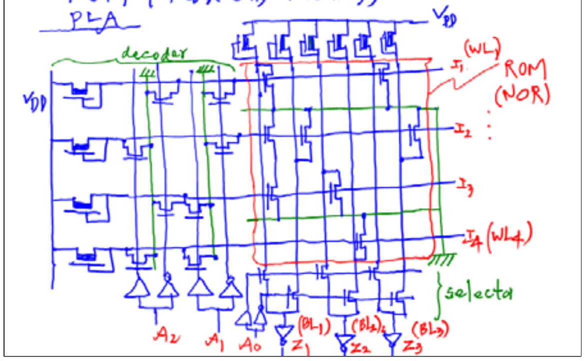
EE 222 Lecture 8 Feb. 1, 2018

ROM (Read Only Memory) – nMOS ROM

PLA



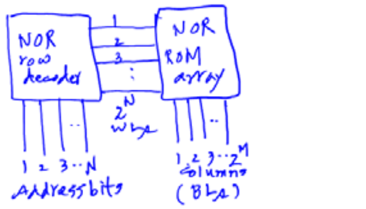
$$I_1 = \overline{\bar{A}_2 + \bar{A}_1}$$
$$I_2 = \overline{A_2 + \bar{A}_1}$$
$$I_3 = \overline{\bar{A}_2 + A_1}$$
$$I_4 = \overline{A_2 + A_1}$$

NOR array

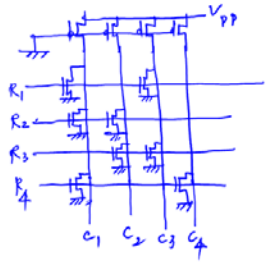
$$\bar{Z}_1 = A_0\,\overline{I_1 + I_2 + I_3} + \bar{A}_0 I_2$$
$$\bar{Z}_2 = A_0 \bar{I}_3 + \bar{A}_0 \bar{I}_4$$
$$\bar{Z}_3 = A_0 \bar{I}_1 + \bar{A}_0 \bar{I}_2$$

NOR array gated by $A_0, \bar{A}_0$

NOR decoder – NOR ROM Array

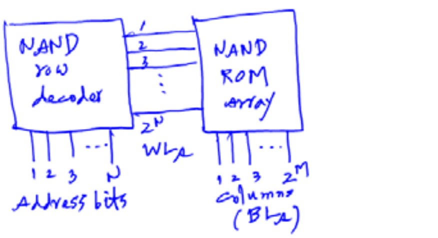


4 × 4 NOR-based ROM

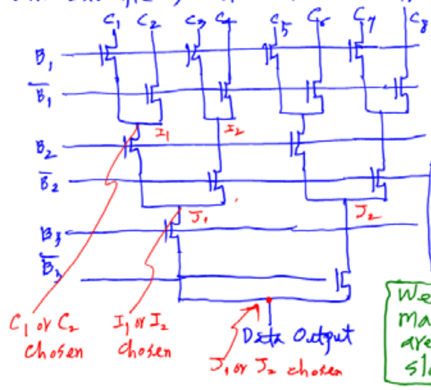


R_1	R_2	R_3	R_4	C_1	C_2	C_3	C_4
1	0	0	0	0	1	0	1
0	1	0	0	0	0	1	1
0	0	1	0	1	0	0	1
0	0	0	1	0	1	1	0

NAND-row decoder – NAND ROM Array

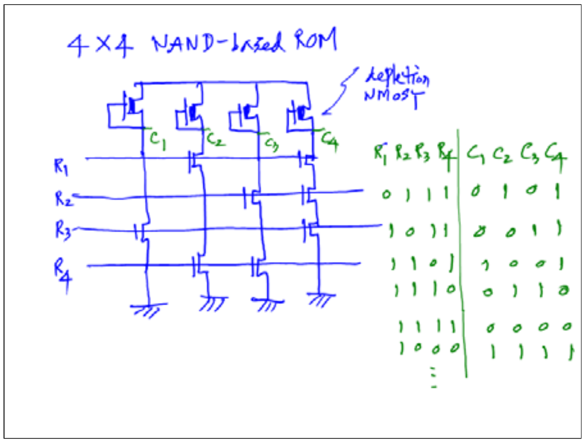


An Example of Column Decoder Circuit (Binary Tree Decoder)



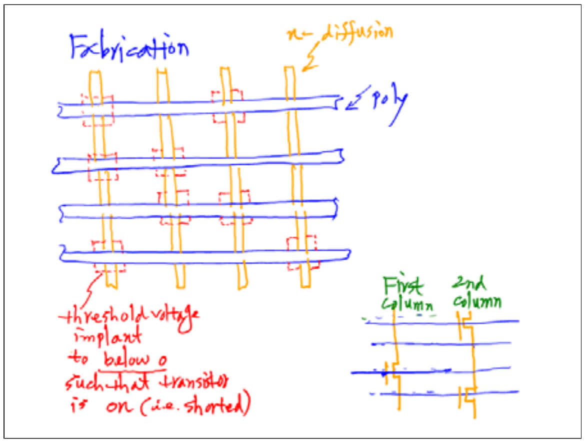
strength
compact,
less area
less power

Weakness
Many NMOSTs are in series
slow !



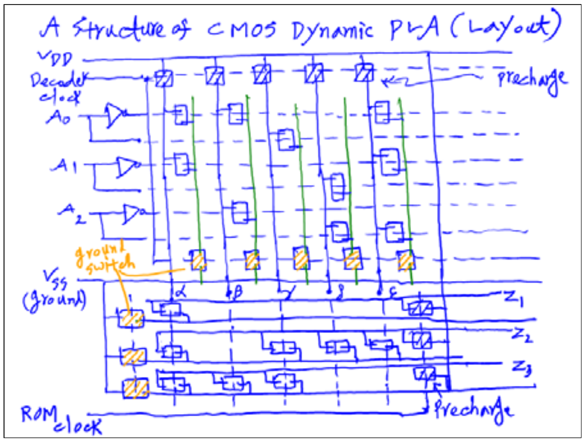
4×4 NAND-based ROM

R_1	R_2	R_3	R_4	C_1	C_2	C_3	C_4
0	1	1	1	0	1	0	1
1	0	1	1	0	0	1	1
1	1	0	1	1	0	0	1
1	1	1	0	0	1	1	0
1	1	1	1	0	0	0	0
1	0	0	0	1	1	1	1
⋮							

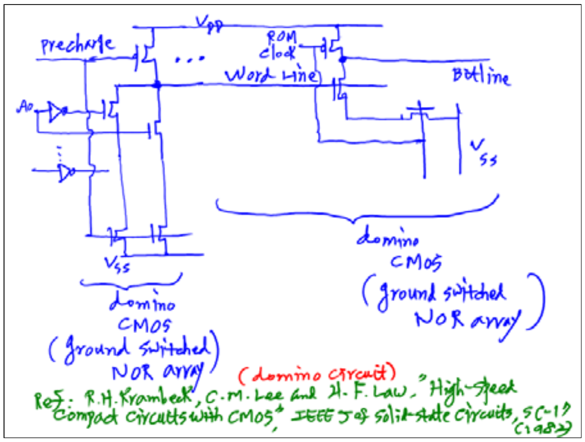


Fabrication

threshold-voltage implant to below 0 such that transistor is on (i.e. shorted)



A Structure of CMOS Dynamic PLA (Layout)

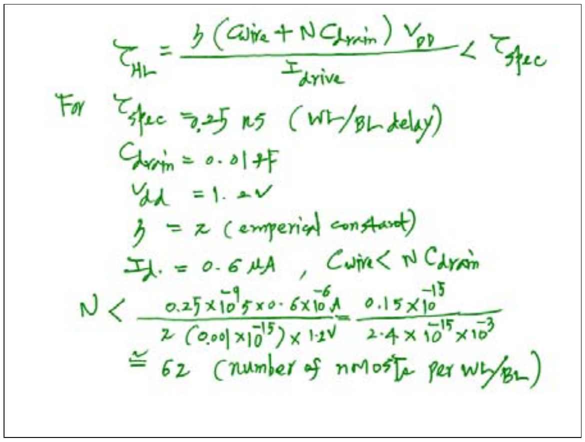


domino CMOS (ground switched NOR array)

domino CMOS (ground switched NOR array)

(domino circuit)

Ref: R.H. Krambeck, C.M. Lee and H.F. Law, "High-Speed Compact Circuits with CMOS", IEEE J of Solid-State Circuits, SC-17 (1982)



$$\tau_{HL} = \frac{\gamma\,(C_{wire} + N C_{drain})\,V_{DD}}{I_{drive}} < \tau_{spec}$$

For $\tau_{spec} = 0.25\ ns$ (WL/BL delay)

$C_{drain} = 0.01\ fF$

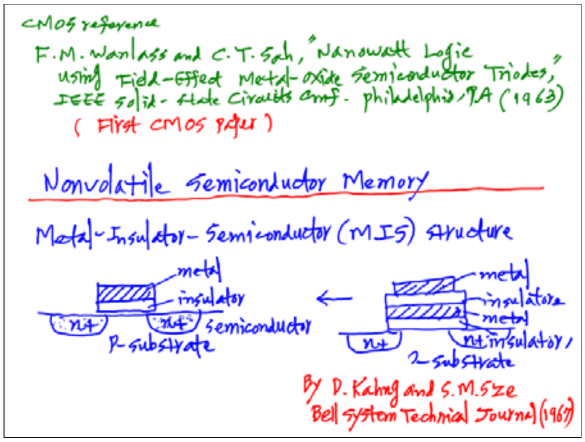
$V_{dd} = 1.2V$

$\gamma = 2$ (emperical constant)

$I_d = 0.6\ \mu A$, $C_{wire} < N C_{drain}$

$$N < \frac{0.25\times10^{-9}s\times0.6\times10^{-6}A}{2\,(0.001\times10^{-15})\times1.2V} = \frac{0.15\times10^{-15}}{2.4\times10^{-15}\times10^{-3}}$$

$\cong$ 62 (number of nMOSTs per WL/BL)



CMOS reference

F.M. Wanlass and C.T. Sah, "Nanowatt Logic using Field-Effect Metal-Oxide Semiconductor Triodes," IEEE Solid-State Circuits Conf. Philadelphia, PA (1963)

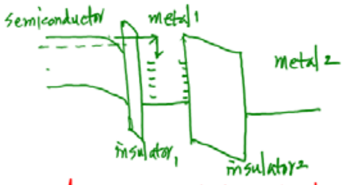
(First CMOS paper)

Nonvolatile Semiconductor Memory

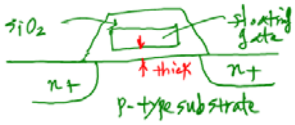
Metal-Insulator-Semiconductor (MIS) structure

By D. Kahng and S.M. Sze
Bell System Technical Journal (1967)

Operation Principle of metal-insulator-metal-insulator-semiconductor nonvolatile memory

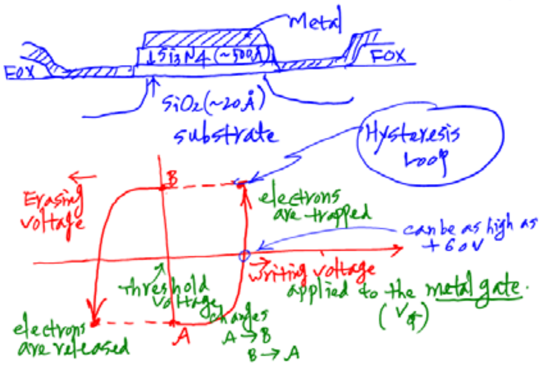


Electrons are injected into the floating gate (metal 1) by tunneling through insulator 1, and are stored semi-permanently.

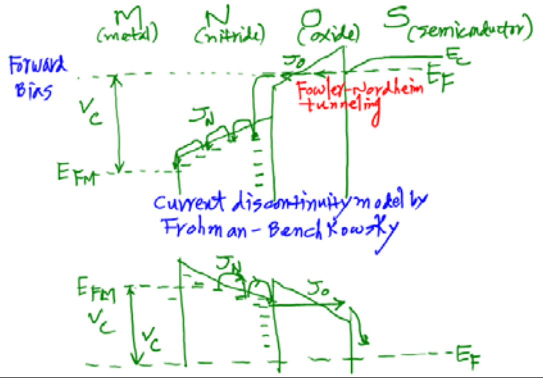


electrons are injected into the 'floating gate' by crossing the oxide (SiO_2) potential barrier and are stored in the floating gate.
Thus, named FAMOS (Floating-gate Alanche injection MOS)

MNOS Structure (Metal Nitride Oxide Semicon)



Energy Band Diagram



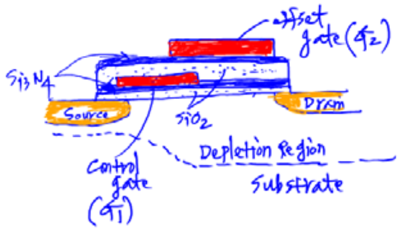
$$J_N = C_N E_N^2 \exp\left(-E_1/E_N\right)$$

$$J_{ox} = C_{ox} E_{ox}^2 \frac{\pi c kT/E_{ox}}{\sin(\pi kT/E_{ox})} \exp\left(-E_2/E_{ox}\right)$$

E_N = electric field in the nitride layer
E_{ox} = " oxide layer
k = Boltzman's constant
T = temp in [°K]
J_N = nitride layer current density
J_o (J_{ox}) = oxide layer " $\boxed{V_g = E_{ox} t_{ox} + E_N t_N}$

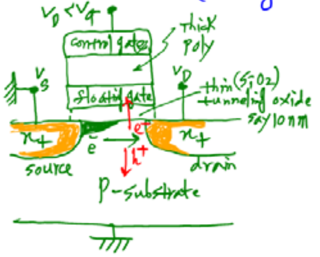
e.g. $C_{ox} = 10^{-5}\ A/V^2$ $C_N = 3.5 \times 10^{-10}\ A/V^2$
$E_2 = 2.54 \times 10^8\ V/cm$ $E_1 = 1.2 \times 10^8\ V/cm$
$t_{ox} = 50\ Å$ $t_N = 1000\ Å$

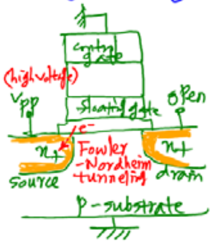
Stacked Gate Tetrode Proposed by H. G. Dill & T. N. Toombs (1969)

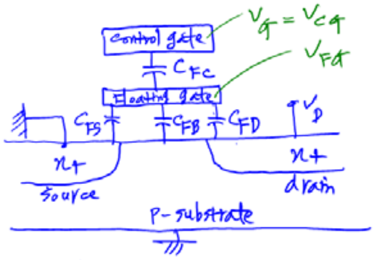


Flash Memory

Memory cell is a transistor with a floating gate whose threshold voltage can be programmed (changed) repeatedly by applying an electric field (through V_G voltage) to its gate.







$$C_{total} = C_{FC} + C_{FS} + C_{FB} + C_{FD}$$

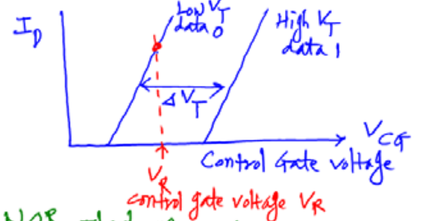
$$V_{FG} = \frac{Q_{FG}}{C_{FG}} + \frac{C_{FC}}{C_{total}} V_{CG} + \frac{C_{FD}}{C_{total}} V_D$$

Q_{FG} = charge stored in the floating gate

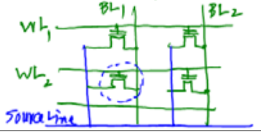
$$\underline{V_T(CG)} = \frac{C_{total}}{C_{FC}} \underline{V_T(FG)} - \frac{Q_{FG}}{C_{FC}} - \frac{C_{FD}}{C_{FC}} V_D$$

$$\Delta V_T(CG) = -\frac{\Delta Q_{FG}}{C_{FC}}$$

between "0" & "1".



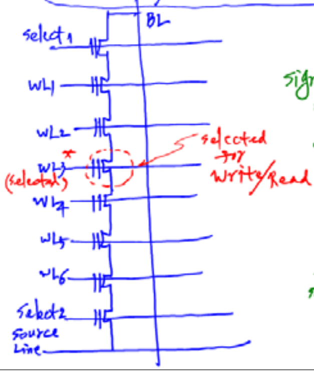
NOR Flash Memory



Signal	Erase	Program	Read
	operation		
BL_1	open	6V	1V
	open	0V	0V
Sourceline	12V	0V	0V
WL_1	0	0	0
WL_2	0	12V	5V

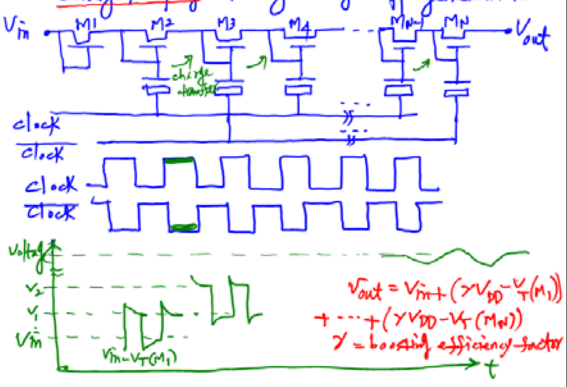
NOR array faster but many contacts (area ↑)

NAND array slower but compact (area ↓)



Signal	Erase	Program	Read
	operation		
BL1	open	0V	1V
WL1	0	10V	5V
2	0	10V	5V
3	0	20V	0V
4	0	10V	5V
5	0	10V	5V
6	0	10V	5V
sourceline	20V V_{pp}	0V	0V
select line2	open	0V	5V
p-well	20V	0V	0V
n-tub	20V	0V	0V

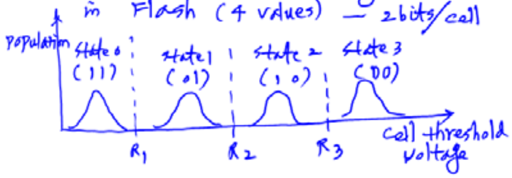
Charge Pumping for High Voltage V_{pp} generation



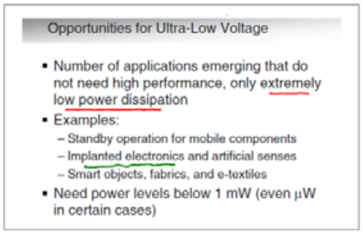
$$V_{out} = V_{in} + (\gamma V_{DD} - V_T(M_1)) + \cdots + (\gamma V_{DD} - V_T(M_N))$$

γ = boosting efficiency factor

Multi-level-cell Threshold Voltage Distribution in Flash (4 values) — 2bits/cell



Subthreshold operation

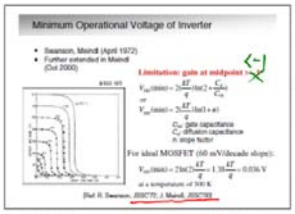


Opportunities for Ultra-Low Voltage

- Number of applications emerging that do not need high performance, only extremely low power dissipation
- Examples:
  - Standby operation for mobile components
  - Implanted electronics and artificial senses
  - Smart objects, fabrics, and e-textiles
- Need power levels below 1 mW (even µW in certain cases)

Slide 11.4
Although keeping the power density constant is one motivation for the continued search to lower the EOP, another, maybe even more important, reason is the exciting applications that only become feasible at very low energy/power levels. Consider, for instance, the digital wristwatch. The concept, though straightforward, only became attractive once the power dissipation

[R1]



Minimum Operational Voltage of Inverter

- Swanson, Meindl (April 1972)
- Further extended in Meindl (Oct 2000)

Limitation: gain at midpoint > −1

<−1

$V_{DD}(\min) = 2\frac{kT}{q}\ln\left(2+\frac{C_d}{C_{ox}}\right)$

or

$V_{DD}(\min) = 2\frac{kT}{q}\ln(1+n)$

C_{ox}: gate capacitance
C_d: diffusion capacitance
n: slope factor

For ideal MOSFET (60 mV/decade slope):

$V_{DD}(\min) = 2\ln(2)\frac{kT}{q} = 1.38\frac{kT}{q} = 0.036\,\text{V}$

at a temperature of 300 K

[Ref: R. Swanson, JSSC72; J. Meindl, JSSC00]

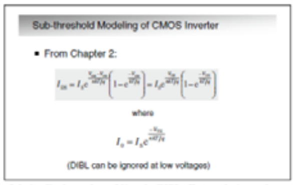
Slide 11.5
The question of the minimum operational voltage of a CMOS inverter was addressed in a landmark paper [Swanson72] in the early 1970s – published even before CMOS integrated circuits came in vogue! For an inverter to be regenerative and to have two distinct steady-state operation points (a "1" and a "0"), it is essential that the absolute value of the gain of the gate in the transient region be larger than 1. Solving for those conditions leads to an expression for V_{min} equal to $2(kT/q)\ln(1+n)$, where n is the slope factor of the transistors. One important observation is that V_{min} is proportional to the operational temperature T. Cooling down a CMOS circuit to temperatures close to absolute zero (e.g., liquid Helium), makes operation at mV levels possible. (Unfortunately, the energy going into the cooling more than often offsets the gains in operational energy.) Also, the closer the MOS transistor operating in sub-threshold mode gets to the ideal bipolar transistor behavior, the lower the minimum voltage. At room temperature, an ideal CMOS inverter (with a slope factor of 1) could marginally operate at as low as 36 mV!

gain < −1

|gain| > 1

$$I_{DS} = I_S e^{\frac{V_{GS}-V_{th}}{nKT/q}}\left(1 - e^{\frac{-V_{DS}}{KT/q}}\right) = I_0 e^{\frac{V_{GS}}{nKT/q}}\left(1 - e^{\frac{-V_{DS}}{KT/q}}\right)$$

where $I_0 = I_S e^{\frac{-V_{th}}{nKT/q}}$



Sub-threshold Modeling of CMOS Inverter

- From Chapter 2:

$$I_{DS} = I_S e^{\frac{V_{GS}-V_{Th}}{nkT/q}}\left(1-e^{-\frac{V_{DS}}{kT/q}}\right) = I_0 e^{\frac{V_{GS}}{nkT/q}}\left(1-e^{-\frac{V_{DS}}{kT/q}}\right)$$

where

$$I_0 = I_S e^{\frac{-V_{Th}}{nkT/q}}$$

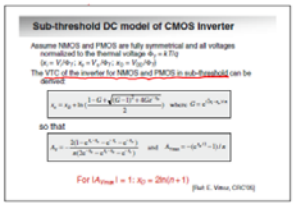
(DIBL can be ignored at low voltages)

Slide 11.6
Given the importance of this expression, a quick derivation is worth undertaking. We assume that at these low operational voltages, the transistors operate only in the sub-threshold regime, which is often also called the *weak-inversion* mode. The current–voltage relationship for a MOS transistor in sub-threshold mode was presented in Chapter 2, and is repeated here for the sake of clarity. For low values of V_{DS}, the DIBL effect can be ignored.

$x_i = V_i/\phi_T$ $x_o = V_o/\phi_T$ $x_D = V_{DD}/\phi_T$

thermal voltage

$$x_o = x_D + \ln\left(\frac{1-G+\sqrt{(G-1)^2+4Ge^{-x_D}}}{2}\right),\quad G = e^{\frac{2x_i - x_D}{n}}$$



Sub-threshold DC model of CMOS Inverter

Assume NMOS and PMOS are fully symmetrical and all voltages normalized to the thermal voltage $\Phi_T = kT/q$
($x_i = V_i/\Phi_T$; $x_o = V_o/\Phi_T$; $x_D = V_{DD}/\Phi_T$)

The VTC of the inverter for NMOS and PMOS in sub-threshold can be derived:

$$x_o = x_D + \ln\left(\frac{1-G+\sqrt{(G-1)^2+4Ge^{-x_D}}}{2}\right) \quad \text{where } G = e^{(2x_i - x_D)/n}$$

so that

$$A_V = -\frac{2(1-e^{x_o-x_D}-e^{-x_o}-e^{-x_D})}{n(2e^{-x_D}-e^{x_o-x_D}-e^{-x_o})} \quad \text{and} \quad A_{Vmax} = -(e^{x_D/2}-1)/n$$

For $|A_{Vmax}| = 1$: $x_D = 2\ln(n+1)$

[Ref: E. Vittoz, CRC'05]

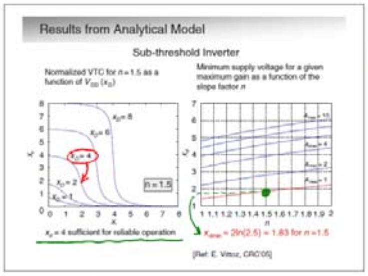
Slide 11.7
The (static) voltage transfer characteristic (VTC) of the inverter is derived by equating the current through the NMOS and PMOS transistors. The derivation is substantially simplified if we assume that two devices have exactly the same strength when operating in sub-threshold. Also, normalizing all voltages with respect to the thermal voltage Φ_T leads to more elegant expressions. Setting the gain to −1 yields the same expression for the minimum voltage as was derived by Swanson.

$K_p = K_n$

$K_x = \mu_x\left(\frac{W}{L}\right)_x \times C_{ox}$

$$A_V = -\frac{2(1-e^{(x_o-x_D)}-e^{-x_o}-e^{-x_D})}{n(2e^{-x_D}-e^{(x_o-x_D)}-e^{-x_o})}$$

$|A_{Vmax}| = 1,\ x_D = 2\ln(n+1)$



Results from Analytical Model

Sub-threshold Inverter

x_o = 4 sufficient for reliable operation

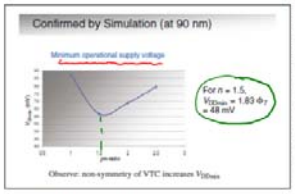
x_{Dmin} = 2ln(2.5) = 1.83 for n = 1.5

[Ref: E. Vittoz, CRC'05]

thermal voltage leads to reasonable noise margins (assuming $n = 1.5$). This is approximately equal to 100 mV.

$x_D = V_{DD}/\phi_T$ $\phi_T = \frac{kT}{q}$ (thermal voltage)

$x_D = 4 \simeq 100\text{mV}$ $= 26\text{mV}$ at T = 300°K (room temp.)



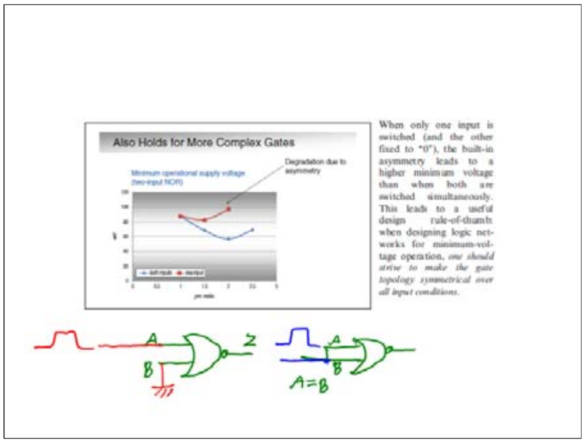
Confirmed by Simulation (at 90 nm)

For n = 1.5, V_{DDmin} = 1.83 Φ_T = 48 mV

Observe: non-symmetry of VTC increases V_{DDmin}

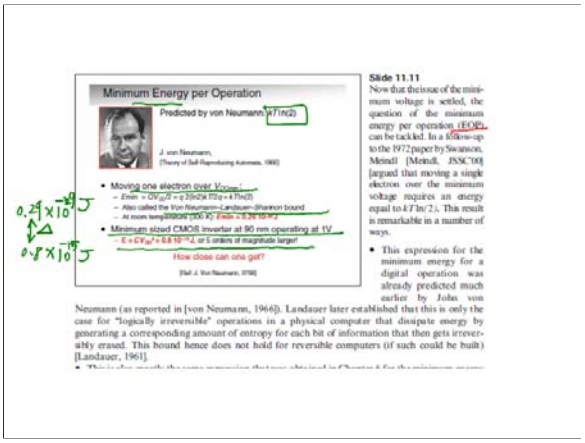
Slide 11.9
Simulations (for a 90 nm technology) confirm these results. When plotting the minimum supply voltage as a function of the PMOS/NMOS ratio, a minimum can be observed when the inverter is completely symmetrical, that is when the PMOS and NMOS transistors have identical drive strengths. Any deviation from the symmetry causes V_{min} to rise. This implies that transistor sizing will play a role in the design of minimum-voltage circuits.

Also worth noticing is that the simulated minimum voltage of 60 mV is slightly higher than the theoretical value of 48 mV. This is mostly owing to the definition of "operational" point. At 48 mV, the inverter is only marginally functional. In the simulation, we assume a small margin of approximately 25%.


Minimum operational supply voltage
(two-input NOR)
Degradation due to asymmetry
both inputs
one input
pn ratio
A
B
Z
A
B
A=B

Also Holds for More Complex Gates

When only one input is switched (and the other fixed to "0"), the built-in asymmetry leads to a higher minimum voltage than when both are switched simultaneously. This leads to a useful design rule-of-thumb when designing logic networks for minimum-voltage operation, *one should strive to make the gate topology symmetrical over all input conditions.*


0.29×10^{-20} J
Δ
0.8×10^{-15} J

Minimum Energy per Operation

Predicted by von Neumann: $kT\ln(2)$

J. von Neumann,
[Theory of Self-Reproducing Automata, 1966]

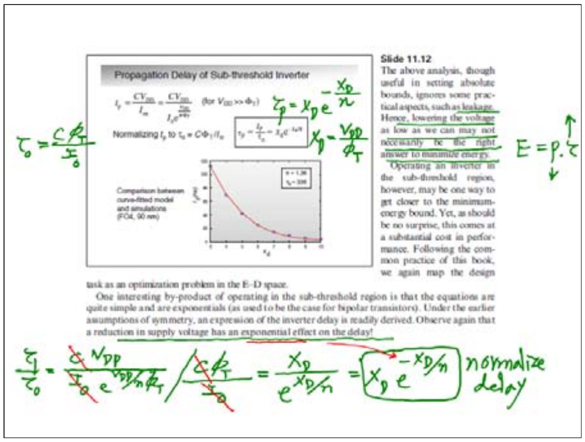
- Moving one electron over $V_{DD(min)}$:
  - $E_{min} = QV_{DD}/2 = q\,2(\ln 2)kT/2q = kT\ln(2)$
  - Also called the Von Neumann–Landauer–Shannon bound
  - At room temperature (300 K): $E_{min} = 0.29\ 10^{-20}$ J
- Minimum sized CMOS inverter at 90 nm operating at 1V
  - $E = CV_{DD}^2 = 0.8\ 10^{-15}$ J, or 5 orders of magnitude larger!

How close can one get?

[Ref: J. Von Neumann, 1966]

Slide 11.11
Now that the issue of the minimum voltage is settled, the question of the minimum energy per operation (EOP) can be tackled. In a follow-up to the 1972 paper by Swanson, Meindl [Meindl, JSSC'00] [argued that moving a single electron over the minimum voltage requires an energy equal to $kT\ln(2)$. This result is remarkable in a number of ways.

- This expression for the minimum energy for a digital operation was already predicted much earlier by John von Neumann (as reported in [von Neumann, 1966]). Landauer later established that this is only the case for "logically irreversible" operations in a physical computer that dissipate energy by generating a corresponding amount of entropy for each bit of information that then gets irreversibly erased. This bound hence does not hold for reversible computers (if such could be built) [Landauer, 1961].


Comparison between curve-fitted model and simulations (FO4, 90 nm)
x_d
$\tau_0 = \frac{C\phi_T}{I_0}$
$\tau_p = x_D e^{-\frac{x_D}{n}}$
$x_D = \frac{V_{DD}}{\phi_T}$
$E = P\cdot\tau$
↑
↓

Propagation Delay of Sub-threshold Inverter

$$t_p = \frac{CV_{DD}}{I_{on}} = \frac{CV_{DD}}{I_0 e^{\frac{V_{DD}}{n\Phi_T}}} \quad (\text{for } V_{DD} >> \Phi_T)$$

Normalizing t_p to $\tau_0 = C\Phi_T/I_0$: $\tau_p = \frac{t_p}{\tau_0} = x_d e^{-x_d/n}$

Slide 11.12
The above analysis, though useful in setting absolute bounds, ignores some practical aspects, such as leakage. Hence, lowering the voltage as low as we can may not necessarily be the right answer to minimize energy.

Operating an inverter in the sub-threshold region, however, may be one way to get closer to the minimum-energy bound. Yet, as should be no surprise, this comes at a substantial cost in performance. Following the common practice of this book, we again map the design task as an optimization problem in the E–D space.

One interesting by-product of operating in the sub-threshold region is that the equations are quite simple and are exponentials (as used to be the case for bipolar transistors). Under the earlier assumptions of symmetry, an expression of the inverter delay is readily derived. Observe again that a reduction in supply voltage has an exponential effect on the delay!

$$\frac{\tau_p}{\tau_0} = \frac{\cancel{C} V_{DP}}{\cancel{I_0}\, e^{V_{DD}/n\phi_T}} \Big/ \frac{\cancel{C}\phi_T}{\cancel{I_0}} = \frac{x_D}{e^{x_D/n}} = \boxed{x_D e^{-x_D/n}}$$ normalize delay

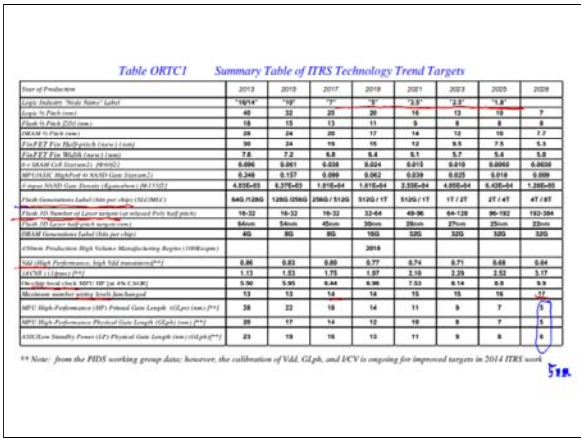
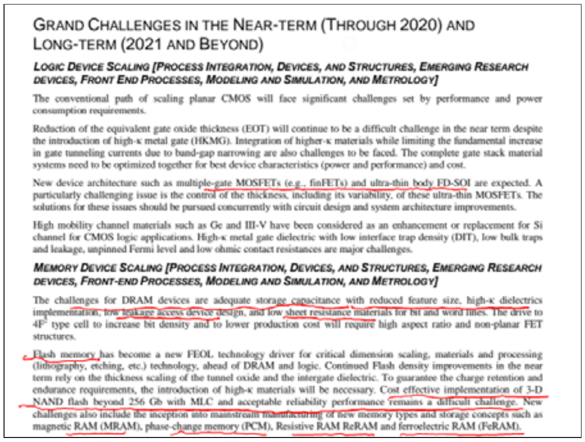


Table ORTC1 Summary Table of ITRS Technology Trend Targets

Year of Production	2013	2015	2017	2019	2021	2023	2025	2028
Logic Industry "Node Name" Label	"16/14"	"10"	"7"	"5"	"3.5"	"2.5"	"1.8"	
Logic ½ Pitch (nm)	40	32	25	20	16	13	10	7
Flash ½ Pitch [2D] (nm)	18	15	13	11	9	8	8	8
DRAM ½ Pitch (nm)	28	24	20	17	14	12	10	7.7
FinFET Fin Half-pitch (new) (nm)	30	24	19	15	12	9.5	7.5	5.3
FinFET Fin Width (new) (nm)	7.6	7.2	6.8	6.4	6.1	5.7	5.4	5.0
6-t SRAM Cell Size(um2) [@60f2]	0.096	0.061	0.038	0.024	0.015	0.010	0.0060	0.0030
MPU/ASIC HighPerf 4t NAND Gate Size(um2)	0.248	0.157	0.099	0.062	0.039	0.025	0.018	0.009
4-input NAND Gate Density (Kgates/mm2) [@155f2]	4.03E+03	6.37E+03	1.01E+04	1.61E+04	2.55E+04	4.05E+04	6.42E+04	1.28E+05
Flash Generations Label (bits per chip) (SLC/MLC)	64G /128G	128G /256G	256G / 512G	512G / 1T	512G / 1T	1T / 2T	2T / 4T	4T / 8T
Flash 3D Number of Layer targets (at relaxed Poly half pitch)	16-32	16-32	16-32	32-64	48-96	64-128	96-192	192-384
Flash 3D Layer half-pitch targets (nm)	64nm	54nm	45nm	30nm	28nm	27nm	25nm	22nm
DRAM Generations Label (bits per chip)	4G	8G	8G	16G	32G	32G	32G	32G
450mm Production High Volume Manufacturing Begins (100Kwspm)				2018				
Vdd (High Performance, high Vdd transistors)[**]	0.86	0.83	0.80	0.77	0.74	0.71	0.68	0.64
1/(CV/I) (1/psec) [**]	1.13	1.53	1.75	1.97	2.10	2.29	2.52	3.17
On-chip local clock MPU HP [at 4% CAGR]	5.50	5.95	6.44	6.96	7.53	8.14	8.8	9.9
Maximum number wiring levels [unchanged]	13	13	14	14	15	15	16	17
MPU High-Performance (HP) Printed Gate Length (GLpr) (nm) [**]	28	22	18	14	11	9	7	5
MPU High-Performance Physical Gate Length (GLph) (nm) [**]	20	17	14	12	10	8	7	5
ASIC/Low Standby Power (LP) Physical Gate Length (nm) (GLph)[**]	23	19	16	13	11	9	8	6

*** Note: from the PIDS working group data; however, the calibration of Vdd, GLph, and I/CV is ongoing for improved targets in 2014 ITRS work*

5nm



Grand Challenges in the Near-term (Through 2020) and Long-term (2021 and Beyond)

Logic Device Scaling [Process Integration, Devices, and Structures, Emerging Research devices, Front End Processes, Modeling and Simulation, and Metrology]

The conventional path of scaling planar CMOS will face significant challenges set by performance and power consumption requirements.

Reduction of the equivalent gate oxide thickness (EOT) will continue to be a difficult challenge in the near term despite the introduction of high-κ metal gate (HKMG). Integration of higher-κ materials while limiting the fundamental increase in gate tunneling currents due to band-gap narrowing are also challenges to be faced. The complete gate stack material systems need to be optimized together for best device characteristics (power and performance) and cost.

New device architecture such as multiple-gate MOSFETs (e.g., finFETs) and ultra-thin body FD-SOI are expected. A particularly challenging issue is the control of the thickness, including its variability, of these ultra-thin MOSFETs. The solutions for these issues should be pursued concurrently with circuit design and system architecture improvements.

High mobility channel materials such as Ge and III-V have been considered as an enhancement or replacement for Si channel for CMOS logic applications. High-κ metal gate dielectric with low interface trap density (DIT), low bulk traps and leakage, unpinned Fermi level and low ohmic contact resistances are major challenges.

Memory Device Scaling [Process Integration, Devices, and Structures, Emerging Research devices, Front-end Processes, Modeling and Simulation, and Metrology]

The challenges for DRAM devices are adequate storage capacitance with reduced feature size, high-κ dielectrics implementation, low leakage access device design, and low sheet resistance materials for bit and word lines. The drive to $4F^2$ type cell to increase bit density and to lower production cost will require high aspect ratio and non-planar FET structures.

Flash memory has become a new FEOL technology driver for critical dimension scaling, materials and processing (lithography, etching, etc.) technology, ahead of DRAM and logic. Continued Flash density improvements in the near term rely on the thickness scaling of the tunnel oxide and the intergate dielectric. To guarantee the charge retention and endurance requirements, the introduction of high-κ materials will be necessary. Cost effective implementation of 3-D NAND flash beyond 256 Gb with MLC and acceptable reliability performance remains a difficult challenge. New challenges also include the inception into mainstream manufacturing of new memory types and storage concepts such as magnetic RAM (MRAM), phase-change memory (PCM), Resistive RAM ReRAM and ferroelectric RAM (FeRAM).