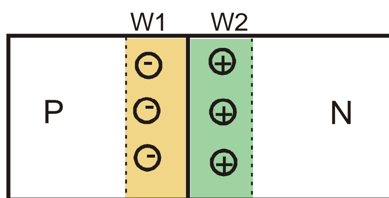
The Power Diode

What makes a difference between the usual P/N semiconductor diode and the power diode? The response is basically: the need to withstand high voltages in the reverse bias condition; this need will pose some constraints on the device structure.

All power devices that must operate at high reverse voltages present at least one P/N junction reverse biased, so the voltage requirements for the P/N junction of the power diode are equally shared by the other power devices, as we will see in the following.

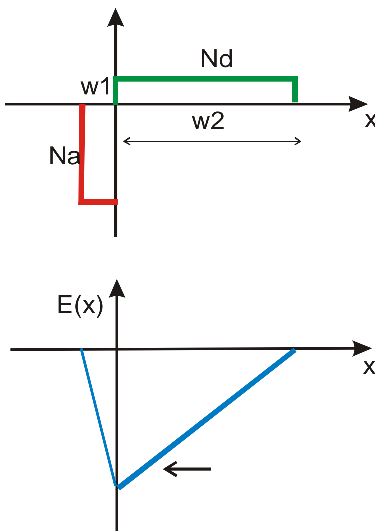
Let's briefly recall, for a P/N junction, the basic relationship for the depletion region and field profile in reverse bias.



the charge balance in the depleted regions W1 and W2 gives: q W1 Na = q W2 Nd, where Na is the concentration of acceptor atoms in P region, and Nd of the donor ones in N region. Then:

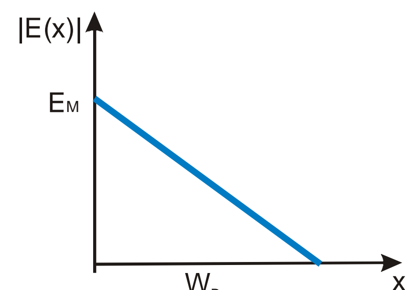
$$\frac{W1}{W2} = \frac{Nd}{Na}$$

If Na >>Nd (P^+N diode) W1 <<W2; then the total depletion width $W_D \cong W2$



Assuming constant doping profile in both P+ and N region (abrupt junction assumption), the field profile is linear, based on the Poisson equation:

$$\left.\frac{dE}{dx}\right|_{W1} = -\frac{q}{\varepsilon} N_A; \qquad \left.\frac{dE}{dx}\right|_{W2} = \frac{q}{\varepsilon} N_D$$



Assuming P+ >>N, the field profile lies in the low-doped side of the junction, and the max E_M value is dependent on the depletion layer width $W_D \cong W2$

$$\frac{E_M}{W_D} = \frac{q}{\varepsilon} N_D \quad \Rightarrow \quad E_M = \frac{q}{\varepsilon} N_D W_D$$

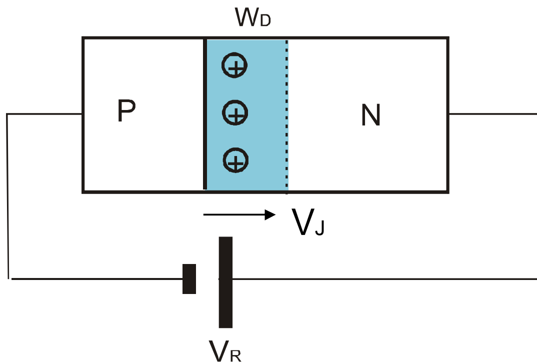
Recalling the expression of the built-in voltage V_{BI} with no external bias:

$$V_{BI} = \frac{KT}{q} \ln \frac{N_A N_D}{n_i^2}$$

and recalling that: $V = -\int E(x)dx \cong \frac{1}{2} E_M W_D$ we obtain: $W_D \cong \sqrt{\frac{2\varepsilon}{qN_D} V_{BI}}$

assuming: $N_A = 10^{18}\ cm^{-3}$
$N_D = 10^{15}\ cm^{-3}$

we obtain: $V_{BI} = 0.8$ V, $W_D = 0.1\ \mu m$

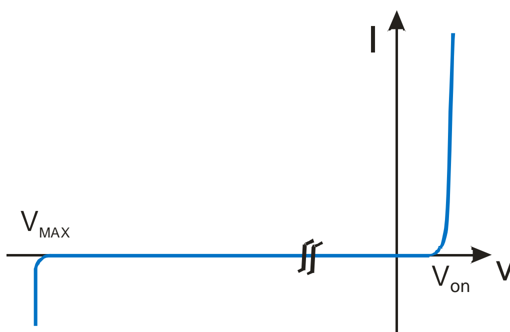


In reverse bias the junction voltage V_J is the sum of $V_{BI} + V_R$.
The built in voltage (about 1V) is negligible with respect to the reverse bias voltages applied, so one can approximate: $V_J \cong V_R$.

$$V_R \cong \frac{1}{2} E_M W_D$$

both the max field E_M and the depletion width W_D increase with the reverse voltage V_r

Breakdown voltage of a P/N junction

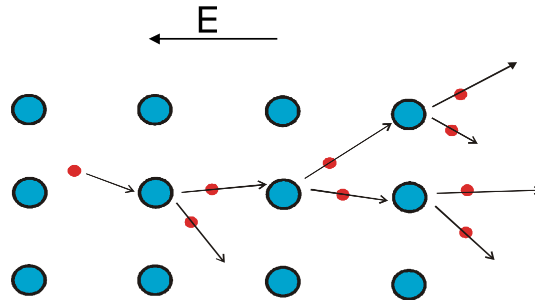


As it is well known, the IV characteristic of a PN junction has a current that depends exponentially from the applied voltage in an exponential vay in forward bias with a low voltage drop V_{ON}. In reverse bias it has a very low current, up to a max voltage V_{MAX} limited by avalanche multiplication, where the current sharply increases in almost unlimited way.

The max reverse voltage V_{MAX} of the junction is called the breakdown voltage V_{BR}, and puts a limit in the max reverse voltage that the junction can withstand.
We will now discuss the dependence of the breakdown voltage from the doping and width of the low doped region of the junction (remember that the field and voltage in reverse bias drops almost completely in the low doped region), to define the constraints posed on that region from the need of an high V_{MAX} in the range of hundreds or thousand of volts.

Avalanche multiplication

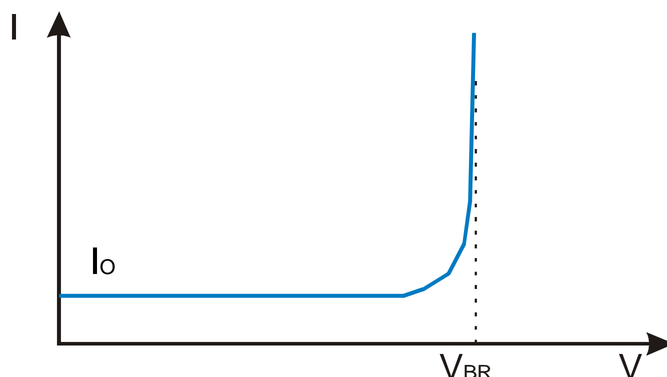
In the depleted region of the junction the mobile carriers travel under the drift effect the electric field E; their mobility is due to the collisions with Si atoms, but in the mean free path between each collision the carriers are accelerated by the field.
If the field value is so high to give to the mobile charge an energy above the energy gap E_G before the collision, the carrier that collides with an atom creates a secondary electron-hole pair generated by impact ionization.



Each secondary carrier can create again an electron-hole pair if that carrier can reach an energy higher than E_G in his mean free path. The result is a number of secondary carriers created by impact ionization that tends to infinite even if the number of primary carriers is very low. This is the effect of avalanche multiplication.

The current due to impact multiplication is usually indicated as $I = M\, I_O$, where I_O is the current when impact ionization is not present, and M is the multiplication factor due to impact ionization.
The factor M can vary from 1 infinite, depending from the percentage of carriers that, under the field E applied, reach an energy high enough to generate a secondary electron-hole pair by impact ionization. The factor M is a function of field E, and the breakdown voltage V_{BR} is reached when M= f(E) becomes infinite.

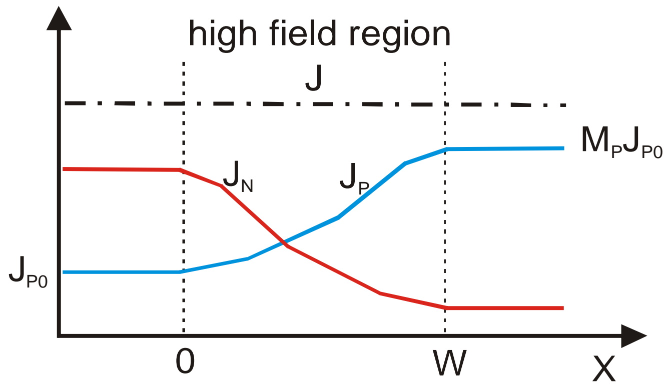
An empirical relation for the multiplication factor M is the Miller expression:



$$M \cong \frac{1}{1-\left(\frac{V}{V_{BR}}\right)^{N}}$$

with $N = 2 - 6$

Let's define more quantitatively the multiplication factor M from the basic transport equations



In the high field the primary carriers (both electrons and holes) moves with saturated velocity v_s , and generate secondary carriers. The generation term in the continuity equation becomes : $G = \alpha_p p\, v_s + \alpha_n n\, v_s$
where α_p and α_n are multiplication coefficients for holes and electrons, depending from field.
The current components are:
$Jp = q\, p\, v_s$; $Jn = -\, q\, n\, v_s$

In the multiplication region:

$$\begin{cases} 1)\ \dfrac{dJ_P}{dx} = qG = \alpha_p J_P(x) + \alpha_n J_n(x) \\ 2)\ \dfrac{dJ_n}{dx} = -qG = -\alpha_p J_P(x) - \alpha_n J_n(x) \\ 3)\ J = J_P(x) + J_n(x) = \cos t. \end{cases}$$

Substituting eq. (3) in (1) one has:

$$\frac{dJ_P(x)}{dx} - (\alpha_p - \alpha_n) J_P(x) = \alpha_n J$$

An useful simplification comes from the assumption of equal ionization coefficients for holes and electrons: $\alpha_P = \alpha_N = \alpha$, where α is a function of E and from E(x) depends on x

$$\frac{dJ_P(x)}{dx} = \alpha(x) J \quad \Rightarrow \quad J_P(x) = J \int_0^x \alpha(x)dx + C$$

Integrating with boundary conditions:
at x = 0 Jp = Jp0
at x=W Jp=MpJp0

$$J_P(W) = J \int_0^W \alpha(x)dx + J_{P0}$$

summing J_{N0} to both members:

$$J_P(W) + J_{N0} \equiv J = J \int_0^W \alpha(x)dx + J_{P0} + J_{N0}$$

Then, by recalling that at x = 0 $J_{P0} + J_{N0} = J_0$, we have for the multiplied current J:

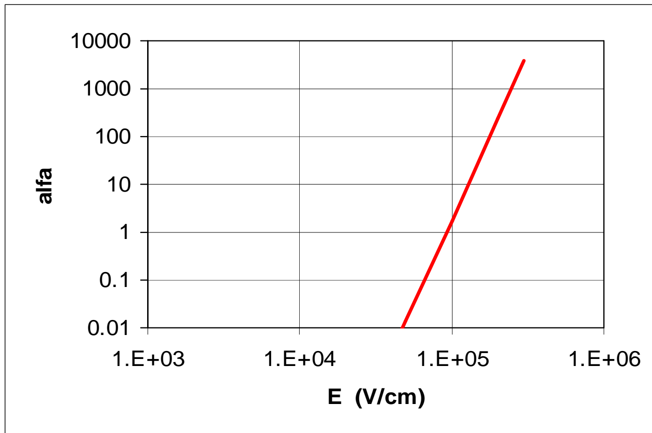
$$J\left[1 - \int_0^W \alpha(x)dx\right] = J_0 \quad \Rightarrow \quad \boxed{\frac{J}{J_0} = M = \frac{1}{1 - \int_0^W \alpha(x)dx}}$$

From the previous equation we see that the breakdown condition M→ ∝ is reached when:

$$\int_0^W \alpha(x)dx = 1$$

(breakdown condition)

α is a a strong function of the electric field E; an approximate expression of α(E) is:
a(E)= 1.8 10^{-35} E ^{7}



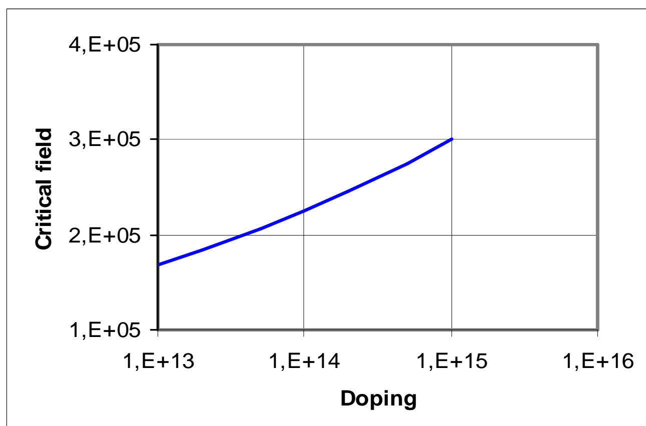
This is a very fast dependence from field E, as it can be seen from the plot.
The value of α can be significant only for fields in the range of 10^5 V/cm (for lower E values there is no appreciable multiplication effect)

The breakdown condition: $\int_0^W \alpha(x)dx = 1$ can be transformed as a function of field by a variable transform *dx* → *dE*

$$\frac{dx}{dE}\int_0^{E_{CR}} \alpha(E)dE = 1$$

is a function of a critical field E_{CR} that gives the first term equal to 1.

where $\frac{dx}{dE}$ is defined by the Poisson eq. $\left[\text{for constant doping}: \frac{dx}{dE} = \frac{\varepsilon}{qN_D}\right]$

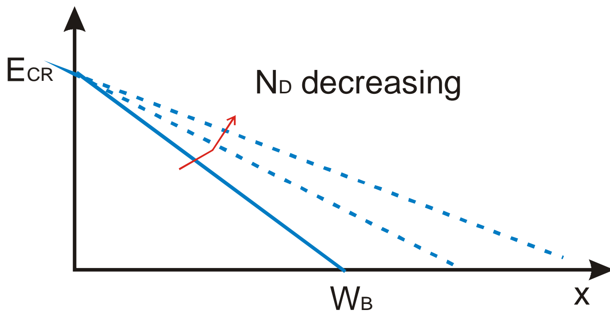


As it can be seen from the plot where the approximate value of α(E) given is used, the critical field E_{CR} is slowly varying with the doping for dopings less than 10^{15} cm^{-3}

For an approximate analysis we can assume

E_{CR} = 2 10^5 V/cm

Assuming as the critical field for avalanche breakdown: E_{CR} = 200 kV/cm one can easily determine the breakdown voltage as a function of doping of the low doped region of the PN junction, assuming a linear field profile down to the abscissa W_B (depleted width at breakdown):



Recalling the relations:

$$E_M = \frac{qW_D N_D}{\varepsilon} \quad (4)$$

$$V_R = \frac{E_M W_D}{2} \quad (5)$$

At breakdown : $V_{BR} = \frac{1}{2} E_{CR} W_B$ (6)

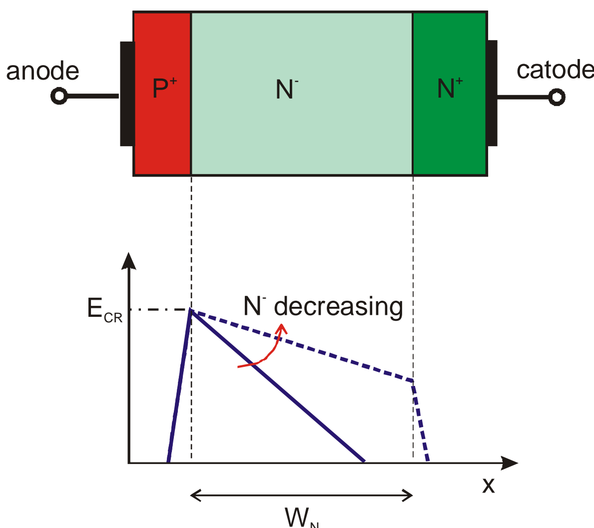
From the Poisson relation (a) we obtain: $$W_B = \frac{E_{CR}\varepsilon}{qN_D} \quad (7)$$ and substituting in (6):

we obtain the V_{BR} dependence on doping as:

$$V_{BR} = \frac{1}{2}\frac{E_{CR}^2\ \varepsilon}{qN_D} \quad (8a)$$

The PIN diode

Up to now we considered only a P+/N junction, not a diode. To have a diode we need to contact both the P and N region with a metal to which are bonded the wires that connect the si chip to the external terminals. To have a good ohmic contact between the metal and the semiconductor, we need to have a high doping at the interface; otherwise we should have a schottky barrier and not an ohmic contact at the Si-metal interface.
So we need to interpose an high doped N^+ layer between the low doped N region and the metal (the P^+ region has already high doping). This region can be the substrate of the chip for low N^- thickness, or a diffused region for thick N^- layers (above 150 μm).



For a given W_N thickness of the low doped layer, and depending on the doping N^-, we can have two cases, as indicated in the side plot:

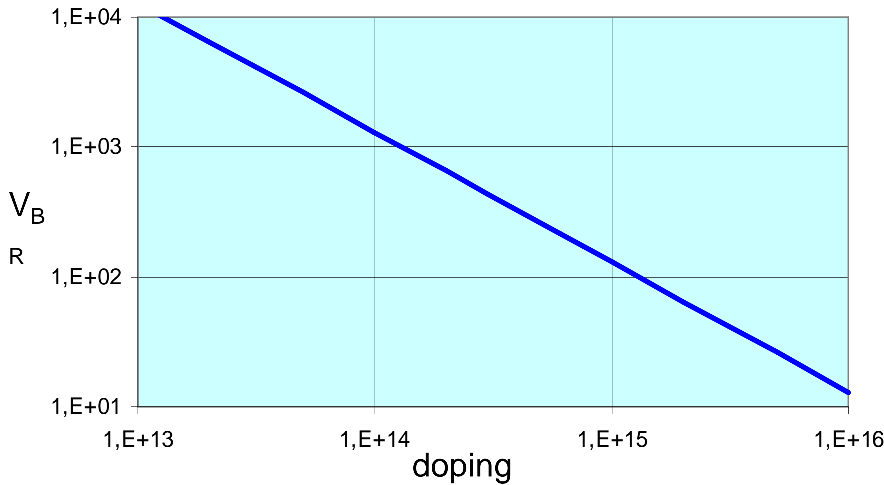
a) the field at breakdown goes to zero in the low doped region (solid line): the depleted region is all inside the N^- region (Non-Punch-Through case - NPT)
b) the field at breakdown reach the high doped N^+ region before going to zero (dashed line): this means that the depletion region reach the high doped region N^+ and then goes to zero (Punch-Through case - PT)

Using the previous relation (8a) for the V_{BR} dependence on doping of the N region, assuming a Non-Punch-Through case (NPT), we obtain:

$$V_{BR} = \frac{1}{2}\frac{E_{CR}^2\ \varepsilon}{qN_D}$$

with: $\varepsilon_{SI} = 10^{-12}$ F/cm
E_{CR} = 200 KV/cm

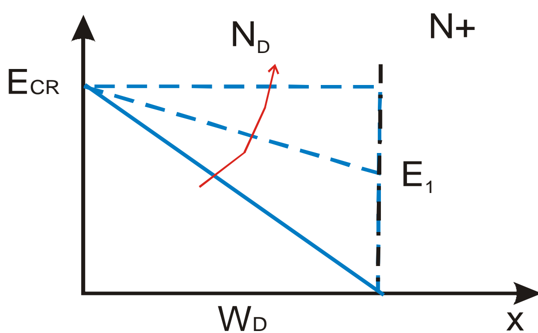
$$V_{BR} = \frac{1.3 \cdot 10^{17}}{N_D}$$



This plot shows a breakdown voltage of more than 1 KV for a N^- region doped 10^{14} cm^{-3}

However, the limitation posed by the thickness W of the N region is not taken into account in this plot; from eq. (d) to account for the NPT case, one need a thickness W_B = 125 μm at breakdown, quite high as we will see in the following.

If the thickness W_D of the low doped N_D region is less than value W_B given by eq. (d) the field profile assumes a trapezoidal shape.



From the limit case of a doping N_D^* for which $W_D = W_B$, if the doping decreases furthermore, then the breakdown voltage merely doubles with respect to the one given by eq. (e) for the N_D^*, i.e.:

for $N_D \to 0$ $\quad V_{BR} = E_{CR} W_D$

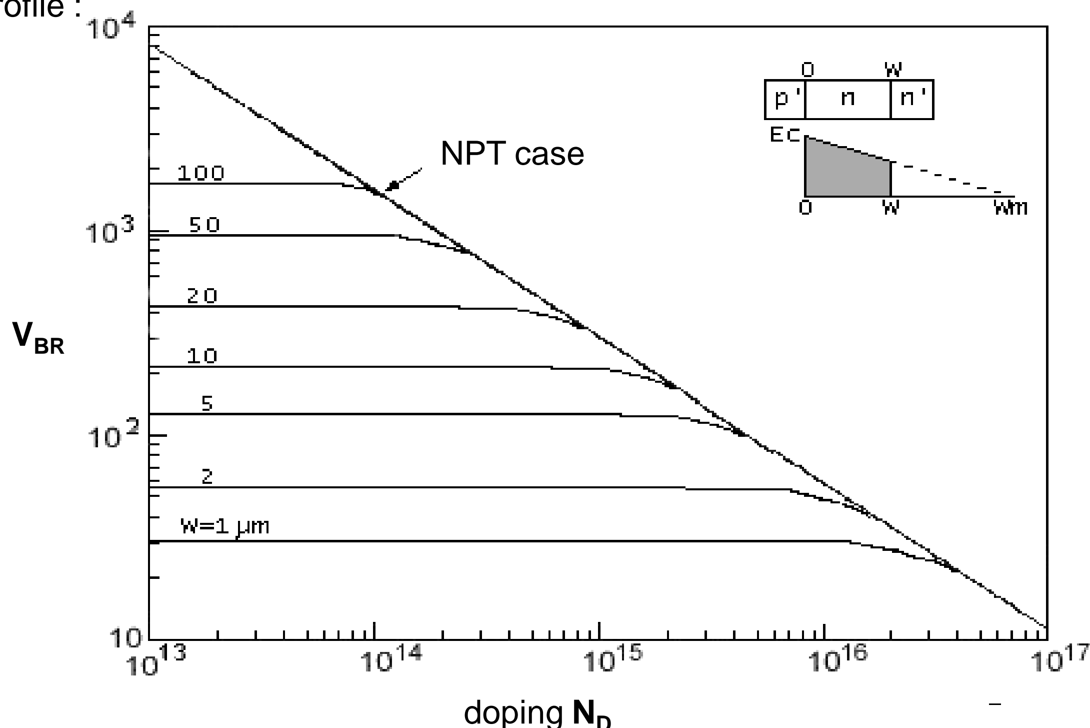
For the more general case of a trapezoidal field profile, we have (PT case):

$$V_{BR} = E_{CR} W_D - \frac{E_{CR} - E_1}{2} W_D$$

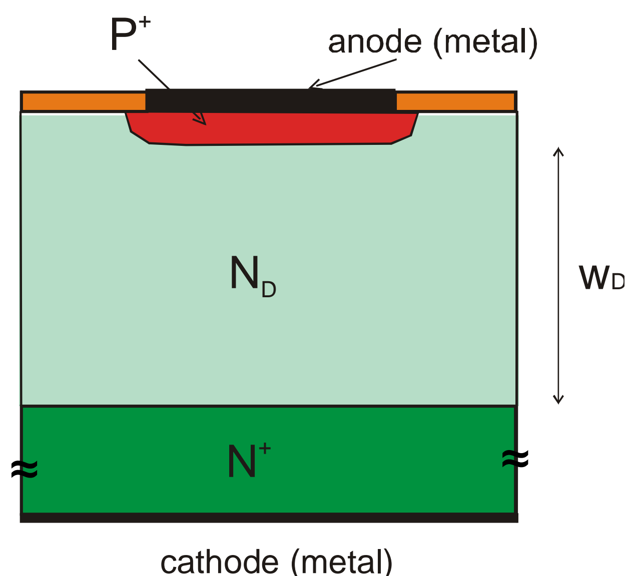
$$V_{BR} = E_{CR} W_D - \frac{qN_D W_D^2}{2\varepsilon} \qquad (8b)$$

Breakdown Voltage for a Punch-Through diode

Using eq. (8b) for the general case case we obtain the breakdown voltage as a function of both the doping N_D and the thickness W_D of the low doped region. In this graphic the asymptote for very low doping corresponds to the rectangular field profile, and the PT line to the triangular field profile :



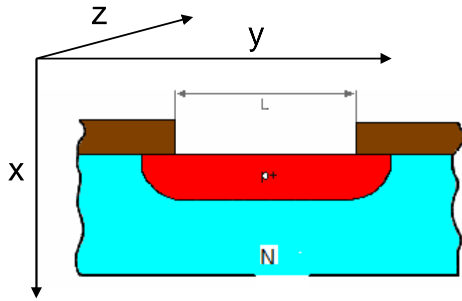
Physical structure of a PIN diode



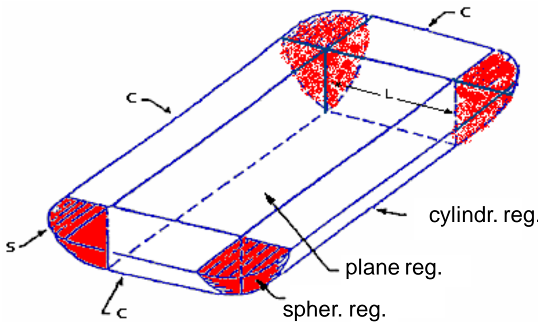
The cross-section of the physical structure of a PIN diode is depicted here.
The N^+ region (not in scale) is the chip substrate (some hundreds of microns, thicker than the active area). The low doped N_D region of thickness W_D is the epitaxial layer, in which the P^+ region is made by implant and thermal diffusion through the field $SI0_2$ oxide. The oxide layer on top protects the lateral sides of the PN junction from an excessive current leakage in reverse bias, because the oxide is a good passivation layer (this is a basic benefit of the so-called planar technology).

However as a result of this technique the PN junction will present a curvature both in the Y and Z directions (assuming the x axis as the vertical depth). This curvature will result in a strong limitation on the breakdown voltage, as we will see in the following.

2D and 3D effects of junction curvature

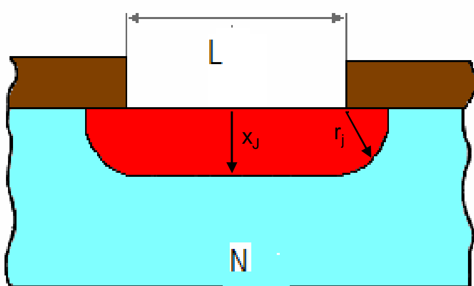


The doping through the oxide window gives rise to a lateral diffusion under the oxide, and creates a 2D geometrical shape of the junction in the X and Y directions if one assumes an infinite length in the Z direction: in that case it is said that the junction has a cylindrical curvature in the y direction.

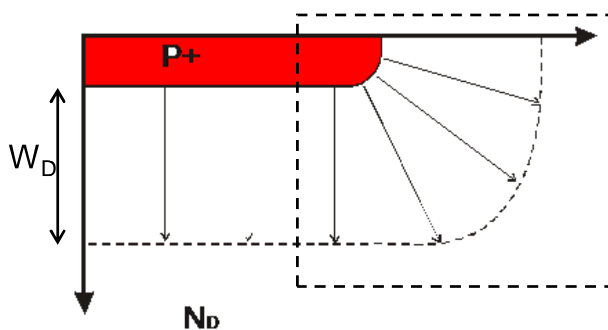


If the length of the window in the Z direction is not infinite, the corners of the window will create after diffusion a spherical junction curvature: it is said that the junction has a spherical curvature because it is bended both in the Y and Z directions.

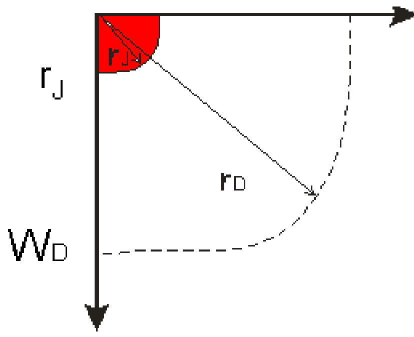
Field analysis for cylindrical junction



Assumptions: a) isotropic diffusion: $r_J = x_J$
b) approximation: $L >> r_j$



the cylindrical region in the dashed area can be analyzed by using radial coordinate in the plane E,r



The field in the cylindrical region can be described again by the Poisson equation in radial coordinates:

$$\frac{1}{r}\frac{d}{dr}(rE) = \frac{Q(r)}{\varepsilon} = -\frac{qN_D(r)}{\varepsilon}$$

Integrating the first term between the generic abscissa r and r_D (where E = 0), we have a power law dependence for the field:

$$E(r) = -\frac{qN_D(r)}{\varepsilon}\left(\frac{r_D^2 - r^2}{r}\right)$$

At the abscissa r_J the field has is max value $E_M|_{CIL}$:

$$\left|E_M\right|_{CIL} = \frac{qN_D}{2\varepsilon}\left(\frac{r_D^2 - r_J^2}{r_J}\right) \qquad (9)$$

For low doping values N_D the depletion layer is large and : $r_D >> r_J$. It is possible in this case to approximate eq. (9) as:

$$E_M\Big|_{CIL} \cong \frac{qN_D}{2\varepsilon}\left(\frac{r_D^2}{r_J}\right) \qquad (10)$$

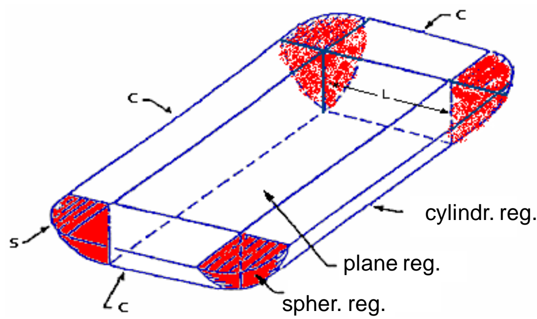
For the plane junction the thickness W_D can be written as: $W_D = r_D - r_J$, and the max field $E_M|_P$ in the same approximation $r_D >> r_J$ can be written as:

$$E_M\Big|_P = \frac{qN_D W_D}{\varepsilon} = \frac{qN_D(r_D - r_J)}{\varepsilon} \cong \frac{qN_D r_D}{\varepsilon} \qquad (11)$$

From eq. (10) and (11) it comes out that the ratio of the max field $E_M|_{CIL}$ in the cylindrical region and $E_M|_P$ in the plane region is larger than 1:

$$\frac{E_M\big|_{CIL}}{E_M\big|_P} \cong \frac{r_D}{2r_J} >> 1 \qquad (12)$$

Field analysis for spherical junction



For the spherical region of the junction the field can be again described by the Poisson equation in radial coordinates:

$$\frac{1}{r^2}\frac{d}{dr}(r^2E) = \frac{Q(r)}{\varepsilon} = -\frac{qN_D(r)}{\varepsilon}$$

Integrating between r and r_D we extract the field E(r), and we obtain the $E_M|_{SP}$ as :

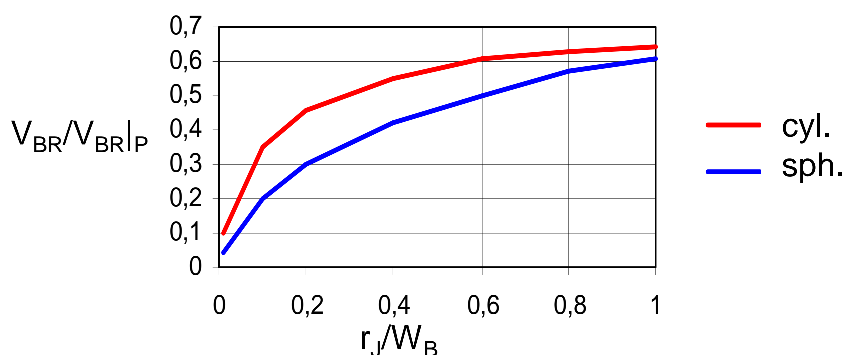
$$E(r) = -\frac{qN_D(r)}{3\varepsilon}\left(\frac{r_D^3 - r^3}{r^2}\right) \Rightarrow \left|E_M\right|_{SP} = \frac{qN_D}{3\varepsilon}\left(\frac{r_D^3 - r_J^3}{r_J^2}\right)$$

In the case $r_D >> r_J$ $\quad \left|E_M\right|_{SP} \cong \frac{qN_D}{3\varepsilon}\left(\frac{r_D^3}{r_J^2}\right)$

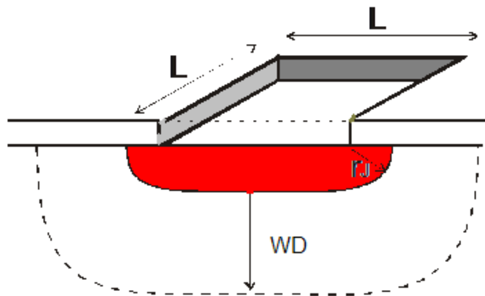
$$\frac{E_M|_{SP}}{E_M|_{CIL}} \cong \frac{2r_D}{3r_J} >> 1 \qquad (13)$$

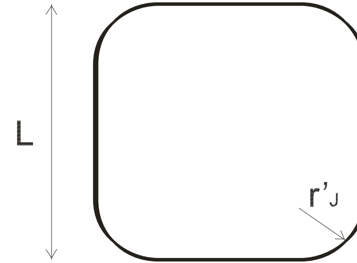
Breakdown voltage for cylindrical or spherical junction

From eq. (12) and (13) it comes out that $E_M|_{CIL}$ and furthermore $E_M|_{SP}$ are higher than $E_M|_P$ of the plane junction, so the value E_{CR} for the breakdown will be reached in the cylindrical (or spherical) region when the $E_M|_P$ is still less than the E_{CR} value.
Assumed equipotential the anode and cathode planes, the V_{BR} value for the cylindrical (or spherical) regions will be lower than the one $V_{BR}|_P$ of the plane region, and the current will increase to infinite even if it is flowing only in the portion of the junction subjected to premature breakdown.
By integrating the expressions of E(r) for the two cases analyzed one can obtain the following results for $V_{BR}|_{CIL}$ or $V_{BR}|_{SP}$ normalized to $V_{BR}|_P$, as a function of the junction curvature radius r_J normalized to the plane depletion layer at breakdown W_B



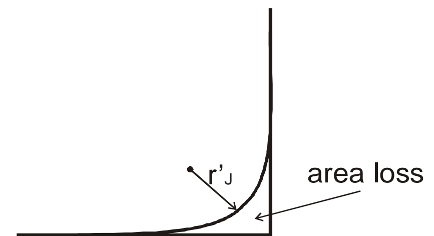
The decrease of $V_{BR}|_{SP}$ respect to $V_{BR}|_{CYL}$ can be eliminated by making the curvature radius r'_J in the Y,Z plane (chip surface) much higher than the one r_J of the cylindrical region. This is possible because the r'_J value is determined by the mask design of the window opening, and one can easily made $r'_J >> r_J$.
This latter is limited by the diffusion depth of the P^+ region, that is usually less than 10 μm, and is usually much less than the depletion layer W_B at breakdown.





The limitation for high r'_J comes out from the area loss with respect to the one of a square window: the assumptions then are: $L >> r'_J > W_B$; then $r'_J >> r_J$.

As an example: $x_J = r_J = 10\mu m$, $W_D = 50\mu m$, $r'_J = 200\ \mu m$
Percentage area loss < 2 10^{-2} if L = 1mm.

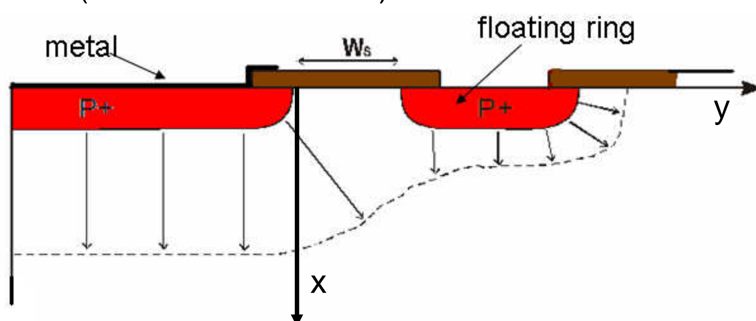


Junction termination techniques

To overcome the V_{BR} decrease due to the lateral junction curvature the two most widely used techniques (both for the power diode and the other high voltage devices) are:

a) Floating field rings
b) Field plates

The first one need to create some P+ diffused regions (rings) around the P^+ anode region as indicated in the plot (here only one ring is sketched), left "floating" with respect to the anode bias (not metal contacted).



Assuming the potential in the y direction not altered by the presence of the ring, this latter will be actually biased by the potential value as far as the depletion region reach (and trespass) it.

Using the Poisson eq. for the plane case in the y direction (eq. 14), starting from the lateral boundary of the P+ anode, and integrating it with the boundary conditions:
V= 0 for y=0; V= $-V_R$ for y= W_D (depletion width at V_R, assumed equal to $W_D|_X$), we have:

$$\frac{d^2V}{dy^2} = -\frac{dE}{dy} = -\frac{q}{\varepsilon}N_D \quad (14) \qquad |V(y)| = V_R - \frac{q}{\varepsilon}N_D\left(W_D y - \frac{y^2}{2}\right) \quad (15)$$

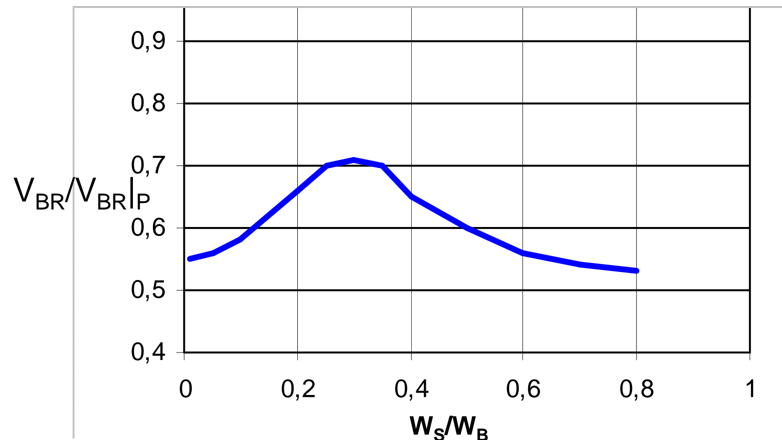
When V_R is increased, the depletion region W_D extends in the y direction and reaches the floating ring at W_S before breakdown if $W_S < W_B$.
Then the voltage bias of the field ring at $y = W_S$ will be:

$$|V_{FR}| = V_R - \frac{q}{\varepsilon} N_D \left(W_D W_S - \frac{W_S^{\,2}}{2} \right)$$

Two limiting cases are:

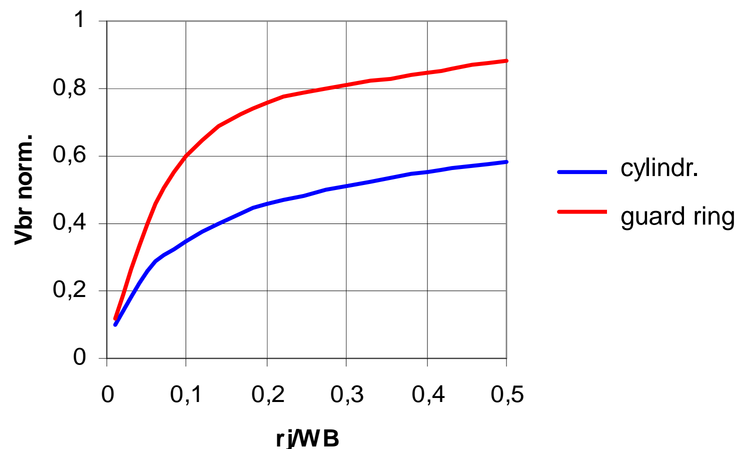
a) for $W_S \to 0$ $V_{FR} = V_R$ Then the floating ring has the same voltage of the main junction and the diode V_{BR} is limited by the $V_{BR}|_{CIYL}$ of the floating ring because its junction curvature is the same as for the main junction.

b) for $W_S \to W_B$ $V_{FR} = 0$. Then the ring is not biased up to the breakdown voltage and this latter is the $V_{BR}|_{CYL}$ of the main junction, i.e. the main junction is not protected by premature breakdown by the presence of the guard ring.

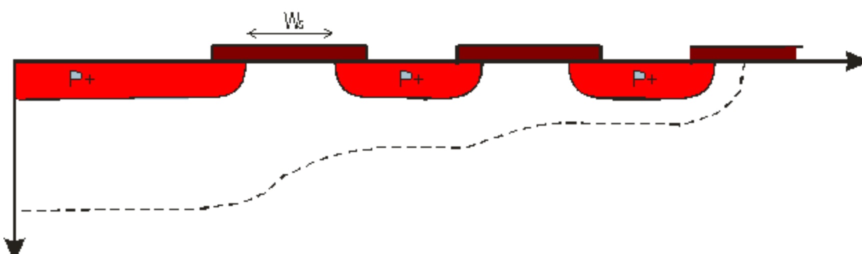


The best choice, as seen in the plot (for the case r_J/W_B=0.4), is $W_S \cong W_B/3$.

The increase of V_{BR} using a floating ring is dependent on the curvature radius r_J of the junction (the depth and curvature of the ring is assumed equal to that of the main junction because the ring is made with the same diffusion process as the main junction).



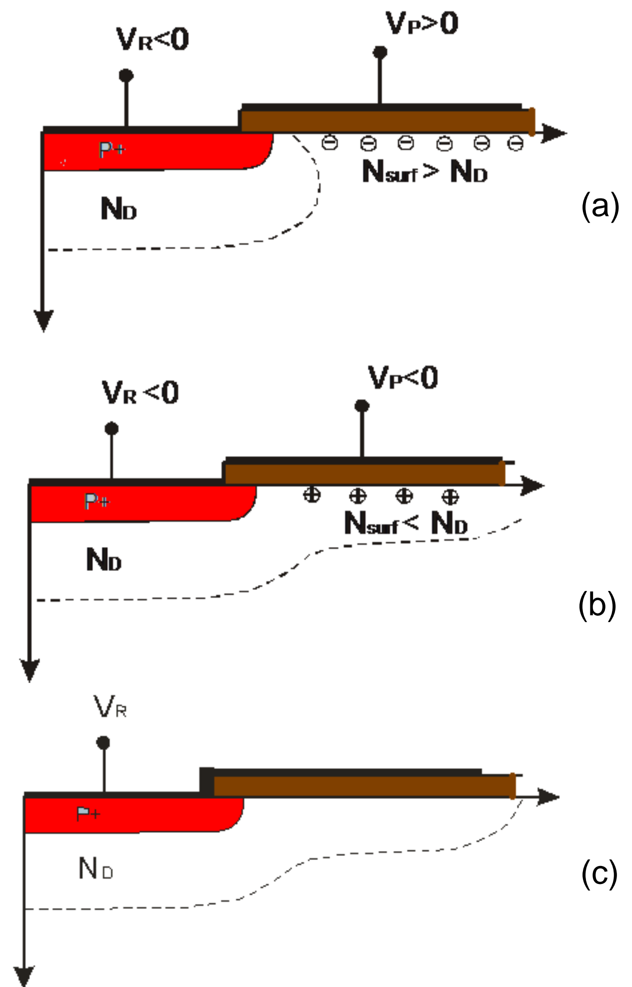
To increase the V_{BR} near the value of the plane case for the case of low r_J/W_B ratios, we need to use multiple floating rings (about 4 or 5 rings are usually employed) to reduce the lateral field for each ring respect to the preceding one.



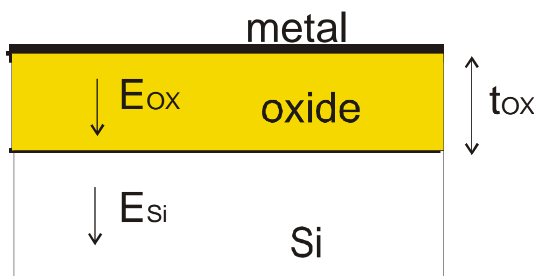
b) Field plate

This technique uses a metal plate over the oxide and the charge induced on the Si surface to modify the lateral field distribution.
If the metal plate over the oxide is positively biased (case a), the surface is in the accumulation state and $N_{SURF}>N_D$; $V_{BR}<V_{BR}|_{CYL}$
If the metal plate is negatively biased (case b), $N_{SURF}<N_D$, and a depletion layer due to the plate bias develops under the plate thus reducing the lateral field. In case (b) the equipotential lines extend under the plate and the potential distribution is more similar to the plane case, so $V_{BR} > V_{BR}|_{CYL}$.
In the practical case (c) the metal plate over the oxide is connected to the metal of the anode junction, so the bias of the field plate is the same as the V_R of the junction.

How to evaluate the equivalent voltage induced on the Si surface by the field plate?



The critical parameter for evaluation the voltage bias at the Si surface is is the oxide thickness t_{OX} of the field plate. We can evaluate the field across the oxide using the Gauss law:



$$E_{OX} \cdot \varepsilon_{OX} = E_{SI} \cdot \varepsilon_{SI}$$

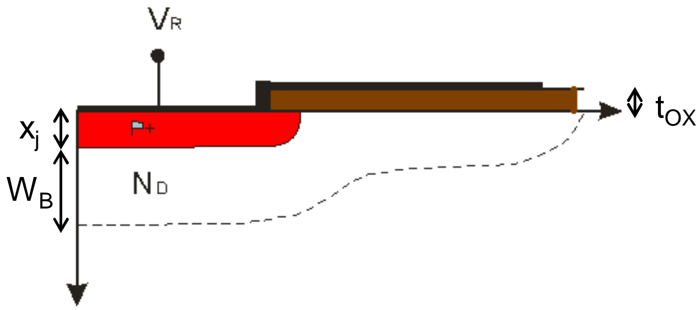
$$E_{OX} = E_{SI} \frac{\varepsilon_{SI}}{\varepsilon_{OX}} = E_{SI} \frac{12}{4} = 3E_{SI}$$

Gauss law tells that the field in the oxide is 3 times the field in the Si material: that means that to evaluate the voltage drop across the thickness t_{OX} we can consider an equivalent Si layer with a thickness $3t_{OX}$. ($\Delta V = E_{SI}t_{SI} = E_{OX}t_{OX}$).
For negative bias of the field plate, the field at the end of the plate corresponds to the case of a diffused Si P^+ layer with an equivalent thickness $3t_{OX}$, and then with a curvature radius $r_J = 3t_{OX}$ and with a voltage bias $V_R - \Delta V$
If we use a thin oxide, the voltage induced on the Si surface is near to the one of the main junction, but the field due to the curvature radius at the end of the field plate will be very strong, inducing a premature breakdown at the end of the field plate.
If we use a thick oxide, the curvature will be quite small, but the voltage drop under the plate will be large and the voltage at the Si surface will be much less than the one of the main junction, creating a premature breakdown at the lateral junction side .

Example: consider two possible cases:

a) x_J=10μ W_B=100μ t_{OX} = 1μ r_{JFP} = 3μ

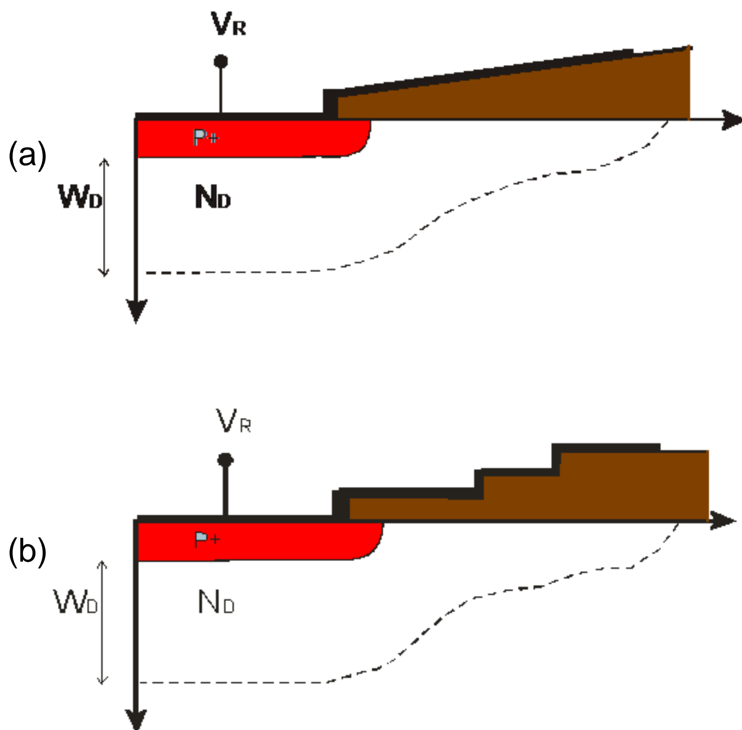
b) x_J=10μ W_B=100μ t_{OX} = 10μ r_{JFP} = 30μ



Assuming a reverse bias V_R = 1000 V, we obtain:

a) For E_{CR} SI = 20 V/μm E_{OX} =60V/μm
ΔV_{OX} = 60 V V_{FP} = 940 V
con r_{JFP} = 3μ << r_J = 10μ the breakdown develops at the end of the field plate

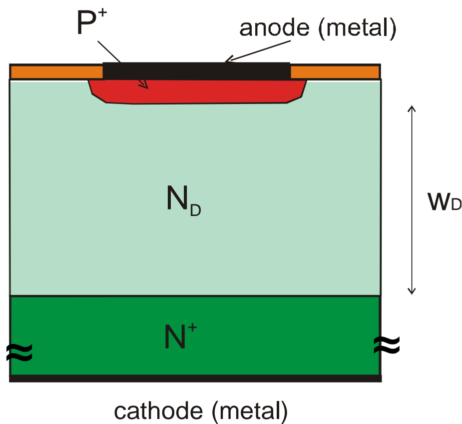
b) For E_{CR} SI = 20 V/μm E_{OX} =60V/μm
ΔV_{OX} = 600 V V_{FP} = 400 V
con r_{JFP} = 30μ > r_J = 10 μ the breakdown develops at the main junction lateral side



The ideal solution would be to make an oxide thickness that increases in the y direction going far from the main junction, as indicated in fig. (a). This however is not workable for the usual planar technology:

A more practical solution is to make an oxide thickness that is increasing in steps in moving far from the main junction, as indicated in the fig. (b).The above can be obtained by depositing over the thermal oxide the next oxide layers by CVD with different film thickness and windows openings

Ratings for the PIN power diode



The main ratings for a power diode from its datasheet are:

- Max reverse voltage in OFF state V_{RMAX}
- Max current in ON state I_{MAX}
- Voltage drop in ON state at nominal current V_{ON}
- Dynamic power losses in a) turn-on
  b) turn-off

These ratings are mainly defined by the following structure parameters:

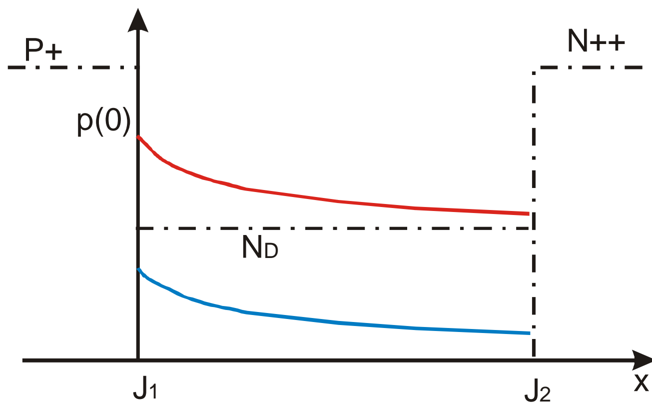
- N_D doping of the low doped region
- W_D thickness of the low doped region
- lifetime τ of the low doped region

From the analysis of the breakdown voltage of the PIN diode we know that in order to allow an high V_{RMAX} the structure must have a low doping N_D and a relatively thick W_D (for example N_D about 10^{14}cm^{-3} and W_D about 50 μm to have a V_{RMAX} of 1000V).
What are the effects of the presence of the N_D layer of thickness W_D on the static and dynamic performances of the diode?

I-V characteristics of the PIN diode

The presence of two N regions (the low doped N_D of the epitaxial layer and the high doped N^{++} of the substrate), as well as the operation at high current levels, will induce some changes in the static I-V curves with respect to the usual ones of the signal P/N diode.
In the structure we can see two junctions, J_1 and J_2; the presence of the N_D/N^{++} junction J_2 and the low level of doping in the N_D region, have an important effect on the I-V characteristics (as well as for the dynamic behavior, as we will see).

The holes injected in that region from the P^+ layer are usually called “minority carriers” because they are injected in a N doped region, where the majority carriers are electrons. However if the doping N_D of this region is quite low, it can happens that the holes density injected from the (high doped) P^+ region can have an higher value than the one of electrons that is equal to N_D at ambient temperature (all donors are ionized and give a free electron).



This plot reports the holes (minority carriers) distribution in the low doped N_D layer, for a diffusion length $L=(D\tau)^{1/2} > W_D$.

In the plot we can distinguish two different modes of operation:

a) low injection when the injected carriers (holes) are lower than the doping N_D (blue curve)
b) high injection when the injected carriers are higher than the doping N_D (red curve)

In the neutral N_D region the electric charge balance gives:

N_D +p (positive charge) = n (negative charge) (a)

We have low injection if: $p \ll N_D$ then from (a) $n \cong N_D$

We have high injection if: $p > N_D$ then from (a) $n \cong p > N_D$

In low injection regime $n(x) = N_D$ there is no voltage drop on J_2 and $V = V(J_1)$. The J-V characteristic of the diode is the ideal one:

$$J = J_0\left[\exp\left(\frac{V}{V_T}\right) - 1\right] \quad (16)$$

In high injection regime $p(0) > N_D$ let's recall the expression of p(0):

$$p(0) = p_o \exp\left(\frac{V_{J1}}{V_T}\right) = \frac{n_i^2}{N_D}\exp\left(\frac{V_{J1}}{V_T}\right) \quad (17)$$

As an example:
for $V = V_{J1} = 0.6V$, assuming $N_D = 10^{14}$ cm^{-3}, $ni^2 = 2\cdot10^{20}$ cm^{-3}, we have from eq. (16): $p(0) = 2\cdot10^{15} \gg N_D$. Then we are in high injection because: $p(0) = n(0) \gg N_D$. If the diffusion length $L_P > W_D$ also $p(W) = n(W) > N_D$.

Due to an electron concentration $n(W_D) > N_D$, for the Boltzmann equation there will be a voltage drop V_{J2} across the junction J_2 to allow this electron increase; J_2 will be directly biased and will inject electrons from the substrate N^{++} into the N_D region.

The overall diode voltage will be the sum $V_{PIN} = V_{J1} + V_{J2}$
Let's evaluate the I-V relationship in that case

In the assumption $L_P >> W_D$ we can approximate the holes distribution as flat over the thickness W_D, and we have: $p(W_D) \cong n(W_D) > N_D$.
From the Boltzmann eq. (17) applied to the abscissa W_D, in the case $n(W_D) > N_D$, we have:

$$n(W_D) = N_D \exp\left(\frac{V_{J2}}{V_T}\right)$$

indicating that a voltage V_{J2} develops across J_2 in order to have $n(W_D) > N_D$

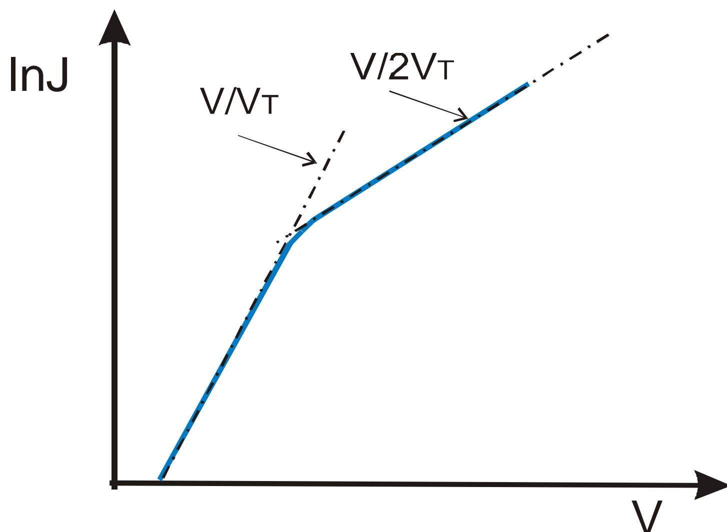
The current density J of the diode can be related to the carrier densities p(0) or n(W) as:

$$J = \frac{Q_P}{\tau} \cong \frac{qp(0)W_D}{\tau} = J_{P0} \exp\left(\frac{V_{J1}}{V_T}\right) \cong \frac{qn(W_D)W_D}{\tau} = J_{N0} \exp\left(\frac{V_{J2}}{V_T}\right)$$

$$\frac{J}{J_{P0}} = \exp\left(\frac{V_{J1}}{V_T}\right); \quad \frac{J}{J_{N0}} = \exp\left(\frac{V_{J2}}{V_T}\right); \quad \ln J - \ln J_{P0} = \frac{V_{J1}}{V_T}; \quad \ln J - \ln J_{N0} = \frac{V_{J2}}{V_T};$$

for a variation ΔJ: $\Delta \ln J = \frac{\Delta V_{J1}}{V_T}$; $\Delta \ln J = \frac{\Delta V_{J2}}{V_T}$; $2\Delta \ln J = \frac{\Delta(V_{J1}+V_{J2})}{V_T} = \frac{\Delta V_{PIN}}{V_T}$

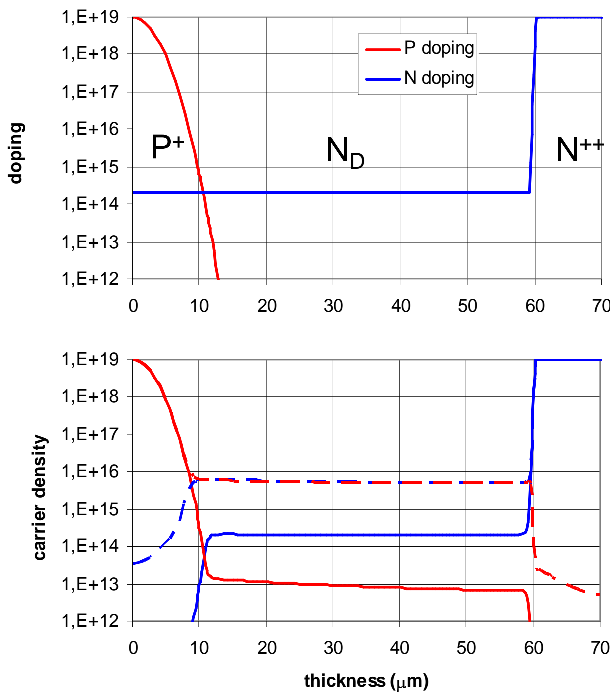
$$\Delta \ln J = \frac{\Delta V_{PIN}}{2V_T} \Rightarrow \ln J_2 - \ln J_1 = \frac{V_{PIN2} - V_{PIN1}}{2V_T} \Rightarrow \boxed{J = J_0 \exp\left(\frac{V_{PIN}}{2V_T}\right);} \quad (18)$$



In high injection the voltage drop V_{PIN} across the diode increases with respect to the low injection case because the slope of the I-V curve in the log scale is $1/2V_T$ instead $1/V_T$ as for the the ideal PN diode. This will increase the power dissipation in the d.c. case.

The point of change in slope is determined by the temperature (through the V_T term), and on the doping N_D: the lower is N_D the lower the voltage V at which the high injection start to develop; for N_D in the range of 10^{14} the diode is in high injection starting from V = 0.6 V

The assumption of a constant carrier distribution in the low doped region in high injection can be checked by the results of a 1D numerical analysis of a PIN diode made by the PC1D numerical simulator.
The following plots present the results of the simulation for a typical power diode with parameters:



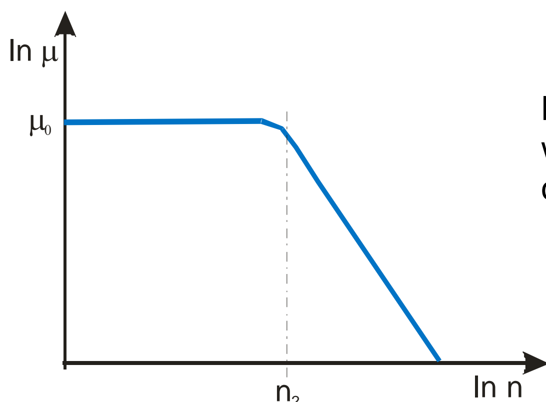
P^+ : gaussian distribution
J_1 = 10 μm J_2 = 60 μm
N_D = 2 10^{14} W_D = 50 μm
N^{++} : uniform doping 10^{19}
τ = 1 μs

The upper plot presents the doping concentrations of the PIN diode.
The bottom plot presents the carrier distributions:
in low injection (V= 0.45 V, solid lines) show the electrons at N_D value, and the holes at much lower value;
in high injection (V = 0.7 V, dashed lines) the holed and electron distributions are almost identical, with a quite constant value, higher than the N_D value.

The presence of a high concentration of both electrons and holes in the low doped region will strongly reduce the voltage drop across this region, respect to the ohmic case where only majority carriers were present; let's evaluate the voltage drop across that region:

In the approximation : p(x) = n(x) = n_a = cost. (ambipolar concentration), the carrier will move only with a drift field because the diffusion gradient is negligible, and the total diode current in the epi region will be : $J = q(\mu_n n + \mu_p p)E = q(\mu_n + \mu_p)n_a E = q\mu_a n_a E$ (19) where μ_a is the ambipolar mobility.
As the hole and electron mobilities, also the ambipolar mobility is dependent from the doping (or carrier) concentration; a simple expression of the mobility dependence on the doping concentration in low injection is:

$$\mu = \frac{\mu_0}{1 + \frac{N_D}{n_2}} \qquad \text{with} \quad n_2 \cong 10^{17}$$



In high injection the doping is substituted by n = p = n_a and we have for the ambipolar mobility the dependence from the carrier concentration as:

$$\mu_a = \frac{\mu_{a0}}{1 + \frac{n_a}{n_2}} \qquad (20)$$

From eq. (19): $J = q\mu_a n_a E = q\mu_a n_a \frac{V_D}{W_D}$ the voltage drop V_D across the low doped region is:

$$V_D = J\frac{W_D}{q\mu_a n_a} = \frac{W_D}{qn_a\mu_{a0}\frac{1}{1+n_a/n_2}} \quad (21)$$

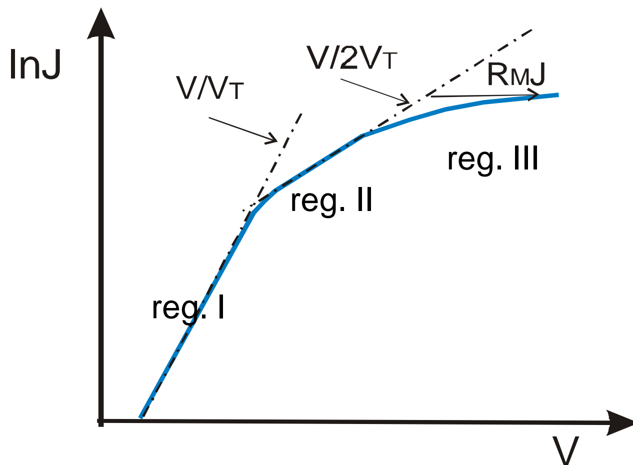
in the limit $n_a >> n_2$ (22), we have:

$$V_D = J\frac{W_D}{q\mu_{a0}n_2} = R_M I \quad (23)$$

where: $$R_M = \frac{W_D}{qA\mu_{a0}n_2}$$

The resistance R_M is much less than the "ohmic" resistance R_D of the epi region:

$$R_D = \frac{W_D}{qA\mu_0 N_D}$$



because n_2 is much higher than the doping N_D.

The decrease in resistance of the low doped region is called "*Conductivity modulation*" and is due to the high injection regime ($p=n>>N_D$).
In the I-V curve we can identify 3 regions: the first 2 exponential, and the 3rd ohmic, in the high current range (the normal operating range for power diode)

Trade-off between V_{ON} and V_{BR}

Even if the R_M value of the "conductivity modulated" epi region is much lower than the "ohmic" value, the voltage drop due to this resistance constitutes a significant increase of the total diode voltage $V_{PIN} = V_{J1} + V_{J2} + R_M I$.
The doping and thickness of the low doped epi region are defined according to the breakdown voltage V_{BR} needed, so we can express R_M as a function of V_{BR}. Recalling eq. (21) for R_M, with reference to a triangular field profile up to breakdown (NPT case) , assuming: $V_{BR} = E_{CR}W_D/2 = 10^{-5}$ V, $n_2 = 10^{17}$, $\mu_{a0} = 2000$ we have:

$$R_M = \frac{10^{-5}V_{BR}}{qA\mu_{a0}n_2} = \frac{V_{BR}}{3\cdot 10^4 A} \quad \text{(NPT)}$$

indicating that R_M increases linearly with the breakdown voltage V_{BR}.

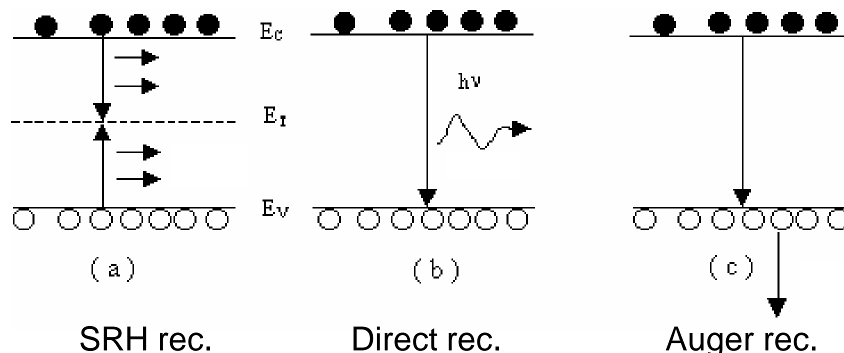
As an example: for V_{BR}= 1000V, assuming A=0.5cmx0.5cm → R_M = 0.13 Ω; with a current of 10 A the voltage drop $R_M I$ = 1.3 V, that makes the overall diode voltage over 2 V with a d.c. power dissipation of more than 20 W.

For the case of almost constant field profile up to W_D (in the limit of very low N_D values) we have $V_{BR} = E_{CR}W_D$, and the value of R_M is half the value of the NPT case.

Recombination lifetime

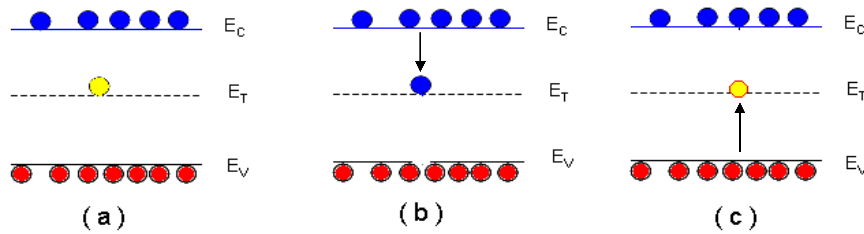
Up to now we did not consider the effect of carrier lifetime in the PIN diode analysis, due to the ideal approximation of a constant carrier profile P(x) = n(x) = const. The above assumption implies a carrier diffusion length L (and a recombination lifetime τ) $\rightarrow \infty$.
Let's now consider the (very important) role of carrier lifetime both in the I-V static and in the dynamic behaviour.
Carrier generation and recombination are processes that take place when the semiconductor is not in thermal equilibrium, (i.e. $p \cdot n = n_i^2$). When in a region the carriers are less than the equilibrium value ($p \cdot n < n_i^2$) there will be a net generation of carriers (typically in the depleted region of a reverse biased PN junction, the generation is responsible of the leakage current in reverse bias). When in a region the carriers are larger than the equilibrium value ($p \cdot n > n_i^2$) there will be a net recombination of carriers.
The carrier recombination plays a paramount role in both the static and dynamic behaviour of any bipolar device (not only PIN diode), and the recombination is defined through the recombination lifetime parameter, that indicates the mean time required for the carrier to disappear (because it recombines with an hole). We will briefly recall the main aspects of carrier recombination.
In all semiconductors we have three main recombination mechanisms to allow the energy and momentum balance required to eliminate an electro-hole pair from the mobile carriers (schematically indicated in the following figure that reports the electrons in the Conduction band and holes in the Valence band, separated by the bandgap E_G):

a) SRH (Shockley-Read-Hall) recombination that require a recombination center (or trap) located in the forbidden gap to allow the energy and momentum exchange between electron and holes, needed to eliminate both carriers from the carrier population.
b) Direct recombination that takes place when an electron and hole can match exactly their momentum so they disappear and the excess energy is released through electromagnetic radiation by an optical photon
c) Auger recombination that transfer the excess energy and momentum to a third carrier that need to be present in a triple collision process.



In Si, the most important is the SRH recombination (a), because the direct recombination (b) is unlikely to happen in semiconductors with indirect bandgap like Si, and the Auger recombination is important only at high carrier concentrations (above 5 10^{17}) to leave probability for a triple collision mechanism. So we will concentrate on SRH recombination.

The SRH recombination requires a third component – recombination center or trap – located in the forbidden gap – step (a) -, that can temporarily capture an electron (minority carrier in a P region) coming from the conduction band - step (b) - and then let it recombinate with a hole coming from the valence band – step (c) (the dual process is also possible if we consider the holes as minority carriers in N region).



The main parameters that determine the recombination process are:

a) the energy level E_T of the trap in the forbidden gap (referred to the midgap energy E_i)
b) the capture cross sections σ_p and σ_n respectively for holes and electrons
c) the density N_T of the trap

Both parameters are dependent on the actual source of the center (an impurity atom or a defect in the crystal); with this mechanism the net recombination rate U for the excess carriers (electrons or holes) present in a region ($pn \neq n_i^2$) can be described by a very general relationship that can be specified for all operation regimes.

The recombination rate U can be described by the following general equation:

$$U = \frac{pn - n_i^2}{\tau_{n0}\left(p + p_i \exp\left(\frac{E_i - E_T}{kT}\right)\right) + \tau_{p0}\left(n + n_i \exp\left(\frac{E_T - E_i}{kT}\right)\right)} \qquad (22)$$

where: $\tau_{n0} = \frac{1}{\sigma_n v_{th} N_T}$ $\tau_{p0} = \frac{1}{\sigma_p v_{th} N_T}$ are reference electron and holes lifetimes dependent on the thermal carrier velocity ν_{th}, on cross section σ and trap density N_T

Defining in general way the lifetime as $\tau = \frac{\Delta p}{U}$ (or $\frac{\Delta n}{U}$)

and referring for reference to a N region, where an excess Δp of carrier is injected, with $p = p_0 + \Delta p$, $n = N_D + \Delta n$, $p_0 N_D = n_i^2$, $p_0 << N_D$, substituting in (22), we have the following relationship for lifetime :

$$\tau = \frac{\tau_{n0}\left(p_0 + \Delta p + p_i \exp\left(\frac{E_i - E_T}{kT}\right)\right) + \tau_{p0}\left(N_D + \Delta n + n_i \exp\left(\frac{E_T - E_i}{kT}\right)\right)}{\Delta n + N_D} \qquad (23)$$

The lifetime depends on the recombination center through τ_{p0}, τ_{n0}, E_T, from doping through n_0 (= N_D), from injection level through Δn, Δp, and from temperature through n_i , kT.

a) low injection limit

in that case Δp , $\Delta n << N_D$, and eq. (23) becomes:

$$\tau_L = \tau_{p0}\left(1 + \frac{n_i}{N_D}\exp\left(\frac{E_T - E_i}{kT}\right)\right)$$

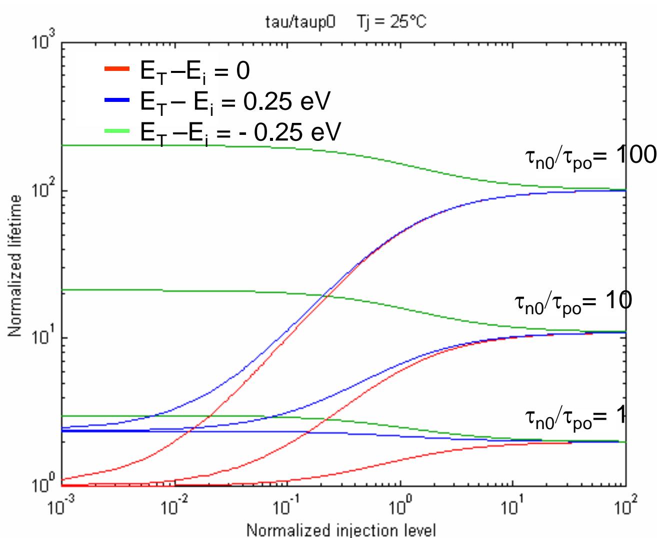
For the case of an energy level at midgap ($E_T = E_i$) : $\tau_L = \tau_{p0}$

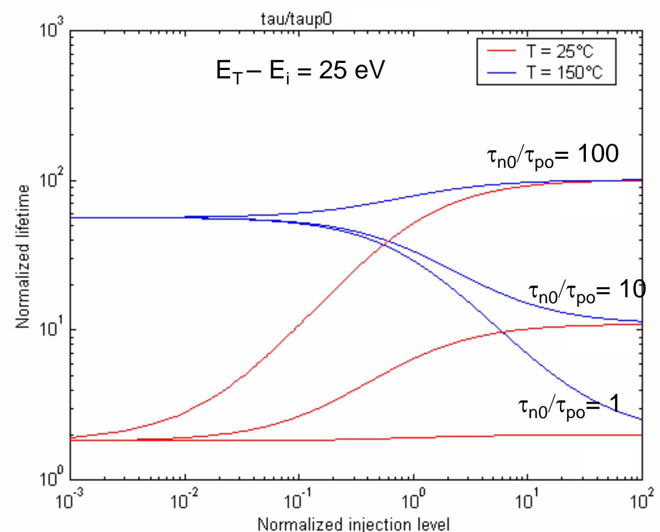
b) high injection limit

in that case $\Delta p = \Delta n >> N_D$, and eq. (23) becomes:

$$\tau_H = \tau_{p0} + \tau_{n0}$$

The lifetime as a function of injection level I = p_0/N_D is plotted in the two graphs below for different ratios of τ_{n0}/τ_{p0}. The first one presents the lifetime dependence on the energy level of the trap, and the second one from temperature.
For N doped Si, typically the energy level of the center lies in the upper bandgap, and τ_{n0} is about 5 -10 times τ_{p0}, so the lifetime increases both with injection level and with temperature; however the actual behavior depends on the recombination center present in the crystal, and it may be changed with the processes used for lifetime control, as we will see.

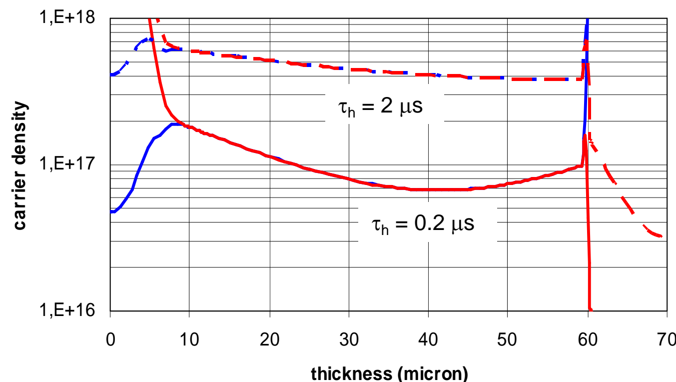




Lifetime dependence of the PIN diode behaviour

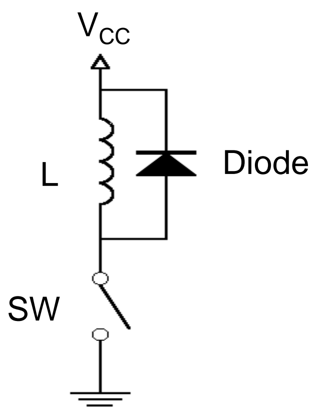
a) Static behaviour

The carrier lifetime affects the I-V curve mainly in region III. If the lifetime is not very high, the carrier distribution in the high injection regime is no longer a constant one, but tends to reduce in the middle of the epi region.
This behaviour can be noted in the carrier distribution plots of the following graph, obtained by PC1D for the same power diode as before.
For a high level lifetime of 2 μs (dashed line curves) one can note that the carrier distribution are much more flat than for the case of the low lifetime value of 0.2 μs, (solid lines) where the carriers reduce at 1/3 of the initial value in the center of the low doped region. In the latter case the approximation (22) do not holds, and the voltage drop is described by eq. (21) instead of eq. (23), with an average value of n_a lower than in the former case (the R_M value is higher in the latter case)



b) Dynamic behaviour

To allow a fast dynamic operation of the PIN diode, we need a fast change of the mobile carriers concentrations in the low doped region: this in turn will ask for a low lifetime to allow a quick redistribution of carries during the transient (in contrast with the d.c. case).
The basic test circuit for both the turn-ON and turn-OFF transients is the one schematically reported:



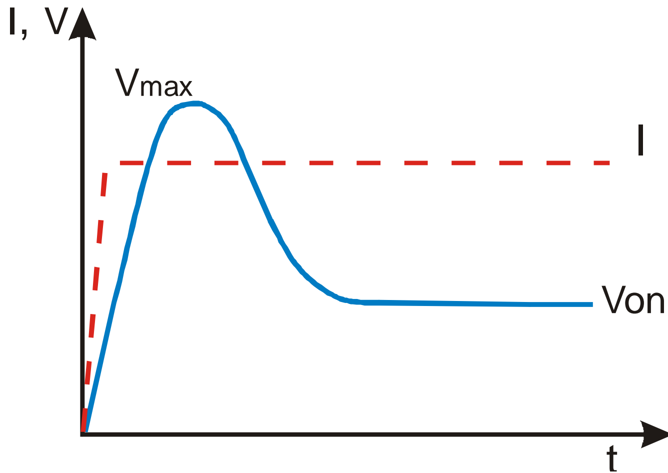
The ideal switch SW of the circuit is actually done by a power device (like a MOS): during the first time T_{ON} when SW is closed, the inductance L charges with a linear current ramp i(t) = $(L/V_{CC})t$; the diode is reverse biased by the voltage $V_{CC} - V_{ON(SW)}$.

During the time T_{OFF} SW opens, and the current $I(T_{ON})$ will start to flow into the diode, inducing a TURN-ON transient into it, after which the current I will flow into it (assuming that the T_{OFF} is sufficiently short to leave the current in L practically unaltered).

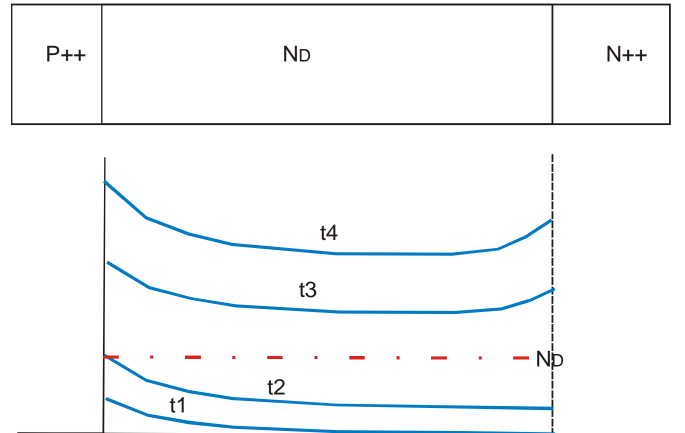
After the T_{OFF} time the switch SW closes again during a second T_{ON} time; the voltage drop across the inductance change abruptly from $-V_{ON}$ to $V_{CC} - V_{ON(SW)}$, and the diode becomes reverse biased: this will start the TURN-OFF transient, after which the diode becomes an open circuit and all the current flows into the switch SW.

Turn-on transient

When the diode is brought in turn-on by a current forced into it (red line in plot (a)), the diode voltage increases much above the V_{ON} voltage corresponding to the current I in steady state: this is because the holes injected in the low doped region (plot (b)) do not reach at the short times t_1 ,t_2 the high injection level (above N_D) needed to bring the N_D region in conductivity modulation (and to reduce its resistance R_D). As a consequence, the voltage drop across the diode increases to V_{MAX} (a value that can be of some tens of volts) and then it decreases to V_{ON} at times t_4 needed to establish the steady state n,p carrier distribution. The V_{MAX} increase is as higher as faster is the dI/dt slope of the forced current with respect to the carrier lifetime.



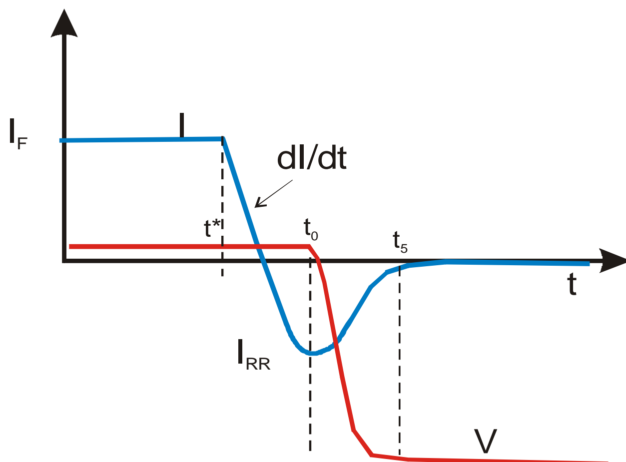
(a) current and voltage during turn-off



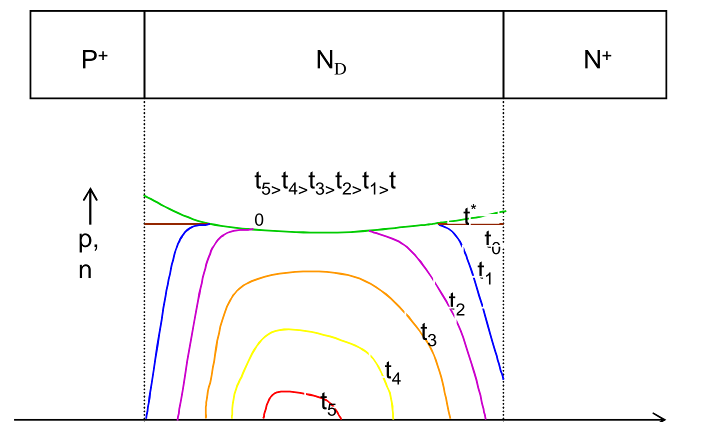
(b) hole distribution in N_D region

Turn-off transient

In real turn-off cases, when the switch closes at t=t*, the current in SW will increase almost linearly between t* and t_0, and then the diode current will linearly decrease.
Due to the charge still present in the N_D region when the diode current is crossing the zero value, the diode will remain in its conducting state (the diode voltage is dependent on the carrier distribution at the PN and NN^+ junctions) when the current is negative. Only after time t_1 the two junctions start to be depleted from carriers and two depleted regions develop, that increase their width toward the center of N_D region, until all the stored charge is removed and the diode will sustain the full reverse voltage V_R. During the time interval $t_0 - t_5$ there will be a substantial power dissipation due to the reverse current I_{RR} and to the growing reverse voltage of the diode.



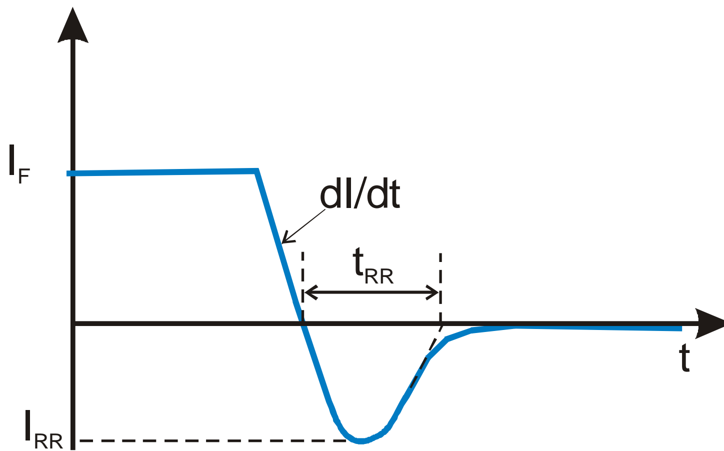
(a) current and voltage during turn-off



(b) carrier distributions in the N_D region

The diode switching transient in turn-off is named "Reverse recovery", indicating the time needed by the diode to establish a steady-state reverse behavior, i.e. an open circuit. With reference to a typical reverse recovery curve reported in figure, the main parameters for the reverse recovery transient are:

a) the on-state forward current I_F
b) the max current I_{RR} in reverse conducting stage
c) the slope dI/dt of the current
d) the time t_{RR} needed to recover the non conducting state after the current changes from positive to negative (usually it is determined by the time difference between the zero current point and the time corresponding to 10% of I_{RR}).



We will evaluate the effect of lifetime on the relevant parameters with some approximations, i.e. assuming the current plot during time t_{RR} as of triangular shape.

The total stored charge Q_{RR} in the low doped region to be removed during the reverse recovery is obtained by integrating the current during the t_{RR} time.
Assuming a triangular shape for the current curve we have from geometrical considerations:

$$Q_{RR} \cong \frac{I_{RR} t_{RR}}{2} \quad (24) \qquad I_{RR} \cong \frac{di}{dt}\frac{t_{RR}}{2} \quad (25)$$

The reverse charge Q_{RR} must be equal to the charge stored in forward conduction: this latter can be related to the lifetime τ through the charge control diode model: $I = Q/\tau$. Then:

$$Q_{RR} = Q_F = I_F \tau \quad (26)$$

and, by substituting (24), (25) in (26) we have the following expression for I_{RR}:

$$I_{RR} \cong \frac{2\tau I_F}{t_{RR}} = \frac{\tau I_F}{I_{RR}}\frac{di}{dt} \Rightarrow I_{RR} \cong \sqrt{\tau I_F \frac{di}{dt}} \quad (27)$$

or the following expression for t_{RR}:

$$I_{RR} \cong \frac{2\tau I_F}{t_{RR}} = \frac{\tau I_F}{I_{RR}}\frac{di}{dt} \Rightarrow t_{RR} = \frac{2\tau I_F}{t_{RR}\frac{di}{dt}} = \sqrt{\frac{2\tau I_F}{\frac{di}{dt}}} \quad (28)$$

From eq. (27) and (28) we found that both the peak reverse current I_{RR} and the reverse recovery time t_{RR} increase with the square root of lifetime, so a low lifetime value is needed to have a fast switching during the turn-off. The peak reverse current also increases if the forward current I_F increases, or the current slope dI/dt decreases (fast commutation).
The diodes that have a fast turn-off transient (low I_{RR} and low t_{RR}) are called "fast recovery (FR)" diodes. For these diodes one need to reduce the (rather high) lifetime value of the low doped region to allow a fast carrier recombination.

The technologies used to introduce recombination centers in the epi region are called "lifetime control techniques", and are usually needed for most bipolar devices if fast transient are required. In any case the need for a low lifetime is in contrast with the need of a conductivity modulation for the low doped region, so a trade-off between voltage drop and switching transient is a key point of all bipolar devices.

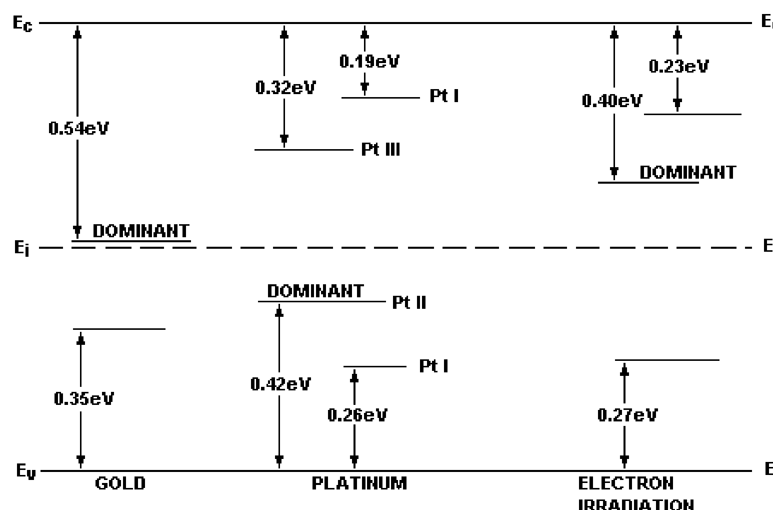
In particular for the PIN diodes with high reverse voltage rating, this is the main choice to be done:
- if the diode has to be used in low frequency operation the d.c. power losses have to be considered most relevant and a low voltage drop should be chosen to reduce the d.c. power losses: then a high τ is needed
- if the diode has to be used in high frequency operation, a fast switching should be chosen to reduce the dynamic power losses, and a low τ is required for a fast recovery diode.

Lifetime control techniques

The most used techniques for lifetime control are:

a) the introduction of impurity atoms, usually heavy metal atoms, like gold (Au) or platinum (Pt), by thermal diffusion in the epi layer prior the realization of the PIN diode
b) the irradiation with high energy electrons (or more recently with light ions - hydrogen or helium)

The energy levels of the recombination centers due to these techniques are indicated in figure: for Au and Pt the dominant one is near the midgap, while for electrons is in the upper bandgap



The irradiation of electrons or light ions can be done after making the PIN diode, because the accelerated ions are cable to pass through the metal or oxide layer and do not require too much thermal treatment after irradiation (only an heating of about 300°C is enough to reduce the residual crystal damage due to irradiation.
Instead the doping with heavy metal atoms require high temperature treatment (about 800 -900 °C) that can not be done after making the diode structure.

A further advantage of irradiation with H or He ions is the ability of reducing the lifetime only in a limited thickness of the layer, because the recombination centers (due to vacancy in the crystal) are mostly generated at the depth where the ion atoms are stopped, and then the position of the lifetime decrease can be modified by changing the energy of the irradiated ions.

The position of the lifetime minimum can give a better trade-off between voltage drop in the on-state and reverse recovery energy losses during switching; the best trade-off is obtained by a minimum lifetime position near the P^+N junction.