CMOS VLSI and Aspects of ASIC Design

PRESENTED BY:

Dr. Shasanka Sekhar Rout

SYLLABUS

Session 1
Lecture: Metal Oxide Semiconductors (MOS)
- Overview of CMOS Module
- Introduction of MOS devices
- n-MOS, p-MOS and CMOS
- MOS operation

Session 2
Lecture: CMOS Gates and Characteristics
- CMOS Logic Circuits
- CMOS Inverter and I-V Characters

Session 3
Lecture: Design of Complex Functions
- Design of Complex Functions using CMOS

Session 4
Lecture: MOS Cells
- Structure of MOS cells
- Threshold Voltage and DC transfer characteristics
- Introduction to IC Design Cycle and layout

Session 5
Lecture: CMOS functioning and parameters
- CMOS as a switch, estimation of parasitic values
- rise and fall times
- power dissipation
    - ❖ Dynamic and Static
    - ❖ Low power architecture

Session 6
Lecture: Technology Scaling
- ✓ Device sizing, rationed and non-rationed logic
- ✓ Concepts of Interconnect
- ✓ Transistor Scaling and interconnect Scaling

Session 7
Lecture: CMOS Logic Circuits
- ✓ Design of logic circuit, fabrication steps
- ✓ CMOS Combinational Logic circuits
- ✓ Design of NAND and NOR Gates
- ✓ Transmission Gates

Session 8
Lecture: IC Fabrication using CMOS Technology
- ✓ CMOS Technology, wafer formation, photolithography.
- ✓ Well and channel formation, Silicon dioxide and gate oxide.

Session 9
Lecture: Memory using CMOS
- ✓ Design of Complex circuits
- ✓ SRAM, DRAM & current Mirror
- ✓ Memory cell and their read write operation.

Session 10
Lecture: Advanced Trends in Backend VLSI
- ✓ Overview of CMOS Fabrication Steps
- ✓ Sequential CMOS logic
- ✓ Circuit Design of Latches and Flip-Flops
- ✓ Introduction to FinFET technology

Session 11
Lecture: CMOS Design Issues
- ✓ Design issue like Delays, Cross talk , Velocity saturation,
- ✓ Body effect, latch up, electro migration, hot carrier and Channel Length modulation

Session 12
Lecture: Aspects of ASIC Design
- ❖ Introduction of ASIC Design,
- ❖ An overview of Backend VLSI Design Flow – Libraries, Floorplanning, Placement, Routing, Verification, Testing.
- ❖ Flow Diagram, Specifications and Schematic cell Design.
- ❖ Spice simulation. Analysis of analog and digital circuits, Circuits elements, AC and DC analysis.

Session 13
Lecture: Characteristics of Different Technologies
- ❖ Transfer Characteristics, Transient responses, Noise analysis of current and voltage based on different technologies.

Session 14
Lecture: Back-end Processes
- ❖ Fabrication methods of circuit elements, Layout design of different cells.
- ❖ Library cell designing, NAND, NOR, NOT, X-OR etc.
- ❖ Circuit Extraction, Electrical rule check, LVS, Post layout Simulation and parasitic extraction.

Session 15 & 16
Lecture: Back-end Designing and Analysis
- ❖ Design format, timing analysis, back annotation and post layout simulation of silicon circuit
- ❖ Study of design Issues Antenna effect, Electro migration effect body effect, Inductive and capacitive cross talk, Drain punch through, etc.
- ❖ Placement of cells, placing of I/O blocks, Initialization of floor planning, routing and creating design data base
- ❖ DFT Guidelines, Test Pattern and BIST
- ❖ ASIC design implementation

Session 3 & Session 7

Lecture: Design of Complex Functions

- Design of Complex Functions using CMOS

Lecture: CMOS Logic Circuits

✓ Design of logic circuit, fabrication steps

✓ CMOS Combinational Logic circuits

✓ Design of NAND and NOR Gates

✓ Transmission Gates

CMOS Logic Circuits

- CMOS logic circuits are widely used in digital electronics due to their high efficiency, low power consumption, and scalability.
- These circuits use both NMOS and PMOS transistors to implement various logic functions.

Fabrication Steps:

- ✓ Starting Material Preparation (Wafer Production)
- ✓ Oxidation
- ✓ Photolithography
- ✓ Etching
- ✓ Doping
- ✓ Gate Formation
- ✓ Source/Drain Creation
- ✓ Interconnect Formation
- ✓ Multilevel Metallization
- ✓ Passivation
- ✓ Packaging

CMOS Combinational Logic circuits

- Combinational logic circuits, or gates, which perform Boolean operations on multiple input variables and determine the outputs as Boolean functions of the inputs, are the basic building blocks of all digital systems.

CMOS Combinational Logic Circuits

Design of CMOS NOR2 (Two-Input NOR) Gate:

V_A	V_B	V_{out}
low	low	high
low	high	low
high	low	low
high	high	lower

A CMOS NOR2 gate and its complementary operation: Either the nMOS network is on and the pMOS network is off, or the pMOS network is on and the nMOS network is off

Layout of the CMOS NOR2 gate:

Design of CMOS NAND2 (Two-Input NAND) Gate:

V_A	V_B	V_{out}
low	low	high
low	high	high
high	low	high
high	high	low

Layout of the CMOS NAND2 gate:

Complex CMOS Logic Gates: Design (A(D+E)+BC)'

A complex CMOS logic gate realizing the Boolean function (A(D+E)+BC)'

Layout of Complex CMOS Logic Gates:

- If we can minimize the number of diffusion-area breaks both for nMOS and for pMOS transistors, the separation between the polysilicon gate columns can be made smaller, which will reduce the overall horizontal dimension and, hence, the circuit layout area. The number of diffusion breaks can be minimized by changing the ordering of the polysilicon columns.

Stick-diagram layout of the complex CMOS logic gate, with an arbitrary ordering of the polysilicon gate columns

Layout of Complex CMOS Logic Gates:

- A simple method for finding the optimum gate ordering is the Euler-path approach: find an Euler path in the pull-down graph and a Euler path in the pull-up graph with identical ordering of input labels, i.e., find a common Euler path for both graphs. The Euler path is defined as an uninterrupted path that traverses each edge (branch) of the graph exactly once.

Finding a common Euler path in both graphs for n-net and p-net provides a gate ordering that minimizes the number of diffusion breaks and, thus, minimizes the logic-gate layout area. In both cases, the Euler path starts at (x) and ends at (y)

Layout of Complex CMOS Logic Gates:

Optimized stick-diagram layout of the complex CMOS logic gate

CMOS Implementation of the XOR Function:

AND-OR-INVERT (AOI) and OR-AND-INVERT (OAI) gates:

- The pull-down net of the AOI gate consists of parallel branches of series-connected nMOS driver transistors. The corresponding p-type pull-up network can simply be found using the dual-graph concept.

An AND-OR-INVERT (AOI) gate and the corresponding pull-down net

AND-OR-INVERT (AOI) and OR-AND-INVERT (OAI) gates:

- The pull-down net of the OAI gate consists of series branches of parallel-connected nMOS driver transistors, while the corresponding p-type pull-up network can be found using the dual-graph concept.

An OR-AND-INVERT (OAI) gate and the corresponding pull-down net

Design of CMOS Full-Adder Circuit:

- The sum-out and carry-out signals of the full adder are defined as the following two combinational Boolean functions of the three input variables: *A, B, and C*.

$$\begin{aligned} \text{sum_out} &= A \oplus B \oplus C \\ &= ABC + A\overline{B}\,\overline{C} + \overline{A}\,\overline{B}C + \overline{A}\,\overline{C}B \\ \text{carry_out} &= AB + AC + BC \end{aligned}$$

Gate-level schematic of the one-bit full-adder circuit

Design of CMOS Full-Adder Circuit:

Transistor-level schematic of the one-bit full-adder circuit with total of 14 nMOS and 14 pMOS transistors

Pseudo-nMOS Gates:

- One possible approach to reduce the number of transistors is to use a single pMOS transistor, with its gate terminal connected to ground. With this simple pull-up arrangement, the complex gate can be implemented with much fewer transistors.

The pseudo-nMOS implementation of the OAI gate

CMOS Transmission Gates (TG) (Pass Gates):

- ✓ The CMOS transmission gate consists of one nMOS and one pMOS transistor, connected in parallel. The gate voltages applied to these two transistors are also set to be complementary signals. As such, the CMOS TG operates as a bidirectional switch between the nodes A and B which is controlled by signal C.

- ✓ If the control signal C is logic-high, i.e., equal to V_{DD}, then both transistors are turned on and provide a low-resistance current path between the nodes A and B.

- ✓ If, on the other hand, the control signal C is low, then both transistors will be off, and the path between the nodes A and B will be an open circuit. This condition is also called the high-impedance state.

CMOS Transmission Gates (TG)

Four different representations of the CMOS transmission gate (TG)

CMOS Transmission Gates (TG)

- The implementation of CMOS transmission gates in logic circuit design usually results in compact circuit structures which may even require a smaller number of transistors than their standard CMOS counterparts.
- The operation of the **multiplexer** can be understood quite easily: If the control input S is logic-high, then the bottom TG will conduct and the output will be equal to the input B.
- If the control signal is low, the bottom TG will turn off and the top TG will connect the input A to the output node.

Two-input multiplexer circuit implemented using two CMOS TGs

CMOS Transmission Gates (TG)

Eight-transistor CMOS TG implementation of the XOR function

CMOS Transmission Gates (TG)

Six-transistor CMOS TG implementation of the XOR function

Session 5

Lecture: CMOS Functioning and Parameters

- ❖ CMOS as a switch, estimation of parasitic values
- ❖ Rise and fall times
- ❖ Power dissipation
  - ✓ Dynamic and static
  - ✓ Low power architecture

CMOS Functioning and Parameters

CMOS as A Switch:

- In CMOS technology, transistors can function as electronic switches to switch between on and off states allows CMOS transistors to implement logic gates, memory cells, and other digital circuit elements.
- The CMOS inverter is a fundamental building block that demonstrates the switching behavior of NMOS and PMOS transistors.

  NMOS Transistor: Connected between the output and ground.

  PMOS Transistor: Connected between the output and the supply voltage

CMOS Functioning and Parameters

Advantages of Using CMOS Transistors as Switches:

- **Low Power Consumption**: Because one of the transistors is always off.
- **High Noise Margin**
- **Scalability**: To create highly ICs with a large number of transistors.
- **Efficiency**: Efficient switching, making ideal for digital logic applications.

Applications of CMOS Switches:

- **Digital Logic Gates**: Implementing basic logic gates (AND, OR, NOT, NAND, NOR, XOR, XNOR) using CMOS technology.
- **Memory Cells**: Building memory elements like SRAM cells.
- **Multiplexers and Demultiplexers**: Switching circuits used in selecting and routing data signals.
- **Flip-Flops and Latches**: Basic storage elements in sequential logic circuits.
- **Analog Switches**: CMOS transistors can also be used in analog applications for switching signals in analog multiplexers and sample-and-hold circuits.

CMOS Functioning and Parameters

Estimation of Parasitic Values:

- The simplest model which can be used to represent the resistive and capacitive parasitic of the interconnect line consists of one lumped resistance and one lumped capacitance.
- Assuming that the capacitance is discharged initially, and assuming that the input signal is a rising step pulse at time $t = 0$, the output voltage waveform of this simple RC circuit is found as:

$$V_{out}(t) = V_{DD}\left(1 - e^{-\frac{t}{RC}}\right)$$

- The rising output voltage reaches the 50%-point at $t = t_{PLH}$, thus, we have:

$$V_{50\%} = V_{DD}\left(1 - e^{-\frac{\tau_{PLH}}{RC}}\right)$$

- The propagation delay for the simple lumped RC network is found as:

$$\tau_{PLH} \approx 0.69\,RC$$

CMOS Functioning and Parameters

Delay-Time Definitions: The delay times calculated using C_{load} may slightly overestimate the actual inverter delay.

Propagation delay

- The propagation delay times t_{PHL} and t_{PLH} determine the input-to-output signal delay during the high-to-low and low-to-high transitions of the output, respectively. By definition, t_{PHL} is the time delay between the $V_{50\%}$-transition of the *rising* input voltage and the $V_{50\%}$-transition of *the falling* output voltage. Similarly, t_{PLH} is defined as the time delay between the $V_{50\%}$-transition of *the falling* input voltage and the $V_{50\%}$-transition of the *rising* output voltage.

$$V_{50\%} = V_{OL} + \frac{1}{2}\left(V_{OH} - V_{OL}\right) = \frac{1}{2}\left(V_{OL} + V_{OH}\right)$$

- Thus, the propagation delay times t_{PHL} and t_{PLH} are found:

$$\tau_{PHL} = t_1 - t_0$$

$$\tau_{PLH} = t_3 - t_2$$

- The average propagation delay t_p of the inverter characterizes the average time required for the input signal to propagate through the inverter.

$$\tau_P = \frac{\tau_{PHL} + \tau_{PLH}}{2}$$

Rise time & fall time

- ✓ The rise time t_{rise} is defined here as the time required for the output voltage to rise from the $V_{10\%}$ level to $V_{90\%}$ level. Similarly, the fall time t_{fall} is defined here as the time required for the output voltage to drop from the $V_{90\%}$ level to $V_{10\%}$ level. The voltage levels $V_{10\%}$ and $V_{90\%}$ are defined as:

$$V_{10\%} = V_{OL} + 0.1 \cdot (V_{OH} - V_{OL}) \qquad \tau_{fall} = t_B - t_A$$

$$V_{90\%} = V_{OL} + 0.9 \cdot (V_{OH} - V_{OL}) \qquad \tau_{rise} = t_D - t_C$$

Power Dissipation

- ✓ Instantaneous power (drawn from supply):

$$P(t) = i_{DD}(t)V_{DD}$$

- ✓ Energy Consumed over a time interval T:

$$E = \int_0^T i_{DD}(t)V_{DD}dt$$

- ✓ Assuming periodic input and output waveforms, the average power dissipated by any device over one period can be found as follows:

$$P_{avg} = \frac{E}{T} = \frac{1}{T}\int_0^T i_{DD}(t)V_{DD}dt$$

Power Dissipation

- ✓ The average power dissipation of the CMOS inverter can be calculated as the power required to charge up and charge down the output load capacitance.

$$P_{avg} = \frac{1}{T}\left[\int_0^{T/2} V_{out}\left(-C_{load}\frac{dV_{out}}{dt}\right)dt + \int_{T/2}^{T}\left(V_{DD}-V_{out}\right)\left(C_{load}\frac{dV_{out}}{dt}\right)dt\right]$$

$$P_{avg} = \frac{1}{T}\left[\left(-C_{load}\frac{V_{out}^2}{2}\right)\Bigg|_0^{T/2} + \left(V_{DD}\cdot V_{out}\cdot C_{load} - \frac{1}{2}C_{load}V_{out}^2\right)\Bigg|_{T/2}^{T}\right]$$

$$P_{avg} = \frac{1}{T}C_{load}V_{DD}^2$$

$$P_{avg} = C_{load}\cdot V_{DD}^2\cdot f$$

Power Dissipation

Power Analysis:

- ✓ Static/Leakage Power Analysis
- ✓ Dynamic Power Analysis.

- Power Density of the Integrated Circuit increase exponentially with every Technology generation

  $P_{TOTAL} = P_{DYNAMIC} + P_{LEAKAGE}$

  $P_{DYNAMIC} = P_{SWITCHING} + P_{SHORT_CIRCUIT}$

- **Leakage Power (Static Power):** Leakage at almost all junctions due to various effects
  - Reverse Biased Diode Leakage
  - Gate Induced Drain Leakage
  - Gate Oxide Tunnelling
  - Sub-threshold Leakage

- **Switching Power:** When signal change their state, energy is drawn from the power supply to charge up the loadcapacitance from 0 to VDD

- **Short Circuit Power (Crowbar Power/ EM Rush Through Power):** Finite non-zero rise and fall times of transistors which causes a direct current path between the Power and Ground`

Power Dissipation

$$E = \int_0^t (V_{DD} I_{leak} + CV^2{}_{DD} f_c) dt$$

Total Power Dissipation

$$\int_0^t V_{DD} I_{leak} dt$$

Static Power Dissipation

Dynamic Power Dissipation

$$\int_0^t CV^2{}_{DD} f_c dt$$

Minimize I_{leak} by:

- Lower operating voltage
- Fewer leaking transistors

Minimize I_{switch} by:

- Lower operating voltage
- Less switching capacitance
- Less switching activity

Power Dissipation

Low Power Architecture:

- Low power architecture is essential in modern electronics, especially for portable and battery-operated devices like smartphones, wearables, and IoT.
- These architectures focus on reducing power consumption while maintaining or enhancing performance.
- Several strategies and design principles are employed to achieve low power consumption in digital circuits.
- The techniques used in low power architecture are:

1. **Voltage Scaling:** Lowering the voltage reduces power consumption
2. **Clock Gating:** Disables the clock signal to portions of the circuit that are not in use, effectively reducing dynamic power consumption
3. **Power Gating:** Disconnects the power supply to certain parts of a circuit when they are not in use. This technique reduces both dynamic and leakage power
4. **Multi-Threshold CMOS (MTCMOS):** Low-threshold transistors are used for high-speed operation, and high-threshold transistors are used for low-leakage, low-power states
5. **Body Biasing:** Adjusts the threshold voltage of transistors dynamically by changing the voltage of the substrate (body) to reduce leakage power
6. **Clock and Power Domains**: Isolate different parts of a circuit with separate clocks and separate power supplies control over power consumption

Session 6

Lecture: Technology Scaling

- ✓ Device sizing, rationed and non-rationed logic
- ✓ Concepts of Interconnect
- ✓ Transistor Scaling and interconnect Scaling

Technology Scaling

Technology scaling:

- It refers to the process of shrinking the dimensions of semiconductor devices to fit more transistors onto a single chip.
- This practice has been central to the semiconductor industry's ability to improve the performance, power efficiency, and cost-effectiveness of integrated circuits (ICs).
- The concept is often associated with **Moore's Law**, which predicts that the number of transistors on a chip will double approximately every two years.

Key Aspects of Technology Scaling:

- **Geometric Scaling:** Reducing the length of the transistor gate
- **Electrical Scaling:** Lowering the supply voltage to reduce power consumption and heat generation
- **Performance and Power Scaling:** Increasing the operating frequency of the transistors to enhance the overall performance of the IC

Technology Scaling

Device sizing, rationed and non-rationed logic:

- These are the key concepts in digital circuit design to optimize the performance, power consumption, and area of digital circuits.

Device Sizing:

- Device sizing refers to adjusting the dimensions of the transistors in a CMOS circuit, specifically their width (W) and length (L).
- The width of a transistor directly affects its drive strength, which influences the circuit's performance, power consumption, and area.

Rationed Logic:

- It refers to logic circuits where the sizing of the transistors is carefully chosen to achieve a specific ratio, often to balance the performance of PMOS and NMOS transistors or to optimize the power-delay product.
- To achieve similar rise and fall times, the PMOS transistor is often sized 2-3 times wider than the NMOS transistor because the electron mobility in NMOS transistors is higher than the hole mobility in PMOS transistors.
- This approach is commonly used in CMOS inverters and other basic gates to balance performance, optimized power consumption and improved signal integrity.

Technology Scaling

Non-Rationed Logic:

- Non-rationed logic does not follow specific sizing ratios between PMOS and NMOS transistors.
- Instead, the sizing is determined by other design considerations, such as area constraints, specific performance requirements, or power consumption targets.

Use Cases:

1. **Non-Critical Paths**: In paths where performance is not critical, non-rationed logic can be used to save area and power.
2. **Custom Logic**: In custom-designed circuits, where specific performance or power characteristics are needed, non-rationed logic allows for more flexibility.
3. **Special Applications**: In certain low-power or low-speed applications, non-rationed logic might be used to minimize area and power without stringent performance requirements.

Technology Scaling

Concepts of Interconnect:

- Interconnects are the wiring and signal pathways that connect various components on an integrated circuit (IC).
- As IC technology scales down, interconnects become increasingly critical in determining the overall performance, power consumption, and reliability of a chip.

Types of Interconnects:

a. Local Interconnects

- Connect nearby transistors within a small region of the IC.
- Typically found within standard cells in a cell-based design.

b. Intermediate Interconnects

- Connect different blocks or modules within an IC.
- Span moderate distances and may traverse multiple cells.

c. Global Interconnects

- Connect distant blocks or modules across the entire IC.
- Often used for clock distribution, power, and critical signal routing.

Technology Scaling

Concepts of Interconnect:

Interconnect Delay:

- **RC Delay**: Delay caused by the resistance and capacitance of the interconnect. Dominates in intermediate and local interconnects.
- **RLC Delay**: Includes inductance and is significant for global interconnects at high frequencies.
- **Elmore Delay Model**: A simplified method for estimating RC delay in interconnects.

Advanced Interconnect Technologies:

✓ 3D Ics

✓ Optical Interconnects

✓ Carbon Nanotubes (CNTs) and Graphene

Design Techniques for Interconnect Optimization:

❖ Wire Sizing

❖ Wire Spacing

❖ Buffer Insertion

❖ Layer Assignment

Technology Scaling

Transistor Scaling and Interconnect Scaling:

- The reduction of the size, i.e., the dimensions of MOSFETs, is commonly referred to as ***scaling.*** There are two basic types of size-reduction strategies:

  a. Full scaling (also called constant-field scaling)

  b. Constant voltage scaling

- Scaling of MOS transistors is concerned with systematic reduction of overall dimensions of the devices as allowed by the available technology.
- It is easy to recognize that the scaling of all dimensions by a factor of $S > 1$ leads to the reduction of the area occupied by the transistor by a factor of S^2.

Technology Scaling

a. Full scaling (also called constant-field scaling):

This scaling option attempts to preserve the magnitude of internal electric fields in the MOSFET, while the dimensions are scaled down by a factor of S.

Quantity	Before Scaling	After Scaling
Channel length	L	$L' = L/S$
Channel width	W	$W' = W/S$
Gate oxide thickness	t_{ox}	$t_{ox}' = t_{ox}/S$
Junction depth	x_j	$x_j' = x_j/S$
Power supply voltage	V_{DD}	$V_{DD}' = V_{DD}/S$
Threshold voltage	V_{T0}	$V_{T0}' = V_{T0}/S$
Doping densities	N_A N_D	$N_A' = S \cdot N_A$ $N_D' = S \cdot N_D$

$$C_{ox}' = \frac{\varepsilon_{ox}}{t_{ox}'} = S \cdot \frac{\varepsilon_{ox}}{t_{ox}} = S \cdot C_{ox}$$

$$I_D'(lin) = \frac{k_n'}{2} \cdot \left[2 \cdot (V_{GS}' - V_T') \cdot V_{DS}' - V_{DS}'^2\right]$$

$$= \frac{S \cdot k_n}{2} \cdot \frac{1}{S^2} \cdot \left[2 \cdot (V_{GS} - V_T) \cdot V_{DS} - V_{DS}^2\right] = \frac{I_D(lin)}{S}$$

$$I_D'(sat) = \frac{k_n'}{2} \cdot (V_{GS}' - V_T')^2 = \frac{S \cdot k_n}{2} \cdot \frac{1}{S^2} \cdot (V_{GS} - V_T)^2 = \frac{I_D(sat)}{S}$$

$$P' = I_D' \cdot V_{DS}' = \frac{1}{S^2} \cdot I_D \cdot V_{DS} = \frac{P}{S^2}$$

Quantity	Before Scaling	After Scaling
Oxide capacitance	C_{ox}	$C_{ox}' = S \cdot C_{ox}$
Drain current	I_D	$I_D' = I_D/S$
Power dissipation	P	$P' = P/S^2$
Power density	$P/Area$	$P'/Area' = P/Area$

Technology Scaling

b. Constant voltage scaling:

Here, all dimensions of the MOSFET are reduced by a factor of S, as in full scaling. The power supply voltage and the terminal voltages, on the other hand, remain unchanged.

$$I_D'(lin) = \frac{k_n'}{2} \cdot \left[2 \cdot (V_{GS}' - V_T') \cdot V_{DS}' - V_{DS}'^2\right]$$

$$= \frac{S \cdot k_n}{2} \cdot \left[2 \cdot (V_{GS} - V_T) \cdot V_{DS} - V_{DS}^2\right] = S \cdot I_D(lin)$$

$$I_D'(sat) = \frac{k_n'}{2} \cdot (V_{GS}' - V_T')^2 = \frac{S \cdot k_n}{2} \cdot (V_{GS} - V_T)^2 = S \cdot I_D(sat)$$

$$P' = I_D' \cdot V_{DS}' = (S \cdot I_D) \cdot V_{DS} = S \cdot P$$

Quantity	Before Scaling	After Scaling
Dimensions	W, L, t_{ox}, x_j	reduced by S $(W' = W/S, \ldots)$
Voltages	V_{DD}, V_T	remain unchanged
Doping densities	N_A, N_D	increased by S^2 $(N_A' = S^2 \cdot N_A, \ldots)$

Quantity	Before Scaling	After Scaling
Oxide capacitance	C_{ox}	$C_{ox}' = S \cdot C_{ox}$
Drain current	I_D	$I_D' = S \cdot I_D$
Power dissipation	P	$P' = S \cdot P$
Power density	$P/Area$	$P'/Area' = S^3 \cdot (P/Area)$

Problems Due to Scaling

- This large increase in current and power densities may eventually cause serious reliability problems for the scaled transistor, such as **electromigration, hot-carrier degradation, oxide breakdown, electrical over-stress (EOS)** etc.
- **Electromigration:** Displacement of the atoms as a result of current flowing through a conductor which causes voids and failures in a device.
- **Hot-carrier degradation:** Heating inside the channel of the MOSFET by the electric field during circuit operation. The accumulation of damage resulting degradation in device behavior due to hot carrier injection.
- **Oxide breakdown:** If the oxide electric field perpendicular to the surface is larger than a certain breakdown field, the silicon-dioxide layer may sustain permanent damage during operation, leading to device failure.
- **Electrical over-stress (EOS):** Thermal damage that may occur when an electronic device is subjected to a current or voltage that is beyond the specification limits of the device.

Problems Due to Scaling

Short-Channel Effects:

- As a working definition, a MOS transistor is called a short-channel device if its channel length is on the same order of magnitude as the depletion region thicknesses of the source and drain junctions. Alternatively, a MOSFET can be defined as a short-channel device if the effective channel length L_{eff} is approximately equal to the source and drain junction depth x_j. The short-channel effects that arise in this case are attributed to two physical phenomena:

  (i) The limitations imposed on electron drift characteristics in the channel,

  (ii) The modification of the threshold voltage due to the shortening channel length.

Narrow-Channel Effects:

- MOS transistors that have channel widths W on the same order of magnitude as the maximum depletion region thickness x_{dm} are defined as narrow-channel devices. The most significant narrow-channel effect is that the actual threshold voltage of such a device is larger than that predicted by the conventional threshold voltage.

Problems Due to Scaling

Channel hot-electron:

- The channel hot-electron (CHE) effect is caused by electrons flowing in the channel region, from the source to the drain. This effect is more pronounced at large drain to source voltages, at which the lateral electric field in the drain end of the channel accelerates the electrons. The electrons arriving at the Si-SiO_2 interface with enough kinetic energy to overcome the surface potential barrier are injected into the oxide. Electrons and holes generated by impact ionization also contribute to the charge injection. Note that the channel hot-electron current and the subsequent damage in the gate oxide are localized near the drain junction.

- Hence, new MOSFET technologies based on smaller device dimensions must carefully account for the hot-carrier effects and also ensure reliable long-term operation of the devices. Other reliability concerns for small-geometry devices include interconnect damage through **electromigration**, **electrostatic discharge (ESD)** (sudden flow of electricity between two electrically charged objects caused by contact, an electrical short or dielectric breakdown) and **electrical over-stress (EOS).**

THANK YOU