CMOS Layout

Layout basically transfers circuit to be implemented to Silicon. Hence it is tightly linked to silicon wafer processing. Hence we will cover this now in brief and then move on to do a primitive layout of inverter. Then we will look at design rules and implementation of other gates and layout in CMOS technology.

Basic CMOS Process

Lithography

MOS device structure shown earlier is implemented in silicon with help of layout masks. The layout has mask names which usually indicate either a function or a material layer. These geometrical patterns for each layer on the layout have to be transferred to Silicon. This is achieved by the lithography step in processing. The basic principle is to use a photosensitive material (X ray sensitive for X-ray lithography, e-beam sensitive for e-beam lithography) for exposing the layout geometries.

The photosensitive material called resist is deposited on Silicon Wafer which possibly has one or more materials grown or deposited on it as shown.

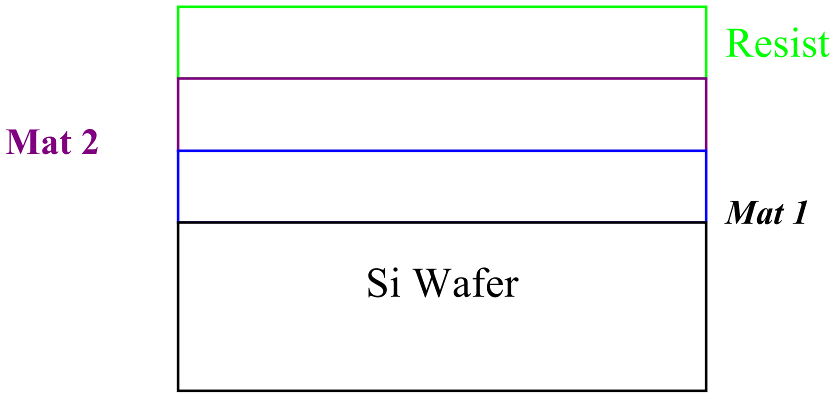


Figure: Silicon Cross Section after resist deposition

The resist is deposited on Silicon wafer by spinning the wafer at high speed with droplets of dissolved resist. The solvent is then evaporated by baking the wafer at about 50° - 60° C (Soft bake). The resist is typically 0.1-1µm thick. Then ultraviolet light is shone through the mask plate as shown.

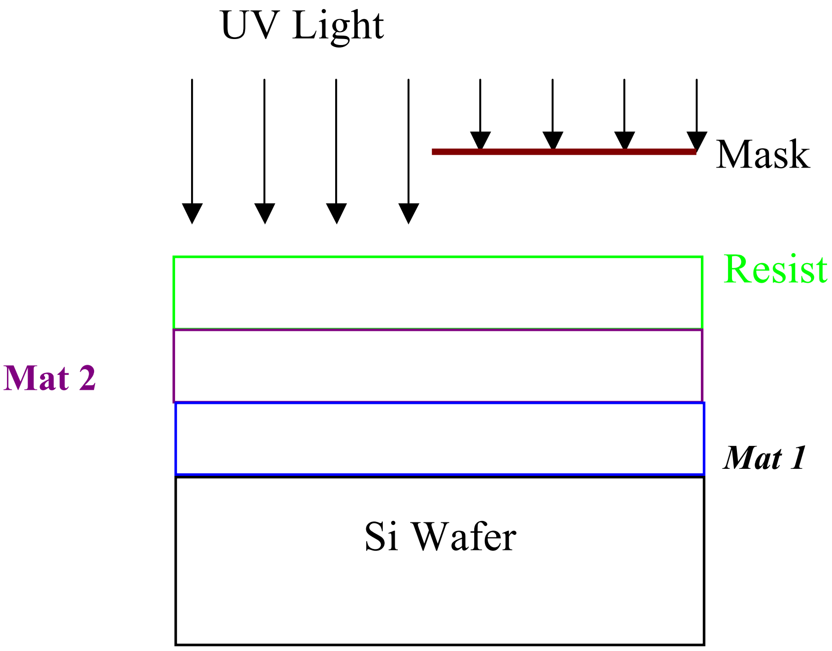


Figure : The process of exposure

This UV light comes from a very complicated optical system to ensure uniform intensity over the large extent of wafer and proper focusing on resist. The positive resist exposed to UV light softens and can be dissolved in a developer to expose mat 2. Now, mat 2 and mat 1 in the exposed area can be selectively etched to do processing in the masked Silicon.

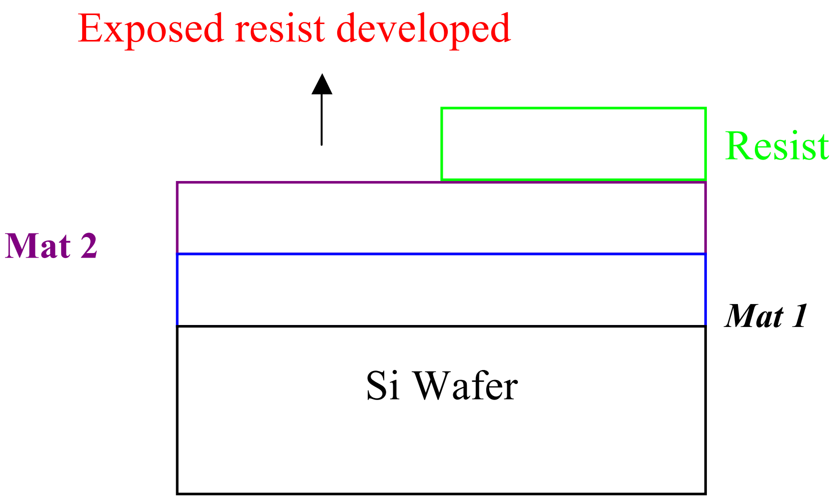


Figure : Cross Section after development

However, due to diffraction of light around mask edges, nonplanar surfaces and slight over or under exposures, it is difficult to get exact replica of open feature transferred to the resist. Hence actual mask sizes on layout are not transferred to silicon exactly and there is always variability. This variability has become a serious concern below 90nm technology where impact on circuit performance due to this is serious.

One more consequence for designers arises from the fact that features on a layer cannot align exactly with the previous layer. Hence to ensure that the circuit functions as expected, the layout sizes may need to be adjusted. Another option is to achieve self alignment where lithography has no impact. However, this is not possible for every layer.

Etching

The normal step, which follows lithography, is etching. Etching is a selective removal of material from the wafer. The basic principle is that the materials to be etched will only be removed from exposed areas and will not be removed from the areas protected by unexposed resist.

Ideally, any material other than the material to be etched should not be affected in this process. The silicon wafer is in the state shown after mat 2 has been etched after the development previously shown. Note that sometimes the edge etched is not vertical and there is significant undercut in mat 2 under resist.

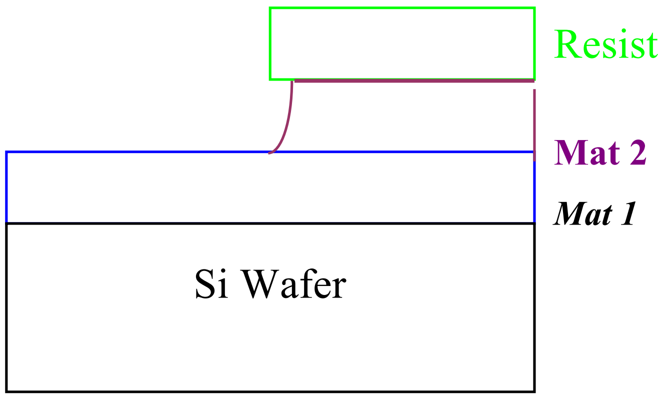


Figure : Profile after etching mat 2

This undercut in etching is a problem which increases some features and decreases others and hence some adjustments to layout are needed as a consequence.

Deposition

One would like to have uniform cover of deposited material on the surface of the wafer. It is quite easy to get uniformity if the surface on which deposition is done is planar. However, a typical deposition pattern with a CVD isotropic deposition system will be as shown.

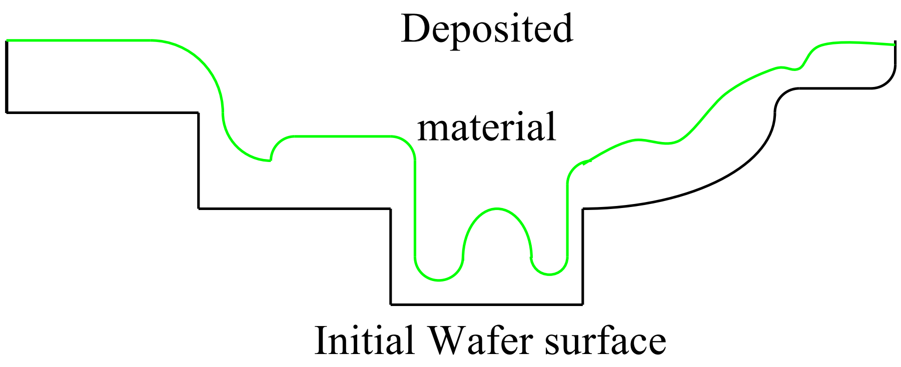


Figure : Profile after deposition on non-planar surface

One can easily notice non-uniformity of the deposited layer. In fact, the layer is very thin on the vertical walls of initial surface. This may pose some problems for metal layer reliability due to higher current density. This sometimes dictates a higher line width for metal lines than minimum allowed width.

Oxidation

This is one of the most important steps in Silicon Processing as oxide can be used as a block for impurity diffusion, as a dielectric for capacitance or interlayer insulation, or as a protective layer to guard Silicon against contamination. Oxide can also be selectively grown using nitride as a mask as Nitride oxidation is much slower than Silicon oxidation. Thus the areas of Silicon covered by nitride will not oxidize.

The oxide is normally grown in dry oxygen, steam, or high-pressure oxygen in 750°c to 1200°c range. The main reactions are

$$Si + O_2 \rightarrow SiO_2$$

$$Si + 2H_2O \rightarrow SiO_2 + 2H_2$$

The wet oxidation is much faster than dry oxidation. There are certain general guidelines on which methods to use.

(i) When very high quality, thin oxide is needed such as gate oxide in MOS processing, only dry oxidation is used. Nowadays Nitrated oxides are used.

(ii) When good thick oxide is needed dry oxidation followed by wet oxidation is used. This reduces the time for oxidation by a factor of about 10. This is the only method for growth of good quality thick oxides of thickness around 1μm in practice. The high-pressure oxidation can also be used to enhance the growth rate.

(iii) Deposited oxide layers are normally used for inter-metal isolation, since metals such as Aluminum cannot sustain high oxidation temperatures.

There are many consequences of oxidation which affect designers or alter processing somewhat. These will now be discussed briefly.

(1) While oxidizing, about 50% of Si layer is consumed. Hence the oxide layer will have higher elevation than original Si and will eventually lead to non-planar starting surfaces.

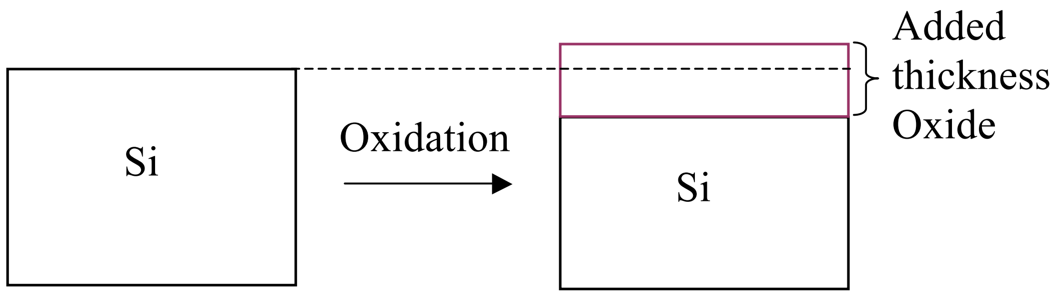


Figure : Increased Elevation due to oxidation

(2) Nitride is an effective mask for oxidation. If it were a perfect mask, the oxidation after patterning nitride will look as shown.

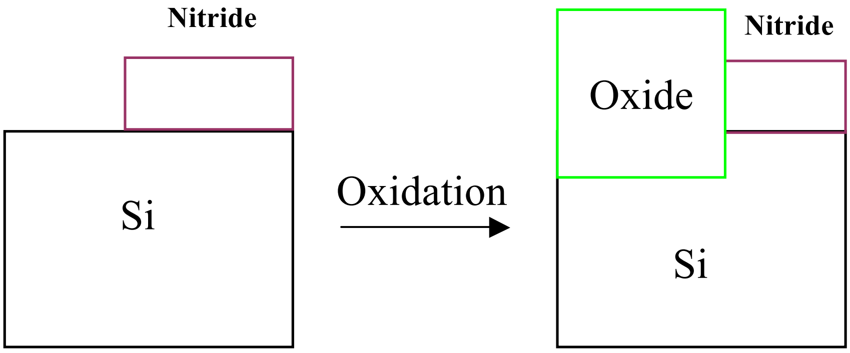


Figure : Selective Oxidation with Perfect Nitride

In reality, due to high stresses near oxide-nitride interface, oxide encroaches under nitride and lifts it off to produce bird's beak structure.

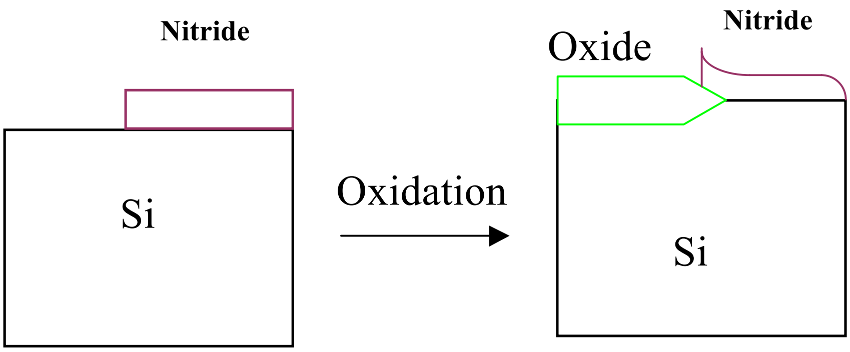


Figure : Bird's Beak Formation

This has very profound implications for width of MOS devices as shown.

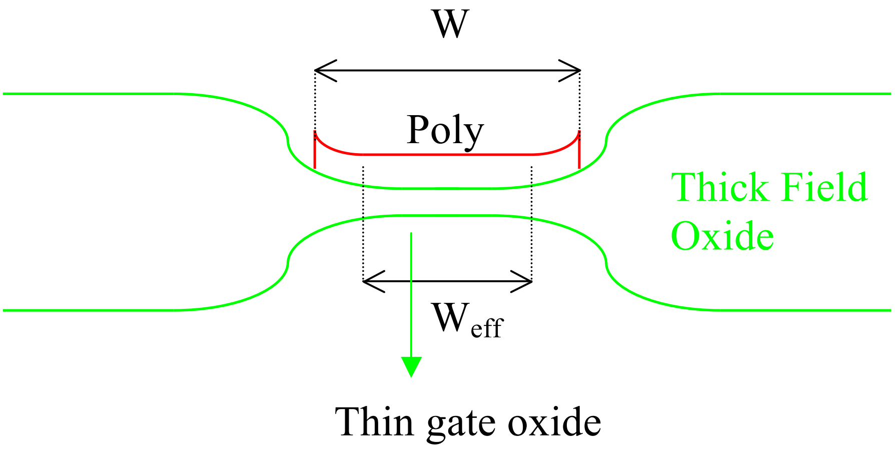


Figure : Width Reduction due to Bird's Beak

The thickness of oxide away from W_{eff} is larger than gate oxide thickness. Since the threshold voltage increases with increasing oxide thickness, there will be width reduction. Since the oxide thickness under poly gradually increases, different parts away from W_{eff} will progressively be turned on with increasing gate voltages. Hence, a modelling problem. For technology below 0.18μm, shallow trench isolation is used instead to avoid these problems.

(3) Impurity Segregation

Due to high temperature during oxidation, the impurities tend to diffuse deeper. The redistribution, however, is affected by the presence of oxide – Si interface. The concentration of Boron near the interface is reduced in Si compared to normal concentration and affects the threshold voltage. The concentration of P and As is increased as shown.

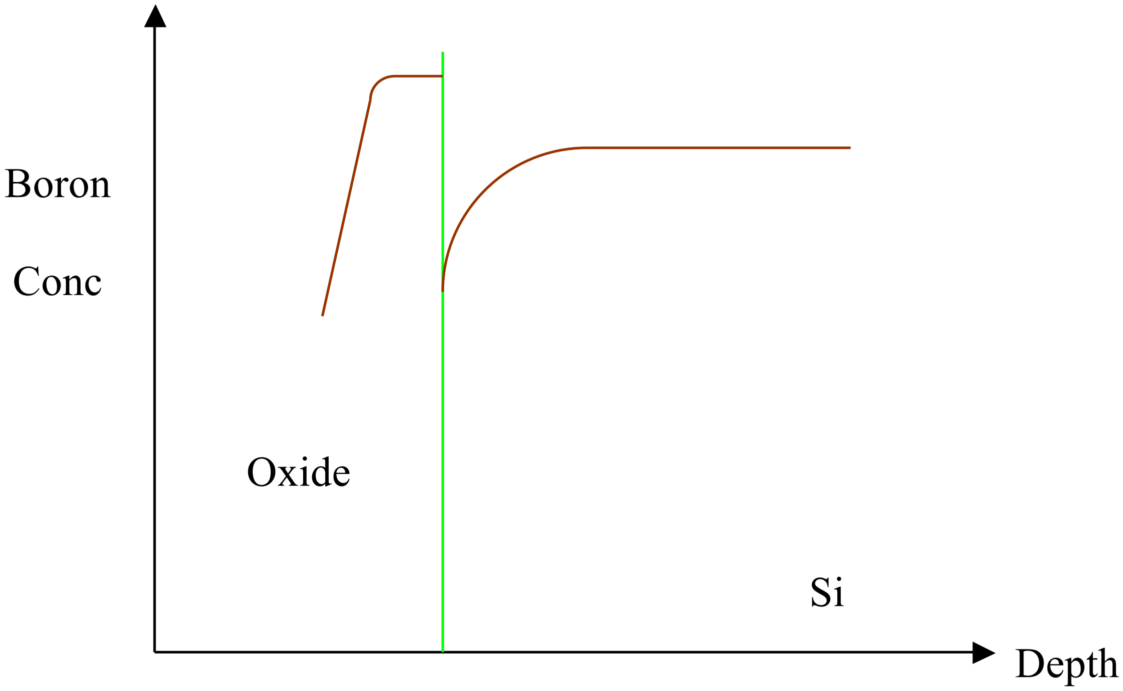


Figure : Boron Redistribution due to Segregation

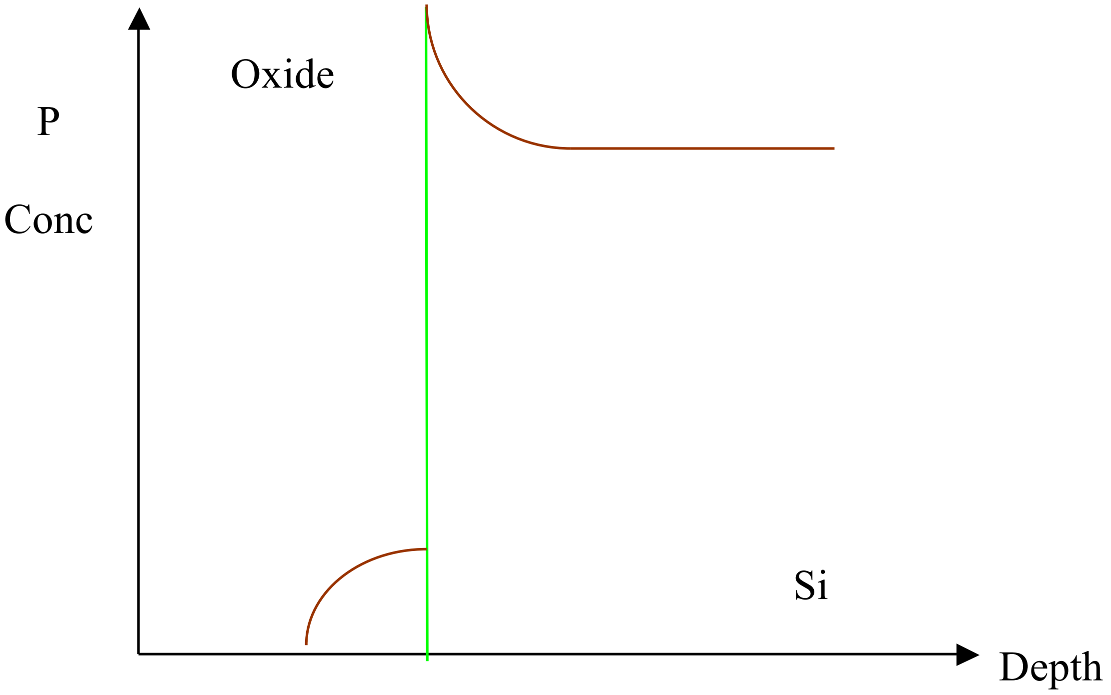


Figure : Phosphorus Redistribution due to Segregation

This may have some effects on threshold voltages when Boron threshold adjust implant is used. Also, high pressure oxidation alters the level of segregation.

Ion Implantation

This is a process of introducing impurity or other ions into Si by giving them enough energy to penetrate into Si. Higher energy ions will penetrate deeper into Silicon. Hence the ion energy and number of ions to be introduced measured by dose determine the impurity profile. Normally this step is used to selectively dope silicon where the parts covered by resist will not allow implant to pass through.

If the scattering of impurity projectiles is random, it is easy to find the average depth of penetration (range R_p), standard deviation in the depth (ΔR_P), and standard deviation perpendicular to beam direction (ΔR_L). The typical values of range are in 0.01μm - 2μm interval.

For P, As and Sb, the distribution of ions is a Gaussian. For Boron, it is found that the distribution has a long tail which is modelled by a Pearson IV or V distribution function. High-energy Oxygen implants are popular for forming oxide deep inside Silicon for Silicon on Insulator (SOI) devices.

Precise control of doping profile and integrated doping concentration is possible by ion implants. However, it is not totally free of problems, which will be discussed in brief.

(1) Channeling: The incident ion beam could sometimes be perpendicular to one of the crystallographic planes in Si. In this case, due to less scattering along the path, the ion may penetrate 2 to 3 times deeper than the normal depth predicted by random scattering. To avoid this, the beam may be sent at 3° - 7° to the normal.

(2) Damage – Crystalline structure of Si is damaged by implants as Silicon ions/atoms are scattered randomly and displaced from their normal positions. This damage is to the extent that Silicon becomes amorphous in a short span. Hence a post implant anneal becomes essential to give enough thermal energy to Si to restore original crystal structure. Normally, an anneal of about 30 minutes at 850°c - 900°c is adequate to remove the crystalline damage and activate impurities by moving them into substitutional sites. Nowadays, rapid thermal annealing is often used as thermal budget is quite limited for short channel MOSFETs. RTA for 30 seconds at around 980°c is typically sufficient. SPIKE anneal is also used in case of tight thermal budget. This, however, will redistribute the impurities by diffusion and some of the implant advantages are adversely affected.

Diffusion

This is a natural phenomenon due to the tendency for impurities to move from a region of high concentration to a region of low concentration. Hence if some impurities are introduced near Silicon surface, at higher temperature they will diffuse deeper into Silicon. Diffusion was a very popular way of selectively introducing impurities in Silicon before implants came along.

Due to slow impurity diffusion in SiO2, oxide was used as a mask for selected introduction of impurities.

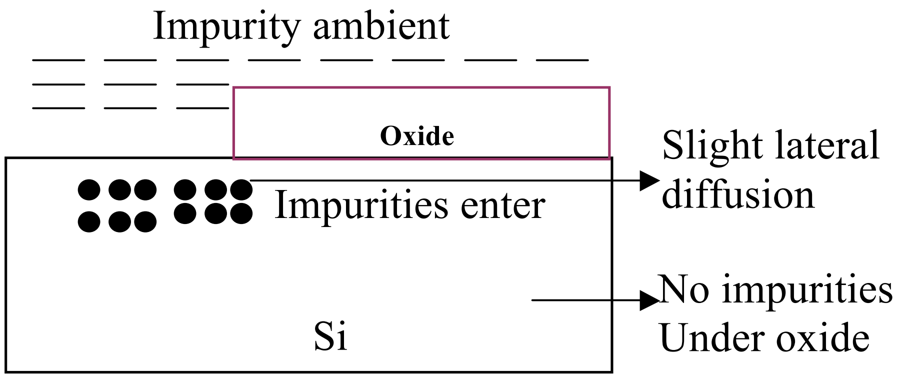


Figure : Oxide as Diffusion Mask

Even if this technique is not so much used in the present implant age, the diffusion is still a part of processing life. The post implant anneal makes the implanted impurities redistribute by diffusion.

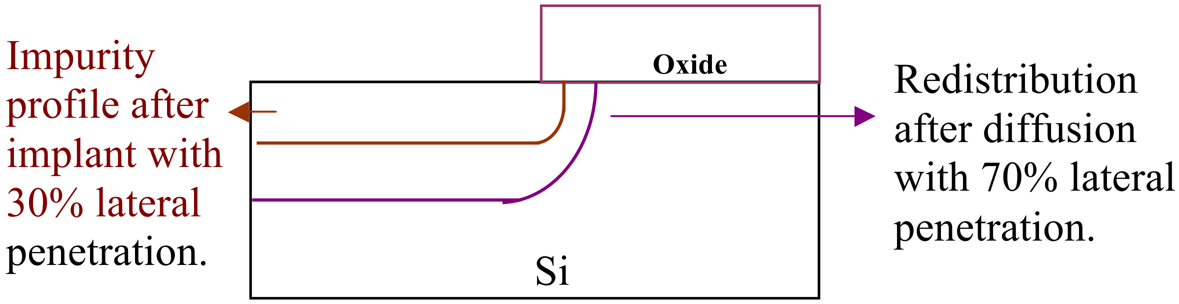

Impurity profile after implant with 30% lateral penetration.
Oxide
Redistribution after diffusion with 70% lateral penetration.
Si

Figure : Redistribution of Implanted Impurities after anneal

The diffusion takes place any time Silicon is heated. The presence of lateral diffusion at about 70% of depth decides some of the critical device features and inherently limits some of the device features and properties.

For very deep diffusions, the profile is a near Gaussian. For very shallow ones, it is a near error function. The characteristic depths are determined by the diffusion coefficient D which is a primary exponential function of the temperature and the time *t* of diffusion. $\sqrt{Dt}$ is typically the characteristic depth for the distribution. The diffusion coefficient D is a very strong function of concentration of vacancies and interstitial within the crystal. It also depends on the presence of other impurities and the other materials. One such effect we saw in segregation before. The other one is shown in the figure.

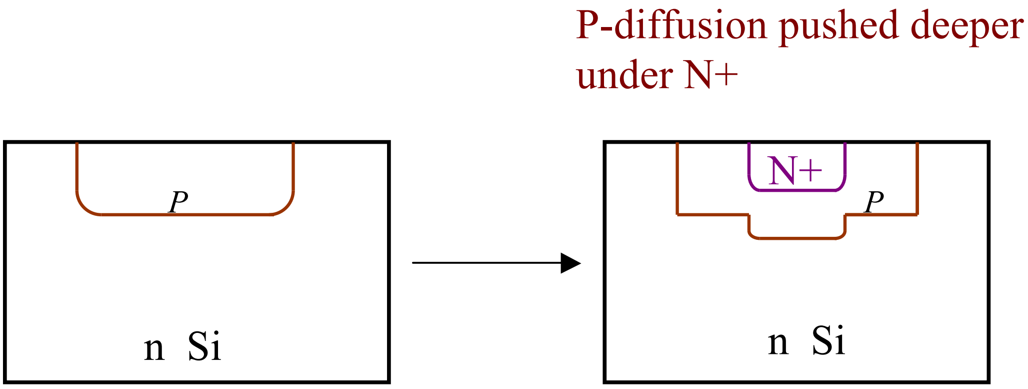
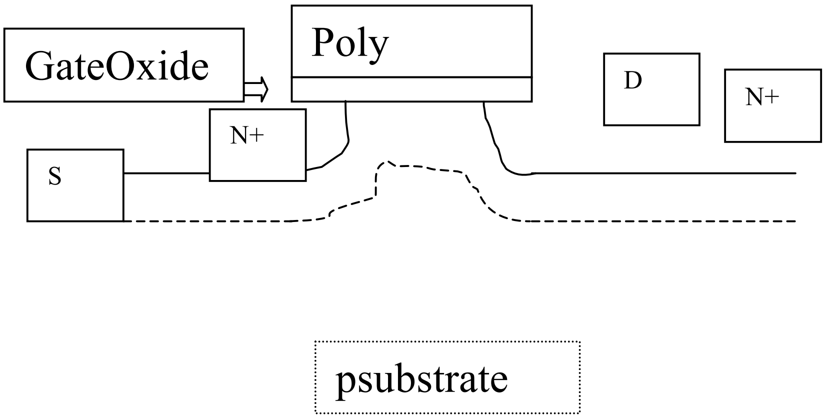


Figure : Effect of Impurity on Diffusion

Layout of Stand-alone CMOS Inverter

Earlier, we drew a circuit of CMOS inverter without worrying about how to make these devices and connect them on Silicon. Let us now discuss how many different layers will be needed on the layout at the minimum. Normally, many more layers are needed as there are additional steps and more than one metal layer.

We need to achieve the device structure below for both n and p devices.



(1) The starting substrate will normally have one type of doping. Hence to make both p and n channel devices, we need to create an n type regions in p-type substrate. This level is labeled as NWEL colored in brown.

(2) There should then be two more layers to form the n-channel and p-channel devices. Also, there should be a layer that defines gate regions.

(3) There should be another layer, which will make holes in the insulators or other materials to connect metal to appropriate regions in Si e.g. connecting V_{DD} to source of the p-channel device.

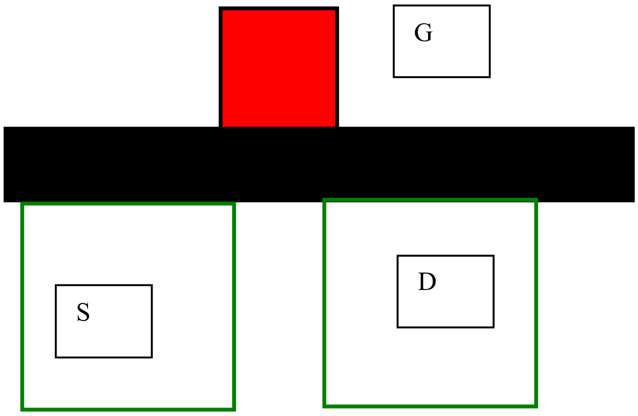
(4) A layer for metal which can carry V_{DD}, V_{SS}, input and output signal to the outside world.

The mask is determined by process limitations.

(1) Consider forming source-drain-gate structure. Let us say we form S/D first. Then we grow gate oxide and put gate on the top. For MOSFET, gate must overlap S/D. As these are different masks, that will mean gate should be wider by a large amount so that even if lithography does not align properly the gate, it will still overlap S/D. Hence if x is lithography misalignment possible, y is separation between S and D junctions under gate, then gate would need to be at least (y+2x) wide. Of this, only distance y induces channel, the rest

just gives extra load to slow circuit down. Hence a bright trick is applied in processing.

In this method, the gate is first formed and S/D implant energy is chosen so that it does not go through the polysilicon gate. Hence only the exposed silicon regions on the two sides of the gate receive implant but S/D mask is a sum of source, channel and drain area. The source drain regions automatically attach to the gate as shown. This has made 45nm technology today feasible.



(2) As many devices are integrated on any chip, isolation region with thick oxide is formed wherever there are no channel/S/D regions.

(3) S/D or some other regions need to connect to metal. For this, a contact is placed in S/D areas. As despite of misalignment due

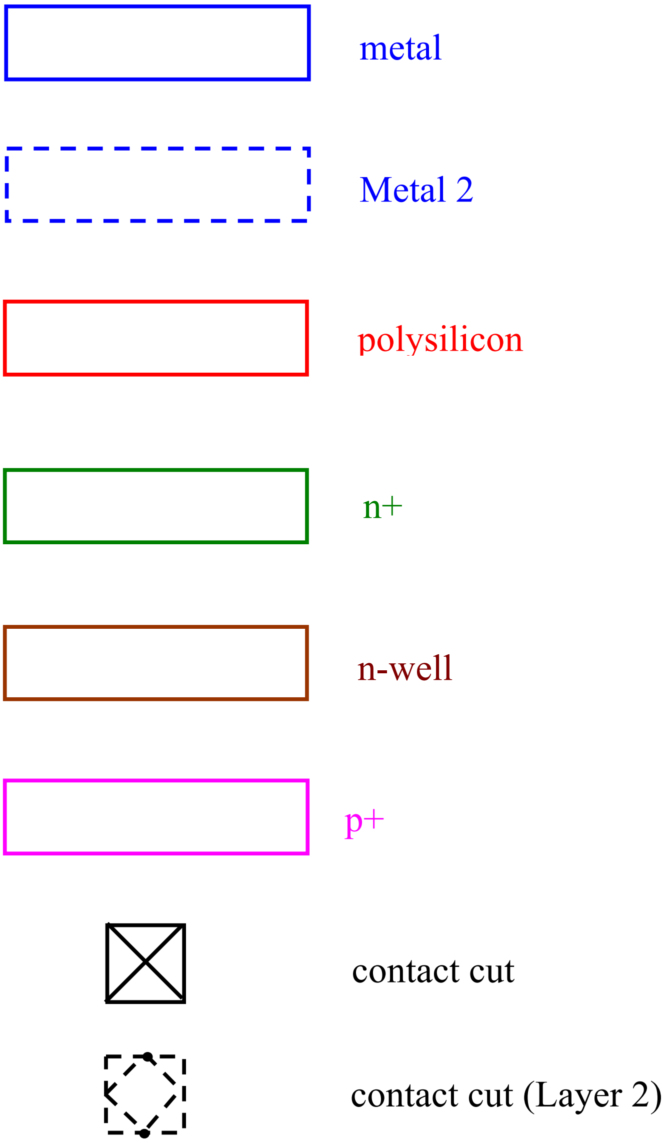
to lithography, the contact must be in this region, S/D areas have to be larger than the contact by a certain amount.

(4) Similarly, as metal must cover full contact despite of misalignment, metal mask is larger than the contact by the same amount.

Note that it is possible for polysilicon, which forms the gate material, to be used as a short interconnect if many inputs are connected together. Also source or drain regions can be used as short interconnects if they are connected together for the same channel devices. Also, note that many other considerations may add some process complexities or masks to the above basic set. The color convention is now given in the table. The convention for poly, N type active area, contact and metal is universal. Others may vary a lot.

Layer	Color
NWEL	Brown
POLY	Red
N+	Green
P+	Pink/Purple
CONT	Black
METAL	Blue

Layout diagram colour code



metal

Metal 2

polysilicon

n+

n-well

p+

contact cut

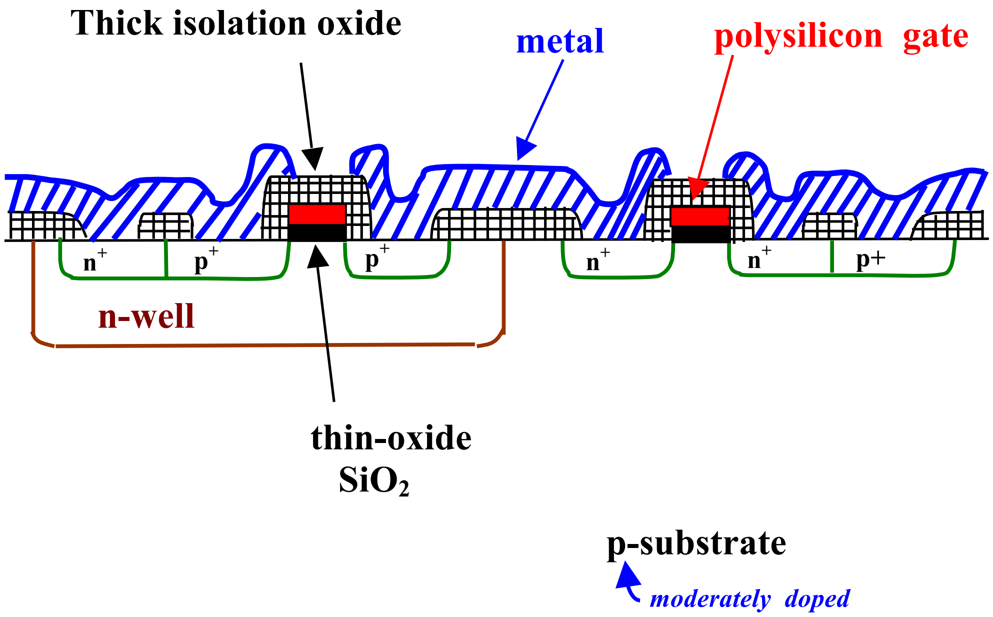
contact cut (Layer 2)

Let us now evolve a primitive, stand alone, layout of the inverter.

Notice that an n+ (Also denotes by NSD – n type source drain) area is added which is butting to p+ (PSD) and connected to V_{DD}. This is extra geometry added for reliability. Similar area is also needed with n+. We have already looked at the process cross section along the line drawn. The sizes of these geometries and spacing are based design rules, which now follow. The design rules are largely determined by the processes.

Cross-section within silicon of this inverter along a line at the centre of the layout is as shown. For one process, the additional contacts are not included for simplicity.

n-well process

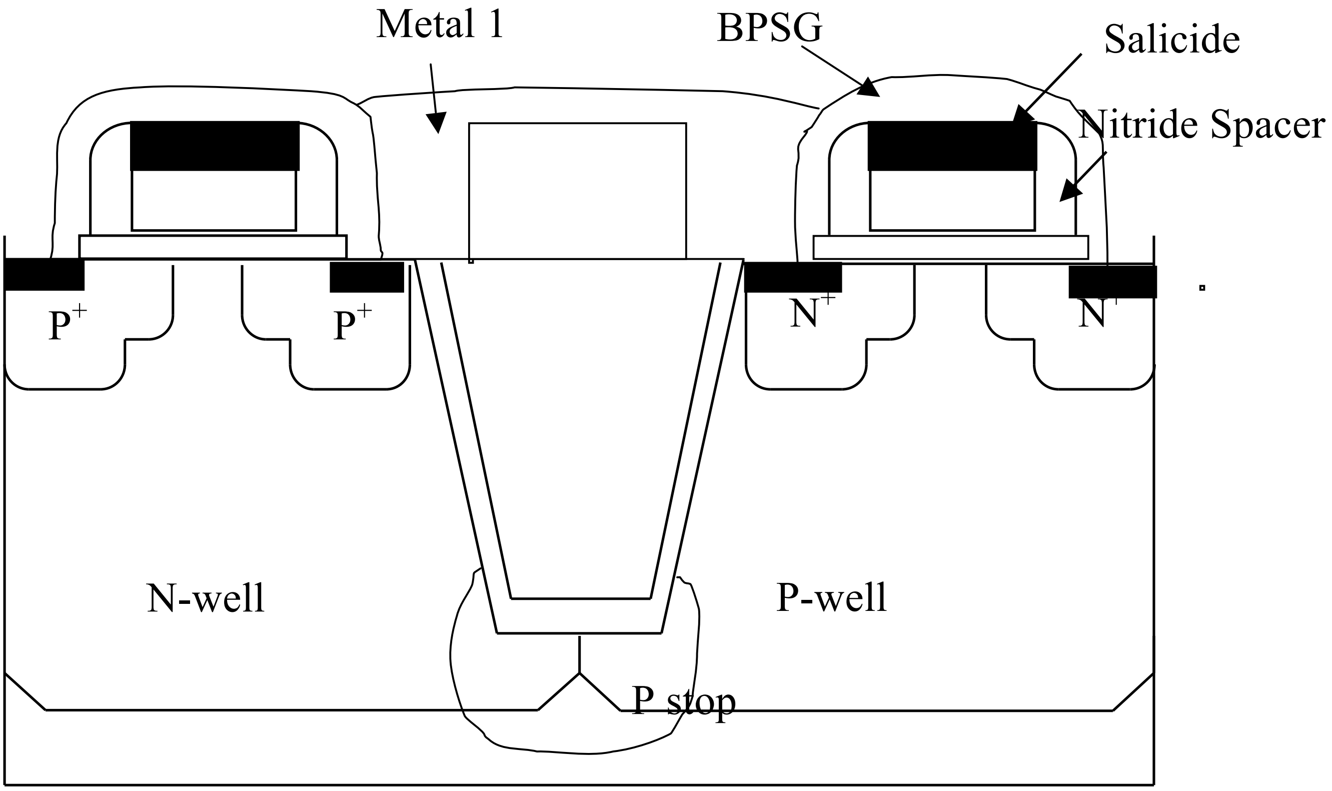


Other processes :

p-well, twin-tub,

SOI (silicon on insulator)

Oxide or sapphire or magnesium aluminate



Modern CMOS inverter process based cross section

Process Flow of simple CMOS process

(1) Starting p-type wafer – deep Nwell formation to create p-channel MOSFETs.

(2) Selective thick isolation oxide where the is no N+ or P+ regions.

(3) Clear P+ and N+ areas and grow thing gate oxide.

(4) Deposit polysilicon and patern poly/gate mask to form gates of MOSFETs.

(5) Open only N+ areas in the resist and implant As. As stops in resist in P+ areas, isolation and gate giving self alighnment.

(6) Open only P+ areas in the resist and implant B. B stops in resist in N+ areas, isolation and gate giving self alighnment.

(7) Deposit thick BPSG oxide. This trovides inslulation between metal that will fill contact. All channel poly is covered by this thick oxide. Hence metal can run over polysilicon without connecting to it.

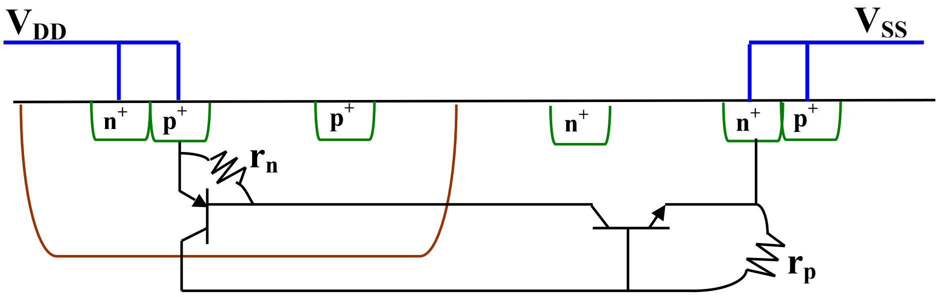
(8) Open contacts using contact mask.

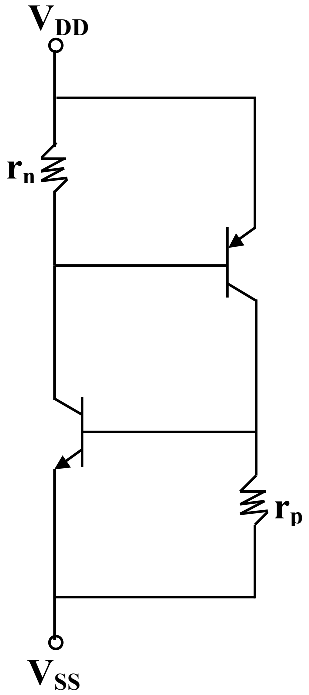
(9) Deposit metal 1 and pattern metal 1 mask to remove unwanted metal.

(10) Deoposit thick oxide and open vias to connect to Metal 1.

(11) Deposit metal 2 and pattern metal 2 mask to remove unwanted metal.

latch-up

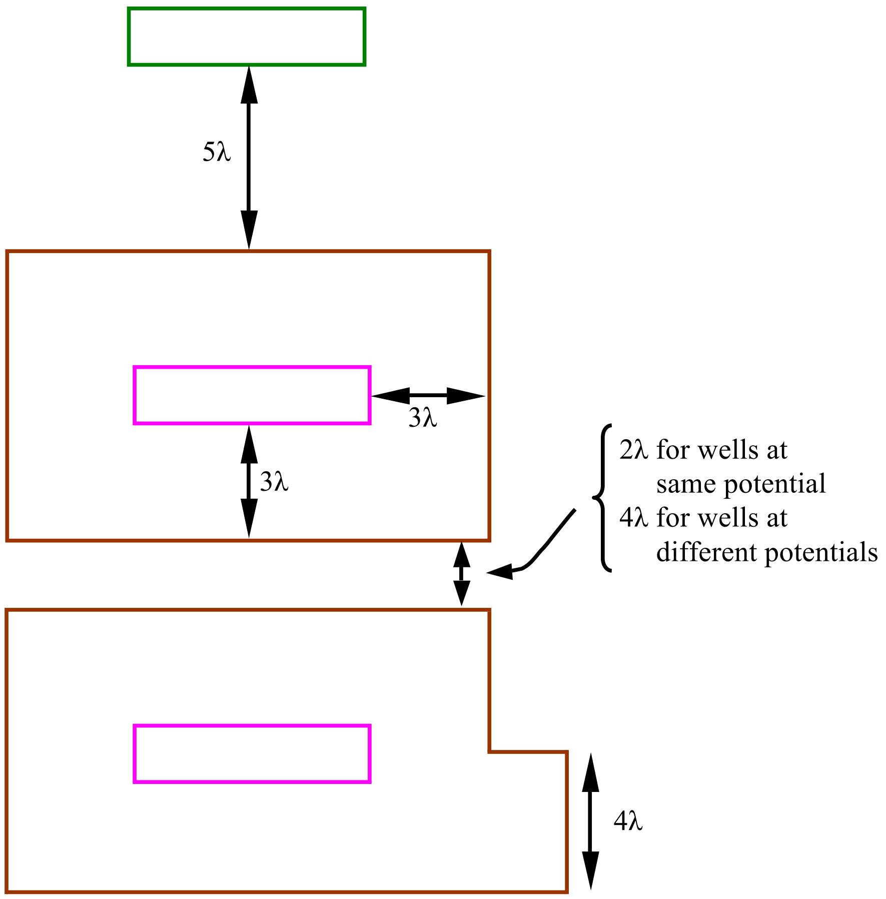




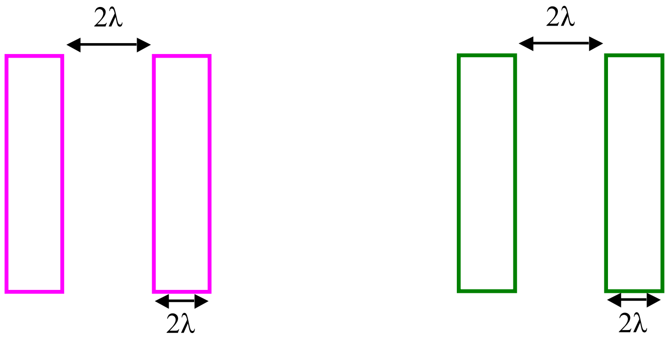
What can we do to suppress this? Gives rise to large design rules and extra features. The design rules are technology independent using a parameter λ which shrinks as technology scales.

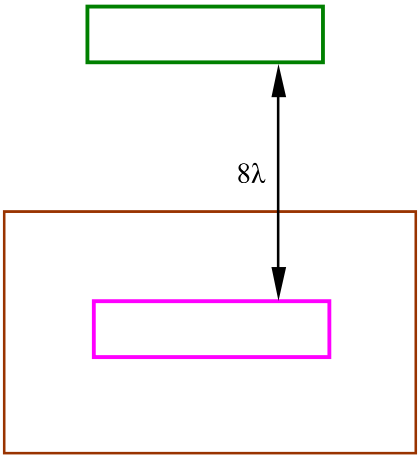
Design Rules

n-well spacing

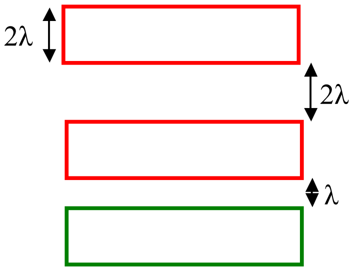


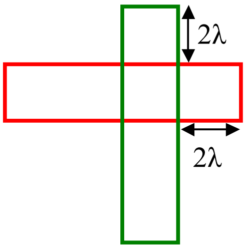
P+(PSD)/N+(NSD) spacing



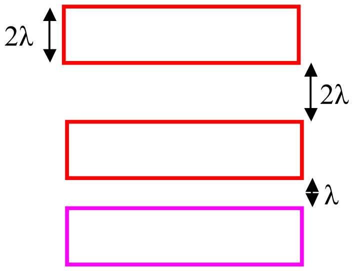


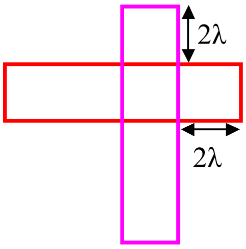
Polysilicon-n+ spacing



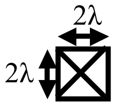


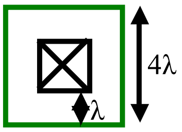
Polysilicon-p^+ spacing

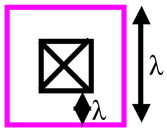


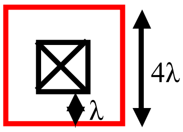


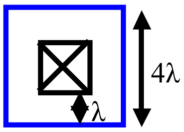
contact spacing

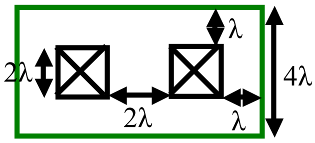


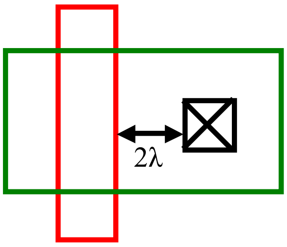


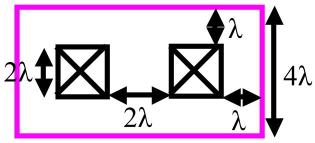


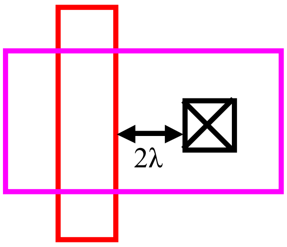


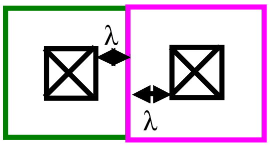


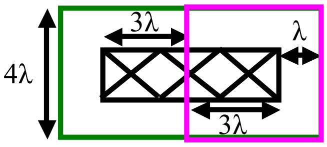


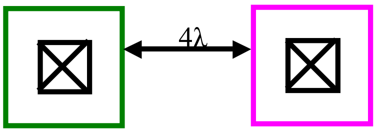




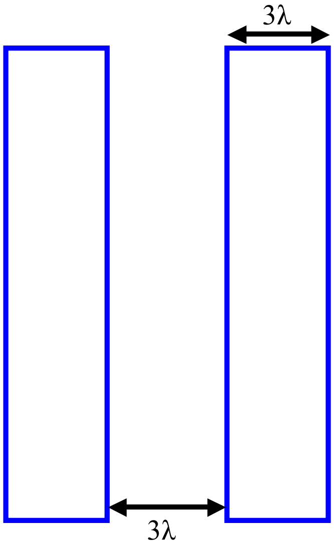



λ
λ

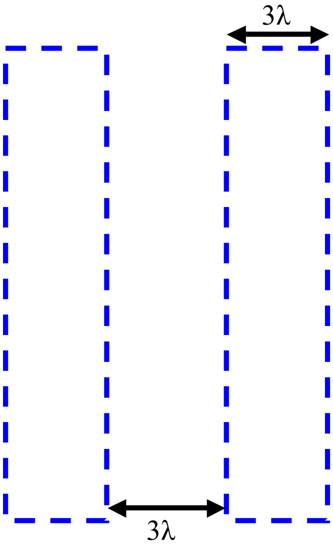

3λ
λ
4λ
3λ


4λ

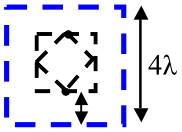
metal spacing



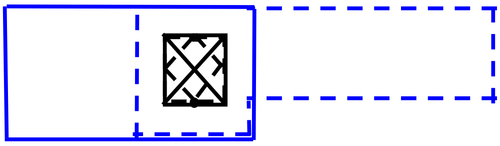
metal 2 spacing



Metal Layer 2 contact



Metal 1 to Metal 2 connection



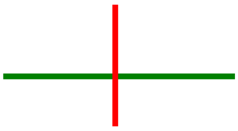
Stick diagram

A simple representation of layout as a plan.

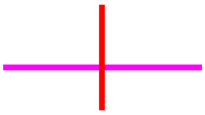
Useful to decide how to do actual physical layout.

— (blue)	metal
— (red)	polysilicon
— (green)	N+
— (Pink or purple)	P+
×	contact cut

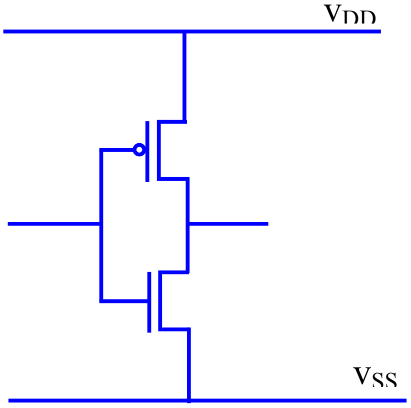
A transistor is formed whenever a polysilicon wire crosses a diffusion wire

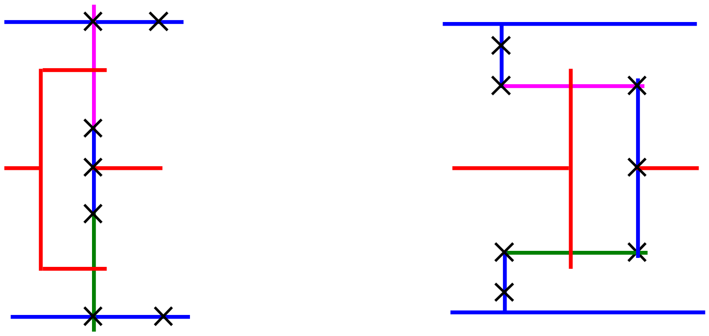


n-type transistor



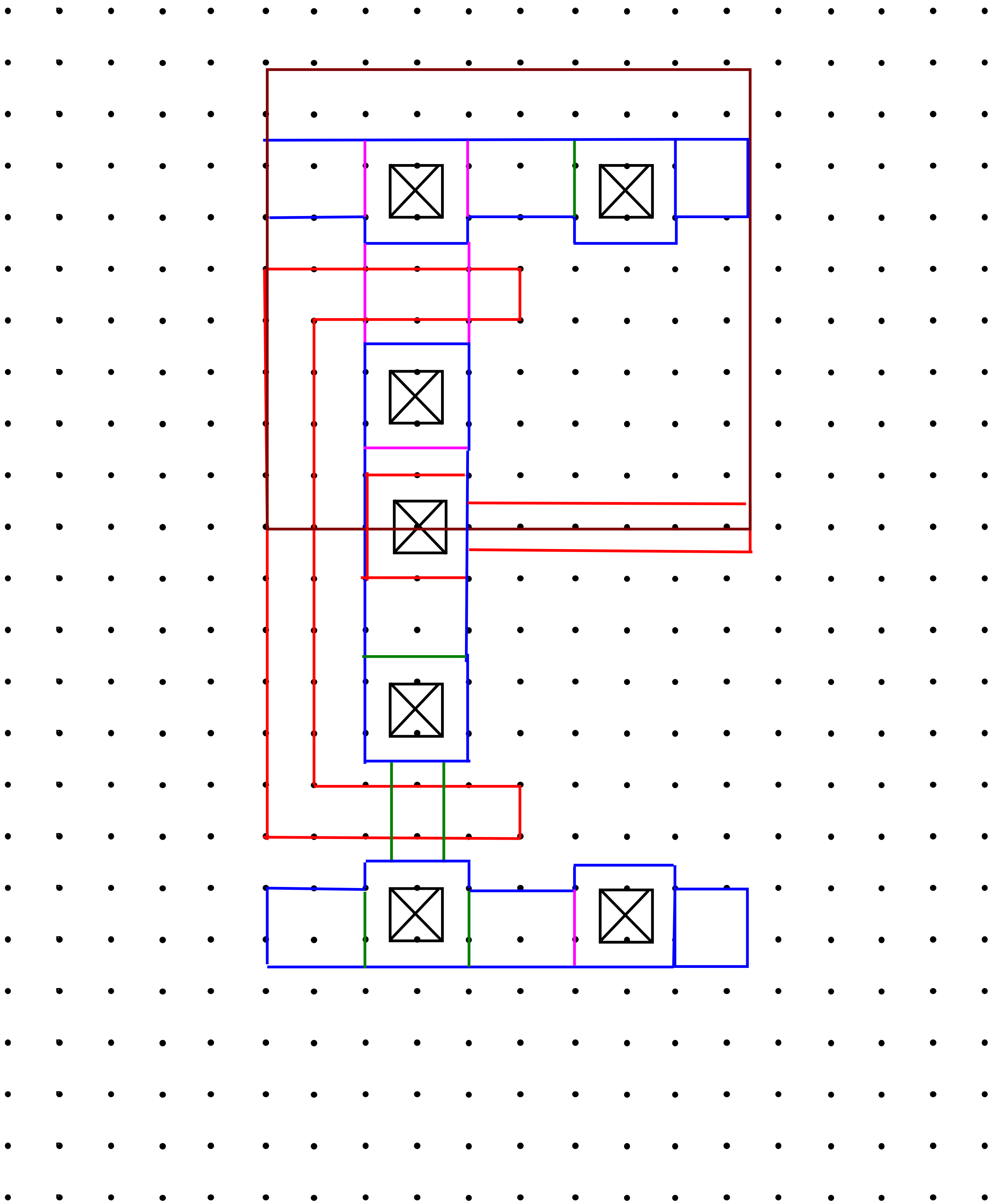
p-type transisto

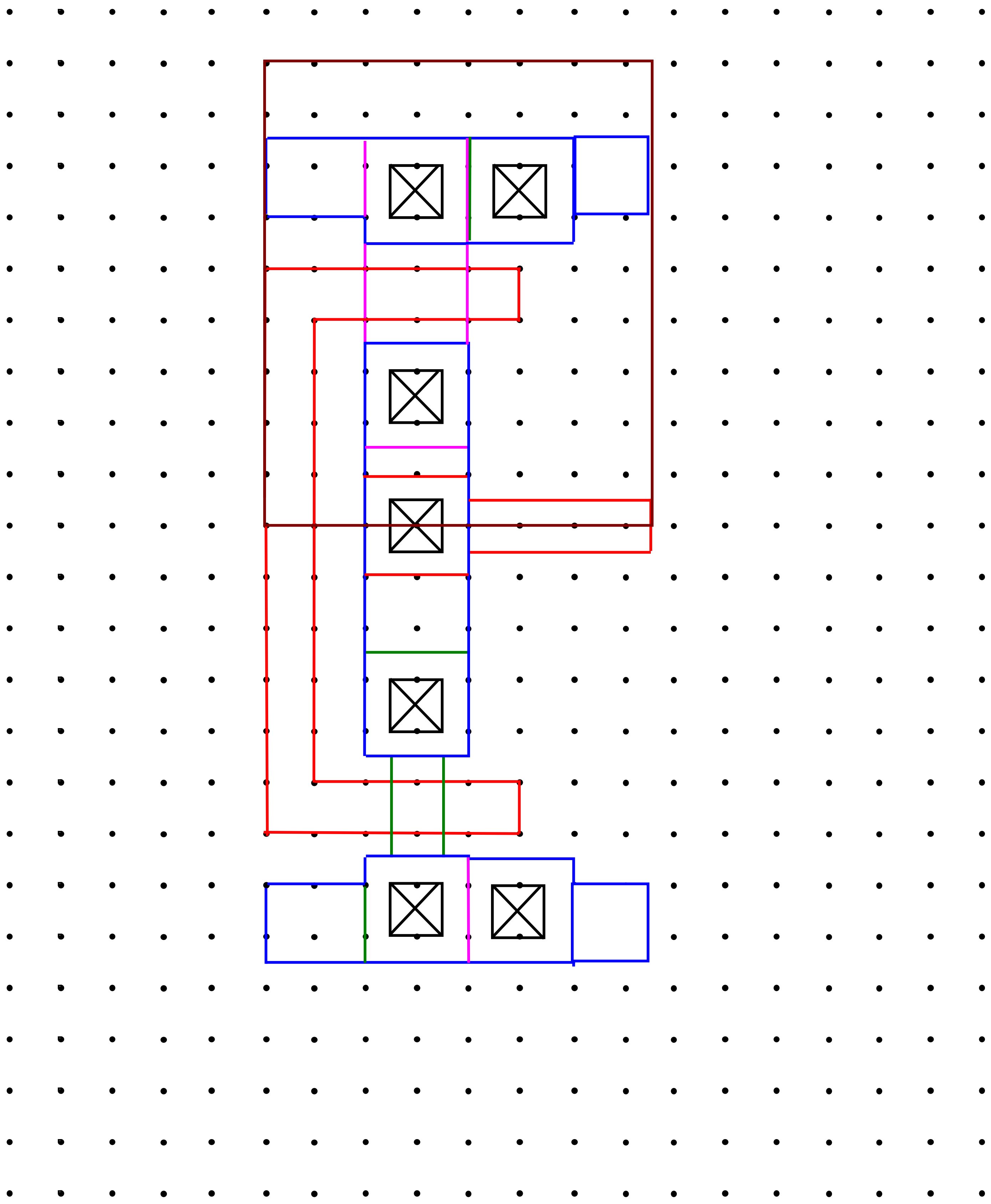


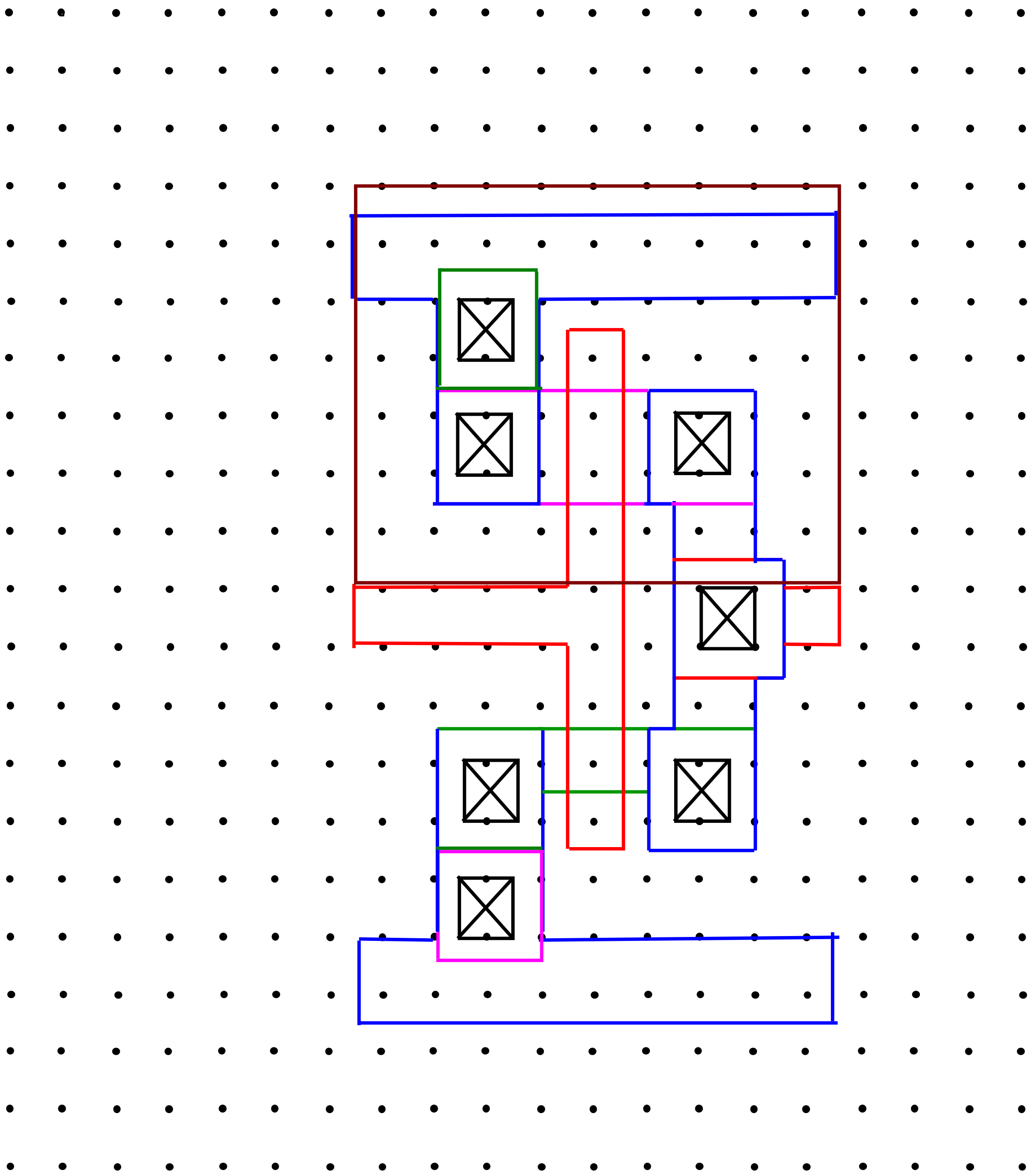


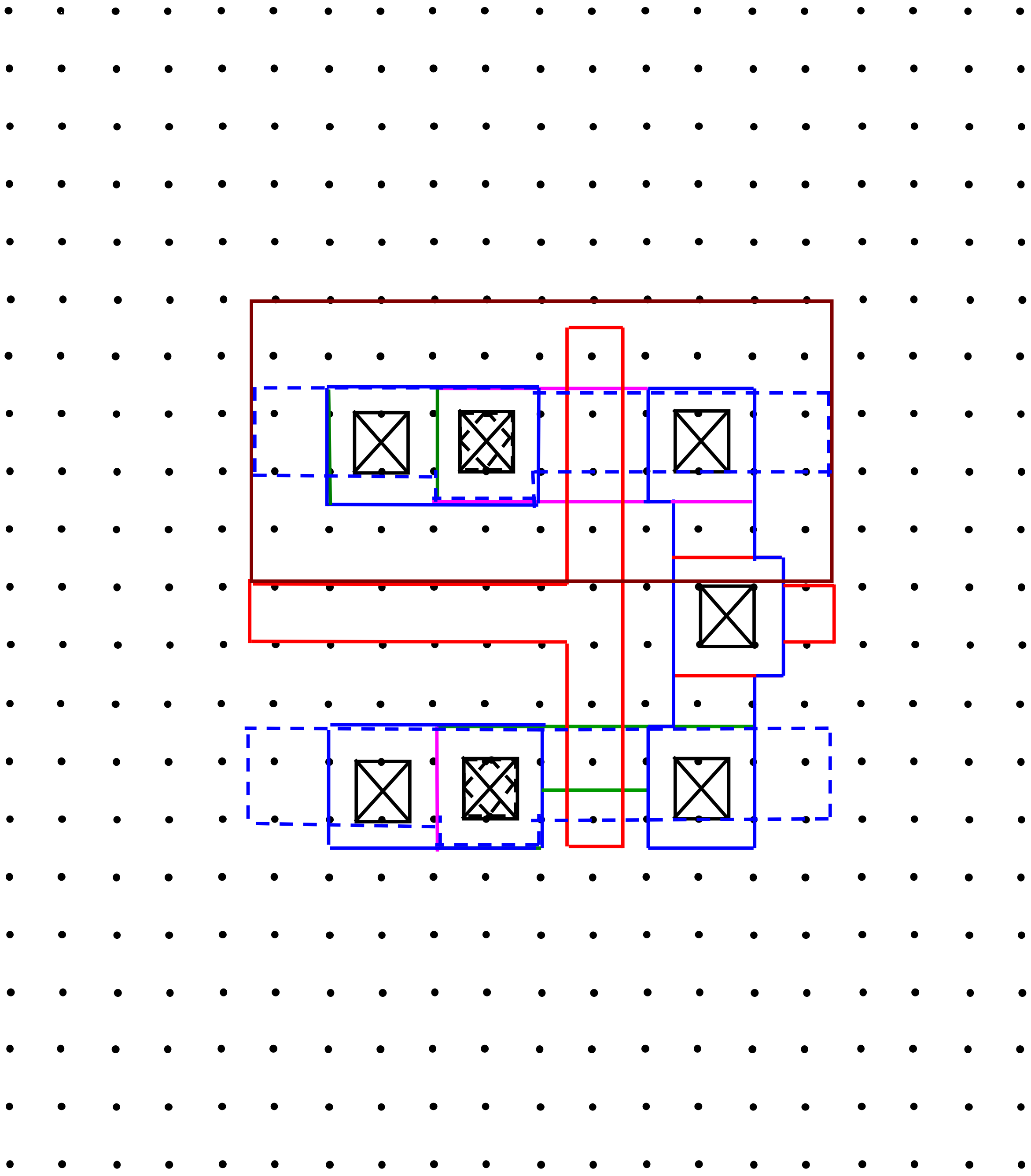
We will do full layout on the right. Try doing yourself on the left.

P. 35 layout separates the substrate and well contacts and wastes area. Such layouts can be compacted by using butting contacts on P. 36.





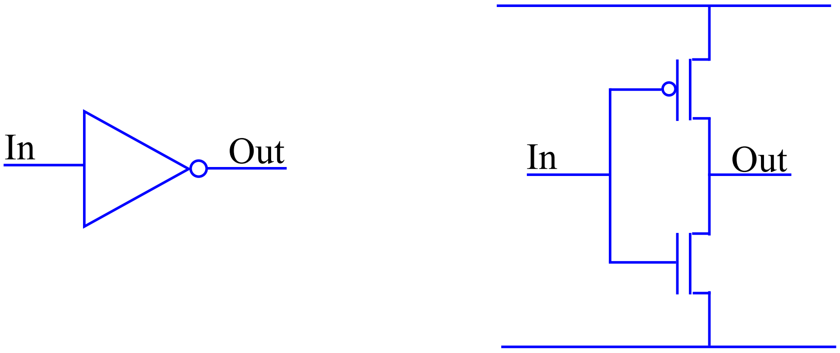




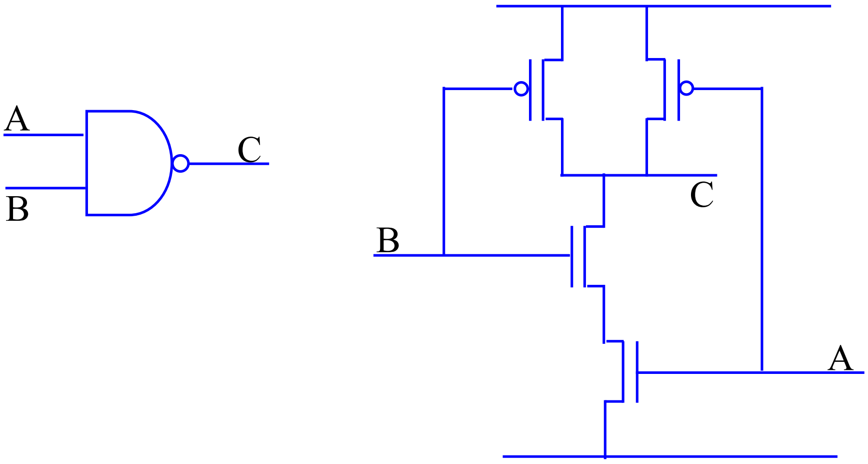
It is noted here that only M2 can only be connected to M1 through a via, and not directly to active/poly region (See illustration).

After discussing inverter and layout representations, we now move to general CMOS logic circuits. The design is carried out by deciding n-type pull down path that sets login zero. P-type devices are complement of n-type network as will be clear when we look at designs. All power and delay aspects apply to these CMOS logic implementations also.

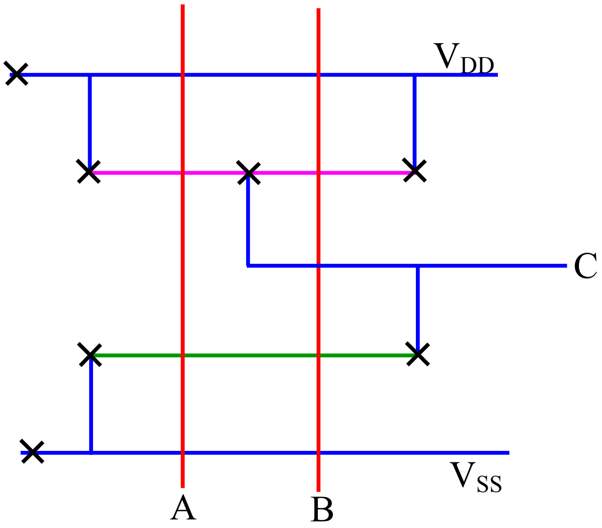
CMOS logic



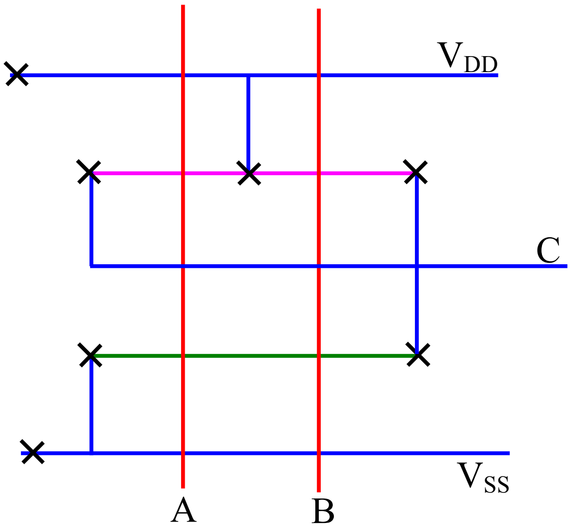
2-input NAND Gate

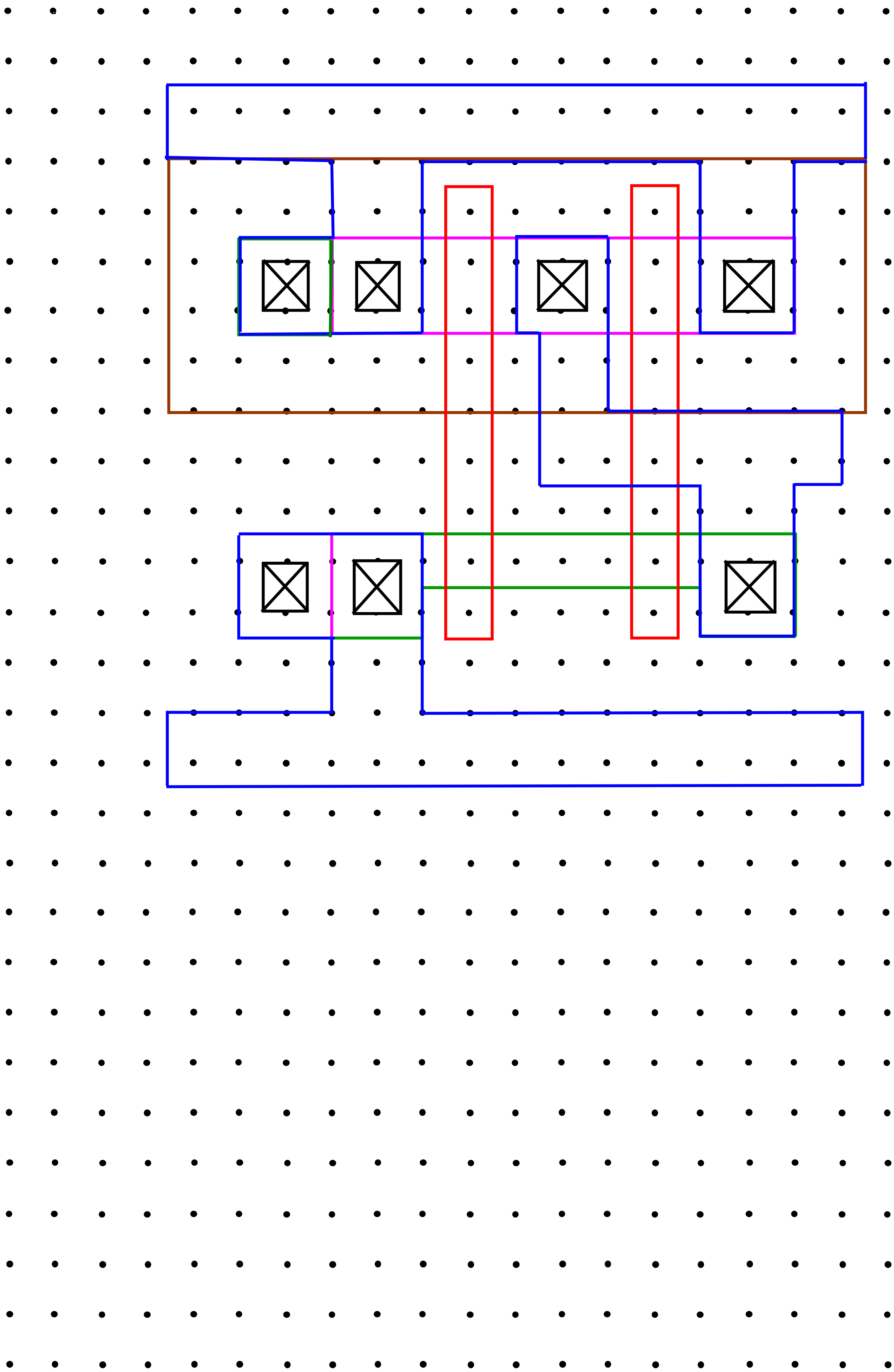


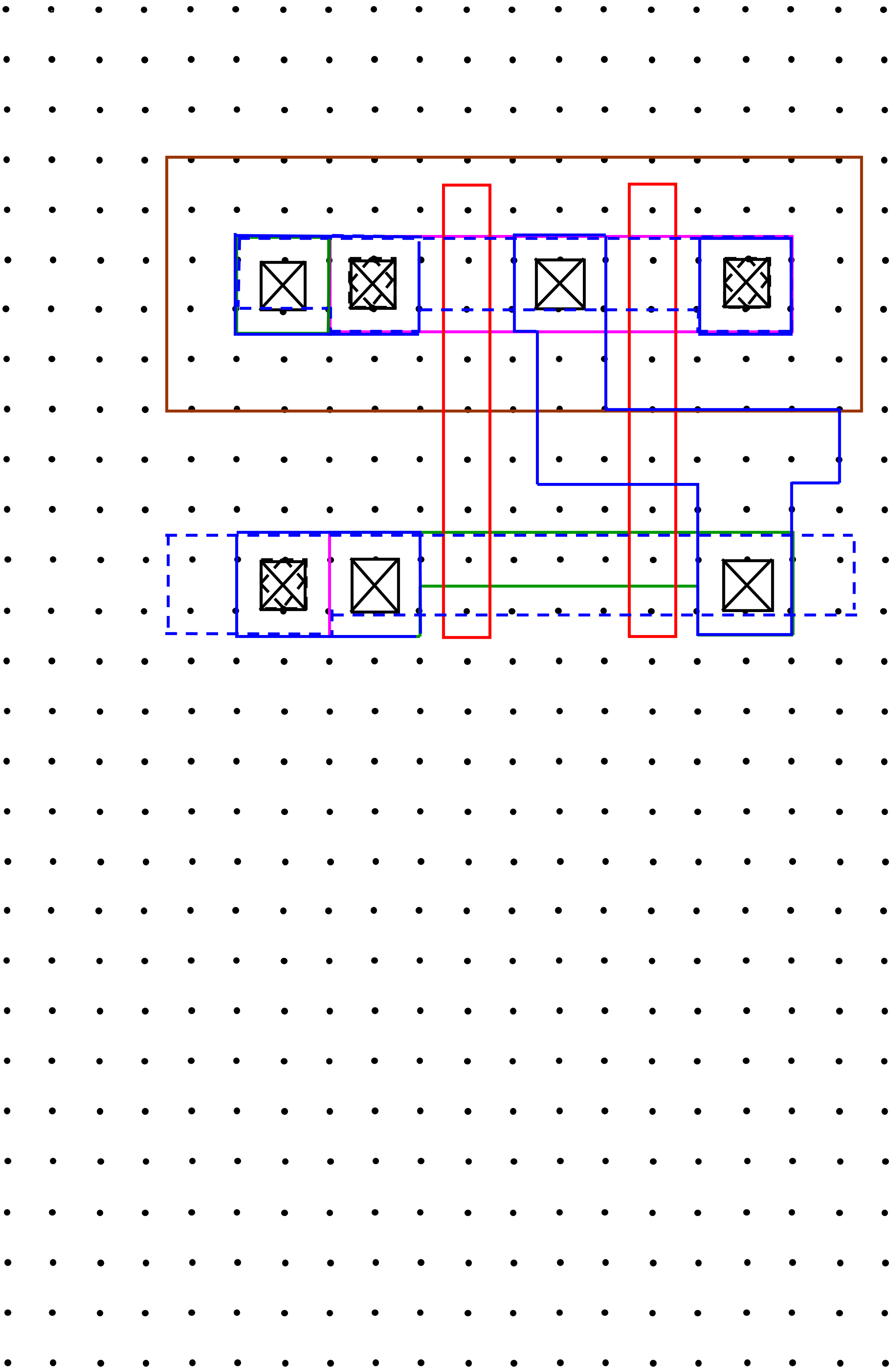
Sizing for timing: If NAND gate above needs to have the same delay as an inverter driving identical load, the n-channel device width has to be doubled as two n-channel devices are in series, effectively double device resistance and the delay. Similar considerations can be applied to other gates. But it is the longest series connected path that determines delay and sizing.



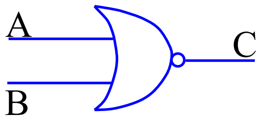
$$C = \overline{A \cdot B}$$

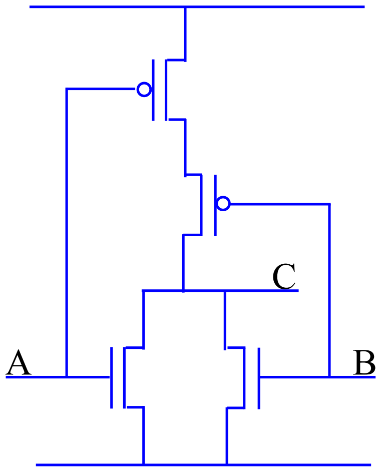


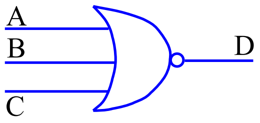


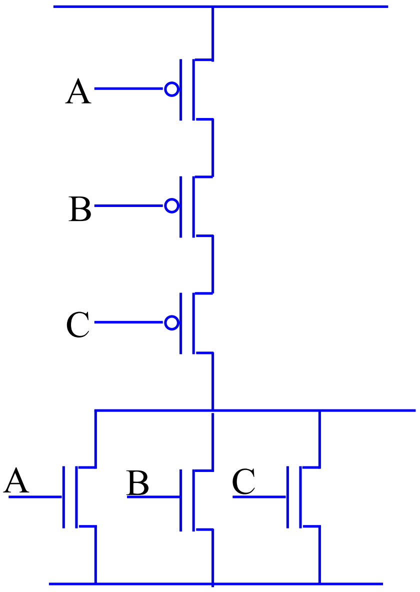


2- and 3- input NOR

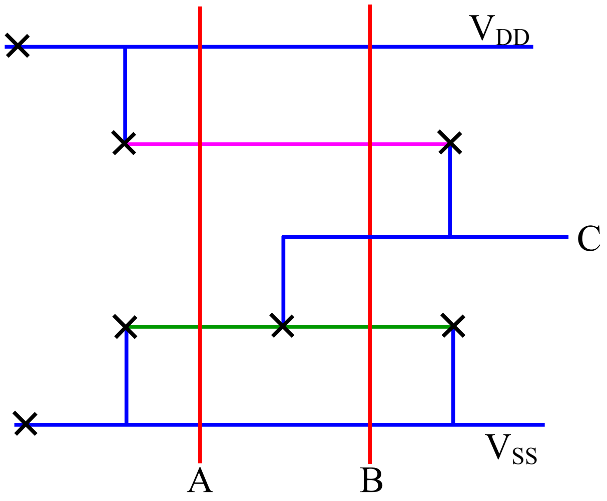




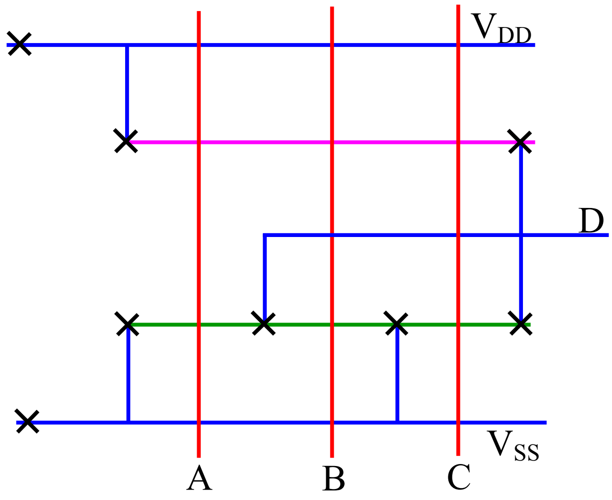




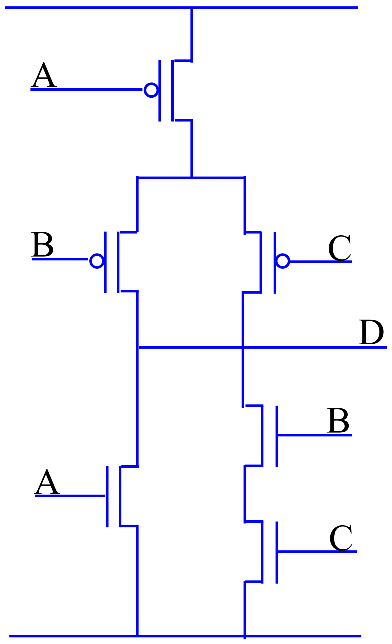
$C = \overline{A + B}$

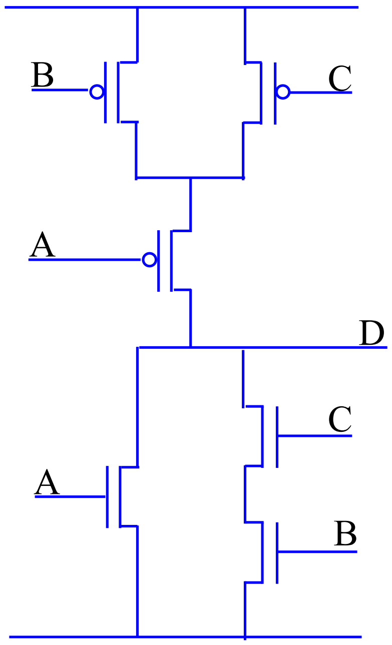


$D = \overline{A + B + C}$

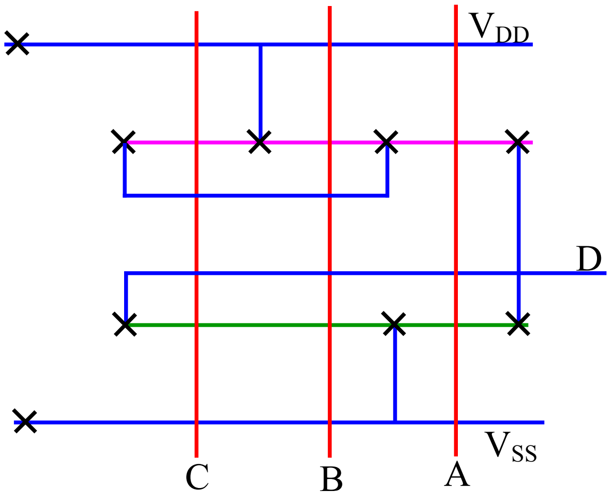


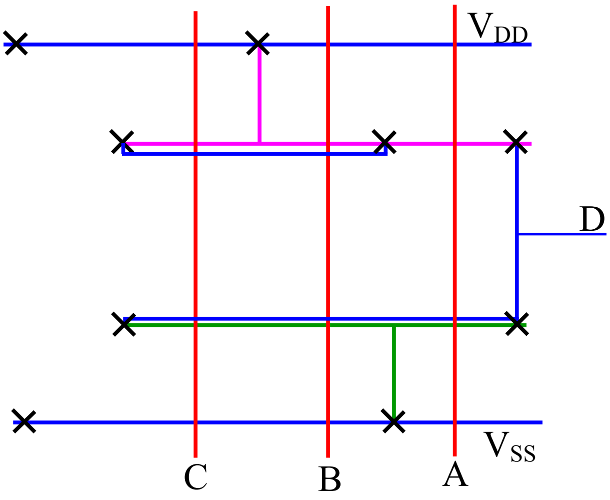
$$D = \overline{A + B \cdot C}$$

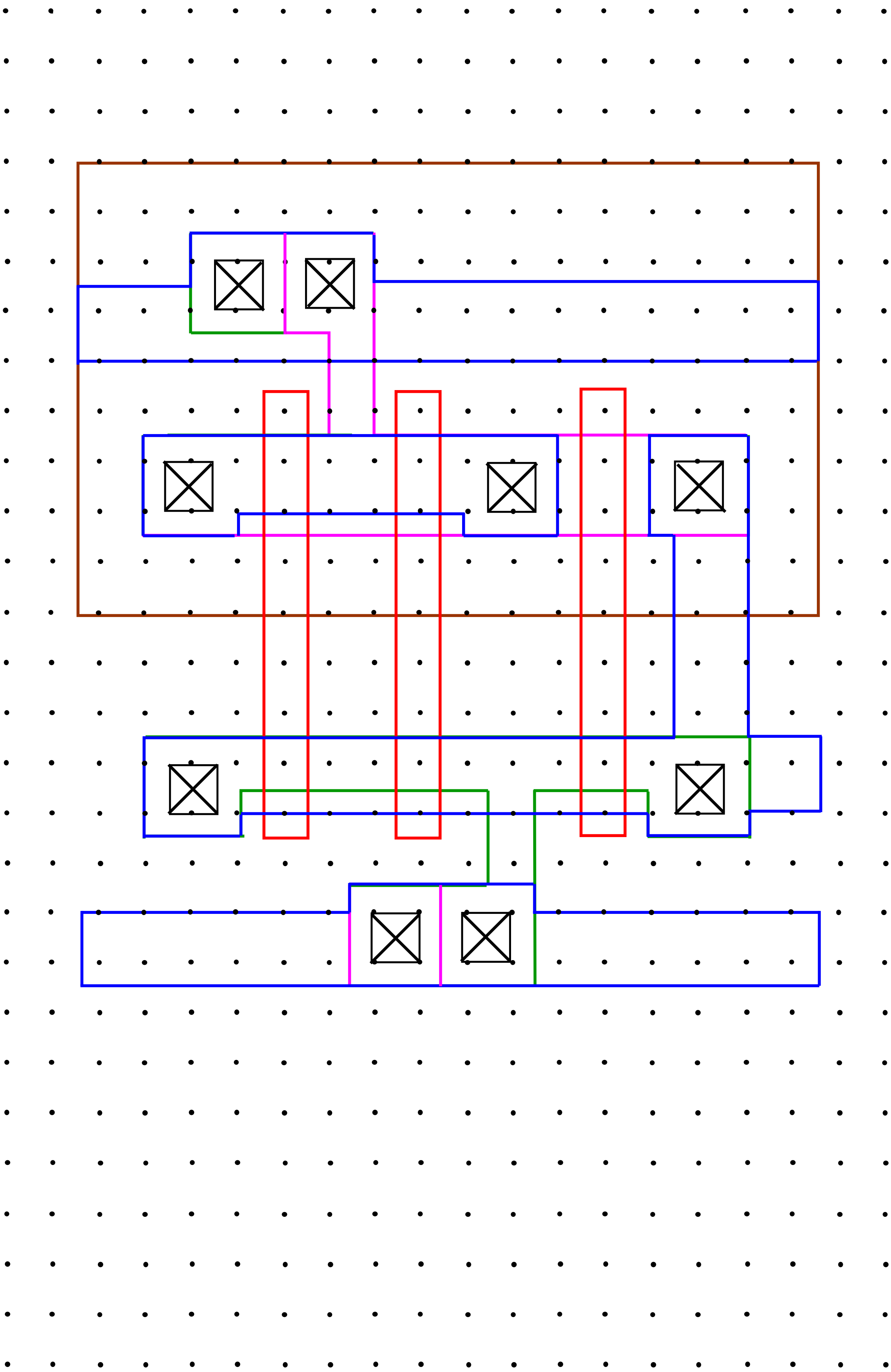


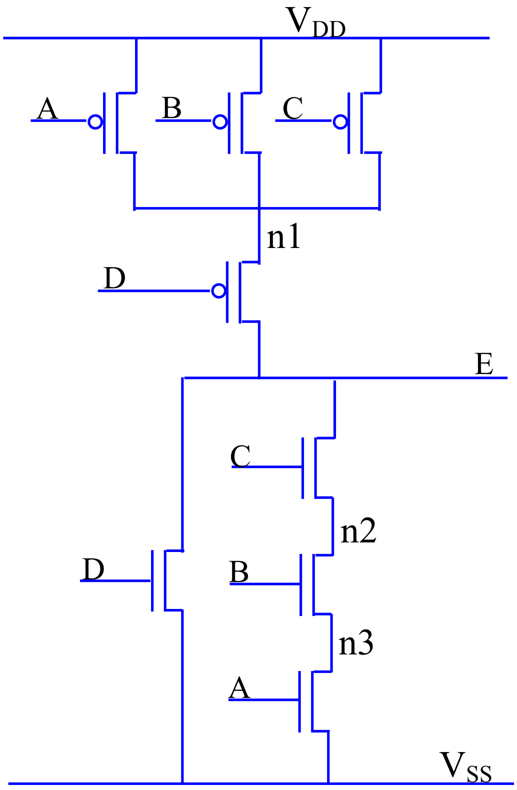


$$D = \overline{A + B \cdot C}$$

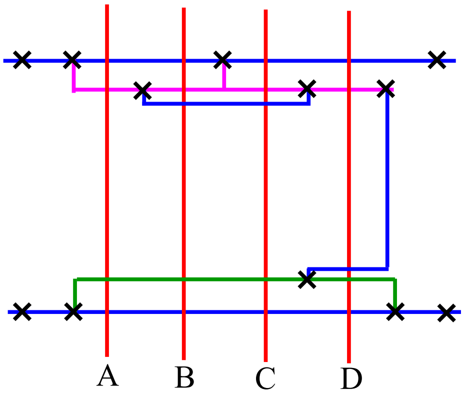








V_DD
A
B
C
n1
D
E
C
n2
D
B
n3
A
V_SS



A
B
C
D

XNOR : $\overline{(A + B) \cdot \overline{AB}}$