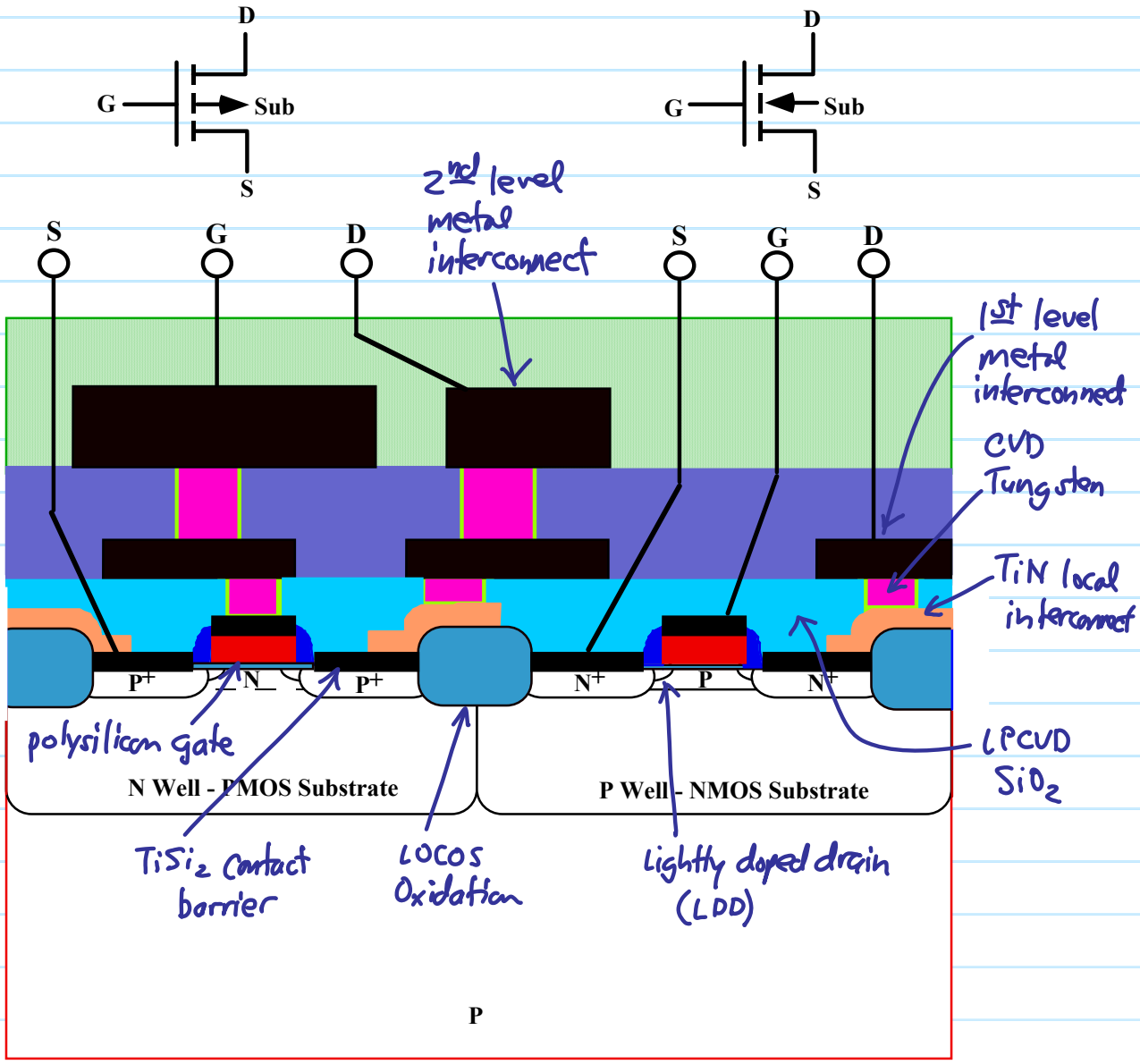
Typical mid-2000's CMOS Process (good down to ~0.25μm)

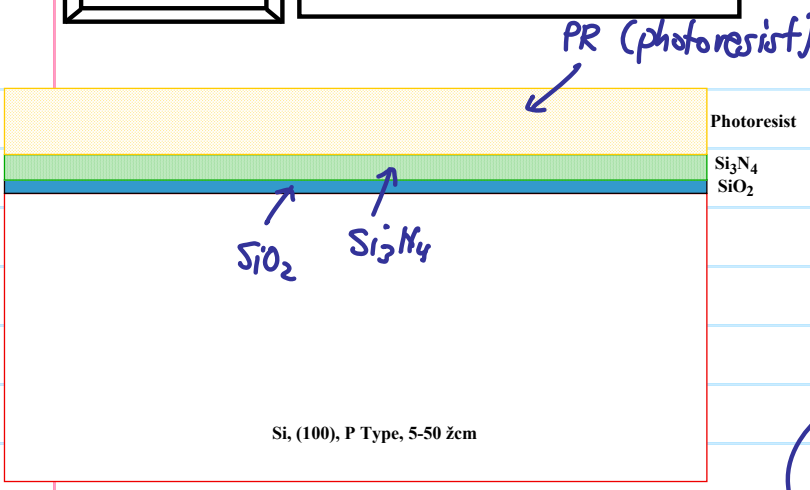


Comments:

① Above cross-section uses LOCOS oxidation for device isolation → this limits it to ~0.25μm processes ⇒ need trench isolation for smaller channel lengths

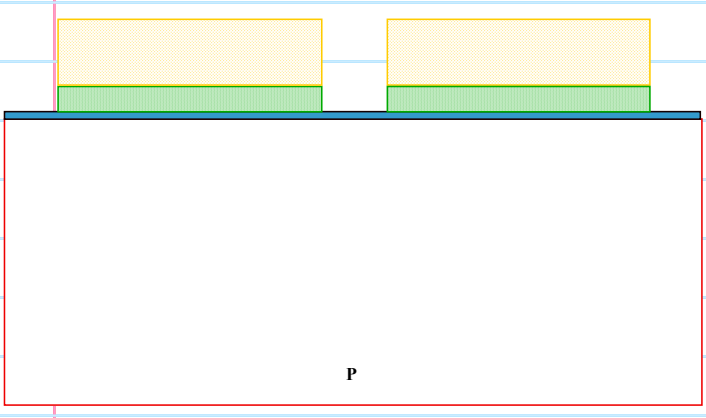
② $TiSi_2$ contact barriers over S/D junctions prevent metal spiking into the junctions.

③ Lightly doped drain (LDD) junctions reduce E-fields that if too large generate "hot e⁻'s" that can degrade device performance → a reliability problem.
↳ also help to reduce punchthrough

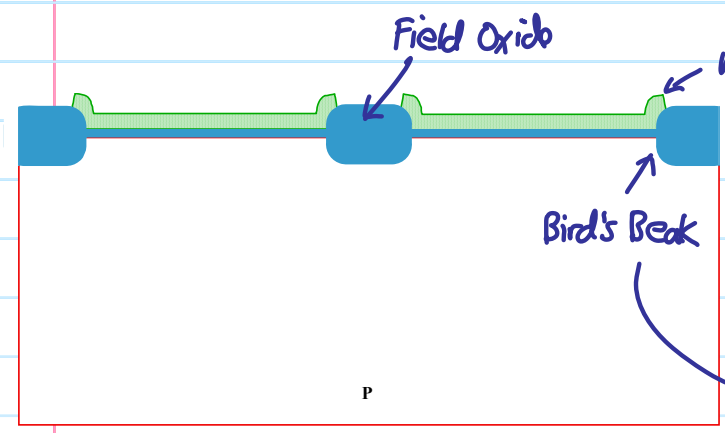


- Substrate selection: moderately high resistivity, (100) orientation, P type.
- Wafer cleaning, thermal oxidation (≈ 40 nm), nitride LPCVD deposition (≈ 80 nm), photoresist spinning and baking (≈ 0.5 - 1.0 μm).

sized to minimize stress on Si substrate

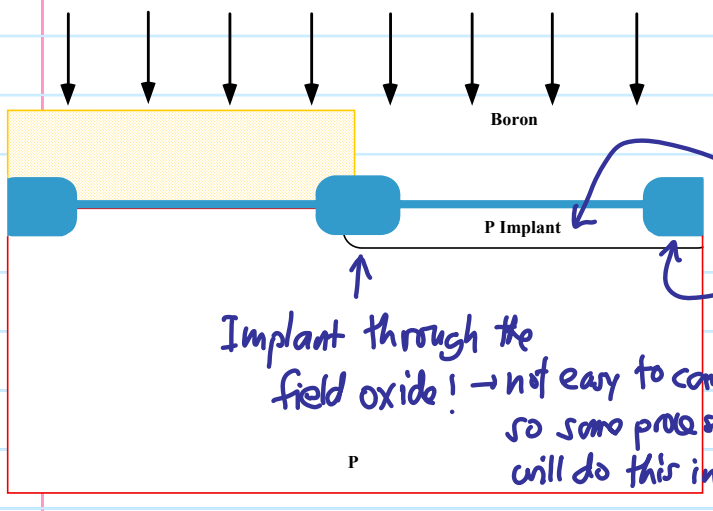


- Mask #1 patterns the active areas. The nitride is dry etched.



- Field oxide is grown using a LOCOS process. Typically 90 min @ 1000°C in H_2O grows ≈ 0.5 μm.

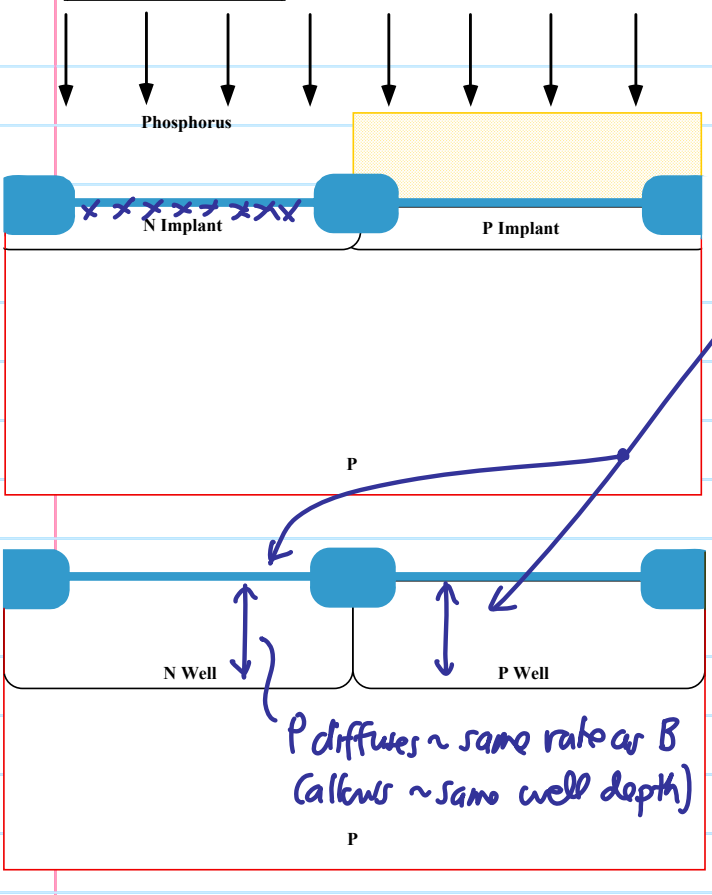
encroaches on the device causing problems that we will discuss (shallow trench isolation a better approach!)



This single implant realizes both the well & the channel-stop layer

- Mask #2 blocks a B+ implant to form the wells for the NMOS devices. Typically 10^{13} cm^{-2} @ 150-200 KeV.

↳ $5 \times 10^{16} - 10^{17}$ cm^{-3}



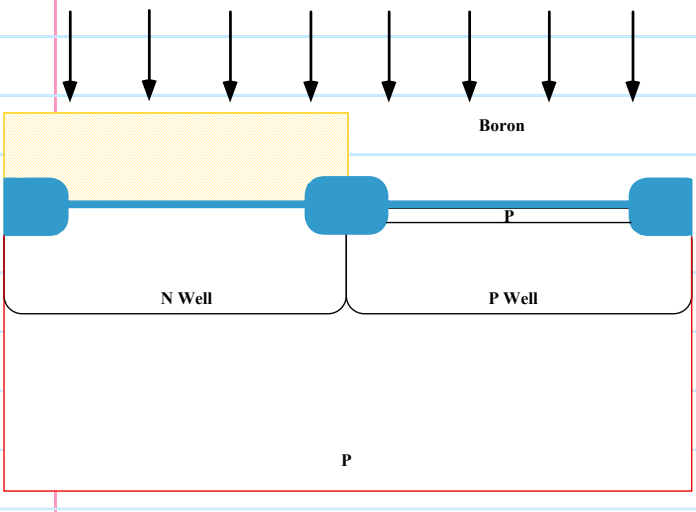
(antimony)
As or Sb can also be used, but they are heavier, so require higher III energy, plus P diffuses @ same rate as B

- Mask #3 blocks a P+ implant to form the wells for the PMOS devices. Typically 10^{13} cm^{-2} @ 300+ KeV.

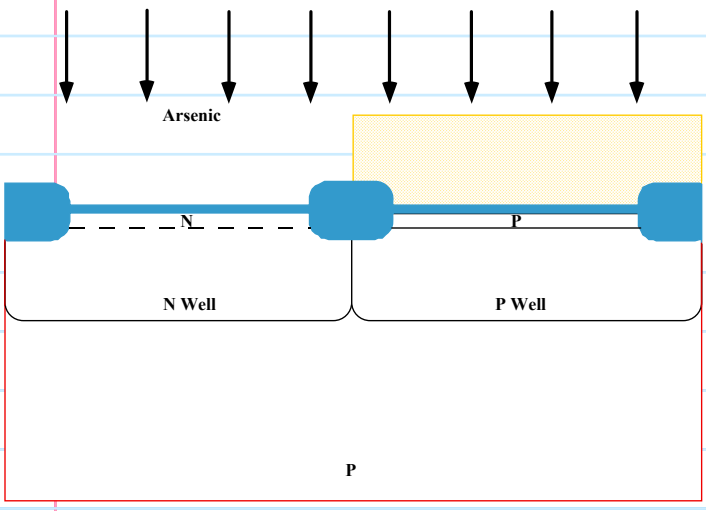
Very high energy will damage the Si surface. Must anneal to repair the damage.

- A high temperature drive-in produces the "final" well depths and repairs implant damage. Typically 4-6 hours @ 1000°C - 1100°C or equivalent Dt.

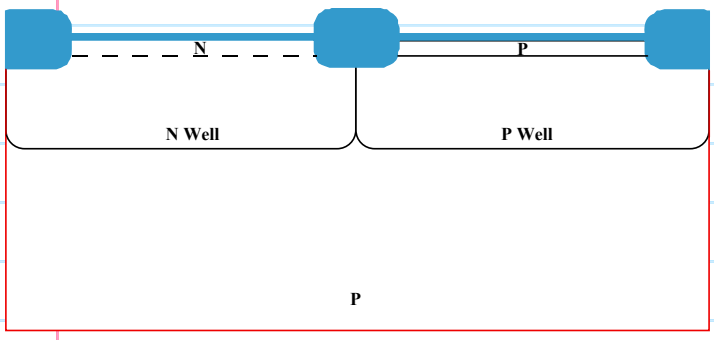
P diffuses ~ same rate as B
(allows ~ same well depth)



- Mask #4 is used to mask the PMOS devices. A V_{TH} adjust implant is done on the NMOS devices, typically a 1-5 x 10^{12} cm^{-2} B+ implant @ 50 - 75 KeV.

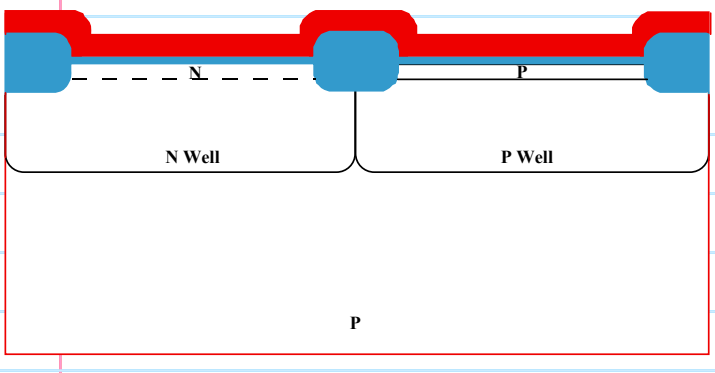


- Mask #5 is used to mask the NMOS devices. A V_{TH} adjust implant is done on the PMOS devices, typically 1-5 x 10^{12} cm^{-2} As+ implant @ 75 - 100 KeV.

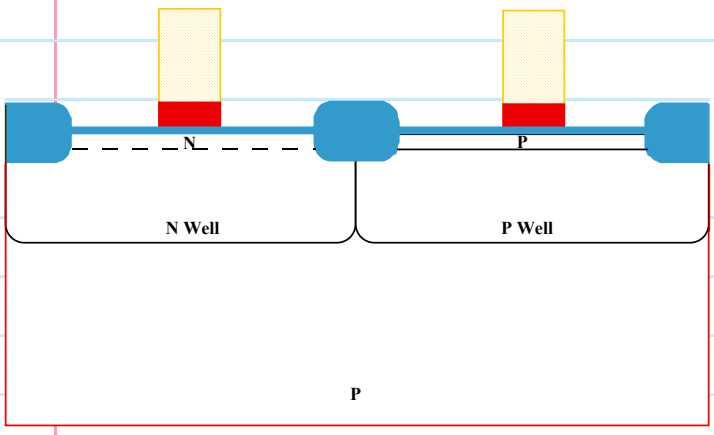


- The thin oxide over the active regions is stripped and a new gate oxide grown, typically 3 - 5 nm, which could be grown in 0.5 - 1 hrs @ 800 °C in O2.

This is done because the previous oxide is damaged after numerous implants.

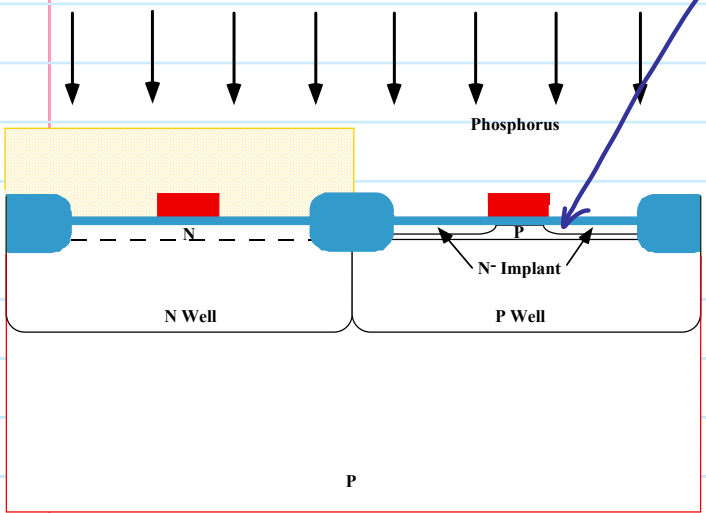


- Polysilicon is deposited by LPCVD (≈ 0.5 μm). An unmasked P+ or As+ implant dopes the poly (typically 5 x 10^{15} cm^{-2}).

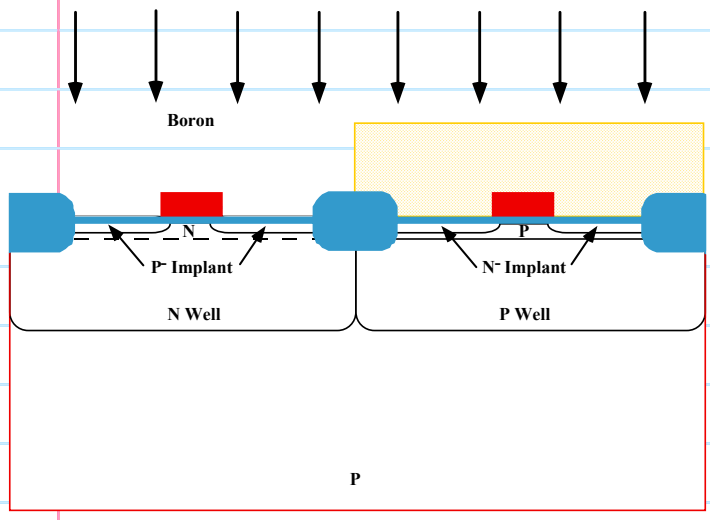


- Mask #6 is used to protect the MOS gates. The poly is plasma etched using an anisotropic etch.

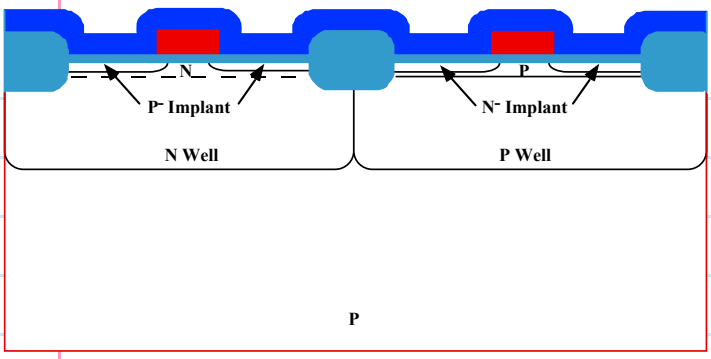
Want the S/D regions near the gate to be
① shallow (i.e., x_j = small) to minimize punchthrough
② lightly-doped to minimize E-fields & enhance device reliability



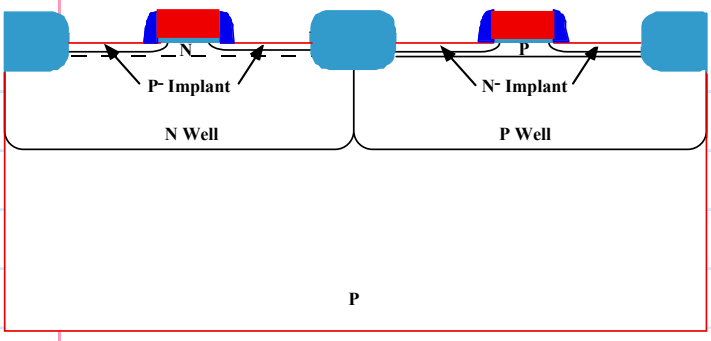
- Mask #7 protects the PMOS devices. A P+ implant forms the LDD regions in the NMOS devices (typically 5 x 10^{13} cm^{-2} @ 50 KeV).



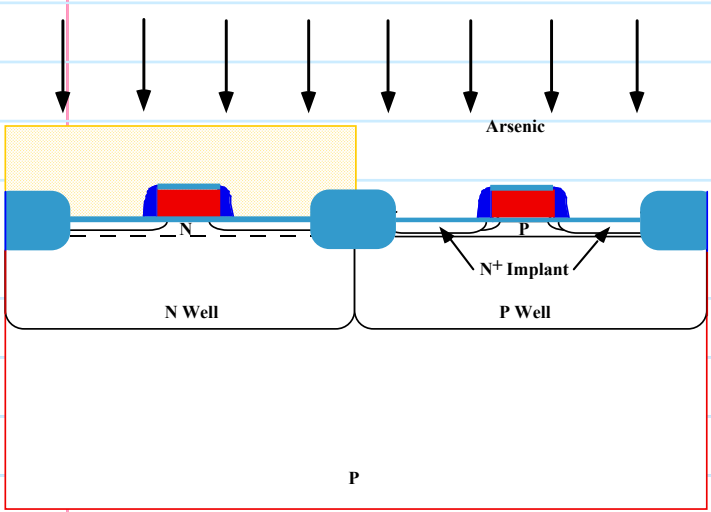
- Mask #8 protects the NMOS devices. A B+ implant forms the LDD regions in the PMOS devices (typically 5 x 10^{13} cm^{-2} @ 50 KeV).



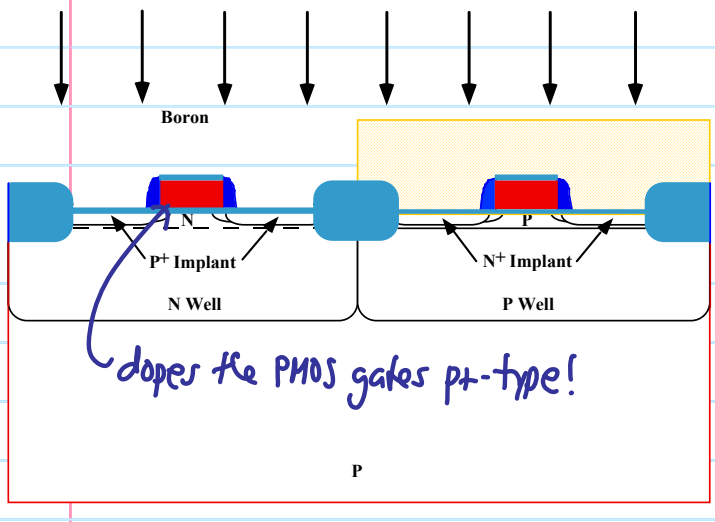
- Conformal layer of SiO2 is deposited (typically 0.5 μm).



- Anisotropic etching leaves "sidewall spacers" along the edges of the poly gates.

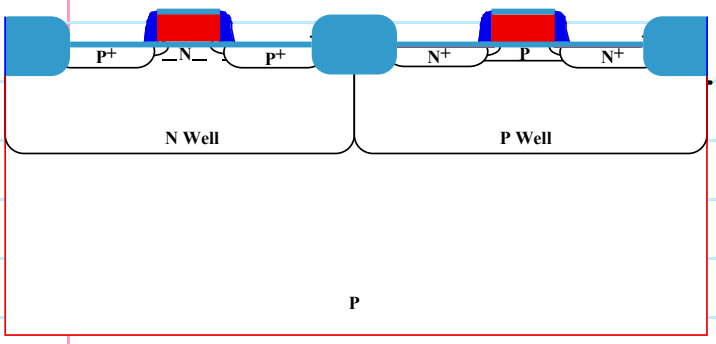


- Mask #9 protects the PMOS devices, An As+ implant forms the NMOS source and drain regions (typically 2-4 x 10^{15} cm^{-2} @ 75 KeV).

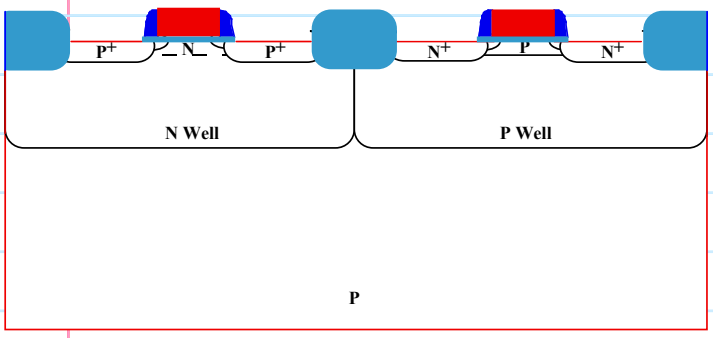


- Mask #10 protects the NMOS devices, A B+ implant forms the PMOS source and drain regions (typically 1-3 x 10^{15} cm^{-2} @ 50 KeV).

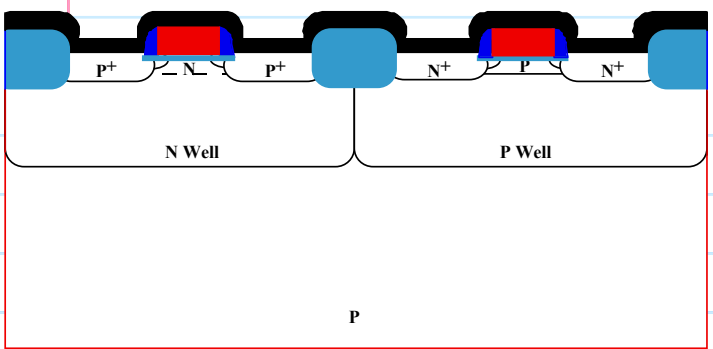
Also makes the PMOS gates p+.



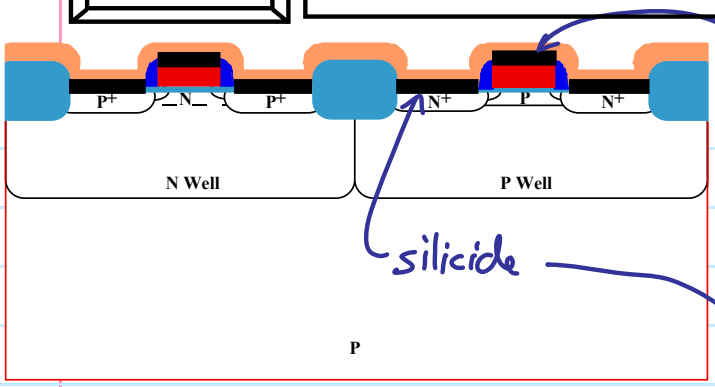
- A final high temperature anneal drives-in the junctions and repairs implant damage (typically 30 min @ 900°C or 1 min RTA @ 1000°C).



- An unmasked oxide etch allows contacts to Si and poly regions.



- Ti is deposited by sputtering (typically 100 nm).

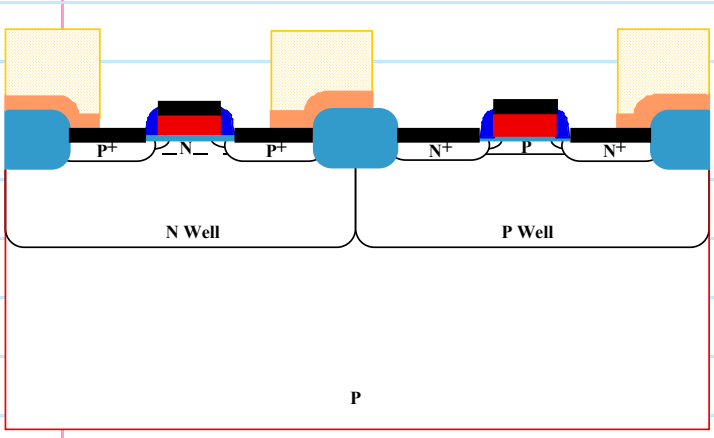


polycide

silicide

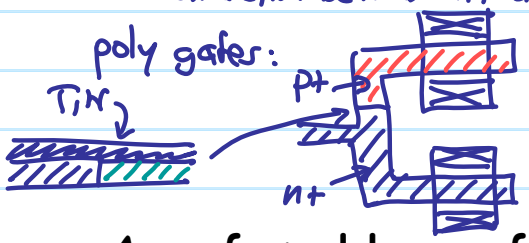
- The Ti is reacted in an N_2 ambient, forming $TiSi_2$ and TiN (typically 1 min @ 600 - 700°C).

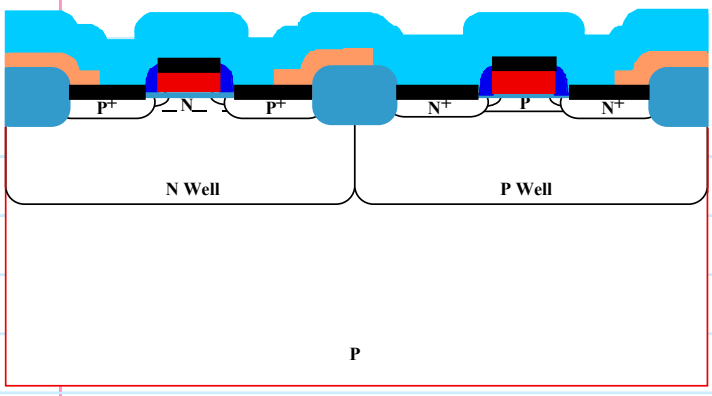
reduce the sheet resistance of the poly-Si gate & diffused junctions



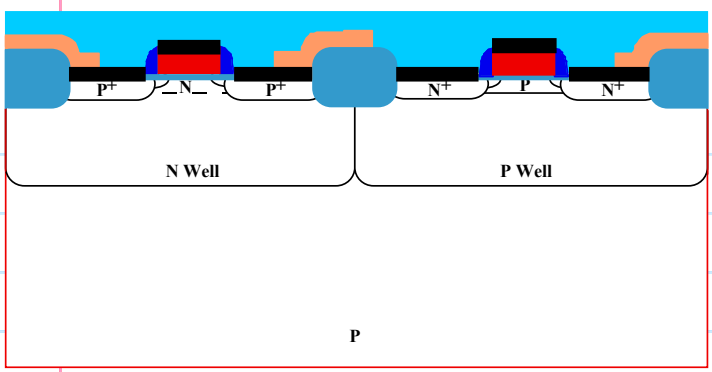
- Mask #11 is used to etch the TiN, forming local interconnects.

These TiN interconnects now allow interconnection between n+ and p+ poly gates:

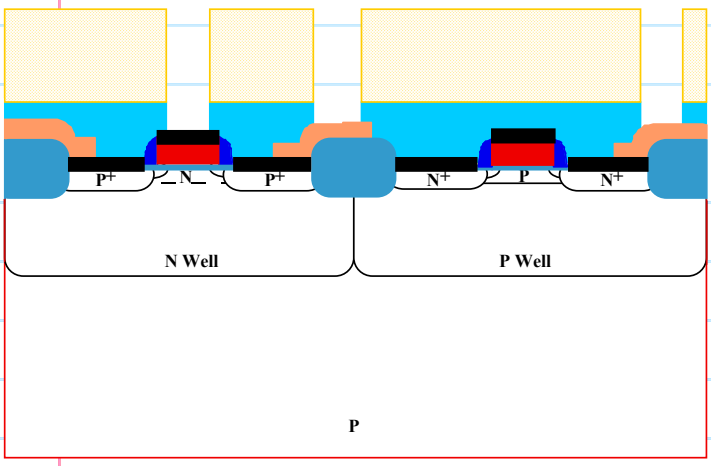




- A conformal layer of SiO_2 is deposited by LPCVD (typically 1 μm).

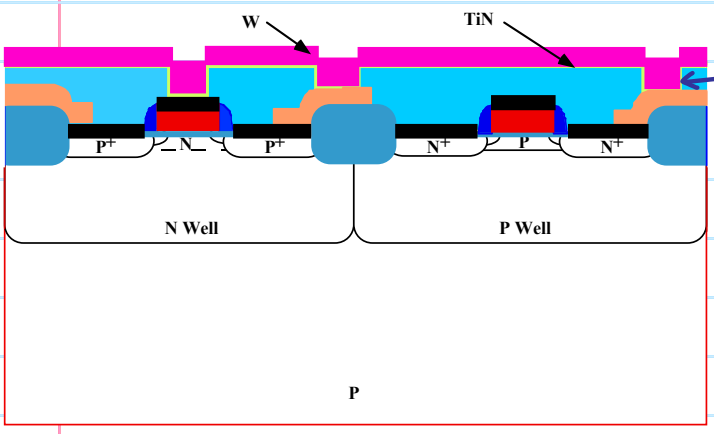


- CMP is used to planarize the wafer surface.



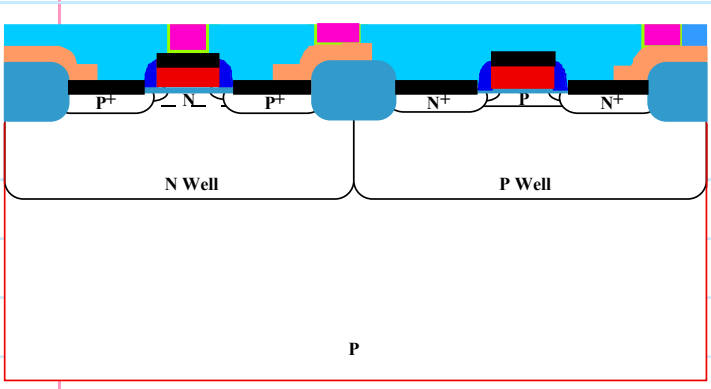
- Mask #12 is used to define the contact holes. The SiO_2 is etched.

① Helps w metal adhesion.
② Serves as a barrier against pitting & spiking of metal into the silicide.

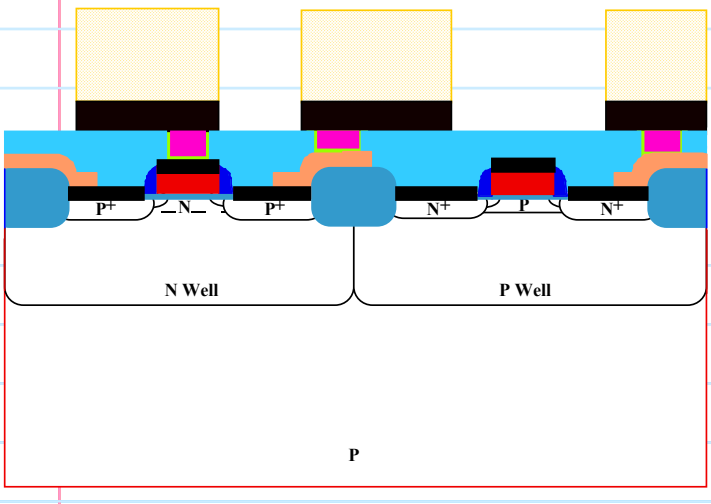


(the green layer)

- A thin TiN barrier layer is deposited by sputtering (typically a few tens of nm), followed by W CVD deposition.



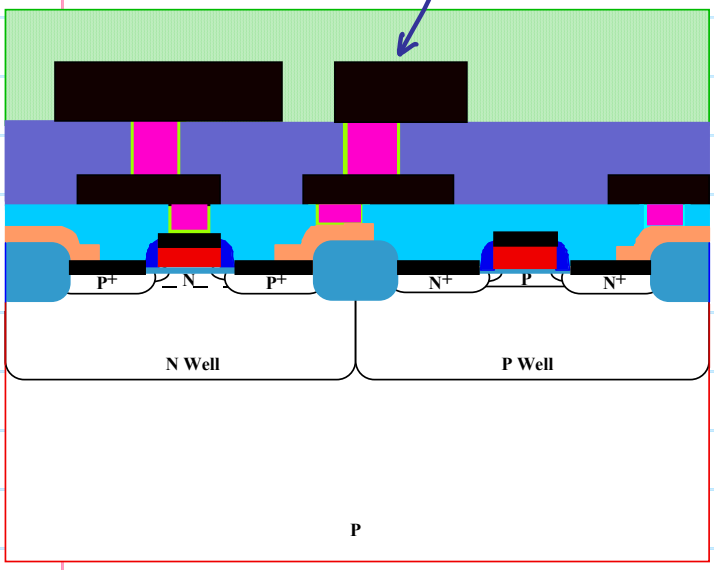
- CMP is used to planarize the wafer surface, completing the damascene process.



or AlCu for better resistance against electromigration

- Al is deposited on the wafer by sputtering. Mask #13 is used to pattern the Al and plasma etching is used to etch it.

Electroplated Cu for higher levels of metal.



- Intermetal dielectric and second level metal are deposited and defined in the same way as level #1. Mask #14 is used to define contact vias and Mask #15 is used to define metal 2. A final passivation layer of Si_3N_4 is deposited by PECVD and patterned with Mask #16.

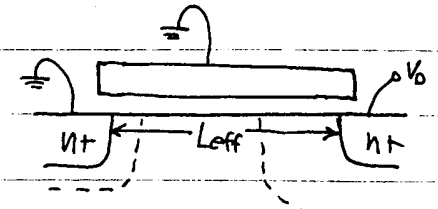
There will be more levels of metal → Cu.
Use a dual-damascene process.

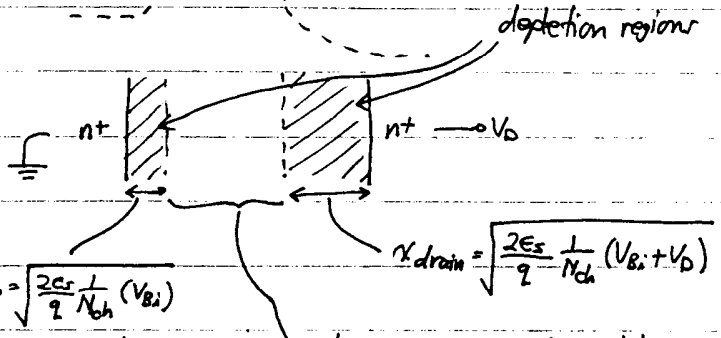
<u>Differences from older long-channel CMOS</u>:

1. Very thin lightly doped drain (LDD) and source junctions; very small x_j
2. Very thin gate oxide
3. Much higher channel & substrate doping
   - ↳ So why these differences? <u>Answer</u>: punchthrough!

Punchthrough (bulk)

⇒ the simple model:





$$\left[V_{Bi} = \frac{kT}{q}\ln\frac{N_{D(or)}}{n_i} + \frac{kT}{q}\ln\frac{N_A}{n_i}\right]_{ch}$$

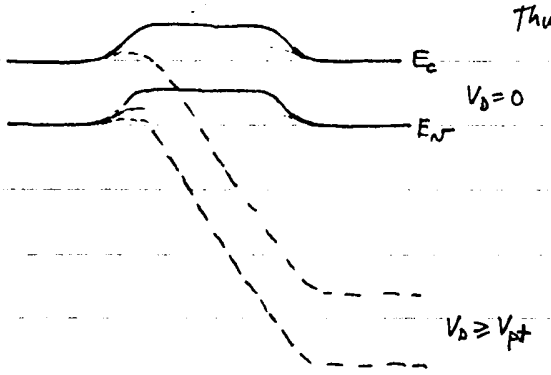
↑ 0.56V

when = 0 ⇒ punchthrough!

⇒ when $x_s + x_D > L_{eff}$ ⇒ punchthrough!

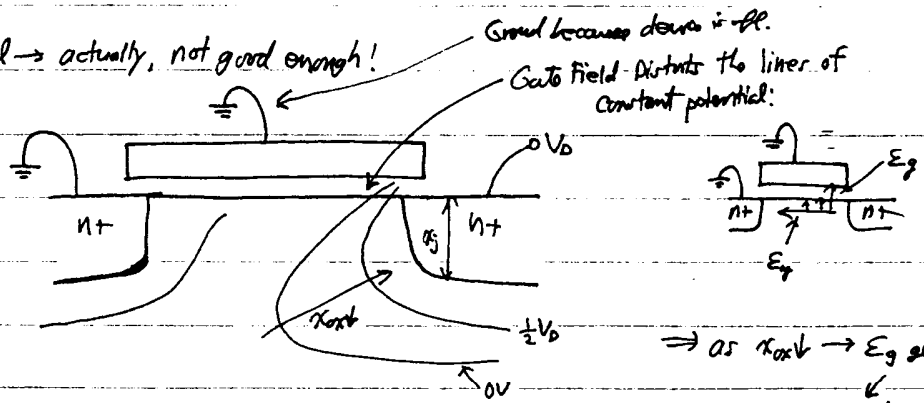
Thus, as $L_{eff}\downarrow$ → punchthrough $I_D\uparrow$

↑ as we go to short-channel devices!



Above: simple model → actually, not good enough!

⇒ why?



⇒ as $x_{ox}\downarrow$ → E_g gets larger

↙ punchthrough is suppressed close to the surface

∴ if $x_j\downarrow$ → D/S depletion regions suppressed by gate

∴ less punchthrough I_D

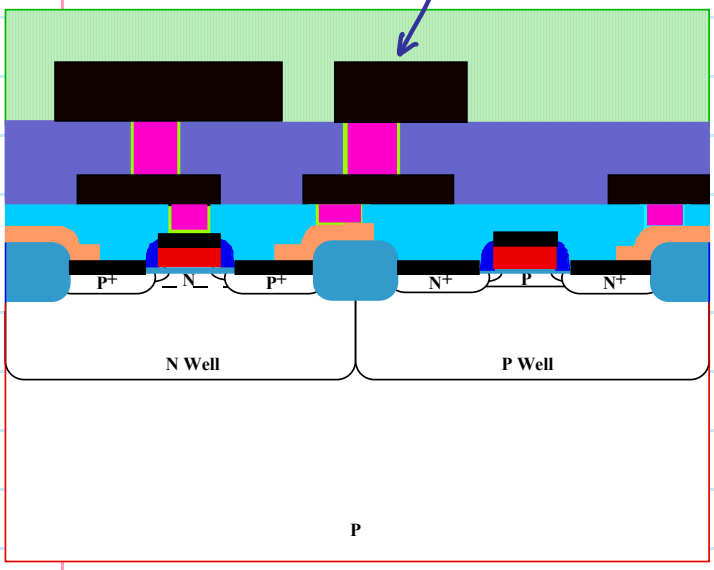
Thus, bulk punchthrough depends on:

	L ↓	x_j ↓	x_{ox} ↓	N_{ch} ↑
punch-through	worse	better	better	better

↑ since x_{drain} ↓ when N_{ch} ↑ (decreases depletion layer width)

⇒ these dictate the technology as device L gets shorter!

Electroplated Cu for higher levels of metal.



- Intermetal dielectric and second level metal are deposited and defined in the same way as level #1. Mask #14 is used to define contact vias and Mask #15 is used to define metal 2. A final passivation layer of Si_3N_4 is deposited by PECVD and patterned with Mask #16.

There will be more levels of metal → Cu.
Use a dual-damascene process.

<u>Differences from older long-channel CMOS</u>:

1. Very thin lightly doped drain (LDD) and source junctions; very small x_j
2. Very thin gate oxide
3. Much higher channel & substrate doping
   - ↳ So why these differences? <u>Answer</u>: punchthrough!

<u>Remarks</u>:

- More advanced technologies also require a thinner gate oxide, for reasons explained above
  - ↳ SiO_2 can no longer satisfy, since the gate oxide thicknesses below 1 nm have leakage issues
  - ↳ Newest technologies use ALD high-k oxides (e.g., HfO_2)
- More advanced technologies must reduce as much as possible the interconnect capacitance ⇨ low-k inter-metal dielectrics
- The use of LOCOS in this process limits it to 0.18-0.25 μm channel lengths
  - ↳ Any smaller and the LOCOS encroachment it too much
  - ↳ Plus, the topography makes lithography for smaller channel lengths more difficult