The University Of Sheffield.

# Part 6: CMOS

CMOS principles

CMOS scaling

Technology issues

Speed and power dissipation

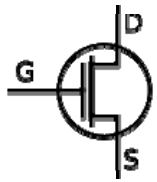
Use of strain

Gate architecture

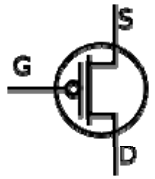
CMOS devices

# CMOS

NMOS



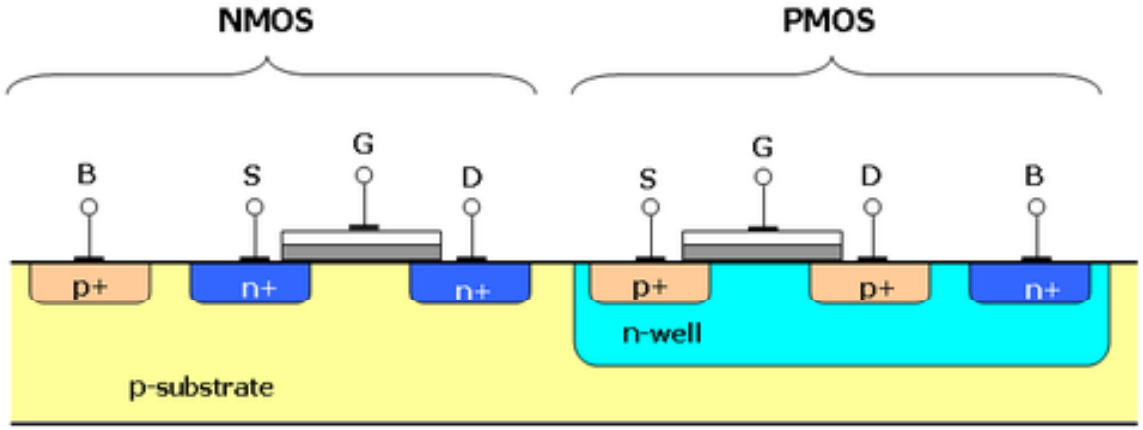

PMOS




***Complementary*** MOS is the basis of modern microprocessor technology

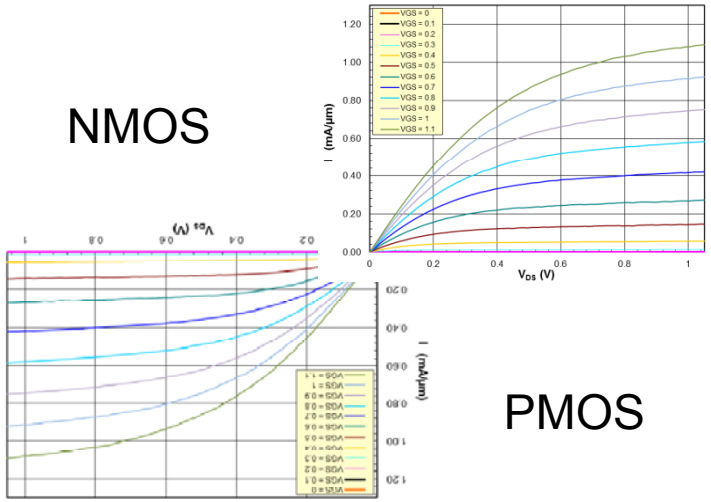
CMOS uses a 'complementary pair of n and p –type *enhancement mode* FETs in which the doping arrangements are reversed.

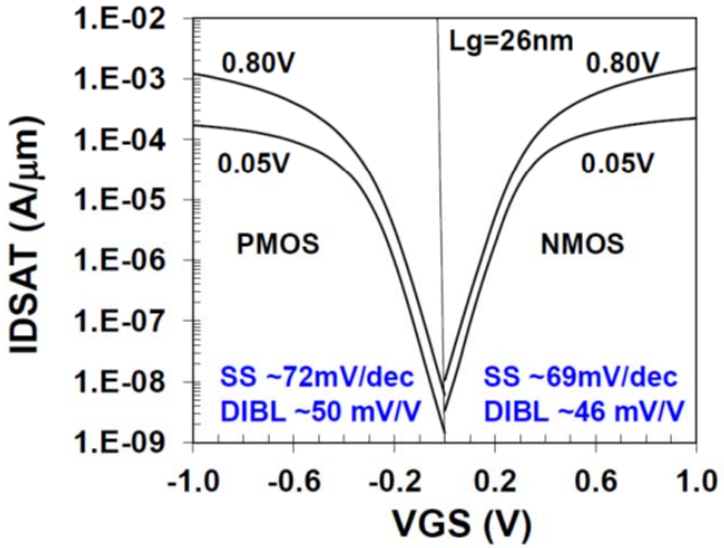
These FETs require the application of a gate voltage to turn on. The direction of the gate voltage is however different for the type types.



# CMOS

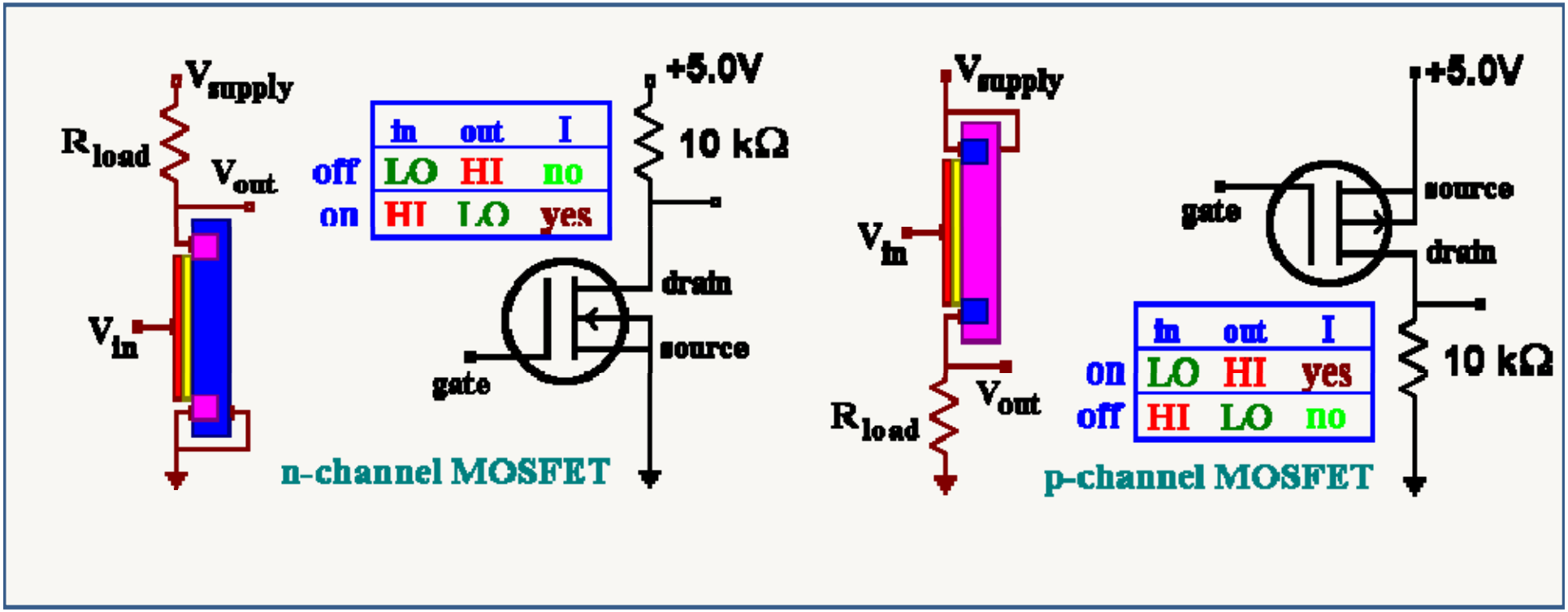
It is important that the two FETs are ‘complementary’ in characteristics. The most important factor is that the threshold voltages are made nearly equal (and of course opposite sign).




Normal to plot the characteristics together as shown on the right - this is rather ideal and often not achievable in reality

# CMOS




During any part of the logic sequence one device will be raised above threshold to make in conduct (on-state) whilst the other is in the off state. Combining two devices together means in principle there is no static power drain at stand-by.

# CMOS

The CMOS ‘pair’ also offers low power for switching as well as low static power dissipation

$$\tau = \frac{C_g \Delta V_{DD}}{I}$$

Switching time is governed by the gate capacitance and supply voltage (typically 2-3x the lowest $V_T$ of the pair)

A further advantage is that CMOS uses a lower transistor count than other approaches (particularly of emitter coupled logic which it replaced)

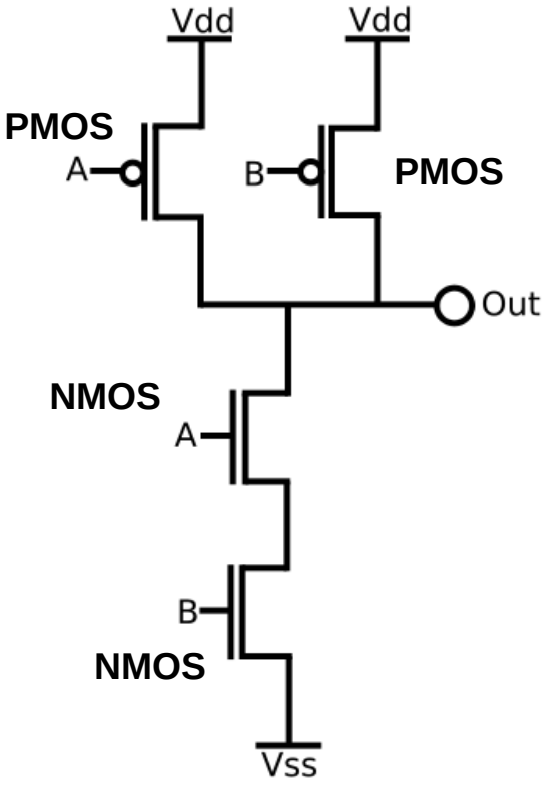
It is also highly integratable with a geometry which naturally allows a closely spaced pair with surface interconnection

# CMOS

## NAND logic gate- basis of modern digital logic

If both of the A and B inputs are high, then both the NMOS transistors will conduct whilst neither of the PMOS transistors will conduct, and a conductive path will be established between the output and ground, giving a zero.

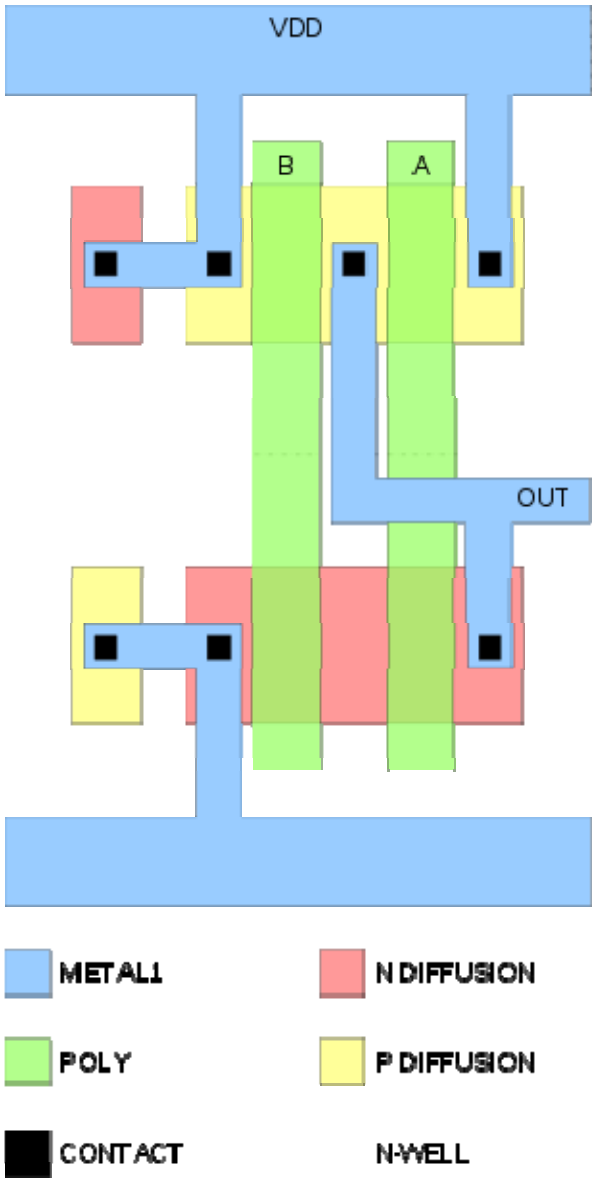
If either of the A or B inputs is low, one of the NMOS transistors will not conduct, but one of the PMOS transistors will and a conductive path will be established between the output and $V_{dd}$ (voltage source), bringing the output high



# CMOS

The result is a conductive path between the output and $V_{dd}$ ; this gives a '1'.

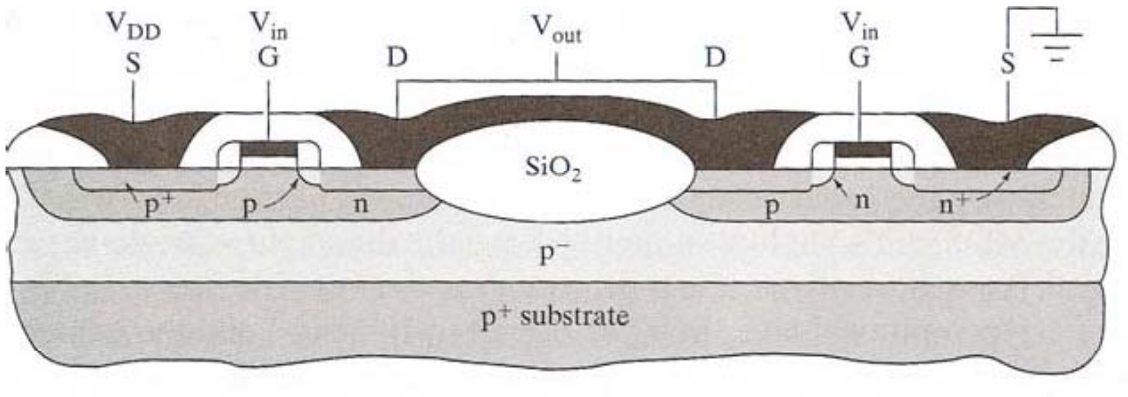
CMOS has fast low-to-high and high-to-low output transitions. This is because the transistors have low resistance when switched on. Also the output signal swings the full voltage between low and high rails, making for a strong symmetric response.

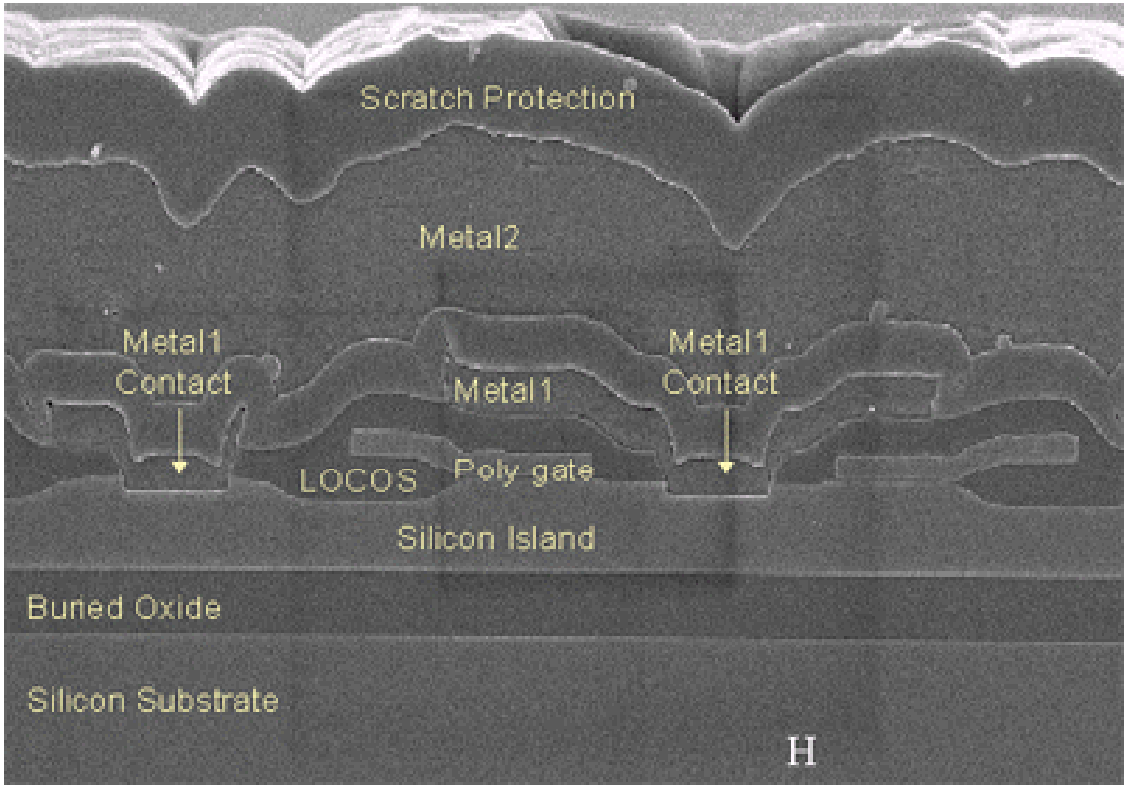



**NAND circuit (plan view)**

# CMOS

CMOS needs to form an n and p channel MOSFETs on the same substrate. Uses ion implantation doping (compensation doping)



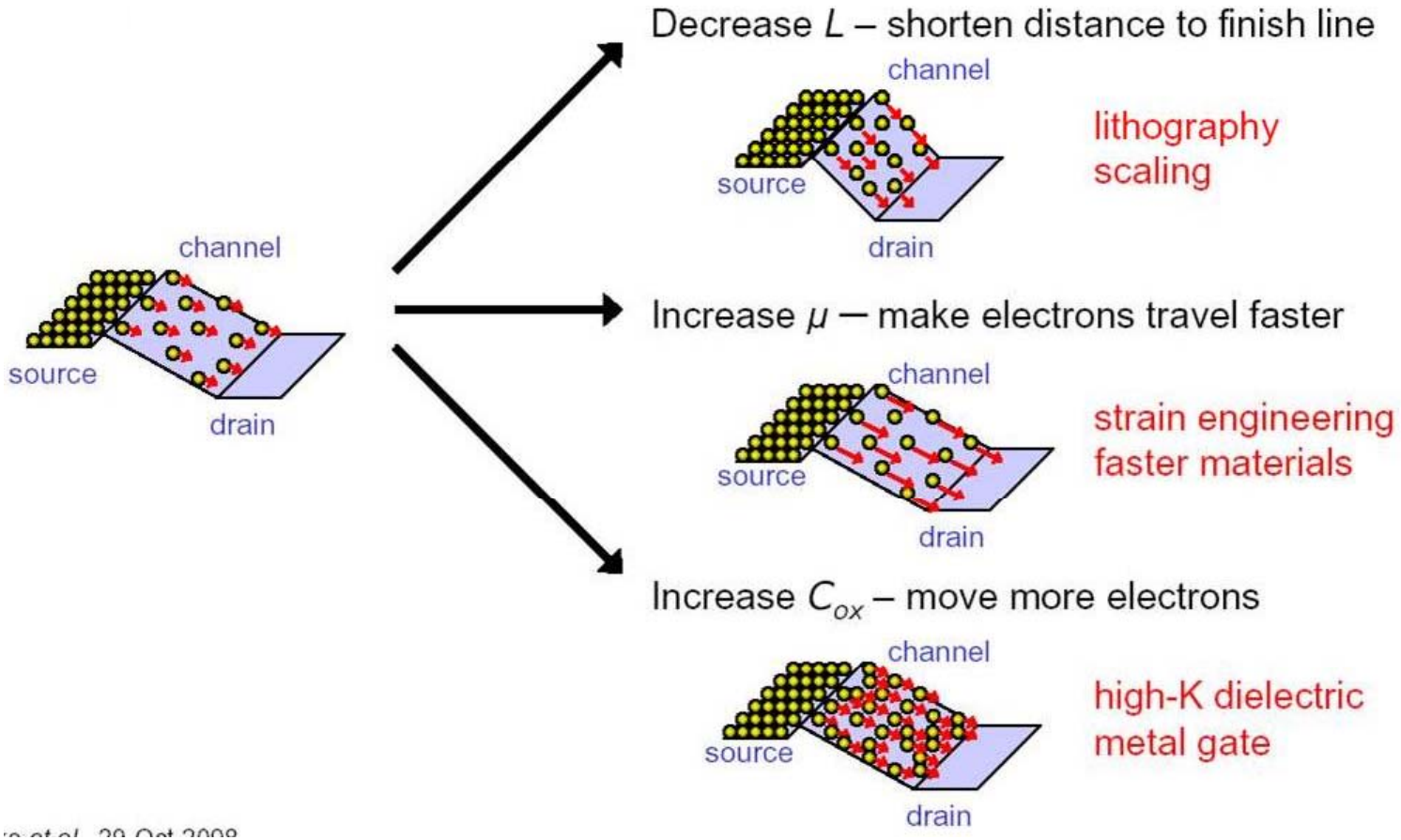
# CMOS

## MOSFET drain current

$$I_D = \frac{Z\mu C_{OX}}{L}\left[(V_{GS} - V_T)V_{DS} - \frac{V_{DS}^2}{2}\right]$$

High performance- would like to have a large $I_D$ for a small $V_{GS}$

→ **Increase $C_{ox}$ and μ, reduce L**



# CMOS

So for high speed we need to reduce L, increase μ and increase $C_{ox}$

**However!**

Reducing L has an affect on $V_T$

$C_{OX}$ depends on the oxide thickness, $t_{OX}$ , which cannot be further reduced without causing excessive gate leakage current

To increase mobility μ, we have to move from simple silicon based devices

**Solutions**

- Decrease L and control $V_T$ (suppress the short-channel effect
- Increasing μ via mobility improvement: SiGe, III-V, Graphene
- Reduce gate leakage via using thicker, but higher-*k* dielectrics

# CMOS

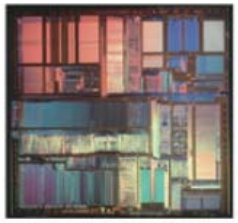
## Scaling in modern CMOS

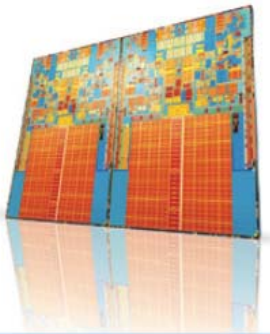
Process of systematic shrinking of device size to make transistors smaller and the circuits more dense

Scaling can apply to the device dimensions such as gate length and gate oxide thickness, basic device characteristics such as $g_m$ and $f_T$ and also operating parameters such as $V_{DD}$ and power consumption

All part of the industry philosophy: to get better performance for the same area. Essentially an economic drive: offer better performance, or same performance at less cost- make more money

# CMOS

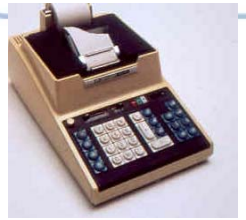




Intel® 4004 processor
Introduced 1971
Initial clock speed
108 KHz
Number of transistors
2,300
Manufacturing technology
10µ

Intel® Pentium® processor
Introduced 1993
Initial clock speed
66 MHz
Number of transistors
3,100,000
Manufacturing technology
0.8µ

Quad-Core Intel® Xeon® processor (Penryn)
Dual-Core Intel® Xeon® processor (Penryn)
Quad-Core Intel® Core™2 Extreme processor (Penryn)
Introduced 2007
Initial clock speed
> 3 GHz
Number of transistors
820,000,000
Manufacturing technology
45nm



92K IPS

250M IPS

76B IPS

**40 years: 220 times smaller- 1 million times faster**

# CMOS

## CMOS Device Scaling Demonstration

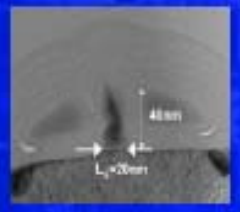
90nm Node
2003

50nm Length
(IEDM2002)

65nm Node
2005

30nm
Prototype
(IEDM2000)

45nm Node
2007



20nm Prototype
(VLSI2001)

32nm Node
2009

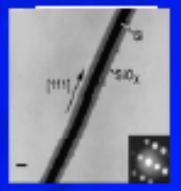

15nm Prototype
(IEDM2001)

22nm Node
2011

10nm Prototype
(DRC 2003)

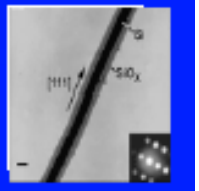
16 nm node
2013

11nm node
2015



TBD

8 nm node
2017



TBD

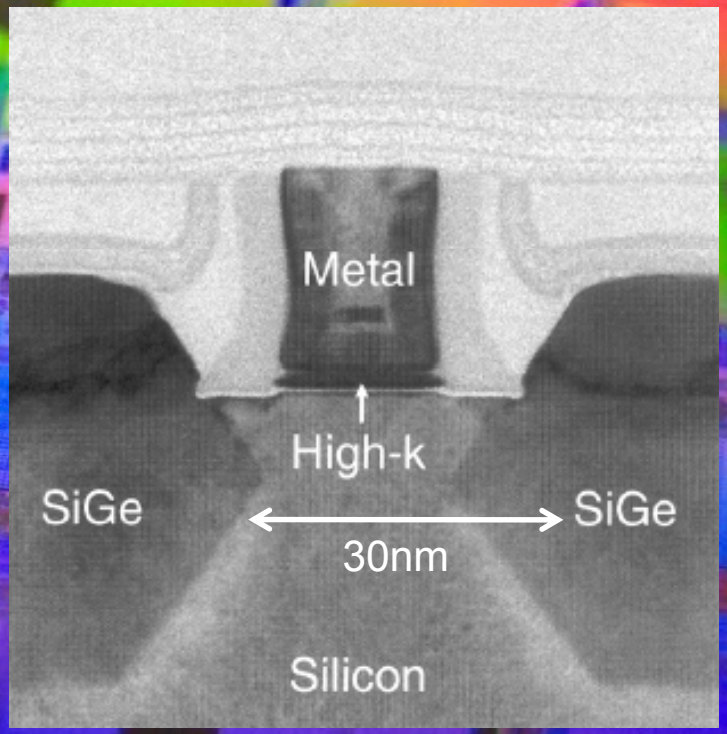
Intel research devices scale to 10nm (16nm node)
Channel engineering solutions (Nanowires/Nanotubes) are being investigated to extend device scaling through end of next decade

intel

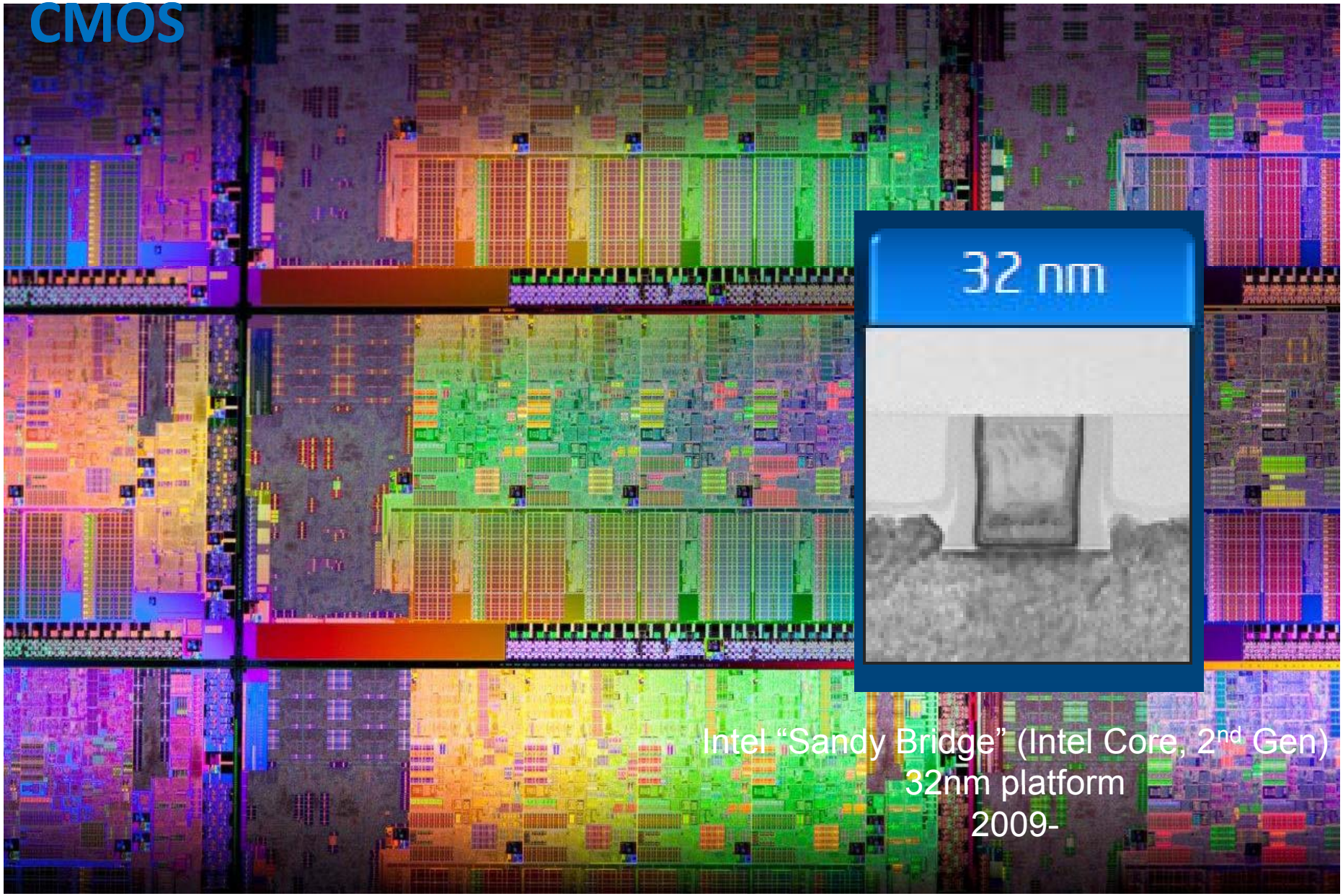
*Source: Intel; Morales and Lieber*
*Science*, **279**, 208, 1998

The University Of Sheffield.

CMOS

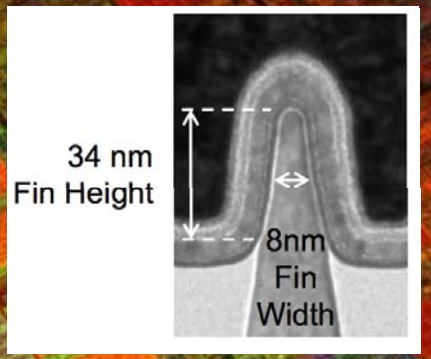


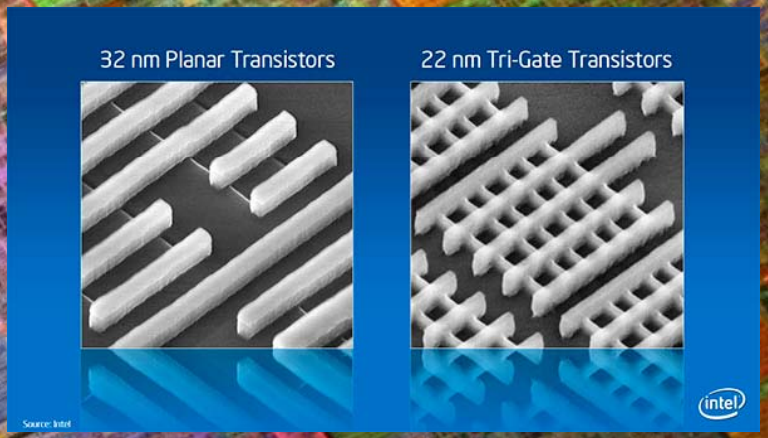

Intel “Santa Rosa” (Core 2 duo)
45nm platform
2007-2010


CMOS
32 nm
Intel "Sandy Bridge" (Intel Core, 2nd Gen)
32nm platform
2009-

The University Of Sheffield.

# CMOS

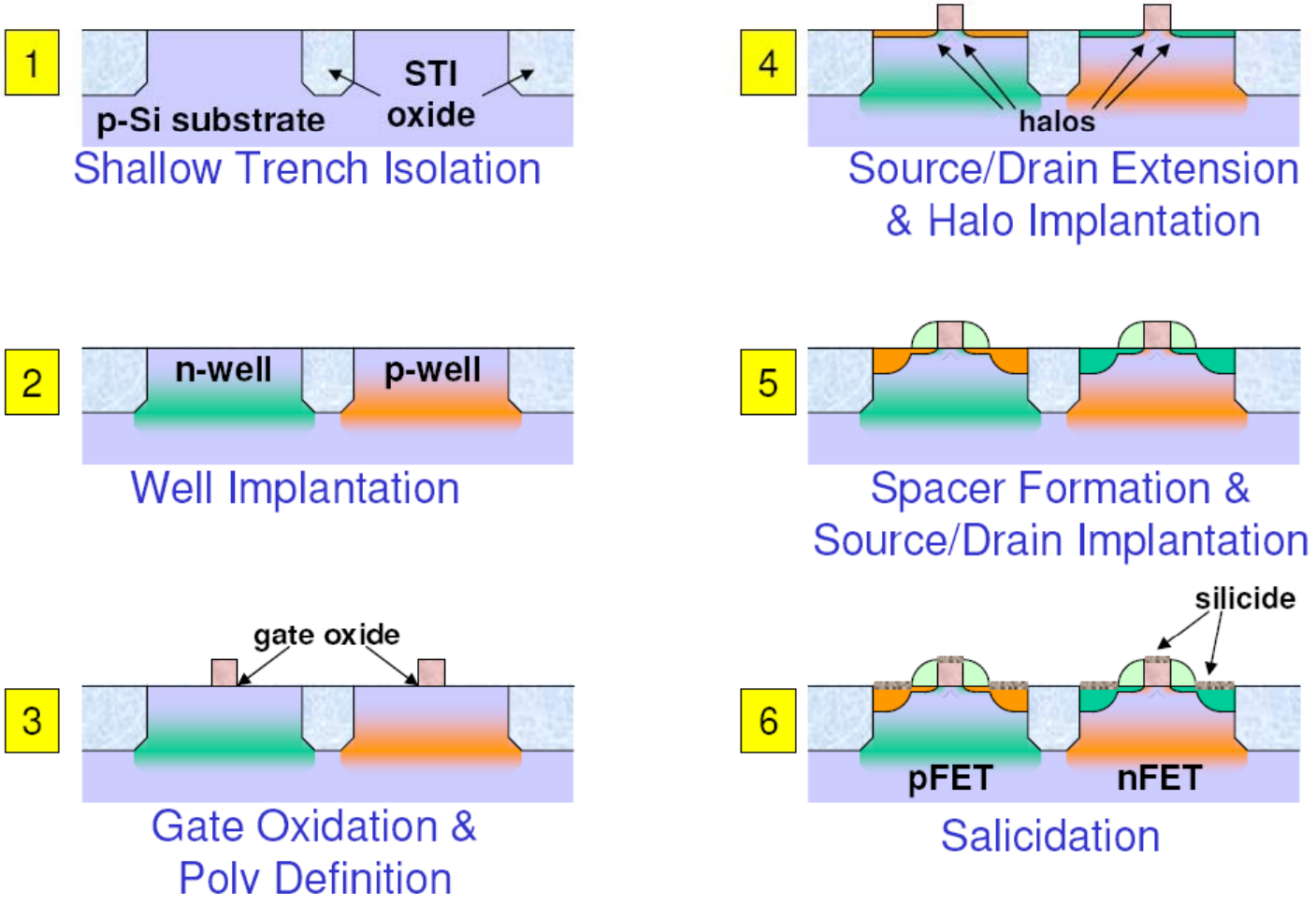



Intel “Haswell” (4th Generation core)
22nm platform
2013-

# CMOS

## Deep Submicron FET Fabrication Sequence




*Reproduced from Avago Technologies*

# CMOS

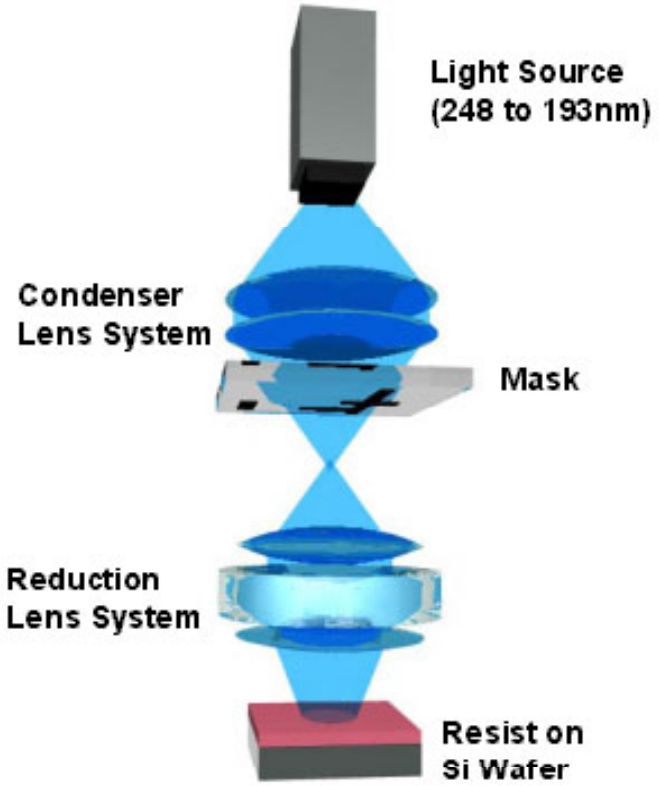
## Lithography

Optical image transfer from an opaque mask to the semiconductor surface

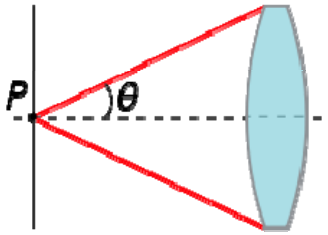
Uses photoresit- a light sensitive polymer which undergoes chemical changes on exposure. The light exposed regions may then be stabilised by means of a developer. These regions may then be dissolved in solvent.

Any optical system has a minimum resolution (*W*) due to diffraction. Two waves separated by λ/4 will destructively interfere.

NA= numerical aperture = n. Sinθ




$$W = \frac{\lambda}{4.NA}$$



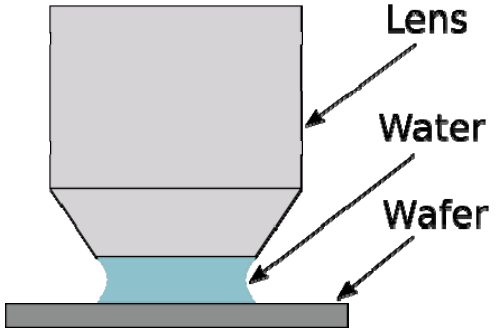
# CMOS

Many advanced lithography systems operate close to the near-field condition ($\theta \rightarrow 90^{o}$). $\sin\theta$ more typically $\approx 0.5$

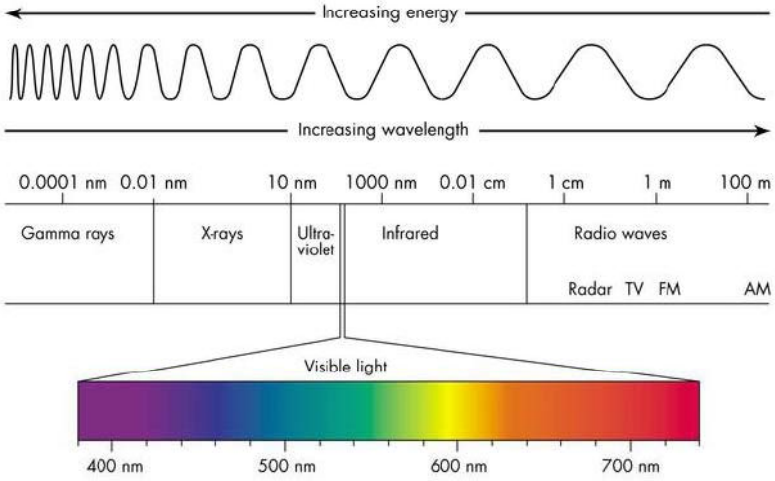
Also can make use of an immersion lens where the optical medium is some fluid (water, oil)

Air: n=1, Water: n=1.33, Silicone Oil: n= 1.45

NA typically ~0.5-0.7




The other way to improve the resolution has been to reduce the wavelength of the light



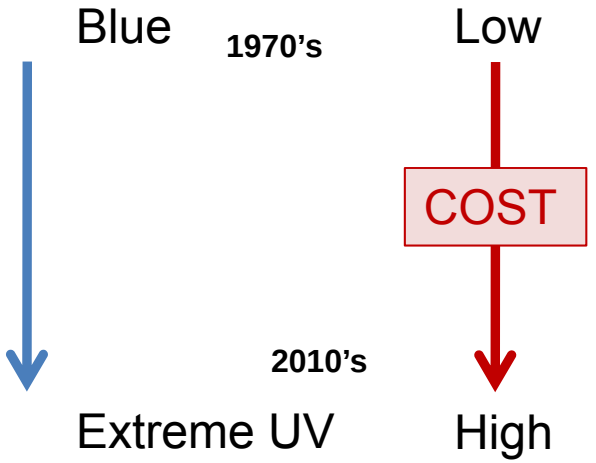
# CMOS

436 nm (G-line of mercury arc lamps)

365nm (I-line of mercury arc lamps)

248 nm (krypton-fluoride excimer laser).

193 nm (Argon-fluoride laser)




Yet even with E-UV there is a resolution issue

$$W = \frac{193x10^{-9}}{4.0.7} = 69nm$$

**How to make 22nm gates? On option is to move from optical lithography**

- e-beam – de Broglie wavelength of an electron (at typical acceleration voltages) is in the 100pm range
- x-rays – wavelengths in the 10's of pm range

# CMOS

E-beam is frequently used in nanolithography, but it is a sequential process in which the beam is made to scan the surface. It is relatively slow for large areas and would be impossibly slow for a whole 12 or 14 inch wafer

X-ray optics are very difficult to construct and as a result x-ray lithography is as yet only a technique for the lab.

For these reasons the industry prefers to use optical lithography, albeit at extreme UV wavelengths. Currently developing 157nm Fluorine excimer laser sources for lithography (but many issues).

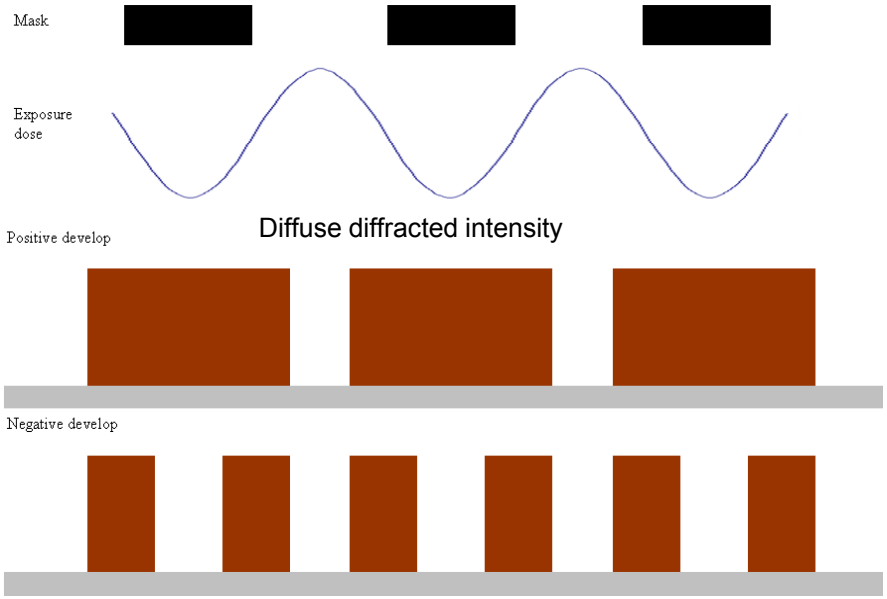
It has also found ways to ‘live’ with the resolution issues.

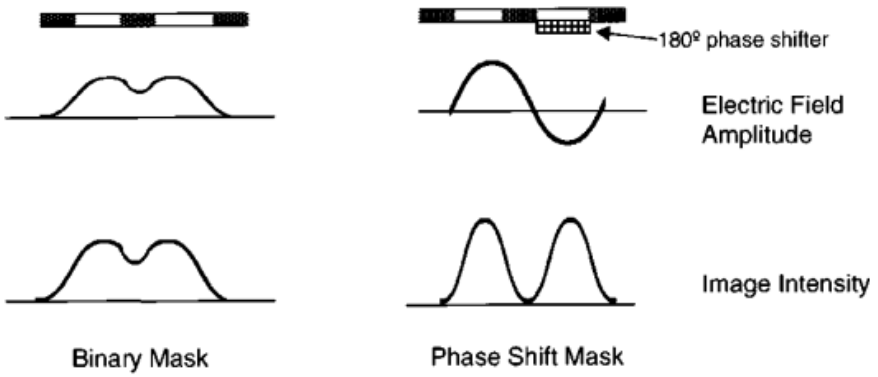
# CMOS

One such ‘trick’ is double patterning

The sample is exposed once to photoresit and processed in a positive develop, then exposed again and a negative develop is performed. As a result the resolution can be halved. This can be performed again and again (triple, quadruple patterning)

Another is to use a special phase-shift mask. This is now quite common in industry



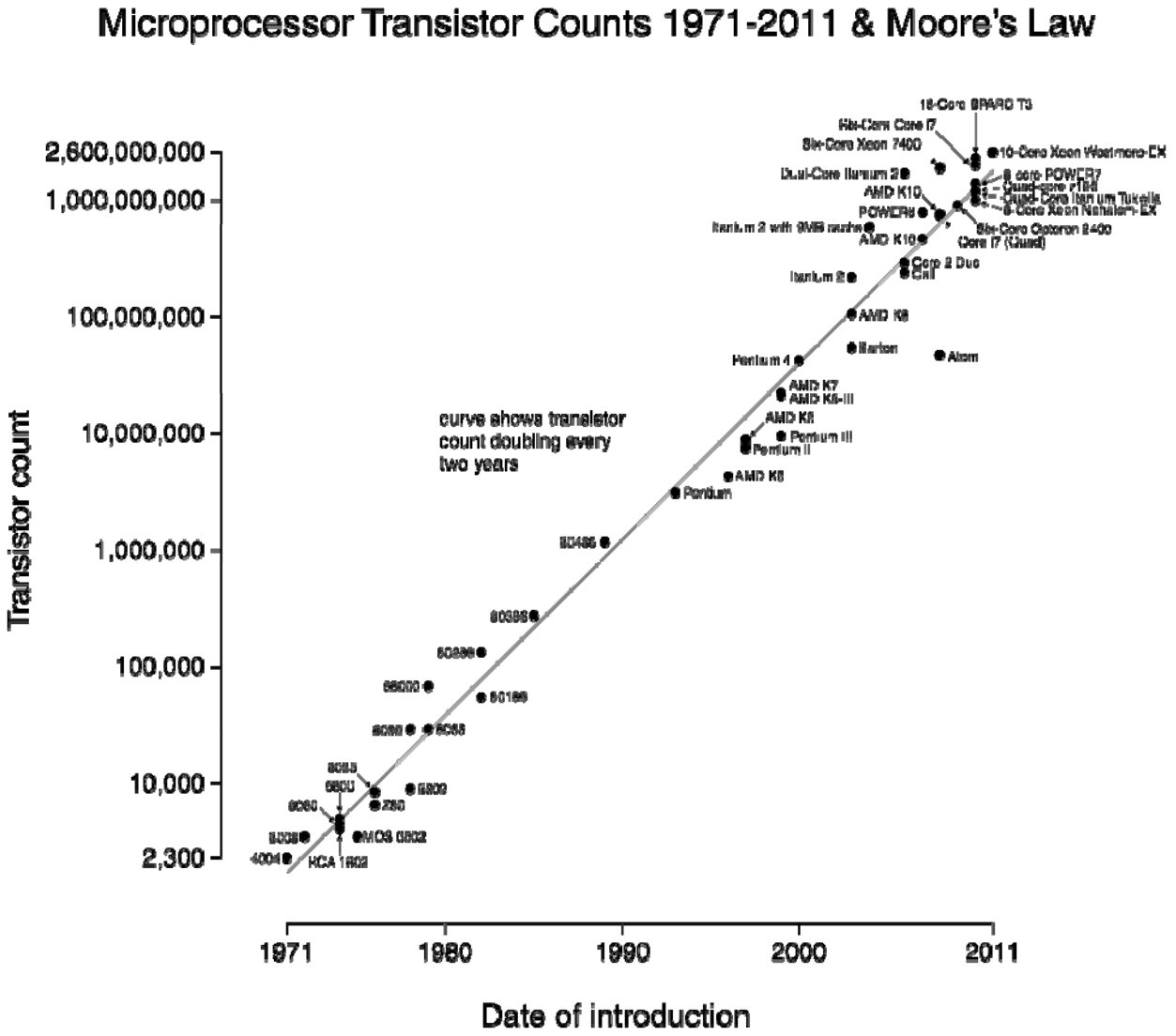



# Si CMOS: Moore's Law

No of transistors per chip doubling every 2 years

We are currently at the 22nm node. Plans are in place for the next 16nm node.

What happens after that?

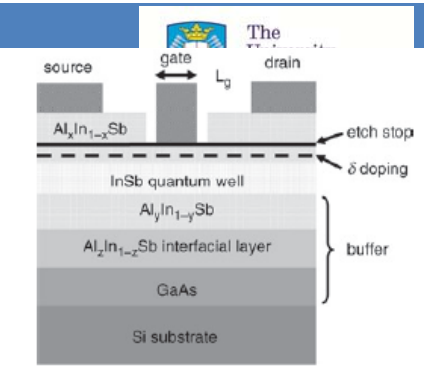


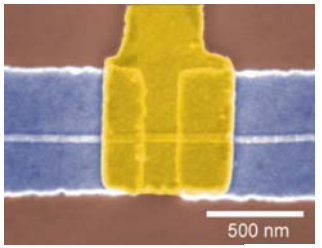
# CMOS

## Moore’s law continues- more Moore

Use different channel materials (e.g: III-V, or graphene) or different geometries e.g: nanowire

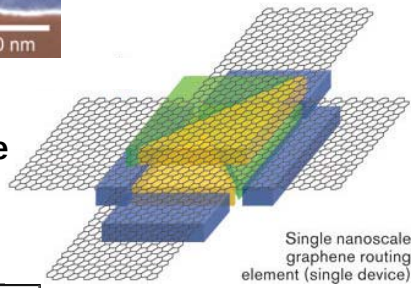
III-V (InSb channel)




Nanowire

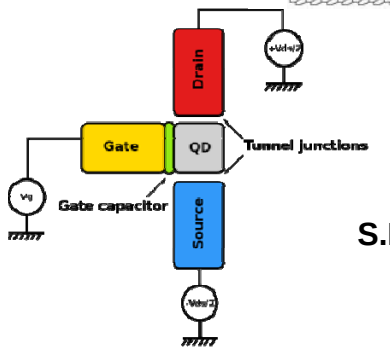
Graphene




## Moore's law ends- more than Moore

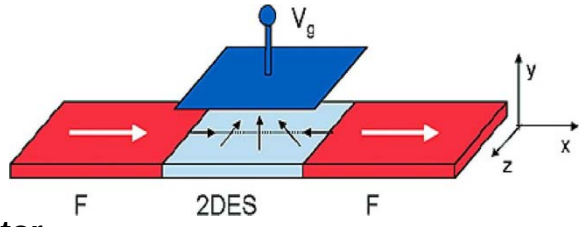
CMOS is replaced by other switching technologies

Single electron transistor, where the gate controls the rate of tunnelling through a quantum dot




S.E.T.

Spin valve, where the gate controls the spin direction of the electron



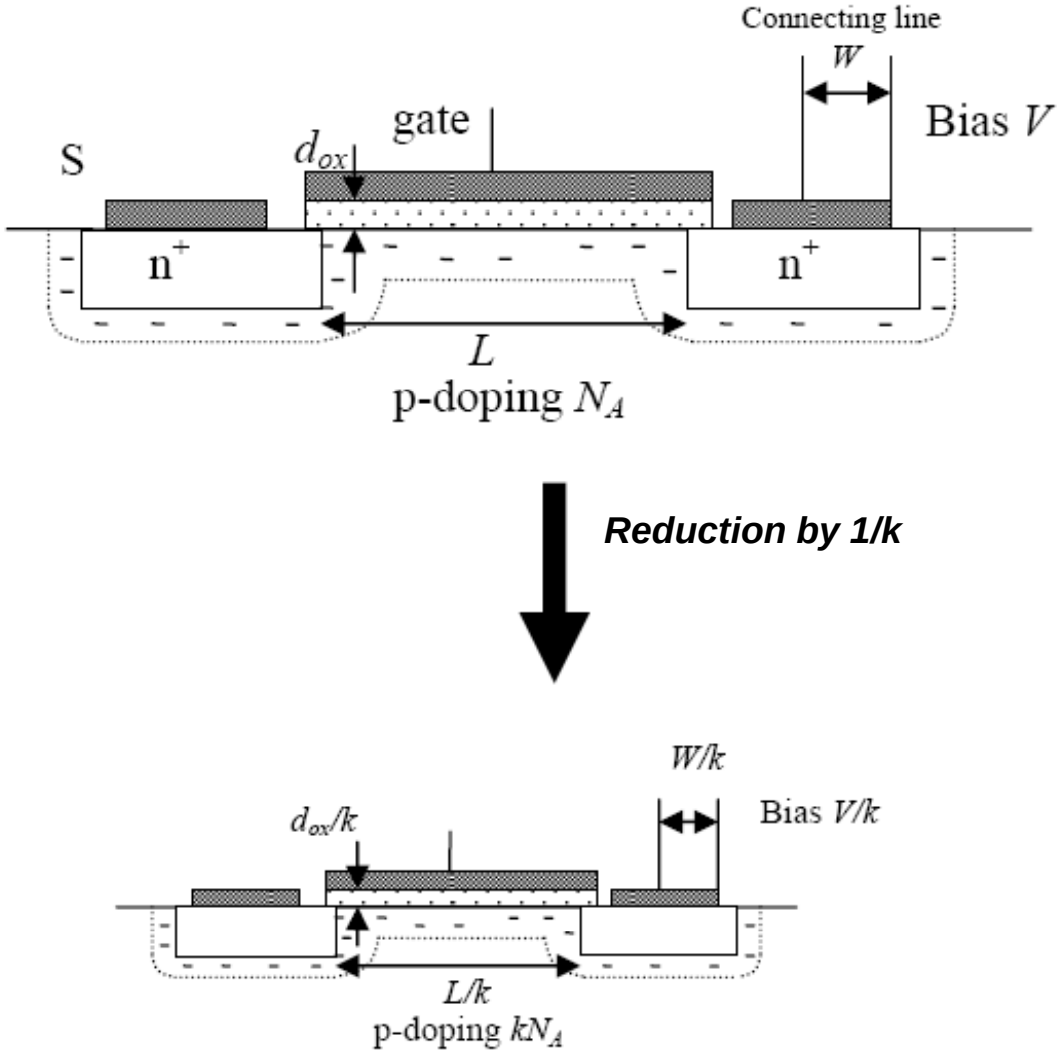

Spin gate transistor

# CMOS

## CMOS scaling

All dimensions scaled down by a factor 1/k (k>1). Also all voltages (constant field)

Scaling the voltage also ensures a constant electric field to avoid breakdown and reduce power dissipation




1) **Total gate capacitance**

$$C_{OX}^{*} \propto \frac{1}{d_{ox}} LZ$$

Scaling: $C^{*}_{OX} \rightarrow C^{*}_{OX}/k$

# CMOS

**2) Gate capacitance per unit area** $C_{OX} \rightarrow C_{OX}k$

**3) Threshold voltage** $V_T = -|V_{FB}| + 2|V_B| + \frac{(2q\varepsilon_S N_A |2V_B|)^{1/2}}{C_{OX}}$

**4) Drain current** $I_D \propto C_{OX}\frac{Z}{L}V^2 \rightarrow C_{OX}k\frac{Z/k}{L/k}.\frac{V^2}{k^2} \rightarrow \frac{I_D}{k}$

**5) Switch delay** $\propto L \propto \frac{1}{k}$ **6) Power dissipation** $\propto I_D V \propto \frac{1}{k^2}$

**7) Power-delay product (figure of merit)** $\propto \frac{1}{k^2}.\frac{1}{k} \propto \frac{1}{k^3}$

| **Parameter** | $Z, L, d_{ox}$ | $V_{DS}, V_T$ | $C^*_{OX}$ | $C_{OX}$ | $I_D$ | DC power | Switch time | Power-delay product |
|---|---|---|---|---|---|---|---|---|
| **Scaling** | $1/k$ | $1/k$ | $1/k$ | $k$ | $1/k$ | $1/k^2$ | $1/k$ | $1/k^3$ |

# CMOS

## Scaling Impact on MOSFET Performance Parameters

$$g_m = -\frac{Z\mu C_{OX}}{L}\left[V_{GS} - V_T\right] \propto C_{OX}\frac{Z}{L}V \rightarrow kC_{OX}\frac{Z/k}{L/k}V/k \rightarrow g_m$$

So $g_m$ maintained despite increase in $C_{OX}$ -gate capacitance per unit area due to thinner oxide.

$$f_T = \frac{\mu}{2\pi L^2}(V_{GS} - V_T) \propto \frac{V}{L^2} \rightarrow \frac{V/k}{L^2/k^2} \rightarrow kf_T$$

so. $f_T$ improved by a factor *k*

If scaling of V is not maintained (gets more difficult for smaller devices) then $f_T$ will increase at a even faster rate ($k^2 f_T$).

However, the velocity tends to saturate

$$f_T = \frac{v_{sat}}{2\pi L} \rightarrow kf_T$$

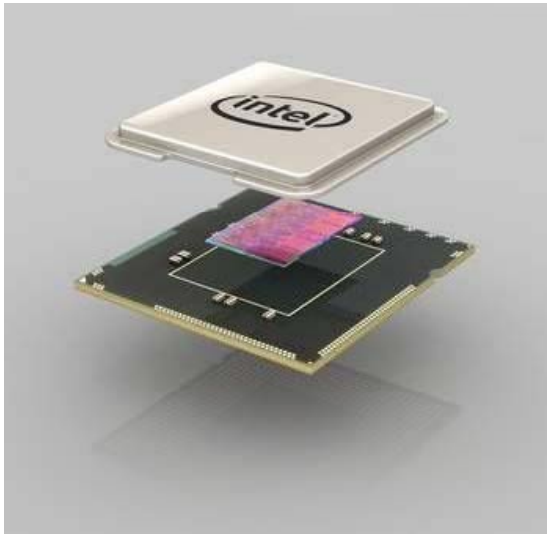
Reassuringly- same result

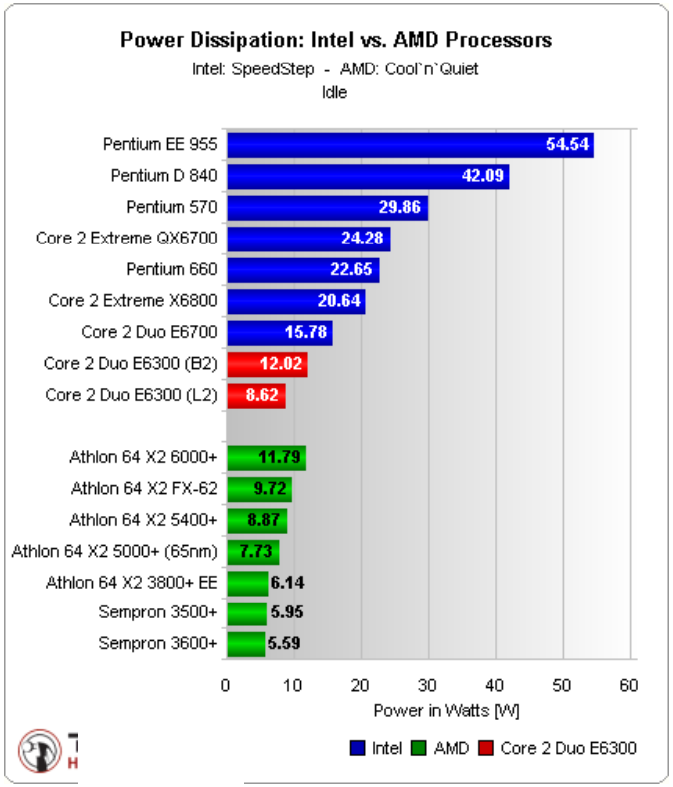
# CMOS

## Power considerations

At low frequency the power dissipation ($I_D V_D$) scales with $1/k^2$.

Power dissipation is a serious problem in microprocessors so this $1/k^2$ reduction is very useful. However, since the chip area also scales with $1/k^2$, the power density remains constant. Harder to dissipate heat from a smaller devices: requires good radial heat conduction



# CMOS

## Power considerations

A well as MOSFET static power dissipation, which comes primarily from gate leakage we also have **dynamic power dissipation** during the switching cycle

$$P_D = \alpha C V_{DD}^2 f$$

α - switching probability (not all gates working at the same time)

C – total capacitance

$V_{DD}$ – voltage swing. f – clock frequency (latest processors 3-4GHz)

From the previous arguments the total power should scale as $1/k^3$ and the power density by 1/k. However, at very small dimensions the voltage stops scaling causing power density to increase.

Also if we increase the operating frequency we increase the power and power density

# CMOS

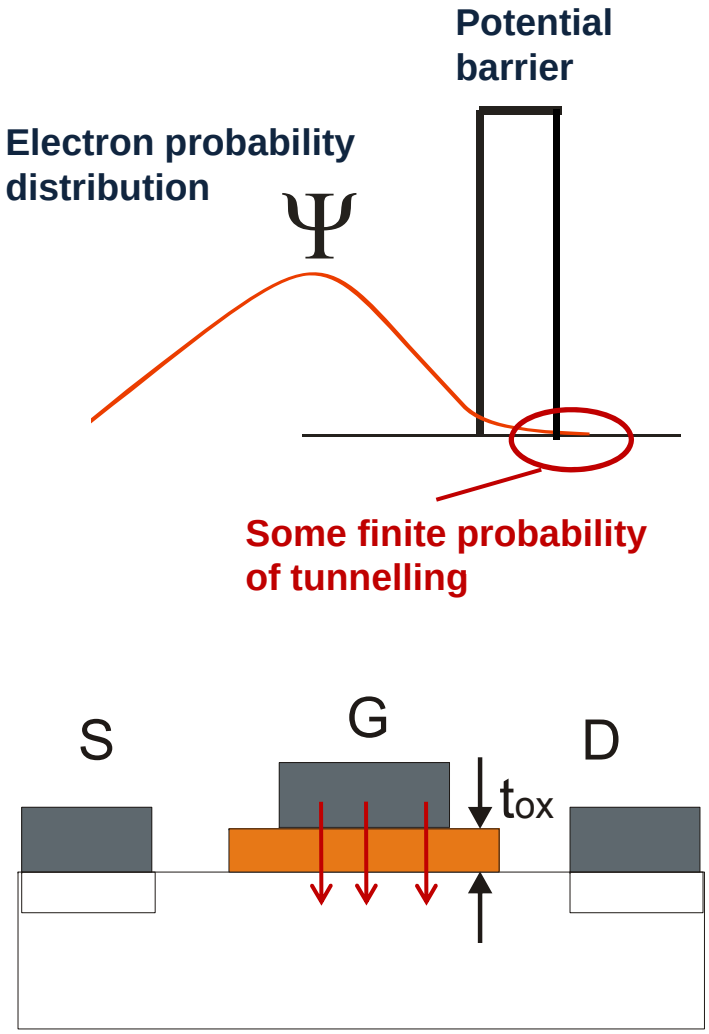
## Limits to Scaling

Reduce all dimensions by 1/k including the oxide thickness

As this thickness reduces below a few nm then electron tunnelling can occur

The tunnelling of electrons from the gate to the channel causes gate leakage which leads to excess power dissipation. This is a quantum mechanical effect due to the small dimensions.

The magnitude of this current depends exponentially on the oxide thickness and has recently become a problem as gate oxide thicknesses are reducing towards just a few atomic layers thick.



# CMOS

## Approaches to circumvent the tunnel limit

- Accept thicker oxide and continue scaling and design out effects of thicker oxide (e.g dual gate FETs )

- Use an insulator with higher dielectric constant to achieve the same capacitance at a greater distance. Such dielectrics are known as a high-k (*k* ≡ $\varepsilon_r$ *relative dielectric constant)*

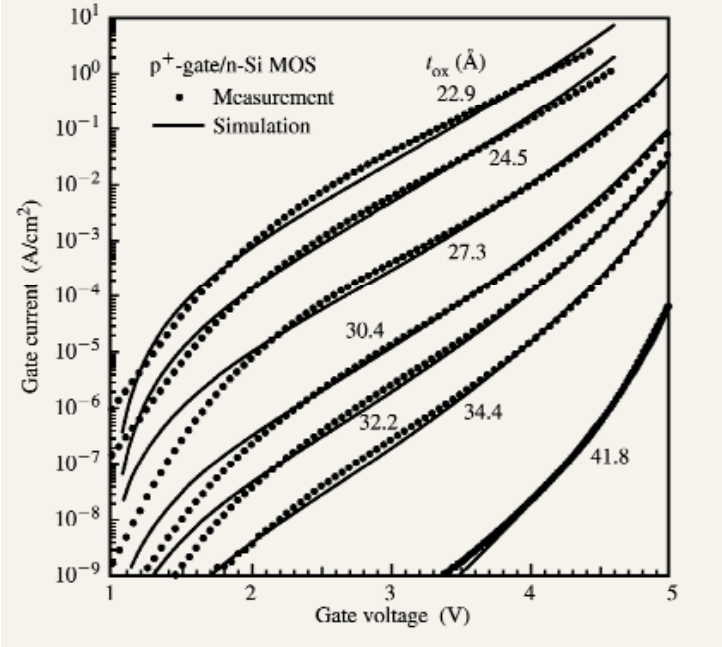
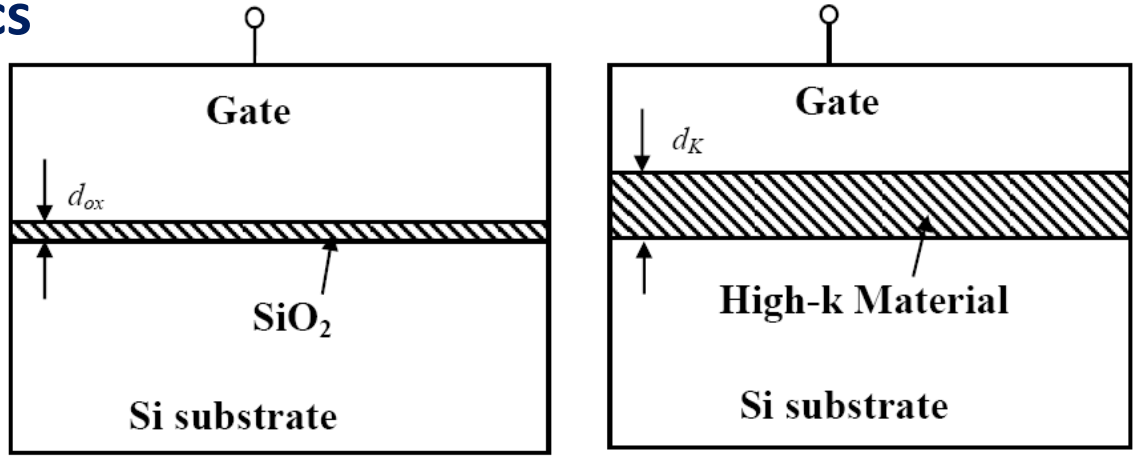



**Figure 14**

Measured and simulated $I_G$–$V_G$ characteristics under accumulation conditions of p+-gate/n-Si MOS devices with oxides ranging from 22.9 to 41.8 Å. The thickness is determined using the QM scheme.

# CMOS

## High –k dielectrics




From the 45nm node onwards a a thicker layer of high –k hafnium oxide is used to reduce tunnelling. $HfO_2$ has 5x the $\varepsilon_r$ of $SiO_2$

Further ongoing research to find even higher k-materials

| *Dielectric* | *Dielectric constant (bulk)* |
|---|---|
| Silicon dioxide ($SiO_2$) | 3.9 |
| Silicon nitride ($Si_3N_4$) | 7 |
| Aluminum oxide ($Al_2O_3$) | ~10 |
| Tantulum pentoxide ($Ta_2O_5$) | 25 |
| Lanthanum oxide ($La_2O_3$) | ~21 |
| Gadolinium oxide ($Gd_2O_3$) | ~12 |
| Yttrium oxide ($Y_2O_3$) | ~15 |
| Hafnium oxide ($HfO_2$) | ~20 |
| Zirconium oxide ($ZrO_2$) | ~23 |

# CMOS

CMOS device is in the ON state when $V_G > V_T$. We have already said that $V_T$ should scale as 1/k.

In practice, however, the dimensions have reduced much faster than the operating voltage. **What is the problem?**

**Operating Voltage**

($V_{DD}$-$V_T$) must be large enough to provide sufficient $I_D$ for acceptable performance .

- For low 'off-state' currents $V_T \geq 120$ mV hence voltage scaling cannot be maintained

- Logic swing must be ≥ 4kT (100 mV) to maintain 2 distinct logic states, otherwise can get errors due to thermal noise

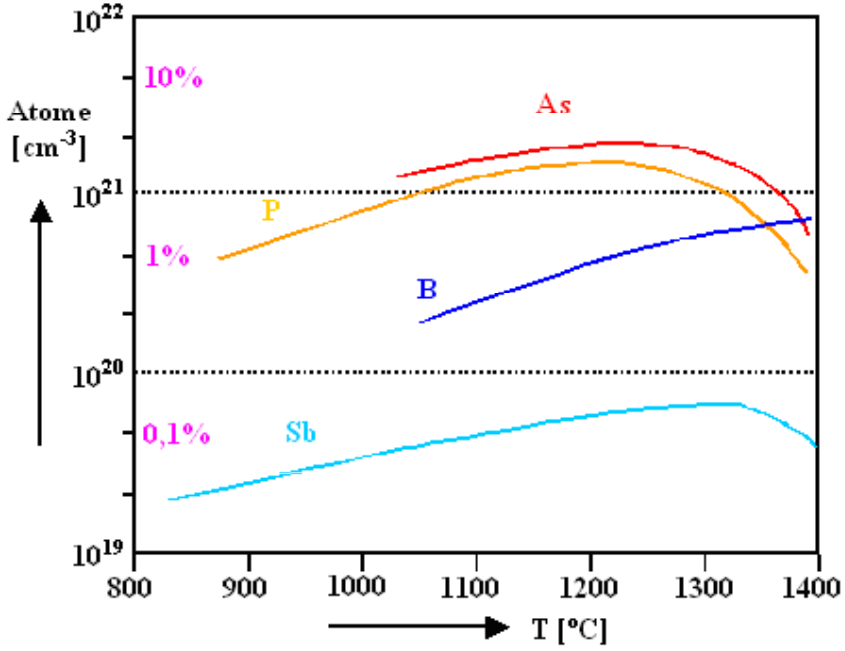
# CMOS

## Doping Limits

Doping concentrations need to increase as part of the scaling process: same number but smaller device size

There is however a solubility limit above which no more dopant atoms can be accommodated.

This happens at the source and drain contacts first. Also spreading of the dopants (out-diffusion) occurs during the annealing cycles and this is worse for high doping. Practically we are at those limits now.




Solubility limits for various dopants in silicon

# CMOS

## Gate Lengths

The technological limit here is the lithography, as discussed before. Further developments may get us down to 10nm resolution, but not without much more work and ever greater cost. Difficult to see optical techniques going beyond that limit.

Gate lengths of < 10 nm will result in the problem of direct tunnelling between the source and drain contacts, so this is probably the limit of the MOSFET approach anyway.

**We are not far from that limit (2-3 years)**

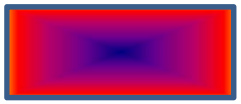
# CMOS

## Interconnect Delays

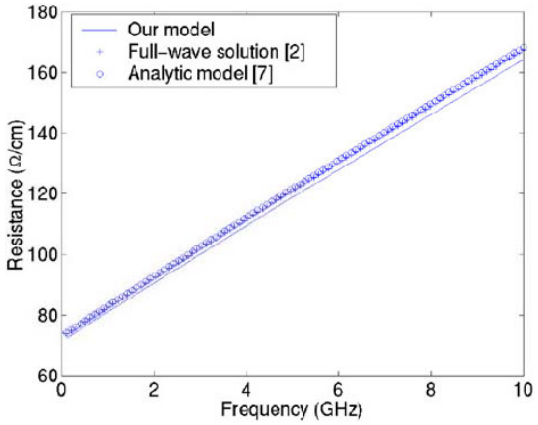
As devices become smaller as the interconnect wires become thinner. This gives rise to more resistance.

Also at high frequencies we the skin effect in which current becomes confined towards the surface of the conductor

Low frequency



High frequency




Calculation showing the increase in resistance due to the skin effect

Additionally the wires and metal contacts are now closer together and subject to more capacitive coupling.

# CMOS

Such capacitances are proportional to the dielectric constant of the insulator. Need low-k materials ($<k_{Si02}$ i. e <3.5) for this application

| Film | *k* range | Deposition method |
|---|---|---|
| FSG | 3.2-4.1 | PECVD |
| Polyimides | 3.1-4.0 | Spin-on |
| HSQ | 2.5-3.3 | Spin-on |
| MSQ | 2.0-3.0 | Spin-on |
| SiOCH | 2.2-3.5 | CVD, PECVD |
| BCBs | 2.6-2.8 | Spin-on |
| Fluorinated polyimides | 2.5-2.9 | Spin-on |
| Diamond-like carbon | 2.7-3.4 | PECVD |
| Spin-on organics | 2.0-3.2 | Spin-on |

A simplistic expression for the interconnect delay is given by:

$$\tau = R_{tr}(C_G + cl) + rl(C_G + cl) = (R_{tr} + rl)(C_G + cl)$$

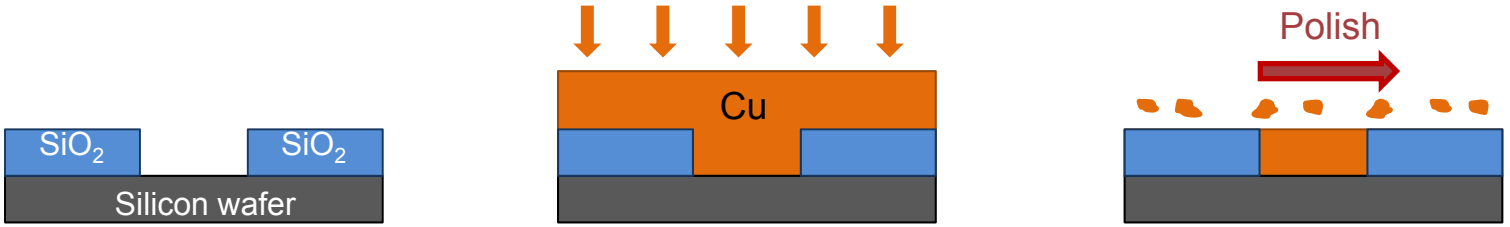
Where $R_{tr}$ = 'on' resistance of transistor , $C_G$ = transistor gate capacitance and r, c = resistance and capacitance of the interconnects per unit length

Clearly the interconnect delay could increase substantially if r and c increase with reduced dimensions or increased operating speed

# CMOS

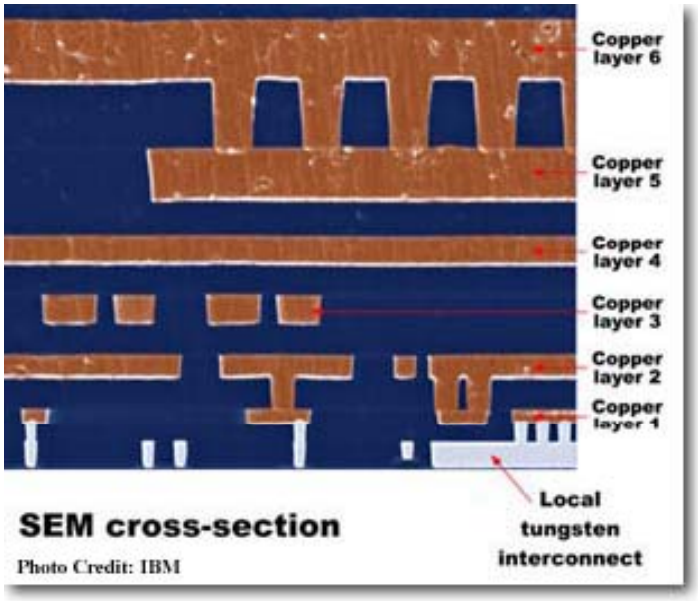
Early CMOS used Al interconnects for ease of deposition and etch. Cu has a lower resistivity, however Cu is difficult to work with:

Difficult to etch, needed a new process- Damascene.




Also Cu reacts with Si to form $CuSi_2$ leading to contact resistance and adhesion problems. Needs special interfacial layers.

To reduce k, use a low-k dielectric such fluorosilicate glass



# CMOS

Despite these improved technologies interconnect power losses are still a major factor in limiting CMOS performance

I**nterconnect power losses are predicted to dominate beyond the 22nm node (2013) unless this is specifically addressed.**

**Hot Electron Effects**

These occur at short channel lengths, in which the field is high

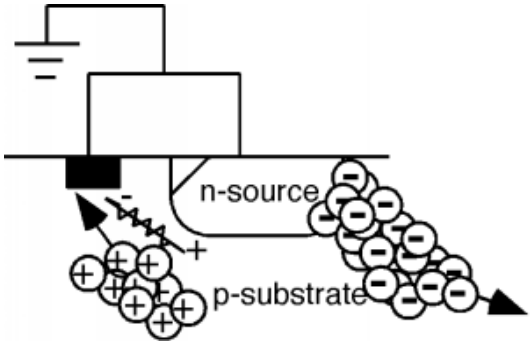
Electrons accelerate in the high electric field and become ‘hot’ (energetic)

These hot electrons can surmount the gate dielectric barrier causing increased gate leakage. Process is called *thermionic assisted tunnelling*

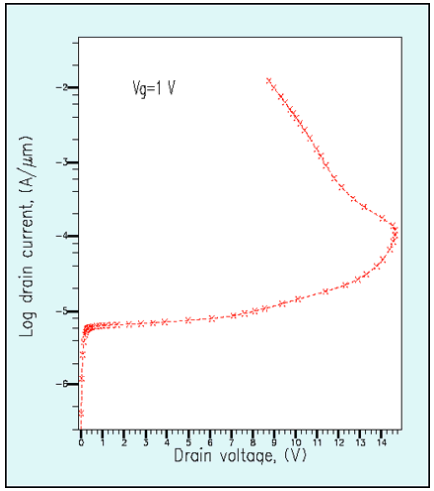
# CMOS

These electrons can enter and get trapped in the dielectric causing a time dependent increase in the threshold voltage, $V_T$

Hot electrons can also produce impact ionisation (see later) in which secondary electrons increase the drain current with increased $V_D$ leading to increased output conductance $g_D$

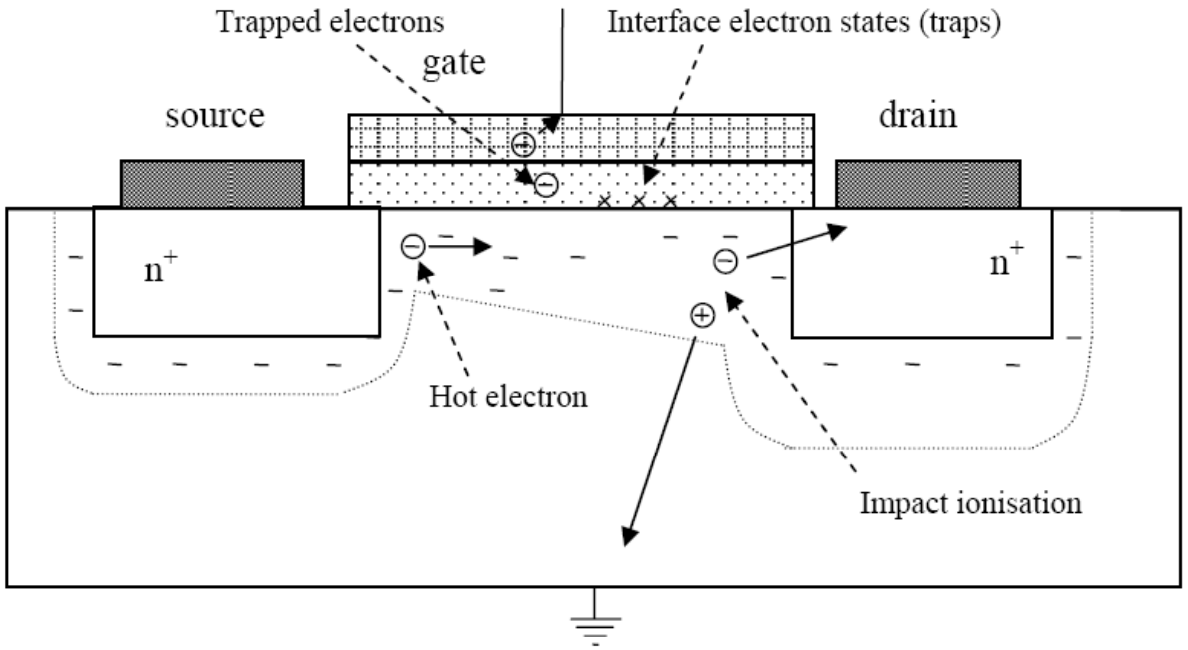



Secondary holes are also produced by this process, and these drift towards the substrate. These holes can causing noise or cause latch-up problems in CMOS



# CMOS

These effects cause the device characteristics to change with time and hence can lead to poor reliability.




Devices are tested under higher than normal operating conditions (voltage and temperature) to assess reliability. Latch-up (a low impedance path) can develop with time and can show up under such testing.

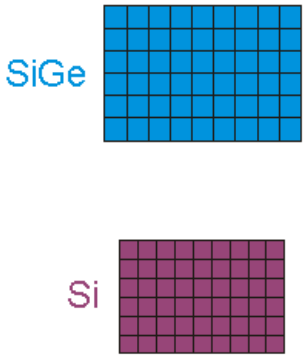
# CMOS

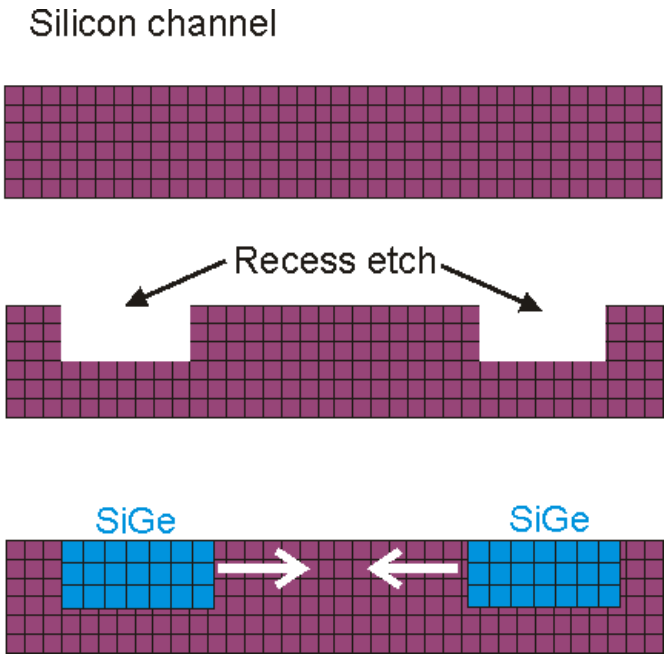
Hot carrier effects are less problematic for p-channel MOS because the holes have a lower mobility (less energy) and they have a higher gate oxide barrier to surmount.

## Channel mobility and strain

Starting with the 90nm node, strained Si channels have been introduced to improve the mobility of electrons and holes.

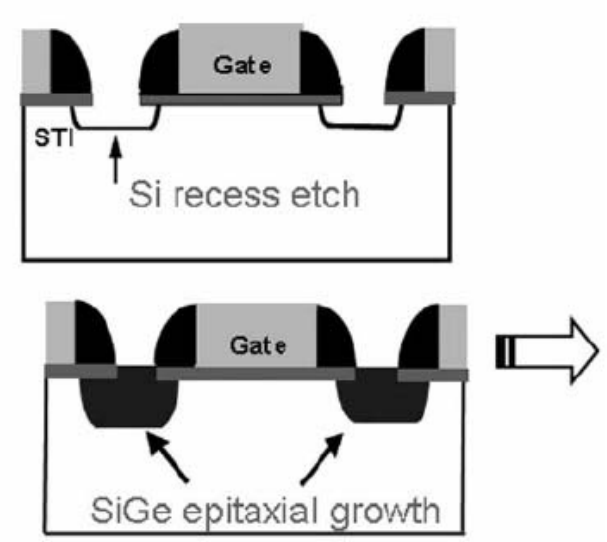
The strain has been introduced by using a SiGe alloy which has a larger lattice constant (crystal spacing) than the Silicon

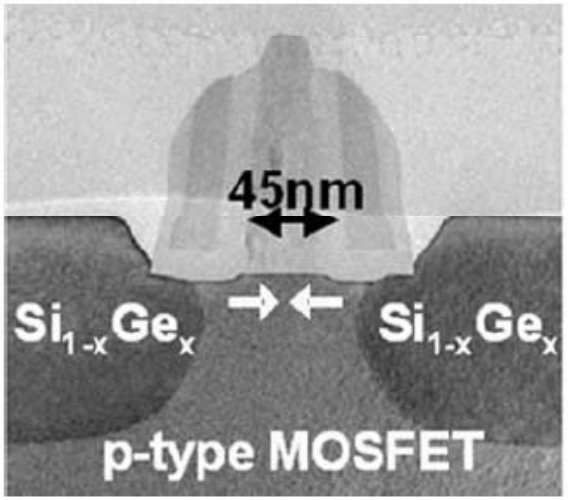


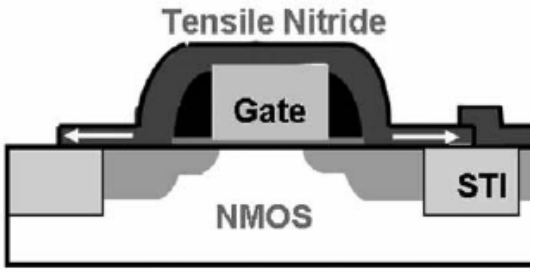
# CMOS

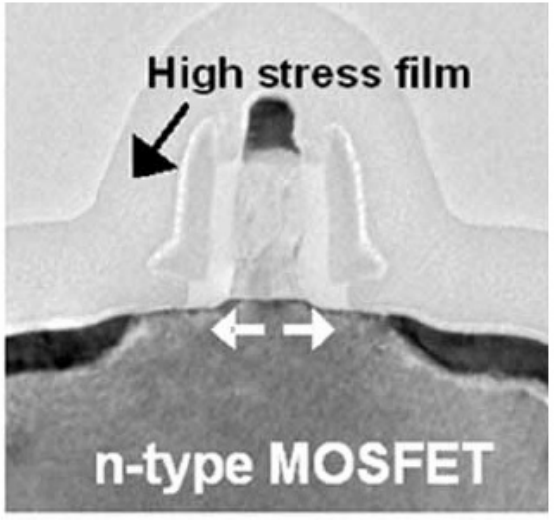
**Compressively strained P MOSFET**



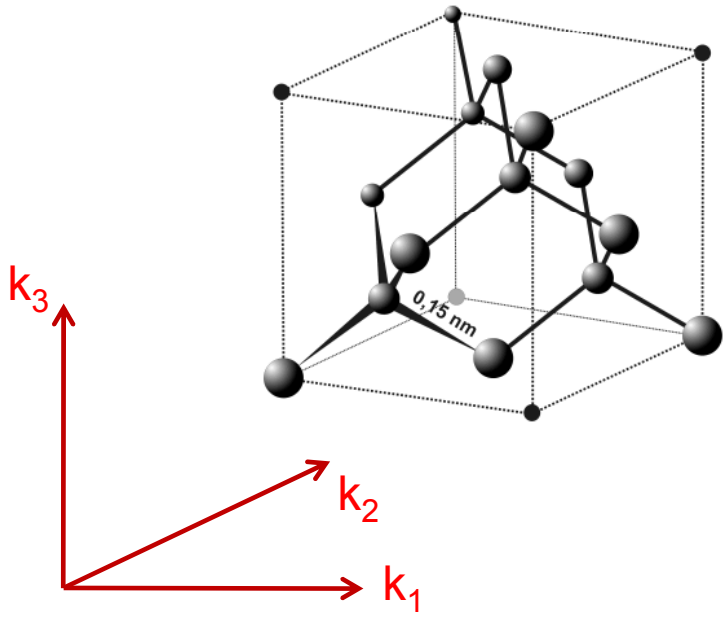

**Tensile strained N-MOSFETs**

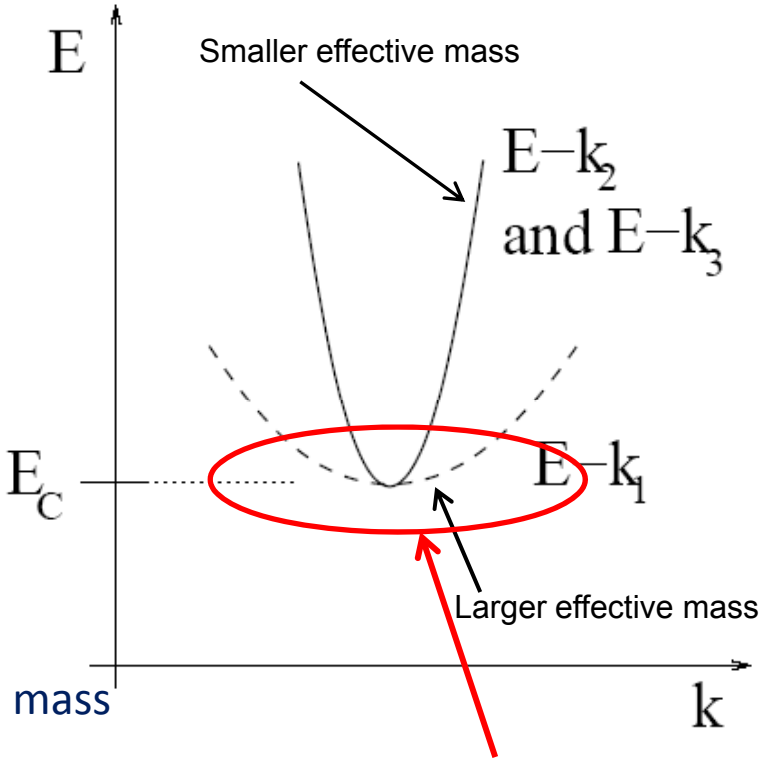


# CMOS

Using strain exploits the fact that m* is direction dependent in semiconductor crystals. This is because electrons can move through the crystal more easily in certainly directions



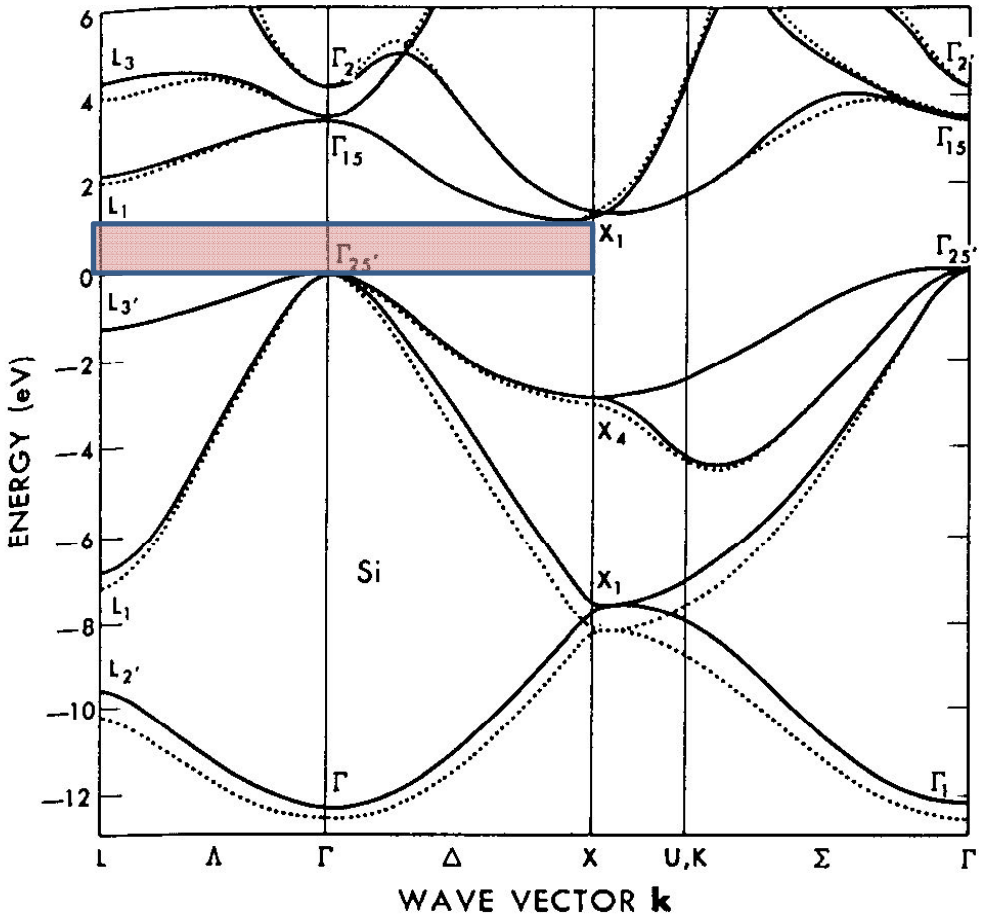

- $k_1$ is a <110> direction. This has a large effective mass
- $k_2$ and $k_3$ are orthogonal directions and have a smaller effective mass

We can think of these conduction band minima as valleys

# CMOS

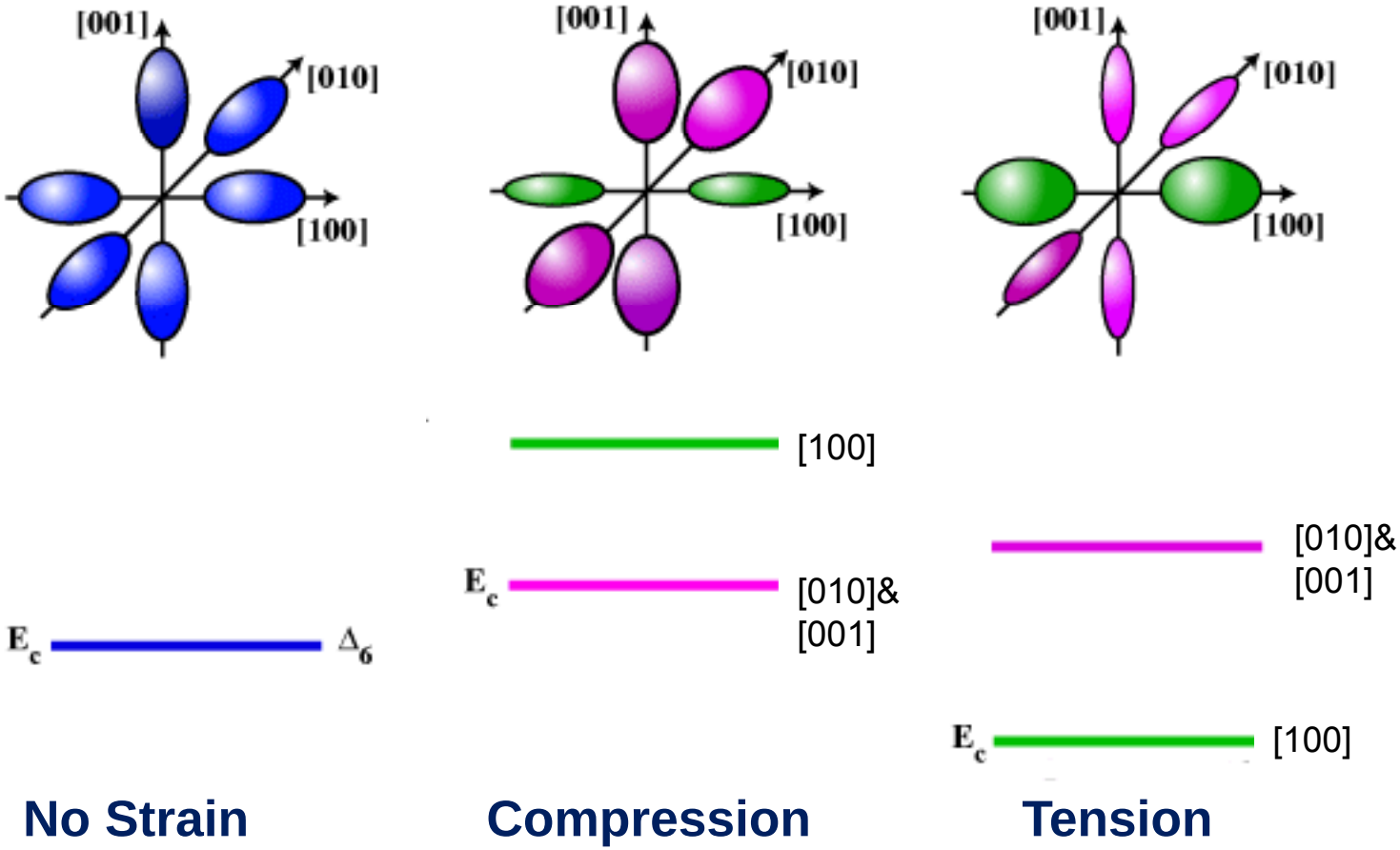
The full band structure is very complex.

We refer to the three reciprocal space directions as
Γ, X and L



# CMOS

Consider 6 conduction band minima (two for spatial each direction)



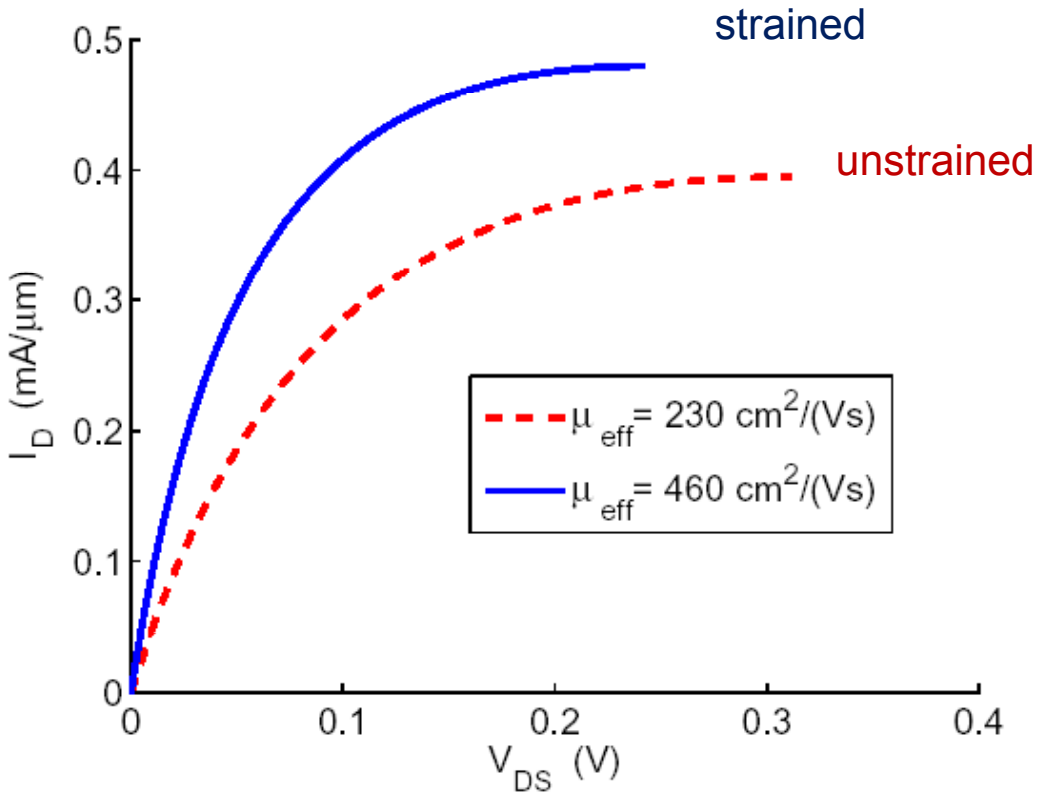

Application of strain can split & raise/lower the bands on orthogonal axes. This can be used to bring a low effective mass band below that of a high effective mass one, increasing the overall mobility

# CMOS

Silicon -unstrained $m_e$*= 0.26

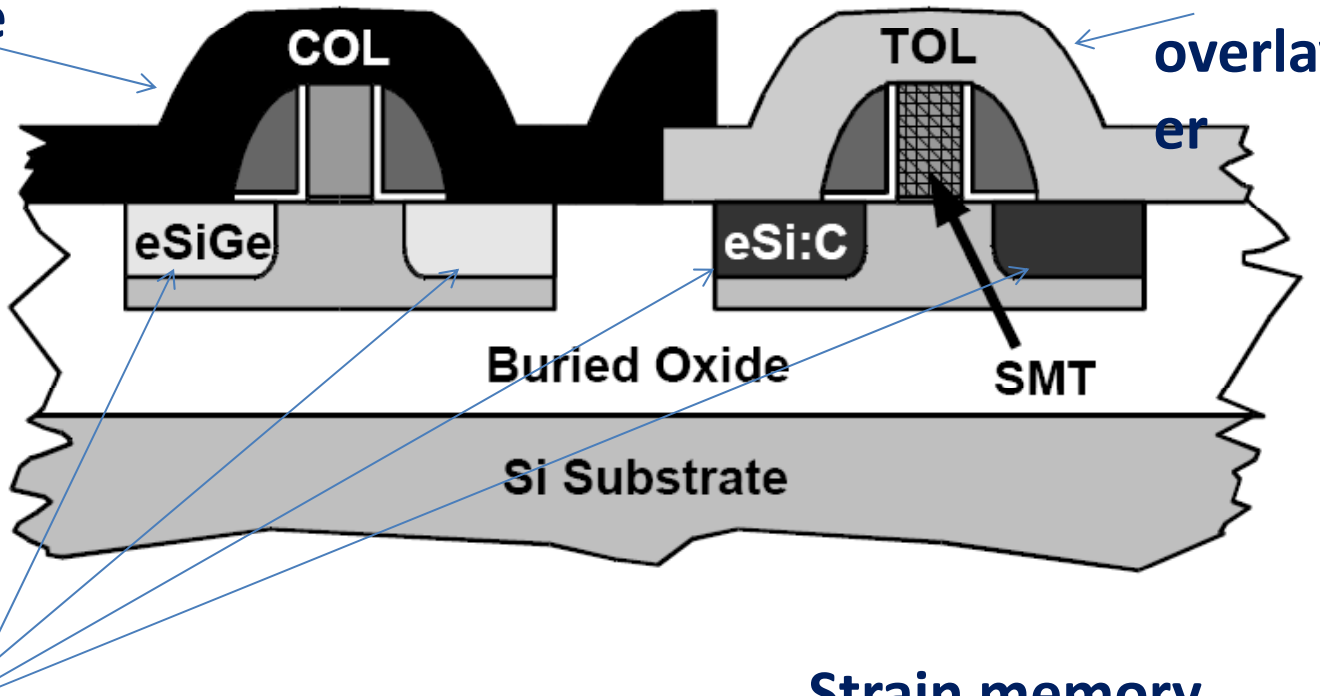
If we strain and bring a lower effective mass band down below energy of the other bands then the lowest band has an $m_e$*=0.19

**Potential ~40% improvement in effective mass, translating to a 40% improvement in $I_D$**



# CMOS

Latest generation: more extreme strain!




**COL-compressive overlayer**

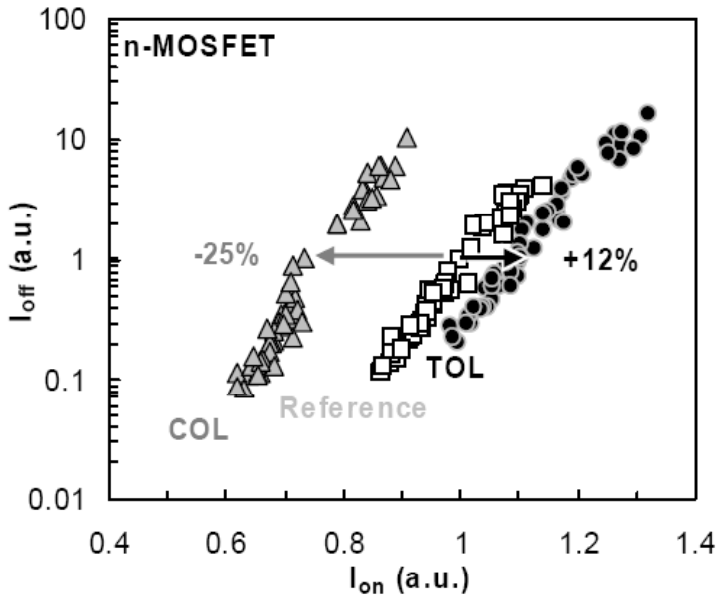
**TOL-tensile overlayer**

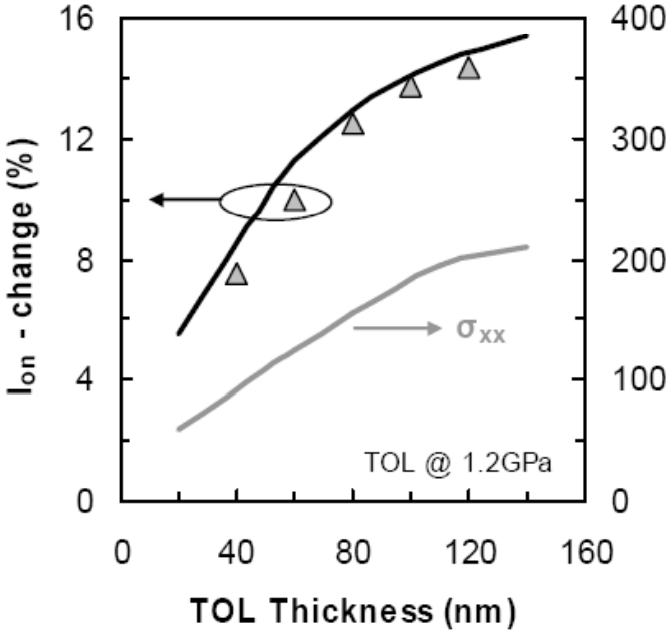
**Source-drain stressors**

**Strain memory technology (re-crystalisation)**

# CMOS

One commonly used performance measure for CMOS is to look at the ratio of the **ON** current divided by the **OFF** current. Usually this needs to be >1000




For NMOS  a tensile strained overlayer (TOL) improves the ON current, whilst a compressive strained overlayer reduces it. The effect is opposite for PMOS.

# CMOS

A combination of strain techniques is now being used to give performance enhancements.

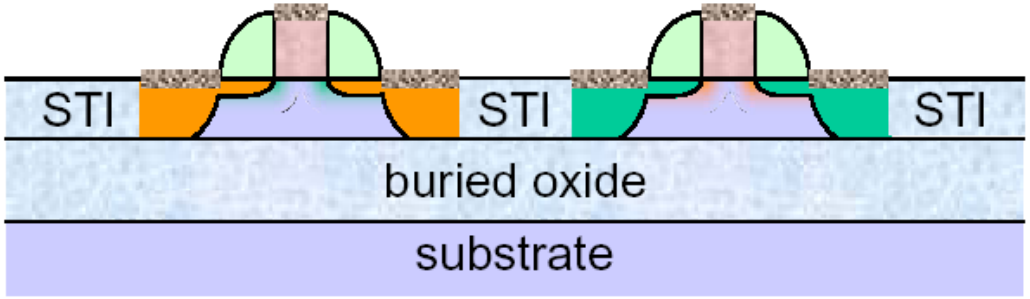
There is still the possibility to go further using more effective or even new types of stressor.

| $I_{on}$-gain | n-MOSFET | p-MOSFET |
|---|---|---|
| eSi:C | (6 %) | - |
| TOL | 12 % | - |
| SMT | 10 % | - |
| sSOI | (10 %) | - |
| COL | - | 35 % |
| eSiGe | - | 25 % |
| Total Theory | 22 % | 60 % |
| Total Experiment | 20 % | 72 % |

Note the big improvement is for the performance of the pMOS

# CMOS

Previous work has discussed FETs on doped substrates (usually p-type). Modern CMOS tends to use silicon-on-insulator (SOI) as a substrate which has a thin oxide layer implanted into the substrate



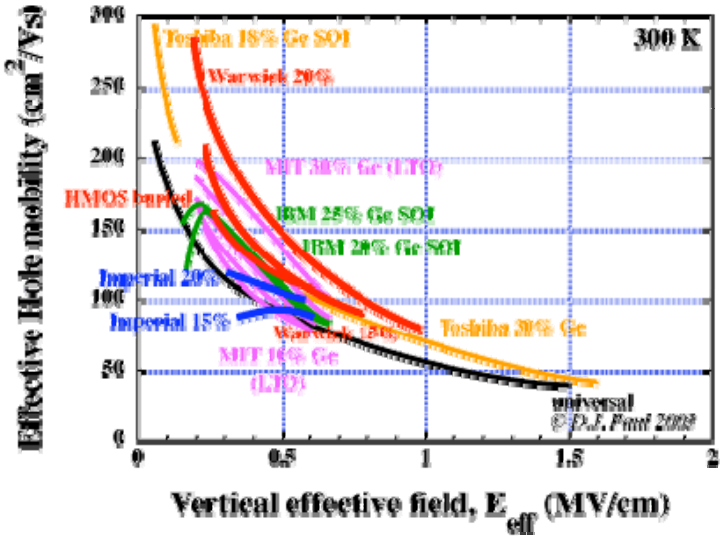

The oxide creates an insulating layer below the channel. This device then has no major parasitic elements coming from the source and drain interacting with the substrate (no *floor*).

One problem however is that the upper Si layer is degraded by the oxide implantation. Newer technology uses wafer bonding techniques

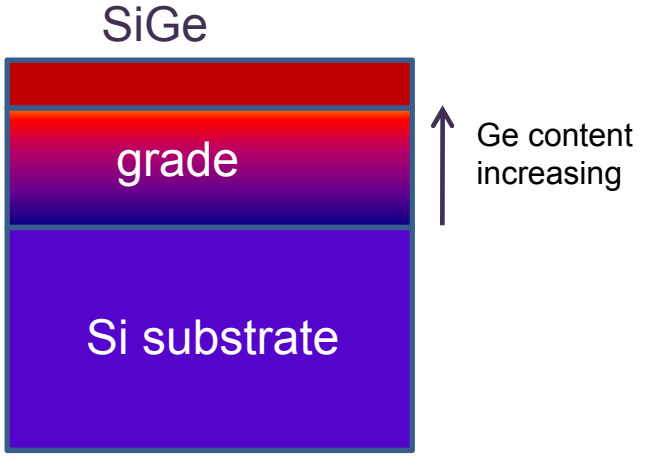
# CMOS

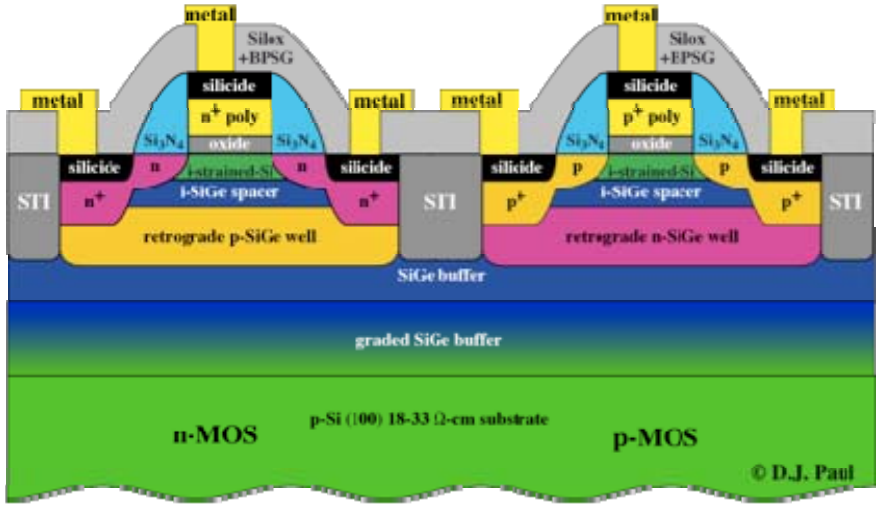
Strained channels (metamorphic)

Increasing the SiGe content of devices gives a major improvement in mobility, particularly for the holes



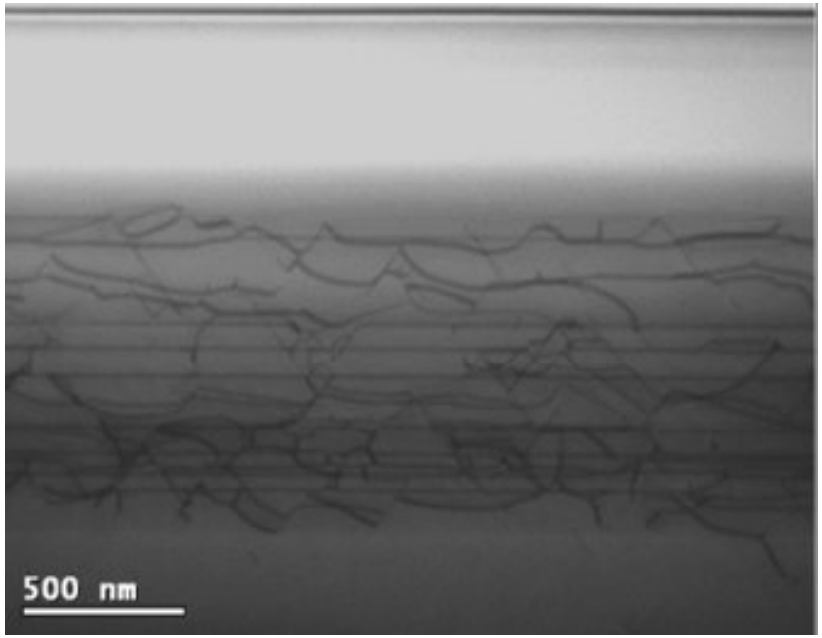

This SiGe based approach is called Heterostructure MOS (HMOS)

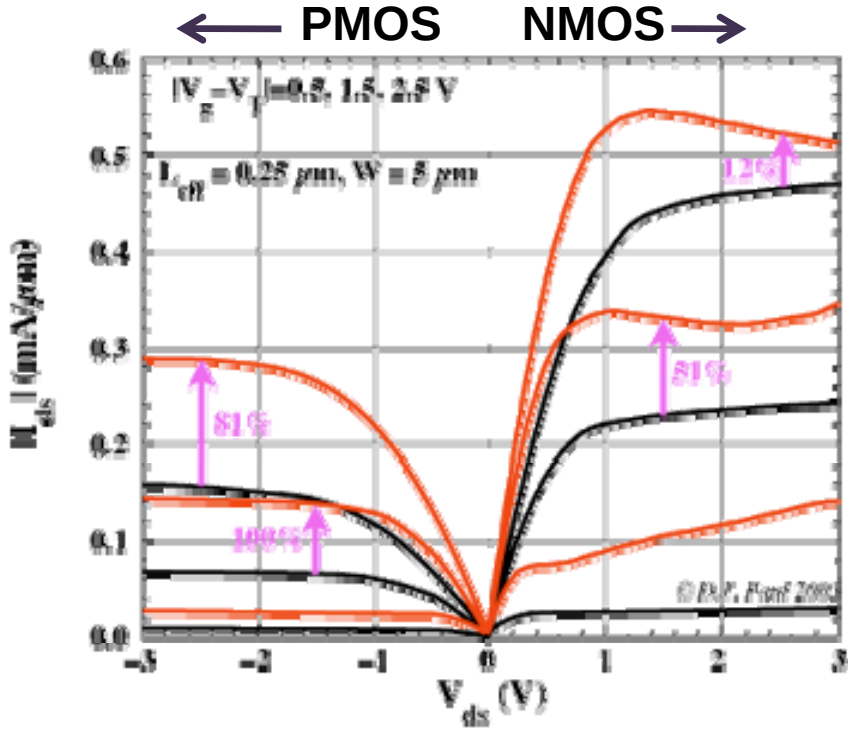


# CMOS

Both n-type and p-type devices are improved, but the major benefit is for p-type. This results in a more ‘complementary’ CMOS device

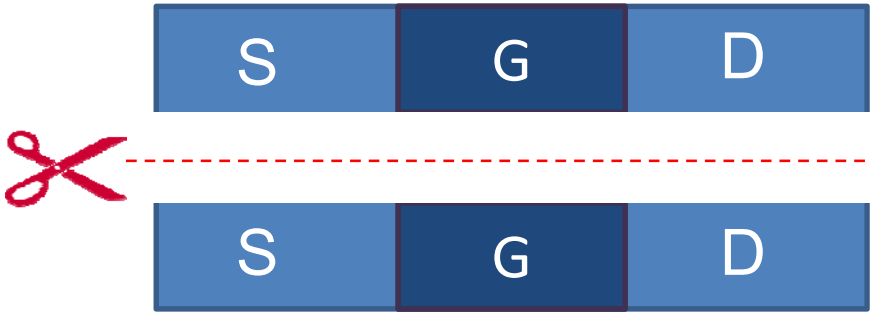



A major issue with this technology is controlling the dislocations which come from the grading from Si to SiGe

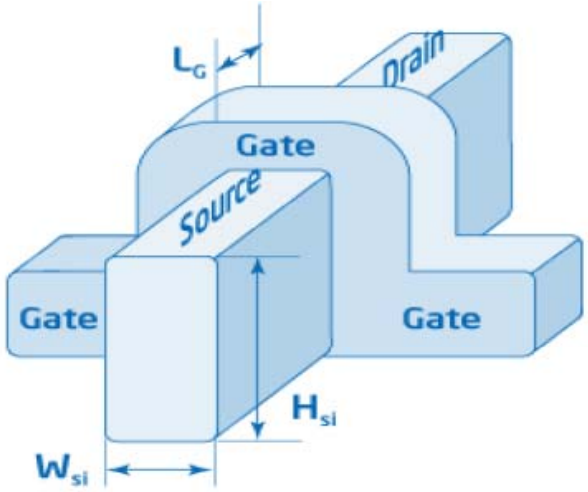
# Multi-gates and Fin gates

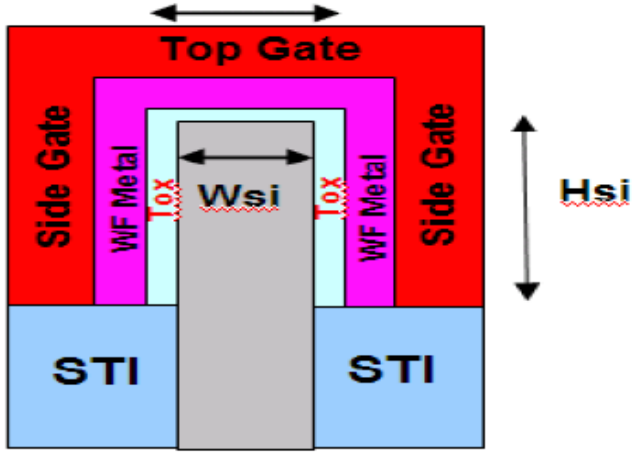
At small dimensions a multi-channel structure better than a single channel one with a single gate




The spread of the field is reduced leading to increased $I_d$ and reduced $V_T$

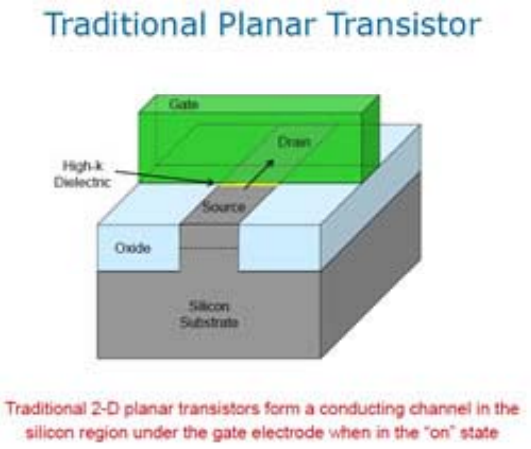
A 3D gate performs even better in this respect (called a Fin gate)



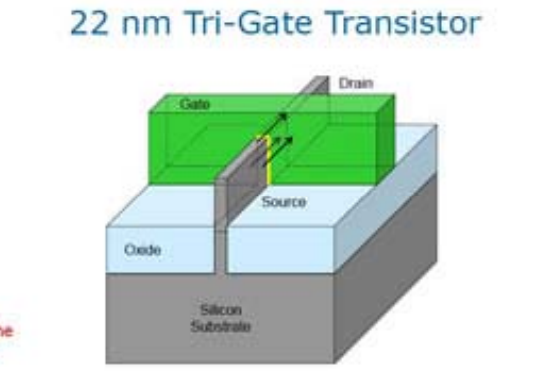



# Multi-gates and Fin gates

Better still is to have multi-gates **and** Fins.

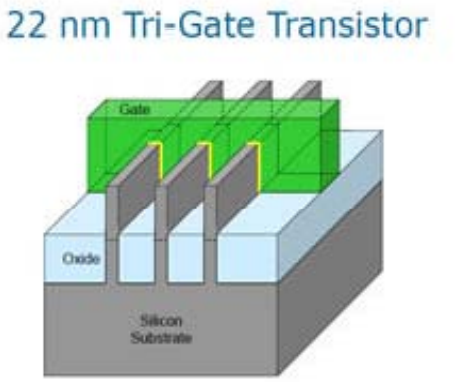



Traditional 2-D planar transistors form a conducting channel in the silicon region under the gate electrode when in the "on" state




3-D Tri-Gate transistors form conducting channels on three sides of a vertical fin structure, providing "fully depleted" operation

This approach is known to show a significant increase in drain current for the same gate voltage (higher transconductance)

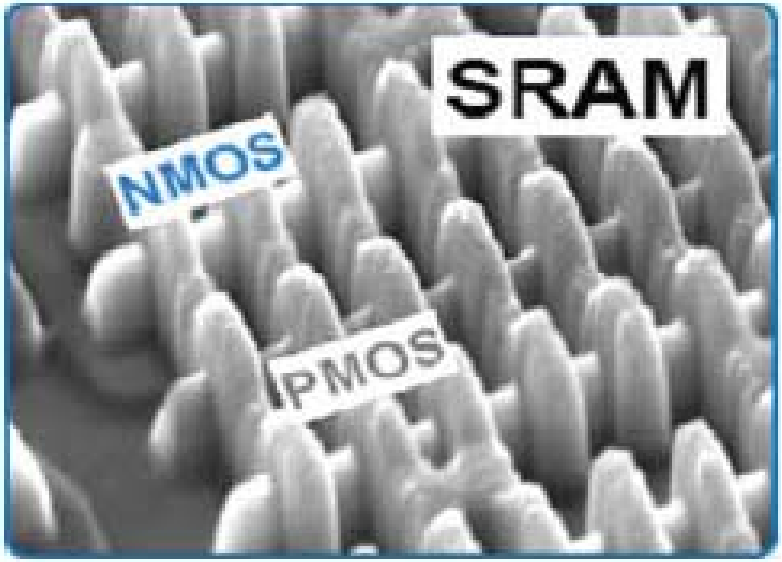


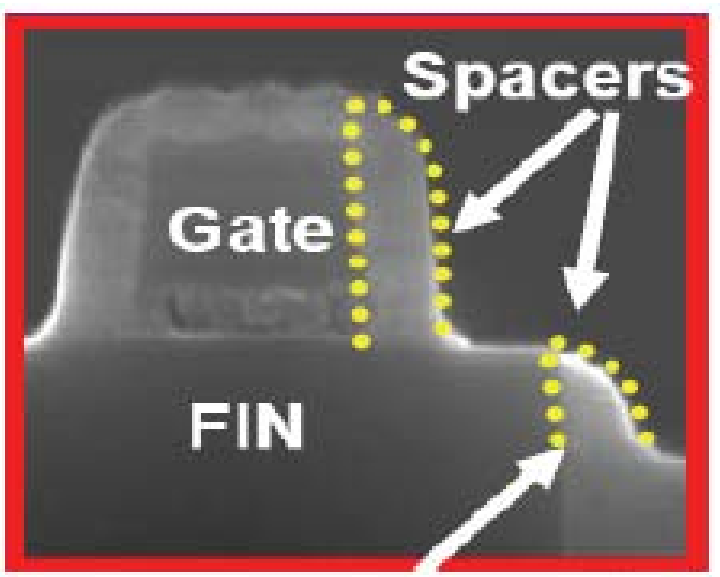

Tri-

Gate transistors can have multiple fins connected together to increase total drive strength for higher performance

# Multi-gates and Fin gates

Intel call this the ‘Tri-gate’ approach and use this on their 22nm node processors



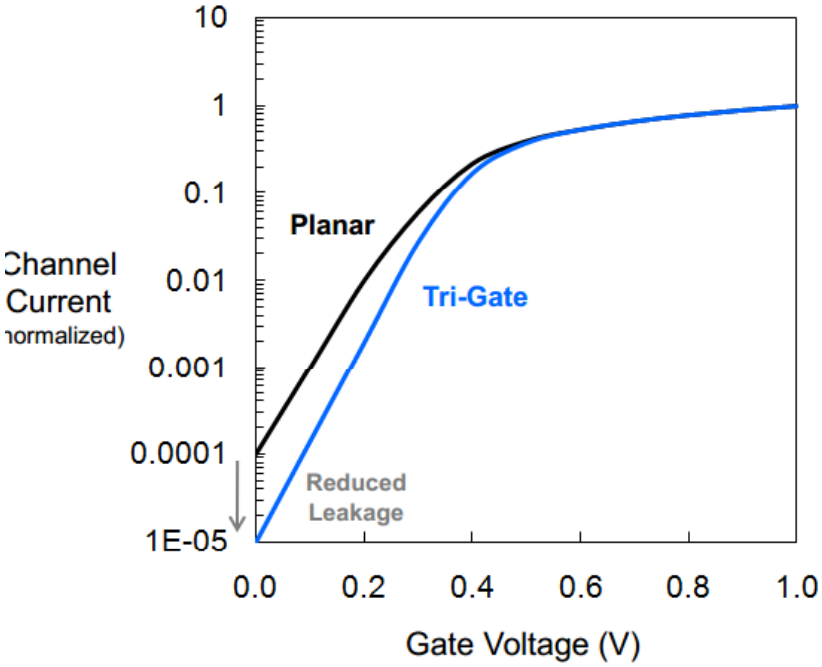



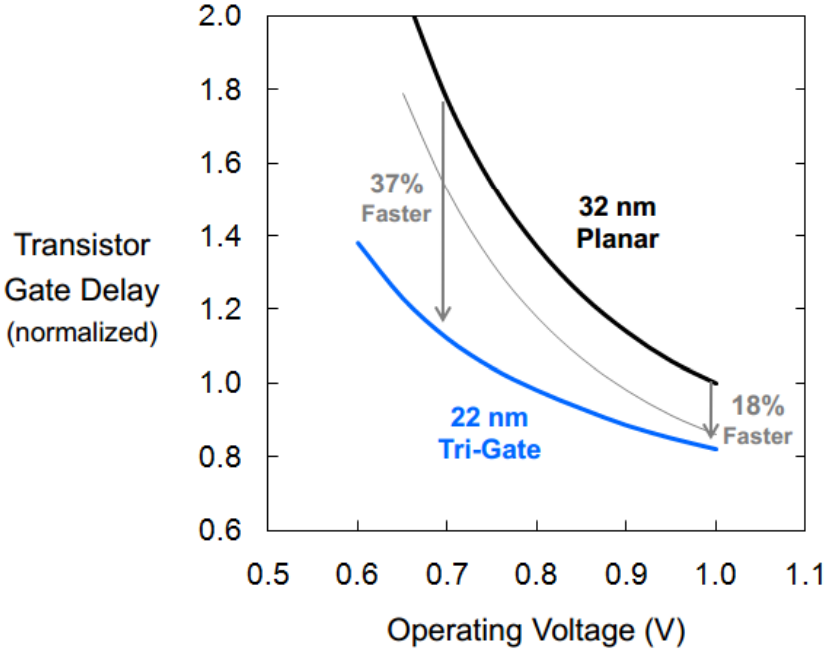

Very little in the way of performance data is available at the moment- keeping many of the details under wraps

# CMOS

According to Intel the Tri-gate electrode increases the inversion layer area to allow **a reduced** drive current. Process cost additions are only 2-3%

Gives ‘unprecedented’ gains in operating speed at low voltage



# CMOS Processors

## Processor Types

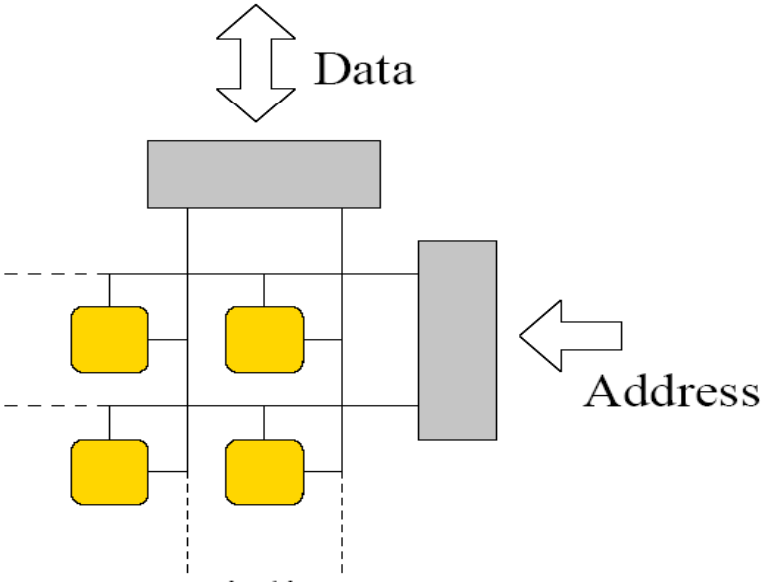
- **Full-custom:** All circuits/transistor layouts, optimized for a specific application.
- **Standard-cell:** Arrays of small function blocks (gates, FFs) automatically placed and routed.
- **Gate array:** Partially prefabricated wafers customized with metal layers.
- **Field-programmable gate array (FPGA):** Prefabricated chips with field-programmable switches
- **Microprocessor (CPU, MCU):** Instruction set interpreter customized through software.
- **Domain Specific Processor**: (DSP, NP, GPU).

# CMOS Memory

Semiconductor memory is critically important in modern digital systems. There is a strong drive to increase capacity, reduce cost and reduce access times. There are two types of memory in common usage, both based on the Si MOSFET

- **Flash Memory-** Computer BIOS, Memory sticks, SD cards etc
- **Dynamic random access memory (DRAM)-** PCs, embedded systems

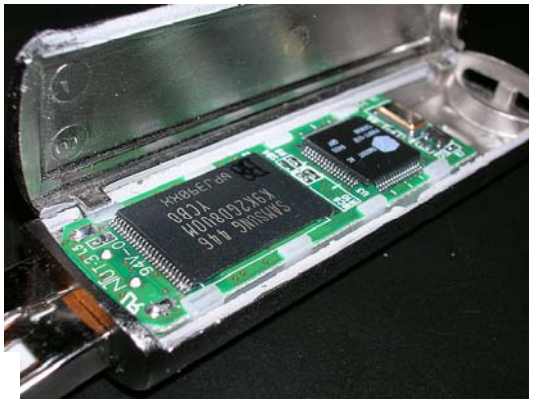
Both have the same basic function in that the gate is used to 'write' information and the source-drain is used to 'read'. However the gate structure and source connection differs between the types



# CMOS Memory

## FLASH

Write a few times, store without power, read infrequently




## DRAM

Write/Read frequently, store needs power.

# CMOS Memory

Traditional hard disk drive

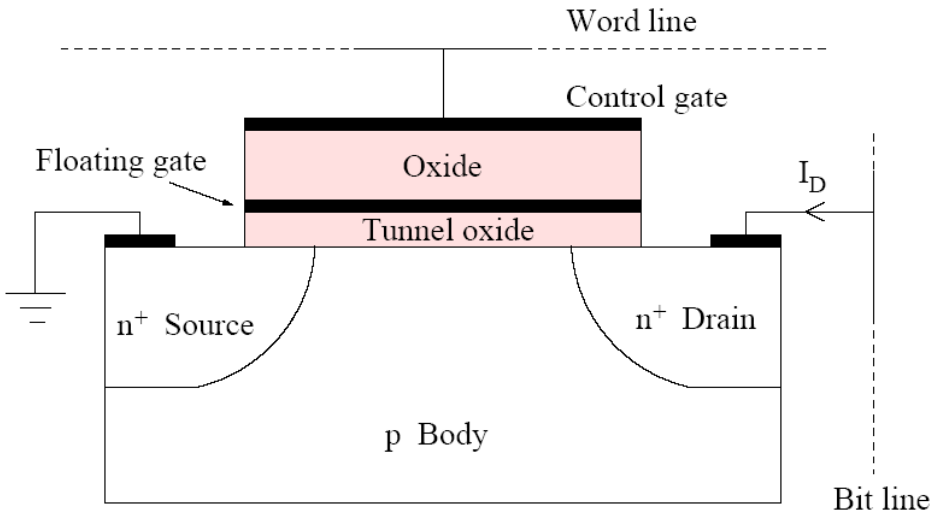
Solid state hard drive

| | Solid State Memory Storage | Hard Disk Storage |
|---|---|---|
| Spin-up time | Instantaneous. | Up to few sec. |
| Random access time[45] | About 0.1 | 5–10 ms |
| Read latency time[46] | Very low | High |
| Defragmentation | Not needed. | Needed, for optimum performance |
| Acoustic levels | SSDs have no moving parts and make no sound | HDDs have moving parts (heads, spindle motor) and have varying levels of sound depending upon model |
| Mechanical reliability | A lack of moving parts virtually eliminates mechanical breakdowns | HDDs have many moving parts that are all subject to failure over time |
| Weight and size[52] | The weight of flash memory and the circuit board material are very light compared to HDDs | Not as light as SSDs) |
| Write longevity | Flash memory: limited number of writes over the life of the drive. SSDs based on DRAM do not have a limited number of writes. | Magnetic media do not have a limited number of writes. |
| Cost | High: $1.00–2.00 per GB | Low: $0.01-0.05/GB |
| Power consumption | 1/2 to 1/3 the power of HDDs | 2-10 Watts |

# CMOS Memory

A Flash memory cell has two gates separated by a thick oxide layer

The idea is to place (write) or remove (erase) electrons on the floating gate

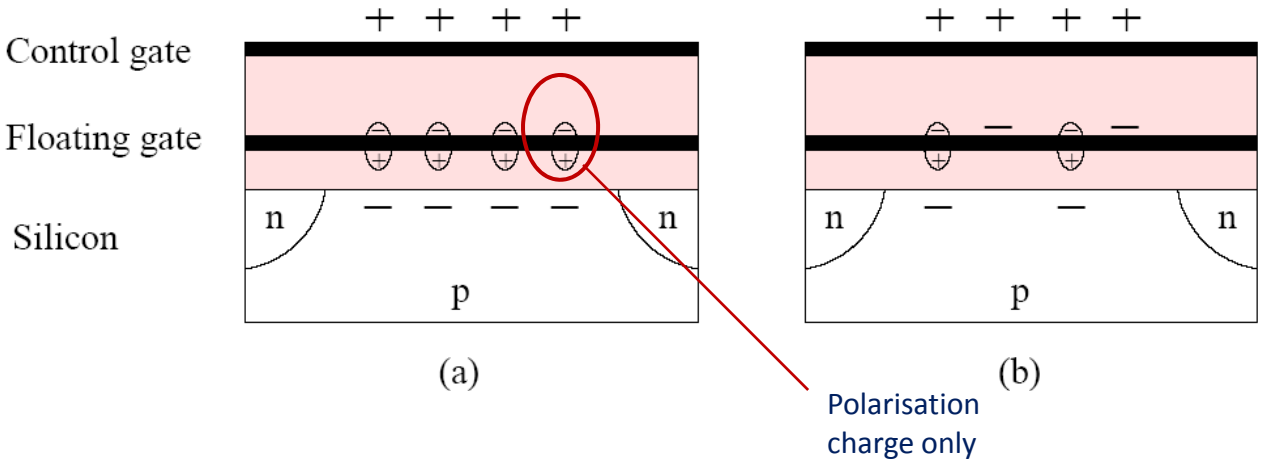



The charge on the floating gate alters the threshold voltage of the FET and with appropriate bias can change the magnitude of drain current . A current sensing circuit on the bit line interprets this drain current as either a ‘1’ or a ‘0’.

# CMOS Memory

**(a) Equilibrium**- no net charge on the floating gate

**(b) Electrons added to the floating gate**- reduction in electron density in the channel




No net charge on the system.
Therefore:

$$Q_n + Q_F + Q_G = 0$$

Channel Float Gate

If the floating charge is changed by $\Delta Q_F$ then to maintain a given $Q_n$ means that $Q_G$ must change by - $\Delta Q_F$

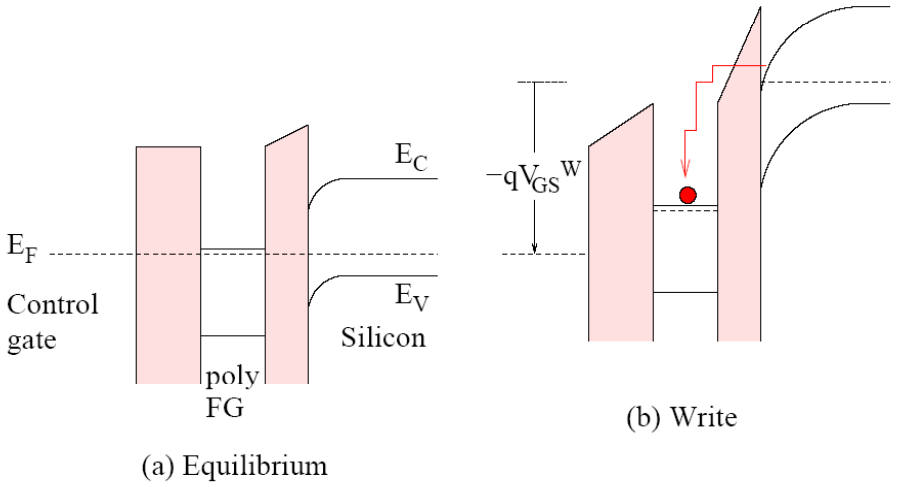
# CMOS Memory

Associated change in the threshold voltage

$$\Delta V_T = \frac{\Delta Q_G}{C_{ox}} = -\Delta Q_F \frac{t_{ox}}{\varepsilon_{ox}}$$

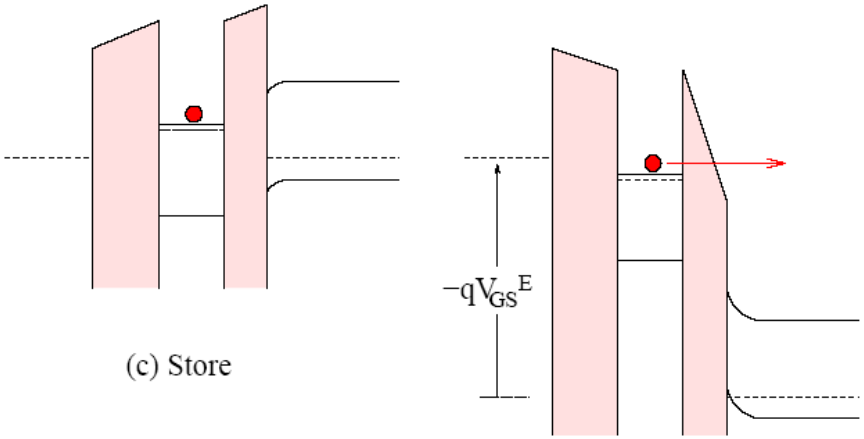
The read voltage is set up so that it is just above $V_T$ when $Q_F = 0$. In this case the transistor turns on and registers a '1'. However if charge is added to $Q_F$ then $V_T$ is raised and the transistor turns off(='0')

Programming of the floating gate occurs by applying a high positive voltage pulse to the control gate. Under these conditions electrons can tunnel through the gate oxide using 'field assistance'



# CMOS Memory

To erase the cell, a large negative voltage is applied to the control gate and electrons tunnel out to the gate.




The use of a rapid high voltage pulse to erase gives us the term '**FLASH memory**'

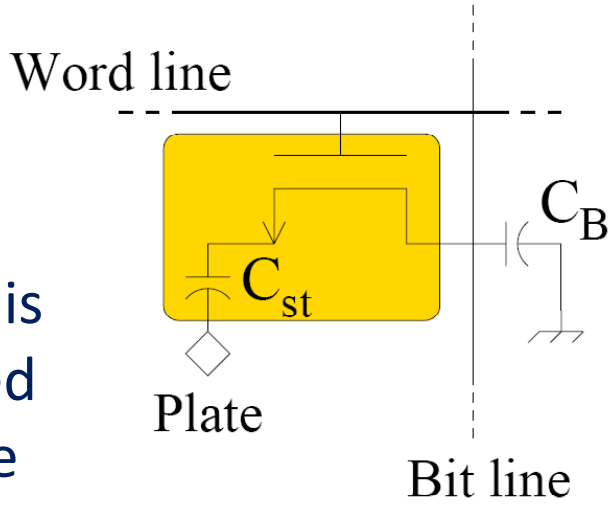
In the store state, the time for electrons to tunnel out of the floating gate is very long- practically almost infinite. In practice there are anomalous leakage mechanisms which increase with stress and with write-erase operations.

Solid state drives have low access times, a low latency, no moving parts- high reliability. But at present limited to about 100K write-erase-cycles

# CMOS Memory

In a DRAM the MOSFET acts as a switch between a bit line and a storage capacitor, $C_{ST}$

This is connected to the bit line when the FET is turned on by the gate. The bit line is connected to earth only via a bit line capacitor. Its voltage will change in response to a change in the charge state on CB, which depends on whether the FET is on or off and whether there is charge on $C_{ST}$




The plate terminal is usually held at $V_{DD}/2$. To **WRITE** a '1', the bit line voltage is pre-set to $V_{DD}+ V_T$ and when the word line is enabled, the transistor turns on raising the source voltage to $V_{DD}$ and so the voltage across $C_{ST}$ is $V_{DD}/2$

# CMOS Memory

In the **READ** operation, the bit line is preset to $V_{DD}/2$ and the transistor is turned on since the voltage on $C_{ST}$ is also $V_{DD}/2$, then there is **no change**. Circuitry interprets no change as a '1'

To **WRITE** a '0' the bit line is set to zero volts and the transistor is turned on. This sets the source voltage to zero and so $V_{ST} \rightarrow -V_{DD}/2$

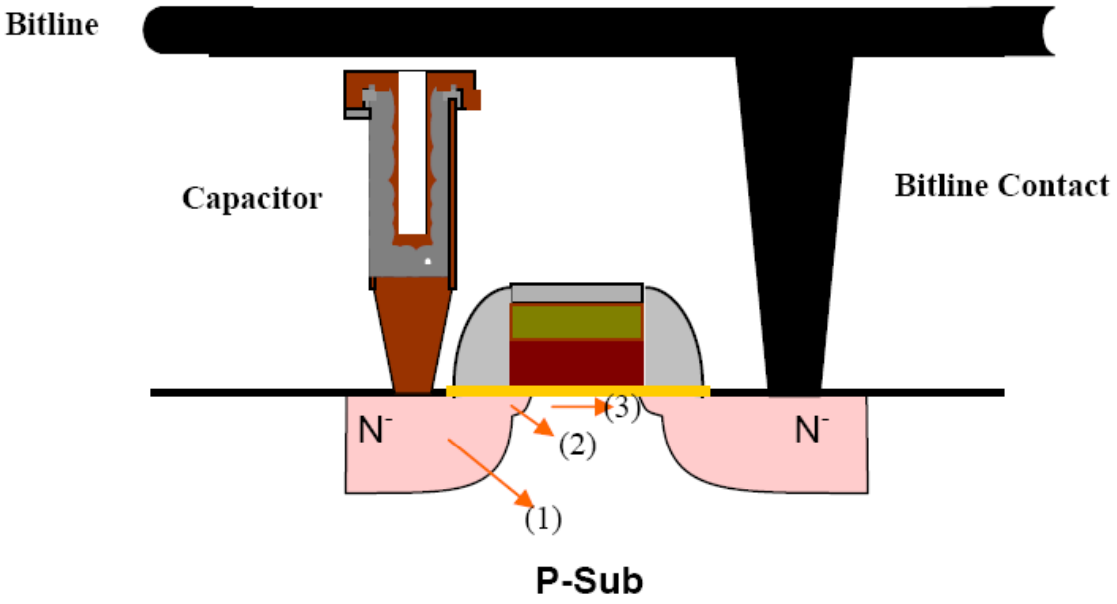
To **READ** the zero again the bit line is set at $V_{DD}/2$ and the transistor is turned on. Since $V_{ST} = -V_{DD}/2$ the bit voltage falls to zero (or $<< V_{DD}/2$). This change is interpreted as a '0'

However at this point the charge on $C_{ST}$ is depleted and would need to be replenished by writing another zero. Because of this need to replenish after reading a zero, this is called **dynamic** random access memory

# CMOS Memory

However this is not the only reason why this needs to be refreshed. Additional mechanisms include:

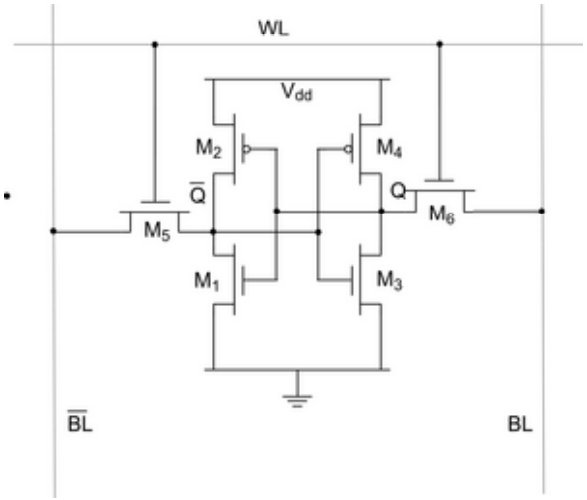
- Reverse junction leakage from the source to the substrate (1)
- Gate induced drain leakage (2)
- Sub-threshold  channel leakage  (3)




Typical DRAM refresh rates are in the range of 10-500µs.

# CMOS Memory

DRAM is being superseded by Static RAM (SRAM), which uses a bi-stable latch circuit to remove the need for immediate re-write. Improvement in speed comes at the expense of a 6 transistor circuit element!




These days Flash memory and DRAM/SRAM both have access times in the range of a few nsec. However DRAM/SRAM does not need the high voltage pulses needed for write and erase (10's of nsec) and therefore has a much lower latency.

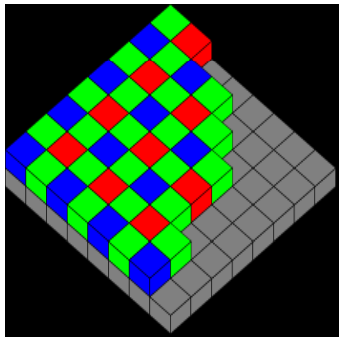


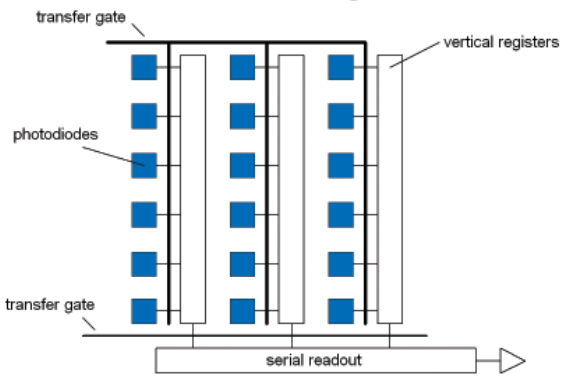
DRAM /SRAM also does not degrade with write/erase operations.

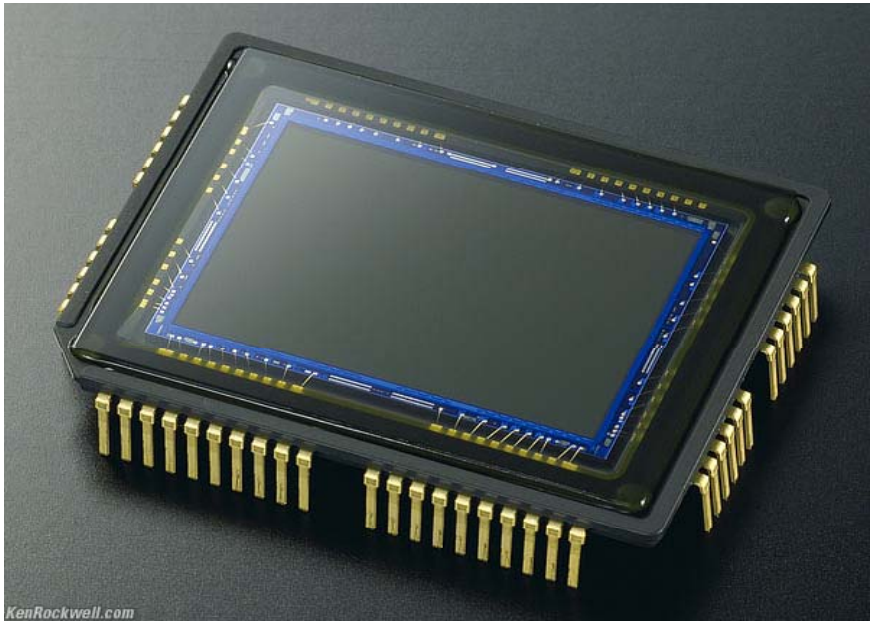
# MOS imaging

CMOS image sensors works by combining Si operating as a photodiode with MOS charge storage

There are two types: **charge coupled devices (CCDs)** and the newer **CMOS image sensors**

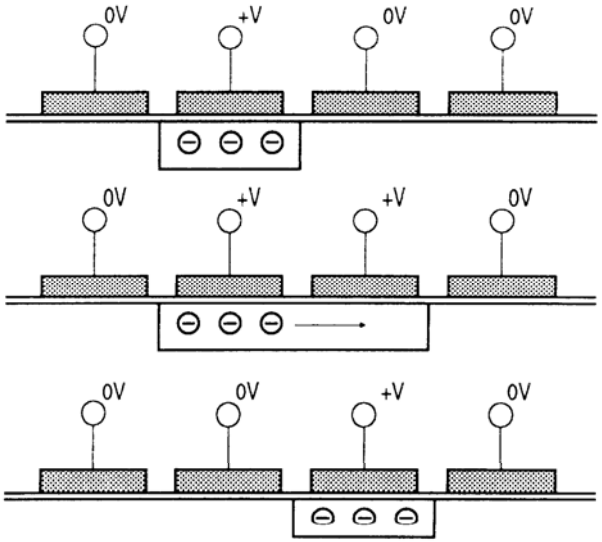





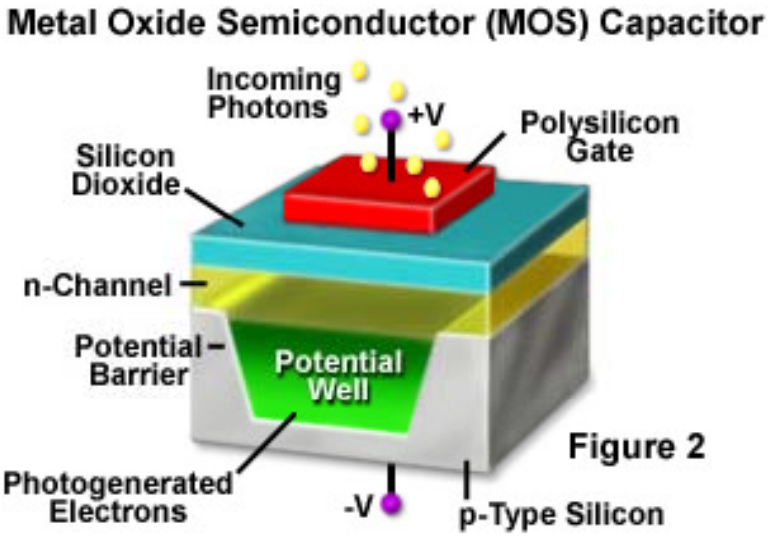



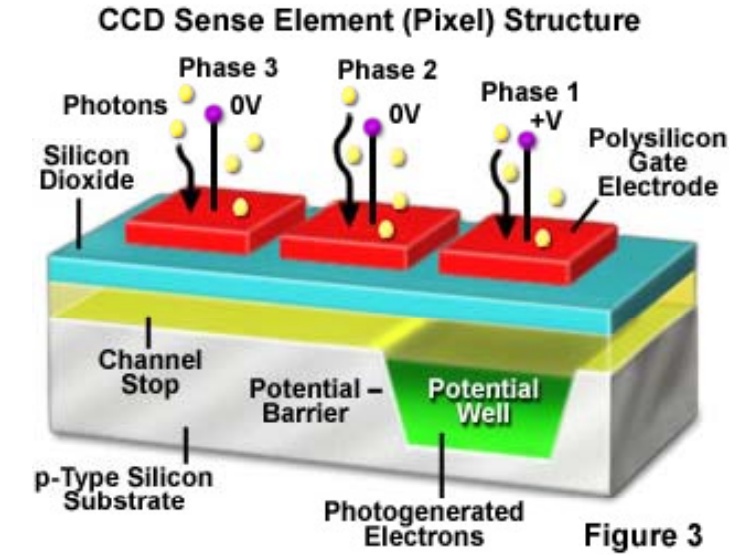
# CCD

CCDs make use of a MOS capacitors ability to charge up under illumination. The charge is stored in a potential well




Charge is then moved around the structure by sequentially applying voltages to the gates



# CCD

The shutter opens and closes giving an accumulated charge in each well. These charges are then shifted row by row across the parallel register under control of clock signals. Rows of charge packets are transferred in sequence into the serial shift register.

Charge contents of pixels in the serial register are transferred one pixel at a time to an on-chip amplifier, which boosts the electron signal. This signal is then processed to give the image

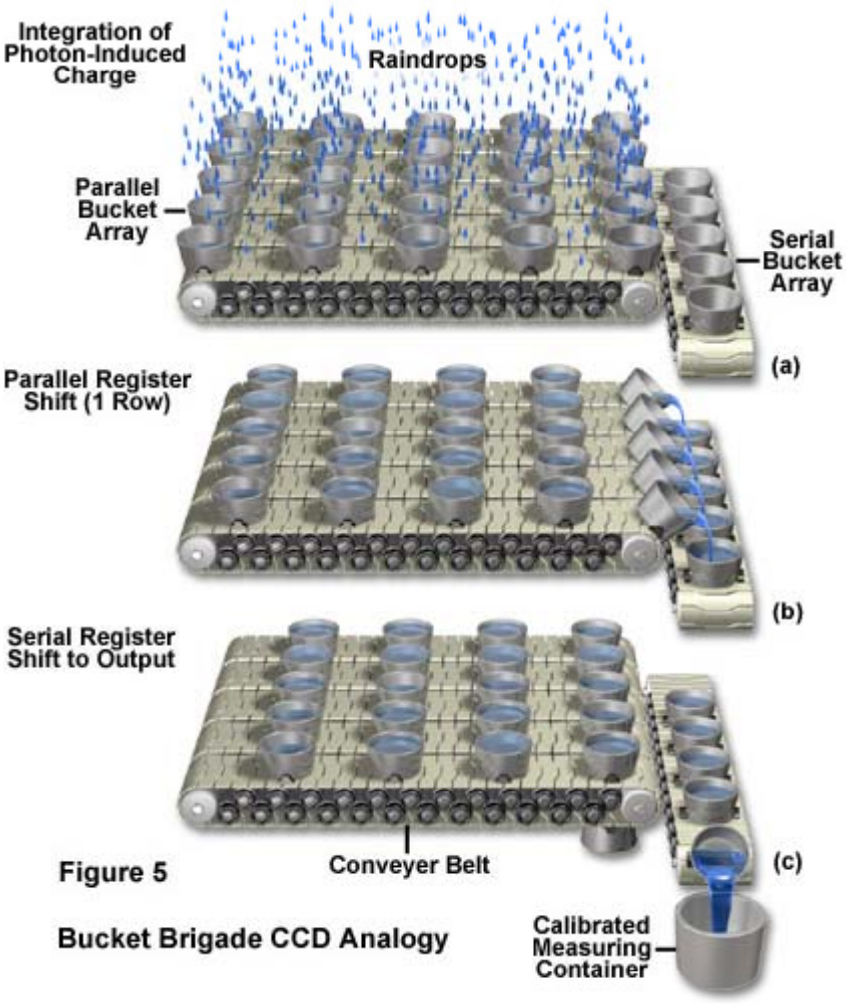



Figure 5

Bucket Brigade CCD Analogy

# CMOS Image sensor

CMOS image sensors combine a Si photodiode with CMOS transistors to amplify the photovoltage.

Direct transfer of charge to voltage allows the device to perform a number of processing and control functions on chip, including timing, exposure control, analog-to-digital conversion, white balance and even some initial image processing algorithms.

CMOS image sensors are often described as ‘Active pixel’ devices

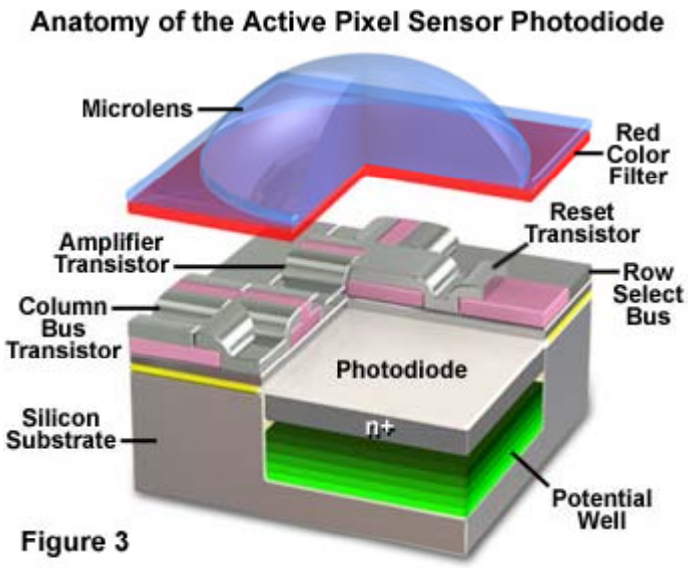
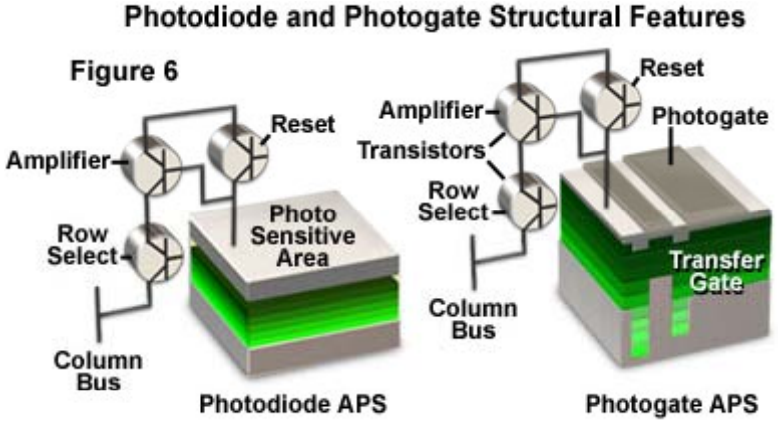


Figure 3



The University Of Sheffield.

# CCD versus CMOS image sensor

| | CCD | CMOS Image sensor |
|---|---|---|
| Pixel size | Limited | Very small: driven by CMOS developments |
| Frame rate | Limited by the speed of emptying its registers | Inherently fast & can increase speed by only outputting changing levels |
| Low light | Better signal to noise. All charge gets transferred with very little loss | More sensitive, but suffers from noise when voltage sampling |
| Electronic shuttering | Charge can only be emptied by reading | Data can be cleared without read out |
| Complexity | Read-out complexity: 5 or more supply voltages at different clock speeds .Significantly higher power consumption | Lower overall cost due to on-chip processing. Single-voltage power supply. |
| | | |