Appendix 3.1: Thermal Analysis

In Part 3, Sec. 3.4 we have discussed the problem of minimizing thermal distortion in wideband amplifiers. We have seen that the requirements of wideband design influence the choice of bias and consequently influence the thermal stability of the amplifier rather unfavorably. Whilst some aspects of good thermal stability and low thermal distortion can be solved by using differential amplifier configurations, biasing with constant current sources, and employing DC feedback techniques, there are still problems regarding the system's behavior under overdrive and overload conditions or the exposure to extreme ambient temperatures. Although the technology development offered us FETs and MOSFETs, devices with a negative thermal coefficient (drain current decreasing with increasing temperature), the bipolar technology is still used for top-class performance, mainly because of the recent low supply voltage trends (linearity and dynamic range can be a problem with devices having a relatively high V_{gs} threshold compared to V_{be}). But in spite of low supply voltages, really fast devices are tiny and can run seriously hot even at 1 mA current. With the number of transistors within a package increasing at an ever faster rate, it is not uncommon to find ICs with a nominal working temperature of 60°C, and junction temperatures of 120°C and more.

Here we are going to discuss some aspects of predicting the thermal stability, first after the circuit is powered up, and then upon transient overdrive. Consider the example of a simple single transistor amplifier, Fig. A3.1.1a. Let us first analyse its electrical conditions which result in the device's idle power dissipation, modeling the transistor's thermal behavior by a thermo-electric circuit analogy such as the one shown in Fig. A3.1.1b. Upon the result obtained we shall analyse the thermal dynamics after a transient overdrive and try to predict the maximum biasing level which would still prevent thermal runaway after the overdrive.

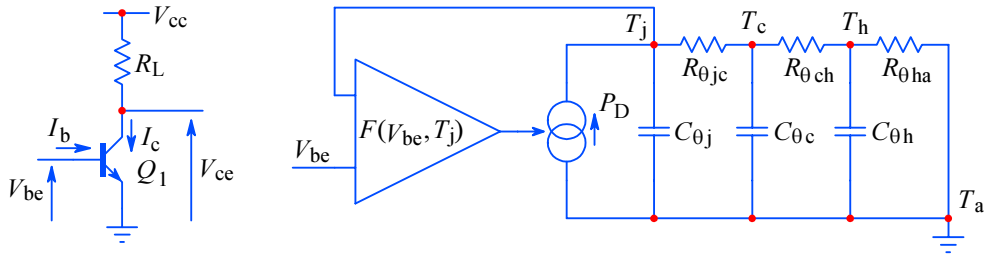

Fig. A3.1.1: **a)** A simple transistor amplifier. The electrical parameters define the power dissipation and the device's temperature increase above ambient temperature. **b)** The circuit's thermo-electric analogy is modeled by the transfer function $F(V_{be}, T_j)$, which determines the power dissipation P_D. Power dissipation is modeled as energy flow ('current'). The $R_{\theta i}C_{\theta i}$ are thermal (not electrical!) time constants: $C_{\theta j}$ is the thermal capacitance of the semiconductor's die, $R_{\theta jc}$ is the thermal resistance between the semiconductor junction and the transistor case, $C_{\theta c}$ is the thermal capacitance of the case, $R_{\theta ch}$ is the thermal resistance from the case to the heatsink (if there is one), $C_{\theta h}$ is the heatsink's thermal capacitance and $R_{\theta ha}$ is the thermal resistance from the heatsink to the ambient. T_j, T_c, T_h, and T_a are the junction, case, heatsink and ambient (air) temperatures, respectively, all modeled as 'potentials' referenced to the 'ground' T_a. P_D is in [W], $R_{\theta i}$ are in [K/W], $C_{\theta i}$ are in [Ws/K], and T_i are in [K].

The electrical parameters which set the Q_1 power dissipation are its power supply voltage V_{cc}, its collector load resistance R_L, and the bias dependent working point, denoted by the collector to emitter voltage V_{ce} and the collector current I_c; the power dissipated by the device is simply the product of its voltage and current:

$$P_D = V_{ce} I_c = (V_{cc} - I_c R_L)\, I_c = V_{cc} I_c - I_c^2 R_L \qquad \text{(A3.1.1)}$$

This means that the power dissipated by the device can be calculated as the difference between the total power supplied by the power source, $V_{cc} I_c$, and the power delivered to the load, $I_c^2 R_L$.

Eq. A3.1.1 is an inverted parabolic (quadratic) function (shown in Fig. A3.1.2), having a peak at $V_{ce} = V_{cc}/2$, where $I_c = I_{max}/2 = V_{cc}/2R_L$, and decreasing both below and above this point.

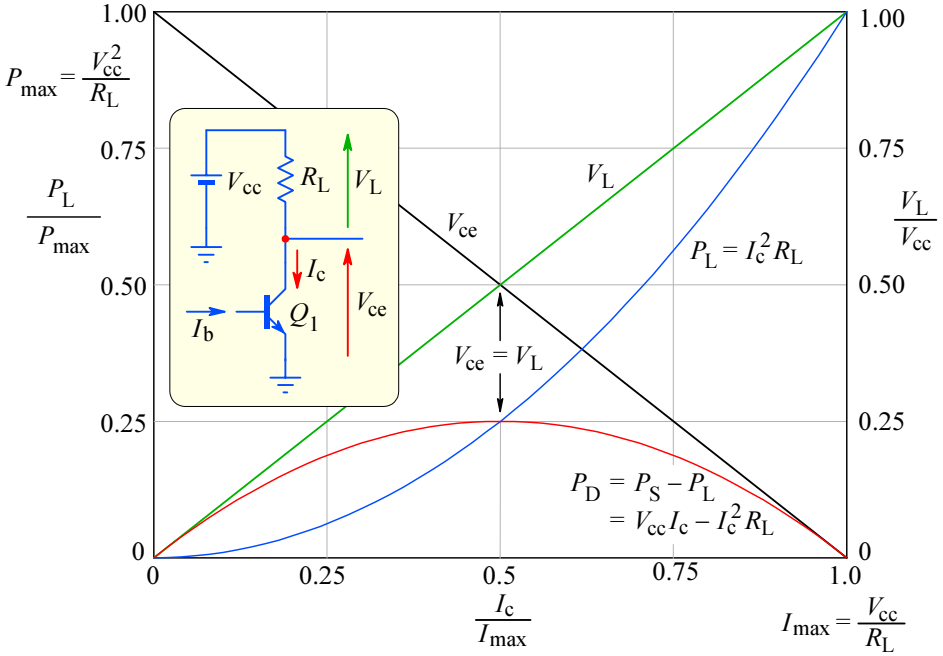

Fig. A3.1.2: The load power P_L and the transistor's dissipated power P_D as functions of the collector current I_c. The dissipated power P_D can be calculated as the difference between the total power supplied from the power source, $P_S = V_{cc} I_c$, and the power delivered to the load, $P_L = I_c^2 R_L$.

As we have discussed in Part 3, we would prefer to bias the transistor at the maximum of the power dissipation, which would maximize the available dynamic range, and at the same time minimize the thermal problems, since the top of the parabola is essentially flat, resulting in very little change of the device's temperature with small signals.

Unfortunately, in order to achieve a high bandwidth we need a low value of R_L, a high V_{cb} (for a low C_μ), and a high I_c (for a low r_e, and for the ability to drive a low impedance load). This means that our transistors will usually be biased by an $I_c \approx 0.25\, I_{max}$. Also, the base would usually be voltage driven (from a low impedance signal source). All those conditions will reflect unfavorably on the thermal stability.

We start our analysis with the well known relation between the base-emitter voltage V_{be} and the collector current I_c, and we are going to examine the dependence

of this relation with device's junction temperature T_j. The base voltage, which gives the required collector current, is:

$$V_{be} = \frac{k_B T_j}{q_e} \ln\left(\frac{I_c}{I_s} - 1\right) \tag{A3.1.2}$$

This equation sets the voltage–biasing condition, which is potentially unstable owed to two factors: a highly nonlinear I_c to V_{be} relationship, and a **negative** temperature coefficient of V_{be} (see Fig. A3.1.4). Of course, constant current biasing can be employed to enforce thermal stability by design, but in many cases we want the output voltage to be independent of the load. In such cases current biasing would limit the output current, so thermal stability must be provided in a different way (usually by adding a small 'degeneration' resistor in the emitter).

At first glance, from Eq. A3.1.2 it would seem that for a given I_c the V_{be} should increase linearly with temperature. Not in an actual transistor, however, because I_s is also highly temperature-dependent. An often quoted figure is that I_s doubles every 8 K:

$$I_s = I_{sn} \cdot 2^{(T_j - T_n)/8} \tag{A3.1.3}$$

but in SPICE and similar circuit simulator programs, a more accurate model is used:

$$I_s = I_{sn} \left(\frac{T_j}{T_n}\right)^3 e^{\left[\frac{E_g q_e}{k_B}\left(\frac{1}{T_n} - \frac{1}{T_j}\right)\right]} \tag{A3.1.4}$$

In both equations I_{sn} is the nominal saturation current ($\approx 10^{-14}$ A for a typical silicon small signal transistor at 'room temperature' $T_n = 24 + 273$ K). The effective energy gap E_g in silicon is about 1.1 eV (1 eV = 1.602×10^{-19} VAs). In Fig. A3.1.3 the first approximation is shown by **a**, and the second approximation is shown by **b**:

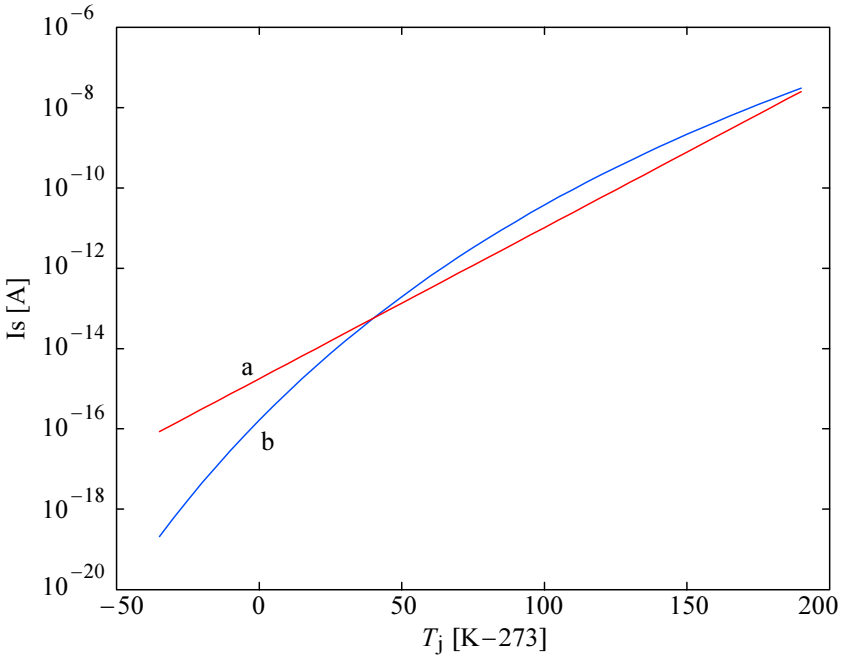

Fig. A3.1.3 : **a)** The transistor saturation current is temperature-dependent, often quoted as doubling every 8 K; **b)** a more accurate model (as in SPICE). Note that both models are empirical, based only partially on device physics.

By including the SPICE model for I_s into Eq. A3.1.2 we obtain the combined temperature dependence, as shown in Fig. A3.1.4, along with both of its components, each normalized to its own 'nominal' value at the room temperature. It is obvious that the influence of I_s dominates over the thermal voltage component V_T. As a result the complete temperature dependence of V_{be} can be approximated by the often quoted temperature coefficient of $\approx -2\,\mathrm{mV/K}$.

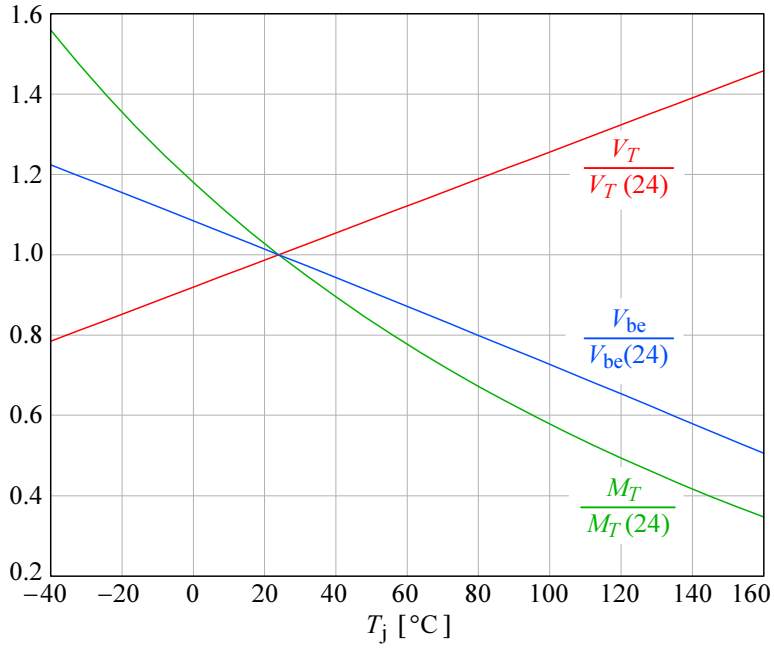

Fig. A3.1.4: The temperature dependence of the base–emitter voltage V_{be} of a typical small signal transistor for a given collector current (1 mA) is a function of two components. One is the 'thermal voltage' $V_T = k_B T_j / q_e$ which increases with temperature. The other is the natural logarithm of the collector to saturation current ratio $M_T = \ln(I_c/I_s - 1)$, which decreases with temperature because of I_s increase. Each function is shown as normalized to its own value at room temperature (24°C). Since $V_{be} = V_T\, M_T$, and, obviously, the temperature dependence of I_s dominates, the resulting negative temperature coefficient of V_{be} is $\approx -2\,\mathrm{mV/K}$. If the transistor's base is driven from a low impedance voltage source the collector current will increase with temperature, thus increasing also the dissipated power, increasing in turn the junction temperature even further. If not controlled, this positive feedback results in thermal runaway, leading eventually to the destruction of the device.

From Eq. A3.1.1, Eq. A3.1.2 and Eq. A3.1.4 we can derive the thermo-electric model transfer function, $F(V_{be}, T_j)$, which governs the power dissipation, P_D. To do so, we need to express the junction temperature T_j as a function of P_D. This we achieve by considering the various $R_{\theta i}$ and $C_{\theta i}$ shown in Fig. A3.1.1, which are driven by P_D, and by solving the transfer function of this thermal network.

In the thermo-electric model analogy of Fig. A3.1.1b the power dissipation P_D is represented by an energy ('current') flowing through the device's thermal resistances: the junction to case resistance $R_{\theta jc}$, the case to heatsink resistance $R_{\theta ch}$, and the heatsink to ambient resistance $R_{\theta ha}$.

In the usual electrical circuit a 1 A current flowing through a 1 Ω resistor produces a 1 V voltage drop. Since the dimension of $R_{\theta i}$ is *kelvin/watt* [K/W], in the thermo-electric circuit analogy the power dissipation of 1 W 'flowing' through a

thermal resistance of 1 K/W, would result in a temperature difference of 1 K. In this way we have a thermal equivalent of Ohm's law:

$$\Delta T = P_{\mathrm{D}}\, R_{\theta} \tag{A3.1.5}$$

The capacitances $C_{\theta\mathrm{j}}$, $C_{\theta\mathrm{c}}$ and $C_{\theta\mathrm{h}}$ represent the semiconductor die, the case, and the heatsink masses with their specific thermal capacitances, which depend on the materials which they are made of, as well as on their volume to surface ratio. The dimension of $C_{\theta\mathrm{i}}$ is *watt second/kelvin* [Ws/K]. In accordance with the thermo-electric circuit analogy, under transient conditions the thermal capacitances prevent their appropriate node temperatures from jumping instantly to the value determined by $R_{\theta\mathrm{i}}$, and instead rise (or fall) exponentially, just like a voltage at the node of an electrical RC network. For an increase in power dissipation:

$$\Delta T(t) = \Delta T(0)\left(1 - \mathrm{e}^{-t/R_{\theta}C_{\theta}}\right) \tag{A3.1.6}$$

and for a decrease:

$$\Delta T(t) = \Delta T(0)\,\mathrm{e}^{-t/R_{\theta}C_{\theta}} \tag{A3.1.7}$$

The ground potential is the ambient (air) temperature, T_{a}. If there is no heatsink we can simply short the nodes T_{h} and T_{a} and eliminate $R_{\theta\mathrm{ch}}$ and $C_{\theta\mathrm{h}}$. But note that now the case to air thermal resistance $R_{\theta\mathrm{ca}}$ is much higher than was the sum $R_{\theta\mathrm{ch}} + R_{\theta\mathrm{ha}}$. Note also that T_{a} inside the system box is often substantially higher then that of the air outside!

Owing to the exponential interdependence of T_{j} and P_{D}, we can not write an explicit equation as a solution; instead, we have to solve the system of equations iteratively, one after another. Or, as we are used to do in electrical circuits, we can write a set of differential equations and solve it in one of the usual ways.

But first we must calculate the electrical and thermal initial conditions. For a chosen DC operating point and assuming that the system is allowed enough time to stabilize, the semiconductor junction temperature T_{j} will be:

$$T_{\mathrm{j}} = T_{\mathrm{a}} + P_{\mathrm{D}}(R_{\theta\mathrm{ha}} + R_{\theta\mathrm{ch}} + R_{\theta\mathrm{jc}}) \tag{A3.1.8}$$

However, owed to the presence of thermal capacitances, T_{j} follows P_{D} with a certain time delay.

Let us say that we would like to solve the thermal system numerically. We write a set of differential equations which we shall integrate by executing an iterative loop in small time increments. In order to achieve a low integration error Δt must be a small fraction of the smallest time constant in the system. In the thermo-electric model analogy, we are using the well known equations $dV/dt = i/C$ and $i = V/R$, in which we have substituted V by T, and i by P_{D}, whilst R and C are substituted by their thermal equivalents, $R_{\theta\mathrm{i}}$ and $C_{\theta\mathrm{i}}$. In this way the time variable comes into the equation through the thermal time constants.

There are three equations, one for each node. The first equation represents the state at the semiconductor junction node, the second is for the case node, and the last is for the heatsink node:

$$P_{\mathrm{D}} = C_{\theta\mathrm{j}} \frac{\Delta T_{\mathrm{j}}}{\Delta t} + \frac{T_{\mathrm{j}} - T_{\mathrm{c}}}{R_{\theta\mathrm{jc}}} \tag{A3.1.9}$$

$$\frac{T_{\mathrm{j}} - T_{\mathrm{c}}}{R_{\theta\mathrm{jc}}} = C_{\theta\mathrm{c}} \frac{\Delta T_{\mathrm{c}}}{\Delta t} + \frac{T_{\mathrm{c}} - T_{\mathrm{h}}}{R_{\theta\mathrm{ch}}} \tag{A3.1.10}$$

$$\frac{T_{\mathrm{c}} - T_{\mathrm{h}}}{R_{\theta\mathrm{ch}}} = C_{\theta\mathrm{h}} \frac{\Delta T_{\mathrm{h}}}{\Delta t} + \frac{T_{\mathrm{h}} - T_{\mathrm{a}}}{R_{\theta\mathrm{ha}}} \tag{A3.1.11}$$

Here each temperature T_{i} is the node temperature at the current time instant, and each ΔT is the difference between the temperature values at the present and the previous instant. If Δt is small ΔT will also be small compared to the temperature difference across each thermal resistance, so we can assume a quasi-static situation at thermal resistances and account for the temperature change only at the next instant.

The only problem that remains now is to find the values of time constants of the actual system, and other parameters. The electrical parameters can be found in the transistor data sheets supplied by the manufacturer. Most manufacturers provide the thermal parameters for each type of case; these are sometimes given in a separate section of the product catalog.

For a silicon transistor with the die size 3×3×0.5 mm enclosed in a TO-5 metal case the typical values are:

$$C_{\theta\mathrm{j}} = 0.004\,\mathrm{Ws/K} \qquad R_{\theta\mathrm{jc}} = 30\,\mathrm{K/W} \qquad C_{\theta\mathrm{c}} = 0.06\,\mathrm{Ws/K} \tag{A3.1.12}$$

For a star-shaped black anodized aluminum heatsink for the TO-5 case the typical values are:

$$R_{\theta\mathrm{ch}} = 0.5\,\mathrm{K/W} \qquad C_{\theta\mathrm{h}} = 1.5\,\mathrm{Ws/K} \qquad R_{\theta\mathrm{ha}} = 4.5\,\mathrm{K/W} \tag{A3.1.13}$$

An attentive reader might ask why is it important to account for the $C_{\theta\mathrm{j}}$ when its value is so low. True, a high P_{D} step would cause T_{j} to increase almost instantly. But after the thermal shock $C_{\theta\mathrm{j}}$ 'discharges' (cools down) through a relatively high $R_{\theta\mathrm{jc}}$ and this time constant is comparable to the discharge time of $C_{\theta\mathrm{c}}$. Even if the thermal shock (owed to the signal) is short the large thermal time constant keeps the junction temperature high. So when the next signal pulse arrives (even if of a lower value) it could cause overheating and a consequent thermal runaway. An adequate cooling rate is therefore vital for regaining thermal balance.

To understand this we must realize that the system cooling is a linear function of temperature ($\Delta T_{\mathrm{ca}}/R_{\theta\mathrm{ca}}$), whilst heating is exponential, as shown in Fig. A3.1.5. Those two functions have two intercept points: the lower one is the point of absolute thermal stability, defined by the chosen DC operating point, whilst the upper one is the critical point. After power up the device's temperature is low and the heating power is higher than cooling, so the temperature rises towards the stable point. If a signal pulse moderately increases the temperature, the device will cool down to the stable point again once the signal has vanished. But once the device's junction was

heated by the signal beyond the critical point, the device produces more heat by its own bias alone than it is capable of dissipating, and thermal runaway is inevitable.

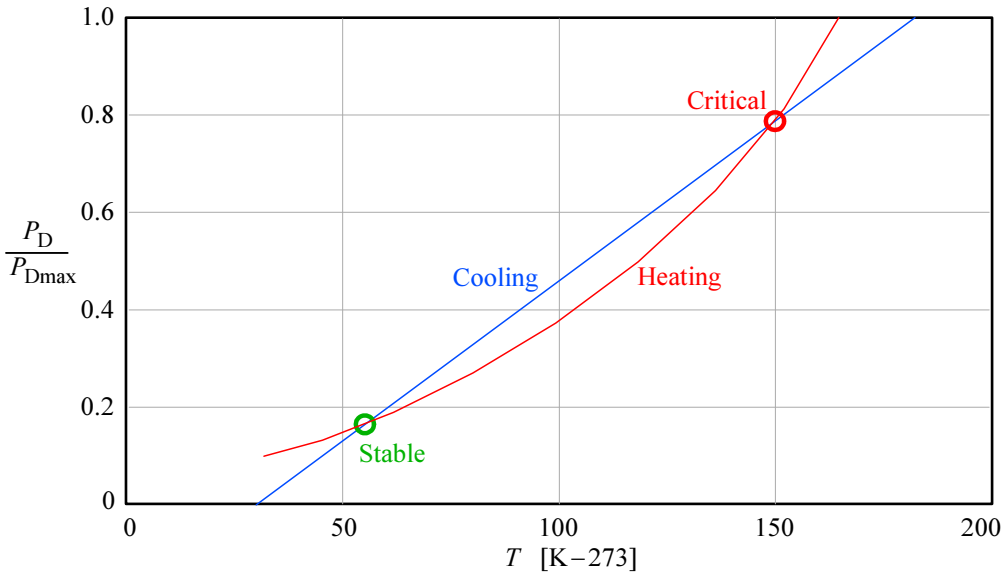

Fig. A3.1.5: The cooling is a linear function of temperature; the heating is exponential. The lower intercept point is thermally stable, the upper is the critical point above which the system's self-heating is higher than its cooling ability, therefore it undergoes a thermal runaway. The vertical scale is the power produced and dissipated, normalized to the power with which the system would reach the maximum allowable temperature (180+273 K for a silicon transistor).

Unfortunately, in most manufacturers' data sheets only the thermal resistances are given, so we are forced to find the thermal capacitance values either experimentally, by measuring the thermal time constants, or by digging deep into manuals on material properties.

Note that for some transistors in plastic cases the total junction to air thermal resistance $R_{\theta ja}$ can be as high as 140 K/W! Metal cases have much lower thermal resistance, about 35 K/W for a typical TO-5 case, and even less with a heatsink; but for a wideband design a heatsink would add too much (electrical) capacitance to the collector circuit, since the collector is often in contact with the metal case. It is therefore a matter of judicious design balance, whether we run the transistor hot to achieve low r_e (with high I_c) and low C_μ (with high V_{cb}) and then risk to spoil everything with some 30 pF of the heatsink stray capacitance, or do we rather settle for lower current and voltage and improve the bandwidth by low stray capacitance.

In modern circuits with surface mounted discrete components, the problem is worse still, since the transistor epoxy case has a high thermal resistance. A four-layer printed circuit board with wide ground-plane and power-plane area on inner layers can serve as a good heatsink, but then the stray capacitance can then be high, too, therefore large etched areas around critical nodes and properly terminated transmission line interconnections are mandatory.

In integrated circuits the thermal problem can be a nightmare, since there are hundreds of transistors packed close together in a plastic case; not all of them will be required to transfer the signal at full bandwidth, but some of them could be running hot, generating high thermal gradients across the chip. Large thermal pads, or a large

number of pins connected to ground can be used to lower R_θ. Instead of plastic, a ceramic case can be used if the system's performance can justify the cost increase.

It is time to make a simple example of the thermal runaway problem.

We have a transistor in a TO-5 case ($P_{\mathrm{Dmax}} = 0.5\,\mathrm{W}$ without a heatsink) loaded by $R_\mathrm{L} = 100\,\Omega$ and supplied by a DC voltage of $V_\mathrm{cc} = 15\,\mathrm{V}$ (the configuration is shown in Fig. A3.1.1a). Such an amplifier would have its dissipation parabola peak at a collector current $I_\mathrm{c} = 75\,\mathrm{mA}$ ($I_\mathrm{c}R_\mathrm{L} = V_\mathrm{cc}/2$); however, owing to the limit imposed by P_{Dmax} the maximum allowable current is only $I_\mathrm{c} = 50\,\mathrm{mA}$. If we are to drive this amplifier by a large signal we must ensure an adequate linear dynamic range, therefore the DC bias current should be lower still, say, $I_\mathrm{c} = 30\,\mathrm{mA}$. This should be low enough, since we do not want to drive the amplifier by more than some 20 % off the optimum bias level. Nevertheless, an occasional accidental overdrive might occur; if it exceeds the bias by some 60 % the system comes uncomfortably close to the maximum power dissipation.

In Fig. A3.1.6 we have plotted the junction and case temperature of such a transistor. Upon established initial DC bias, and consequently the initial temperatures, the transistor is driven by a signal pulse of different amplitudes, but lasting much shorter than the transistor's dominant thermal time constant, $R_\theta C_\theta$.

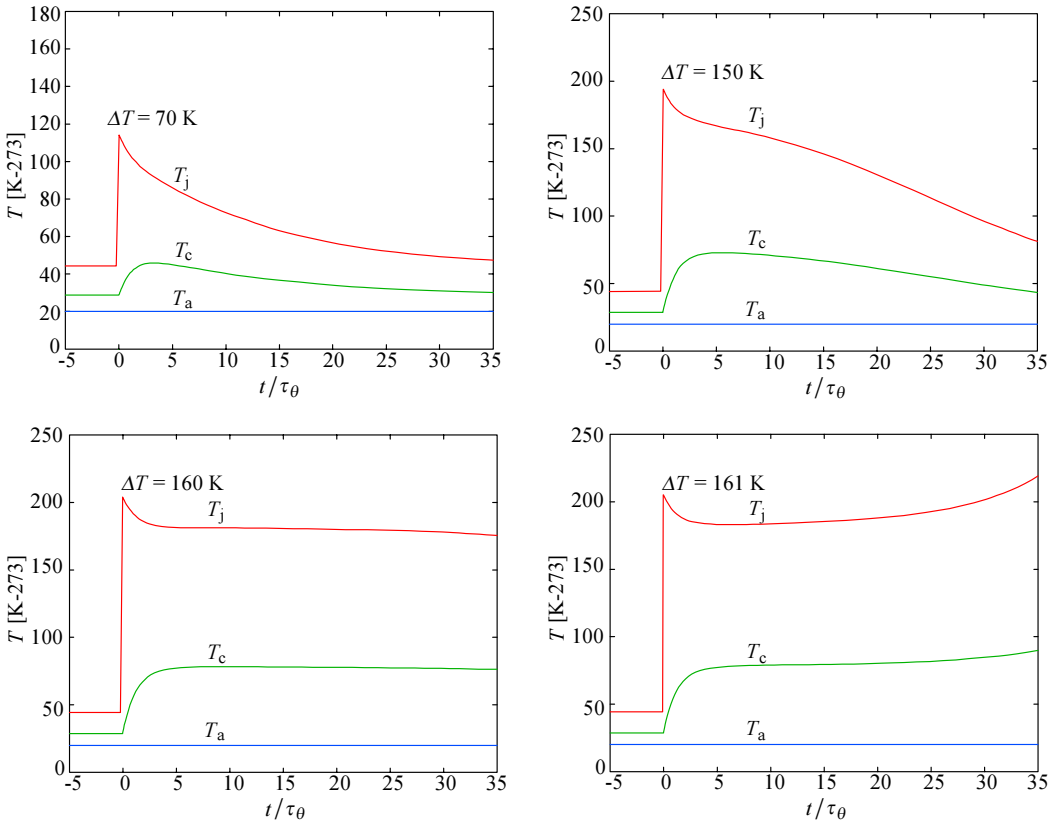

Fig. A3.1.6: The junction and case temperature of a transistor with a defined initial DC operating conditions, exposed to a short signal pulse (with a duration much shorter than the dominant thermal time constant τ_θ) applied at $t = 0$. If the resulting thermal shock is low, the device quickly regains its initial conditions. But when the thermal shock is high the recovery can last much longer. Close to the critical stability point, if the shock exceeds the limit by just a small margin the device will experience a thermal runaway. It is also evident that even a low thermal shock, repeated before the device recovers, will have the same consequences.

If the resulting thermal shock is low the device recovers in a relatively short time; note, however, the two different time constants (one short and one long) governing the junction temperature T_j. This becomes even more pronounced if the thermal shock is higher.

When approaching the critical stability point, the recovery time becomes very long, and also the system's sensitivity to temperature is very high: by exceeding the critical stability by only a very low margin, the device will experience a thermal runaway and be permanently damaged.

Thermal behavior of semiconductor devices can, of course, be also examined by circuit simulator programs. Note, however, that in SPICE and similar programs the temperature is a 'global' parameter, affecting all the components of the simulated circuit simultaneously. Although it can be adjusted by the user, or made to step within a selected range, it can not be made signal dependent, as it is in reality, without altering the device models. Recently exactly such a solution was described in [Ref. A3.1.1] with some very interesting results on modeling signal distortion in amplifiers.

Finally, we need to mention the effect of temperature on the dynamic properties of transistors, namely through the charge mobility parameters. Charge drift and diffusion are both manifestations of the random thermal motion of charge carriers. Consequently, the mobility μ of electrons and holes and their associated diffusion coefficients D are not independent:

$$\frac{D_h}{\mu_h} = \frac{D_e}{\mu_e} = \frac{k_B\, T}{q_e} \qquad \text{(A3.1.14)}$$

This is known as the Einstein's equation. It can be justified by considering the statistical quantum-mechanical implications of the energy equilibrium state in a non-uniformly doped semiconductor [Ref. A3.1.2]. This means that transistor's bandwidth will change with temperature.

References:

[A3.1.1] *C. Bateman* : Simulating Power MOSFETs,
Electronics World, Dec. 2004, pp. 10-20

[A3.1.2] *P.E. Gray* & *C.L. Searle*, Electronic Principles: Physics, Models and Circuits,
J. Wiley & Sons, 1969