Switching Energy in CMOS Logic: How far are we from physical limit?

Saibal Mukhopadhyay

Arijit Raychowdhury

Professor: Kaushik Roy

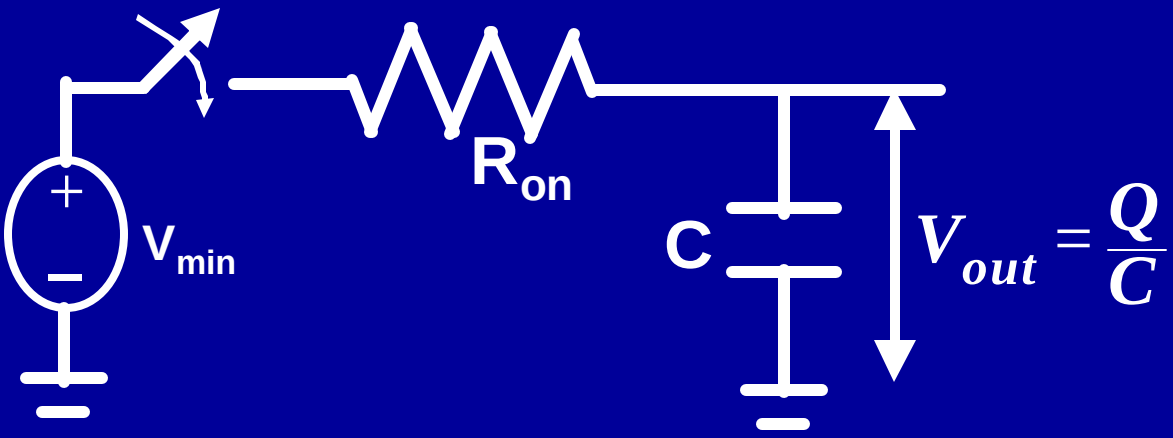
Dept. of Electrical & Computer Engineering

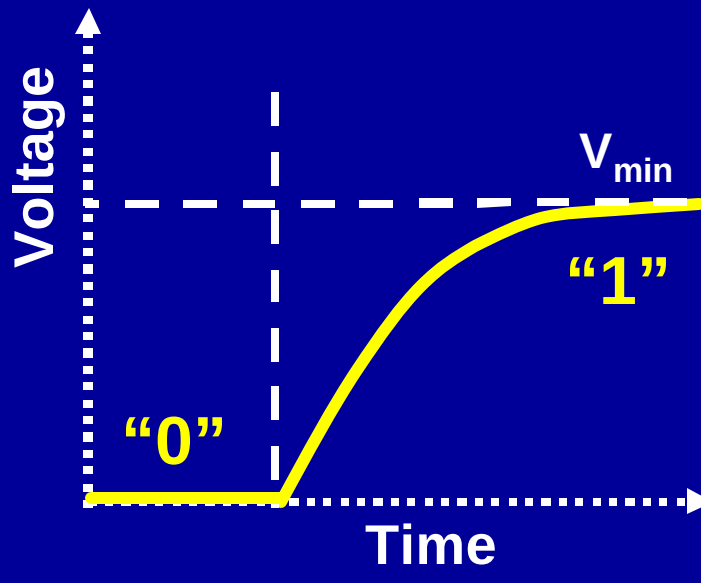
Purdue University

Outline

- **Switching energy in charge transfer based Digital Logic**
  - Basics and Physical Limits
- **Practical consideration for switching energy in CMOS Logic**
  - **Static requirements**
  - **Dynamic requirements**
  - **System considerations**
- **What can we do to reduce switching energy ?**
- **Summary**

Charge Based Digital Logic

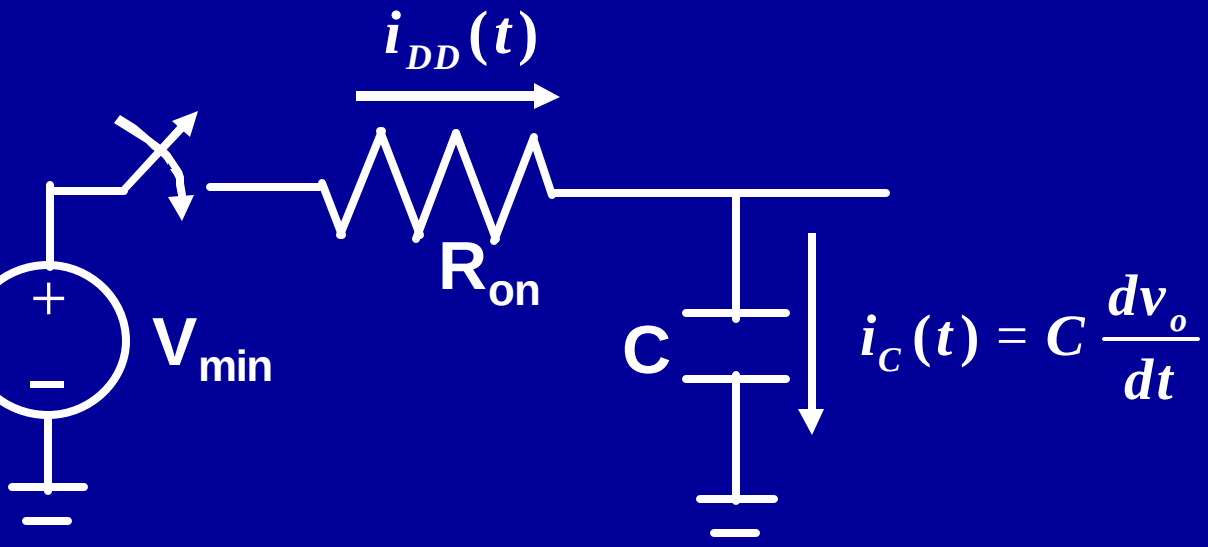


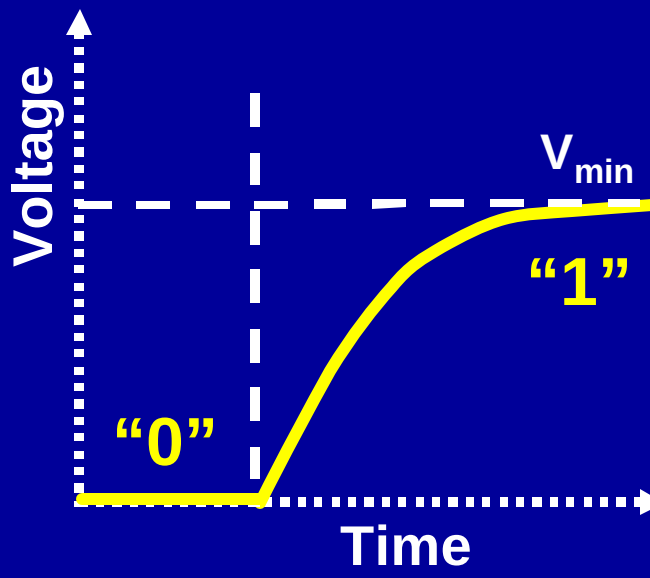


Key principles in the charge based digital logic

1. Representation of digital states
   Logic "0": No Charge in the capacitor
   Logic "1": Charge stored in the capacitor
2. Change of digital state
   Charge/dis-charge capacitor through a resistor

Switching Energy





$$E_{Total} = \int_0^{\infty} i_{DD}(t) V_{min}\, dt = \int_0^{V_{DD}} C V_{min}\, dv_0 = C V_{min}^2$$

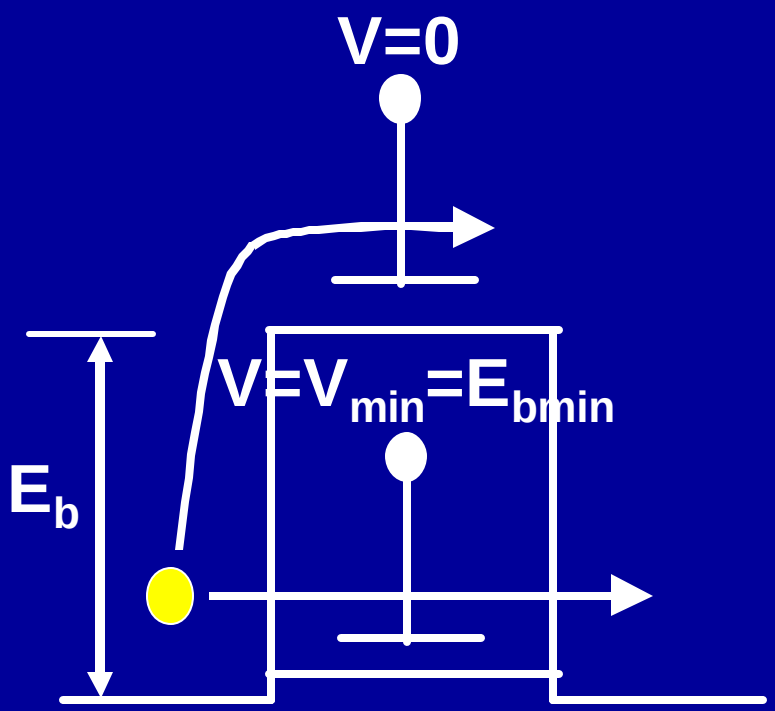
$$E_{Cap} = \int_0^{\infty} i_C(t) v_0(t)\, dt = \int_0^{V_{DD}} C v_0\, dv_0 = \frac{1}{2} C V_{min}^2$$

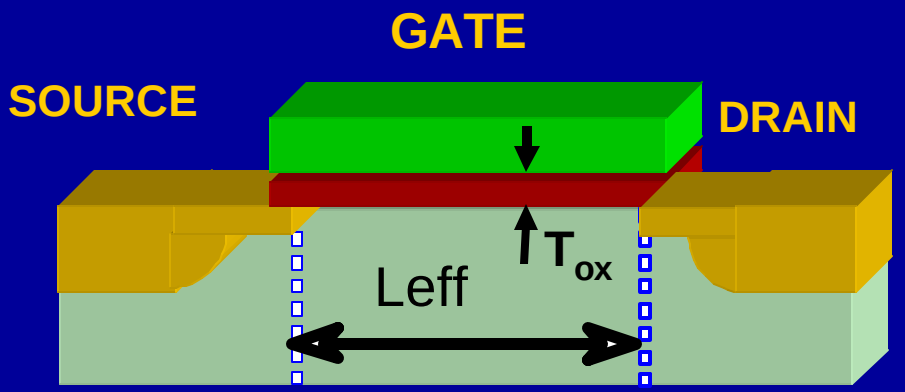
$$\therefore E_{diss}(0 \rightarrow 1) = E_{Total} - E_{Cap} = \frac{1}{2} C V_{min}^2$$

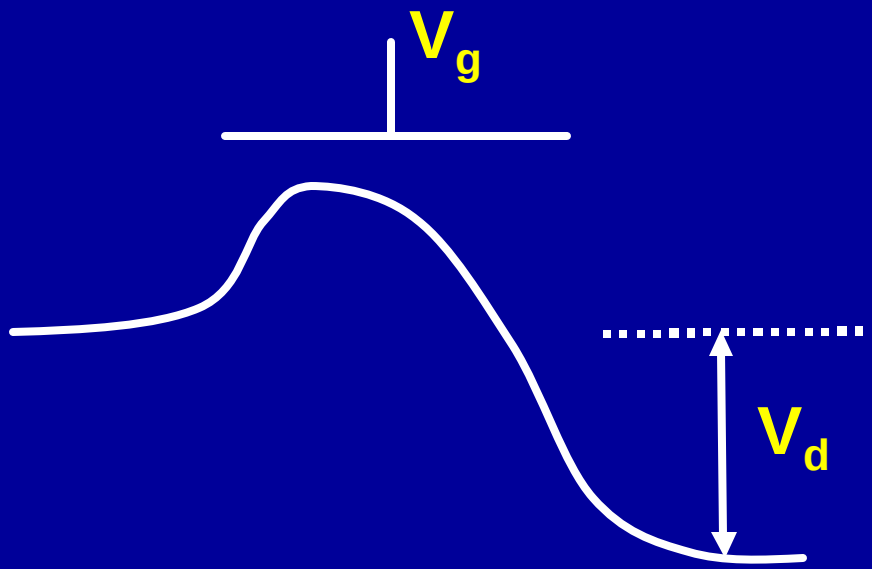
$$E_{diss} = C V_{min}^2 = Q V_{min}$$

Switching energy can be minimized by reducing Q and/or V_{min}

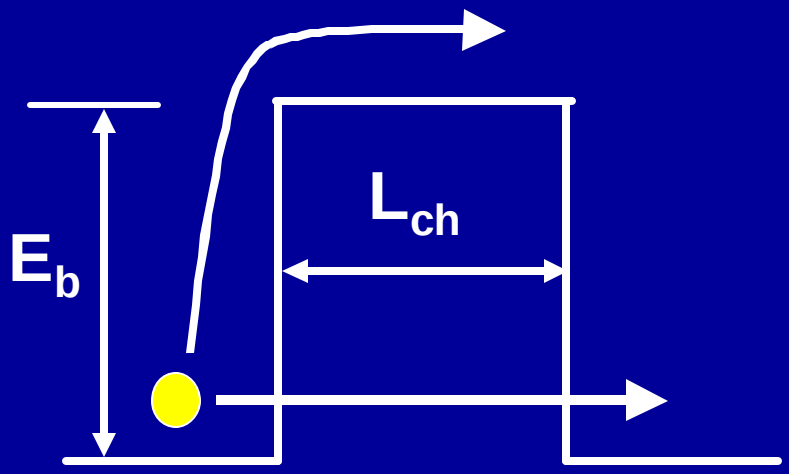
Physical Medium for Computation: Barrier Model







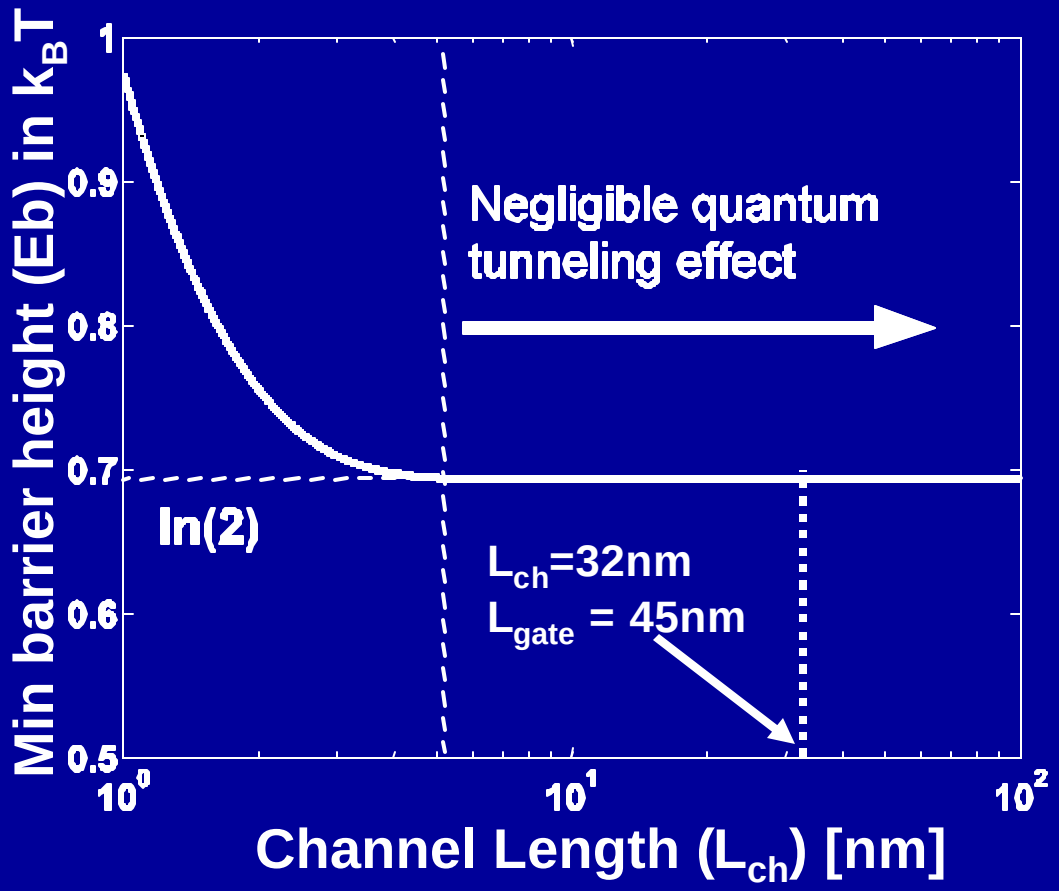
Minimum Barrier Height: Zhirnov's Model



$$P_{err} = P_{err\_cl} + P_{err\_QM} - P_{err\_cl}P_{err\_QM}$$

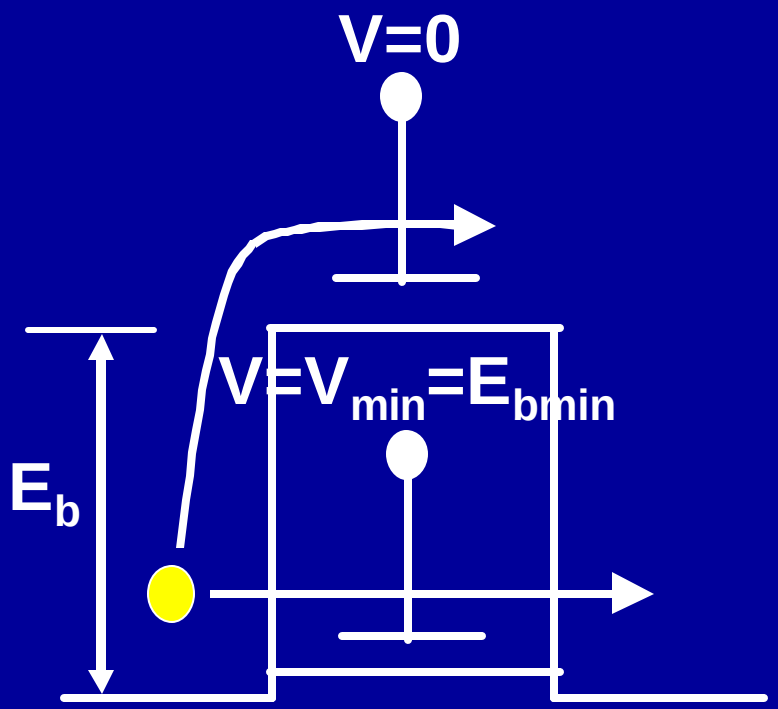
$$For \;\; L_{ch} > 10nm$$

$$P_{err} \sim \exp\left(-E_b / k_B T\right) \Rightarrow E_b = k_B T \ln\left(1/P_{err}\right)$$



Minimum barrier height = $E_{bmin} \sim k_B T \ln(2)$

Minimum Operating Voltage and Switching Energy



- **Minimum operating voltage**

  $V_{min} \sim k_B T \ln(2)$

- **Minimum switching energy**

  $E_{diss} = Q_{min} V_{min} = q k_B T \ln(2) \sim 0.7 k_B T$

Switching energy for an minimum sized inverter designed using in 45nm gate length devices ~ $35000 k_B T$

Why are we so far from the limit?

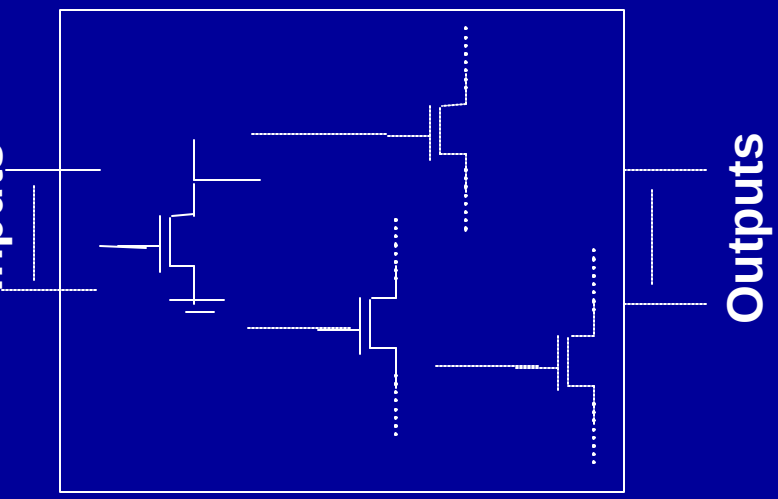
1. Can we operate with $V_{min} \sim K_B T \ln 2$?

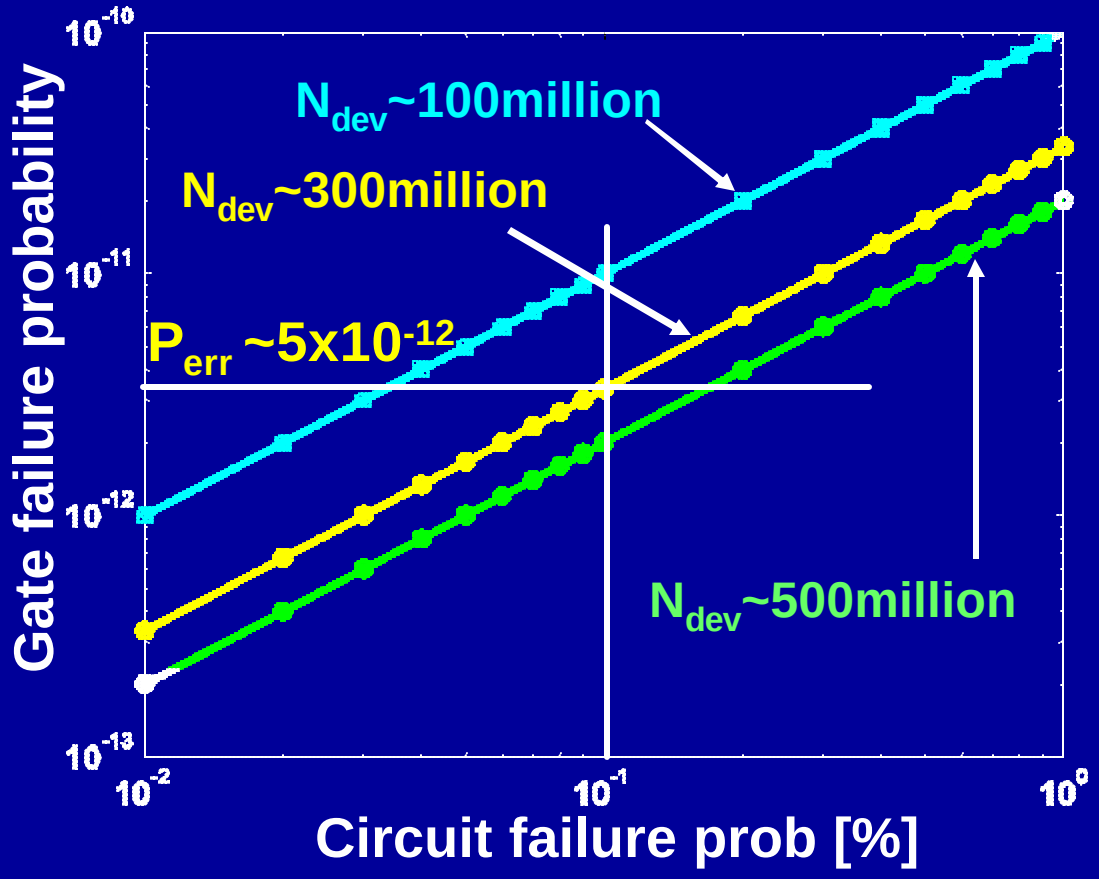
2. Can we operate with $Q_{min} = q$?

Outline

- **Switching energy in charge transfer based Digital Logic**
  - Basics and Physical Limits
- **Practical consideration for switching energy in CMOS Logic**
  - **Static requirements**
  - **Dynamic requirements**
  - **Circuit/System considerations**
- **What can we do to reduce switching energy ?**
- **Summary**

Reliability of Circuit Operation





# of devices = N_{dev}

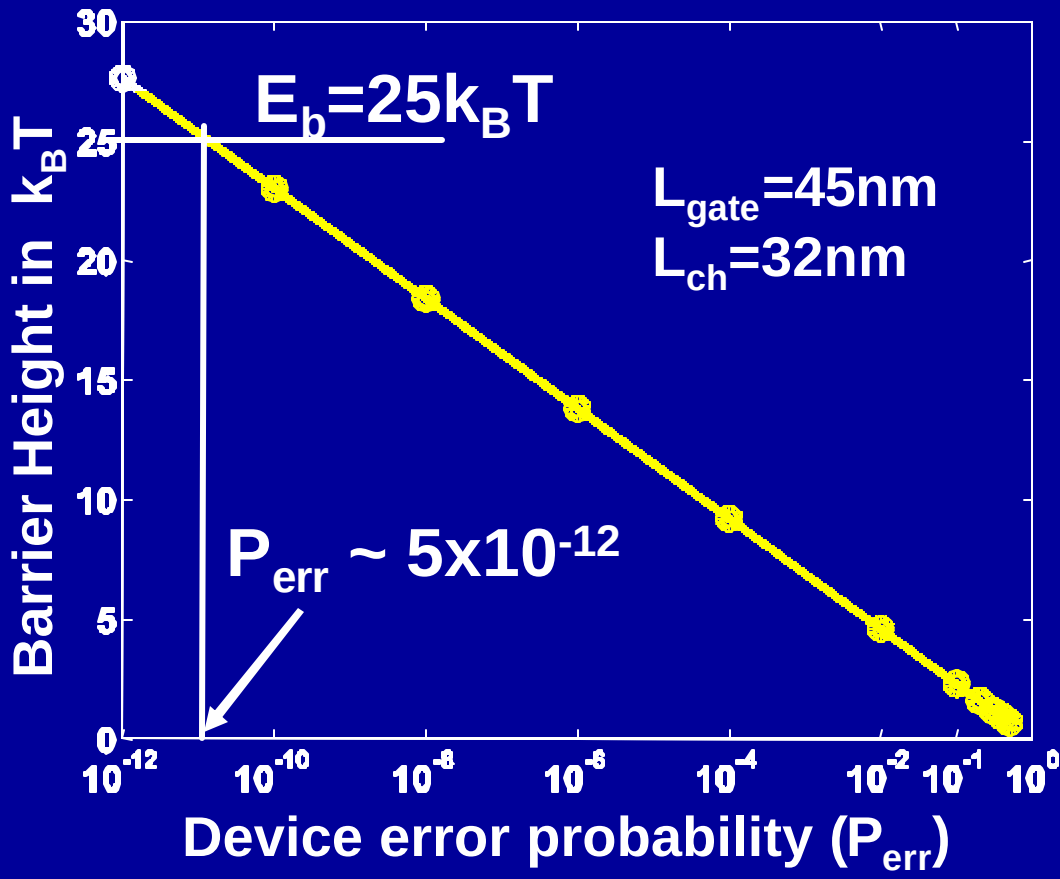
Prob. of error of a single gate = P_{err}

Prob. of error of the circuit = $P_{circ} = 1 - (1-P_{err})^{Ndev}$

Reliable operation of the circuit imposes stronger constraint on the reliability of the gate operation

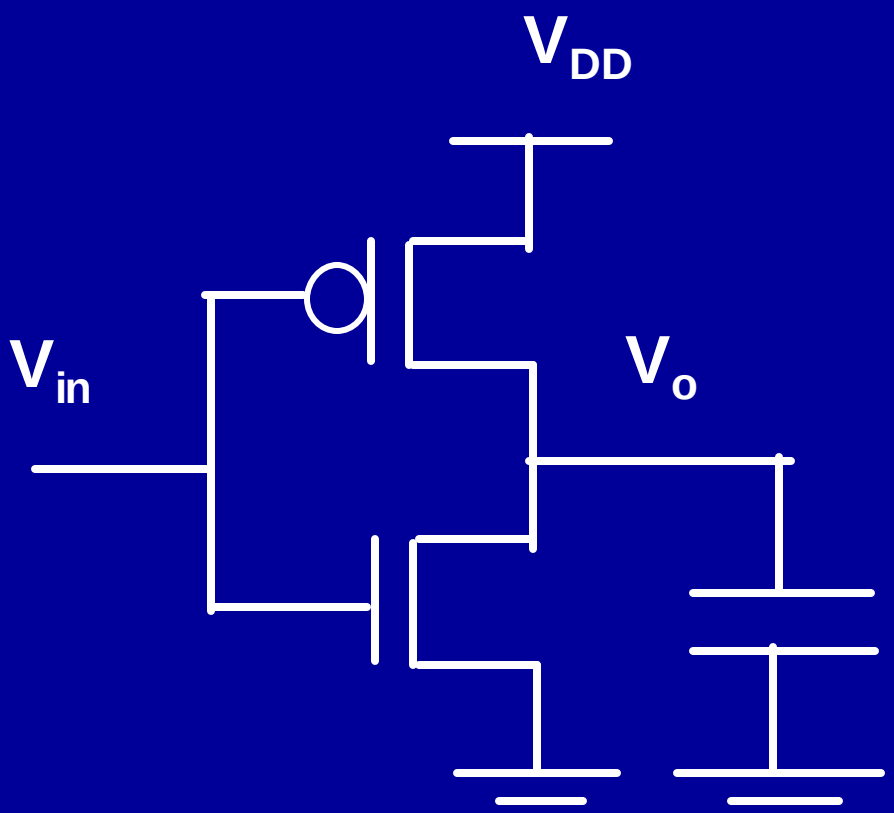
Reliable Operation for a Device

- Reliable operation requires a higher barrier
  - P_{err} = 0.5
    => E_b= 0.7k_BT
  - P_{err} = 5x10^{-12}
    => E_b= 25k_BT
- 0.1% failure rate for a circuit of 300 million devices => V_{min}~25k_BT



$k_BTln(2)$ → Reliability → $25k_BT$

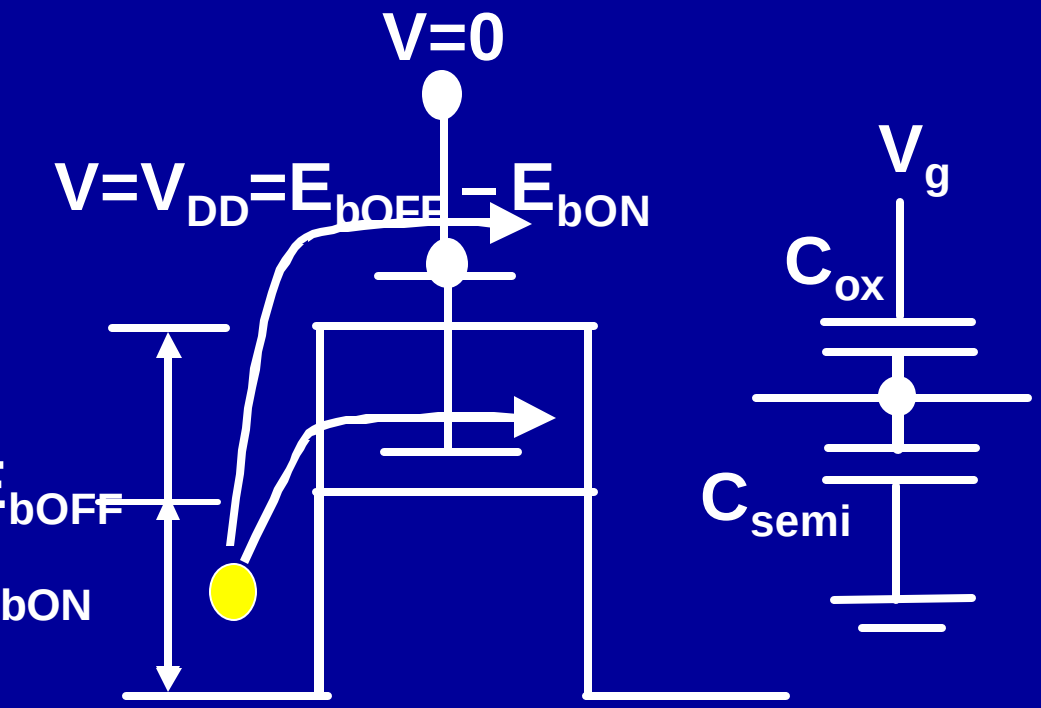
CMOS Logic: Physical Model

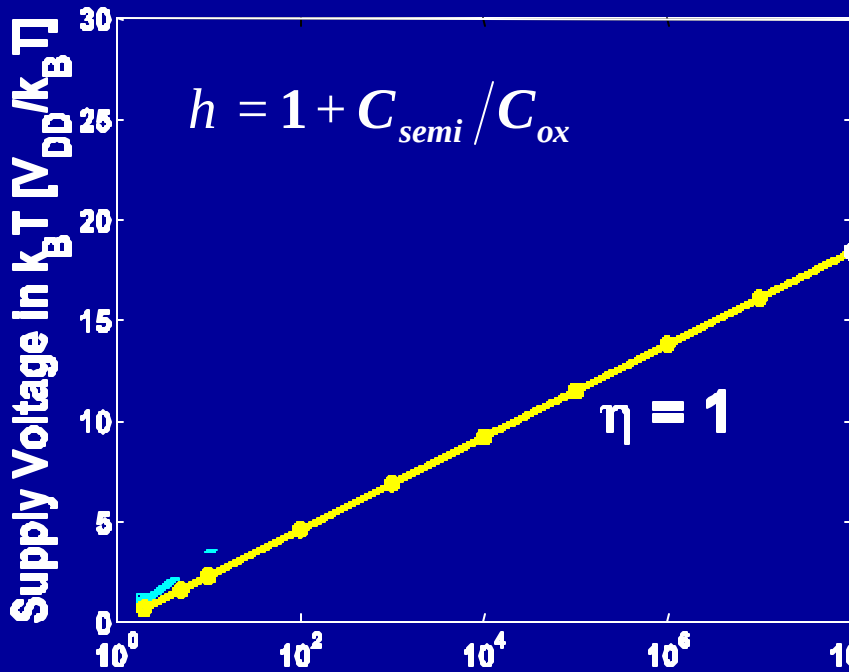


Vin	Vout
“0” ◦ 0V	“1” ◦ V_{DD}
“1” ◦ V_{DD}	“0” ◦ 0V

CMOS logic operates based on presence or absence of charge and not on localization of charge

Operation of MOS Device

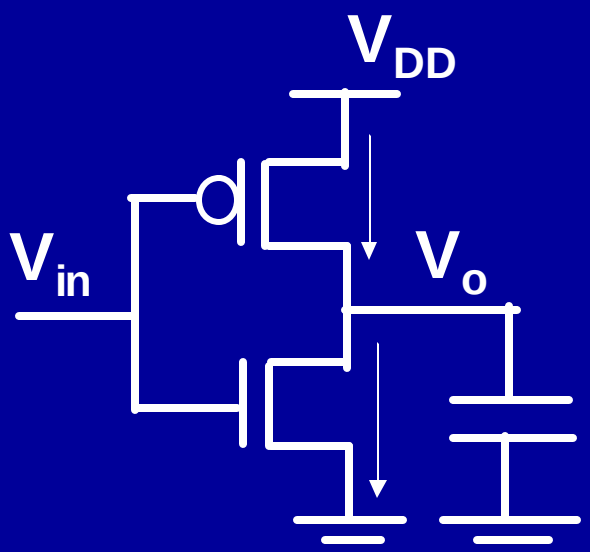


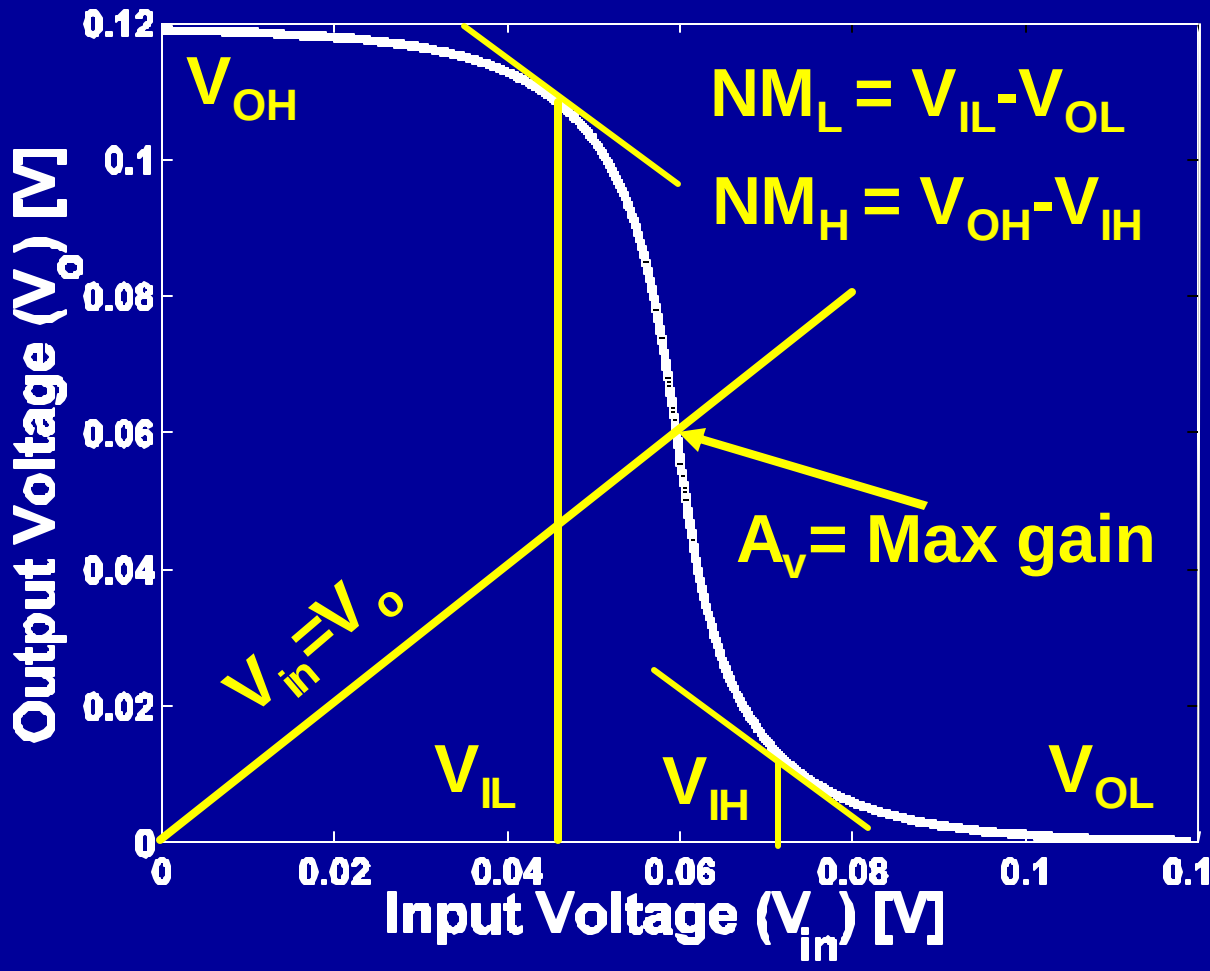


$$V_{DD} = h\frac{k_B T}{q} ln\left(\frac{p_{on}}{p_{off}}\right)$$

Operation with a larger p_{on}/p_{off} requires a higher supply voltage

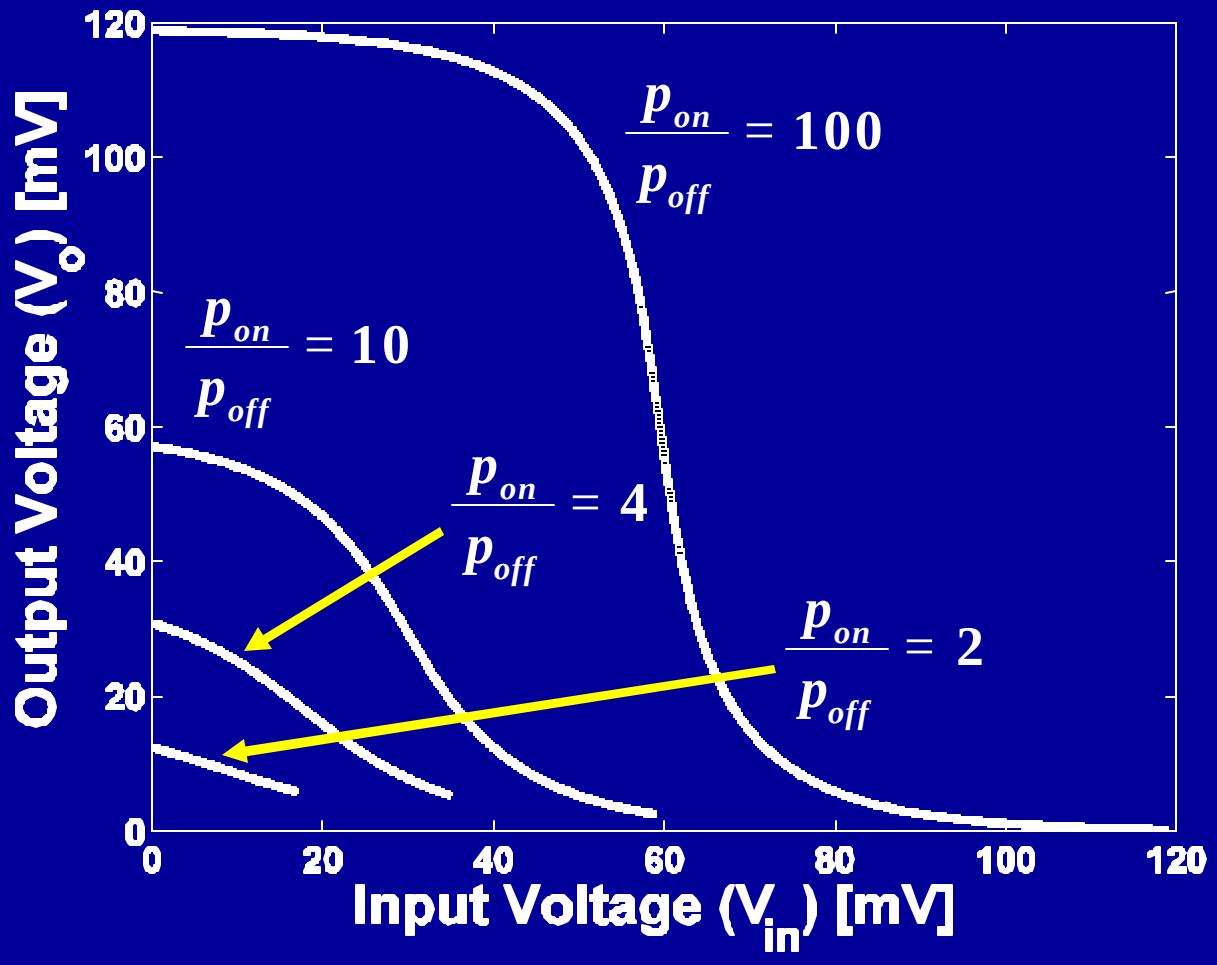
Operation of CMOS Logic





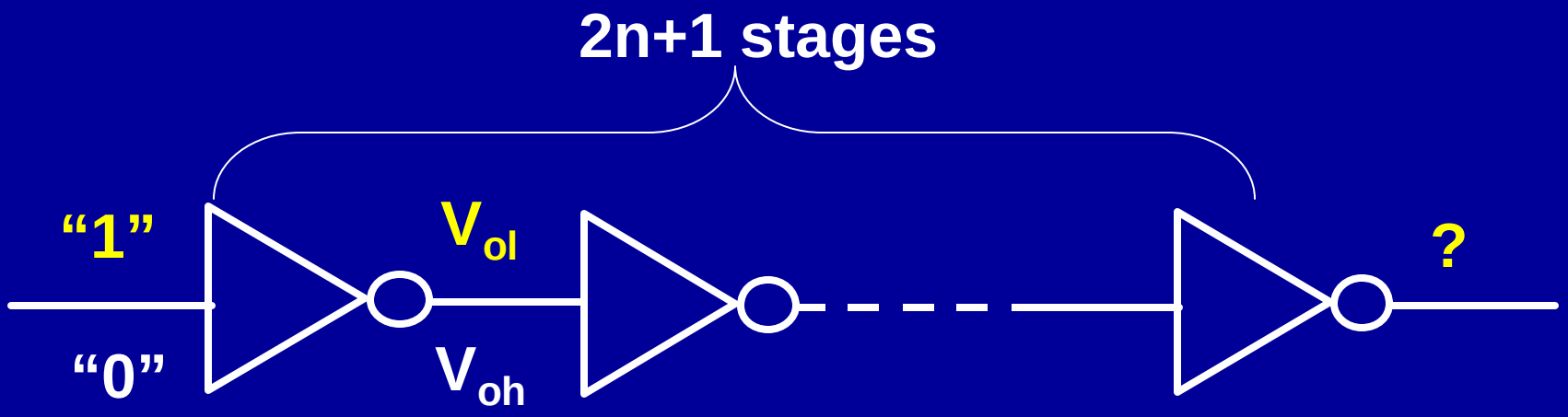
$$V_{in} = 0.5h\frac{k_B T}{q} \times \left(ln\left(\frac{p_{on}}{p_{off}} \right) + ln\left(\frac{1 - exp\left(-q\left(V_{DD} - V_0 \right)/k_B T \right)}{1 - exp\left(-qV_0/k_B T \right)} \right) \right)$$

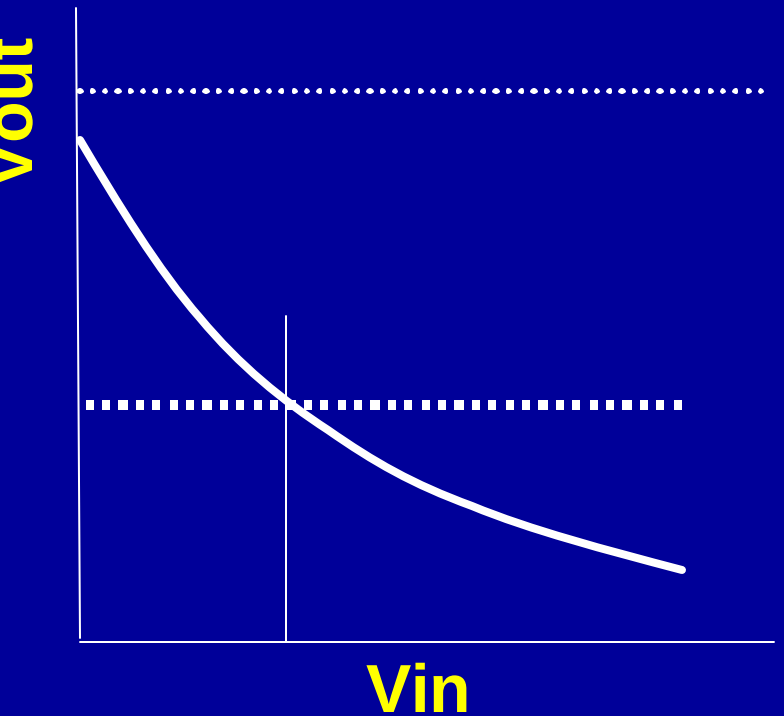
Operation of CMOS Logic



Higher p_{on}/p_{off} improves maximum gain and noise margin

Operation of CMOS Logic





$$Vin = V_{DD}/2 - D$$

$$Vo(1) = Vin(2) = V_{DD}/2 + DA_v$$

M

$$Vo(2n+1) = V_{DD}/2 - D(-1)^{2n+1}A_v^{2n+1}$$

$$if\ A_v < 1, as\ n \to \infty, V_O \to V_{DD}/2$$

$$if\ A_v > 1, as\ n \to \infty, V_O \to V_{OH}$$

Operation of CMOS Logic

distinguishability

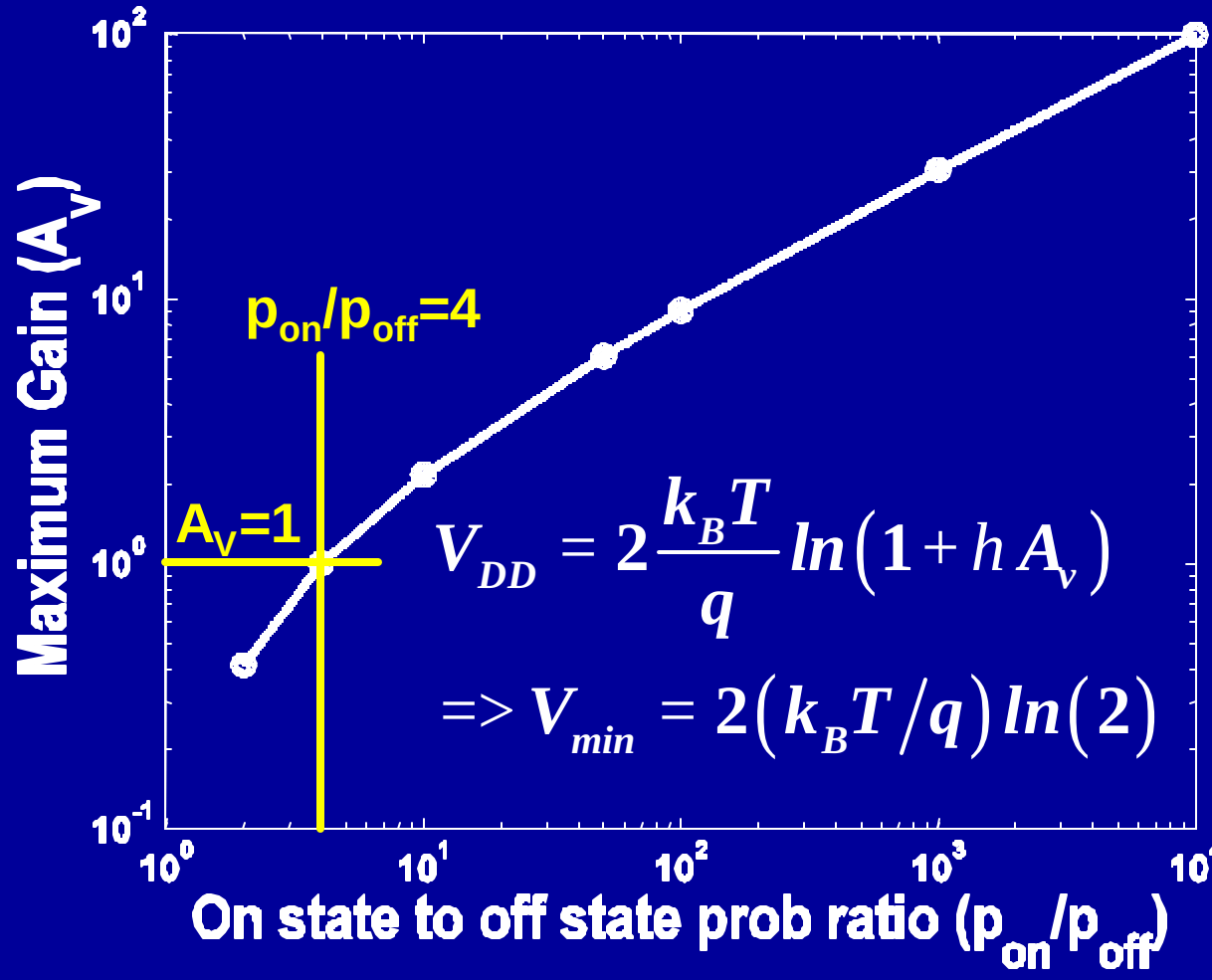
=> Gain (A_V) > 1

for CMOS inverter

Minimum p_{on}/p_{off} is *"4"*

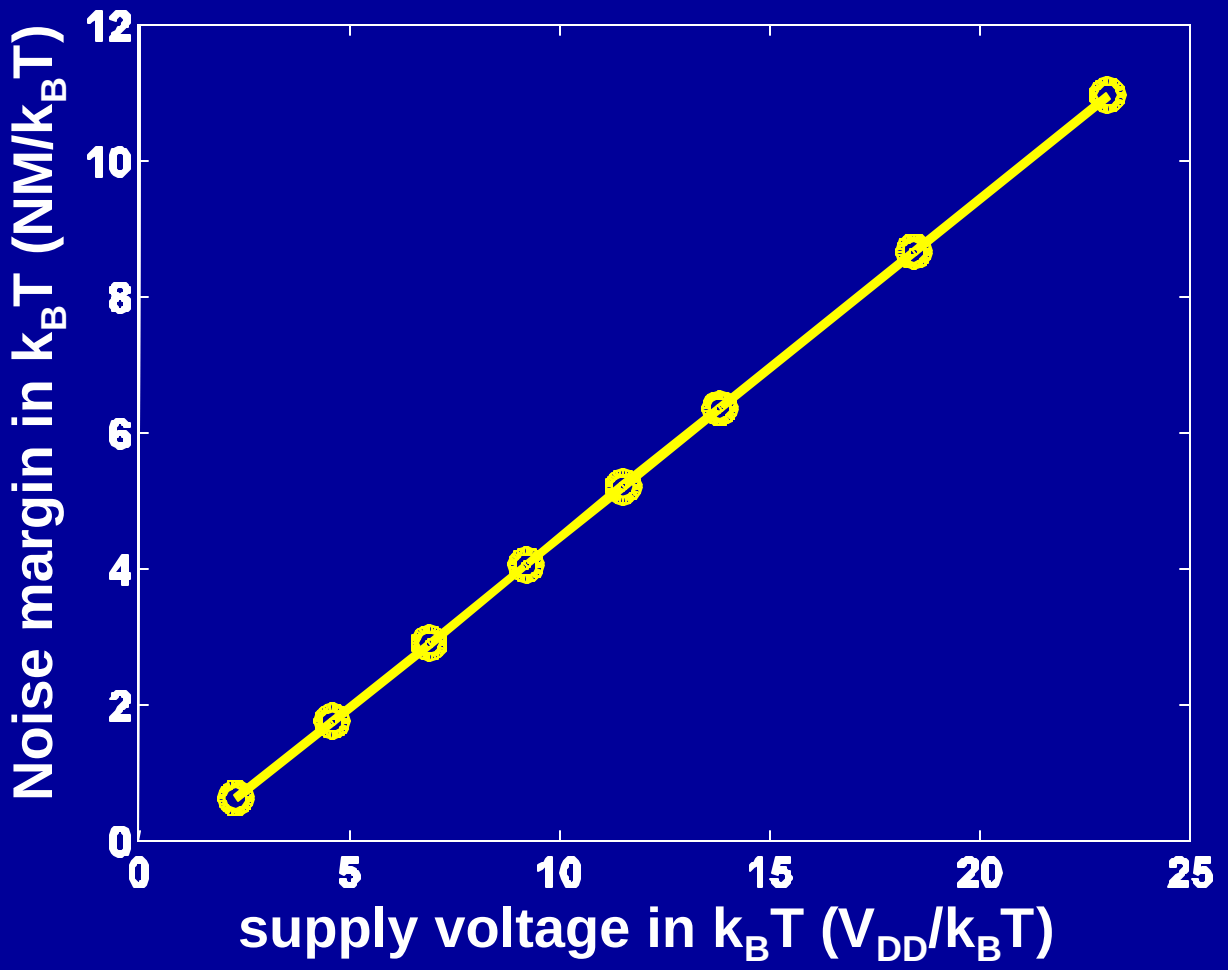
and

not "2"



$V_{min} = k_B T \ln(2)$ Device to Inverter $V_{min} = 2k_B T \ln(2)$

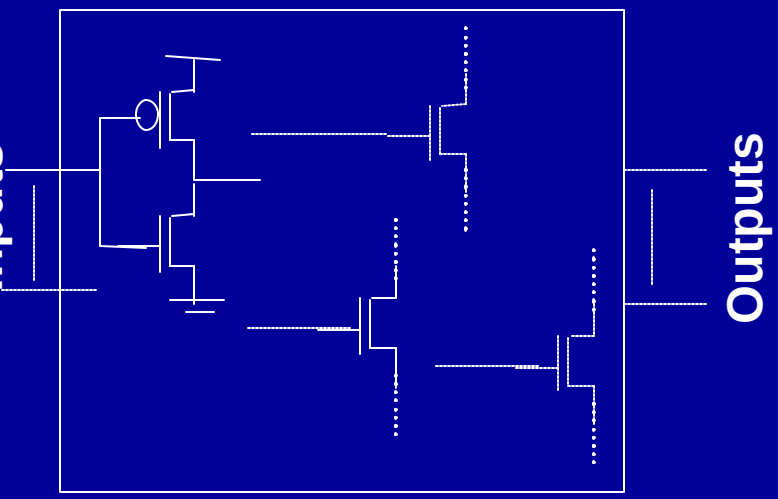
Operation of CMOS Logic

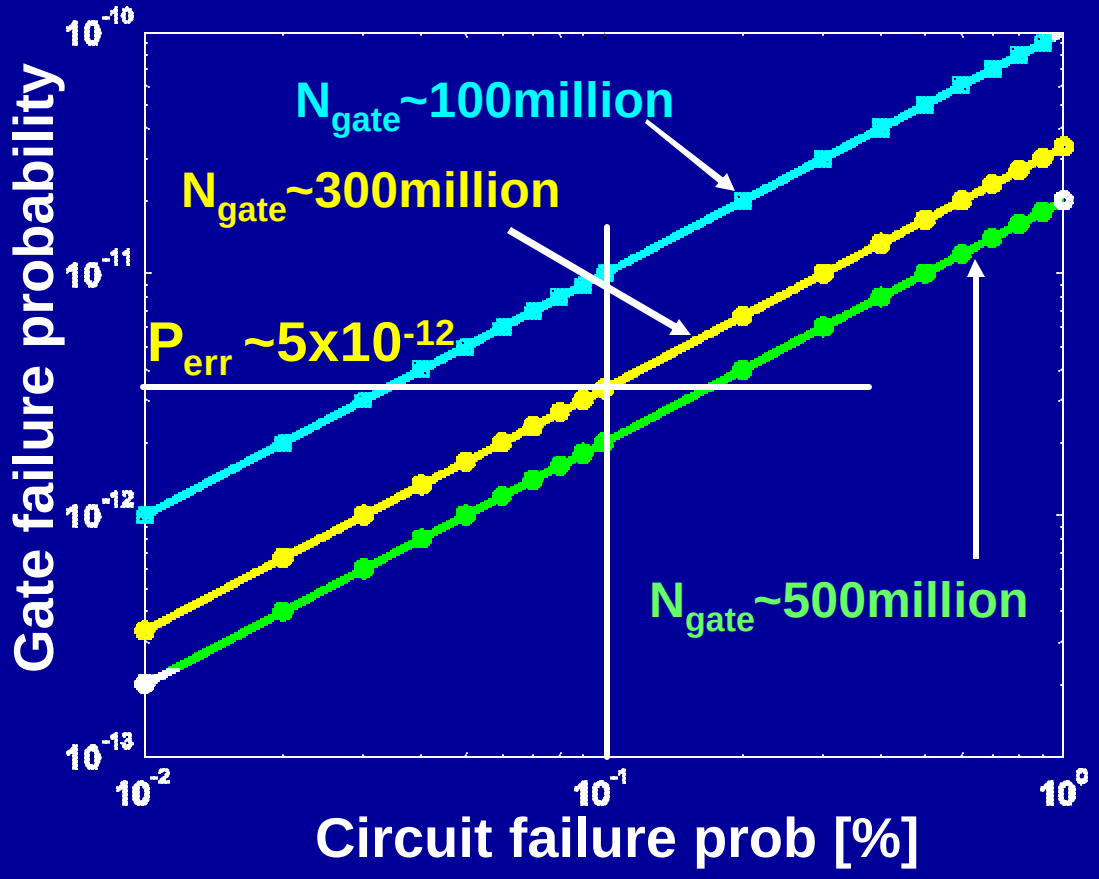


To ***prevent spontaneous change of state*** noise margin needs to be at least higher than k_BT

$$\Rightarrow V_{DD} > 3k_BT$$

Reliability of Circuit Operation





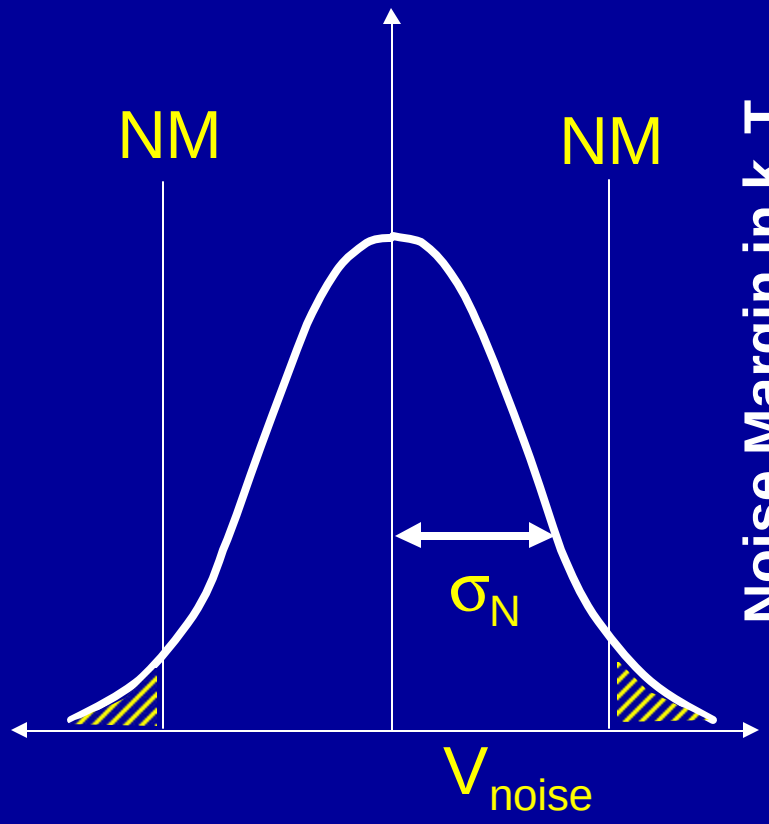
# of gates = N_{gate}

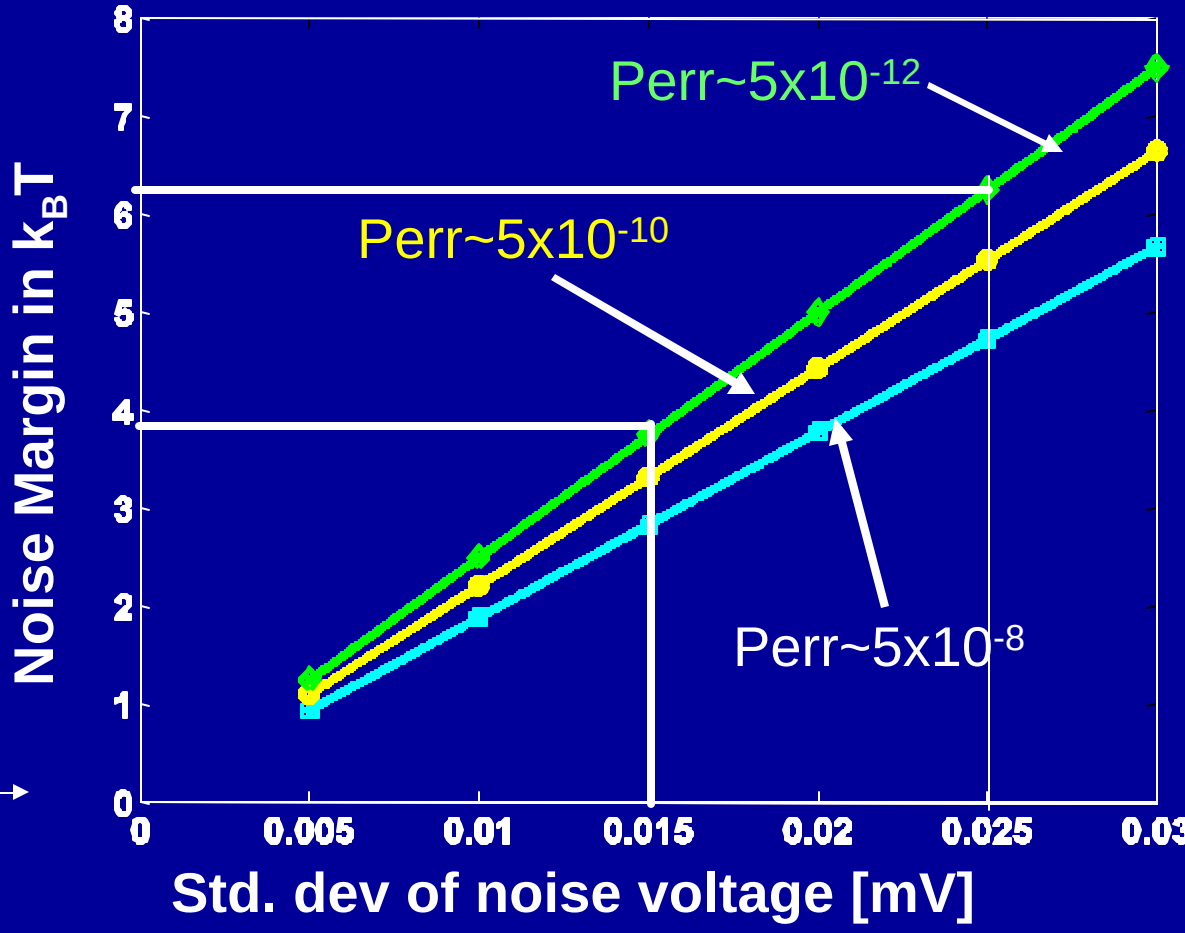
Prob. of error of a single gate = P_{err}

Prob. of error of the circuit = $P_{circ} = 1 - (1-P_{err})^{Ndev}$

Reliable operation of the circuit imposes stronger constraint on the reliability of the gate operation

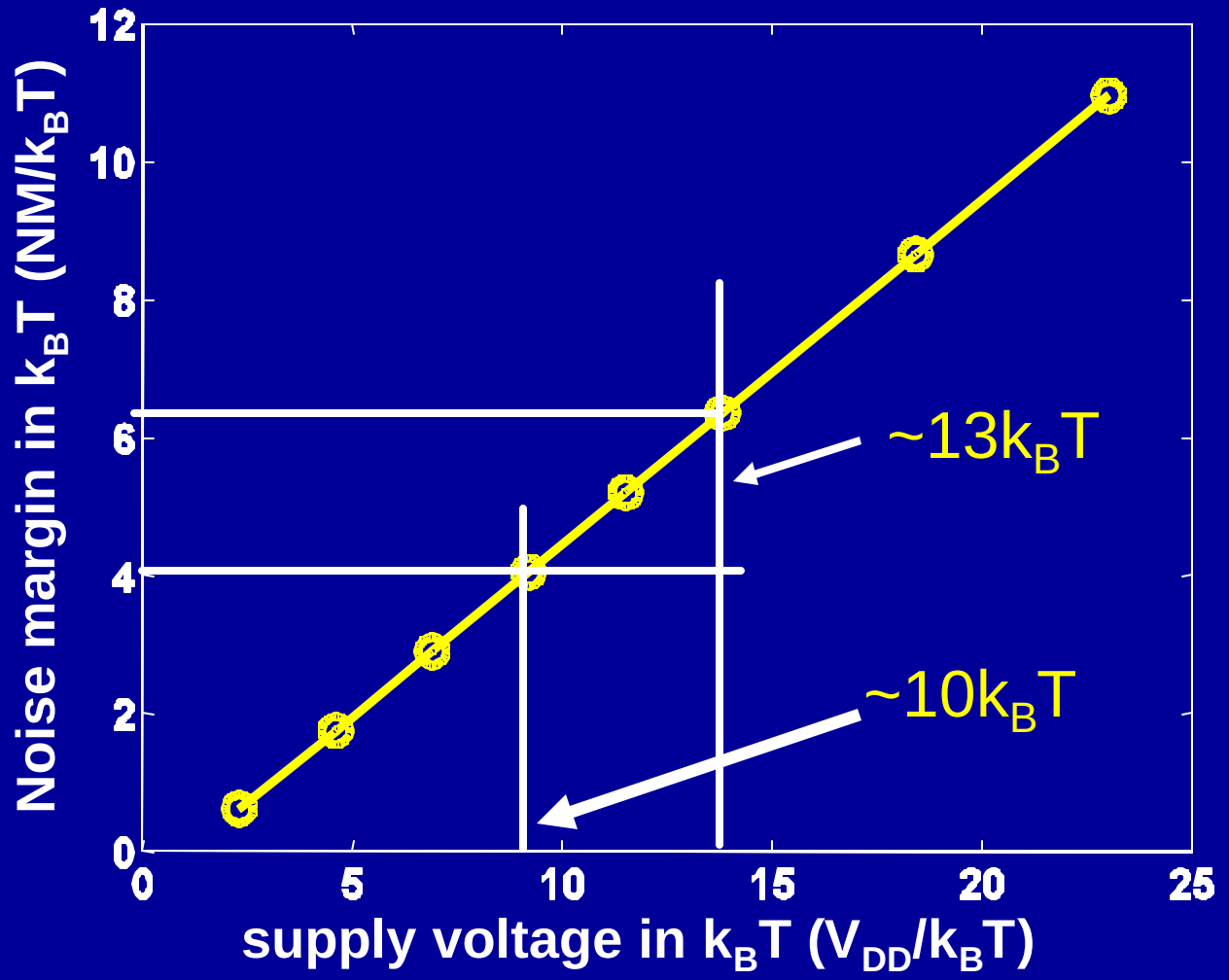
Reliability of CMOS Inverter Operation





Higher noise requires a larger noise margin for reliable operation

Reliability of CMOS Inverter Operation



$V_{min} = 2k_BT\ln(2)$

Reliability

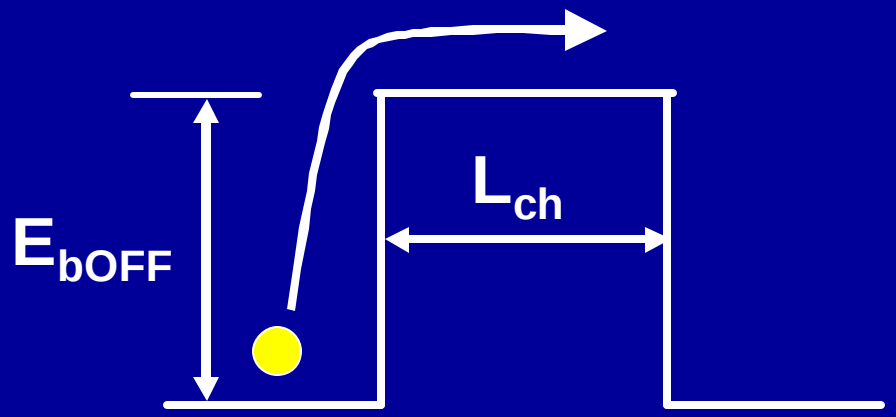
$V_{min} = 10k_BT$

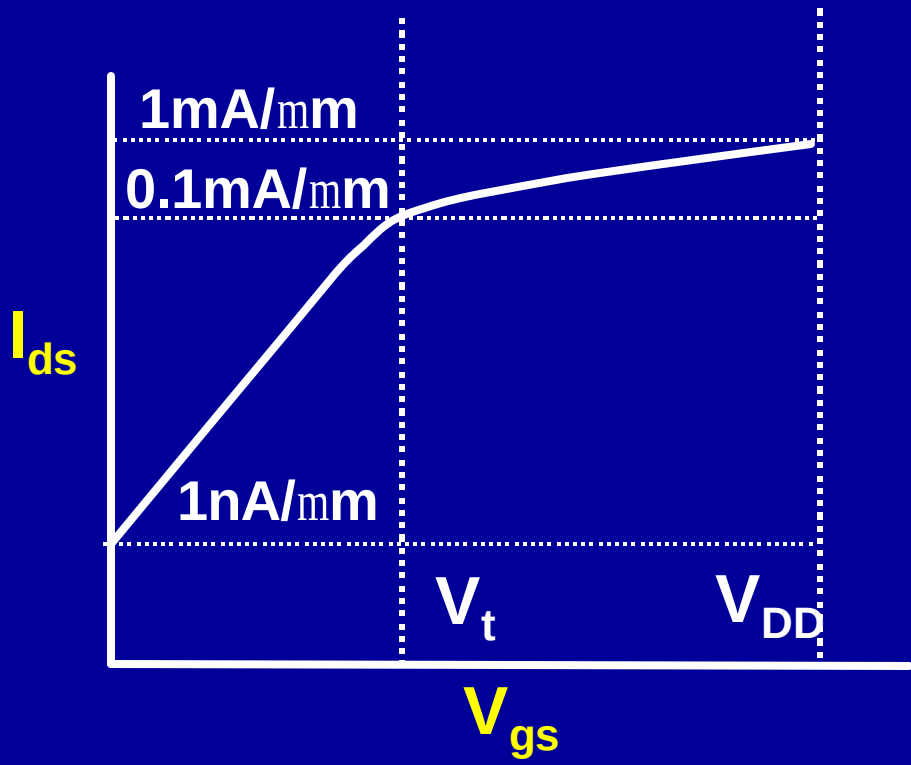
Operations of CMOS Logic

1. It is a “single well - double barrier” system.
2. Presence or absence of charge at the “well” determines the logic state
3. At both logic states, the well is strongly coupled to V_{DD} or GND through a “on” device

The “driven” nature of CMOS logic makes it reliable even at very low voltage operation

Limit of p_{off} : Leakage Power





$$I_{leak} = I_0 \exp\left(-\frac{E_{bOFF}}{k_B T}\right) = I_0 p_{off}$$

$$I_{leak} \sim 1nA/\mu m \Rightarrow p_{off} \sim 10^{-5}$$

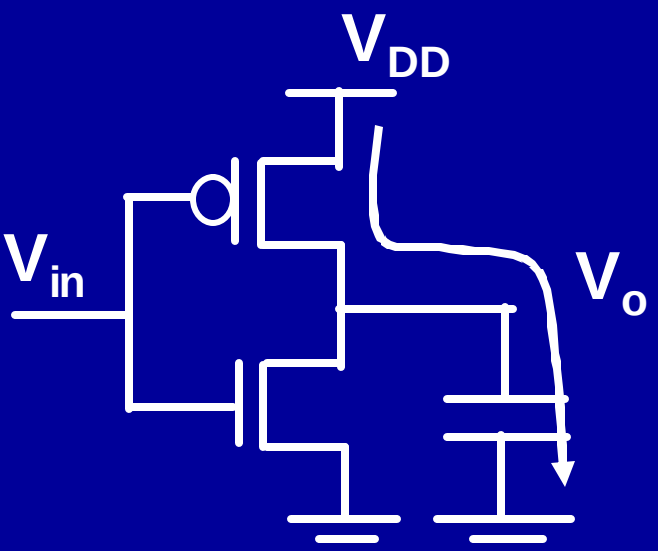
$$\Rightarrow E_{bOFF} = k_B T \cdot \ln\left(10^5\right) \sim 11 k_B T$$

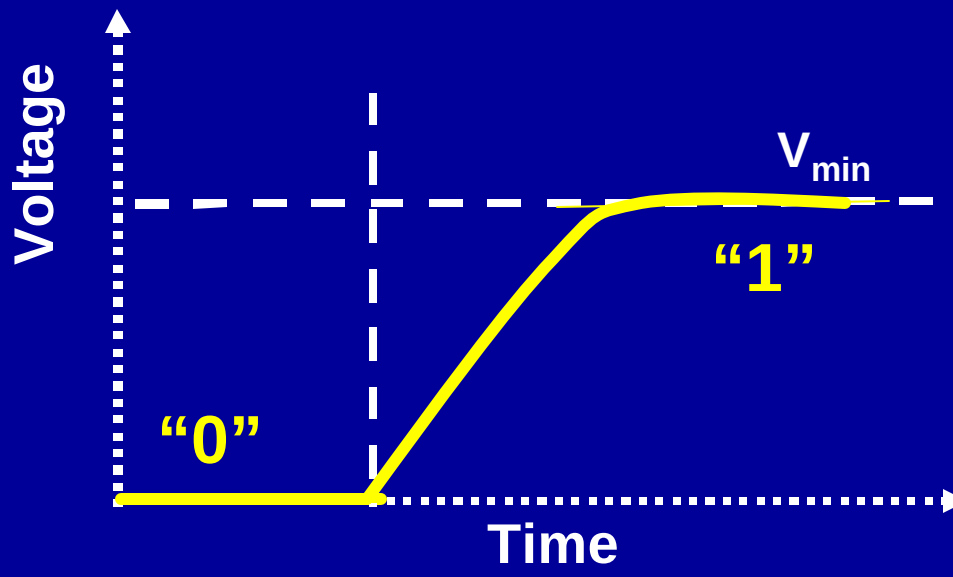
E_{bOFF} ~ $11k_BT$ helps to meet a leakage requirement of 1nA/㎛

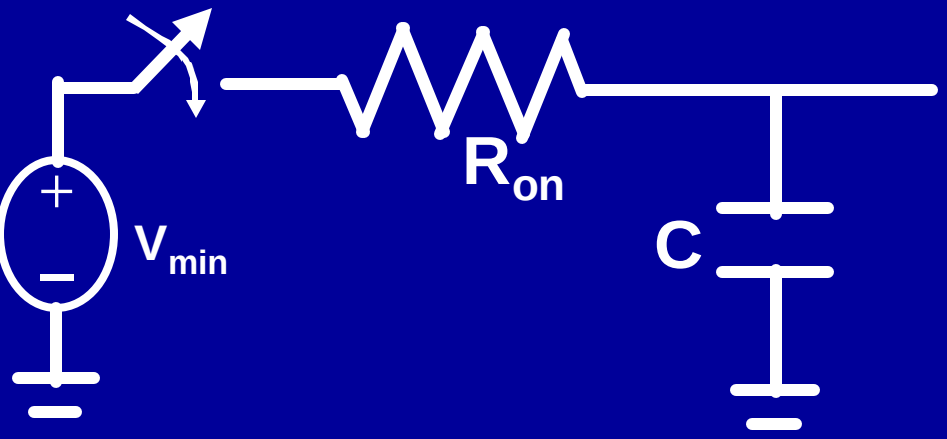
Outline

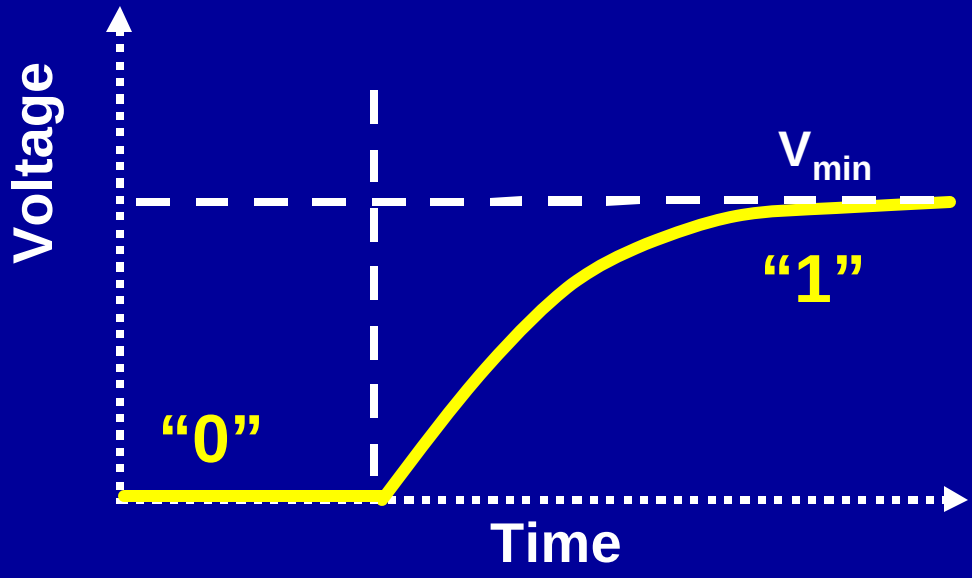
- Switching energy in charge transfer based Digital Logic
  - Basics and Physical Limits
- Practical consideration for switching energy in CMOS Logic
  - Static requirements
  - Dynamic requirements
  - Circuit/System considerations
- What can we do to reduce switching energy ?
- Summary

Delay in CMOS Logic

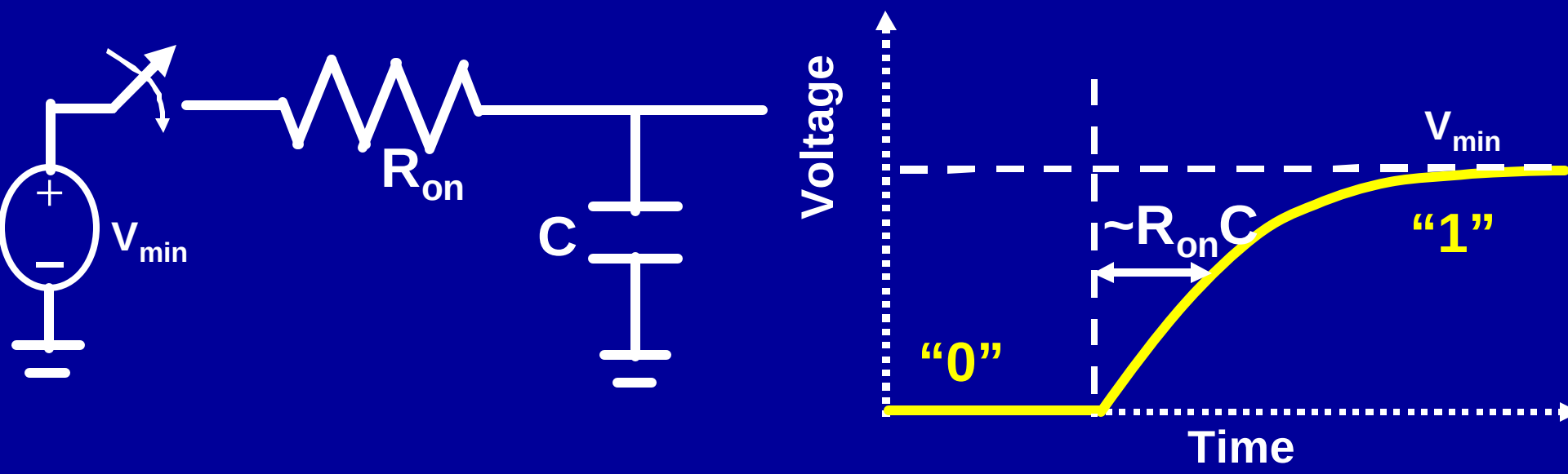






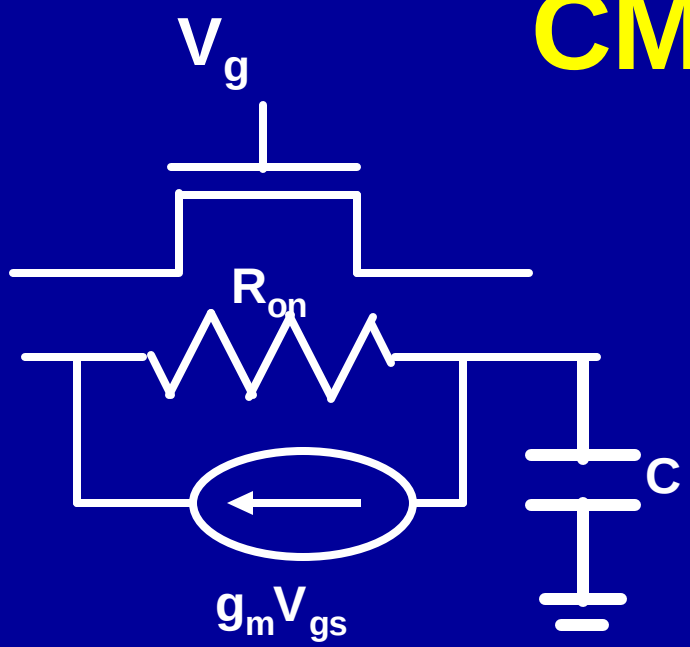


Delay and Switching Energy



- **Delay through an RC circuit**
  - **Independent of applied voltage V_{min}**
  - **Lower C reduces both delay and switching energy : key principle in technology scaling**

Delay and Switching Energy : CMOS Logic



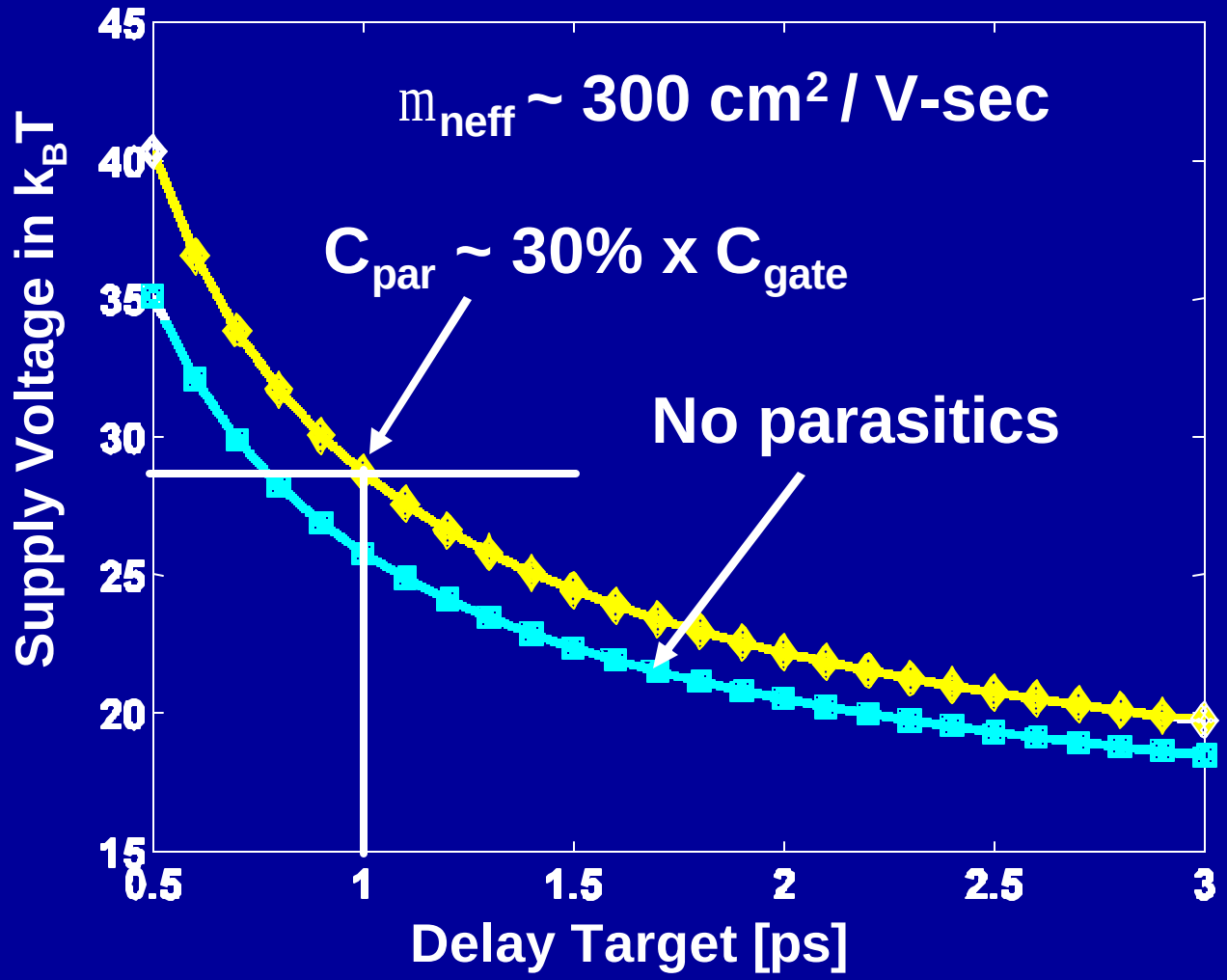
The dependence of R_{on} on the applied gate bias makes delay and energy correlated for CMOS

$$C_{gate} V_{DD} = t\, I_{on}$$

$$For: W_P = 2W_N = 2L_{min}$$

$$\left(1 + \frac{C_{par}}{C_{ox}}\right) 3L_{min}^2 C_{ox} V_{DD} = t m_{\text{eff}} \frac{L_{min}}{2L_{min}} C_{ox} \left(V_{DD} - h\frac{E_{bOFF}}{q}\right)^2$$

Impact of Delay on Minimum V_{DD}

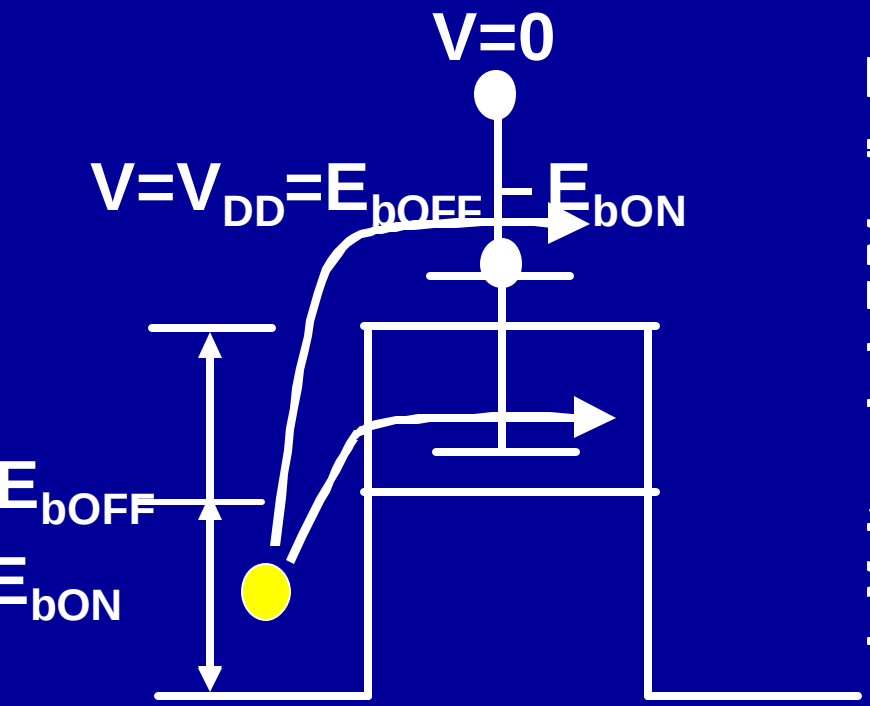


V_{min}= 10k_BT

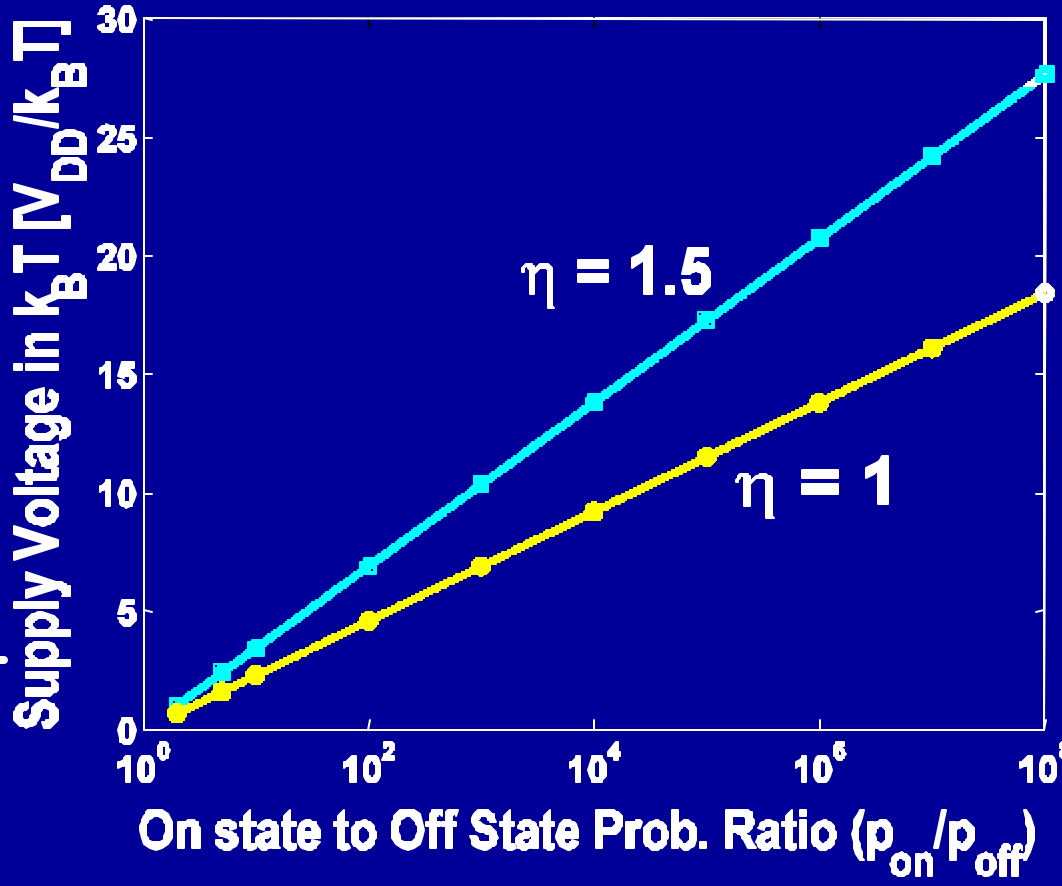
Delay (1ps)

V_{min}= 28k_BT

Non-ideal subthreshold slope

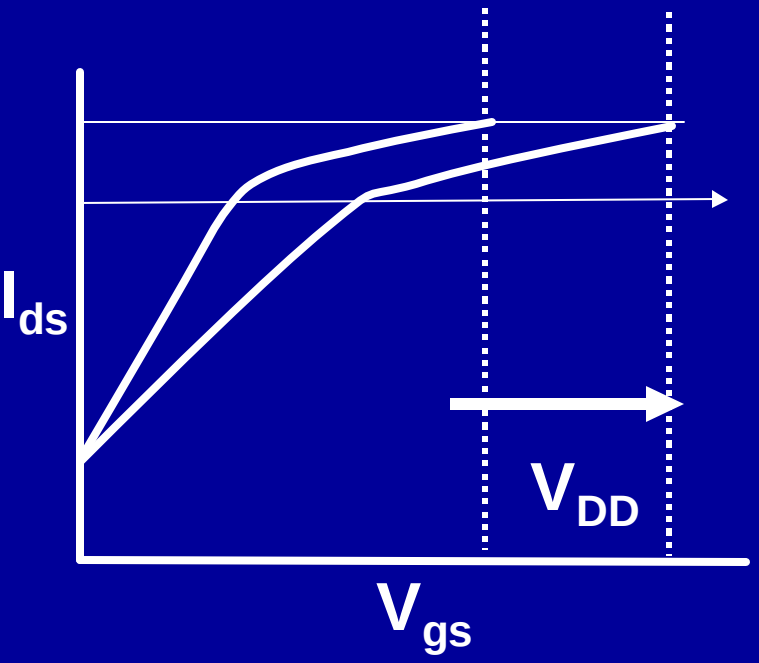


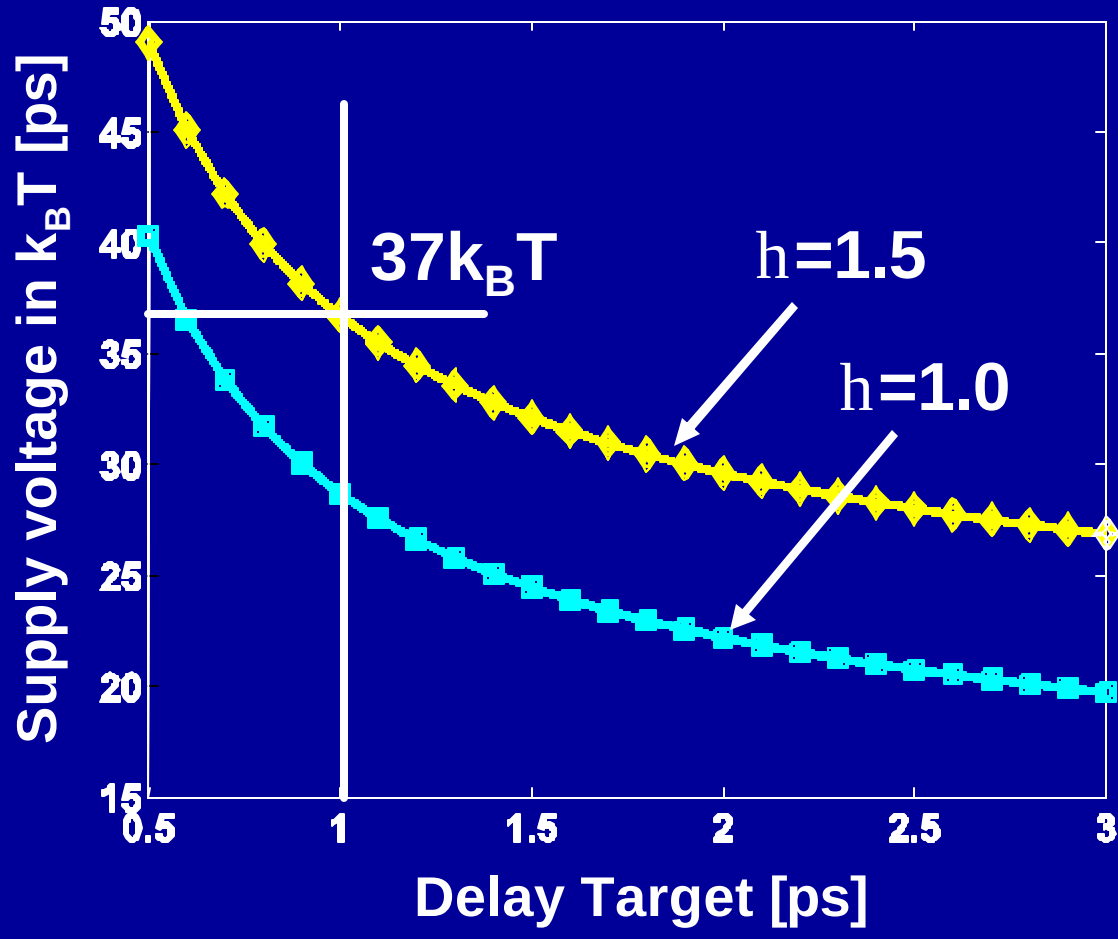
$$V_{DD} = h\frac{k_B T}{q} ln\left(\frac{p_{on}}{p_{off}}\right)$$



A larger subthreshold slope requires a higher V_{DD} to achieve a pon/poff

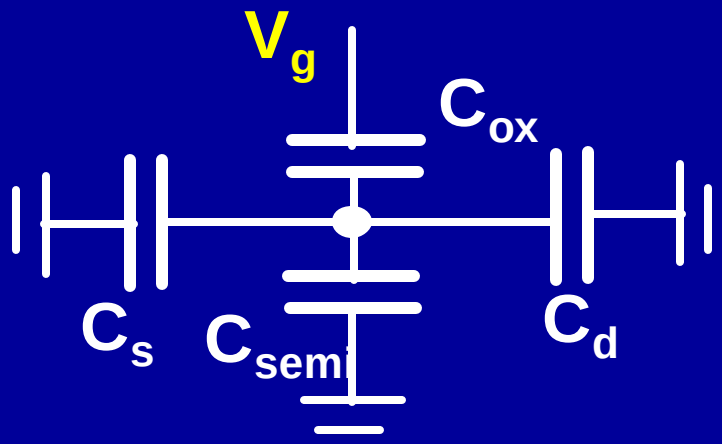
Non-ideal subthreshold slope



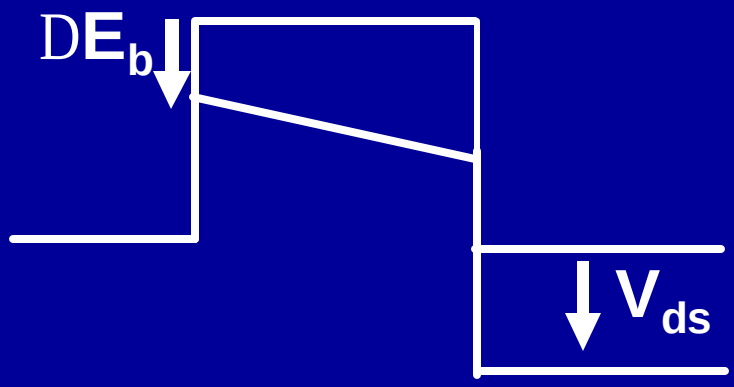


Non-ideal subthreshold slope increases the V_{DD} required to achieve a certain delay

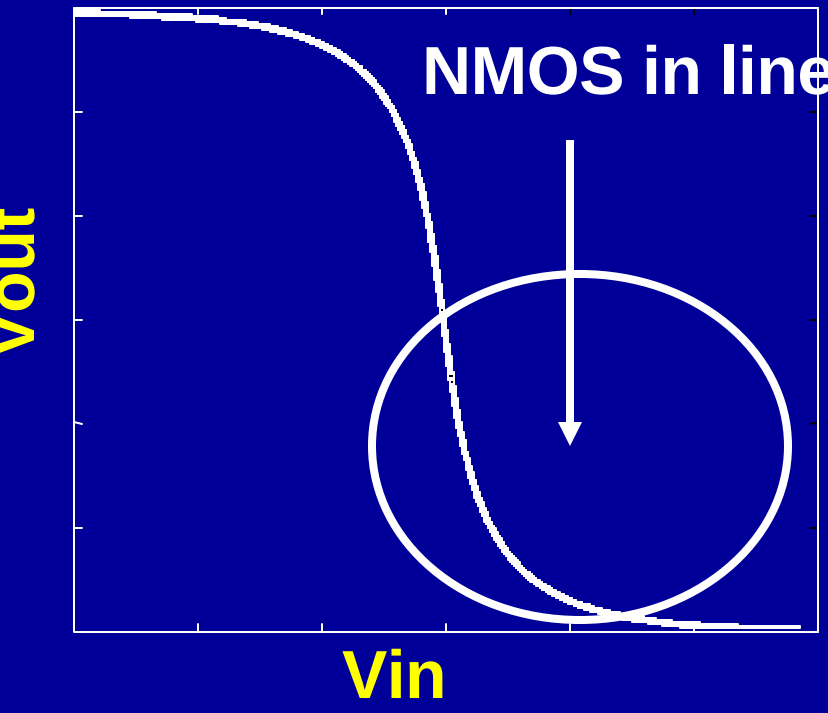
2-D Electrostatics

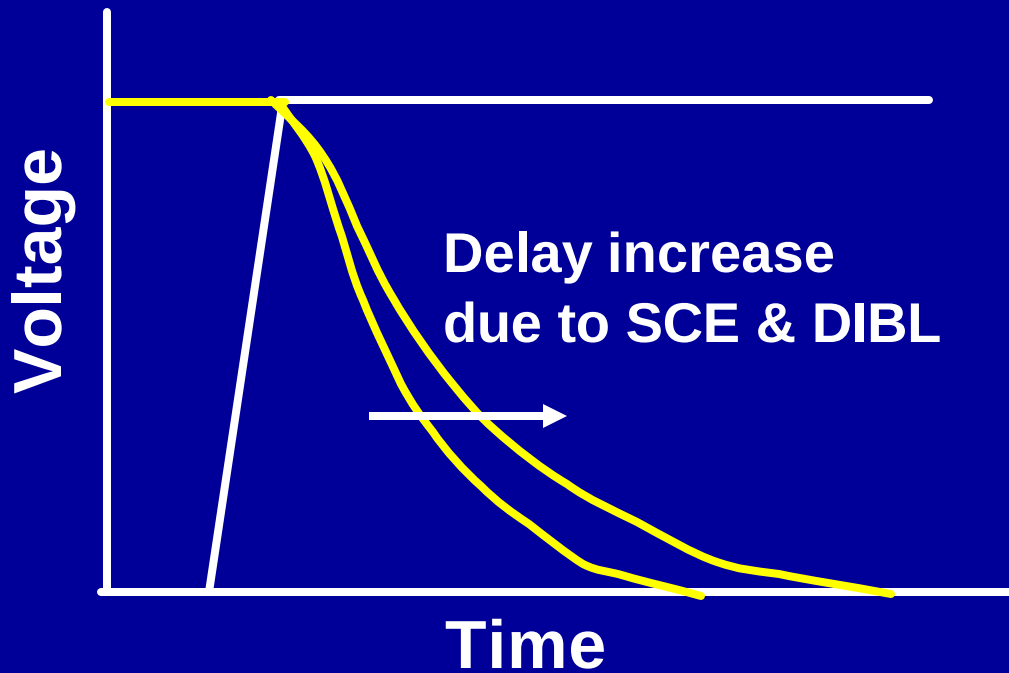


Degraded Sub-slope

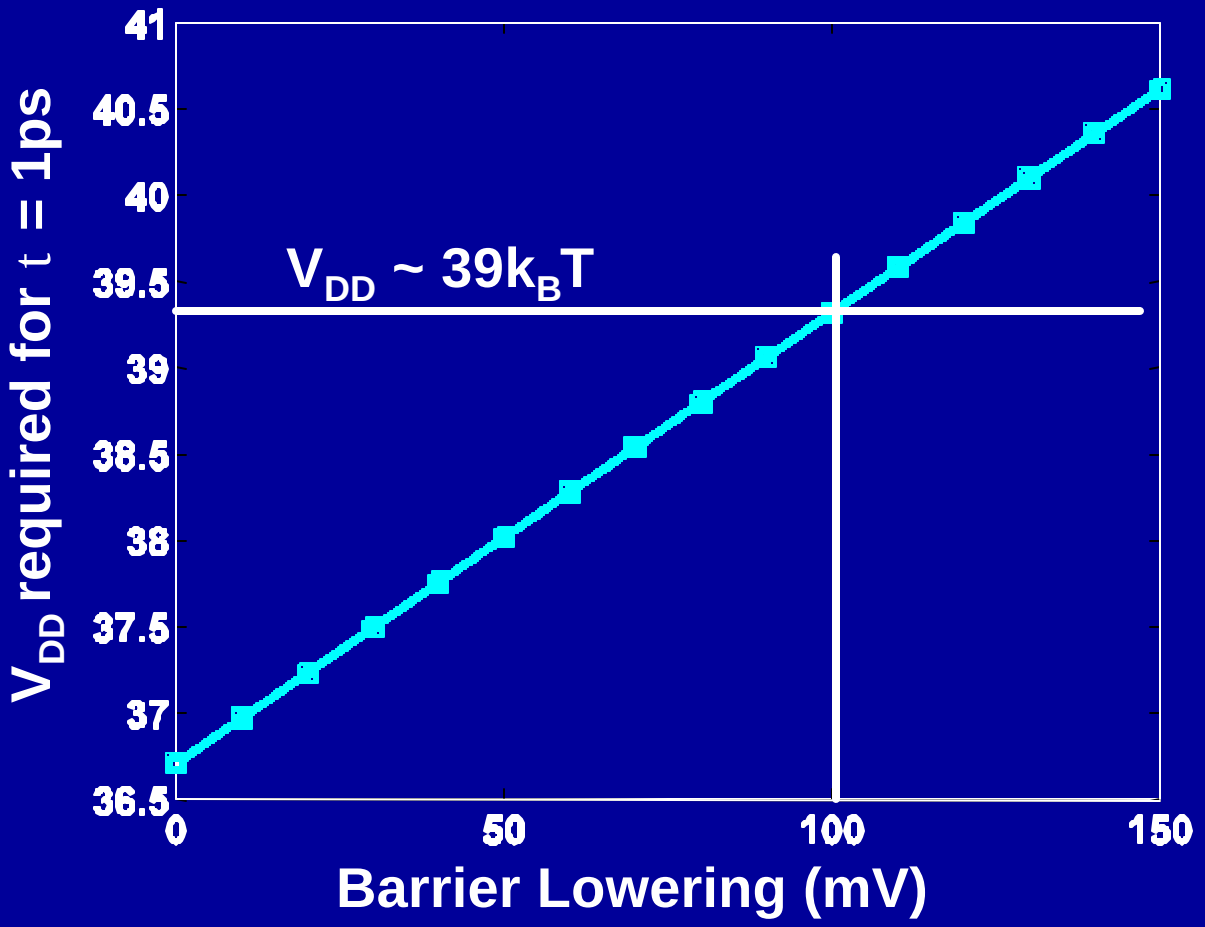


Drain Induced Barrier Lowering



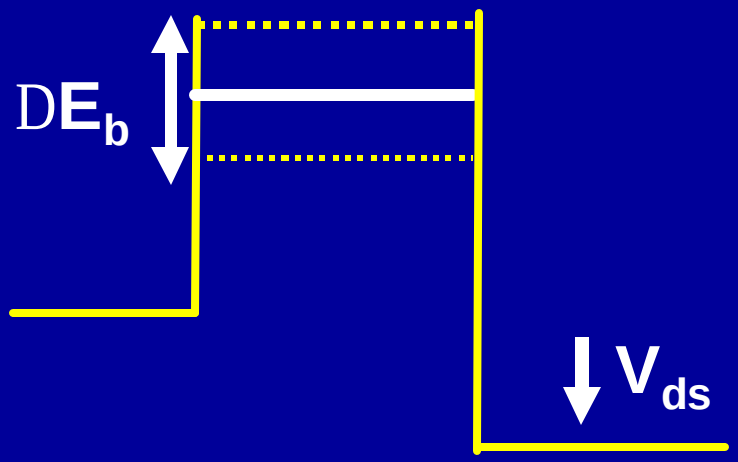


2-D Electrostatics

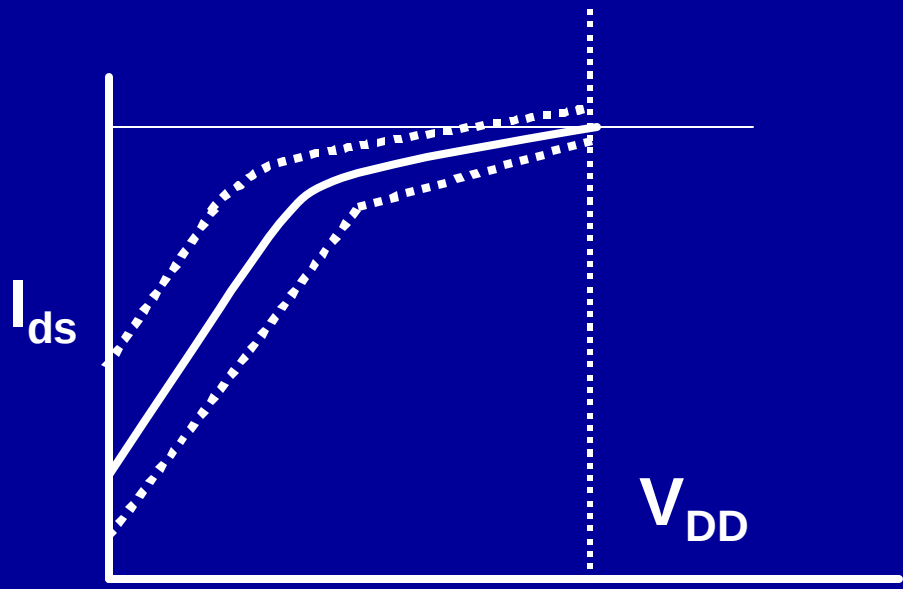


Under same leakage power 2-D effect increases the V_{DD} required to achieve a target delay

Process Variability



Variation in Process Parameters

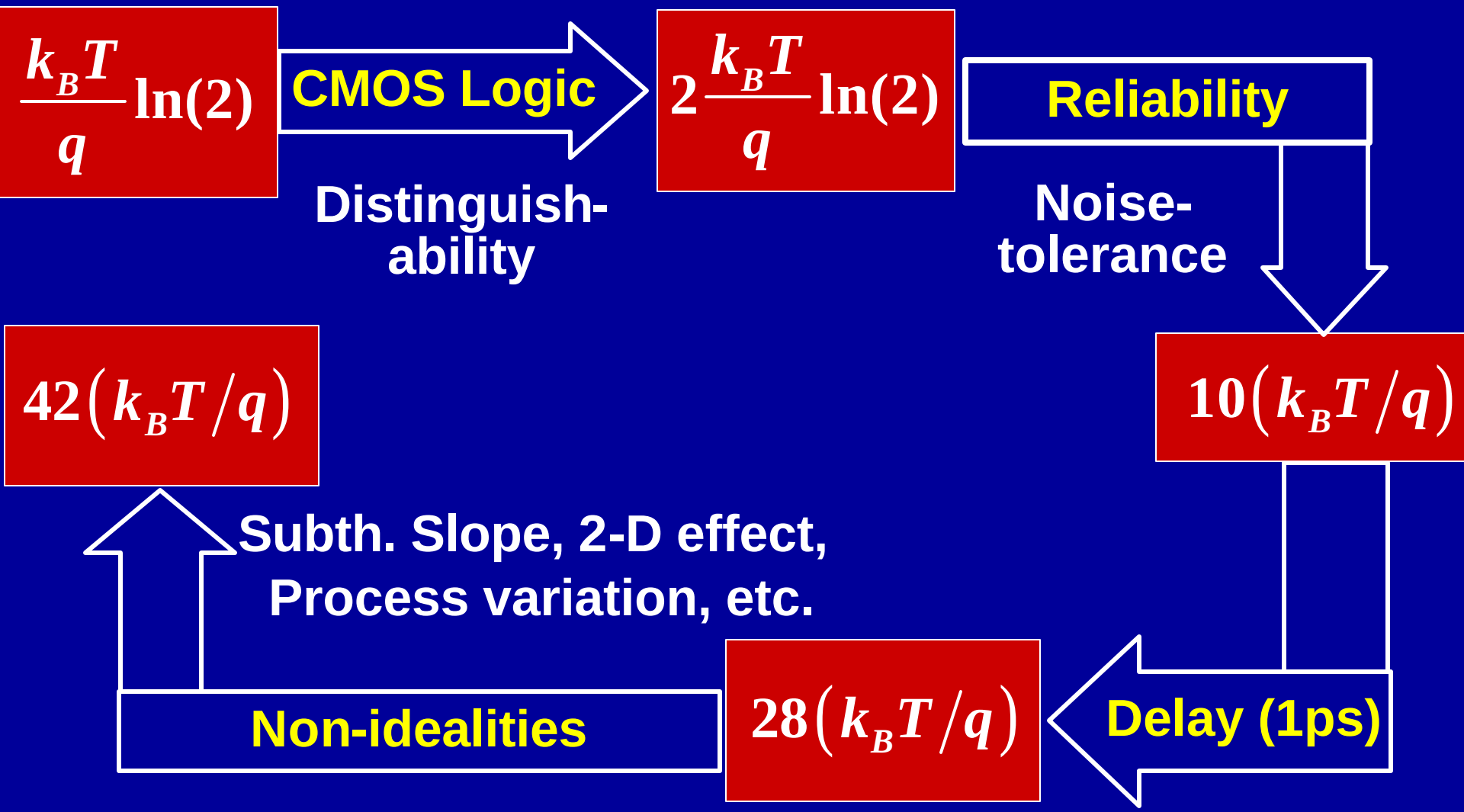


- Leakage ~ p_{off} variation
- Reliability ~ p_{on}/p_{off} variation
- Delay ~ variation in E_{bOFF} will change the delay

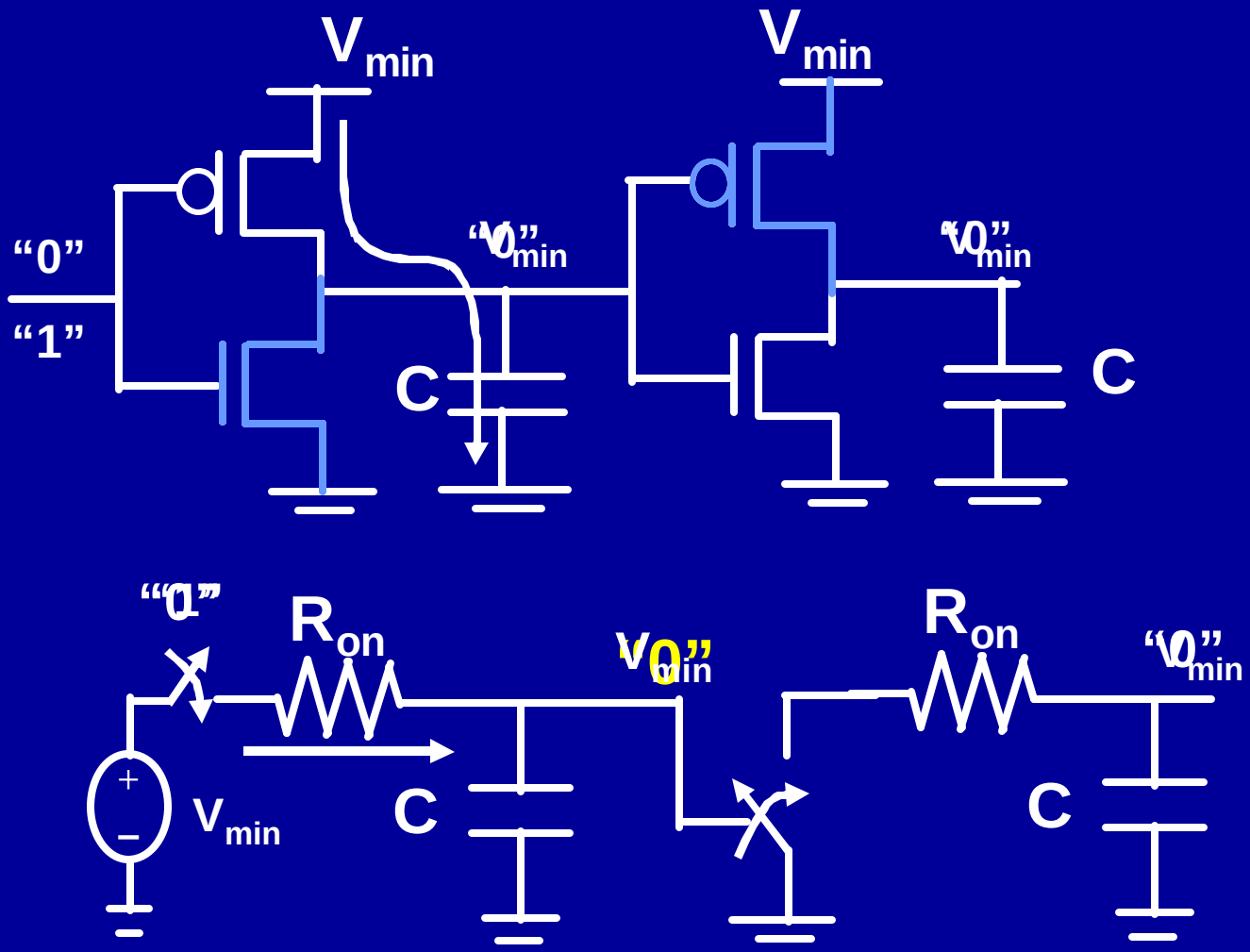
The designed E_{bOFF} and V_{DD} needs to be increased to account for the effect of variation

$\pm$ 10% variation in E_{bOFF} => V_{DD} ~ $42k_BT$

Why We are using V_{DD} much larger than the $k_BT\ln(2)$ limit?

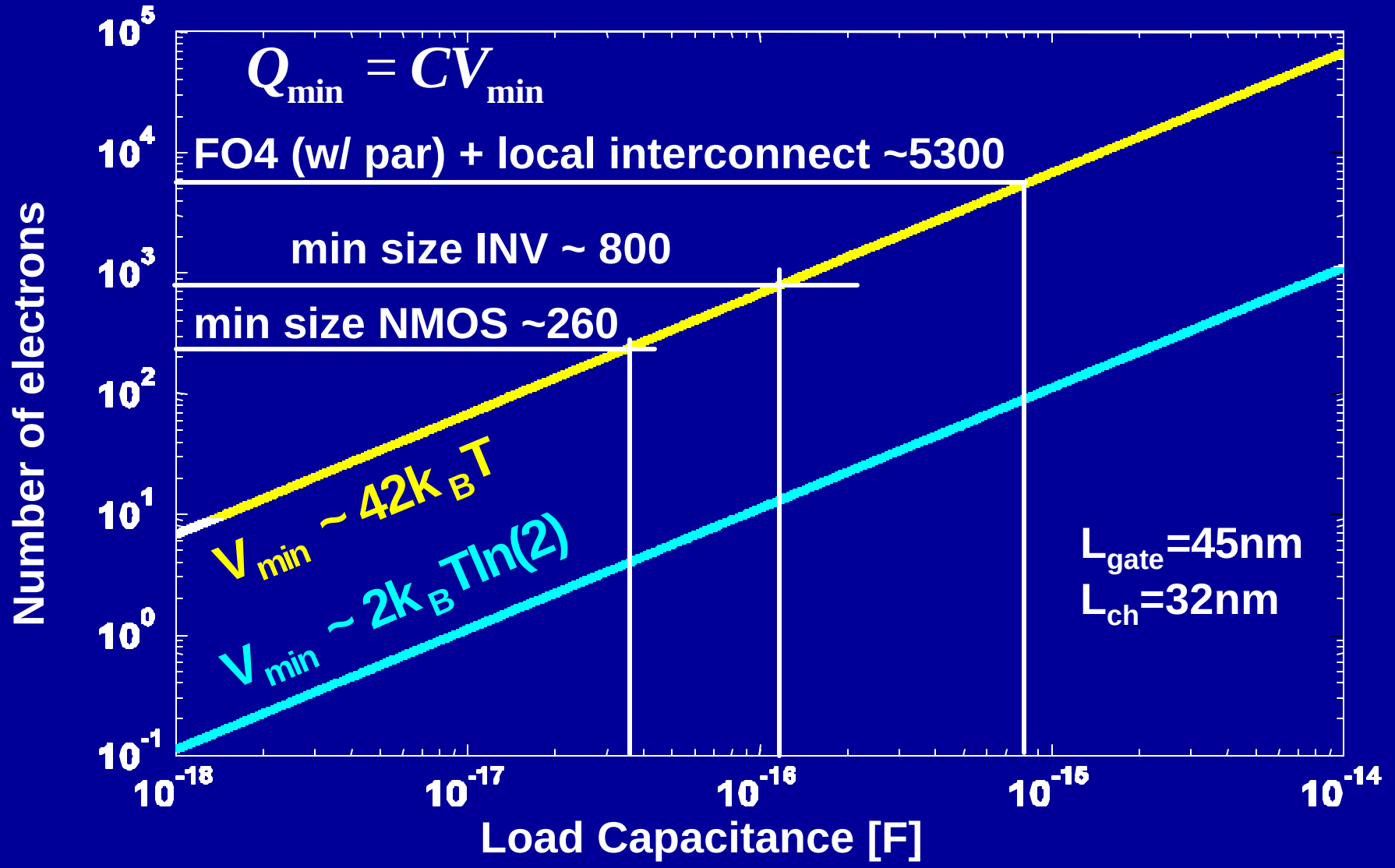


Drivability in Digital Logic



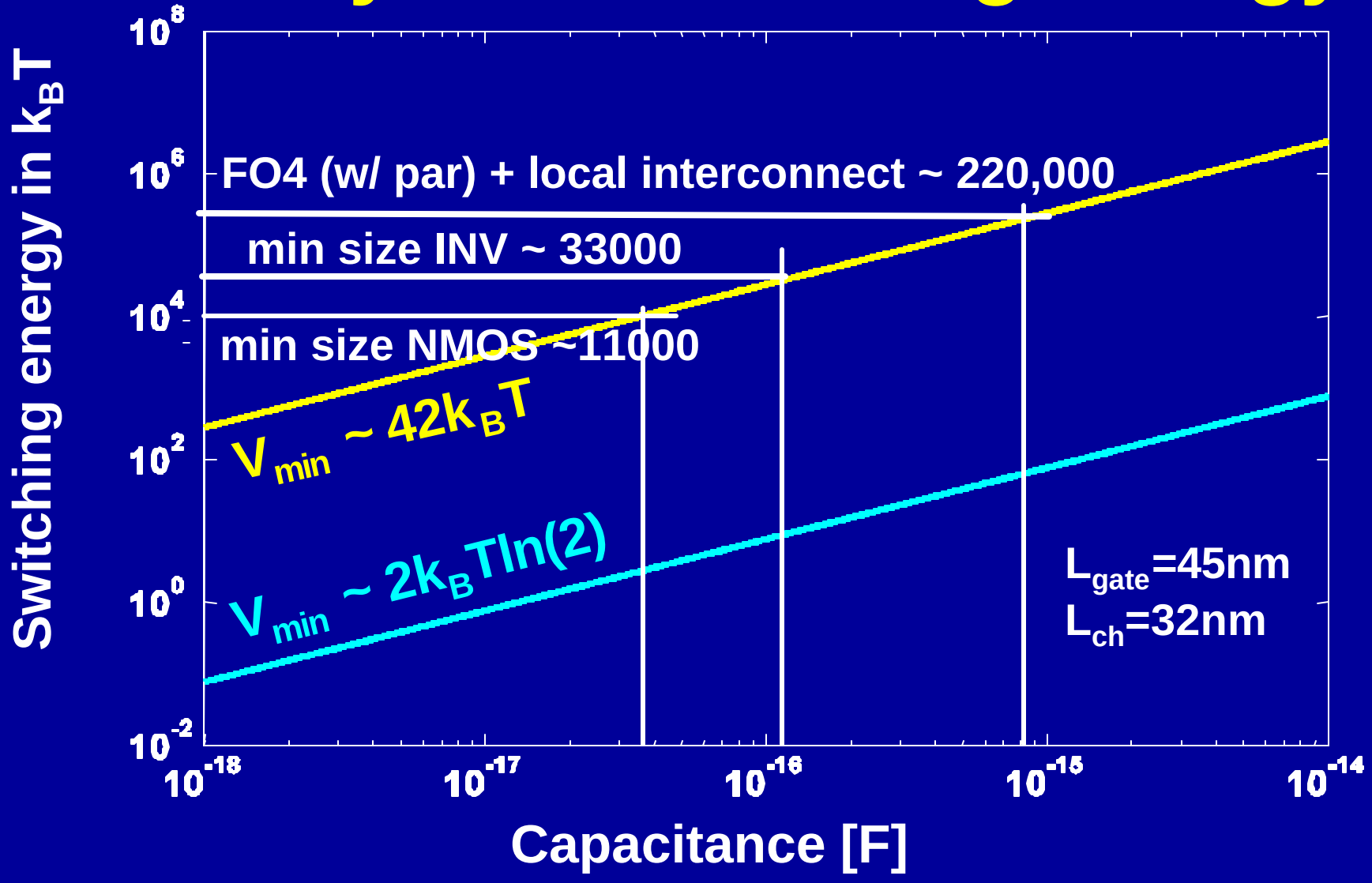
Vmin needs to be developed across a finite capacitance for driving the next gate

Drivability and Minimum Charge



Drivability requirement does not allow to operate with a single electron for CMOS logic operation

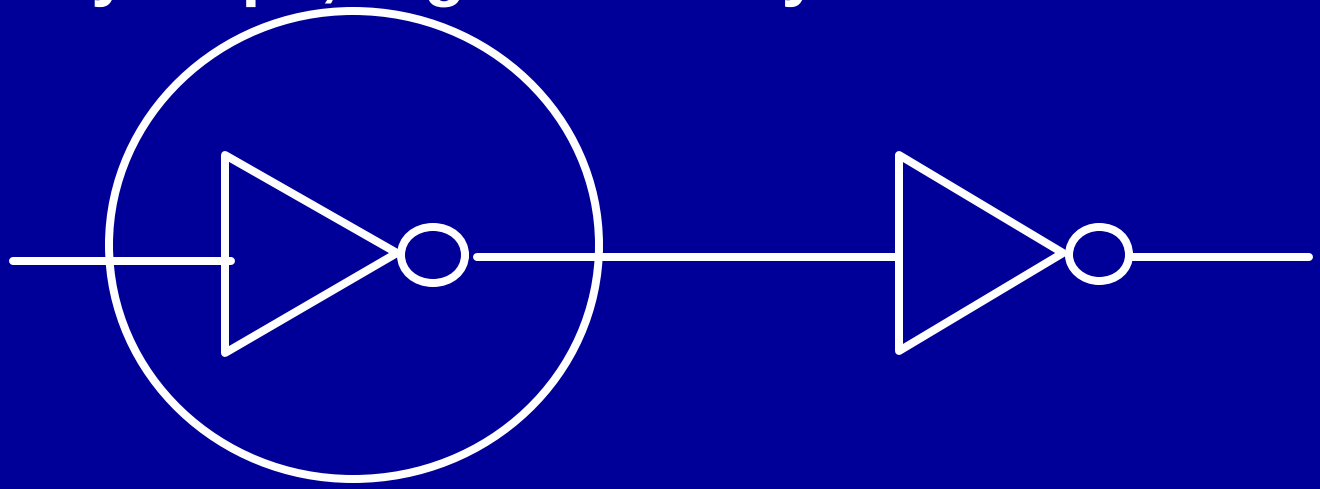
Drivability and Switching Energy



Drivability requirement increases the minimum switching energy for an inverter to ~ 33,000 k_BT

Switching Energy in CMOS Logic

Delay ~ 1ps, High reliability



$k_B T \ln(2)$ → **Delay/Reliability** → $42 k_B T$

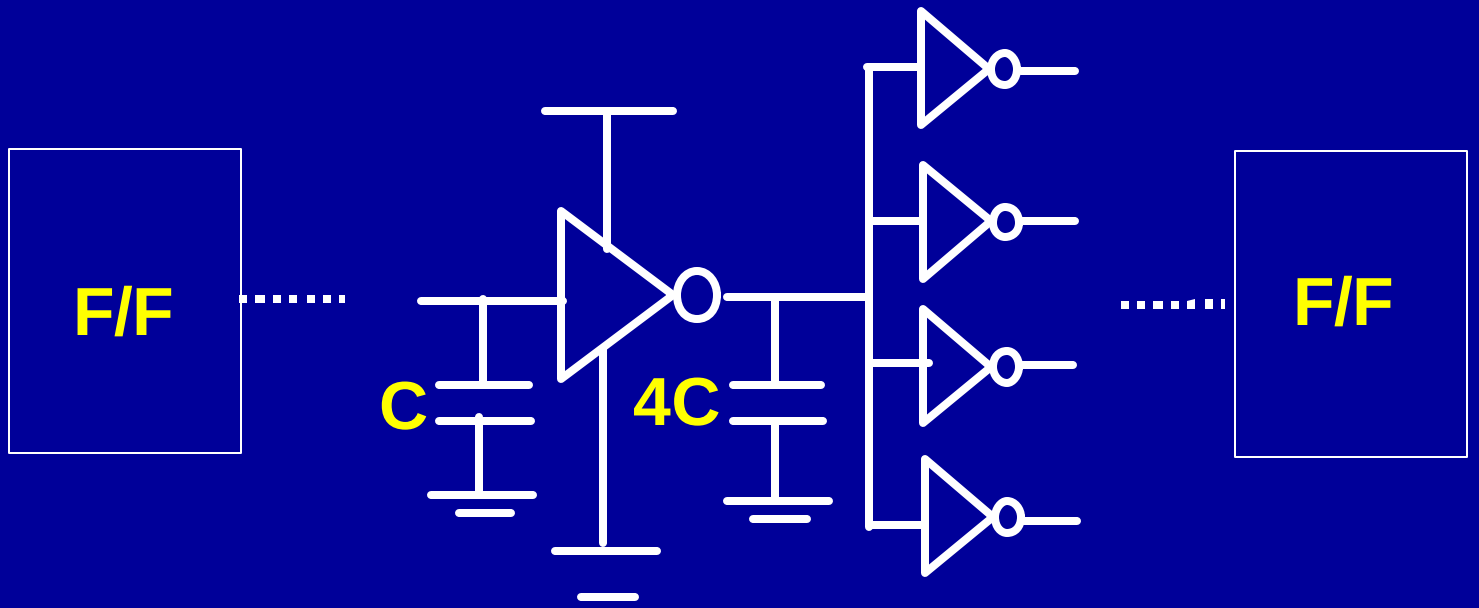
Drivability

$33000 k_B T$

Outline

- **Switching energy in charge transfer based Digital Logic**
  - Basics and Physical Limits
- **Practical consideration for switching energy in CMOS Logic**
  - **Static requirements**
  - **Dynamic requirements**
  - **Circuit/System considerations**
- **What can we do to reduce switching energy ?**
- **Summary**

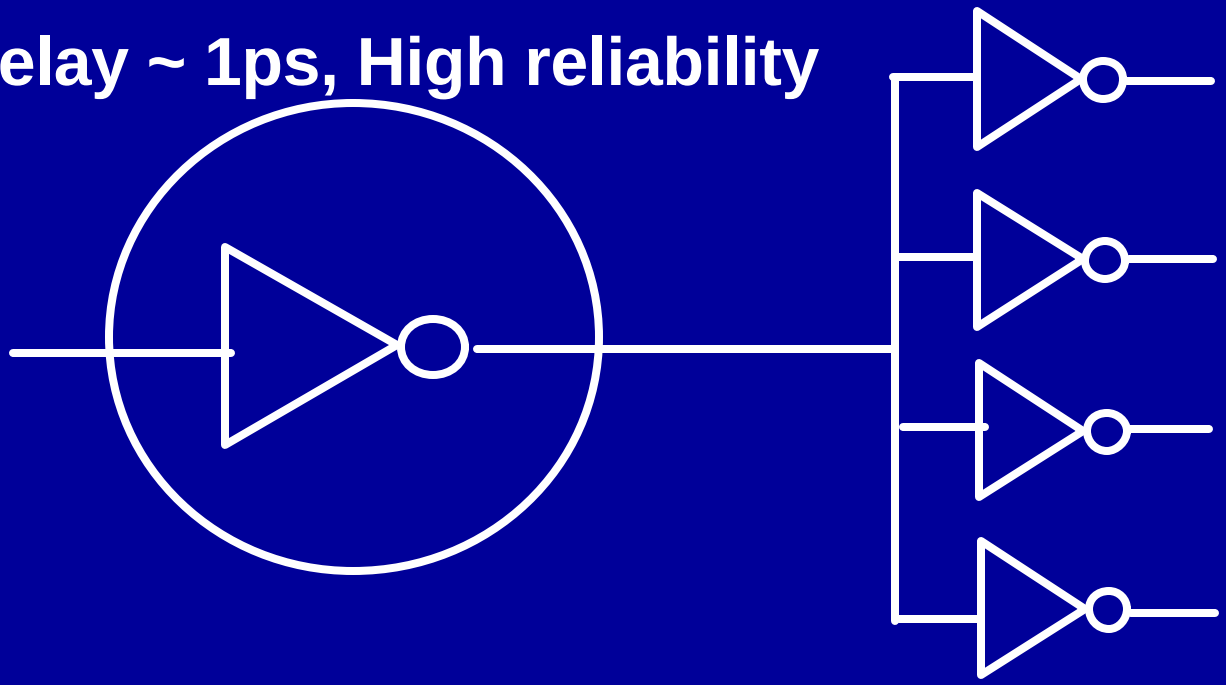
Operation of CMOS Circuits



- For logic operation a gate has to drive more than one gates in a CMOS logic
- Typical fanout is assumed to be 4

Switching Energy in CMOS Logic

Delay ~ 1ps, High reliability



$k_BTln(2)$

Delay/Reliability

$42k_BT$

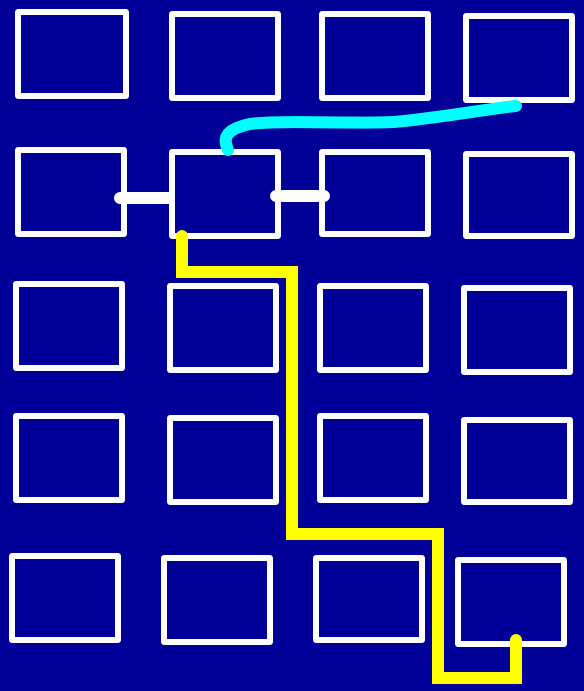
Drivability

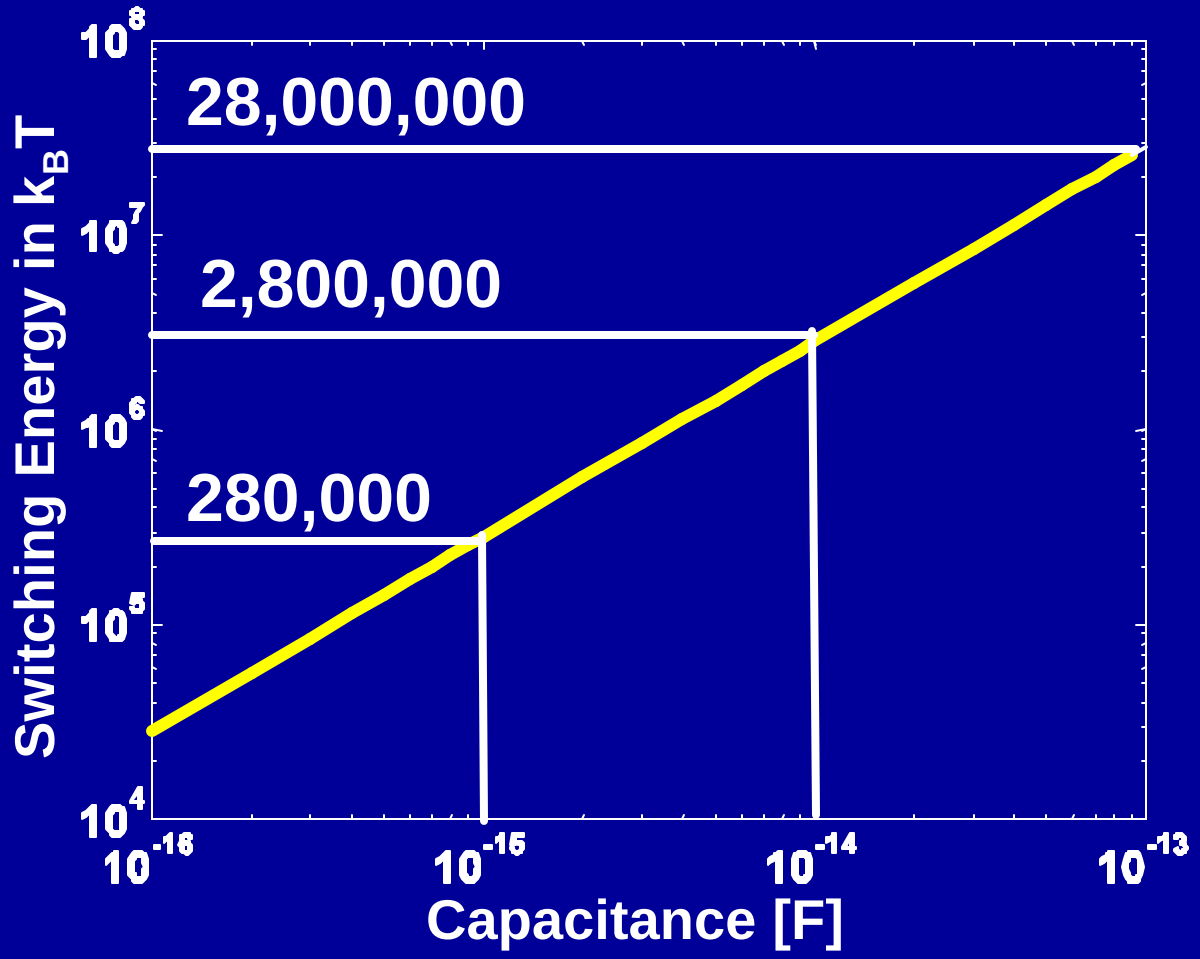
$33000k_BT$

FO4

$220{,}000k_BT$

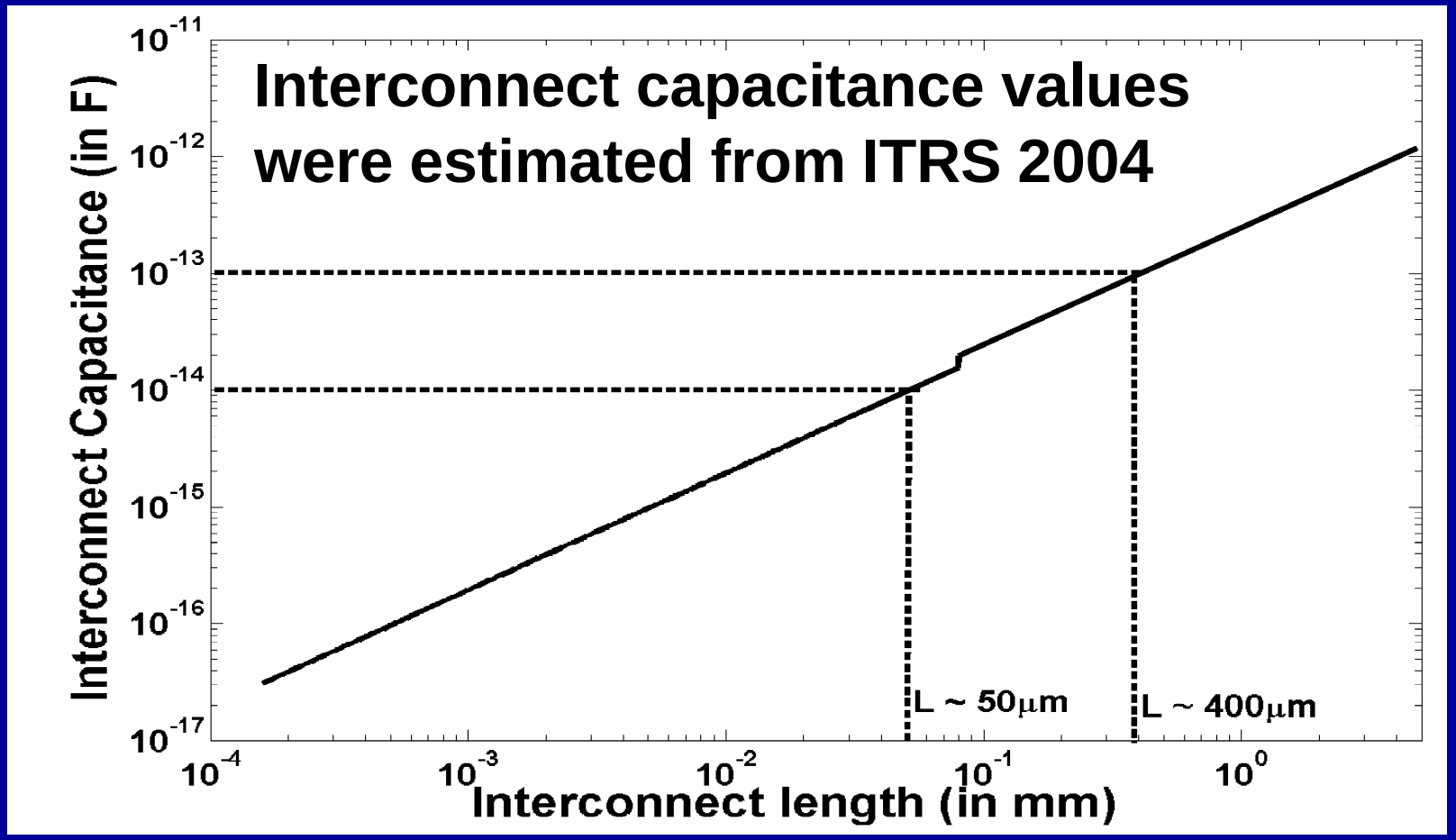
Switching Energy for a System





Driving “long” interconnects can significantly increase the switching energy

Switching Energy for a System



Interconnect of length ~400㎜ has 100 fF of cap which requires ~28,000,000 $k_B T$ to switch

How many long interconnects exists in an Integrated Circuits?

- For a logic block of 'N' elements (say inverters) the total number of external interconnects : $T = kN^p$

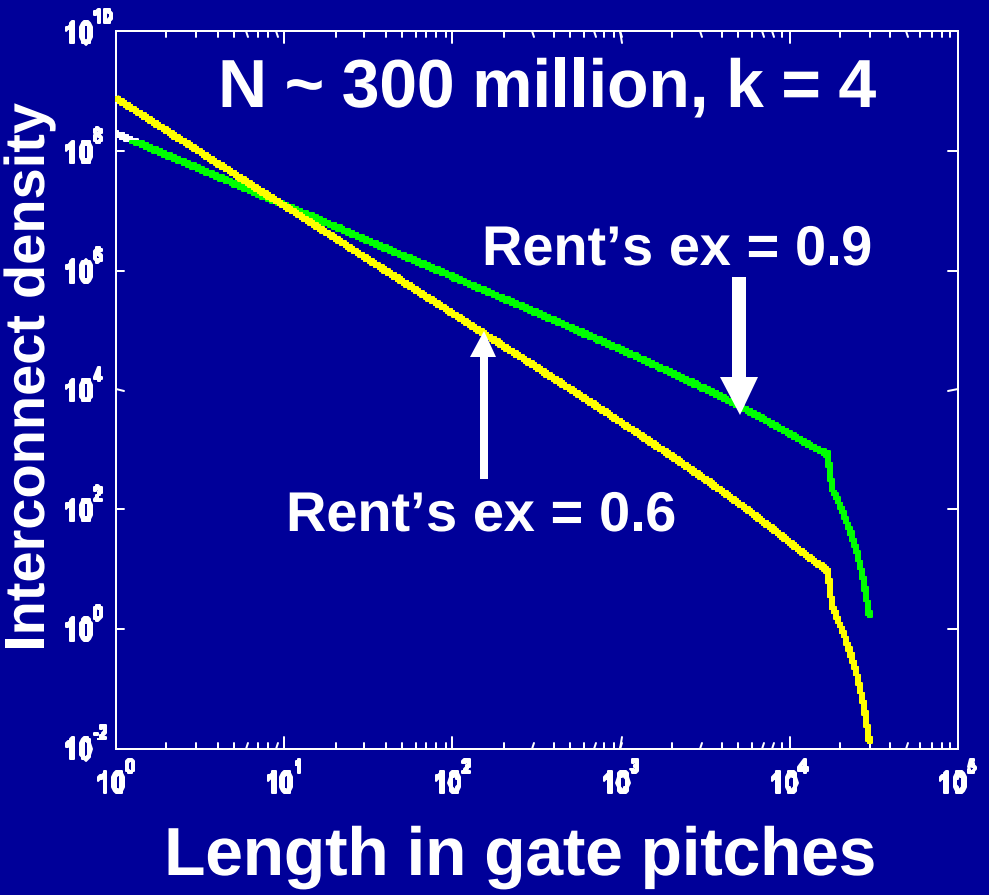
  p = Rent's exponent – represents the balance between local and global interconnects
- Rent's rule $\rightarrow$ Int. conn. length distribution

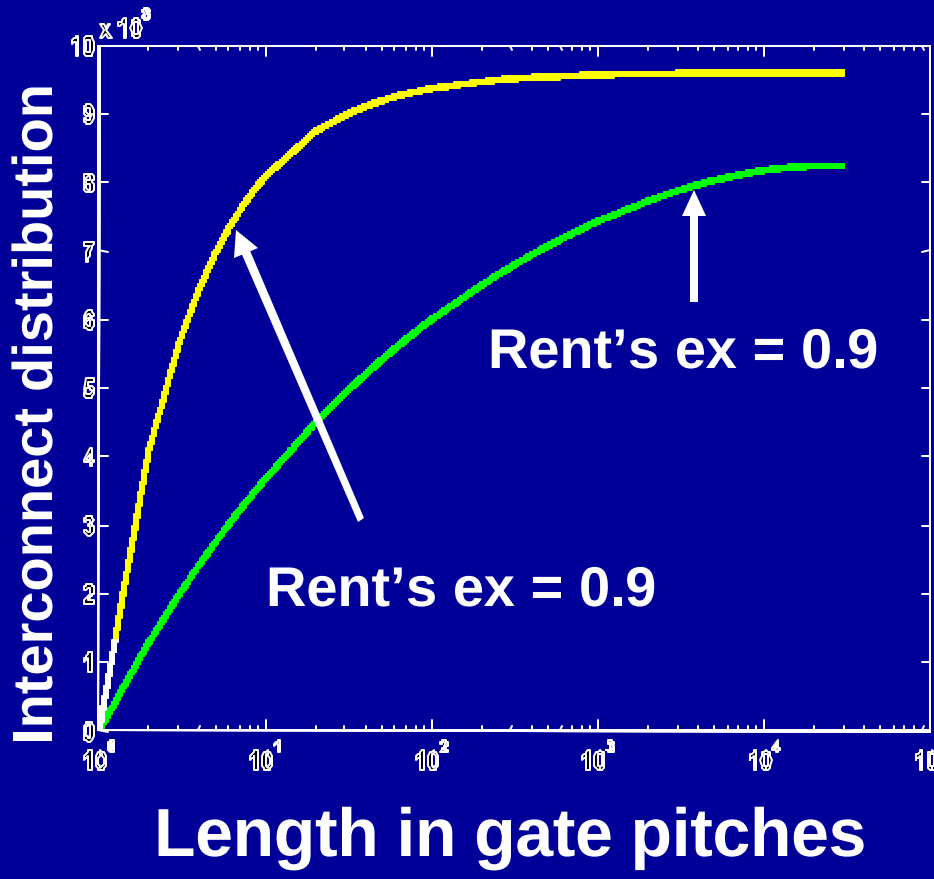
  Density = $i(l)$ = # of Int with length 'l' s.t. $a < l < b$

  Distribution = $I(l)$ = # of Int with length less than 'l'
- Wiring capacitance can be calculated from interconnect length distribution

1. **Feynman Lectures on Computation, pages 277-282**
2. **W.E. Donath, IBM J. Res. Develop. 25, 152 (1981)**
3. **J.A. Davis, et. al, IEEE TED, vol. 45, March 1998, pp:580 - 597**

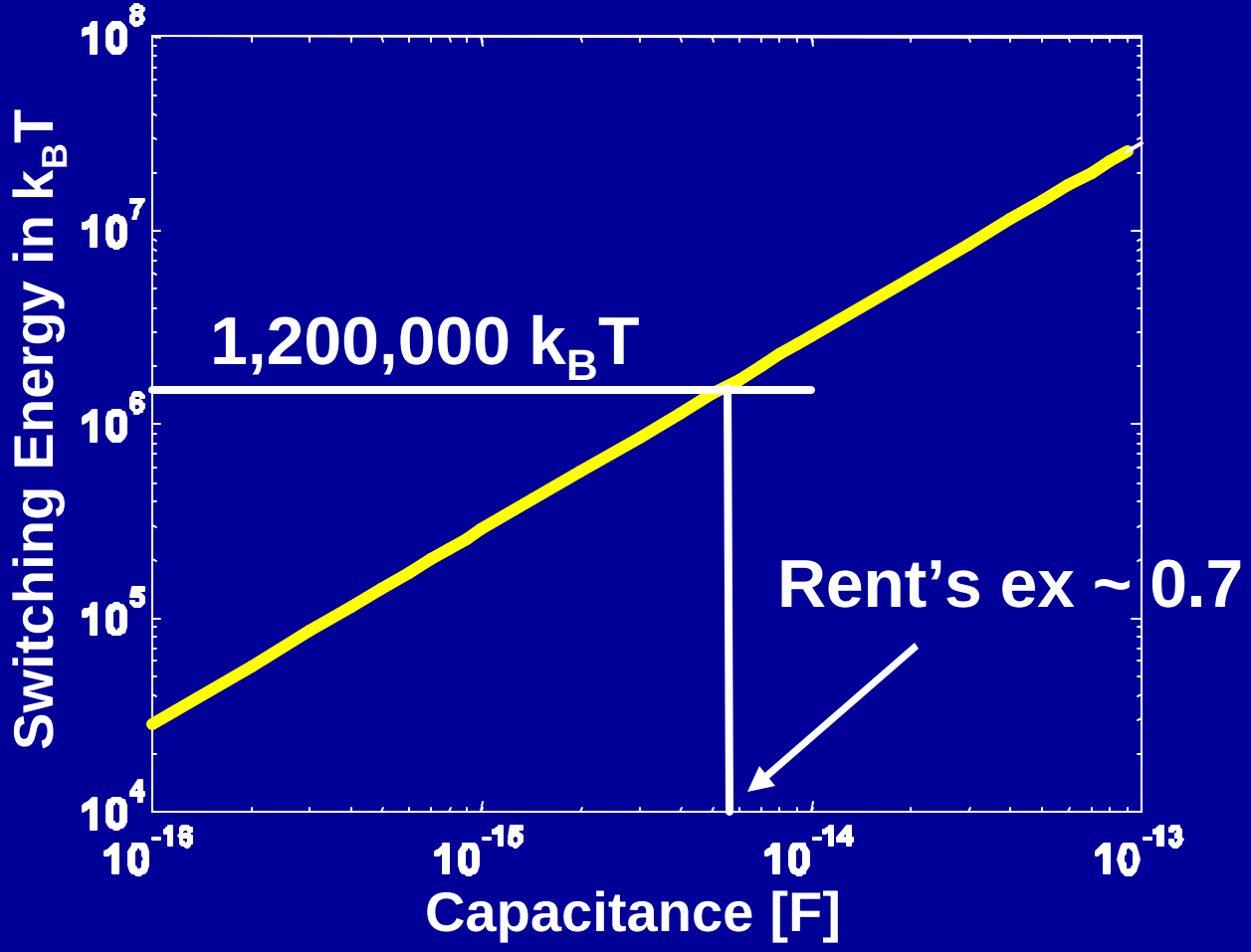
Distribution of Interconnect





A higher Rent's exponent indicates a higher number of global interconnects

Switching Energy for a System



Interconnect (or wiring) capacitance can increase the average switching energy of a gate to ~1,200,000 kBT

Practical Limits in Switching Energy in CMOS Systems

Physical Limit: $k_B T \ln(2)$

↓

Requirement for Computation: 33,000 $k_B T$

Reliability, Speed and Drivability

↓

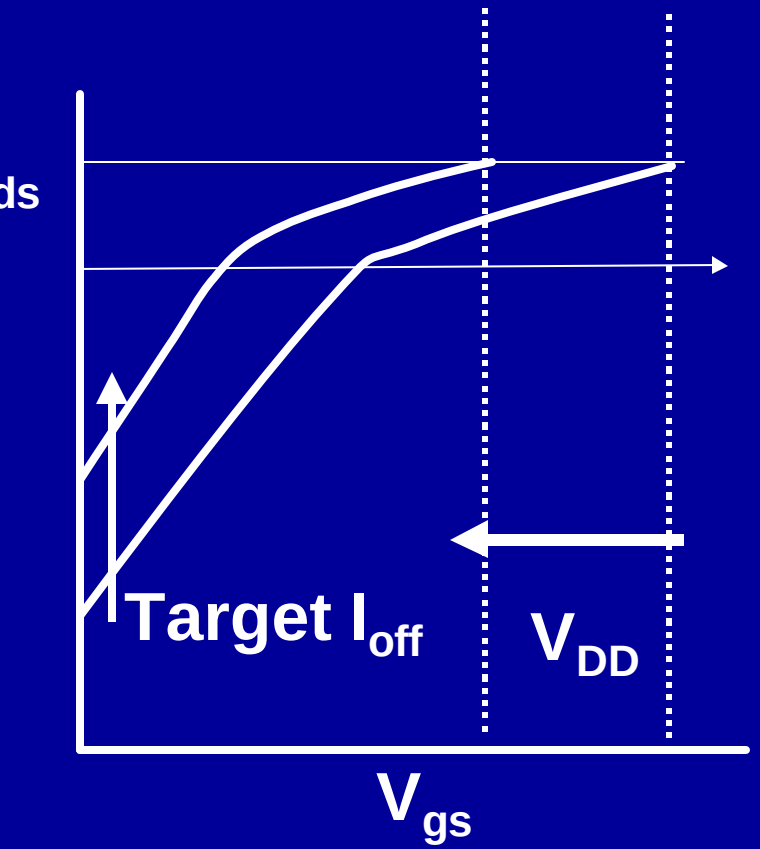
Requirement for Communication:

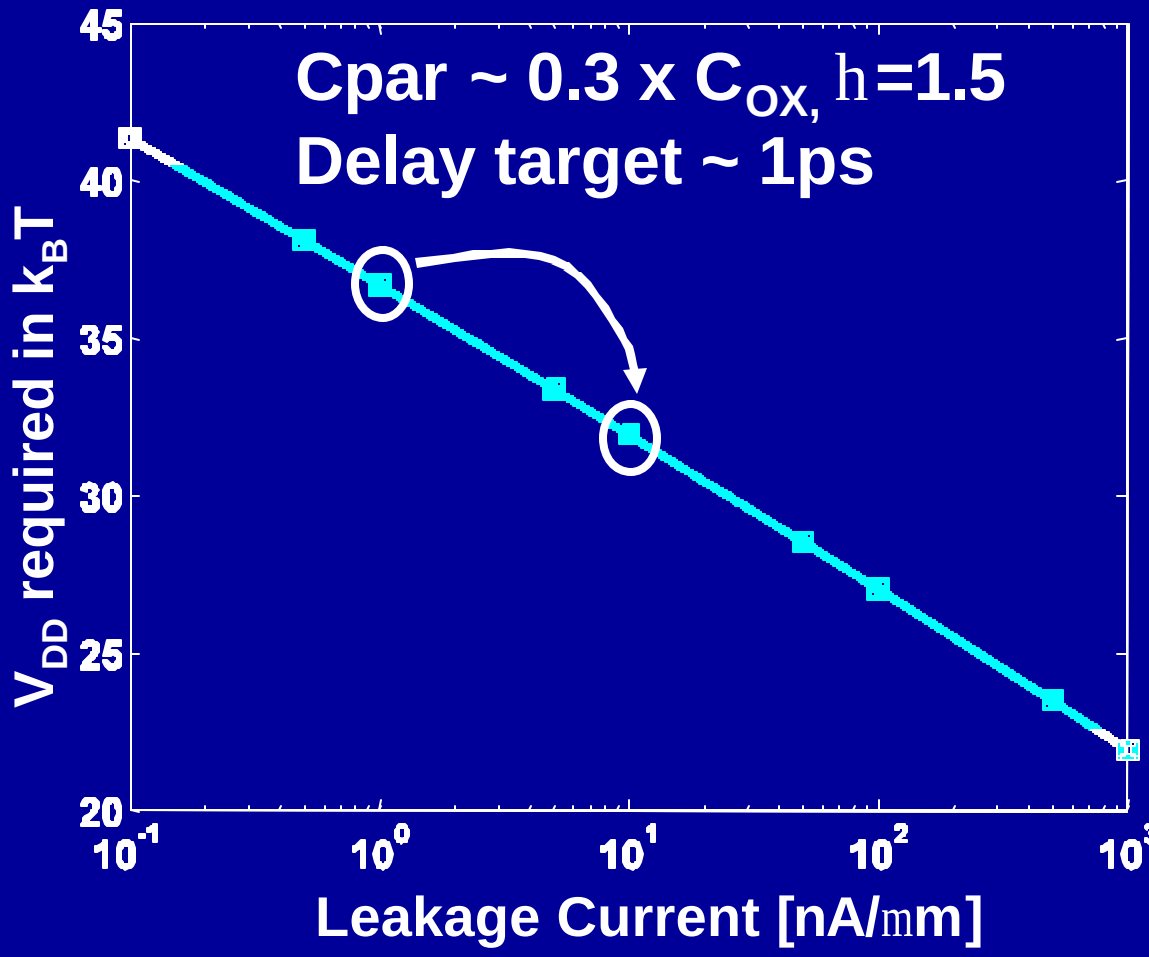
1,200,000 $k_B T$

Local and global communication

How can we reduce the practical switching energy limit?

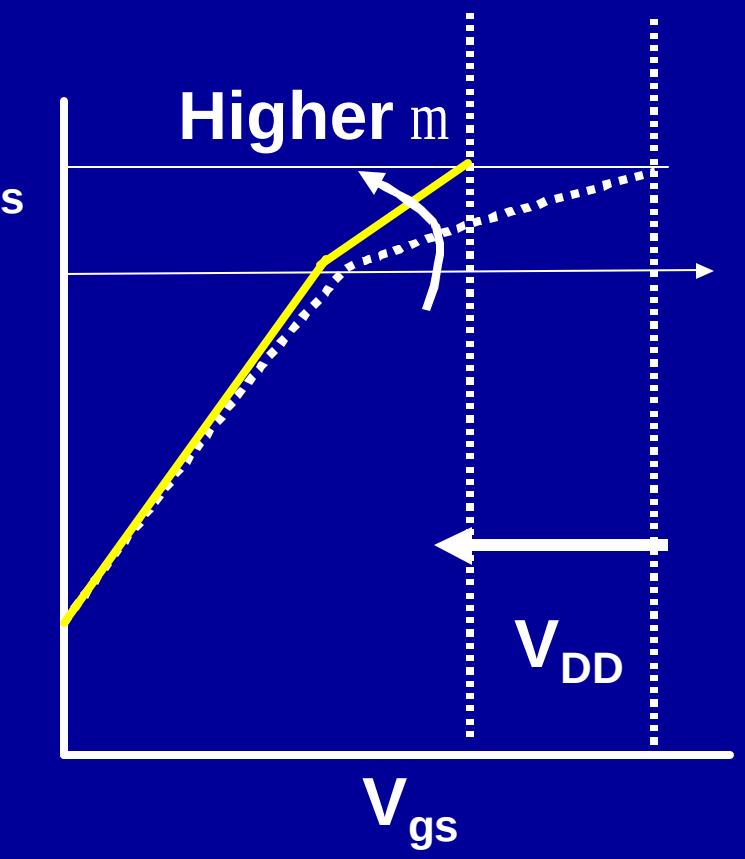
Switching Energy and Leakage Power Trade-off

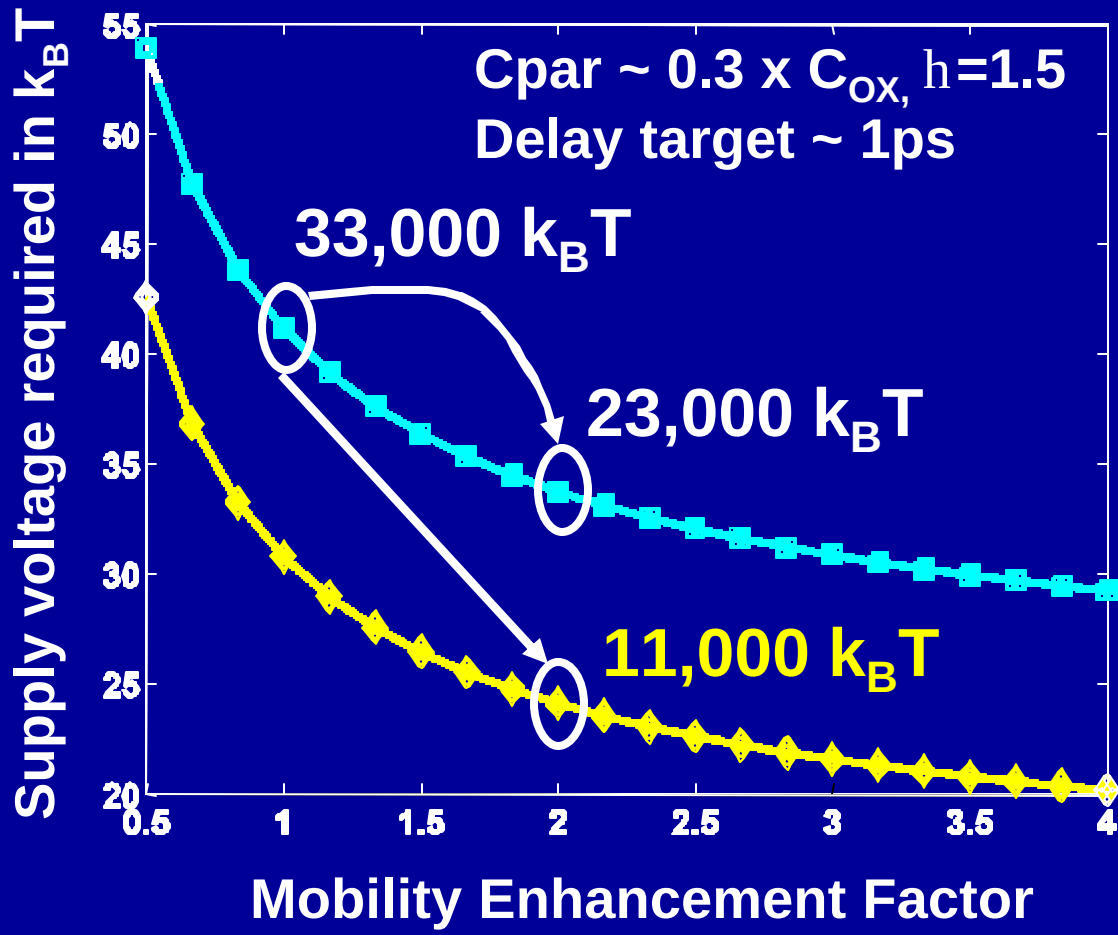




Operating at 10X higher leakage can reduce the switching energy from 33,000k_BT to 23,000 k_BT

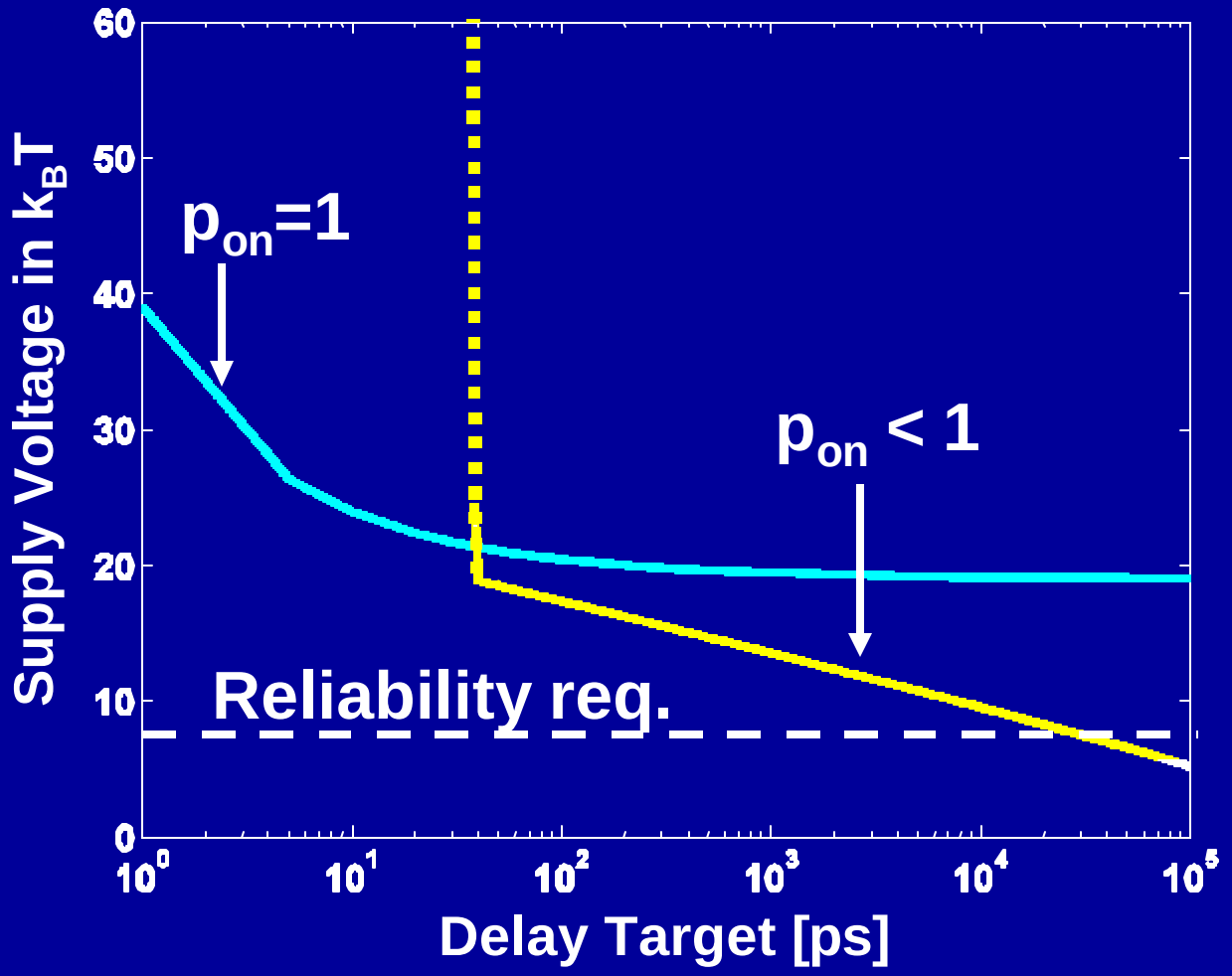
Can Higher Mobility help?





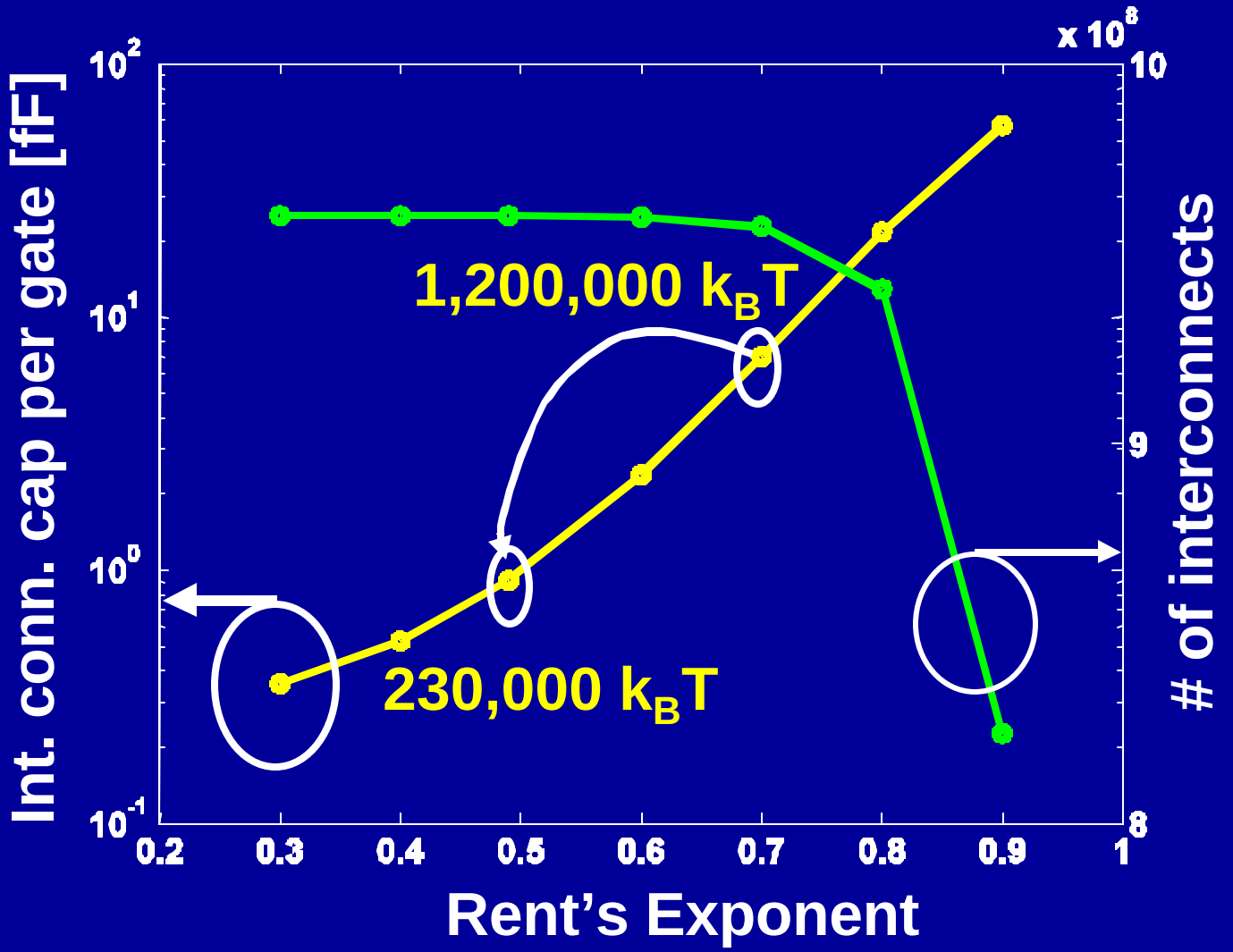
Devices with higher mobility and higher leakage target can reduce switching energy

Switching Energy and Delay Trade-off



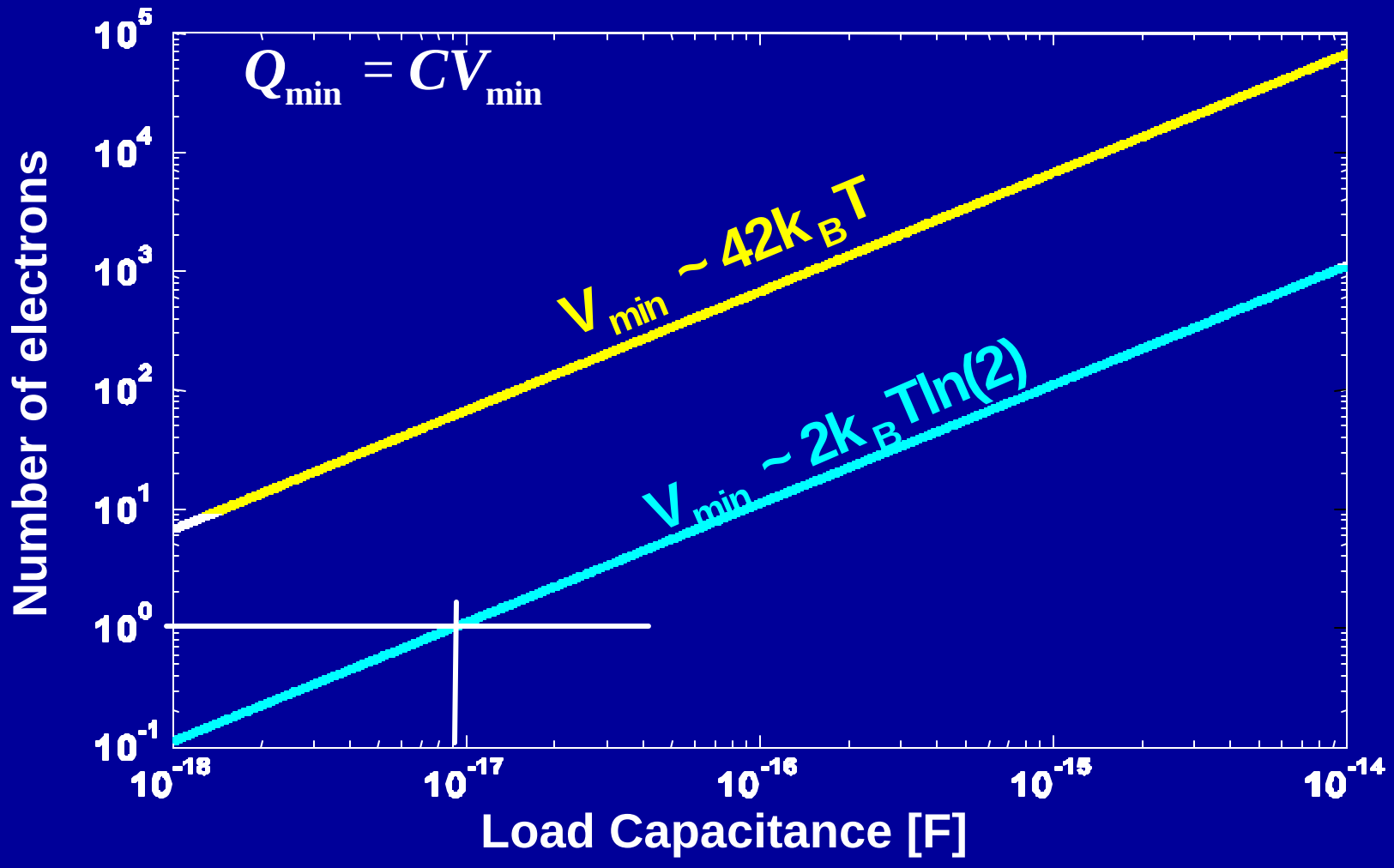
For delay targets > 100ps subthreshold operation is more energy efficient

Switching Energy for a System



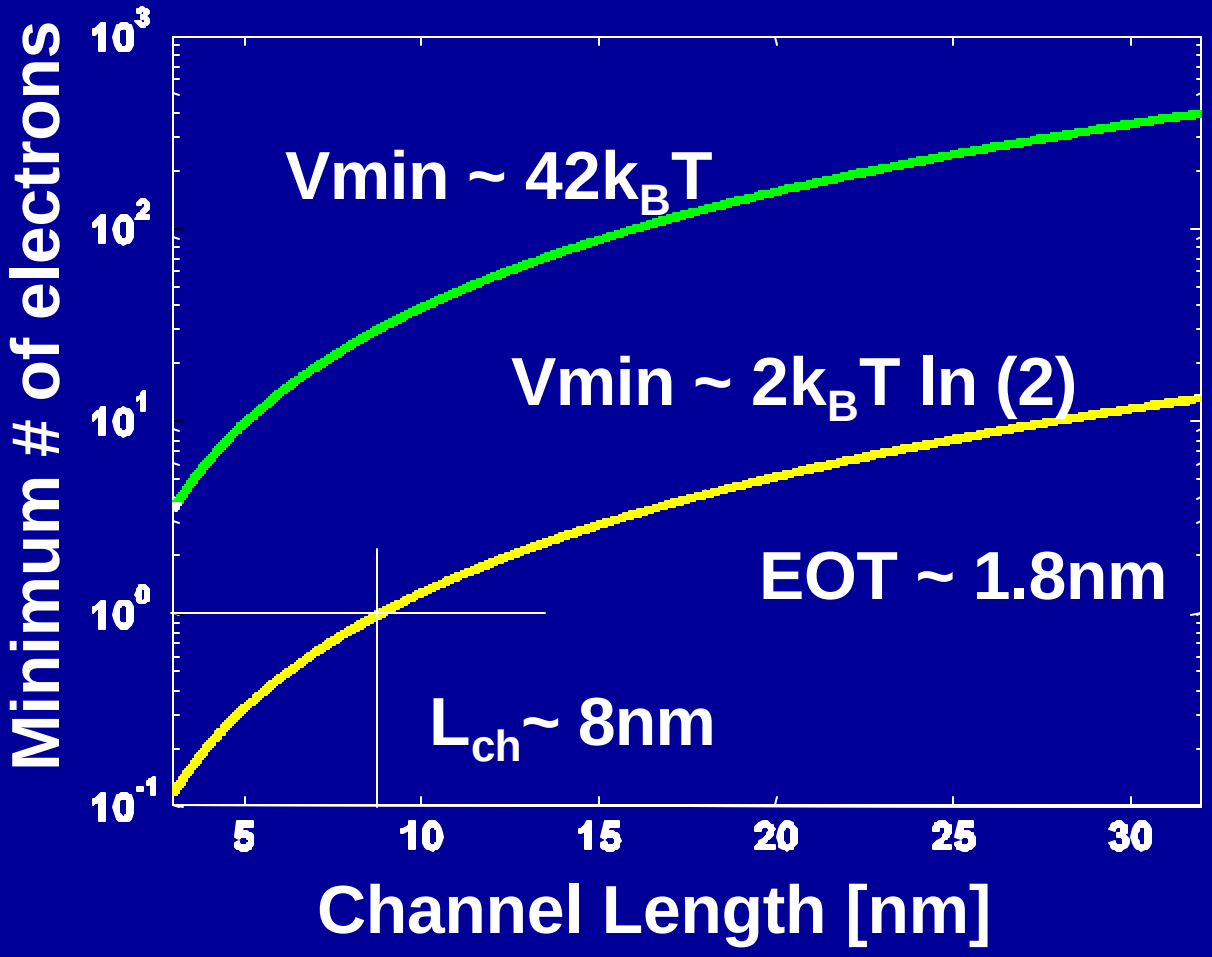
Reducing the number of local interconnects can significantly reduce the system switching energy

Single Electron Operation in CMOS



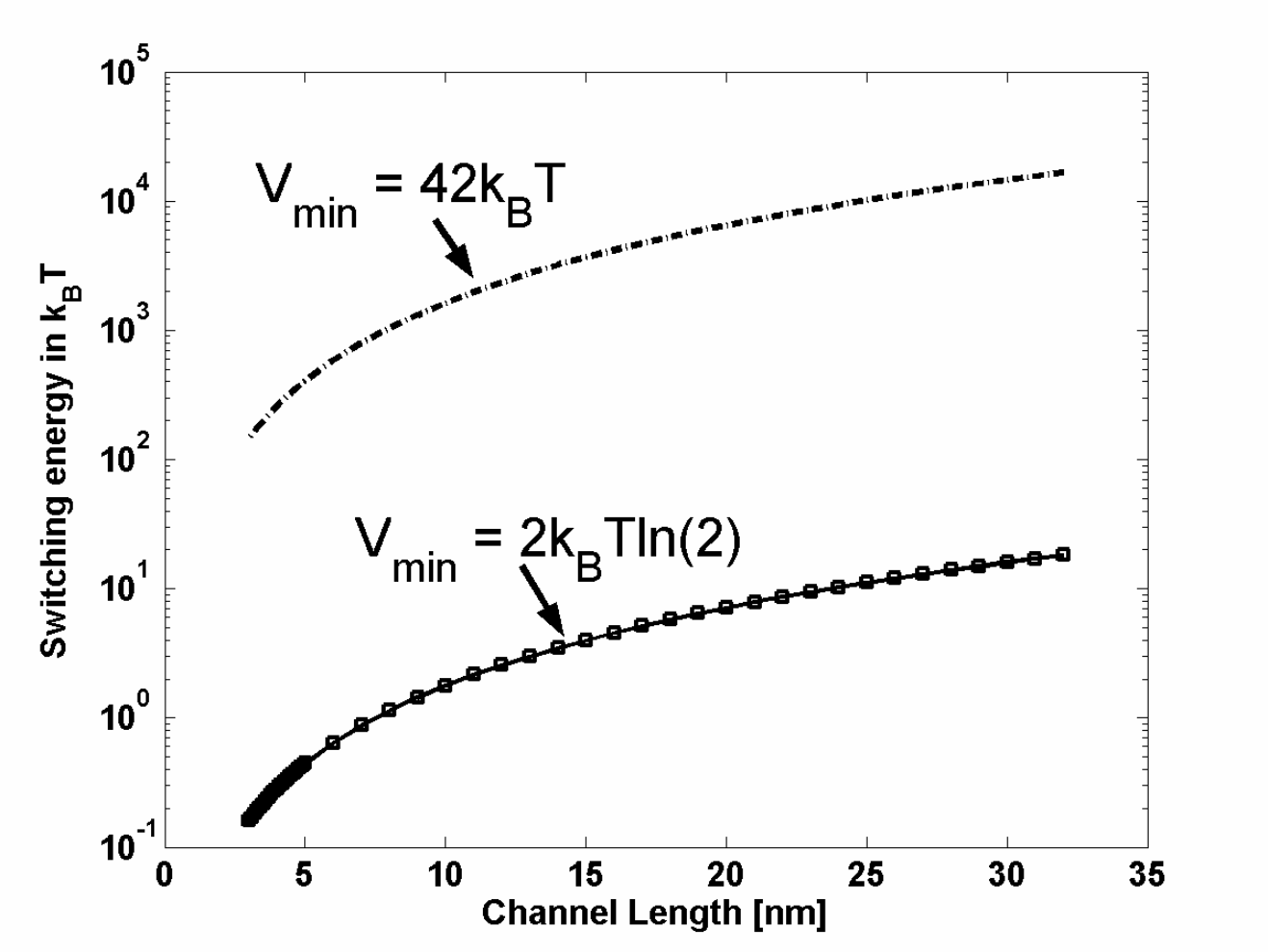
Single electron operation at room temperature is only possible if C < 9aF

Scaling and Single Electron Operation in CMOS



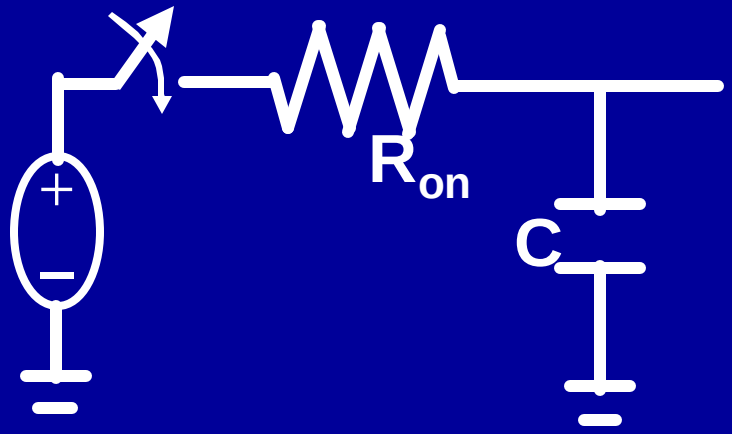
Single electron operation in CMOS logic is possible for L < 8nm

Scaling and Switching Energy



Scaling helps to reduce switching energy even if the supply voltage remains the same

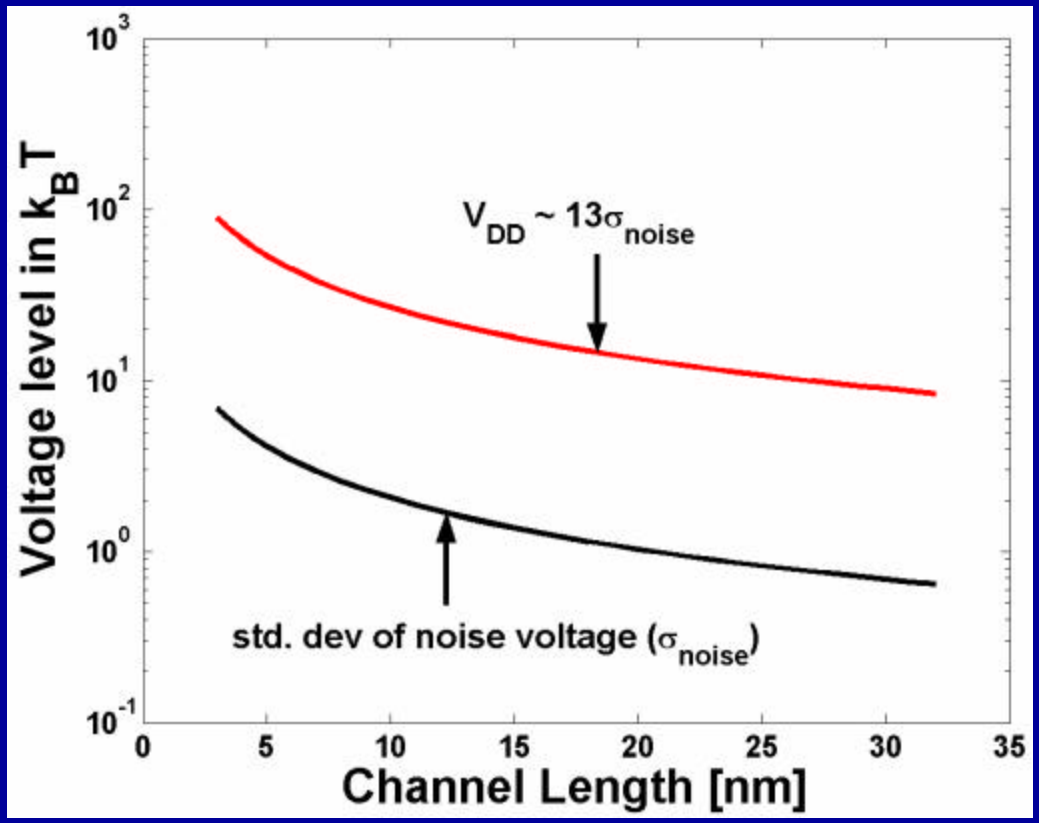
Scaling and Thermal Noise



$$V_{noise} = s_{noise} = \sqrt{\frac{4k_B TR}{RC}} = \sqrt{\frac{4k_B T}{C}}$$

$$For: C = 0.1\,fF \Rightarrow s_{noise} = 12mV$$

$$For: C = 9aF \Rightarrow s_{noise} = 43mV$$



Increase in thermal noise at lower capacitance can reduce the energy benefit of scaling

Summary

1. Can we operate with $V_{min} \sim K_B T \ln 2$?
   - Reliability
   - Delay
   - sub. slope, 2-D effects, variability etc.
2. Can we operate with $Q_{min} = q$?
   - Drivability
   - Parasitic and Interconnect capacitance

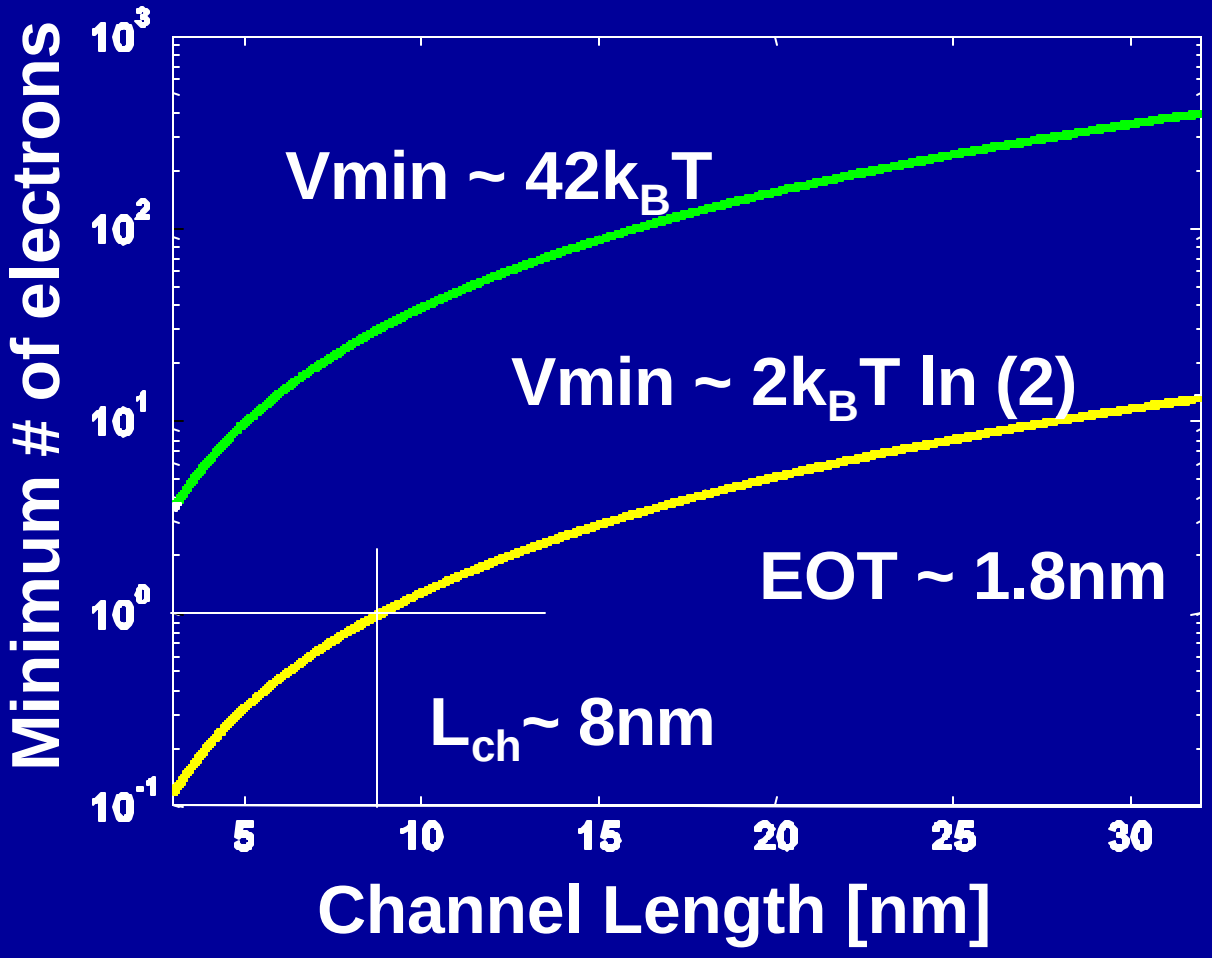
Device/Circuit/System level investigations can reduce the practical limit of switching energy, but it is very difficult to achieve the physical limit in CMOS logic

References

1. V. Zhirnov et. al, Proceedings of the IEEE, vol. 91, Nov 2003 pp. 1934 – 1939.
2. J. D. Meindl, Proceedings of the IEEE, Vol.83, April 1995, pp.:619 - 635
3. W.E. Donath, IBM J. Res. Develop. 25, 152 (1981)
4. J.A. Davis, et. al, IEEE TED, vol. 45, March 1998, pp:580 – 597 (two consecutive papers)
5. L. B. Kish, *Phys. Lett. A*, vol. 305, pp. 144–149, 2002.
6. Y. Taur and T. H. Ning, Fundamentals of Modern VLSI Devices, Cambridge University Press, 1998

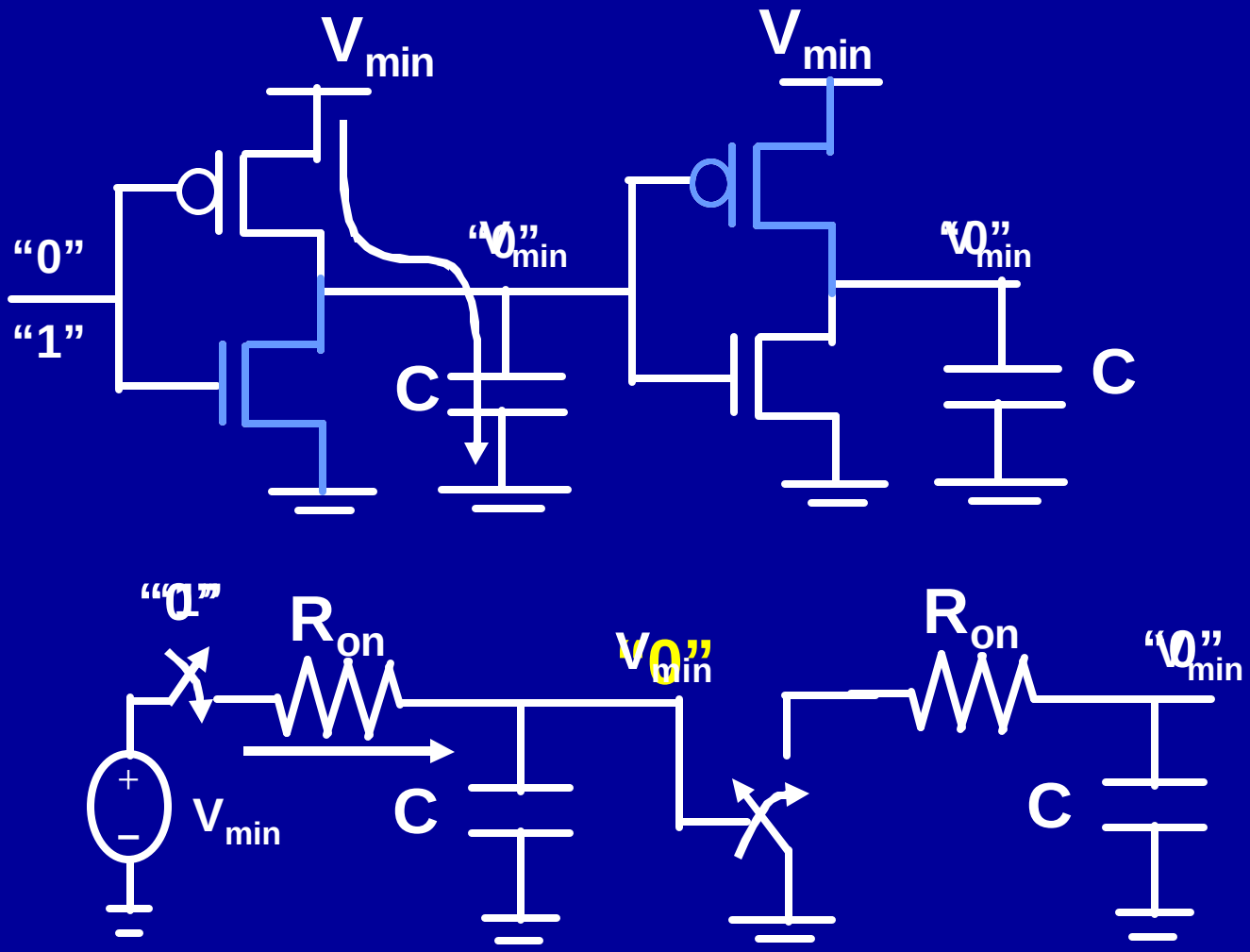
Questions and Answers

Scaling and Single Electron Operation in CMOS



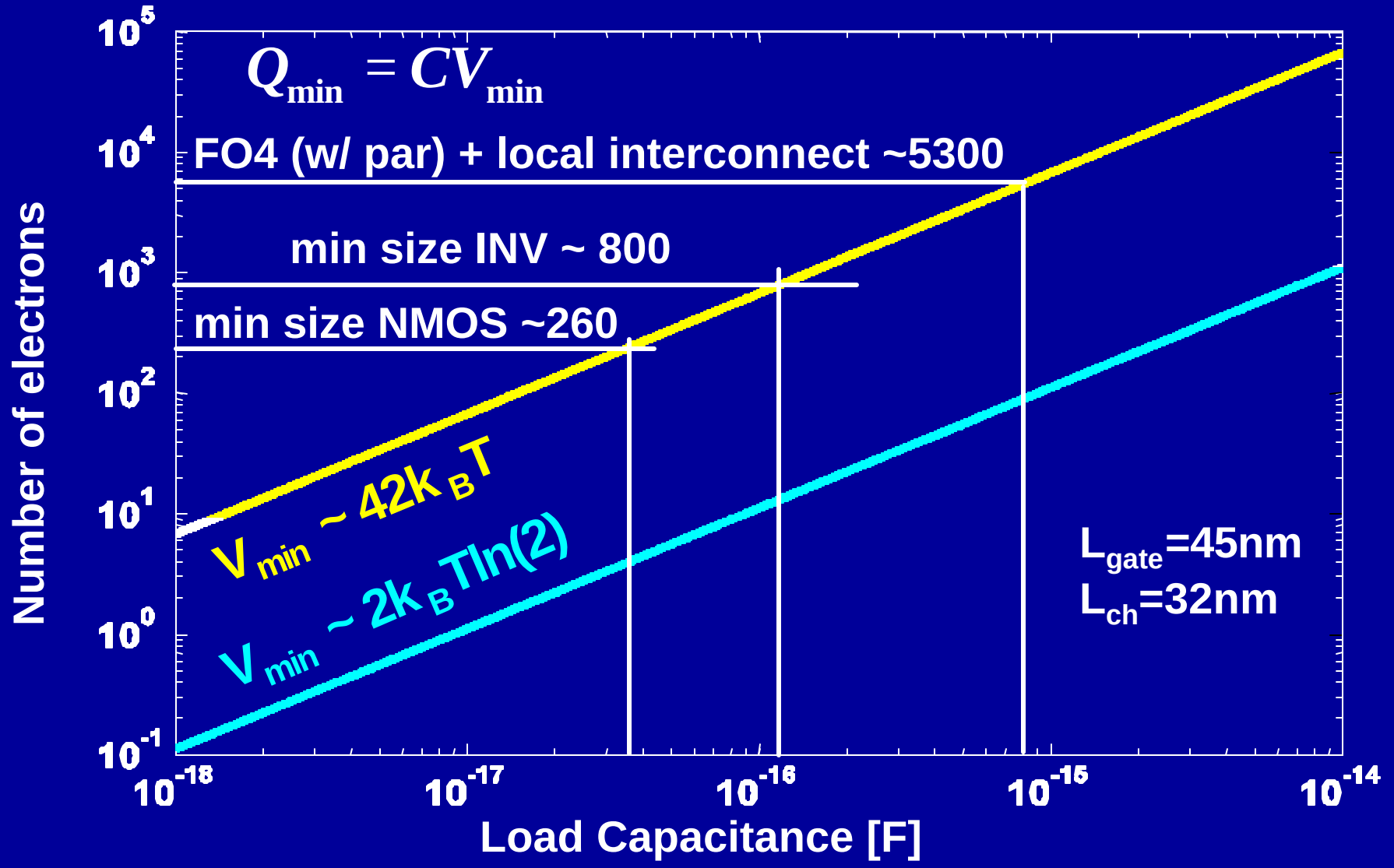
Single electron operation in CMOS logic is possible for L < 8nm

Drivability in Digital Logic



Vmin needs to be developed across a finite capacitance for driving the next gate

Drivability and Minimum Charge



Drivability requirement does not allow to operate with a single electron for CMOS logic operation