Principle of Semiconductor Devices Part II: Field Effect Transistors and MOSFETs

For the online version of these notes with better formatting, please visit
https://illusion.blog/notes/principle-of-semi-devices/part-2/1/

1. MOS Capacitor Charge Distribution

About the structure of MOSFET, MOS capacitor, the charge, electric field and potential distribution in accumulation, depletion and inversion modes.

About this series

This is a series of notes based on the video course ***Principle of Semiconductor Devices Part II: Field Effect Transistors and MOSFETs***.

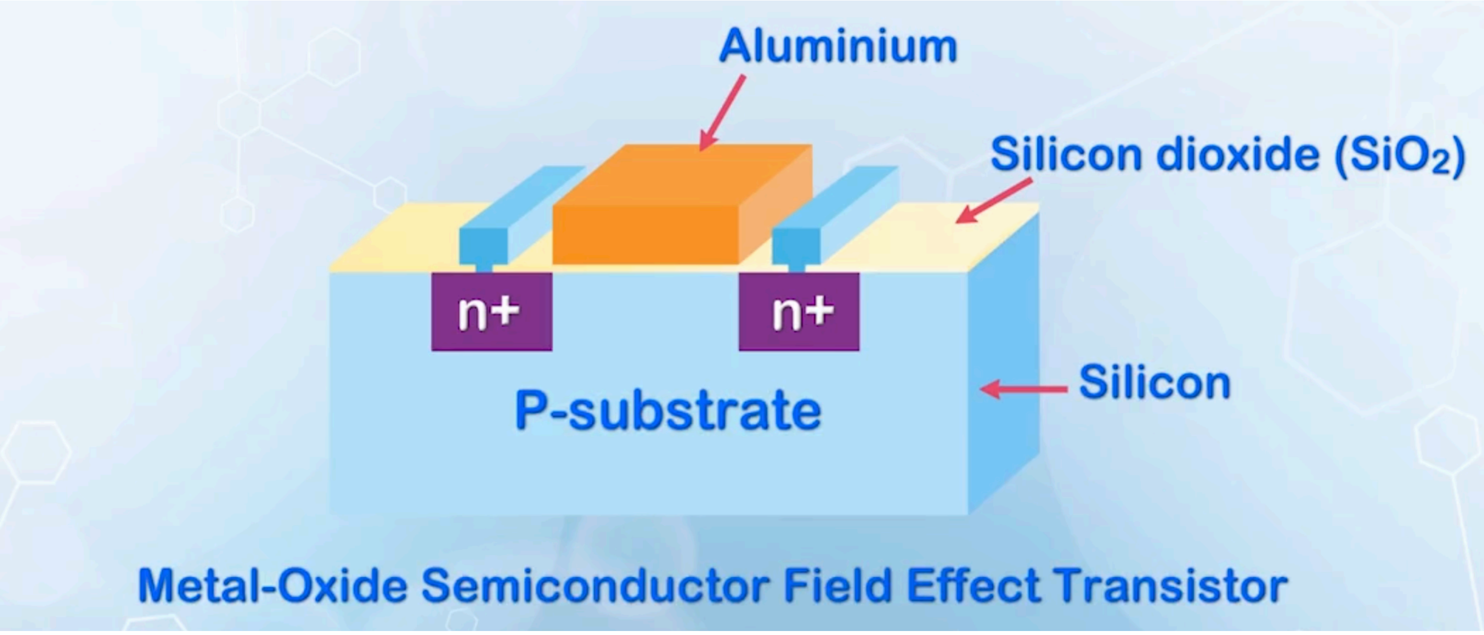
Repost videos with Chinese subtitles is available on bilibili.

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

Structure of a MOSFET

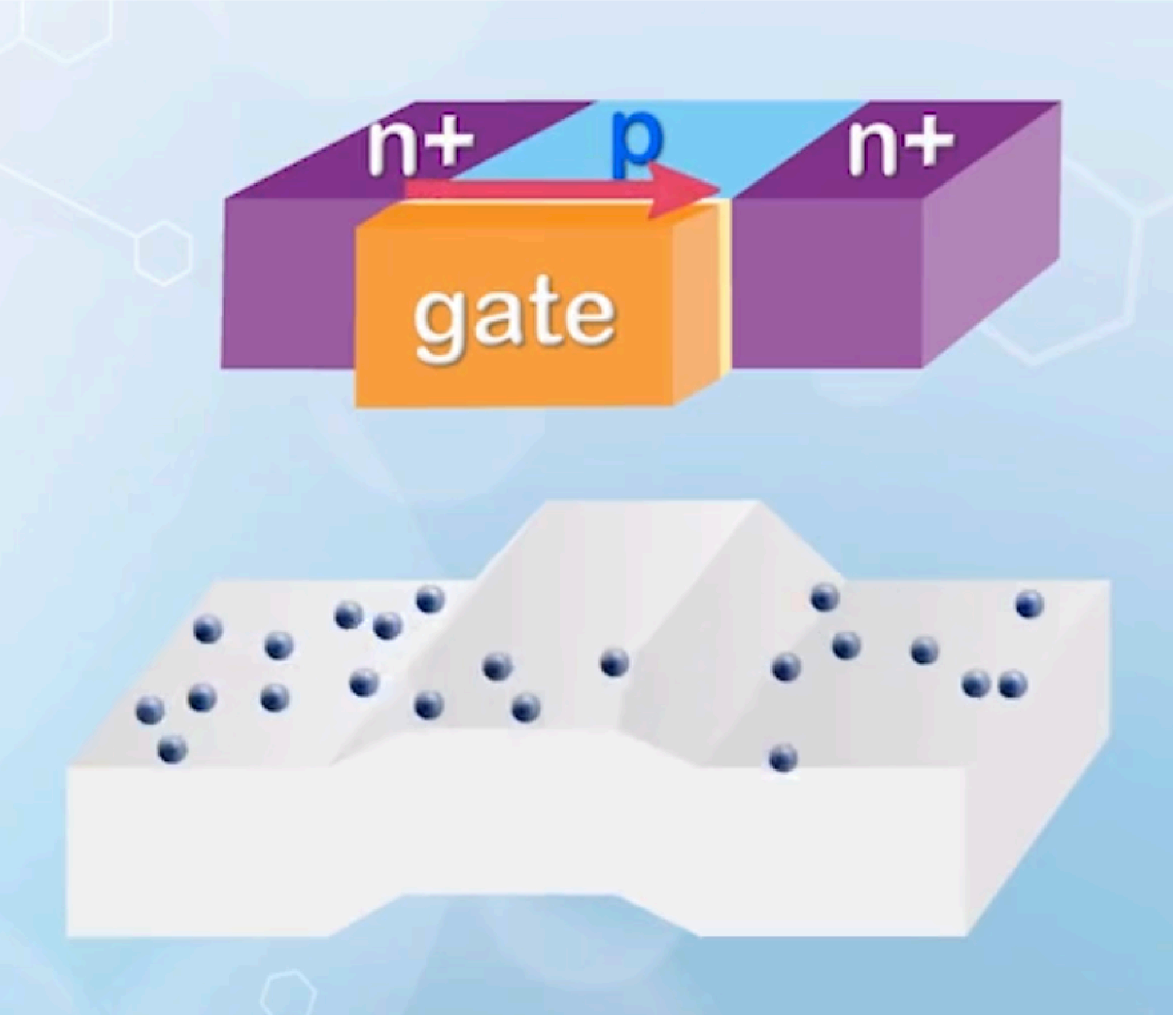
When controlling a BJT, we supply a controlled current flow to the P region (NPN). This is equivalent to a electrical connection to the P region through a resistor.

We can also apply a controlled voltage to the P region through a capacitor, forming a **field effect transistor (FET)**, classified similar to BJTs into N-type and P-type FETs.



Actual materials may differ, but we always assume the dielectric is silicon dioxide (SiO_2), and the substrate is silicon (Si), as they are the most widely used material systems.

The substrate is mainly used for mechanical support, and can be removed for analysis.



The effect of voltage through a capacitor can only penetrate a small distance into the semiconductor, so conduction occurs only in a thin layer near the oxide-semiconductor interface.

With positive voltage applied to one of the N+ terminals relative to the other, electrons will be drained to the terminal.

The terminal with higher voltage is called the **drain**, and the other terminal is called the **source**.

In BJT, the base region is considered as a conductor, and the voltage at two ends of the base region is the same. But in MOSFET, the capacitor drops the voltage from the front to the back, so the voltage must be considered separately. Thus the MOSFET is a **four-terminal device**. The terminal of the controlling capacitor is called the **gate**, and the terminal connected to the opposite end of the gate is called the **substrate** or **body**.

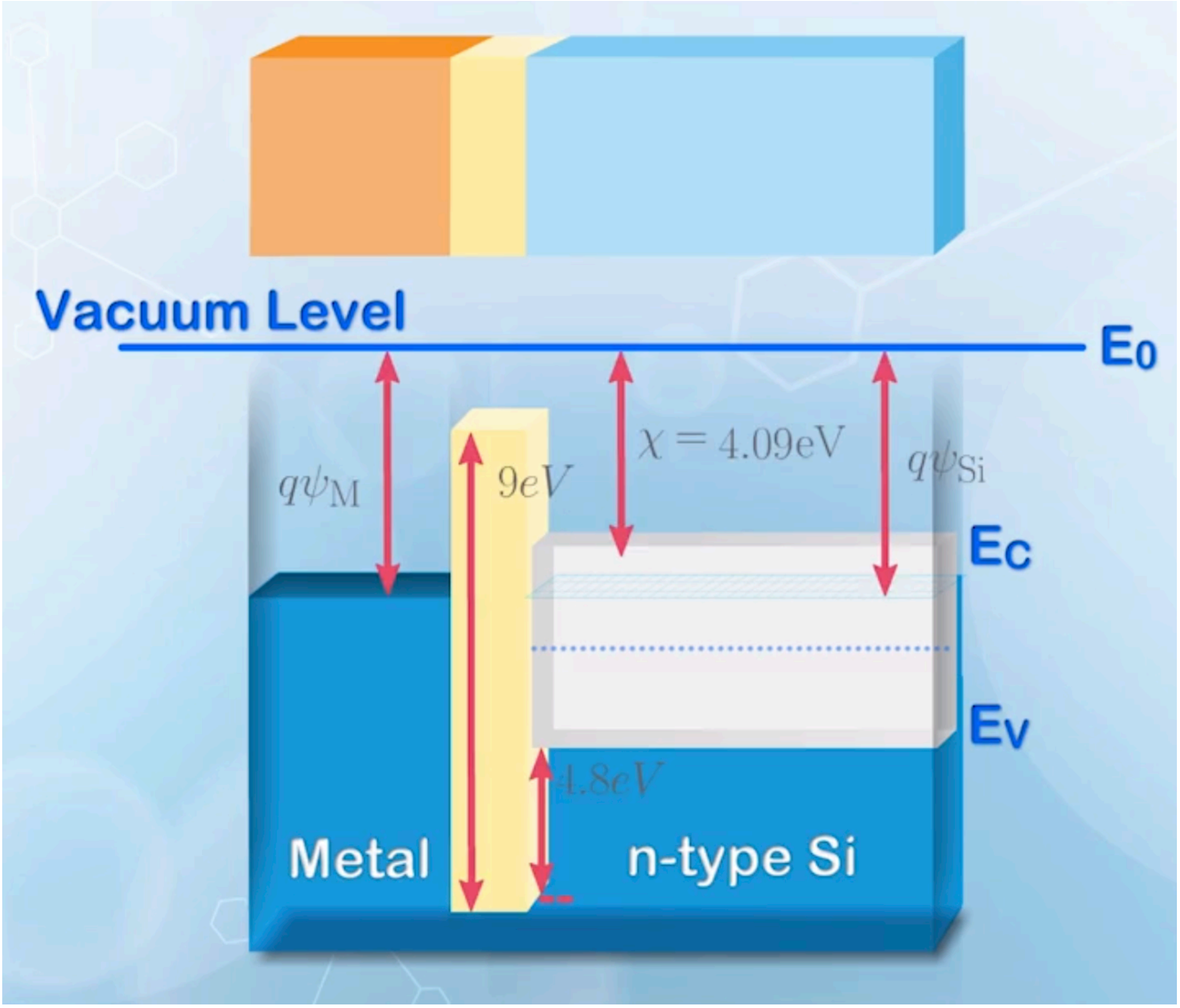
Ignoring the source and drain, the substrate and gate together form a **MOS capacitor**, and it is a **two-terminal device**.

Structure of a MOS Capacitor

The MOS capacitor is formed by sandwiching silicon dioxide between a metal layer and the silicon substrate.

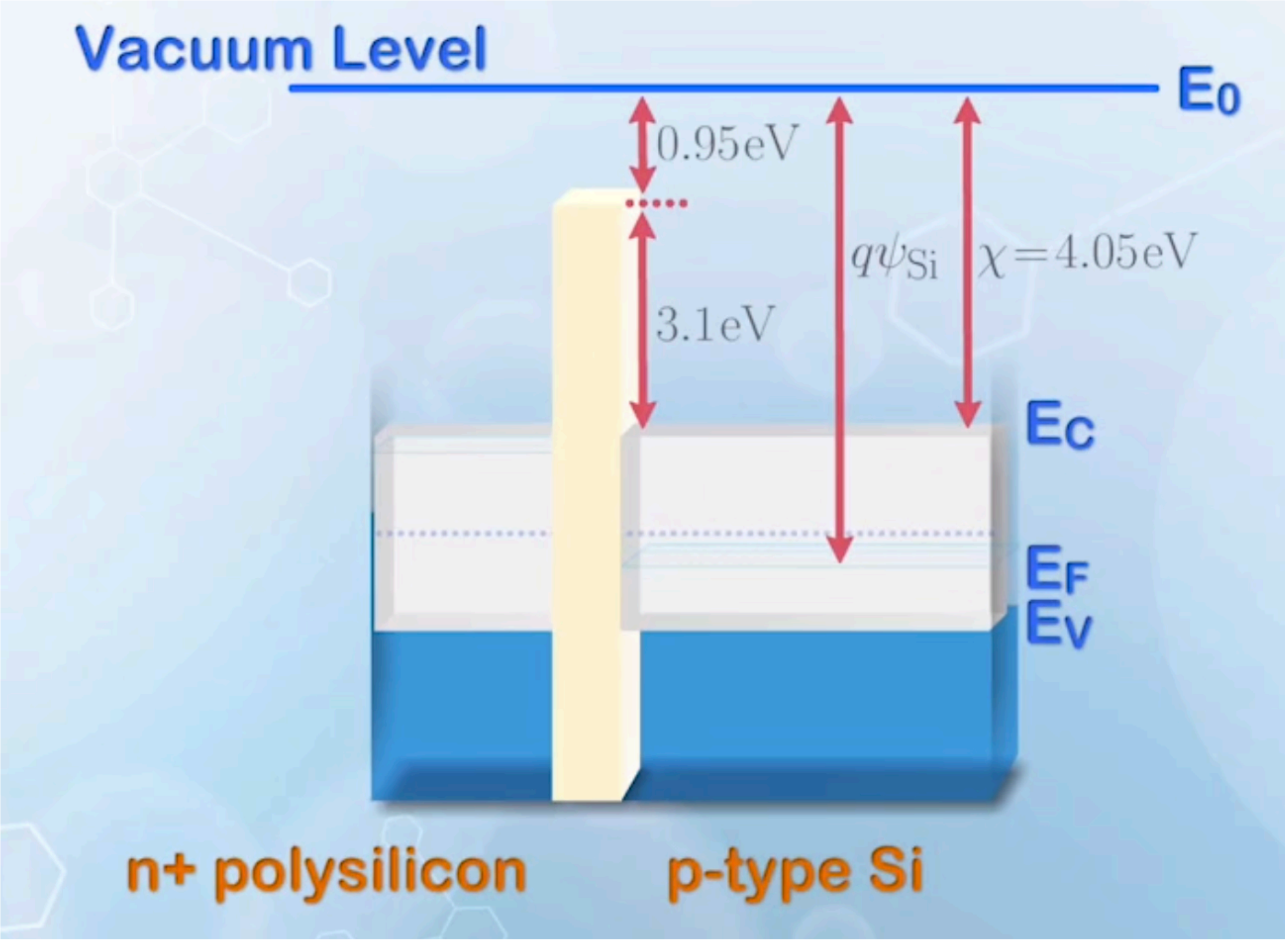
- About the band diagram:
    - **Metal**: the band structure is just represented by its Fermi level
    - **Oxide**: similar to semiconductor, but with a larger band gap
    - Vacuum level is taken as a common reference
    - Metal work function $q\psi_{\mathrm{M}}$: from vacuum level to Fermi level in metal
    - Electron affinity of oxide: from vacuum level to conduction band edge in oxide ($\sim 0.95\,\mathrm{eV}$)
    - Electron affinity of silicon χ: from vacuum level to conduction band edge in silicon ($\chi = 4.09\,\mathrm{eV}$)
    - Work function of silicon $q\psi_{\mathrm{Si}}$: from vacuum level to Fermi level in silicon
    - Assuming the work function of metal is the same as that of silicon, and the substrate is **moderately doped p-type silicon**

    - For n-type silicon:

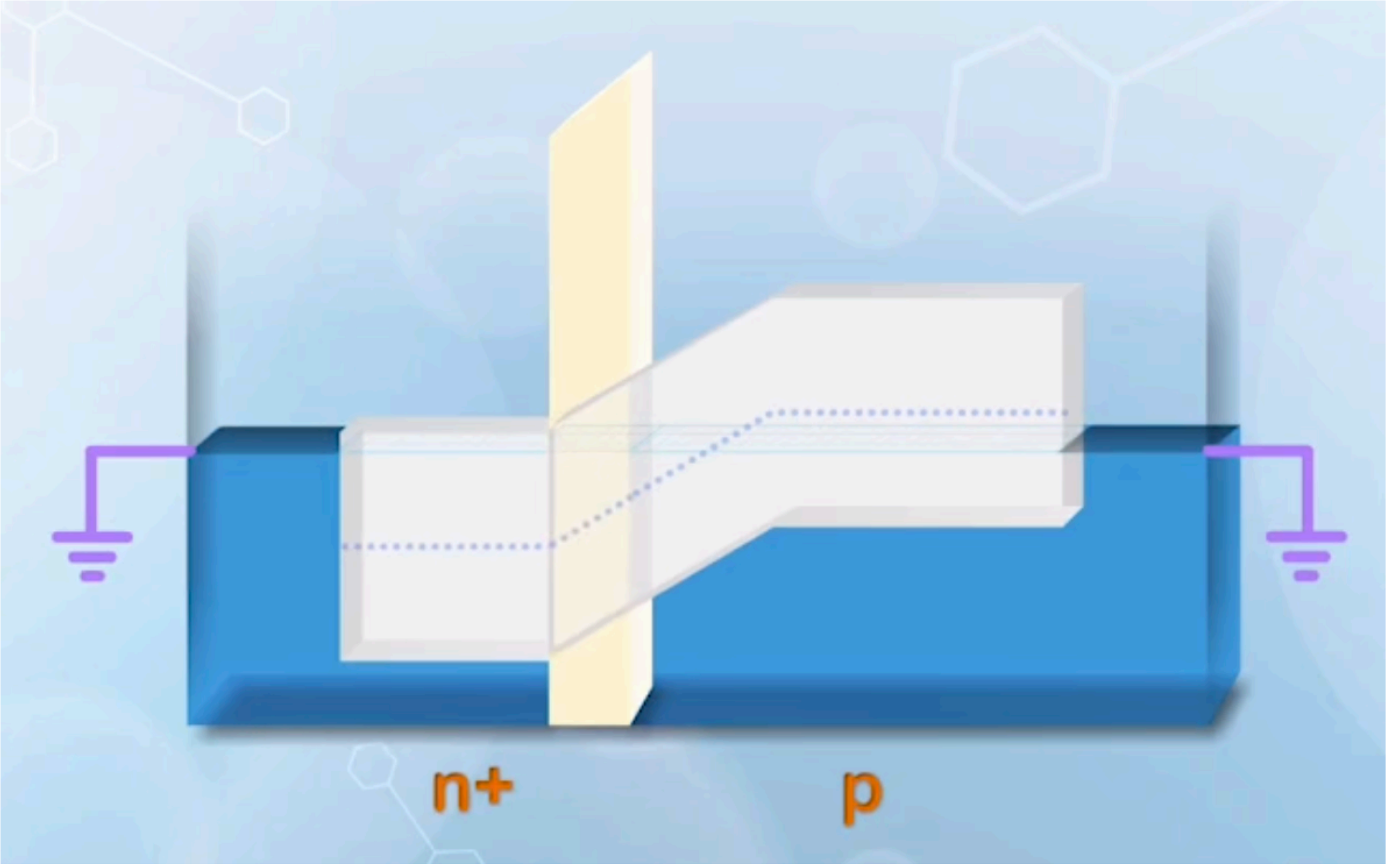


- Heavy doped polysilicon has replaced metal as the gate electrode since around 1970, which can withstand high temperature that simplified fabrication process. The properties of polysilicon have been discussed in part 1
- More recently, the gate electrode has become a very complex multiple material gate stack, but we will be focusing on the MOS capacitor with **polysilicon gate** in this course.

- **Flat band condition**: all energy bands in the structure are flat



- When the two ends are connected to the same potential, or ground, the system is at thermal equilibrium, and the Fermi levels on both sides are aligned
- The band diagram is similar to a PN junction, treating the N+ polysilicon as N+ silicon, except that there is an insulator in the depletion region



- The N+ region is much more heavily doped than the P region, so the depletion region in the N+ side can be ignored, and the band diagram for the gate terminal is considered flat
- We only care the Fermi level for the gate terminal, thus the N+ region can be treated as a metal
- The energy difference between two sides are similar to qV_{bi}, but is dropped across the oxide layer and the silicon substrate
    - **Surface band bending ϕ_{S}**: the potential drop on the substrate side, measures the downward bending of the energy bands at the oxide-semiconductor interface relative to the flat band condition, using ϕ instead of ψ to distinguish the reference of voltage to the ground system between that to the vacuum level
    - The amount of voltage drop across the oxide V_{ox} is measured by the energy difference qV_{ox} at the two ends of the oxide
    - **No external voltage can be measured** between the two ends of the MOS capacitor, and the internal voltage is balanced by the voltage drop at the contacts
    - But **there are charges on the two sides of the oxide layer**
- Two achieve flat band condition, an external voltage V_{G} must be applied to the gate terminal to raise the energy band at the gate
    - This voltage is **negative**
    - **Negative voltage moves the energy bands upward, and positive voltage moves the energy bands downward**

  - It is denoted as V_{FB} (value is **negative**), the **flat band voltage**, equal to the difference in Fermi levels on the two sides of the capacitor in flat band condition, or the V_{bi} of the PN junction with the oxide layer removed
  - For metal gate

$$V_{\mathrm{FB}} = \psi_{\mathrm{M}} - \psi_{\mathrm{Si}}$$

    equal to the built-in potential of the metal-semiconductor contact without the oxide layer

Accumulation Mode

- Same amount of charges with opposite polarity will be accumulated on the two sides of a capacitor when a voltage is applied

- Assume **the gate is metal**

  - The accumulation of charges is achieved by either accumulating or depleting electrons on the metal-insulator interface
  - The electron concentration on the metal side is very high, assuming it is infinite
  - All accumulated or depleted charges only appear in an infinitely thin sheet that the interface, described by a δ function

- On the semiconductor side, there are two ways to supply the required charges:

  - Through the motion of **conduction band electrons**
  - Through the motion of **valence band holes**
  - One of the two will dominate, depending on the applied voltage

- The flat band is used as a reference, because there are no charges on either side of the oxide layer

$$V_{\mathrm{G}} = V_{\mathrm{FB}}$$

- When V_{G} is more negative than V_{FB}, or $V_{\mathrm{G}} < V_{\mathrm{FB}}$

  - The surface band bending is upward, $\phi_{\mathrm{S}} < 0$
  - Negative charges are accumulated at the gate terminal, by accumulating electrons on the metal-oxide interface
  - Same amount of positive charges are accumulated on the silicon side

    - By accumulating holes in the valence band, or driving away electrons to leave excess fixed charge from the atom core of the dopants
    - The substrate is p-type, so there are fewer electrons to drive away, but many holes to accumulate on the semiconductor side
    - **Accumulation** dominates (by accumulating majority carriers)
  - All accumulated charges stays within a very thin layer near the oxide interface, we can assume that they behave like sheet charge with a minimal thickness, resembling a parallel plate capacitor
    - The accumulated charge is

$$\begin{aligned} Q_{\text{acc}} &= C_{ox}(V_{\text{G}} - V_{\text{FB}}) & & \\ C_{ox} &= \frac{\varepsilon_{ox}}{t_{ox}} & & \text{normalized with respect to area} \\ \varepsilon_{ox} &= \varepsilon_0 \varepsilon_{r(ox)} & & \varepsilon_{r(ox)} \approx 3.9 \text{ for SiO}_2 \end{aligned}$$

      where t_{ox} is the oxide thickness
    - Q under flat band is zero, so only a negative voltage beyond V_{FB} is used to accumulate charges
    - A negative sign should be added to indicate the correct polarity of positive charges on the semiconductor side

$$Q = -C_{ox}(V_{\text{G}} - V_{\text{FB}}) \quad \text{for } V_{\text{G}} < V_{\text{FB}}$$

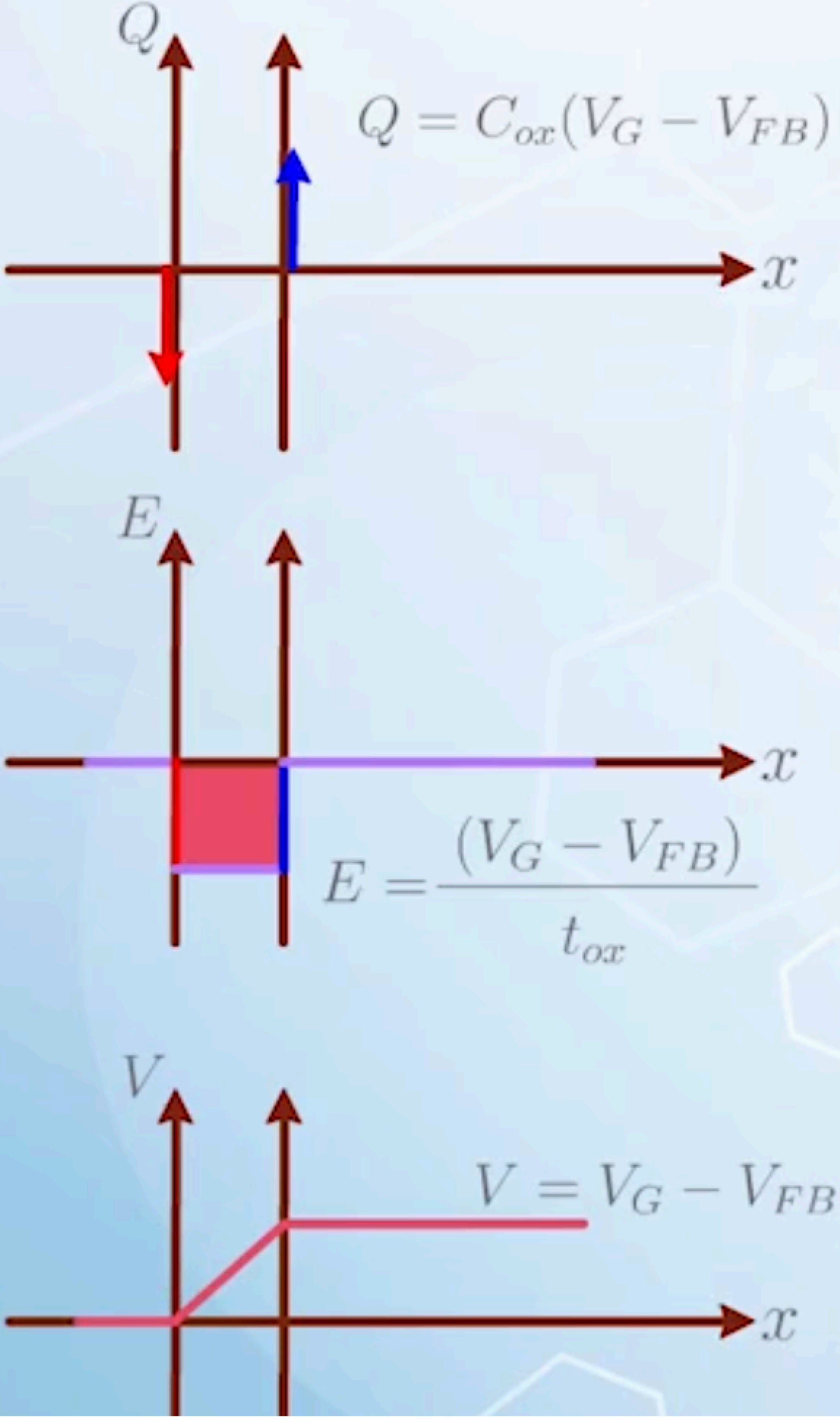
      but we usually just use the magnitude of Q_{acc}, considering the sign of charge separately
  - Integrating the charge density with Poisson's equation will give the electric field

$$\begin{aligned} E_{ox} &= \frac{C_{ox}(V_{\text{G}} - V_{\text{FB}})}{\varepsilon_{ox}} \\ &= \frac{V_{\text{G}} - V_{\text{FB}}}{t_{ox}} \end{aligned}$$

    or just voltage over distance
  - Integrating the electric field will give the potential difference

$$V = V_{\text{G}} - V_{\text{FB}}$$

    corresponding to the shape of the band diagram in the **oxide layer**
  - All the voltage drop in capacitor is **dropped across the oxide**



$$Q = C_{ox}(V_G - V_{FB})$$

$$E = \frac{(V_G - V_{FB})}{t_{ox}}$$

$$V = V_G - V_{FB}$$

Depletion Mode

- Starting from the flat band condition, we apply a slightly more positive voltage to the gate, $V_{\mathrm{G}} > V_{\mathrm{FB}}$
- The energy band on the silicon side will bend downward, giving a $\phi_{\mathrm{S}} > 0$
- It will cause positive charges to accumulate on the gate side, by driving away electrons from the metal-oxide interface, or equivalently accumulating holes in metal
- On the semiconductor side, negative charges must accumulate to balance the positive charges on the gate side
  - By accumulating electrons in the conduction band
  - Or driving away holes in the valence band, leaving the negative core dopant ions in the crystal lattice
  - Again, the substrate is p-type, so there are many holes to drive away, but few electrons to accumulate
    - Driving away holes dominates
    - The p-type substrate is usually **moderately doped**, the density of negative core dopant ions contributed by doping N_A is a few orders of magnitudes lower than the carrier concentration at the metal, or N+ polysilicon gate terminal
    - The charge will spread over some distance, similar to the depletion region of a N+/P junction on the lightly doped P side, instead of a thin sheet in the accumulation mode
    - Additionally, the charge cannot move around, as they come from fixed ionized dopant atoms
    - This is called the **depletion mode**
- We cannot use the parallel plate capacitor model here anymore
  - Assume the depletion region width on the silicon side is x_d
  - Then the accumulated charge is

$$Q_{\mathrm{D}} = qN_A x_d \quad \text{normalized with respect to area}$$

  - Integrating the charge density with Poisson's equation will give the electric field

    - There will be a discontinuity in electric field at the oxide-semiconductor interface, because of the permittivity factor in the equation

      $$E = \frac{Q}{\varepsilon}$$

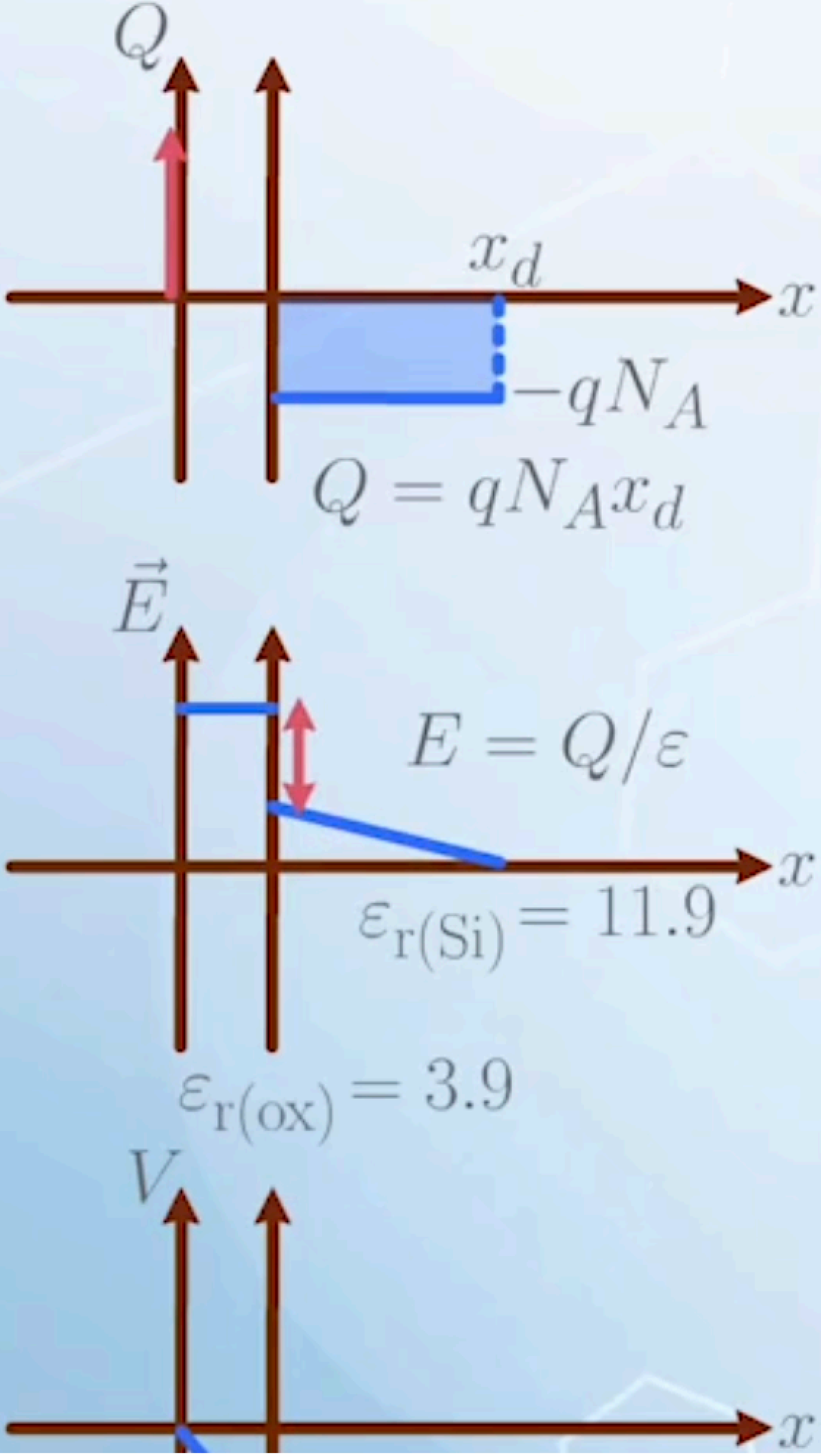
    - At the interface, Q needs to be divided by two different permittivity values

      $$\varepsilon_{r(ox)} = 3.9$$
      $$\varepsilon_{r(\text{Si})} = 11.9$$

      There will be a drop when the electric field leaves the oxide layer and enters the silicon substrate

    - The electric field is similar to a PIN diode mentioned earlier
  - Integrating the electric field will give the voltage difference

- When the discontinuity is removed, it still corresponds to the shape of the band diagram



$Q = qN_A x_d$

$E = Q/\varepsilon$

$\varepsilon_{\mathrm{r(Si)}} = 11.9$

$\varepsilon_{\mathrm{r(ox)}} = 3.9$

- When we continue to increase V_{G} to cause more band bending, the mid-gap energy may become lower than the Fermi level at the oxide-semiconductor interface
  - $\phi_{\mathrm{S}} > \phi_{\mathrm{B}}$, where ϕ_{B} is the difference between the Fermi level and the mid-gap energy in the bulk semiconductor away from the interface

$$\phi_{\mathrm{B}} = \frac{kT}{q} \ln \frac{N_A}{n_i}$$

  - The Fermi level at the interface is closer to the conduction band than the valence band, meaning that the hole concentration is lower than the electron concentration at the interface
  - In this case, accumulating electrons in the conduction band may be more effective than driving away more holes in the valence band, after many holes have already been depleted
  - But the amount of electrons are still small compared to holes at the end of the depletion region, so the depletion mechanism continues, and the depletion region continues to extend
  - The band diagram, electric field and potential extends in a similar way

Inversion Mode

- When we continue to increase V_{G}, the band banding at the interface may cause $\phi_{\mathrm{S}} > 2\phi_{\mathrm{B}}$
- The separation between the conduction band and the Fermi level at the interface will become equal to, or smaller than the separation between the Fermi level and the valence band in the bulk silicon

Why?

- The separation between the conduction band and the Fermi level at the interface is

$$\frac{E_G}{2} + q\phi_{\mathrm{B}} - q\phi_{\mathrm{S}}$$

- The separation between the Fermi level and the valence band in the bulk silicon is

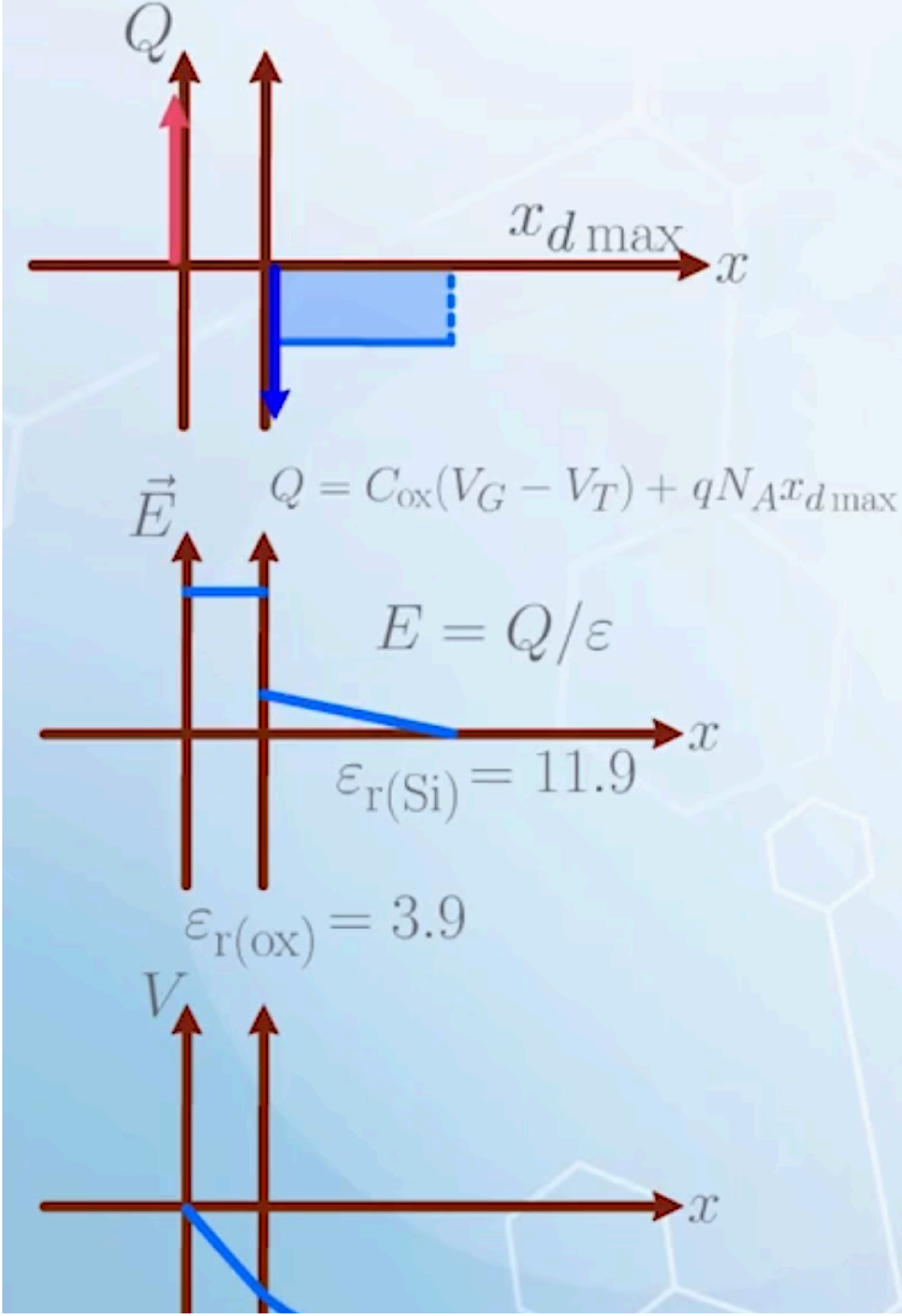
$$\frac{E_G}{2} - q\phi_\mathrm{B}$$

- When $\phi_\mathrm{S} > 2\phi_\mathrm{B}$, the first separation is smaller than the second separation

- The electron concentration at the interface will be higher than the hole concentration anywhere in the bulk
- Now, supplying the required negative charge on tge semiconductor side by accumulating electrons will be more effective
  - The depletion region will stop expanding, and its width reaches a maximum value $x_{d\mathrm{max}}$
- Further increasing V_G will cause gathering of electrons at the oxide-silicon interface
- Electrons can move around compared to the fixed dopant ions, forming a thin layer of negative charges
- This is called the **inversion mode**, as the surface region is now inverted to N-type
- To calculate the charge density on the silicon side, we have to calculate both the dopant ions and the thin sheet of electrons
  - The accumulated electrons will not become significant until ϕ_S exceeds $2\phi_\mathrm{B}$
  - Denote the voltage at this point as V_T, and it is a fixed number
  - All extra voltage beyond V_T will be used to accumulate electrons, similar to the accumulation mode
  - The gathered electrons again behave like a sheet of charges, the equation of parallel plate capacitor can be used again
  - The total charge density on the silicon side is given by

    $$Q_\mathrm{total} = C_{ox}(V_\mathrm{G} - V_\mathrm{T}) + qN_A x_{d\mathrm{max}} \quad \text{for } V_\mathrm{G} > V_\mathrm{T}$$

    the first term is the charge formed by electrons, denoted as Q_inv
- The majority carrier of the P type silicon at the silicon dioxide interface has changed from holes to electrons, thus the name **inversion**, or **strong inversion**
- V_T marks the boundary of V_G between depletion mode and inversion mode
- Again, integrating the charge density with Poisson's equation will give the electric field, and integrating the electric field will give the potential difference, similar to depletion mode



$Q = C_{\text{ox}}(V_G - V_T) + qN_A x_{d\,\max}$

$E = Q/\varepsilon$

$\varepsilon_{\text{r(Si)}} = 11.9$

$\varepsilon_{\text{r(ox)}} = 3.9$

- Note that the discontinuity in electric field at the oxide-semiconductor interface will be larger than that in depletion mode, because of the electron sheet charge, regarded as a δ function, will cause a sudden change after integration

When the Gate is Polysilicon...

The charge at the gate side formed by metal is much simpler, because it can be regarded as sheet charge for both negative and positive charges, without depletion consideration.

When the gate is N+ polysilicon, the doping is usually a few orders of magnitude higher than the substrate doping.

The depletion region in the heavily doped side is much smaller than that in the lightly doped side, so even the depletion charge on the gate side can be considered as sheet charge to simplify the analysis.

In this case, the behavior of the N+ polysilicon can be treated as metal, so we can focus on the analysis on the behavior of charges on the bulk silicon side, which is more important.

2. Calculating MOS Capacitor Charge and Capacitance

About charge and capacitance calculation in an MOS capacitor under accumulation, depletion, and inversion modes, and the dynamic capacitance behavior in inversion mode.

Accumulation and Depletion Charge

- Under accumulation mode, the charge is composed of accumulated holes, and is given by:

$$\begin{aligned} Q_{\text{acc}} &= C_{ox}(V_{\text{G}} - V_{\text{FB}}) \qquad && \text{for } V_{\text{G}} < V_{\text{FB}} \\ C_{ox} &= \frac{\varepsilon_{ox}}{t_{ox}} \\ \varepsilon_{ox} &= \varepsilon_0 \varepsilon_{r(ox)} && \varepsilon_{r(ox)} \approx 3.9 \end{aligned}$$

  - Once the gate oxide thickness is given, C_{ox} is known
  - V_{FB} is similar to the built-in potential in a metal-semiconductor contact or in a PN junction

    - For metal oxide semiconductor

$$V_{\mathrm{FB}} = \psi_{\mathrm{M}} - \psi_{\mathrm{Si}}$$

    - For N+ polysilicon oxide semiconductor, assume the Fermi level of N+ silicon is very close to the conduction band edge

$$-qV_{\mathrm{FB}} = \frac{E_G}{2} + q\phi_{\mathrm{B}}$$
$$V_{\mathrm{FB}} = -(0.55\,\mathrm{V} + \frac{kT}{q}\ln\frac{N_A}{n_i})$$

      and is known once the doping concentration of the P substrate N_A is given

- Under depletion mode, the charge is composed of ionized dopant atoms, and is given by:

$$Q_{\mathrm{D}} = qN_A x_d \quad \text{for } V_{\mathrm{FB}} < V_{\mathrm{G}} < V_{\mathrm{T}}$$

  - To calculate x_d, it is similar to calculating the depletion width in a PN junction
  - Integrating the Poisson equation in the depletion region gives:

$$E(x) = -\frac{qN_A}{\varepsilon_{\mathrm{Si}}}(x - x_0)$$
$$\phi_s = \frac{qN_A}{2\varepsilon_{\mathrm{Si}}}x_d^2$$

  - Once ϕ_s is known, x_d can be calculated as:

$$x_d = \sqrt{\frac{2\varepsilon_{\mathrm{Si}}\phi_s}{qN_A}}$$

  - The calculation of ϕ_s is discussed in the next section
- Under inversion mode, the charge is composed of both mobile electrons at the interface, and the ionized dopant atoms in the depletion region:

$$Q_{\mathrm{total}} = C_{ox}(V_{\mathrm{G}} - V_{\mathrm{T}}) + qN_A x_{d\mathrm{max}} \quad \text{for } V_{\mathrm{G}} > V_{\mathrm{T}}$$

Calculate Surface Band Bending

To solve the depletion charge, we need to find out what ϕ_{S} is.

- When an MOS capacitor is operating under depletion mode, the depletion region in the silicon substrate can be considered as another insulator

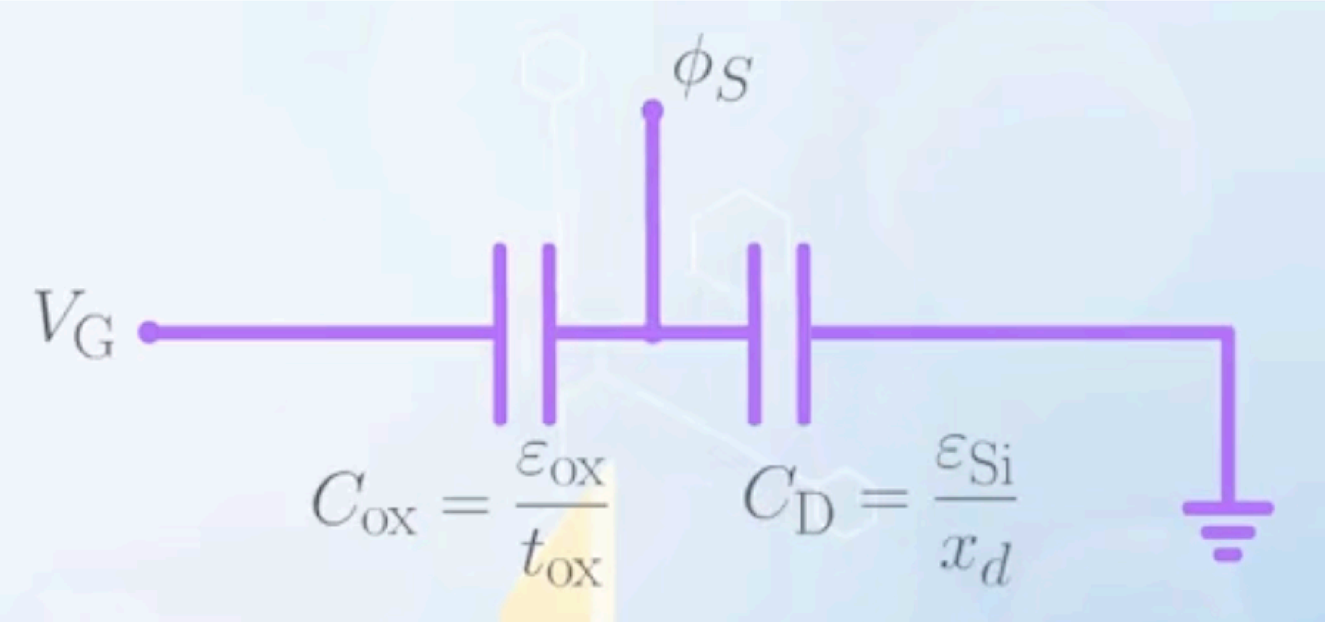
- The MOS capacitor can be considered as two capacitors connected in series
  - One is the oxide capacitor, with normalized capacitance

$$C_{ox} = \frac{\varepsilon_{ox}}{t_{ox}}$$

  - The other one is the silicon depletion capacitor, with normalized capacitance

$$C_{\mathrm{D}} = \frac{\varepsilon_{\mathrm{Si}}}{x_d}$$

  - When V_{G} is applied to the C_{ox} end, and the C_{D} end is grounded, the voltage is divided between the two capacitors according to the capacitance values, and ϕ_{S} is the potential between the two capacitors


ϕ_S
V_G
$C_{OX} = \frac{\varepsilon_{OX}}{t_{OX}}$
$C_D = \frac{\varepsilon_{Si}}{x_d}$

  - In the language of engineers, ϕ_{S} is usually called the **surface potential**
  - ϕ_{S} is given by

$$\phi_{\mathrm{S}} = \frac{C_{ox}}{C_{ox} + C_{\mathrm{D}}} V_{\mathrm{G}}$$

    and we also have

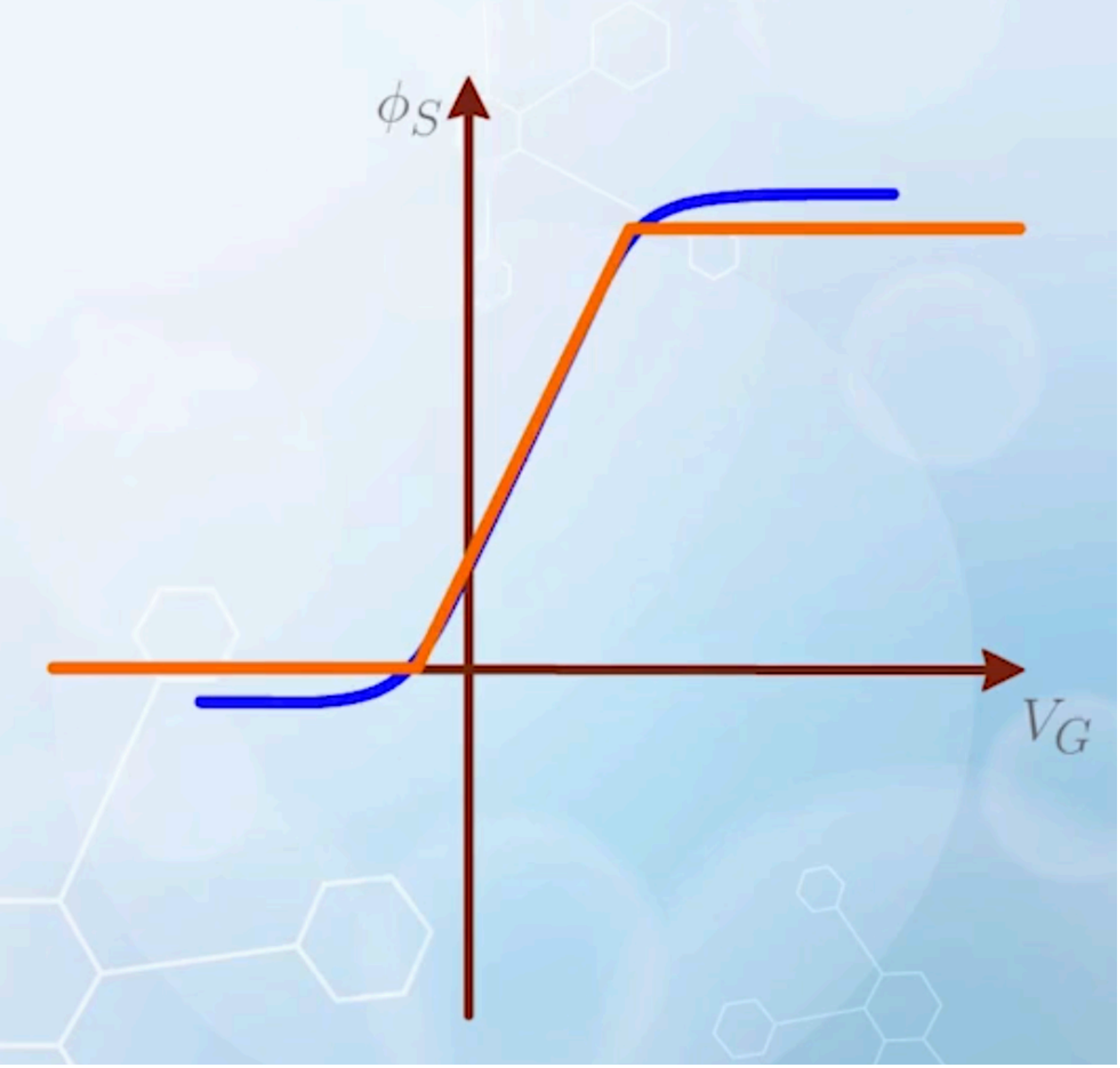
$$C_{\mathrm{D}} = \frac{\varepsilon_{\mathrm{Si}}}{x_d}, \quad x_d = \sqrt{\frac{2\varepsilon_{\mathrm{Si}}\phi_s}{qN_A}}$$

  - With some magical algebraic manipulation, we can obtain an expression of V_{G} as a function of ϕ_{S}

$$V_{\mathrm{GB}} = \phi_{\mathrm{S}} + \frac{\sqrt{2qN_A\varepsilon_{\mathrm{Si}}}}{C_{ox}} \sqrt{V_{\mathrm{th}} e^{-\frac{\phi_{\mathrm{S}}}{V_{\mathrm{th}}}} + \phi_{\mathrm{S}} - V_{\mathrm{th}} + e^{-\frac{2\phi_{\mathrm{B}}}{V_{\mathrm{th}}}} \left(V_{\mathrm{th}} e^{\frac{\phi_{\mathrm{S}}}{V_{\mathrm{th}}}} - \phi_{\mathrm{S}} - V_{\mathrm{th}} \right)}$$

but no close form expression can be obtained for ϕ_{S} as a function of V_{G}, because the given expression is a transcendental function with ϕ_{S} in both the exponential and polynomial terms

- By plotting the above expression, and swapping the horizontal and vertical axes, we can obtain a graph of ϕ_{S} as a function of V_{G}, which can be approximated with three straight lines



- From left to right
  - The first line segment corresponds to the accumulation mode, where $\phi_{\text{S}} \approx 0$, which represents no band bending
  - The second line segment corresponds to the depletion mode
  - The third line segment corresponds to the inversion mode
- The two turning points
  - $A(V_{\text{FB}}, 0)$

        - $B(V_\mathrm{T}, 2\phi_\mathrm{B})$
    - In depletion mode, ϕ_S can be approximated with a straight line between points A and B:

      $$\phi_\mathrm{S} = \frac{1}{n}(V_\mathrm{G} - V_\mathrm{FB})$$

      where $1/n$ is the slope of the graph
    - The slope can also be derived from the series capacitor model:

      $$\begin{aligned} \Delta\phi_\mathrm{S} &= \frac{C_{ox}}{C_{ox} + C_\mathrm{D}} \Delta V_\mathrm{G} \\ \Rightarrow \frac{1}{n} &= \frac{C_{ox}}{C_{ox} + C_\mathrm{D}} \\ \Rightarrow n &= 1 + \frac{C_\mathrm{D}}{C_{ox}} && \text{ideality factor, usually } 1 < n < 2 \end{aligned}$$

    - By approximating the slope to be a constant, we sort of assume C_D to be bias-independent, taking up an average value
    - Ideally, $n = 1$, meaning the gate voltage can directly control ϕ_S
        - Happens when C_{ox} is infinitely large, or C_D is very small
        - It is a very important value for MOSFETs
- Now back to x_d

  $$\begin{aligned} x_d &= \sqrt{\frac{2\varepsilon_\mathrm{Si}\phi_\mathrm{S}}{qN_A}} \\ &= \sqrt{\frac{2\varepsilon_\mathrm{Si}}{qN_A} \frac{(V_\mathrm{G} - V_\mathrm{FB})}{n}} \end{aligned}$$

- Finally the depletion charge can be calculated as:

  $$Q_\mathrm{D} = qN_A x_d = \sqrt{2qN_A\varepsilon_\mathrm{Si}\frac{(V_\mathrm{G} - V_\mathrm{FB})}{n}} \quad \text{for } V_\mathrm{FB} < V_\mathrm{G} < V_\mathrm{T}$$

  It shows that Q_D has a square root dependence on V_G in depletion mode.

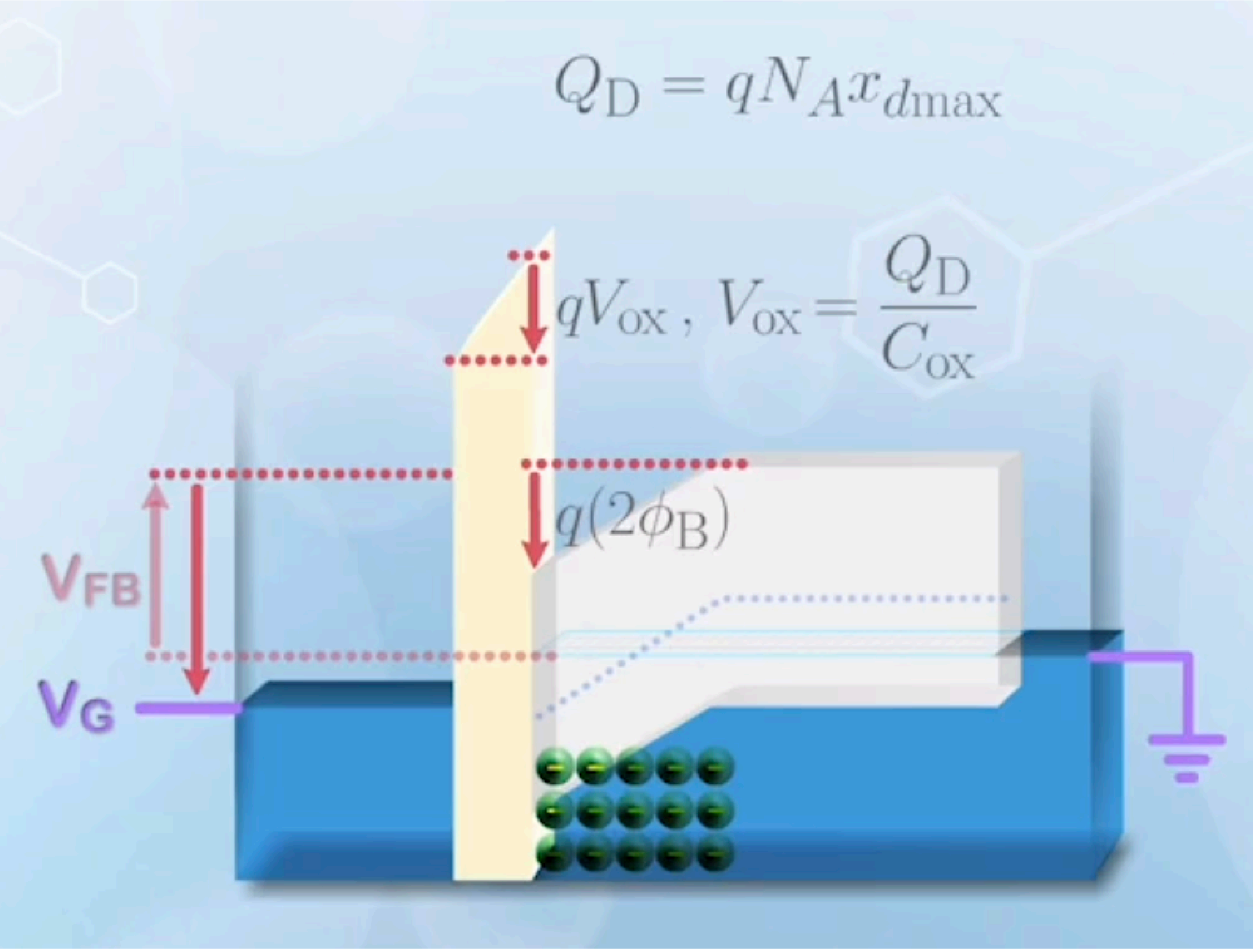
Threshold Voltage and Inversion Charge

In the depletion mode, charge is given by:

$$Q_\mathrm{inv} = C_{ox}(V_\mathrm{G} - V_\mathrm{T}) + qN_A x_{d\mathrm{max}} \quad \text{for } V_\mathrm{G} > V_\mathrm{T}$$

$$x_{d\max} = \sqrt{\frac{2\varepsilon_{\mathrm{Si}}(2\phi_{\mathrm{B}})}{qN_A}}$$
$$\phi_{\mathrm{B}} = \frac{kT}{q}\ln\frac{N_A}{n_i}$$
$$\Rightarrow qN_A x_{d\max} = \sqrt{4qN_A\varepsilon_{\mathrm{Si}}\phi_{\mathrm{B}}}$$

We have to obtain V_{T} to calculate Q_{inv}.


$Q_{\mathrm{D}} = qN_A x_{d\max}$
qV_{ox}, $V_{\mathrm{ox}} = \frac{Q_{\mathrm{D}}}{C_{\mathrm{ox}}}$
$q(2\phi_{\mathrm{B}})$
VFB
VG

1. We give $V_{\mathrm{G}} = V_{\mathrm{FB}}$ to achieve flat band condition
2. We apply additional gate voltage to reach V_{T}
3. The additional gate voltage is dropped across the oxide layer and the silicon depletion region
    - The voltage dropped across the oxide layer is

$$V_{ox} = \frac{Q_{\mathrm{Dmax}}}{C_{ox}} = \frac{\sqrt{4qN_A\varepsilon_{\mathrm{Si}}\phi_{\mathrm{B}}}}{C_{ox}}$$

    - The voltage dropped across the silicon depletion region is

$$\phi_S = 2\phi_B$$

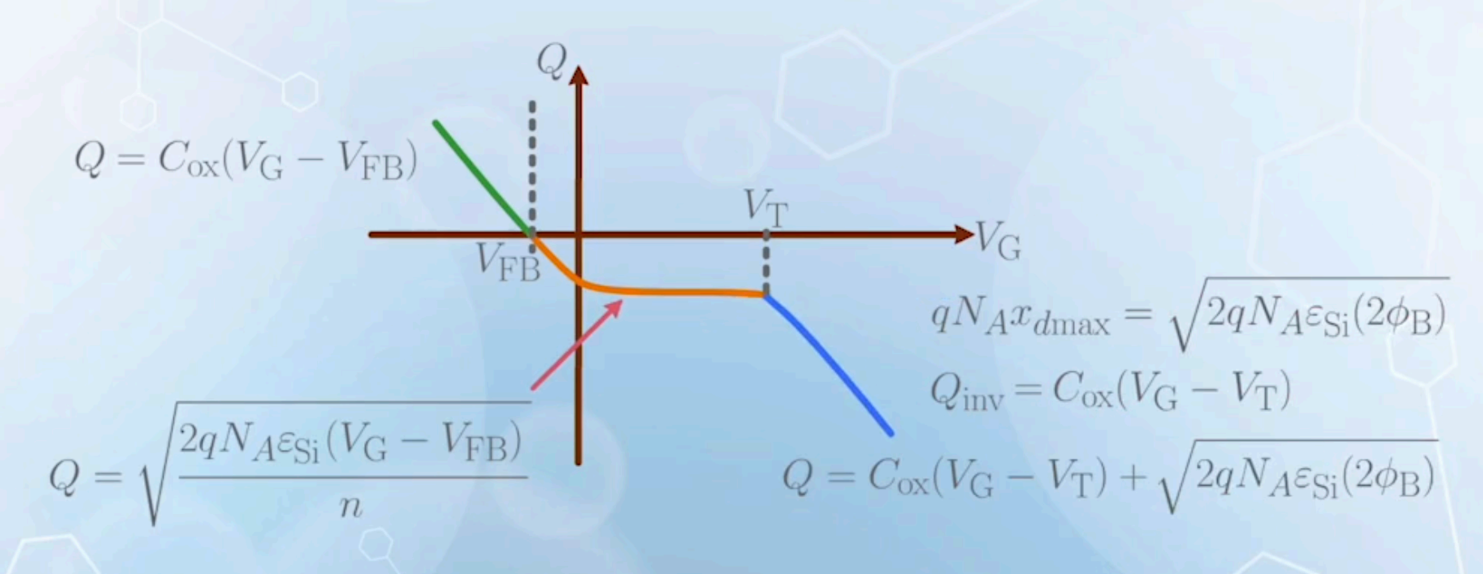
4. We now obtain the threshold voltage as

$$\begin{aligned} V_T &= V_{FB} + V_{ox} + \phi_S \\ &= V_{FB} + \frac{\sqrt{4qN_A\varepsilon_{Si}\phi_B}}{C_{ox}} + 2\phi_B \end{aligned}$$

   Note that V_{FB} is negative for N+ polysilicon gate on P substrate, and the signs of V_{ox} and ϕ_S is always the same

Now we can calculate the total charge in strong inversion mode as a function of V_G.

With the expression of Q in all three modes, we can plot the charge-voltage characteristics of an MOS capacitor (charge polarity following the silicon side):



Or, to avoid the polarity, we can plot the absolute value of charge versus gate voltage.

On the gate side, the same amount of charge with opposite polarity is induced, following the dependence of charges in the silicon substrate. The distribution of the charge is assumed to be a sheet charge at the metal-oxide interface.

Equilibrium Capacitance of a MOS Capacitor

The charge in the MOS capacitor generally has very non-linear behavior, thus we need to use the differential form to calculate the capacitance:

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

To find out the capacitance, the task is to find out where the fluctuating charge δQ appears, when a small varying voltage δV is applied to the gate.

Once done, we can graphically determine the capacitance by treating the location of δQ to be the two terminals of a linear capacitor.

- In accumulation mode

$$Q = C_{ox}(V_{\mathrm{G}} - V_{\mathrm{FB}})$$
$$C = \frac{\mathrm{d}Q}{\mathrm{d}V_{\mathrm{G}}} = C_{ox}$$

  - Graphically, the small signal charge δQ appears at the two sides of the oxide, thus the capacitance is simply C_{ox}
- In depletion mode

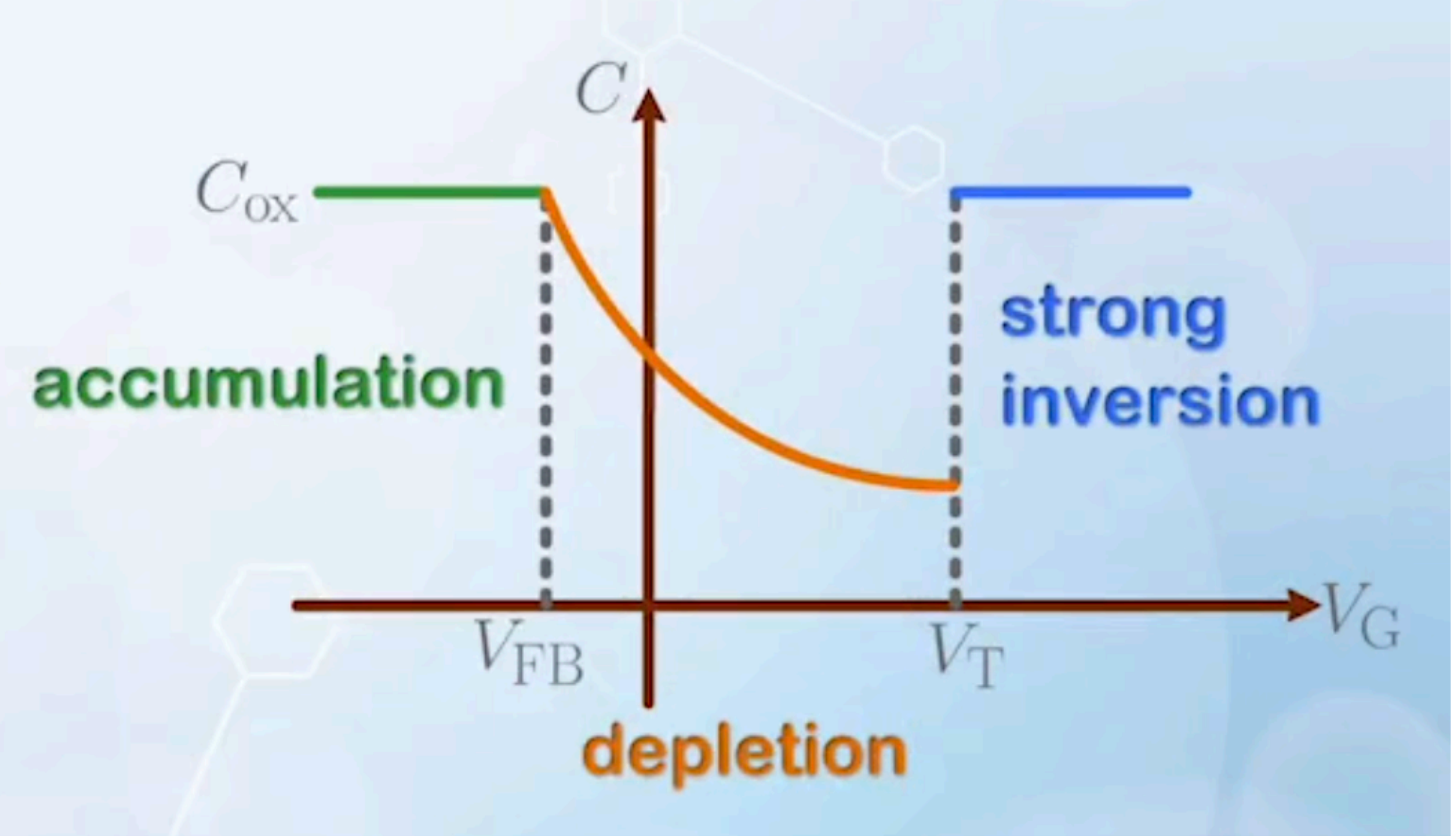
$$Q = \sqrt{2qN_A\varepsilon_{\mathrm{Si}}\frac{(V_{\mathrm{G}} - V_{\mathrm{FB}})}{n}}$$
$$C = \frac{\mathrm{d}Q}{\mathrm{d}V_{\mathrm{G}}} = \sqrt{\frac{2qN_A\varepsilon_{\mathrm{Si}}}{n}}\frac{1}{2\sqrt{V_{\mathrm{G}} - V_{\mathrm{FB}}}}$$

  - Graphically, the small signal charge δQ appears at the metal-oxide interface, and the edge of the depletion region in silicon
    - As the depletion width x_d expands with higher biasing voltage V_{G}, the capacitance decreases
    - The shape is similar to the capacitance of a reverse-biased PN junction, but inversed horizontally because the voltage is applied on the gate side (N+ side)
      - They both come from the fact that the depletion width increases with higher biasing voltage
- In inversion mode

$$Q = C_{ox}(V_{\mathrm{G}} - V_{\mathrm{T}}) + \sqrt{4qN_A\varepsilon_{\mathrm{Si}}\phi_{\mathrm{B}}}$$
$$C = \frac{\mathrm{d}Q}{\mathrm{d}V_{\mathrm{G}}} = C_{ox}$$

  - Graphically, the small signal charge δQ appears at the two sides of the oxide again, thus the capacitance returns to C_{ox}

The capacitance-voltage characteristics of an MOS capacitor can be plotted as:



This is a simplified model. The actual $C-V$ characteristics is smoother, without abrupt transitions

Dynamic Capacitance in the Inversion Region

The $C-V$ curve describes the expected behavior under voltage equilibrium condition. But the inversion capacitance can vary with the measurement setup.

- To measure the MOS capacitance, a DC ramp-up voltage is applied, to define the biasing conditions on the voltage axis of the graph
- A small AC signal is superimposed to the system to measure the capacitance
- The DC ramp-up voltage can be fast or slow, the same for the AC measurement signal frequency, so there are four possible measurement setups
- When both the DC and AC signals are slow, the measured capacitance is the equilibrium capacitance described above
- When the DC signal is slow, but the AC signal is fast
  - The capacitance remains the same after reaching the threshold voltage
- When both the DC and AC signals are fast
  - The capacitance continues to decrease after reaching the threshold voltage
- The condition with fast DC signal but slow AC signal does not make sense, because if the DC signal ramps up before the AC signal can finish a cycle, the AC signal is no longer considered

AC

- The resistance is very high in the conduction band, electrons cannot be supplied from the conduction band. The electrons are supplied by thermal generation
    - When the holes are depleted by V_{G}, recombination rate decreases, but the generation rate remains the same
    - More electrons are generated, and holes are removed through the ground terminal connected to the substrate
    - The current path is though **thermal generations** and **majority carrier motions in the valence band**
- The thermal generation process is relatively slow, in range of milliseconds
- If measurement is performed using frequencies higher than $\mathbf{kHz}$, generation may not be able to catch up
- If electrons cannot be generated fast enough, the depletion region will continue to expand beyond $x_{d\mathrm{max}}$ to supply the charge
- When ramp-up stops, the system will stabilize, electrons will be generated, and depletion width will return to $x_{d\mathrm{max}}$

- When the ramp-up is slow, the band diagram is allowed to stabilize, and depletion width is kept at $x_{d\mathrm{max}}$ for $V_{\mathrm{G}} > V_{\mathrm{T}}$
    - When slow AC signal is applied, electrons can be generated or removed through thermal generation and recombination, the equilibrium capacitance C_{ox} is measured
    - When fast AC signal is applied, there is no time for generation and recombination, thus more holes will be depleted or recovered at a distance of $x_{d\mathrm{max}}$ from the interface, the measure capacitance will be the series combination of C_{ox} and C_{D} with a thickness of $x_{d\mathrm{max}}$
- When the ramp-up is fast, the band diagram cannot stabilize, and depletion width continues to expand beyond $x_{d\mathrm{max}}$ for $V_{\mathrm{G}} > V_{\mathrm{T}}$
    - The measured capacitance will be smaller thant the equilibrium capacitance in the threshold condition
    - The depletion width will increase with larger V_{G}, leading to smaller capacitance
    - Also called the **deep depletion** mode
- What is **high** frequency?
    - It depends on the thermal generation rate
    - For a good silicon crystal, it is in the $\mathbf{kHz}$ range

    - Compared with modern electronics operating in MHz and GHz range, it is considered very slow
    - Additionally, we seldom do measurements at such low frequencies, as the result may be easily disrupted by physical noises such as vibration
    - Most measurements are done at a frequency of $100\,\mathrm{kHz}$ or higher, thus the most commonly observed capacitance is usually the case with slow DC ramp-up and fast AC measurement signal
    - The low frequency $C-V$ curve is very difficult to obtain in a MOS system with a silicon substrate, unless the substrate has lots of defects to enhance the generation-recombination rate

3. Charge Coupled Device (CCD) and CMOS Active Pixel Sensors

About charge coupled devices (CCDs), including its operation principle, architecture, design considerations, performance, and comparison with CMOS active pixel sensors (APS).

CCD Operation Principle

When using PN junctions to detect optical signals, converting light into an instantaneous current is subject to noise and delay. For applications like digital cameras, the current is usually integrated over a short period of time, and collected charge is used to represent the light intensity. This can be achieved by connecting a capacitor to the PN junction.

$$\Delta V = \frac{1}{C} \int_0^{\Delta t} I_{\text{photo}}$$

The combination of a capacitor and a diode, in a more compact form, is just a MOS capacitor. Operating a MOS capacitor as a device to manipulate charges is called **Charge Coupled Device (CCD)**.

- CCD is just a MOS capacitor operated in **deep depletion** mode
- A fast pulse with $V_{\mathrm{G}} > V_{\mathrm{T}}$ is applied to the gate, electrons cannot be supplied quickly enough, and the depletion region expands beyond $x_{d\max}$

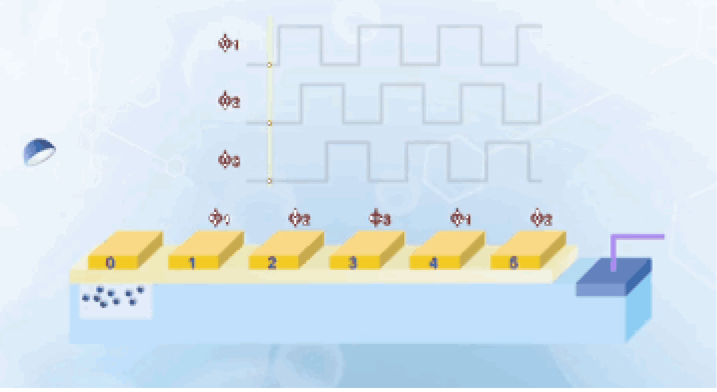
    - The surface potential ϕ_s is given by the potential divider circuit between the **gate oxide capacitor** and the **silicon depletion capacitor**
    - After the pulse ends, electrons will be generated, the depletion region shrinks back to $x_{d\text{max}}$, and ϕ_S drops back to $2\phi_\text{B}$ (all excess potential is dropped across the oxide, similar to accumulation mode)
- **Potential well / quantum well**: the difference between the initial surface potential and the final surface potential
    - Can be considered as a container of charges
    - When empty, the depth is $\phi_S - 2\phi_\text{B}$, where ϕ_S is the surface potential right after the pulse
    - When electrons are generated, they will gradually fill the potential well, decreasing the clearance depth
    - When the well is full, ϕ_S returns to $2\phi_\text{B}$, and the amount of charge stored is

      $$Q = C_{ox}(V_\text{G} - V_\text{T})$$

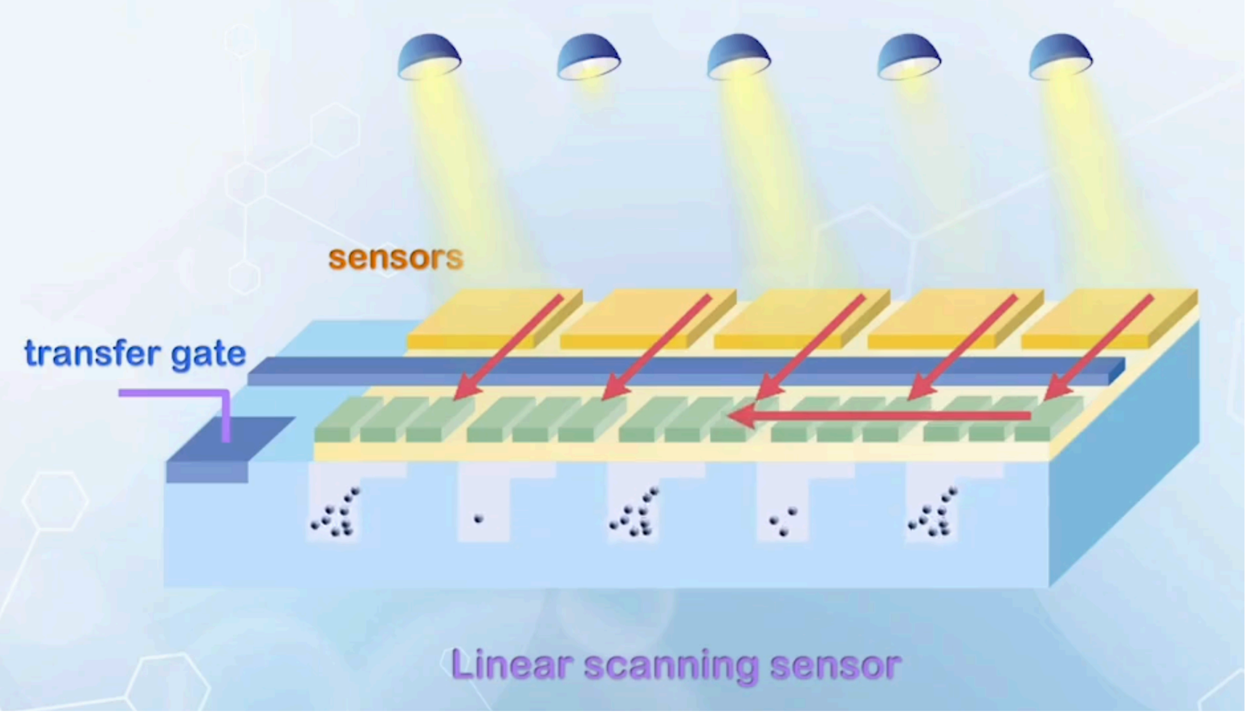
      or the equilibrium value of inversion electrons charge for a given V_G
- A short period of time after the pulse, say $20\,\mu\text{s}$, the potential well can be considered empty, as the generation rate is in the ms range
    - In this period, if a light is shone to the capacitor, electron-hole pairs will be generated, and electrons will be collected in the potential well by the electric field, and holes will be driven away to the ground at the body contact
    - Amount of electrons collected will be proportional to the light intensity and exposure time
    - By measuring the charge, we can find out the intensity of the light
    - This charge has to be measured quickly in μs range, so that the thermal generation process does not add any generated electrons to it

CCD Image Sensor Architecture

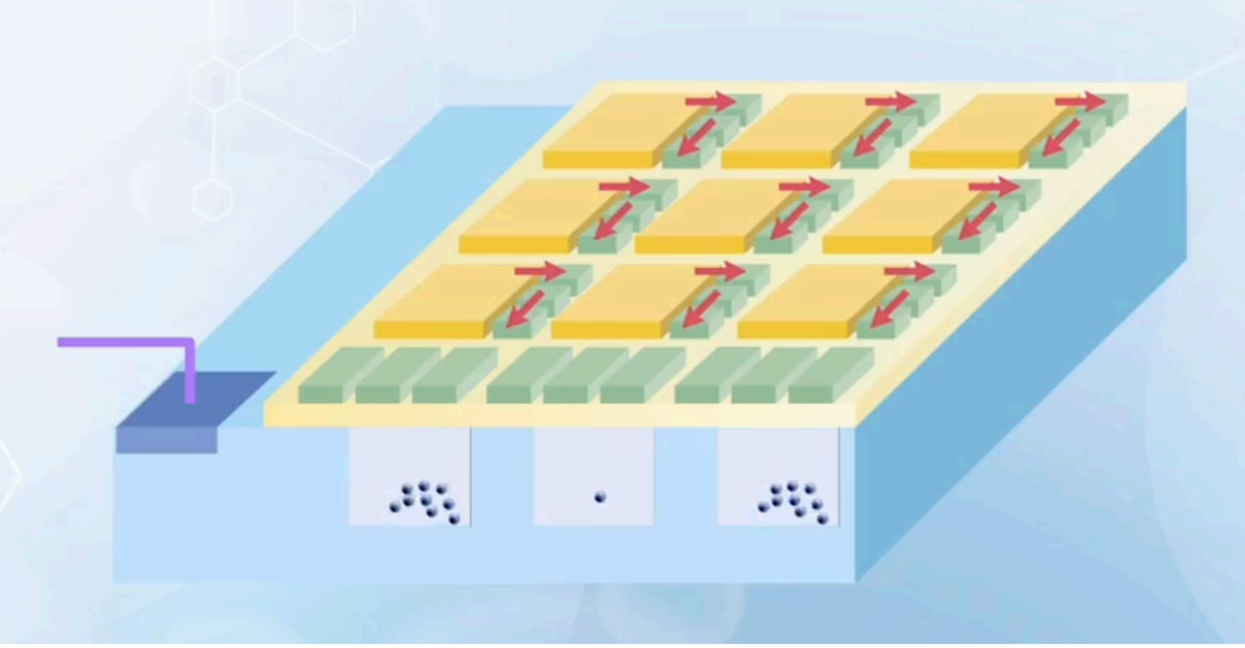
- Generated charges have to be transported to the edge of the sensor array to be processed
    - This can be achieved by putting multiple CCDs side by side
    - By applying some specific voltage patterns to the gates, they can be used to transport charges from one side to the other
        - Consider CCD on the leftmost side as the light sensor (C0), and the rest covered by metal to block light

        - To sense a optical light, a voltage is applied to C0 to create a quantum well
        - Once the sampling period is over, the light is cut off
        - To transfer the charge to the next CCD (C1), a higher voltage V^+ is applied to the gate of C1, creating a deeper quantum well
        - The voltage on C0 is reduces at the same time, forcing the electrons from the C0 to enter C1 (Or that the more positive voltage on C1 attracts the electrons from C0)
        - After all electrons are transferred to C1, and voltage on C0 reduce to 0, the same V^+ is applied to C2, so that the electrons will be shared by C1 and C2
        - The by reducing the voltage on C1 to 0, all electrons will be forced to C2
        - The process is repeated until the charges reach the edge of the array
        - The most common way to implement this is with a three-phase clock
        - 

    - Once the electrons are transported to the end of the array, the signal is read out by a charge-to voltage converter
    - From a logical point of view, this is a **shift register**, which shifts the zero or one from the optical sensor, from one end to the other
        - It carries more than just binary 0 and 1, but a continuous range of charge values from empty to the situation when the well is full
    - CCD shift register cells do not need to sense optical signals, thus can be made smaller to save space
        - Higher voltage is needed for the same electron storage capacity without overflowing the well
- Some applications
    - **Linear scanning sensor**
        - Sensors capture a single line of optical image
        - At the end of a sensing period, the signal is transferred to the bottom CCD shift register, then passed to the edge of array for processing



- 2D array is needed to capture 2D image without mechanical scanning
  - Optical signals capture by the sensor are first transferred to vertical shift registers, then to horizontal shift registers at the edge of the array for processing



- To construct a colored digital camera, two more features are needed
  - **Micro lens**: added on the top of each pixel to focus light onto the sensor
  - **Color filters**: RGB filters forming a Bayer pattern, only one color is sampled at each pixel. For each color, missing pixels are calculated by some weighing functions based on neighboring pixels, instead of direct measurements
    - Real resolution of a CCD is around one third of the image stored in the file

CCD Design and Structural Optimization

To optimize the performance of CCD image sensors

- **Deep enough potential well** so that electrons will not overflow under strong light
    - A high gate voltage is needed, which may cause the gate oxide to break down
    - **Thick oxide** has to be used
    - The depth is determined by $\phi_S - 2\phi_\mathrm{B}$
        - ϕ_S is given by

$$\phi_S = \frac{C_{ox}}{C_{ox} + C_\mathrm{D}}(V_\mathrm{G} - V_\mathrm{FB})$$

        - To increase ϕ_S, we can either increase C_{ox} or reduce C_D
        - Increasing C_{ox} means reducing oxide thickness, which is contradictory to the requirement of high breakdown voltage
        - Thus, we have to reduce C_D by **light doping**
        - Another benefit of light doping is that the depletion region can extend deeper, ensuring all optical signals are absorbed

CCD Image Sensor Performance

- Data in a CCD array is read out serially, or we cannot directly access information in a particular pixel, and the readout process is relatively slow
- Closely packed CCD sensors can potentially cause interference between pixels, creating ghost images and aliasing
    - **Ghost images** happen when a CCD camera is exposed to a bright scene, followed by a dark scene
        - Under strong light, some electron-hole pairs may be generated in the neutral region if the depletion region is not deep enough
        - These electron-hole pairs are not separated as in the depletion region, and moves around randomly
        - When the next dark scene is captured, the number of electron hole pair generated is relatively small, with density lower than the stray electrons.
        - Some stray electrons generated from the previous bright scene without recombining may be collected, so that a ghost image from the previous time frame remains in the current frame
    - **Aliasing** happens when a CCD array is exposed to a scene with high spacial contrast

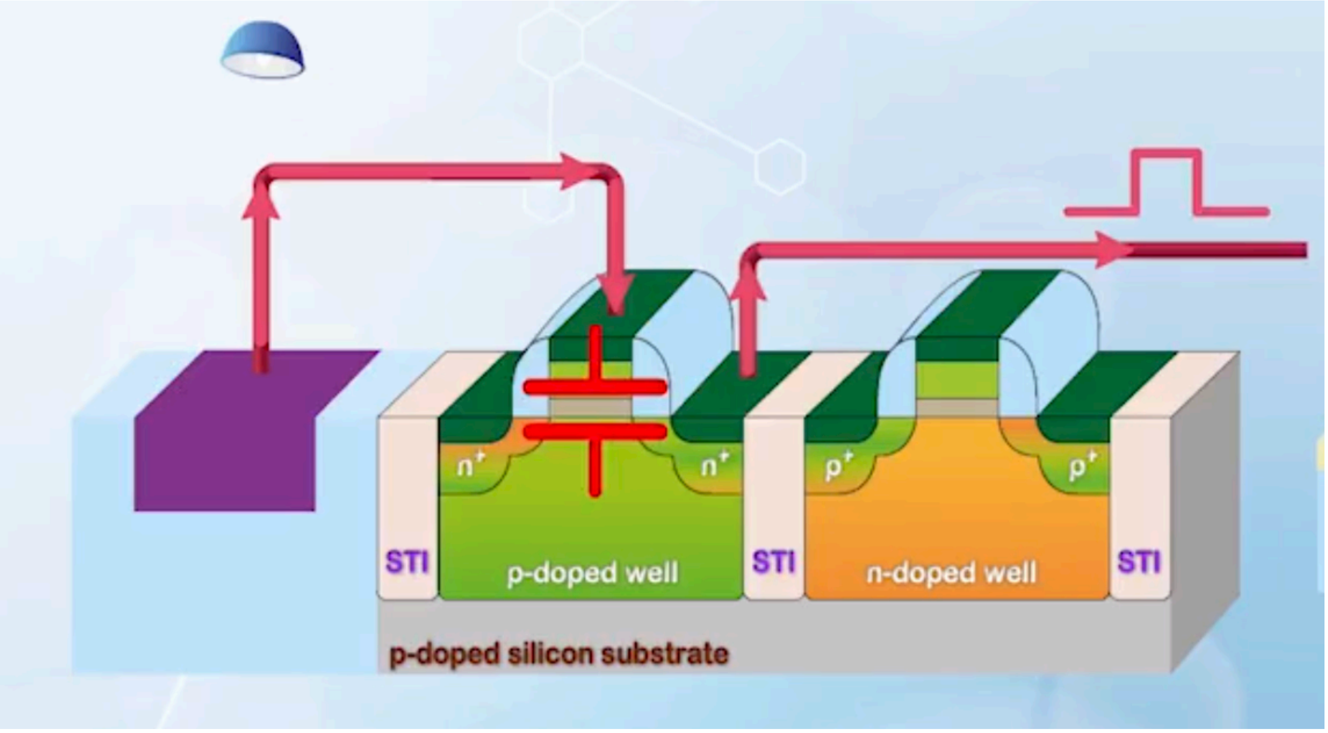
        - At the bright side, some electrons are generated in the neutral region, and accidentally move to the next sensor, which is in the dark side
        - The bright side of a image may flood the dark side, causing aliasing
- Circuits are needed to control the device (e.g. three-phase clock), and encode images
    - All these are usually done by **low voltage** CMOS circuits
    - These devices are optimized with **thin gate oxide** and **heavy substrate doping**
    - But CCD sensors need to use **high voltage** to generate deep potential wells, and uses **thick oxide** and **light substrate doping**
    - Thus, CCD technology and its control circuits are difficult to be optimized together on a single chip
    - A CCD system is usually made up of **three chips**: CCD sensor array, control circuitry, and output formatting circuitry
    - Resulting **high cost** and **large form factor**

CMOS Active Pixel Sensors

More recent digital image applications have migrated to the **CMOS active pixel sensor** technology, or **CMOS APS**.

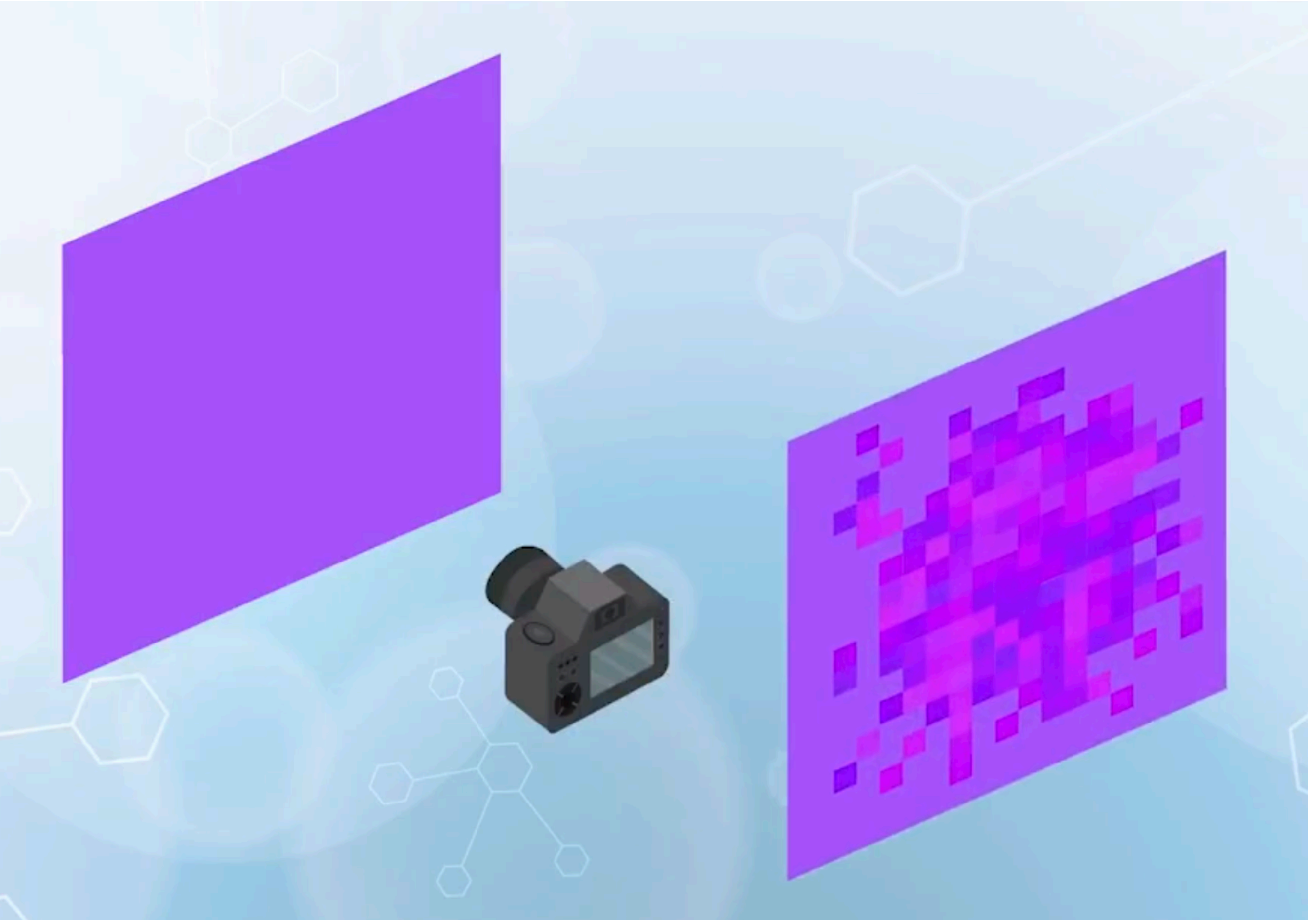
Main problem of CCDs is its difficulty to integrate the image processing and control circuits together with the sensor array, due to different ways of optimization.

- CCD has the signal processing in the **charge domain**, so that the information is being processed in the same silicon substrate before being converted to the required data format
- In CMOS APS, the charge information is **converted to voltage immediately** after being sensed, and passed around through metal wires on top of the chip
    - Does not require a potential well to store the charge, and the charge can be moved immediately to a external capacitor and be converted to voltage



  - Main disadvantage
    - The loss during signal conversion
    - A large wire loading capacitance that the converted voltage has to drive
    - Thus an amplifier is needed at each charge to voltage conversion point
    - CMOS APS need an amplifier at each pixel, while CCDs only need one at the output
    - With continuous transistor size reduction, it is possible to add a few transistors in each pixel to form an amplifier
- Charge is read out at each pixel, so that individual pixel data can be accessed directly instead of serially in CCDs
- Raw image quality of CMOS APS is usually lower than CCDs
  - **Fill factor**: the portion of the sensing area in a pixel
    - CMOS APS has lower fill factor because more non-sensing area is needed to form the amplifier
  - The mismatch of the amplifiers and voltage conversion errors at each pixel may cause **fixed pattern noise** (the image generated due to pixel mismatch when capturing a blank

image with uniform color)



- Recent development in image processing algorithms can effectively compensate the image quality, and all compensation circuitry can be integrated on the same chip, enabling CMOS APS to compete with CCDs in high-end imaging market

4. MOS Capacitor With a Source

About CMOS technology, MOS capacitor with a source and its capacitance characteristics, and body effect.

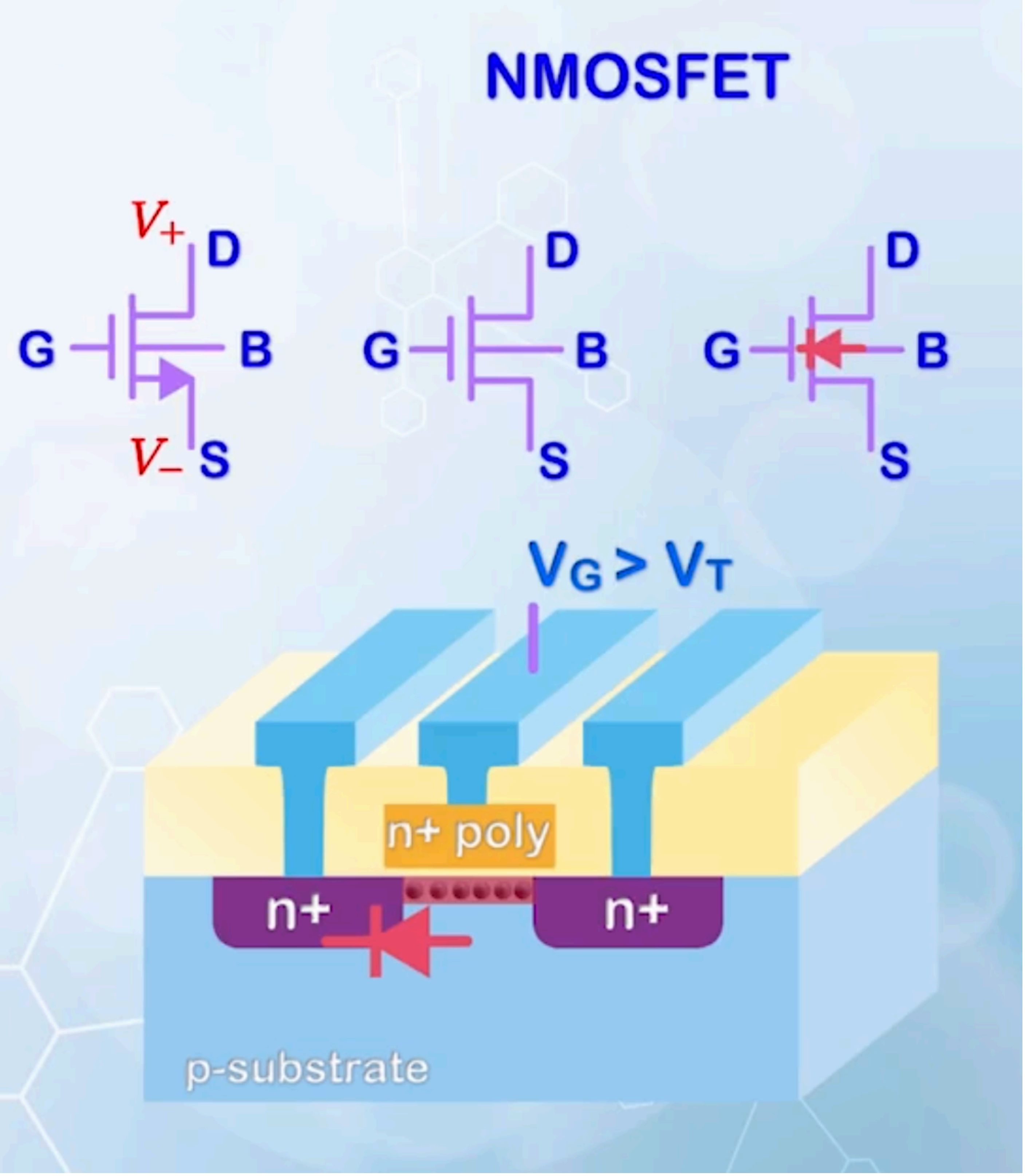
Introduction to CMOS Technology

- In contrast to vertically constructed BJTs, source and drain regions in MOSFETs are place horizontally
    - We do not need to worry about the carrier loss due to recombination in lateral BJTs, as the conduction only takes place in a thin layer near the oxide-silicon interface
    - **Channel**: the conduction layer in the MOSFET

    - It is relatively easy to place N channel devices and P channel devices on the same substrate

- For a **NMOSFET**

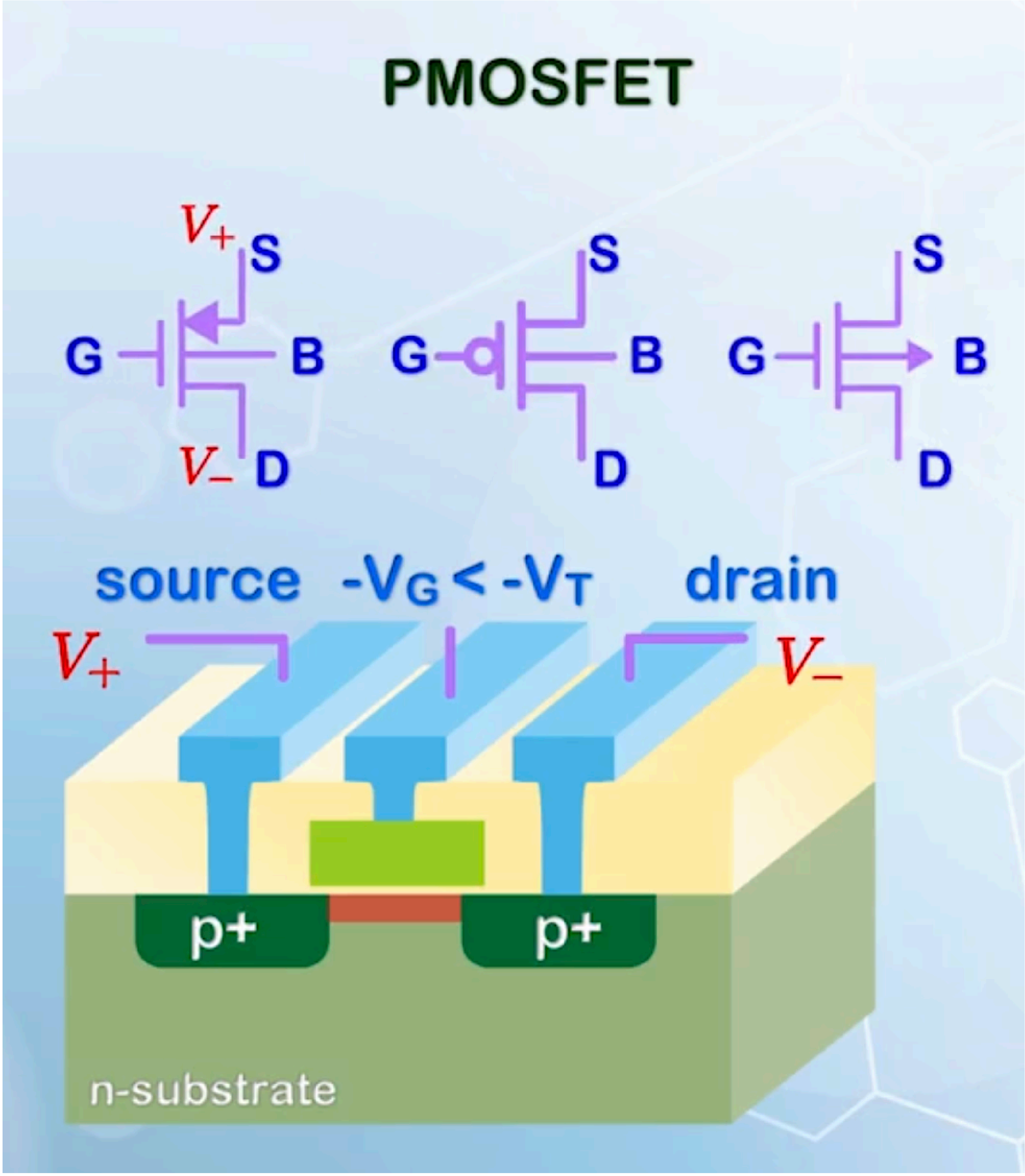
    - N+ silicon for source and drain
    - P type substrate
    - Usually N+ polysilicon gate
    - When $V_{\mathrm{G}} > V_{\mathrm{T}}$, a conductive electron path is formed underneath the oxide substrate interface, and the source and drain are connected
    - N in NMOSFET indicates the conducting carriers are electrons
    - Unlike BJT, MOSFETs have a symmetric structure
        - Drain and source can only be determined by relative voltage applied
        - **Source** is the terminal with the **more negative voltage**
        - If voltage is known, an arrow marks the source, following the polarity of PN junction (**from P pointing to N / the direction of current flow**)
        - Sometimes, source and drain may not be distinguishable, and some alternative symbols may be used
        - When the arrow is placed on the substrate terminal, it points to the gate, as the substrate is P type



- For a **PMOSFET**
  - P+ silicon for source and drain
  - N type substrate
  - When a voltage more negative than the threshold voltage is applied to the gate, a conductive hole path is formed underneath the oxide substrate interface, and the source

  and drain are connected

  - **Source** is the terminal with **more positive voltage** than the drain, as it is where the holes flow from
  - Early PMOSFET gates were also made up of N+ polysilicon due to process simplicity, but it will cause an un-optimized flat band voltage, resulting in undesired large negative threshold voltage
  - Before the use of multiple material gate stack, P+ polysilicon gates were used in most cases to provide symmetry between NMOSFETs and PMOSFETs



- To add a PMOSFET beside an NMOSFET on the same substrate (P type), certain regions must be converted to N type, forming a **well**

- NMOSFETs are faster and can provide higher current drive than PMOSFETs at the same size, NMOSFETs are used more frequently

    - NMOSFETs are usually constructed on the substrate to reduce the overhead of well formation

- More advanced processing may place both NMOSFETs and PMOSFETs in separated wells for better performance optimization, but with some chip area penalty

- To ensure proper operation, the P substrate is usually connected to ground, or the lowest potential in the circuit to prevent the turn-on of any junction between the substrate and the source / drain, as the source / drain voltage may change during operation

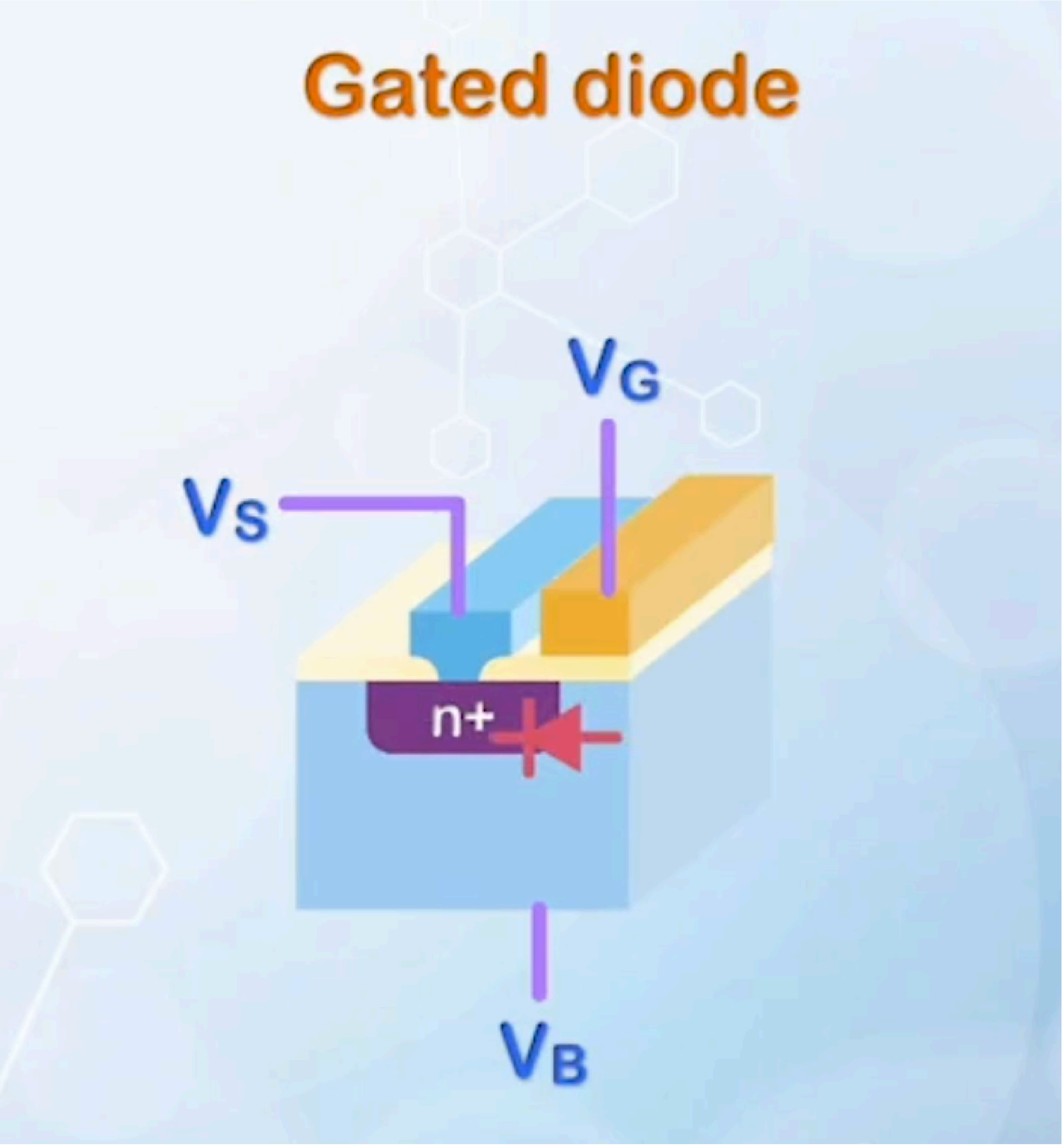
- Similarly, the N well is usually connected to the highest potential in the circuit, usually the power supply, for the same reason

- **Complementary MOSFET Technology (CMOS)**: the technology putting both NMOSFETs and PMOSFETs on the same wafer

    - Advantage: turn on a switch and turn off the other with the same voltage, allowing the implementation of very compact logic gates with very low standby power
    - BJT is popular for discrete devices due to high current drive
    - CMOS is dominant for integrated circuits due to low power consumption and high density

MOS Capacitor With a Source

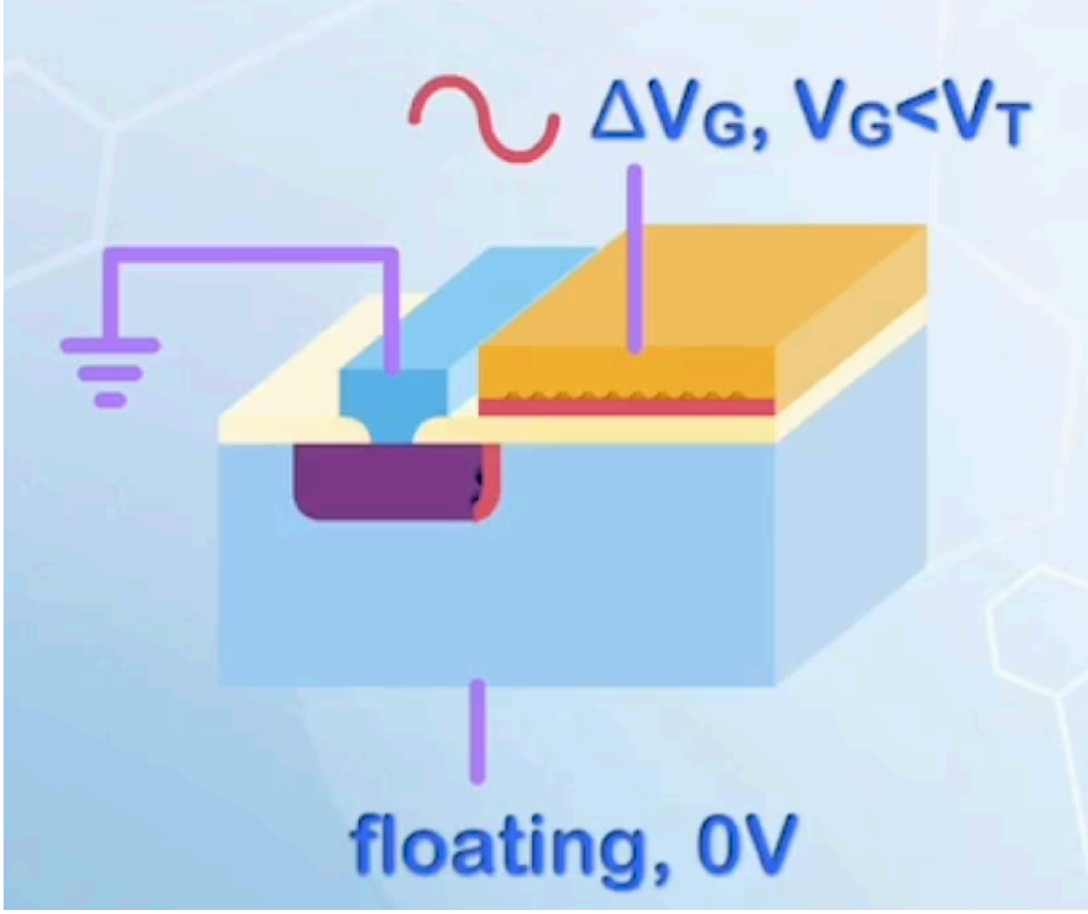
Now we add source and drain regions to the MOS capacitor structure discussed previously. To keep it simple for now, assume the source and the drain are connected together, and both can be labeled as source. By symmetry, it can be considered as a three terminal device, usually called a **gated diode**, as it comprises a gate and a diode.



To measure the capacitance of this structure

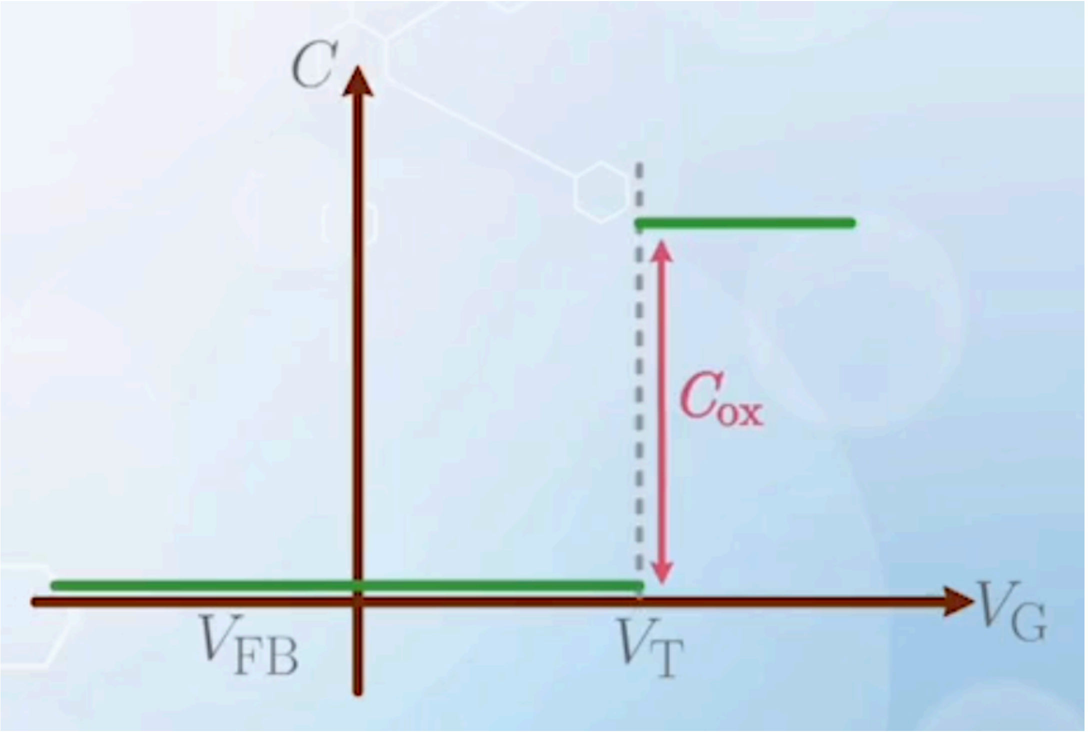
- When **source** is floating, we measure the capacitance between the gate and the substrate, which is the same as the MOS capacitor discussed previously, and the source region has no effect on the capacitance

    - Reminder: generation is slow, under normal measurement conditions, only **high frequency** $C-V$ characteristics can be measured
- When **substrate** is floating, we measure the capacitance between the gate and the source
    - Take source as the reference voltage and assume it to be grounded
    - Substrate will pick up the voltage of source, otherwise, a current flow will be established between the source and the substrate, which is not allowed by the floating substrate condition
    - Now we need to find out where the fluctuating charge δQ appears
        - At the gate terminal, all charge appear at the gate-oxide interface
        - At the N+ source terminal, the carriers responsible for conduction are electrons
            - They cannot enter the substrate when $V_\mathrm{G} < V_\mathrm{T}$, because the substrate is considered to be an insulator to these electrons
                - The capacitance can be considered to be formed between the N+ source boundary and the gate electrode
                - There are almost no overlap between the location where δQ appears at the gate and the source, so the capacitance is very small

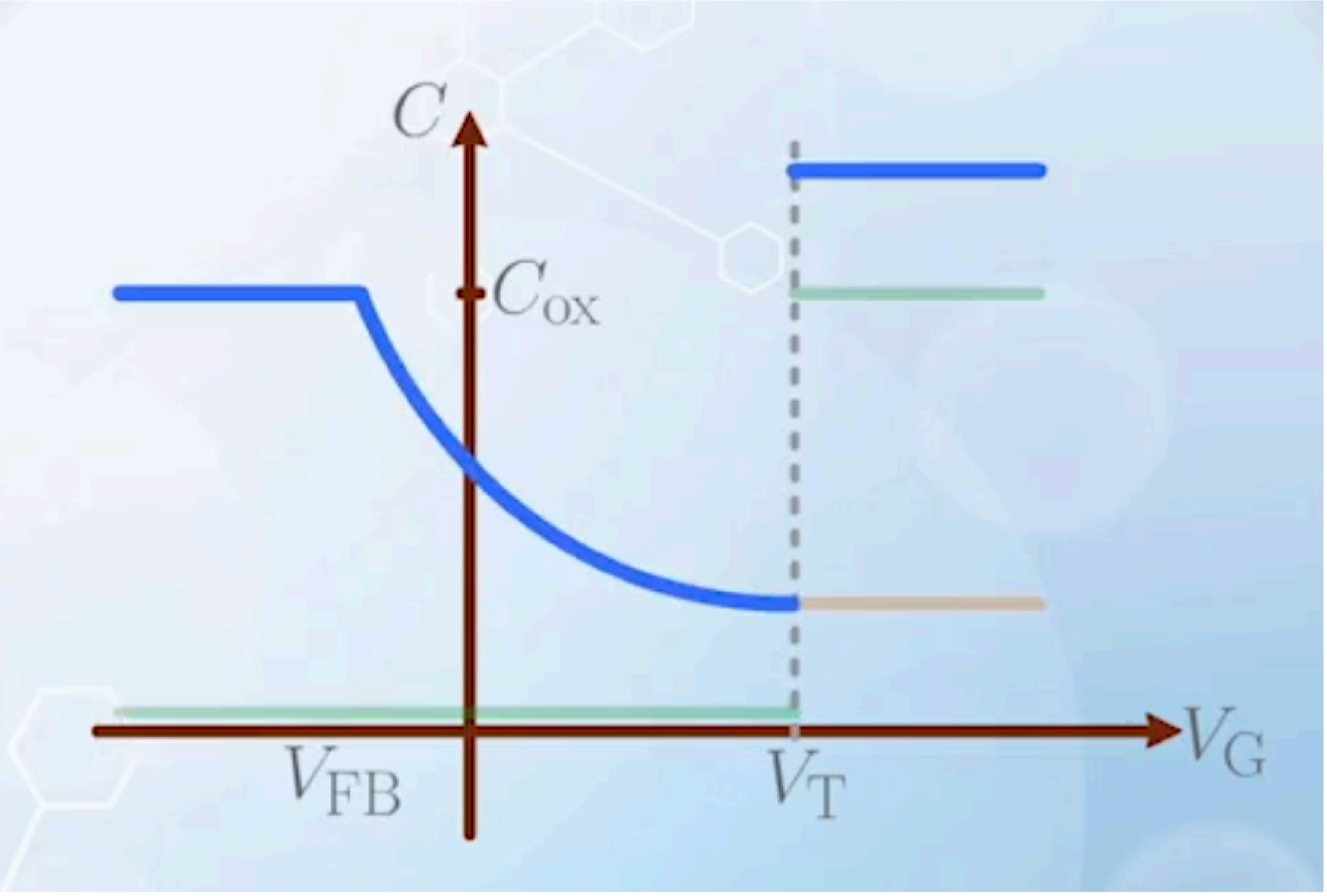


            - When $V_\mathrm{G} > V_\mathrm{T}$, the channel region is inverted, and a conductive path is formed which connects the source and the substrate
                - δQ can reach the silicon interface at the substrate

        - All of a sudden, an additional capacitance of C_{ox} is added to the system when V_{G} increases beyond V_{T}
        - There is a sudden jump in the capacitance value at V_{T}

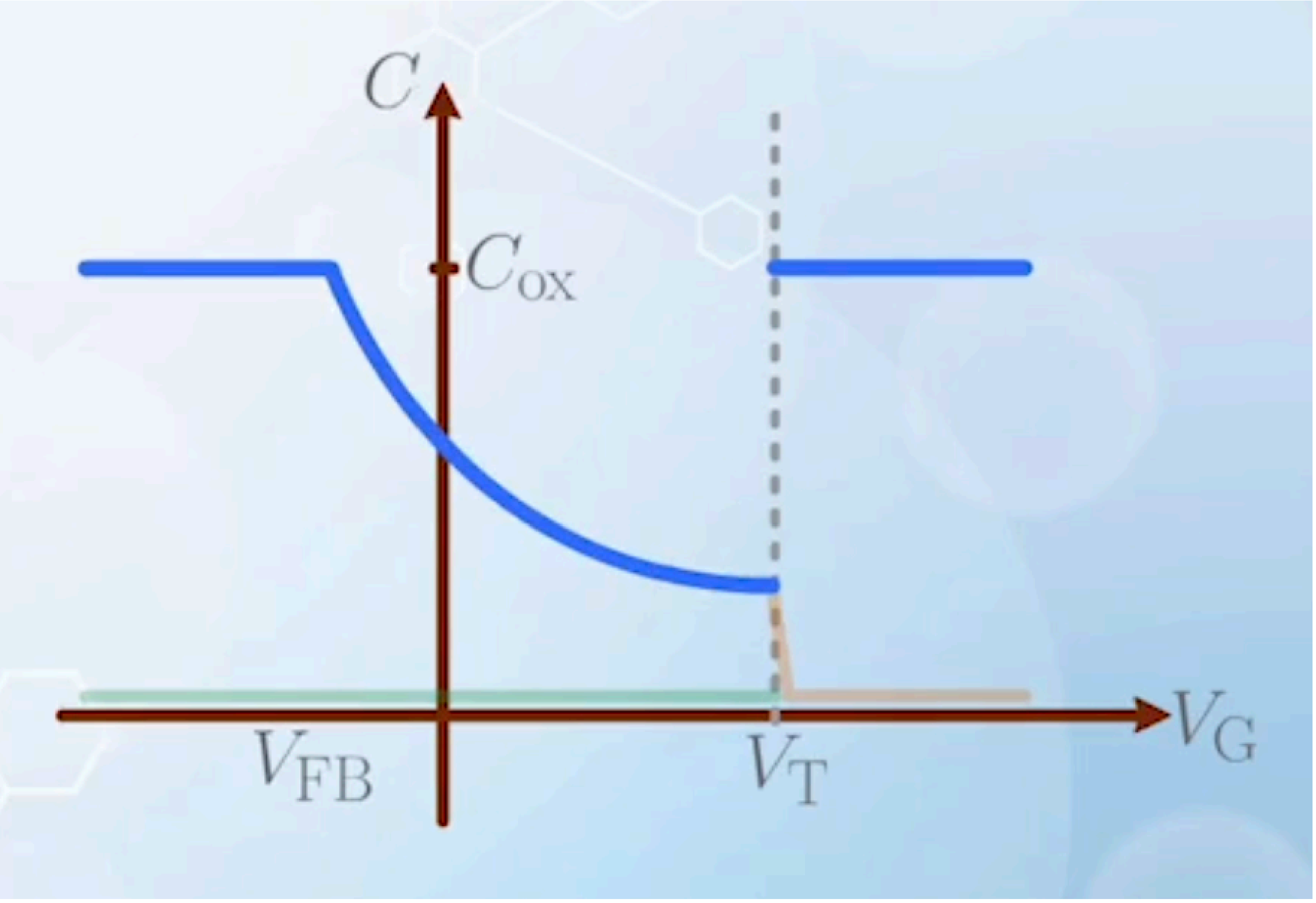


- Connecting **source and substrate** as a single terminal, we measure the capacitance between the gate and the combined source / substrate terminal
    - The resulting capacitance is the sum of the capacitance measured with floating source and floating substrate



- Two capacitors are assumed to have no interaction
- In reality, when the inversion layer is formed and grounded, the connection from the substrate to the gate electrode is cut off, and the **gate to substrate capacitance** becomes zero

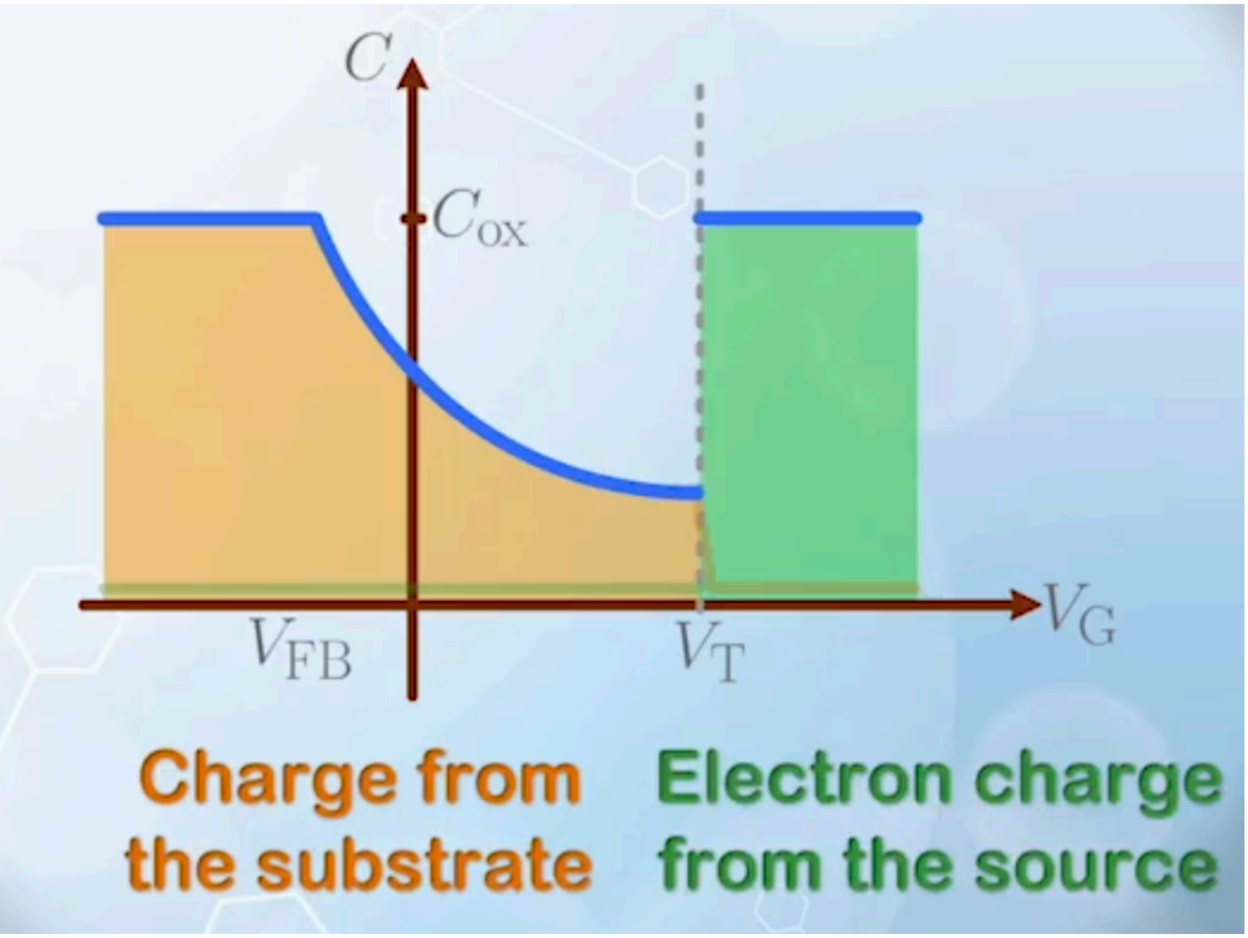
- The resulting capacitance is like

  It is just the low-frequency $C - V$ characteristic of a MOS capacitor

  - This is because when a source is added to the MOS capacitor and grounded, it can supply electrons to the channel when necessary without the generation process, thus the low-frequency characteristics will always be observed

- The curve shall be divided into two parts at V_T



  - Before V_T: the charge is provided by the substrate
  - After V_T: the charge is electrons provided by the source

Effects of Body Voltage

- When the substrate and source are connected to different voltages, mainly about negative substrate voltage relative to the source in an NMOSFET, or positive substrate voltage relative to the source in a PMOSFET, otherwise, the substrate-source diode will be forward biased and conduct current, which is not controlled by the gate
- For NMOSFET, when source, drain, and substrate voltages are all zero
  - A flat band voltage $V_\mathrm{G} = V_\mathrm{FB0}$ is applied to the gate, all energy bands are flat, and the Fermi level in the silicon substrate aligns
  - When a voltage is applied to the gate, causing the energy band to bend by $2\phi_\mathrm{B}$, the threshold voltage is reached, gate voltage is V_T0
- Now we apply a negative voltage to the substrate, keeping the gate voltage at V_FB0
  - This V_B raises the energy band at the P substrate, causing the Fermi level to break between the source and substrate

    - A band bending between the gate and the substrate also occurs because of the extra voltage difference, and the MOSFET is no longer at flat band condition
    - To bring the energy bands back to flat, a more negative voltage must be applied to the gate, and now the gate voltage is $V_{\mathrm{FB}} = V_{\mathrm{FB0}} + V_{\mathrm{B}}$, where V_{B} is negative
    - The voltage required to reach the **same band bending** ($2\phi_{\mathrm{B}}$) from the back to the front is now $V_{\mathrm{T0}} + V_{\mathrm{B}}$
        - However, previously at this band bending condition, the barrier between the source and the channel almost disappears, and electrons can easily flow from the source to the channel. But the negative V_{B} raises the channel energy for electrons
        - Also, the electrons generated in the channel will be drained away, as the drain and source is connected to a higher voltage, and the inversion charge cannot be collected
        - The system remains at **depletion**
    - To match the **threshold condition** with the case when $V_{\mathrm{B}} = 0$, an extra band bending of V_{B} is required, in addition to the previous band bending of $2\phi_{\mathrm{B}}$, and the new threshold voltage is marked as V_{T}, with band bending of $2\phi_{\mathrm{B}} - V_{\mathrm{B}}$ (minus sign as V_{B} is negative)

Threshold Voltage With a Substrate Bias

- To achieve threshold condition, more band bending from the substrate to the gate is required
- More band bending results in a larger depletion width, given by a new x_d, and $x_d > x_{d\mathrm{max}}$
- Recall that

$$V_{\mathrm{T}} = V_{\mathrm{FB}} + \phi_{\mathrm{S}} + V_{ox}$$

- Now

$$\begin{aligned} V_{\mathrm{FB}} &= V_{\mathrm{FB0}} + V_{\mathrm{B}} \\ \phi_{\mathrm{S}} &= 2\phi_{\mathrm{B}} - V_{\mathrm{B}} \end{aligned}$$

    - The sum of V_{FB} and ϕ_{S} remains unchanged
- All difference comes from V_{ox}

$$
\begin{aligned}
V_{ox} &= \frac{Q_{\mathrm{D}}}{C_{ox}} \\
Q_{\mathrm{D}} &= qN_A x_d \\
x_d &= \sqrt{\frac{2\varepsilon_{\mathrm{Si}}}{qN_A}(2\phi_{\mathrm{B}} - V_{\mathrm{B}})} \\
\Rightarrow Q_{\mathrm{D}} &= \sqrt{2qN_A\varepsilon_{\mathrm{Si}}(2\phi_{\mathrm{B}} - V_{\mathrm{B}})} \\
\Rightarrow \Delta Q_{\mathrm{D}} &= Q_{\mathrm{D}} - Q_{\mathrm{D0}} \\
&= \sqrt{2qN_A\varepsilon_{\mathrm{Si}}}\left(\sqrt{2\phi_{\mathrm{B}} - V_{\mathrm{B}}} - \sqrt{2\phi_{\mathrm{B}}}\right)
\end{aligned}
$$

- Putting all together

$$
\begin{aligned}
V_{\mathrm{T}} &= V_{\mathrm{T0}} + \frac{\Delta Q_{\mathrm{D}}}{C_{ox}} \\
&= V_{\mathrm{T0}} + \frac{\sqrt{2qN_A\varepsilon_{\mathrm{Si}}}}{C_{ox}}\left(\sqrt{2\phi_{\mathrm{B}} - V_{\mathrm{B}}} - \sqrt{2\phi_{\mathrm{B}}}\right) \\
&= V_{\mathrm{T0}} + \gamma\left(\sqrt{2\phi_{\mathrm{B}} - V_{\mathrm{B}}} - \sqrt{2\phi_{\mathrm{B}}}\right) \\
\text{where} \quad \gamma &= \frac{\sqrt{2qN_A\varepsilon_{\mathrm{Si}}}}{C_{ox}} \quad \text{body factor}
\end{aligned}
$$

  heavy doping or smaller C_{ox} will give rise to a higher body factor

- $2\phi_{\mathrm{B}}$ is usually approximately $0.7\,\mathrm{V}$ for doped silicon, similar to the turn-on voltage of silicon PN junctions and BJT

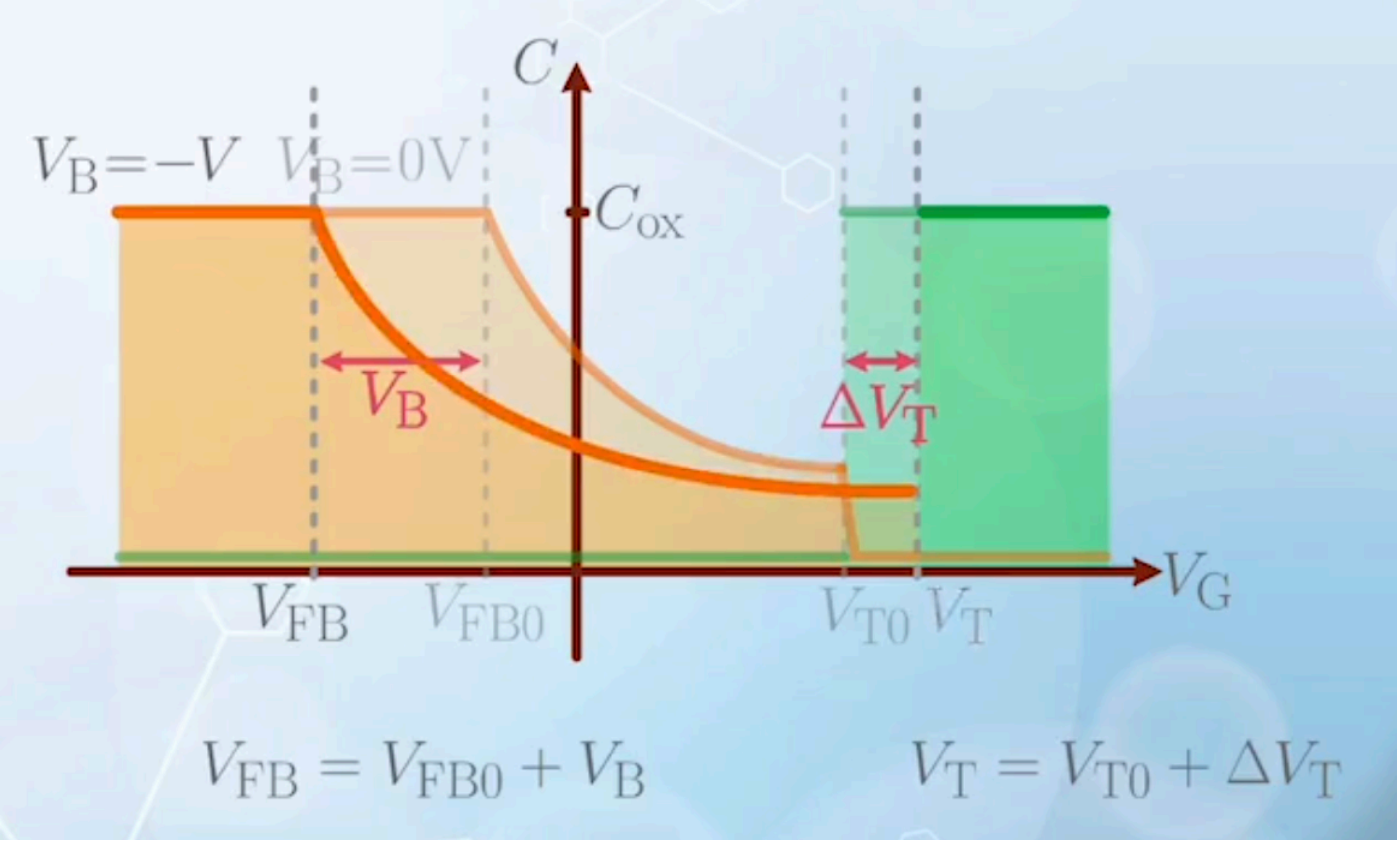
$$
V_{\mathrm{T}} = V_{\mathrm{T0}} + \gamma\left(\sqrt{0.7 - V_{\mathrm{B}}} - \sqrt{0.7}\right)
$$

  which shows that V_{T} increases with the square root of the substrate bias magnitude $|V_{\mathrm{B}}|$

MOS Capacitor With a Body Bias

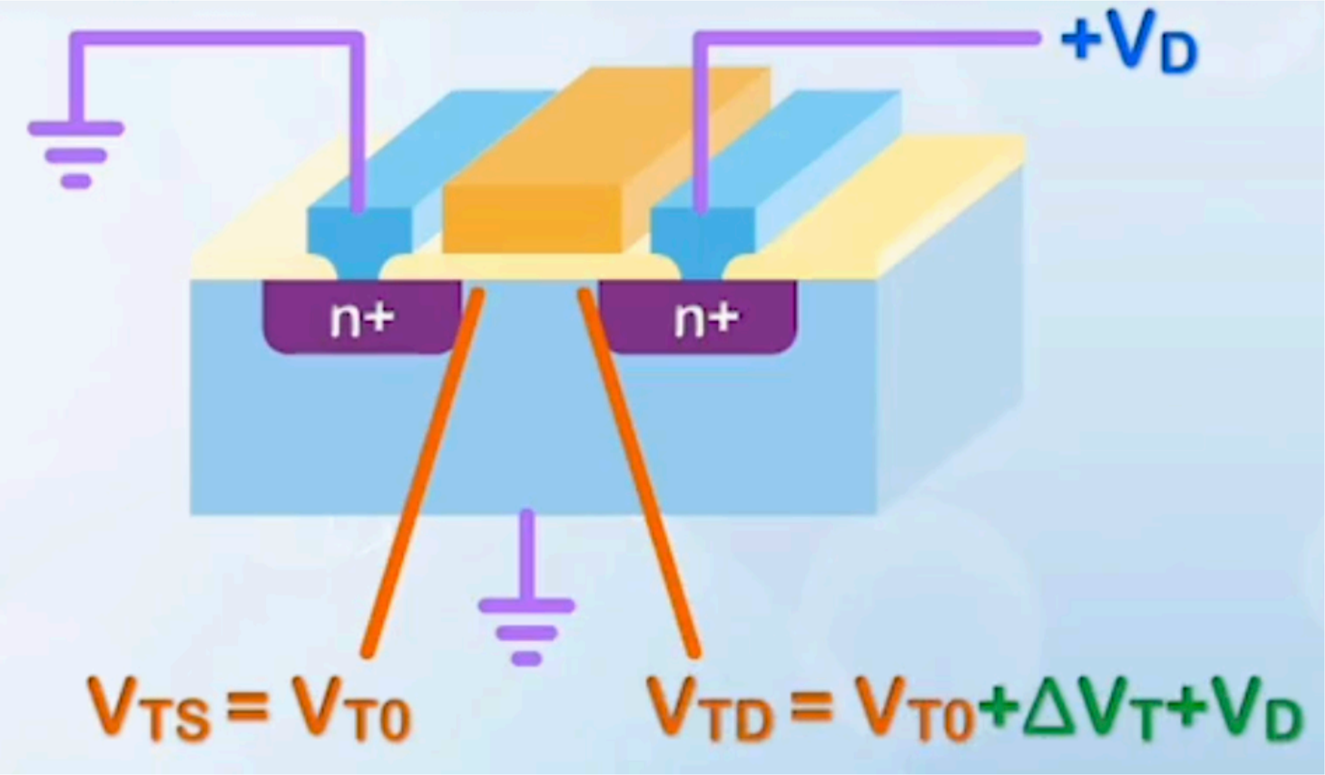
- When a negative body bias is applied to the substrate, the curve of the first part of the $C - V$ characteristics shifts left by $-V_{\mathrm{B}}$
- When x_d reaches $x_{d\max}$, the inversion layer cannot be formed yet, the source and drain terminals will force the capacitor to enter the **deep depletion region**, the depletion region continues to expand, and the capacitance continues to decrease, until V_{G} is the new threshold voltage $V_{\mathrm{T}} = V_{\mathrm{T0}} + \Delta V_{\mathrm{T}}$

- After that, the inversion layer is formed, and the capacitance jumps back to C_{ox}



- If we add a positive voltage V to the source, and the substrate is grounded, the $C - V$ curve will shift right by V as a whole, and the new V_T becomes $V_{T0} + |V| + \Delta V_T$
  - This (**grounded substrate and positive N+ terminal voltage**) is a more common setup in NMOSFET
  - Usually, $V_{TS} < V_{TD}$, as the source and substrate are usually grounded without any voltage difference, while the drain may have a positive voltage applied (the substrate is at

a negative voltage relative to the drain)



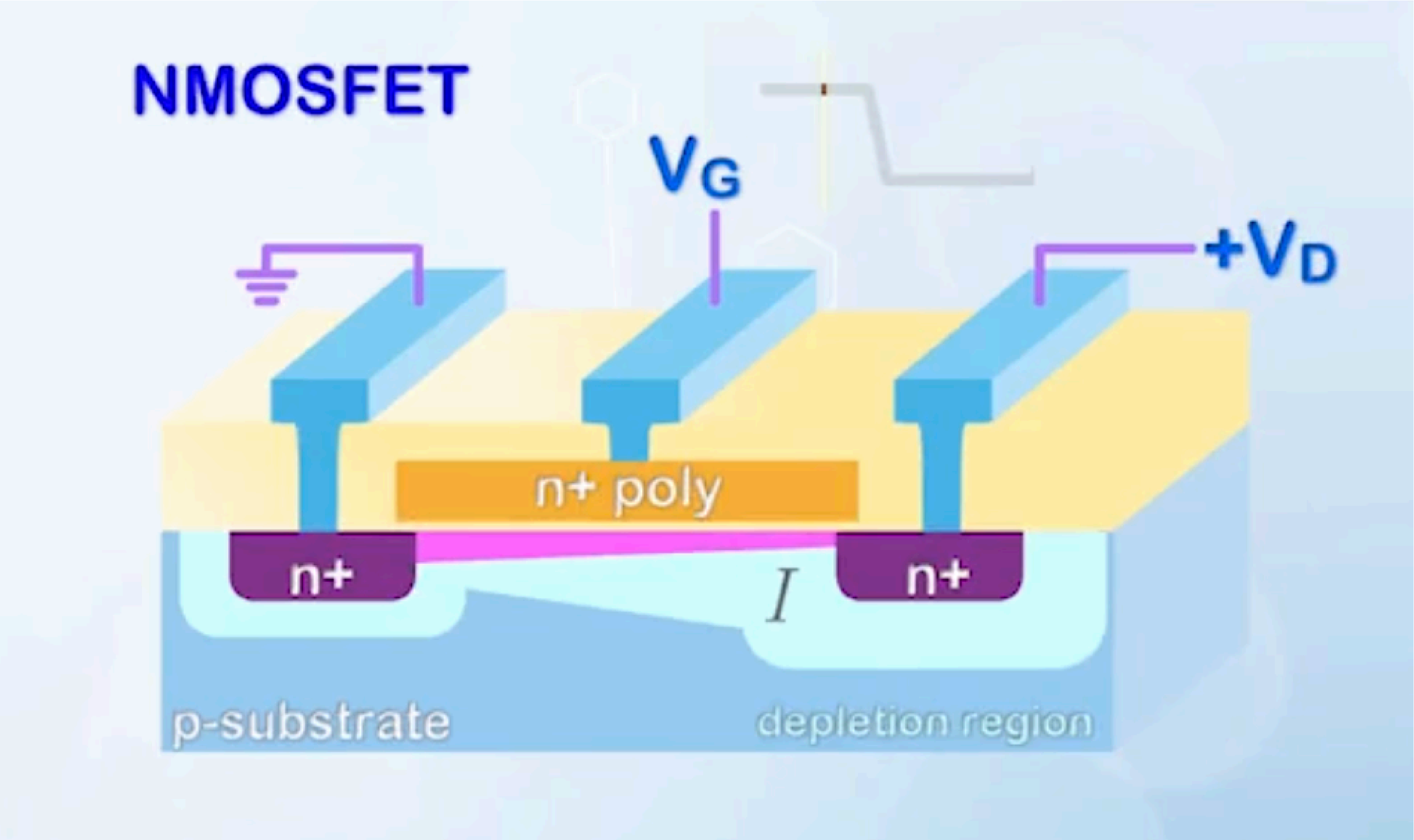
- This effect is usually ignored by approximating V_{T} anywhere in the channel to be V_{T0}

5. Classical MOSFET Turn-on Current

About how to calculate the current of a MOSFET when it is turned on, based on the classical pinchoff model, the channel length modulation effect, and some discussions about the inconsistencies and limitations of the model.

Charge and Velocity of Channel Carriers

MOSFET is basically a switch controlled by gate. When $V_{\mathrm{G}} < V_{\mathrm{T}}$, electrons from the source cannot enter the channel, and no current flows. When $V_{\mathrm{G}} > V_{\mathrm{T}}$, electrons from the source can enter the channel, and a conduction path is formed. A current will flow if a positive V_{G} is applied (NMOSFET).



Assume source and substrate are grounded

- As a switch, I_D is mostly assumed to be 0 when $V_\mathrm{G} < V_\mathrm{T}$
- When $V_\mathrm{G} > V_\mathrm{T}$, inversion electrons form in the channel, and electrons will be attracted from the source to drain by the positive V_D
    - The electric field is the main driving force, and the current is mainly **drift current**
    - The current depends on the **number of charge** available, and the **velocity of these charges**
    - The current at any location y is

      $$I(y) = Q(y)v(y)$$

      where $Q(y)$ is the amount of charge at location y
    - Defining y as the coordinate from the source to drain
    - There is no accumulation or removal of charges, the current flows remains constant along the channel
        - The region with more charges will have lower velocity, and vice versa
    - Starting from a small V_D close to 0
        - For a MOSFET with channel length L and width W, total charge under the gate is

$$C_{ox}WL(V_{\mathrm{G}} - V_{\mathrm{T}})$$

    - The charge per unit length is the above divided by L

$$Q(y) = C_{ox}W(V_{\mathrm{G}} - V_{\mathrm{T}})$$

- When V_{D} increases, the voltage across the gate capacitor near the drain will be reduced to $V_{\mathrm{G}} - V_{\mathrm{D}}$, so the charge per unit length near the drain becomes

$$Q(y) = C_{ox}W(V_{\mathrm{G}} - V_{\mathrm{T}} - V_{\mathrm{D}})$$

  which means, the charge near the drain is less than that near the source

    - The voltage in the channel somewhere between source and drain is $V(y)$, so the charge per unit length at location y is

$$Q(y) = C_{ox}W(V_{\mathrm{G}} - V_{\mathrm{T}} - V(y))$$

      V_{T} should be a function of y, as increasing $V(y)$ is similar to decreasing V_{D}, but the effect is ignored for now

- The velocity of electrons are usually assumed to be proportional to the electric field

$$v(y) = \mu E(y)$$

  where μ is the mobility of carriers. For electrons in silicon

$$\mu_{e(b)} = 1400\,\mathrm{cm}^2\mathrm{V}^{-1}\mathrm{s}^{-1}$$

  It is the mobility when electrons are moving **inside the bulk silicon**

    - When electrons are moving **near the silicon-oxide interface**, the mobility is lower due to more scattering

$$\mu_{e(s)} = 600\,\mathrm{cm}^2\mathrm{V}^{-1}\mathrm{s}^{-1}$$

    - For holes in silicon

$$\mu_{h(b)} = 450\,\mathrm{cm}^2\mathrm{V}^{-1}\mathrm{s}^{-1}$$
$$\mu_{h(s)} = 250\,\mathrm{cm}^2\mathrm{V}^{-1}\mathrm{s}^{-1}$$

- The current is given by

$$I_{\mathrm{D}} = C_{ox}W(V_{\mathrm{G}} - V_{\mathrm{T}} - V(y))\mu E(y)$$

Linear Region Current Equation

- The electric field is the negative gradient of voltage

$$E(y) = \frac{\mathrm{d}V(y)}{\mathrm{d}y}$$

- Now the current equation becomes

$$I_{\mathrm{D}}\mathrm{d}y = C_{ox}W\mu(V_{\mathrm{G}} - V_{\mathrm{T}} - V(y))\mathrm{d}V(y)$$

  There should be a negative sign in the equation, but let's focus on the magnitude for now

- Integrating both sides from source to drain

$$\begin{aligned}\int_{\text{source}}^{\text{drain}} I_{\mathrm{D}}\mathrm{d}y &= \int_{\text{source}}^{\text{drain}} C_{ox}W\mu(V_{\mathrm{G}} - V_{\mathrm{T}} - V)\mathrm{d}V \\ \int_{0}^{L} I_{\mathrm{D}}\mathrm{d}y &= \int_{0}^{V_{\mathrm{D}}} C_{ox}W\mu(V_{\mathrm{G}} - V_{\mathrm{T}} - V)\mathrm{d}V \\ I_{\mathrm{D}}L &= C_{ox}W\mu\left[(V_{\mathrm{G}} - V_{\mathrm{T}})V_{\mathrm{D}} - \frac{V_{\mathrm{D}}^2}{2}\right] \\ I_{\mathrm{D}} &= \mu C_{ox}\frac{W}{L}\left[(V_{\mathrm{G}} - V_{\mathrm{T}})V_{\mathrm{D}} - \frac{V_{\mathrm{D}}^2}{2}\right]\end{aligned}$$

- A simpler approach is
  - The current is also given by

$$I_{\mathrm{D}} = Q_{\mathrm{avg}}v_{\mathrm{avg}}$$

  - The average charge per unit length is just the average of the charge at source and drain

$$Q_{\mathrm{avg}} = C_{ox}W\left(V_{\mathrm{G}} - V_{\mathrm{T}} - \frac{V_{\mathrm{D}}}{2}\right)$$

  - As of the average velocity

$$v_{\mathrm{avg}} = \mu E_{\mathrm{avg}}$$

  - If we assume the electric field is uniform along the channel

$$E_{\mathrm{avg}} = \frac{V_{\mathrm{D}}}{L}$$

- In reality, the electric field near the source is lower, and that near the drain is higher, as electrons move faster near the drain
- However, integrating E over the channel length must give V_{D} still, so the average electric field is still V_{D}/L
- Therefore,

$$I_{\mathrm{D}} = C_{ox} W \left(V_{\mathrm{G}} - V_{\mathrm{T}} - \frac{V_{\mathrm{D}}}{2} \right) \mu \frac{V_{\mathrm{D}}}{L}$$

  which is the same as the previous result

Saturation Region Current Equation

As V_{D} increases, I_{D} will increase until $V_{\mathrm{D}} = V_{\mathrm{G}} - V_{\mathrm{T}}$. If we still follow the previous current equation, when V_{D} exceeds $V_{\mathrm{G}} - V_{\mathrm{T}}$, the current will start to decrease, and eventually become 0. But from measurements, the current of a MOSFET eventually saturates and becomes a constant with a high enough V_{D}.

- For the inversion charge distribution

$$C_{ox} W (V_{\mathrm{G}} - V_{\mathrm{T}} - V(y))$$

  to be valid, $V_{\mathrm{G}} - V_{\mathrm{T}} - V(y)$ must be positive. Otherwise, the channel will be depleted, and there will be no charge for conduction

  - This will happen at the drain when V_{D} becomes larger than $V_{\mathrm{G}} - V_{\mathrm{T}}$
  - A **pinchoff region**, where the channel is depleted, will form near the drain
  - The channel can be separated into two different regions by the $V(y) = V_{\mathrm{G}} - V_{\mathrm{T}}$ point
    - **Gradual channel region**: the region where $V(y) < V_{\mathrm{G}} - V_{\mathrm{T}}$, inversion charge exists, and the channel behaves like a conductor
    - **Pinchoff region**: the region where $V(y) > V_{\mathrm{G}} - V_{\mathrm{T}}$, inversion charge is depleted, and the channel becomes an insulator
  - All drain voltage beyond $V_{\mathrm{G}} - V_{\mathrm{T}}$ will be dropped across the pinchoff region
  - **!!!INCONSISTENCY MENTIONED!!!**

> In the **pinchoff region**, the channel is depleted, and $Q(y) = 0$. Meanwhile, the current is given by $I_{\mathrm{D}} = Q(y)v(y)$. How can there be a current if there is no charge?
>
> This inconsistency will be resolved later, and for now, we just forget about the pinchoff region, and assume the drain is moved to the $V_{\mathrm{G}} - V_{\mathrm{T}}$ point, with V_{D} picking up the value of $V_{\mathrm{G}} - V_{\mathrm{T}}$.

- Assuming the pinchoff region is very small compared to the channel length, and the length of the gradual channel region can be approximated as L
- Then, the current in the pinchoff condition is calculated with the same equation as before, but with V_{D} replaced by $V_{\mathrm{G}} - V_{\mathrm{T}}$

$$
\begin{aligned}
I_{\mathrm{Dsat}} &= \mu C_{ox} \frac{W}{L} \left[(V_{\mathrm{G}} - V_{\mathrm{T}})(V_{\mathrm{G}} - V_{\mathrm{T}}) - \frac{(V_{\mathrm{G}} - V_{\mathrm{T}})^2}{2} \right] \\
&= \frac{1}{2} \mu C_{ox} \frac{W}{L} (V_{\mathrm{G}} - V_{\mathrm{T}})^2
\end{aligned}
$$

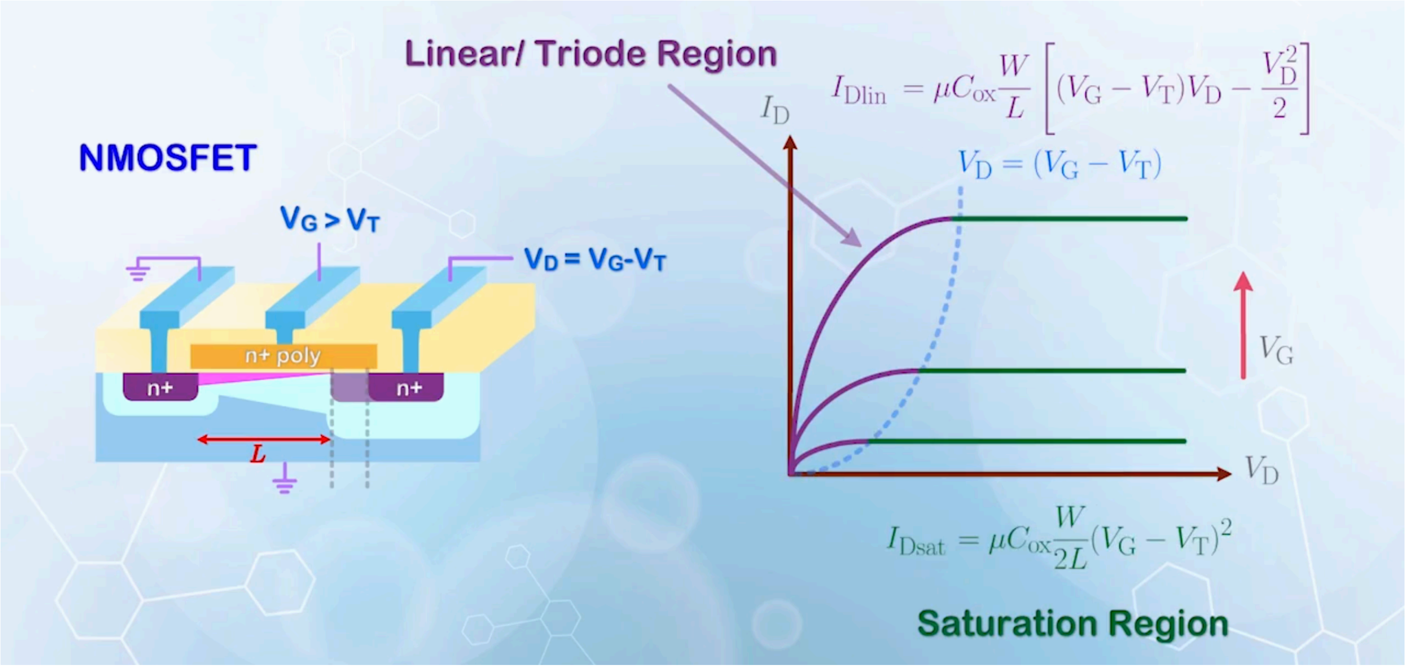
  Any V_{D} beyond $V_{\mathrm{G}} - V_{\mathrm{T}}$ will be dropped across the pinchoff region, and will not affect the current

- Therefore, the final current equation is

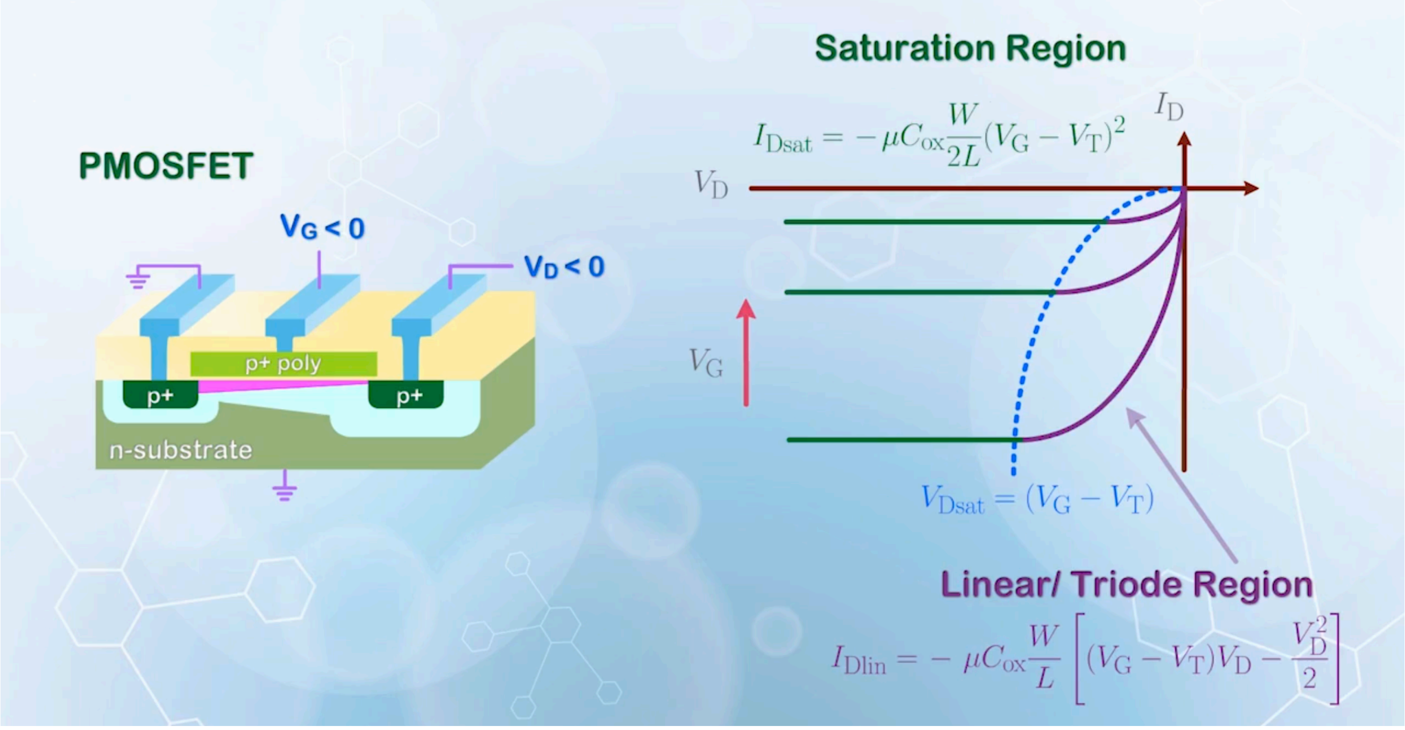
$$
\begin{aligned}
I_{\mathrm{Dlin}} &= \mu C_{ox} \frac{W}{L} \left[(V_{\mathrm{G}} - V_{\mathrm{T}}) V_{\mathrm{D}} - \frac{V_{\mathrm{D}}^2}{2} \right] && V_{\mathrm{D}} < V_{\mathrm{G}} - V_{\mathrm{T}} \\
I_{\mathrm{Dsat}} &= \frac{1}{2} \mu C_{ox} \frac{W}{L} (V_{\mathrm{G}} - V_{\mathrm{T}})^2 && V_{\mathrm{D}} \geq V_{\mathrm{G}} - V_{\mathrm{T}}
\end{aligned}
$$

  - The first region is called the **linear region** or **triode region**, and the second region is called the **(current) saturation region**
  - The separation between the two regions is

$$
V_{\mathrm{Dsat}} = V_{\mathrm{G}} - V_{\mathrm{T}}
$$


Linear/ Triode Region
$I_{\mathrm{Dlin}} = \mu C_{\mathrm{ox}} \frac{W}{L}\left[(V_{\mathrm{G}} - V_{\mathrm{T}})V_{\mathrm{D}} - \frac{V_{\mathrm{D}}^2}{2}\right]$
NMOSFET
I_{D}
$V_{\mathrm{D}} = (V_{\mathrm{G}} - V_{\mathrm{T}})$
$V_G > V_T$
$V_D = V_G - V_T$
n+ poly
n+
n+
L
V_{G}
V_{D}
$I_{\mathrm{Dsat}} = \mu C_{\mathrm{ox}} \frac{W}{2L}(V_{\mathrm{G}} - V_{\mathrm{T}})^2$
Saturation Region

- Same applies to PMOSFET, with voltages and currents become negative relative to the source
    - Negative sign is added to the current, as the current in PMOSFET flows out of the drain


Saturation Region
I_{D}
$I_{\mathrm{Dsat}} = -\mu C_{\mathrm{ox}} \frac{W}{2L}(V_{\mathrm{G}} - V_{\mathrm{T}})^2$
PMOSFET
V_{D}
$V_G < 0$
$V_D < 0$
p+ poly
p+
p+
n-substrate
V_{G}
$V_{\mathrm{Dsat}} = (V_{\mathrm{G}} - V_{\mathrm{T}})$
Linear/ Triode Region
$I_{\mathrm{Dlin}} = -\ \mu C_{\mathrm{ox}} \frac{W}{L}\left[(V_{\mathrm{G}} - V_{\mathrm{T}})V_{\mathrm{D}} - \frac{V_{\mathrm{D}}^2}{2}\right]$

Channel Length Modulation

Previously, we assumed the drain voltage beyond V_{Dsat} does not affect I_{D}. However, in reality, increasing V_{D} beyond V_{Dsat} will extend the pinchoff region further into the channel, and the length of the gradual channel region will be reduced. As a result, I_D will increase. This is the **channel length modulation** effect, describing how V_D affects I_D by changing the effective channel length.

- In saturation region, the current is given by

$$I_\mathrm{Dsat} = \frac{1}{2}\mu C_{ox}\frac{W}{L_\mathrm{ch}}(V_\mathrm{G} - V_\mathrm{T})^2$$

  where L_ch is the length of the region where the channel voltage increases from 0 to V_Dsat
- Replacing L_ch with $L - \Delta L$, and perform some mathematical magic, we have

$$\begin{aligned} I_\mathrm{Dsat} &= \frac{1}{2}\mu C_{ox}\frac{W}{L - \Delta L}(V_\mathrm{G} - V_\mathrm{T})^2 \\ &= \frac{1}{2}\mu C_{ox}\frac{W}{L}\frac{1}{1 - \frac{\Delta L}{L}}(V_\mathrm{G} - V_\mathrm{T})^2 \\ &= I_\mathrm{Dsat0}\frac{1}{1 - \frac{\Delta L}{L}} \\ &= I_\mathrm{Dsat0}\left(1 + \frac{\Delta L}{L} + \left(\frac{\Delta L}{L}\right)^2 + \cdots\right) \\ &\approx I_\mathrm{Dsat0}\left(1 + \frac{\Delta L}{L}\right) \quad \Delta L \text{ is assumed to be very small} \end{aligned}$$

- To calculate ΔL, we use Poisson's equation
  - In the pinchoff region, there are only the depleted charge from the dopant ions
  - Therefore, the charge density is

$$\rho = qN_A$$

  - Integrate twice to get the voltage difference across the pinchoff region

$$V = \frac{qN_A}{2\varepsilon_\mathrm{Si}}(\Delta L)^2$$

  - The voltage across the pinchoff region is $V_\mathrm{D} - V_\mathrm{Dsat}$
  - Thus,

$$\Delta L = \sqrt{\frac{2\varepsilon_\mathrm{Si}}{qN_A}(V_\mathrm{D} - V_\mathrm{Dsat})}$$

  - Putting it back

$$I_{\text{Dsat}} = I_{\text{Dsat0}} \left(1 + \frac{1}{L} \sqrt{\frac{2\varepsilon_{\text{Si}}}{qN_A}(V_{\text{D}} - V_{\text{Dsat}})} \right)$$

- **!!!INCONSISTENCY BACK AGAIN!!!**

  For the same reason mentioned before, ΔL given here is physically incorrect, as assuming the pinchoff region only contains depleted charge is physically incorrect.

- I_{Dsat} has a square root dependence on V_{D} beyond V_{Dsat}
- As the range of V_{D} is limited, we can linearize the equation

  $$I_{\text{Dsat}} \approx I_{\text{Dsat0}}(1 + \lambda V_{\text{D}})$$

  where λ is the **channel length modulation parameter**
- The slope in the $I_{\text{D}} - V_{\text{D}}$ curve is then given by $I_{\text{Dsat}}\lambda$, and the output resistance is

  $$r_{\text{o}} = \frac{1}{\lambda I_{\text{Dsat0}}}$$

  which is similar to the output resistance of a BJT, which is given by

  $$r_{\text{o}} = \frac{V_A}{I_{\text{C}}}$$

  with V_A replaced by $1/\lambda$ and I_{C} replaced by I_{Dsat0}
  - Therefore, $1/\lambda$ is sometimes called the V_A of a MOSFET, and the unified output resistance equation is

    $$r_{\text{o}} = \frac{V_A + V_{\text{D}}}{I_{\text{Dsat}}} \approx \frac{V_A}{I_{\text{Dsat}}}$$

    assuming $V_A \gg V_{\text{D}}$

Inconsistencies in the Pinchoff Model

We have derived different equations for the MOSFET current in linear and saturation regions based on the pinchoff model.

- For $V_{\text{D}} < V_{\text{G}} - V_{\text{T}}$

  $$I_{\text{Dlin}} = \mu C_{ox} \frac{W}{L} \left[(V_{\text{G}} - V_{\text{T}})V_{\text{D}} - \frac{V_{\text{D}}^2}{2} \right]$$

- For $V_{\mathrm{D}} > V_{\mathrm{G}} - V_{\mathrm{T}}$

$$I_{\mathrm{Dsat}} = \frac{1}{2}\mu C_{ox}\frac{W}{L}(V_{\mathrm{G}} - V_{\mathrm{T}})^2(1 + \lambda V_{\mathrm{D}})$$

- There are some inconsistencies
  - **Mathematical inconsistencies**: using the two equations, the current at $V_{\mathrm{D}} = V_{\mathrm{G}} - V_{\mathrm{T}}$ is not continuous
    - Actually, the equation given for the saturation region is only used to illustrate the effect of V_{D} on I_{D}, and is never used to calculate I_{Dsat} besides finding the slope after differentiation
    - The simplest way to correct for the discontinuity is just substitute V_{D} with $V_{\mathrm{D}} - V_{\mathrm{Dsat}}$, but **who cares?**
    - We just assume the channel length modulation effect is very small, and use I_{Dsat0} to approximate I_{Dsat}
  - **Physical inconsistencies**: according to the pinchoff model, $Q(y) = 0$ for y in the pinchoff region. To obtain a finite current, $v(y)$ must be infinite. However, based on the knowledge of relativity theory, the speed of light is the limit of all measurable speed, so the velocity of electrons cannot be infinite
- However, it is still a good approximation to reality in some special cases
- These are the limitations of the pinchoff model, more accurate models will be discussed later

6. Subthreshold Region Current and Effect of Body Bias

About the subthreshold behaviors of MOSFETs, the turn-on characteristics, comparison between MOSFETs and BJTs, and the effect of body bias and substrate depletion charge on MOSFET I-V characteristics.

MOSFET Subthreshold Region

For a properly designed MOSFET, the V_{T} is usually positive for N MOSFET, and negative for P MOSFET. We will assume it is the case for the rest of this section, even though there are some exceptions.

We have derive the current equation for MOSFET when $V_{\mathrm{G}} > V_{\mathrm{T}}$. Consider NMOSFETs. When $0 < V_{\mathrm{G}} < V_{\mathrm{T}}$, we usually say that the MOSFET is operating in the **subthreshold region**. Many people assumes the current in this region is zero, but it is not true. In fact, the subthreshold current is very important, and determines how well a MOSFET can be turned off in modern electronic circuits.

- Consider the potential barrier V_{bi} between the source-body junction
- When $V_{\mathrm{G}} = V_{\mathrm{T}}$, the potential barrier almost disappears, and the MOSFET turns on
- Before the MOSFET turns on, it is similar to a BJT, with V_{G} changing the barrier height, and the change of barrier height is determined by the surface band bending ϕ_{s}
- In a BJT, the collector current under forward active mode is given by

$$I_{\mathrm{C}} = I_{\mathrm{C0}} e^{\frac{qV_{\mathrm{BE}}}{kT}}$$

- Similarly, the subthreshold current of a MOSFET can be expressed as

$$I_{\mathrm{D}} = I_{\mathrm{D0}} e^{\frac{q\phi_{\mathrm{s}}}{kT}}$$

  where I_{D0} is a constant, and is usually measured instead of calculated
- When $V_{\mathrm{G}} > V_{\mathrm{T}}$, I_{D} is limited by the resistance of the channel, instead of the source to body potential barrier
- From previous section, we know that in depletion region,

$$\phi_{\mathrm{S}} = \frac{1}{n}(V_{\mathrm{G}} - V_{\mathrm{FB}})$$

- The subthreshold region also operates in depletion region, as we have $0 < V_{\mathrm{G}} < V_{\mathrm{T}}$, so we can substitute ϕ_{S} into the equation of I_{D}, and get

$$I_{\mathrm{D}} = I_{\mathrm{D0}} e^{\frac{q(V_{\mathrm{G}} - V_{\mathrm{FB}})}{nkT}}$$

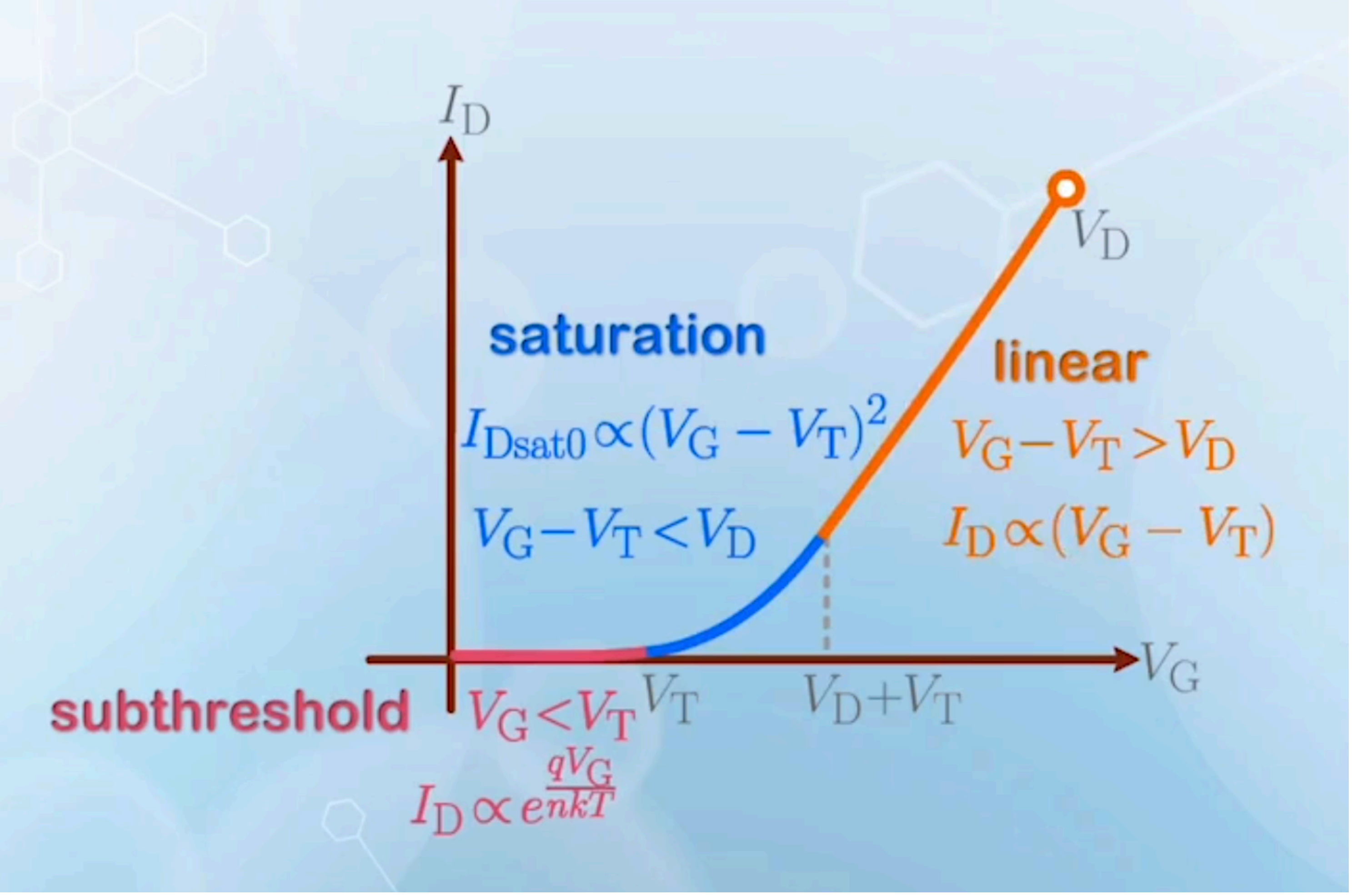
- V_{FB} is also a bias independent constant for a given MOSFET, it can be lumped into I_{D0}, and we can finally get

$$I_{\mathrm{D}} = I_{\mathrm{D0}} e^{\frac{qV_{\mathrm{G}}}{nkT}}$$

Turn-on Characteristics

We have now obtained all the equations for drain current of a MOSFET from subthreshold to strong inversion, we can now study the characteristics of a MOSFET when it is turned on or off with V_{G}.

Plotting the $I_{\mathrm{D}} - V_{\mathrm{G}}$ characteristics of a NMOSFET in linear scale:



- Before V_G reaches V_T, the MOSFET is in subthreshold region, and I_D increases exponentially with V_G
  - Because the current is relatively small in this region, it is difficult to observe its characteristics in a linear scale
- After V_G exceeds V_T, but before it reaches $V_D + V_T$, the MOSFET is in the saturation region, and I_D has a square dependence on V_G
- After V_G exceeds $V_D + V_T$, the MOSFET is in the linear region, and I_D increases linearly with V_G
- When measured with a larger V_D
  - The curve remains more or less the same in subthreshold and saturation region, because V_D has no effect on I_{Dsat} besides insignificant contributions of channel length modulation effect
  - The linear region extends further, as the transition point $V_G = V_{D2} + V_T$ shifts right
- There are some similarities between this graph, and the $I - V$ characteristics of BJTs and PN junctions
  - The turn-on voltage for silicon junctions is assumed to be $0.7\,\mathrm{V}$

- To observe the subthreshold characteristics more clearly, we can plot the same data in a semi-log scale
- The subthreshold region now becomes a straight line, showing the exponential dependence of I_D on V_G
- It is similar to the BJT Gummel Plot, or the $\log I_\text{C} - V_\text{BE}$ graph of a BJT
  - The part beyond V_on of a BJT **cannot** be used, as it will cause a large current flowing through the base
  - In MOSFET, we **can** use this region, as the gate blocks the current with the insulating oxide
- The slope of the subthreshold region is

  $$\text{slope} = \frac{q}{nkT}\log(e)$$

  It measures how abrupt a MOSFET can be turned on, but the swing is more commonly used

  $$\text{Subthreshold Swing} = \frac{1}{\text{slope}} = \frac{nkT}{q}\ln(10)\ (\text{mV}/\text{dec})$$

  - At room temperature, the subthreshold swing is approximately $60n\,\text{mV}/\text{dec}$, where n is the ideality factor given by

    $$n = 1 + \frac{C_\text{D}}{C_{ox}}$$

Subthreshold Swing

The subthreshold swing indicates the ratio between the on state current and off state current of a MOSFET. This is because once V_T is fixed, the leakage current at $V_\text{G} = 0$ is determined by how fast V_G can turn off the current, controlled by the subthreshold swing.

- For example, $S = 80\,\text{mV}/\text{dec}$, and $V_\text{T} = 0.8\,\text{V}$
  - This means for every $80\,\text{mV}$ decrease in V_G, I_D decreases by a factor of 10
  - From $V_\text{G} = 0.8\,\text{V}$ to $V_\text{G} = 0\,\text{V}$, there are 10 such steps, so the current decreases by a factor of 10^{10}
  - If V_T is reduced to $0.4\,\text{V}$, the current will only decrease by a factor of 10^5, 10^5 times larger than before, which is significant
- The actual leakage current may be larger than the predicted value due to other effects
  - The leakage current from the drain to substrate may define the lowest bound of the leakage current

        - This current is independent of V_{G}, and appears to be flat in the $\log I_{\mathrm{D}} - V_{\mathrm{G}}$ plot
- As $S = 60n\,\mathrm{mV/dec}$, we need to minimize n to build a good switch
    - n is given by

$$n = 1 + \frac{C_D}{C_{ox}}$$

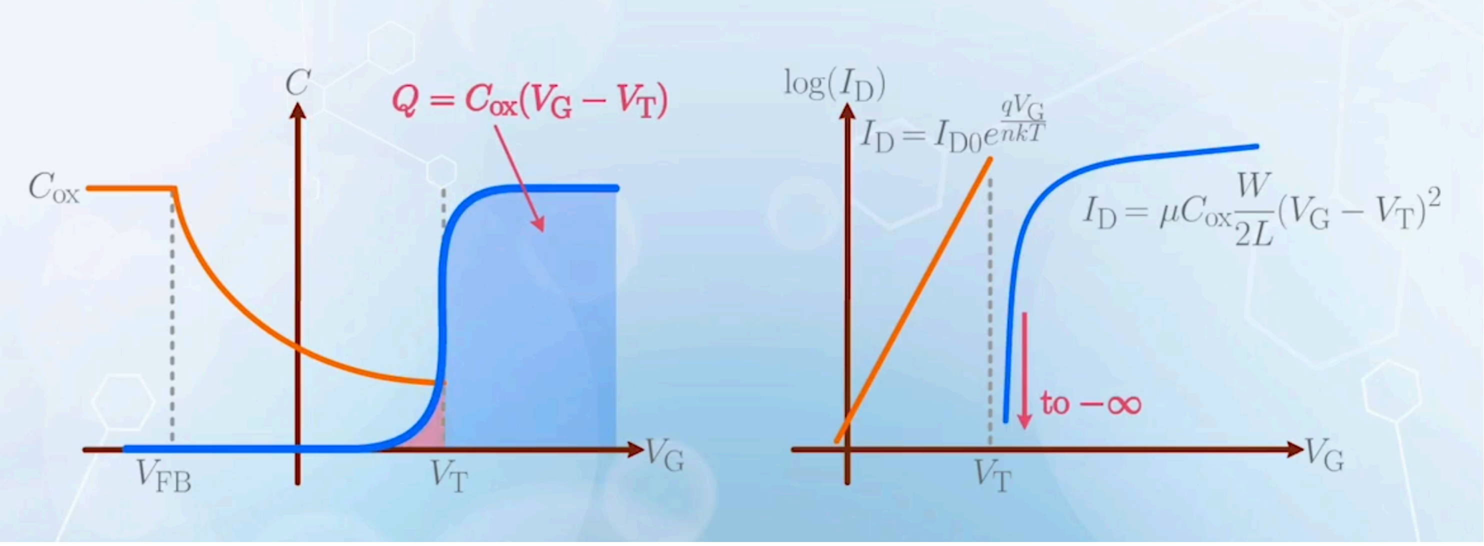
    - This can be done by maximizing C_{ox} and minimizing C_D
        - Reducing C_{D} by lowering the substrate doping concentration is limited by another constraint, which will be discussed later
        - It is more common to increase C_{ox} by using a thinner oxide layer
- The best achievable subthreshold swing at room temperature is approximately $60\,\mathrm{mV/dec}$, when $n = 1$, just the case of BJTs

Similar to the linear plot, I_{D} measured under different V_{D} overlaps in the subthreshold and saturation region, and start to separate in the linear region.

Current at the Threshold Voltage

Combining the subthreshold current equation and the strong inversion equations, we will observe a discontinuity at $V_{\mathrm{G}} = V_{\mathrm{T}}$, as the strong inversion current equation predicts $I_{\mathrm{D}} = 0$ at V_{T}.

This discontinuity occurs because we used $Q = C_{ox} V_{\mathrm{G}} - V_{\mathrm{T}}$ to calculate the inversion charges at the channel near the source end, which means the inversion charge below threshold is zero, and abruptly appears when V_{G} exceeds V_{T}.


C
$Q = C_{\mathrm{ox}}(V_{\mathrm{G}} - V_{\mathrm{T}})$
C_{ox}
V_{FB}
V_{T}
V_{G}
$\log(I_{\mathrm{D}})$
$I_{\mathrm{D}} = I_{\mathrm{D0}} e^{\frac{qV_{\mathrm{G}}}{nkT}}$
$I_{\mathrm{D}} = \mu C_{\mathrm{ox}} \frac{W}{2L}(V_{\mathrm{G}} - V_{\mathrm{T}})^2$
to $-\infty$
V_{T}
V_{G}

In reality, the inversion charge appears before threshold, as there are always electrons in the conduction band to prevent the current at V_T from going to 0.

Handling the current at V_T requires tedious mathematics to solve the surface potential, and there are some other more advanced courses that deal with the formulation of a continuous current voltage equations from the subthreshold to the strong inversion regions.

The main takeaway of this section is that there is a small region around V_T that we do not know how to calculate the current.

MOSFET v.s. BJT

BJTs and MOSFETs are usually considered very different devices operating with different principles. However, are actually very similar. They are both comprised of the sam PNP / NPN structure, and MOSFETs in subthreshold region operates very similarly to BJTs.

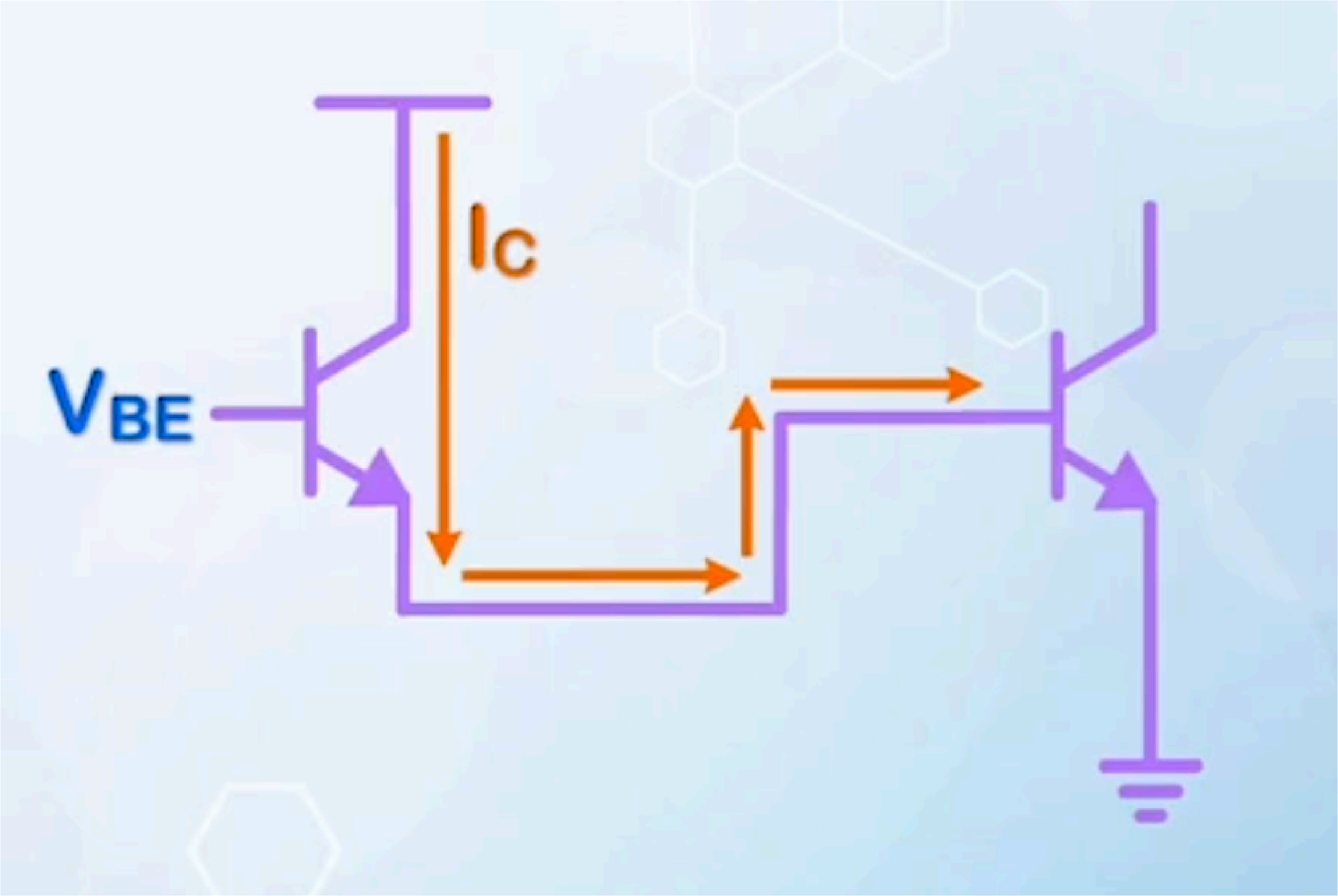
When considering the performance of a device, we not only consider its output, but also the loading device introduced to operate it. More specifically, the speed of a device is determined by the speed to charge the input capacitance of similar devices to the required voltage through its current.

If input capacitance is C, and has to be charged to voltage V with a driving current I to achieve transition, the delay is given by

$$\text{delay} = \frac{CV}{I}$$

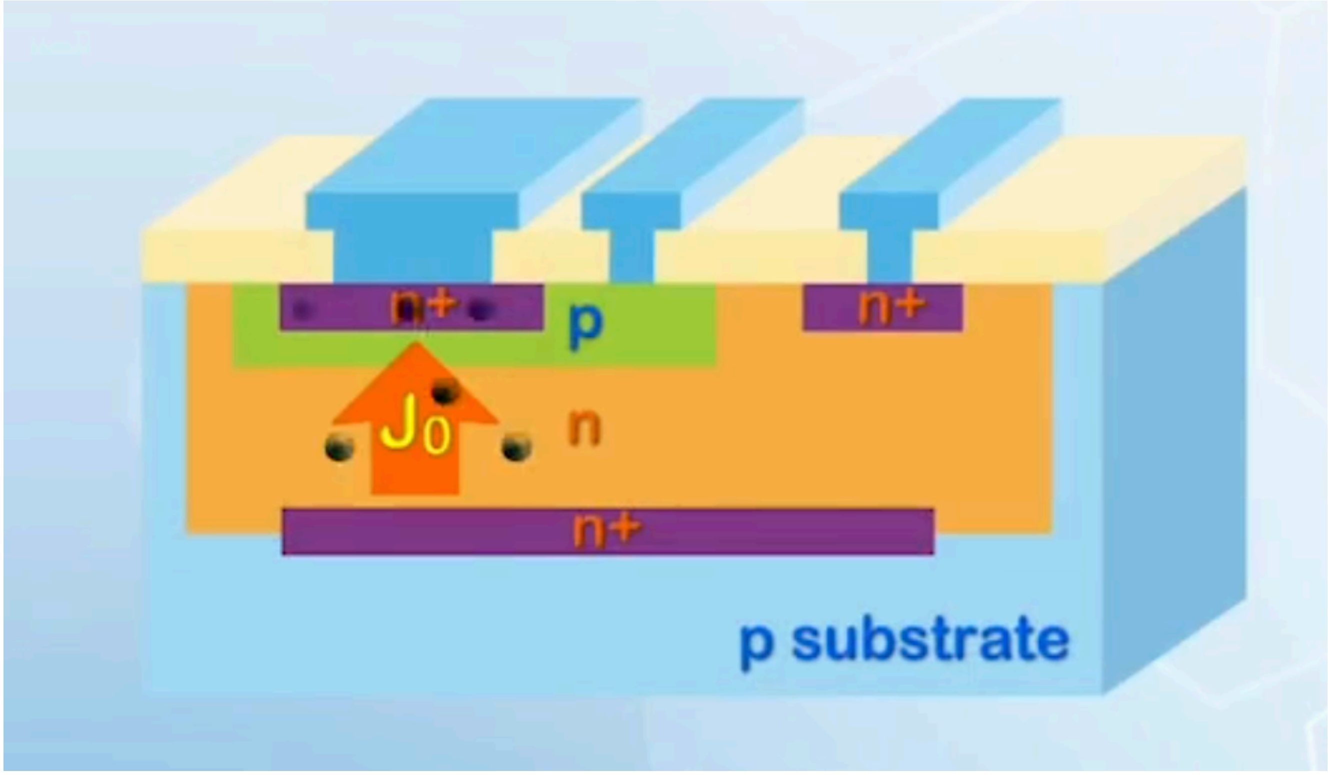
Or the speed can be characterized by $\frac{I}{C}$ for given power supply.

- Consider a BJT driving itself



  - Assume a specific current density J_0 through the BJT
  - The size of a BJT is mainly determined by the emitter area given by W and L
  - The current flows vertically, and is given by

$$I = J_0 W L$$



- The input capacitance is mainly determined by the base-emitter junction capacitance C_π
- The capacitance is given by

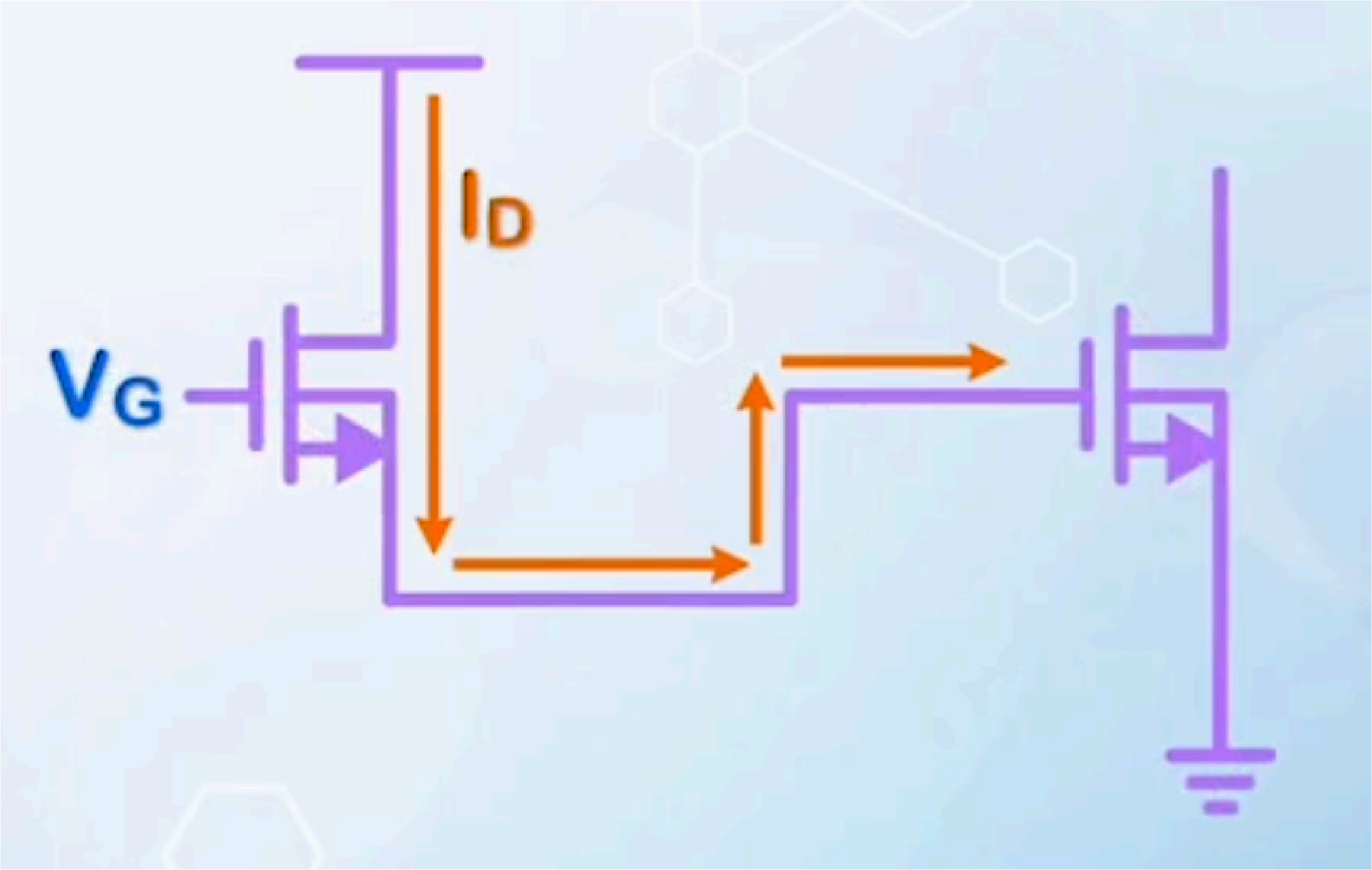
  $$C = C_{\pi 0} W L$$

  where $C_{\pi 0}$ is the normalized base-emitter capacitance per unit area
- The speed:

  $$\text{speed} \sim \frac{I}{C} = \frac{J_0 W L}{C_{\pi 0} W L} = \frac{J_0}{C_{\pi 0}}$$

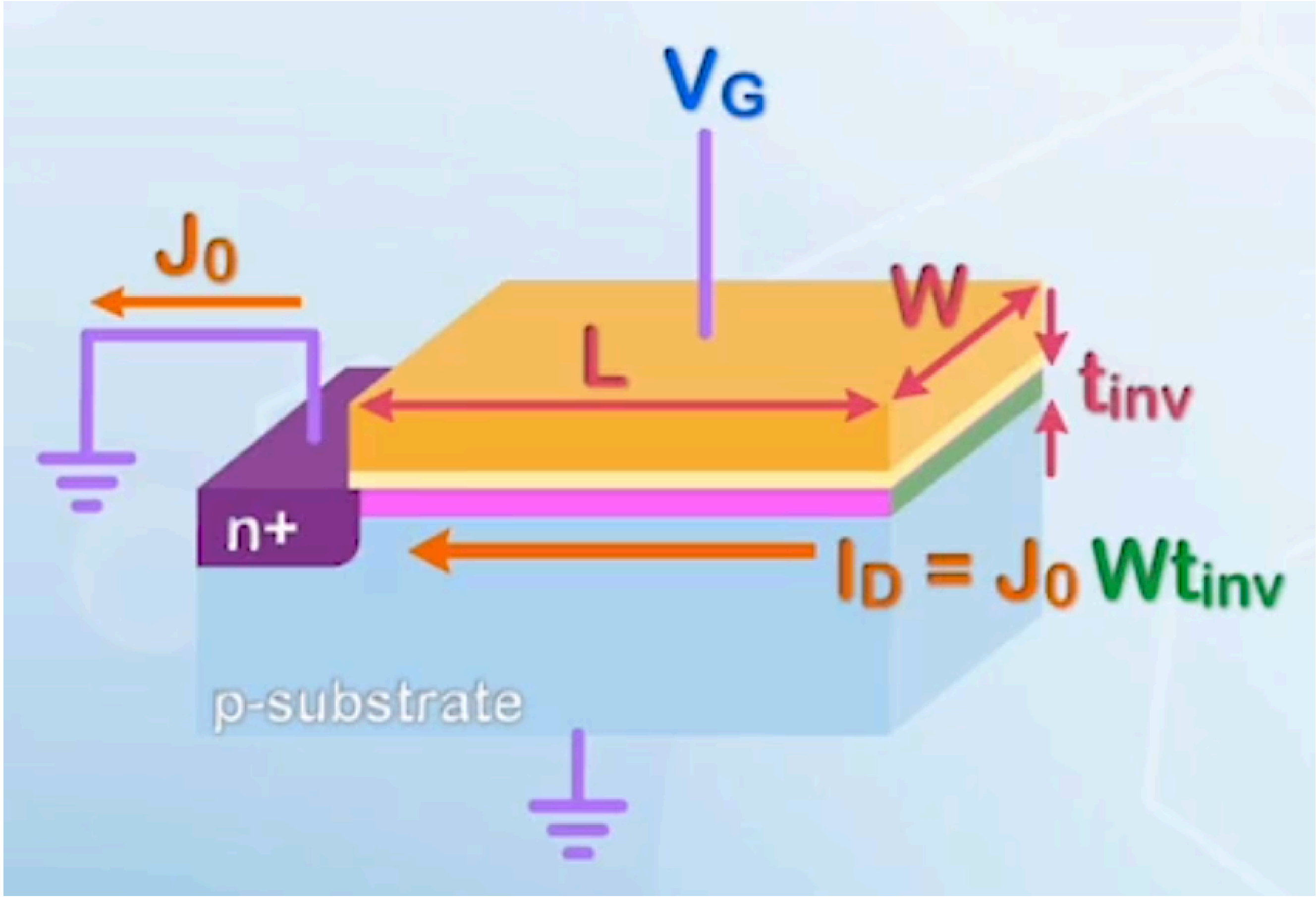
- Reducing the size of the BJT does not affect its speed, as both current and capacitance scale with area

- Consider a MOSFET driving itself



  - Assume it has the same current density J_0
  - The input capacitance is $C_{ox}WL$
  - However, the current flows horizontally through the channel, with a cross-sectional area of Wt_{inv}, where t_{inv} is the thickness of the inversion layer
  - The current is given by

$$I = J_0 W t_{\text{inv}}$$



- The speed:

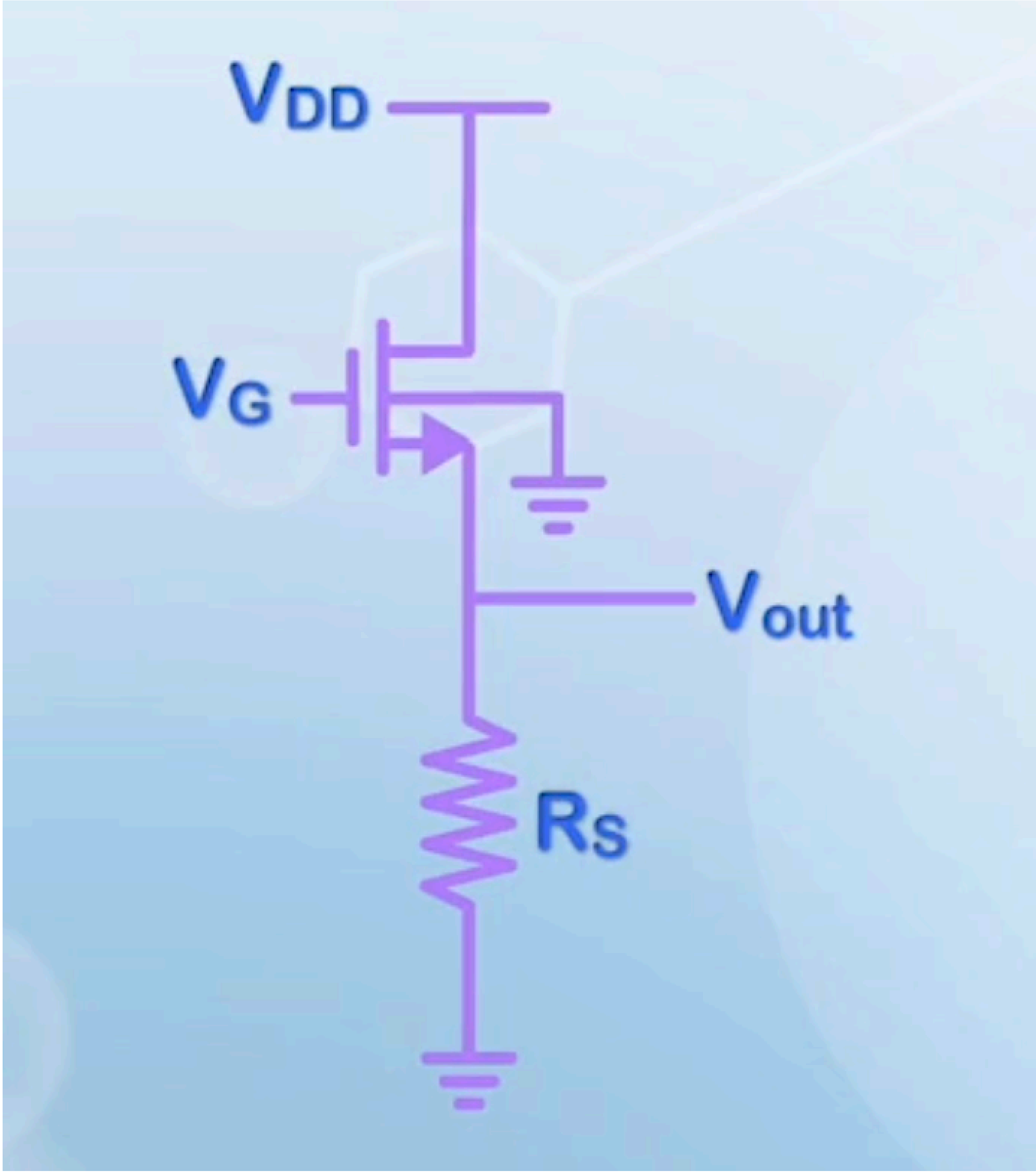
$$\text{speed} \sim \frac{I}{C} = \frac{J_0 W t_{\text{inv}}}{C_{ox} W L} = \frac{J_0 t_{\text{inv}}}{C_{ox} L}$$

- Because t_{inv} is in the order of a few nanometers, which is much smaller than L, it makes a MOSFET very slow to drive itself
- To increase the speed, we have to either increase the current drive, or decrease the loading capacitance
  - This is why we operate MOSFETs at a higher V_{G} to provide more current flow
- Early day MOSFETs operating in **subthreshold** region have such a low driving current that they are considered not usable for any meaningful applications
- When we scale down the MOSFET, L decreases, decreasing the capacitance and increasing the speed, making size reduction advantageous for MOSFETs
- In SOTA MOSFETs, L and t_{inv} are getting very close, enabling MOSFET circuits to be used in subthreshold region
- The increase in speed when scaling down is mainly contributed by the reduction of loading capacitance, instead of the increase in current, making MOSFETs suitable for integrated circuits with closely packed transistors and small parasitic capacitive loading

- When driving external elements with high capacitive load, BJTs with large cross-sectional area is still more desirable
  - This is why BJTs are more popular as a discrete element to function as a driver for large external loads

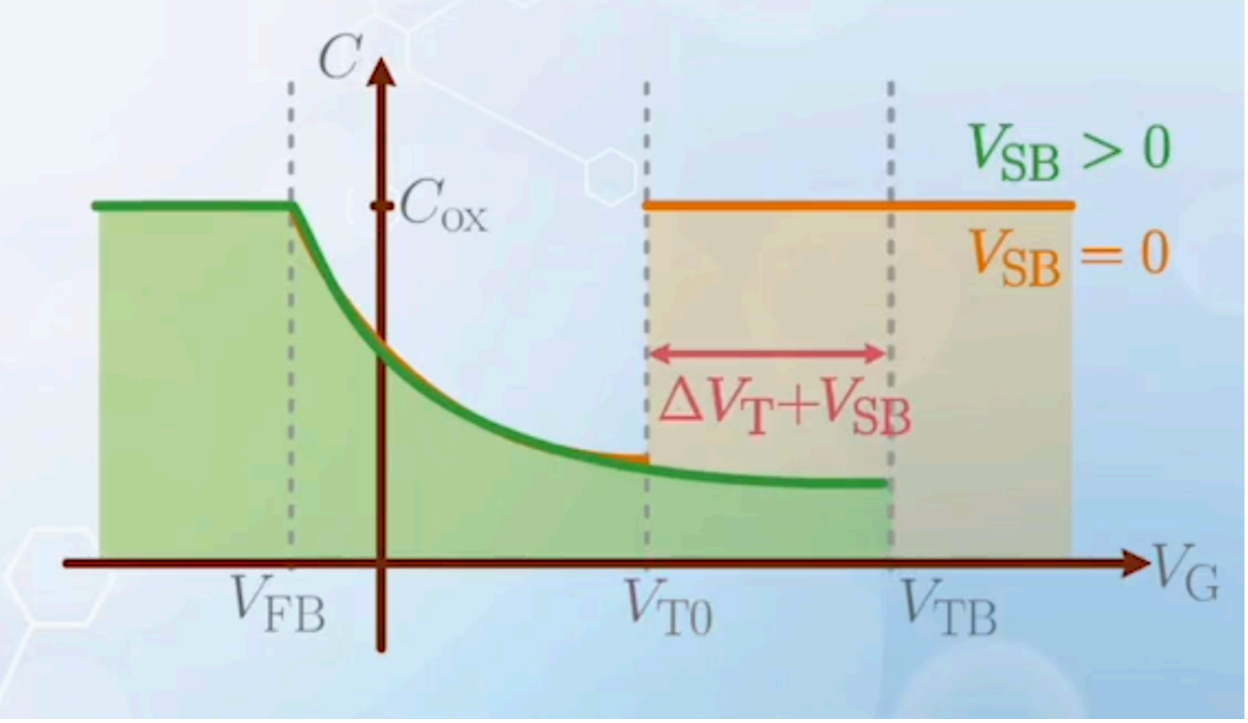
I-V Characteristics with Substrate Bias

Up to now, we have assumed that the source and substrate of a MOSFET are connected together and grounded. However, in some applications, like source follower circuits, source voltage may be higher than the substrate voltage, or effectively a negative substrate bias is applied to the MOSFET.



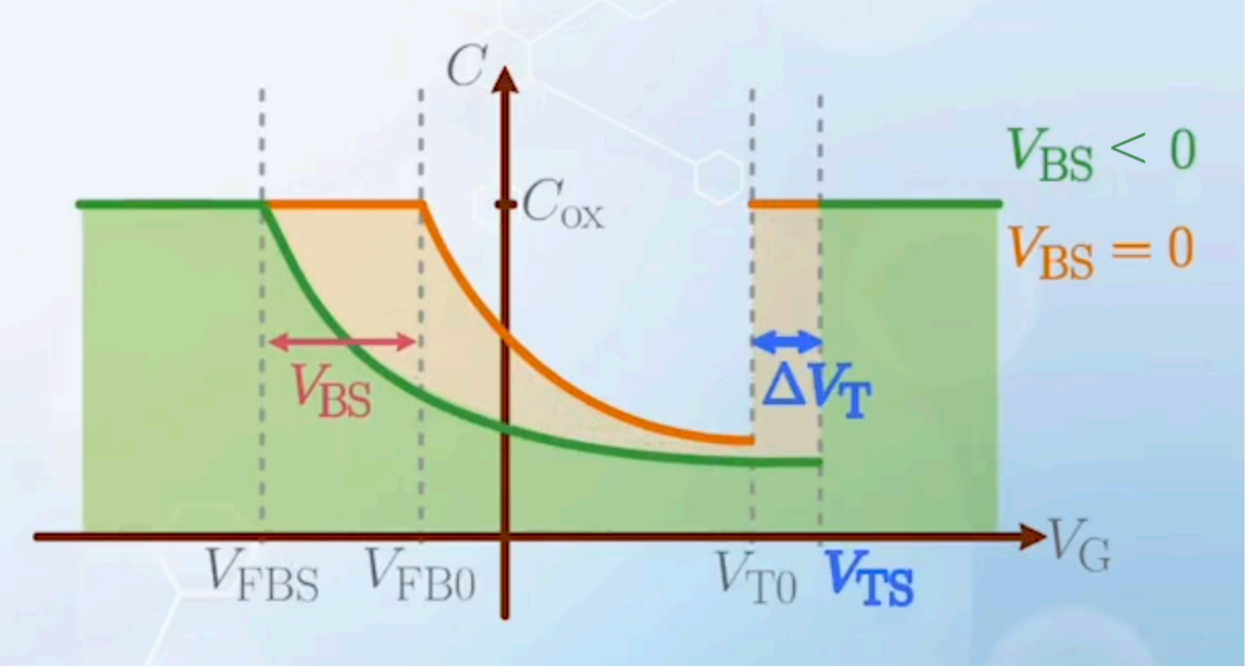
- When source and body voltages are different, we need to pick a reference
  - In the source follower circuit, we can pick the source voltage as reference, and V_{T} is labeled as V_{TB} with respect to the body voltage
    - V_{G}, V_{D}, and V_{S} are also labeled as V_{GB}, V_{DB}, and V_{SB} respectively, to indicate they are measured with respect to the body voltage

    - The $C - V$ characteristics now is



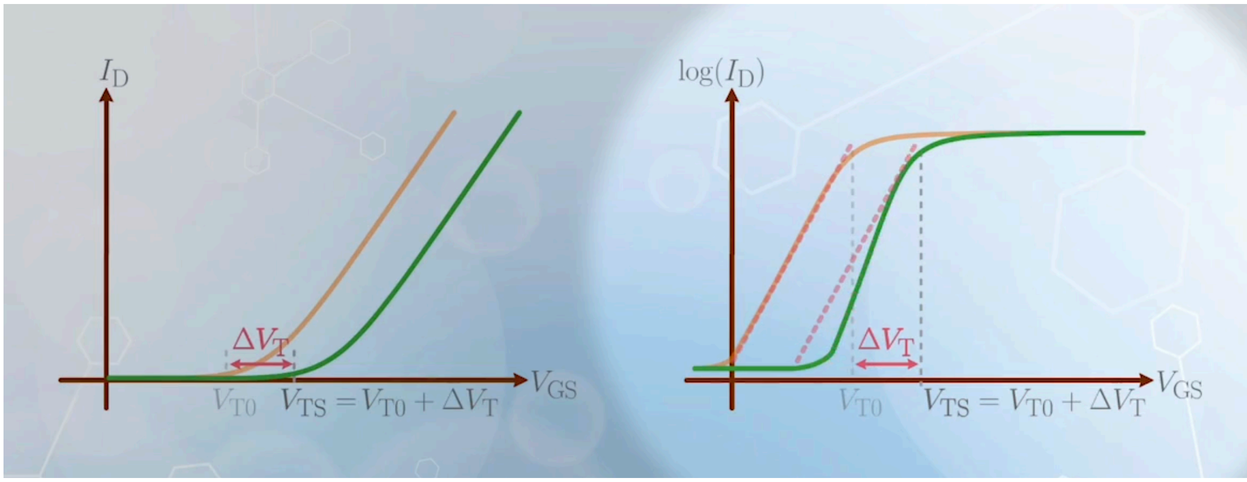
$$\Delta V_T = \gamma\left(\sqrt{2\phi_B + V_{SB}} - \sqrt{2\phi_B}\right)$$

- Or we can pick the source voltage as reference
    - Label V_T V_G, V_D, and V_B as V_{TS}, V_{GS}, V_{DS}, and V_{BS}
    - In MOSFETs, we are more interested in the inversion electrons in the channel, and these electrons come from the source, thus this reference is more commonly used
    - The $C - V$ characteristics now is



Its effect is mainly the change in V_T

- The $I-V$ and $\log I - V$ graph will be shifted right by ΔV_T



- As the capacitance with body bias in the depletion mode is smaller compared to no body bias, and the capacitance is a series of C_{ox} and C_D, thus C_D decreases (because the body voltage increases the depletion width at the same gate voltage).
  $n = 1 + C_\mathrm{D}/C_{ox}$, thus n decreases, and the subthreshold slope becomes steeper with negative body bias, but the lower limit of I_D may still be subjective to the drain junction leakage

Substrate Depletion Charge Effect

When deriving the current equations of MOSFETs, we have assumed V_T to be constant along the channel.

$$I_\mathrm{D} = Q(y)v(y)$$
$$Q(y) = C_{ox}W(V_\mathrm{G} - V_\mathrm{T} - V(y))$$

However, this is not true, and V_T is a function of the channel voltage. We may also consider the body effect:

$$V_\mathrm{T}(V(y)) = V_\mathrm{FB0} + 2\phi_\mathrm{B} + \gamma\sqrt{2\phi_\mathrm{B} + V(y) - V_\mathrm{B}}$$
$$\gamma = \frac{\sqrt{2qN_A\varepsilon_\mathrm{Si}}}{C_{ox}}$$

Putting it back to I_D, and follow the previous derivation steps, we have

$$\int_0^L I_\mathrm{D}\mathrm{d}y = \int_0^{V_\mathrm{D}} C_{ox}W\left(V_\mathrm{G} - \left(V_\mathrm{FB0} + 2\phi_\mathrm{B} + \gamma\sqrt{2\phi_\mathrm{B} + V(y) - V_\mathrm{B}}\right) - V(y)\right)\mu\mathrm{d}V$$
$$I_\mathrm{D} = \mu C_{ox}\frac{W}{L}\left((V_\mathrm{G} - V_\mathrm{FB0} - 2\phi_\mathrm{B})V_\mathrm{D} - \frac{V_\mathrm{D}^2}{2} - \frac{2}{3}\gamma\left((2\phi_\mathrm{B} - V_\mathrm{B} + V_\mathrm{D})^{3/2} - (2\phi_\mathrm{B} - V_\mathrm{B})^{3/2}\right)\right)$$

Again, we assume the range of $V(y)$ is limited, and we can approximate $V_\text{T}(V(y))$ as

$$\begin{aligned} V_\text{T}(V(y)) &\approx V_\text{FB0} + 2\phi_\text{B} + \gamma\sqrt{2\phi_\text{B} - V_\text{B}} + \frac{\gamma}{2\sqrt{2\phi_\text{B} - V_\text{B}}} V(y) \\ &= V_\text{TS} + \frac{\gamma V(y)}{2\sqrt{2\phi_\text{B} - V_\text{B}}} \end{aligned}$$

and

$$\begin{aligned} Q(y) &= C_{ox} W \left(V_\text{G} - V_\text{TS} - \left(1 + \frac{\gamma}{2\sqrt{2\phi_\text{B} - V_\text{B}}} \right) V(y) \right) \\ &= C_{ox} W (V_\text{G} - V_\text{TS} - \alpha V(y)) \\ \text{where } \alpha &= 1 + \frac{\gamma}{2\sqrt{2\phi_\text{B} - V_\text{B}}} \end{aligned}$$

Following the previous derivation steps, we can get

$$I_\text{D} = \mu C_{ox} \frac{W}{L} \left((V_\text{G} - V_\text{T}) V_\text{D} - \frac{\alpha V_\text{D}^2}{2} \right)$$

Usually, $1 < \alpha < 2$, and is called the **substrate / bulk charge factor**, which determines how strong V_T varies along the channel.

$$\begin{aligned} I_\text{D} &= \mu C_{ox} \frac{W}{L} \left((V_\text{G} - V_\text{T}) V_\text{D} - \frac{\alpha V_\text{D}^2}{2} \right) \\ &= \mu C_{ox} \frac{W}{L} \left(V_\text{G} - (V_\text{T} + \frac{\alpha - 1}{2} V_\text{D}) - \frac{V_\text{D}}{2} \right) V_\text{D} \\ &= \mu C_{ox} \frac{W}{L} \left(V_\text{G} - V_\text{T(new)} - \frac{V_\text{D}}{2} \right) V_\text{D} \\ \text{where } V_\text{T(new)} &= V_\text{T} + \frac{\alpha - 1}{2} V_\text{D} \end{aligned}$$

which means the equation assumes V_T increases linearly with V_D.

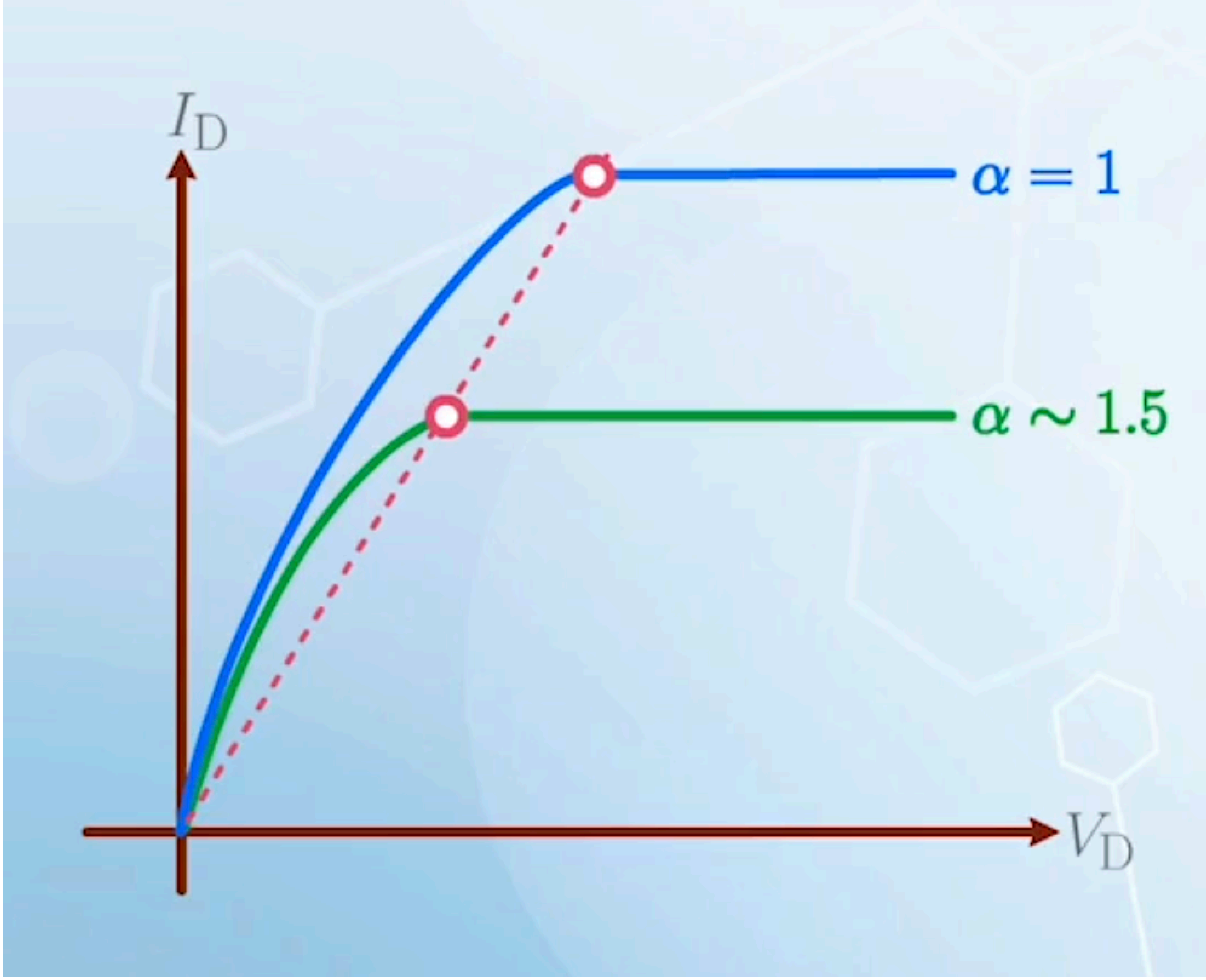
After modifying the linear region, we also need to modify the saturation region, just by finding the peek of the quadratic equation.

$$I_\text{Dsat} = \frac{1}{2\alpha} \mu C_{ox} \frac{W}{L} (V_\text{G} - V_\text{T})^2$$

$$V_\text{Dsat} = \frac{V_\text{G} - V_\text{T}}{\alpha}$$

In traditional long channel transistors, $\alpha \approx 1.5$. By ignoring the V_{T} variations along the channel, we may overestimate the current by 50%

No matter the value of α, the point $(V_{\mathrm{Dsat}}, I_{\mathrm{Dsat}})$ lies on the same straight line joining the origin and the point when $\alpha = 1$, because V_{Dsat} and I_{Dsat} are both scaled down by α.



The change in V_{T} only affects the current calculations, but not the turn-on characteristics, as they are determined by the V_{T} at the source.

7. Mobility Degradation and Carrier Velocity Saturation

About mobility degradation in MOSFETs, and the carrier velocity saturation model.

Effect of Gate Voltage on Carrier Motion

The previous given equations to calculate I_D have assumed that the carrier mobility μ is a constant and bias independent. However, for carriers moving near the interface between the bulk silicon and the gate oxide.

Electrons moving in the channel are under the influence of two electric fields: the vertical field from the gate, attracting it to move vertically towards the gate oxide, and lateral field from the drain, attracting it to move laterally towards the drain.

The electron will bounce off the silicon and oxide interface a few times before reaching the drain. If the vertical field is strong, electrons will bounce more times off the surface. The collision between the electron and the interface is inelastic, causing energy loss and reducing the velocity of the electron. Therefore, it will take longer time for electrons to reach the drain.

Thus, the mobility μ is not a constant, and decreases with an increasing vertical field, or increasing V_G. This is the **mobility degradation** due to vertical electric field.

To correct for this effect, μ need to be modified as $\mu(V_\mathrm{G})$.

The Effective Vertical Electric Field

To quantify the variation of μ caused by V_G, we need to calculate the vertical electric field.

Electrons at different distance from the interface will experience different vertical electric field. To simplify the problem, the average electric field is used, assuming it to be the electric field experienced by all electrons.

- x: the distance from the interface into the bulk silicon, in the vertical direction
- $n(x)$: the electron concentration at location x
- $Ex(x)$: the vertical electric field component E_x at location x
- The average vertical electric field is

$$\begin{aligned} E_\mathrm{eff} &= \frac{\int n(x) E_x(x) \mathrm{d}x}{\int n(x) \mathrm{d}x} \\ &= \frac{Q_\mathrm{inv}}{2\varepsilon_\mathrm{Si}} + \frac{Q_\mathrm{B}}{\varepsilon_\mathrm{Si}} \end{aligned}$$

where Q_{inv} is the inversion charge density, and Q_{B} is the depletion charge density, relabeled from Q_{D} in the MOSFET capacitance section to avoid confusion with the drain charge

  - An intuitive derivation of this equation:
  - The average electric field can be considered the electric field experienced by an average electron
  - **The average electron** is the one with half of the electrons in the channel above it, and half below it
  - The electric field starts at a positive charge at the game, and terminates at a negative charge at the substrate
  - Electric field terminating above the average electron will not be experienced by the average electron
  - Therefore, only the electric field terminating **below** the average electron will be experienced by the average electron
  - Once the charge Q terminating the electric field below the average electron is known, the electric field can be calculated as $E = \frac{Q}{\varepsilon_{\mathrm{Si}}}$
  - The charge Q is comprised of two parts:
    - **Half of the inversion charge Q_{inv}**, since only half of the electrons are below the average electron
    - The **entire depletion charge Q_{B}**, since the inversion layer is very thin compared to the depletion region
  - Thus the equation is obtained
- We need to further express E_{eff} as a function of V_{G}
  - The inversion charge density is

$$Q_{\mathrm{inv}} = C_{\mathrm{ox}}(V_{\mathrm{G}} - V_{\mathrm{T}})$$

  - The depletion charge can be calculated with the following equation

$$V_{\mathrm{T}} = V_{\mathrm{FB}} + 2\phi_{\mathrm{B}} + \frac{Q_{\mathrm{B}}}{C_{\mathrm{ox}}}$$

    as the threshold voltage V_{T} can be obtained from measurements rather than calculation

$$\begin{aligned} Q_{\mathrm{B}} &= C_{\mathrm{ox}}(V_{\mathrm{T}} - (V_{\mathrm{FB}} + 2\phi_{\mathrm{B}})) \\ \text{approximating } \phi_{\mathrm{B}} &= 0.35\,\mathrm{V} \\ \text{and } V_{\mathrm{FB}} &= -(0.55\,\mathrm{V} + \phi_{\mathrm{B}}) \\ \Rightarrow V_{\mathrm{FB}} + 2\phi_{\mathrm{B}} &= -0.2\,\mathrm{V} = -V_{\mathrm{a}} \\ Q_{\mathrm{B}} &= C_{ox}(V_{\mathrm{T}} + V_{\mathrm{a}}) \end{aligned}$$

V_a is dependent of gate material and substrate doping

- The effective vertical electric field can now be expressed as

$$\begin{aligned} E_\mathrm{eff} &= \frac{C_{ox}(V_\mathrm{G} - V_\mathrm{T})}{2\varepsilon_\mathrm{Si}} + \frac{C_{ox}(V_\mathrm{T} + V_\mathrm{a})}{\varepsilon_\mathrm{Si}} \\ &= \frac{\varepsilon_{ox}(V_\mathrm{G} - V_\mathrm{T})}{2\varepsilon_\mathrm{Si} t_{ox}} + \frac{\varepsilon_{ox}(V_\mathrm{T} + V_\mathrm{a})}{\varepsilon_\mathrm{Si} t_{ox}} \\ \text{approx. } \varepsilon_\mathrm{Si} &= 11.9\varepsilon_0 \approx 3\varepsilon_{ox} = 3 \times 3.9\varepsilon_0 \\ \Rightarrow E_\mathrm{eff} &= \frac{(V_\mathrm{G} - V_\mathrm{T})}{6t_{ox}} + \frac{(V_\mathrm{T} + V_\mathrm{a})}{3t_{ox}} \end{aligned}$$

  where $V_\mathrm{a} = 0.2\,\mathrm{V}$

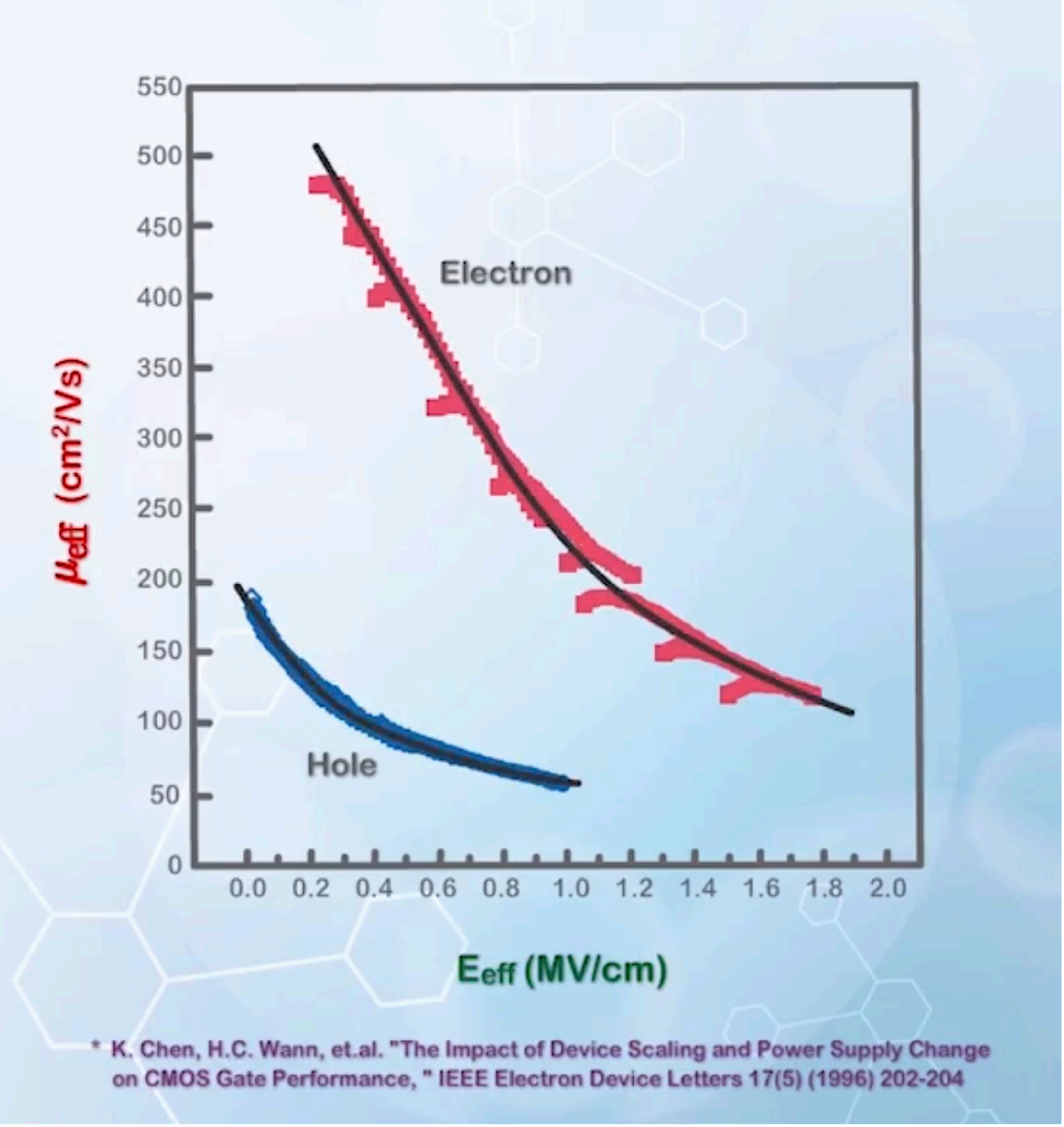
- Once V_G, V_T and t_{ox} is known, the effective vertical electric field can be calculated

Calculating Effective Mobility

We still need to obtain the relation between μ_eff and E_eff.

There are many theories predicting the effective mobility based on the microscopic effects, but none of them fit the experimental data well.

In practice, engineers rely on experimental calibration to obtain the effective mobility. Experimental data are collected are plotted:



* K. Chen, H.C. Wann, et.al. "The Impact of Device Scaling and Power Supply Change on CMOS Gate Performance, " IEEE Electron Device Letters 17(5) (1996) 202-204

Regardless of the gate oxide thickness and the substrate doping concentrations, they all fall onto a single curve.

As the results are very consistent, an empirical equation is more practical to use, rather than complex theoretical models.

The empirical equation is given as

$$\mu_{\text{eff}} = \frac{\mu_0}{1 + (E_{\text{eff}}/E_0)^{\nu}}$$

and a widely used set of parameters for this model is

	Electrons	Holes
$\mu_0 \quad (\text{cm}^2/(\text{V} \cdot \text{s}))$	670	160
$E_0 \quad (\text{MV}/\text{cm})$	0.67	0.7
ν	1.6	1.0

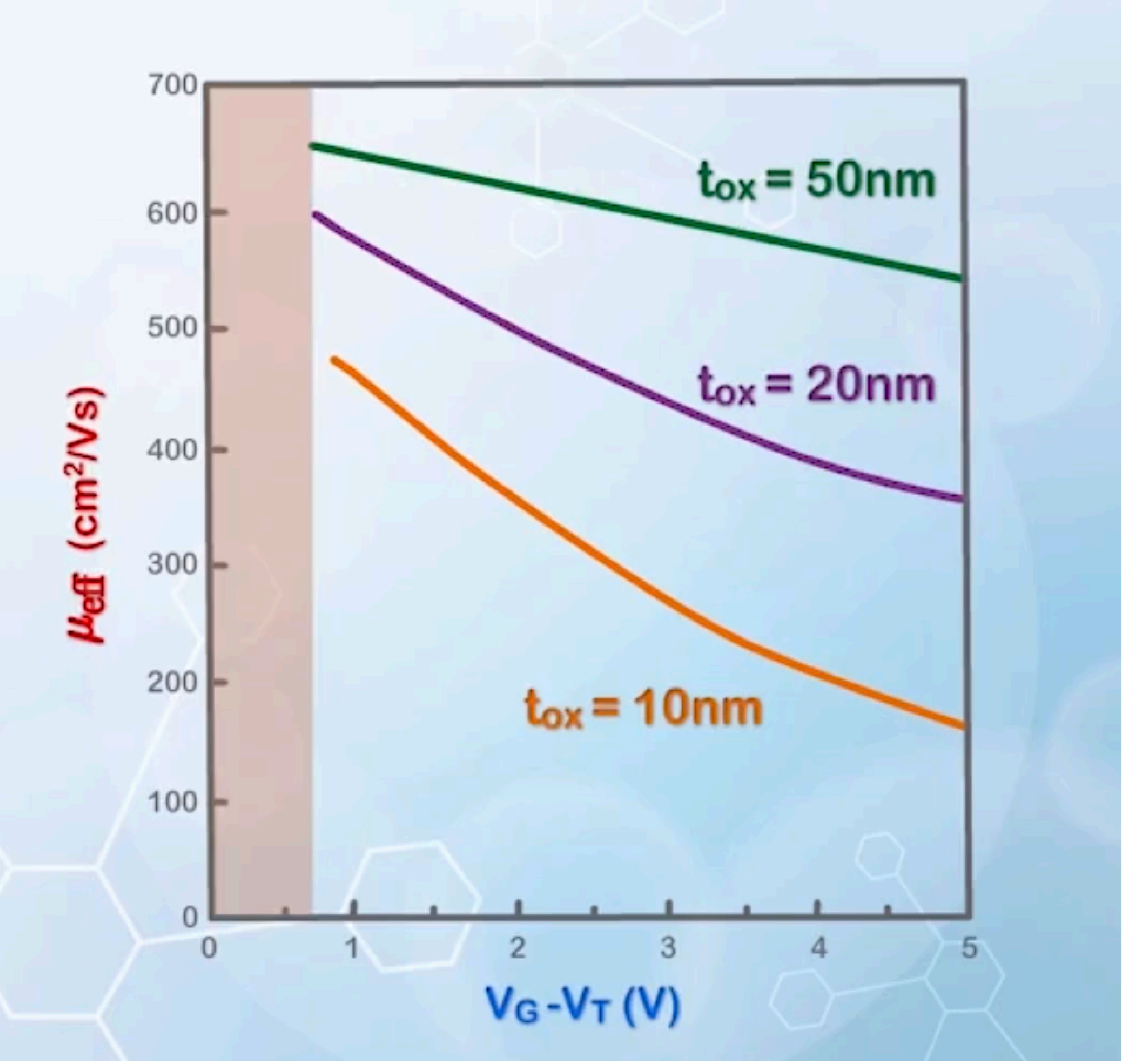
Some other sets of parameters are also used, due to variations in the fabrication process and the physical structure.

This is the **universal mobility model**, as it can fit different sets of data very well.

The relationship between μ_{eff} and $V_{\text{G}} - V_{\text{T}}$ can be expressed as

$$\mu_{\text{eff}} = \frac{\mu_0}{1 + \left(\frac{(V_{\text{G}} - V_{\text{T}})}{6E_0 t_{ox}} + \frac{(V_{\text{T}} + V_{\text{a}})}{3E_0 t_{ox}}\right)^{\nu}}$$

Now we can plot μ_{eff} versus V_{G} for different oxide thicknesses:



Note that the model may not be accurate for V_G close to V_T, as we use $C_\mathrm{ox}(V_\mathrm{G} - V_\mathrm{T})$ to approximate the inversion charge, giving 0 at $V_\mathrm{G} = V_\mathrm{T}$, which is not true, as discussed in previous sections. But for most of the part, it is valid.

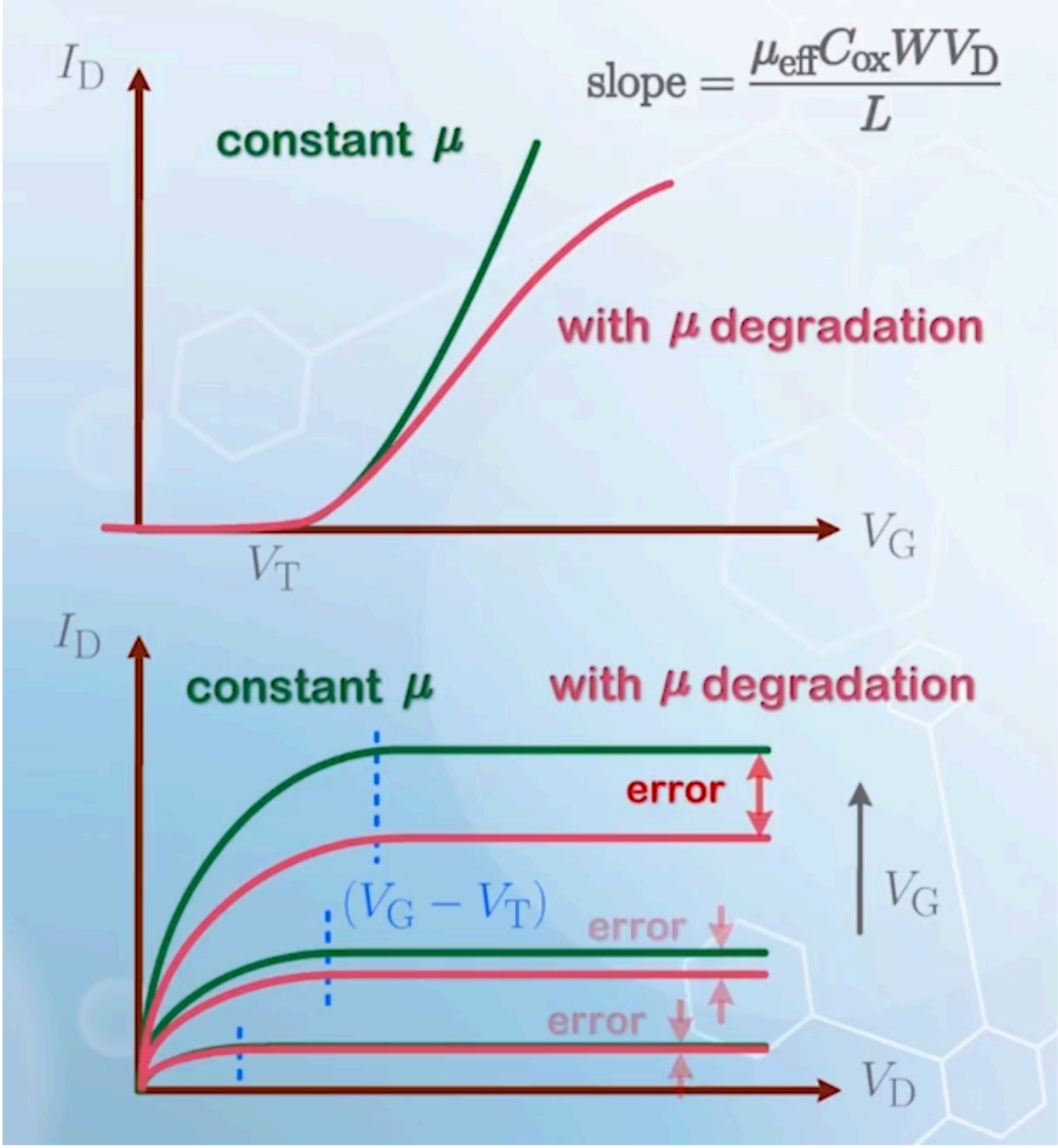
From the graph, we can see that when t_{ox} is large, assuming μ to be a constant won't introduce significant error. But when t_{ox} is small, using a constant μ will introduce an error of more than 50% at high or low V_G values. This is because for a given power supply voltage V_dd, the range of V_G is fixed, and according to the equation of E_eff

$$E_\mathrm{eff} = \frac{(V_\mathrm{G} - V_\mathrm{T})}{6t_{ox}} + \frac{(V_\mathrm{T} + V_\mathrm{a})}{3t_{ox}}$$

thinner oxide thickness will lead to larger E_{eff} variation for the same V_{G} variation, causing larger μ_{eff} variation.

Therefore, the mobility degradation effect may not be important for older MOSFETs with thick oxide, but it is very important for modern MOSFETs with thin oxide.

The $I_{\text{D}} - V_{\text{G}}$ and $I_{\text{D}} - V_{\text{D}}$ characteristics considering mobility degradation is shown below:



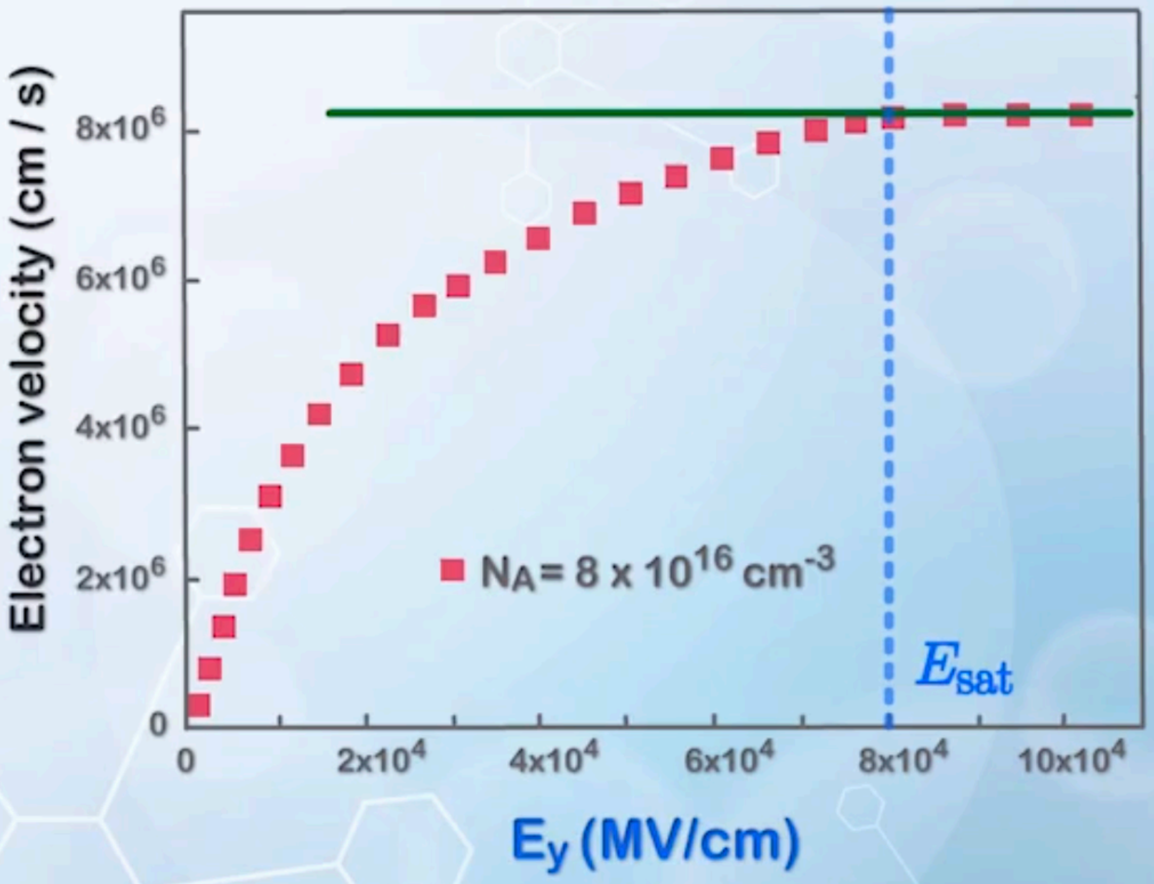
Compared to constant μ, mobile degradation will cause a downward bending in the $I_{\mathrm{D}} - V_{\mathrm{G}}$ curve, as opposed to the linear increase in the constant μ case. And in the $I_{\mathrm{D}} - V_{\mathrm{D}}$ curve, all current will be overestimated if mobility degradation is not considered, with larger error at higher V_{G}.

Carrier Velocity Saturation Model

Up to now, we have linearly related the carrier velocity v to the electric field E

$$v = \mu_{\mathrm{eff}} E$$

E is given by $\frac{V_{\mathrm{D}}}{d}$, and can be very large when d is small enough, without a theoretical limit. The maximum achievable electric field is determined practically by the strength of the medium, and a discharge will occur if the medium cannot withstand the electric field. However, the speed of carriers is bounded by the speed of light, according to the theory of relativity. Therefore, the linearity between v and E cannot hold forever, and the velocity will saturate with a high enough E. This is **carrier velocity saturation**.



* F. Assaderaghi, D. Sinitsky, J. Bokor, P.K.Ko, H. Gaw and C. Hu, "High-Field Transport of Inversion-Layer Electrons and Holes Including Velocity Overshoot", IEEE Transaction on Electron Devices, Vol. 44, no. 4, April 1997, pp. 664-671

From measurements, carrier velocity saturates at a certain electric field E_{sat}, and the saturated velocity for electrons and holes are approximately

$$\begin{aligned} \text{electrons} \quad v_{\text{sat}} &\approx 8 \times 10^6 \text{ cm/s} \\ \text{holes} \quad v_{\text{sat}} &\approx 6 \times 10^6 \text{ cm/s} \end{aligned}$$

To include this effect in drain current calculation, we need to derive a new expression for carrier velocity $v = f(E_{\text{y}})$ with respect to the lateral electric field E_{y}.

The easiest way is to use a straight line to connect the two known points: (0,0) and $(E_{\text{sat}}, v_{\text{sat}})$, giving

$$v = \begin{cases} \mu E_y & E_y < E_{\text{sat}} \\ v_{\text{sat}} & E_y \geq E_{\text{sat}} \end{cases}$$

To be more accurate, the slope of the curve, or mobility, should decrease when the lateral electric field increases, so a better expression would be

$$v = \begin{cases} \frac{\mu_\text{eff}}{1+\frac{E_y}{E_\text{sat}}} E_y & E_y < E_\text{sat} \\ v_\text{sat} & E_y \geq E_\text{sat} \end{cases}$$

The slope will be decreasing with increasing E_y

$$\begin{aligned} \text{slope} &= \frac{\mathrm{d}v}{\mathrm{d}E_y} \\ &= \frac{\mu_\text{eff} E_\text{sat}^2}{(E_\text{sat} + E_y)^2} \\ \text{slope}|_{E_y=0} &= \mu_\text{eff} \\ \text{slope}|_{E_y=E_\text{sat}} &= \frac{\mu_\text{eff}}{4} \end{aligned}$$

providing better fitting for the data.

We have the measured v_sat, so E_sat can be calculated as

$$\begin{aligned} v_\text{sat} &= \frac{\mu_\text{eff}}{2} E_\text{sat} \\ \Rightarrow E_\text{sat} &= \frac{2 v_\text{sat}}{\mu_\text{eff}} \end{aligned}$$

where μ_eff is given by

$$\mu_\text{eff} = \frac{\mu_0}{1 + (E_\text{eff}/E_0)^\nu}$$

To derive the new drain current equation considering velocity saturation

$$\begin{aligned} I_\text{D} &= Q(y) v(y) \\ &= C_{ox} W (V_\text{G} - V_\text{T} - V(y)) \frac{\mu_\text{eff}}{1 + \frac{\mathrm{d}V/\mathrm{d}y}{E_\text{sat}}} \frac{\mathrm{d}V}{\mathrm{d}y} \\ I_\text{D} \mathrm{d}y &= \left(\mu_\text{eff} C_{ox} W (V_\text{G} - V_\text{T} - V(y) - \frac{I_\text{D}}{E_\text{sat}} \right) \mathrm{d}V \\ \int_0^L I_\text{D} \mathrm{d}y &= \int_{V_\text{S}}^{V_\text{D}} \left(\mu_\text{eff} C_{ox} W (V_\text{G} - V_\text{T} - V(y) - \frac{I_\text{D}}{E_\text{sat}} \right) \mathrm{d}V \\ I_\text{D} L &= \mu_\text{eff} C_{ox} W \left[(V_\text{G} - V_\text{T}) V_\text{D} - \frac{1}{2} V_\text{D}^2 \right] - \frac{I_\text{D}}{E_\text{sat}} V_\text{D} \\ I_\text{Dlin} &= \frac{\mu_\text{eff} C_{ox} \frac{W}{L} \left[(V_\text{G} - V_\text{T}) V_\text{D} - \frac{1}{2} V_\text{D}^2 \right]}{1 + \frac{V_\text{D}}{E_\text{sat} L}} \end{aligned}$$

Compared to the classical model, there is just an additional factor of $\frac{1}{1+\frac{V_\text{D}}{E_\text{sat}L}}$.

As for the saturation region, we define V_{Dsat} as the drain voltage that causes the lateral electric field near the drain to reach E_{sat}

$$I_{\text{Dsat}} = WC_{ox}(V_{\text{G}} - V_{\text{T}} - V_{\text{Dsat}})v_{\text{sat}}$$

V_{Dsat} is the point where $I_{\text{Dlin}} = I_{\text{Dsat}}$

$$\begin{aligned}
\frac{\mu_{\text{eff}}C_{ox}\frac{W}{L}\left[(V_{\text{G}} - V_{\text{T}})V_{\text{Dsat}} - \frac{1}{2}V_{\text{Dsat}}^2\right]}{1 + \frac{V_{\text{Dsat}}}{E_{\text{sat}}L}} &= WC_{ox}(V_{\text{G}} - V_{\text{T}} - V_{\text{Dsat}})v_{\text{sat}} \\
\mu_{\text{eff}}E_{\text{sat}}\left[(V_{\text{G}} - V_{\text{T}})V_{\text{Dsat}} - \frac{1}{2}V_{\text{Dsat}}^2\right] &= (LE_{\text{sat}} + V_{\text{Dsat}})(V_{\text{G}} - V_{\text{T}} - V_{\text{Dsat}})v_{\text{sat}} \\
\mu_{\text{eff}}E_{\text{sat}}\left[(V_{\text{G}} - V_{\text{T}})V_{\text{Dsat}} - \frac{1}{2}V_{\text{Dsat}}^2\right] &= (LE_{\text{sat}} + V_{\text{Dsat}})(V_{\text{G}} - V_{\text{T}} - V_{\text{Dsat}})\frac{\mu_{\text{eff}}E_{\text{sat}}}{2} \\
2(V_{\text{G}} - V_{\text{T}})V_{\text{Dsat}} &= LE_{\text{sat}}(V_{\text{G}} - V_{\text{T}}) + V_{\text{Dsat}}(V_{\text{G}} - V_{\text{T}} - LE_{\text{sat}}) \\
\Rightarrow V_{\text{Dsat}} &= \frac{(V_{\text{G}} - V_{\text{T}})LE_{\text{sat}}}{V_{\text{G}} - V_{\text{T}} + LE_{\text{sat}}}
\end{aligned}$$

Carrier Velocity Saturation v.s. Pinchoff

Comparing the two set of equations

- The pinchoff model

$$\begin{aligned}
I_{\text{Dlin}} &= \mu_{\text{eff}}C_{ox}\frac{W}{L}\left[(V_{\text{G}} - V_{\text{T}})V_{\text{D}} - \frac{V_{\text{D}}^2}{2}\right] \\
I_{\text{Dsat}} &= \frac{1}{2}\mu_{\text{eff}}C_{ox}\frac{W}{L}(V_{\text{G}} - V_{\text{T}})^2 \\
V_{\text{Dsat}} &= V_{\text{G}} - V_{\text{T}}
\end{aligned}$$

- The carrier velocity saturation model

$$\begin{aligned}
I_{\text{Dlin}} &= \frac{\mu_{\text{eff}}C_{ox}\frac{W}{L}\left[(V_{\text{G}} - V_{\text{T}})V_{\text{D}} - \frac{V_{\text{D}}^2}{2}\right]}{1 + \frac{V_{\text{D}}}{E_{\text{sat}}L}} \\
I_{\text{Dsat}} &= WC_{ox}(V_{\text{G}} - V_{\text{T}} - V_{\text{Dsat}})v_{\text{sat}} \\
V_{\text{Dsat}} &= \frac{(V_{\text{G}} - V_{\text{T}})LE_{\text{sat}}}{V_{\text{G}} - V_{\text{T}} + LE_{\text{sat}}}
\end{aligned}$$

- In the **linear region**, the two models are very similar, with just an additional factor in the carrier velocity saturation model that reduces the increase in I_{D} in the linear region
- The new V_{Dsat} is similar to have two resistors of LE_{sat} and $V_{\text{G}} - V_{\text{T}}$ in parallel, which will be smaller than both of them

- Carrier velocity saturation will kick in before pinchoff occurs, preventing pinchoff from happening
- This is because the carrier velocity saturation limits the maximum carrier velocity, preventing the carrier density from reaching zero at the drain end, as $I_\mathrm{D} = Qv$ is a constant
- This removes the inconsistency in the pinchoff model that requires infinite carrier velocity

The problem of the pinchoff model is that it does not provide a clear physical image for the ending of the gradual channel region, and we arbitrarily picked $V_\mathrm{G} - V_\mathrm{T}$ as the pinchoff point.

By providing a more physical definition for the end of the gradual channel region, the carrier velocity saturation model avoids reaching the pinchoff condition, and thus solved the problem.

Also, the saturation current in the carrier velocity saturation model is also physically derived, instead of extending the flat portion of the linear region equation of the drain current, and I_Dsat predicted by the carrier velocity saturation model is always smaller than that predicted by the pinchoff model.

The same substrate charge factor α that considers the effect of V_T along the channel, can also be added to the equations from the velocity saturation model, giving

$$\begin{aligned} I_\mathrm{Dlin} &= \frac{\mu_\mathrm{eff} C_{ox} \frac{W}{L}\left[(V_\mathrm{G} - V_\mathrm{T})V_\mathrm{D} - \frac{\alpha V_\mathrm{D}^2}{2}\right]}{1 + \frac{V_\mathrm{D}}{E_\mathrm{sat} L}} \\ I_\mathrm{Dsat} &= W C_{ox} (V_\mathrm{G} - V_\mathrm{T} - \alpha V_\mathrm{Dsat}) v_\mathrm{sat} \\ V_\mathrm{Dsat} &= \frac{(V_\mathrm{G} - V_\mathrm{T}) L E_\mathrm{sat}}{V_\mathrm{G} - V_\mathrm{T} + \alpha L E_\mathrm{sat}} \end{aligned}$$

8. I-V Characteristics With Carrier Velocity Saturation Model

About the comparisons between the pinchoff model and the carrier velocity saturation model for MOSFET I-V characteristics, with varying channel length, gate voltage, and gate oxide thickness.

Carrier Velocity Saturation in Long Channel MOSFET

The carrier velocity saturation model is more accurate in terms of physics, but the pinchoff model is equivalent to the velocity saturation model under special circumstances, and the velocity saturation model can also be converted to the pinchoff model when the channel is very long.

- In the linear region, consider the carrier velocity saturation model:

  $$I_{\mathrm{Dlin}} = \frac{\mu_{\mathrm{eff}} C_{ox} \frac{W}{L} \left[(V_{\mathrm{G}} - V_{\mathrm{T}}) V_{\mathrm{D}} - \frac{V_{\mathrm{D}}^2}{2} \right]}{1 + \frac{V_{\mathrm{D}}}{E_{\mathrm{sat}} L}}$$

  When L is very large, the $\frac{V_{\mathrm{D}}}{E_{\mathrm{sat}} L}$ term can be ignored, and the equation reduces to the pinchoff model in the linear region
- Same thing applies to V_{Dsat}. With carrier velocity saturation model:

  $$V_{\mathrm{Dsat}} = \frac{(V_{\mathrm{G}} - V_{\mathrm{T}}) E_{\mathrm{sat}} L}{(V_{\mathrm{G}} - V_{\mathrm{T}}) + E_{\mathrm{sat}} L}$$

  This is the parallel combination of $(V_{\mathrm{G}} - V_{\mathrm{T}})$ and $E_{\mathrm{sat}} L$. When L is very large, $E_{\mathrm{sat}} L$ becomes very large, and V_{Dsat} approaches $(V_{\mathrm{G}} - V_{\mathrm{T}})$, which is the same as the pinchoff model.
- In the saturation region, consider the carrier velocity saturation model:

  $$\begin{aligned} I_{\mathrm{Dsat}} &= W C_{ox} (V_{\mathrm{G}} - V_{\mathrm{T}} - V_{\mathrm{Dsat}}) v_{\mathrm{sat}} \\ &= W C_{ox} v_{\mathrm{sat}} (V_{\mathrm{G}} - V_{\mathrm{T}} - \frac{(V_{\mathrm{G}} - V_{\mathrm{T}}) E_{\mathrm{sat}} L}{(V_{\mathrm{G}} - V_{\mathrm{T}}) + E_{\mathrm{sat}} L}) \\ &= W C_{ox} v_{\mathrm{sat}} \frac{(V_{\mathrm{G}} - V_{\mathrm{T}})^2}{(V_{\mathrm{G}} - V_{\mathrm{T}}) + E_{\mathrm{sat}} L} \end{aligned}$$

  When L is very large, $E_{\mathrm{sat}} L$ becomes very large, and the equation reduces to

  $$I_{\mathrm{Dsat}} = W C_{ox} v_{\mathrm{sat}} \frac{(V_{\mathrm{G}} - V_{\mathrm{T}})^2}{E_{\mathrm{sat}} L}$$

  also, we have

  $$v_{\mathrm{sat}} = \frac{\mu_{\mathrm{eff}} E_{\mathrm{sat}}}{2}$$

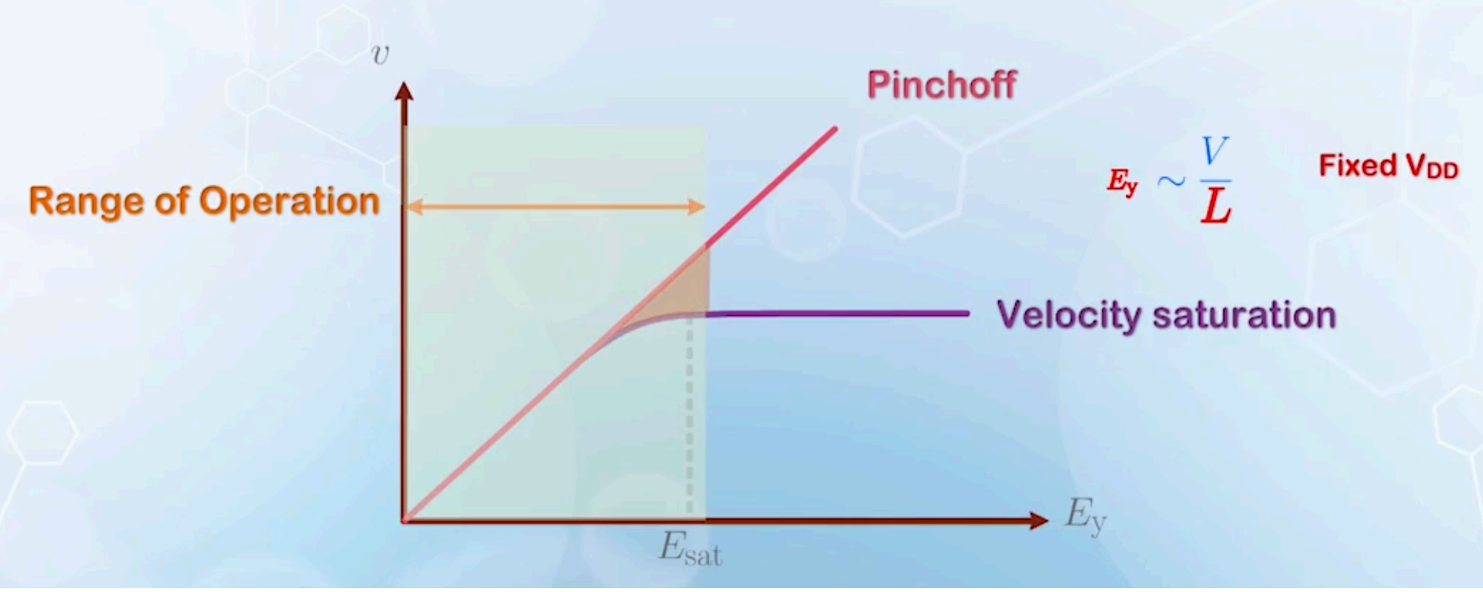
  therefore,

  $$I_{\mathrm{Dsat}} = \frac{\mu_{\mathrm{eff}} C_{ox} \frac{W}{L} (V_{\mathrm{G}} - V_{\mathrm{T}})^2}{2}$$

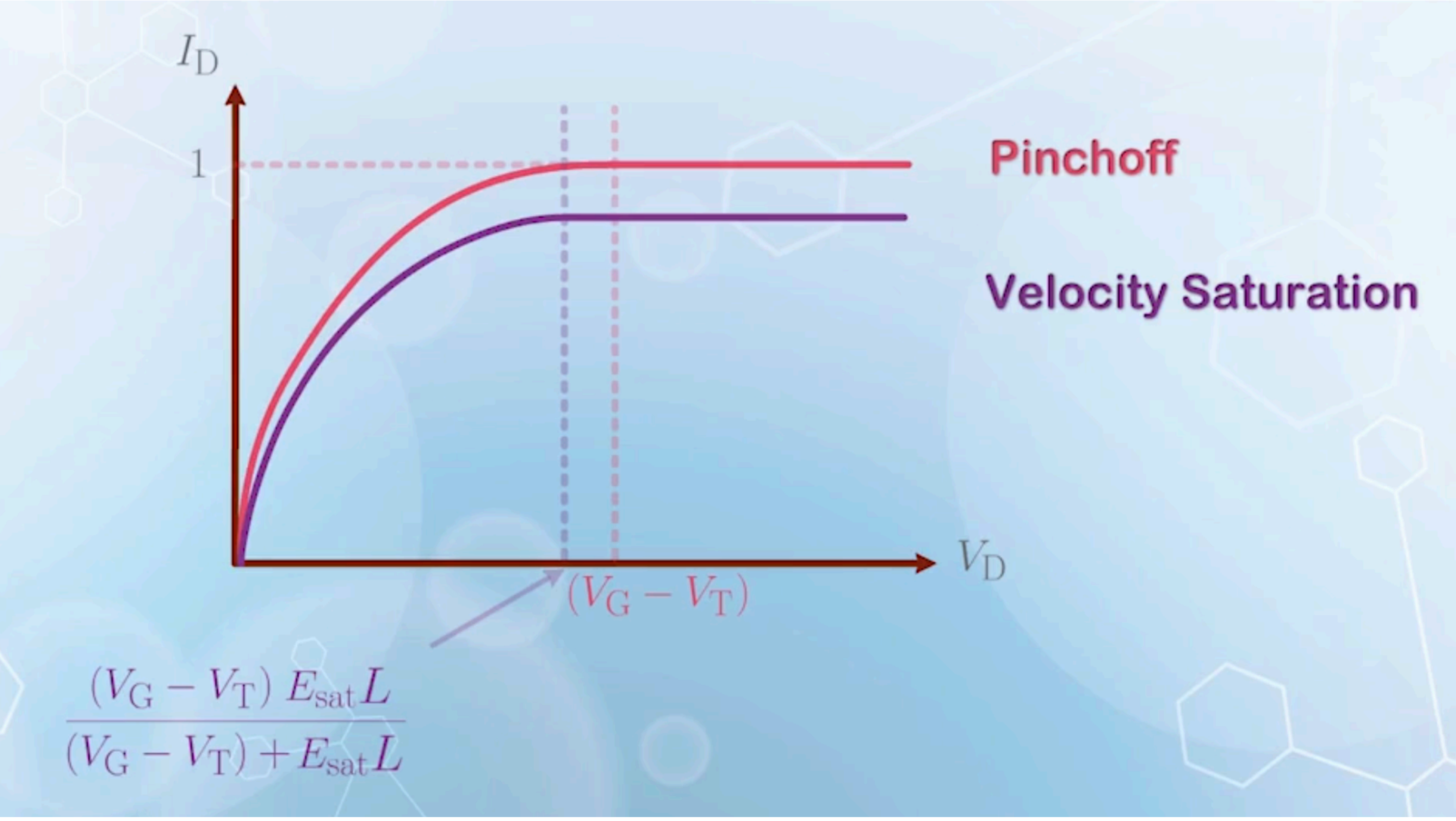
  This is the same of the pinchoff model in the saturation region.

Why this happens? Consider the $v - E_y$ characteristics of the two models. In the region where v is almost proportional to E_y, the two models are very similar. Errors only occur when E_y is large enough to make v approach v_{sat}.

As E_y is given by $\frac{V}{L}$, the range of operation of a MOSFET is limited to the region with a smaller E_y for a longer L. Then the errors caused by the difference between the two models are limited.



Now consider the $I_D - V_D$ characteristics. When $E_{\text{sat}}L$ is small, V_{Dsat} is determined by the E_{Dsat} term in the parallel model, which will be very small. When L increases, V_{Dsat} increases, and the current can further increase before entering saturation. When $L \to \infty$, the two curves are the same.



Characteristics With Size Reduction

To determine whether L can be considered long enough to use the pinchoff model, we need to consider L as well as the range of operation voltages.

The L term always appears with E_{sat} as $E_{\text{sat}}L$. Thus the behavior of the equations does not only depend on L, but the combination of L and E_{sat}, or, its relative magnitude to $V_{\text{G}} - V_{\text{T}}$.

When $E_{\text{sat}}L$ is small enough, V_{Dsat} is basically dominated by the $E_{\text{sat}}L$ term, and I_{Dsat} term is now

$$\begin{aligned} I_{\text{Dsat}} &= WC_{ox}v_{\text{sat}}(V_{\text{G}} - V_{\text{T}} - E_{\text{sat}}L) \\ &\approx WC_{ox}v_{\text{sat}}(V_{\text{G}} - V_{\text{T}}) \end{aligned}$$

if we further ignore the $E_{\text{sat}}L$ term.

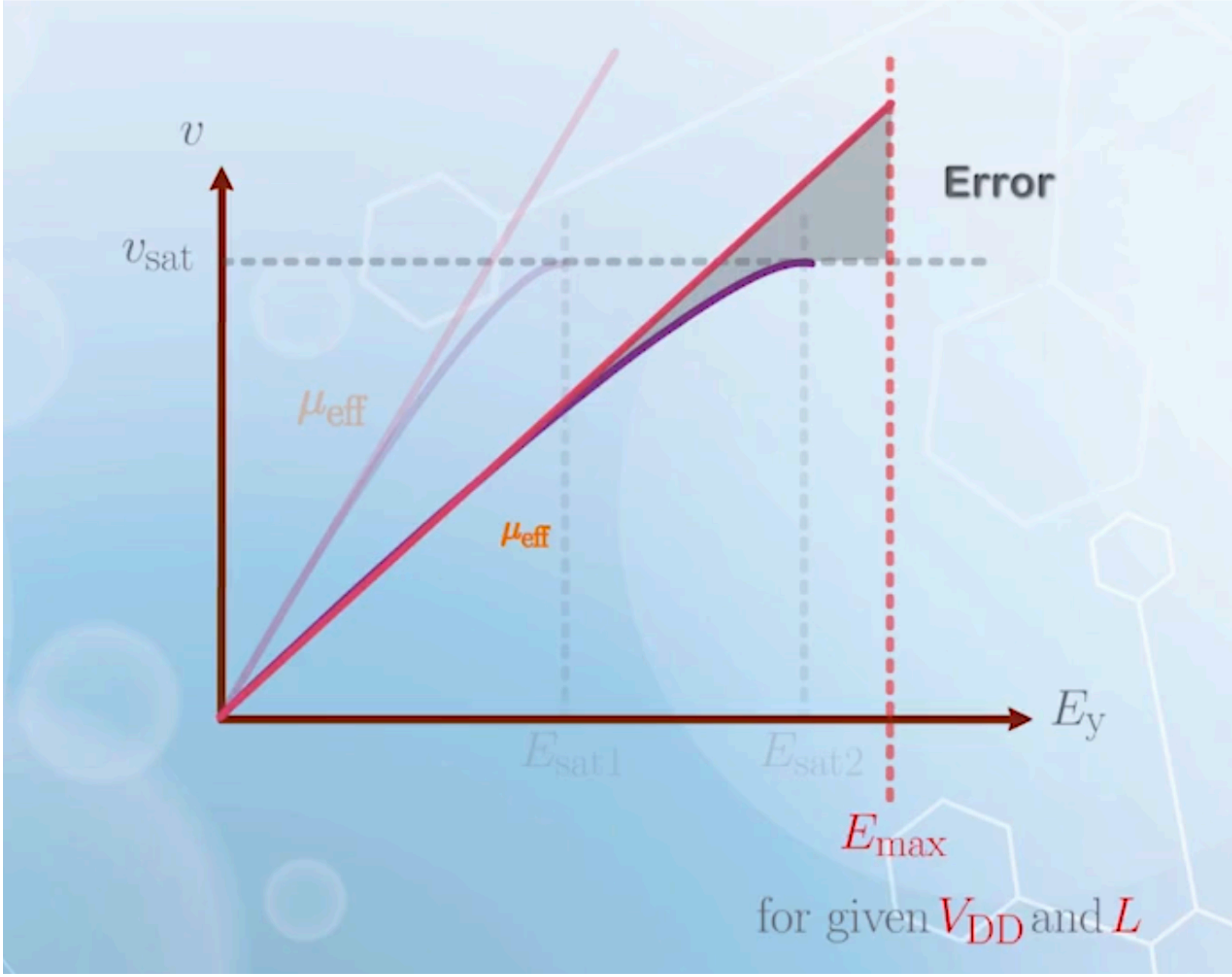
These are the characteristics of the MOSFETs are becoming, with the reduction of dimensions in the state of art technologies.

But there will always a region where $V_{\text{G}} - V_{\text{T}} \ll E_{\text{sat}}L$, and pinch off model is applicable. This will happen if V_{G} is just above V_{T}, making $V_{\text{G}} - V_{\text{T}}$ very small, by limiting the power supply voltage.

Also, because E_{sat} always appear with L, making E_{sat} large will allow the use of pinchoff model even for small L.

$$E_{\text{sat}} = \frac{2v_{\text{sat}}}{\mu_{\text{eff}}}$$

E_{sat} can be made large by making μ_{eff} small. For a same voltage, this can be achieved by using a thin gate oxide to increase vertical electric field. Thus the pinchoff model can also be used with a relatively short channel if the gate oxide is thin enough.



MOSFET Current With Gate Voltage

The pinchoff model predicts that in the saturation region, the MOSFET current varies with the square of $(V_{\text{G}} - V_{\text{T}})$.

$$I_{\mathrm{Dsat}} = \frac{\mu_{\mathrm{eff}} C_{ox} \frac{W}{L} (V_{\mathrm{G}} - V_{\mathrm{T}})^2}{2} \propto (V_{\mathrm{G}} - V_{\mathrm{T}})^2$$

Increasing V_{G} with an equal step will increase the spacing between the I_{Dsat} curves.

But with carrier velocity saturation model

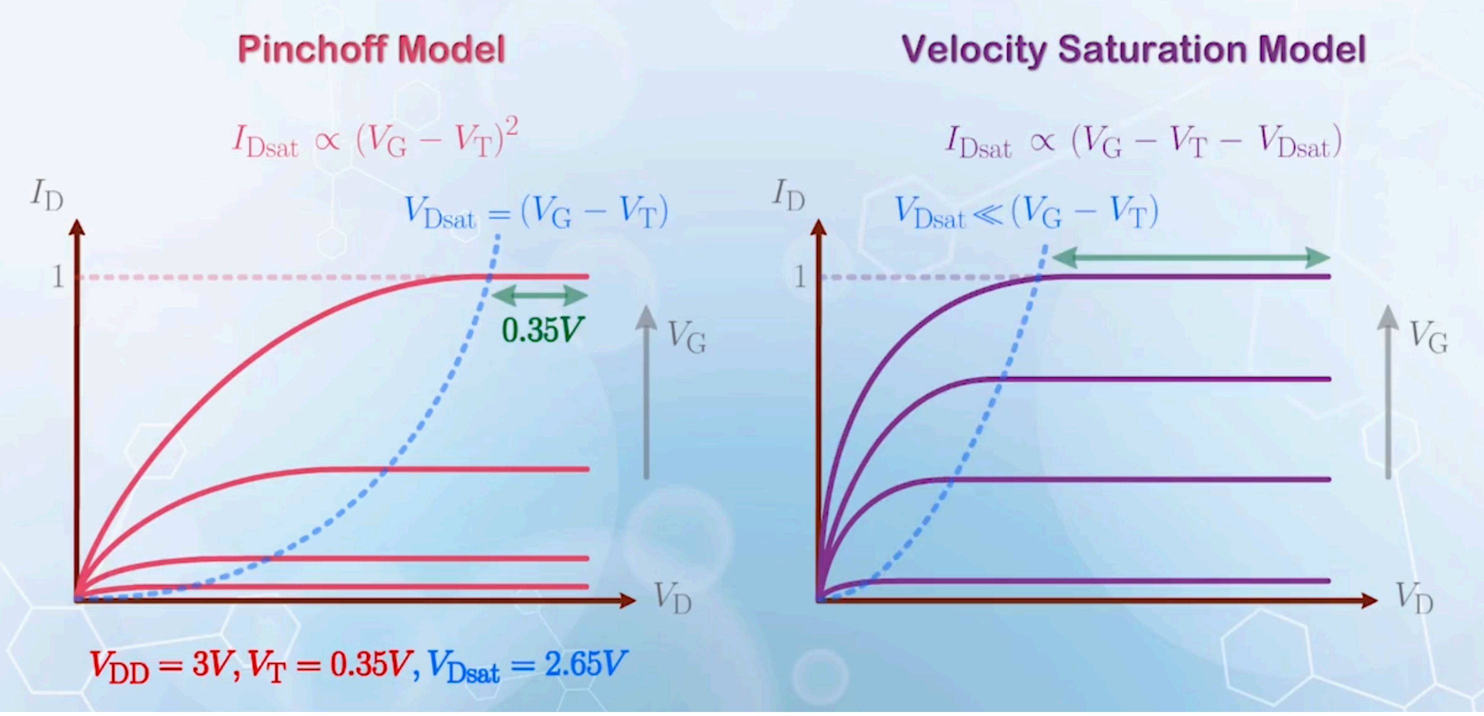
$$\begin{aligned} I_{\mathrm{Dsat}} &= W C_{ox} v_{\mathrm{sat}} (V_{\mathrm{G}} - V_{\mathrm{T}} - V_{\mathrm{Dsat}}) \\ &= W C_{ox} v_{\mathrm{sat}} \frac{(V_{\mathrm{G}} - V_{\mathrm{T}})^2}{(V_{\mathrm{G}} - V_{\mathrm{T}}) + E_{\mathrm{sat}} L} \\ &\propto (V_{\mathrm{G}} - V_{\mathrm{T}} - V_{\mathrm{Dsat}}) \end{aligned}$$

This will give a more uniform spacing between the I_{Dsat} curves when V_{G} is increased with an equal step.

As pinchoff model is an approximation of the carrier velocity saturation model with long channel length, it means I_{Dsat} for long channel MOSFETs has a quadratic dependence on V_{G}, but this dependence becomes more linear as the channel length is reduced, which is the case we observe in most state of the art devices.

Also, consider the span of saturation region, which is determined by V_{Dsat}.

For example, with a $\mathbf{3\,V}$ power supply, and a $\mathbf{0.35\,V}$ threshold voltage, V_{Dsat} predicted by the pinchoff model is $\mathbf{2.65\,V}$, giving a saturation region of $\mathbf{0.35\,V}$. Practically, we see a much larger saturation region, as the carrier velocity saturation model predicts a much smaller V_{Dsat}.



MOSFET Current Dependence on Channel Length

With the pinchoff model, $I_{\text{Dsat}} \propto \frac{1}{L}$. When $L \to 0$, $I_{\text{Dsat}} \to \infty$. But with the carrier velocity saturation model, I_{Dsat} has no explicit dependence on L, which only implicitly affects I_{Dsat} through V_{Dsat}.

When $L = 0$,

$$V_{\text{Dsat}} = \frac{(V_{\text{G}} - V_{\text{T}})E_{\text{sat}}L}{(V_{\text{G}} - V_{\text{T}}) + E_{\text{sat}}L} = 0$$

$$I_{\text{Dsat}} = WC_{ox}v_{\text{sat}}(V_{\text{G}} - V_{\text{T}} - V_{\text{Dsat}}) = WC_{ox}v_{\text{sat}}(V_{\text{G}} - V_{\text{T}})$$

Which one is more correct?

The ∞ current predicted by the pinchoff model comes from the calculation of E given by $\frac{V}{L}$. Also, the pinchoff model assumes that the carrier velocity is proportional to E. So when $L \to 0$, $E \to \infty$, and thus $v \to \infty$, leading to an infinite current. This is not true, as discussed with the carrier velocity saturation model, which is more physically consistent.

The increase in the saturation current by reducing channel length is smaller according to the carrier velocity saturation model, and the maximum achievable current is limited by $WC_{ox}v_{\text{sat}}(V_{\text{G}} - V_{\text{T}})$ when $L \to 0$.

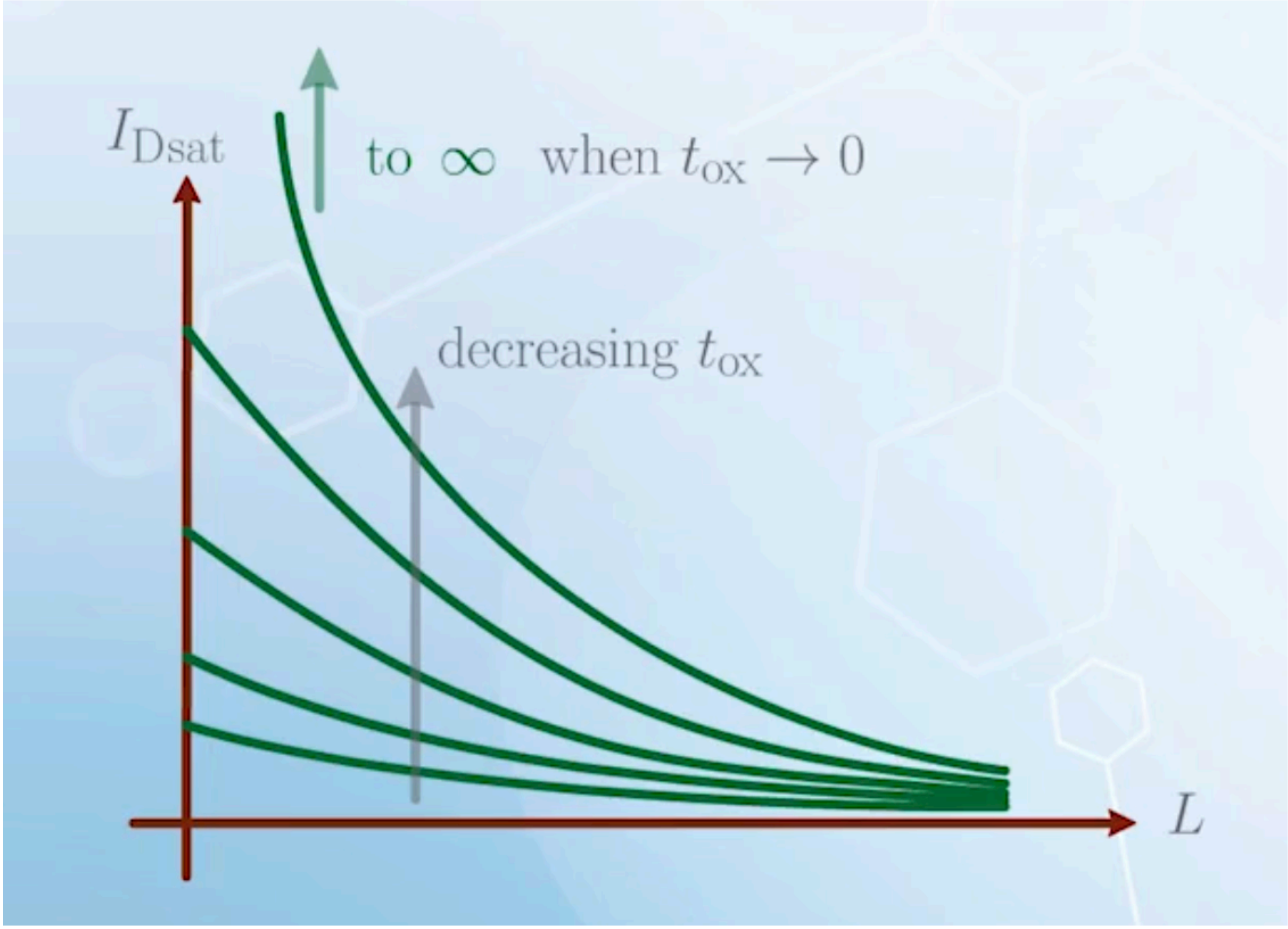
Also, base on the pinchoff model, scaling W and L together will not affect I_{Dsat}, as the $\frac{W}{L}$ term remains constant. This assumption may be valid for long channel MOSFETs, but not in general. When W is halved, I_{Dsat} will be halved, but when L is halved, I_{Dsat} cannot be doubled, thus the overall current will be reduced.

MOSFET Current Dependence on Gate Oxide Thickness

Reducing the gate oxide thickness t_{ox} increases C_{ox}, thus increases I_{Dsat} in both models.

But in the carrier velocity saturation model, reducing t_{ox} also decreases μ_{eff} due to mobility degradation, thus increases E_{sat}, which increases V_{Dsat}, and further decreases I_{Dsat}. This counteracts the increase in I_{Dsat}, making the increase smaller than that predicted by the pinchoff model.

This effect will be reduced when L is decreased, and when $L \to 0$, $I_{\text{Dsat}} \propto \frac{1}{t_{ox}}$



When L is large, decreasing t_{ox} does not significantly increase I_{Dsat} due to the counteracting effect of increasing E_{sat}. But when L is small, decreasing t_{ox} significantly increases I_{Dsat}.

When t_{ox} is very large, reducing L has very small head room to increase the current. But reducing t_{ox} creates more head room for I_{Dsat} to increase with reducing L.

Theoretically, the current can be infinity if $t_{ox} \to 0$, or $C_{ox} \to \infty$.

t_{ox} and L should be reduced together to achieve the maximum I_{Dsat} with the size reduction.

The Physical Effect of Channel Length

The current is given by

$$I = Q(y)v(y)$$

In the pinchoff model, reducing L mainly increases $\mu\frac{V_{\mathrm{D}}}{L}$, which can be considered as increasing $v(y)$.

In the carrier velocity saturation model, the carrier velocity $v(y)$ is limited by v_{sat}, which is independent of L. L only appears implicitly in V_{Dsat}, which belongs to the $Q(y)$ part.

The saturation current is mainly determined by the charge density near the drain. When L is reduced, required V_{Dsat} to increase the channel electric field to reach E_{sat} is also reduced. Therefore, more charges can move with the velocity of v_{sat}, increasing the current. In extreme case with $L = 0$, a very small $V_{\mathrm{Dsat}} \to 0$ will cause the electric field to reach E_{sat}, and all the channel charges coming out of the source can move with the velocity of v_{sat}. This gives the maximum current achievable with the reduction of channel length

$$I_{\mathrm{Dsat}} = WC_{ox}v_{\mathrm{sat}}(V_{\mathrm{G}} - V_{\mathrm{T}})$$

I-V Calculation Example

Given the following parameters:

$$\begin{aligned} t_{ox} &= 7\,\mathrm{nm} \\ W &= 1\,\mu\mathrm{m} \\ L &= 0.35\,\mu\mathrm{m} \\ V_{\mathrm{G}} &= 3\,\mathrm{V} \\ V_{\mathrm{D}} &= 1.5\,\mathrm{V} \\ V_{\mathrm{T}} &= 0.4\,\mathrm{V} \\ v_{\mathrm{sat}} &= 8 \times 10^6\,\mathrm{cm/s} \end{aligned}$$

Calculate I_{D}.

The universal mobility model gives

$$\mu_{\mathrm{eff}} = \frac{\mu_0}{1 + (E_{\mathrm{eff}}/E_0)^\nu}$$

	Electrons	Holes
$\mu_0 \quad (\mathrm{cm}^2/(\mathrm{V}\cdot\mathrm{s}))$	670	160

	Electrons	Holes
E_0 (MV/cm)	0.67	0.7
ν	1.6	1.0

and E_{eff} is given by

$$E_{\text{eff}} = \frac{(V_{\text{G}} - V_{\text{T}})}{6t_{ox}} + \frac{(V_{\text{T}} + V_{\text{a}})}{3t_{ox}}$$

We now have

$$\mu_{\text{eff}} = 256\,\text{cm}^2/(\text{V} \cdot \text{s})$$

Then, E_{sat} is given by

$$\begin{aligned} E_{\text{sat}} &= \frac{2v_{\text{sat}}}{\mu_{\text{eff}}} \\ &= 0.625\,\text{MV/cm} \end{aligned}$$

Now V_{Dsat} is given by

$$\begin{aligned} V_{\text{Dsat}} &= \frac{(V_{\text{G}} - V_{\text{T}})E_{\text{sat}}L}{(V_{\text{G}} - V_{\text{T}}) + E_{\text{sat}}L} \\ &= 1.19\,\text{V} \\ &< 1.5\,\text{V} = V_{\text{D}} \end{aligned}$$

The device is in saturation region.

Finally, I_{Dsat} is given by

$$\begin{aligned} I_{\text{Dsat}} &= WC_{ox}v_{\text{sat}}(V_{\text{G}} - V_{\text{T}} - V_{\text{Dsat}}) \\ &= 557\,\mu\text{A} \end{aligned}$$

With the pinchoff model, the device would still be in linear region, and the calculated current would be $1\,\text{mA}$.

9. Transistor Scaling

About the scaling of MOSFETs, including Moore’s law, scaling rules, ITRS, economic considerations, scaling limits, and effects of scaling on parasitics.

The Moore’s Law

> There is plenty of room at the bottom. – Richard Feynman, 1959

By making transistors smaller, they can be placed closer, electrons carrying information needs to travel shorter distances, which makes the circuit faster. The power required to move these electrons is also reduced, due to smaller loading, leading to the reduction of power consumption. Furthermore, the cost can be reduced, as the entire system can be implemented in a smaller area of an IC. The process of transistor size reductions is often called **technology scaling**. It is a rare situation that brings about benefits without any drawbacks.

Gordon Moore, the co-founder of Intel, predicted in 1965 that the number of transistors in an integrated circuit would double for every year. This phenomenon is often referred to as **Moore’s Law**. In 1975, he revised his prediction to a doubling every two years. The exact number does not matter, it just predicts that the number of components increases exponentially with time.

So far, the increase in the number of transistors over the past years follows the Moore’s law quite closely, but it may not due to the correctness of Gordon Moore’s prediction. As people believe that if they cannot keep up 😨, someone else will 🤣👉. Therefore, engineers are working hard to make it happen. The correctness of Moore’s law can be considered as a self-fulfilling prophecy.

The Scaling Rule

What parameters can be scaled in the scaling process? Reducing the dimensions in an arbitrary way may not maintain the transistor characteristics, and the performance may not be optimized. A set of scaling rules is needed to guide the scaling process.

In 1974, Robert Dennard from IBM proposed a set of scaling rules.

As carrier motion is mainly determined by the electric field, Robert Dennard proposed that the scaling process should keep the electric field unchanged. To achieve this, geometric dimensions, voltages, and sometimes doping concentrations need to be scaled together.

Suppose the geometric dimensions are scaled down by a factor of $k < 1$. To achieve the same electric field, the main parameters that need to be scaled are:

- Gate length $L_{\mathrm{G}} \to kL_{\mathrm{G}}$
- Gate width $W_{\mathrm{G}} \to kW_{\mathrm{G}}$
- Gate oxide thickness $t_{\mathrm{ox}} \to kt_{\mathrm{ox}}$
- Power supply voltage $V_{\mathrm{DD}} \to kV_{\mathrm{DD}}$

Then the performance after the scaling is (velocity saturation model is used, as scaling concerns small devices):

Parameter	Scaling Result
L_{G}, W_{G}, t_{ox}, V_{DD}	k
I_{Dsat}	$W_{\mathrm{G}}C_{ox}(V_{\mathrm{G}} - V_{\mathrm{T}})v_{\mathrm{sat}} \to kk^{-1}k = k$
$I_{\mathrm{Dsat}}/W_{\mathrm{G}}$	1
C_{G}	$\varepsilon_{ox}W_{\mathrm{G}}L_{\mathrm{G}}/t_{ox} \to kk/k = k$
Switch delay τ	$C_{\mathrm{G}}V_{\mathrm{DD}}/I_{\mathrm{D}} \to kk/k = k$
Clock f	$1/\tau \to 1/k$
Chip area	not directly resulting from scaling measured by another factor $\alpha > 1$ for most cases in the past
# Transistor N (number per chip)	α/k^2

Parameter	Scaling Result
Power / chip	$NC_{\mathrm{G}}V_{\mathrm{DD}}^2 f \to \alpha/k^2 k k^2 \cdot 1/k = \alpha k$

Some notes:

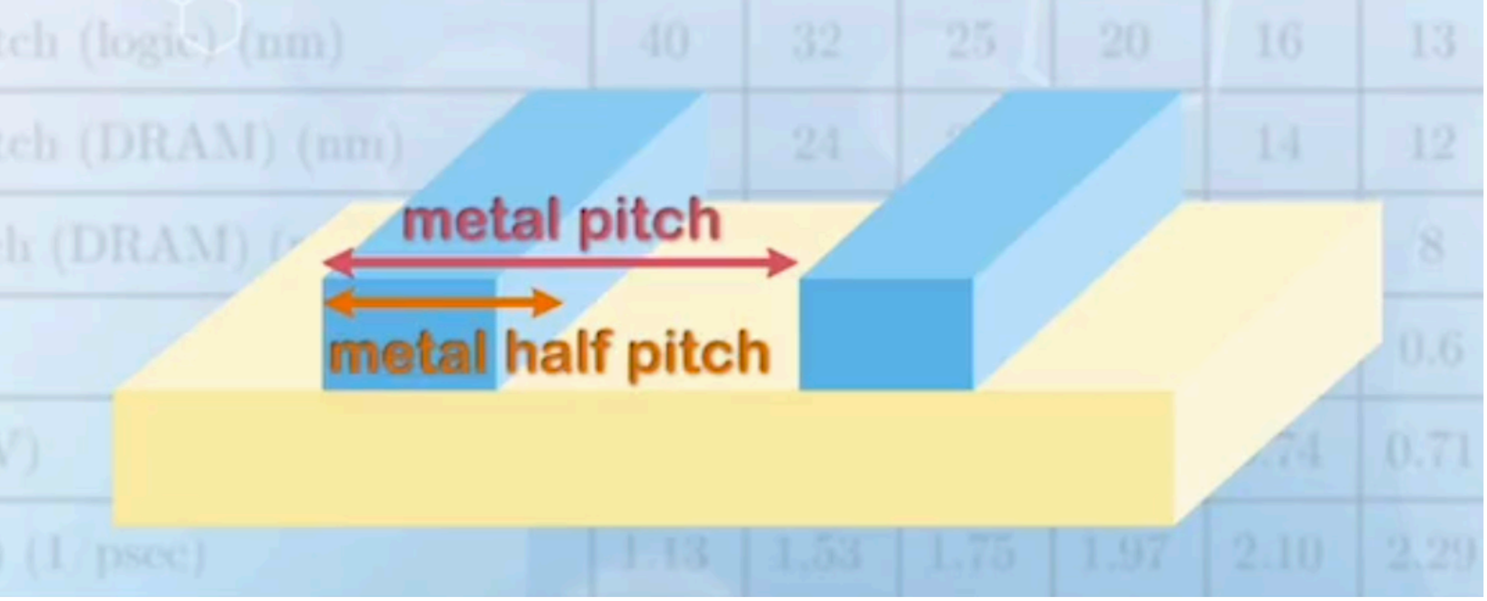
- I_{Dsat} is **decreased** when scaling down
    - The performance improvements mainly comes from the reduction in loading, rather than higher current drive
- The power dissipation per chip is composed of leakage power and dynamic power
    - $NC_{\mathrm{G}}V_{\mathrm{DD}}^2 f$ gives the dynamic power to charge and discharge the load capacitance
    - In the early days, the leakage power is not very serious, and only dynamic power is considered

The International Technology Roadmap for Semiconductors (ITRS)

Moore law has become the consensus of the semiconductor industry. Some people even predict that beating Moore law may result in new products that the market cannot absorb, and hence causing engineers to lose their jobs, and falling behind the Moore law will lead to nothing to buy and shrink the distribution chain. No matter what, sustaining Moore law is an important target for many companies.

The International Technology Roadmap for Semiconductors (ITRS) is introduced to guide the semiconductor technology development. But it did not provide the solution.

The smallest dimension of the MOSFET is usually the gate length. In some applications with regular patterns, such as memory arrays, the spacing between the MOSFET may be better represented by the metal line spacing, and the technology node may also refer to the metal half pitch (half the distance between two metal lines with minimum line width, separated by the minimum distance).



However, the actual implementation of the technology varies among different companies, and the technology nodes have no direct relationship with the gate length or metal pitch. In fact, they may not correspond to any physical dimension of the MOSFET. The technology node only represents the technology generation over time, like an agreed version number among different companies.

The dimension of the gate length was smaller than the technology node before the 22 nanometer technology node. Starting around the 16 nanometer technology node, the gate length has become larger than the number given by the technology node. The dimension of each technology node is reduced to about 70% of the previous technology.

The EOT stands for equivalent oxide thickness. As multiple materials are used to form the gate dielectric, it specifies the equivalent oxide thickness when the gate dielectric is composed of pure Silicon dioxide to achieve the same gate capacitance. **For the same dielectric material thickness, a material with a higher dielectric constant will give a smaller EOT**. The EOT reduction is slowed down due to the difficulties in finding a thin material that can still behave as an insulator

The power supply voltage V_{DD} reduction also more or less stops due to the noise margins and other issues that are difficult to overcome at low supply voltages.

With more transistors packed on the chip of the same area, more connections are needed. In order to allow more metal lines to be laid on the chip and not increasing the area, the number of interconnect layers also increases with the technology scaling.

The ITRS also specifies some performance related parameters, such as the current density.

By predicting the performance of the MOSFETs, IC designers and application engineers will be able to start planning for the system before the technology of a specific node is available.

With the success of the ITRS to guide the technology over the past 20 years, the ITRS issued a final roadmap in 2016 because following Moore law has become very difficult. Different kinds of roadmaps, such as the International Roadmap for Devices and Systems (IRDS) are proposed to succeed the ITRS.

The Economy of Scaling

With the scaling process, the computational cost to customers is reduced, but the cost to follow the Moore's law is increasing, which includes manufacturing, testing, researching and development costs. The exponential increase in capital costs has offset the benefits of the scaling process, such that only very few companies desires to continue the scaling process.

To calculate the cost:

- The number of dice per wafer is given by

  $$N_{\text{die}} = \frac{\pi(\frac{D}{2})^2}{A_{\text{die}}} - \frac{\pi D}{\sqrt{2A_{\text{die}}}}$$

  where D is the diameter of the wafer, and A_{die} is the area of each die.
- The die cost is given by

  $$C_{\text{die}} = \frac{C_{\text{wafer}}}{N_{\text{die}} \times Y_{\text{die}}}$$

  where C_{wafer} is the cost per wafer, and Y_{die} is the yield.
  - The wafer cost increase due to the use of more advanced technology is relatively small, and can be well compensated with the increase in the number of of available dice
- The final chip cost including test and packaging is given by

  $$C_{\text{var}} = \frac{C_{\text{die}} + C_{\text{test}} + C_{\text{pkg}}}{Y_{\text{test}}}$$

  as some chips may be lost due to testing and packaging.
  - This is the **variable cost**, as the total production cost changes with the number of chips produced

- There is also a **fixed cost** for producing the chips, which is a one-time investment, independent of the number of chips produced.
  - Mostly the cost of the mask to transfer the patterns from design to wafer
  - To include this to the individual chip cost, it has to be divided by the number of chips produced N

The number of chips per wafer is a very important factor for cost reduction. The wafer size is getting bigger over the years to accommodate a larger number of chips per wafer. The wafer size has increased from 1 inch ($25\,\mathrm{mm}$) to 12 inches ($300\,\mathrm{mm}$). The $450\,\mathrm{mm}$ wafer has been proposed, but there is still quite resistance due to the potentially insufficient return.

The Scaling Limit

The scaling process is still being pursued, but it cannot continue forever. Ultimately, it will be limited to atom size. But before reaching that limit, more macroscopic problems may slow down the process.

Consider the $5\,\mathrm{nm}$ technology node. This may not represent the actual dimension, but we can use it just as a reference.

As the lattice constant of silicon is around $0.5\,\mathrm{nm}$, there are only 10 silicon atoms in the lateral dimension of a $5\,\mathrm{nm}$ MOSFET. Furthermore, the equivalent thickness of the gate dielectric material is around $1\,\mathrm{nm}$, containing only 2–3 atom layers of silicon dioxide atoms. This is why further scaling using silicon dioxide as the gate dielectric seems to be almost impossible.

Even we can make such devices, many physical limitations will be manifest at such small dimensions.

In a volume of $5\,\mathrm{nm} \times 5\,\mathrm{nm} \times 5\,\mathrm{nm}$ cube, there are only a few thousand of silicon atoms. Even doping the silicon to $1 \times 10^{19}\,\mathrm{cm}^{-3}$, the number of dopant atoms in the volume is still a single digit. The assumption of uniform dopant distribution may no longer hold. If one or two dopant atoms are missing, the percentage error is huge, causing significant device to device variations, and making the performance unpredictable.

Besides the dopant variation, any small error such as edge roughness or other geometrical imperfections will be significantly amplified, causing significant variations.

In terms of processing, **photolithography** is the most critical step in defining the small geometries. It basically uses a beam of light to transfer a computer-drawn pattern to a photosensitive material on the wafer through a mask.

In this process, the feature size defined cannot be smaller than the half of the wavelength of the light used. To define very small features, extreme ultraviolet light is used, which has many other problems itself, such as loss through the lens, low penetration depth, and large diffraction angle that causes a less precise geometry.

To pattern small dimensions without using light with ultra short wavelengths, engineers are using a technique called **immersion lithography**, which fills the space between the lens and the wafer with a liquid medium, so that the wavelength of the light is compressed. But such systems are very expensive.

Further development of lithography systems for dimensions beyond **5 nm** will cause significant capital investment, and the return is becoming uncertain.

Effects of Scaling on Parasitics

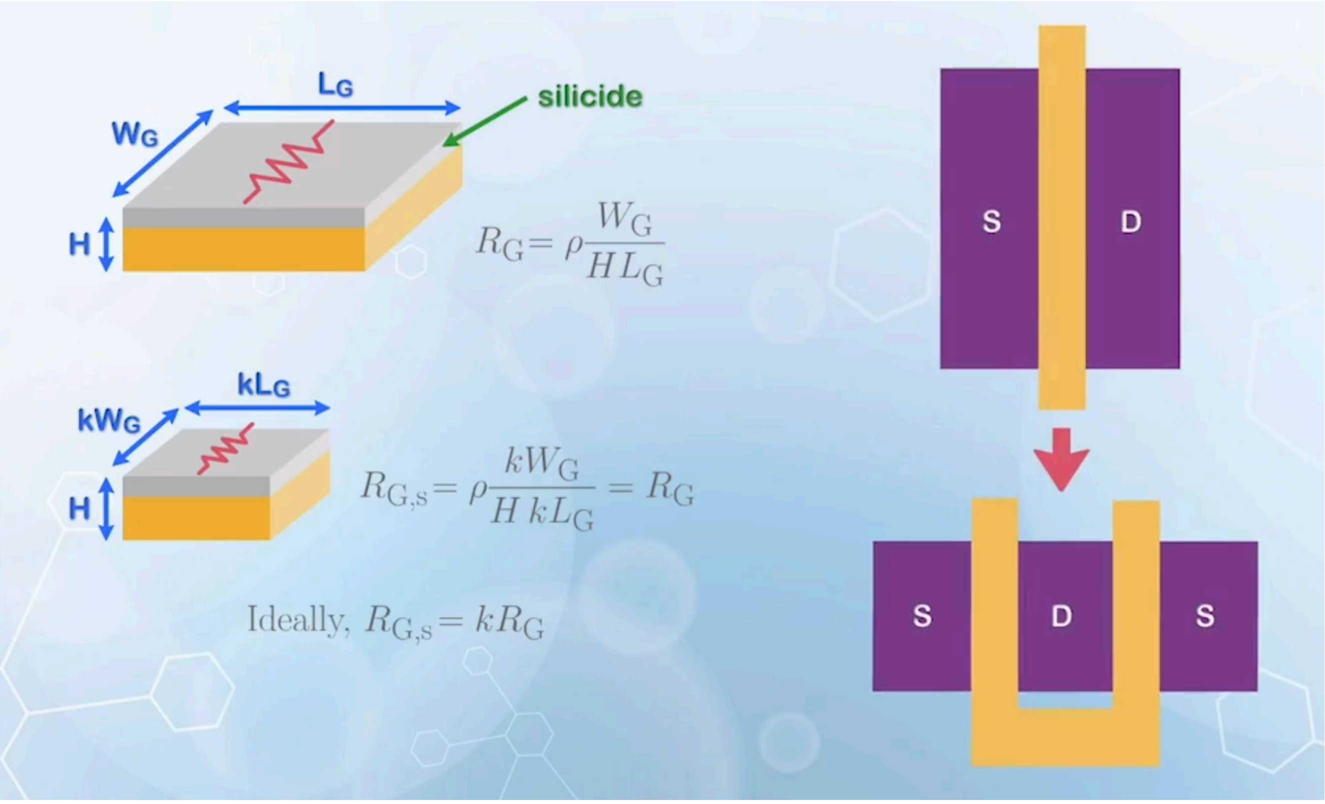
A MOSFET comes with some parasitic components.

- The gate resistance
    - Assume the height of the gate H is more or less constant during scaling
        - In real process, it is sort of achieved by the use of silicide, a semi-metallic material, to increase the conductivity
    - With W_G and L_G scaled down by k, the gate resistance R_G is given by

      $$R_\mathrm{G} = \rho \frac{W_\mathrm{G}}{HL_\mathrm{G}}$$
      $$R_\mathrm{G,s} = \rho \frac{kW_\mathrm{G}}{HkL_\mathrm{G}} = R_\mathrm{G}$$

      It remains more or less a constant during scaling
    - Ideally, it should scale by k to achieve better reduction in loading

- To reduce the resistance, a parallel layout is used for wide MOSFETs

  L_G
  silicide
  W_G
  H
  $R_G = \rho \frac{W_G}{H L_G}$
  kL_G
  kW_G
  H
  $R_{G,s} = \rho \frac{kW_G}{H\,kL_G} = R_G$
  Ideally, $R_{G,s} = kR_G$
  S
  D
  S
  D
  S

- The gate capacitance
  - It is given by

$$C_{\mathrm{G}} = \frac{\varepsilon_{ox} W_{\mathrm{G}} L_{\mathrm{G}}}{t_{ox}}$$

  - After scaling, it becomes kC_{G}
  - The RC delay at the input is scaled by a factor of k

$$R_{\mathrm{G,s}} C_{\mathrm{G,s}} = kR_{\mathrm{G}} C_{\mathrm{G}}$$

  - Using parallel layout can decrease the gate resistance, but may increase the gate capacitance, if not carefully handled

- The source and drain resistance
  - The sheet resistance component due to N+ doping is given by

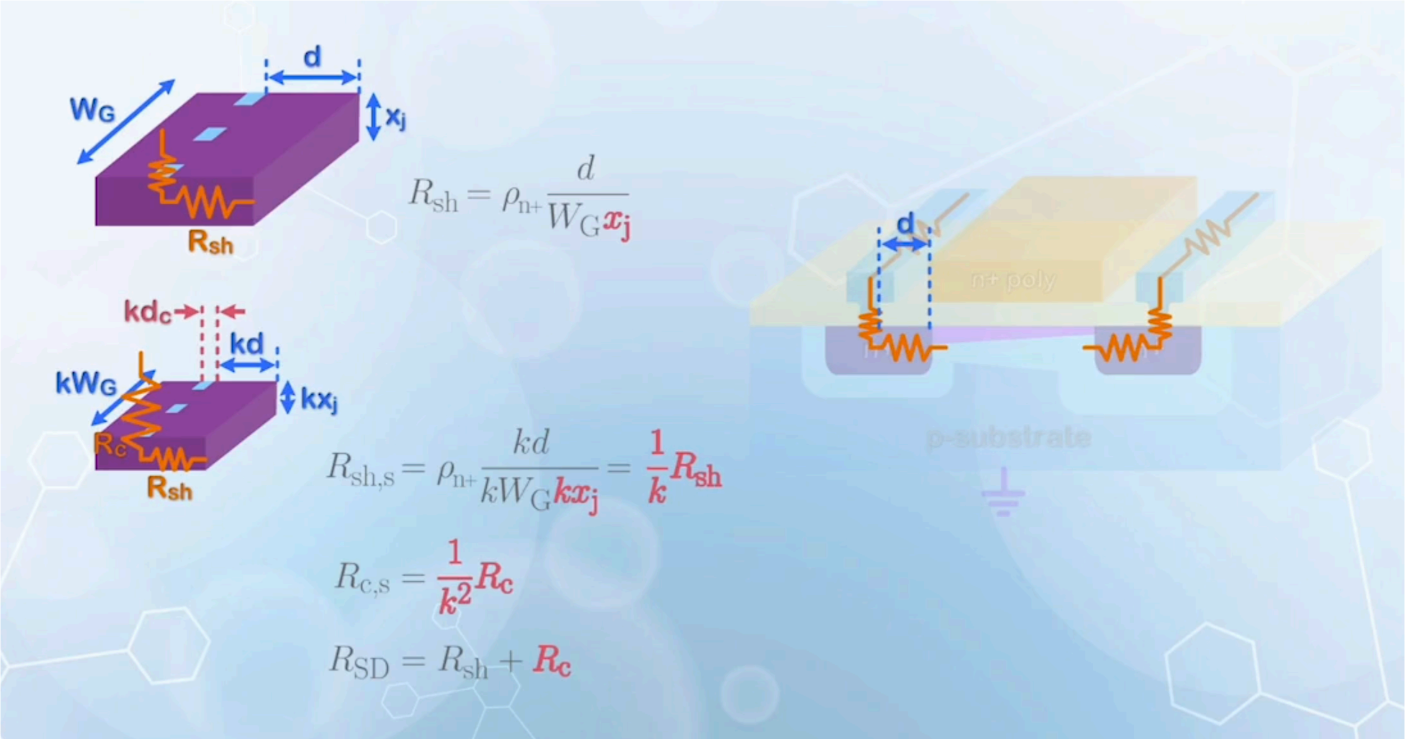
$$R_{\mathrm{sh}} = \rho_{\mathrm{n+}} \frac{d}{W_{\mathrm{G}} x_{\mathrm{j}}}$$

  - The metal resistance is considered to be very small and can be ignored

- Assuming the junction depth x_j remains unchanged after scaling, the sheet resistance remains unchanged
- But when a shallower junction is used, the series resistance increases
- The contact size is also reduced, leading to higher contact resistance R_c

  $$R_\mathrm{SD} = R_\mathrm{sh} + R_\mathrm{c}$$

  It has become the major component in the source and drain series resistances, which will keep increasing with the scaling process
- The increase in the source and drain resistance has the most adverse effect in the scaling process, as it significantly offsets the advantage of lower resistance in the MOSFET due to channel length reduction



$$R_\mathrm{sh} = \rho_\mathrm{n+}\frac{d}{W_\mathrm{G}x_\mathrm{j}}$$

$$R_\mathrm{sh,s} = \rho_\mathrm{n+}\frac{kd}{kW_\mathrm{G}kx_\mathrm{j}} = \frac{1}{k}R_\mathrm{sh}$$

$$R_\mathrm{c,s} = \frac{1}{k^2}R_\mathrm{c}$$

$$R_\mathrm{SD} = R_\mathrm{sh} + R_\mathrm{c}$$

- The source and drain capacitance
  - It is given by

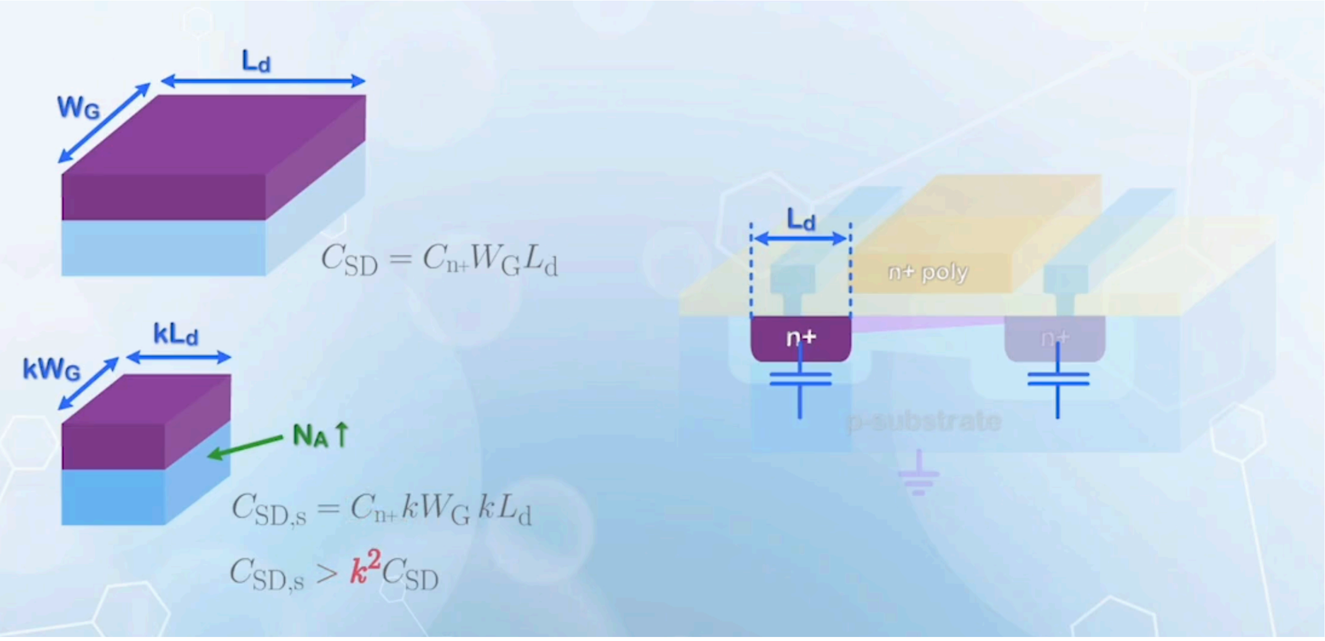
    $$C_\mathrm{SD} = C_\mathrm{n+}W_\mathrm{G}L_\mathrm{d}$$

  - When the lateral dimension of a MOSFET is reduced by k, the capacitance becomes

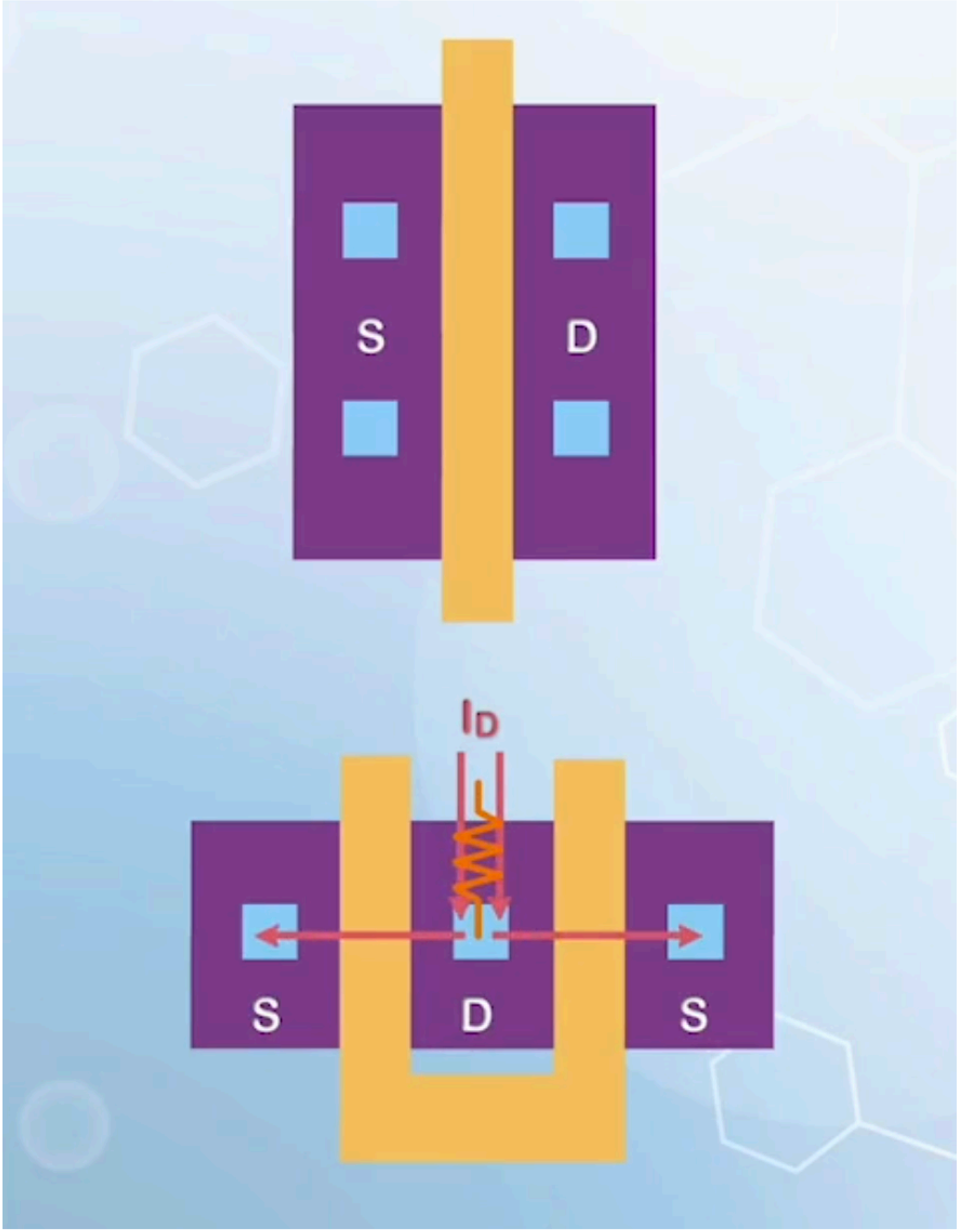
    $$C_\mathrm{SD,s} = C_\mathrm{n+}kW_\mathrm{G}kL_\mathrm{d} = k^2C_\mathrm{SD}$$

  - But the substrate doping usually increases with scaling

    - Recall that the reverse PN junction capacitance increases with higher doping on the lightly doped side
    - The increase in the substrate doping will cause an increase in the junction capacitance, offsetting some of the reduction of the source and drain parasitic capacitance due to scaling



- The parasitic source and drain capacitance can be reduced by using the parallel layout, as the drain region can be shielded by two transistors
    - Also increases the effective resistance through the source or drain, as the two MOSFETs share the same contact vias
    - The effect of drain contact will be amplified, as the amount of current flowing through it will be doubled



10. Short Channel MOSFET

About the challenges and solutions with short channel MOSFET design, including threshold voltage scaling, source / drain charge sharing, drain induced barrier lowering and punchthrough.

Short Channel MOSFET Design

Besides overcoming fabrication challenges, proper device design is needed to achieve desired performance of small MOSFETs.

The main purpose of a transistor is to serve as an switch, and a good switch should have a high current drive when turned on, and zero when turned off.

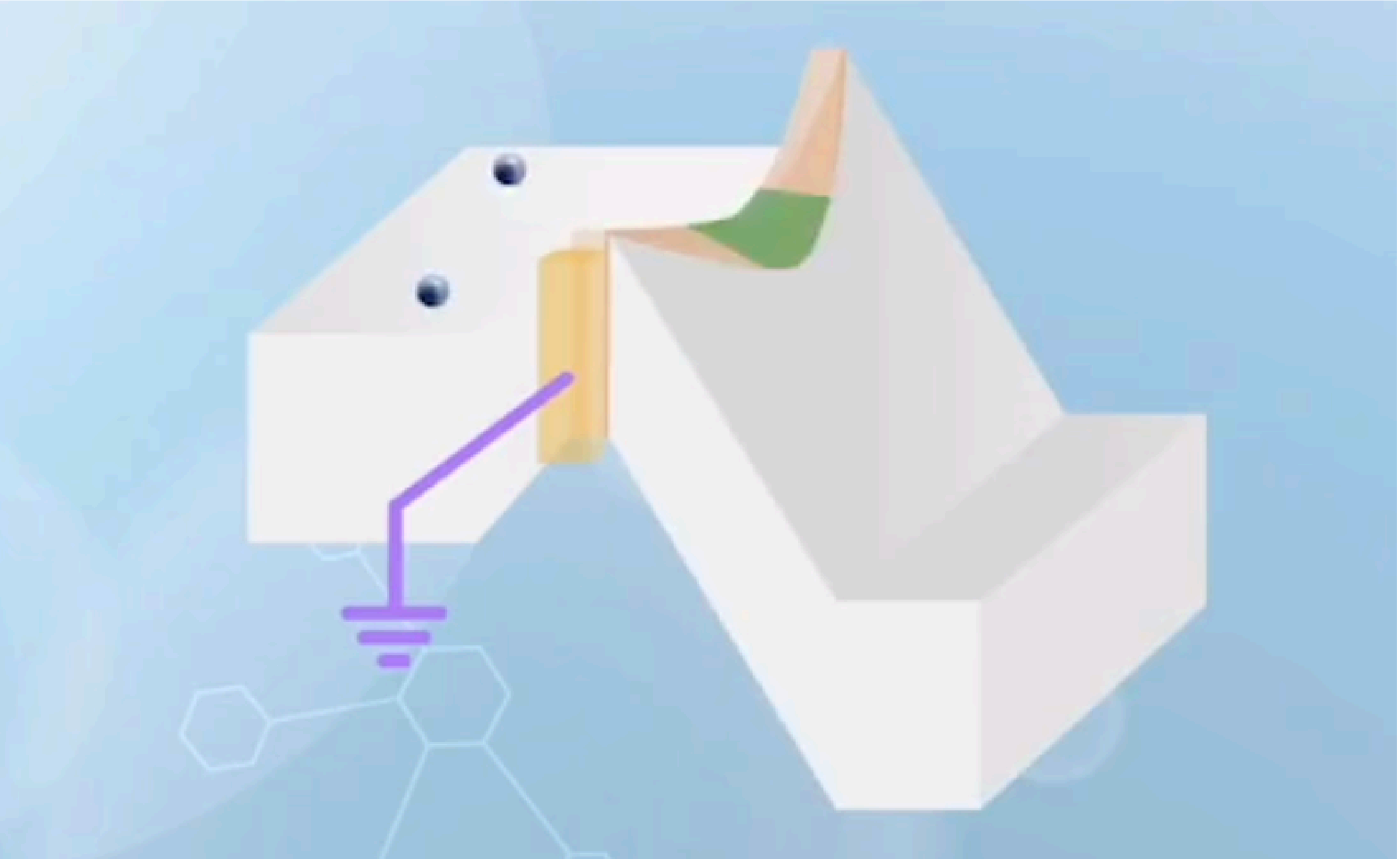
When channel length is short, giving high current when turned on is not a problem, the problem is how to turn it off completely.

Consider a long MOSFET with the electric potential of the channel controlled by the gate. When the gate is grounded, energy band in the channel is pulled up, with no electrons flowing in the channel. When drain voltage is applied, it pulls down the band through the drain terminal. There is a competition between the gate terminal holding the band high and drain voltage pulling the band down. In a long channel MOSFET, the gate controls most part of the channel, and the effect of the drain voltage on the channel is very small, so the gate can effectively turn off the MOSFET, holding the energy band high.

However, in a short channel MOSFET, the control of gate voltage over the channel is reduced, due to the its smaller area over the channel. Relatively, the effect of drain voltage on the channel energy band increases, causing the source to body energy barrier to be lowered, and introducing an additional leakage current that cannot be stopped by the gate.

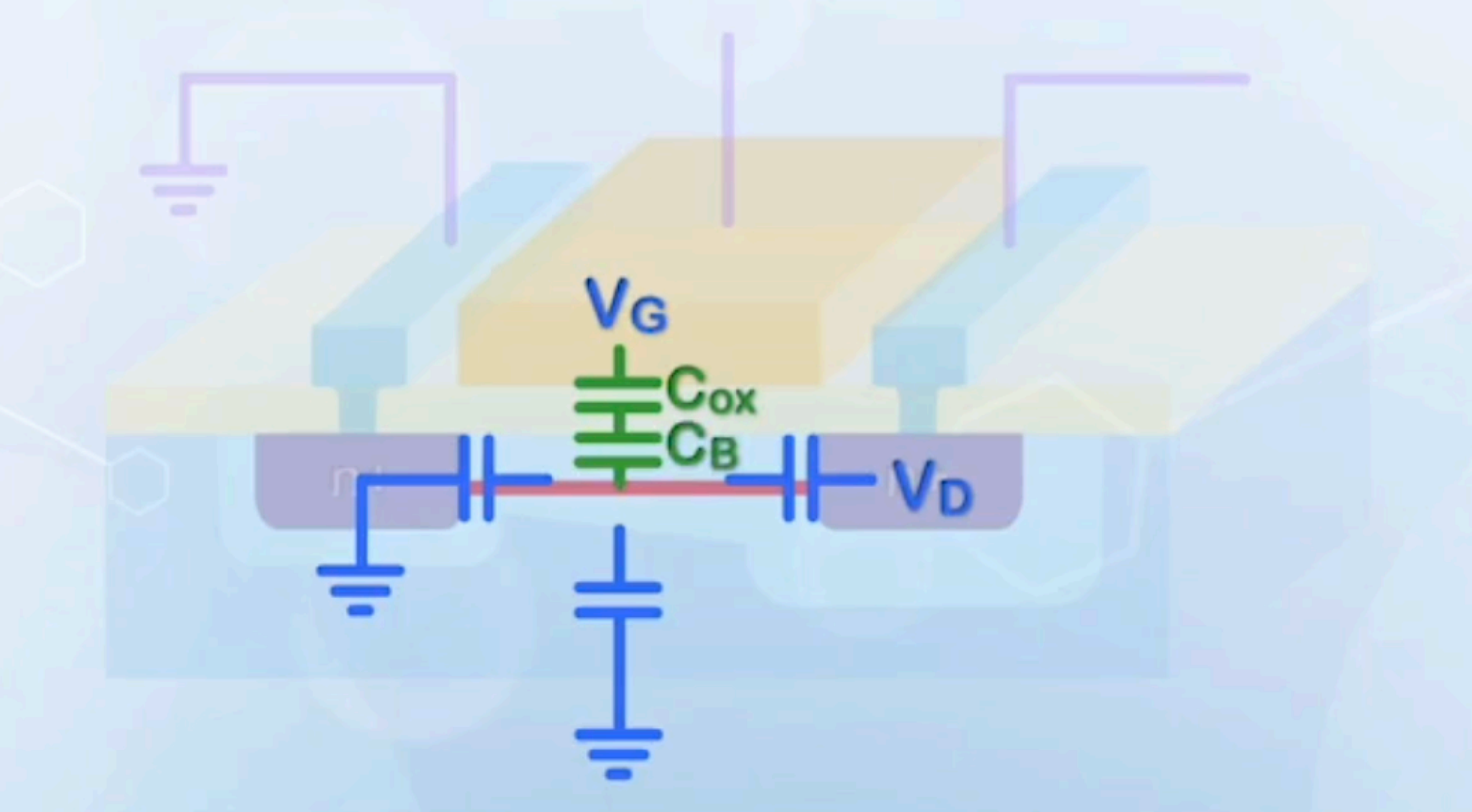
When channel is short enough, the gate loses control of the channel. The MOSFET cannot be turned off, and electron will flow once a drain voltage is applied.

Along the depth of the channel, the gate control is still relatively strong near the gate region. At the bottom od the depletion region under the gate, the body terminal also helps to hold the energy band high when grounded. Therefore, the point that is under the strongest influence of the drain voltage is **somewhere between the gate oxide to silicon interface, and the body terminal**. The leakage current will be the highest though this particular point.



In terms of electrostatic control, when the MOSFET is turned off, the region below the gate is depleted, and is considered insulator. The gate, source, drain and body terminals can only control the potential of the region through capacitive coupling. The potential at any point in the depletion region is the weighted sum of the four terminal voltages, and the relative influence of the terminals are determined by the associated capacitances.

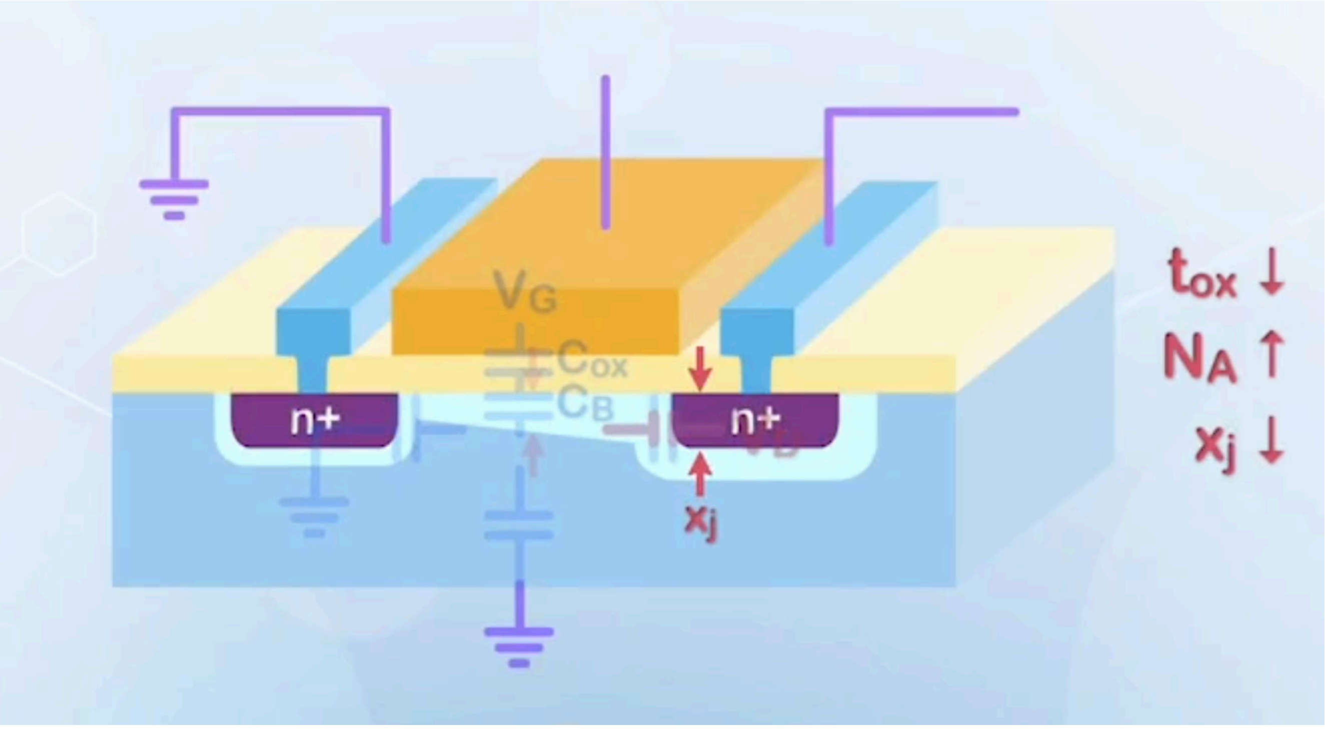
At the channel near the gate, the gate capacitance is C_{ox}. Moving deeper, the capacitance associated with the gate terminal becomes C_{ox} in series with the depletion region capacitance C_{B}, and the series capacitance is smaller than C_{ox}, weakening the gate control relative to the drain. In this case, the gate terminal may not be able to turn off the MOSFET at that location.



To increase the gate control over the channel

- Increase C_{ox}
    - By **reducing oxide thickness**
- Reduce the depletion region width
    - So that the weakest point is closer to the gate
    - Effectively, increases C_{B}
    - Using **heavy substrate doping**
- Reduce the drain capacitance
    - By **reducing drain junction depth $\boldsymbol{x_j}$**

- Effective side wall area of the drain capacitance is reduced

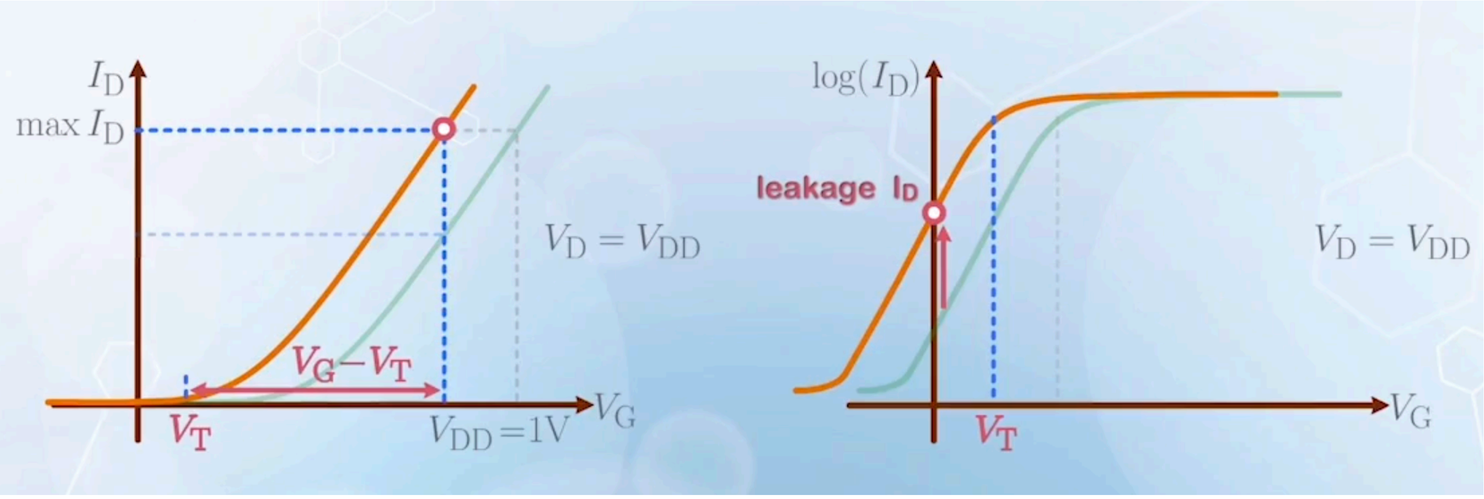


Threshold Voltage Scaling

The choice of V_{T} has a significant impact of the speed and power tradeoff in CMOS technology. For the same technology node, the CMOS process can be further divided into two types: low power (LP), decreasing leakage current, and high performance (HP), increasing on current. The difference is mainly in the choice of V_{T}.

The maximum current drive of a MOSFET is determined by $V_{\mathrm{G}} - V_{\mathrm{T}}$, or the **gate overdrive voltage**. The reduction of power supply voltage has significantly limited the range of usable $V_{\mathrm{G}} - V_{\mathrm{T}}$. Thus, there is a desire to reduce V_{T} together with V_{DD} scaling to provide more overdrive.

However, reducing V_{T} will cause the $\log(I_{\mathrm{D}}) - V_{\mathrm{G}}$ curve to shift left, significantly increase the leakage current at $V_{\mathrm{G}} = 0$.



An Example

Given: $V_{DD} = 1\,\text{V}$, initial threshold voltage $V_{T1} = 0.7\,\text{V}$, reduced threshold voltage $V_{T2} = 0.3\,\text{V}$, subthreshold swing $S = 80\,\text{mV/decade}$.

The saturation current:

$$\begin{aligned} \max(V_G - V_{T1}) = 0.3\,\text{V} \quad &\Rightarrow \max(I_{D1}) \propto 0.3\,\text{V} \\ \max(V_G - V_{T2}) = 0.7\,\text{V} \quad &\Rightarrow \max(I_{D2}) \propto 0.7\,\text{V} \\ \max(I_{D2}) = 2.33&\max(I_{D1}) \end{aligned}$$

The leakage current:

$$I_{\text{leakage}} \xrightarrow{V_{T2}=V_{T1}-0.4\,\text{V}} 10^5 I_{\text{leakage}}$$

To reduce V_T without increasing leakage current, the subthreshold swing S must be improved (**lowered**, as the swing is $\frac{1}{\text{slope}}$, and we would like to increase the slope). It is equivalent to reduce the ideality factor n.

$$n = 1 + \frac{C_D}{C_{ox}}$$

We can use extremely thin gate oxide to increase C_{ox}, but reducing C_D by reducing substrate doping is not preferred, as it will weaken the gate control over the region under the gate, leading to a even large leakage current under high V_D due to drain induced barrier lowering and punchthrough.

Ideally, $n = 1$, $S = 60\,\text{mV/decade}$, and with $V_T = 0.3\,\text{V}$, on-off current ratio can be as high as 10^5.

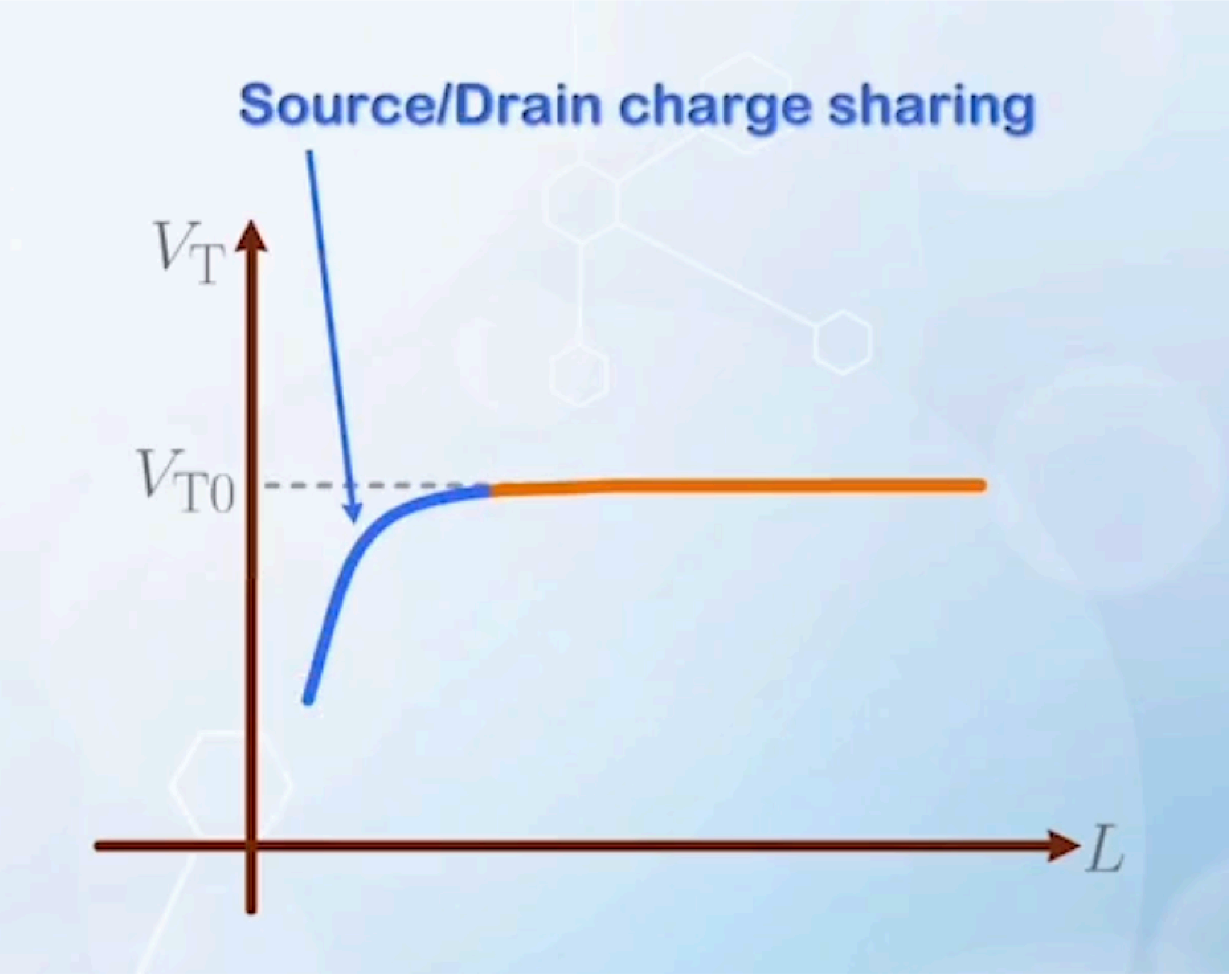
In the past, when calculating transistor power, we only calculate the dynamic power that charges up the loading capacitors. With the reduction of V_{T}, the off state leakage current increases exponentially even when MOSFETs are not switching, which can become dominant in modern ICs. Therefore, MOSFETs with multiple V_{T} values may be used in an IC to tradeoff between power dissipation and speed. A low V_{T} can be used in high speed parts that switch frequently, while a high V_{T} can be used for other parts that switch less frequently to reduce leakage current.

Source / Drain Charge Sharing

When the channel length is short enough, we say the MOSFET is experiencing **short channel effects**. The description of the MOSFET becoming a poor switch is qualitative, a more quantitative way is needed to compare the performance of MOSFETs with short channel length behaves differently from long channel MOSFETs.

The **short channel effect** is the dependence of V_{T} with respect to the dimension and voltage for very short channel MOSFETs.

In classic theory, V_{T} is a constant, independent of channel length. At very small dimensions, V_{T} becomes a function of channel length L and terminal voltages V. This behavior was first explained by the **source / drain charge sharing model**.



Without source and drain, the threshold voltage is given by

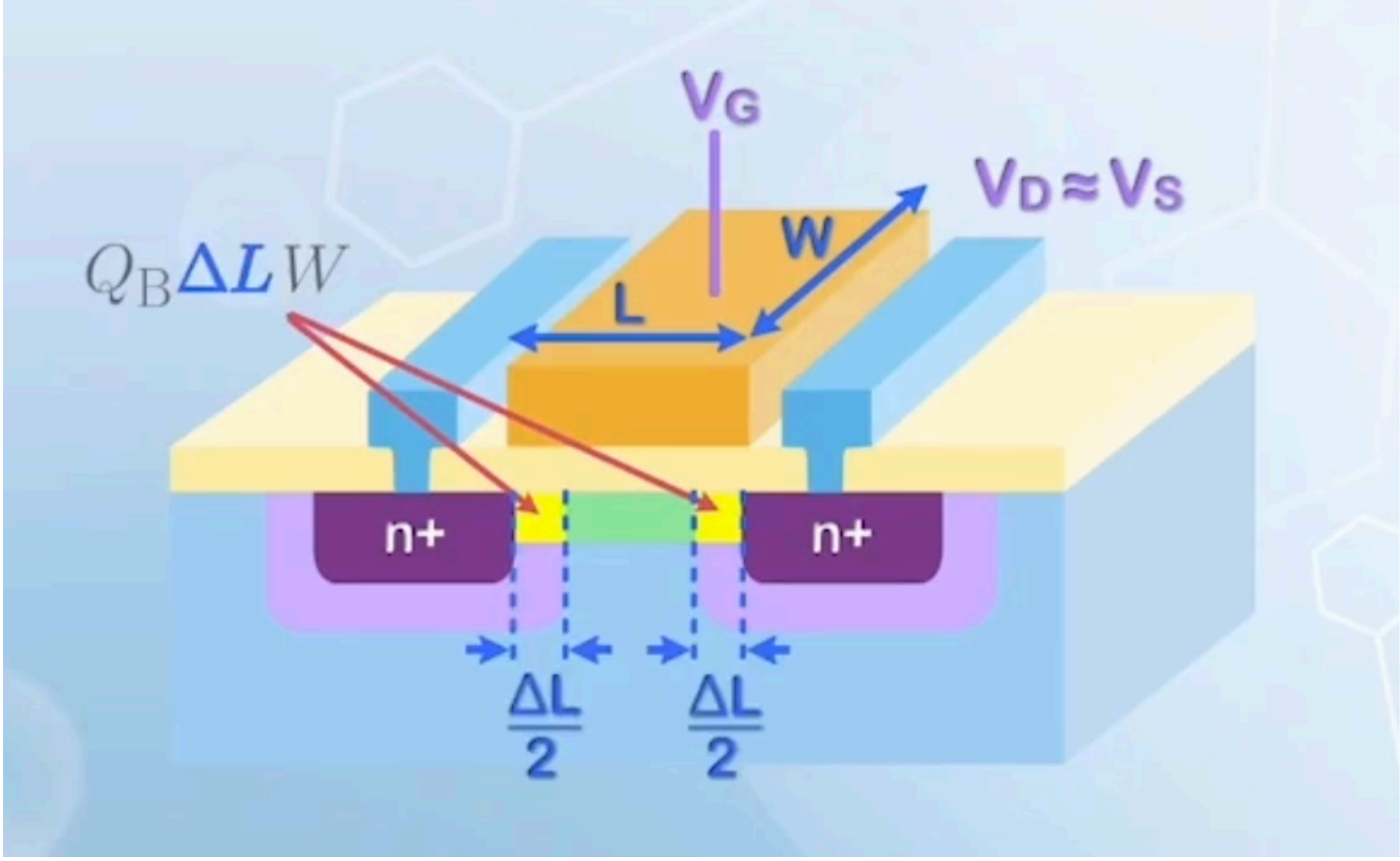
$$V_{\mathrm{T}} = V_{\mathrm{FB}} + 2\phi_{\mathrm{B}} + \frac{Q_{\mathrm{B}}}{C_{ox}}$$

The last term is normalized with respect to area. Should be $\frac{Q_{\mathrm{B}}WL}{C_{ox}WL}$ with physical dimensions included.

When source and drain are present, the source and drain depletion region have already depleted some part of the original $Q_{\mathrm{B}}WL$, even under flat band condition.

Assume $V_{\mathrm{D}} \approx V_{\mathrm{S}}$ to avoid the non uniform charge distribution along the channel. With part of the channel already depleted, a smaller gate voltage is needed to reach threshold condition, and V_{T}. will drop.

The amount of charge depleted by the source and drain is expressed as $Q_\mathrm{B}\Delta LW$, where ΔL is the effective length of charge provided by the source and drain.



The new threshold voltage is

$$\begin{aligned} V_\mathrm{T} &= V_\mathrm{FB} + 2\phi_\mathrm{B} + \frac{Q_\mathrm{B}WL - Q_\mathrm{B}W\Delta L}{C_{ox}WL} \\ &= V_\mathrm{FB} + 2\phi_\mathrm{B} + \frac{Q_\mathrm{B}}{C_{ox}}\left(1 - \frac{\Delta L}{L}\right) \end{aligned}$$

To put it simply, the source and drain helps the gate by providing some extra charge to the channel, such that a smaller gate voltage is needed to reach the same threshold condition.

In the expression, $Q_\mathrm{B}\Delta L$ is independent of L, and the contribution of source and drain is very small when L is large. However, when L is small, the contribution becomes significant, and V_T drops noticeably.

Most circuits rely on a stable V_T to determine the turn-on and turn-off conditions of MOSFETs, and V_T variations with the MOSFET dimensions are not desired. So when the V_T drop becomes significant, we consider the MOSFET no longer functioning at that channel length.

To reduce source / drain charge sharing:

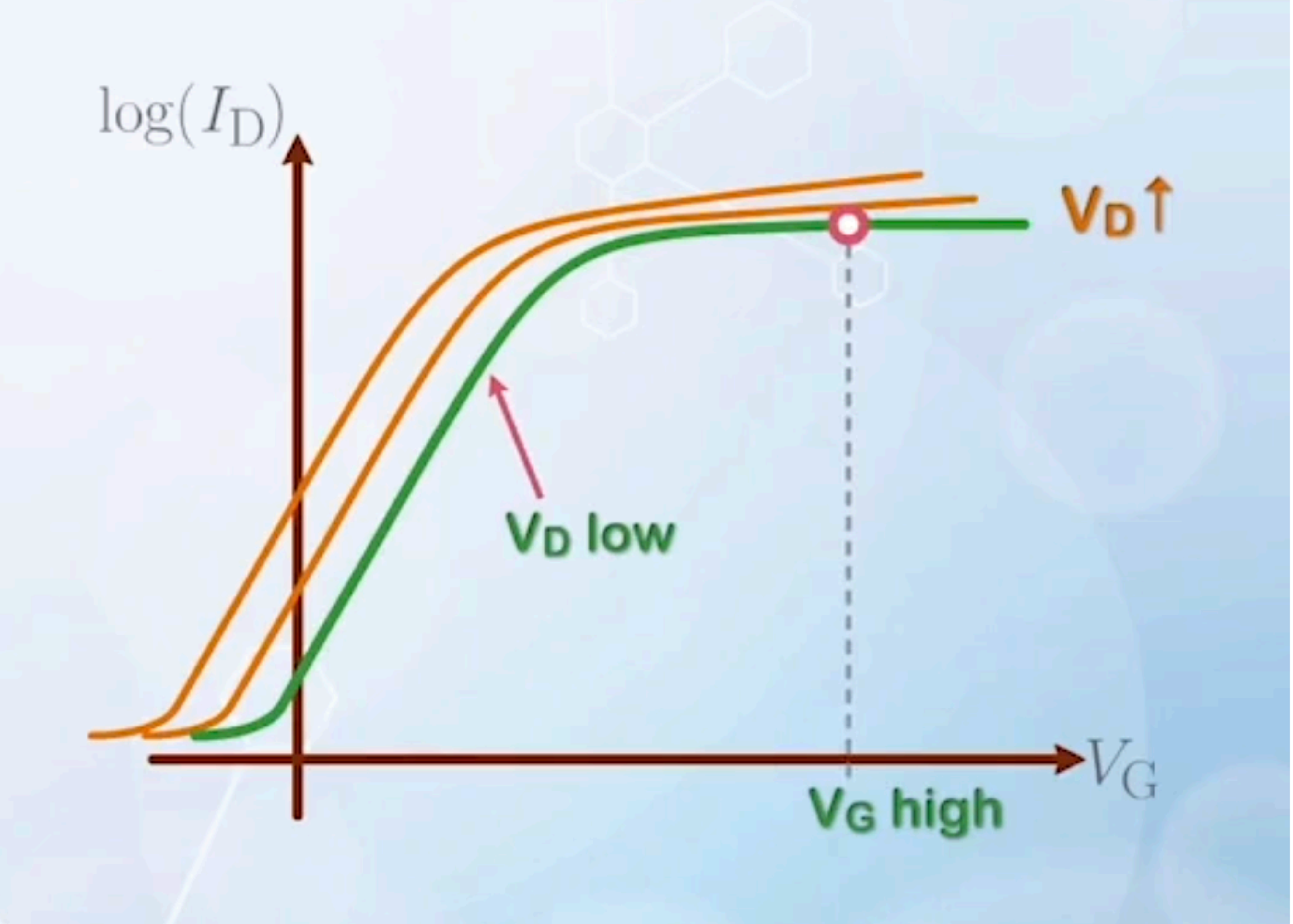
- Increase C_{ox} by **reducing oxide thickness**
  - The effect of source / drain contributed charge on V_{T} is reduced
- Decrease ΔL by **heavy substrate doping**
  - Reduces the proportion of source and drain depleted charge relative to the gate depleted charge
- **Make the source and junction shallower**
  - Reduces x_j
  - Similar to increasing substrate doping to reduce the charge contributed by source and drain

Drain Induced Barrier Lowering (DIBL)

The threshold voltage V_{T} of short channel MOSFETs also reduces with increasing drain voltage V_{D}, instead of remaining constant as in long channel MOSFETs. This is due to **drain induced barrier lowering (DIBL)**.

To turn on a MOSFET, a gate voltage is needed to reduce the source to channel barrier to cause conduction. In a long channel MOSFET, the drain voltage has limited penetration through the channel, and will not affect the source to channel barrier. When the channel length is short enough, the drain voltage can lower the source barrier. In this case, the gate voltage needed to reach threshold condition is lowered. The higher the drain voltage, the more reduction in the threshold voltage is observed.

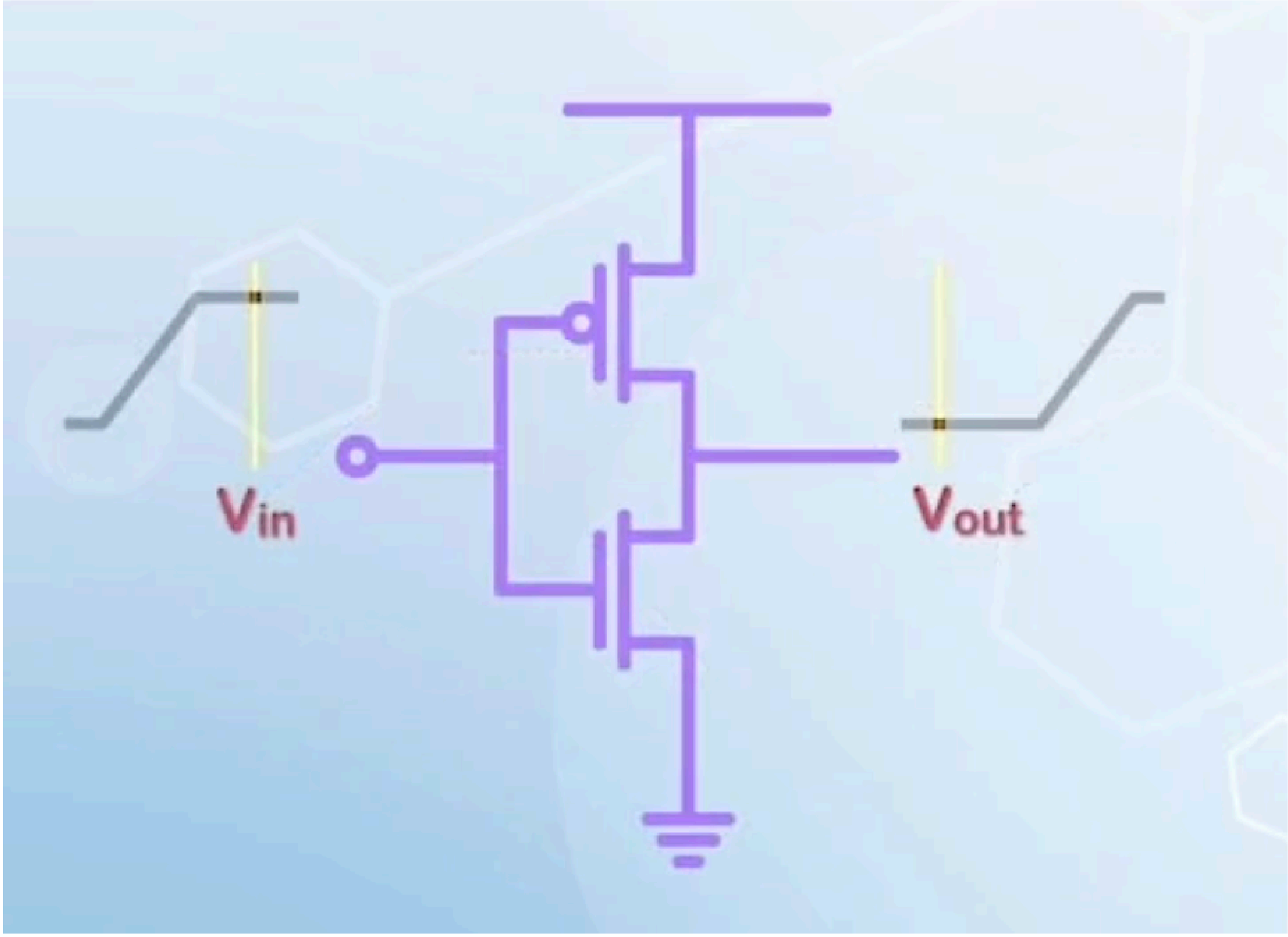
The subthreshold curve of a short channel MOSFET at different V_{D} can be plotted.



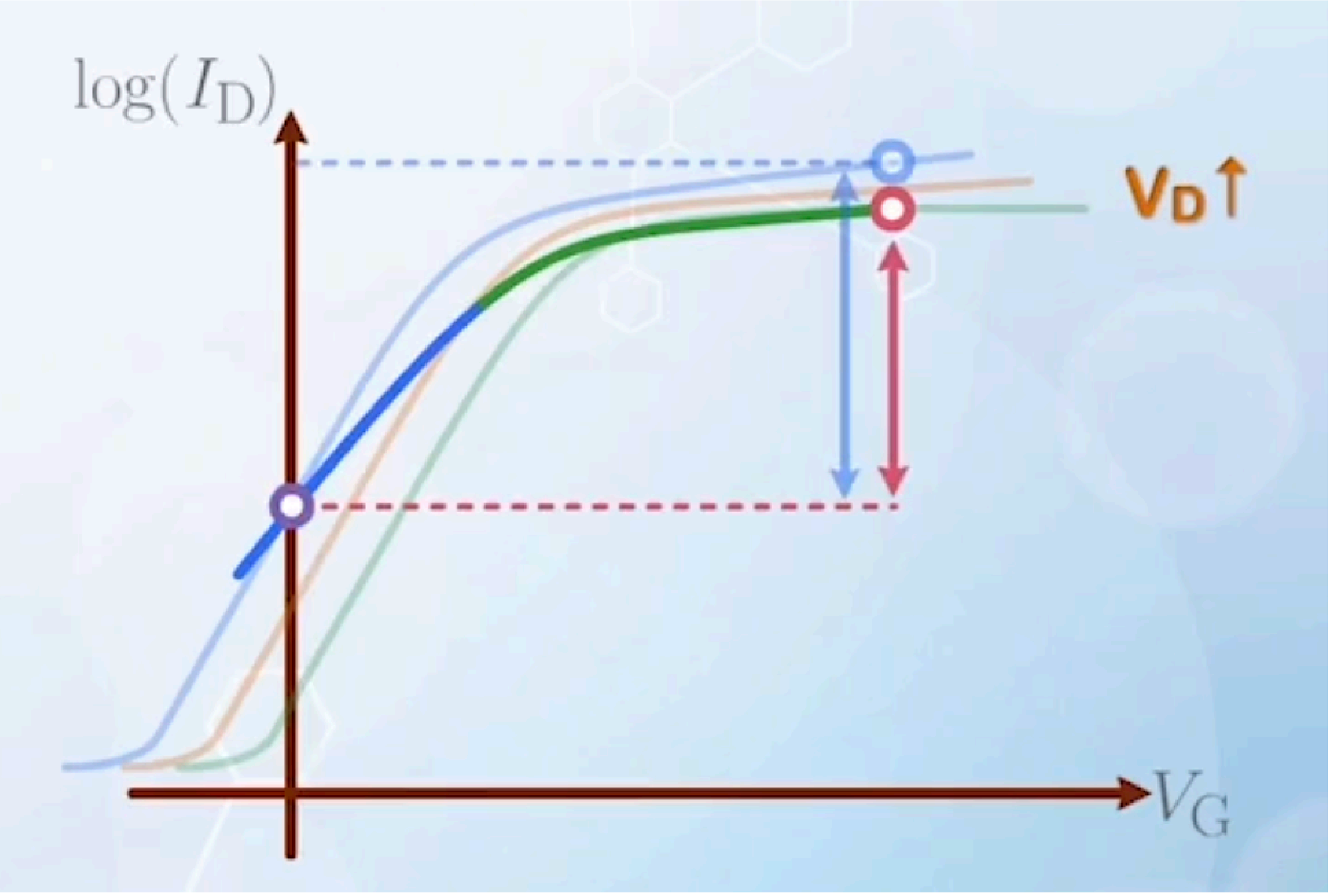
With increasing V_D, the curve shifts to the left because of the reduction in V_T.

The DIBL effect can also dynamically effect the on-off current ratio.

Consider a common CMOS inverter:



- When input is high, NMOSFET is on, and PMOSFET is off
  - Output is pulled to ground
  - Drain voltage of NMOSFET is close to ground
  - NMOSFET follows the curve with high V_T when V_G is high
- When input is low, NMOSFET is off, and PMOSFET is on
  - Output is pulled to V_{DD}
  - Drain voltage of NMOSFET is high
  - NMOSFET follows the curve with high V_D, and V_T is reduced by DIBL
  - Off state leakage current takes the higher value
- Overall, the NMOSFET follows the low current curve when high current is needed, and the high current curve when low current is needed
  - Resulting a much smaller on-off current ratio than that predicted under static condition
  - DIBL causes significant degradation in switching performance



To reduce DIBL, it is equivalent to increase the control of gate over the channel

- Reduce oxide thickness to increase C_{ox}
- Increase substrate doping to reduce depletion region width, increasing C_B
- Reduce drain junction depth to reduce drain capacitance

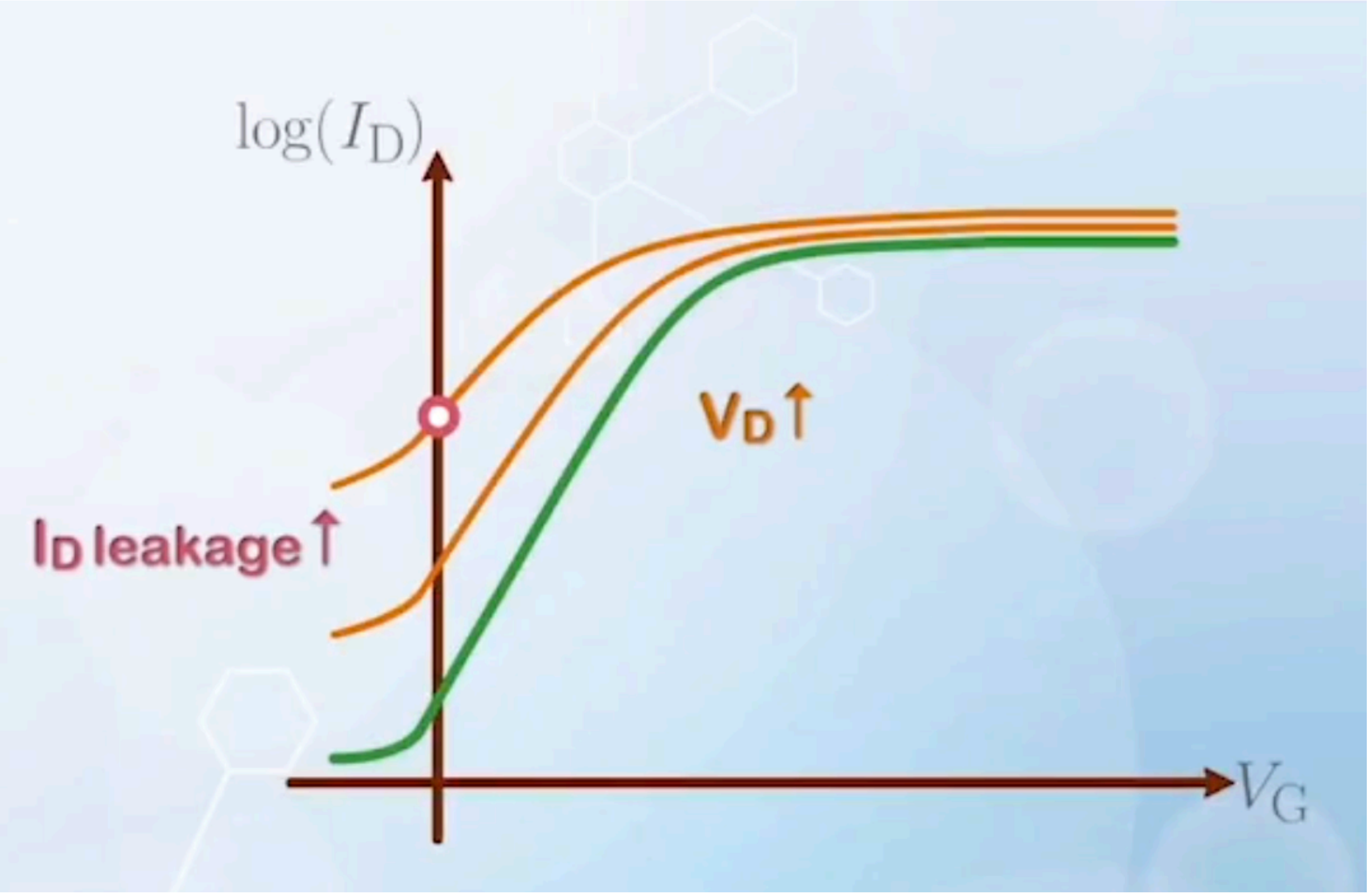
Punchthrough

In DIBL, the drain voltage **does not fully lower** the source to channel barrier, and the gate terminal still controls the energy band of the region under the gate.

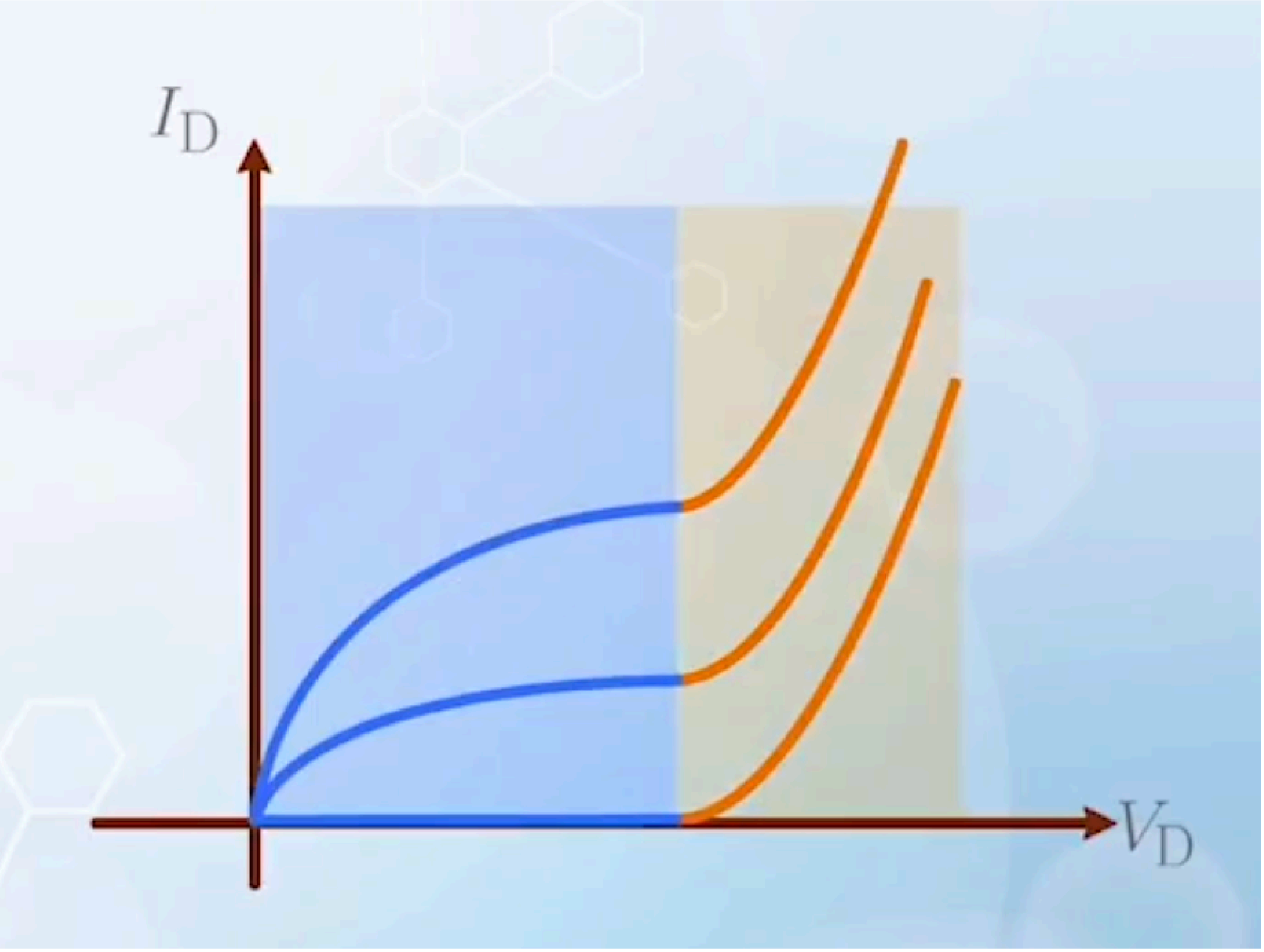
When the channel length is further reduced, the drain voltage may overtake the gate voltage in controlling the energy band under the gate. This is called **punchthrough**.

The control of gate over the channel decreases along the depth of the depletion region, as the capacitance associated with the gate terminal decreases. When channel length is short enough, some region below the surface may be fully controlled by the drain voltage instead of the gate voltage.

The gate has little effect on the energy band at that location, and the conduction is fully controlled by the drain voltage.



With punchthrough, the leakage current increases with increasing drain voltage. Subthreshold swing also increases, as punchthrough weakens the gate control over the drain current. The curve may become flat at high V_D, meaning the current becomes independent of gate voltage, and the gate fully loses control of the drain current.



When V_D is not high enough, the gate still has some control over the channel, and the MOSFET can still be considered functional. However, when V_D increases beyond a certain value, the depletion region of the source touches the depletion region of the drain, the gate loses control, and the conduction is directly controlled by the drain voltage.

To reduce punchthrough, again the gate control over the energy band should be enhanced, while the drain control should be weakened. In addition, the width of drain depletion region should be reduced with heavier substrate doping.

Conclusion

Short channel MOSFETs may become poor switches due to difficulties in turning off. The threshold voltage and the subthreshold swing play an important role in determining the on and off state current of MOSFETs.

In short channel MOSFET design, we want to enhance the gate control over the channel, while reducing the drain control. This can be achieved by

- Thin gate oxide thickness
- Heavy substrate doping
- Shallow source / drain junctions

11. Features in Modern MOSFET Structures

About modern MOSFET design, including source / drain extensions, high-k gate dielectric, metal gate technology, source / drain and channel engineering, and strain engineering.

Modern MOSFET Structure

MOSFETs in the sub-100nm dimensions have many features not found in conventional MOSFETs.

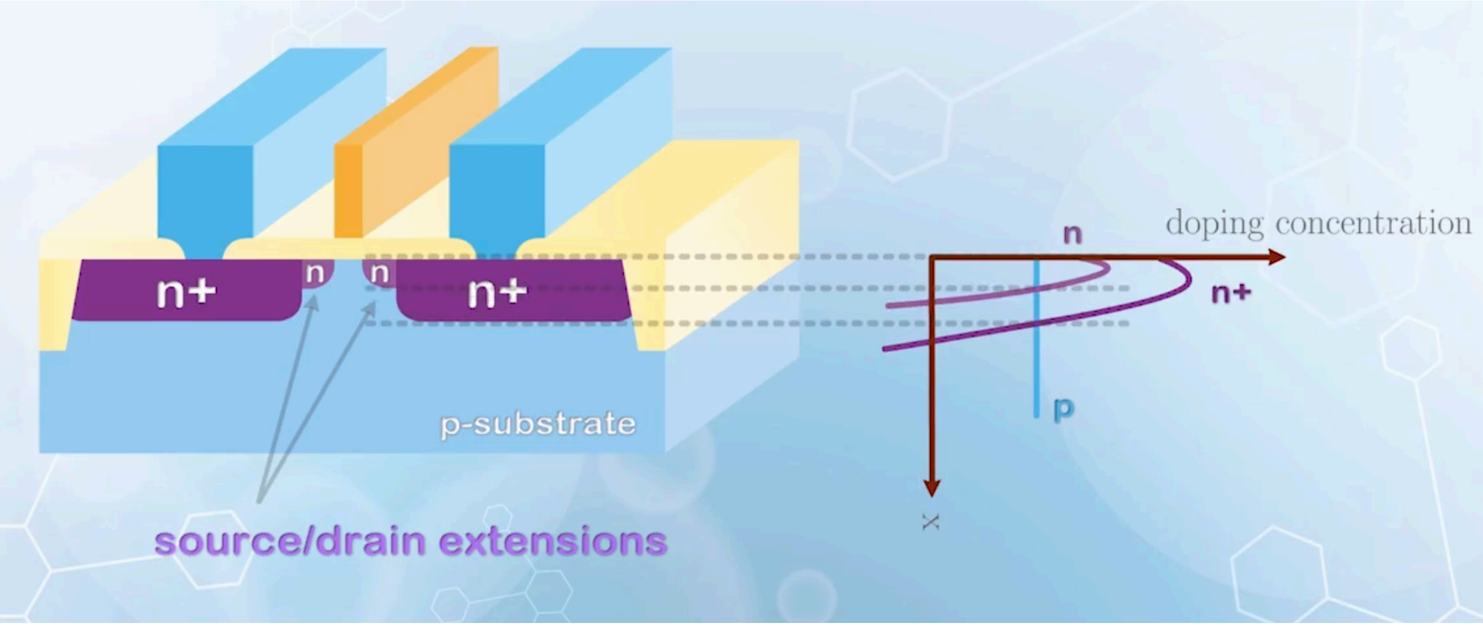
The main features of short channel MOSFETs include:

- Thin oxide
- Heavy substrate doping
- Shallow source and drain junctions

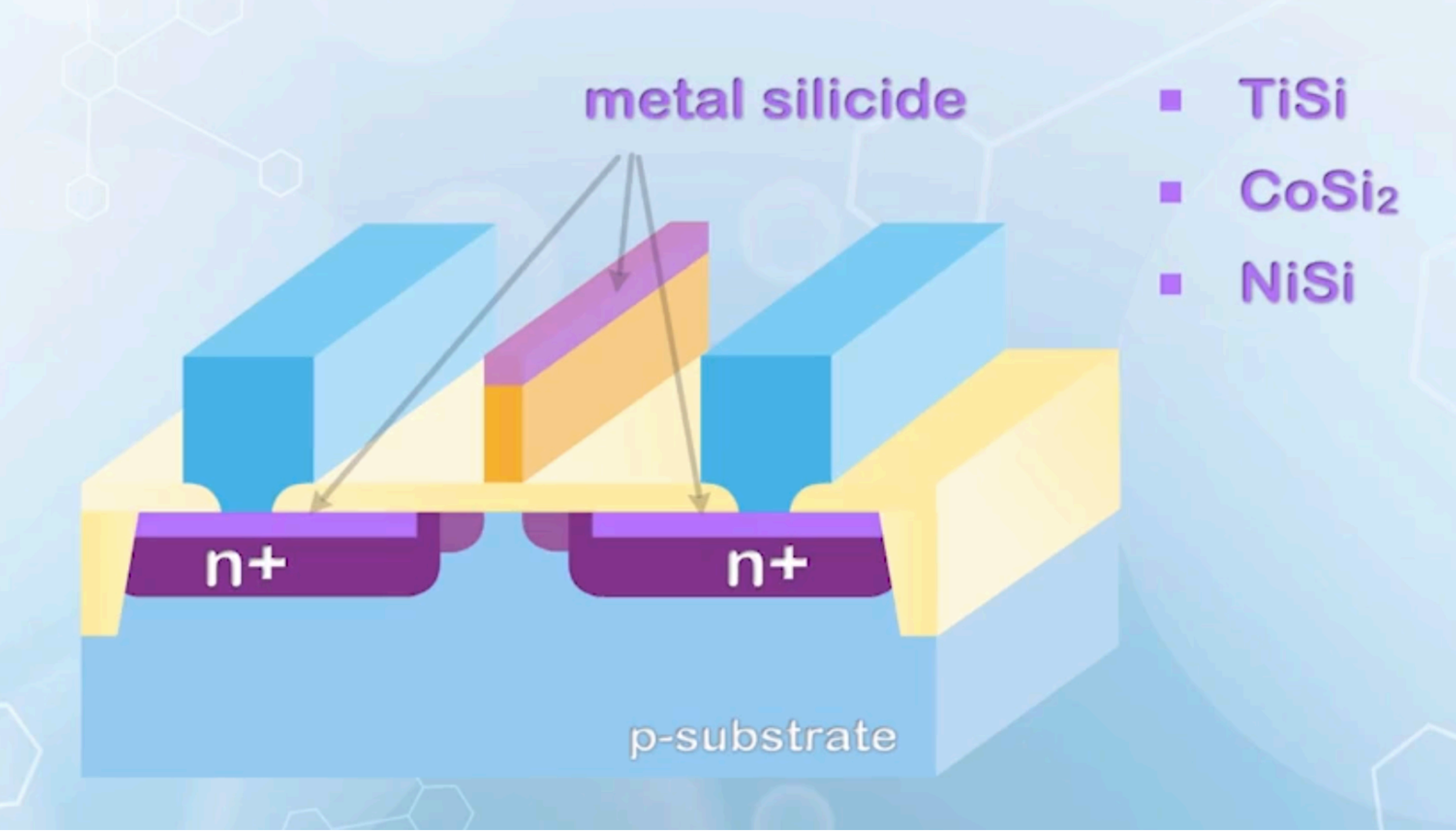
When channel length is reduced, the height of the polysilicon gate is usually maintained to give a large cross sectional area for gate resistance reduction. The gate of a sub-100nm MOSFET usually has a high aspect ratio.

When source and drain junctions are made shallow, the source and drain resistance may increase. As a compromise, the source and drain regions are only made shallow near the channel, and becomes deeper again once further away from the channel region. The shallow regions are called **source / drain extensions**.

The source / drain extensions are usually more lightly doped, as they are formed by ion implantation. The junction depth is determined by the distance where the N doping drops to the background P type doping concentration. A lower doping concentration makes the cross point closer to the surface, resulting in a shallower junction depth.



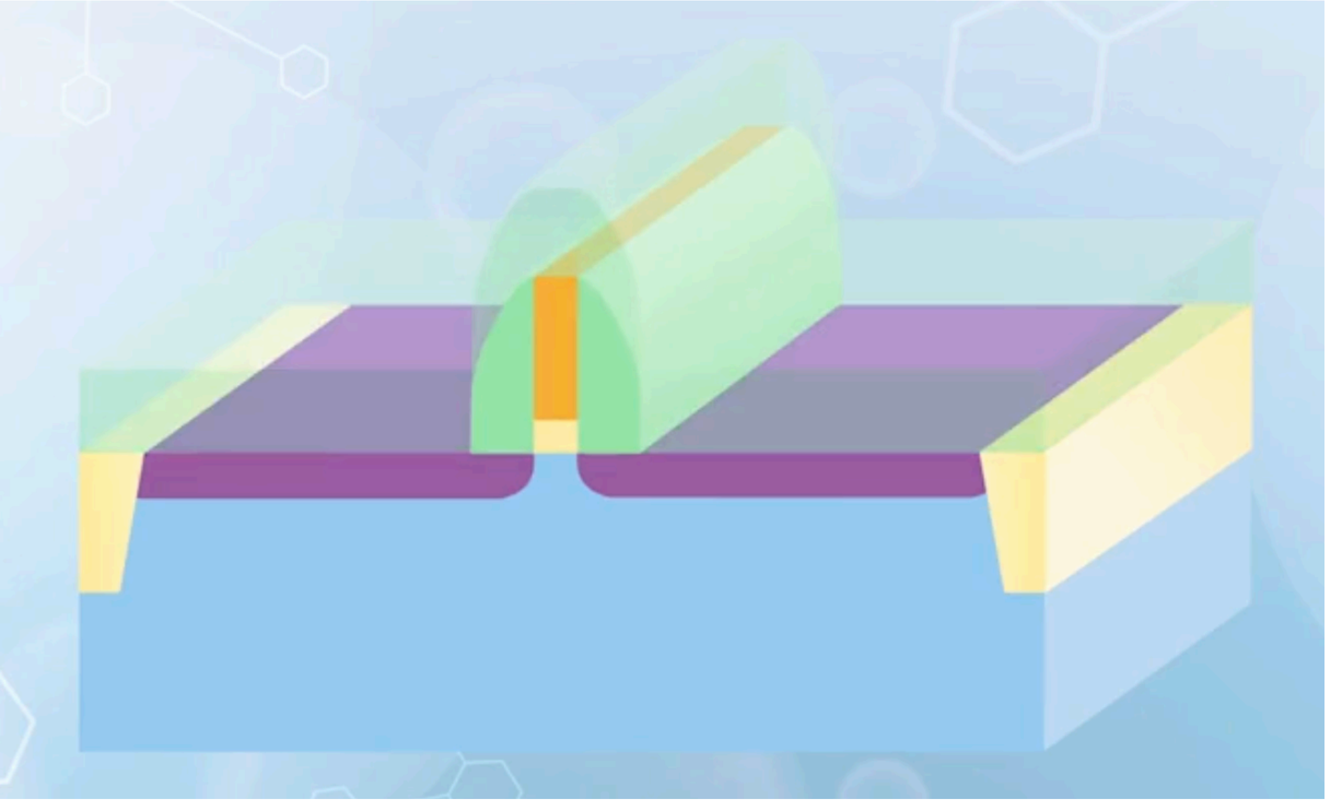
To further reduce the resistance, **silicide technology** was introduced around 250nm technology node. Silicide is a compound formed between silicon and a metal, usually titanium (TiSi), cobalt ($CoSi_2$) or nickel (NiSi). It is a semi-metal with higher conductivity than heavily doped silicon, but slightly higher than that of a metal. Using silicide instead of metal is mainly due to it's process simplicity, as it can be formed selectively on top of silicon.



To form the source / drain extension and the silicide, a spacer technology is used, resulting a layer of insulator formed on the sidewall of the gate.

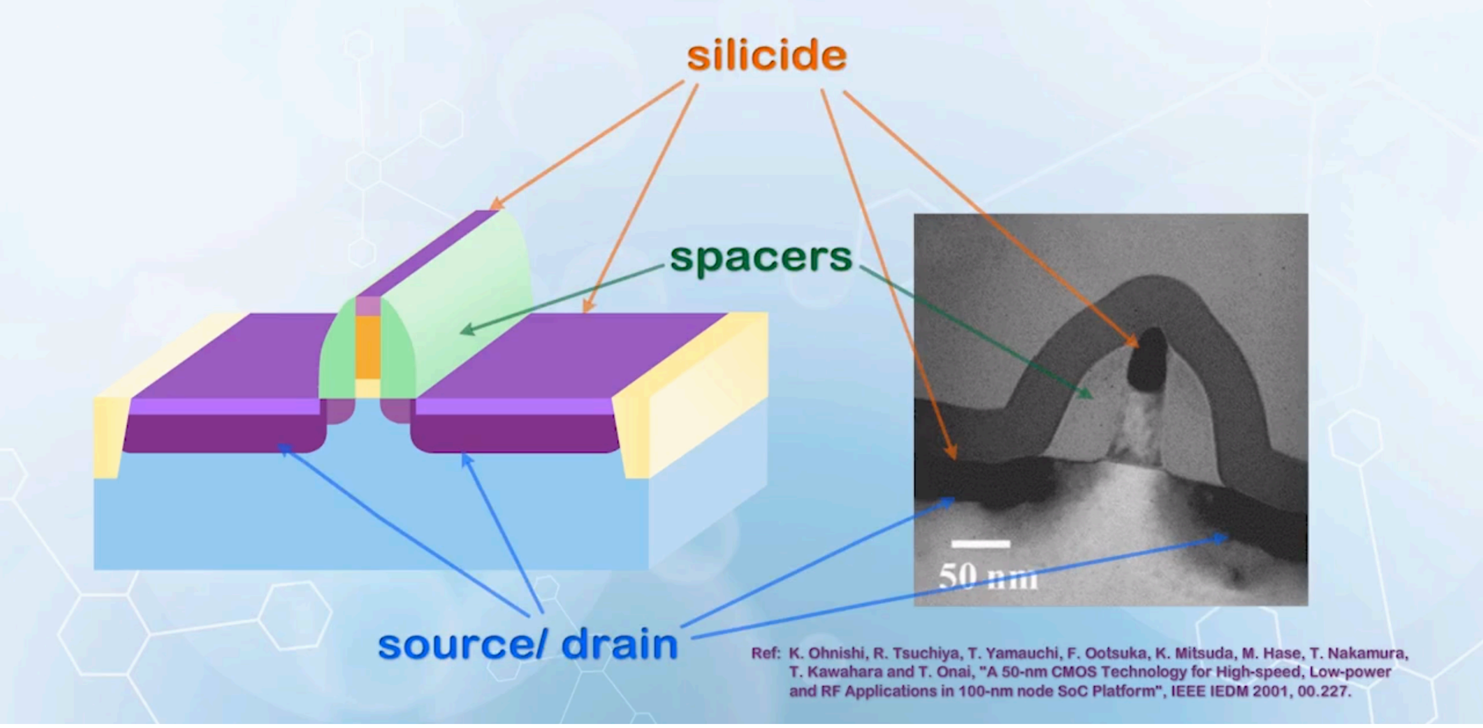
The process is briefly described as follows:

1. The gate is defined first
2. Shallow ion implantation is performed to form the source / drain extensions
3. A thin insulator, usually silicon dioxide or silicon nitride, is deposited conformally over the entire wafer
4. Anisotropic or directional etching is used to move the deposited insulator from top of the wafer
    - Due to the height difference between the insulator next to the gate, an insulator called spacer will be left over next to the gate



5. Second ion implantation is performed, forming the deep source and drain regions
    - The shallow source / drain extension regions are protected by the spacers
6. Thin metal layer is deposited onto the wafer
7. Place the wafer at elevated temperature, and the metal reacts with the underlying silicon, forming silicide. No reaction takes place in the space region
8. Remaining metal is moved with chemical etching, leaving the silicide on the source, drain and gate regions.

This process which forms silicide on the source, drain and gate regions simultaneously is called **self aligned silicide process (SAlicide process)**



Ref: K. Ohnishi, R. Tsuchiya, T. Yamauchi, F. Ootsuka, K. Mitsuda, M. Hase, T. Nakamura, T. Kawahara and T. Onai, "A 50-nm CMOS Technology for High-speed, Low-power and RF Applications in 100-nm node SoC Platform", IEEE IEDM 2001, 00.227.

High-k Gate Dielectric

The traditional approach to CMOS transistor scaling (i.e., reducing gate oxide thickness, increasing substrate doping, decreasing source / drain junction depth) works very well for large devices. But at 45nm technology node, the gate oxide thickness is already reduced to 1.2nm, or about 4 layers of atoms in the oxide layer. Further reduction of gate oxide thickness is extremely difficult. At such thickness, the gate leakage current may become very high, and the gate oxide may not be able to serve as an insulator.

However, if we trace back to the reason why we use a thin oxide layer, it is to increase the gate capacitance for stronger gate control. The normalized gate capacitance is given by

$$C_{ox} = \frac{\kappa \varepsilon_0}{t_{ox}}$$

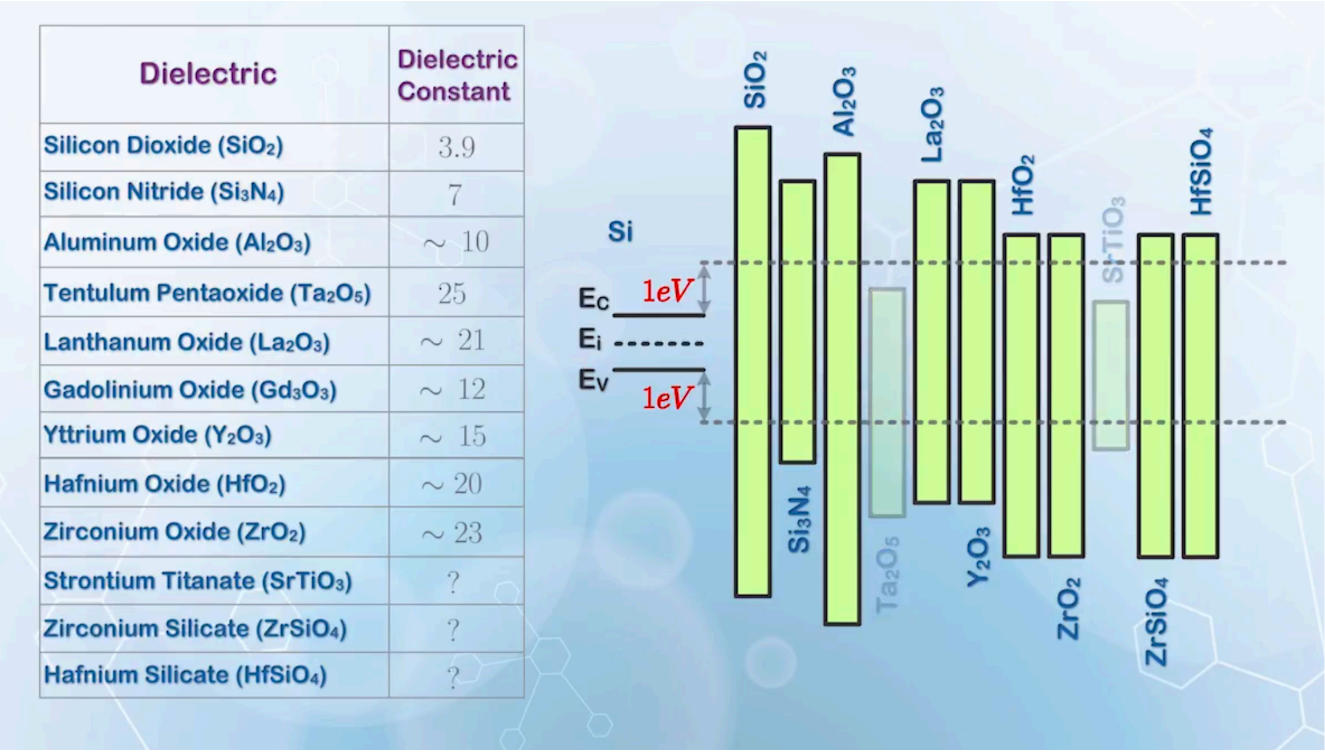
where κ is the dielectric constant of the dielectric material.

When reducing the oxide thickness is no longer feasible, we can instead increase the dielectric constant of the gate dielectric material. This leads to the search for another material as an alternative to replace silicon dioxide as the gate dielectric.

The new material should have the following properties:

- The dielectric should have a high barrier for both electrons and holes

  - To prevent charge carriers from jumping over the insulator, an offset larger than $1\,\mathrm{eV}$ is required for both the conduction band and valence band



Dielectric	Dielectric Constant
Silicon Dioxide (SiO_2)	3.9
Silicon Nitride (Si_3N_4)	7
Aluminum Oxide (Al_2O_3)	$\sim$ 10
Tentulum Pentaoxide (Ta_2O_5)	25
Lanthanum Oxide (La_2O_3)	$\sim$ 21
Gadolinium Oxide (Gd_3O_3)	$\sim$ 12
Yttrium Oxide (Y_2O_3)	$\sim$ 15
Hafnium Oxide (HfO_2)	$\sim$ 20
Zirconium Oxide (ZrO_2)	$\sim$ 23
Strontium Titanate ($SrTiO_3$)	?
Zirconium Silicate ($ZrSiO_4$)	?
Hafnium Silicate ($HfSiO_4$)	?

- The material should be able to withstand the high processing temperature without decomposition
- The material should be able to form a smooth and low defect interface with silicon to reduce channel carrier mobility degradation
  - This is very difficult compared to silicon dioxide, which is known to form the best interface with silicon
  - The solution: forming one to two layers of silicon dioxide atoms on the silicon substrate first, before depositing the high-k dielectric material

Among other choices, silicon nitride ($\mathrm{Si_3N_4}$) is first used in production, due to its compatibility with the silicon process. But its k value is not high enough.

After another round of elimination, hafnium based oxide ($\mathrm{HfO_2}$, $\kappa > 20$) stands out as the most widely used gate dielectric material.

By using thicker hafnium oxide to achive the same gate capacitance as that with thin silicon dioxide, the gate leakage current can be reduced by a factor of $10^4 - 10^5$

The **equivalent oxide thickness (EOT)** is used to compare among different gate dielectric materials. It is the thickness of silicon dioxide required to give the same normalized capacitance as that of the gate capacitor formed by high k materials.

$$
\begin{aligned}
C_{\mathrm{G}} = \frac{\kappa \varepsilon_0}{t_{\mathrm{g}}} &= \frac{\kappa_{ox}\varepsilon_0}{\mathrm{EOT}} \\
\Rightarrow \mathrm{EOT} = \frac{\kappa_{ox}}{\kappa} t_{\mathrm{g}}
\end{aligned}
$$

While using high k gate dielectric can increase the gate capacitance without causing significant gate leakage current, it is not the same as reducing the gate dielectric thickness.

The capacitance given by $C_{\mathrm{G}} = \frac{\kappa\varepsilon_0}{t_{\mathrm{g}}}$ assumes the parallel plates of the capacitor have an infinite area, ignoring the fringing electric field. In nano CMOS, it may not be the case, as the lateral dimension can be comparable to the gate dielectric thickness, and the fringing field can become significant.

When the gate dielectric is thin, the gate blocks most of the electric field from the drain. When the k value increases, the dielectric thickness is also increased to give the same EOT. This causes more electric field from the drain to affect the channel potential, weakening the gate control. Thus using a high k gate dielectric can increase the gate capacitance and reduce the gate leakage, it may not be as effective as what is predicted by the simple model.

Metal Gate Technology

The overall capacitance from the gate electrode to the channel of the MOSFET not only consists of the capacitance of the gate dielectric, but also includes the capacitance due to the finite thickness of the channel inversion charge C_{inv}, which we do not have much control over and is simply ignored. The polysilicon gate also contributes to a series capacitance C_{poly} due to the polysilicon gate depletion effect.

So far, we have assumed that the polysilicon gate is much more heavily doped than the substrate, and its depletion region width can be ignored. But when the gate oxide thickness is reduced to a few atom layers, the polysilicon gate depletion region thickness can no longer be ignored.

If we only increase C_{ox}, but do nothing with C_{poly}, the effect will not be obvious. To eliminate the polysilicon gate depletion at the gate electrode, metal gates are adopted at around 45nm technology node. As a second benefit, metal gates also reduces the gate resistance of the MOSFET.

Choosing the metal for gate electrode needs more careful consideration. In addition to the gate capacitance, it also affects the threshold voltage of the MOSFET, as the work function of the metal affects the flat band voltage V_{FB}, which is a component of the threshold voltage V_{T}.

In conventional CMOS technology, the work function of the polysilicon gate is adjusted by doping, so that we have N+ polysilicon gate for NMOSFET and P+ polysilicon gate for PMOSFET.

If we use a metal with work function similar to that of N+ polysilicon, V_{T} of PMOSFET will be too negative, making it difficult to turn on. Similarly, if we use a metal with work function similar to that of P+ polysilicon, V_{T} of NMOSFET will be too positive.

To achieve symmetrical V_{T} for both NMOSFET and PMOSFET, metal with work function near the middle of the band gap of silicon can be used, but such symmetrical V_{T} ($\approx +\mathbf{0.55\,V}$ for NMOSFET and $\approx -\mathbf{0.55\,V}$ for PMOSFET) is too high for both types of MOSFETs. With the reduction in supply voltage, V_{T} needs to be scaled together to maintain a reasonable on state current, and the increase in V_{T} caused by the use of single type metal for the gate electrode is not desirable.

To achieve optimum V_{T}, two different metals with different work functions are needed. Some choices include:

- Tantalum nitride **TaN** for NMOSFET
    - With **low** work function, and is closer to the **conduction band** of silicon
- Tungsten nitride **WN** for PMOSFET
    - With **high** work function, and is closer to the **valence band** of silicon

While structurally simple, such method significantly increases the fabrication complexity.

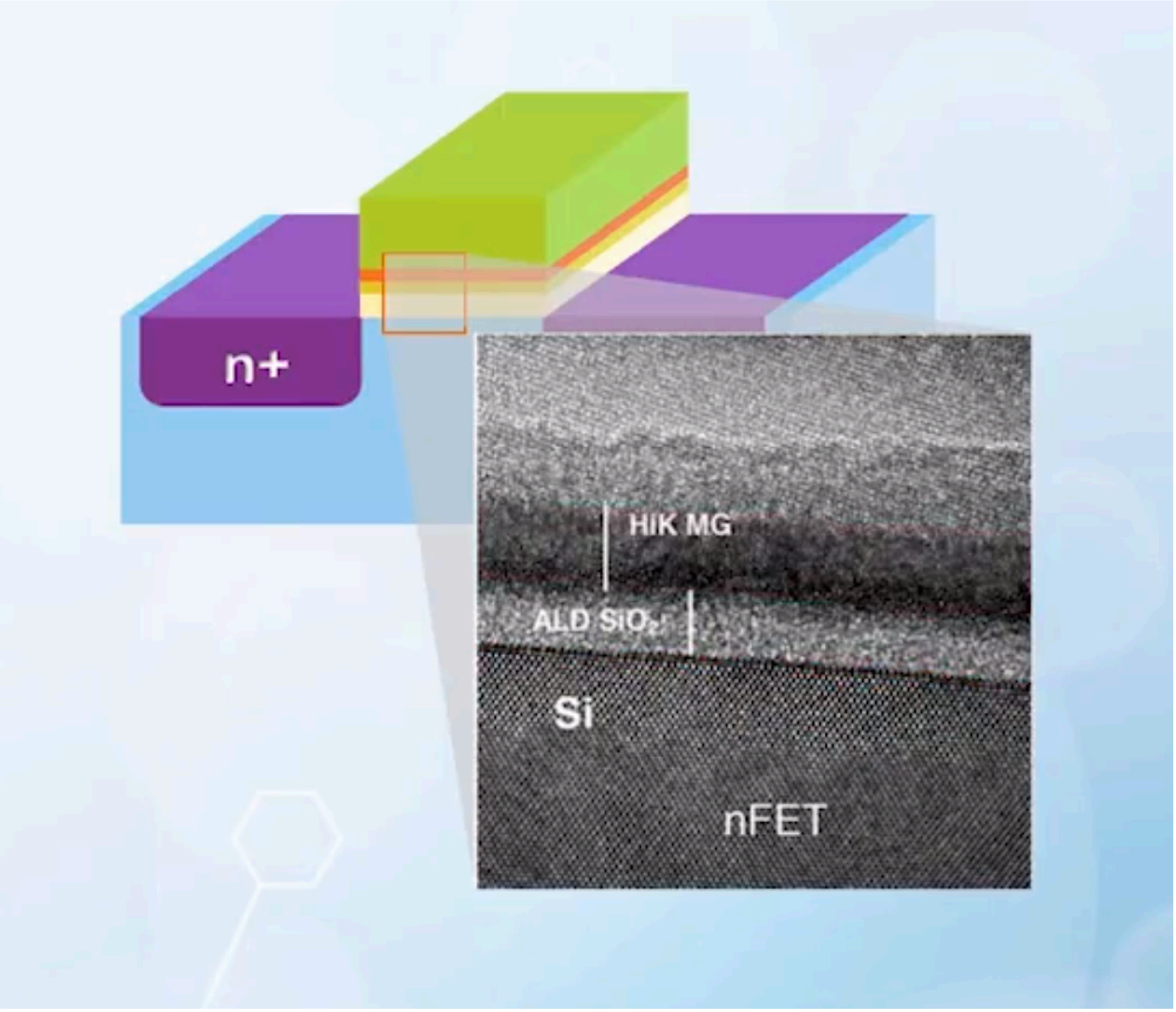
If using two different metals is not desirable, **dipoles** can be used to adjust the the threshold voltage.

Dipoles at the oxide-electrode interface can be introduced by some nano-scale capping layers before depositing the metal gate. Some choices include:

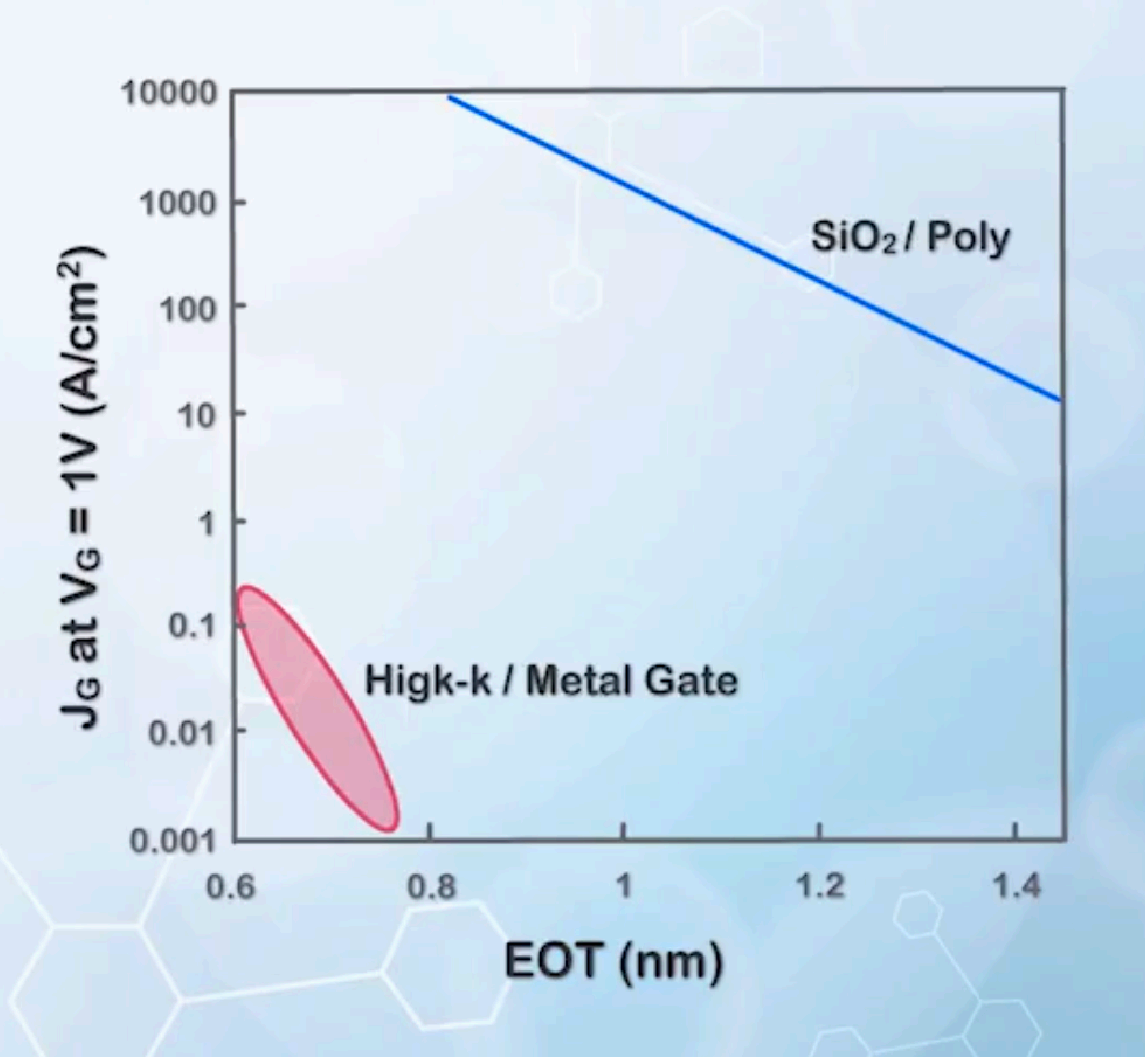
- Lanthanum oxide ($\mathbf{LaOx}$) for NMOSFET
- Aluminum incorporated hafnium oxide ($\mathbf{Al + HfOx}$) for PMOSFET

The gate electrode of the MOSFETs are not formed by a single material, but a complex, multi-material gate stack, together with the high k gate dielectric. Features of such gate stack include:

- A smooth interface to achieve high channel carrier mobility
- A high k gate dielectric layer to achieve high gate capacitance with low gate leakage current
- Proper metal work function to achieve the correct V_{T}



By switching to such high-k / metal gate structure, we can achieve a lower EOT with lower gate leakage current. This is why it becomes the dominant MOSFET technology in spite of its high cost.



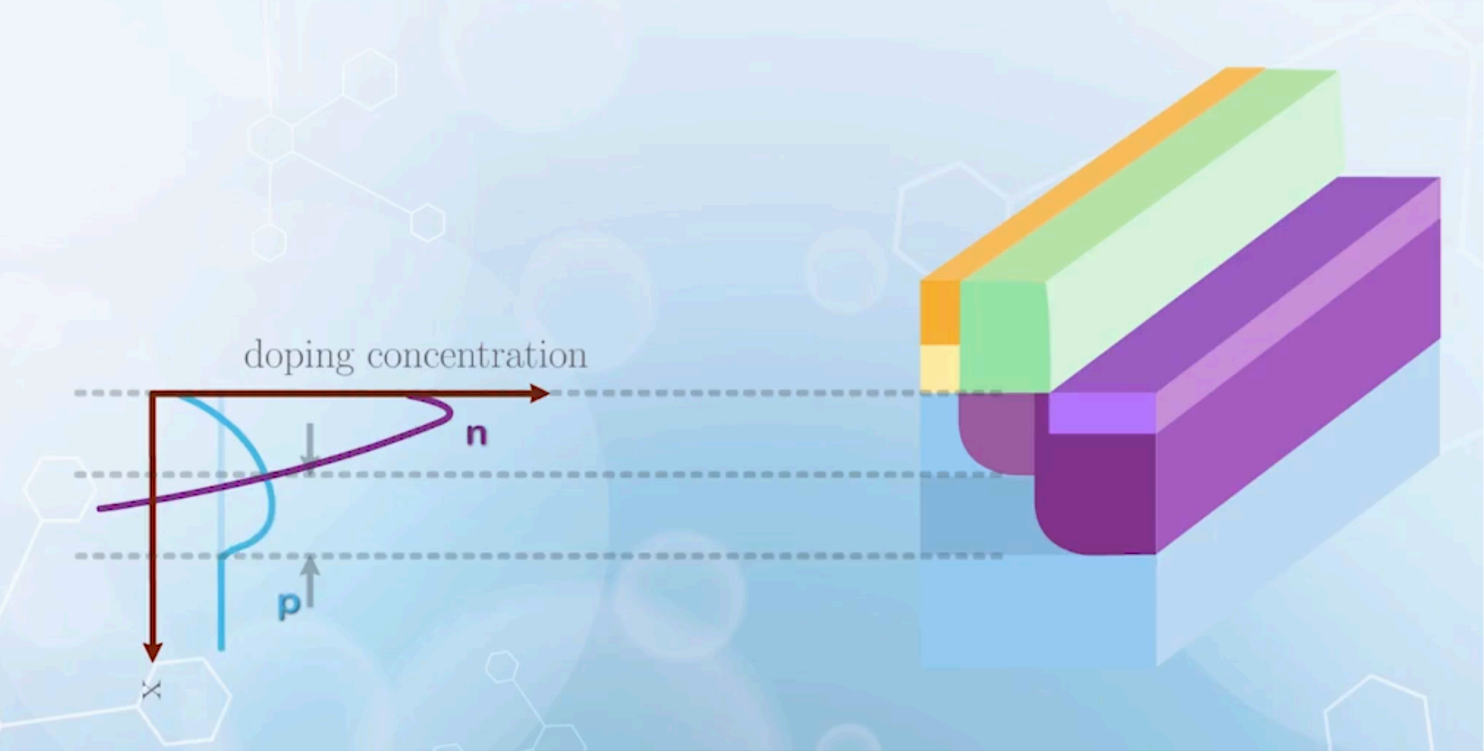
Source / Drain and Channel Engineering

Controlling the source and drain junction depth is no easy task. As there are many high temperature processing cycles during the formation of the MOSFET, and the dopant ions (especially for small dopant ions like boron B) tend to diffuse down to the substrate, hence increasing the junction depth.

Dopant diffusion is one of the main issues in forming shallow junctions, and the most straightforward way to reduce the diffusion rate is to use heavier dopant atoms, such as indium In as the p-type dopant and antimony Sb as the n-type dopant. However, due to the size mismatch between these large dopant atoms and the silicon atoms, the solid solubility of these dopants in silicon is usually only around $10^{19}\ \mathrm{cm}^{-3}$. This results in low doping concentration and high source and drain resistance.

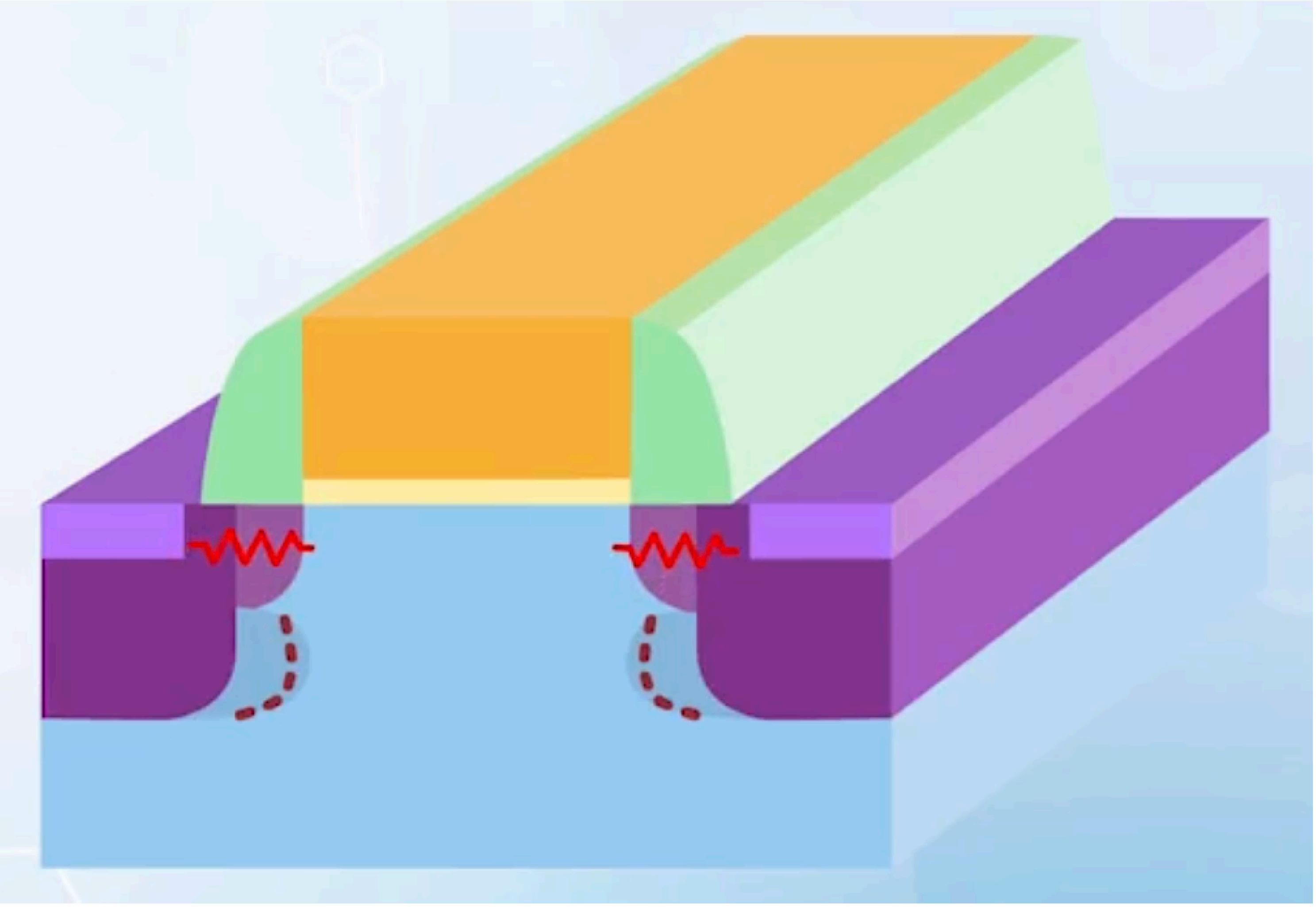
Another approach to reduce junction depth is to raise the background doping concentration, so that the distance for the source and drain doping concentrations to drop to the background level is reduced. But this also increases the V_{T} of the MOSFET. To compensate for this, we can increase the doping **only in the region below the channel**, and reduce the doping of the channel near the surface, to maintain the average doping concentration approximately the same. The region or well with lower doping near the surface of the channel, and high doping concentration in the region further below the surface is called the **retrograde well**.

Drawback of this method is the decreased depletion width of the PN junction formed by source / drain and the substrate, will increase the junction capacitance between source / drain and the substrate terminal. Thus, the heavy doping is usually confined to the region **below the source / drain extensions**, and **above the deep source / drain junction depth**.



To further optimize the doping profile of the heavily doped region without affecting V_{T}, heavy doping can be applied only to the region below the source / drain extensions. Such structure with a heavy P doping below the N type source / drain extensions is called **halo structure** or **pocket implanted structure**. Another benefit of the halo structure is, by increasing the doping concentration near the source / drain extensions, the depletion width of the deep source / drain is reduced, allowing the source and drain to be placed closer together without causing punchthrough.

We can also use shorter source / drain extensions, and reduce the series resistance of the source and drain.



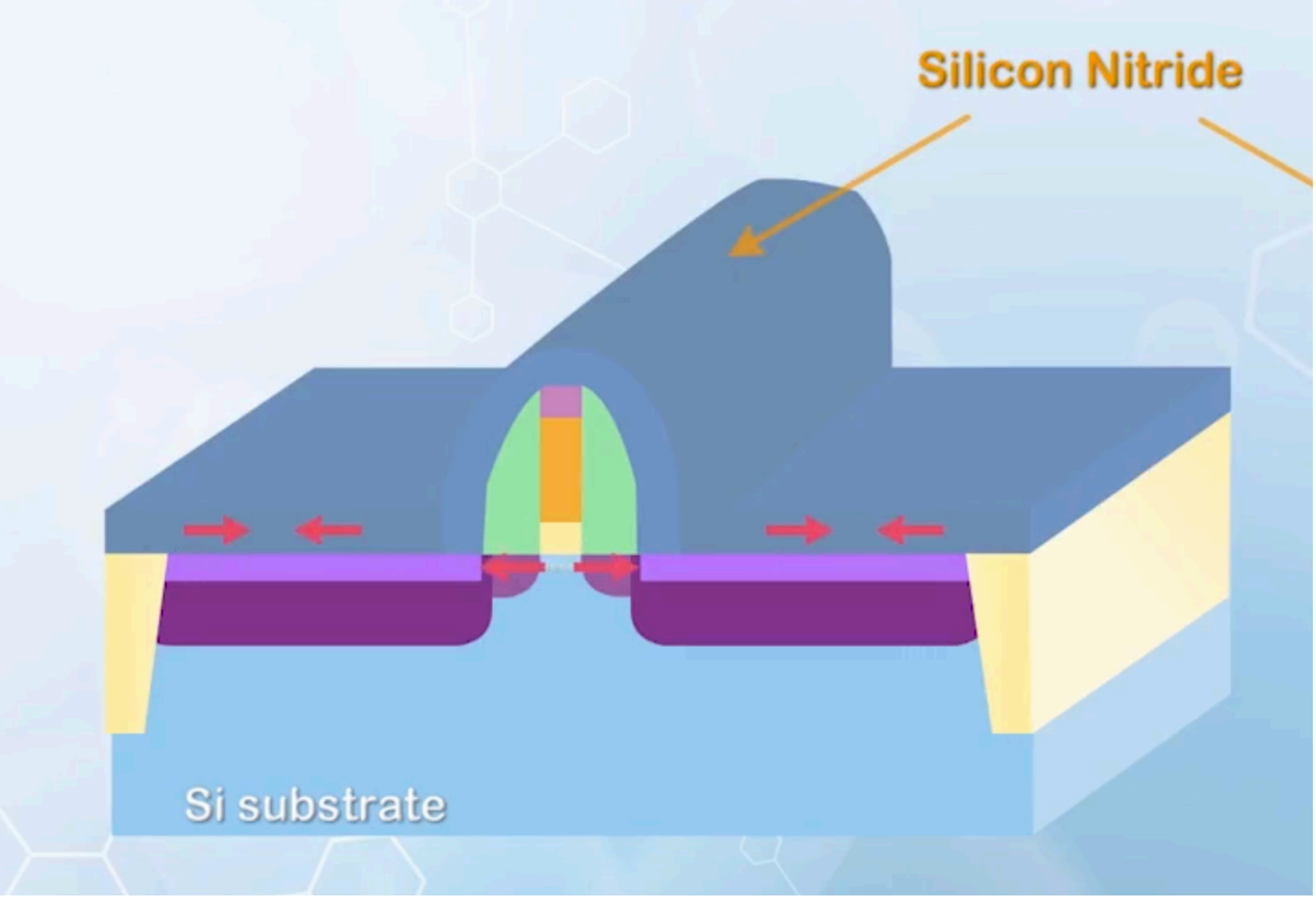
Strain Engineering

All above methods are used to reduce short channel effects. But it is also important to increase the on state current of extremely scaled MOSFETs.

Strain engineering is introduced to increase the channel carrier mobility without changing its structure or using a new material system. Understanding strain engineering requires knowledge of quantum mechanics and the energy momentum ot E-k space, which is not covered in this course. Instead, a more intuitive explanation is given here, which may not be entirely correct in strict physics sense.

When electrons move in silicon crystal, they will experience a stronger attraction at the location of an atom, which slows them down. If spacing between atoms are increased, electrons can move a longer distance before being slowed down by the atoms, resulting in an increase in the average speed of the carriers. When atoms in the silicon crystal are forced to stay at a larger distance apart, the silicon crystal is said to be under **tensile strain**.

This can be achieved by growing a thin layer of silicon on top of a similar lattice structure, but a larger atomic spacing. In early days, silicon germanium (SiGe) layer, which has a similar structure to silicon, but larger average atomic spacing due to larger germanium atoms, is used to form strained silicon. More recently, tensile strain is achieved with materials with smaller atomic spacings, like silicon nitride, on top of the source and drain region of the MOSFET. By exerting a compressive stress in the source and drain regions, tensile strain is induced in the channel region. This can almost double the mobility.



For PMOSFET, the hole mobility can be increased by compressive strain instead of tensile strain, as the nature of hole movement is to move along the bonds between the silicon atoms. It is easier to find an electron to take place of the hole, hence causing the hole to move more smoothly.

Compressive strain can be achieved by implanting germanium to the source and drain regions of the MOSFET, causing a crystal expansion in those regions, which exerts a compressive stress on the channel region. A double increase in carrier mobility is also possible with this method.

On / off Current Tradeoffs

The current is given by

$$I_{\mathrm{D}} = Q(y)v(y)$$
$$Q \propto C_{ox}(V_{\mathrm{G}} - V_{\mathrm{T}})$$

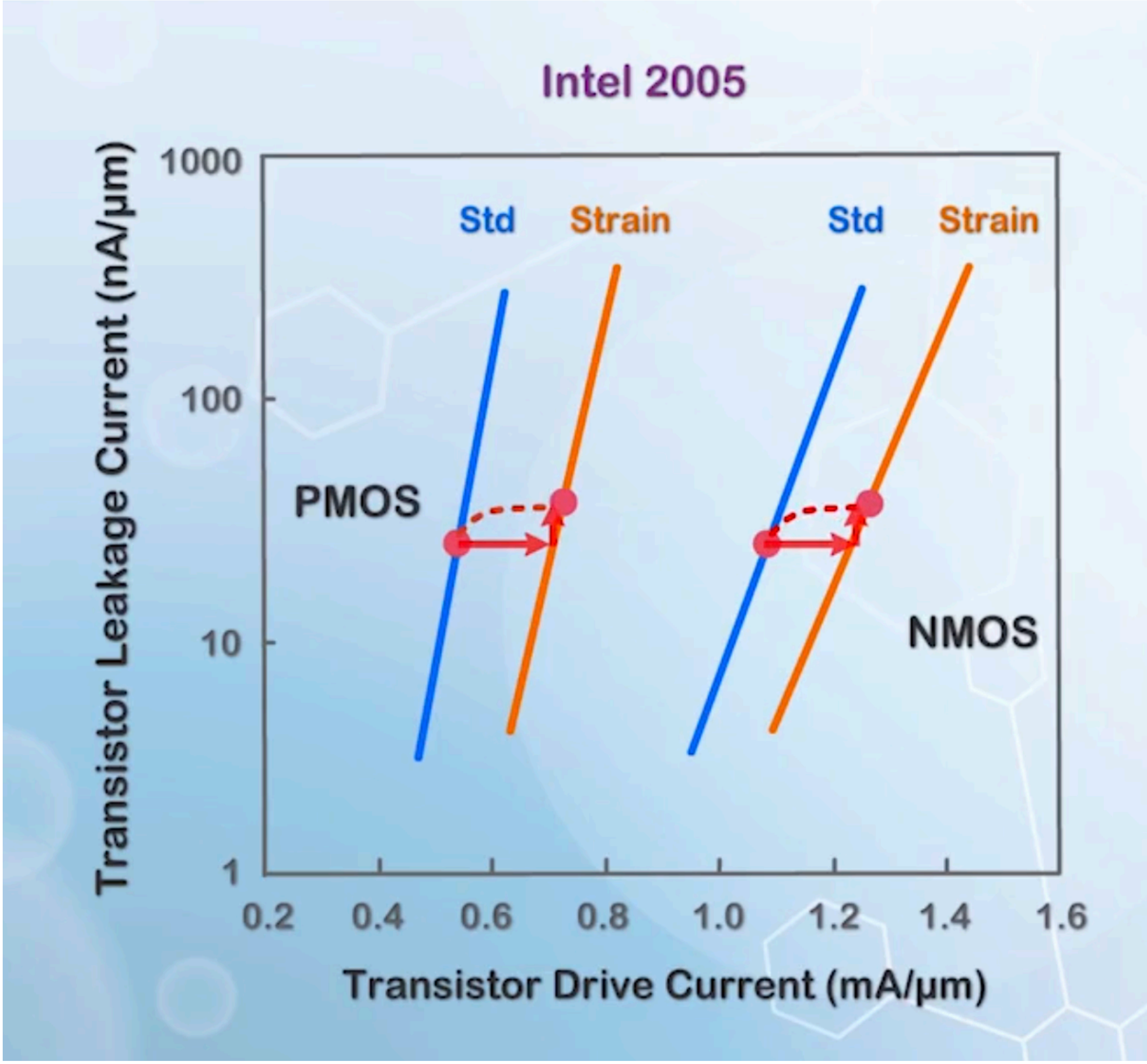
Increasing current by increasing C_{ox} has been discussed before, but we can also increase Q by decreasing V_{T}.

However, decreasing V_{T} will also increase the off state current $I_{\mathrm{D\ off}}$ exponentially, as every reduction of $60n\,\mathrm{mV}$ in V_{T} will increase $I_{\mathrm{D\ off}}$ by a factor of 10. The on state current increases linearly with V_{T} reduction, but the off state current increases exponentially. Thus, this is not desirable.

The velocity $v(y)$ before saturation can be approximated by

$$v = \mu_{\mathrm{eff}} \frac{V_{\mathrm{D}}}{L}$$

Increasing v by decreasing L has also been discussed before, and strain engineering can also increase the effective mobility μ_{eff}, hence increasing v. This approach is more desirable, as it only increases the off state leakage current linearly instead of exponentially, thus the on-off current ratio is maintained.



The increase in the on state current is represented by the lateral shift, while the increase in off state current is represented only by a very small vertical shift. It is similar to increasing the width of the MOSFET to increase the current, but without increasing the input capacitance and the footprint taken up by the MOSFET.

12. Advanced Nano-CMOS Technologies

About more advanced nano-CMOS device structures, including silicon-on-insulator MOSFETs, double-gate and multi-gate MOSFETs, tunneling MOSFETs, junctionless transistors, and 2D material based transistors.

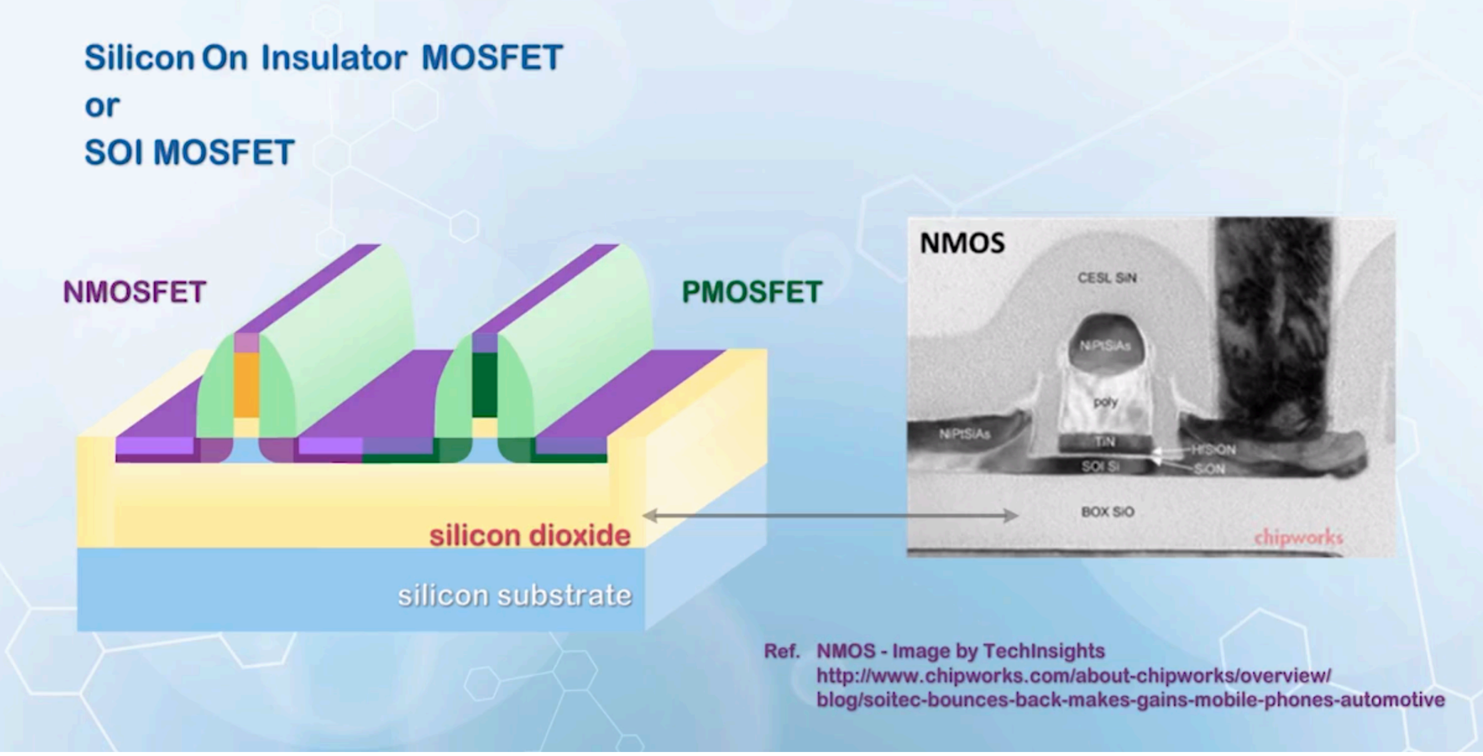
Silicon-on-insulator MOSFETs

Conventional planar MOSFET structures has been used till around 32nm technology node. In 22nm technology node, Intel announced the use of a 3D multi gate transistor architecture. Many new device concepts are proposed at the same time to enhance the current CMOS technology. Some entirely new application, such as flexible electronics, also require new device structures.

In a MOSFET, only a thin layer of channel is useful for the switching functionality, and the structure below the channel is only for physical support. Even worse, it provides a leakage path between the source and drain by creating regions under weak gate control and enhances the short channel effects. If the substrate can be removed, the leakage path can be eliminated, and the MOSFET will be more resistant to the short channel effects. This can be achieved by cutting the top part of a MOSFET and placing it on an insulator, usually silicon dioxide. Such structure is called an **silicon-on-insulator (SOI) MOSFET**

Benefits of SOI MOSFETs include:

- Regions with weak gate control are eliminated, leading to stronger gate control over the current path between the source and drain regions
    - The SOI structure can be scaled to very small dimensions when the silicon film is thin enough
- Forming shallow source and drain junctions will be easy, as the junction depth automatically follows the silicon film thickness
- The insulator isolates the transistor from the substrate, leading to smaller junction capacitance between the source / drain and the substrate
    - Smaller parasitic capacitance load, faster speed
- Elimination of the substrate allows the isolation of MOSFETs by dividing them into islands, without the need of wells
    - NMOSFET and PMOSFET can be placed next to each other with closer spacing
    - Chip area can be reduced



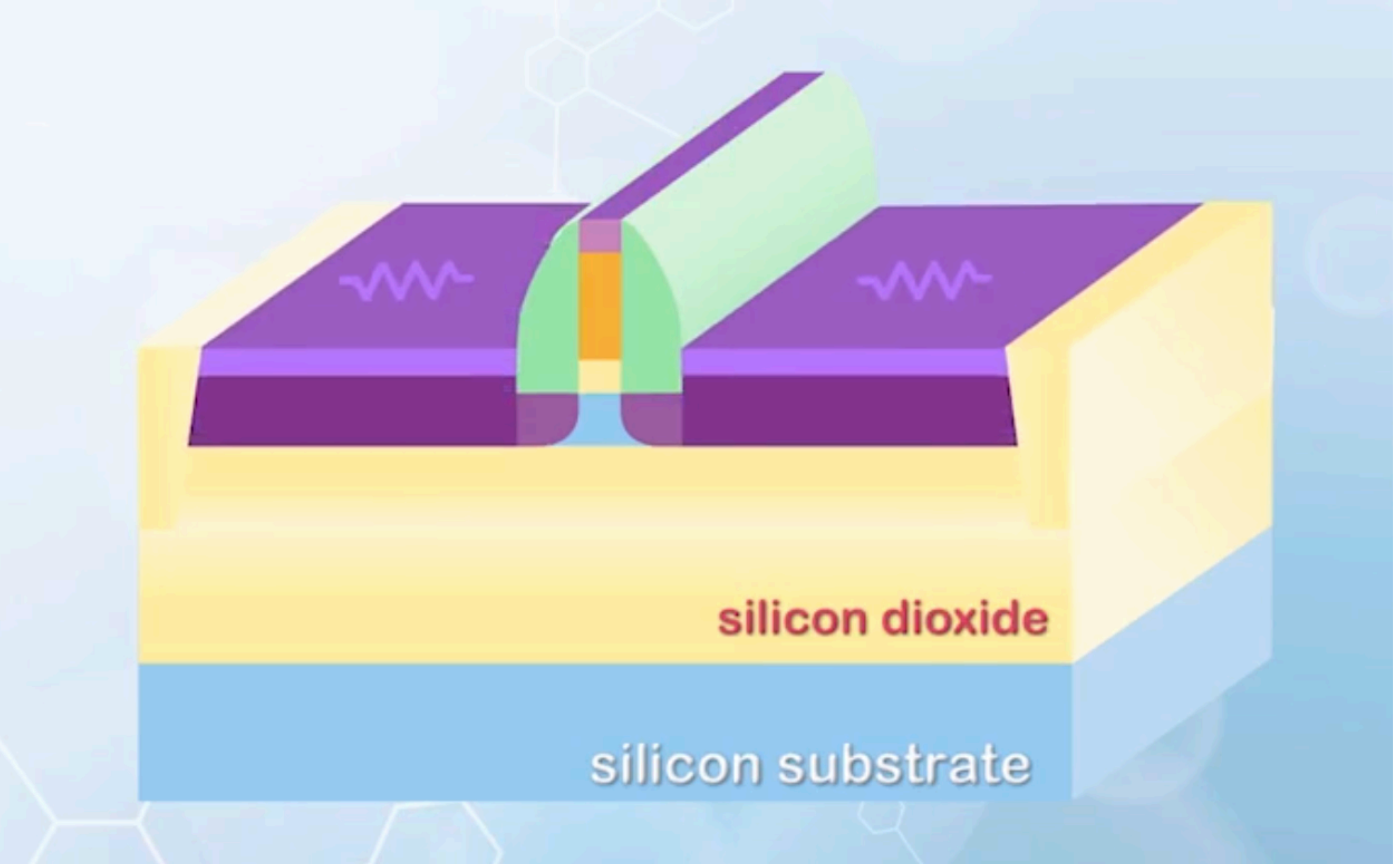
Ref. NMOS - Image by TechInsights
http://www.chipworks.com/about-chipworks/overview/blog/soitec-bounces-back-makes-gains-mobile-phones-automotive

Difficulty: Forming high quality silicon film on insulator is not easy. Depositing silicon onto the insulator to form the silicon film is not applicable, as it gives polycrystalline silicon rather than single crystal silicon.

The most common method is two use two silicon wafers with oxide grown on the surface, than flip one of the wafers over and bond the two wafers together. The top wafer is then thinned down to the desired silicon film thickness. In the process, the oxide surface need to be cleaned to ensure free of particles to avoid the formation of voids. The thinning process also needs to be precise to avoid damaging the remaining silicon film.

As a result, SOI wafers are more expensive than conventional silicon wafers.

The series resistance of SOI MOSFETs is usually high, due to thin silicon film thickness. To reduce the resistance, it is common to increase the height of **source and drain relative to the channel region**, forming a raised source / drain or recessed channel structure.

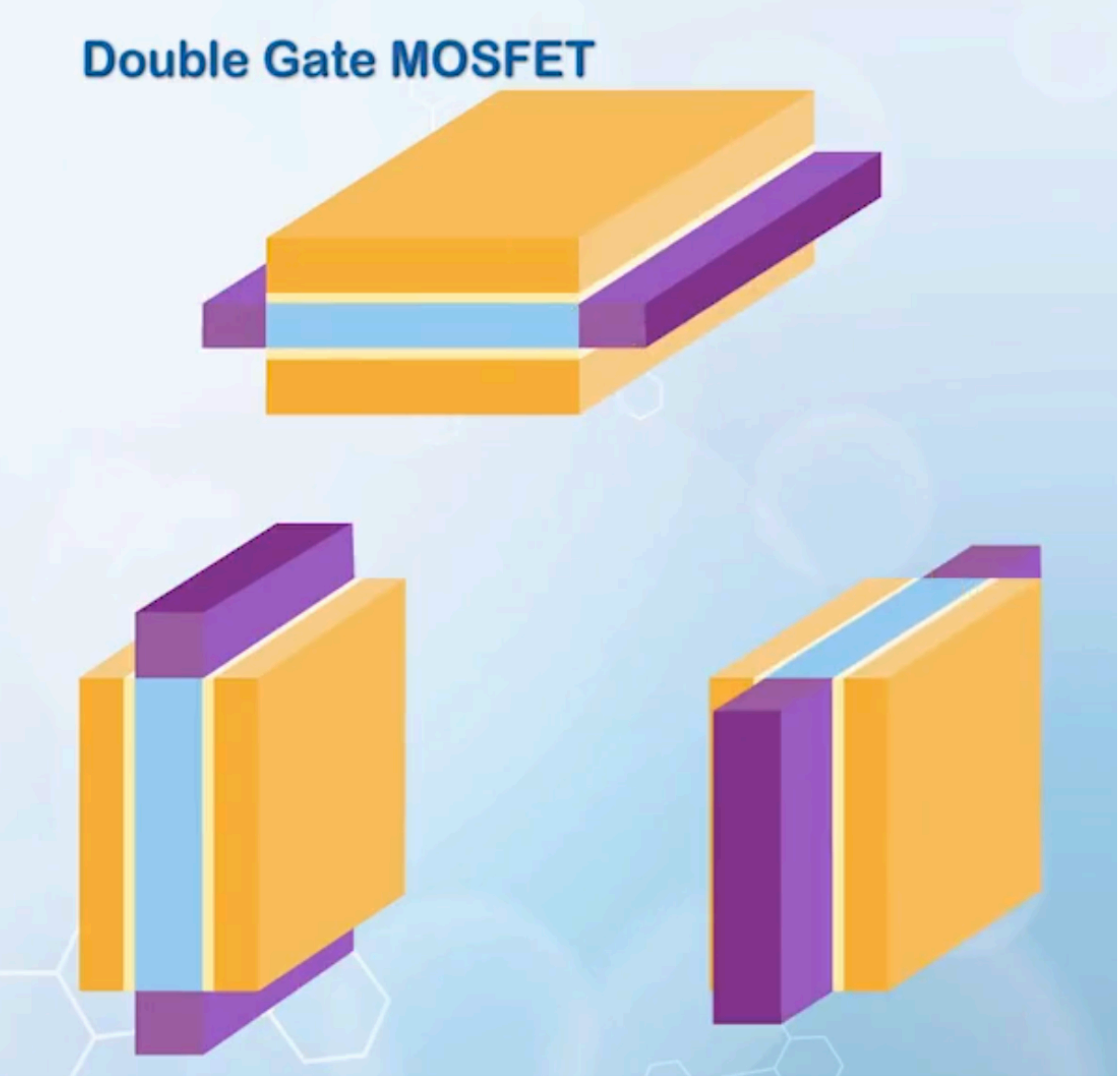


Double-gate MOSFETs and FinFETs

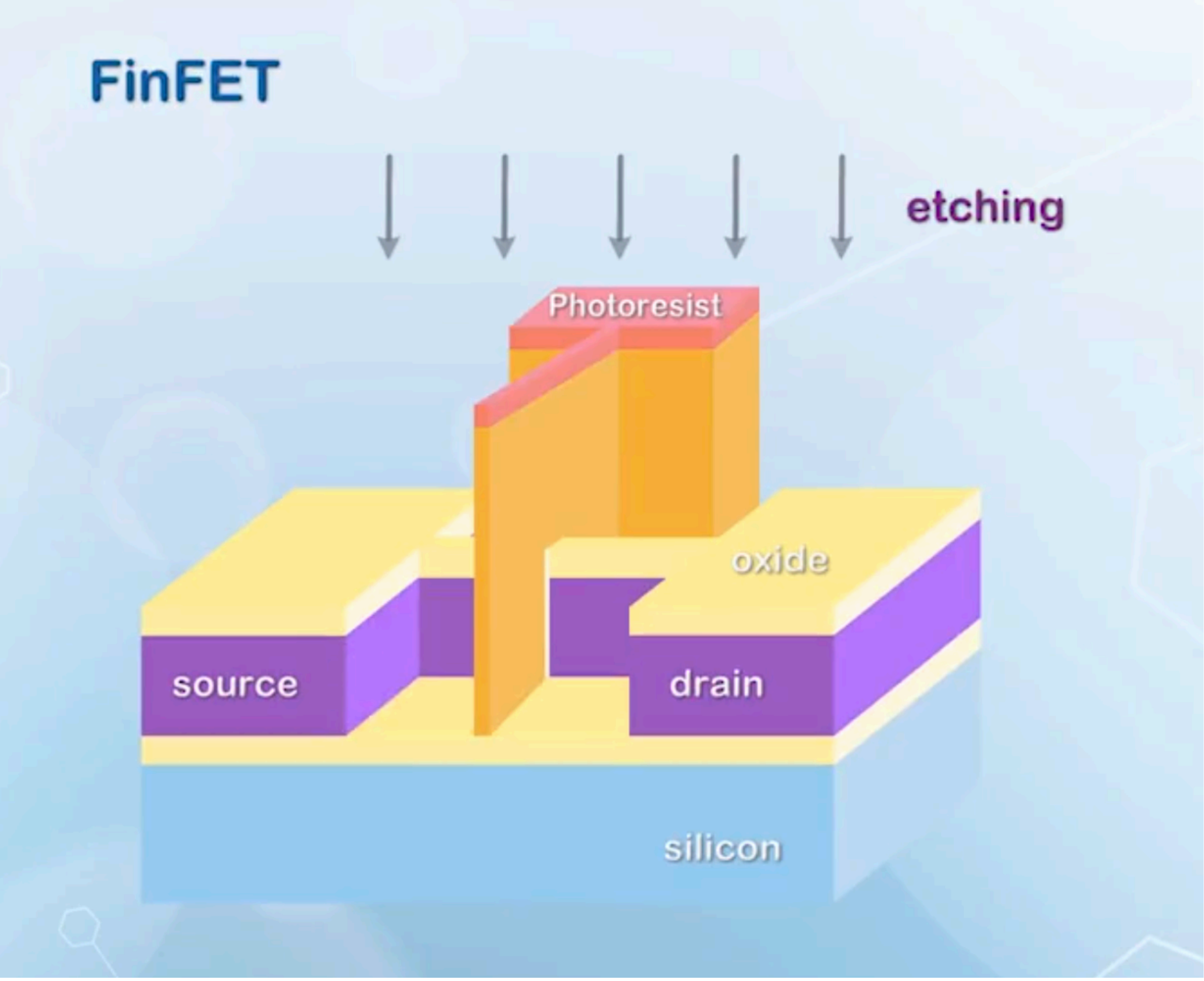
By adding another gate at the bottom of the silicon channel, we can increase the gate control. This forms a **double-gate MOSFET** structure.

Double-gate MOSFETs have three possible orientations:

- Horizontal channel, with gates on the top and bottom
    - First proposed structure
    - Difficult to connect and align the top and bottom gates precisely
- Vertical channel, with source and drain at the bottom and top, while gates are on the two sides
    - Also difficult to fabricate
    - Is not compatible with existing routing and placement methodology in integrated circuits
- Horizontal channel, with gates on the two sides
    - The dominant approach to implement today



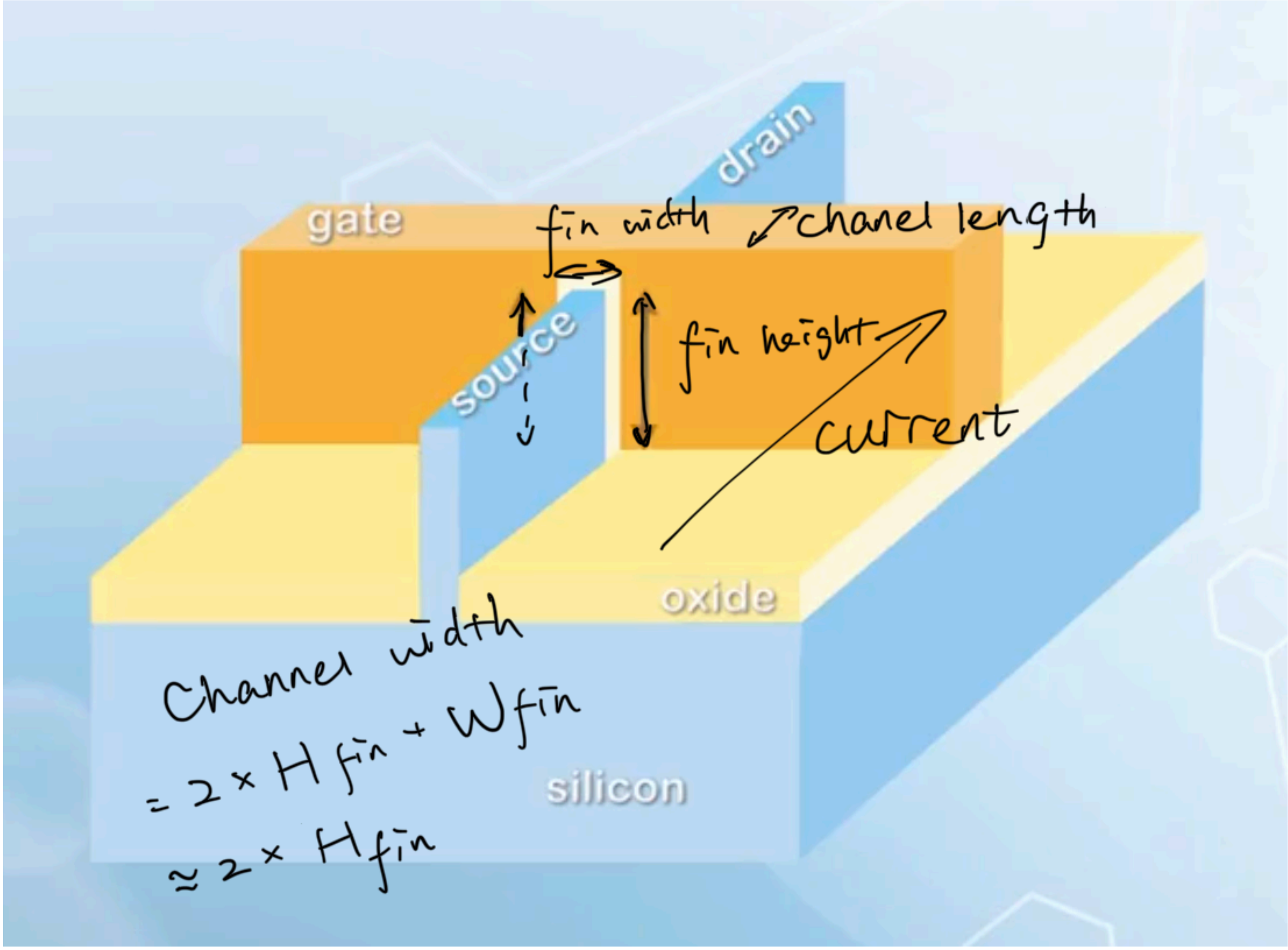
Once of the implementation of the structure is the **FinFET** structure.



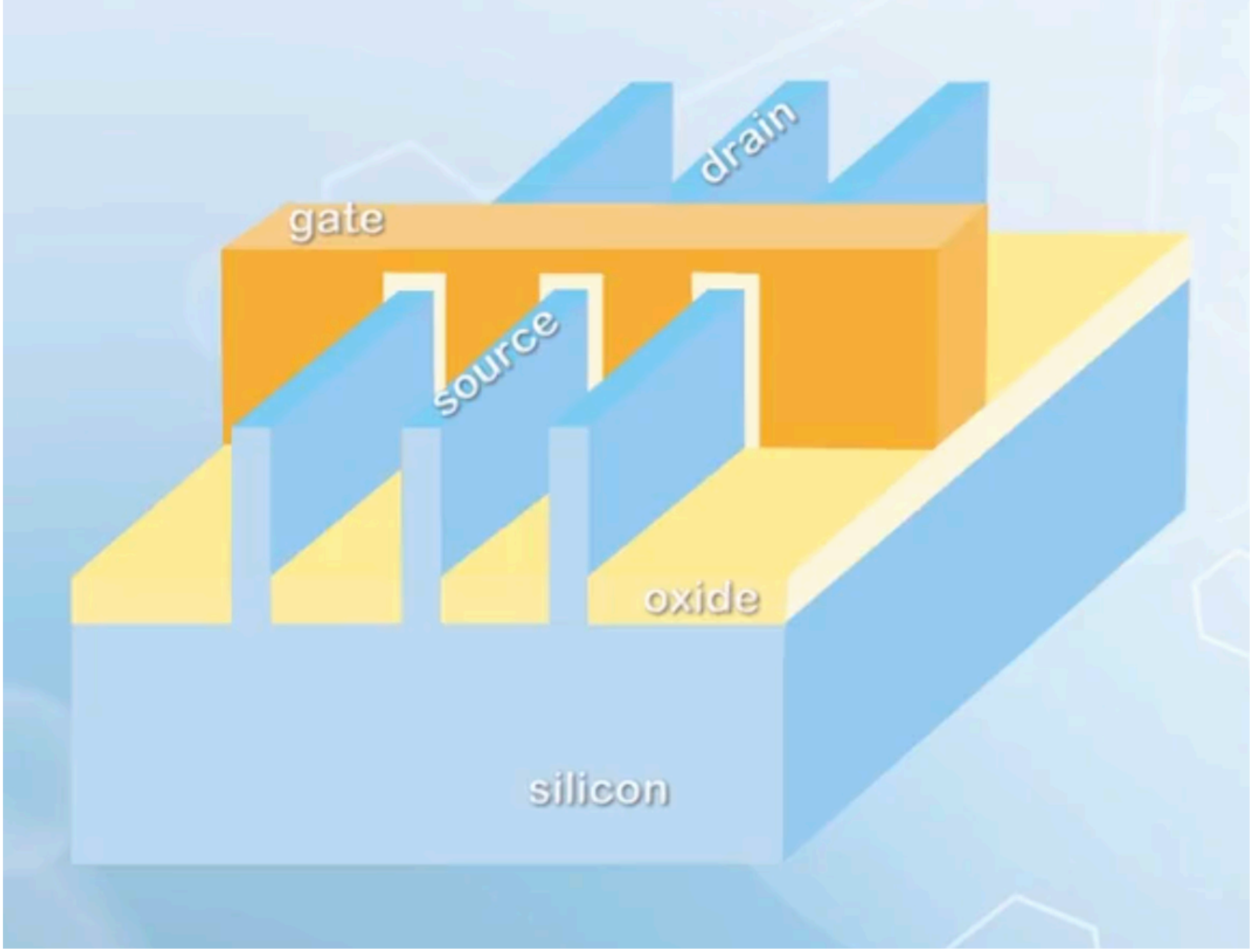
The two gates can be formed together in a single etching process, and the structure is also compatible with existing planar integrated circuit design methodology.

To avoid the expensive SOI technology, the channel region is of a FinFET is formed from the bulk silicon wafer etched down from the surface, with the channel region protected from etching. The resulting fin is connected to the substrate, which is different from the initially proposed structure, with the fin sitting on an insulator.

Channel forms on the two sides of the fin, and the channel width is two times the height of the fin. This cannot be changed by photolithography, and all FinFETs have the same channel width.



To scale the current of the FinFET using different channel widths, a number of channels can be placed side by side in parallel. So the FinFET causes a digitalization of the the channel width with a increment defined by two times the fin height.



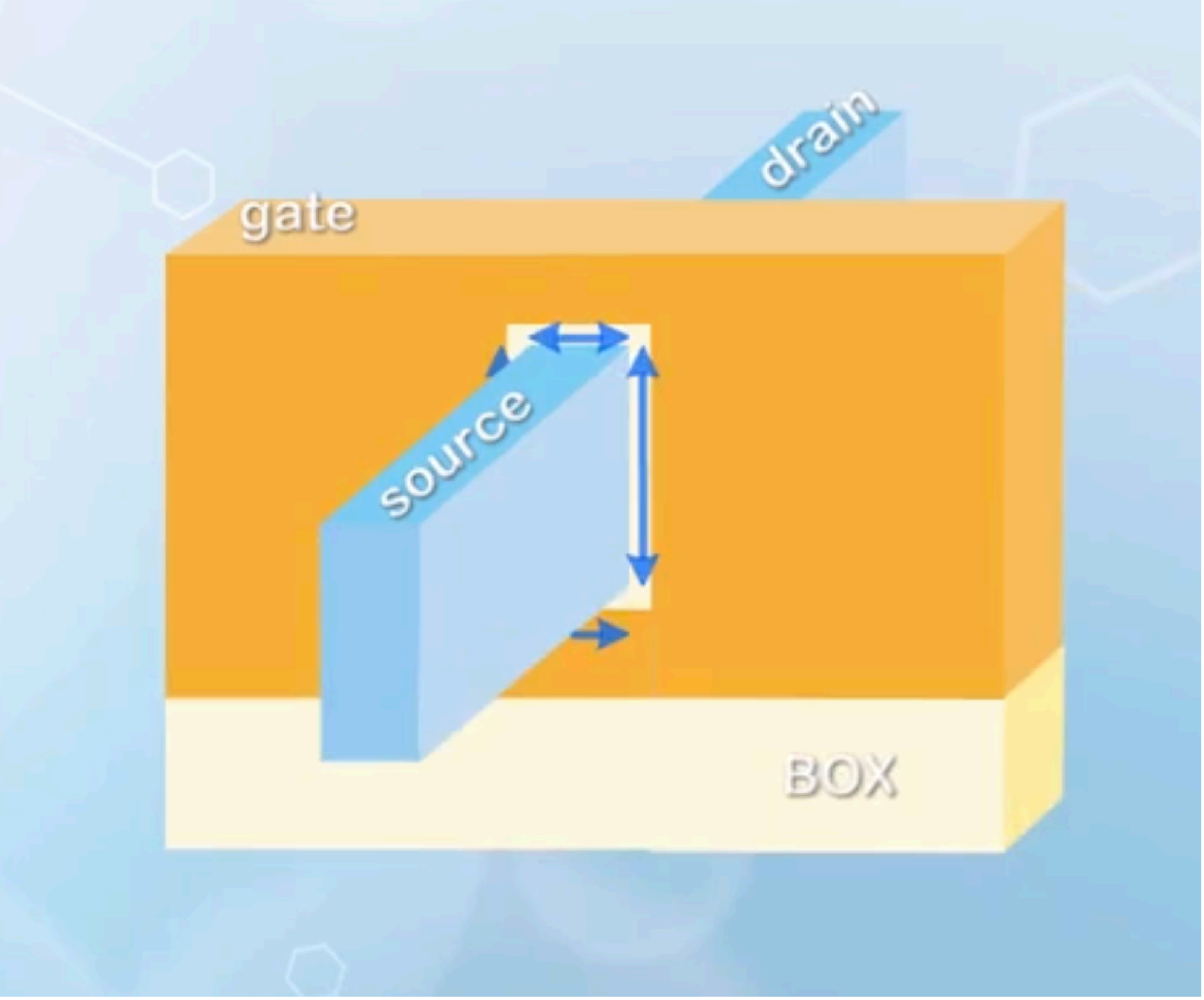
FinFET have two times the current density of a single gate MOSFET, and two times the gate capacitance. It is equivalent to two single gate MOSFETs with half the channel thickness folded together, but it consumes a smaller footprint than a single gate MOSFET with two times the channel width.

Multi-gate MOSFETs

The development of FinFETs marks the use of 3D FET structures to replace the planar MOSFET structures. With channel region standing out from the substrate surface, it is possible to conduct current on more than one surface. In this case, more than two gates can be formed, and it comes the era of multi-gate MOSFETs.

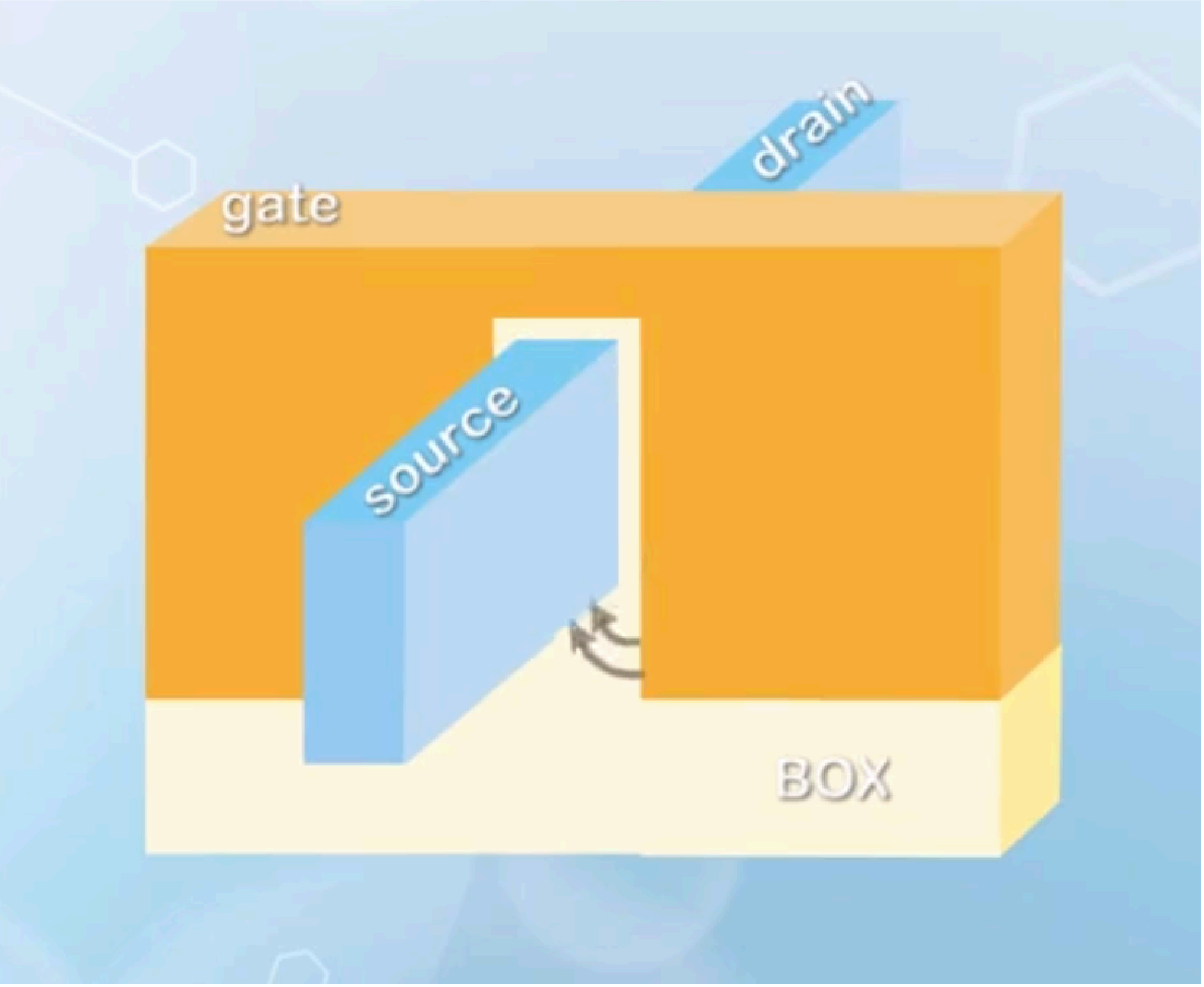
The top region of FinFETs are too narrow to be used, thus they are considered double-gate MOSFETs. With a wider top surface used for conduction, a **tri-gate MOSFET** can be formed. It is probably the only production ready technology beyond the double-gate FinFETs.

Ultimately, the strongest gate control can be achieved by surrounding a rectangular channel with gates on all four sides, forming a **gate-all-around (GAA) MOSFET** structure.

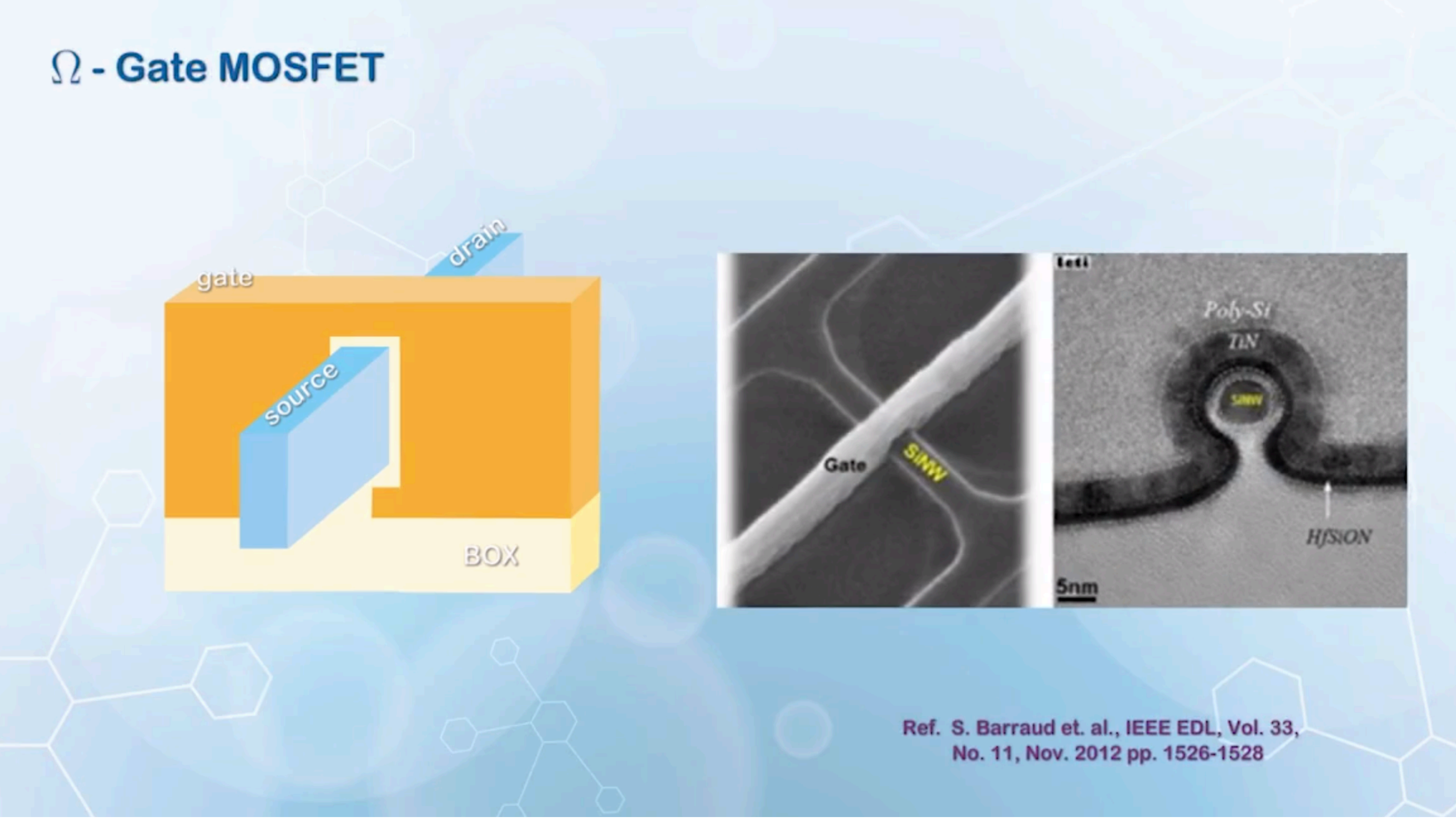


This structure is difficult to fabricate, as the bottom gate is covered by the channel, and cannot be etched together with the other three gates.

As a compromise between process complexity and better gate control, some structures extend the side wall gate below the supporting surface. The extended gate geometry provides capacitive coupling to the bottom surface for better gate control through the fringing electric field. This structure is called **π gate MOSFET**.



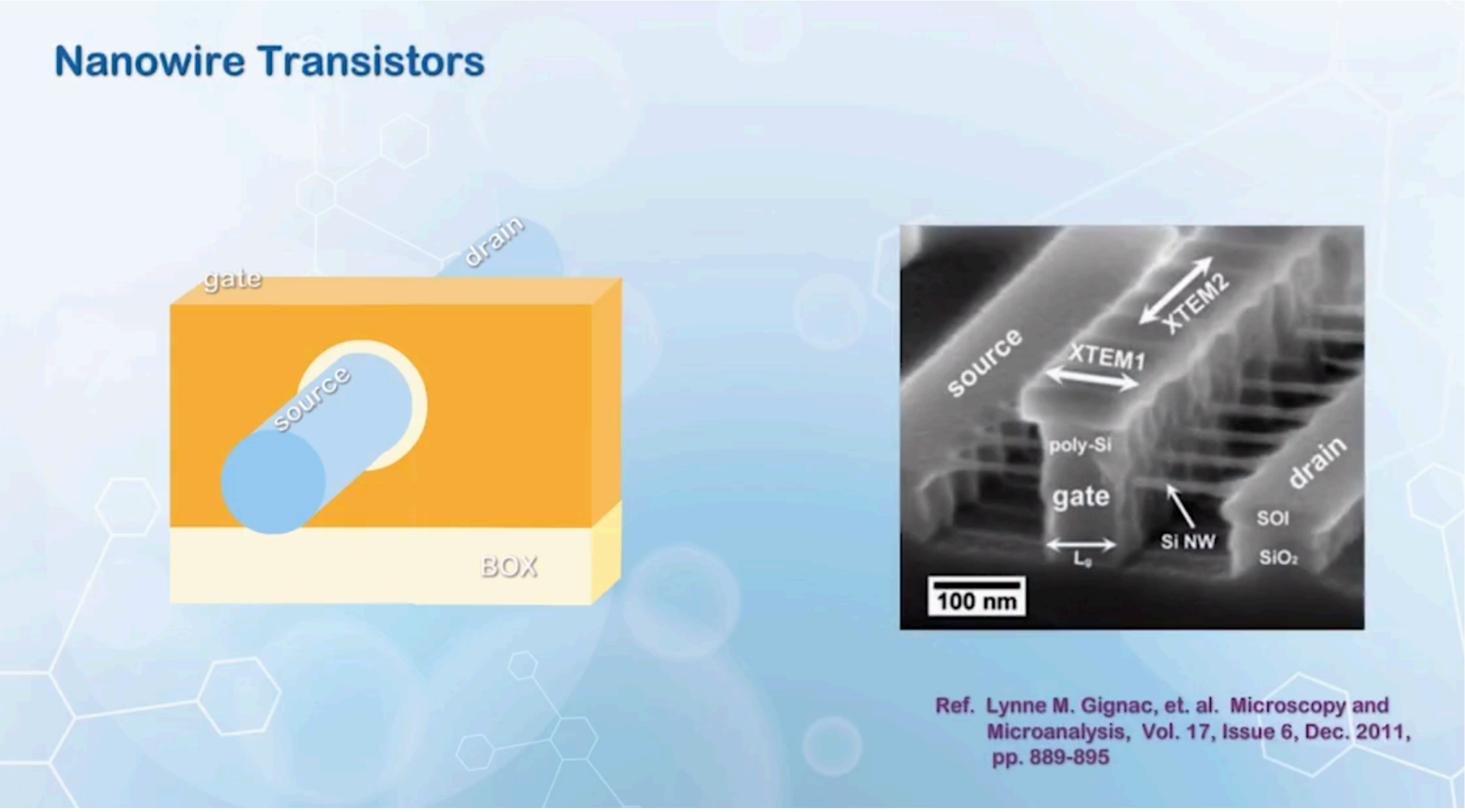
Some over etch can also be used to extend a little bit of the side wall gate to the bottom surface to increase the gate coupling, forming a **Ω gate MOSFET** structure.



Ω - Gate MOSFET

Ref. S. Barraud et. al., IEEE EDL, Vol. 33, No. 11, Nov. 2012 pp. 1526-1528

The problems with multi-gate structures formed with a rectangular channel cross section is the corner regions. The sharp corners cause high electric field concentration, leading to reliability issues. More advanced structures use a circular channel, with the gate wrapping around the channel. The diameter of the circular channel is usually very small, to avoid the center being too far away from the gate, which may lead to high off state leakage current. Such MOSFETs are called **nanowire transistors**.



Nanowire Transistors

Ref. Lynne M. Gignac, et. al. Microscopy and Microanalysis, Vol. 17, Issue 6, Dec. 2011, pp. 889-895

To calculate the $I-V$ relationship of the nanowire transistor, some parameters need to be modified:

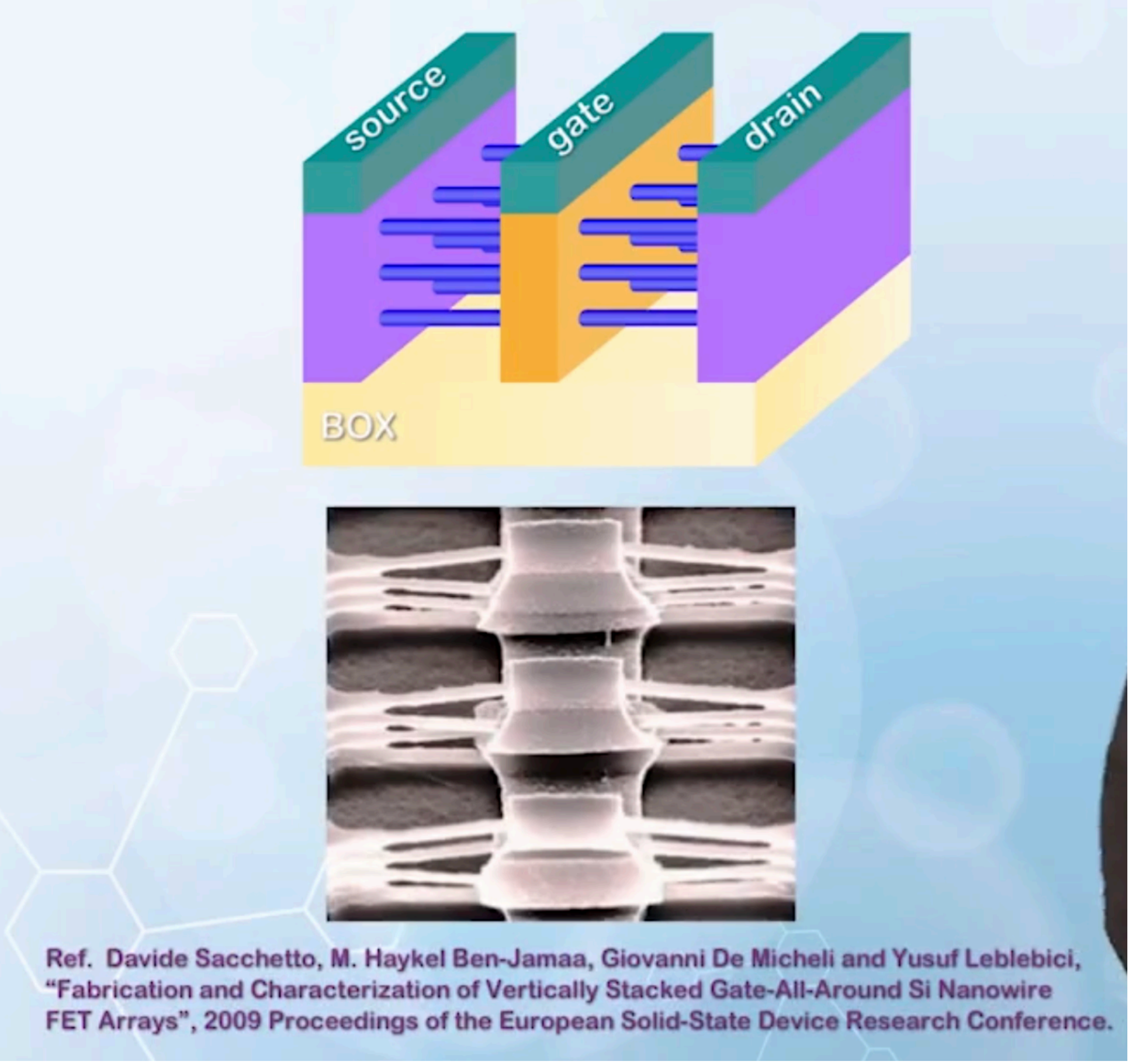
- The width is given by the circumference of the nanowire cross section

$$W_{\mathrm{NW}} = 2\pi R$$

  where R is the radius of the nanowire
- The effective oxide thickness should be used to calculate gate capacitance, due to the non-planar gate geometry. It is given by

$$t_{\mathrm{di}_{\mathrm{eff}}} = R \ln\left(1 + \frac{t_{\mathrm{di}}}{R}\right)$$

The formation of gate-all-around nanowire transistors usually requires the use of a suspended wire, so that the gate can go underneath the channel. With suspended wires, it is possible to stack a few nanowires together to form a 3D array of nanowire transistors, allowing high drive current over a small footprint. But they are still in exploration stage, and very far away from production yet.



Ref. Davide Sacchetto, M. Haykel Ben-Jamaa, Giovanni De Micheli and Yusuf Leblebici, "Fabrication and Characterization of Vertically Stacked Gate-All-Around Si Nanowire FET Arrays", 2009 Proceedings of the European Solid-State Device Research Conference.

Tunneling MOSFETs

To reduce the leakage current of MOSFETs, engineers have introduced many changes to the conventional CMOS technology, including new material systems and new device structures. Another approach to suppress the short channel effects is to change the physics.

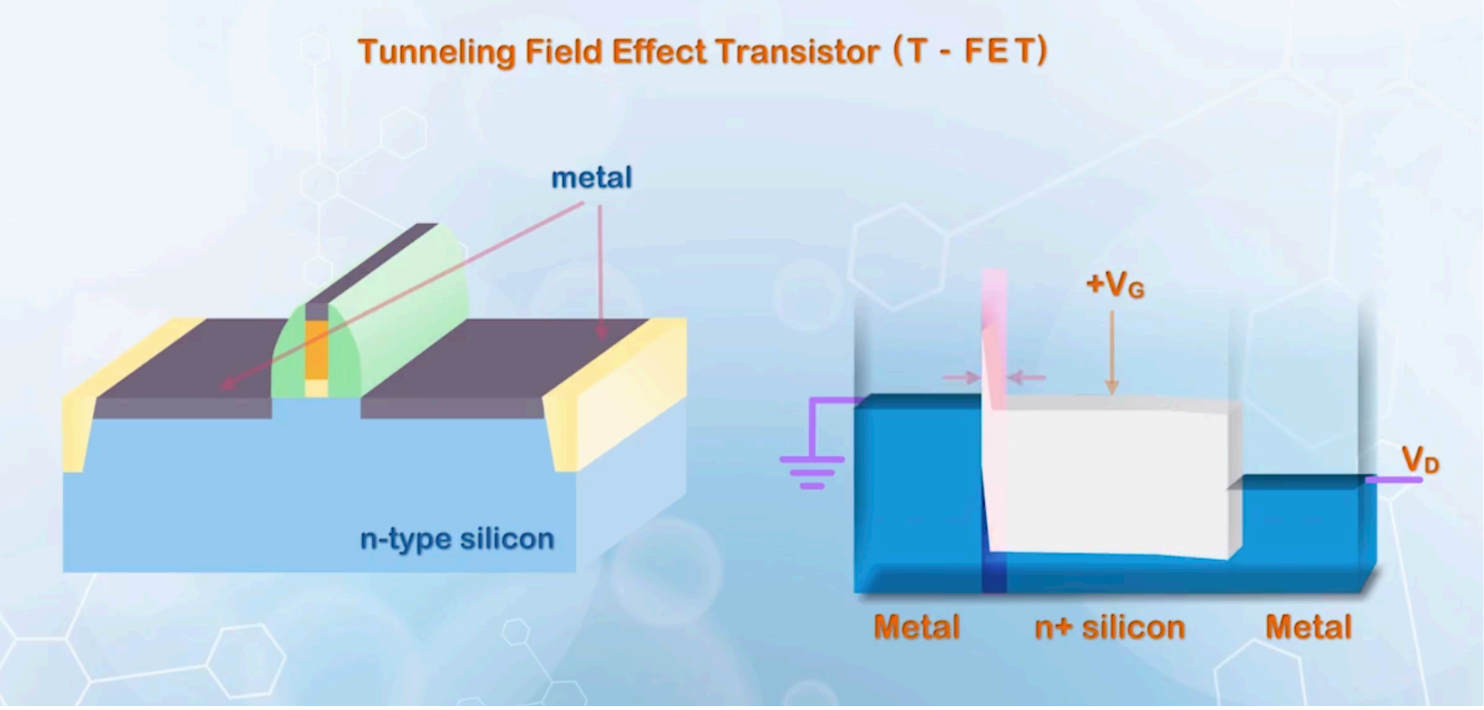
The on-off state current ratio is dictated by the subthreshold swing, which is $60\,\mathrm{mV/dec}$ under room temperature in ideal case. This physical limit is due to the Fermi-Dirac statistics when PN junction is used to inject carriers from high concentration region to low concentration region, or the

thermionic emission process. To overcome the limit, we may use **carrier tunneling** instead of thermionic emission to inject carriers.

The most important characteristics of a transistor is the exponential increase in current with applied gate voltage. In conventional MOSFETs, thermionic emission is the most common phenomenon used.

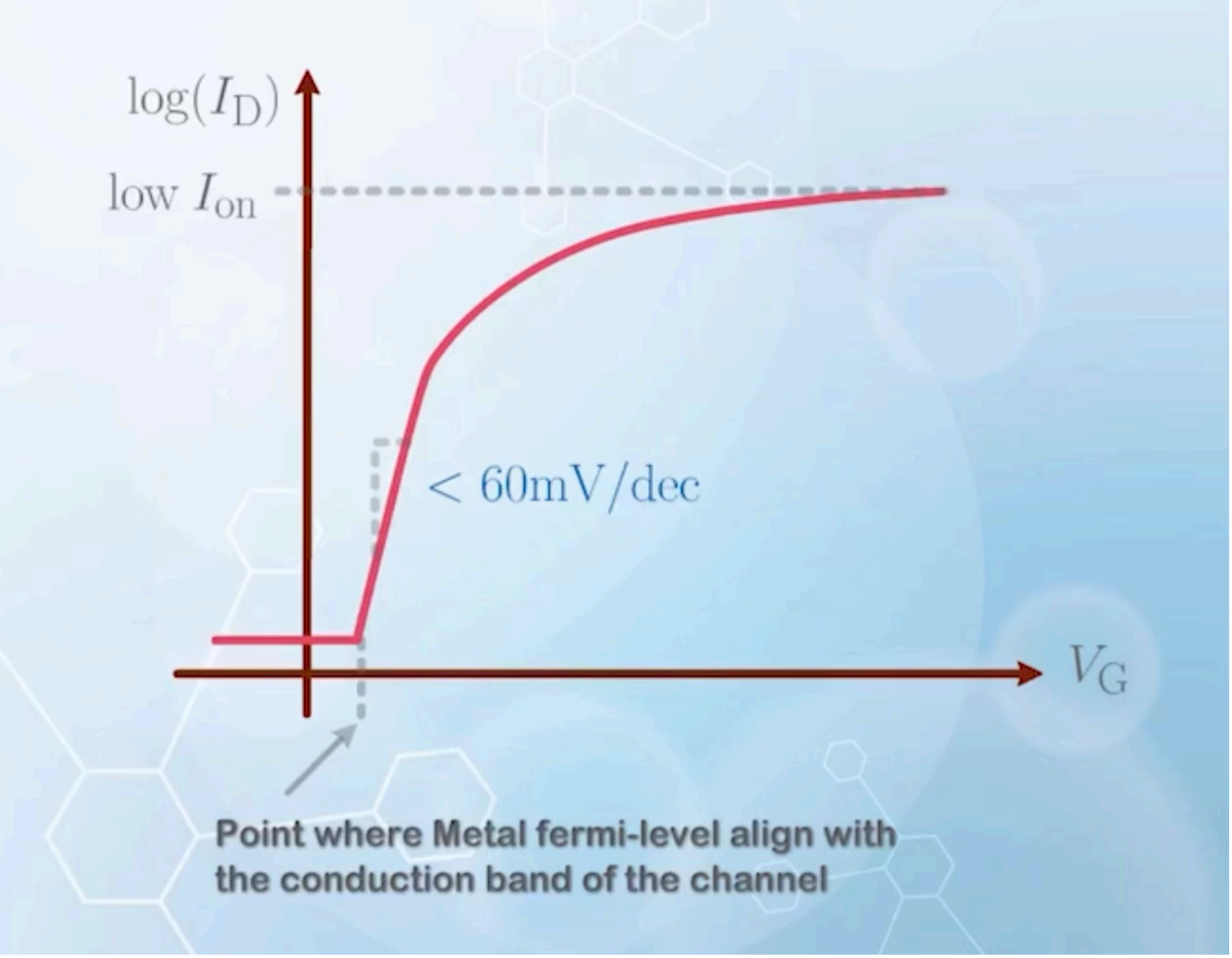
However, breakdown can also cause a very sharp turn-on, and the voltage to cause Zener breakdown of a metal-semiconductor contact can be adjusted to a low value. This is the principle of a **tunneling MOSFET (T-FET)**.

Early T-FETs use metal to form the source and drain with N type silicon substrate as the channel. When a positive gate voltage is applied, the barrier at the metal-semiconductor junction may become thin enough for electrons to tunnel through the barrier. The probability of electrons tunneling though the barrier can be modulated by the gate voltage.

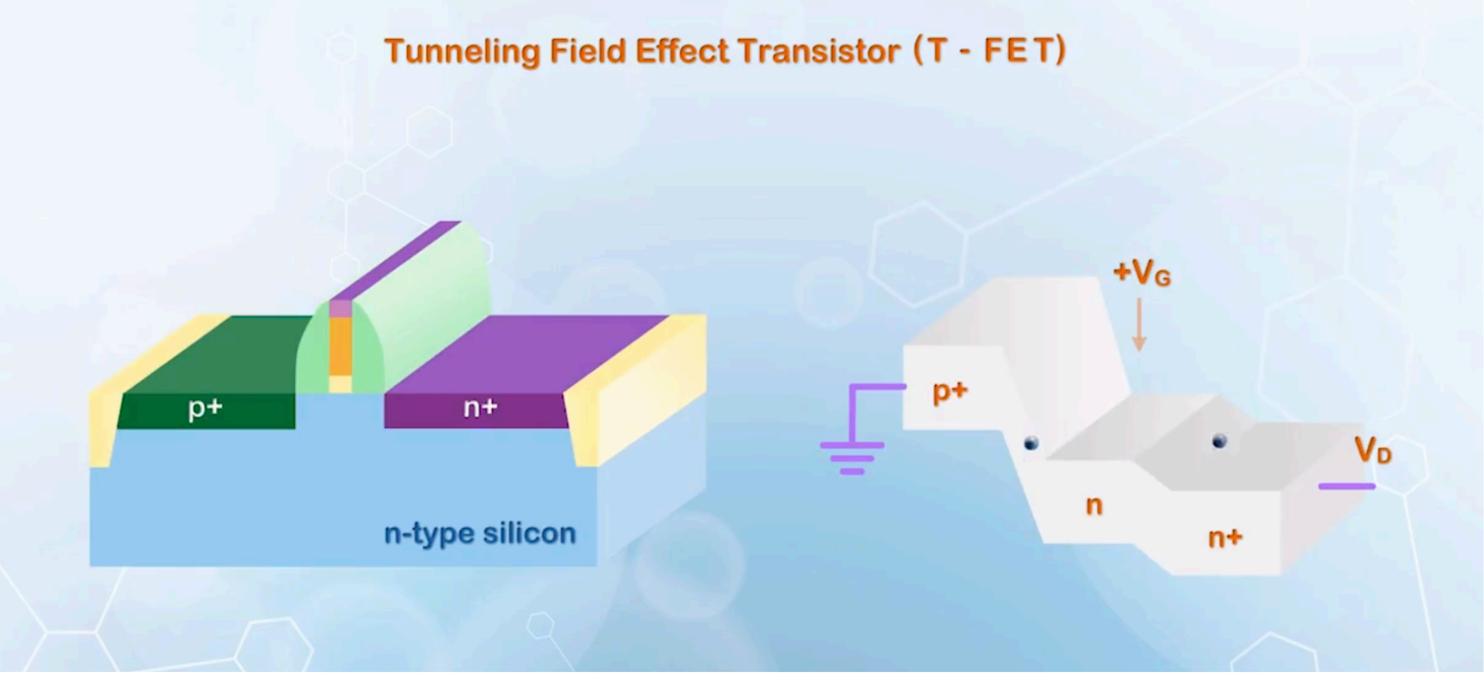


There is a specific point corresponding to the alignment of the metal Fermi level and the conduction band edge of the N silicon channel. Before this point, there is no state for electrons from the metal to tunnel to, unless given extra energy. The tunneling current is expected to be very small. Beyond this point, electrons will suddenly find available states to tunnel to, and a significant increase in current is expected. Measurements also show a very rapid turn-on, with subthreshold swing less than $60\,\mathrm{mV/dec}$ at room temperature. However, it only applies to a very small portion of the turn-on curve, and the turn-on rate drops quite rapidly outside the gate voltage range corresponding to the alignment point.

T-FETs has a barrier independent of the scaling process, thus are relatively easier to turn off, and do not have barrier lowering issues like DIBL. However, the problem shifts to the on state current, which is relatively low due to high contact resistance.



More recent implementations of T-FETs have switched from using metal source and drain to using semiconductor material. By properly biasing the gate, source and drain terminals, electrons from the valance band of the source can tunnel through to the conduction band of the channel, and collected by the drain.



By using semiconductor materials for both the source and body, the junction property can be more easily adjusted by doping.

To improve injection efficiency, some compound semiconductor material systems are used to form source and channel regions, and significant progress has been made to achieve a steep subthreshold slope with subthreshold swing smaller thant $60\,\mathrm{mV/dec}$ at room temperature over a wide range of gate voltages.

Junctionless Transistors

The high on-off current ratio is the main design goal of nano scale transistors. Scaling down helps the on state current to increase linearly, but the off state current also increases exponentially. A simple solution is to use a longer channel MOSFET with higher V_T, but it increases capacitance loading and footprint. The strategy is: **increase the effective channel length without changing physical geometry**, leading to the idea of **junctionless transistors**.

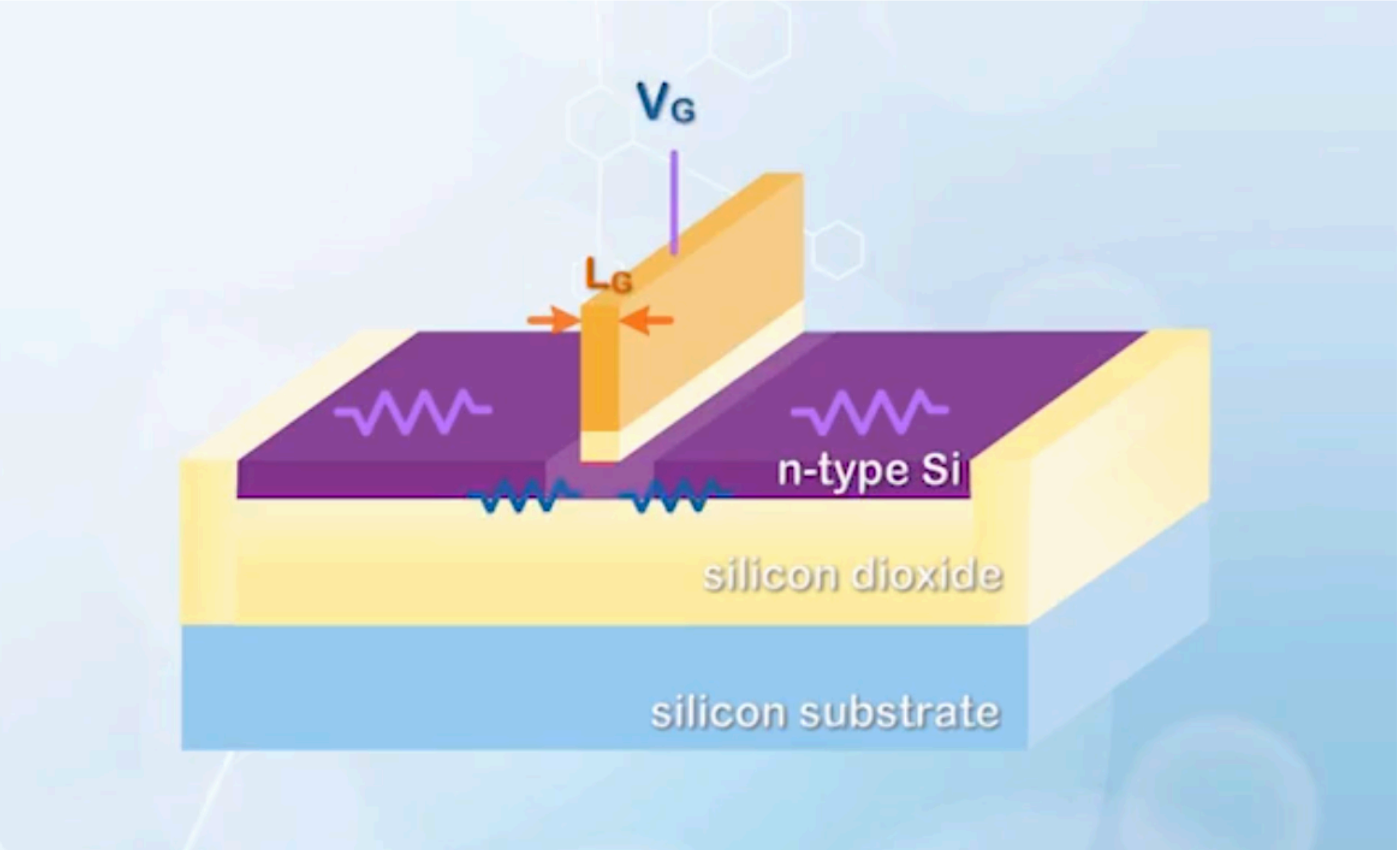
For a junctionless transistor, the doping concentration for the channel, source and drain regions are the same, without physical boundaries for the channel. The work function of the gate material is carefully chosen to completely deplete the channel when $V_\mathrm{G} = 0$. For example, if an N type semiconductor is used for the transistor, the gate material should have a large work function, so that its Fermi level is close to the valance band edge of the semiconductor channel. In this case, the gate material mimics a P+ gate, depleting the N type channel when put in contact.

Because the channel is physically boundary-less, the depletion region can extend outside the gate edges, into source and drain. The channel borrows some regions of the source and drain to effectively increase the channel length when the transistor is turned off.

When a positive voltage is applied to the gate, it is similar to forward biasing the channel PN junction, causing the depletion region to shrink. It restores the original carrier concentration in the channel, and conduction occurs. The positive voltage effectively eliminated the channel length.

The problem: this kind of device need extremely thin channel region, or a very small cross-sectional area. Otherwise, channel region far away from the gate will have a significant leakage current. With such small cross sectional area, the series resistance may be very high. Moreover, the highest doping concentration for the channel, source and drain is limited, as too high doping concentration will make it difficult to fully deplete the channel at $0\,\mathrm{V}$. This further increases the series resistance.

A compromise is to increase doping at two ends of the source and drain regions where contacts are made. But some part in the middle still needs to remain at lower doping concentration, despite not covered by the gate, otherwise the depletion region cannot expand beyond the gate edges. These less heavily doped regions still contribute to series resistance when the transistor is turned on.



Although the increased series resistance reduced its competitiveness, its relatively simple structure without any junction makes it a good vehicle for concept demonstrations.

2D Transistors and Carbon Nanotube FETs

When discussing SOI technology, advantages of ultra thin channel region to reduce short channel effects have been covered. Ultimately, we would like to have a single layer of atoms to form the channel region.

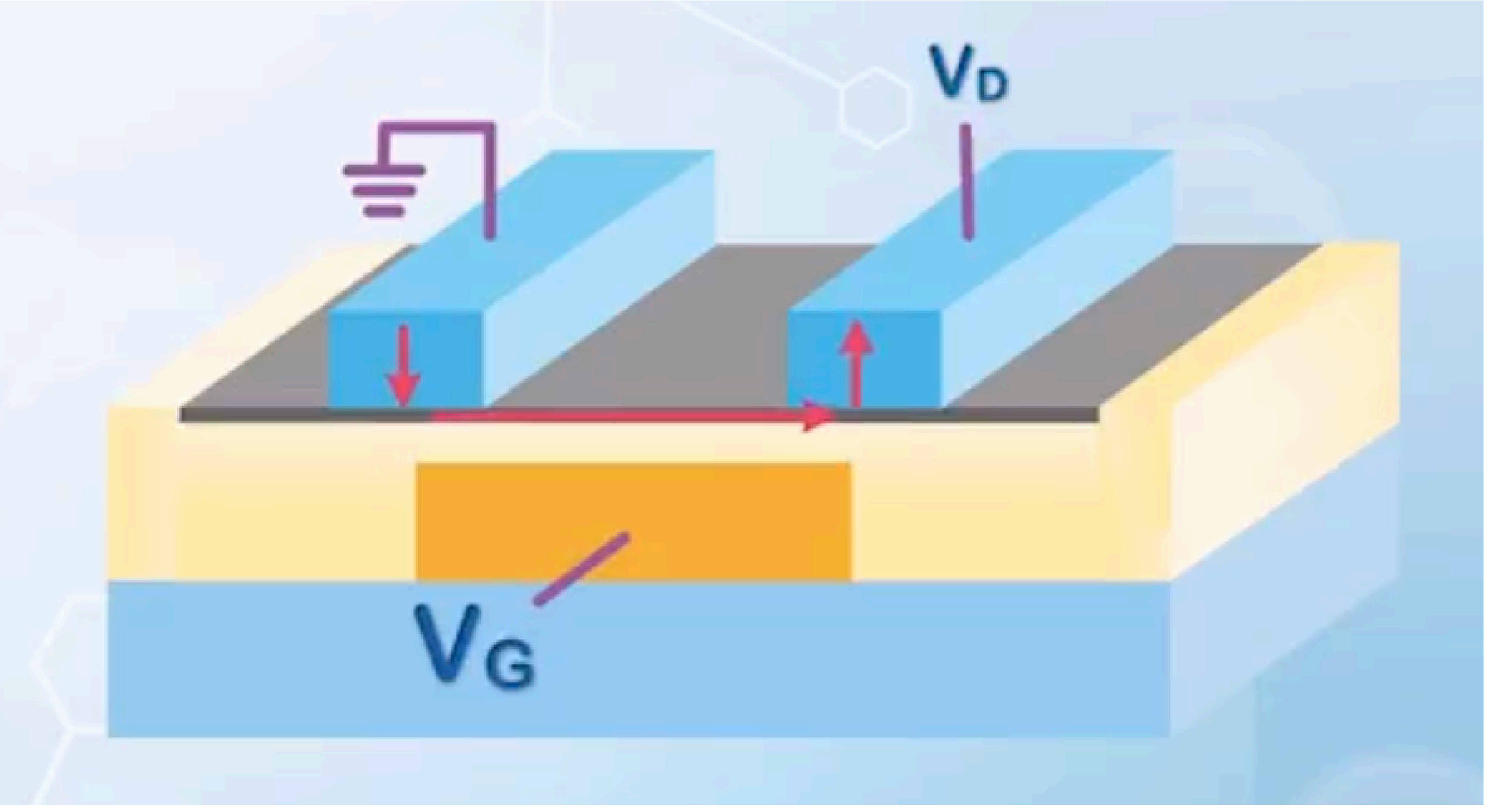
To make a single layer of atoms, we can use some 2D materials. Unlike silicon, a mature method to dope a 2D material to make it P or N type is not available, thus a junctionless transistor structure is usually used.

Among the 2D materials, graphene is most well understood. However, graphene behaves like a conductor, and is difficult to deplete with gate voltage. Therefore using graphene to fabricate 2D transistors is difficult, and it is more desirable to use a semiconductor material as the channel of the transistor. This includes:

- Molybdenum disulfide ($\mathbf{MoS_2}$) for N channel devices
- Tungsten diselenide ($\mathbf{WSe_2}$) for P channel devices

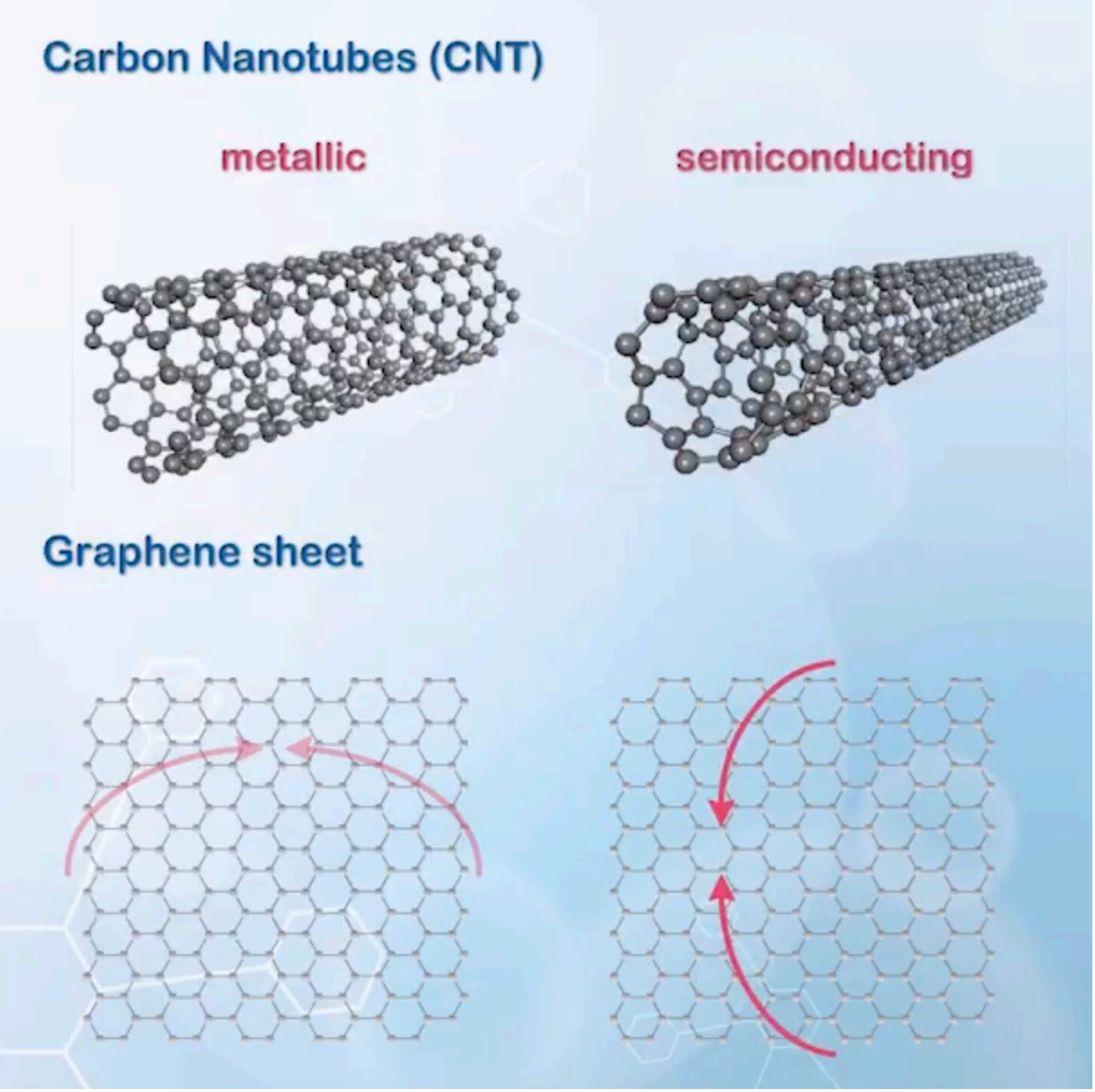
As the 2D materials cannot be doped, **very small separation** between the source, drain and gate is needed, to avoid the undoped high resistance regions. Most experimentally demonstrated 2D transistors use the **bottom gate structure** to provide an overlap between the gate and source / drain contacts, eliminating the highly resistive undoped regions.

When gate voltage is applied, carriers are attracted by the gate to move downward from the metal contact, conduct through the bottom interface, then move up to the electrode.

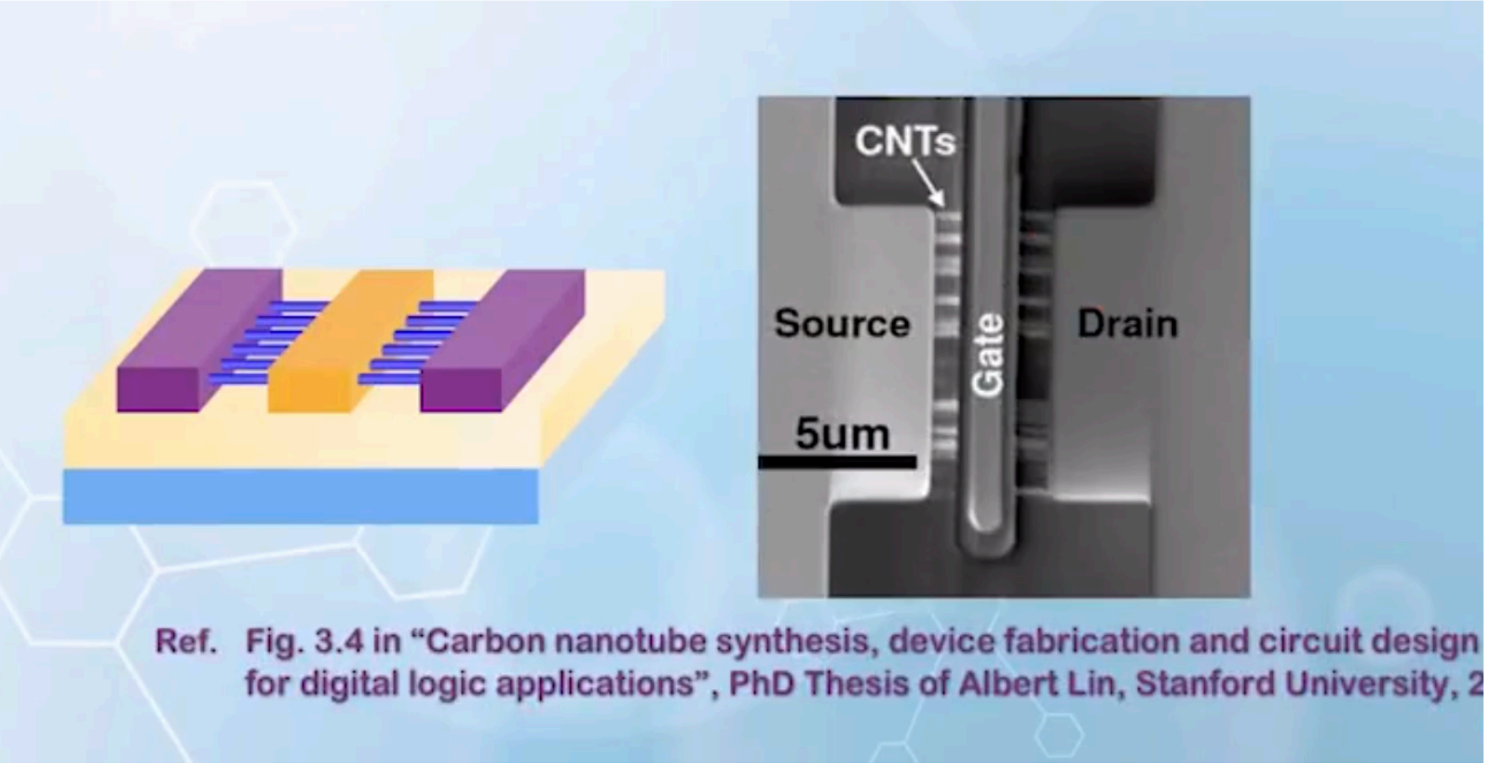


The contact between the metal electrode and the 2D material is another important issue, as contacts usually have the highest resistance in a heterogenous system. As the 2D material behaves like a semiconductor, the contact is expected to be a metal-semiconductor contact. Based on the work function of the metal, the mechanism for carrier injection can be either thermionic emission for linear ohmic contacts, or tunneling for Schottky contacts, with the latter expected to dominate.

Rolling up a 2D material forms a 1D tube. For example, rolling up a graphene sheet gives a **carbon nanotube (CNT)**. A graphene sheet can be rolled up in two different orientations, and depending on the rolling orientation, the CNT can be either metallic or semiconducting.



The most natural structure for a CNT transistor is the gate-all-around nanowire transistor structure. However, forming a gate-all-around structure with CNT is very difficult. Most carbon nanotube FETs demonstrated place CNTs side by side, which is similar to a planar transistor.



Ref. Fig. 3.4 in “Carbon nanotube synthesis, device fabrication and circuit design for digital logic applications”, PhD Thesis of Albert Lin, Stanford University, 2

Another problem with carbon nanotubes is that synthesized CNTs usually contain a mixture of metallic and semiconducting CNTs, and they are hard to separate. If the channel is formed with a number of CNTs bundled together, it is likely that some metallic CNTs are included, shorting the source and drain. The remove the metallic CNTs, high current is usually passed through the transistor to burn out the metallic CNTs before the transistor is used. In the case of a single CNT transistor, having a metallic CNT can declare the transistor non-functional. Some redundancy or luck is required to get a working circuit made of CNT MOSFETs.