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

\[ V_{out} = \frac{Q}{C} \]
Switching Energy

Switching energy can be minimized by reducing $Q$ and/or $V_{\text{min}}$.

$$E_{\text{Total}} = \int_0^\infty i_{DD}(t)V_{\text{min}}\,dt = \int_0^{V_{DP}} CV_{\text{min}}\,dV_0 = CV_{\text{min}}^2$$

$$E_{\text{Cap}} = \int_0^\infty i_C(t)v_0(t)\,dt = \int_0^{V_{DP}} Cv_0\,dV_0 = \frac{1}{2}CV_{\text{min}}^2$$

$$E_{\text{diss}}(0 \rightarrow 1) = E_{\text{Total}} - E_{\text{Cap}} = \frac{1}{2}CV_{\text{min}}^2$$

$$E_{\text{diss}} = CV_{\text{min}}^2 = QV_{\text{min}}$$
Physical Medium for Computation: Barrier Model

\[ V = V_{\text{min}} = E_{\text{bmin}} \]

\[ V = 0 \]

\[ V_g \]

\[ V_d \]
Minimum Barrier Height: Zhirnov’s Model

Channel Length ($L_{ch}$) [nm] vs. Minimum barrier height ($E_b$) in $k_B T$

- $P_{err} = P_{err\_cl} + P_{err\_QM}$
- $-P_{err\_cl}P_{err\_QM}$
- For $L_{ch} > 10\text{nm}$

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

Minimum barrier height = $E_{b_{min}} \sim k_B T \ln(2)$
Minimum Operating Voltage and Switching Energy

- Minimum operating voltage
  \[ V_{\text{min}} \sim k_B T \ln(2) \]
- Minimum switching energy
  \[ E_{\text{diss}} = q V_{\text{min}} = q k_B T \ln(2) \sim 0.7 k_B T \]

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 $V_{\text{min}} \sim K_B T \ln 2$?

2. Can we operate with $Q_{\text{min}} = q$?
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
# of devices = \( N_{\text{dev}} \)

Prob. of error of a single gate = \( P_{\text{err}} \)

Prob. of error of the circuit = \( P_{\text{circ}} = 1 - (1 - P_{\text{err}})^{N_{\text{dev}}} \)

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
  - \( P_{\text{err}} = 0.5 \)
    \[ \Rightarrow E_b = 0.7k_B T \]
  - \( P_{\text{err}} = 5 \times 10^{-12} \)
    \[ \Rightarrow E_b = 25k_B T \]
- 0.1% failure rate for a circuit of 300 million devices \( \Rightarrow V_{\text{min}} \sim 25k_B T \)

\[ k_B T \ln(2) \quad \text{Reliability} \quad 25k_B T \]
CMOS logic operates based on presence or absence of charge and not on localization of charge.
Operation of MOS Device

\[ V = V_{DD} = E_{b\text{OFF}} - E_{b\text{ON}} \]

\[ V = 0 \]

\[ V_{DD} = \eta \frac{k_B T}{q} \ln \left( \frac{p_{on}}{p_{off}} \right) \]

Operation with a larger \( p_{on}/p_{off} \) requires a higher supply voltage
Operation of CMOS Logic

\[ V_{in} = 0.5\eta \frac{k_B T}{q} \times \ln \left( \frac{p_{on}}{p_{off}} \right) + 0.5\eta \frac{k_B T}{q} \times \ln \left( \frac{1 - \exp \left( -q \frac{V_{DD} - V_0}{k_B T} \right)}{1 - \exp \left( -q V_0 / k_B T \right)} \right) \]
Higher $p_{on}/p_{off}$ improves maximum gain and noise margin.
Operation of CMOS Logic

2n+1 stages

Vin = \( V_{DD}/2 - \Delta \)

Vo(1) = Vin(2) = \( V_{DD}/2 + \Delta A_v \)

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

if \( A_v < 1 \), as \( n \rightarrow \infty \), \( V_O \rightarrow V_{DD}/2 \)

if \( A_v > 1 \), as \( n \rightarrow \infty \), \( V_O \rightarrow V_{OH} \)
**Operation of CMOS Logic**

**distinguishability**

\[ \Rightarrow \text{Gain (} A_V \text{)} > 1 \]

for CMOS inverter

Minimum \( p_{on}/p_{off} \) is “4” and not “2”

\[ V_{min} = k_B T \ln(2) \]

**Device to Inverter**

\[ V_{min} = 2k_B T \ln(2) \]
To prevent spontaneous change of state noise margin needs to be at least higher than $k_B T$.

$$=> V_{DD} > 3k_B T$$
Reliability of Circuit Operation

\[ \text{# of gates} = N_{\text{gate}} \]

\[ \text{Prob. of error of a single gate} = P_{\text{err}} \]

\[ \text{Prob. of error of the circuit} = P_{\text{circ}} = 1 - (1-P_{\text{err}})^{N_{\text{dev}}} \]

Reliable operation of the circuit imposes stronger constraint on the reliability of the gate operation
Higher noise requires a larger noise margin for reliable operation.
Reliability of CMOS Inverter Operation

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

$V_{\text{min}} = 10k_B T$
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

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

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

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

$E_{bOFF} \sim 11k_B T$ helps to meet a leakage requirement of $1\text{nA/\mu m}$
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

\[ V_{in} \rightarrow V_{DD} \rightarrow V_{o} \]

\[ V_{min} \rightarrow R_{on} \rightarrow C \]

Voltage vs. Time

"0"

"1"

Voltage vs. Time

"0"

"1"
• Delay through an RC circuit
  – Independent of applied voltage $V_{\text{min}}$
  – 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.

$C_{gate}V_{DD} = \tau I_{on}$

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

$$
\left(1 + \frac{C_{par}}{C_{ox}}\right)3L_{min}^2 C_{ox} V_{DD} = \tau \mu_{eff} \frac{L_{min}}{2L_{min}} C_{ox} \left( V_{DD} - \eta \frac{E_{bOFF}}{q} \right)^2
$$
Impact of Delay on Minimum $V_{DD}$

- $\mu_{\text{neff}} \sim 300 \text{ cm}^2/\text{V-sec}$
- $C_{\text{par}} \sim 30\% \times C_{\text{gate}}$
- No parasitics

Supply Voltage in $k_B T$

Delay Target [ps]

$V_{\text{min}} = 10k_B T$

Delay (1ps)

$V_{\text{min}} = 28k_B T$
Non-ideal subthreshold slope

\[ V_{DD} = \eta \frac{k_B T}{q} \ln \left( \frac{p_{on}}{p_{off}} \right) \]

A larger subthreshold slope requires a higher \( V_{DD} \) to achieve a \( \text{pon/loff} \)
Non-ideal subthreshold slope

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

$\eta = 1.5$

$\eta = 1.0$

$37k_B T$
2-D Electrostatics

Degraded Sub-slope

Drain Induced Barrier Lowering

NMOS in linear

Delay increase due to SCE & DIBL
Under same leakage power 2-D effect increases the $V_{DD}$ required to achieve a target delay

$V_{DD} \sim 39k_B T$
Process Variability

- Leakage $\sim p_{\text{off}}$ variation
- Reliability $\sim p_{\text{on}}/p_{\text{off}}$ variation
- Delay $\sim$ 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 $E_{bOFF}$ $\Rightarrow V_{DD} \sim 42k_B T$
Why We are using $V_{DD}$ much larger than the $k_B T \ln(2)$ limit?

$\frac{k_B T}{q} \ln(2)$  
CMOS Logic  
Distinguish-ability

$2 \frac{k_B T}{q} \ln(2)$  
Reliability  
Noise-tolerance

$42 \left( \frac{k_B T}{q} \right)$
Subth. Slope, 2-D effect, Process variation, etc.
Non-idealities

$28 \left( \frac{k_B T}{q} \right)$

$10 \left( \frac{k_B T}{q} \right)$

Delay (1ps)
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.

Load Capacitance [F]

Number of electrons

\[ Q_{\text{min}} = CV_{\text{min}} \]

FO4 (w/ par) + local interconnect \( \sim 5300 \)

min size INV \( \sim 800 \)

min size NMOS \( \sim 260 \)

\[ V_{\text{min}} \sim 42k_B T \]

\[ V_{\text{min}} \sim 2k_B T \ln(2) \]

\[ L_{\text{gate}} = 45\text{nm} \]

\[ L_{\text{ch}} = 32\text{nm} \]
Drivability and Switching Energy

Drivability requirement increases the minimum switching energy for an inverter to $\sim 33,000 \, k_B T$

- FO4 (w/ par) + local interconnect $\sim 220,000$
- min size INV $\sim 33000$
- min size NMOS $\sim 11000$
- $V_{min} \sim 42k_B T$
- $V_{min} \sim 2k_B T \ln(2)$

$L_{gate} = 45\text{nm}$
$L_{ch} = 32\text{nm}$
Switching Energy in CMOS Logic

Delay ~ 1ps, High reliability

$k_B T \ln(2)$ Delay/Reliability

$42k_B T$

$33000k_B T$

Drivability
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

- Delay/Reliability: $k_B T \ln(2)$
- Drivability: $42k_B T$
- FO4: $220,000k_B T$
- $33000k_B T$
Driving "long" interconnects can significantly increase the switching energy
Interconnect of length $\sim 400 \mu m$ has 100 fF of cap which requires $\sim 28,000,000 \, k_B T$ 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 → Int. conn. length distribution

\[ \text{Density} = i(l) = \# \text{ of Int with length ‘}l\text{‘ s.t. } a < l < b \]

\[ \text{Distribution} = I(l) = \# \text{ of Int with length less than ‘}l\text{‘} \]

• Wiring capacitance can be calculated from interconnect length distribution

1. Feynman Lectures on Computation, pages 277-282
A higher Rent’s exponent indicates a higher number of global interconnects.
Interconnect (or wiring) capacitance can increase the average switching energy of a gate to $\sim 1,200,000 \text{ k}_B\text{T}$.
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?
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.
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 $C < 9\text{aF}$

$$Q_{\text{min}} = CV_{\text{min}}$$

$$V_{\text{min}} \sim 42k_B T$$

$$V_{\text{min}} \sim 2k_B T \ln(2)$$
Scaling and Single Electron Operation in CMOS

Single electron operation in CMOS logic is possible for $L < 8\text{nm}$.
Scaling helps to reduce switching energy even if the supply voltage remains the same.
Scaling and Thermal Noise

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

1. Can we operate with $V_{\text{min}} \sim K_B T \ln 2$ ?
   - Reliability
   - Delay
   - sub. slope, 2-D effects, variability etc.

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

Questions and Answers
Single electron operation in CMOS logic is possible for $L < 8\text{nm}$.

Scaling and Single Electron Operation in CMOS

- $V_{\text{min}} \sim 2k_B T \ln (2)$
- $V_{\text{min}} \sim 42k_B T$
- $EOT \sim 1.8\text{nm}$
- $L_{\text{ch}} \sim 8\text{nm}$

Minimum # of electrons vs. Channel Length [nm]
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.