Principle of Semiconductor Devices Part I: Semiconductors, PN Junctions and Bipolar Junction Transistors

For the online version of these notes with better formatting, please visit
https://illusion.blog/notes/principle-of-semi-devices/part-1/1/

1. From Atom to Band Diagram

About energy band, different materials, carrier motion, and the water analogy.

About this series

This is a series of notes based on the video course ***Principle of Semiconductor Devices Part I: Semiconductors, PN Junctions and Bipolar Junction Transistors***.

Repost videos with Chinese subtitles is available on bilibili.

Images, unless otherwise stated, are taken from the course videos.

Energy Bands

- When atoms are put together, orbits at the same energy level are combined and form a energy band across the entire crystal.
- Electrons can stay at a wider range of energies instead of just one particular level.
- **Valence Band**: the energy band where electrons are shared.
    - All spaces that can hold electrons are filled with electrons.
- **Conduction Band**: the first empty energy band above the valence band.
    - Almost empty.
- **Band Gap**: the energy difference between the conduction band and the valence band.

    - No electron states exist in this energy range.

Metal, Insulator, Semiconductor

- **Metal**: valence band is only partially filled.
    - Electrons can move freely to conduct electricity.
- **Insulator**: full valence band and empty conduction band.
    - Cannot conduct electricity.
    - With energy supplied, electrons can be excited to the conduction band.
- **Semiconductor**: similar band diagram as insulator, but with a smaller band gap.
    - Room temperature is enough to excite a considerable amount of electrons to the conduction band.
- Material with a full valence band and an empty conduction band can be insulator or semiconductor, depending on the size of band gap and temperature of the working environment.

Carrier Motion

- **Diffusion**: particles move from high concentration to low concentration.
- **Drift**: particles move under the influence of an electric field.
- **They can co-exist**: electrons carry charges, accumulation of negative charges raises energy of electrons, which leads to drift and diffusion at the same time.
- Motion of electrons and holes is needed for electrons to move in an energy band, and the motion of the smaller quantity is easier to observe.

- In metals, electrons move in the partially filled band, and either electrons or holes are counted.
- In semiconductors, conduction happens both in the conduction band and the valence band, and carriers cannot be mixed. Current is the sum of both electrons and holes.
- The total number of carriers in the semiconductor is still much smaller than that in metals.

The Water Analogy

- Electrons when voltage is applied -> water molecules moving down the potential
    - **Tend to stay at the lower part of energy band**
- Holes when voltage is applied -> holes moving up the potential
    - **Tend to stay near the top part of energy band**

- Deficiency: water molecules and bubbles are electrically neutral, while electrons and holes are charged.

2. Calculating Carrier Concentrations

About density of states, Fermi-Dirac distribution, carrier density calculation, effective density of states, Boltzmann approximation, and water analogy for the bandgap.

- The density of carriers in a solid semiconductor depends on three factors:
  - density of state available in an energy band
  - size of band gap
  - temperature of the operation environment

Density of States

- **State**: a space to hold one electron, becomes a hole without an electron.
- Density of states are not uniformly distributed in an energy band.
  - Fewer states closer to the band gap
  - More states further away from the band gap
- $D(E)$: Distribution of states as a function of energy (density of states)
- E_C: lowest energy in the conduction band
- E_V: highest energy in the valence band
- $E_g = E_C - E_V$: band gap energy
  - Exact value of E_C and E_V does not matter, only the difference matters.
  - $E_g \approx 1.1\,\text{eV}$ for silicon

Fermi-Dirac Distribution

- Without energy supplied (absolute zero temperature), all electrons stay at lowest energy states.
- With energy supplied:
  - The distribution of carriers in the states of a band is governed by the Fermi-Dirac distribution function:

$$f_e(E) = \frac{1}{1 + e^{\frac{E - E_F}{kT}}}$$

the probability that an electron state at energy E is occupied by an electron.
    - E_F: Fermi energy level, a reference energy level
    - k: Boltzmann's constant
    - T: absolute temperature in Kelvin

More on Fermi-Dirac Distribution

When $T = 0\,\mathrm{K}$

$$f_e(E) = \begin{cases} 1, & E < E_F \\ 0, & E > E_F \end{cases}$$

It is an abrupt function.

When $T > 0\,\mathrm{K}$, the transition of $f_e(E)$ from 1 to 0 is smoothened.

Thus, E_F is **defined** as the energy level at which the probability of occupancy is 1/2, regardless of T.

- **Materials with a band gap**: E_F lies within the band gap, due to the symmetry of the probability distribution. ($\frac{1}{1+e^x} + \frac{1}{1+e^{-x}} = 1$)
- **In the valence band**: the probability of holes occupying a state

$$f_h(E) = 1 - f_e(E) = \frac{1}{1 + e^{\frac{E_F - E}{kT}}}$$

Calculating Carrier Density

- $n(E) = D(E) \cdot f_e(E)$: density of electrons at energy E, usually for conduction band
- $p(E) = D(E) \cdot f_h(E)$: density of holes at energy E, usually for valence band
- Total number:

$$n = \int D(E) \cdot f_e(E)$$
$$p = \int D(E) \cdot f_h(E)$$

$p = n$ because every electron excited to the conduction band leaves a hole in the valence band (electron-hole pair generation).

    - For pure or intrinsic silicon, $n = p = n_i$

Equivalent (Effective) Density of States

- We do not care about $D(E)$, we only care about the total number of carriers.
- Simplify:
  - Assume energy band is narrow, all states are at E_C or E_V.
  - Define an equivalent density of states N_C (total number of states if they are all located at E_C) and N_V (similar for valence band).
  - Then:

$$n = N_C \cdot f_e(E_C)$$
$$p = N_V \cdot f_h(E_V)$$

$$N_C = 2\left(\frac{2\pi m_e kT}{h^2}\right)^{3/2}$$
$$N_V = 2\left(\frac{2\pi m_h kT}{h^2}\right)^{3/2}$$

    - m_e: effective mass of electrons
    - m_h: effective mass of holes
    - h: Planck's constant
- For silicon at room temperature ($T = 25\,^\circ\mathrm{C}$):

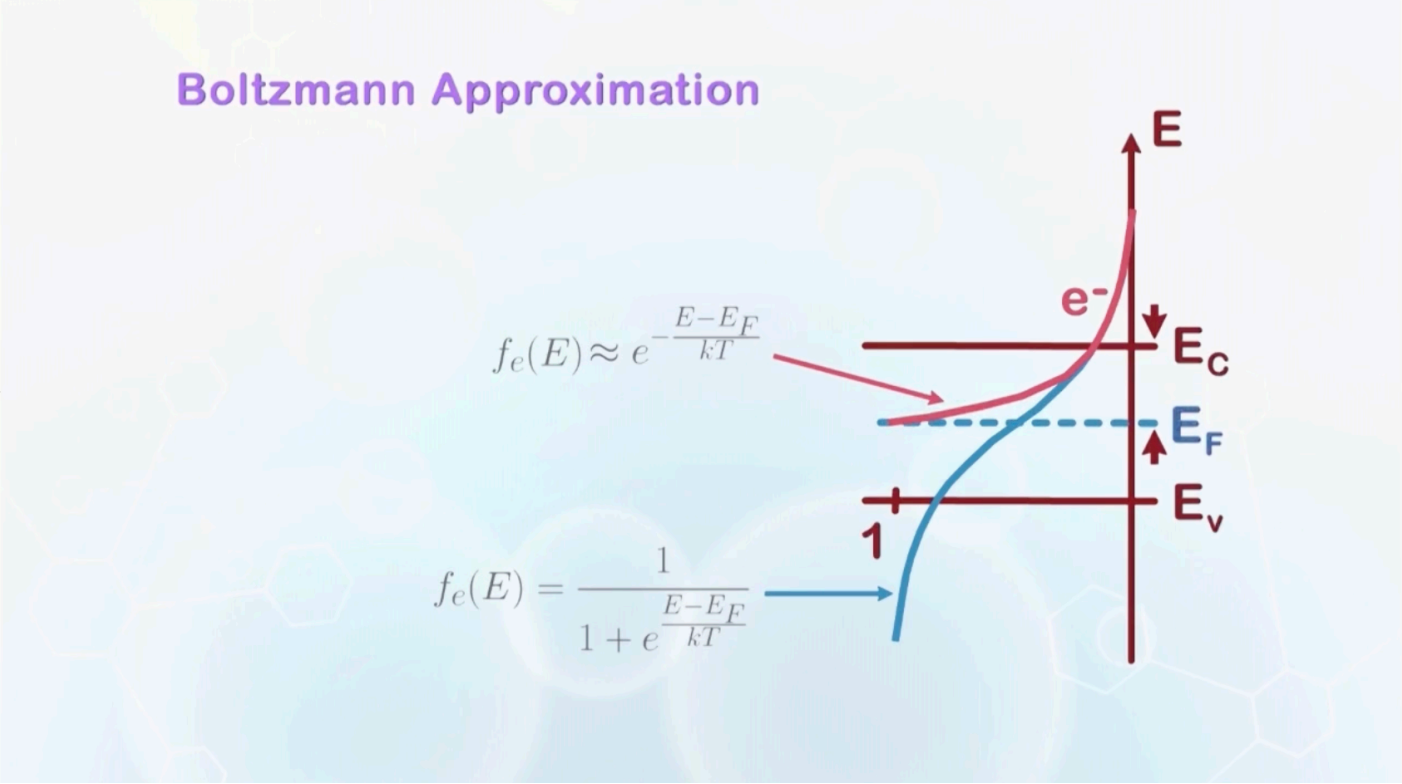
$$N_C = 2.8 \times 10^{19}\,\mathrm{cm}^{-3}, N_V = 1.09 \times 10^{19}\,\mathrm{cm}^{-3}$$

Boltzmann Approximation

- $kT \approx 0.026\,\mathrm{eV}$ at $25\,^\circ\mathrm{C}$, $E - E_F \gg kT$
- Thus,

$$f_e(E) = \frac{1}{1 + e^{\frac{E-E_F}{kT}}} \approx e^{-\frac{E-E_F}{kT}}$$

for conduction band when E_F is not too close to the conduction band.



- After the approximation, $f_h(E) = 1 - f_e(E)$ is no longer valid, and

$$f_h(E) \approx e^{-\frac{E_F - E}{kT}}$$

- Finally, the intrinsic electron concentration and the intrinsic hole concentration:

$$n = N_C \times f_c(E_C) = N_C \times e^{-\frac{E_C - E_F}{kT}} = n_i$$
$$p = N_V \times f_h(E_V) = N_V \times e^{-\frac{E_F - E_V}{kT}} = p_i$$

- Multiplying the two equations:

$$np = n_i^2 = N_C N_V e^{-\frac{E_g}{kT}}$$

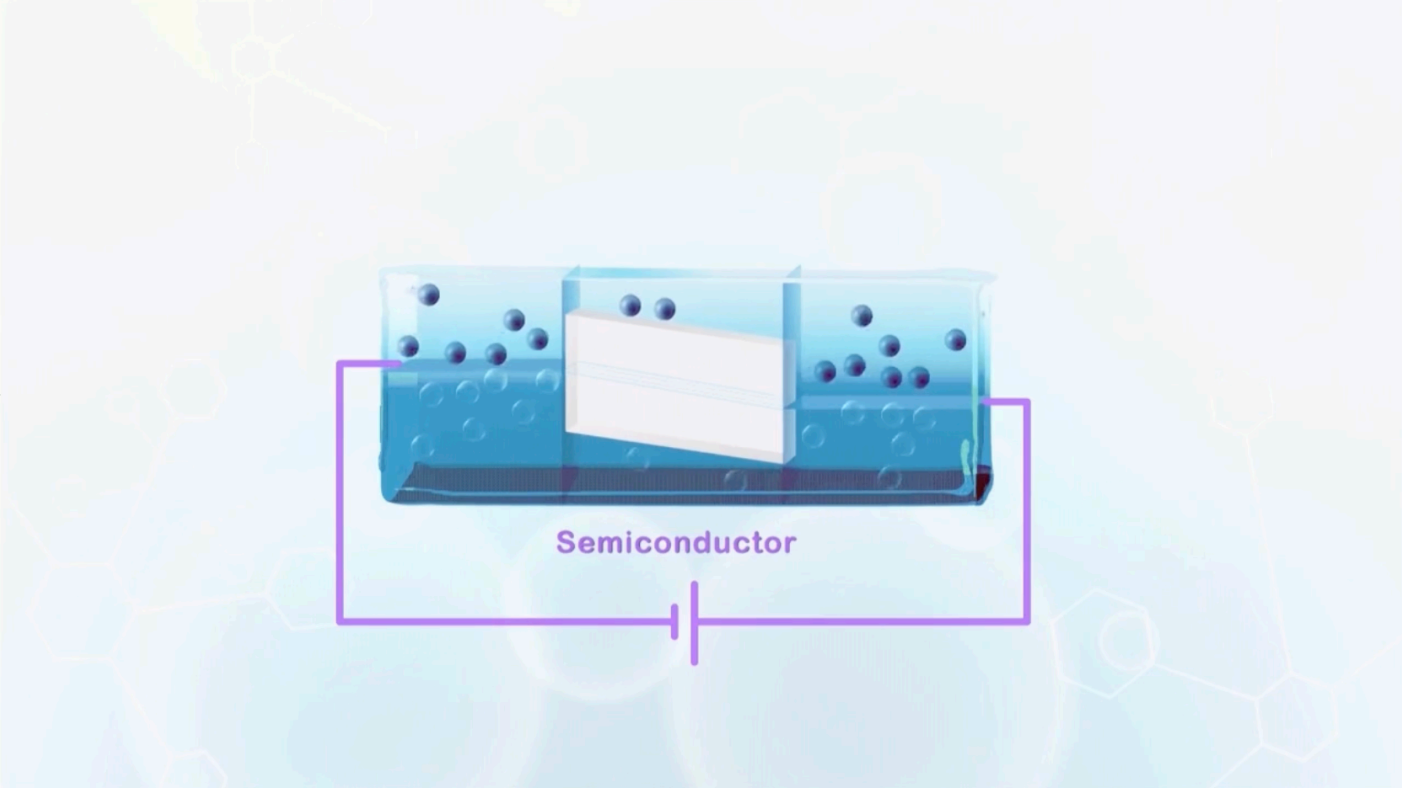
  This shows that the number of carriers increases with temperature, and decreases with band gap.
- Number of carriers of silicon at room temperature:

$$n_i = 1.45 \times 10^{10}\,\mathrm{cm}^{-3} \sim 10^{10}\,\mathrm{cm}^{-3}$$

Water Analogy for the bandgap

- A light, hollow, closed box partially filled with water.

- At the water to air interface, the probability to find a water molecule is **0.5**. This is the Fermi level of the box.
- Uniform external potential -> placing the box in a larger water tank where the water level represents the external potential.
- The box will float, aligning the water level inside the box with the water level outside -> The Fermi level is a reference energy level with respect to the surrounding.
- The band gap is a solid box without water molecules dropped inside the box, it will float in water, and the plane separating the floating part and sinking part is the Fermi level.
- The solid box has cracks, water molecules can jump above through the box through the cracks -> electrons excited from valence band to conduction band.
- External voltage applied -> external water level changes -> Fermi level at the two ends of the semiconductor changes -> current flows.



- Battery only controls the two ends, inside the semiconductor, the Fermi level is subject to the properties of the material.

3. Effects of Doping

About doping, dopant states, and how to calculate carrier density and locate the Fermi level in doped semiconductors.

Doping of Silicon

- Common group V elements used for doping: P and As
  - Others may not be suitable due to size mismatch
- Group V elements give an extra electron which is forced into the conduction band
  - Donor dopant
  - N-type silicon, N for **N**egative carriers
- Common group III elements used for doping: B
- Group III elements have one less electron, creating a hole in the valence band
  - Acceptor dopant
  - P-type silicon, P for **P**ositive carriers

Dopant States

- Donor dopants have an extra positive charge in the nucleus
  - Stronger attraction to electrons
  - Some electrons can stay in the bandgap where they are not allowed in intrinsic silicon
  - Introduces a **donor energy level E_D**
- Acceptor dopants have one less positive charge in the nucleus
  - Weaker attraction to electrons
  - Introduces a loosely bound hole, electrons can stay a bit further away from the nucleus at the **acceptor energy level E_A**
- $\text{Density}_{\text{Si}} \gg \text{Density}_{\text{dopants}}$ due to solid solubility limit in the amount of dopants that can be taken by silicon **before it loses intrinsic properties**
  - Locations of donor sites are very sparse
  - $E_C - E_D$ is very small, electrons can be easily excited and move around in the conduction band
  - Once the electron comes out of the dip, it may not be distinguished among others
    - We can assume all electrons coming from donor dopants will be delocalized and become electrons in the conduction band
- In N-type silicon, electrons significantly outnumber holes because of donor dopants
  - Electrons are the majority carriers
  - Holes are the minority carriers
- P-type silicon is similar, but with holes as majority carriers and electrons as minority carriers

Carrier Density

- When doped with both donor and acceptor dopants:
  - **Recombination**: electrons from donor dopants can recombine with holes from acceptor dopants
  - Cancel each other out instead of adding up the number of carriers
- When dopants are added to the system originally at thermal equilibrium (take donor dopants as an example):
  - Holes can more easily recombine with electrons
  - Newly added electrons $\gg$ number of holes
    - Number of electrons does not change much
    - Number of holes decreases significantly
  - At new thermal equilibrium:
    - $pn = n_i^2$ (always the case at thermal equilibrium)
    - $n = p + N_D$ (every hole **must come from an electron**)
    - $N_D \gg p$ (Majority carriers come from dopants)
    - Thus, $p = n_i^2 / N_D$

Locating the Fermi Level

- For N-type silicon, with additional electrons in the conduction band, the probability of finding an electron in the conduction band increases
  - Temperature is the same, the shape of the Fermi-Dirac distribution does not change
  - The Fermi level must have moved up to increase the probability of finding an electron in the conduction band
  - The Fermi level of intrinsic silicon is marked as E_i
  - $$\begin{aligned} n &= N_C e^{-\frac{E_C - E_F}{kT}} \\ n &= N_D \end{aligned}$$
    $$\Rightarrow E_C - E_F = kT \ln \frac{N_C}{N_D}$$
  - Equivalent density of states N_C does not change with doping
  - $$\begin{cases} n = N_C e^{-\frac{E_C - E_F}{kT}} & \text{in N-type silicon} \\ n_i = N_C e^{-\frac{E_C - E_i}{kT}} & \text{in intrinsic silicon} \end{cases}$$
  - Dividing the two equations:

$$n = n_i e^{\frac{E_F - E_i}{kT}}$$

or

$$E_F - E_i = kT \ln \frac{n}{n_i}$$

- For P-type silicon, the Fermi level moves down
  - $$\begin{aligned} p &= N_V e^{-\frac{E_F - E_V}{kT}} \\ p &= N_A \\ \Rightarrow E_F - E_V &= -kT \ln \frac{N_V}{N_A} \end{aligned}$$
  - $$\begin{aligned} &\begin{cases} p = N_V e^{-\frac{E_F - E_V}{kT}} & \text{in P-type silicon} \\ p_i = n_i = N_V e^{-\frac{E_i - E_V}{kT}} & \text{in intrinsic silicon} \end{cases} \\ \Rightarrow \quad & p = n_i e^{\frac{E_i - E_F}{kT}} \\ \Rightarrow \quad & E_i - E_F = kT \ln \frac{p}{n_i} \end{aligned}$$

Water Analogy for Fermi Level

- Water level represents the Fermi level
- When dopants are added, the water level rises or falls
- For N-type silicon, the weight of the box representing the semiconductor increases, pushing the box down and the water level up
- For P-type silicon, the weight of the box decreases, causing the box to rise and the water level to fall

4. PN Junction Formation

About formation of the PN junction, band diagram, depletion region width calculation, one-sided PN junction, and whether we can measure the built-in potential.

Doping of the PN Junction

- **Diode**: Putting P-type and N-type silicon together, forming a PN junction.
- Recall **recombination**:
  - When electrons coming from donors meet holes from acceptor, recombination happens.

    - After recombination, location of dopants carries a small positive charge, and location of acceptors carries a small negative charge.
- When we put P-type silicon on the left, and N-type silicon on the right:

```
   P   |   N
. . . .|. . . .
```

  we assume that recombination happens in an orderly manner, starting from the interface of the two types of silicon (the junction).

```
   P   |   N
. . . ⊣+ . . .
```

```
   P   |   N
. . - ⊣+ + . .
    ^^^^^^^
   depletion region
```

  areas near the junction are depleted of free carriers, forming a **depletion region**, where generally assumed to have no carriers.
    - Covalence bounds are similar to intrinsic silicon.
    - Assumption is not accurate, as there are still carriers in intrinsic silicon
    - The number is much smaller than the number of those added by doping, so we can ignore them.
- As recombination happens, charges accumulates, which tend to oppose further diffusion of carriers across the junction.
- An equilibrium will be reached eventually, with a particular depletion width.
- Simplifications we made:
    - **Abrupt junction approximation**: Assumes there is a clear boundary between P-type and N-type silicon, where a transition region (both donors and acceptors exist) exists in reality.
    - **Depletion approximation**: Assumes the depletion region is fully devoid of charge carriers and there is a clear boundary between the depletion region and the neutral region.

Band Diagram of the PN Junction

- When PN junction is formed, the Fermi levels of P-type and N-type silicon must align at equilibrium.
- In P side, $E_i > E_F$
- In N side, $E_i < E_F$
- The built-in potential V_{bi} satisfies:

$$qV_{\text{bi}} = qV_p + qV_n$$

  where $qV_p = E_{i_P} - E_F$, $qV_n = E_F - E_{i_N}$

- $$\begin{cases} E_{i_P} - E_F = kT \ln \frac{N_A}{n_i} \\ E_F - E_{i_N} = kT \ln \frac{N_D}{n_i} \end{cases}$$

- Therefore,

$$V_{\text{bi}} = \frac{kT}{q} \ln \frac{N_A N_D}{n_i^2}$$

Calculating the Depletion Region Width

- Depletion region width $x_d = x_p + x_n$
- Charge neutrality:

$$N_A x_p = N_D x_n$$

- Charge density:

$$\rho(x) = \begin{cases} -qN_A, & -x_p < x < 0 \\ +qN_D, & 0 < x < x_n \\ 0, & \text{elsewhere} \end{cases}$$

- Using Poisson's equation:

$$\nabla^2 V = -\frac{\rho}{\varepsilon}$$

- In 1D:

$$\frac{d^2V}{dx^2} = -\frac{\rho}{\varepsilon}$$

- Integrate twice:

$$V_{\text{bi}} = \frac{qN_Ax_p^2}{2\varepsilon_{\text{Si}}} + \frac{qN_Dx_n^2}{2\varepsilon_{\text{Si}}}$$

- Solve with equation (1) and (2):

$$x_p = \sqrt{\frac{2\varepsilon_{\text{Si}}V_{\text{bi}}}{q} \cdot \frac{N_D}{N_A(N_A+N_D)}}$$

$$x_n = \sqrt{\frac{2\varepsilon_{\text{Si}}V_{\text{bi}}}{q} \cdot \frac{N_A}{N_D(N_A+N_D)}}$$

- Finally:

$$x_d = \sqrt{\frac{2\varepsilon_{\text{Si}}V_{\text{bi}}}{q}\left(\frac{1}{N_A}+\frac{1}{N_D}\right)}$$

  where

$$V_{\text{bi}} = \frac{kT}{q}\ln\frac{N_AN_D}{n_i^2}$$

If you are an idiot just like me...

When $a > 0$ and $b > 0$,

$$\begin{aligned}&\sqrt{\frac{a}{b(a+b)}} + \sqrt{\frac{b}{a(a+b)}} \\ =&\frac{\sqrt{a^2}+\sqrt{b^2}}{\sqrt{ab(a+b)}} \\ =&\sqrt{\frac{a+b}{ab}}\end{aligned}$$

One-Sided PN Junction

- Doping on two sides are asymmetrical
  - Usually, a PN junction is formed by counter doping to convert part of a material to the opposite type
  - Counter doping concentration is much higher to minimize the background dopants
- In the depletion region expression, higher doping concentration term can be removed, and the depletion region width is mainly controlled by the lightly doped side

    - Graphically, the depletion region extends much more into the lightly doped side (PN junction is one-sided)

Measuring the Built-in Potential

- The built-in potential cannot be measured directly with a voltmeter
- The potential difference will be canceled out by the contact potential when the voltmeter is connected

5. Derivation of the Ideal PN Junction Diode Equation

About carrier statistics with respect to locations, external voltages, and the diffusion current with applied voltage.

Carrier Statistics with Respect to Locations

- In a P+/N junction:
    - Carrier concentration
        - $p_{p0} \cong N_A$
        - $n_{p0} \cong n_i^2/N_A$
        - $n_{n0} \cong N_D$
        - $p_{n0} \cong n_i^2/N_D$
        - Denotations: $\mathbf{CarrierType}_{\mathbf{Region}}$, and 0 means thermal equilibrium
        - We draw the graph in log scale as carrier concentration varies a lot
    - $p_{p0} > n_{n0}$ because of heavier doping on the P side
- Carrier concentration in the depletion region is not a constant, but varies with location
    - It is difficult to calculate, so we simply ignore it, assuming it is very small compared to the neutral regions

Carrier Motion at Thermal Equilibrium

> **Discussing only electrons**
>
> We will only be discussing electrons from now on, but the discussion applies to holes as well.

- No net current flow at thermal equilibrium, but carriers are not stationary
- In the conduction band
  - On the N side, higher electron concentration pushes electrons to **diffuse** to the P side
  - Some of the electrons are driven back by **drift**
  - There are many electrons on the N side, but most with energy lower than E_{Cp0} (E_C of the P side at thermal equilibrium). These electrons cannot move to the P side
  - Electrons that can move freely across the two sides are those with energy higher than E_{Cp0}
  - On the P side, concentration of these high-energy electrons are n_{p0}
  - On the N side, concentration of electrons is governed by the Fermi-Dirac distribution, where $E_{Cp0} - E_F$ is identical to that of the P side, so the concentration is also n_{p0}

Fermi-Level Under External Voltage

- The PN junction symbol

- **Positive Bias/Forward Bias**: Positive voltage V_A on the P side with respect to the N side
- **Negative Bias/Reverse Bias**: Negative voltage V_A on the P side with respect to the N side
- At **reverse bias**, alignment of the Fermi levels on the two sides is broken
  - Fermi level on the P side is raised by $q|V_A|$ with respect to the N side
  - Fermi level is the same across the neutral regions, but there is a discontinuity across the depletion region
  - In the depletion region, there are externally injected carriers, so the equilibrium cannot be assumed. Simple Fermi-Dirac distribution with one Fermi level does not apply
  - Must be handled with some advanced concepts of quasi-Fermi levels
  - For now, we will ignore the depletion region and focus on the neutral regions

Energy Band Bending Under Reverse Bias

- When a reverse bias is applied, we need to find the new barrier height and the new depletion region width
- For the barrier height:
  - At thermal equilibrium, the barrier height is qV_{bi}
  - Under reverse bias, the barrier height is raised by $q|V_A|$
  - New barrier height: $V_{Bh} = q(V_{\text{bi}} - V_A)$ as V_A is negative

    Why negative?

    V_A is the potential of the P side with respect to the N side, so it is negative under reverse bias

- For the depletion region width:
  - Charge neutrality still holds

$$N_A x_p = N_D x_n$$

  - The second equation becomes

$$V_{Bh} = V_{\text{bi}} - V_A = \frac{qN_A x_p^2}{2\varepsilon_{\text{Si}}} + \frac{qN_D x_n^2}{2\varepsilon_{\text{Si}}}$$

  - Solving the two equations, we get

$$x_d = \sqrt{\frac{2\varepsilon_{\text{Si}}(V_{\text{bi}} - V_A)}{q}\left(\frac{1}{N_A} + \frac{1}{N_D}\right)}$$

  - Assuming $N_A \gg N_D$ (P+/N junction), we have

$$x_d \approx \sqrt{\frac{2\varepsilon_{\text{Si}}(V_{\text{bi}} - V_A)}{qN_D}}$$

Carrier Concentration at Reverse Bias

- When an electron enters the edge of the depletion region from the P side, it moves to the N side due to the slope in the band diagram (caused mainly by **drift**)
  - Electron concentration at the edge of the depletion region on the P side is lower

    - More electrons from the rest of the P side **diffuse** to the edge of the depletion region
- When we draw the graph of carrier concentration
    - The concentration on the P side becomes lower as x moves to the depletion region
    - The concentration on the N side remains the same (n_{n0}) as the concentration is much higher than that caused by the few electrons coming from the P side
    - Similarly for holes, the concentration on the N side becomes lower as x moves to the depletion region, while the concentration on the P side remains the same (p_{p0})

Diffusion Current Under Reverse Bias

- Current calculation is achieved by counting the number of electrons crossing a particular location per unit time
- It is much easier to do so at locations with lower carrier concentration
- For electrons, we calculate the current on the P side near the edge of the depletion region
    - The driving force is mainly **diffusion**
    - Current density for electrons (current per unit area):

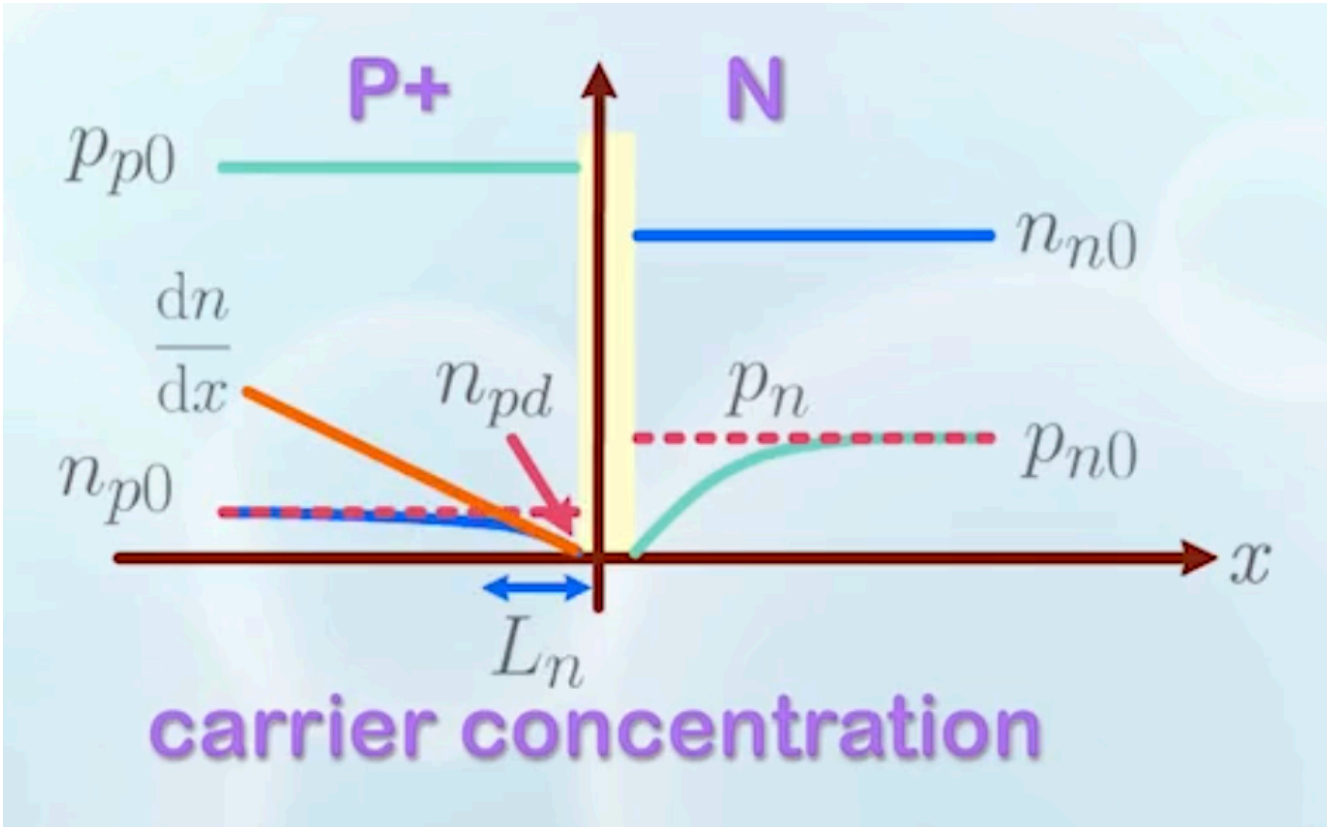
$$J_{n,\text{diff}} = qD_n \frac{\mathrm{d}n}{\mathrm{d}x}$$

      where D_n is the diffusion coefficient of electrons, indicating how mobile electrons are in the medium
        - When $\mathrm{d}n/\mathrm{d}x$ is large, it means carrier concentration differs a lot, so more carriers are moving across
    - The corresponding current for holes is

$$J_{p,\text{diff}} = -qD_p \frac{\mathrm{d}p}{\mathrm{d}x}$$

      where D_p is the diffusion coefficient of holes
        - The negative sign is because holes move in the opposite direction to electrons
- In the graph, $\mathrm{d}n/\mathrm{d}x$ is the slop of electron concentration on the P side near the edge of the depletion region (this marks a straight line)
    - n_{pd}: electron concentration at the edge of the depletion region on the P side. It is not a constant, and varies with the applied voltage V_A

- L_n: the distance the straight line intersects the n_{p0} line.



  It represents the average distance an electron can diffuse before recombining. We can assume it is a known constant at this stage.
- With these two notations, we have

$$\frac{dn}{dx} = \frac{n_{pd} - n_{p0}}{L_n}$$

  and

$$J_{n,\text{diff}} = qD_n \frac{n_{pd} - n_{p0}}{L_n}$$

  In this equation, all are known constants except n_{pd}

Reverse Bias Current of a PN Junction

- A few more assumptions:
  - The carrier statistics in the neutral **N** region is not affected by the bias voltage, as the number of electrons added is very small compared to the number of electrons already there
  - The depletion region is very small compared to the neutral regions, so we can ignore its thickness

- With these assumptions, n_{pd} equals to the carrier concentration in the neutral N region with energy higher than E_{Cp}
  - This is because concentration cannot change abruptly from P side to N side, as we've ignored the depletion region thickness
  - Calculate carrier concentration with Fermi-Dirac distribution

$$n_{pd} = n_{p0} e^{-\frac{E_{Cp} - E_{Cp0}}{kT}} = n_{p0} e^{-\frac{q|V_A|}{kT}} = n_{p0} e^{\frac{qV_A}{kT}}$$

- Finally,

$$\begin{aligned} J_{n,\text{diff}} =& qD_n \frac{n_{pd} - n_{p0}}{L_n} \\ =& qD_n \frac{n_{p0}}{L_n} \left(e^{\frac{qV_A}{kT}} - 1 \right) \end{aligned}$$

- The same for holes:

$$\begin{aligned} J_{p,\text{diff}} =& -qD_p \frac{p_{n0} - p_{nd}}{L_p} \\ =& qD_p \frac{p_{n0}}{L_p} \left(e^{\frac{qV_A}{kT}} - 1 \right) \end{aligned}$$

- Combine the two:

$$\begin{aligned} J =& J_{n,\text{diff}} + J_{p,\text{diff}} \\ =& q \left(D_n \frac{n_{p0}}{L_n} + D_p \frac{p_{n0}}{L_p} \right) \left(e^{\frac{qV_A}{kT}} - 1 \right) \end{aligned}$$

Conditions Under Forward Bias...

- Everything is the same as reverse bias, except that V_A is now positive
- When $V_A > V_{\text{bi}}$, the depletion region disappears, and the PN junction becomes a resistor, all voltage drops across the neutral regions
- We can denote

$$I_0 = q \left(D_n \frac{n_{p0}}{L_n} + D_p \frac{p_{n0}}{L_p} \right)$$

  which is a constant for a particular PN junction once the doping concentration is known
- The final equation is

$$I_D = I_0 \left(e^{\frac{qV_A}{kT}} - 1 \right)$$

6. Carrier Actions in a PN Junction

About how minority and majority carriers move in a PN junction with external voltage applied, and the difference between short and long diodes.

Review

$$I_D = I_0 \left(e^{\frac{qV_A}{kT}} - 1 \right)$$

$$I_0 = q \left(D_n \frac{n_{p0}}{L_n} + D_p \frac{p_{n0}}{L_p} \right)$$

Mechanism for Minority Carriers

- Doping is asymmetrical, so the current is dominated by the minority carriers on the lightly doped side
  - In the P+/N junction, holes on the N side dominate the current
  - Electrons generally move faster than holes in silicon, so which diode is faster?
    - A. P+/N
    - B. P/N+
    - Answer: @@B@@
    - Analysis: @@We want electrons to dominate the current, and electrons are the minority carriers on the P side, so we make the P side the lightly doped side.@@

Mechanism for Majority Carriers

What's the driving force for majority carriers?

- When an electron move from N to P, it has two effects:
  - Reduces concentration of electrons on the N side, causing diffusion
  - Results charge imbalance, creating a positive charge behind, causing drift
- The current density:
  - $J_{n,\text{diff}} = qD_n \frac{\mathrm{d}n}{\mathrm{d}x}$
  - $J_{n,\text{drift}} = q\mu_n nE$ where μ_n is the electron mobility (indicates how easily electrons move in the material)

- On the N side, electrons are majority carriers, so n is very large, while $\mathrm{d}n/\mathrm{d}x$ is small. Thus **drift** is the main driving force

- In general:
  - **Diffusion** is the main driving force to move **minority** carriers, because even a large E cannot cause a large current with the limited number of carriers
  - **Drift** is the main driving force to move **majority** carriers, because there are many carriers to move to balance out the effect of the charge imbalance

Short Diode Current

- When electrons get pushed from N to P, recombination rate will increase with a tendency to restore the equilibrium
- On average, an electron need to travel L_n before recombination
- When the diode is short, the neutral region length W_p and W_n may be smaller than L_n and L_p
  - $$W_p < L_n \quad \text{and} \quad W_n < L_p$$
  - Note that the subscript of L is different from that of W, as the subscript of L indicates the type of area, and the subscript of W indicates the type of carrier
- In this case, no combination can happen before an electron reaches the end of the diode, and the carrier distribution is a straight line
- At the two ends, the diode contacts with metal, and the carrier concentration is determined by metal, as it has a large number of carriers
- The metal will **force** carrier concentration to become n_{p0} and p_{n0} at the two ends
- The current density becomes:

$$J_{n,\text{diff}} = qD_n \frac{n_{p0}}{W_p}\left(e^{\frac{qV_A}{kT}} - 1\right)$$

$$J_{p,\text{diff}} = qD_p \frac{p_{n0}}{W_n}\left(e^{\frac{qV_A}{kT}} - 1\right)$$

- The electron current on the P side must be supported by the continuous electron **influx** from the N side, which is driven by **drift**, thus the drift current on the N side equals the diffusion current on the P side
- Similarly, the drift current of holes on the P side equals the diffusion current on the N side
- The total current density is the combination of the two, and must be a constant across the diode

Why Carrier Distribution is a Straight Line

- Within a short diode, no combination will happen, so the current flow must be a constant
- $J_{n,\text{diff}} \propto \frac{dn}{dx}$
- Thus, $\frac{dn}{dx}$ is a constant, and $n(x)$ is a straight line
- Similar analysis also applies to holes

Carrier Recombination in Long Diode

- In a long diode, carriers can recombine before reaching the end of the diode
- The carrier distribution is no longer a straight line
- The diffusion current decreases as x moves away from 0
- The total current must be a constant, so the decrease in diffusion current must be compensated by an increase in drift current
- The mechanism behind the increase in drift current:
  - When an electron recombines with a hole, the carriers disappear, leving a **negative** charge behind (a hole is eliminated in the P side neutral region)
  - The negative charge is quickly removed by the nearby holes
  - Eventually, the missing hole must be compensated by an **externally supplied** hole coming from the battery connected to the P+ side
  - This creates the drift current
  - The current can be considered as:
    - By **holes** from the end of P side to the recombination point
    - By **electrons** from the recombination point to the end of N side
    - It's like current changing lanes at the recombination point
    - Whenever a recombination happens, a drift current must be added to the **left**, but not the right, of the recombination point, thus the drift current increases as x moves away from 0

Does Recombination Increase or Decrease Current?

Answer: Recombination increases the current.

- Carriers face higher resistance when moving through low concentration areas

- When electrons move on the N side, they face low resistance, and when they move to the P side, they face higher resistance
- Recombination allows current to switch carriers from electrons to holes, which face lower resistance on the P side
- It provides a mechanism for a higher current flow

- By looking at the carrier distribution graph:
  - If there is no recombination, the carrier distribution is a straight line, and $L_n \to \infty$, $L_p \to \infty$, so the gradient is small
  - If there is recombination, L_n and L_p become smaller, so the gradient becomes larger
  - A larger gradient means a larger current
  - This additional current introduced by recombination is referred to as **recombination current**

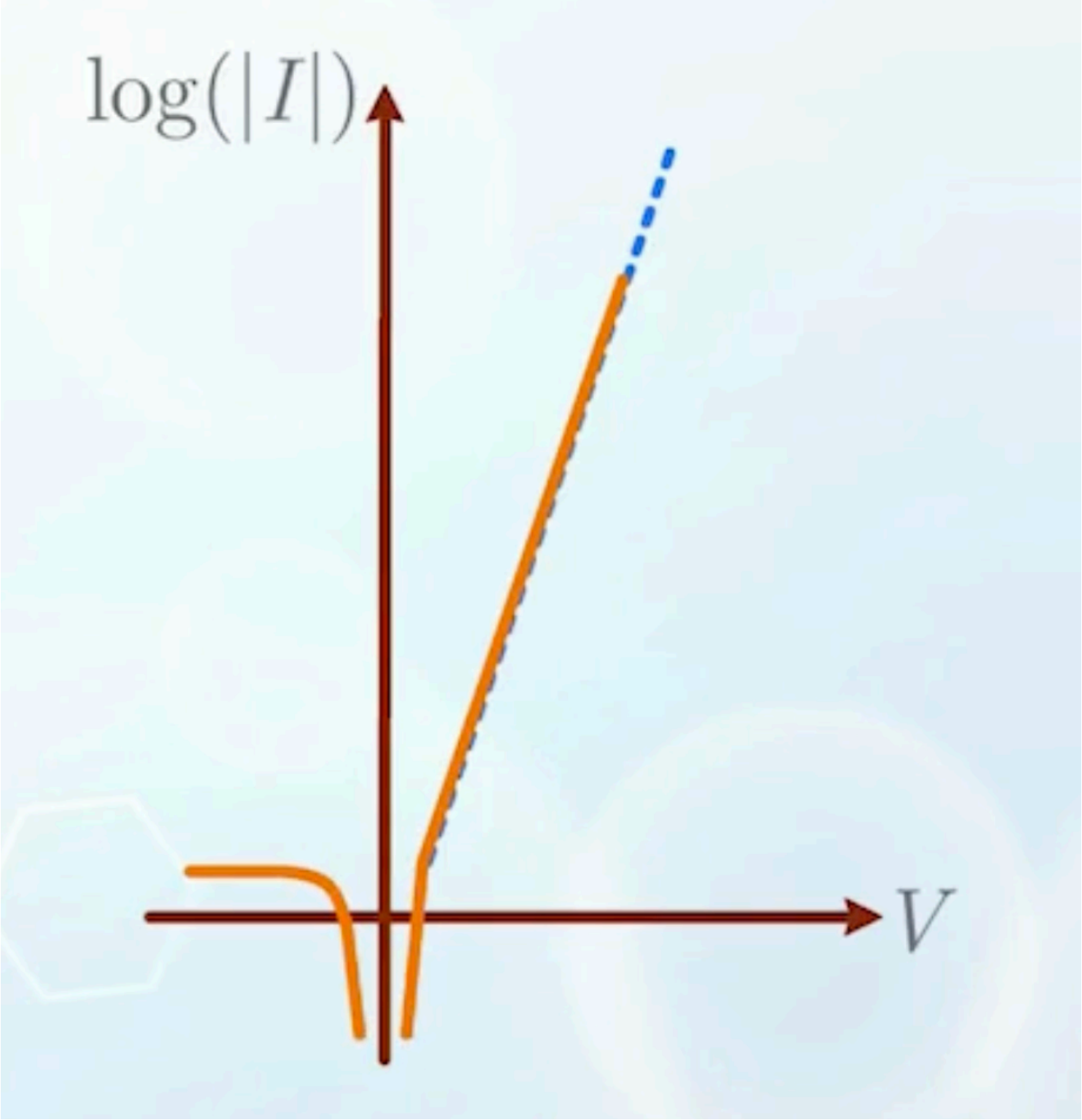
7. Real PN Junction Charateristics

About non-ideal PN junction characteristics, PN junction turn-on, breakdown, temperature effects, and how to design a PN junction.

Ideal Diode Current in Log Scale

$$I = I_D(e^{\frac{qV_A}{kT}} - 1)$$

$$\log(|I|) = \log I_D + \log(|e^{\frac{qV_A}{kT}} - 1|)$$



- When V_A is small, -1 takes dominance, and I decreases quickly in log scale
- When V_A is large, $e^{\frac{qV_A}{kT}}$ takes dominance, and I increases linearly in log scale
- The slope of the linear region can indicate how abruptly we can turn on the diode with applied voltage
    - $$\text{slope} = \frac{q}{kT} \log e$$

- The unit of slope in log scale is a bit difficult to express, so we tend to refer

$$\text{swing } (S) = \frac{1}{\text{slope}} = \frac{kT}{q} \ln 10$$

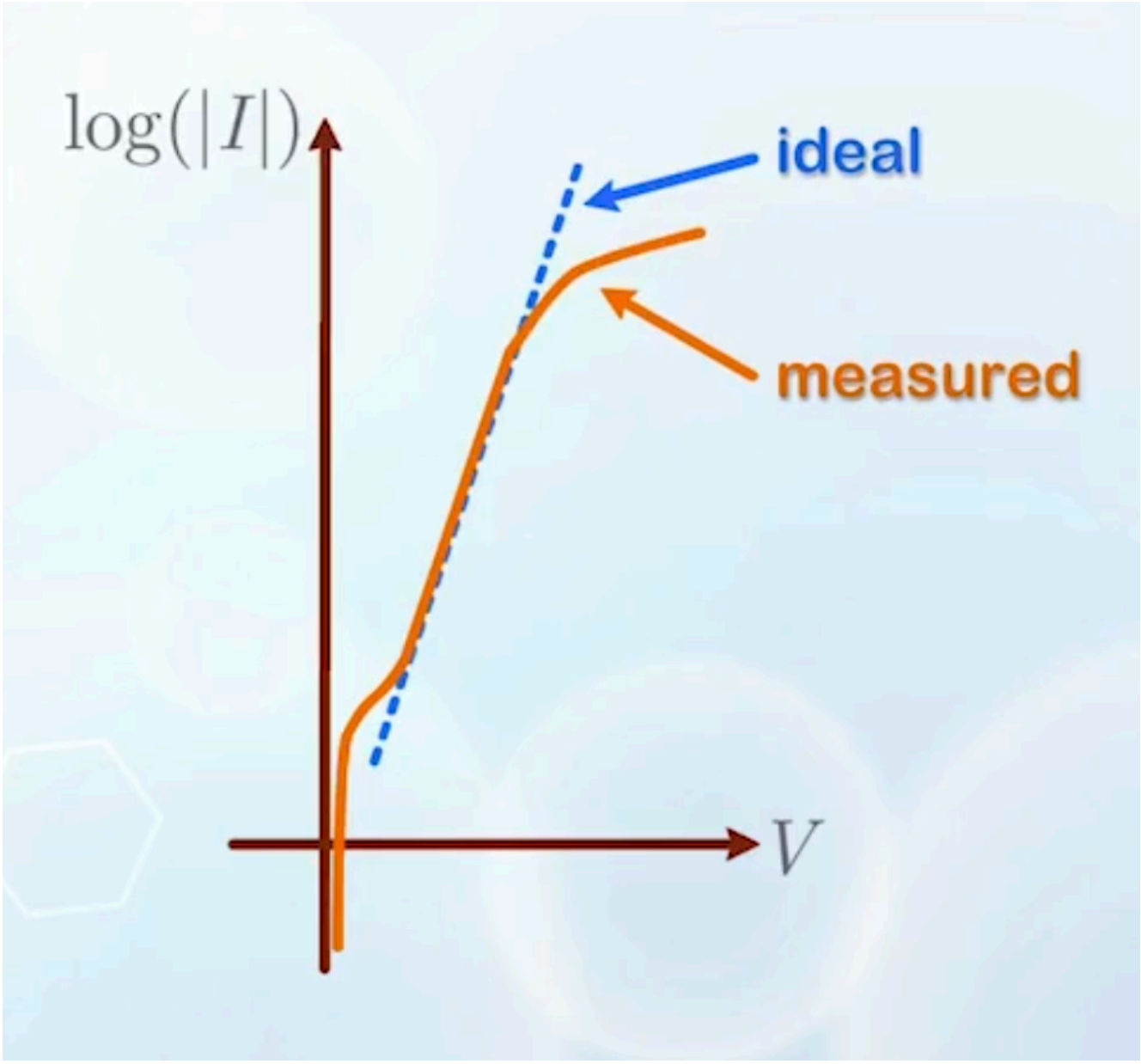
- Once T is known, S is fixed
- **A Very Important Reference**

  At room temperature:

  - $S = 60\,\text{mV/dec}$
    - This means that for every increase of $\mathbf{60\,mV}$ in the applied voltage V_A, the current I increases by 10 times (one decade).
    - **Remember this value**
    - A smaller swing means a steeper slope, which means the diode can be turned on more abruptly with a smaller change in voltage
  - $V_{\text{th}} = \frac{kT}{q} = 25\,\text{mV}$
    - Or $\mathbf{26\,mV}$, depending on the round off method
    - The **thermal voltage**
    - With the same unit as voltage

Recombination Current in the Depletion Region

- The characteristics of a real diode differs from the real one
  - At low current, measured current is usually higher than the ideal one
  - When V_A increases to close to V_{bi}, the increase in the measured current slows down

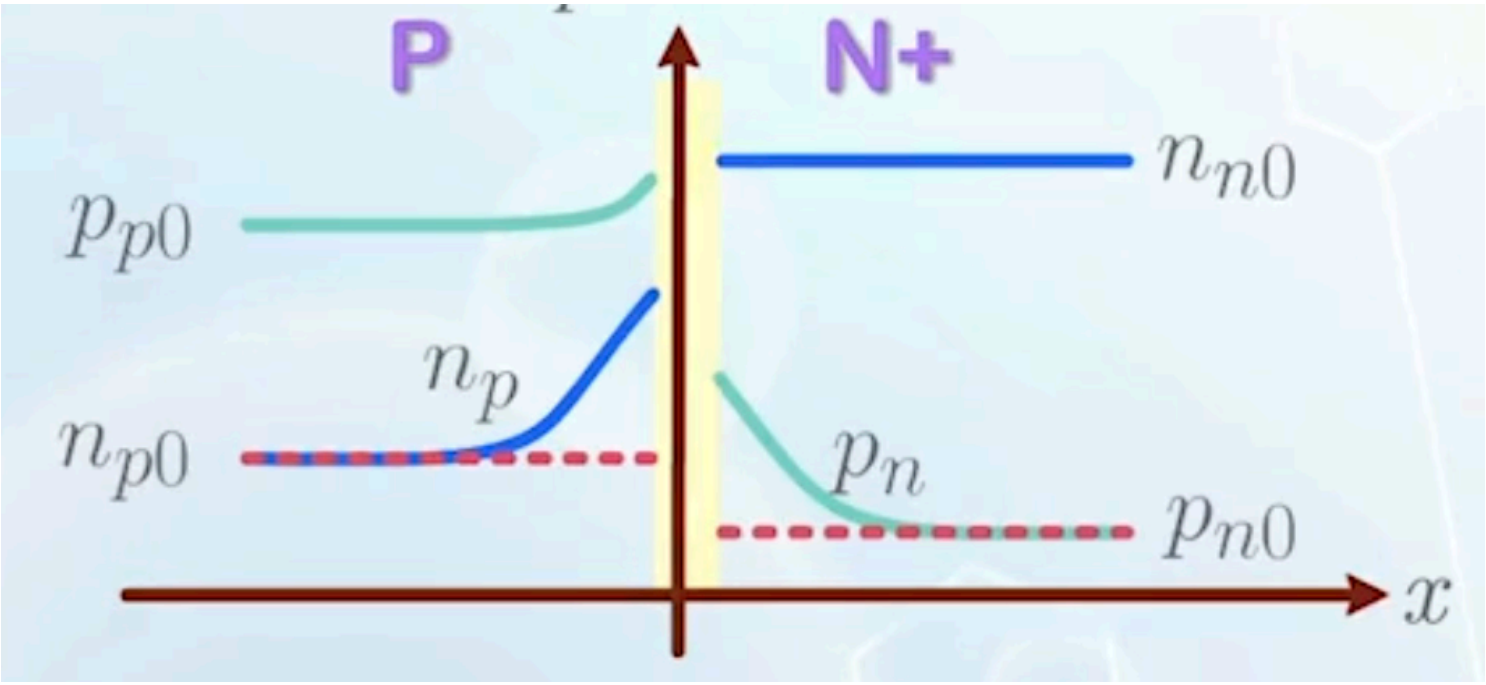


- The ideal equation assumes no recombination happens in the depletion region
- However, in reality, recombinations happen in the depletion region
- As discussed in the previous chapter, recombination is a mechanism that encourages a higher current flow
- Therefore, recombination in the depletion region leads to higher majority carrier current
  - This current can be ignored when V_A is large
  - When V_A is small, this current is observable

High Level Injection

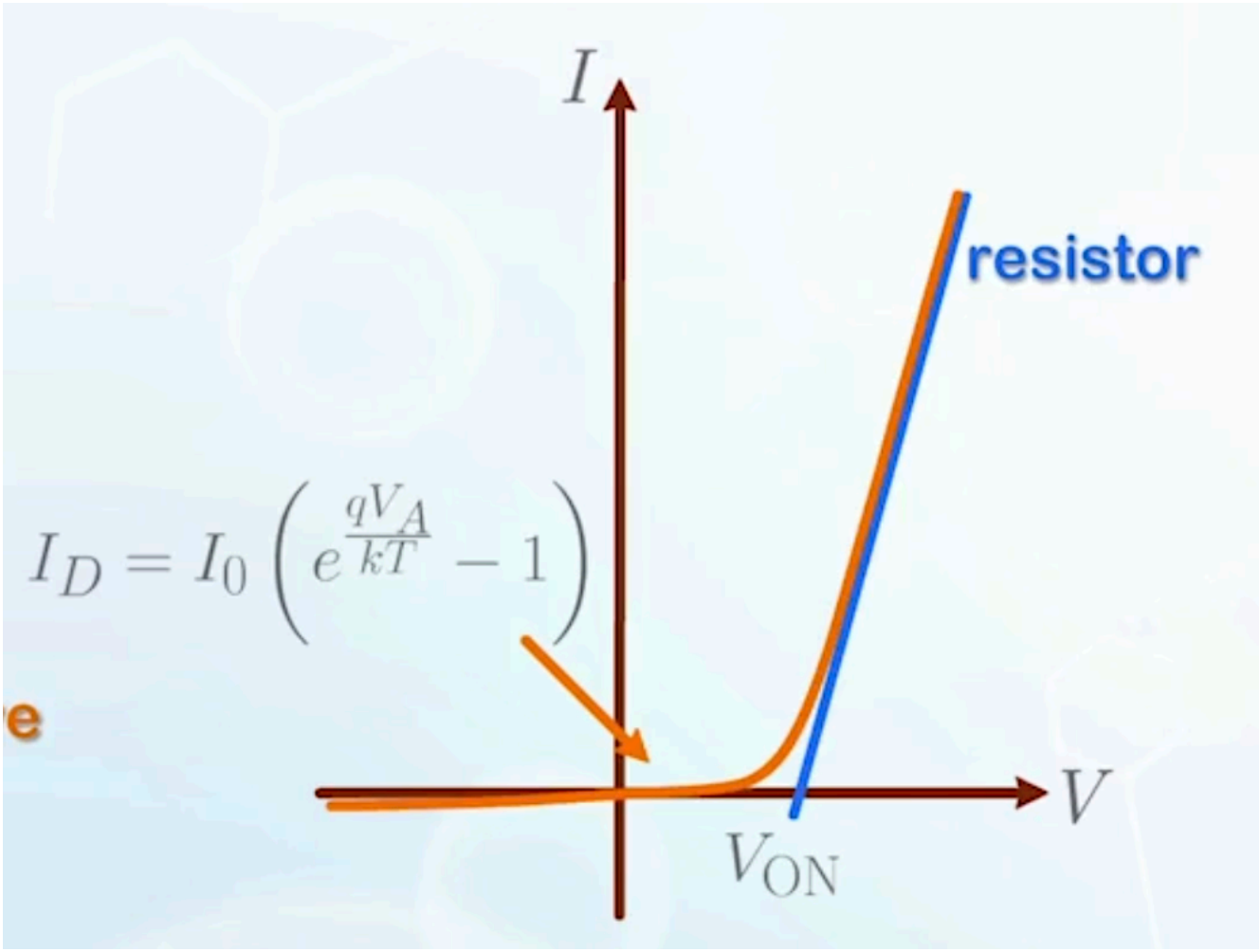
- The ideal equation assumes the number of injected minority carriers does not effect the majority carrier concentration

- However, when V_A is large, the number of injected minority carriers can be comparable to the majority carrier concentration at equilibrium
- On the P side:
  - Electrons are injected from the N side
  - High level electron injection causes accumulation of negative charges
  - This accumulation of negative charges repels holes, causing holes to accumulate on the P side near the depletion region
  - This creates a diffusion force which drives holes away **against** its motion in the ideal forward bias case
  - This is the high-level injection effect
  - The accumulation of holes is insignificant when V_A is small
  - But can be very significant when V_A is large
- The high-level injection effect always takes place first **at the lightly doped side** (because the majority carrier concentration is lower)



Complete PN Junction Turn-on

- When we further increase V_A beyond V_{bi}, the depletion region disappears
- As discussed in Chapter 1.5, the diode now behaves like a resistor
  - I becomes proportional to V_A
  - Gives a straight line in linear scale
  - Shows saturating behavior in log scale

- When you see the current increases with voltage, the diode is most likely to be **in the resistive region already**
    - So the ideal diode equation is only valid for the $I-V$ curve **close to zero**, or **before the diode turns on**
- **Fully turn-on region**: the resistive region when the depletion region disappears
- **Turn-on voltage V_{ON}**: Measured by extending the straight line in $I-V$ curve to cut the V axis
    - 

    - The turn-on voltage for a silicon diode is assumed to be around $\mathbf{0.7\,V}$

PN Junction Breakdown

- The ideal equation shows that the current remains small for any negative V_A
- However, in reality, when V_A is negative and large enough, the current increases rapidly in the negative direction
- This is called **reverse breakdown**
- Two different mechanisms can cause reverse breakdown
    - **Avalanche breakdown**

        - A high electric field is created in the depletion region when V_A is negative and large
        - This high electric field can accelerate carriers to very high speeds
        - These high-speed carriers can collide with atoms in the crystal lattice and generate electron-hole pairs
        - When the reverse bias is high enough, the additional electron-hole pairs, together with the original carriers, can create more collisions and generate even more electron-hole pairs, resulting in a large reverse current
        - It is like a avalanche caused by a small snowball
    - **Zener breakdown**
        - Recall the equation used to calculate the depletion region width:

$$x_d = \sqrt{\frac{2\varepsilon_{\text{Si}} V_{\text{bi}}}{q}\left(\frac{1}{N_A} + \frac{1}{N_D}\right)}$$

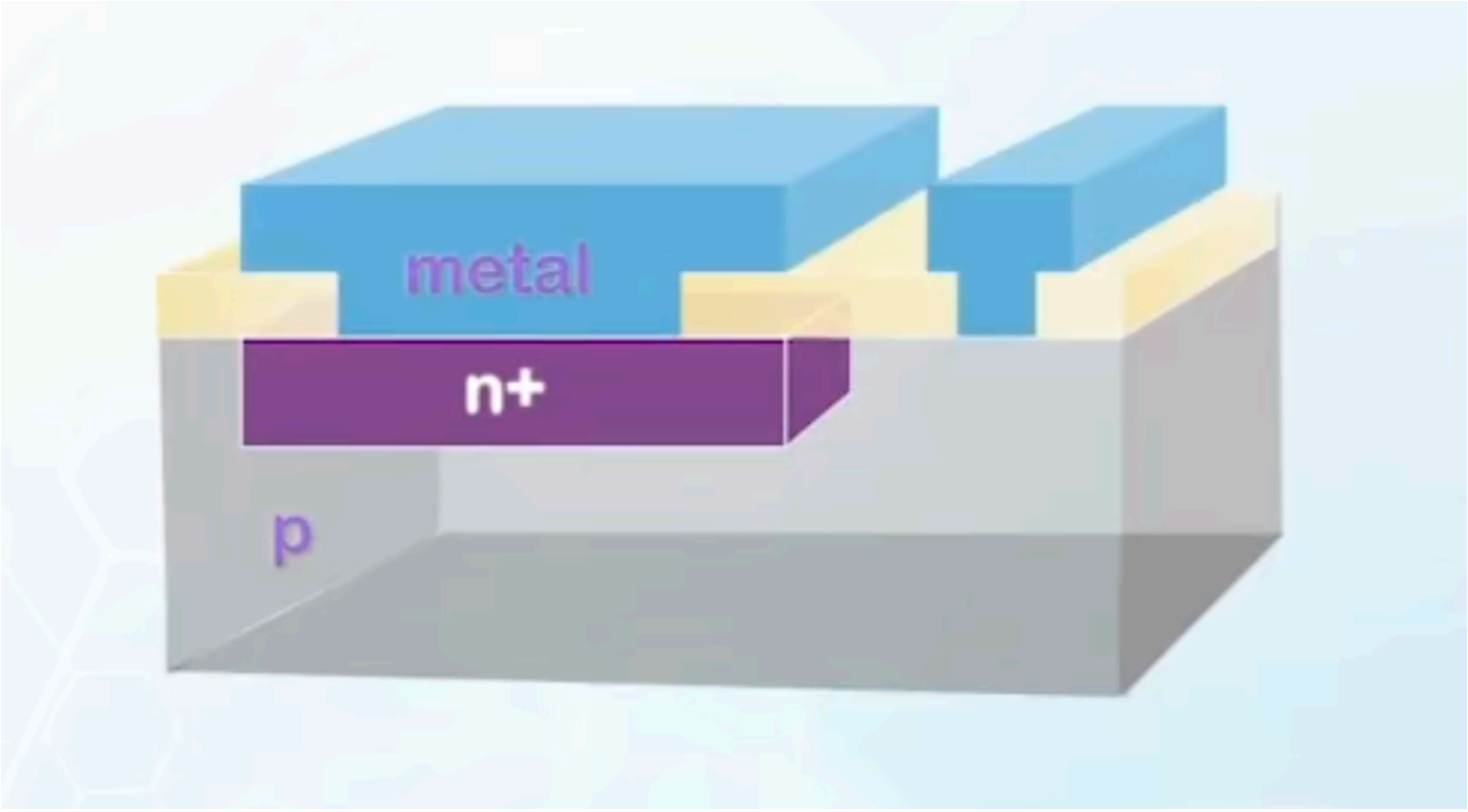
        - When the doping concentration is very high, the depletion region width can be very small
        - For a narrow junction with high reverse bias voltage, the lateral separation between the conduction band and the valence band can be very small
        - This allows electrons in the valence band to tunnel through the energy barrier to the conduction
        - This tunneling effect can create a large reverse current
- Breakdown may **not** be destructive
    - If the current is limited, the diode can recover after the reverse bias is removed
    - What really destroys the diode is the heat generated by the large current, melting the junction

Temperature Effects

- At high temperature, the impact of carriers added by doping is less significant
    - Both sides of the junction behave more like intrinsic semiconductors
    - V_{bi} and V_{ON} decreases
    - The junction will be less effective as a rectifier, more like a resistor
    - I_D increases, leading to a higher reverse saturation current
    - The slope of the ideal region decreases, leading to a larger swing S, making the diode more conductive, the current becomes less sensitive to voltage change, and is more **difficult** to turn **off**

- **Rectifying properties are degraded**

PN Junction Design



- The current follows a more complicated 2D or 3D pattern
- For the heavily doped side, the resistance is relatively low and close to ideal, it can be doped as heavily as possible, subjective to the solid solubility limit
- The lightly doped side controls the properties of the diode
  - If we simply connect metal to the lightly doped side, the series resistance will be very high
  - The metal-semiconductor contact resistance is usually very high
  - And the current may concentrate on a small area, due to the non-uniform resistance distribution
- We wish to decrease the resistance by adopting a higher doping concentration, but this may significantly decrease the breakdown voltage
- Design goal: select a high enough doping concentration to reduce the series resistance, while maintaining a high enough breakdown voltage

---

- The breakdown condition is related to the maximum electric field in the depletion region

- The avalanche breakdown voltage can be plotted against the doping concentration on the lightly doped side
- Once the doping concentration passes a certain value, the Zener breakdown will take over
  - Once the **electric field/slope of the band** is known, the lateral separation between the conduction band and the valence band (tunneling distance) can be calculated
- It is important to understand how the electric field changes with doping concentration

---

- Recall how we calculate the depletion region width
  - $$x_{p/n} \approx \sqrt{\frac{2\varepsilon_{\text{Si}}(V_{\text{bi}} - V_A)}{q}\frac{1}{N_{A/D}}}$$
  - According to Gauss's law, the electric field in the depletion region is

    $$\frac{\mathrm{d}E}{\mathrm{d}x} = \frac{\rho}{\varepsilon_{\text{Si}}}$$

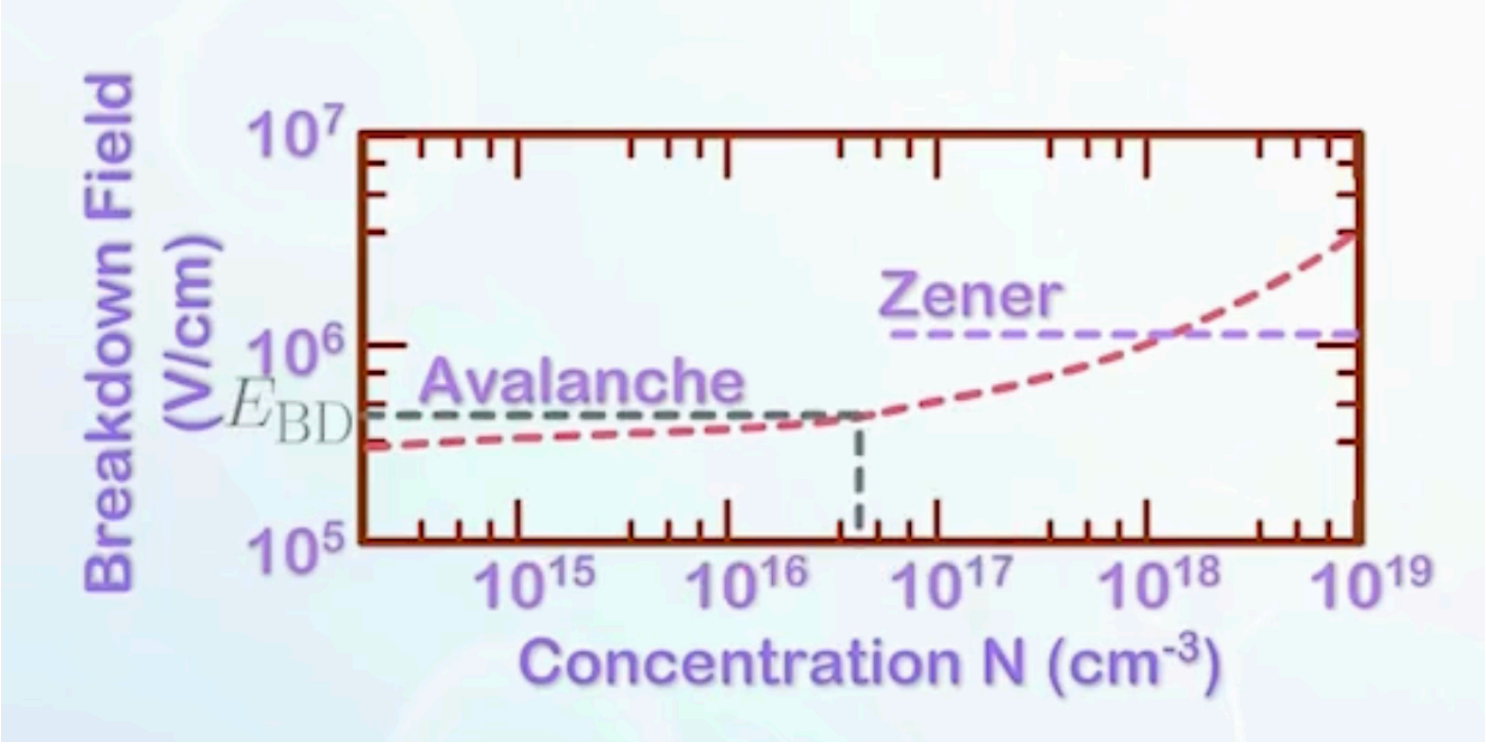
    Thus

    $$\vec{E}_{\text{max}} = -\frac{qN_{A/D}x_{p/n}}{\varepsilon_{\text{Si}}}$$

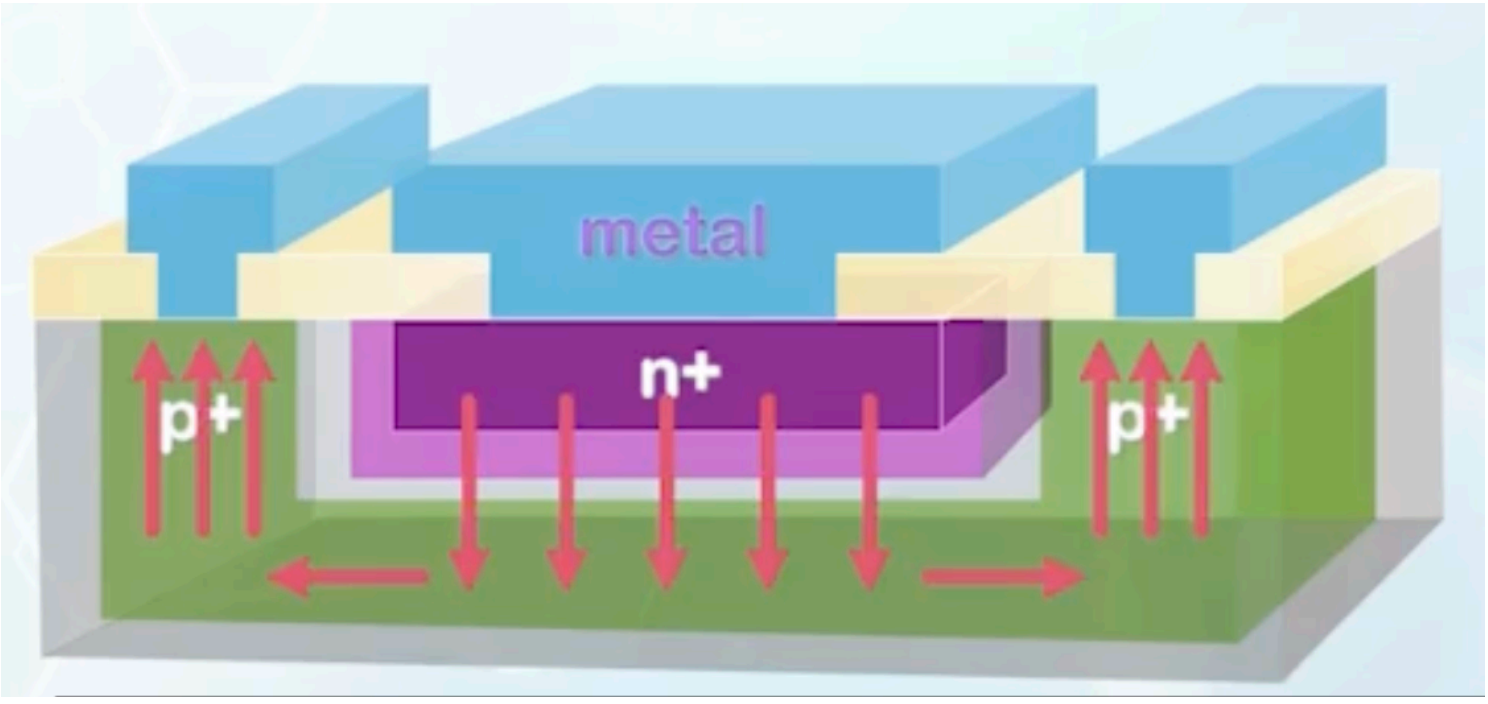
  - Now we have

    $$\vec{E}_{\text{max}} = -\sqrt{\frac{2qN_{A/D}(V_{\text{bi}} - V_A)}{\varepsilon_{\text{Si}}}}$$

- Now we have the relationship between the maximum electric field and the doping concentration
- To choose a proper doping concentration, we begin with an arbitrary concentration
- Then we can calculate $|\vec{E}_{\text{max}}|$
- We compare $|\vec{E}_{\text{max}}|$ with E_{BD}, which can be found in the avalanche breakdown curve
- If $|\vec{E}_{\text{max}}| < E_{\text{BD}}$, the doping concentration is acceptable
- Else, we need to reduce the doping concentration and repeat the process
- As an engineer, make sure to leave enough margin



- Besides the depletion region, a diode also consists of neutral regions
- The neutral regions contribute little to the rectifying properties of the diode, but contribute to the series resistance
    - We can increase the doping concentration to reduce the resistance
- We can also add another P contact to make the resistance more uniform at different locations



8. PN Junction Switching and Model

About the charge storage effects, PN junction diode models, and parameter extraction.

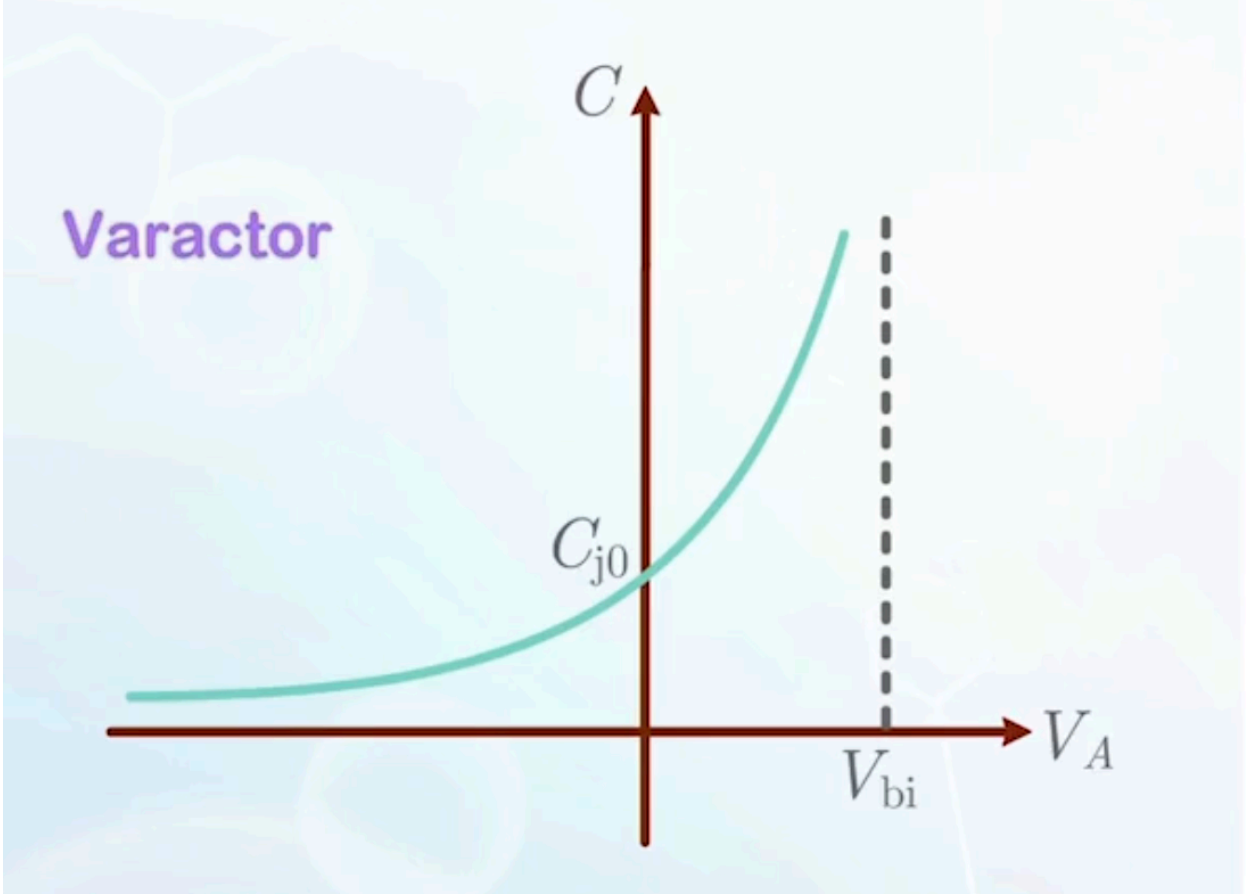
Reverse Bias Junction Capacitance

- When switching a diode with time varying voltage, response may be delayed
  - Charges need to be accumulated to reach steady state
  - The delay can be modeled as a **capacitance** within the diode
- Under reverse bias, the diode can be modeled as an insulator sandwiched between two conductors, or a parallel-plate capacitor
  - The depletion region is the insulator
  - The P and N regions are the conductors
  - Its capacitance is denoted as C_j
  - $$C_j = \frac{\varepsilon_{\text{Si}} A}{W_d}$$

    where A is the cross-sectional area of the junction, and W_d is the width of the depletion region.
  - We normalize it with respect to area so we can drop A
  - Substituting W_d from previous section, we have

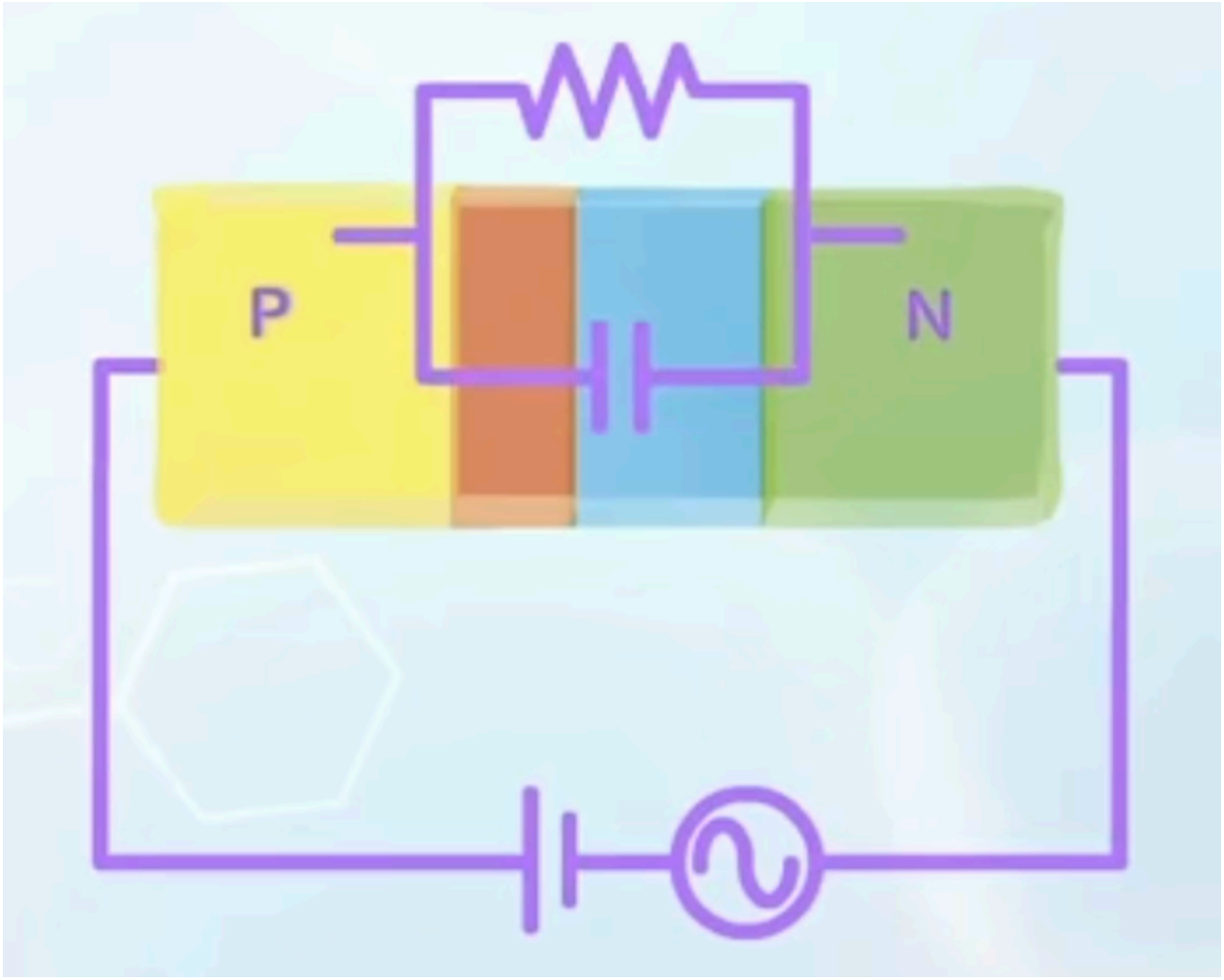
$$\begin{aligned} C_j &= \frac{\varepsilon_{\text{Si}}}{\sqrt{\frac{2\varepsilon_{\text{Si}}(V_{\text{bi}} - V_A)}{q}\left(\frac{1}{N_A} + \frac{1}{N_D}\right)}} \\ &= \frac{\varepsilon_{\text{Si}}}{\sqrt{\frac{2\varepsilon_{\text{Si}} V_{\text{bi}}}{q}\left(\frac{1}{N_A} + \frac{1}{N_D}\right)}} \frac{1}{\sqrt{1 - \frac{V_A}{V_{\text{bi}}}}} \\ &= \frac{C_{j0}}{\sqrt{1 - \frac{V_A}{V_{\text{bi}}}}} \end{aligned}$$

    C_{j0} is bias voltage independent, representing the junction capacitance at equilibrium (zero bias).
  - The function shows that
    - Lower doping concentration leads to lower junction capacitance, due to wider depletion region
    - Capacitance of a diode can be changed by applying different voltages, making it a **varactor** (variable capacitor), which can select a signal at a specific frequency, but rejecting others.

- 

Forward Bias Diffusion Capacitance

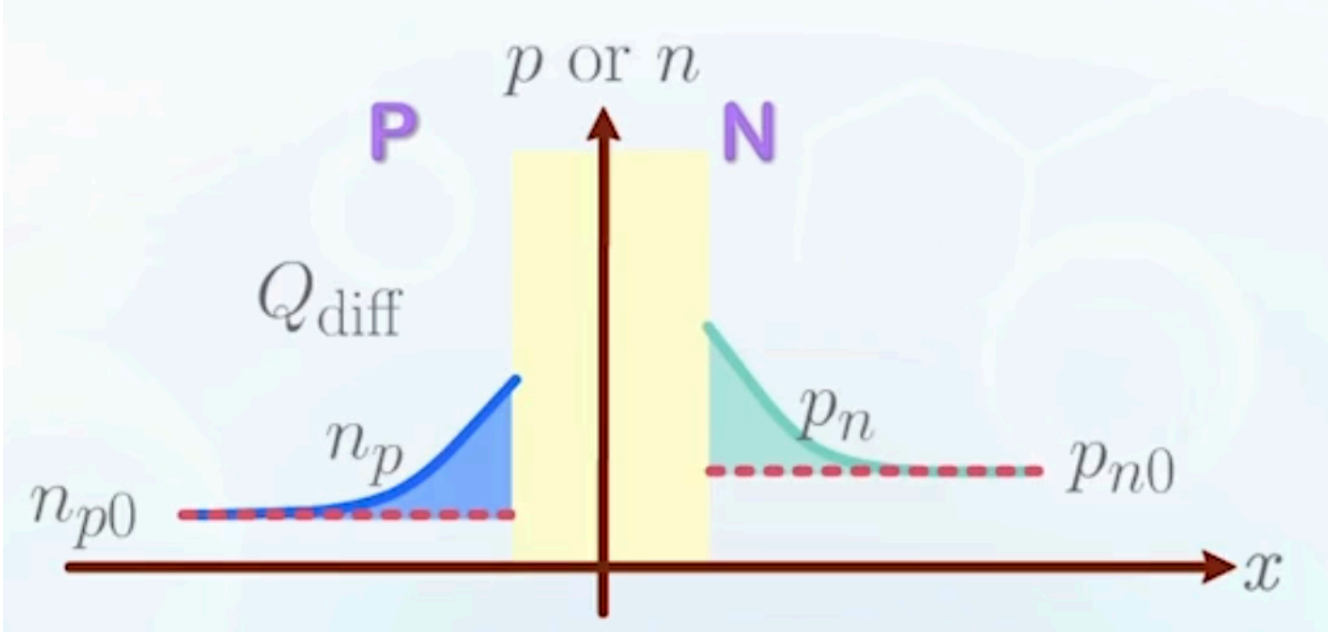
- Under forward bias, a current conduction path is established
  - The parallel-plate capacitor model is no longer valid
  - The diode can be modeled as a resistor in parallel with a capacitor



- For non-linear capacitors (as the case here), capacitance is defined as

$$C = \frac{\mathrm{d}Q}{\mathrm{d}V}$$

  where Q is the charge stored in the diode, and V is the applied voltage.
- A PN junction under forward bias stores two kinds of charge
  - The charges in the depletion region, which has already been modeled as C_j in the previous section
  - The charges of excess minority carriers injected from the opposite sides, temporarily stored before leaving (small under reverse bias, so we ignored it)

- 

- Carrier concentration at thermal equilibrium is not counted as stored charge since the system is electrically neutral
- The excess minority carrier charge (**diffusion charge**) in P side is

$$Q_{\text{diff},n} = \frac{1}{2} q (n_{pd} - n_{p0}) L_n$$

  which is the size of the shaded area in the graph approximated as a triangle
  - Assuming injected carriers $\gg$ equilibrium carriers, we can drop n_{p0}

$$\begin{aligned} Q_{\text{diff},n} &= \frac{1}{2} q n_{pd} L_n \\ &= q \frac{L_n}{2} n_{p0} e^{\frac{qV_A}{kT}} \end{aligned}$$

- Adding the contribution from N side, we have the total diffusion charge

$$Q_{\text{diff}} = \frac{q}{2} (L_n n_{p0} + L_p p_{n0}) e^{\frac{qV_A}{kT}}$$

- Differentiating it with respect to V_A, we have the diffusion capacitance

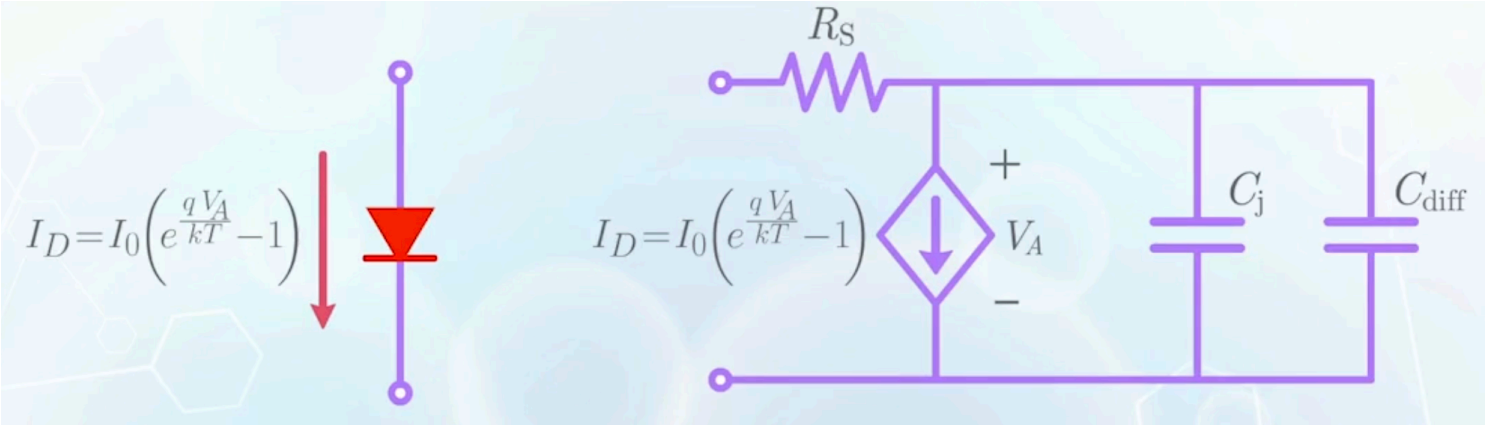
$$C_{\text{diff}} = \frac{\mathrm{d}Q_{\text{diff}}}{\mathrm{d}V_A} = \frac{q^2}{2kT} (L_n n_{p0} + L_p p_{n0}) e^{\frac{qV_A}{kT}}$$

  Or simplified as

$$C_{\text{diff}} = \frac{qQ_{\text{diff}}}{kT} = \frac{Q_{\text{diff}}}{V_{\text{th}}}$$

  where $V_{\text{th}} = \frac{kT}{q}$ is the thermal voltage

Large Signal PN Junction Model



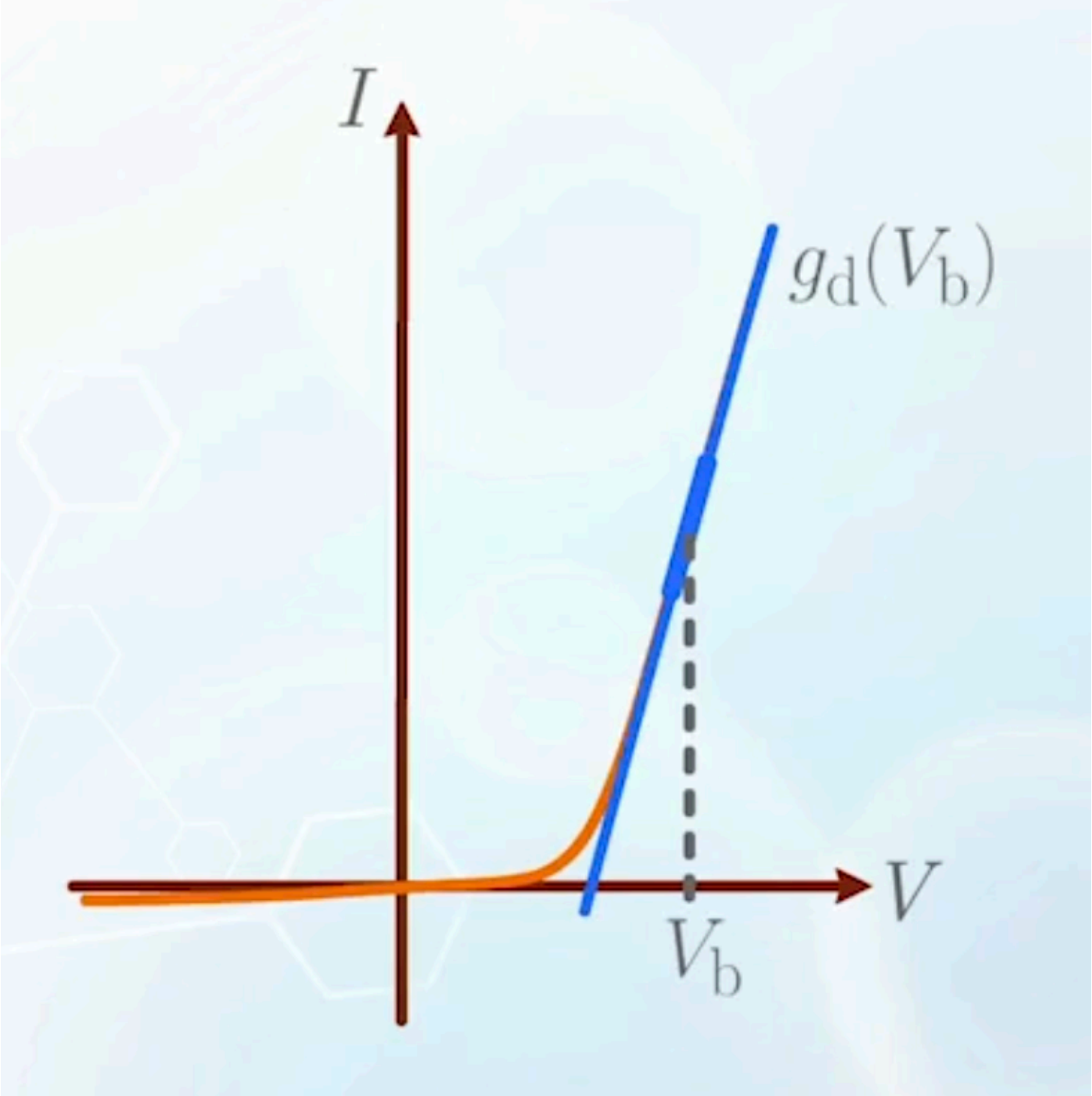
- The ideal PN junction current-voltage relationship is

$$I_D = I_0\left(e^{\frac{qV_A}{kT}} - 1\right)$$

  which can be represented as a voltage-controlled current source
- The neutral regions can be represented as resistor R_S
- To account for the charge storage effects, we add two capacitors in parallel
  - C_j for junction capacitance
  - C_{diff} for diffusion capacitance
  - C_j dominates under reverse bias, while C_{diff} dominates under forward bias, so we may only need to calculate one of them depending on the bias condition

Small Signal PN Junction Model

When PN junction can only be operated over a small region around a bias voltage V_b



- The $I-V$ relationship can be approximated as the tangent of curve at that bias point
- The voltage-controlled source can be replaced with a resistor whose conductance is $G = g_d(V_b)$
- g_d can be obtained by differentiating ideal diode equation with respect to voltage

$$
\begin{aligned}
g_d(V_b) &= \left.\frac{\mathrm{d}I_D}{\mathrm{d}V_A}\right|_{V_A=V_b} \\
&= \left.\frac{\mathrm{d}I_0\left(e^{\frac{qV_A}{kT}}-1\right)}{\mathrm{d}V_A}\right|_{V_A=V_b} \\
&= \frac{q}{kT} I_0 e^{\frac{qV_b}{kT}} \quad \text{(simplified by dropping -1)} \\
&= \frac{I_D(V_b)}{V_{\text{th}}}
\end{aligned}
$$

- All other components (R_S, C_j, C_{diff}) are evaluated at V_b and becomes a fixed value

PN Junction Diode Parameter Extraction

A sample data sheet provided by diode manufacturer:

CHARACTERISTICS

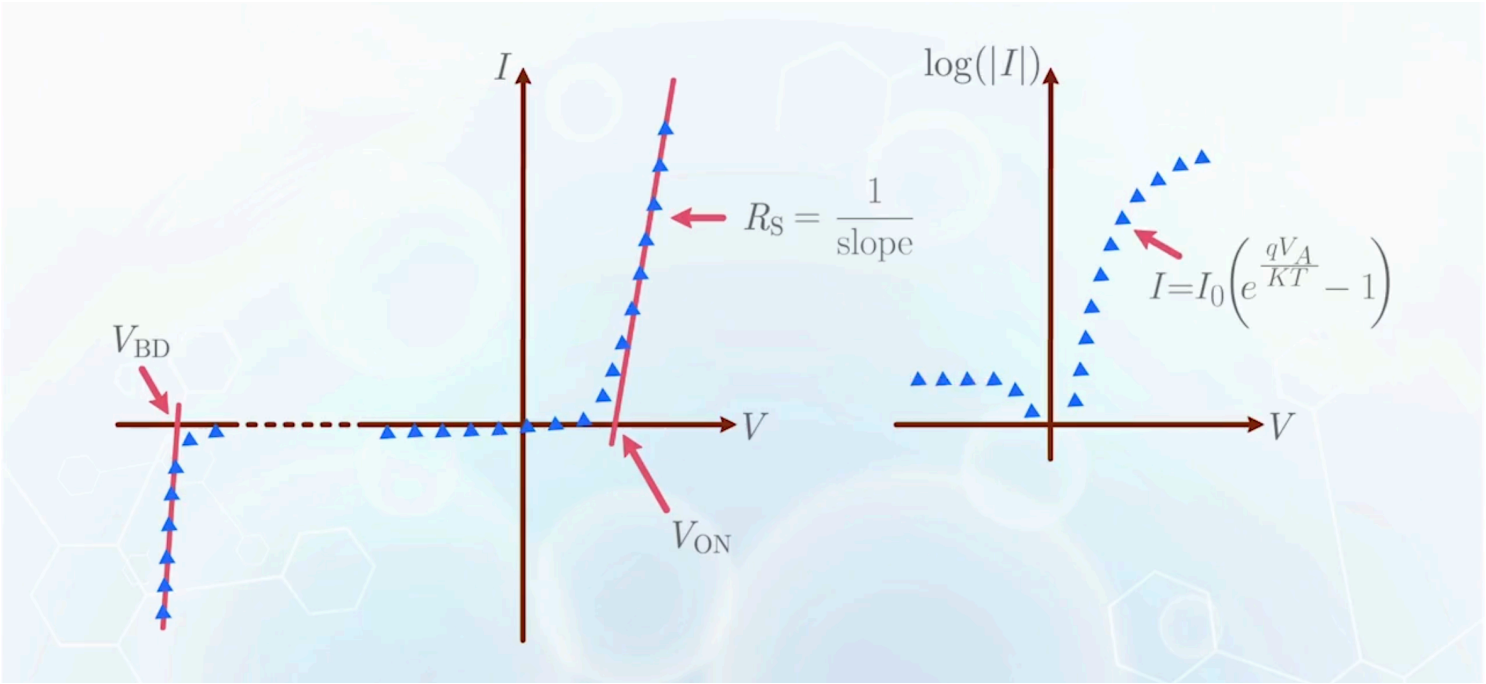
$T_j = 25\,^\circ\text{C}$ unless otherwise specified$

SYM.	PARAMETER	CONDITIONS	MAX.	UNIT
V_F	forward voltage	$I_F = 100\,\text{mA}$	1	V
I_R	reverse current	$V_R = 200\,\text{V},\ T_j = 150\,^\circ\text{C}$	100	nA
C_d	diode capacitance	$f = 1\,\text{MHz},\ V_R = 0$	2	pF
V_{BR}	reverse breakdown voltage		300	V
R_S	series resistance	$V_F = 2\,\text{V}$	3	Ω
n	ideality factor	$V_F = 0.5\,\text{V}$	1.05	
t_{rr}	reverse recovery	when switched from	50	ns

SYM.	PARAMETER	CONDITIONS	MAX.	UNIT
	time	$I_F = 300\,\text{mA},\ R_L = 100\,\Omega$		

These values are mesured instead of calculated, and the process is called **parameter extraction**.

It is the process to find the values of these unknown parameters so that the values predicted by the model give the best fit to the experimental data



- I_0 can be read from the reverse saturation current in the log scale graph
- In the forward bias region, before the diode fully turns on, the ideal diode equation may not match the experimental data, due to non-ideal effects
  - The ideality factor n is introduced to account for these effects
  - The modified ideal diode equation is

$$I_D = I_0\left(e^{\frac{qV_A}{nkT}} - 1\right)$$

  - For an ideal diode, $n = 1$, while for a real diode, n is typically between 1 and 2

9. Optical Property of PN Junctions

About the optical properties of PN junctions, how to apply them to photo detectors, solar cells, and LEDs, and how to design these devices for better performance.

Optical Properties of Semiconductors

- Electron-hole pairs will be generated when energy is supplied
  - The energy can come from thermal energy
  - But can also come from light (particles called photons)
    - The energy is related to its wavelength
    - $$\text{energy} = \frac{hc}{\lambda}$$
    - With given units:

$$\text{energy} = \frac{1.24}{\lambda\,(\mu\text{m})}(\text{eV})$$

- If the photon energy is higher than the bandgap energy, it will be absorbed and excite an electron from valance band to conduction band
- If the photon energy is lower than the bandgap energy, it will not be absorbed, but transmitted through the material
  - Explains why transparent materials are usually insulators with large bandgap
- Wavelength of visible light is from 0.4 to $0.7\,\mu\text{m}$, corresponding bandgap is from 3.1 to $1.8\,\text{eV}$
  - Silicon absorbs light from infrared to the entire visible spectrum
- When electron-hole pairs are generated by light absorption, generated carriers tend to move together by diffusion, so **the net current is 0**. They will eventually recombine in a region without light
- But if light is shining on the depletion region of a PN junction, the electric field will cause electrons and holes to move in opposite directions, leading to a current when both sides of junction are shorted, having the same voltage
  - The short-circuit current I_{SC} or I_{photo}, proportional to illumination intensity
  - The current flows from N side to P side, so it is **negative** according to the convention of PN junction diode current equation
  - With bias voltage applied, the total current is the sum of **bias-induced** PN junction current and the **photo-induced** optical current:

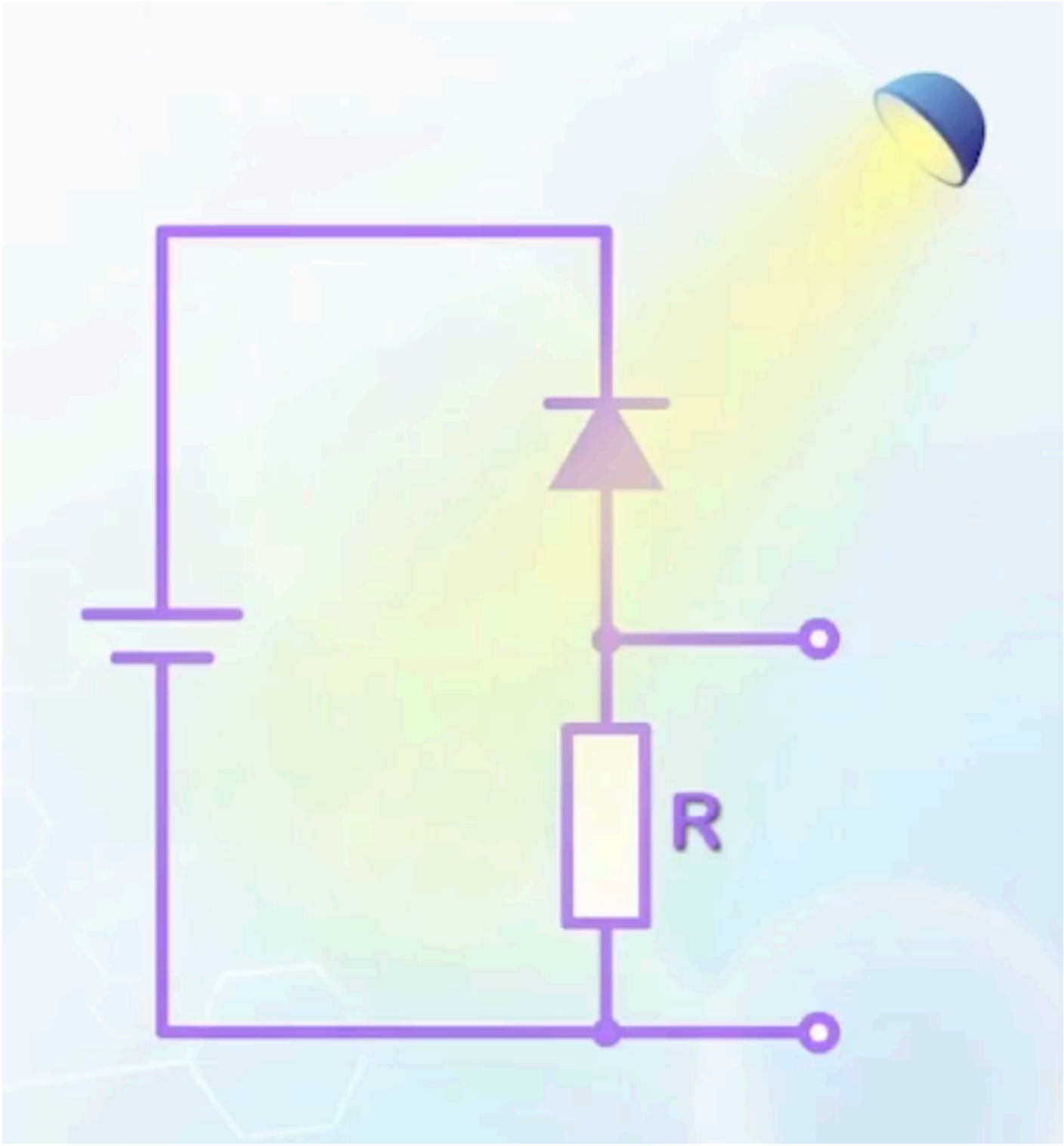
$$I = I_{D0}\left(e^{\frac{qV_A}{kT}} - 1\right) - I_{\text{photo}}$$

  - In the I-V curve:
    - At higher light density, the curve shifts downwards

- The curve cuts the V axis at open-circuit voltage V_{OC}
  - Happens when the reverse I_{photo} is balanced out by the forward diode current at V_{OC}

Photo Detectors

- A simple photo detector circuit



  - A reverse bias is applied, and the diode current is equal to I_{photo} because the reverse current is very small

- Without light, the output voltage is **0** because the diode behaves like an open circuit
- With light, the output voltage is:

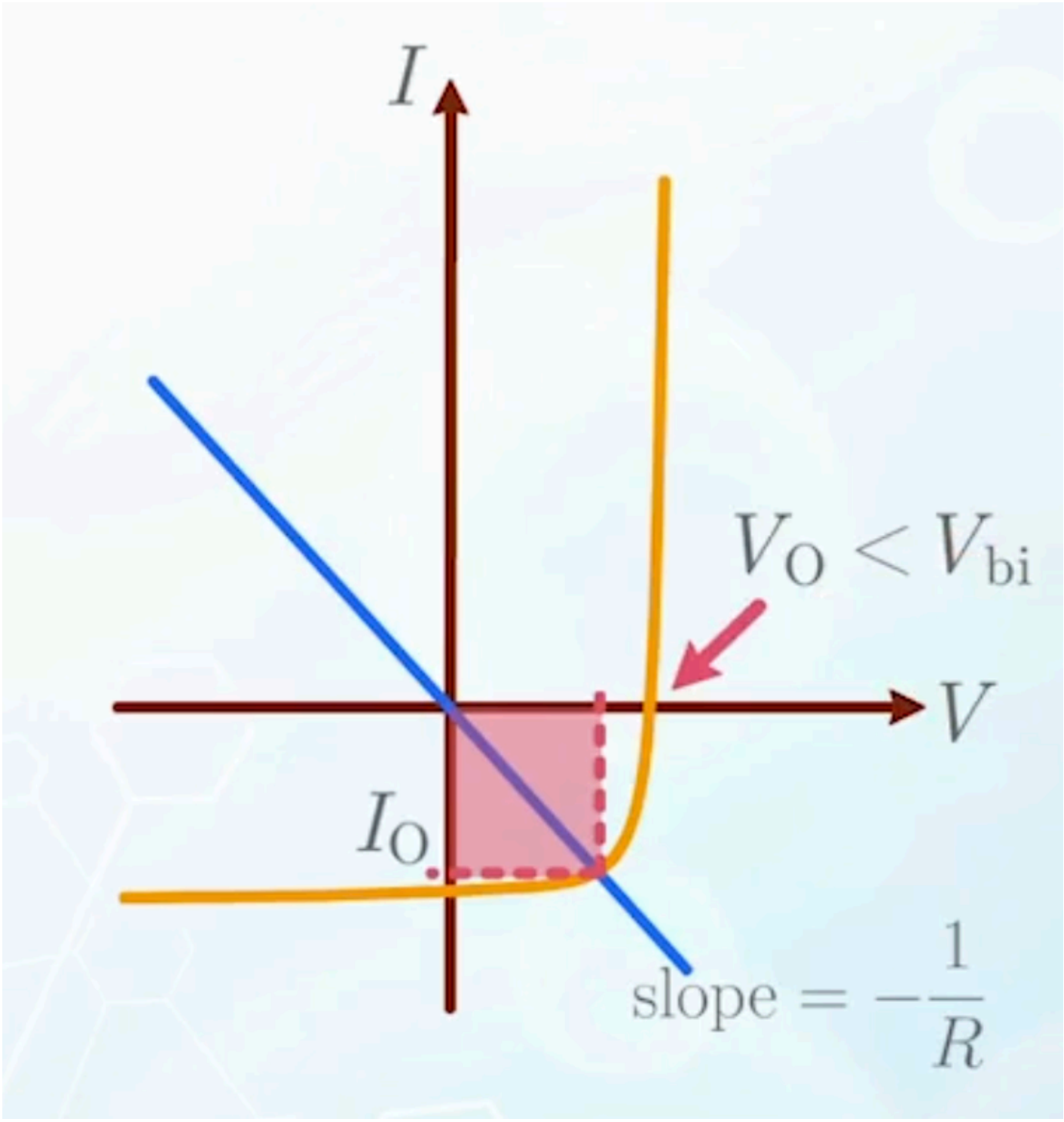
$$V_{\text{out}} = I_{\text{photo}} R$$

This means the output voltage is proportional to I_{photo}, thus proportional to the light intensity

Solar Cells

- In a photo detector circuit, the PN junction is forced to work in reverse bias mode, or the 3rd quadrant of the I-V curve, where power is consumed (like a resistor)
  - The current flows from P side to N side, **following** the voltage polarity
- In a solar cell, a current is generated to flow from the N side terminal to the P side terminal
  - The current is fed to the resistor and generates a forward bias within the diode
  - The diode now works in the **4th quadrant** of the I-V characteristic, where the direction of current flow is **opposite** to the voltage polarity
  - Converts optical energy to electrical energy, is a solar cell
  - The operation point of the circuit is determined by the I-V curve of the diode, and the resistor load

- Define P -> N to be positive:



  - (V_O, I_O) operation point
  - Shaded area: power
  - Maximum voltage from a solar cell is limited by V_{bi}
    - For silicon based solar cell, usually $< 1\,V$, as V_{bi} has to be smaller than bandgap energy

Why?

- Fermi level on both side has to be within the bandgap
- Fermi level must align at equilibrium
- It is only possible when $qV_{\text{bi}} < E_G$

- For higher voltage output, multiple cells are connected in series
- The current, voltage, and power output provided by a solar cell can vary significantly with the illumination intensity and the load resistance, thus a **voltage regulator** is usually required to stablize the output voltage

Solar Cell Materials

- Solar cells fabricated using single crystal silicon is still quite expensive
  - As single crystal silicon is produced from molten silicon at very high temperature, which is energy intensive and polluting
- Light can only penetrate to a small depth from the top, on a PN junction on a wafer surface
  - A large portion of substrate material below the penetration depth is only used as mechanical support
- In second generation solar cells, thin film technology is used to form a thin solar cell on a low-cost substrate, usually glass
  - Silicon on the supporting substrate is formed by depositing the material using silicon atom carrying gas, such as Silane ($\mathbf{SiH_4}$), which can be performed at a much lower temperature (400 – 600 °C instead of ~1400 °C)
  - The deposited film will have either polycrystalline or amorphous structure, due to the mismatch of atomic spacing between the supporting substrate and the thin-film material
    - Depositing silicon at low temperature usually leads to amorphous structure, and the grain size increases with temperature
    - At sufficiently high temperature, polycrystalline structure can be formed on the supporting substrate
    - Grain boundaries exist regardless of polycrystalline or amorphous structure, which cause scattering when carriers move from one grain to another, leading to worse performance than single crystal silicon solar cells
  - The thin-film material will also have defects and stability issues
  - But the cost is much lower and can cover a much larger area

- Third generation solar cells
    - Using organic materials
    - Low temperature production
    - Short turnaround time
    - Easy disposal
    - Low efficiency
    - Low reliability inherent to organic materials

Solar Cell Design

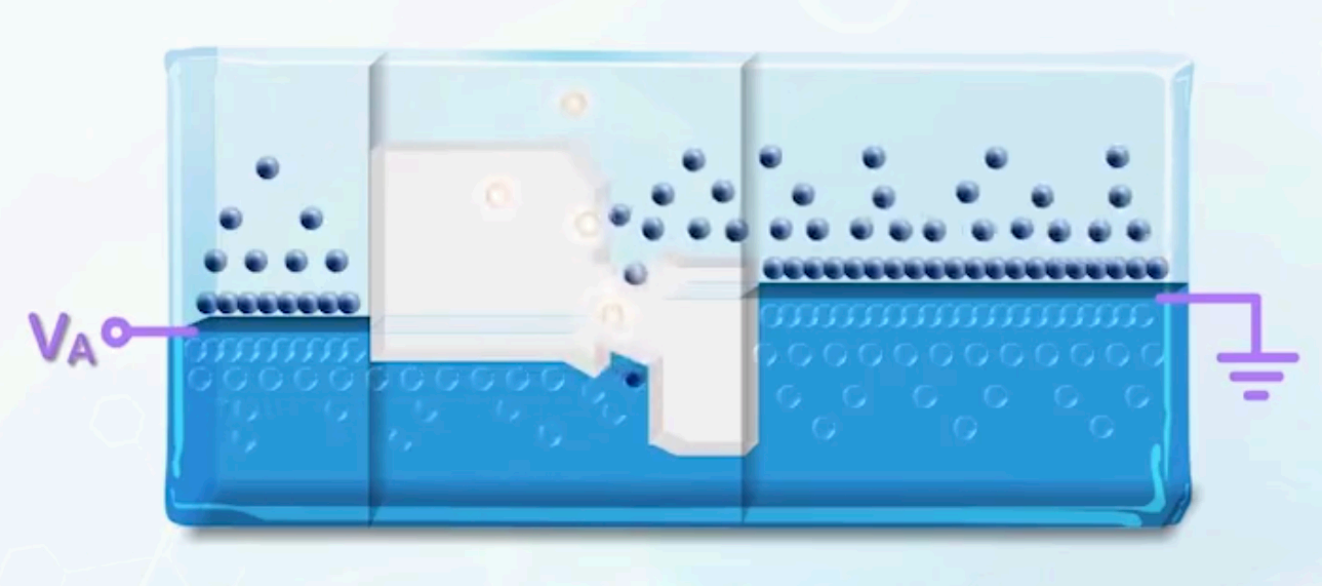
Although solar cells are basically a simple PN junction, some design optimizations are needed to achieve high energy conversion efficiency.

- When light shines on the solar cell, part of the light is absorbed, while part is reflected
    - To increase absorption and reduce reflection, an **anti-reflection coating** is usually applied to the surface
        - This is why solar cells look dull from the light absorbing surface
- To allow maximum amount of light to enter the depletion region, the **top layer of PN junction has to be thin**, so that the depletion region is closer to the top of the wafer
- The depletion region has to extend **deep enough** to absorb all the light entering the diode
    - This can be achieved by using **lightly doping** on one side of the junction
    - But it also decreases V_{bi}, limiting the maximum possible voltage
    - To achieve both high V_{bi} and deep depletion region, an **intrinsic/lightly doped layer** can be inserted between two **heavily doped P+ and N+** regions, forming a **PIN diode**
        - In this structure, the intrinsic layer will be totally depleted, and the thickness of the intrinsic layer can be used to adjust the depletion width
        - On the other hand, V_{bi} is determined by the two heavily doped side, and can be adjusted to be close to the bandgap voltage
- Contacts have to be added to the top and bottom layers **without blocking the light**
    - So they are usually arranged into strips to minimize the distance from any location on the top layer of the surface of the junction to the metal line
    - Another approach is to use transparent electrodes to cover the entire surface

Light Emitting Diodes (LEDs)

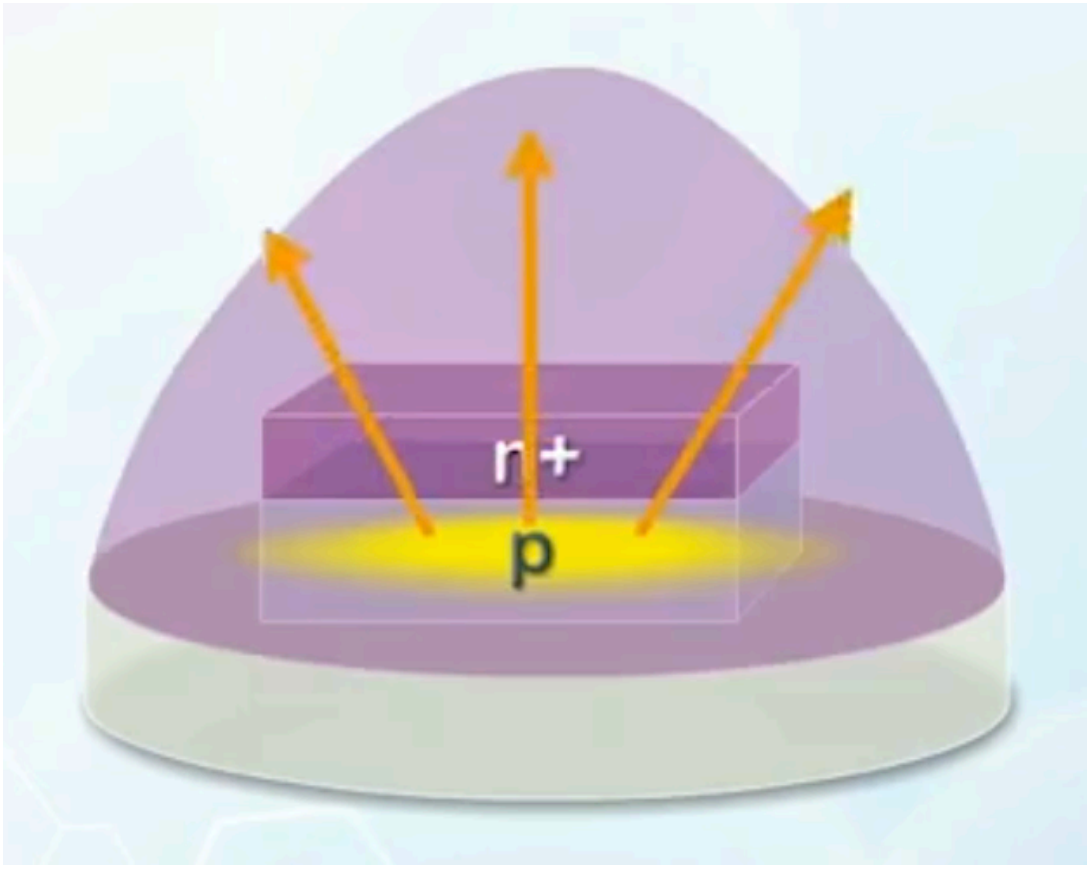
- Light is absorbed during electron-hole generation

- The reverse combination process can emit light energy
  - A semiconductor can be used as a light source when substantial recombination occurs
  - For example, light is emitted in a long PN junction under forward bias because of recombination
- We usually cannot see the emitted light
  - Recombination takes place deep inside the silicon crystal
  - The energy of emitted photons is, in general, governed by E_G
    - Silicon has a bandgap of $\mathbf{1.1\,eV}$, corresponding to infrared light, which is invisible to human eyes
    - We need materials with larger bandgap to emit visible light
    - Organic materials can also be used to form OLEDs
  - Due to the use of large bandgap materials, LEDs usually have a larger built-in potential, thus resulting in a larger turn on voltage V_{ON}
    - Typically, V_{ON} is about $\mathbf{3\,V}$ for a blue LED
- Designing LED structure to achieve high efficiency
  - We want to have all electrons and holes recombine **before reaching the end of the diode** -> **a very long diode**
  - However, a long diode will push the region of light generation deep inside the solid, making it difficult for light to escape to the surface
  - To force more recombination over a shorter neutral region, a **heterogeneous** structure using materials of different bandgaps is often used

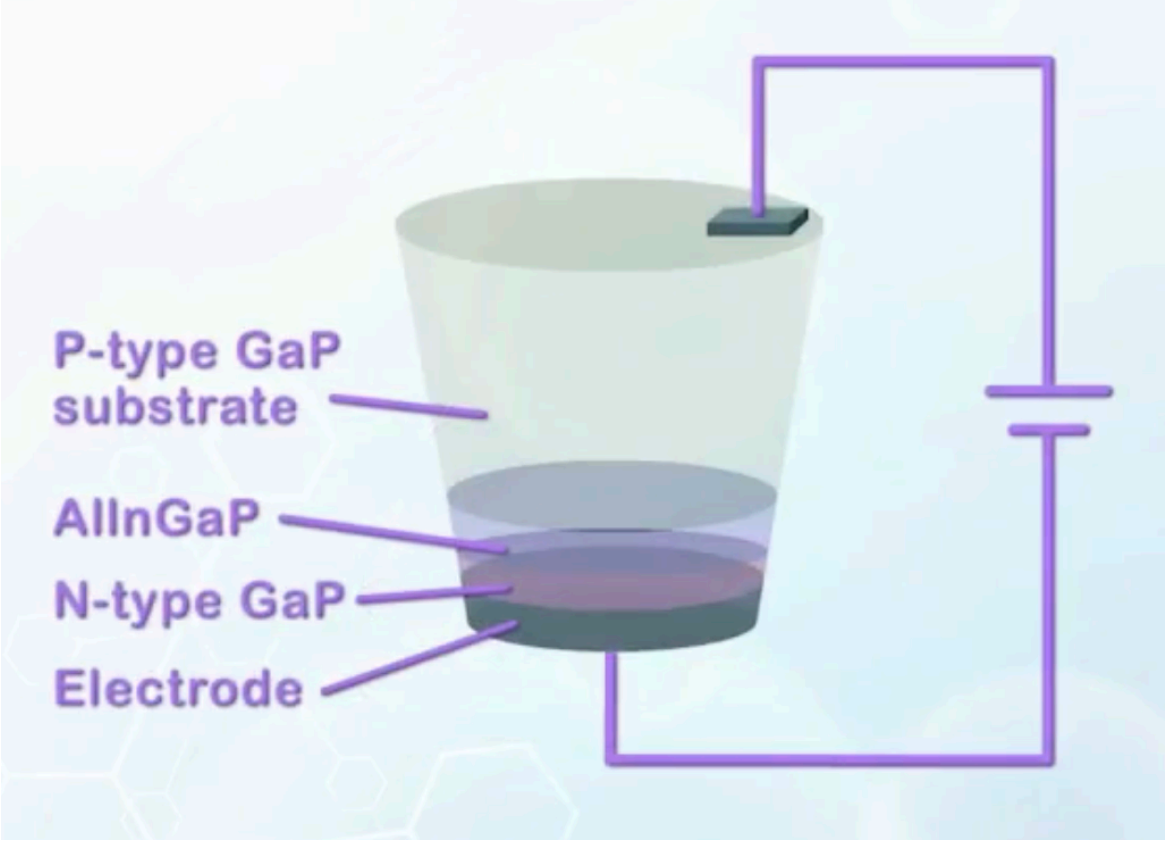


    - Electrons and holes are injected and trapped by the higher bandgap material, leading to more recombination in a shorter distance

        - All currents are recombination current to ensure efficient conversion from electrical energy to optical emission
- The shape of LEDs also need to be carefully designed
    - In most LED applications, light emits only in a particular direction
        - Light generated in the LED radiates in all directions
        - The **bottom surface should be reflective** to increase the light intensity in a particular direction of interest
        - At the same time, **reflection from the top surface should be avoided**
        - When light reaches the top surface, total reflection may occur if the incident angle is larger than the critical angle, which will reduce the intensity of emitted light
        - To avoid total internal reflection
            - The semiconductor material may be structured into a **dome shape**, so that most radiating light will be emitted in a direction perpendicular to the top surface
            - Or use some capping material with a similar reflective index to that of the light emitting material, to shape the upper surface



        - In case of LED with small light-emitting area, it can be structured like a **trapezoid**, with slope sides to confine the direction of emitted light



10. Metal Semiconductor Contacts

About work function, metal-semiconductor contacts (Schottky contacts and ohmic contacts), derivation of Schottky diode I-V characteristics, and comparison between Schottky diodes and PN junction diodes.

Work Function of a Material

- When metal and semiconductor contacts, how to draw the band diagram for such a heterogeneous system, as the metal and semiconductor have very different band structures?
  - Bandgap of **semiconductor**: Valence band, bandgap, conduction band
  - Bandgap of **metal**: Partially filled band, electrons fill the states up to **Fermi level**
    - Normally represent the band diagram only by its Fermi level
    - States above Fermi level are mostly empty, states below Fermi level are mostly filled
  - Electron energy is related to the electronic potential which is relative, a **common reference** is needed to align the two systems in contact
    - Electrons moving in vacuum (free space) without any influence of any charged body are considered to have the same electronic potential, and defined as **zero** potential energy

- When they are brought **closer** to a charged body (e.g., the nucleus of an atom), they lose electronic potential, and are considered to have **negative** potential energy
- For metal, the Fermi level is the highest electronic potential energy with a high chance of finding an electron in a state
- **Work-function $q\phi_M$**: The required energy to bring an electron from the Fermi level of the metal to set it free
  - ϕ_M: the corresponding voltage of the metal work-function
  - A well-defined material property for a **metal**
- **Vacuum (Energy) Level**: The energy level at which electrons are free from the attraction of the atom
- For **semiconductor**, the work-function is not a fixed value, but depends on doping
  - **Electron Affinity χ**: The energy difference between the vacuum level and the conduction band edge
  - For silicon

$$\chi_{Si} \approx 4.05\,\text{eV}$$

  - Metal may be considered as a semiconductor with zero bandgap, thus the work-function or electron affinity carries similar meanings

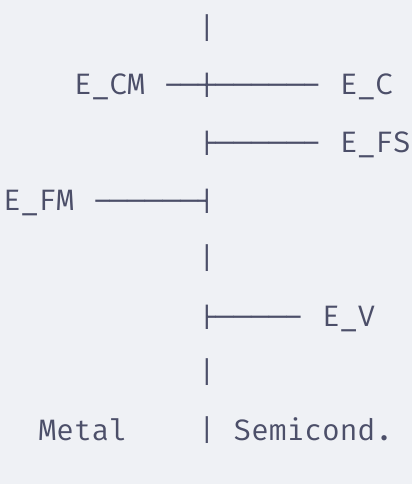
$$q\phi_M = \chi$$

- When two materials are put together, the vacuum level is taken is the common reference to align the two materials
  - Electrons at vacuum level can move freely between the two materials without a sudden change in energy
  - At the interface, the two materials are fused together, with an abrupt change in the band structure
    - Most electrons at the metal are located below the Fermi level, they must gain enough energy to overcome the barrier at the interface before moving to the conduction band of the semiconductor
    - The barrier height is an important parameter to determine the properties of the metal-semiconductor contact
  - **Schottky barrier $q\phi_B$**: The energy barrier formed at the metal-semiconductor interface
  - **Schottky diode**: A two terminal device that consists of a metal and a semiconductor, forming a Schottky barrier

- An electron from the metal can enter the valence band of the semiconductor only if there are spaces (holes) available, equivalent to have a hole moving from the semiconductor to the metal
- Holes moving from metal to the semiconductor must also overcome the barrier between the metal Fermi level and the semiconductor valence band edge

Formation of a Schottky Contact

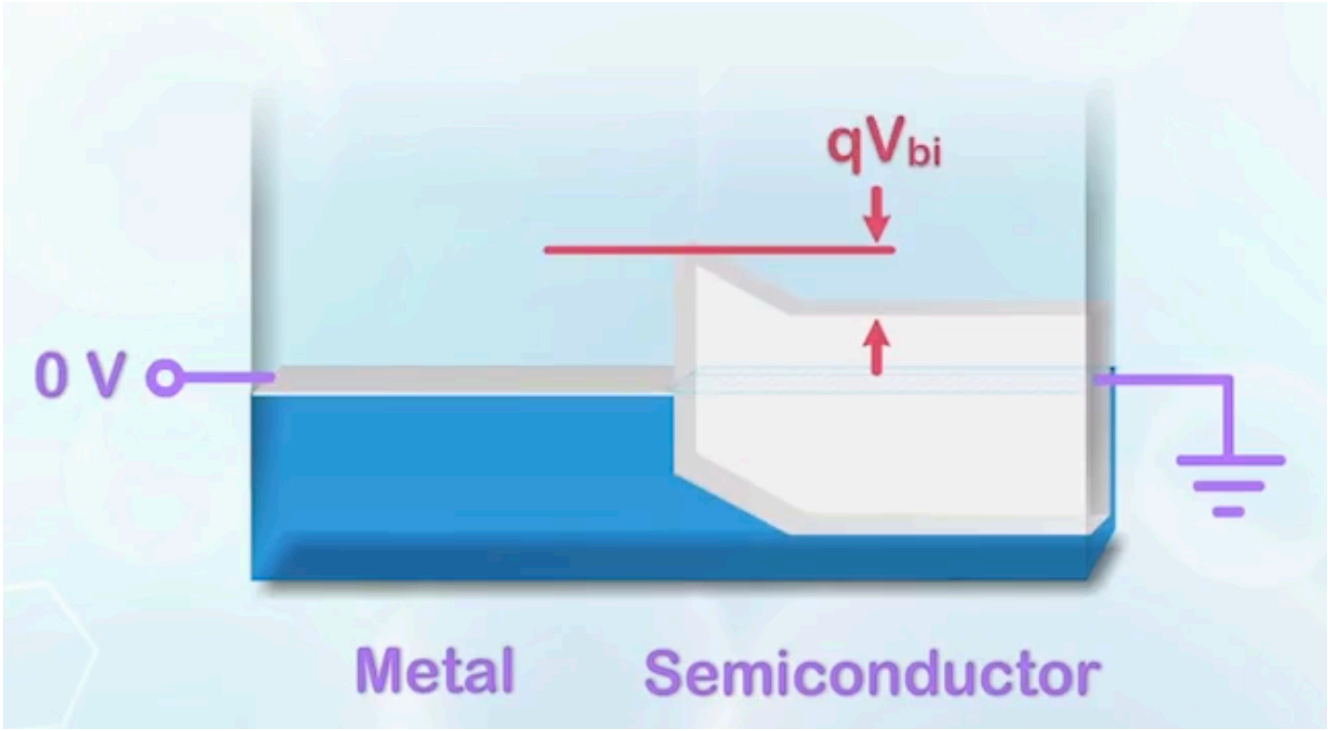
Case study - Assume the metal Fermi level is close to the middle of the N type silicon bandgap



- E_{CM}: The energy level at the conduction band edge of the semiconductor at the **metal interface**

- Under this condition, $E_{FS} > E_{FM}$, the semiconductor has a higher electron concentration at energy above E_{CM}
  - Electrons will flow from the semiconductor to the metal
  - Metal will absorb any amount of electrons from the semiconductor without changing its Fermi level
  - Moved electrons from the semiconductor leave behind positively charged ions, a **depletion region** is formed
  - Charge in the depletion region will attract carriers in the metal with opposite charge to accumulate at the surface to achieve charge neutrality
  - Once depletion region is formed, it is more difficult for electrons to move from semiconductor to metal
  - Finally, **thermal equilibrium** is achieved with a fixed depletion width

- At thermal equilibrium, the system is subject to external potential, like connecting a battery with $\mathbf{0\,V}$ at the two ends of the Schottky diode
  - The Fermi level must align with external voltage
  - qV_{bi} is formed to balance out the tendency of electron flow from the semiconductor to the metal
  - Similar to a P+/N junction, except that the P+ region is formed by a material with almost zero bandgap and a very large number of carriers



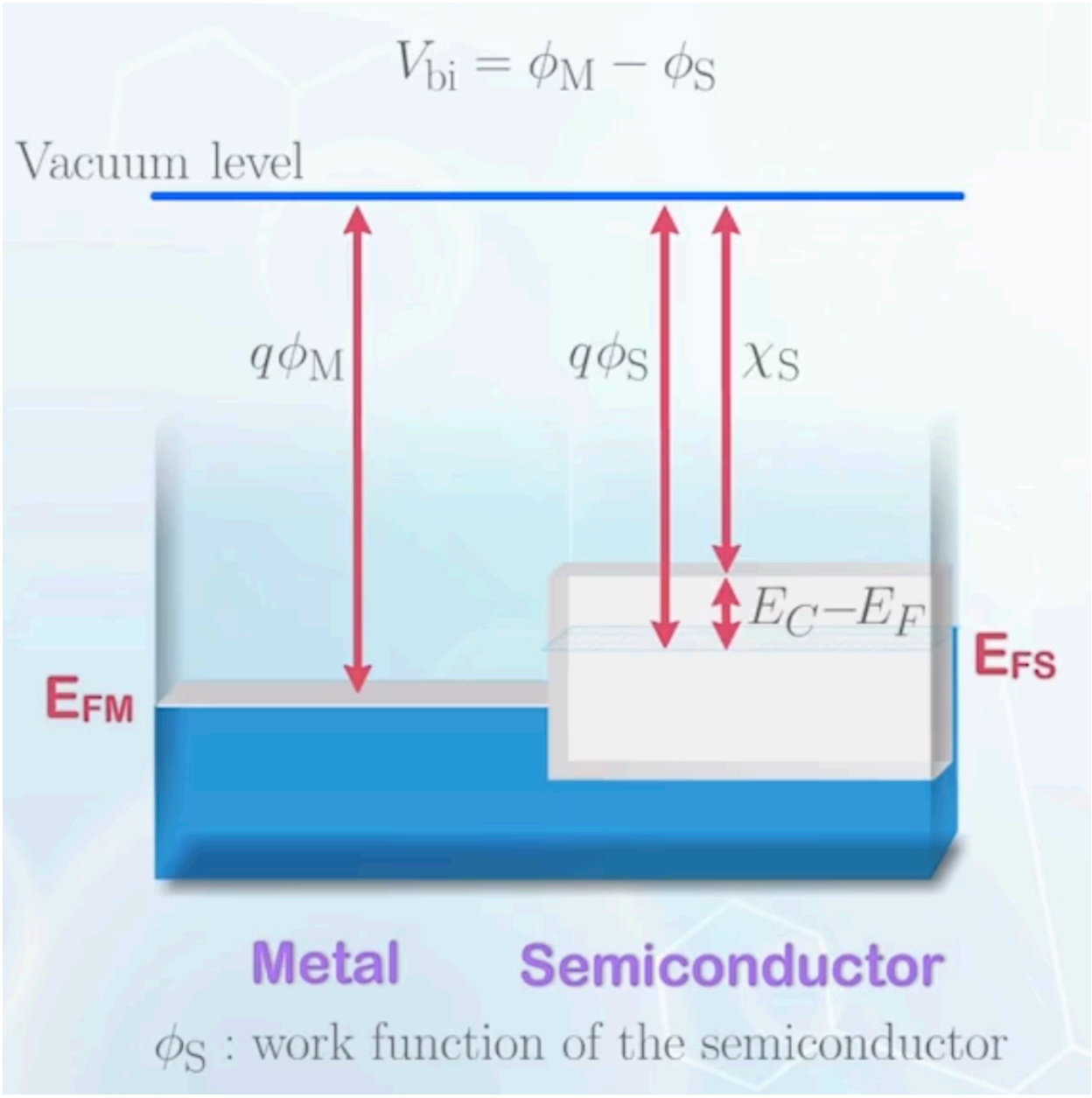
Band Diagram of a Schottky Contact at Thermal Equilibrium

- To find out the **built-in potential V_{bi}** and the **depletion region width**
  - Built-in potential is equal to the difference in Fermi levels

$$qV_{\text{bi}} = E_{FS} - E_{FM}$$

    - When locating the Fermi levels, we cannot make reference to the conduction band or the valence band, unlike a PN junction

- Taking the **vacuum level** as the common reference



$$\begin{aligned} qV_{bi} &= q\phi_M - q\phi_S \\ &= q\phi_M - (\chi_S + (E_C - E_F)) \\ &= q\phi_M - (\chi_S + \frac{E_G}{2} - (E_F - E_i)) \end{aligned}$$

For silicon

$$\phi_S = 4.08\,\text{V} + 0.55\,\text{V} - \frac{kT}{q}\ln\frac{N_D}{n_i}$$

Thus

$$V_{bi} = \phi_M - 4.63\,\text{V} + \frac{kT}{q}\ln\frac{N_D}{n_i}$$

- To calculate the depletion region width
    - Similar to a PN junction, but on the metal side, the charge is given by a δ function, representing crowding of carriers at the surface
    - Or, the metal side can be treated as a super doped P type silicon, with infinite carrier supply and infinitely small depletion region width
    - Using the same method as PN junction, we have

$$x_n = \sqrt{\frac{2\varepsilon_{\text{Si}} V_{\text{bi}}}{qN_D}}$$

      It is the same as PN junction, with the P side depletion width assumed to be zero

Carrier Motion of a Schottky Contact at Thermal Equilibrium

- On the metal side, the current conduction takes place in **one single energy band**, and we count either electrons **or** holes, and resistance for both is very small
- On the semiconductor side, current conduction takes place in **both conduction band and valence band**
    - Assume the semiconductor is N type
    - Hole concentration in the valence band is very small, therefore the hole resistance is very large, unfavorable for hole current conduction
    - Electron concentration in the conduction band is relatively large, except for a small region near the junction
    - The current conduction is mainly through electrons (or **majority** carriers of the semiconductor)
        - Metal-semiconductor contact is usually referred to as a **majority carrier device/unipolar device**, as current conduction is only contributed by either electrons **or** holes, but not both
- By replacing the P side of a PN junction with a metal, it removes the high resistance region in the conduction band for electrons
    - Once the carriers gain enough energy to overcome the potential barrier, they can be easily on the metal side
    - Encourages smoother current flow, but makes calculation more difficult
        - In a PN junction, we rely on the diffusion in the high resistance region that limits the carrier motion to count the carriers

        - Without the high resistance region, calculation will become more complex as there are lots of carriers moving very fast at all locations in the conduction band
        - We cannot find a very obvious mechanism that controls the current flow
        - We have to calculate the current with some **probabilistic** methods
- To calculate the current
    - Assume the electrons are moving at some thermal velocity v_{th}, and we only consider the lateral component of the velocity $v_{\text{th}x}$
    - Assume the depletion region width is small and can be ignored, or the neutral region is directly connected to the metal
    - Electrons above energy level E_{CM} are free to move in any directions
        - Consider electrons on the metal side near the interface with energy above E_{CM}
        - Assume half of them are moving towards the semiconductor, and half of them are moving away from the semiconductor
            - n_{M}: the number of electrons above E_{CM} on the metal side
            - **Current flow from the metal side to the semiconductor side**

$$I_{\text{M}} = -qA(\frac{n_{\text{M}}}{2})v_{\text{th}x}$$

        - Similarly, for electrons on the semiconductor side near the interface with energy above E_{CM}
            - n_{S}: the number of electrons above E_{CM} on the semiconductor side
            - **Current flow from the semiconductor side to the metal side**

$$I_{\text{S}} = qA(\frac{n_{\text{S}}}{2})v_{\text{th}x}$$

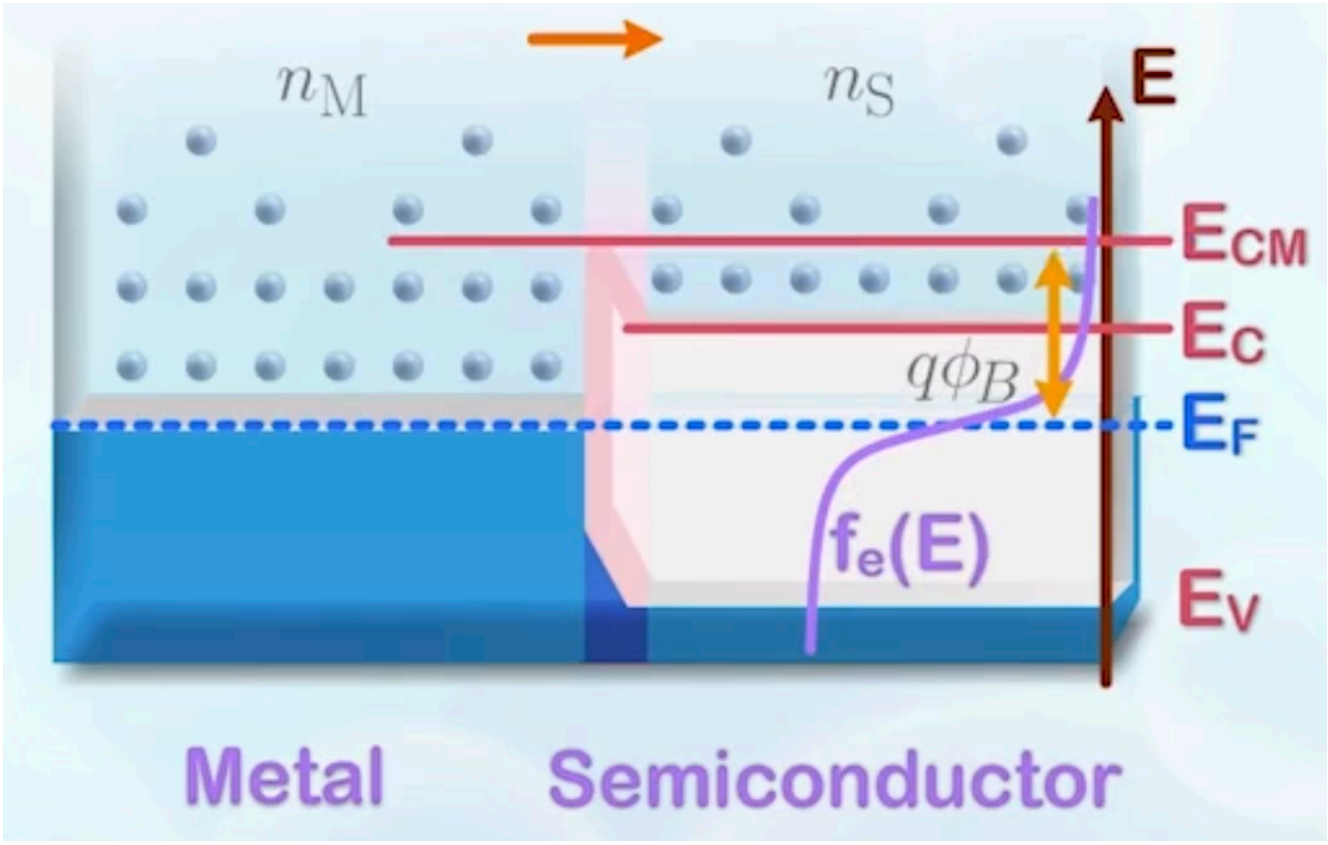
        - The total current

$$\begin{aligned} I_{\text{MS}} &= I_{\text{M}} + I_{\text{S}} \\ &= qA\frac{n_{\text{S}} - n_{\text{M}}}{2}v_{\text{th}x} \end{aligned}$$

    - At thermal equilibrium

$$\begin{aligned} I_{\text{MS}} &= 0 \\ n_{\text{S0}} &= n_{\text{M0}} \\ n_{\text{S0}} &= N_C e^{-\frac{q\phi_B}{kT}} \\ &= n_{\text{M0}} \end{aligned}$$

where ϕ_B is the Schottky barrier height



Schottky Diode Current-Voltage Characteristics

- Under reverse bias
  - On the metal side, the Fermi level is raised, but n_M remains unchanged, because the barrier height ϕ_B is not affected by the bias
  - On the semiconductor side, n_S decreases as E_{CM} is moved further away from the Fermi level

$$n_S = n_{S0} e^{\frac{qV_A}{kT}} < n_{S0} \quad (V_A < 0)$$

  - The total current

$$\begin{aligned} I_{MS} &= I_M + I_S \\ &= \frac{1}{2} q A v_{thx} (n_S - n_M) \\ &= \frac{1}{2} q A v_{thx} n_{S0} \left(e^{\frac{qV_A}{kT}} - 1 \right) \\ &= \frac{1}{2} q A v_{thx} N_C e^{-\frac{q\phi_B}{kT}} \left(e^{\frac{qV_A}{kT}} - 1 \right) \\ &= K T^2 e^{-\frac{q\phi_B}{kT}} \left(e^{\frac{qV_A}{kT}} - 1 \right) \end{aligned}$$

where

$$KT^2 = \frac{1}{2} q A v_{\text{th}x} N_C$$

- The temperature dependence comes from both N_C and $v_{\text{th}x}$
- K is the Richardson constant. For silicon

$$K \approx 120\,\text{A} \cdot \text{cm}^{-2} \cdot \text{K}^{-2}$$

- Grouping all the bias independent terms, we have

$$I_{\text{MS}} = I_0 \left(e^{\frac{qV_A}{kT}} - 1 \right)$$

which is similar to the PN junction diode equation, except that I_0 is not dependent on the doping concentration, but the Schottky barrier height

$$\begin{aligned} I_0 &= KT^2 e^{-\frac{q\phi_B}{kT}} \\ &= \frac{1}{2} q A v_{\text{th}x} N_C e^{-\frac{q\phi_B}{kT}} \end{aligned}$$

- When the reverse bias further increases, the Schottky diode may enter the **breakdown region**
  - It is easier for electrons to tunnel through the bandgap in the depletion region as the Schottky diode only consists of half of the depletion region of a PN junction
  - The breakdown voltage of a Schottky diode is usually **lower** than that of a PN junction diode, and the breakdown mechanism is usually **Zener breakdown**

- Under forward bias, the analysis is similar
  - The built-in potential is reduced, but the Schottky barrier height remains unchanged
  - E_{CM} is moved closer to the Fermi level on the semiconductor side, thus increasing n_{S}
  - The current

$$\begin{aligned} n_{\text{M}} &= n_{\text{M0}} = n_{\text{S0}} \\ n_{\text{S}} &= n_{\text{S0}} e^{\frac{qV_A}{kT}} > n_{\text{S0}} \quad (V_A > 0) \\ I_{\text{MS}} &= I_{\text{M}} + I_{\text{S}} \\ &= \frac{1}{2} q A v_{\text{th}x} (n_{\text{S}} - n_{\text{M}}) \\ &= \frac{1}{2} q A v_{\text{th}x} n_{\text{S0}} \left(e^{\frac{qV_A}{kT}} - 1 \right) \\ &= \frac{1}{2} q A v_{\text{th}x} N_C e^{-\frac{q\phi_B}{kT}} \left(e^{\frac{qV_A}{kT}} - 1 \right) \\ &= KT^2 e^{-\frac{q\phi_B}{kT}} \left(e^{\frac{qV_A}{kT}} - 1 \right) \\ &= I_0 \left(e^{\frac{qV_A}{kT}} - 1 \right) \end{aligned}$$

is exactly the same as the reverse bias case, and very similar to the PN junction diode equation, except for the expression of pre-exponential term I_0

Schottky v.s. PN Junction Diode

There are some significant differences in performance between Schottky diodes and PN junction diodes

- In a PN junction, the turn-on and turn-off requires the increase or decrease of carriers in the neutral region, represented by the diffusion capacitance C_{diff}
- In a Schottky diode, the electrons moving from the semiconductor to the metal are simply absorbed, without the need of storage of carriers
  - C_{diff} is eliminated
  - **Very fast turn-on/turn-off**
- The built-in potential of Schottky diodes is usually lower than that of PN junction diodes, as it only consists of half of the PN junction depletion region
- Metal can absorb any amount of electrons
  - **No high-level injection** is observed in a Schottky diode
- Schottky diodes in general has lower series resistance, as metal is used as one of the terminals
- The common turn-on voltage for silicon Schottky diodes is around $\mathbf{0.3\,V}$, **lower** than the $\mathbf{0.7\,V}$ of silicon PN junction diodes
  - More suitable for **very high-speed circuits**
  - Also used when **a lower turn-on voltage is desired**, such as the clamping device in a bipolar junction transistor
- **However**, the low breakdown voltage makes Schottky diodes unsuitable for rectifier circuits to block high voltage input sources

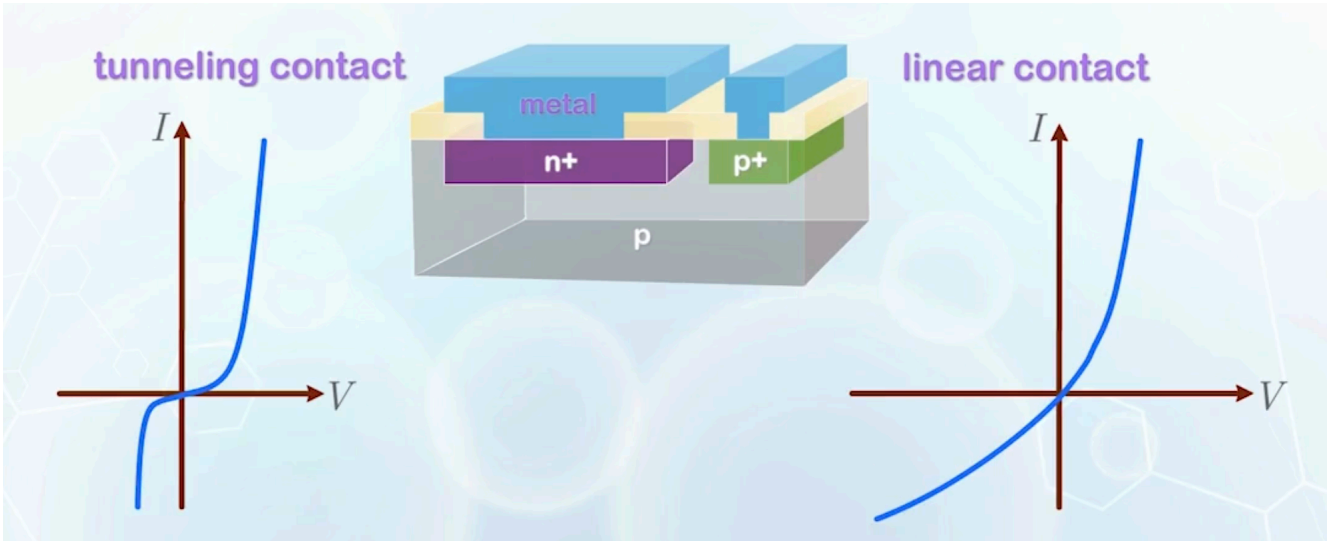
Ohmic Metal Semiconductor Contacts

Schottky contacts only allow current flow in one direction, which may be a problem when using metal to connect different devices

- **Ohmic contact**: A metal-semiconductor contact that allows current to flow in both directions
- To allow bi-directional current flow, we want to increase the reverse current, making the characteristics more symmetrical about the origin of zero applied voltage
- Two different methods

1. **Linear ohmic contact**: Allows electrons to move freely across the barrier and significantly increase the reverse current
    - Metal with Fermi level close to the **conduction band** edge of an **N type** semiconductor
    - Metal with Fermi level close to the **valence band** edge of a **P type** semiconductor
    - **Limitations**: Using a single metal, we can only make ohmic contact to either N type or P type semiconductor, but not both, as the metal having a small barrier to the conduction band will have a large barrier to the valence band, and vice versa
2. **Tunneling contact**: Lowering the breakdown voltage to a very small value, by doping the semiconductor very heavily. At the same time, the metal Fermi level should be very close to the **valence band edge of an N type** semiconductor, or the **conduction band edge of a P type** semiconductor
    - Causes a very large built-in potential
    - The depletion region width is extremely small due to heavy doping
    - Electrons can easily tunnel through the very thin barrier even under a small reverse bias
    - **Less linear** than the linear contact

- With both linear and tunneling contacts, **a single metal** can be used to form ohmic contact to **both N type and P type** semiconductors
    - Using metal with work-function close to the valence band edge of the semiconductor
        - Forms linear ohmic contact with a P type semiconductor
        - Forms tunneling contact with a heavily doped N type semiconductor
    - To ensure ohmic contact, the semiconductor to be placed in contact with the metal is usually doped heavily, for both tunneling and reduction of resistance



The internal built-in potential of a PN junction will be balanced out by the band bending at the metal-semiconductor contacts, thus the built-in potential cannot be measured.

11. Basic Bipolar Junction Transistor in Forward Active Mode

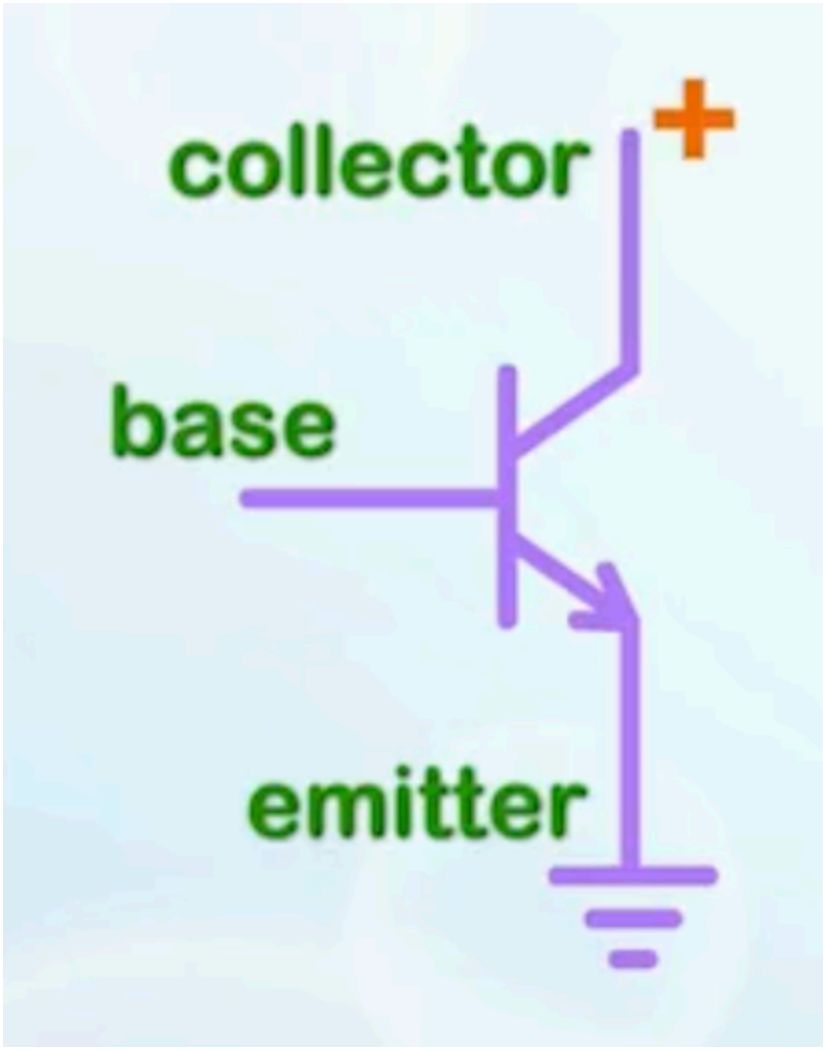
About the formation of bipolar junction transistors (BJT), their operation principles, and discussions on current components, current-voltage characteristics, and design considerations in the forward active mode.

Formation of the Bipolar Junction Transistor (BJT)

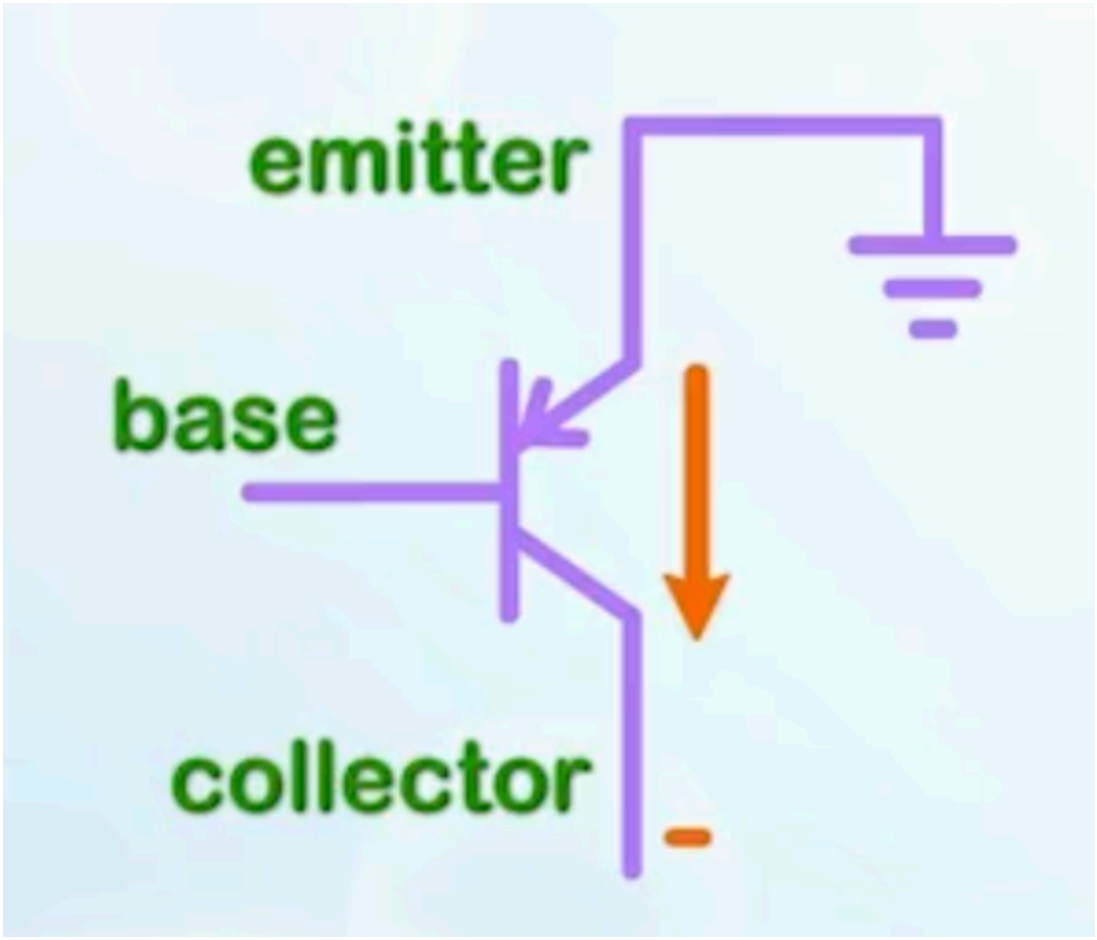
The most important elements in modern circuits are switches and amplifiers, both of which can be constructed by extending the fundamental PN junction structure.

- Adding another N region on the other side of a P region, we have two N regions separated by a P region, with a lot of free moving electrons separated by the P region
  - When the barrier is lowered by applying **positive voltage** to the P region, electrons able to move across the barrier increases exponentially
  - Forms a three terminal device with electron flow between the two outer terminals controlled by the middle terminal
  - This is a **bipolar junction transistor (BJT)**, specifically an NPN transistor
  - **Base(B)**: the middle P region
  - Different voltages are usually applied to the two N regions
    - **Emitter(E)**: the N region from where the carriers enter the base
    - **Collector(C)**: the N region receiving the carriers from the base

- The emitter differs from the collector by **having a much heavier doping**

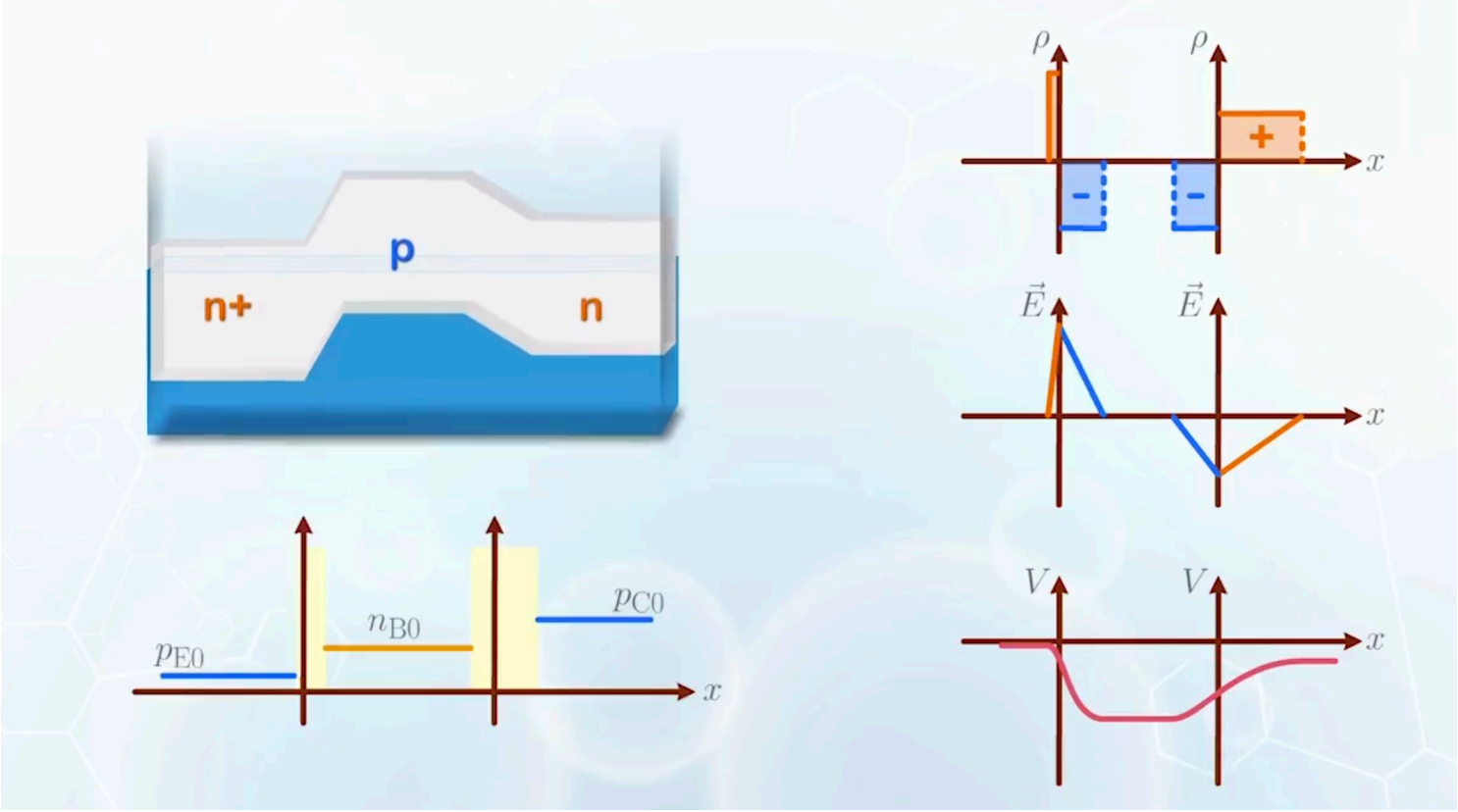


- Similarly, a PNP transistor can be formed by adding a N region between two P regions



- In both symbols, the arrow indicates the direction of the current flow between the **base** and the **emitter**
- **From now on...**

  We will be using NPN transistors for the following analysis, unless otherwise specified.



The emitter is usually the most heavily doped region, while the collector is the most lightly doped region.

Basic Operation Principles in the Forward Active Mode

Different voltages are applied to the three terminals to operate a BJT. The voltages can forward bias or reverse bias the PN junctions. Four possible operation modes are possible, as there are two PN junctions.

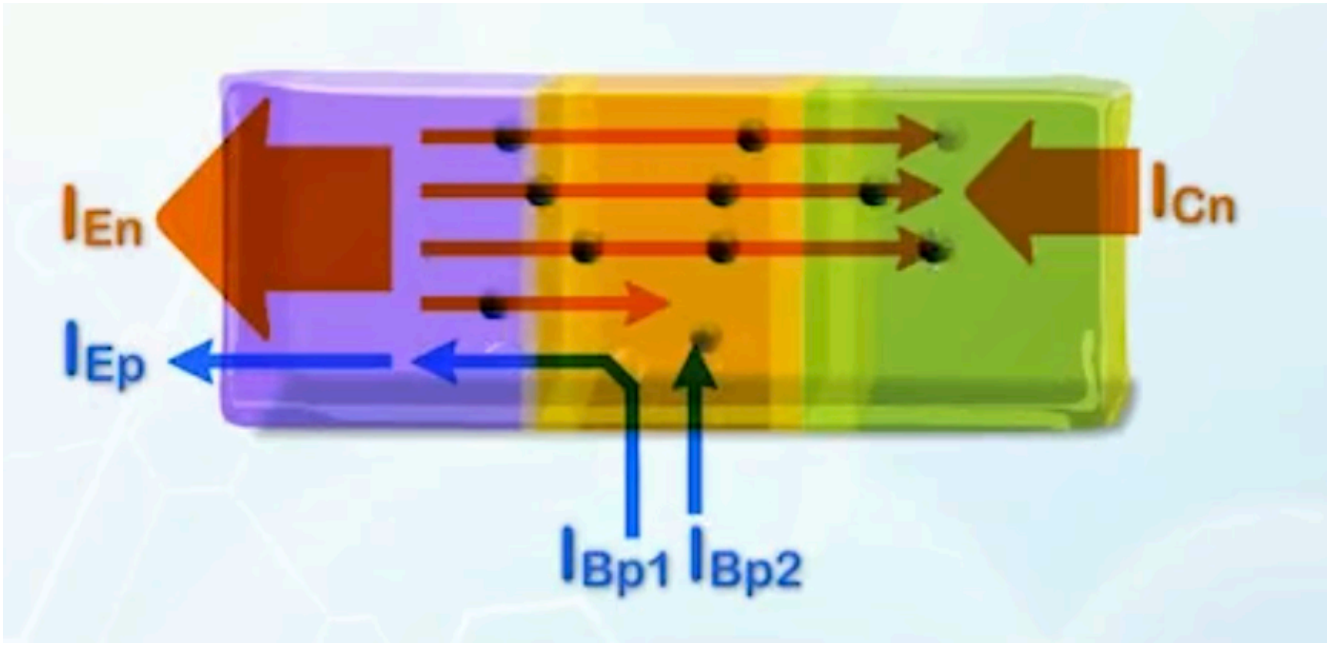
Base-Emitter Junction	Base-Collector Junction	Operation Mode
Forward Biased	Reverse Biased	Forward Active
Reverse Biased	Forward Biased	Reverse Active
Forward Biased	Forward Biased	Saturation
Reverse Biased	Reverse Biased	Cutoff

The case with the base-emitter junction forward biased, and the base-collector junction reverse biased, is the most important operation mode, called the **forward active operation region**. It is the most common way to connect a BJT in a circuit.

- In forward active mode
  - Apply a positive voltage to the base and collector, taking the emitter as the reference ground
  - Apply a more positive voltage to the collector to reverse bias the base-collector junction
  - The amount the electrons that can flow from the emitter to the collector is controlled by the base-emitter voltage as described earlier
  - The minority carrier concentrations
    - The base-emitter junction is forward biased, so the number of carriers on the two sides of the base-emitter depletion region increases
    - The base-collector junction is reverse biased, so the number of carriers on the two sides of the base-collector depletion region decreases
    - In a properly designed BJT, the base width is usually **very small**, we have to use the **short-diode** carrier distribution in the neutral region of the base, drawing a straight line
  - Ideally, when a positive voltage is applied to forward bias the base-emitter junction, all the electron current from the emitter should flow to the collector without leaking to the base
    - This is not possible, as the base contact forms part of the PN junction terminal, and some current always flows to the base when a forward bias voltage is applied to it
    - Both base current and the collector current comes from the emitter, we want to **maximize** the collector current and **minimize** the base current
      - Based on the geometry, it can be achieved by making the base contact small, making the collector contact large
      - By putting the base contact sideways, the momentum of the electrons also favor electron motion to the collector region
      - By isolating the **collector** voltage, so that it will not affect the voltage at the base
        - The reverse biased collector to base contact serves this purpose, as any voltage applied to the collector will be dropped across the base to collector depletion region without changing the base voltage

        - The collector will passively receive electrons reaching it, and sweep them away with the electric field in the base to collector depletion region
    - In terms of current, the BJT can also be used as an amplifier as a small base current can lead to a large collector current

Current Components in the Forward Active Mode

- The entire operation of the BJT is governed by the base-emitter junction
    - At the emitter
        - I_{Ep}: the hole current due to diffusion at the emitter, where $_{E}$ for emitter, and $_{p}$ for holes
        - I_{En}: the majority carrier electron current at the emitter that supplies electrons moving to the base, where $_{n}$ for electrons
    - In the base region
        - I_{Bp1}: the majority carrier hole current that supplies the emitter hole current I_{Ep}
            - I_{Bp1} and I_{Ep} are the same current that are named differently at different locations
    - In the collector region
        - I_{Cn}: most electrons entering the base from the emitter will enter the collector
    - A small amount of electrons will recombine with holes in the base region, causing an extra hole current to enter the base to replenish the recombined holes, marked as I_{Bp2}
    - There is another reverse bias current between the collector and the base, but it is very small and will be ignored
    - 

- With current components identified, we can define some performance related parameters

- **Emitter injection efficiency γ**: measures the ratio of the useful portion of the total emitter current

$$\gamma = \frac{\textbf{useful emitter current}}{\textbf{total emitter current}}$$

  - Among the emitter current, I_{En} is the useful portion that contributes to the collector current, while $I_{\mathrm{E}p}$ is just part of the leakage current $I_{\mathrm{B}p1}$ to the base
  - Therefore,

$$\gamma = \frac{I_{\mathrm{E}n}}{I_{\mathrm{E}n} + I_{\mathrm{E}p}}$$

    Ideally, γ is 1, and typically achievable value is 0.999
- **Base transport factor α_T**: measures the portion of electrons injected from the emitter that reaches the collector
  - In reality, some electrons will recombine with holes at the base, forming $I_{\mathrm{B}p2}$ and leak away instead of reaching the collector

$$\alpha_T = \frac{I_{\mathrm{C}n}}{I_{\mathrm{E}n}}$$

  - Ideally, α_T is 1, and typically achievable value is 0.99
- **Common-base current gain α**: measures the ratio between the collector current and the emitter current with the base voltage referenced to be the ground
  - It is given by

$$\alpha = \frac{I_{\mathrm{C}}}{I_{\mathrm{E}}} = \frac{I_{\mathrm{C}n}}{I_{\mathrm{E}n} + I_{\mathrm{E}p}} = \gamma\alpha_T$$

  - Ideally, α is 1, meaning all emitter current becomes collector current without leakage to the base. Its typically achievable value is 0.99, mainly limited by α_T
- **Common-emitter current gain β**: measures the ratio between the collector current and the base current
  - It is given by

$$\beta = \frac{I_{\mathrm{C}}}{I_{\mathrm{B}}} = \frac{I_{\mathrm{C}n}}{I_{\mathrm{B}p1} + I_{\mathrm{B}p2}} = \frac{\alpha}{1-\alpha}$$

  - Ideally, $\beta = \infty$. Typically, it is around 100 to 200

Current-Voltage Characteristics in the Forward Active Mode

The emitter is usually used as the ground reference. The change in the collector current and the base current can be derived as a function of the collector and the base voltages.

- For V_{BE}, the current flowing through the base-emitter junction is just the PN junction diode current

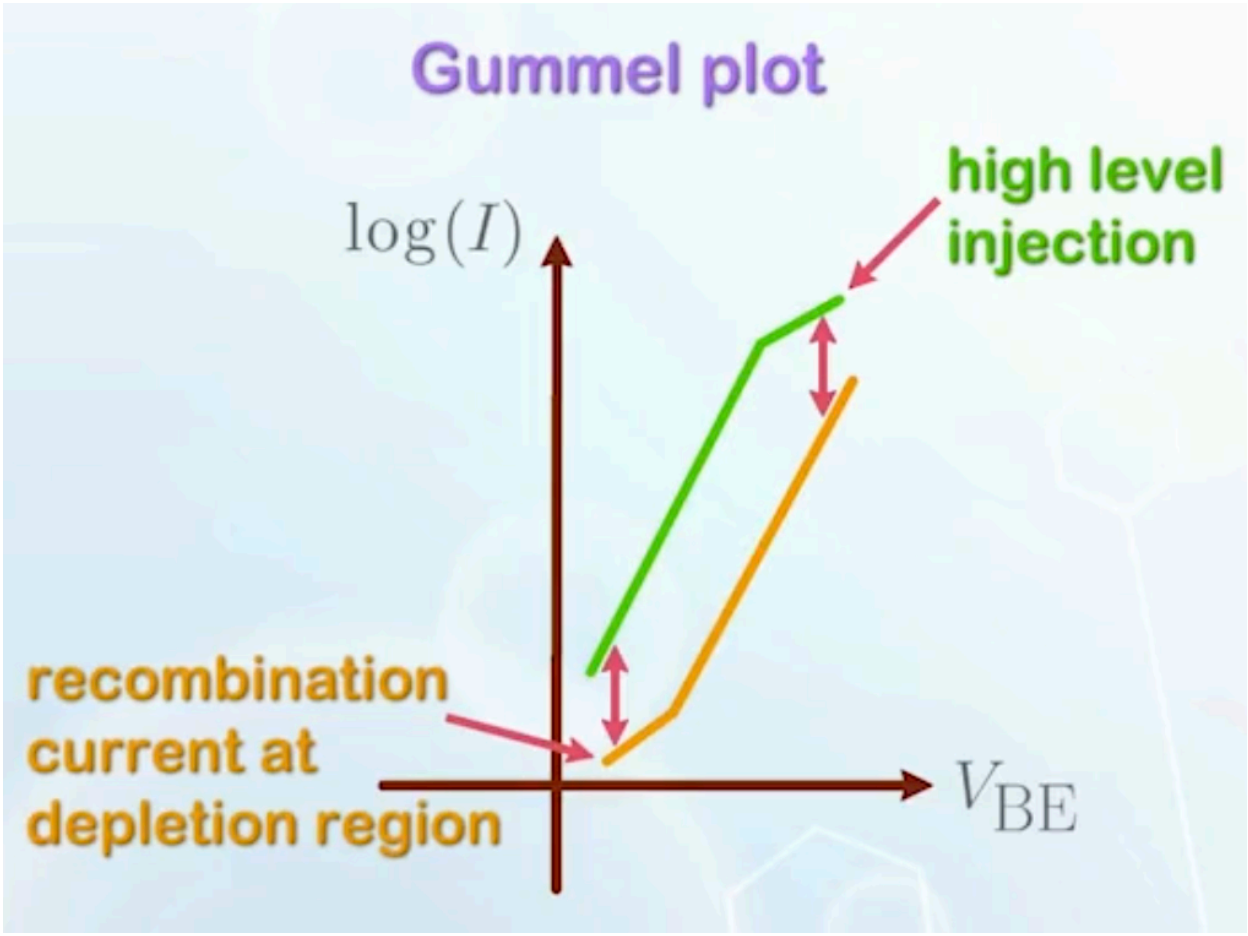
$$I_{\mathrm{E}} = I_{\mathrm{E0}} \left(e^{\frac{qV_{\mathrm{BE}}}{kT}} - 1 \right)$$

- The collector current and the base current are just the partition of the emitter current

$$\begin{aligned} I_{\mathrm{C}} &= \alpha I_{\mathrm{E}} = \alpha I_{\mathrm{E0}} \left(e^{\frac{qV_{\mathrm{BE}}}{kT}} - 1 \right) \\ I_{\mathrm{B}} &= (1-\alpha) I_{\mathrm{E}} = (1-\alpha) I_{\mathrm{E0}} \left(e^{\frac{qV_{\mathrm{BE}}}{kT}} - 1 \right) \\ \frac{I_{\mathrm{C}}}{I_{\mathrm{B}}} &= \frac{\alpha}{(1-\alpha)} = \beta \end{aligned}$$

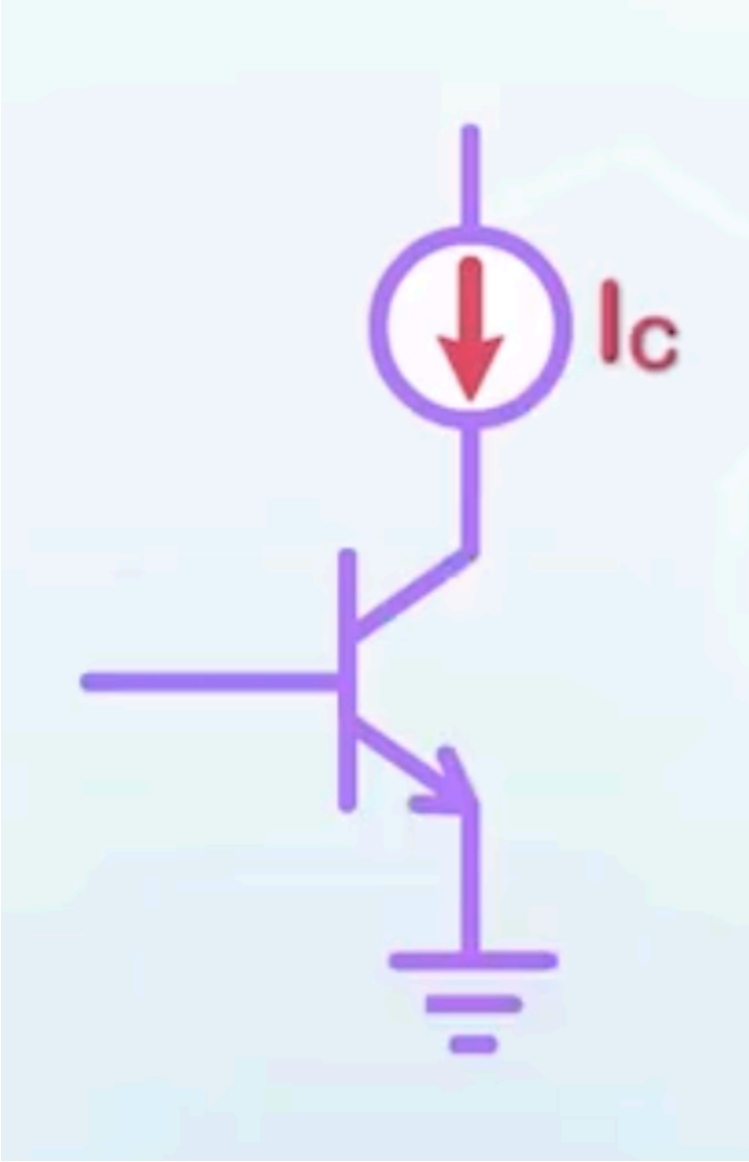
  - Both currents have an exponential dependence on the base-emitter voltage V_{BE}
  - Plotting the current in log scale, we obtain the Gummel plot
    - The vertical distance between the two lines indicates the current gain β, which is constant over a wide range of the curves
    - At very small V_{BE}, recombination current in the depletion region becomes dominant, affecting the base current first as it is smaller
    - At very large V_{BE}, high-level injection effects takes place, leading to a slower increase in the collector current

- These two effects lead to a reduction in β at both ends of the voltage range



- Using V_{BE} as the independent variable, the range of operation is very small and limited to V_{bi}
- In analog circuits, we often need to fix some biasing value so that the transistor will be operating in a known condition
- If we fix V_{BE} to provide the operation condition of a BJT, a small noise can cause a large fluctuation in the current due to the exponential relationship

    - Thus the **current** of a BJT is usually fixed by a known current source



    - Plotting β as a function of $\log I_C$, it will be a constant over many orders of magnitude of the collector current
- For V_{CE}, as the collector voltage is isolated from the base-emitter junction, it has no effect on the collector current
    - The collector current is constant versus V_{CE} except at very small V_{CE}, where the base-collector junction may no longer be reverse biased
    - The collector current will only change when the base current or V_{BE} is changed
    - The change in the collector current is proportional to the base current, with β as the multiplier. But if V_{CE} is used to change the current, the spacing of the current will be very non-uniform

Emitter and Base Design of the BJT

The basic criteria of a BJT are simply maximizing the collector current while minimizing the base current for a given emitter current. For an ideal switch, I_{Bp1} and I_{Bp2} should be zero. This is equivalent to maximizing the emitter injection efficiency γ and the base transport factor α_T.

- For γ

$$\gamma = \frac{I_{\mathrm{E}n}}{I_{\mathrm{E}n} + I_{\mathrm{E}p}} = \frac{1}{1 + \frac{I_{\mathrm{E}p}}{I_{\mathrm{E}n}}}$$

  - To maximize γ, we need to minimize the ratio $\frac{I_{\mathrm{E}p}}{I_{\mathrm{E}n}}$
  - $I_{\mathrm{E}n}$ and $I_{\mathrm{E}p}$ are just electron and hole current of the PN junction

$$\frac{I_{\mathrm{E}p}}{I_{\mathrm{E}n}} = \frac{D_{p\mathrm{E}} p_{n0} L_n}{D_{n\mathrm{B}} n_{p0} L_p}$$

    - Replace the diffusion length with the length of the emitter and the base, as the emitter and base regions of a BJT are usually short
    - Substitute p_{n0} and n_{p0}
    - Finally we have

$$\frac{I_{\mathrm{E}p}}{I_{\mathrm{E}n}} = \frac{D_{p\mathrm{E}} N_{\mathrm{B}A} W_{\mathrm{B}}}{D_{n\mathrm{B}} N_{\mathrm{E}D} W_{\mathrm{E}}}$$

    - Among all the components, the one with the largest tuning range is the doping concentrations
    - To minimize $\frac{I_{\mathrm{E}p}}{I_{\mathrm{E}n}}$, the emitter doping can be designed to be three or more orders of magnitude larger than the base doping

$$1000 N_{\mathrm{B}A} < N_{\mathrm{E}D} \Rightarrow \gamma \approx 0.999$$

  - For α_T

$$\alpha_T = \frac{I_{\mathrm{C}n}}{I_{\mathrm{E}n}} = 1 - \frac{I_{\mathrm{B}p2}}{I_{\mathrm{E}n}}$$

    - $I_{\mathrm{E}n}$ is the PN junction diffusion current, given by the slope of the electron distribution in the base region
      - Assuming $n_{\mathrm{BC}} \ll n_{\mathrm{BE}}$

$$I_{\mathrm{E}n} \approx q D_{n\mathrm{B}} \frac{n_{\mathrm{BE}}}{x_{\mathrm{B}}}$$

where n_{BE} is the electron concentration of the base near the edge of the depletion region of the base-emitter diode, and x_{B} is the width of neutral region at the base
        - We can substitute x_{B} with W_{B} when the depletion region is small compared with the base width
    - To calculate $I_{\mathrm{B}p2}$
        - **Base recombination lifetime τ_{B}**: the average time a carrier will stay in the base region before recombining
        - The combination current is given by

$$\begin{aligned} I_{\mathrm{B}p2} &= q \times (\textbf{number of carriers recombined per second}) \\ &= q \times \frac{\textbf{number of carriers}}{\tau_{\mathrm{B}}} \\ &\approx q\frac{n_{\mathrm{B}}}{\tau_{\mathrm{B}}} \\ &\approx q\frac{n_{\mathrm{BE}}W_{\mathrm{B}}}{2\tau_{\mathrm{B}}} \quad (\textbf{number of carriers given by area under the line}) \end{aligned}$$

    - Putting everything together, we have

$$\alpha_T = 1 - \frac{W_{\mathrm{B}}^2}{2D_{n\mathrm{B}}\tau_{\mathrm{B}}}$$

        - All terms are difficult to adjust except for W_{B}, therefore the most effective way to maximize α_T is to make the base width small
        - Crystal quality in the base region should be good enough to minimize carrier recombination, increasing τ_{B}
        - Typically α_T is only about 0.99, making it more critical than γ to achieve a high transistor gain
- With $\gamma \sim 0.999$, and $\alpha_T \sim 0.99$, we have $\alpha \sim 0.99$ and $\beta \sim 100$

12. Non-Ideal Effects in Bipolar Junction Transistor

About the other three operation modes of BJT, and some non-ideal effects, including the Early effect, base punchthrough, and breakdown.

Reverse Active/Cutoff/Saturation Modes

- **Reverse active mode**: when the base-emitter junction is reverse biased and the base-collector junction is forward biased, electrons are injected from the collector to the base, then to the emitter. Operation is similar to forward active mode
    - The BJT still functions as a switch, which is controlled by the base-collector voltage
    - But the doping is not optimized for this mode, electron injection from the collector to the base has to be accomplished by a large hole current from the base to the collector
    - γ is very small and β is usually less than 1
    - Like using a large amount of running water to control the water flow in a small pipe
- **Cutoff mode**: when both the base-emitter junction and the base-collector junction are reverse biased, no current can flow through the junction. The connection between any two terminals of the BJT can be considered as open circuit
- **Saturation mode**: when both the base-emitter junction and the base-collector junction are forward biased, both emitter and collector inject electrons into the base
    - The minority carrier concentration in the base represents the superposition of the forward active and the reverse active cases, and its slope indicates the net current flow
    - We can calculate the carrier concentrations at the edges of the depletion regions of the base near the emitter and collector as:

$$n_{\mathrm{BE}} = n_{\mathrm{B0}} e^{\frac{qV_{\mathrm{BE}}}{kT}}$$
$$n_{\mathrm{BC}} = n_{\mathrm{B0}} e^{\frac{qV_{\mathrm{BC}}}{kT}}$$

    - If $V_{\mathrm{C}} = V_{\mathrm{E}}$, then $n_{\mathrm{BC}} = n_{\mathrm{BE}}$, the electron injections from the two sides balance out, no current flow through the device
        - Excess electrons at the base will recombine with holes, resulting a base hole current $I_{\mathrm{B}p2}$
        - The base hole current will flow to both the emitter and collector, given by the slope of the minority carrier concentration at the emitter and the collector
        - The currents are mainly hole currents from the base to both the emitter and collector
        - When $n_{\mathrm{BE}} \neq n_{\mathrm{BC}}$, a net electron current flow will be observed in addition to the hole current from the base

Current-Voltage Characteristics in the Saturation Mode

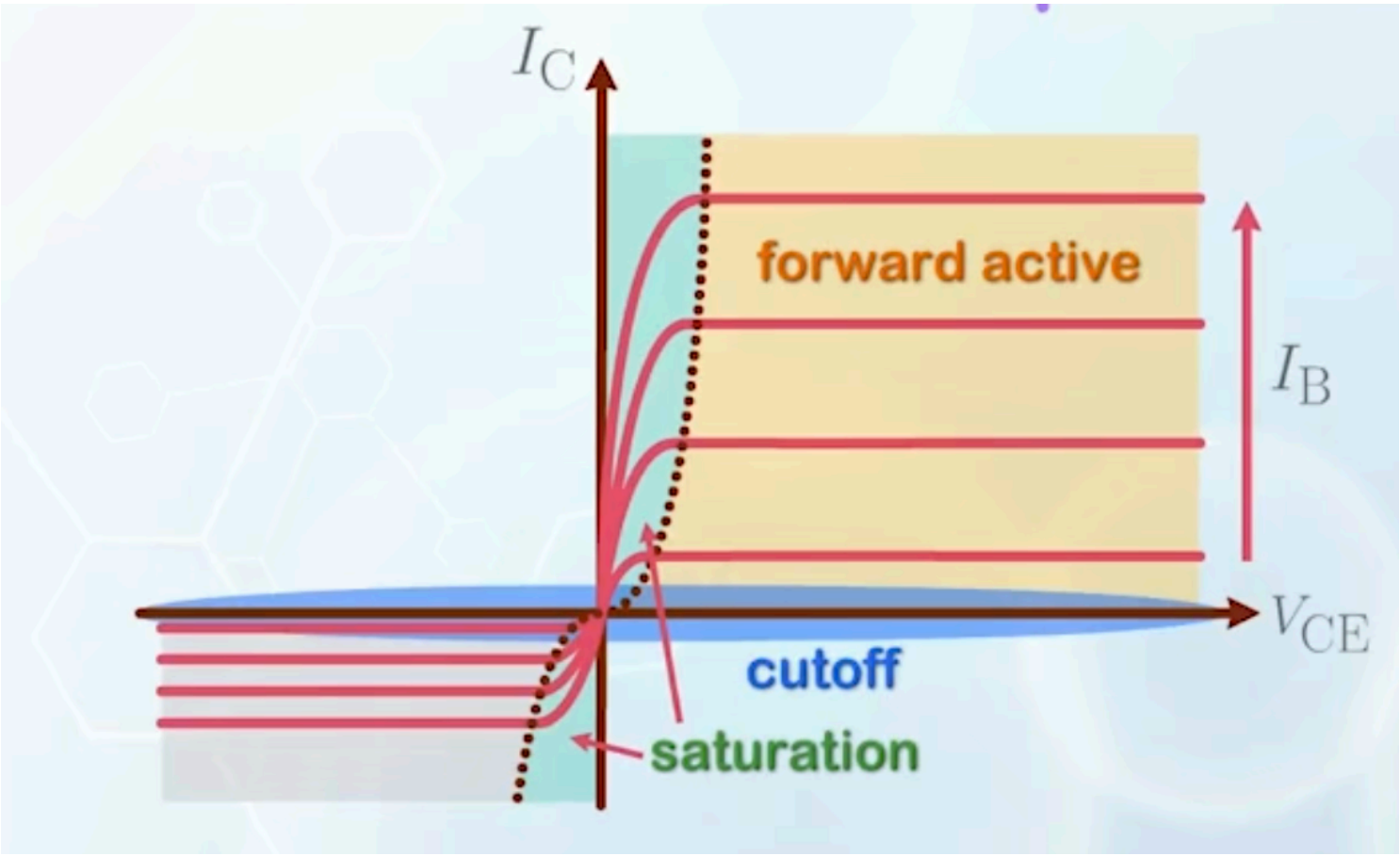
- From the forward active mode, we decrease V_{CE} until it reaches the same as the base voltage, entering the saturation mode

- Further reduction of V_{CE} will cause electrons to be injected from the collector to the base, balancing out some of the electron injection from the emitter to the base, leading to a decrease in the collector current I_{C}
- The hole current from the base to the collector increases, and contributes to a negative collector current that subtracts from the original collector current
- The collector current in the $I_{\mathrm{C}} - V_{\mathrm{CE}}$ curve is expected to drop when V_{CE} is reduced below V_{BE}
- From the device perspective, the boundary between the forward active mode and the saturation mode is the point where $V_{\mathrm{CE}} = V_{\mathrm{BE}}$, which is called the saturation voltage V_{CEsat}
- Measurement results show that I_{C} stays more or less constant after entering the saturation mode, until a much lower value than V_{BE}
    - Because carrier injection follows exponential scale instead of linear scale
    - For example

$$
\begin{aligned}
V_{\mathrm{BE}} &= 0.5\,\mathrm{V} \\
V_{\mathrm{CE}} &= 0.2\,\mathrm{V} \\
\Rightarrow V_{\mathrm{BC}} &= 0.3\,\mathrm{V} \\
\Rightarrow \frac{n_{\mathrm{BE}}}{n_{\mathrm{BC}}} &= e^{\frac{q(V_{\mathrm{BE}} - V_{\mathrm{BC}})}{kT}} = e^{\frac{0.3q}{kT}}
\end{aligned}
$$

$$
\begin{aligned}
&e^{\frac{qV}{kT}} \text{ increases by } 10\times \text{ when} \\
&V \text{ increases by } 60\,\mathrm{mV} \\
\Rightarrow &\frac{n_{\mathrm{BE}}}{n_{\mathrm{BC}}} > 1000
\end{aligned}
$$

    - The electron injection from the collector to the base can be considered minimal, and the characteristics still follow those of the forward active mode
    - The drop in I_{C} will not be noticeable until V_{CE} drops to a very small value (e.g., $0.1\,\mathrm{V}$)
- Based on the observed data, V_{CEsat} is usually assigned a fixed value of $0.2\,\mathrm{V}$, instead of a dependent value determined by V_{BE}.



Difference between BJT and MOSFET Saturation Regions

The saturation region of BJT is different from that of MOSFET. The saturation region of BJT corresponds to the linear region of MOSFET, while the saturation region of MOSFET corresponds to the forward active region of BJT.

The Early Effect

Ideally, the collector voltage does not affect the collector current, thus I_C is a flat line in the $I_C - V_{CE}$ curve in the forward active mode. But in reality, the collector voltage does have an effect on the collector current

- The collector current is given by the slope of minority carrier concentration at the base

$$I_C \approx qD_{nB}\frac{n_{BE}}{x_B}$$
$$x_B = W_B - x_{dBE} - x_{dBC}$$

  - When the collector voltage increases

$$\begin{aligned} V_{\mathrm{CB}} \uparrow &\Rightarrow x_{d\mathrm{BC}} \uparrow \\ &\Rightarrow x_{\mathrm{B}} \downarrow \\ &\Rightarrow I_{\mathrm{C}} \uparrow \end{aligned}$$

leading to a finite slope in the $I_{\mathrm{C}} - V_{\mathrm{CE}}$ curve

- The characteristics are similar to the superposition of a BJT and a resistor in parallel, creating a permanent collector to emitter leakage path, hurting the gain of the transistor
- The resistor is called the output resistance r_O

$$
\begin{aligned}
r_O &= \frac{1}{\text{slope}} \\
&= \left(\frac{\mathrm{d}I_\mathrm{C}}{\mathrm{d}V_\mathrm{CE}} \right)^{-1}
\end{aligned}
$$

$$
\frac{\mathrm{d}I_\mathrm{C}}{\mathrm{d}V_\mathrm{CE}} = \frac{\mathrm{d}I_\mathrm{C}}{\mathrm{d}x_\mathrm{B}} \cdot \frac{\mathrm{d}x_\mathrm{B}}{\mathrm{d}V_\mathrm{CE}}
$$

$$
x_\mathrm{B} = W_\mathrm{B} - x_{d\mathrm{BE}} - x_{d\mathrm{BC}}
$$

$$
\begin{aligned}
\frac{\mathrm{d}I_\mathrm{C}}{\mathrm{d}x_\mathrm{B}} &\approx -qD_{n\mathrm{B}} \frac{n_\mathrm{BE}}{x_\mathrm{B}^2} \\
&= -\frac{I_\mathrm{C}}{x_\mathrm{B}} \\
&\approx -\frac{I_\mathrm{C}}{W_\mathrm{B}}
\end{aligned}
$$

(because x_B is no longer in differential,
we can safely approximate it with W_B)

$$
\begin{aligned}
x_{d\mathrm{BC}} &= \sqrt{\frac{2\varepsilon_\mathrm{Si}(V_\mathrm{bi} + V_\mathrm{CE})}{q} \frac{N_{CD}}{N_{\mathrm{B}A}(N_{CD} + N_{\mathrm{B}A})}} \\
&= \sqrt{\frac{2\varepsilon_\mathrm{Si}}{q} \frac{N_{CD}}{N_{\mathrm{B}A}(N_{CD} + N_{\mathrm{B}A})}} \left(\sqrt{V_\mathrm{bi}} + \frac{V_\mathrm{CE}}{2\sqrt{V_\mathrm{bi}}} \right) \\
\frac{\mathrm{d}x_\mathrm{B}}{\mathrm{d}V_\mathrm{CE}} &= -\frac{\mathrm{d}x_{d\mathrm{BC}}}{\mathrm{d}V_\mathrm{CE}} \\
&= -\sqrt{\frac{\varepsilon_\mathrm{Si}}{2qV_\mathrm{bi}} \frac{N_{CD}}{N_{\mathrm{B}A}(N_{CD} + N_{\mathrm{B}A})}}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\mathrm{d}I_\mathrm{C}}{\mathrm{d}V_\mathrm{CE}} &= \frac{I_\mathrm{C}}{W_\mathrm{B}} \sqrt{\frac{\varepsilon_\mathrm{Si}}{2qV_\mathrm{bi}} \frac{N_{CD}}{N_{\mathrm{B}A}(N_{CD} + N_{\mathrm{B}A})}} \\
&= \frac{I_\mathrm{C}}{V_{EA}} \\
\Rightarrow I_\mathrm{C} &= I_{\mathrm{C}0} \exp\left(\frac{V_\mathrm{CE}}{V_{EA}} \right) \\
&\approx I_{\mathrm{C}0} \left(1 + \frac{V_\mathrm{CE}}{V_{EA}} \right)
\end{aligned}
$$

- If all collector current at different I_B are extrapolated back to the negative V_CE axis, they will intersect at the same point $-V_{EA}$. V_{EA} is called the **Early voltage**

- The slope is given by

$$\frac{\mathrm{d}I_\mathrm{C}}{\mathrm{d}V_\mathrm{CE}} = \frac{I_\mathrm{C}}{V_{EA} + V_\mathrm{CE}}$$

Because V_{EA} is usually very large with typical value of $300\,\mathrm{V}$, the V_CE in the denominator can be ignored for low voltage transistor applications

- The output resistance can be expressed as

$$r_O = \frac{V_{EA}}{I_\mathrm{C}}$$

- V_{EA} of a BJT is not calculated, but measured or supplied by the device specifications

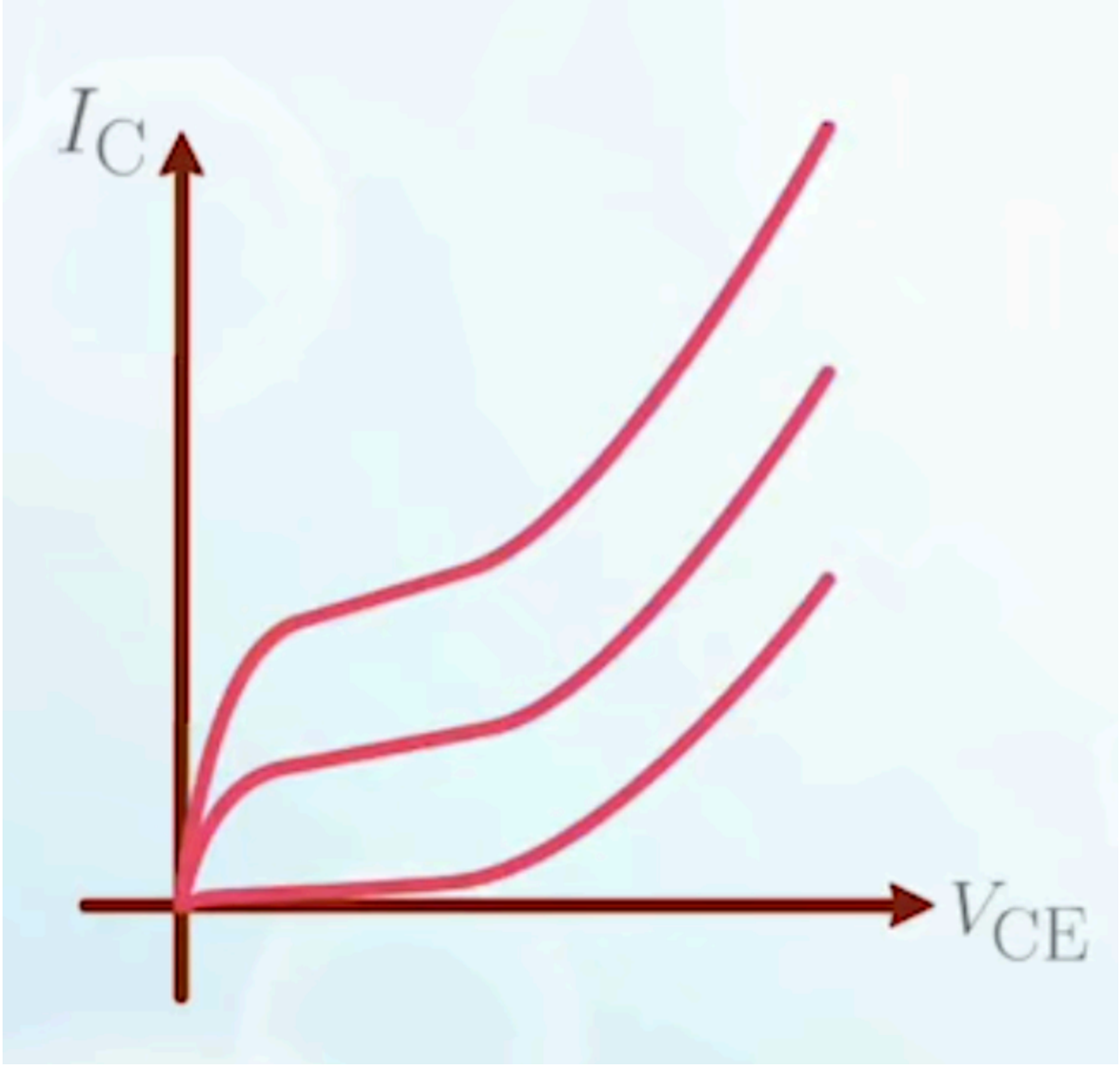
Base Punchthrough

The early effect reduces the gain of the BJT, and should be minimized. This can be done if the change in x_B is small compared to the original x_B, or to minimize $\frac{\Delta x_\mathrm{B}}{x_\mathrm{B}}$

- Easiest way: increase the base width W_B
  - Not applicable, because it will increase more recombination in the base, reducing α_T
- Another choice: minimize the change
  - $x_{d\mathrm{BC}}$ is give by

$$\begin{aligned} x_{d\mathrm{BC}} &= \sqrt{\frac{2\varepsilon_\mathrm{Si}(V_\mathrm{bi} + V_\mathrm{CE})}{q} \frac{N_{CD}}{N_{\mathrm{B}A}(N_{CD} + N_{\mathrm{B}A})}} \\ &= \sqrt{\frac{2\varepsilon_\mathrm{Si}(V_\mathrm{bi} + V_\mathrm{CE})}{q} \frac{1}{N_{\mathrm{B}A}(1 + \frac{N_{\mathrm{B}A}}{N_{CD}})}} \end{aligned}$$

  - This can be achieved by heavy base doping and light collector doping
  - This is why the collector is usually doped more lightly than the base
- For a BJT with extremely small base region, the depletion region at the base-emitter junction and the base-collector junction can touch each other
  - **Base punchthrough**: in extreme case of Early effect, the base start to lose control of the electron flow, and is unable to stop the current flow
  - Once the depletion regions touch, V_CE can directly lower the barrier height, and the current show an exponential dependency on V_CE

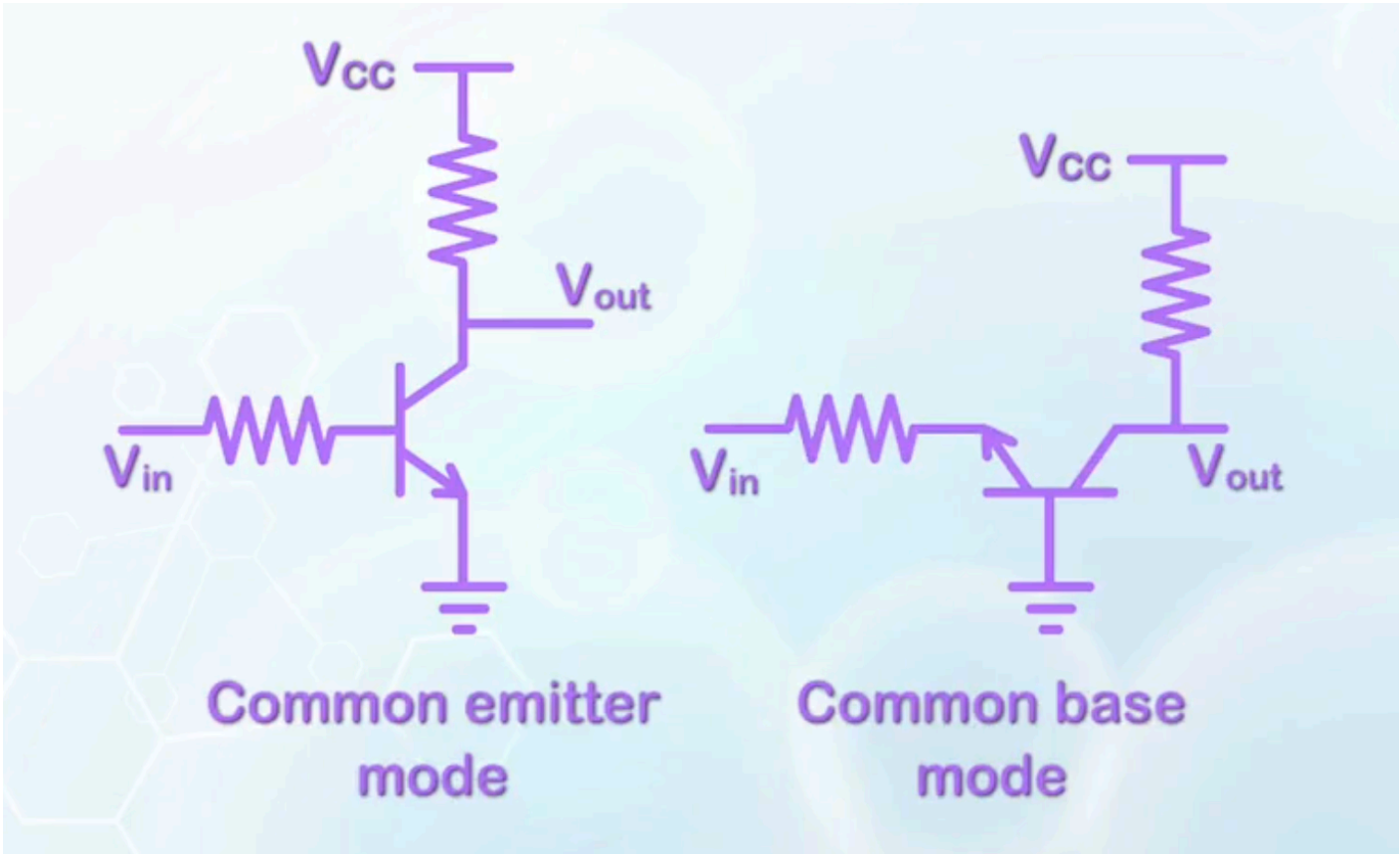


- Base punchthrough limits the minimum base width of a BJT
- To avoid base punchthrough, the same method used to reduce Early effect can be applied
  - Large base width
  - Heavy base doping and light collector doping

BJT breakdown

When a BJT is used as a switch, it is usually placed between two high voltage nodes to control the current flow. The highest voltage the BJT can withstand is limited by the breakdown voltage of the BJT.

- There are two ways to connect a BJT as a switch



  - In common base mode
    - When a high voltage is applied to the emitter, the BJT is in cutoff mode, and no current flows
    - When a negative voltage is applied, a current will flow through the load
    - The maximum voltage the BJT can withstand is determined by the breakdown voltage of the reverse biased **base-collector junction**
    - But the breakdown voltage **also** depends on the **input**, which can be high, low, or open
      - In most cases, the worst case with lowest breakdown voltage is when the input is open
      - This lowest breakdown voltage is denoted as BV_{CBO}: **breakdown voltage in common-base configuration with emitter open**
      - In this case, the BJT just has the base-collector junction reverse biased, and the breakdown voltage is the same as that of a PN junction
      - To increase breakdown voltage, the most effective way is to reduce doping on the lightly doped side (the collector side)
  - In common emitter mode
    - Again, the breakdown voltage is the lowest when the input is open

- The breakdown voltage is denoted as BV_{CEO}: **breakdown voltage in common-emitter configuration with base open**
- $BV_{\text{CEO}} < BV_{\text{CBO}}$ due to a possible feedback mechanism
  - Assume the base-collector junction is reverse biased at a particular voltage such that for every 10 electrons entering the base-collector diode, there is a probability of generating one electron/hole pair
  - Under normal condition, this voltage is not high enough to cause breakdown
  - But in common-emitter configuration, the hole generated in the impact ionization process will enter the base
  - As the base is open, the hole can only enter the emitter
  - To allow the hole to enter the emitter, an associated voltage results at the floating base
  - The ratio of the electron to hole current when the base emitter junction is lowered, is related to β
  - Assume $\beta = 100$, a single hole entering the emitter will cause 100 electrons to be injected from the emitter to the base
  - These 100 electrons will enter the collector, causing 10 more electron/hole pairs to be generated
  - This positive feedback will cause a very high collector current, even though the initial voltage is not high enough to cause breakdown
- To increase the breakdown voltage
  - Reduce the chance to generate the first electron/hole pair by reducing the electric field, achievable by reducing the collector doping
  - Reduce β to trade off between gain and breakdown voltage

Summary of BJT Design

- Very high emitter doping relative to base, to increase γ
- Very small base width, to increase α_T
  - So that the BJT behaves more like an ideal switch with high gain
- Base width cannot be too small, limited by the Early effect and base punchthrough
- Collector doping should be low to reduce Early effect and increase breakdown voltage

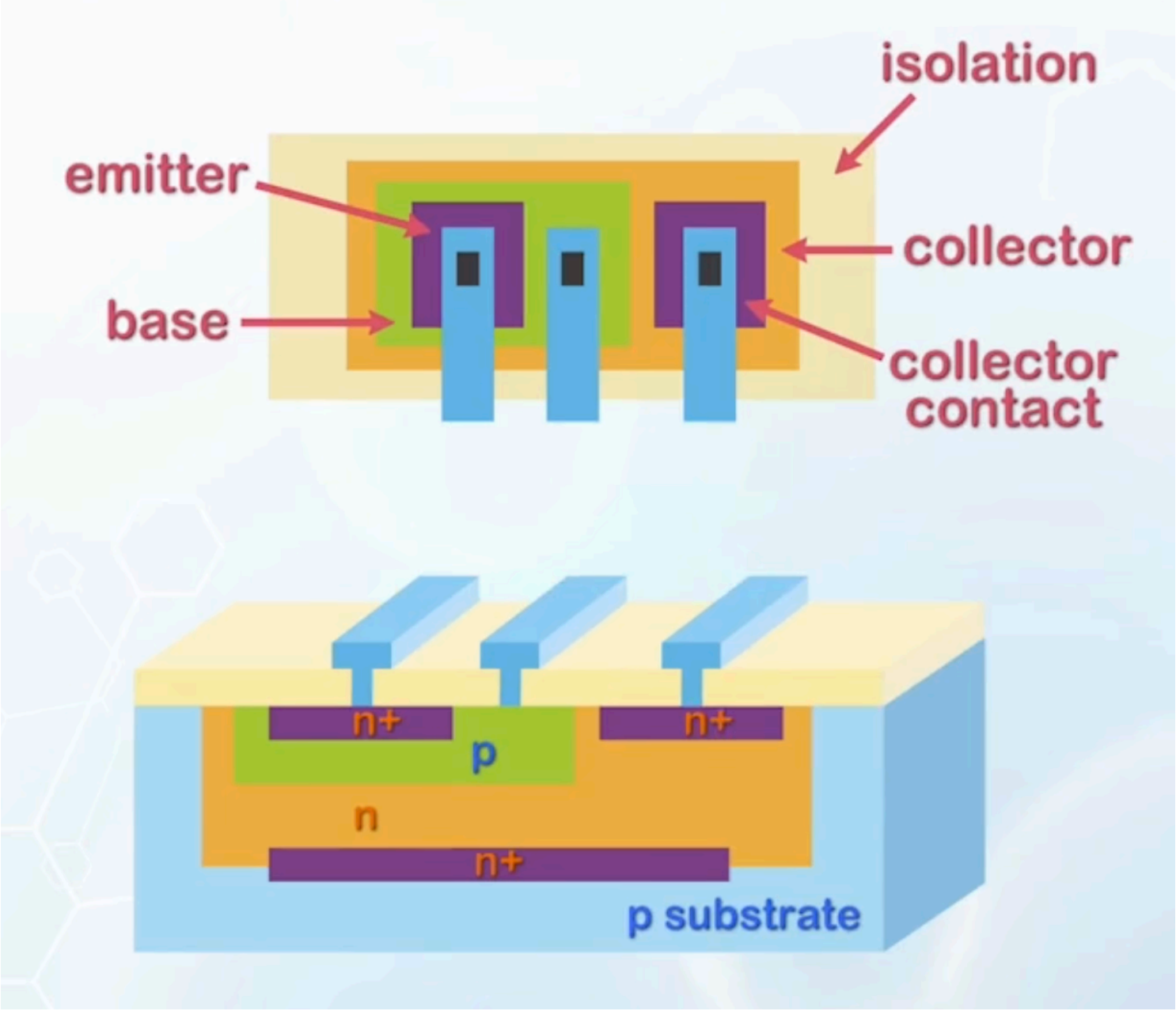
13. Physical Structure and Switching

About the physical structure of BJTs, and the switching transient behavior.

Physical BJT Design

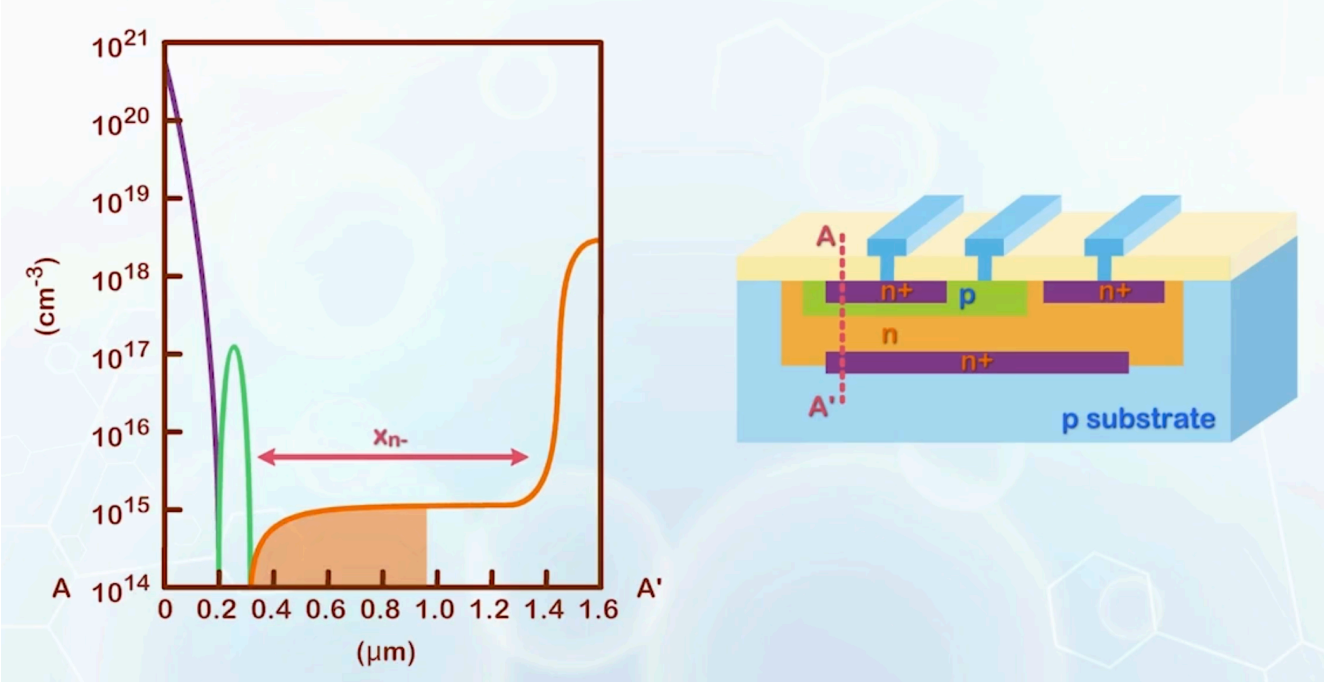
On a physical wafer, BJTs can be placed horizontally or vertically. The latter is more common in integrated circuits, as doping is usually performed by ion implantation from the front side of the wafer, and it is easier to form heavily doped region near the surface.

- The emitter is usually placed on the top, followed by the base, and the lightly doped collector at the bottom
- The emitter can be directly connected from the top of the wafer
- The base must extend to the surface for external connection
- Similarly, the collector must also extend to the surface for external connection
    - The collector resistance will be very high because it is lightly doped
    - To reduce the resistance, a buried N+ layer is used, and an N+ region at the top of the collector region is formed during the emitter doping to provide an ohmic contact between the metal and the N- silicon collector
- Isolate BJT from other devices with a P region, usually the silicon substrate itself



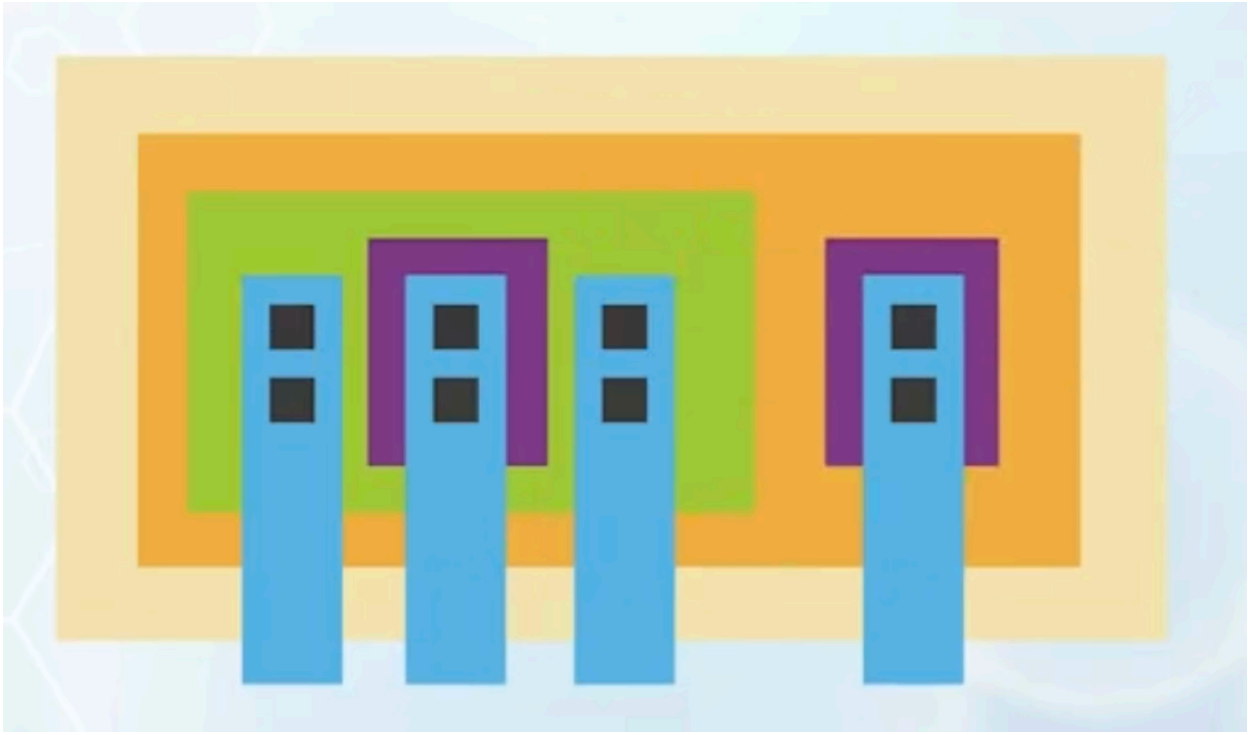
- The BJT only takes up a small area of the entire structure, and a large portion of area is used for device isolation and contacts
- Plot the doping concentration along the vertical cutline

- The curve:



- The emitter is doped as heavily as possible ($10^{20}\ \mathrm{cm}^{-3}$) to maximize γ
- The base is narrow to reduce α_T, but not too narrow to avoid punchthrough
- The collector is lightly doped to reduce the Early effect, to avoid base punchthrough, and to increase the breakdown voltage, usually limited by the fabrication process ($10^{15}\ \mathrm{cm}^{-3}$)
- The base doping should be selected carefully to trade off between γ and r_O, typically around $10^{17}\ \mathrm{cm}^{-3}$
- At the end of the collector is the heavily doped N+ region to reduce collector resistance
  - It is another trade off:
    - When the reverse bias voltage of the base-collector junction is increased, the depletion region width x_{dC} expands,
    - If the N- region is too short, the depletion region will touch the N+ region before the maximum operation voltage, the depletion region will stop expanding, forcing the reverse voltage to drop across a smaller depletion width, increasing the electric field and reducing the breakdown voltage. The average effective doping in the depletion region also increases
    - If the N- region is too long, electrons still have to travel a long distance in the lightly doped region, increasing the collector resistance
- Besides the collector resistance, there is also a base resistance, and the holes need to travel a long distance to recombine with electrons in the emitter
  - The resistance may be very high because the base is very thin
  - To reduce the base resistance, the lateral dimension of the base should be kept small, to minimize the distance between the base contact and all locations at the base

- Adding another base contact to the left may also help, but that requires extra area
- In terms of layout, a long emitter is preferred to reduce the base resistance



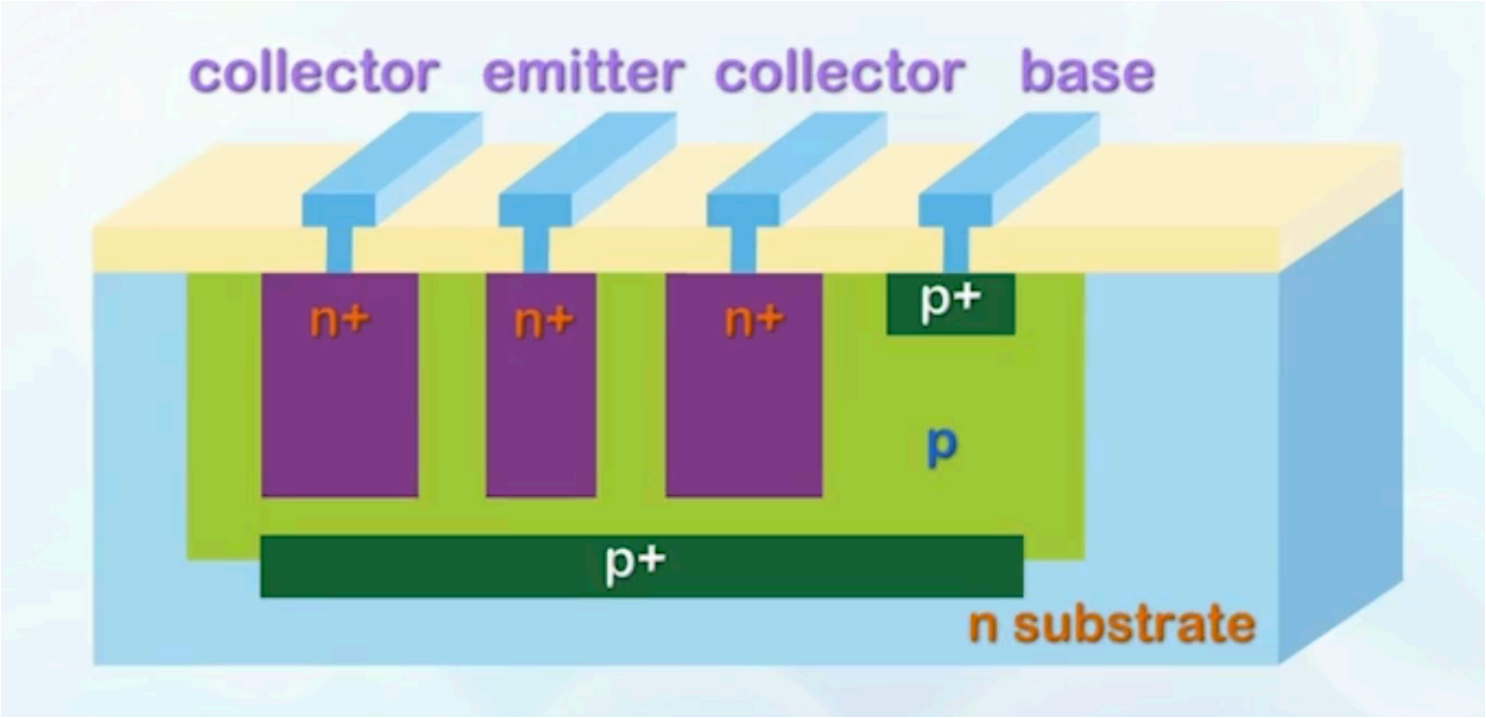
Lateral BJT

BJTs can also be laid out laterally on the wafer surface, as a lateral BJT.

- The region at the bottom of the structure must be P type, otherwise the N regions will be shorted without any barriers in between
- But if the substrate is P type, all the bases of lateral BJTs on the substrate will be shorted together
- Therefore, we have the P region forming the base **extended** to the region below the emitter and the collector, and build the whole device on an **N substrate**
- If breakdown voltage is not a concern, the collector can also be heavily doped to N+, and is formed together with the emitter, to simplify the fabrication process
- The separation between the emitter and the collector is kept small to reduce recombination at the base. This can be achieved by placing the base contact **outside** the main BJT structure
- Electrons can be injected to the base through the bottom of the N+ region. These electrons have to travel a much longer distance to reach the collector, and is easily lost to recombination
    - Thus the base transport factor α_T is usually very small in lateral BJTs
    - To improve the gain, the side wall area of the emitter should be increased relative to the bottom area, by making both the emitter and collector junctions deep, and the lateral

dimension of the emitter small. An additional collector region can also be added to the other side of the emitter

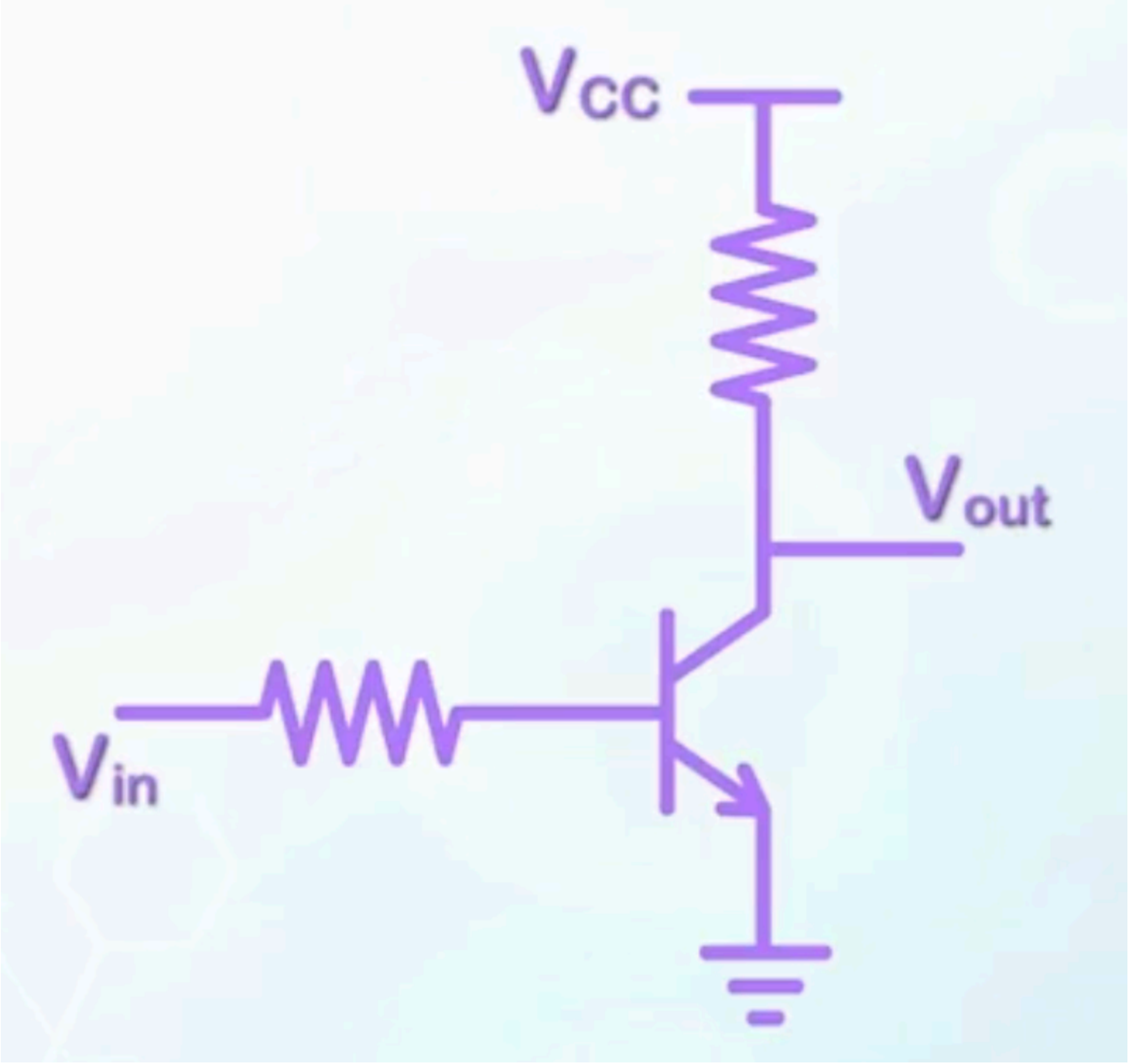
- The depth of the base region is no longer restricted by the collector region, as in the vertical BJT, a P+ region can be introduced at the bottom of the base to reduce the base resistance



The vertical dimension can be more precisely controlled relative to the lateral dimension in IC processing, the BJT technology is still dominated by vertical BJTs.

Turn on Transient

Consider BJT connected as an inverter.



- When a BJT is switched, it will switch between different operation modes, and has to be companied by the change in the amount of minority carrier charge
- Assume $V_{CC} = 5\,\text{V}$, V_{in} is negative, the BJT is in cutoff
  - The BJT is behaving like an open-circuit, V_{out} is the same as V_{CC}
  - When the base voltage is switched to a logic high at time $t = 0$, the base terminal is initially held at $0\,\text{V}$ by the capacitance
  - The initial base current is $I_B = \frac{V_{CC}}{R_B}$

Why VCC instead of Vin?

> The logic high of V_{in} is usually just V_{CC}

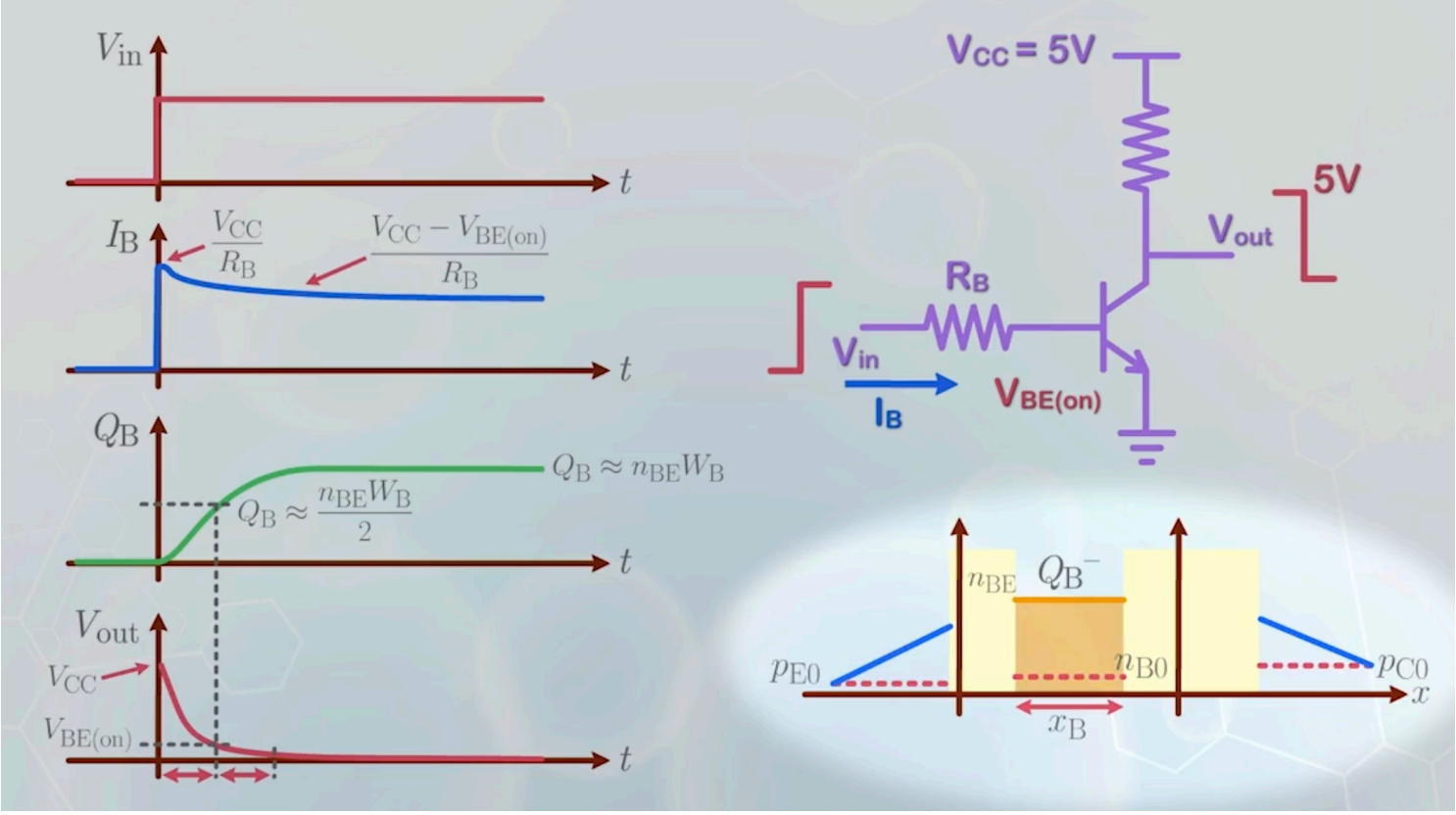
- Over a short period of time, the BJT will enter the forward active region
  - The voltage across the base-emitter junction will be equal to $V_{\text{BE(on)}}$, typically assumed to be $\mathbf{0.7\,V}$
  - All voltage beyond the value will be dropped across the resistor R_{B}, thus

$$I_{\text{B}} = \frac{V_{\text{CC}} - V_{\text{BE(on)}}}{R_{\text{B}}}$$

- Now there are excess negatively charge electrons (the part below the minority carrier concentration distribution line) stored temporarily in the base region, and a base current is needed to supply holes to the base
- The amount of accumulated holes at the base Q_{B}^{+} will be the same as the amount of accumulated electrons Q_{B}^{-}. Q_{B} is called the **base charge**
- Supplying electrons from the emitter is fast due to the large electron current
- Supplying holes from the base contact is slow due to the small base current
- Insufficient Q_{B}^{+} will slow down electron motion due to electrostatic repulsion, thus the overall Q_{B} will increase gradually over time
- To calculate Q_{B} at steady state condition, we use the electron concentration, or Q_{B}^{-} at the base, but **the speed is really limited by Q_{B}^{+}**
- Over time, I_{C} increases, causing V_{out} to decrease
- Assume V_{out} will drop to $\mathbf{0}$
  - The base current continues to supply the base charge Q_{B}, and saturates at

$$Q_{\text{B}} = n_{\text{BE}} x_{\text{B}} \approx n_{\text{BE}} W_{\text{B}}$$

  - The time at which the BTJ switches from forward active to saturation is marked in the timing diagram when the base charge is half of its saturation value
  - At this point, V_{out} will drop to $V_{\text{BE(on)}}$, and takes about the same amount of time to drop the remaining voltage to achieve full saturation, determined by the time to fully charge the base with a constant base current
- The supply of the base charge is usually used as a measure for the speed of BJT switching



Turn off Transient

Similar to turn on, the delay for turn off is caused by the time to remove the accumulated base charge. It is actually more significant than the turn on delay.

- V_{in} becomes 0 at time $t = 0$
- The charge at the base has to be removed before the BJT can enter forward active, then cutoff
- The removal of the accumulated holes is achieved by the base current passing through the base resistor to ground
    - The minority carrier charge in the neutral base region near the base-emitter junction sustains the junction in forward bias, and junction voltage remains at $V_{BE(on)}$, and this voltage is the driving force for the base current
    - The base current is given by

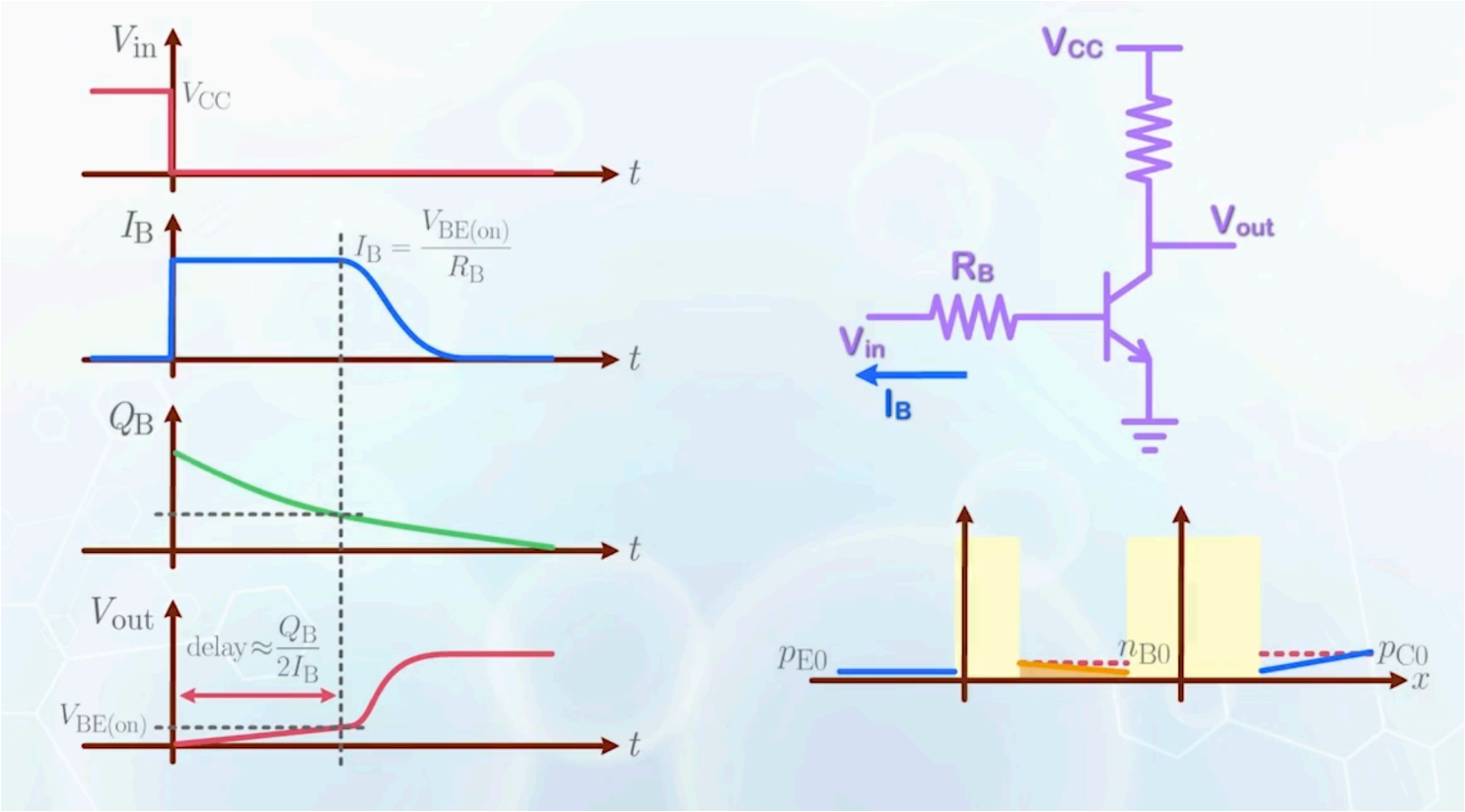
$$I_B = \frac{V_{BE(on)}}{R_B}$$

        - $V_{BE(on)}$ is much smaller than V_{CC}, thus it takes a longer time to remove the base charge
    - The base current will remain constant for a long period of time, until the BJT enters forward active

- After that, I_{B} will decrease, while Q_{B} continues to decrease
- V_{out} starts at 0 at saturation, and increases to about $V_{\mathrm{BE(on)}}$ when the BJT enters forward active, then rise to V_{CC} when the BJT enters cutoff
  - There is a long delay before the output voltage rises
  - The approximation delay to reach the forward active mode is given by

$$\text{delay} \approx \frac{Q_{\mathrm{B}}}{2I_{\mathrm{B}}}$$

  - The overall speed of a BJT when connected in the common-emitter configuration is usually limited by turn off



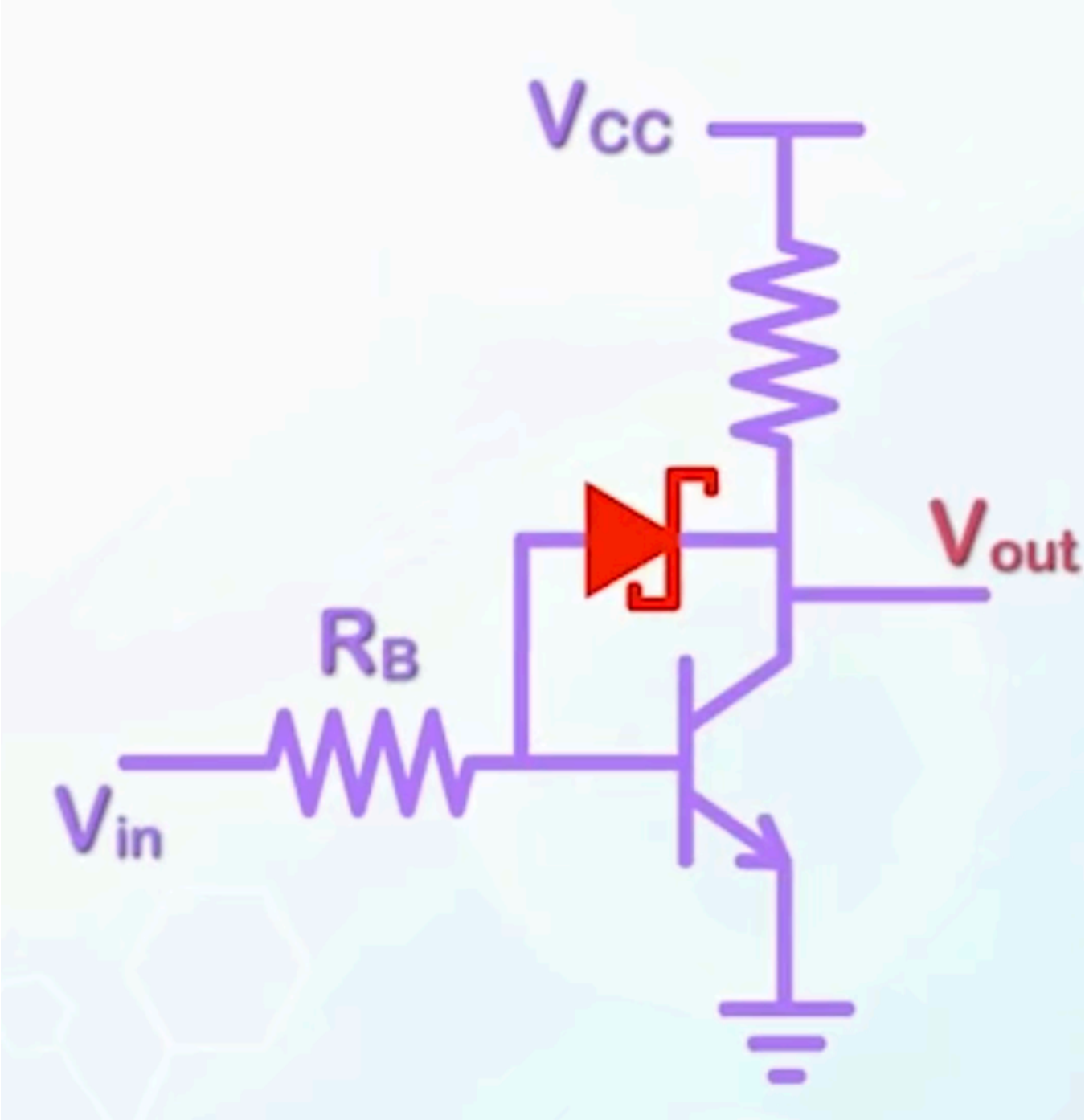
Schottky BJT

Compared to switching from cutoff to forward active, switching from forward active to cutoff only increases the output voltage range by $0.7\,\mathrm{V}$, but requires removing or accumulating the same amount of base charge.

By limiting the minimum voltage to $0.2\,\mathrm{V}$ instead of ground, the long delay can be eliminated.

Due to the exponential dependence, $0.2\,\mathrm{V}$ can create more than 10^3 times difference between n_{BE} and n_{BC}, enough to eliminate most of the base charge.

A Schottky diode can be connected between the base and collector to prevent the base-collector junction from going into deep saturation.



- When $V_{\text{out}} > V_{\text{BE(on)}}$, the Schottky diode is open circuited, and the BJT operates normally in forward active
- The turn on voltage for the Schottky diode is about $\mathbf{0.3\,V}$, lower than that of a silicon PN junction
- When the BJT enters shallow saturation, V_{out} drops to $\mathbf{0.4\,V}$, the Schottky diode turns on, and shorts the base to the collector, with $\mathbf{0.3\,V}$ offset
- The collector voltage will be maintained at $\mathbf{0.4\,V}$, preventing deep saturation

- The amount of base charge is kept about half of that without the Schottky diode
- The initial delay in turn off will be eliminated, and the overall switching speed is improved, and the trade off is the reduced output voltage swing
- To physically implement the Schottky diode, place a metal with Fermi level close to the valence band edge of Silicon at the base-collector junction
    - It will form a Ohmic contact with the base, and a Schottky barrier with the collector
    - Most commonly used metal for this structure is gold

14. Bipolar Junction Transistor Models

About modeling BJTs under different operation modes, small signal model for amplifier design, frequency response and structural optimization.

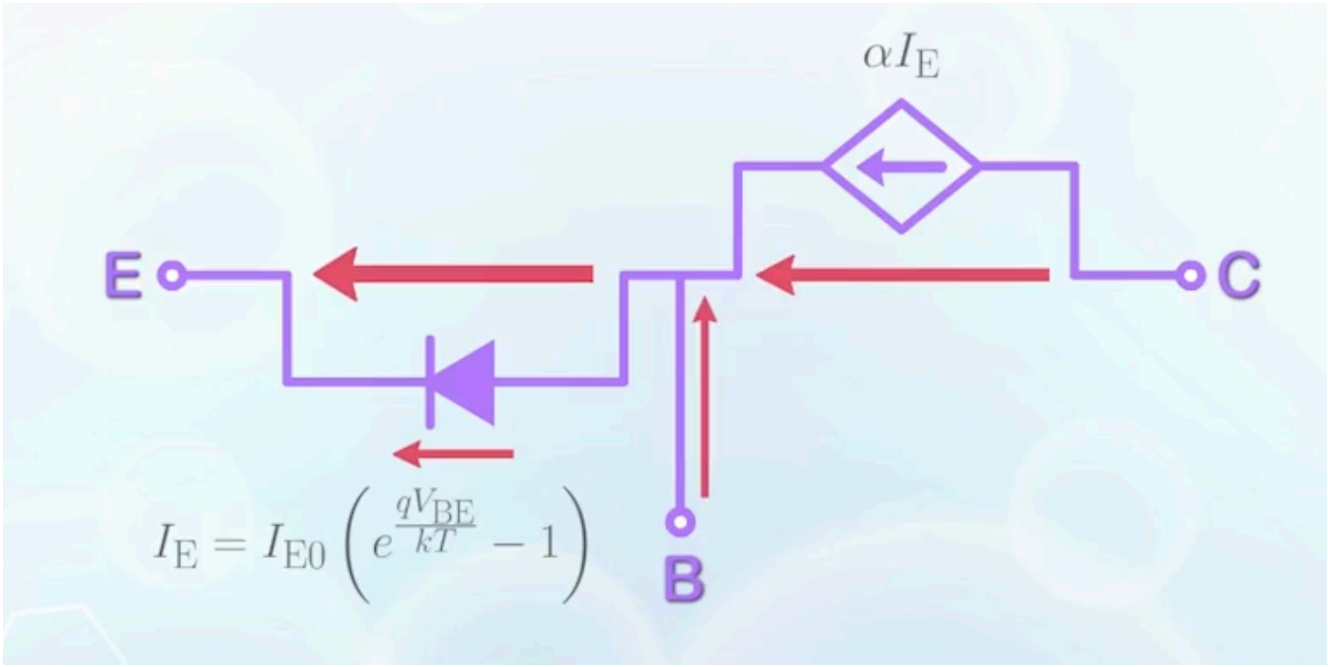
The Ebers-Moll Model

- To describe a BJT in **forward active mode** in steady state:
    - The operation of a BJT is mainly based on the PN junction diode between the base and emitter, we place a diode between the base and emitter in the model
    - The emitter current is just the diode current given by

$$I_{\mathrm{E}} = I_{\mathrm{E0}} \left(e^{\frac{qV_{\mathrm{BE}}}{kT}} - 1 \right)$$

    - A current controlled current source is placed between the collector and emitter, to represent α part of the emitter current that reaches the collector
    - The difference between I_{E} and I_{C} is the base current

- The model describes exactly the BJT operation in forward active mode in steady state

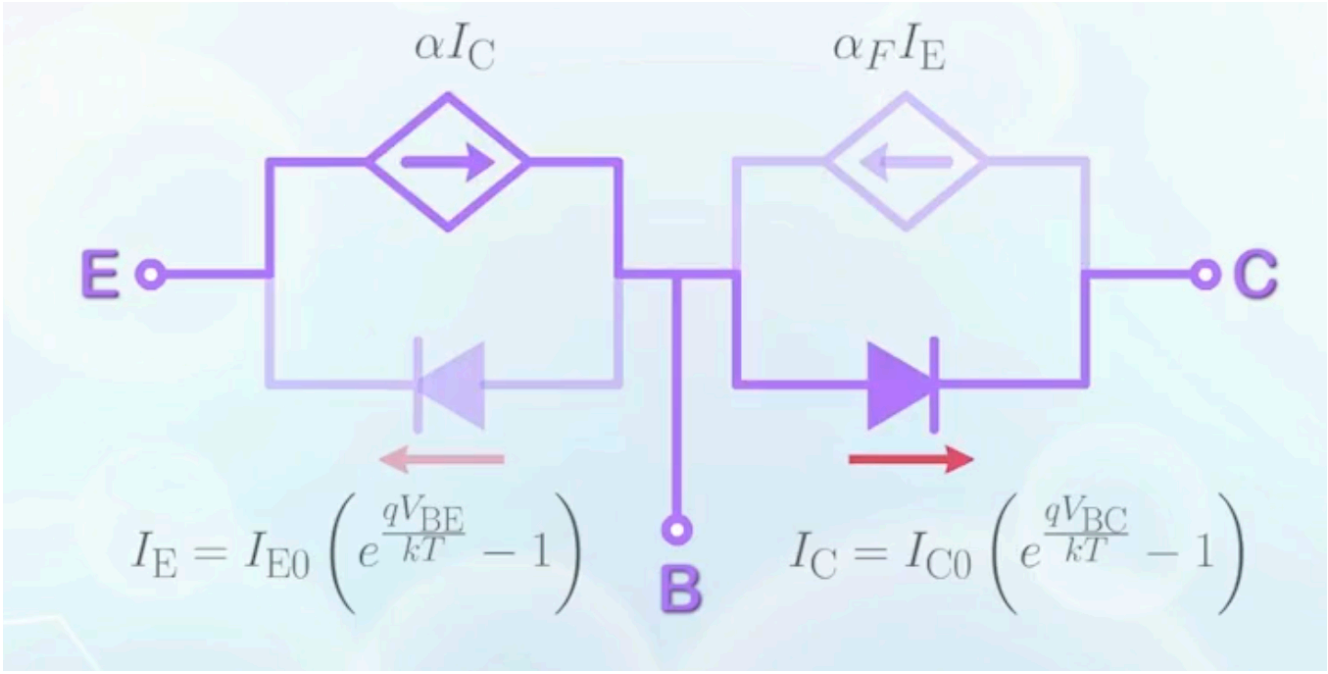


- In **reverse active mode**:
  - The transistor actions come from the base-collector junction, which is another PN junction diode
  - The collector current is again given by the ideal diode equation

$$I_C = I_{C0}\left(e^{\frac{qV_{BC}}{kT}} - 1\right)$$

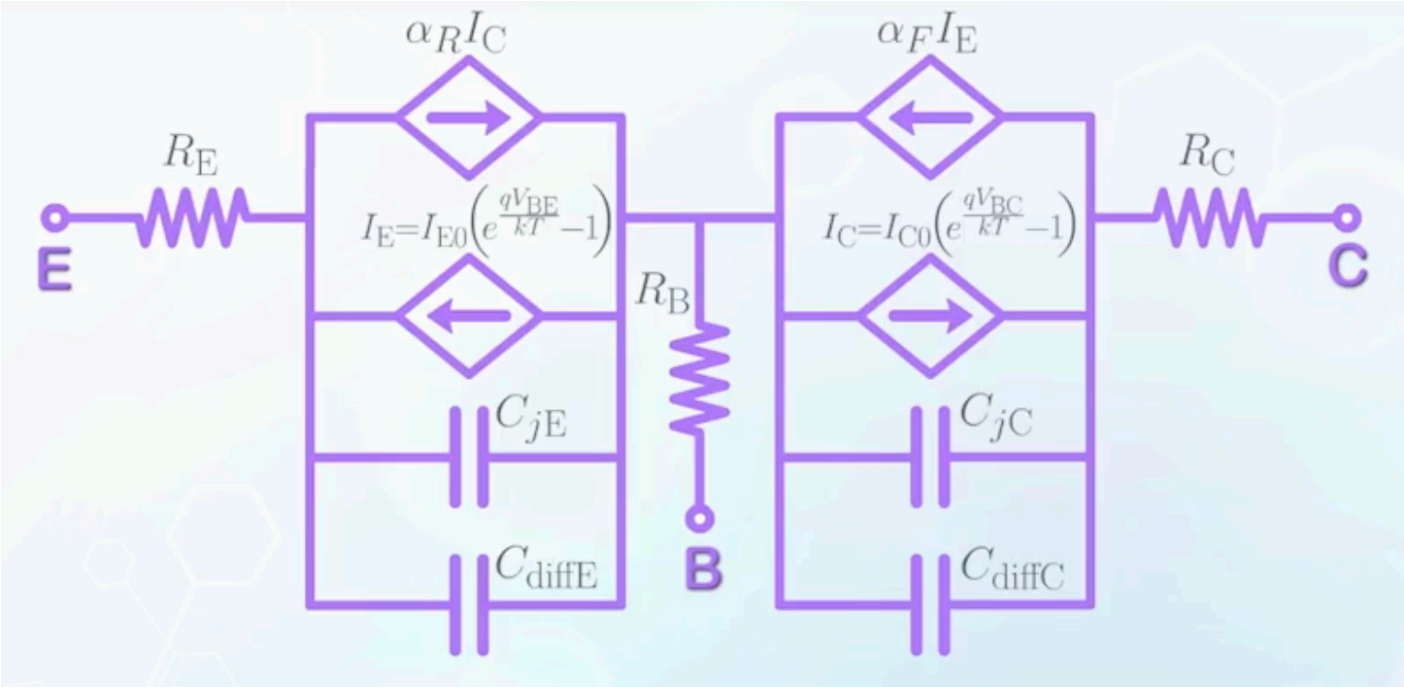
  - A current controlled current source is placed between the emitter and collector, to represent α_R part of the collector current that reaches the emitter
    - The α in reverse active mode is different from that in forward active mode
    - To distinguish the two, we denote the reverse active mode one as α_R, and the forward active mode one as α_F

- The model for BJT under reverse active mode:



  - In forward active mode, the base-collector diode is in reverse bias, and can be regarded as an open circuit, $I_C \approx 0$. Thus, components related to the reverse active mode can be ignored
  - The same applies to reverse active mode, where the forward-active-mode-related components can be ignored
- In **cutoff mode**, all diodes are considered as open circuits, and no current flows
- In **saturation mode**, both diodes are forward biased, and injects current to the base
  - The net current is determined by the relative magnitudes of the injection currents
- The model now can describe the BJT operation in all four modes
- It should be noted that the diodes in the model only represent the function of the voltage-controlled current source in the diode model
- To capture the dynamic behavior, the **full model of the diode**, as discussed in the PN junction section should be included
- The currents though the base, emitter and the collector terminals are different in the case of BJT, the base, emitter and collector resistances have to be associated with the respective

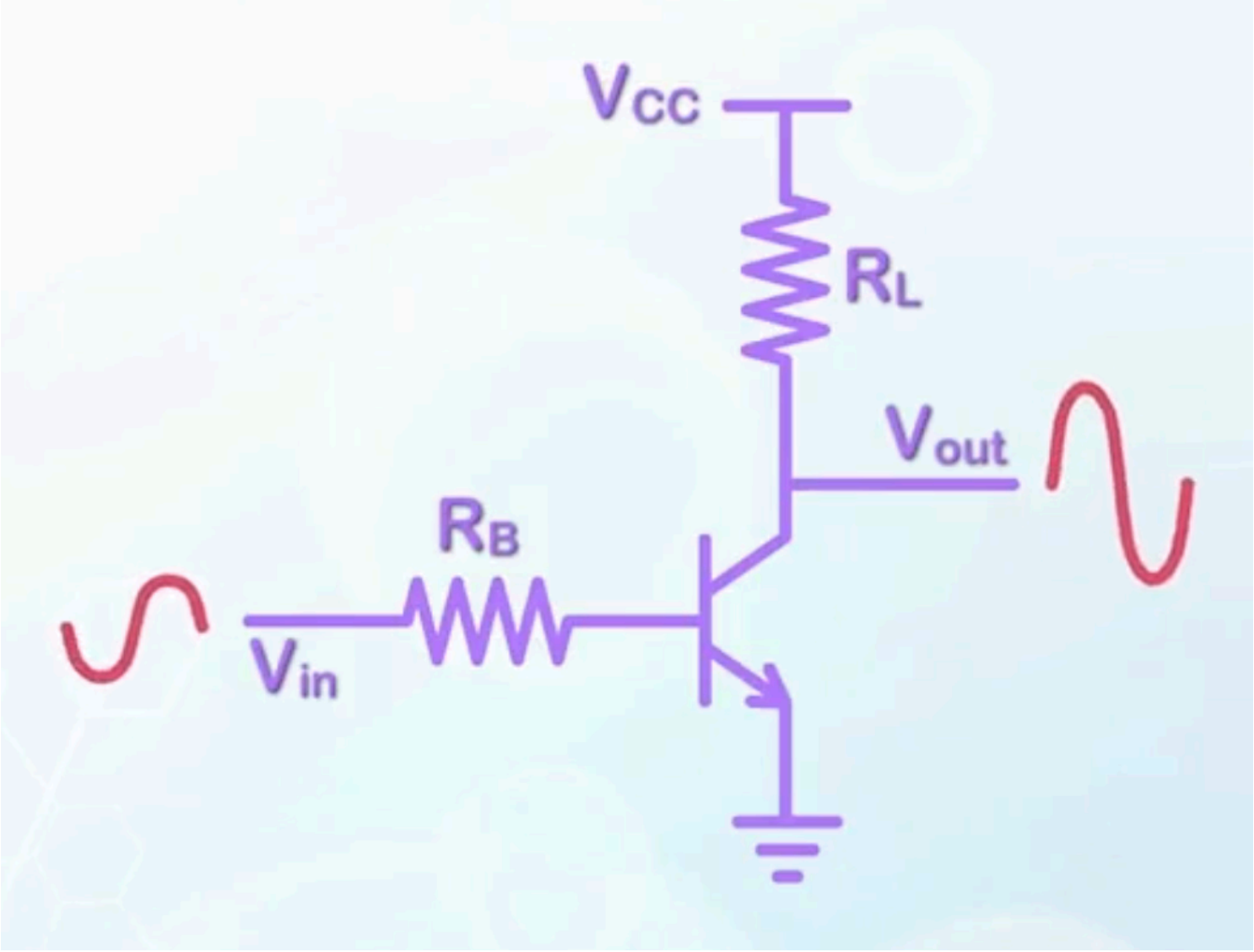
terminals



- This is still an over-simplified situation, as it assumes **the diodes operate independently**
- More detailed models should be able to capture the interaction between the two junctions

The Small Signal Model

Besides digital switches, BJTs can also be used as amplifiers, common in signal processing circuits.



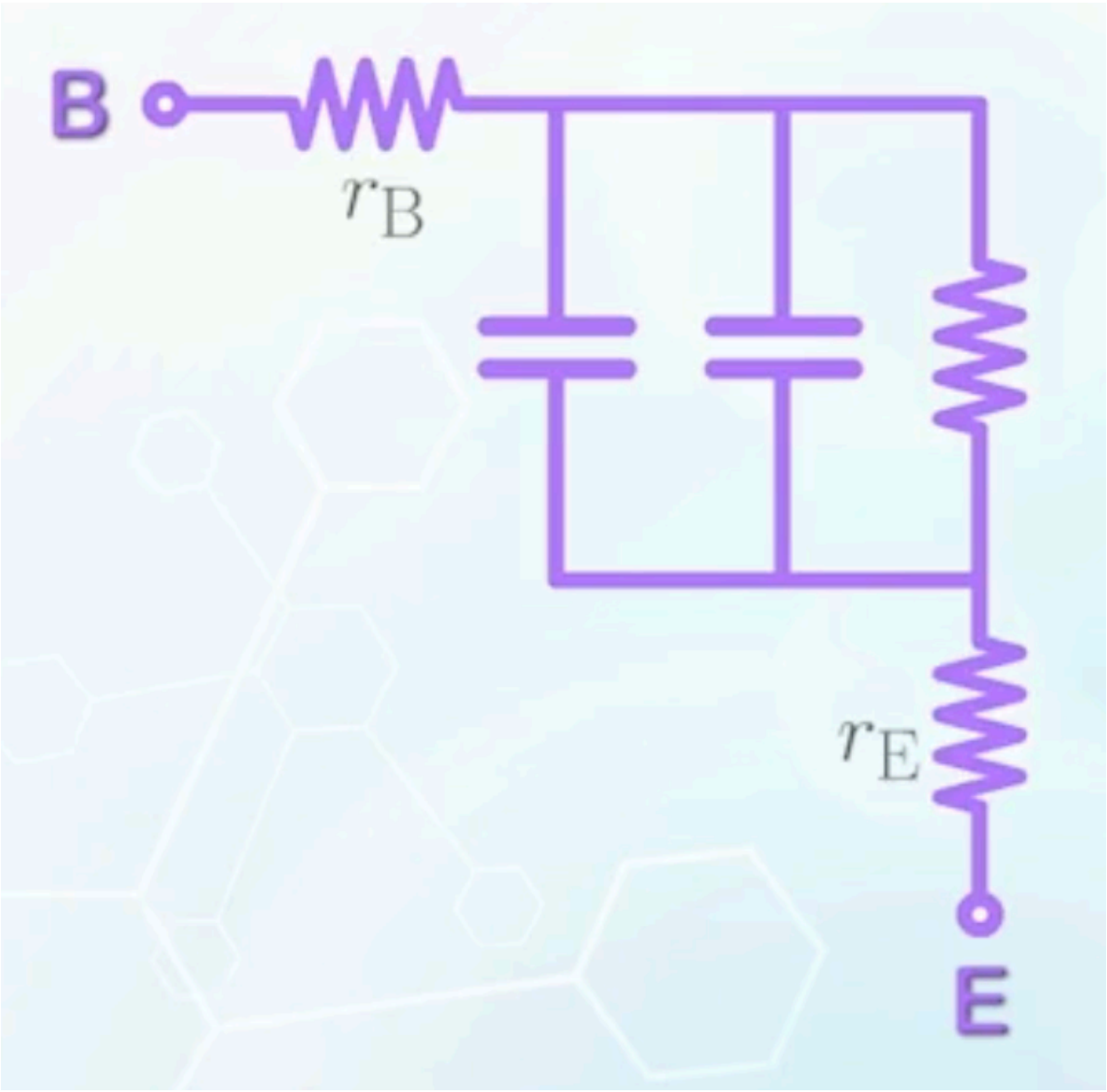
- A very small varying voltage can be applied to the input
- Based on the exponential relationship between I_C and V_{BE}, a small change in V_{BE} can cause a large change in I_C
- The collector current change will be converted to the output voltage, multiplied by the load resistance R_L

$$\Delta V_{out} = -\Delta I_C R_L$$

  A small input voltage change can lead to a large, inverted output voltage change
- The input signal is assumed to be small, or it will saturate the waveform, leading to clipping
- For a very small signal input, we can linearize the circuit around a particular biasing point, usually in the forward active mode

- To construct the small signal model
  - Start from the PN junction between the base and emitter

- Copy the small signal diode model from the PN junction section



- With the emitter assumed to be grounded, the resistance r_π measures the reciprocal of the slope of the I_B-V_{BE} curve at the biasing point

$$\begin{aligned} r_\pi &= \left(\frac{\mathrm{d}I_B}{\mathrm{d}V_{BE}}\right)^{-1} \\ &\approx \left(\frac{\mathrm{d}}{\mathrm{d}V_{BE}} I_{B0} e^{\frac{qV_{BE}}{kT}}\right)^{-1} \\ &= \left(\frac{q}{kT} I_{B0} e^{\frac{qV_{BE}}{kT}}\right)^{-1} \\ &= \frac{V_{th}}{I_B} \end{aligned}$$

- C_{jE} is the base-emitter junction capacitance at the biasing point

$$C_{j\mathrm{E}} = \frac{C_{j\mathrm{E}0}}{\sqrt{1 - \frac{V_{\mathrm{BE}}}{V_{\mathrm{bi}}}}}$$

- C_π is the diffusion capacitance due to the modulation of charge at the base

$$C_\pi = \frac{\mathrm{d}Q_{\mathrm{B}}}{\mathrm{d}V_{\mathrm{BE}}} = \frac{Q_{\mathrm{B}}}{V_{\mathrm{th}}}$$

  - In BJT, we prefer to express C_π in terms of I_{C} instead of Q_{B}

$$\begin{aligned} Q_{\mathrm{B}} &\approx q\frac{n_{\mathrm{BE}}W_{\mathrm{B}}}{2} \\ I_{\mathrm{C}} &\approx qD_{n\mathrm{B}}\frac{n_{\mathrm{BE}}}{W_{\mathrm{B}}} \\ \Rightarrow \frac{Q_{\mathrm{B}}}{I_{\mathrm{C}}} &\approx \frac{W_{\mathrm{B}}^2}{2D_{n\mathrm{B}}} && \textbf{(with a unit of time)} \\ &= \tau_{\mathrm{B}} && \textbf{(base transit time)} \end{aligned}$$

  - Physically, τ_{B} measures how long the carriers are stored in the base, or equivalently, how long it takes for the carriers to move from the emitter to the collector through the base, and its usually in the $\mathbf{10\,ps}$ range
  - Now back to C_π:

$$C_\pi = \frac{Q_{\mathrm{B}}}{V_{\mathrm{th}}} = \tau_{\mathrm{B}}\frac{\mathrm{d}I_{\mathrm{C}}}{\mathrm{d}V_{\mathrm{BE}}} = \tau_{\mathrm{B}}\frac{I_{\mathrm{C}}}{V_{\mathrm{th}}}$$

  - The value $\frac{\mathrm{d}I_{\mathrm{C}}}{\mathrm{d}V_{\mathrm{BE}}}$ measures the change in the collector current with respect to the base-emitter voltage, and is called the **transconductance g_m** because it has the unit of conductance, but its value is determined by the voltage of other terminals that it is **not connected to**

$$\begin{aligned} g_m &= \frac{I_{\mathrm{C}}}{V_{\mathrm{th}}} \\ C_\pi &= g_m\tau_{\mathrm{B}} \end{aligned}$$

  - Transconductance represents the slope of the linearized characteristics of BJT: $\mathbf{slope} = g_m$

- Between the collector and the emitter, the current is controlled by the signal at the base-emitter junction
  - It can be represented by a current controlled current source

$$I_{\mathrm{C}} = \beta I_{\mathrm{B}}$$

  - Or with a voltage controlled current source

$$I_{\mathrm{C}} = g_m v_{\mathrm{BE}}$$

where a lowercase v is used to represent the small change in voltage

- It should be noted that

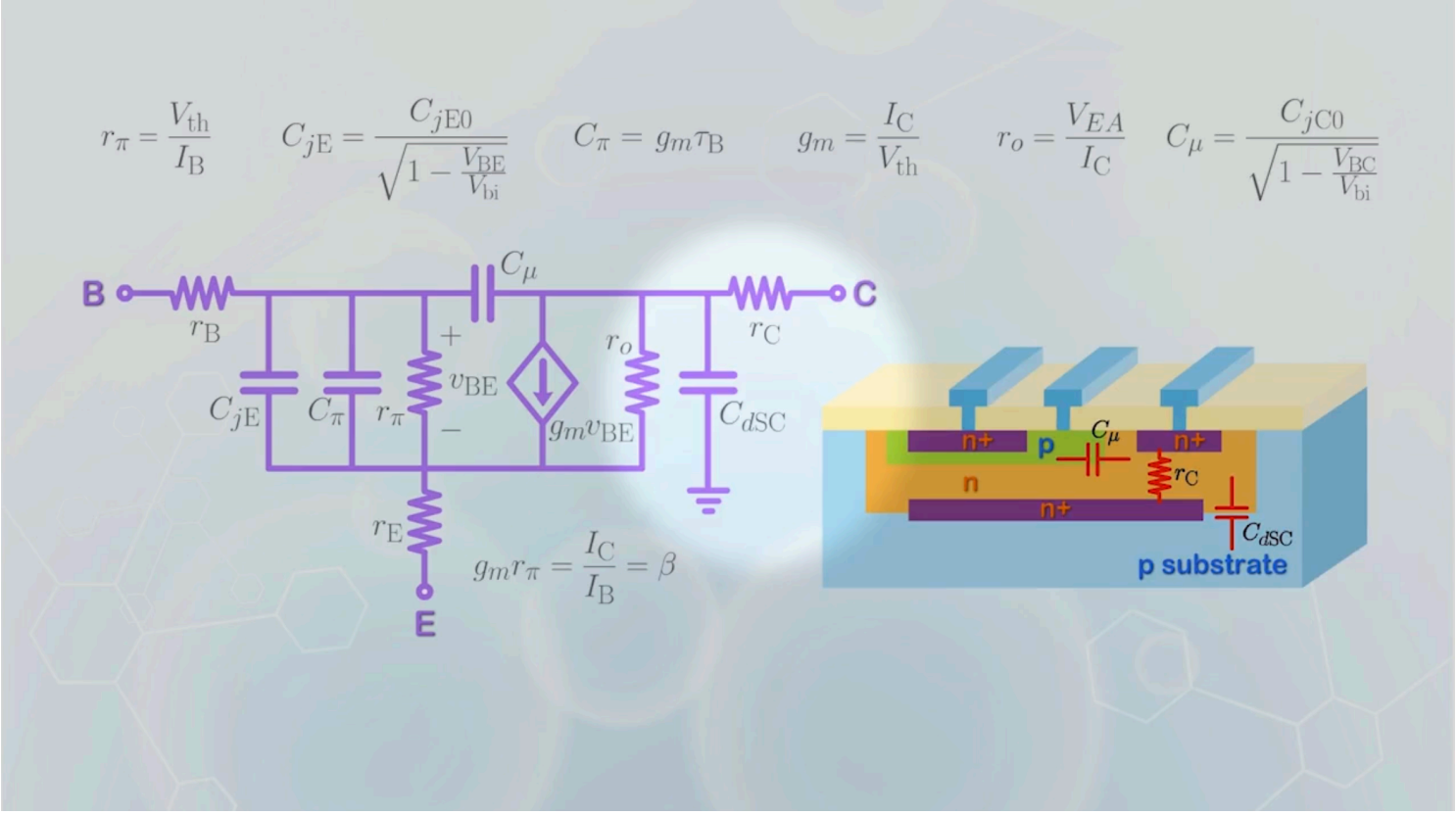
$$\begin{aligned} r_\pi &= \frac{V_{\mathrm{th}}}{I_{\mathrm{B}}} \\ g_m &= \frac{I_{\mathrm{C}}}{V_{\mathrm{th}}} \\ \Rightarrow g_m r_\pi &= \frac{I_{\mathrm{C}}}{I_{\mathrm{B}}} = \beta \end{aligned}$$

- In addition to the current source, that gives the collector current, the Early effect contributes a resistor r_O in parallel, connecting between the collector and emitter

- Between the base and collector, there is a reverse bias junction capacitance C_μ
  - It is the depletion capacitance between the base-collector junction

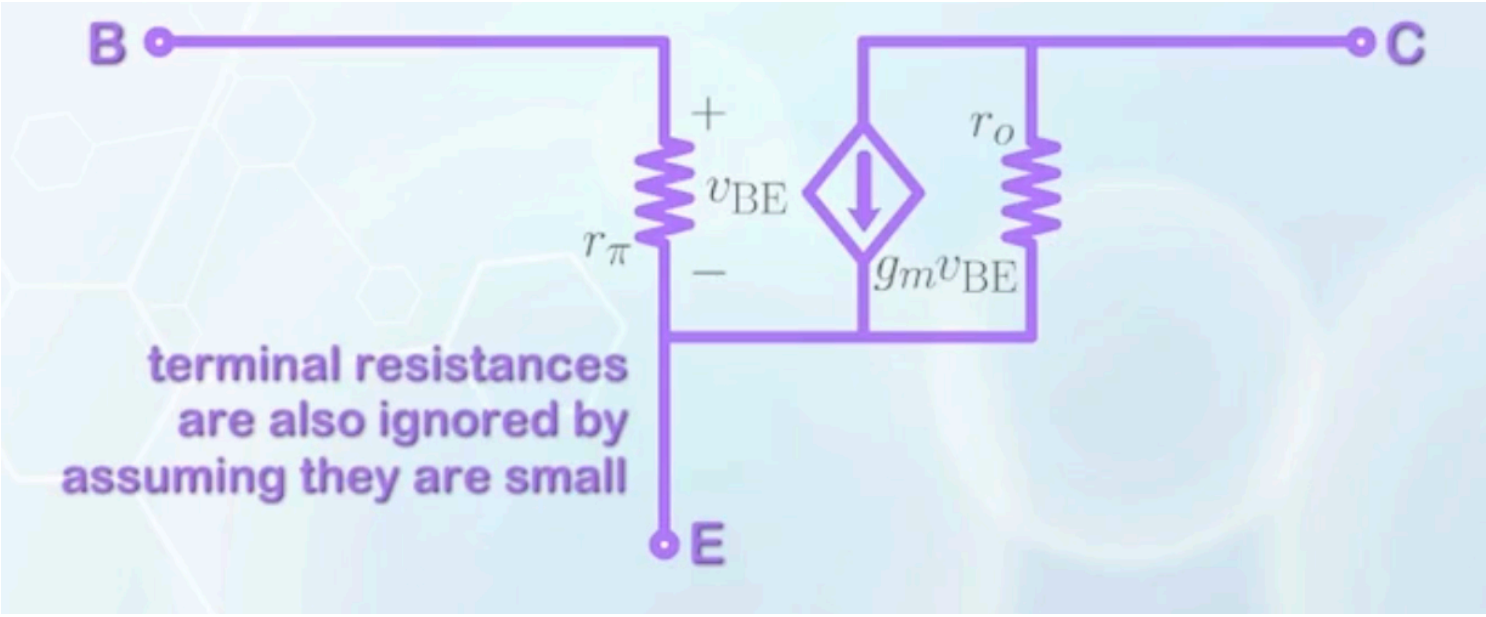
$$C_\mu = \frac{C_{j\mathrm{C0}}}{\sqrt{1 - \frac{V_{\mathrm{BC}}}{V_{\mathrm{bi}}}}}$$

  - C_μ may also include the change of Q_{B} as a function of V_{C} due to the reduction of base neutral region width, which is usually small enough to be ignored
- There is another series resistance for the collector r_{C} between the collector and metal contact
- The collector is surrounded by a P substrate for isolation, introducing a reverse bias junction capacitance $C_{d\mathrm{SC}}$ between the collector and the substrate



Frequency Response

For a transistor used as an amplifier, the output can follow input when the input frequency is low enough. In such case, the capacitors in the small signal model can be considered as open circuits, and removed from the model. The gain of the transistor becomes constant.



When frequency increases, it is possible that before a carrier can reach the collector, the signal has already been reversed. In such case, the collector current does not respond to the change of the input

signal.

The gain of the amplifier **drops** when the input frequency becomes higher than a certain value, due to the **capacitance**, limiting the frequency range a BJT can function properly.

- The speed of a BJT, in terms of the highest frequency it can handle, is usually characterized by the **transit frequency f_T** of the transistor
  - f_T is defined as the frequency at which the current gain β drops to unity (1x)
  - In the $\beta - \log(f)$ graph, β is supposed to be flat over a wide range of low frequency operation, and starts to drop at some high frequency (corner frequency)
  - The drop in beta is linear with the log of frequency
  - Extending the linear drop line to the frequency at which $\beta = 1$, we get the transit frequency f_T
  - At f_T, the collector current becomes the same as the base current, and the amplifier is no longer amplifying the input signal
  - f_T can be calculated using the small signal model in the Fourier domain (C -> $j\omega C$), and $\beta = \frac{I_C}{I_B}$ becomes a function of ω
  - Skipping the math details, the final result is

$$
\begin{aligned}
f_T &= \frac{g_m}{2\pi C_{\text{in}}} \\
&= \frac{g_m}{2\pi(C_{j\text{E}} + C_\mu + C_\pi)} \\
&= \frac{g_m}{2\pi(C_{j\text{E}} + C_\mu + g_m \tau_\text{B})}
\end{aligned}
$$

    C_{in} is the input capacitance, and $C_{j\text{E}}$ and C_μ does not very that much
  - We care about how f_T varies with I_C
    - g_m is related to I_C

$$
g_m = \frac{I_\text{C}}{V_\text{th}}
$$

    - At very small I_C, g_m is very small, thus f_T is more or less proportional to g_m, or I_C, because the denominator is dominated by $C_{j\text{E}} + C_\mu$
    - When $g_m \tau_\text{B}$ becomes comparable, or larger than the junction capacitances, $C_{j\text{E}} + C_\mu$ may be ignored, and

$$
f_T \approx \frac{1}{2\pi\tau_\text{B}}
$$

- The maximum f_T is limited by the base transit time τ_B, which is natural because the collector cannot experience any change in current if the carriers have not reached the collector before the input signal changes, even though the electrons are moving back and forth at the base-emitter junction
- With τ_B in the picosecond range, usual values of f_T is in the range of $50 - 100\,\text{GHz}$
- Further increasing I_C will cause f_T to drop, because β is shown to decrease when BJT enters high-injection region

Structural Optimization

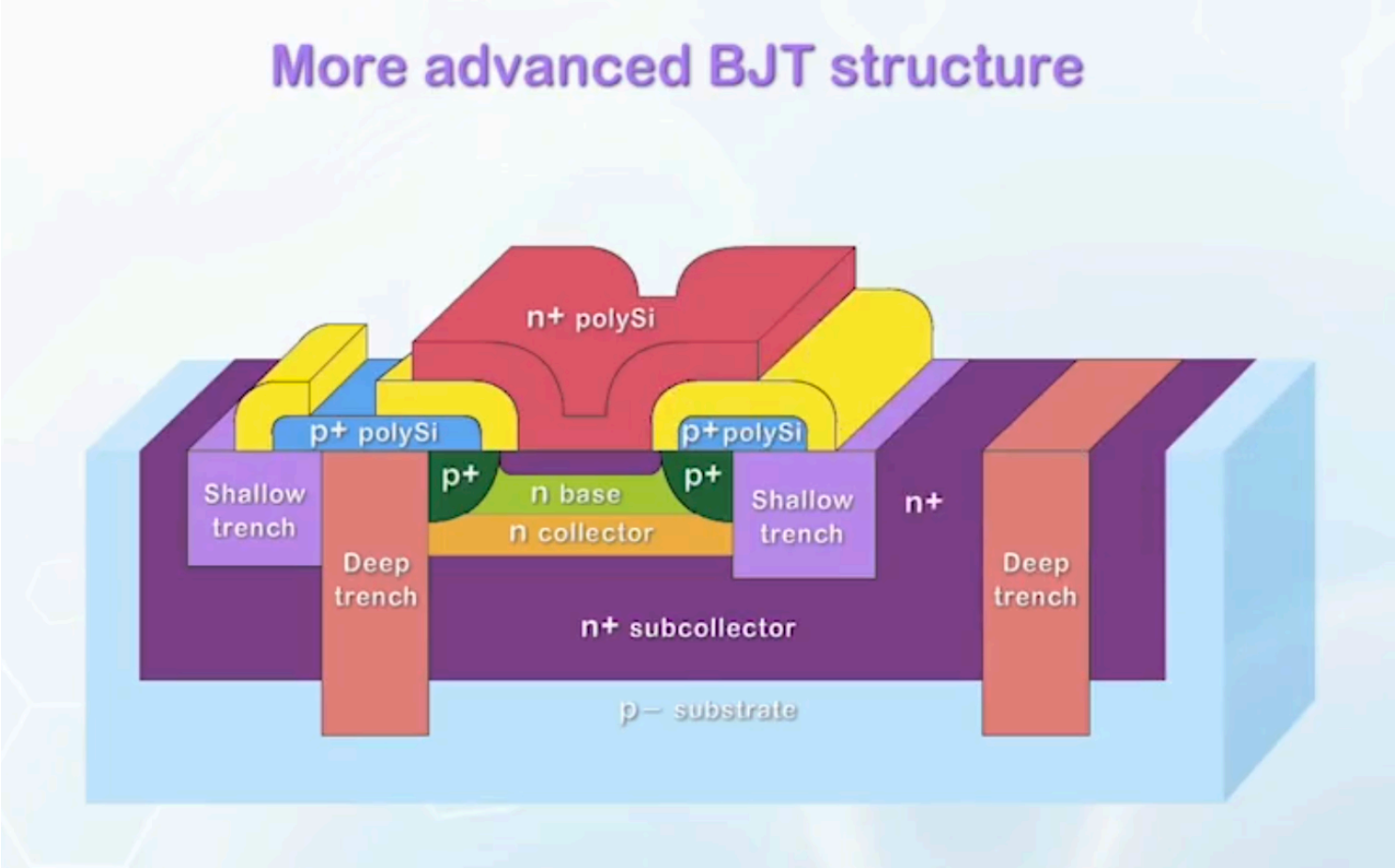
Compared to MOSFET, BJTs are more commonly used to make discrete devices instead of integrated circuits.

The advantages of BJTs include

- High current drive capability
- Sustains high voltage

However, we usually have either NPN or PNP BJTs, but not both, on a single wafer, compared to the complementary MOSFET pairs. Possible complementary BJT structures are complex and not a popular choice.

To improve the performance of a BJT circuit, it is important to reduce the parasitic elements.



- The emitter is extended by adding an N+ polysilicon layer, giving a long emitter length to reduce the diffusion hole current from the base to the emitter
- A P+ region is added to the base outside the transistor region to reduce the base resistance
- The P+ polysilicon layer is used to contact the P+ base, to avoid including contact that consumes large area in the transistor structure
- Trench isolation filled with insulator is used to isolate different regions instead of just a reverse junction, to avoid the parasitic junction capacitance