatively, we can first sum two currents of opposite TCs and then allow the result to flow through a resistor [9]. We pursue the latter here.

The bandgap core is typically realized as illustrated in Figure 2(a), where the emitter areas of Q_1 and Q_2 differ by a factor of n , and amplifier A_1 adjusts the gate voltage of M_1 and M_2 to equalize V_X and V_Y . We thus obtain

$$V_{BE1} = V_{BE2} + |I_{D2}|R_1. \quad (1)$$

Hence,

$$V_T \ln \frac{I_{D1}}{I_{S1}} = V_T \ln \frac{I_{D2}}{I_{S2}} + |I_{D2}|R_1, \quad (2)$$

where I_{S1} and I_{S2} denote the emitter saturation currents of Q_1 and Q_2 , respectively. Viewing Q_1 as a unit and Q_2 as n units in parallel, we have $I_{S2} = nI_{S1}$ and

$$|I_{D2}|R_1 = V_T \ln n, \quad (3)$$

where M_1 and M_2 are assumed to be identical. The voltage across R_1 is therefore proportional to the absolute temperature (PTAT) and so are the drain currents of M_1 and M_2 if R_1 has a zero TC.

It is possible to make I_{D1} and I_{D2} independent of the temperature by attaching two resistors from X and Y to the ground [see Figure 2(b)] [9]. Let us formulate the circuit's behavior, assuming that $R_2 = R_3$. Since $V_X \approx V_Y$, (1) still holds, and the current through R_1 is still equal to $(V_T \ln n)/R_1$. Summing this current and that through R_3 , we have

$$|I_{D1}| = |I_{D2}| = \frac{V_T \ln n}{R_1} + \frac{|V_{BE1}|}{R_3} \quad (4)$$

$$= \frac{1}{R_3} \left(\frac{R_3}{R_1} V_T \ln n + |V_{BE1}| \right). \quad (5)$$

The two terms on the right-hand side of (5) represent currents with opposite TCs. For $|I_{D2}|$ to have a TC of zero, we select $(R_3/R_1)V_T \ln n$ to be approximately $17V_T$ [12]. Now, as depicted in Figure 2(c), this current is copied and applied to a resistor to yield a nominally constant output voltage [9],

$$V_{out} = \frac{R_L}{R_3} \left(\frac{R_3}{R_1} V_T \ln n + |V_{BE1}| \right). \quad (6)$$

The key to the circuit's low-voltage operation is that V_{out} can be arbitrarily small even though $|V_{BE}| + 17V_T \approx 1.2$ V at $T = 25^\circ$ C.

Design Issues

The topology of Figure 2(c) entails several issues. First, noting that the TC of

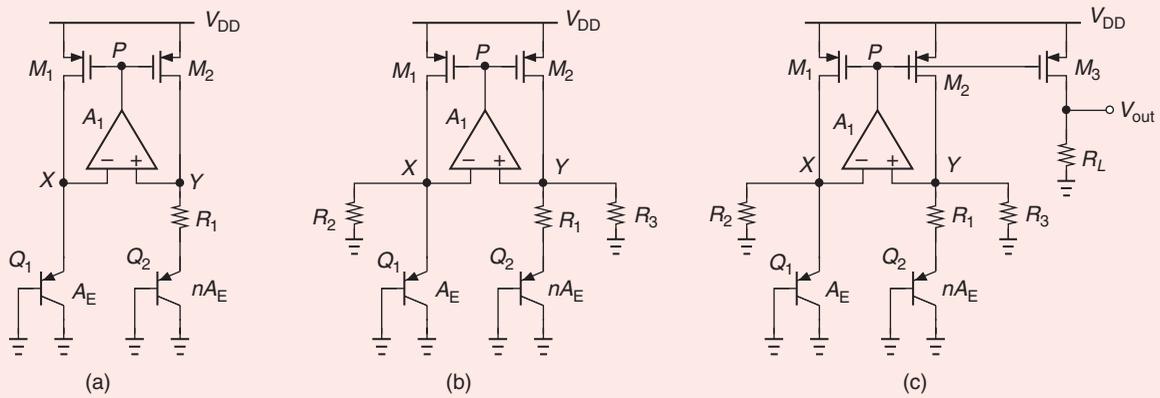


FIGURE 2: (a) A basic bandgap core. (b) The addition of resistors to create constant currents. (c) The addition of an output branch.

the base-emitter voltage, $\partial V_{BE}/\partial T$, is around $-1.5\text{mV}/^\circ\text{C}$, we expect $|V_{BE}|$ to be relatively large at low temperatures. With a worst-case V_{DD} of 0.95V , this leaves little voltage headroom for M_1 and M_2 . We must therefore select relatively large bipolar transistors and low collector currents to ensure a moderate $|V_{BE}| = V_T \ln(I_C/I_S)$.

Second, as T goes from 0°C to $+100^\circ\text{C}$, $|V_{BE}|$ in Figure 2(c) drops by roughly 150mV , whereas V_{out} remains relatively constant. The resulting difference between the drain-source voltages of $M_{1,2}$ and M_3 leads to a substantial error in I_{D3} and, hence, a large variation in V_{out} . We will resolve this issue through the use of a regulated cascode structure.

Third, the offset of A_1 , V_{OS1} , in Figure 2(c) introduces an error in V_{out} . We have [13]

$$V_{out} = \frac{R_L}{R_3} \left[|V_{BE1}| + \frac{R_3}{R_1} V_T \ln n - \left(1 + \frac{R_3}{R_1} \right) V_{OS1} \right]. \quad (7)$$

The contribution of V_{OS1} can be minimized by maximizing $\ln n$ —a remedy that costs chip area.

Fourth, the 40-dB supply-rejection requirement imposes a lower bound on the operation amplifier (op amp) gain in Figure 2(c). As explained next, A_1 must reach several hundred.

Fifth, the bandgap core of Figure 2(b) and (c) can indefinitely remain off after powerup if V_X and V_Y begin from zero and A_1 loses the ability to control V_P . The core

must therefore incorporate a start-up circuit.

Core Design

The design of the core presented in Figure 2(a) begins with the choice of the

bipolar transistors' dimensions and emitter area ratio, n . From the issues outlined in the previous section, we note that the limited voltage headroom makes it desirable to minimize $|V_{BE}|$ and, hence, maximize the emitter

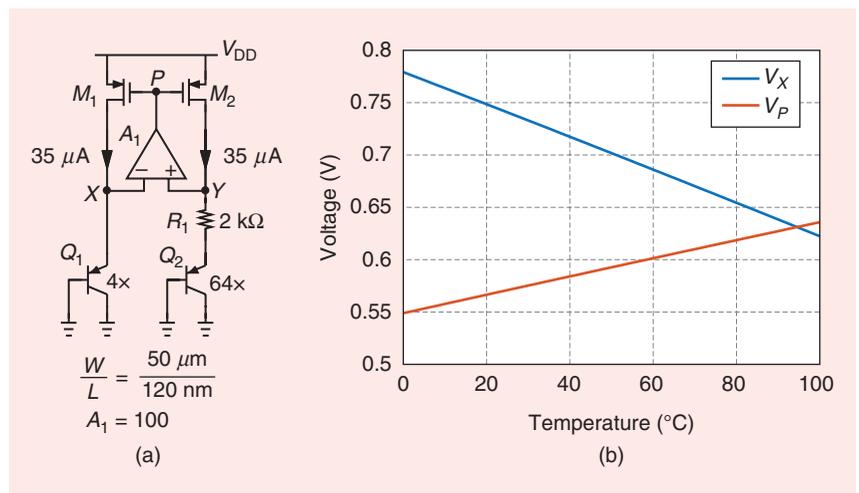


FIGURE 3: (a) The preliminary core design and (b) its internal voltages versus T .

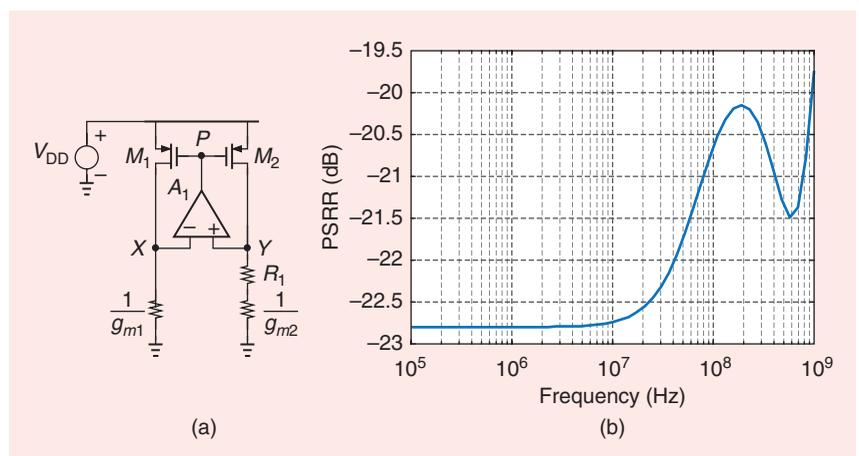


FIGURE 4: (a) A test setup for studying PSRR and (b) the PSRR of the basic core.

Next, we investigate the core's supply rejection by constructing the setup displayed in Figure 4. The supply voltage varies by a small amount, ΔV_{DD} , and Q_1 and Q_2 are replaced with their small-signal resistances. Note that $1/g_{m1} = 1/g_{m2}$ because Q_1 and Q_2 carry equal currents. If I_{D1} and I_{D2} change by ΔI_D , we have

$$\begin{aligned} \Delta V_Y - \Delta V_X &= \Delta I_D \left(R_1 + \frac{1}{g_{m2}} \right) - \Delta I_D \frac{1}{g_{m1}} \\ &= \Delta I_D R_1, \end{aligned} \quad (9)$$

and, hence, $\Delta V_P = A_1 \Delta I_D R_1$. In a well-designed circuit, we expect ΔI_D to be small and $V_{GS1,2}$ to be relatively constant, which predicts that $\Delta V_P \approx \Delta V_{DD}$. It follows that

$$\Delta I_D \approx \frac{\Delta V_{DD}}{A_1 R_1}. \quad (11)$$

We now ask, which quantity is the "output" of interest here? Since the drain current of M_1 and M_2 is eventually copied and applied to a resistor [e.g., R_L in Figure 2(c)] to generate the reference voltage, we define the PSRR as

$$\text{PSRR} = \frac{\Delta V_{DD}}{\Delta I_D R_L} \quad (12)$$

$$\approx \frac{A_1 R_1}{R_L}. \quad (13)$$

Moreover, if R_1 sustains a voltage of $V_T \ln n \approx 72 \text{ mV}$ and R_L an output voltage of 500 mV, we have $R_1/R_L = 0.14$. It follows that

$$\text{PSRR} = 0.14 A_1. \quad (14)$$

For 40 dB of rejection, A_1 must exceed 700. In practice, the PSRR is plotted as the inverse of the previous quantities, i.e., as the magnitude of the transfer function from V_{DD} to the output of interest.

For initial PSRR simulations, we simply multiply the voltage variation across R_1 by $1/0.14$, arriving at the plot presented in Figure 4(b). For supply-perturbation frequencies up to tens of megahertz, the PSRR is around -23 dB, which agrees with (14). At higher frequencies, $C_{GS1} + C_{GS2}$ in Figure 4(a) couples the V_{DD} changes to C_{GD1} and C_{GD2} , causing V_X and V_Y to bounce. The PSRR

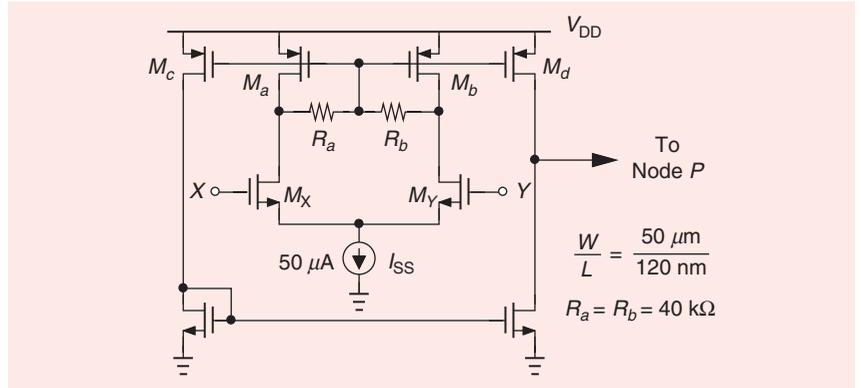


FIGURE 9: A two-stage op amp for use in the bandgap reference.

is far short of the desired value, necessitating further design efforts.

Op Amp Design

Since the op amp in Figure 3(a) must operate with input CM levels as high as 780 mV, we select an NMOS input stage for it. The simplest implementation is a five-transistor operational transconductance amplifier (OTA), as presented in Figure 5. We assume $W/L = 50 \mu\text{m}/120 \text{ nm}$ for all of the transistors. With a tail current of $50 \mu\text{A}$, the op amp provides a gain of about 20, and M_X and M_Y exhibit a minimum source voltage of 350 mV at $T = 100^\circ\text{C}$, which is sufficient for I_{SS} . However, at $T = 0^\circ\text{C}$, both $|V_{BE}|$ and $|V_{TH1,2}|$ take on large values, possibly pushing M_X into the triode region and lowering the op amp gain. Figure 6 plots V_X , V_A , and V_P versus T , demonstrating that $V_X - V_A$ keeps M_X in saturation. Figure 7

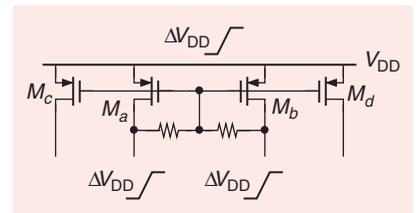


FIGURE 10: Paths from V_{DD} to the internal nodes of the two-stage op amp.

presents how the PSRR responses at $T = 0^\circ\text{C}$, 50°C , and 100°C illustrate a degradation at low temperatures.

An interesting observation in Figure 7 is that the low-frequency PSRR is around -35 dB at $T = 0^\circ\text{C}$, whereas (14) would yield $1/(0.14 A_1) \cong -9 \text{ dB}$ for $A_1 = 20$. Why is the performance better than expected? In the analysis leading to (14), we assumed that the op amp must multiply $V_Y - V_X$ by A_1 to adjust V_P and allow it to track V_{DD} . However, in the circuit of Figure 5, the OTA provides an additional

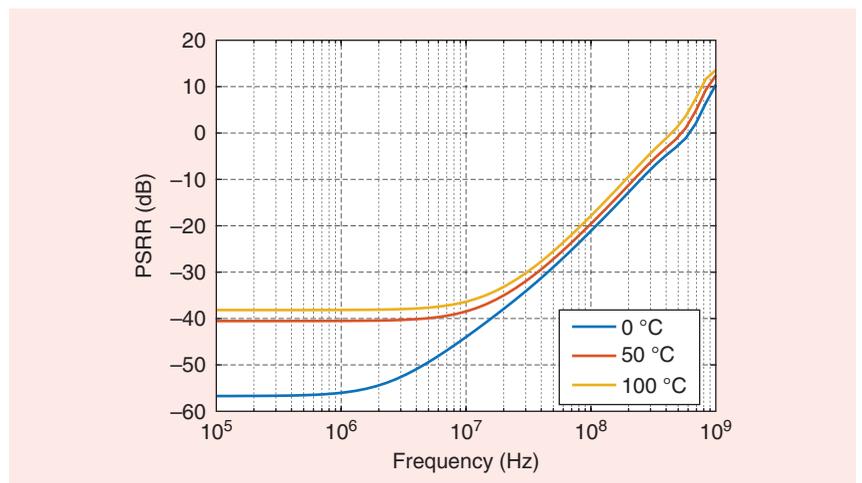


FIGURE 11: The PSRR of the core with a two-stage op amp.

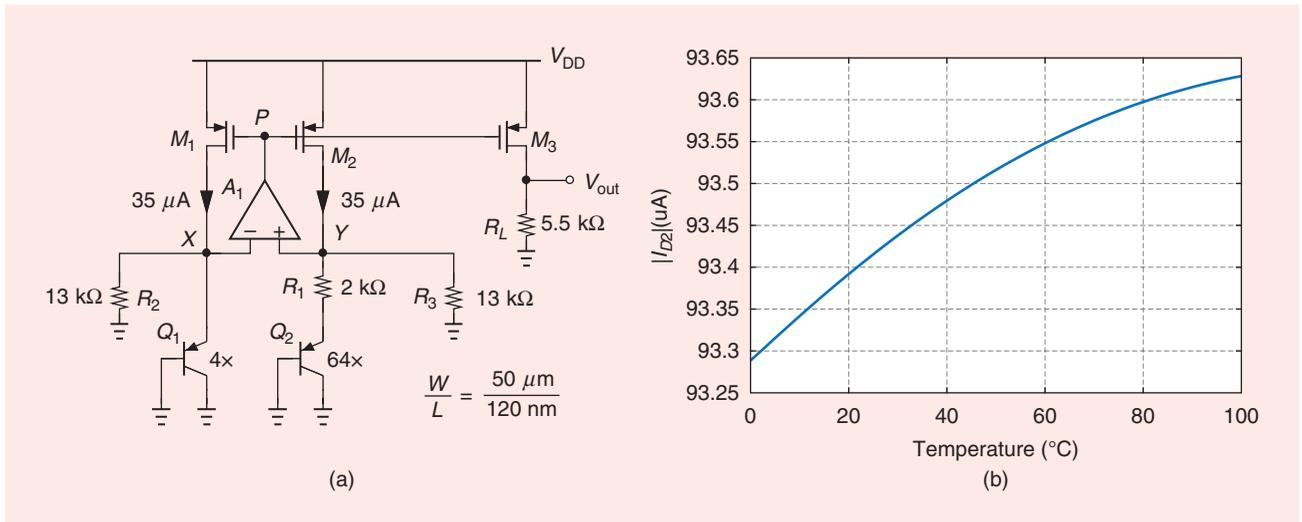


FIGURE 12: (a) The bandgap reference with R_2 and R_3 added and (b) the drain current of M_2 versus T .

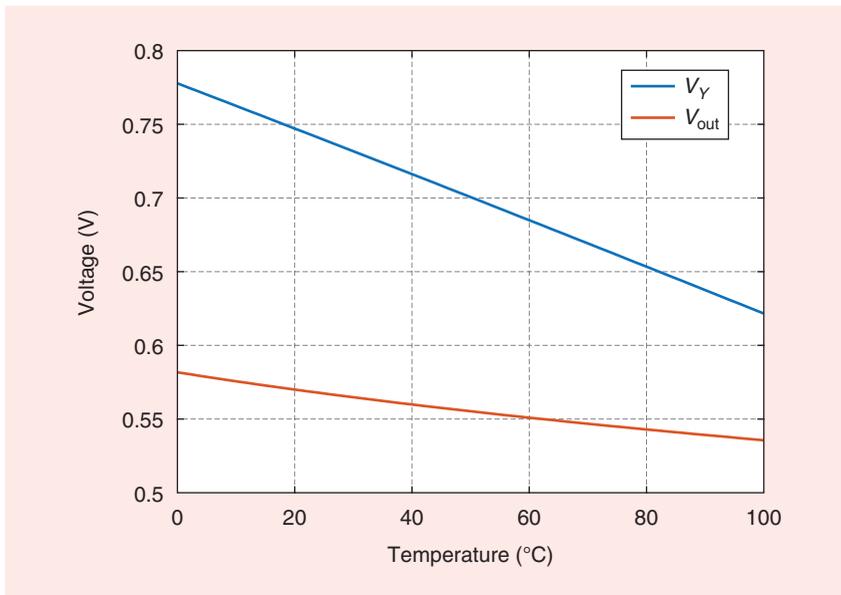


FIGURE 13: The voltages at node Y and at the output of the bandgap circuit versus T .

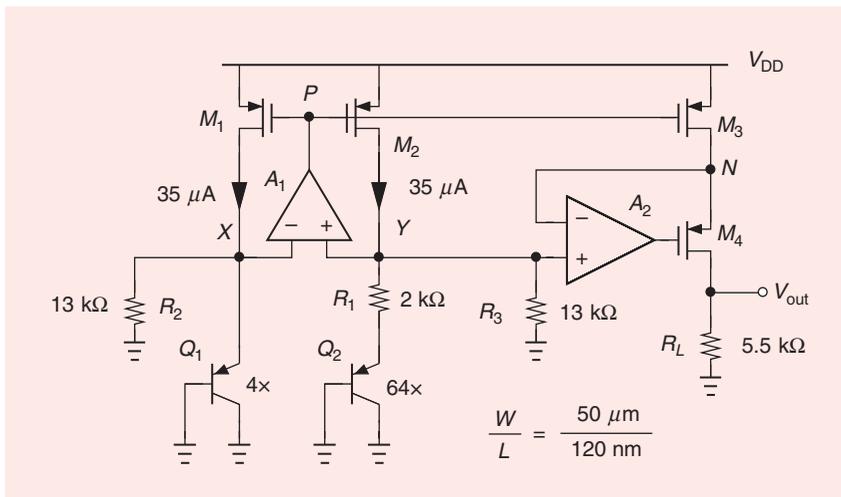


FIGURE 14: The bandgap circuit with a regulated cascode.

tracking mechanism: If V_{DD} varies by ΔV_{DD} , so do V_A and V_P (Figure 8). In essence, the OTA's PMOS active load bootstraps P to V_{DD} .

That $\Delta V_P \approx \Delta V_A$ can be seen by noting that, in the absence of asymmetries within the op amp, $V_P = V_A$ if $V_X = V_Y$. This property is generally considered a drawback of the five-transistor OTA, but it proves useful here. Since the V_{DD} perturbation is directly applied to node P by M_a and M_b , the op amp provides only additional correction.

The PSRR degradation at low temperatures calls for a higher op amp gain and, hence, a two-stage topology. Figure 9 illustrates a simple design wherein the output CM level of the first stage is set by R_a and R_b to be equal to one PMOS source-gate voltage below V_{DD} . This method also defines the bias currents of the second stage by forming current mirrors. Thus, the total supply current is about 100 μ A. This op amp offers a gain of 380 at $T = 0^{\circ}\text{C}$ and 320 at $T = 100^{\circ}\text{C}$, improving the PSRR according to (14).

Does the effect illustrated in Figure 8 exist in the two-stage op amp as well? If $V_X \approx V_Y$ and V_{DD} changes by ΔV_{DD} , then so do the drain voltages of M_a and M_b (Figure 10). We observe that the gate-source voltages of M_c and M_d remain relatively constant, introducing little change at their drains and, hence, in V_P . In other

words, this op amp does not provide the direct cancellation mechanism of the five-transistor OTA. Figure 11 plots the bandgap core PSRR in the presence of the two-stage op amp.

Complete Bandgap Reference

As prescribed by Figure 2(b), we must attach equal resistors from nodes X and Y to the ground to change the nature of I_{D1} and I_{D2} from PTAT to temperature-independent quantities. This requires that $(R_3/R_1)V_T \ln n$ be around $17V_T$ and $R_2 = R_3 \approx 6R_1$ if $n=16$. Figure 12 displays the modified circuit as well as $|I_{D2}|$ versus T . Resistors R_2 and R_3 are rounded up to $13\text{ k}\Omega$, and the op amp is implemented as the two-stage topology of Figure 9. We note that $|I_{D2}|$ changes by less than 0.4% across our temperature range of interest.

In the next step, we copy I_{D2} and apply the result to R_L , forming the reference voltage, V_{out} . According to (6), an output voltage of 0.5 V requires $R_L/R_2 \approx 0.42$ because the quantity within the parenthesis is around 1.2 V. It follows that $R_L = 5.5\text{ k}\Omega$. Plotted in Figure 13 is V_{out} versus T , exhibiting a variation of 40 mV.

Why does V_{out} drift so much even though I_{D2} is fairly constant? This error arises from the temperature-dependent difference between V_{DS2} and V_{DS3} and the channel-length modulation of M_2 and M_3 . From Figure 13, we note that $V_Y - V_{out}$ is equal to 200 mV at $T=0^\circ\text{C}$ and 85 mV at $T=100^\circ\text{C}$. The relatively long transistor channels still prove inadequate in obtaining an acceptably small current mismatch between M_2 and M_3 .

The error due to channel-length modulation is suppressed if we guarantee that the drain voltage of M_3 tracks that of M_2 . This can be accomplished by a regulated cascode structure. Figure 14 illustrates the idea of comparing these voltages by means of op amp A_2 and adjusting the gate voltage of M_4 accordingly. As V_Y falls with T , so does V_N , leaving less voltage headroom for M_4 and requiring that its overdrive voltage increase. Even if M_4 operates in

the triode region (e.g., at high temperatures), the loop gain provided by A_2 still ensures that $V_N \approx V_Y$. Op amp A_2 is realized as in Figure 9 but with $W/L = 25\ \mu\text{m}/120\ \text{nm}$ for all of the transistors. Figure 15 plots V_{out} versus T , revealing a variation of about 2.5 mV. The total supply current is around 0.5 mA. We have therefore met all of the specifications except for the supply rejection.

Figure 16 depicts the PSRR. Owing to the high op amp gain, the low-frequency value satisfies our target. Beyond 10 MHz, however, the PSRR degrades because the gain of A_1 in Figure 14 begins to fall. This is expected as the op amp's low-bias currents yield a high output resistance, about 45 k Ω , which, along

with $C_{GS1} + C_{GS2} + C_{GS3} = 0.35\ \text{pF}$, creates an open-loop pole in the vicinity of 10 MHz. To improve the PSRR, we add a simple low-pass filter to the output node. Figure 17 depicts the filter and the resulting PSRR.

Output Noise and Offset

In most applications, the noise of bandgap references proves critical. A circuit using I_{D2} or V_{out} in Figure 14 as a reference can experience performance degradation due to their noise. Figure 18 plots the noise voltage in V_{out} . At frequencies up to several hundred megahertz, it is dominated by the flicker noise of M_2 and M_3 . At 1 GHz, these devices and the first stage of the op amp contribute significant thermal noise.

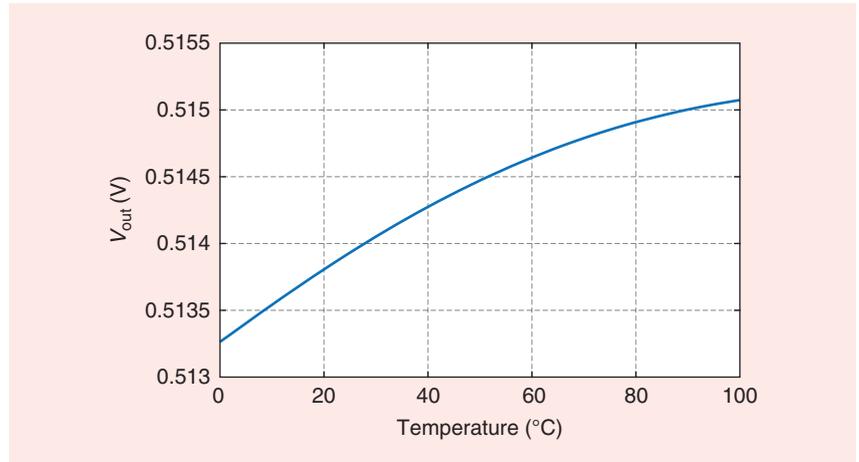


FIGURE 15: The output voltage of the final design versus T .

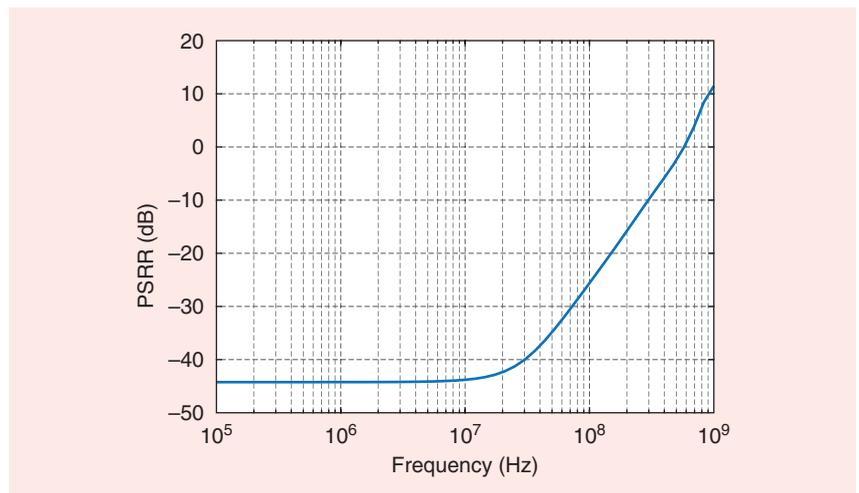


FIGURE 16: The PSRR of the final design.

If the output noise is unacceptably high for a given application, we can pursue three methods to reduce it. First, to lower the flicker noise, the channel areas of M_1 – M_3 can be increased while maintaining their W/L ratio. Second, the widths and bias currents of M_1 – M_3 and the areas of Q_1 and Q_2 can be scaled up, and the values of all of the resistors can

be scaled down by the same factor to reduce the output thermal noise. This remedy trades noise for area and power. Third, the output low-pass filter can incorporate larger capacitors, but at the cost of area.

The op amp offset, V_{OS} , arises primarily from the first stage in Figure 9. Writing the threshold mismatch as $\Delta V_{TH} = A_{VTH}/\sqrt{WL}$ and assuming

$A_{VTH} \approx 3\text{ mV}$, we obtain $V_{OS} \approx 1.7\text{ mV}$. From (7), this translates to an error of about 5 mV in V_{out} , which is a reasonable amount.

Start-Up and Transient Response

To ensure that the reference generator of Figure 14 reaches the desired state when the circuit is turned on, we must examine the initial behavior of the core and the two-stage op amp. Suppose V_X and V_Y are zero at powerup. Then, M_X , M_Y , M_a , and M_b in Figure 9 remain off, and so do M_c and M_d . We recognize that node P floats, possibly keeping M_1 and M_2 off as well. This degenerate state can be avoided if we add a means to prohibit V_P from staying high when the circuit turns on.

Whether a bandgap remains off or not depends on a number of factors. Capacitive coupling paths from V_{DD} to the internal nodes, e.g., to X and Y in Figure 14, can turn on the circuit. Also, the slope of the V_{DD} ramp and the temperature may encourage or impede the start-up process.

To ensure start-up, we can employ a timing circuit to initially keep node P in Figure 14 low. Figure 19(a) illustrates the idea of drawing a current from P by M_S as node G tracks the V_{DD} ramp and exceeds the transistor threshold. After V_{DD} stabilizes, V_G returns to zero. The drawback of this approach is that if V_{DD} takes, for example, 1 ms to ramp up, then R_1 and C_1 must be extremely large to permit G to go high.

We instead explore a different line of thought: if the circuit remains inactive after V_{DD} rises, then V_{out} in Figure 14 is zero and can be compared to a reference roughly representing its desired value. The result can then enable a mechanism to draw current from P . As depicted in Figure 19(b), amplifier A_3 compares V_{out} to V_r and turns on M_S if the former is well below the latter. The amplifier is implemented as the OTA displayed in Figure 5, except W/L is chosen to be equal to $5\text{ }\mu\text{m}/30\text{ nm}$ for all of the transistors. To accommodate the offset of A_3 , we select

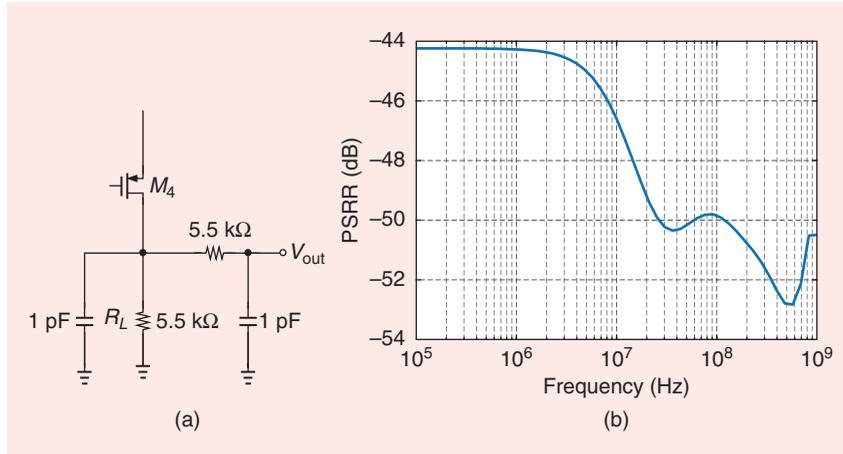


FIGURE 17: (a) The addition of a low-pass filter and (b) the resulting PSRR.

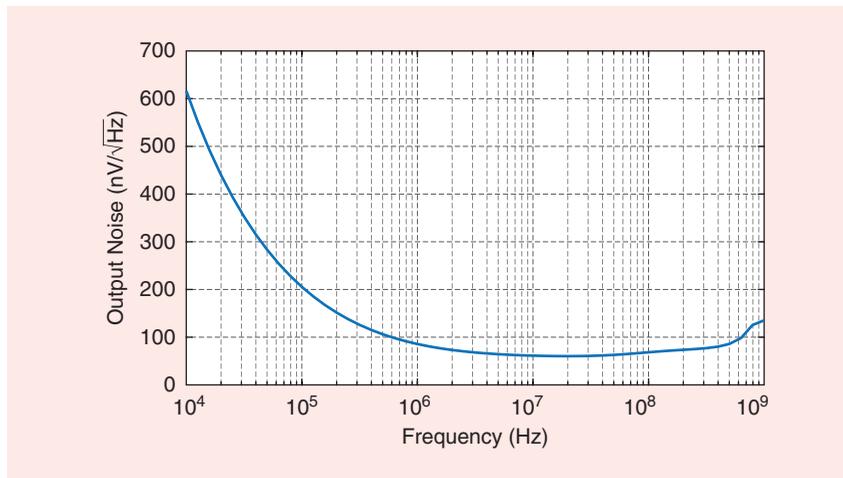


FIGURE 18: The output noise voltage of the bandgap reference.

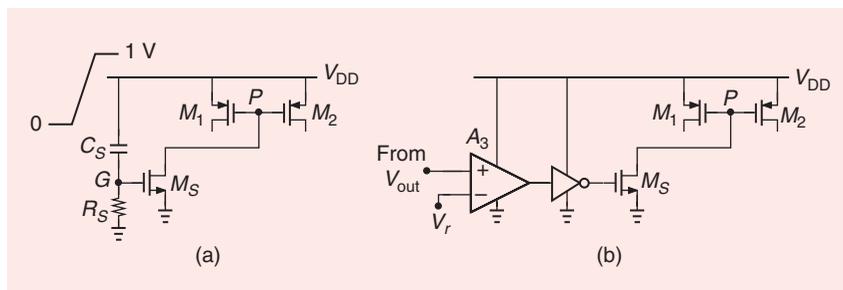


FIGURE 19: The start-up circuits using (a) a timing mechanism and (b) a high-gain comparison method.

(continued on p. 16)

of ac coupling attenuates the low-frequency component of the signal, causing baseline wander, which may result in the loss of data in wireline communication. To restore this low-frequency component for a random binary sequence or, equivalently, to remove the baseline wander, we employ a comparator with quantized feedback in conjunction with superposition.

Acknowledgment

I would like to thank Hossein Shakiba (the author of [2]) for his insights and discussions that inspired this article.

References

- [1] B. E. Boser, "Offset control," in *EECS 247: Lecture Notes 27*, Berkeley, CA: Univ. of California, 2002, pp. 1–15. [Online]. Available: <https://inst.eecs.berkeley.edu/~n247/fa07/lectures/L27.pdf>

- [2] M. H. Shakiba, "A 2.5Gb/s adaptive cable equalizer," in *Proc. Dig. Tech. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, Feb. 1999, pp. 396–397. doi: 10.1109/ISSCC.1999.759317.
- [3] F. Waldhauer "Quantized feedback in an experimental 280-Mb/s digital repeater for coaxial transmission," *IEEE Trans. Commun.*, vol. 22, no. 1, pp. 1–5, Jan. 1974. doi: 10.1109/TCOM.1974.1092055.
- [4] A. Sheikholeslami, "Circuit intuitions: Equalizer circuit," *IEEE Solid State Circuits Mag.*, vol. 12, no. 1, pp. 6–7, Winter 2020. doi: 10.1109/MSSC.2019.2952233.

SSC

THE ANALOG MIND (continued from p. 12)

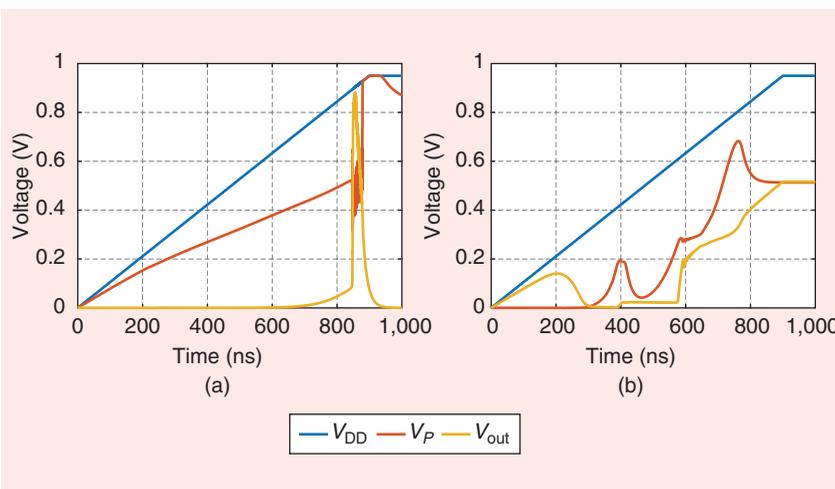


FIGURE 20: The bandgap internal waveforms (a) without and (b) with the start-up circuit.

$V_r \approx 0.4\text{ V}$ and generate it from V_{DD} by a resistive divider. Figure 20(a) and (b) plots V_{DD} , V_P , and V_{out} for a 900-ns V_{DD} ramp before and after the start-up circuit is added, respectively. We note that the former fails (in fact, it oscillates with a period of several microseconds) whereas the latter settles properly. For a 1-ms V_{DD} ramp, the bandgap turns on with or without the start-up mechanism.

References

- [1] R. J. Widlar, "Some circuit design techniques for linear integrated circuits," *IEEE Trans. Circuit Theory*, vol. 12, no. 4, pp. 586–590, Dec. 1965. doi: 10.1109/TCT.1965.1082512.
- [2] R. J. Widlar, "New developments in IC voltage regulators," *IEEE J. Solid-State Circuits*, vol. 6, no. 1, pp. 2–7, Feb. 1971. doi: 10.1109/JSSC.1971.1050151.
- [3] A. P. Brokaw, "A simple three-terminal IC bandgap reference," *IEEE J. Solid-State Circuits*, vol. 9, no. 6, pp. 388–393, Dec. 1974. doi: 10.1109/JSSC.1974.1050532.
- [4] R. A. Blauschild, P. A. Tucci, R. S. Muller, and R. G. Meyer, "A new NMOS temperature-

- stable voltage reference," *IEEE J. Solid-State Circuits*, vol. 13, no. 6, pp. 767–774, Dec. 1978. doi: 10.1109/JSSC.1978.1052048.
- [5] Y. P. Tsividis and R. W. Ulmer, "A CMOS voltage reference," *IEEE J. Solid-State Circuits*, vol. 13, no. 6, pp. 774–778, Dec. 1978. doi: 10.1109/JSSC.1978.1052049.
- [6] E. A. Vittoz and O. Neyroud, "A low-voltage CMOS bandgap reference," *IEEE J. Solid-State Circuits*, vol. 14, no. 3, pp. 573–577, June 1979. doi: 10.1109/JSSC.1979.1051218.
- [7] R. Gregorian, G. Wegner, and W. E. Nicholson, "An integrated single-chip PCM voice codec with filters," *IEEE J. Solid-State Circuits*, vol. 16, no. 4, pp. 322–333, Aug. 1981. doi: 10.1109/JSSC.1981.1051596.
- [8] K. E. Kujik, "A precision reference voltage source," *IEEE J. Solid-State Circuits*, vol. 8, pp. 222–226, June 1973. doi: 10.1109/JSSC.1973.1050378.
- [9] H. Banba et al., "A CMOS bandgap reference circuit with Sub-1-V operation," *IEEE J. Solid-State Circuits*, vol. 34, no. 5, pp. 670–674, May 1999. doi: 10.1109/4.760378.
- [10] C. J. B. Fayomi et al., "Sub-1-V CMOS bandgap reference design techniques: A survey," *Analog Integr. Circuits Signal Process.*, vol. 62, no. 2, pp. 141–157, Feb. 2010. doi: 10.1007/s10470-009-9352-4.
- [11] H. Neuteboom, B. M. J. Kup, and M. Janssens, "A DSP-based hearing instrument IC," *IEEE J. Solid-State Circuits*, vol. 32, no. 11, pp. 1790–1806, Nov. 1997. doi: 10.1109/4.641702.
- [12] B. Razavi, "The bandgap reference," *IEEE Solid State Circuits Mag.*, vol. 8, no. 3, pp. 9–12, Summer 2016. doi: 10.1109/MSSC.2016.2577978.
- [13] B. Razavi, *Design of Analog CMOS Integrated Circuits*, 2nd ed., New York, NY, USA: McGraw-Hill, 2017.

SSC