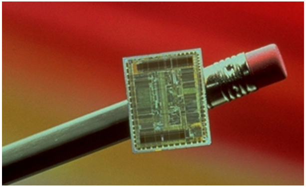
TELECOM ParisTech

INSTITUT Mines-Télécom



ICS904/EN2 : Design of Digital Integrated Circuits

From the bases of CMOS digital circuits to the economical choices. . .

Yves MATHIEU

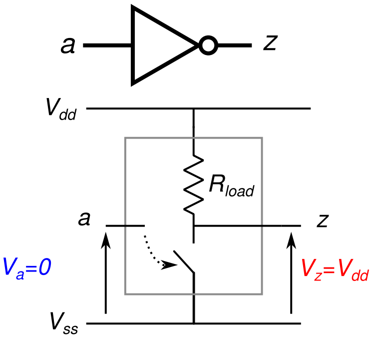
yves.mathieu@telecom-paristech.fr

université PARIS-SACLAY

Building a logic gate

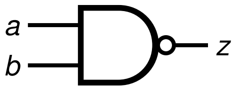
What do we need ?

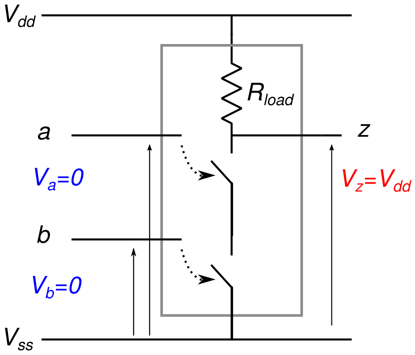
- A ground reference : V_{ss}
- A power supply : V_{dd}
- An electrical definition for logic values : $0 \equiv V_{ss}$, $1 \equiv V_{dd}$
- A resistive load : R_{load}
- A switch **controlled** by a voltage (referenced to V_{ss})
  - $V_{control} = 0$ Open switch
  - $V_{control} = V_{dd}$ Closed switch



Let's try to build a more complex gate . . .

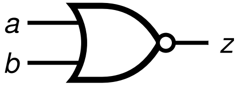
The two input NAND gate

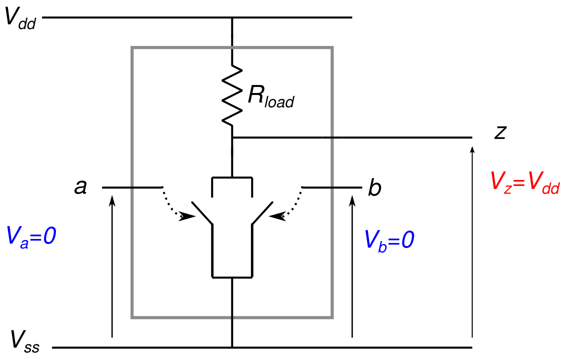




Let's try other gates. . .

The two input NOR gate





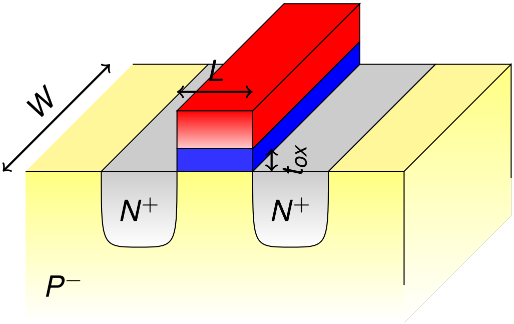
Very simple structures

but. . .

- A permanent current flows through the gate when the logic output is **0** :
  - The only usefull power consumption should be linked to the **activity** of gates not to their state. . .
- Physicists do not know how to realize ideals switches (at reasonable operating temperatures) :
  - The **0** logic level doesn't reach V_{ss}.
  - Safe operation of the gate is not garanteed.

The MOS transistor

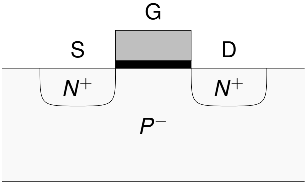
Metal Oxide Semiconductor



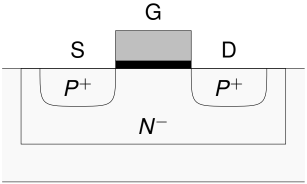
CMOS transistors

Complementary Metal Oxide Semiconductor

- Gate : G, Drain : D, Source : S, Threshold Voltage : V_T
- With $V_{TN} > 0$ and $V_{TP} < 0$



nMOS transistor



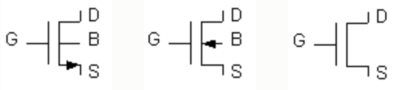
pMOS transistor

- N channel
- Electrons current
- Conduction if $V_{gs} > V_{TN}$

- P channel
- Holes current
- Conduction if $V_{gs} < V_{TP}$

Simplistic model of the NMOS transistor

Schematic Symbols



Cut-off region

$$\text{If } V_{GS} \leq V_{TN} \text{ then } I_{DS} = 0$$

Conduction region (Saturation region)

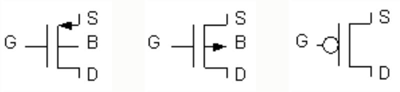
$$\text{If } V_{GS} > V_{TN} \text{ then } I_{DS_{max}} = K_n \cdot (V_{GS} - V_{TN})^2$$

Technological and geometrical factors

$$K_n = \frac{1}{2}\mu_{0N} \cdot C'_{ox}\frac{W_N}{L_N}$$

Simplistic model of the PMOS transistor

Schematic symbols



Cut-off region

$$\text{If } V_{GS} \geq V_{TP} \text{ then } I_{DS} = 0$$

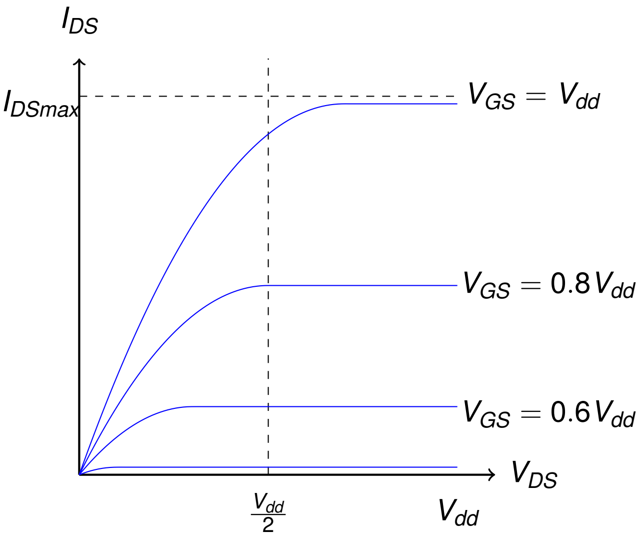
Conduction region (Saturation region)

$$\text{If } V_{GS} < V_{TP} \text{ then } I_{DS_{max}} = -K_p \cdot (V_{GS} - V_{TP})^2$$

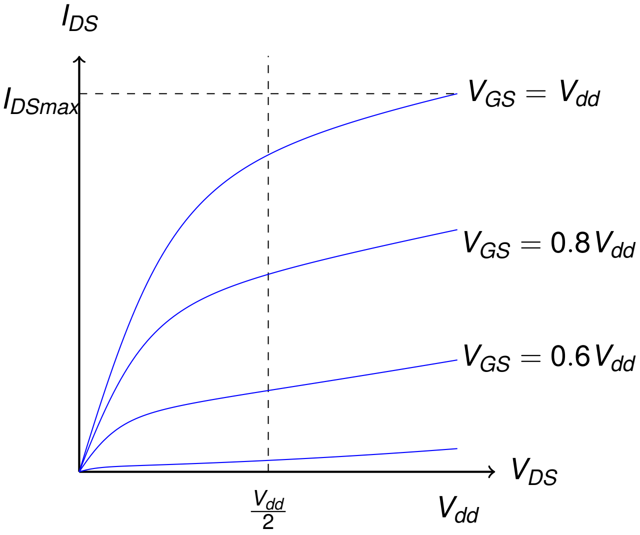
Technological and geometrical factors

$$K_P = \frac{1}{2}\mu_{0P} \cdot C'_{ox} \frac{W_P}{L_P}$$

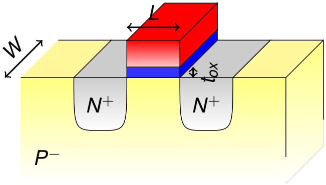
$I_{DS} = f(V_{GS}, V_{DS})$



28nm node. NMOS transistor



T_{ox} : gate oxide thickness



$C_{ox} = 1.15fF$

$R_{dsON} = 510.00\Omega$

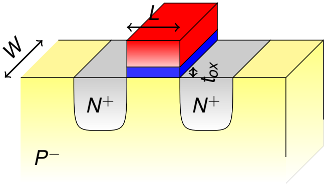
$I_{dsmax} = 880.00\mu A$

$L = 65.00nm$

$W = 1000.00nm$

$t_{ox} = 1.30nm$

L : Gate Length



$C_{ox} = 1.15fF$

$R_{dsON} = 510.00\Omega$

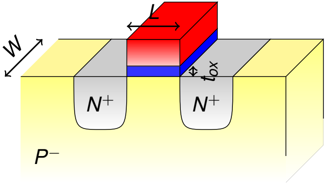
$I_{dsmax} = 880.00\mu A$

$L = 65.00nm$

$W = 1000.00nm$

$t_{ox} = 1.30nm$

W : Gate Width



$C_{ox} = 1.15fF$

$R_{dsON} = 510.00\Omega$

$I_{dsmax} = 880.00\mu A$

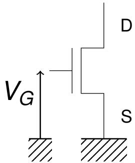
$L = 65.00nm$

$W = 1000.00nm$

$t_{ox} = 1.30nm$

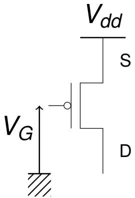
MOS transistors and logic levels

Two non ideal electronic switches



nMOS Transistor with Source connected to Ground.

- $V_G = V_{ss}$
  $\Rightarrow$ open switch
- $V_G = V_{dd}$
  $\Rightarrow$ closed switch



pMOS Transistor with Source connected to power supply

- $V_G = V_{ss}$
  $\Rightarrow$ closed switch
- $V_G = V_{dd}$
  $\Rightarrow$ open switch

CMOS logic

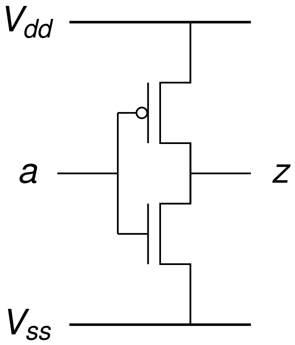
CMOS invertor

- Boolean input value $a = 0$
  - $\rightarrow$ $V_a = 0$
    - $\rightarrow$ nMOS cutoff
    - $\rightarrow$ pMOS conducting
  - $\rightarrow$ $V_z = V_{dd}$

$\rightarrow$ boolean output value $z = 1$

- boolean input value $a = 1$
  - $\rightarrow$ $V_a = V_{dd}$
    - $\rightarrow$ nMOS conducting
    - $\rightarrow$ pMOS cutoff
  - $\rightarrow$ $V_z = 0$

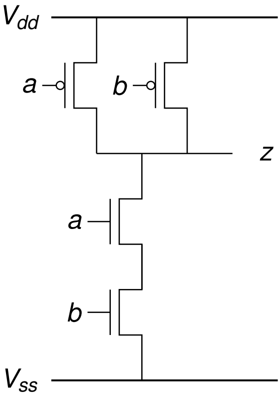
$\rightarrow$ Boolean output value $z = 0$



No static power consumption (first order approximation)

CMOS logic

The two input NAND gate



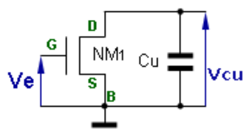
MOS switches efficiency

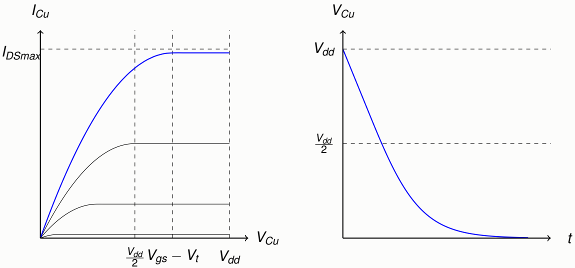
CMOS gates : parasitic capacitive loads

- Internal nodes of a logic gate have parasitic capacitances :
  - gate-source or gate-drain capacitance (gate oxyde).
  - source-bulk ou drain-bulk capacitance (reverse biased diodes)
- The output of a logic gate is connected to :
  - Wires (parasitic capacitance to the ground, supply, or between nodes)
  - Inputs of other logic gates (transistors gate node)
- All these item may be summed to an equivalent unique parasitic capacitance C_{par} :
  - that should be charged during rising edges of the gate output
  - that should be discharged during falling edges of the gate output
- The computation time of a gate is directly linked to the charge et de discharge times.

is the NMOS transistor an efficient switch ?

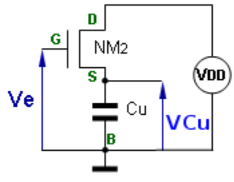
NMOS transistor discharging a capacitor C_u

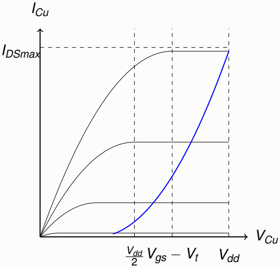


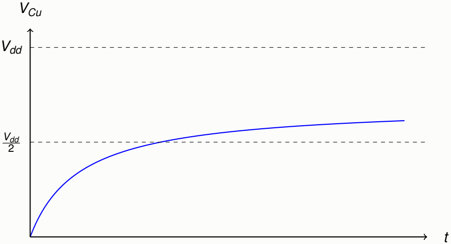


is the NMOS transistor an efficient switch ?

NMOS transistor charging a capacitor C_u

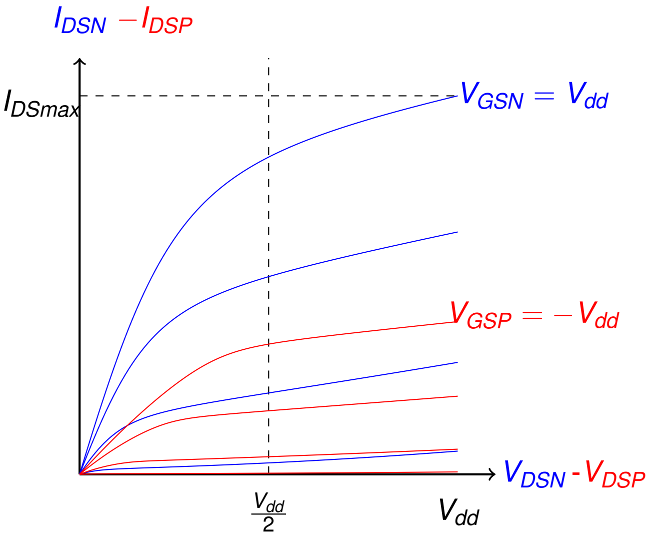






NMOS versus PMOS

Equal geometries transistors



NMOS versus PMOS

- NMOS drain-source current is a electron current.
- PMOS drain-source current is a hole current.
- Charge mobility are different because componants are different
- With identical geometries and identical voltages (V_{gs}, V_{ds}), I_{ds} currents are different.
- PMOS transistor is less efficient than NMOS transistor

Stacked switches : body effect

- Threshold voltage value is linked to V_{sb} voltage

switches efficiency ?

A few common-sense rules

- NMOS preferred usage is discharging capacitors (generation of logic zeros)
- PMOS preferred usage is charging capacitors. (generation of logic ones)
- For equal efficiency, PMOS transistors should have larger width than NMOS transistors.
- Transistors stacks should be limited to 3 or 4 transistors.

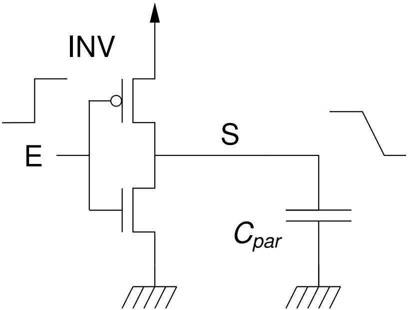
Performance criterions

- **Area/cost** :
  The smaller the chip is, the better the efficiency of production and therefore the lower the manufacturing cost.
  - Using smaller transistors (technology evolution)
  - Using less transistors (architectural choices)
- **Speed** :
  Faster logic gates implies larger processing power.
  - How to increase the clock frequency ?
- -> **Power consumption** :
  Computation means power consumption.
  - How to minimize this power consumption ? (Internet Of Things) . . .)
  - How to evacuate the dissipated heat (Servers for cloud). . .)

Computation time of a logic gate

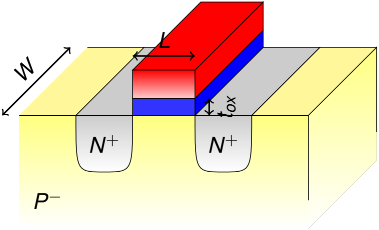
The simple case of a rising edge at the input of an invertor

- Hypothesis 1 : The rising edge has a null duration.
- Hypothesis 2 : The only parasitic capacitance taken into account is the gate capacitance.
- Hypothesis 3 : The current flowing through the transistors for charge or discharge of parasitic capacitance C_{par} is roughly equal to $I_{DS_{max}}$



Computation time of a logic gate

MOS transistor



Current through the conducting transistor

$$I_{DS_{max}} = K_n \cdot (V_{dd} - V_{TN})^2 \text{ with } K_n = \frac{1}{2}\mu_{0N} \cdot C'_{ox}\frac{W_N}{L_N}$$

Parasitic capacitance of the transistor gate

$$C_{ox} = C'_{ox} W_N \cdot L_N$$

Computation time of a logic gate

Computation time of an invertor

Current equation for the parasitic capacitance

$$I_{C_{par}} = C_{par} dV_{C_{par}} / dt$$

The NMOS transistor acts as a current source

$$I_{C_{par}} \approx I_{DSmax} = K_n \cdot (V_{dd} - V_{tn})^2$$

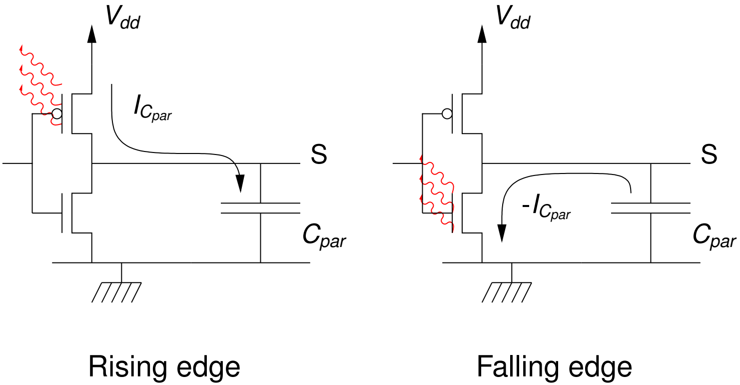
Discharge from V_{dd} to 0

$$t_{comp} = C_{par} \frac{\Delta V}{I_{DSmax}} = C_{par} \frac{V_{dd}}{K_n \cdot (V_{dd} - V_{tn})^2}$$

Encreasing the power-supply voltage in order to increase speed (overclocking) ? (bad way)

Power consumption of CMOS logic

Dissipated energy versus stored energy



Power consumption of CMOS logic

Energy balance

Charging : Energy comes from the power supply

$$E_{V_{dd}} = C_{par} \int_0^{V_{dd}} V_{dd} \, \mathrm{d}V_s = C_{par} {V_{dd}}^2$$

Discharging : Stored energy in the capacitance

$$E_{Cpar} = C_{par} \int_0^{V_{dd}} V_s \, \mathrm{d}V_s = C_{par} \frac{{V_{dd}}^2}{2}$$

$C_{par} \frac{{V_{dd}}^2}{2}$ dissipation whatever the edge

Power consumption of a full chip

- Let C_{chip} be the overall parasitic capacitance of the chip.
- Let F_{clk} be the operating frequency of the chip clock (synchronous logic)
- Let T_{act} (activity) be the mean transition probability of signals during a single cycle of the clock ($T_{act} \approx 0.3$)

Overall power consumption of the chip

$$P_{circuit} \approx T_{act} F_{clk} C_{chip} {V_{dd}}^2$$

What do you think now of overclocking ?

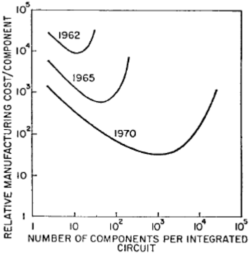
A bit of history

Moore's "law(s)"

Intel 1969 - 106 employees (2015 - 80000 employees)
https://commons.wikimedia.org/wiki/File:Intel_Mountain_View_in_1969.jpg

A bit of history

Moore's "law(s)"



- Gordon Moore, cofounder of Intel.
- Gordon Moore "Cramming More Components onto Integrated Circuits," Electronics, pp. 114–117, April 19, 1965.
- 1965 : « The complexity for minimum component costs has increased at a rate of roughly a factor of two per year »

http://www.cs.utexas.edu/~fussell/courses/cs352h/papers/moore.pdf
http://www.intel.com/content/www/us/en/history/museum-gordon-moore-law.html

A bit of history

Moore's "law(s)"

- This empirical observation became a prediction.
- This prediction became a roadmap for silicon foundries.
  - Research and development expenditure adjustment. . .
  - Investments expenditure adjustment for foundries. . .
  - . . . in order to follow the roadmap.
- Moore'law widened to other key parameters :

  Processing power of . . . double every . . . years
  Power consumption of . . . is divided by two every . . . years

Moore's laws were exponential laws, followed during more than four decades.

Technology evolution model

"Theoritical downsizing"

- Technology "nodes" :
  - A technology node is defined by the minimum gate length of the transistor (90nm, 65nm, 40nm, 28nm, . . .)
  - For each new node silicon founders try to reduce the transistor area with a factor of **2**
  - *Foundries* are investing billions of dollars in order to follow this objective . . .
- A linear reduction factor of $\beta = \sqrt{2}$ is used :
  - The width W and the length L of transistors are divided by β.
  - The oxide thickness T_{OX} is divided by β.
  - The power supply voltage V_{dd} is divided by β.
  - The threshold voltage V_T of the transistor is divided by β.

Theoritical downsizing

Performance evolutions

Parasitic capacitances as a function of β

$$C_{par}(\beta) = (W/\beta)(L/\beta)(\beta C'_{ox}) = \frac{C_{par}}{\beta}$$

Energy consumption of a gate as a function of β

$$E_{gate}(\beta) = \frac{C_{par}}{\beta}(\frac{V_{dd}}{\beta})^2 = \frac{E_{gate}}{\beta^3}$$

Computation time of a gate as a function of β

$$t_{comp}(\beta) = \frac{t_{comp}}{\beta}$$

Theoritical downsizing

Reducing costs and power consumption

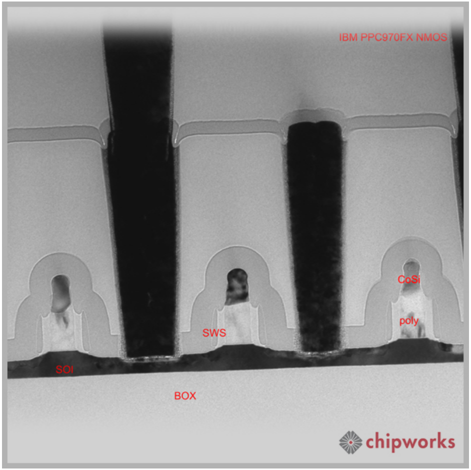
- Keeping the clock frequency constant.
  - $F_{clk}(\beta) = F_{clk}$
- The area reduction implies a price reduction
  - $Area(\beta) = \frac{Area}{\beta^2}$
- Power consumption is lower.
  - $P_{chip}(\beta) = T_{act} F_{clk} \frac{E_{chip}}{\beta^3} = \frac{P_{chip}}{\beta^3}$
- This strategy is particularly interesting for mobile systems :
  - Transition from high-end devices to low-end devices (smartphones),
  - New usages for ultra-low power devices (Internet Of Thinks).

Theoritical downsizing

Enhancing computational power

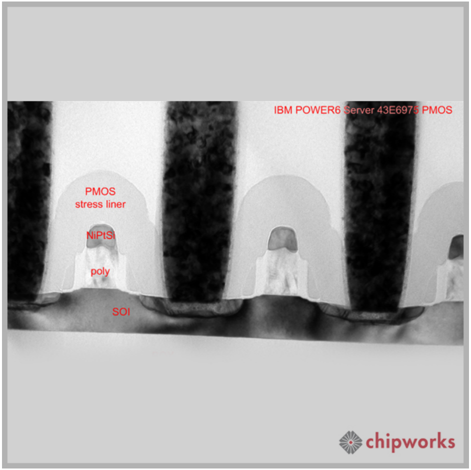
- Using maximum achievable frequency
  - $F_{clk}(\beta) = \beta F_{clk}$
- Using maximum achievable complexity (more transistors with the same area)
  - $Area(\beta) = Area$
- Power consumption doesn't change
  - $P_{chip}(\beta) = P_{chip}$
- This strategy is usefull for server CPUs.
  - Computational power is enhanced using higher frequencies.
  - Computational power is enhanced using parallelism.

MOS Transistor : 2004



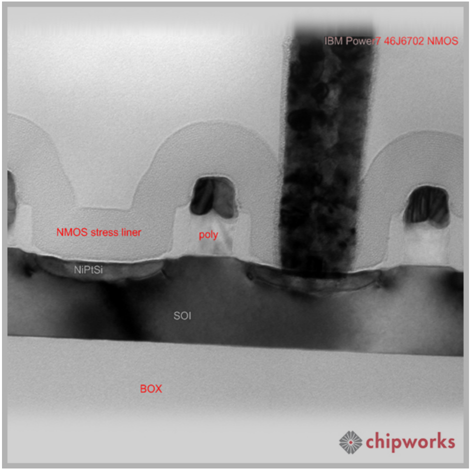
PPC970fx (90nm)

MOS Transistor : 2008



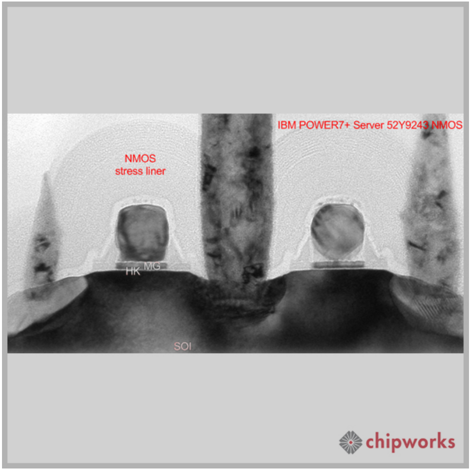
Power6 (65nm)

MOS Transistor : 2011



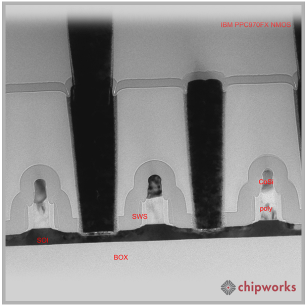
Power7 (45nm)

MOS Transistor : 2013

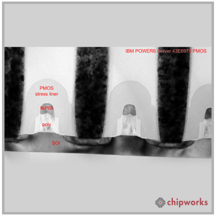


Power7+ (32nm)

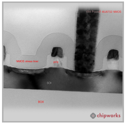
Using the same scale



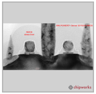
2004
90nm
PPC970fx



2009
65nm
Power6



2011
45nm
Power7



2013
32 nm
Power7+

Photo Credits

The images are from the analysis of the evolution of IBM technologies made by Chipworks Inc.
The analysis as well as the original images were available in 2014 here :
`http://www.chipworks.com/en/technical-competitive-analysis/resources/blog/ibm-continues-major-source-chip-innovation/`

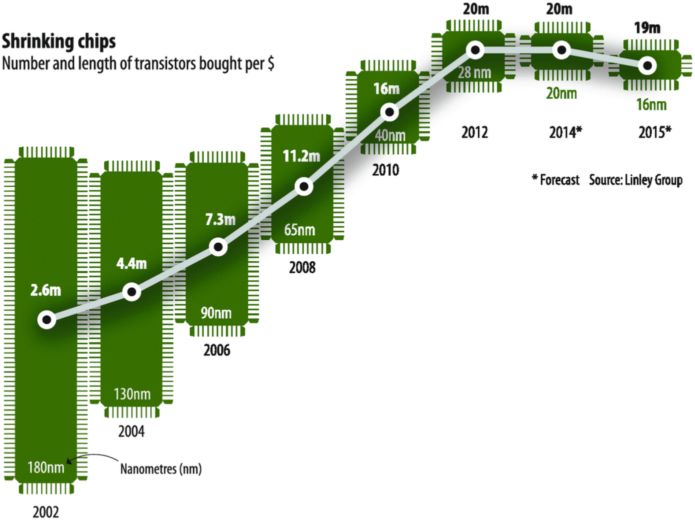
Practical downsizing

What are today problems ?

- For CPU, frequencies have reached their maximal values (from 3 to 4 GHz) at the beginning of the century. This is due to the maximum heat dissipation of the chips.
- When lowering the power-supply voltage we can no longer reach the "ideal switch" approximation : chips have larger and larger static dissipation added to the computation dissipation.
- Technologists must use more and more complex (costly) manufacturing processes to continue to follow the "Moore's Law"..
- At the end of the previous century, some predicted the end for "Moore's law" for scientific reasons (MOS transistor physics), it seems, since 2014, that the main difficulty is economical.

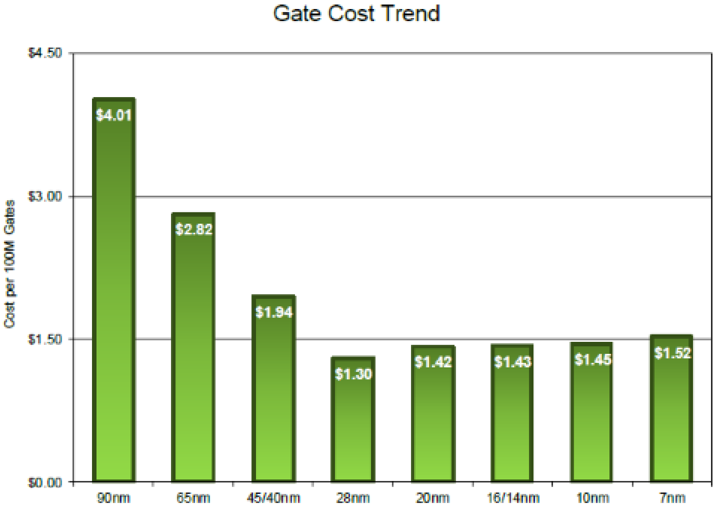
Technology downsizing

(forecast 2013) The end of Moore's law ?



Technology downsizing

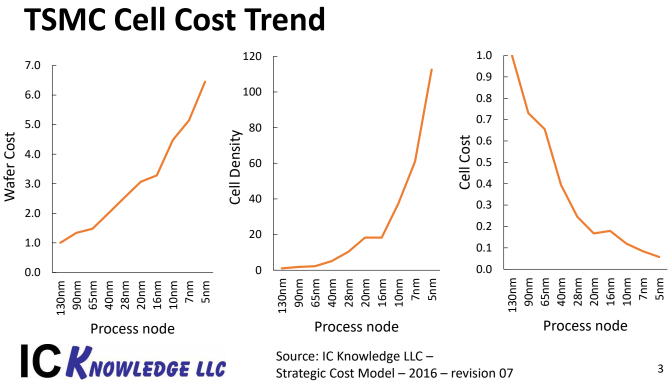
(forecast 2016) Controversy continues



Source: International Business Strategies, Inc.

Technology downsizing

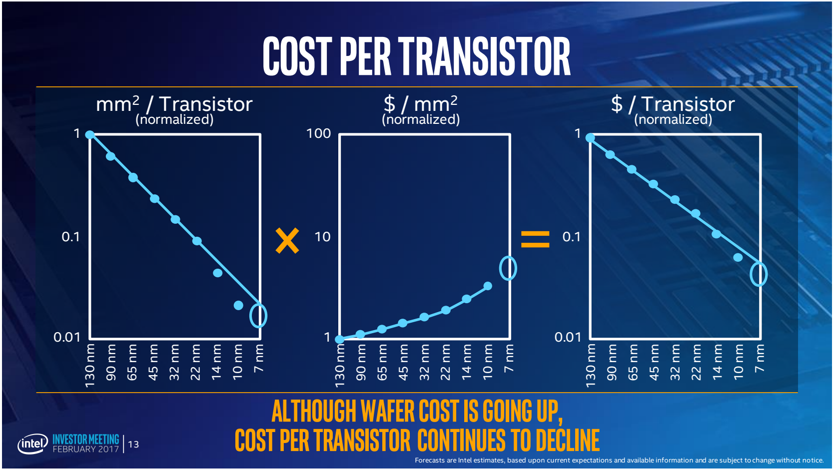
(2016) TSMC technologies



TELECOM ParisTech

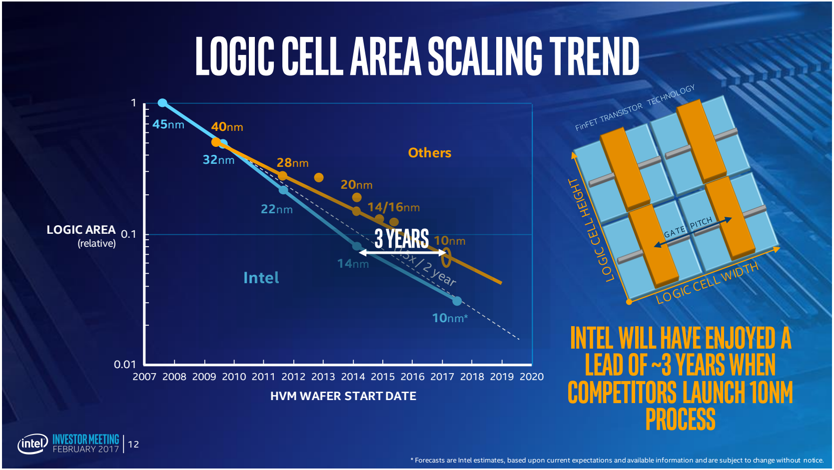
Technology downsizing

(2017) Intel Investor Meeting



Technology downsizing

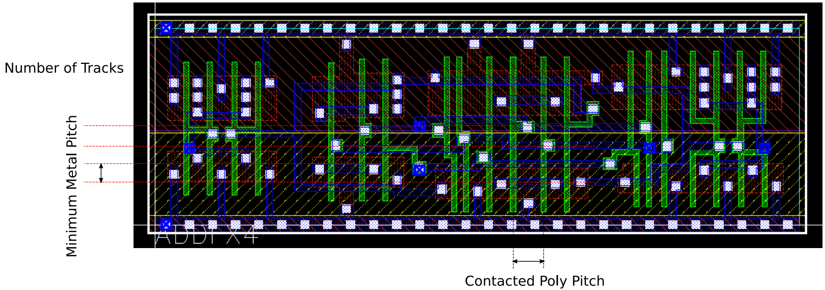
(2017) Gate length is no longer a good metric



Technology downsizing

Concept of "Standard node"

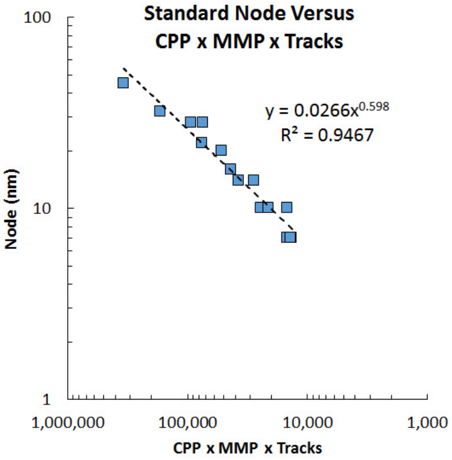
- $Cell\_area \propto CPP * MMP * Tracks$



Technology downsizing

"Standard Node Versus area formula"

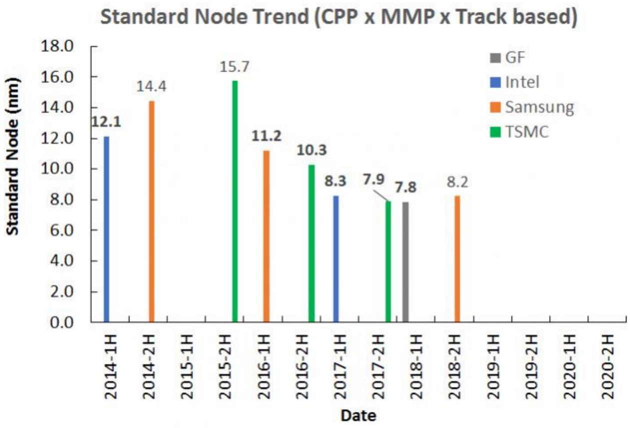
- source :https ://www.semiwiki.com
- 54 processes from 12 companies

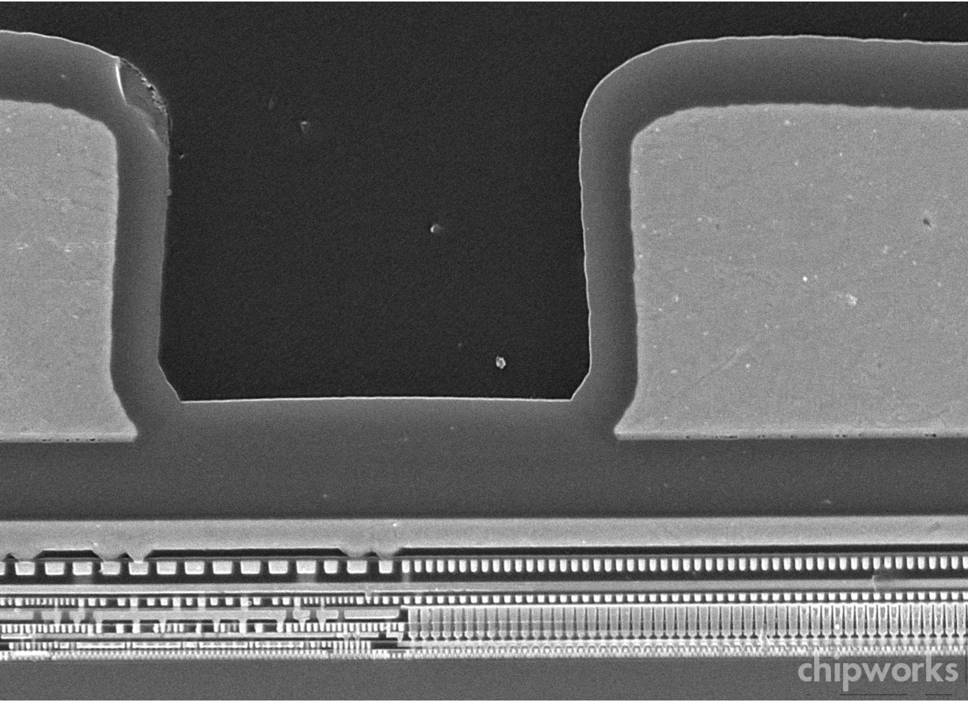


Technology downsizing

"Standard Node By Company"

- source :https ://www.semiwiki.com
- Annoucements : Intel(14nm, 10nm) Tsmc(16nm, 10nm, 7nm)





Gordon Moore Fishing

source https://commons.wikimedia.org/wiki/File:Gordon_moore_fishing.jpg