The Abdus Salam
International Centre for Theoretical Physics

United Nations
Educational, Scientific
and Cultural Organization

International Atomic
Energy Agency

310/1780-17

ICTP-INFN Advanced Tranining Course on
FPGA and VHDL for Hardware Simulation and Synthesis
27 November - 22 December 2006

Basic Gates

Jorgen CHRISTIANSEN
PH-ED
CERN
CH-1221 Geneva 23
SWITZERLAND

BASIC CMOS digital building blocks

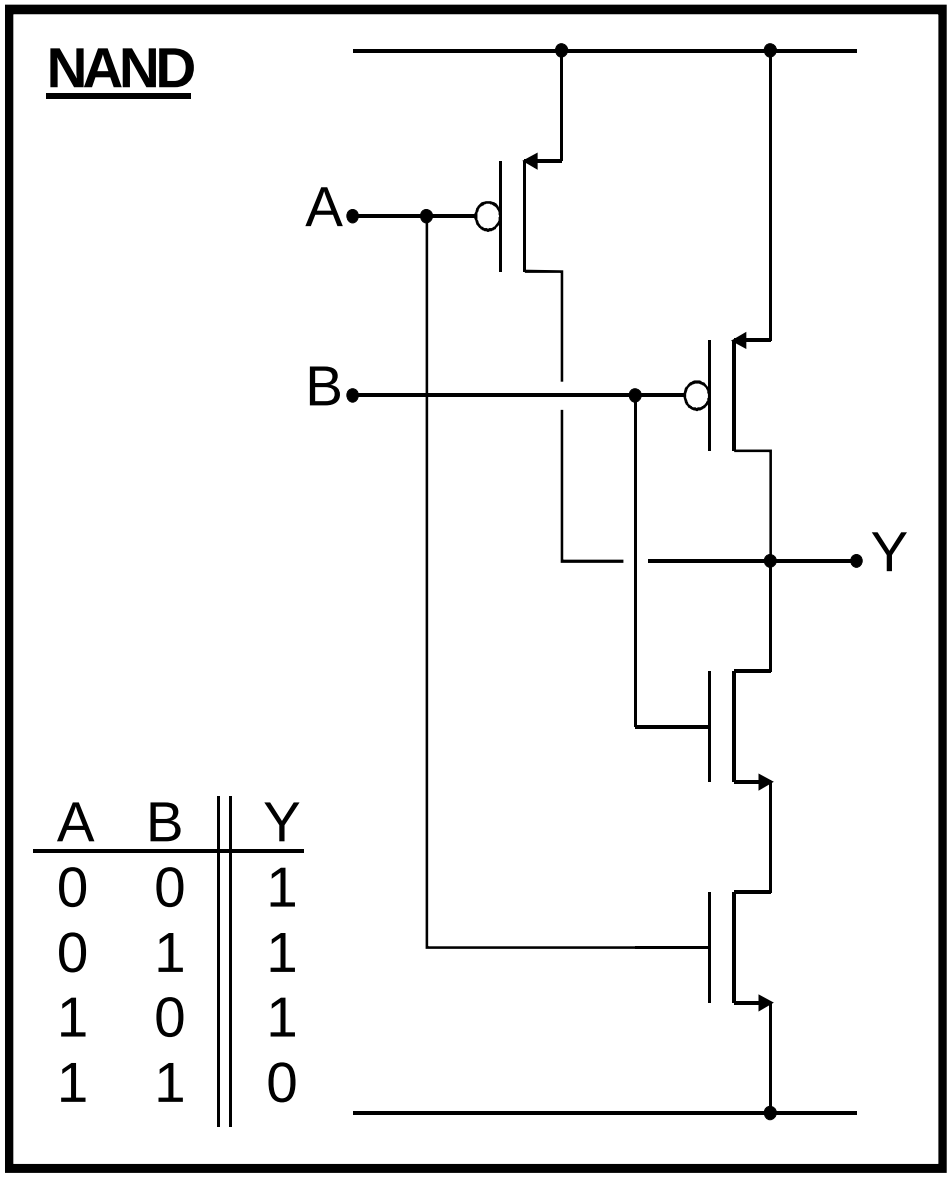
Paulo Moreira &
Jorgen Christiansen
CERN - Geneva, Switzerland

This part is compressed set of transparencies
from Paulo Moreira:
http://paulo.moreira.free.fr/

Outline

- Part 2: BASIC CMOS digital building blocks (from Paulo Moreira)
  - Gates: NAND, NOR, PASS
  - Sequential: Latch, Flip-Flop
  - Interconnects
  - Memory: ROM, RAM, PROM, FLASH , ,
  - A bit about Delay and phase locked loops

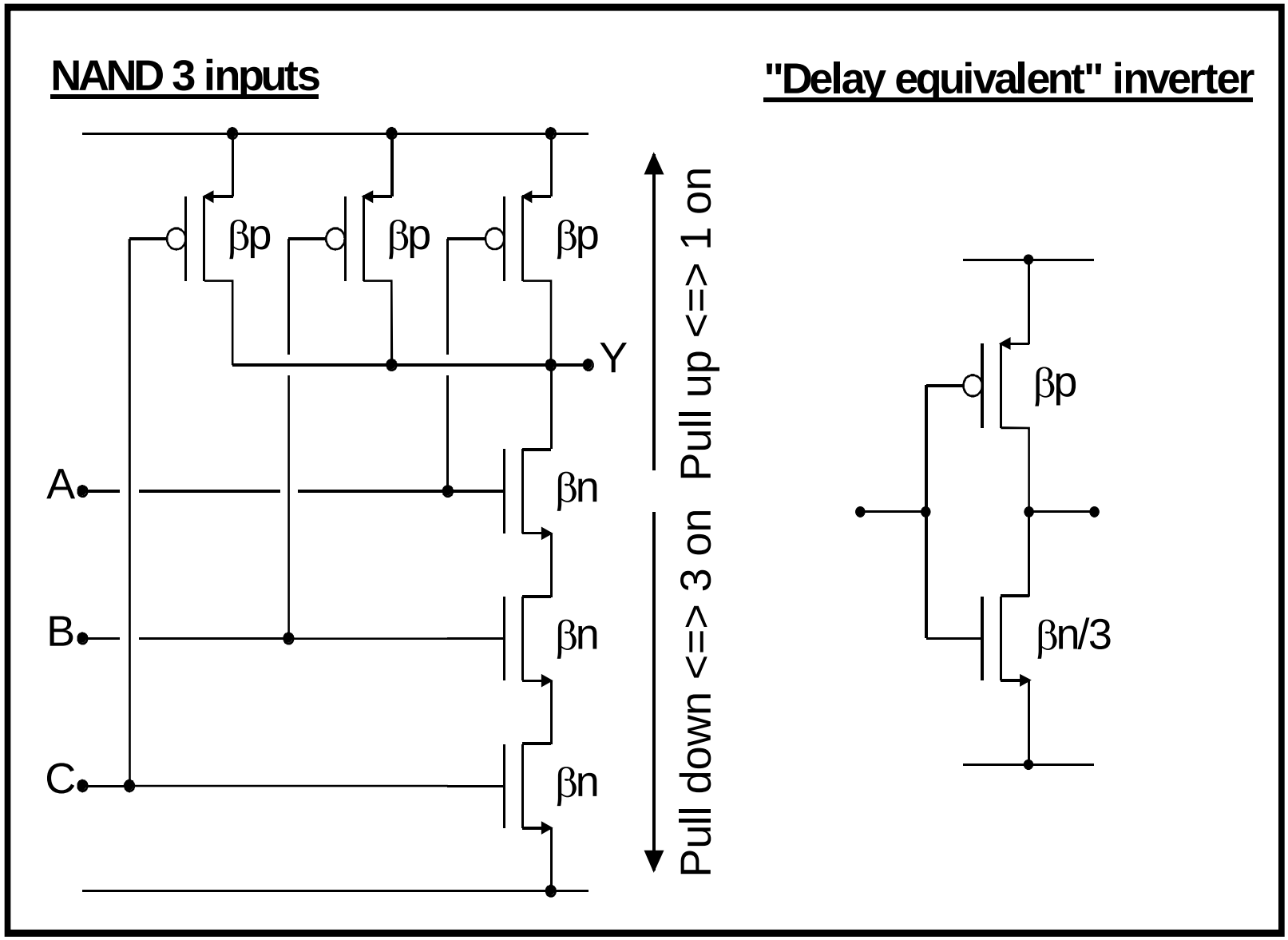
NAND 2-inputs



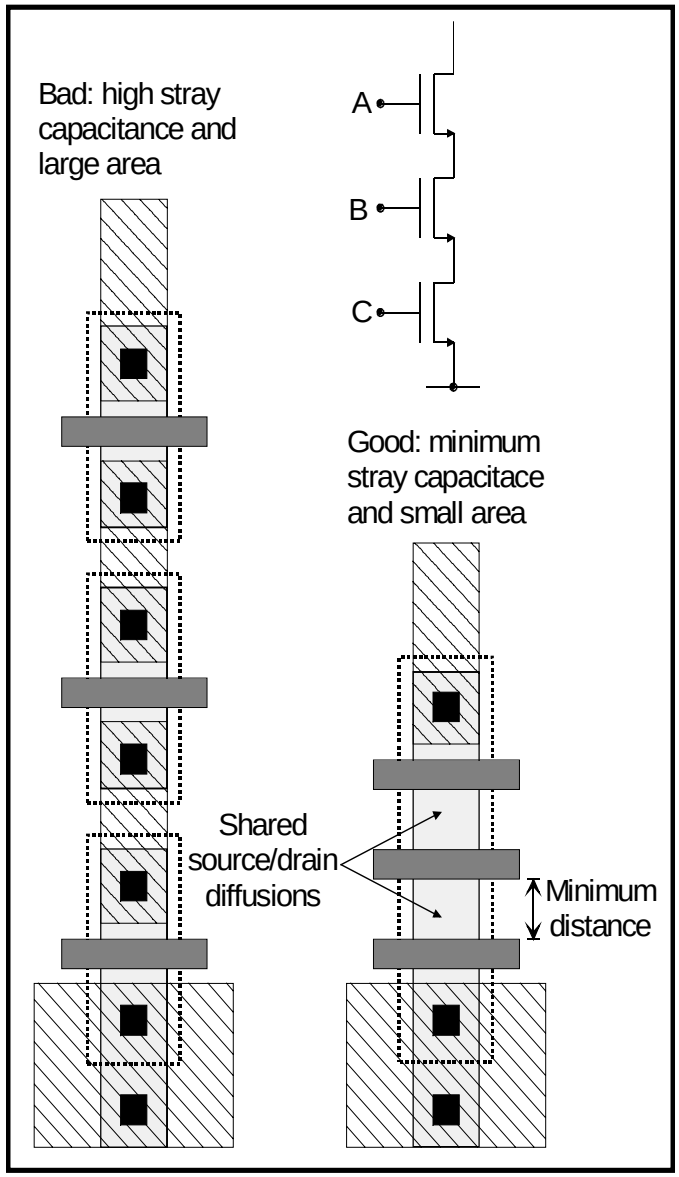
A	B	Y
0	0	1
0	1	1
1	0	1
1	1	0

"Gates are inverters in disguise!"

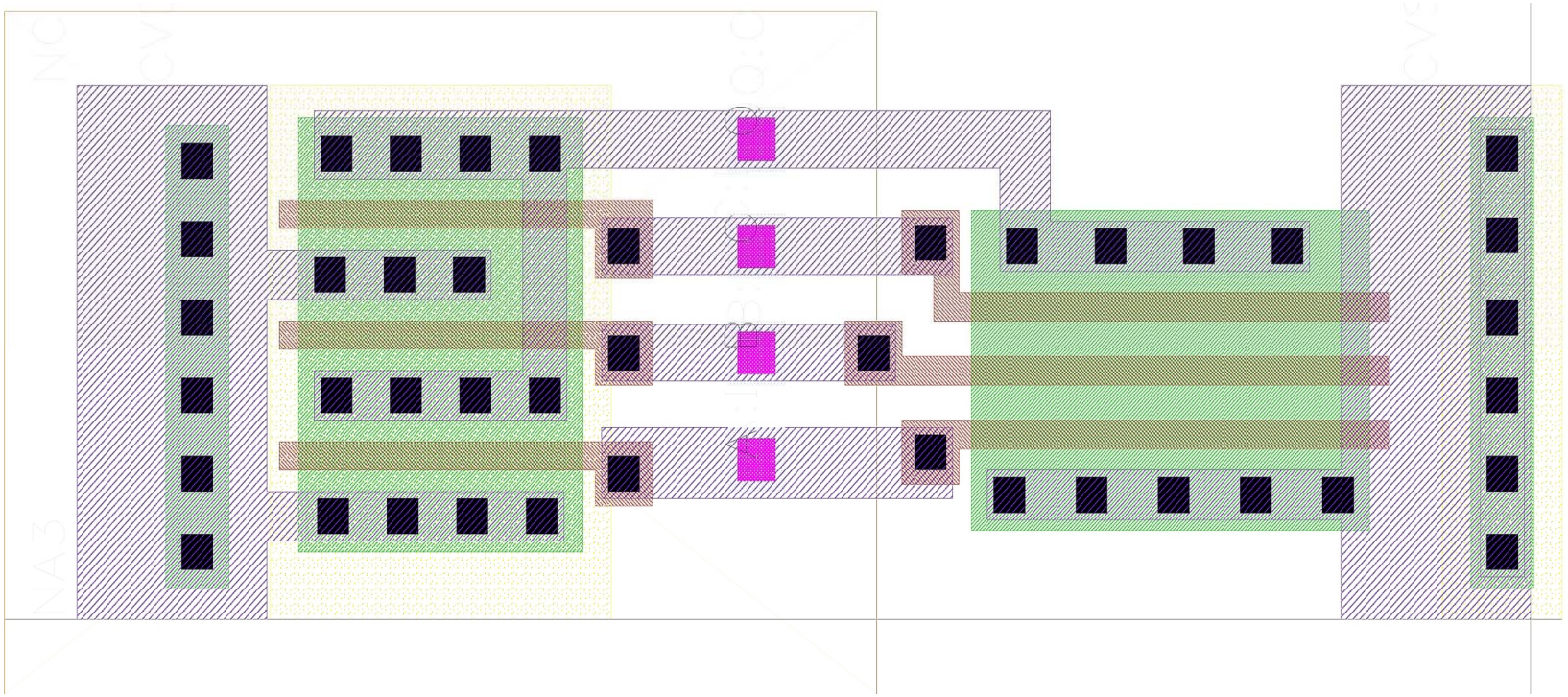
NAND 3-inputs



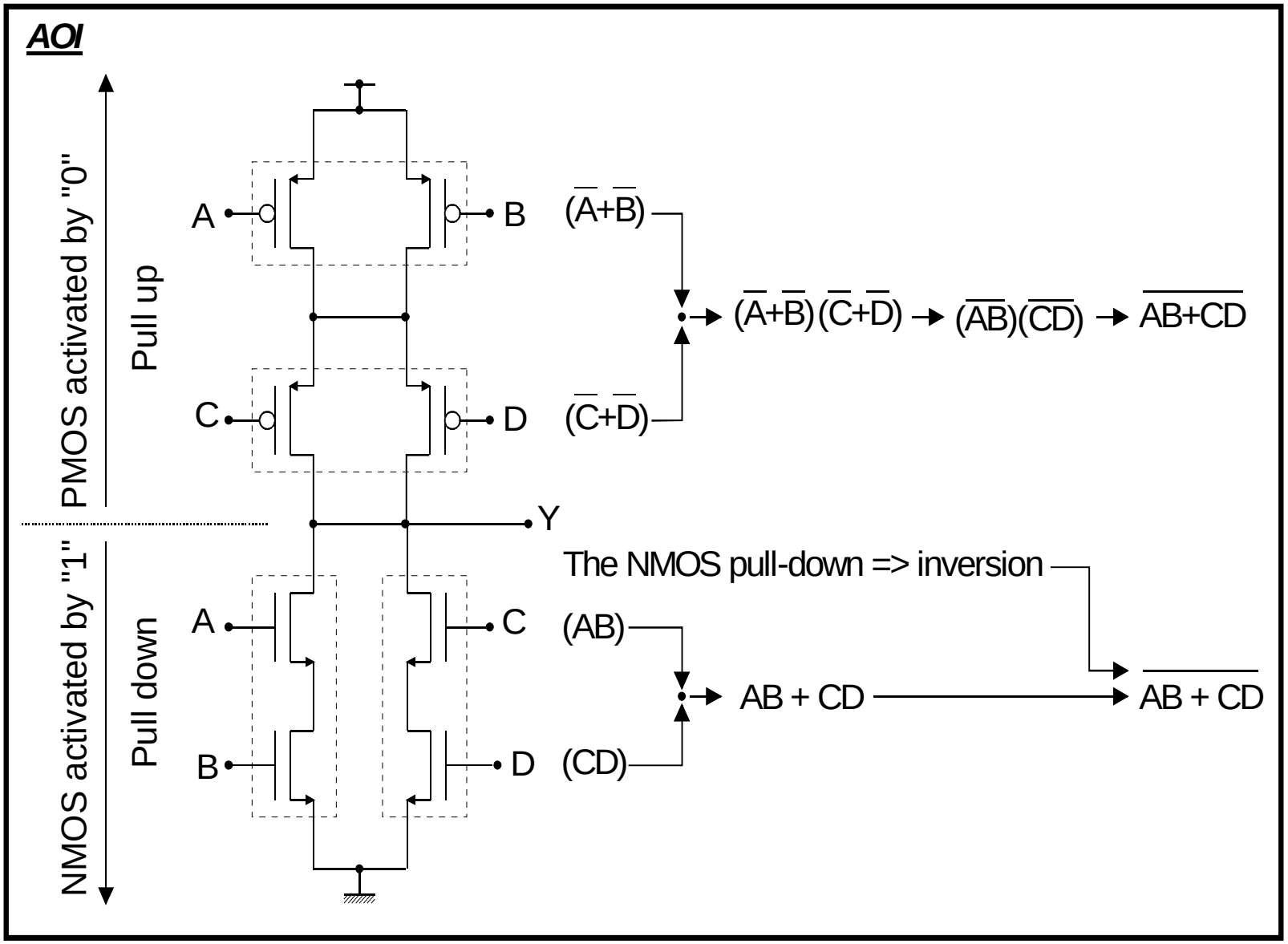
NAND 3-inputs



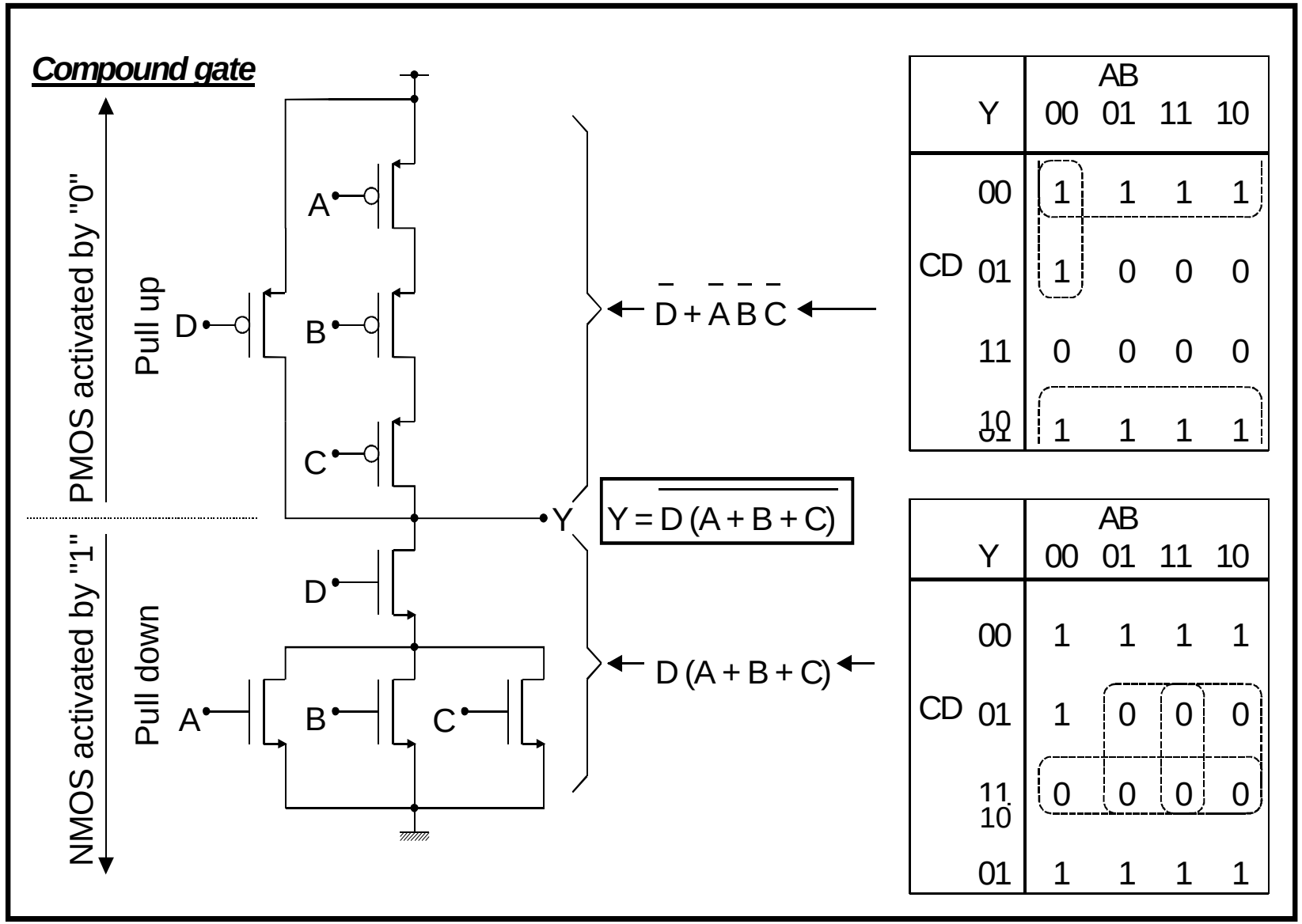
NAND 3-inputs



“Reading” CMOS gates



Designing CMOS gates



Complex CMOS gates

- Can a compound gate be arbitrarily complex?
  - NO, propagation delay is a strong function of fan-in:

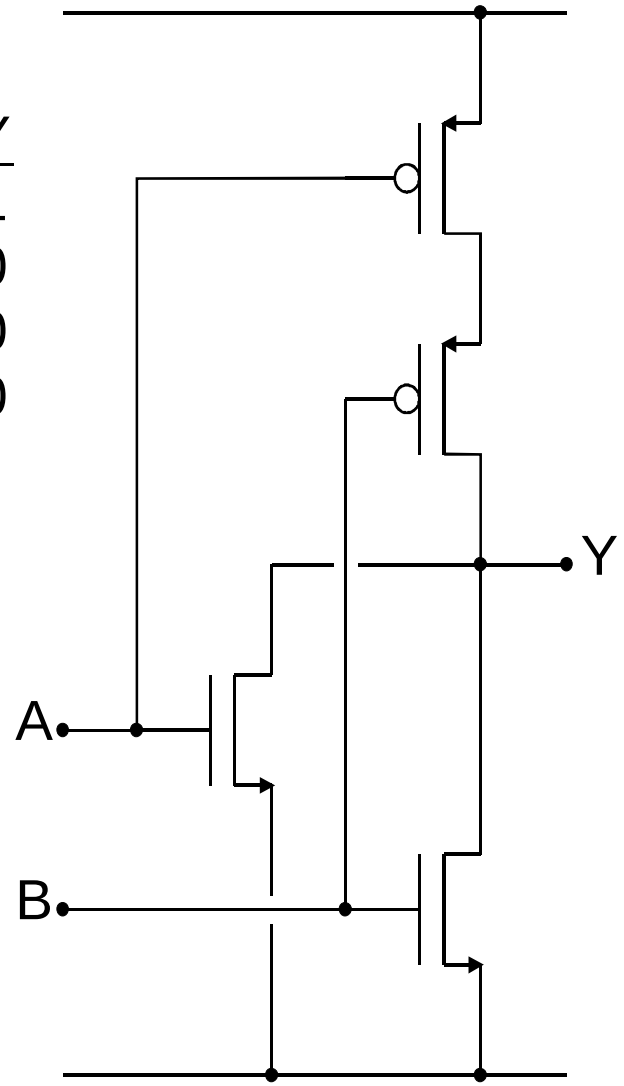
$$t_p = a_0 \cdot FO + a_1 \cdot FI + a_2 \cdot (FI)^2$$

  - FO $\Rightarrow$ Fan-out, number of loads connected to the gate:
    - 2 gate capacitances per FO + interconnect
  - FI $\Rightarrow$ Fan-in, Number of inputs in the gate:
    - Quadratic dependency on FI due to:
      - "RC" delay formed by MOST channels resistance and the capacitance of the source and drain diffusions
  - Avoid large FI gates (Typically FI $\leq$ 4)

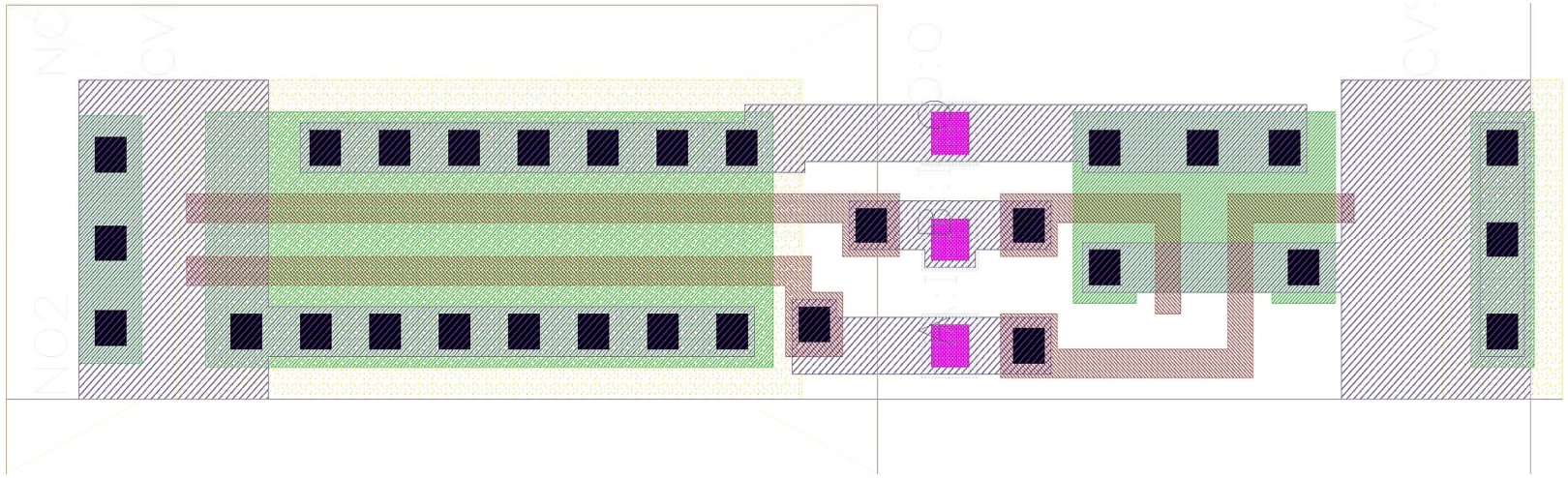
NOR 2-inputs

NOR

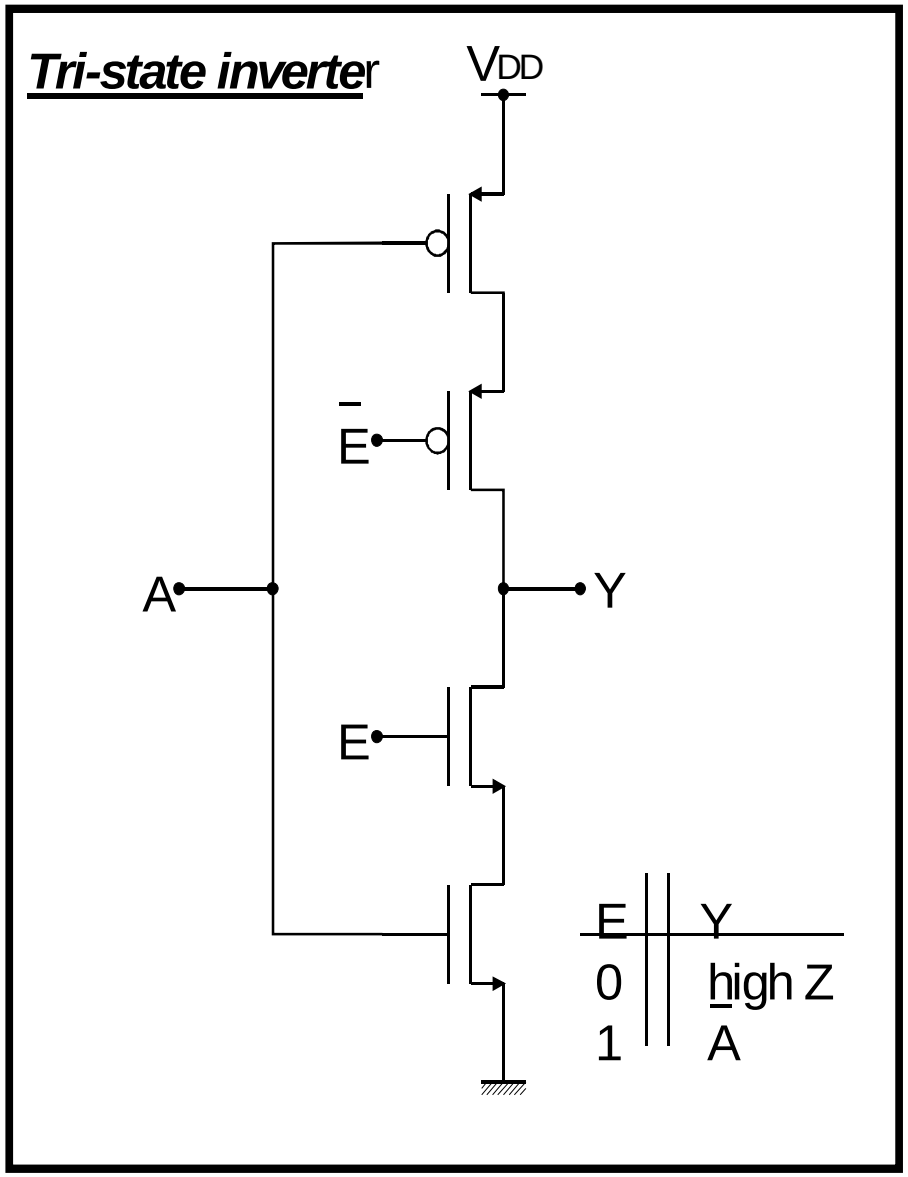
A	B	Y
0	0	1
0	1	0
1	0	0
1	1	0



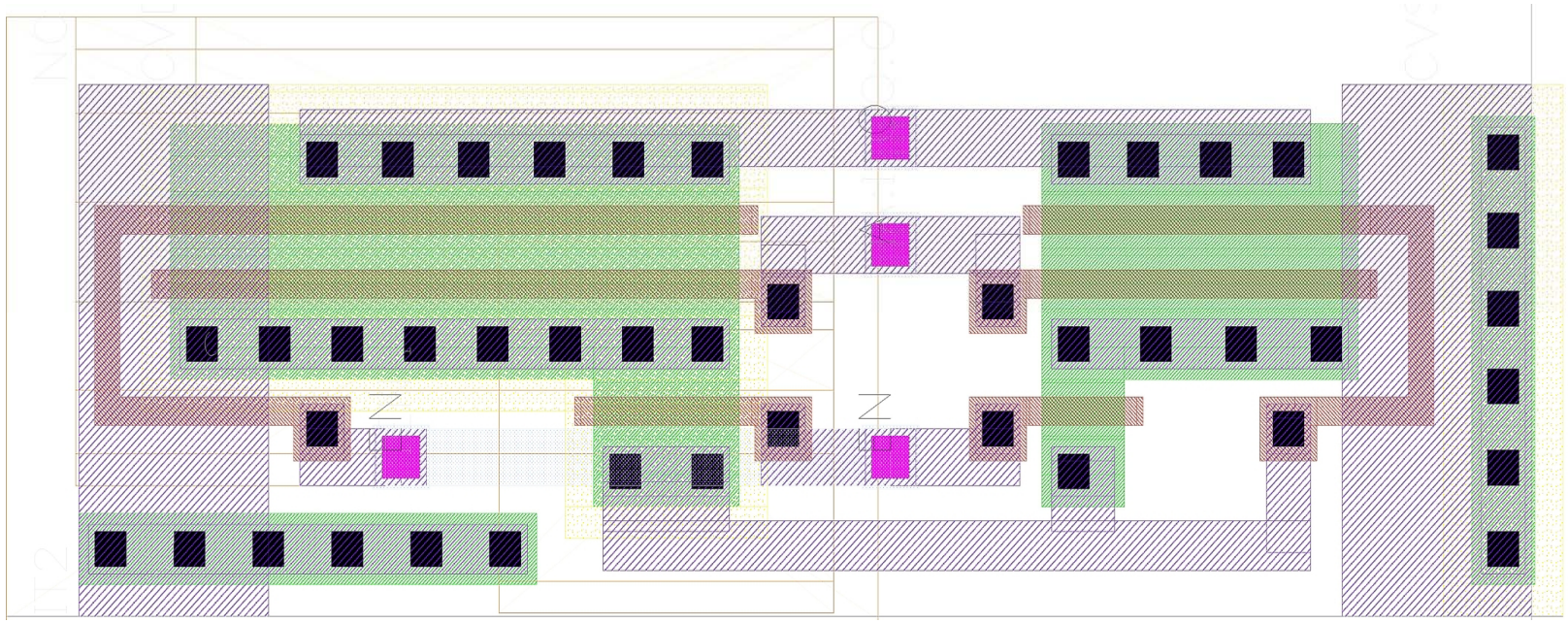
NOR 2-inputs



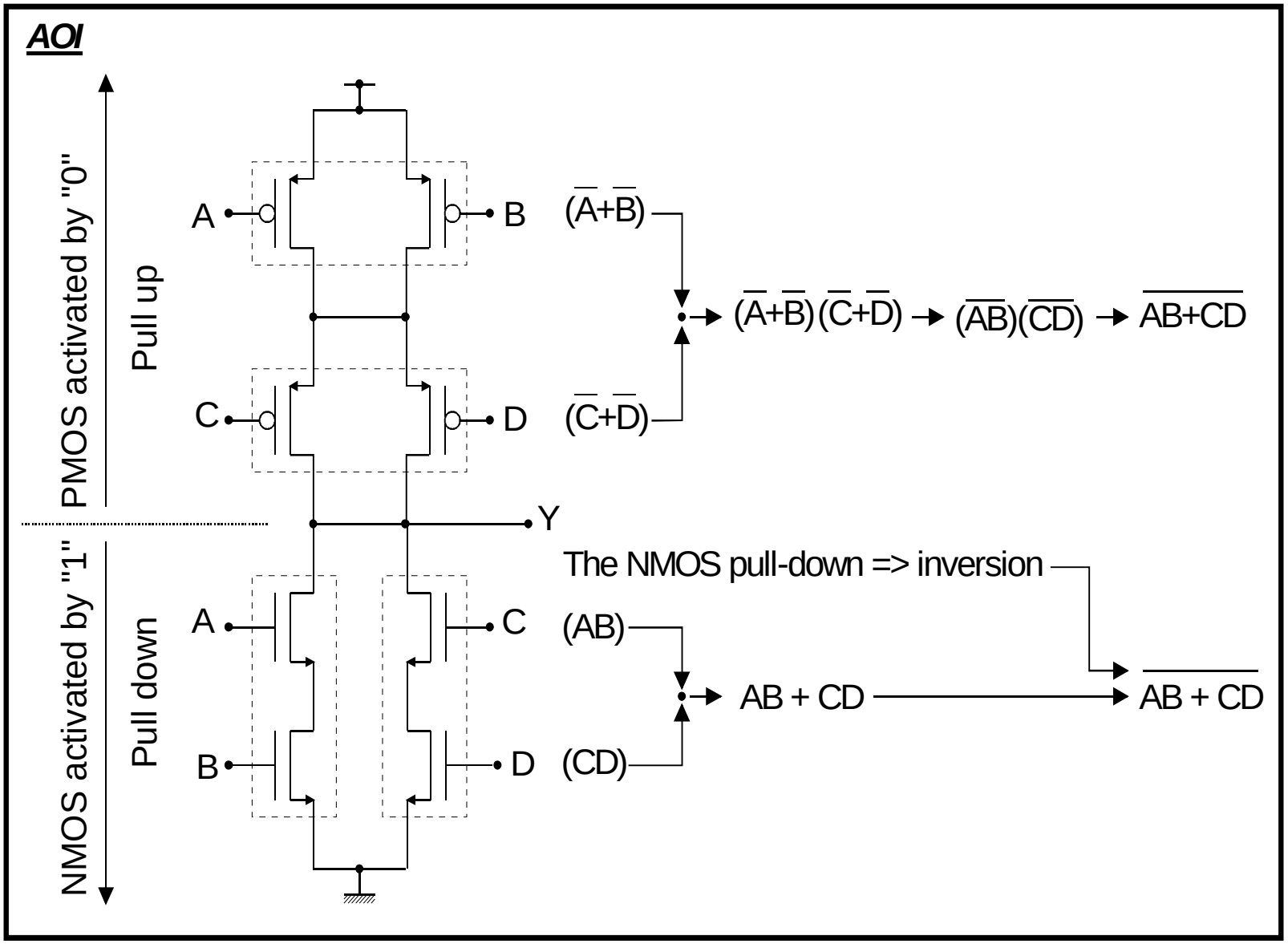
Tri-State Inverter



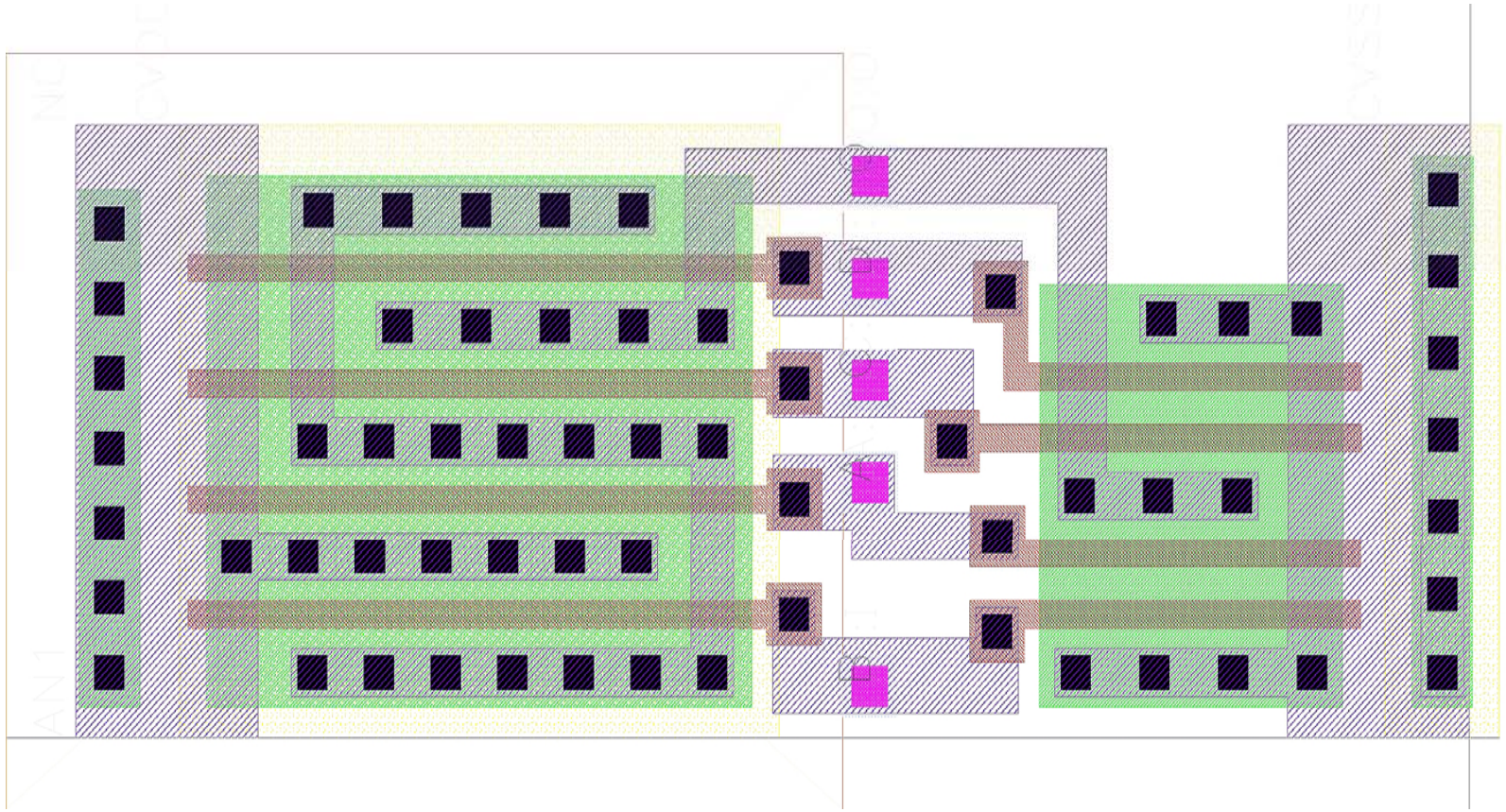
Tri-State Inverter



AOI 4-inputs


AOI
PMOS activated by "0"
Pull up
NMOS activated by "1"
Pull down
A
B
C
D
Y
$(\overline{A}+\overline{B})$
$(\overline{C}+\overline{D})$
$(\overline{A}+\overline{B})(\overline{C}+\overline{D}) \rightarrow (\overline{AB})(\overline{CD}) \rightarrow \overline{AB+CD}$
The NMOS pull-down => inversion
(AB)
(CD)
AB + CD
$\overline{AB + CD}$

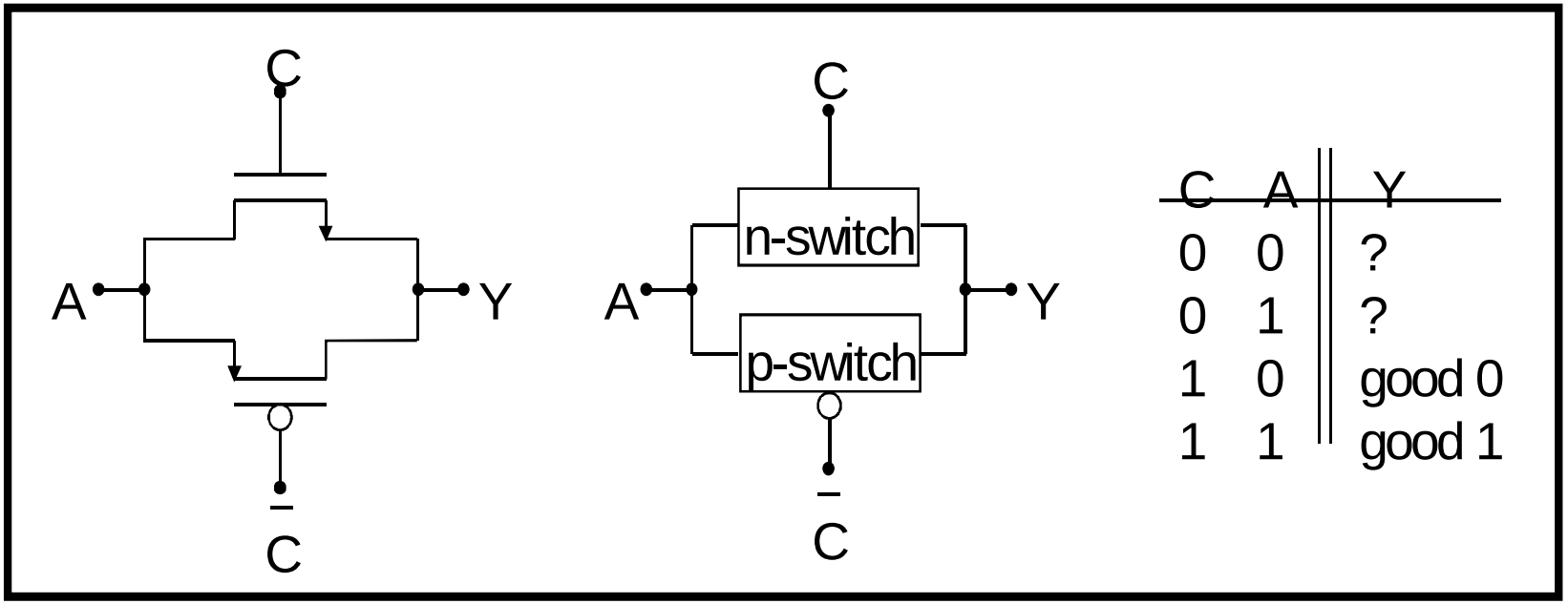
Complex CMOS gates: AOI



“An useful complement”

- The pass gate switch
- Regions of operation
- Pass gate delay

The CMOS pass gate

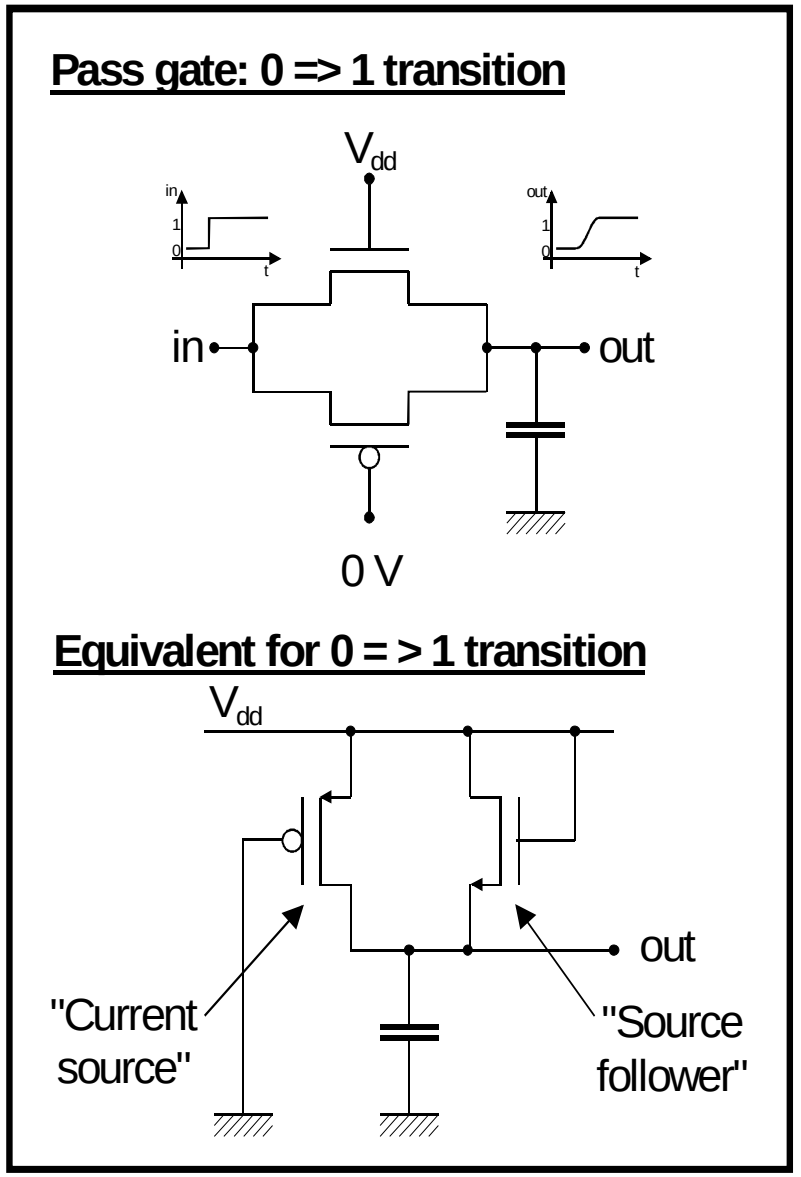


C	A	Y
0	0	?
0	1	?
1	0	good 0
1	1	good 1

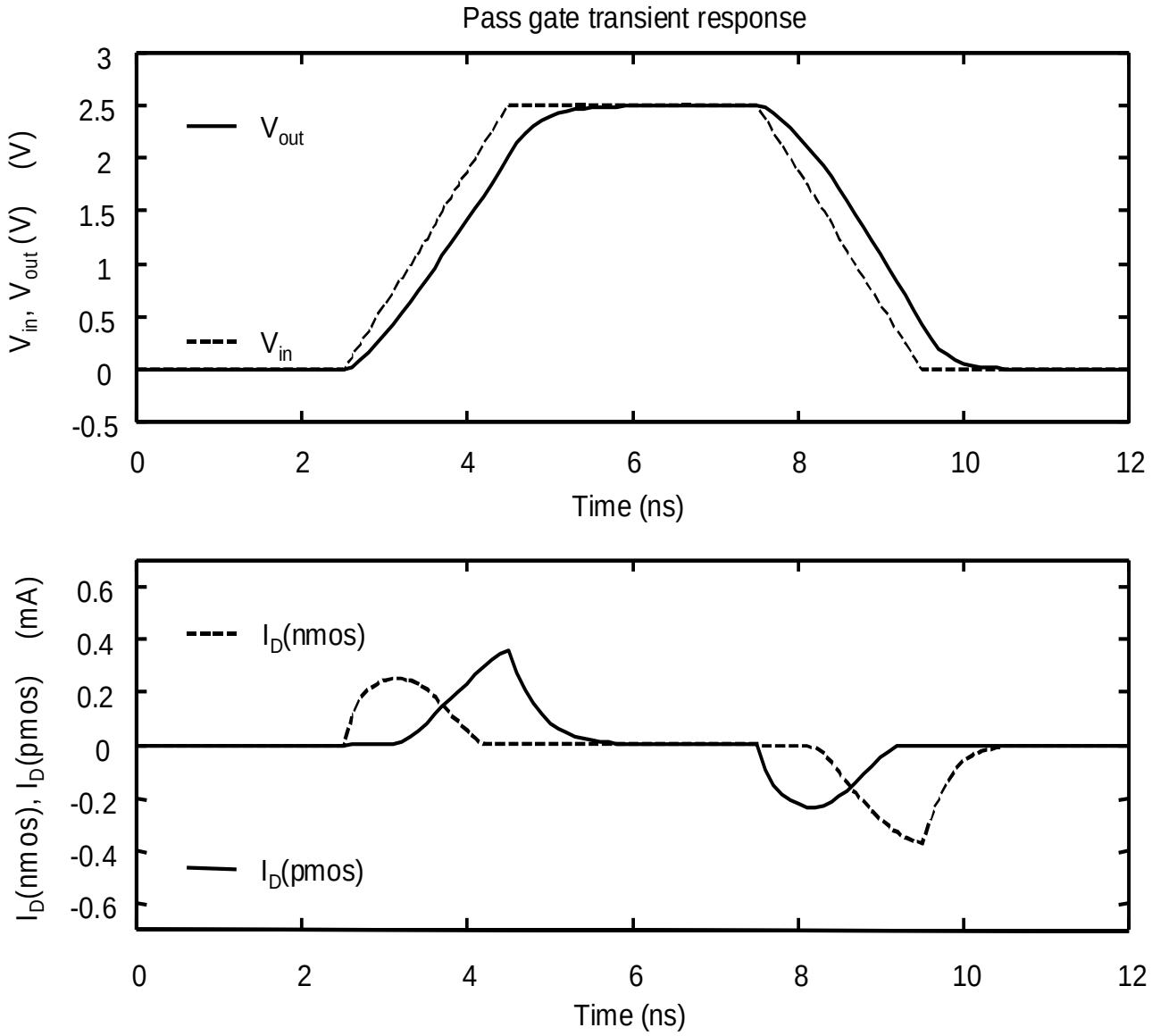
The CMOS pass gate

Regions of operation:
"0" to "1" transition

- NMOS:
  - source follower
  - $V_{gs} = V_{ds}$ always:
    - $V_{out} < V_{dd}-V_{TN} \Rightarrow$ saturation
    - $V_{out} > V_{dd}-V_{TN} \Rightarrow$ cutoff
  - $V_{TN} > V_{TN0}$ (bulk effect)
- PMOS:
  - current source
  - $V_{out} < |V_{TP}| \Rightarrow$ saturation
  - $V_{out} > V_{TP} \Rightarrow$ linear



The CMOS pass gate



The CMOS pass gate

- Regions of operation: “0” to “1” transition

$V_{out} < \lvert V_{TP} \rvert$	NMOS and PMOS saturated
$\lvert V_{TP} \rvert < V_{out} < V_{dd} - V_{TN}$	NMOS saturated, PMOS linear
$V_{out} > V_{dd} - V_{TN}$	NMOS cutoff, PMOS linear

- Regions of operation: “1” to “0” transition

$V_{out} > Vdd - V_{TN}$	NMOS and PMOS saturated
$V_{dd} - V_{TN} > V_{out} > \lvert V_{TP} \rvert$	NMOS linear, PMOS saturated
$V_{TP} > V_{out}$	NMOS linear, PMOS cutoff

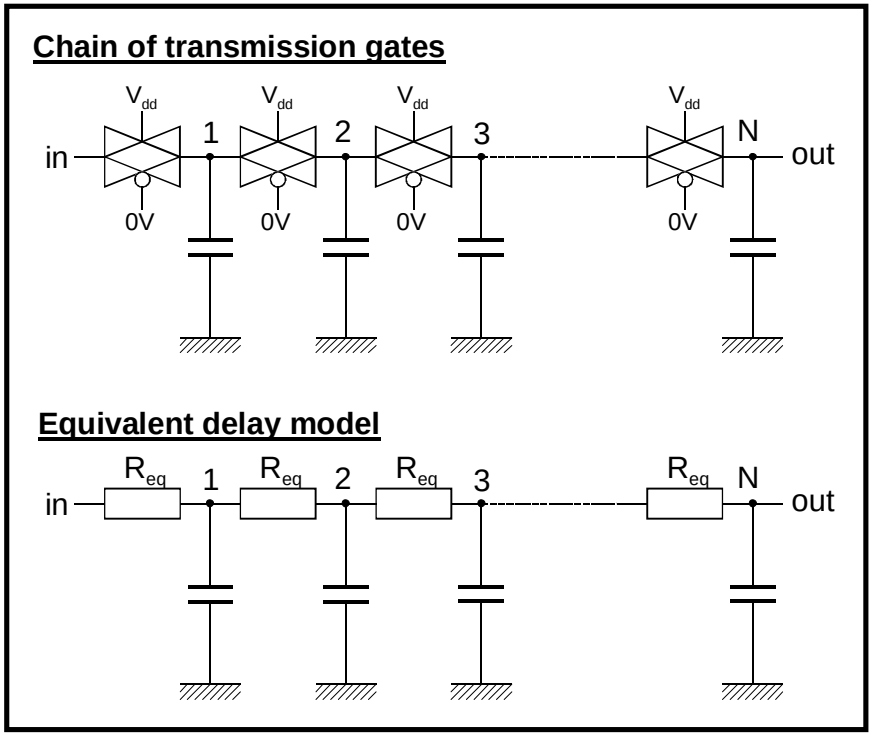
- Both devices combine to form a good switch

The CMOS pass gate

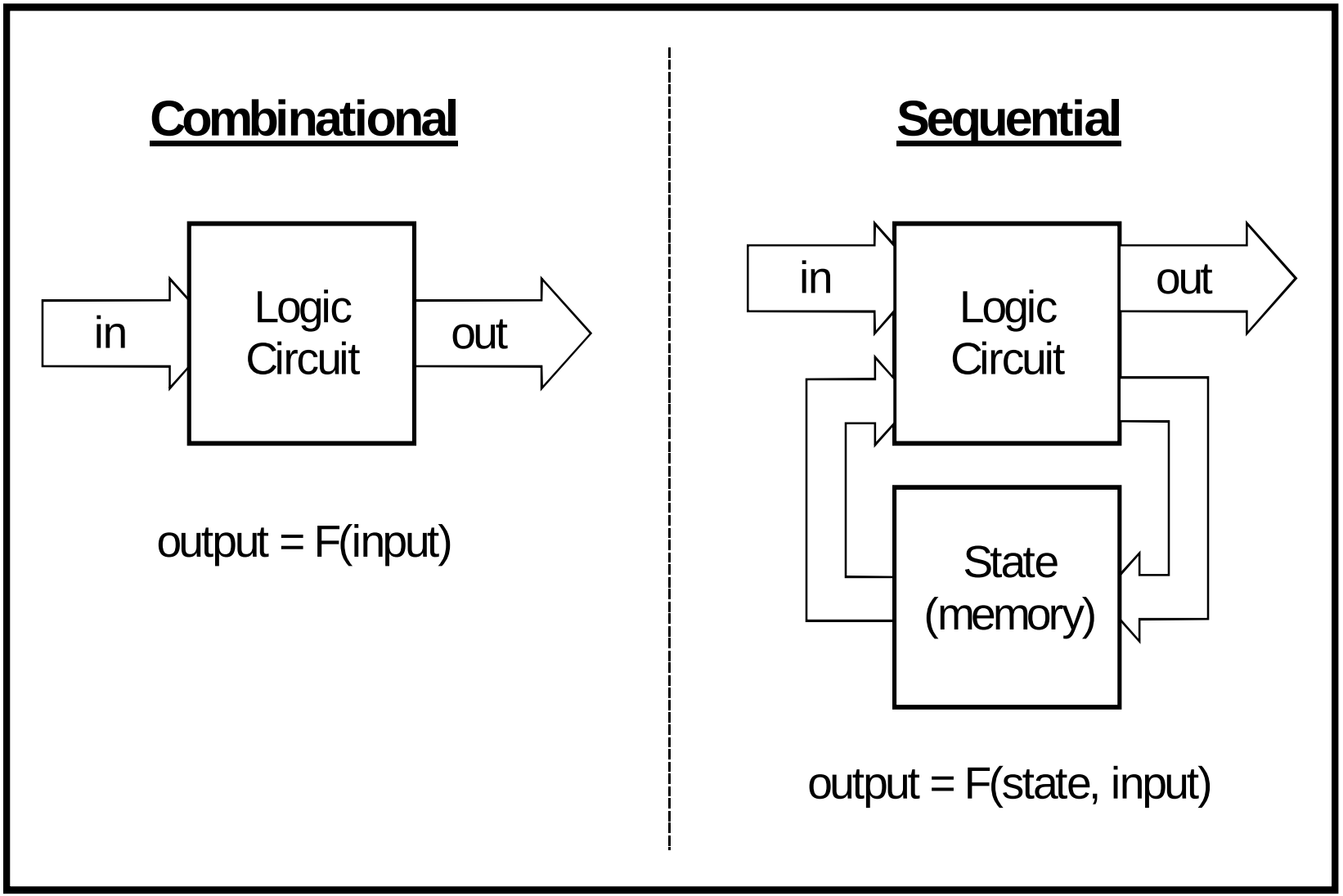
- Delay of a chain of pass gates:

$$t_d \propto C \cdot R_{eq} \cdot \frac{N \cdot (N+1)}{2}$$

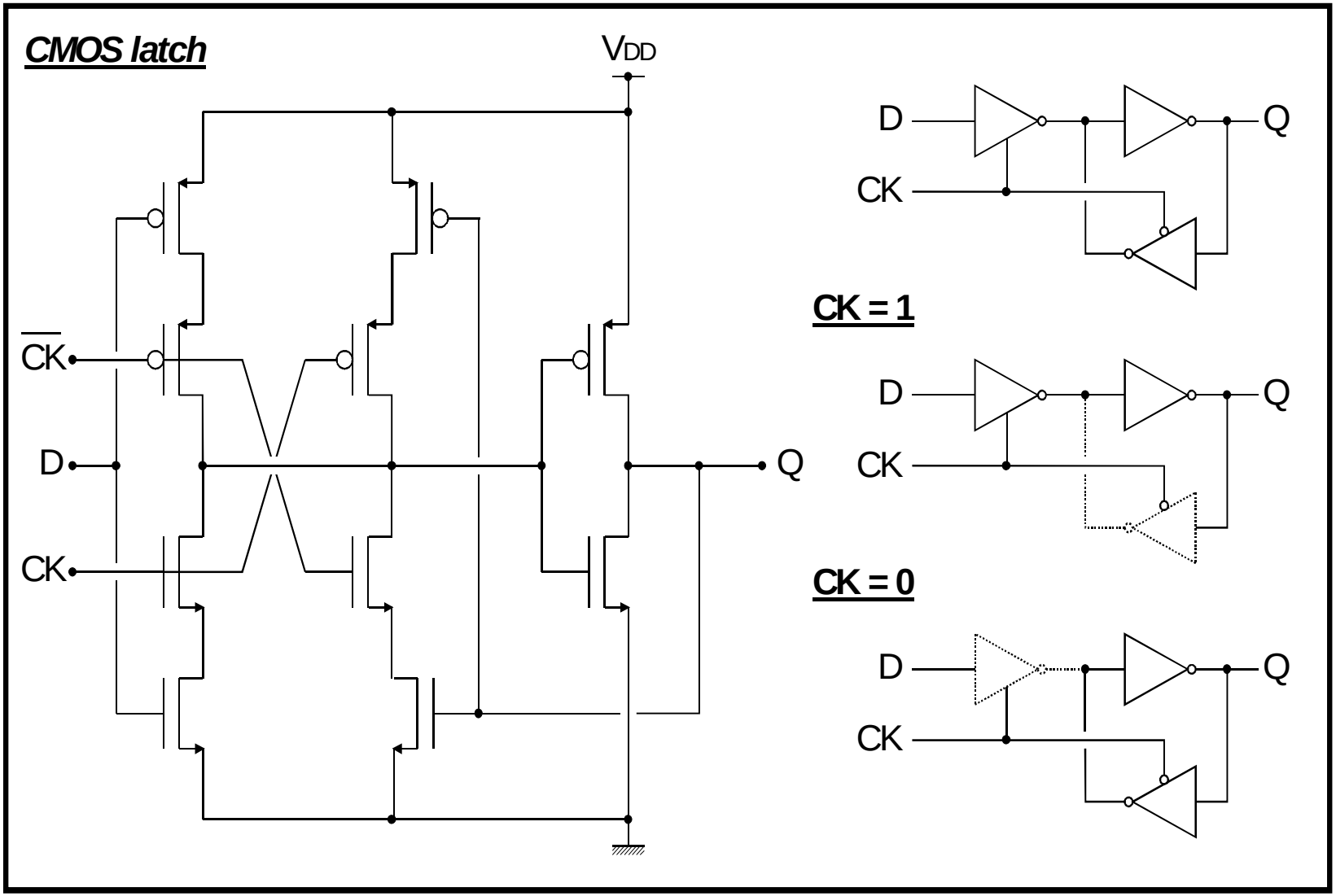
- Delay proportional to N^2
- Avoid N large:
  - Break the chain by inserting buffers
- Warning:
  - A pass gate provides no power gain or buffering
  - All the work is done by the previous gate
  - It really looks like a simple switch
  - Used extensively in some FPGA architectures for programmable wiring



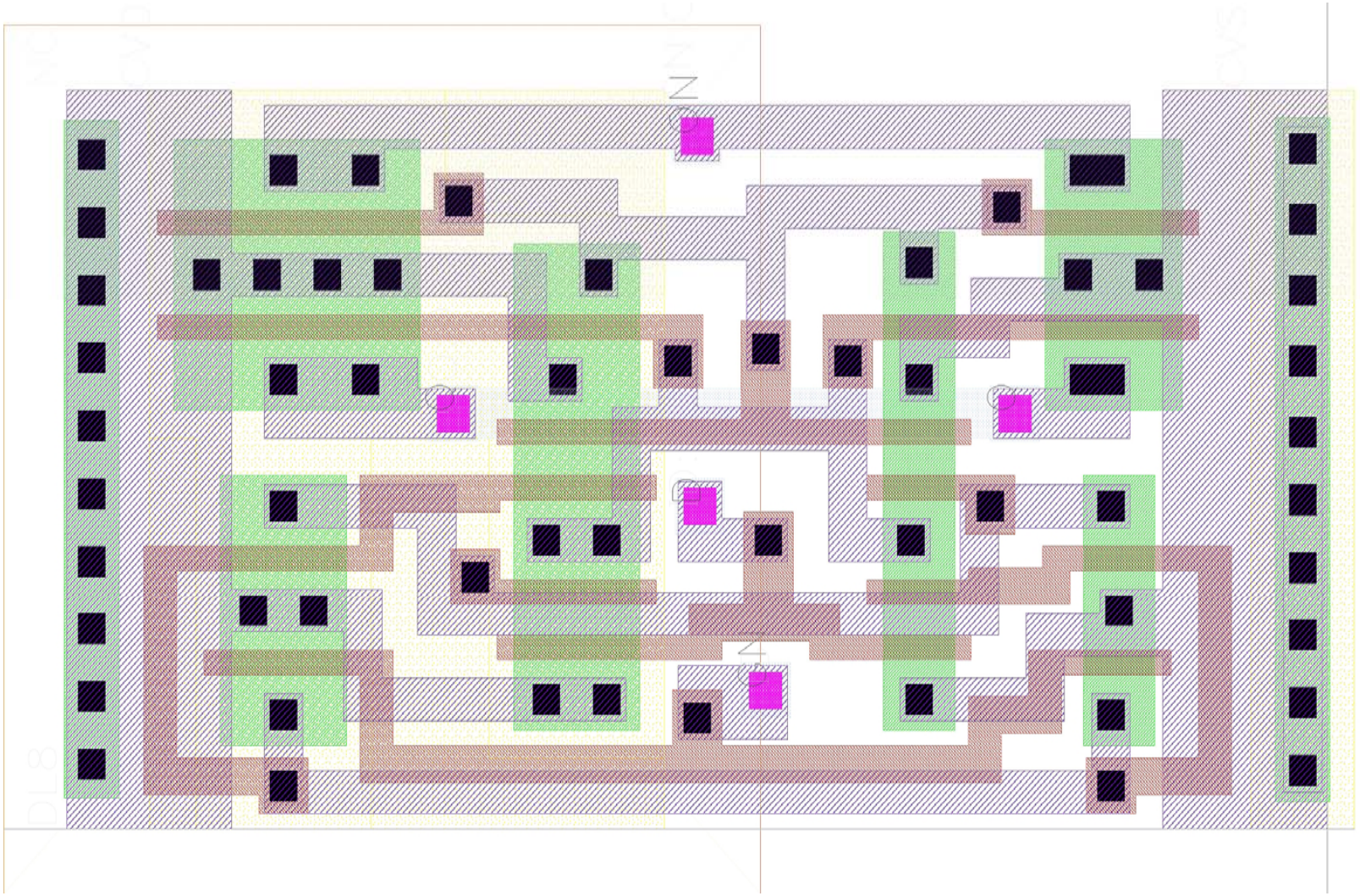
Sequential circuits



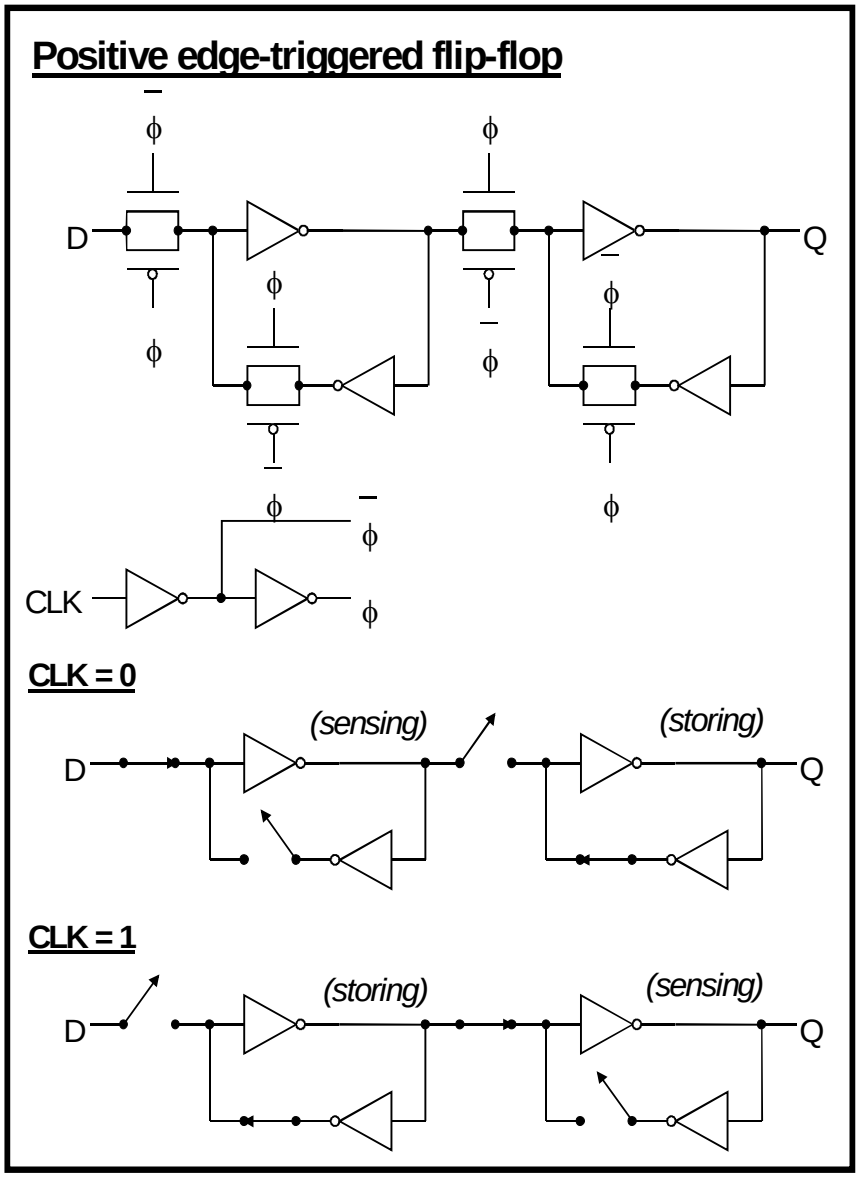
Latch



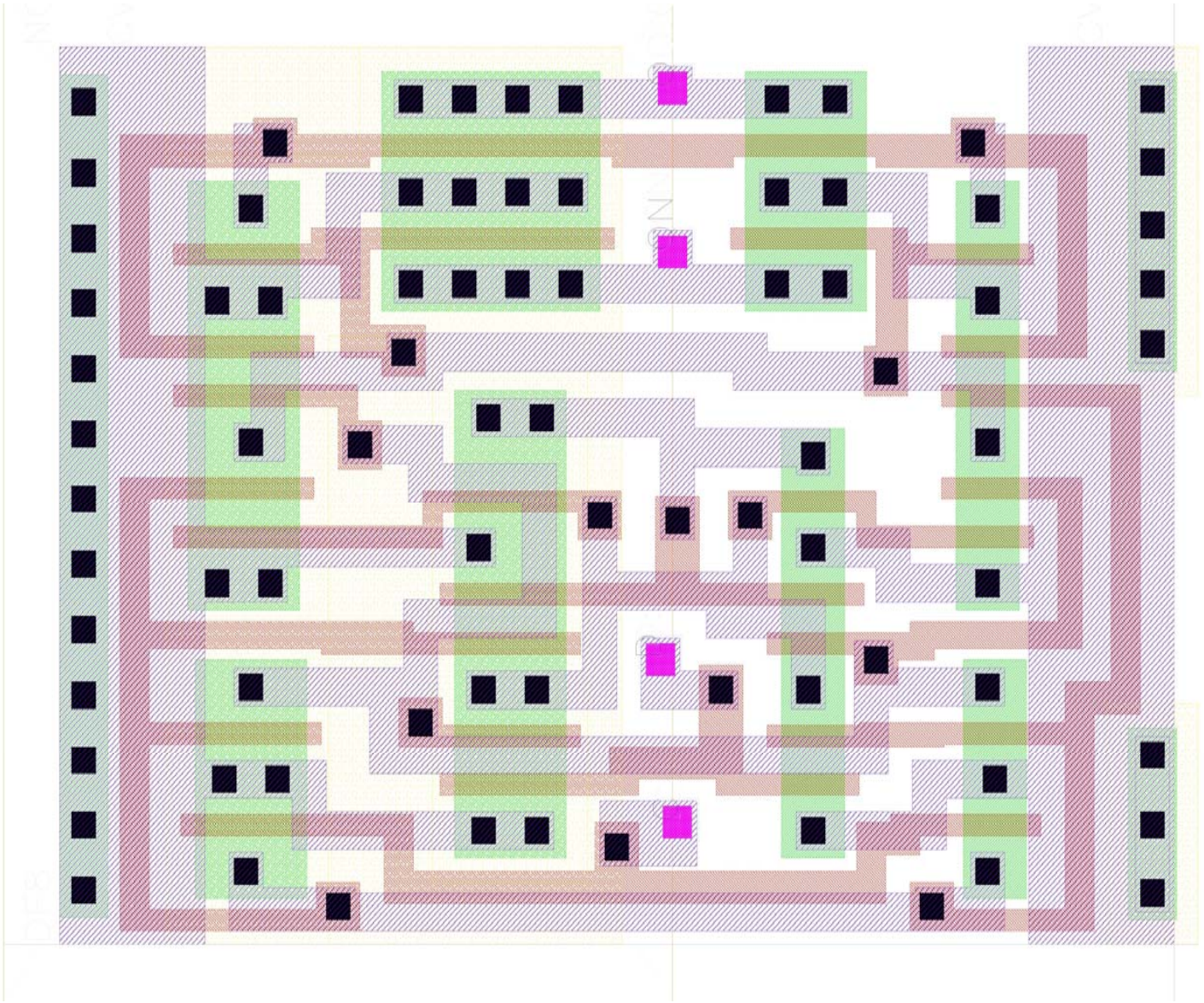
Latch



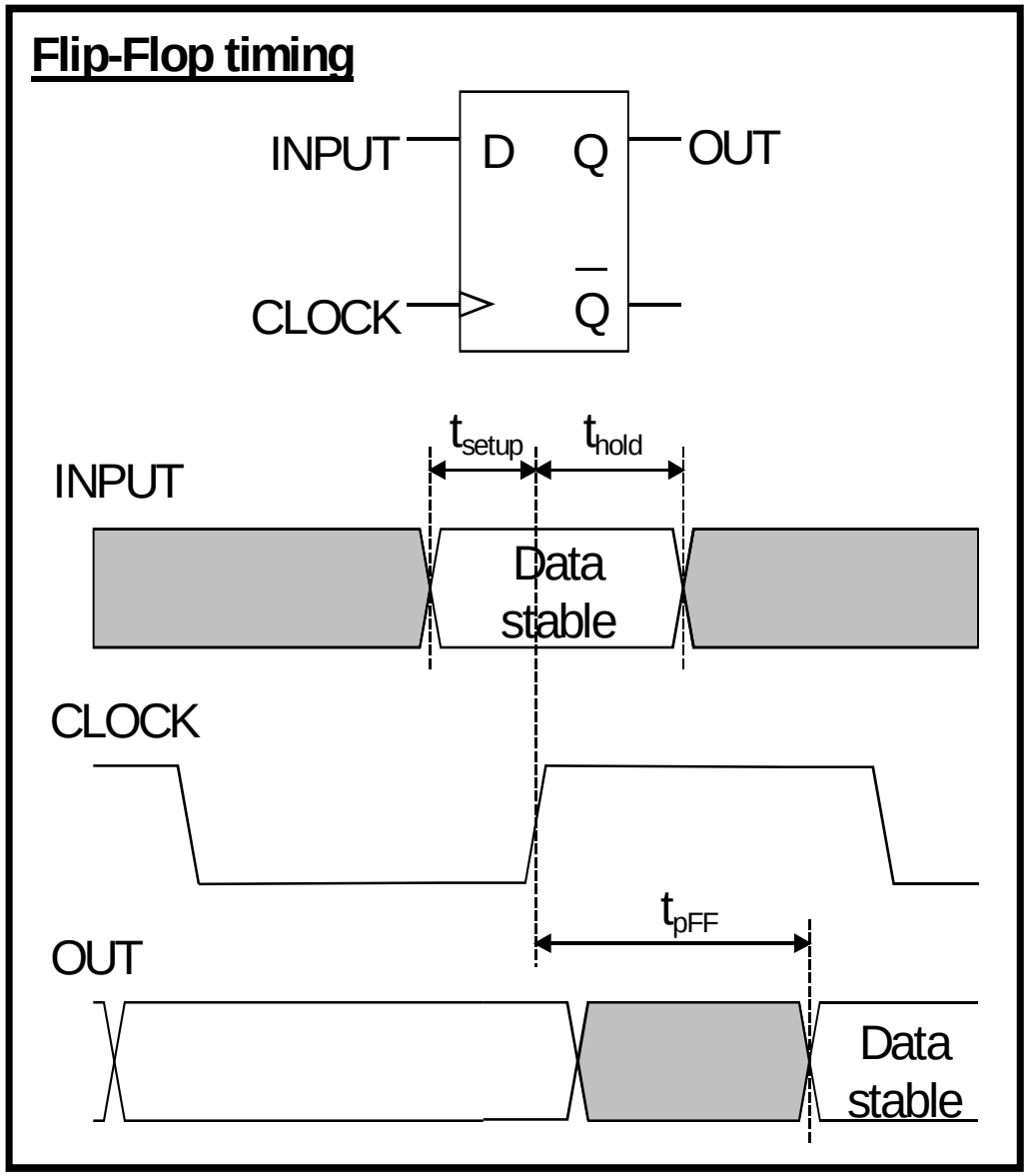
D Flip-Flop


Positive edge-triggered flip-flop
D
Q
CLK
CLK = 0
(sensing)
(storing)
CLK = 1
(storing)
(sensing)

D Flip-Flop

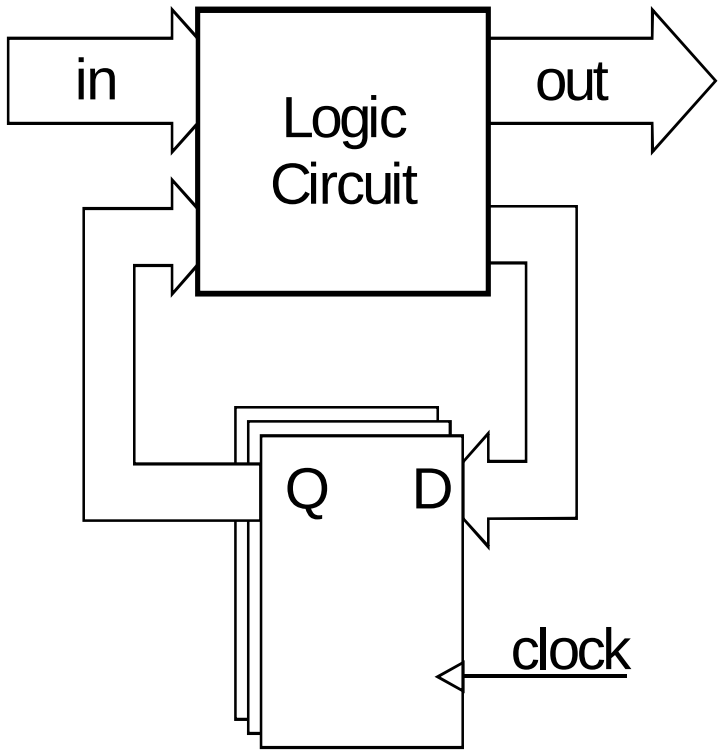


Timing constraints



State machine timing

Maximum clock frequency



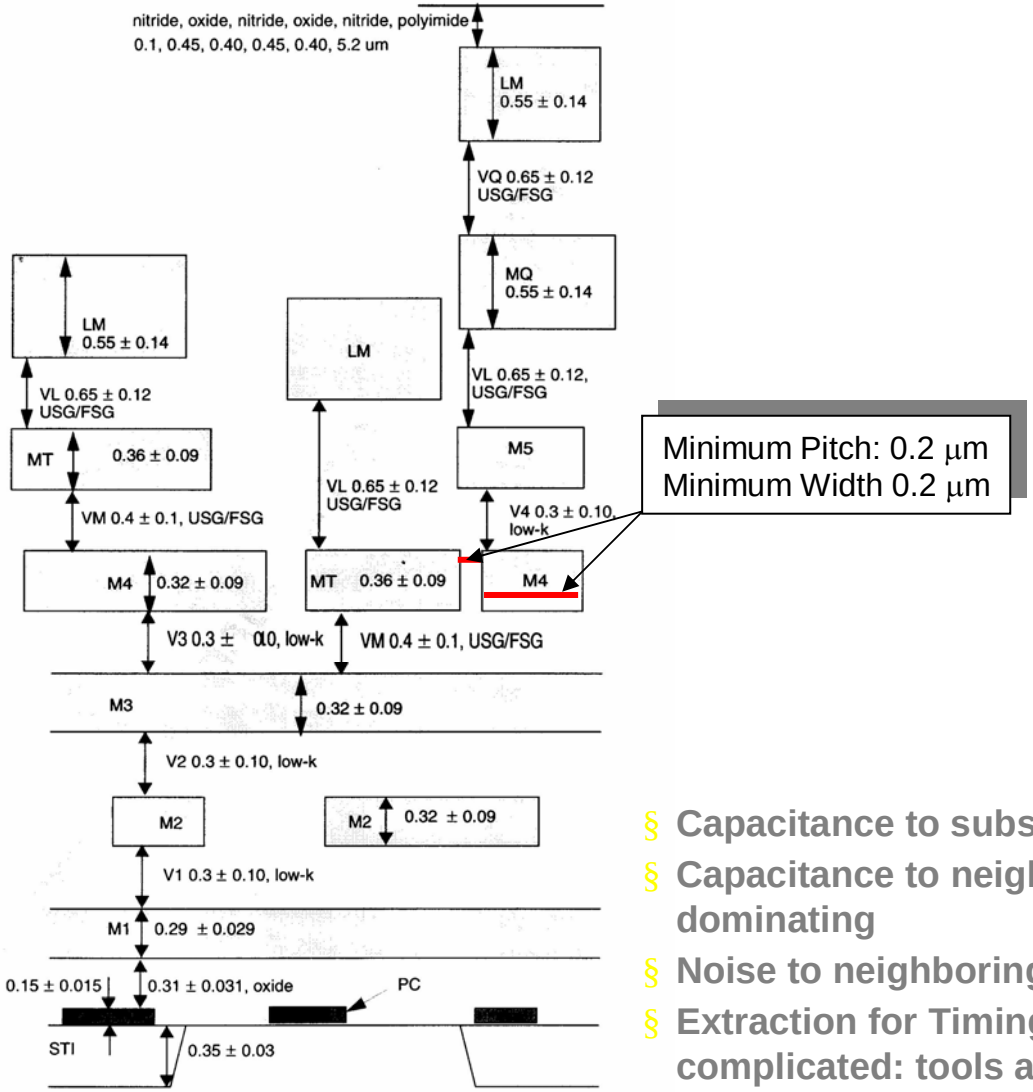
$$T_{min} > t_{pFF} + t_{p,comb} + t_{setup}$$

$$f_{max} = 1/T_{min}$$

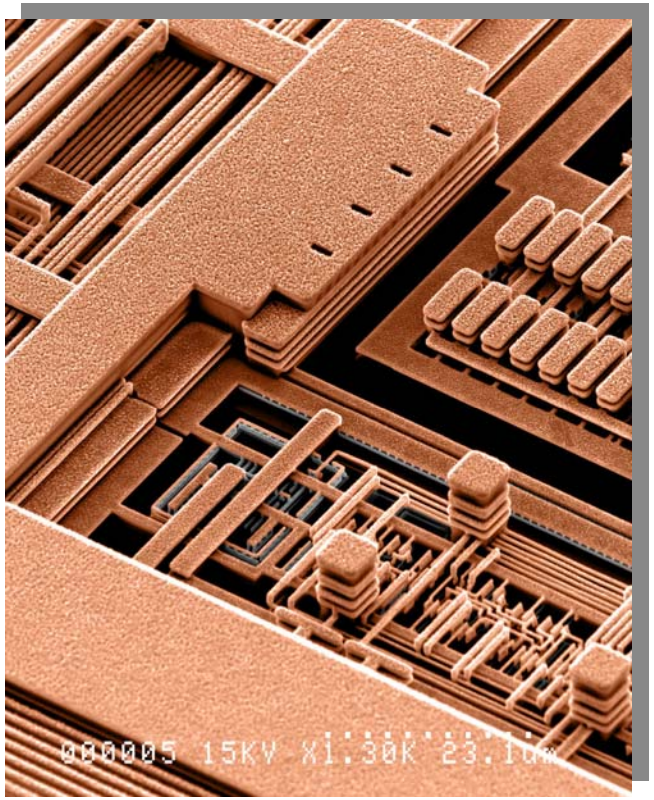
Interconnects

- The previous result assumes that signals can propagate instantaneously across interconnects
- In reality interconnects are metal or polysilicon structures with associated resistance and capacitance.
- That, introduces signal propagation delay that has to be taken into account for reliable operation of the circuit

Interconnects



Minimum Pitch: 0.2 μm
Minimum Width 0.2 μm



- Capacitance to substrate becomes irrelevant
- Capacitance to neighboring signal becomes dominating
- Noise to neighboring signal also not negligible
- Extraction for Timing simulation horribly complicated: tools absolutely mandatory

Interconnects

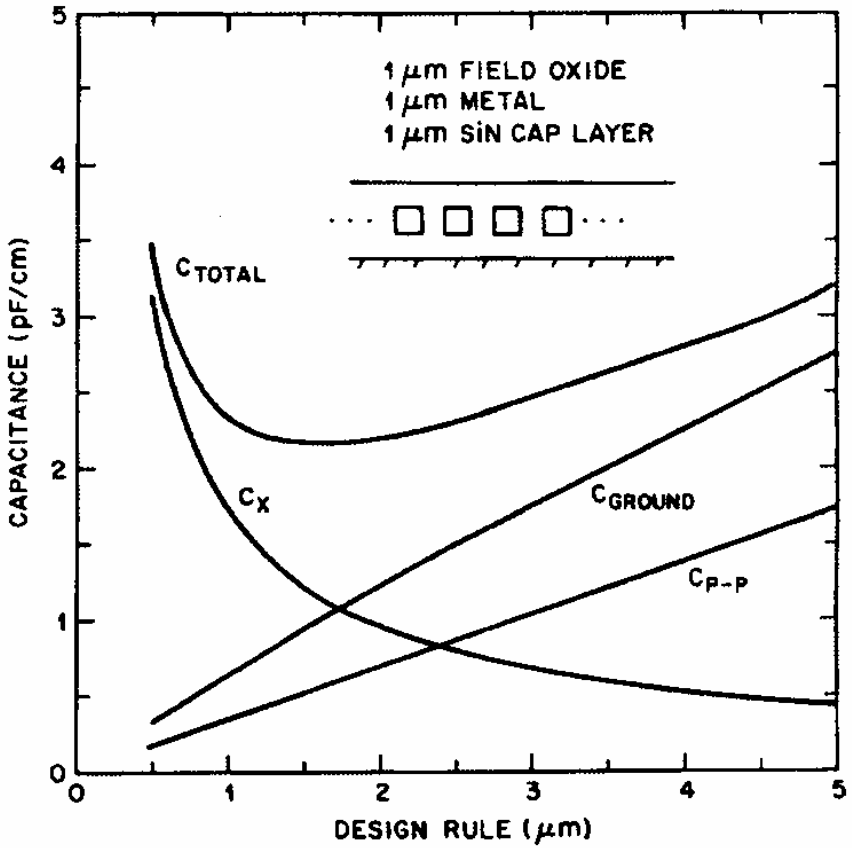
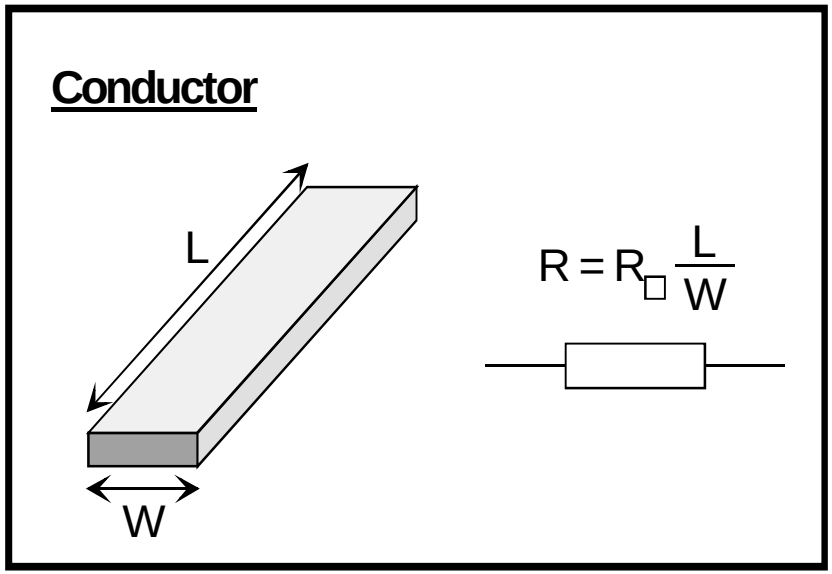


Figure 3.10: Interconnect capacitance including wire-to-wire capacitance [Schaper83]. (© 1983 IEEE)

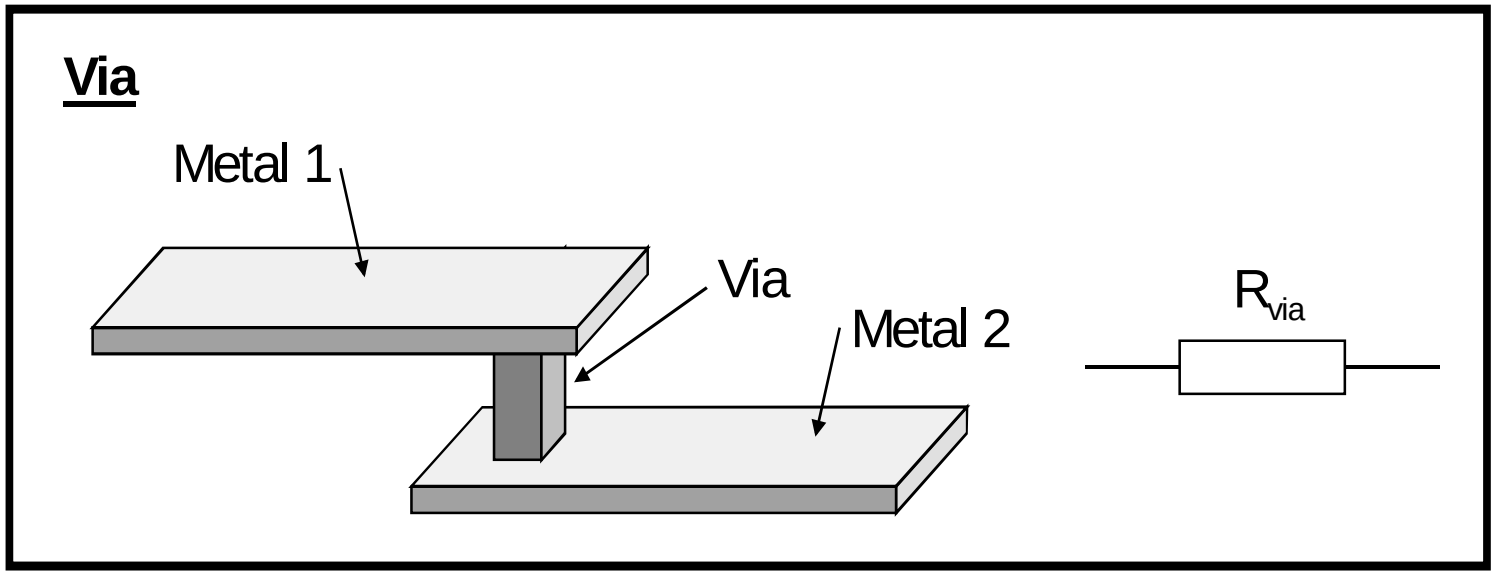
Interconnects



Film	Sheet resistance (Ω/square)
n-well	310
p+, n+ diffusion (salicided)	4
polysilicon (salicided)	4
Metal 1	0.12
Metal 2, 3 and 4	0.09
Metal 5	0.05

(Typical values for an advanced process)

Interconnects

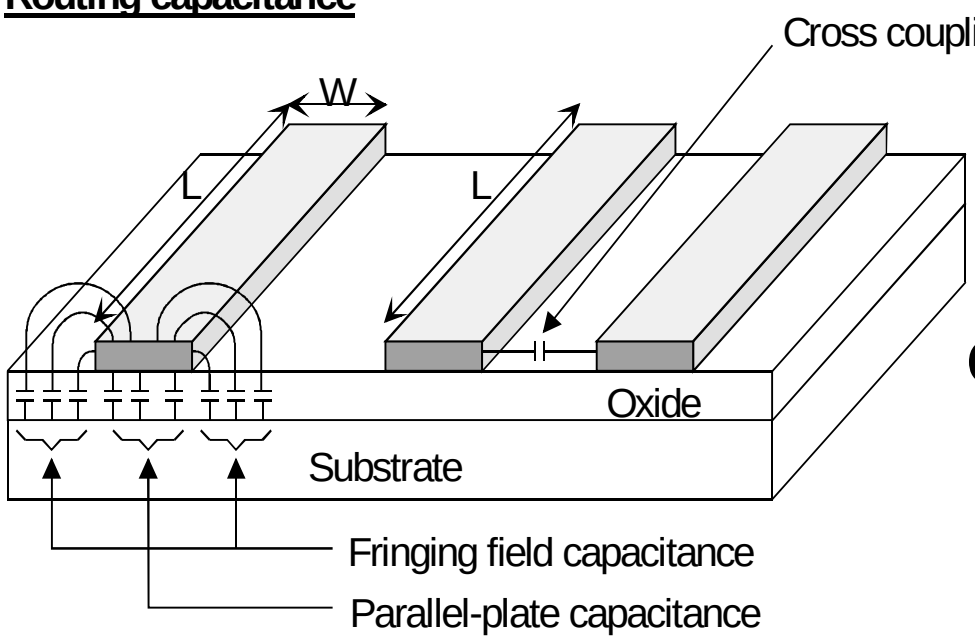


- Via or contact resistance depends on:
  - The contacted materials
  - The contact area

Via/contact	Resistance (Ω)
M1 to n+ or p+	10
M1 to Polysilicon	10
V1, 2, 3 and 4	7

Interconnects

Routing capacitance

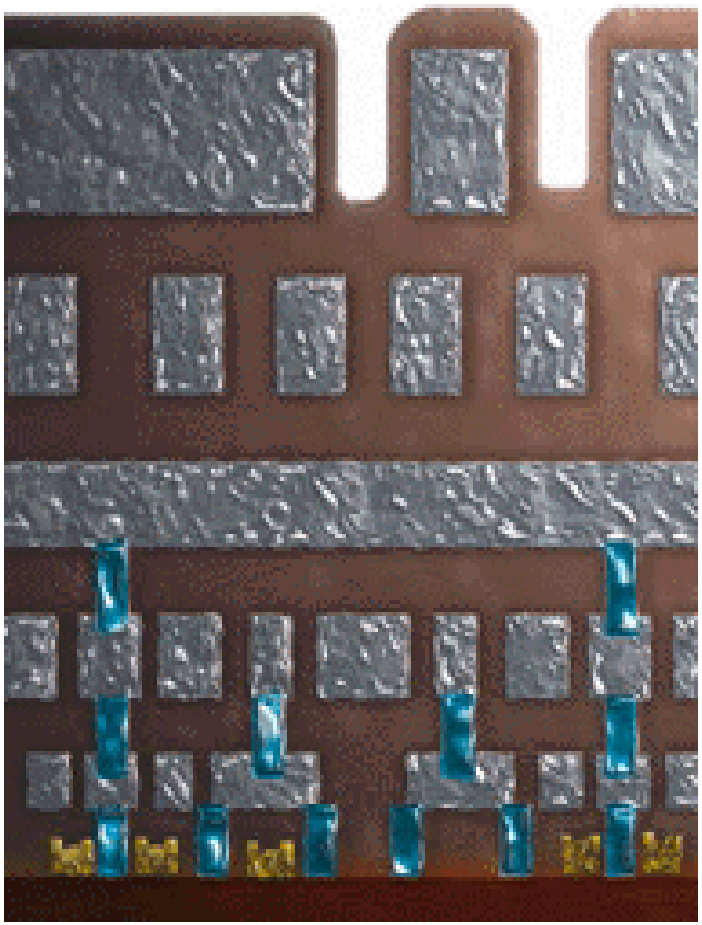


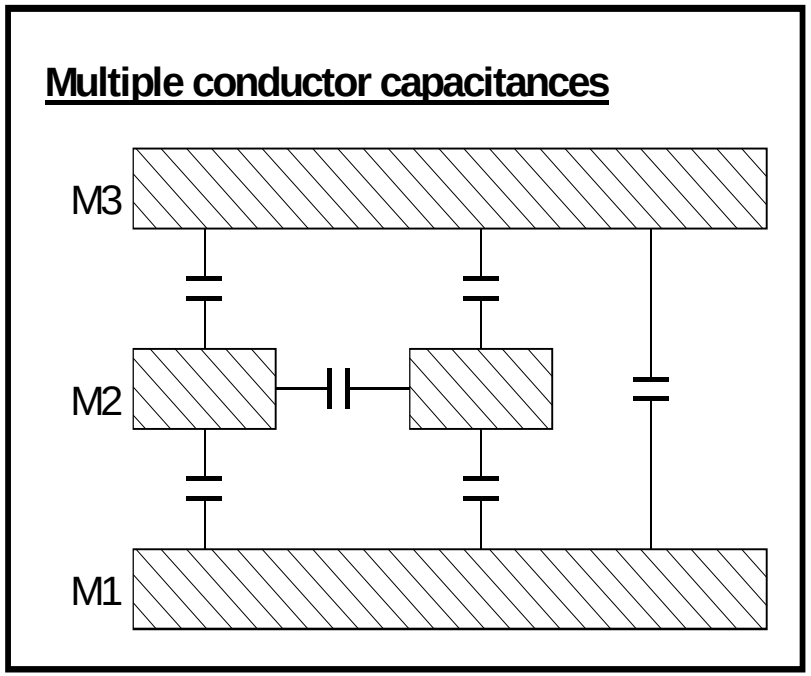
$$C = C_{f0}\, 2\, L + C_{p0}\, W\, L$$

$$C_x = C_{x0}\, L$$

Interconnect layer	Parallel-plate (fF/μm^2)	Fringing (fF/μm)
Polysilicon to sub.	0.058	0.043
Metal 1 to sub.	0.031	0.044
Metal 2 to sub.	0.015	0.035
Metal 3 to sub.	0.010	0.033

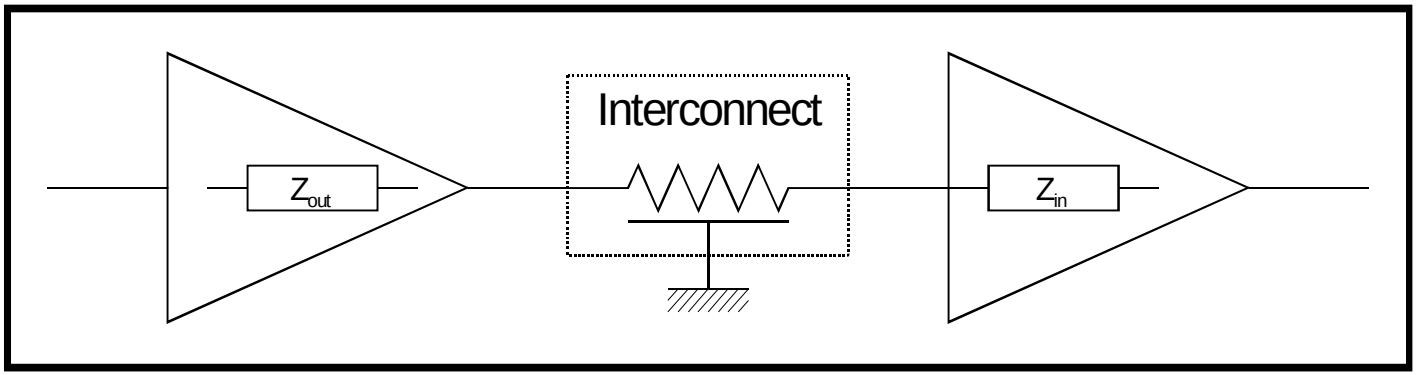
Interconnects





- Three dimensional field simulators are required to accurately compute the capacitance of a multi-wire structure

Interconnects

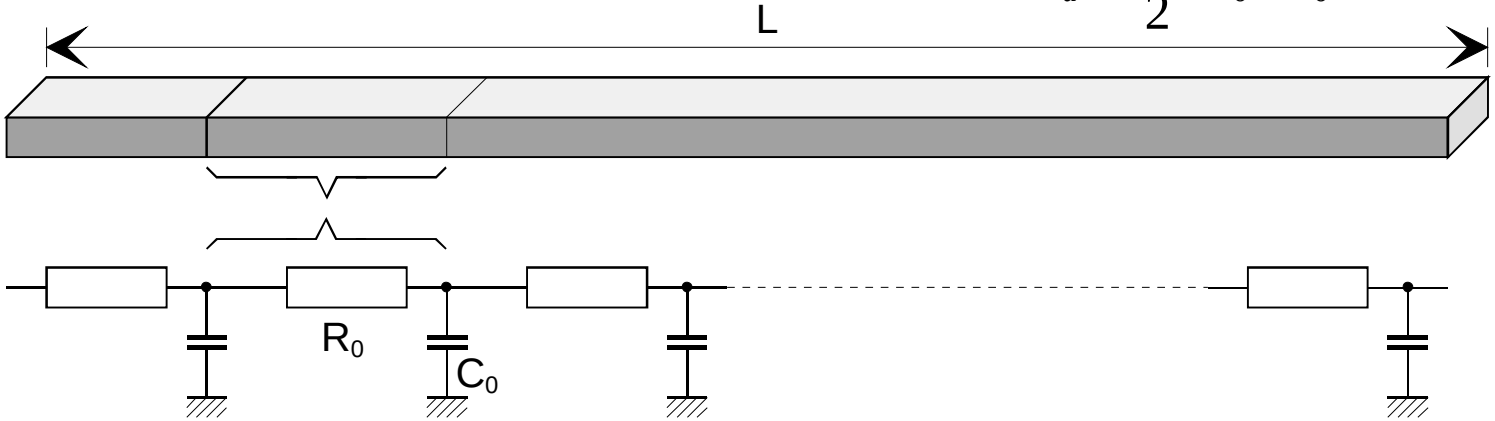


- Delay depends on:
  - Impedance of the driving source
  - Distributed resistance/capacitance of the wire
  - Load impedance
- Distributed RC delay:
  - Can be dominant in long wires
  - Important in polysilicon wires (relatively high resistance)
  - Important in salicided wires
  - Important in heavily loaded wires

Interconnects

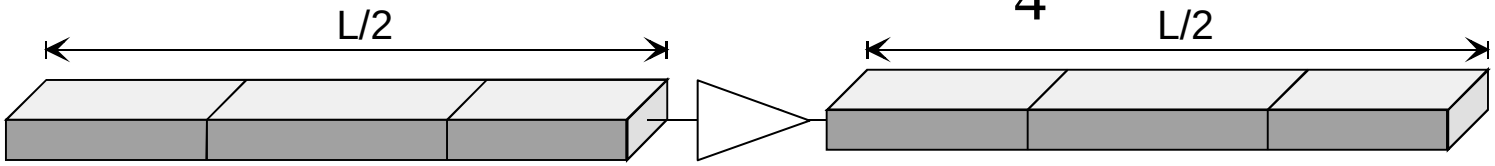
Long line

$$t_d = \frac{1}{2} \cdot R_0 \cdot C_0 \cdot L^2$$



Delay optimization

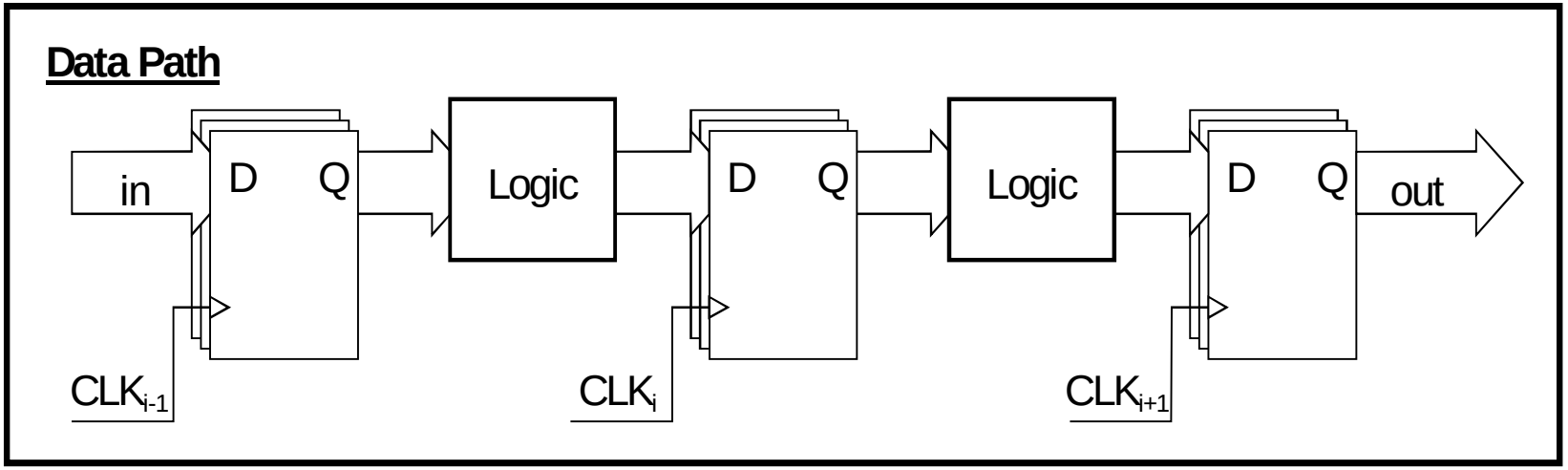
$$t_d = \frac{1}{4} R_0 \cdot C_0 \cdot L^2 + t_{buff}$$



Clock distribution

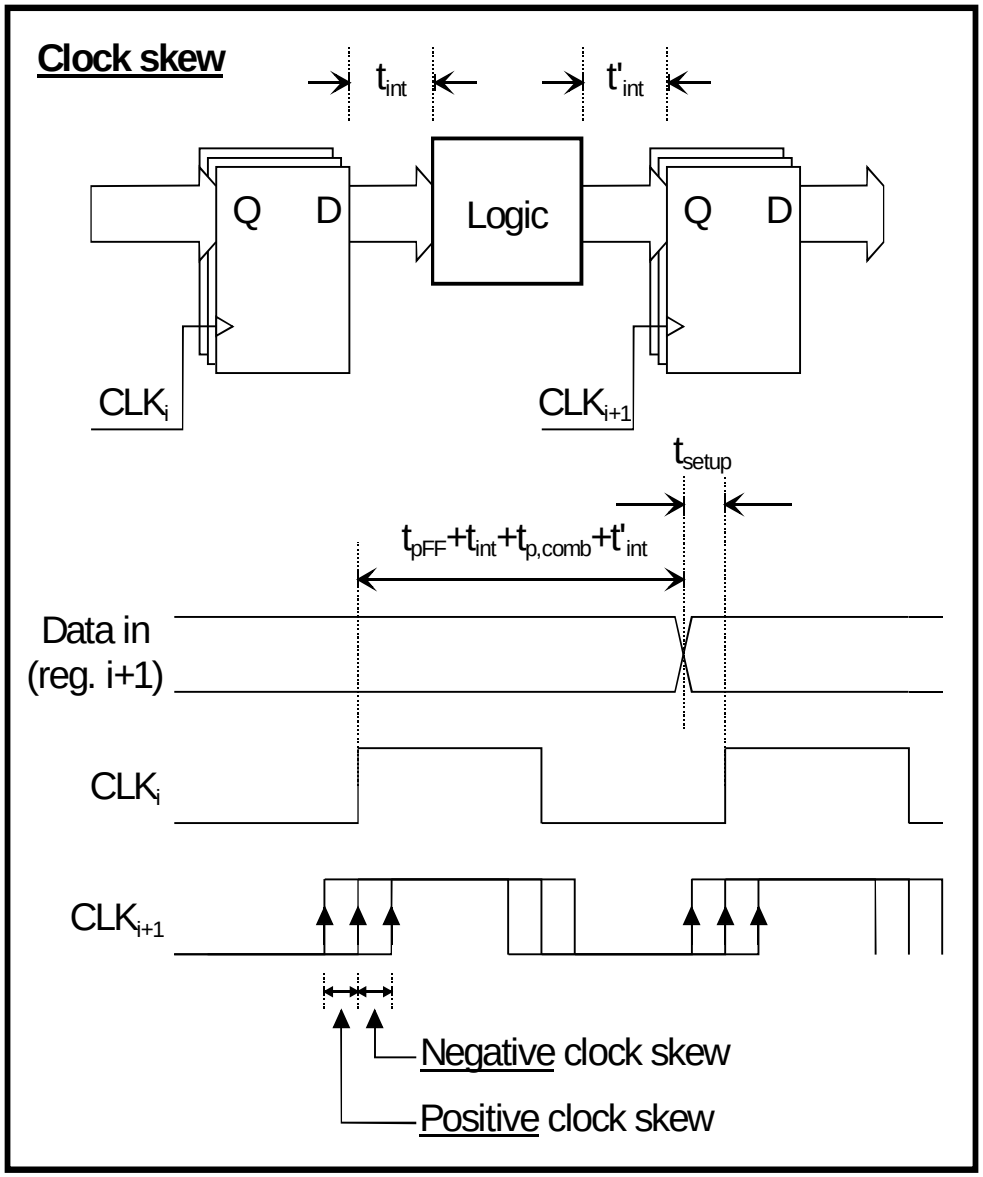
- Clock signals are “special signals”
- Every data movement in a synchronous system is referenced to the clock signal
- Clock signals:
  - Are typically loaded with high fanout
  - Travel over the longest distances in the IC
  - Operate at the highest frequencies

Clock distribution



- "Equipotential" clocking:
  - In a synchronous system all clock signals are derived from a single clock source ("clock reference")
  - Ideally: clocking events should occur at all registers simultaneously … = t(clk$_{i-1}$) = t(clk$_i$) = t(clk$_{i+1}$) = …
  - In practice: clocking events will occur at slightly different instants among the different registers in the data path

Clock distribution



Clock distribution

- Skew: difference between the clocking instants of two “sequential” registers:

  Skew = $t(CLK_i) - t(CLK_{i+1})$

- Maximum operation frequency:

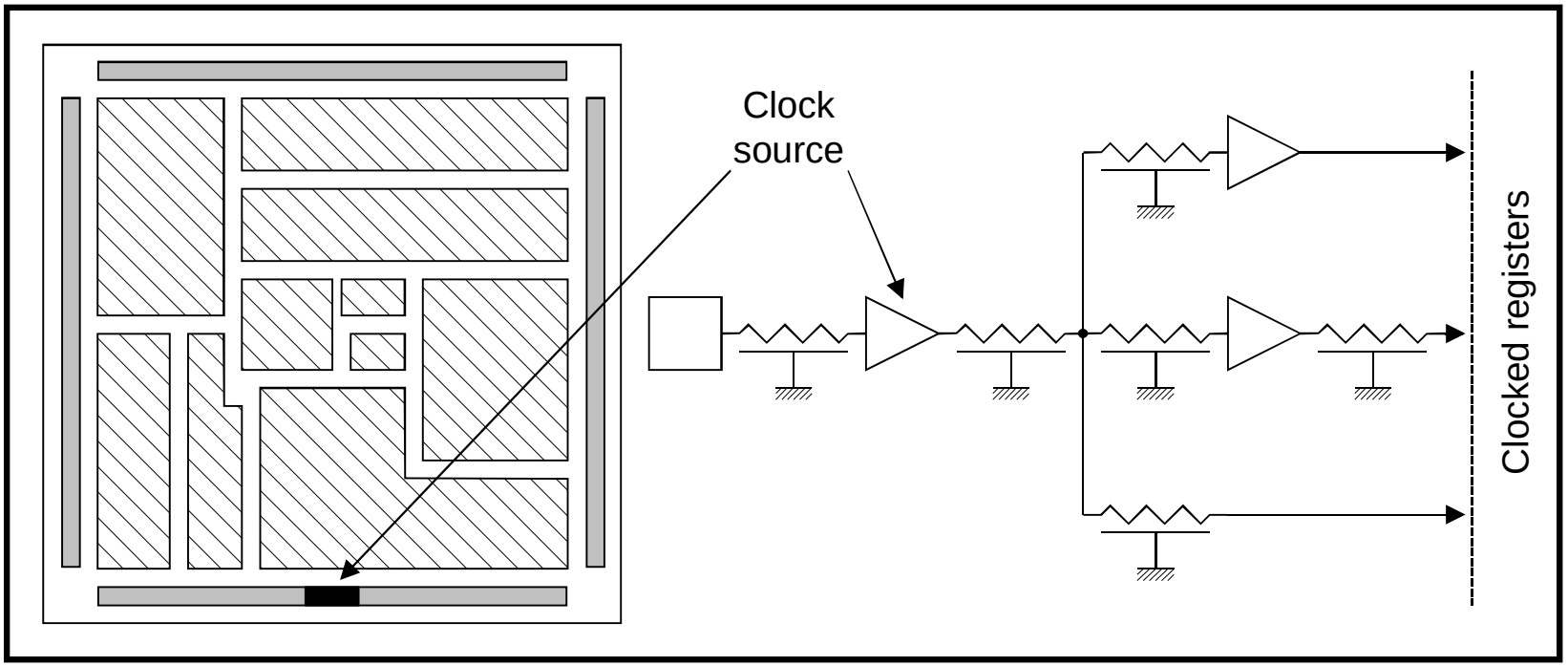
$$T_{\min} = \frac{1}{f_{\max}} = t_{dFF} + t_{\text{int}} + t_{p,comb} + t'_{\text{int}} + t_{setup} + t_{skew}$$

- Skew > 0, decreases the operation frequency
- Skew < 0, can be used to compensate a critical data path BUT this results in more positive skew for the next data path!

Clock distribution

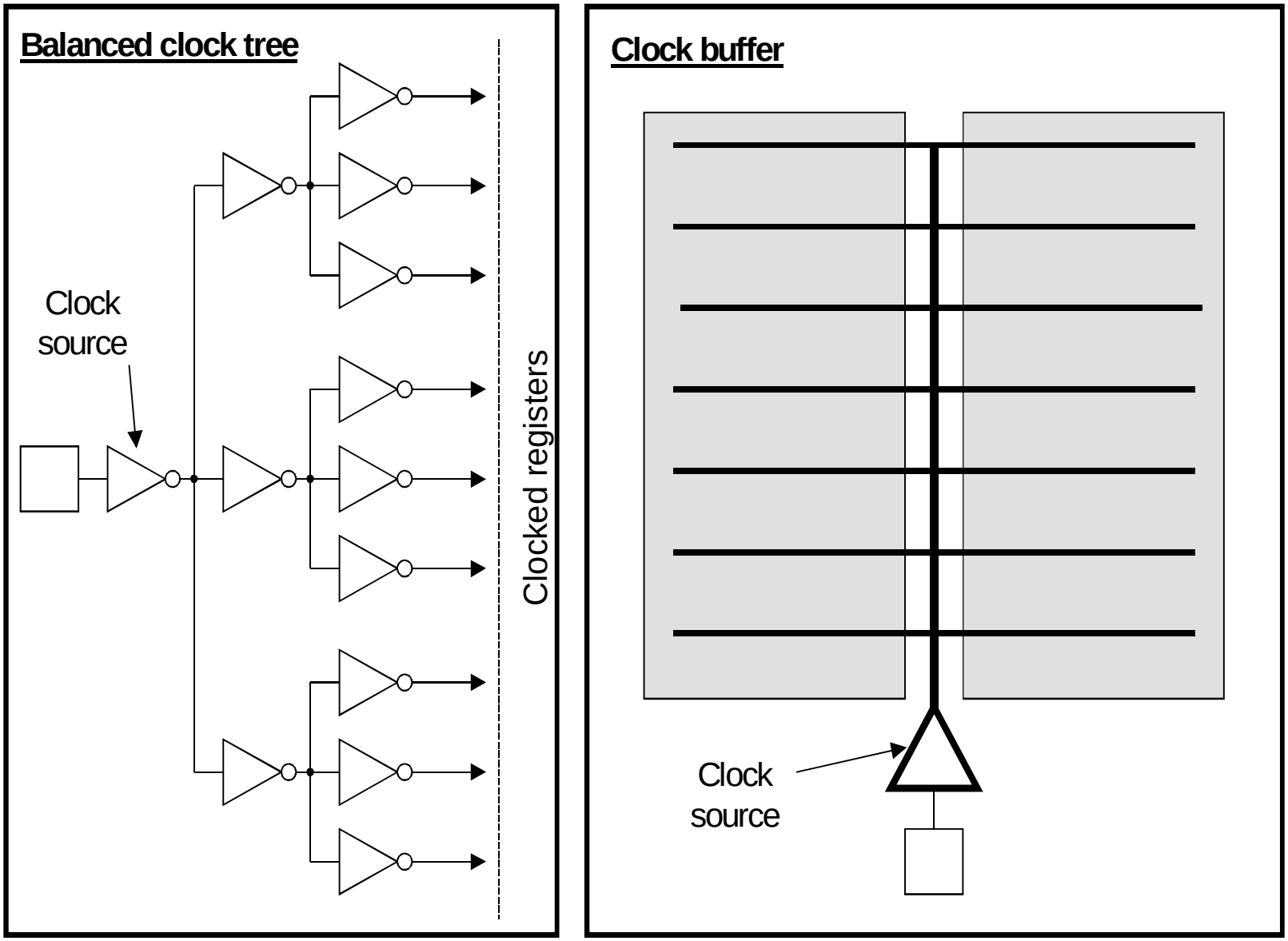
- Different clock paths can have different delays due to:
  - Differences in line lengths from clock source to the clocked registers
  - Differences in passive interconnect parameters:
    - line resistance/capacitance, line with, …
  - Differences in delays in the active buffers:
    - Different driving strength
    - Different loading
    - Differences in active device parameters:
      - threshold voltages, channel mobility;
- In a well designed and balanced clock distribution network, the distributed clock buffers are the principal source of clock skew

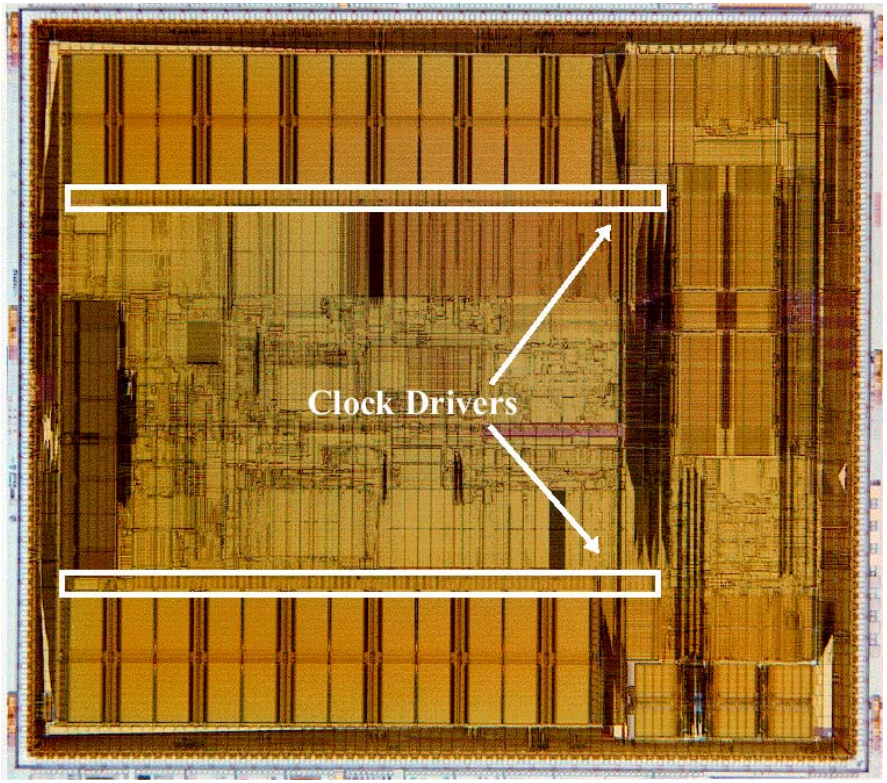
Clock distribution

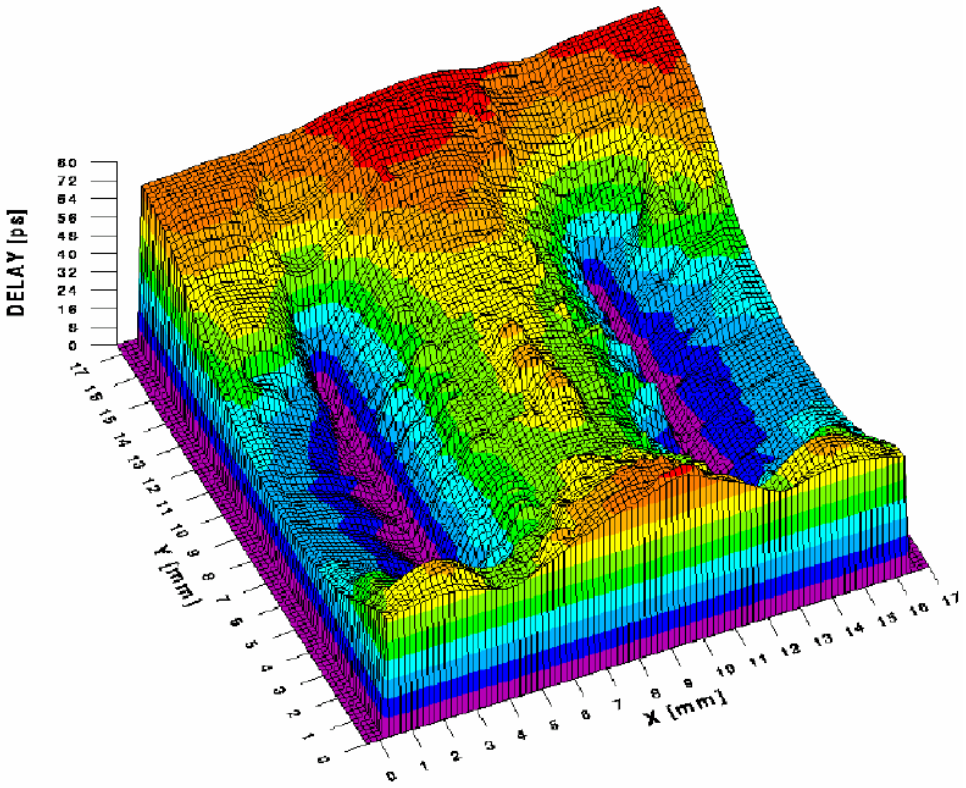


- Clock buffers:
  - Amplify the clock signal degraded by the interconnect impedance
  - Isolate the local clock lines from upstream load impedances
  - Loading should be equalized

Clock distribution




Clock Drivers


DELAY [ps]
60
72
84
96
48
40
32
24
16
8
0
Y [mm]
X [mm]

Memory

- Memory classification
- Write/read cycle
- Memory architecture
- Read-only memories
- Nonvolatile read-write memories
- Read-write memories
- Sense amplifiers

- Note: Majority of transistors in chips today are used for memory (Even in a microprocessor)

Memory classification

- Memory: logic element where data can be stored to be retrieved at a later time
- Read-Only Memory (ROM)
  - The information is encoded in the circuit topology
  - The data cannot be modified: it can only be read
  - ROM's are not volatile. That is, removing the power source does not erase the information contents of the memory.

Memory classification

- Read Write Memories (RWM)
  - RWM's allow both reading and writing operations
  - RWM can be of two general types:
    - Static: the data is stored in flip-flops
    - Dynamic: the data is stored as charge in a capacitor
  - Both types of memories are volatile, that is, data is lost once the power is turned off
  - Dynamic memories require periodic "refresh" of its contents in order to compensate for the charge loss caused by leakage currents in the memory element

Memory classification

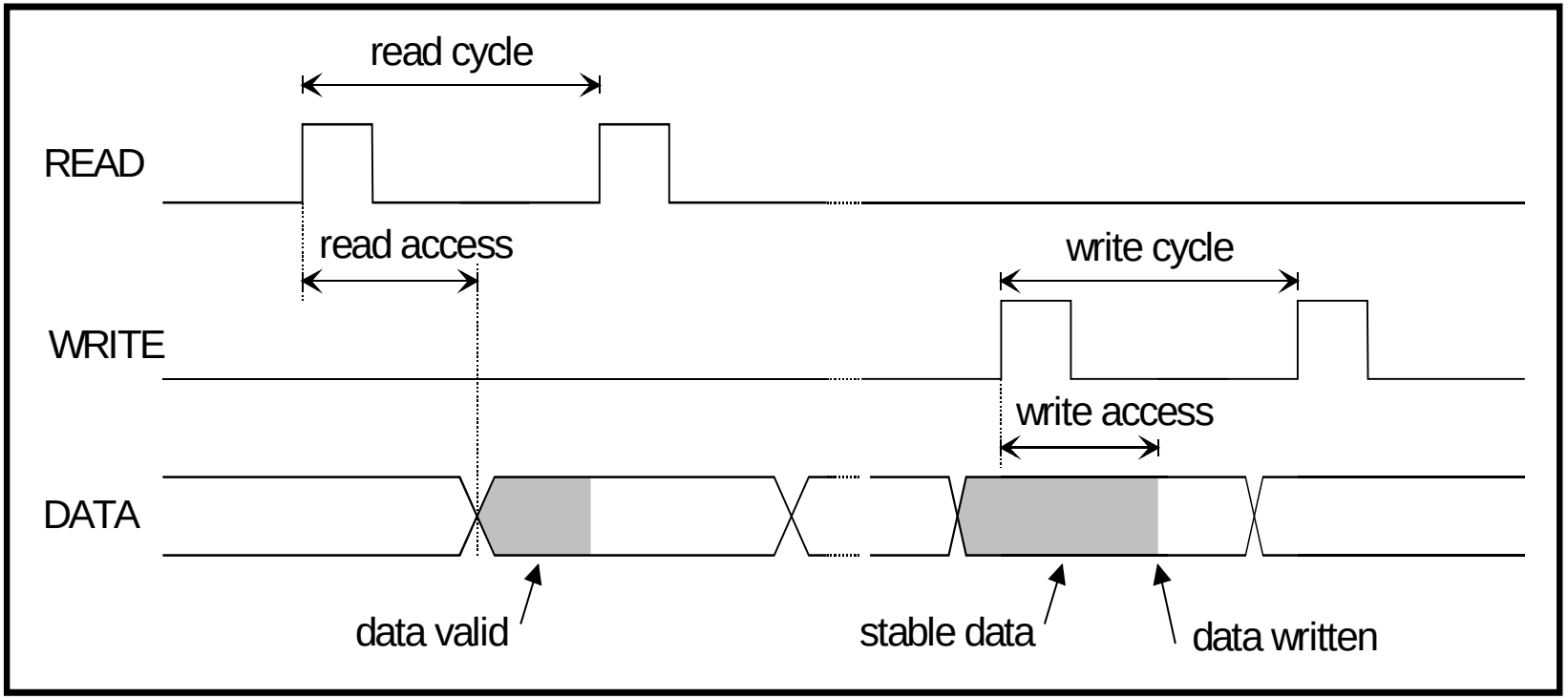
- Nonvolatile Read-Write Memories (NVRWM)
  - These are non volatile memories that allow write operations
  - However:
    - The write operation takes substantially more time than the read operation
    - For some types of NVRWM's, the write operation requires special lab equipment
  - Examples of such memories are:
    - EPROM (Erasable Programmable Read-Only memory)
    - E^2PROM (Electrically Erasable Programmable Read-Only Memory)

Memory classification

- Memories can also be classified according to the way they allow access to the stored data:
  - Random Access: memory locations can be read or written in a random order
  - First-In First-Out (FIFO): The first word to be written is the first word to be read
  - Last-In First-Out (LIFO): The last word to be written is the first word to be read (stack)
  - Shift Register: information is streamed in and out. It can work either as a FIFO or as a LIFO

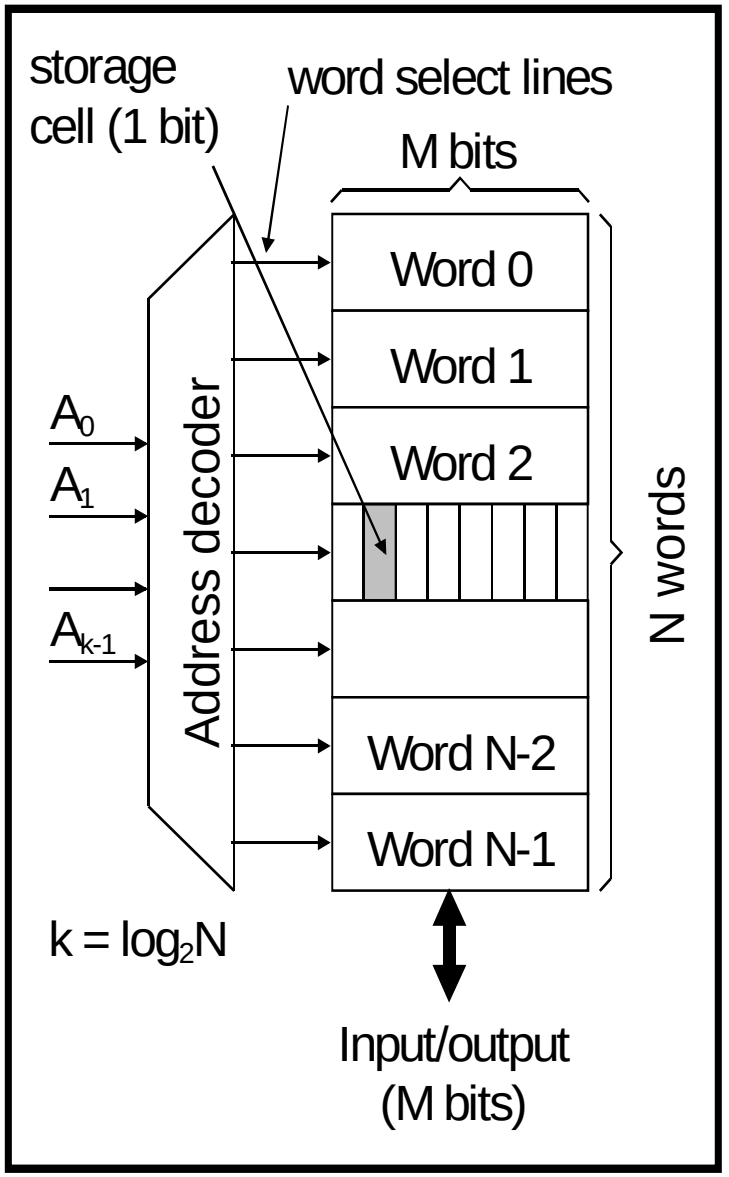
Write/read cycle

- Read-access time: delay between read request and data valid
- Write-access time: delay between write request and the actual writing
- Read or write cycle time: minimum time required between successive read or write operations



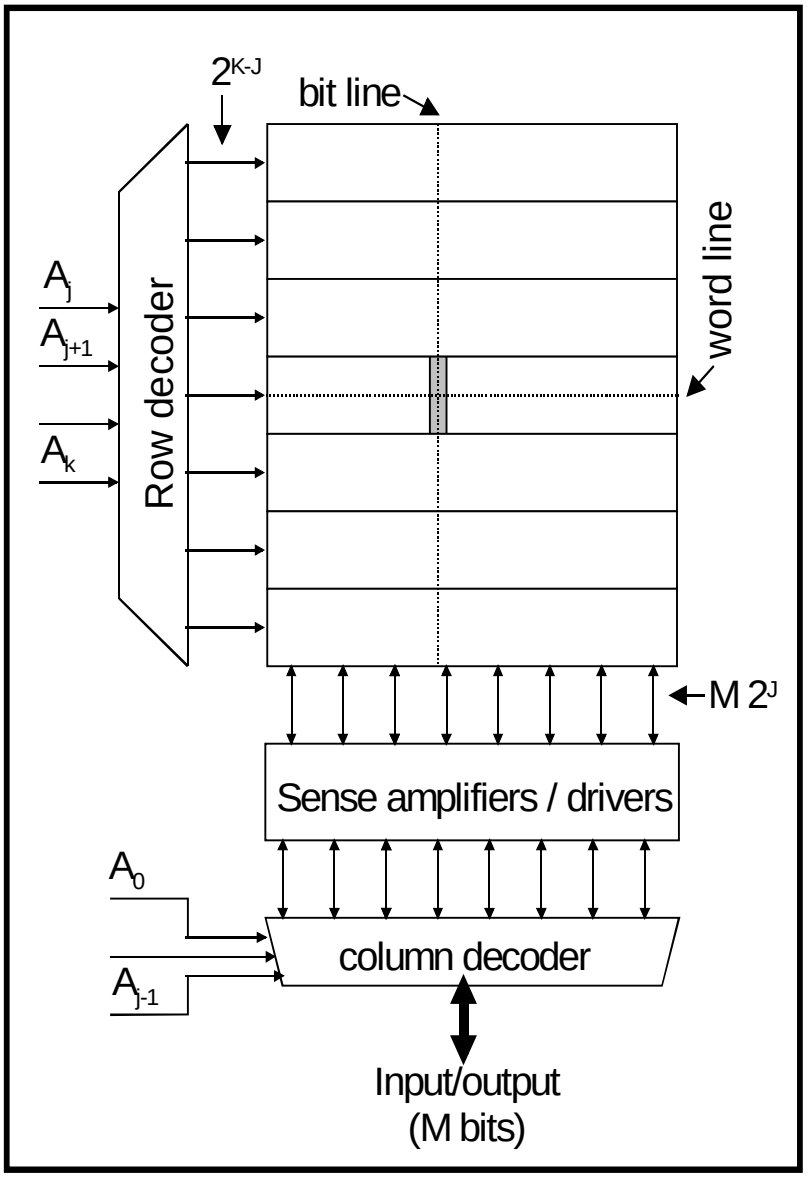
Memory architecture

- The memory is organized in N words, each of M bits wide
- One word at a time is selected for read/write using a select signal
- A decoder is used to convert a binary encoded address into a single active word select line
- This structure is not practical, it results in very big aspect ratios



Memory architecture

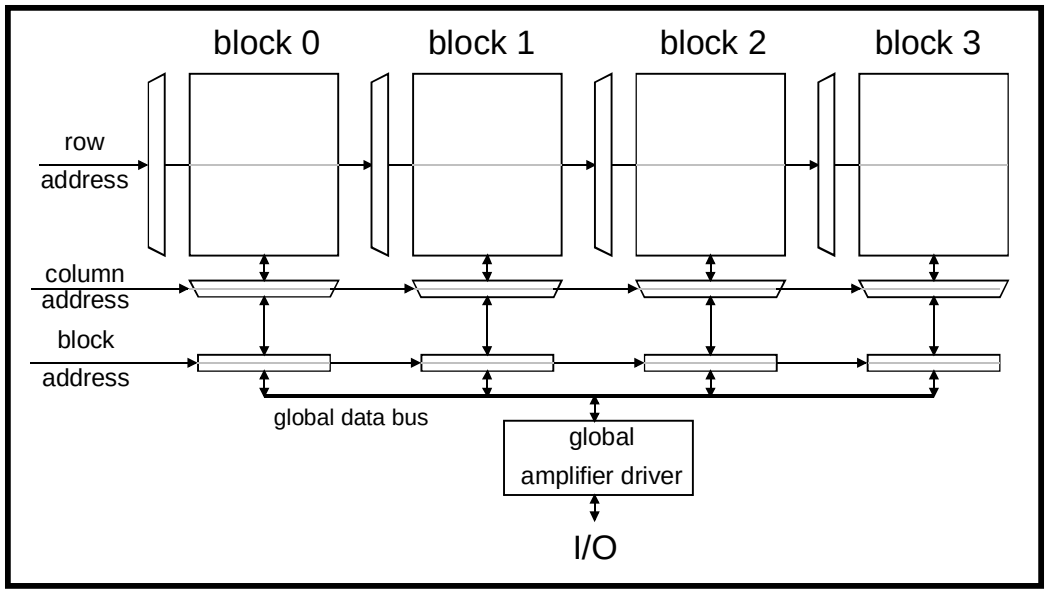
- Memories are organized to be almost square in layout:
  - Multiple words are stored in the same row and selected simultaneously
  - The correct word is then selected by the column decoder
  - The word address is split in two fields:
    - row address: enables one row for R/W
    - column address: selects a word within a row
  - Even this structure is impractical for memories bigger than 256Kbits



Memory architecture

- The silicon area of large memory cells is dominated by the size of the memory core, it is thus crucial to keep the size of the basic storage cell as small as possible
- The storage cell area is reduced by:
  - reducing the driving capability of the cell (small devices)
  - reducing the logic swing and the noise margins
- Consequently, sense amplifiers are used to restore full rail-to-rail amplitude

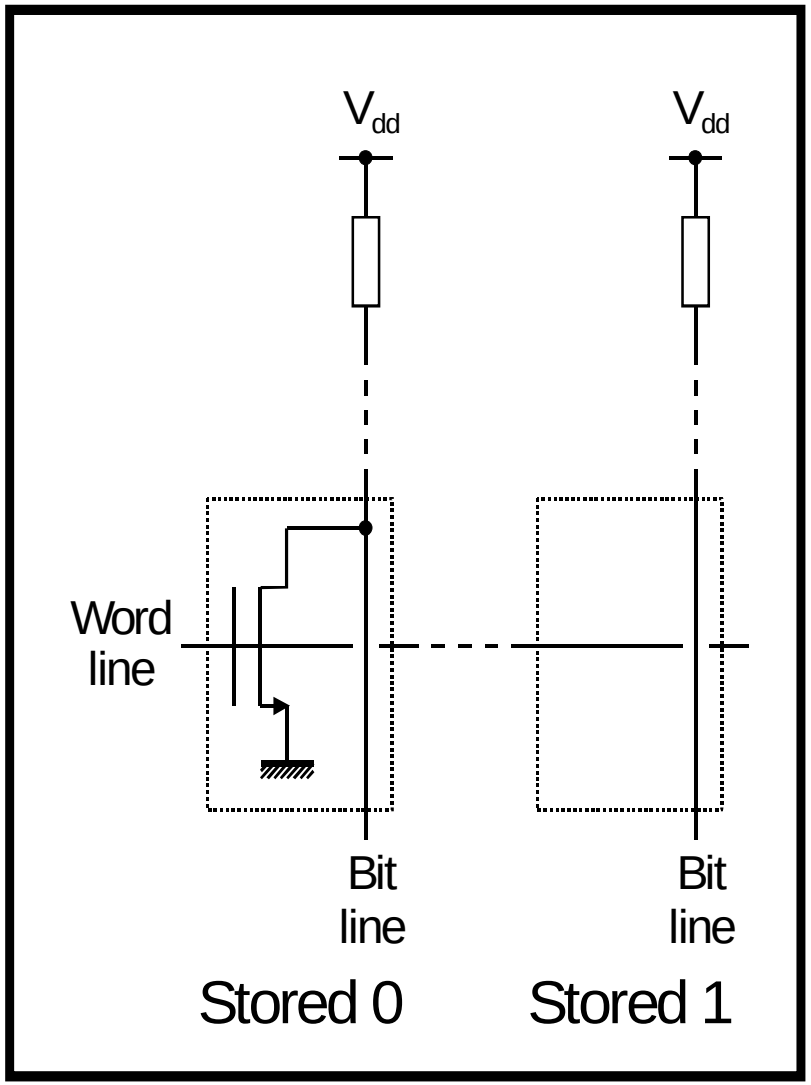
Memory architecture



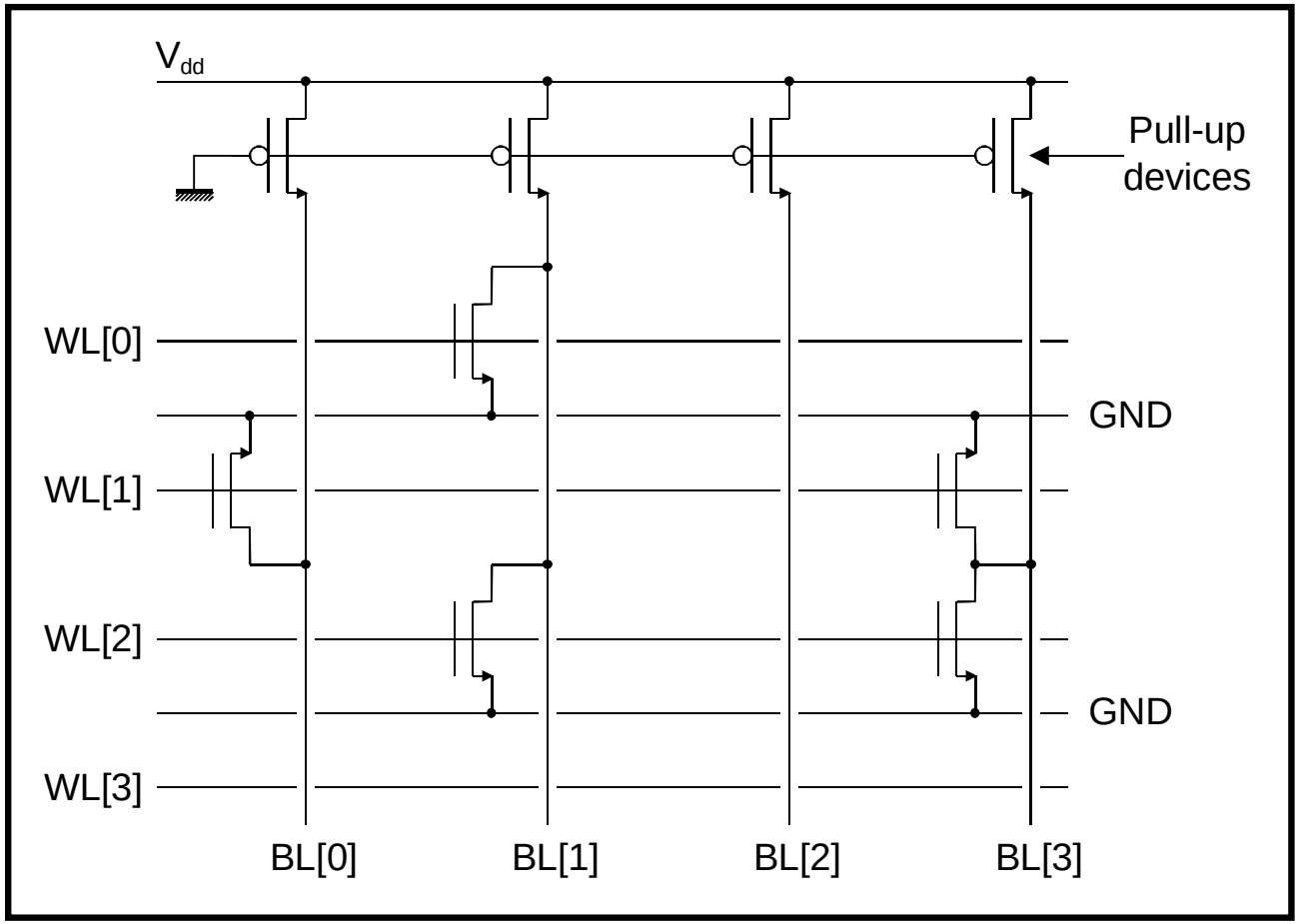
- Large memories start to suffer from speed degradation due to wire resistance and capacitive loading of the bit and word lines
- The solution is to split the memory into “small” memory blocks
- That allows to:
  - use small local word and bit lines $\Rightarrow$ faster access time
  - power down sense amplifiers and disable decoders of non-active memory blocks $\Rightarrow$ power saving

Read-only memories

- Because the contents is permanently fixed the cell design is simplified
- Upon activation of the word line a 0 or 1 is presented to the bit line:
  - If the NMOS is absent the word line has no influence on the bit line:
    - The word line is pulled-up by the resistor
    - A 1 is stored in the “cell
  - If the NMOS is present the word line activates the NMOS:
    - The word line is pulled-down by the NMOS
    - A 0 is stored in the cell
- The NMOS isolates the bit from the word line



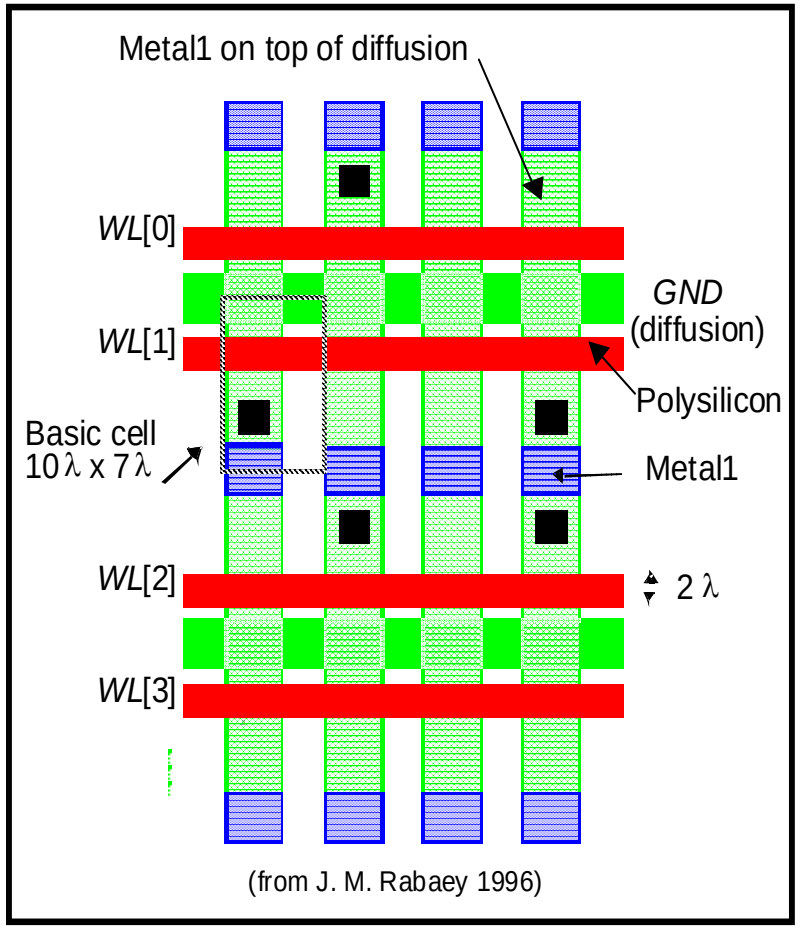
Read-only memories



- A ground contact has to be provided for every cell
  - a ground rail has to be routed through the cell
  - the area penalty can be shared between two neighbor cells:
    - the odd rows are mirrored around the horizontal axis

Read-only memories

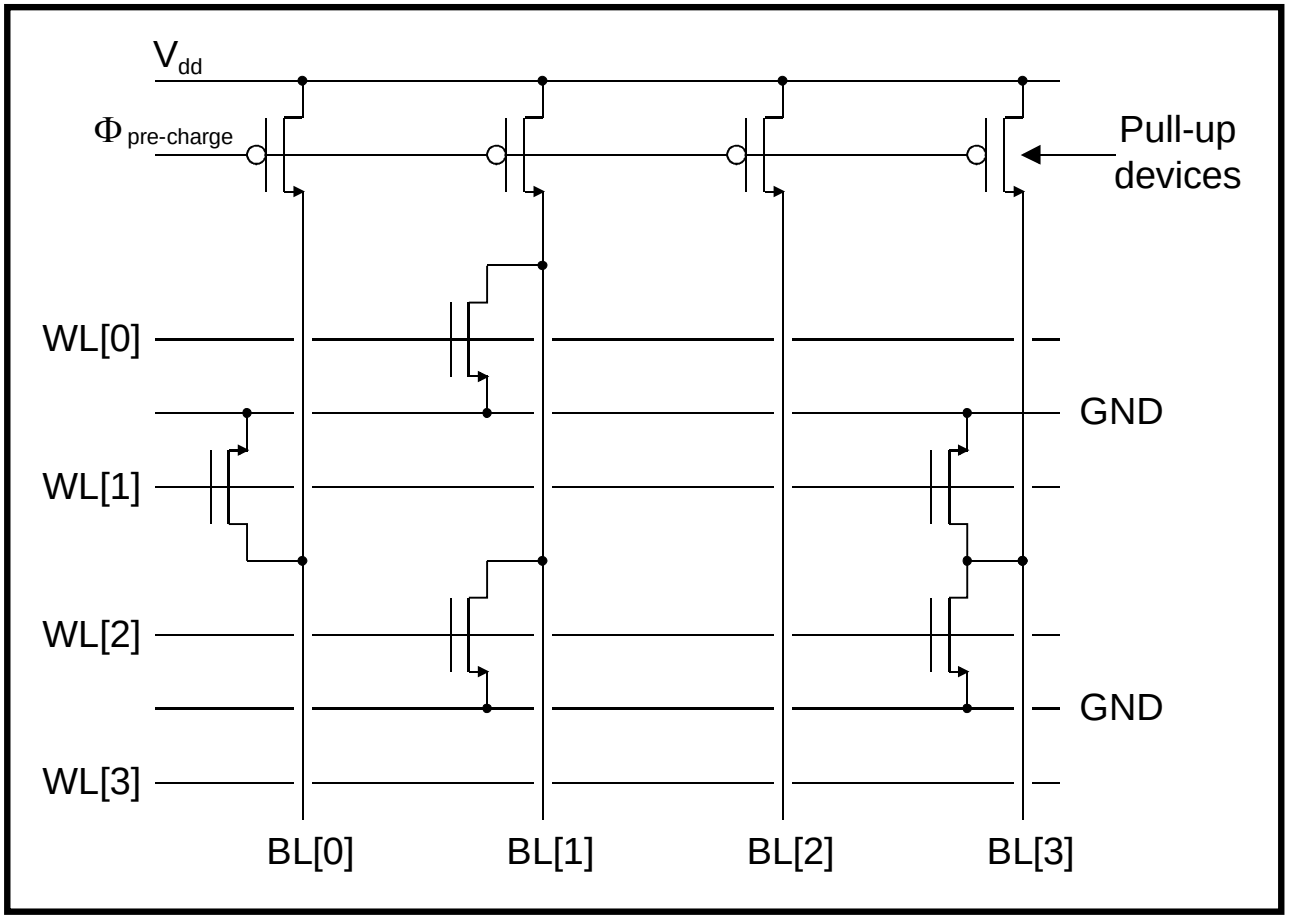
- Use close to minimum size pull-down devices to:
  - make the cell size small
  - reduce the bit line capacitance
- R(pull-up) > R(pull-down) to:
  - ensure adequate low level
- Since for large memories the bit line capacitance can be of the order of pF's, low to high transitions will be slow
- A wider pull-up device can be used resulting in a higher V_{OL}
  - this reduces the noise margin but speeds the low-to-high transition
  - to interface with external logic, a sense amplifier is required to restore the logic levels
  - an inverter with adjusted switching threshold can be used as a sense amplifier



- 0 ⇒ metal-to-diffusion contact
- 1 ⇒ no metal-to-diffusion contact
- only the contact mask layer is used to program the memory array

Read-only memories

- Disadvantages:
  - V_{OL} depends on the ratio of the pull-up/pull-down devices
  - A static current path exists when the output is low causing high power dissipation in large memories
- Solution:
  - Use pre-charged logic
  - Eliminates the static dissipation
  - Pull-up devices can be made wider
  - This is the most commonly used structure
- Operation:
  - The bit lines are first pre-charged by the pull-up devices
  - during this phase the word lines must be disabled
  - Then, the word lines are activated (word evaluation) during this phase the pull-up devices are off

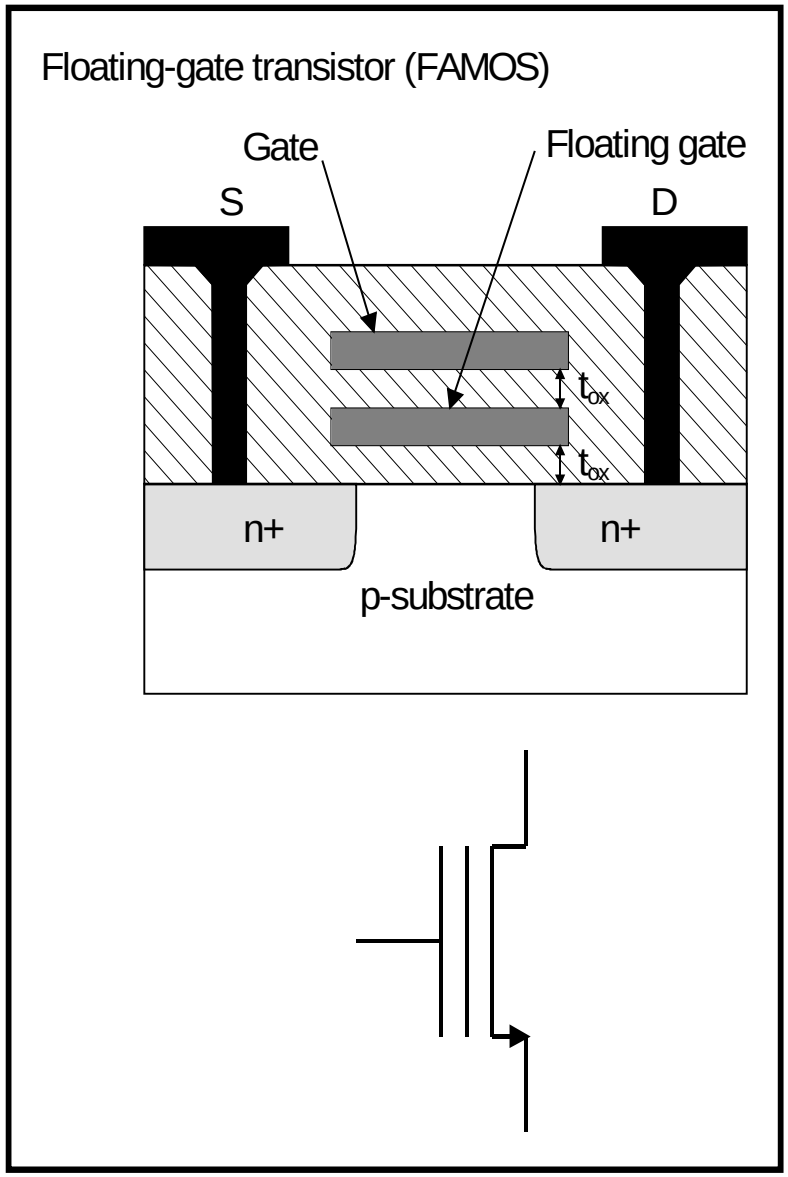


Nonvolatile read-write memories

- The same architecture as a ROM memory
- The pull-down device is modified to allow control of the threshold voltage
- The modified threshold is retained "indefinitely":
  - The memory is nonvolatile
- To reprogram the memory the programmed values must be erased first
- The "heart" of NVRW memories is the Floating Gate Transistor (FAMOS)
- Basis for FLASH memory

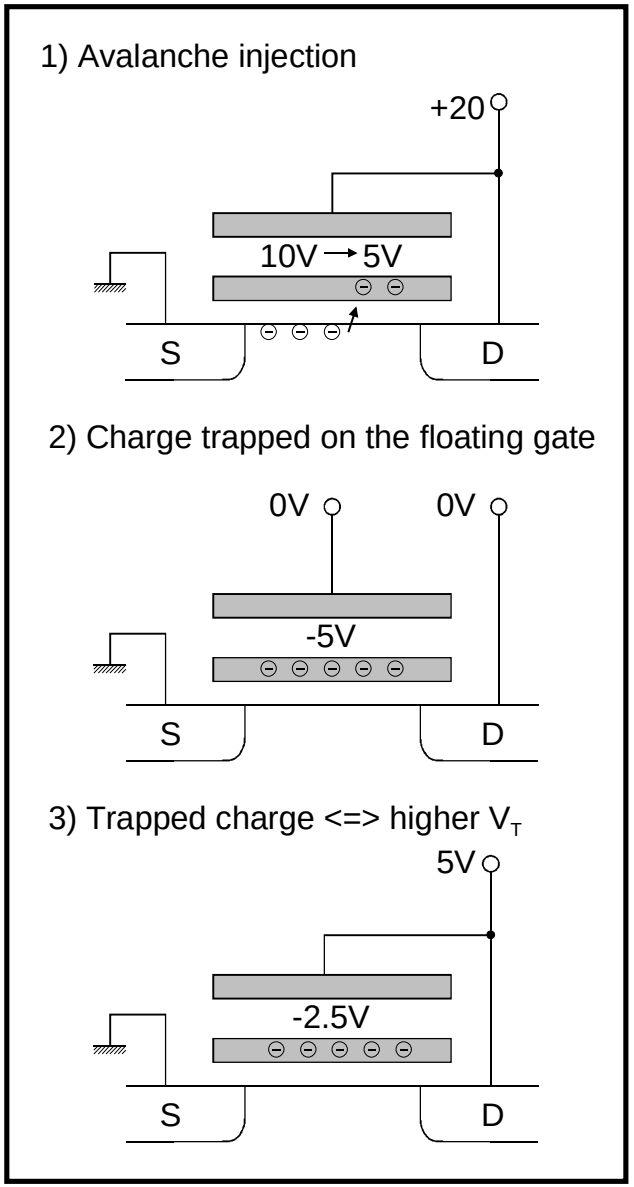
Nonvolatile read-write memories

- A floating gate is inserted between the gate and the channel
- The device acts as a normal transistor
- However, its threshold voltage is programmable
- Since the t_{ox} is doubled, the transconductance is reduced to half and the threshold voltage increased



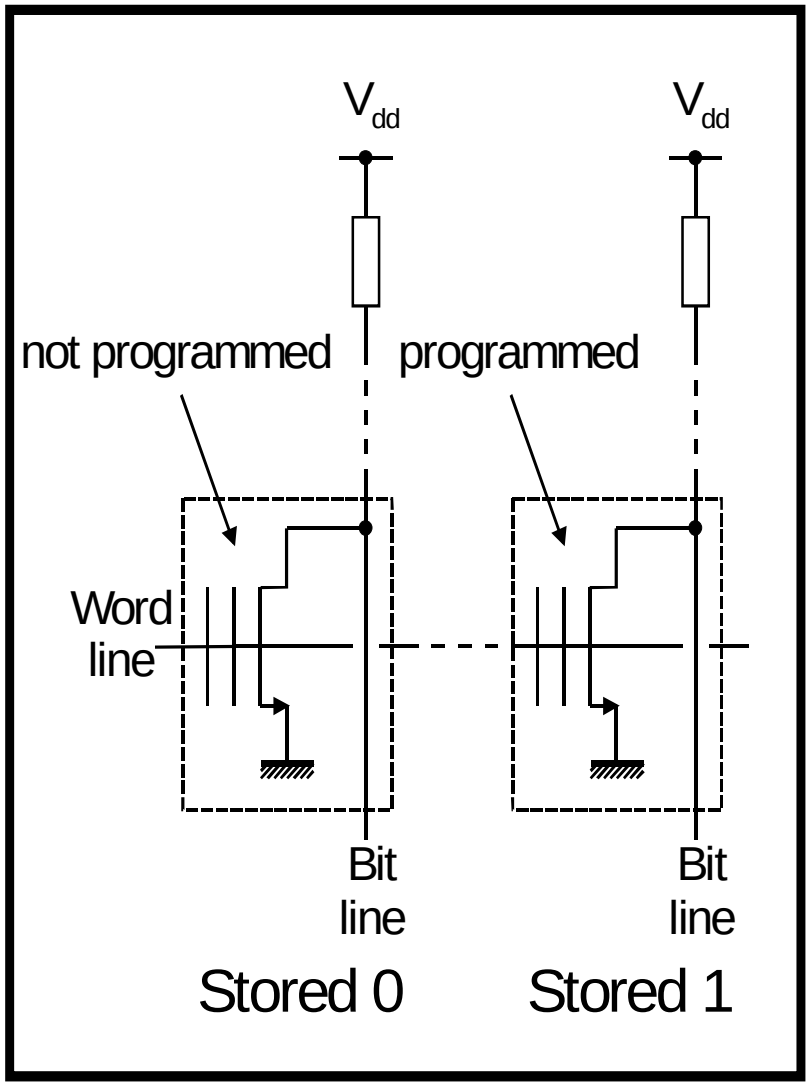
Nonvolatile read-write memories

- Programming the FAMOS:
  - A high voltage is applied between the source and the gate-drain
  - A high field is created that causes avalanche injection to occur
  - Electrons traverse the first oxide and get trapped on the floating gate (t_{ox} = 100nm)
  - Trapped electrons effectively drop the floating gate voltage
  - The process is self limiting: the building up of gate charge eventually stops avalanche injection
  - The FAMOS with a charged gate is equivalent to a higher V_T device
  - Normal circuit voltages can not turn a programmed device on



Nonvolatile read-write memories

- The non-programmed device can be turned on by the word line thus, it stores a “0”
- The word line high voltage can not turn on the programmed device thus, it stores a “1”
- Since the floating gate is surrounded by SiO_2, the charge can be stored for many years

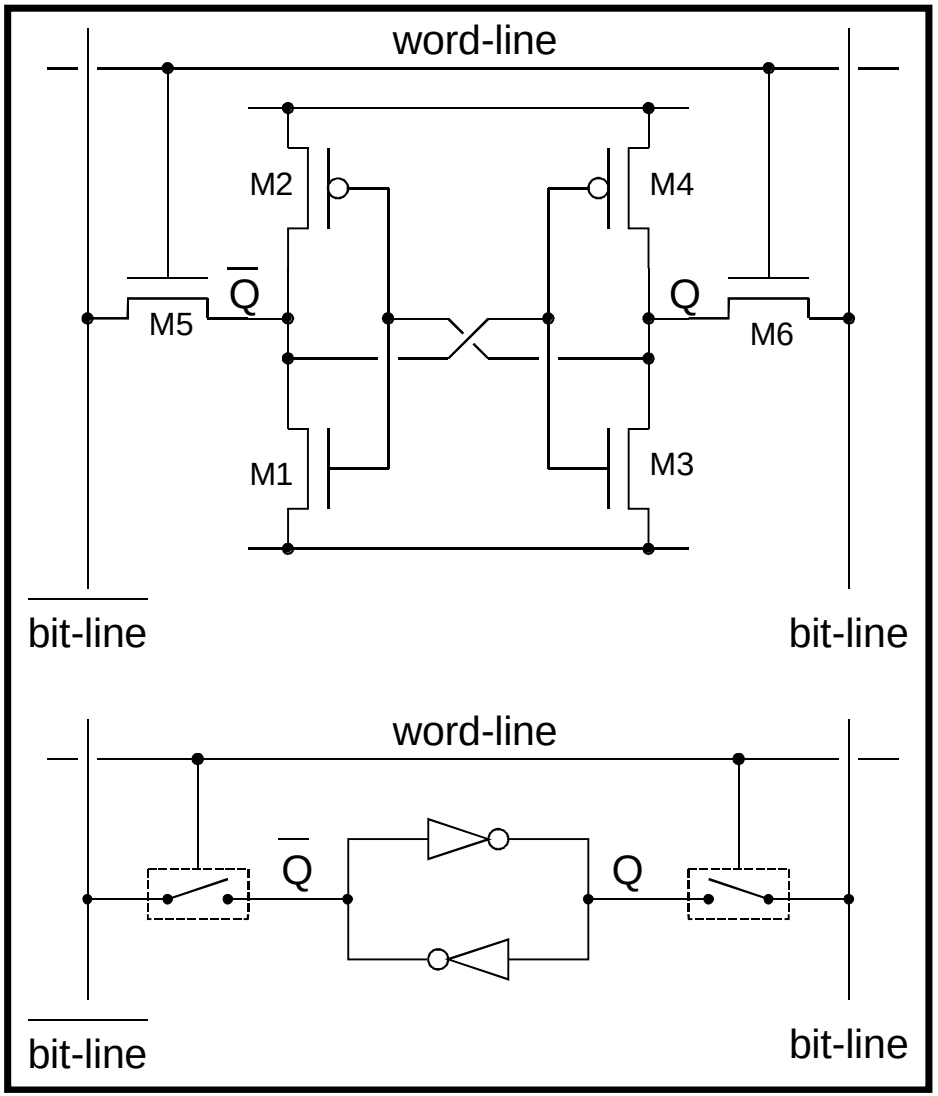


Nonvolatile read-write memories

- Erasing the memory contents (EPROM):
  - Strong UV light is used to erase the memory:
    - UV light renders the oxide slightly conductive by direct generation of electron-hole pairs in the SiO_2
  - The erasure process is slow (several minutes)
  - Programming takes 5-10μs/word
  - Number of erase/program cycles limited (<1000)
- Electrically-Erasable PROM (E^2PROM)
  - A reversible tunneling mechanism allows E^2PROM's to be both electrically programmed and erased

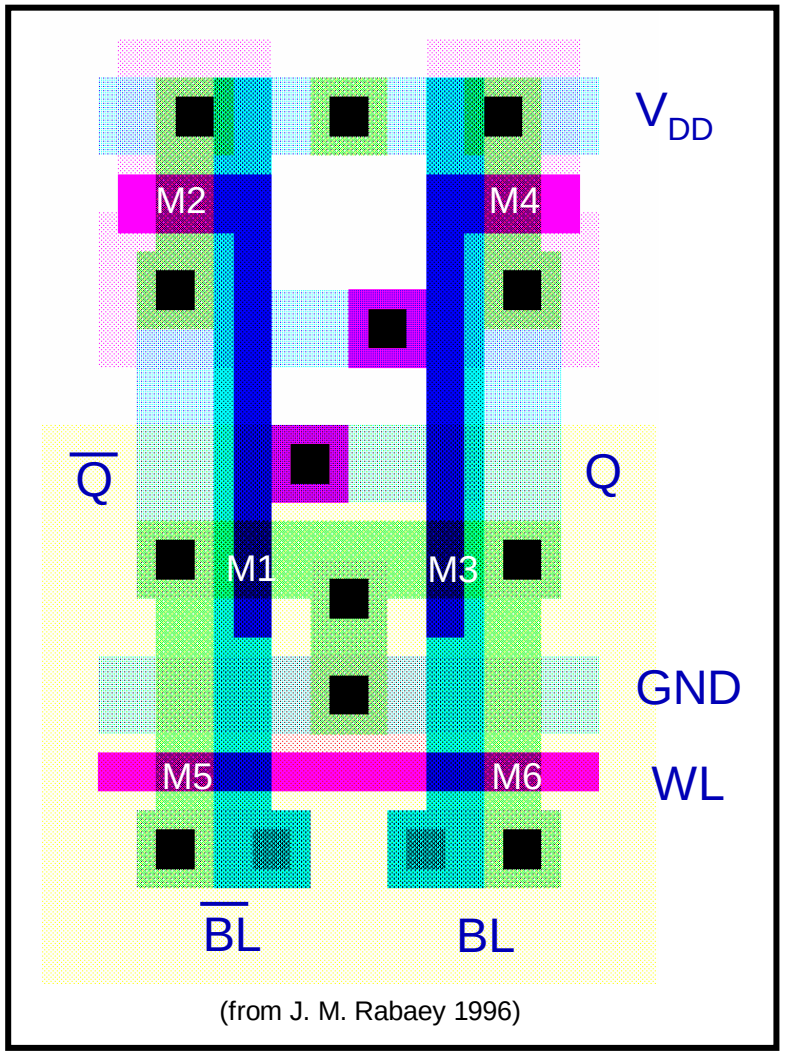
Read-write memories

- Static Read-Write Memories (SRAM):
  - data is stored by positive feedback
  - the memory is volatile
- The cell use six transistors
- Read/write access is enabled by the word-line
- Two bit lines are used to improve the noise margin during the read/write operation
- During read the bit-lines are pre-charged to $V_{dd}/2$:
  - to speedup the read operation
  - to avoid erroneous toggling of the cell



Read-write memories

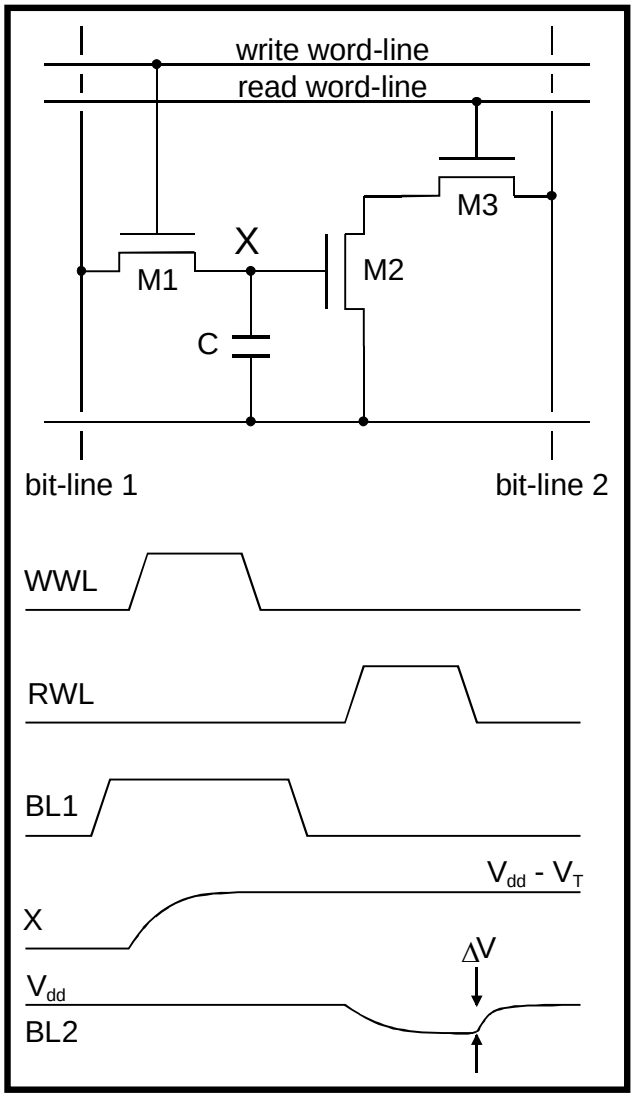
- SRAM performance:
  - The read operation is the critical one:
    - It involves discharging or charging the large bit-line capacitance through the small transistors of the cell
  - The write time is dominated by the propagation delay of the cross-coupled inverter pair
  - The six-transistor cell is not area efficient:
    - It requires routing of two power lines, two bit lines and a word line
    - Most of the area is taken by wiring and interlayer contacts



(from J. M. Rabaey 1996)

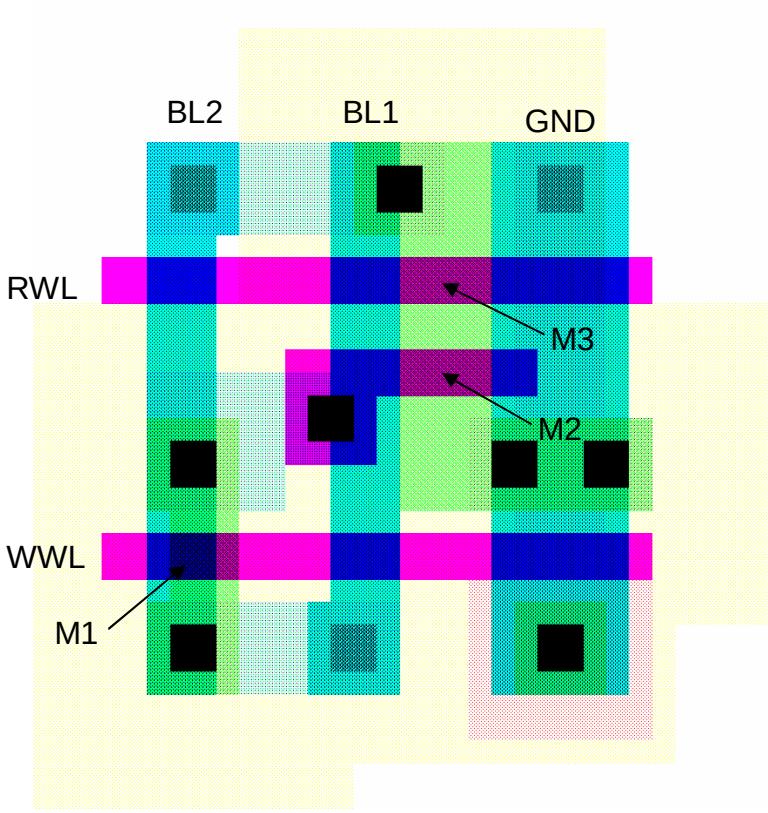
Read-write memories

- Dynamic Random-Access Memory (DRAM)
  - In a dynamic memory the data is stored as charge in a capacitor
- Tree-Transistor Cell (3T DRAM):
  - Write operation:
    - Set the data value in bit-line 1
    - Assert the write word-line
    - Once the WWL is lowered the data is stored as charge in C
  - Read operation:
    - The bit-line BL2 is pre-charged to Vdd
    - Assert the read word-line
    - if a 1 is stored in C, M2 and M3 pull the bit-line 2 low
    - if a 0 is stored C, the bit-line 2 is left unchanged



Read-write memories

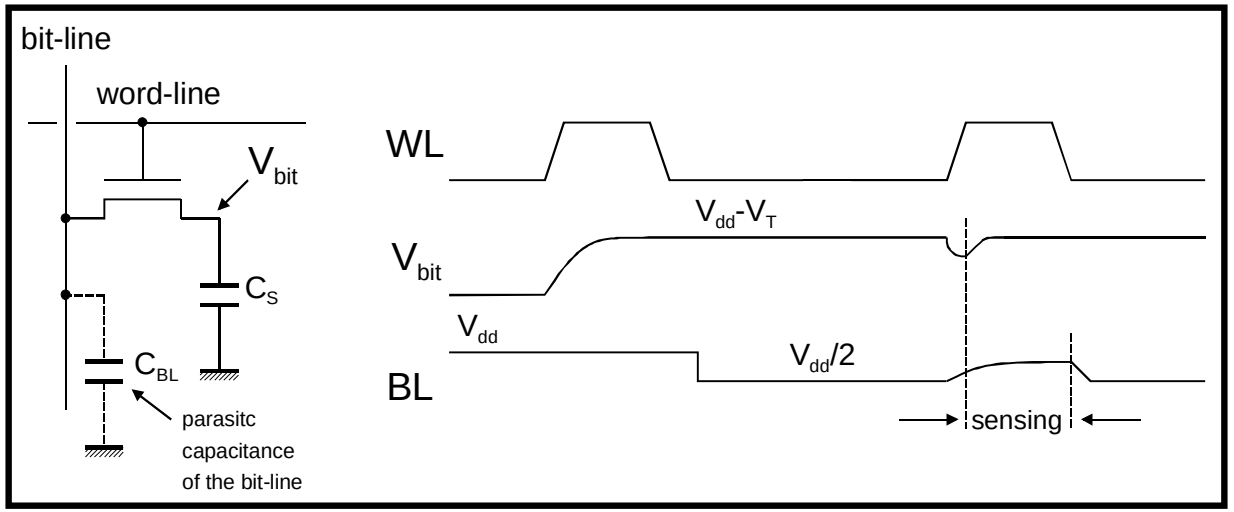
- The cell is inverting
- Due to leakage currents the cell needs to be periodically refreshed (every 1 to 4ms)
- Refresh operation:
  - read the stored data
  - put its complement in BL1
  - enable/disable the WWL
- Compared with an SRAM the area is greatly reduce:
  - SRAM $\Rightarrow$ 1092 λ^2
  - DRAM $\Rightarrow$ 576 λ^2
  - The are reduction is mainly due to the reduction of the number of devices and interlayer contacts



(from J. M. Rabaey 1996)

Read-write memories

- One-Transistor dynamic cell (1T DRAM)
  - It uses a single transistor and a capacitor
  - It is the most widely used topology in commercial DRAM's
- Write operation:
  - Data is placed on the bit-line
  - The word-line is asserted
  - Depending on the data value the capacitance is charged or discharged

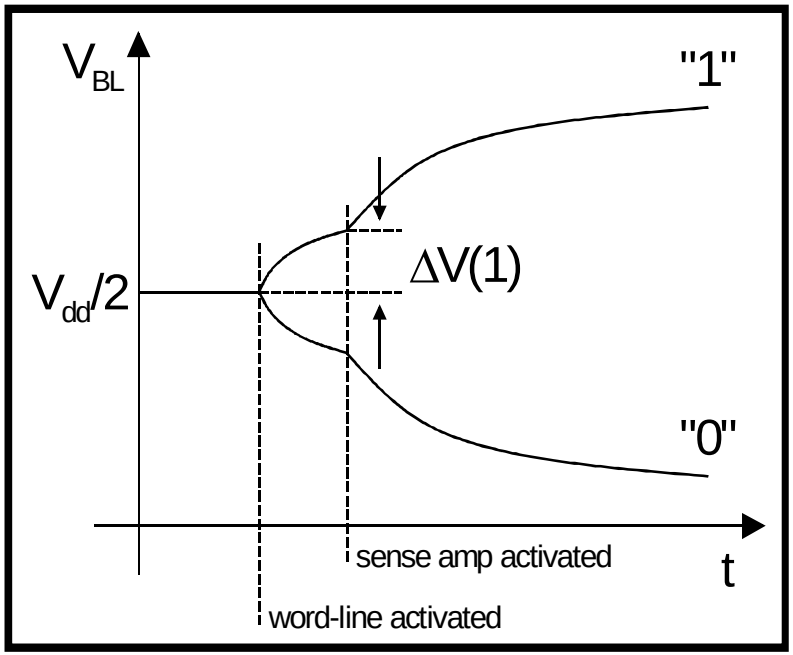


Read-write memories

- Read operation:
  - The bit-line is pre-charged to $V_{dd}/2$
  - The word-line is activated and charge redistribution takes place between C_S and the bit-line
  - This gives origin to a voltage change in the bit-line, the sign of which determines the data stored:

$$\Delta V = \left(V_{BIT} - \frac{V_{dd}}{2}\right)\frac{C_S}{C_S + C_{BL}}$$

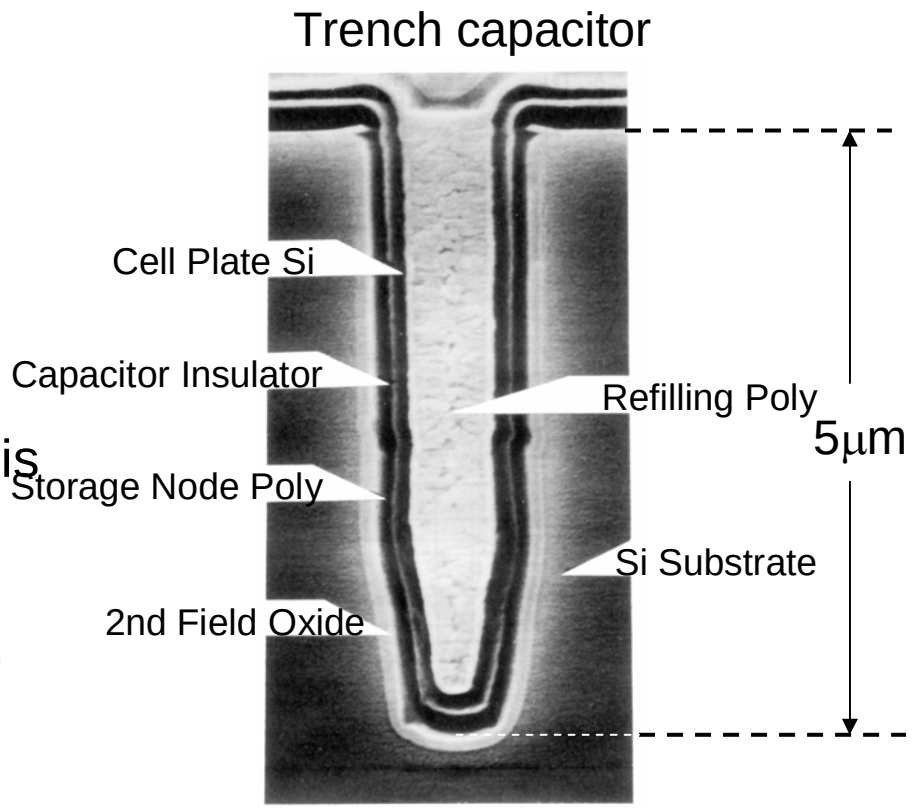
  - C_{BL} is 10 to 100 times bigger than $C_S \Rightarrow$ $\Delta V \cong 250mV$



  - The amount of charge stored in the cell is modified during the read operation
  - However, during read, the output of the sense amplifier is imposed on the bit line restoring the stored charge

Read-write memories

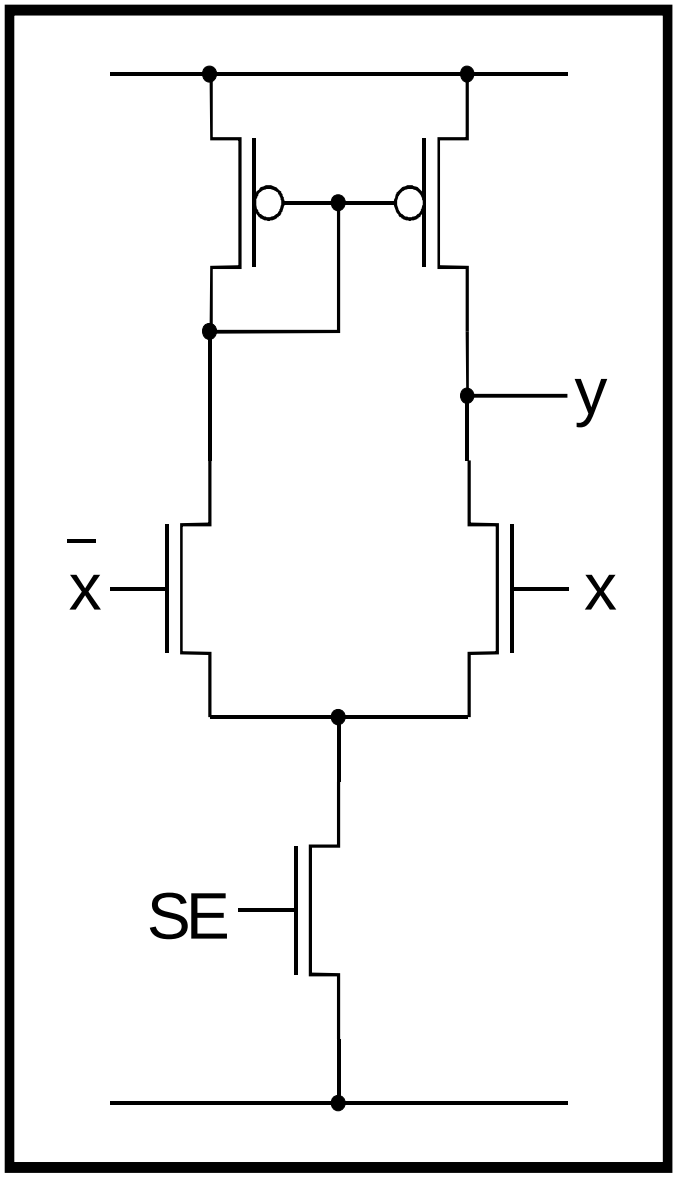
- Contrary to the previous cases a 1T cell requires a sense amplifier for correct operation
- Also, a relatively large storage capacitance is necessary for reliable operation
- A 1 is stored as V_{dd}-V_T. This reduces the available charge:
  - To avoid this problem the word-line can be bootstrapped to a value higher than Vdd



(from T. Mano et al., 1987)

Sense amplifiers

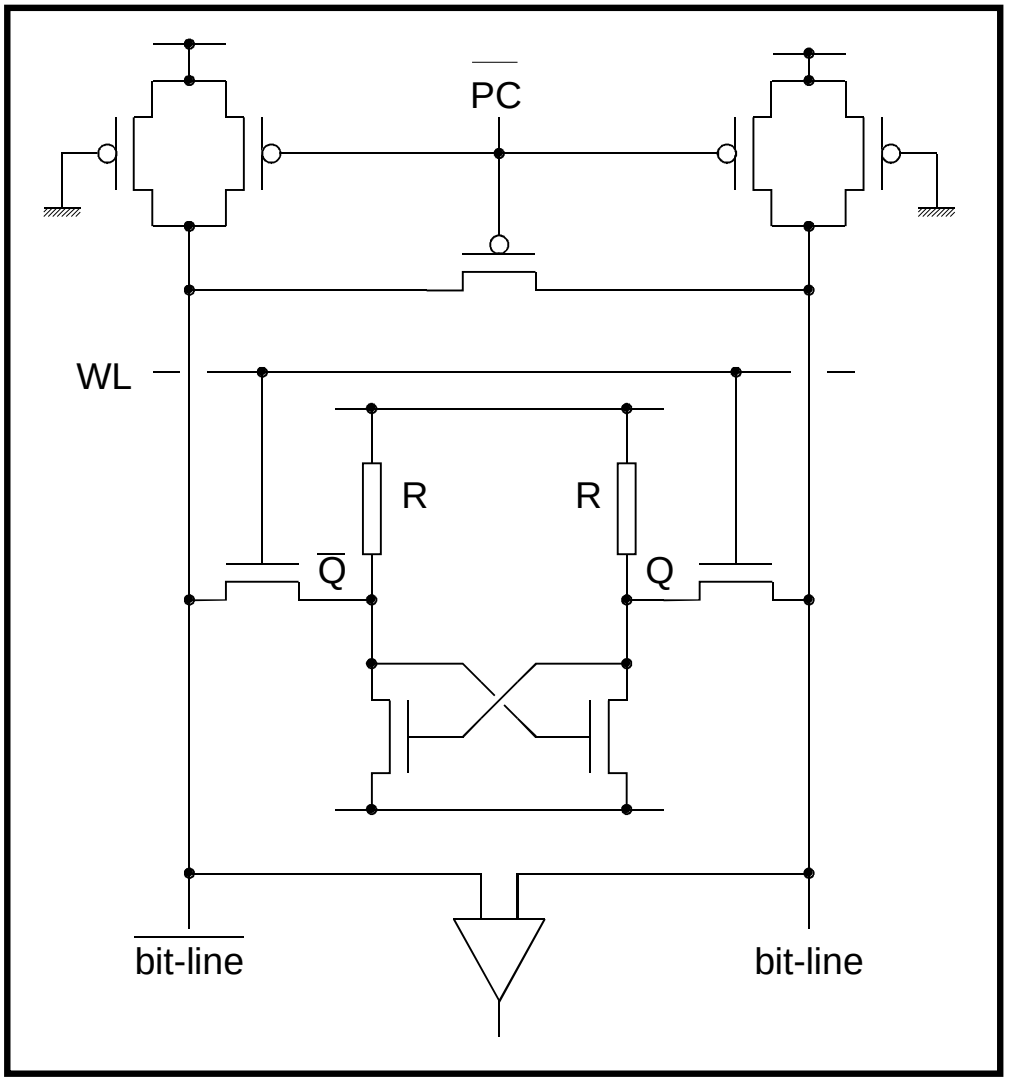
- Sense amplifiers improve the speed performance of the memory cell:
  - they compensate for the low driving capability of the cells
- Contribute to power reduction by allowing to use low signal swings on the heavily capacitive bit-lines
- They perform signal restoration in the refresh and read cycles of 1T dynamic memories
- They can be differential or single ended



Sense amplifiers

SRAM read cycle:

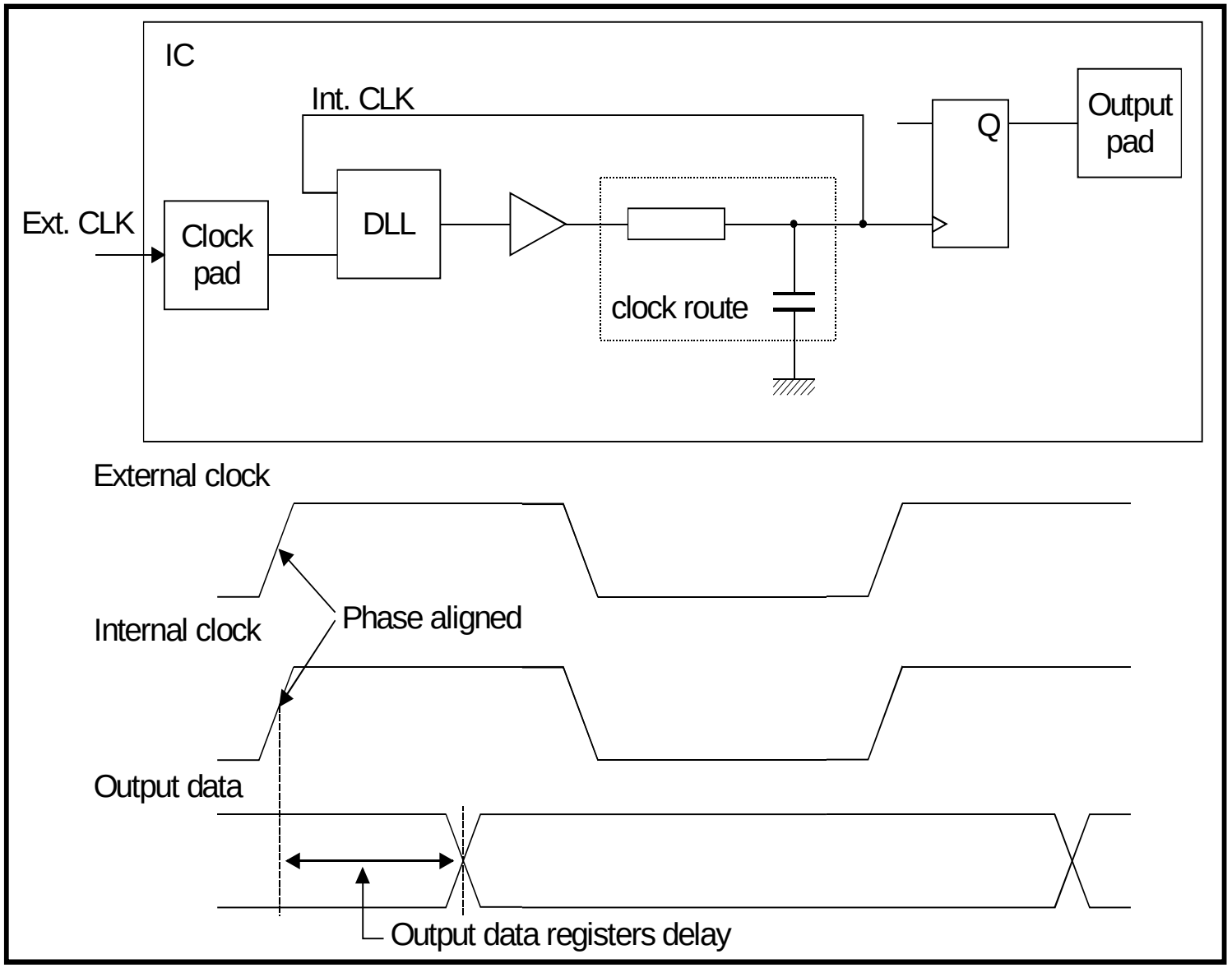
- pre-charge:
  - pre-charge the bit-lines to V_{dd} and make their voltages equal
- Reading:
  - disable the pre-charge devices
  - enable the word lines
  - once a minimum (≅0.5V) signal is built up in the bit-lines the sense amplifier is turned on
- The grounded PMOS loads limit the signal swing and facilitate the next pre-charge



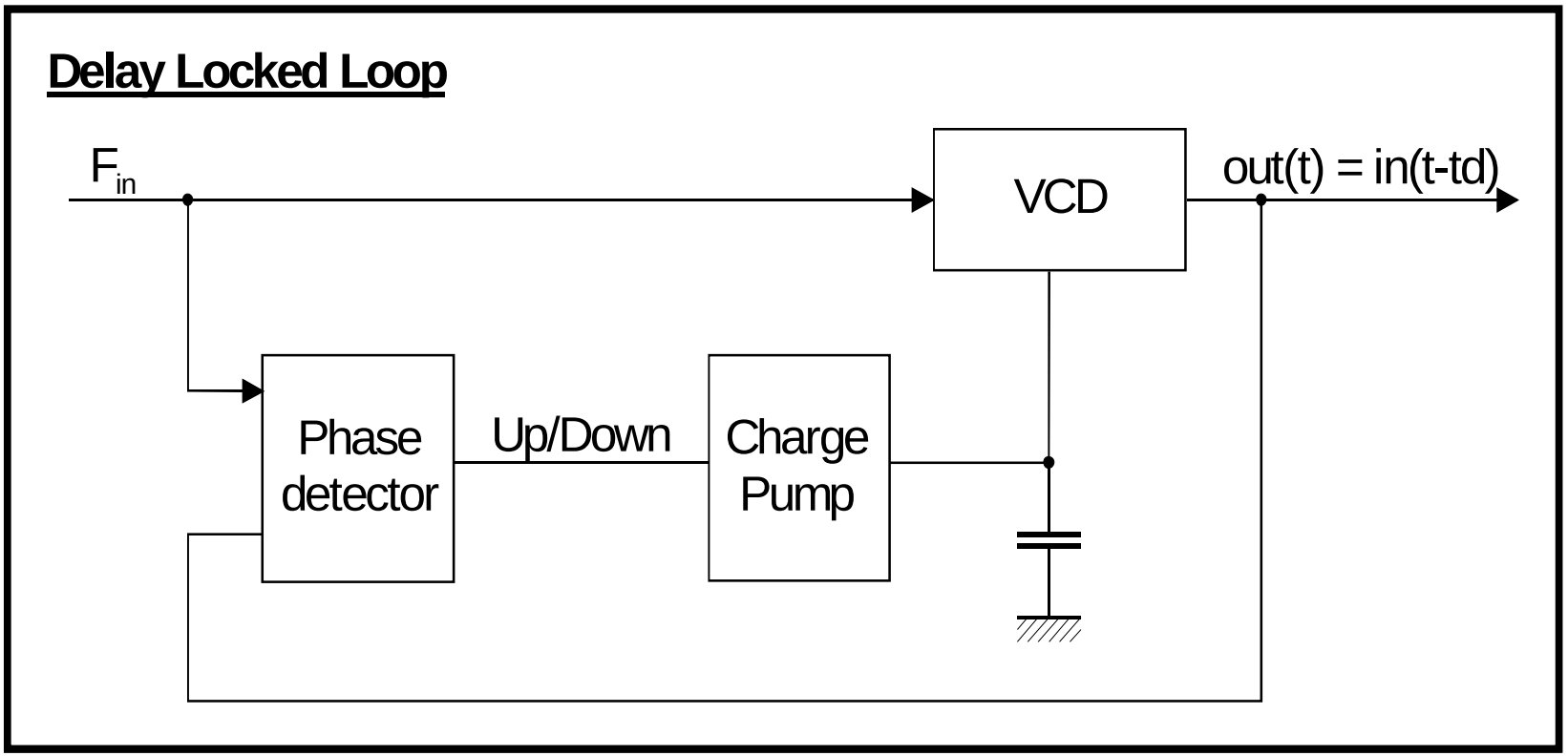
Delay and Phase locked loops

- Allows on chip clock multiplication
- Allows precise timing control inside chip
- Allows precise timing control with outside world

Delay locked loops

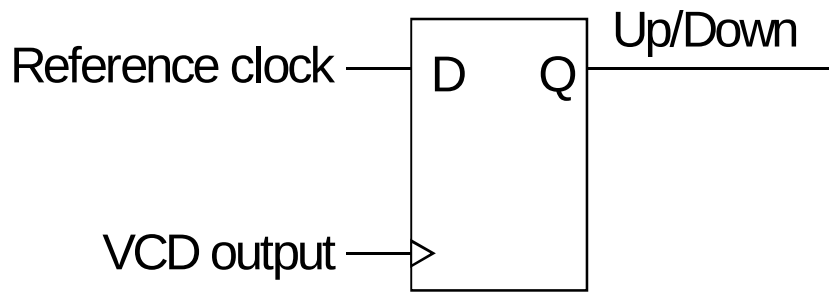


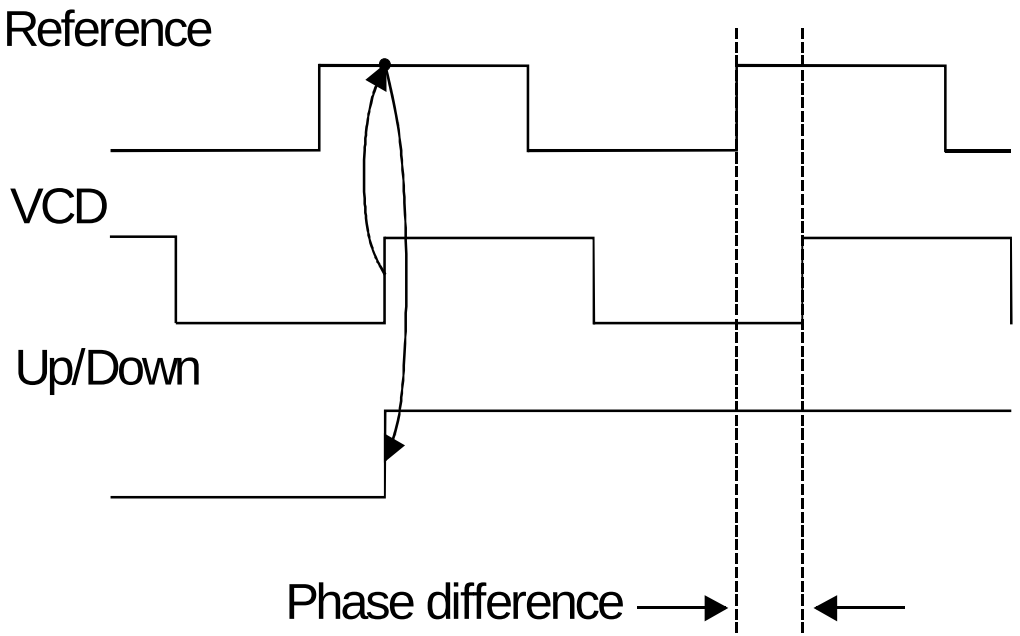
Delay locked loops



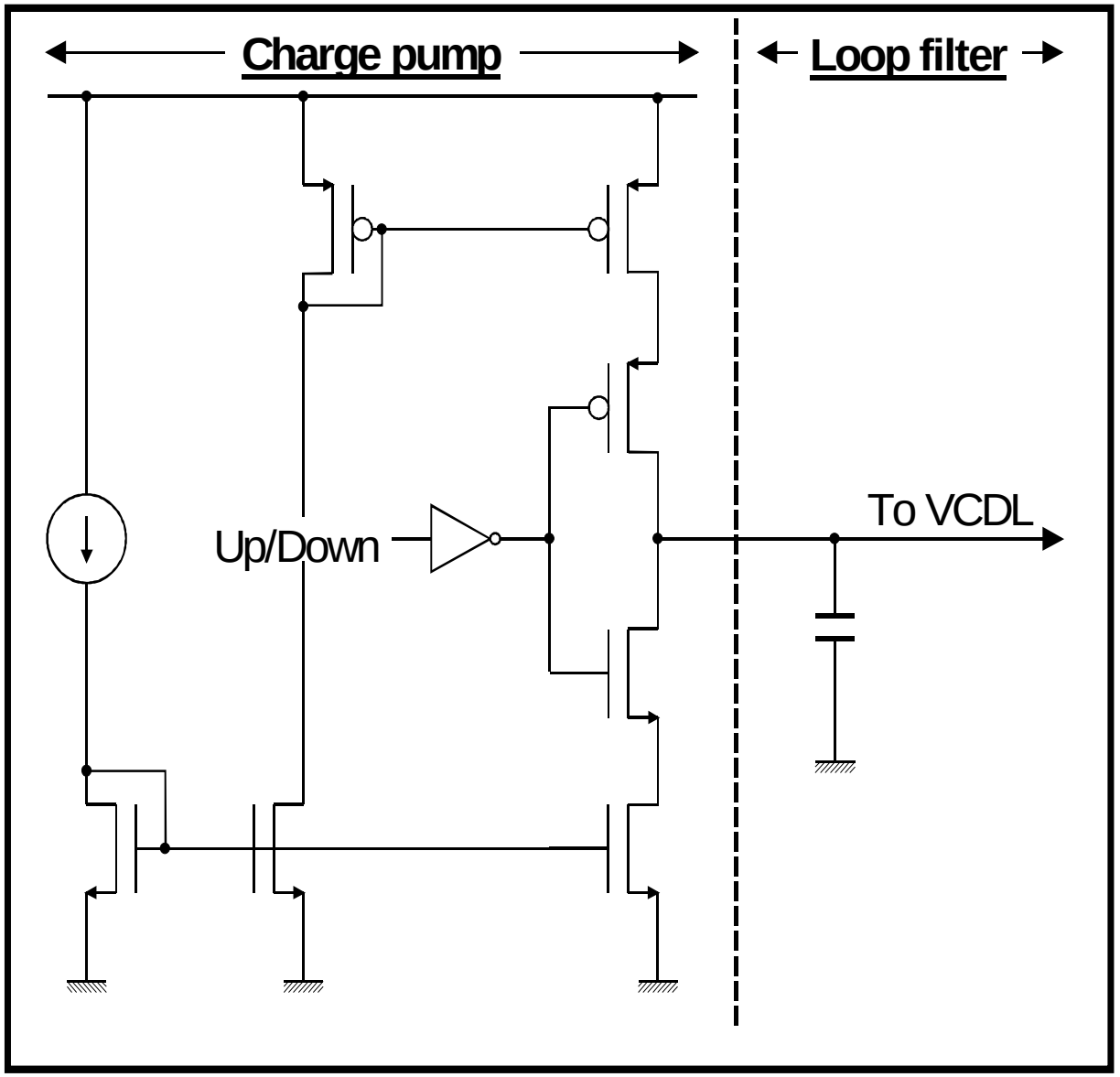
Delay locked loops

Phase detector

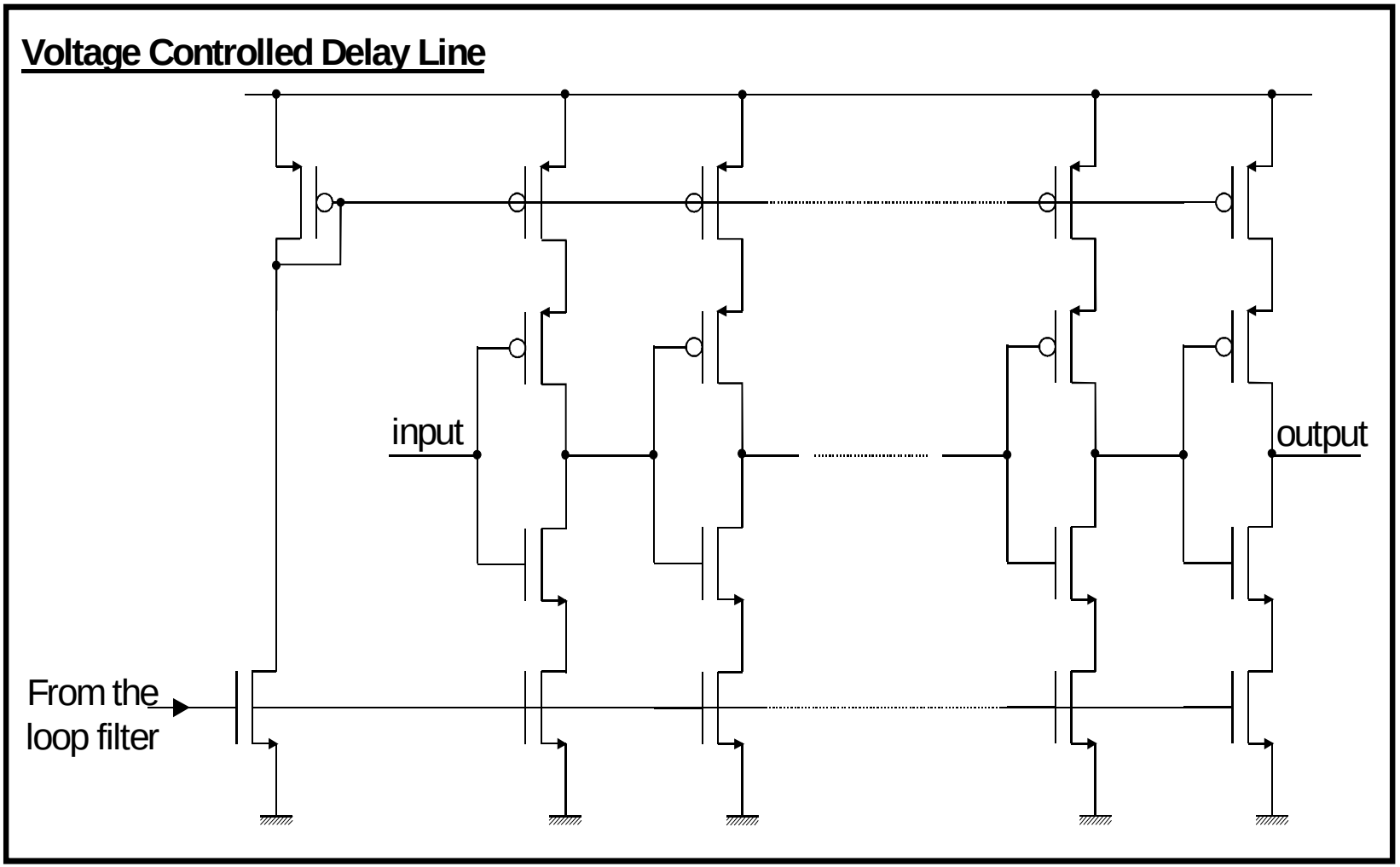




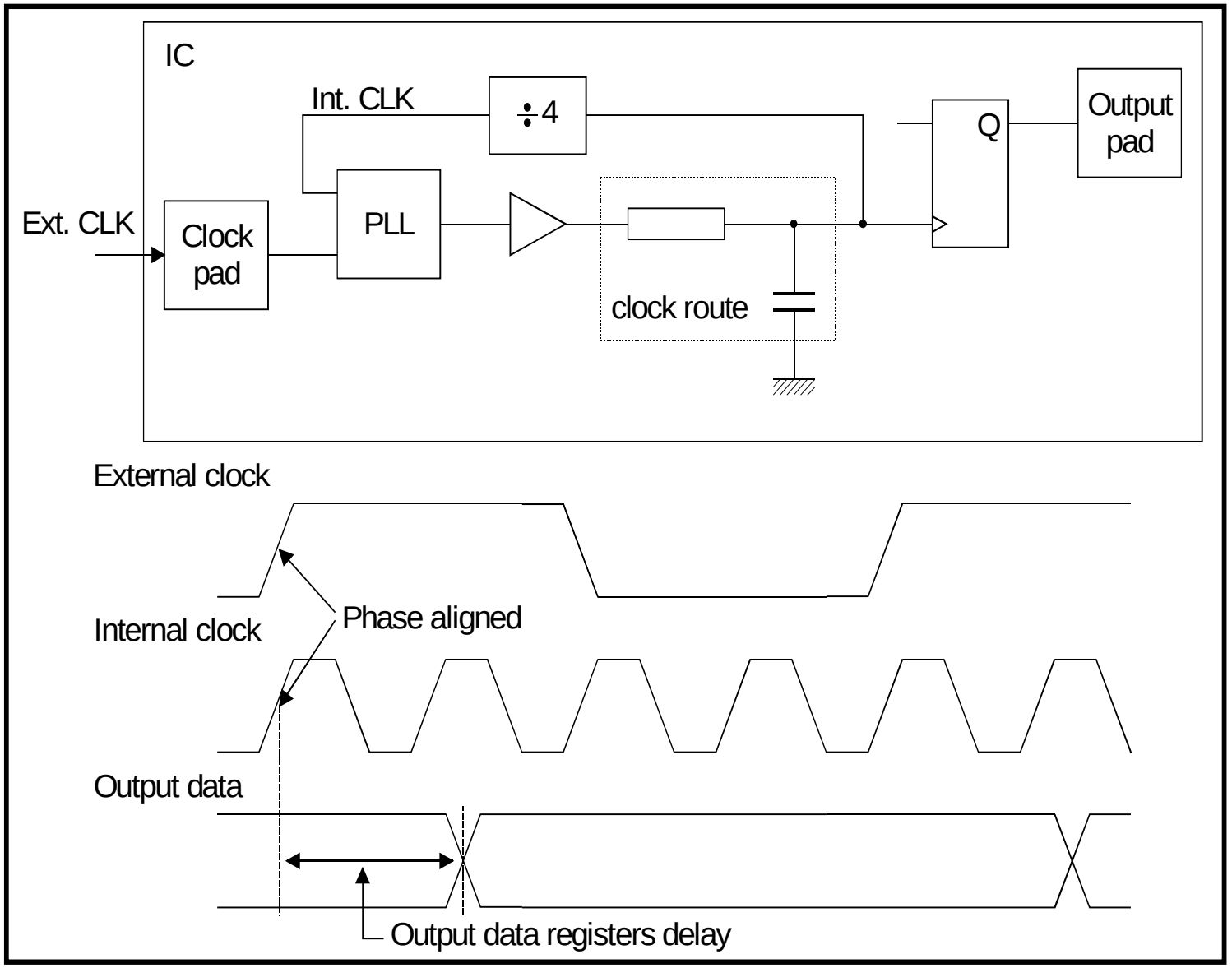
Delay locked loops



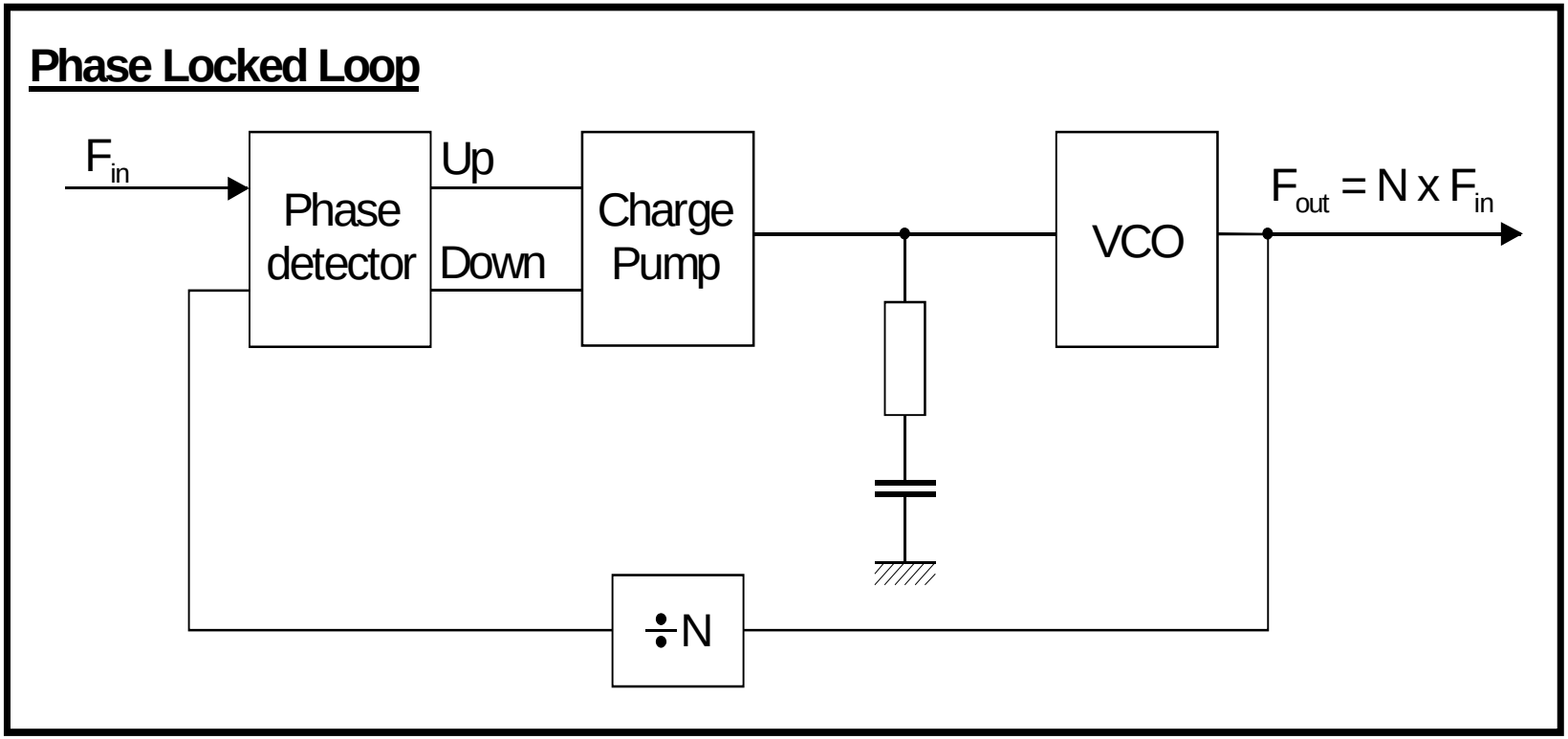
Delay locked loops



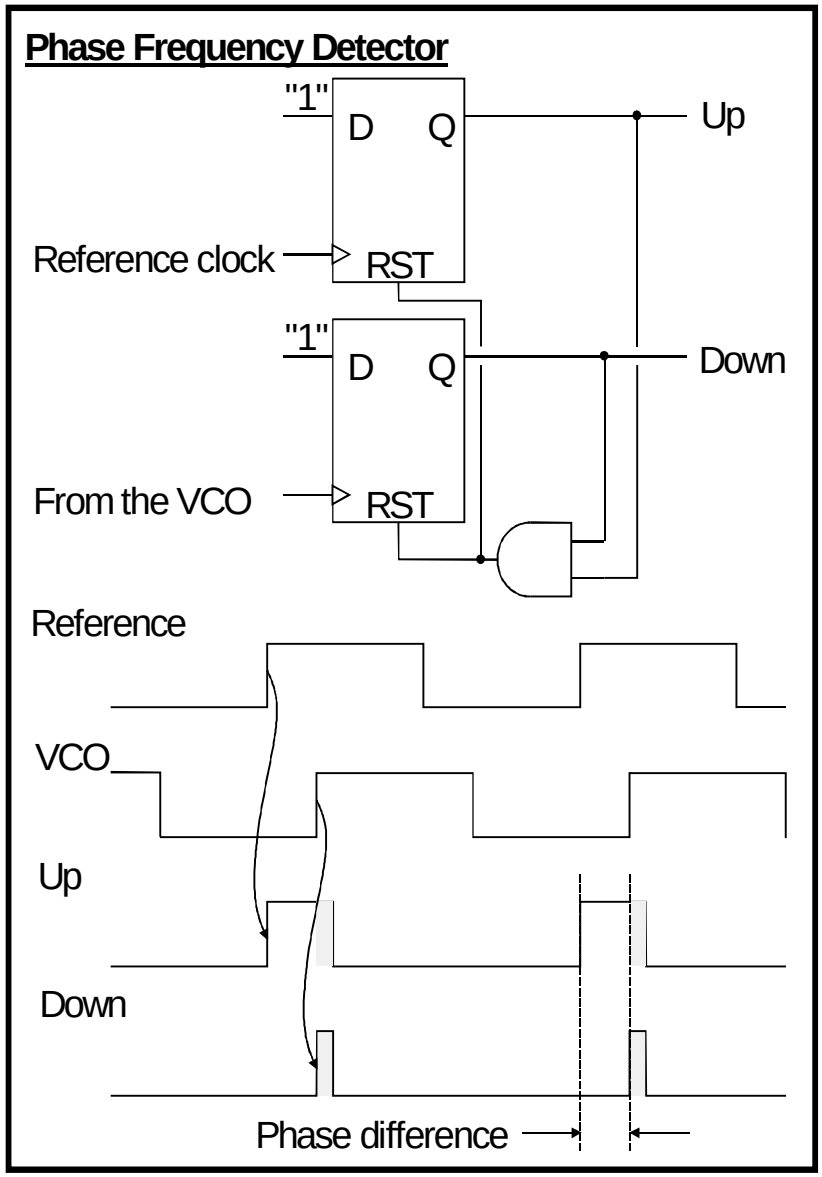
Phase Locked Loops



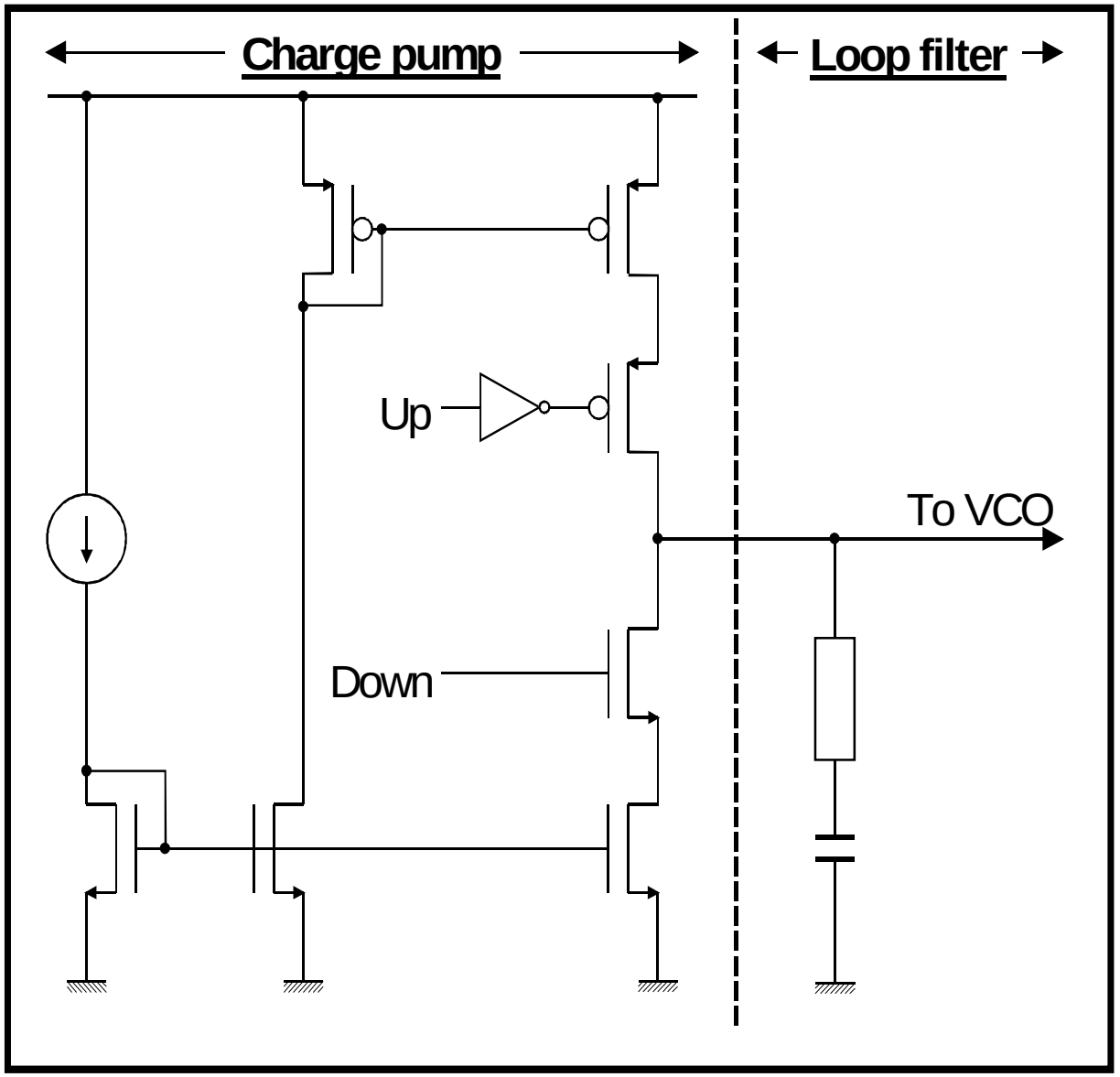
Phase locked loops



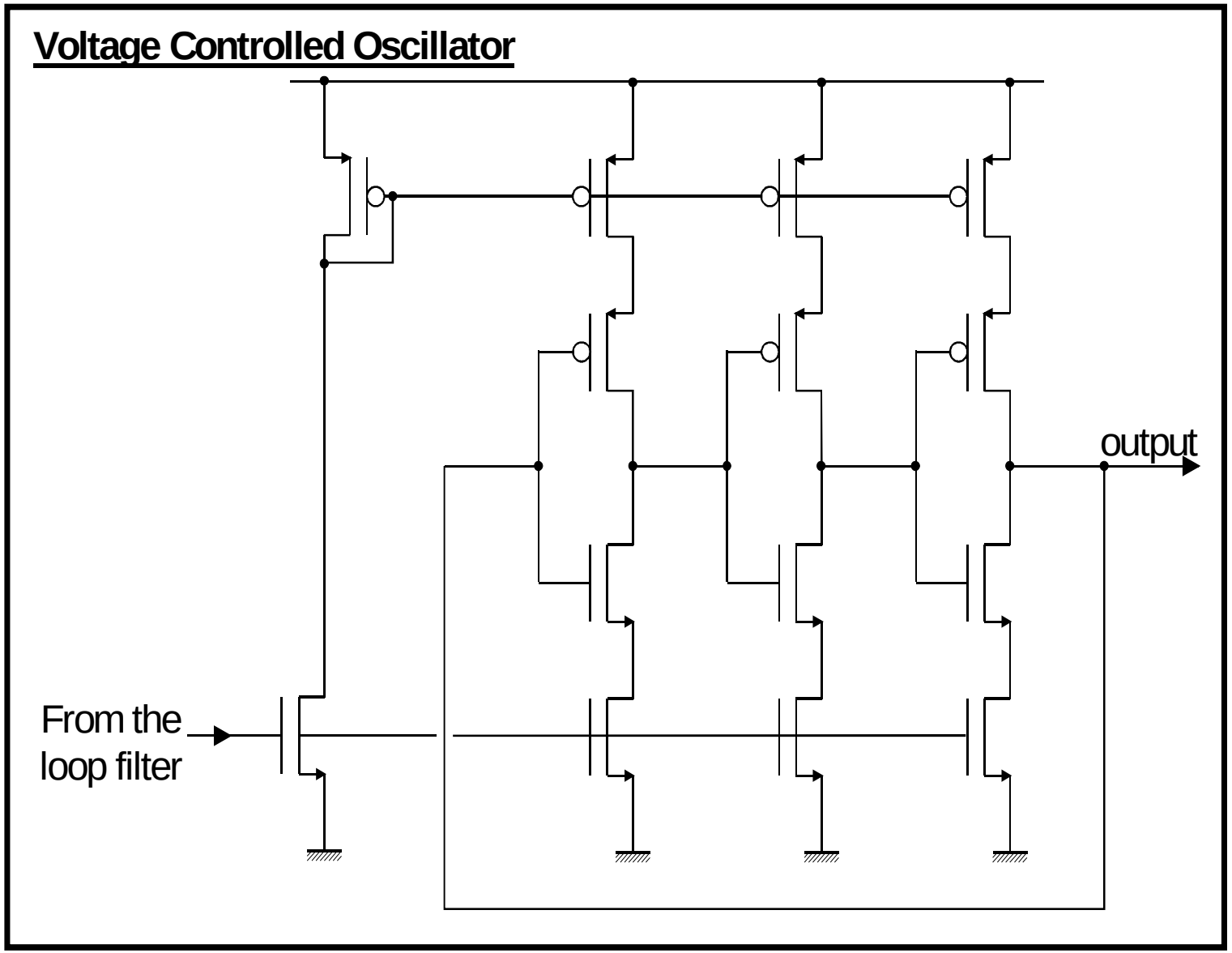
Phase locked loops



Phase locked loops



Phase locked loops



Recommended reading

Analogue

- **Design of Analog Integrated Circuits and Systems**
  by Kenneth R. Laker, Willy M.C. Sansen
  McGraw-Hill Higher Education; ISBN: 007036060X
- **Design of Analog CMOS Integrated Circuits**
  by Behzad Razavi
  McGraw-Hill Higher Education; ISBN: 0072380322

Digital

- **Digital Integrated Circuits: A Design Perspective**
  by Jan M. Rabaey
  Prentice Hall; ISBN: 0131786091
- **Principles of Cmos VLSI Design**
  by Neil H. E. Weste, Kamran Eshraghian
  Addison-Wesley Pub Co; ISBN: 0201533766

Analogue / Digital

- **CMOS Circuit Design, Layout, and Simulation**
  by R. Jacob Baker, Harry W. Li, David E. Boyce
  IEEE Press Series on Microelectronic Systems; IEEE; ISBN: 0780334167

Modeling

- **Operation & Modeling of the MOS Transistor**
  by Yannis Tsividis
  McGraw-Hill Higher Education; ISBN: 0070655235

On the Web

This Course:

http://paulo.moreira.free.fr

CERN Tutorials:

The CERN web site includes video recordings of lectures on engineering and physics.

http://humanresources.web.cern.ch/humanresources/external/training/tech/special/ELEC2002.asp

http://www.cern.ch/TechnicalTraining/special/FEED2002.asp

EMAIL:

Paulo.Moreira@cern.ch