VLSI II – Summary

written by Stephan Senn, D-ITET

24 Aug 2006

Contents

* belongs to VLSI III

Clocking of Synchronous Circuits

Timewise Scattering of Clock Signals

- unequal switching thresholds of transistors
- unbalanced interconnect delays
- supply noise by ground bounce and supply droop
- PTV and OCV
- crosstalk from switching activities
- unevenly distributed fanouts and load capacitances
- unevenly distributed drive strengths and buffers

Timing Quantities

- *Clock Skew t_{sk}:* inaccuracy of the same clock edge arriving at different locations within a given clock domain

  $t_{sk} = t_{di\,rcv} - t_{di\,xmt}$ $t_{di\,rcv}$: delay for receiving circuit $td_{i\,xmt}$: delay for sending circuit

- *Clock Jitter t_{ji}:* inaccuracy of consecutive clock edges arriving at the same location
- *Clock Distribution Delay t_{di}:* time lag measured from when a clock edge appears at the clock source until a state transition actually takes place in response to that edge

Data Call Window and Data Valid Window: see p. 3ff.

- Data Call Window: aka aperture window, consisting of setup and hold time, stable required time window
- Data Valid Window: consisting of propagation and contamination delays, actual stable time window
- Hold Margin: margin of exceeding the hold time
- Setup Margin: margin of exceeding the setup time
- Skew Margin: minimum of the hold and the setup margins with respect to the clock
- Noise Margin: minimum of the hold and the setup margins
- General Rule: *The Data Valid Window fully encompasses the Data Call Window!*
- Concluding Rule: *Clock skew and jitter do not exceed setup and hold margin within a clock domain or at any operating condition.*

Clocking Disciplines

- Single-Edge-Triggered One-Phase Clocking
- Dual-Edge-Triggered One-Phase Clocking
- Symmetric Level-Sensitive Two-Phase Clocking
- Unsysmmetric Level-Sensitive Two-Phase Clocking
- Single-Wire Level-Sensitive Two-Phase Clocking
- (Level-Sensitive One-Phase Clocking and Wave Pipelining)

Single-Edge-Triggered One-Phase Clocking (SETOPC)

- $T_{cp} = T_{clk}$
- Bistables: Single-Edge-Triggered Flip-flops (SETFF)
- Timing Analysis:
  - longest path: $t_{lp} = \max(t_{pd\,ff} + t_{pd\,c} + t_{su\,ff})$
  - shortest path: $t_{sp} = \min(t_{cd\,ff} + t_{cd\,c} - t_{ho\,ff})$
  - setup condition: $\max|t_{sk}| \leq T_{cp} - t_{lp}$
  - hold condition: $\max|t_{sk}| \leq t_{sp}$
  - slack time: $t_{slack} = T_{cp} - (t_{lp} + \max|t_{sk}|)$
- Conclusion:
  - inherently sensitive to clock skew
  - shift registers (and therefore also scan paths) are very vulnerable to clock skew
  - Design For Test (DFT) with Scan-Chains: full scan, partial scan, Automatic Test Pattern Generation (ATPG), special scan-flip-flops in use
  - Benefits of Scan-Chains:
    - checking of the proper functioning of the chaining flip-flops
    - serially read out the circuit's state
    - serially put the circuit into a user-defined state
    - combinational logic test by applying suitable test vector sets (in and out shifting)
  - Mixing Logic Families: no way! dangerous!

Dual-Edge-Triggered One-Phase Clocking (DETOPC)

- $T_{cp} = \frac{1}{2} T_{clk}$
- Bistables: Dual-Edge-Triggered Flip-flops (DETFF)
- Timing Analysis:
  - longest path: $t_{lp} = \max(t_{pd\,ff} + t_{pd\,c} + t_{su\,ff})$
  - shortest path: $t_{sp} = \min(t_{cd\,ff} + t_{cd\,c} - t_{ho\,ff})$
  - setup condition: $\max|t_{sk}| \leq T_{cp} - t_{lp}$
  - hold condition: $\max|t_{sk}| \leq t_{sp}$
  - slack time: $t_{slack} = T_{cp} - (t_{lp} + \max|t_{sk}|)$

- Conclusion:
  - the same problems like SETOPC
  - better energy efficiency due to half-frequency clocking in comparison to SETOPC (overall power reduction of about 10-20%)[1]
  - computer memory and mainboard design: Double Data Rate (DDR)

Symmetric Level-Sensitive Two-Phase Clocking (SLSTPC)

- $T_{cp} = T_{clk} = T_1 + T_2 + T_3 + T_4$
- Bistables: Latches
- Timing Analysis:
  - run time for C_{21}: $t_{21} \leqslant T_1 + T_3 + T_4 = T_{cp} - T_2$
  - run time for C_{12}: $t_{12} \leqslant T_1 + T_2 + T_3 = T_{cp} - T_4$
  - total run time condition: $\max(t_{12} + t_{21}) \leqslant T_{cp}$
  - setup conditions:

    $$\max|t_{sk}| \leqslant T_{cp} - T_2 - \max(t_{pd\,la2} + t_{pd\,C21} + t_{su\,la1})$$

    $$\max|t_{sk}| \leqslant T_{cp} - T_4 - \max(t_{pd\,la1} + t_{pd\,C12} + t_{su\,la2})$$

    $$0 \leqslant T_{cp} - \max(t_{pd\,la1} + t_{pd\,C12} + t_{pd\,la2} + t_{pd\,C21})$$
  - hold conditions:

    $$\max|t_{sk}| \leqslant T_2 + \min(t_{cd\,la2} + t_{cd\,C21} - t_{ho\,la1})$$

    $$\max|t_{sk}| \leqslant T_4 + \min(t_{cd\,la1} + t_{cd\,C12} - t_{ho\,la2})$$
- Conclusion:
  - time borrowing
  - sizing of the non-overlap intervals
  - area overhead due to extra design effort[2]
  - no HDL synthesis support at the moment
  - geometrically and electrically similar nets preferable
  - relaxed timing: improvement of energy-efficiency and switching noise

1 higher capacitive loads than SETOPC
2 Note: Two latches occupy more chip area than a single flip-flop!

Unsymmetric Level-Sensitive Two-Phase Clocking (ULSTPC)

- $T_{cp} = T_{clk} = T_1 + T_2 + T_3 + T_4$
- Bistables: Latches
- Timing Analysis:
  - run time for C_{21}: $t_{21} \leqslant T_1 + T_3 + T_4 = T_{clk} - T_2$
  - run time for left part: $t_l \leqslant T_2 + T_3$
  - total run time condition: $\max(t_l + t_{21}) \leqslant T_{cp}$
  - setup conditions:

    $\max|t_{sk}| \leqslant T_{cp} - T_2 - \max(t_{pd\,la2} + t_{pd\,C21} + t_{su\,la1})$

    $\max|t_{sk}| \leqslant T_2 - T_3 - \max(t_{pd\,la1} + t_{su\,la2})$
  - hold conditions:

    $\max|t_{sk}| \leqslant T_2 + \min(t_{cd\,la2} + t_{cd\,C21} - t_{ho\,la1})$

    $\max|t_{sk}| \leqslant T_4 + \min(t_{cd\,la1} - t_{ho\,la2})$
- Conclusion:
  - the same pros and cons like SLSTPC
  - unproductive region T_2
  - wide skew margins possible
  - Level Sensitive Scan Design (LSSD): see p. 18

Single-Wire Level-Sensitive Two-Phase Clocking (SWLSTPC)

- $T_{cp} = T_{clk} = T_1 + T_3$
- Bistables: Latches
- Timing Analysis:
  - run time for C_{21}: $t_{21} \leqslant T_1 + T_3$
  - run time for C_{12}: $t_{12} \leqslant T_1 + T_3$
  - total run time condition: $\max(t_{12} + t_{21}) \leqslant T_{clk}$
  - setup conditions:

    $\max|t_{sk}| \leqslant T_{cp} - \max(t_{pd\,la} + t_{pd\,c} + t_{su\,la})$ with $t_{cp\,c} = \max(t_{cp\,C21}, t_{cp\,C12})$

    $0 \leqslant T_{cp} - \max(2\,t_{pd\,la} + t_{pd\,c})$
  - hold conditions:

    $\max|t_{sk}| \leqslant \min(t_{cd\,la} + t_{cd\,c} - t_{ho\,la})$ with $t_{cd} = \min(t_{cd\,C21}, t_{cd\,C12})$
- Conclusion:
  - as sensitive to skew and jitter as SETOPC
  - bistables with embedded logic in use

Containment of Clock Skew

- Problems of Slow Clock Ramps:
  - Disparities of transistor thresholds across clocked cells lead to clock skew. This effect gets even worse when the clock ramps are very slow.
  - Correct behaviour and accurate timing of flip-flops and other components are put at risk.
  - Sluggish clocks tend to inflate setup and hold times.
- General Guidance: Waveforms of clocks must be kept clean and sharp!
- Clock Distribution: see p. 26
  - collective buffer:
    - as short wires as possible
    - central driver
    - reasonably wide wires
    - use low-resistance metal layers
    - avoid unnecessary layer changes
    - equalize delays
  - distributed buffer tree: make use of clock tree generators
    - hierarchical partition: balance the clock loads
    - local subtrees and reasonable sized clock buffers
    - use low-resistance metal layers
    - dummy loads, extra buffer insertion and detours for early branches
  - hybrid approaches: distributed clock tree and clock meshes
- Analysis: Static Timing Analysis (STA)

Input / Output Timing

- Overview:

I/O Timing:	**Data Call Window**	$t_{su\ inp}$	$t_{ho\ inp}$	**Data Valid Window**	$t_{pd\ oup}$	$t_{cd\ oup}$
friendly	narrow	small	close to zero or negative	wide	small	large, as close to $t_{pd\ oup}$ as possible
unfriendly	wide	large	large	narrow	large	close to zero

- Improvements for more friendly Timing:
  - using output registers
  - use input registers only in combination with extra contamination delays (adding artificial contamination delays)

- reduce clock distribution delay t_{di}:
  - $t_{su\,inp} = \max(t_{pd\,b} + t_{su\,ff}) - t_{di}$
  - $t_{ho\,inp} = \max(-t_{cd\,b} + t_{su\,ff}) + t_{di}$
  - $t_{pd\,out} = \max(t_{pd\,d} + t_{pd\,ff}) + t_{di}$
  - $t_{cd\,out} = \max(t_{cd\,d} + t_{cd\,ff}) + t_{di}$
- driving I/O registers from an early clock: reduce data call window
- tapped (delayed) domain's clock: clock feeded through a chip and then provided to the board for better timing characteristics
- Delay Locked Loop (DLL) or Phase Locked Loop (PLL)

Clock Gating

- Idea: Any toggling of the clock input of a disabled flip-flop amounts to wasting energy in discharging and recharging the associated node capacitances for nothing.
- Clock Gating Circuits: see p. 38f.
- Requirements:
  - no latency: the enable input must affect the next active clock edge
  - gated clock output: free of hazards
  - enable input: must be immune to hazards
  - standard timing requirements for the enable input: setup and hold times
  - small propagation delay of the clock
  - low power dissipation for gated mode
  - the same duty cycle[3] for gated and non-gated clocks
  - reset: well-defined value at reset!

Summary: see p. 41ff.

3 Duty cycle: $D = \frac{\tau}{T_{cp}}$ with τ: active phase

Acquisition of Asynchronous Data

Basic Problems

- Data Inconsistency
- Synchronizer Metastability

Data Inconsistency with Vectored Acquisition

- Vectored Acquisition: two clock domains are traversed by two or more electrical lines that form together a data, control or status word
- Crossover State: mixing of new with old bits, cannot be avoided with the help of single synchronizer flip-flops
- Counteractive Measures:
  - two-way sampling with comparison: see p. 51
  - unit distance coding: hamming distance one or less
  - handshaking: with request (REQ) and acknowledge (ACK)
    - transition signalling (or two-phase protocol, two-stroke protocol, NRZ[4] protocol): p. 53
    - level signalling (or four-phase protocol, four-stroke protocol, RZ[5] protocol): p. 53
    - partial or unsymmetric handshaking: p. 55

Data Inconsistency with Scalar Acquisition

- Scalar Acquisition: two clock domains are traversed by only one electrical line
- Problem: bad synchronization (see p. 57)
- Solutions:
  - use input flip-flop
  - generate complementary signal at the input (no transmission)
- Remaining Problem: marginal triggering (see next section!)
- Possible Solutions against Marginal Triggering for Slow Clocks:
  - dual-edge triggering for synchronizers
  - tapped delay lines
  - conversion of serial into parallel data transmission for lowering data transmission rate
  - analogue PLL for clocking synchronizer faster and in phase

Synchronizer Metastability

- Problem: As no fixed timing relationship can be guaranteed between data and clock, the input signal will occasionally toggle in the immediate vicinity of a clock event, thereby ignoring the requirement that data must remain stable throughout the bistable's data call window.
- Visual behaviour: oscillations, glitching, slow ramping, in general excessive delays
- Effects: Metastability resolution leads to unpredictable excessive delays! The metastability

4 stands for: Non-Return-to-Zero
5 stands for: Return-to-Zero

resolution time is orders of magnitude bigger than the propagation delay: $t_{mr} > t_{pd}$. See p. 61!

- Known as: timing violation or marginal triggering
- There are no counteractive measures that avoid marginal triggering, however.
- Statistical Model to estimate the Synchronization Failure:

$$MTBE = \frac{e^{K_2 t_{al}}}{K_1 f_{clk} f_d}$$

$$t_{al} = T_{clk} - \max(t_{pdc}) - t_{suff} \quad \text{with} \quad T_{clk} = \frac{1}{f_{clk}}$$

t_{al}: time-wise allowance for resolving metastability

K_1, K_2: parameters

f_{clk}: clocking frequency

f_d: average edge rate of asynchronous data signal

Assumption: Clock and data signal are independent of each other!

- Containment of Metastable Behaviour: Keep the number of synchronization operations as small as possible and allow as much time as practical for any metastable condition to resolve.
  - select flip-flops with good metastability resolution
  - drive synchronizers with fast-switching clock: fast ramps and narrow data call window
  - free synchronizers from unnecessary loads
  - estimate reliability at the system level by rising the ratio clock period to longest propagation delay
  - remove combinational delays from synchronizers (two-stage or multi-stage synchronizers)
  - lower clock frequency at the consumer end
- Plesiochronous Systems: Systems where data producer and consumer are clocked from separate oscillators that operate at the same nominal frequency. Note that the presented statistical model is not valid. These systems make use of PLLs and DLLs.

Parametric Design Verification

Verification Tools (see also p. 87)

- **Dynamic Methods:**
  - Gate-Level Simulation
- **Static Methods:**
  - Code Inspection
  - Electrical Rule Check (ERC)
  - Timing Verification or Static Timing Analysis (STA)
- **Rules:**
  - Always make use of all of them.
  - Always check the actual operating conditions against those for timing verification.
  - Stick to realistic figures!
  - "Simulation is based on models, and models are wrong."
- **Known Timing Problems:**
  - inadequate clock waveform
  - insufficient setup and hold times: longest and shortest paths
  - excessive clock skew and distribution delays
  - slow ramps on informations signals
  - asynchronous signals in conjunction with poor synchronizers

Electrical Rule Check (ERC)

Inspection of netlists in order to find anomalies in the circuit structures

- unconnected power and ground fragments
- shorted power and ground nodes
- missing drivers on output pad
- cell inputs left open
- permanent bus contention
- MOSFET terminals left unconnected or shorted together

Code Inspection

Inspection of VHDL or Verilog code

- naming conflicts
- inconsistent index and/or address ranges
- false endian type: big or little endian
- driver conflicts
- no established clocking scheme in use

- zero latency loops
- high-fanout or high-fanin circuit structures: large centralized multiplexer, busses, etc.
- large centralized control structure

Accuracy of Timing Analysis

- **Cell Delays:**
  - *input-to-output pathes:* pin-to-pin delay model, individual characterization of input-to-output pathes for falling and rising edges, state-dependent delays possible (normally not modelled)
  - *transient behaviour:* described by setup and hold times as well as contamination and propagation delays, not captured in standard simulation models (t_{cd} is replaced by t_{pd}, zero ramping!)
  - *load dependencies:* prop/ramp model, adequate for CMOS circuits before 0.5µm

    $$t_{pd} = t_{id} + t_{ed} = t_{id} + r_{cap} C_{ext} = t_{cd} + r_{cap} C_{ext}$$

    t_{id}: intrinsic delay (corresponds to the contamination delay t_{cd})

    t_{ed}: extrinsic delay with load factor r_{cap} and output capacitance C_{ext}
  - *input waveform dependencies:* sophisticated input slope model with look-up tables
  - *statistical variations:*
    - Process Temperature Voltage (PTV) Variations: variations on a processed wafer

      $$t_{derated} = t_{nominal} \cdot K_P K_\theta K_V$$

      K_P(fast,typical,slow), $K_\theta(\theta_j)$, $K_V(U_{dd})$: derating curves
    - On-Chip Variations (OCV): variations on a die, for technologies smaller than 130nm
  - *trip point and library characterization:* see p. 77f.
- **Interconnect Delays and Layout Parasitics:**
  - lumped RC network model (there are also others)
  - leg: resistance-capacitance pair, piece of wire of uniform electrical characteristics or very nearly so
  - parasitic resistances of a leg k:

    $$R_k = R_{sq} \frac{l_k}{w_k} + \frac{1}{m(k)} R_{plug\ i,j}$$

    m(k): number of parallel plugs at leg k
  - parasitic capacitances of cells and interconnect lines:

    $$C_{ext} = \sum_{j=1}^{J} C_{inp}(j) + C_{line}$$

    $$C_{line} = \sum C_{plate} + \sum C_{fringe} + \sum C_{lateral} \approx l\,(w\,c_{plate\,M1\,bulk} + 2\,c_{fringe\,M1\,bulk})$$

    J: number of cells being driven
  - Min/Max Timing Verification: no way! too pesimistic!

Static Timing Analysis

- Positive Aspects:
  - works without stimuli and expected responses
  - no coverage problem
- Negative Aspect:
  - accuracy: Delay figures also depend on data patterns not only on geometric layouts and therefore compromise STA.
- Scheme: see p. 73

Circuit Simulation

- Preconditions for better Simulation:
  - The simulation must be carried out at the gate-level.
  - Zero-latency loops are not allowed.
  - The simulator must be properly set up to report unsettled nodes, if any.
  - The longest and the shortest delay must be exercised along every signal propagation path.
  - Include necessary checks in the HDL code.
- Unsettled Node: A situation where a circuit gets clocked before it has settled to a steady state. The current state is only temporarily stable. That is the reason why it is not detected.
- Transistor Level Simulation: too slow, no meaningful reporting
- Checks: They are delegated to a circuit model!

Signal Integrity

- Random Noise:
  - Conductive Coupling: noise picked up by a wire
  - Electromagnetic Coupling: electromagnetic fields, crosstalk between adjacent on-chip lines
  - Common Impedance Coupling: current fluctuations and noise voltage building
- Two Important Requirements:
  - level restoration: Restore level!
  - clear difference between 1 and 0 states: Keep 1 and 0 apart!
- Noise Margin U_{nm}: see p.95
  - $U_{nml} = U_{il} - U_{ol}$
  - $U_{nmh} = U_{oh} - U_{ih}$
  - $U_{nm} = \min(U_{nml}, U_{nmh})$
  - typical value: $U_{nm} = 0.43V$
  - forbidden intervals
- Ground Bounce and Supply Droop:
  - working principle and modelling: see p. 96ff.
  - roots of noise in circuits: pad drivers (simultaneous output switching noise, SSO noise or SSN), clocking, buffer contention
  - trend: Switching noise is a critical issue in VLSI and is bound to become even more important in the future. This is because output pin counts continue to grow while both supply voltages and switching times shrink.
- Counteractive Measures:
  - *reduction of series impedances:*
    - ground and power pads: sufficient number of pads
    - package choice: prefer solder ball, general guide: low inductance and low impedance
    - optimum pinout: special lead frames, corner and center pinning
    - bypass capacitor and on-chip bypass capacitor: for momentarily supply energy for switching activities, high resonance frequency[6], low equivalent series inductance (ESL), low equivalent series resistance (ESR)
    - low impedance supply routing: see p.107
    - solid PCB ground and power connections: avoid wrapping even for test purpose
  - *separation of noise pollutors from potential victims*
    - physical isolation: heavy noise in padframe (ESD protection, pad drivers, clock buffers, heavy loads, etc.), low noise in core (input buffers, level shifter, etc.)
    - electrical decoupling of core and padframe: separate pads, split power lead frames, split

6 definition: $f_0 = \frac{1}{\sqrt{LC}}$

and interspersed supplies

- electrical decoupling of switching pads from nonswitching ones: sepcial pad drivers (see p.109)

- *avoidance of excessive switching currents*
  - sizing of drivers: not stronger than absolutely necessary, testing: high currents, speed and noise are proportional!
  - slew rate control: negative feedback circuits, other approaches: voltage references, selectively disable circuit sections
  - soft switching drivers: minimizing simultaneous switching of n- and p-MOSFETs by special circuits
  - staggered switching: avoid simultaneous switching of pads by switching over a lapse of time
  - Low Voltage Differential Signalling (LVDS): reduce common mode noise, absence of current spikes, less vulnerable to PTV variations
- *safeguard noise margins*
  - level shifters at inputs: core supply
  - special attention for asynchronous reset and gated clocks: look for ample noise margins, use Schmitt triggers
  - handling of unused inputs: set to well-defined reference
  - optimum switching threshold: designing on transistor-level
  - noise analysis: coverage problem, accurancy

Gate- and Transistor-Level Design

Naming and Counting Conventions: see p. 172f.

MOSFET Models

- **The Sah Model aka Shockley Model: L > 2µm**

	n-type:	**p-type:**
Subthreshold Region (Cut-Off Region)	$I_d = 0$ $U_{gs} \leqslant U_{thn}$	$I_d = 0$ $U_{sg} \geqslant U_{thp}$
Linear Region (Triode Region)	$I_d = \frac{1}{2}\left[2\beta_n (U_{gs} - U_{thn}) U_{ds} - U_{ds}^2\right]$ $0 < U_{ds} < U_{gs} - U_{thn}$	$I_d = \frac{1}{2}\left[2\beta_p (U_{thp} - U_{gs}) U_{ds} - U_{ds}^2\right]$ $U_{gs} - U_{thp} < U_{ds} < 0$
Saturation Region (Active Region)	$I_d = \frac{1}{2}\beta_n (U_{gs} - U_{thn})^2$ $0 < U_{gs} - U_{thn} \leqslant U_{ds}$	$I_d = \frac{1}{2}\beta_p (U_{thp} - U_{gs})^2$ $U_{ds} \leqslant U_{gs} - U_{thp} < 0$

MOSFET gain factor:

$$\beta_{n,p} = \frac{\mu_{n,p}\epsilon_{ox}}{t_{ox}} \cdot \frac{W_{n,p}}{L_{n,p}} = \beta_{sqn,sqp} \cdot \frac{W_{n,p}}{L_{n,p}}$$

$\beta_{Sqn,sqp}$: process gain factor

Sign-Rule for p-type MOSFETs:

- 1. alternative: Change *ds* and *gs* to *sd* and *sg*! Set $U_{th} > 0$!
- 2. alternative: Change the sign for U_{sg} and U_{th}!

Note: Enhancement and Depletion Type

- **The Shichman-Hodges Model: L > 2µm**

	n-type:
Subthreshold Region (Cut-Off Region)	$I_d = 0$ $U_{gs} \leqslant U_{thn}$
Linear Region (Triode Region)	$I_d = \frac{1}{2}\left[2\beta_n (U_{gs} - U_{thn}) U_{ds} - U_{ds}^2\right](1 + \lambda U_{ds})$ $0 < U_{ds} < U_{gs} - U_{thn}$
Saturation Region (Active Region)	$I_d = \frac{1}{2}\beta_n (U_{gs} - U_{thn})^2 (1 + \lambda U_{ds})$ $0 < U_{gs} - U_{thn} \leqslant U_{ds}$

λ: channel length modulation factor

- **The Alpha-Power Law Model: L < 2µm**

	n-type:	**p-type:**
Subthreshold Region (Cut-Off Region)	$I_d = 0$ $U_{gs} \leqslant U_{thn}$	$I_d = 0$ $U_{sg} \geqslant U_{thp}$
Linear Region (Triode Region)	$I_d = I'_{d0n}$ $U_{ds} < U'_{d0n}$	$I_d = I'_{d0p}$ $U_{ds} > U'_{d0p}$
Saturation Region (Active Region)	$I_d = \left(2 - \frac{U_{ds}}{U'_{d0n}}\right)\frac{U_{ds}}{U'_{d0n}} I'_{d0n}$ $U_{ds} \geqslant U'_{d0n}$	$I_d = \left(2 - \frac{U_{ds}}{U'_{d0p}}\right)\frac{U_{ds}}{U'_{d0p}} I'_{d0p}$ $U_{ds} \leqslant U'_{d0p}$

$$I'_{d0n} = P_{cn}\frac{W_n}{L_n}(U_{gs} - U_{thn})^{\alpha_n} = I_{d0n}\cdot\left(\frac{U_{gsn} - U_{thn}}{U_{dd} - U_{thn}}\right)^{\alpha_n} \quad , \quad I_{d0n} = I_d(U_{gs} = U_{dd})$$

$$U'_{d0n} = P_{vn}(U_{gs} - U_{thn})^{\frac{\alpha_n}{2}} = U_{d0n}\cdot\left(\frac{U_{gsn} - U_{thn}}{U_{dd} - U_{thn}}\right)^{\frac{\alpha_n}{2}} \quad , \quad U_{d0n} = U_{dd} - U_{thn}$$

$$I'_{d0p} = -P_{cp}\frac{W_p}{L_p}(U_{thp} - U_{gs})^{\alpha_p} = I_{d0p}\cdot\left(\frac{U_{thp} - U_{gs}}{U_{dd} + U_{thp}}\right)^{\alpha_p} \quad , \quad I_{d0p} = I_d(U_{gs} = U_{dd})$$

$$U'_{d0p} = -P_{vp}(U_{thp} - U_{gs})^{\frac{\alpha_p}{2}} = U_{d0n}\cdot\left(\frac{U_{thn} - U_{gs}}{U_{thp} - U_{dd}}\right)^{\frac{\alpha_p}{2}} \quad , \quad U_{d0p} = U_{dd} - U_{thp}$$

$\alpha_{n,p}$: velocity saturation index (between 1 and 2)

Properties: strong longitudinal fields (≈0.7V/µm), also vertical fields, mobility degradation due to velocity saturation

Note: Attention no channel length modulation included!

- **Second Order Effects:**
  - *short channel effect:* distorted fields due to drain-body and source-body fields, threshold voltages tend to drop below $L_g < 500$nm
  - *narrow channel effect:* distorted fields due to field implants and oxide fields
  - *back gate effect:* shift of the threshold voltage $U_{th\,n}$ [$U_{th\,p}$] towards more positive [negative] values by $U_{sb\,n} > 0$ [$U_{sb\,p} < 0$]
  - *subthreshold conduction:* $I_d \propto 10^{\frac{U_{gs} - U_{thn}}{S}} W_n$ for n-type in subthreshold region with $S = 70...120\,mV/dec$
  - *thermal dependencies:* thermal dependencies of transistor parameters
  - *geometric distortions:* lateral diffusion, overetching of the gate material, etc.

- **Typical Values:**
  - $\mu_n = 400...700 \frac{cm^2}{Vs}$
  - $\mu_p = 100...300 \frac{cm^2}{Vs}$

CMOS Logic

- **Ratioed and Ratioless Circuits:**

  A circuit or subcircuit is ratioless, if the geometric sizes and current drive capabilities of its transistors do not affect its logic function as specified in a truth table, for instance.

- **Antagonistic Structure:** p-channel pull-up network and n-channel pull-down network

  Antagonism between pull-down and pull-up networks is entirely sufficient for a fully complementary static CMOS gate. Structural duality is a stronger criterion that implies electrical antagonism.

  Note that there are also self-dual structural, so called identical structures, which are not amenable to series / parallel decomposition.

- **Duality:** drawing rules

CMOS Logic Gates: see p. 128ff.

- **NAND Gate**
- **NOR Gate**
- **AOI Gates**
- **Branch-Based Logic**
- **Transmission Gates and other Three-State Gates**
- **Parity Gates: XOR and EQV**
- **Adder and Ripple-Carry-Adder**
- **Latches**
  - Switched Feedback Loop
  - Overpowered Feedback Loop
  - Jamb Latch
  - Clocked Inverters Connected Back to Back
  - Seesaw with Input Gating
- **Function Latches**
- **Single-Edge-Triggered Flip-Flops**
- **Dual-Edge-Triggered Flip-Flops**
- **Static RAM (S-RAM)**
- **Dynamic RAM (D-RAM)**
- **Snapper, Schmitt Trigger, Tie-Off-Cell, Filler-Cell, Level-Shifters, Adjustable Delay Lines**

Busses and Three-State Nodes

- Avoidance of Interconnect Overhead: use of multi-driver nodes or busses
- Problems:
  - bus contention: stationary contentions and transisent conflicts
  - floating
- Solutions:
  - pull-ups and pull-downs: static power dissipation, static currents, slow ramp times, switching noise
  - active snappers: marginal triggering, spurious signals
  - central access control (better than distributed access control)

Safe Designs with Transmission Gates

- Problems:
  - no drive capability (no amplification)
  - no level restoration: no regeneration of depressed voltage levels
  - backward signal propagation (bidirectionality)
  - leads to testability problems
- Design Rules:
  - viewed from outside: invisible bidirectionality
  - no driving conflicts
  - no floating of any node
  - restored output levels
  - no repercussions from outputs to inputs or to state registers
  - characterization with a truth table
  - no more than three MOSFETs in series

Interfacing with Mechanical Contacts

- Debouncing
- Solutions:
  - SR-Seesaws
  - Snapper
  - Synchronizer Circuit with Sampling Facility
  - Microcomputer (with Software, Debouncing Algorithm)

The CMOS Inverter: see p. 117ff.

- Operating Principle: 1 → 0

	n-type:	**p-type:**
Operating Region A	cut-off	linear
Operating Region B	saturation	linear
Operating Region C	saturation	saturation
Operating Region D	linear	saturation
Operating Region E	linear	cut-off

- Characteristics of CMOS Circuits: level restoration due to amplification
- Optimal Gate Widths: $\frac{W_p}{W_n} = \sqrt{\frac{\mu_n}{\mu_p}}$
- Propagation Delay: $t_{pd} \propto \frac{C_k U_{dd}}{(U_{dd} - U_{th})^{\alpha}}$ with $C_k = 2C_m + C_l$
  - for long transistors: L > 3µm, α = 2
  - for short transistors: L < 3µm, α < 2

Energy Efficiency and Heat Removal

Figures of Interest

	General Purpose Processor	Low Power Circuit
min. feature size	65 – 100nm	65 – 200nm
die size	110 – 150mm^2	< 10mm^2
wiring level	7 – 9	< 6
transistor count	40 – 130M	< 1M
supply current	30 – 85A	< 1mA
clock frequency	1.8 – 3.6GHz	< 500MHz
power dissipation	40 – 120W	< 1mW
power density	30 – 100W/cm^2	< 5mW/cm^2

- kitchen hotplate: 7W/cm^2
- battery-operated circuits: problem how to get the energy in
- high-performance circuits: problem to get heat out

Energy in CMOS Circuits

- **Charging and Discharging of Capacitive Loads:**

$$E_{chk} = \frac{\alpha_k}{2} C_k U_{dd}^2 \quad , \quad C_k \approx C_{gate\,k} + C_{wire\,k}$$

  - activity factor α_k: indicates how many times per computation cycle a given node k switches from one logic state to the opposite one when averaged over many such computation cycles
  - amount of computation cycles for one full charge-discharge cycle: $\frac{2}{\alpha_k}$
  - glitch-induced activity: values in excess of 6
  - temporal and spatial correlations: LSB and MSB
  - evaluation of switching activities: from logic simulation, from statistical analysis of the gate-level netlist (see p. 192)

- **Crossover Currents:**

$$E_{crk} \approx \sigma_k E_{chk}$$

  - depends on many parameters:
    - supply voltage U_{dd}
    - geometric sizes of the transistors: W and L
    - electrical characteristics of the transistors: U_{th}, $\beta_{\square}$, ...
    - waveform of input signal rise and fall times
    - waveform of output signal: load

- rule: The slower the input transitions to a gate, the more energy gets wasted in crossover currents.
- $\sigma_k = 0.05...1.5$
- general guideline:
  - make signal rise and fall times approximately the same
  - make them comparable to the propagation delay of a typical gate from the cell library being used

- **Driving of Resistive Loads:**

$$E_{rrk} = P_{rrk} T_{cp} = \frac{U_{dd}^2}{R_k} \rho_k T_{cp}$$

  - Pure CMOS logic circuits provide no direct current paths from VDD to VSS. Departures (e.g. pull-up, pull-down, PLL, DLL, pad, etc.) exist, however.
  - ρ_k: on-time-ratio

- **Leakage Currents:**

$$I_{lk} = \sum_{chip} (I_{ds\,off} + I_{db\,rev} + I_{bb\,rev} + I_{g\,tun})$$

$$I_{lk} \approx \frac{\Delta I_{ds\,off}}{\Delta W} \cdot \sum_{g=1}^{G} W_g = 10^{\frac{U_{gs\,off} - U_{th}}{S}} \cdot \sum_{g=1}^{G} W_g$$

$$E_{lk} \approx T_{cp} I_{lk}$$

  - $I_{ds\ off}$: leakage currents through subthreshold conduction between drain and source
  - $I_{db\ rev}$: subsurface currents through reverse-biased drain-bulk junctions
  - $I_{bb\,rev}$: leakage currents through reverse-biased well-well and substrate-well junctions
  - $I_{gt\ un}$: leakage current through electron tunneling through the gate oxide (gate leakage)

  - importance: from the 0.25µm technology on
  - subthreshold slope: $S = 70...120\,mV/dec$

- **Total Energy Dissipation:**

$$E_{cp} = \sum_{k=1}^{K} (E_{chk} + E_{crk} + E_{rrk}) + E_{lk} \approx U_{dd}^2 \cdot (1+\sigma) \cdot \sum_{k=1}^{K} \frac{\alpha_k}{2} C_k + T_{cp} \cdot \left(U_{dd}^2 \cdot \sum_{k=1}^{K} \frac{\rho_k}{R_k} + U_{dd} \cdot \frac{\Delta I_{ds\,off}}{\Delta W} \cdot \sum_{g=1}^{G} W_g \right)$$

Designer's Choice

- High-Speed Design: fast circuit operation:
- Low-Power Design: low switching energy per computation cycle, low leakage current
- Low-Activity Design: low static current, problem for devices that operate in short bursts (e.g. wrist watches and alike)
- Other Requirements: reliable operation, compatibility within a logic family, etc.

CMOS Voltage Scaling and Fabrication Process Choice to lower Energy Dissipation

- ever thinner gate oxides: lower supply voltage in order to avoid breakdown but also lower threshold voltage and therefore more subthreshold conduction
- extra circuit for dynamic voltage and frequency scaling

- advance fabrication process
- propagation delay: $t_{pd} \propto \frac{C_k U_{dd}}{(U_{dd} - U_{th})^{\alpha}}$

Counteractive Measures to avoid dynamic Energy Dissipation (see p. 203ff.)

- Activity Reduction at the Algorithm and Architecture Level
- Energy-Efficient Clocking
- Activity Reduction at the RTL and Logic Level
- Cut Down Parasitic Effects at the Electrical and Physical Level

Counteractive Measures to avoid Leakage (see p. 207ff.)

- Fast Architecture built from Low-Leakage Cells
- Variable-Threshold CMOS (VTCMOS): Dynamic Back-Biasing (DBB)
- Multi-Threshold CMOS (MTCMOS)
- Super Cut-Off CMOS (SCCMOS)
- Triple-S Logic
- Virtual Power / Ground Rails Clamp
- Leakage Control Transistors (LECTOR)

Heat Flow and Heat Removal

Junction (Die) Temperature:

$$\theta_j = \theta_a + (R_{th\,jc} + R_{th\,cs} + R_{th\,sa})$$

θ_a : ambient temperature

$R_{th\,jc}$: thermal resistance between junction and IC package

$R_{th\,cs}$: thermal resistance between IC package and heat sink

$R_{th\,cs}$: thermal resistance between heat sink and ambient

- Heat Transfer: conduction, convection and radiation
- Heat Sinks: heat pipes, forced air cooling, heat-conducting compound
- Typical Values:
  - $\theta_j = 85...125\,°C$
  - $R_{th\,jc} = 0.3\,°K/W$
  - $R_{th\,ca} = R_{th\,cs} + R_{th\,sa} = 0.25...0.55\,°K/W$
- Rule: Lowering the thermal resistance generally leads to higher costs!

Contributions to Node Capacitance: see p. 215f.

Absolute Bound for Switching Energy

- Minimum Supply Voltage: $U_{dd} = 0.036V$
- Minimum Energy Dissipation per Cycle: $E_{ch} = 0.036eV$

Physical Design

Design Verification

- Design Rule Check (DRC): exhaustive DRC?
- Layout Extraction and Backannotation
- Layout versus Schematic (LVS)
- Functional Equivalence Checking: HDL checking
- Post-Layout Timing Verification
- Power Grid Analysis
- Signal Integrity Analysis
- Post-Layout Simulation

Design Rule Check (DRC)

aka Design Rules, Layout Rules

- Minimum Width Rules
- Minimum Intralayer Spacing Rules
- Minimum Interlayer Spacing Rules
- Minimum Enclosure Rules
- Minimum Extension Rules
- Maximum Width Rules: for contacts and vias (stipple contact or via)
- Minimum Occupancy Rules: for CMP

Electrical Properties

- Sheet Resistance R_{sh}:

$$R = R_{sh} \cdot \frac{l}{w} \quad \text{with} \quad R_{sh} = \frac{\rho}{h}$$

h : height l : length
w : width

- Some technical Terms:
  - Silicide: compound between Si and more electropositive elements (e.g. $TiSi_2$, $TaSi_2$, ...), for improved electrical connectivity
  - Polysilicide: silicide layer placed over polysilicon
  - Salicide: Self-Aligned Silicide

Interlayer Connection

- Contact: connection between a metal and a silicon layer
- Via: connection between two metal layers
- Stacked Contacts / Vias
- Staggered Contacts / Vias

- Layer Counting: from the bottom (substrate, bulk) to the top
- Properties:
  - *first-level metal layer:* for intracell routing
  - *intermediate metal layer:* for extracell routing
  - *higher-level metal layers:* thicker, withstands higher currents, for distributing of power, clock and other critical nets
- today: Over-the-Cell Routing instead of Routing Channels

Floorplanning and Place & Root

- Guidelines:
  - reflect functional organization
  - greatly determines performance and costs of the final solution
- Floorplanning includes:
  - partitioning into major building blocks
  - number and anticipated sizes, shapes and placement of all such blocks
  - package selection and pin / pad utilization
  - wide busses and electrically critical signals
  - clock domains
  - voltage domains
  - on-chip power and clock distribution schemes
- Building Process:
  - identify the major building blocks: subdivision of the system into blocks and subblocks, on-chip- versus board-level
  - establish a pin budget: corelimited versus padlimited design
  - find a relative arrangement of all major building blocks: respect clock and voltage domains, maximize area efficiency, minimize long busses and wires, minimize total delays and critical paths, use local connections, pay attention to macro- and megacells
  - plan power, clock and signal distribution: pay attention to voltage losses, elctromigration, interconnect delays (RC), series impedances, cross coupling due to capacitances and inductances
  - place and route: reoptimization and rebuffering of synthesized gate-level netlists after initial placement, Engineering Change Order (ECO), “sewing kits” for prototypes
  - chip assembly: place all building blocks, interconnect them, prepare padframe and connect it with the core

Packaging

- What is the Package used for?
  - protection against mechanical stress and other environmental attacks
  - expansion of connector geometry

  - provide electrical connection with the surrounding circuitry with particular emphasis on low impedance for power and ground nets (supply rails)
  - carry away the thermal power while keeping die at an acceptable temperature
  - improve handling during shipping and board assembly
- Package Types: SIP, DIP, DIL, PGA, SOP, SOIC, TSOP, SSOP, LLCC, LCC, QFP, FQFP, SOJ, LDCC, JLCC, LGA FLGA, BGA, FBGA
- Building Process:
  - wafer sorting: Process Control Monitor (PCM)
  - wafer testing: probe card
  - wafer sawing: blue film
  - encapsulation
  - final testing
  - die and wire bonding: terraced package, staggered pads, double bond, sometimes special pins for ground and power, wire bonding rules (see p. 252)
- advanced packaging techniques:
  - flip chip technique with ball grid arrays (bumps): mechanical interposer or direct attachment to the Printed Circuit Board (PCB)
  - folded flexiprints
  - Multi-Chip Module (MCM):[7] System-in-Package (SiP) and tiny Surface Mounted Devices (SMD)
  - chip stacking and cubing
- selecting a packaging technique:
  - available space
  - number of pins necessary
  - geometrical issues of the die and the package
  - electrical characteristics
  - maximum power dissipation, ambient temperature, thermal resistance
  - resistance against mechanical, thermal and other environmental stress
  - expected lifetime, aging and reliability
  - graded product sortiment
  - facility to replacement
  - board mounting technique
  - required equipment for packaging, mounting and testing
  - yield losses due to packaging and mounting operations
  - packaging and mounting costs

7 compare this approach with System-on-Chip (SoC)

Layout in more Detail

- Justification for Manual Layout Aka Rectangle Pushing:
  - very high production volumes
  - library devlopment (e.g. RAMs, ...)
  - specific requirements
  - analog circuit parts (e.g. pad drivers, ...)
  - integrated sensors and actuators (especially MEMS devices)
  - test structures for PCM
- Requirements:
  - minimum area (maximum density)
  - maximum performance
  - maximum fabrication yield
  - avoidance of parasitic effects
- Standard cells for Full-Custom Fabrication
- Sea-of-Gates Macros for Semi-Custom Fabrication
- Manhattan and Boston Geometry
- MOSFET Device fabricated in a twin-well Process: see p. 258ff.
  - structure
  - parasitic devices
  - grid-matrix layout, stick diagram

Electrostatic Discharge (ESD)

- Example: Walking across a synthetic carpet can generate voltages from 100V to 30kV under worst case conditions.
- Two Destructive Effects:
  - dielectric breakdown
  - local overheating often followed by melting
- Counteractive Measures:
  - handling precautions
  - on-chip ESD protection: input, output and supply protection
- on-chip ESD Protection:
  - desired properties: zero on-state resistance, zero area requirement, no parasitics, triggering only at ESD events, absorption of infinite amounts of energy, instantaneous turn-on, clamp voltage just above the operating supply voltage
  - collecting ESD protection circuits in order to save chip area
  - avalanche-triggered snapback BJT, grounded-gate NMOS (ggNMOS): see p. 271f.

Electromigration

- long-time effect
- Attention: Not the same as fusing!
- wear-out phenomenon that affects metal conductors subject to excessive current densities
- disintegration tends to follow lattice imperfections (e.g. Grain boundaries, impurities, dislocations, ...)
- Working Principle:
  - thermally agitated metal ions above one-half of the melting point
  - metal ions are pushed along by the impact of flowing electrons
  - diminishing of the cross section
  - further increase of current density
  - vicious cycle: severing of wires
- Blech Effect: Shorter lines can withstand higher current densities!
- Typical Values for Al: 5 to 10mA/μm^2 (kA/mm^2)

Latch-Up

- Definition: short current path between power and ground
- Working Principle: Thyristor aka Silicon-Controlled Rectifier (SCR), see p.274
- Two Conditions for Enabling Latch-Up:
  - positive feedback
  - current gain of BJTs bigger than unity
- Counteractive Measures:
  - use body ties and butted contacts: guard bar and guard rings for minority and majority carriers (see p.278)
  - keep current amplification of positive feedback loop low: sizing the parasitic BJTs
- Industry Practice:
  - fore core cells: one pair of body ties for every two to eight MOSFETs
  - for input and output pads: four guard ring structure, avoid placing of n- and p-type transistors next to eachother

Geometric Quantities

- Gate Length L_g: $L_{drawn} < L_{eff}$
- Minimum Feature Size: $F = \min(L_{drawn})$
- Minimum Half Pitch: $F = \frac{1}{2}\min(P)$
- Minimum Lithographic Square: F^2
- ***x***nm ***y***M***z***P CMOS process: *y layers of metal plus z layers of polysilicon with x=F*

VLSI Economics and Project Management

Models of Industrial Cooperation

- Systems assembled from Standard Parts alone
  - systems are built from standard components:
    - Small Scale Integration (SSI)
    - Medium Scale Integration (MSI)
    - Large Scale Integration (LSI)
    - Application Specific Standard Product (ASSP)
  - fast turnaround times
  - very difficult business today:
    - less ability to innovation
    - exposed design know-how
    - dependence on IC vendors
    - low integration density
    - high manufacturing costs
    - poor agility
- Systems built around Program-Controlled Processors:
  - systems designed around a microprocessor, a digital signal processor or another component that executes program instructions sequentially
  - software development
  - unlimited agility
  - very fast turnaround times
- Systems designed on the basis of Programmable Logic Devices (PLD):
  - configurable logic is used: Complex PLD (CPLD), Simple PLD (SPLD) and Field Programmable Gate Arrays (FPGA)
  - fairly coherent EDA package
  - virtual components prepared by independent IP vendors
  - absence of time-consuming manufacturing cycle
- Systems designed on the Basis of Semi-Custom ASICs:
  - only a small subset of all layers is custom-made
  - long turnaround times (weeks or months)
- Systems designed on the Basis of Full-Custom ASICs:
  - all layers are custom-made
  - highly specialized partners
  - very long turnaround times (months)

- independent cell library vendors (not part of the foundry design kits)

Handoff Points and Manufacturing Services for ASIC Design Data

1. Algorithm Design	Behavioural Model Handoff: software code with performance- or timing-related constraints
2. Architecture Design	Architecture Handoff: HDL models with performance- or timing-related constraints
3. RTL Design	RTL Handoff: e.g. VHDL or Verilog code
4. Gate-Level Synthesis	Netlist Handoff: e.g. Verilog code
5. Back-End Design	Full-Layout Handoff: GDS II or CIF
6. Chip Finishing: seal ring, alignment marks, ...	
7. Mask or Reticle Preparation	
8. Wafer Processing	
9. Process Control Monitoring (PCM)	
10. Wafer Testing (Probe Card)	
11. Wafer Sawing	
12. Encapsulation	
13. Final Testing	

popular handoff points: at steps 2 or 3

- **Integrated Device Manufacturer:**
  - covers: all steps
  - full testing by the manufacturer with customer support
  - responsibility for yield on behalf of the manufacturer
  - payment: either one-time payment or sales price for each delivered chip, only the correctly working chips are paid!
- **ASIC Manufacturer:**
  - covers: steps 6 to 13
- **Silicon Foundry:**
  - covers: steps 6 to 9
  - testing: limited to PCM, but no chip testing
  - no IC testing and no packaging
  - full responsibility on behalf of the customer
  - ideal for prototype production
- **Design House:**
  - covers: steps 2 to 5 (depends on handoff point)

- acts between customer and manufacturer
- system design starting from different handoff points

- **Customer Owned Tooling (COT):**
  - covers: steps 8 and 9
  - testing: limiting to PCM
  - nearly all responsibility on behalf of the customer
  - ideal for large volume production

Virtual Components (VC)

- HDL models for synthesis, intellectual property modules or IP modules
- Problems:
  - copyright protection
  - quality and usage of VCs
  - service and support
- Business Models:
  - *one-time payment for unlimited usage:* money against technology transfer, more popular
  - *payment for each delivered unit:* royalty fee to the vendor for each unit delivered

Cost of ASIC Implementation

- **Non-Recurring Costs:** do not depend on the quantity of produced items
  - project management
  - circuit specification
  - purchase and assimilation of VCs, if any
  - circuit design
  - design verification
  - preparation of testbenches and vectors for simulation
  - CAE/CAD related expenses
  - Non-Recurring Engineering Costs: sign-off procedure, preparation of masks, preparation of probe cards, preparation for fabrication (setting up production lines and testing equipment)
  - prototype fabrication and testing
  - product qualification: life-cycle test, JEDEC[8] standards, ...
  - redesigns, if any
- **Recurring Costs:** depend on the quantity of produced items
  - semiconductor wafers
  - wafer processing
  - volume testing

8 **J**oint **E**lectron **D**evice **E**ngineering **C**ouncil

- packaging
- royalties for VCs, if any
- board or other substrate space
- external catalogue parts, if any
- assembly

- **Total Costs per Item (board, chip or alike):**

$$c = \frac{c_0}{n} + c_1 \qquad c_1 = c_f + c_t + c_p$$

c_0 : non-recurring expenses
c_1 : recurring expenses
c_f : expenses for fabrication
c_t : expenses for testing
c_p : expenses for packaging
n : amount of working items

Important Rule:

- n small : non-recurring costs c_0 dominate!
- n very large : recurring costs c_1 dominate!

- **Fabrication Costs c_f per Item:**

$$c_f = \frac{c_{\mathrm{wp}}}{y_f n_m} \approx c_{\mathrm{wp}} \cdot \left(1 + \frac{DA_c}{\alpha}\right)^{\alpha} \cdot \frac{A_d}{\pi\left(\frac{d_w}{2} - \sqrt{A_d}\right)^2}$$

$$n_m = \frac{\pi}{A_d} \cdot \left(\frac{d_w}{2} - \sqrt{A_d}\right)^2$$

$$y_f = \left(1 + \frac{DA_c}{\alpha}\right)^{-\alpha}$$

c_{wp} : costs for purchasing and processing of one raw wafer
y_f : fabrication yield
D : defect density (typical values: 0.006...0.012mm^{-2})
α : clustering factor (typical values: 2...3)
A_c : total area of one die minus unused spaces (cut lines and alike)
A_d : total area of one die
d_w : diameter of a wafer

- **Wafer Costs c_{wp} in more detail:**

$$c_{\mathrm{wp}} \approx c_{wr} + n_{ls} c_{ls}$$

c_{wr} : price of a raw wafer and other fixed contributions
n_{ls} : amount of lithography pattern steps
c_{ls} : average cost of a lithography pattern step

- **Fabrication of small Quantities:**
  - Multi-Project Wafer (MPW)
  - Multi-Level Masks: sharing of masks
  - Electron Beam Lithography
  - LASER Programming
  - Hardwired FPGA
  - Cost Trading: manufacturers pay parts of the NRE in exchange for higher item costs

The Market Aspect

- price and costs
- key factors for a successful product:
  - pops up just in time
  - is an improvement rather than revolutionary
  - fits into the company's product portfolio
  - fits into the company's manufacturing strategy
  - reduce costs
- commercialization stages and market priorities: (see p. 314)
  - technology-driven commercialization
  - new-market-driven commercialization
  - product- and process-improvement-driven commercialization
  - end-game commercialization
- customer's expectations: price, operating costs, benefits, ownership, ...
- offering services versus offering products
- product grading

Making a Choice

- Decision Process: see p. 322ff.
- Arguments in favour of ASIC Design:
  - push towards advanced products:
    - reduced parts count
    - improved reliability
    - reduced space requirements
    - package decision
    - superior performance of dedicated hardware
    - interchip optimizations
    - tight control over parasitic circuit elements
    - improved energy efficiency
    - improved innovation
    - protection of know-how
    - tamper-proof
    - opportunity of implementing test strategy
  - push towards cost reduction:
    - reduced assembly costs
- Arguments against ASIC design:

- little flexibility
- long turnaround times
- small expected sales volumes
- need of highly specialized design engineers
- multiple partners involved
- stronger technical challenges and financial risks
- IC experience in a stable environment
- technical compromises
- bad product grading

Keys to successful VLSI Design

see p. 327ff.

Evaluating Business Partners

- Silicon Vendor
- Design House
- Fabrication Process
- Design Kits and Cell Libraries

A Primer on CMOS Technology*

Electrical Conduction of Material

	Insulator	**Semiconductor**	**Metal**
Bandgap	wide (>5eV)	medium (≈1eV)	small (<1eV) or overlapped
Resistivity ρ	$10^{8}\Omega m$	$10^{3}\Omega m$ @ T=20°C	$10^{-6}\Omega m$
Valence Band	filled	filled	partially filled
Conduction Band	empty	empty	empty

Doping of Semiconductor Materials

- n-type material: donors produce additional electrons
- p-type material: acceptors produce additional holes
- syntax: [doping type]$^{\text{doping concentration}}$
  - doping concentration: + strong, - light
  - doping type: n, p

PN-Junctions

- building of depletion region with built-in voltage U_{bi} (in thermal equilibrium)
- metallurgical junction: borderline between p- and n-region
- two antagonistic currents:
  - current due to diffusion (concentration gradient): builds up potential difference
  - current due to drift (electrical field): attenuates potential difference
- operating regions:
  - forward bias: $U_{pn} + U_{bi} > 0$ (U_{bi} typical negative)
  - reverse bias: $U_{pn} < 0$
- Ohmic Contact: contacts between metals and heavily doped semiconductors
- Schottky Junction: contact between metals and lightly doped semiconductors

MOSFET (p. 7ff. VLSI III Script)

- MOS: Metal Oxide Semiconductor
- FET: Field-Effect Transistor
- Operation Regions:

Operation Regions	
• Thermal Equilibrium	• Linear Regime
• Cut-Off	• Pinch-Off
• Weak Inversion	• Saturation
• Strong Inversion	• Super Cut-Off

CMOS[9] Fabrication Flow of a State-of-the-Art 1P3M Twin-Well (Bulk) Process

Front-End-Of-Line (FEOL)	
1. initial wafer: 700-800µm thick, 200 or 300mm diameter, lightly p-doped epitaxial layer 2. n-well formation: positive photoresist as protection layer 3. p-well formation: same mask as in previous step (negative photoresist in use) 4. active area definition: active areas covered with nitride layer 5. **Shallow Trench Isolation (STI)**: dry etching of deep trenches (450nm), nitride layer as protection layer, trenches filled with SiO_2, **Chemical Mechanical Polishing (CMP)** 6. gate stack formation and patterning: delicate oxidation step (4-5nm), followed polysilicon film (200nm), dry etching	7. n-channel source / drain extensions: **lightly doped protractions**[10] called Lightly Doped Drains (LDD), for better control of short channel effects, gate as shield (**self-aligned gate**) 8. p-channel source / drain extensions 9. sidewalls or oxide spacers: oxide layer deposited and etched away in order to form an insulating wall on either face of the poly gate 10. n-type doping: sidewall spacers as protection, positive photoresist as protection layer 11. p-type doping: same mask as in previous step (negative photoresist in use) 12. salicidation: self-aligned silicide (salicide), highly conductive film
Back-End-Of-Line (BEOL)	
13. first interlevel dielectric: SiO_2 layer, CMP planarization 14. contact plug formation: **tungsten plugs** 15. deposition and patterning of first metal layer (**subractive metallization**) 16. second interlevel dielectric 17. via plug formation: tungsten plug 18. deposition and patterning of second metal layer	19. third interlevel dielectric followed by via plug formation 20. deposition and patterning of third metal layer: thicker (0.9µm) and larger minimum width and spacings required as for lower layers 21. overglass and bond pad openings: final layer of silica with etched openings for the pads note: two additional masks per metal layer needed

9 Complementary Metal Oxide Semiconductor
10 in German: Hinausziehen

Photolithography

- masks made of glass, quartz and fused silica sputtered with Cr
- Big Problem:
  - Quartz and fused silica materials is opaque below 195nm.
  - Shorter wavelengths are absorbed more strongly and reflected poorly by most materials.
- Traditional Optical Lithography: E-, G-, H- and I-lines (546-364nm)
- Deep UV Lithography:
  - KrF_6: 248nm
  - ArF_6: 193nm

Process Control Monitors (PCM)

Resolution Enhancement Techniques (RET)

- Phase Shift Masks (PSM)
- Optical Proximity Correction (OPC)
- Computerized Resolution Enhancement: PSM and OPC
- Immersion Lithography
- Post-Optical Lithography (Next-Generation Lithography):
  - Extreme UV Lithography (EUV)
  - Electron Beam Lithography: Electron Beam Direct Write (EBDW), Electron Projection Lithography (EPL)
  - Imprint Lithography

New Technologies

- Silicon On Insulator (SOI):

Pros:	*Cons:*
• no wells necessary • no parasitic BJTs therefore no latch-up • no need for body-ties • superior layout density • reduced sensitivity to radiation and higher operating temperatures • faster operation and / or better energy efficiency • smaller junctions surfaces reduces parasitic source and drain capacitances	• floating body effect: trapped charges in the channel • counteractive measures: extreme thin silicon layer: Ultra-Thin-Body SOI (UTBSOI) or Depleted Substrate Transistors (DST)

- Strained Silicon and SiGe Technology:
  - tensile stress: stress due to the expansion of lattice structure

- compressive stress: stress due to the compression of lattice structure
- two sources for stress: incorporating large Ge atoms into the narrower Si lattice, applied mechanically in the packaging process

• New Interconnect Materials and Interlevel Dielectrics:

- requirements for interconnect materials: low RC delay, less electromigration, reduced vulnerability to corrosion (passivation layer needed), etc.
- requirements for interlevel dielectrics: low ε_r, withstand stresses for CMP and wire bonding, process compatibility, thermal stability, low moisture absorption, etc.
- new materials:

	today and in the past:	*future:*
Interconnect Material	Al	Cu in a damascene process (sealed with an extra liner acting as diffusion barrier)
InterLevel Dielectric (ILD)	SiO_2	organosilicate glasses, organic synthetics, enclosure of air bubbles

Technology Outlook*

Moore's Laws

- *First Law:* The capacity of DRAMs quadruples approximately every three years.
- *Second Law:* The capital requirements for a DRAM fab grow by a factor of 1.8 or so over the three years that separate one memory generation from the next.

Sturtevant's Law

According to experts in the field, optical lithography has always been anticipated to come to an end six or seven years into the future.

Two important Directions for the Future

- Moore's law will slow down and eventually come to a standstill by the time atomic and quantum scales are approached.
- Moore's law will remain in place, but only by shifting over to some radically different kind of technology.

International Roadmaps for Semiconductors (ITRS)

Market Aspects

- General Aspect:
  - How much computing and communication does the world really need? (amount)
  - How much is it prepared to pay for it? (price)
- Expectations and Wishes from the Customers:
  - ease of use
  - safety and comfort
  - economic advantages
  - social status
  - fun and entertainment
- Big Problem: rising Non-Recurring Costs (NRC) of IC manufacturing

Scaling of CMOS Technology

- Motto: smaller, faster, cheaper
- Point of Time for Technology Change:
  Expected Production Volume $\times + \Delta$ Savings for Fabrication of a Die...

  ...with new Process Technology $\geq + \Delta$ new Fabrication Acquisition Costs
- Robert Dennard in 1972: voltage levels will shrink linearly to maintain constant electric fields, benefits for gate delay and energy efficiency

Device Scaling

- Smaller Gate Delays: $t_{pd} \propto \frac{C_k U_{dd}}{I_{ds\,on}}$
- Higher Current Drive: $I_{ds\,on} \propto \frac{\mu \epsilon_{ox}}{t_{ox}} \frac{W}{L_g} (U_{dd} - U_{th})^{\alpha}$
- Lower Dissipated Energy per Switching Event: $E_{chk} \propto C_k U_{dd}^2$
- Low Overall Leakage Current: $I_{lk} \propto 10^{\frac{-U_{th}}{S}}$

Scaling Facilities I						
	L_g	t_{ox}	U_{dd}	U_{th}	x_j	N_a, N_d
typical value	100nm	3nm	1.5V	0.4V	40nm	$10^{18}cm^{-3}$
far future	10nm	0.5nm	0.6V	0.2V	5nm	$10^{19}cm^{-3}$
scaling trend	↓	↓	↓	↓	↓	↑
minimum leakage	↑			↑		
high current drive	↓	↓	↑	↓		
small gate delay		↑ (C_k ↓)	↓			
low switching energy		↑	↓			

Scaling Facilities II		
	prime motivations:	*limiting factors:*
L_g	• higher layout density • higher current drive • higher switching speed	• cost of lithography • short channel effects
t_{ox}	• maintain current drive • maintain gain factor	• reliability[11] • gate dielectric tunneling
U_{dd}	• improve energy efficiency • avoid gate dielectric breakdown	• threshold voltage • delay and ramp times
U_{th}	• maintain current drive	• off-state leakage
x_j	• suppress short channel effects	• series resistance
N_a, N_d	• scale depth of depletion layer	• loss of source and drain isolation

11 meaning here: Beständigkeit

Technology for the Future

- Dielectric Interconnect Materials: low ϵ_r
- Dielectric Gate Materials: high ϵ_r
  - capacity: $C = \frac{\epsilon_r \epsilon_0}{d}$
  - tunneling problems leads to leakage
  - Equivalent Oxide Thickness (EOT): $EOT = t_{ox} \cdot \frac{\epsilon_{SiO_2}}{\epsilon_{ox}}$
  - new materials: Hafnium Oxide (HfO_2), Strontium Titanate ($SrTiO_3$), Zirconium Oxide (ZrO_2)
- Rebirth of Metal Gates:
  - Problems with Polysilicon: penetration of dopens, depletion effect
  - Possible Solution: poly-Si/HfO_2/Si
  - Problems with Metal Gates: no threshold tuning and (no self-aligning process)
- Vertical Integration:
  - Chip Stacking or Cubing Tap
  - Problem: heat evacuation
- New Device Topologies:
  - Double-Gate MOSFET (DG-MOSFET)
  - fin-FET
  - gate-all-around transistor (surround gate transistor, pillar FET)
- New Semiconductor Materials:
  - High Mobility Semiconductors: Germanium On Insulator (GOI), Gallium Arsenide (GaAs), Indium Phosphide (InP), Silicon Germanium (SiGe)
  - Wide Bandgap Semiconductors: Silicon Carbide (SiC), Gallium Nitride (GaN)
  - Metallic Materials: graphitic films (graphenes)
  - Polymer Semiconductors: plastic circuits, OLED, RFID
- Data Storage: Non-Volatile RAM Types
  - Flash RAM: Non-Volatile-RAM (NOVORAM)
  - Phase Change RAM (PRAM): Ovonic Unified Memories (OUM)
  - Ferroelectric RAM (FeRAM)
  - Magnetic RAM (MRAM)
- Data Processing:
  - Carbon Nanotubes
  - Nanojunctions
  - Molecular Electronics
  - Crossbar Logic

- Magnetic Flux Quantum Device: Josephson Junction, Rapid Single Flux Quantum (RSFQ)
- Quantum Cellular Arrays: quantum dots in a square
- other Quantum Devices
- Quantum Computing

Circuit Design Methodology for the Future

- Productivity in the Design Process:
  - High-Level Synthesis Tools: incremental design (model year model) or Virtual Components (CV)
  - On-Chip Firmware Facility
  - Formal Verification
  - Reliability and Fault Tolerance: redundant hardware, self diagnosis, error correction, self programming and self replication
- Architecture Design:
  - Domination of Interconnect Delay: wire planing
  - System on a Chip (SoC) versus Multi-Chip Module (MCM)
- Circuit Testing:
  - more sophisticated fault models
  - sophisticated current handling: switching current, quiescent current, static leakage current
  - costs not allowed to grow with circuit complexity
  - testing of circuits with high clocking frequencies