



Πολυτεχνείο Κρήτης Τμήμα Ηλεκτρονικών Μηχανικών και
Μηχανικών Υπολογιστών

Σχεδίαση CMOS image sensor για βιοιατρικές εφαρμογές

Στεφανάκης Διονύσιος
Α. Μ. 2000030118

Μέλη της Εξεταστικής Επιτροπής:
Επίκουρος Καθηγητής Ματτίας Μπούχερ (Επιβλέπων)
Αναπληρωτής Καθηγητής Κώστας Μπάλας
Καθηγητής Μηχάλης Ζερβάκης

Χανιά, 17 Ιουλίου 2008

English:

Technical University of Crete

Subject: Design of CMOS image sensor for medical applications
by Stefanakis Dionysios

Professors: Matthias Bucher, Costas Balas, Michalis Zervakis

Περίληψη

Παρουσιάζεται η προμελέτη για την σχεδίαση ενός αισθητήρα εικόνας σε τεχνολογία CMOS (CMOS image sensor).

Ένας CMOS image sensor αποτελείται από τα εξής στοιχεία:

- 1) Φωτοευαίσθητο υλικό συμβατό με πυρίτιο
- 2) Διακόπτες (Reset, Select)
- 3) Ακόλουθος τάσης (Source follower)
- 4) (CDS)
- 5) ADC

Η επιλογή φωτοευαίσθητου υλικού γίνεται με βάση κριτηρίων που εξάγονται μαθηματικά όπως: Φωτορεύμα, απόκριση χρόνου, περιοχή απογύμνωσης, χωρητικότητα, “quantum efficiency”, “responsivity” και θορύβου. Επίσης, μελετάται η διαφορετική συμπεριφορά του φωτοευαίσθητου υλικού υπό διαφορετικές συνθήκες φωτισμού.

Έχοντας επιλέξει ένα κατάλληλο φωτοευαίσθητο υλικό, μελετάται το κύκλωμα, το οποίο έχει την καλύτερη απόδοση, σχετικά για την εξαγωγή της τιμής του φωτοσήματος. Παρουσιάζονται βασικά κυκλώματα εξαγωγής τιμής του φωτοσήματος, τα λεγόμενα pixel circuits. Τα κυκλώματα εξαγωγής κατηγοριοποιούνται σε CCD και CMOS pixel. Για το κύκλωμα εξαγωγής CMOS APS, μελετάται η χρονική απόδοση για την εξαγωγή του φωτοσήματος. Συγκεκριμένα υπολογίζεται η χρονική απόδοση για τρεις τρόπους εξαγωγής του φωτοσήματος. Οι τρεις τρόποι για την εξαγωγή ενός φωτοσήματος είναι οι ακόλουθοι: εξαγωγή του φωτοσήματος σαν φορτίο, εξαγωγή του φωτοσήματος σαν ρεύμα και εξαγωγή του φωτοσήματος σαν τάση. Ο χρόνος εξαγωγής που προκύπτει καθορίζει τον χρόνο έκθεσης του φωτοευαίσθητου υλικού στον φωτισμό. Έχοντας επιλέξει τον τρόπο εξαγωγής του φωτοσήματος συνεχίζουμε με τον σχηματισμό πλέγματος από pixel (pixel array).

Η τιμή του φωτοσήματος είναι αναλογική και θα πρέπει να μετατραπεί σε ψηφιακή. Η μετατροπή αυτή επιτυγχάνεται με τα ADC. Η τοποθέτηση των ADC πραγματοποιείται είτε στα pixels, είτε στο κάτω μέρος της στήλης εξόδου του pixel είτε ξεχωριστά για όλο τον αισθητήρα. Κάθε τρόπος τοποθέτησης απαιτεί συγκεκριμένα χαρακτηριστικά από τον ADC. Για αυτόν το λόγο παρουσιάζουμε διάφορα ADC και τα χαρακτηριστικά τους. Τα ADC κατηγοριοποιούνται σε Nyquist και σε oversampling. Τα Nyquist ADC σε συνδυασμό με το κύκλωμα εξαγωγής CMOS APS απαιτούν ένα πρόσθετο κύκλωμα, το CDS.

Ο αισθητήρας μπορεί να καταστραφεί στην περίπτωση που ένας απλός CMOS APS χρησιμοποιείται στο τομέα της βιοιατρικής, όπου χρησιμοποιούνται ακτίνες X. Μελετώντας άλλα συστήματα απεικόνισης στην βιοϊατρική παρατηρούμε την χρήση ενός υλικού τύπου scintillator, το οποίο προστατεύει τον αισθητήρα από καταστροφή. Αυτό το υλικό έχει την ιδιότητα να μετατρέπει τις ακτίνες X σε ορατό φως (μήκος κύματος μεταξύ 400 και 650 nm). Έτσι αναζητήσαμε ένα υλικό τύπου scintillator, που να είναι συμβατός με το φωτοευαίσθητο υλικό μας. Η μελέτη έδειξε, ότι ένα κατάλληλο υλικό είναι το CSI:Tl. Επίσης αποδείξαμε, ότι ο scintillator είναι συμβατός με την CMOS τεχνολογία και ότι χρειάζεται μόνο μια επιπλέον επίστρωση για την ενσωμάτωση του στην CMOS τεχνολογία. Τέλος, υλοποιήσαμε ένα CMOS APS 2x2 pixel σύστημα σε σχηματικό επίπεδο με το πρόγραμμα CADENCE.

Table of contents

CHAPTER 1. INTRODUCTION.....	7
CHAPTER 2. OPTICS AND SEMICONDUCTOR.....	9
2.1 INTRINSIC AND EXTRINSIC SEMICONDUCTORS	9
2.1.1 Fermi-Dirac.....	9
2.1.2 Density of states.....	9
2.1.3 Carrier concentrations.....	9
2.1.4 Conductivity	11
2.1.4.1 Total current	12
2.1.5 Recombination-Generation of carriers.....	13
2.1.6 The continuity equation	16
2.1.7 Photoconductivity	17
2.1.7.1 Constant light intensity.....	19
2.1.7.2 Impulse response	20
2.1.7.3 Sinusoidal steady-state response	20
2.2 FORMING JUNCTIONS	21
2.2.1 Pn Junction.....	22
2.2.1.1 Zero Bias (Equilibrium)	22
2.2.1.2 Depletion region	24
2.2.1.3 Bias Cases.....	25
2.2.1.3.1 Forward bias.....	25
2.2.1.3.2 Reverse bias	26
2.2.1.4 Generation-recombination in pn junction.....	27
2.2.1.4.1 Forward bias.....	27
2.2.1.4.2 Reserve bias	30
2.2.1.4.3 Dark current	31
2.2.1.5 Photocurrent in reverse bias	31
2.3 REFERENCES.....	34
CHAPTER 3. PHOTODETECTORS BASED ON SILICON	35
3.1 NOISE SOURCES	35
3.2 BULK SEMICONDUCTOR	36
3.3 DIODES	39
3.3.1 PN Junction.....	39
3.3.2 PIN diode	43
3.4 BIPOLAR PHOTOTRANSISTOR	46
3.5 MOS CAPACITOR (PHOTO-GATE).....	47
3.5.1 Depletion approximation for MOS	52
3.5.2 Photogate (PG).....	55
3.6 REFERENCES.....	57
CHAPTER 4. PIXEL AND ARRAY IMPLEMENTATIONS.....	58
4.1 PHOTOGATE BASED PIXEL ARCHITECTURES (CCD)	58
4.1.1 Inline transfer CCD.....	59
4.1.2 Frame transfer CCD.....	60
4.1.3 Full frame CCD	60
4.1.4 Properties of CCD.....	60
4.1.4.1 Quantum efficiency	61
4.1.4.2 Charge collection efficiency.....	61

4.1.4.3 Charge transfer efficiency	61
4.1.4.4 Noise.....	61
4.1.4.5 Response speed.....	62
4.2 PHOTODIODE BASED PIXEL ARCHITECTURES (CMOS).....	62
4.2.1 Active pixel (APS).....	63
4.2.1.1 (PD) Photodiode type APS.....	64
4.2.1.2 (PG) Photogate type APS	64
4.2.1.3 Logarithmic APS.....	64
4.2.1.4 CTIA APS pixels.....	64
4.2.1.5 (PPD) Pinned photodiode pixel.....	65
4.2.1.6 TFA pixels	65
4.2.1.7 (CAPS) Complementary active pixels.....	65
4.2.2 APS readout methods	65
4.2.2.1 Charge-mode pixels.....	65
4.2.2.2 Dynamic range	69
4.2.2.3 Current-mode pixels	69
4.2.2.4 Voltage-mode pixels.....	70
4.2.2.5 Unified model of pixel information rate.....	71
4.2.3 Pixel systems	72
4.2.3.1 PPS array system	72
4.2.3.2 APS array systems.....	72
4.2.3.2.1 Chip-level ADC	73
4.2.3.2.2 Column-level ADC	73
4.2.3.2.3 Pixel-level ADC	74
4.3 REFERENCES	74
CHAPTER 5. ANALOG TO DIGITAL CONVERTERS	75
5.1 NYQUIST RATE CONVERSION	77
5.1.1 Quantization.....	78
5.1.2 Nyquist ADC Errors	79
5.1.2.1 Quantization error.....	79
5.1.2.2 Differential Nonlinearity (DNL)	79
5.1.2.3 Missing Codes	79
5.1.2.4 Integral Nonlinearity (INL)	80
5.1.2.5 Offset and Gain Error	80
5.1.2.6 Aliasing	80
5.1.3 ADC architectures	82
5.1.3.1 Successive Approximation	82
5.1.3.2 Flash	84
5.1.3.2.1 Pipeline A/D Converters	85
5.1.3.3 Integrating ADC	86
5.1.3.3.1 Single slope.....	87
5.1.3.3.2 Dual slope	87
5.1.4 Bandwidth and Resolution Relationship.....	88
5.1.5 Designing A/D Performance.....	89
5.1.6 Performance Constraints.....	90
5.2 OVERSAMPLING ADC	91
5.2.1 PCM converter.....	91
5.2.2 PWM converter.....	92
5.2.3 Modulation ADC's	93
5.2.3.1 Delta modulation	93

5.2.3.2 Sigma delta Modulation	94
5.2.3.2.1 Sigma integrator (DAI)	95
5.2.3.2.2 Modulation and noise shaping	98
5.2.3.2.3 Decimating and Filtering	102
5.2.3.2.3.1 Averaging	102
5.2.3.2.3.2 Decimation.....	103
5.2.3.2.3.3 Sigma delta decimator	108
5.2.3.2.3.4 Overview of the sigma delta modulation and modulator	109
5.3 REFERENCES.....	111
CHAPTER 6. DENTAL RADIOGRAPHY WITH CMOS APS.....	112
6.1 MEDICAL IMAGING	112
6.1.1 Scintillator in medical imaging.....	114
6.2 CMOS AND SCINTILLATOR	118
6.3 SENSOR IMPLEMENTATION	121
6.3.1 CMOS APS design	122
6.3.1.1 Switch design	127
6.3.1.2 Source follower	129
6.3.2 CDS design	130
6.3.2.1 Fixed Pattern Noise (FPN)	131
6.3.3 Pixel and Readout testing	133
6.3.4 2x2 Pixel	137
6.4 REFERENCES.....	154
CHAPTER 7. CONCLUSIONS	156
APPENDIX A	158
A.1 LIGHT AND CRYSTAL OPTICS	158
A.2 ELECTROMAGNETIC WAVE PROPAGATION	158
A.2.1.1 Refraction index	161
A.2.1.2 Wave function	161
A.2.2 Energy of a wave	164
A.2.3 Propagation in isotropic media	165
A.2.4 Reflection, Refraction and Transmissivity.....	167
A.3 CRYSTAL CONSTRUCTION AND SEMICONDUCTORS	168
A.3.1 Bravais lattices	168
A.3.2 Miller indices	171
A.3.3 The Reciprocal Lattice.....	172
A.3.4 Wigner-Seitz Cell	173
A.3.5 Brillouin Zones	174
A.3.6 Particle theories.....	176
A.3.6.1 The E-k diagram (Dispersion curve)	178
A.3.6.2 Bragg's Law	182
A.3.6.3 Bloch's Theorem	183
A.4 REFERENCES	184
APPENDIX B.....	185
B.1 INTRODUCTION TO THE Z-TRANSFORM.....	185
B.1.1 Laplace relationship	188
REFERENCES.....	189
APPENDIX C	190

C.1 TRANSISTORS AND MODELS.....	190
C.1.1 Bipolar transistor (BJT)	190
C.2 MOS TRANSISTOR (MOST)	191
C.2.1 Charge sheet models	193
C.2.1.1 Complete charge sheet model.....	193
C.3 BSIM MODELS	198
C.3.1 Berkeley Short-Channel Igfet Models (BSIM's)	198
C.3.1.1 BSIM model.....	198
C.3.1.2 BSIM2.....	201
C.3.1.3 BSIM3.....	201
C.3.1.4 BSIM4.....	212
C.3.1.4.1 Basic parameter list of the BSIM4.2.1model	213
C.4 REFERENCES	217

Page intently left blank

Chapter 1. Introduction

The chapters in this assertion are organized in entities that describe all the necessary different parts, from the photodiode to the ADC, which a CMOS APS system consists of. In chapter 6 an overview of the CMOS APS system and an application is presented. A description of the chapter's content follows:

- **Chapter 2** reintroduces the reader to familiar and well known facts of the junction. Mainly diodes and their interaction with light are analyzed. The photocurrent as well the capacitance of a diode is extracted. Readers who are unfamiliar with the basic physic concepts are encouraged to read appendix A.
- **Chapter 3** explores and outlays all available photosensitive devices based on silicon that may be a candidate for a sensor. All devices are analyzed based on their characteristics which are noise contributions, sensing speed and sensing amount.
- **Chapter 4** employs the best and most common photosensitive devices, discussed in chapter 3, in a pixel scheme. The photosensitive characteristics are not enough to ensure a satisfying photosignal accommodation. For that reason various pixel implementations exist and they in turn depend on the method on how the photosignal is read out. The read out methods distinct CDD from CMOS based sensors. The chapter begins by introducing the CDD pixels and their common pixel arrays implementations. Afterwoods CMOS based pixel implementations are presented. The readout methods and characteristics of the APS and PPS pixel scheme are analyzed. Three different ways (charge, current and voltage mode) of reading the photosignal are inspected. Lastly pixel arrays of APS and PPS and the necessary requirents on the analog to digital convertes are presented.
- **Chapter 5** presents the analog to digital converters (ADC's). ADC's are categorized into Nyquist and oversampling converters. To understand ADC's, an introduction into digitization of analog signals is made. A first look into the conversation errors in the digitation process gives the reader the intuition on the problems that might occur in an ADC archicecture. Hence the Nyquist the errors that might occur in Nyquist converters are presented. Nyquist converters errors that also exist in Oversampling converters. Therefore the reader has gained a general overview on what to pay attention when dealing with ADC's. Nyquist ADC architectures are presented next. If designing a sensor the ADC errors that may exsit in spefic architecture is important but also of importance is the size and components used. For that reason we also present the implementation of the architecture. Each of the presented Nyquist arhcitecures is designed to function for certain speed of conversation. In conjunction with chapter 4 a suitable ADC can be chosen. In the oversampling architecture section we present the basic idea behind oversampling. Lastly one off the most common and best ADC architecture is presented the Sigma delta ADC.
- **Chapter 6** introduces to the reader to medical imaging. X-ray imaging as well as associated common x-ray sensors is presented. Scintillator placed on top of the sensors is the main method to acquire x-ray images. Therefore the scintillator

material is examined. An analysis of scintillator compatibility with the CMOS process follows, which shows satisfactory results. The CMOS APS pixel design procedure is introduced. The design includes the components used in the pixel as well the components used for noise suppression outside the pixel (CDS). Finally a first implementation on the schematic level is done.

Chapter 2. Optics and Semiconductor

2.1 Intrinsic and extrinsic semiconductors

To study the electrical conduction we concentrate our analysis to the transfer of particles from the valence to the conduction band. First we must find the number of electrons or holes in the conduction or valence band. To do this we use two fundamental theories, the Fermi-Dirac and the density of states.

2.1.1 Fermi-Dirac

It's a statistical view that describes the electrons or holes in a solid. In particular the probability that an energy level E (state) is been occupied is given by

$$F(E) = \frac{1}{e^{\frac{(E-E_f)}{kT}} + 1} \quad (2.1)$$

where k is the Boltzmann constant and E_f is Fermi energy. E_f is a guideline that indicates that energies below E_f will be fully occupied and energies above E_f will be completely empty. A more closer investigation of $F(E)$ shows that for all temperatures T , $F(E)$ is ranging from 1 as maximum to 0 as minimum and has an average value of 0.5.

2.1.2 Density of states

We found the probability of states, but how many states are there? The answer lies in determining the density per unit volume in the material $Z(E)$. As we analyzed in the preceding chapter energy is the equivalent of momentum in space. The latter as showed is depended of the wave's property in the crystal. Therefore defining a function, known as the density function $Z(E)$, which takes in account the momentum and energy, we can determine the density. For this the effective mass concept is incorporated into the density function. The effective mass describes the electron movement through the crystal. Without going into much detail, it has been found that electrons as well as holes behaves similar to a free particle in space, except that it's mass is different:

$$m^* = \hbar^2 \left[\frac{d^2}{dk^2} \right]^{-1} \quad (2.2)$$

Then

$$Z(E) = 4\pi \left(\frac{2m_e^*}{\hbar^2} \right)^{\frac{3}{2}} E^{\frac{1}{2}} \quad (2.3)$$

2.1.3 Carrier concentrations

If $F(E)$ represents the possibility of a state being occupied then $1-F(E)$ represents the possibility of a state not being occupied. In other words $F(E)$ stands for electrons where as $1-F(E)$ stands for holes. Said this we are now ready to calculate the number of

electrons, also known as the carrier concentration. Integrating (2.1) and (2.3) over all the conduction band E_c we get

$$n = \int_{E_c}^{\infty} F(E)Z(E)dE \quad (2.4)$$

$$n = \int_{E_c}^{\infty} \frac{1}{e^{\frac{(E-E_f)}{kT}} + 1} 4\pi \left(\frac{2m_e^*}{\hbar^2} \right)^{\frac{3}{2}} E^{\frac{1}{2}} dE \quad (2.5)$$

Defining the range $E'=E - E_c$ and $E_c \geq E > E_f \Rightarrow E - E_c \geq E_f - E_c$ gives us for electrons (n-type)

$$n = 2 \left(\frac{2\pi m_e^*}{\hbar^2} \right)^{\frac{3}{2}} e^{\frac{(E_f - E_c)}{kT}} \quad (2.6)$$

And for holes (p-type)

$$p = 2\pi \left(\frac{2\pi m_h^*}{\hbar^2} \right)^{\frac{3}{2}} e^{\frac{(E_v - E_f)}{kT}} \quad (2.7)$$

When we say the material is n-type, it is meant that the majority carriers are electrons and minority carriers are holes. From appendix A it is obvious due to the mechanism that

$$n_i^2 = np \quad (2.8)$$

where n_i^2 is called the intrinsic carrier concentration. A semiconductor is called intrinsic when the mechanism for creating conduction particles is mainly due to the undoped material. In extrinsic materials the carrier concentrations can be increased by introducing compactable dopand's in the intrinsic silicon. These dopand's come mainly from the III or IV group of the periodic table. The mostly used are Phosphorus and Boron, and are called donors and acceptors. The composites SiP and SiB have then a special property: A weak link exists between the atom and the “unnecessary” created ions. By “unnecessary” it is meant, that the electron can freely roam in the neighborhood of the Phosphorus atom and the hole can attract easily an electron from adjoined atoms, see figure 2.1. Therefore Phosphorus increases electrons movements and Boron increases holes movements. As a result only a slight change in temperature is enough to generate conductivity, this can also be seen from how close to the conduction/valence energy band the dopant energy levels are.

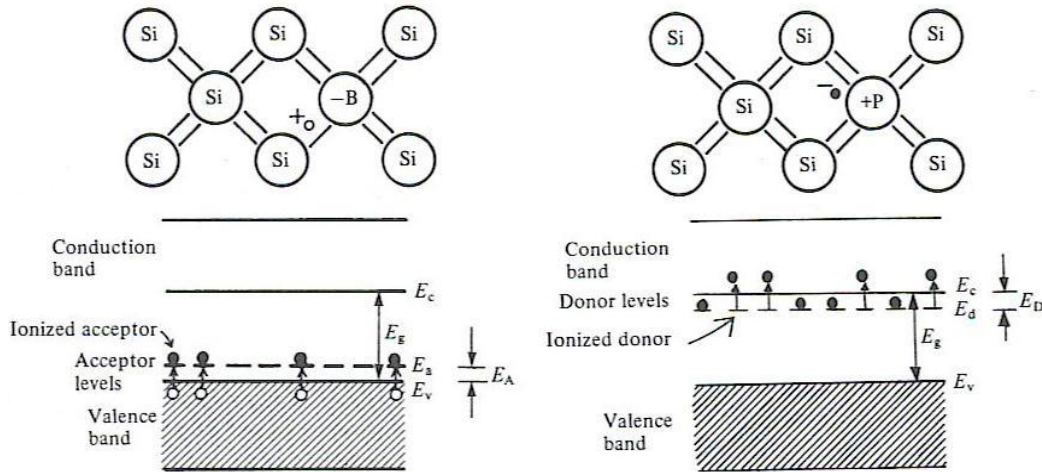


Figure 2.1 Composites a) acceptor

Figure from [1]

b)donor

If we define the common term in (2.6) and (2.7) as $N_c = 2 \left(\frac{2\pi m_e^*}{h^2} \right)^{\frac{3}{2}}$, we can express the carrier concentration in the extrinsic semiconductor as

$$n = N_D = N_c e^{\frac{(E_f - E_c)}{kT}} \quad (2.9)$$

$$p = N_A = N_c e^{\frac{(E_v - E_f)}{kT}} \quad (2.10)$$

(2.9) and (2.10) can also be expressed with the intrinsic carrier concentration defining

$$e^{\frac{(E_f - E_c)}{kT}} = e^{\frac{(E_f - E_i)}{kT}} e^{\frac{(E_i - E_c)}{kT}} \text{ and } n_i = n \text{ when } E_f = E_i$$

$$n = N_D = n_i e^{\frac{(E_f - E_i)}{kT}} \quad (2.11)$$

$$p = N_A = p_i e^{\frac{(E_i - E_f)}{kT}} \quad (2.12)$$

2.1.4 Conductivity

When an electric field is applied (battery for instance) a force $-qE_x$, which stops the random movements of electrons due to temperature, and a net movement towards the anode can be observed. In general random movements of electrons and holes exist but their net current production is zero. If n is the electron density and p_x is the momentum along the direction x we have

$$-nqE_x = \frac{dp_x}{dt} \quad (2.13)$$

One might think that momentum is constantly increasing but taking in account the collision with defects and phonons, the momentum reaches a steady state. Surviving electrons that don't collide in a time span dt , form the electron concentration n , which can be written as:

$$-\frac{dn(t)}{dt} = \frac{1}{\tau} n(t) \Rightarrow n(t) = n_0 e^{-\frac{t}{\tau}} \quad (2.14)$$

where τ is called mean free time and with that one can calculate the probability of an electron coalition. The latter is equal to $\frac{dt}{\tau}$. In (2.11) we found the total momentum of the electrons traveling via electric field, taking into account the coalition probability we get:

$$-\frac{p_x}{\tau} - nqE_x = 0 \quad (2.15)$$

Averaging momentum for a single electron $\overline{p_x} = \frac{p_x}{n}$ we get:

$$\overline{p_x} = -q\tau E_x \quad (2.16)$$

Therefore a constant net drift velocity exists:

$$u_D = \frac{\overline{p_x}}{m_e^*} = \frac{-q\tau E_x}{m_e^*} \quad (2.17)$$

Concentration of electron n is defined per unit area hence $A=1$ and because A being unity we get that the current density is

Semiconductor	μ_e (cm^2/Vs)	μ_h (cm^2/Vs)
Si	1350	480
Ge	3900	1900
GaAs	8500	480
InP	4600	150
GaP	450	150
InAs	3300	460
CdTe	1050	100

Figure 2.2 Some mobility values for various semiconductors. Values from [2].

$$J = -nqu_D \quad (2.18)$$

Combining (2.14) and (2.15) we rediscover Ohm's law

$$J = \sigma E_x \quad (2.19)$$

where $\sigma = \frac{nq^2\tau}{m_e^*}$ the electron

conductivity and it depends mainly on the mobility defined as

$$\mu_e = \frac{q\tau}{m_e^*} \quad (2.20)$$

The same procedure holds for holes and the total drift current through the semiconductor is then

$$J = (nq\mu_e + pq\mu_h)E_x = \sigma E_x \quad (2.21)$$

2.1.4.1 Total current

Drift current is due to electric field and hence no current should exist without electric field. The truth is there is another activity of current present called the diffusion current. To find out more about this activity we remind our self's that in the semiconductor construction we didn't mention any anomalies that may happen in the creation of such material. In the real world carrier concentrations can be affected by defects or by natural disturbance of the carrier concentration by introducing dopants, like we mentioned earlier. This means that no matter by what anomaly, the carrier concentration isn't uniformly spread in the semiconductor. The result is a non-uniform carrier concentration, which causes a current flow analogous with carrier distribution. The latter is been characterized by a varying gradient concentration of electrons along the semiconductor length l . Therefore we make different approach to calculate the current density. We estimate the average rate of electrons R . The estimation is made by setting a cross-section and measure the flow of electrons to the right and to the left of it.

$$R_{right} = \frac{1}{2}n(l)\frac{l}{\tau} = \frac{1}{2}n(l)u \quad (2.22)$$

$$R_{left} = \frac{1}{2}n(-l)\frac{l}{\tau} = \frac{1}{2}n(-l)u \quad (2.23)$$

The total flow R is then

$$R = R_{left} - R_{right} = \frac{1}{2}u(n(-l) - n(l)) \quad (2.24)$$

The density function $n(x)$ can be approximated by a Taylor series neglecting the higher order derivatives we get:

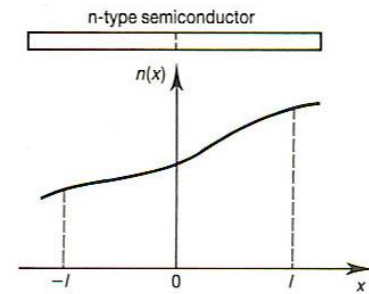


Figure 2.3 Non uniform carrier concentration. Figure from [2]

$$R = \frac{1}{2}u \left[\left(n(0) - l \frac{dn}{dx} \right) \left(n(0) + l \frac{dn}{dx} \right) \right] = -D_n \frac{dn}{dx} \quad (2.25)$$

where $D_n = lu$ is the electron diffusivity, also called the diffusion coefficient, or diffusion constant. The same coefficient D_p can be extracted for a p-type semiconductor. The coefficient D in general can be expressed otherwise by noting that by kinetic theory one can write

$$\frac{1}{2} m_e^* u^2 = \frac{1}{2} kT \quad (2.26)$$

Substituting $\frac{l}{\tau} = u$ and by (2.18) we get

$$D_n = lu = \frac{kT}{q} \mu_e \quad (2.27)$$

$$D_p = lu = \frac{kT}{q} \mu_h \quad (2.28)$$

The last two equations are called the Einstein relations. We are now in position to write the current density as

$$J_n = q D_n \frac{dn}{dx} \quad (2.29)$$

$$J_p = -q D_p \frac{dp}{dx} \quad (2.30)$$

When in addition an electric field is applied then the total current constitutes of the drift current and the diffusion current:

$$n - type: \quad J_n = q \mu_n n E + q D_n \frac{dn}{dx} \quad (2.31)$$

$$p - type: \quad J_p = q \mu_p p E - q D_p \frac{dp}{dx} \quad (2.32)$$

It's obvious that the total current is then

$$J = J_n + J_p \quad (2.33)$$

2.1.5 Recombination-Generation of carriers

In the equations established so far, we have seen that one major contributor to current flow is temperature; electrons gain momentum by absorbing energy created by heat. In appendix A we described the mechanism of light absorption and how electrons, or any particle, get advantage of it. Here we will analyze how we can get advantage of this phenomenon, namely producing excess current thanks to light energy. The first step is to know the available range of energy for our material. We also remember from appendix A that electromagnetic wave (here visible light) travels at a frequency f_v and that the energy associated with it is $E = h \cdot f_v$. Let's say we choose silicon as our material then again from appendix A we know it has band gap E_g . Let's put these together:

$$h f_v = \frac{hc}{\lambda} \geq E_g \quad (2.34)$$

This equation tells us that energy bigger than the band gap of the material will not be absorbed. Therefore photons will see the material as transparent and simply pass through it. An equivalent expression of energy is that which is containing the wavelength λ . The maximum frequency or wavelength, that can be detected (absorbed) is known as the cut off frequency or cut off wavelength respectively. Using (A.17), (A.38) and (2.32) we can set up an absorption-cutoff wavelength graph (figure 2.4) for various semiconductors.

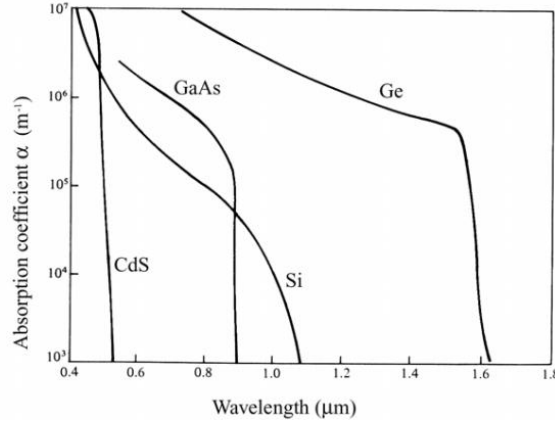


Figure 2.4 Absorption and wavelength graph. One can easily see the cutoff wavelength for various semiconductors. Si has a cut- off wavelength of 1106 μm, hence $E_{Si}=1,124$ eV @ 300K. Figure from [3].

The next step is to define the excess current produced, which brings us to the concept of generation and recombination. When an electron absorbs a photon (energy) it promotes itself to the conduction band. This is called generation of carrier, and the rate at which this phenomenon happens is expressed as g_i . After τ time (carrier lifetime) the electron recombines with a hole and this rate is denoted as r_i . In equilibrium an equal numbers of free carriers exist and hence the concentrations of electrons n_0 and holes p_0 are equal to the rates:

$$g_i = r_i = Bn_0p_0 \quad (2.35)$$

where B is a constant. When light shines on the semiconductor extra carriers are generated and therefore more carriers recombine. This increases the equilibrium rate g_0 by g_i . The carrier concentration then are influenced by increase carriers Δ_n and Δ_p , where of course $\Delta_n = \Delta_p$. Lastly r_0 increases by r_i , trying to balance the change. Exactly when r_0 reaches r_i , we have from (2.33) that

$$r_0 + r_i = B(n_0 + \Delta_n)(p_0 + \Delta_p) \quad (2.36)$$

The total change in hole concentration is then

$$\frac{dp}{dt} = g_0 + g_i - (r_0 + r_i) \quad (2.37)$$

When the change in concentration is zero we have maximum generation and from (2.36) and (2.37) we get

$$g_0 = B(n_0 + \Delta_n)(p_0 + \Delta_p) - Bn_0p_0 \quad (2.38)$$

To solve this we simply note that when we have for instance n-type semiconductor $n_0 \gg p_0$ then $\Delta_n \Delta_p$ is zero.

$$g_0 \cong Bn_0\Delta_p = \frac{p-p_0}{\frac{1}{Bn_0}} \quad (2.39)$$

where p is the optically generated hole concentration. The denominator is the minority carrier lifetime and is symbolized τ_p for holes or τ_n for electrons. Lifetime is measured in time units and expresses the time an excess carrier remains free before recombining. After r increased to balance the change, we have reached steady state and then

$$g_0 = r_0 = \frac{p-p_0}{\tau_p} \quad (2.40)$$

Rearranging (2.38) gives us p

$$p = p_0 + \tau_p g_0 \quad (2.41)$$

The light source is suddenly switched off, say at $t=0$, where (2.41) holds. Then after a long time $p(t=\infty)=p_0$. So we have two known conditions for the optically generated hole concentration:

$$p(t) = \begin{cases} p_0 + \tau_p g_0, & t = 0 \\ p_0, & t = \infty \end{cases} \quad (2.42)$$

Using the total change equation (2.36) we can measure the amount of optically generated hole and consequently the lifetime:

$$\frac{dp}{dt} = g_i - (r_0 + r_i) \quad (2.43)$$

Using (2.33) and (2.38) in (2.40)

$$\frac{dp}{dt} = -\frac{p-p_0}{\tau_p} \quad (2.44)$$

The solution with (2.42) as the boundary conditions is

$$p(t) = p_0 + \tau_p g_0 e^{\frac{-t}{\tau_p}} \quad (2.45)$$

The same result holds for minority concentration consisting of electrons in a p-type semiconductor. (2.45) holds only for direct band gap semiconductors, where electrons recombine directly with holes. In the recombination process the electron recombines with the hole in the valence band (interband). By the law of conservation, the electron must set free the energy gained from the photon absorption. This is done in two different ways and therefore we have two different recombination processes:

- Direct transition. Photon wave vector coincide with K-vector at Γ . Energy set free through photon emission.
- Indirect transition. Photon wave does not agree with K-vector at Γ . Energy is set free by creation and destruction of phonons.

Consequently with direct transitions one could implement laser diodes, since we got radiative emissions. The material chosen for this are obviously direct band gap materials. As for materials with indirect transition they would make poor lasers because of their low emission probability. In Non radiative materials, like silicon, electrons recombine through impurity centers. These centers are either “little” energy states or lay in the forbidden energy zones. The mechanism is that the electron falls into one centre and is kept there till a hole passes by or is created, with which it can recombine. This process makes calculating the constant B in the generation, recombination equation (2.35) very difficult. It has been done and the result for some semiconductors is presented in the figure 2.5:

Group(s)	Element/ compound	Direct/ indirect	E_g (eV)	Readily doped n- or p-type	B ($m^3 s^{-1}$)	λ_g (nm)
IV	C	i	5.47			227
	Si	i	1.12	Yes	1.79×10^{-21}	1106
	Ge	i	0.67	Yes	5.25×10^{-20}	1880
IV-VI	SiC	i	3.00	Yes		413
III-V	(hex. α)					
	AlP	i	2.45			506
	AlN	i	5.90	No		210
	AlSb	i	1.50			826
	AlAs	i	2.16			574
	GaN	d	3.40	No		365
	GaP	i	2.26	Yes	5.37×10^{-20}	549
	GaAs	d	1.43	Yes	7.21×10^{-16}	861
	InN	d	2.40			516
	InP	d	1.35	Yes	1.26×10^{-15}	918
	InAs	d	0.35		8.50×10^{-17}	3540
	InSb	d	0.18		4.58×10^{-17}	6870
	ZnO	d	3.20	No		387
	ZnS(α)	d	3.80	No		326
	ZnS(β)	d	3.60	No		344
II-VI	ZnSe	d	2.28	No		480
	ZnTe	d	2.58	No		544
	CdS	d	2.53	No		490
	CdSe	d	1.74	No		712
	CdTe	d	1.50	Yes		826

Figure 2.5 Properties of various semiconductors including B. Figure from [1]

2.1.6 The continuity equation

The continuity equation incorporates all three basic phenomena's in the semiconductor, namely drift, diffusion and generation-recombination. Once again we take an arbitrary part or slice of the semiconductor and study the occurring phenomena's.

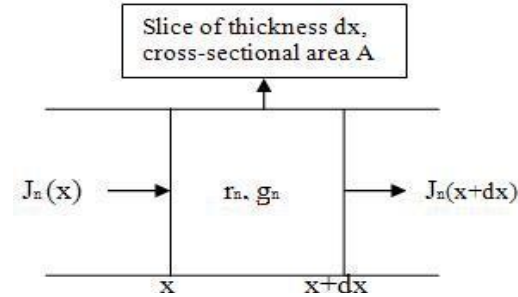


Figure 2.5 Studying drift, diffusion and recombination processes in an infinitesimal area of a semiconductor.

There are four aspects that one should include in the calculations:

1. The number of electrons flowing into the slice at x .
2. The generation rate in the slice
3. The number of electrons flowing out of the slice at $x+dx$
4. The recombination rate within the slice

As we saw in the previous section we calculated the change in hole concentration,

let's calculate the change in electron concentration in the area A , then we have

$$\frac{\partial n}{\partial t} A dx = \left[\frac{1}{-q} \frac{J_n(x)A}{dx} - \frac{2}{-q} \frac{J_n(x+dx)A}{dx} \right] + (\overset{3}{g_n} - \overset{4}{r_n}) A dx \quad (2.46)$$

$J_n(x + dx)$ can be expressed with a Taylor series and (2.46) can be written as

$$\frac{\partial n}{\partial t} = \frac{\partial J_n(x)}{-q \partial x} + (g_n - r_n) \quad (2.47)$$

This is called the electron continuity equation. To express both electron and hole concentration in one equation one has to observe that due to doping a p-type semiconductor has as majority carriers holes and as minority carriers electrons. Consequently expressing the minority concentration n in a p-type using (2.31) gives us

$$\frac{\partial n_p}{\partial t} = n_p \mu_e \frac{\partial E}{\partial x} + \mu_e E \frac{\partial n_p(x)}{\partial x} + D_n \frac{\partial^2 n_p(x)}{\partial x^2} + (g_n - r_n) \quad (2.48)$$

The same procedure can be done for holes. Then (2.45) and (2.45) for holes are

$$\frac{\partial p}{\partial t} = \frac{\partial J_p(x)}{-q \partial x} + (g_n - r_p) \quad (2.49)$$

$$\frac{\partial p_n}{\partial t} = -p_n \mu_h \frac{\partial E}{\partial x} - \mu_h E \frac{\partial p_n(x)}{\partial x} + D_p \frac{\partial^2 p_n(x)}{\partial x^2} + (g_p - r_p) \quad (2.50)$$

The solutions to (2.48) and (2.50) are difficult to find. Using Poisson equation and the boundary conditions one can make approximations based on assumptions. To understand this let's take the case where minority carriers are injected at one end in a p-type semiconductor as shown in Figure 2.6 a). Let's assume no field exists then g_n and E will be zero. This situation is equivalent of having a slice far away from the injected edge. Hence changes in concentration are not time depended anymore and the partial derivatives have no meaning, in other words (2.48) becomes a steady state equation:

$$\frac{dn_p}{dt} = 0 = D_n \frac{d^2 n_p(x)}{dx^2} - r_n \quad (2.51)$$

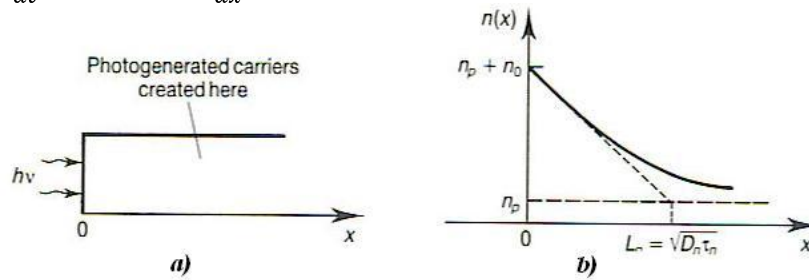


Figure 2.6 Approximation of carrier diffusion length with continuity equation. Figure from [2].

From (2.52) we have that (2.51) is

$$D_n \frac{d^2 n_p(x)}{dx^2} = \frac{n - n_0}{\tau_n} = \frac{\Delta n_p(x)}{\tau_n} \quad (2.52)$$

A solution for continuity equation is now feasible by choosing for the injected edge $x=0$, where the minority concentration is $n_p = n_p(0)$, and at far away from the edge $x=\infty$, where the minority concentration is steady (zero increase) $n_p = n_p$. With these boundary conditions we find the final version of the continuity equation:

$$n_p(x) = n_p + (n_p(0) - n_p) e^{-\frac{x}{L_n}} \quad (2.53)$$

where $L_n = \sqrt{D_n \tau_n}$ is the diffusion length, representing the length an electron can travel before it recombines. Figure 2.6 b) shows an example of light injection which causes a rapid increase in minority carrier concentration (electrons) and the decay of it if the light source is switched off. With no light and no electric field the generated carriers travel solely due to diffusion. Light injection and diffusion was also exploited in the Haynes-Shockley experiment, where the minority lifetime was extracted by measuring the carrier concentration decay.

Note: n_p or p_n can be read as carrier concentration of electrons (n/p) in a p/n-type semiconductor

2.1.7 Photoconductivity

If a uniformly illumination falls upon the semiconductor, meaning the increase in generation rate is position independent, additional electron-hole pair will be created. As a result we have new levels of carrier concentrations, expressed like

$$n = n_0 + \Delta n \text{ and } p = p_0 + \Delta p \quad (2.54)$$

Assumed we have p-type semiconductor where $n_0 \ll p_0$. Also we assume low light injection, and then $\Delta n, \Delta p \ll p_0$ holds. From (2.39) we get

$$r_n \cong g_n \cong B p_0 \Delta n = \frac{n-n_0}{\frac{1}{B n_0}} = \frac{n-n_0}{\tau_n} = \frac{\Delta n}{\tau_n} \quad (2.55)$$

The total change in electron concentration in time can now be viewed as the optical generation rate of electrons minus the rate of recombined electrons. Mathematically this can result also from the continuity equation (2.48) setting terms with electron concentrations to zero, and replace the generation rate with the optical one G_0 . This substitution is done because the generation is assumed only to be due to optical means, namely ideally we suppose no thermal generation is taking place. Hence

$$\frac{dn}{dt} = G_0 - \frac{\Delta n}{\tau_n} \quad (2.56)$$

The p-type semiconductor is connected by a external source and the electron and hole densities are driven by the electric field of the source. Hence we got drift current and the following holds:

$$J_h = q \mu_h p E \text{ and } J_n = q \mu_e n E \quad (2.57)$$

The total current density is

$$J = J_n + J_p = q(\mu_h n + \mu_e p) E = \sigma E \quad (2.58)$$

where σ the conductivity is

$$\sigma = q(\mu_h n + \mu_e p) \quad (2.59)$$

The photoconductivity is then defined as the difference between the conductivity with optical injection and the conductivity without optical injection also called the conductivity σ_0 :

$$\Delta \sigma = \sigma - \sigma_0 = q(\mu_h \Delta n + \mu_e \Delta p) \Rightarrow \sigma_0 = q(\mu_h n_0 + \mu_e p_0)$$

The total current is found by multiplying the density with the area A as shown in figure 2.7

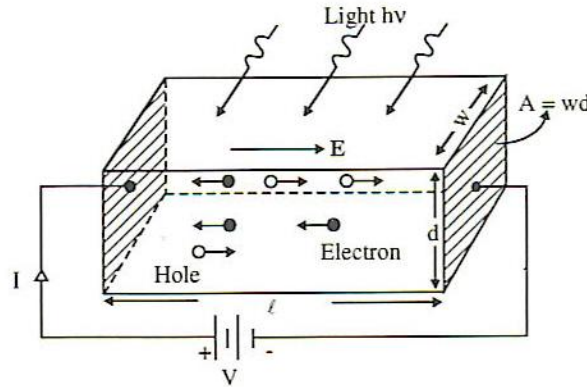


Figure 2.7 Uniform optical generation in a p-type semiconductor. Figure from [4]

where can be seen that $A=wd$ and then we have

$$I = JA = \sigma EA = \frac{\sigma AV}{l} \quad (2.60)$$

The photocurrent ΔI is defined as the difference between the total current in the presence of optical creation of electron-holes (excitation) and the dark current $I_0 = \frac{\sigma_0 AV}{l}$.

$$\Delta I = I - I_0 = \Delta \sigma \frac{A}{l} V = q(\mu_h \Delta n + \mu_e \Delta p) \frac{A}{l} V \quad (2.61)$$

2.1.7.1 Constant light intensity

If G_0 is independent of time then at steady state, we obtain

$$\Delta n = G_0 \tau_n \quad (2.62)$$

Since each broken bond produces an electron-hole pair we have $\Delta n = \Delta p$. Therefore the photocurrent is substituting (2.62) into (2.61)

$$\Delta I = q(\mu_h + \mu_e) G_0 \tau_n \frac{A}{l} V \quad (2.63)$$

Holes have usual a much slower mobility then electrons and therefore (2.63) can be expressed with the transit time of electrons. Analyzing the transit time gives

$$\tau_t = \frac{l}{u_n} = \frac{l}{\mu_n E} = \frac{l^2}{\mu_n V} \quad (2.64)$$

Applying (2.64) to (2.63) we have

$$\Delta I = q(G_0 l A) \frac{\tau_n}{\tau_t} \quad (2.65)$$

The term $G_0 l A$ is the total number of electron-hole pairs created per second in a volume lA . The ratio recombination time (carrier lifetime) to transit time $\frac{\tau_n}{\tau_t}$ is the photoconductive gain, which as we will see in chapter three is an important criteria for collecting photocurrent. We turn now to the optical generation rate: This is equal to the number of injected photons per second. The latter is equal to radiant flux, which is defined as photon flux $\frac{P_{opt}}{hf_v}$ per unit volume lwd multiplied by the quantum efficiency η .

$$G_0 = \eta \frac{\frac{P_{opt}}{hf_v}}{lwd} \quad (2.66)$$

The quantum efficiency η is the fraction of photons that results into creation of electron-hole pairs. From appendix A, we know that the light is “split” into a refracted and into a reflected part. As a result the passing light (refracted) is the part that actually generates the electron-pair. For this reason we define two quantum efficiencies: The internal and the external quantum efficiency. The internal quantum efficiency η_{int} is the proportion of electron-hole recombination's per second that result in photon emission. The number of radiative recombination's per second will be $\frac{1}{\tau_{rr}}$, while the total number of recombinations per second must be $\frac{1}{\tau_{rr}} + \frac{1}{\tau_{nr}}$, so

$$\eta_{int} = \frac{\frac{1}{\tau_{rr}}}{\frac{1}{\tau_{rr}} + \frac{1}{\tau_{nr}}} = \frac{\frac{1}{\tau_{rr}}}{1 + \frac{\tau_{rr}}{\tau_{nr}}} \quad (2.67)$$

Hence for non-radiative recombinations η_{int} is quite low, which is true for silicon based materials. This also means that the non-radiative lifetime is much bigger than the radiative lifetime, a property that should be noticed. Having defined η_{int} we are now able to derive the external quantum efficiency η . It is multiplied by the absorbance which includes the reflectance mentioned above; see equation (A.45). Putting it all together:

$$\eta = \eta_{int}(1 - R)(1 - e^{-ad}) \quad (2.68)$$

The primary injected photocurrent I_{ph} is defined as

$$I_{ph} = q\eta \frac{P_{opt}}{hf_v} \quad (2.69)$$

Finally the photocurrent is then

$$\Delta I = q\eta \frac{P_{opt} \tau_n}{h f_v \tau_t} \quad (2.70)$$

The sensitivity, meaning the amount of change of photocurrent due to change in optical power (radiant power) is expressed with the current responsivity R (A/W).

$$R = \frac{\Delta I}{P_{opt}} = q \frac{\eta \tau_n}{h f_v \tau_t} \quad (2.71)$$

2.1.7.2 Impulse response

It is the same case as in (2.45) for a n-type semiconductor. Substituting the generation rate with the optical generation rate and using (2.61) we get

$$\Delta n = n(t) - n_0 = \tau_n g G_0 e^{\frac{-t}{\tau_n}} \quad (2.72)$$

$$\Delta I(t) = q \mu_n G_0 \tau_n \left(\frac{AV}{l} \right) e^{\frac{-t}{\tau_n}} \quad (2.73)$$

Modeling the optical light as an impulse function and using the continuity equation, like we did to derive (2.56) we get

$$G_0(t) = g_0 \delta(t) \quad (2.74)$$

$$\frac{\partial}{\partial t} \Delta n(t) = g_0 \delta(t) - \frac{\Delta n(t)}{\tau_n} \quad (2.75)$$

Supposing that the semiconductor is at equilibrium at the beginning and the concentrations satisfy at time t $\delta n(0_+) - \delta n(0_-) = g_0$. Hence no excess carrier concentrations are at start. Then

$$\delta n(t) = g_0 e^{\frac{-t}{\tau_n}} \quad (2.76)$$

2.1.7.3 Sinusoidal steady-state response

We can model the optical intensity in two ways: One being with sinusoidal signal such that

$$G_0(t) = G_0 \cos \omega t \quad (2.77)$$

Choosing once again the continuity equation and substituting the generation rate we get

$$\frac{\partial}{\partial t} \Delta n(t) = G_0 \cos \omega t - \frac{\Delta n(t)}{\tau_n} \quad (2.78)$$

The response to (2.77) can be found by defining

$$\delta n(t) = \text{Re}(\delta n e^{-i\omega t}) \text{ and } G_0 \cos \omega t = \text{Re}(G_0 e^{-i\omega t}) \quad (2.79)$$

Then

$$\delta n = \frac{G_0 \tau_n}{1 - i\omega \tau_n} \quad (2.80)$$

Final act to find the solution is to take the real part of (2.80)

$$\delta n(t) = \text{Re} \left(\frac{G_0 \tau_n}{1 - i\omega \tau_n} \right) e^{-i\omega t} \Rightarrow \quad (2.81)$$

$$\delta n(t) = \frac{G_0 \tau_n}{\sqrt{1 - \omega^2 \tau_n^2}} \cos(\omega t - \varphi) \quad (2.82)$$

Where φ is the delay in phase of the the ac response and is equal to $\tan^{-1}(\omega \tau_n)$.

The other way of modeling is include ac and dc components that the optical generation rate might have

$$G_0(t) = G_0(1 + m \cos \omega t) \quad (2.83)$$

where m is the modulation index. The response of the excess carrier concentration is then

$$\delta n(t) = G_0 \tau_n \left[1 + \frac{m}{\sqrt{1 + \omega^2 \tau_n^2}} \cos(\omega t - \varphi) \right] \quad (2.84)$$

Inserting the modulated optical signal into its power expression yields

$$P(t) = P_{opt}(1 + m \cos(\omega t)) \quad (2.85)$$

The photocurrent response is then

$$I(t) = I_p \left[1 + \frac{m}{\sqrt{1 + \omega^2 \tau_n^2}} \cos(\omega t - \varphi) \right] \quad (2.86)$$

where

$$I_p = q\eta \frac{P_{opt} \tau_n}{h f_v \tau_t} \quad (2.87)$$

If we take the mean root square of the optical power and replacing the cosine term with $1/\sqrt{2}$,

$$p_{rms} = \frac{m P_{opt}}{\sqrt{2}} \quad (2.88)$$

The rms photocurrent is then

$$i_p = q\eta \frac{P_{opt} \tau_n}{h f_v \tau_t} \frac{m}{\sqrt{1 + \omega^2 \tau_n^2}} \quad (2.89)$$

This concludes our investigation of optical generation rate modeling. Figure 2.8 shows all cases

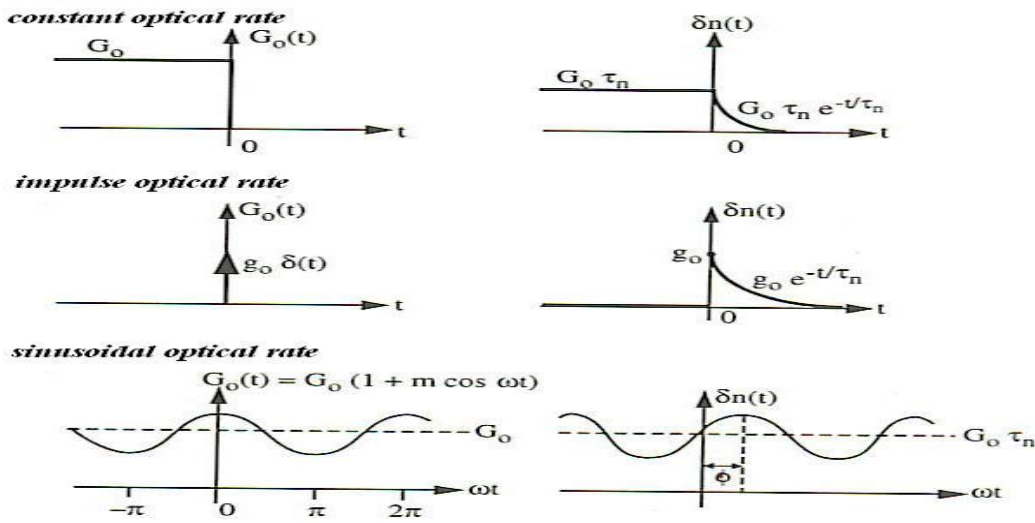


Figure 2.8 Uniform optical generation in a p-type semiconductor. Figure from [4]

2.2 Forming Junctions

Bringing together dissimilar materials opens up new generation of devices. The simplest one consist of two dissimilar materials (n-type and p-type) that are forming a junction. This junction is the well known diode or pn junction. The diode is a very important device and is used in a large variety of applications, be it in optoelectronics as laser, in

electronics as diode or, and this interest us the most, as photodetection device. Also formed from more than two junctions are the transistors, like the bipolar, JFET and MOS transistor. There are many devices which are capable of detecting light. The most common detecting devices are analytical described in the third chapter. For this reason the understanding of junctions is very important.

2.2.1 Pn Junction

2.2.1.1 Zero Bias (Equilibrium)

In forming the junction pn, like the name implies, n-type and p-type material are brought together. At the junction the transition from one type of concentration to the other plays an important role, there tow major transitions

1. Abrupt
2. Linearly graded

The first is an assumption that the concentration changes instantly and the second is that the concentration change linearly. To simplify the analysis abrupt junctions are mostly used to generate a mathematical model of their behavior. When the junction is formed the holes in the p-type diffuse to the n-type and simultaneously the electrons from the n-type will diffuse into p-type. Note that the diffused carriers were majority carriers in their type of semiconductor and are now minority carrier. For instance holes in p-type diffuse to n-type. The rearranged carriers cause an electric field to form, since we got suddenly two opposite charges. This electric field is the same electric field that causes drift current and hence minority carriers are transferred from one to the other side. Like we saw the direction of drift current (electric field) is in the opposite way of diffusion and thus competes with the diffusion current. In one point in time their fields are equal and no more diffusion or drift is observed. At this state the junction is said to be in equilibrium and the formed potential across the junction is called the built in potential. Because this process is a little bit confusing the figure 2.9 explains in a more optical way.

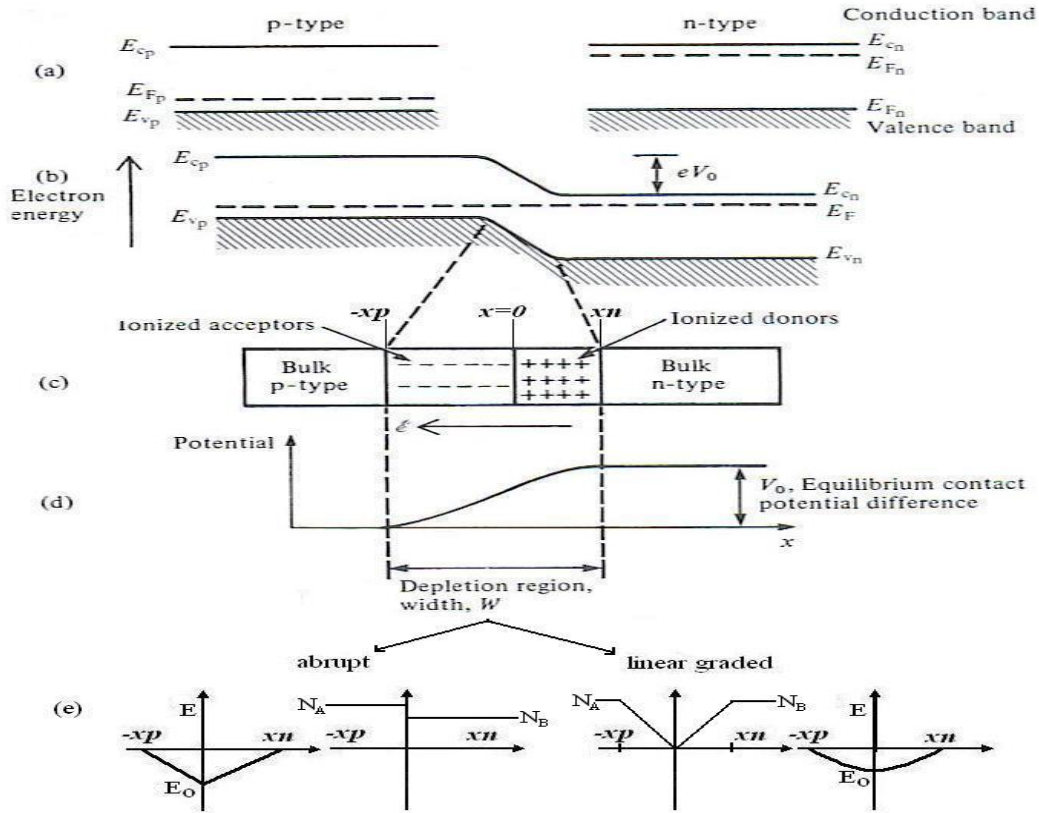


Figure 2.9 Formation of pn junction a) Separate pieces and their energy levels b) Brought together and the new energy levels, abrupt changes c) diffusion and drift carrier "battle" d) equilibrium after the "battle". (e) Electric field for abrupt and linear junction. Figure from [1].

We can calculate the built-in potential V_{bi} . This can be done by observing that the energy levels for the conduction and the valence band are different but the Fermi energy stays the same. Since the concentrations depend on the difference between the energy conduction valence and the Fermi level we can rewrite (2.9) for electrons in both types.

$$n_{n0} = N_c e^{\frac{(E_f - E_{cn})}{kT}} \quad \text{"in n-type"} \quad (2.90)$$

$$n_{p0} = N_c e^{\frac{(E_f - E_{cp})}{kT}} \quad \text{"in p-type"} \quad (2.91)$$

where 0 denotes the concentrations in equilibrium. From figure 2.9 we see that the energy difference $E_{cp} - E_{cn} = qV_{bi}$, therefore taking the logarithm on both sides of (2.90) and (2.91) yields

$$V_{bi} = \frac{E_{cp} - E_{cn}}{q} = \frac{1}{q} kT \log_e \frac{n_{n0}}{n_{p0}} \quad (2.92)$$

$$\Rightarrow V_{bi} = E_{cp} - E_{cn} = \frac{kT}{q} \log_e \frac{n_{n0}}{n_{p0}} \quad (2.93)$$

Using the continuity equation and Poisson's equation one gets in one dimension that

$$\frac{dE}{dx} = -\frac{d^2V}{dx^2} = \frac{\rho}{\epsilon\epsilon_0} = \frac{q(N_D - N_A)}{\epsilon\epsilon_0} \quad (2.94)$$

where N_D and N_A are the majority carrier concentration and ϵ and ϵ_0 are the dielectric constants of permittivity in silicon and air respectively. Using this result and (2.8) in the junction region one gets

$$V_{bi} = \frac{D_p}{\mu_h} \log_e \frac{N_A N_D}{n_i^2} = \frac{D_n}{\mu_e} \log_e \frac{N_A N_D}{n_i^2} = \frac{kT}{q} \log_e \frac{N_A N_D}{n_i^2} \quad (2.95)$$

where $\frac{kT}{q}$ is often called the thermal voltage Φ_T .

2.2.1.2 Depletion region

The depletion region is defined as the region where there are only fixed ions after the drift and diffusion flows “stripped” away the minority carriers as one can see from figure 2.9. The remaining ions are fixed since they cannot move and are called ionized impurities. Also from figure 2.9 the region can be defined from $-x_p$ to x_n . Using (2.94) we find that

$$\frac{dE}{dx} = \begin{cases} -\frac{q}{\epsilon\epsilon_0} N_A, & -x_p \leq x < 0 \\ \frac{q}{\epsilon\epsilon_0} N_D, & 0 < x \leq x_n \end{cases} \quad (2.96)$$

For $E=0$ $-x_p=x_n=x$ and $E=E_0$ like one can see from figure 2.7 (e) then at $x=0$

$$\int_0^{E_0} dE = -\frac{q}{\epsilon\epsilon_0} N_A \int_{-x_p}^0 dx \quad (2.97)$$

$$E_0 = -\frac{q}{\epsilon\epsilon_0} N_A x_p = \frac{q}{\epsilon\epsilon_0} N_D x_n \quad (2.98)$$

$$E_0 = \frac{q}{\epsilon\epsilon_0} N_A x_p + \frac{q}{\epsilon\epsilon_0} N_D x_n \quad (2.99)$$

Integrating (2.98) gives us the associated potentials

$$V_{bi} = \frac{q}{\epsilon\epsilon_0} N_A x_p^2 + \frac{q}{\epsilon\epsilon_0} N_D x_n^2 = V_p + V_n \quad (2.100)$$

Analyzing the term $x_p^2 + x_n^2$ and noting that $x_p + x_n = W$ and $N_D x_n = N_A x_p$ we get

$$V_{bi} = \frac{q}{2\epsilon\epsilon_0} N_A x_p (x_p + x_n) \quad (2.101)$$

(2.101) can be rearranged to give

$$x_p = \frac{W N_D}{N_A + N_D} \text{ and } x_n = \frac{W N_A}{N_A + N_D} \quad (2.102)$$

Substituting (2.102) in (2.101) we have

$$V_{bi} = \frac{q}{2\epsilon\epsilon_0} \frac{N_A N_D}{N_A + N_D} W^2 \quad (2.103)$$

From this one can easily calculate the depletion width W

$$W = \left(\frac{2\epsilon\epsilon_0 V_{bi} (N_A + N_D)}{q N_A N_D} \right)^{\frac{1}{2}} \quad (2.104)$$

From another point of view the depletion region can be viewed as being a capacitor with a dielectric constant. The stored charge Q_j , j stands for junction, on either side of the junction is

$$|Q_j| = A q N_d x_n = A q N_a x_p \quad (2.105)$$

Using (2.102) and (2.104) we can write

$$|Q_j| = \frac{A q N_d N_a x_n W}{N_A + N_D} = A \left(\frac{2\epsilon\epsilon_0 (V_{bi} - V) (N_A + N_D)}{q N_A N_D} \right)^{\frac{1}{2}} \quad (2.106)$$

Differentiating (2.106) with respect to V gives us

$$C_j = \frac{dQ_j}{dV} = \frac{A}{2} \left[\left(\frac{2q\epsilon\epsilon_0}{(V_{bi} + V)} \right) \left(\frac{(N_A N_D)}{(N_A + N_D)} \right) \right]^{\frac{1}{2}} \quad (2.107)$$

Where A is the junction capacitance and V is the voltage of a source connected to the junction, also called biasing the junction. This leads us to the concept of biasing a pn junction. As one can see from (2.107) biasing influences the capacitance of the junction and it has been found that for abrupt $C \sim V^{-\frac{1}{2}}$ and for linear graded junctions $C \sim V^{-\frac{1}{3}}$. The biasing of a junction and the associated behavior will be explained next.

2.2.1.3 Bias Cases

In this section we will examine the pn junction's behavior under bias conditions. The first will be the forward bias and the second will be the reversed bias condition. These conditions will give us the basic understanding of what currents dominates and how we extract these currents for our advantage.

2.2.1.3.1 Forward bias

As the figure shows the source (battery) is connected to the anode and cathode of the junction. Therefore electrons are injected into the n-type region and holes into the p-type region. The conduction and valence bands react to this and bend with the result of lowering the barrier height from qV_{bi} to $q(V_{bi} - V)$. The latter has the effect that electrons and holes can now easier diffuse to the opposite regions p and n respectively, where they turn into minority carriers. Because the quantities of carrier crossing the junction are large we have a rather large diffusion current. On the other hand the drift current remains the same. In other words in applying a source, to make a junction be forward biased, we managed to manipulate the balance of diffusion and drift current. The outcome is the increase in diffusion current. Let's calculate the new minority carrier concentrations. For this we use (2.92) with the voltage V of the source:

$$n_p = n_{n0} e^{-\left(\frac{q(V_{bi} - V)}{kT}\right)} = n_{p0} e^{\left(\frac{qV}{kT}\right)} \quad (2.108)$$

The difference in concentration $\Delta n(x)$ as a function of distance from the junction is

$$\Delta n(x) = n_p(x) - n_{p0} \quad (2.109)$$

At zero distance the difference is

$$\Delta n(0) = n_p - n_{p0} \quad (2.110)$$

From (2.108) and (2.110) we have

$$\Delta n(0) = n_{p0} \left[e^{\left(\frac{qV}{kT}\right)} - 1 \right] \quad (2.111)$$

With the definition of diffusion length (2.111) gets

$$\Delta n(x) = \Delta n(0) \left[e^{-\left(\frac{x}{L_n}\right)} - 1 \right] \quad (2.112)$$

Differentiating (2.112)

$$\frac{d\Delta n(x)}{dx} = -\frac{\Delta n(x)}{L_n} e^{-\left(\frac{x}{L_n}\right)} \quad (2.113)$$

If $\frac{d\Delta n(x)}{dx} = \frac{dn(x)}{dx}$ holds we can use the electron diffusion density, which results in

$$J_n = \frac{qD_n \Delta n(0)}{L_n} e^{-\left(\frac{x}{L_n}\right)} \quad (2.114)$$

Combining (2.111) and (2.114)

$$J_n = \frac{qD_n}{L_n} n_{p0} \left[e^{\left(\frac{qV}{kT}\right)} - 1 \right] \quad (2.115)$$

With a similar procedure one can show that for hole diffusion

$$J_p = \frac{qD_p}{L_p} p_{n0} \left[e^{\left(\frac{qV}{kT}\right)} - 1 \right] \quad (2.116)$$

The total diffusion observed across the junction is then

$$J = J_n + J_p = J_s \left[e^{\left(\frac{qV}{kT}\right)} - 1 \right] \quad (2.117)$$

$$\text{where } J_s = \frac{qD_p}{L_p} p_{n0} + \frac{qD_n}{L_n} n_{p0} \quad (2.118)$$

For large values of V the junction equation (2.117) becomes

$$J = J_s \left[e^{\left(\frac{qV}{kT}\right)} \right] \quad (2.119)$$

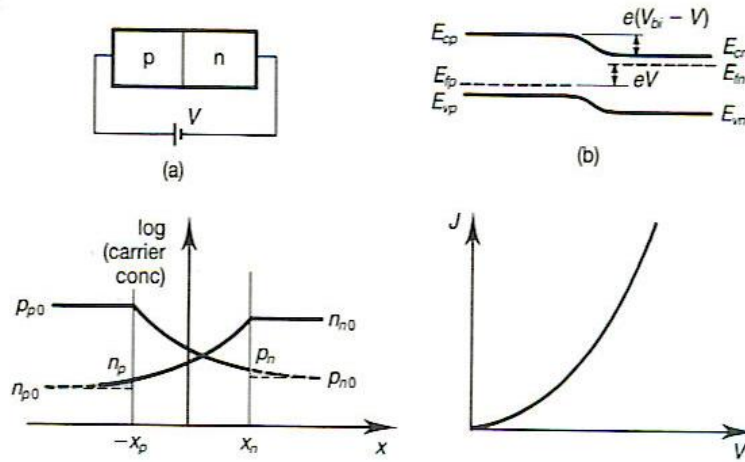


Figure 2.10 pn junction a) Connecting anode(p) and cathode(n) to the source's positive and negative pole b) Barrier lowering allows carrier concentrations to diffuse. The diode equation (2.119) has an exponential nature. Figure from [1]

2.2.1.3.2 Reverse bias

It is similar to the forward bias case with almost everything reversed. Electrons are now being injected into the p region and holes into the n-region. This raises the barrier from qV_{bi} to $q(V_{bi} + V)$. For this reason there is no diffusion current observed across the junction since the carriers do not have the energy to “jump” over the new barrier. The injected carrier concentrations are minority concentrations which driven solely due to drift. This needs further explanation: The external source injects minority carriers which will “jump” over the junction, due to the electric field across the depletion region. For instance looking at figure 2.9 (c) a hole is injected from the right and is accelerated through the electric field. In addition, the external source also lowers the diffusion component, and therefore strengthens the drift component. At the junction there are gathered minority carriers, attracted by the ionized atoms. With the applied source these are then swept to the poles and new minority carriers are then extracted from the vicinity

of the depletion region. Exactly this “stripes” away additional minority carriers and hence the depletion region grows accordingly to the force of the source. We mentioned that there is no diffusion observed, this is true for the junction. Truly no carriers jump over the junction with diffusion, but the diffusion component is responsible that the injected minority carriers made it to the depletion region. For this reason we acknowledge that two mechanisms define the total current in the reverse bias case:

$$J = J_n + J_p = J_s \left[e^{\left(\frac{qV}{kT}\right)} - 1 \right] \quad (2.120)$$

For large values of the voltage in reverse bias (2.120) becomes

$$J = -J_s \quad (2.121)$$

Thus $-J_s$ is called the saturation voltage and because it has a small value the diode doesn't conduct much.

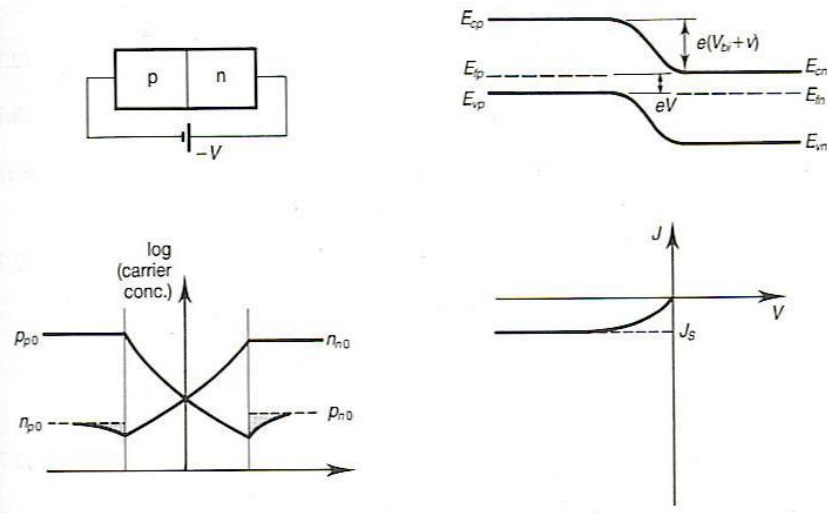


Figure 2.11 pn junction a) Connecting anode(p) and cathode(n) to the source's negative and positive pole b) Barrier heightening produces more minority carriers that drift across the junction. The saturation current (2.121) for large voltages remains constant.

Figure from [2].

2.2.1.4 Generation-recombination in pn junction

2.2.1.4.1 Forward bias

To understand the procedure of light interaction we look at the generation and recombination of currents in the pn junction. Especially important is the depletion region, where the generated carriers are swept across it and therefore easily detected. As before we begin analyzing the generation recombination processes:

1. Electron capture at an energy centre
2. Electron emission from occupied centre
3. Hole capture at an occupied centre
4. Hole release from an empty centre

The band gap concentration N_t has an energy centers E_t and for which the rate R_a of capture of electron by process is defined as

$$R_a = c_n n N_t (1 - F(E_t)) \quad (2.122)$$

where $1 - F(E_t)$ is the probability of a trap being unoccupied and c_n is the capture coefficient of electrons with typical values of $10^{18} \text{ cm}^3 \text{ s}^{-1}$. One centre can be occupied by one electron, hence R_a is proportional to the number of unoccupied centers, which is the product $N_t(1 - F(E_t))$. For the second process, R_b is the rate of electron emissions and is equal to

$$R_b = e_n N_t F(E_t) \quad (2.123)$$

Similarly for the third and forth process the capture and emission rates are given by

$$R_c = c_p p N_t F(E_t) \quad (2.124)$$

$$R_d = e_p p N_t (1 - F(E_t)) \quad (2.125)$$

where c_p and e_p are the capture and emission coefficients respectively.

In thermal equilibrium the number of emitted electrons is equal to that of the captured holes. Therefore $R_b = R_a$. Thus

$$c_n n (1 - F(E_t)) = e_n F(E_t) \quad (2.126)$$

Using (2.11) and the Fermi equation in (2.125) give

$$e_n = c_n n_i e^{\frac{(E_t - E_i)}{kT}} \quad (2.127)$$

Similar for holes

$$e_p = c_p n_i e^{\frac{(E_i - E_t)}{kT}} \quad (2.128)$$

For non-equilibrium (2.126) and (2.127) do not hold. We denote G as the generation rate for which in steady state the number of electrons in the conduction band or holes in the valence band are equal. Then

$$G = R_a - R_b \quad (2.129)$$

$$G = R_c - R_d \quad (2.130)$$

Combining the last two equations with the four rates R_a, R_b, R_c and R_d yields:

$$c_n n (1 - F(E_t)) - e_n F(E_t) = c_p p F(E_t) - e_p (1 - F(E_t)) \quad (2.131)$$

Setting $c_n = c_p = c_0$ and remembering the relationship $\cosh x = \frac{(e^x + e^{-x})}{2}$ we can solve for $F(E_t)$:

$$F(E_t) = \frac{n + n_i e^{\frac{(E_i - E_t)}{kT}}}{n + p + 2n_i \cosh\left(\frac{(E_t - E_i)}{kT}\right)} \quad (2.132)$$

Like mentioned above the generation and the recombination rate are equal, in example $G=U$. Using (2.126) – (2.129) we get

$$U = c_0 N_t \left[n (1 - F(E_t)) - n_i e^{\frac{(E_t - E_i)}{kT}} F(E_t) \right] \quad (2.133)$$

After some minor calculations we get

$$U = \frac{c_0 N_t (pn - n_i^2)}{n + p + 2n_i \cosh\left(\frac{(E_t - E_i)}{kT}\right)} \quad (2.134)$$

From (2.97) important conclusions can be made for various bias conditions. We know that in equilibrium $pn = n_i^2$ and therefore $U=0$. For the forward bias case we have an increase in carrier concentration $pn \geq n_i^2$ and the diode will try to balance the concentrations with the generation-recombination mechanism of capturing. For the n-side (2.11) we have

$$n_n p_n = n_{n0} p_{n0} e^{\frac{(qV)}{kT}} = n_i^2 e^{\frac{(qV)}{kT}} \quad (2.135)$$

Substituting (2.132) into (2.133)

$$U = \frac{c_0 N_t n_{n0} p_{n0} e^{\left(\frac{qV}{kT}-1\right)}}{n_{n0} + p_{n0} + 2n_i \cosh\left(\frac{(E_t - E_i)}{kT}\right)} \quad (2.136)$$

With (2.136) some aspects of the recombination rate can be made. Looking at the denominator we observe two facts. First For all semiconductors n_i and \cosh are fixed values for given trap level. Second U is maximum when $n_{n0} + p_{n0}$ is minimum. This happens when $n_{n0} = p_{n0}$ and occurs when E_i is halfway between E_{fp} and E_{fn} in the depletion region. Then

$$n_{n0} = p_{n0} = n_i^2 e^{\left(\frac{qV}{2kT}\right)} \quad (2.137)$$

Substituting (2.135) in (2.136) we have:

$$U_{max} = \frac{c_0 N_t n_i e^{\left(\frac{qV}{kT}-1\right)}}{2 \left[e^{\left(\frac{qV}{2kT}\right)} + 1 \right]} \quad (2.137)$$

Simplifying even further for $E_t = E_i$ and for values of $V \geq \frac{4kT}{q}$ (2.137) becomes

$$U_{max} = \frac{c_0 N_t n_i}{2} e^{\left(\frac{qV}{2kT}\right)} \quad (2.138)$$

The recombination current density J_{gr} is defined

$$J_{gr} = \int_0^W q U dx \quad (2.139)$$

Substituting U_{max} we get

$$J_{gr} = \frac{q W n_i}{2\tau} e^{\left(\frac{qV}{2kT}\right)} \quad (2.140)$$

where we define $J_R = \frac{q W n_i}{2\tau}$ and τ is the effective recombination lifetime and is equal to

$$\tau = \frac{1}{c_0 N_t} \quad (2.141)$$

The total current for the forward bias is then

$$J = J_s e^{\left(\frac{qV}{kT}\right)} + J_R e^{\left(\frac{qV}{2kT}\right)} \quad (2.142)$$

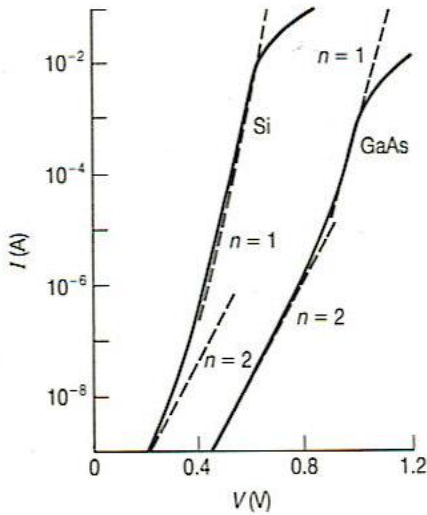


Figure 2.12 Ideality factor for Si and GaAs. Figure from [2]

The result is that the diffusion current dominates over the generation-recombination current. The reason is the difference in the exponential term. One should note that for low values this isn't the case. Finally we see that the forward current depends on the first exponential term with the form

$$J \propto e^{\left(\frac{qV}{nkT}\right)} \quad (2.143)$$

where n is the ideality factor. It shows what kind of current dominates. Typically values are $n=1$ (diffusion current dominates) and $n=2$ for recombination current. We can see the ideality factor in figure 2.12.

2.2.1.4.2 Reverse bias

In reverse bias the, carrier concentration decrease and $pn < n_i^2$ within the depletion region. The mechanism is now emission of electrons and holes. Using (2.133) gives a negative recombination rate, namely $-U$. This means that generation rate is positive and responsible for current in the reverse bias junction. Since $n_i > p, n_i > n$ we have that (2.133) is

$$G = \frac{c_0 N_t n_i}{2 \cosh\left(\frac{(E_t - E_i)}{kT}\right)} \quad (2.144)$$

G is maximal when $E_t = E_i$, exactly like in the forward case. With arithmetic calculation like in the calculation for U , it can be shown that G falls off exponentially as E_t moves away from E_i . The total current due to generation is then

$$J_{gr} = \int_0^W q G dx = \frac{qWn_i}{\tau} \quad (2.145)$$

The total current is then

$$|J| = J_S + \frac{qWn_i}{\tau} \quad (2.146)$$

In figure 2.13 we see a more realistic diode. We can see that the p region is heavily doped (p^+) and has much smaller dimension, then our ideal case. Remember we want the absorption of light to be maximal. Hence build the diode with its depletion region close to length (depth) where maximum absorption is happening. This depth is close to the surface. For this reason also $N_A \gg N_D$ and

$$x_n = \left(\frac{2\epsilon\epsilon_0 V}{qN_D}\right)^{\frac{1}{2}} \quad \text{and} \quad x_p = \left(\frac{2\epsilon\epsilon_0 V}{qN_A^2}\right)^{\frac{1}{2}}$$

Also from (2.144) we have

$$|J| = \frac{qD_p}{L_p} p_{n0} + \frac{qWn_i}{\tau} = q \sqrt{\frac{D_p}{\tau_p}} \frac{n_i^2}{N_D} + \frac{qWn_i}{\tau} \quad (2.147)$$

Also from (2.145) we have that n_i is small for silicon and the diffusion current dominates and the reverse current saturates at

$$|J| = J_S \quad (2.148)$$

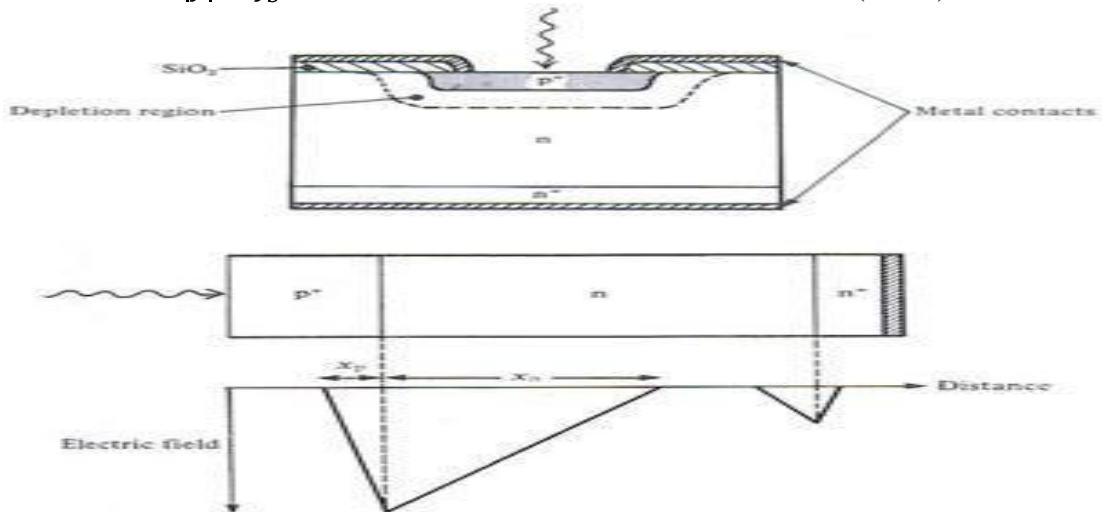


Figure 2.13 A np+ diode for light detection [1]

2.2.1.4.3 Dark current

Equation (2.121), also known as dark current, is in reality an overestimated current because there are two major factors that contribute to the dark current. These are the generation-recombination and the tunneling currents. We saw that the generation-recombination current is given by (2.147). From the equation we note that the diffusion current is dependent on n_i^2 whereas the generation current is proportional to n_i . Therefore diffusion current has a band gap and temperature dependency whereas the generation-recombination has not. At low temperature J_{gr} dominates and the total reverse current I_R is proportional to the reverse voltage V_R by $I_R \propto V_R^{\frac{1}{2}}$. This can be also seen from the depletion region dependency on V_R . At about 175 °C the diffusion current becomes dominant as the current saturates. Figure 2.14 shows dark currents at the same level of temperature for various semiconductors.

Also important are tunneling currents which are present in the diode, but their effect is only valid when high fields are involved. Therefore they are relevant only in avalanche photodiodes (APD's).

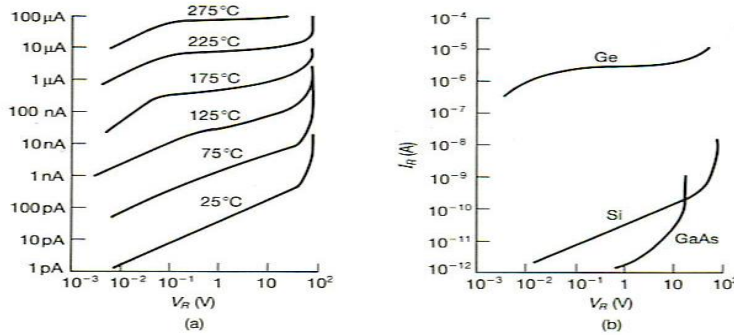


Figure 2.14 a) dark current variations with temperature b) dark current in various semiconductors with constant temperature. Figure from [2].

2.2.1.5 Photocurrent in reverse bias

To derive the photocurrent for the reversed biased pn junction we once again look at the photon absorption mechanisms. These are depicted in Figure 2.15 and a short description follows

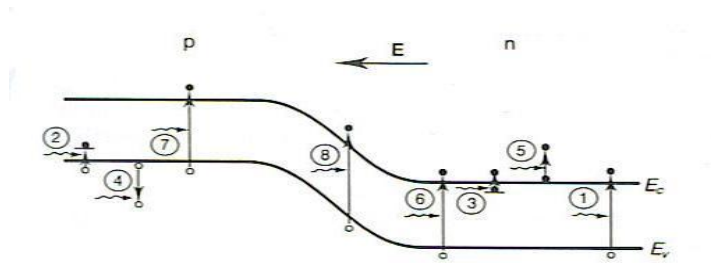


Figure 2.15 Photon absorption mechanisms. Figure from [2]

1. If Photon energy is greater or equal to the band gap we got electron-hole pairs creations far away from the depletion region. These are driven by diffusion and therefore recombine after travelling the diffusion length. The result is no charge is detected.

2. A state that does occur but is not very significant is the creation of holes due to absorption of energy less than the band gap. The holes are released via acceptor states, which are mostly occupied anyway.
3. Same as 2 but with donor states.
4. Free carrier absorption within the valence band. This elevates electrons to higher states and lowers holes to lower states. No extra carriers are produced.
5. Same as 4, with the difference that it happens in the conduction band.
6. Same as 1, except that the creation of pairs happens near the depletion region, which separates the electron and hole in such way that the electron becomes minority carrier in the n-type and the hole by diffusion is swept away through the electric field to the p-type side. As a result we got more majority carriers now in the p-type which leads to stripping away more minority carriers. Henceforth an additional charge q is detected.
7. Same as 6 with an electron swept across the junction and increasing the minority concentration in the n-type.
8. In the depletion region when an electron-hole pair is created, they are split by the electric field, and swept to the side of their charge type, namely to the n and p region. One might think that this contributes two charges to the current but the reality is that the particles q and h move not all the way the junction. The net flow contributes only one q of charge.

The conclusion is that only mechanisms six to eight contribute to the photocurrent and number eight is the most desired, since it doesn't have a long time delays.

Another detour before we actually extract the photocurrent must be done. This time we seek the value of power absorption in the pn junction. Suppose the pn junction is illuminated from p-type side. This type of assumption is practical when considering real diodes like the one shown in figure 2.13. The power is the power of the electromagnetic wave travelling inside the semiconductor. Like explained in appendix A, electromagnetic wave is only a portion of the original wave prior to entering the surface of the semiconductor. Consequently the equation (A.45) holds also when defining the refracted wave as starting point of the decay of its power. At depth x we can express then the remaining (absorbed) power as

$$P(x) = P_0(1 - R)e^{-ax} \quad (2.149)$$

where a the absorption coefficient.

We turn now to the depletion region current. Defining x_p as the beginning and x_n as the end of the depletion region, then according to (2.149) the power for the depletion region is

$$P(x_p) - P(x_n) = P_0(1 - R)[e^{-ax_p} - e^{-ax_n}] \quad (2.150)$$

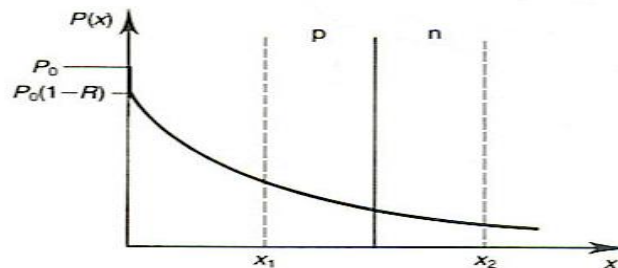


Figure 2.16 Photon absorption depth. Figure from [2]

The current density generated in the depletion region is then

$$J_0^{DR} = \frac{q\eta_{int}}{hf_v} \Phi_0 (1 - R) [e^{-ax_p} - e^{-ax_n}] \quad (2.151)$$

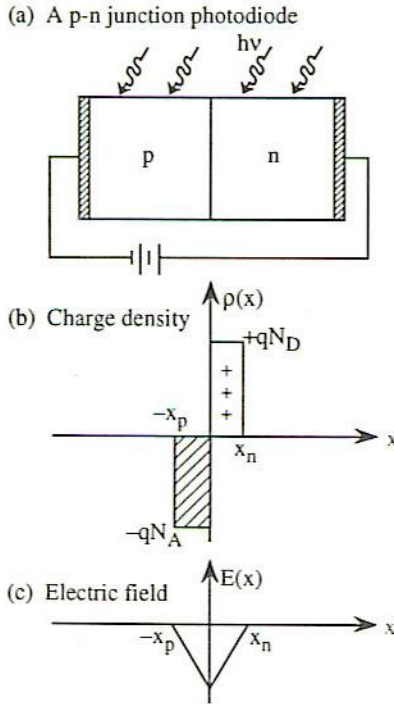


Figure 2.17 Pn junction with uniform light.
Figure from [4]

This expression has the advantage of not containing any dimensions of the diode. Note that multiplying by the area where light is impinging, turns (2.151) into an equation containing P_0 and current I_{DR} . This current like we mentioned above is not the only current contributing to the photocurrent, current by diffusion is also present. This current will be present on both sides of the depletion region. Therefore we are setting up the origin of the axial system to be at the depletion region, see figure 2.17.

Using the continuity equation (2.52) and representing the generation rate as the optical generation rate. For the n-type side we have:

$$D_p \frac{d^2 p_n(x)}{dx^2} - \frac{p_n - p_{n0}}{\tau_h} + g(x) \quad (2.152)$$

Solving for $g(x)$ and using (2.151) we get that the optical generation rate at depth x is

$$g(x) = \frac{\eta_{int} \Phi_0 (1 - R) a}{hf_v} e^{-ax} \quad (2.153)$$

To find a solution to (2.152) we assume uniform optical generation rate G_0 , in example we assume a constant rate at any depth. With this a solution is

possible by calculating the homogeneous solution and the particular solution:

$$\Delta p_n(x) = c_1 e^{\frac{-(x-x_n)}{L_p}} + c_2 e^{\frac{(x-x_n)}{L_p}} + G_0 \tau_p \quad (2.154)$$

If the n region is very long we can set $c_2 = 0$. One can see then as $x \rightarrow \infty$ $\Delta p_n(x)$ approaches the total photogenerated holes $G_0 \tau_p$. At $x = x_n$ the hole concentration is held steady by the voltage bias V

$$p_n(x = x_n) = p_{n0} e^{\frac{qV}{kT}} \quad (2.155)$$

Therefore we obtain the homogenous solution

$$\Delta p_n(x = x_n) = p_{n0} \left(e^{\frac{qV}{kT}} - 1 \right) \quad (2.156)$$

The final solution with the particular solution is

$$\Delta p_n(x) = \left[p_{n0} \left(e^{\frac{qV}{kT}} - 1 \right) - G_0 \tau_p \right] e^{\frac{-(x-x_n)}{L_p}} + G_0 \tau_p \quad (2.157)$$

We know from (2.116) that hole diffusion is defined like $J_p = \frac{qD_p}{L_p} p_{n0} \left[e^{\left(\frac{qV}{kT} \right)} - 1 \right]$.

Noticing from (2.112) and (2.116) that $J_p(x)$ can be expressed otherwise by rearranging the terms and by partial derivative we can get an equivalent expression [4]

$$J_p(x) \cong -qD_p \frac{\partial}{\partial x} \Delta p_n(x) \quad (2.158)$$

Substituting (2.157) into (2.158) we get

$$J_p(x) = \frac{qD_p}{L_p} \left[p_{n0} \left(e^{\frac{qV}{kT}} - 1 \right) - G_0 \tau_p \right] e^{\frac{-(x-x_n)}{L_p}} \quad (2.159)$$

At the boundary x_n , $J_p(x)$ becomes:

$$J_p(x_n) = \frac{qD_p}{L_p} p_{n0} \left(e^{\frac{qV}{kT}} - 1 \right) - qG_0 L_p \quad (2.160)$$

Where we can see that the first term is due to voltage bias and the last term is due to optical generation. A similar approach is made for the current density at $x = -x_p$.

$$J_n(-x_p) = \frac{qD_n}{L_n} n_{p0} \left(e^{\frac{qV}{kT}} - 1 \right) - qG_0 L_n \quad (2.161)$$

Once again to find the current I we multiply by the area A of the diode, note that A is now the cross-sectional area of the diode since we took the origin at the depletion region. Hence

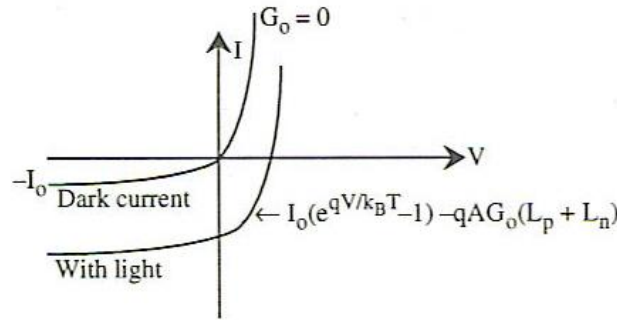
$$I = A[J_p(x_n) + J_n(-x_p)] = I_0 \left(e^{\frac{qV}{kT}} - 1 \right) - qG_0 A(L_n + L_p) \quad (2.162)$$

where

$$I_0 = qA \left(\frac{D_p}{L_p} p_{n0} + \frac{D_n}{L_n} n_{p0} \right) \quad (2.163)$$

is the diode reverse current. The last term of (2.162) is the photocurrent we sought.

Figure 2.18 shows that the current increases due to the optical generation rate increase.



2.18 Optical generation rate increases the current in the reverse biased diode. Figure from [4]

2.3 References

- [1] Wilson, John and Hawkes, John., *Optoelectronics and introduction*. s.l. : Prentice Hall, 1998.
- [2] Wood, David., *Optoelectronics Semiconductor Devices*. s.l. : Prentice Hall International Series in Optoelectronics, 1994.
- [3] Sze, S M., *Physics of semiconductor devices*. New York : John Wiley and Sons, 1981.
- [4] Chuang, Shun Lien., *Physics of optoelectronics devices*. s.l. : A Wiley-Interscience Publication John Wiley & Sons Inc., 1995.
- [5] Klingshirm, C F., *Semiconductor Optics*. s.l. : Springer, 1997.

Chapter 3. Photodetectors based on Silicon

3.1 Noise sources

A detectors principle is to convert optical signals into electronic signals. The conversation is done with the following steps:

- Creation of electron-hole pairs: More precisely the carrier generation and recombination processes.
- Assemble and identification of electrons and holes and amplify them, to achieve a best possible gain in current
- Construction of a desirable output for various external components.

The main features of a photodetector should be *sensitivity to light for a distinct wavelength, response speed, and the detectors noise*.

The basic equations for the photoconductive devices have been written in chapter two, one important aspect of choosing one detector over the other are the noise sources that can exist in them. The better one eliminates noises, the better radiant flux (radiant power) can be detected. There are three kinds of noises to lookout:

1. **Shot noise** is caused by random arrival time of the signal carriers (electron, photon, etc.). If not taken proper care of this noise, phenomena's like random fluctuations in the measurement signal could be observed.
2. **Flicker noise** or also known as 1/f noise, appears in various forms, for example just to name two; impurities in a conductive channel generation and recombination noise in a transistor (base current). It's always accompanied with direct current.
3. **Thermal noise, Johnson noise, or Nyquist noise** is the noise generated by the thermal changes due to the charge carriers (the electrons) inside an electrical conductor in equilibrium, which happens regardless of any applied voltage. Thermal noise is considered to be white, implying that the power spectral density is equal throughout the frequency spectrum. Further the amplitude of the signal resembles a Gaussian probability density function.

Despite that, generally noise is often in practise described by the *noise equivalent power (NEP)*. NEP is the radiant flux for which a signal-to-noise ratio of unity is provided at the output of a given optical detector at a given data-signaling rate or modulation frequency, operating wavelength, and effective noise bandwidth. It's a nonlinear function of the bandwidth and is valid only if the dark-current noise dominates the noise level. There can be differences in the definition of NEP, depending on the manufacture or authors, for instance some define NEP for a 1 Hz bandwidth, when defined like that, NEP has watts units.

$$NEP = \frac{S_n \sqrt{NB_{eff}}}{R_{ph}} \quad (3.1)$$

where S_n noise spectral density ($A / Hz^{1/2}$), NB_{eff} the effective noise bandwidth. Similar the *specific directivity* D^* of a photodetector is a measure to characterize performance. Like NEP it's normalized to unit area and unit bandwidth. It's defined as:

$$D^* = \frac{\sqrt{A^* NB_{eff}}}{NEP} \quad (3.2)$$

where A the detectors area and NB_{eff} as described above. Hence it can be expressed as the reciprocal of NEP, since with substitution we get:

$$D^* = \frac{R_{ph} \sqrt{A}}{S_n}$$

3.2 Bulk Semiconductor

In a bulk semiconductor, like described in section 3.1, the energy of the photons is absorbed and this energy must be balanced through mechanism that ensures the energy conservation of the system. In a semiconductor the main mechanism is the interband transition. The mechanism deals with the surplus of energy by shifting the generation rate and therefore increasing electron-hole pairs. The result is a current flowing through the semiconductor. If we have an intrinsic semiconductor the carrier's mobility and lifetime must be sufficient to reach the edges. Carriers move freely through the material in random directions. The collection of the current, namely the created carriers, is been done by ohmic contacts. To attract the carriers to the contacts, an external field (voltage) is applied, consequently biasing the direction towards the contacts. The current generation doesn't depend on the applied voltage. One could increase the strength of the field yielding more carriers. This only forces more carriers to move to the contacts but not be created. The current therefore depends on the incident radiant density P_{ph} . To evaluate this we start with the equation that gives the numbers of photons per second in radiant flux. Like stated above the radiant flux is a function of the depth x . Conclusively $\frac{P_{ph} x}{hv}$.

Further the numbers of electron-hole pairs created per second is obtained by multiplying by the quantum efficiency η , so

$$\frac{\eta P_{ph} x}{hv} \quad (3.3)$$

Combining (2.149) and (2.153) we get

$$g_o(x) = \frac{\eta \alpha P_{ph}(0)}{h\nu LD} e^{-\alpha x} = \frac{\eta \alpha P_{ph}(0)}{h\nu LD} (1 - e^{-d\alpha}) \quad (3.4)$$

Taking in account that the absorption coefficient is inversely proportional to the thickness then since $d \gg 1/\alpha$, the term in the exponent becomes zero and (3.20) resolves to

$$g_0(x) = \frac{\eta P_{ph}}{h\nu LD} \quad (3.5)$$

which we already saw in (2.66). If we integrate over the whole thickness d of the semiconductor the generation rate per unit volume is determined as:

$$G_0 = \int_0^d g_o(x) dx \quad (3.6)$$

The photogenerated process involves increase in carriers such that at all time $\Delta p = \Delta n$ holds. This is logical assumption, since electron-hole pairs are created, thus must be equal in increase. Further if the optical generation rate is uniformly distributed over the semiconductor we can assume that the increase in carrier concentration is analogous to the light. Hence the generation of carriers equals $G_n(t) = G_{I_0}$, with G_{I_0} as the light intensity. The extra carrier generation, namely the excess carriers, exists as long the light source is impinging with constant intensity at the semiconductor. Each excited carrier recombines after time τ to the valence band and simultaneously another is released to the conduction band. The conclusion is simple: an excess carrier (Δp , Δn) depends on G_{I_0} as well on τ . Hence $\Delta n = G_{I_0}\tau$. Combining the above we can state that

$$R = \frac{\Delta p}{\tau} = \frac{\Delta n}{\tau} = g_o(x) = G = \frac{\eta P_{ph}' w L}{\hbar \omega w L D} \quad (3.7)$$

where R is the recombination rate. The conductivity, as we know, in a semiconductor, is the due to the densities and mobility's of the n and p concentrations:

$\sigma = nq\mu_n + pq\mu_p$, where q = the charge. Also the conduction current density is given as

$$J_n = q\mu_n nE, \quad J_p = q\mu_p pE \text{ and the total current } J = J_n + J_p = \sigma E = \sigma \frac{V_{applied}}{L}$$

By substitution σ with the extra conductivity $\Delta\sigma$ and the n , p concentration with the excess carriers (Δp , Δn) we get the expression:

$$\Delta\sigma = q\mu_p \Delta n + q\mu_n \Delta p \quad (3.8)$$

By substituting σ by $\Delta\sigma$ and the concentration (n, p) with (Δp , Δn) we can analyze the current density due to the excess carriers as follows

$$\begin{aligned} \Delta J &= \Delta\sigma \frac{V_{applied}}{L} = \left(\frac{q\mu_n \eta P_{ph}' w L \tau}{\hbar \omega w L D} + \frac{q\mu_p \eta P_{ph}' w L \tau}{\hbar \omega w L D} \right) \frac{V_{applied}}{L} \\ &= \frac{q\mu_n \eta P_{ph}' w L \tau}{\hbar \omega D} \left(1 + \frac{\mu_p}{\mu_n} \right) \frac{V_{applied} \tau}{L} \end{aligned} \quad (3.9)$$

The light induced current will be the current density J multiplied by the cross-sectional area $A = WD$.

$$\Delta I = \Delta J A = \frac{q\eta P_{ph}' w \mu_n}{\hbar \omega} \left(1 + \frac{\mu_p}{\mu_n} \right) \frac{V_{applied}}{L} \quad (3.10)$$

The current is defined as $I_{ph} \equiv \frac{\eta q P_{ph}}{\hbar \omega}$ and the for whole semiconductor the total radiant flux is $P_{ph} = P_{ph}' WL$ then (3.15) can be rewritten as

$$\begin{aligned}\Delta I &= I_{ph} \left(1 + \frac{\mu_p}{\mu_n} \right) \frac{V_{applied} \mu_n \tau}{L^2} \stackrel{E=\frac{V}{L}}{=} I_{ph} \left(1 + \frac{\mu_p}{\mu_n} \right) \frac{E \mu_n \tau}{L} \\ &= I_{ph} \left(1 + \frac{\mu_p}{\mu_n} \right) \frac{v_n \tau}{L} = I_{ph} \left(1 + \frac{\mu_p}{\mu_n} \right) \frac{\tau}{t_r}\end{aligned}\quad (3.11)$$

where t_r is the average time a carrier needs to traverse the length of the device, called transient time. The gain of a bulk semiconductor photodetector is defined as the carrier collection rate at the contacts to the carrier generation rate. Mathematically

$$G = \frac{\frac{\Delta I}{I_{ph}}}{\frac{q}{g_0 w L D}} \quad (3.12)$$

Substituting (3.17) with the alternative expression of the transient time, the gain becomes

$$t_r = \frac{L}{u_n} = \frac{L}{\mu_n E} = \frac{L}{\mu_n V_{applied}} = \frac{L^2}{\mu_n V_{applied}} \Rightarrow G = \frac{\tau(\mu_n + \mu_p) V_{applied}}{L^2} \quad (3.14)$$

Conclusively the gain of a semiconductor of length L is a factor of three parameters: the recombination time, the charge mobilities and the strength of the applied field. For small lengths silicon devices gains up to 10^6 are expected.

We have mentioned various noises that can exist in a bulk semiconductor, to derive expressions for the NEP we modulate the optical intensity as a sinusoidal signal $G_0(t) = G_0(1 + m \cos \omega t)$. In chapter two we derived the rms value of the photocurrent:

$$i_{o_{rms}} = q\eta \frac{P_{rms}}{h\nu} \left(\frac{\tau_n}{\tau_t} \right) \frac{1}{\sqrt{1 + \omega^2 \tau^2}} \quad (3.15)$$

For the coming analysis of the noise factors we need to transform the photocurrent from the time to the frequency domain. This is been done by the Fourier transform function. From [3] we restate the equations:

$$\text{Average Power over time period } T \quad P = \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} i^2(t) dt = \int_0^\infty S(f) df$$

$$\text{Spectral density function} \quad S(f) = \frac{2}{T} |i(f)|^2$$

$$\text{Thermal noise} \quad \langle i_T^2 \rangle = \frac{4k_B T B}{R} \quad (3.16)$$

Flicker noise (generation-recombination noise)

$$\langle i_{GR}^2 \rangle = \frac{4qI_{ph} \frac{\tau}{\tau_t} B}{1 + (2\pi f \tau)^2} = \frac{4qBI_{ph}}{1 + (\omega_0^2 \tau^2)} \frac{\tau}{t_r} \quad (3.17)$$

Finally we now can define the most important component for defining the NEP, the signal to noise ratio SNR.

$$\left(\frac{S}{N}\right)_{power} = \frac{i_{rms}^2}{\langle i_T^2 \rangle + \langle i_{GR}^2 \rangle} = \frac{\eta m^2 \frac{P_{ph}}{h\nu}}{8B \left[1 + \left(\frac{k_B T}{q R_{ph} I_{ph}} \right) \left(\frac{\tau}{\tau_i} \right) 1 + \omega^2 \tau^2 \right]} \quad (3.17)$$

$$= \frac{m^2 I_{ph}}{8qB} \left(1 + \frac{\mu_p}{\mu_n} \right) \left[1 + \frac{k_B T}{q} \frac{\tau_i}{\tau} \frac{1}{I_{out} R_{ph}} 1 + \omega^2 \tau^2 \right]^{-1}$$

Solving for P_{ph} we find that the average power P_{ph} impinging onto the detector must be at least

$$P_{ph}^{min} = \frac{4\hbar\omega B(SNR)}{\eta M^2 \left(1 + \frac{\mu_p}{\mu_n} \right)} \times \left\{ 1 + \left[1 + \frac{M^2 k_B T}{2q^2 R_p B(SNR)} \left(\frac{\tau_r}{\tau} \right)^2 (1 + \omega^2 \tau^2) \right]^{\frac{1}{2}} \right\} \quad (3.18)$$

Setting $SNR=1$ and taking the rms value ($P_{ph}^{min} * \frac{1}{\sqrt{2}}$) we get the NEP:

$$NEP = \frac{2\sqrt{2}\hbar\omega B}{\eta M \left(1 + \frac{\mu_p}{\mu_n} \right)} \left\{ 1 + \left[1 + \frac{M^2 k T}{2q^2 B R_p} \left(\frac{\tau_r}{\tau} \right)^2 (1 + \omega^2 \tau^2) \right]^{\frac{1}{2}} \right\} \quad (3.19)$$

NEP indicates the minimum detectable radiation level; the smaller the NEP value, the better the performance. Here we see that NEP can be quite high. If we decrease the shunt resistor R_p we increase thermal noise attribution, and therefore the total noise as also NEP will indicate.

To summarize the features of a bulk semiconductor device as photodetector

- High gain possible
- Slow response
- High noise

3.3 Diodes

3.3.1 PN Junction

Easy to fabricate with the bulk silicon complementary metal oxide semiconductor (CMOS) process is the photodiode. It is a good photodetector compared with the previous one. Its principle lies within the creation of a junction of two different semiconductor types: one n-type and one p-type. We will repeat the basic principles of the pn junction, for further details please refer to chapter two. The formation of the depletion region is due to the battle between the drift and diffusion currents. These currents when equally strong, are the reason of the formation of the depletion region or in other words the formation of the region of space charge. It is called like that because at the boundaries of

the junction one side is been negatively and the other is been positively charged. The result of this electric field is the creation of an electric field (net field) which pushes all free charges out of the depletion region to the positive or negative charged n or p junction region. Hence no further motion of carriers is detected and the PN junction is said to be in equilibrium. This also explains the potential, also called the build in potential, which a PN junction may have without applying any external source. When light impinges on the junction diode, electron-hole pairs are created everywhere, especially in the depletion region the carriers are swept to the N and P regions like explained above. Holes and Electrons generated in the N and P regions diffuse to the depletion region and hence are transferred to the opposite side from their origins. Carriers are detected in the depletion region either as photocurrent or as photovoltage.

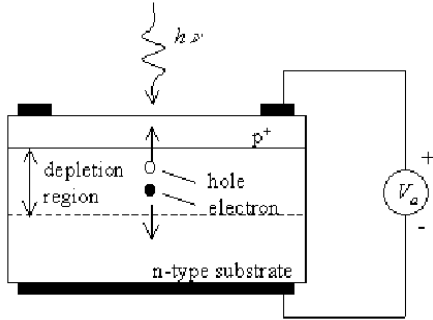


Figure 3.1. Pn junction. Figure from [7]

Therefore detection increases by making the depletion layer thicker such that more photons are been absorbed. On the other hand making these changes one has to consider that this increases the transit time of the carriers which in turn slows down the response time of the photodetector. Thus a trade-off between the response time and quantum efficiency exists. Connecting an voltage source at the PN junction terminals we can set it in forward or in the

reserved bias mode. The forward bias mode increases the carrier diffusion from one side (P region) to the other (N region). Thus setting up an external field larger than the build in potential ψ causes a current to flow from the P to the N region and the depletion region doesn't change in size. In the reserved bias mode the depletion region is increased due to the supply of additional potential in the depletion region and hence more photons can be detected. In both modes the depletion region can be seen as capacitance called the junction capacitance C_j . This capacitance is not proportional to the applied voltage V_a but varies with small changes in the voltage, therefore a small signal analysis yields that

$$C_j = A \sqrt{\left[\frac{e\epsilon_0\epsilon_r}{2\left(\frac{1}{N_A} + \frac{1}{N_D}\right)(|\psi| - V_a)} \right]} \quad (3.38)$$

with A area of the depletion region, N_A and N_D the number of acceptor/donor atoms per m^3 , e the electronic charge magnitude $1.60 \times 10^{-19} C$, ϵ_0 permittivity of a vacuum $8.85 \times 10^{-12} Fm^{-1}$, $\epsilon = \epsilon_0\epsilon_r$ the relative permittivity. C_j is regulated through the applied voltage, it is greater for forward than for reserved bias, and decreasing drastically with increasing reserve bias. In both modes to detect the photocurrent generated we measure an external resistance called the load resistance R_L . For both modes one can either measure the voltage (photovoltaic mode) or either the current across the resistance (photoconductive mode). Both modes are shown with their equivalent circuits respectively.

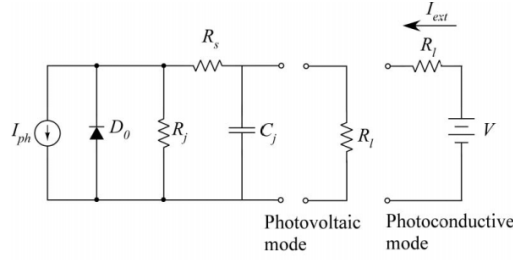


Figure 3.2. Pn equivalent models. Figure from [10]

To explain the terms photovoltaic and photoconductive a little bit more one has only to look at the I-V characteristics of a PN junction, as shown in chapter two in figure 2.18. Taking the short circuit I_{sc} and the the open circuit V_{oc} one can see that these points are crossover points of the current and the applied voltage. The negative but steady increasing part of the current is called photoconductive. The quadrant in which V_{oc} is positive and I_{sc} is negative is called photovoltaic, because the junction is producing power.

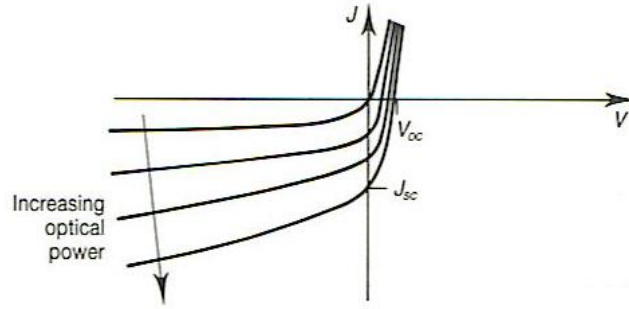


Figure 3.3. Finding the short circuit current and open circuit voltage of a pn junction. Figure from [2]

The current I_{ext} is been given as follows:

$$I_{ext} = I_{ph} - I_0 \left(e^{\frac{-V_a + I_{ext}(R_s + R_j)}{V_T}} - 1 \right) + \frac{V_a - I_{ext}(R_s + R_j)}{R_j} \quad (3.20)$$

where I_0 the minimal current (leakage current) of the diode D_0 and V_T the defined thermal voltage $\frac{kT}{q}$. Since R_j is usually very large the last term of (3.20) can be cancelled out. In

the photovoltaic mode to measure the voltage due to the incident optical power P_{ph} we proceed to an open-circuit voltage V_{oc} measure of the photodiode:

$$V_{oc} = V_a(I = 0) = V_T \ln \left(\frac{I_{ph}}{I_0} + 1 \right) = V_T \ln \left(\frac{q\eta P_{ph}}{\hbar\omega I_0} + 1 \right) \quad (3.21)$$

As we can see from (3.2) the strong internal electric field in the depletion region is the reason for the fast response of the PN junction. Three important time factors that make up a fast response are

1. the diffusion time in the N and P regions
2. the drift time, also called the charge collection time t_c , in the depletion region
3. the time spend on the load resistance and the parasitic diode capacitance

The first factor is important for weakly absorbing structures (i.e. when the diffusion length is less than the absorption depth) and is expressed as described in [3].

$$\begin{aligned} \text{p-on-n} \quad t_d &= \frac{1}{13} \left(\frac{3}{\alpha} - 0.54 \rho_n^{\frac{1}{2}} (V_{bi} + V_a)^{\frac{1}{2}} \times 10^{-4} \right)^2 \\ \text{n-on-p} \quad t_d &= \frac{1}{36.4} \left(\frac{3}{\alpha} - 0.32 \rho_p^{\frac{1}{2}} (V_{bi} + V_a)^{\frac{1}{2}} \times 10^{-4} \right)^2 \end{aligned}$$

The second factor is defined as $t_c = \frac{w}{u}$ where w is the depletion width and u the carrier saturation velocity. The third factor is the time taken to discharge the junction capacitance C_j through the load resistor R_L . The time can be calculated with a simple RC model $\rightarrow t_{RC} = 2.2 R_L C_j$. To manipulate some of the above factors with the purpose of achieving a high response speed we can do the following:

- Reducing the drift time of the PN junction by applying high value of reserve bias
- A strong reserved biased junction minimizes most of the parasitic capacitance

Unfortunately limitation to the drift speed of the carriers exists because it is proportional to $\sqrt{|V_a|}$ so care has to be taken when trying to improve the drift time.

We will further analyse the noise parameters in the PN junction. Assuming that the optical intensity is modulated as $P(t) = P_{ph}(1 + m \cos \omega t)$ then the average photocurrent is then described by

$$I_{ph} = \frac{q\eta P_{ph}}{\hbar\omega} \xrightarrow{P(t)=m*P_{ph}} \frac{P(t)=m*P_{ph}}{I_{ph}^{rms} = I_{ph} * \frac{1}{\sqrt{2}}} \rightarrow \langle i^2 \rangle_{ph}^{rms} = \frac{q\eta m P_{ph}}{\hbar\omega \sqrt{2}} \quad (3.22)$$

Two noises prevail: the shot and the thermal noise. Sources of the shot noise are

1. The background current I_B generated by random light sources unrelated to the incident light.
2. the dark current I_D due to the reserve saturation current of the diode in the depletion region and the thermal generation of electron-hole pairs.
3. the photocurrent I_{ph} itself

They can be interpreted as independent random process each contributing to the shot noise. Taking the sum of all we have

$$\langle i_{sh}^2 \rangle = 2q(I_{ph} + I_B + I_D)B \quad (3.23)$$

where B the bandwidth of the photodiode. The resistances R_j , R_l and R_i (R_i is the resistance needed if the PN junction is connected to another circuit with an finite input resistance) are combined to an equivalent resistance R_{eq} which makes up for the total thermal noise. Substituting in (3.16) R_{eq} for R we get:

$$\langle i_{th}^2 \rangle = \frac{4kT}{R_{eq}} B \quad (3.24)$$

Again we can model the PN junction with the various contributions of noise. We have all the necessary information to calculate the SNR of the PN junction:

$$SNR = \frac{\frac{1}{2} \left(\frac{q\eta m P_{ph}}{\hbar\omega} \right)^2}{2q(I_B + I_D + I_{ph})B + 4kT \frac{1}{R_{eq}} B} \quad (3.25)$$

Setting $I_{eq} = I_B + I_D + I_{ph} + 2kT \frac{1}{qR_{eq}}$ we find the minimal radiant flux P_{ph} necessary to obtain a specific SNR as follows:

$$P_{ph}^{\min} = \frac{2\hbar\omega B}{\eta m^2} (SNR) \left\{ 1 + \left[1 + \frac{m^2 I_{eq}}{qB(SNR)} \right]^{\frac{1}{2}} \right\} \quad (3.26)$$

Setting SNR equal to one and taking the RMS of the P_{ph}^{\min} we find the NEP

$$NEP = \frac{\sqrt{2}\hbar\omega B}{\eta m} \left\{ 1 + \left[1 + \frac{m^2 I_{eq}}{qB} \right]^{\frac{1}{2}} \right\} \quad (3.27)$$

This is the NEP expression derived for a PN Junction where the quantum efficiency remains unchanged. We can see that in a alternate implementation of a PN junction one can improve parameters that influences the quantum efficiency and therefore minimizes further the NEP.

3.3.2 PIN diode

A good example is the PIN diode. It consist of a p-type and an n-type region which are separated by a layer of intrinsic semiconductor symbolized i, hence the name P-I-N diode. Compared to the PN junction who has a thin depletion region layer the PIN junction has a wide depletion region and the absorption is a function of depth in the material. For a light injected with an radiant flux intensity of $I_{opt} \left(I_{opt} = \frac{P_{opt}}{A} \left(\frac{Watt}{cm^2} \right) \right)$, the generation rate is

$$G(x) = 1 - R \eta_i \left(\frac{I_{opt}}{h\nu} \right) \alpha e^{-\alpha x} \quad (3.28)$$

the injected electrons per unit area is found by taking integrating the generation rate having as boundaries the fact that minimum and maximum generated electron-hole pairs are 0 and infinite respectively:

$$\Phi_0 = \int_0^\infty G \ x \ dx = 1 - R \eta_i \frac{I_{opt}}{h\nu} \quad (3.29)$$

where $\frac{I_{opt}}{h\nu}$ is the number of photons injected per unit area per second and η_i the internal quantum efficiency for the probability of creating an electron-hole pair for each incident

photon. The overall photocurrent is the sum of a drift current created inside the depletion region and the diffusion current created at the P and N regions. The diffusion current is trying to equilibrate the uneven distribution of carriers and as such the unnecessary carriers (hole minority) is diffusing into the reserve-biased junction. Hence the overall current due to light is

$$J_{tot} = J_{dr} + J_{diff} \quad (3.30)$$

Making the hypothesis that P region has no thickness at all we can look at the contribution in the intrinsic region $0 < x < W$. The drift current is due to generation rate over the depletion region, hence:

$$J_{dr} = -q \int_0^W G(x) dx = -q \Phi_0 (1 - e^{-\alpha W}) = -q (1 - R) \eta_i \left(\frac{I_{opt}}{h\nu} \right) (1 - e^{-\alpha W}) \quad (3.31)$$

where minus sign indicates the direction of the current is towards $-x$. The diffusion current due to the minority current density in the N region can be expressed with the below relation [3]. The equation is derived from the continuity equation for semiconductors as explained in [3]

$$J_{diff} = -q D_p \frac{\partial}{\partial x} P_n(x) \quad (3.32)$$

For a steady state situation, the concentration gradient far from the material exists and for that reason setting

$$0 = \frac{\partial P_n}{\partial t} = G(x) - \frac{\delta P_n}{\tau_p} - \frac{1}{q} \frac{\partial}{\partial x} J_p(x) \quad (3.33)$$

To find the excess hole concentration we solve the second order non-homogenous differential equation:

$$\frac{\partial^2}{\partial x^2} \delta P_n - \frac{1}{L_p^2} \delta P_n = -\frac{1}{D_p} G(x) \quad (3.34)$$

D_p is known to be the diffusion coefficient which gives a the relation between diffusivity, mobility and temperature. For a one dimensional motion the coefficient for electrons its

$D_n = \frac{kT}{e} \mu_n$ and for hole its $D_p = \frac{kT}{e} \mu_p$. As known from mathematics the total solution

for equations like (3.34), is the sum of two solutions: the solution of the homogeneous and a particular one, therefore we get

$$\delta P_n = \underbrace{A e^{-\frac{(x-W)}{L_p}}}_{homogeneous} + \underbrace{C e^{-\alpha x}}_{particular} \quad (3.35)$$

Taking the limit $\lim_{\delta P_n \rightarrow 0} \delta P_n$ the term $e^{-\frac{(x-W)}{L_p}}$ becomes infinite and for that reason can be canceled out. Finding the constant C is done by substituting simply G(x) in (3.34). The constant is found to be

$$C = \frac{\Phi_0}{D_p} \frac{\alpha L_p^2}{1 - \alpha^2 L_p^2} \quad (3.36)$$

the boundary condition at depth W gives

$$\delta P_n W = P_{n0} \left(e^{\frac{qV}{kT}} - 1 \right) - P_{n0} \quad (3.37)$$

and therefore the A coefficient can be determined

$$A = -P_{n0} - Ce^{-\alpha W} \quad (3.38)$$

Since we determined all unknown variables the general solution is rewritten as

$$\delta P_n x = -P_{n0} + Ce^{-\alpha W} e^{-\frac{(x-W)}{L_p}} + Ce^{-\alpha x} \quad (3.39)$$

the hole current density on the n side is then

$$J_{diff} = -qD_p \frac{d}{dx} P_n x \Big|_{x=W} = qD_p \alpha \left(1 - \frac{1}{\alpha L_p} \right) Ce^{-\alpha W} - q \frac{D_p}{L_p} P_{n0} = -q \left(\frac{\Phi_0 \alpha L_p}{1 - \alpha L_p} e^{-\alpha W} + \frac{D_p}{L_p} P_{n0} \right) \quad (3.40)$$

Combining (3.12) and (3.22) the total current at x=W becomes

$$J = J_{dr} + J_{diff} = -q\Phi_0 \left(1 - \frac{1}{1 + \alpha L_p} e^{-\alpha W} \right) - \frac{D_p}{L_p} P_{n0} \quad (3.41)$$

The analysis up to now had the purpose to find the total photocurrent. Like promised at the beginning, we now are ready to rewrite the quantum efficiency as a function of all parameters which can be fine-tuned to obtain the optimum or at least a high quantum

efficiency of a PIN junction. Since the quantum efficiency is written as $\eta = \frac{J}{I_{opt}} \frac{q}{h\nu}$

substituting $I_{opt} = \frac{P_{opt}}{A}$ we get

$$\eta = \frac{J}{\frac{P_{ph}}{A\hbar\omega}} = \frac{J}{\frac{\Phi_0}{1-R}} = 1 - R \left(1 - \frac{1}{1 + \alpha L_p} e^{-\alpha W} \right) \quad (3.42)$$

As we can see this type of photodetector fulfill certain criteria's to achieve optimum quantum efficiency:

1. Small reflection coefficient R is desirable
2. Diffusion length L_p and diffusion width W must be large with respect to the absorption depth $1/\alpha$.
3. The diffusion current must be held minimum otherwise the device's response speed decreases dramatically.

Point three in the above list, denotes the fact that there exists a contradiction between making the depletion region too wide and having better quantum efficiency. One must be careful in designing the PIN junction because having a large W automatically means

having a large diffusion current too, which results in a lesser quality of quantum efficiency!

3.4 Bipolar Phototransistor

Generally noted we can state that responsivity in field effect transistors is worse than in bipolar phototransistors. The reason of it is due to its larger channel, which the current flows through. Another difference is that during sensing, bipolar transistors can provide current gain; also its big collector area is capable of collecting more photons than other transistors. When bipolar is chosen as a phototransistor its base terminal is set to float (floating terminal). The function of producing photocurrent I_{ph} is similar to a pnp diode, it consist of sweeping the photogenerated holes of the reverse-biased collector-base junction into the collector. The potential in the emitter-base region is increased by electrons produced in the base region and then pushed by the collector's potential in to the base. The increase of the emitter-base junction potential is responsible of the phenomena that holes from the emitter travel across the base and into the collector region, where they are picked up as additional current. The fact the base terminal is floating the emitter current I_E is equal to the collector current I_C :

$$I_c = I_E = I_{CEO} = (1 + h_{FE})(I_{ph} + I_{eq}) \quad (3.43)$$

where h_{FE} is the static common emitter current gain, I_{eq} represents the background current and dark current, and I_{CEO} is the collector-emitter current with open base. The gain of a phototransistor is $1 + h_{FE}$. Similarly to the photodiode the current of the active junction of a phototransistor contributes to shot noise. At low temporal frequencies, the output noise power is

$$\overline{i_o^2} = 2qI_c \left(1 + \frac{2h_{fe}^2}{h_{FE}} \right) B \quad (3.44)$$

where the small signal h_{fe} is approximately equal to h_{FE} and B the bandwidth of the phototransistor. The total base current is zero cause the junction and recombination currents are balancing the photocurrent, namely the dark and the background current. The latter currents in turn contribute each to the shot noise with spectral density $2q(\frac{I_c}{h_{FE}})$, this

becomes apparent at the output as $4qh_{fe}^2 \left(\frac{I_c}{h_{FE}} \right)$. The collector current contributes from its

part a spectral density of $2qI_c$. We have already noted that the current can flow through a larger area; this is the reason why we have greater signal power but also for our disadvantage, a bigger noise factor. If we modulate the signal for impinging light in paragraph 3.2 the RMS photocurrent will be expressed like

$$i_{ph} = (1 + h_{fe}) \frac{q\eta MP_{ph}}{\hbar\omega\sqrt{2}} \quad (3.45)$$

Since SNR is $\text{RMS}_{\text{output}}/\text{RMS}_{\text{noise}}$ we have

$$SNR = \frac{\frac{1}{2}(1+h_{fe})^2 \left(\frac{q\eta MP_{ph}}{\hbar\omega} \right)^2}{2qI_c \left(1 + \frac{2h_{fe}^2}{h_{FE}} \right) B} \quad (3.46)$$

To find the minimum value of the average optical power P_{ph}^{\min} for a given SNR we proceed like in equations (3.25) and (3.27).

$$P_{ph}^{\min} = \frac{2\hbar\omega B(1+h_{fe}) \left(1 + \frac{2h_{fe}^2}{h_{FE}} \right) (SNR)}{\eta M^2 (1+h_{fe})^2} \times \left\{ 1 + \sqrt{1 + \frac{M^2 (1+h_{fe})^2 I_{eq}}{qB(SNR)(1+h_{FE}) \left(1 + \frac{2h_{fe}^2}{h_{FE}} \right)}} \right\} \quad (3.66)$$

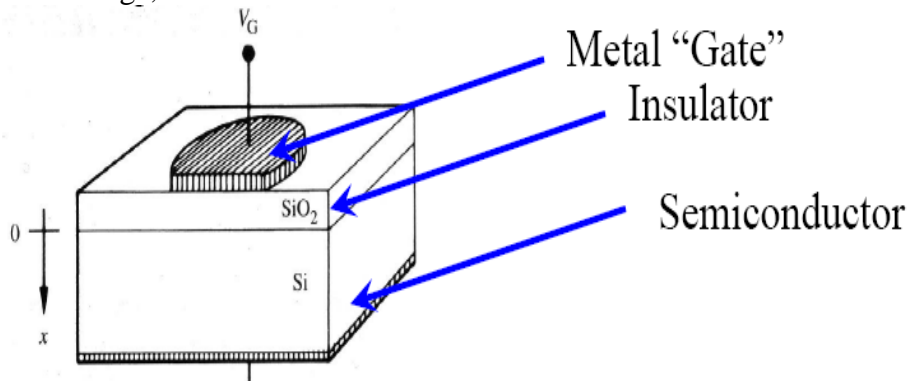
The NEP is found from the relationship $MP_{ph}^{\min} / \sqrt{2} SNR = 1$ and with the approximation that $1 \ll h_{fe} = h_{FE}$

$$NEP = \frac{2\sqrt{2}\hbar\omega B}{\eta M} \times \left\{ 1 + \sqrt{1 + \frac{M^2 I_{eq}}{2qB}} \right\} \quad (3.46)$$

The result is a adjustable noise – gain phototransistor that can be controlled via the common-emitter current gain h_{FE} . Disadvantage is its large collector-base junction which slows down the response, compared to the photodiode. More precisely the above mentioned junction space acts like a capacitance thus the current that must cross the junction, is being slowed down.

3.5 MOS Capacitor (Photo-gate)

Capacitors are widely used with in digital logic circuits, DRAM storage units (storing charge) or simply supply a capacitance for an analog integrated circuit. Also capacitors are known to be the building block for the most common transistor produced – the MOS transistor. The substrate is normally taken to be grounded and the “Gate” electrode can be biased with a voltage, V_G .



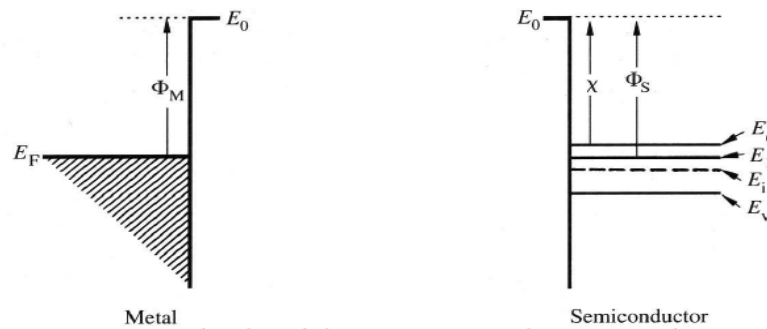


Figure 3.4. A MOS overview, also shown the associated potentials and energies

E_0 = Vacuum Energy Level. The minimum energy an electron must have to free itself from the material.

Φ_M = “Work function” of the metal. This is the energy difference from the Fermi energy (average energy) of an electron in the metal to the vacuum energy level.

Φ_S = “Work function” of the semiconductor. This is the energy difference from the fermi energy (average energy) of an electron in the semiconductor to the vacuum energy level. Note that this energy depends on doping since E_F depends on doping.

χ = Electron Affinity of the semiconductor. This is the energy difference from the conduction band minimum in the semiconductor to the vacuum energy level. Note that this energy does NOT depend on doping

$(E_C - E_F)_{FB} = \Phi_S - \chi$ in the quasi-neutral region where the bands are not bent or are in “flat band”

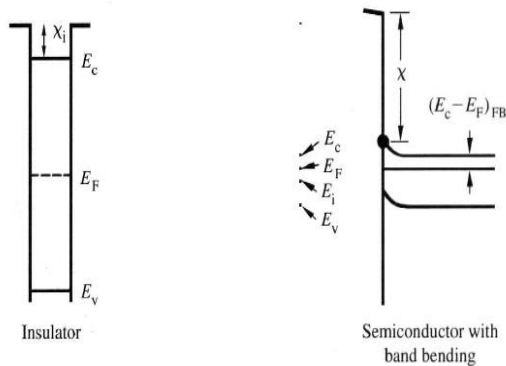


Figure 3.5 MOS' silicon and Insulator region energies differences. Here examined as an isolated case.

Likewise, if no charges are stored on the “plates” (metal and semiconductor regions near the insulator) of the capacitor, the bands are not bent in the insulator nor in the semiconductor. Note that the assumption of an equipotential surface in the metal implies that a perfect conductor can not support an electric field (electrostatics).

The insulator is simply a very wide bandgap, intrinsically doped semiconductor characterized by an electron affinity, χ_i . Later on we will see that due to charge distributions the semiconductor can have an electric field near the insulator that forces the energy bands to bend near the insulator-semiconductor interface. Since the insulator prevents any current from flowing, when we bring the materials together, the fermi-energy must be flat.

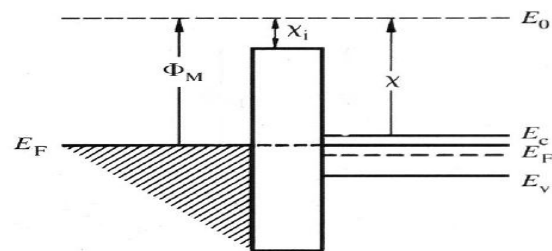


Figure 3.6. MOS' silicon, Insulator and metal region energies and their differences. Here side by side comparison.

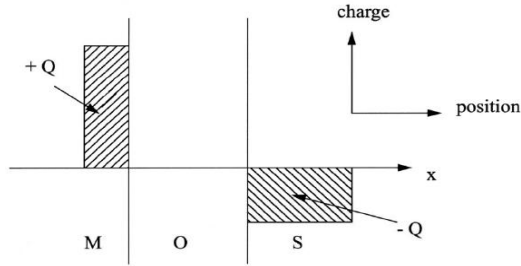


Figure 3.7. MOS'es silicon, Insulator and metal region and their charges.

A positive voltage on the gate puts positive charge on the gate electrode. Gauss's law forces an equal negative charge to form near the semiconductor-insulator interface. Charge separated by a distance creates an electric field across the insulator. In the case of a bias voltage V_G at the gate (metal) we have: For all V_G the Fermi level *in the each layer* remains flat due to zero current through the structure. The applied bias separates the Fermi levels at the metal and semiconductor ends by qV_G

$E_{F(Metal)} - E_{F(Semiconductor)} = -qV_G$. If the semiconductor is grounded:

- metal side Fermi level moves downward if $V_G > 0$
- metal side Fermi level moves upward if $V_G < 0$

Since there are no charges in the oxide we can apply Poisson's equation for the oxide:

$$\frac{dE_{oxide}}{dx} = \rho = 0 \Rightarrow E_{oxide} = Const.$$

$$V = \int E_{oxide} dx \Rightarrow x \text{ is linear to Potential}$$

Hence the potential varies linearly with x , therefore analogous does the energy bands.

The following cases will show the different state of function of the MOS Capacitor for N-type and P-type semiconductors. Afterwards we explain the photo-gate which is based on the MOS principles. For the analysis the following assumptions have been made:

- Metal is an equipotential region.
- Oxide is a perfect insulator with zero current flow.
- Neither oxide nor oxide-semiconductor interface have charge centers.
- Semiconductor is uniformly doped.
- An ohmic contact has been established on the back side of the wafer.
- Analysis will be one-dimensional.
- The semiconductor is thick enough to have a quasi-neutral region (where electric field is zero and all energy bands are flat).
- The energy relationship exist: $\Phi_M = \Phi_S = \chi + (E_C - E_F)F_B$

For an N-type semiconductor we have the phenomena's (operation modes):

- **Accumulation** happens in the case of $V_G > 0$. This causes the lowering of the Fermi-energy ($E = -qV$). Physically this means that holes are pushed away from the semiconductor surface, leaving a depletion layer consisting of ionized acceptors (negative charge) near the surface. Now the insulator has an electric field across it that terminates almost immediately in the near perfectly conducting metal, but

terminates over a finite distance in the semiconductor of “finite resistivity”. Since $n = n_i e^{(E_F - E_i)/kT}$, the electron concentration in the semiconductor near interface increases.

- **Depletion** when $V_G < 0$. Fermi-energy ($E = -qV$) is raised. Electrons are pushed away from the semiconductor surface, leaving a depletion layer consisting of ionized donors (positive charge) near the surface. Once again the insulator has an electric field across it. From $n = n_i e^{(E_F - E_i)/kT}$ we see a decrease in electron concentration in the semiconductor near the interface.
- **Inversion** occurs for higher magnitudes of bias $V_G < 0$, namely V_G equals a voltage called threshold voltage V_T . First when $V_G < 0$ the fermi-energy near the interface crosses the intrinsic energy and the “type” of material swaps from n-type to p-type (in the area of the interface). The charge model indicates that positive charge must be created in the semiconductor near the interface. This charge is in the form of ionized donors and holes. Near the surface the donor concentration must equal the hole concentration, so we can write,

$$\begin{aligned} P_{\text{interface}} &= N_D \\ P_{\text{interface}} &= n_i e^{(E_i - E_F)/kT} \end{aligned} \quad (3.47)$$

The onset of inversion occurs for a voltage called the threshold voltage V_T . Calculating the charge distribution as a function of position in the semiconductor we find that inversion takes place when

$$E_{i-\text{INTERFACE}} - E_{i-\text{BULK}} = 2 (E_F - E_{i-\text{BULK}}) \quad (3.48)$$

- **Strong Inversion** occurs for higher magnitudes of $V_G < 0$. In this case the hole concentration increases near the surface.

In the case of an P-type semiconductor we have the same phenomena's with reserved V_G and charge values. Below a graphical summary of a P-type is illustrated.

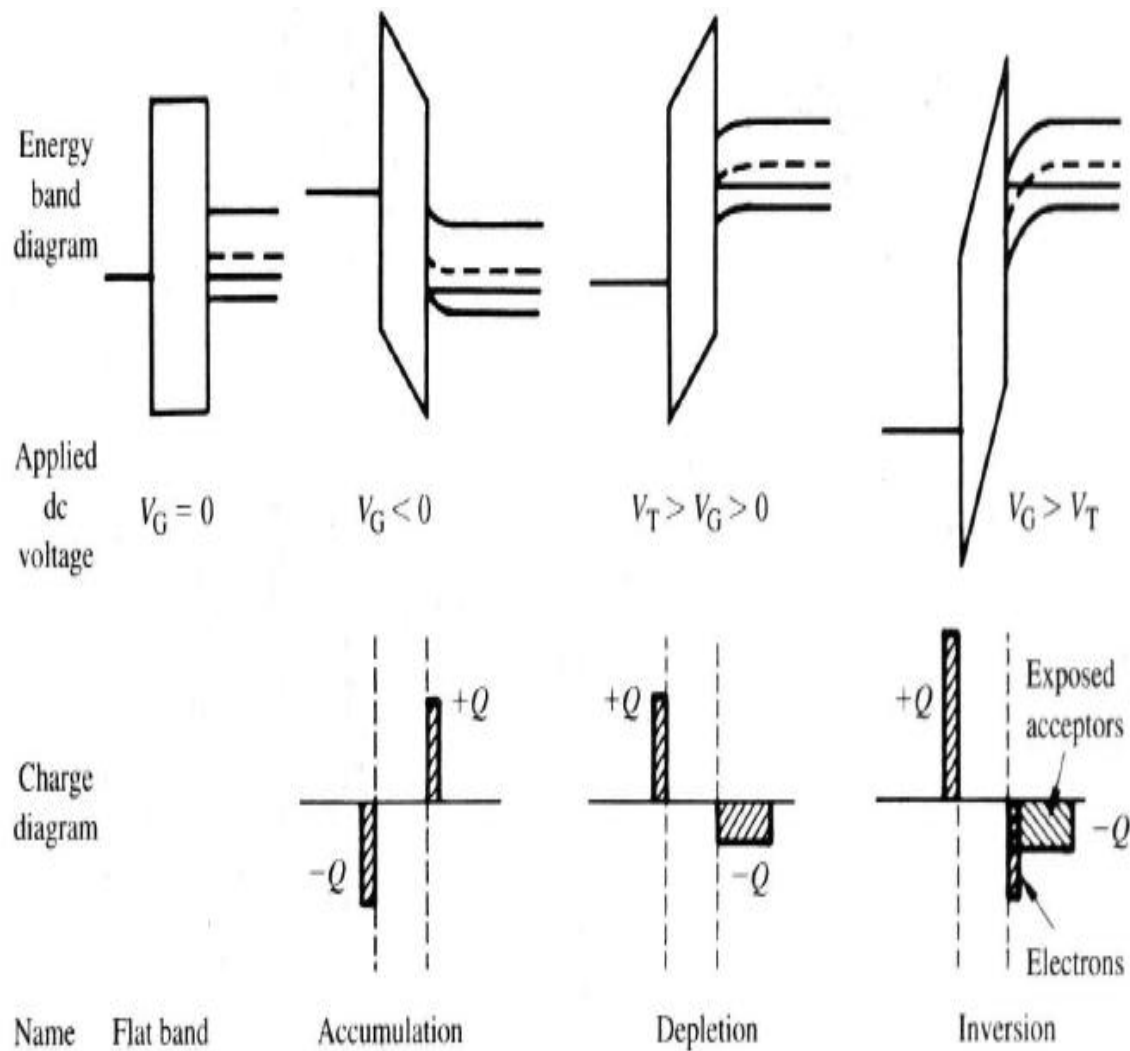
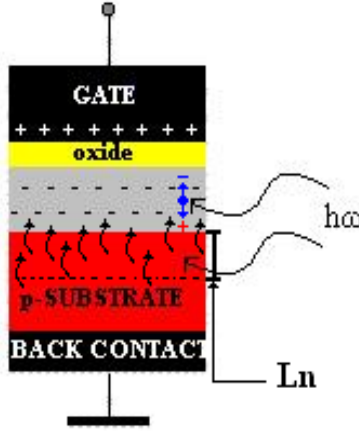


Figure 3.8. MOS basic operation functions.

A photo-gate, which is typically a MOS capacitor with a P-type semiconductor, is operating in depletion mode. This mode is desirable since we want to detect the photo generated carriers (electrons) in the depletion region. If the diffusion length is long enough to reach the depletion region then carriers generated in the P-type semiconductor contribute to the charge accumulation, see figure 3.8.



DEPLETION MODE

Figure 3.9. Electron-hole pair separation in MOS

Equations of (3.49) are depicted in figure 3.10. We have that the doped concentrations in the semiconductors are given by the equations:

$$P_{Bulk} = n_i e^{\frac{(E_{i-Bulk} - E_F)}{kT}} = N_A, \quad n_{Bulk} = n_i e^{\frac{(E_F - E_{i-Bulk})}{kT}} = N_D$$

Hence taking the logarithm, and solving for the energy difference between the Fermi and the interface potential we get

$$\phi_F = \begin{cases} \frac{kT}{q} \ln \left(\frac{N_A}{n_i} \right) & p\text{-type} \\ -\frac{kT}{q} \ln \left(\frac{N_D}{n_i} \right) & n\text{-type} \end{cases} \quad (3.72)$$

To calculate the carriers we must proceed to an analysis which will show the potential differences in the semiconductor surfaces. Let's define $\phi(x)$ the electrostatic potential inside the semiconductor at a certain depth x . Be aware that $\phi(x)$ is measured from the oxide interface. We define

$$\phi(x) = \frac{1}{q} [E_{i-BULK} - E_i(x)] = \text{electrostatic potential}$$

$$\phi_S = \frac{1}{q} [E_{i-BULK} - E_{i-Interface}(x)] = \text{surface potential}$$

$$\phi_F = \frac{1}{q} [E_{i-BULK} - E_F(x)]$$

(3.49)

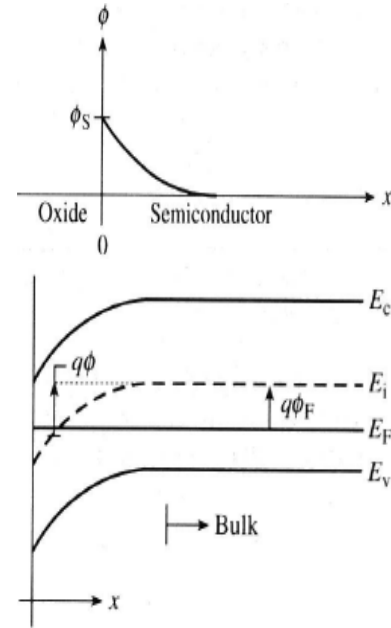


Figure 3.10 MOS surface potential with difference to Fermi potential.

The above definitions allow us to define the depletion width W in depletion mode of the MOS capacitor.

3.5.1 Depletion approximation for MOS

For calculating W , a depletion region approximation is necessary. The procedure is analogous for the pn junction in chapter two, but is repeated since it includes dealing with oxide, where charges are almost negligible. The approximation makes assumptions about the charges in a semiconductor: *As within the depletion region the semiconductor is depleted of mobile carriers, the charge density in that region is due to the ionized acceptors. Outside the depletion region, the semiconductor is assumed neutral (no*

charge!). Also the electric field is zero outside the depletion region, since a non-zero field would cause the mobile carriers to redistribute. So we define that the depletion region does not contain mobile carriers so that there can be an electric field. Density in metal is very high and therefore its potential is low, as a result the potential across the metal is left out. The steps involved are:

I. Define the metal-semiconductor interface at ($x = 0$) and the edge of the depletion region at ($x = W$).

II. Find the charge density in the semiconductor. Thus

$$\rho(x) = q(p - n + N_D - N_A) \cong -qN_A \quad \text{for } 0 < x < W, \quad \rho(x) = 0 \quad \text{for } x > W$$

III. Find the electric field in the semiconductor. The charge in the semiconductor is balanced by the charge in the metal, so only an electric field exists around the metal-semiconductor interface. To find the electric field as function of the position we use Gauss law

$$E(x) = -\frac{qN_A}{\epsilon_s}(W - x) \quad 0 < x < W, \quad E(x) = 0 \quad x > W$$

ϵ_s is the dielectric constant of the semiconductor

The maximum value of the electric field is located at the interface, therefore we

$$\text{set } x = 0 \text{ in the above equation, such that } E(x=0) = -\frac{qN_A W}{\epsilon_s} = -\frac{Q_A}{\epsilon_s}$$

Q_A is the total charge (per unit area)

IV. Express the potential across the semiconductor as a function of the depletion layer width. Since the component of electric field in any direction is the negative of rate of change of the potential in that direction $E_x = -d\phi/dx$ or $E_x = -dV/dx$, we can derive the potential simply by integrating the electric field, thus

$$\left\{ \begin{array}{ll} \phi(x) = 0 & x < 0 \\ \phi(x) = \frac{qN_A}{2\epsilon_s}(W^2 - (W - x)^2) & 0 < x < W \\ \phi(x) = \frac{qN_A W^2}{2\epsilon_s} & W \leq x \end{array} \right.$$

V. Express the width W through the applied voltage. When in thermal equilibrium a voltage is applied, the built-in potential, ϕ_i , equals the total potential difference across the semiconductor. One can verify this by setting

$$\phi(x = -\infty) - \phi(x = 0) = \phi_i - V_{\text{applied}}.$$

Therefore we can establish a relationship between the semiconductor potential at the surface, the applied voltage and the depletion layer width. Thus we can determine the depletion layer width by the

boundary condition: $\phi_i - V_{\text{applied}} = -\phi(x=0) = \frac{qN_A W^2}{2\epsilon_s}$. Solving for W we get

$$W = \sqrt{\frac{2\epsilon_s \phi_i - V_{\text{applied}}}{qN_A}} \quad (3.50)$$

The above approximation is generally applied on semiconductors. To find the depletion width in the MOS capacitor we follow the same procedures like above. Of course parameters are adopted to fit the approximation and besides of the approximations assumptions, the following guidelines for a MOS capacitor analysis are established:

1. Full depletion approximation holds
2. Generally to derive MOSFET models one must make the assumption that the inversion layer charge is proportional with the applied gate voltage and that in addition, for our capacitor model, the inversion layer charge is zero at and below the threshold voltage. This is described by
$$Q_{\text{inv}} = C_{\text{ox}}(V_G - V_T) \quad \text{in depletion region}$$
$$Q_{\text{inv}} = 0 \quad \text{outside depletion region}$$

The proportionality constant between the charge and the applied voltage is therefore expected to be the gate oxide capacitance C_{ox} . This assumption also implies that the inversion layer charge is located exactly at the oxide-semiconductor interface.

Hence for the MOS capacitor the depletion width approximation is

- I. W is defined in the same manner as in the normal approximation.

II. **Charge density:** $Q_W = -qN_A W$

III. **Electric field:** Since the electric field is the sum of the electric field in the semiconductor at the interface, E_s , and the field in the oxide, E_{ox} . We have

$$E_s = \frac{qN_A W}{\epsilon_s} \quad \text{and} \quad E_{\text{ox}} = \frac{qN_A W}{\epsilon_{\text{ox}}}$$

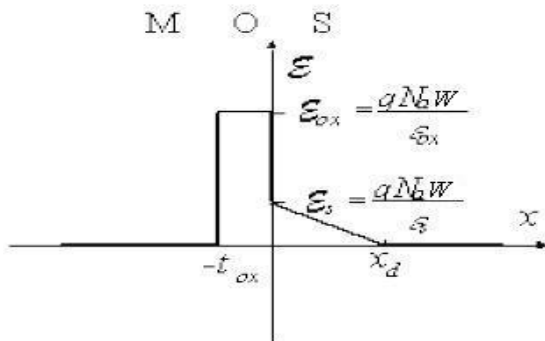


Figure 3.11 An example of electric field at the interface in a NMOS semiconductor. Note the abrupt change due to the change of the dielectric constant.

At the oxide-semiconductor interface the field in the oxide is larger than the field in the semiconductor, this is due to the permittivity of each compound for oxide its $\epsilon_{\text{ox}} = 3.9\epsilon_0$ and for the semiconductor it's $\epsilon_s = 11.9\epsilon_0$! This is the cause why the electric field changes abruptly at the interface. Further away from the interface and into the semiconductor the electric field changes linearly due to the constant doping density and outside the depletion region or at its edge the field is considered to be zero.

IV. **Potential:** Integrating the electric field yields the potential at the surface, with surface we mean at the top of the depletion region. $\phi_s = \frac{qN_A W^2}{2\epsilon_s}$.

V. **W through the applied voltage:** As we know in the accumulation mode there no forming of the depletion region, but additional charge (Q_{inv}) is present in the inverse layer when MOS capacitor is in inversion mode. Additional charge is been made when the electron density at the surface exceeds the hole density in the substrate, N_A , this is achieved when applying gate voltage which in turn gradually ads charge. The threshold voltage V_T therefore is the voltage for which the electron density at the surface equals N_A . The latter corresponds to the situation where the total potential across the surface equals twice the bulk potential ϕ_F . This in turn means that the depletion width is bounded to a potential of $0 < \phi_s \leq 2\phi_F$, then W is defined as

$$W = \sqrt{\frac{2\epsilon_s \phi_s}{qN_A}} \quad (3.51)$$

3.5.2 Photogate (PG)

From here on we derive the potentials characteristics of the MOS capacitor, keep in mind that PG and MOS capacitor are the same. Finally with (3.51) we are able to derive the potential of the gate Voltage V_G . We have an electric Field and charges across the insulator (oxide), hence a potential V_{ox} is created. If we see this phenomenon from another point of view, we can describe the system gate-insulator-surface potential as a capacitor. Remember a capacitor is described as $C = \frac{Q}{V_c}$. Here Q is the charge density

$Q_w = -qN_A W$ and C is the capacitance of the insulator, namely C_{ox} . Now V_G is partially distributed as V_{ox} and V_s . Thus

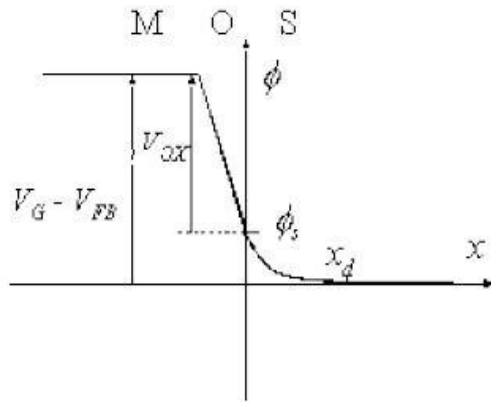


Figure 3.12. Potential in a N-MOS Semiconductor.

$$V_{ox} = \frac{qN_A W}{C_{ox}} \text{ and } \boxed{V_G = V_{ox} + V_s} \quad (3.52)$$

(V_s is Φ_s) holds. The right part of (3.52) is already calculated and the last step is to include the charge that will be generated due to impinging light. This charge contributes to the accumulation of charge near the surface, hence the surface potential increases.

$$V_G = \frac{\sqrt{2qN_A \epsilon_{si} \psi_s} + |Q_{sig}|}{C_{ox}} + \phi_s \quad (3.52)$$

Next we derive the quantum efficiency of the the MOS capacitor (PG). The procedure is analogous to the photodiode, except for the fact that no diffusion region exists above the junction. Therefore the diffusion current J_{diff} is only due to electrons in

the P-type substrate and drift current is due to optical generation rate. Electrons are minority in P-type, hence like in the p-n junction we calculate the concentration at the boundary of the depletion region:

$$n(W_D) = n_{p0} e^{\frac{-\psi_s}{U_T}}$$

Making the same calculations like with the p-n junction we derive the quantum efficiency for MOS capacitor:

$$\eta = (1 - R) \left(1 - \frac{1}{1 + \alpha L_n} e^{-\alpha W} \right) \quad (3.53)$$

The result is almost identical with the PIN diode, difference is, that the electron diffusion length is the major contributor of the quantum efficiency.

Important noise sources of the MOS capacitor are the dark current, shot noise and transfer noise. Dark current are the minority carriers that due to temperature variations, slowly add up to the potential well. Shot noise comes from two sources, one of the random arrivals of Photons and the other from the variations in dark current. Luckily shot noise from Photons as well as temperature variations can be characterized by a Poisson random process. Finally we take a look at the charge collection when light hits the P-type substrate MOS capacitor (in depletion mode): During a time interval, called integration time t_{int} , an optical power P_{ph} is responsible for an excess collection of charge (electrons). This excess charge constitutes the photocurrent I_{ph} . Hence

$$|Q_{sig}| = I_{ph} t_{int} = \frac{q \eta P_{ph}}{\hbar \omega} t_{int} \quad (3.54)$$

The stored charge $|Q_{sig}|$ must be, with one or the other way, transported to other components, so that a measure of light can be done. Below we will see that this comes with a cost of transporting noise sources also to the measuring.

Noise calculations is done considering that dark current and the photocurrent have uniform power spectral density of $2q(I_{ph} + I_D) A^2/Hz$. The rms current is easily calculated as:

$$\sqrt{\tilde{i}_n^2} = \sqrt{2q \left(\frac{I_{ph} + I_D}{2t_{int}} \right)} = \sqrt{q \left(\frac{I_{ph} + I_D}{t_{int}} \right)} \quad (3.55)$$

where $1/(2t_{int})$ is the bandwidth with sampling rate $1/t_{int}$. Because of $\sqrt{\tilde{i}_n^2}$ there some charge fluctuation at the end of the integration time and is given by:

$$Qn = \sqrt{\frac{q(I_{ph} + I_D)}{t_{int}}} t_{int} = \sqrt{q(I_{ph} + I_D)} \quad (3.56)$$

The SNR expressed with the associated charges in the MOS is

$$SNR = \left(\frac{Q_{sig}}{Qn} \right)^2 = \frac{I_{ph}^2 I_D^2}{q(I_{ph} + I_D) t_{int}} = \frac{q t_{int} \left(\frac{\eta P_{ph}}{\hbar \omega} \right)^2}{\frac{\eta P_{ph}}{\hbar \omega} + I_D} \quad (3.57)$$

Again the minimum power required to achieve an arbitrary SNR

$$P_{ph}^{min} = \frac{\hbar \omega (SNR)}{2 \eta t_{int}} \left\{ 1 + \left[1 + \frac{4 I_D t_{int}}{q (SNR)} \right]^{-\frac{1}{2}} \right\} \quad (3.58)$$

We know that NEP is defined with SNR=1 hence:

$$NEP = \frac{\hbar \omega}{2 \eta t_{int}} \left\{ 1 + \left[1 + \frac{4 I_D t_{int}}{q} \right]^{-\frac{1}{2}} \right\} \quad (3.59)$$

We mentioned above the problem of transporting the charge to other electric medium or devices. To get the charge out of the MOS capacitor we must channel it through it, hence the conductance of the MOS capacitor gets involved. Each time a transfer occurs we add thermal noise due to conductance to the output. Since a channel to output exist the conductance thermal noise is integrated over the bandwidth, resulting in a conductance-independent kT/C noise. Lastly one can see that combining the above noise with the noise due to charge fluctuations, that the overall output noise increases by the rms value of $\sqrt{kTC_{EQ}}$ for each transfer.

3.6 References

- [1] Pecht, Orly Yardid and Cummings, Ralph Etienne. *CMOS imagers: From phototransuction to image processing*. Dortrecht (NL) : Kluwer academic publisher, 2004.
- [2] Wood, David. *Optoelectronics Semiconductor Devices*. s.l. : Prentice Hall International Series in Optoelectronics, 1994.
- [3] Chuang, Shun Lien. *Physics of optoelectronics devices*. s.l. : A Wiley-Interscience Publication John Wiley & Sons Inc., 1995.
- [4] Klingshirm, C F. *Semiconductor Optics*. s.l. : Springer, 1997.
- [5] Wilson, John and Hawkes, john. *Optoelectronics and introduction*. s.l. : Prentice Hall, 1998.
- [6] Cheo, P K. *Fiber Optics and Optoelectronics, 2nd. Ed.* s.l. : Prentice-Hall International, 1990.
- [7] Van Zeghbroeck, Bart V., *Principles of Semiconductor Devices and Heterojunctions*. s.l. : Prentice Hall, 2008. 0130409049 .

Chapter 4. Pixel and array implementations

Clearly one can see that the best choices for light sensing imagers depend on the light detecting devices, which were analyzed in the third chapter. There, the best devices for sensing were those with good noise suppression, satisfying speed and best light conversation, just to name a few criteria's. By comparing only the three mentioned criteria's one can see easily that the photodiode and the photogate qualify for the light detecting devices. In fact, the photogate is being used since the 1970 as sensing devices in an array of photogates, called CCD. Photodiodes instead, are been used just the last two decades with Complementary Metal-Oxide Semiconductor as the CCD's counterpart.

4.1 Photogate based pixel architectures (CCD)

Like mentioned above charge-coupled devices (CCD) is an array of photogates, which share the same substrate. A difference is the gate which consists of three overlapping gates is one pixel. We will explain this overlapping mode a little bit later on. The principle of the light interaction with a photogate was explained in chapter three. Here we will explain the transfer and the characteristics that the light sensing device with photogates will have. Starting from the charge transfer we assume a more ideal model, namely a not overlapping gate, but a closely spaced gate model. This situation is shown in figure (4.1 (c))

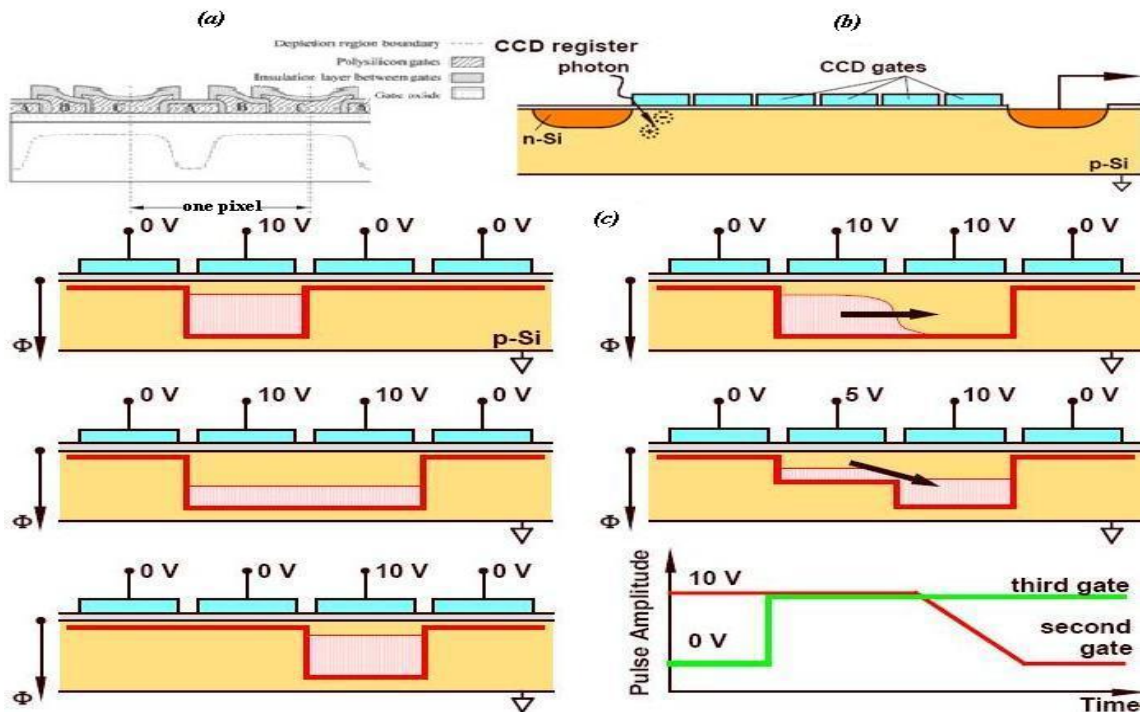


Figure 4.1 (a) Overlapping gates (b) Photogate as CCD (c) CCD principle of transferring the charge. Figures from [1] and [10].

Assume the second gate is biased at 10 V (4.1 (c) top left). This has the effect that under the gate charges can be accumulated; see the potential well formed right under it. The

charge cannot be moved elsewhere if not another gate creates a second potential well. This situation corresponds to the third gate raised voltage and its value is 10 V. The result is the charge distribution over the new enlarged potential well (4.1 (c) top right). If the second gate decreases its voltage to 0 V, then as it makes the transition, the charge distribute like the water flow in a waterfall (middle-right). This is logical since the potential well of the second photogate shrinks. At the end of the transition all the charge is transferred into the potential well of the third gate. This procedure transfers continues until the charge is transferred out of the sensing array, where it is observed by voltage measuring. At the right of the bottom in the figure (4.1 (c)) we see the signal, or also called pulses, necessary to make the transfer just described above. The problem is that when each photogate is generating electrons, due to photon induced electron-hole pair, then the time to transfer each charge of the photogates becomes large, and hence we get a low frame rate. A more sophisticated method to increase the speed is to set some of the CCD gates only as storing or shifting operators of charge. To do this some modifications has to be done to the normal structure of the CCD gate as shown in the figure below.

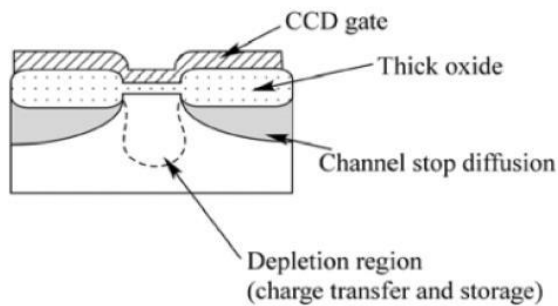


Figure 4.2 (a) Modified CCD to keep charge concealed and ready to shift .Figure from [10]

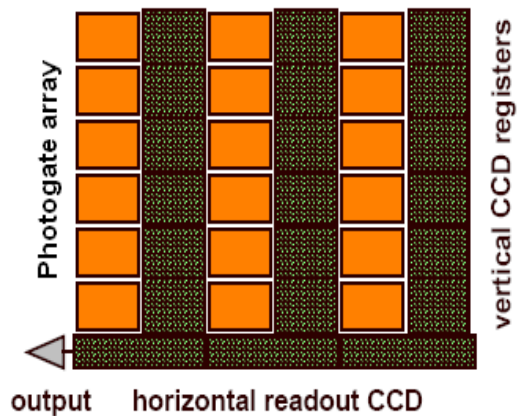


Figure 4.3 Inline CCD structure. Figure from [1]

camcorders. One of its shortcomings is the time the VCCD and HCCD lines need to empty the charges, an important role in this matter plays the length of those lines. Hence the exposure time depends on the discharge time of the lines. Another shortcoming of this architecture is the small fill factor. The latter is defined as the ratio of the device's dedicated sensing area to the actual sensing area. One can see that half of area of the inline CCD consists of non-sensing gates.

Two basic concepts exists which take advantage of the modified gate, the inline transfer CCD and the frame transfer CDD structure.

4.1.1 Inline transfer CCD

The structure is an array of inline CCD (figure 5.1 (c)) next to each other. These form vertical lines and between each, a vertical shift registers (VCCD's) is been put. The VCCD lines connect to a horizontal line (HCCD) from which the charge is readout. During the exposure time the charge is accumulated. After this all photogates are emptied and simultaneously the associated charged are transferred to the VCCD's registers. In the coming exposure time the charge in the VCCD's are transferred to the HCCD line and eventually transferred to the output. This CCD structure can be implemented on a very small chip size, an example are

4.1.2 Frame transfer CCD

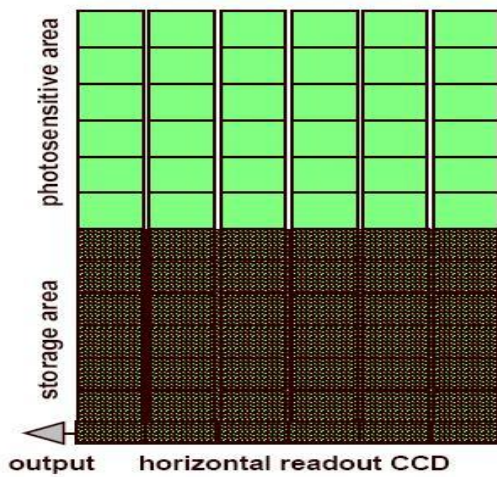


Figure 4.4 Frame transfer structure. Figure from [1]

In this structure the sensing area is divided into two parts: A light sensing area and a buffer like area. Both areas are equal in size. The photogates in the light sensing area are replaced with photodiodes and hence no vertical line is used to transfer the charge collected. The buffer area on the other hand is still photogate made, and therefore holds any charge collected during exposure from the photodiodes. One might think that additional or extra charges can be produced if the buffer area is exposed to light. That is the reason why the buffer area is covered with a thin layer of metal, reflecting all impinging light.

Metallic materials like aluminum for example are known for their high reflectance. After the exposure time the charge is stored in the buffer area and at the next exposure time the charge is shifted towards the horizontal line, where it can be transferred to the output. The main advantage of this architecture is its big fill factor, large chip size implementation and the compatibility with standard technology. Due to repeated transfer, charges might get lost or trapped and blurring (smear) effects might appear.

4.1.3 Full frame CCD

The principle of this structure is the same as in frame transfer, with the difference of the buffer area. The generated electrons are directly transferred during a read out cycle, or also called read out time. In that time the array cannot be exposed to light or do anything else than transferring charge. To prevent any such things from taking place a mechanical shutter is needed during the read out. This architecture has been used in still frame imaging.

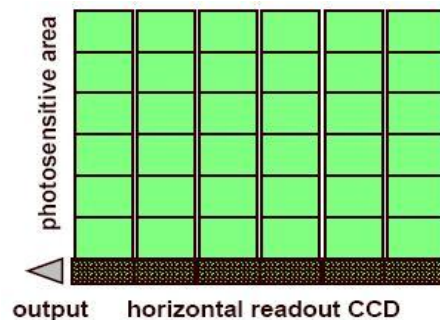


Figure 4.5 Frame transfer structure. Figure from [1].

4.1.4 Properties of CCD

This section deals with the performance in a CCD. Let's review the process which almost all architectures go through: First we have charge generation by the impinging photon, second the charge is collected by the nearest potential well and lastly the charge "bucket" is transferred through the array. In parallel one might look at this process and recon that for each state in the process a similar defined performance must hold. This is somewhat true; the performance depends on

4.1.4.1 Quantum efficiency

Quantum efficiency η is the numbers of charges generated per incident photon. We have this subject discussed in the second chapter.

4.1.4.2 Charge collection efficiency

The position at which the charges are created is very important. For instance if the charge is created near the potential well, mainly near the surface, experience a strong force of the electric field and therefore are collected to the potential well. In contrast a charge created deep in the substrate experience a weak force of the electric field and one possibility is that it diffuses into a surrounding photogate (pixel). The charge diffusion into another pixel is also called “split event”. Another possibility is that the charge may be trapped or may recombine. Trapping and recombining under this circumstance is called “partial event”. Both events are the cause of the blurring or smearing of the image, a very undesirable effect in precision devices such as in a medical imager. As a result the metric to the above events is the charge collection efficiency η_{CC} .

4.1.4.3 Charge transfer efficiency

Charge transfer efficiency η_{CT} is defined as the ratio of the charge transferred to the initial charge stored. In other words its charge left when it's transported from one gate (initial charge Q_0) to the other gate (neighbor gate). The remaining charge, assumed to have been transported n times in the registers, can be expressed mathematically as

$$Q_n = Q_0 \eta_{CT}^n \quad (4.1)$$

Charge transfer exists due of three phenomena's the (thermal) diffusion, self-induced drift and fringing field drift. If the charge amount is small diffusion is the main cause of transfer. Hence the storage of the gate decreases exponentially with time equal to the diffusion time constant τ_{th} .

$$\tau_{th} = \frac{4L^2}{\pi^2 D_n} \quad (4.2)$$

where L is the space between the centers of two gates. For large amount of charges the dominate factor self-induced drift is the main contributor for charge transfer. The fringing field drift is an independent factor since it is not influenced by charge intensity, for that reason once the charge is transferred; they are being influenced by this factor.

4.1.4.4 Noise

Noise always corrupt the desired signal, in this case transfer efficiency has to be as good as possible, to obtain a minimized distorted output. The factors which are responsible of corrupting the transfer efficiency are

- the dark current
- the transport speed
- the interface traps

The origin of dark current is the thermal generation of carriers in the depletion region, the diffusion and the recombination current. If the photogates voltage is set to depletion mode, immediately after the forming of the depletion region, some thermally generated minority carriers are produced and these fill the slowly the potential well. These carriers lower the ability of fully charge the potential well and therefore spoil, in a minor but important degree, the transferring of charge. The corruption can be minimized if the clock frequency which are pulsing the voltages of the gates is kept high. Contrary a too high frequency could also have negative effects on the transfer. The charges could be trapped in an intermediate state near the interface (a defect in the lattice called dangling bonds) and later on been released.

The choice of frequency plays a role in the trapping mechanism of charge, at high frequencies the charge could be trapped easily and because of high switching it is most likely that it is not released at the same time as the charges are being transferred. It can also happen that the trapped charge is released when another charge is stored in its area. Therefore it contributes in the wrong way to the charge. As a result the overall process of transfer is quite sensitive to from losing a charge, till to have more charge then actually sensed. Summing it all together we see that due to trapping there can be variations among pixel; this is true especially at high frequencies. The “trapping noise” is a major contributor to the fixed-pattern noise (FPN). FPN is defines the various noise sources that may arise from pixel variations, studies on that subject is still be done. A noise called shot noise influences at midrange frequencies, shot noise occurs when the finite number of particles that carry energy in an optical device like the photogate, is small enough to give rise to detectable statistical fluctuations in a measurement. The magnitude of this noise depends on the photocurrents average value in time. When the average current changes quickly shot noise becomes more apparent. At low frequencies dark current is more dominant than any other noise source.

4.1.4.5 Response speed

In addition due to the charge transfer problem the transfer time and hence the respond speed is reduced. The result is seen in the image produced by the CCD, it can contain blurring, a kind of combination of a delayed or and distorted image. The reduction of trapped charges can be reduced by two ways: One being by maintaining a background charge (fat zero), meaning to have the traps filled at all times, and the other one being to modify again the photogate structure. This new technology is called buried channel CCD (BCCD). In this technology the transfer efficiency (respond time) is somewhat smaller than the original CCD.

4.2 Photodiode based pixel architectures (CMOS)

Architectures with photodiodes in the pixel design could also be used with photogates as we see saw in the CCD section. The reason why we separate the pixel architectures according to the sensing devices is simple due their common use with these architectures. One could easily switch the sensing devices and adapt the pixel architecture to them. For now let's stick to our concept of pixel architectures. In photodiodes architecture we further divide it into active pixels (APS) and into passive pixels (PPS). The latter consist

of a photodiode and a transistor as a switch to couple the produced charges to the output line. Two conventions are made the output line is called a the bit or column line and the transistor the row or select transistor, these name conventions become more apparent when we view the hole pixel area. Even though PPS have a big fill factor the shortcomings of it are significant. Some of these are the low sensitivity and the high noise due to the capacitance of the bit line.

4.2.1 Active pixel (APS)

Active pixel (APS) is a pixel design in which each pixel contains at least one active transistor. The basic idea is to buffer the light injected charge with a transistor also called source follower, and at the right timing release the charge. The transistor also amplifies the signal coming from the photodiode and is activated only during readout, which means the power dissipation is less than CCD's. . In the process of a transistor the non-uniformity of a wafer causes threshold voltage variations amongst transistors. This is a major problem with CMOS APS, because it creates, almost naturally, a high FPN noise. This noise can be canceled out with suitably chosen circuits that subtract voltage differences. A popular circuit for this purpose is the correlated double sampling circuit (DDS). Comparing CDD against CMOS yields the following points:

Pro's and Con's	
CCD	CMOS
<ul style="list-style-type: none"> • Low noise • Smaller pixel size • Low dark current • 100% Fill factor • Higher sensitivity • Shutter • High weight • Higher cost • Higher power consumption 	<ul style="list-style-type: none"> • Lower power consumption • Single power supply • High integration capability • Low fill factor • Lower cost • Single master clock • Random access • Low weight • Compact size

Table 4.1

The increased demand for more integration due to rapidly advance in technology, the CMOS APS were favored over other implementations. This “boom” happened in the 90's of the last century, and since then a variety of APS appeared. A glimpse into the spectrum of APS pixel design follows below:

4.2.1.1 (PD) Photodiode type APS

First described by Noble in 1968 it consists of three transistor, each responsible for a different task. The tasks are: resetting the photodiode's voltage via a reset transistor, buffering the photodiode voltage onto a vertical-column bus (column line) via a source follower. This type of sensing device is simple and yet gets more sophisticated with scaling. Hence, PD APS is satisfactory choice, and depending on the application and the process technology in use it can achieve mid low or high-performance. A very versatile APS.

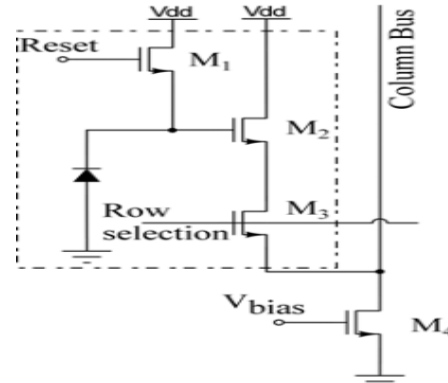


Figure 4.6 PD APS. Figure from [3]

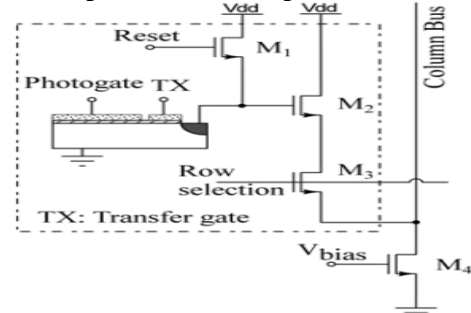


Figure 4.7 PG APS. Figure from [3]

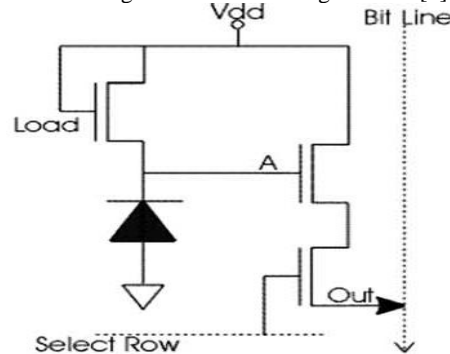


Figure 4.8 logarithmic APS. Figure from [3]

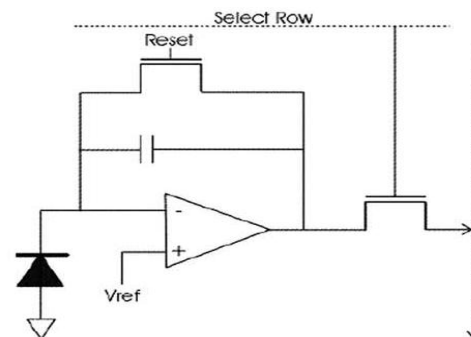


Figure 4.9 CTIA APS. Figure from [3]

4.2.1.2 (PG) Photogate type APS

Like mentioned before the switching of sensing devices, here the exchange of the PD with the photogate (PG), brings the APS pixel architecture closer to the CCD architecture. Therefore integration transport and readout properties are the same as in CCD.

4.2.1.3 Logarithmic APS

In some applications a desirable output is a non-linear one, this will allow the increase on the intrascene dynamic range. It must be pointed out that they suffer from large FPN, high temperature dependence and low output swing. For this reason they are not been used.

4.2.1.4 CTIA APS pixels

Another method of suppressing FPN is used in this architecture. Here a transimpedance amplifier (CTIA) is used directly coupled to the photodiode and the transistors are used as operating "handles".

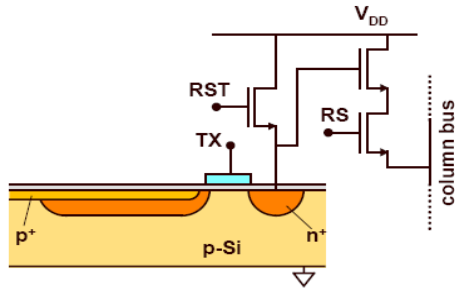


Figure 4.10 Logarithmic APS. Figure from [1]

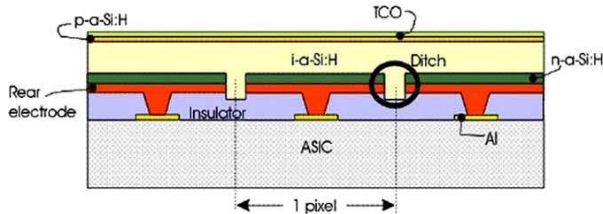


Figure 4.11 TFA APS. Figure from [3]

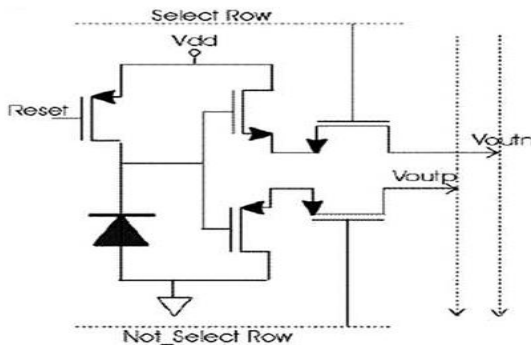


Figure 4.12 CAPS. Figure from [3]

4.2.1.5 (PPD) Pinned photodiode pixel

They are more sensitive and have lower dark current than the PG photopixels. The transfer gate TX has the same purpose like in the photogate; by pulsing it, it transfers the charge packet to the floating diffusion region n^+ .

4.2.1.6 TFA pixels

Another implementation that offers more sensitivity with the help of a material that has much higher absorption coefficient than silicon. This material is deposited on the ASIC as a thin layer. TFA are used in high demands in applications with high dynamic range.

4.2.1.7 (CAPS) Complementary active pixels

This architecture achieves very low power consumption and a high low-voltage operation capability by operating in complementary way, note the replacement of the reset transistor with a PMOS. Like the PD APS, they can be high integrated.

4.2.2 APS readout methods

An important role as to how to read out the produced charge is played by the circuits connected to the sensing device. These circuits are called readout circuits and are characterized by the information rate they can achieve. The information rate is the rate at which the APS pixel transduce optical signal into electrical signal. This is done in two fashions. The first method would be to integrate the photocurrent onto a capacitor during a sampling period, and the second method is to transduce the photocurrent waveform continuously into an output current or into a voltage signal. For that reason the readout circuits decide the usefulness, measured by the pixel information rate, of a certain pixel APS readout architecture.

4.2.2.1 Charge-mode pixels

All architectures in which the photocurrent is used in form of charge is called charge-mode pixel. Three phases governs a charge-mode pixel, which are executed like seen in the order below:

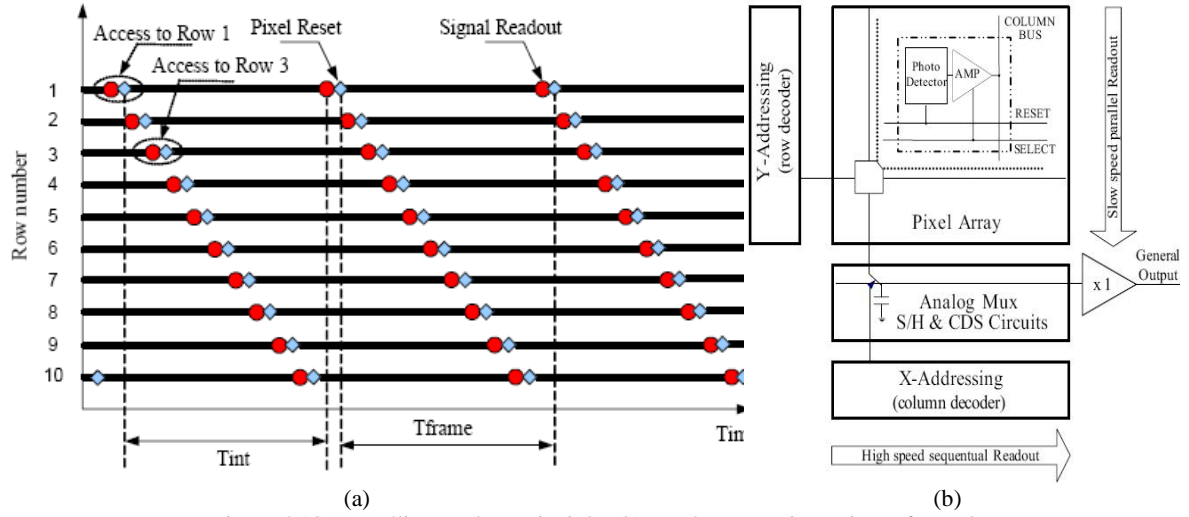
- Reset phase
- Integration phase
- Readout phase

Like mentioned the PD APS is a traditional three cell transistor with a photodiode, for that reason it's also called a 3T cell pixel. The capacitances connected to the sensing node (node between reset transistor and photodiode) are the photodiodes capacitance, the gate capacitance of the source follower and the source capacitance of the reset transistor. The last two capacitances are called parasitic capacitances, because they are not from the same device but do contribute negatively to the device operating purpose. The source follower isolates the sensing node from the column line capacitance and therefore allows the photodiode to accumulate charge free from external distortions. As a result the source follower reduces the pixel to pixel variations, which would contribute to fixed pattern noise (FPN).

In the reset phase the reset transistor is switched on and the sensing node is set to a high value. This value is then transferred to the column line and finally ends up in the CDS circuit, located at the bottom of each column line.

During the integration phase the reset transistor is off. Also during this phase the photocurrent produced due to photon collection, discharges the detection node. Remember that the forward bias current has direction from the positive to the negative pole of the photodiode whereas the reserve bias current (and the photocurrent) flows in the opposite direction of the forward bias current.

In the readout phase after the integration time, the new value of the photodiodes capacitance is send to the CDS circuit where it is been subtracted from the previously stored reset value. The pixels cannot be read all at once and therefore a method, called rolling readout, is used. This method is used in image sensors that have large array sizes to ensure that all pixels are exposed to the image for the same amount of time. Reading a pixel involves resetting the row that the pixel is in and then once the integration time has elapsed accessing the voltage that is left in the pixel. With multiple rows, each row needs to be read out sequentially causing the last row to be read out at sometime later than the first. The resetting of the rows, therefore, also needs to happen sequentially to maintain a common integration time for every row. The resetting of the row is therefore equivalent to opening the shutter of a normal camera and the reading of the row is the equivalent of closing it again. Because both signals are sequentially scrolling through the rows of the array this method is called rolling readout, see figure 5.13



To study only the procedure of charge transfer to the readout circuit we must make a simplification of the PD APS structure. The new structure consists only of the photodiode and a reset transistor. It is actually the PD APS version (Figure 4.6) with only the reset transistor. To get an estimation of the charge-mode pixel we calculate the input-referred noise for each phase. Input-referred noise is used to model the circuit's output noise. This may sound strange but in the real world output noise may affect the input and because we cannot measure the input noise we refer the output noise to the input. Hence the name: input-referred noise. All input-referred noise for each noise are then summed together to form the overall input-referred noise. Let's start with the reset phase which is much smaller the settling time of the photodiode. This observation leads us to calculate the noise using temporal analysis, described in [4] and reported by [10]. The results are taken from [10]. The mean-square noise voltage at the end of the reset period is

$$\overline{V_n^2(t_r)} = \frac{1}{2} \frac{kT}{C_{out}} \left(1 - \frac{t_{th}}{(t_r - t_1 - t_{th})^2} \right) \quad (4.3)$$

where t_r is the reset time, t_1 is the time that the reset transistor operates in the above threshold region, and t_{th} is the time required to charge the detecting node capacitance C_{out} till the thermal voltage of the reset transistor. If the reset time is in order of microseconds and for that reason about three times larger then t_r and t_{th} , which is usually the case, then (4.3) can be written as

$$\overline{V_n^2(t_r)} = \frac{1}{2} \frac{kT}{C_{out}} \quad (4.4)$$

In the integration time the reset transistor is turned off, therefore since the photodiode is working in the mid range, the only contributor is the shot noise. With constant capacitance we get at the end of the phase the following result

$$\overline{V_n^2(t_{int})} = \frac{q(I_{ph} + I_B)}{C_{out}^2} t_{int} \quad (4.5)$$

where t_{int} is the integration time and I_B is the background current. During the readout time only noise from the readout circuits is present, therefore for the pixel only the first to stages contribute. To sum the two mean square voltages we rely on statistics on information. In statistics on information a information submitted but divided via different “channels” is called mutual information $I(X;Y)$. In other words mutual information is a measure of dependence between the signals X and Y . Mutual information is difficult to be expressed mathematically but can be related to the average uncertainty when joint mutual information $p(x,y)$ and marginal distributions $p(x)$ and $p(y)$ are given. If that is the case we get

$$I(X;Y) = \sum_{x,y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)} \quad (4.6)$$

Having established 5.6 we can express the rate of information R at the output of the system as $I(X;Y)$ per sample times the sampling rate f_s :

$$R = f_s I(X;Y) \quad (4.7)$$

Of course a restriction exists on how much information can be transferred in a channel. This restriction measure is the capacity of the channel. Most channels can be modeled as channels in which the noise can be added and has Gaussian distribution. If S is the input signal power, N the noise power, and Δf is the bandwidth of the channel then

$$C = \Delta f \log_2 \left[1 + \frac{S}{N} \right] \frac{\text{bits}}{\text{second}} \quad (4.8)$$

Using (4.6) on (4.5) and (4.4) we have that for the charge mode pixel

$$I_{qm} = \frac{1}{2t_{int}} \log_2 \frac{\frac{\sigma_s^2 I_{ph}^2}{C_{out}^2} t_{int}^2}{\left(\frac{kT}{2C_{out}} + \frac{q(I_{ph} + I_B)}{C_{out}^2} t_{int} \right)} \quad (4.9)$$

$$I_{qm} = \Delta f \log_2 \left[1 + \frac{\sigma_s^2 I_{ph}^2}{(2kTC_{out}\Delta f + 2q(I_{ph} + I_B))\Delta f} \right] \quad (4.10)$$

The bandwidth of the pixel is $1/(2t_{int})$, denoted as Δf , with sampling rate $1/(t_{int})$. We already mentioned that the transit time t_r is shorter than the integration time t_{int} , henceforth (4.9) is a function of t_{int} , I_{ph} , C_{out} , T and I_B .

As measurement showed the information rate of the charge-mode pixel is independent of the bandwidth Δf . A fact that also can be seen in equation 4.9, I_{qm} has not its maximum when t_{int} is infinitive. At the first glance one might think long integration time gives the most qualitative image results. In charge-mode the best information rate and therefore the best image result is obtained at finite integration time t_{int} . In figure 4.14 average values are used. These values are

- photocurrent of $I_{ph}=100$ fA
- background current $I_B=2$ fA
- $C_{out}=10$ fF, $T=300$ K
- contrast power $\sigma_s^2=0.1$

With the help of these average values an optimum integration time for maximal information rate time can be found. It's around $0.5 \cdot 10^{-4}$ second. Of course with scaling of the photodiode one must change the values according to the technology process.

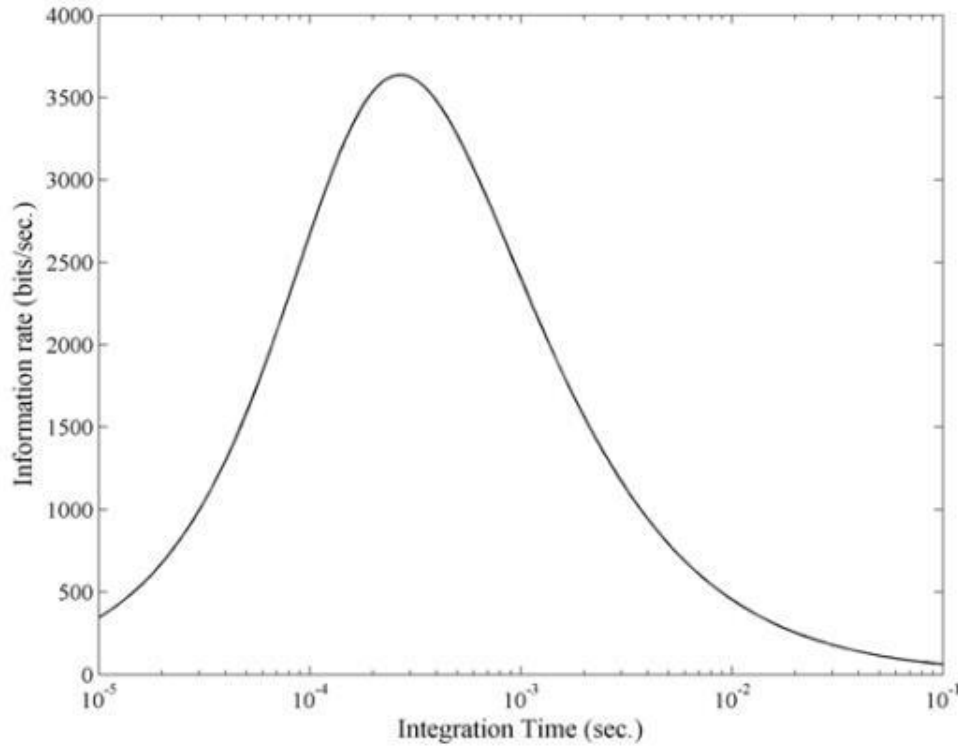


Figure 4.14 In Charge-mode pixel maximum information rate is achieved at finite integration time. Figure from [10].

4.2.2.2 Dynamic range

Another important factor, besides of information rate, is the dynamic range. The dynamic range measures the illumination capability of a sensor to react to a range of light intensities. Pictures differ by application, but all should capture bright objects and distinguish them among other dark or bright objects. Mathematically, intrascene dynamic range is expressed as:

$$DR = 20 \log \frac{S}{N} \quad (4.11)$$

where S is the saturation level and N the root mean square (rms) read noise floor measure. APS have a dynamic range around 65-75dB. This isn't so bad compared to the human dynamic range which is around 90 dB and compared to the dynamic range of the camera film which is around 80dB. Still an increase in dynamic range is desirable; this can be done in two ways. The first way is to decrease noise; this in exchange increases the range towards darker scenes. The second way is to expand the saturation level of incident light and thus increase the range towards brighter scenes. Many methods were developed to deal with the darker and brighter scenes, the reader is been directed to [10].

4.2.2.3 Current-mode pixels

In this mode the photocurrent is directly converted into a current signal. Therefore the pixel consists solely of the photodiode. In addition a switch to control the output of the

photodiode, is added and is usually implemented with a transistor, see logarithmic pixel. Another addition to the bulk PD could be the connection of an amplifier, as seen in the CTIA pixel. Both implementations are based on continuous sampling. A simplified version of the PPS pixel to study the noise contribution and therefore its response capability is showed in figure 5.14 (c).

One part of noise consists of shot noise originating from the PD, and the other noise is from the active load circuitry (current load from the attached amplifier or other circuit). The variance of each source of noise (PD and circuitry) is $2q(I_{ph} + I_B)\Delta f$ and therefore the overall variance is

$$\bar{I}_n^2 = 4q(I_{ph} + I_B)\Delta f \quad (4.12)$$

Using once again (4.6) to find the information rate and considering $\Delta f = \frac{1}{2t_{int}}$, we have

$$I_{im} = \Delta f \log_2 \left[1 + \frac{\sigma_s^2 I_{ph}^2}{4q(I_{ph} + I_B)\Delta f} \right] \quad (4.13)$$

A maximum is reached at infinite bandwidth

$$\begin{aligned} I_{im} &= \lim_{\Delta f \rightarrow \infty} \Delta f \log_2 \left[1 + \frac{\sigma_s^2 I_{ph}^2}{4q(I_{ph} + I_B)\Delta f} \right] \Rightarrow \\ I_{im} &= \frac{\sigma_s^2 I_{ph}^2}{\frac{4q(I_{ph} + I_B)}{\ln 2}} \end{aligned} \quad (4.14)$$

4.2.2.4 Voltage-mode pixels

The photocurrent is converted using a linear resistor. The mean square voltage is given by

$$\overline{V_{sig}^2} = \sigma_s^2 I_{ph}^2 R^2 \quad (4.15)$$

The photodiode as shown in the current-mode, contributes a noise variance of $2q(I_{ph} + I_B)\Delta f$. The noise variance of a resistor is known to be $4kTR\Delta f$. (4.14) becomes then

$$\overline{V_{sig}^2} = (4kTR + 2q(I_{ph} + I_B)R^2)\Delta f \quad (4.16)$$

The information rate deduced with help of (4.16) and (4.6) is

$$I_{vm} = \Delta f \log_2 \left[1 + \frac{\sigma_s^2 I_{ph}^2 R^2}{(4kTR + 2q(I_{ph} + I_B)R^2)\Delta f} \right] \quad (4.17)$$

The maximum information rate becomes

$$\begin{aligned} I_{vm} &= \lim_{\Delta f \rightarrow \infty} \Delta f \log_2 \left[1 + \frac{\sigma_s^2 I_{ph}^2 R^2}{(R + 2q(I_{ph} + I_B)R^2)\Delta f} \right] \Rightarrow \\ I_{vm} &= \frac{1}{\ln 2} \frac{\sigma_s^2 I_{ph}^2}{\frac{4kT}{R} + 2q(I_{ph} + I_B)} \end{aligned} \quad (4.18)$$

4.2.2.5 Unified model of pixel information rate

Keeping the bandwidth and integration time fixed and assuming two cases: One being that the photocurrent I_{ph} is much lower than I_B and the other one being that the photocurrent I_{ph} is much higher than I_B , we can construct a unified model of information rate. Then from the three information rates (4.9), (4.14) and (4.17) we have

$$I = \begin{cases} \Delta f \log_2 \left[1 + \frac{\sigma_s^2 I_{ph}^2}{a \Delta f} \right], & I_{ph} \ll I_B \\ \Delta f \log_2 \left[1 + \frac{\sigma_s^2 I_{ph}}{b \Delta f} \right], & I_{ph} \gg I_B \end{cases} \quad (4.19)$$

where a and b take different values according to mode of the pixel:

Mode/constant	a	b
Charge-mode	$2KTC_{out}\Delta f + 2qI_B$	$2q$
Current-mode	$4qI_B$	$4q$
Voltage-mode	$4kT/R + 2qI_B$	$2q$

Table 4.2

The result is shown in the figure below, in which the information rate depends on the amount of photocurrent.

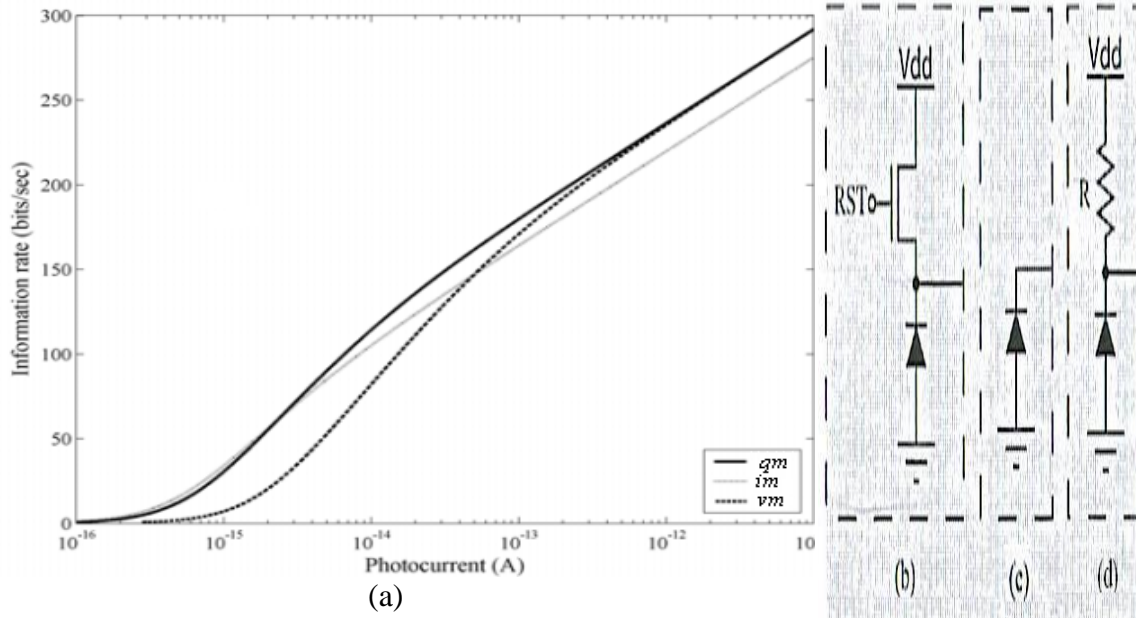


Figure 4.15 (a) Unified information rate for the three pixel modes. (b) Charge mode (c) Current-mode (d) Voltage mode

We can see immediately that the charge-mode pixel is the best choice for maximum information rate in the whole range of photocurrent.

4.2.3 Pixel systems

A two dimensional sensor has a similar structure to a (S)DRAM chip. As known, RAM's have to lines, one vertical and one horizontal. With these lines one can easily access a specific cell of choice and readout its contents, which is called random access.

4.2.3.1 PPS array system

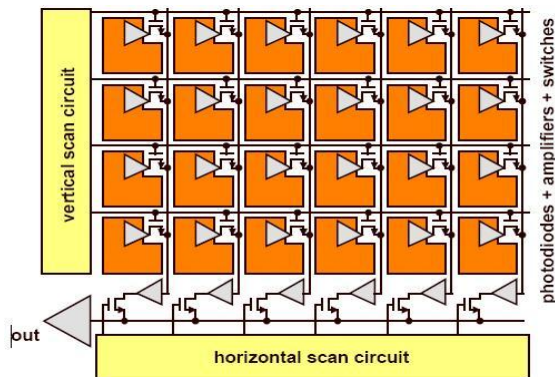


Figure 4.16 PPS array with column amplifiers. Figure from [1]

The first array consisted of bulk PPS's outlaid on an area. The PPP's are connected to the column line, to the select transistor (switch) and to the vertical and horizontal control circuits. As one can see from figure 4.16 the select transistor is used to address each PPS pixel. The PPS array is simple, but introduces more noise that it actually produces, and this weakens the photocurrent signal.

The reason is because during readout, there is a mismatch between the small capacitance of the photodiode and the large capacitances of the column and the vertical access line. An improved version with amplifier placed at the end of the column line, amplifies the photocurrent, but also amplifies the noise.

4.2.3.2 APS array systems

A better array system can be build with APS. To demonstrate why, we list the major differences between the APS and the PPS array:

	Pro's	Con's
APS	<ul style="list-style-type: none"> Fast readout time Pixel scalability Reduced capacitance yields lower read noise and High signal to noise ratio (SNR) 	<ul style="list-style-type: none"> Fill factor ranges between 50% and 70% Low quantum efficiency Higher FPN due to more transistors
PPS	<ul style="list-style-type: none"> Fill factor can reach 90% High quantum efficiency 	<ul style="list-style-type: none"> Slow readout time Scaling is difficult FPN can be very high, if amplifiers for each column are used. This is usually the case.

Table 4.3 Differences in pixel arrays

Scalability and fast readout time is a feature that permits the use of digital conversation on the same die. The conversation of the photodiodes analog signal is been done by an

analog to digital converter, also called for short ADC converter. Three different versions exist of where to place the ADC block, and analogous to the version, there are certain requirements that the system must fulfill.

4.2.3.2.1 Chip-level ADC

Only one ADC for the entire array is used, depicted in figure 4.17. The result is the limitation of the maximum pixel frequency, especially if high accuracy is desirable. For example for more than 12 bits, internal components of the ADC cannot mismatch, since this will result in inaccuracy issues. In parallel to accuracy the ADC must operate in high-speed, since all pixels must be converted during just a few clock rates (remember the short times for each phase in the charge-mode pixel).

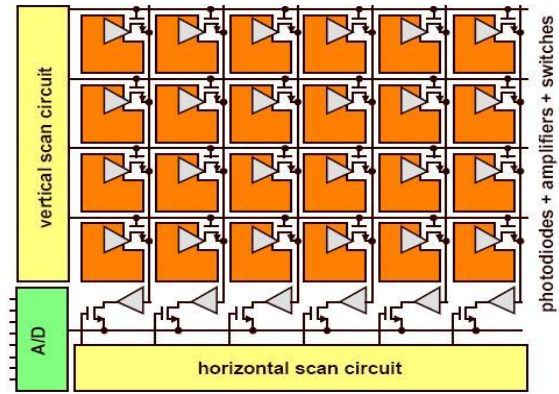


Figure 4.17 ADC at array. Figure from [1]

4.2.3.2.2 Column-level ADC

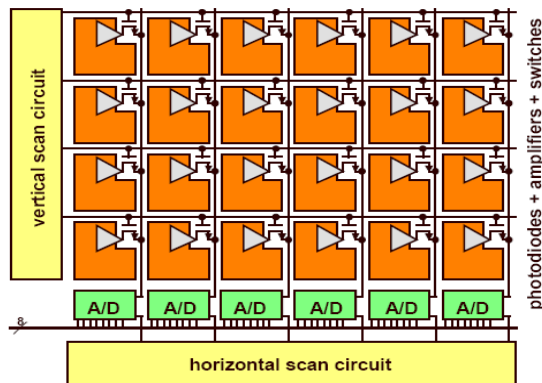


Figure 4.18 ADC at each column. Figure from [1]

An ADC is placed at each column, and each ADC is responsible for the conversion of the column line, therefore the speed requirements fall according to the existent number of column line in the array. For instance in figure 4.18, the speed for each ADC is a sixth of that of the ADC in the chip-level ADC implementation. In other words the frame rate x resolution is increased by 6. This implementation has some drawbacks:

1. Fitting an ADC in a column width of just a few microns is difficult. Even this feasible, most of the ADC functions are degraded for the fitting purpose. As a result mismatch can occur more often contributing to the FPN.
2. More ADC's means more power consumption
3. Area of the array can increase if the ADC need more space then the column widths
4. Due to point 2. for higher frequencies, a thermal managing system becomes necessary

4.2.3.2.3 Pixel-level ADC

Placing an ADC converter in the pixel turns the bus from an analog column line to a digital line one. This method allows parallel readout of the APS. The speed of the ADC is drastically reduced since it only converts the pixel value and the requirements on accuracy can be 8 or 12 bits. As a result, no correction circuits like CDS are needed and noise is been minimized. Hence SNR is very high. Furthermore the power consumption is decreased because the ADC only operates when the pixel is accessed.

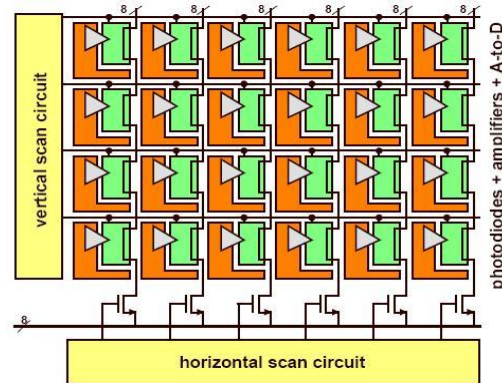


Figure 4.19 ADC in pixel. Figure from [1]

Of course there are some minor disadvantages to this implementation. The first would be that a satisfying ADC implementation must be found to actually fit in the pixel, without degrading the fill factor of the pixel. As a result the technology of the implementation must be small enough to allow pixel-level ADC. Most reported implementations of pixel-level ADC can be found working in CMOS nanometer processes. Another minor disadvantage is the slight increase of the area due to area adaptation (ADC) and the larger column lines.

The choice of an array system is dependent on the application where the CMOS sensor is used and therefore choosing an array implementation over the other must be studied. As we saw ADC's play an important role in the array of choice. In the next chapter we study different types of ADC's. Knowing ADC's characteristics, means knowing to which pixel array system it can be associated with and how it influences the array readout time.

4.3 References

- [1] Coghill, John and THEUWISSEN, Albert., "Digital Imaging TECHNOLOGY 101" s.l. : DALSA Corp.
- [2] Pecht, Orly Yardid and Cummings, Ralph Etienne., *CMOS imagers: From phototransuction to image processing*. Dordrecht (NL) : Kluwer academic publisher, 2004.
- [3] Bigas, M, et al., "Review of CMOS image sensors." *Microelectronics Journal*, s.l. : Elsevier, September 6, 2005, Vol. 37 , pp. 433–451
- [4] Tian, Hui, Fowler, Boyd and Gamal, Abbas El., "Analysis of temporal noise in CMOS photodiode active pixel sensor." *IEEE, s.l. : J. Solid-State Circuits*, Jan 2001, Issue 1, Vol. 36, pp. 92-101
- [5] Pearson, Michael., "ON-THE FLY IMAGING PARAMETER ADJUSTMENTS." s.l. : U.S.A Patents, 2001- 08-890343US.

Chapter 5. Analog to digital Converters

A historical view of sampling and coding might introduce terms and facts of this subject and finally give the reader the basic understanding for the continuing analysis of converters. (If you familiar with the sampling theory you may skip this part).

In the late 18th century the French mathematician Fourier unknowingly laid the groundwork for A/D conversion. Every data conversion technique relies on sampling. Sampling is known as the procedure where one is looking at the input signal at regular intervals and creating a digital word for it. When this is done for each value of the input signal we can characterize the result as a sampled waveform of the input. Thanks to Nyquist we know that this actually works. Harry Nyquist discovered while working at Bell Laboratories in the late '20s and wrote a landmark paper¹ describing the criteria for what we know today as sampled data systems. Nyquist taught us that for periodic functions, if you sampled at a rate that was at least twice as fast as the signal of interest, *then no information (data) would be lost upon reconstruction*. And since Fourier had already shown that all alternating signals are made up of nothing more than a sum of harmonically related sine and cosine waves, for example audio signals are *periodic functions* and can be sampled without lost of information following Nyquist's instructions. This became known as the *Nyquist frequency*, which is the highest frequency that may be accurately sampled, and is one-half of the *sampling frequency*. For example the standardized sampling frequency for the audio CD (compact disc) is 44.1 kHz. Taking in account what we just described as the the Nyquist frequency then the theoretical sampling frequency would be is 22.05 kHz.

Even while the Nyquist frequency is a powerful discovery, it isn't without its problem. The problem is called *aliasing* frequencies. Following the Nyquist criteria (Nyquist frequency) guarantees that no information will be lost; it does not, however, guarantee that no information *will be gained*. Although by no means obvious, the act of sampling an analog signal at precise time intervals is an act of *multiplying* the input signal by the sampling pulses. This introduces the possibility of generating "false" signals indistinguishable from the original. In other words, given a set of sampled values, we cannot relate them specifically to one unique signal. For instance we observe the aliasing phenomena in Figure 1, it clearly indicates that the samples could have been from any of the three different waveforms and from all possible sum and difference frequencies between the sampling frequency and the one being sampled.

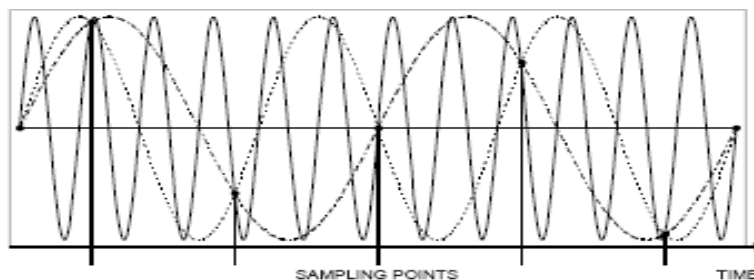


Figure 5.1. False Sampling

All such false waveforms that fit the sample data are called “aliases.” Claude Shannon proved how to prevent us from aliases. Shannon is recognized as the father of information theory: in 1948 at Bell Laboratories as a young engineer, he defined an entirely new field of science. As a 22 year-old student at MIT he showed in his master’s thesis how the algebra invented by the British mathematician George Boole in the mid-1800s, could be applied to electronic circuits. Since that time, *Boolean algebra* has been the backbone of digital logic and computer design. Shannon studied Nyquist’s work closely and came up with a simple extension. He observed (and proved) that if you restrict the input signal’s bandwidth to less than one-half the sampling frequency *then no errors due to aliasing are possible*. So bandlimiting your input to no more than one-half the sampling frequency *guarantees no aliasing*. The only problem is that it isn’t possible. To realize the Shannon limit based on the Nyquist criteria a filter with infinite slope is needed. This is impossible at least for now! Nobody can guarantee that any noise or signal will be greater than the Nyquist frequency. The question that arises here is how to eliminate aliases since the implementing of the anti-aliasfilter cannot be made. There is actually another way, a backdoor, of achieving aliasing free results. The trick is if you cannot restrict the input bandwidth so aliasing does not occur, then increase the sampling frequency until the aliasing products that do occur, do so at ultrasonic frequencies. These are then effectively dealt with by a simple single-pole filter. This is where the term “oversampling” comes in. An example of it is the audio industry which went, in just a few years, from the CD system standard of 44.1 kHz, and the pro audio quasi-standard of 48 kHz, to 8-times and 16-times oversampling frequencies of around 350 kHz and 700 kHz respectively.

In this section we describe the idea of coding an analog signal, the motivation behind coding was already mentioned in the preface. One simple and also very old but efficient technique to implement the coding of the analog signal is the pulse code modulation technique. From the diagram below one can easily see that it consists of three blocks: A sampler, a “quantizator” and finally a coder-block. The output of a PCM system is a series of digital words, where the word-size is determined by the available bits. In the Quantization paragraph it will be clear why also they are called PCM A/D converters.

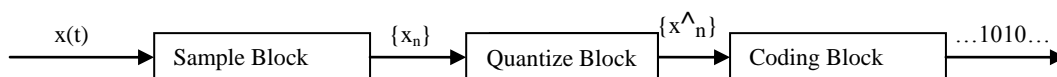


Figure 6-2. Diagram of a PCM system

Most A/D converters can be classified into two groups according to the sampling rate criteria. Nyquist rate converters, such as a successive approximation register (SAR), double integration, and oversampling converters, sample analog signals which have maximum frequencies slightly less than the Nyquist frequency, $f_N = f_s / 2$, where f_s is the sampling frequency. Meanwhile, oversampling converters perform the sampling process at a much higher rate, $f_N \ll F_s$, where F_s denotes the input sampling rate.

5.1 Nyquist Rate Conversion

Let's review and analyze the sampling theorem in the case of the converters. We know that digital conversion of a signal is traditionally described as either as uniform sampling or either as quantization in amplitude. In the first case the signal is a continuous time signal is sampled at uniformly spaced time intervals T . The samples, $x[n]$, of the continuous time signal $x(t)$ can be represented as $x[n]=x(nT)$. In the frequency domain, the sampling process creates periodically copies, also called images, of the original continuous time signal. Mathematically this can be described as follows:

$$X_s(f) = \frac{1}{T_s} \sum_{k=-\infty}^{\infty} X(f - kf_s) \quad (5.1)$$

where $X_s(f)$ represents the spectrum of the sampled signal, and $X(f)$ is the spectrum of the original continuous time signal. To know how this might look like in the frequency domain, where most of analysis take place, see Figure 5.3

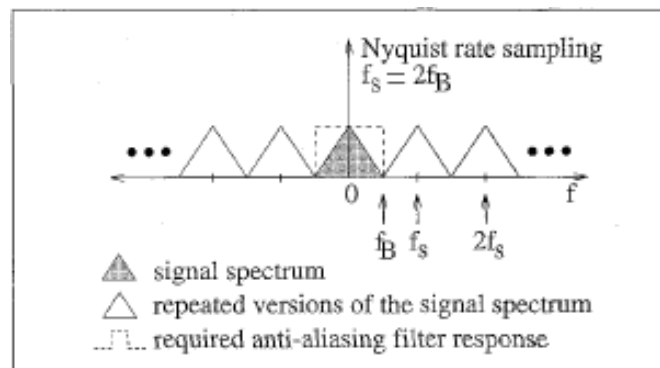


Figure 5.3. showing the band-limited input signal spectrum, copies of the input due to sampling and the needed anti-aliasing filter response. $f_s = 2f_B$, f_B is the bandwidth of the signal

So we conclude that the signal can be reconstructed back to continuous time if the repeated versions of the signal spectrum do not overlap. This occurs if the signals spectrum is band limited to half the sampling rate. For the previous example the signal with bandwidth f_B must be sampled at a rate greater than twice its bandwidth $f_s \geq 2f_B$.

Aliasing, as we mentioned before, is the interference between the repeated versions of the signal spectrum and prevents reconstruction of the signal. An anti-aliasing filter is often used, even in the case when we know that the signal is nominally band limited to $f_s/2$, just to ensure that the signal is indeed band limited. For instance speech has a bandwidth of 4 kHz and thus can be sampled at 8 kHz. Even so, there is some residual signal energy above 4 kHz which is responsible that aliasing occurs at 8 kHz sampling rate. To prevent aliasing we put an anti-aliasing filter, a continuous time analog filter, before the sampler. The case where $f_s = 2f_B$ is known as Nyquist rate sampling, and clearly the cutoff frequency off the anti-aliasing filter must be very sharp, namely for our example it has to be $f_B = f_s/2$. Because this is in practice very difficult we will later explain how to relax the cutoff frequencies. The sampling process, an invertible operation, also automatically involves the discretization or better known as the quantization of the signal.

5.1.1 Quantization

After sampling, the signal samples must also be quantized in amplitude to a finite set of output values. Transfer characteristics of quantizers or A/D converters with an input signal sample, $x[n]$, and an output, $y[n]$, can be seen in Figure 5.2. However quantization is a non-invertible process, since finite number of output amplitude values are mapped to infinite numbers of input amplitude values. The finite numbers of bits at the output produce the digital code words each one of which is corresponding to a quantized output amplitude. For instance, for the 1 bit AD converter of Figure 5.4c, the output levels V and $-V$ can be mapped to digital codes “1” and “0.” The digital code words are said to be in pulse code modulation (PCM) format.

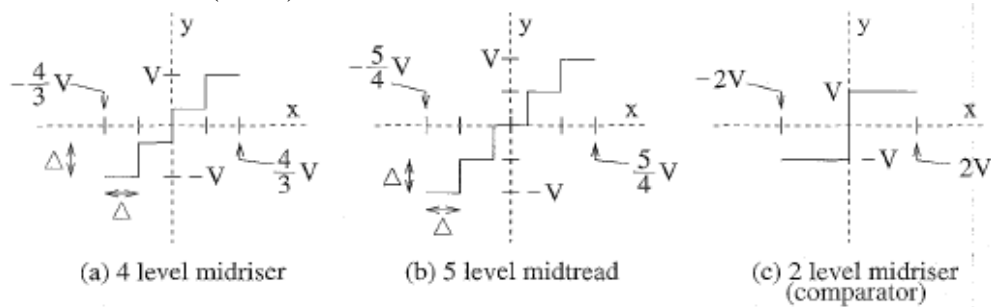


Figure 5.4. Typical transfer characteristics of quantizers

From another point of view the quantized output amplitude values can be considered as the output of an ideal digital to analog converter (DAC) whose inputs are the corresponding digital code words. An ADC or quantizer with Q output levels is said to have N bits of resolution where $N = \log_2(Q)$. As should be clear from Fig 5.4, for an ADC with Q quantization levels, only input values separated by at least $\Delta = 2V/(Q-1)$ can be distinguished or resolved to different output levels. N digital bits are needed to encode the Q code words corresponding with each output level. The difference between the binary digital codes for two adjacent output levels is one least significant bit (LSB) of the overall N bit codeword. Consequently, a difference in input amplitudes corresponds to a one LSB difference in the digital output code words. To represent some properties of the quantization we demonstrate with our examples of quantization that an even number of levels in the transfer characteristic is often more desirable than one with an odd number. In Figures 5.4a-4b we represent a four level (2 bit) “midriser” and a five level (roughly 2 bit) “midtread” ADC. Unlike the midtread ADC, the midriser ADC does not contain a zero output level for a zero input value, effectively creating a DC offset that may be undesirable in some applications. Note that the midriser needs to have an even number of output levels to produce a completely symmetric transfer curve, whereas the midtread needs an odd number of output levels. Thus the midriser ADC’s symmetric characteristic, with an even number of levels, is more desirable since the number of output levels, Q , can be made a power of two and encoded with exactly $N = \log_2(Q)$ bits. In contradiction to the symmetric midtread ADC with odd numbers, the number of output levels Q , must be odd, and so cannot be made a power of two and encoded as efficiently. The number of bits needed will be $N = \log_2(Q-1) + 1$, where $Q-1$ is chosen a power of two. If the number of levels for the midtread ADC is forced to be a power of two by using only $Q-1$ levels,

it will no longer have a symmetric transfer characteristic and will distort large amplitude symmetric input signals (e.g., a sinusoid). This distortion, of course, may be negligible when the number of output levels is very large. The conclusion of the above comparison is that one should carefully choose whether a midriser or midtread quantizer transfer characteristic should be used. For the two level quantizer, a midtread characteristic will not be able to represent both positive and negative output levels, and so will severely distort a signal containing samples of both polarities. For this 2 level case, a midriser characteristic, shown in Figure 5.4c, will almost always be used.

5.1.2 Nyquist ADC Errors

The midriser of course is an ideal case in analog signal conversion and often there are errors that distort, shift or even “swallow” voltage levels. These voltage levels are the known voltage staircase or simply called staircase. We will mention briefly the basic errors, like explained thoroughly in [1], that can occur. For more insight please refer to [1].

5.1.2.1 Quantization error

This error in the quantization step is defined as the difference of the analog input voltage from the output voltage.

$$Q_e = V_{in} - V_{staircase} \quad (5.2)$$

The staircase voltage is given as

$$V_{staircase} = D_w \frac{V_{ref}}{2^N} = D_w V_{LSB} \quad (5.3)$$

where D_w is the digital word, LSB the least significant bit in the word formation, V_{LSB} the voltage equivalent of the LSB in the staircase and V_{ref} the reference voltage or otherwise interpreted as the range of voltage for which the ADC must convert.

5.1.2.2 Differential Nonlinearity (DNL)

DNL is the difference between the actual word produced and the ideal word. This observed with the step width:

$$DNL = Actual\ step\ width - Ideal\ step\ width \quad (5.4)$$

The steps can be interpreted as voltage deviations, which we already know, can be illustrated with the LSB voltage.

5.1.2.3 Missing Codes

As we see from the staircase, a step (level) accords to a voltage, and the voltage in turn to digital output. When DNL indicates to be more or less than one LSB, then in this case, a missing code can occur.

5.1.2.4 Integral Nonlinearity (INL)

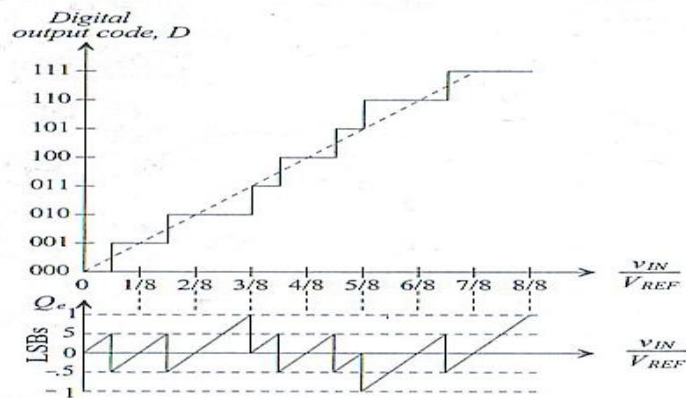
With all other possible errors set to zero, INL is defined as the difference between the data converter code transition points and the straight line, which goes through the end transition points of the ideal converter

5.1.2.5 Offset and Gain Error

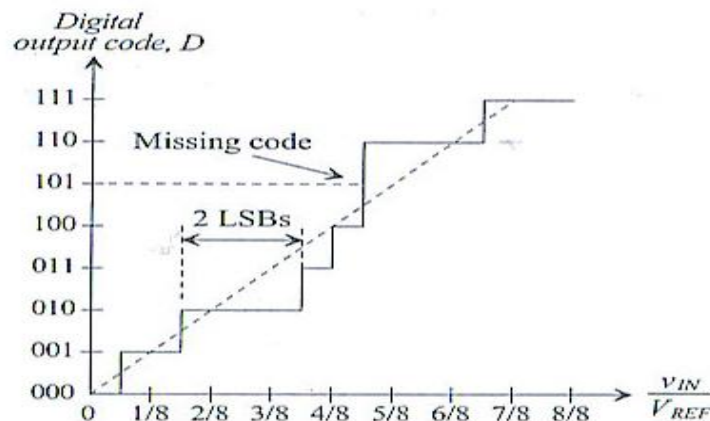
Offset error exist if the straight line though the end transitions is shifted and a results in a difference of 1LSB. As a result the non-ideal and the transition line are parallel and separated by 1 LSB. On the other hand gain error, like offset error, results in a mismatch of the transition line, but gain error starts from the ideal position. This is the reason that its slope differs from the ideal slope.

5.1.2.6 Aliasing

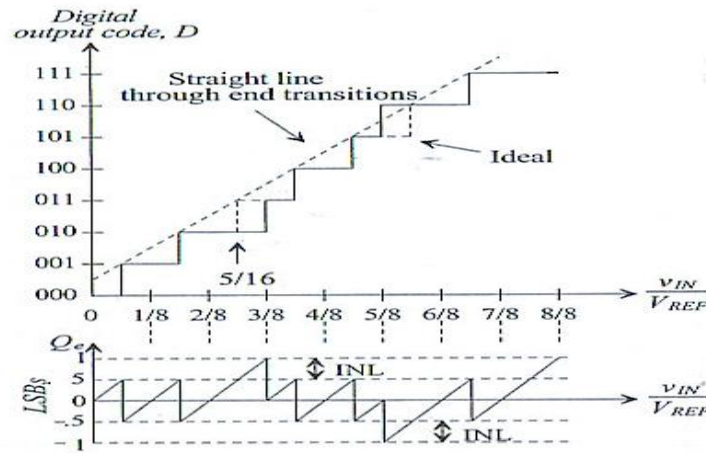
Like mentioned before, it is the result of overlapping frequencies when the sampling frequency is not twice as the Nyquist frequency. To completely understand the errors, it is best to represented them graphically:



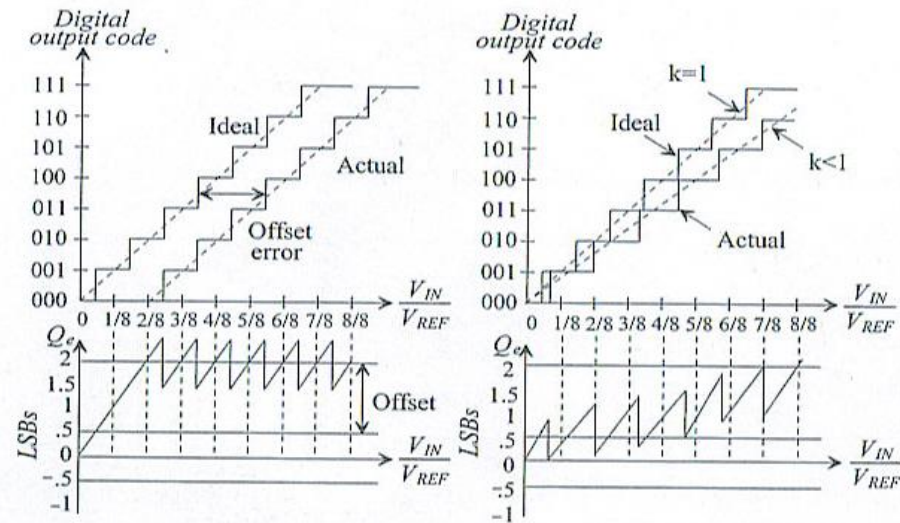
(a) DNL error. A profound error in Nyquist ADC's. Its deviation from the ideal step widths is described with the associated voltage differences (expressed in LSB).



(b) Missing Codes possible when LSB is more or less than one.



(c) Integral Nonlinearity is similar to DNL, with the difference that it describes the difference from ideal quantization error, expressed in LSB.



(d) Offset error and gain error are deviations from the ideal straight line.

Figure 5.5. All important errors in a Nyquist rate Converter. Shown in the figures are errors that cause wrong quantization outputs (digital codes). The example is for an input voltage of 5V, and to achieve an 8-bit resolution, the reference voltage is $V_{REF} = 5/8 = 0.625V$. Figures from [1]

Now that we introduced sampling and quantization, the two key operations of an A/D converter, we are now ready to have a first glance into some of the most common A/D conversion implementations. The representations of different A/D converters will help us understand the main reason of choosing a specific A/D converter beyond other. A general overview of Nyquist rate converters actions can be thought as follows:

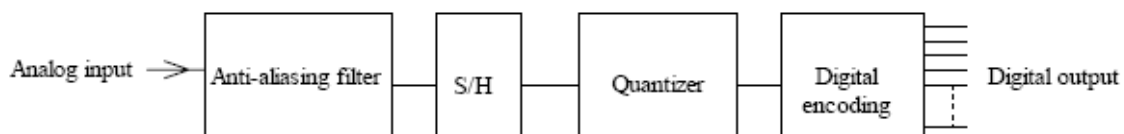


Figure 5.6. Generalized diagram of the Nyquist rate converters. Note the Store and Hold (S/H) device performs the sample function. Sometimes amplification is needed due to glitches in the circuit of the device and thus in that occasion an S/H is called a SHA.

5.1.3 ADC architectures

5.1.3.1 Successive Approximation

One of the earliest and most successful analog-to-digital conversion techniques is the successive approximation. The heart of any A/D circuit is a comparator. A comparator is an electronic block whose output is determined by comparing the values of its two inputs. If the negative input exceeds the positive input, the output swings negative otherwise if the positive input is larger than the negative input then the output swings positive. Therefore if a reference voltage is connected to one input and an unknown input signal is applied to the other input, you now have a device that can *compare* and tell you which is larger. Thus a comparator gives you a “high output” (which could be defined to be a “1”) when the input signal exceeds the reference, or a “low output” (which could be defined to be a “0”) when it does not. A comparator is the heart in the successive approximation technique as Figures 5 and 6 shows. The idea is quite simple; the

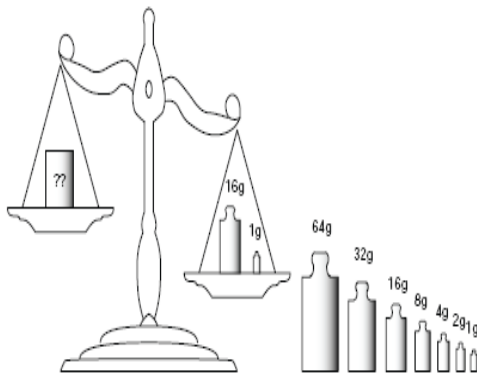


Figure 5.6. Successive approximation idea.

circuit evaluates each sample and creates a digital word representing the closest binary value. It is analogous to a gold miner's assay scale, or a chemical balance as seen in Figure 6. The scale comes with a set of graduated weights, each one half the value of the preceding one, such as 1 gram, $\frac{1}{2}$ gram, $\frac{1}{4}$ gram, $\frac{1}{8}$ gram, etc. To measure its mass of the unknown weight you compare the unknown sample against these known values by first placing the heaviest weight on the scale. If it tips the scale, you remove it; if it does not you leave it and go to the next smaller value. If that value tips the scale, you remove it, if it does not, you leave it and go to the next lower value, and so on until you reach the smallest weight that tips the scale. Finally if the scale does not tip when you put the last weight on, then you put the next highest weight back on, and that is your best answer. In other words the sum of all the weights on the scale is the closest value you can resolve. Returning back to our successive approximation A/D converter the number of steps is equal to the bits available, for example, a 16-bit system requires 16 steps for *each sample*. The analog sample is successively compared to determine the digital code (like the weights on the scale), beginning with the determination of the biggest (most significant) bit of the code. To compare the scale example with what the converter is doing we can say that a “0” was assigned to each weight removed, and a “1” to each weight remaining – in essence creating a digital word equivalent to the unknown sample, with the number of bits equaling the number of weights. And the quantizing error will be no more than $\frac{1}{2}$ the smallest weight (or $\frac{1}{2}$ *quantizing step*). We next illustrate an example of SAR. A 16-bit would be too lengthy as an example but 3-bit is feasible.

The SAR algorithm is as follows:

- a. An applied “1” forces the shift register to perform a shift to the right. Hence the bits in the register, represented as $B_{N-1} - B_0$, are permuted moving for instance a one to the right.

- b. The initial condition of the SAR is that the MSB of the digital output, represented as D_{N-1} , is set to one and all remaining bits are zero.
- c. The output of the SAR controls the DAC. For the initial condition at which the SAR output is 100, the reference voltage is set as $DAC_{out} = V_{ref}/2$.
- d. The comparator at the bottom then compares the DAC_{out} to the input voltage, yielding a zero for less and 1 for greater the input. For 1 the D_{N-1} is reset to 0 and for a 0 output D_{N-1} remains unchanged.
- e. Having a result of the comparator, the shift register is pulsed to perform another shift. This is equal like saying a conversation is performed, go to the next one.
- f. The next bit in the SAR is set to one D_{N-2} , and analogous with the value of previous bit D_{N-1} , the reference voltage of the DAC is set either to $V_{ref}/2$ or $3V_{ref}/4$.
- g. Again the input voltage is compared to the reference voltage of the DAC and a 1 is resulting for greater and a 0 for less than the input value. The comparator output sets as in step d. the D_{N-2} to 1 or 0.
- h. Repeating this procedure yields the correct digital output code for the input. Of course since it takes N steps, N as for the desired resolution, the input signal must be hold till the conversation is done. This is done by a store and hold circuit (S/H).

As before the 3-bit example is best presented graphically.

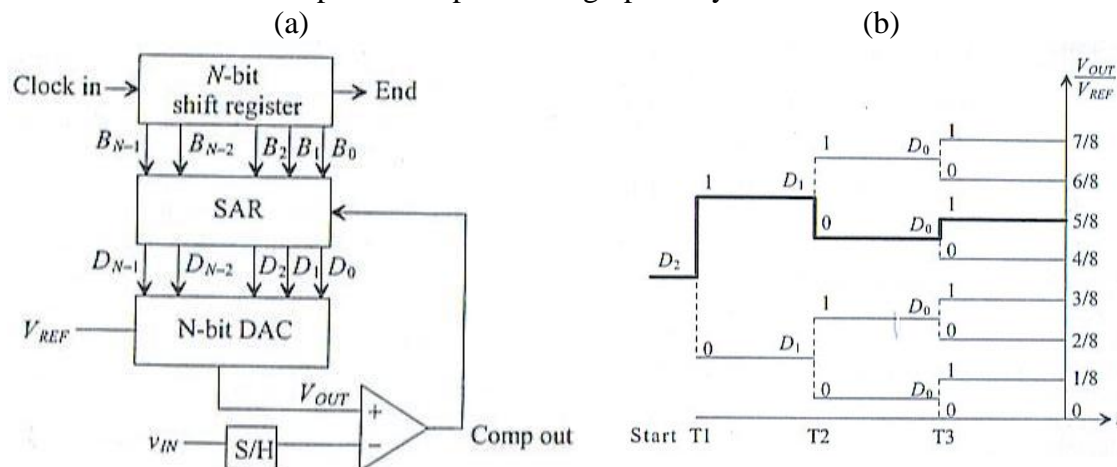


Figure 5.6. (a) Logical diagram of a successive approximation converter. (b) The 3-bit SAR. The comparator guides the search through the possible outputs. This is the analogy to the gold miner scale that removes or adds weights. Figures from [1].

As stated earlier the successive approximation technique must repeat this cycle for each sample. Even with today's technology, this is a very time consuming process and is still limited to relatively slow sampling rates. On the other hand we have to give it some credits since a 16-bit successive converter, brought us to the 44.1 kHz digital audio world.

5.1.3.2 Flash

The *flash A/D converters*, also known as parallel ADC's uses the distributed sampling to achieve a high conversion speed. These are the simplest and potentially the fastest of the entire ADC's available. The flash ADC's do not need explicit front end sample and hold circuits and their performance is determined primarily by that of their constituent comparators. Since comparators do not achieve much higher speeds than sample and hold amplifiers (SHA's), flash ADC's can operate faster than front end SHA's. An N-bit flash ADC consists of 2^N resistors and $2^N - 1$ comparators arranged as in Figure 5.7 Each comparator has a reference voltage from the resistor string which is 1 LSB higher than that of the one below it in the chain. For a given input voltage, all the comparators below a certain point will have their input voltage larger than their reference voltage and a "1" logic output, and all the comparators above that point will have a reference voltage larger than the input voltage and a "0" logic output. The $2^N - 1$ comparator outputs therefore behave in a way analogous to a mercury thermometer, and the output code at this point is sometimes called a "thermometer" code. Since $2^N - 1$ data outputs are not really practical, they are processed by a decoder to generate an N-bit binary output.

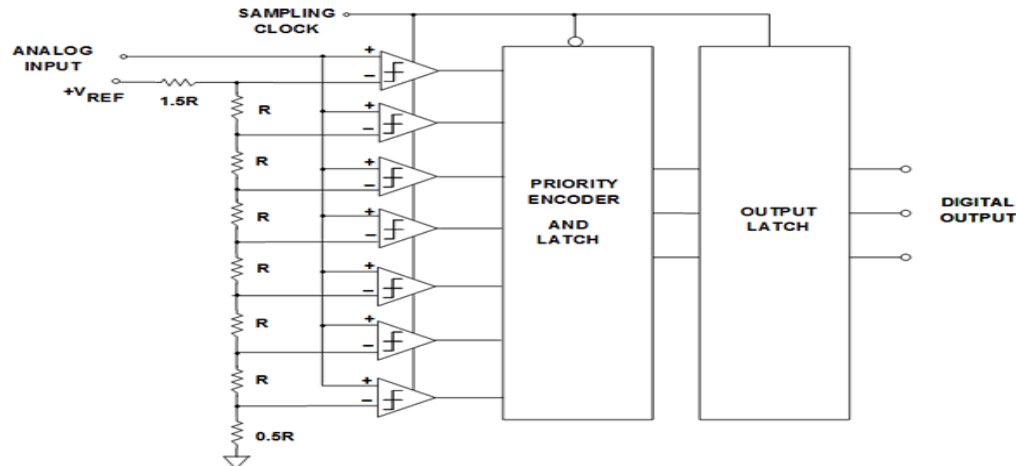


Figure 5.7. A 3-bit flash parallel converter

The input signal is applied to all the comparators at once, so the thermometer output is delayed by only one comparator delay from the input, and the encoder N-bit output by only a few gate delays on top of that, so the process is very fast. In addition, the individual comparators provide an inherent "sample-and-hold" function, so theoretically a flash converter does not need a separate SHA, provided the comparators are perfectly dynamically matched. In practice, however, the addition of a proper external sample-and-hold usually enhances the dynamic performance of most flash converters because of the inevitable slight timing mismatches which occur between comparators. Because the flash converter uses large numbers of resistors and comparators and is limited to low resolutions, and for it to be fast, each comparator must run at relatively high power levels. Hence, the problems of flash ADCs include limited resolution, high power dissipation because of the large number of high speed comparators (especially at sampling rates greater than 50 MSPS), and relatively large (and therefore expensive) chip sizes. In addition, the resistance of the reference resistor chain must be kept low to supply

adequate bias current to the fast comparators, so the voltage reference has to source quite large currents (typically $> 10 \text{ mA}$).

5.1.3.2.1 Pipeline A/D Converters

The most popular ADC architecture for sampling rates from a few megasamples per second (MS/s) up to 100MS/s+, with resolutions from 8 bits at the faster sample rates up to 16 bits at the lower rates is the pipelined analog-to-digital converter (ADC). Particular because of the widespread sampling rate a wide range of applications, including CCD imaging, ultrasonic medical imaging, digital receiver, base station, digital video (for example, HDTV), xDSL, cable modem, and fast Ethernet, are using the pipeline technique. The general concept of *pipeline ADC's* is that in each stage it carries out an operation on a sample, provides the output for the following sampler and once that sampler has acquired the data begins the same operation on the next sample. As every stage incorporates a sample and hold function, the analog data is preserved, allowing different stages to process different samples concurrently. Thus, the conversion rate depends on the speed of only one stage, usually the front end. While the concurrent operation of pipelined converters makes them attractive for high speeds, their extensive linear processing of the analog input relies heavily on operational amplifiers, which are relatively slow building blocks in analog design. A main advantage of the for an N-stage pipeline converter is its high throughput. After an initial delay of N clock cycles, one conversion will be completed per clock cycle. The disadvantage is having the initial N clock cycle delay before the first digital output appears. A slight error in the first stage propagates through the converter and results in a much larger error at the end of the conversion. A more specific example is a 12-bit pipelined ADC showed in Figure 8. In the example the analog input V_{IN} is

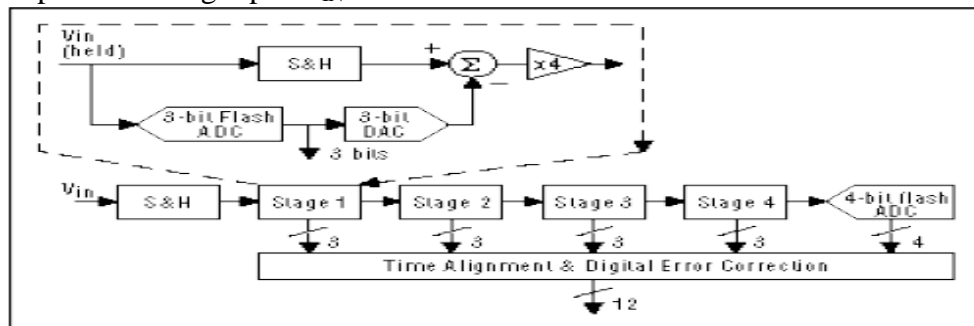


Figure 5.8. 12-bit pipelined ADC

first sampled and held constant by a sample-and-hold (S&H), while the flash ADC in stage one quantizes it to 3 bits. The 3-bit output is then fed to a 3-bit DAC (accurate to about 12 bits), and the analog output is subtracted from the input. The "remainder" is then gained up by a factor of 4 and fed to the next stage (stage two). This gained-up remainder continues through the pipeline, providing 3 bits per stage until it reaches the 4-bit flash ADC, which resolves the last 4LSB bits. Because the bits from each stage are determined at different points in time, all the bits corresponding to the same sample are time-aligned with shift registers before being fed to the digital-error-correction logic. Note that as soon as a certain stage finishes processing a sample, determining the bits and passing the residue to the next stage, it can start processing the next sample due to the sample-and-

hold embedded within each stage. Like in a CPU pipelining is the main reason for the high throughput. Considering each sample has to propagate through the entire pipeline before all its associated bits are available for combining in the digital-error-correction logic, data latency is a major concern and has always to be associated with pipelined ADCs. Digital Error Correction is often needed in most modern pipelined ADCs. It reduces the accuracy requirement of the flash ADCs (namely that of the individual comparators). In Figure 8, the 3-bit remainder at the summation-node output has a dynamic range $1/8$ that of the original stage-one input (V_{IN}), but the subsequent gain is only 4. Therefore, the input to stage two occupies only half the range of the 3-bit ADC in stage two (in the case when there is no error in the first 3-bit conversion in stage one). In the case when an analog input close to the trip point of this comparator is applied and one of the comparators in the first 3-bit flash ADC has a significant offset, then an incorrect 3-bit code and hence an incorrect 3-bit DAC output would result, producing a different remainder. Albeit, it can be proven that, as long as this gained-up remainder doesn't over-range the subsequent 3-bit ADC, the LSB code generated by the remaining pipeline when added to the incorrect 3-bit MSB code will give the correct ADC output code. The implication is that none of the flash ADCs in Figure 1 has to be as accurate as the entire ADC. Forsooth, the 3-bit flash ADCs in stages one through four require only about 4 bits of accuracy. The digital error correction will not correct for errors made in the final 4-bit flash conversion. However, any error made here is overwhelmed by the large (4^4) cumulative gain preceding the 4-bit flash, requiring the final stage to be only more than 4-bits accurate. When we look our example in Figure 8, howbeit each stage generates 3 raw bits, cause of the interstage gain is only 4, each stage (stages one to four) resolves only 2 bits. A third bit, called "1-bit overlap" between adjacent stages, is used to reduce the size of the remainder by one half, allowing extra range in the next 3-bit ADC for digital error correction. The effective number of bits of the whole ADC is therefore $2 + 2 + 2 + 2 + 4 = 12$ bits. The component accuracy is a matter very delegate and important when designing a pipeline A/D.

Errors like gain or linearity in the individual DAC and gain amplifiers are not corrected by the digital error correction. Cause in the subsequent stages error terms are divided down by the preceding interstage gain(s), less accuracy is needed there (for instance, 10-bit for stage two, 8-bit for stage three, and so forth), then the S/H's and the DAC which need 12-bit accuracy. This event is often further enhanced and analyzed to further save power by making the pipelined stages progressively smaller. In most pipelined ADCs designed with CMOS or BiCMOS technology, the S&H, the DAC, the summation node, and the gain amplifier are usually implemented as a single *switched-capacitor* circuit block called a multiplying DAC (MDAC). The major factor limiting MDAC accuracy is the inherent capacitor mismatch. On the other hand a purely bipolar implementation would be more complicated and would suffer mainly from resistor mismatch in the current source DAC and the interstage gain amplifier. For the first couple of stages often some form of capacitor and resistor trimming or digital calibration is required, to achieve 12 bits accuracy or higher.

5.1.3.3 Integrating ADC

In this structure a digital counter is incremented when the input signal is reached. The latter is been done by integrating the input signal from 0 to the actual input. There are two versions of integrating ADC, called single and dual slope converters.

5.1.3.3.1 Single slope

The basic logic blocks that the single slope ADC consist of are: An operational amplifier in the negative feedback arrangement, a comparator, a latch and a counter. This is shown in figure 5.9. The actual value of the input is proportional to the number of clock pulses. The integrator ideally starts at zero and linearly increases with a slope that depends on the gain integrator. The reference voltage is negative so that the integrator output is positive. When the output of the integrator surpasses the value of the input value the control logic pulses the latch to store the counters value. The time that a output value is produced depends on the magnitude of the input value, because the larger the input the larger the time the integrator takes to reach that value. Hence the frequency of operation must be much greater or at least equal to the frequency of the existing maximum input value for a conversation to take place. Therefore to have a short conversation time the clock frequency must be very high, about 2^N clock cycles. Then the conversation time is described as

$$t_C = \frac{V_{in}}{V_{ref}} 2^N T_{CLK} \quad (5.5)$$

5.1.3.3.2 Dual slope

The architecture aims of overcoming the clock, resistors and capacitors issues of the single slope by simply integrating twice once for the input signal and once for the reference voltage. The latter is not exactly rising voltage integration but more a decreasing voltage integration. What was meant by decreasing is that after the input signal is integrated the signal is not reset to make another identical integration, like in the single slope, but instead the counter is saved in the latch and after that, the input voltage is disconnected, and the reference voltage is connected to the negative feedback topology. The discharge of the integrator value follows and the voltage is compared to the reference voltage and each time its value is surpassed the counter is increased. The number of counts is the digital output code. One might think that this architecture is not superior to the single slope, but looking at it from another prospective one discovers the opposite. The nonidealities, like capacitor and resistor mismatch, that exist in the single slope result in inaccuracies. By using the same clock for both slopes theses nonidealities are cancelled out.

$$\begin{aligned} \text{charging} \quad : \quad V_C &= \frac{1}{C} \int_0^{T_1} \frac{V_{in}}{R} dt = \frac{V_{in}}{R} T_1 \\ \text{and discharging: } V_C &= \frac{V_{in}}{R} T_1 - \frac{1}{C} \int_0^{T_2} \frac{V_{ref}}{R} dt = 0 \\ &\Rightarrow V_{in} T_1 = V_{ref} T_2 \\ \text{and the relationship holds } \frac{\text{counter output}}{2^N} &= \frac{V_{in}}{V_{ref}} \end{aligned} \quad (5.6)$$

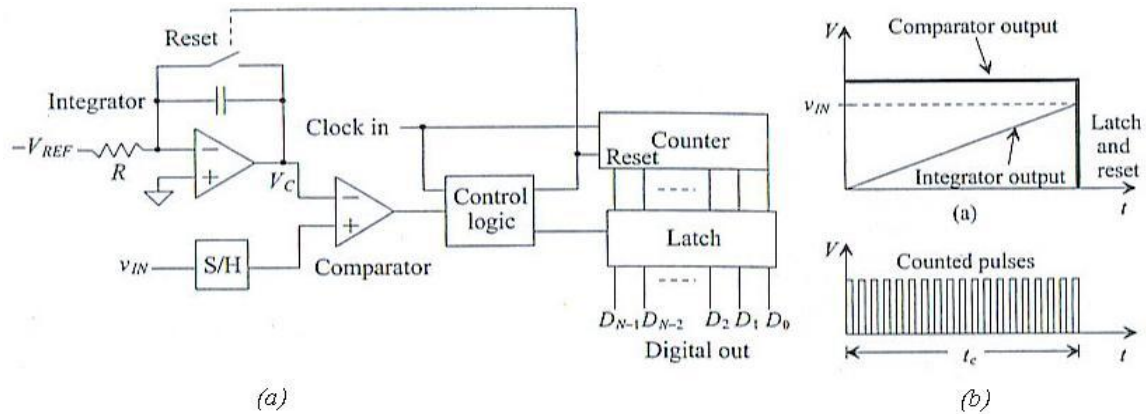


Figure 5.9. (a) Circuit component overview (b) Integrating the input value until overflow. Number of Pulses are the digital code. Figure from [1]

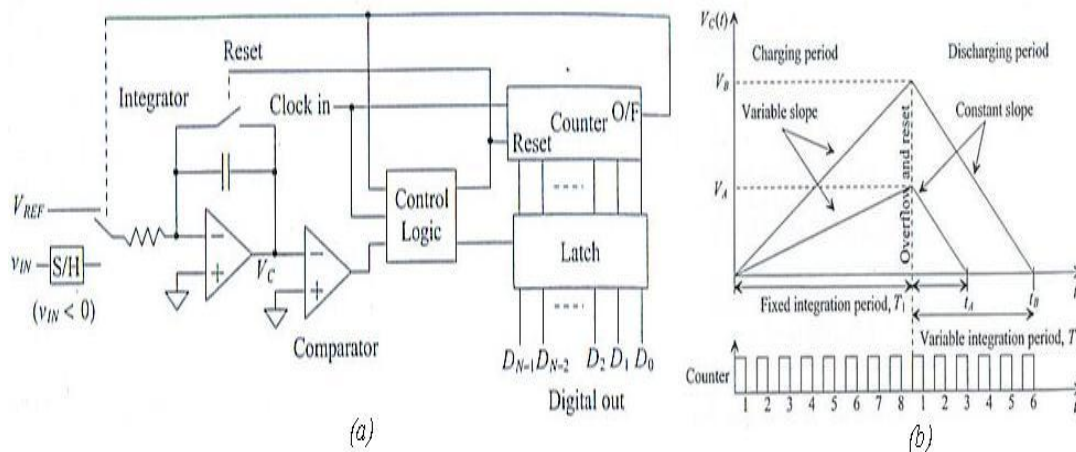


Figure 5.10. (a) Circuit component overview (b) Integrating the input value and then the discharging the same integrator value with V_{ref} as overflow value. Number of discharge pulses is the digital code. Figure from [1]

5.1.4 Bandwidth and Resolution Relationship

From the above we can see that most conventional A/D converters, such as the successive approximation, subranging, pipelined and flash converter types quantize signals sampled at, or slightly above, the Nyquist rate. Consequently, these converters are Nyquist rate PCM converters. Their main characteristic is to provide tradeoffs among signal bandwidth, output resolution, and the complexity of the analog and digital hardware. A comparison between sigma-delta and the Nyquist conversion A/D's based on their qualitative bandwidth and resolution is shown in Figure 9. As latter it becomes understandable sigma-delta A/D converters attain the highest resolution for relatively low signal bandwidths. Consequently, sigma-delta techniques are often used in speech applications where the signal bandwidth is only 4 kHz and where up to 14 bits of resolution may be needed. Also sigma-delta ADCs are popular for digital audio applications, where the signal bandwidth is 20-24 kHz and where high fidelity audio requires 16-18 bits of resolution. Pipelined and Flash converters, on the other hand, may be used for broadcast video applications where the signal band is about 5 MHz, but the resolution required is only about 8 bits.

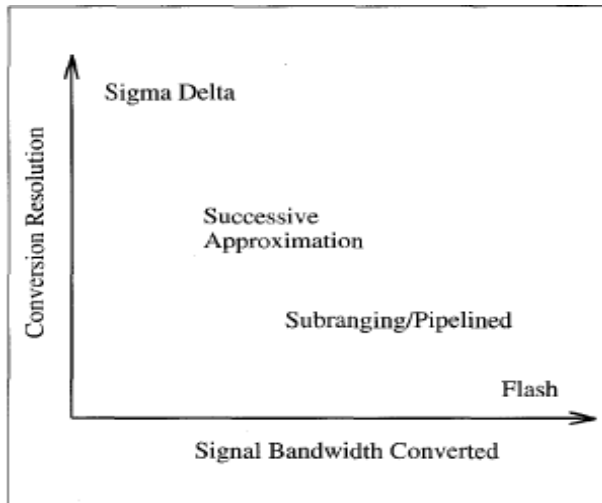


Figure 5.11. Bandwidth and Resolution tradeoffs of different ADC's

5.1.5 Designing A/D Performance

Having made a proper introduction in the sampling and quantization processes, we now examine the A/D converter and characterize its performance. The diagrams in Figure 5.4 show the transfer characteristic for typical quantizers with input x and output y . Assume that the maximum and minimum quantized output values always are V and $-V$. The least significant bit (LSB) of an ADC with Q quantization levels is equivalent to $2V/(Q-1)$. For both the midriser and

midtread type of ADCs of 5.4, the magnitude of the quantization error ($e = y - x$) between the output and input does not exceed half a LSB, mathematically $|e| \leq \Delta/2$, given that $|x| \leq V + \Delta/2$. Under these conditions, the quantizer or the ADC is characterized as not overloaded. Otherwise for $|x| > V + \Delta/2$ (therefore $|e| \leq \Delta/2$), the ADC is said to be overloaded. Unfortunately the quantizer embedded in any ADC is a non-linear system, what makes its analysis difficult. The quantizer is often linearized, to make the analysis amenable. This is done by modeling it by a noise source, $e[n]$, added to the signal $x[n]$, to produce the quantized output signal $y[n]$ which is finally expressed as: $y[n] = x[n] + e[n]$.

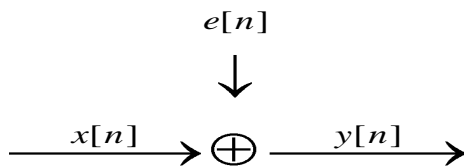


Figure 5.12. Output representation of a quantizer

To handle the added noise, by the quantizer, in the analysis of the system, we assume that the noise process, $e[n]$, has some statistical properties. These are eventually justified and shown to be true by the many observations of the output statically characteristics. Mainly we adapt that the signals through the quantizer device are represented by a stationary random process,

such that all error samples, are stationary random process samples. Also the error sequence $e[n]$, must be uncorrelated with the sequence $x[n]$, such that the output $y[n]$ can be seen as the sum of uncorrelated signals, $x[n]$ and $e[n]$ respectively. The error process is modeled by a probability density function (pdf) which distributes the $e[n]$ uniformly over the range of quantization error, namely $\pm\Delta/2$. From this we conclude, for a simplified analysis the error must be a white noise process. Under the assumption when N is large, in example the quantizer is not overloaded, then the successive signal values are not excessively correlated.

To derive the quantization noise ratio, we consider an N bit ADC with $Q = 2^N$ quantization levels, with a quantization step $\Delta = 2V / (Q-1) = 2V / (2^N-1)$. Since we deal with uniform Gaussian distribution we have zero mean $e[n]$ and variance or power of σ_e^2 . Finally it should be stressed that the signals power σ_x^2 is taken from a random process with a zero mean value. Combining σ_e^2 and σ_x^2 we can derive the SNR:

$$\sigma_{\epsilon}^2 = \frac{\Delta^2}{12} = \frac{\left(\frac{2V}{2^N-1}\right)^2}{12} \cong \left(\frac{2V}{2^N}\right)^2 \quad (5.7)$$

A more rigorous analysis on the quantization noise can be found in [2]. From the definition of the signal to noise ratio we conclude:

$$SNR \equiv 10 \log \left(\frac{PowerOfSignal}{PowerOfNoise} \right) = 10 \log \left(\frac{\sigma_x^2}{\sigma_{\epsilon}^2} \right) = 10 \log \left(\frac{\sigma_x^2}{V^2} \right) + 4.77 + 6.02 \text{ (dB)} \quad (5.8)$$

Equation 5.8 shows that the SNR improves by 6 dB for every increment of N, and this accord to an increase of one bit in resolution. So an ADC that has an improved SNR by x dB has an increased resolution by $x/6dB$ bits. The range of an Nyquist ADC is confined by its SNR value, to find out this range consider a general input in the form of a sinusoidal wave $V_{sin} = A \sin \omega t$. The mean square of the input is then $V_{sin}^2 = \frac{1}{2\pi} \int_0^{2\pi} A^2 \sin^2 \omega t$. This in turn is equal to the power of the sinusoidal input and therefore the SNR becomes

$$SNR_{dB} = 10 \log \left[\frac{\left(\frac{A^2}{2}\right)}{\left(\frac{A^2}{3 \cdot 2^{2N}}\right)} \right] = 10 \log \left(\frac{3 \cdot 2^{2N}}{2} \right) = 6.02N + 1.76dB \quad (5.9)$$

In reality the value of SNR is reduced by previously introduced errors that may exist, so equation 5.9 represents the ideal case. For example if we have a missing code, or equivalently one less LSB, then this results immediately in a loss of resolution of 1 bit. In other words the SNR drops by 6dB and the new SNR is $SNR_{dB} = 6.02N + 1.76dB - 6dB = 6.02N - 4.24dB$.

5.1.6 Performance Constraints

In every type of Nyquist rate converter the resolution depends on the accuracy of the devices, each sample is quantized at the full precision of the converter. Therefore implanting the ADC with a certain technology such as integrating circuits on VLSI chips, one must sure that the accuracy can easily be met. Like seen resistances and capacitors play an important role in defining the reference voltage, therefore any mismatching issues must be minimized. If an ADC is to convert a signal to 8 bits resolution then the capacitors matching must met the requirements of accuracy to be at least 2^8 . In general this is difficult and an optimum of 0.1% mismatch is achieved. If the resolution is to be higher than 8 bits, then methods to match the components (R or C) such as laser trimming or calibration must be considered. Furthermore anti-aliasing filters much have a sharp cut-off frequency, something that is impossible to construct. Hence guard bands to relax the frequency cut-off frequency must be inserted into the filter. This can further reduce resolution.

5.2 Oversampling ADC

An introduction into oversampling can be done by introducing first the PCM and second the PWM converter. Both ADC's are left out intently because they form a "bridge" between oversampling and Nyquist converters. Having understood basic principles of oversampling we will explain the most common used oversampling converter, the sigma delta converter.

5.2.1 PCM converter

The PCM converter can operate at Nyquist rate and also at higher frequencies. An example of PCM is the SAR ADC. The digital output code is defined by the available bits to shift. In the PCM system the total quantization noise σ_e^2 is the same as described in equation 5.7. For higher sampling frequencies the difference between the Nyquist rate and the oversampling frequencies lies in the distribution of their power spectral densities. The noise is considered to be white noise and therefore we have a uniformly spreading of the noise over the sampling frequencies f_{over} and f_{Nyquist} .

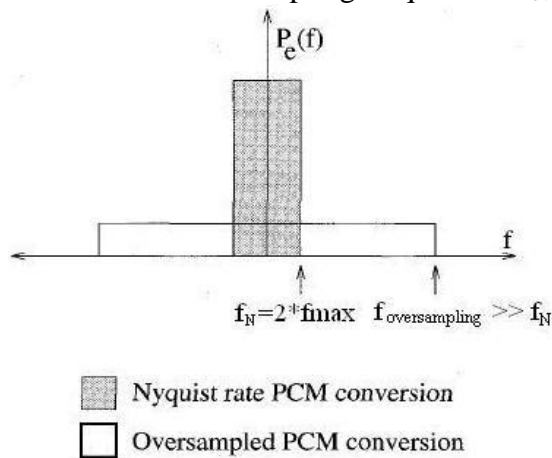


Figure 5.13. Oversampling and power spectrum distribution

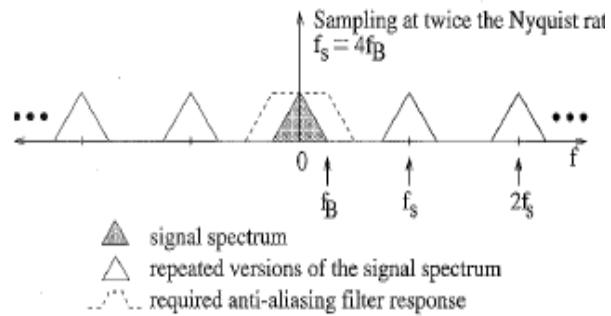


Figure 5.14 Image replicas with oversampling. Observe that the filter cut-off can be set up to $f_{\text{sampling}} / 2$.

The advantage of oversampling is having less noise in the band of interest ($f_B = f_{\text{max}}$). After the sampling is performed the samples are passed through a low pass filter (LPF), where the image replications due to sampling are removed and only the content in the frequency range of interest $\{-f_{\text{max}}, f_{\text{max}}\}$ are allowed to pass through. The result isn't useful since we cannot compare it with to normal frequencies, hence the output must be downsampled to Nyquist frequency. Imagine having a LP disc turning very fast, the pitch would be so high that we couldn't hear the music. The goal is to downsample so that we can reconstruct it. Furthermore the decimator (Low Pass Filter and Downsampler) cancels additional unwanted noise. The decimator also increases the resolution of the ADC. More information of decimators is presented in the sigma delta section.

To find the SNR of the oversampled PCM converter let's reconsider what happens. From Figure 5.12 and taking the z-transform of the input, output and quantization error we have

$$Y(z) = X(z) + E(z) \quad (5.10)$$

The transfer function $H(z)$ of the input and the error are taken to be 1. To account for non ideal transfer function we rewrite equation 5.10 :

$$Y(z) = H_x(z)X(z) + H_E(z)E(z) \quad (5.11)$$

Next we calculate the power densities of $E(z)$ and $X(z)$. To do that we need to remind our self's that the input is always busy and with that we can conjunct the input with a stationary random process. In other words we model the output being equal to $P(f)|H(f)|^2$. In this case the densities become

$$\begin{aligned} P_{xy}(f) &= P_x(f)|H_x(f)|^2 \\ P_{ey}(f) &= P_e(f)|H_e(f)|^2 \end{aligned} \quad (5.12)$$

As mentioned we assume white noise then $P_e(f) = \frac{\sigma_\epsilon^2}{f_s}$. Also assuming ideal transfer functions $H(z)=1$ and cut-off frequency of the LPF to be ideal at f_{\max} then

$$\sigma_{ey}^2 = \int_{-f_{\max}}^{f_{\max}} P_{ey}(f)df = 2 \int_0^{f_{\max}} P_{ey}(f)df = 2 \int_0^{f_{\max}} \frac{\sigma_\epsilon^2}{f_s} df = \sigma_\epsilon^2 \left[\frac{2f_{\max}}{f_s} \right] \quad (5.13)$$

This result proves that the noise in the band of interest is less than in a Nyquist rate PCM. The noise is distributed over the whole range equally, see figure 5.13. Hence we can assume the noise power σ_x^2 does not change and equal that of an Nyquist rate PCM. The SNR is then

$$\begin{aligned} SNR_{dB} &= 10 \log \left(\frac{\sigma_x^2}{\sigma_{ey}^2} \right) = 10 \log(\sigma_x^2) - 10 \log(\sigma_\epsilon^2) - 10 \log \left(\frac{f_{oversampling}}{2f_B} \right) \Rightarrow \\ SNR_{dB} &= 10 \log(\sigma_x^2) - 10 \log(\sigma_\epsilon^2) - 3.01r = 6.02N + 1 - 76 + 10 \log OSR \end{aligned} \quad (5.14)$$

The term $\frac{f_{oversampling}}{2f_B}$ is called oversampling ratio (OSR) and setting it equal to 2^r (r is the OSR rate) we can show that doubling the OSR results in an increase of the SNR by 3dB or in an increase of half a bit. The oversampled PCM shows that with little more circuitry for controlling the signals in higher frequencies, higher resolution and a more relaxed requirement on the anti-aliasing filter are achieved. The disadvantage is that the digital filter must attenuate the out of band noise very effectively, so that when downsampling we do not downshift the noise too.

5.2.2 PWM converter

The PWM consists of an integrator and a comparator. The integrator produces a constant integration and discharge of a suitable arbitrary voltage. Hence the integrator output results in a triangular repeated waveform. The repetition interval equals the sampling rate. The comparator simply compares the input signal to the triangular waveform at all times. The outcome is a string of logical pulses (0 or 1) which represents the digital code. Of course the timing sets the digital word length. The integrator as well as the comparator is less insensitive to components mismatch, the opposite is true for the components of other Nyquist converters. This is PWM's main advantage.

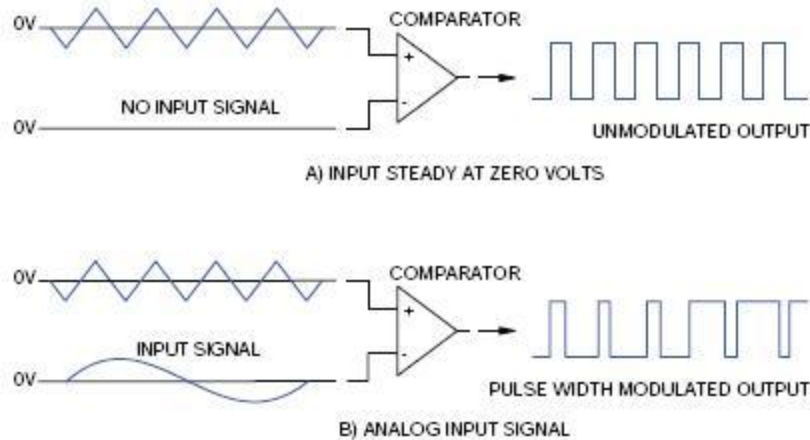


Figure 5.15. PWM analog modulation (a) Output of zero input signal (b) Output to an non-zero input

An example of PWM functionality is shown in figure 5.15 for zero input, and then the output is 50% of time high and 50% of the time low, taking the mean of the output pulses yields an overall output of zero volts. For a non-zero input the input, the input is constantly compared to the triangular wave and the output is held 1 if the input exceeds the triangular wave and 0 if not. Hence we got an output with its width modulated as time proceeds. The average of the output pulses is the input voltage. For instance the output is 50% on 6V and 50% at 0V which results in an input voltage of 3 V. The PWM codes every sample of the input with the comparator to a single value or to one bit. Hence for every sample PWM's comparator acts like 1-bit ADC. This one bit conversion method is also used in the sigma delta converter.

5.2.3 Modulation ADC's

First we introduce the close “cousins” of sigma delta or better to the foregoers from which sigma delta came to be. Delta sigma is that foregoer. Therefore an introduction into Delta Sigma is necessary. In the next sections we introduce the basic characteristics of sigma delta modulation and its parts. Finally we explain the decimator process which follows sigma delta modulation.

5.2.3.1 Delta modulation

Delta modulation quantizes the difference between two sequential samples, instead of parts of the input signal. The basic idea behind that is that two close samples do not change radically, hence their difference ($x(t) - \hat{x}(t)$) remains constant. Consequently all differences between samples are similar. These differences are the input signal to the delta modulator. Each difference (delta) is being quantized with an one bit ADC (comparator), which is the output $y(t)$. The output is a quantized version of the difference; hence by integrating it we get the original analog difference. This signal is then subtracted from the next difference. For this reason feeding the output back to the input

and subtracting it from the input (difference) can give a prediction of the future difference. From another perspective the system autocorrects itself or adapt itself to the new conditions. Shown in figure 5.16 (a) is a sinusoidal signal as input. The differences of the input signal become more and more positive as the signal rises from 0 to its peak value. Hence the quantized values follow the input.

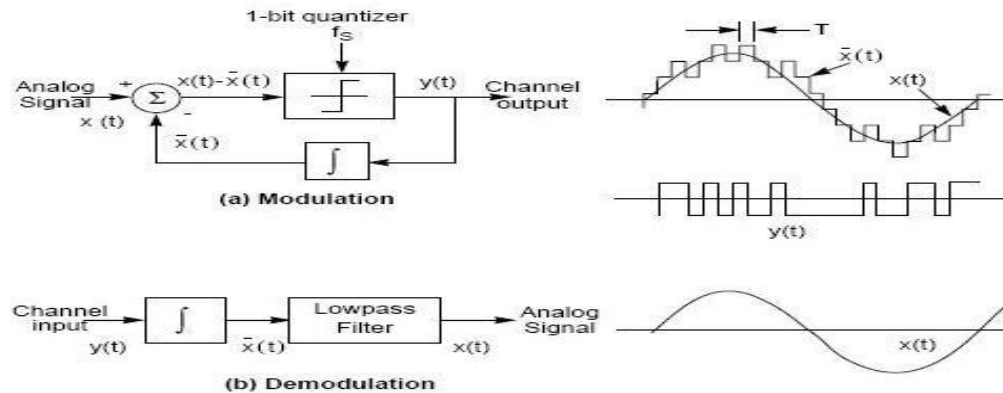


Figure 5.16. Delta modulation (a) Modulation (b) Demodulation

The demodulation process is nothing more to translate the quantized version back to their analog values and filter the result with a LPF to obtain the original input. Their disadvantage is that for rapid changing signals the, the output cannot follow the input. The reason for this is the non-adaptive step size, like mentioned it's constant. This is corrected in the DPCM version of delta modulation.

5.2.3.2 Sigma delta Modulation

Looking at the delta modulation we see that two integrators are necessary. The integration is a linear operation and therefore the modulation integrator can be moved before the summation node. Combining then the two integrators into one, yield the following structure:

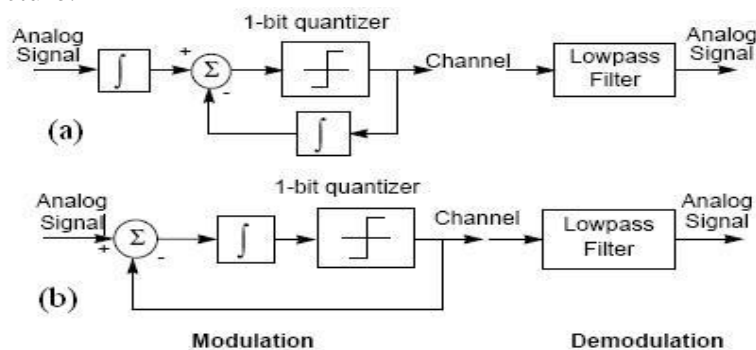


Figure 5.17. Linear integration operations allows to (a) move modulation integrator before sigma (b) combine two integrators into one

The origin of the name sigma delta should now be obvious; it's because of the placement of the integrator sigma (figure 5.17 (a)) before the delta modulation. The big advantage is now that quantization of an integrated signal is being done and therefore allows modulating of fast changing signals.

5.2.3.2.1 Sigma integrator (DAI)

DAI stands for discrete analog integration and its main feature is that it uses feedback to improve SNR of the ADC. How does feedback improve the resolution will be clear soon. For the time being recall (see equation 5.14) that increasing the OSR leads to an increase in bit resolution. The increase in OSR is also called averaging. The problem with that is large amount of oversampling (averaging) needed to achieve a certain resolution. For instance to have an increase in resolution of only 6-bits requires 4096 samples. For a 1MHz signal bandwidth we would need a sampling frequency of 8.192 GHz! Feedback lowers the OSR to achieve the same increase in resolution as without feedback. The circuit used to implement the DAI is shown in figure 5.18 (a).

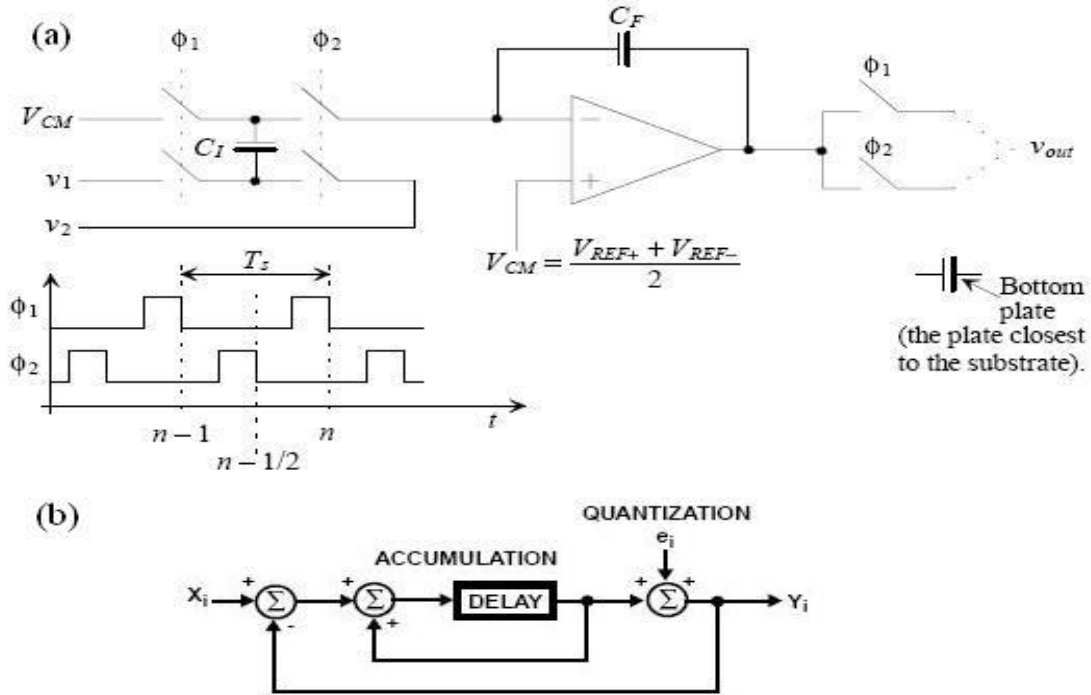


Figure 5.18. (a) Circuit implementation of the DAI. Logical block implementation of DAI (Accumulation block)

To avoid jitter and capacitor C_1 mismatch the circuit must be clocked with non-overlapping clocks ϕ_1 and ϕ_2 . To examine the function of the DAI let's start when ϕ_1 is high. The switches connected to ϕ_1 close at $t=n-1$, letting C_1 charge and the charge at the end of the "high" time is:

$$Q_1 = C_1(V_{CM} - v_1[(n-1)T_s]) \quad (5.15)$$

The output of the integrator is

$$V_{out}[(n-1)T_s] \quad (5.16)$$

When ϕ_2 is high the capacitor C_1 is again charged and the charge at the end of the "high" period ($t=n-1/2$) is

$$Q_2 = C_1(V_{CM} - v_2[(n-\frac{1}{2})T_s]) \quad (5.17)$$

Before we proceed let's have a quick review of inverting op-amps. An inverting op-amp consists of a feedback connection between the output and the negative (inverting)

terminal. In the feedback loop a device with impedance Z_1 exist and impedance Z_2 is connected to the negative terminal.

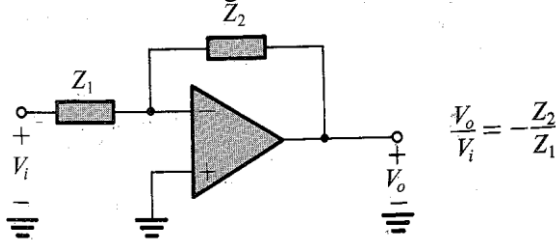


Figure 5.20. Inverting op-amp. Figure [3]

In that case the output is equal to $\frac{V_o}{V_i} = -\frac{Z_2}{Z_1}$. The transfer function is defined by the impedance ratio. Letting Z_1 be R and Z_2 be $1/sC$, then we get an RC integrator with $\frac{V_o}{V_i} = -\frac{1}{j\omega CR}$. The RC integrator is also called continuous-time

integrator. The output of the integrator is then $V_{out}(t) = V_C - \frac{1}{RC} \int_0^t u_1(t)$,

where $i_1 = \frac{u_1(t)}{R}$. The role of R in the integrator is to provide a current (V_R/R) to the capacitor in the negative feedback. If we replace the resistance R with a capacitance C_1 , and place switches between C_1 then we get a switched capacitor integrator. In relationship to the resistor, the equivalent current through the capacitor is $I_{AB} = \frac{C_s(V_A - V_B)}{f_{ck}^{-1}} = C_s(V_A - V_B)f_{clk}$.

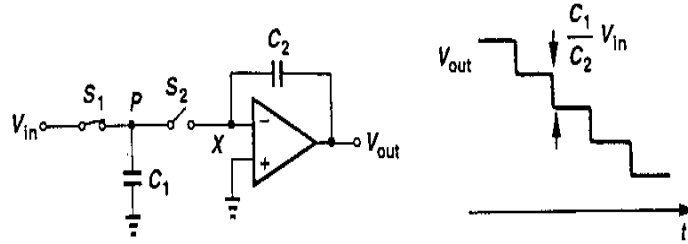


Figure 5.20. Discrete time integrator with non-inverting terminal to the ground. Figure from [4]

In every clock cycle C_1 absorbs a charge equal to $C_1 V_{in}$ when S_1 is on and transfers it to C_2 when S_2 is on. Node X is considered virtual ground. For a constant input voltage V_{in} , the output changes by $V_{in} C_1 / C_2$ every clock cycle see figure 5.19. The final value of V_{out} after every clock cycle is $V_{out}(kT_{clk}) = V_{out}[(k-1)T_{clk}] - V_{in}[(k-1)T_{clk}] \frac{C_1}{C_2}$. Returning to our DAI we see that the non-inverting terminal is set to V_{CM} . The difference between the charges $Q_2 - Q_1$ is transferred, like the charge C_1 in figure 5.20, to the feedback capacitor C_F . As a result an output voltage change occurs, similar that to figure 5.20. This change is written as

$$(V_{out}(nT_{clk}) - V_{out}[(n-1)T_{clk}])C_F = C_1 (V_1[(n-1)T_{clk}] - V_2[(n-\frac{1}{2})T_{clk}]) \quad (5.18)$$

Transforming equation 5.18 to the z-domain

$$V_{out}(z)(1 - z^{-1}) = \frac{C_1}{C_F} (V_1(z)z^{-1} - V_2(z) * z^{-1/2}) \quad (5.19)$$

The transfer function is then when ϕ_1 is connected as the output

$$V_{out}(z) = \frac{C_1}{C_F} \frac{(V_1(z)z^{-1} - V_2(z)z^{-\frac{1}{2}})}{(1-z^{-1})} \quad (5.20)$$

If we connect ϕ_2 as output then as seen in figure 5.18 (a) the output is delayed by $T_{clk}/2$ and therefore

$$Q_1 = C_1(V_{CM} - v_1[(n-1/2)T_{clk}]) \quad (5.21)$$

When ϕ_2 is high the capacitor C_1 is again charged and the charge at the end of the “high” period is

$$Q_2 = C_1(V_{CM} - v_2[nT_{clk}]) \quad (5.22)$$

$$(V_{out}(nT_{clk}) - V_{out}[(n-1)T_{clk}])C_F = C_1(V_1[(n-1/2)T_{clk}] - V_2[nT_{clk}]) \quad (5.23)$$

The transfer function is then when ϕ_2 is connected as the output

$$V_{out}(z) = \frac{C_1}{C_F} \frac{(V_1(z)z^{-1/2} - V_2(z))}{(1-z^{-1})} \quad (5.24)$$

With $V_2(z) = V_{CM}$ equation 5.24 becomes

$$H(z) = \frac{V_{out}(z)}{V_1(z)} \frac{C_1}{C_F} \frac{z^{-1/2}}{(1-z^{-1})} \quad (5.25)$$

The frequency response of $H(z)$ is found by substituting $z = e^{j2\pi\frac{f}{f_s}}$ then

$$H(z) = \frac{z}{z-1} = e^{j2\pi\frac{f}{f_s}} \frac{1}{e^{j2\pi\frac{f}{f_s}} - 1} = e^{j2\pi\frac{f}{f_s}} \frac{1}{-1 + \cos 2\pi\frac{f}{f_s} + j \sin 2\pi\frac{f}{f_s}}$$

From the relationship $\left| \frac{1}{a+jb} \right| = \frac{1}{\sqrt{a^2+b^2}}$ we get the magnitude

$$|H(f)| = \frac{1}{\sqrt{2(1-\cos 2\pi\frac{f}{f_s})}} \quad (5.26)$$

and the phase

$$\angle H(f) = 2\pi\frac{f}{f_s} - \left(\pi\frac{f}{f_s} + \frac{\pi}{2} \right) = 180\frac{f}{f_s} - 90 \text{ (degrees) for } 0 < f < f_s \quad (5.27)$$

Below is the graphical interpretation in the z-plane and frequency plane (s-plane) respectively. To understand the transformation relationship between z-transform and Laplace see appendix z-transform.

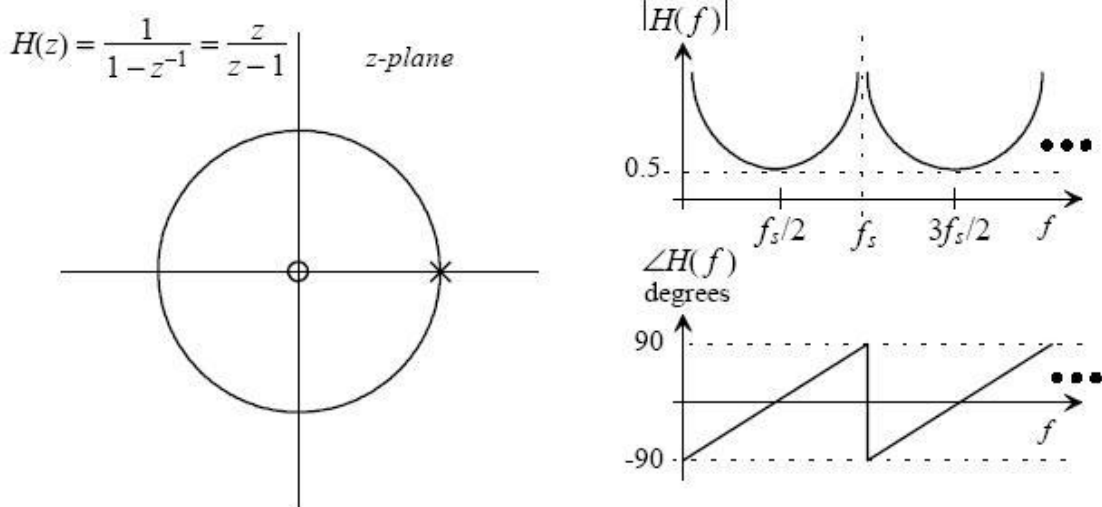


Figure 5.21. Graphical representation of DAI's transfer function. Figure from [2]

We can now establish a block diagram showing the discrete analog integrator functions.

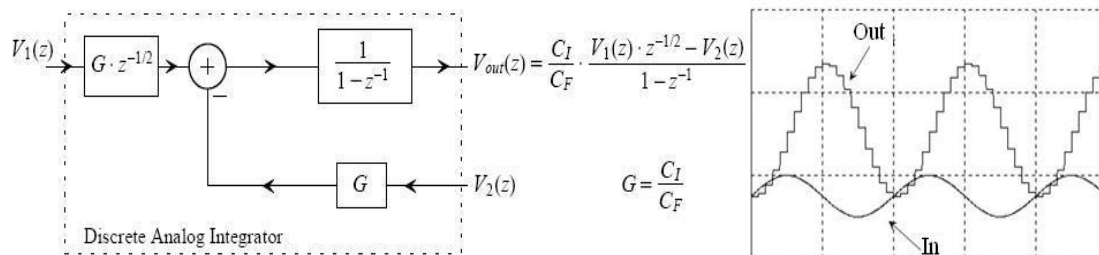


Figure 5.22. Block diagram of DAI's and example of output. Figure from [2]

From the figure above we can clearly see that when ϕ_2 switches or no switches are used, the input signals are delayed by half clock cycle ($z^{-1/2}$) and when ϕ_1 switches are used the input signals are delayed by a clock cycle (z^{-1}). The input signal V_1 is always enhanced by the gain set by the capacitor ratio C_I/C_F . Next we will examine why this topology is suited to be a modulator and we will calculate the resulting SNR.

5.2.3.2.2 Modulation and noise shaping

To derive any conclusions about the functionality of the modulator we look at the modulation block which includes the DAI block and the quantization block (comparator). First we redraw the block diagram including the previously found transfer function $H(z)$.

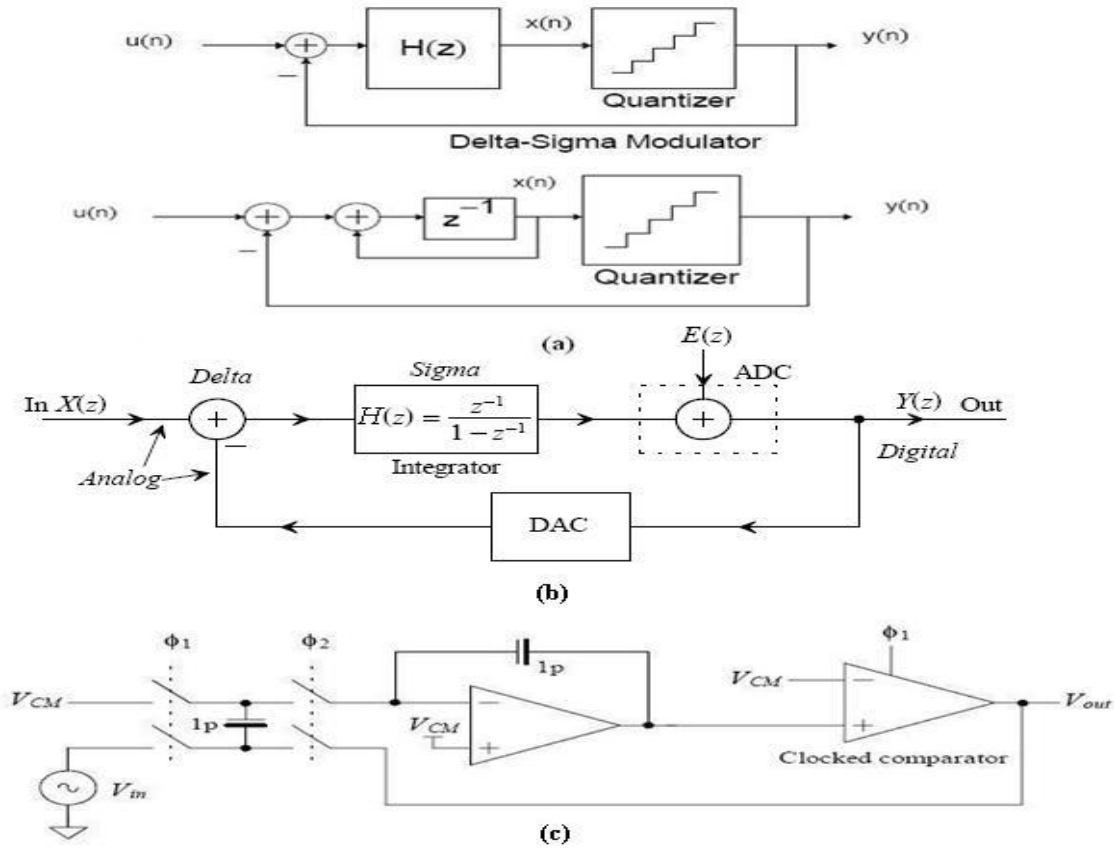


Figure 5.23. (a) Block diagram of the modulator. (b) Overview with modulator (sigma integrator) (b) Circuit implementation. Figures from [2]

Now we are ready to find the SNR, the first block diagram in figure 5.23 holds also for delta modulation. Here we see the importance of the integrator in both modulation types, if $H(z)$ is large, the quantization noise in the range of where $H(z)$ is used, usually at $(-f_{\max}, f_{\max})$, is almost zero. You can verify this with the following “block diagram” equations:

$$S_{TF}(z) = \frac{Y(z)}{U(z)} = \frac{H(z)}{1+H(z)} \quad (5.28)$$

$$N_{TF}(z) = \frac{Y(z)}{E(z)} = \frac{1}{1+H(z)} \quad (5.29)$$

$$Y(z) = S_{TF}(z)U(z) + E(z)N_{TF}(z) \quad (5.30)$$

Substituting in the equations above the transfer function found for the DAI $Y(z)$ is

$$S_{TF}(z) = \frac{Y(z)}{U(z)} = \frac{\frac{1}{z-1}}{1+\frac{1}{z-1}} = z^{-1} \quad (5.31)$$

$$N_{TF}(z) = \frac{Y(z)}{E(z)} = \frac{1}{1+\frac{1}{z-1}} = 1 - z^{-1} \quad (5.32)$$

From 5.30 we get

$$Y(z) = z^{-1}X(z) + (1 - z^{-1})E(z) \quad (5.33)$$

The meaning of $Y(z)$ can be clarified if we examine the term $(1 - z^{-1})$. Suppose to have transfer function $H'(z) = (1 - z^{-1})$. Hence

$$H(z) = \frac{Y(z)}{X(z)} = (1 - z^{-1}) \Rightarrow Y(z) = X(z) - z^{-1}X(z) \quad (5.34)$$

In the time domain it is written as

$$Y[nT_{clk}] = X[nT_{clk}] - X[(n-1)nT_{clk}] \quad (5.35)$$

The equation shows that the function of $H(z)$ is differentiator. A differentiator is a device which implements the derivative of signal. For instance suppose we have the signal $x(t) = \sin(2\pi ft) = \sin(\omega t)$. The derivative of the signal is $\omega \cos(\omega t)$. So the derivative of a sinewave is a cosine wave whose amplitude is proportional to the original $x(t)$ sinewave's frequency. The frequency magnitude response of an ideal differentiator's is a straight line increasing with frequency ω . The frequency magnitude response of non ideal differentiator looks more like a half-circle. Let's examine if equation 5.34 has such an frequency respond.

$$H(z) = \frac{Y(z)}{X(z)} = (1 - z^{-1}) = \frac{z-1}{z} \quad (5.36)$$

Therefore $|H(z)|$ and $\angle H(f)$ is

$$|H(f)| = \frac{1}{\sqrt{2(1 - \cos 2\pi \frac{f}{f_s})}} \text{ and } \angle H(f) = \frac{\pi}{2} - \pi \frac{f}{f_s} \text{ for } 0 < f < f_{\text{sampling}} \quad (5.37)$$

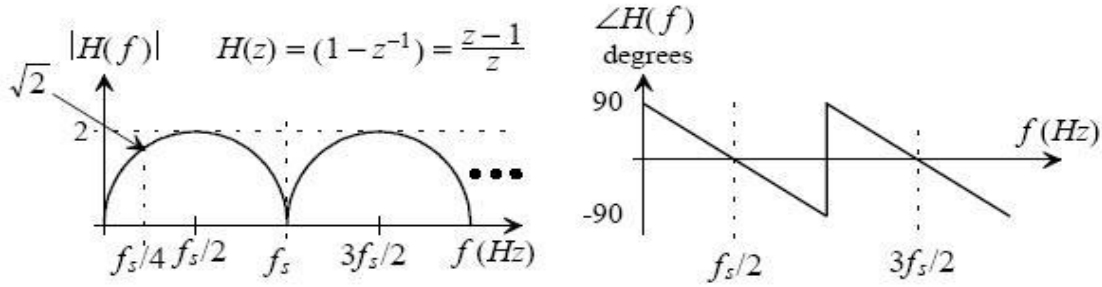


Figure 5.23. Frequency and magnitude response of differentiator. Figure from [2]

So coming back to equation 5.33 we conclude that the input signal simply passes through the modulator with a delay while the quantization noise is differentiated. To see the impact of the noise to the system we analyze the noise transfer equations, for that we concentrate on two equations 5.32 and 5.33.

$$(1 - z^{-1})E(z) = N_{TF}(z)V_{Qe}(f) = \left(1 - e^{-j2\pi \frac{f}{f_{\text{sampling}}}}\right) \frac{V_{LSB}}{\sqrt{12f_{\text{sampling}}}} \quad (5.38)$$

Where an analysis of $E(z) = V_{Qe}(f) = \frac{V_{LSB}}{\sqrt{12f_{\text{sampling}}}}$ is found in [2]. The PSD of the noise transfer $N_{TF}(z)$ is

$$|N_{TF}(z)|^2 |V_{Qe}(f)|^2 = 2 \left(1 - \cos(2\pi \frac{f}{f_{\text{sampling}}})\right) \frac{V_{LSB}^2}{12f_{\text{sampling}}} \quad (5.39)$$

5.35 is called the first order modulation noise. One must be aware that modulation noise is the new noise that is the outcome of the differentiated quantization noise. It is also an unwanted source added to the input signal and can have large magnitude, see figure 5.24. Luckily after passing through an LPF most of the noise is filtered out, and less noise then in oversampling ADC remain.

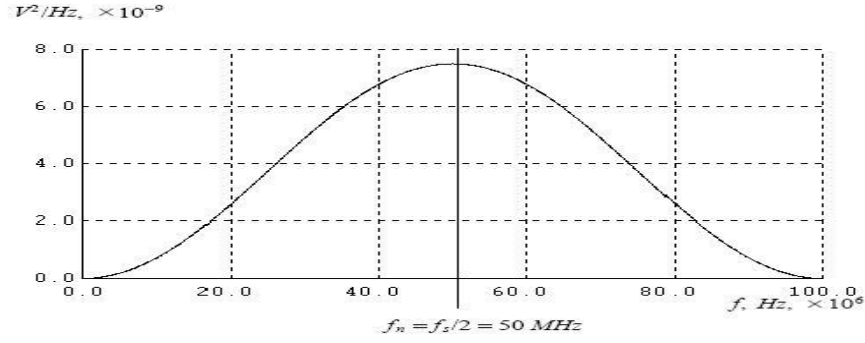


Figure 5.24. Modulation noise and modulation magnitude example. Figure from [2]

To find the SNR we find the RMS quantization noise below f_{sampling} , namely in the band of interest. Firstly we simply integrate over the band of interest the modulation noise using (5.39)

$$V_{Q,RMS}^2 = 2 \int_0^{f_{\text{max}}} |N_{TF}(z)|^2 |V_{Qe}(f)|^2 df = 2 \frac{V_{LSB}^2}{12f_{\text{sampling}}} 4 \int_0^{f_{\text{max}}} \sin^2 \pi \frac{f}{f_{\text{sampling}}} df \quad (5.40)$$

where $f_{\text{max}} = f_{\text{sampling}} / 2\text{OSR}$ and for small values in the sinusoidal term we get

$$\sigma_{ey}^2 = V_{Q,RMS}^2 = \frac{V_{LSB}^2}{\sqrt{12}f_{\text{sampling}}} \frac{\pi}{\sqrt{3}} \frac{1}{K^2} \quad (5.41)$$

Assuming peak input is a sinusoidal wave with a peak value of $2^N (\Delta/2)$ then the SNR becomes:

$$\sigma_x^2 = \left(\frac{\Delta 2^N}{2\sqrt{2}} \right)^2$$

$$SNR_{dB} = 10 \log \left(\frac{\sigma_x^2}{\sigma_{ey}^2} \right) = 10 \log \left(\frac{3}{2} 2^{2N} \right) - 20 \log \left(\frac{\pi}{\sqrt{3}} \right) + 20 \log \left(K^2 \right) \Rightarrow$$

$$SNR_{dB} = 6.02N + 1 - 76 - 5.17 + 30 \log(OSR) \quad (5.42)$$

As a result every doubling in the oversampling ratio OSR results in 1.5 bits increase in resolution or equivalently in an increase of the SNR by 9 dB. In the figure below we have a comparison between normal oversampling and modulation oversampling.

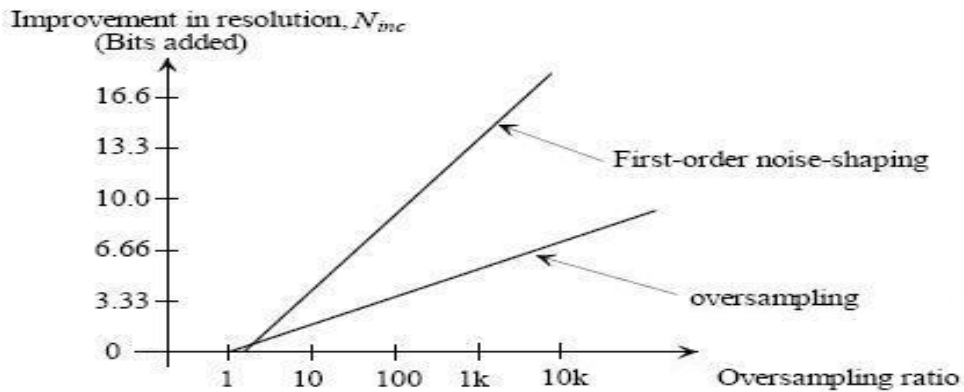


Figure 5.25. Comparison of oversampling and first order oversampling

Why it is called also first order oversampling will be clear in the next sections. To summarize the results, we can state that modulation with sigma delta results in increase of resolution with lesser OSR and also that the noise is pushed to higher frequencies, if the band of interest is not very large compared to the oversampling frequency. Hence the alternative name of sigma delta modulators: noise shapers.

5.2.3.2.3 Decimating and Filtering

Before we begin with decimation we define averaging which is the basic function of the decimator.

5.2.3.2.3.1 Averaging

We saw earlier that by oversampling the SNR is increased by half a bit, see equation 5.14. If we rewrite 5.14 we can express the increased resolution as

$$N_{inc} = \frac{10 \log OSR}{6.02} \quad (5.43)$$

We will show how the increase in resolution is achieved by averaging. To average means to calculate the mean of a sum. This is exactly what we will do next. Assume two ideal 8-bit ADC's like in figure 5.26 (a), instead of performing a division to find the mean, we sum the two outputs. This works by looking at the 8 higher bits in the sum or alternative to perform a shift right operation of the sum. Mathematically the output is

$$y(nT_{CLK}) = x(nT_{CLK}) + x[(n-1)T_{CLK}] \quad (5.44)$$

Care must be taken because if the input frequency is not the sampling frequency then many outputs value can average to zero. Imagine a sinusoidal input with frequency half of that of the sampling frequency, then the average of values from the period $[0, T/2]$ cancel out with the values of the period $[T/2, T]$, resulting in average of zero! The summation logic, or also called digital filter, in the z-domain is shown in 5.26 (b)

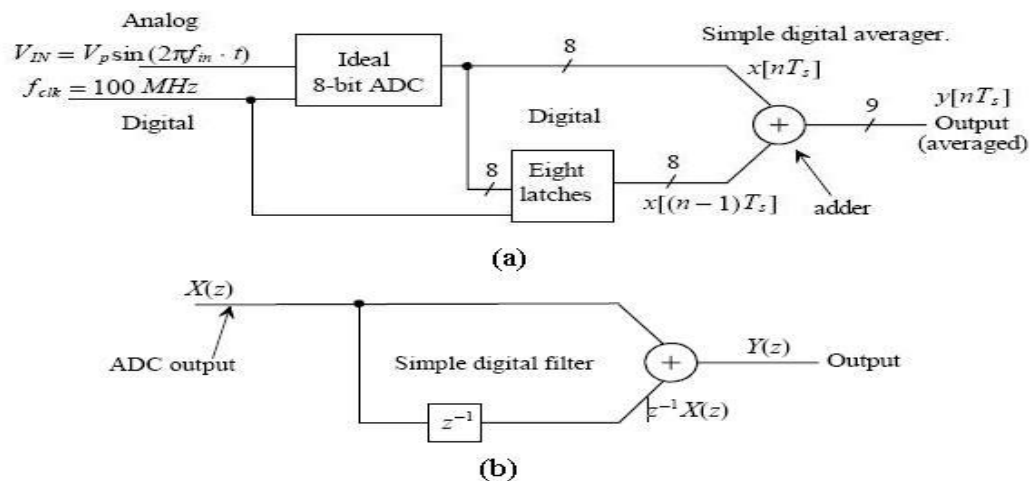


Figure 5.26. Implementing averaging (a) Block diagram of averaging (summation) (b) Equivalent of (a) in the z-domain.

The transfer function of the summation in the z-domain is found to be

$$H(z) = 1 + z^{-1}$$

which is similar to we encountered before in equation 5.36 except the sign of the phase response is now a mirrored version of figure 5.23. For the sake of clarity it is depicted in figure 5.27.

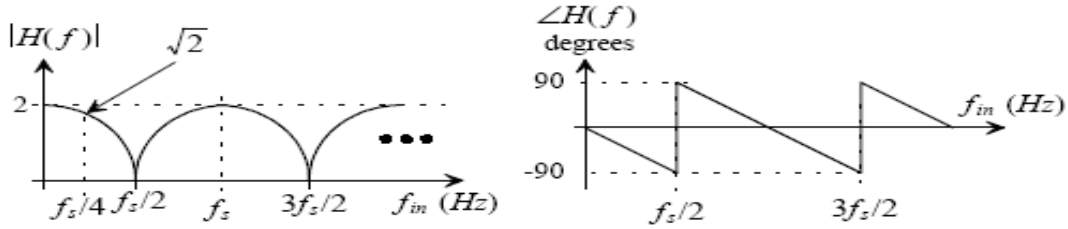
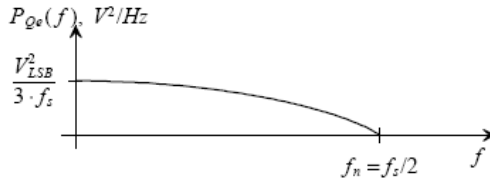


Figure 5.27. Magnitude and phase response of the z-domain averaging method (digital filter response). Figure from [2]

The simple digital filter has the following impact on the input:

- The conclusion of the analysis is that the summation does something remarkable to the input frequency. It attenuates them, remember the example for input frequencies half of the sampling frequency, here we can verify this by looking what happens to the magnitude when passing the summation (average) block. At $f_{\text{sampling}}/2$ we see zero magnitude!
- Another remark on the product of the noise PSD with the filter response is the attenuation of the power spectral density to zero at $f_{\text{sampling}}/2$.



- Increase in signal amplitude by the number of points averaged. Therefore we have an increase in the signals power and in turn an increase in SNR (less noise).
- To attenuate the signal we must confine the input frequency to less to $f_{\text{sampling}}/4$. The reason for this is that the magnitude at $f_{\text{sampling}}/4$ is $\sqrt{2}$. As known the RMS noise is found by dividing the power term of the signal by $\sqrt{2}$. Therefore at $f_{\text{sampling}}/4$ the input signal remains unchanged, while for input frequencies less than $f_{\text{sampling}}/4$ we benefit from averaging. If K is the number of points averaged then we can write the condition:

$$f_{\text{required}} = B = \frac{f_{\text{sampling}}}{K} = \frac{f_{\text{Nyquist}}}{K} \quad \text{with } f_{\text{in}} \leq B \quad (5.45)$$

Equation 5.45 should not be confused with the OSR, even the when we mean oversampling we mean averaging, but oversampling without averaging doesn't lead to any attenuation of the quantization noise.

5.2.3.2.3.2 Decimation

We saw in the last section that the input frequency must be less according to 5.45. The reduction of the frequency is what is called decimation. The word decimation can be

translated into “every tenth”. The goal of the decimation is to lowpass filter and downsample the input signal. Mathematically expressed this looks like

$$y[Ki * T_{clk}] = \sum_{n=K(i-1)}^{K(i-1)+K-1} \frac{x[nT_s]}{K} \quad (5.46)$$

and a more graphically illustration is:

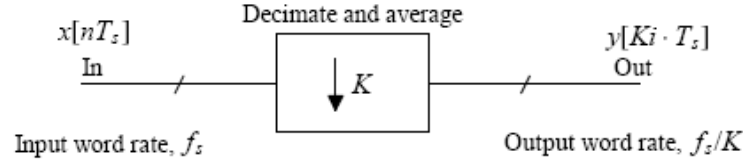


Figure 5.27. Decimation principle. Figure from [2]

Recognizing that the summation term is nothing more than averaging, we translate 5.46 into the z-domain:

$$H(z) = \frac{Y(z)}{X(z)} = \frac{1}{K} \sum_{n=0}^{K-1} z^{-n} = \frac{1}{K} (1 + z^{-1} + z^{-2} + \dots + z^{1-K}) \quad (5.47)$$

With arithmetic means we bring 5.47 in the form of

$$H(z) = \frac{1}{K} \frac{1 - z^{-K}}{1 - z^{-1}} \quad (5.48)$$

Setting $z = e^{j2\pi f_{sampling}}$ we can find the magnitude and frequency response of the transfer function.

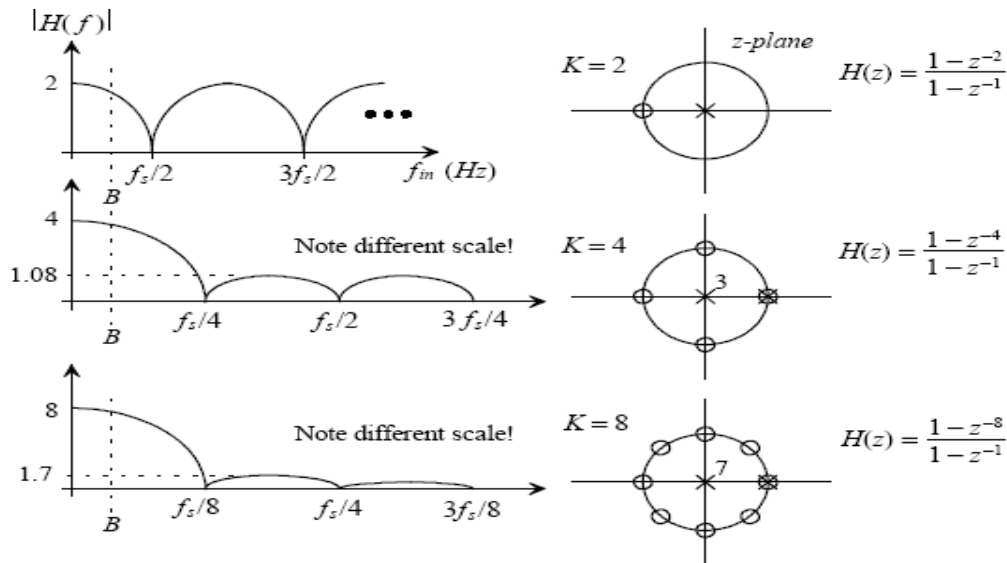
The magnitude response is:

$$|H(f)| = \frac{1}{K} \frac{1 - e^{jK2\pi \frac{f}{f_{sampling}}}}{1 - e^{j2\pi \frac{f}{f_{sampling}}}} = \frac{1 - e^{j2\pi f_{required}}}{1 - e^{j2\pi \frac{f}{f_{sampling}}}} \quad (5.49)$$

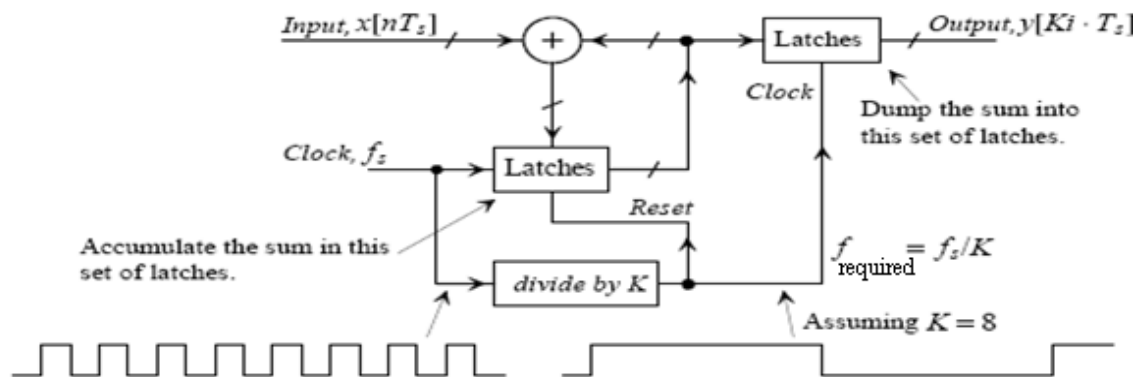
The frequency response is:

$$\begin{aligned} |H(f)| &= \frac{1}{K} \frac{\sqrt{2(1 - \cos K2\pi(f/f_{sampling}))}}{\sqrt{2(1 - \cos 2\pi(f/f_{sampling}))}} = \frac{1}{K} \frac{\sin(K\pi(f/f_{sampling}))}{\sin(\pi(f/f_{sampling}))} \\ &= \frac{\text{sinc}(K\pi(f/f_{sampling}))}{\text{sinc}(\pi(f/f_{sampling}))} \end{aligned} \quad (5.50)$$

The frequency response is a ratio of sinc functions and that's why this averaging filter is called a sinc filter. Again a graphically interpretation in the z-domain is shown in the figure below



A possible implementation of the transfer function $H(z)$ can be done by the accumulate and dumb circuit.



One can see the clearly from figure 5.27 that if the averaging number increases the attenuation increases. The main lobe is the first big lobe from 0 to $f_s/8$ and the second lobe (sidelobe) is the lobe from $f_s/8$ to $f_s/4$. Their ratio defines the attenuation capability of the decimation filter. For $K \geq 3$ the relationship

$$\left| \frac{\text{Main lobe}}{\text{side lobe}} \right| = K \sin \left(\frac{1-5\pi}{K} \right) \quad (5.51)$$

holds [2]. The drawback is to achieve good attenuation, cascades of many accumulate and dump filters must be implemented. To find the desired attenuation of the filter it is shown [2] that

$$\left| \frac{Main\ lobe}{side\ lobe} \right| = \left| K \sin \left(\frac{1-5\pi}{K} \right) \right|^L \approx 13L\ dB \text{ and the droop at } f_s/4 \text{ is } Droop = \left| K \sin \left(\frac{\pi}{2K} \right) \right|^{-L} \approx (-3.9)L\ dB \quad (5.52)$$

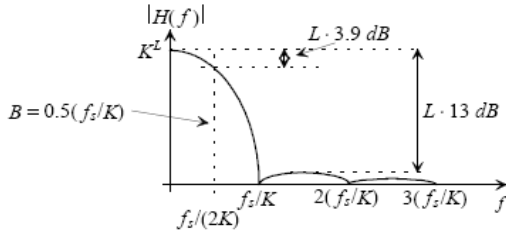


Figure 5.29. Amount of attenuation done.

To achieve even further attenuation we must increase the averaging number. The main problem with that is that the latches increases, see accumulate and dumb circuit. To decrease the amount of necessary circuitry needed, various methods have been developed, we present one of them next.

To reduce complexity a simple method is used: the splitting of the transfer function in two parts:

$$H(z) = \left[\frac{1-z^{-k}}{1-z^{-1}} \right]^L = \overbrace{\left(\frac{1-z^{-k}}{1-z^{-1}} \right)^L}^{L \text{ differentiators}} \overbrace{\left(\frac{1}{1-z^{-1}} \right)^L}^{L \text{ integrators}} \quad (5.52)$$

The first part is the $(1-z^{-k})^L$ and the second part is the $\left(\frac{1}{1-z^{-1}} \right)^L$. Analyzing the transfer function of these parts leads to the following observation:

- a) $H(z) = (1-z^{-k})^L$ is a digital differentiator, see equation 5.36 and the associated discussion. Rearranging the terms we can write $H(z) = 1 - z^{-k} = \frac{z^k - 1}{z^k}$. This filter is called a comb filter, an example out of the many existing finite impulse (FIR) response digital filters. Its main characteristic is that applying a unit amplitude impulse to the comb filter causes the output of the comb to be one at the time of appliance and zero at other times.

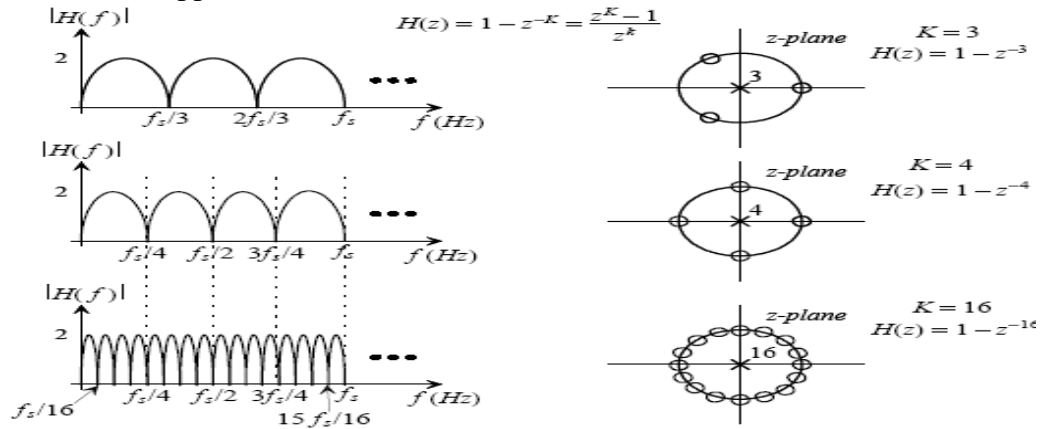


Figure 5.30. Comb filter frequency response. Figure from [2]

- b) $H(z) = \left(\frac{1}{1-z^{-1}} \right)^L$ is a digital integrator, see equation 5.26 and the associated discussion. This filter is an example of infinite impulse response (IIR) digital filter. A unit impulse at the input forces the output of the filter to be one at all times.

The comb filter's demands a clock frequency equal K, see for instance Figure 5.30 the case when K=16. Because the integrator block precedes the comb filter block, the output clock frequency of the integrator clock must be K less if comb filter is to be driven. This in turn reduces the registers used in the comb filters to one. The result is less complexity and easier implementation. One problem that might occur is the aliasing of the lobes.

Some frequency components could “survive” the decimation process (side-lobes). To avoid this we can resample the output of the first decimation process with a smaller frequency. Assume downsampling to Nyquist rate is achieved in the first decimation process to $f_s/K=2B$ ($K=8$). Still some lobes are not eliminated between $f_s/8$ and B . Resampling with $2*(f_s/K)$ and passing it through a half-band digital filter eliminates the remaining lobes. The next figures clarify what is meant by side-lobes aliasing and by resampling the output:

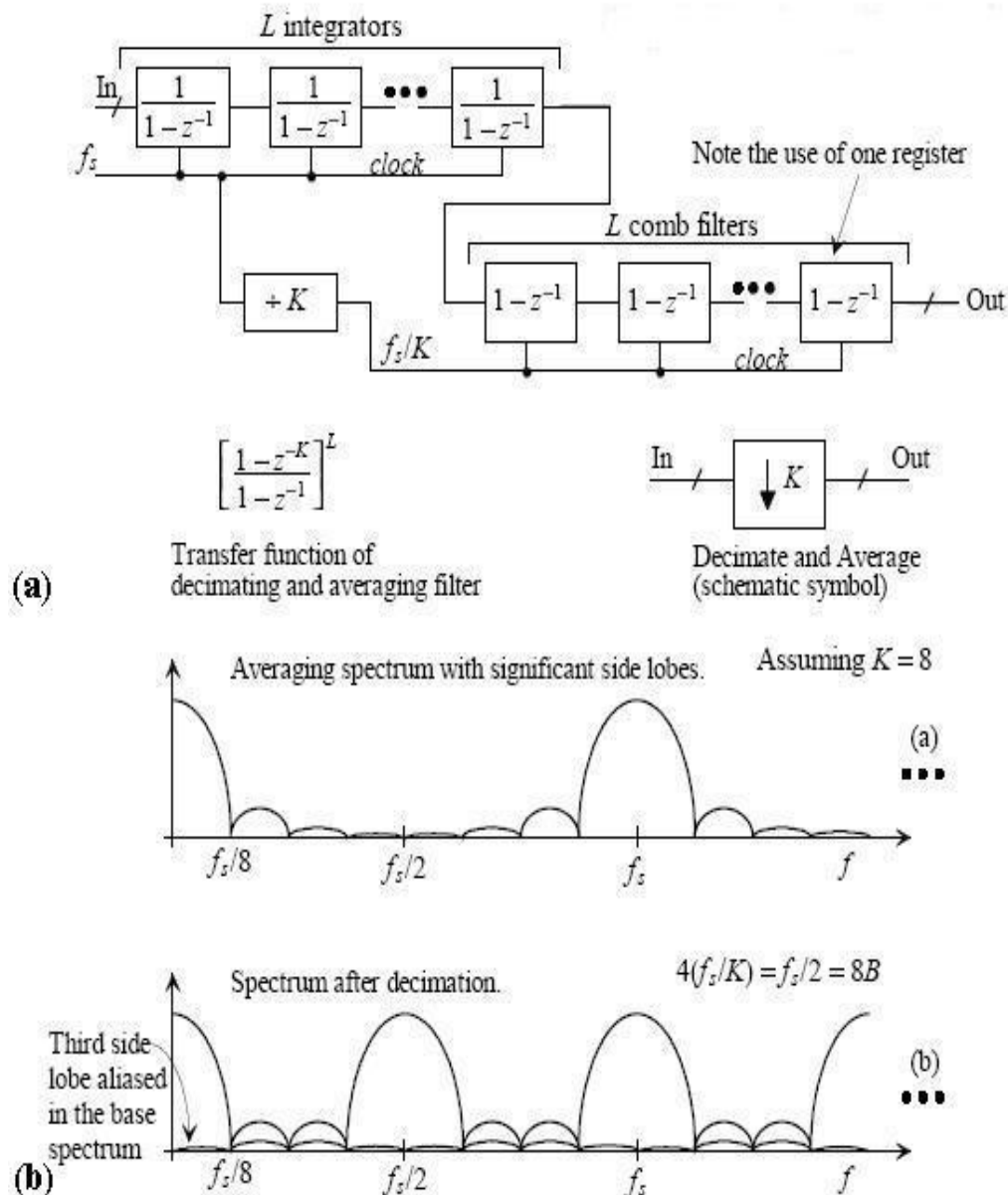


Figure 5.31. (a) Filter implementation with cascaded integrators and differentiators (b) Aliasing in the decimation process. Figures from [2]

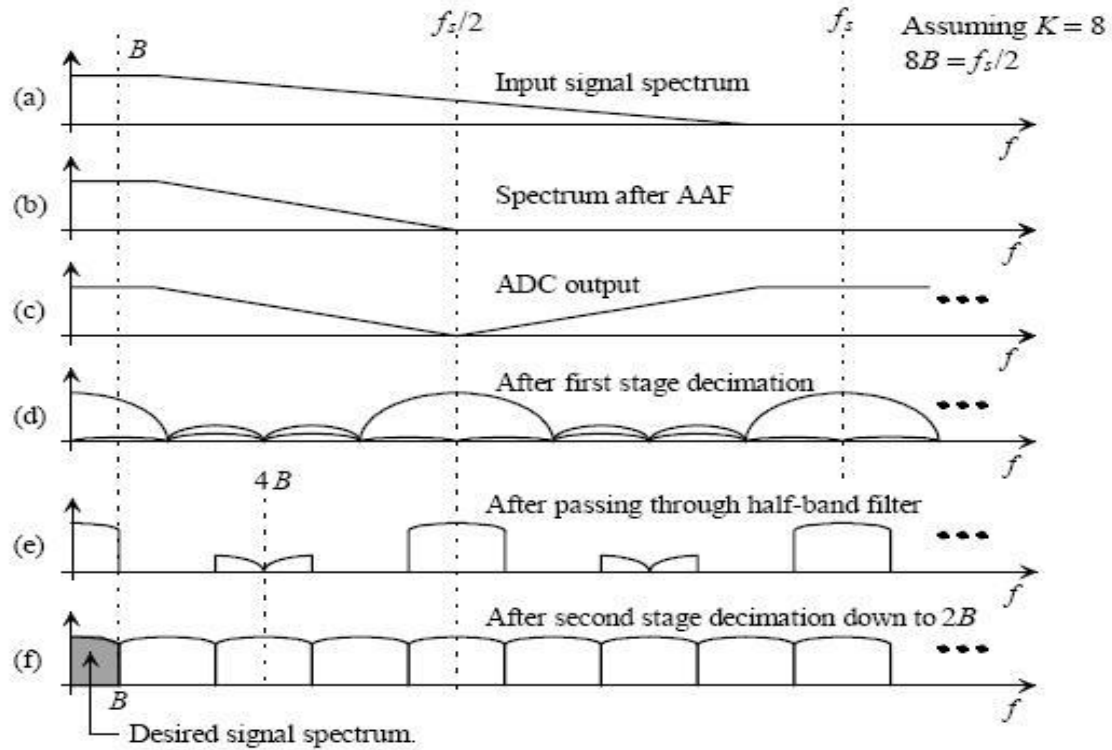
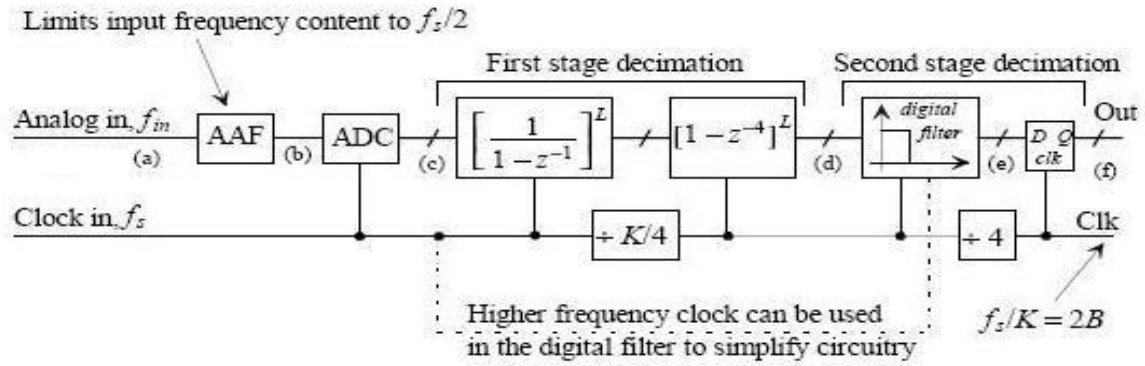


Figure 5.31. Overview of the decimation process. Figures from [2]

5.2.3.2.3.3 Sigma delta decimator

We turn now to the decimator needed for the sigma delta modulator. From the equation 5.42 we see the SNR increase of the modulator. Rearranging the terms yields:

$$N_{inc} = \frac{30 \log OSR - 5.17}{6.02} \quad (5.53)$$

As we mentioned the modulator uses a one bit DAC for its linearity. If for example a 4-bit DAC is used the modulator is called multibit modulator. In the case of a 1-bit ADC the increased resolution would be $N_{inc} + 1$. It has been shown [2] that the word size increases by $\log_2 K$ in each integrator stage. Hence

$$L \log_2 K \geq \frac{30 \log K - 5.17}{6.02} \rightarrow L = 1 + M \quad (5.54)$$

where M is the order of the modulator. For the first order 1-bit sigma delta modulator L=2 is been used. Therefore the transfer function is

$$H(z) = \left[\frac{1}{K} \frac{1-z^{-K}}{1-z^{-1}} \right]^2 \quad (5.55)$$

An example of a modulator with a decimation filter is shown in the below figure. The sinc filter is clocked at 100Mhz/16 and the transfer function are derived from the equations 5.55 and 5.50.

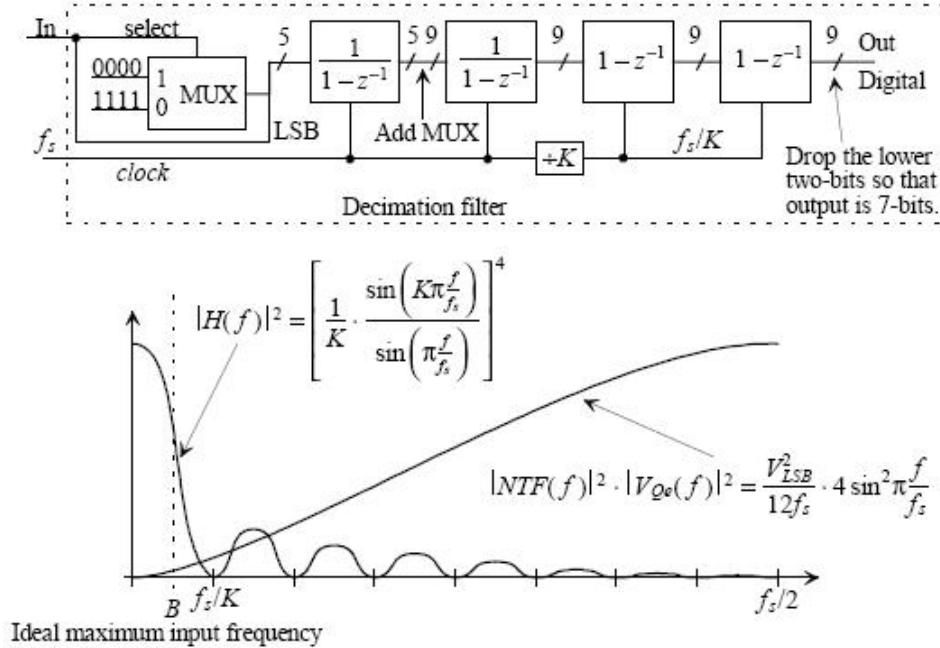


Figure 5.32. Example of decimation of the modulator output. Figure from [2]

The SNR with a sinc filter can be found by finding the RMS quantization noise in the band of interest, according to [2] this is

$$SNR_{dB} = 6.02N + 1 - 76 - 3.01 + 30\log(OSR) \quad (5.55)$$

We see only an increase of 2.16 dB compared to the ideal SNR from equation 5.42. Note that we neglected the aliasing that can occur in the decimation process, which degrade the SNR.

5.2.3.2.3.4 Overview of the sigma delta modulation and modulator

For reasons of clarity the complex system is presented in more qualitative manner. The following figure shows the overall process that exists in the sigma delta.

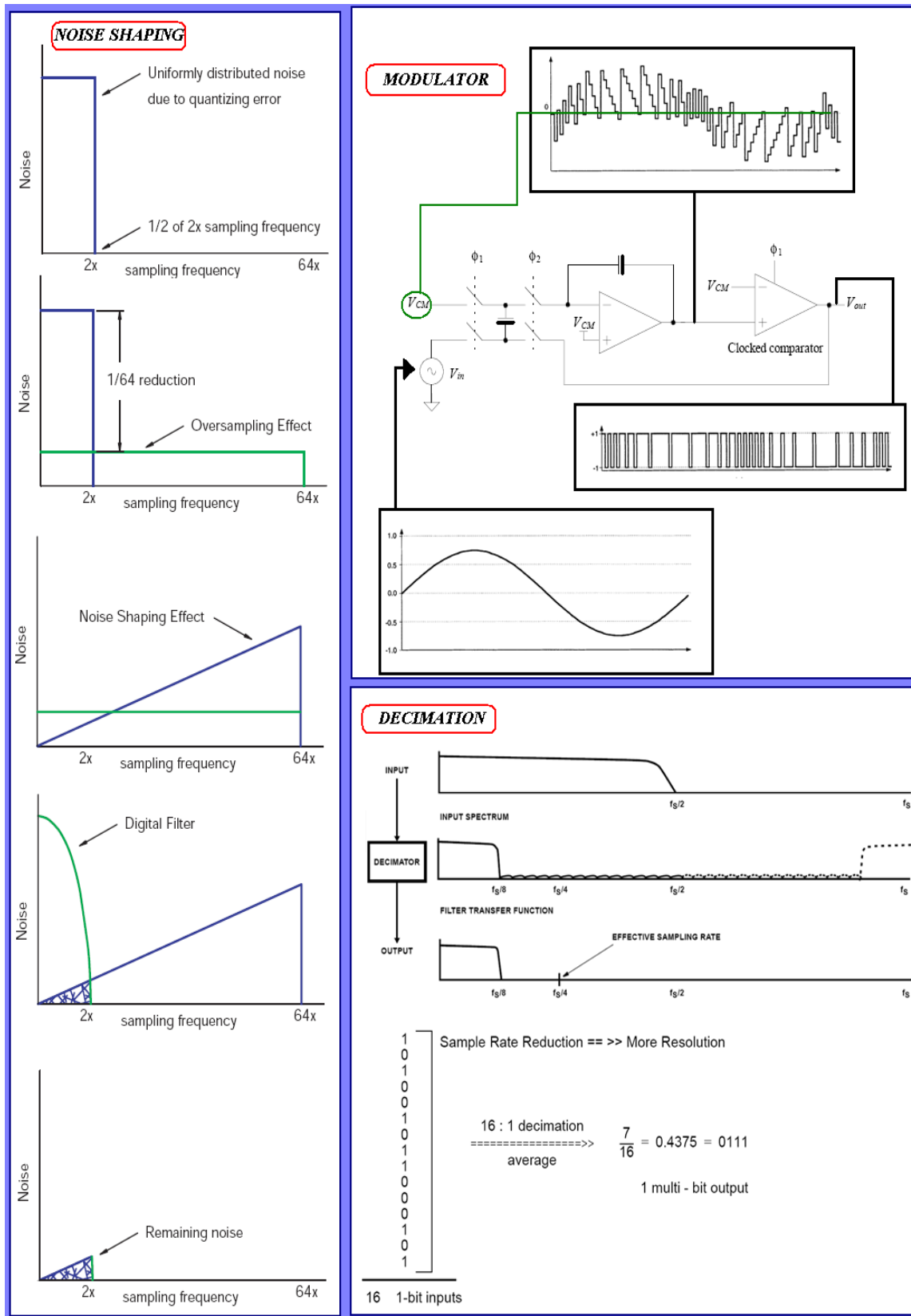


Figure 5.33. Sigma delta overview.

5.3 References

- [1] Baker, R Jacob., *CMOS circuit Design, Layout and Simulation*. s.l. : IEEE Press, Wiley Interscience, 2005. Vol. Second Edition. 0-471-70055-X.
- [2] Baker, R Jacob., *CMOS Mixed-Signal Circuit Design*. s.l. : IEEE Press, 2002. 0-471-22754-4.
- [3] Sedra, Adel S and Smith, Kenneth C., *Microelectronic Circuits*. New York : OXFORD UNIVERSITY PRESS, 2004.
- [4] Razavi, Behzad., *Design of Analog CMOS Integrated Circuits*. s.l. : McGraw Hill, 2001- 0-07-238032-2.
- [5] Sangil, Park., "Principles of Sigma-Delta modulation for analog-to digital converters." *Motorola Digital Signal Processors*. s.l. : Motorola. Vol. APR/D, 1
- [6] Sansen, Willy M C., *Analog design essentials*. s.l. : Springer, 2006. 100-387-25746-2.
- [7] Aziz, Pervez M, Sorensen, Henrik K and Van der Spiegel, Jan., "An Overview of Sigma-Delta Converters How a 1 bit ADC achieves more then 16 bit resolution." *Signal Processing Magazine*, s.l. : IEEE, September 1996, Issue 1, Vol. 13, pp. 61-84.
- [8] Bohn, Dennis., "Digital Dharma of Audio A/D Converters." s.l. : Rane Corporation, 2003. Vol. RaneNote, 137. 09948.
- [9] Norsworthy, R Steven, Schreier Richard and Temes Gabor C., *Delta-Sigma Data Converters*. s.l. : IEEE Press, 1997. 0-7803-1045-4.

Chapter 6. Dental radiography with CMOS APS

Since 1895 when x-rays were discovered, film has been the primary medium for capturing, displaying and storing radiographic images. Amorphous silicon or CCD arrays covered with scintillator are nowadays widely used as detectors in medical imaging. The steadily increasing demand for more integration of circuits on a chip is pushing CCD and Amorphous silicon out of bounds, and the reason for this is because they are difficult to integrate, not to mention the associated cost of integration. Besides that their main advantage is that they can be constructed for large area of imaging, which is a necessity. CMOS technology slowly catches up to this requirement, with more smaller technology advantages over the years, CMOS imagers may be also used in large array design.

6.1 Medical imaging

We have seen in Appendix A Figure A.9 that the radiation can be separated into two types ionizing and non-ionizing radiation. The reason why a radiation is called ionizing is because when it interacts with an atom, it can remove tightly bound electrons from their orbits with the result that the atom becomes charged or also called ionized. Gamma rays and beta particles are examples of ionizing radiation. Non-ionizing radiation on the other hand does not have enough energy to “knock out” an electron out of the atom. Microwaves and visible light are examples of non-ionizing radiation. They have just enough energy to cause electrons via photon absorption to rise to higher orbits. This phenomenon is also called excitation. We didn’t analyze the impact on atoms that radiation might have, this was intently sealed away such that it can be brought together with the subject of medical imaging. To say more about the latter we begin with the x-ray creation. x-rays are created within by an electrode pair setup, usually a cathode and an anode. The latter is placed inside a glass vacuum tube. The cathode is a wire that is been

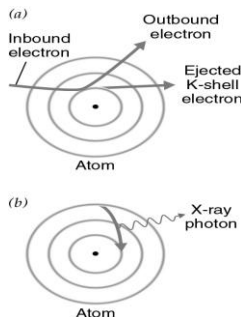


Figure 6.1
Characteristic x-ray

heated and the current that passes through it, heats it up so much that electrons are extracted. These then are being attracted by the positively charged anode which is in the form of a flat disc made off tungsten. Therefore the electrons travel in the tube at speed, defined by the voltage difference between the cathode and anode. The goal is to achieve high speed so that when the extracted electrons collide with the tungsten electrons, electrons from the lower tungsten orbital are pushed out of the atom. What happens next is that electrons from higher orbitals replenish the knocked out electrons. Hence electrons falling from the conduction band to the valence band release photons with high energy levels, these are the so called x-rays or

known as fluorescence x-ray. Another way of x-ray creation is known as Bremsstrahlung (brake emitting). In this situation a high speed electron approaches and penetrates the atom. The atom's nucleus may attract the electron just enough to alter its course. Like a comet whipping around the sun, the electron slows down and changes direction as it speeds past the atom. This "braking" action causes the electron to emit excess energy in the form of an

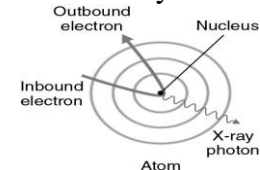


Figure 6.2 Bremsstrahlung
x-ray

x-ray photon. In the creation process of x-rays, some x-rays are emitted via Bremsstrahlung and some as characteristic x-rays. Luckily the tube is designed in that way that low energy photons are filtered away. Usually a slit in the tube is made in a certain way that only high energy particles can go through it. The result is an x-ray machine capable for medical imaging. x-ray imaging has a wide spectrum of application, where either one dimensional (1D) two dimensional (2D) or either three-dimensional (3D) imaging is used. In radiation imaging the penetration power of high energy radiation enables deep level imaging of objects rather than only surface imaging. 1D imaging covers all spectroscopy and profiling applications. 2D imaging captures constructs images with arrays of pixel sensors or other scanning methods. 3D imaging could be obtained by various methods including image processing of 2D captures.

A medical image is achieved by providing spatial mapping of some parameter, feature, or process within a biological entity. The image is constructed by attaining information about the observed object. This can be done in two ways: One being transmission and the other being the emission of information. The first method is the basic known setup, x-rays source, object and a detector. In the emission case the source emits inside body. An example of x-ray imaging is the gamma-ray imaging, where a gamma-ray isotope is injected in negligible amounts into the patient. The injected isotope is chosen according to metabolism of the organ under study. The goal is to acquire information about the spatial distribution of the source within the body. This can be done by tracking the radioactive change in the tissues. The number of photons detected in a detector is proportional to a weighted integral of the activity contained. Two variations of imaging exist:

- a) PET (positron emission tomography) is based on the detection of back-to-back 512 keV photons from the annihilation of positrons (emitted by the isotope injected into the patient) with electrons in the neighboring tissue.
- b) SPECT (single photon emission computed tomography) uses the radioactive emission pattern of the injected into the patient's body isotope to estimate its distribution. The gamma-rays produced are monochromatic emissions and their energies lay between a range of 60 keV and 511 keV.

Transmission emissions are x-ray images which are based on x-ray attenuation by the human body. The image is constructed by the absorption of parts of the body with different densities. Bremsstrahlung x-rays in transmission emissions are filtered because of their continuous energy distribution, while characteristic x-rays have discrete energy distribution, which are easier detected. The energies used are about 10 keV for mammography and 10-70 keV for dental and chest radiography. As one might guess the energy used in x-ray imaging is always chosen in relation to the type and structure of the object to be imaged. An example of transmission emission imaging is computed tomography CT. It consists of an arc of individual x-ray detectors that monitor the transmitted x-ray flux. The entire system rotates around the patient and continuously acquires data.

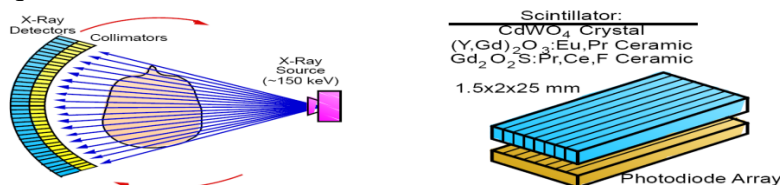
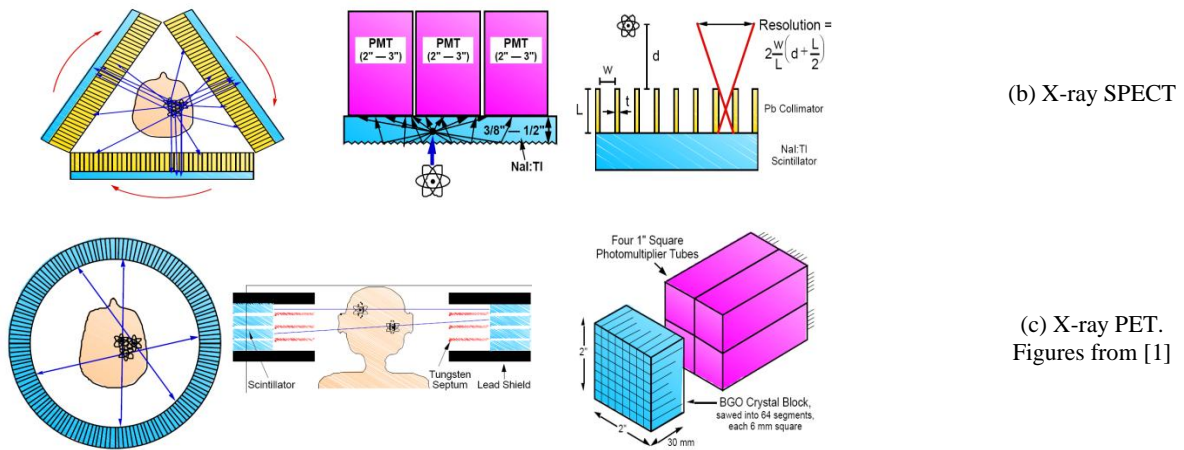


Figure 6.2
(a) X-ray CT



Without going into further details about the method used in CT imaging we focus on the x-ray detectors that all systems described so far use. Two methods distinguish x-ray detection, direct and indirect detection. In direct X-ray detection, the detector converts the absorbed x-ray directly into a charge signal. For energies below 100 keV, this is possible if the detection medium has good absorption efficiency. Further increase in detection can be done if the atomic number (Z) of the material used is high. Film, which is primary used in planar x-ray imaging, as well as silicon have low absorption efficiency when dealing with energies above 20 keV. Furthermore silicon is sensitive to radiation, it could heavily ionize the substrate and as a result damage the integrated devices. For that reason an intermediate medium is used, this kind of acquiring a x-ray image is called indirect detection. The medium is usually scintillator. One might notice from figures 6.3 that scintillators are used in every imaging system mentioned here. Hence it is necessary to introduce the reader to the scintillator material.

6.1.1 Scintillator in medical imaging

Scintillators are materials that absorb high energy particles and convert the atomic energy interaction into visible light. The materials of which they are made can be organic or inorganic solids, liquids and gases. Inorganic solids are mostly used in medical imaging systems and therefore the discussion is confined to inorganic scintillators. The physical function of inorganic scintillators are complex, it includes relaxation and initial electron excitation, thermalization and trapping of electrons and holes and excitation of the luminescent center. Many radiative emissions exist in scintillators, when high energy photons strikes the scintillator atom, a inner shell hole and an energetic primary electron result, followed by radiative decay and non-radiative decay. Non radiative decay can be Auger recombination or inelastic electron-electron scattering. These phenomena's last from 10^{-15} to 10^{-13} seconds. After the collision electron energies fall below the ionizing threshold. Then the energized electrons and holes distribute themselves by intraband transitions, to the ground levels and top levels of the conduction and valence band respectively. The charge carrier becomes trapped on defects and impurities, by the crystal lattice relaxation or by Coulomb attraction. The latter is responsible of the formation of impurity-bound excitons. To see what the reason for excitation from luminescent center

is we must look at the nature of the material of which scintillators are made. There are two types of material:

- Activated materials (activators) have a well-defined impurity ion injected into the lattice. The luminescent center therefore is made up of these impurities. Concentrations of this kind of impurity are in the order of 0.1%. For instance a material with activator is NaI:TI.
- Stoichiometric materials in which the concept of an identifiable luminescent center is no longer applicable. In this kind of the property, luminescence could be due the lattice itself or due to a constituent of the lattice, which however cannot be regarded in isolation, (immersed in a host medium). Examples of materials are $\text{Bi}_4\text{Ge}_{30-2}$ (BGO), CeF_3 , PbWO_4 and BaF_2 .

A clear distinctive separation between the presented materials is nowadays not possible since the quest for better scintillators demands sometime the coexistence of several of types of materials. An example of mixed materials is the crystal $\text{Ce}_x\text{La}_{1-x}\text{F}_3$.

Inorganic scintillators contain activators and hence we will focus only on activators and their efficiency of excitation of luminescent. Several factors seen in the second chapter like the quantum efficiency are used here to determine the efficiency of activated scintillators. We present here the basic factors that composite the luminescent efficiency.

- Quantum efficiency η_q (see chapter 2)
- Energy efficiency η_e , which is a measure of energy lost in the process
- Conversion efficiency β is partial quantum efficiency that describes the number of electrons and holes formed upon absorption of the x-ray.
- Transfer efficiency S is the probability that the excited electrons and holes reach the activator material
- Luminescent quantum yield Q is an efficiency measure of the photon created by the excited luminescent center

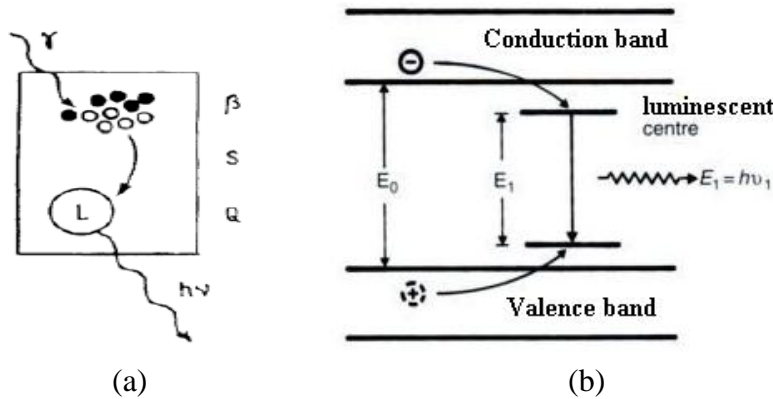


Figure 6.4 (a) Scintillation principle and the associated parameters (b) Introducing luminescent centers with activators. Figure (a) and (b) from [2] and [3] respectively.

Having all necessary factors we can define the following, as explained in [2]:

$$\eta_Q = \beta S Q \text{ and } 0 \leq \eta_Q, \beta, S, Q \leq 1 \quad (6.1)$$

Every factor is independent and therefore every factor can be analyzed separately. Suppose the minimum energy to create an electron-hole pair is ξ_{\min} , then for every $\xi > \xi_{\min}$ we can write

$$\beta = \frac{\xi_{min}}{\xi} \quad (6.2)$$

The strategy is to calculate first ξ_{min} and then the remaining energy factor, due to energy losses, ξ . A theory for calculation of ξ_{min} is done by Robbins. He showed that $\xi_{min} = \chi E_g$, where χ in Robbins assumption is the electron affinity of the material. To calculate ξ , Robbin adapted a random statistical model to account for the energy losses. The model defines a parameter K as

$$K = \frac{\text{rate of energy loss of optical photons}}{\text{rate of energy loss by ionization}} \approx 0.244 * 10^4 \left(\frac{1}{e_{\infty}} - \frac{1}{e_0} \right) \left(\frac{\hbar \omega_{LO}}{1 - 5E_g} \right)^{3/2} \quad (6.3)$$

where e_{∞} and e_0 are the high frequency and static dielectric constant and ω_{LO} is the longitudinal optical photon frequency. ξ becomes then

$$\xi = E_i (1 + K + 2L_f) \quad (6.4)$$

where E_i is the threshold energy. Assuming $\chi = 2.3$ like Robbin did, and with ξ and ξ_{min} defined as described above, β can be immediately defined from (6.2) as

$$\beta = \frac{1}{1 + 0.65K} \quad (6.5)$$

With (7.5) the quantum efficiency η_q and the energy efficiency are approximated as (see [2])

$$\eta_q = \frac{1}{1 + 0.65K} \quad (6.6)$$

$$\eta_e = \frac{1}{2.3} \frac{h\nu}{E_g} \eta_q \quad (6.7)$$

Equation (6.7) tells us immediately that if the emission energy is equal to the bandgap and the quantum efficiency is unity, the maximum energy efficiency of a scintillator is 1/2.3 or 43%.

The following table shows the parameters for common scintillator materials

	(a) $h\nu/2.3E_g$	(b) η_q	(c) η_e
Cs:TI	0,155	0,96	0,15
CsI:pure	0,27	0,029	0,007
NaI:TI	0,22	0,52	0,11
BGO	0,22	0,09	0,02
CaF ₂ :Eu	0,10	0,67	0,067
CeF ₃	0,17	0,077	0,013
CeP ₃ O ₁₄	0,19	0,08	0,015
LuPO	0,17	0,34	0,06

Table 6.1 Scintillator factors (parameters) that determines the transfer rate S

From Table 6.1 we conclude that the when E_g is 2.3 (a) has maximum values which ranges between 0.13 and 0.27. An average value to expect for a good scintillator would be about 20. Of course 20 is an ideal value and usually lower due to low luminescence yield, poor conversation or poor transfer. Transfer is the most important parameter when dealing with real scintillator materials because little is known about it. In activated materials the dopant concentrations is low and most of the absorbed energy is stored in the lattice. Therefore the transfer S is solely the transfer from the lattice to the luminescent center. All electron-hole pairs, excitons, mobile or self-trapped charges can

by heat acquire energy and “jump” to the conduction band, see Figure 6.1, and thereby contribute to the luminescent center excitation. The problem is that the lay down a distance till the find a luminescent center and as a result many electrons can vanish by other means and this decreases the transfer S. Electrons can reach a luminescent center by

- Radiative transfer
- Sequential charge trapping
- Excitonic transfer

Most Scintillators use trapping as the main mechanism of transfer but excitonic transfer can contribute a lot to transfer. Excitonic transfer occurs by the formation of localized or delocalized excitons. The self trapped excitons (STE) are transported to the activator. The “transporter” is a Forster type, a nonradiative emission that depends on the distance of the emitter and the absorber. Hence a local STE can contribute with high gain, while a delocalized STE has minor contribution. Since one only can guess the distribution of STE in the crystal, the transfer S cannot be well measured. To derive the transfer S we refer it to the light output of the scintillator. The number of photons produced by the scintillator depends on the number of electron-hole pairs generated when a photon with energy E_γ is absorbed.

$$n_p = n_{e-h}SQ = \frac{E_\gamma}{2.3E_g} \beta SQ \quad (6.8)$$

Then the light output L (photons per MeV) can be written as

$$L = \frac{n_p}{E_\gamma} = \frac{10^6}{2.3E_g} \beta SQ \quad (6.9)$$

The luminescence quantum yield (Q) is in general a well-known quantity, particularly for the case of weak electron-lattice coupling, which is characteristic of rare-earth ions.

A further characterization of scintillator is its speed. The speed depends on the transfer (R_{tra}) and emission processes (R_r). The slower part of the process defines the speed. Both R_{tra} and R_r have a exponential decay time. This is a property of a difference of exponentials which, describe the rise and decay time of the luminescence in the system of centers fed by a transfer. We must note that the emission rate R_r has a maximum rate which limits the performance of the scintillator. On the other hand the transfer rate can be slow but still contribute to the overall efficiency of the scintillator. For electric dipole emission the radiative decay time is given $\tau_r = 1/R_r$ nsec. It follows, like described in [2], that

$$\tau_r = 1 - 5 * 10^{-5} \frac{\lambda^2}{\frac{1}{9}f(n^2+2)^2 n} \quad (6.10)$$

where f is the oscillator strength of the transition, λ the wavelength of the transition (nm) and n the refraction index. From equation(6.10) one can see that from all visible wavelengths that a scintillator might emit, the shortest wavelength (blue) increases the speed the most. In Table 6.2 follows some of the most important parameters for choosing a scintillator.

	E_g (eV)	ξ/E_g	T (nsec)	L (ph/MeV)	β	Q	S	η_q
NaI:Tl	5.9	2.6	230/1000	38000	0.88	1	0.59	0.52
CsI:Tl	6.4	2.4	800/6000	65900	0.97	1	0.99	0.96

BGO	5	3.3	400	8200	0.69	0.13	1	0.094
CaF ₂ :Eu	12.2	3.6	940	24000	0.63	1	1	0.67
CeF ₃	10.4	3.8	20	3200	0.61	1	0.13	0.077
CeP ₃ O ₁₄	8.7	7.8	30	4000	0.30	1	0.27	0.08
LuPO	8.7	6.1	25	17200	0.37	1	0.92	0.34

Table 6.2. Important scintillator values. Data from [2].

6.2 CMOS and scintillator

The purpose is to design a CMOS x-ray imager that is capable of taking x-ray images for dental radiography. Only recently digital x-ray imaging is an alternative to x-ray imaging with silver films. Digital imaging offer many advantage over traditional x-ray imaging. The following table shows some of the differences between digital and silver film medical imaging:

<i>(Silver) Film based imaging</i>	<i>Digital imaging</i>
Density: Degree of darkness of the exposed film	Brightness: Same as density in film based imaging
Latitude: Distinguishable densities made by the measured exposures.	Dynamic Range: The numerical range of each pixel.
Film speed: Radiation needed to produce the average density needed, also depended on the films sensitivity. Less radiation is needed if the film is fast.	Linearity: A linear or direct relationship between exposure and image density.
Contrast: Compared areas have different densities, which is called contrast. High contrast images have few shades of gray will in the contrary low contrast will have more shades of gray.	Contrast Resolution: Capability to distinguish small density differences.
Resolution: Capability to separate small close spaced objects. This is measured in line pairs per millimeter	Spatial Frequency and Modulation transfer function (MTF): The first is the same as the resolution in film based imaging. MTF is a measure of image fidelity as a function of spatial frequency. In simple words it's a measure of how close the image is to the actual object.
Radiographic mottle: Uneven densities in the exposed film.	Background Electronic Noise: Leakage current or current that not represents any image information but is an inherent property of the devices used.
Sharpness: Ability of a radiograph to distinguish boundaries of densities.	Signal to noise ratio: Like seen many times in previous chapters, it is the fraction of the output to the noise.

Table 6.3. Differences in terminologies between tradiotinal and digital x-ray imaging.

In dental radiography a intensity peak ranging from 40 keV to 60 keV are produced. A silicon wafer with standard thickness (525 μm) absorbs about 3.38% of 60 keV x-ray energy, this is not suitable at all. To increase detection an x-ray scintillation layer is placed on top on the photodiodes. In this way standard a CMOS APS sensor can be used, without needing any modifications. Of course modifications exist in the fabrication process, where a suitable scintillator must be laid on top on the photodiodes. The choice of scintillator is something that needs to be discussed. In most application where the goal

is to detect rather small amount of radiation, like in dental radiography (10-70 keV), the scintillator must satisfy two criteria: One being having a maximum light yield, so that the quantum efficiency of the x-ray detector is about 80% and the other one criteria is that the scintillator must be compactable with CMOS process. Scintillator measurements has been taken and compared, see [4]. It has been found that CsI:Tl satisfy the above mentioned requirements, and therefore is the most used scintillator in applications which use array systems for sensing. The fabrication process on how to couple a scintillator material with a sensing array is described in [5]. In the following text we just regenerate the most important facts. For the start we know from appendix A that silicon has a cutoff frequency that is 1100 nm. Scintillators emission wavelength ranges between 400 to 650 nm. This is means that the photodiode should have its maximum responsivity in the scintillators wavelength range. As one can see from Table 6.4 the emission peak wavelength for CsI:Tl is

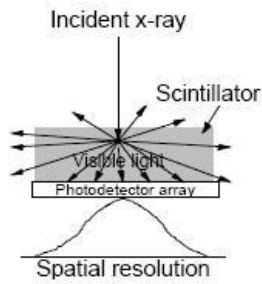
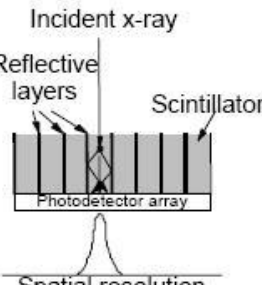
Material	Density (g/cm ³)	80% absorption (μm)	Photons or e/h pairs per MeV	n	Decay time (ns)	Effective Atomic Number	Emission wavelength (nm)	
CsI(Tl)	4.51	400	65900	1.74	10 ³	54	560	 <p>Figure 6.5 (a) Simple placement of CsI(Tl) and the resulting spatial resolution. Figures from [5]</p>
Silicon	2.33	200000	277000	4	0	14	-	 <p>Figure 6.5 (b) Interjection of reflective layers to guide the scintillated light. Figures from [5]</p>
SiO ₂	-	-	-	1.46	-	-	-	

Table 6.4. Properties of CMOS and Scintillation materials

about 550 nm, so the photodiode must be “tuned” to have its peak sensitivity about the same wavelength. When a x-ray is absorbed by the scintillator visible light is produced and is distributed in random manner, in all directions. Spatial resolution becomes then a problem since it is maximum where the light source is and hence the nearest pixel to this source has maximum output (see Figure 6.5 (a)). Of course not only one x-ray exist, and hence many light sources result. The overlapping resulting light can then produce wrong images. From Figure 6.5 (b) we see the resulting spatial resolution is increased for every

pixel if the scintillator is placed inside a well with the walls being made of a reflective material. The material proposed in [5] is aluminum, it has a high reflectivity, low density and low atomic number. Thus aluminum can be used to coat the scintillator even on top, since its low atomic number quarantines us that the x-ray will penetrate through the aluminum walls but restrains visible light through reflection to the well. Nonetheless we must note that a portion of light inside the well is lost due to the reflection mechanism. The mathematical formulation of the reflective efficiency of aluminum can be formulated with the following definitions:

- Light that reaches the photodiode is denoted as L_{pd}
- Emitted light by the scintillator is denoted as L_R
- R_A is the ratio between the area of the photodiode and the area of the scintillator
- R_{loss} is the percentage of losses in each reflection

Then

$$\frac{L_{pd}}{L_R} = \frac{1}{1-(1-R_A)(1-R_{loss})} \quad (6.11)^1$$

Equation 6.11 is useful to study the impact of light efficiency when changes of the material dimensions are made. We have used the term “tuned” to indicate that the photodiode below the scintillator must be adjusted to detect 560 nm, this means that the photodiode depletion layer depth should be around 0.75 μm [6].

The fabrication steps of coating the scintillators are described by [7]. In this approach two levels of dies are used to construct the detector, a photodetector matrix die and silicon cavity die. The cavities in the silicon are opened using DRIE (Deep Reactive Ion Etching) technique. Without going into further details (for more information see [7]) the basic steps of the technique can be seen in the Figure 6.6 (a)-(f). Note SU-8 is a photoresist material.

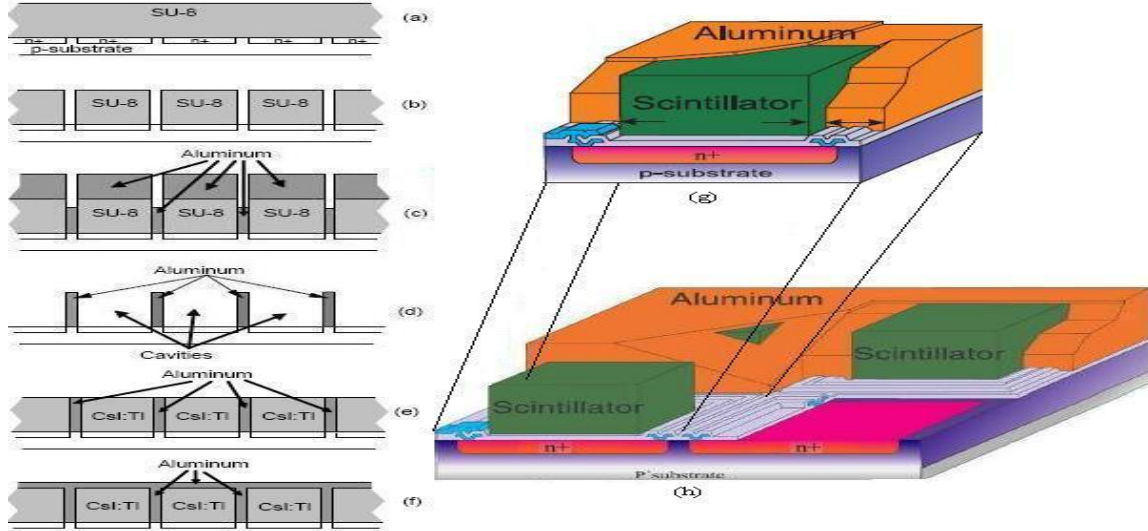


Figure 6.6 (a) Deposit of SU-8 above the CMOS detector (b) SU-8 pattern production of (c) Aluminum is evaporated (d) Complete removal of SU-8 (e) CsI:Tl is inserted into cavities (f) Aluminum is evaporated on top (g) One pixel (photodiode) overview (h) 2x2 overview of scintillator & CMOS system. Figures from [5] and [8].

¹ The relationship is referred in [8] and the we are only reproducing the result

6.3 Sensor implementation

In this section we seek to construct a CMOS APS image sensor with the help of electronic design automation software (EDA). With EDA it is possible to design and simulate a CMOS APS image sensor, which is called schematic. The schematic is really a pseudo-implementation and therefore symbols for components, connections and various sources are used to represent their existence in a design. In the schematic all components function according to their model. A component can have many models, like a car can be compactable with different engines, but only one model is representing a component at a time. This is the equivalent of saying many engines can be inserted into a car, but the car cannot run with many engines, only one engine can be used with a car at one time. The choice of model to use with a certain component is up to the analog designer. An example of a symbol used for the BSIM transistor can be found in the appendix “Transistors and BSIM model”. After verification through schematic simulation that the system is functioning properly the design must be translated on the silicon wafer. This is done with the layout design, also supported by most of the existing EDA software. The following sensor implementation is done only in schematic. The layout should be done only if the designer is confident that the schematic is final, meaning fully functional and complete in all desired perspectives. The layout design can be send to foundry where the design is been “printed” on the wafer. Therefore a good schematic design makes up the most important work of practically any implementation.

The goal is to implement a CMOS APS sensor that can be used for dental radiography. In dental radiography, a film based material is inserted into the oral area of the patient. The film material is covering the tooth area for examination. The patient usually presses the material against the tooth while the dentist is taking the x-ray. The material covering the tooth area is always on the opposite side where the x-ray machine stands, see Figure 6.4. Looking at Figure 6.4 we see that the area of the material is not very big, hence we are restricted to about an area of $2 \times 1,5 \text{ cm}^2$. The restricted area is a factor that must be met by the available CMOS technology of $0.35 \mu\text{m}$. As a result before we proceed we must introduce some technology aspects, which in the end allow us to calculate the pixel area needed for the “dental sensor”. The EDA software that we use for the schematic design is CADENCE IC 5141, it is precise, has fast simulation time and has a vast range of sub-applications. It is worth mentioning that high demanding corporations and top academic institutions are using CADENCE, on top on that many design kits (technology process files) have as priority to work with CADENCE first. One out of many companies that provide design kits is Austriamicrosystems (AMS), which developed a $0.35 \mu\text{m}$ design kit for CADENCE. The $0.35 \mu\text{m}$ design kit is also available at EUROPRACTICE IC Service. Furthermore Austriamicrosystems is developing a photodiode model available for the $0.35 \mu\text{m}$ process, data for this model can be found in the documentation files of the photodiode. For the calculation of the area and as well for the pixel design the following data from the photodiode and from the $0.35 \mu\text{m}$ process documentation file are retrieved and presented:

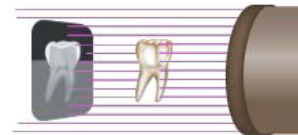


Figure 6.7. Film position in dental x-ray imaging. Figure from [9]

Capacitance of photodiode	0.08 fF/ μm^2
Responsivity of Photodiode at 550nm	290 mA/W
Dark current	45pA at 27° Celcius
Minimum photodiode size	6 μm x 6 μm
0.35 μm process supply voltage	3.3V
Additional process steps for an inorganic anti-reflecting layer (ARC) on top of the photodiode	1 step
0.35 μm transistor model	BSIM3v3

Table 6.5. Important parameters of the 0.35 μm technology

6.3.1 CMOS APS design

To calculate the number of pixels that might fit into an area of 2x1.5 cm² we examine the minimum pixel size and the capacitance of the photodiode. The minimum photodiode size is 36 μm^2 . We must account additional space for the reset, column select and common drain (source follower) transistors. According to [28] the best layout achievable for a pixel in the 0.35 μm process is a square from which the photodiode occupies $\frac{3}{4}$ of the pixel in the form of an L (more precisely a right rotated by 90° degree L “ – “). As a result the minimum pixel size is 36*1,25 = 45 μm^2 . To take a precaution we start using only half of the available sensor area, namely 1.5 cm², remember that the CDS and ADC blocks, which are at the bottom of each column, can take a large amount of space. We seek to find the number of pixels when minimum photodiode size is used. Let's denote the area of the pixel as A_{pix} , and $A_{pdarray}$ as the available photodiode area. Knowing that 1 μm =0.0001cm we have

$$\#pixels = \frac{A_{pdarray}}{A_{pix}} = \frac{1.5 \cdot 10^{-4}}{0.0045 \cdot 0.0045 \cdot 10^{-4}} \approx 74074 \quad (6.12)$$

If we take a larger pixel size, say 100x100 μm^2 then

$$\#pixels = \frac{A_{pdarray}}{A_{pix}} = \frac{1.5 \cdot 10^{-4}}{0.01 \cdot 0.01 \cdot 10^{-4}} \approx 15000 \quad (6.14)$$

As one can see from Table 6.6 a 100x100 μm^2 pixel size can only be chosen to meet the lowest available image resolution (90x60 pixels). Even for the smallest pixel size (6.12) the resulting array can provide the second lowest image resolution available (180x120). Hence we are forced to reverse the procedure and choose a suitable image resolution and then calculate the size. To choose between image resolutions we present the most common image resolution available in Table 6.6.

Pixels (px) /cm	3.75 x 2.5	7.5 x 5	11.4 x 7.6	15.2 x 10.2	22.9 x 15.2	30.5 x 20.3	45.7 x 30.5	61 x 40.6	91.4 x 61	121.9 x 81.3	182.9 x 121.9
90 x 60 px 5 400 px	WEB only	WEB only	20 dpi	15 dpi	10 dpi	8 dpi	5 dpi	4 dpi	3 dpi	2 dpi	1 dpi
180 x 120 px 21 600 px	WEB only	WEB only	40 dpi	30 dpi	20 dpi	15 dpi	10 dpi	8 dpi	5 dpi	4 dpi	3 dpi

360 x 240 px 86 400 px	240 dpi	WEB only	80 dpi	60 dpi	40 dpi	30 dpi	20 dpi	15 dpi	10 dpi	8 dpi	5 dpi
540 x 360 px 194 400 px	360 dpi	180 dpi	120 dpi	90 dpi	60 dpi	45 dpi	30 dpi	23 dpi	15 dpi	11 dpi	8 dpi
720 x 480 px 345 600 px	480 dpi	240 dpi	160 dpi	120 dpi	80 dpi	60 dpi	40 dpi	30 dpi	20 dpi	15 dpi	10 dpi
900 x 600 px 540 000 px	600 dpi	300 dpi	200 dpi	150 dpi	100 dpi	75 dpi	50 dpi	38 dpi	25 dpi	19 dpi	13 dpi
1350 x 900 px 1 215 000 px	900 dpi	450 dpi	300 dpi	225 dpi	150 dpi	113 dpi	75 dpi	56 dpi	38 dpi	28 dpi	19 dpi
1800 x 1200 px 2 160 000 px	1200 dpi	600 dpi	400 dpi	300 dpi	200 dpi	150 dpi	100 dpi	75 dpi	50 dpi	38 dpi	25 dpi
2700 x 1800 px 4 860 000 px	1800 dpi	900 dpi	600 dpi	450 dpi	300 dpi	225 dpi	150 dpi	113 dpi	75 dpi	56 dpi	38 dpi
3600 x 2400 px 8 640 000 px	2400 dpi	1200 dpi	800 dpi	600 dpi	400 dpi	300 dpi	200 dpi	150 dpi	100 dpi	75 dpi	50 dpi
5400 x 3600 px 19 440 000 px	3600 dpi	1800 dpi	1200 dpi	900 dpi	600 dpi	450 dpi	300 dpi	225 dpi	150 dpi	113 dpi	75 dpi
7200 x 4800 px 34 560 000 px	4800 dpi	2400 dpi	1600 dpi	1200 dpi	800 dpi	600 dpi	400 dpi	300 dpi	200 dpi	150 dpi	100 dpi
9000 x 6000 px 54 000 000 px	6000 dpi	3000 dpi	2000 dpi	1500 dpi	1000 dpi	750 dpi	500 dpi	375 dpi	250 dpi	188 dpi	125 dpi
13500 x 9000 px 121 500 000 px	9000 dpi	4500 dpi	3000 dpi	2250 dpi	1500 dpi	1125 dpi	750 dpi	563 dpi	375 dpi	281 dpi	188 dpi
18000 x 12000 px 216 000 000 px	12000 dpi	6000 dpi	4000 dpi	3000 dpi	2000 dpi	1500 dpi	1000 dpi	750 dpi	500 dpi	375 dpi	250 dpi

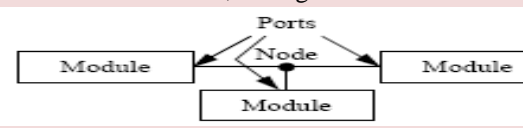
Table 6.6. Image pixel resolutions in the form of pixels, dpi and the area (cm) that they occupy

We choose the lowest available resolution available with the lowest impact on the area of the sensor. The 360 x 240 (Video CD resolution) yields 84000 pixels and hence comparing it with equation 6.12, one concludes that the resulting area will not be bigger than whole sensor area of 3 cm². The exact area of the Video CD resolution is

$$A'_{pdarray} = 86400 * (0.0045 * 0.0045) = 1.7496 \text{ cm}^2 \quad (6.15)$$

Which leaves us an area of 1.2504 cm^2 for implementing the ADC, CDS, control and memory circuitry. The ADC and the CDS circuit contain each an operational amplifier, the ADC can contain also a comparator. Both comparator and amplifier can have up to 20 transistors each. If the majority of the transistors are considered to be at minimum dimensions ($W, L=0.35\mu\text{m}$), then $(20+40)*0.35\mu\text{m}^2 = 0.000735 \text{ cm}^2$ is really a little area compared to the area of the photodiode array, and like mentioned in chapter four, it is the capacitors and resistors in the ADC or in any circuit that take up most of the area. Now that we have calculated and verified that a 360×240 image sensor can be implemented on a 3 cm^2 area, we focus onto the photodiode. Like mentioned before the photodiode is still under development by AMS and after many hours of investigating we came to the conclusion that it cannot be used yet. The problem of using the AMS photodiode consisted mainly in the compatibility with CADENCE and not in the model itself. Future releases of the AMS photodiode have a good chance to work “off the shelf”. Nonetheless we keep the data and the analysis derived because of its $0.35\mu\text{m}$ process parameters. Instead of waiting for a next release of the photodiode we turned to model the photodiode with verilog-ams. Verilog-ams is a language similar to VHDL, it is the direct successor of the hardware description language HDL. In other words it is a flow and control language aimed for electronic analog and mixed-signal implementations. The verilog-ams code was implemented based on the implementation in [11]:

	Code	Description
1	<code>`include "disciplines.h"</code>	<code>`</code> is similar to <code>#</code> in C (HeaderFile, bibliaries). The command <code>“include”</code> signals the compiler to include a file (This is often called a directive). The discipline file is a standard file that contains variable declarations which have certain properties (thermal, electrical and others see IEEE 1364 1995). For each property default values (behaviors) are defined, called natures. Furthermore functions to access the standard values are also defined in <code>“disciplines.h”</code> . The access functions are used to determine the allowed range of the inserted value, result(s) of a standard handling process, etc.
2	<code>`include "optical_sf.v"</code>	If a new type is needed, include the necessary information in the discipline file or include a new <code>“.h”</code> file, describing the new features. This is been done for the optical discipline and nature, which is not contained in the IEEE 1364 1995 standards.
		<i>“New disciplines defined in in the optical_sf.v file”</i>
3	<code>Discipline optical_sf</code>	
4	<code>potential Illuminance;</code>	Define <code>optical_sf</code> as a new discipline that has the property of potential.
5	<code>enddiscipline</code>	
		<i>“New nature is appended to the discipline.h file”</i>
6	<code>nature Illuminance</code>	Define the property of this new discipline as

		having
7	units = Cd;	capacitance as units
8	access=LP;	and access function LP
9	Ifdef CHARGE_ABSTOL	If a tolerance is been defined
10	abstol= CHARGE_ABSTOL;	then set it to be the new default value
11	else	else
12	abstol=1e-14;	set it to the default value
13	endif	
14	endnature	
15	<pre>module pd(ilight, tano, tcat,tj);</pre>	<p>The module section provides the separation between different entities, or better called components. Hence one defines the internal functions of a device, say a resistance R, in the module block, so that the compiler can distinguish an R from a C module. Every module can have in, out or inout connections (ports) and with these connections different modules can connect to each other, see figure below.</p>  <p>Figure 6.8. Ports can connect to modules and other entities (Node here). Figure from [12].</p> <p>A module declaration is followed by its naming and its basic ports. After that follows the description of the properties of the module ports, possible parameter and variable declarations (parameters of the module and temporary variables used in the analog block), and in the end the analog begin block.</p>
16	inout tano,tcat;	define as input and output tano, tcat
17	in ilight;	as in ilight
18	electrical tano, tcat;	tano and tcat have the electrical property
19	optical_sf ilight;	while ilight is the new defined potential
20	thermal tj;	tj is has thermal properties
21	parameter real CD=3fF;	Parameters value are set according to the AMS
22	parameter real RLEAK=1M;	photodiode data
23	parameter real SENSITIVITY=0.29;	
24	parameter real IDARK0=45p;	
25	parameter real IDARK_DT=45.0;	
26	real tempc, ir, ic, ip, idark, id;	Define variables that defines the module ports

27	analog begin	<p>The analog begin is responsible of the functionality of the module. Only one analog block can exist in a module. Here all kinds of function can be used to set the signals of the module. Functions seen here are</p> <ul style="list-style-type: none"> • The access functions: (<i>temperature calculation, potential calculation, optical power calculation, power calculation</i>), • Build in mathematical functions: (<i>derivative “ddt”, power of x^y “pow”, absolute value “abs”</i>) • Potential access function: The potential between to ends of a branch • Flow access: The current that flows through a branch, defined by two ends. • Branch access: It is not a function but modifies a branch defined with a potential or flow access function. This done with the contribution operator <+.
28	tempc=Temp(tj)-273.0;	Tempc takes the translated value of tj (Kelvin to C°)
29	ir=V(tano,tcatt)/RLEAK;	Leakage current created by pd resistance
30	ic=CD*ddt(V(tano,tcatt));	Current from the pd capacitance
31	ip=-SENSITIVITY*LP(ilight);	Photocurrent derived from (2.71)
32	idark=IDARK0*pow(10.0(tempc/IDARK_DT));	From the diode equation in the absence of light
33	id=ir+ic+ip+idark;	The forward current is the sum of each current
34	I(tano,tcatt) <+ id;	Set current through module to be the forward current
35	Pwr(tj) <+ abs(id)*V(tano,tcatt);	Power consumed by the module
36	end	
37	endmodule	

The problem with the verilog-ams version of the photodiode is that CADENCE’s verilog-ams compiler doesn’t handle new disciplines, and therefore this kind of photodiode could not work either. The choices were narrowed down to only one: To simulate the discharge of the sensing node due to photocurrent with a piecewise linear source. The piecewise linear source (PWL) can output certain set voltages in various time periods. Therefore it is possible to simulate at least the behavior of photodiode in an APS pixel setup. For this setup we neglect the reset transistor and of course the photodiode and replace the two with a PWL source. The rest of the setup is like in a normal APS configuration scheme. In the “remaining” APS scheme the following components exist: A source follower and a column-out select transistor (SEL). The neglected reset as well as the SEL transistor, are switches. Switches are used everywhere, from the design of switched capacitor circuits to the design of ADC circuits. In the APS system, a switch has to be designed so that the time constant to charge nodes (the sensing node for instance) is not large and of course the passing voltage is not affected by the switch itself. As a result it is very important to understand switches, which is what we examine next.

6.3.1.1 Switch design

As seen from the APS circuit a MOS transistor is used as a switch. This can happen because a MOS can be on while carrying zero current and its gate voltage is independent of the source and drain voltage. Hence clocking the gate of a MOST, turns the MOST on and off. There is however a problem, the MOST isn't an ideal device and gate to source and gate to drain capacitance as well as other non-idealities exist. In figure below an overview of the existing non-idealities is presented.

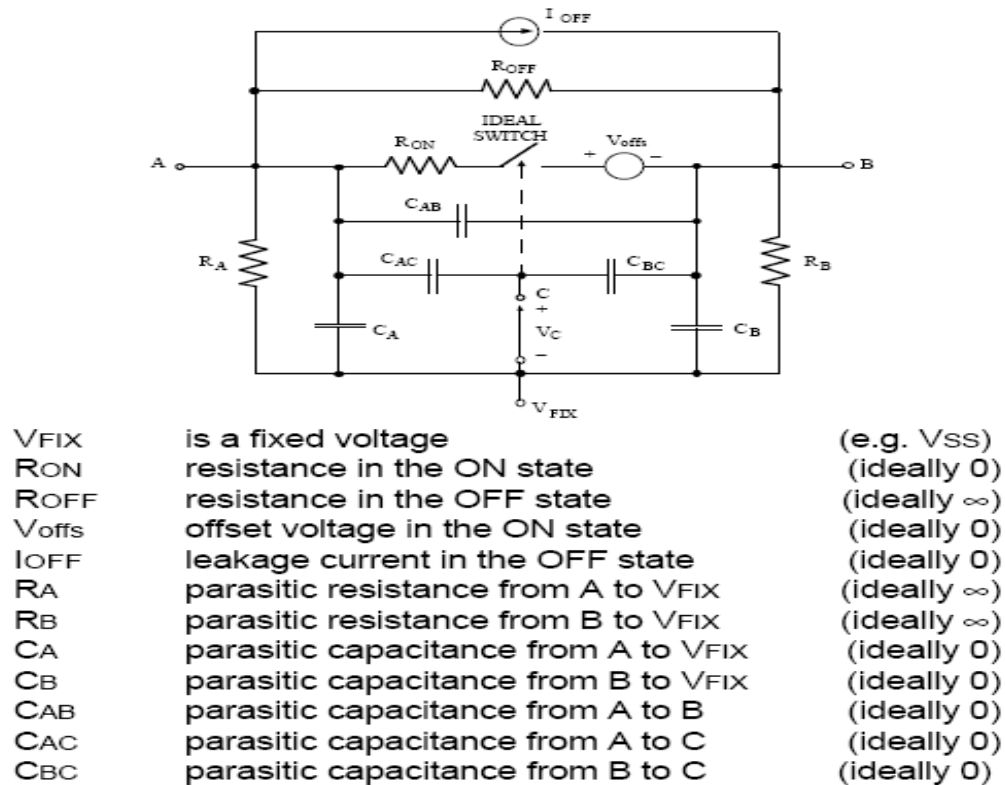


Figure 6.9. Non idealities in a MOST switching device. Figure from [13]

The parasitic capacitance and resistance are very small² and are quickly charged. The most important factor in designing a MOST switch is its ON resistance R_{ON} . The offset voltage usually has little impact on the overall result. When a non-zero voltage is transferred to the gate of the MOST (CLK=high), a short circuit between the source and the drain occurs and the voltage is transferred from one end (drain/source) to the other (source/drain). It doesn't matter which of the two terminals, drain or source, is connected to the voltage to be transferred, because the MOST is a symmetrical device. When the gate is at 0 volt (CLK=low), R_{off} is not infinitive but is big and hence no voltage can be transferred. Looking at this from another perspective we can state that when the gate voltage is 0 no inversion channel exists and no current is raised at the other terminals, hence no voltage can occur. When the CLK is high the amount of voltage transferred depends solely on the R_{ON} resistance. The ON resistance is the drain source resistance of the MOST, hence from the transconductance $g_{ds}=1/r_{ds}$ we get

² This is when not working in the RF (Ghz) range where the capacitance become short circuits.

$$R_{ON} = \frac{1}{g_{ds}} = \frac{1}{k_{n,p} \frac{W}{L} (V_{GS} - V_{TH} - V_{DS})} \quad (6.16)$$

When $V_D = V_S$ (6.16) then

$$R_{ON} = \frac{1}{g_{ds}} = \frac{1}{k_{n,p} \frac{W}{L} (V_{GS} - V_{TH})} \quad (6.17)$$

If $V_G = \text{CLK}_{\text{high}} = V_{DD}$ and the voltage to be transferred is denoted as V_{in} then

$$R_{ON} = \frac{1}{g_{ds}} = \frac{1}{k_{n,p} \frac{W}{L} (V_{DD} - V_{in} - V_{TH})} \quad (6.18)$$

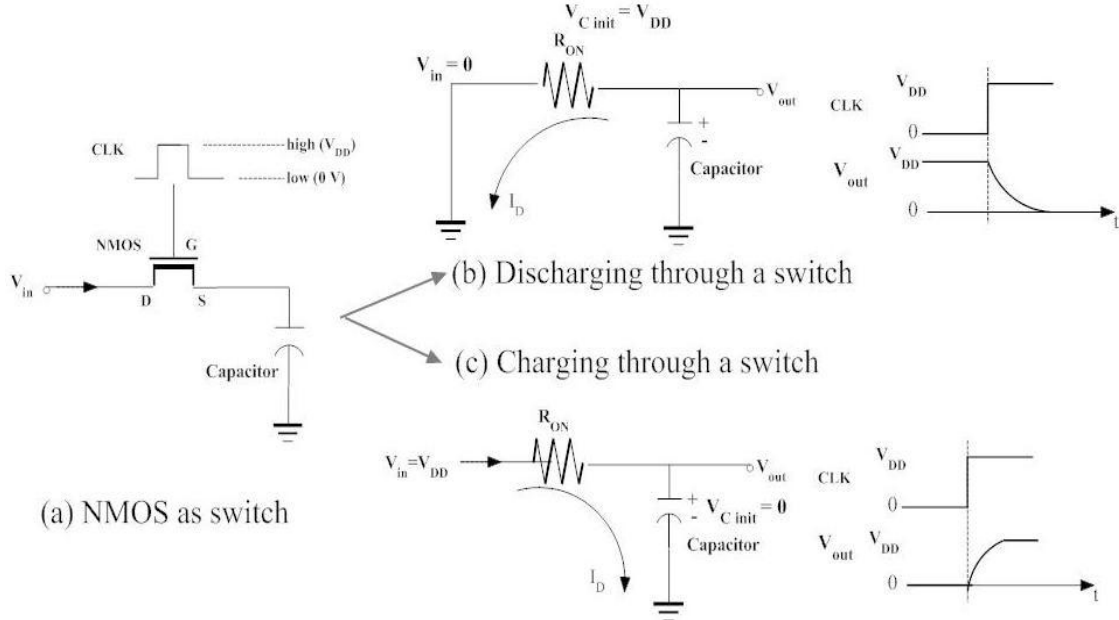


Figure 6.10 (a) Switch principle and the associated effects when $V_G \neq 0$ (b) Capacitor discharges through R_{ON} and (c) charges through R_{ON} . Bulk not seen in the figure is grounded.

The switch can be implemented with a NMOS. Using the data provided for the $0.35\mu\text{m}$ process we find the following values: typical gain factor is $170 \mu\text{A}/\text{V}^2$ and threshold voltage for short channel is 0.6V . With these data we can calculate the R_{on} . Taking $W/L=1$ and $V_{in}=2\text{V}$ we get:

$$R_{ON} = \frac{10^6}{170(3.3-2-0.6)} \approx 8.4 \text{ k}\Omega \quad (6.19)$$

The resulting RC circuit seen in Figure 6.10 is known to have time constant τ . The same time constant exists in a simple RC circuit and therefore the time that the switch needs to charge or discharge the capacitor is equal to

$$\tau = 5 * R_{on} C \quad (6.20)$$

The capacitance per pixel, denoted as C_{pix} , is the capacitance of the photodiode C_{pd} multiplied by A_{pix} .

$$C_{pix} = A_{pix} * C_{pd} = 45 \mu\text{m}^2 * 0.08 \text{ fF}/\mu\text{m}^2 = 3.6 \text{ fF} \quad (6.21)$$

Hence the capacitor with $V_{in}=2\text{V}$ is fully charged at $t=30.24\text{ps}$. Knowing the integration time of the pixel to be in the order of milliseconds and the rows of the pixel array to be 340 we conclude that the time constant is more than enough to read out every pixel.

6.3.1.2 Source follower

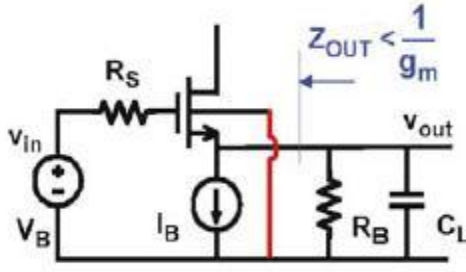


Figure 6.11 MOST in common drain configuration. Common is the drain to the output and input! Figure from [14].

Looking at Figure 6.11 we see that the output of the common drain is at the source and hence the output is analogous to the transconductance g_m and the resistance r_{ds} of the transistor. The transistor is being biased by a current and therefore the current through the transistor is kept constant, which results also in a constant V_{GS} . The latter can be explained by applying a small signal to the gate, the source then will respond equally by changing its potential by the same change that occurred at the gate. The voltage gain of the transistor is therefore unity.

This is why the common drain is also called a voltage buffer. If for example a voltage is applied to the gate (V_{in}) the output at the source shows

$$V_{in} - V_{GS} = V_{out} \quad (6.22)$$

It has been proven [15], that the input impedance of a source follower is high while $R_{out} = 1/(g_m + g_{mb})$ is low. Therefore the source follower offers also isolation; this is exactly the reason why a source follower is used in the APS system. It's "job" is to track the input (sensing node) while isolating it from external voltages. To calculate the bias current needed to "track" a voltage of 2 V we use the following equations. We know that the current through a saturated transistor is described by

$$I_D = k' \frac{W}{L} (V_{GS} - V_T)^2 \quad (6.23)$$

where the channel modulation effect is neglected. Solving 6.23 for V_{GS} we get

$$V_{GS} = \sqrt{V_T - \frac{I_D}{K' \frac{W}{L}}} \quad (6.24)$$

where V_T is given by (B-5a), instead of V_{to} . When the NMOS body terminal must be connected to the substrate (ground) V_{to} cannot be used in (6.24). The 0.35 μm process is p-substrate based, hence $V_{BS} \neq 0$. This effects the gain of the source follower which not one but $1/n$. Despite the non unity gain, we continue to calculate the bias current. The chosen configuration to bias the follower is a NMOS current mirror which mirrors 1, meaning no current is multiplied but only mirrored. Looking at the APS we see that the voltage drop across the CD must not affect the current mirror. The current mirror must at all time's bias the CD, hence we construct must a current mirror that met this requirement. How do we accomplish that? For our purpose we use the BSIM model but being open minded we can "steal" and apply a very important concept, the inversion coefficient (IC). In the EKV model, see [16], [17] and [18] the IC is defined as

$$IC = \frac{I_D}{I_{spec}}, \text{ where } I_{spec} = 2U_T^2 n \beta \quad (6.25)$$

where $U_T = \frac{k*T}{q}$, $n = 1 + \frac{\gamma}{1+2\sqrt{\Psi_0+V_p}}$, $\beta = \mu_{n,p} C_{ox} \frac{W}{L}$, V_p the pinch off voltage, Ψ_0 the surface potential at zero bias, γ the body effect coefficient, k the Boltzmann constant, T the temperature and q the electron charge. For an IC of 10 the channel is on the onset of strong inversion. The value of I_{spec} can easily be found. Solving equation (6.25) for I_D , we get the necessary current for which the non diode-connected transistor stays in saturation

($I_C=10$). Saturation of the non diode-connected transistor depends also on its V_{DS} , this voltage drop must be small so that the voltage path: supply voltage $\rightarrow V_{DS}$ (of the source follower) \rightarrow Switch (SEL), is always greater than voltage drop of the non diode-connected transistor. If this isn't the case the non diode-connected transistor enters either the linear region or either shut offs, needless to mention that this situation is not allowed in current mirrors. From equation(6.22) and(6.24) one can see that the bias current must be held small, so that the output follows the input with a difference of V_T . This difference between output and input is unavoidable. Arithmetic calculations using(6.25),(6.24) and(6.23) showed that a current of 50 nA satisfies a minimum voltage droop (V_T) between input and output of the source follower. To avoid short channels effect and to satisfy a minimum V_{DS} across the non-diode connected transistor we raised the dimensions of the current mirror transistors to $W=5$ and $L=10$ μm respectively. All this might sound a little confusing and for that reason we summarize the requirements and methods of building a suitable current mirror for biasing a source follower with 2 V input.

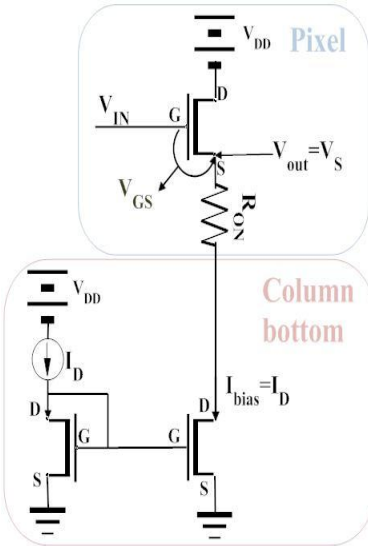


Figure 6.12. Current mirror for biasing source follower. Note that the bias circuit is not in the pixel but at the bottom of each column. Bulk not seen in the figure is grounded.

#	Steps	Findings
A	Find necessary I_D for which (6.24) yields V_T	$I_{bias} = I_D = 50 \text{ nA}$
B	Set desired input in(6.22).	2 V
C	Estimate voltage drop in the path Supply \rightarrow CD \rightarrow Switch for V_{OUT_MAX} and V_{OUT_MIN} of the CD.	$3.3 - (A-4,0) - (R_{ON} * I_{bias})$ $\rightarrow \max[1.899 \text{ V}]$ $\rightarrow \min[3.299 \text{ V}]$
D	Substitute in(6.25) $I_C=10$ and I_D form A	Done
E	Select I_{spec} such that the two transistors in the mirror configuration have current gain 1 $(\frac{W}{L})_2 / (\frac{W}{L})_1 = 1$ – Also set W and L having long channel values (eliminating short channel effects).	Best current and voltage response found for $(\frac{W}{L})_2 = \frac{0.35}{10} \mu m$ $(\frac{W}{L})_1 = \frac{0.35}{10}$
F	Check in E the dimensions, so that minimum Voltage drop V_{DS} across the non diode-connected transistor is achieved. This and(6.25) assures us that the current mirror stays in saturation.	V_{DS} across the non diode-connected transistor for which the transistor is in saturation (mirrors the current) is minimum at 150 mV. Quite small compared to requirements in step C.

Table 6.7. Determining the characteristics of current mirror and source follower

6.3.2 CDS design

The CDS's purpose is to eliminate the column FPN noise by subtracting the reset signal from the sensing node signal. The subtracting is called double sampling. FPN consist of offset and gain components which increases with illumination but degrades the image the most at low illumination.

6.3.2.1 Fixed Pattern Noise (FPN)

Under uniform illumination FPN is typically reported as the standard deviation of the spatial variation in pixel outputs (temporal noise is excluded) as a portion of voltage swing, expressed in percentage. Values can range between 0.1% and 4%.

Four sources of variation cause FPN:

1. Column variations due to transistor variations (threshold voltages of transistors differ slightly) and ibias variation. Illumination is assumed constant
2. Pixel variation, dark current variations. Illumination is assumed constant
3. Offset variation that effect the 1 and 2 due to varying illumination
4. Gain variation due to amplifiers that are connected to the photodiode (PPS for example).

With different device parameters it's difficult to analyze FPN. To account for the deviations from the device's nominal parameters we accord to each parameter a random variable. For instance a device can have r parameter values, then the variables will be $Z_1, Z_2, Z_3, Z_4, \dots, Z_r$. The mean value of a variable denoted as z_i , expresses the nominal value of the parameter. For instance the following statement shows a deviation by Δ of the parameter under examination.

$$Z_i = z_i + \Delta Z_i \quad (6.26)$$

When dealing with small deviations and constant illumination we can approximate the pixel output voltage as a function of the device parameters by

$$V_o(Z_1, Z_2, \dots, Z_r) \approx u_0(z_1, z_2, \dots, z_r) + \sum_{i=1}^r \left. \frac{\partial u_0}{\partial z_i} \right|_{z_1, z_2, \dots, z_r} \Delta Z_i \quad (6.27)$$

where $u_0(z_1, z_2, \dots, z_r)$ is the nominal output voltage and the partial derivative $\frac{\partial u_0}{\partial z_i}$ is the rate of change of u_0 . Hence the change is what we seek and therefore the variation in V_o is due to the partial derivative term in(6.27):

$$\Delta V_o = \sum_{i=1}^r \left. \frac{\partial u_0}{\partial z_i} \right|_{z_1, z_2, \dots, z_r} \Delta Z_i \quad (6.28)$$

FPN is then defined as the variable change in nominal change in the output voltage. In other words it is the standard deviation σ_{V_o} of V_o . Lastly we can derive the standard deviation when assuming that ΔZ_i 's are uncorrelated:

$$\sigma_{V_o} = \sqrt{\sum_{i=1}^r \left. \frac{\partial u_0}{\partial z_i} \right|_{z_1, z_2, \dots, z_r}^2 \sigma_{Z_i}^2} \quad (6.29)$$

It is assumed, even not true not all, that FPN type one and two are uncorrelated. If they are uncorrelated, FPN source one and two can be analyzed separately. Column variations (Column FPN) and Pixel variations (Pixel FPN) are then two separate variations described by two separate variables. Let's define Z_1, Z_2, \dots, Z_c variables for column FPN and Z_1, Z_2, \dots, Z_p variables for the pixel FPN. Then

$$\Psi = \sum_{i=1}^c \left. \frac{\partial u_0}{\partial z_i} \right|_{z_1, z_2, \dots, z_c} \Delta Z_i \quad \text{Column FPN} \quad (6.30)$$

$$\Omega = \sum_{i=1}^p \left. \frac{\partial u_0}{\partial z_i} \right|_{z_1, z_2, \dots, z_p} \Delta Z_i \quad \text{Pixel FPN} \quad (6.31)$$

Using(7.29) we can state that the standard deviation of column and pixel FPN is

$$\sigma_{V_o}^2 = \sigma_{\Psi}^2 + \sigma_{\Omega}^2 \quad (6.32)$$

Offset and gain variations can be pact in a single statement that influences the pixel output voltage

$$u_0 = hj_{ph} + u_{os} \quad (6.33)$$

where h is the gain, j_{ph} the density of the photocurrent and u_{os} the offset voltage of amplifiers if used. As showed in [19] the resulting deviations can be expressed with

$$\sigma_H^2 * j_{ph} \quad \text{as gain FPN} \quad (6.34)$$

$$\sigma_{V_{os}}^2 \quad \text{as gain FPN} \quad (6.35)$$

where(6.35) is usually equal to the op-amp offset. To calculate the FPN one might first see the parameters that FPN effects. The below parameters are usually enough to calculate each FPN type:

parameter	sensitivity	effect on FPN
A_D	$\frac{j_{ph}^{t_{int}}}{C_f}$	pixel/gain
i_{dc}	$\frac{t_{int}}{C_f}$	pixel/offset
v_{os}^{op}	1	column/offset
C_f	$\frac{i_{dc}^{t_{int}} + C_{ol}v_T}{C_f^2}$	column/offset
	$\frac{A_D j_{ph}^{t_{int}}}{C_f^2}$	column/gain
v_T	$\frac{C_{ol}}{C_f}$	column/offset
C_{ol}	$\frac{v_T}{C_f}$	column/offset

Figure 6.9. Parameters to calculate FPN sources. Figure from [19]

In [19] an analysis is performed to calculate the APS FPN. It has been proven that the below scheme cancels the following FPN sources:

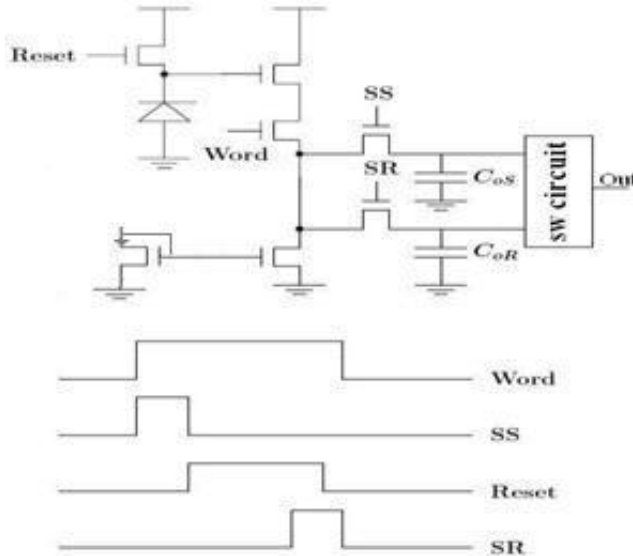


Figure 6.12. Half of the CDS system is placed in the APS pixel. The sw circuit, at the bottom of each column, is the other half of the CDS system. Figure from [19].

Offset variations due to

- threshold variations from the reset transistor
- threshold variations from the CD transistor
- reset noise due to capacitance of reset transistor
- mismatch in transistor dimensions of the CD transistor
- I_{bias} variations

Disadvantage is that it adds a reset noise during reset and readout. The whole noise is equal to $kT/2C_D$.

As one sees the CDS circuit is implemented with two sampling schemes which sample the signal and the reset value of the pixel during column selection. The two signals are then feed to a switched capacitor circuit (sw circuit) which in turn subtracts the two signals from each other. The switched capacitor circuit is implemented like the DAI in chapter five, with the difference that the capacitor for sampling is grounded. In other bibliography the sw circuit can be implemented in various ways, we chose a simple and minimum size sw, see Figure 6.10.

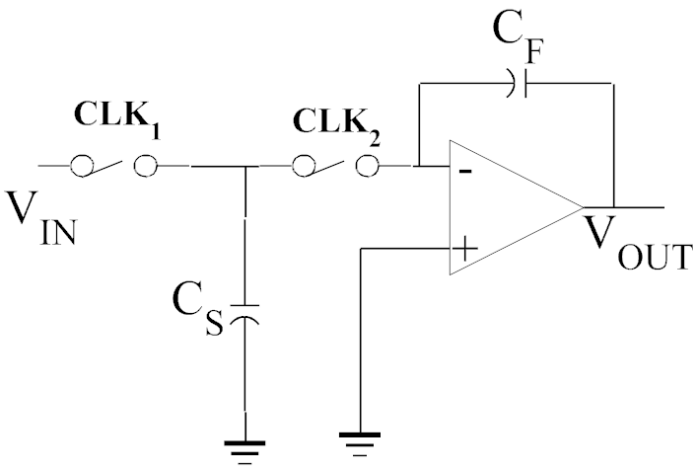


Figure 6.13. The proposed sw circuit

Fortunately this simple sw is available in the CADENCE functional library. Hence we use the CADENCE sw instead of building one. The parameters for the sw are the same as indicated in Figure 6.10, namely C_S , C_F and CLK (clock is made complementary in the sw block, hence only one clock as input is sufficient). The sw component has several signals to be set which are:

Switched Capacitor characteristics:

Terminals	
vout_p, vout_n:	output terminals [V,A]
vin_p, vin_n:	input terminals [V,A]
vphi:	switching signal [V,A]
Instance Parameters	
cap_in =	input capacitor value
cap_fb =	feedback capacitor value
vphi_trans =	transition voltage of vphi

Description

The switched capacitor integrator makes one transition between its internal capacitor at each rising clock vphi. That means that the clock frequency must be high to integrate the differential signal at its ports (vin_p, vin_n). Can be operated for non differantial signal grounding vout_n.

6.3.3 Pixel and Readout testing

The ADC that we choose is the one provided by CADENCE. It is an 8-bit Nyquist rate ADC, which has the following parameters:

8-bit /ADC characteristics:

Terminals	
vin:	[V,A]
vclk:	[V,A]
vd0... vd7:	data output terminals [V,A]

Instance Parameters	
mismatch_fact	= maximum mismatch as a percentage of the average value []
vlogic_high	= [V]
vlogic_low	= [V]
vtrans_clk	= clk high-to-low transition voltage [V]
vref	= voltage that voltage is done with respect to [V]
tde1, trise, tfall	= {usual} [s]

Description

This ADC comprises 8 comparators. An input voltage is compared to half the reference voltage. If the input exceeds it, bit 7 is set and half the reference voltage is subtracted. If not, bit 7 is assigned zero and no voltage is subtracted from the input. Bit 6 is found by doing an equivalent operation comparing double the adjusted input voltage coming from the first comparator with half the reference voltage. Similarly, all the other bits are found.

Now we are ready to implement a one pixel test, namely a pixel with all necessary read out circuitry (bias current, CDS and an ADC). Below is the schematic of the one pixel test cell view (schematic). Its separated by two pictures so that the reader sees most of the details.

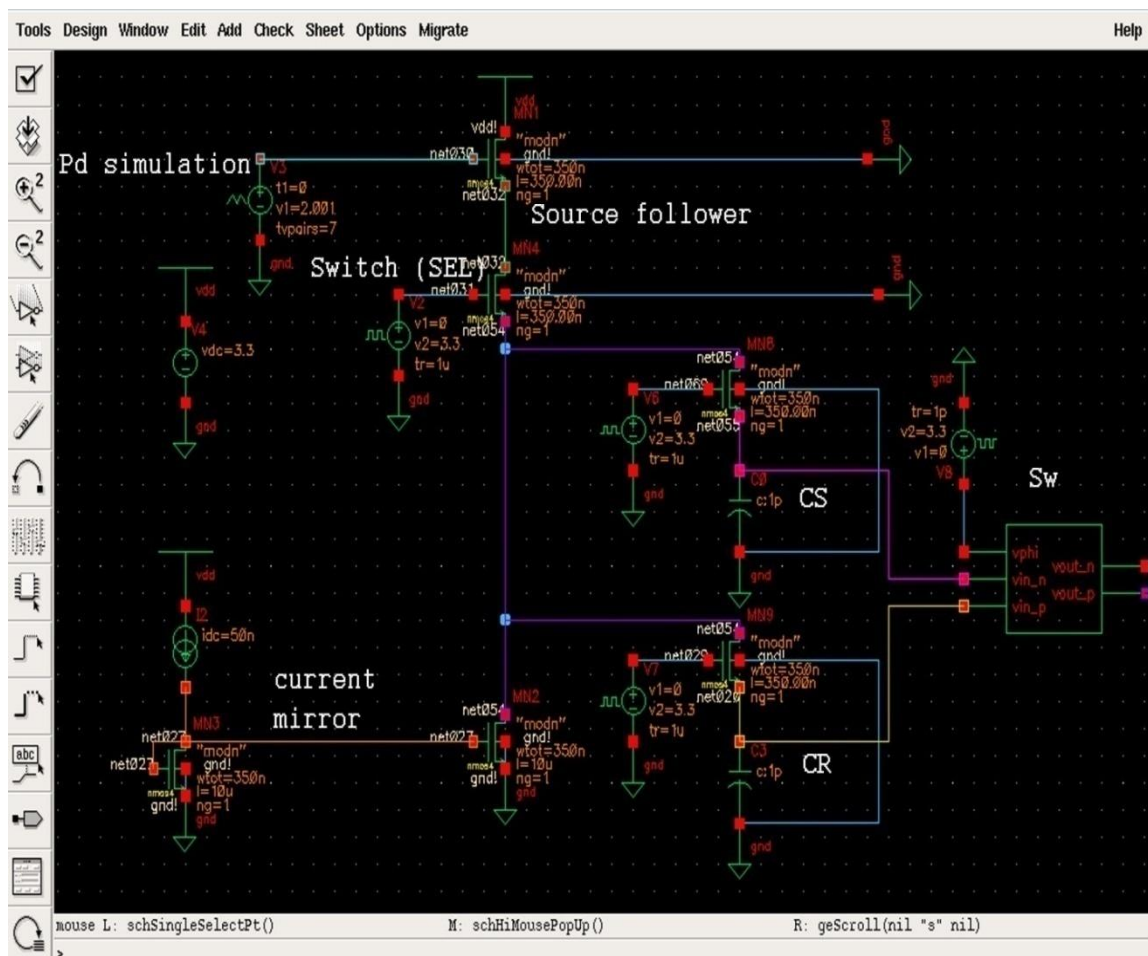


Figure 6.14. A one pixel implementation. Seen are the basic parts: Pd (simulated with a voltage piecewise linear source), Source follower, the bias current (implemented with current mirror), Reset signal sampler (CR), Signal signal sampler (CS) and the switched capacitor circuit.

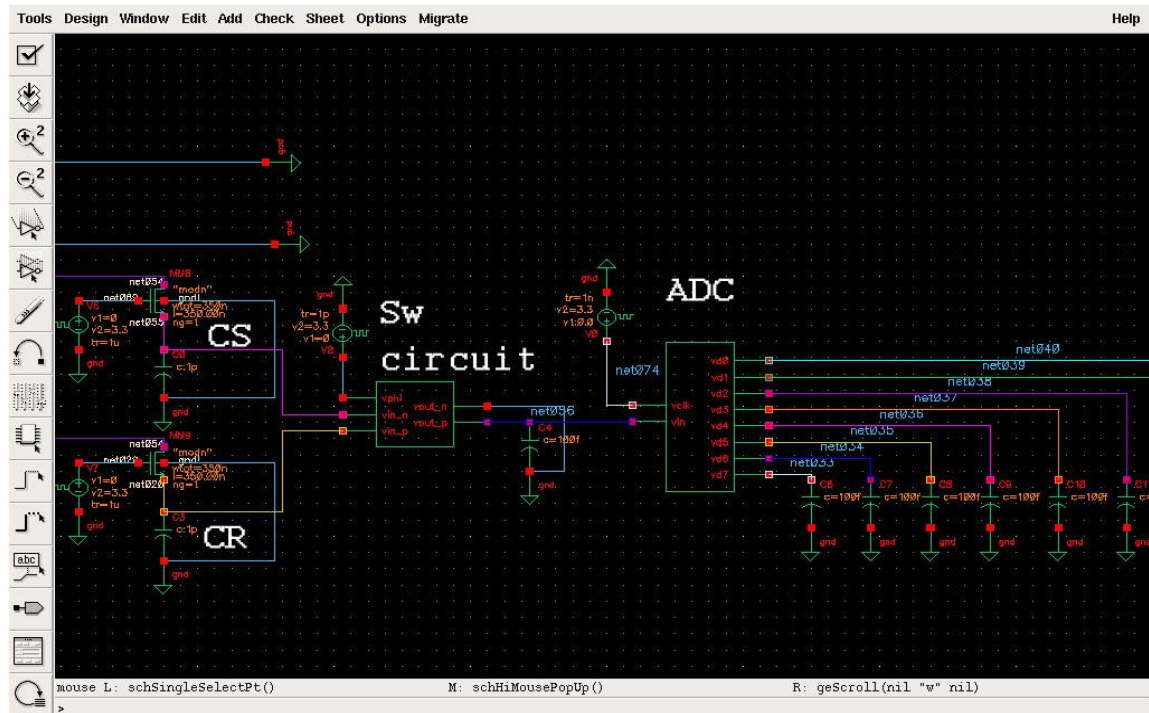
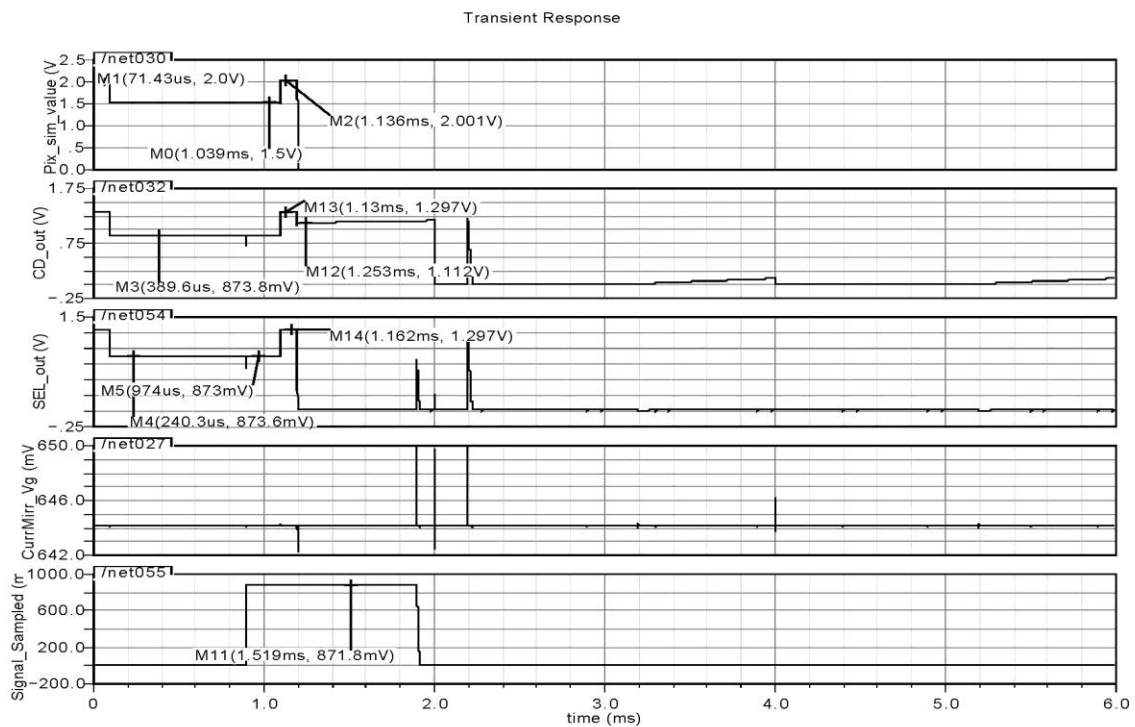


Figure 6.15. A one pixel implementation. At the end is the ADC circuit.

The one pixel test circuit was simulated with Spectre by CADENCE. The analysis is chosen to be a transient response of duration 6ms. All outputs in the plot are labeled according to their functions.



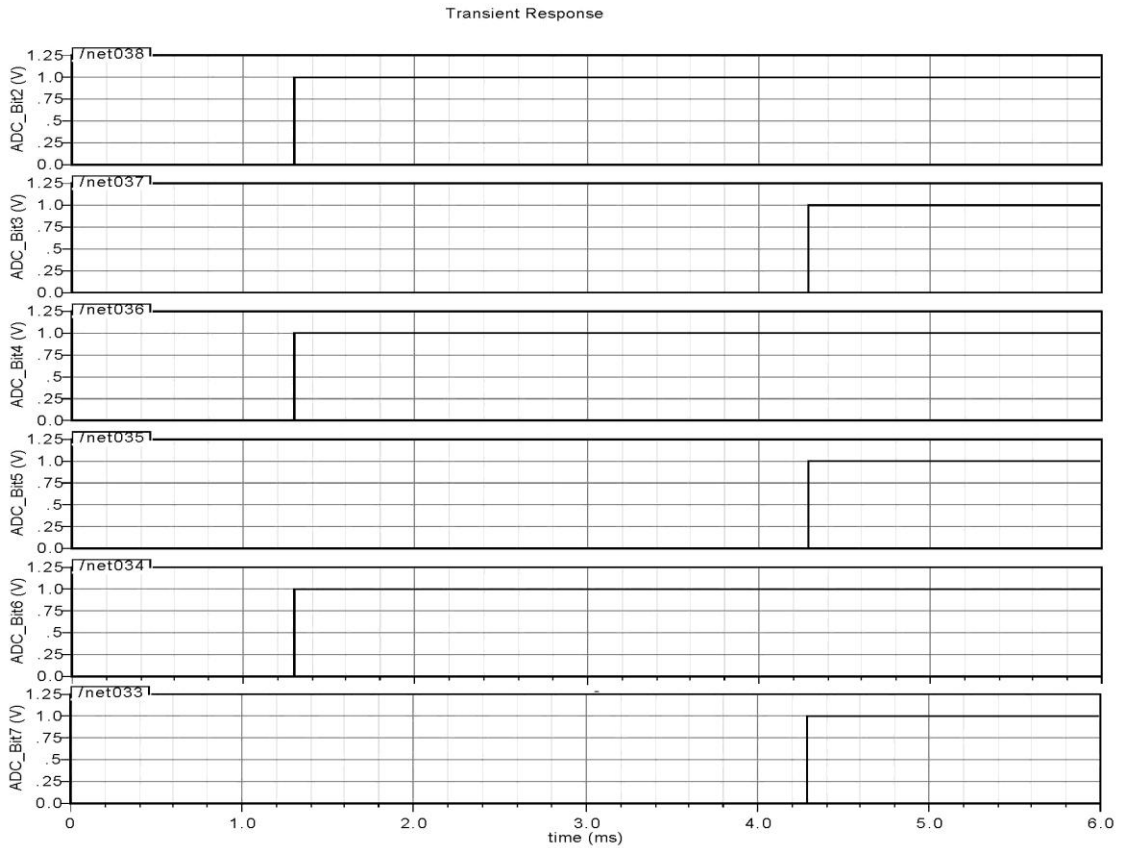
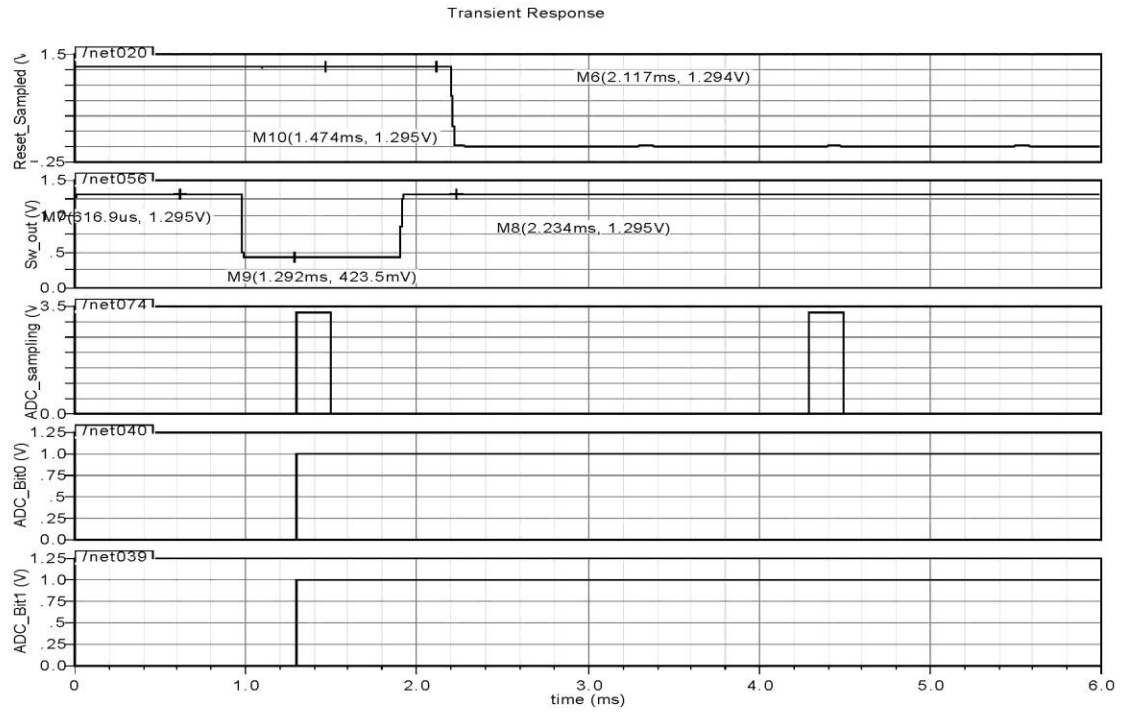


Figure 6.12. Simulation results of the one pixel system

The results clearly show that the APS system is working. To verify we look at the different building blocks of the APS and interpret the resulted plot:

- The Pd signal has duration of 1-2 ms from which 0.1ms is reserved for resetting the photodiode and an additional 0.1ms is for sampling the pd signal. The remaining 1ms is the integration time, remember the procedure for a reading out a photodiode array is to reset, integrate and read out the pixel by this sequence.
- From the start we see a problem: The source follower follows the Pd signal but with a voltage difference bigger then we initially calculated. We calculated a voltage drop of 0.6 V but in the simulation a voltage drop of 0.704 V appears, see window labeled CD_out. This “error” could not be an error at all, but more a capacitance problem due to switching (clock feedthrough phenomena). Switching activity of the SS and SR switches is the root of this problem. Unfortunately this cannot be corrected. Despite that, the CD is following the input well.
- The signals sampled onto the CR and CS are subtracted from each other with the switched capacitor circuit. This state seems to have problems again with the source follower. Reset voltage is $2V - 0.704V = 1.296V$, Signal voltage is $1.5 - 0.6282 V = 0.8718 \text{ mV}$, difference with Sw circuit is found to be $1.296.0.8718 V = 0.4235 \text{ mV}$ (see Signal_sampled, Reset_sampled and Sw_out).
- The ADC samples the output of the sw circuit after the pd signal is sampled, hence we set the ADC to sample at about 1.2 ms.
- The result of the ADC must be an 8-bit representation of the sw output. Scanning through ADC_Bit7 to ADC_Bit0 we find the word to be $01010111_{(bin)} \rightarrow 87_{(dec)}$. The maximum value of the source follower is 1.296 V, this is the reference voltage of the ADC, then the V_{LSB} (see chapter five) is $V_{LSB} = \frac{V_{ref}}{2^N} = \frac{1.296}{2^8} = 0.0050625 V$. The digital form of the sw output is then $\frac{0.4235}{0.0050625} = 83.6$. The ADC allows only integers and hence $83_{(dec)} \rightarrow 1010011_{(bin)}$.
- The conclusion is that voltage drops in the source follower followed by potential capacitor leakages makes the one pixel APS system to not work correctly. The error is acceptable when one compares the theoretical result (87) with the one we got from simulation (83). We simulated the above system many times and presented here the worst case scenario which is results in an error off 4 codes.

6.3.4 2x2 Pixel

To make a more realistic simulation and to discover possible problems when building an array of pixels, we constructed a tiny APS sensor. It consists of 4 pixels laid out on a matrix of 2 rows by 2 columns. We saw back in chapter four that's is generally not a good idea to “steal” space from the sensing device (photodiode) and combining the results from the one pixel system, we decided to take the sampling process out of the photodiode and into the switched capacitor. With other words instead of sampling the reset and pd signal in the pixel, it is sampled at the bottom of the column, before the switched capacitor circuit. We also have to change the way we sample the signals, since we took the sampling process out of the pixel. As a result we change the way the select transistor (switch) samples. The select transistor in the pixel must sample now signals, once the reset and once the pd signal. The reset signal is at the beginning of the APS

function, then follows a 1 ms shut off (integration time) and finally we sample again. The select clock, which commands the sampling transistor (SEL), must be different or delayed for each row. The reason for that is to make a rolling out scheme, remember after row one is been reset, row two starts resetting while row one integrates. To make things easier we have build components out of the basic building blocks of the APS, seen in Figure 4.13 (b) of chapter 4 (see also one pixel system). We chose to “wrap” the following circuits into components:

- a) Pixel (CD, SEL transistor)
- b) Ibias (Current mirror)
- c) CDS&ADC (sampling circuitry for reset and pd signal, sw circuit and the 8-bit ADC)

In the following pages we present the above mentioned components

- a) Pixel

Schematic:

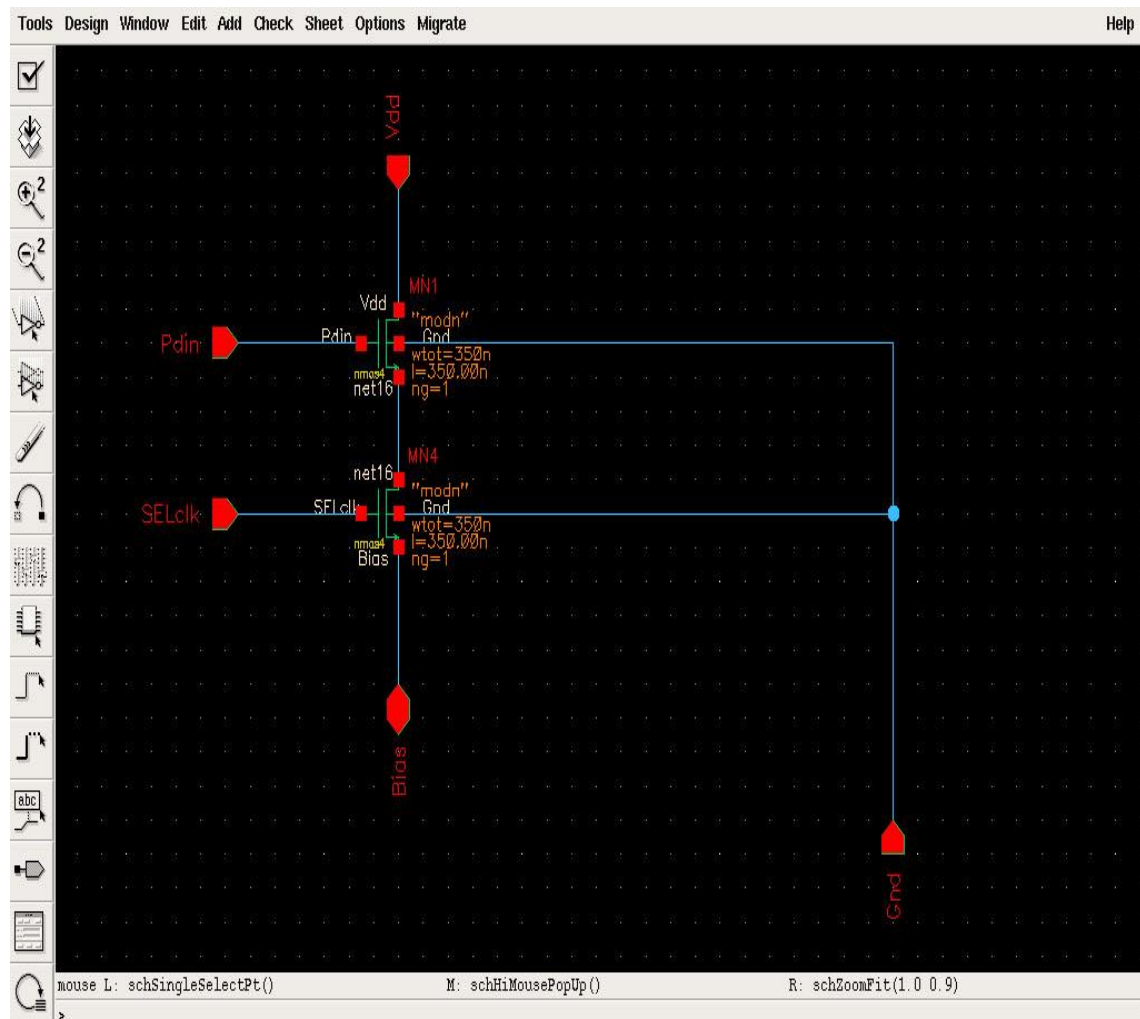


Figure 6.16. Schematic of Pixel

Component:

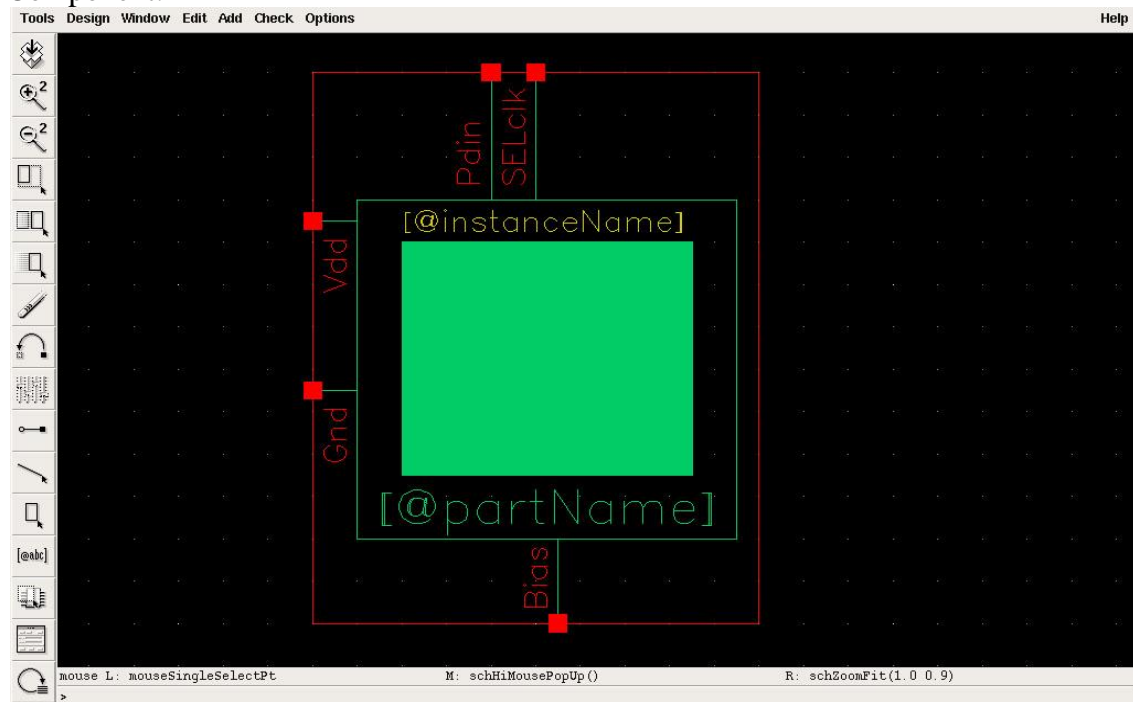


Figure 6.17. Component of Pixel Schematic.

b) Ibias

Schematic:

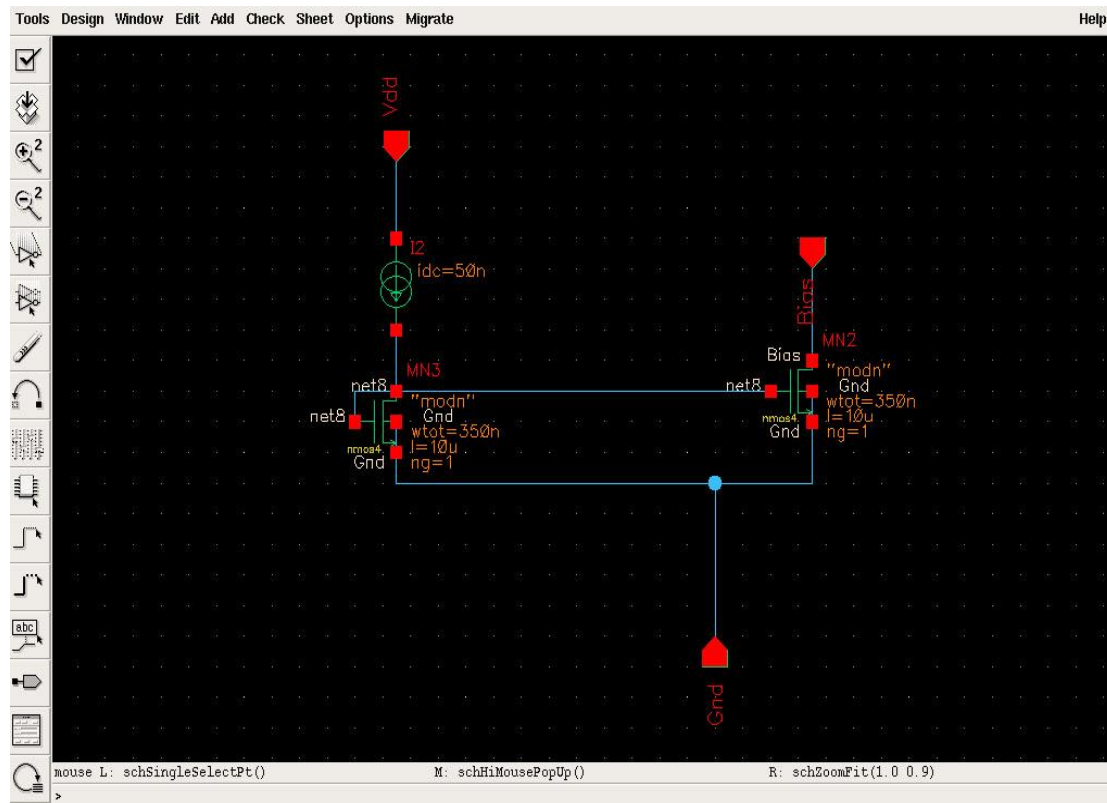


Figure 6.18. Schematic of Bias

Component:

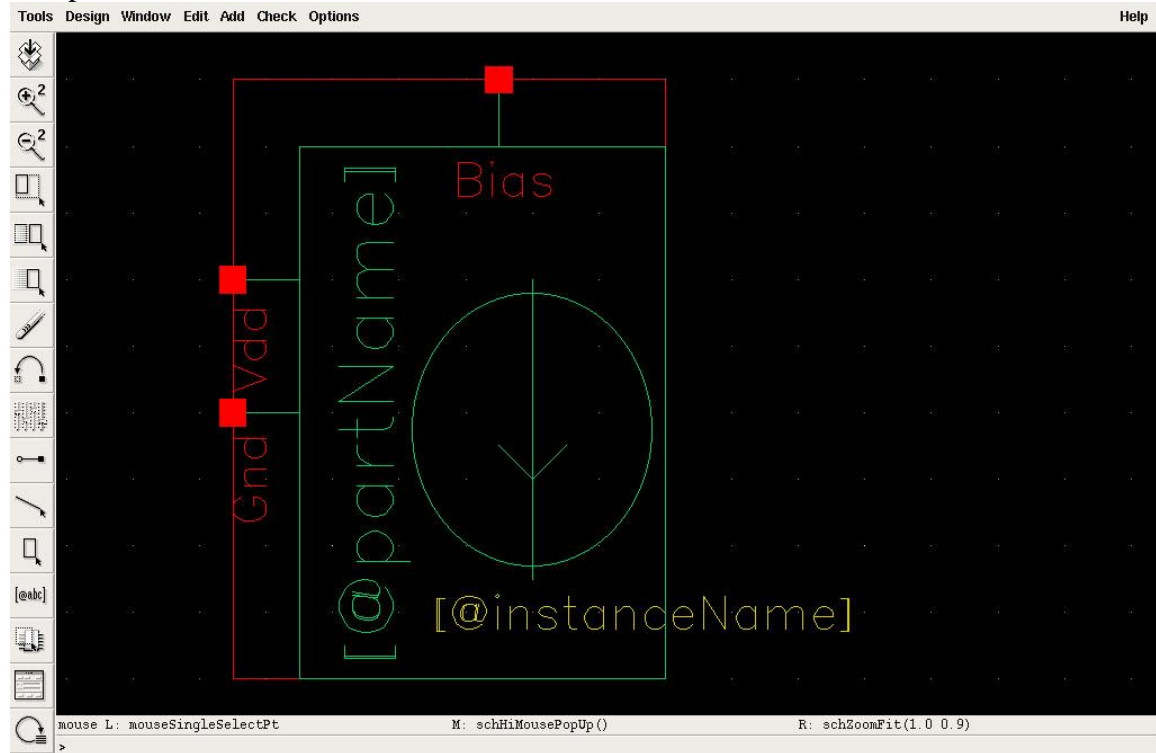


Figure 6.19. Component of Bias Schematic

c) CDS&ADC

Schematic:

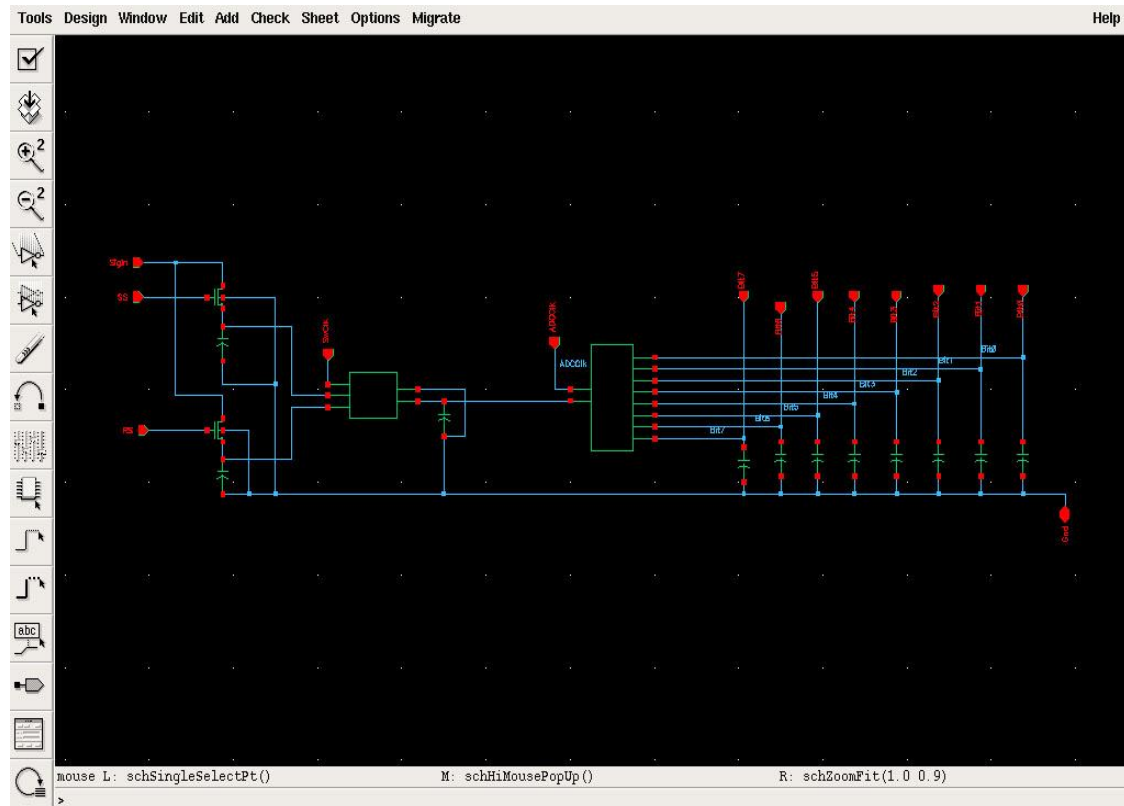


Figure 6.20. Schematic of CDS and ADC

Component:

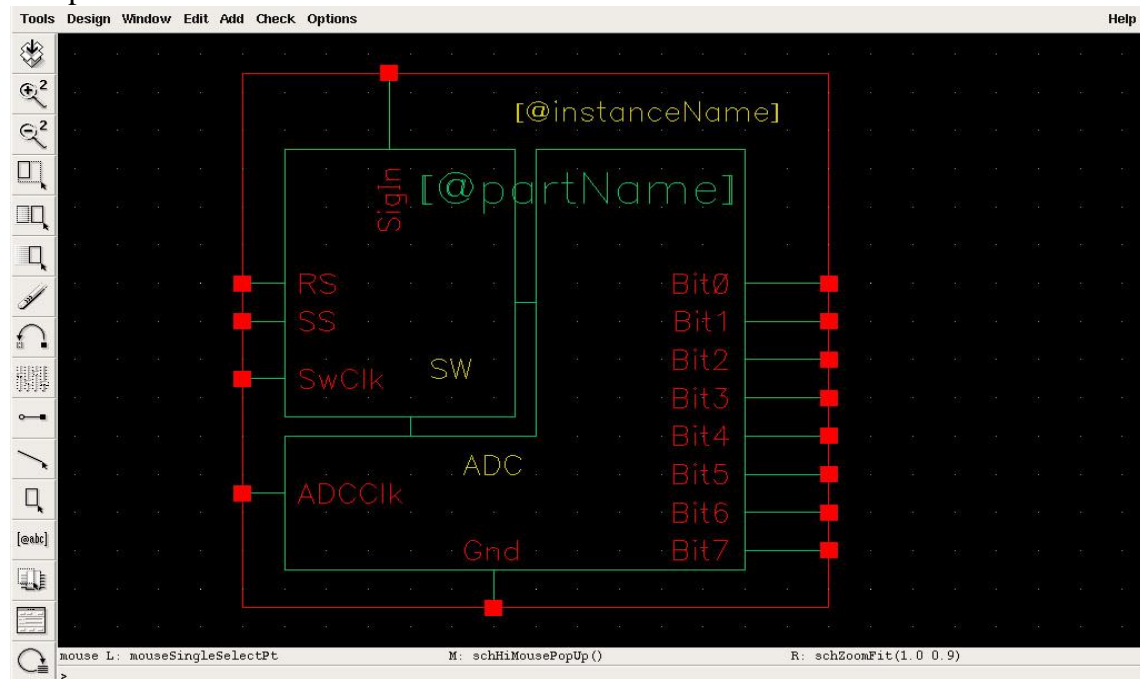


Figure 6.21. Component of CDS and ADC

Now that we have a matrix of pixels, a control unit must be build to drive the necessary clocks of the different pixels high and low at right times. For simulation purpose we build

a simple circuitry containing pulsed voltage sources and piecewise linear voltage sources. These sources are the easiest way to control the array, of course for bigger arrays there is no way around and one must build a proper control unit. The roll out procedure discussed in chapter 4 must be implemented, and for this purpose the clocks must be set correctly. In the roll out scheme first all pixels in the first row are reset, integrated and finally read out. Shortly after the pixels in the first row are reset the pixel in the second row follows the same procedure. We made the control circuit a component named controlclk:

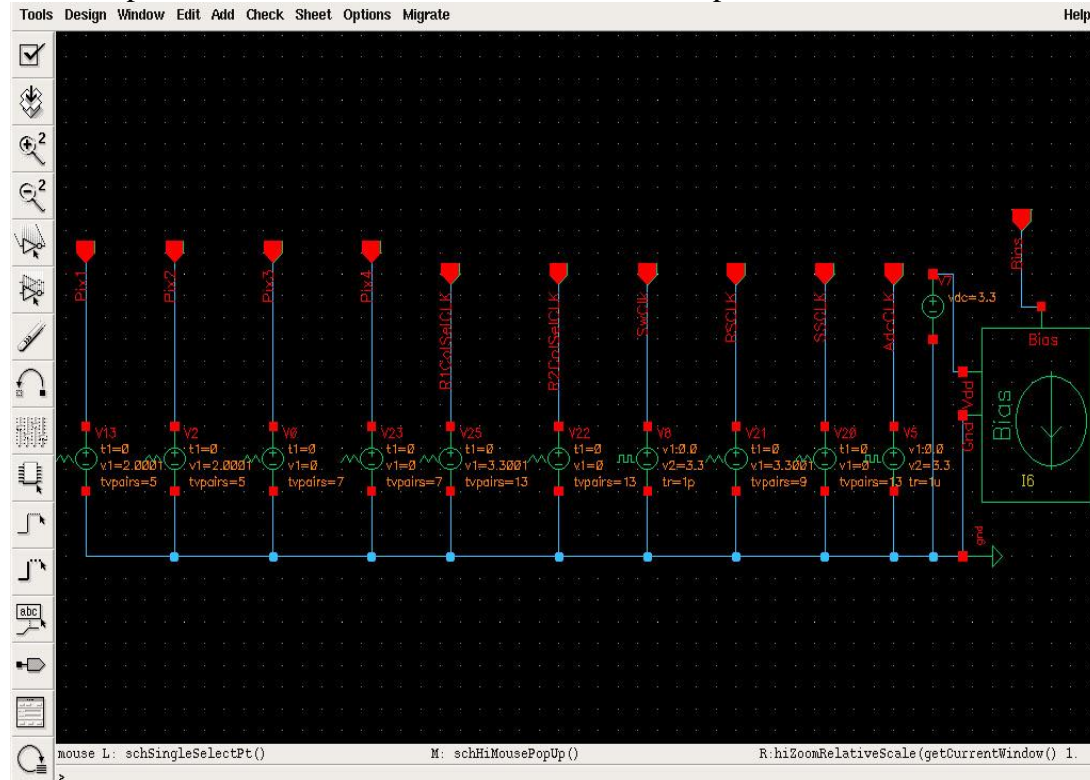


Figure 6.22. Schematic of the control unit

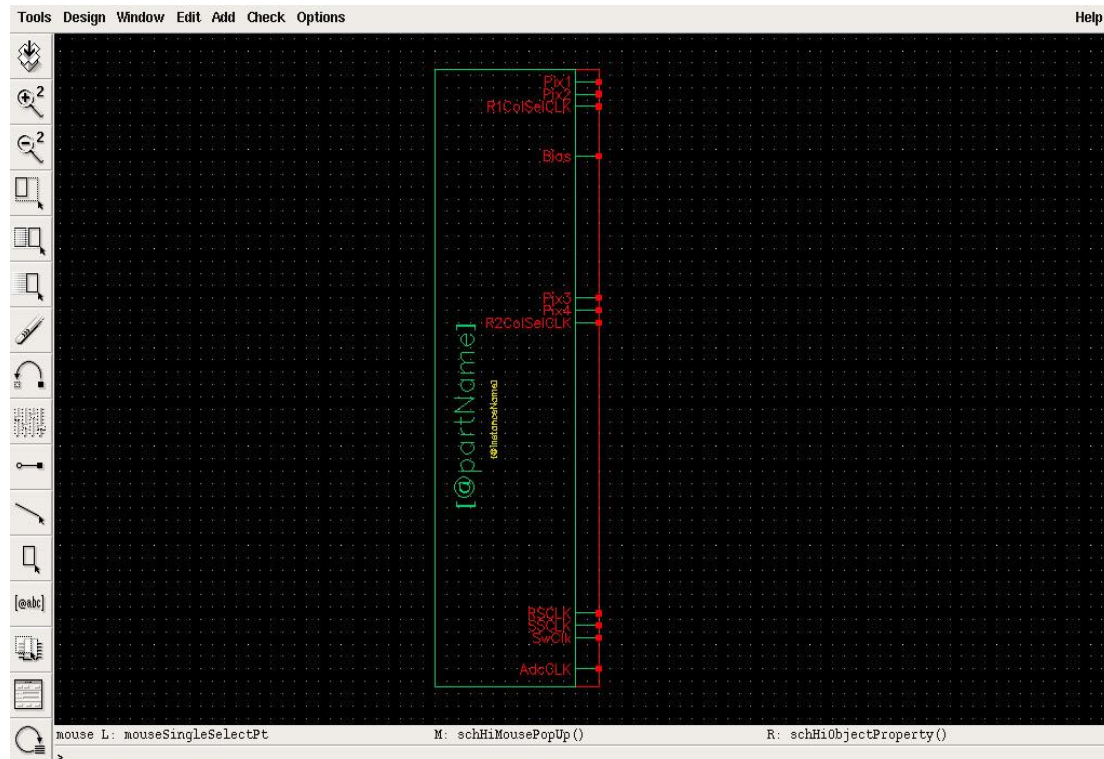


Figure 6.23. Component of the control unit

Finally we combine all components in one schematic, the 2x2 pixel schematic. Because of its size we place different views of the schematic.

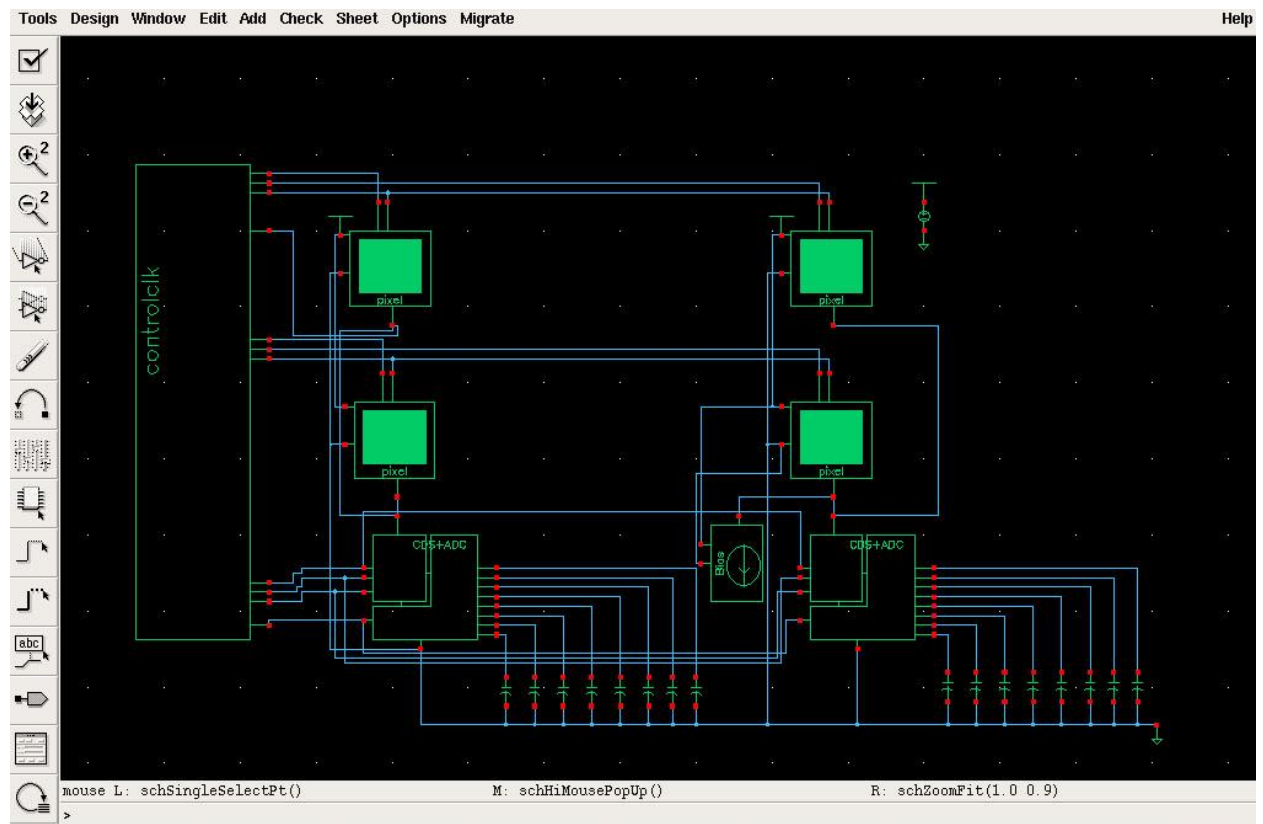


Figure 6.24. Overview of the 2x2 pixel array

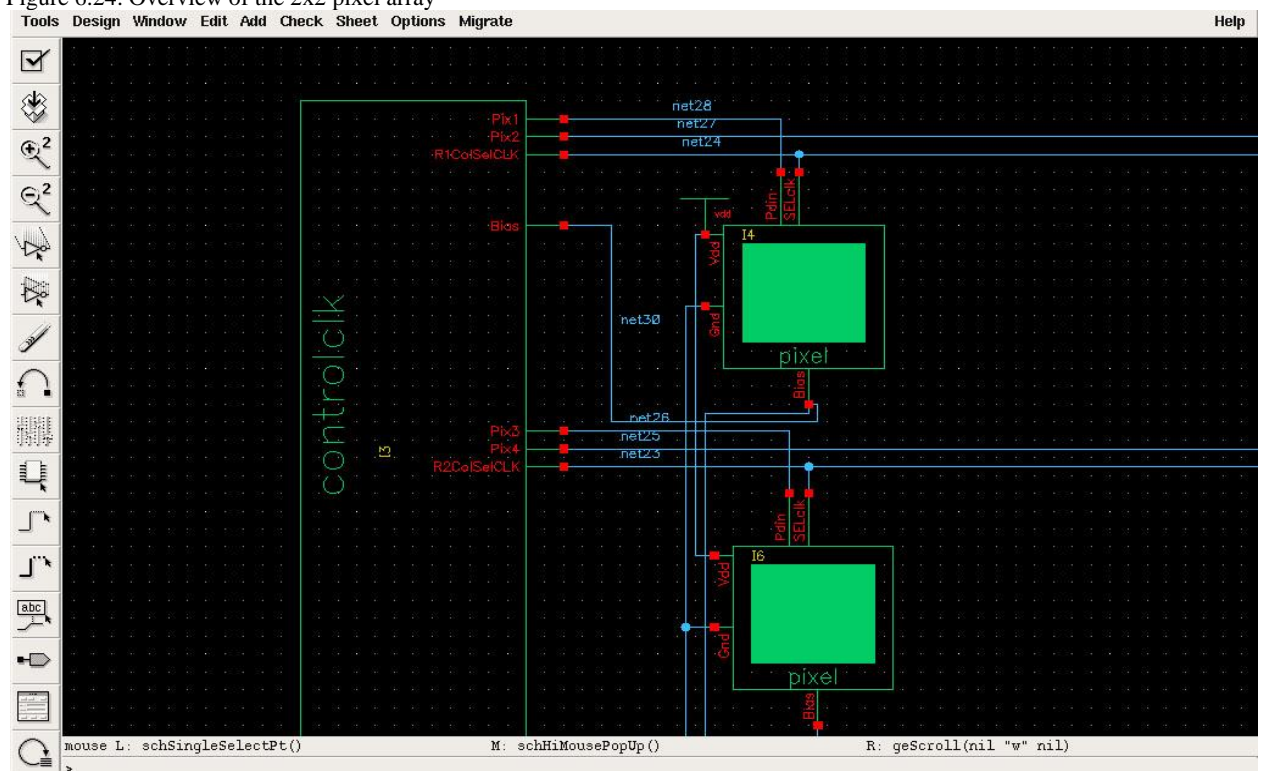


Figure 6.25. Part of the 2x2 pixel array

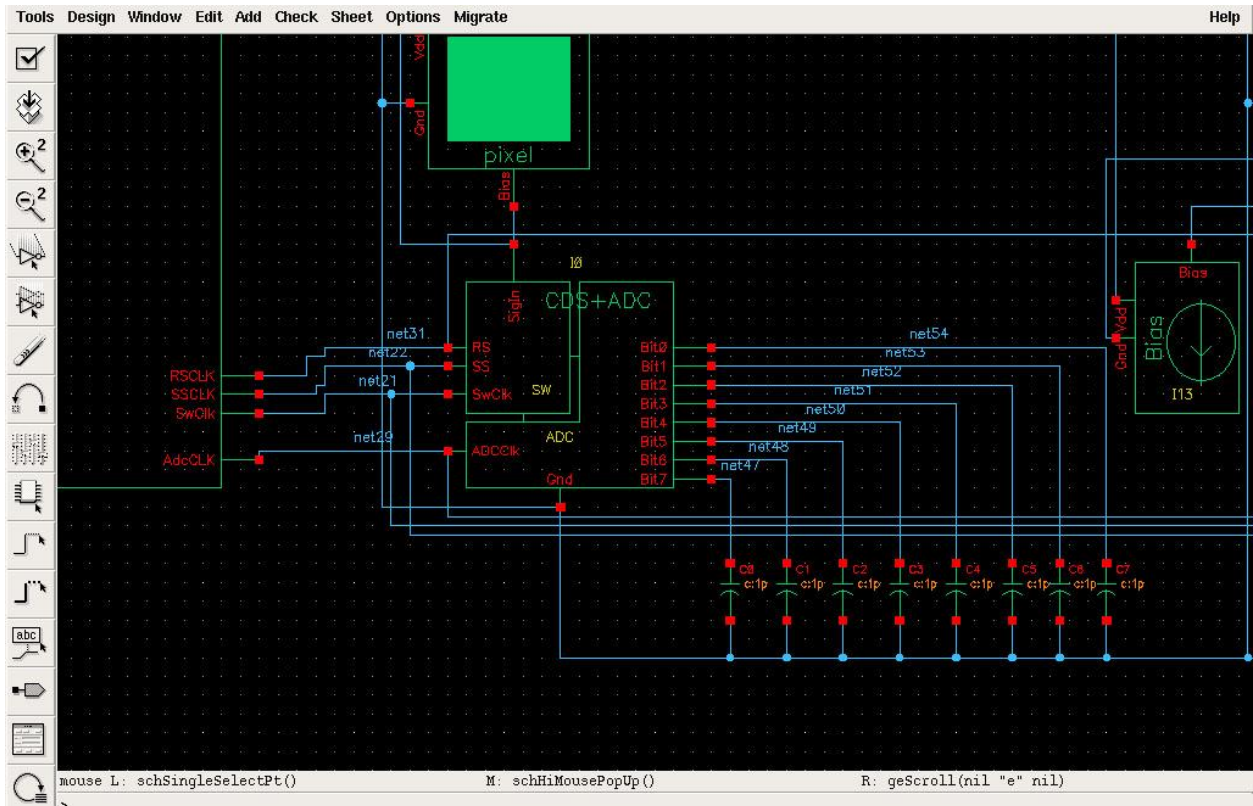


Figure 6.26. Part of the 2x2 pixel array

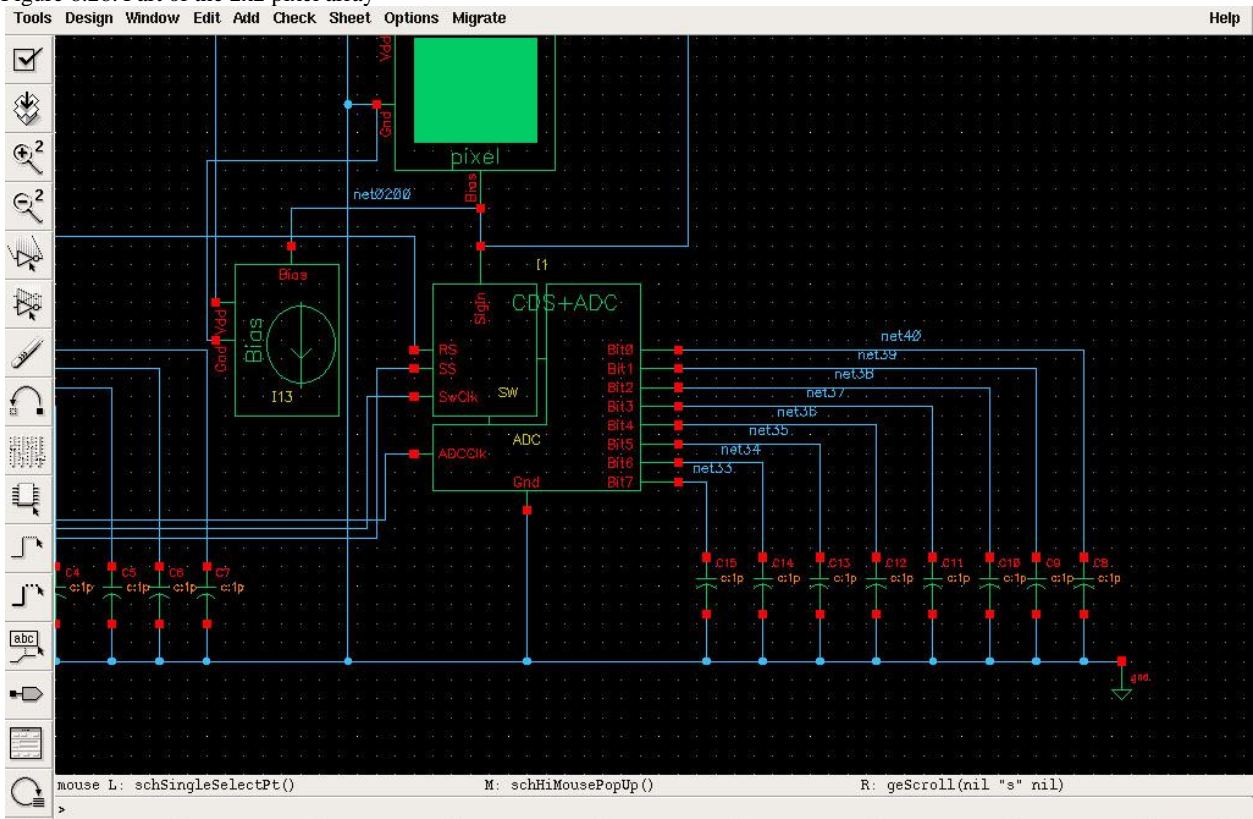


Figure 6.27. Part of the 2x2 pixel array

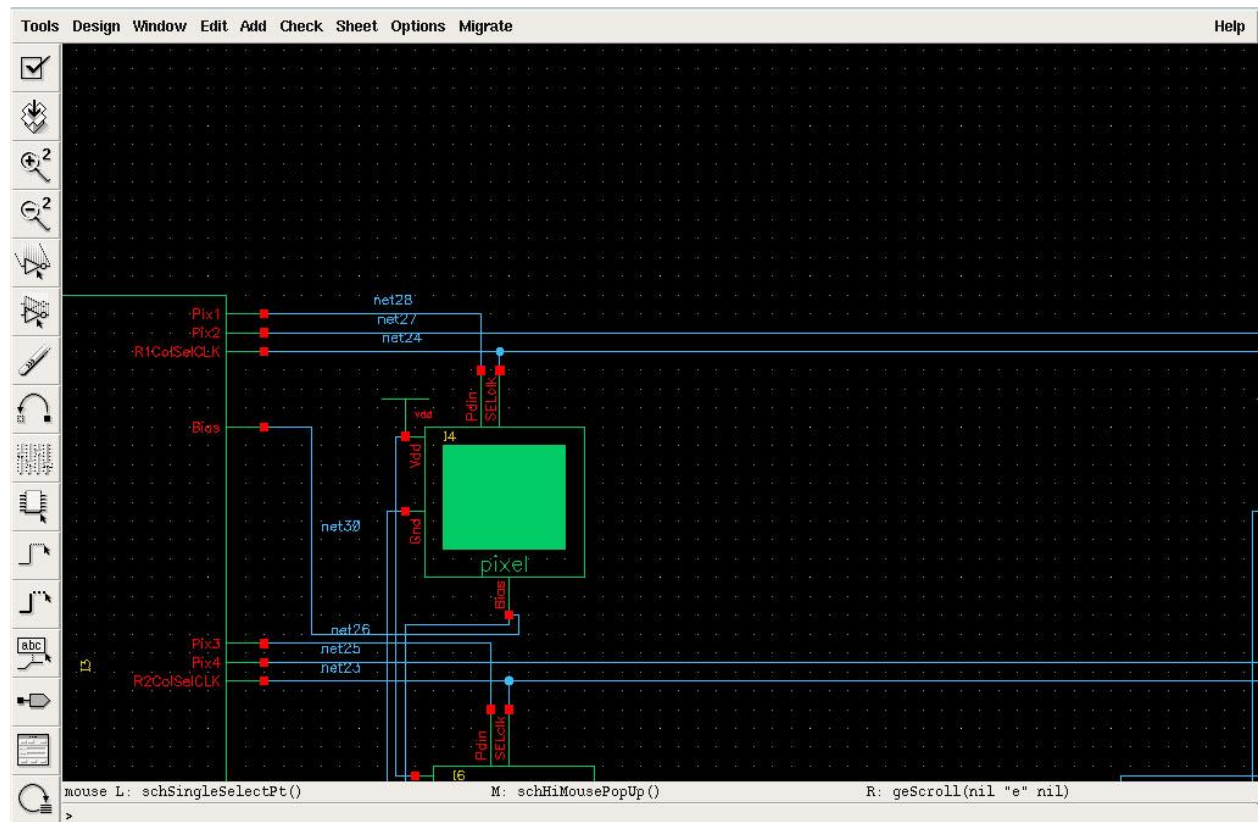
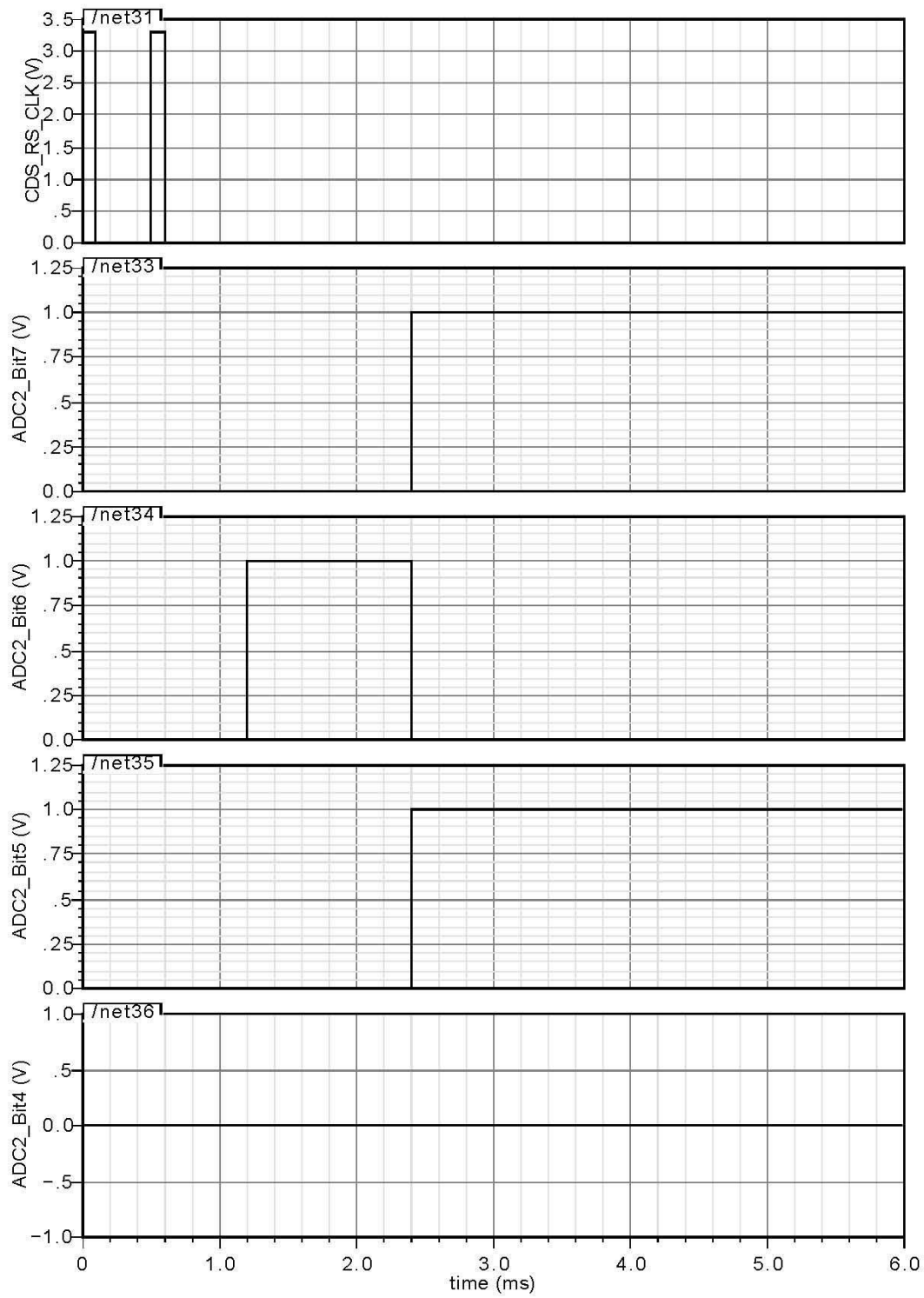


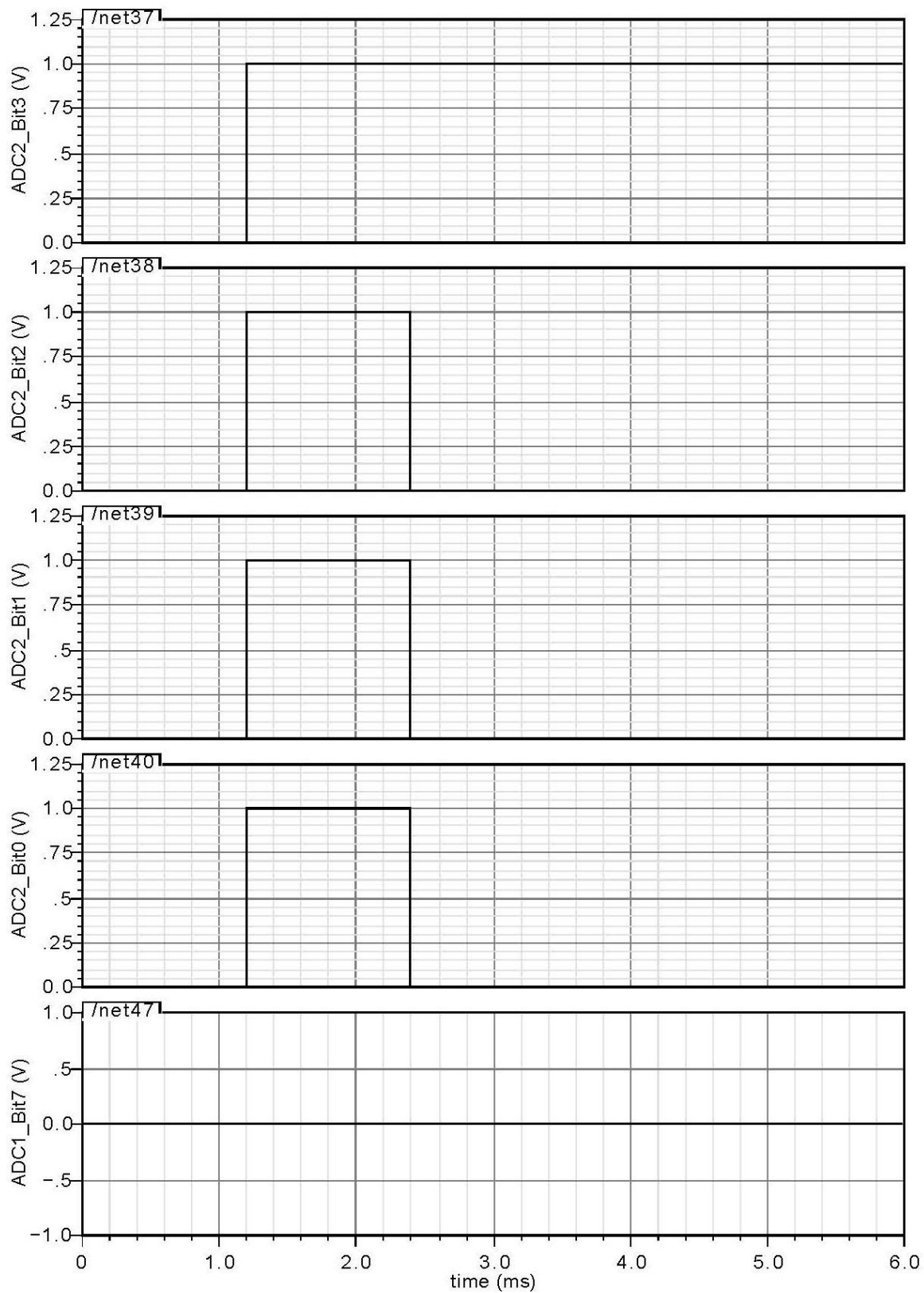
Figure 6.28. Part of the 2x2 pixel array

We simulated the 2x2 pixel with Spectre and chose an transient response analysis of 6 ms. The plots of the simulation are shown in the following pages:

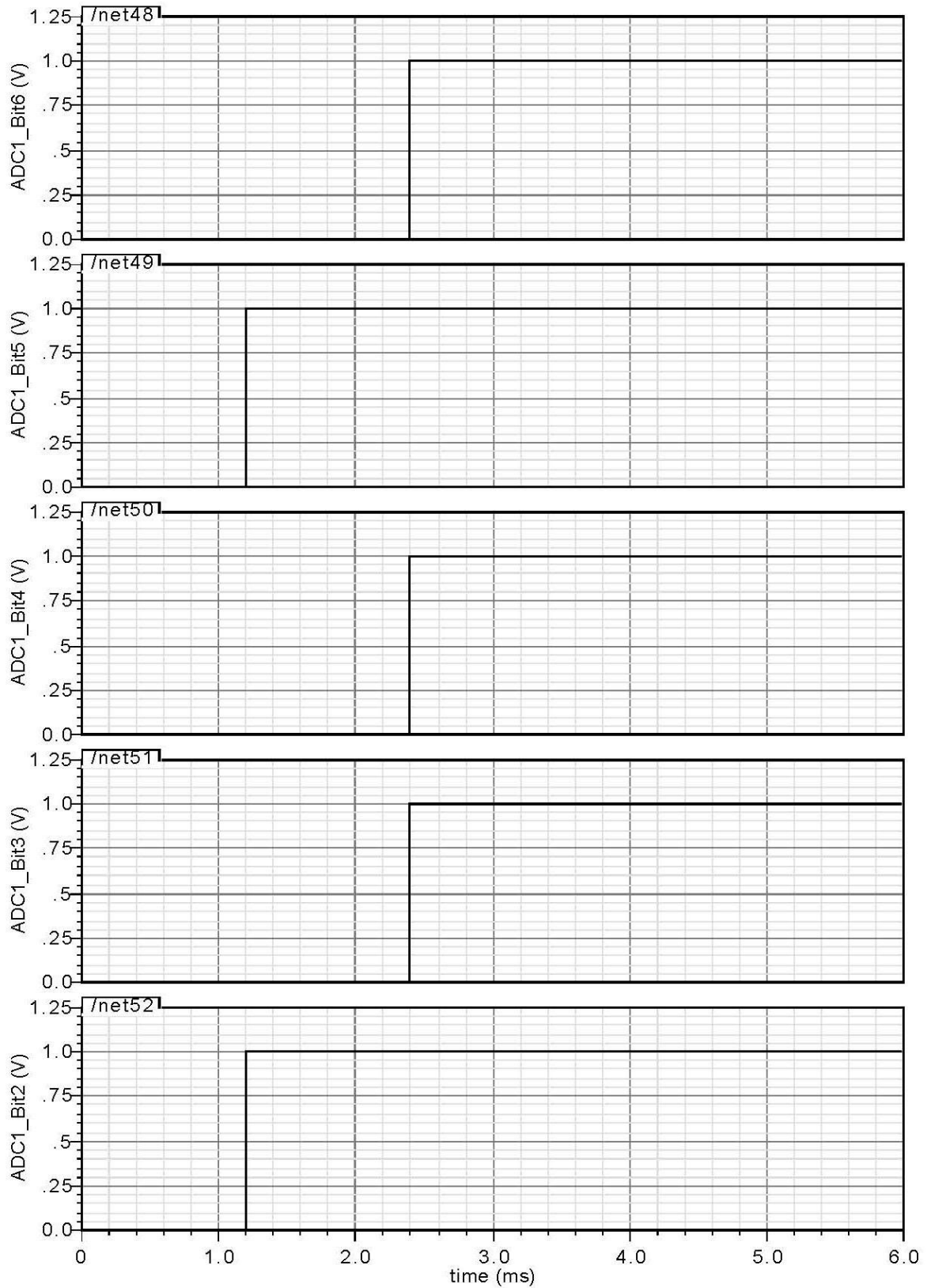
Transient Response



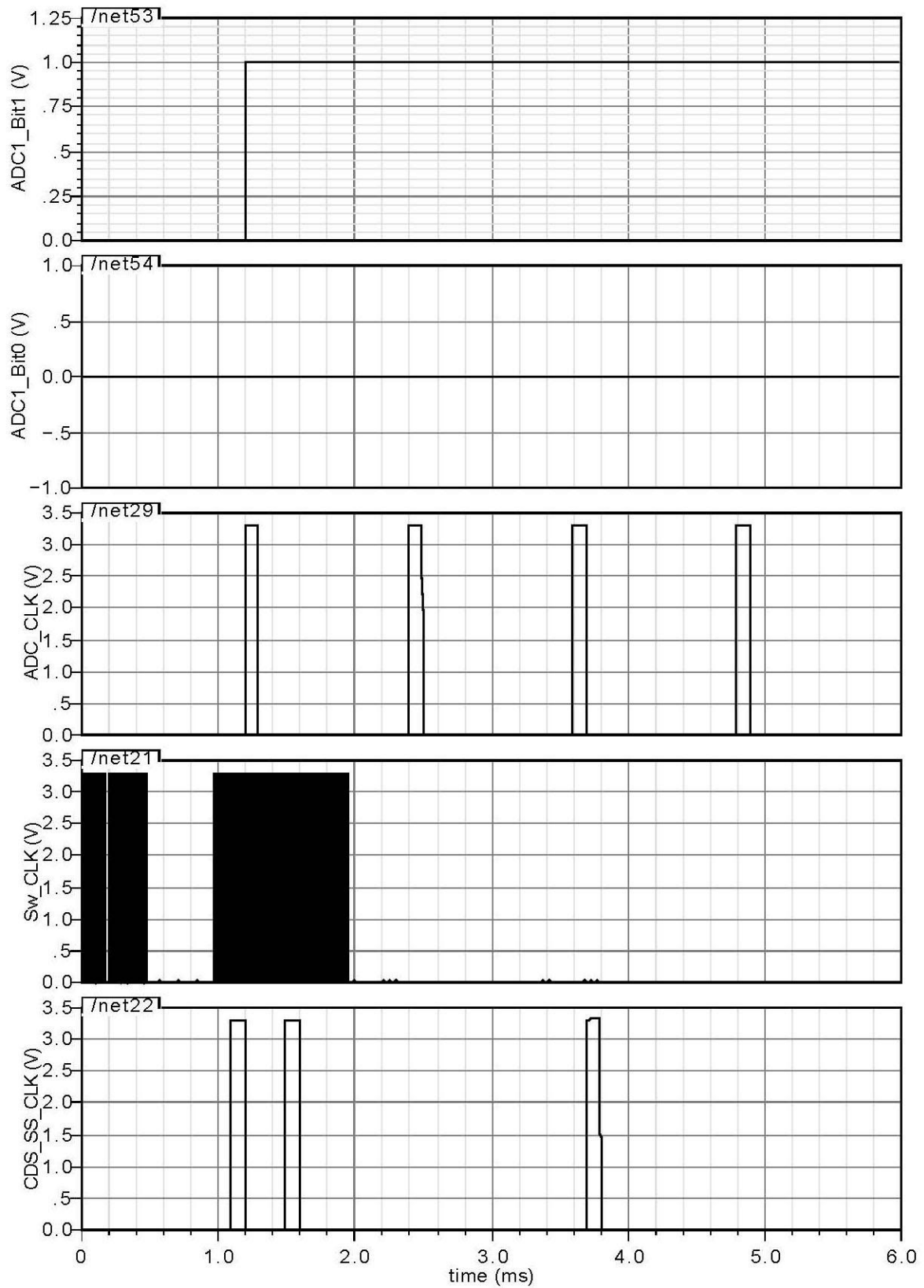
Transient Response



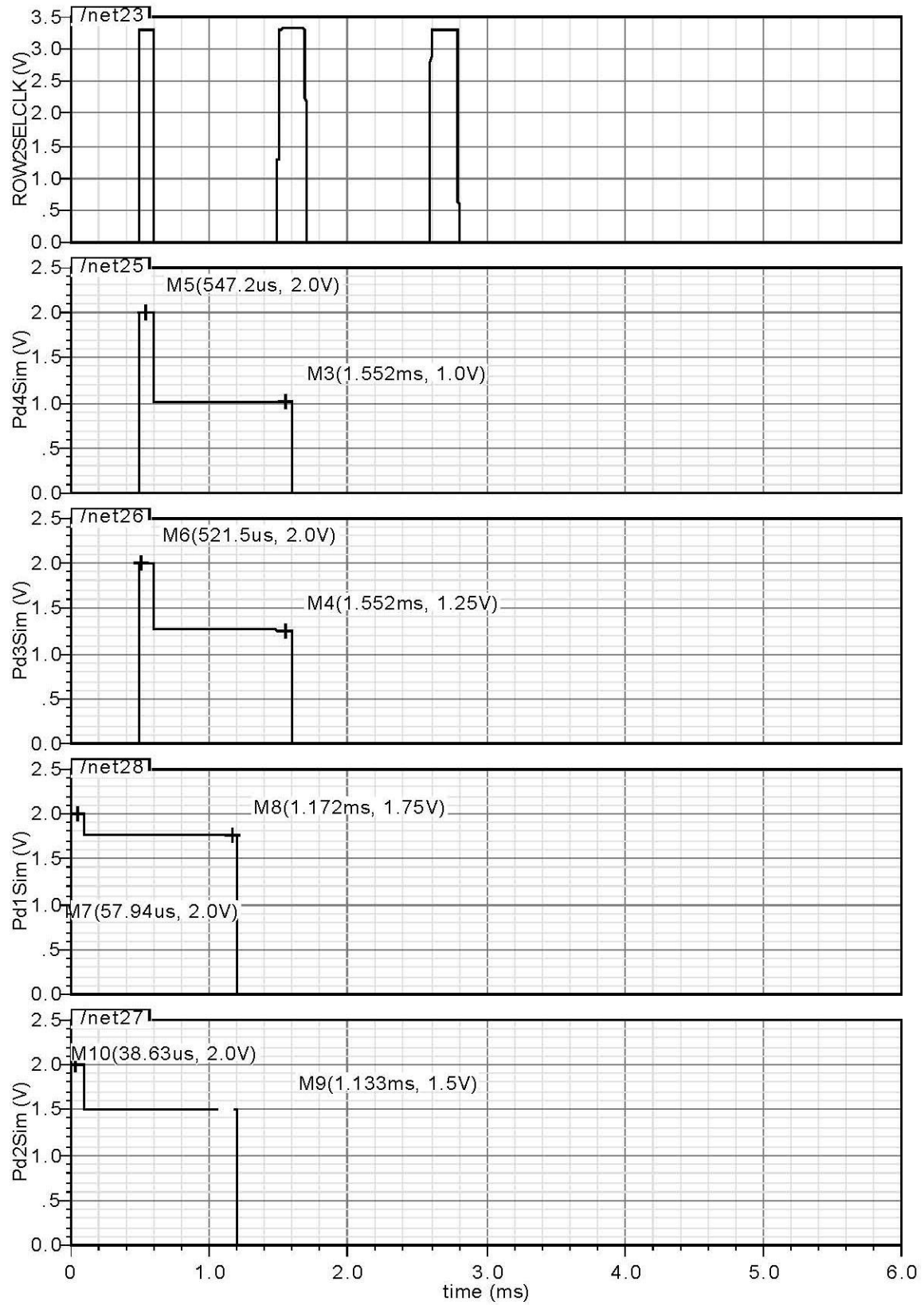
Transient Response



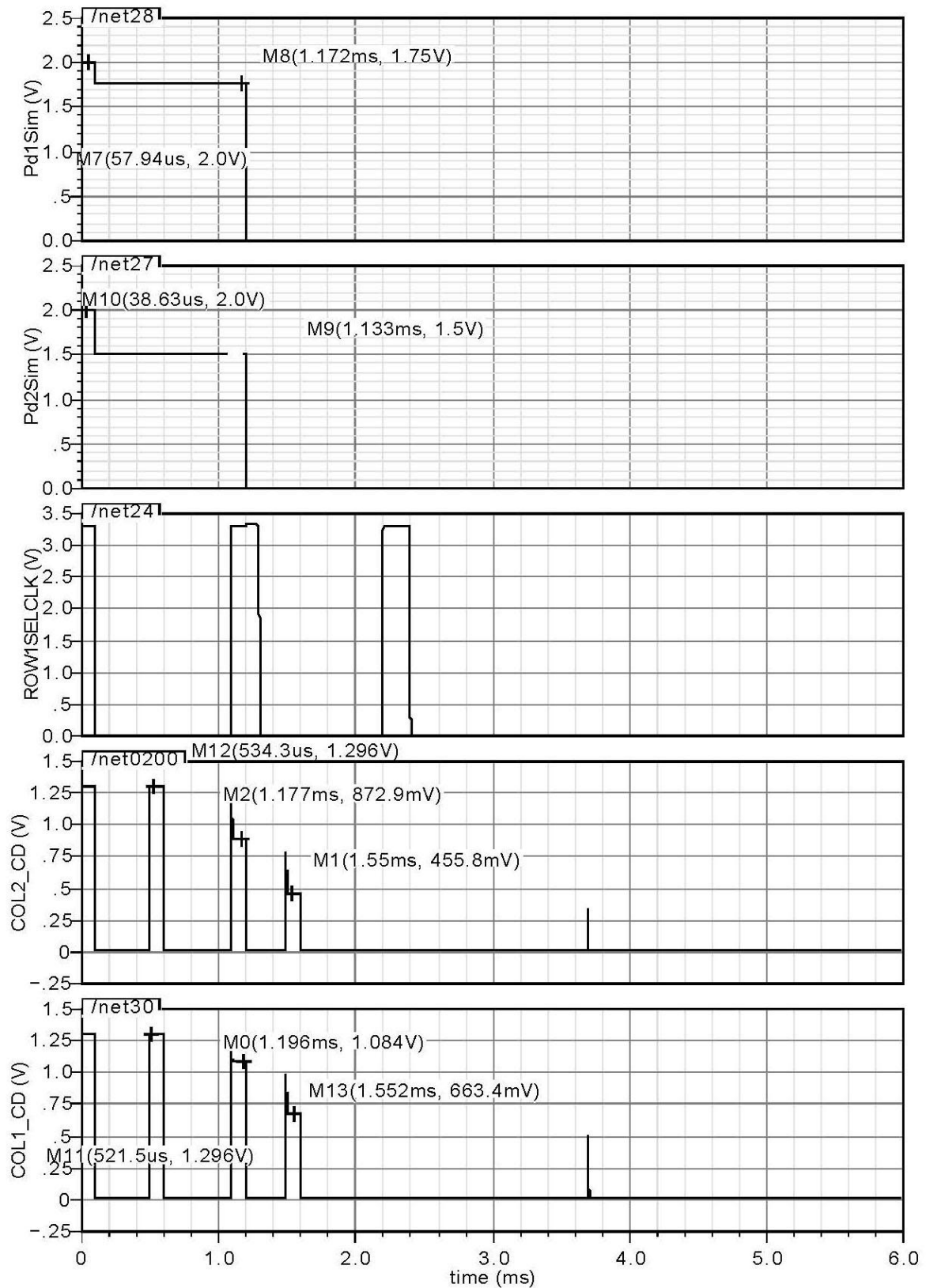
Transient Response



Transient Response



Transient Response



Looking at Figure 6.21 the top left pixel is pixel one, at top-right we have pixel two, at bottom left we got pixel three and at bottom right we have pixel four. PdXSim is the simulated value of the sensing node, which is the input to the source follower. PdXSim is seen on the plot for the four different pixels and have the following name convention Pd1Sim, Pd2Sim, Pd3Sim and Pd4Sim. It worth to mention that the clock signals for sampling are very important, they must be set correctly or the 2x2 array will produces garbage values. To verify we check if the output string of the ADC's match the difference seen in the source followers. The switched capacitor circuit output is not available anymore to us since it is part of the component CDS&ADC. Therefore we use the output of the CD and do a little math to find the difference between the pd signal and the reset signal. To clarify the situation we present the verification procedure in a table form:

Reset (V)	Signal (V)	A D C o u t	Differ- ence (V)	Difference $\frac{\text{Difference}}{V_{\text{LSB}}}$ $V_{\text{ref}}=(0.0050625)$	Binary ("theoretical ")	Binary (plot "found")	Dec (plot)
Pd1Sim (Pixel 1)							
COL1_CD= 1.296	COL1_CD=1.08 4	1	0.212	41.876543→ 41	00101001	00100110	38
Pd2Sim (Pixel 2)							
COL2_CD= 1.296	COL2_CD= 0.8729	2	0.4231	83.575308→ 83	01010011	01001111	79
Pd3Sim (Pixel 3)							
COL1_CD= 1.296	COL1_CD= 0.6634	1	0.6326	124.95802→ 124	01111100	01111110	126
Pd4Sim (Pixel 4)							
COL2_CD= 1.296	COL2_CD= 0.4558	2	0.8402	165.96543→ 165	10100101	10101000	168

Table 6.8. Verification of the data output of the ADC, calculated "theoretically" and compared to the simulation data obtained from the plots³

The results from Table 6.8 direct us to the same conclusions that we had made in the one pixel test, namely that we have an error off maximum 4 codes.

A last remark must be made on how the the 2x2 APS sensor is detecting the incoming x-ray. The CMOS APS system is designed for intra oral x-ray imaging and cannot be connected via cables to the x-ray tube. So how do we detect the x-ray? For this purpose a detection area on the sensor is been reversed. The detection area will contain a detection block which will start the controlclk signals and the controlclk component triggers the read out of the pixels. The detection block is simply another pixel which is not covered with scintillator, but with a material that blocks, not all, but most of the x-rays. A proposed detection block is one described by Fossum [21], see figure 6.29:

³ V_{LSB} is 0.0050625V

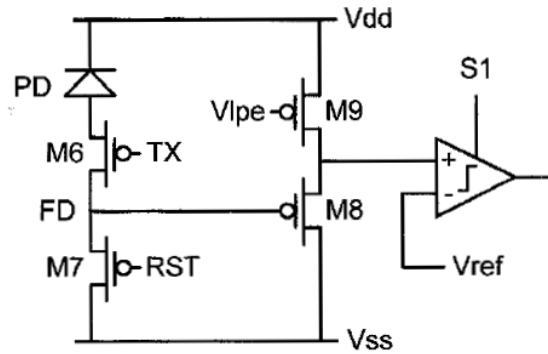


Figure 6.29. X-ray detection block

6.4 References

- [1] Moses, William W., "Scintillator requirements for medical imaging." *Scintillator requirements for medical imaging*. Berkeley, California, United States of America : University of California.
- [2] Eempicki, A., "THE PHYSICS OF INORGANIC SCINTILLATORS." *Journal of Applied Spectroscopy*. 1995, Vol. 62, 4.
- [3] Graham, Donald T and Cloke, Paul J., *Principles of Radiological Physics*. s.l. : Elsevier Health Science, 2003. 0443070733.
- [4] Ramos, N F, et al., "CMOS X-ray Imager for Dental Radiography." Bucharest : International Conference on Information Society Technologies for Broadband Europe, 2002, pp. 326-326.
- [5] Rocha, J G, et al., "Scintillating microcavities for x-ray imaging sensors." *MicroMechanics Europe Workshop*. September 2006.
- [6] Rocha, J G, et al., "X-ray detector based on a bulk micromachined photodiode combined with a scintillating crystal." *Micromechanics Microengineer*, s.l. : Institute of physics publishing, June 13, 2003, pp. S45-S50.
- [7] Kleinmann, P, Linnros, J and Peterson, C S., "An x-ray imaging Pixel based on a scintillating guides screen." *IEEE TRANSACTION ON NUCLEAR SCIENCE*. August 2000, Vol. 47, 4, pp. 1483-1486.
- [8] Rocha, J G, et al., "Comparison between bulk micromachined and CMOS x-ray detectors." *Sensors and Actuators*, s.l. : Elsevier, April 24, 2004, Issue 115, Vol. A, pp. 215-220. 10.1016.
- [9] Parks, Edwin T and Williamson, Gail F., "Digital Radiography:An overview." *The journal of contemporary dental practice*, November 15 2002, Issue 4, Vol. 4.
- [10] Pecht, Orly Yadiot and Cummings, Ralph Etienne., *CMOS imagers: From phototransduction to image processing*. Dordrecht (NL) : Kluwer academic publisher, 2004.
- [11] Pecheux, Francois, Christophe, Lallement and Alain, Vachoux., "VHDL-AMS and Verilog-AMS as alternative hardware description languages for efficient modeling of multidiscipline systems." *IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN OF INTEGRATED CIRCUITS AND SYSTEMS*, February 2005, Issue 2, Vol. 24, pp. 204-225.

- [12] Kundert, Kenneth S and Zinke, Olaf., *The designer's guide to Verilog AMS*. Boston, Dordrecht, London : Kluwer Academics Publisher's, 2004. 1-4020-8044-1-
- [13] Baschiroto, Andrea., "Charge injection and Clock feedthrough analysis." Lecce : University of Lecce, 2000.
- [14] Sansen, Willy M C., *Analog design essentials*. s.l. : Springer, 2006. 100-387-25746-2.
- [15] Razavi, Behzad., *Design of Analog CMOS Integrated Circuits*. s.l. : McGraw Hill, 2001- 0-07-238032-2.
- [16] Bucher, Matthias, Lallement, Christophe and Krummenacher, Enz C., "Accurate MOS Modelling for Analog Circuit Simulation Using the EKV Model." *IEEE Int. Symp. on Circuits & Systems (ISCAS'96)*. May 1996, pp. 703-706.
- [17] Bucher, Matthias., "MOS model for analog design of CMOS integrated Circuits." *lecture 2*. Xania : RFIC lab of Technical University of Crete, 2007. www.rfic.tuc.gr.
- [18] Bucher, Matthias, Kazazis, Dimitris and Krummenacher, François., "Geometry- and Bias-Dependence of Normalized Transconductances in Deep Submicron CMOS." *Workshop on Compact Models, NANOTECH 2004*. March 8-11, 2004.
- [19] Gamal, El A., "Fixed Pattern Noise." *Lecture Notes 6*. s.l. : University of Yale, 2005.
- [20] Tian, Hui, Fowler, Boyd and Gamal, Abbas El., "Analysis of temporal noise in CMOS photodiode active pixel sensor." *IEEE, s.l. : J. Solid-State Circuits*, Jan 2001, Issue 1, Vol. 36, pp. 92-101
- [21] E. R. Fossum, R. H. Nixon, D. Schick, "A 37x28mm 2 600k-Pixel CMOS APS Dental X-ray Camera-on-a-Chip with Self-Triggered Readout", *ISSCC*, Feb 1998, Issue 3, Vol. 11.

Chapter 7. Conclusions

This assertion dealt with the study of the requirements needed for a CMOS image sensor. Most information presented is about the different parts that are necessary to construct the essential parts of a CMOS image sensor. This was a difficult task, even if the CMOS image sensors exist commercially for more than ten years now, no complete description of them exist. Most authors focus on the pixel cell of the CMOS image sensor or on any other CMOS based pixel implementation, but do not describe the sensor as a whole.

We described in chapter 2 the photodiode interaction with illumination and some construction aspects. Photocurrent and capacitance of the photodiode were also discussed. Important conclusions for the photodiode characteristics can be made, for example the magnitude of the photocurrent must be quite high (10^{-6} A) while the dark current must be low (10^{-15} A). Another conclusion is that for visible light detection the junction depth of the photodiode must be shallow. These are important features that a designer must know, after all the photodiode is the heart of the sensor. One must note that these and other important aspects described in chapter 2 and in appendix A are hard to find. Furthermore we analyzed other sensing devices like the photogate which can also be employed into a CMOS process pixel. Hence we mentioned various sensing devices so that one can choose freely among sensing devices without having narrowed down the analysis to only one device. In chapter 6 we even showed a verilog-ams code implementation of the photodiode, which can be integrated in almost any simulator, without having the need to actually construct or order a photodiode. Despite not having no success with the verilog-ams code in CADENCE it can be very useful to any designer that wants to approach the photodiode on a pseudo-implementation level.

In chapter 4 we saw how the information rate impacts on the read out mode of the CMOS APS and PPS pixel. Choices are made according to the best information rate available. We showed that the charge mode is best rate, but this isn't necessary. Current mode is lately also used. Furthermore we showed how the APS array system is built and that read out is performed by the rolling out technique. The rolling out technique gives clues on how to implement the control circuit in the APS system.

In chapter 5 we analyzed almost every ADC that can be employed with a CMOS APS. Analyzed were quantization errors, component mismatches, speed and size characteristics. A trade off exists between speed and size, more speed means bigger size and bigger size means more mismatches. Hence we showed that a designer must choose according to the size and the speed requirements of the application.

Chapter 6 showed that scintillator is compatible with the CMOS process and hence can be used in x-ray imaging. The CMOS integrating process of scintillator is described and needs only one additional step. Therefore costs of construction are at a reasonable level. Furthermore we studied various scintillator materials and have drawn conclusions on which are the best scintillator materials that can be used for medical imaging. The choice was CsI:Tl, even though we could have chosen another scintillator, its light yield and wavelength output matches perfectly with the photodiode. The design of the CMOS APS included analysis of the transistor (switches and source follower). These components were constructed to function at minimum length (minimum area). The

design procedure for each transistor (switch and source follower) were explained thoroughly. CDS and the ADC circuit were built in circuits from CADENCE. With these we showed the procedure for building a first implementation of a 2x2 APS pixel system. Even if the results (shown in Table 6.8) are a bit disheartening, we gathered most information on designing a CMOS APS sensor. Fine tuning of devices, precise clock pulsing (sampling) and a further study on the impact of capacitor charges must be made to gain further precision. Furthermore we illustrated the challenges and procedures to verify the design simulation results. In the sixth chapter the analog designer can find a first insight on CMOS sensor and its requirements.

The assertion is made with the hope of extending it in the future, and become a complete guide from schematic to layout design. The future steps include:

1. Correct function of the 2x2 APS pixel on the schematic level.
2. Implement a full CDS and ADC circuit. The challenge here is to build an operational opamp and a comparator which have the following characteristics: phase margins 100°, low slew rate, high open loop gain (feedback configuration) and a minimum size.
3. Layout design
4. Implementation of the 2x2 system at a foundry
5. Testing the system under real conditions

An alternative future implementation could be with another technology, namely with the 50 nm technology. With 50 nm the pixel area devoted for the three transistors shrinks so much that an ADC can be included. Hence the ADC can be integrated at pixel-level, instead at the column level. For that reason we mentioned in chapter 5 the sigma delta ADC. It has been proven that a sigma delta converter can be integrated at pixel-level. A CDS and Nyquist ADC circuit would take too much space. Of course the APS ceases to be APS, since the sigma delta ADC would be directly connected to the photodiode and the read out method of the photodiode's signal would be the current mode. Other problems arise and must be dealt with when dealing with nm technology. For instance gate tunneling occurs and this affects the statistical and matching behaviour of the transistor. The resulting gate current is known to degrade the transistor's function and therefore a general degradation of the sigma delta ADC function is to be expected. Furthermore in [1] it is proven that due to very small threshold voltages in the 45 nm technology, switches stop to function and a new approach to switching must be considered. One possible approach is the use of switching op-amps. Hence with 45 nm technology basic design procedures must be reconsidered, and further testing of the pixel is necessary.

As one can see the possibilities are endless and therefore a study of the CMOS APS system was unavoidable.

References:

[1] Sansen, Willy M C., *Analog design essentials*. s.l. : Springer, 2006. 100-387-25746-2

APPENDIX A

A.1 Light and Crystal Optics

To understand semiconductors and light interaction we review some basic knowledge in Optoelectronics and Semiconductor Physics. These are necessary for latter chapters in which we will make use of the basic properties and equations derived in this chapter. For instance the p-n diode can be used to detect light, and hence it is useful to know what governs the p-n diode and how light interacts with it. We shall begin with the basics of light and then go on to semiconductor physics.

A.2 Electromagnetic wave propagation

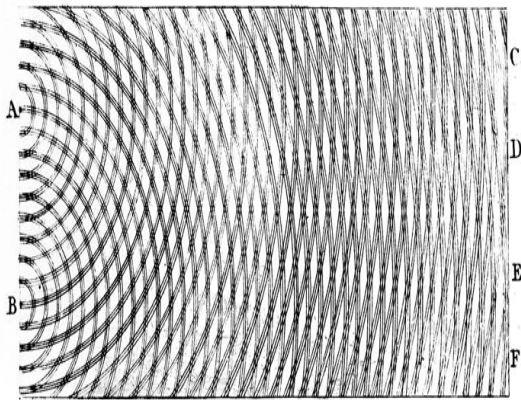


Figure A.1 Spherical wavefronts

Various Light-theories existed already in the ancient world. Indian, Hellenistic and Muslim theories contributed and exploited the idea that light must consist something tiny, like atoms, and that's why they can make their way easily through mediums. Later on, in the middle ages, this given ancient property was extended to that light must have a wave like property. This property helped Young to explain interference and diffraction of light. On the other hand other situations where light and matter interact and exchange energy was still a mystery.

Around 1850, Faraday and later on Maxwell saw that light induces changes to the magnetic and electric field of a medium. Therefore a light must be composed of these fields. Maxwell tried to find out how the fields are related and how they propagate with each other in light. For that purpose he set up all the known equation at that time. Even today they are known as the Maxwell equations:

<i>Integral form</i>	<i>Differential form</i>	<i>Law</i>
<i>Total charge and current</i>		
1. $\oint E dA = \frac{Q_{encl}}{\epsilon_0}$	1. $\nabla E = \frac{\rho_{tot}}{\epsilon_0}$	Gauss Electric
2. $\oint B dA = 0$	2. $\nabla B = 0$	Gauss Magnetic
3. $\oint_{\partial S} E dl = -\frac{d\Phi_{B,S}}{dt}$	3. $\nabla \times E = -\frac{d\Phi_{B,S}}{dt}$	Farraday induction
4. $\oint B dl = \mu_0 \left(I_c + \epsilon_0 \frac{d\Phi_E}{dt} \right)$	4. $\nabla \times B = \mu_0 J_{tot} + \mu_0 \epsilon_0 \frac{\partial E}{\partial t}$	Ampère Circuital
<i>Free charge and current</i>		
5. $\oint_S D dA = Q_{free,S}$	5. $\nabla D = \rho_{free}$	Gauss electric displacement
6. $\oint B dA = 0$	6. $\nabla B = 0$	Gauss Magnetic

7. $\oint_{\partial S} E \, dl = -\frac{d\Phi_{B,S}}{dt}$	7. $\nabla \times E = -\frac{\partial B}{\partial t}$	Faraday induction Ampère Circuital
8. $\oint_{\partial S} H \, dl = I_{free,S} + \frac{d\Phi_{E,S}}{dt}$	8. $\nabla \times H = J_{free,S} + \frac{d\Phi_{E,S}}{dt}$	

Table A.1

To apply the above laws to measure the fields, we must confine the space in which the measures are been taken. Also since light is “invisible”, namely untouchable, we use our imagination to construct areas through which light passes trough, and try to explain with Maxwell equations the nature of light. Take a look at the following:

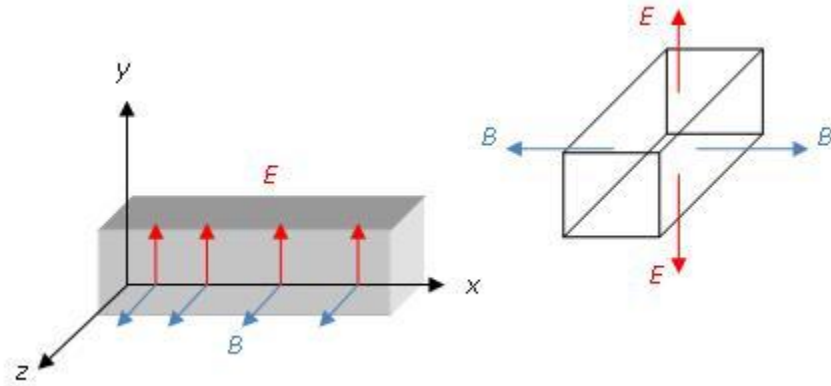


Figure A.2. The electric Field and the magnetic field in an area: volume of rectangular. As one can see the magnetic as well as the electronic field are equal in both directions.

From the above figure we can see that magnetic fields cancel each other out in our defined area (Gaussian surface) and thus the magnetic Law of Gauss holds. The second Law of the Maxwell equations holds also, meaning that $E=B=0$. This result also implies that the first two equations can be valid only if the field E is perpendicular to B . The wave in that case is called a traverse wave. We know that light travels and that it has a wave property. Once again we construct an imaginary area, called plane, and set it in a space. The Law of Faraday requires an amount time, say a small amount dt , so we can measure the increase in flux (Φ_e) of the magnetic field. Setting up multiple planes at the direction of propagation x we simulate how the light travels. These fronts are also called wave fronts. We assume that each front appears at a distance $c*dt$.

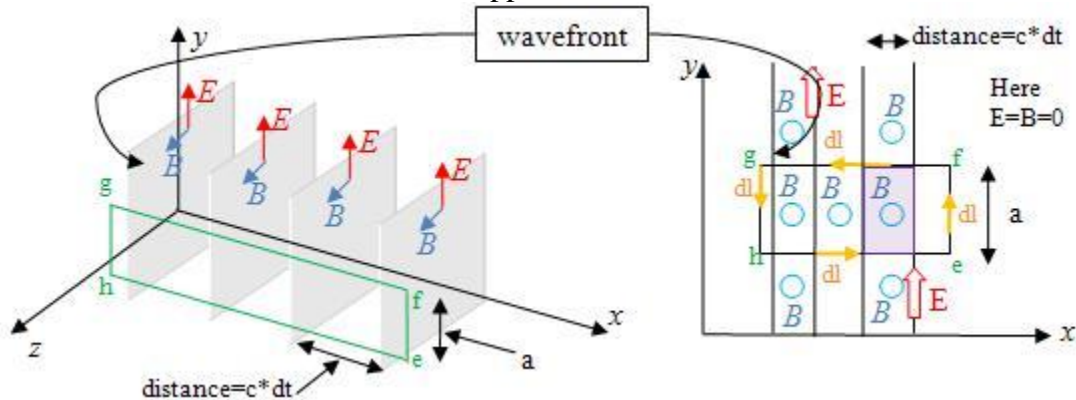


Figure A.3. Depicting how the wave is analyzed. Setting areas in time equal to $c*dt$ to calculate Maxwell's first two equations. *Assumption is made that before the wave front $E=B=0$ while behind B, E is uniform.*

The magnetic field changes in the z direction, so integrating the rectangle, counterclockwise (efgh), we find that only gh (a) produces a non-zero integral. Hence the overall field is $\oint E dl = -Ea$ (dl is opposite to E). The induced by the electric field, magnetic flux, is therefore increased by $\frac{d\Phi_B}{dt} = Bacdt$. Thus the forth equation of Maxwell yields:

$$\begin{aligned} -Ea &= Bacdt \\ \Rightarrow E &= cB \end{aligned}$$

The last equation of Maxwell is easily described. One must simply see that there is no conduction current, hence: $\stackrel{ic=0}{\Rightarrow} \oint B dl = \mu_0 \epsilon_0 \frac{d\Phi_E}{dt}$.

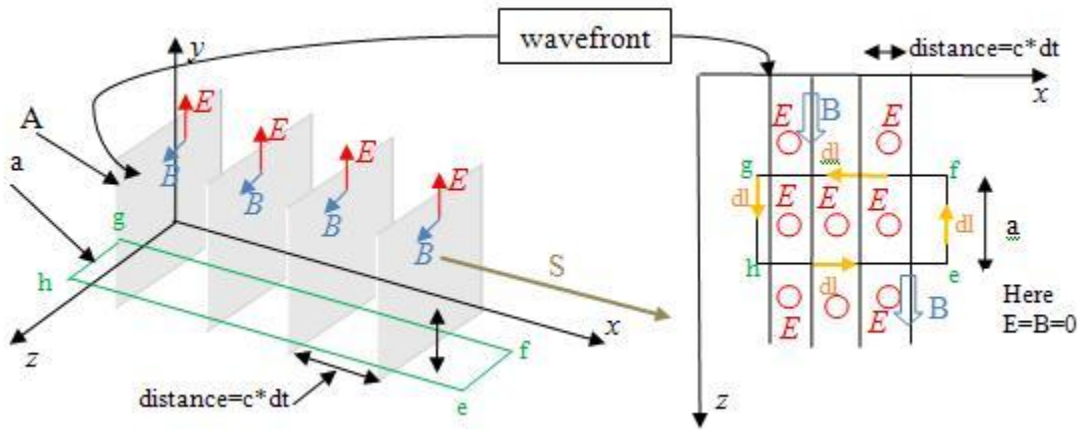


Figure A.4. Continued wave analysis.

Following the same procedure like for the electric field, we see that only a contributes to the magnetic field yielding (dl parallel with B). Hence

$$\oint B dl = Ba$$

The last equation by noticing that $\frac{d\Phi_E}{dt} = Eac$ is then

$$\oint B dl = Ba = \mu_0 \epsilon_0 \frac{d\Phi_E}{dt} = \mu_0 \epsilon_0 Eac \Rightarrow \mu_0 \epsilon_0 Ec$$

Since we found that $E = cB$ we have a relation that gives **the speed of light c** :

$$c = \sqrt{\frac{1}{\mu_0 \epsilon_0}} \quad (\text{A.1})$$

where ϵ_0 and μ_0 are the permeability and permittivity of free space, respectively. Substituting the values yield $c \sim 3 * 10^8 \frac{m}{s}$.

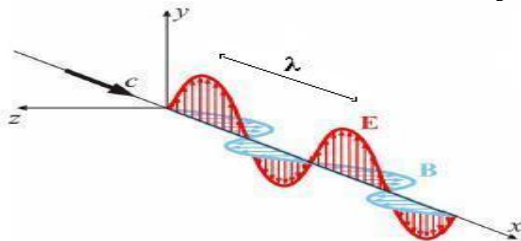


Figure A.5

Light as electromagnetic sinusoidal wave. Wave is travelling through air with the speed of light c Red color is the electric field and blue color is the magnetic field. Figure from [1]

A.2.1.1 Refraction index

Another interesting fact arises from the two Maxwell equations: If we replace the speed of light with the speed of light u , which an electromagnetic wave may have in a medium, we can find the permeability and permittivity of that medium. Defining $\mu = k_m \mu_0$ (relative permeability) with k_m the relative permeability and $B = \mu_0 \epsilon_0 E c \rightarrow B = \epsilon \mu E u$ we have:

$$u = \sqrt{\frac{1}{\mu \epsilon}} = \sqrt{\frac{1}{k k_m}} \sqrt{\frac{1}{\mu_0 \epsilon_0}} = \sqrt{\frac{c}{k k_m}} \quad (\text{A.2a})$$

u 's decency is clearly on the relative permeability k_m . For dielectric material we have $k_m=1$. u then becomes

$$u = \sqrt{\frac{1}{k}} = \sqrt{\frac{1}{\mu_0 \epsilon_0}} = \frac{c}{\sqrt{k}} \text{ valid for } k > 1, u < c \quad (\text{A.2b})$$

Rearranging the terms yields the refractive index n of the material

$$\frac{c}{u} = \sqrt{k} = n \quad (\text{A.3})$$

A.2.1.2 Wave function

Maxwell proved that the fields displacements can be represented by the known displacement equation of a mechanical wave :

$$\nabla^2 y(x, t) = \frac{\partial^2}{u^2 \partial t^2} y(x, t) \rightarrow \nabla^2 (\mathcal{E}, B) = \frac{\partial^2}{c^2 \partial t^2} (\mathcal{E}, B) \quad (\text{A.2.4})$$

We will study the ideal case where there is only one wavelength. In reality light is a mixture of different wavelengths (waves) and only if it passes through medium(s) like for instance our eye some wavelengths are filtered. We refer to such medium as Polaroid's and the event as polarization of light, which we will discuss later. Equation A.4 shows that fields oscillate with frequency f_v and hence values of the field(s) are periodic. The distance between two identical values is called a wavelength λ . By convention λ is always referred to the λ of the electric field. The wavelength for vacuum is denoted as λ_o , and is related to the speed of the fields in the direction x by

$$c = \lambda_o f_v \quad (\text{A.2.5})$$

For other medium to find the speed of propagation we rearrange equation (A.3) and with use of A.5 we get:

$$u = \frac{c}{n} = f_v \frac{\lambda_o}{n} \quad (\text{A.2.6})$$

where n is the refraction index of the medium and λ is the wavelength in the medium.

As we already proved the fields E and B are oscillating perpendicularly, this simplifies finding the mathematical expression of propagation light. Usually, for simplicity, we refer to the electric field propagation since the magnetic component behaves similarly. To simplify even further we take the simplest expression that satisfy all of the Maxwell conditions. Therefore waves are represented by sinusoidal waves:

$$E(x, t) = E_0 \cos(\omega t - kx + \varphi)$$

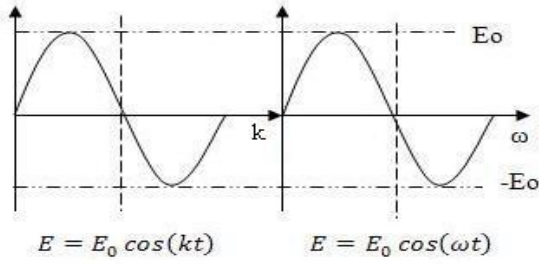


Figure A.6. Wave travelling through space. It can be expressed with the wavenumber k or with the angular frequency ω .

This equation describes how E varies in time (t) as it propagates in a direction (x). E_0 is the amplitude of the wave, ω ($2\pi f_v$) the angular frequency, k is the wavenumber ($k = \frac{2\pi}{\lambda}$) and φ the phase constant. The cosine term is called the phase of the wave. For demonstration purposes if we take $x=0$ or $t=0$ assuming $\varphi=0$ we get the figures left.

If E_0 for x and t then $\varphi \neq 0$. The phase velocity of the wave is given $u = \frac{\omega}{k} = f_v \lambda$. Now in reality waves are, like mentioned before, a combination of different wavelengths. The ideal case here is for a single wavelength or also called monochromatic wave. The packet of waves is propagating; hence this packet also has a velocity: the group velocity. By inspection of two waves it was been found that the group velocity is

$$u_g = \frac{\partial \omega}{\partial k} \quad (\text{A.7})$$

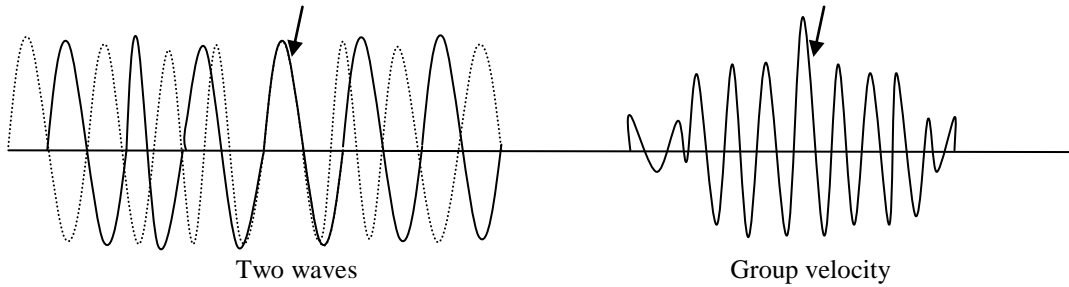


Figure A.7. The group velocity concept.

This is something to be remembered because we get back to it in the semiconductor. To complete and to generalize wave function (A.4) we include plane waves moving in arbitrary directions. For this we use the wave vector k : k 's magnitude ($|k| = \frac{2\pi}{\lambda}$) is the reciprocal of wavelength and the vector indicates direction of propagating. To clear things out, let's define r as a vector starting from the origin to the point (x, y, z) in space. For a two-dimensional case for instance, where wave fronts are traveling in a direction with angle θ to the x axis, k is defined as

$$k = ik_x + jk_y$$

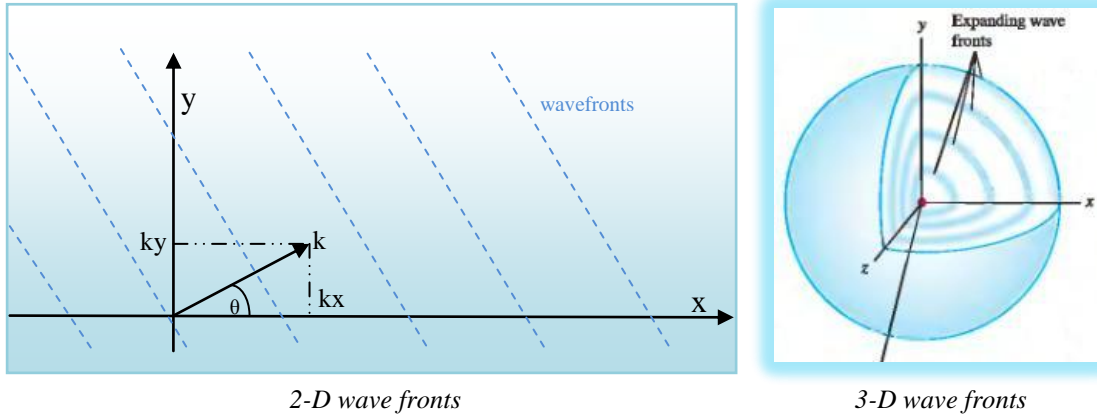


Figure A.8. Expanding wavefronts in 2D and 3D.

multiplying \mathbf{r} we have

$$\mathbf{k} * \mathbf{r} = (ik_x + jk_y)(ix + jy) = xk_x + yk_y = xk \cos \theta + yk \sin \theta \quad (\text{A.8})$$

Substituting \mathbf{r} for the direction \mathbf{x} in (A.4) and $\mathbf{k} * \mathbf{r}$ with A.8 we get

$$\begin{aligned} E(x, y, z, t) &= E_0 \cos(\omega t - \mathbf{k} * \mathbf{r} + \varphi) \Rightarrow \\ E(x, y, t) &= E_0 \cos(\omega t - k \cos \theta - yk \sin \theta + \varphi) \quad (\text{A.9}) \end{aligned}$$

Spherical waves, like that in figure A.8, are constructed by a point source of light and radiate in isotropic mediums uniformly and in anisotropic mediums they may experience change in their direction of propagating or change the irradiance. In both cases we start to analyze this phenomena's with the equivalent equation of A.9

$$E = \frac{A}{r} \cos(\omega t - \mathbf{k} * \mathbf{r}) \quad \text{A.10}$$

Where A is the source strength and $1/r$ "tracks" the decrease in amplitude from the source. A more general expression can be derived by setting $\varphi=0$ and observing that in (A.9) the angle $\mathbf{k} * \mathbf{r}$ is always negative and the sign depend mainly on \mathbf{r} . We define $\mathbf{k} * \mathbf{r}$ as function φ and then (A.10) can be written in the time domain as

$$E(\mathbf{r}, t) = \check{x}E_x \cos(\omega t - \varphi_x) + \check{y}E_y \cos(\omega t - \varphi_y) + \check{z}E_z \cos(\omega t - \varphi_z) \quad (\text{A.11})$$

where \check{x}, \check{y} and \check{z} are unity vectors pointing in the directions x, y and z respectively. (A.11) is a convenient way of expressing the electric field since it results from the phasor expression $E(\mathbf{r}, \omega)$ ($E(\mathbf{r}, t) = \text{Re}[E(\mathbf{r}, \omega)] e^{-i\omega t}$). Hence the phasor expression $E(\mathbf{r})$ is

$$E(\mathbf{r}, \omega) = \check{x}E_x e^{i\varphi_x} + \check{y}E_y e^{i\varphi_y} + \check{z}E_z e^{i\varphi_z} \quad (\text{A.12})$$

which is used in many situations.

We mention isotropic and anisotropic materials. To explain the terms we look once again to Maxwell equations. We described only four out of the eight equations. These four equations describe light interactions with materials where constants like the ϵ permittivity and μ permeability are truly constants. These materials are known as isotropic, whereas materials in which the permittivity changes are called anisotropic. The

reason why we mention this is because semiconductor is an isotropic medium, as one already may know.

A.2.2 Energy of a wave

Another key aspect is the energy of an electromagnetic wave. From figure A.4 we see that the plane front has an area A. The density of energy on one front is

$$u = \frac{1}{2} \epsilon_0 E^2 + \frac{1}{2\mu_0} B^2 \quad (\text{A.13})$$

Because $B = \frac{E}{c} = \sqrt{\mu_0 \epsilon_0} * E$ we have

$$u = \epsilon_0 E^2 \quad (\text{A.14})$$

Density is defined for a volume and as we know, two sequential plane fronts are $c*dt$ away from each other. We define the volume between them as $dU = A * c * dt$, or simply the energy that flows through an area A in time dt. Now power density (S) is $\frac{\text{energy}}{m^2 * sec}$ so

$$S = \frac{1}{A} \frac{dU}{dt} = \epsilon_0 c E^2 = \frac{\epsilon_0}{\sqrt{\epsilon_0 \mu_0}} E^2 = \sqrt{\frac{\epsilon_0}{\mu_0}} E^2 = \frac{EB}{\mu_0} \quad (\text{A.15})$$

Hence
 $S = \frac{EB}{\mu_0} = \frac{E \mu_0 H}{\mu_0} = EH$ for free space and the direction of the energy (Pointing vector) is simply

$$S = E \times H \quad (\text{A.16})$$

A last property added to light was that given by Alfred Einstein, namely that of the wave-particle duality. The duality concept opened a new research area the quantum physics. The word quantum has its origin from Max Planck research on black body radiation. As he has shown the body emits, even in absence of light, discrete bundles or packets of energy. These packets were called quanta. The problem was that Planck didn't connect it with the electromagnetic wave theory. Einstein showed that the energy must travel like light as particles. This particle of light was given the name photon. The energy of quanta is defined as

$$E = hf_v \quad (\text{A.17})$$

Where h is Planck's constant, and f_v the frequency of the wave. With the last equation we can set up a spectrum off all electromagnetic waves in vacuum (Figure A.9).

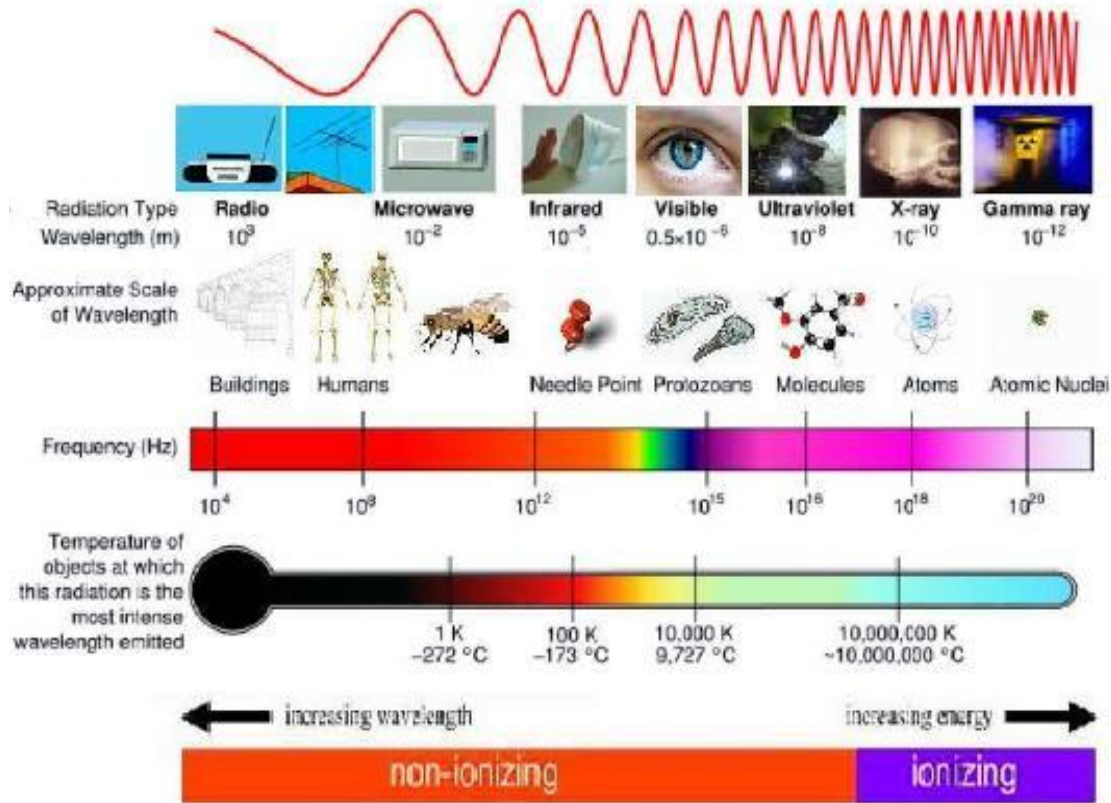


Figure A.9 Electromagnetic spectrum with its associated frequencies, temperatures and wavelengths.

Next we will show that analysis of isotropic medium reveal some properties of light interaction with semiconductors.

A.2.3 Propagation in isotropic media

With the phasor (A.12) of the electric field we redefine Maxwell equations, assuming current density J and charge density ρ to be zero then the first four equations:

$$\nabla \times E(r) = i\omega B \quad (\text{A.18})$$

$$\nabla \times H(r) = -i\omega D \quad (\text{A.19})$$

$$\nabla B(r) = 0 \quad (\text{A.20})$$

$$\nabla D(r) = \rho(r) = 0 \quad (\text{A.21})$$

where D is known as the displacement factor and H is known as magnetic field intensity or either as magnetic field strength or either as auxiliary magnetic field or as magnetizing field. In simple words H represents the vorticity (spin). These complex fields vectors have the spatial dependency $\exp(ik \cdot r)$ in common. Hence

$$k \times E = \omega B \quad (\text{A.22})$$

$$k \times H = -\omega D \quad (\text{A.23})$$

$$k \cdot H = 0 \quad (\text{A.24})$$

$$k \cdot E = 0 \quad (\text{A.25})$$

From the last two equations we see that k is perpendicular to B and D but this doesn't hold for anisotropic media. We will confine our analysis here to isotropic media only where

$$D = \epsilon E \text{ and } B = \mu H \quad (\text{A.26})$$

and revisiting equations (A.22) to (A.25) we see that D and B are parallel to E and H respectively. The plane wave is the direction of the wave vector k . Also in isotropic media the polarization does not change as it propagates through the homogeneous medium, so equations (A.22) to (A.25) with (A.26) become

$$k \times E = \omega \mu H \quad (\text{A.27})$$

$$k \times H = -\omega \epsilon E \quad (\text{A.28})$$

$$k \cdot H = 0 \quad (\text{A.29})$$

$$k \cdot E = 0 \quad (\text{A.30})$$

From these equations the following conclusion can be made:

1. k is perpendicular to both E and H
2. $k \times E$ is in the direction of H and $k \times H$ in direction of $-E$
3. E, H and k are perpendicular to each other

Taking the cross product of (A.27) by k and using (A.30) we have

$$k \times (k \times E) = (k \cdot E)k - k^2 E = -k^2 E = \omega \mu k \times H = \omega \mu (-\omega \epsilon E) \quad (\text{A.31})$$

Using (A.28) in (A.31)

$$(k^2 - \omega^2 \mu \epsilon) E = 0 \quad (\text{A.32})$$

From plain mathematics we exclude the solution $E=0$ hence

$$k^2 = \omega^2 \mu \epsilon \quad (\text{A.33})$$

Lastly substituting (A.2a) ($\frac{1}{u} = \sqrt{\mu \epsilon}$) and (A.3) in (A.33)

$$k = \frac{\omega}{c} n \quad (\text{A.34})$$

In most semiconductors $\mu = \mu_0$ and the permittivity is complex satisfying the Kramers-Kronig relations. An analytical explanation is found in [2]. It's sufficient to say that in real materials, the polarization does not respond instantaneously to an applied field. This causes dielectric loss, which can be expressed by a permittivity that is both complex and frequency dependent:

$$\epsilon(\omega) = \epsilon'(\omega) + i\epsilon''(\omega) \quad (\text{A.35})$$

Since permittivity is complex then also refraction index n is complex

$$\bar{n}(\omega) = n(\omega) + i\kappa(\omega) \quad (\text{A.36})$$

Then (A.34) becomes

$$k = \frac{\omega}{c} \bar{n}(\omega) = \frac{\omega}{c} [n(\omega) + i\kappa(\omega)] \quad (\text{A.37})$$

It has been found that the absorption coefficient a is half of the imaginary part of k (A.37). Using (A.5):

$$a(\omega) = 2 * \text{Im} \kappa = 2 \frac{\omega}{c} \kappa(\omega) = \frac{4\pi}{\lambda} \kappa(\omega) \quad (\text{A.38})$$

where we used (A.5) and $k(\omega)$ is also known as k the extension coefficient.

Assuming a plane wave travels in the z direction and using the phasor expressions for E and H we have

$$E = \check{x} E_0 e^{ikz} = \check{x} E_0 \exp\left(i \frac{2\pi}{\lambda} nz - \frac{a}{2} z\right) \quad (\text{A.39})$$

$$H = \check{y} \frac{E_0}{\eta} e^{ikz} = \check{y} \frac{\bar{n} E_0}{\eta_0} \exp\left(i \frac{2\pi}{\lambda} nz - \frac{a}{2} z\right) \quad (\text{A.40})$$

where $\eta_0 = \sqrt{\frac{\mu_0}{\epsilon_0}} = 120\pi$ the characteristic impedance of the free space.

Since H is complex due to \tilde{n} , S also will be complex. Taking the real part of E and H we obtain the optical power density

$$S = \frac{1}{2} \text{Re}[E \times H] = \tilde{z} \frac{n}{2\eta_0} |E|^2 e^{-az} \quad (\text{A.41})$$

From (A.41) we see that the wave propagates and decays exponentially as it moves away from the origin of axes, here in the z direction. A more important result is the rate of decay, which is determined by the absorption coefficient a . The absorption coefficient a will be very useful in analyzing photoconductive materials.

A.2.4 Reflection, Refraction and Transmissivity

The concept of reflection and refraction is here summarized in the Figure below, for more insight on this matter refer to [3].

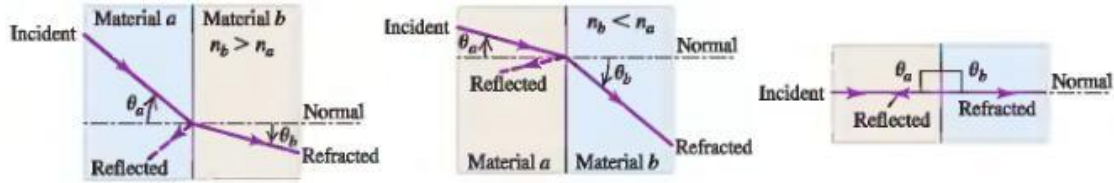


Figure A.10. Reflection and refraction on different materials

Snell's Law: $\frac{\sin \theta_a}{\sin \theta_b} = \frac{n_b}{n_a}$	Phenomena's when refraction index $n_b < n_a$	Phenomena's when angle of incident ray is zero (normal axis)
when refraction index $n_b > n_a$		

The absorption coefficient a , plays also a role in the reflectivity of a semiconductor, writing $\kappa = \frac{a\lambda}{4\pi}$ and defining the reflectance R for zero incident angle to be

$$R = \left| \frac{n_0 - \tilde{n}}{n_0 + \tilde{n}} \right|^2 = \frac{(n-1)^2 + \kappa^2}{(n+1)^2 + \kappa^2} \quad (\text{A.42})$$

where n_0 is the refraction index of the air.

The absorbance A for a slap with thickness d is described as a function of transmissivity T and reflectivity (reflectance) R . $A = A \cdot T - R$, where

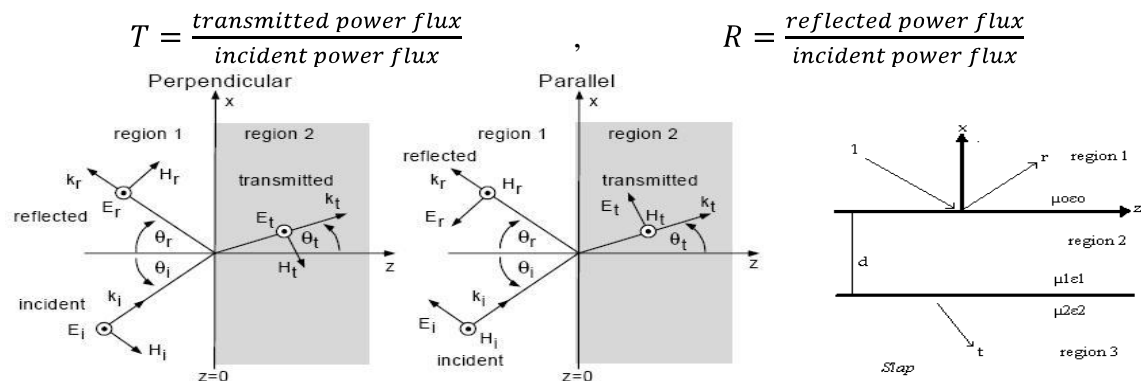


Figure A.11. Showing all kinds of phenomenas possible by incident light on a material.

In [2] it was proven that the following expression holds

$$A = 1 - R - T = \frac{(1-R_{01})(1-e^{-ad})}{(1-R_{01}e^{-ad})^2} (1 + R_{01}e^{-ad}) \quad (\text{A.43})$$

where R_{01} is the reflectivity between region0 and region1. Also

$$1 = R_{01} + T_{01} \quad (\text{A.44})$$

If $ad \gg 1$ and region2 is the same as region0 then (A.43) can be simplified into

$$A \cong (1 - R_{01})(1 - e^{-ad}) \quad (\text{A.45})$$

A.3 Crystal construction and semiconductors

When atoms come together to form a solid material they can be placed in repeating order such they can form a lattice. Connection between atoms can exists and hence they are been repeated throughout the material. When atoms come together in such manner its said that the atoms have a periodicity. Materials in which atoms are placed in a high ordered repeated structure are called crystalline. If the atoms are spread randomly in the material with little periodicity or none at all is called amorphous crystalline. If we have various lattices which are repeated with many periodicities in the material, then we speak of a polycrystalline.

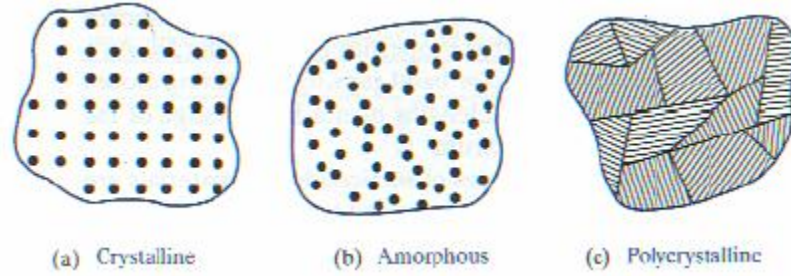


Figure A.12. Crystal structures in semiconductors. Figure from [1].

From what is presented until now we can conclude that crystals are categorized by their crystal structure and the underlying lattice. While some crystals have a single atom placed at each lattice point, most crystals have a combination of atoms associated with each lattice point. The combination of atoms is also called the basis.

A.3.1 Bravais lattices

To be capable to fill the whole space of the material we need to define special repeated lattices called Bravais lattices. To construct the lattice we define three unit vectors \vec{a}_1, \vec{a}_2 and \vec{a}_3 , and a set of integers c_1, c_2 and c_3 so that each lattice point, is identified by a vector, also called the translation vector:

$$\vec{r} = c_1\vec{a}_1 + c_2\vec{a}_2 + c_3\vec{a}_3 \quad (\text{A.46})$$

From the above it is clear that not only one bravais lattice can be defined, but accordingly to the dimensions of space, many ways of describing each point in the lattice can exist, for example in two dimensions, there are five distinct Bravais lattices, while in three

dimensions there are fourteen. The lattices in two dimensions are the square lattice, the rectangular lattice, the centered rectangular lattice, the hexagonal lattice and the oblique lattice as shown in Figure A.13

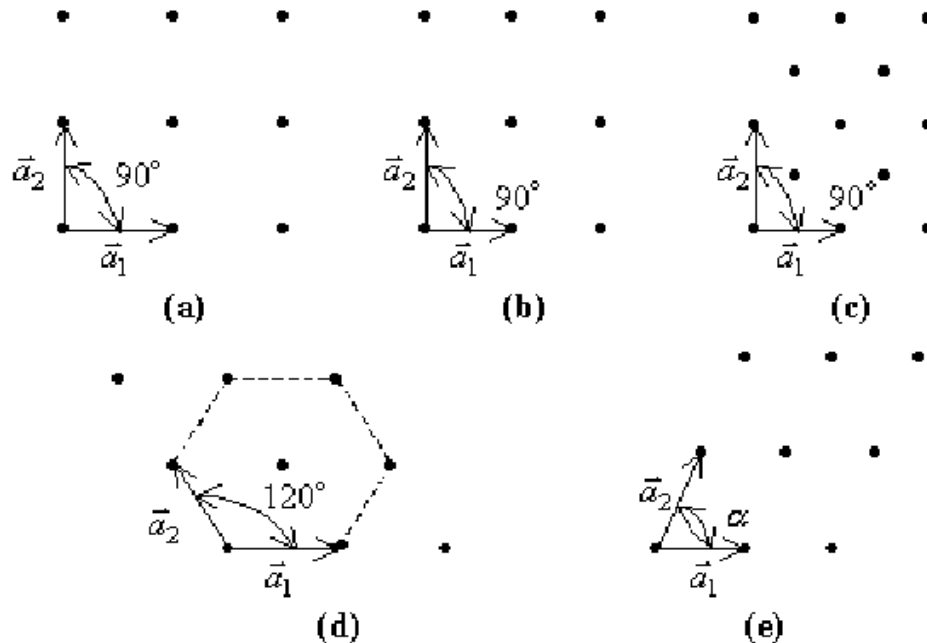


Figure A.13. The five possible bravais lattices in a 2-D crystal (a)cubic (b)rectangular (c) centered rectangular (d)hexagonal and (e)oblique. Figure from [4].

The lattices are grouped together according to their similarities in symmetry. For example the rectangular and the centered rectangular group because all the lattice points, except the atom in the center centered rectangular lattice, can be constructed in the same way. Therefore the centered rectangular lattice is been constructed with the same lattice vectors as the rectangular with exception of the addition of the centered atom. Some minimal definition like primitive cell and unit cell are necessary to organize and design crystal structures. The primitive cell is the smallest part of the lattice that which repeated would reconstruct the entire crystal structure. The unit cell is needed to produce a primitive cell. It has the property of being the smallest volume possible defined by unit vectors. Although they may look identical they don't have to be the same.. In the following example two primitive cells can be defined (defined by the unit vectors \vec{a} and \vec{b}). One can see that both have minimum volume, so both cells are valid.

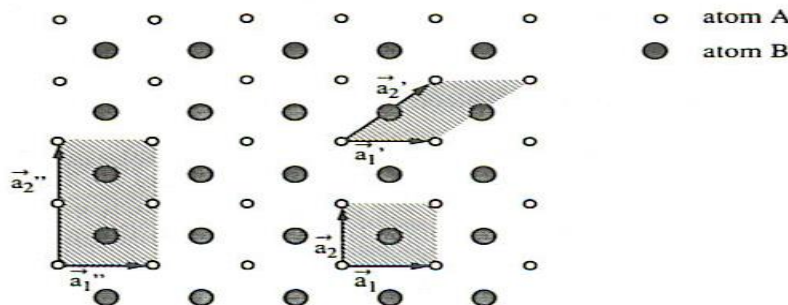


Figure A.14. Primitive cells definition. Figure taken from [5].

The fourteen bravais lattices (for the three dimensional case) are the important ones and are been displayed for each system

System	Number of Lattices	Lattice Symbol	Restriction on crystal cell angle
Cubic	3	P or sc, I or bcc, F or fcc	$a=b=c$ $\alpha = \beta = \gamma = 90^\circ$
Tetragonal	2	P, I	$a=b \neq c$ $\alpha = \beta = \gamma = 90^\circ$
Orthorhombic	4	P, C, I, F	$a \neq b \neq c$ $\alpha = \beta = \gamma = 90^\circ$
Monoclinic	2	F, C	$a \neq b \neq c$ $\alpha = \beta = 90^\circ \neq \gamma$
Triclinic	1	P	$a \neq b \neq c$ $\alpha \neq \beta \neq \gamma$
Trigonal	1	R	$a=b=c$ $\alpha = \beta = \gamma < 120^\circ, \neq 90^\circ$
Hexagonal	1	P	$a=b \neq c$ $\alpha = \beta = 90^\circ$ $\gamma = 120^\circ$

P - Primitive: simple unit cell

F - Face-centred: additional point in the centre of each face

I - Body-centred: additional point in the centre of the cell

C - Centred: additional point in the centre of each end

R - Rhombohedral: Hexagonal class only

The cubic lattices are an important subset of these fourteen Bravais lattices since a large number of semiconductors are cubic. The three cubic Bravais lattices are the simple cubic lattice, the body-centered cubic lattice and the face-centered cubic lattice. Since all unit vectors identifying the traditional unit cell have the same size, the crystal structure is completely defined by a single number. This number is the lattice constant, a . An example of different systems of bravais lattices is shown in Figure A.15

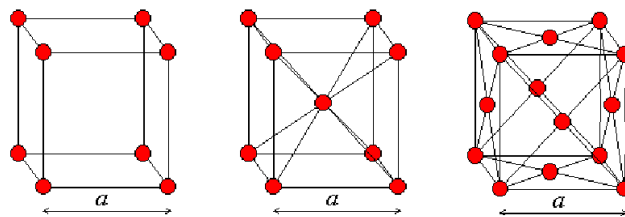


Figure A.15. (a)normal cubic (b)body-centered cubic(bcc)(c)face-centered cubic(fcc)lattice. Figure from [4].

The bonding in semiconductors is covalently and always forms a tetrahedral structure. As can be seen in the Figure A.1.3 only the four outer electrons in silicon or germanium contribute to the bond, and these interact with one another to form an elongated electron cloud associated with each electron. The clouds are as far as possible away from the nucleus and maintain between them an angle of approximately $109^\circ 28'$.

Therefore when this arrangement is repeated we always have a tetrahedral shape, or better known as the diamond structure. When different atoms are used, then they bond in a similar way as in the diamond structure, but despite the electron contributions one atom is acting as the central point where the other atoms bond. This is called zincblende structure. An example is shown in figure A.16.

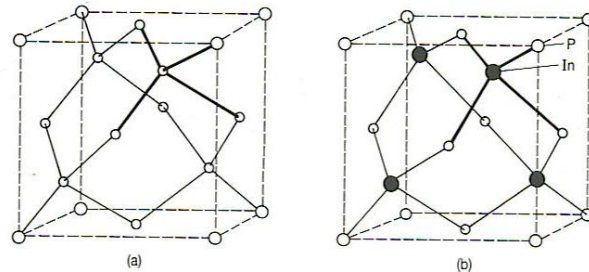


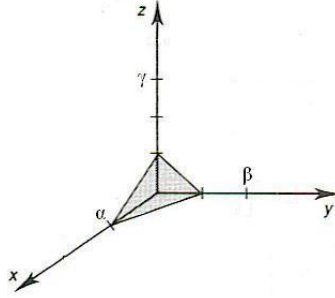
Figure A.16. Zincblende structures. Figure from [5].

The reason of analyzing the structures of crystals becomes clearer in the next paragraphs. For now we can calculate, based on what we described so far, the package density of a crystal. It's an important issue when choosing a semiconductor for an application.

Structure	Radius	Atoms in a unit cell	Packaging density $\left(\frac{\text{Volume of atoms}}{\text{Volume of unit cell}} \right)$
Simple Cubic	$\frac{a}{2}$	1	$\frac{\pi}{6} = 52\%$
Body Centered Cubic	$\frac{\sqrt{3}a}{4}$	2	$\frac{\sqrt{3}\pi}{8} = 68\%$
Face Centered Cubic	$\frac{\sqrt{2}a}{4}$	4	$\frac{\sqrt{2}\pi}{6} = 74\%$
Diamond	$\frac{\sqrt{3}a}{8}$	8	$\frac{\sqrt{3}\pi}{16} = 34\%$

A.3.2 Miller indices

From figure A.15 (b) we can see that spacing along the face of the cube is different that across the body of the cube. Therefore we can state that the choice of a particular plane can be different from another since the one includes a slightly different atomic layout then the other plane. Hence the importance of the electrical properties lies within the crystal orientation. To identify the orientation of a crystal we need a method which is independent of position in space and the orientation can be applied anywhere in the crystal structure. The suitable method is the Miller indices. The best way to explain the Miller indices is with example. Shown in the below is a crystal plane scaled with the lattice constant α , β and γ .



The crystal intercepts the x, y and z axes at α , $\beta/2$ and $\gamma/3$. The steps involved to calculate the Miller indices are:

- Interpret the intercepts as a fraction of the unit distance α , β and γ , here α , $\beta/2$ and $\gamma/3$.
- Remove the unit symbols
- Take reciprocals, here 1,2,3, such that it can be reduced any further i.e. they cant be divided by a common divider. If we had 2,4,6 \rightarrow 1,2,3.
- Remove the commas and enclose the result with parentheses 1,2,3 \rightarrow (123)

(123) is the Miller indices of the example.

A.3.3 The Reciprocal Lattice

Since the real space is described by the wave vectors k or quasi momenta $\hbar k$ we want to translate the real space vectors so that we do not have to work with the whole k -space, but only with the unit cell. The space in the unit cell can be defined by new lattices vectors b_i like:

$$w_1 = \frac{2\pi}{V_{uc}} \vec{a}_2 \times \vec{a}_3 \quad (\text{A.47})$$

$$\text{where } V_{uc} = \vec{a}_1 \cdot \vec{a}_2 \times \vec{a}_3 \quad (\text{A.48})$$

the volume of the unit cell, then the reciprocal lattice is described by the general translation vector

$$G = c_1' w_1 + c_2' w_2 + c_3' w_3 \quad (\text{A.49})$$

In real space a periodic function which is smooth can be described by $f(R+r) = f(R)$ with r defined by (A.1.1). If we write $f(R)$ in the sense of a Fourier series which sums all vectors of the reciprocal lattice, we get:

$$f(R) = \sum_G f_G e^{iGR} \quad (\text{A.50})$$

with

$$f_G = V_{uc}^{-1} \int_{uc} f(\mathbf{r}) e^{-i\mathbf{G}\cdot\mathbf{r}} d\mathbf{r} \quad (\text{A.51})$$

From observation if we substitute \mathbf{r} with $\mathbf{r} + \mathbf{R}$ then in (A.51) $e^{-i\mathbf{G}\cdot(\mathbf{r}+\mathbf{R})} = e^{-i\mathbf{G}\cdot\mathbf{r}} e^{-i\mathbf{G}\cdot\mathbf{R}}$. Therefore only certain wave vectors \mathbf{k} which satisfy $e^{-i\mathbf{G}\cdot\mathbf{R}} = 1$ always fulfill the scalar product $\mathbf{R}\cdot\mathbf{G} = 2\pi m$ with $m = 0, \pm 1, \pm 2, \dots$.

From the above we can express every \mathbf{k} vector in the unit cell, described by the vectors \mathbf{w}_i , and just shift the result for the whole crystal according to Noether's theorem [4] for infinitesimal translations.

$$\mathbf{k} \rightleftharpoons \mathbf{k} + \mathbf{G} \quad (\text{A.52})$$

The theorem also states that the energy is conserved with the translation. Hence any translation from real space with momentum $\hbar\mathbf{k}$ to reciprocal space is conserved. Consequently any \mathbf{k} -vector outside the unit cell can be shifted inside the unit cell and vice versa, thus only the unit cell vectors are sufficient to describe the crystal.

A.3.4 Wigner-Seitz Cell

Although the above unit cell is not used, a more practical method to work in reciprocal space is been used. Introduction to the most used cell, the Brillouin zone, is the construction of the Wigner-Seitz cell, a primitive cell which displays the full symmetry of the lattice. The Figure below shows the construction of a Wigner-Seitz cell. In reciprocal space, the Wigner-Seitz cell is also a Brillouin zone and we shall use it to construct Brillouin zones later.

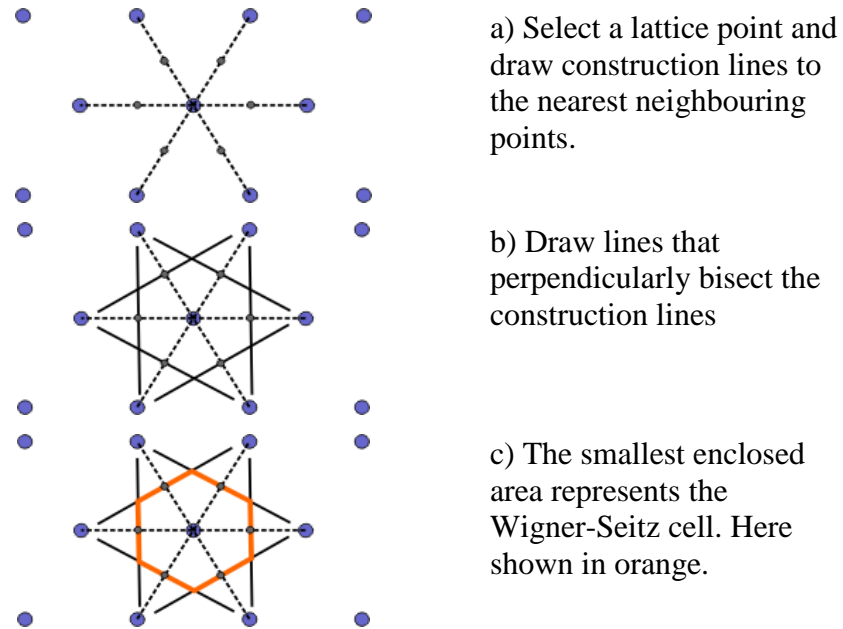


Figure A.17 Construction of the Wigner-Seitz cell <http://britneyspears.ac/physics/crystals/wcrystals.htm>

A.3.5 Brillouin Zones

In the reciprocal lattice the Brillouin zone is defined as the volume within the Wigner-Seitz cell. At the boundaries of the Brillouin zone, the Bragg diffraction condition in the reciprocal lattice must be satisfied.

$$k' = k + G \quad (\text{A.53})$$

Where k' is the wavevector of the diffracted wave and k is the incident wavevector and G is a reciprocal lattice vector. Squaring (A.8) gives:

$$k'^2 = k^2 + 2kG + G^2 \quad (\text{A.54})$$

and assuming the wave is elastically scattered, then $k'^2 = k^2$. Equation (A.54) becomes $2k \times G = -G^2$. Since G is a lattice vector then $-G$ is rewritten as

$$2kG = G^2 \quad (\text{A.55})$$

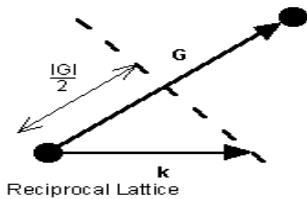
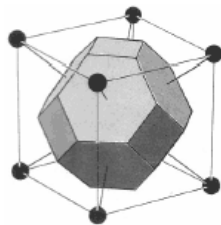


Figure A.18 The geometric interpretation of the Bragg diffraction condition that gives rise to Brillouin zone boundaries. Figure <http://britneyspears.ac/physics/crystals/wcristals.htm>

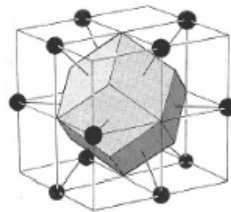
The geometric interpretation of equation (A.55) is that the diffraction condition is satisfied if k lies in the plane that perpendicularly bisects the lattice vector G . The construction method for the Brillouin zones (BZ), is the same as for the construction for the Wigner-Seitz cell (WS) mentioned above but using the reciprocal lattice space. Further Brillouin zones can be constructed by taking the next nearest set of lattice points from the starting point and repeating the process. Since the lattice and reciprocal lattice are related, the WS cell defined in real space and the WS in k -space are also related. In particular, the WS defined in the bcc real space lattice gives a fcc BZ in reciprocal lattice and vice versa.

Lattice Real Space



bcc WS cell

Lattice k-space



fcc BZ

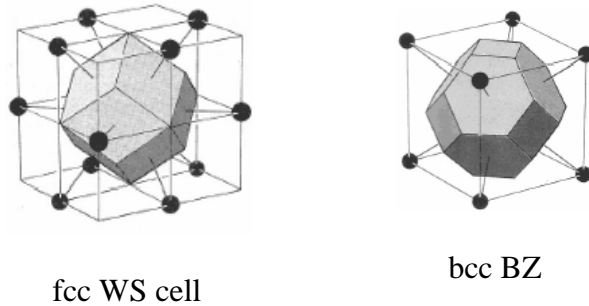


Figure A.19 The transformation of the WS cell of bcc lattice in real space transforms to a Brillouin zone in a fcc lattice in reciprocal space while the WS cell of a fcc lattice transform to a Brillouin zone of a bcc lattice in reciprocal space. Figure <http://britneyspears.ac/physics/crystals/wcrystals.htm>

Points of symmetry on the Brillouin zone are given particular importance especially when determining the bandstructure of the material.

Electrons in the semiconductor are perturbed by the potential of the crystal. The bandstructure of the semiconductor are the allowed energies that the electrons can have. These bands of energy vary with k-space (reciprocal lattice space). Therefore, points of high-symmetry on the Brillouin zone have specific importance. Perhaps the most important, at least for optoelectronic devices, is at $\mathbf{k} = \mathbf{0}$ which is known as the gamma point, the symbol for this point is Γ .

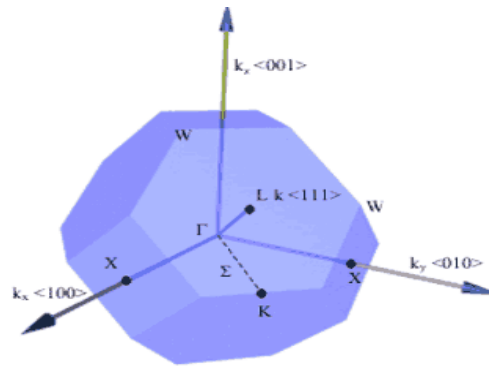


Figure A.20 Some important symmetry points on the Brillouin zone of a fcc crystal (real space) and directions of planes. Figure <http://britneyspears.ac/physics/crystals/wcrystals.htm>

Symmetry Point	\mathbf{k}
Γ	$k_x = 0, k_y = 0, k_z = 0$
X	$\{k_i = \pm 2\pi/a, k_j = 0, k_k = 0\}$
L	$\{k_x = \pm\pi/a, k_y = \pm\pi/a, k_z = \pm\pi/a\}$

K-vectors of the important symmetry points for the fcc crystal structure. The indices for X points are a cyclic permutation of axes. E.g. If $i=x$ then $j=y, k=z$. If $i=y$ then $j=z$ and $k=x$, etc.

A.3.6 Particle theories

We know the concept of wave-particle duality exist, and that the momentum of a particle is

$$p = \frac{E}{c} = \frac{h\nu}{c} = \frac{h}{\lambda}$$

To calculate the where this particle might be in space is not an easy task, since this was one of the fundamental question which created quantum physics. Light as a bundle of waves as it travels may interfere with itself (different waves) with two outcomes: a) destructive b) constructive. The question is where in time and space does this happen? The answer includes defining a probability that allows us to limit the problem. In quantum physics Ψ is the wave function of a mechanical wave, describing a wave in time and space. Very similar to the wave function used in Optics. In 3-D space Ψ may also be complex Ψ^* and the probability to find one particle in space volume, defined by dx, dy and dz is called the probability function:

$$\int_{-\infty}^{\infty} \Psi \Psi^* dx dy dz = \int_{-\infty}^{\infty} |\Psi|^2 dx dy dz = 1 \quad (\text{A.56})$$

This still doesn't explain where these interferes occur. Heisenberg proved by repeating experiments by measuring particles momentum and position, that no matter how exact the experiments are, there is always an uncertainty Δx where a particle might be. This is known as the uncertainty principle. The Figure below shows what is meant by that:

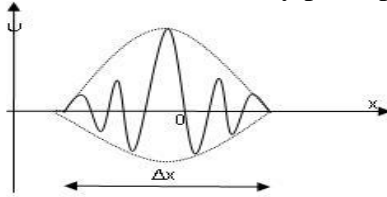


Figure A.21. Group velocity.

Schrodinger equation is the incorporation of Ψ and the energy, also called potential V , to express the space and time dependency that is associated with particle:

$$\frac{\hbar^2}{2m} \left(\frac{\partial^2 \Psi}{\partial x^2} + \frac{\partial^2 \Psi}{\partial y^2} + \frac{\partial^2 \Psi}{\partial z^2} \right) - \nabla \Psi = -i\hbar \frac{\partial \Psi}{\partial t} \quad (\text{A.57})$$

where $\Psi = \Psi(x, y, z, t)$ and $V = V(x, y, z, t)$.

Most analyses ignore t ($t=0$) due its difficulty to find an actual solution. Ruling out t in (A.57) reduces (A.57) to:

$$\frac{d^2 \Psi}{dx^2} + \frac{2m}{\hbar^2} [E - V(x)] \Psi(x) = 0 \quad (\text{A.58})$$

where E is the total energy of the particle. (A.58) remains a difficult equation. Schrödinger equation can be applied to find the states of a particle, such as an electron, in materials like semiconductors and metals. Therefore analysis is been focused at atomic level. A solution to (A.58) can be extracted introducing a simplified model; this model is called potential well: We imagine that only two absolute potential values exist, namely

$$V(x) = \begin{cases} 0, & x = 0 \\ \infty, & x = L \end{cases} \quad (\text{A.59})$$

These two values are called the boundary conditions. A particle cannot jump over a infinitive barrier (∞), hence it is trapped in the well. Inside the potential well (A.58) becomes

$$\frac{d^2 \Psi(x)}{dx^2} + \frac{2m}{\hbar^2} E \Psi(x) = 0 \quad (\text{A.60})$$

A possible solution is

$$\Psi(x) = A \sin kx + B \cos kx \quad (\text{A.61})$$

where A, B are constants and

$$k^2 = \frac{2mE}{\hbar^2} \quad (\text{A.62})$$

Looking what happens (A.62) at the boundaries we have that $\psi(x) = 0$ must hold or else there is $\psi(x) \neq 0$ outside the boundaries, which is unacceptable.

$$\psi(x) = 0 \begin{cases} B = 0 @ x = 0 \\ \sin kx = 0 @ x = L \end{cases} \Rightarrow \psi(x) = A \sin kx \text{ with } k = \frac{n\pi}{L}, n = 1, 2, 3, \dots \quad (\text{A.63})$$

Substituting (A.62) we get

$$\left(\frac{2E_n}{\hbar^2}\right)^{\frac{1}{2}} = \frac{n\pi}{L} \Rightarrow E_n = \frac{n^2 \hbar^2}{8mL^2} \quad (\text{A.64})$$

As one can see the energy is quantized because $kx=kL$ must be multiple of π . This is the reason why n is called the quantum number. With the probability function (A.58) one can find A and rewrite (A.63)

$$\psi_n = \left(\frac{2}{L}\right)^{\frac{1}{2}} \sin \frac{n\pi x}{L} \quad (\text{A.65})$$

For every quantum number correspond one ψ and one E , these are called the eigenfunctions and energy eigenvalues respectively. Both ψ and E are called eigenstate. The one dimensional well can be expanded to a 3D cube well with side L :

$$E_n = \frac{n^2 \hbar^2}{8mL^2}, k_1 = \frac{n_1 \pi}{L}, k_2 = \frac{n_2 \pi}{L}, k_3 = \frac{n_3 \pi}{L} \text{ and } n^2 = n_1^2 + n_2^2 + n_3^2 \quad \text{and}$$

$$\psi_n = \left(\frac{8}{L}\right)^{\frac{1}{2}} \sin k_1 x \sin k_2 y \sin k_3 z \quad (\text{A.66})$$

Figure below is an example of a 3D case:

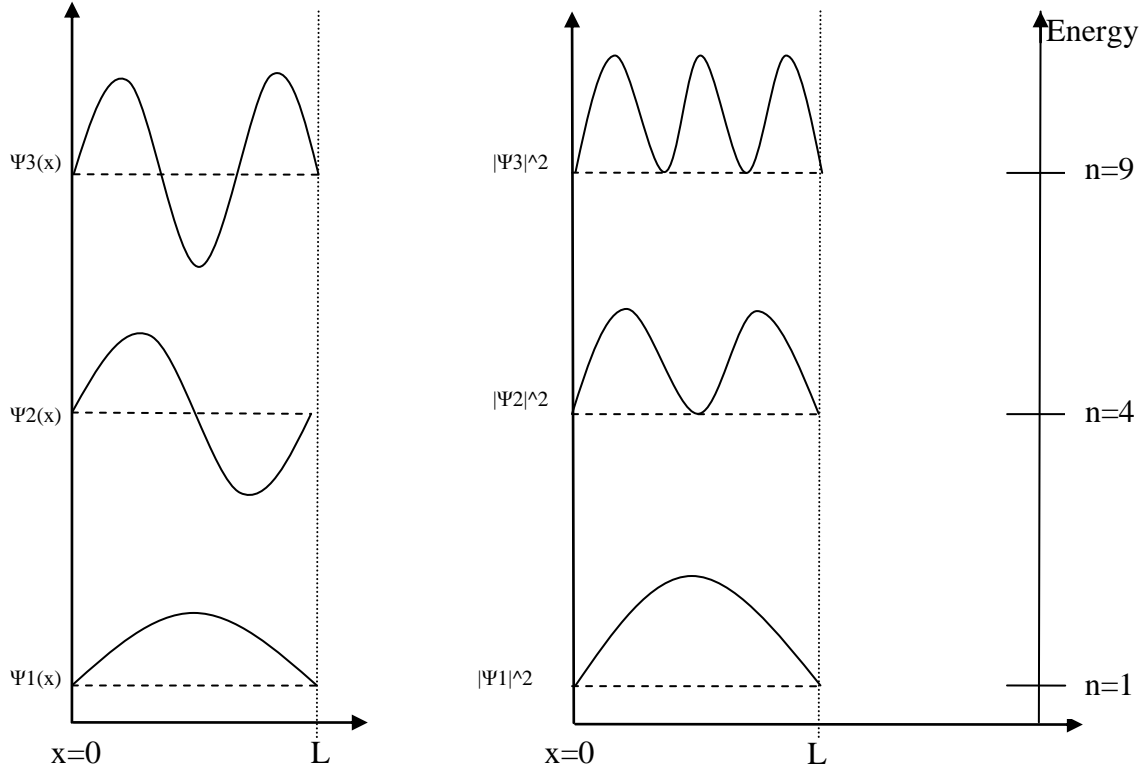


Figure A.22. Three different potentials in a well.

Because more variables appear in the 3D case it's been observed that the eigenfunction can have also "substates", which we call spin m_s . To distinguish two electrons having the same eigenfunction we associate in which direction the electrons in that state spin. Spin

values are -0.5 and 0.5. Finally one can define a state in the 3D potential by the four variables, called quantum numbers (n_1, n_2, n_3 and m_s). Another important rule in the creation of solids is Pauli's rule: It forbids electrons to have the same quantum numbers. This explains why in solids there are gaps in the energy bands. When atoms come together their energy bands “fuse” into each other, because of Pauli's rule, there are not allowed to “spread” into the forbidden zone (in the forbidden zone where no eigenstate exists). Therefore allowed energy levels are shared between atoms, which are called energy bands. Let's examine this phenomenon a little bit closer.

A.3.6.1 The E-k diagram (Dispersion curve)

Knowing the energy levels that a solid might have, allows us to model and extract some of its desired charge properties. In a solid, atoms are closely spaced and thereby interact with each other, resulting in a splitting of energy levels. As all levels are split this gives the formation of energy bands, each band being a collection of closely spaced sublevels as in Figure A.23. A short introduction into the Kronig-Penney model will illustrate the importance of the Brillouin zone. Assume we have a A.D crystal consisting of positive ions with a distance a to each other, like shown in figure A.23

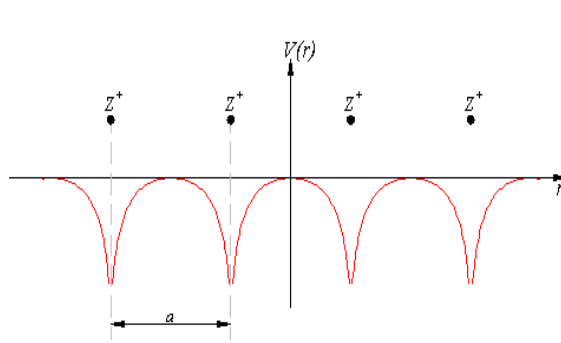


Figure A.23 An A.D crystal example

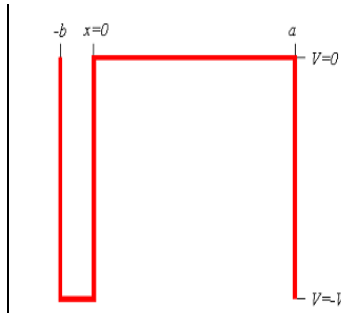


Figure A.24 The simplified Kronig-Penney model of the A.D crystal

Applying Bloch's theorem, we only need to find a solution for a single period, thus simplifying the problem and the potential function is approximated by a rectangular potential

The ions with their charge create an electromagnetic field, which, as known, influence the electrons. This potential is described mathematically by a periodic function with period a . According to Bloch's theorem, the wavefunction solution of the Schroedinger equation if the potential is periodic is

$$\psi(x) = e^{ikx} u(x) \quad \text{with } u(x) \text{ a periodic function}$$

At the Brillouin zone boundary, one finds energy degeneracy that result in a shift in energy given by:

$$E = \frac{\hbar^2 k^2}{2m} \pm |U_k|$$

This energy gap between Brillouin zones is known as the band gap. Also near the edges of the lattice there are certain problems with the boundary conditions, thus to overcome

these problems we represent the ion lattice as a ring following the Born-von Karman boundary conditions. If L is the length of the lattice so that $L \gg a$ (a the distance between two atoms or the period of the potential), we can consider the neighborhood of one ion linear, and the wavefunction of the electron remains unchanged. The result is that boundary conditions become circular boundary conditions:

$$\psi(0) = \psi(L)$$

Assume there are N numbers of Ions in the lattice, and then the relation $aN = L$ holds. Substituting the boundary condition in Bloch equation for periodical lattice we get a quantization of 'electron' wave vector k :

$$\begin{aligned} \psi(0) &= e^{ik0} u(0) = e^{ikL} u(L) = \psi(L) \\ u(0) &= e^{ikL} u(Na) \rightarrow e^{ikL} = 1 \\ \Rightarrow kL &= 2\pi n \rightarrow k = \frac{2\pi}{L} n, n = 0, \pm 1, \pm 2, \dots, \pm \frac{N}{2} \end{aligned} \quad (\text{A.67})$$

where k is the wavevector. As we can see the quantization of the wave gives us a condition where discontinuities between allowed and forbidden energy bands exists.

The corollary of the above condition is due to the periodicity of crystals for a given momentum, many levels of energy are possible, and that some energy might not be available at any momentum. The collection of all possible energies and momenta is known as the band structure of a material. Properties of the band structure define whether the material is an insulator, semiconductor or conductor. The highest energy bands, namely the valence and conduction bands and their associated band gap are crucial in determining the electronic properties of a solid. The more electrons that there are in the conduction band, the greater the conductivity of the material. Equation 1.67 can be written in a more general form:

$$k = \pm \frac{n\pi}{a} \quad (\text{A.68})$$

To see how the Brillouin concept is connected to the Kronig-Penney model we construct the first Brillouin zone of a simple cubic lattice. We have :

$$a_1 = (a, 0, 0), \quad a_2 = (0, a, 0), \quad a_3 = (0, 0, a)$$

From the construction of the Wigner-Seitz cell we find the orthogonals b_i 's

$$b_1 = \left[\frac{2\pi}{a}, 0, 0 \right] \quad b_2 = \left[0, \frac{2\pi}{a}, 0 \right] \quad b_3 = \left[0, 0, \frac{2\pi}{a} \right]$$

The result is that the first Brillouin zone is a cube, like in figure A.24, which extends in all three directions : $-\frac{\pi}{a} \leq k_i \leq +\frac{\pi}{a}$ $i = x, y, z$. The latter and from 1.68 we have:

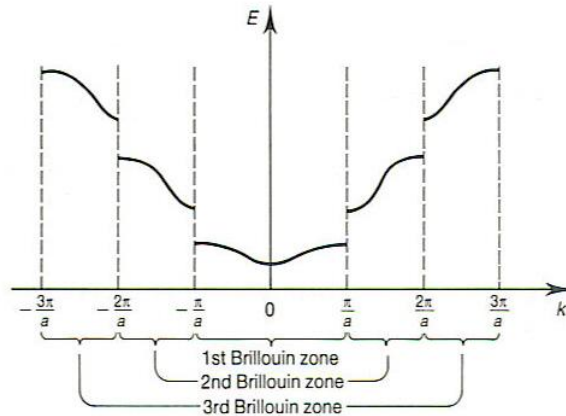


Figure A.25 The energy graph of the Kronig-Penny model and the associated Brillouin zones is shown. Figure from [5]

There also exist higher zones of Brillouin zones and they are formed like shown in the Figure below:

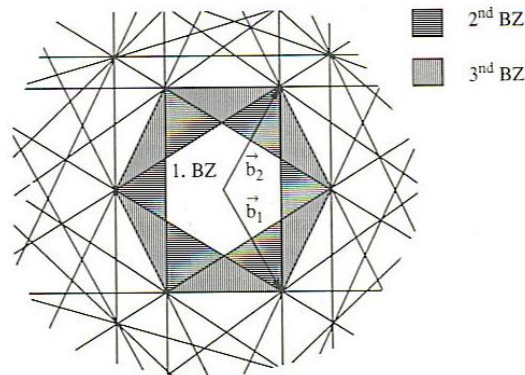


Figure A.26 Construction of the first three Brillouin zones. Figure from [4]

The periodicity of 1.68 and the fact that we work in reciprocal space allows us to shift all energy-wavevector curves in the first Brillouin zone, which gives a so called reduced zone presentation or E-k diagram. Also the diagram often is called dispersion relation due to its representation of allowed and forbidden momentum with relation to the wavevector in the system (here A.D crystal).

The above method has been done for the 3-D crystal. We represent only the outcome of three important E-k diagrams of Ge, Si and GaAs.

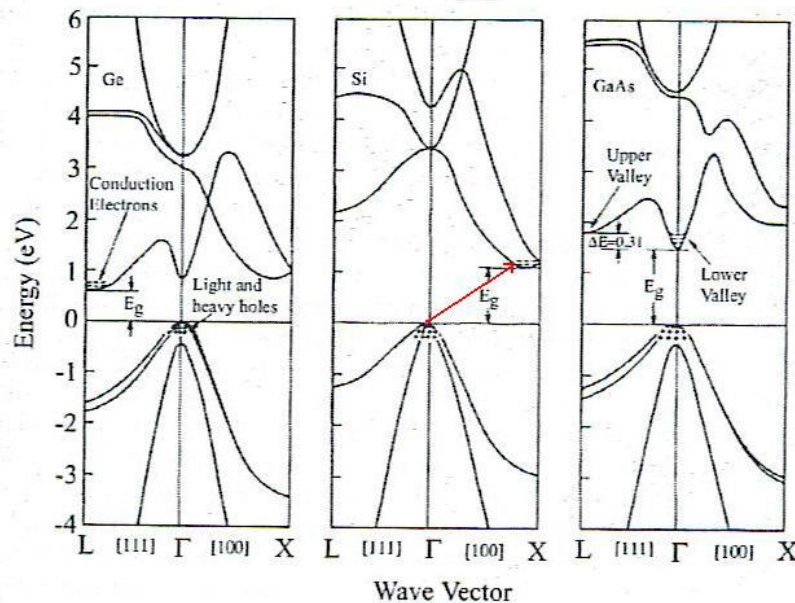


Figure A.27 Energy band structure along $\langle 100 \rangle$ and $\langle 111 \rangle$ axes. The x axis is the k vector. Figure from [6], original source [9].

Why did is a E-k diagram of any importance? The reason is simple, one can see the band gap of the highest energy bands, namely that of the valence and that of the conduction band. Normally in all solids the lower energy bands are filled with electrons, but when an electron, for example through optical stimulation of the atom, is transferred to the conduction band, then it is free to move through the crystal. Therefore it contributes to the conduction of the solid and hence the name conduction band. The value of the band gap defines the kind of the solid. For example isolators have a big band gap and conductors like metal have small values. The value of the band gap of semiconductors lays somewhere between the value of isolators and conductors. Semiconductors are further characterized as indirect or direct band gap materials. The difference between them is the way electrons are promoted to the conduction band. In indirect materials, like Silicon, the electron needs additional energy to make it to the conduction band while for direct materials this isn't needed. The most common form of this additional energy is a phonon, which is created by quantized lattice vibration. Another way to see this is observing in figure A.27 (Si case) that the minimum energy at the point ($E=0$ (y axis), k vector= Γ (x axis)) is not the bandgap energy E_g . Hence for an electron to reach the conduction band, in figure A.27 the red arrow, it needs an additional energy. Next, we will investigate how particles like electrons or holes, contribute to the overall conductivity of the material. To present how electric flow is possible we introduce an even more simplified version of the E-k diagram, see figure A.27. It contains only two states, the valence and the conduction band. An electron is promoted to the conduction band through

- Thermal increase of the material
- Optical stimulation
- Applied Electric field (increases electron collisions with the lattice \rightarrow heat)

The result is that an empty state (hole) is created, which another electron will occupy. Therefore we get an chain reaction and this results in hole movement in the valence band

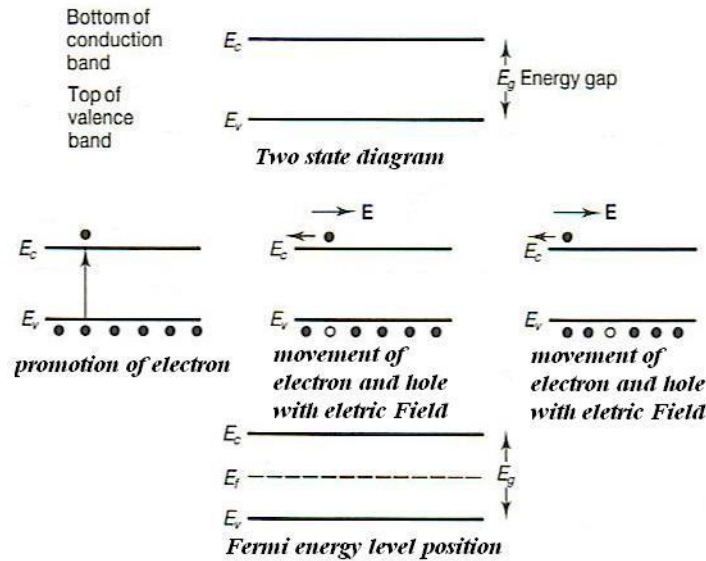


Figure A.28 Simplified E-k diagram to explain the electric flow mechanism. Figure from [5]

A.3.6.2 Bragg's Law

$$n\lambda = 2d \sin \theta \quad (\text{A.69})$$

derived by the English physicists Sir W.H. Bragg and his son Sir W.L. Bragg in 1913 to explain why the cleavage faces of crystals appear to reflect X-ray beams at certain angles of incidence (theta, θ). The variable d is the distance between atomic layers in a crystal, and the variable λ is the wavelength of the incident X-ray beam (see applet); n is an integer

This observation is an example of X-ray wave interference, commonly known as X-ray diffraction (XRD), and was direct evidence for the periodic atomic structure of crystals postulated for several centuries. The Braggs were awarded the Nobel Prize in physics in 1915 for their work in determining crystal structures beginning with NaCl, ZnS and diamond. Although Bragg's law was used to explain the interference pattern of X-rays scattered by crystals, diffraction has been developed to study the structure of all states of matter with any beam, e.g., ions, electrons, neutrons, and protons, with a wavelength similar to the distance between the atomic or molecular structures of interest.

Bragg's Law can be derived by considering the conditions necessary to make the phases of the beams coincide when the incident angle equals and reflecting angle. The rays of the incident beam are always in phase and parallel up to the point at which the top beam strikes the top layer at atom z (Figure A.29). The second beam continues to the next layer where it is scattered by atom B. The second beam must travel the extra distance $AB + BC$ if the two beams are to continue traveling adjacent and parallel. This extra distance must be an integral (n) multiple of the wavelength (A.69) for the phases of the two beams to be the same:

$$n\lambda = AB + BC \quad (\text{A.70})$$

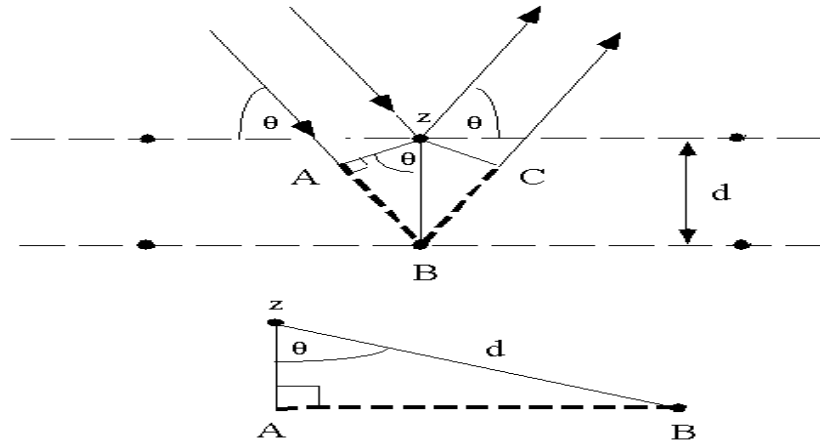


Figure A.29. Bragg angle calculation

Recognizing d as the hypotenuse of the right triangle Abz , we can use trigonometry to relate d and q to the distance $(AB + BC)$. The distance AB is opposite q so,

$$AB = d \sin \theta \quad (\text{A.71})$$

$$(\text{A.70}) \xrightarrow{AB=BC} n\lambda = 2AB \quad (\text{A.72})$$

From (A.70) and (A.72) we get

$$n\lambda = 2d \sin \theta$$

The location of the surface does not change the derivation of Bragg's Law.

A.3.6.3 Bloch's Theorem

Bloch's theorem applies to any periodic potential. For our interest a crystal, like a semiconductor, has this periodic potential. Let R be any vector in a lattice. Let Ψ be a single electron solution to the 1D Schrödinger equation

$$\frac{\hbar^2}{2m} \frac{\partial^2}{\partial x^2} \psi + U(r)\Psi(r) = E\Psi(r) \quad (\text{A.73})$$

where $U(r+R)=U(r)$ for all R belong to the lattice. Then there exists a wavevector k in the reciprocal lattice and a periodic function $u_k(r)$ such that $u_k(r+R)=u_k(r)$ such that Ψ is of the form

$$\Psi = e^{(ikr)} u_k(r) \quad (\text{A.74})$$

The result is that the solution (eigenfunction) in a material with periodic potential is a plane wave modulated by the periodicity of the crystal $u_k(r)$.

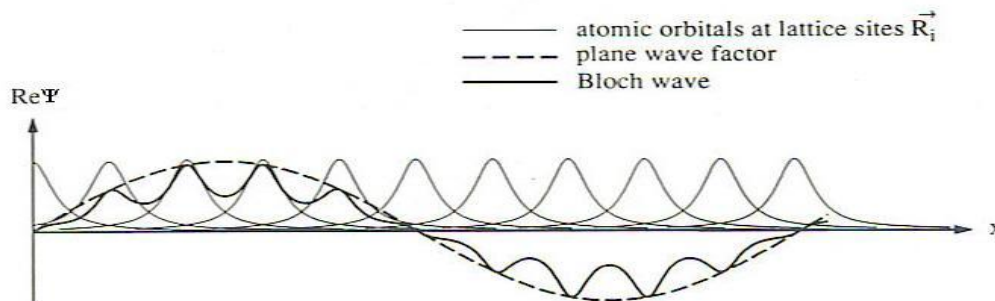


Figure A.30. Bloch wave in a one dimensional potential.

Since Ψ is periodic in reciprocal space, E is also periodic.

$$E(k, i) = E(k + G, i) \quad (\text{A.75})$$

where the translation vector G . We know from the analysis of the Schrödinger equation that that band gap occur at the boundaries

$$k = \frac{n\pi}{a}, n = 1, 2, 3 \dots \text{and } a \text{ is the lattice constant}$$

Therefore because (A.75) is periodic, the energies at different zones defined by k can be shifted, with use of G , into one single zone. This zone is also called the Brillouin zone, or reduced zone. An important fact arises when we examine the boundaries: The group velocity of the waves become zero

$$u_g = \frac{1}{\hbar \partial k_k}_{i=\pm \frac{\pi}{a}} = 0 \quad (\text{A.72})$$

From which we get two solutions of the wave, one having its maxima of $|\Psi|^2$ at potential minima and the other having is maximum $|\Psi|^2$ at maximum potential. This result is illustrated in the figure below.

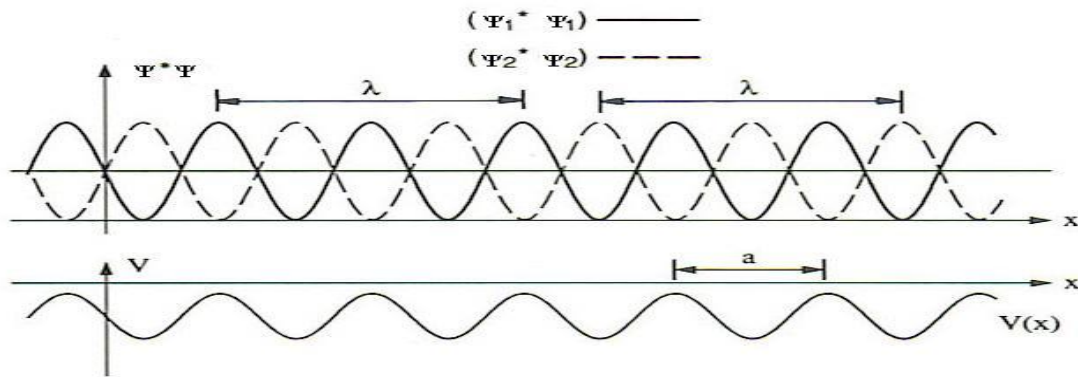


Figure A.31. Two waves having the same magnitude of k but different potential.

A.4 References

- [1] Balas, Konstantin., "Introduction to Optoelectronics "Εισαγωγή στην Οπτοηλεκτρονική"." Chania "Χανιά" : Technical University of Crete "Πολυτεχνείο Κρήτης", 2006.
- [2] Chuang, Shun Lien., *Physics of optoelectronics devices*. s.l. : A Wiley-Interscience Publication John Wiley & Sons Inc., 1995.
- [3] Young, Hough D., Freedman, Roger A and Ford, Levis A., *Univeristy Physics*. s.l. : Pearson, Addison Wesley, 2007.
- [4] Klingshirm, C F., *Semiconductor Optics*. s.l. : Springer, 1997.
- [5] Wood, David., *Optoelectronics Semiconductor Devices*. s.l. : Prentice Hall International Series in Optoelectronics, 1994.
- [6] Percht, Orly Yardly and Cummings, Ralph Etienne., *CMOS imagers: From phototransuction to image processing*. Dortrecht (NL) : Kluwer academic publisher, 2004.
- [7] Sze, S M., *Physics of semiconductor devices*. New York : John Wiley and Sons, 1981.

APPENDIX B

B.1 Introduction to the z-transform

The z-transform is used mainly in the characterization of digital filters. We introduce here the basic understating of the z-transform and its common application in the decimator filter used in the sigma delta ADC. The z-transform is a conversation between the discrete time-domain and the complex frequency domain. An analogy is to sample a signal in the time domain. The goal, for every transformation, is similar to that of the transformation of number to its logarithm equivalent. One must look at it as always being a tool for making complex calculations easier. This is done by simply transforming the expression from one domain to the other, and then with the transformed expression, calculations can be lesser complex. For digital filter the transformation (relationship) between the Laplace transform and the z-transform is of particular interest and therefore we are interested solely in that region of transformation. Here are the basic transforms used:

The two sided $(-\infty \leq n \leq +\infty)$ z-transform with n as integer:

$$X(z) = Z\{x|n|\} = \sum_{n=-\infty}^{\infty} x(n)z^{-n}$$

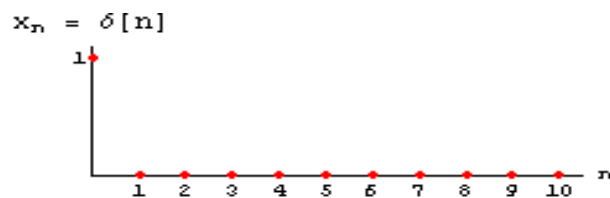
Example:

$$x_n = \delta(n) = \begin{cases} 1, & \text{if } n = 0 \\ 0, & \text{if } n \neq 0 \end{cases}$$

Z-transform:

$$X(z) = Z\{x|n|\} = \sum_{n=0}^{\infty} x(n)z^{-n} = 1 + \sum_{n=0}^{\infty} 0 * z^{-n} = 1$$

Graphically:



Since the z-transform is a summation of discrete values one must look at the region of convergence (ROC) to confirm if the transformation can exist.

Example:

$$x_n = 0.5^n \text{ expanding it to } (-\infty, +\infty) \text{ yields : } x_n = (\dots, 0.5^{-2}, 0.5^{-1}, 1, 0.5^1, 0.5^2, \dots) \Rightarrow$$

$$x_n = (\dots, 2^2, 2, 1, 0, 0.5^1, 0.5^2, \dots)$$

Z-transform:

$$X(z) = Z\{x|n|\} = \sum_{n=-\infty}^{\infty} x(n)z^{-n} \leq \infty$$

Impossible for any z to convergence!

Example:

$x_n = 0.5^n u(n)$ expanding it to $(-\infty, +\infty)$ yields : $x_n = (\dots, 0, 0, 1, 0.5^1, 0.5^2, \dots) \Rightarrow$

$$x_n = (\dots, 0, 0, 1, 0.5^1, 0.5^2, \dots)$$

Z-transform:

$$X(z) = Z\{x[n]\} = \sum_{n=0}^{\infty} x(n)z^{-n} = \sum_{n=0}^{\infty} 0.5^n z^{-n} = \sum_{n=0}^{\infty} \left(\frac{0.5}{z}\right)^n = \frac{1}{1 - 0.5z^{-1}}$$

Therefore it convergence for $|1 - 0.5z^{-1}| \leq 1$ or for any $|z| > 0.5$

The above example shows that the condition convergence can be graphically interpreted as for most of the cases when the z condition forms a circle. For the above example the ROC is

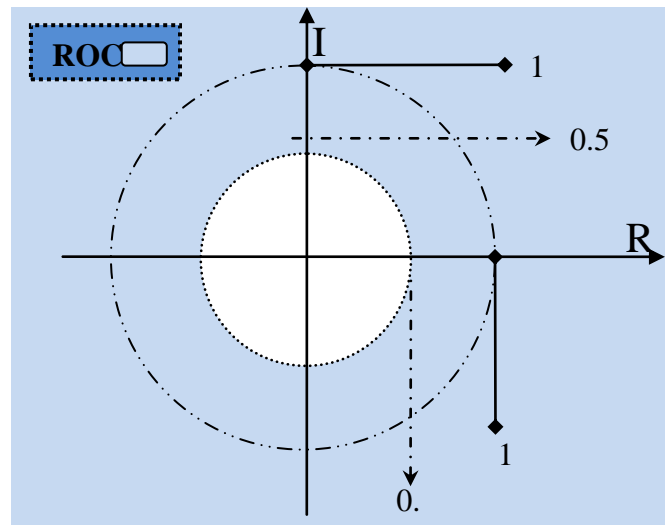


Figure B.1. Region of converge example

Common z-transforms:

1	$\delta[n]$	1	all z
2	$\delta[n - n_0]$	z^{-n_0}	$z \neq 0$
3	$u[n]$	$\frac{1}{1 - z^{-1}}$	$ z > 1$

4	$-u[-n-1]$	$\frac{1}{1-z^{-1}}$	$ z < 1$
5	$nu[n]$	$\frac{z^{-1}}{(1-z^{-1})^2}$	$ z > 1$
6	$-nu[-n-1]$	$\frac{z^{-1}}{(1-z^{-1})^2}$	$ z < 1$
7	$n^2u[n]$	$\frac{z^{-1}(1+z^{-1})}{(1-z^{-1})^3}$	$ z > 1$
8	$-n^2u[-n-1]$	$\frac{z^{-1}(1+z^{-1})}{(1-z^{-1})^3}$	$ z < 1$
9	$n^3u[n]$	$\frac{z^{-1}(1+4z^{-1}+z^{-2})}{(1-z^{-1})^4}$	$ z > 1$
10	$-n^3u[-n-1]$	$\frac{z^{-1}(1+4z^{-1}+z^{-2})}{(1-z^{-1})^4}$	$ z < 1$
11	$a^n u[n]$	$\frac{1}{1-az^{-1}}$	$ z > a $
12	$-a^n u[-n-1]$	$\frac{1}{1-az^{-1}}$	$ z < a $
13	$na^n u[n]$	$\frac{az^{-1}}{(1-az^{-1})^2}$	$ z > a $

14	$-na^n u[-n-1]$	$\frac{az^{-1}}{(1-az^{-1})^2}$	$ z < a $
15	$n^2 a^n u[n]$	$\frac{az^{-1}(1+az^{-1})}{(1-az^{-1})^3}$	$ z > a $
16	$-n^2 a^n u[-n-1]$	$\frac{az^{-1}(1+az^{-1})}{(1-az^{-1})^3}$	$ z < a $
17	$\cos(\omega_0 n) u[n]$	$\frac{1-z^{-1}\cos(\omega_0)}{1-2z^{-1}\cos(\omega_0)+z^{-2}}$	$ z > 1$
18	$\sin(\omega_0 n) u[n]$	$\frac{z^{-1}\sin(\omega_0)}{1-2z^{-1}\cos(\omega_0)+z^{-2}}$	$ z > 1$
19	$a^n \cos(\omega_0 n) u[n]$	$\frac{1-az^{-1}\cos(\omega_0)}{1-2az^{-1}\cos(\omega_0)+a^2z^{-2}}$	$ z > a $
20	$a^n \sin(\omega_0 n) u[n]$	$\frac{az^{-1}\sin(\omega_0)}{1-2az^{-1}\cos(\omega_0)+a^2z^{-2}}$	$ z > a $

Table B.1. Common z-transforms

B.1.1 Laplace relationship

A simple method of obtaining the relationship between the Laplace and the z-transform can be achieved by substituting $s = j\omega$ in the Laplace and $z = e^{j\omega t}$ in the z-transform respectively:

$$F(s) = \int_0^\infty e^{-st} f(t) dt \Rightarrow F(s) = \int_0^\infty e^{-j\omega t} f(t) dt \quad (\text{B.1})$$

$$F(z) = \sum_{n=0}^\infty f(nt) z^{-n} \Rightarrow F(z) = \sum_{n=0}^\infty f(nt) e^{-j\omega t} \quad (\text{B.2})$$

The Laplace transform can be interpreted with the s-plane (frequency plane), while the z-transform, like seen before, can be interpreted with ROC (real and imaginary plane). If we had one plane (s plane) of one transform and want to translate it into the other (z plane), the following rules must be followed:

- i. *The left part of the s-plane translates within the unit circle*
- ii. *The distance f_s along the real frequency axis translates into the periphery of the unit circle*
- iii. *A pole outside the unit circle results in unstable system*
- iv. *When dealing with multiple poles the first pole marks a marginal function of the system, but other poles increase the instability of the system.*
- v. *Poles inside the unit circle give stable system*
- vi. *Without affecting the systems stability a zero can be placed anywhere*

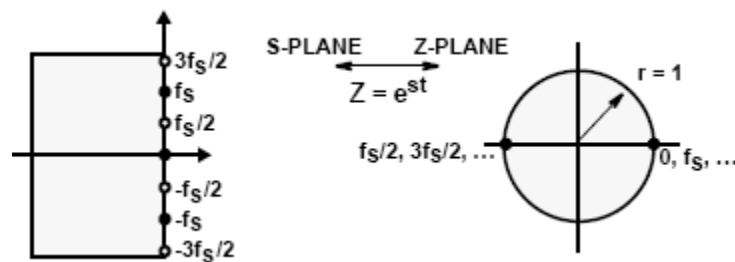


Figure B2. An example of transform translation

An important addition to table APPXX-1 is the characterization that a z^{-1} in a z-transform translates into a unit time delay in the time domain.

References

- [1] Zervakis, Michael E., "Introduction To Digital Signal Processing." *Digital Signal Processing Notes-Set1*. Chania : Digital Image & Signal Processing Laboratory of Technincal University of Crete, 2007.
- [2] Jarman, David., "A Brief Introduction to Sigma Delta Conversion." *Application Note*. s.l. : Intersil, 1995. Vol. AN9504.

APPENDIX C

C.1 Transistors and models

In this section we will investigate in very short manner two basic transistors. All of the transistors are formed with 3 junctions, consisting of n or p-type materials. The basic idea of all transistors is explained in the figure C.1 where terminals are labeled A, B and C. External circuits are connected via terminal A and B. The property is that the transistor is determined by the input of terminal C. Ideally the input shouldn't take any power from the circuit that provides it. The input then should have two properties: One being that the

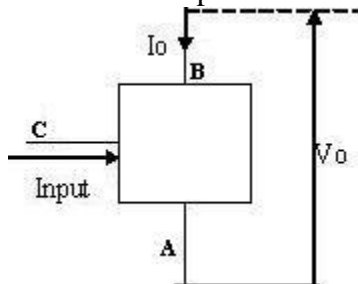


Figure C.1 General transistor

input impedance is zero and the input must be a controllable current, and the other is being the input impedance being infinite and the input must be a controllable voltage between A and C. Hence we got two types of transistor: the field effect transistor which can only be used as high impedance device, while the other the bipolar transistor is controlled better by voltage. The MOS transistor is the most widely used transistor but more importantly we will implement with MOS, hence we will emphasize on MOST.

C.1.1 Bipolar transistor (BJT)

A BJT consists of two pn junctions back-to-back with the distance between the two junctions being very small in the order of a few microns. The resulting structure is either pnp or npn. The function of a BJT (npn) is as follows: If a voltage is applied between the base and the emitter, we achieve in that way a more positive base in relation to the emitter. In this case a current crosses the emitter-base junction. Because the emitter is heavily doped, much more than the base, the emitter-base current consist of electrons and the direction of the current is towards the base. If now a voltage is applied between the base and the collector, so to make the pn junction reversed biased, then the diffusion length of the minority carriers in the base is larger than the active region. Therefore electrons are influenced by diffusion and make their way from the emitter to the base where they are swept into the collector.

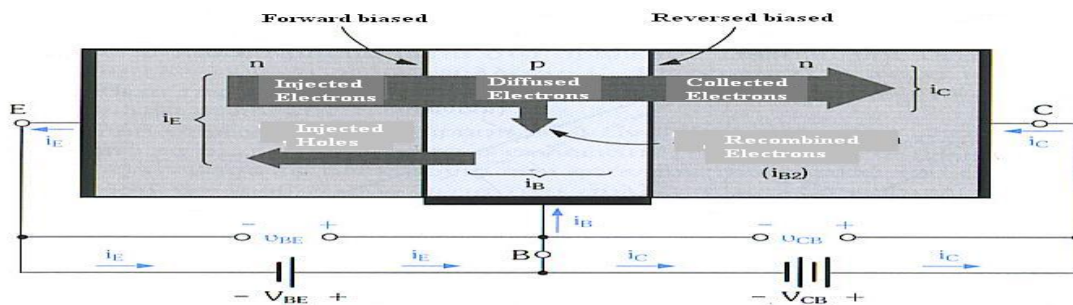


Figure C.2 BJT transistor showing all currents and voltages (Sedra & Smith, 2004)

We present here only the most important equations that exist in the transistor. They are also depicted in the figure above.

Collector current is found like the current in a pn junction and is equal to

$$i_c = I_S e^{\frac{V_{BE}}{V_T}} \quad (C.1)$$

The base current is equal to $i_B = i_{B1} + i_{B2}$, where i_{B1} is the diffused and i_{B2} is the injected current. It's found that i_{B2} is proportional to $e^{\frac{V_{BE}}{V_T}}$ and therefore

$$i_b = \frac{i_c}{\beta} = \frac{I_S}{\beta} e^{\frac{V_{BE}}{V_T}} \quad (C.2)$$

where the β constant depends on the transistor.

The collector current is the sum

$$i_E = i_c + i_B \Rightarrow i_E = \frac{\beta+1}{\beta} I_S e^{\frac{V_{BE}}{V_T}} \quad (C.3)$$

With arithmetic means we find

$$i_c = \alpha i_E, \alpha = \frac{\beta+1}{\beta} \text{ and } \beta = \frac{\alpha}{\alpha+1} \quad (C.4)$$

C.2 MOS transistor (MOST)

As we saw for the BJT transistor, we can manipulate the outcome currents with different junctions. The same idea is used in the MOS transistor, but with a difference that the current can be controlled symmetrically. The basic structure is shown in figure C.3. There are two heavily doped regions (n^+ , p^+) which extend due to fabrication process under the gate. The gate is made out of polysilicon(metal) and is therefore heavily doped (n or p type). For this reason it is isolated and shielded from other parts of the MOS with material known as silicon oxide. The body also known as substrate and is either p or n . Hence the name: metal oxide semiconductor. For example figure C.3 shows a n -type MOS.

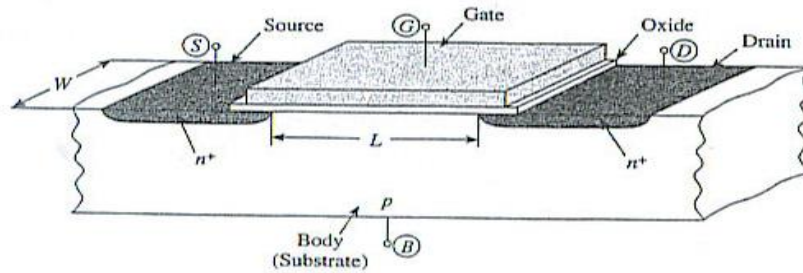


Figure C.3 An overview of an n-channel MOSFET. Figure from (1)

In the figure V_G is positive and attracts electrons from the heavy doped regions n^+ regions. The situation is similar to the situation in the MOS capacitor where we called the MOS being in inversion mode. The electrons form a layer which is called inversion layer. The substrate (p -type) and the n^+ regions form a np junction in reversed bias. Let's assume the case where the drain potential is more positively (has higher potential) then the source. Then the depletion region between the substrate and the drain is bigger than the source-body counterpart. For this reason some electrons can get loose and build up at

the source. Raising more the potential will lose and attract even more electrons. To distinguish between more or less electron concentrations with applied voltage, regions of inversion are defined. These regions are called weak, moderate and strong inversion. Applying in strong inversion a positive voltage between the terminals source and gate V_{DS} , we force the electrons to move via the inversion layer, or better called now channel, to travel from source to the drain. Hence the names at the drain the electrons vanishes and at the source the electrons are replenished. Exactly this process causes a current to flow, which is analogous to the applied voltage V_{DS} . An interesting fact occurs when the voltage exceeds a voltage called pinch-off voltage: the current saturates since the potential at the drain is so high that it drains all electrons nearby, with the result that it shortens the inversion channel (pinch off). With the new voltage definition we clearly can divide the I-V curve shown in figure C.4 into two regions the saturation region and the non-saturation region.

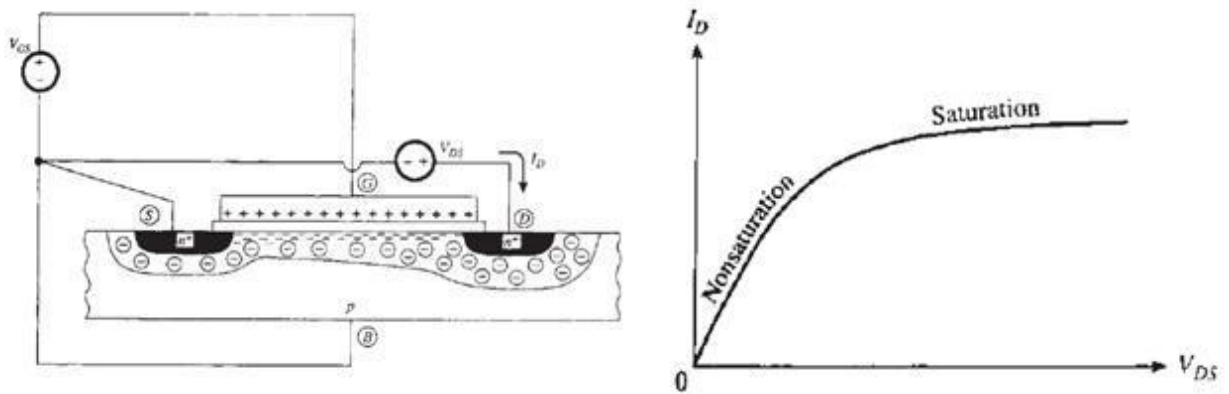


Figure C.4. An overview of a biased n-channel MOSFET and the I-V plot. Figure from (1)

If we connect the substrate with a negative pole and the drain with a positive then we will witness a effect called body effect. This effect stripes away electrons from the source and hence the current in the channel decreases. Note that the substrate and the two n^+ regions are effectively shielded by the depletion region. If that's not the case, namely the substrate- n^+ junction is not in reverse bias, we have substrate conductance, also called having a back gate, which can destroy the MOS transistor (if the transistor is destroyed or not, depends solely on the applied voltage). For normal operation it is recommended to keep $V_{GB} > V_{SB}$. The body effect changes as voltage changes, therefore expressions, found in (1), for the body effects are defined. The important fact to take away from the body effect is the fact that it influences the transition voltages (V_W , V_M , V_S) of the operation regions of the transistor. Most influenced by the body effect is the region between the weak and moderate/strong inversion, there the threshold voltage (onset voltage or separation voltage) is not well defined as in the MOS capacitor. In the MOS capacitor the threshold voltage V_{t0} is found by setting $V_{SB}=0$. Due to the influence, the threshold voltage changes and its real value must be recalculated. The new threshold voltage is called extrapolated threshold, and is called threshold voltage V_T . The V_T is then expressed with the terminal voltage dependency for instance in strong inversion (1)

$$V_T = V_{T0} + \gamma(\sqrt{V_{SB} + \Phi_0} - \sqrt{\Phi_0}) \quad (C.5a)$$

where γ is the body effect coefficient and expresses how much change in the onset or separation voltages (V_W , V_M , V_S) will be for a given V_{SB} . The “zero bias $V_{SB}=0$ ” threshold is defined as

$$V_{T0} = V_{FB} + \Phi_0 + \gamma\sqrt{\Phi_0} \quad (C.5b)$$

where γ is defined as

$$\gamma = \frac{\sqrt{2q\epsilon_s N_A}}{C'_{ox}} \quad (C.6)$$

Next we will define the various way of expressing the channel current.

C.2.1 Charge sheet models

To derive the current equation I_D due to the inversion channel in MOS transistor terminal voltages, geometrical dimensions, charges and potentials have to be defined for a certain region of operation. We assume that transistor parts are perfectly shielded and therefore

$$I_G = 0 \quad (C.7)$$

$$I_B = 0 \quad (C.8)$$

also, we have uniform charge distribution in the channel. This allows us to write the charges as charge (Q) per unit area (example ΔA) at a certain point (x) in the channel:

$$\Delta Q'_I = \frac{\Delta Q_I}{\Delta A} \xrightarrow{\lim_{\Delta A \rightarrow 0} \Delta Q'_I} Q'_I = \frac{d(\Delta Q'_I)}{dA} \quad (C.9)$$

$$\Delta Q'_B = \frac{\Delta Q_B}{\Delta A} \xrightarrow{\lim_{\Delta A \rightarrow 0} \Delta Q'_B} Q'_B = \frac{d(\Delta Q'_B)}{dA} \quad (C.10)$$

$$\Delta Q'_G = \frac{\Delta Q_G}{\Delta A} \xrightarrow{\lim_{\Delta A \rightarrow 0} \Delta Q'_G} Q'_G = \frac{d(\Delta Q'_G)}{dA} \quad (C.11)$$

In general the regions of operation can be separated by looking at which end of the channel is more inverted. Assuming having a n-channel MOS transistor, the drain is the more inverted. Table C.1 can also be used in p-channel:

Region	More heavy inverted channel end condition
Strong inversion	End is strong inverted
Moderate inversion	End is moderate inverted
Weak inversion	End is weak inverted

Table C.1 Region of operation of MOS

C.2.1.1 Complete charge sheet model

The first term of the title reveals the purpose of this section: Finding an expression of the current which is valid in all regions of operation. The rest of the title refers to how to model for the current, namely taking a part (sheet) of the inversion channel and establish charge equations for which immediately the current can be found. This is depicted in figure C.5

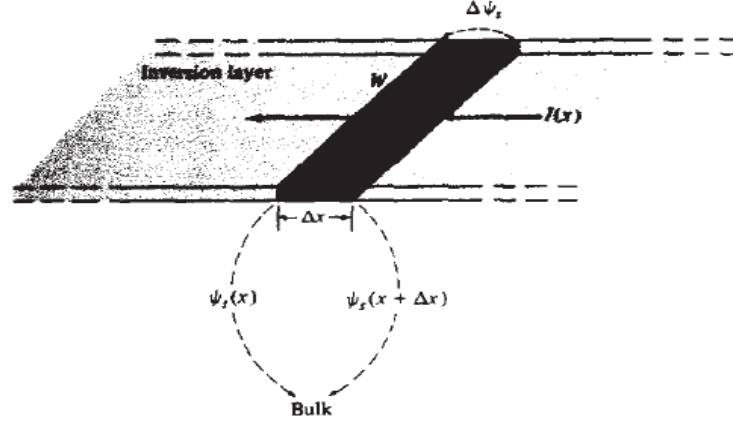


Figure C.5 The charge sheet model, showing relations surface-channel-bulk potentials. Figure from (1)

As we saw many times the current is the result of the drift and the diffusion component. Hence

$$I(x) = I_{drift}(x) + I_{diff}(x) \quad (C.12)$$

In figure C.5 we can deduce the relationship

$$\Delta\psi_s(x) = \psi_s(x + \Delta x) - \psi_s(x) \quad (C.13)$$

where $\psi_s(x)$ is the potential from the channel to the bulk. In (1) analytic expressions are derived for the drift and diffusion current as well for the coming equations, for the currents it has been shown that they have been expressed by the equations (C.7), (C.10) and (C.10). We represent only the results here:

$$I_D = I_{DS} = \underbrace{\mu W (-Q'_I) \frac{d\psi_s(x)}{dx}}_{I_{drift}(x) \text{ "IDS1"}} + \underbrace{\mu W \phi_t \frac{dQ'_I}{dx}}_{I_{diff}(x) \text{ "IDS2"}} \quad (C.14)$$

where ϕ_t is the thermal voltage. Integrating (C.12) from 0 to L, L as the length of the channel, results in two currents I_{DS1} and I_{DS2} . Next expressing Q'_I as a function of ψ_s will help us to evaluate I_{DS1} and I_{DS2} . This can be done by setting ψ_s and V_{CB} as independent variables then

$$Q'_I = -C'_{ox}(V_{GB} - V_{FB} - \psi_s + \frac{Q'_B}{C'_{ox}}) \quad (C.15)$$

where Q'_B is the charge due to ionized atoms in the depletion region and γ is the body effect coefficient defined in (C.6). Substituting Q'_B in (C.15) gives

$$Q'_I = -C'_{ox}(V_{GB} - V_{FB} - \psi_s + \gamma\sqrt{\psi_s}) \quad (C.16)$$

With substitution of (C.16) into (C.14)

$$I_{DS1} = \frac{W}{L} \mu C'_{ox} \left[(V_{GB} - V_{FB})(\psi_{sL} - \psi_{s0}) - \frac{1}{2}(\psi_{sL}^2 - \psi_{s0}^2) - \frac{2}{3}\gamma(\psi_{sL}^{\frac{3}{2}} - \psi_{s0}^{\frac{3}{2}}) \right] \quad (C.17)$$

$$I_{DS2} = \frac{W}{L} \mu C'_{ox} \left[\phi_t(\psi_{sL} - \psi_{s0}) + \phi_t \gamma (\psi_{sL}^{\frac{1}{2}} - \psi_{s0}^{\frac{1}{2}}) \right] \quad (C.18)$$

Solving for ψ_{s0} and ψ_{sL} we have

$$\psi_{s0} = V_{GB} - V_{FB} - \gamma \sqrt{\psi_{s0} + \phi_t e^{\frac{\psi_{s0} - 2\Phi_F - V_{SB}}{\phi_t}}} \quad (C.19)$$

$$\psi_{sL} = V_{GB} - V_{FB} - \gamma \sqrt{\psi_{sL} + \phi_t e^{\frac{\psi_{sL} - 2\Phi_F - V_{DB}}{\phi_t}}} \quad (C.20)$$

A plot of the surface potential at the source versus the drain is showed in figure C.4. The previous outcome shows the weakness of the analysis. We supposed that the electric field

in the channel is vertical, but in reality there is also a horizontal component increasing as one gets nearer to the drain. This effect is neglected in the analysis and the potential transits smoothly from one region to the other region of operation. In reality the currents are discontinuous at the boundaries of transition. For long channel devices this isn't very harmful as long as L stays long, but for short channel devices the inaccuracy at the vicinity of the transition regions increases. A two dimensional analysis (vertical and horizontal electric field) would be far too complex and as shown the one-dimensional approach brings good results. Nonetheless it's a good approximation on which many models can be built, so one could refer to it as the "mother-approach". The derived models ("siblings") are all trying to get around the problem of the $3/2$ and $1/2$ power of (C.17), each in a different way. The most important of these derived models are shortly referred here:

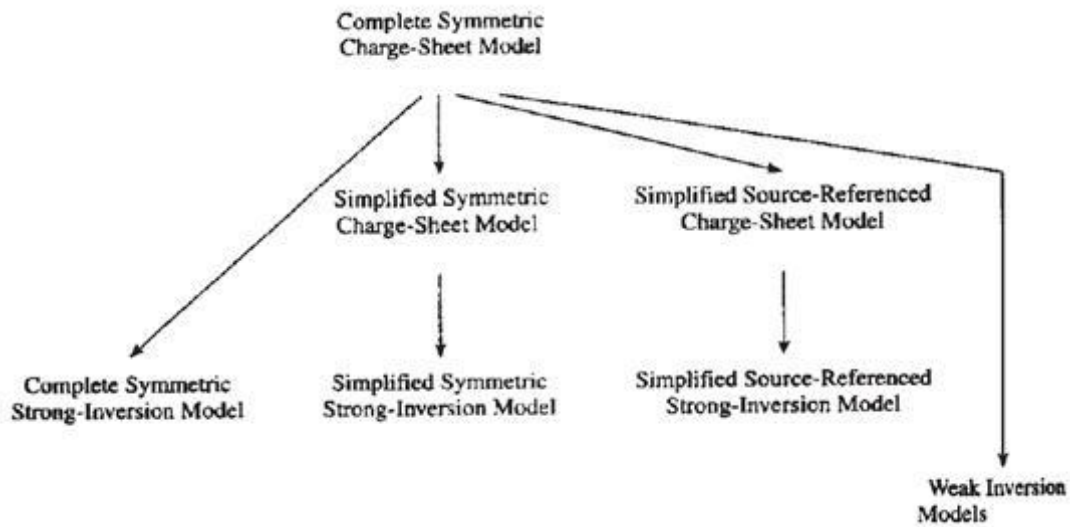


Figure C.6 The charge sheet model and his siblings, showing relations surface-channel-bulk potentials. Figure from (1)

Model	Characteristics/Function
1. Simplified Charge sheet model	<ul style="list-style-type: none"> • Takes advantage of the little change of slope $\frac{Q'_B}{C'_{ox}}$. • Expressing the slope with the surface potential along the channel. • Accuracy depends on choosing appropriate surface potential expression and position • Former leads to source-reference model (ψ_{s0}) and symmetric model ($\psi_{se} = \psi_{sa}$). • $I_{DS} = \underbrace{I_{DS1}}_{I_{Forward}} - \underbrace{I_{DS2}}_{I_{Reverse}} = \underbrace{\mu \frac{W}{L} \frac{Q'_{I0}{}^2}{2nC'_{ox}} - \phi_t Q'_{I0}}_{I_F} + \underbrace{\mu \frac{W}{L} \frac{Q'_{IL}{}^2}{2nC'_{ox}} - \phi_t Q'_{IL}}_{I_R}$

(C.21)	
2. Quasi-Fermi Potentials	<ul style="list-style-type: none"> • Drift and Diffusion currents (C.9)-(C.11) combined into a single expression $Q_I \wedge' (dV/dx)$. Where V is the “quasi Fermi potential”, independent of the channel thickness. • Requires complex calculations
3. Complete Symmetric Strong Inversion Model	<ul style="list-style-type: none"> • Version of simplified charge sheet model for strong inversion • Results are mainly functions of terminal voltages, instead of potentials. Inserting $\psi_{s0} = \Phi_0 + V_{SB}$ and $\psi_{sL} = \Phi_0 + V_{DB}$ into (C.17) gives the appropriate equation for the new I_{DS}. Due to non-uniformities of the substrate, Φ_0 is only an approximation and set to $\Phi_0 = 2\Phi_F + \Delta\Phi$, where $\Delta\Phi \approx 6\Phi_t$. • Major role of defining regions in saturation plays the pinch-off voltage $V_P = \left(-\frac{\gamma}{2} + \sqrt{\frac{\gamma^2}{4} + V_{GB} - V_{FB}}\right)^2 - \Phi_0$ (C.22) • Saturation current takes different values, analogous in which region it is defined. Setting in the new I_{DS} $I_{DS} _{V_{DB}=V_P} = I'_{DS \text{ Forward Sat.}} \quad I_{DS} _{V_{SB}=V_P} = I''_{DS \text{ Reverse Sat.}}$ <p style="text-align: center;">and</p> $I_{DS} \begin{cases} I_{DS \text{ Non-Sat.}} , V_{DB} \leq V_P \\ I'_{DS} , V_{DB} > V_P \end{cases}$

This can be better seen graphically in the following figure

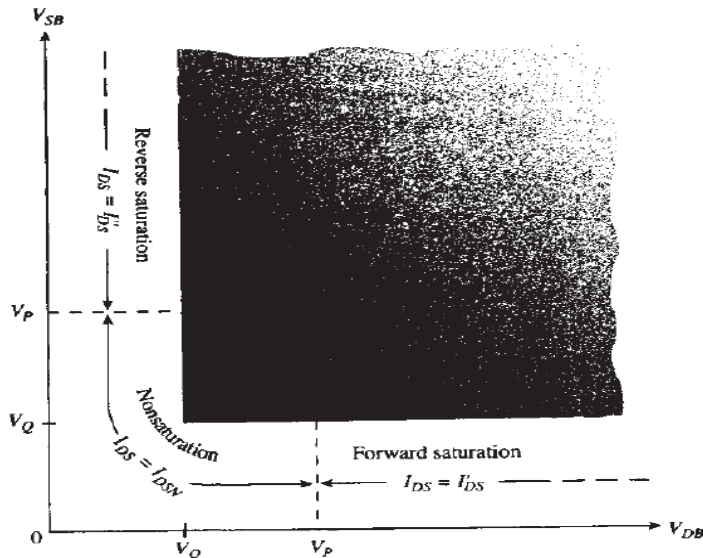


Figure C.7. The strong inversion regions and its associated currents. Figure from (1)

4. Simplified Symmetric Strong	<ul style="list-style-type: none"> • Assumes the diffusion component to be small and therefore it can be neglected. Then I_{DS} can be written as
--------------------------------	---

Inversion Model	$I_{DS} = \mu \frac{W}{L} \frac{(Q'_{I0}{}^2 - Q'_{IL}{}^2)}{2nC'_{ox}} \Rightarrow I'_{DS} = \mu \frac{W}{L} C'_{ox} \frac{n}{2} (V_P - V_{SB})^2 \text{ and}$ $I''_{DS} = -\mu \frac{W}{L} C'_{ox} \frac{n}{2} (V_P - V_{DB})^2 \text{ where } n \text{ is the slope of the}$ <p>difference between the pinch off voltage V_P and its associated potential ψ_{sa} for a given V_{GB}.</p> <ul style="list-style-type: none"> Using the approximation of V_P : $V_P \approx \frac{V_{GB} - V_{T0}}{n}$. The current I_{DS} is then rewritten as $I_{DS} = \mu \frac{W}{L} C'_{ox} \left[(V_P - V_{T0})(V_{DB} - V_{SB}) - \frac{n}{2} (V_{DB}^2 - V_{SB}^2) \right] \quad (C.23)$ <p>Seen in (C.23) the symmetry, namely that the current is equally influenced with terminal voltages V_{DB} and V_{SB}. Taking the derivative of (C.24) with the terminal voltage V_{DB} and setting $V_{DB} = V_P$ we get the forward current. With the same procedure for V_{SB} we get the reverse current.</p>
5. Simplified Source Referenced Strong Inversion Model	<ul style="list-style-type: none"> The Simplified Charge sheet model in conjunction with (C.16) and (C.17) can give us an expression of the drift and diffusion current from which it is distinctly seen that the drain surface appears only in a difference with the surface potential. Hence the name of the model. With the alternative expression of the potentials ψ_{s0} and ψ_{sL} and defining $V_{DB} - V_{SB} = V_{DS}$, $V_{G4} - V_{SB} = V_{GS}$ we present the outcome of the calculation of the current $I_{DS} = \frac{W}{L} \mu C'_{ox} \left[(V_{GS} - V_T _{V_{SB}}) V_{DS} - \frac{a}{2} V_{DS}^2 \right] \quad (C.24)$ <p>where a is a inaccuracy factor given $a = 1 + \gamma / (2\sqrt{\psi_{sa}})$ and $V_T _{V_{SB}}$ is the threshold voltage when $V_{SB} = V_{CB}$.</p> <ul style="list-style-type: none"> This I_{DS} has its maximum value at V_{DS}' which is found to be $V_{DS}' = \frac{V_{GS} - V_T}{a} \quad (C.25)$ <ul style="list-style-type: none"> With V_{DS}' the current then can be divided into two parts, a saturation part and a non-saturation part: $I_{DS} \begin{cases} \frac{W}{L} \mu C'_{ox} \left[(V_{GS} - V_T) V_{DS} - \frac{a}{2} V_{DS}^2 \right], & V_{DS} \leq V_{DS}' \\ \frac{W}{L} \mu C'_{ox} \left[\frac{(V_{GS} - V_T)^2}{2a} \right], & V_{DS} > V_{DS}' \end{cases} \quad (C.26)$ <p>Figure C.7 shows the I_{DS}-V_{DS} curve.</p> <ul style="list-style-type: none"> If a is taken to close to one then (C.26) becomes the well known quadratic MOSFET current equation

$$I_{DS} \begin{cases} \frac{W}{L} \mu C'_{ox} \left[(V_{GS} - V_T) V_{DS} - \frac{1}{2} V_{DS}^2 \right], & V_{DS} \leq V'_{DS} \\ \frac{W}{L} \mu C'_{ox} \left[\frac{(V_{GS} - V_T)^2}{2} \right], & V_{DS} > V'_{DS} \end{cases}$$

(C.27)

C.3 BSIM models

As we saw from the short analysis above the current I_{DS} is a very important and calculations of it demand the continuity of the derivatives. This is especially true for the moderate region, where the current is difficult to derive. For that reason various models for the transistor have occurred. These models are designed to approximate the current in all regions of inversion and to be as accurate and simple as possible. The approximation approximates also various phenomena's not mentioned until now, like the capacitor effects in MOSFET or the geometry influence. The capacitances for instance can affect greatly the current. Therefore the models are also called interpolation models, since they take into account all possible deviations.

A first simulation program for MOS transistors was SPICE, since then a number of models have been developed. Besides of public models also proprietary models have been developed, for example HSPICE. The most widely used, well documented and known model is the UC Berkeley's SPICE. Originally, SPICE came with three MOS models known as level 1, level 2, and level 3. The first level of the MOS model was a first order implementation of the currents (see above analysis of currents). This simplified model was rarely used, whereas level two and level three are improved levels of level one and are used widely. Level two and three contain about 15 DC parameters each and can be used up to the 1 μm channel length of transistor technology. Their characteristics are:

- For varying length the drain current could be fit, with reasonable accuracy (about 5% RMS error).
- Little advanced fitting capability for analog application
- One parameter for fitting the subthreshold region (weak inversion region)
- They can't vary the mobility degradation with back-bias, so the fits to I_{ds} in the saturation region at high back-bias are not very good
- Problems occur in interpolating over device geometry
- For narrow devices they perform poorly

The difference between model two and three are not big, even it is said that level two is physically based whereas level three is more semiempirical based. One could prefer level three because of its speed and its smoother transition region approximations.

C.3.1 Berkeley Short-Channel Igfet Models (BSIM's)

C.3.1.1 BSIM model

Many disadvantages of level three and two are reconsidered in the BSIM model. Dependencies, like the geometry of the device, are incorporated into the equations to get

better results. Hence most equations are built from scratch to include missing or needed phenomena's not existing in the previous levels. On the other hand the rewriting concept had also to be simple, for example in the BSIM and BSIM2 model each parameter (except for a very few) is written as a sum of three terms

$$Parameter = Par_0 + \frac{Par_L}{L_{eff}} + \frac{Par_W}{W_{eff}} \quad (C.28)$$

where Par_0 accounts for the zero-order, the second term is length dependence and the third term is the width dependence. L_{eff} and W_{eff} are the effective channel width and length. From (C.28) one sees that more parameters are produced then in the level two and three, hence BSIM has about 54 DC parameters. The advantages of the BSIM model over the previous model levels is better fitting

- for submicron channel lengths
- over a wider range of geometries
- in the subthreshold region
- for nonzero back-bias

It does have however some drawbacks: It's unable to fit over a large geometry variation, and it still isn't very useful for analog application. To justify that statement we only must look at the higher neglected terms of Taylor expansion around $1/L_{eff}$ and $1/W_{eff}$. In order to fit better for varying geometries these higher terms must be included, which are not in the BSIM model. Another drawback is the lack of fitting G_{ds} , and as mentioned it's doubtful if this is usable in the analog domain. Last but not least no default parameter values are set if left empty or undefined; this can cause weird behavioral of the device. This doesn't stop BSIM to be superior to the previous foregoing models.

The BSIM model contains parameters in the following categories:

- Operating Point Parameters (*shown in table C.2*)
- Model Description
- Length/Width Sensitivity Parameters
- DC Model Equations
- AC Model Equations
- Temperature Equations (*user defined only!*)

Parameter Name		Parameter Description	Unit
1	ID	Drain current	A
2	VGS	Gate-source voltage	V
3	VDS	Drain-source voltage	V
4	VBS	Body-source voltage	V
5	VTH	Threshold voltage	V
6	VDSAT	Drain-source saturation voltage	V
7	GM	Transconductance (dIds/dVgs)	mho

8	GDS	Transconductance (dI_{ds}/dV_{ds})	mho
9	GMBS	Transconductance (dI_{ds}/dV_{bs})	mho
10	CBD	Body-drain capacitance	F
11	CBS	Body-source capacitance	F
12	CGSOV	Gate-source overlap capacitance	F
13	CGDOV	Gate-drain overlap capacitance	F
14	CGBOV	Gate-bulk overlap capacitance	F
15	QG	Total gate charge	C
16	QB	Total bulk charge	C
17	QD	Total drain charge	C
18	QS	Total source charge	C
19	CGGB	Total capacitance reflected on the gate	F
20	CBGB	Bulk-gate capacitance	F
21	CDGB	Drain-gate capacitance	F
22	CSGB	Source-gate capacitance	F
23	CGBB	Gate-bulk capacitance	F
24	CBBB	Total capacitance reflected on the bulk	F
25	CDBB	Drain-bulk capacitance	F
26	CSBB	Source-bulk capacitance	F
27	CGDB	Gate-drain capacitance	F
28	CBDB	Bulk-drain capacitance	F
29	CDDB	Total capacitance reflected on the drain	F
30	CSDB	Source-drain capacitance	F
31	CGSB	Gate-source capacitance	F
32	CBSB	Bulk-source capacitance	F
33	CDSB	Drain-source capacitance	F

34	CSSB	Total capacitance reflected on the source	F
35	PWR	Power	W
36	LIN	Linear region flag	s
36	BKDWN	Breakdown flag	s

Table C.2 Basic parameters of BSIM model. Data from
<http://www.ece.uci.edu/eceware/cadence/cspiceref/chap8.html#1040320>

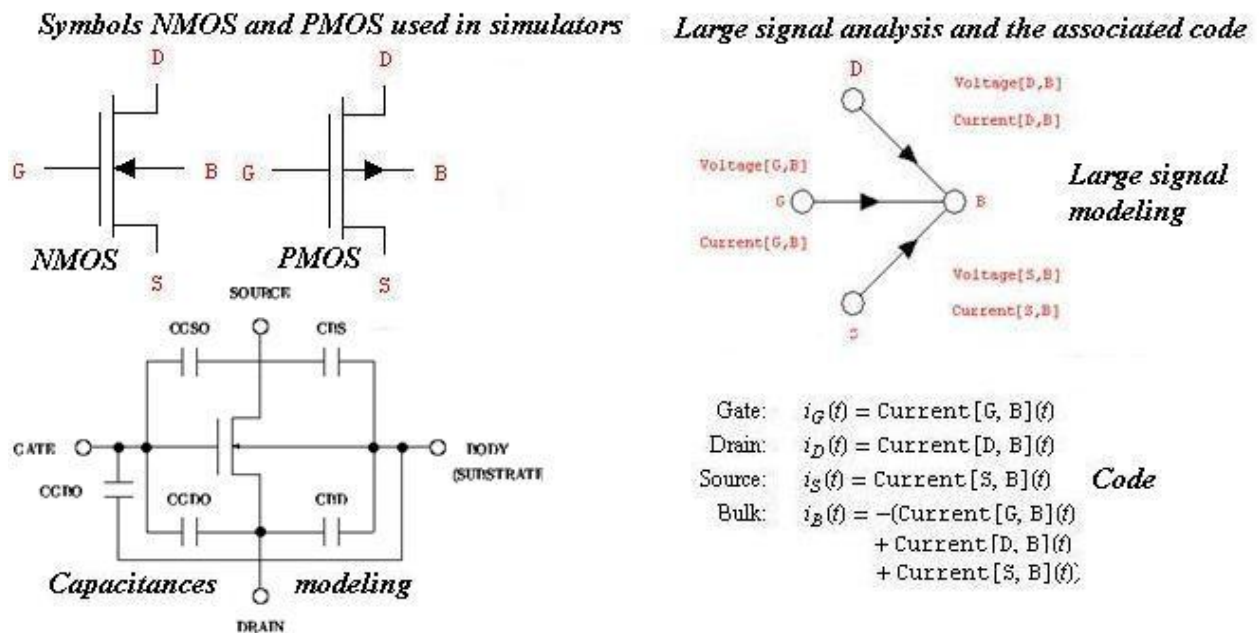


Figure C.8. Example of symbols used in a simulator and an example of representing parameters

C.3.1.2 BSIM2

The BSIM2 version was actually meant as an improvement version of the BSIM version, since at beginning a lot of “bugs” were fixed. These “bugs” were

- removing unnecessary parameters, that had no influence on the overall outcome
- wrong varying currents which were a function of certain parameters
- adding unphysical fitting parameters, and parameters to allow fitting of Gds

Having done all the fixing, the parameter number rose up to 100, subthreshold fitting became more accurate, and the decision was made to make the fixed BSIM the BSIM2 version. However problems like the fitting of large geometry variations and other drawbacks of BSIM remained.

C.3.1.3 BSIM3

Quickly UC Berkeley became aware of the shortcomings of the BSIM2 model and announced the upcoming of the BSIM3 model, which will cover the errors. Even though the name implies a continuing based on the previous models, it became clear that another rewriting was necessary. For that reason the BSIM3 model is completely different from its predecessors. It overthrows the width dependence and parameters for G_{ds} , introduced in BSIM and BSIM2, and adopt instead, the geometry dependence, which directly are referred in the model equations, as in (Spice) level two and three. In addition, BSIM3 is a more physically based model, with about 30 fitting parameters, and more parameters than ever! An outstanding future is the manipulation of parameters without affecting the fitting parameters. With all this features it is a strong candidate for analog applications. The poor fitting over a wide range of geometries still lacks, but future modifications will address this problem. Therefore subversions of the BSIM3 emerged. One the subversions BSIM3v3 has been chosen as the first industry standard model for circuit simulation and is supported by EIA Compact Model Council (CMC), a consortium of 20 companies including *IBM, Intel, TI, Motorola, Lucent, AMD, Hitachi, Philips, Infineon, TSMC, Cadence, Avanti, etc.*

Major Physical Mechanisms of BSIM3v3 are

- Short/Narrow Channel Effects on Threshold Voltage
- Non-Uniform Doping Effects
- Mobility Reduction Due to Vertical Field
- Bulk Charge Effect
- Carrier Velocity Saturation
- Drain Induced Barrier Lowering (DIBL)
- Channel Length Modulation (CLM)
- Substrate Current Induced Body Effect (SCBE)
- Parasitic Resistance Effects
- Quantum Mechanic Charge Thickness Model
- Unified Flicker Noise Model

Most important parameters of the BSIM3 version 3 are displayed next:

Name	Description	Units	Default
NMOS	N-channel type model		yes
PMOS	P-channel type model		no
Idsmod	Ids model		8
Version	model version		3.22
Mobmod	mobility model selector		1
Capmod	capacitance model selector		1
Noimod	noise model selector		1

Paramchk	model parameter checking selector		0
Binunit	bin unit selector		1
Rg	gate resistance	ohms	0
Rsh	drain and source diffusion sheet resistance	ohms/sq	0.0
Nj	bulk P-N emission coefficient		1.0
Xti	junction current temp. exponent		3.0
Js	gate saturation current	A/m2	10 ⁻⁴
Jsw	sidewall junction reverse saturation current	A/m2	0.0
Lint	length offset fitting parameter (binning parameter; see Note)	m	0.0
Ll	coefficient of length dependence for length offset	mLln	0.0
Lln	power of length dependence of length offset		1.0
Lw	coefficient of width dependence for length offset	mLwn	0.0
Name	Description	Units	Default
Lwn	power of width dependence of length offset		1.0
Lwl	coefficient of length and width cross term for length offset	m(Lwn+Lln)	0.0
Wint	width offset fitting parameter (binning parameter; see Note)	m	0.0
Wl	coefficient of length dependence for width offset	mWln	0.0
Wln	power of length dependence of width offset		1.0
Ww	coefficient of width dependence for width offset	mWwn	0.0
Wwn	power of width dependence of width offset		1.0
Wwl	coefficient of length and width cross term for width offset	m(Wwn+Wln)	0.0

Tnom	parameter measurement temp.	°C	25
Trise	temperature rise above ambient	°C	0
Tox	oxide thickness	m	1.5×10^{-8}
Cj	zero-bias bulk junction bottom capacitance	F/m ²	5.0×10^{-4}
Mj	bulk junction bottom grading coefficient		0.5
Cjsw	zero-bias bulk junction sidewall capacitance	F/m	5.0×10^{-10}
Mjsw	bulk junction sidewall grading coefficient		0.33
Pb	bulk junction potential	V	1.0
Pbsw	sidewall junction potential	V	1.0
Xt	doping depth	m	1.55×10^{-7}
Vbm	maximum applied body bias	V	-5.0
Vbx	V _{th} transition body voltage	V	calculated parameter
Xj	metallurgical junction depth	m	1.5×10^{-7}
Dwg	coefficient of Weff's gate dependence (binning parameter; see Note)	m/V	0.0
Dwb	coefficient of Weff's body dependence (binning parameter; see Note)	m/V(1/2)	0.0
Nch	channel doping concentration	1/cm ³	1.7×10^{17}
Name	Description	Units	Default
Nsub	substrate doping concentration	1/cm ³	6.0×10^{16}
Ngate	poly-gate doping concentration	1/cm ³	depends
Gamma1	body effect coefficient near interface	V(1/2)	depends
Gamma2	body effect coefficient in the bulk	V(1/2)	depends
Alpha0	1st parameter of impact ionization current (binning parameter; see Note)	m/V	0.0
Beta0	2nd parameter of impact ionization current (binning parameter; see Note)	V	30.0

Vth0	zero-bias threshold voltage (binning parameter; see Note)	V	depends
K1	first order body effect coefficient (binning parameter; see Note)	$V(1/2)$	depends
K2	second order body effect coefficient (binning parameter; see Note)		depends
K3	narrow width effect coefficient (binning parameter; see Note)		80.0
K3b	body effect coefficient of K3 (binning parameter; see Note)	$1/V$	0.0
W0	narrow width effect W offset (binning parameter; see Note)	m	2.5×10^{-6}
Nlx	lateral non-uniform doping effect (binning parameter; see Note)	m	1.74×10^{-7}
Dvt0	short channel effect coefficient 0 (binning parameter; see Note)		2.2
Dvt1	short channel effect coefficient 1 (binning parameter; see Note)		0.53
Dvt2	short channel effect coefficient 2 (binning parameter; see Note)	$1/V$	-0.032
Dvt0w	narrow width effect coefficient 0 (binning parameter; see Note)	$1/m$	0.0
Dvt1w	narrow width effect coefficient 1 (binning parameter; see Note)	$1/m$	5.3×10^6
Dvt2w	narrow width effect coefficient 2 (binning parameter; see Note)	$1/V$	-0.032
Name	Description	Units	Default
Cgso	gate-source overlap capacitance, per channel width	F/m	depends
Cgdo	gate-drain overlap capacitance, per channel width	F/m	depends
Cgbo	gate-bulk overlap capacitance, per channel length	F/m	0.0

Xpart	flag for channel charge partition		0.0
DROUT	DIBL effect on Rout coefficient binning parameter; see Note)		0.56
Dsub	DIBL effect coefficient in subthreshold region binning parameter; see Note)		(fixed by DROUT)
Ua	linear Vgs dependence of mobility (binning parameter; see Note)	m/V	2.25×10^{-9}
Ua1	temperature coefficient of Ua	m/V	4.31×10^{-9}
Ub	quadratic Vgs dependence of mobility (binning parameter; see Note)	(m/V) ²	5.87×10^{-19}
Ub1	temperature coefficient of Ub	(m/V) ²	-7.61×10^{-18}
Uc	body-bias dependence of mobility (binning parameter; see Note)	m/V ² 1/V	-4.65×10 ⁻¹¹ Mobmod=1, 2 -0.0465 Mobmod=3
Uc1	temperature coefficient of Uc	m/V ² 1/V	-5.6×10 ⁻¹¹ Mobmod=1,2 -0.056 Mobmod=3
U0	low-field mobility at T=Tnom (binning parameter; see Note)	cm ² /Vs	670.0 NMOS 250.0 PMOS
Ute	temperature coefficient of mobility		-1.5
Rdsw	source drain resistance per width (binning parameter; see Note)	ohms × μmWr	0.0
Prwg	gate bias effect coefficient of Rdsw (binning parameter; see Note)	1/V	0.0
Prwb	body effect coefficient of Rdsw (binning parameter; see Note)	1/V	0.0
Wr	width dependence of Rds (binning parameter; see Note)		1.0
Prt	temperature coefficient of Rdsw	ohms × μm	0.0
Name	Description	Units	Default
Vsat	saturation velocity at T=Tnom (binning parameter; see Note)	m/s	8.0×10^4
At	temperature coefficient of Vsat	m/s	3.3×10^4

A0	bulk charge effect coefficient for channel length (binning parameter; see Note)		1.0
Keta	body-bias coefficient of bulk charge (binning parameter; see Note)	1/V	-0.047
Ags	gate bias coefficient of Abulk (binning parameter; see Note)	1/V	0.0
A1	first non-saturation factor for PMOS (binning parameter; see Note)	1/V	0.0
A2	second non-saturation factor for PMOS (binning parameter; see Note)		1.0
B0	bulk charge effect coefficient for channel width (binning parameter; see Note)	m	0.0
B1	bulk charge effect width offset (binning parameter; see Note)	m	0.0
Voff	threshold voltage offset (binning parameter; see Note)	V	-0.08
Nfactor	subthreshold swing factor (binning parameter; see Note)		1.0
Cdsc	D/S and channel coupling capacitance (binning parameter; see Note)	F/m ²	2.4×10^{-4}
Cdscb	body-bias dependence of Cdsc (binning parameter; see Note)	F/V/m ²	0.0
Cdscl	drain-bias dependence of Cdsc (binning parameter; see Note)	F/V/m ²	0.0
Cit	interface state capacitance (binning parameter; see Note)	F/m ²	0.0
Eta0	subthreshold region DIBL coefficient (binning parameter; see Note)		0.08
Etab	body-bias coefficient for DIBL effect (binning parameter; see Note)	1/V	-0.07
Pclm	channel-length modulation coefficient (binning parameter; see Note)		1.3
Pdiblcl	first Rout DIBL effect coefficient		0.39

Name	Description	Units	Default
Pdiblc2	second Rout DIBL effect coefficient		0.0086
Pdiblc b	body effect coefficient of DIBL correction parameters	1/V	0
Pscbe1	first substrate current body effect	V/m	4.24×10^8
Pscbe2	second substrate current body effect	m/V	10–5
Pvag	Vg dependence of Rout coefficient (binning parameter; see Note)		0.0
Delta	effective Vds parameter (binning parameter; see Note)	V	0.01
Kt1	temperature coefficient of Vth	V	–0.11
Kt1l	channel length sensitivity of Kt1	V×m	0.0
Kt2	body bias coefficient of Kt1		0.022
Cgsl	light doped source-gate region overlap capacitance	F/m	0.0
Cgdl	light doped drain-gate region overlap capacitance	F/m	0.0
Ckappa	coefficient for lightly doped region overlap capacitance	F/m	0.6
Cf	fringing field capacitance	F/m	
Clc	constant term for short channel model	m	0.1×10^{-6}
Cle	exponential term for short channel		0.6
Dlc	length offset fitting parameter from C-V	m	Lint
Dwc	width offset fitting parameter from C-V	m	Wint
Nlev	Noise model level		-1
Gdwnoi	Drain noise parameters for Nlev=3		1
Kf	flicker (1/f) noise coefficient		0.0
Af	flicker (1/f) noise exponent		1.0
Ef	flicker (1/f) noise frequency exponent		1.0
Em	flicker (1/f) noise parameter	V/m	4.1×10^7

Noia	noise parameter A		1.0×1020 NMOS 9.9×1018 PMOS
Name	Description	Units	Default
Noib	noise parameter B		5.0×104 NMOS 2.4×103 PMOS
Noic	noise parameter C		$-1.4 \times 10-12$ NMOS 1.4×1012 PMOS
Imax	explosion current	A	10.0
wVsubfwd	substrate junction forward bias (warning)	V	infinite
wBvsub	substrate junction reverse breakdown voltage (warning)	V	infinite
wBvg	gate oxide breakdown voltage (warning)	V	infinite
wBvds	drain-source breakdown voltage (warning)	V	infinite
wIdsmax	maximum drain-source current (warning)	A	infinite
Toxm	gate oxide thickness tox value at which parameters are extracted	m	
Vfb	DC flat-band voltage	V	depends
Noff	CV parameter in VgsteffCV for weak-to-strong inversion region		1.0
Voffcv	CV parameter in VgsteffCV for weak-to-strong inversion region		1.0
Ijth	diode limiting current	A	depends
Alpha1	substrate current parameter	1/V	0.0
Acde	exponential coefficient for charge thickness in the accumulation and depletion regions (binning parameter; see Note)	m/V	1.0
Moin	coefficient for the gate-bias dependent surface potential (binning parameter; see Note)	V(1/2)	15.0
Tpb	temperature coefficient of pb	V/K	0.0

Tpbsw	temperature coefficient of pbsw	V/K	0.0
Tpbswg	temperature coefficient of pbswg	V/K	0.0
Tcj	temperature coefficient of cj	1/K	0.0
Tcjsw	temperature coefficient of cjsw	1/K	0.0
Tcjswg	temperature coefficient of cjswg	1/K	0.0
Name	Description	Units	Default
Llc	coefficient of length dependence for CV channel length offset	mLln	DC Ll
Lwc	coefficient of width dependence for CV channel length offset	mLwn	DC Lw
Lwlc	coefficient of length and width cross-term for CV channel length offset	mLwn + LLn	DC Lwl
Wlc	coefficient of length dependence for CV channel width offset	mWln	DC Wl
Wwc	coefficient of width dependence for CV channel width offset	mWwn	DC Ww
Wwlc	coefficient of length and width cross-term for CV channel width offset	mWln + Wwn	DC Wwl
wPmax	maximum power dissipation (warning)	W	infinite
Acm	area calculation method		-1
Calcacm	flag to use Acm when Acm=12		0
Hdif	length of heavily doped diffusion (ACM=2,3 only)	m	0
Ldif	length of lightly doped diffusion adjacent to gate (ACM=1,2)	m	0
Wmlt	width diffusion layer shrink reduction factor		1
Xw	accounts for masking and etching effects	m	0
Xl	accounts for masking and etching effects	m	0
Rdc	additional drain resistance due to contact resistance	Ohms	0

Rsc	additional source resistance due to contact resistance	Ohms	0
Vfbcv	flat-band voltage parameter for capmod=0 only	F/m	-1.0
B3qmod	BSIM3 charge model (0 for Berkeley, 1 for Hspice Capmod = 0)		0
Cjswg	S/D (gate side) sidewall junction capacitance	F/m	Cjsw
Pbswg	S/D (gate side) sidewall junction built in potential	V	Mjsw
Mjswg	S/D (gate side) sidewall junction grading coefficient		Pbsw
Name	Description	Units	Default
Is	bulk junction saturation current	A	1e-14
Nqsmode	non-quasi-static model selector		0
Elm	non-quasi-static Elmore constant parameter		5.0
Rd	drain resistance	Ohms	0
Rs	source resistance	Ohms	0
Flkmod	flicker noise model selector		0
Tlev	temperature equation selector (0/1/2/3)		0
Tlevc	temperature equation selector for capacitance (0/1/2/3)		0
Eg	band gap	eV	1.16
Gap1	energy gap temperature coefficient alpha	V/oC	7.02e-4
Gap2	energy gap temperature coefficient beta	K	1108
Cta	Cj linear temperature coefficient	1/oC	0
Ctp	Cjsw linear temperature coefficient	1/oC	0
Pta	Vj linear temperature coefficient	1/oC	0

Ptp	Vjsw linear temperature coefficient	1/oC	0
Trd	Rd linear temperature coefficient	1/oC	0
Trs	Rs linear temperature coefficient	1/oC	0
Wmin	binning minimum width (not used for binning; use BinModel)	m	0
Wmax	binning maximum width (not used for binning; use BinModel)	m	1
Lmin	binning minimum length (not used for binning; use BinModel)	m	0
Lmax	binning maximum length (not used for binning; use BinModel)	m	1
AllParams	DataAccessComponent-based parameters		

C.3.1.4 BSIM4

Another version of the BSIM was needed in the submicron-technology addressing new and accuracy important phenomena's. The main modeling subjects in BSIM4 are listed below

- Basic IV model overview
 - *Vth* model for pocket/retrograde technologies
 - *Vgsteff*
 - Bulk charge (*Abulk*) model
 - Mobility models
 - *Rout* model
- GIDL current model
- Bias-dependent *Rds(V)* model, internal or external
- Gate (equivalent) *Tox* and dielectric constant, and quantum mechanical charge-layer thickness model
- RF and High-speed model
 - Intrinsic input resistance (*Rii*) model
 - Non-Quasi-Static (NQS) model
 - Holistic and noise-partition thermal noise model
 - Substrate resistance network
 - Flicker noise model
- Geometry calculation (Layout-dependent parasitics) model
- Asymmetrical source/drain junction diode model
- I-V and breakdown model
- Gate dielectric tunneling current model

C.3.1.4.1 Basic parameter list of the BSIM4.2.1model

Parameter name	Description	Default value	*Binnable?	Note
VTH0 or VTHO	Long-channel threshold voltage at $V_{bs}=0$	0.7V (NMOS) -0.7V (PMOS)	Yes	Note-4
VFB	Flat-band voltage	-1.0V	Yes	Note-4
PHIN	Non-uniform vertical doping effect on surfacepotential	0.0V	Yes	-
K1	First-order body bias coefficient	$0.5V^{1/2}$	Yes	Note-5
K2	Second-order body bias coefficient	0.0	Yes	Note-5
K3	Narrow width coefficient	80.0	Yes	-
K3B	Body effect coefficient of K3	0.0 V ⁻¹	Yes	-
W0	Narrow width parameter	2.5e-6m	Yes	-
LPE0	Lateral non-uniform doping parameter at $V_{bs}=0$	1.74e-7m	Yes	-
LPEB	Lateral non-uniform doping effect on K1	0.0m	Yes	-
VBM	Maximum applied body bias in VTH0 calculation	-3.0V	Yes	-
DVT0	First coefficient of short-channel effect on V_{th}	2.2	Yes	-
DVT1	Second coefficient of short-channel effect on V_{th}	0.53	Yes	-
DVT2	Body-bias coefficient of short-channel effect on V_{th}	-0.032V ⁻¹	Yes	-
DVTP0	First coefficient of drain-induced V_{th} shift due to for long-channel pocket devices	0.0m	Yes	Not modeled if binned DVTP0 ≤ 0.0
DVTP1	First coefficient of drain-induced V_{th} shift due to for long-channel pocket devices	0.0V ⁻¹	Yes	-

Parameter name	Description	Default value	*Binnable?	Note
DVT0W	First coefficient of narrow width effect on V_{th} for small channel length	0.0	Yes	-
DVT1W	Second coefficient of narrow width effect on V_{th} for small channel length	5.3e6m ⁻¹	Yes	-
DVT2W	Body-bias coefficient of narrow width effect for small channel length	-0.032V ⁻¹	Yes	-
U0	Low-field mobility	0.067 m ² /(Vs) (NMOS); 0.025 m ² /(Vs) PMOS	Yes	-
UA	Coefficient of first-order mobility degradation due to vertical field	1.0e-9m/V for MOBMOD =0 and 1; 1.0e-15m/V for MOBMOD =2	Yes	-
UB	Coefficient of secon-order mobility degradation due to vertical field	1.0e-19m ² / V ²	Yes	-
UC	Coefficient of mobility degradation due to body-bias effect	-0.0465V ⁻¹ for MOB- MOD=1; - 0.0465e-9 m/V ² for MOBMOD =0 and 2	Yes	-
EU	Exponent for mobility degradation of MOBMOD=2	1.67 (NMOS); 1.0 (PMOS)		-
VSAT	Saturation velocity	8.0e4m/s	Yes	-

Parameter name	Description	Default value	*Binnable?	Note
CDSC	coupling capacitance between source/ drain and channel	2.4e-4F/m2	Yes	-
CDSCB	Body-bias sensitivity of Cdsc	0.0F/(Vm2)	Yes	-
CDSCD	Drain-bias sensitivity of CDSC	0.0(F/Vm2)	Yes	-
PCLM	Channel length modulation parameter	1.3	Yes	-
PDIBLC1	Parameter for DIBL effect on Rout	0.39	Yes	-
PDIBLC2	Parameter for DIBL effect on Rout	0.0086	Yes	-
PDIBLCB	Body bias coefficient of DIBL effect on Rout	0.0V-1	Yes	-
DROUT	Channel-length dependence of DIBL effect on Rout	0.56	Yes	-
PSCBE1	First substrate current induced body- effect parameter	4.24e8V/m	Yes	-
PSCBE2	Second substrate current induced body- effect parameter	1.0e-5m/V	Yes	-
PVAG	Gate-bias dependence of Early voltage	0.0	Yes	-
DELTA (δ in equation)	Parameter for DC V_{dseff}	0.01V	Yes	-
FPROUT	Effect of pocket implant on Rout deg- radation	0.0V/m0.5	Yes	Not mod- eled if binned FPROUT not positive
PDITS	Impact of drain-induced V_{th} shift on Rout	0.0V-1	Yes	Not mod- eled if binned PDITS=0; Fatal error if binned PDITS negative

Parameter name	Description	Default value	*Binnable?	Note
A0	Coefficient of channel-length dependence of bulk charge effect	1.0	Yes	-
AGS	Coefficient of V_{gs} dependence of bulk charge effect	0.0V-1	Yes	-
B0	Bulk charge effect coefficient for channel width	0.0m	Yes	-
B1	Bulk charge effect width offset	0.0m	Yes	-
KETA	Body-bias coefficient of bulk charge effect	-0.047V-1	Yes	-
A1	First non-saturation effect parameter	0.0V-1	Yes	-
A2	Second non-saturation factor	1.0	Yes	-
WINT	Channel-width offset parameter	0.0m	No	-
LINT	Channel-length offset parameter	0.0m	No	-
DWG	Coefficient of gate bias dependence of W_{eff}	0.0m/V	Yes	-
DWB	Coefficient of body bias dependence of W_{eff} bias dependence	0.0m/V ^{1/2}	Yes	-
VOFF	Offset voltage in subthreshold region for large W and L	-0.08V	Yes	-
VOFFL	Channel-length dependence of VOFF	0.0mV	No	-
MINV	V_{gsteff} fitting parameter for moderate inversion condition	0.0	Yes	-
NFACTOR	Subthreshold swing factor	1.0	Yes	-
ETA0	DIBL coefficient in subthreshold region	0.08	Yes	-
ETAB	Body-bias coefficient for the sub-threshold DIBL effect	-0.07V-1	Yes	-
DSUB	DIBL coefficient exponent in sub-threshold region	DROUT	Yes	-
CIT	Interface trap capacitance	0.0F/m ²	Yes	-

Parameter name	Description	Default value	*Binnable?	Note
PDITSL	Channel-length dependence of drain-induced V_{th} shift for Rout	0.0m-1	No	Fatal error if PDITSL negative
PDITSD	V_{ds} dependence of drain-induced V_{th} shift for Rout	0.0V-1	Yes	-

*Binning means piece-wise geometry analysis. Having an infinite number of bins means to have a non-scalable model or unphysical device model. Therefore saying that a transistor is “binnable” it is meant that the parameters are extracted by fitting electrical data at fixed geometry. For a “non-binnable” model parameters are extracted by fitting electrical data over geometry at fixed bias.

The latest BSIM models and documentations can be found at website <http://www-device.eecs.berkeley.edu> .

C.4 References

1. **Tsividis, Yannis.** *The MOS transistor 2nd Ed.* New York Oxford : Oxford University Press, 1999.
2. **Hu, Chenming.** BSIM4 MOSFET Model for Circuit Simulation. Berkeley : Department of EECS University Of California Berkeley.