

EE105 Reader

Ali M. Niknejad

Copyright ©2022 by Ali Niknejad

PUBLISHED BY ALI NIKNEJAD

<http://rfic.eecs.berkeley.edu/~niknejad/index.html>

EDITED BY TARIK FAWAL

<http://www.tarikfawal.com>

Licensed under the Creative Commons Attribution-NonCommercial 3.0 Unported License (the “License”). You may not use this file except in compliance with the License. You may obtain a copy of the License at <http://creativecommons.org/licenses/by-nc/3.0>. Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an “AS IS” BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

First printing, January 17, 2020

Latest printing, August 21, 2022

DOWNLOAD THE LATEST VERSION AT THE LINK BELOW:

<https://tfawal.github.io/files/edance.pdf>

Foreword

I took this course with Professor Niknejad in the Fall of 2021. I had transferred from the City College of San Francisco, and just spent my first year at Cal online during the COVID-19 pandemic. I did not adapt well to the online format, and my first year was rough. Having gone through the EECS 16A/B series during that year, now EE105 was my first upper division course. I was very excited for it, but also completely unprepared for the challenge that it would present.

The materials in this course are dense, and they comes at you fast if you don't already have a lot of experience with circuits. I failed my mid-term exam miserably. I was a bit shell-shocked from the experience, and I stopped going to class. I decided to either drop the course, or just fail the rest of it on purpose so I could retake it.

Then one day I got an email from Professor Niknejad. He was checking in to see if I was alright, and asked about why I hadn't shown up in a few weeks. This was the first time that a Cal professor had reached out to me—they are usually very busy and hard to get a hold of. Over that last year I have gotten to know Ali a bit more, and see how busy he really is. As I write this now, his gesture means even more to me. I was so stressed out when he contacted me, that I wrote a brutally honest reply. I told him my plan to drop the class, and went on a rant about the course was not working for me. I complained that everything went too fast, and this book needed some editing to make it more accessible to people like me. I had not been to school for fifteen years before returning in 2017, and I always felt like I was playing catch up with all the brilliant students at Cal.

I regretted my email only a few minutes after sending it. I figured Ali would be offended by it, but to my surprise he wrote a very in-depth reply. He put himself on my level by telling my how he struggled with the material as a student, and offered words of encouragement to continue; affirming that I would not fail the course. I decided to stick with the class and work very hard, and I wound up doing much better on the final.

About a month after the class ended, I began revisiting the book. The material in it is actually really great, and I found it much easier to understand the second time around. This gave me an idea for a project. I wanted to edit the book in order to make it more accessible for future students that might have a similar experience in struggling to grasp the material, like I did the first time around. Fortunately, Professor Niknejad agreed to let me do this.

I hope that the work I put in to this book benefits everyone, but in particular I am hoping to help people out that are in a similar situation that I was. The bulk of revisions were dedicated to; creating an in-depth index, re-organizing the figures so that they are on the same (or next) page when they are referenced; creating a glossary of terms for a quick reference, and adding in additional mathematical steps to make certain derivations more clear.

Any important terms that are highlighted in **bold** can be found in the index, and other words that are *italicized* can usually be found in the glossary of terms. The glossary itself contains both definitions of words and concepts, and their accompanying equations. The glossary is meant to summarize difficult concepts into a few sentences, using ordinary language.

If you are a bit rusty on circuit analysis, then for a review it is *highly* recommended to read Professor Niknejad's EECS 16A reader. In particular, you should review all circuit theorems, and make sure that you are very familiar with Thévenin and Norton equivalence. You can download it at <http://rfic.eecs.berkeley.edu/niknejad/edogs.html>.

I would like to thank Professor Niknejad for his help in my academic career. For those taking his class, take advantage of being able to learn circuits from a master of the field. If you are struggling, remember my story and don't give up. Never give up. Good luck with the course, and I hope you find the work put into this material useful in your success.

Tarik Fawal

August 18, 2022



Contents

1	Linear Time-Invariant Systems	19
1.1	Chapter Preview	19
1.2	Example: Analyzing an RLC Circuit	20
1.2.1	Example: Low Pass Filter (LPF)	20
1.2.2	LPF the "hard way"	20
1.3	Linear Time-Invariant Systems	24
1.3.1	Linear Time-Invariant Definition	24
1.4	The Complex Exponential Technique	25
1.4.1	Why introduce complex numbers?	25
1.4.2	Complex Exponential	25
1.4.3	Euler's Relations and The Circle	26
1.4.4	The Magic of Sinusoids	26
1.5	The Low Pass Filter Again	28
1.5.1	LPF Example: The "soft way"	28
1.6	Generalization to any Linear Circuit	30
1.6.1	"Proof" for Linear Systems	30
1.6.2	General Complex Exponential	30
1.7	Time Domain Characterization of Linear Systems	31
1.7.1	Unit Area Rectangular Function	31
1.8	Back to Time Domain: Impulse Response	34
1.8.1	Convolution Operation	34
1.9	Step and Pulse Response	36
1.9.1	Step Response	37
1.9.2	Sinusoidal Response	38
1.9.3	Pulse Response	38

1.10 Frequency Domain Characterization of Linear Systems	38
1.10.1 Relation to Complex Exponential	39
1.10.2 De-convolving the Convolution	39
1.10.3 Fourier Series and Transform	42
1.10.4 Frequency Domain Interpretation	42
1.10.5 But Most Systems are Non-Linear ...?	44
1.11 Chapter Summary	45
2 AC Circuits and AC Analysis	47
2.1 Chapter Preview	47
2.2 Transfer Functions	48
2.2.1 Transfer Function Concept	48
2.2.2 Voltage and Current Gain	50
2.2.3 Trans-Impedance and Trans-Admittance	51
2.2.4 Impede the Currents!	51
2.2.5 Admit the Currents!	51
2.3 Transfer Function Poles/Zeros	52
2.3.1 Complex Transfer Function	52
2.3.2 Poles and Zeros	52
2.4 AC Circuits Review	53
2.4.1 Phasors	53
2.4.2 Capacitor I-V Phasor Relation	54
2.4.3 Inductor I-V Phasor Relation	55
2.4.4 Direct Calculation of H (no DEs)	55
2.5 Example AC Circuit Problems	56
2.5.1 LPF Example: Again!	56
2.5.2 Bigger Example (no problem!)	56
2.5.3 Series RLC Circuits	57
2.6 Bode Plots	60
2.6.1 Finding the Magnitude and Phase (Quickly!)	60
2.6.2 Bode Plot	61
2.7 Power Flow in AC Circuits	64
2.7.1 Power Flow with Phasors	65
2.7.2 More Power to You!	65
2.7.3 Understanding AC Power Flow	66
2.8 Chapter Summary	67
3 Introduction to Semiconductors	69
3.1 Chapter Preview	69
3.2 Conduction in an Ideal Metal "Gas"	70
3.2.1 Ohm's Law	70
3.2.2 Ohm's Law Revisited	70
3.2.3 Conductivity of a Gas	70
3.2.4 Mobility	72
3.2.5 Conduction in Metals	72
3.2.6 Wave Nature of Electron	73
3.2.7 Scattering in Metals	73

3.2.8	Summary of Conduction	73
3.3	Introduction to Semiconductors	74
3.3.1	Resistivity for a Few Materials	74
3.3.2	Electronic Properties of Silicon	75
3.3.3	States of an Atom	77
3.3.4	Energy Band Diagram	77
3.3.5	Model for Good Conductor	79
3.3.6	Bond Model for Silicon ($T = 0\text{ K}$)	79
3.3.7	Bond Model for Silicon ($T > 0\text{ K}$)	79
3.3.8	Holes?	80
3.3.9	Yes, Holes!	81
3.3.10	More About Holes	82
3.4	Intrinsic Carrier Concentration	82
3.4.1	Thermal Equilibrium (Pure Si)	82
3.4.2	Generation Statistics	83
3.5	Doping with Impurities	83
3.5.1	Doping with Group V Elements	83
3.5.2	Donor Accounting	83
3.5.3	Doping with Group III Elements	85
3.5.4	Mass Action-Law (Again)	85
3.5.5	Compensation	86
3.6	Drift Currents	87
3.6.1	Thermal Equilibrium	87
3.6.2	Drift Velocity and Mobility	87
3.6.3	Mobility vs. Doping in Silicon at 300°K	88
3.7	Diffusion Currents	89
3.7.1	Diffusion	89
3.7.2	Diffusion Equations	90
3.7.3	Einstein Relation	91
3.7.4	Total Current and Boundary Conditions	91
4	IC Resistors and Capacitors and Electrostatics	93
4.1	Chapter Preview	93
4.2	IC Fabrication: Si Substrate	94
4.2.1	IC Fabrication: Oxide	94
4.2.2	IC Fabrication: Ion Implantation	94
4.3	IC Resistors	95
4.3.1	“Diffusion” Resistor	95
4.3.2	Poly Film Resistor	95
4.3.3	Ohm’s Law	96
4.3.4	Sheet Resistance (R_{\square})	96
4.3.5	Using Sheet Resistance (R_{\square})	97
4.4	Review of Electrostatics	98
4.4.1	Electrostatics Review	98
4.4.2	Electrostatics in 1D	98
4.4.3	Electrostatic Potential	99
4.4.4	Boundary Conditions	100

4.5 IC Capacitors	101
4.5.1 MIM (Metal-Insulator-Metal) Capacitor	101
4.5.2 Review of Capacitors	101
4.5.3 A Non-Linear Capacitor	102
4.5.4 Small Signal Capacitance	103
4.5.5 Example of Non-Linear Capacitor	103
5 PN Junctions in Equilibrium	105
5.1 Chapter Preview	105
5.2 Structure	106
5.3 Carrier Concentration and Potential	106
5.4 PN-Junction in Equilibrium	108
5.4.1 PN-Junction: Overview	108
5.4.2 PN-Junction Currents	110
5.4.3 PN-Junction Fields	111
5.4.4 "Exact" Equation for Fields	111
5.4.5 Depletion Approximation	112
5.4.6 Total Depletion Width	115
5.4.7 Contact Potential	116
5.4.8 PN-Junction Capacitor	117
5.5 Reverse Biased PN Junction	117
5.5.1 Voltage Dependence of Depletion Width	117
5.5.2 Charge Versus Bias	118
5.5.3 Derivation of Small Signal Capacitance	118
5.5.4 Physical Interpretation of Depletion Cap	118
5.5.5 A Variable Capacitor (Varactor)	119
6 PN Junction Currents	121
6.1 Chapter Preview	121
6.2 Qualitative Overview of Diode Currents	122
6.2.1 Diode under Thermal Equilibrium	122
6.2.2 Reverse Bias	122
6.2.3 Forward Bias	123
6.2.4 Diode I-V Curve	123
6.2.5 Fabrication of IC Diodes	124
6.3 Carrier concentration in Non-Equilibrium	124
6.3.1 Equilibrium versus Non-Equilibrium	124
6.3.2 Recombination-Generation Centers	125
6.3.3 Thermal Generation/Recombination Rate	125
6.4 Current Continuity Equation	126
6.4.1 Excess Carrier Continuity	126
6.4.2 Diffusion Currents	127
6.4.3 Excess Carrier Current Flow	127
6.5 Forward Biased PN-junctions	128
6.5.1 Minority Carriers at Junction Edges	128
6.5.2 Minority Carrier Concentration Distribution	130
6.5.3 Diode Diffusion Currents	131
6.5.4 Height Analogy	131

6.6	Diode Small Signal Model	132
6.6.1	Diode Capacitance	132
6.6.2	Complete Small-Signal Model	133
6.7	Photonic Applications	134
6.7.1	Solar Cells	134
6.7.2	<i>PN</i> -junction Solar Cells	135
6.7.3	Equations for Optical Generation	135
6.7.4	Short Circuit Current	136
6.7.5	Open Circuit Voltage	137
6.7.6	Light Emitting Diodes (LEDs)	137
6.7.7	LED Materials and Structure	139
6.8	References	140
7	MOS Capacitor	141
7.1	Chapter Preview	141
7.2	MOS Capacitor Structure	142
7.2.1	MOS Capacitor	142
7.2.2	Metal-Oxide-Semiconductor Junction	143
7.2.3	Gate Materials and Contact Potential	144
7.3	MOS Regions of Operation	145
7.3.1	Fields and Charge at Equilibrium	145
7.3.2	Flat-Band Voltage, $V_{GB} = V_{FB}$	145
7.3.3	Accumulation, $V_{GB} < V_{FB}$	146
7.3.4	Depletion, $V_{GB} > V_{FB}$	146
7.3.5	Inversion, $V_{GB} = V_T$	147
7.4	MOS Device Threshold Voltage	148
7.4.1	Threshold Voltage Definition	148
7.4.2	Derivation of V_T	148
7.5	Fields and Potential in Oxide/Substrate	149
7.5.1	Fields in oxide/substrate	149
7.5.2	Potential Variations in Oxide/Substrate	150
7.6	Charge-Voltage (Q-V) and Capacitance-Voltage (C-V) Curves	151
7.6.1	Q - V Curve for MOS Capacitor	151
7.6.2	MOS C - V Curve	151
7.6.3	C - V Curve Equivalent Circuits	151
7.6.4	Numerical Example	152
8	MOSFET Device Physics	155
8.1	Chapter Preview	155
8.2	Device Layout and Cross Section	156
8.2.1	MOSFET Top View and Layout	156
8.2.2	MOSFET Overview	157
8.2.3	MOSFET Flavors: PMOS and NMOS	158
8.2.4	CMOS Technology	159
8.2.5	Circuit Symbols	159

8.3	NMOSFET Large Signal Models and Regions of Operation	160
8.3.1	Cut-off, $V_{GS} < V_T$	160
8.3.2	Inversion, $V_{GS} > V_T$ and $V_{DS} > 0$	160
8.3.3	Observed Behavior: $I_{DS} - V_{GS}$	161
8.3.4	Observed Behavior: $I_{DS} - V_{DS}$	162
8.4	Derivation of MOSFET $I - V$ Curve	162
8.4.1	MOSFET "Linear" Region	162
8.4.2	MOSFET as a Variable Resistor	163
8.4.3	Approximate Derivation of Inversion Charge Variation	164
8.4.4	Drift Velocity and Drain Current	165
8.4.5	Square Law "Exact" Derivation	165
8.5	Understanding Current Saturation (Pinch Off)	168
8.5.1	The Saturation Region	168
8.5.2	Pinch Off	168
8.5.3	Square-Law Current in Saturation	170
8.5.4	Actual Saturation Current	170
8.5.5	Channel Length Modulation	171
8.5.6	Summary: Regions of Operation	172
8.6	The Complementary PMOS Device	172
8.6.1	PMOS Device	172
9	MOS Transistor Small-Signal Models	175
9.1	Chapter Preview	175
9.2	Introduction to Amplifiers	176
9.2.1	A Simple Circuit: A MOS Amplifier	176
9.2.2	Common Source Amplifier	176
9.3	Operating Points	178
9.3.1	Small Signal vs. Large Signal	178
9.3.2	Selecting the Output Bias Point	178
9.3.3	Finding the Input Bias Voltage	180
9.3.4	Applying the AC Voltage: The "Hard Way"	180
9.3.5	Small-Signal Simplifications	181
9.4	Amplifier Design and Analysis Using the Small-Signal Approach	182
9.4.1	Small-Signal Current	182
9.4.2	The MOS Transconductance	183
9.4.3	Output Resistance r_o	184
9.4.4	Small-Signal Model for MOSFET	185
9.4.5	Small-Signal Analysis Steps: The "Easy Way"	185
9.5	PMOS Amplifier and Small-Signal Model	188
9.5.1	Small-Signal PMOS Model	188
9.5.2	PMOS Amplifier	188
9.6	What We've Ignored	189
10	Complete MOS Small-Signal Model	191
10.1	Chapter Preview	191
10.2	MOSFET Capacitance in Saturation	192
10.2.1	MOSFET Cross Section	192

10.2.2	Gate-Source Capacitance C_{GS}	192
10.2.3	Gate-Drain Capacitance C_{GD}	193
10.2.4	Drain/Source-Bulk Capacitances C_{DB} and C_{SB}	193
10.2.5	Three Terminal Small Signal Model Including Capacitors	194
10.3	Back-Gate Effect	195
10.3.1	All MOSFETs have a "Back Door"	195
10.3.2	Body Bias Affects: V_T - DC Signals	195
10.3.3	Role of the Substrate Potential - AC Signals	196
10.4	Complete Four Terminal MOS Small-Signal Model	197
10.4.1	Four-Terminal Small-Signal Model	197
10.4.2	Complete Small-Signal Model PMOS	197
11	Bipolar Junction Transistors	199
11.1	Chapter Preview	199
11.2	Overview of BJT	200
11.2.1	Ideal BJT Structure	200
11.2.2	Actual BJT Cross-Section	201
11.2.3	BJT Layout	202
11.2.4	BJT Schematic Symbol	203
11.3	Observed $I-V$ Characteristics	204
11.3.1	Base-Emitter Voltage Control	204
11.3.2	Collector Characteristics (Sweep I_B)	205
11.4	BJT Physics without Equations	206
11.4.1	Transistor Action	206
11.4.2	Diffusion Currents	206
11.4.3	BJT Currents	207
11.5	BJT Device Physics with Equations	209
11.5.1	Collector Current	209
11.5.2	Base Current	209
11.5.3	Current Gain	209
11.6	Large Signal and Small-Signal Equivalent Circuits	210
11.6.1	Ebers-Moll Model	210
11.6.2	Ebers-Moll Equivalent Circuit	210
11.6.3	Forward Active Region and the Early Effect	211
11.6.4	Simplified Ebers-Moll	212
11.6.5	Small-Signal Model	212
12	Single-Stage Amplifiers	215
12.1	Chapter Preview	215
12.2	Review Two-Port Amplifiers	216
12.2.1	One-Port Equivalent Models	216
12.2.2	Small-Signal Two-Port Models	216
12.2.3	Input Impedance Z_{in}	218
12.2.4	Output Impedance Z_{out}	218

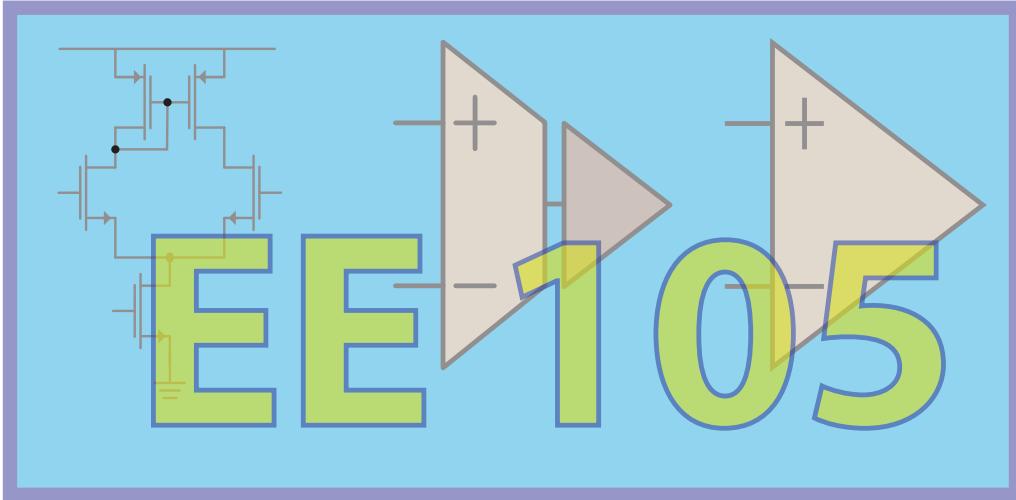
12.3 Single-Stage MOS Amplifier Family	220
12.3.1 Isolating the Bias Points	220
12.3.2 DC Coupled vs AC Coupled Amplifiers	222
12.3.3 Common Gate (CG) Amplifier	222
12.3.4 Common Gate AC Model	222
12.3.5 CG as a Current Amplifier: Find A_i	224
12.3.6 CG Input Impedance	224
12.3.7 CG Output Impedance	226
12.3.8 CG Two-Port Model	227
12.3.9 Common-Gate as a "Voltage Amplifier"	228
12.4 Common Drain Amplifier	230
12.4.1 CD DC Operating Point	230
12.4.2 CD Voltage Gain	231
12.4.3 CD Output Resistance	232
12.5 Impact of Body Effect	233
12.6 Chapter Summary: Amplifiers $\rightarrow G_m/V/I$	235
13 Current Mirrors and Biasing	237
13.1 Chapter Preview	237
13.2 High Load Impedance in Amplifiers	238
13.2.1 Load Impedance	238
13.2.2 Headroom Limitations	238
13.2.3 Achieving High Gain	239
13.2.4 Transistor Current Source	240
13.2.5 Transistor Process / Temperature Variations	240
13.3 The Basic Current Mirror	241
13.3.1 Diode Connected Device	241
13.3.2 Diode Connected – Small-Signal Model	242
13.3.3 The Integrated "Current Mirror"	243
13.3.4 Current Mirror with Multiplication Ratio	244
13.3.5 Current Mirror as Current Source	246
13.3.6 Small-Signal Resistance of Current Source	246
13.4 Improved Current Source: The Cascode Mirror	247
13.4.1 Improved Current Sources	247
13.4.2 Effect of Source Degeneration	247
13.4.3 Cascode (or Stacked) Current Source	248
13.4.4 Drawback of Cascode Current Source	248
13.5 Current Sources and Sinks	250
13.5.1 Generating Multiple Outputs	250
13.6 Example: Source-Follower with Real Current Source	251
13.6.1 Common Drain AC Schematic	252
13.6.2 CD Voltage Gain With Real Current Source	252
13.7 Generation of Current References	253
13.7.1 Constant G_m Reference Current	253

14	Frequency Response of Amplifiers	257
14.1	Chapter Preview	257
14.2	Frequency Response General Considerations	258
14.3	Review MOS Parasitic Capacitors	259
14.4	Common-Source Amplifier Frequency Response	260
14.4.1	Common Source Amplifier: De-Coupling and Coupling Capacitors	260
14.4.2	Typical Passband Frequency Response	261
14.4.3	Common-Source Voltage Amplifier	262
14.5	Common-Source: Brute Force Frequency Response Calculation	263
14.6	The Miller Theorem	264
14.6.1	The Miller Effect	264
14.6.2	The Miller Effect of a Capacitor	264
14.7	Common-Source: Miller Approach Frequency Response Calculation	265
14.8	Common Drain Amplifier Frequency Response	267
14.8.1	Voltage Gain Across C_{gs}	267
14.8.2	Bandwidth of Source Follower	268
14.8.3	Miller Summary	268
14.9	Common Gate Amplifier Bandwidth	269
14.10	Device Unity Gain Frequency f_T	271
14.10.1	Unity-Gain Frequency f_T	271
14.10.2	Frequency Response of Multistage Amplifiers	272
14.11	The Method of Open-Circuit Time Constants (OCTC)	273
14.11.1	OCTC Assumptions	273
14.11.2	Procedure to Find b_1	273
14.11.3	Finding the Thévenin Resistance	273
14.12	Example Calculation: The Common-Source Amplifier Dominant Pole	274
14.12.1	Applying OCTC to CS Amplifier	274
14.12.2	OCTC: $R_{C_{gs}}$	274
14.12.3	OCTC: $R_{C_{gd}}$	274
14.12.4	OCTC: $R_{C_{ab}}$	275
14.12.5	Applying OCTC to CS Amplifier	275
14.12.6	Practical Approach	275
14.13	Chapter Summary	276
15	Multi-Stage Amplifiers	277
15.1	Chapter Preview	277
15.2	The Need for Multi-stage Amplifiers	278
15.3	Review Amplifier Input/Output Impedance Characteristics	279
15.3.1	Impedance "Match"	279
15.4	Common-Source Cascades	280
15.4.1	Two-Stage Voltage Amplifier	280
15.4.2	CS Cascade Analysis	280
15.4.3	CS Cascade Frequency Response	281
15.4.4	Bandwidth Extension	282

15.5 Common-Source with Capacitive Load	283
15.5.1 Common-Source with a Large Capacitive Load	283
15.6 Common-Source Common-Gate Cascade (Cascode)	285
15.6.1 The Common-Source / Common-Gate Cascade	285
15.6.2 Merged CS + CG = Cascode	285
15.6.3 Cascode Gain	286
15.6.4 Cascode Bandwidth	286
15.6.5 Cascode Bandwidth - Small Signal	287
15.6.6 Cascode Biasing	288
15.6.7 Example: Complete Amplifier Design	288
15.7 "Big Circuit" Example	290
15.7.1 Cutting Through the Complexity	290
15.7.2 Eliminate More Clutter	291
16 Differential Amplifiers	293
16.1 Chapter Preview	293
16.2 Motivation for Differential Operation	294
16.3 Birth of Symmetric Source-Coupled Pair	295
16.3.1 Goal: Differential Transconductance G_m	295
16.3.2 Gain of Source Degenerated CS	296
16.3.3 A Symmetric Source Coupled Pair is Born	297
16.4 Differential Operation	299
16.4.1 MOS Differential Pair: DC Bias	299
16.4.2 MOS Differential-Pair: Differential Inputs	300
16.5 Small-Signal Differential Circuits	303
16.5.1 General Differential Drive	303
16.5.2 Pure Differential-Mode Excitation	304
16.5.3 Pure Common-Mode Excitation	304
16.5.4 Differential-Mode "Half Circuit"	305
16.5.5 Complete Small-Signal Differential and Common-Mode Models	306
16.5.6 Common-Mode Operation - Real Current Source	307
16.5.7 Differential and Common-Mode Gain	307
16.5.8 Small-Signal Common-Mode Operation	308
16.5.9 Current-Source Loads	308
16.6 Common-Mode Rejection Ratio and DC Offsets	309
16.6.1 Common-Mode Rejection Ratio	309
16.6.2 Common-Mode Gain with Mismatched R_D	309
16.6.3 Common-Mode Gain with Mismatch of g_m	310
16.6.4 DC Offset	310
16.7 Current Mirror Load	312
16.7.1 Differential Input, Single-End Output	312
16.7.2 Current Mirror Load	312
16.7.3 Common-Mode Gain	314
16.8 Appendix: Large Signal Derivation Steps	316

17	Op-Amp Feedback and Frequency Response	317
17.1	Chapter Preview	317
17.2	Introduction to Feedback	318
17.2.1	Feedback Control	318
17.2.2	Negative Feedback Block Diagram	318
17.2.3	Electronic Feedback	320
17.2.4	History	320
17.3	Why Feedback?	322
17.3.1	Precision Analog	322
17.3.2	Other Benefits of Feedback	322
17.3.3	Loop Gain	322
17.3.4	Noise Rejection	323
17.3.5	Positive Feedback	323
17.4	Circuit Models for Op-Amps	324
17.4.1	Practical Op-Amps	324
17.4.2	World's Simplest Op-Amp Model	324
17.4.3	General Model of Amplifier	325
17.4.4	"Physical" Op-Amp Model	325
17.4.5	Operational Transconductance Amplifier (OTA)	326
17.4.6	Op-Amp Capacitance	326
17.4.7	Transconductance Amplifier Model	327
17.5	Gain/Bandwidth Trade-off	328
17.5.1	Op-Amp Gain / Bandwidth	328
17.5.2	Driving Capacitive Loads	328
17.5.3	Open-Loop Frequency Response	329
17.5.4	Bandwidth Extension	330
17.5.5	Gain / Bandwidth Product in Feedback	331
17.5.6	Unity Gain Feedback Amplifier	331
17.5.7	How Feedback Broadbands an Amplifier	331
17.5.8	Back to Circuit Model	332
17.5.9	Turning a Current Source into a Resistor	333
17.6	Feedback and Stability	334
17.6.1	Stability	334
17.6.2	Non-Dominant Poles	334
17.6.3	Oscillation	334
17.6.4	Instability	335
18	Non-Ideal Operational Amplifiers	337
18.1	Chapter Preview	337
18.2	Offset Voltage and Currents	338
18.2.1	Offset Voltage	338
18.2.2	Trimming of Offset Voltage	339
18.2.3	Input Bias Currents and Offset Currents	339
18.2.4	Effect of Input Bias Current in Amplifier Circuit	340
18.2.5	Reducing the Effect of Input Bias Currents	341
18.3	Op-Amp Frequency Response	342
18.3.1	Frequency Response of Open-Loop Op-Amp	342
18.3.2	Frequency Response of Closed-Loop Op-Amp	342

18.4	Voltage Swing and Slewing	344
18.4.1	Output Saturation	344
18.4.2	Slew Rate	345
18.4.3	Origin of Slew Rate Limit	347
18.4.4	Full-Power Bandwidth	347
18.5	Noise and Distortion	348
18.5.1	Op-Amp Distortion	348
18.5.2	Op-Amp Noise	350
18.6	Datasheets	351
19	References	355
A	Appendix: Calculus Review	357
A.1	Geometric Series	357
A.2	Taylor Series	357
B	Appendix: Linear Algebra Review	359
B.1	Linear Dependence and Independence	359
B.1.1	Proof of Sine and Cosine's Linear Independence	359
C	Appendix: Common Drain Output Resistance	361
D	Appendix: Glossary of Terms	363
D.1	A	363
D.2	B	363
D.3	C	364
D.4	D	365
D.5	E	366
D.6	F	367
D.7	G	369
D.8	H	369
D.9	I	369
D.10	L	370
D.11	M	370
D.12	N	371
D.13	O	371
D.14	P	372
D.15	Q	372
D.16	R	373
D.17	S	373
D.18	T	374
D.19	V	376
D.20	W	376



1. Linear Time-Invariant Systems

1.1 Chapter Preview

In this chapter we will review linearity in the context of an Linear Time-Invariant (LTI) system by making analogies with discrete finite dimensional systems. Many people have familiarity with concepts such as orthogonality and eigenfunctions from linear algebra. It is not critical to have this background, so these references can be skipped. Most importantly, we will show that complex exponential functions are eigenfunctions of linear systems, and they span the space of solutions. We will define the time-domain impulse response and convolution operator, and in a very natural way show the relation between the impulse response and the transfer function in the frequency domain. This will lead us into a preview of the frequency domain representation of linear systems, which we will pick up in the next chapter.

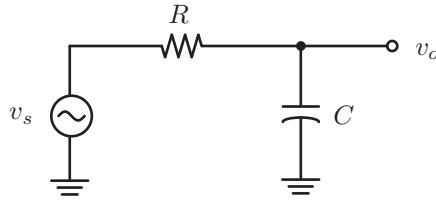


Figure 1.1: A simple RC low-pass filter.

1.2 Example: Analyzing an RLC Circuit

1.2.1 Example: Low Pass Filter (LPF)

Let's begin with an example calculation. Suppose we have a **low pass filter** as shown in *Fig. 1.1*. If we drive this system with a source v_s and observe the output at v_o , we can write the governing equations as follows by using KVL around the loop:

$$\begin{aligned} v_o(t) &= v_s(t) - v_r(t) \\ &= v_s(t) - i(t)R \end{aligned}$$

Recall the current in a capacitor:

$i(t) = C \frac{dv_c(t)}{dt}$

Current in a capacitor as a function of time (1.1)

Since the current is the same through the resistor and capacitor, and the voltage at the output is the same as the voltage across the capacitor, we have:

$$v_o(t) = v_s(t) - RC \frac{dv_o(t)}{dt}$$

Or, in terms of the circuit time constant $\tau = RC$:

$$v_o(t) = v_s(t) - \tau \frac{dv_o(t)}{dt} \quad (1.2)$$

Let's suppose that the input signal is given by $v_s(t) = V_s \cos(\omega t)$. We know that in steady-state the output amplitude and phase will change:

$$v_o(t) = \underbrace{K \cdot V_o}_{V_o} \cos(\omega t + \phi) \quad (1.3)$$

How do we find the change in the amplitude and phase of the signal?

1.2.2 LPF the "hard way"

The following is the wrong way to solve the problem. It's important to review it to see how painful it can be even for a simple example. We start by plugging the known form of the output (*Eq. 1.3*) into the LHS of the governing equation (*Eq. 1.2*), and verify that it satisfies the equation:

$$\begin{aligned} V_o \cos(\omega t + \phi) &= v_s(t) - \tau \frac{d}{dt} \left(v_o(t) \right) \\ &= V_s \cos(\omega t) - \tau \frac{d}{dt} \left(V_o \cos(\omega t + \phi) \right) \quad \text{Substituting for } v_s(t) \text{ and } v_o(t) \\ &= V_s \cos(\omega t) + \tau \omega V_o \sin(\omega t + \phi) \end{aligned}$$

Rearranging, we have:

$$V_s \cos(\omega t) = V_o \cos(\omega t + \phi) - \tau \omega V_o \sin(\omega t + \phi) \quad (1.4)$$

Recall the following trigonometric identities:

$$\begin{aligned} \cos(x+y) &= \cos x \cos y - \sin x \sin y && \text{Sum of angles for cosine} \\ \sin(x+y) &= \sin x \cos y + \cos x \sin y && \text{Sum of angles for sine} \end{aligned}$$

Applying the identities to Eq. 1.4:

$$\begin{aligned} V_s \cos(\omega t) &= V_o (\cos \omega t \cos \phi - \sin \omega t \sin \phi) - \tau \omega V_o (\sin \omega t \cos \phi + \cos \omega t \sin \phi) \\ &= V_o \cos \omega t \cos \phi - V_o \sin \omega t \sin \phi - V_o \tau \omega \sin \omega t \cos \phi + V_o \tau \omega \cos \omega t \sin \phi \end{aligned}$$

Factoring, we have:

$$V_s \cos(\omega t) = V_o \cos(\omega t) (\cos \phi - \tau \omega \sin \phi) - V_o \sin(\omega t) (\sin \phi + \tau \omega \cos \phi) \quad (1.5)$$

Since sine and cosine are linearly independent¹ functions:

$$a_1 \sin(\omega t) + a_2 \cos(\omega t) = 0 \quad (1.6)$$

This implies that $a_1 = 0$ and $a_2 = 0$ must be zero independently. Applying the linear independence gives us

$$-V_o \sin \phi - V_o \tau \omega \cos \phi = 0 \quad (1.7)$$

and

$$\tan \phi = -\tau \omega \quad (1.8)$$

The phase response is therefore

$$\phi = -\tan^{-1} \tau \omega \quad (1.9)$$

Likewise we have

$$V_o \cos \phi - V_o \tau \omega \sin \phi - V_s = 0 \quad (1.10)$$

$$V_o (\cos \phi - \tau \omega \sin \phi) = V_s \quad (1.11)$$

$$V_o \cos \phi (1 - \tau \omega \tan \phi) = V_s \quad (1.12)$$

$$V_o \cos \phi (1 + (\tau \omega)^2)^{1/2} = V_s \quad (1.13)$$

$$V_o (1 + (\tau \omega)^2)^{1/2} = V_s \quad (1.14)$$

The amplitude response is therefore given by

$$\frac{V_o}{V_s} = \frac{1}{\sqrt{1 + (\tau \omega)^2}} \quad (1.15)$$

We can see that both the **amplitude** and **phase response** are a strong function of frequency. At very low frequencies, $\omega \approx 0$, the signal passes undisturbed. At very high frequencies $\omega \rightarrow \infty$, the signal experiences infinite attenuation and is effectively shorted out by the capacitor.

¹See Appendix B for a review of linear dependence/independence, and a proof for sine and cosine are LI.

LPF Magnitude and Phase Response

Plots of the magnitude and phase response are shown in *Fig. 1.2(a & b)*. The magnitude plot shows that for signal frequencies below the cutoff frequency, $1/\tau$, the signal experiences little attenuation. In fact, at the passband edge frequency of $1/\tau$, the magnitude is by definition lower by $1/\sqrt{2}$, which corresponds to half the power. Likewise, the phase response is zero at low frequencies, meaning the signal is not delayed, but even at the cutoff frequency the delay is 45° , and asymptotically reaches 90° .

As is common practice in electrical engineering, we more often plot the magnitude on a log-log scale with units of dB, as shown in *Fig. 1.2(c)*. On the deciBel (deci = 10) scale, we take the base $10 \log_{10}$ and multiply by 10 to get rid of fractional parts. The "Bel" part is in honor of the inventor of the phone,. The dB scale allows us to expand the passband and more clearly understand the behavior of the transfer function as a function of frequency. The multiplication by 10 is just for convenience.

But you'll notice that we often multiply log by 20 rather than 10 in the definition of dB. Why is that? Remember that power is proportional to voltage squared:

$$\text{dB} = 10 \log \left(\frac{V_o}{V_s} \right)^2 = 20 \log \left(\frac{V_o}{V_s} \right) \quad (1.16)$$

which means that if we take the log of voltage, we can just multiply by 2 o convert it into power. At various frequencies we have:

$$\omega = 1/\tau \rightarrow \left(\frac{V_o}{V_s} \right)_{\text{dB}} = -3 \text{dB} \quad (1.17)$$

$$\omega = 100/\tau \rightarrow \left(\frac{V_o}{V_s} \right)_{\text{dB}} = -40 \text{dB} \quad (1.18)$$

$$\omega = 1000/\tau \rightarrow \left(\frac{V_o}{V_s} \right)_{\text{dB}} = -60 \text{dB} \quad (1.19)$$

Observe that the slope of signal attenuation is 20 dB/decade in frequency beyond the passband. Alternatively, if you double the frequency, the attenuation changes by 6 dB, or 6 dB/decade. This is an important rule of thumb when understanding the frequency response of a system.

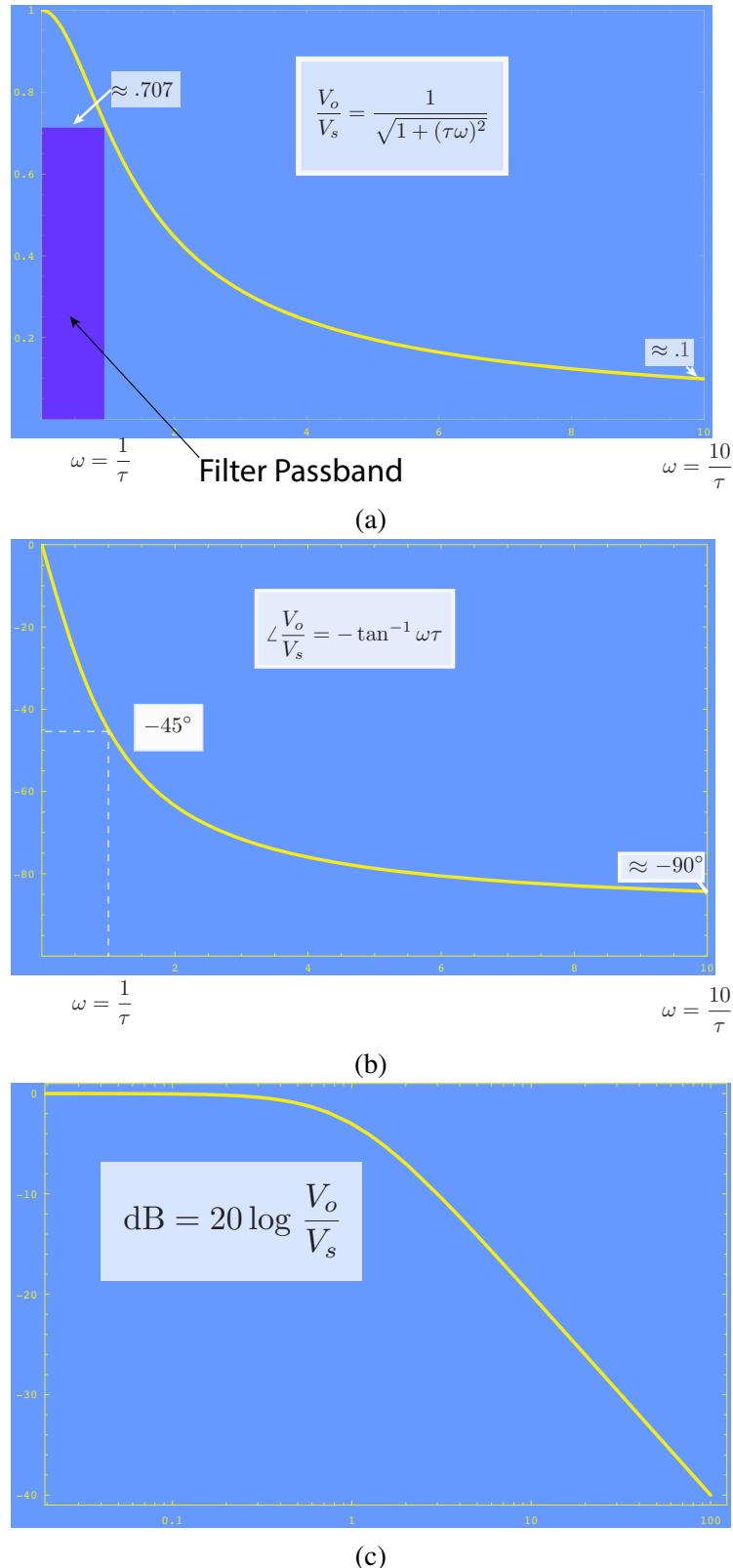


Figure 1.2: The (a) magnitude and (b) phase response of the low-pass filter. The (c) log-log plot of the magnitude response of a low-pass filter (dB scale).

1.3 Linear Time-Invariant Systems

1.3.1 Linear Time-Invariant Definition

A system \mathbf{L} is linear if it satisfies the following requirement:

$$\mathbf{L}[\alpha x(t) + \beta y(t)] = \alpha \mathbf{L}[x(t)] + \beta \mathbf{L}[y(t)] \quad (1.20)$$

In this notation, \mathbf{L} is an operator that corresponds to the action of the system at hand. If we input $x(t)$ an input into our system, the system outputs $y(t)$. In a **linear system**, if we scale the input, we also scale the output by the same amount. Also, if we take a superposition of inputs, the output is the same superposition of the inputs.

Now, a system is **time invariant** if its response is not a function of time. In other words, the system does not "age" and it does not care when you apply the input. If you apply an input today and observe the output, and come back tomorrow and put in the same input, the output should look the same. Mathematically, we can summarize this behavior with the following equations:

$$y(t) = \mathbf{L}[x(t)] \quad (1.21)$$

$$\mathbf{L}[x(t - \tau)] = y(t - \tau) \quad (1.22)$$

Notice that **time shifting** the input just corresponds to applying the same time shift to the output. Another way to think of time-invariance is to say that the system has no "clock" or time reference, or that the transfer function is not a function of time. It does not matter when you apply the input. The transfer function is going to be the same.

If you are familiar with discrete linear system, you know that linearity and superposition are fundamental properties.

Vectors and Matrices Become Functions and Integrals

In finite dimensional discrete linear systems, inputs and outputs are defined in terms of input and output vectors. For LTI systems, we can think of the inputs and outputs as infinite dimensional vectors, but instead of a countably infinite set of elements we have a uncountably large number of time instants that define the input and output. In other words, we represent the input and output as continuous functions of time. We will see that in LTI systems, we can also represent linear systems by basis functions:

$$\phi_1(t), \phi_2(t), \dots \quad (1.23)$$

and instead of a matrix representation (summation) we will have an integral representation called the **Convolution Integral**. Eigenvectors have the direct analog of Eigenfunctions:

$$\mathbf{L}[\phi_n(t)] = \lambda_n \phi_n(t) \quad (1.24)$$

And the concept of an **Orthonormal Basis** maps directly into LTI systems:

$$x(t) = \sum_n \phi_n(t) x_n \quad (1.25)$$

$$x(t) = \int \phi(st)x(t)dt \quad (1.26)$$

We can therefore perform **Eigenfunction** expansion of an arbitrary input and study the response of a linear system to an arbitrary input by using our knowledge of how the system responds to the eigenfunctions. In other words, the eigenvalues of the system, or the spectral expansion, will be the most important aspect of a linear system.

If this doesn't make any sense, don't worry. We'll cover all of this in a step-by-step fashion. This is only meant as a teaser for people who have seen linear system theory previously.

1.4 The Complex Exponential Technique

We begin by introducing the complex exponential, and we will show that complex exponentials are eigenfunctions of linear systems. This means that they play an immensely central role in linear system analysis.

1.4.1 Why introduce complex numbers?

Complex exponentials actually make things less complex. To see this, first note that integration and differentiation are trivial with complex exponentials:

$$\frac{d}{dt} e^{j\omega t} = j\omega e^{j\omega t} \quad (1.27)$$

$$\int e^{j\omega x} dx = \frac{1}{j\omega} e^{j\omega x} \quad (1.28)$$

So now any ordinary **differential equation** (ODE), a sum of differentiation operators, is reduced now to trivial algebraic manipulations. In fact, we'll show that you don't even need to directly derive the ODE by using **phasors** (a phasor is essentially a shorthand notation for a complex number). The key is to observe that the current/voltage relation for any element can be derived for **complex exponential** excitation.

1.4.2 Complex Exponential

From Euler's we have an important relationship between complex exponentials and ordinary sin and cos functions:

$$e^{jx} = \cos x + j \sin x \quad (1.29)$$

If take the magnitude of this quantity, it's unity

$$|e^{jx}| = \sqrt{\cos^2 x + \sin^2 x} = 1 \quad (1.30)$$

This means that for any value of x , the magnitude of a complex exponential is 1. Thus, $e^{j\phi}$ is a point on the unit circle at an angle of ϕ from the x -axis.

Any complex number z (Fig. 1.3), expressed as having a real and imaginary part $z = x + jy$, can also be interpreted as having a magnitude and a phase. The magnitude $|z| = \sqrt{x^2 + y^2}$ and the phase $\phi = \angle z = \tan^{-1} y/x$ can be combined using the complex exponential

$$x + jy = |z|e^{j\phi} \quad (1.31)$$

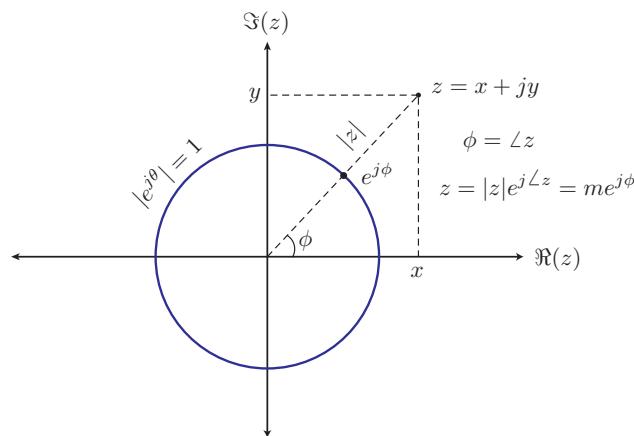


Figure 1.3: A general complex number z can be represented by its real and imaginary components, or its magnitude and phase.

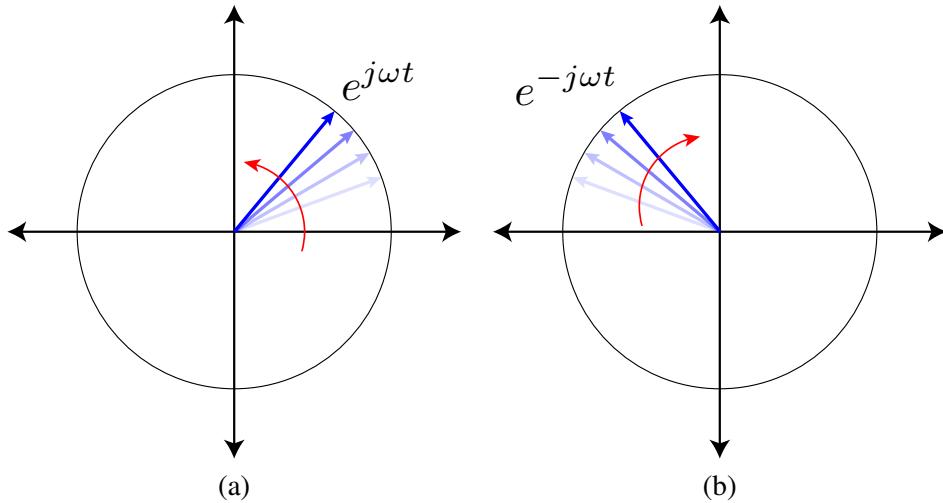


Figure 1.4: (a) The complex exponential $e^{j\omega t}$ rotates around the unit circle counter-clockwise at a rate of ω , or once every $2\pi/\omega$ seconds. Likewise (b) $e^{-j\omega t}$ rotates around the unit circle clockwise. Click [e^{j\omega t} here](#) and [e^{-j\omega t} here](#) to see the animations.

1.4.3 Euler's Relations and The Circle

The argument to $e^{j\omega t}$ is just linearly increasing with a slope of ω , and completes 2π radian every $\frac{2\pi}{\omega}$ seconds. This implies that $e^{j\omega t}$ is nothing but a point rotating on a circle on the complex plane. The real part and imaginary parts are just projections of the circle, which by trigonometry we know equal the cosine and sine functions.

We can also express cos and sin in terms of e as follows

$$\cos x = \frac{e^{jx} + e^{-jx}}{2} \quad (1.32)$$

$$\sin x = \frac{e^{jx} - e^{-jx}}{2j} \quad (1.33)$$

which shows that two counter rotating complex exponentials sum to sin and cos. We can visualize these relations as shown in *Fig. 1.4*. Be sure to click the links to see an animation, especially for the sum of two counter rotating complex exponentials in *Fig. 1.5*.

1.4.4 The Magic of Sinusoids

What were going to show is that when a linear, time invariant (LTI) circuit is excited by a sinusoid (see *Fig. 1.6*), its output is a sinusoid at the same frequency. Only the magnitude and phase of the output differ from the input. Sinusoids are very special functions for LTI systems, in other words they are eigenfunctions. The "**Frequency Response**" is a characterization of the input-output response for sinusoidal inputs at all frequencies.

Since most periodic (non-periodic) signals can be decomposed into a summation (integration) of sinusoids via Fourier Series (Transform), the response of a LTI system to virtually any input is characterized by the frequency response of the system. We'll return to this point later.

Complex Exponential is Powerful

To find steady state response we can excite the system with a complex exponential. Here our system is any general LTI system. In terms of circuits, it is any linear circuit consisting of resistors, capacitors, inductors, mutual inductors, transformers, and linear dependent sources. At

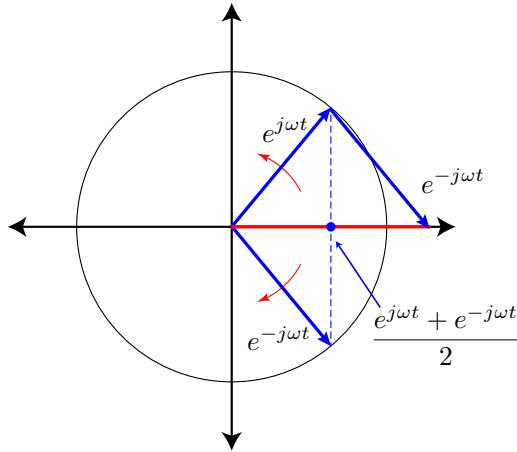


Figure 1.5: We visualize the function \cos by building it through vectorwise addition of $e^{j\omega t}$ and $e^{-j\omega t}$. Click [here](#) to see the animation.

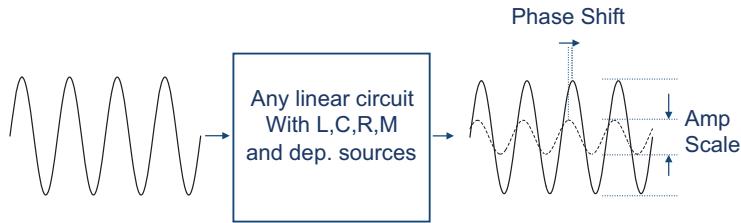


Figure 1.6: In an LTI system, if we drive it with a sinusoid, the output is a sinusoid at the same frequency, with a new amplitude and phase.

any frequency, the system response is characterized by a single complex number H , as shown in Fig. 1.7. The magnitude response is given by $|H(\omega)|$ and the phase response is given by $\angle H$. We see that the complex exponential is an "eigenfunction" of the system. It is used to probe the system. If we characterize the response to all eigenfunctions, we can completely characterize the system. Because a sinusoid is a sum of complex exponentials (and because of linearity!), we can also probe a system by applying a real sinusoidal input. This is what we do in the lab.

H is the Transfer Function

In summary, $e^{j\omega}$ is an eigenfunction for any linear system (circuit), and $H(\omega)$ is the eigenvalue of the system. There's a continuous spectrum of eigenvalues $H(\omega)$. The linear system is completely characterized by the spectrum of eigenvalues, or in the frequency domain by the **transfer function** $H(\omega)$. We also often write $H(\omega)$ in the following form (explained shortly):

$$H(j\omega) \quad (1.34)$$

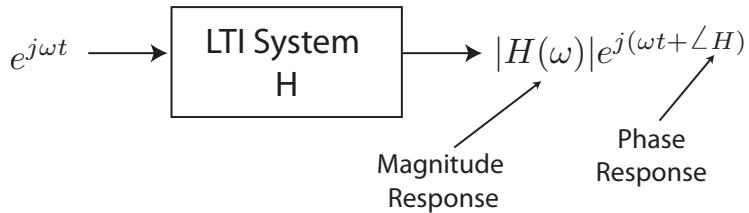


Figure 1.7: The complex exponential response of a linear system is another complex exponential, with an additional magnitude and phase factor.

1.5 The Low Pass Filter Again

1.5.1 LPF Example: The "soft way"

Remember that earlier we tried to analyze the low-pass filter using sinusoids. Now let's excite the system with a complex exponential. Recall that the sinusoidal response is related to the exponential response so we can always recover the sinusoidal response from the complex exponential response

$$v_s(t) = v_o(t) + \tau \frac{dv_o}{dt} \quad (1.35)$$

Derivatives are simply products with $j\omega$

$$v_s(t) = V_s e^{j\omega t} \quad (1.36)$$

Grouping terms

$$v_o(t) = |V_o| e^{j(\omega t + \phi)} = V_o e^{j\omega t} \quad (1.37)$$

Now substitute into the original equation

$$V_s e^{j\omega t} = V_o e^{j\omega t} + \tau \cdot j\omega \cdot V_o e^{j\omega t} \quad (1.38)$$

and divide out the non-zero common factors ($|e^{j\omega t}| = 1$)

$$V_s = V_o (1 + j\omega \cdot \tau) \quad (1.39)$$

So with a few lines of algebra, we have the transfer function:

$$\frac{V_o}{V_s} = \frac{1}{1 + j\omega \cdot \tau} \quad (1.40)$$

The system is characterized by the complex function

$$H(\omega) = \frac{V_o}{V_s} = \frac{1}{1 + j\omega \cdot \tau} \quad (1.41)$$

The magnitude and phase response match our previous calculation:

$$|H(\omega)| = \left| \frac{V_o}{V_s} \right| = \frac{1}{\sqrt{1 + (\omega\tau)^2}} \quad (1.42)$$

$$\angle H(\omega) = -\tan^{-1} \omega \tau \quad (1.43)$$

Clearly, using the complex exponential is much faster and easier than the earlier approach where we used sinusoids directly and the final answer is the same. It's not obvious why the transfer function for a complex exponential matches the one for the sinusoidal excitation. Let's explore this further.

Why did it work?

If we push the complex exponential through the system and take the real part of the output, then why is that the same as the "real" sinusoidal response? One argument to explain why the complex exponential works in place of a sinusoid is to observe that sine and cosine are simply imaginary and real parts of the complex exponential. So any signal (current or voltage) can be written in the following equivalent forms

$$s(t) = S_o \cos(\omega t + \phi) = S_o \Re[e^{j(\omega t + \phi)}] \quad (1.44)$$

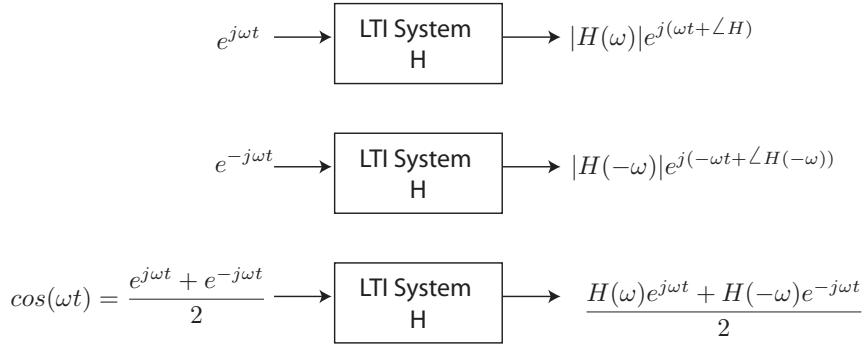


Figure 1.8: Sinusoidal response of an LTI system by superposition.

For example, if we excite our system with an input $z(t)$ that is complex, producing an output $y_z(t)$, and if we take the real or imaginary part of the signals, we have

$$y_z(t) = \mathbf{L}(z(t)) \quad (1.45)$$

Now take the real part of both sides and use the fact that the order of the operator \Re and our system \mathbf{L} can be interchanged:

$$\Re[y_z(t)] = \Re[\mathbf{L}(z(t))] = \mathbf{L}(\Re[z(t)]) \quad (1.46)$$

If this argument does not convince you, there's a slightly more complicated way to show this must be true.

And yet another perspective

As shown in *Fig. 1.8*, another way to see this is to observe that the system is linear so that we can push through two inputs separately and sum the outputs:

$$y = \mathbf{L}(s_1(t) + s_2(t)) = \mathbf{L}(s_1(t)) + \mathbf{L}(s_2(t)) \quad (1.47)$$

Similarly, to find the response to a sinusoid, we can find the response to $e^{j\omega t}$ and $e^{-j\omega t}$, and sum the results. For a linear system represented by real circuit elements, if the input is real function of time, the output must also be a real function:

$$y(t) = \frac{H(\omega)e^{j\omega t} + H(-\omega)e^{-j\omega t}}{2} \quad (1.48)$$

The only way for two complex numbers to sum to a real number is if they are complex conjugates. That means the second term is the conjugate of the first, which implies that

$$|H(-\omega)| = |H(\omega)| \quad (1.49)$$

Or in other words, the magnitude response is an even function of frequency. Likewise, we have

$$\angle H(-\omega) = -\angle H(\omega) = -\phi \quad (1.50)$$

Or the phase function has to be an odd function of frequency. Therefore the output is given by

$$y(t) = \frac{|H(\omega)|}{2} \left(e^{j(\omega t + \phi)} + e^{-j(\omega t + \phi)} \right) \quad (1.51)$$

or more simply

$$|H(\omega)| \cos(\omega t + \phi) \quad (1.52)$$

1.6 Generalization to any Linear Circuit

1.6.1 "Proof" for Linear Systems

For an arbitrary linear circuit, in other words a black box containing an arbitrary number of inductors, capacitors, resistors, mutual inductors, or linear dependent sources, we can decompose the system into linear sub-operators, like multiplication by constants, time derivatives, or integrals:

$$y = \mathbf{L}(x) = ax + b_1 \frac{d}{dt}x + b_2 \frac{d^2}{dt^2}x + \dots + c_1 \int x + c_2 \iint x + \dots \quad (1.53)$$

For a complex exponential input this simplifies to:

$$y = \mathbf{L}(e^{j\omega t}) = ae^{j\omega t} + b_1 \frac{d}{dt}e^{j\omega t} + b_2 \frac{d^2}{dt^2}e^{j\omega t} + \dots + c_1 \int e^{j\omega t} + c_2 \iint e^{j\omega t} + \dots \quad (1.54)$$

$$y = ae^{j\omega t} + b_1 j\omega e^{j\omega t} + b_2 (j\omega)^2 e^{j\omega t} + \dots + c_1 \frac{e^{j\omega t}}{j\omega} + c_2 \frac{e^{j\omega t}}{(j\omega)^2} + \dots \quad (1.55)$$

Note that every term is of the form $e^{j\omega t}$ times a constant, which when grouped together gives

$$y = e^{j\omega t} \underbrace{\left(a + b_1 j\omega + b_2 (j\omega)^2 + \dots + c_1 \frac{1}{j\omega} + c_2 \frac{1}{(j\omega)^2} + \dots \right)}_{H} \quad (1.56)$$

We've grouped together all the terms multiplying $e^{j\omega t}$ and note that at a given frequency, they sum to a complex number H . The amplitude of the output is the magnitude of the complex number H and the phase of the output is the phase of the complex number H .

$$y = e^{j\omega t} \underbrace{\left(a + b_1 j\omega + b_2 (j\omega)^2 + \dots + \frac{1}{j\omega} + \frac{1}{(j\omega)^2} + \dots \right)}_{H} \quad (1.57)$$

Written more compactly, we have

$$y = e^{j\omega t} |H(\omega)| e^{j\angle H(\omega)} \quad (1.58)$$

or equivalently

$$\Re[y] = |H(\omega)| \cos(\omega t + \angle H(\omega)) \quad (1.59)$$

1.6.2 General Complex Exponential

Looking back at our derivation, it's clear that a complex exponential is also an eigenfunction of our system, and the periodic complex exponential is just a special case. We can make the same argument for a general input of the form e^{st} :

$$y = \mathbf{L}(x) = ax + b_1 \frac{d}{dt}x + b_2 \frac{d^2}{dt^2}x + \dots + c_1 \int x + c_2 \iint x + \dots \quad (1.60)$$

$$y = e^{st} \underbrace{\left(a + b_1 s + b_2 s^2 + \dots + \frac{1}{s} + \frac{1}{s^2} + \dots \right)}_{H(s)} \quad (1.61)$$

Later we'll see that $H(s)$ can be used to solve a problem involving the **transient response** of a system (response to an initial input) whereas $H(j\omega)$ is used to find the **steady-state response** to a sinusoidal input.

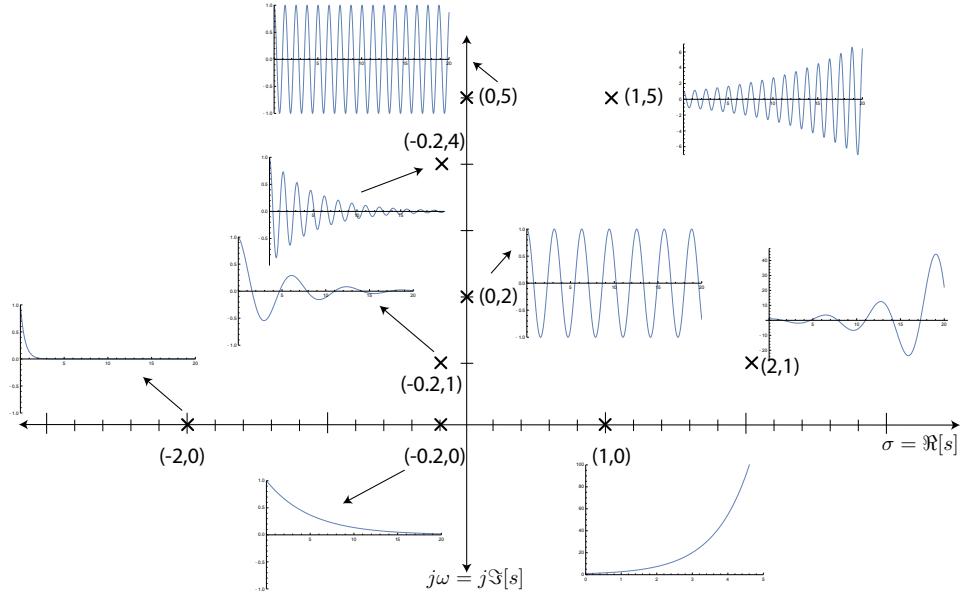


Figure 1.9: On the complex plane, the location of the argument s to e^{st} corresponds to either sinusoidal harmonic oscillation (on the imaginary axis), or exponential growth or decay (on the real axis). Other off-axis locations are a combination of decaying or growing exponential sinusoidal waveforms.

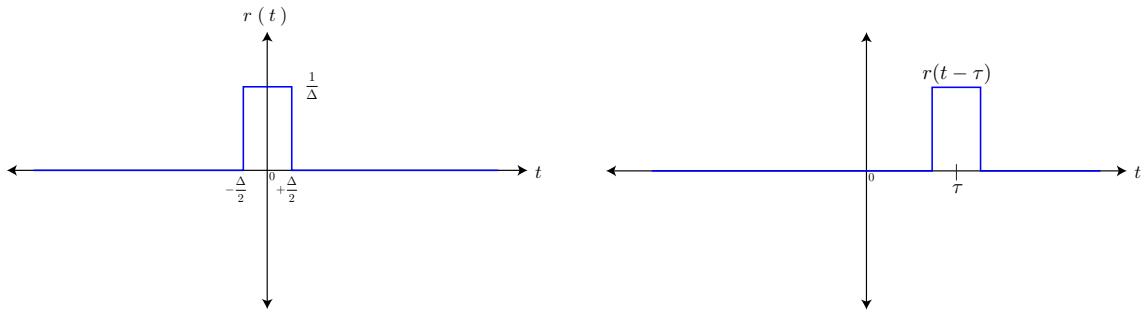


Figure 1.10: The unit rectangular function, centered at the origin or shifted by τ .

Complex Plane

In Fig. 1.9 we show the "s-plane", and the "x" marks on the diagram indicate particular complex values of s , and the effect these values have on the shape of the response. Notice that the real axis corresponds to the exponential envelope growth or decay. The imaginary axis corresponds to the rate of oscillation. On the real axis, there's no oscillation at all. On the imaginary axis, there's no decay. This plot shows that the general exponential e^{st} can represent different kinds of functional behavior including steady-state oscillation (sinusoidal response) and other decaying and growing oscillation amplitudes.

1.7 Time Domain Characterization of Linear Systems

1.7.1 Unit Area Rectangular Function

A very convenient function, the **unit rectangular function**, shown in Fig. 1.10, is defined as follows. Consider a rectangular function with unit area:

$$r(t) = \begin{cases} \frac{1}{\Delta} & -\Delta/2 < t < \Delta/2 \\ 0 & \text{otherwise} \end{cases} \quad (1.62)$$

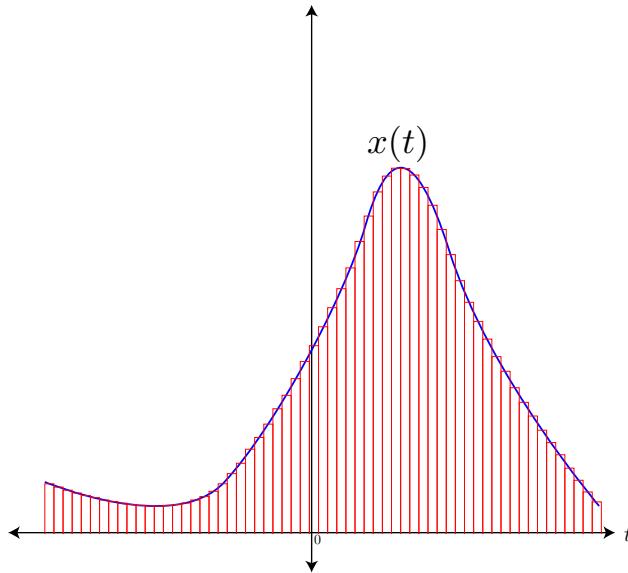


Figure 1.11: Representation of a general function by a sum of shifted rectangular functions.

By definition, we have a unit area function

$$\int_{-\infty}^{\infty} r(t)dt = 1 \quad (1.63)$$

Function Approximation

As shown in *Fig. 1.11*, we can approximate any function as a sum of rectangular shifted functions

$$x(t) \approx \sum_i x(i\tau)r(t - i\tau)\Delta \quad (1.64)$$

As the rectangles become smaller and smaller, the representation becomes more accurate. The key question is what happens as we take the limit of a very small time step Δ ?

The Delta Function $\delta(t)$ – A Strange Beast

Now imagine taking the limit as $\Delta \rightarrow 0$

$$\lim_{\Delta \rightarrow 0} r(t) = \delta(t) \quad (1.65)$$

What do we get? Is this thing even a function? A pictorial representation of the limiting process is shown in *Fig. 1.12*. We see that the function becomes more and more narrow and taller and taller. No, it's not an ordinary function.² The **delta function** was originally introduced by Oliver Heaviside ~1900, Dirac co-invented it and gets credit as it's known as the "Dirac Delta Function".³

Dirac Delta Sifting Property

As shown in *Fig. 1.13*, one interesting thing about the Delta function is that it can be used to basically pick out a certain value of a function (because it's zero except at one location):

$$x(\tau) = \int \delta(t - \tau)x(t) dt \quad (1.66)$$

²Mathematicians have made sense of it. Now it's known as a distribution or generalized function.

³Heaviside was derided by the mathematicians of the day for his methods. He often got the right answer without being mathematically rigorous, but that's because he was way ahead of everyone in his techniques, and it took a while for the mathematical world to catch up.

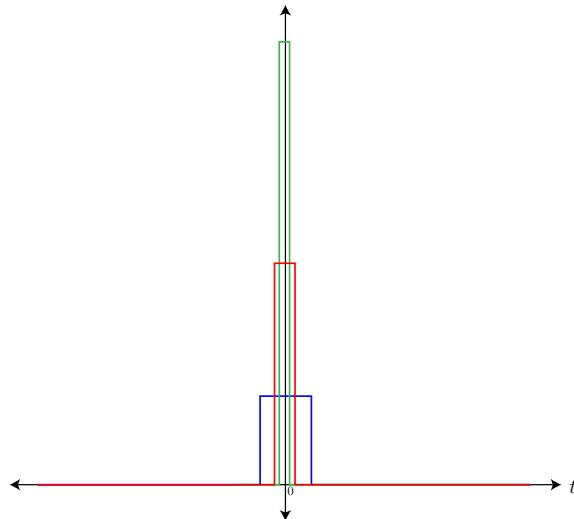


Figure 1.12: The "Delta Function" can be obtained through a limiting process by making the interval of the rectangular function shorter and shorter.

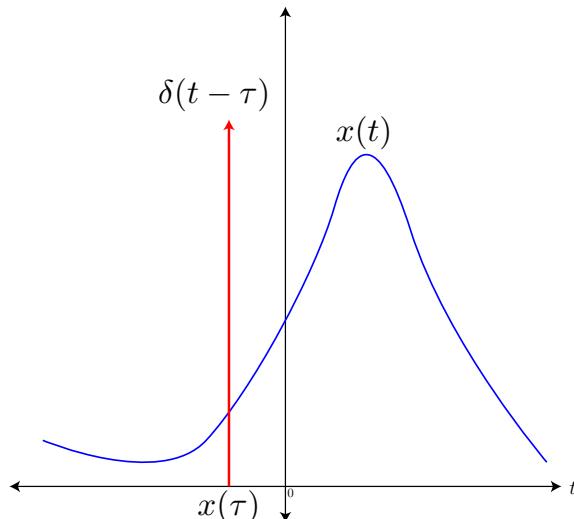


Figure 1.13: The "sifting" property of the Delta function yields the function value at any point.

This is not obvious unless you take the limit and then it becomes clear that the integral converges to the value of the function. So the Dirac Delta function is better defined in terms of this property, and many functions satisfy this function and can be used to define the Delta function.

Recall that we started by using the rectangular function to approximate any arbitrary function. As we take the limit, we are reconstructing our function point-by-point using weighted sums of Delta functions. To see this, simply start with the approximated function as before

$$x(t) \approx \sum_i x(i\tau) r(t - i\tau) \Delta \quad (1.67)$$

As we take the limit, we expect that our approximation becomes more exact and we have the following integral:

$$x(t) = \int x(\tau) \delta(t - \tau) d\tau \quad (1.68)$$

In some ways this is not really a deep result in the sense of saying that a function is a continuous stream of real numbers, and any particular value can be "sifted" out by multiplying by a delta and

integrating around the delta. If you think of the function as a vector, the delta function is like the vector $e_n = [0 \cdots 0 1 0 \cdots 0]^T$ with the 1 in the n'th position.

1.8 Back to Time Domain: Impulse Response

Up to now we studied the sinusoidal steady-state response using the complex exponential. But we are now armed with the right tools to tackle the **general response** of a linear system to an arbitrary input in time. Using the fundamental properties of linearity and superposition, we have

$$y(t) = \mathbf{L}[x(t)] = \mathbf{L} \left[\int \delta(t - \tau)x(\tau)d\tau \right] \quad (1.69)$$

$$y(t) = \mathbf{L}[x(t)] = \int \mathbf{L}[\delta(t - \tau)x(\tau)]d\tau = \int x(\tau)\mathbf{L}[\delta(t - \tau)]d\tau \quad (1.70)$$

Let's define the impulse function as the response of a linear system to a Delta function applied at time τ

$$h(t, \tau) \equiv \mathbf{L}[\delta(t - \tau)] \quad (1.71)$$

Notice that in a time-invariant system, it should not matter when we apply the Delta function, as long as we shift the output. So if $h(t)$ is the response to a Delta function at time 0, $h(t - \tau)$ is the response to an input applied at time τ . So in summary, for a linear time-invariant system (LTI), we only need to characterize the system response to a Delta function once and for all to find

$$h(t) \equiv \mathbf{L}[\delta(t)] \quad (1.72)$$

This means that $h(t)$ contains all the information about the linear system. If we call the Dirac Delta function an "impulse", then for obvious reasons, we call $h(t)$ the "**impulse response function**".

1.8.1 Convolution Operation

We have demonstrated that any linear system input/output response can be represented by the following integral:

$$y(t) = \int_{-\infty}^{\infty} h(t - \tau)x(\tau)d\tau \quad (1.73)$$

Think of this as the sum of the input weighted by the impulse response function. The operation is known as the convolution operation and symbolically defined by the operator "*":

$$y(t) = h(t) * x(t) \quad (1.74)$$

The symbol * is the **convolution operator**. You can also change variables $x = t - \tau$ and write it as:

$$y(t) = \int_{\infty}^{-\infty} h(x)x(t - x)(-dx) = \int_{-\infty}^{\infty} h(x)x(t - x)dx \quad (1.75)$$

In this equation, x is just a dummy variable. We can substitute τ to make the equation look the same as before (recall from calculus that $\int_b^a x = -\int_a^b x$):

$$y(t) = \int_{-\infty}^{\infty} x(t - \tau)h(\tau)d\tau \quad (1.76)$$

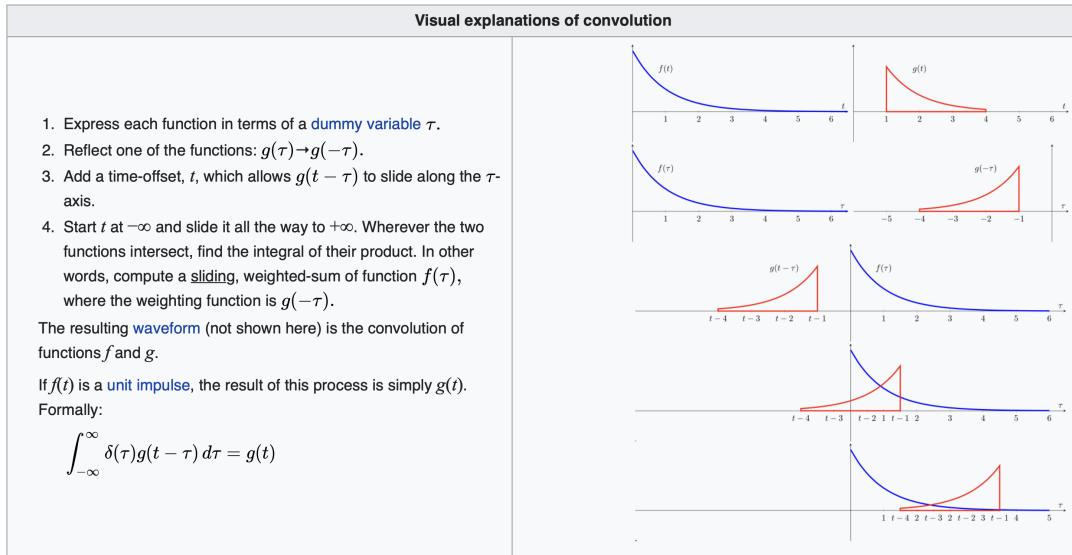


Figure 1.14: The convolution operation visualized. See Wikipedia for full details and an animation.

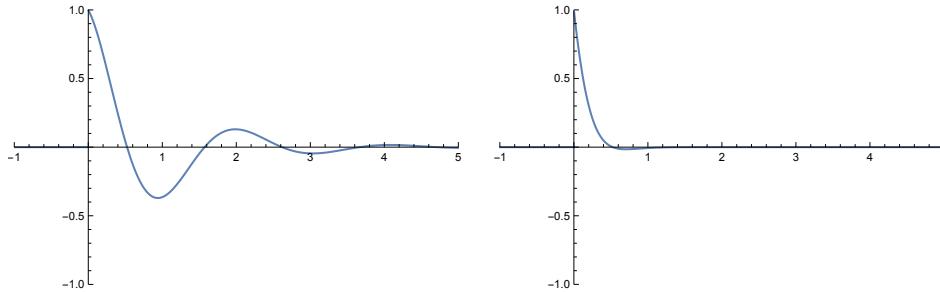


Figure 1.15: Two hypothetical impulse response functions are shown. One is long and oscillates, the second is short and delays exponentially.

Convolution Visualization

It's useful to visualize the convolution operation. Note that the convolution operation takes in two functions and produces a new function. The operation is described nicely on Wikipedia⁴ reproduced here in *Fig. 1.14*. As described in the figure, we first express each function in terms of the dummy variable (of integration) τ . Next, we reflect one of the functions about the time axis. For example, $h(\tau)$ becomes $h(-\tau)$. The physical explanation for this will be described shortly. Next we compute a sliding weighted-sum of the function $x(\tau)$ with $h(t - \tau)$. Only points where the two functions intersect contribute significantly to the waveform.

Memory

For example, if $h(t)$ is mostly zero except over a small interval, The fact that the impulse response function has time duration is an indication that the system has memory. The current output depends not only on the current input, but also past values of the input. The longer the duration of the impulse response function, the more "memory" the system has. Capacitors and inductors store energy and act as memory in circuits. They store energy in the magnetic and electric fields, and resonance can occur because these forms of energy can be exchanged. The impulse response function for a system with no memory is just another delta function! Two examples are shown in *Fig. 1.15* to illustrate a system with a longer vacillating impulse response versus a system that has a

⁴<https://en.wikipedia.org/wiki/Convolution>

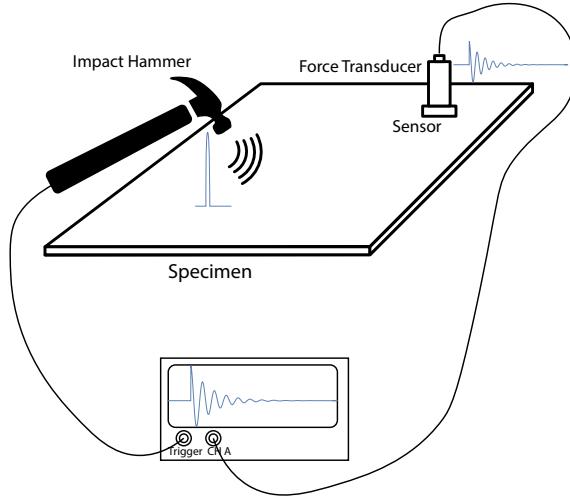


Figure 1.16: A measurement setup to record the impulse response of a mechanical system.

short and exponentially decaying impulse response.

This should help you understand why we needed to flip the impulse response in the convolution operation (or equivalently the input function). Notice first that the impulse function is necessarily zero until time zero. This is because the system is causal and there can be no output if there's no input. Next, notice that by flipping the impulse response and taking a sliding weighted sum with the input, we're including the contribution of past inputs on the current input. If the impulse response is short and decaying exponentially, then only a small amount of past inputs influence the current input. On the other hand, if the impulse response is longer, then the system remembers the history of the input for a longer duration and we must include those terms to figure out the current output.

The Delta Function "Hammer": Mechanical Analogy

Another way to look at a delta function is to think of it like hitting your system with a hammer (see *Fig. 1.16*). It is an impulsive input that excites the system. After such an excitation, the system will "vibrate" in its natural modes and these vibrations will eventually die out. The time that the vibrations last is related to the depth of the system's memory. The vibrations are related to the fact that the system can store kinetic and potential energy and these modes complement each other. Loss in the system causes this stored energy to eventually dissipate. In fact, one way to measure the impulse response is exactly with the setup shown. The hammer is implemented an electromechanical solenoid, and the actuation signal is fed to the trigger of the oscilloscope (or other recording device). The vibrations are picked up by a force transducer and carefully recorded, yielding the impulse response.

1.9 Step and Pulse Response

Consider a system with an impulse response shown in *Fig. 1.17*. This is in fact the impulse response function for a low-pass filter:⁵

$$h(t) = u(t) \frac{1}{\tau} e^{-t/\tau} \quad (1.77)$$

Notice that the system has a time constant $\tau = RC$ that corresponds to the amount of time the system responds to an impulse input ($\tau = 1/3$). Here $u(t)$ is the step function (zero until time zero, and then unity thereafter). As discussed earlier, the impulse response is necessarily zero until time

⁵See tables 1.21 and 1.22 of transforms later in this lecture.

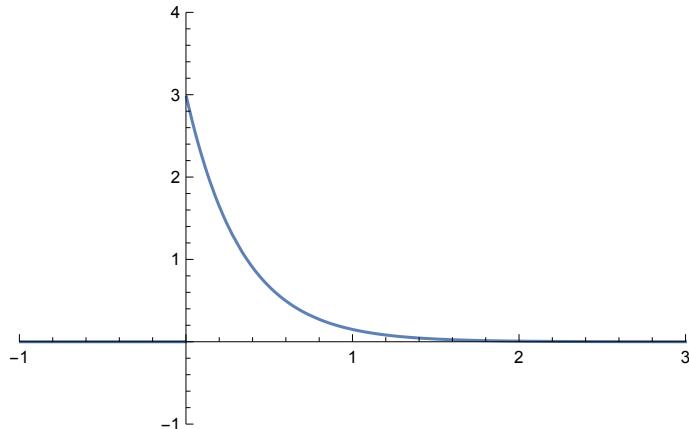


Figure 1.17: The impulse response for the low-pass filtered analyzed earlier in the frequency domain.

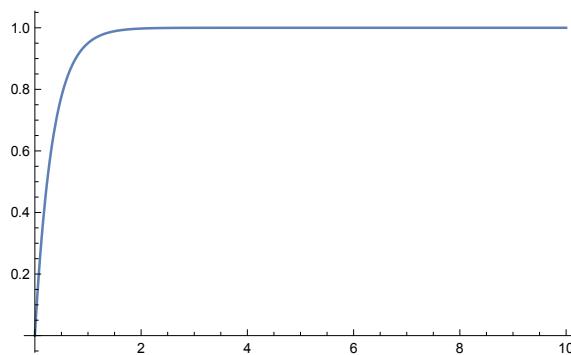


Figure 1.18: The step response of a low-pass filter.

zero, and the $u(t)$ enforces this condition. If you now imagine an arbitrary input, each time point is a weighted sum of the current input and past inputs, with an exponentially decaying amplitude for past inputs. Only the recent $\sim \tau$ inputs really make an impact, and earlier inputs are "forgotten".

1.9.1 Step Response

A very interesting input is the "**step input**", or application of $u(t)$ to our system. From what we know, we can calculate the response to a step by using a convolution:

$$y(t) = h(t) * u(t) = (1 - e^{-3t}) u(t) \quad (1.78)$$

The response is plotted in *Fig. 1.18*. The step has a very sharp transition at the input, but at the output it's smoothed out because each output point is an average of the current input and past inputs. Since at the moment of the transition, all past inputs are zero, the output cannot track the input instantaneously.

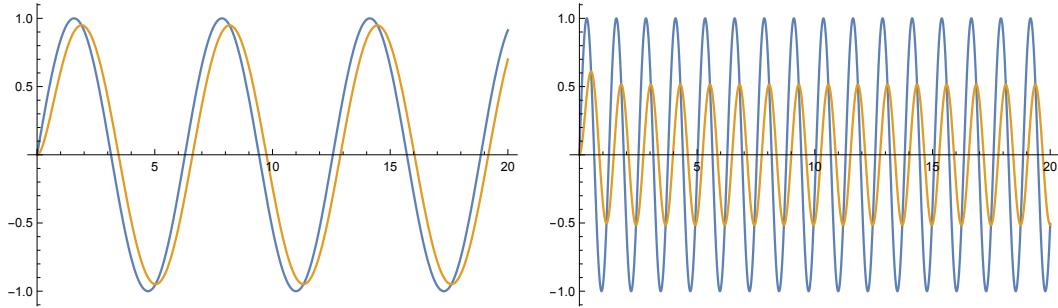


Figure 1.19: The response of a "slow" and "fast" sinusoid. Slow means the period of oscillation is smaller than the filter characteristic time constant τ . Fast means the opposite.

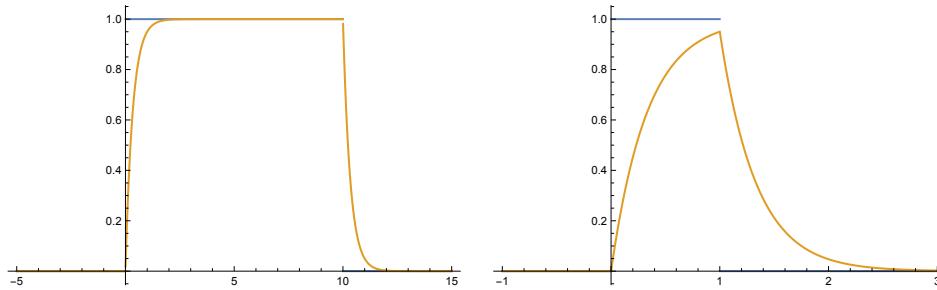


Figure 1.20: The **pulse response** of a low-pass filter. Note that the shorter pulse has been expanded for clarity (see the time scale).

1.9.2 Sinusoidal Response

As shown in *Fig. 1.19*, if the input changes slower than the time scale of τ , or if the frequency of the input is lower than roughly $1/\tau$, then the output is a smoothed version of the input (each time point is blurred by the width of the impulse response). Fast transitions are smoothed out and don't appear at the output. If the input changes very quickly, several cycles of the input are averaged out and produce only a small output. This is how the filter works as high frequency signals are attenuated by the low-pass filter. If the oscillation frequency is very slow with respect to τ , the impulse response "looks like" a delta function with no memory and just passes the current input to the output.

1.9.3 Pulse Response

Now let's see what happens to a pulse. By superposition, this is just two unit-step responses, so we can easily reconstruct the response as shown in *Fig. 1.20*. If the pulse is wide, then the output has enough time to settle and the output pulse is a faithful representation of the input. If the pulse width is shorter than the time-constant of the filter, there is insufficient time for the output to settle, and the pulse is distorted. In high speed communication systems, we need to ensure the system has enough **bandwidth** to allow the highest frequency pulses to get through.

1.10 Frequency Domain Characterization of Linear Systems

We started this chapter by discussion of the steady-state sinusoidal response of a linear system. We found that we can use complex exponential inputs to quickly and efficiently calculate the output, and by sweeping the frequency, we obtain the **spectral response** of the system $H(s)$ (the complex eigenvalues) that completely characterizes the system. Next we find that in the time domain, the impulse response is a way to "probe" a linear system to determine its behavior. In other words

both $H(s)$ and $h(t)$ are very special responses that characterizes the system. One comes from the complex exponential input e^{st} and the other from the application of a delta function. How are these functions $H(s)$ and $h(t)$ related?

1.10.1 Relation to Complex Exponential

Recall that we found that the complex exponential is an eigenfunction for our linear system:

$$\mathbf{L}[e^{st}] = H(s)e^{st} \quad (1.79)$$

But that means that we can apply the convolution integral to the input e^{st} and equate the two outputs:

$$\mathbf{L}[e^{st}] = H(s)e^{st} = \int_{-\infty}^{\infty} h(t-\tau)e^{s\tau}d\tau \quad (1.80)$$

A simple manipulation shows that:

$$\begin{aligned} H(s) &= \int h(t-\tau)e^{s\tau}e^{-st}d\tau \\ &= \int h(t-\tau)e^{-s(t-\tau)}d\tau \\ &= \int_{\infty}^{-\infty} h(x)e^{-sx}(-dx) \end{aligned} \quad \text{Let } x = t - \tau$$

Thus, we have:

$$H(s) = \int_{-\infty}^{\infty} h(x)e^{-sx}dx \quad (1.81)$$

This equation shows us that $H(s)$ and $h(t)$ are indeed related in a very special way.

Laplace Transform

We derived a very important relation between the transfer function and the impulse response which is known as the **Laplace Transform** \mathcal{L} :

$$H(s) = \int_{-\infty}^{\infty} h(t)e^{-st}dt = \mathcal{L}\{h(t)\} \quad (1.82)$$

This can be interpreted as transforming a **time-domain** function to the "**s-domain**", which is the general complex **frequency-domain**. In this book we will not use Laplace Transforms explicitly, instead we will mostly use the transfer function in the frequency domain: $H(j\omega)$, which is nothing but evaluating the transfer function $H(s)$ for $s = j\omega$. Even though we won't make extensive use of the Laplace Transform, we nevertheless spent some time to highlight these important concepts and relations because they show a deep connection between the frequency domain response $H(j\omega)$ and the time domain response.

1.10.2 De-convolving the Convolution

In fact, one of the most important relations is how convolution in time translates to the frequency domain. Consider the (complicated) convolution operator and take the results into the Laplace domain:

$$Y(s) = \mathcal{L}\{x(t) * h(t)\} = \mathcal{L}\left\{\int h(t-\tau)x(\tau)d\tau\right\} \quad (1.83)$$

Use the definition of the Laplace Transform

$$Y(s) = \int \left(\int h(t - \tau)x(\tau)d\tau \right) e^{-st} dt \quad (1.84)$$

Notice that if we make a simple change of notation and call $y = t - \tau$ the new dummy variable, we have:

$$\begin{aligned} Y(s) &= \int \int h(\underbrace{t - \tau}_y) x(\tau) e^{-st} d\tau dt \\ &= \int \int h(y) e^{-s(y+\tau)} x(\tau) dy d\tau \quad t = y + \tau \end{aligned}$$

Each integral is only a function of one variable, so we can separate them out:

$$Y(s) = \left(\int h(y) e^{-sy} dy \right) \left(\int x(\tau) e^{-s\tau} d\tau \right) \quad (1.85)$$

Or more simply, the convolution is a (simple) product in the Laplace domain:

$$Y(s) = H(s)X(s) \quad (1.86)$$

This is a very important result. Not only is it more efficient to work directly in the **Laplace domain** (or frequency domain) when computing the response of a linear system, but it is also much easier to visualize what's going on. A filter can be visualized very simply in the frequency domain. For an arbitrary input, if we compute its Laplace (or Fourier, see below) Transform, we can find the output by simply multiplying the two and doing an inverse transform. In many cases, we don't even need to do the inverse transform, which is not a very simple calculation to carry out. We simply use the concept of frequency domain to better understand our system.

Laplace Transform Properties

The properties of the Laplace Transform are summarized in *Fig. 1.21*. Most are easily derived and are provided here just for reference. We just proved the most important property related to the convolution.

Laplace Transform Table

From Wikipedia⁶, the table in *Fig. 1.22* shows some common Laplace Transform "pairs", or the s-domain representation for some common functions. This table is important because in practice we hardly ever evaluate the Laplace Transform integral, especially in the reverse direction. Instead, the lookup table is used for common functions and special techniques, such as partial fraction expansions, can be used to put a rational function into standard form. We showed earlier that most linear systems do in fact generate a rational function transfer function.

⁶https://en.wikipedia.org/wiki/Laplace_transform

	Time domain	s domain	Comment
Linearity	$af(t) + bg(t)$	$aF(s) + bG(s)$	Can be proved using basic rules of integration.
Frequency-domain derivative	$tf(t)$	$-F'(s)$	F' is the first derivative of F with respect to s .
Frequency-domain general derivative	$t^n f(t)$	$(-1)^n F^{(n)}(s)$	More general form, n th derivative of $F(s)$.
Derivative	$f'(t)$	$sF(s) - f(0^+)$	f is assumed to be a differentiable function , and its derivative is assumed to be of exponential type. This can then be obtained by integration by parts
Second derivative	$f''(t)$	$s^2 F(s) - sf(0^+) - f'(0^+)$	f is assumed twice differentiable and the second derivative to be of exponential type. Follows by applying the Differentiation property to $f'(t)$.
General derivative	$f^{(n)}(t)$	$s^n F(s) - \sum_{k=1}^n s^{n-k} f^{(k-1)}(0^+)$	f is assumed to be n -times differentiable, with n th derivative of exponential type. Follows by mathematical induction .
Frequency-domain integration	$\frac{1}{t} f(t)$	$\int_s^\infty F(\sigma) d\sigma$	This is deduced using the nature of frequency differentiation and conditional convergence.
Time-domain integration	$\int_0^t f(\tau) d\tau = (u * f)(t)$	$\frac{1}{s} F(s)$	$u(t)$ is the Heaviside step function and $(u * f)(t)$ is the convolution of $u(t)$ and $f(t)$.
Frequency shifting	$e^{at} f(t)$	$F(s-a)$	
Time shifting	$f(t-a)u(t-a)$	$e^{-as} F(s)$	$u(t)$ is the Heaviside step function
Time scaling	$f(at)$	$\frac{1}{a} F\left(\frac{s}{a}\right)$	$a > 0$
Multiplication	$f(t)g(t)$	$\frac{1}{2\pi i} \lim_{T \rightarrow \infty} \int_{c-iT}^{c+iT} F(\sigma) G(s-\sigma) d\sigma$	The integration is done along the vertical line $\text{Re}(\sigma) = c$ that lies entirely within the region of convergence of F . ^[21]
Convolution	$(f * g)(t) = \int_0^t f(\tau)g(t-\tau) d\tau$	$F(s) \cdot G(s)$	
Complex conjugation	$f^*(t)$	$F^*(s^*)$	
Cross-correlation	$f(t) * g(t)$	$F^*(-s^*) \cdot G(s)$	
Periodic function	$f(t)$	$\frac{1}{1 - e^{-Ts}} \int_0^T e^{-st} f(t) dt$	$f(t)$ is a periodic function of period T so that $f(t) = f(t+T)$, for all $t \geq 0$. This is the result of the time shifting property and the geometric series .

Figure 1.21: Table of Laplace Transform properties from the Wikipedia site.

Function	Time domain $f(t) = \mathcal{L}^{-1}\{F(s)\}$	Laplace s-domain $F(s) = \mathcal{L}\{f(t)\}$	Region of convergence	Reference
unit impulse	$\delta(t)$	1	all s	inspection
delayed impulse	$\delta(t-\tau)$	$e^{-\tau s}$		time shift of unit impulse
unit step	$u(t)$	$\frac{1}{s}$	$\text{Re}(s) > 0$	integrate unit impulse
delayed unit step	$u(t-\tau)$	$\frac{1}{s} e^{-\tau s}$	$\text{Re}(s) > 0$	time shift of unit step
ramp	$t \cdot u(t)$	$\frac{1}{s^2}$	$\text{Re}(s) > 0$	integrate unit impulse twice
n th power (for integer n)	$t^n \cdot u(t)$	$\frac{n!}{s^{n+1}}$	$\text{Re}(s) > 0$ ($n > -1$)	Integrate unit step n times
q th power (for complex q)	$t^q \cdot u(t)$	$\frac{\Gamma(q+1)}{s^{q+1}}$	$\text{Re}(s) > 0$ $\text{Re}(q) > -1$	[27][28]
n th root	$\sqrt[n]{t} \cdot u(t)$	$\frac{1}{s^{\frac{1}{n}+1}} \Gamma\left(\frac{1}{n} + 1\right)$	$\text{Re}(s) > 0$	Set $q = 1/n$ above.
n th power with frequency shift	$t^n e^{-\alpha t} \cdot u(t)$	$\frac{n!}{(s+\alpha)^{n+1}}$	$\text{Re}(s) > -\alpha$	Integrate unit step, apply frequency shift
delayed n th power with frequency shift	$(t-\tau)^n e^{-\alpha(t-\tau)} \cdot u(t-\tau)$	$\frac{n! \cdot e^{-\tau s}}{(s+\alpha)^{n+1}}$	$\text{Re}(s) > -\alpha$	Integrate unit step, apply frequency shift, apply time shift
exponential decay	$e^{-\alpha t} \cdot u(t)$	$\frac{1}{s+\alpha}$	$\text{Re}(s) > -\alpha$	Frequency shift of unit step
two-sided exponential decay (only for bilateral transform)	$e^{-\alpha t }$	$\frac{2\alpha}{\alpha^2 - s^2}$	$-\alpha < \text{Re}(s) < \alpha$	Frequency shift of unit step
exponential approach	$(1 - e^{-\alpha t}) \cdot u(t)$	$\frac{\alpha}{s(s+\alpha)}$	$\text{Re}(s) > 0$	Unit step minus exponential decay
sine	$\sin(\omega t) \cdot u(t)$	$\frac{\omega}{s^2 + \omega^2}$	$\text{Re}(s) > 0$	Bracewell 1978, p. 227
cosine	$\cos(\omega t) \cdot u(t)$	$\frac{s}{s^2 + \omega^2}$	$\text{Re}(s) > 0$	Bracewell 1978, p. 227

Figure 1.22: Table of Laplace Transform pairs from the Wikipedia site.

1.10.3 Fourier Series and Transform

In previous courses, you may have learned that you can represent a periodic function in time as a **Fourier series**

$$x(t) = \sum_{n=-\infty}^{\infty} c_n e^{j2\pi n f_0 t} \quad (1.87)$$

In this equation $f_0 = 1/T$ is the fundamental frequency of the function of period T . For many practical signals of interest, we can go back and forth between the time or frequency domain, in the sense that the coefficients c_n are a complete representation of the signal (invertible). The **FFT (fast fourier transform)** is an efficient way to compute the Fourier Series. We know that the magnitude of the coefficients c_n determine the amount of power concentrated around a certain frequency. If a function varies gradually in time with a time scale say less than T , then coefficients above $n = f_0 T$ will be smaller in magnitude.

The Fourier Transform is a natural extension of the Fourier Series into the continuous time domain and it works for non-periodic functions (it has a line spectrum for periodic functions)

$$X(\omega) = \mathcal{F}\{x(t)\} = \int_{-\infty}^{\infty} e^{-j\omega t} x(t) dt \quad (1.88)$$

And the inverse transform is given by

$$x(t) = \mathcal{F}^{-1}\{X(\omega)\} = \int_{-\infty}^{\infty} e^{j\omega t} X(f) df \quad (1.89)$$

For most signals, this is a well defined transformation allowing us to think of a signal in either the time domain or frequency domain. Notice that the Fourier Transform is related to the Laplace Transform by

$$\mathcal{F}\{x(t)\} = X(s)|_{s=j\omega} = X(j\omega) \quad (1.90)$$

This brings us full circle and shows that the frequency response $H(j\omega)$, the sinusoidal steady-state response, is in fact related to the Laplace Transform of the impulse response of our system. While $H(s)$ is general, $H(j\omega)$, or the Fourier Transform of the impulse response, is a special case that applies to the periodic steady-state response.

1.10.4 Frequency Domain Interpretation

If we transform the input function from the time domain to the frequency domain, then we can simply say that every frequency component in the input gets multiplied and phase shifted by the corresponding frequency domain component of the transfer function, as shown in Fig. 1.23.

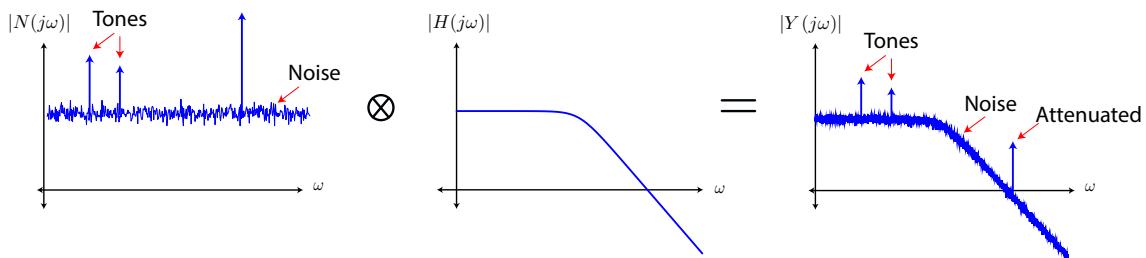


Figure 1.23: The frequency domain representation of a signal plus noise (left). The signal consists of tones (delta-functions in frequency domain) and noise (flat spectrum). The signal is low-pass filtered, which is equivalent to multiplication by the frequency domain transfer function shown in the middle, resulting in the filtered signal and reduction in noise (right).

This follows from the properties of convolution in time becoming multiplication in the frequency domain. Understanding a filter is much, much easier when it is analyzed in the frequency domain. In this example we see a frequency domain representation of a signal with three tones, shown as Delta functions. They are delta functions because the input signal consists of pure tones with a very long duration, and so all the energy is concentrated in these two frequency bands. We also show that the signal is accompanied by "white noise", shown as a flat spectrum signal. Although we have not proven this yet, it can be shown that noise is a wideband signal and can be modeled as a flat signal in the frequency domain.

If we low-pass filter the signal and noise, we expect that we will reduce the contribution of the noise. In this example we suppose that the third tone is undesired, so we select the corner frequency of the filter just beyond the second tone, so that we do not attenuate the desired tones. After filtering, we would expect that the overall noise of the signal can be reduced without affecting the signals of interest. This is shown by simply multiplying the frequency domain representation of the signal and noise with the filter response. We see the noise contribution beyond the corner is reduced dramatically, and the third tone is also attenuated.

Audio Example

In this example, the input signal consists of a pure tone (261.626 Hz) and its third harmonic, which is delayed with respect to the fundamental. Also, the signal is corrupted by a lot of noise content. The signals of interest are shown in *Fig. 1.24*. Now consider the signal + noise with noise power $10\times$ lower (Signal-Noise Ratio (SNR) = 10 dB), as shown in *Fig. 1.25*.

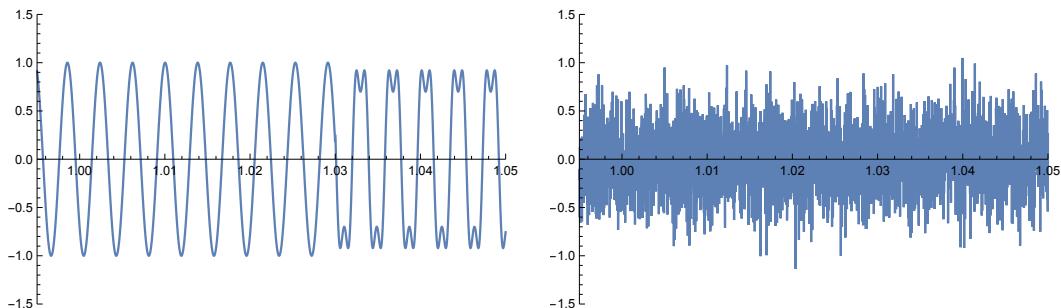


Figure 1.24: A audio signal consisting of a tone and the third harmonic (delayed by 1 second). The signal will be added to the noisy signal shown on the right.

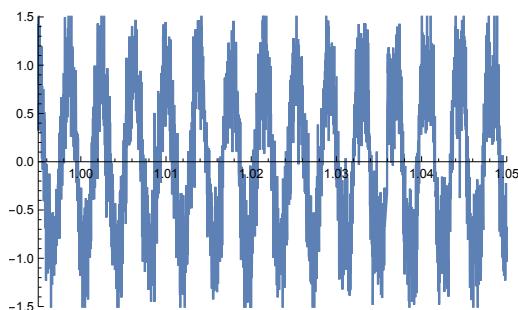


Figure 1.25: The noisy signal is shown. The noise impacts both the amplitude and zero crossings of the signal.

Click the following links to listen to the signals:

- [Tone + Third Harmonic, no noise](#)
- [Signal + Noise](#)
- [Just the "white" noise](#)

Now if we filter the signal + noise, we can reduce the noise without impacting the signal (since the signal is mostly at low frequency in this case):

- [Low-pass filtered signal + noise](#)

This is used extensively in practice to improve the signal quality. To improve the signal quality further, we build a **band-pass filter** to pick out just the tone of interest. The first filter picks up the first tone, and eliminates most of the noise.

- [Band-pass filtered centered around first tone](#)

Now we apply a bandpass filter with a wider passband, allowing both tones to pass. We pickup both tones, but of course the downside is we hear more noise.

- [Wideband filtered signal + noise](#)

1.10.5 But Most Systems are Non-Linear ...?

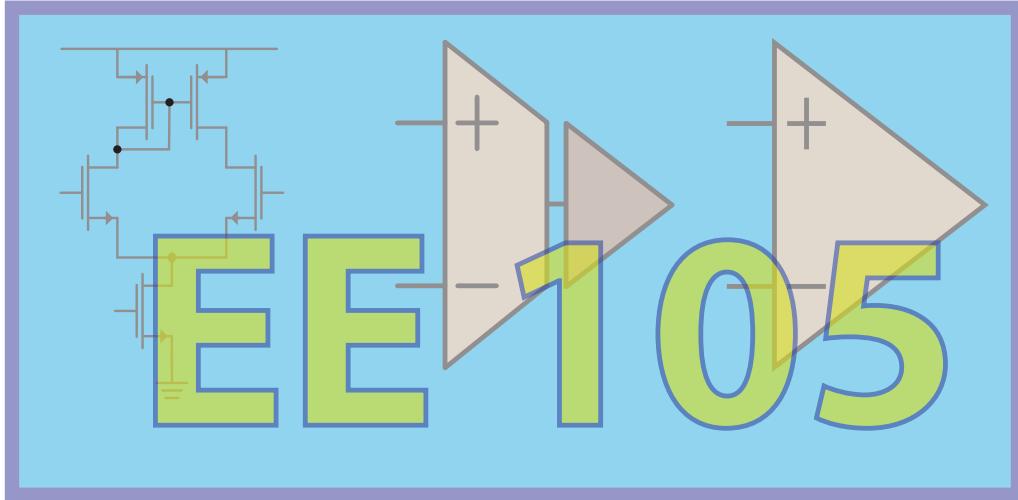
As we have demonstrated throughout this chapter, linear system theory is powerful and very useful for design and analysis. But you may object and say that most real systems are non-linear, so why is this stuff useful?

When systems are non-linear, we often "linearize" them about an operating point, and use linear theory to understand the response to "small" perturbations. We will be doing this extensively in this course. Soon you will find that the silicon transistor and diodes are non-linear, so we will build "small-signal" models that are linear. This is useful as long as the signals of interest don't deviate too much from the operating point. If they do, we need non-linear techniques. Even in these situations, we can gain useful insights from the small-signal approximation, and it is the starting point for design and analysis.

1.11 Chapter Summary

This chapter has taken us from the frequency domain AC style analysis to the time domain impulse response, and then all the way back full circle to the frequency domain transfer function.

Frequency response allows us to completely characterize a system using the sinusoidal response. More generally, the frequency response is characterized by the complex exponential response $H(s)$. The impulse response $h(t)$ completely characterizes a system in the time-domain, and we found a connection between the two approaches. The transfer function $H(s)$ is the Laplace Transform of the impulse response $h(t)$. The Fourier Transform is the Laplace transform evaluated on the imaginary axis, and AC circuit theory is how we calculate (or measure) the transfer function in practice. Convolution in time simplifies to multiplication in the frequency domain (and vice versa). Many concepts as calculations are easier to carry out in the frequency domain (filters), and as an engineer your intuition for the frequency domain will increase throughout your career.



2. AC Circuits and AC Analysis

2.1 Chapter Preview

This chapter builds on the previous by focusing solely on AC circuits, or the circuit response in sinusoidal steady state. We will also focus on circuits in particular and derive an even simpler procedure for analysis based on "phasors", which is just a shorthand way to do the complex exponential response that we discussed in the previous chapter. We formally introduce the concept of the AC transfer functions in circuits as ratios of signals, either voltage or current, in the frequency domain. We will see that for circuits of interest, we can completely characterize the transfer function in terms of a finite number of "poles" and "zeros". We will also discuss the Bode Plot technique for estimating the transfer function and touch briefly on AC power.

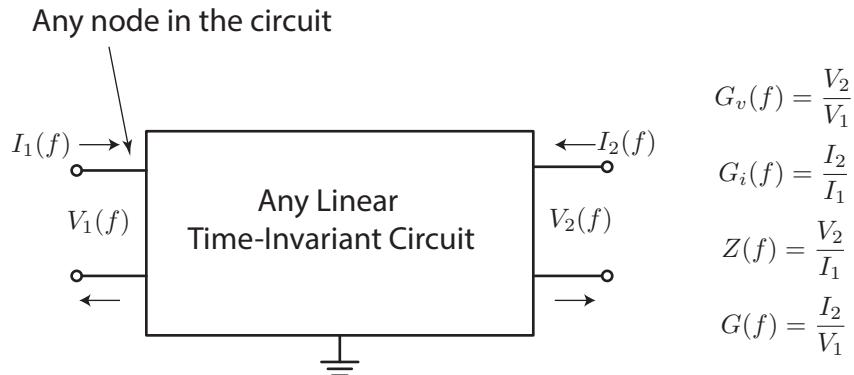


Figure 2.1: Between two ports in a linear time-invariant (LTI) circuit, we can define four transfer functions. In this chapter we will focus on the frequency domain transfer functions.

2.2 Transfer Functions

2.2.1 Transfer Function Concept

In many situations, we are interested in the transfer of voltage or current from one terminal pair, or port, to another port in the circuit. As shown in the black-box *Fig. 2.1*, we can identify these ports by explicitly labeling them and clearly identifying the voltages across and the currents into and out of these ports. Note that some ports may have terminals in common, such as a common-ground for ground referenced signals.

The key is that we drive a pair of nodes with a sinusoidal current or voltage – *the input*, and observe the output across a load at another pair of terminals – *the output*. The ratio is the transfer function. For a fixed frequency, it's just a complex number. From the last chapter, we know it's the Laplace transform of the impulse response evaluated on the $j\omega$ axis.

Example 1 – ECG Amplifier

The **electrocardiogram** (ECG) is a measurement of the voltage between two points on the body. For example, as shown in *Fig. 2.2*, the voltage across the chest measured from the "right arm" (RA) to the "left arm" (LA) is what's known as lead I. The ground reference is often the "right leg" (RL). This voltage is the "input voltage" and we need to amplify it and reject any **common-mode** noise (see Section 16.6), especially 60 Hz or 50 Hz AC line voltage, which is often orders of magnitude larger than the desired signal. The output voltage may be single-ended and referenced to a different ground, as shown in the figure. In this scenario, we define the transfer function as:

$$G_v(f) = \frac{V_{out}}{V_1 - V_2} \quad (2.1)$$

The variation in frequency is due to the frequency content of the heartbeat signal and possibly filters applied to the signal intentionally to reduce the impact of noise.

Example 2 – Mic/Speaker

The next example is deceptively simple because on first inspection of *Fig. 2.3(a)*, you may conclude that there's an error in the drawing, because the negative input of the inverting amplifier is shorted due to the "virtual ground" connection. In fact, were it not for the microphone source resistance, as shown in *Fig. 2.3(b)*, this would be the case. But the signal is amplified and filtered by the presence of the capacitor in the inverting amplifier configuration.

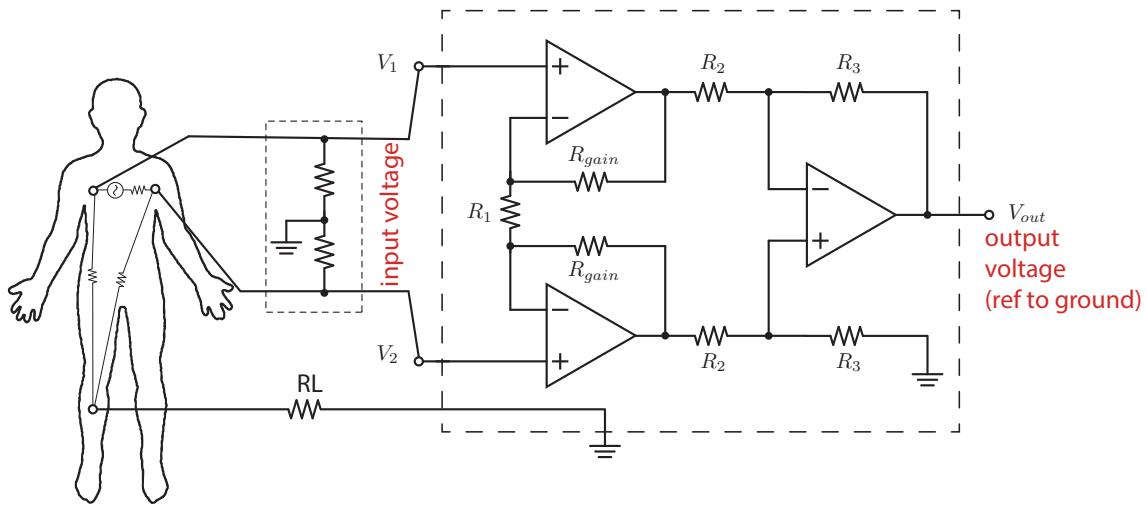


Figure 2.2: An ECG amplifier is an example of a transfer function from a non-ground referenced port to an output port that is referenced to ground, or a differential to single-ended transfer function.

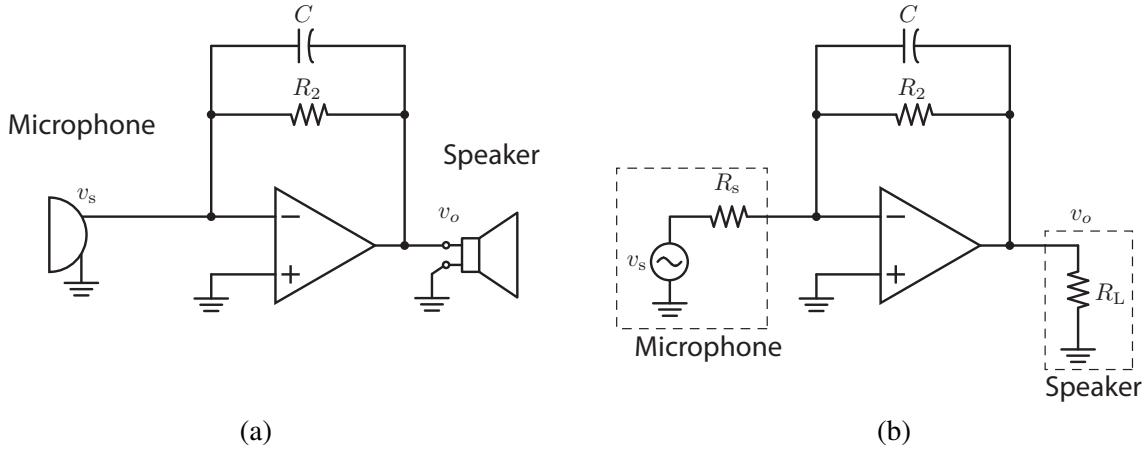


Figure 2.3: The transfer function from a microphone to the output of the amplifier is actually a current to voltage transfer function. Note the voltage to voltage transfer function is undefined for an ideal op-amp since the negative input terminal is a virtual ground.

Low frequency signals will see a gain of:

$$G_v(f) = \frac{v_o}{v_s} = -\frac{R_2}{R_s} \quad (2.2)$$

On the other hand, very high frequency signals will see a very small gain, because as the capacitor becomes more conductive than the resistor R_2 , or in other words $\omega C \gg R_2$, then the gain is simply:

$$G_v(f) \sim -\frac{\frac{1}{\omega C}}{R_2} = \frac{-1}{\omega C R_2} \quad (2.3)$$

In this example, the presence of the source impedance changed everything. In general we must be careful to include the potential impact of the source/load impedances.

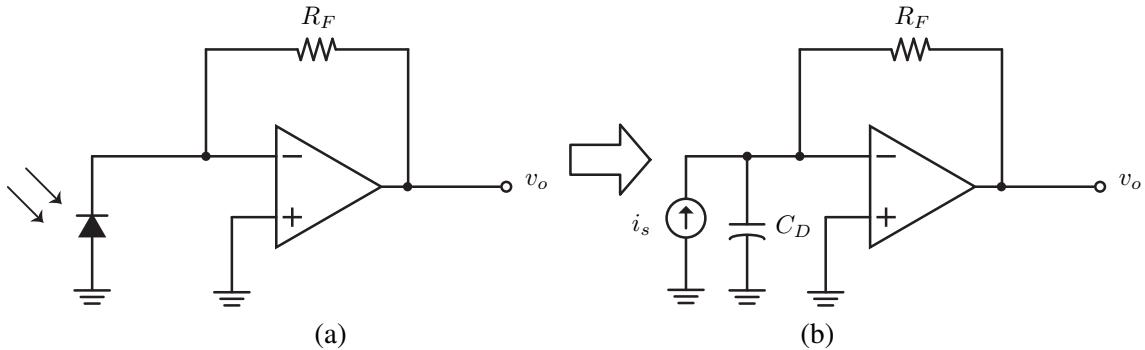


Figure 2.4: (a) A photodetector is used to convert photons to current, which are then converted into a voltage through the transimpedance amplifier (TIA). (b) The equivalent circuit model.

Example 3 – Photodetector

For many applications, we need to process a voltage, so a "transimpedance" amplifier is needed to convert the current to the voltage through R_F as shown.¹

Shown in Fig. 2.4(a), the source is a **photodiode**, a device that has an output current proportional to the number of impinging photons. Again, upon first inspection it seems that the diode is "shorted" to ground through the virtual ground connection, but the diode is better modeled as a current source. The input current is therefore the signal of interest, as shown in Fig. 2.4(b). It is important to include the **source capacitance** (C_D), which limits the high frequency performance. One benefit of the amplifier is that the signal across C_D is mostly shorted, and therefore the impact is not as severe as it would be otherwise.

2.2.2 Voltage and Current Gain

The voltage (current) gain is just the voltage (current) transfer function from one port to another port (see *Fig. 2.5*):

$$G_v(\omega) = \frac{V_2}{V_1} = \left| \frac{V_2}{V_1} \right| e^{j(\phi_2 - \phi_1)} \quad (2.4)$$

$$G_i(\omega) = \frac{I_2}{I_1} = \left| \frac{I_2}{I_1} \right| e^{j(\varphi_2 - \varphi_1)} \quad (2.5)$$

If $G > 1$, the circuit has voltage (current) gain. If $G < 1$, the circuit has loss or attenuation. It's a common misconception that only "active" circuits can have gain. We will demonstrate in *section 2.5.3* that *RCL* circuits in particular are passive, and can provide voltage and/or current gain. Another common example is the ubiquitous **transformer**, which provides gain through mutual coupling. It may surprise you that even *RC* circuits can provide gain, although not as easily as with inductors and transformers!

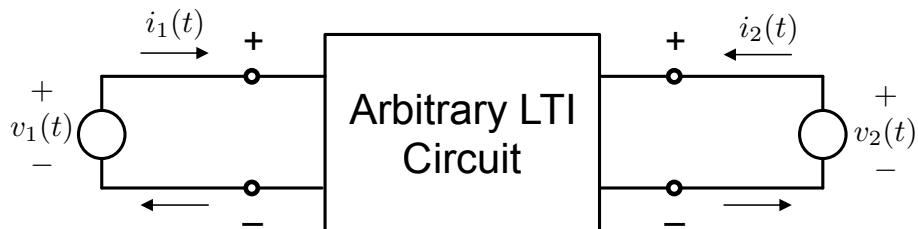
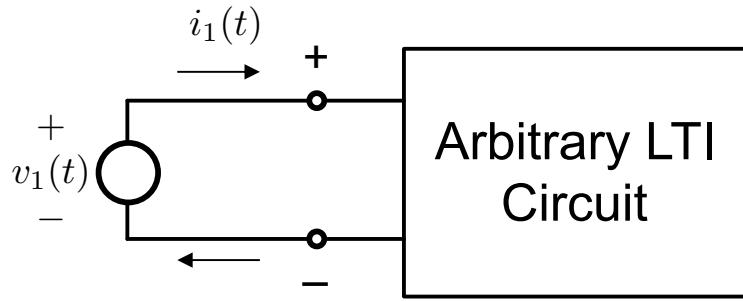


Figure 2.5: An arbitrary **two-port** consisting of a linear time-invariant circuit.

¹Why not simply use a resistor in parallel with the diode to convert the current into a voltage?

Figure 2.6: An arbitrary **one-port** linear time-invariant circuit.

2.2.3 Trans-Impedance and Trans-Admittance

Current/voltage gain are unitless quantities, but we can define a signal as a voltage or a current, or both (or even a power, or the product). If the transfer involves voltage to current or vice versa, we must be careful to define the units. We have a **trans-impedance** (TIA) amplifier gain

$$J(\omega) = \frac{V_2}{I_1} = \left| \frac{V_2}{I_1} \right| e^{j(\varphi_2 - \varphi_1)} [\Omega] \quad (2.6)$$

Less commonly, we have a **trans-admittance** (TAA) amplifier:

$$K(\omega) = \frac{I_2}{V_1} = \left| \frac{I_2}{V_1} \right| e^{j(\varphi_2 - \varphi_1)} [S] \quad (2.7)$$

2.2.4 Impede the Currents!

Suppose that the "input" is defined as the voltage of a terminal pair (*port*) and the "output" is defined as the current into the port (see Fig. 2.6).

$$v(t) = V e^{j\omega t} = |V| e^{j(\omega t + \varphi_v)} \quad (2.8)$$

$$i(t) = I e^{j\omega t} = |I| e^{j(\omega t + \varphi_i)} \quad (2.9)$$

The **impedance** Z is defined as the ratio of the phasor voltage to phasor current ("self" transfer function):

$$Z(\omega) = H(\omega) = \frac{V}{I} = \left| \frac{V}{I} \right| e^{j(\varphi_v - \varphi_i)} \quad (2.10)$$

We are of course familiar with the concept of impedance Z , but even though we may not think about it as a transfer function, it is indeed a transfer function.

2.2.5 Admit the Currents!

Similarly, suppose that the "input" is defined as the current of a terminal pair (port) and the "output" is defined as the voltage into the port. The **admittance** Y is defined as the ratio of the current to voltage ("self" transfer function)

$$Y(\omega) = H(\omega) = \frac{I}{V} = \left| \frac{I}{V} \right| e^{j(\varphi_i - \varphi_v)} \quad (2.11)$$

2.3 Transfer Function Poles/Zeros

2.3.1 Complex Transfer Function

Following our derivation from the last chapter, we excite a system with an input voltage (current) x and define the output voltage y (current) to be any node voltage (branch current). For a complex exponential input, the "transfer function" from input to output is given by:

$$H \equiv \frac{y}{x} = \left(a + b_1 j\omega + b_2 (j\omega)^2 + \dots + \frac{c_1}{j\omega} + \frac{c_2}{(j\omega)^2} + \dots \right) \quad (2.12)$$

We can write this in **canonical form** as a rational function:

$$H(\omega) = \frac{n_1 + n_2 j\omega + n_3 (j\omega)^2 + \dots}{d_1 + d_2 j\omega + d_3 (j\omega)^2 + \dots} \quad (2.13)$$

This form of the equation shows that the numerator and denominator are in general polynomial equations, and polynomial equations are characterized by either their coefficients or more importantly their roots.

“s” Complex Plane

You may hear people talking about transfer functions as a function of complex “s” rather than frequency

$$H(s) = \frac{(z_1 - s)(z_2 - s)\dots}{(p_1 - s)(p_2 - s)\dots} \quad (2.14)$$

As we learned, this is a generalization (Laplace domain) of frequency. When doing AC analysis, we are deriving the function along the imaginary axis (sinusoidal steady-state response)

$$H(s = j\omega) = \frac{(z_1 - j\omega)(z_2 - j\omega)\dots}{(p_1 - j\omega)(p_2 - j\omega)\dots} \quad (2.15)$$

This is why you may see introduce this notation: $H(j\omega)$

2.3.2 Poles and Zeros

For most circuits that we deal with, the transfer function can be shown to be a **rational function**:

$$H(\omega) = \frac{n_1 + n_2 j\omega + n_3 (j\omega)^2 + \dots}{d_1 + d_2 j\omega + d_3 (j\omega)^2 + \dots} \quad (2.16)$$

The behavior of the circuit can be extracted by putting the transfer function into **standard form**, and finding the roots of the numerator and denominator:

$$H(\omega) = \frac{(z_1 - j\omega)(z_2 - j\omega)\dots}{(p_1 - j\omega)(p_2 - j\omega)\dots} = \frac{\prod(z_i - j\omega)}{\prod(p_i - j\omega)} \quad (2.17)$$

Or in another form (**DC gain explicit**):

$$H(\omega) = G_0(j\omega)^K \frac{(1 - j\omega\tau_{z1})(1 - j\omega\tau_{z2})\dots}{(1 - j\omega\tau_{p2})(1 - j\omega\tau_{p2})\dots} \quad (2.18)$$

$$= G_0(j\omega)^K \frac{\prod(1 - j\omega\tau_{z,i})}{\prod(1 - j\omega\tau_{p,i})} \quad (2.19)$$

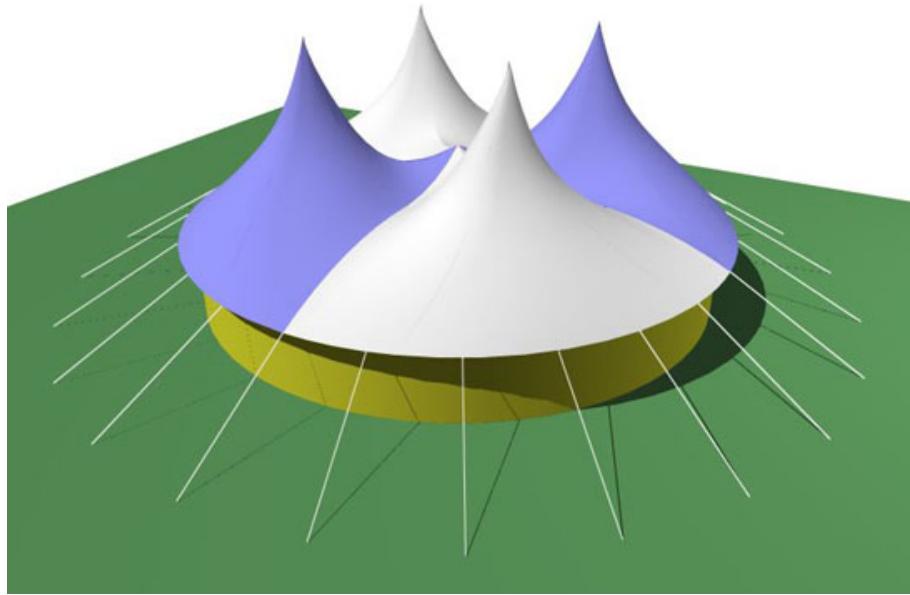


Figure 2.7: A tent has poles and "zeros", or locations of stakes. At the stakes the elevation is at ground (zero), whereas at the poles the elevation hits a local maximum.

Building Tents: Poles and Zeros

The roots of the numerator are called the “**zeros**” since at these frequencies, the transfer function is zero. These are the “stakes in the ground” of our tent, see *Fig. 2.7*. The roots of the denominator are called the “**poles**”, since at these frequencies the transfer function peaks (like a pole in a tent). While the frequency response requires us to keep track of the transfer function over a continuous frequency spectrum, this perspective gives us a new insight. We actually can completely characterize a system by a finite number of poles and zeros. Come to think of it, this should be obvious because our circuit is built out of a finite number of components (L , C , M , and dependent sources). So the number of poles and zeros should be related to the number of elements with “memory”, such as inductors and capacitors.

2.4 AC Circuits Review

2.4.1 Phasors

With our new confidence in complex numbers, we can go full steam ahead and work with them directly. We can even drop the time factor, $e^{j\omega t}$, since it will cancel out of the equations. In other words, the part that matters is the amplitude and phase shift of the resulting complex exponential, not the complex exponential itself. Let’s define this part as the “**phasor**”. We excite system with a phasor: $\hat{V}_1 = V_1 e^{j\varphi_1}$ and the response will also be phasor: $\hat{V}_2 = V_2 e^{j\varphi_2}$.

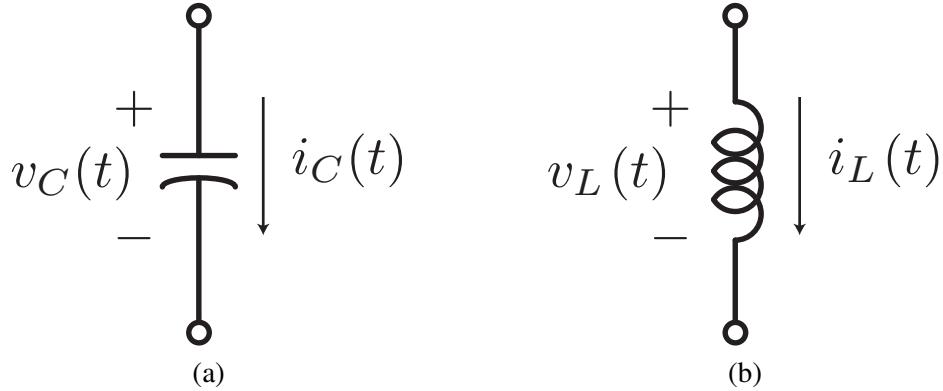


Figure 2.8: (a) Capacitor and (b) inductor I - V relations in the time domain.

2.4.2 Capacitor I-V Phasor Relation

To find the phasor relation for current and voltage for a capacitor, begin with the " $I - V$ " relation in the time domain (see *Fig. 2.8*):

$$i_c(t) = C \frac{dv_C(t)}{dt} \quad (2.20)$$

Now for sinusoidal steady state we substitute:

$$i_c(t) = I_c e^{j\omega t} \quad (2.21)$$

and,

$$v_c(t) = V_c e^{j\omega t} \quad (2.22)$$

This leads to:

$$I_c e^{j\omega t} = C \frac{d}{dt} [V_c e^{j\omega t}] \quad (2.23)$$

or,

$$CV_c \frac{d}{dt} e^{j\omega t} = j\omega CV_c e^{j\omega t} \quad (2.24)$$

Let's eliminate the common $e^{j\omega t}$ factor:

$$I_c e^{j\omega t} = j\omega C V_c e^{j\omega t} \quad (2.25)$$

Or more simply, the phasor relation:

$$I_c = j\omega CV_c \quad (2.26)$$

This is "**Ohm's law**" in the frequency-domain for a capacitor.

2.4.3 Inductor I-V Phasor Relation

Let's now find the phasor relation for current and voltage in an inductor using the same procedure:

$$v(t) = L \frac{di(t)}{dt} \quad (2.27)$$

For sinusoidal steady-state, this implies:

$$Ve^{j\omega t} = L \frac{d}{dt}[Ie^{j\omega t}] \quad (2.28)$$

or,

$$LI \frac{d}{dt} e^{j\omega t} = j\omega LI e^{j\omega t} \quad (2.29)$$

Cancelling the common factor:

$$Ve^{j\omega t} = j\omega LI e^{j\omega t} \quad (2.30)$$

This is the phasor relation for an inductor, or "Ohm's Law" in the frequency domain for an inductor:

$$V = j\omega LI \quad (2.31)$$

2.4.4 Direct Calculation of H (no DEs)

To directly calculate the transfer function (impedance, trans-impedance, etc) we can generalize the circuit analysis concept from the “real” domain to the “phasor” domain. We drop the time dependence $e^{j\omega t}$ (common factor that cancels out) and just keep track of amplitude and phase (phasor). With the concept of **impedance (admittance)**, we directly analyze a circuit without explicitly writing down any differential equations. We are free to use KVL, KCL, mesh analysis, loop analysis, or nodal analysis where inductors and capacitors are treated as complex resistors, or more correctly as an impedance (or admittance).

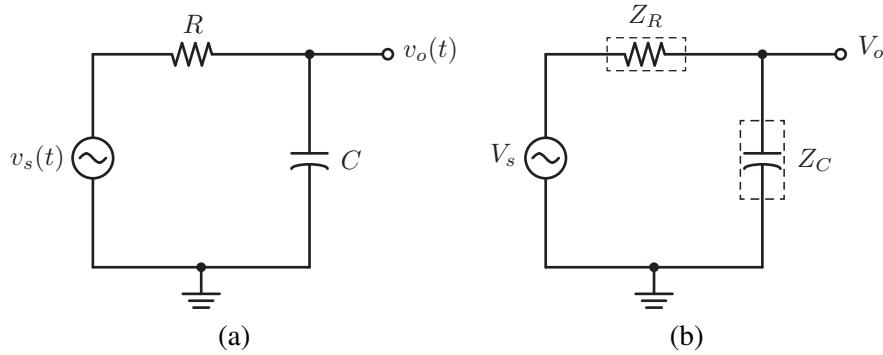


Figure 2.9: (a) The low-pass filter in the time domain and (b) an AC equivalent circuit.

2.5 Example AC Circuit Problems

2.5.1 LPF Example: Again!

In Fig. 2.9, we show a "real" time-domain circuit and the corresponding frequency-domain "phasor" circuit. Instead of setting up the DE in the time-domain, we prefer to do calculations directly in the frequency domain using phasors, treating the capacitor as an imaginary "resistance" or impedance. We know the impedances: $Z_R = R$ and $Z_C = \frac{1}{j\omega C}$.

The fastest way to solve the problem is to say that the LPF is really a **voltage divider**:

$$H(\omega) = \frac{V_o}{V_s} = \frac{Z_C}{Z_C + Z_R} = \frac{\frac{1}{j\omega C}}{R + \frac{1}{j\omega C}} = \frac{1}{1 + j\omega RC} \quad (2.32)$$

2.5.2 Bigger Example (no problem!)

Consider a more complicated example shown in Fig. 2.10. To find v_o , you can do **nodal analysis** or use **network theorems** to simplify the circuit. Let's find the **Thévenin equivalent** from C_2 's perspective. The open-circuit voltage is just the first voltage divider:

$$v_{th} = \frac{Z_{C_1}}{R_1 + Z_{C_1}} v_s \quad (2.33)$$

The impedance looking into the circuit from C_2 is given by:

$$Z_{th} = R_2 + R_1 || Z_{C_1} \quad (2.34)$$

Therefore we have:

$$v_o = \frac{Z_{C_2}}{Z_{C_2} + Z_{th}} V_{th} \quad (2.35)$$

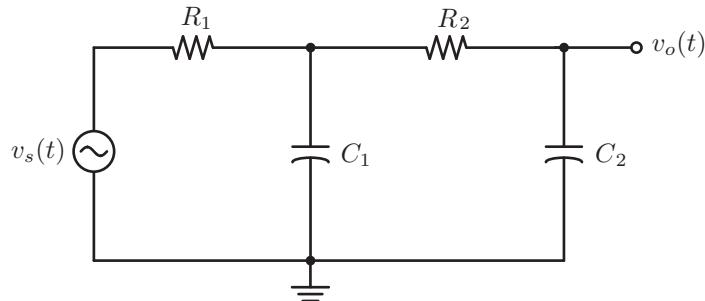
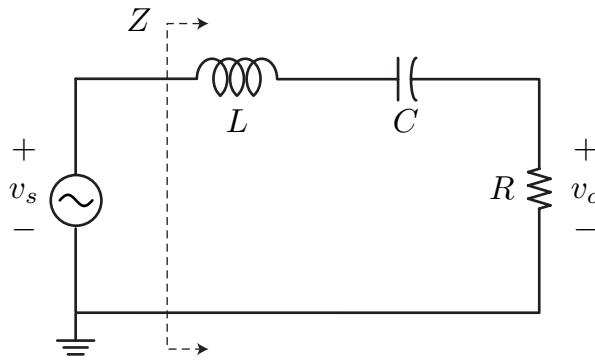


Figure 2.10: A circuit with two resistors and two capacitors can be analyzed quickly using nodal analysis or circuit theorems by treating the capacitors as impedances.

Figure 2.11: A series RLC circuit.

2.5.3 Series RLC Circuits

The RLC circuit shown in *Fig. 2.11* is simple but extremely useful and rich in application. Since the circuit elements are all in series, the **input impedance** seen by the source is simply given by:

$$\begin{aligned} Z &= j\omega L + \frac{1}{j\omega C} + R \\ &= R + j\omega L + \frac{j\omega L}{j^2\omega^2 LC} \\ &= R + j\omega L \left(1 - \frac{1}{\omega^2 LC} \right) \end{aligned}$$

The impedance is purely real at the **resonant frequency**, or when $\Im(Z) = 0$, which happens at a frequency $\omega = \pm \frac{1}{\sqrt{LC}}$. At resonance the impedance takes on a minimal value. See *Fig. 2.12* for phasor diagrams of resonance, including before and after resonance occurs.

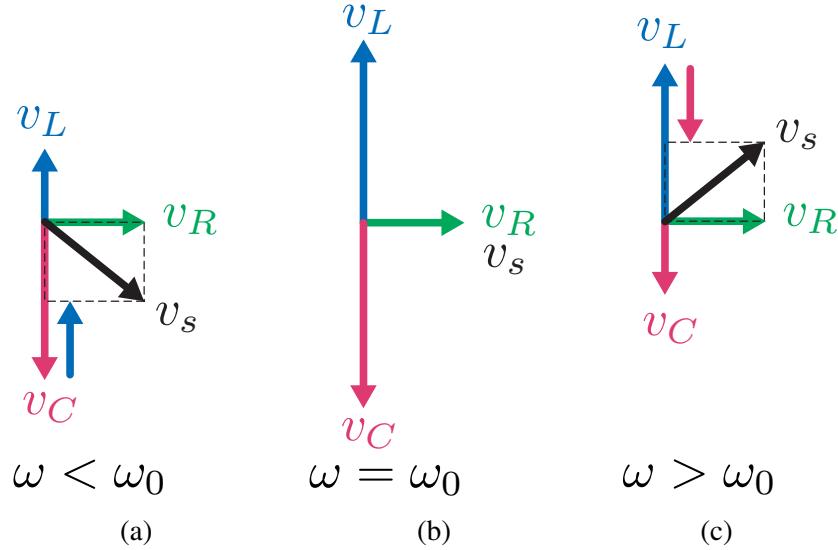


Figure 2.12: The phasor diagram shows the relations between the voltages across each component. Note that the inductor/capacitor voltages are exactly 180° degrees out of phase. (a) Below resonance, the reactance of the capacitor dominates. (b) At resonance, these reactances are equal and so the voltages across the circuit sum to v_R , so the entire source voltage is applied to the resistor R . (c) Above resonance, the reactance of the inductor dominates.

Series Resonance

It is worthwhile to investigate the cause of resonance, or the cancellation of the reactive components due to the inductor and capacitor. Since the inductor and capacitor voltages are always 180° out of phase, and one reactance is dropping while the other is increasing, there is clearly always a frequency when the magnitudes are equal. Resonance, defined as ω_0 , occurs when $\omega L = \frac{1}{\omega C}$. Now we define the resonant frequency of this circuit as:

$$\omega_0 = \frac{1}{LC} \quad (2.36)$$

Quality Factor

So what is the magic about this circuit? The first observation is that at resonance, the voltage across the reactances can be larger, in fact much larger, than the voltage across the resistors R (again, see Fig. 2.12 and observe the length of the vectors). In other words, this circuit has **voltage gain**. Of course it does not have **power gain**, for it is a **passive circuit** after all. The voltage across the inductor is given by:

$$v_L = j\omega_0 Li = j\omega_0 L \frac{v_s}{Z(j\omega_0)} = j\omega_0 L \frac{v_s}{R} = jQ \times v_s \quad (2.37)$$

where we have defined a circuit **Q (quality) factor** at resonance as:

$$Q = \frac{\omega_0 L}{R} \quad (2.38)$$

Voltage Multiplication

It's easy to show that the same voltage multiplication occurs across the capacitor (the reactances are equal at resonance after all)

$$v_C = \frac{1}{j\omega_0 C} i = \frac{1}{j\omega_0 C} \frac{v_s}{Z(j\omega_0)} = \frac{1}{j\omega_0 RC} \frac{v_s}{R} = -jQ \times v_s \quad (2.39)$$

This voltage multiplication property is the key feature of the circuit that allows it to be used as an **impedance transformer**.²

More of Q

We can re-write the Q factor in several equivalent forms owing to the equality of the reactances at resonance

$$Q = \frac{\omega_0 L}{R} = \frac{1}{\omega_0 C} \frac{1}{R} = \frac{\sqrt{LC}}{C} \frac{1}{R} = \sqrt{\frac{L}{C}} \frac{1}{R} = \frac{Z_0}{R} \quad (2.40)$$

where we have defined the $Z_0 = \sqrt{\frac{L}{C}}$ as the characteristic impedance of the circuit.

Resonant Circuit Transfer Function

Let's now examine the transfer function of the circuit. Combining L and C as a series impedance, we have a simple voltage divider:

$$H(j\omega) = \frac{v_o}{v_s} = \frac{R}{j\omega L + \frac{1}{j\omega C} + R} \quad (2.41)$$

$$H(j\omega) = \frac{j\omega RC}{1 - \omega^2 LC + j\omega RC} \quad (2.42)$$

²It is important to distinguish this Q factor from the intrinsic Q of the inductor and capacitor. For now, we assume the inductor and capacitor are ideal.

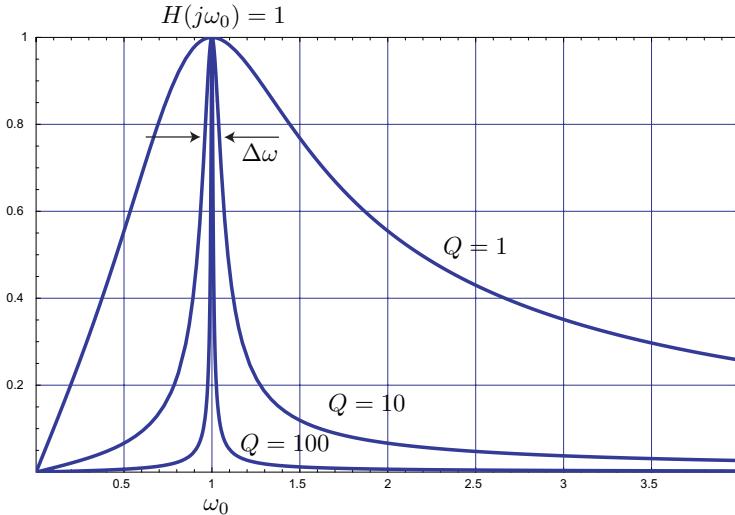


Figure 2.13: The transfer function from source to the resistor voltage as a function of frequency. Depending on the component values (circuit Q), the transfer function exhibits either low or high selectivity.

Obviously, the circuit cannot conduct DC current, so there is a zero in the transfer function. The denominator is a quadratic polynomial. It is worthwhile to put it into a standard form that quickly reveals important circuit parameters:

$$H(j\omega) = \frac{j\omega_L^R}{\frac{1}{LC} + (j\omega)^2 + j\omega_L^R} \quad (2.43)$$

Using the definition of Q and ω_0 for the circuit:

$$H(j\omega) = \frac{j\omega \frac{\omega_0}{Q}}{\omega_0^2 + (j\omega)^2 + j\frac{\omega\omega_0}{Q}} \quad (2.44)$$

Factoring the denominator with the assumption that $Q > \omega$ gives us the complex poles of the circuit:

$$s^\pm = -\frac{\omega_0}{2Q} \pm j\omega_0 \sqrt{1 - \frac{1}{4Q^2}} \quad (2.45)$$

The poles have a constant magnitude equal to the resonant frequency:

$$|s| = \sqrt{\frac{\omega_0^2}{4Q^2} + \omega_0^2 \left(1 - \frac{1}{4Q^2}\right)} = \omega_0 \quad (2.46)$$

Circuit Bandwidth and Selectivity

As we plot the magnitude of the transfer function (Fig. 2.13), we see that the selectivity of the circuit is also related inversely to the Q factor.

In the limit that $Q \rightarrow \infty$, the circuit is infinitely selective and only allows signals at resonance ω_0 to travel to the load. Note that the peak gain in the circuit is always unity, regardless of Q , since at resonance the L and C together disappear and effectively all the source voltage appears across the load. The selectivity of the circuit lends itself well to filter applications. To characterize the peakiness, let's compute the frequency when the magnitude squared of the transfer function drops

by half:

$$|H(j\omega)|^2 = \frac{\left(\omega \frac{\omega_0}{Q}\right)^2}{(\omega_0^2 - \omega^2)^2 + \left(\omega \frac{\omega_0}{Q}\right)^2} = \frac{1}{2} \quad (2.47)$$

This happens when:

$$\left(\frac{\omega_0^2 - \omega^2}{\omega_0 \omega / Q}\right)^2 = 1 \quad (2.48)$$

Solving the above equation yields four solutions, corresponding to two positive and two negative frequencies. The peakiness is characterized by the difference between these frequencies, or the bandwidth, given by:

$$\Delta\omega = \omega_+ - \omega_- = \frac{\omega_0}{Q} \quad (2.49)$$

The normalized bandwidth is inversely proportional to the circuit Q :

$$\frac{\Delta\omega}{\omega_0} = \frac{1}{Q} \quad (2.50)$$

You can also show that the resonance frequency is the geometric mean frequency of the 3 dB frequencies:

$$\omega_0 = \sqrt{\omega_+ \omega_-} \quad (2.51)$$

2.6 Bode Plots

2.6.1 Finding the Magnitude and Phase (Quickly!)

The magnitude of the response can be calculated quickly by using the property of the mag operator:

$$|H(\omega)| = \left| G_0(j\omega)^K \frac{(1 - j\omega\tau_{z1})(1 - j\omega\tau_{z2}) \cdots}{(1 - j\omega\tau_{p1})(1 - j\omega\tau_{p2}) \cdots} \right| \quad (2.52)$$

Separating out the terms:

$$= |G_0| \omega^K \frac{|1 - j\omega\tau_{z1}| |1 - j\omega\tau_{z2}| \cdots}{|1 - j\omega\tau_{p1}| |1 - j\omega\tau_{p2}| \cdots} \quad (2.53)$$

The magnitude at DC depends on G_0 and the number of poles/zeros at DC. If $K > 0$, the DC gain is zero. If $K < 0$, DC gain is infinite. Otherwise if $K = 0$, then DC gain is simply G_0 . The phase can be computed quickly with the following formula:

$$\angle H(\omega) = \angle G_0(j\omega)^K \frac{(1 - j\omega\tau_{z1})(1 - j\omega\tau_{z2}) \cdots}{(1 - j\omega\tau_{p1})(1 - j\omega\tau_{p2}) \cdots} \quad (2.54)$$

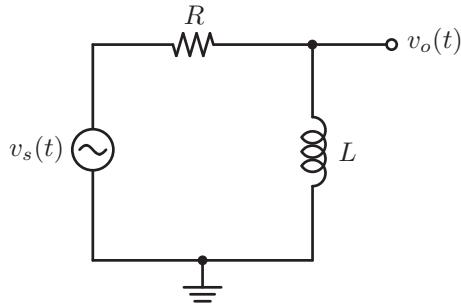
Using the properties of the phase operator:

$$= \angle G_0 + \angle(j\omega)^K + \angle(1 - j\omega\tau_{z1}) + \angle(1 - j\omega\tau_{z2}) + \cdots - \angle(1 - j\omega\tau_{p1}) - \angle(1 - j\omega\tau_{p2}) - \cdots \quad (2.55)$$

Note that the second term is simple to calculate for positive frequencies:

$$\angle(j\omega)^K = K \frac{\pi}{2} \quad (2.56)$$

Interpret this as saying that multiplication by j is equivalent to rotation by 90 degrees.

Figure 2.14: A simple *RL* high-pass filter.

2.6.2 Bode Plot

A **Bode Plot** is simply a "sketch" of the log-log plot of the magnitude and phase response of a circuit (impedance, transimpedance, gain, ...). It provides insight into the behavior of a circuit as a function of frequency without detailed calculations. The log expands the scale so that breakpoints in the transfer function are clearly delineated. In feedback circuit design, Bode plots are used to "compensate" circuits for stability.

Example: High-Pass Filter Bode Plot

The **high-pass filter** shown in *Fig. 2.14* is easily analyzed using a voltage divider:

$$H(j\omega) = \frac{j\omega L}{R + j\omega L} = \frac{j\omega \frac{L}{R}}{1 + j\omega \frac{L}{R}} \quad (2.57)$$

In terms of normalized time constants:

$$H(\omega) = \frac{j\omega\tau}{1 + j\omega\tau} \quad (2.58)$$

It is always useful to sanity check your intuition versus equations you derive. For example, for this circuit we expect (and confirm) the following behavior at various frequencies. At very high frequencies, the signal must pass through as the inductor is an open circuit:

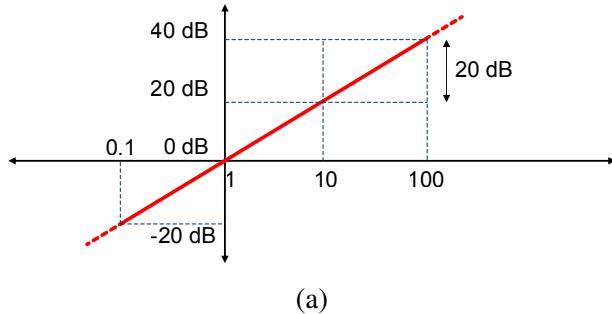
$$\omega \rightarrow \infty \quad |H| \rightarrow \left| \frac{j\omega\tau}{j\omega\tau} \right| = 1 \quad (2.59)$$

At DC the output is shorted to ground by the inductor:

$$\omega \rightarrow 0 \quad |H| \rightarrow \frac{0}{1+0} = 0 \quad (2.60)$$

At the breakpoint, the inductance provided the same impedance as the resistor, but the voltage is not half due to the phase difference:

$$\omega = \frac{1}{\tau} \quad |H| = \left| \frac{j}{1+j} \right| = \frac{1}{\sqrt{2}} \quad (2.61)$$



(a)

Figure 2.15: Sketch of the magnitude response for $j\omega\tau$.**HPF Magnitude Bode Plot: Numerator**

Recall that log of a product is the sum of logs. So we have:

$$|H(\omega)|_{dB} = \left| \frac{j\omega\tau}{1+j\omega\tau} \right|_{dB} = |j\omega\tau|_{dB} + \left| \frac{1}{1+j\omega\tau} \right|_{dB} \quad (2.62)$$

Let's focus on the numerator first. For this term, $|j\omega\tau|_{dB}$, is a straight line, and it increases by 20 dB/decade. This term crosses the 0 – dB point at the "breakpoint", or $\omega = 1/\tau$, as shown in *Fig. 2.15*:

$$\omega\tau = 1 \Rightarrow |j\omega\tau|_{dB} = 0 \text{ dB} \quad (2.63)$$

HPF Bode Plot: Denominator

Likewise, if we focus on the denominator (second term), we have:

$$\left| \frac{1}{1+j\omega\tau} \right|_{dB} = 0 \text{ dB} - |1+j\omega\tau|_{dB} \quad (2.64)$$

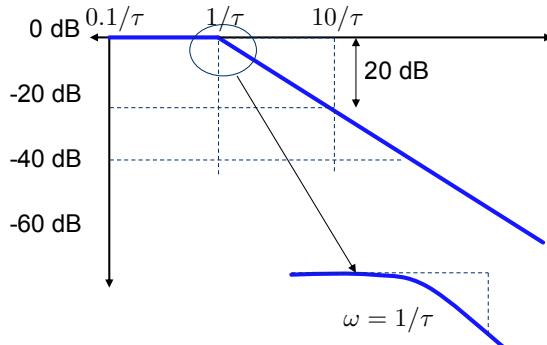
Focus on these three frequency ranges:

$$\omega \ll \frac{1}{\tau}$$

$$\omega \gg \frac{1}{\tau}$$

$$\omega = \frac{1}{\tau}$$

This is just a low-pass filter, and we can approximate the plot asymptotically as shown in *Fig. 2.16*.



(b)

Figure 2.16: Sketch of the magnitude response for a low pass function with a breakpoint at $1/\tau$.

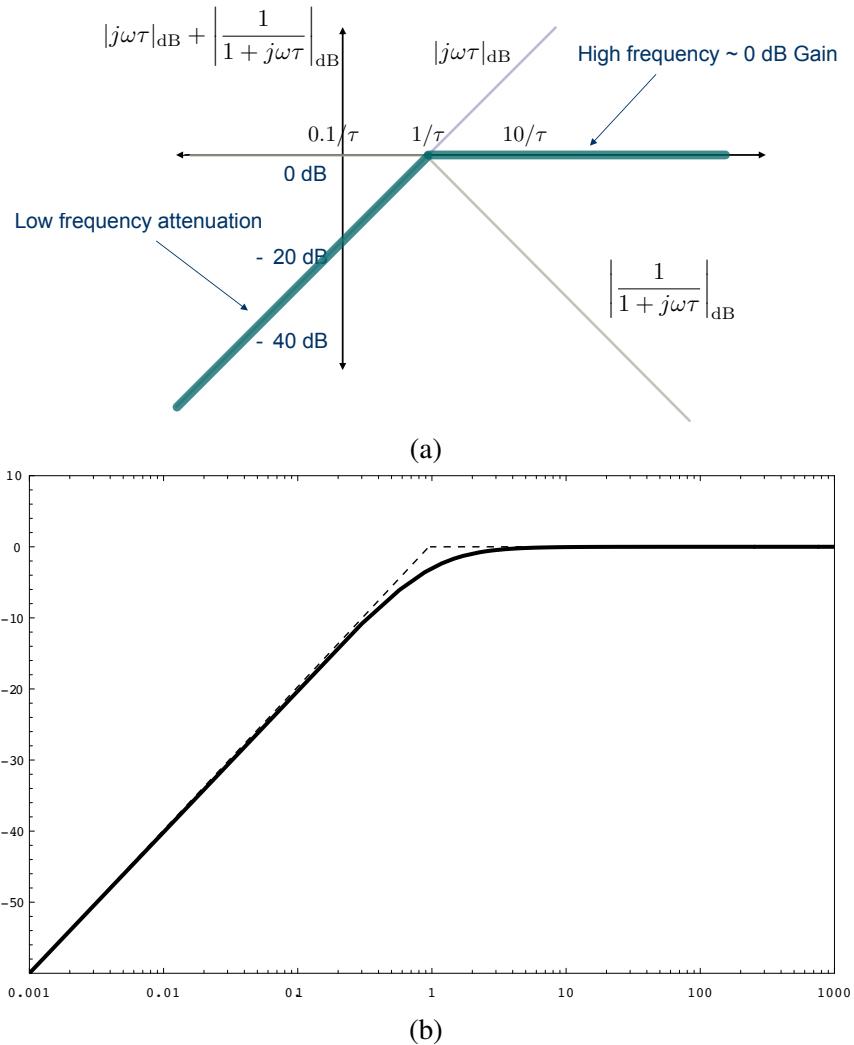


Figure 2.17: (a) Composite Bode plot for the high-pass filter obtained by summing the shown component parts. (b) A plot of the exact transfer function for the high-pass filter.

Composite Bode Plot

The **composite** is simply the sum of each component. You have the numerator term:

$$|j\omega\tau|_{dB} \quad (2.65)$$

and the denominator term:

$$\left| \frac{1}{1+j\omega\tau} \right|_{dB} \quad (2.66)$$

and together they form two plots that we join together as shown in Fig. 2.17 (a):

$$|j\omega\tau|_{dB} + \left| \frac{1}{1+j\omega\tau} \right|_{dB} \quad (2.67)$$

For comparison, we also provide an overlay of the approximate curve and the actual plot (Fig. 2.17 (b)), and we see that the approximation is very accurate away from break point. At the break point there is a 3 dB error.

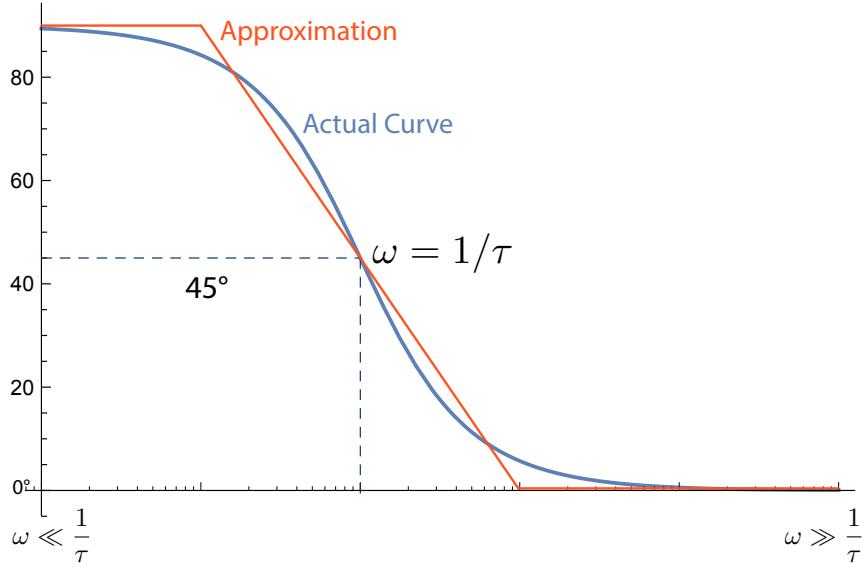


Figure 2.18: The approximate and exact phase response of the high-pass filter.

HPF Phase Plot

Phase can be naturally decomposed as well:

$$\angle H(\omega) = \angle \frac{j\omega\tau}{1+j\omega\tau} = \angle j\omega\tau + \angle \frac{1}{1+j\omega\tau} = \frac{\pi}{2} - \tan^{-1}\omega\tau \quad (2.68)$$

The first term is simply a constant phase of 90° while the second term is the arctan function. Estimate the arctan function as shown using straight lines. The rule of thumb is that the line should break at a frequency of one tenth the break point of the circuit, and end moving at a symmetrical point $10\times$ higher than the break point. See Fig. 2.18.

2.7 Power Flow in AC Circuits

The **instantaneous power** flow into any element is the product of the voltage and current:

$$P(t) = i(t)v(t) \quad (2.69)$$

For a periodic excitation, the **average power** is:

$$P_{av} = \int_T i(\tau)v(\tau)d\tau \quad (2.70)$$

In terms of sinusoids we have:

$$\begin{aligned} P_{av} &= \int_T |I| \cos(\omega t + \varphi_i) |V| \cos(\omega t + \varphi_v) d\tau \\ &= |I| \cdot |V| \int_T (\cos \omega t \cos \varphi_i - \sin \omega t \sin \varphi_i) \cdot (\cos \omega t \cos \varphi_v - \sin \omega t \sin \varphi_v) d\tau \\ &= |I| \cdot |V| \int_T d\tau \cos^2 \omega t \cos \varphi_i \cos \varphi_v + \sin^2 \omega t \sin \varphi_i \sin \varphi_v + c \sin \omega t \cos \omega t \\ &= \frac{|I| \cdot |V|}{2} (\cos \varphi_i \cos \varphi_v + \sin \varphi_i \sin \varphi_v) = \frac{|I| \cdot |V|}{2} \cos(\varphi_i - \varphi_v) \end{aligned}$$

2.7.1 Power Flow with Phasors

The result of our calculation shows that the average power is simply given by

$$P_{av} = \frac{|I| \cdot |V|}{2} \cos(\varphi_i - \varphi_v) \quad (2.71)$$

The phase term is sometimes denoted as the "**power factor**" and written as PF :

$$P_{av} = \frac{|I| \cdot |V|}{2} \cdot PF \quad (2.72)$$

Note that if $(\varphi_i - \varphi_v) = \frac{\pi}{2}$, then $P_{av} = \frac{|I| \cdot |V|}{2} \cos(\pi/2) = 0$. This means that inductors and capacitors don't dissipate any energy.

An important lesson from this calculation is that since power is a non-linear function, we cannot simply take the real part of the product of the phasors:

$$P \neq \operatorname{Re}[I \cdot V] \quad (2.73)$$

In fact, the correct answer can be deduced from our previous calculation:

$$\begin{aligned} P &= \frac{|I| \cdot |V|}{2} \cos(\varphi_i - \varphi_v) \\ &= \frac{1}{2} \operatorname{Re}[I \cdot V^*] \\ &= \frac{1}{2} \operatorname{Re}[I^* \cdot V] \end{aligned}$$

2.7.2 More Power to You!

In terms of the circuit impedance we have:

$$\begin{aligned} P &= \frac{1}{2} \operatorname{Re}[I \cdot V^*] \\ &= \frac{1}{2} \operatorname{Re}\left[\frac{V}{Z} \cdot V^*\right] \\ &= \frac{|V|^2}{2} \operatorname{Re}[Z^{-1}] \\ &= \frac{|V|^2}{2} \operatorname{Re}\left[\frac{Z^*}{|Z|^2}\right] \\ &= \frac{|V|^2}{2|Z|^2} \operatorname{Re}[Z^*] \\ &= \frac{|V|^2}{2|Z|^2} \operatorname{Re}[Z] \end{aligned}$$

Check the result for a real impedance (resistor) to verify we have not made any errors. Also, in terms of current:

$$\begin{aligned} P &= \frac{1}{2} \operatorname{Re}[I^* \cdot V] \\ &= \frac{1}{2} \operatorname{Re}[I^* \cdot I \cdot Z] \\ &= \frac{|I|^2}{2} \operatorname{Re}[Z] \end{aligned}$$

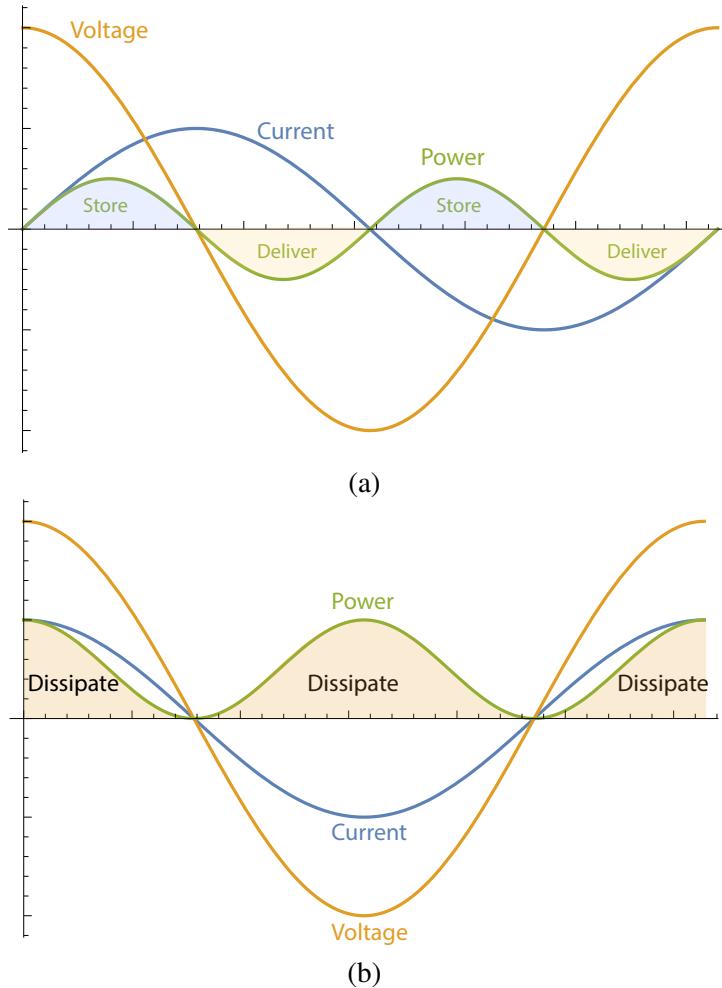


Figure 2.19: (a) The current and voltage wave forms in a reactive circuit are 90° out of phase whereas (b) for a resistive circuit, they are in phase.

2.7.3 Understanding AC Power Flow

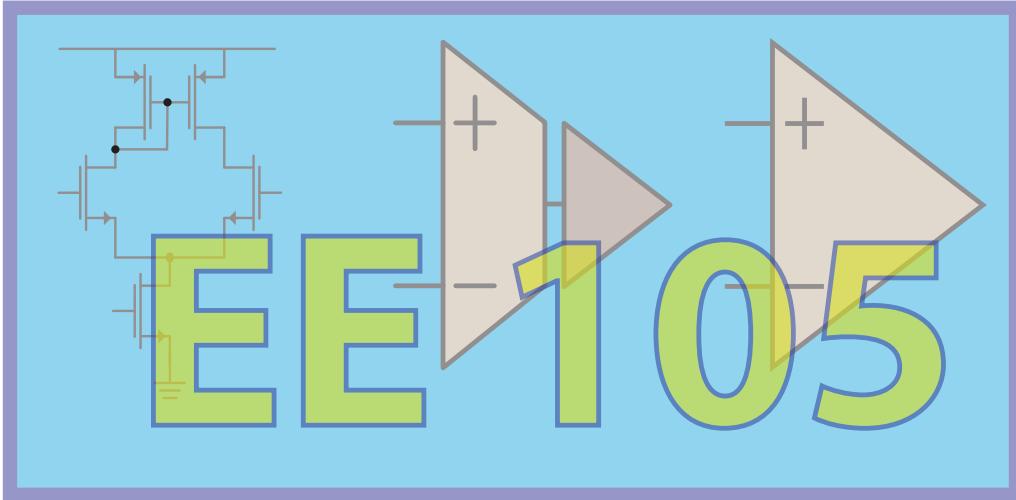
Notice that if the voltage and the current are 90° out of phase (inductors and capacitors), then their product is such that during one half the cycle power flows in, and in the other half cycle power flows out, as illustrated in *Fig. 2.19*. This means there is no net power flow into these components.

For a resistor, on the other hand, since current and voltage are in phase, it doesn't matter what direction current flows, it is always in phase with the voltage and dissipating power. Also, power flows twice per cycle!

2.8 Chapter Summary

In this chapter we demonstrated that phasor analysis allows us to treat all *LCR* circuits as simple “resistive” circuits by using the concept of impedance (admittance). While the frequency response allows us to completely characterize a system, we also showed that we can also characterize the system by the locations of its poles and zeros.

The Bode plot is a very useful tool for sketching the transfer function in the log-log domain once we know the location of the poles and zeros. Ultimately, we will learn that the location of poles and zeros is the key to understanding how a linear system will respond to inputs, or even initial conditions (charges on capacitors, currents in inductors).



3. Introduction to Semiconductors

3.1 Chapter Preview

In this chapter we will build models for conductors and semiconductors, starting with the simplest model of all, an ideal metal with a "gas" of electrons swarming about and study conduction in this ideal system. Surprisingly, many properties of real conductors are captured with this model. Next, we will delve into semiconductors, and learn why the model breaks down. Then we will briefly look at the structure of semiconductors, and develop a simple bond model that can account for free carriers; especially when a material is "doped", or when impurities are added to a solid to modulate its conductivity. With this model in hand, we'll discuss drift and diffusion currents. Drift currents are already familiar conduction of electrons in response to a field, or Ohm's Law, whereas diffusion current is due to concentration gradients, which play an important role in devices like diodes. Some of the concepts in this chapter are referenced from [6], pages 133-142.

3.2 Conduction in an Ideal Metal "Gas"

3.2.1 Ohm's Law

One of the first things we learn in electrical engineering and physics is Ohm's law:

$$V = I \times R \quad (3.1)$$

Is this trivial? Maybe what's really going on is the following:

$$V = f(I) = f(0) + f'(0)I + f''(0)I^2/2 + \dots \approx f'(0)I \quad (3.2)$$

In the above **Taylor expansion**, if the voltage is zero for zero current, and the current is small, then this is generally valid for a reasonably smooth function. The range of validity (radius of convergence) is the important question. It turns out to be VERY large!

3.2.2 Ohm's Law Revisited

In physics we learned that the current density is proportional to the electric field:

$$J = \sigma \mathcal{E} \quad (3.3)$$

Is this also trivial? Well, it's the same as Ohm's law, so the questions are related. In Eq. 3.3, σ is the **conductivity**, which is the reciprocal of **resistivity**, ρ . Recall the definitions of both conductance and resistance:

$$R = \frac{V}{I} = \rho \frac{\ell}{A} \quad \text{Definition of resistance} \quad (3.4)$$

$$G = R^{-1} = \frac{I}{V} = \rho^{-1} \left(\frac{\ell}{A} \right)^{-1} = \sigma \frac{A}{\ell} \quad \text{Definition of conductance} \quad (3.5)$$

Also recall that we can define the electric field as the potential over a certain displacement, and the current density is the current flow through a cross-sectional area. For a rectangular solid:

$$J = \sigma \mathcal{E} = \sigma \frac{V}{\ell} = \frac{I}{A}$$

Equating the last two terms from the above equation and solving for V gives us Ohm's law:

$$V = I \times \frac{\ell}{\sigma A} = I \times \sigma^{-1} \frac{\ell}{A} = I \times \rho \frac{\ell}{A} = I \times R$$

Is it not strange that current (velocity) is proportional to force? Where does conductivity come from?

3.2.3 Conductivity of a Gas

Electrical conduction is due to the motion of positive and negative charges. For example, for water with pH=7, the concentration of hydrogen H+ ions (and OH-) is:

$$10^{-7} \text{ mole/L} = 10^{-10} \text{ mole/cm}^3 = 10^{-10} \times 6.02 \times 10^{23} \text{ cm}^{-3} = 6 \times 10^{13} \text{ cm}^{-3} \quad (3.6)$$

Typically, the concentration of charged carriers is much smaller than the concentration of neutral molecules. The motion of the charged carriers (electrons, ions, molecules) gives rise to electrical conduction.

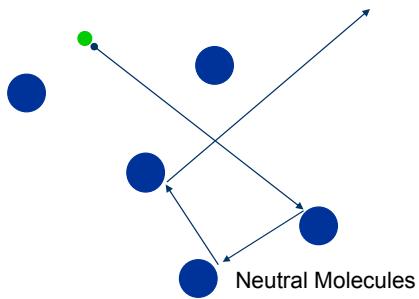


Figure 3.1: Simple model for a charged particle like an electron moving in a random array of atoms or molecules. We assume the particle moves freely until it encounters an atom or molecule, causing it to scatter off in a new direction with a new speed.

Collisions in Gas

At a temperate T , each charged carrier will move in a random direction and velocity until it encounters a neutral molecule or another charged carrier. Since the concentration of charged carriers is much less than molecules, it will most likely encounter a molecule. For a gas, the molecules are widely separated (~ 10 molecular diameters). After colliding with the molecule, there is some energy exchange and the charge carrier will come out with a new velocity and new direction.

Memory Loss in Collisions

Schematically our model thus far is shown in *Fig. 3.1*. The key point is the initial velocity and direction is lost (randomized) after a few collisions.

Application of Field

When we apply an electric field, during each "free flight", the carriers will gain a momentum of $\mathcal{E}qt$. Therefore, after t seconds, the momentum gained during "free flight", if there is no collisions, is given by:

$$M\mathbf{u} + \mathcal{E}qt \quad (3.7)$$

If we take the average momentum of all particles at any given time, we have:

$$M\bar{\mathbf{u}} = \frac{1}{N} \sum_j (M\mathbf{u}_j + \mathcal{E}qt_j) \quad (3.8)$$

In this equation, N is the number of carriers (of charge), $M\mathbf{u}_j$ is the initial momentum before collision, and $\mathcal{E}qt_j$ is the momentum gained from the field in the time t_j from the last collision.

Random Things Sum to Zero!

When we sum over all the random velocities of the particles, we are averaging over a large number of random variables with zero mean, the average is zero:

$$M\bar{\mathbf{u}} = \frac{1}{N} \sum_j \left(\underbrace{M\mathbf{u}_j}_{\text{Initial momentum before collision.}} + \underbrace{\mathcal{E}qt_j}_{\text{Momentum gained from the field.}} \right) \quad (3.9)$$

This allows us to ignore the first sum, leading to:

$$M\bar{\mathbf{u}} = \frac{1}{N} \sum_j \mathcal{E}qt_j = \mathcal{E}q\tau \quad (3.10)$$

So the current is given by:

$$\mathbf{J} = \mathbf{Nq}\bar{\mathbf{u}} = \mathbf{Nq} \left(\frac{\mathcal{E} \mathbf{q} \tau}{\mathbf{M}} \right) = \mathbf{Nq}^2 \frac{\tau}{\mathbf{M}} \mathcal{E} = \sigma \mathcal{E} \quad (3.11)$$

3.2.4 Mobility

From the previous derivation, we see that the average momentum gain from the field is given by:

$$\mathbf{J} = \mathbf{J}^+ - \mathbf{J}^- = e \left(\frac{N^+ e \tau^+}{M^+} - \frac{-N^- e \tau^-}{M^-} \right) \mathcal{E} \quad (3.12)$$

In many situations we'd like to find the average speed gained from the field, which is defined as the **mobility** (μ):

$$\mu = e^2 \left(\frac{N^+ \tau^+}{M^+} + \frac{N^- \tau^-}{M^-} \right) \quad (3.13)$$

So we see that even though we apply an electric field, applying a force to electrons, they don't accelerate like free electrons. Instead they gain only a fixed amount of momentum, not linearly increasing in time as predicted for free electrons.

A good analogy is the following. Imagine driving a car on a busy freeway, a stop-and-go situation. Every time you have the opportunity to move, you accelerate a certain amount of time but very soon the car in front of you stops, forcing you to hit the breaks. So even though you're accelerating, or trying to accelerate forward, you're actually just able to gain momentum for brief periods of time in between the stops. In a similar fashion, when electrons or other charges accelerate under the application of a field, they only do so for a short period of time, in between collisions, and so they can only gain a bit of momentum on average before they are forced to stop and try again, rather than continuously gaining momentum from the field such as in free space.

Negative and Positive Carriers

Since current is contributed by positive and negative charge carriers:

$$\mathbf{J} = \mathbf{J}^+ - \mathbf{J}^- = e \left(\frac{N^+ e \tau^+}{M^+} - \frac{-N^- e \tau^-}{M^-} \right) \mathcal{E} \quad (3.14)$$

Or in terms of mobility:

$$\sigma = e^2 \left(\frac{N^+ \tau^+}{M^+} + \frac{N^- \tau^-}{M^-} \right) \quad (3.15)$$

3.2.5 Conduction in Metals

Can we apply this simple "gas" model to a conductor? The high conductivity of metals is due to large concentration of free electrons. These electrons are not attached to the solid but are free to move about the solid. In other words, we have an "electron gas". In metal sodium, for example, each atom contributes a free electron:

$$N = 2.5 \times 10^{22} \text{ atoms/cm}^3$$

From the measured value of conductivity (easy to make in the lab), we can back calculate the **mean free-time**:

$$\tau = \frac{\sigma m}{Ne^2} = \frac{(1.9 \times 10^{17})(9 \times 10^{-28})}{(2.5 \times 10^{22})(23 \times 10^{-20})} = 3 \times 10^{14} \text{ sec} \quad (3.16)$$

A Deep Puzzle

This value of mean free-time is surprisingly long. The mean velocity for an electron at room temperature is about:

$$\frac{mv^2}{2} = \left(\frac{3}{2}\right) kT v = 3 \times 10^7 \text{ sec} \quad (3.17)$$

At this speed, the electron travels a distance of $v\tau = 3 \times 10^{-7}$ cm. The molecular spacing between adjacent ions is only 3.8×10^{-8} cm. Why is it that the electron is on average zooming by 10 positively charged ions?

3.2.6 Wave Nature of Electron

The free carrier can penetrate right through positively charged host atoms, as shown schematically in Fig. 3.2a. **quantum mechanics** explains this. For a periodic arrangement of potential functions, the electron does not scatter. The influence of the crystal is that it will travel freely with an **effective mass** different from the **rest mass** or **free-electron mass**. So why does it scatter at all?

3.2.7 Scattering in Metals

At temperature T , the atoms are in random motions in the host atoms (see Fig. 3.2b), and so the potential function experienced by charged carriers is not periodic, but quasi-periodic.

Even at extremely low temperatures, the presence of an impurity upsets the periodicity, as shown in Fig. 3.2c. These two mechanisms, vibrations in the crystal (**phonons** in the parlance of solid-state physics) and impurities are the source of **scattering**, and thus resistance, in metals.

3.2.8 Summary of Conduction

Using our ideal gas model for conductors, we have the following result:

$$\sigma = e^2 \left(\frac{N^+ \tau^+}{M^+} + \frac{N^- \tau^-}{M^-} \right) \quad (3.18)$$

The conductivity is determined by the density of free charge carriers (both positive and negative), the charge of carrier (e), the effective mass of carrier (different inside solid), and the mean relaxation time (the time of the memory loss, usually the time between collisions). This turn is determined by several mechanisms, e.g. scattering by impurities and scattering due to vibrations in crystal.

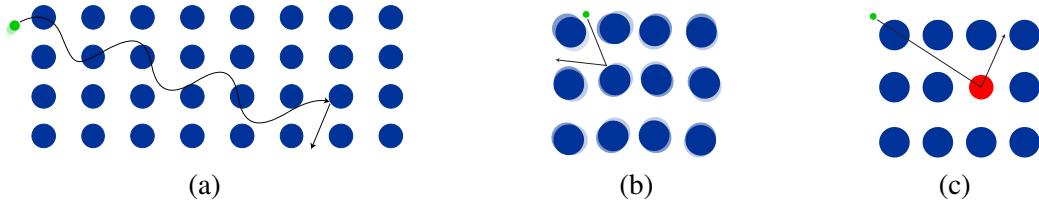


Figure 3.2: (a) Measurements show that the average distance electrons move in a crystal is much larger than the inter-atomic distance, indicating that the electrons somehow move through the nuclei unperturbed. In fact, a free electron in a periodic arrangement of atoms can move freely around the crystal as a "free particle" with an *effective mass* that takes into account the potential of the nuclei in the crystal. In other words, the electron does not scatter simply because it encounters an atom. (b) The random thermal motion of atoms in a crystal (*lattice vibrations*) disturbs the periodicity of the array and results in scattering of free carriers. (c) The presence of an impurity atom (shown in red) also disturbs the periodicity of the array and results in scattering.

3.3 Introduction to Semiconductors

3.3.1 Resistivity for a Few Materials

Let's take a careful look at the conductivity for a few different materials:

- Pure copper, 273K 6.4×10^7 S/m
- Pure copper, 373K 4.46×10^6 S/m
- Pure germanium, 273K 0.5 S/m
- Pure germanium, 500K 833 S/m
- Pure water, 291K 4×10^{-6} S/m
- Seawater 4 S/m

This list is very interesting in several ways. First of all we see that copper, a good conductor, is in fact a fantastic conductor, with a conductivity several orders of magnitude larger than a material like germanium. Germanium is a "semi" conductor, or just **semiconductors**, for obvious reasons. It conducts, but very poorly. Also, whereas temperature has a relatively minor impact on the conductivity of copper, decreasing by about 30% when the temperature is increased by 100°, germanium seems to be much more sensitive. The trends are also opposite as germanium is becoming more conductive, by orders of magnitude, with increasing temperature ! Why?

Finally, we have pure water, which is an insulator, because in pure water there are essentially no free charges. But sea water (and human tissue) is moderately conductive, but it doesn't have the same sensitivity to temperate as a semiconductor.

We have several mysteries on hand! What gives rise to this enormous range? Why are some materials semi-conductive? Why the strong temperature dependence for semiconductors?

Periodic Table of Elements

Have a look at the periodic table reproduced here in *Fig. 3.3*. You may recognize that many semiconductors are either Group III, IV, or V elements. What is special about Group III, IV, and V elements? Recall that the number of outer shell electrons in atoms is related to its group number. Noble gases are unique in that they have complete shells and this is in fact why they are so stable. Group IV elements in particular have four electrons in their outer shell. This is why when they are grouped together, they bond and share electrons, four electrons on average.

Group →	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Period ↓	1	H															He	
2	Li	Be											5	6	7	8	9	Ne
3	Na	Mg											13	14	15	16	17	Ar
4	K	Ca	Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	31	32	33	34	35	Kr
5	Rb	Sr	Y	Zr	Nb	Tc	Ru	Rh	Pd	Ag	Cd	In	50	51	52	53	54	Xe
6	Cs	Ba	La	Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg	81	82	83	84	85	Rn
7	Fr	Ra	Ac	* 104	105	106	107	108	109	110	111	Rg	112	113	114	115	116	117
				Db	Sg	Bh	Hs	Mt	Ds	Rs	Cn	Nh	Fl	Mc	Lv	Ts	Og	
				* 58	59	60	61	62	63	64	65	Dy	67	68	69	70	71	
				* 90	91	92	93	94	95	96	97	Bk	98	Cf	99	100	101	102
												Es	Fm		102	103		

Figure 3.3: Periodic table of elements. Note that many semiconductors are either group 4 elements or combinations of group 3 and 5 elements. Credit: By Offnfopt - Own work, Public Domain, <https://commons.wikimedia.org/w/index.php?curid=62296883>

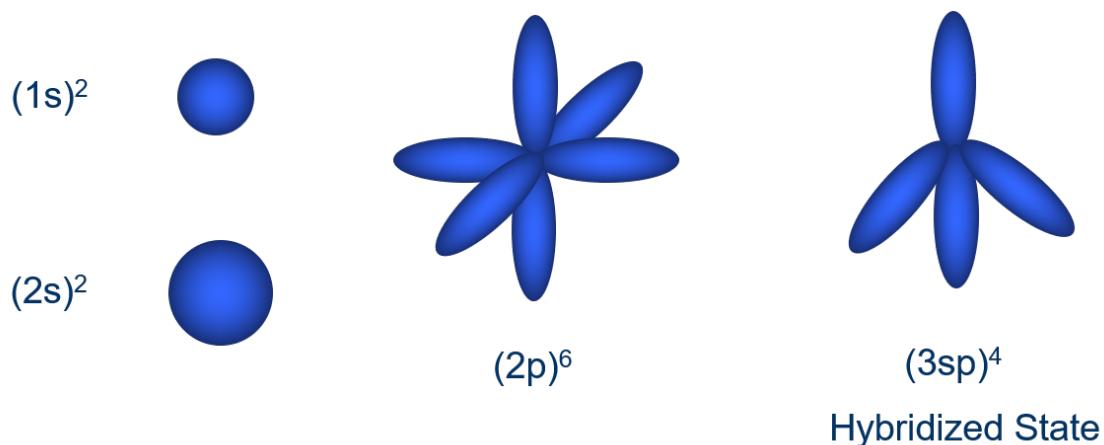


Figure 3.4: The s and p orbitals of the isolated silicon atoms and the $3sp$ hybridized state consisting of a linear combination of s and p orbitals. When silicon atoms form covalent bonds, the hybridized state is preferred and results in a lower energy configuration.

3.3.2 Electronic Properties of Silicon

We will focus on silicon, but a lot of what we have to say applies to most semiconductors. Silicon is the most commonly used semiconductor, so it is the most relevant example. From chemistry, we know that silicon is in Group IV element, like carbon. It has a total of fourteen electrons, arranged into the following orbital structure shown in *Fig. 3.4*:

- Atom electronic structure: $1s^2 2s^2 2p^6 3s^2 3p^2$
- Crystal electronic structure: $1s^2 2s^2 2p^6 3(sp)^4$
- Diamond lattice, with 0.235 nm bond length

Notice that the s and p orbitals combine when the silicon atoms combine into a hybridized state. This deserves a bit of explanation. When silicon atoms are far apart, the electrons occupy the usual orbital states you're familiar with from chemistry. The last two electrons are in the p orbital, which can hold a total of 6. But when the atoms are brought into close proximity, the outer shell valence electrons feel the influence of two nuclei. Therefore, it is logical to assume that the electrons should spend more time in between the nuclei, essentially forming covalent bonds. If the **Pauli exclusion principle** didn't apply, then all the electrons would occupy the p orbitals in between the silicon atoms, leaving some empty and some more than full! But alas, only two electrons per orbital (with opposite spin).

On the other hand, the geometry of the p orbitals shown in *Fig. 3.4* would be ideal if silicon could bond to six neighbors. But this does not happen, because there would be too many valence electrons since each atom already has two $3p$ electrons, since silicon only needs four additional electrons to form a complete shell. On the other hand, the hybridized $3(sp)^4$ state (linear combination of s and p orbitals) contains four electrons and can bond with four neighbors as shown. This scenario is how silicon atoms combine to form a crystal, with the geometry more complicated than a simple cube arising from these bond angles.

Also, silicon is a very poor conductor at room temperature. Why?

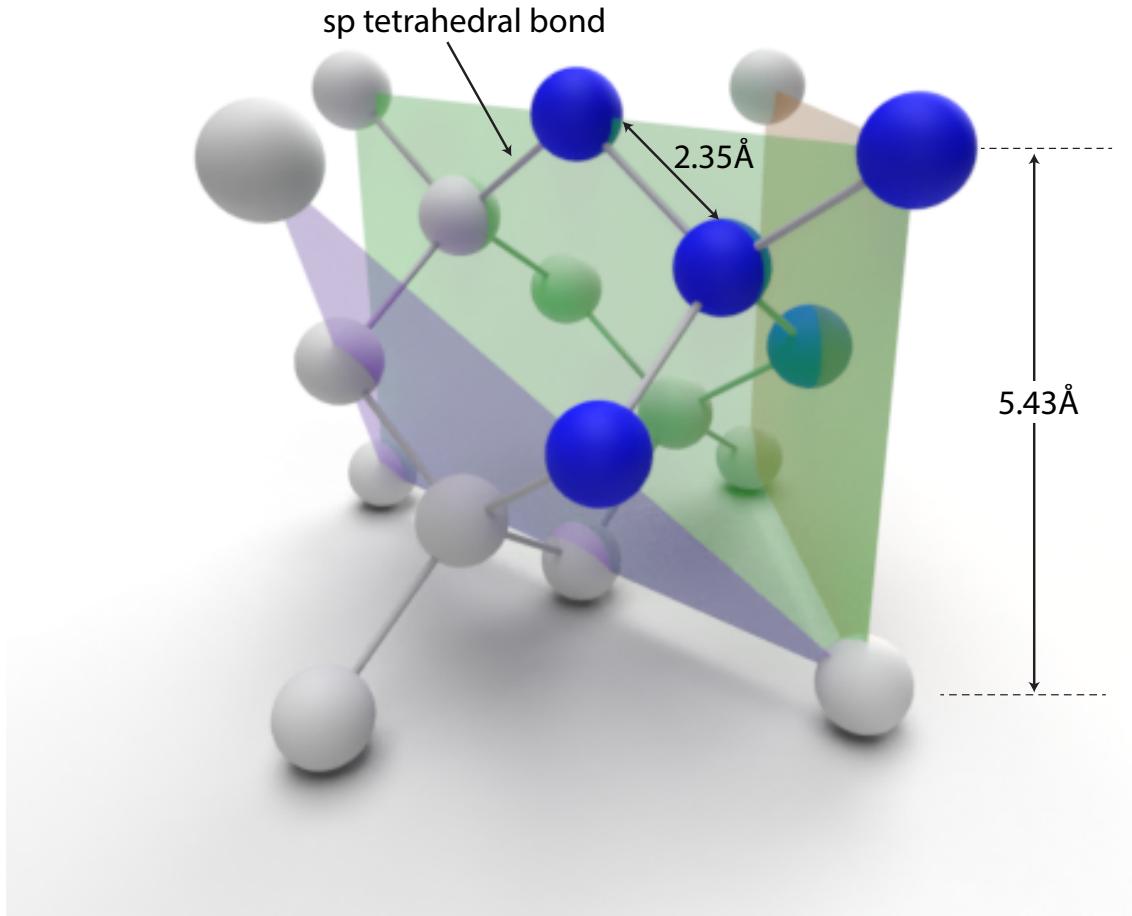


Figure 3.5: The crystal structure of silicon. All atoms are silicon, and have covalent bonds with their four neighbors. The colored atoms in this 3D model represent a unit cell that is repeated across the crystal. The symmetry planes are also highlighted. A 3D model can be viewed and manipulated online <https://sketchfab.com/3d-models/silicon-crystal-lattice-73e292f32ffe4ca490e166faeba317e7>.

Si Diamond Structure

Silicon atoms arrange in a nice crystal structure known as the **diamond structure** shown in *Fig. 3.5*. In this figure, spheres represent silicon atoms with one particular unit cell highlighted. The planes in the figure correspond to the symmetry planes of the crystal. Usually the crystal is cut along one of these symmetry planes.

Notice that each silicon atom bonds to four neighbors using a covalent (shared electron) bond. In a **covalent bonds**, the electrons are shared among a group of host atoms, giving up their identity and loyalty to a fixed atom as they are shared among a group of nuclei, forming a more stable overall structure than if they remained in their lone atomic orbitals. The inner core electrons, on the other hand, are bound to the host nucleus and don't participate in conduction (the interest of this chapter) unless very energetic particles such as x-rays are incident on the crystal.

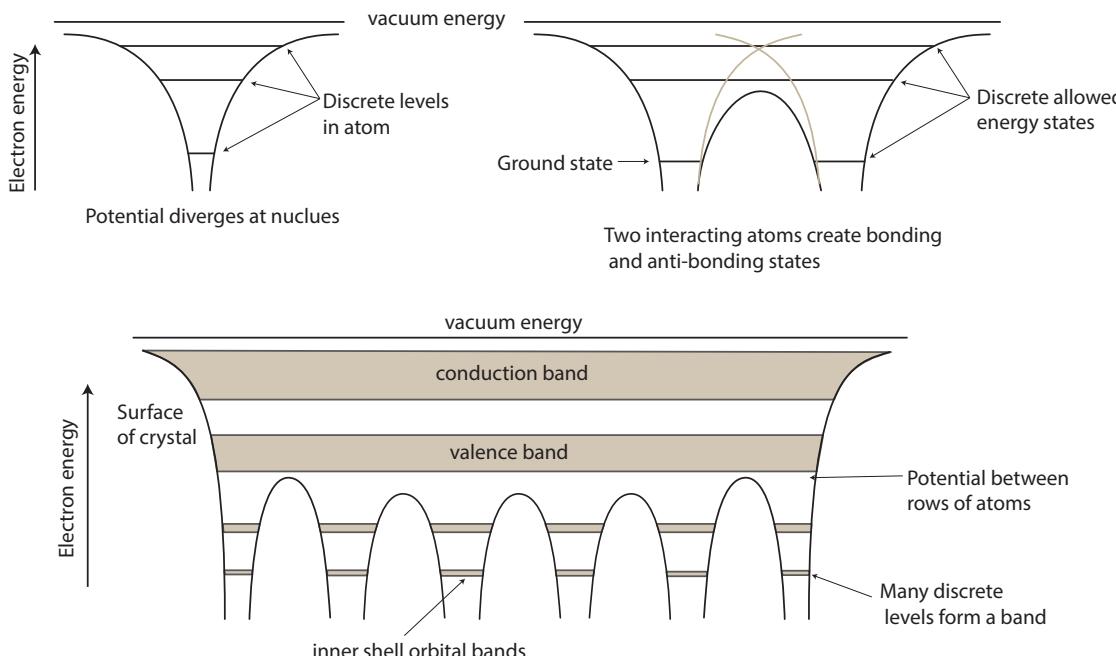


Figure 3.6: On the top left we have the discrete energy levels for isolated atoms. Electrons are only found to have discrete energy level with forbidden energy levels in between allowed states. We now imagine bringing two atoms closer and closer together as shown on the top right. Due to their interaction, the energy levels split into bonding and non-bonding states. When many atoms are involved (bottom), these splittings become very fine forming an almost continuous band of allowed energy levels. The top two bands called the conduction and valence band are the most relevant to our discussion. (Figure adapted from D. Neamen, *Semiconductor Physics And Devices*, 2003. [4])

3.3.3 States of an Atom

From quantum mechanics and chemistry, we know that for an atom the allowed energy levels for an atom are discrete (2 electrons can occupy a state since with opposite spin), as shown in *Fig. 3.6(a)*. There is an energy gap between states that electrons cannot occupy. From solid-state physics, we have an important result that when a collection of atoms are brought together, the energy levels split as shown in *Fig. 3.6(b)*. If there are a large number of atoms, the discrete energy levels form a nearly “continuous” band. The gap between the energy levels may widen or even disappear, giving rise to different material properties.

3.3.4 Energy Band Diagram

In the energy diagram discussed, the lines represent allowed energy levels and the gaps represent levels that are simply not supported, they are forbidden. This arises from the wave properties of the electron. When an electron “orbits” an atom, we know that its orbit corresponds to an integer number of **de Broglie**¹ wavelengths. In essence, when in such orbits, the electron wave experiences constructive interference whereas in other orbits it experiences destructive interference. Because the square magnitude of the “wave” is the probability to find the electron, we find it is more probable for electrons to occupy these orbits. This is a very simple perspective, but it got people like **Niels Bohr**² to plunge ahead and develop a model for an atom. In a collection of “free” electrons confined to the crystal “box”, the spatial confinement results in a discrete energy (and

¹See https://en.wikipedia.org/wiki/Louis_de_Broglie for more information about Louis de Broglie.

²See https://en.wikipedia.org/wiki/Niels_Bohr for more information about Niels Bohr.

momentum) spectrum. The allowed energy levels are essentially a result of the boundary condition that requires that electrons are confined to the box with effectively zero probability to appear outside of the box. When boundary conditions are imposed on the wave function, a discrete spectrum of allowed states results.

In a similar way, for a collection of atoms, the forbidden energy states cannot be occupied by electrons. Lower energy states correspond to "**valence band**" electrons, or electrons that are bound to host atoms. Higher energy states, in the "**conduction band**" are "free" electrons that take part in conduction, as they are free to move around the crystal. In essence, the conduction band electrons have sufficient energy to leave the nucleus and roam the crystal. The host atoms become ionized and charged as a result. The electrons act like free electrons in a vacuum, albeit with a different mass.³

The gap between the conduction and valence band determines the conductive properties of the material (see *Fig. 3.7*). In insulators, the gap between the conduction band and valence band, or the **band-gap**, is quite large, $\sim 4 - 8 \text{ eV}$. Therefore, it takes an enormous amount of energy to pull electrons away from host atoms. At room temperature, electrons only have an energy of about 25 meV , so an 4 eV band gap is nearly an infinite one to overcome.

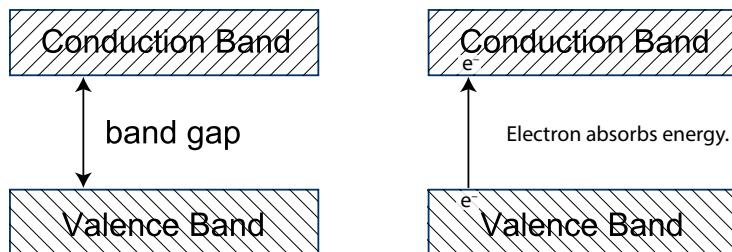


Figure 3.7: (left) The band model approximates the discrete energy states in the valence and conduction band as a continuum. Electrons in the valence band cannot conduct current as they are in covalent bonds. Electrons in the conduction band can respond to external fields and result in net current flow. (right) Electrons in the valence band can gain energy from photons or phonons (lattice vibrations) and move into the conduction band. The phonon or photon must have sufficient energy to cross the *band gap*, which is a forbidden state.

In a metal, the conduction band is partially filled and therefore many electrons are free to move about, resulting in good conduction properties. The band gap is either very small, or partially overlapping with the valence band, resulting in no discernible band gap. And this leaves semiconductors, which are materials that have a "medium sized" gap, about $\sim 1 \text{ eV}$. This means that the barrier to become a free electron is high, but not impossible. In a given solid of semiconductor, there's always a few lucky electrons that will be in the conduction band. In fact, the probability of occupying a certain energy level is related to the **Boltzmann distribution** modified to account for the fact that electrons are **Fermions**, and cannot occupy the same energy levels, resulting in **Fermi-Dirac statistics**.

As discussed previously, thermal energy is very low on average ($\sim 25 \text{ meV}$), but there are a few (very few) high energy electrons. So the conductivity of semiconductors is poor, but not zero. As shown in *Fig. 3.7*, electrons in the valence band can gain energy from incoming **photons** or from the **lattice vibrations** in the crystal (phonons) and if the energy is sufficiently large, they can absorb it and go into the conduction band.

³The electrons are still bound to reside in the crystal, and a significant amount of energy is required to remove these electrons from the crystal due to the fact that the entire crystal is charge neutral and leaving the crystal means charge separation

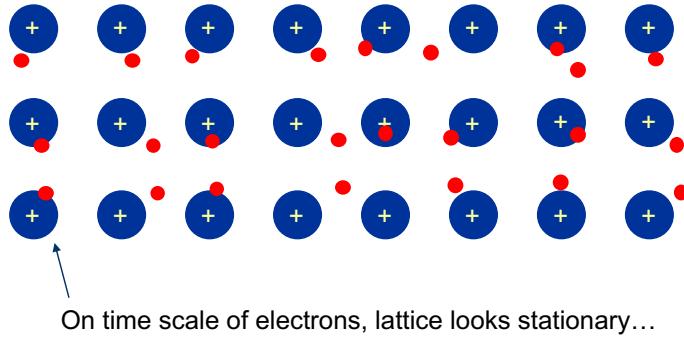


Figure 3.8: Model for good conductor has many ionized atoms arranged in a crystal surrounded by clouds of free electrons that "glue" the atoms together.

3.3.5 Model for Good Conductor

Given this background, we can now model a good conductor as a collection of atoms that are all ionized forming a “sea” of electrons can wander about crystal, shown in *Fig. 3.8*. The electrons are the “glue” that holds the solid together. They spend most of their time in between the ionized nuclei, effectively shielding the nuclei from each other, lowering the energy of the system. But they are not confined to bonds, they are free to move around the crystal and contribute to conduction. Since they are “free”, they respond to applied fields and give rise to *ohmic* currents. They also respond to incoming photons and give rise to the optical properties such as the fact that these materials are opaque and have shiny surfaces. Electrons at the surface of the structure readily accept optical photons and re-radiate them.

Semiconductor Bond Model

Unlike a conductor, a semiconductor does not have many free electrons. This is because all of the electrons are forming bonds in the crystal. How can we free electrons? Either by giving them enough energy to break free (photons or very high energy phonons), or we have to introduce impurities in the crystal structure that have a different number of bonding electrons. We will discuss this in detail next.

3.3.6 Bond Model for Silicon ($T = 0 K$)

Let’s build a simple 2D model for silicon as shown in *Fig. 3.9*. The circles represent the nuclei and the lines represent electrons that are forming covalent bonds, or residing in the $3(sp)^4$ orbitals. For simplicity, we squashed the orbitals to lie on a plane.

Each bond holds two electrons with opposite spin. Each atom contributes four valence electrons to the orbital and borrows four electrons from each of its four neighbors. Thus, each atom experiences a lower energy state corresponding to having a full orbital, despite being a group IV element. All the electrons are busy orbiting the nuclei, and none can contribute to current conduction.

3.3.7 Bond Model for Silicon ($T > 0 K$)

Now let’s suppose we increase the temperature, which means the crystal lattice can absorb energy in the form of vibrations.⁴ Occasionally enough energy can be absorbed by a valence electron to be broken free from the lattice structure. In the energy **band model**, we know the energy absorbed has to be at least as large as the band-gap. Once an electron becomes a conduction band electron, it’s

⁴Quantum mechanically, these vibrations can be treated as particle “phonons” (in analogy with photons) which can exchange energy with the electrons.

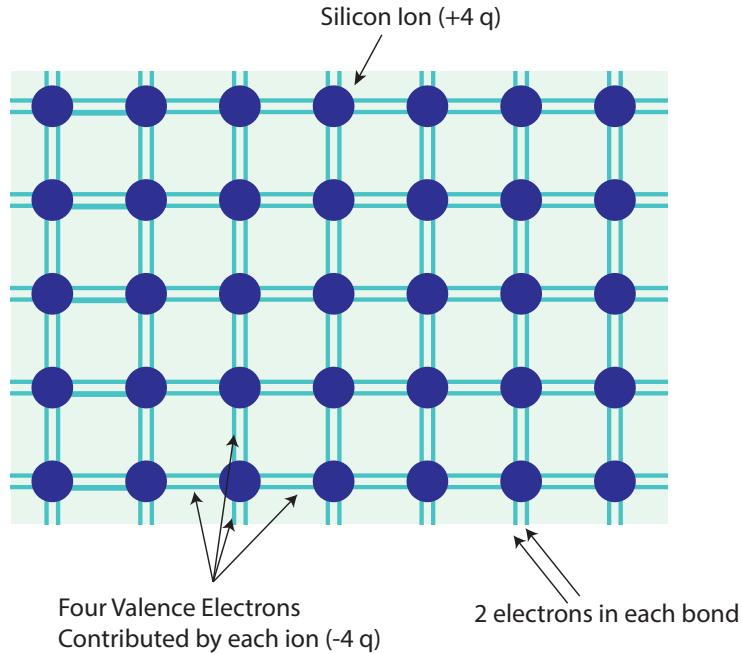


Figure 3.9: Simplified model for the silicon crystal at $T = 0^\circ\text{K}$.

free to roam the crystal like a free electron (see *Fig. 3.10*). The mass of the electron is not the same as the mass of an electron in a pure vacuum. That's because the electron will still feel the influence of the crystal potential, which varies periodically. It turns out that we can define an "effective" mass for the electron and treat it as if it were free and in vacuum, free from the influence of the **crystal potential**.

3.3.8 Holes?

When we formed a free electron, we leave behind a broken bond. This bond means the host atom has net charge. But this charge corresponding to the vacancy is "fixed", right? Notice, as illustrated in *Fig. 3.11*, that the vacancy (hole) left behind can be filled by a neighboring electron. But if the neighboring electron jumps into this vacancy, it forms its own vacancy while filling the other one. This chain reaction of vacancies moving around is complex as it corresponds to the motion of many

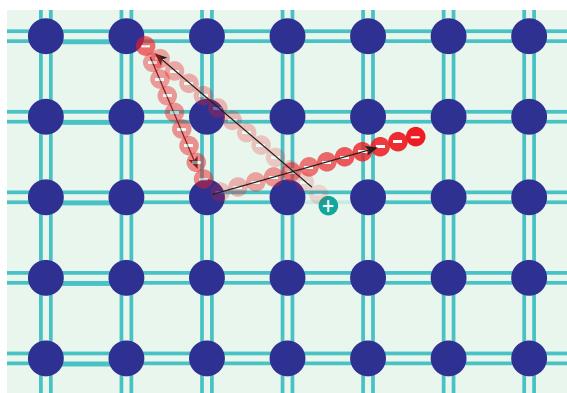


Figure 3.10: In the silicon crystal, when a bond is broken, we create both a free electron and a free hole. The hole is a fictitious particle that behaves like a real particle in every way. It represents the motion of many electrons in the valence band.

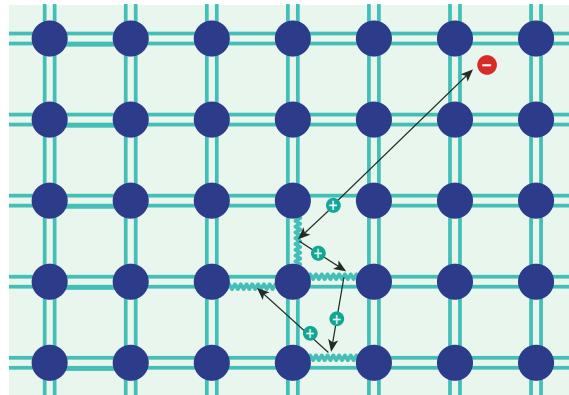


Figure 3.11: Illustration of how a broken covalent bond moves around the crystal like a free particle. Each wiggly bond represents a valence band electron moving into the vacancy created by the hole.

valence band electrons. An easy way to model it is to treat it like there is a positive charge traveling around. We call this positive (fictitious⁵) particle a "**hole**". It turns out that we can treat the holes as legitimate particles with positive mass and of course positive charge (because the broken bond leaves behind a net charge of +1).

3.3.9 Yes, Holes!

Using results from solid state theory, it can be shown that the net motion of many electrons in the valence band can be equivalently represented as the motion of a hole. The picture is more complicated because electrons near the valence band edge have negative effective mass! This is why the hole model works as it does, something we're not going to worry about in this chapter (or book). Another common analog is to say a hole is like a bubble in a bottle. It can move around as if it were a free particle, but in fact its motion is the result of the motion of a lot of water molecules. If we completely fill the bottle up, there is no opportunity for a bubble to float around. In the same way, a filled valence band does not contribute current. Even if the electrons are moving (due to motion in the bond), the net current is zero

$$\sum_{\text{Filled Band}} (-q)v_i = 0 \quad (3.19)$$

Now let's take a partially filled band and play a trick. Let's say a partially filled band is a full band minus some electrons that we need to take out of the valence band to make it full:

$$J_{vb} = \sum_{vb} (-q)v_i = \sum_{\text{Filled Band}} (-q)v_i - \sum_{\text{Empty States}} (-q)v_i \quad (3.20)$$

Since the first term is the full band and it does not contribute any current, the resulting current is due to the electrons we subtracted out to make a partially filled band. In other words, the current is due to the empty states and even though the charges are negative, they act like positive charges:

$$J_{vb} = - \sum_{\text{Empty States}} (-q)v_i = \sum_{\text{Empty States}} qv_i \quad (3.21)$$

⁵Calling it real or fake is almost a philosophical question. If it looks like a duck, swims like a duck, and quacks like a duck, then it probably is a duck.

3.3.10 More About Holes

When a conduction band electron encounters a hole, the process is called **recombination**. The electron can fill the void, so we say that the electron and hole annihilate one another thus depleting the supply of free carriers. This happens by chance as free electrons roam about and encounter holes. So this recombination process depletes the crystal of both free electrons and free holes. In thermal equilibrium it stands to reason that there must be a **generation** process to counterbalances to produce a steady stream of carriers. This is of course the thermal generation we considered before.

3.4 Intrinsic Carrier Concentration

3.4.1 Thermal Equilibrium (Pure Si)

In thermal equilibrium, there must be a balance between generation and recombination, otherwise a material would have more and more carriers spontaneously over time due to generation. To determine the **thermal equilibrium** concentration of holes, p_0 , and electrons n_0 , first observe that every time we create a free electron, we also create a free hole. Likewise, every time a hole and electron recombine, both the free hole and free electron carrier concentration drop together. So that means that $n_0 = p_0$.

Since the generation process is thermal in nature, we expect that the generation rate to be a strong function of temperature.

Let's say G represents the total about of free carrier generation per unit of time, or the rate of carrier generation:

$$G = G_{th}(T) + G_{opt} \quad (3.22)$$

Note that in general we include thermal and optical carrier generation. For the sake of simplicity, we'll ignore the optical generation. Now what about the recombination rate? You might expect the rate to be proportional to the concentration of electrons and holes as follows:

$$R = k(n \times p) \quad (3.23)$$

Simply stated, the chance that an electron and a hole will meet and annihilate should depend on how many free electrons and or holes there are, with some proportionality constant k that might be quite complicated. In other words, it is logical to assume that the rate will increase with the more carriers we have around.

Getting back to thermal equilibrium, if we assume the rate of generation and recombination are balanced, we can say:

$$G = R \quad (3.24)$$

Substituting for the recombination rate, and ignoring the optical generation rate:

$$k(n \times p) = G_{th}(T) \quad (3.25)$$

The right hand side is a function of temperature. We are not concerned with the details of how it varies for now, but we can say it is a constant for a fixed T . We call it n_i^2 because it has the units of concentration of carriers squared:

$$n \times p = G_{th}(T)/k = n_i^2(T) \quad (3.26)$$

Since $n = p$ for an intrinsic (undoped, or pure) silicon crystal, we call it the **intrinsic carrier concentration** n_i . One can calculate the value of n_i using techniques from solid-state physics and the value is approximately given by

$$n_i(T) \cong 10^{10} \text{cm}^{-3} \text{ at } 300\text{K} \quad (3.27)$$

We call this the **law of mass action**:

$$p \cdot n = n_i^2 \quad \text{Law of mass action} \quad (3.28)$$

3.4.2 Generation Statistics

The rate at which carriers are generated is a strong function of the material band-gap E_g (the distance between the valence band and conduction band), because it takes that much energy to promote an electron (or that much energy to break the bond). Since the mechanism for generation is *thermal* (vibrations), the hotter the temperature, the more generation. It can be shown that the rate of generation is given by the following equation (approximately):

$$n_i \propto e^{-E_g/2kT} \quad \text{Intrinsic carrier concentration} \quad (3.29)$$

Recall that kT is proportional to the thermal energy and E_g is the band gap. The above equation says that the number of carriers at an “excess” energy above the valence band depends strongly on how much thermal energy we have in the system.

The origin of this equation can be explained using a simple analogy. Instead of a band gap, imagine a wall. Let’s say that the wall is 6.5 feet tall and anyone with a height over 5.8 feet can hop over the wall. If we take a classroom full of students and plot the distribution of heights, we get a normal distribution. The average height is 5.3 feet and we have marked the local of 5.8, the necessary energy required to get over this wall. If we integrate under the curve to find the total number of students tall enough, we can see that it’s not unreasonable that the integral over the tail of the distribution is approximately exponential as we have claimed. If we raise the barrier height, exponentially fewer students can make it over the wall.

3.5 Doping with Impurities

The term **doping** means that we add impurities to the crystal to change its properties. As we will show, the presence of impurities can have a profound impact on the properties of a semiconductor.

3.5.1 Doping with Group V Elements

Let’s start with a group V element on the periodic table, for example Phosphorous (P) or Arsenic (As). Because these are group V elements, they do not fit naturally in the crystal. That is, unless they get rid of one of their valence electrons. Then they can “fit in” and present to be silicon (group IV) atoms, and form covalent bonds with their neighbors. We are assuming that the number of group V elements is substantially smaller than the number of group IV elements, so they would be more or less always surrounded by silicon atoms. In our bonding model, we can say that the group V element fits into the crystal as an ionized atom by donating an electron to the crystal, as shown in Fig. 3.12. We call these dopants **donors**.

3.5.2 Donor Accounting

Each ionized donor will contribute an extra “free” electron. The material is charge neutral, so the total charge concentration must sum to zero:

$$\rho = \underbrace{-qn_0}_{\text{free electrons}} + \underbrace{qp_0}_{\text{free holes}} + \underbrace{qN_D}_{\text{ions (immobile)}} = 0 \quad \text{Charge neutrality} \quad (3.30)$$

By the Law of Mass-Action ($n \times p = n_i^2(T)$):

$$-qn_0 + q \frac{n_i^2}{n_0} + qN_D = 0$$

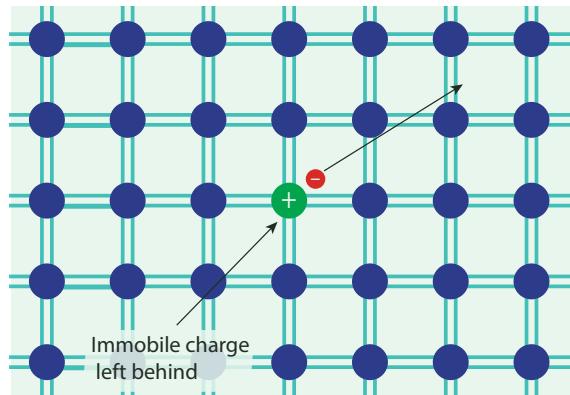


Figure 3.12: When a group V dopant (*donor*) is added to the silicon crystal, it readily gives up one electron in order to form a covalent bond with its neighbors. The extra electron is released and becomes a free carrier.

This leaves us with an equation involving only the density of free electrons:

$$-qn_0^2 + qn_i^2 + qN_D n_0 = 0$$

When we solve the quadratic pretty quickly since it's quadratic:

$$n_0^2 - N_D n_0 - n_i^2 = 0$$

The general solution is given by:

$$n_0 = \frac{N_D \pm \sqrt{N_D^2 + 4n_i^2}}{2} \quad \text{Free electrons in doping Si} \quad (3.31)$$

Only the positive root is physically valid, because it implies that the number of free electrons is larger than N_D . Recall that we know there were n_i free electrons before the dopants, and then we introduced the dopants and every dopant donates a free electron. So why is the free electron concentration not simply $N_D + n_i$? Because, now the rate of recombination is higher, and that is exactly what the solution to the quadratic is telling us. So clearly, only the positive root is physically valid, and we have:

$$n_0 = \frac{N_D + \sqrt{N_D^2 + 4n_i^2}}{2} \quad (3.32)$$

For most practical situations: $N_D \gg n_i$. This means that we can simplify the quadratic to:

$$n_0 = \frac{N_D + N_D \sqrt{1 + 4\left(\frac{n_i}{N}\right)^2}}{2} \approx \frac{N_D}{2} + \frac{N_D}{2} = N_D \quad (3.33)$$

The final result is so obvious you might wonder why we bothered with all the questions. It's true, if you assume the doping concentration is very high, then the thermally generated electrons don't matter and we can just assume that the number of free electrons is the same as the number of dopants.

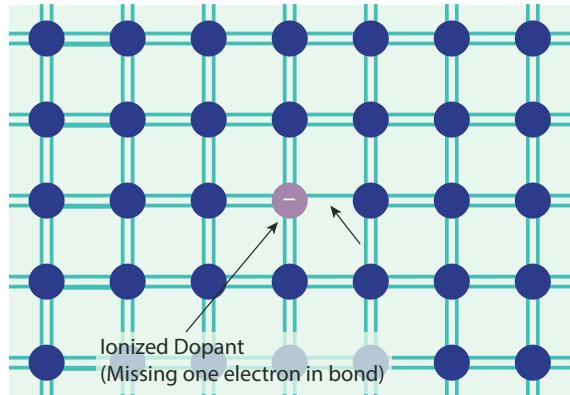


Figure 3.13: When a group III dopant (*acceptor*) is added to the silicon crystal, it needs an electron to form a covalent bond with its neighbors, which it readily accepts from a nearby bond. This electron, coming from another bond, creates a free hole.

3.5.3 Doping with Group III Elements

So we have seen that adding a group V element doping element increases the number of free electrons. How about if we want to increase the number of holes? Well, by symmetry, let's try a group III element. Boron, for example, has three bonding electrons. So if it were surrounded by group IV elements like silicon, it would be sorely missing an electron. What can happen is that another electron can come to the rescue and make Boron a very happy atom. It will be a negatively charged ion (because it has an extra electron attached to it), but it can now form covalent bonds with its neighbors in an orderly fashion. But where does that electron come from? A neighboring atom, for example, can lose its electron, forming a hole, and that hole can now roam about as a free hole. The model is illustrated in *Fig. 3.13*, and shows that the addition of a group III element creates a negatively charged ion and a free hole.

We don not need to repeat the math. We can just use the final result of the group V case, and take note that if the doping concentration is much larger than n_i , then the number of free holes will simply equal to the number of ionized dopants. At normal room temperature this number is more or less all of them.⁶. We call group III dopants as “**acceptors**”, because they accept an electron to form a covalent bond as opposed to group V dopants, which “donate” an electron and are known as “**donors**” as we noted above.

3.5.4 Mass Action-Law (Again)

The balance between generation and recombination means that in any given scenario, the product of the concentration of holes and electrons is constant.

$$p_o \cdot n_o = n_i^2 \quad (3.34)$$

When a process creates holes and electrons asymmetrically, as we have just seen with dopants, then these extra free carriers will drive down the number of free carriers of opposite type due to recombination. Recall that at ($T = 300\text{K}$, $n_i = 10^{10}\text{cm}^{-3}$). Now for the N-type doping case, the concentration of electrons is approximately equal to the number of ionized dopants N_D^+ , or approximately equal to the number of dopants:

$$n_0 = N_D^+ \cong N_D \quad (3.35)$$

⁶Freeze out is a condition you may encounter if you cool a semiconductor. In this scenario, there is not enough energy to ionize the dopants, and the results of this section are no longer valid.

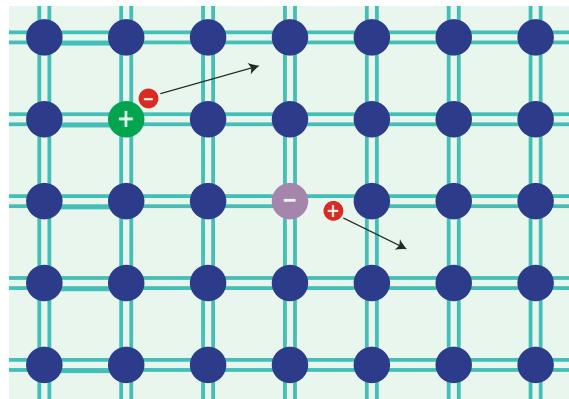


Figure 3.14: Silicon crystal doped by both an acceptor and donor dopant results in the creation of both free holes and electrons.

And that means that the number of holes has been reduced due to the increased presence of electrons:

$$p_0 = \frac{n_i^2}{N_D} \quad (3.36)$$

The same is true for a P-type scenario:

$$p_0 = N_A^- \cong N_A \quad (3.37)$$

These extra holes drive down the concentration of electrons P-type case:

$$n_0 = \frac{n_i^2}{N_A} \quad (3.38)$$

3.5.5 Compensation

What happens if we dope with *both* donors and acceptors? Who wins out? Well, it depends on the doping level. Essentially we create both free electrons for every donor dopant and free holes for every acceptor dopant. Take a look at Fig. 3.14.

If we dope with more donors than acceptors, and $N_D > N_A$, then the material is “N-type”, because the action of the free electrons is more than the free holes. In other words, the free electron density is given by:

$$n_o = N_D - N_A \gg n_i \quad (3.39)$$

And the free holes are approximately given by the law of Mass-Action:

$$p_o = \frac{n_i^2}{N_D - N_A} \quad (3.40)$$

On the other hand, if we have more acceptors than donors, or $N_A > N_D$, then we can approximate the number of free holes by:

$$p_o = N_A - N_D \gg n_i \quad (3.41)$$

And the number of free electrons by:

$$n_o = \frac{n_i^2}{N_A - N_D} \quad (3.42)$$

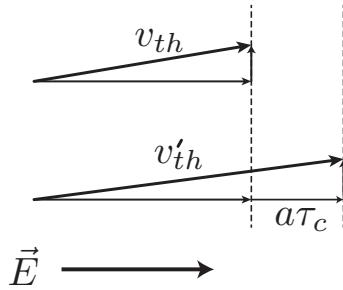


Figure 3.15: The drift velocity will increase by the application of an electric field, on average by an amount $a\tau_c$, where a is the acceleration from the field and τ_c is the time between collisions.

3.6 Drift Currents

Given what we have learned about a semiconductor, we can see that it has some resemblance to the ideal gas model we developed earlier. In particular, we have some concentration of “free” carriers that roam the crystal with thermal energy and can respond to an external field.

3.6.1 Thermal Equilibrium

The rapid, random motion of holes and electrons is quite fast because these particles are so light. The “**thermal velocity**” is approximately $v_{th} = 10^7$ cm/s, and the number of collisions happens on a time scale of approximately $\tau_c = 10^{-13}$ s. The thermal velocity can be approximated by equating the thermal energy with the kinetic energy

$$\frac{1}{2}m_n^*v_{th}^2 = \frac{1}{2}kT \quad (3.43)$$

During this time, the average carrier moves a distance of λ

$$\lambda = v_{th}\tau_c \quad (3.44)$$

Which is approximately

$$\lambda = 10^7 \text{cm/s} \times 10^{-13} \text{s} = 10^{-6} \text{cm} \quad (3.45)$$

(hole case)

3.6.2 Drift Velocity and Mobility

Now, as before, we apply an electric field \mathcal{E} and charge carriers accelerate, at least for τ_c seconds before they collide. As shown in Fig. 3.15, during this time they gain momentum from the field:

$$v_{dr} = a\tau_c = \frac{q\mathcal{E}}{m}\tau_c \quad (3.46)$$

Or in other words, the **drift velocity** is proportional to the field \mathcal{E} :

$$v_{dr} = \frac{q\tau_c}{m}\mathcal{E} = \mu_n\mathcal{E} \quad (3.47)$$

We define a **mobility** for both the electrons, μ_n and holes, μ_p and expect they will be different due to the difference in the effective mass between holes and electrons.

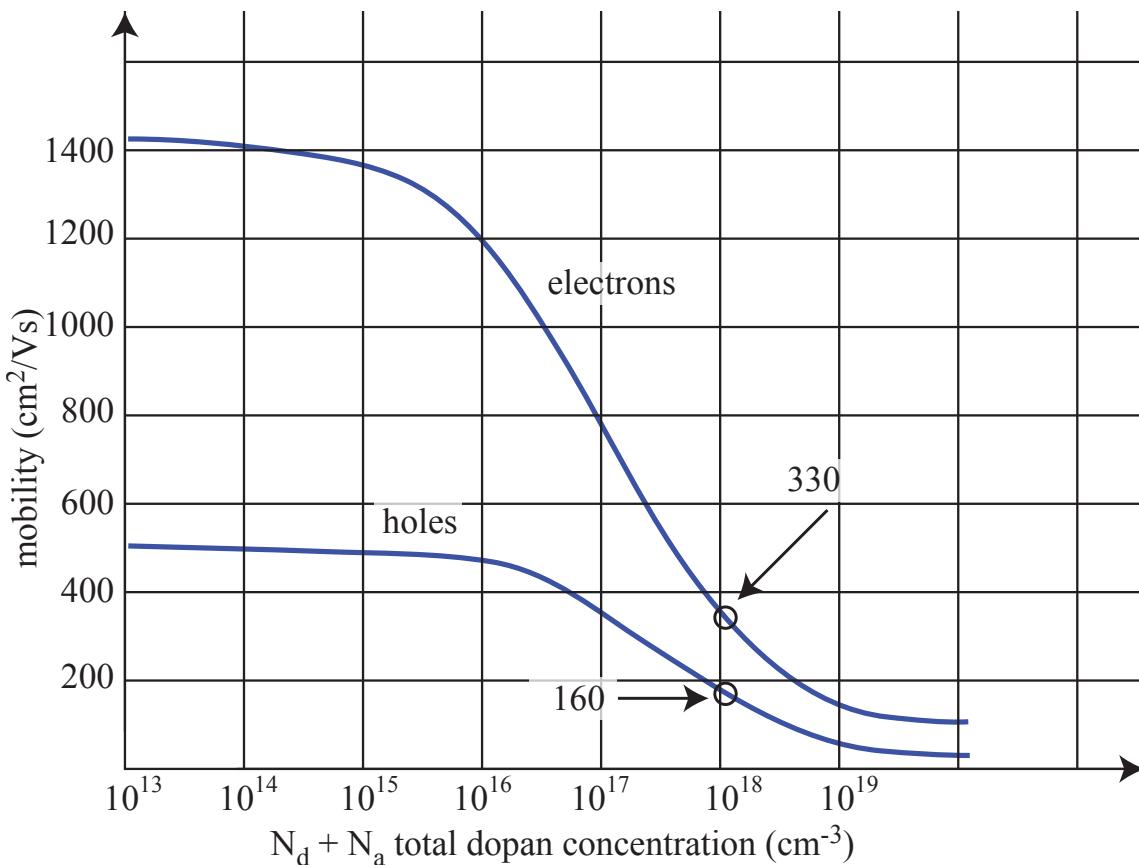


Figure 3.16: Mobility of electrons and holes in silicon as a function of *total* doping concentration.

3.6.3 Mobility vs. Doping in Silicon at 300°K

As you might expect, the mobility should depend on temperature and also on the total dopant concentration, because the more dopants we introduce, the more scattering sites. This is born out by measurement and theory, as shown in *Fig. 3.16*, which shows that the mobility is highest for undoped silicon and then begins to drop as we introduce more and more dopants. A typical values for electrons is about $\mu_n = 1000\text{cm}^2/\text{Vs}$ and for holes $\mu_p = 400\text{cm}^2/\text{Vs}$. Electrons go faster as they are “lighter”.

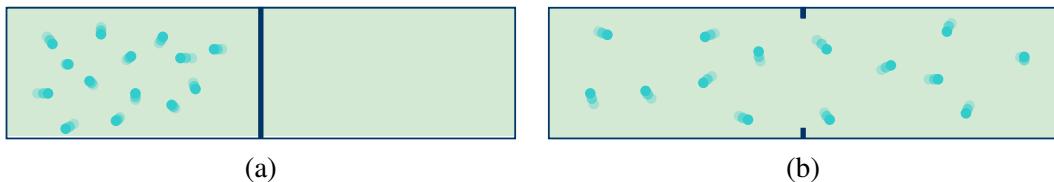


Figure 3.17: (a) Particles moving randomly in a box with a partition are confined to the left partition. (b) When we open the partition, after a brief time we find particles distributed equally in both partitions.

3.7 Diffusion Currents

We are all familiar with diffusion currents in everyday life, for example the fragrance of a bottle of perfume spreads across the room through the random motion of perfume molecules in air. This motion is due to the fact that even though perfume molecules move randomly due to thermal energy, they move in directions from high concentration to low concentration. Why does this happen exactly?

3.7.1 Diffusion

Let's do a simple thought experiment to figure out why **diffusion** flows from high to low concentration. In Fig. 3.17, imagine that we fill the left chamber with a gas at temperature T . If we suddenly remove the divider, what happens? The gas will fill the entire volume of the new chamber. How does this occur?

Let's zoom in to the motion of the molecules (Fig. 3.18). The net motion of gas molecules to the right chamber was due to the concentration gradient. If each particle moves on average left or right with equal probability, then eventually half will be in the right chamber. If the molecules were charged (or electrons), then there would be a net current flow. It is clear that the diffusion current flows from high concentration to low concentration, because in areas of higher concentration there are more particles moving both left and right, which is larger than say the particles moving left and right in the regions of low concentration.

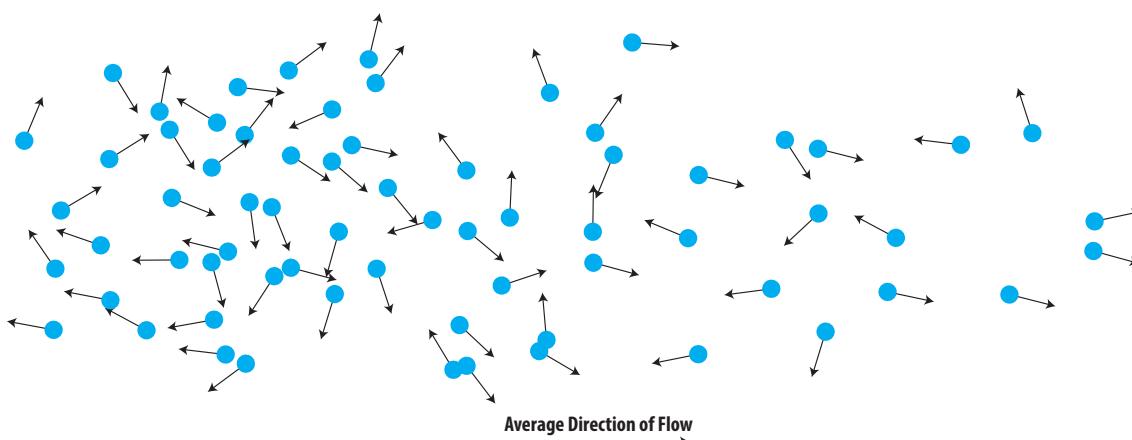


Figure 3.18: Particles in random thermal motion move on average in every direction with equal probability. If there's a concentration gradient, then regions of high concentration contribute more carriers moving towards regions of low concentration compared to particles from low concentrations, thus leading to diffusion.

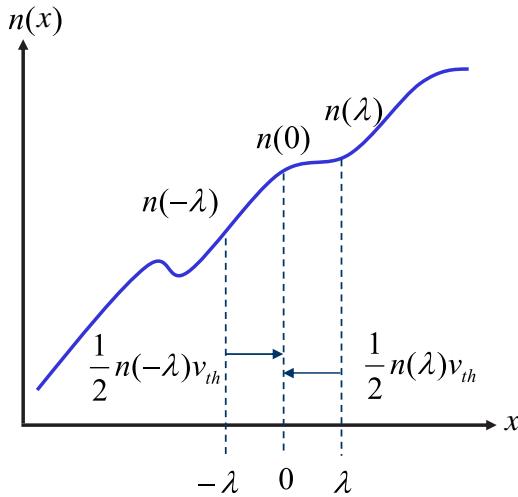


Figure 3.19: The distribution of electron concentration $n(x)$ leads to a diffusion current flow in the direction from high concentration to low concentration. To show this, we consider how many particles cross a given point, $x = 0$ for example, in a unit of time. λ is the average distance moved by particles in thermal motion.

3.7.2 Diffusion Equations

We can find an equation for diffusion currents of electrons if we carefully consider the collective average motion. With reference to *Fig. 3.19*, assume that the **mean free-path** is λ . Let's find the flux of carriers crossing the $x = 0$ plane by imagining how many particles will cross the origin in a time equal to the average inter-collision time. If we move left and right a distance λ , then all particles moving in the right direction (towards the origin), will cross the origin as they travel on average a distance λ . The flux of electrons is therefore the number of particles multiplied by the distance they move

$$dF = \frac{1}{2}v_{th}(n(-\lambda) - n(\lambda)) \quad (3.48)$$

In the above equation, the factor $1/2$ arises because only half the particles from the left move right, and likewise half the particles on the right move left. If the distance λ is small, and n is a smooth function, we can estimate the amount of carriers at each location by a simple Taylor series expansion:

$$F = \frac{1}{2}v_{th}\left(\left[n(0) - \lambda \frac{dn}{dx}\right] - \left[n(0) + \lambda \frac{dn}{dx}\right]\right) \quad (3.49)$$

Which shows that it's only the gradient of the concentration that contributes to current:

$$F = -v_{th}\lambda \frac{dn}{dx} \quad (3.50)$$

Now, since electrons carry charge, the current flows in the opposite direction:

$$J = -qF = qv_{th}\lambda \frac{dn}{dx} \quad (3.51)$$

3.7.3 Einstein Relation

The thermal velocity at a given temperature T is given by:

$$\frac{1}{2}m_n^*v_{th}^2 = \frac{1}{2}kT \quad (3.52)$$

And therefore the mean free path λ is given by:

$$\lambda = v_{th}\tau_c \quad (3.53)$$

The term we are interested in from the diffusion current equation is the product of $v_{th}\lambda$:

$$v_{th}\lambda = v_{th}^2\tau_c = kT \frac{\tau_c}{m_n^*} = \frac{kT}{q} \frac{q\tau_c}{m_n^*} \quad (3.54)$$

This leads to:

$$J = qv_{th}\lambda \frac{dn}{dx} = q \left(\frac{kT}{q} \mu_n \right) \frac{dn}{dx} \quad (3.55)$$

The term in the parenthesis is known as the **diffusion constant**, and it is related to the mobility:

$D_n = \left(\frac{kT}{q} \right) \mu_n$

Einstein relation
(3.56)

3.7.4 Total Current and Boundary Conditions

When both drift and diffusion are present, the total current is given by the sum:

$$J = J_{drift} + J_{diff} = q\mu_n n E + qD_n \frac{dn}{dx} \quad (3.57)$$

In resistors, the carrier is approximately uniform and the second term is nearly zero. For currents flowing uniformly through an interface (no charge accumulation), as shown in *Fig. 3.20*, the field is discontinuous. Nevertheless the current must flow through the interface (unless charge is accumulating):

$$J_1 = J_2 \quad (3.58)$$

This implies that:

$$\sigma_1 E_1 = \sigma_2 E_2 \quad (3.59)$$

So the electric field is discontinuous by the inverse ratio of conductivity:

$$\frac{E_1}{E_2} = \frac{\sigma_2}{\sigma_1} \quad (3.60)$$

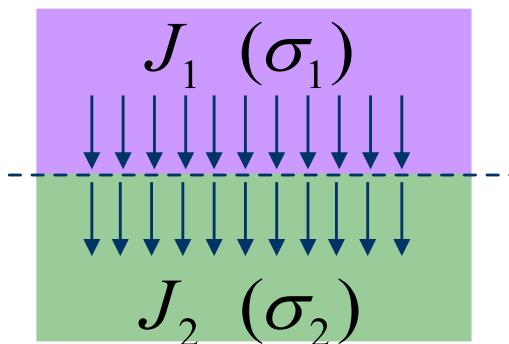
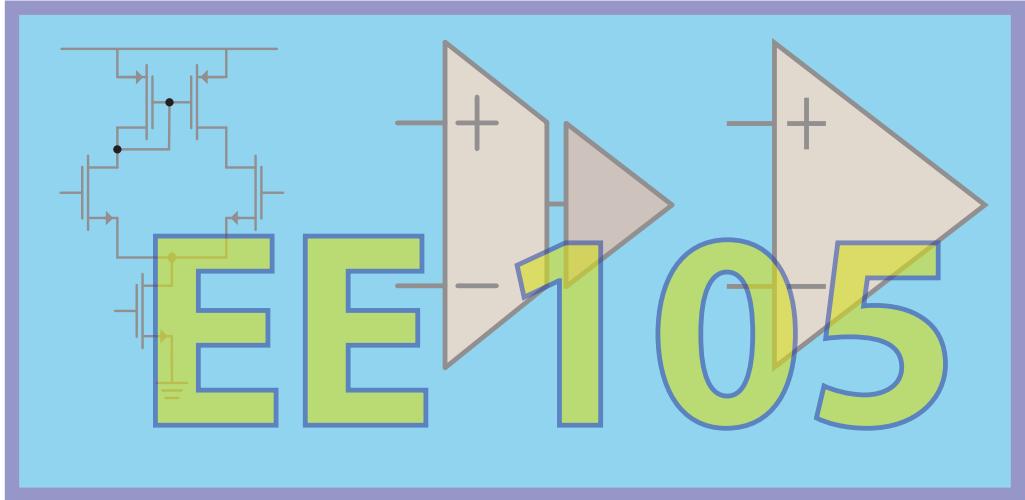


Figure 3.20: Uniform current flowing through a discontinuity in the conductivity must be accompanied by a discontinuity in the electric field.



4. IC Resistors and Capacitors and Electrostatics

4.1 Chapter Preview

In this short chapter we discuss integrated circuit device fabrication, in particular fabrication of resistors. The fabrication steps also apply to capacitors, diodes and transistors, devices which we will cover in the following chapters. Next, we will review electrostatics, a topic of integral importance in the next chapter. Readers familiar with electrostatics can skip or skim these sections. Finally, we discuss capacitors and in particular non-linear capacitors, or devices that have a non-linear charge-voltage characteristic. We shall see in the next two chapters that reverse-biased diodes are essentially non-linear capacitors.

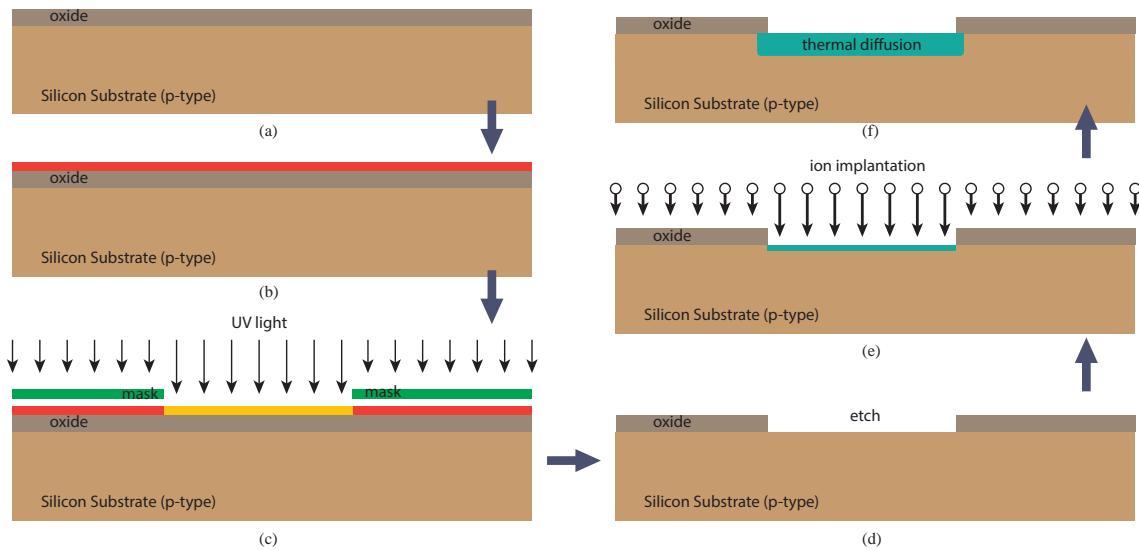


Figure 4.1: The fabrication steps in a typical photolithography process to create a diffusion region in the silicon substrate. The steps include (a) oxide growth and (b) deposition of a photoresist over the surface; (c) selective exposure to UV using a mask; (d) chemical etching to wash away regions exposed to UV; (e) ion-implantation; (f) diffusion to allow implants to diffuse into the structure.

4.2 IC Fabrication: Si Substrate

For most semiconductors, a pure *ingot* of a silicon (*Si*) crystal cut into thin wafers is the starting material for fabrication. The *Si* wafer is extremely pure (~ 1 part in a billion impurities). As we learned in the last chapter, this level of purity is necessary in order to selectively dope the semiconductor to alter its properties. The *Si* crystal structure has a density of about 5×10^{22} , and we typically dope regions of the substrate with doping levels ranging from $10^{14} - 10^{18}$. We want the unintentional dopants (impurities) to be about one to two orders of magnitude less dense $\sim 10^{12}$ in order to realize a high quality device.

Si wafers are polished to about $700 \mu\text{m}$ thick with a mirror finish. The *Si* forms the substrate for the IC, and all the structures are built on the thin films on the top layer of the substrate.

4.2.1 IC Fabrication: Oxide

Si has a native oxide, SiO_2 . SiO_2 (Quartz) is extremely stable and very convenient for fabrication. It is an insulator, so it can be used for house interconnection on top of the substrate (wires interconnecting various devices for example). It can also be used for selective doping by using SiO_2 to block doping in regions where we don't want to dope, and by chemically etching openings or windows using **photolithography** (described below) to dope regions. These openings allow ion implantation into selected regions, and SiO_2 can block ion implantation in other areas.

4.2.2 IC Fabrication: Ion Implantation

The steps involved in ion implantation are shown in Fig. 4.1. We start with a *Si* substrate. It is usually pre-doped to be *P*-type, but this depends on the details of the process. We grow a layer of oxide (thermally) by heating the substrate and flowing oxygen onto the surface. Next, we cover the surface with **photoresist**, a material that changes its chemical properties based on exposure to light. We use photolithography to selectively expose regions of the photoresist using a UV source and a lithography mask. Next, we chemically etch (using an acid such as *HF*) the surface, removing

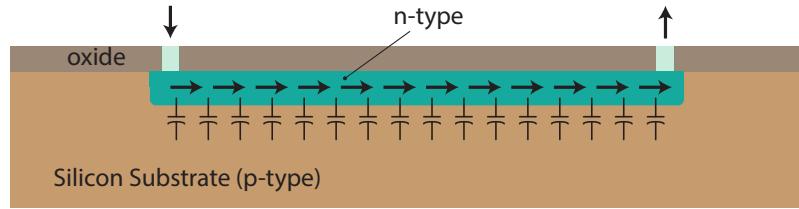


Figure 4.2: Cross-section of an IC diffusion resistor in a *P*-type region is made with an *N*-type diffusion region and contacts.

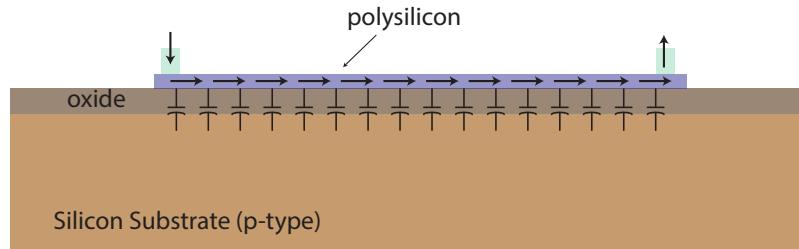


Figure 4.3: Cross-section of a polysilicon resistor is made with a thin film of material on top of the oxide.

areas exposed to UV light.¹ Finally, we introduce ion implantation (*N*-type), whereby we inject dopants onto the entire surface of the wafer. Regions with the SiO_2 are protected and block the ions, whereas the window openings allow the ions to be implanted into the silicon substrate. Next, a higher temperature is used to allow the dopants to diffuse into the substrate and to form bonds with the silicon crystal.

4.3 IC Resistors

4.3.1 “Diffusion” Resistor

Using ion implantation and diffusion, we can create a so-called "**diffusion resistor**", shown in *Fig. 4.2*. The thickness and dopant concentration of resistor is set by process steps, whereas the shape of the resistor is set by design (layout). Metal contacts are connected to ends of the resistor. You may be wondering why current would flow through a thin diffusion layer rather than through the silicon substrate, which is comparatively much larger (wider and thicker). The reason is that the resistor is DC isolated from the substrate, and only high frequency AC currents can flow from the diffusion region into the substrate. In other words, there is a distributed capacitor from the resistor to the substrate, which is usually connected to ground potential. This happens because the diffusion region is *N*-type and the substrate is *P*-type. This ***PN-junction*** should be **reverse biased** at all times (meaning the diffusion resistor should never be biased at a negative potential with respect to the substrate). We will learn the details of *PN*-junctions in the next chapter.

4.3.2 Poly Film Resistor

To lower any **parasitic capacitance**, we should build the resistor further away from substrate, as shown in *Fig. 4.3*. This is done by depositing a thin film of “poly” *Si* (heavily doped) material on top of the oxide, rather than directly on the substrate. This thin film resistor is not made of the crystalline form of silicon, but rather it is a **polycrystalline** form, which is a material made with many fragments of crystals, of varying sizes and arranged in different orientations. For our

¹This kind of photoresist is "positive" because exposure to light chemically alters the material so that it reacts with an acid. There's also "negative" photoresist materials too that don't react with the acid when exposed.

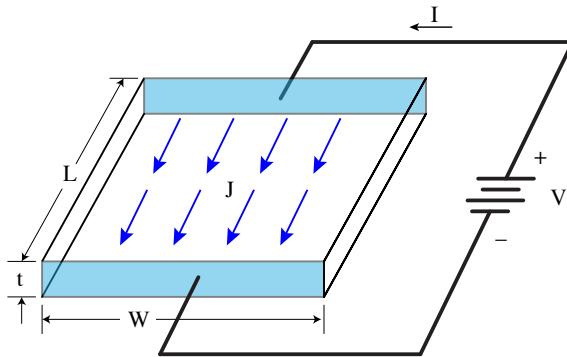


Figure 4.4: A slab of silicon (or any material) forms a resistor and the current flows due to drift. The fields inside the material arise due to the applied external voltage.

purposes, the material behaves similar to a crystal and we can model it as a doped semiconductor. The process technology details set the thickness and resistance of this layer (say $10\Omega/\square$, see Section 4.3.4)

4.3.3 Ohm's Law

With reference to Fig. 4.4, consider the current I in terms of the current density J :

$$I = JA = JtW = (\sigma\mathcal{E})tW$$

The voltage V dropped across the resistor can be written in terms of electric field: $E = V/L$. Putting all of this together, we have:

$$I = \left(\sigma \frac{V}{L}\right)tW = \left(\frac{\sigma t W}{L}\right)V = \frac{1}{R}V$$

This is nothing but Ohm's Law, with the material resistance given by:

$$R = \frac{L}{W} \frac{1}{\sigma t} \quad \text{Material resistance} \quad (4.1)$$

4.3.4 Sheet Resistance (R_\square)

The resistance of a rectangular slab of material can be written in a way as to distinguish process specific parameters (such as doping levels and material thickness) from layout (geometry) parameters, such as length and width:

$$R = \frac{\text{resistivity}}{\text{thickness}} = \frac{\rho L}{Wt} = \left(\frac{\rho}{t}\right)\left(\frac{L}{W}\right)$$

Since IC resistors have a specified fixed thickness and resistivity – something not under the control of the circuit designer – we lump these parameters into a constant R_\square :

$R = R_\square \left(\frac{L}{W}\right)$

Material resistance in terms of sheet resistance parameter (4.2)

Notice that R_\square is a convenient way to capture the process specific parameters. This term is known as the **sheet resistance**. Even though it has units of resistance, or Ω , we emphasize that it is a sheet resistance by the symbol \square , since R_\square is the resistance of a square ($L = W$ implies $L/W = 1$).

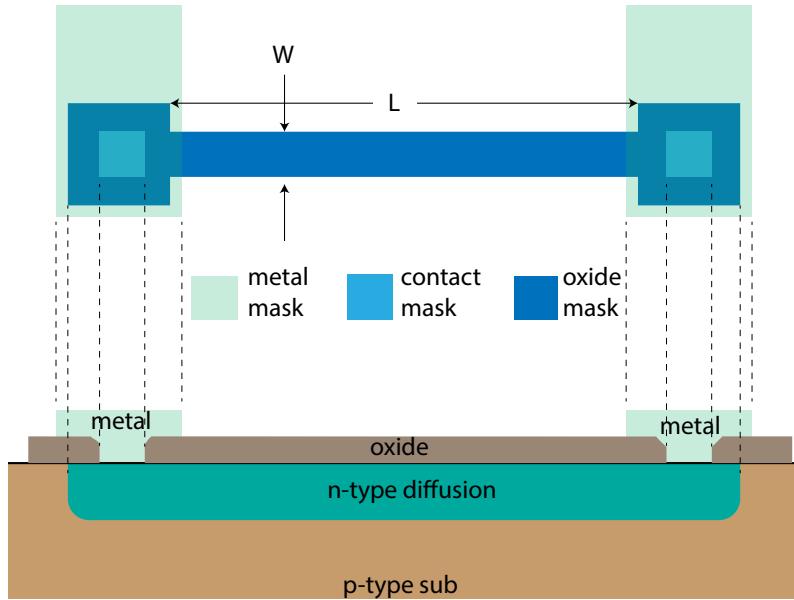


Figure 4.5: Layout and cross section of a resistor using a "dog bone" layout to introduce contacts on either side which define the terminals of the resistor.

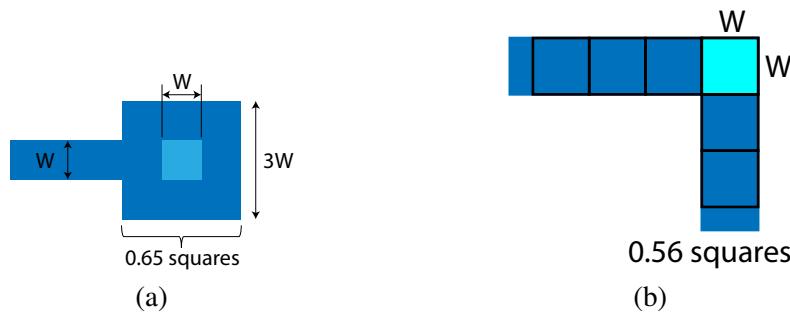


Figure 4.6: (a) Due to non-uniform current flow, a contact region contributes approximately only 0.65 squares of resistance. (b) Similarly, current turning a corner only contributes 0.56 squares of resistance.

4.3.5 Using Sheet Resistance (R_{\square})

Consider the ion-implanted (or “diffused”) IC resistor shown in *Fig. 4.5*. We show both the cross-section and the top view. To figure out the resistance, we just count the number of “squares” from end to end and multiply by R_{\square} . On the other hand, we know that there are regions where the current density J “turns” or spreads, and so not all of square contains equal current. Contact regions also have non-uniform current distribution as the current spreads out from a small via opening to the entire width of the resistor at one end, and then the inverse process occurs at the other end. To account for the non-uniform current flow, we find that some simple rules of thumb can be applied. As shown in *Fig. 4.6*, when turning corners, only about half (0.56 squares) of the metal contributes resistance, since some of the current will flow on the “inner track” and little current flows on the “outer track”. Likewise, when current spreads out from a via, only about 65% of the region contributes resistance, and a good rule of thumb for “dog bone” layout resistors is to count the contact region as 0.65 squares.

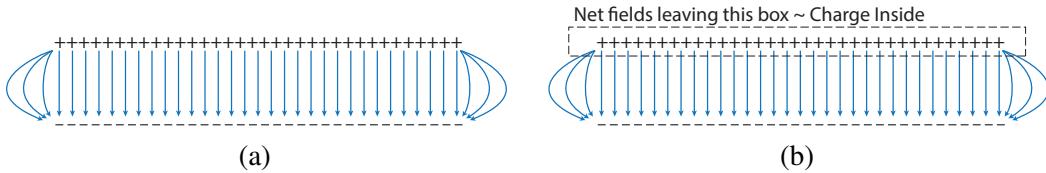


Figure 4.7: (a) Electric field lines leave positive charges and end on negative charges. (b) If we draw a box around the positive charges, we will find that the net integral of the electric field is positive and outward on the surface of the box.

4.4 Review of Electrostatics

4.4.1 Electrostatics Review

Electric fields go from positive charge to negative charge (by convention). Electric field lines *diverge* and *converge* on charge (see Fig. 4.7). This is captured in **Poisson's equation** below:

$$\nabla \cdot \mathcal{E} = \frac{\rho}{\epsilon} \quad \text{Poisson's equation} \quad (4.3)$$

In words, if the electric field changes magnitude, then there has to be some charge involved. The result is that in a charge free region, the electric field must be constant!

Gauss' law equivalently says that if there is a *net* electric field leaving a region, there has to be positive charge in that region (see Fig. 4.7(b)). If we integrate the electric field over a closed surface, then the result will be the net charge inside:

$$\oint \mathcal{E} \cdot dS = \frac{Q}{\epsilon} \quad \text{Gauss' Law, closed surface} \quad (4.4)$$

We can derive this from the divergence of charge by integrating over the volume of the closed surface. Integration of the charge density over the volume is the total charge:

$$\oint_V \nabla \cdot \mathcal{E} dV = \oint_V \frac{\rho}{\epsilon} dV = \frac{Q}{\epsilon} \quad (4.5)$$

The volume integral can be converted into a surface integral using the **divergence theorem**:

$$\oint_V \nabla \cdot \mathcal{E} dV = \oint_S \mathcal{E} \cdot dS = \frac{Q}{\epsilon} \quad (4.6)$$

4.4.2 Electrostatics in 1D

Fortunately, most of the problems we will be solving are idealizations in one dimension, and everything simplifies in 1-D. The divergence operator is a simple derivative:

$$\nabla \cdot \mathcal{E} = \frac{d\mathcal{E}}{dx} = \frac{\rho}{\epsilon} \quad (4.7)$$

Which is equivalent to saying that the incremental electrical field is related to an increment of charge:

$$d\mathcal{E} = \frac{\rho}{\epsilon} dx \quad (4.8)$$

If we integrate over the region of interest, we have the total electric field:

$$\mathcal{E}(x) = \mathcal{E}(x_0) + \int_{x_0}^x \frac{\rho(x')}{\epsilon} dx' \quad (4.9)$$

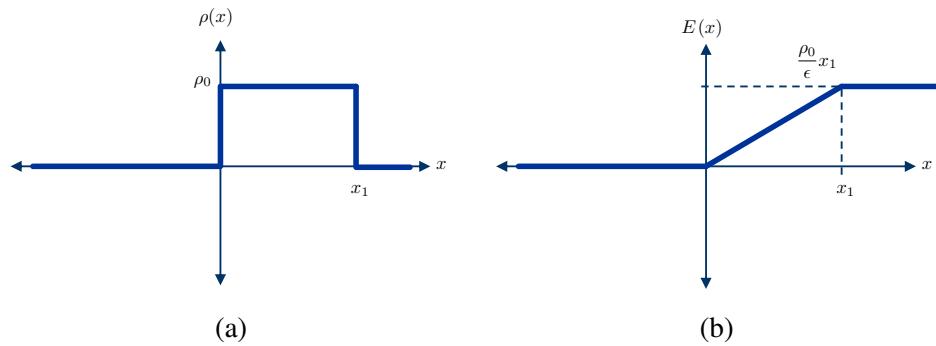


Figure 4.8: (a) In this hypothetical example, the charge density is uniform over a region from the origin to x_1 . We assume that the field is zero at $x = 0$. (b) Integration of the charge density yields the electric field, which increases linearly up to x_1 , and remains constant thereafter.

As an application of these equations, one that will become important in the next chapter, consider a uniform charge distribution as shown in *Fig. 4.8(a)*. Integrating the charge density, we get a linear function. The field increases linearly, as shown in *Fig. 4.8(b)*.

$$\mathcal{E}(x) = \int_0^x \frac{\rho(x')}{\varepsilon} dx' = \frac{\rho_0}{\varepsilon} x \quad (4.10)$$

4.4.3 Electrostatic Potential

The electric field (force) is related to the potential (energy):

$$\mathcal{E} = -\frac{d\varphi}{dx} \quad (4.11)$$

The negative sign says that field lines go from high potential points to lower potential points, and it takes work to push a positive charge against the field:

$$F_e = q\mathcal{E} = -e \frac{d\phi}{dx} \quad (4.12)$$

Note that as shown in *Fig. 4.9(a)*, an electron should “float” to a high potential point since it has negative charge.

$$\varphi(x) - \varphi(x_0) = - \int_C \mathcal{E} \cdot d\vec{l} \quad (4.13)$$

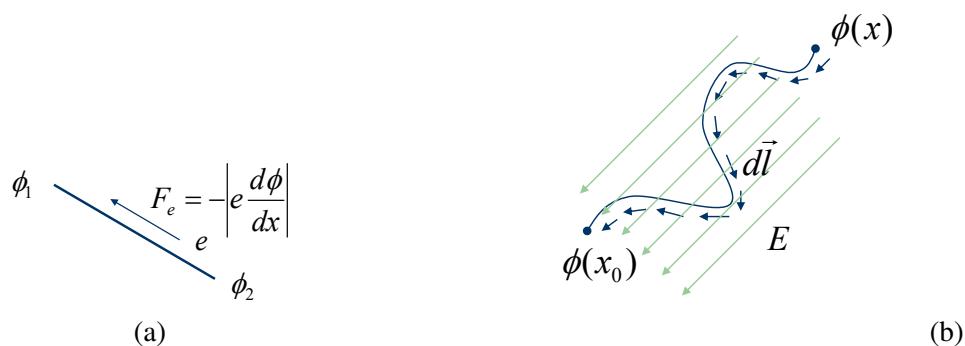


Figure 4.9: (a) The force experienced by an electron is related to the gradient of the potential from point 1 to point 2. (b) The work to move a charge against a field is the line integral of the the path over the field, with only paths parallel to the field contributing to the integral.

The potential is the integral of the field, as shown in *Fig. 4.9(b)*. The integral is path independent for static fields because only the path in the direction of the field contributes, and motion perpendicular to the field does not require any energy expenditure (there are no forces in perpendicular directions). In one dimension, this is a simple integral:

$$\varphi(x) - \varphi(x_0) = - \int_{x_0}^x \mathcal{E}(x') dx' \quad (4.14)$$

Going the other way, we have Poisson's equation in 1D:

$$\frac{d^2\varphi(x)}{dx^2} = -\frac{\rho(x)}{\epsilon} \quad (4.15)$$

4.4.4 Boundary Conditions

Potential must be a continuous function. If not, the fields (forces) would be infinite. Electric fields need not be continuous. We have already seen that the electric fields diverge on charges. In fact, across an interface shown in *Fig. 4.10* we apply Gauss' Theorem:

$$\oint \epsilon \mathcal{E} \cdot dS = -\epsilon_1 \mathcal{E}_1 S + \epsilon_2 \mathcal{E}_2 S = Q_{inside} \quad (4.16)$$

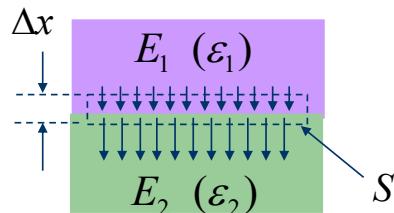


Figure 4.10: The boundary conditions at the junction of two materials is determined by applying Guass' Law to a pillbox shaped region surrounding the border, and letting the height of the pillbox Δx diminish.

We have assumed that the field is uniform across the interface. As the "pillbox" around the interface becomes smaller and smaller, the charge inside must be zero:

$$Q_{inside} \xrightarrow{\Delta x \rightarrow 0} 0 \quad (4.17)$$

Simplifying the equations, we have :

$$-\epsilon_1 \mathcal{E}_1 S + \epsilon_2 \mathcal{E}_2 S = 0 \quad (4.18)$$

Or:

$$\frac{\mathcal{E}_1}{\mathcal{E}_2} = \frac{\epsilon_2}{\epsilon_1} \quad (4.19)$$

The field is discontinuous across the boundary, which implies charge density at the surface! The surface charge arises due to the different levels of electric polarization of the dielectric materials (since ϵ is different).

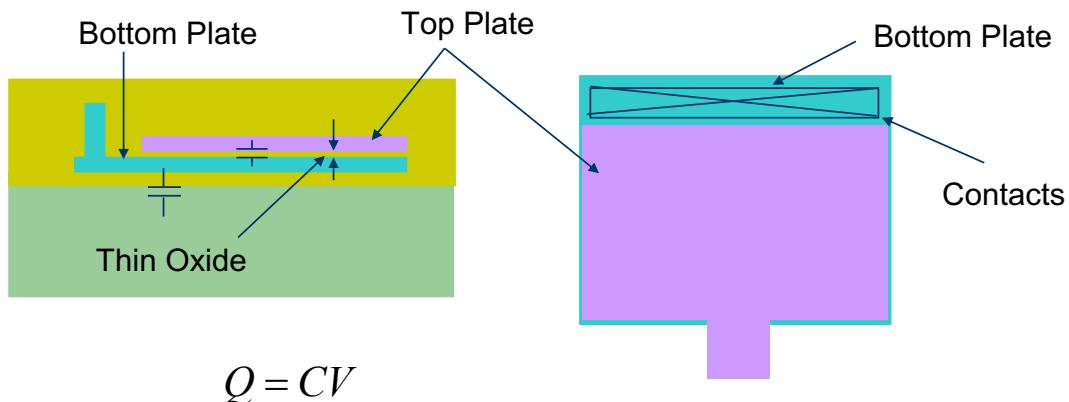


Figure 4.11: Integrated circuit capacitors are made by overlapping two large plates, made from interconnect metallization or using polysilicon and doped silicon regions. The latter types of capacitors are known as MOS capacitors and will be covered in Chapter 7.

4.5 IC Capacitors

4.5.1 MIM (Metal-Insulator-Metal) Capacitor

By forming a thin oxide and metal (or polysilicon) plates, an **MIM capacitor** is formed (*Fig. 4.11*). Metal plates make a higher quality capacitor, but for certain applications, polysilicon gates are sufficient. Contacts are made to top and bottom plates. Parasitic capacitance exists between bottom plate and substrate.

4.5.2 Review of Capacitors

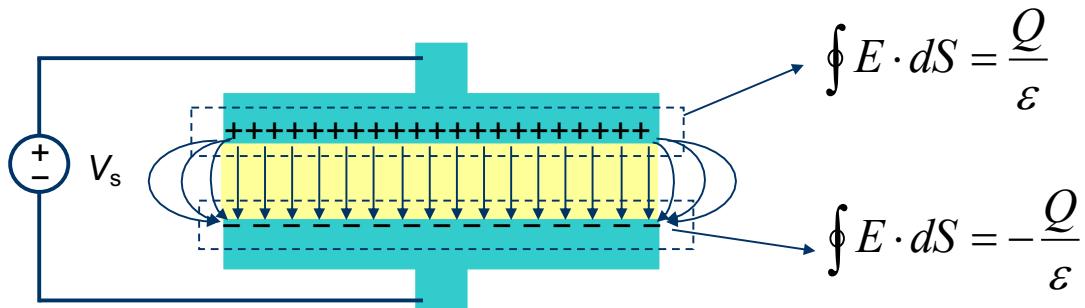


Figure 4.12: The charges on a parallel plate capacitor are necessarily equal since the region between the plates is devoid of charge (insulator).

Armed with electrostatics, we can analyze capacitors in detail. For example, it is easy to show that the charge on the top and bottom plate are equal (see *Fig. 4.12*). This is true since the field is uniform in the charge free insulator, thus the same fields that leave the top plate must enter the bottom plate.

If we integrate the field from the bottom plate to the top plate, we obtain the applied voltage on the capacitor. The field is constant, so we have:

$$\int \mathcal{E} \cdot dl = \mathcal{E}_0 t_{ox} = V_s \longrightarrow \mathcal{E}_0 = \frac{V_s}{t_{ox}} \quad (4.20)$$

By Gauss' Law, we can relate the field to the charge on the plate:

$$\oint \mathcal{E} \cdot dS = \mathcal{E}_0 A = \frac{Q}{\epsilon} \longrightarrow \frac{V_s}{t_{ox}} A = \frac{Q}{\epsilon} \longrightarrow Q = CV_s \quad (4.21)$$

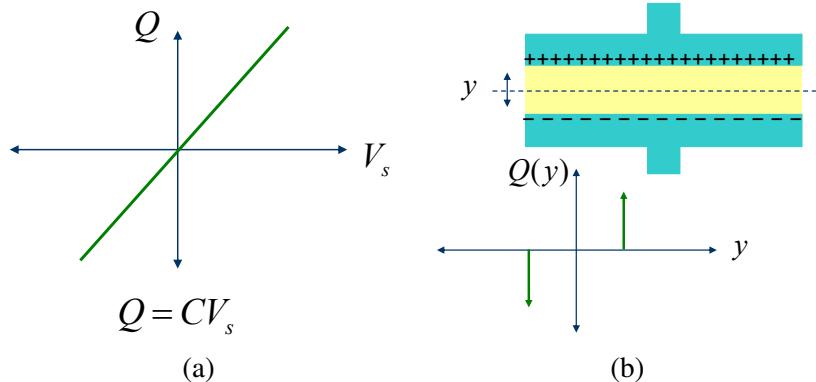


Figure 4.13: (a) An ideal capacitor constructed with perfect metals has a linear Q - V relation and (b) all the charges reside at the surface of the plates.

This leads to the well known charge-voltage relationship shown in Fig. 4.13(a):

$$C = \frac{A \epsilon}{t_{ox}} \quad \text{Capacitance between two parallel plates} \quad (4.22)$$

For an ideal capacitor constructed with metal plates, all charge must be at surface of the plates. Recall that the field inside of an ideal metal is zero, so that means all the charge is at the surface. If we plot the charge as a function of position, then we obtain two delta functions, because of the infinite concentration of charge (see Fig. 4.13(b)).

4.5.3 A Non-Linear Capacitor

We will soon meet capacitors that have a non-linear Q-V relationship. If plates are not ideal metal, the charge density can penetrate into surface. For example, suppose the charge is uniformly distributed through the thickness of the plates as shown in *Fig. 4.14*.

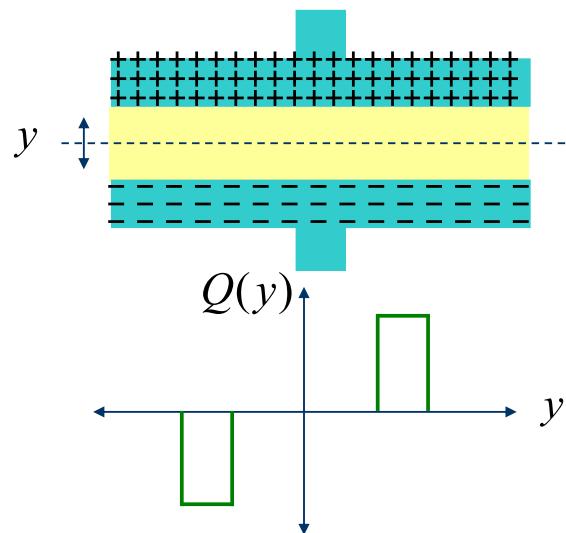


Figure 4.14: A capacitor structure whereby the charges are distributed uniformly along the plates, something similar to the depletion region that we will study in the next chapter.

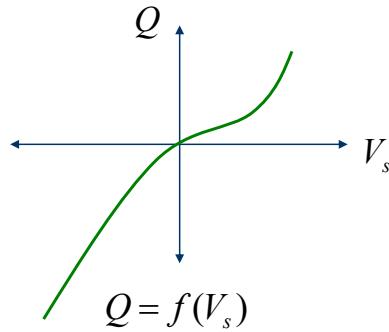


Figure 4.15: A hypothetical non-linear capacitor as is defined by its $Q = f(V)$ relation.

For a **non-linear capacitor**, we have a general relationship between the charge and the voltage (see *Fig. 4.15*):

$$Q = f(V_s) \neq CV_s \quad (4.23)$$

We can't identify a capacitance like a linear $Q - V$ capacitor. But now imagine that we apply a small signal on top of a bias voltage. If the signal is very small, we can do a Taylor Series expansion:

$$Q = f(V_s + v_s) \approx f(V_s) + \frac{df(V)}{dV} \Big|_{V=V_s} \cdot v_s \quad (4.24)$$

$f(V_s)$ is a constant amount of charge. The incremental charge is therefore:

$$Q = Q_0 + q \approx f(V_s) + \frac{df(V)}{dV} \Big|_{V=V_s} \cdot v_s \quad (4.25)$$

4.5.4 Small Signal Capacitance

To find the "small-signal capacitance", break the equation for total charge into two terms:

$$Q = Q_0 + q \approx f(V_s) + \frac{df(V)}{dV} \Big|_{V=V_s} v_s \quad (4.26)$$

$$q = \frac{df(V)}{dV} \Big|_{V=V_s} v_s = Cv_s \quad (4.27)$$

$$C \equiv \frac{df(V)}{dV} \Big|_{V=V_s} \quad (4.28)$$

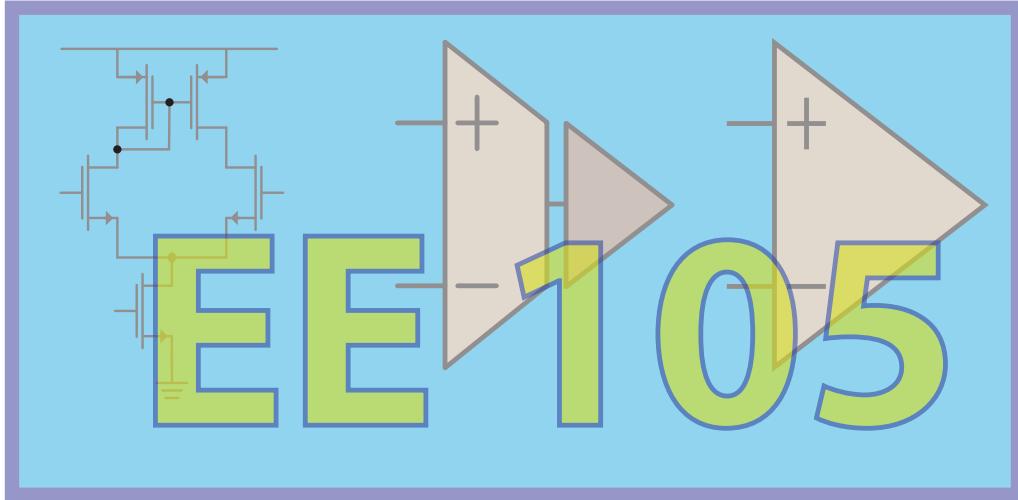
4.5.5 Example of Non-Linear Capacitor

We will see that for a *PN*-junction, the charge is a function of the reverse bias:

$$Q_j(V) = -qN_Ax_p \sqrt{1 - \frac{V}{\varphi_b}} \quad (4.29)$$

The small signal capacitance is therefore given by:

$$C_j(V) = \frac{dQ_j}{dV} = \frac{qN_Ax_p}{2\varphi_b} \frac{1}{\sqrt{1 - \frac{V}{\varphi_b}}} = \frac{C_{j0}}{\sqrt{1 - \frac{V}{\varphi_b}}} \quad (4.30)$$



5. PN Junctions in Equilibrium

5.1 Chapter Preview

We will see in this chapter and the following that the simple *PN*-junction structure provides a rich variety of functionality. It can be used as a function of bias, acting as a "one way" gate for current, rectifying AC signals and allowing AC-to-DC conversion, converting light photons into electrical energy (solar cell), and performing the opposite functionality of creating light from electricity in a very efficient manner in light-emitting diodes (LEDs). This simple structure is also the foundation of lasers and photo-detectors, used to transport energy around the globe using thin fibers of glass, imaging inside of the body, detecting x-rays and other high energy particles, among many other functions. Diodes are also used to detect extremely high frequency signals, demodulating information in communication receivers, or used as switches to allow a transmitter and receiver to share the same antenna. It is no exaggeration that the diode is one of the most versatile circuit elements. To understand these myriad of applications, we must first discover the properties of the junction under a forward and reverse bias. In this chapter we focus on the reverse bias, which is simpler to analyze.

We begin by considering that a concentration variation in carriers across a junction (such as a massive amount of electrons in an *N*-region compared to a dearth of electrons in a *P*-region) leads to a built-in potential difference between these regions. We apply this to a *PN*-junction at thermal equilibrium to find its properties, including its ability to store charge as a non-linear capacitor. In the next chapter, we will see how the diode conducts current.

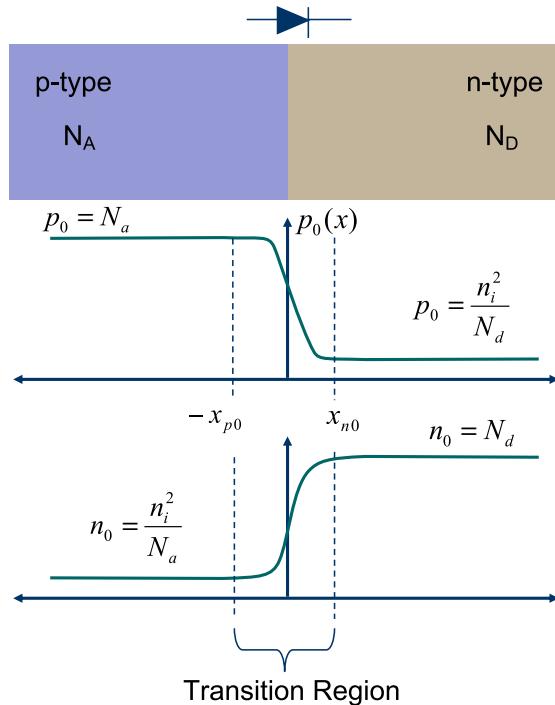


Figure 5.1: A *PN*-junction is the interface between a *P*-type and *N*-type doped semiconductor. There are large concentration gradients between the two sides, with orders of magnitude more holes on the *P*-side compared to the *N*-side. The same is true of electrons on the *N*-side. Even though the junction is abrupt, we shall see that the concentration must change smoothly over a "transition region" from one region to the other.

5.2 Structure

As shown in *Fig. 5.1*, a *PN*-junction is simply the junction between two regions doped as *N*-type and *P*-type. When terminals are added to both sides, it is known as a *PN*-junction diode, or just a diode.

5.3 Carrier Concentration and Potential

In **thermal equilibrium**, under the assumption that there are no external fields, we expect the electron and hole current densities to be zero. Recall that there might be a lot of motion due to thermal energy, but on average the currents should be zero. If there is a steady-state concentration gradient that is somehow maintained, then that implies there is a constant diffusion current, which must be canceled by a corresponding drift current. For the electrons we have:¹

$$J_n = 0 = qn_0\mu_n\mathcal{E}_0 + qD_n \frac{dn_o}{dx} \quad (5.1)$$

This equality between drift and diffusion implies that there must be a non-zero field \mathcal{E}_0 . Since the electric field is the gradient of potential, we also see that there is a corresponding change in potential due to the concentration gradient:

$$\frac{dn_o}{dx} = -\left(\frac{\mu_n}{D_n}\right)n_o\mathcal{E}_0 = \left(\frac{q}{kT}\right)n_o \frac{d\varphi_0}{dx} \quad (5.2)$$

¹Why not the sum of the electron and hole currents?

We have used Einstein's relation $kT/q = D/\mu$ to simplify the equation. The differential in voltage is related to the differential in concentration:

$$d\varphi_0 = \left(\frac{kT}{q} \right) \frac{dn_o}{n_0} = V_{th} \frac{dn_o}{n_0} \quad (5.3)$$

We have an equation relating the potential to the carrier concentration. Integrating this relation:

$$\varphi_0(x) - \varphi_0(x_0) = V_{th} \ln \frac{n_0(x)}{n_0(x_0)} \quad (5.4)$$

This shows that the potential difference is related to the logarithm of the concentration difference, with $V_{th} = kT/q$ as the scaling constant (about 26 mV at room temperature).

It is customary (and confusing) to define the potential reference to be intrinsic Si, or in other words, if an *N*-type region forms a junction with intrinsic silicon, it will develop a potential φ derived from this equation:

$$n = n_i e^{\varphi_0(x)/V_{th}} \quad (5.5)$$

$$\varphi_n = V_{th} \ln \left(\frac{n}{n_i} \right) \quad (5.6)$$

If we do a similar calculation for holes, we need to carefully account for the fact that there are fewer electrons (n_p) than intrinsic silicon (n_i), making the potential negative:

$$\varphi_p = V_{th} \ln \left(\frac{n_p}{n_i} \right) \quad (5.7)$$

$$\varphi_p = V_{th} \ln \left(\frac{n_i^2}{p \cdot n_i} \right) \quad (5.8)$$

or:

$$\varphi_p = V_{th} \ln \left(\frac{n_i}{p} \right) = -V_{th} \ln \left(\frac{p}{n_i} \right) \quad (5.9)$$

Since this is an equilibrium situation, the Law of Mass Action should hold. With these definitions of potential, we can express the free carrier density as n_i times an exponential factor that takes the potential of the *P*-type or *N*-type region into account. As expected, the Law of Mass Action is upheld:

$$n_0(x)p_0(x) = n_i^2 e^{-\varphi_0(x)/V_{th}} e^{\varphi_0(x)/V_{th}} = n_i^2 \quad (5.10)$$

The Doping Changes Potential

Due to the logarithmic nature of the potential, the potential changes linearly for exponential increase in doping:

$$\varphi_0(x) = V_{th} \ln \frac{n_0(x)}{n_i(x_0)} = 26 \text{ mV} \cdot \ln \frac{n_0(x)}{n_i(x_0)} \approx 26 \text{ mV} \cdot \ln 10 \log \frac{n_0(x)}{10^{10}} \quad (5.11)$$

Since working with factors of ten is convenient when dealing with doping levels, we can convert these equations into a base 10 logarithm:

$$\varphi_0(x) \approx 60\text{mV} \log \frac{n_0(x)}{10^{10}} \quad (5.12)$$

$$\varphi_0(x) \approx -60\text{mV} \log \frac{p_0(x)}{10^{10}} \quad (5.13)$$

For example, to quickly calculate the potential of a *P*-type region with a concentration of 10^{16}cm^{-3} holes, use the fact that we have 6 orders of magnitude more holes than intrinsic, so the potential is $6 \times 60\text{mV} = -360\text{mV}$. A negative voltage means we are dealing with a *P*-type material. An *N*-type materials has a positive potential with respect to intrinsic Si.

5.4 PN-Junction in Equilibrium

5.4.1 PN-Junction: Overview

As illustrated in *Fig. 5.2*, it is useful to do a thought experiment and consider what happens before we reach **equilibrium**. Consider when the junction shown in *Fig. 5.1* is first formed. Due to the abrupt nature of the junction, we expect very large diffusion currents to flow. Mobile charges transfer near the junction, especially majority carriers diffuse in large numbers. The currents are large due to the large concentration gradients. There are orders of magnitude more electrons in the *N*-type region, and orders of magnitude more holes in the *P*-type region. So it is natural to assume that these carriers will diffuse to the other side, becoming minority carriers. However, since these mobile carriers become minority carriers in the new region, they will not penetrate far due to recombination.

The key observation is that initially the two regions are charge neutral. But as holes and electrons are charge carriers, as they cross the junction, they disturb the delicate charge balance and introduce net charge into each region. As a result, a voltage difference builds up between regions. This voltage gets larger and larger as more mobile carriers diffuse across the barrier. This creates a field at the junction that causes drift currents to oppose the diffusion current. This will ensue until the field is sufficiently large to produce a drift current that can counter the diffusion current. In thermal equilibrium, drift current and diffusion must balance each other on average. We will also show that near the junction a "transition region" develops so that the density of holes $p(x)$ and electrons $n(x)$ change smoothly from one side to the other.

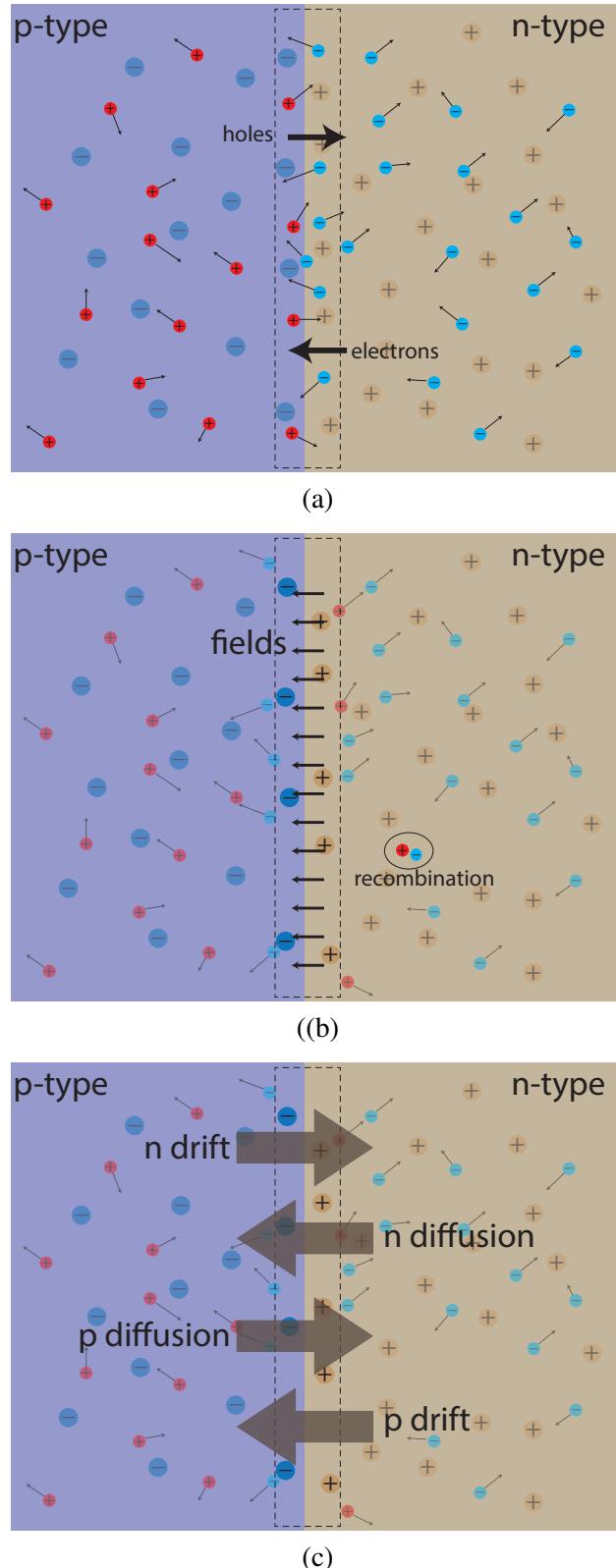


Figure 5.2: Here we outline the non-equilibrium situation when a contact between the N-type and P-type regions are formed. (a) Diffusion currents flow due to the large concentration gradients. (b) Charge imbalance leads to a potential barrier between the two regions. (c) This internal *built-in* potential leads to an equilibrium state with zero net current, whereby diffusion currents are balanced by drift currents.

5.4.2 PN-Junction Currents

Considering the *PN*-junction in thermal equilibrium, the currents have to be zero. So, we have for the electrons the same set of equations we derived earlier:

$$J_n = 0 = qn_0\mu_n\mathcal{E}_0 + qD_n \frac{dn_o}{dx} \quad (5.14)$$

$$qn_0\mu_n\mathcal{E}_0 = -qD_n \frac{dn_o}{dx} \quad (5.15)$$

Now let's focus on the electric field, rather than the potential:

$$\mathcal{E}_0 = \frac{-D_n \frac{dn_o}{dx}}{n_0 \mu_n} = -\frac{kT}{q} \frac{1}{n_0} \frac{dn_o}{dx} \quad (5.16)$$

We can also state the same applies for the holes:

$$\mathcal{E}_0 = \frac{D_p \frac{dp_o}{dx}}{p_0 \mu_p} = \frac{kT}{q} \frac{1}{p_0} \frac{dp_o}{dx} \quad (5.17)$$

Since the fields must be finite, the concentration of electrons and holes cannot change abruptly. Therefore, a smooth variation is expected in the "transition region". This is shown schematically in *Fig. 5.3*.

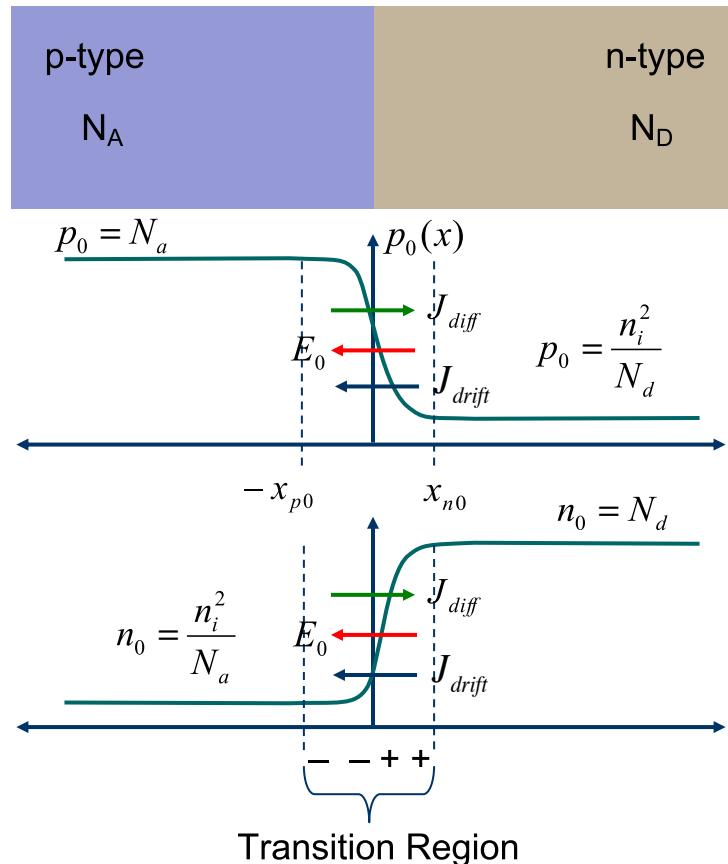


Figure 5.3: The concentration of electrons and holes changes smoothly from one region to the other. The net charge at the boundary "space charge" region leads to an internal electric field which produces a drift current that balances the equilibrium diffusion current.

5.4.3 PN-Junction Fields

With reference to *Fig. 5.3*, let's define the transition region width as x_{p_0} and x_{n_0} , where the first subscript denotes the region type, and the second subscript emphasizes thermal equilibrium. The **transition regions** are defined as the region near the junction where the majority carrier concentrations deviate from the equilibrium values for an isolated doped semiconductor. To solve for the electric fields, we need to write down the charge density in each region:

$$\rho_0(x) = q(p_0 - n_0 + N_D - N_A) \quad (5.18)$$

On the *P*-side of the junction, there are very few electrons and only acceptors:

$$\rho_0(x) \approx q(p_0 - N_A) \quad \text{for } -x_{p_0} < x < 0 \quad (5.19)$$

Since the hole concentration is decreasing on the *P*-side, the net charge is negative:

$$N_A > p_0 \quad (5.20)$$

and so

$$\rho_0(x) < 0 \quad (5.21)$$

Analogous to the *P*-side, the charge on the *N*-side is given by:

$$\rho_0(x) \approx q(-n_0 + N_D) \quad \text{for } 0 < x < x_{n_0} \quad (5.22)$$

The net charge here is positive since:

$$N_D > n_0 \quad (5.23)$$

So we have:

$$\rho_0(x) > 0 \quad (5.24)$$

These facts are summarized in *Fig. 5.3*. We see a field arising that begins with the positive charges (*N*-side), and ends on the negative charges (*P*-side) inside the transition region. By definition, the transition region is therefore the region where the charge density is non-zero. Note that normal *N*-type and *P*-type materials are charge neutral, because ionized dopant charges are neutralized by an equal number of free carriers. In the transition region, the free carriers are depleted (due to diffusion) and so there is net charge.

5.4.4 "Exact" Equation for Fields

Given the above approximations, we now have an expression for the **charge density** in each region of the semi-conductor:

$$\rho_0(x) \cong \begin{cases} q(n_i e^{-\varphi_0(x)/V_{th}} - N_A) & , -x_{p_0} < x < 0 \\ q(N_D - n_i e^{\varphi_0(x)/V_{th}}) & , 0 < x < x_{n_0} \end{cases} \quad (5.25)$$

We also have the following result from electrostatics:

$$\frac{d\mathcal{E}_0}{dx} = -\frac{d^2\varphi}{dx^2} = \frac{\rho_0(x)}{\epsilon_s} \quad (5.26)$$

Notice that the potential φ appears on both sides of the equation, making it difficult to solve the problem. We can find a much simpler equation to solve if we make a simplifying assumption about the charge in the transition region, known as the depletion approximation.

5.4.5 Depletion Approximation

The **depletion approximation** is simply that all free carriers are "depleted" in the transition region, making the entire region consist of ionized dopants. Under this assumption, the charge density expressions are independent of the potential, and are constant:

$$\rho_0(x) \cong \begin{cases} -qN_A & , -x_{p_0} < x < 0 \\ (+)qN_D & , 0 < x < x_{n_0} \end{cases} \quad (5.27)$$

This implies the rate of change of the electric field is constant:

$$\frac{d\mathcal{E}_0}{dx} = \frac{\rho_0(x)}{\epsilon_s} \quad (5.28)$$

The solution for electric field is now easy and simply given by the integration of the charge over the region of interest:

$$\mathcal{E}_0(x) = \int_{-x_{p_0}}^x \frac{\rho_0(x')}{\epsilon_s} dx' + \mathcal{E}_0(-x_{p_0}) \quad (5.29)$$

Is the depletion approximation reasonable? Note that in equilibrium, doped silicon is charge neutral because for every ionized dopant, there is a free carrier (of opposite sign). As we derived earlier, the number of free carriers diminishes *exponentially* as we go into the depletion region. The potential voltage changes across the junction, a result we derived from stating that in equilibrium the net current is zero. The potential change is in fact the potential barrier that must arise as to counter the diffusion of carriers across the *PN*-junction. From the *N*-side (higher potential) we move to the *P*-side, and the potential drops by $\sim .7V$ (depending on the doping). For every $60mV$ drop in potential, there is a $10\times$ reduction in free carriers. Since the free carrier concentration starts at the doping level and drops by $10\times$ for the first $60mV$ drop, we see that neglecting the free carrier concentration is actually quite a reasonable assumption to make and should only result in a very small error in the following calculations.

Returning to the solution of the problem, since charge density is a constant:

$$\mathcal{E}_0(x) = \int_{-x_{p_0}}^x \frac{\rho_0(x')}{\epsilon_s} dx' = -\frac{qN_A}{\epsilon_s}(x + x_{p_0}) \quad (5.30)$$

If we start from the *N*-side we get the following result:

$$\mathcal{E}_0(\overbrace{x_{n_0}}) = \int_x^{x_{n_0}} \frac{\rho_0(x')}{\epsilon_s} dx' + \mathcal{E}_0(x) = \frac{qN_D}{\epsilon_s}(x_{n_0} - x) + \mathcal{E}_0(x) \quad (5.31)$$

or:

$$\mathcal{E}_0(x) = -\frac{qN_D}{\epsilon_s}(x_{n_0} - x) \quad (5.32)$$

Since we impose that $\mathcal{E}_0(x_{n_0}) = 0$ outside of the transition region. This is also reasonable because in the neutral *PN*-junction regions, the fields are zero under equilibrium (no external fields are applied either).

Plot of Fields In Depletion Region

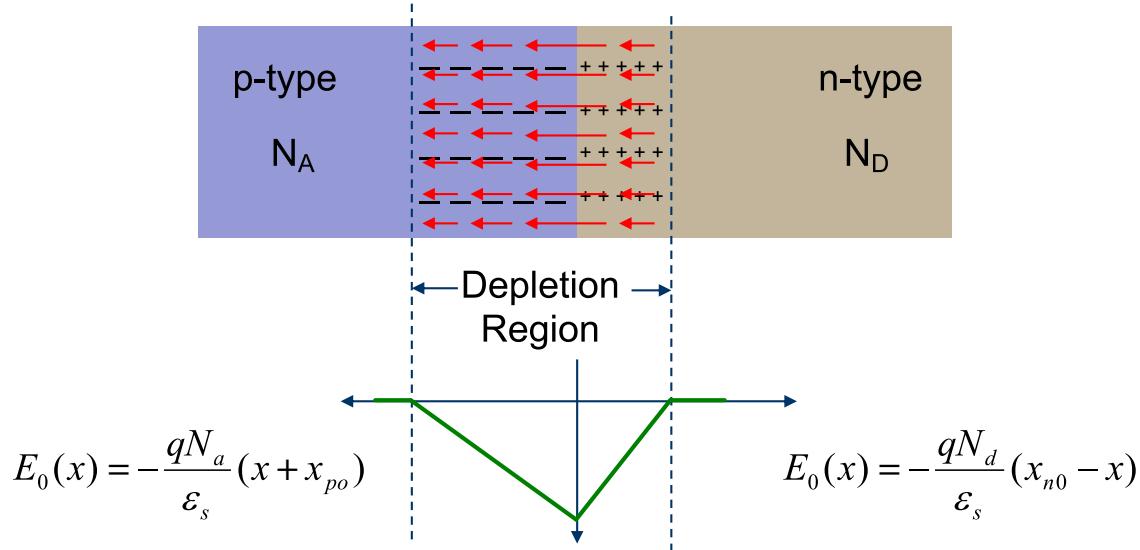


Figure 5.4: The electric fields in the depletion region are negative (they point left since the positive charges are on the right and the negative charges are on the left). Here we assume zero fields in the neutral P-region and N-region. The peak fields occur at the junction.

In Fig. 5.4 we plot the electric fields across the PN-junction. Note that the electric field is zero outside of the depletion region, and the electric field grows uniformly as we enter the depletion region. If we start on the P-side, the fields gradually grow from zero to a peak value at the interface of the region. Then as we enter the N-side, the electric field grows back down to zero as we enter the neutral N-region.

There are several noteworthy facts to take away from this plot. First the width of the depletion region is not necessarily symmetric between the N-side and P-side. Also, the slope of the electric field is larger in one region than the other. Finally the negative peak of the field always occurs at the junction. Why is that?

All of these facts are explained by the difference in doping levels between the N-type and P-type side. The region with the higher doping will have a smaller depletion region because it can support more charge per unit volume. Also, the peak is always at the junction or transition point because fields originate because of charge. From Gauss' law we know that if we were to draw an imaginary box around the P-region, with one face of the box intersecting the neutral P-region, and the other face crossing the depletion region, then as we move the second face to the right into the P-region, we are covering more and more charge inside the box, causing the fields crossing the boundary to be necessarily larger. This is true until the boundary crosses the origin and enters the N-region. Now the amount of charge begins to decrease because we have both positive and negative charges in the box. Finally, when the box reaches the neutral N-region, it must contain exactly zero charge, or stated another way, the amount of charge uncovered in the depletion regions must be equal and opposite in magnitude.

Continuity of the Electric Field Across the Junction

The electric fields diverges on charge. For a sheet charge at the interface, the electric field could be discontinuous. In our case, the depletion region is only populated by a background density of fixed charges so the electric field is continuous. In other words, across the interface we have:

$$\mathcal{E}_0^n|_{x=0} = -\frac{qN_A}{\epsilon_s}x_{po} = -\frac{qN_D}{\epsilon_s}x_{no} = \mathcal{E}_0^p|_{x=0} \quad (5.33)$$

Which is just another way of saying:

$$qN_A x_{p_0} = qN_D x_{n_0} \quad (5.34)$$

Or in words, the total fixed charge in the N -region equals the total fixed charge in the P -region. This is of course something we already anticipated from Gauss' Law.

Potential Across Junction

From our earlier calculation we know that the potential in the N -region is higher than the P -region. We also noted that the potential has to smoothly transition from high to low in crossing the junction. We know that the physical origin of the potential difference is due to the charge transfer that occurred due to the concentration gradient. Let's integrate the field to get an analytical expression for the potential:

$$\varphi(x) = \varphi(-x_{p_0}) + \int_{-x_{p_0}}^x \frac{qN_A}{\epsilon_s} (x' + x_{p_0}) dx'$$

This integral is easy to perform:

$$\varphi(x) = \varphi_p + \frac{qN_A}{\epsilon_s} \left(\frac{x'^2}{2} + x' x_{p_0} \right) \Big|_{-x_{p_0}}^x$$

We arrive at potential on P -side:

$$\varphi_0^p(x) = \varphi_p + \frac{qN_A}{2\epsilon_s} (x + x_{p_0})^2$$

Also performing the integral on N -side:

$$\varphi_0^n(x) = \varphi_n - \frac{qN_D}{2\epsilon_s} (x - x_{n_0})^2$$

The potential *must* be continuous at interface (field finite at interface)

$$\varphi_0^n(0) = \varphi_n - \frac{qN_D}{2\epsilon_s} x_{n_0}^2 = \varphi_p + \frac{qN_A}{2\epsilon_s} x_{p_0}^2 = \varphi_0^p(0)$$

Solve for Depletion Lengths

We have two equations and two unknowns. We are finally in a position to solve for the depletion depths:

$$\varphi_n - \frac{qN_D}{2\epsilon_s} x_{n_0}^2 = \varphi_p + \frac{qN_A}{2\epsilon_s} x_{p_0}^2$$

$$qN_A x_{p_0} = qN_D x_{n_0}$$

The solutions to these equations is given by:

$$x_{n_0} = \sqrt{\frac{2\epsilon_s \varphi_{bi}}{qN_D} \left(\frac{N_A}{N_A + N_D} \right)}$$

Depletion width on N-side (5.35)

and:

$$x_{p_0} = \sqrt{\frac{2\epsilon_s \varphi_{bi}}{qN_A} \left(\frac{N_D}{N_A + N_D} \right)}$$

Depletion width on P-side (5.36)

Where the total potential difference between the N -type and P -type side is **built-in potential**:

$$\varphi_{bi} \equiv \varphi_n - \varphi_p > 0 \quad (5.37)$$

Sanity Check

Does the above equation make sense? Let's say we dope one side very highly. Then physically we expect the depletion region width for the heavily doped side to approach zero:

$$x_{n_0} = \lim_{N_D \rightarrow \infty} \sqrt{\frac{2 \epsilon_s \varphi_{bi}}{q N_D} \frac{N_A}{N_D + N_A}} = 0 \quad (5.38)$$

Also for the depletion region on the *P*-side, the width should be independent of the doping level N_D :

$$x_{p_0} = \lim_{N_A \rightarrow \infty} \sqrt{\frac{2 \epsilon_s \varphi_{bi}}{q N_A} \frac{N_D}{N_D + N_A}} = \sqrt{\frac{2 \epsilon_s \varphi_{bi}}{q N_A}} \quad (5.39)$$

We see that the entire depletion width is dropped across *P*-region, as expected.

5.4.6 Total Depletion Width

The sum of the depletion widths is the “space-charge region”:

$$X_{dep_0} = x_{p_0} + x_{n_0} = \sqrt{\frac{2 \epsilon_s \varphi_{bi}}{q} \left(\frac{1}{N_A} + \frac{1}{N_D} \right)} \quad \text{Total depletion region width} \quad (5.40)$$

By definition, this region is essentially depleted of all mobile charge. Due to high electric field, carriers move across region at velocity saturated speeds. For example:

$$X_{dep_0} = \sqrt{\frac{2 \epsilon_s \varphi_{bi}}{q} \left(\frac{1}{10^{15}} \right)} \approx 1 \mu \quad (5.41)$$

implies:

$$\mathcal{E}_{pn-junction} \approx \frac{1V}{1\mu} = 10^4 \frac{V}{cm} \quad (5.42)$$

The expression for the built-in potential is given by:

$$\varphi_{bi} \equiv \varphi_n - \varphi_p = V_T \left(\ln \left(\frac{N_D}{n_i} \right) + \ln \left(\frac{N_A}{n_i} \right) \right)$$

$$\varphi_{bi} = V_T \ln \left(\frac{N_D N_A}{n_i^2} \right) \quad \text{Built-in potential} \quad (5.43)$$

In Eq. 5.43, V_T (also often symbolized as ϕ_T) is known as **thermal voltage**, and it is equivalent to $kT/q \approx 26mV$. Plugging in some typical numerical values for doping, we arrive at:

$$\varphi_{bi} = 26mV \ln \frac{N_D N_A}{n_i^2} = 60mV \times \log \frac{10^{15} 10^{15}}{10^{20}} = 600mV \quad (5.44)$$

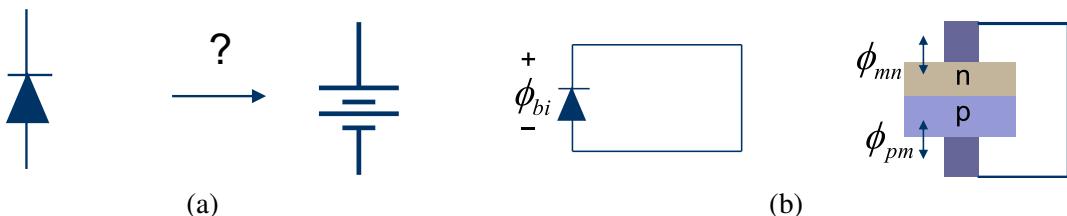


Figure 5.5: (a) One may question if a *PN*-junction is like a battery that can supply power to a circuit. (b) If we short circuit a *PN*-junction, no current flows because the contact potentials between the metal and the semiconductor perfectly balances the built-in potential.

5.4.7 Contact Potential

Have we invented a battery?

Can we harness the *PN*-junction and turn it into a battery (see *Fig. 5.5(a)*). It seems that we have a perpetual motion machine because our solution to the problem so far does not require any input energy (light for example), and yet we have a potential barrier that maybe can be harnessed to do work. Why is it that the built-in potential is not a battery?

Unfortunately this cannot be the case. When we attach a wire to the *PN*-junction, we actually are replaying the same physics that we just described. For example, a metal is kind of like an *N*-type region, but with very heavy doping. So it will form a built-in potential, called the **contact potential**, between the *P*-type and *N*-type regions. Let's call the contact potential with the *N*-type region as φ_{mn} and the contact potential with the *P*-type region as φ_{mp} .

If we short circuit the *PN*-junction (*Fig. 5.5(b)*), then going around the loop we have a KVL loop:

$$\varphi_{bi} + \varphi_{mn} + \varphi_{pm} = 0 \quad (5.45)$$

We can re-write this as:

$$\varphi_{bi} = \varphi_{mn} - \varphi_{pm} \quad (5.46)$$

This shows that the net built-in potential between the *NP*-region is the same, whether we use the metal potential as the reference or intrinsic silicon as the reference as we did earlier.

Stated another way, a certain amount of energy is required to take an electron from the semiconductor and put it in the metal, and then back into the semiconductor. This energy must exactly balance out the built-in potential barrier between the N and P regions, because we are simply providing another pathway for carriers to diffuse across the barrier. If this pathway did not present a potential barrier, then we would not be in thermal equilibrium and there would be a net current flow across the junctions from the wire.

Contact Potential Between Materials

In summary, the contact between a *PN*-junction creates a potential difference. Likewise, the contact between two dissimilar metals creates a potential difference.² When a metal semiconductor junction is formed, a contact potential forms as well. In fact, forming a good "ohmic" contact between a semiconductor and a metal requires special engineering that we usually ignore in an elementary coverage of *PN*-junctions.

²Proportional to the difference between the **work functions**.

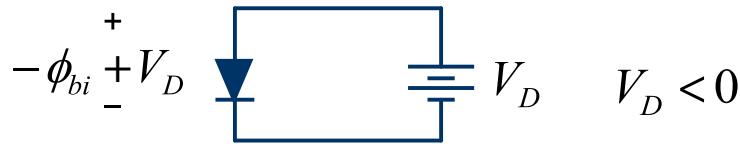


Figure 5.6: A voltage V_D is applied across the PN -junction. When the voltage $V_D < 0V$, we say the junction is reverse biased.

5.4.8 PN-Junction Capacitor

Under thermal equilibrium, the PN -junction does not draw any current. But notice that a PN -junction stores charge in the space charge region (transition region). Since the device is storing charge, it is acting like a capacitor. Positive charge is stored in the N -region, and negative charge is stored in the P -region:

$$-Q_p = qN_A x_{p0} = qN_D x_{n0} = (+)Q_n \quad (5.47)$$

5.5 Reverse Biased PN Junction

What happens if we “reverse-bias” the PN -junction, as shown in Fig. 5.6? Since no current is flowing, the entire reverse biased potential is dropped across the transition region.³ To accommodate the extra potential, the charge in these regions must increase. If no current is flowing, the only way for the charge to increase (decrease) is to grow (shrink) the depletion regions.

5.5.1 Voltage Dependence of Depletion Width

We can redo the math, but in the end we realize that the equations are the same except that we replace the built-in potential with the effective reverse bias ($V_D < 0V$). The depletion regions *grow* (this is how we check our sign convention):

$$x_n(V_D) = \sqrt{\frac{2\epsilon_s(\varphi_{bi} - V_D)}{qN_D} \left(\frac{N_A}{N_A + N_D} \right)} = x_{n0} \sqrt{1 - \frac{V_D}{\varphi_{bi}}} \quad (5.48)$$

and:

$$x_p(V_D) = \sqrt{\frac{2\epsilon_s(\varphi_{bi} - V_D)}{qN_A} \left(\frac{N_D}{N_A + N_D} \right)} = x_{p0} \sqrt{1 - \frac{V_D}{\varphi_{bi}}} \quad (5.49)$$

So the total depletion width is given by:

$$X_{dep}(V_D) = x_p(V_D) + x_n(V_D) = \sqrt{\frac{2\epsilon_s(\varphi_{bi} - V_D)}{q} \left(\frac{1}{N_A} + \frac{1}{N_D} \right)} \quad (5.50)$$

We can simplify this equation and write the depletion region width in terms of the zero bias case, X_{dep_0} , and the applied reverse bias voltage:

$$X_{dep}(V_D) = X_{dep_0} \sqrt{1 - \frac{V_D}{\varphi_{bi}}} \quad (5.51)$$

³As we will see in the next chapter, a very small amount of current will flow under reverse-bias conditions.

5.5.2 Charge Versus Bias

As we increase the reverse bias, the depletion region grows to accommodate more charge ($V_D < 0V$):

$$Q_j(V_D) = -qN_A x_p(V_D) = -qN_A x_{p_0} \sqrt{1 - \frac{V_D}{\varphi_{bi}}} \quad (5.52)$$

More charge is stored as we increase the applied voltage, but charge is *not* a linear function of voltage. This is a non-linear capacitor that we alluded to in the previous chapter. We can define an incremental "**small signal**" **capacitance** for small signals by breaking up the charge into two terms:

$$Q_j(V_D + v_D) = Q_j(V_D) + q(v_D) \quad (5.53)$$

In the above equation the voltage v_D is a small incremental voltage, $|v_D/V_D| \ll 1$.

5.5.3 Derivation of Small Signal Capacitance

We perform a Taylor Series expansion about a fixed operating point V_D :

$$Q_j(V_D + v_D) = Q_j(V_D) + \left. \frac{dQ_j}{dV} \right|_{V_D} v_D + \dots \quad (5.54)$$

The linear term looks like a regular linear capacitor. This term is the small-signal capacitance:

$$C_j = C_j(V_D) = \left. \frac{dQ_j}{dV} \right|_{V=V_D} = \left. \frac{d}{dV} \left(qN_A x_{p_0} \sqrt{1 - \frac{V}{\varphi_{bi}}} \right) \right|_{V=V_R} \quad (5.55)$$

$$C_j = \frac{qN_A x_{p_0}}{2\varphi_{bi} \sqrt{1 - \frac{V_D}{\varphi_{bi}}}} = \frac{C_{j0}}{\sqrt{1 - \frac{V_D}{\varphi_{bi}}}} \quad (5.56)$$

The capacitance at zero bias can be written as:

$$C_{j0} = \frac{qN_A x_{p_0}}{2\varphi_{bi}} = \frac{qN_A}{2\varphi_{bi}} \sqrt{\left(\frac{2\epsilon_s \varphi_{bi}}{qN_A} \right) \left(\frac{N_D}{N_A + N_D} \right)} = \sqrt{\frac{q\epsilon_s}{2\varphi_{bi}} \frac{N_A N_D}{N_A + N_D}} \quad (5.57)$$

5.5.4 Physical Interpretation of Depletion Cap

Notice that the expression on the right-hand-side of Eq. 5.57 is almost just the (inverse) of the depletion width in thermal equilibrium. A simple manipulation shows that:

$$C_{j0} = \epsilon_s \sqrt{\frac{q}{2\epsilon_s \varphi_{bi}} \left(\frac{1}{N_A} + \frac{1}{N_D} \right)^{-1}} = \frac{\epsilon_s}{X_{dep_0}} \quad (5.58)$$

This looks like a parallel plate capacitor, with capacitance per unit area $C = \epsilon/d$, where d is the distance between the plates. Is that a coincidence? No, it is actually related to the fact that the incremental charge comes from the edges of the depletion region (as it grows into the N -type and P -type regions). So the new incremental charges are separated by a distance of X_{dep_0} , as if there were plates of metal at the edges of the depletion region. The depletion region itself looks like a dielectric of permittivity of ϵ_s (for silicon $\epsilon_s = 11.7\epsilon_0$. With this observation, we can state that the small-signal incremental capacitance for an applied reverse-bias voltage V_D is given by:

$$C_j(V_D) = \frac{\epsilon_s}{X_{dep}(V_D)} \quad (5.59)$$

5.5.5 A Variable Capacitor (Varactor)

A good application of the reverse-biased *PN*-junction is a "varactor", or a variable capacitor. This is used in almost every radio to tune an oscillator to a frequency for transmission and reception of radio waves (for example in a modulator and demodulator). This is true from a humble AM or FM radio to the most sophisticated modem used for wireless data communication. As shown in *Fig. 5.7*, the capacitance varies in a non-linear way with applied bias V_c . For modern frequency synthesizers, this is not an issue because the applied voltage on the *PN*-junction is derived from a feedback loop that ensures the oscillator is generating zero crossings at a rate consistent with the desired output frequency.⁴

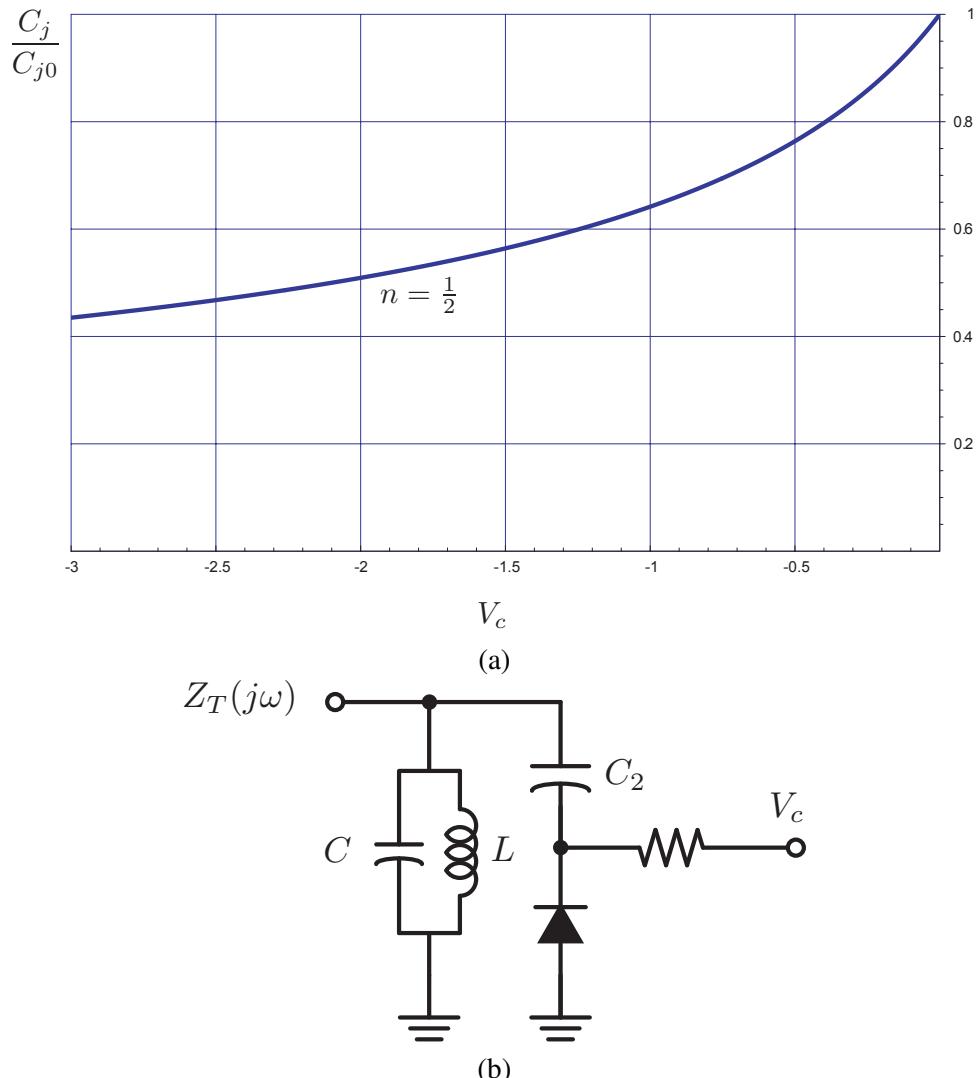
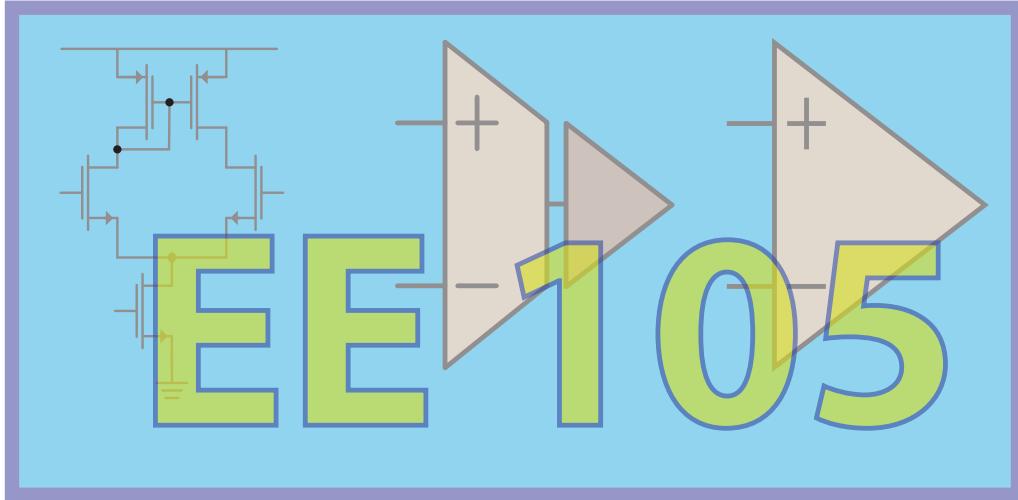


Figure 5.7: (a) The capacitance as a function of reverse bias voltage. This result holds for an abrupt junction. (b) This structure is often used to tune the center frequency of an *LC* tank to build a voltage controlled oscillator. The control voltage V_c is the reverse bias DC voltage, and C_2 is needed to DC isolate the varactor from the tank.

⁴This is done by comparing edge arrival times to a precision lower frequency reference signal derived from either the received signal transitions or an accurate clock reference derived from a crystal oscillator.



6. PN Junction Currents

6.1 Chapter Preview

This chapter is a continuation of the previous chapter, and it builds upon the knowledge that you have gained with regards to *PN*-junctions. The main focus is to derive the current-voltage ($I - V$) relations for a diode. We will demonstrate that diodes are "one way streets" allowing current to flow in one direction (from *P* to *N*), but essentially shut off and behave as open circuits when current flows in the other direction. To derive the $I - V$ curve, we need to cover a lot of preliminary material, such as the carrier concentration in non-equilibrium, in other words when $np \neq n_i^2$. We will use the **current continuity equation** to investigate how **minority carriers** diffuse across a region of **majority carriers**, and find that the diode $I - V$ curve is essentially determined by how minority carriers behave as they are injected into the *P* and *N* regions. Finally, in the last part of this chapter, we will investigate photovoltaic cells, or simply solar cells, and light-emitting diodes (LEDs), two important applications of *PN*-junctions.

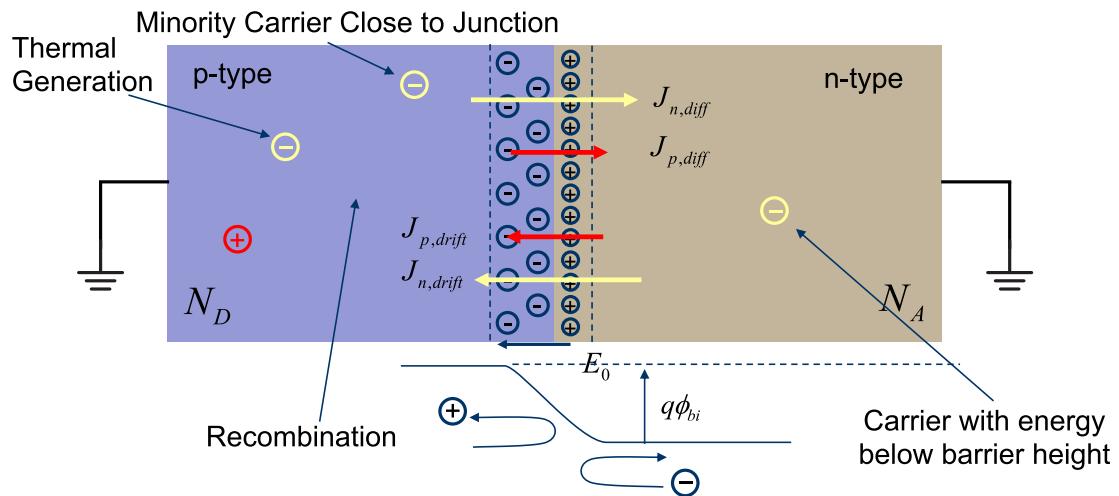


Figure 6.1: In equilibrium, with no applied fields, the current in a *PN*-junction is zero because drift and diffusion current cancel. The built-in potential only allows a small diffusion current to flow, and drift currents due to minority carriers flow in the opposite direction.

6.2 Qualitative Overview of Diode Currents

6.2.1 Diode under Thermal Equilibrium

Under thermal equilibrium (*Fig. 6.1*), when there is no bias, we know that the net current is zero. The diffusion current is small, because very few carriers have enough energy to penetrate the barrier. Drift current is small since minority carriers are few and far between, and only minority carriers generated within a diffusion length can contribute to the current. Other minority carriers recombine before they reach the junction.

6.2.2 Reverse Bias

There are two important points to keep in mind. First, minority drift current is independent of the barrier height, or the built-in potential between the *PN*-junction. On the other hand, diffusion current is a strong (exponential) function of the barrier height. Since a reverse bias causes an increased barrier to diffusion, as shown in *Fig. 6.2*, the diffusion current is reduced exponentially. The drift current does not change, so the net result is a small reverse current, known as the "**saturation current**".

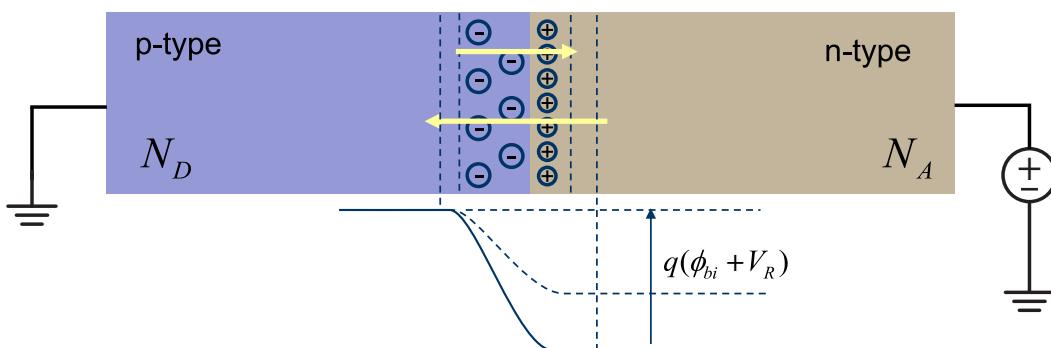


Figure 6.2: When a reverse bias is applied across a *PN*-junction, the diffusion current reduces (exponentially), and so drift currents overcome diffusion currents. The drift currents are usually very small since they are a result of minority carriers that are generated near the junction.

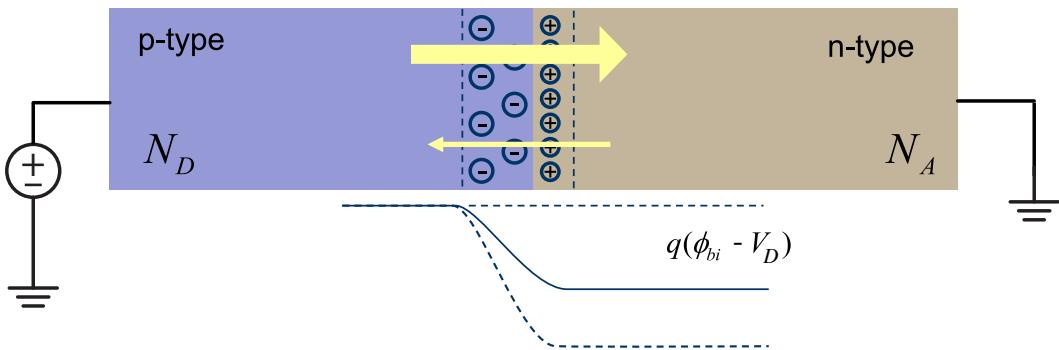


Figure 6.3: In forward bias, the diffusion current increases (exponentially) and dominates over the drift current from minority carriers. Majority carriers from the *P*-region are injected into the *N*-type region and become minority carriers. The same happens to electrons injected from the *N*-side.

6.2.3 Forward Bias

With reference to *Fig. 6.3*, a forward bias causes an exponential increase in the number of carriers. These carriers have sufficient energy to penetrate the barrier, and so the diffusion current *increases* exponentially. Because the drift current does not change, the net result is a large forward current.

6.2.4 Diode I-V Curve

As shown in *Fig. 6.4*, the **diode I-V relation** is an exponential function, as we have discussed qualitatively. This exponential is due to the **Boltzmann distribution** of carrier number versus energy. For a reverse bias voltage, the current saturates to I_S , the bias independent drift current due to minority carriers.

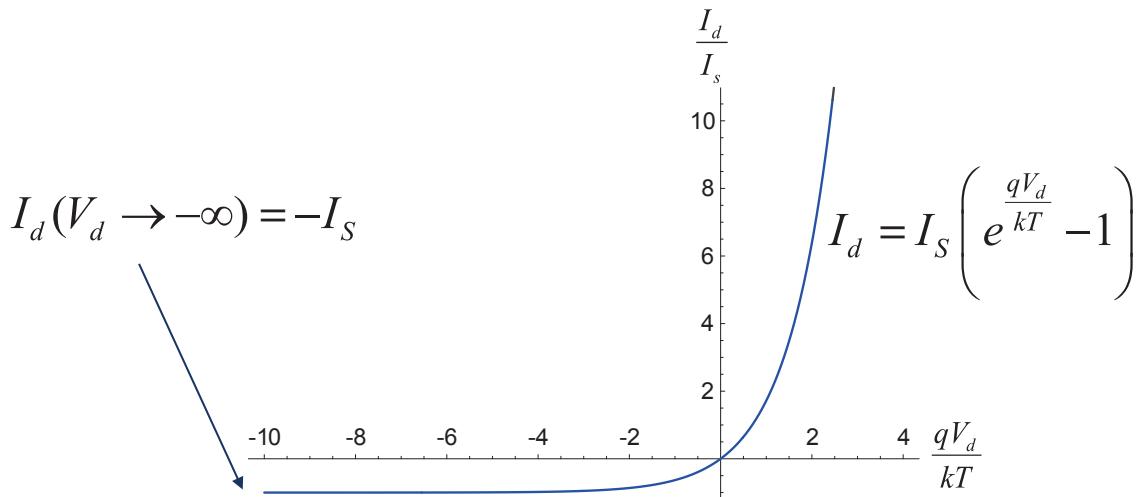


Figure 6.4: The *I-V* curve for a diode. The diode voltage is normalized to kT/q and the current is normalized to the diode saturation current I_S . In forward bias, the current increases exponentially whereas in reverse bias the current saturates to a small value of I_S .

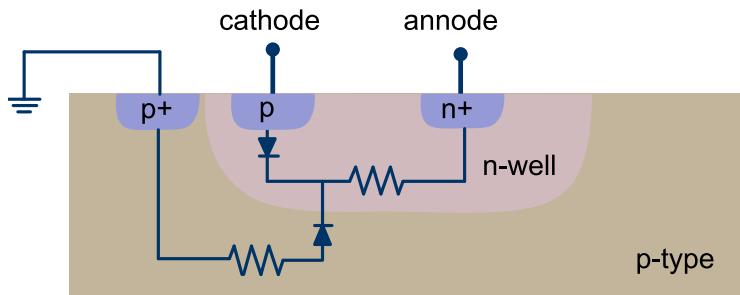


Figure 6.5: Cross section of a fabricated IC diode. The N -well is used to isolate the diode from the substrate, and the diffusion regions inside the N -well are used for the diode structure. The cathode is connected to a P region that forms the PN -junction with the N -well, and the N^+ region is used to make an ohmic contact with the N -well. The substrate is biased to the most negative voltage to prevent the parasitic diode (N -well to substrate) from forward biasing.

6.2.5 Fabrication of IC Diodes

Our drawings of PN -junctions so far has been very idealized. The actual structure of a PN -junction is slightly more complex, as shown in *Fig. 6.5*. To fabricate a diode, we first start with P -type substrate (typically). We create an "N-well" diffusion region to house the diode. Within the well, we counter dope to create P and N^+ diffusion regions to form the cathode and anode terminals of the diode. The N^+ region may seem puzzling at first, but it is needed to create an ohmic contact with the metal region at the anode. Note that we actually end up with two diodes, one between the cathode and anode, as desired, and one from the substrate to the N -well region, which is a parasitic diode. Why not use this diode? The reason is that in an integrated circuit, we have many components and we want them to be isolated from each other. Since all circuits fabricated share the silicon substrate, the substrate node is usually not used. But a diode is a diode, and to isolate the devices from one another, we must reverse bias the unwanted diodes. In this case, the N -well must be reverse biased with respect to the substrate. Typically the substrate is tied to the lowest potential (ground for example) using substrate "taps" or body contacts, which are P^+ regions as shown. One final word about a practical diode. We have drawn in resistors to remind ourselves that P -type and N -type regions are resistive. If we are not careful, resistive parasitics will add up and dominate the device behavior.

6.3 Carrier concentration in Non-Equilibrium

6.3.1 Equilibrium versus Non-Equilibrium

The Law of Mass Action says that in equilibrium, the product of the concentration of holes and electrons is constant

$$np = n_i^2$$

As you may recall, we derived (in a hand wavy fashion) this equation by asserting that in equilibrium, the rate of generation and recombination should be equal. We identified the left-hand side as proportional to the recombination rate (the more carriers, the more recombination):

$$R = c(np)$$

and the right hand side as the thermal generation rate:

$$G = \text{constant} = f(T)$$

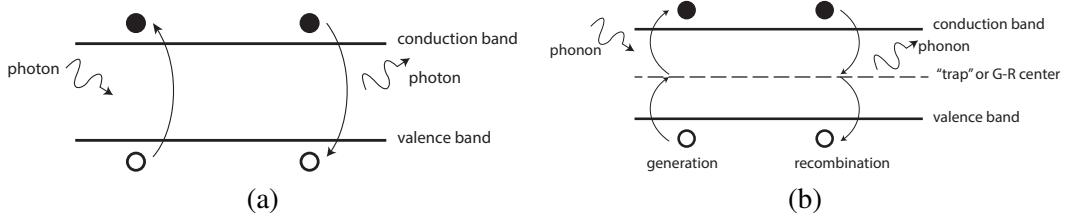


Figure 6.6: (a) Direct generation and recombination by aid of an optical photon with energy exceeding the band gap. (b) Trap assisted generation and recombination through vacancies, defects, or impurities with approximately mid-band gap energy levels. In silicon, the process in (b) is much more common.

By equating these two, we have the Law of Mass Action:

$$R = G \Rightarrow np = n_i^2(T) \quad \text{Law of Mass Action} \quad (6.1)$$

Since the thermal generation rate is essentially a constant for a fixed temperature, if an external process drives up the concentration of say one type of carriers, in particular minority carriers, then we can say that we are in a non-equilibrium situation because now there is an imbalance between recombination and generation:

$$R \neq G$$

The rate of thermal generation is the same as before, but now we have an excess of minority carriers:

$$R > G$$

What happens? The extra carriers should recombine at a higher rate and in steady-state we should return to equilibrium. What drives the dynamics?

6.3.2 Recombination-Generation Centers

Up to now we have assumed that all generation and recombination occurs through a direct mechanism (shown on the left of Fig. 6.6), whereby valence band electrons accept energy (thermal or optical) and are raised into the conduction band. The inverse process is of course recombination. This **direct band gap** mechanism, while possible, is very unlikely in silicon. In reality most recombination and generation occurs through an **indirect band gap** mechanism shown on the right of the figure. The energy level near the middle of the band represents a "trap", usually an impurity or missing atom in the crystal, that can assist both processes by temporarily trapping an electron and then releasing it into the valence band (and vice versa). This process captures and releases **phonons** (crystal vibrations) rather than **optical photons**.

6.3.3 Thermal Generation/Recombination Rate

For a hole in an N -type material, we find the rate of generation is given by:

$$\left. \frac{\partial \Delta p}{\partial t} \right|_{\text{thermal R-G}} = -c_p N_T \Delta p \quad (6.2)$$

In the above equation we have partial derivatives, because there are many mechanisms that generate carriers. We are just focusing on the thermal rate. Δp is the excess minority carrier concentration above equilibrium:

$$\Delta p = p - p_0 \quad (6.3)$$

Here the rate is negative when $\Delta p > 0$, because recombination will outweigh generation when we create an excess of minority carriers. Note that the rate depends on a constant c_p , the density of traps N_T , and the number of excess minority carriers Δp .

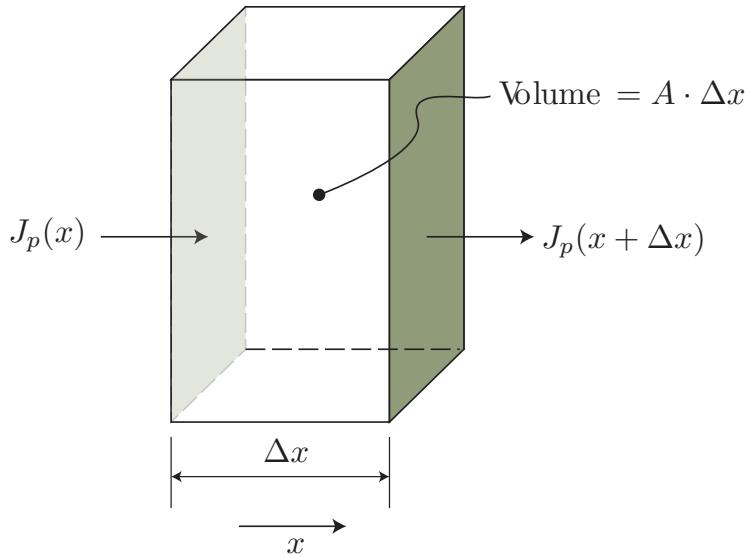


Figure 6.7: Current flow into a region must satisfy the continuity equation. If the current density varies spatially, it must be accompanied by generation or recombination.

Minority Carrier Lifetime

It is common practice to lump the constants $c_p N_T$ into a term called the **minority carrier lifetime** τ_p , because from the above equation it is clear that their product has units of inverse time:

$$\left. \frac{\partial \Delta p}{\partial t} \right|_{\text{thermal R-G}} = -c_p N_T \Delta p = -\frac{\Delta p}{\tau_p} \quad (6.4)$$

The minority carrier lifetime depends very strongly on the quality of the material, and it can vary by orders of magnitude for a poorly controlled process to a well controlled one (from nanoseconds to microseconds).

Majority Carriers versus Minority Carriers

Eqs. 6.4 can be interpreted as the "lifetime" of a minority carrier. We know minority carriers "live" in a pool of majority carriers, and they annihilate upon meeting majority carriers. When you're a minority carrier, there is a good chance that you are going to recombine with a majority carrier, especially when there are more traps, N_T . The more minority carriers there are, the faster the rate of recombination. Note that the rate of generation is fixed (towards equilibrium), whereas the recombination rate is higher due to the excess carriers. The same equation is of course valid for electrons in a *P*-type region:

$$\left. \frac{\partial \Delta n}{\partial t} \right|_{\text{thermal R-G}} = -\frac{\Delta n}{\tau_n} \quad (6.5)$$

6.4 Current Continuity Equation

6.4.1 Excess Carrier Continuity

Consider a "box" of current flowing shown in *Fig. 6.7*. Note that if the current density varies spatially (assume variations along yz are negligible), then there must be a net rate of accumulation (or depletion) of carriers. Applying this to excess minority carriers, we have:

$$J_p(x + \Delta x) - J_p(x) = \Delta x q \frac{\partial \Delta p}{\partial t} = \Delta x q \frac{\Delta p}{\tau_p} \quad (6.6)$$

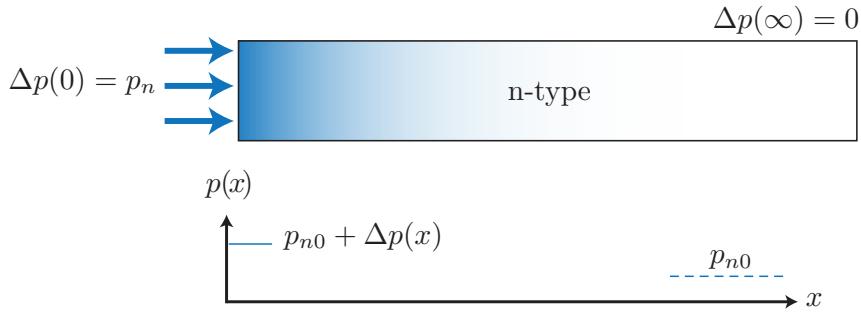


Figure 6.8: Excess minority carrier charge is continuously injected at the origin into an N -type region. Far away from the injection, we expect carriers to return to equilibrium values. Using these boundary points, we would like to solve for the minority carrier distribution in the N -type region.

This equation is stating something quite obvious. If the number of minority carriers is changing within a region, then there must be either generation or recombination in that region. We know from our previous discussion that the excess minority carrier population varies as $\Delta p/\tau_p$. Now taking $\Delta x \rightarrow 0$, we have the following differential equation:

$$\frac{dJ_p}{dx} = q \frac{\Delta p}{\tau_p} \quad (6.7)$$

6.4.2 Diffusion Currents

Now we also know that if there is a spatial variation in the concentration of carriers, then we also have to account for a diffusion current:

$$J_p = qD_p \frac{d\Delta p}{dx} \quad (6.8)$$

Putting this all together, we have

$$\frac{dJ_p}{dx} = \frac{d(qD_p \frac{d\Delta p}{dx})}{dx} = qD_p \frac{d^2\Delta p}{dx^2} = q \frac{\Delta p}{\tau_p} \quad (6.9)$$

Simplifying the equation, and defining a new constant L_p , we have:

$$\frac{d^2\Delta p}{dx^2} = \frac{\Delta p}{D_p \tau_p} = \frac{\Delta p}{L_p^2} \quad (6.10)$$

The constant $L_p = \sqrt{D_p \tau_p}$ is known as the **diffusion length** for reasons which will become clear shortly.

6.4.3 Excess Carrier Current Flow

Let's suppose that some process, yet to be determined, continually creates an excess of minority carriers at the origin of an N -type semiconductor (shown in Fig. 6.8). In other words, the boundary conditions are:

$$\Delta p(0) = p_n \quad (6.11)$$

and:

$$\Delta p(\infty) = 0 \quad (6.12)$$

Under these conditions we assume that the "disturbance" at x is maintained over all time (steady-state). We also assume that far away from the disturbance thermal equilibrium conditions persist. What is the distribution of carriers as we leave the origin?

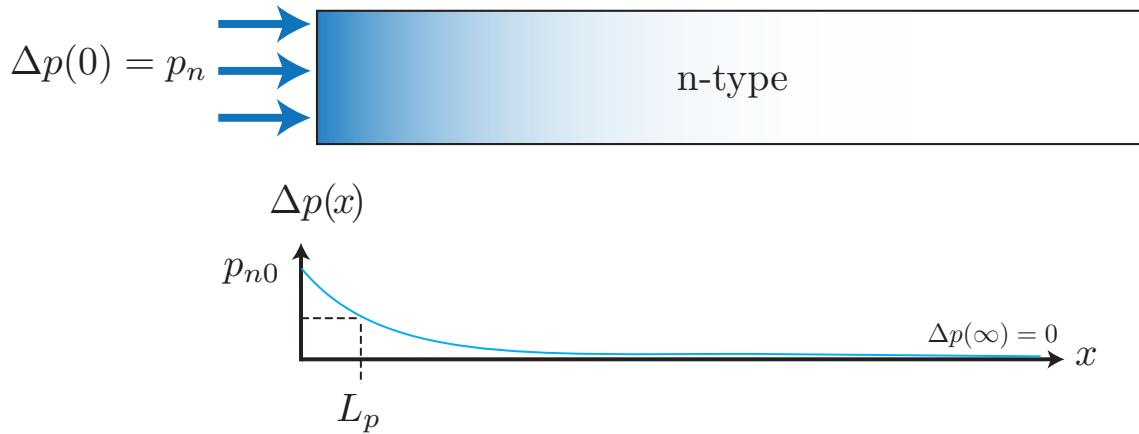


Figure 6.9: Minority carrier distribution as a function of position when excess carriers are continuously injected at the origin.

Minority Carrier Distribution

The general solution, shown in *Fig. 6.9*, is given by:

$$\Delta p(x) = Ae^{-x/L_p} + Be^{x/L_p} \quad (6.13)$$

Applying boundary conditions, we have $B = 0$ and $A = p_n$:

$$\Delta p(x) = p_n e^{-x/L_p} \quad (6.14)$$

This also satisfies the $\Delta p(\infty)$ boundary condition. It is important to realize that this is a steady-state solution (there is no time dependence).

Majority and Minority Carrier Distribution

We can find the current flow due to this excess minority concentration by calculating the diffusion current:

$$J_p(x) = -qD_p \frac{d\Delta p}{dx} = \frac{p_n}{L_p} e^{-x/L_p} \quad (6.15)$$

The above equation shows that the excess minority carriers diffuse into the *N*-type region a distance of a few diffusion lengths, L_p . Note that the current density varies with x . Yet the current must be continuous, because charge is not piling up as a function of time. Recall that this is a steady-state distribution. This means that if we evaluate the current at $x = 0$, that is the same current found throughout the *N*-type region.

Where is the additional current coming from? It's coming from majority carriers who are "coming in" to annihilate the minority carriers. They contribute a current which is exactly $1 - J_p(x)$.

6.5 Forward Biased PN-junctions

6.5.1 Minority Carriers at Junction Edges

When a diode is at zero bias, we know that the diffusion and drift currents balance. If we forward bias the junction, then we are lowering the barrier to diffusion current flow, which means more

minority carriers will be injected. The exact derivation leads to the following expression for the minority carrier concentration at boundaries of the depletion region:

$$\frac{p_n(x = x_n)}{p_p(x = -x_p)} = e^{-(\text{Barrier energy})/kT} \quad (6.16)$$

The result is known as **Boltzmann's Law**:

$$\frac{p_n(x = x_n)}{N_A} = e^{-q(\varphi_{bi} - V_D)/kT}$$

Boltzmann's Law
(6.17)

In Eq. 6.17, V_D is the forward bias voltage that we apply, and φ_{bi} is the built-in potential. This implies an exponentially larger number of minority carriers that are injected into the regions.

We actually derived this equation in the last chapter under thermal equilibrium, but it is important to realize that we are now in a non-equilibrium situation. So strictly speaking, our derivation does not apply. In practice, it can be shown that the law is valid under "**low level injection**" conditions. In other words, when the density of minority carriers injected remains much smaller than the number of majority carriers, the law is valid. Using Boltzmann's Law, the **minority carrier concentrations** at the edges of the depletion region are given by:

$$p_n(x = x_n) = N_A e^{-q(\varphi_B - V_D)/kT} \quad \text{Minority } e^- \text{ concentration, edge of depletion} \quad (6.18)$$

(6.19)

$$n_p(x = -x_p) = N_D e^{-q(\varphi_B - V_D)/kT} \quad \text{Minority } h^+ \text{ concentration, edge of depletion} \quad (6.20)$$

Here are some important things to remember about these equations. In particular, the subscript notation is confusing at first, so please study each term in the above equation carefully.

- Note 1: N_A and N_D are the majority carrier concentrations on the *other* side of the junction.
- Note 2: We can reduce these equations further by substituting $V_D = 0V$, and verifying that we satisfy thermal equilibrium conditions.
- Note 3: These equations are valid assuming that $p_n \ll N_D$ and $n_p \ll N_A$.

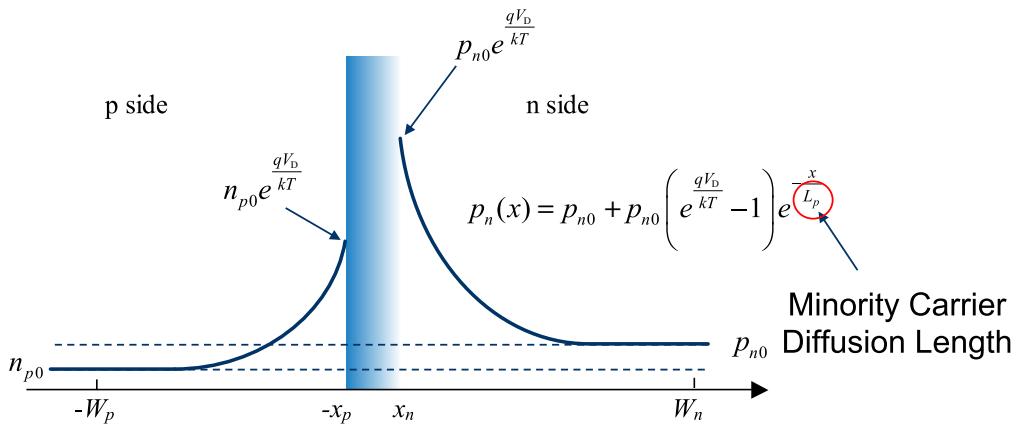


Figure 6.10: Distribution of minority carriers in a *PN*-junction under forward bias. Forward bias increases the diffusion current and causes excess minority carriers (above thermal equilibrium values) to be injected into the *p* and *n* type regions. The minority carriers diffuse and recombine and exhibit an exponentially decaying profile.

6.5.2 Minority Carrier Concentration Distribution

With an excess of minority currents injected into the *N* and *P* regions, we can use our previously derived result to find the **concentration distribution**, which is the carrier concentration as a function of x . Note that the currents decay exponentially with a characteristic length corresponding to the diffusion length. The solutions are plotted in Fig. 6.10. A special case of great practical interest is when virtually none of the diffusing holes and electrons recombine. In other words, the material has a very long minority carrier lifetime. The solution to the equations in this case can be re-derived by taking the limit of large τ .

The solution, shown in Fig. 6.11, is that we obtain straight lines for the concentration profile. To see this, solve the differential equation we derived earlier but set τ_p and τ_n to very large times. This also happens if the minority carrier diffusion lengths are much larger than the diode width: $L_{n,p} \gg W_{n,p}$.

The form of the solutions to a linear concentration profile can be inferred if we note that without recombination, the current must be uniform. In other words, the slope of the minority carrier profile *must* be constant. That means that the carrier distribution is linear.

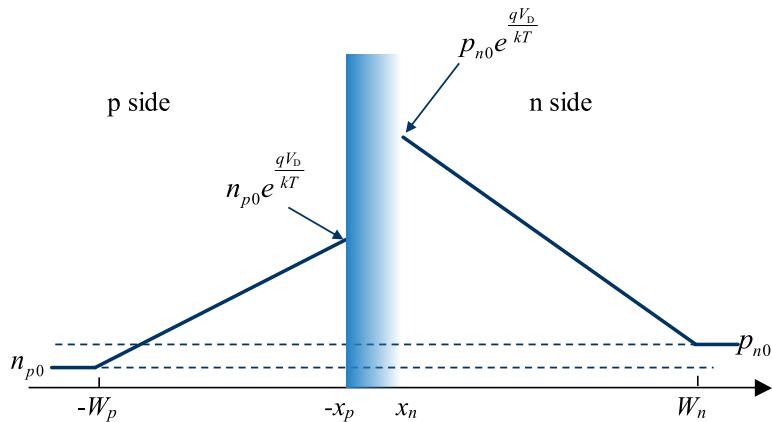


Figure 6.11: Distribution of minority carriers in a *PN*-junction under forward bias for the case of very long minority carrier lifetime. The minority carriers diffuse but do not recombine with majority carriers, leading to linear gradients in the profile.

6.5.3 Diode Diffusion Currents

Given the profile of minority carrier currents in each region, it is rather straightforward to find the overall currents for holes and electrons. Here we use the linear profile case to illustrate the calculations:

$$\frac{dn_p}{dx}(x) \approx \frac{n_{p0} e^{\frac{qV_A}{kT}} - n_{p0}}{-x_p - (-W_p)} \quad (6.21)$$

Note that the electron minority carrier concentration at equilibrium is given by:

$$n_{p0} = \frac{n_i^2}{N_A} \quad (6.22)$$

This leads to electron diffusion current:

$$J_{ndiff} = q D_n \frac{dn_p}{dx} \Big|_{x=-x_p} \approx q \frac{D_n}{W_p} n_{p0} \left(e^{\frac{qV_A}{kT}} - 1 \right) \quad \text{Diffusion current, electrons} \quad (6.23)$$

and hole diffusion current

$$J_{pdiff} = -q D_p \frac{dp_n}{dx} \Big|_{x=x_n} \approx -q \frac{D_p}{W_n} p_{n0} \left(1 - e^{\frac{qV_A}{kT}} \right) \quad \text{Diffusion current, holes} \quad (6.24)$$

Also, since electrons have negative charge, both the electron and hole currents sum together in phase, despite going in different directions. This gives a total current of:

$$J_{diff} = q n_i^2 \left(\frac{D_p}{N_D W_n} + \frac{D_n}{N_A W_p} \right) \left(e^{\frac{qV_A}{kT}} - 1 \right) \quad \text{Diffusion current, total} \quad (6.25)$$

6.5.4 Height Analogy

Imagine a "wall" that is a certain height (say 6 feet tall), and imagine that only the kids in a playground who are taller than that certain height can "jump" the barrier. This is the shaded area under the **normal distribution**, or the "error function" plotted on the left of Fig. 6.12. From the plot on the right, it is clear that as we decrease the height of the wall exponentially more kids can jump the fence, because the shaded area is growing exponentially.

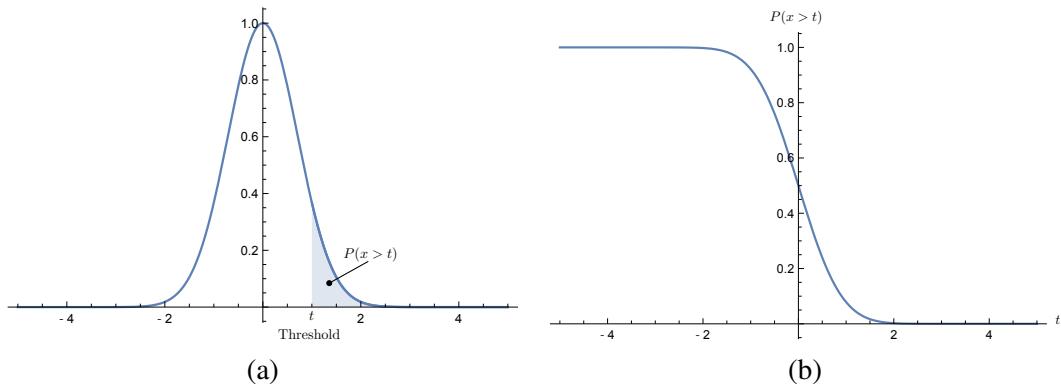


Figure 6.12: (a) The normal distribution function (**Gaussian distribution**) is approximately the variation in heights between students on a playground. The shaded area represents the fraction of students who are taller than a certain threshold. (b) The integration of the normal distribution as a function of threshold (the shaded area) shows that lowering the barrier (moving left along the curve) leads to an exponential increase in the number of students exceeding a threshold in height.

6.6 Diode Small Signal Model

In many applications, we are concerned with the diode response to a **small signal**. In these situations, we usually bias the diode with a DC voltage to the desired **operating point**, either in forward or reverse biased region. Then we apply a small incremental signal to the diode. A good example is when the diode is used as a switch. In this case we either turn the diode on and rely on its low impedance to pass small signals, or turn it off to isolate small signals from the rest of the circuit. Many mobile phones use a diode exactly as described above to share a single antenna between the receiver and transmitter.

To understand how the forward biased diode responds to small signals, the $I - V$ relation of a diode is first approximated:

$$I_D + i_D = I_S \left(e^{\frac{q(V_d+v_d)}{kT}} - 1 \right) \approx I_S e^{\frac{qV_d}{kT}} e^{\frac{qv_d}{kT}} \quad \text{Diode model, small-signal} \quad (6.26)$$

For $x = v_d/kT/q$ and $v_d \ll kT/q$ (small-signals less than 26 mV), we can linearize the exponential using a Taylor Series expansion:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$$

In other words, the total current can be written in terms of the **DC bias** current and the incremental AC current:

$$I_D + i_D \approx I_D \left(1 + \frac{qv_d}{kT} + \dots \right)$$

From the perspective of the "small-signal", the current flows in proportion to the applied voltage v_d . In other words, the diode looks like a conductor:

$$i_D \approx \frac{qv_d}{kT} = g_d v_d \quad (6.27)$$

6.6.1 Diode Capacitance

We have already seen that a reverse biased diode acts like a capacitor, because the depletion region grows and shrinks in response to the applied field. The capacitance in forward bias requires some additional derivations, but in practice we can use the following approximation for the forward bias:

$$C_j = A \frac{\epsilon_s}{X_{dep}} \approx 1.4 C_{j0} \quad (6.28)$$

On the other hand, another important charge storage mechanism comes into play in forward bias. Minority carriers injected into P and N regions "stay" in each region for a while. On average, the additional charge is stored in the diode, and that must be accounted for.

Charge Storage

As shown in Fig. 6.13, increasing forward bias increases minority charge density (the area under the curve goes up). By charge neutrality, the source voltage must supply equal and opposite charge. A detailed analysis yields:

$$C_d = \left(\frac{1}{2} \right) \left(\frac{qI_d}{kT} \right) \tau \quad (6.29)$$

In the above equation, τ is either the time to cross junction (short junction), or the minority carrier lifetime (diffusion length shorter than the width of the diode).

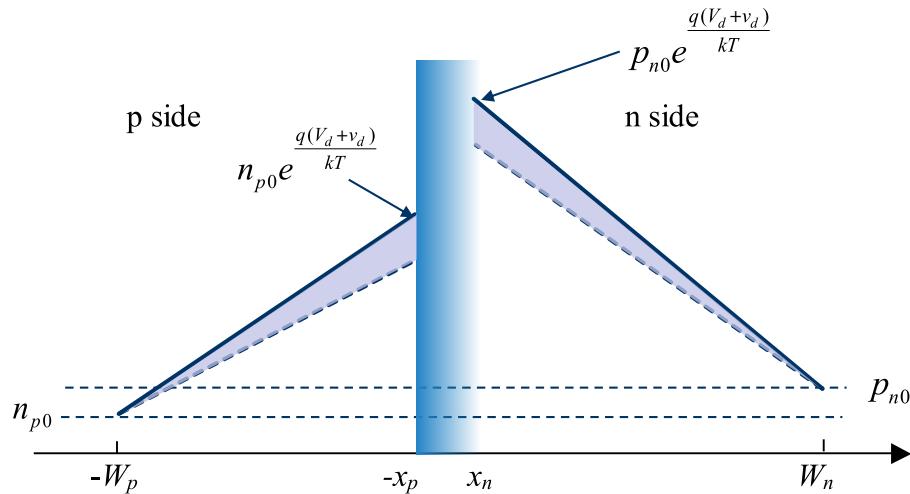


Figure 6.13: Change in minority carrier distribution in a forward biased diode when the forward bias is increased. The dashed line represents the profile for a lower current, and the area between the curves represents the extra charge stored that must flow into the junction.

Consider another way to understand why there is additional capacitance in the diode. Imagine a voltage that abruptly transitions from forward bias to reverse bias. If the diode were just a conductor, it would immediately shutoff, and the current would collapse to zero instantaneously. Even accounting for the junction capacitance, the time to "discharge" the diode must also account for the fact that the diffusion current is due to a spatial distribution of minority carriers in the P and N regions. These distributions cannot collapse to equilibrium values instantaneously. If we stop injecting carriers into the junction, then eventually the distribution will return to the equilibrium value, and the time is proportional to how long it takes for excess minority carriers to "disappear", either through recombination or by leaving the diode. This is why the capacitance is directly proportional to this time τ , and also to the current. The more current the diode is conducting, the more charge that is stored.

6.6.2 Complete Small-Signal Model

In Fig. 6.14, we summarize everything about the diode small-signal model in the forward biased region in the form of an equivalent circuit. We have seen that from the perspective of the small-signal source, the diode appears as a resistor in parallel with a physical capacitor. The model is valid for a forward biased diode excited by a small *incremental* signal. It consists of three distinct components: **small-signal resistance**, **junction capacitance**, and **diffusion capacitance**. Strictly speaking, these components are non-linear as a function of operating point. But for a fixed operating point, any small deviations around it "see" linear components.

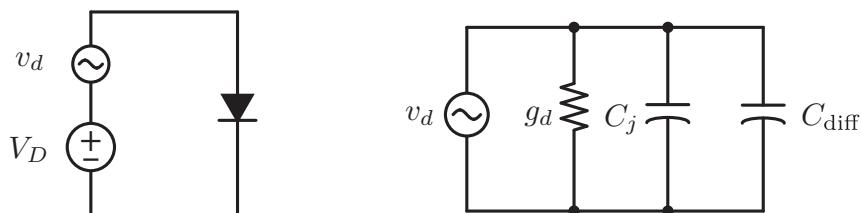


Figure 6.14: The complete small-signal model for a diode in forward bias includes diode conductance and charge storage capacitance due to the junction stored charge (depletion region width) and minority carrier stored charge.



Figure 6.15: A typical solar panel is made from hundreds to thousands of diodes acting as photo-voltaic cells.

6.7 Photonic Applications

6.7.1 Solar Cells

A **solar cell** is a *PN*-junction designed for converting light energy (photons) into electric energy (energized electrons). It is also known as a **photovoltaic cell** (PV cell). Nowadays it is a common sight to see solar cells on rooftops and in solar farms (Fig. 6.15). Typically, efficiency numbers range from 10% to 30% and 1m^2 of area can generate at 25W of power on average (150W peak). Also, the energy cost to fabricate a solar cell is steadily improving, and most solar panels will be energy neutral after a few years of operation.

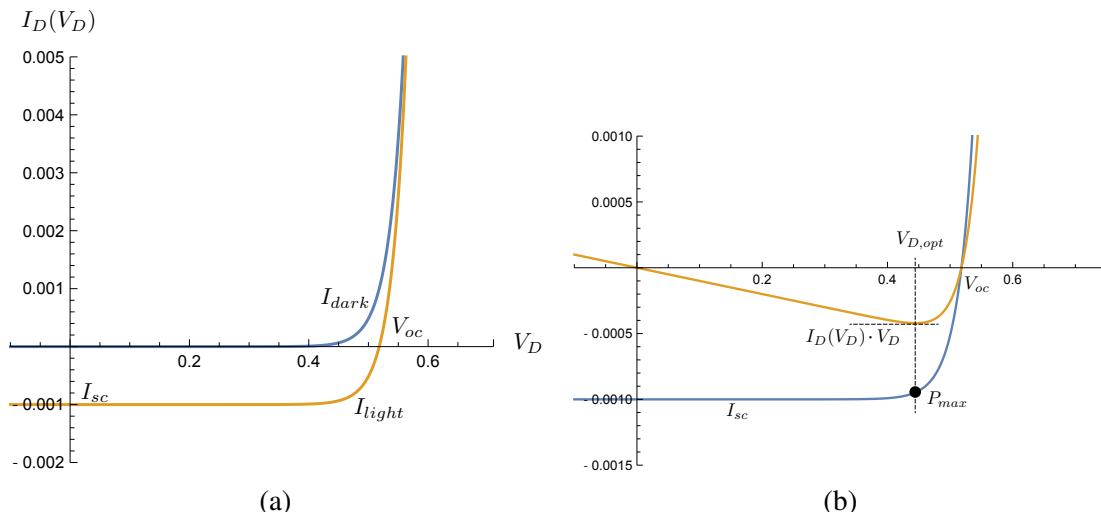


Figure 6.16: (a) Modified diode I - V curve to account for the photo-current. (b) A zoom in to the fourth quadrant where we can extract power from the photo-current. The important parameters include the *short circuit current*, the *open circuit voltage*, and the point of maximum power P_{max} .

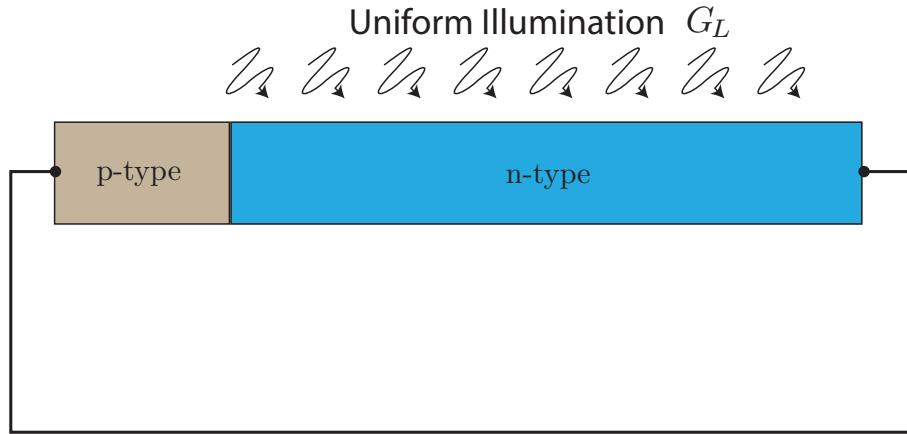


Figure 6.17: A simple model for a solar cell (PN-junction) under equilibrium. We assume the N-region is illuminated uniformly and calculate the minority carrier profile in the region to calculate the short circuit current.

6.7.2 PN-junction Solar Cells

The $I - V$ curve of a solar cell is basically the same as a diode with an extra "generation" term due to optical photons generating electron-hole pairs, as shown in *Fig. 6.16(a)*. The basic diode $I - V$ curve is modified to account for the photon generated carriers I_{sc} :

$$I_D = I_S(e^{qV/kT} - 1) - I_{sc} \quad (6.30)$$

Up to now, we have been ignoring this photocurrent. So, it is worthwhile to examine how to make this term significant. Useful energy is only extracted when the $I \cdot V$ product is negative (here in the fourth quadrant), because in this region the solar cell is generating a voltage and current in opposite phase, and it is thereby a generator.¹ If we zoom into the fourth quadrant (*Fig. 6.16(b)*) and examine the product of $I \cdot V$, we can identify three important points. The I_{sc} is the current that flows when the voltage $V = 0V$, or for a short circuited solar cell. Likewise, the voltage V_{oc} is the open circuit voltage, or the voltage we observe when there is no load on the solar panel. Finally, we have also denoted the point P_{max} as the portion of the solar panel that delivers optimal power.

To characterize a solar cell's performance, we need to calculate I_{sc} and V_{oc} using our knowledge of PN-junction diode physics.

6.7.3 Equations for Optical Generation

Suppose for simplicity that we have an N-type semiconductor, and light shines on it with a uniform intensity (see *Fig. 6.17*). This generates electron-hole pairs at a certain rate. From current continuity, we can say that the holes and electrons are generated, and must satisfy the continuity equation:

$$\frac{d^2\Delta p}{dx^2} = \frac{\Delta p}{L_p^2} - \frac{G_L}{D_p} \quad (6.31)$$

This is the same as the equation we derived earlier (see *Eq. 6.10*) for the PN-junction, with the extra term G_L added to account for optical generation. For a uniform layer and constant illumination, there is no x variation:

$$0 = \frac{\Delta p}{L_p^2} - \frac{G_L}{D_p} \quad (6.32)$$

¹This is due to the passive sign convention that we adopt in electrical engineering and physics.

So we can solve for the excess minority carrier concentration:

$$\Delta p = \frac{L_p^2 G_L}{D_p} = G_L \tau_p \quad (6.33)$$

This is generation of holes above the thermal generation rate, and it is due to the incoming light. Essentially, for every hole generated by the light it only lasts an average of τ_p seconds (minority carrier lifetime). So if we multiply the rate of photon flux (and hence electron-hole pair generation) by the average amount of time they stick around, we have the net excess minority carrier concentration. Stated another way, if ten people per minute (G_L term) enter a room and stick around for about five minutes (τ_p), how many people on average are in the room? 10 people/minutes multiplied by 5 minutes means there will be on average 50 people in the room at any given time.

6.7.4 Short Circuit Current

We are now in a position to derive the necessary parameters for a solar cell. For simplicity, assume a very thin p^+ layer (no light illumination on the p^+ side) and optical carrier generation in the N -region. Since there is no applied bias voltage on the PN -junction, the boundary conditions at the edge of the depletion region is given by:

$$\Delta p(0) = 0 \quad (6.34)$$

On the other hand, if we travel far enough into the N -type region, we expect spatial variations to vanish and the number of excess holes should be directly proportional to the incoming light:

$$\Delta p(\infty) = G_L \tau_p \quad (6.35)$$

Using the general solution, and imposing boundary conditions, we have:

$$\Delta p(x) = G_L \tau_p \left(1 - e^{-x/L_p} \right) \quad (6.36)$$

And the corresponding diffusion current is given by:

$$J_p(x) = -q D_p \frac{d\Delta p}{dx} = q \frac{D_p}{L_p} \tau_p G_L e^{-x/L_p} \quad (6.37)$$

This means the short-circuit current is given by:

$$I_{sc} = A J_p(0) = A q L_p G_L \quad (6.38)$$

Since in practice G_L is not uniform, this result is only approximately true. However, it does give us some nice intuition for the important parameters. In fact, the above equation is actually saying something very simple. With reference to Fig. 6.18, only carriers generated within one diffusion length (L_p) of the junction contribute to the current. That is because optically generated minority carriers generated further away don't actually end up reaching the junction. They recombine, and no net current flows.

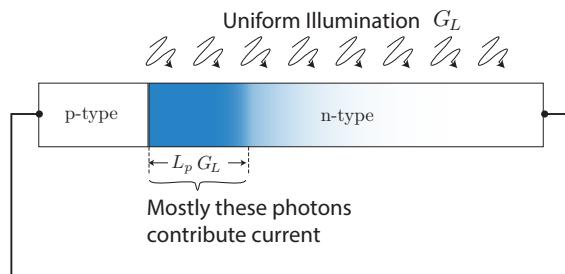


Figure 6.18: Due to recombination, only minority carriers generated within a diffusion length of the depletion region contribute significantly to the short circuit current. Other generated minority carriers recombine and do not contribute net current.

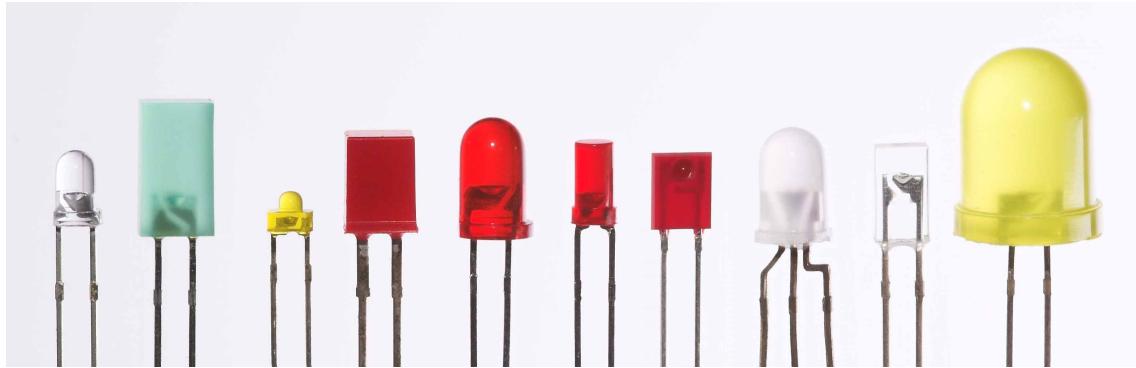


Figure 6.19: Light Emitting Diodes (LEDs) of various colors. [Wikipedia]

6.7.5 Open Circuit Voltage

Now we have an expression for the optically generated current, and we can substitute it into the total diode current:

$$I_D = \left(\frac{A q n_i^2}{N_D} \right) \left(\frac{D_p}{L_p} \right) \left(e^{qV/kT} - 1 \right) - (A q L_p G_L) \quad (6.39)$$

Solving for the open-circuit voltage ($I_D = 0A$), and assuming the exponential dominates:

$$\left(\frac{n_i^2}{N_D} \right) \left(\frac{D_p}{L_p} \right) e^{qV_{oc}/kT} = L_p G_L \quad (6.40)$$

This can be simplified to:

$$V_{oc} = \left(\frac{kT}{q} \right) \ln \left(\frac{\tau_p G_L N_D}{n_i^2} \right) \quad (6.41)$$

To generate a high open-circuit voltage, we need a large τ_p , or a very high quality semiconductors (low defect density).

6.7.6 Light Emitting Diodes (LEDs)

For **LEDs** (see *Fig. 6.19*), it is much more efficient to use a so called "direct band gap" material, which we will define next. Common materials used for LEDs are *InP* and *GaN*. Light is emitted when electrons and holes undergo radiative recombination (in silicon most of the energy is lost to phonons). Let's see why by now defining what a "direct" and "indirect" band gap is.

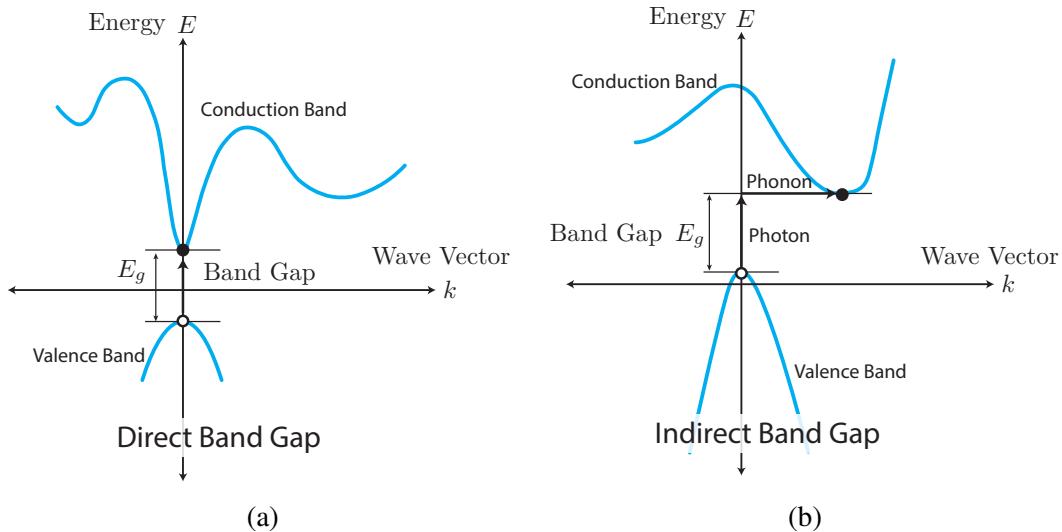


Figure 6.20: (a) In a direct band gap material, such as *GaAs*, the peak of the valence band and the valley of the conduction band align in the momentum space, allowing a photon to be absorbed (or emitted) while conserving momentum. (b) In an indirect band gap material, such as silicon, absorption or emission of a photon requires the assistance of a phonon in order to conserve both energy and momentum.

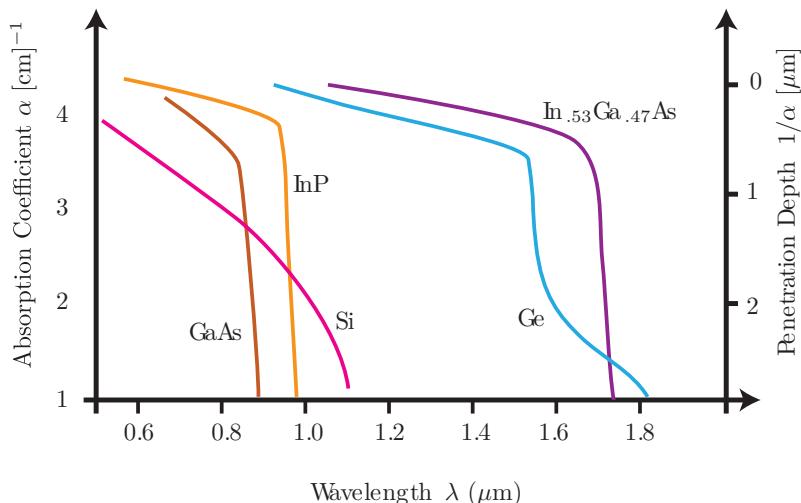


Figure 6.21: Qualitative plot of the absorption coefficient of various materials. Notice that photons of a given wavelength can only be absorbed if they have sufficient energy to overcome the band gap. This explains the sharp rise in the absorption coefficient for shorter wavelengths.

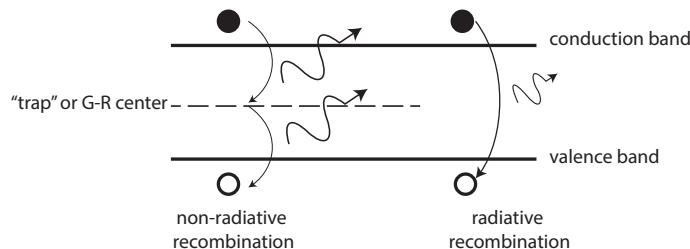


Figure 6.22: In an indirect band gap material, such as silicon, most recombination occurs through a trap (crystal defect or impurity) and generates phonons (left). In a direct band gap material, recombination leads to optical photons (right).

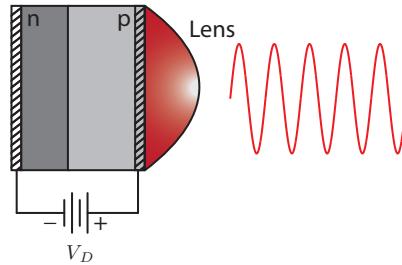


Figure 6.23: An LED is simply a forward biased pn junction made with a direct band gap material and constructed to collect photons. A lens is often used to collimate the light to maximize the intensity. Compare with Fig. 6.19.

Direct versus Indirect band gap Materials

In this book we don't go into the details, but it is important for now to realize that electrons have both energy E and momentum k . In a solid, the allowed energy-momentum states are dictated by quantum mechanics and plotted in an **E-k diagram**, such as the ones shown in Fig. 6.20. In a direct band gap semiconductor, the bands overlap and an electron can absorb a photon *directly*, and become a conduction band electron with the same momentum. In an indirect band gap, the top of the bands do not line up in k space, so the assistance of a phonon is needed for optical generation, making the process less efficient and less likely to occur. Silicon is an indirect band gap material, while many compound semiconductors (group III + V elements together) are direct band gap materials.

It is also important to know how semiconductors absorb light. A qualitative plot of photon absorption based on wavelength is shown in Fig. 6.21. Light intensity is absorbed by a semiconductor as long as the photon energy $E = h\nu = hc/\lambda$ is larger than the band gap. Light intensity drops off exponentially $e^{-\alpha x}$ where α is the **absorption coefficient**.

For the generation of photon, a direct band gap material is preferred. Direct recombination is efficient because k is conserved (momentum). In an indirect band gap, direct recombination is rare because k is not conserved. Most recombination is trap assisted, and does not produce optical photons (as shown in Fig. 6.22). This is the case with silicon, and this why most LEDs are not made with silicon. For efficient light generation, we prefer a direct band gap, and many III – V compound semiconductors have this property. **Compound semiconductors** are made with a mixture of group III and group V atoms, and have similar properties as semiconductors made of group III atoms.

6.7.7 LED Materials and Structure

The structure of an LED *PN*-junction diode is shown in Fig. 6.23. A lens is used to capture photons traveling in different directions, and to accurately align the beam of light for higher intensity. The direct band gap material is forward biased and the current flow results in minority carrier injection, which results in (mostly) radiative recombination. To avoid re-absorbing the photons, the materials and structure are carefully designed to emit light.

Each radiative recombination event generates a photon with energy of the band gap. This allows us to find the wavelength of the photon by equating the energy of a photon $h\nu$ (frequency ν) with the band gap energy:

$$E_g = h\nu = \frac{hc}{\lambda_{LED}} \quad (6.42)$$

Solving for the wavelength of the emitted photon, we have:

$$\lambda_{LED} = \frac{hc}{E_g} = \frac{1.24}{E_g(\text{eV})} \quad (6.43)$$

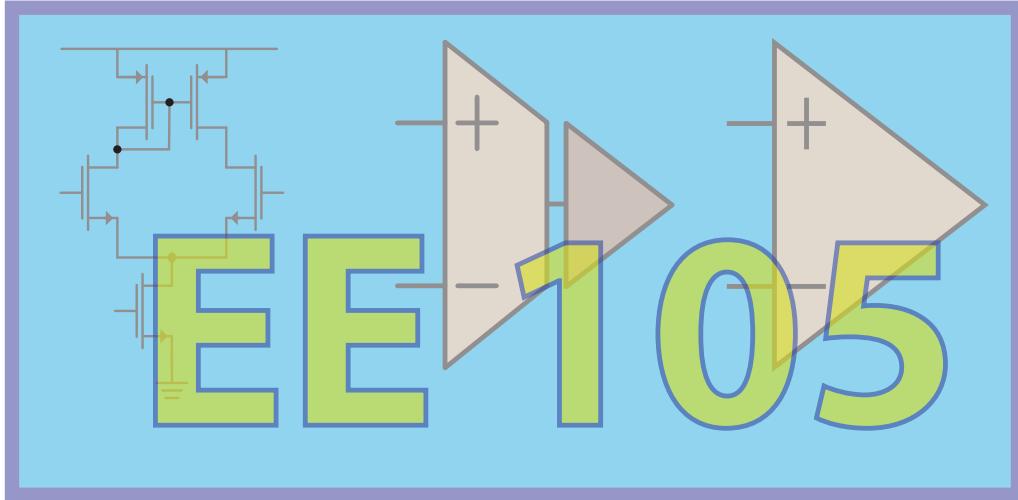
In *Eq. 6.43* we have used common units of eV , because most band gaps are also reported in this energy unit. In *Table 6.1*, we calculate the wavelength based on the reported band gap of several commonly used compound semiconductors. These values can be found in references and on the web, although there is some variation in the literature.

Table 6.1: Commonly used compound semiconductors

	E_g (eV)	λ (μm)	Color
InAs	0.35	3.51	Infrared
InN	0.65	1.91	Infrared
InP	1.34	0.92	Infrared
GaAs	1.43	0.87	Red
GaP	2.26	0.55	Yellow
AlP	2.45	0.51	Green
GaN	3.40	0.37	Blue
AlN	6.20	0.20	UV

6.8 References

Much of this material comes from teaching EE 105 at UC Berkeley for several years. The notes have benefited greatly from Pierret[5] (*Semiconductor Fundamentals: Volume I*), and most recently from Chenming Hu[2] (*Modern Semiconductor Devices for Integrated Circuits*).



7. MOS Capacitor

7.1 Chapter Preview

In this chapter we will be learning about the Metal-Oxide-Silicon (MOS) capacitor (MOS-C) structure (see *Fig. 7.1*). Like most capacitors, it has two plates separated by an insulator. One plate, the "gate", is nominally made of either a metal, or a heavily doped semiconductor that for all intents and purposes acts like a metal. The other plate, the "body", is made of a semiconductor like silicon, and this makes the behavior very different from a regular (metal-insulator-metal) capacitor. It is important to understand this behavior in detail, and the different modes of operations in a MOS-C. We shall find that the MOS-C structure behaves very much like a *PN*-junction in the that it has a built-in potential and a depletion region. This causes the capacitor to have built-in charge, even with zero bias and a thicker effective insulator in certain regions of operation. We will also demonstrate that there is a threshold voltage that, if exceeded, causes the bottom plate surface to "invert" from one type of semiconductor to the opposite kind. When the surface of the bottom plate reaches this inversion state, a conducting charge sheet is formed that acts like the second plate. The threshold voltage is extremely important when we discuss MOS transistors in the next chapter, because this charge sheet will be the channel for current conduction. We will also investigate the fields and potentials in various parts of the device in the various regions of operation. We will conclude the chapter by discussing the quantitative Q - V and C - V curves.

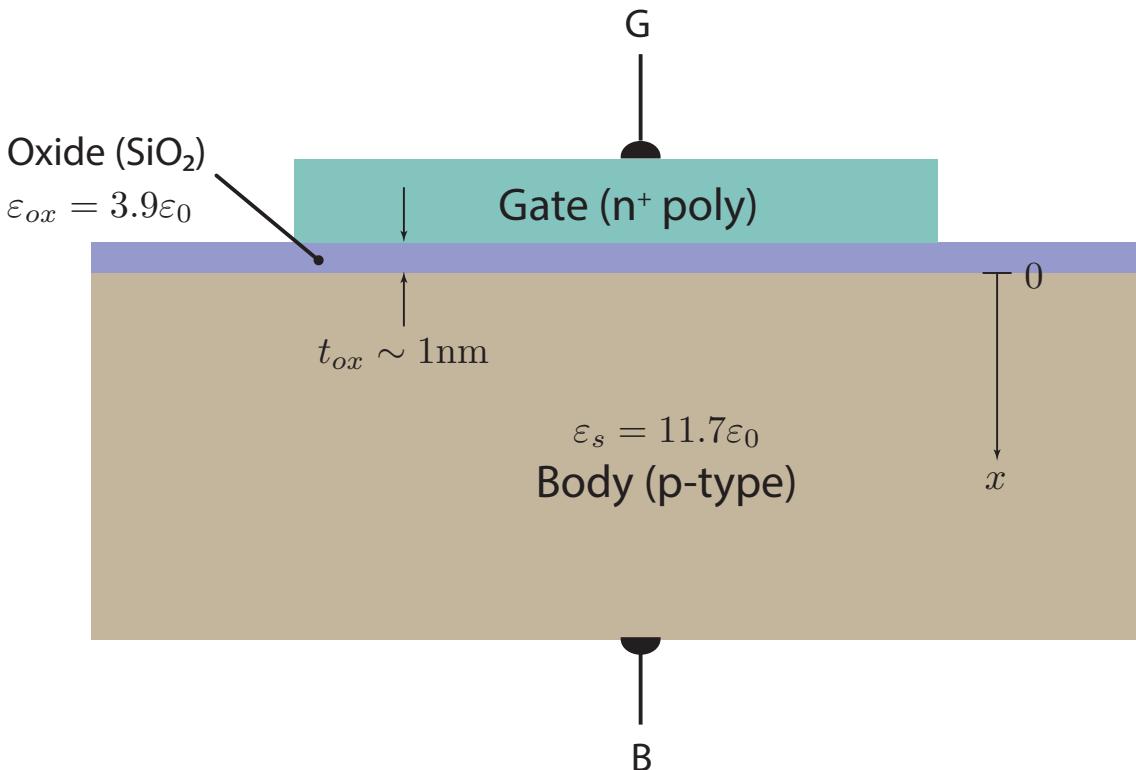


Figure 7.1: The Metal-Oxide-Semiconductor Capacitor (MOS-C) structure. The structure has a "gate" (G) terminal and a "body" (B) terminal. As shown, it is an NMOS-C realized with an N-type polysilicon gate (heavily doped) and a P-type body.

7.2 MOS Capacitor Structure

7.2.1 MOS Capacitor

The **Metal Oxide Semiconductor Capacitor** (MOS-C) structure shown in *Fig. 7.1*, is a sandwich of conductors separated by an insulator. The device has two terminals, the **gate** and the **body**, and a voltage is applied between the gate and body through these terminals. The “metal” gate is more commonly a heavily doped polysilicon (poly-*Si*) layer doped *n*⁺ or *p*⁺. The gate was originally made of metal (e.g. *Al*) until around 1970, but changed to poly-*Si* due to high temperature processing. After 2008, metal gates have been reintroduced for various technical reasons that we will not discuss in this book. Because the body is a semiconductor, there are two possible flavors of MOS-C devices. The **NMOS capacitor** uses a *P*-type substrate, and the **PMOS capacitor** uses an *N*-type substrate. Note that this is not a typo! The body doping is opposite to the flavor, something which we will clarify in this chapter. Finally, the insulator can in theory be any non-conductor, but in practice for silicon MOS structures *SiO*₂ is the oxide of choice. This is because it is native to *Si* (easy to grow, very low surface defects), and it can double as a mask opening during fabrication (blocking implants). Ideally this layer is very thin, with modern devices using a few layers of atoms to realize the lowest thickness possible. Because there is a physical limit to how thin we can make this layer¹, modern devices may use a stack up of layers including higher **dielectric constant** (high "K") insulators. Native *SiO*₂ has a relative permittivity of 3.9, which is about three times smaller than silicon.

¹Due to leakage currents through the gate and the fact that a very thin layer can undergo oxide breakdown since even modest gate-to-body voltages impose large electric fields.

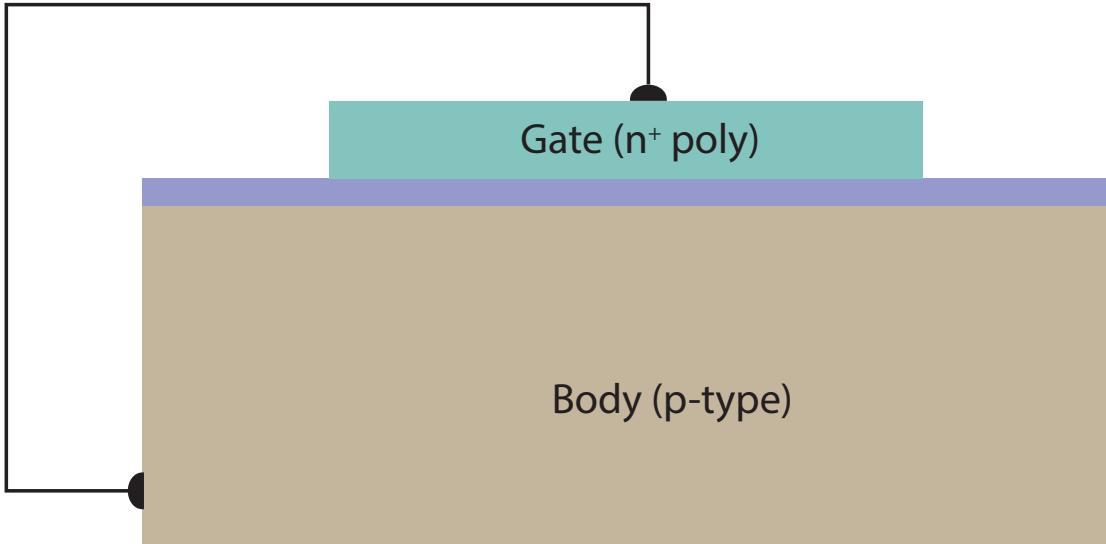


Figure 7.2: The MOS-C with gate/body shorted in equilibrium.

7.2.2 Metal-Oxide-Semiconductor Junction

Consider applying zero volts between the gate and the body, or in other words shorting the gate to the body as shown in *Fig. 7.2*. With this connection, we are allowing currents to flow between the gate and body. In particular, due to diffusion currents, we expect many electrons to leave the gate and enter the body. Likewise, holes in the body will naturally diffuse to the gate. Under thermal equilibrium, current flow will cease and therefore the *N*-type poly gate will rise to a higher potential than the *P*-type substrate, just like in a *PN*-junction diode. Using intrinsic silicon as a reference, the potential of the *P*-type region is given by:

$$\varphi_p = - \left(\frac{kT}{q} \right) \cdot \ln \left(\frac{N_A}{n_i} \right) \quad (7.1)$$

The potential of the *N*-type gate is given by:

$$\varphi_{poly,n^+} = \left(\frac{kT}{q} \right) \cdot \ln \left(\frac{N_{d,poly}}{n_i} \right) \approx 550 \text{ mV} \quad (7.2)$$

Note that the *N*-type gate is heavily doped, so we can approximate the potential as 550 mV if the doping level is not given. In equilibrium, no current can flow because the insulator blocks DC currents. But from our knowledge of *PN*-junctions, and with reference to *Eq. 7.1* and *Eq. 7.2*, we know the flow of majority carrier diffusion will stop due to the **built-in potential difference**:

$$\boxed{\varphi_{bi} = \varphi_{poly,n^+} - \varphi_p} \quad \text{Built-in potential, NMOS-C} \quad (7.3)$$

This potential difference is accompanied by an electric field and fields terminate on charge. Where are these fields and charges?

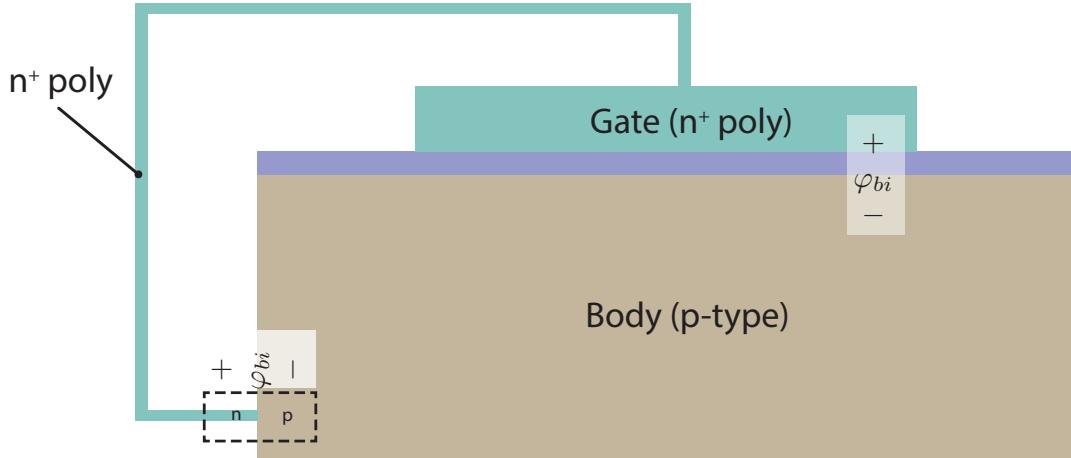


Figure 7.3: The MOS-C structure with gate-to-body shorted using a polysilicon gate material as the routing wire. The built-in potential φ_{bi} that develops across the PN -junction results also in a potential drop across the oxide.

7.2.3 Gate Materials and Contact Potential

What about contact potentials?

Technically, so far we have actually been analyzing the structure shown in *Fig. 7.3*. In other words, we have been assuming that the wires connecting the gate to the body are made of the same material as the gate. There are two built-in potentials that cancel out as we go around the loop. One built-in potential is between the PN -junction (the body to poly- n^+ wire connection), which is just like any other PN -junction that we have studied so far. The other built-in potential is dropped across the oxide layer between the body and gate.

Real Wires

In practice, the wire interconnect is not the poly- n^+ gate material,², but instead a metal like aluminum or copper. The complete structure is shown in *Fig. 7.4*, which shows that there is a p^+ region in the body that forms a contact to the body. Heavily doped diffusion regions form “ohmic” contacts, rather than **rectifying contacts**. Some contact engineering is also required to make the contact between the metal and the gate (not shown). In essence, suitable materials are used to avoid rectifying contacts.

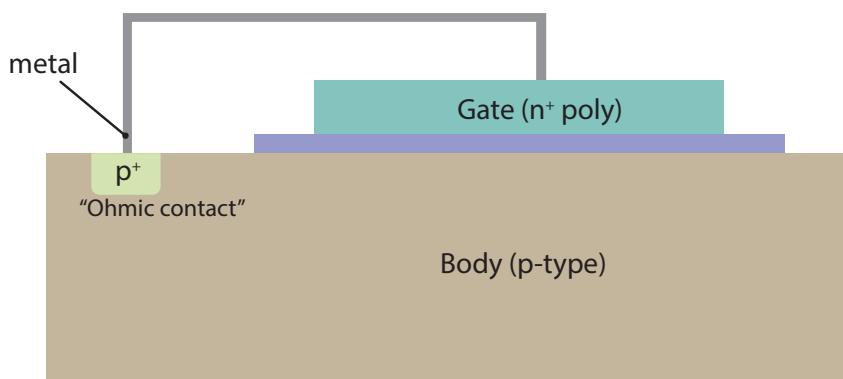


Figure 7.4: In practice, the gate to body connection is made partially with aluminium or copper wires. Special material selection is required (not shown) to avoid making rectifying contacts between the metal layers and the semiconductors. A p^+ heavily doped region acts as a contact point for the body.

²Poly gate is often used for some of the routing in integrated circuits, but most routing is with metal layers

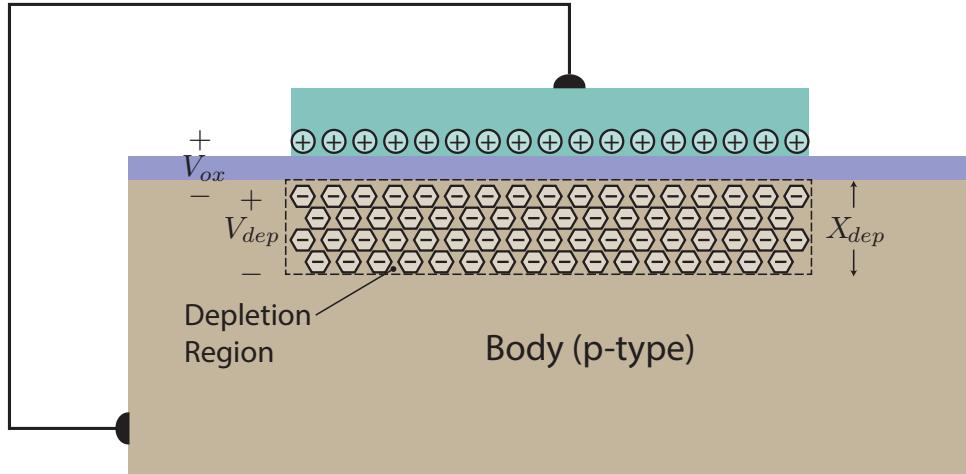


Figure 7.5: In equilibrium, there is a potential drop across the oxide and the silicon surface (depletion region) corresponding to the built-in potential between the gate and body.

7.3 MOS Regions of Operation

7.3.1 Fields and Charge at Equilibrium

As shown in *Fig. 7.5*, in **equilibrium** there is an electric field from the gate to the body due to the difference in the materials that make up the gate and the body. Based on the sign of the voltages, with the gate at a higher potential, we should find a net positive charge on the gate, and negative charge in substrate. Since body is *P*-type, negative charges in the body are the minority carriers, which means they are very small in number. So the charge comes mostly from the formation of the depletion region.

7.3.2 Flat-Band Voltage, $V_{GB} = V_{FB}$

If we apply a bias, we can compensate for this built-in potential to "reset" the capacitor, as evident in *Fig. 7.6*. The required voltage is simply opposite to the built-in potential we saw in *Eq. 7.3*:

$$V_{FB} = -\phi_{bi} = -(\phi_{n^+} - \phi_p) \quad \text{Flat-band voltage, NMOS-C} \quad (7.4)$$

In this scenario the charge on the gate goes to zero and the depletion region disappears. In device physics lingo, the **energy bands** are "flat" under this condition, giving rise to the name "**Flat-Band Voltage**".

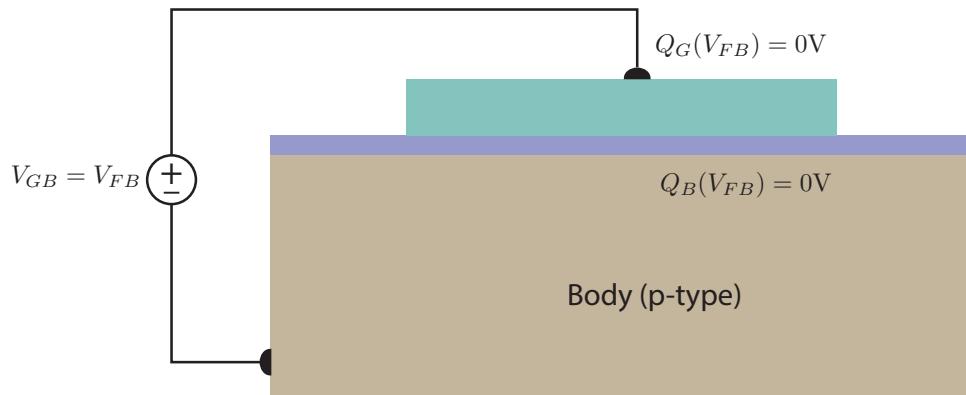


Figure 7.6: The flat-band voltage is defined as the gate-to-body voltage that results in net zero charge and zero fields in the MOS-C structure.

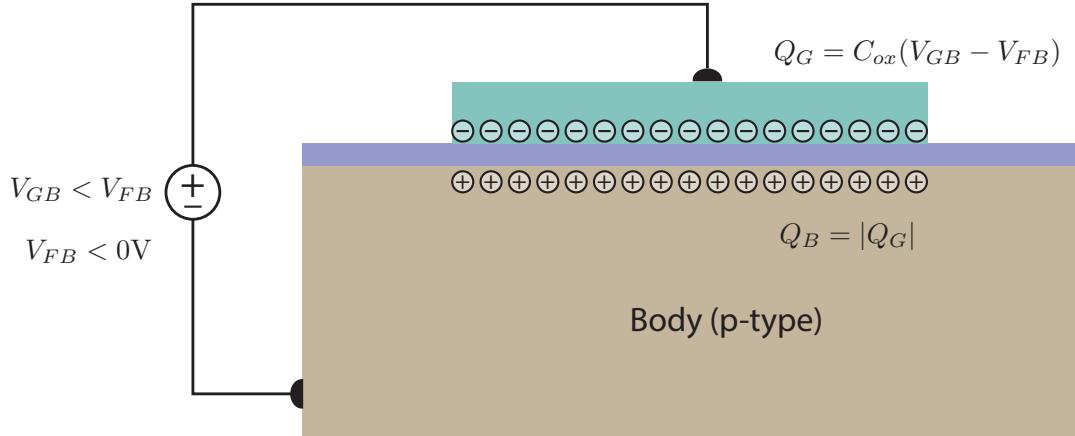


Figure 7.7: In accumulation, a gate-to-body (V_{GB}) voltage less than the flat-band voltage is applied (V_{FB}), resulting in the accumulation of holes at the surface of the semiconductor, which terminate the fields in the oxide that originate from electrons in the gate.

7.3.3 Accumulation, $V_{GB} < V_{FB}$

If we further decrease the potential beyond the “flat-band” condition, we essentially have a parallel plate capacitor, as shown in Fig. 7.7. Plenty of holes and electrons are available to charge up the plates, because we are making the gate more negative than the body. So negative charges (electrons) flow into the gate, and holes can flow to, or **accumulate** on, the surface of the device. The negative bias (strong fields from the oxide) attracts the holes to reside under gate.

7.3.4 Depletion, $V_{GB} > V_{FB}$

In the **depletion** regime, shown in Fig. 7.8, the situation is similar to equilibrium. Since the potential in the gate is higher than the body, the body charge is made up of the depletion region (ionized group-III dopants for the *P*-type body), and there is a potential drop across the body and depletion region. One way to understand this equilibrium scenario is to imagine what happens as the source voltage is increased beyond the flat-band condition. At flat-band, by definition, there is no net charge in the gate or body. As we apply a small voltage, positive charges enter the gate (electrons leave the gate), and negative charges (electrons) enter the body. Since the body is *P*-type, with the majority carriers being holes, most of these electrons will quickly recombine.

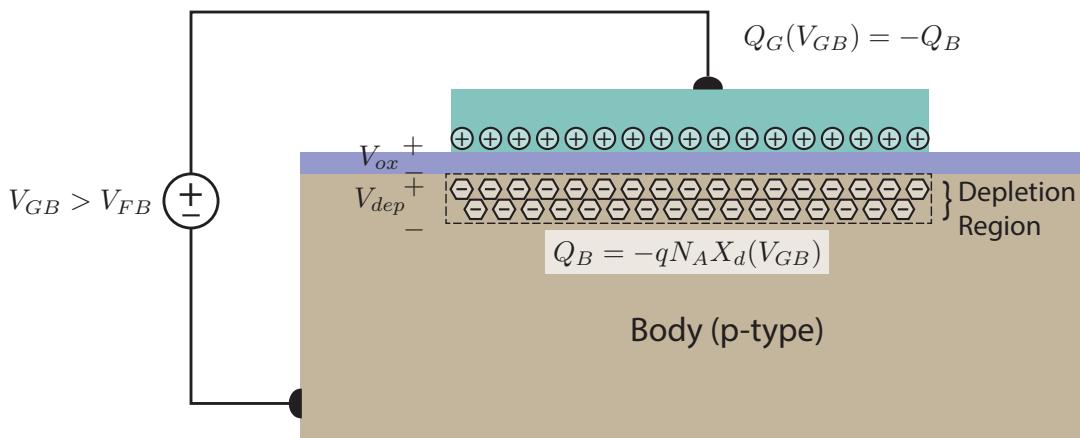


Figure 7.8: In depletion region, a gate-to-body (V_{GB}) voltage greater than the flat-band voltage is applied (V_{FB}). The voltage should not be too large (less than inversion, defined below). In this region, the oxide fields terminate on ionized dopant atoms.

Recall that electric fields diverge (terminate) on charges. Since there is a net positive charge in the gate, we have electric field lines that flow through the oxide and penetrate the silicon. This non-zero body field will move holes away from the surface through the action of drift current. This process will continue as long as field lines penetrate the body. Notice that as holes are repelled from the surface, a net negative charge emerges due to the depletion region. The action of the depletion region is to reduce the field strength, because some of the field lines terminate on the ionized dopants. Eventually, the surface will be completely depleted of positive charge, and all the fields will terminate on ionized dopants. Since the ionized dopants cannot move, not all the fields will terminate at the surface, but rather throughout the depletion region.

7.3.5 Inversion, $V_{GB} = V_T$

In the **inversion** regime, shown in Fig. 7.9, as we further increase the gate voltage, because not all the charge resides at the surface, there is a larger and larger voltage drop across the depletion region. In other words, the gate is "pulling" the surface of the body to a higher potential. A higher surface potential makes it more favorable for electrons to stick around at the surface. Recall that thermal **generation** continually creates electron-hole pairs, and the steady-state number of electron/holes is regulated by **recombination**. At the surface of the depletion region, electron-hole pair generation creates a situation that favors electrons because the holes are repelled from the surface and the electrons are attracted to the surface. Eventually the surface potential increases to a point where the electron density at the surface equals the background ion density:

$$n_s = n_i e^{(q\phi_s/kT)} = N_A \quad \text{Electron density at surface} \quad (7.5)$$

$$\phi_s = -\phi_p \quad (7.6)$$

At this point, the depletion region stops growing (effectively) and the extra charge is provided by the inversion charge at surface. "Inversion" meaning that the surface is effectively *N*-type, even though it started as a *P*-type material.

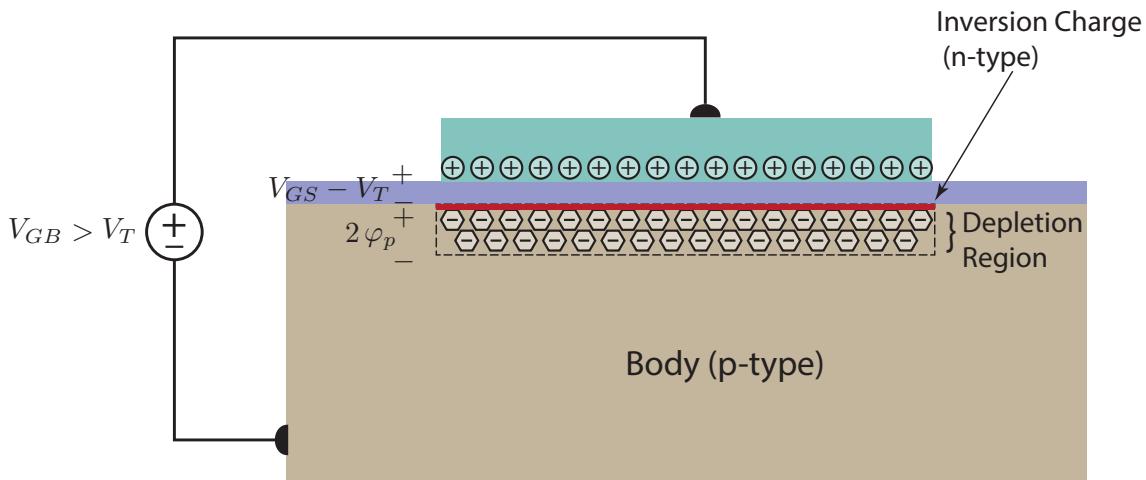


Figure 7.9: Inversion is defined as the gate-to-body (V_{GB}) voltage sufficient to invert the surface of the semiconductor from *P*-type to *N*-type, causing a sheet charge of electrons to accumulate at the surface.

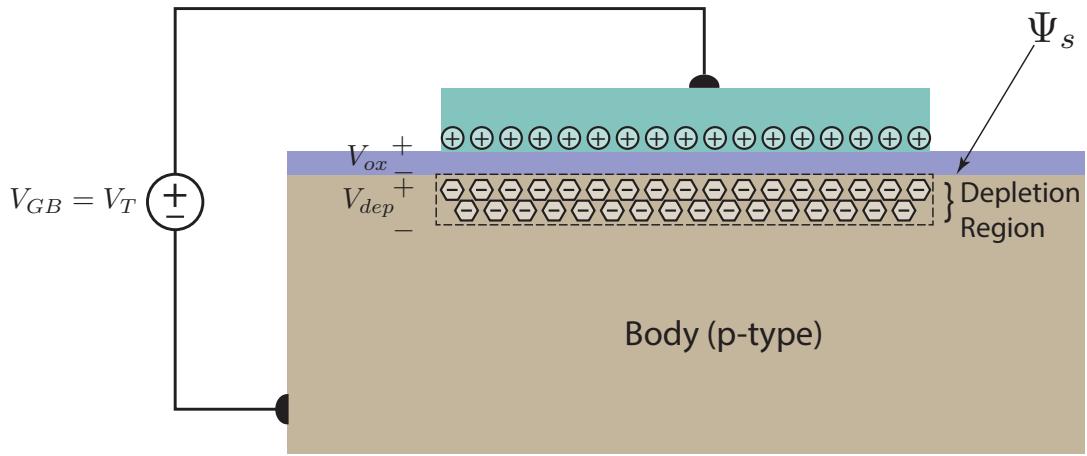


Figure 7.10: To derive the threshold voltage, it is important to realize that the gate-to-body voltage is dropped across both the oxide and the depletion region. In absence of significant inversion charge at the surface, the total gate charge is compensated for entirely by depletion charge.

7.4 MOS Device Threshold Voltage

7.4.1 Threshold Voltage Definition

The **threshold voltage** is defined as the gate-body voltage that causes the surface to change from *P*-type to *N*-type. For this condition, the **surface potential** has to equal the negative of the *P*-type potential. We will show step-by-step that this voltage is equal to:

$$V_{T_n} = V_{FB} - 2\phi_p + \frac{1}{C_{ox}} \sqrt{2q\epsilon_s N_A (-2\phi_p)} \quad \text{Threshold voltage, NMOS-C} \quad (7.7)$$

7.4.2 Derivation of V_T

As shown in *Fig. 7.10*, the gate-to-body voltage is dropped across the oxide and the depletion region:

$$V_{GB} = V_{ox} + V_{dep}$$

To see this, do a KVL loop around the structure when a voltage source is added. Recall from *Fig. 7.3* that the built-in voltage cancels out around the loop since it appears twice with opposite polarity. To cause inversion, we increase the gate voltage until the surface potential becomes sufficiently positive relative to the bulk. Once we have reached the inversion regime the electron density is the same as the body hole density. This means that compared to the bulk, the potential at the surface is at $-2\phi_p$. This voltage is dropped across the depletion region:

$$V_{dep} = -2\phi_p$$

If we apply a gate-to-body voltage of V_{FB} , then we know the fields and charge go to zero. This is a good starting point. Next, the oxide voltage is calculated from the gate/bulk charge. At the edge of inversion, the total bulk charge for inversion is given by:

$$Q_{dep} = \sqrt{2q\epsilon_s N_A (-2\phi_p)}$$

This means the voltage drop across the oxide is given by

$$V_{ox} = Q_{dep}/C_{ox}$$

Putting this all together, we have:

$$\begin{aligned}
 V_{GB} &= V_T = V_{FB} + Q_{dep}/C_{ox} - 2\phi_p \\
 &= V_{FB} - 2\phi_p + Q_{dep}/C_{ox} \\
 &= V_{FB} - 2\phi_p + \frac{1}{C_{ox}} \sqrt{2q\epsilon_s N_A (-2\phi_p)}
 \end{aligned}$$

Notice that we made some simplifying assumptions to arrive here. We assumed zero inversion charge to reach this point, which of course is not quite correct. However, it is a good approximation because the total inversion charge is still quite low, even if the density is high at the surface. We also made the standard depletion approximation, and assumed that the surface depletion region is fully depleted. This follows from the same arguments that we made when deriving the depletion region for the *PN*-junction.

7.5 Fields and Potential in Oxide/Substrate

7.5.1 Fields in oxide/substrate

The electric field in the MOS-C structure is shown schematically and graphically in Fig. 7.11. Notice that the field in the oxide is constant, because the oxide is a good insulator and there are no charges in the oxide. The oxide does exert itself through the permittivity, and because silicon has a dielectric constant that is about three times higher, the fields experience a discontinuity crossing the oxide/body boundary. Also, the fields do not drop to zero in the body instantly because the body charges are distributed throughout the depletion region. Under the assumption of uniform doping, the fields drop linearly until they return to zero at the edge of the depletion region.

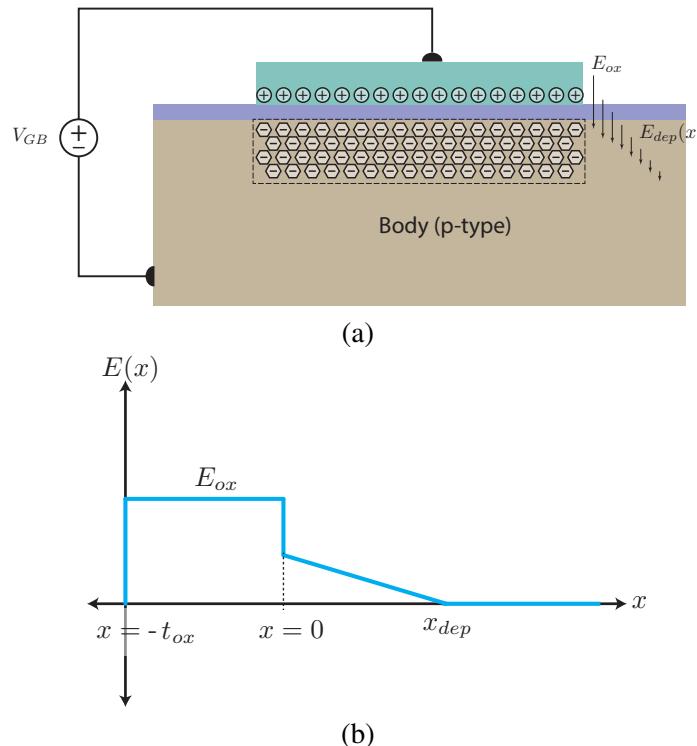


Figure 7.11: The electric field inside the oxide and transistor body shown (a) schematically, and (b) graphed as a function of x . The discontinuity in the electric field is due to the difference in dielectric permittivity between SiO_2 and Si . The linear variation in the depletion region is due to the uniform doping concentration of acceptor atoms.

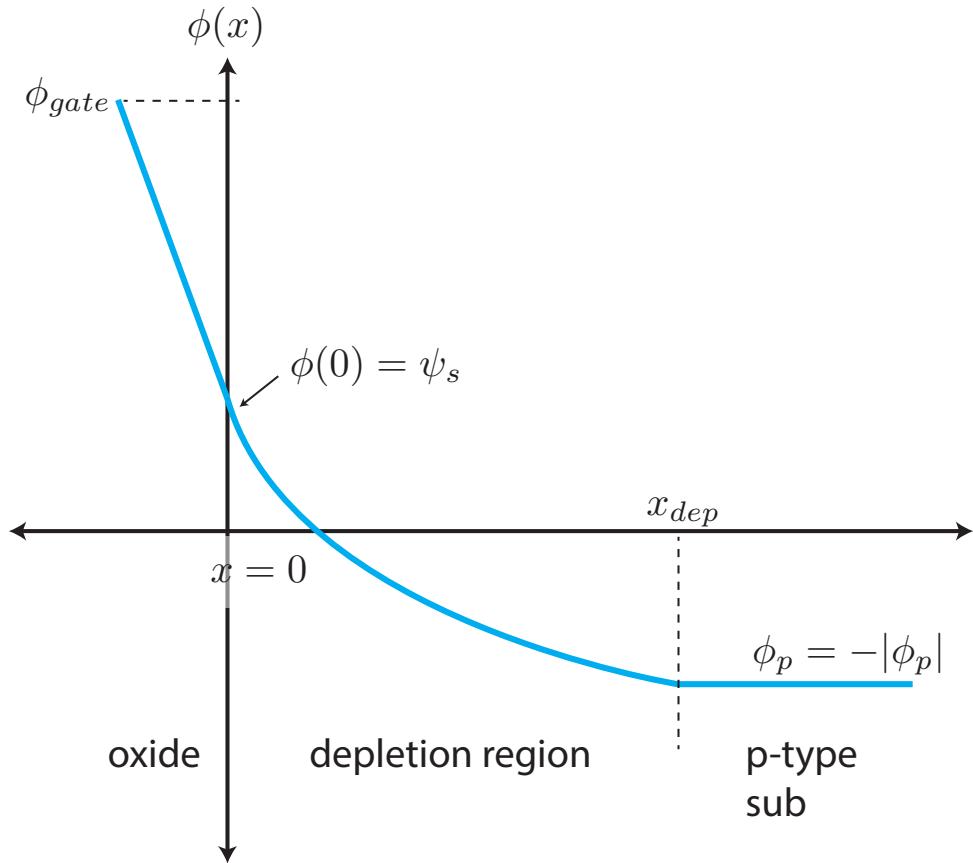


Figure 7.12: The variation of electric potential through the oxide and depletion region.

7.5.2 Potential Variations in Oxide/Substrate

The variation of the potential, the integral of the fields in *Fig. 7.11*, is shown in *Fig. 7.12*. The potential transitions smoothly across the oxide into the body. The potential varies linearly in the oxide, and then quadratically in the body. From this figure it is clear that pulling the gate voltage high pulls the surface potential to a value above the *P*-type region potential, thereby eventually transforming it from *P*-type to *N*-type.

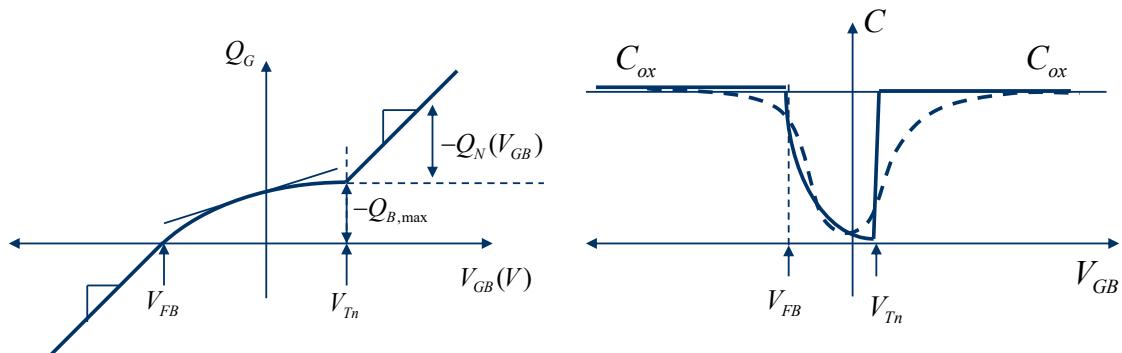


Figure 7.13: The plot of charge vs. voltage Q - V and small-signal capacitance vs. voltage C - V for an MOS-C structure. The breakpoints occur at the flat-band voltage V_{FB} and the threshold voltage V_T . The dashed line in the C - V curve represents a more physical change in capacitance due to higher-order effects not discussed in this chapter.

7.6 Charge-Voltage (Q - V) and Capacitance-Voltage (C - V) Curves

7.6.1 Q - V Curve for MOS Capacitor

A plot of the charge versus voltage (Q - V) is shown in *Fig. 7.13*. The charge is not a linear function of voltage, and so the capacitance curve (C - V), shown in the same figure, is really the small-signal capacitance, which is the derivative of the Q - V relation. In accumulation, the charge is simply proportional to the applied gate-body bias, because the body acts like a second gate. In inversion, the same is true because the surface of the structure is a charge sheet that terminates all additional field lines. Notice that this curve is only valid under **quasi-static** conditions. In other words, we have to allow enough time for thermal generation to supply electrons to the surface. If a rapid voltage (like a step transition) is applied, there is insufficient time for generation and the response would have to come from the depletion region temporarily until sufficient time elapses for surface electron density to build up.

The most complicated part of this curve is the depletion region. The charge grows slower than linear since the voltage is applied over a depletion region, rather than just through the oxide.

7.6.2 MOS C - V Curve

Since the small-signal capacitance is slope of Q - V curve, it is not very difficult to see how the curve shown in the second part of *Fig. 7.13* arises. The capacitance is constant in accumulation and inversion because in these regions the device is acting like a simple parallel plate capacitor. The capacitance in the depletion region is smallest because the voltage drop occurs both across the oxide and across the depletion region. The capacitance is non-linear in the depletion region.

7.6.3 C - V Curve Equivalent Circuits

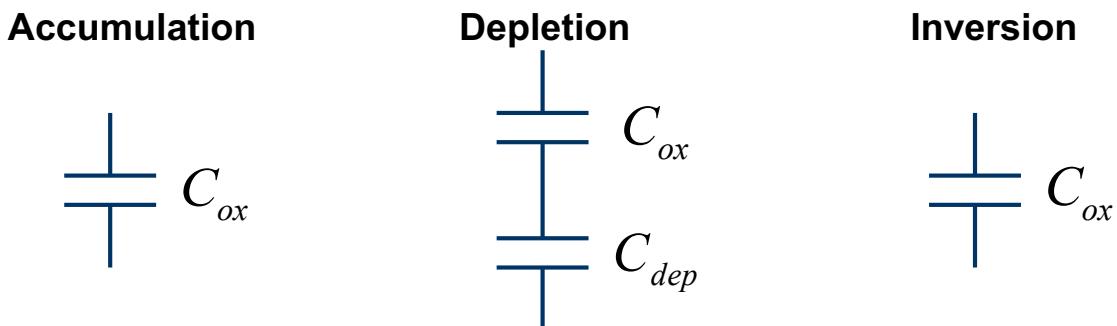


Figure 7.14: The equivalent circuit for a MOS-C capacitor for small-signal perturbations about a fixed operating point in accumulation, depletion, and inversion.

Equivalent small-signal circuits for the MOS-C are summarized in *Fig. 7.14*. In accumulation mode, the capacitance is just due to the voltage drop across t_{ox} , and we can calculate this capacitance per unit area from the parallel plate formula:

$$C_{acc} = C_{inv} = C_{ox} = \frac{\epsilon_{ox}}{t_{ox}} \quad \text{Capacitance, accumulation and inversion regime} \quad (7.8)$$

The equivalent circuit applies in inversion. Let's pause and discuss this in a bit more detail. We are talking about *small-signal* or incremental charges, not total charge. So if the MOS-C is biased in accumulation, and an incremental voltage is applied to the device, the additional charges flow into the surface of the body and gate, and C_{ox} is the relation between incremental charge and voltage.

In inversion, there is already a depletion region, but we are assuming that when an incremental voltage is applied that the depletion region does not grow. Thus, all additional charges will flow to

the surface of the device in the form of the inversion layer. This is why the small-signal equivalent circuit in inversion does not have anything about depletion in it. Anything that is "frozen" in place has no impact on the small-signal response.

In the depletion region, the voltage drop is across the oxide and the depletion region. Let's define the depletion region capacitance as follows:

$$\boxed{C_{dep} = \frac{\epsilon_s}{x_{dep}}} \quad \text{Capacitance, depletion regime} \quad (7.9)$$

Then we can say that in the depletion region, the gate-body incremental voltage "sees" two capacitors in series, as shown in the equivalent circuit:

$$C_{tot} = \frac{C_{dep} C_{ox}}{C_{dep} + C_{ox}} = \frac{C_{ox}}{1 + \frac{C_{dep}}{C_{ox}}} \quad (7.10)$$

$$= \frac{C_{ox}}{1 + \frac{\epsilon_s t_{ox}}{\epsilon_{ox} x_{dep}}} \quad \text{Total series capacitance in depletion regime} \quad (7.11)$$

Since x_{dep} depends on bias, we evaluate x_{dep} at the value of the operating point of the circuit.

7.6.4 Numerical Example

Let's work through the numbers to get a feel for the MOS capacitor. Assume that it is a *P*-type substrate with an oxide thickness of $t_{ox} = 20\text{ nm}$, and a body doping concentration of $N_A = 5 \times 10^{16}\text{ cm}^{-3}$. This gives rise to an oxide capacitance of:

$$C_{ox} = \frac{\epsilon_{ox}}{t_{ox}} = \frac{3.45 \times 10^{-13} \frac{\text{F}}{\text{cm}}}{2 \times 10^{-6} \text{cm}} \approx \boxed{172.5\text{ nF}}$$

With reference to *Eq. 7.1* we can find that:

$$\phi_p = - \left(\frac{kT}{q} \right) \cdot \ln \left(\frac{N_A}{n_i} \right) = -(0.026\text{V}) \cdot \ln \left(\frac{5 \times 10^{16}\text{ cm}^{-3}}{1 \times 10^{10}\text{ cm}^{-3}} \right) \approx \boxed{-401\text{ mV}}$$

Let's first calculate the flat-band voltage. Because we do not have any information about the gate doping level, let's assume that it is **degeneratively doped**, so the potential of the poly- n^+ gate is 550 mV :

$$V_{FB} = -(\phi_{n^+} - \phi_p) = -(550\text{ mV} - (-401\text{ mV})) = -(550\text{ mV} + 401\text{ mV}) = \boxed{-0.952\text{ V}}$$

Next, let's calculate the threshold voltage:

$$\begin{aligned} V_{T_n} &= V_{FB} - 2\phi_p + \frac{1}{C_{ox}} \sqrt{2q\epsilon_s N_A (-2\phi_p)} \\ &= -(0.952\text{ V}) + ((-2)(-0.401)) \\ &\quad + \frac{\sqrt{2(1.602 \times 10^{-19}\text{ C})(1.0359 \times 10^{-12} \frac{\text{F}}{\text{cm}})(5 \times 10^{16}\text{ cm}^{-3}) - 2(-0.401\text{ V})}}{1.725 \times 10^{-7}\text{ F}} \\ &= -0.952\text{ V} + 0.802\text{ V} + 0.668\text{ V} \\ &\approx \boxed{0.52\text{ V}} \end{aligned}$$

Now let's apply a gate-to-body voltage:

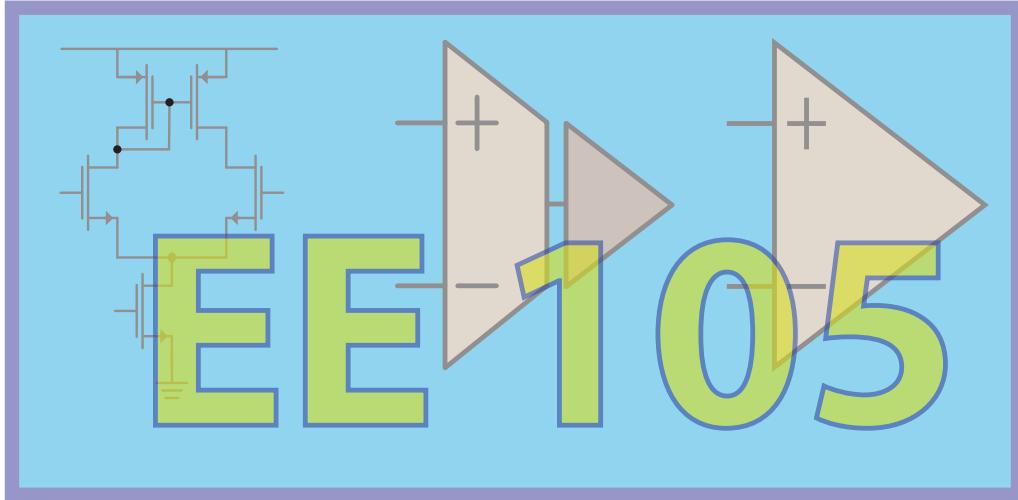
$$V_{GB} = -2.5 < V_{FB}$$

Since the device is in accumulation, the entire voltage drop is across the oxide, and the oxide fields are given by:

$$\begin{aligned}\mathcal{E}_{ox} &= \frac{V_{ox}}{t_{ox}} = \frac{V_{GB} + \phi_{n^+} - \phi_p}{t_{ox}} \\ &= \frac{(-2.5\text{V} + 0.55\text{V} - (-0.401))}{2 \times 10^{-6}\text{cm}} \\ &= \frac{-1.549\text{V}}{2 \times 10^{-6}\text{cm}} \\ &= -8 \times 10^5 \frac{\text{V}}{\text{cm}}\end{aligned}$$

The charge in the substrate (body) consist of holes:

$$\begin{aligned}Q_B &= -C_{ox}(V_{GB} - V_{FB}) \\ &= -(1.725 \times 10^{-7}\text{F}) \cdot (-2.5 - (-0.952\text{V})) \\ &= -(1.725 \times 10^{-7}\text{F}) \cdot (-1.548\text{V}) \\ &\approx 2.67 \times 10^{-7}\text{C/cm}^2\end{aligned}$$



8. MOSFET Device Physics

8.1 Chapter Preview

The naming of the transistor comes from blending two words: **transconductance resistor**. In this chapter we will be studying the MOSFET (pictured in Fig. 8.1), and you will see why the transistor gets its name. A MOSFET, or **MOS-Field-Effect-Transistor**, is a device constructed from a MOS capacitor by adding two extra diffusion regions. This is why it is important that we first studied the MOS-C in the previous chapter.

We begin the chapter by discussing the structure of the MOSFET, discussing the different flavors of MOS technology, including **NMOS** and **PMOS** devices, and **CMOS** technology. The difference between these devices will be discussed in detail. Then we will examine the current-voltage characteristics of a MOSFET, known as the *I-V* curves. Similar to a resistor, we can plot current versus voltage for a MOSFET, but given that the device has three terminals, we must plot families of curves. Once we have an phenomenological understanding of the MOSFET, we derive analytical equations for an ideal "long channel" device. This will lead us into understanding some important concepts in the MOSFET, such as the current saturation mechanism. Finally, we end the chapter by discussing the *PMOS* device, which is completely complementary to the *NMOS* device.

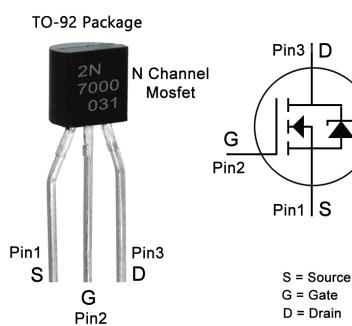


Figure 8.1: The schematic of a 2N7000 MOSFET. This is an *N*-channel device, which means that the "source" carries electrons, and the body is *P*-type.

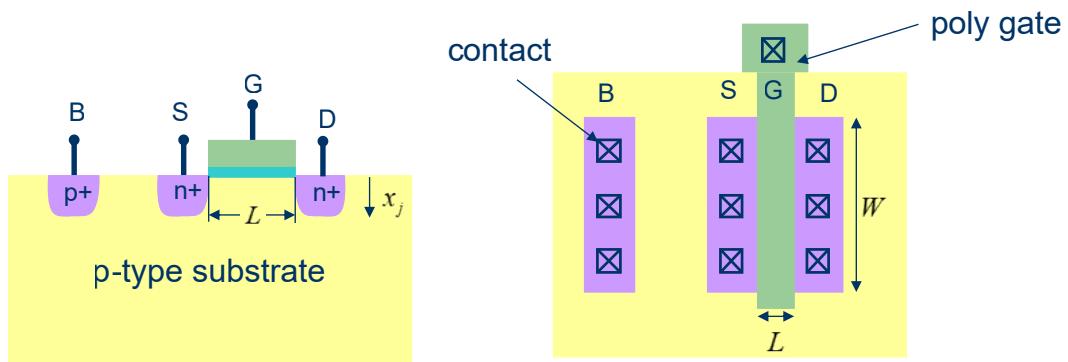


Figure 8.2: Cross section (left) and top-level layout of a MOSFET device (right). In general, the device has four terminals, labeled (G) gate, (B) body, (S) source, and (D) drain. The layout parameters include the width W and length L .

8.2 Device Layout and Cross Section

8.2.1 MOSFET Top View and Layout

The cross-section and top view of a **MOSFET** are shown in *Fig. 8.2*. A MOS transistor is a *four-terminal device* (although usually we only use three terminals) that conducts current between two of its terminals under the control of the third terminal, the "gate" (G). In a MOSFET, the terminals that carry current are known as the "drain" (D) and "source" (S), in analogy with liquid flow. The "source" supplies electrons (in an NMOS), and the *electrons flow from the source to the drain*. Since electrons carry negative charge, the *current flow is from the drain to the source*, which is confusing but standard convention. The fourth terminal is the "body" (B), which establishes a potential reference for the gate terminal. Often the body and source are shorted together, and this leads to a "source referenced" notion of the gate voltage. In general, it is the gate-to-body voltage that determines the operating point of the MOS-C, and therefore the MOSFET.

A top view of the MOSFET shown in *Fig. 8.2*. It is important to note that the dimensions L and W are layout dependent, and can be varied by the circuit designer. These parameters control the size of the channel that forms (from the source to the drain) which allows current to flow when the device is biased in inversion. The parameter W is the **device width**. The larger the width, the more current the transistor can conduct. The device dimension L is the **channel length**. The shorter the channel, the more conductive the channel.

The smallest possible dimension L_{min} is a technology-dependent parameter as manufacturers strive to improve the performance of the device by providing means of fabricating smaller and smaller devices. Today's transistors are fabricated with channel lengths as small as 7 nm but not all applications need such a small device. Generally speaking, smaller channel length leads to a smaller required W to get the same current, and therefore an overall smaller device. This favors high density logic, where millions to billions of transistors are fabricated on a single chip. In analog circuit applications, requirements are different and much larger devices may be employed. Today devices as large as $1\text{ }\mu\text{m}$ are used in power and automotive applications. Short channel devices down to 22 nm are routinely used for **radio frequency** (RF) applications.

Another thing worth mentioning is that MOSFET devices are "planar". In *Fig. 8.2*, note that x_j is the **junction depth**, and the channel will form at the surface of the device. While the metal stack-up (see *Fig. 8.4*) may be three dimensional, the actual transistors all reside in a thin layer at the surface of the body. The circuit designers usually only control the 2D layout of the device, as the third dimension (junction depth for instance) is set by technologists to achieve the best performance. The most important technology parameters, highlighted in *Fig. 8.3*, include the minimum channel length L_{min} , oxide thickness t_{ox} , the device body doping N_A , and the junction depth x_j .

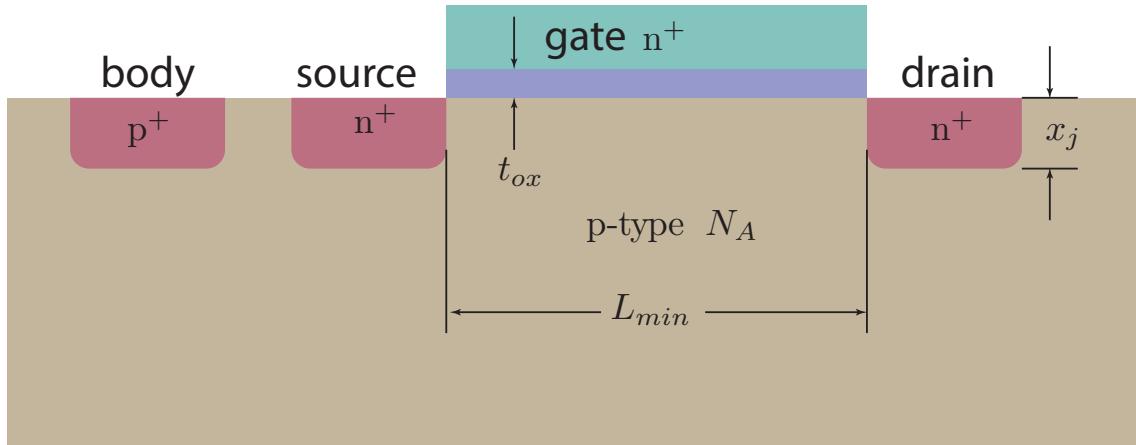


Figure 8.3: Cross section of a MOSFET device. Important technology parameters include the minimum channel length L_{min} , the body doping (N_A), the oxide thickness (t_{ox}), and the junction depth x_j .

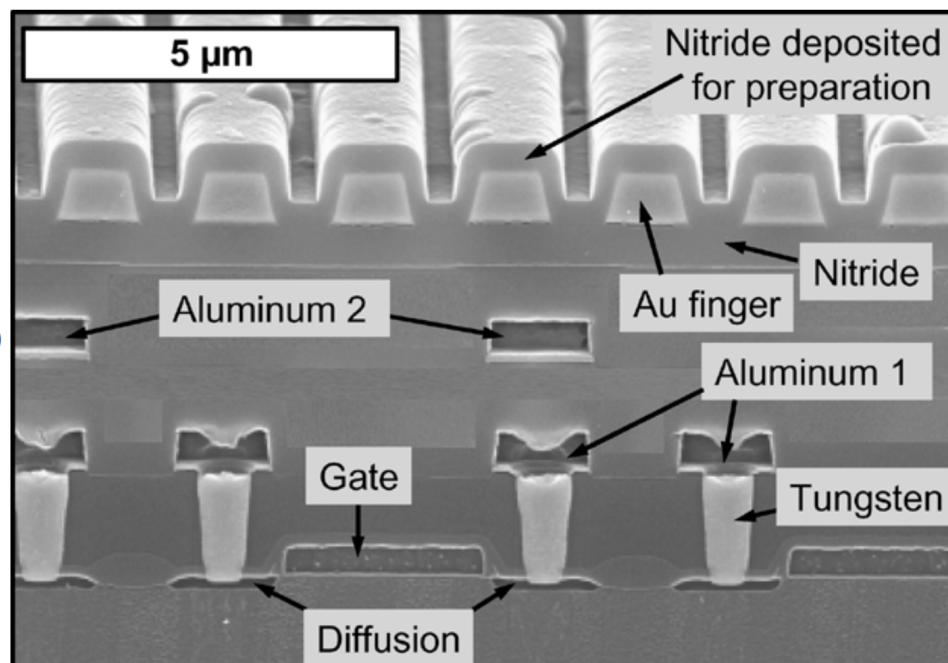


Figure 8.4: SEM cross-section photo of a CMOS chip. (Used without permission from M. Schienle et al., "A fully electronic DNA sensor with 128 positions and in-pixel A/D conversion," in *IEEE Journal of Solid-State Circuits*, vol. 39, no. 12, pp. 2438-2445, Dec. 2004.)

8.2.2 MOSFET Overview

With a solid understanding of a MOS capacitor, we now introduce the MOS transistor, or MOSFET. Note that to make a MOSFET, we start with a MOS capacitor and add two diffusion regions adjacent to the gate, shown in *Fig. 8.3*. Note the doping of these regions is opposite to the body of the device. For a *P*-type substrate, we dope the drain and source *N*-type and make contact with the source and drain terminals. These junctions form reverse biased (under proper biasing conditions) diodes between the source/drain and the body of the transistor. Unless we "turn on" the device, no current can flow between the source and drain because of the presence of two back-to-back diodes.

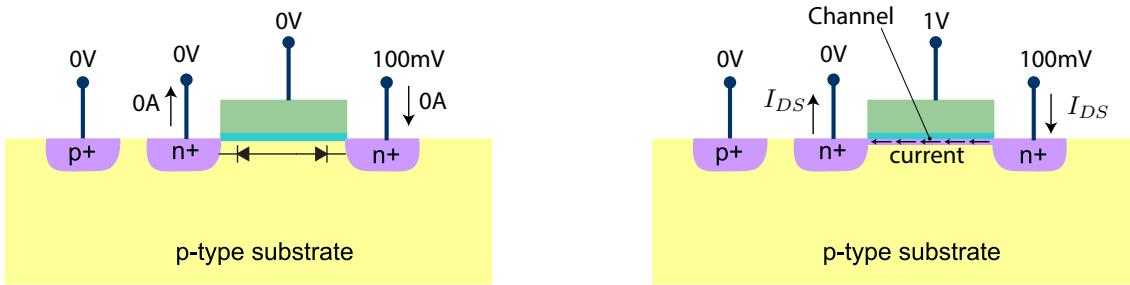


Figure 8.5: Cross section of an *NMOS* device with (a) a zero gate voltage is "off" and cannot conduct current, whereas (b) a sufficiently positive gate voltage is applied to turn "on" the device, allowing electrons to flow from the source to the drain.

When a sufficiently large gate voltage is applied, the surface of the MOS-C is inverted, creating a channel between the source and drain, allowing current to flow (see Fig. 8.5).

It is important to note that the MOSFET is a four terminal device, but in most situations the body terminal is connected to a fixed DC voltage, or shorted to the source. The reason for this will become clear later but for now you can assume that unless otherwise noted, the body of the device is connected to ground for NMOS devices and connected to the supply (or highest potential) for a *PMOS* device. These connections ensure that under normal operating conditions, when the source and drain voltages swing between ground and supply, the diodes remain reverse biased.

8.2.3 MOSFET Flavors: PMOS and NMOS

MOSFETs are also known by other names, namely the **IGFET**, or **Insulated Gate Field-Effect Transistors**. This name is not really in common use, and MOSFET is much more common. FETs are a general class of devices that include other devices such as **Junction-FETs** (J-FET or **JFET**) and **Metal-Semiconductor-FETs**, or **MESFETs**. We will focus on MOSFETs in this book, although the behavior of MOSFETs, JFETs, and MESFETs are all very similar from the high level perspective of design. In fact, we will show in the next chapter that by employing small-signal models, all devices essentially look the same.

For a given technology, it may be possible to fabricate two complementary devices, shown in Fig. 8.6. For example, in this chapter we'll see that we can fabricate a MOSFET on a *P*-type body (often the substrate), resulting in an **NMOS** or **NFET device**, or on an *N*-type body, resulting in a **PMOS** or **PFET device**. These devices are complementary in the sense that the currents and voltages are essentially all the negative of each other. We already saw this with a MOS-C, where for an *N*-type device a sufficiently positive gate voltage results in inversion, whereas for a *P*-type MOS-C, a negative gate-to-body voltage is needed. The same is true of MOSFET devices.

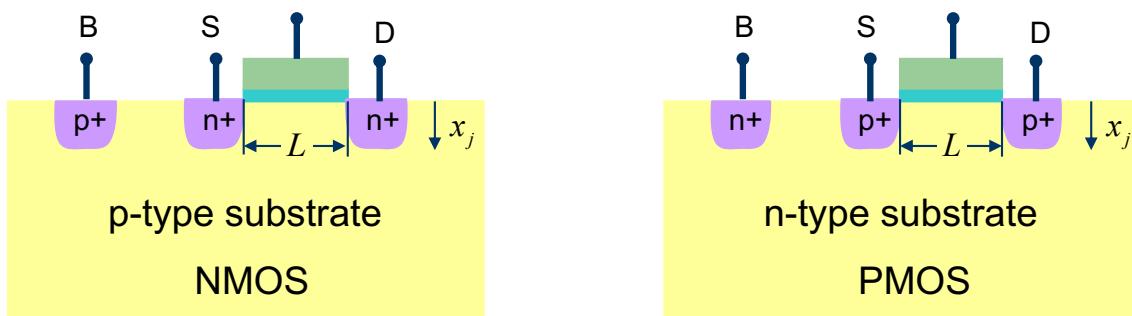


Figure 8.6: Cross section of a *PMOS* device is very similar to an *NMOS* device, except it is realized by using an *N*-type body doping and *p*⁺ doping diffusion regions to define the source and drain.

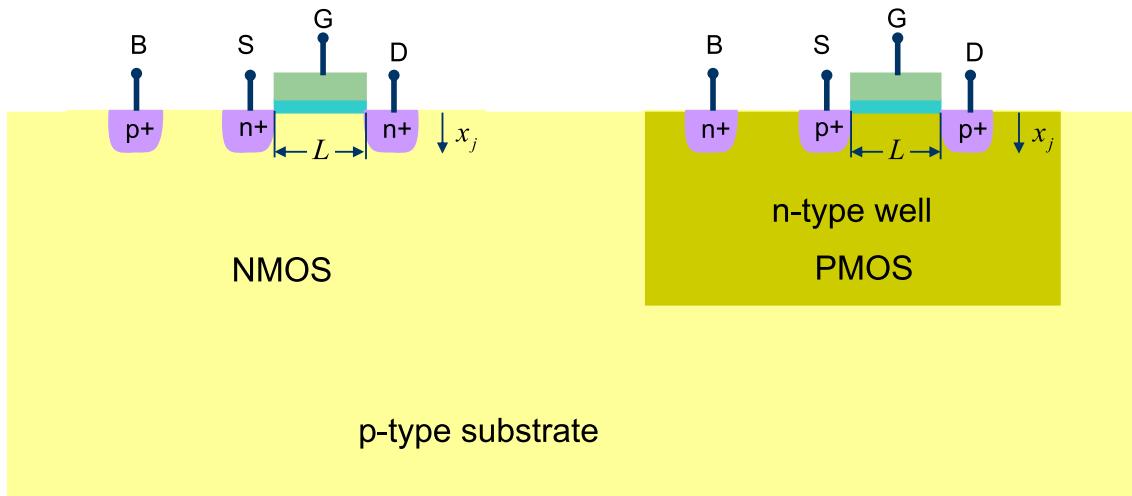


Figure 8.7: *CMOS* technology allows one to build both *NMOS* and *PMOS* devices in the same technology side-by-side as shown. *NMOS* devices are realized directly using the *P*-type substrate as the body whereas the *PMOS* device requires a "well", a region counter-doped *N*-type to realize the *PMOS* body.

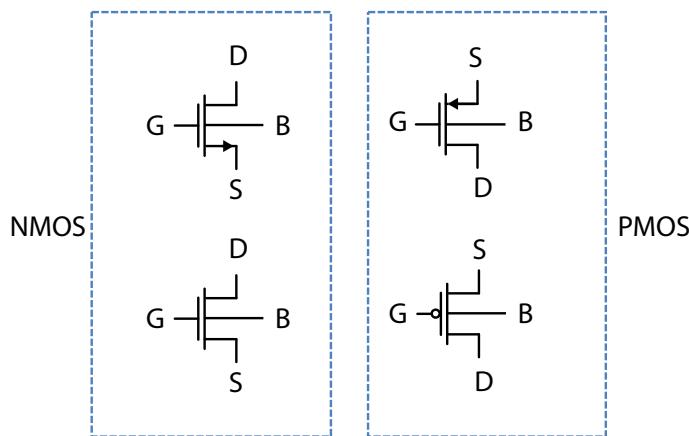


Figure 8.8: Schematic symbols for the *NMOS* and *PMOS* devices.

8.2.4 CMOS Technology

Finally, we should discuss the term "**CMOS**", which stands for "**Complementary MOS**", or a technology capability to fabricate both *NMOS* and *PMOS* devices. *CMOS* technology is now universal for silicon-based technology. As shown in *Fig. 8.7*, both devices are fabricated on the same substrate by creating wells (oppositely doped to the body) to house devices. With a *P*-type substrate shown, the *NMOS* devices can be fabricated directly in the body, and the *PMOS* devices are fabricated in so-called **N-well regions**. These **well-regions** are counter-doped to make them *N*-type. Then *P*-type source/drain junctions are added to make the *PMOS* device. The wells need to be biased such that at all times the well-to-body diode is reverse biased. The well bias is established through the body connection.

8.2.5 Circuit Symbols

In circuit schematics MOSFET symbols are drawn as shown in *Fig. 8.8*. The symbols with the arrows are more common in analog circuit applications, and the symbols with the bubble on the gate for *PMOS* are more favored in digital applications. The bubble emphasizes that the *PMOS* device is the logical "NOT" of the *NMOS* device. In other words, a low voltage is needed to turn on

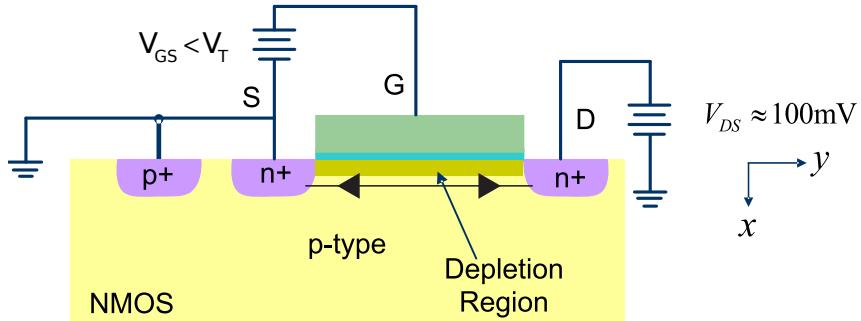


Figure 8.9: The MOS device biased in cut-off, or a gate bias $V_{GS} < V_T$, cannot conduct any current as the source and drain are isolated by two back-to-back diodes and a depletion region.

the *PMOS* device (with respect to the body), whereas for the *NMOS* device a high voltage (relative to the body) is needed.

The arrows in the symbol denote the typical direction of current flow (not electron flow). Since the source and drain are often physically the same, the actual drain/source may be swapped. In other words, the source and drain are really only meaningful when the bias voltages are specified. For an *NMOS* device, to get current to flow from the drain-to-source, the drain should be at a higher potential than the source. The opposite is true for the *PMOS* device.

The body terminal is sometimes omitted from the schematic for clarity. In this situation, it is implicit that the body of all *NMOS* devices is tied to ground, or the lowest potential, and the body of all *PMOS* devices is connected to the supply, or highest potential.

8.3 NMOSFET Large Signal Models and Regions of Operation

We begin by looking at the MOSFET as a MOS-C, and note the importance of the gate-to-body voltage. Next, we will look at the device as a black box in order to understand the drain-source current-voltage behavior. Later we'll use this insight to derive the analytical equations.

8.3.1 Cut-off, $V_{GS} < V_T$

As shown in Fig. 8.9, we can look at the MOSFET as a MOS capacitor with two n^+ diffusion regions on each side. When $V_{GB} < V_T$, the device is either in accumulation or in depletion. Because there are no (or very few) inversion charges at the surface, there is no path for current to flow from the source to the drain. No matter how large we make V_{DS} (up to breakdown limits), current does not flow through two back-to-back diodes that are of opposite polarity.

8.3.2 Inversion, $V_{GS} > V_T$ and $V_{DS} > 0$

On the other hand, if the gate is biased above threshold, the surface is inverted, or for an *NMOS* there is an island filled with electrons that extends from the source to the drain. This inverted region forms a channel of inversion charges (in this case electrons) that connects the drain and source, as shown in Fig. 8.10. The inversion charges originate from n^+ diffusion regions, so this region can form quickly, unlike a MOS-C where the inversion layer was formed thermally. If a drain-source voltage (V_{DS}) is applied, electrons will flow from source to drain.

Note: Don't forget that in an N-channel MOSFET (P-type body) the charge carriers (electrons) flow from $S \rightarrow D$, whereas current flows from $D \rightarrow S$.

In a P-channel MOSFET the charge carriers are holes, and they still flow from $S \rightarrow D$. But now the current also flows from $S \rightarrow D$.

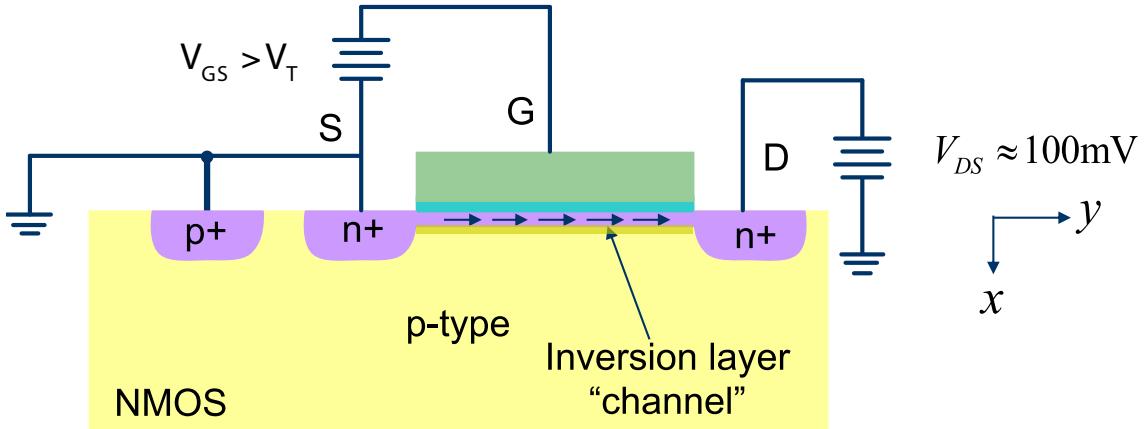


Figure 8.10: When a sufficiently positive gate voltage is applied, $V_{GS} > V_T$, a channel forms between the drain and source, essentially connecting these regions. Electrons can be injected into the channel from the source side, and flow to the drain (biased at a higher potential). Note current flow is in the opposite direction due to negative electron charge.

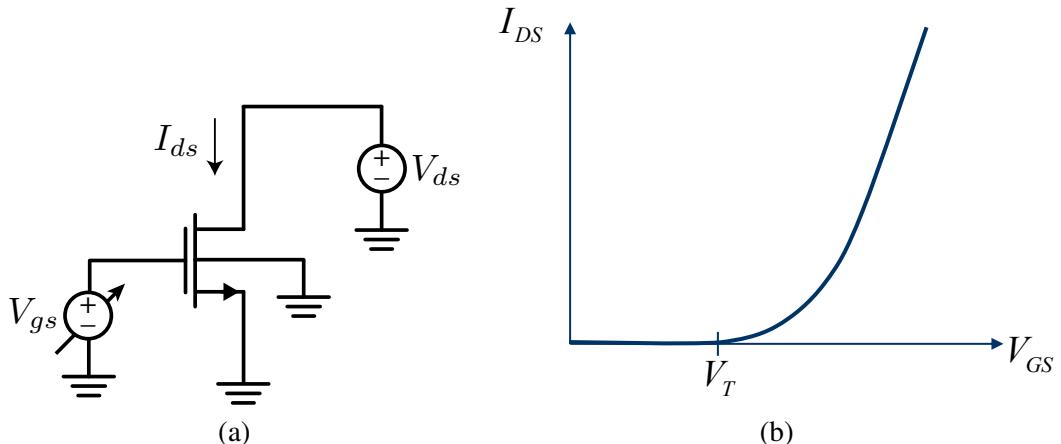


Figure 8.11: The I_{DS} versus V_{GS} characteristics of a MOSFET can be traced out by holding the drain voltage constant and sweeping the gate voltage as shown.

8.3.3 Observed Behavior: $I_{DS} - V_{GS}$

As shown in Fig. 8.11a, suppose we connect an NMOS device drain and gate to a positive voltage $V_{DS} > 0\text{V}$, and sweep the source connected to the gate. In this experiment, we monitor the drain-to-source current I_{DS} . As shown in the resulting **drain current-gate voltage curve** (Fig. 8.11b), the current is initially zero, and the transistor only conducts when the gate-to-source voltage exceeds a threshold. The body of the transistor and source are both tied to ground, so in this case the gate-to-source voltage is also the gate-to-body voltage. We know that when we cross the V_T (threshold voltage) of the device, an inversion layer forms and this must be responsible for the current flow.

We find that the current increases rapidly as we cross the threshold, and it reaches a point where it simply increases linearly with applied gate-to-source voltage. Also, if we zoom in to the region near the threshold, we will observe very small conduction, in a region called "**sub-threshold region**", where minute current flow is possible. We will explain this current flow when we study bipolar junction transistors (BJTs) in Ch. 11. For now, in this chapter we will ignore this region.

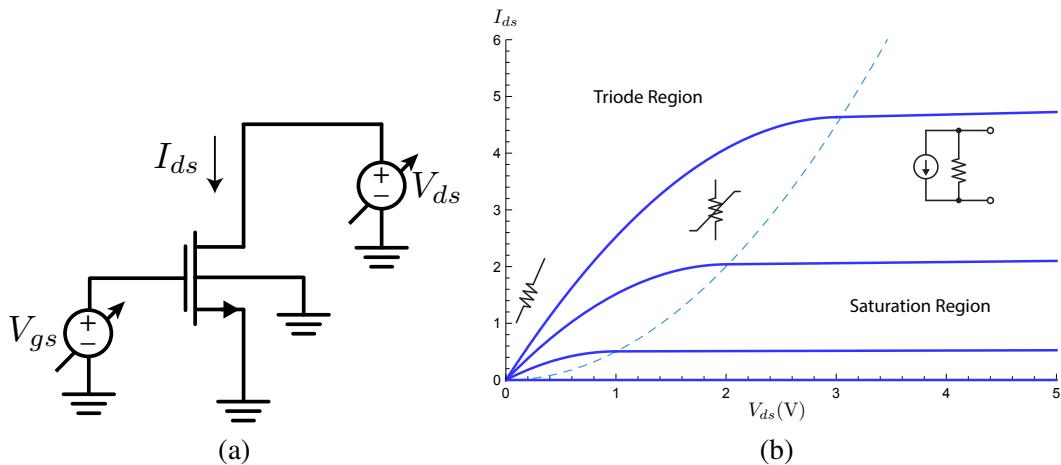


Figure 8.12: (a) The I_{DS} versus V_{DS} characteristics of a MOSFET can be traced out by holding the gate voltage constant and sweeping the drain voltage as shown. (b) A family of curves results from varying the gate voltage in discrete steps. Note that the device characteristics can be divided into two distinct regions. In the "Triode Region" the device acts like a non-linear resistor whereas in the "Saturation Region" the current is relatively constant as a function of V_{DS} , acting like a current source.

8.3.4 Observed Behavior: $I_{DS} - V_{DS}$

Now suppose that we fix the gate voltage to constant values and sweep the drain-to-source voltage, as shown in Fig. 8.12a. What we observe is a **family of curves**, shown in Fig. 8.12b. Each curve corresponds to a particular value of the gate voltage. Just as before, if the gate-to-source voltage is below the threshold, then no current flows and we observe the flat line at $0A$. For a gate-to-source voltage above threshold, we observe an $I-V$ behavior of a non-linear resistor. For low values of drain-to-source voltage, the device is like a normal resistor. As the voltage is increased, the resistance behaves non-linearly and the rate of increase of current slows as the curve bends over.

Eventually the current stops growing and remains essentially constant, acting like a current source. The dashed line separates the current into two regions, called the "**Triode Region**" and the "**Saturation Region**". In the saturation region, the current is saturated at a fixed value and increases very little for variations in the drain-source voltage. We draw schematic symbols in each region to emphasize the behavior of the device, first as a conductor, then a non-linear conductor, and finally as a current source. Now it should be clear why a transistor can act as both a source of transconductance or a resistor.

8.4 Derivation of MOSFET $I - V$ Curve

With an understanding of the channel inversion forming due to the operating point of the MOS-C device, we can now derive the current flow due to the inversion layer. Our derivations are approximate and simple to aid us in gaining insight into the device, not to build accurate models.

8.4.1 MOSFET "Linear" Region

With reference to Fig. 8.10, let's assume the device is biased in inversion to allow current to flow:

$$V_{GS} > V_T \quad (8.1)$$

When the drain-to-source voltage is not too large, we observed that the device behaves like an ordinary conductor. This makes sense if we view the channel as a semiconductor resistor. Let's

calculate the current to verify this. The current in this channel is given by:

$$I_{DS} = -Wv_y Q_N \quad (8.2)$$

In Eq. 8.2, W is the width of the device, v_y is the velocity of charge carriers in the y direction (along the channel), and Q_N is the **inversion charge density** (the charge per unit area) at the surface.

The charge (the free carriers at the surface) is proportional to the voltage applied across the oxide over threshold, or the **overdrive voltage**. "Overdrive" means a bias voltage applied over threshold:

$$\boxed{V_{OD} = V_{GS} - V_T} \quad \text{Overdrive voltage for MOSFET} \quad (8.3)$$

Then the charge is:

$$Q_N = C_{ox}(V_{GS} - V_T) = C_{ox} V_{OD} \quad (8.4)$$

Substituting Eq. 8.4 into Eq. 8.2, we have:

$$I_{DS} = -Wv_y C_{ox}(V_{GS} - V_T) \quad (8.5)$$

If we assume that the channel is of uniform density, then only drift current flows. The **drift velocity** of carriers is proportional to the electric field along the channel:

$$\boxed{v_y = -\mu_n \mathcal{E}_y} \quad \text{Channel drift velocity, MOSFET} \quad (8.6)$$

Notice that this electric field is different from the MOS-C vertical electric field, which is in the x direction, with all field lines terminating on charges. The lateral field is along the channel (y direction), and for a uniform density of charges, is proportional to the drain-to-source voltage:

$$\mathcal{E}_y = -\frac{V_{DS}}{L} \quad (8.7)$$

In Eq. 8.7 L is the channel length, because the voltage is dropped uniformly across the channel. Again, this assumption is only valid if the channel is *uniform*. Making this last substitution completes the derivation of current versus voltage for the MOSFET in the **linear regime**:

$$\boxed{I_{DS} = \frac{W}{L} \mu_n C_{ox} (V_{GS} - V_T) V_{DS}} \quad \text{MOSFET drain current, linear regime} \quad (8.8)$$

The linear region is a special section of the triode region, where the drain current is effectively linearly dependent on the drain voltage. Since current is proportional to voltage, we can now define a **channel conductance**:

$$G_{DS} = \frac{I_{DS}}{V_{DS}} = \mu_n C_{ox} (V_{GS} - V_T) \left(\frac{W}{L} \right) \quad (8.9)$$

Eq. 8.9 shows that the conductance depends on the carrier mobility μ_n , the level of inversion $C_{ox}(V_{GS} - V_T)$, and the device dimensions. For a fixed **aspect ratio** W/L , we get more current by realizing higher mobility μ_n , higher gate-oxide capacitance C_{ox} , and higher bias voltages.

8.4.2 MOSFET as a Variable Resistor

Notice that in the linear region, the current is proportional to the drain-source voltage and the gate-source voltage. For a *fixed gate-source voltage*, the equivalent resistance is given by the reciprocal of the channel conductance, $1/G_{DS}$. We define a **voltage-dependent resistor**:

$$R_{eq} = \frac{V_{DS}}{I_{DS}} = \frac{1}{\mu_n C_{ox} (V_{GS} - V_T)} \left(\frac{L}{W} \right) = R_{\square} (V_{GS}) \frac{L}{W} \quad (8.10)$$

This is a variable resistor that is electronically tunable by adjustment of the gate-to-source voltage. The parameter R_{\square} is the **sheet resistance** of the surface of the MOS transistor. The sheet resistance drops with increasing gate overdrive.

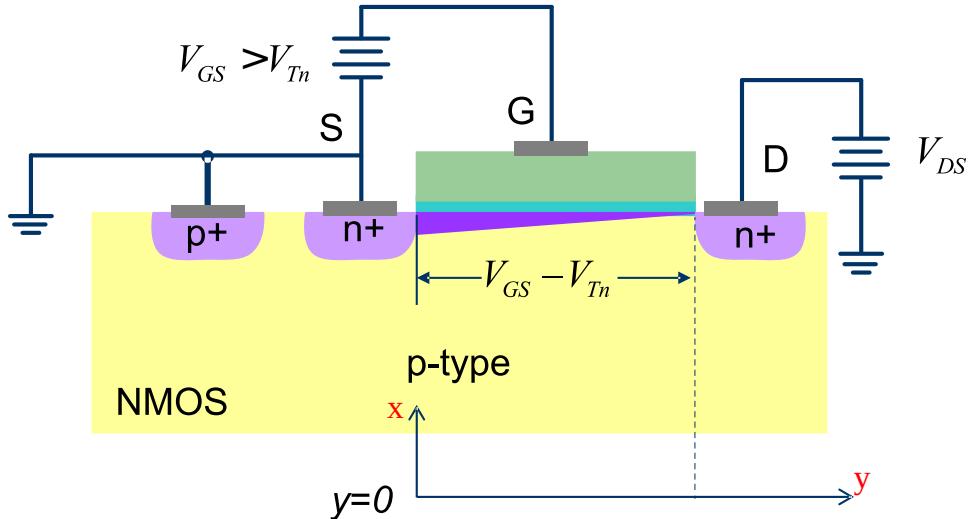


Figure 8.13: The setup to help derive the MOS current flow. Note in particular the non-uniform inversion layer from the source side (maximum inversion) to the drain side (less inversion).

8.4.3 Approximate Derivation of Inversion Charge Variation

Up to now, we have ignored the fact that the level of inversion should vary from the source to the drain. To understand why, consider that the channel covers the surface of the device, and any gate voltage charge will be compensated by inversion charge in the channel. If $V_{DS} \neq 0V$, then the channel potential is not uniform. Because the inversion charge depends on the gate-to-channel voltage, we should expect that the inversion charge to be *maximum* on the **source side** and *lower* on the **drain side**. This is because $V_{DS} > 0V$, as shown in Fig. 8.13.

Let's approximate the inversion charge $Q_N(y)$ along the channel, now a function of position along the channel y . With a channel length of L , by taking an average of the source/drain values we have:

$$Q_N(y) \approx \frac{Q_N(y=0) + Q_N(y=L)}{2} \quad (8.11)$$

At the source side we have:

$$Q_N(y=0) = -C_{ox}(V_{GS} - V_T) \quad \text{MOSFET channel charge, source side} \quad (8.12)$$

While at the drain side we have:

$$Q_N(y=L) = -C_{ox}(V_{GD} - V_T) \quad \text{MOSFET channel charge, drain side} \quad (8.13)$$

Note that the gate-to-drain voltage is lower due to the higher potential at the drain side. To see this, note that the voltage at the drain is given as the voltage at the source plus the V_{DS} :

$$V_D = V_S + V_{DS} \quad (8.14)$$

Substituting Eq. 8.14 for V_D , we can see why V_{GD} must be lower than V_{GS} :

$$V_{GD} = V_G - V_D = V_G - (V_S + V_{DS}) \quad (8.15)$$

$$= V_{GS} - V_{DS} \quad (8.16)$$

This results in less inversion charge at the drain side. With this approximation, the average value of charge is given by:

$$Q_{N_{ave}}(y) \approx -\frac{C_{ox}(V_{GS} - V_T) + C_{ox}(V_{GD} - V_T)}{2} \quad (8.17)$$

Substituting *Eq. 8.16* for V_{GD} into *Eq. 8.17*, we have:

$$Q_{N_{ave}}(y) \approx -\frac{C_{ox}(V_{GS} - V_T) + C_{ox}(V_{GS} - V_{DS} - V_T)}{2} \quad (8.18)$$

Collecting terms, we have two $(V_{GS} - V_T)$ terms, with one in terms of V_{DS} :

$$Q_{N_{ave}}(y) \approx -\frac{C_{ox}(2V_{GS} - 2V_T) - C_{ox}V_{DS}}{2} \quad (8.19)$$

Finally, we arrive at an expression for the **average channel charge**:

$$Q_{N_{ave}}(y) = -C_{ox}(V_{GS} - V_T - \frac{V_{DS}}{2})$$

MOSFET average channel charge (8.20)

As expected, the average charge is lower than the source charge, because the drain voltage is higher than the source. So the vertical electric field is lower on the drain end, resulting in less inversion charge.

8.4.4 Drift Velocity and Drain Current

Now we use mobility to find the drift velocity by approximating the y direction field as before:

$$v_{drift}(y) = -\mu_n \mathcal{E}(y) \approx -\mu_n \left(-\frac{\Delta V}{\Delta y} \right) = \frac{\mu_n V_{DS}}{L} \quad (8.21)$$

This is clearly an approximation valid only under small V_{DS} . Substituting *Eq. 8.21* into *Eq. 8.2* for drift current, we have:

$$I_{DS} = -W v_y Q_N \approx W \mu \left(\frac{V_{DS}}{L} \right) C_{ox} \left(V_{GS} - V_T - \frac{V_{DS}}{2} \right) \quad (8.22)$$

Re-writing this into a form that emphasizes the geometric aspect ratio W/L :

$$I_{DS} = \left(\frac{W}{L} \right) \mu C_{ox} \left(V_{GS} - V_T - \frac{V_{DS}}{2} \right) V_{DS} \quad (8.23)$$

This leads to a family of inverted parabolas for the current as a function of V_{DS} . These curves describe the channel conductance observed in *Fig. 8.12*, wherein the current increases linearly, but then bends as we increase V_{DS} .

We now understand the origin of the bend, since increasing V_{DS} affects both the horizontal field along the channel (which tends to increase the current), but also results in lower levels of inversion in the channel, which tends to decrease the current. These two effects essentially cancel out, and the net result is a curve that is bending and flattening out.

8.4.5 Square Law "Exact" Derivation

It is hard to make claims of "exactness", because the actual equations are very complex. We are still making many assumptions in the following derivation, but what follows will be less hand-wavy and a bit more precise. The key insight is that the threshold voltage varies along the channel, because the **channel potential** (call it $V_{ch}(y)$) is not constant from the source to the drain:

$$V_T(y) = V_T(y=0) + V_{ch}(y) \quad (8.24)$$

The threshold voltage at the drain end is related to the source end as follows:

$$V_T(y=L) = V_T(y=0) + V_D \quad (8.25)$$

The relation in *Eq. 8.25* comes from the fact that the higher drain potential requires a correspondingly higher gate voltage to reach the same level of inversion, exactly by the drain voltage. This means that the inversion charge varies from source to drain:

$$Q_n(y) = C_{ox}(V_G - V_T(y)) \quad (8.26)$$

In other words, if we define V_T as the threshold voltage at the source end ($V_T(y=0) = V_T$), then the inversion charge varies as:

$$Q_n(y) = C_{ox}(V_G - V_T - V_{ch}(y)) \quad (8.27)$$

Using this, and the fact that the \mathcal{E} -field is the spatial derivative of the channel voltage, we have:

$$I_{DS} = \mu_n W C_{ox}(V_G - V_T - V_{ch}(y)) \frac{dV_{ch}}{dy} \quad (8.28)$$

Let's multiply both sides of *Eq. 8.28* by dy , and integrate along the channel. We are converting the integral from position along the channel, to voltage along the channel. Because the current is uniform along the channel, the expression on the left-hand side must evaluate to $I_{DS} \times L$:

$$\int_0^L I_{DS} dy = I_{DS} \times L = \int_0^{V_D} \mu_n W C_{ox}(V_G - V_T - V_{ch}(y)) dV_{ch} \quad (8.29)$$

Resulting in the expression:

$$I_{DS} = \mu_n C_{ox} \frac{W}{L} \left(V_G - V_T - \frac{V_D}{2} \right) V_D$$

Drain current, triode region

(8.30)

Interestingly, this is the same equation we derived earlier. While the result is that same, we have derived it in a way more consistent with the physics without resorting to hand-waving assumptions. It should be noted that we have ignored one thing in this derivation, which is the variation of the threshold voltage due to the bulk depletion charges. In *Eq. 8.27*, we assume the inversion charge just varies linearly with the channel voltage. In practice, as the channel voltage increases from the source end, the depletion region width also changes, which means the threshold voltage is different. This is a subtle point that would over complicate the equations without providing more circuit insight or intuition, and it's common to make this simplification for hand calculations.

The triode region current (*Eq. 8.30*) is plotted in *Fig. 8.14*. We note that the currents peak when $V_{DS} = V_{GS} - V_T$, and then the equations predict lower and lower currents (inverted parabolas). However, in the laboratory we do not observe the current decreasing past the saturation point. This is shown in *Fig. 8.15*, and so there is a fundamental flaw in the derivation that we must address.

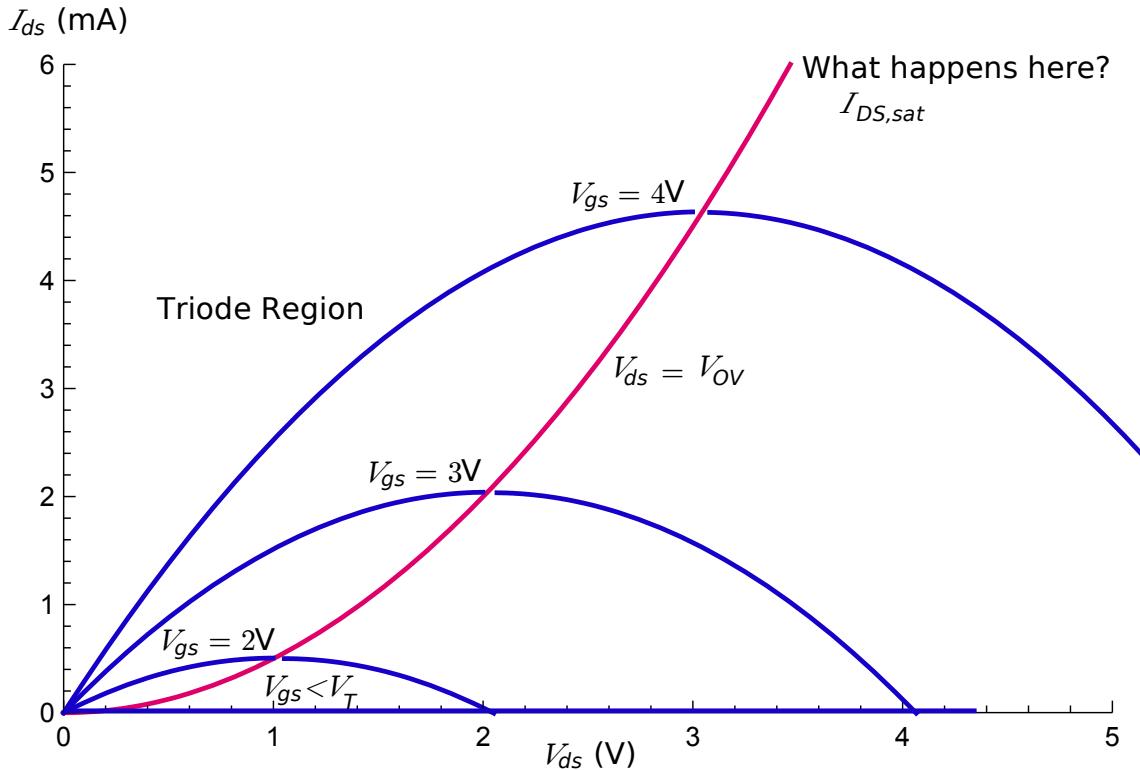


Figure 8.14: The calculated family of MOSFET $I - V$ curves in the triode region. The derivation here is only valid in the triode region and cannot predict the current saturation mechanism.

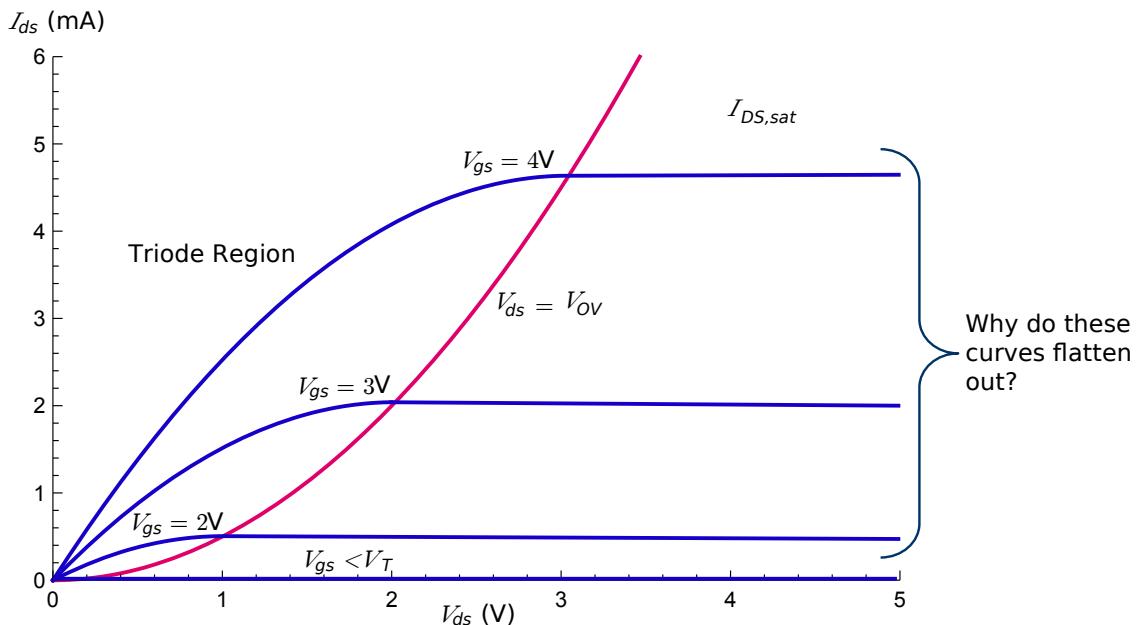


Figure 8.15: The piece-wise calculated MOSFET $I - V$ curves are derived by assuming the current in the saturation region is essentially constant and given by the maximum value predicted from the square-law relations.

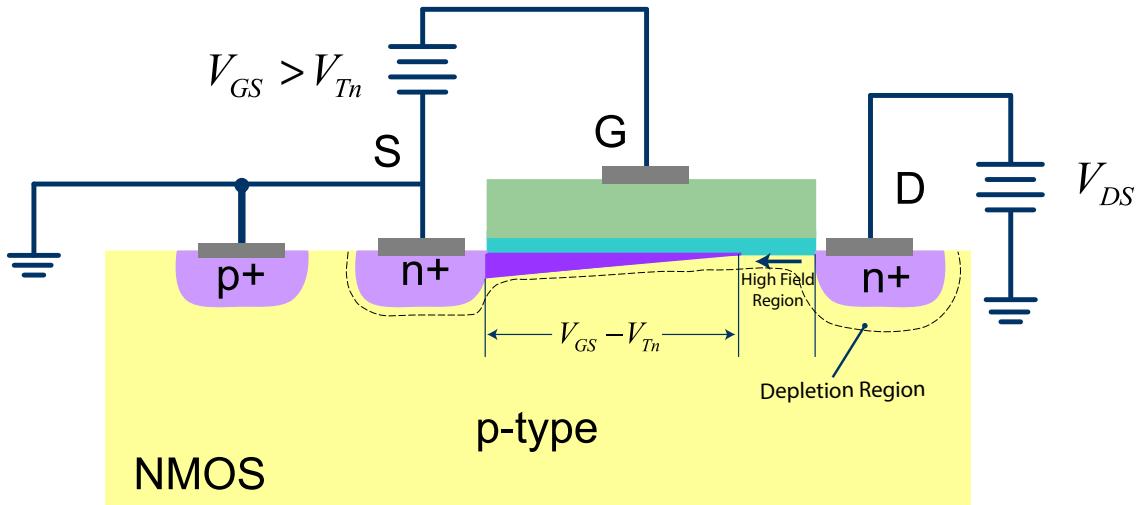


Figure 8.16: When a MOSFET is biased with $V_{DS} > V_{DS,sat}$, the channel is pinched off at a point inside the channel, but near the drain. This creates a very small depletion region between the drain diffusion region and the conductive channel. This region is a high \mathcal{E} -field region.

8.5 Understanding Current Saturation (Pinch Off)

The observed experimental behavior of devices beyond a drain voltage of $V_{DS} = V_{GS} - V_T$ is a flattening of the current, rather than decreasing current. Let's look into this a bit more to understand the origin of this saturation region, shown in *Fig. 8.16*.

8.5.1 The Saturation Region

From our equations, we have that when $V_{DS} > V_{GS} - V_T$, there is no inversion charge at the drain, so what happens? Intuitively, it seems like the channel from the source to the drain is interrupted, or pinched off, so you might think this would lead to zero current again. This intuition is so strong that this region is sometimes called the "pinch-off" region, but the current does not abruptly go to zero as intuition might lead you to believe.

8.5.2 Pinch Off

When $V_{DS} > V_{GS} - V_T$, we can say that the excess voltage $V_{DS} - (V_{GS} - V_T)$ is dropped across a small region between the drain and the channel, shown in *Fig. 8.16*. Let's define the **saturation drain voltage**:

$$V_{DS,sat} = V_{GS} - V_T \quad \text{MOSFET saturation voltage} \quad (8.31)$$

Any voltage above $V_{DS,sat}$ results in a fixed electric field $\mathcal{E}_{d,sat}$ in the channel, and an excess electric field \mathcal{E}_d near the drain end. Since this is a small region, even a modest drain voltage above $V_{DS,sat}$ results in a potentially very large electric field \mathcal{E}_d . This large electric field can propel electrons from the channel into the drain end at high velocity. In this region, increasing the drain voltage does not increase current (appreciably), because the current flow is limited by the supply of electrons from channel side. The supply of electrons is dictated by $\mathcal{E}_{d,sat}$, set by $V_{DS,sat}$.

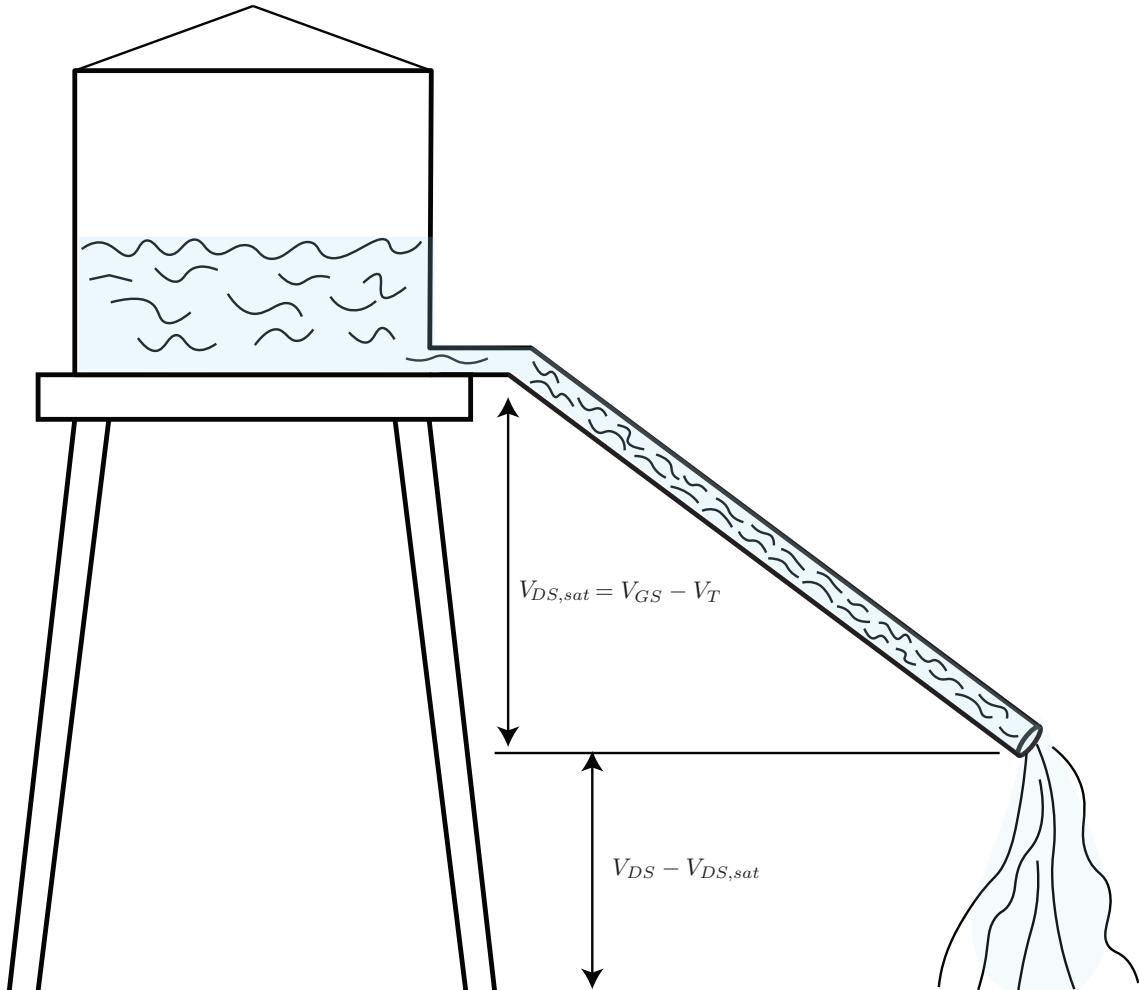


Figure 8.17: A good analog for a MOSFET biased with $V_{DS} > V_{DS,sat}$ is a water pipe flowing down a hill into a waterfall. The region where the water flows on the hill has a potential drop of $V_{DS,sat}$ and the remaining potential $V_{DS} - V_{DS,sat}$, is dropped through the waterfall. The overall current is independent of the waterfall height.

A good analogy is to imagine a stream of water gently flowing down an incline and then into a waterfall, as shown in *Fig. 8.17*. If the slope of the incline is increased, the current flow increases. But the current flow is independent of the height of the waterfall. In our system, the waterfall region is the high \mathcal{E} -field region.

In addition, in real transistors current saturation may happen due to another mechanism, namely velocity saturation. If the field strength in the channel exceeds a threshold, the linear relation between drift velocity and electric field ceases to be valid, and the velocity of carriers begins to saturate. This is shown in *Fig. 8.15*.

8.5.3 Square-Law Current in Saturation

Taking the previous discussion into account, we can say that the current peaks at maximum value when $V_{DS} = V_{GS} - V_T = V_{DS,sat}$. Or in other words:

$$I_{DS,sat} = \frac{W}{L} \mu C_{ox} \left(V_{GS} - V_T - \frac{V_{DS,sat}}{2} \right) V_{DS,sat} \quad (8.32)$$

$$= \frac{W}{L} \mu C_{ox} \left(V_{GS} - V_T - \frac{V_{GS} - V_T}{2} \right) (V_{GS} - V_T) \quad (8.33)$$

Eq. 8.33 can be simplified by expanding the two terms in parentheses on the RHS:

$$\begin{aligned} \left(V_{GS} - V_T - \frac{V_{GS} - V_T}{2} \right) (V_{GS} - V_T) &= V_{GS}^2 - V_{GS} V_T - \frac{1}{2} V_{GS}^2 + \frac{1}{2} V_{GS} V_T - V_{GS} V_T \\ &\quad + V_T^2 + \frac{1}{2} V_{GS} V_T - \frac{1}{2} V_T^2 \\ &= \frac{1}{2} V_{GS}^2 + \frac{1}{2} V_T^2 - V_{GS} V_T \\ &= \frac{1}{2} (V_{GS}^2 - 2V_{GS} V_T + V_T^2) \\ &= \frac{1}{2} (V_{GS} - V_T)^2 \end{aligned}$$

Finally, we arrive at the following form, which is the **saturation current**:

$I_{DS,sat} = \left(\frac{W}{2L} \right) \mu C_{ox} (V_{GS} - V_T)^2$

MOSFET saturation current (8.34)

Notice that in the saturation region, the current is nearly independent of variations in the drain voltage and only depends on the gate voltage. We can think of the transistor as an ideal **voltage-controlled current source** (VCCS), with an output current dictated by the control terminal, the gate. This is the characteristics of an ideal amplifier.

8.5.4 Actual Saturation Current

Now we need to deal with another non-ideality. The measured values of I_{DS} increases slightly with increasing V_{DS} . In other words, the curves are not 100% flat. The physics is complicated, but a simple way to see this is that the channel is getting shorter as the drain voltage depletes away more electrons from the drain end, resulting in a shorter channel length. This phenomenon is known as the **channel length modulation**. We can model this with an additional linear term, λ , that allows the current to increase slightly for larger and larger V_{DS} values:

$I_{DS,sat} = \left(\frac{W}{2L} \right) \mu C_{ox} (V_{GS} - V_T)^2 (1 + \lambda V_{DS})$

Saturation current, with CLM (8.35)

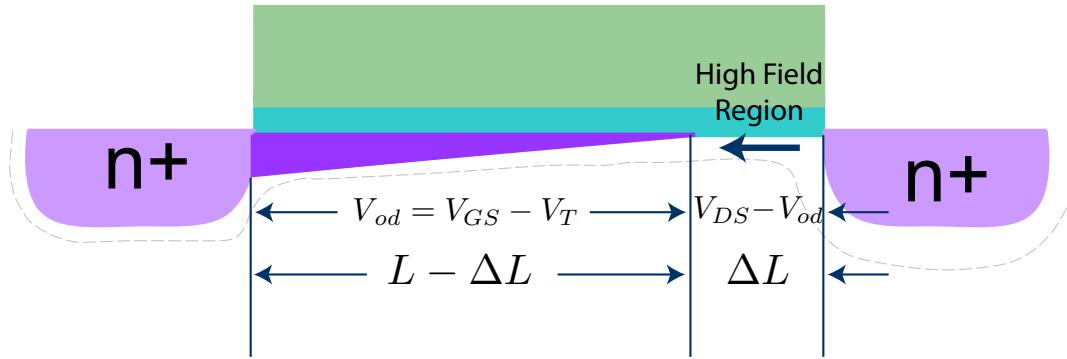


Figure 8.18: Channel Length Modulation (CLM) occurs when the drain voltage is modulated beyond $V_{DS,sat}$. The extra voltage dropped across the high \mathcal{E} -field region causes the effective channel length to shrink by ΔL , which results in an increase in the drain-source current.

8.5.5 Channel Length Modulation

To see why the current increases as a result of the channel length modulation, let's go back to when $V_{DS} = V_{GS} - V_T = V_{DS,sat}$. This is the point in which the channel pinches off near the drain. With a further increase in V_{DS} , we see that the pinch-off point moves toward the source, effectively reducing the channel length from L to $L - \Delta L$, as shown in Fig. 8.18. Since current is inversely proportional to L , we have:

$$I_{DS} = \left(\frac{1}{2}\right) \mu_n C_{ox} \left(\frac{W}{L - \Delta L}\right) (V_{GS} - V_T)^2 \quad (8.36)$$

If this is a **long-channel MOSFET**, then $\Delta L \ll L$, and we can simplify this to:

$$I_{DS} = \left(\frac{1}{2}\right) \mu_n C_{ox} \left(\frac{W}{L}\right) (V_{GS} - V_T)^2 \left(1 - \frac{\Delta L(V_{DS})}{L}\right) \quad (8.37)$$

The *channel length modulation* (CLM) is the change in L as a function of V_{DS} . To first order:

$$\frac{\Delta L(V_{DS})}{L} = \lambda V_{DS} \quad (8.38)$$

Thus, we have arrived at the same result as Eq. 8.35:

$$I_{DS} = \left(\frac{1}{2}\right) \mu_n C_{ox} \left(\frac{W}{L}\right) (V_{GS} - V_T)^2 (1 - \lambda V_{DS}) \quad (8.39)$$

Note that the channel length modulation parameter λ has units of inverse volts, V^{-1} . Sometimes this parameter is called the "**Early Voltage**":

$V_A = \frac{1}{\lambda}$

Early voltage (8.40)

The name "Early" is related to the Early Effect, named after its discoverer James M. Early, first applied to bipolar junction transistors.

8.5.6 Summary: Regions of Operation

Now we will summarize the various conditions that dictate which region of operation a MOSFET is operating in, and the currents associated with them.

Cut-off region: $V_{GS} < V_T$

In cut-off mode the transistor is off, and there is effectively no current that flows¹:

$$I_{DS,off} = 0A \quad (8.41)$$

Linear region: $V_{GS} > V_T, V_{DS} \ll V_{GS} - V_T$

In this region the device acts like a variable conductor, resulting in:

$$I_{DS,lin} = \left(\frac{W}{L} \right) \mu_n C_{ox} (V_{GS} - V_T) V_{DS} \quad (8.42)$$

Triode region: $V_{GS} > V_T, V_{DS} < V_{GS} - V_T = V_{DS,sat}$

In the triode region the channel is inverted, and a drain voltage is applied to allow current to flow. However, the applied V_D does not exceed the gate-overdrive voltage, which governs the current by:

$$I_{DS,tri} = \left(\frac{W}{L} \right) \mu_n C_{ox} \left(V_{GS} - V_T - \frac{V_{DS}}{2} \right) V_{DS} \quad (8.43)$$

Saturation region: $V_{GS} > V_T, V_{DS} > V_{GS} - V_T = V_{DS,sat}$

In saturation the channel is inverted at the source side, but not at the drain side. The current is nearly constant, and only varies minutely due to the CLM effect described earlier:

$$I_{DS,sat} = \left(\frac{W}{2L} \right) \mu_n C_{ox} (V_{GS} - V_T)^2 (1 + \lambda V_{DS}) \quad (8.44)$$

8.6 The Complementary PMOS Device

8.6.1 PMOS Device

So far, we derived the equations for an *NMOS* device. The *PMOS* work exactly the same way, but they have an *N*-type body, and a channel that is made of positive charges (holes). Noted earlier in *Fig. 8.8*, the schematic symbol is also different, with an arrow indicating the direction of current flow from the source to the drain. Notice that holes do flow from the source to drain, so unlike an *NMOS* device, here the arrow and the flow of carriers is the same. The complete square law relation is shown in *Fig. 8.19*. Compared to the *NMOS* device, everything is upside down. This is only if we insist on keeping the same polarity for voltages and currents. Let's define the current as I_{SD} , not I_{DS} . Also V_{SG} and V_{SD} are positive for **forward active regions**, so in terms of these variables, the curves look right-side-up as shown in *Fig. 8.20*.

To derive the *I-V* relation of a *PMOS* device without repeating all the steps above, we can simply invert all voltages and currents and arrive at the same equations. Because we invert the terms, we take an absolute value of the threshold voltage to make the equations identical with the *NMOS* case. For example, the saturation current is given by:

$$I_{SD,sat} = \left(\frac{W}{2L} \right) \mu C_{ox} (V_{SG} - |V_{T_p}|)^2 (1 + \lambda V_{SD}) \quad \text{PMOS saturation current} \quad (8.45)$$

¹This is an approximation we will make in this class. In later courses you will learn about sub-threshold conduction and leakage in the device.

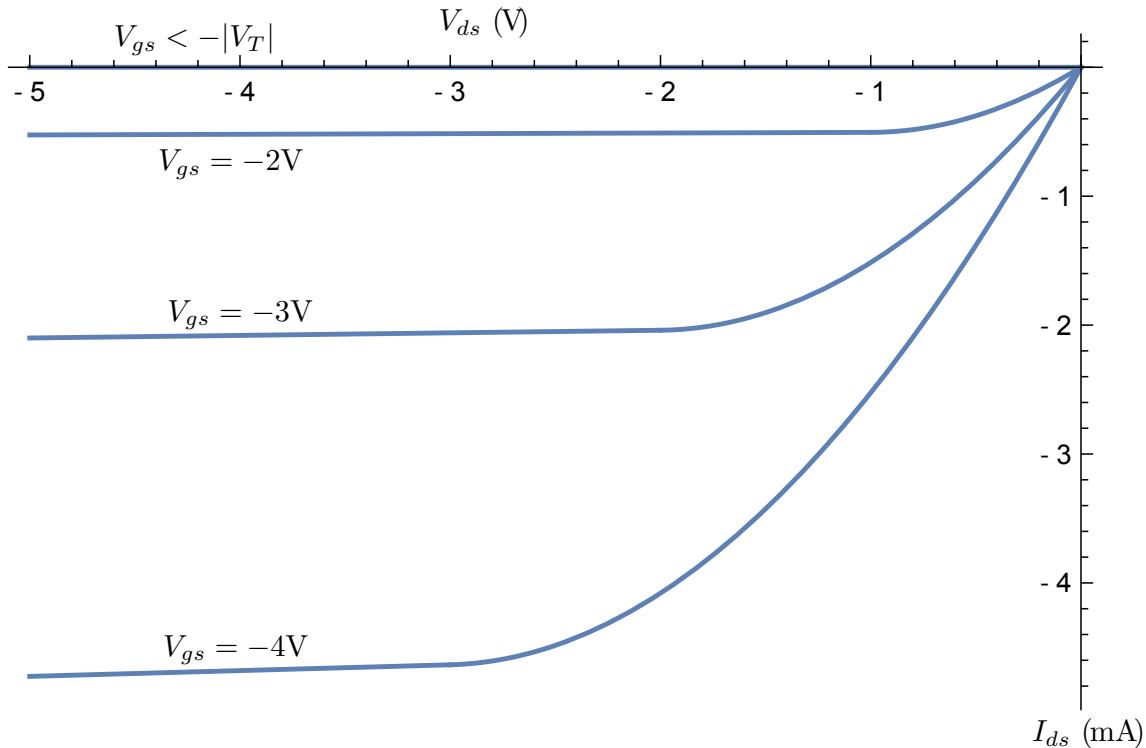


Figure 8.19: PMOS I - V curves are inverted compared to the NMOS curves, because the current flows in the opposite direction (holes versus electrons), and the device is biased with a negative gate-to-source voltage in order to form the hole inversion channel.

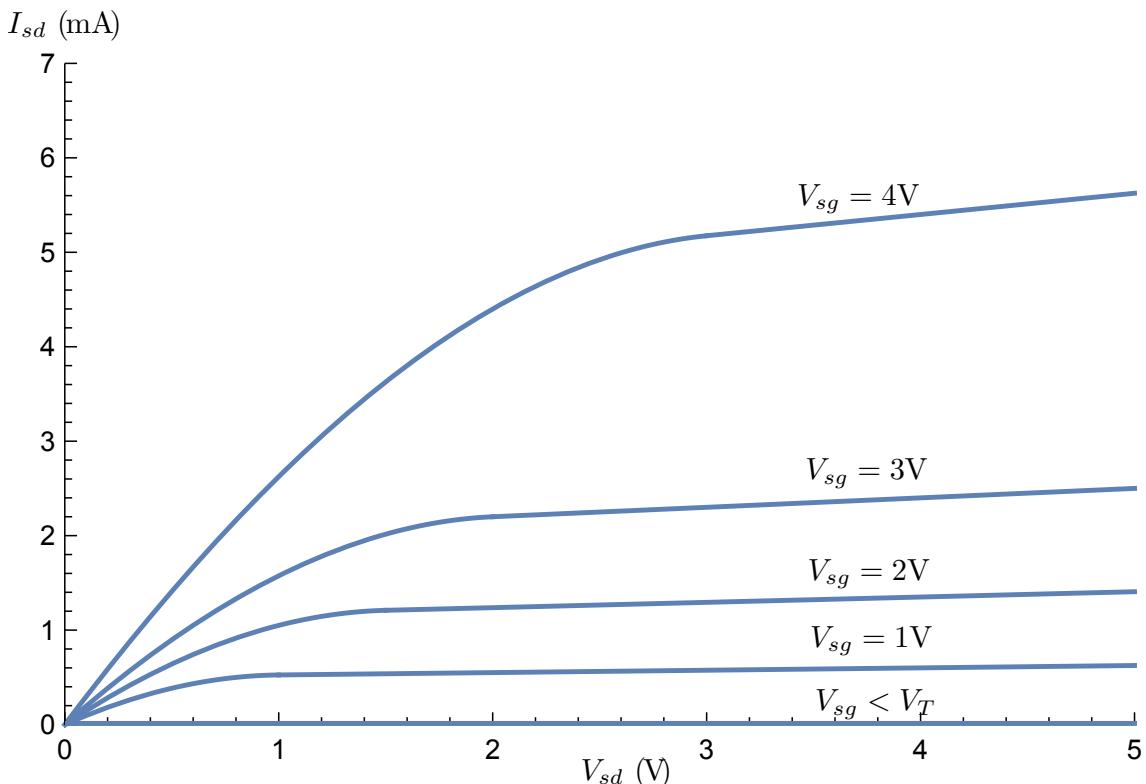
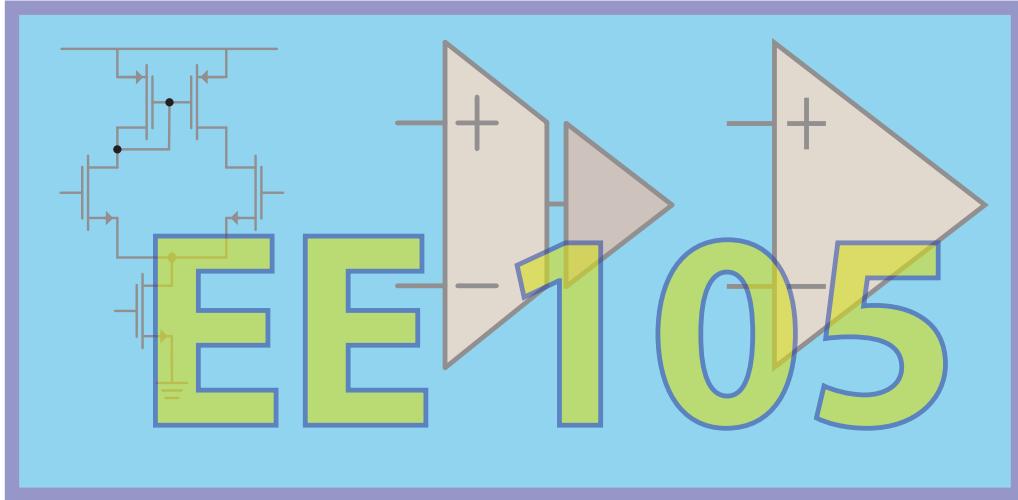


Figure 8.20: PMOS I - V curves "right-side-up" by flipping the orientation of the device voltage and currents.



9. MOS Transistor Small-Signal Models

9.1 Chapter Preview

When the inventors of the bipolar transistor at Bell Labs first built a working device (see *Fig. 9.1*), the next thing they did was to build an audio amplifier to prove that the transistor was actually working. In this chapter we are going to introduce you to your first amplifier, and discuss how we go about analyzing the amplifier using the *I-V* equations we derived in the previous chapter. The amplifier will be a "*common source*" amplifier. We will discuss bias points, also known as the operating point, and what happens when a signal is applied.

Throughout the chapter we will make a very common and important assumption, which leads to the *small signal analysis* technique. This technique simply states that if the signal amplitudes are small (we will define what "small" means), we can analyze the circuit using linear techniques. This leads us to a linearized model of the circuit, which is much easier to analyze. Finally, we will introduce the equivalent linear circuit model for a MOS transistor, and derive its small-signal parameters, such as transconductance and output resistance. In the next chapter, we will cover charge storage effects and modify the model to include capacitors, which are needed to properly account for the frequency response of amplifiers.

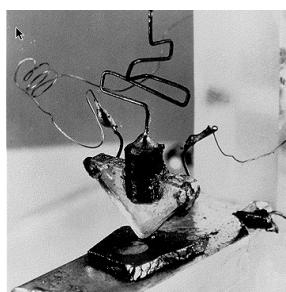


Figure 9.1: The first transistor amplifier was demonstrated on December 23, 1947, by three scientist at Bell Laboratories. The three scientists were Dr. Shockley, Dr. Bardeen, and Dr. Brattain, honored with the Nobel Prize in 1956.

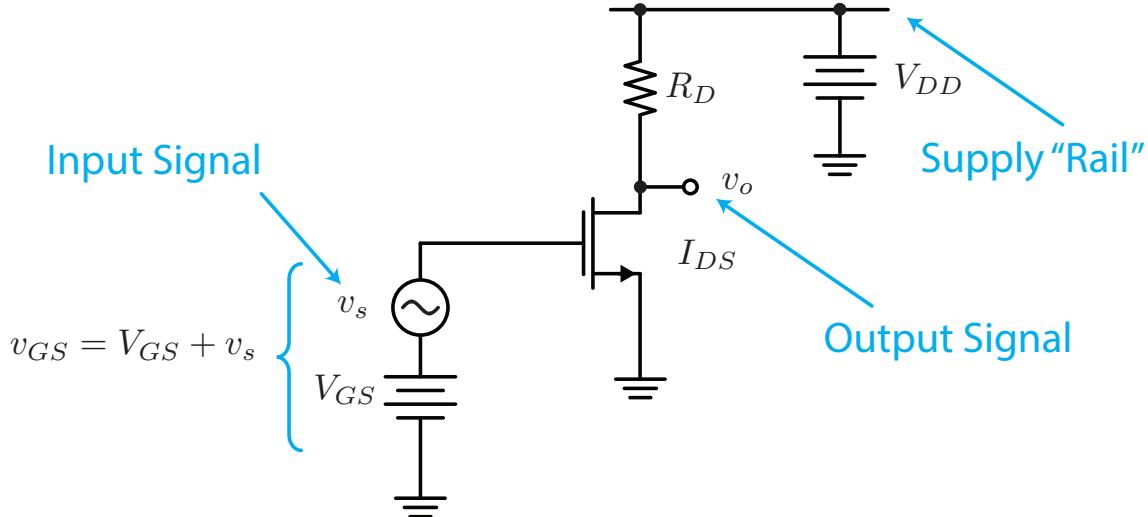


Figure 9.2: A common source transistor amplifier biased with DC voltages V_{GS} and V_{DD} and driven by an AC signal v_s at the input. The output signal is taken at the drain side.

9.2 Introduction to Amplifiers

9.2.1 A Simple Circuit: A MOS Amplifier

A **MOS amplifier** is shown in *Fig. 9.2*. Notice that an NMOS sits at the core. The input signal is applied to the gate, and the output is taken at the drain of the transistor. The source is ground (and the body, although not shown explicitly). MOS devices are biased in the saturation region, so that the output current is only a function of the gate-source voltage (approximately). This means that the transistor is acting like a voltage-controlled current source (VCCS) (see *Fig. 9.3*). In order to turn the MOS "on" and bias it in saturation, the drain is pulled up through a resistor R_D , which is connected to the supply. The "on" current $I_{DS,sat}$ and the resistor R_D are selected together to determine the operating point. On the gate side, the appropriate V_{GS} needs to be supplied to obtain the desired current $I_{DS,sat}$. The input signal v_s is applied in series with this gate bias voltage V_{GS} .

A plot of input and output signals is shown in *Fig. 9.4*. Note that the DC value is subtracted out. In many cases the DC value is not of interest. A good example is an audio amplifier. The speaker only responds to AC signals, and sound is an AC signal. The important point about this plot is the voltage excursions at the output are much larger than the input, providing voltage gain. As shown in *Fig. 9.7*, the key insight is that a gate voltage modulates the source-drain current, and this current can produce a much larger drain-source voltage. This is because the drain-source current only depends on the value of the gate voltage, and not so much on the drain-source voltage, in the saturation region. The goal of this chapter is to derive this plot in a step-by-step fashion.

9.2.2 Common Source Amplifier

The amplifier that we have presented is known as a **common source** (CS) amplifier, because the transistor source terminal is in common with both the input signal and the output signal. Or more simply, the source is grounded to the "common" node, also known as the reference voltage. Another way to understand this is that the "common" node does not carry the signal (both input and output). Since a transistor has three terminals (actually 4, but the body is usually tied to the source), and a voltage is defined by 2 terminals, then you can readily count and see that there are three simple configurations for a single transistor amplifier, "common source", "common drain", and "common gate". In this chapter we'll focus on the common source variety.

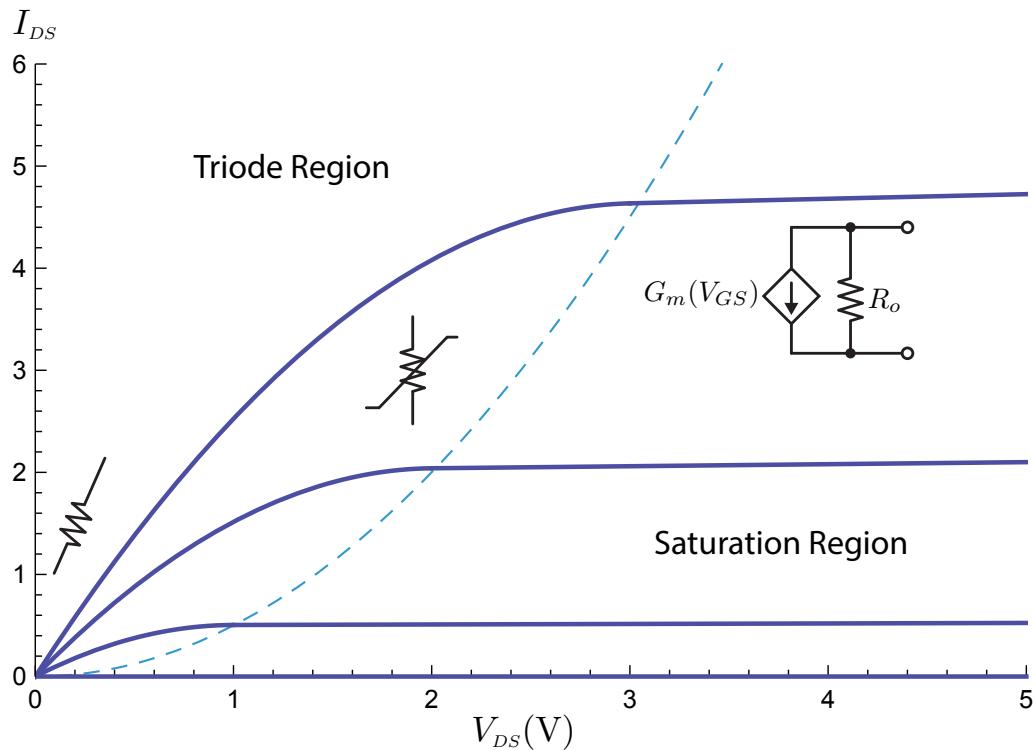


Figure 9.3: The MOS I_{DS} vs V_{DS} family of curves in saturation are nearly flat, indicating that the gate is largely responsible for the current. The device acts like a controlled current source (amplifier).

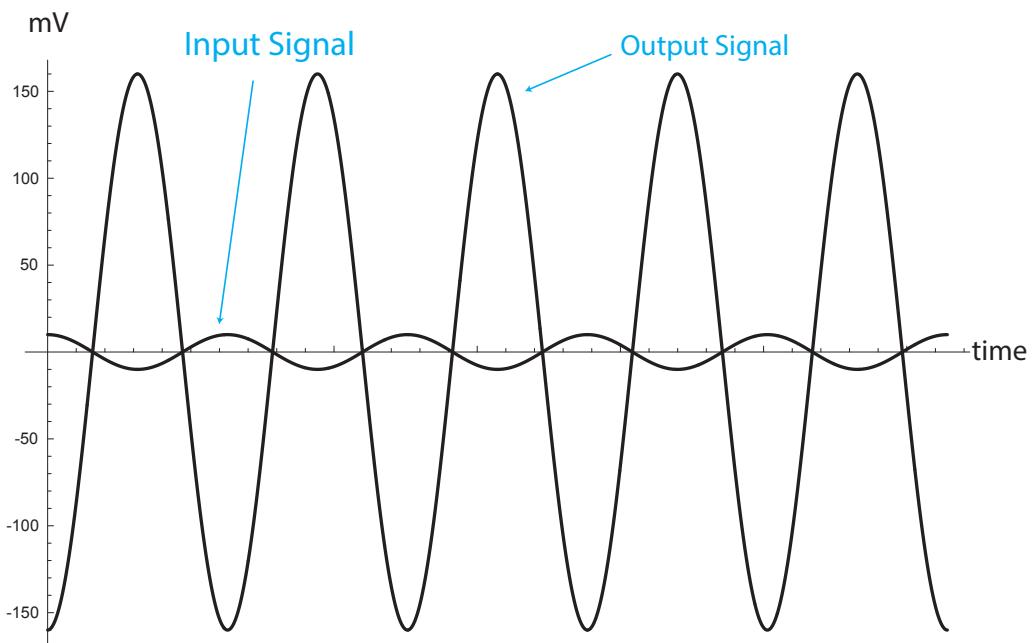


Figure 9.4: The input signal v_s and output signal v_o plotted versus time. The output voltage is an amplified (and inverted) copy of the input. In common source amplifiers the gain is always inverted.

9.3 Operating Points

The bias voltages (DC voltages shown in red in *Fig. 9.5a*) that you apply to the MOSFET determine its **operating point**, also known as the **quiescent point**, or **Q-point** for simplicity. The biasing of the Q-point is illustrated in *Fig. 9.6*. How you bias the MOSFET will make a big difference in the way that it functions. If we operate with a sufficiently high V_{GS} and a sufficiently high V_{DS} to place the device in the saturation region, we can make a very good amplifier.

9.3.1 Small Signal vs. Large Signal

We derived that I_{DS} vs. V_{GS} is a quadratic function for $V_{GS} > V_T$. Therefore large changes in V_{GS} result in quadratically larger changes at the output. However, small changes in V_{GS} (denoted as v_{gs}) will produce linear changes at the output. We can show this using a Taylor series¹ expansion:

$$I_{DS}(V_{GS} + v_{gs}) \approx I_{DS}(V_{GS}) + \frac{dI_{DS}}{dV_{GS}} \cdot v_{gs} \quad (9.1)$$

In many applications of interest, we can make this approximation and assume the signals are small with respect to the bias points. This allows us to convert MOS transistor equations into linear equations, which are much simpler to analyze.

9.3.2 Selecting the Output Bias Point

As shown in *Fig. 9.5b*, we must select the CS amplifier Q points to produce the desired response. A common choice is to bias V_{GS} so that the output is **mid-rail** (between V_{DD} and ground). This is a design choice, not a requirement, but doing so results in a large swing at the output. Notice that the output cannot go larger than V_{DD} , and to provide gain (remain in saturation) it should not go lower than $V_{GS} - V_T$, so any drain-source voltage greater than $V_{DS,sat} = V_{GS} - V_T$ and smaller than V_{DD} is possible. Mid-rail is a good back-of-the-envelope value to maximize the **AC sinusoidal swing** (this assumes $V_{DS,sat}$ is very small). This places a constraint on the DC drain current through the resistor:

$$I_R = \frac{V_{DD} - V_o}{R_D} = \frac{V_{DD} - V_{DS}}{R_D} \quad (9.2)$$

The resistor current flows into transistor:

$$I_R = I_{DS,sat} \quad (9.3)$$

We must ensure that this gives a self-consistent solution (transistor is biased in saturation):

$$V_{DS} > V_{DS,sat} = V_{GS} - V_T \quad (9.4)$$

¹See Appendix A for a review of the Taylor series

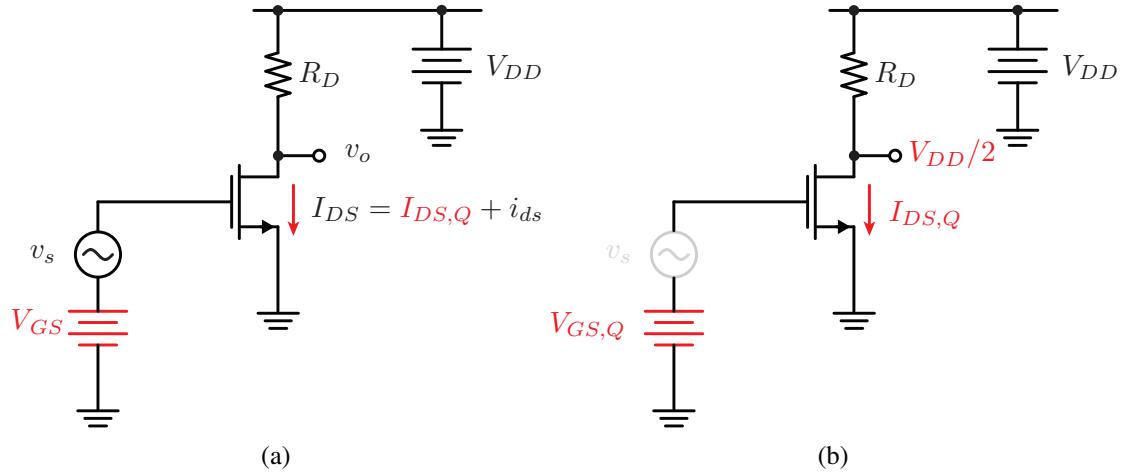


Figure 9.5: MOS amplifier (a) DC setup. The voltage source V_{GS} determines the quiescent current $I_{DS,Q}$, which in turn determines the output (b) DC operating voltage. Notice the DC operating point is independent of the AC input signals.

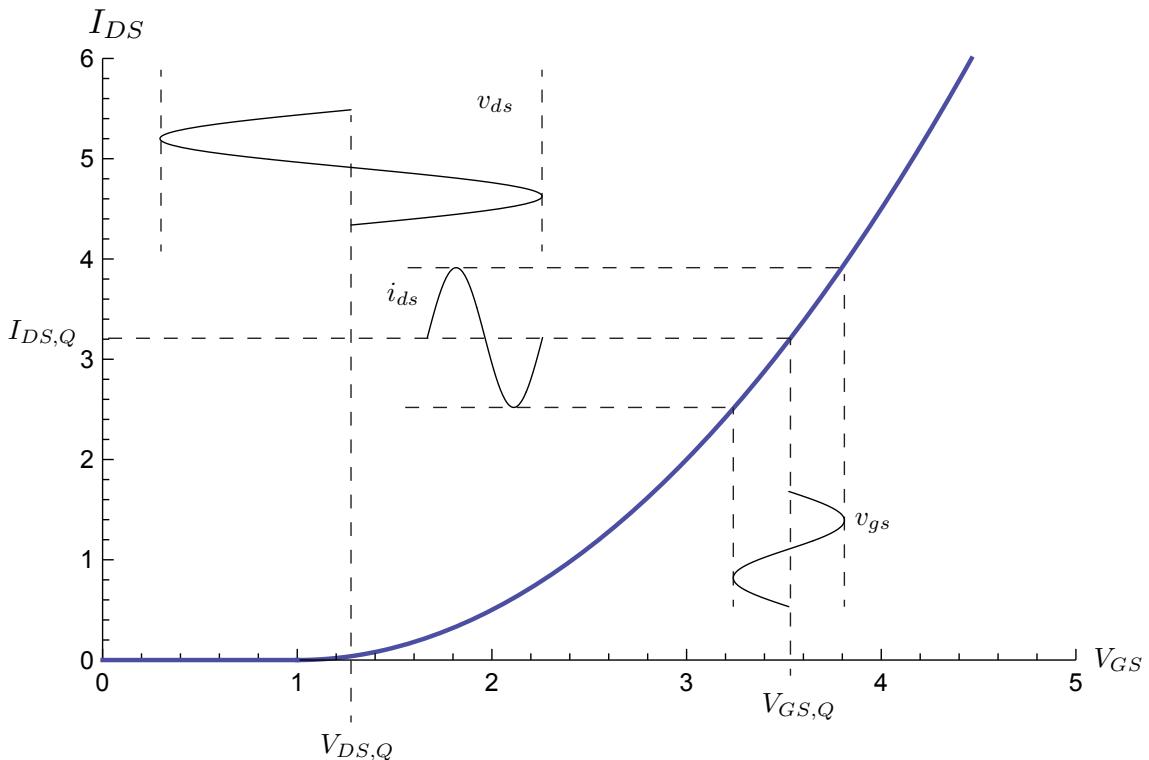


Figure 9.6: To illustrate the transconductance of a MOSFET, we bias the device at a particular operation or *Q point*, and excite it with a small-signal AC around the operating point. The resulting output current can produce a larger drain voltage by using a large load resistor.

9.3.3 Finding the Input Bias Voltage

For now let's ignore the **output resistance** (the term involving λ) to simplify the analysis:

$$I_{DS,sat} = \frac{W}{L} \mu_n C_{ox} \frac{1}{2} (V_{GS} - V_{T_n})^2 \quad (9.5)$$

This means that the current through the resistor is given by:

$$I_{R_D} = \frac{V_{DD}}{2R_D} = I_{DS,sat} = \frac{W}{L} \mu_n C_{ox} \frac{1}{2} (V_{GS} - V_{T_n})^2 \quad (9.6)$$

Note that we assume the voltage drop across the resistor is $V_{DD}/2$ (mid-rail), which was our design choice. To make this a bit less abstract, let's use some typical numbers for a long channel transistor: $W = 40 \mu m$, $L = 2 \mu m$, $R_D = 25 k\Omega$, $\mu_n C_{ox} = 100 \mu A/V^2$, $V_T = 1 V$, and $V_{DD} = 5 V$. Substituting numerical values into Eq. 9.6:

$$\begin{aligned} \frac{V_{DD}}{2R_D} &= \frac{5V}{50k\Omega} = I_{DS,sat} = 100 \mu A \\ &= \frac{W}{L} \mu_n C_{ox} \frac{1}{2} (V_{GS} - V_{T_n})^2 \\ &= \left(\frac{40 \mu m}{2 \mu m} \right)^2 \cdot \left(\frac{100 \mu A}{V^2} \right) \left(\frac{1}{2} \right) (V_{GS} - 1V)^2 \end{aligned}$$

Solving for V_{GS} :

$$\begin{aligned} 100 \mu A &= 20 \cdot 10 \left(\frac{100 \mu A}{V^2} \right) \left(\frac{1}{2} \right) (V_{GS} - 1V)^2 \\ \Rightarrow (V_{GS} - 1V)^2 &= \left(\frac{1}{10} \right) V^2 \end{aligned}$$

Rearranging, we find a value for V_{GS} :

$$V_{GS} = \sqrt{\frac{1}{10}} V + 1V = \boxed{1.32V} \quad (9.7)$$

We need to check the result from Eq. 9.7 to make sure that our assumption of operating point was correct, namely that the device is in saturation. The $V_{DS,sat}$ is given by $V_{GS} - V_T = 0.32V$, and by design $V_{DS} = 2.5V$. So $V_{DS,sat} < V_{DS} = 2.5V$, which confirms our assumption.

9.3.4 Applying the AC Voltage: The "Hard Way"

To find the output signal, we'll take two approaches. The first is the "hard way" because it involves many steps and a lot of approximations. Later we will introduce the "easy way" which leads directly to a linear circuit that we can analyze with known techniques.

In our first approach to solving for the output voltage, we will just use v_{gs} in the equations for the total drain current I_{DS} , and find v_o :

$$V_{GS} = V_{GS,Q} + v_{gs} \quad (9.8)$$

$$v_{gs} = V_s \cos \omega t \quad (9.9)$$

$$V_o = V_{DD} - R_D I_{DS} = V_{DD} - R_D \left[\left(\frac{W}{2L} \right) \mu_n C_{ox} (V_{GS,Q} + v_{gs} - V_T)^2 \right] \quad (9.10)$$

We are neglecting charge storage effects, and we are also ignoring the device output resistance to keep things simple. Now let's solve for the output voltage v_o :

$$V_o = V_{DD} - R_D I_{DS} = V_{DD} - R_D \left[\left(\frac{W}{2L} \right) \mu_n C_{ox} (V_{GS} + v_{gs} - V_T)^2 \right] \quad (9.11)$$

Focusing on the drain current, we note that it can be factored into two terms, the bias point $I_{DS,Q}$ and the signal:

$$I_{DS} = \underbrace{\left(\frac{W}{2L} \right) \mu_n C_{ox} (V_{GS} - V_T)^2}_{I_{DS,Q}} \cdot \left(1 + \frac{v_{gs}}{V_{GS} - V_T} \right)^2 \quad (9.12)$$

The bias current $I_{DS,Q}$ was selected to produce a mid-rail output voltage:

$$V_o = V_{DD} - \underbrace{(R_D I_{DS,Q})}_{\frac{V_{DD}}{2}} \left(1 + \frac{v_{gs}}{V_{GS} - V_T} \right)^2$$

Large-signal output voltage (9.13)

The above equation is the "**Large Signal**" output voltage as a function of the input voltage.

9.3.5 Small-Signal Simplifications

We linearize the output voltage for the small-signal case by expanding $(1+x)^2 = 1 + 2x + x^2$. The last term can be dropped when $x \ll 1$ (the small-signal case):

$$\left(1 + \frac{v_{gs}}{V_{GS} - V_T} \right)^2 = 1 + 2 \left(\frac{v_{gs}}{V_{GS} - V_T} \right) + \cancel{\left(\frac{v_{gs}}{V_{GS} - V_T} \right)^2}^0 \quad (9.14)$$

This leads to the following equation for the output voltage:

$$V_o \approx V_{DD} - (R_D I_{DS,Q}) \left(1 + \frac{2 v_{gs}}{V_{GS} - V_T} \right) \approx V_{DD} - I_{DS,Q} R_D - \frac{2 I_{DS,Q} R_D v_{gs}}{V_{GS} - V_T} \quad (9.15)$$

This can be further simplified by identifying the DC terms (bias) and the signal terms (AC):

$$V_o \approx V_{DD} - (I_{DS,Q} R_D) - \frac{2 (I_{DS,Q} R_D) v_{gs}}{V_{GS} - V_T} \quad (9.16)$$

$$= V_{DD} - \left(\frac{V_{DD}}{2} \right) - \frac{2 \left(\frac{V_{DD}}{2} \right) v_{gs}}{V_{GS} - V_T} \quad (9.17)$$

$$= \frac{V_{DD}}{2} - \frac{V_{DD} v_{gs}}{V_{GS} - V_T} \quad (9.18)$$

As written, the first term in Eq. 9.18 is just the DC bias point, which we designed for maximum swing: $V_{DD} - I_{DS,Q} R_D = \frac{V_{DD}}{2}$. The second term is the **small-signal output voltage**:

$$v_o \approx - \frac{2 I_{DS,Q} R_D v_{gs}}{V_{GS} - V_T} = - \underbrace{\left(\frac{V_{DD}}{V_{GS} - V_T} \right)}_{\text{voltage gain, } A_v} v_{gs}$$

Small-signal output voltage (9.19)

Since $V_{DD} > V_{GS} - V_T$, the output voltage is a larger copy of the input, and we have realized voltage gain! This analysis implies that we should use a high supply V_{DD} , and a low overdrive voltage $V_{DS,sat} = V_{GS} - V_T = V_{OD}$. For example, using numbers from our earlier example, let's take $V_{DD}/(V_{GS} - V_T) = 5V/0.32V$, which gives us a gain of 16. The plot of Fig. 9.4 was generated in this case, and shows a gain of 16. Notice that the gain is "inverting", which is always the case for CS amplifiers.

9.4 Amplifier Design and Analysis Using the Small-Signal Approach

Even though we arrived at a final answer, we did make two key simplifying assumptions. First we neglected the device output resistance. Second, we ignored all charge storage effects. Together, these effects lead to a set of coupled non-linear differential equations, which are painfully difficult to solve in the general case.

In the second approach, called the "*small-signal analysis*", we solve the problem in two steps. First we solve for DC voltages and currents. In this step we ignore AC signal sources. This step is essentially to find the bias point of the MOSFET. Notice that this step is identical to the steps we took earlier. Next we substitute a small-signal model for the MOSFET, and the small-signal models of the other circuit elements. This allows us to use standard AC circuit analysis techniques to arrive at the desired solution, say the voltage gain. This constitutes small-signal analysis. In the next chapter we will add capacitors to allow us to predict the frequency response of amplifiers.

9.4.1 Small-Signal Current

In general, the current I_{DS} is a function of both V_{GS} and V_{DS} . In the small-signal approximation, we begin by decomposing the drain-source current into two terms, the DC and AC parts:

$$I_{DS}(V_{GS}, V_{DS}) = I_{DS,Q} + i_{ds} \quad (9.20)$$

Then we perform a Taylor Series expansion of the current:

$$I_{DS}(V_{GS} + v_{gs}, V_{DS} + v_{ds}) \approx I_{DS,Q} + \frac{\partial I_{DS}}{\partial V_{GS}} \Big|_Q \cdot v_{gs} + \frac{\partial I_{DS}}{\partial V_{DS}} \Big|_Q \cdot v_{ds} + \dots \quad (9.21)$$

If we ignore all terms but the linear, the **AC output signal** is given by the simple relation:

$i_{ds} = g_m v_{gs} + \left(\frac{1}{r_o} \right) v_{ds}$

AC output signal (9.22)

g_m is the **(transfer conductance)**:

$g_m = \frac{\partial i_{DS}}{\partial v_{GS}} \Big|_Q$

Transconductance (9.23)

g_o is the **output conductance**, which is the inverse of the output resistance:

$g_o = \frac{1}{r_o} = \frac{\partial i_{DS}}{\partial v_{DS}} \Big|_Q$

Output conductance (9.24)

The meaning of g_m is illustrated graphically in Fig. 9.7. First we hold V_{DS} constant at $V_{DS,Q}$ and sweep the gate-to-source voltage V_{GS} and plot the current, as shown. We are implicitly assuming that $V_{DS} > V_{DS,sat} = V_{GS} - V_T$ so that the quadratic relation holds. Then we evaluate the slope of the curve at the operating point of interest, $V_{GS,Q}$. This slope represents the small-signal current produced by a small voltage excursion of v_{gs} about the operating point.

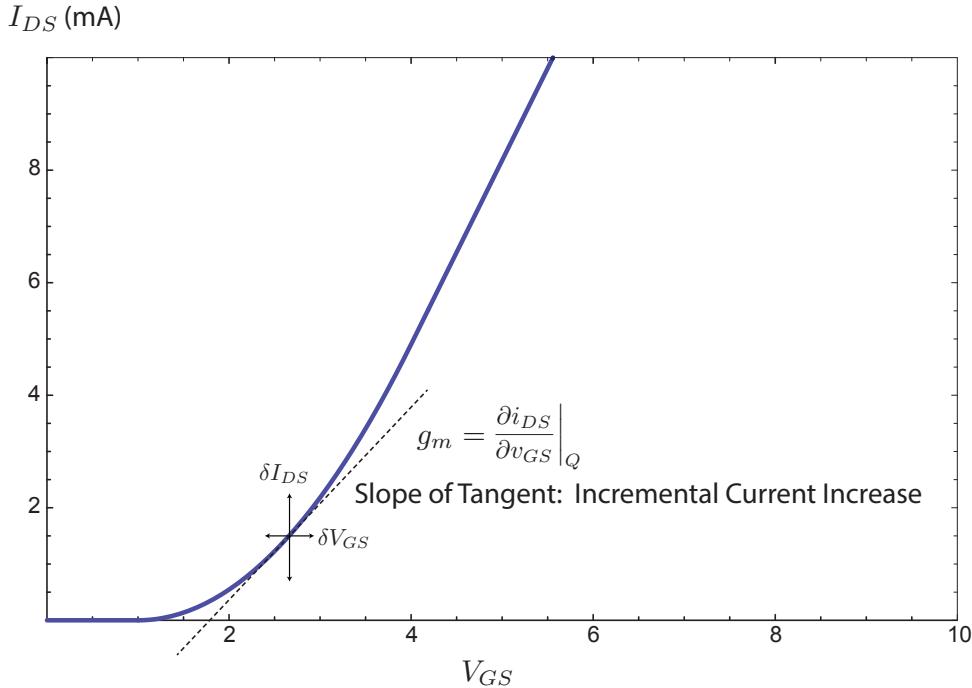


Figure 9.7: The transconductance g_m is the slope of the I_{DS} vs V_{GS} curve for a fixed value of V_{DS} .

9.4.2 The MOS Transconductance

By using the definition of transconductance (the change in drain current due to a change in the *gate-source* voltage), and the MOSFET saturation drain current (*Eq. 8.44*), with *everything else constant*, we evaluate the derivative of $I_{DS,sat}$ with respect to V_{GS} :

$$g_m = \frac{\Delta I_D}{\Delta V_{GS}} \Big|_{V_{GS}, V_{DS}} = \frac{\partial I_{DS}}{\partial V_{GS}} \Big|_{V_{GS}, V_{DS}} = \left(\frac{W}{L} \right) \mu_n C_{ox} (V_{GS} - V_T) \underbrace{(1 + \lambda V_{DS})}_{\approx 0} \quad (9.25)$$

$$= \left(\frac{W}{L} \right) \mu_n C_{ox} (V_{GS,Q} - V_T) = \left(\frac{W}{L} \right) \mu_n C_{ox} (V_{DS,sat}) \quad (9.26)$$

This equation shows that g_m is a strong function of the gate-overdrive voltage ($V_{GS} - V_T$), the device mobility μ , and the gate-oxide capacitance C_{ox} . We can write g_m in other equivalent forms. For instance, by solving for $V_{GS} - V_T$ from *Eq. 8.44* (with $\lambda = 0$), and substituting into *Eq. 9.26*:

$$g_m = \left(\frac{W}{L} \right) \mu_n C_{ox} \sqrt{\frac{2L \cdot I_{DS}}{W \mu C_{ox}}} = \sqrt{\frac{2L \mu^2 W^2 C_{ox}^2 \cdot I_{DS}}{L^2 W \mu C_{ox}}} = \boxed{\sqrt{\left(\frac{2W}{L} \right) \mu C_{ox} I_{DS}}} \quad (9.27)$$

Eq. 9.27 shows that the g_m increases like $\sqrt{I_{DS}}$ with increasing bias. Finally, we can express g_m in a third way, by dividing *Eq. 8.44* by $V_{GS} - V_T$:

$$\frac{I_{DS,sat}}{V_{GS} - V_T} = \frac{1}{2} \underbrace{\left[\left(\frac{W}{L} \right) \mu C_{ox} (V_{GS} - V_T) \right]}_{g_m} = \frac{g_m}{2} \implies \boxed{g_m = \frac{2 I_{DS}}{V_{GS} - V_T}} \quad (9.28)$$

This form may seem the most ambiguous, because it depends on both the bias current and the gate-overdrive, but it turns out to be a very useful way to compare the required g_m to meet certain specifications, as we will show in the following chapters.

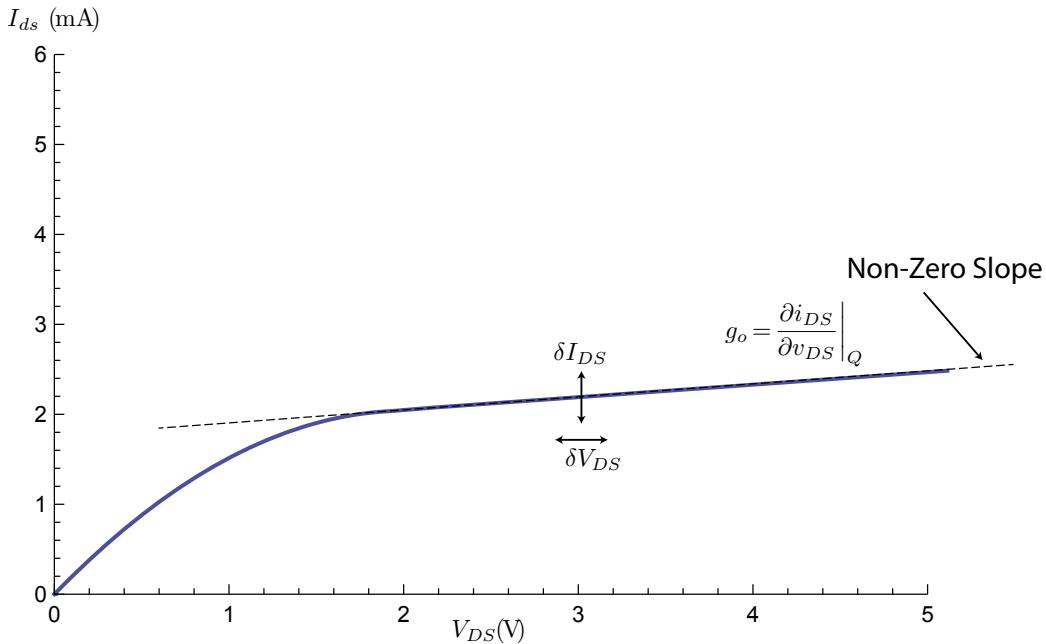


Figure 9.8: The output conductance g_o is the slope of the I_{DS} vs V_{DS} curve for a fixed value of V_{GS} . It's more common to state the output resistance $r_o = g_o^{-1}$.

9.4.3 Output Resistance r_o

We again start with Eq. 8.44, but now to find g_o ($1/r_o$), defined as the change in drain current due to a change in the *drain-source* voltage, with *everything else constant*. In other words:

$$g_o = \left. \frac{\partial i_D}{\partial v_{DS}} \right|_{V_{GS}, V_{DS}} \quad (9.29)$$

As shown in Fig. 9.8, it is more common to state the output resistance, which is the reciprocal of the output conductance:

$$r_o = \left. \frac{\partial i_D}{\partial v_{DS}} \right|_{V_{GS}, V_{DS}}^{-1} \quad (9.30)$$

Since the current in Eq. 8.44 is only a linear function of V_{DS} , the derivative is easy to evaluate:

$$r_o = \frac{1}{\frac{W}{L} \frac{\mu C_{ox}}{2} (V_{GS} - V_T)^2 \lambda} \quad (9.31)$$

If λ is small, then the current is approximately the term in the denominator:

$r_o \approx \frac{1}{\lambda I_{DS}}$

Output resistance (9.32)

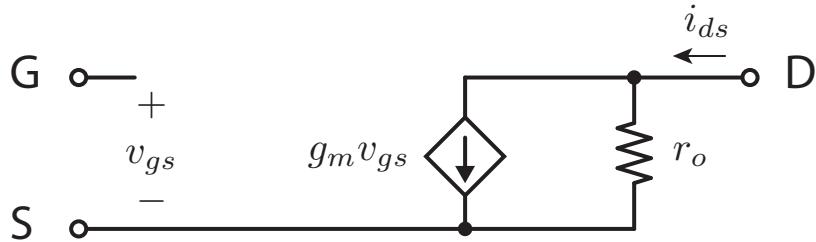


Figure 9.9: The small-signal equivalent circuit model of a NMOS device. This is a three terminal model and does not include any charge storage effects.

9.4.4 Small-Signal Model for MOSFET

The derived equation for the small-signal linearized current, repeated here for convenience:

$$\boxed{i_{ds} = g_m v_{gs} + \left(\frac{1}{r_o} \right) v_{ds}} \quad \text{MOSFET small-signal current} \quad (9.33)$$

If we translate *Eq. 9.33* into a circuit diagram, we obtain the simplified 3-terminal small-signal model for a MOSFET shown in *Fig. 9.9*. In the next chapter, we will develop a more complete model. It is important to note that once you specify the DC operating point of an amplifier, you also fix the small-signal parameters. Thus, to determine the small-signal circuit, you only need to know the DC operating point.

9.4.5 Small-Signal Analysis Steps: The "Easy Way"

Earlier we solved for the small-signal response of the amplifier the "hard way". Now we have introduced a systematic approach. We can summarize the small-signal analysis technique by breaking it down into simple steps:

1. Calculate the bias points using DC sources.
2. Use the bias point to determine the MOSFET region of operation, as well as to calculate small-signal parameters.
3. Turn off the DC sources in the schematic (short DC voltages, open DC currents).
4. Redraw the schematic by using the small-signal model for the MOSFET in place of the transistor.
5. Calculate the gain $\left(\frac{v_{out}}{v_{in}} \right)$ of the circuit, or any other parameter of interest.

Small Signal Analysis: Step 1 and 2

We already found the DC operating point in the earlier part of this chapter. From the DC operating point we can next calculate the **small-signal parameters** of the transistor:

$$\boxed{g_m = \frac{2I_{DS}}{(V_{GS} - V_T)}} \quad \text{Transconductance}$$

$$\boxed{r_o \approx \frac{1}{\lambda I_{DS}}} \quad \text{Output resistance}$$

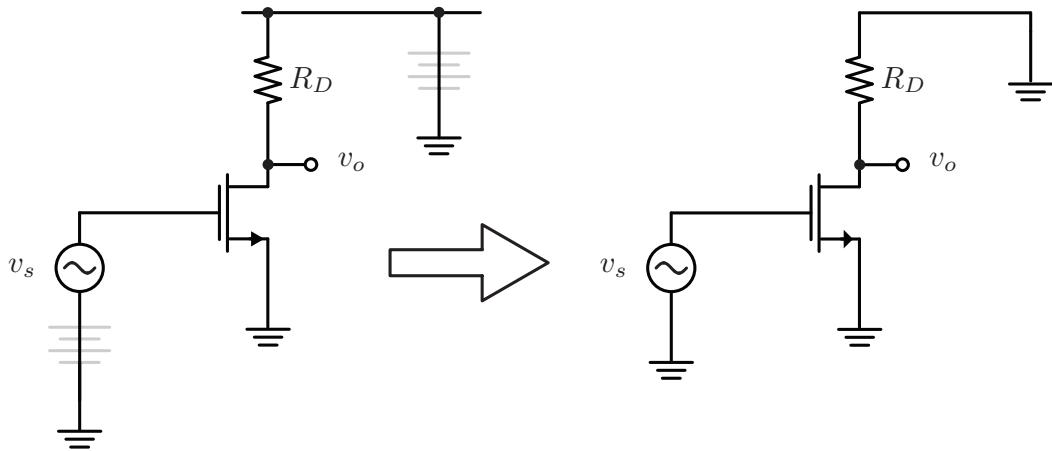


Figure 9.10: The AC model of a the amplifier is produced by shorting out all DC voltage sources.

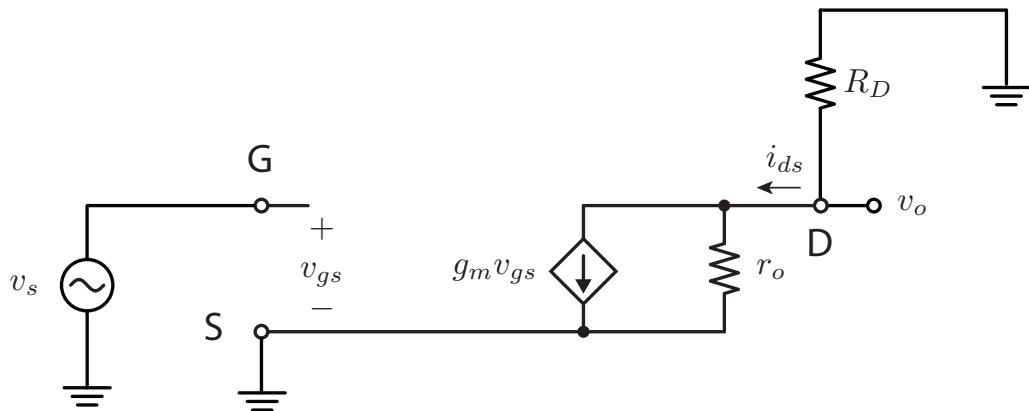


Figure 9.11: The small-signal AC model of the *NMOS* common source amplifier. Note that this is derived by replacing the *NMOS* transistor (*Fig. 9.10*) with its small-signal equivalent circuit (*Fig. 9.9*).

Small Signal Analysis: Step 3

In this step we turn off all DC sources, as shown in *Fig. 9.10*. The DC voltage sources become shorts and DC current sources become opens.

Small Signal Analysis: Step 4

Now we draw a new schematic, by replacing the transistor with its small-signal model, shown in *Fig. 9.11*. As you become more proficient at analyzing transistor circuits, you can sometimes avoid this step. But that will come later, maybe in this class or maybe in your next. For now it's totally okay to just draw out the circuit to familiarize yourself with it.

Instead of a "macro" replacement, you should actually redraw the circuit to simplify the connections and make things as obvious as possible. This will facilitate "inspection" style analysis. For example, you will be able to identify that R_D and r_o are in parallel, so you can treat them as a single equivalent resistor. See *Fig. 9.12*.

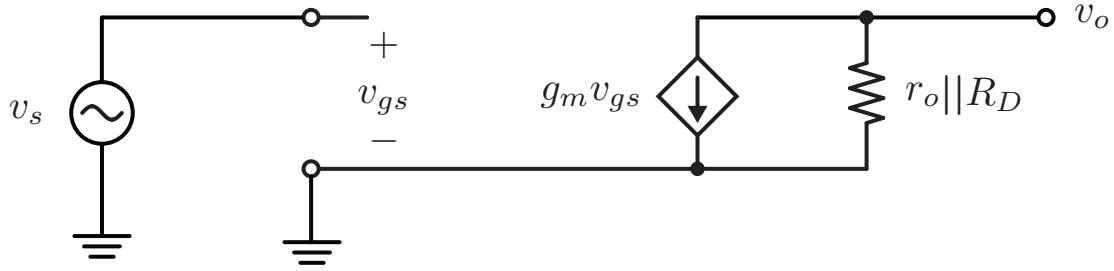


Figure 9.12: A simplified small-signal AC model of the amplifier, derived from *Fig. 9.11*.

Small Signal Analysis: Step 5

Now we are ready to analyze the circuit like any other linear AC circuit. You can find the gain, the input/output impedance, etc. This circuit is very easy to analyze, and the output voltage is given by

$$v_o = -g_m v_{gs} (r_o \parallel R_D) \quad (9.34)$$

Notice that the input signal v_s is actually the same as v_{gs} , so we have

$$A_v = \frac{v_o}{v_s} = -g_m (r_o \parallel R_D) \quad (9.35)$$

We have the device output resistance in this analysis "for free", where we neglected it in the "hard way" analysis we did earlier. Still, the final answer doesn't seem to agree with our earlier analysis. To see why, first write g_m in terms of the transistor bias. In this case, it is convenient to use this form:

$$g_m = \frac{2I_{DS,Q}}{V_{GS,Q} - V_T} \quad (9.36)$$

$$A_v = -\frac{2I_{DS}(r_o \parallel R_D)}{V_{GS} - V_T} \quad (9.37)$$

If we assume $r_o \gg R_D$, then we get the same result as before because $I_D R_D = \frac{V_{DD}}{2}$:

$$A_v \approx -\frac{2I_{DS}R_D}{V_{GS} - V_T} = \frac{2\frac{V_{DD}}{2}}{V_{GS} - V_T} = \frac{V_{DD}}{V_{GS} - V_T} \quad (9.38)$$

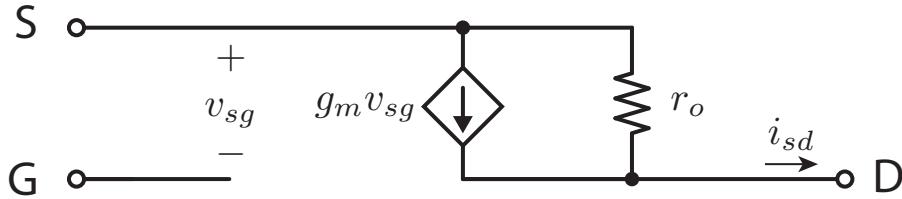


Figure 9.13: The PMOS small-signal equivalent circuit model. Compare with Fig. 9.9.

9.5 PMOS Amplifier and Small-Signal Model

9.5.1 Small-Signal PMOS Model

So far we have focused on the NMOS device. Let's introduce the PMOS small-signal model shown in Fig. 9.13. This model looks like an upside down NMOS model. Remember that the current flows from the source to the drain of a PMOS (from *higher potential to lower potential*). In an NMOS, the drain is at a higher potential, so you must flip the model around. A good way to sanity check is to "wiggle" the gate of the transistor and observe what the model predicts, and make sure the polarity is correct. If we increase the gate voltage for a PMOS, we are reducing the inversion layer conductance, so the current should go down. Note that the small-signal model agrees and shows the current going negative.

9.5.2 PMOS Amplifier

We take the same exact steps as in the NMOS amplifier to produce an amplifier with a PMOS device, shown in Fig. 9.14. Note this is a "Common Source" Amplifier even though the PMOS source connects to DC supply. **DC supplies are AC grounds, so the source is actually AC grounded.** The small-signal model of the amplifier is drawn in Fig. 9.15, and it is identical to the NMOS case. So there's no reason to redo the math.

Which is better, the NMOS or PMOS? Traditionally NMOS amplifiers are preferred, because the electron mobility is higher. This leads to a higher g_m , and thus a higher gain. But PMOS amplifiers are also useful, and they will be used extensively. Namely, when you need to "flip" the DC bias around you can use an cascade of an NMOS and PMOS, and avoid large coupling capacitors. That's jumping ahead, but you will see NMOS and PMOS amplifiers used together in cascade often by the end of this book.

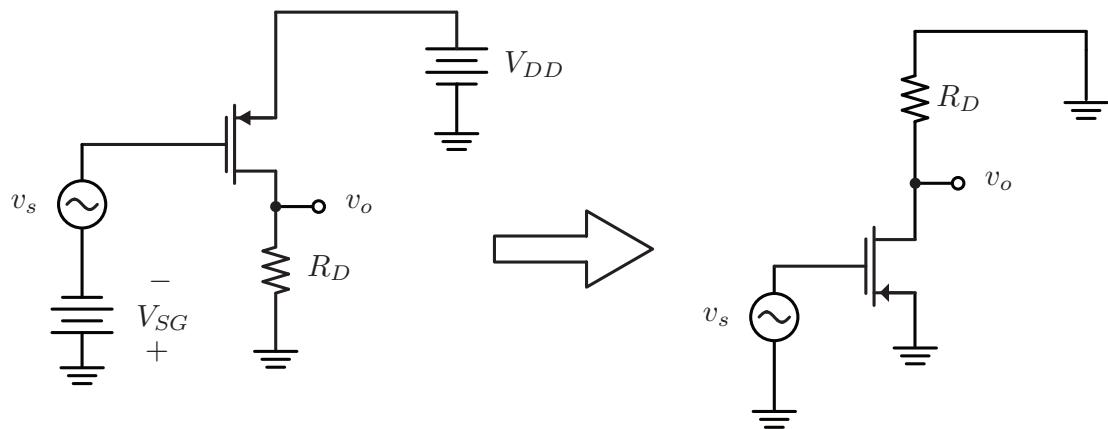


Figure 9.14: The PMOS common-source amplifier and its AC equivalent circuit.

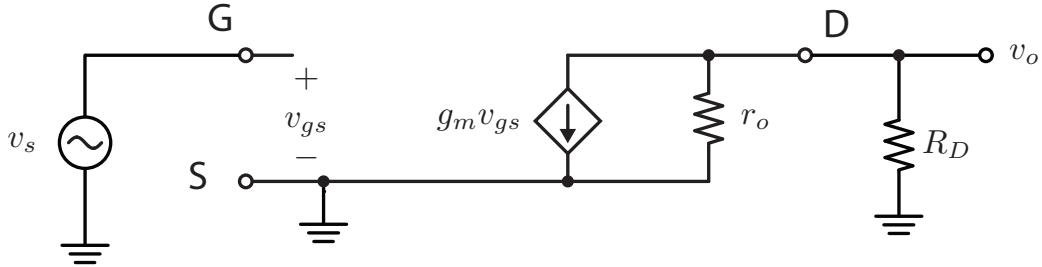
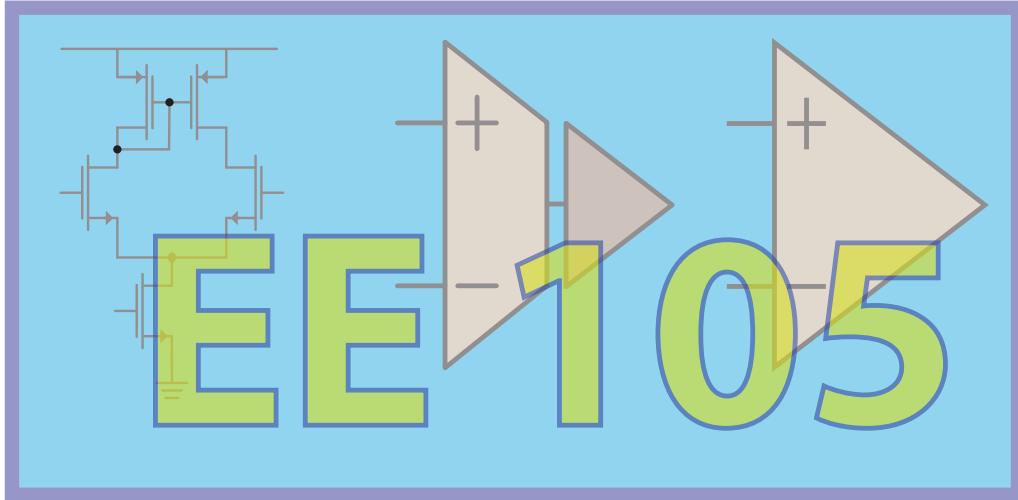


Figure 9.15: The re-drawn PMOS small-signal AC schematic, which is identical to the NMOS case.

9.6 What We've Ignored

We have actually made a lot of approximations up to now. But simplicity is a good thing, and the models we have developed in this chapter will form the core of our analysis and intuition. When we want to model other "second-order" effects, in some cases we must introduce the fourth terminal of the MOSFET, or bring back the body of the transistor. We will show that the body of the transistor acts like a second gate, or "back gate", and cannot be ignored if the body-source voltage is moving around. We also know that the source/drain junctions of the transistor form *PN*-junction diodes with body, which introduces parasitic capacitance. In fact, we also ignored the gate-oxide capacitance. It turns out that we can simply add many of these capacitors to our small-signal model, as we will show in the next chapter.

Also modern transistors are very short channel length devices, and there are many "short channel" effects that we are ignoring. Some of these effects are velocity saturation, and other "high field" effects. Finally, we have been assuming that you have to bias the transistor in the strong inversion region $V_{GS} > V_T$. However, it turns out that even with moderate inversion, or even sub-threshold bias, the transistor still behaves like a transistor for AC signals. This is especially useful in low power circuits which are preferentially biased in moderate or weak-inversion (sub-threshold). This is a more advanced topic that you can learn in your next course on circuit design.



10. Complete MOS Small-Signal Model

10.1 Chapter Preview

In this chapter we make amends for ignoring charge storage effects in the MOS transistor. In particular, we focus on the MOSFET capacitance in the saturation regime, including the Gate-Source C_{gs} , Gate-drain C_{gd} , and drain/source to bulk C_{db} and C_{sb} . This will result in a complete *four terminal* MOS small-signal model.

Some of these capacitances are parasitic to the MOSFET, and they introduce undesirable poles and zeros to the circuit. This has the negative effect of limiting the bandwidth, and slowing the circuit down.

A more subtle effect comes from the so-called "Back-Gate Effect", or how the threshold voltage can be modulated by a source-body bias. We will model this in the small-signal AC circuit by introducing a back-gate transconductance. This will result in the complete *four terminal* MOS small-signal model. This model will be used extensively throughout the rest of this book.

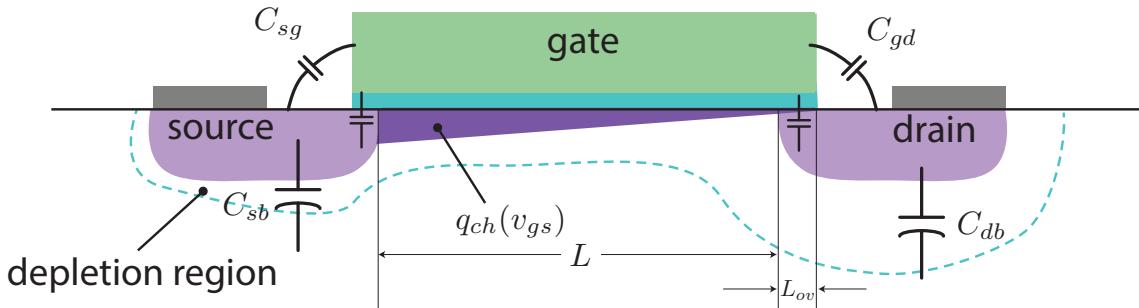


Figure 10.1: Cross section of an MOS transistor highlighting the internal capacitance arising from q_{ch} and the parasitic capacitors due to PN -junctions, overlap and fringing capacitance.

10.2 MOSFET Capacitance in Saturation

10.2.1 MOSFET Cross Section

As shown in the device cross-section of *Fig. 10.1*, MOSFETs have many capacitances. Some are critical to the function of the MOS, such as C_{ox} . However, others are undesired parasitics. Let's analyze where each capacitance comes from step-by-step. For now we will ignore the influence of the back gate, and just assume that the body is AC grounded. In this case, note that channel charge is mostly controlled by gate-source voltage and not the drain voltage.

10.2.2 Gate-Source Capacitance C_{GS}

Since the drain is isolated from the channel (in saturation), the oxide capacitance C_{ox} is formed from the gate to source, as shown in *Fig. 10.2*. But since the inversion charge is "wedge" shaped (inversion decreases as we travel from the source to drain), the **effective capacitance** is lower. A detailed calculation shows that:

$$C_{gs} = \left(\frac{2}{3}\right) W L C_{ox} + C_{ov} \quad \text{Gate-source capacitance in saturation} \quad (10.1)$$

We include a **parasitic overlap capacitance** C_{ov} along source edge of gate:

$$C_{ov} = L_{ov} W C_{ox} \quad \text{MOSFET parasitic overlap capacitance} \quad (10.2)$$

The physical origin of the overlap capacitance is the overlap between the gate and the source, resulting from imperfect alignment between the gate and the source. This causes a diffusion region to grow under the gate. In a "good" transistor, $L_{ov} \ll L$, so $C_{gd} \ll C_{gs}$. The actual value is higher due to fringing fields that leak from the gate to the top of the source and drain junctions.

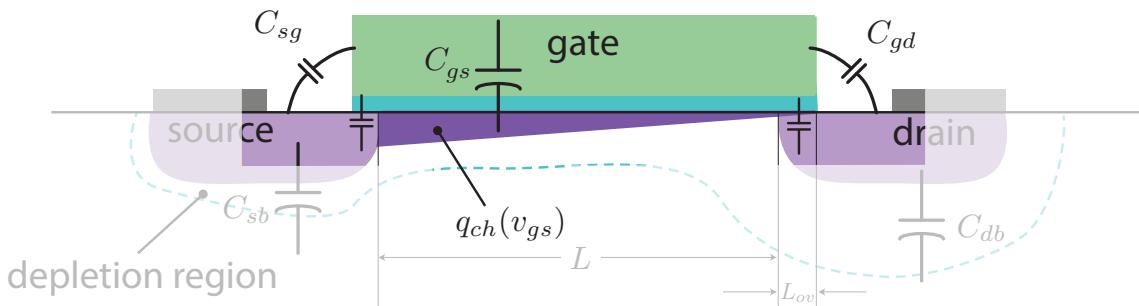


Figure 10.2: The gate-source capacitance is made of two parts; the inversion charge $q_{ch}(v_{gs})$ which is part of the MOS-C internal structure and depends on C_{ox} , and other parasitic capacitors.

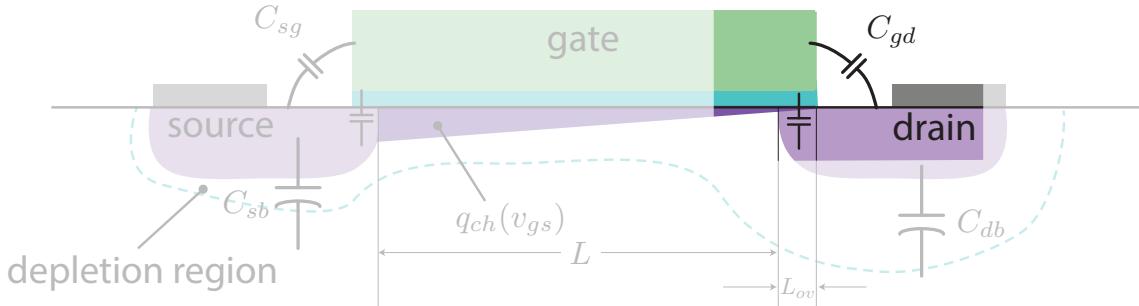


Figure 10.3: The gate-drain capacitance is a parasitic capacitance due to oxide/drain diffusion overlap and fringing terms shown.

10.2.3 Gate-Drain Capacitance C_{GD}

Focusing now on the drain side, *Fig. 10.3*, we see that the same overlap occurs on the drain side. This **gate-drain capacitance** is another undesirable parasitic:

$$C_{gd} = C_{ov} + C_{fringe} = L_{ov} W C_{ox} + C_{fringe} \quad \text{Gate-drain capacitance in saturation} \quad (10.3)$$

Since the MOSFET drain and source are physically identical, this is not surprising.¹ Just keep in mind that this capacitance is not due to change in inversion charge in channel.

10.2.4 Drain/Source-Bulk Capacitances C_{DB} and C_{SB}

The **Drain-Bulk** and **Source-Bulk** capacitances are two more undesired parasitics. They are caused by *PN*-junction depletion regions that isolate the drain/source from the body. The junction forms a "box" inside the body, and so there are five junctions to consider. *Four side-wall junctions* and one *bottom wall junction*, shown in *Fig. 10.4a*.

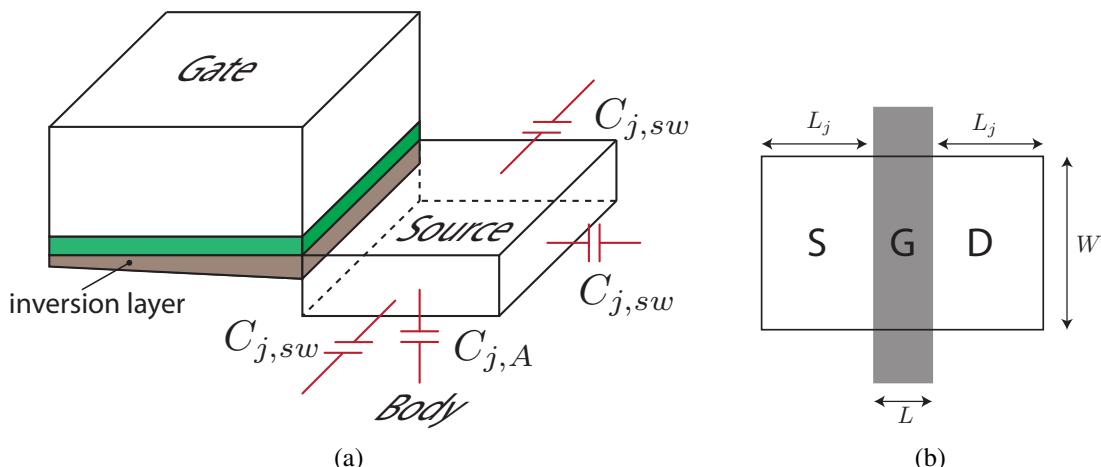


Figure 10.4: (a) The drain/source junctions form *PN*-junctions with the body. Under normal operating conditions, these capacitors are reverse biased. There are "sidewall" contributions $C_{j,sw}$ shown and a bottom plate term $C_{j,A}$. (b) The junction perimeter and area is calculated from the MOS layout and depends on the width W of the device in addition to the length of the junctions L_j .

¹Some special devices use asymmetric source/drain junctions.

The capacitance is defined in terms of the junction areas $A_{s,j}$ and $A_{d,j}$ and the junction perimeters, $P_{s,j}$ and $P_{d,j}$. As shown in Fig. 10.4b, besides the transistor L and W parameters, we need to know the length of the source and drain diffusion regions, L_j , to calculate the area and perimeter. The junction area is simply $L_j \cdot W$ whereas the perimeter is $(2 \cdot L_j + W)$, because the inner junction facing the channel is isolated from the bulk.

$$C_{sb} = \left(\frac{C_{j_0}}{\sqrt{1 + V_{SB}/|\phi_p|}} \right) A_{s,j} + \left(\frac{C_{j,sw_0}}{\sqrt{1 + V_{SB}/|\phi_p|}} \right) P_{s,j} \quad \text{Source-body capacitance} \quad (10.4)$$

$$C_{db} = \left(\frac{C_{j_0}}{\sqrt{1 + V_{DB}/|\phi_p|}} \right) A_{d,j} + \left(\frac{C_{j,sw_0}}{\sqrt{1 + V_{DB}/|\phi_p|}} \right) P_{d,j} \quad \text{Drain-body capacitance} \quad (10.5)$$

Although we have used a **grading coefficient** of $1/2$ for junctions (leading to square roots in the denominator), in reality this is a parameter that you will adjust based on process parameters.

It is important to realize that the source/drain are symmetric. The source/drain is defined by potentials/biasing and the schematic rather than the process. For this reason, the doping parameters for the source and drain are symmetric, and so the calculations for the source and drain to bulk are very much the same. Of course, the area of the junctions may differ based on the device layout. Some special technologies (high breakdown devices) have an asymmetric source/drain, but this is rare.

If the source and body are both AC grounded, then the source-to-bulk capacitance does not play a role in AC response, and it is excluded from the model.

10.2.5 Three Terminal Small Signal Model Including Capacitors

We can now update our three-terminal small-signal model to include three capacitors, as shown in Fig. 10.5. The capacitance C_{gs} is a combination of the gate-oxide capacitance (recall the factor of $2/3$ arising from the uneven charge distribution in the saturation region), and any overlap and fringing capacitance between the gate and the source. On the other hand, the capacitance C_{gd} is all due to overlap, because in the saturation region the drain is isolated from the channel due to pinch-off. The capacitance C_{db} simply accounts for the reverse-biased PN -junction capacitance from the drain diffusion region to the body of the transistor. The capacitor C_{sb} is missing from the model, because it is shorted out in the three terminal model if we assume the source/bulk are tied together. In practice, this may not be true and we should include this capacitor as well. We will cover this in the full four-terminal model.

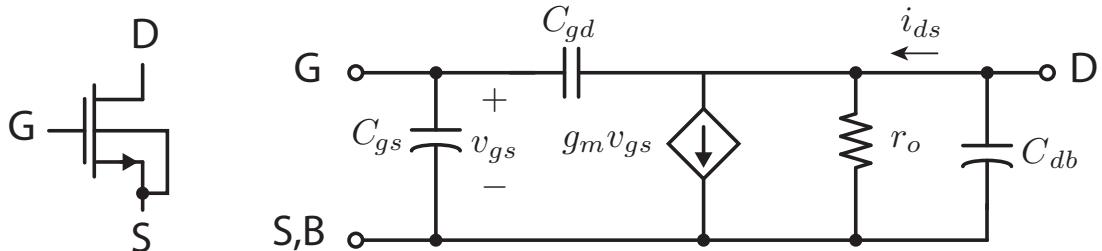


Figure 10.5: The complete three terminal model of a MOSFET with source-body tied together includes capacitance C_{gs} and the parasitic C_{gd} and C_{db} .

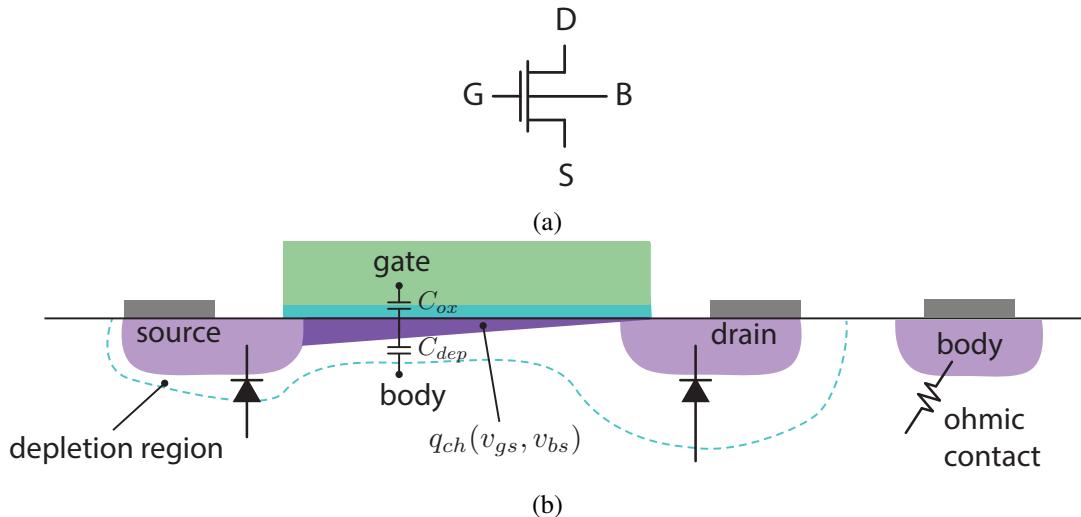


Figure 10.6: (a) The body terminal of a MOSFET is an independent fourth terminal of the device. (b) Variations in the body voltage couple into the channel charge through C_{dep} , in a similar manner as gate voltage variations couple into the channel charge through C_{gs} .

10.3 Back-Gate Effect

10.3.1 All MOSFETS have a "Back Door"

So far we have been ignoring the body terminal of the device, shown explicitly in Fig. 10.6a. The MOS transistor has some symmetry that you should appreciate between the source and drain. It is less obvious, but there is a symmetry between the gate and the body if you think about how the inversion charge depends on both voltages. We will show that the body can act like a "back gate", and control the inversion and thus the current in the device, as shown in Fig. 10.6b. It is something we have ignored, because in many instances the body terminal is simply tied to the source, and the source is sometimes at ground or V_{SS} for NMOS, or supply or V_{DD} for PMOS. What we would like to understand is what happens if there is a DC voltage or an AC voltage swing on the body terminal.

10.3.2 Body Bias Affects: V_T - DC Signals

From our calculations of the MOS capacitor, we know that the body bias V_{SB} has an impact on the channel charge. There is a depletion capacitance between the transistor body and the channel:

$$C_{dep} = \frac{\epsilon_s}{x_{dep,MAX}} \quad (10.6)$$

If we take into account the body voltage with respect to the source, the amount of inversion charge is given by:

$$Q_{inv} = -C_{ox}(V_{GS} - V_T) - C_{dep}V_{BS} \quad (10.7)$$

Here the role of the back-gate is shown explicitly and symmetrically with respect to the front gate. Whereas the front gate, or simply the gate, affects the inversion charge through C_{ox} , the back-gate affects the amount of inversion charge through C_{dep} . Let's factor out C_{ox} and note that $V_{SB} = -V_{BS}$:

$$Q_{inv} = -C_{ox} \left[V_{GS} - V_T + \left(\frac{C_{dep}}{C_{ox}} \right) V_{SB} \right] \quad (10.8)$$

Written in this form, we can lump the back-gate effect into a change in the threshold voltage:

$$V_T(V_{SB}) = V_{T_0} + \left(\frac{C_{dep}}{C_{ox}} \right) V_{SB} \quad (10.9)$$

In Eq. 10.9, V_{T_0} is the zero V_{SB} threshold voltage:

$$V_{T_0} = V_T(0) = V_{FB} - 2\phi_p + \frac{1}{C_{ox}} \sqrt{2q\epsilon_s N_A(-2\phi_p)} \quad (10.10)$$

The change in threshold due to body bias is given by:

$$\Delta V_T = \frac{1}{C_{ox}} \sqrt{2q\epsilon_s N_A} \left(\sqrt{-2\phi_p + V_{SB}} - \sqrt{-2\phi_p} \right) = \gamma \left(\sqrt{V_{SB} - 2\phi_p} - \sqrt{-2\phi_p} \right) \quad (10.11)$$

This is written more compactly in the following form:

$$V_T = V_{T_0} + \gamma \left(\sqrt{V_{SB} - 2\phi_p} - \sqrt{-2\phi_p} \right) \quad \text{Threshold voltage with body bias} \quad (10.12)$$

γ is a **device parameter** in Eq. 10.12:

$$\boxed{\gamma = \frac{\sqrt{2q\epsilon_s N_A}}{C_{ox}}} \quad \text{MOSFET device parameter} \quad (10.13)$$

10.3.3 Role of the Substrate Potential - AC Signals

Since V_{SB} changes the threshold voltage, which in turn changes the drain current, the body acts like a "back-gate", and it should cause an AC current to flow in response to an AC signal between the body and source. The **back-gate transconductance** is defined as:

$$g_{mb} = \frac{\Delta i_D}{\Delta v_{BS}} \Big|_Q = \frac{\partial i_D}{\partial v_{BS}} \Big|_Q \quad (10.14)$$

We can simplify this calculation by using the chain rule and taking advantage of our previous calculation of the change in V_T :

$$g_{mb} = \frac{\partial I_D}{\partial V_{BS}} \Big|_Q = \frac{\partial I_D}{\partial V_T} \Big|_Q \cdot \frac{\partial V_T}{\partial V_{BS}} \Big|_Q \quad (10.15)$$

The threshold voltage is given by Eq. 10.12. Taking the derivatives with V_{BS} and V_T results in the following:

$$g_{mb} = \frac{\partial I_D}{\partial V_{BS}} \Big|_Q = \frac{\partial I_D}{\partial V_T} \Big|_Q \cdot \frac{\partial V_T}{\partial V_{BS}} \Big|_Q = \frac{\gamma g_m}{2\sqrt{-V_{BS} - 2\phi_p}} \quad (10.16)$$

You can also intuitively guess that the answer should in fact be equal to:

$$\boxed{g_{mb} = g_m \left(\frac{C_{dep}}{C_{ox}} \right)} \quad \text{Back-gate transconductance} \quad (10.17)$$

In most transistors $C_{ox} > C_{dep}$, so the back-gate transconductance is smaller.

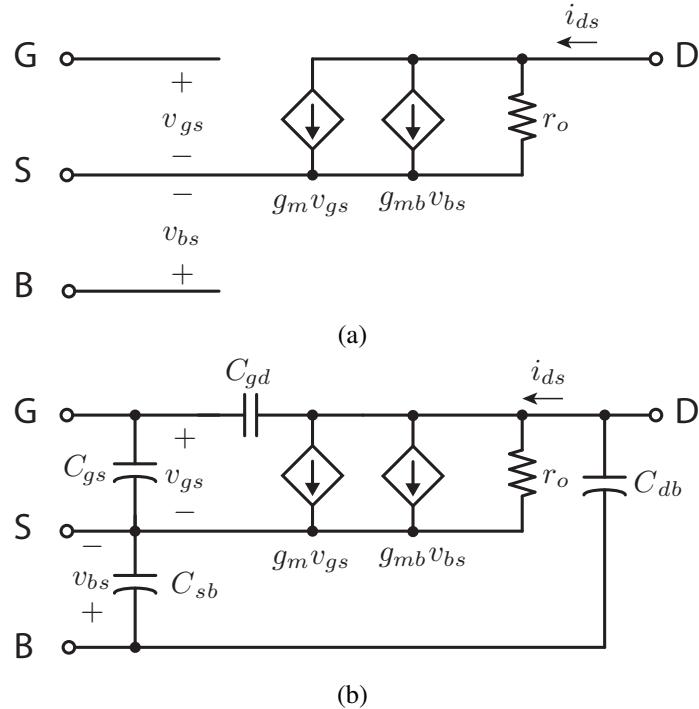


Figure 10.7: The complete small-signal models for an *N*-type MOSFET with (a) no capacitors and with (b) capacitors. The model without capacitors is useful for low frequency calculations.

10.4 Complete Four Terminal MOS Small-Signal Model

10.4.1 Four-Terminal Small-Signal Model

We now complete the small-signal model by noting that the drain-source current depends on v_{gs} (transconductance), r_{ds} (output resistance of device), and also on v_{bs} (back-gate effect):

$$i_{ds} = g_m v_{gs} + \left(\frac{1}{r_o} \right) v_{ds} + g_{mb} v_{bs} \quad \text{Small-signal current with body effect} \quad (10.18)$$

This equation is captured by the equivalent circuit shown in *Fig. 10.7a*. This model naturally shows the symmetry between the front-gate and the back-gate. If we include all the capacitors as well, we arrive at *Fig. 10.7b*, the **complete small-signal model** for an *NMOS* transistor.

10.4.2 Complete Small-Signal Model PMOS

Using the same arguments, we can also conclude that a PMOS device is also described by a similar four-terminal small-signal equivalent circuit model, as shown in *Fig. 10.8*.

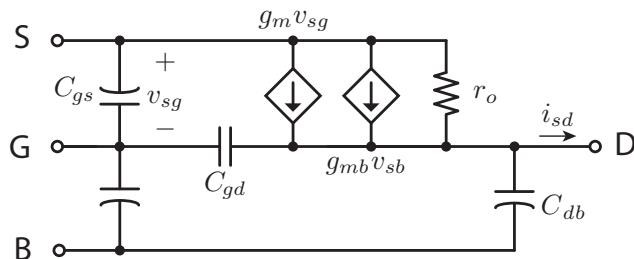
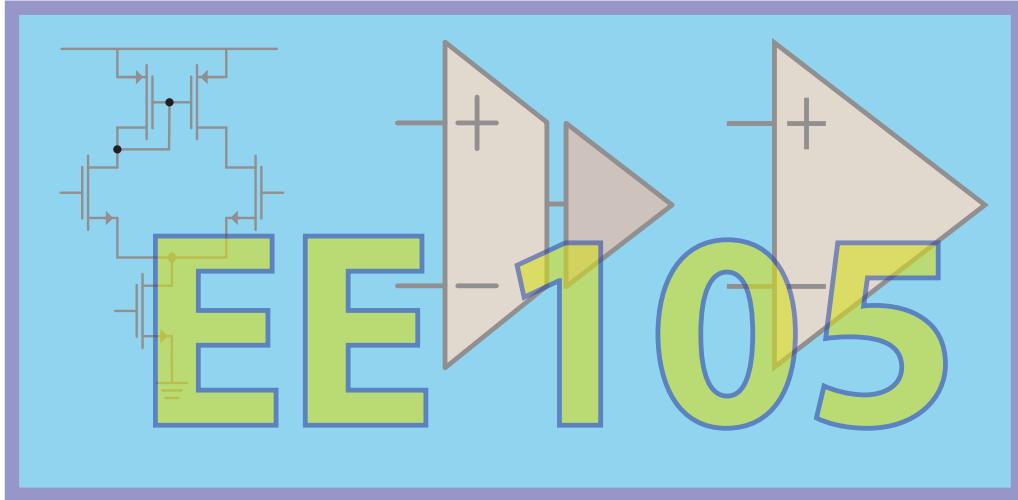


Figure 10.8: The complete small-signal models for an *P*-type MOSFET.



11. Bipolar Junction Transistors

11.1 Chapter Preview

Historically, the first transistor conceived was the MOS device, but the first successful transistor to be built and tested was the bipolar junction transistor, or BJT for short. In fact, the photo shown in the previous chapter (*Fig. 9.1*) was that very transistor. While the CMOS transistors reign supreme in terms of digital logic circuits, offering much lower power and higher density compared to their BJT logic circuit counterparts, BJT transistors continue to play an important role in many applications. For one, they generally have better current handling capability as currents flow through the bulk of the device, as opposed to the surface of an MOS transistor. BJT transistors are also faster in the same lithographic technology node, with the fastest transistors crossing the 1 THz frequency, about 2-3 times faster than CMOS counterparts. But even if you anticipate spending your entire life designing CMOS circuits, there is a good reason to understand bipolar transistor device physics. First and foremost, it is a very interesting device to study with a voltage-current mechanism that is very different from the CMOS transistor. In fact, it is very much a close cousin of the *PN*-junction diode. Second, every CMOS device has a BJT inside of it, like it or not. For low gate voltages, or the so-called sub-threshold region, the BJT dominates and determines the currents.

In this chapter we start with a physical description of the device structure, and then move on to derive the *I-V* characteristics, building on our knowledge of *PN*-junctions. We shall strive to understand the behavior in two ways. First, we will not use any equations, but instead build from our physical understanding of the *PN*-junction. Then we will move on to using equations. We will see that like a MOS device, there is one terminal, the "base", that is "in control" of the current flowing between the other terminals (called the "gate" for MOS). Similar to "source" and "drain", a bipolar transistor has an "emitter" and a "collector". We will conclude the chapter with circuit models of the bipolar transistor, including a "Large Signal" Ebers-Moll model, and then a small-signal model. The small-signal model for a BJT is actually very close to the MOS small-signal model, enabling us to re-use a lot of our background intuition.

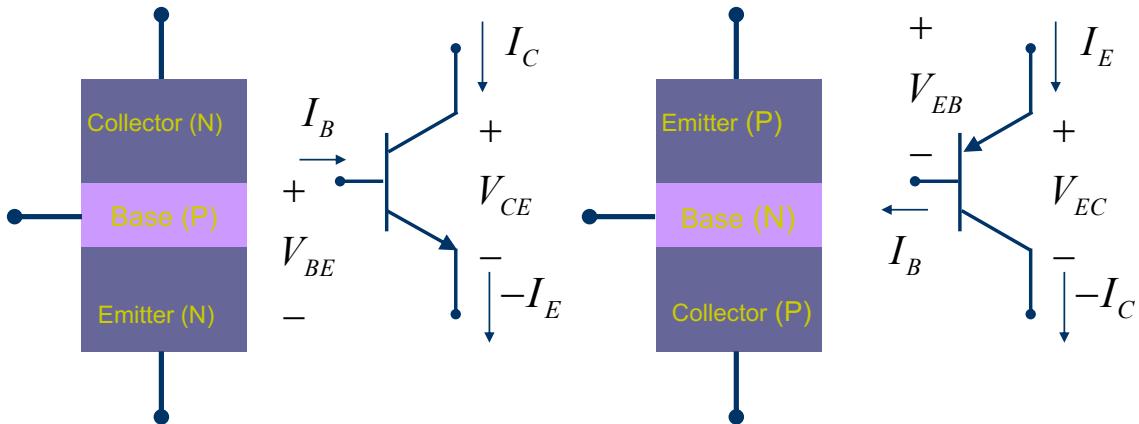


Figure 11.1: Simplified device structure and schematic symbols for *NPN* and *PNP* bipolar junction transistors.

11.2 Overview of BJT

11.2.1 Ideal BJT Structure

The structure of a BJT transistor is deceptively simple. It is just either an *NPN* sandwich, or a *PNP* sandwich, as shown in *Fig. 11.1*. Each region is a doped semiconductor, either *N*-type or *P*-type. The middle part is the **Base** (*B*), and in an *NPN* transistor (similar to NMOS), it's doped *P*-type. Usually a bias voltage is applied across the *NPN* transistor terminals in such a manner such that the top is at a positive voltage, and the bottom at a lower voltage, sometimes ground. The bottom is known as the **Emitter** (*E*), as it emits electrons into the base. The electrons are "collected" at the **Collector** (*C*), at the top terminal. So the current into the collector ideally flows out of the emitter, similar to a MOS device.

But how does current flow at all, since the structure is two back-to-back diodes, pointing in opposite directions? One diode says "one way" North and the other says "one way" South! The two junctions cannot be simultaneously forward biased if the collector is at a high voltage relative to the base, which is the normal **operating point** for the transistor.

The solution lies in making the base very thin so that as minority carriers diffuse across the base from a forward biased **base-emitter junction**, they quickly reach the **base-collector junction**, where they are swept into collector due to the fields from the reverse biased junction, providing a pathway for current to flow. In a well designed transistor, most of the emitter electron current flows into the collector and not the base, so:

$$I_C \approx -I_E \quad (11.1)$$

This means that the base current is very small:

$$I_C \gg I_B \quad (11.2)$$

Because the collector current is a forward-biased diode current of the base-emitter, it has the following form:

$I_C \approx I_S \cdot e^{\frac{qV_{BE}}{kT}}$

BJT collector current (11.3)

This behavior is very much a like a "transistor", because the current through the collector is a function of another terminal-pair, namely the base-emitter. Thus, the BJT device is a **voltage-controlled current source** (VCCS), just like the MOS transistor.

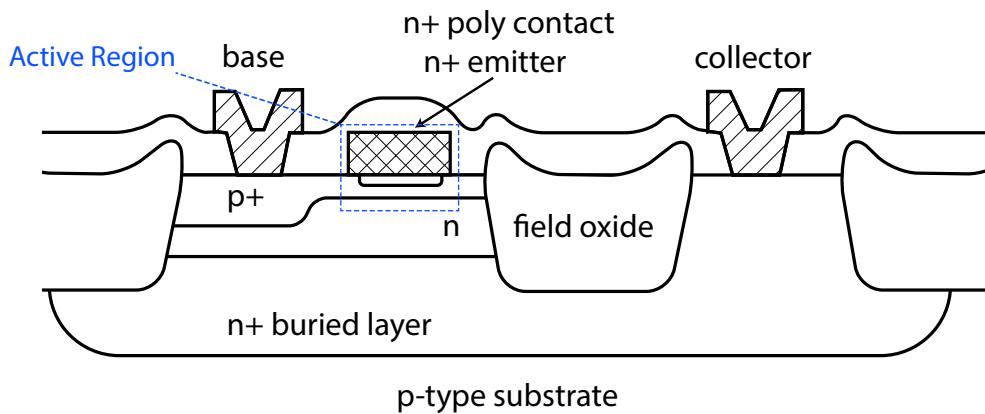


Figure 11.2: Cross-section of a BJT device fabricated in a planar process. The actual transistor active region is under the emitter, with the p^+ base and n^+ buried layer providing low access resistance to the structure.

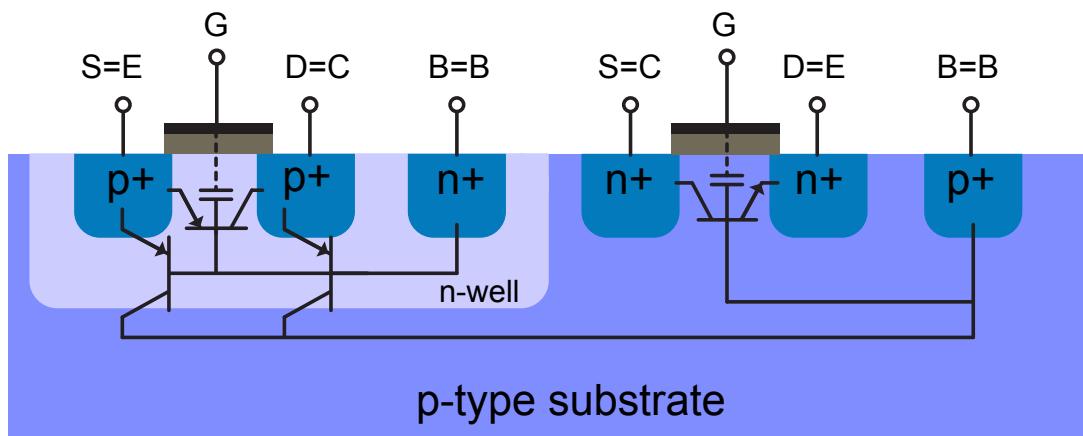


Figure 11.3: Cross-section of a CMOS technology NMOS and PMOS devices. Note that PMOS devices form lateral PNP transistors and substrate transistors. Likewise NMOS devices form lateral NPN devices with all bodies tied to the substrate.

11.2.2 Actual BJT Cross-Section

A real transistor is actually fabricated using a **crystalline** substrate, with various layers grown using diffusion regions and counter-doping. A typical "vertical" BJT is shown in *Fig. 11.2*. This is a complicated figure so let's dissect it piece by piece. First, focus in on the dashed box. This is the "active region" where the transistor actually resides. This region is very much like the *NPN* sandwich of an ideal structure. The collector is the bottom *N*-type region, the middle region is a *P*-type base, and the top is made of an *N*-type emitter. Notice that each layer is doped and so the concentration of dopants necessarily increases from the bottom to the top. The emitter is the most heavily doped, and the collector is the lightest doping. The base is in the middle in terms of doping. This order, collector, then base, and then emitter is chosen for a good reason. Due to symmetry, an N^+PN transistor can be viewed as an upside down transistor with NPN^+ . However, due to the difference in doping levels, this upside down transistor is not a good transistor, which we will see shortly.

Transistors can also be fabricated laterally, similar to NMOS and PMOS devices. In fact, if you examine the cross-section of a MOS device, such as *Fig. 11.3*, it actually has a couple of BJT's hiding in there! Even though the gate couples capacitively to the channel, it can indirectly act like a

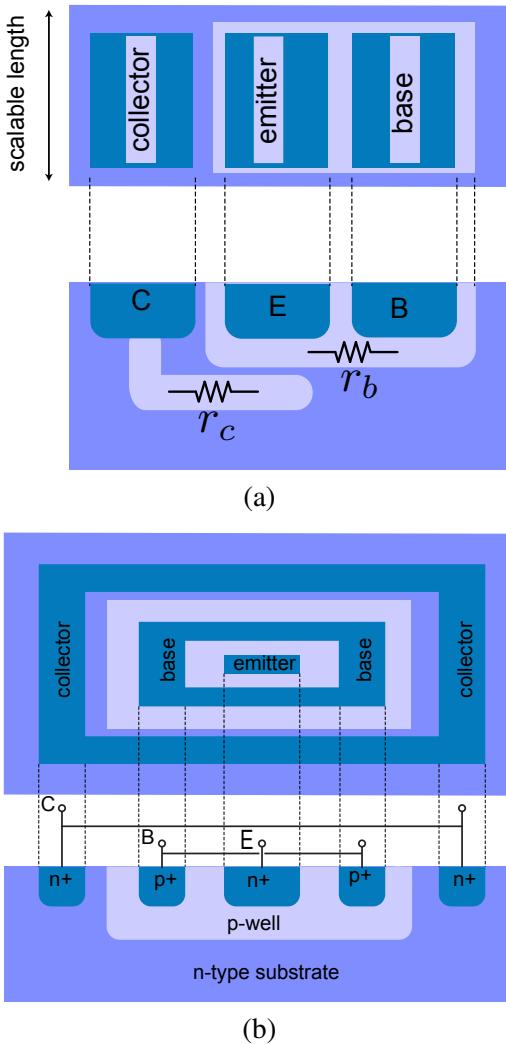


Figure 11.4: (a) A top view of a BJT transistor and corresponding cross-section. The device is scaled by making it wider, or using multiple fingers of a basic device. (b) Top view and cross-section of a BJT with rings of base and collector diffusion regions surrounding the device.

base terminal. These "**lateral**" BJTs tend to have inferior performance because the base width is determined lithographically rather than due to diffusion. One advantage of the BJT structure is in fact the ability to make very thin bases without using lithography. In the early days this was critical for good device performance.

Everything else in the cross-section of the BJT is actually just there to support the device. The n^+ buried layer is a low resistance contact to the collector, and the p^+ region under the base contact serves the same purpose. The emitter is contacted using a highly doped polysilicon film. In fact, film can be used as the emitter itself.

11.2.3 BJT Layout

Let's now view the layout from the top, as shown in *Fig. 11.4a*. This is a unit cell that can grow in width as shown, similar to how we can scale a MOSFET. The emitter area is the most relevant transistor parameter, because the transistor current is injected from the emitter to the base. The base and collector can also form rings around the device to lower the **parasitic resistance** (*Fig. 11.4b*). Although it is not shown in this figure, a multi-finger layout strategy can be used to obtain lower parasitic resistances to the collector and base.

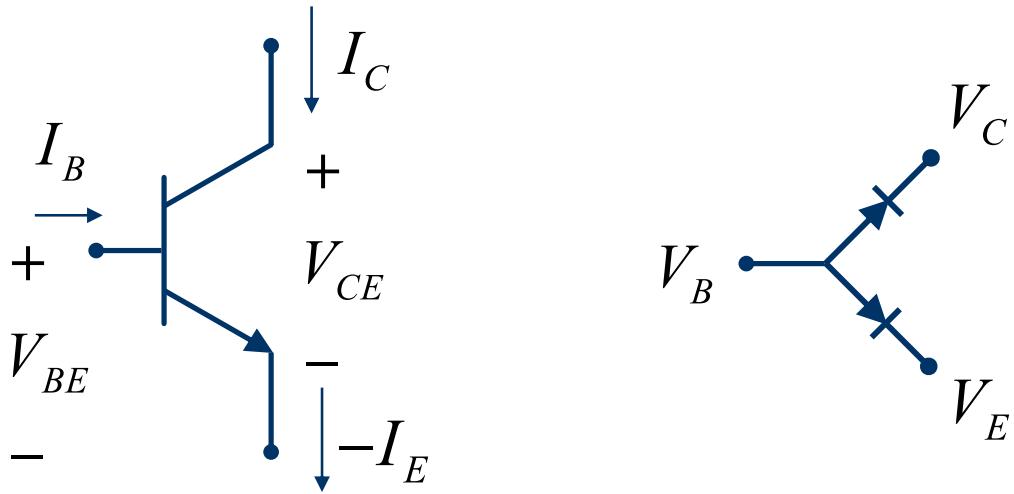


Figure 11.5: Definition of terminal currents (into the device) and junction voltages in a BJT.

11.2.4 BJT Schematic Symbol

The schematic shown in *Fig. 11.5* is for an *NPN* device. The symbol looks a lot like the MOSFET, and the arrow is in the direction of the base-emitter diode. As already noted, the device looks symmetric. But unlike a MOSFET, the collector and emitter cannot be swapped, because of the difference in the doping profiles. In fact, when the BJT is operated in this inverted fashion, it is in the **"reverse active" regime**. In this region the performance is much worse, as we shall explain shortly.

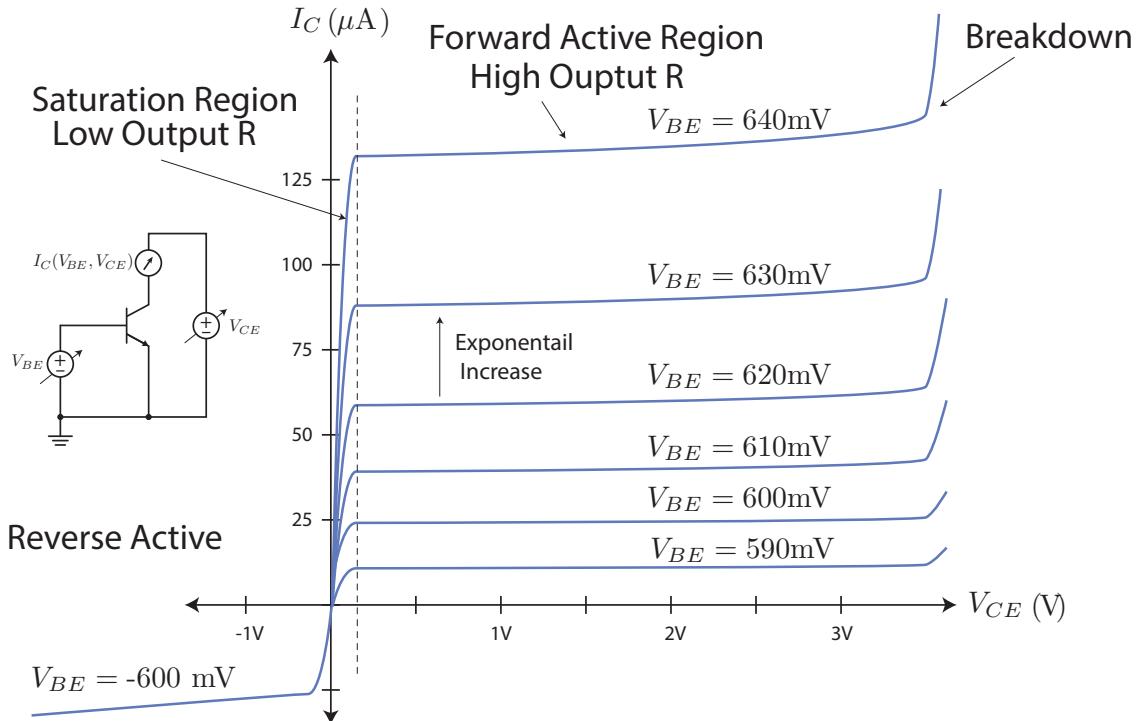


Figure 11.6: The I_C - V_{BE} characteristics of an npn BJT transistor measured by sweeping the V_{CE} voltage (x-axis) and observing the collector current I_C (y-axis). The family of curves are generated by varying the base voltage V_{BE} , as shown in the inset. The current increases exponentially for linear changes in V_{BE} .

11.3 Observed I-V Characteristics

In this section we will show typical **I-V curves** of a BJT transistor, both in terms of a voltage control device and in terms of a current control device. The first set of curves are similar to a MOSFET. But since only a bipolar transistor has a base current, the second perspective is new. One finds experimentally that the following equations describe the behavior of the device:

$$I_C = \beta_F I_B \quad \text{Collector amplification factor} \quad (11.4)$$

$$I_C \approx I_S e^{\frac{qV_{BE}}{kT}} \quad \text{Collector current as a function of emitter voltage} \quad (11.5)$$

These equations show that the **collector current** is an amplified copy of the base current. Alternatively, in terms of voltages, the collector current is an exponential function of the base-emitter voltage.

11.3.1 Base-Emitter Voltage Control

First let's observe the BJT device using voltage control on the **base-emitter junction**, as shown in the inset of Fig. 11.6. For a fixed V_{BE} , we sweep the collector-emitter voltage V_{CE} to generate an I_C - V_{CE} curve, and we observe that the current is approximately flat for a fixed value of V_{BE} . As we increase V_{BE} in discrete steps, we observe a **family of curves**. The collector current increases exponentially as we increase the base-emitter voltage linearly. This is the diode I-V curve we have met before. There is something special and curious about the operation of a BJT. The transistor is a diode between the base-emitter, but the current can be collected at the collector, rather than the base.

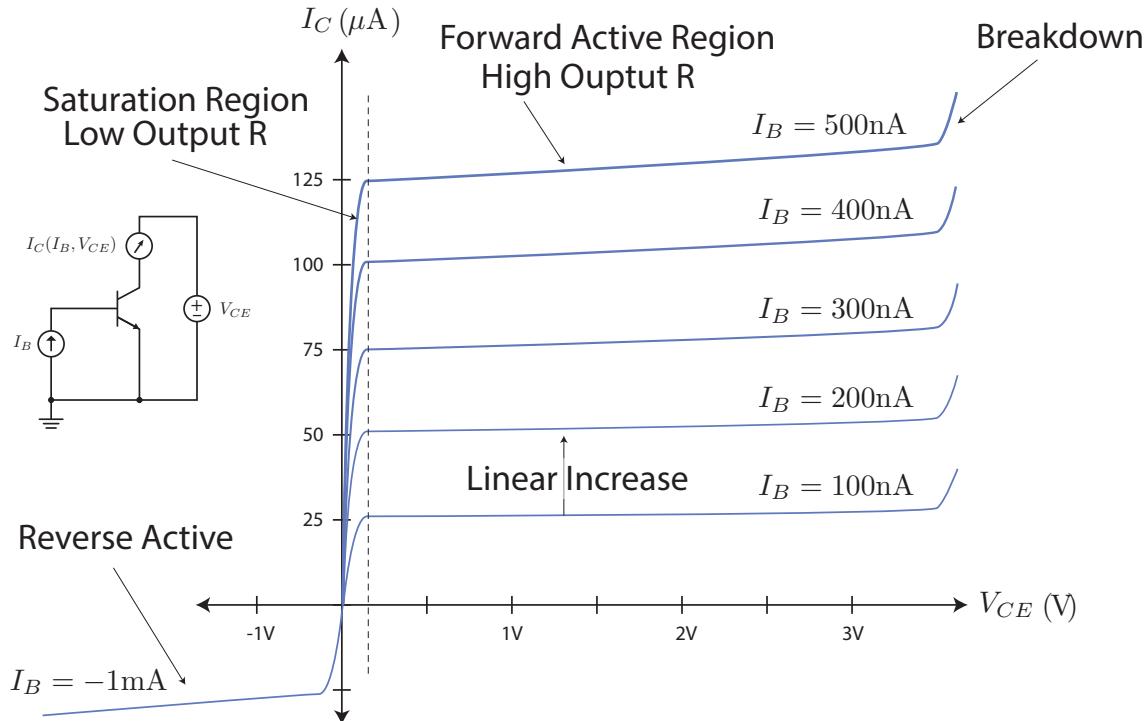


Figure 11.7: The $I_C - V_{CE}$ characteristics of an NPN BJT transistor, measured by sweeping the V_{CE} voltage (x-axis) and observing the collector current (y-axis). The family of curves are generated by varying the base current I_B , as shown in the inset. I_C scales in proportion to I_B linearly, with a current gain defined as β_F .

11.3.2 Collector Characteristics (Sweep I_B)

Now let's view the device as a current controlled device by grounding the emitter and applying a DC voltage V_{CE} . Next we drive the base with a current, I_B , and observe the collector current I_C . The setup is shown in the inset of Fig. 11.7, where we observe that in the "**forward active**" regime, the BJT acts like a **current-controlled current source** (CCCS), with the collector current being a much larger scaled copy of the base current. This behavior is observed over a large range of V_{CE} , but if the collector voltage is dropped below a certain threshold, the device no longer behaves as a current source, and the collector current drops rapidly. What is happening is that in this region the **collector-base junction** is becoming forward biased, and the device is no longer acting like a proper transistor. If the voltage is biased negatively, we enter the *reverse active* region as discussed above.

The key observation is that unlike a MOS device, the BJT has a base current, because the base is not insulated from the transistor like the gate of a MOSFET. However, in a well designed transistor, this base current is a small fraction of the collector/emitter current. Also, because of this base current and KCL, the collector and emitter currents cannot be identical, but they are nearly so.

Now we must address a very confusing point about the nomenclature. In a MOSFET, the forward-active region is also known as the *saturation region*. In a BJT, the region of low V_{CE} is known as the *saturation region*. This is very confusing but it is common practice.

Finally, all transistors have a maximum voltage that you can apply across the collector-emitter before they experience breakdown, or high currents. The same behavior is observed in MOS devices.

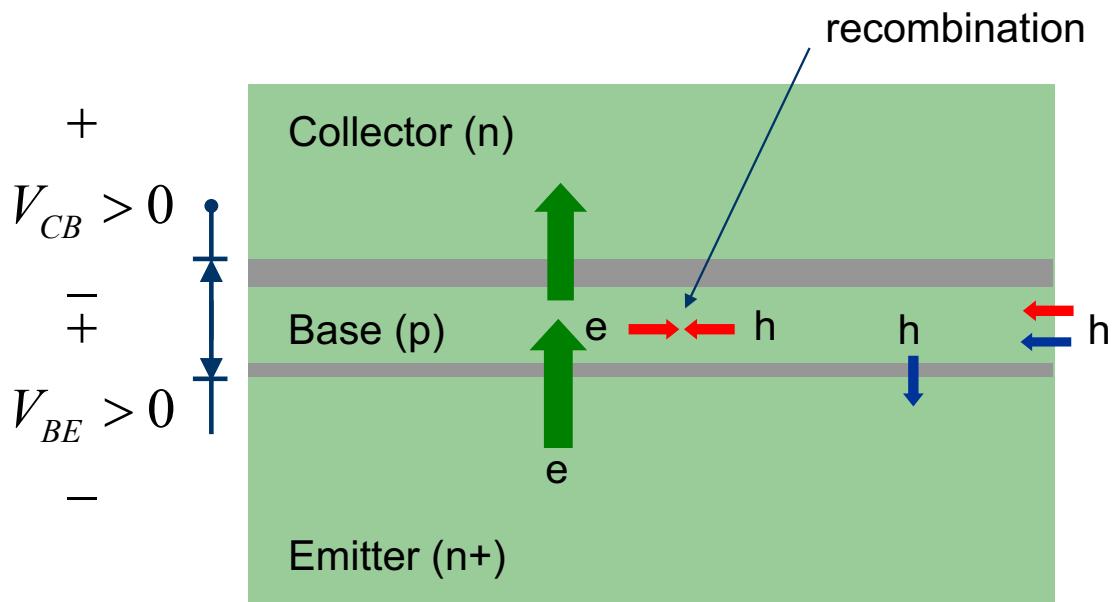


Figure 11.8: Schematic illustration of transistor action. The large arrows show the injection of minority carriers (electrons for *NPN*) into the base, and the subsequent diffusion into the collector region, where they are swept across the reverse-biased junction. The smaller arrows are due to recombination and the correspondingly smaller minority carrier (holes for *NPN*) injection from the base into the emitter.

11.4 BJT Physics *without* Equations

11.4.1 Transistor Action

How does a BJT transistor work its magic? First let's assume that the transistor is in the forward active region. This means that the base-emitter junction is forward-biased, and the collector-base junction is reverse-biased with respect to the base, or the collector is at a higher voltage than the base.

Remember that the base-emitter junction is a *PN*-junction. This means that in forward-bias it will inject electrons into the base due to diffusion, and likewise the base will inject holes into the emitter. This is shown schematically in *Fig. 11.8*. Note that the "electron arrow" is larger than the "hole arrow" because the emitter is heavily doped compared to the base. Now electrons in the base region that were injected from the emitter region can meet one of two fates. They can either cross the base junction and be swept into the collector, or they might recombine with a hole in the base. If the base width is much smaller than the diffusion length of minority carriers in the base, then it is more likely for the electrons to reach the collector, where they are collected, forming the collector current I_C . If most of the emitted electrons are collected in this fashion, $I_C \approx I_E$.

The base current consists of the two portions. The hole current injected into the emitter, and the hole current to support the unlucky but rare electron-hole recombination.

11.4.2 Diffusion Currents

A plot of the minority carriers in the various regions of a bipolar transistor are shown in *Fig. 11.9*. The largest contribution to current is the minority carrier current in the base. In other words, the largest contribution is the electrons injected from the emitter. Also shown is the low concentration of holes in the emitter. Once electrons are collected they are no longer minority carriers, and this is why the minority carrier concentration in the collector is not shown. Here we are interested in gradients in the minority carrier concentration.

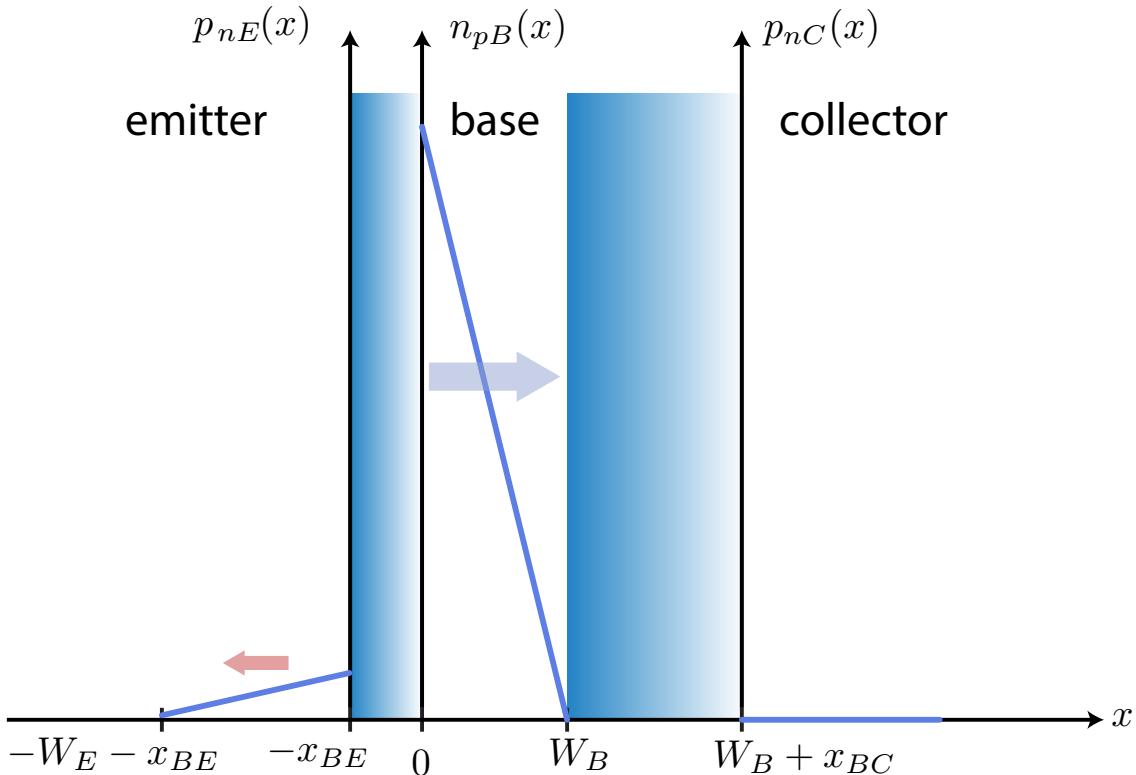


Figure 11.9: Minority carrier profile in an *NPN* transistor. In this figure we assume negligible recombination, resulting in linear minority carrier profiles.

11.4.3 BJT Currents

As we have noted, the collector current is nearly identical to the (magnitude) of the emitter current:

$$I_C = -\alpha_F I_E \quad (11.6)$$

The factor α_F in *Eq. 11.6* is a measure of the number of emitted electrons that are collected. For example, if only 1 in a 1000 recombines in the base, we could say:

$$\alpha_F \sim .999 \quad (11.7)$$

But by Kirchhoff's Law, for steady currents the sum of the currents into a transistor must sum to zero:

$$-I_E = I_C + I_B \quad (11.8)$$

Using the relation between the emitter and collector currents, we have:

$$I_C = -\alpha_F I_E = \alpha_F (I_B + I_C) \quad (11.9)$$

Eq. 11.9 allows us to find the **DC current gain**, or the β_F of a transistor:

$I_C = \left(\frac{\alpha_F}{1 - \alpha_F} \right) I_B = \beta_F I_B$

BJT DC current gain (11.10)

Using values from our earlier calculation, we have:

$$\beta_F = \left(\frac{\alpha_F}{1 - \alpha_F} \right) = \frac{.999}{.001} = 999 \quad (11.11)$$

A good transistor is characterized by a *high value* of β_F . In real transistors an absolute value of β_F is never stated because it varies from device to device. Instead, a range of values are given. A typical $\beta_F = 200$ is a good value to keep in mind. *Never* design a circuit that relies on a particular value of β_F .

Electrons lost to recombination result in an exponentially decaying minority carrier profile, shown in *Fig. 11.10*. So the base current consists of both the recombination current, and the forward-biased injection of holes into the emitter.

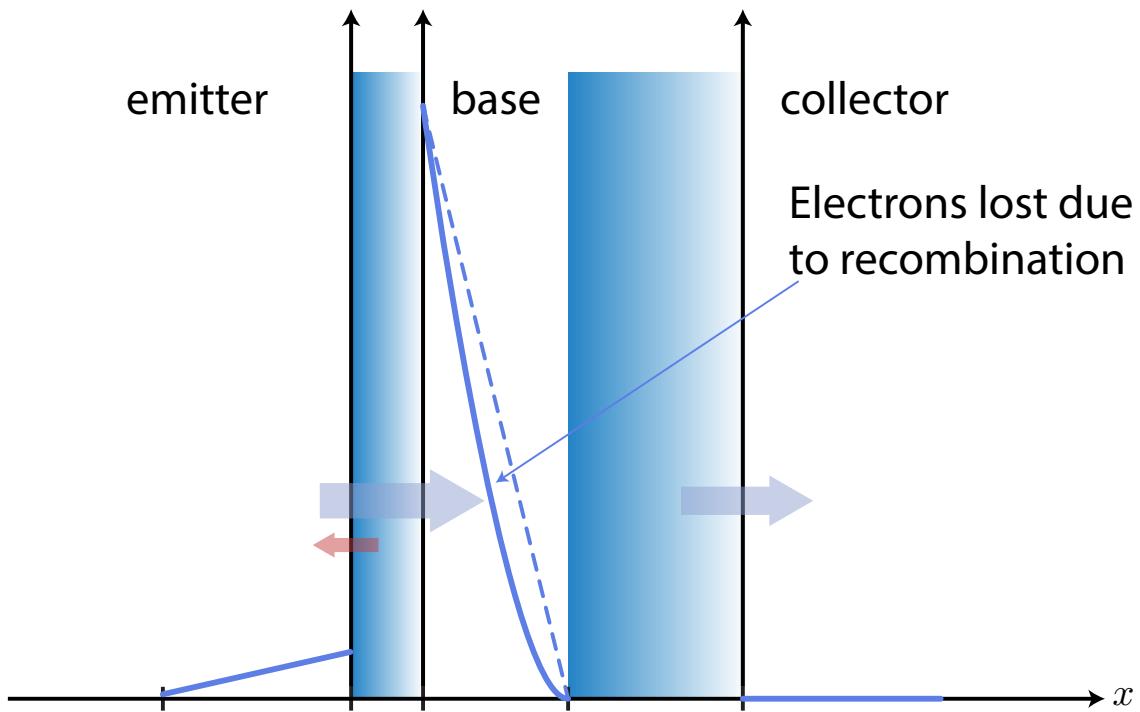


Figure 11.10: Minority carrier profile in an *NPN* transistor. In this figure we assume non-negligible recombination, resulting in a spatially decaying minority carrier profile in the base. The base current is increased as a result, leading to a lower α_F .

11.5 BJT Device Physics with Equations

11.5.1 Collector Current

If you spent a great deal of time understanding the *PN*-junction device physics, it's about to pay off, because all of the equations apply to the BJT. The forward-biased base-emitter junction results in a large diffusion of electrons across base:

$$J_{n_{diff}} = qD_n \frac{dn_p}{dx} = \left(\frac{qD_n n_{p,B_0}}{W_B} \right) e^{\frac{qV_{BE}}{kT}} \quad (11.12)$$

Let's define the **saturation current** as:

$I_S = \left(\frac{qD_n n_{p,B_0} A_E}{W_B} \right)$

Collector saturation current (11.13)

Eq. 11.13 allows us to write the collector current form we have already seen multiple times in this chapter:

$I_C = I_S e^{\frac{qV_{BE}}{kT}}$

Collector current (11.14)

11.5.2 Base Current

Because the base-emitter is forward-biased, there is also a diffusion of holes into the emitter:

$$J_{p_{diff}} = -qD_p \frac{dp_n}{dx} = \left(qD_p p_{n,E_0} W_E \right) \left(e^{\frac{qV_{BE}}{kT}} - 1 \right) \quad (11.15)$$

Because all the holes arise from the base region, the **base current** is responsible for this component:

$I_B = \left(\frac{qD_p p_{n,E_0} A_E}{W_E} \right) \left(e^{\frac{qV_{BE}}{kT}} - 1 \right)$

Base current (11.16)

11.5.3 Current Gain

With equations for the base and collector current, we can find the transistor **forward-active current gain** β_F :

$\beta_F = \frac{I_C}{I_B} = \frac{\left(\frac{qD_n n_{p,B_0} A_E}{W_B} \right)}{\left(\frac{qD_p p_{n,E_0} A_E}{W_E} \right)} = \left(\frac{D_n}{D_p} \right) \left(\frac{n_{p,B_0}}{p_{n,E_0}} \right) \left(\frac{W_E}{W_B} \right)$

Current gain (11.17)

Recall that we wish to maximize this term. We can relate the minority carrier concentrations in the base and emitter to the doping concentrations:

$$\left(\frac{n_{p,B_0}}{p_{n,E_0}} \right) = \frac{\frac{n_i^2}{N_{A,B}}}{\frac{n_i^2}{N_{D,E}}} = \frac{N_{D,E}}{N_{A,B}} \quad (11.18)$$

These equations tell us how to maximize the performance of the transistor. We must dope the emitter as high as possible relative to the base, and the base width should be made as small as possible. The physical reasons are clear and have already been discussed. To recap, a short base width results in a high fraction of minority carriers flowing from the emitter to the collector. A higher emitter doping results in more electron injection (diffusion) from the emitter side rather than hole injection (diffusion) from the base side, which also results in a relatively larger current I_C compared to I_B .

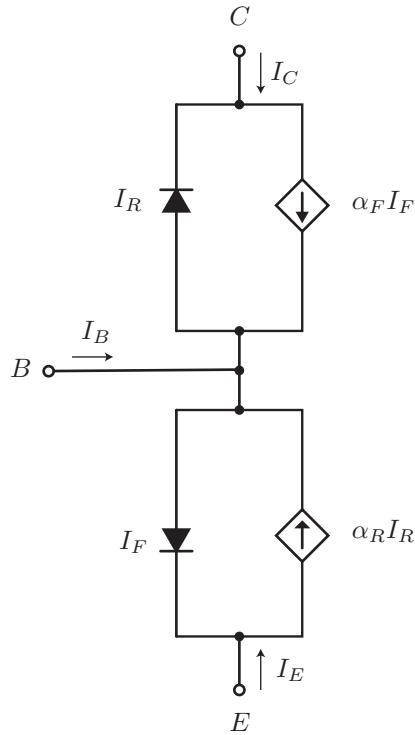


Figure 11.11: The large-signal Ebers-Moll circuit model of a BJT transistor.

11.6 Large Signal and Small-Signal Equivalent Circuits

11.6.1 Ebers-Moll Model

We can now use some symmetry arguments and include both reverse and forward-biased diode currents into the collector and emitter. Recall that α_F represents the fraction of emitter carriers that reach the collector. Likewise, when the transistor is biased in reverse active mode, α_R represents the fraction of injected carriers "collected" at the emitter side. Combining these equations results in the **Ebers-Moll equations**. The first equation is the emitter current:

$$I_E = -I_{ES} \left(e^{qV_{BE}/kT} - 1 \right) + \alpha_R I_{CS} \left(e^{qV_{BC}/kT} - 1 \right) \quad (11.19)$$

The first term on the RHS of Eq. 11.19 is the emitter diode current previously discussed, with the reverse minority carrier current included. The current is negative because the diode is forward biased. The second term is simply the reverse active current. α_R represents the fraction of carriers that reach the emitter (rather than recombining), and is only significant if V_{BC} is forward biased.

Under normal conditions, V_{BC} is negative and the second term contributes only a small amount of current. Likewise, for the collector, we have the following symmetric relations:

$$I_C = \alpha_f I_{ES} \left(e^{qV_{BE}/kT} - 1 \right) - I_{CS} \left(e^{qV_{BC}/kT} - 1 \right) \quad (11.20)$$

It can be shown that the forward active and reverse active regions have a reciprocity relationship that dictates:

$$\alpha_F I_{ES} = \alpha_I I_{CS} \quad (11.21)$$

11.6.2 Ebers-Moll Equivalent Circuit

The **Ebers-Moll model** is shown in Fig. 11.11. It is a physical model of the Ebers-Moll equations, which are large-signal relations between the BJT transistor junction voltages and the resulting currents.

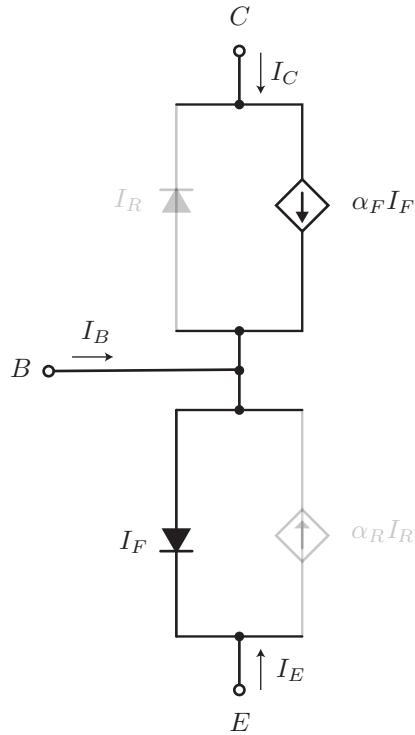


Figure 11.12: The large-signal Ebers Moll circuit model of a BJT transistor in the forward active region. The grayed out portions of the circuit can be neglected in the forward active region.

Note that each junction is represented by a diode: the forward diode I_F , and the reverse diode I_R . When current flows through I_F , a fraction $\alpha_F I_F$ flows into the collector by the action of the controlled current source. This is the normal forward active region. The base current is the difference current, and the reverse-biased diode current flows through I_R . Take note that the model is completely symmetric. This means that if we interchange the role of the collector and emitter, the same physics should apply, but the currents will be different.

11.6.3 Forward Active Region and the Early Effect

We can simplify the equivalent circuit model in the forward active region by noting that I_R carries very little current, since if the BC -junction is reverse-biased, I_R is very small and can be neglected. So we can practically neglect I_R and the controlled source that depends on it, resulting in the simpler model shown in Fig. 11.12. With this approximation, we can write the collector current as:

$$I_C = \alpha_F I_{ES} \left(e^{qV_{BE}/kT} - 1 \right) \quad (11.22)$$

In practice this equation is modified to account for the non-zero output conductance of the device:

$$I_C = \alpha_F I_{ES} \left(e^{qV_{BE}/kT} - 1 \right) \left(1 + \frac{V_{CE}}{V_A} \right) \quad (11.23)$$

In Eq. 11.19, V_A is known as the **Early Voltage**, as introduced in Section 8.5.5. In the MOSFET, increasing the drain-source voltage in saturation resulted in a larger depletion region width on the drain side, and this made the channel effectively shorter. In a BJT, we have a similar effect in that an increase in collector-emitter voltage results in a larger depletion region width on the base-collector junction. This shortens the base width, thereby increasing the current.

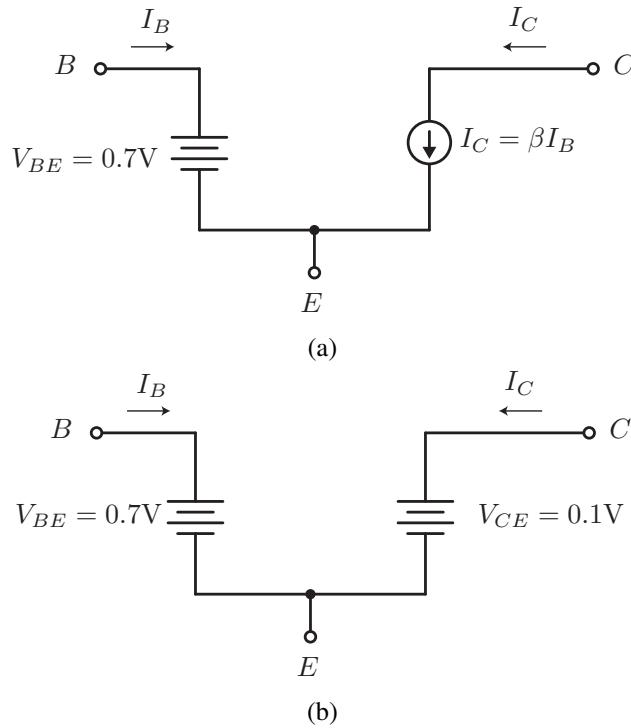


Figure 11.13: Extremely simple models for a BJT device in the (a) forward active and (b) saturation region.

11.6.4 Simplified Ebers-Moll

Even simpler Ebers-Moll models are shown in *Fig. 11.13*. The forward biased diode is treated as a battery with on-voltage of about 0.7V, and the collector current is simply an amplified version of the base current. The model degenerates into two voltage sources in saturation, as both junctions are forward biased, with $V_{CE,sat}$ given by the difference between the diode on-voltages, shown in *Fig. 11.13b*.

11.6.5 Small-Signal Model

The Ebers-Moll models are useful for DC calculations. For most AC circuits, we make use of the equivalent circuit model, known as the hybrid- π model, shown in *Fig. 11.14*. The model is very similar to the small-signal model of 3-terminal MOSFET, with one notable exception. There is a resistance R_π between the base and emitter because unlike a MOSFET, there's a current path for the base current.

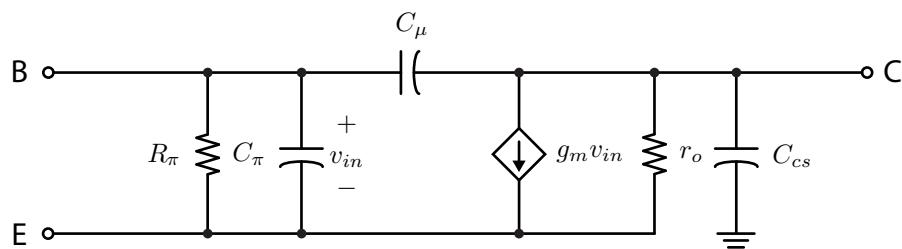


Figure 11.14: The Hybrid- π small-signal equivalent circuit model of a BJT transistor.

It is not very difficult to show that the transconductance of the transistor is given by:

$$g_m = \frac{\partial I_C}{\partial V_{BE}} \Big|_Q = \frac{qI_C}{kT} \quad (11.24)$$

The transconductance here is interesting compared to that of a MOSFET. For the MOSFET we showed (see *Eq. 9.28*):

$$g_{m,MOSFET} = \frac{2I_D}{V_{GS} - V_T} \quad (11.25)$$

For a fixed bias current $I_C = I_D$, the transconductance of a BJT is generally much higher than a MOSFET. This is because when biased in strong inversion, $V_{GS} - V_T \gg kT/q$. Recall that kT/q is about 26 mV at room temperature, and MOS devices in strong inversion are biased with several 100's of mV of overdrive. In fact, if we try to reduce $V_{GS} - V_T$ to boost the transconductance, we are in fact biasing a MOS device as a BJT! The trade-off is that the CMOS device is much slower in sub-threshold region. This is why it is important to understand BJT device physics, even in a CMOS world. The transistor **base resistance** R_π is given by:

$$R_\pi^{-1} = \frac{\partial I_B}{\partial V_{BE}} \Big|_Q = \left(\frac{1}{\beta_F} \right) \frac{\partial I_C}{\partial V_{BE}} \Big|_Q = \frac{g_m}{\beta_F} \quad (11.26)$$

$R_\pi = \frac{\beta_F}{g_m}$

BJT base resistance
(11.27)

This makes sense because the base current is β_F times smaller than the collector current. The resistor, r_o , models the device **output resistance**:

$$r_o^{-1} = \frac{\partial I_C}{\partial V_{CE}} \Big|_Q = \frac{I_C}{V_A} \quad (11.28)$$

$r_o = \frac{V_A}{I_C}$

BJT output resistance
(11.29)

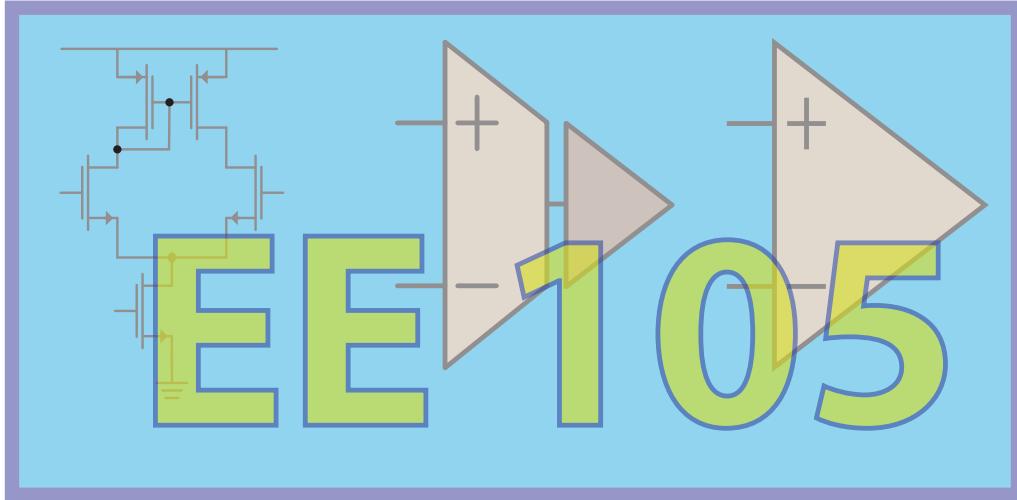
Since the base-emitter and base-collector are *PN*-junctions, there is an associated **junction capacitance**. In the forward active region, $C_\mu = C_{j,bc}(V_{BC})$ is the **base-collector reverse biased junction capacitance**. Likewise C_π contains the forward-biased junction capacitance, but also models the minority charge storage effects we discussed when we analyzed the diode, see Section 6.6.1:

$$C_\pi \approx 1.4C_{j,be0} + C_{diff} \quad (11.30)$$

In *Eq. 11.30*, the factor of 1.4 models the increase in junction capacitance under forward bias. The **diffusion capacitance** is given by a similar expression we presented for a diode:

$$C_{diff} = g_m \tau_F \quad (11.31)$$

The factor τ_F is a technology parameter known as the "**forward transit time**". It is essentially the time it takes a minority carrier to cross the various regions and junction. The physical reason for this is that charge is stored in the base due to the minority carrier diffusion profile in the base. If the base current is increased, it takes a finite amount of time to build a new diffusion profile, a time that corresponds to how long it takes for minority carriers to flow into the base.



12. Single-Stage Amplifiers

12.1 Chapter Preview

This chapter reviews the three important single-stage amplifier topologies. Understanding these single-stage amplifiers is crucial, as we will build on these amplifiers in the following chapters to build more complex and higher performance multi-stage amplifiers. Intimate knowledge of the input and output impedance, and the voltage/current gains is needed in order to move onto more complex amplifiers. Therefore, we will begin by reviewing the fundamentals of amplifiers using two-port models. Then we will show how each single-stage amplifier can be classified accordingly.

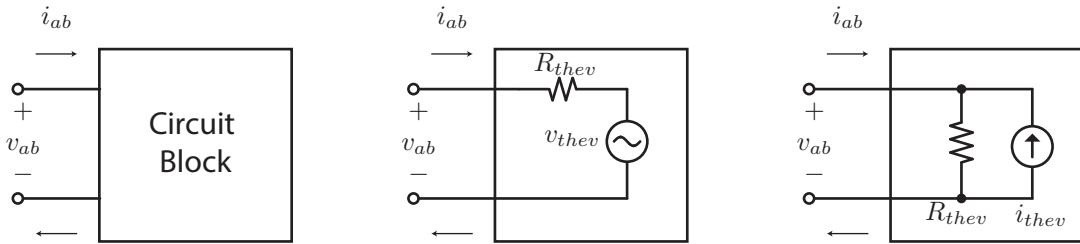


Figure 12.1: A black box one-port linear circuit can be represented as a Thévenin or Norton equivalent.

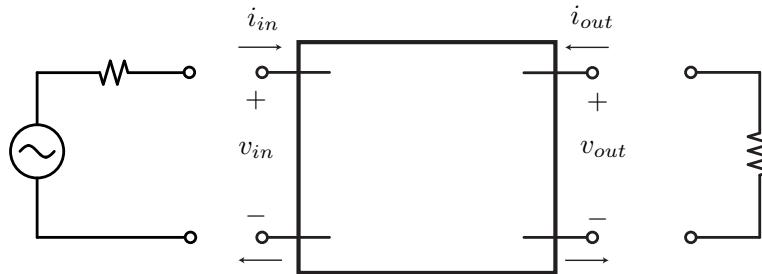


Figure 12.2: Similar to a one-port, a two-port linear circuit can be represented using an equivalent model. If the blackbox is an amplifier, we can choose among many different representations shown in Fig. 12.3 and Fig. 12.4.

12.2 Review Two-Port Amplifiers

12.2.1 One-Port Equivalent Models

Consider any **one-port** linear circuit as a black box with a pair of exposed terminals. We know from basic circuit theory that we can represent this black box as an equivalent **Thévenin or Norton model** (Fig. 12.1). This is also true of amplifiers, but amplifiers have two important differences. First, most amplifiers have an input and an output port, where a "port" is a pair of terminals. Often one terminal is shared, or "common", between the ports, such as the ground connection. The other important difference is that **two-port** amplifiers do not have any sources (of the independent variety) in them. Finally, we assume that an amplifier is a *unilateral* block, meaning the output depends on input but input is independent of output. In other words, signals travel in one direction, but not the other. This is why the schematic symbol for an amplifier is often drawn as a triangle (think of an arrow). In general, though, the output port depends linearly on the current and voltage at the input and output ports, because there is output impedance.

Why are amplifiers *unilateral*? Think about the small-signal model of a MOSFET or BJT. The output depends on the input (v_π for BJTs, or v_{gs} for FETs) through the transconductance, g_m . However, the input is more or less independent of the output, except for the effects of the small "overlap" or junction capacitance (C_μ for BJTs, or C_{gd} for FETs). As long as the "forward" gain g_m is much larger than the reverse, $j\omega C_{gd}$, then the amplifier is unilateral. At low frequencies, this is certainly true and a useful simplifying approximation.

12.2.2 Small-Signal Two-Port Models

A linear two-port box is shown in Fig. 12.2, under the unilateral assumption. This can be represented as a generic amplifier in four different ways, as shown in Fig. 12.3 and Fig. 12.4. Note that signals can be represented as voltage, or as current, or some combination of the two. It is also important to note that any real amplifier has an **input impedance** R_{in} , and an **output impedance** R_{out} . This is because both voltages *and* currents (power) are needed for the circuit to work.

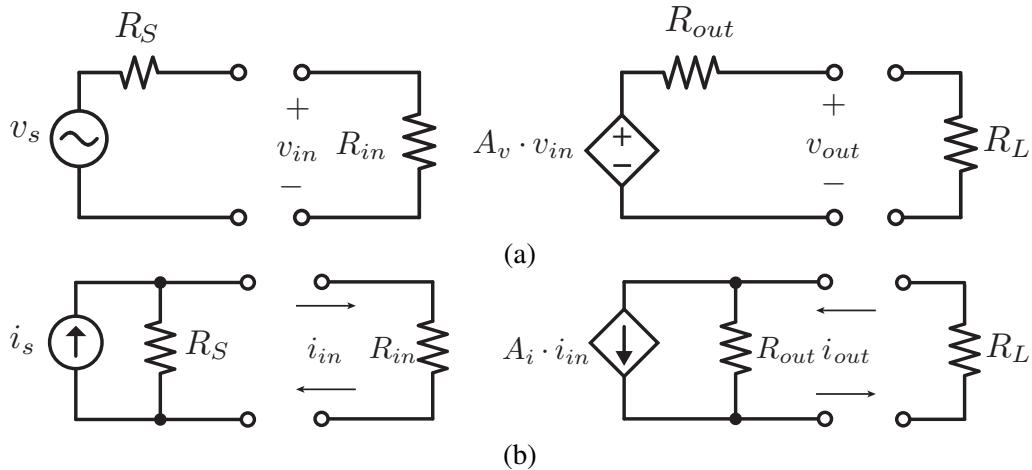


Figure 12.3: (a) A two-port voltage amplifier has internal gain A_v , input impedance R_{in} , and output impedance R_{out} . (b) A two-port current amplifier has internal gain A_i , input impedance R_{in} , and output impedance R_{out} .

Two-Port Small-Signal Voltage Amplifier

Let's begin with *Fig. 12.3a*, or a **voltage amplifier**. This is perhaps the most straightforward amplifier to understand, because both the input and output signals are voltages. Recall that an ideal voltage amplifier has a very large input impedance R_{in} so that it does not "load" the driver circuitry. Then most of the applied voltage will appear across R_{in} rather than the source impedance. Likewise, for similar reasons, the voltage amplifier should have a very small output impedance R_{out} .

Two-Port Small-Signal Current Amplifier

The dual of a voltage amplifier is the **current amplifier**, shown in *Fig. 12.3b*. Now the role of voltage and current are interchanged. The input current into the amplifier is amplified and divided between the load and the internal output impedance R_{out} . For this reason, we would like a current amplifier to have as large of an output impedance R_{out} as possible, in order to maximize the current gain. Also, a current amplifier with a low input impedance can "steal" all the current from a source since a current divider favors the path of highest conductance.

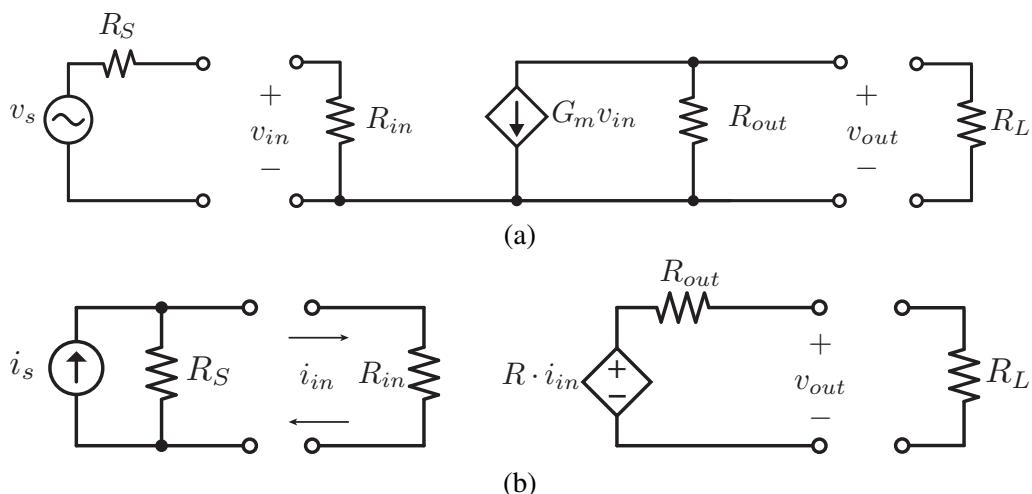


Figure 12.4: (a) A two-port transconductance amplifier has internal gain G_m , input impedance R_{in} , and output impedance R_{out} . (b) A two-port trans-resistance amplifier has internal gain R , input impedance R_{in} , and output impedance R_{out} .

Two-Port Small-Signal Trans-conductance Amplifier

The **trans-conductance amplifier** shown in *Fig. 12.4a* takes an input voltage and generates an output current, just like a conductor. But the signals are at different ports, making it a "trans" conductor, effectively *transporting* the signal from the voltage domain to the current domain. Take a moment to consider that the concept of voltage and current domain only make sense with respect to the source and load impedance. A *good voltage source has a low output impedance with respect to the input impedance of the load*, and a *good current source has a high output impedance with respect to the input impedance of the load*. Then it follows that a transconductance amplifier is effective only if it presents a high impedance, ideally infinite, and generates an output current with high output impedance. A MOS transistor biased in saturation (and a BJT biased in the forward-active region) is almost an ideal transconductor, because its input is a high impedance at low frequencies, and its output is nearly open, due to the large output impedance of a transistor in saturation. Notice that the most general transconductance model has 4 separate terminals, but we've drawn a version with a common node, similar to the transistor hybrid-pi model.

Two-Port Small-Signal Trans-resistance Amplifier

The **trans-resistance amplifier** is an amplifier that converts an input current into a voltage, as shown in *Fig. 12.4b*. It is the dual of the trans-conductance amplifier. It ideally presents a low input impedance at the input, and drives the output with a low impedance voltage. Transresistance amplifiers are also known as **trans-impedance amplifiers**. They are important in fiber optic communication, where the input light signal is a current (proportional to photon flux) that is converted to a voltage by a TIA. Also, many sensors naturally generate a current signal (they have high source impedance), and need to be converted into voltage signals to drive other circuitry, such as **Analog-to-Digital converters**, which expect to be driven by a relatively low source impedance.

12.2.3 Input Impedance Z_{in}

To calculate the input and output impedances of two-ports, we can simply use a test source at the input or output, in order to determine the equivalent load at each port. This is shown schematically in *Fig. 12.5*, where the test sources are labeled as v_x and i_x . When probing the input port of a truly unilateral two-port, the output port plays no role in determining the input impedance. It could be left open-circuited, short-circuited, or terminated in an arbitrary impedance Z . However, in practice circuits are not perfectly unilateral, and so we should have a convention for what to connect to the second output port. It is common to measure the input port impedance while terminating the output port with the load impedance attached, as shown in *Fig. 12.5*. For example, if we excite the input with a test current (voltage) source, and then observe the input voltage (current), we have:

$$Z_{in} = \frac{v_x}{i_x} \Big|_{\begin{array}{l} Z_S \text{ removed,} \\ Z_L \text{ attached} \end{array}} \quad \text{Two-port input impedance using test source} \quad (12.1)$$

12.2.4 Output Impedance Z_{out}

We measure the output impedance in the same manner as the input impedance. Again, if the amplifier were truly unilateral, connecting a source at the output would not affect the input port. However, due to parasitic leakage (say through capacitance), there is a small influence. By convention, the *input* port is terminated with the *source* resistance, and the *output* port is measured, as shown in *Fig. 12.6*. For example, we can excite with a test voltage (current) source, and then measure the output current (voltage):

$$Z_{out} = \frac{v_x}{i_x} \Big|_{\begin{array}{l} Z_L \text{ removed,} \\ Z_S \text{ attached} \end{array}} \quad \text{Two-port output impedance using test source} \quad (12.2)$$

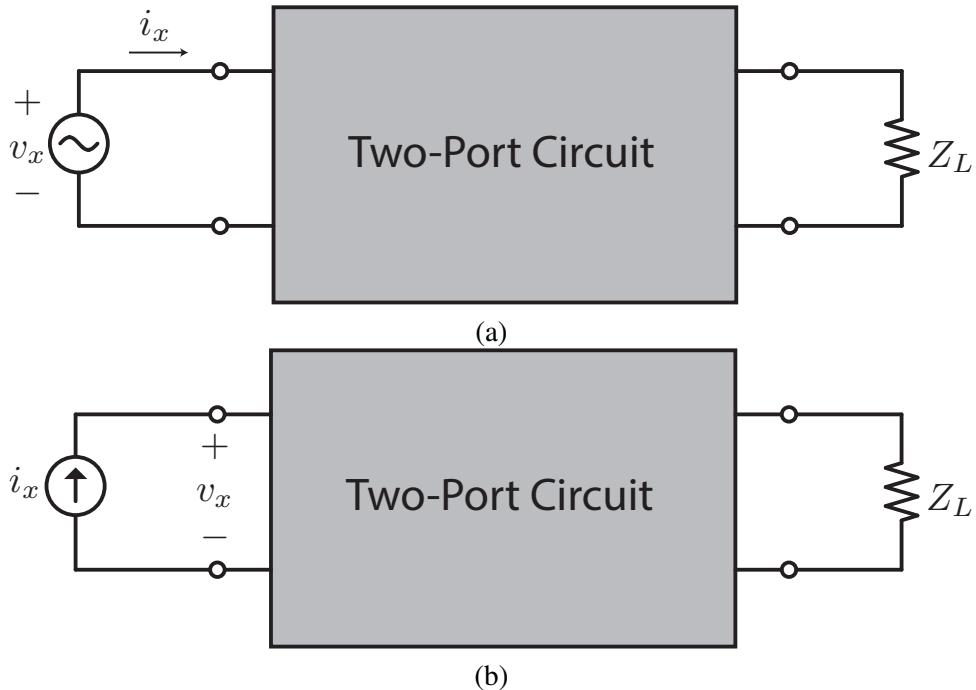


Figure 12.5: To calculate the *input impedance* of a two-port, we apply a test (a) voltage or (b) current source and measure the current/voltage. By convention, we leave the *output resistance* connected to the load.

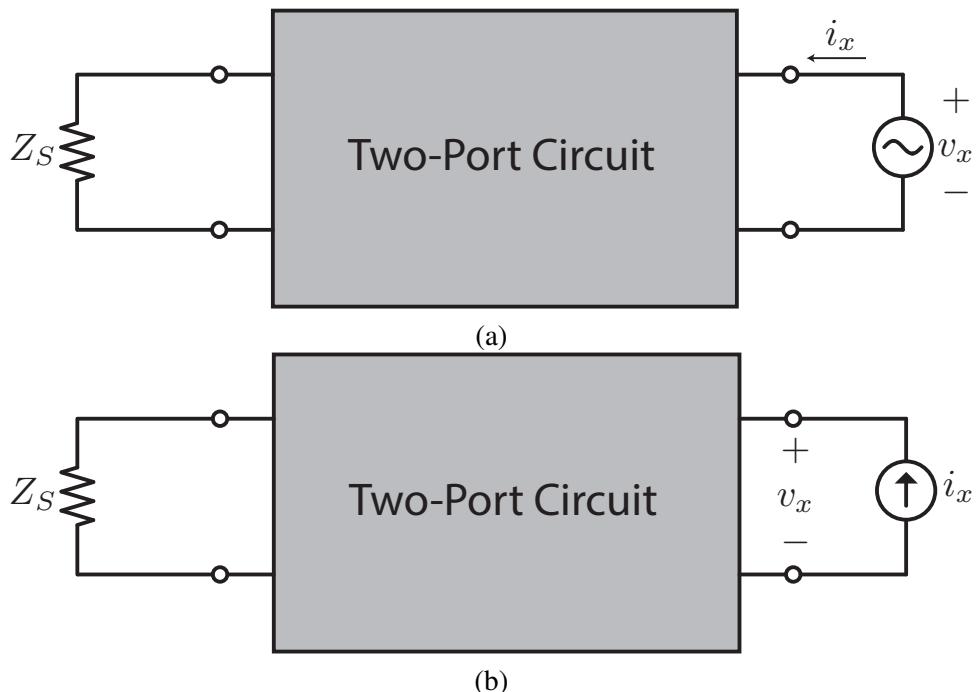


Figure 12.6: To calculate the *output impedance* of a two-port, we apply a test (a) voltage or (b) current source and measure the current/voltage. By convention, we leave the *source resistance* connected to the input.

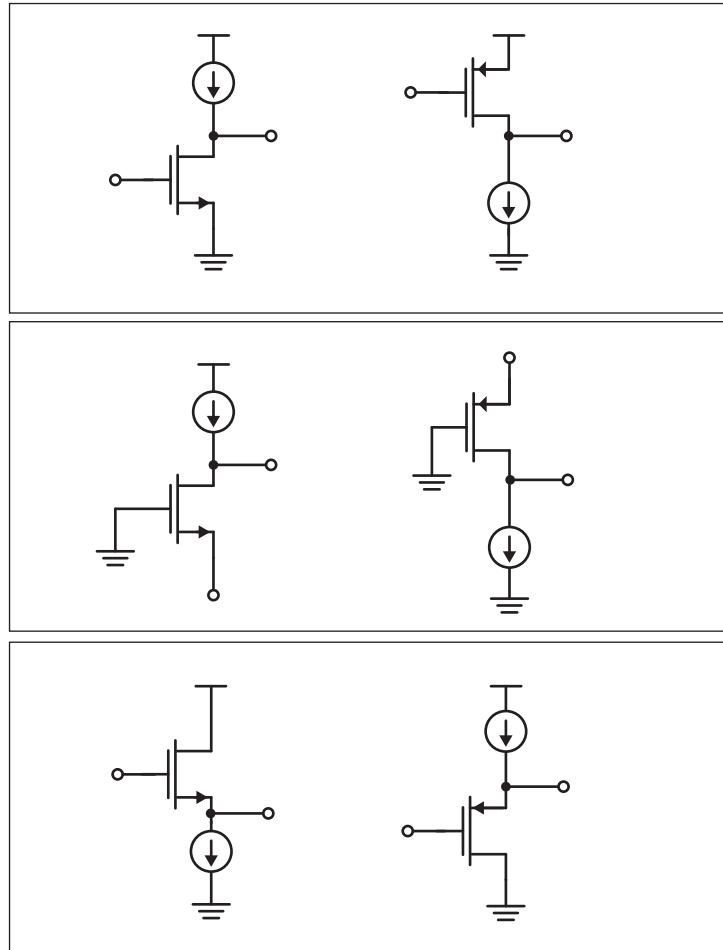


Figure 12.7: Family of NMOS and PMOS single-stage amplifiers. Listing from the top: the Common Source, Common-Gate, and Common Drain Amplifiers. As shown, each is biased with an ideal current source, while in practice we use transistor based current mirrors (see Ch. 13) or resistors to bias the amplifiers.

12.3 Single-Stage MOS Amplifier Family

Because a transistor is a three-terminal device (ignoring the body terminal of a MOSFET), and an amplifier is a four-terminal two-port, constructing an amplifier out of a transistor requires one terminal to be in "common" with both ports. This leads to the family of amplifiers shown in *Fig. 12.7*. The **common source** (CS) amplifier was already studied extensively in the previous chapters on small-signal modeling. We now introduce the **common gate** (CG) amplifier, and the **common drain** (CD) amplifier. The BJT equivalents are the **common emitter** (CE), **common base** (CB), and the **common collector** (CC). In this chapter we will focus on the MOS amplifiers, but with relatively small changes we can apply our knowledge to BJT amplifiers as well. While this is true of the small-signal models, the large-signal behavior, such as the operating point, requires a very different approach.

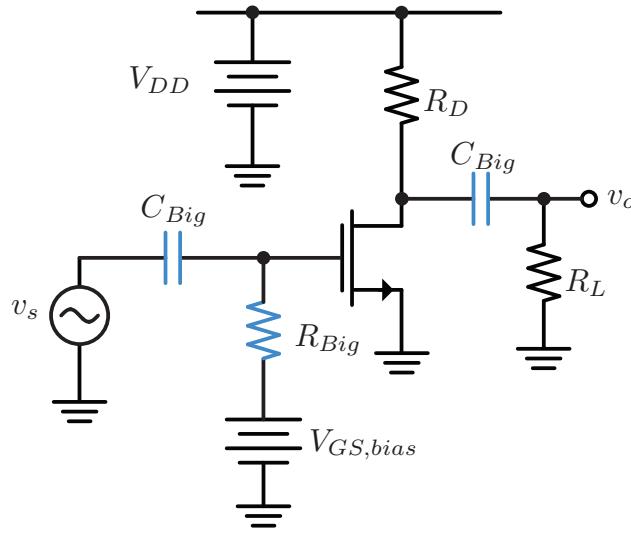
12.3.1 Isolating the Bias Points

When hooking up a source or load to an amplifier, it is easy to change the DC operating point. One solution, adopted with discrete amplifiers, is to use a large "**AC coupling**" capacitor at the input and/or output to isolate the load, as shown in *Fig. 12.8a*. Think of a very large capacitor as a battery. *At AC frequencies, it acts like a short circuit*. One way to emphasize these capacitors as very large is to label them with C_∞ . This indicates that they are ideally infinitely large capacitors, so that

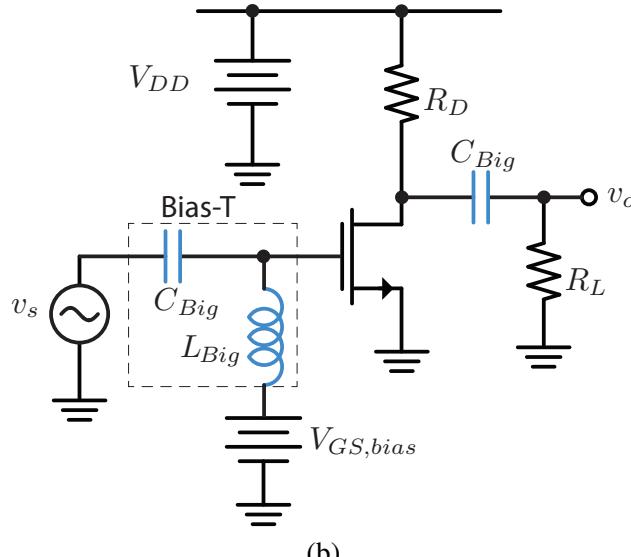
they act as short circuits for any non-zero frequency. Often we also need to *block AC signals* from getting shorted out by bias voltages. One solution adopted (shown in the figure) is a *large resistor*. The resistors are made large enough to avoid loading the amplifier.

"Chokes"

In place of large resistors, infinitely large inductors, or "**chokes**" as they are sometimes called, are also convenient. They work at DC, but at any non-zero frequency they are essentially open circuits. Chokes are also used to *isolate the DC supply or reference voltages* of sensitive circuits from the rest of the (noisy) system, as shown in Fig. 12.8b.



(a)



(b)

Figure 12.8: (a) AC coupling capacitors are large (ideally infinite) and act like batteries, allowing different points in a circuit to be biased to different DC values, effectively blocking DC but allowing AC signals to flow. A large resistor is used to "block" the AC signals without loading the circuit. (b) Large biasing inductors, sometimes called "chokes" are used to apply a DC voltage to a node while acting like an open-circuit at AC frequencies. A combination of an inductor and capacitor forms a bias-T.

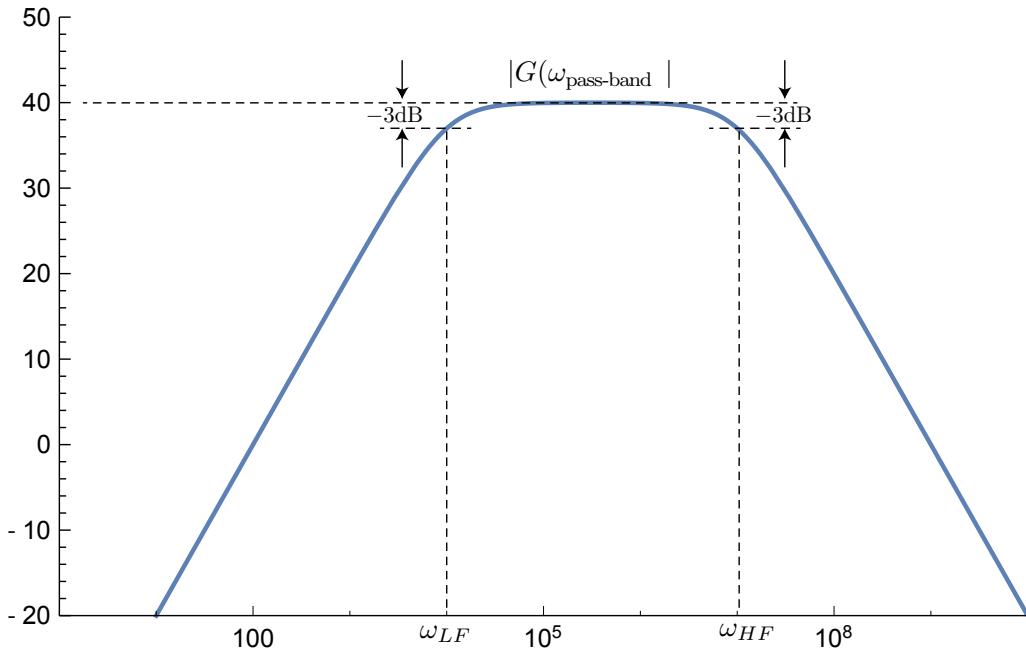


Figure 12.9: Definition of the "mid-band" or "pass-band" of an amplifier occurs at AC frequencies where large capacitors act like shorts (and large inductors are open).

12.3.2 DC Coupled vs AC Coupled Amplifiers

AC coupled amplifiers have no gain at DC because the capacitors are "open" at very low frequencies, or alternatively the chokes short out the signal at DC. The **mid-band** region of an amplifier is shown in *Fig. 12.9*. It is defined as the region where the gain is approximately flat before rolling off due to high frequency effects, and after recovering from the effect of large capacitors. The general rule is to replace large capacitors with shorts, and small capacitors (transistor parasitics for example) as opens when finding the **mid-band gain**.

12.3.3 Common Gate (CG) Amplifier

The common gate amplifier is shown in *Fig. 12.10*. The current source I_Q and voltage source $V_{bias,Q}$ are DC supplies used to *bias the amplifier*. The signal i_s and resistor R_s form the **AC signal source**, and they are isolated by a large capacitor. Likewise, the load R_L is isolated by another capacitor. Notice that without these capacitors, the DC points at the source and drain would shift. The DC operating point of the amplifier is simply given by:

$$I_Q = I_{DS} \quad \text{Common-gate DC operating point} \quad (12.3)$$

Since $V_{bias,Q}$ does not appear in the equation, you may be wondering why we don't simply ground the gate, as the circuit functionality is unaffected. For an ideal current source, operating the gate at ground would bring the source node to a negative voltage. We will learn in the next chapter that real current sources will cease to function under such conditions. The gate bias voltage is there to provide some "**headroom**" to the current source I_Q .

12.3.4 Common Gate AC Model

The **AC equivalent** circuit model is shown in *Fig. 12.11a*, which is obtained from *Fig. 12.10* by simply shorting out all the DC sources. To analyze the circuit, we can replace the transistor with its hybrid-pi equivalent circuit model, shown in *Fig. 12.11b*.

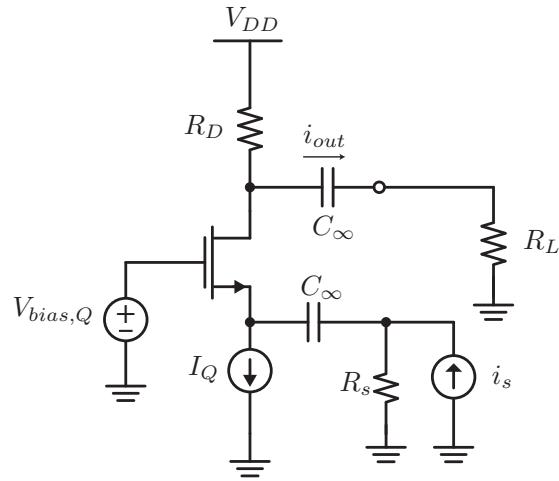


Figure 12.10: A common-gate amplifier driven by a source i_s with source resistance R_s . Capacitors are used to block the source/load DC levels from the amplifier, which is biased through I_Q and $V_{bias,Q}$.

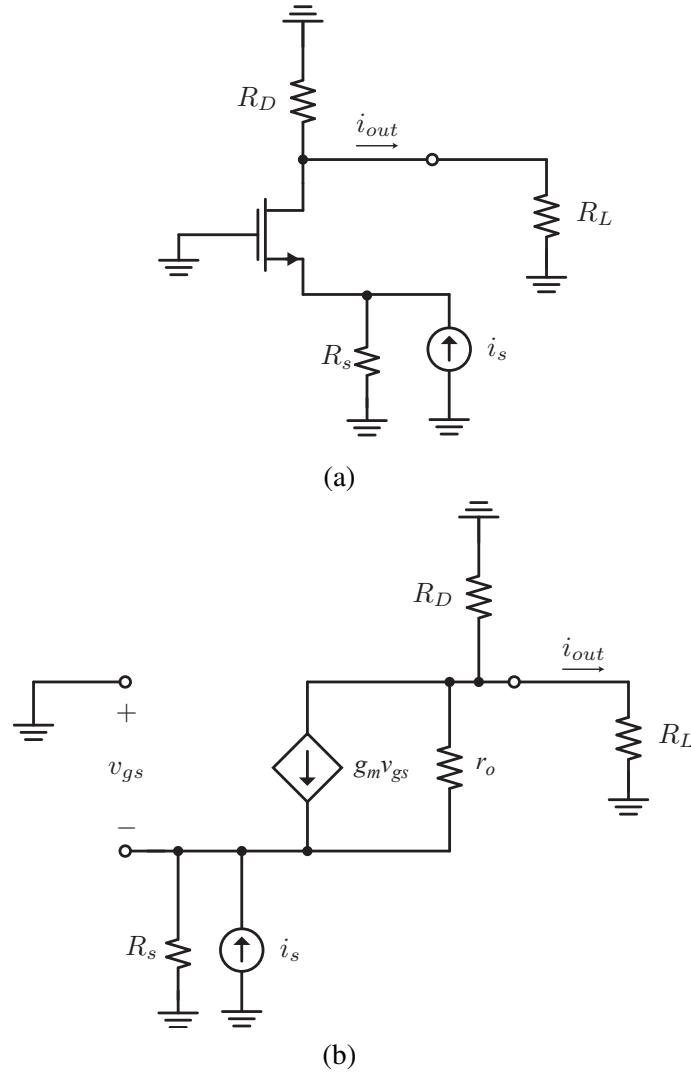


Figure 12.11: (a) The AC equivalent schematic of the common-gate amplifier. (b) The small-signal equivalent circuit of the common-gate amplifier.

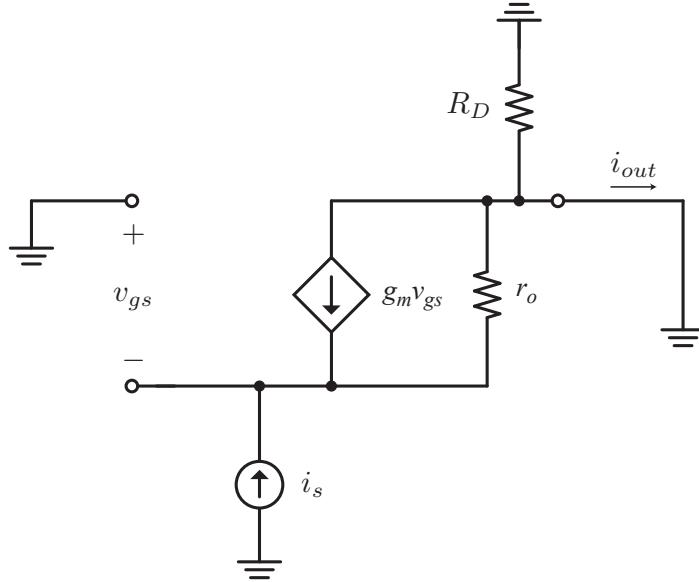


Figure 12.12: To find the intrinsic current gain of a common-gate amplifier, we short the output and drive the input with an ideal current source. The current gain is the current flowing out \$i_{out}\$ divided by \$i_s\$.

12.3.5 CG as a Current Amplifier: Find \$A_i\$

We will first analyze the circuit as a current amplifier. To find the current gain of a current amplifier, we should short-circuit the output and observe the output current (flowing into the short) when driven by an ideal current source. This means we drive it with a current source with zero source conductance, as shown in *Fig. 12.12*. Notice that the output current is the drain. Also notice that the source current must flow through the output because that is the only path for the current to return:

$$i_{out} = i_d = -i_s \quad (12.4)$$

From these observations, it is obvious that the **intrinsic current gain** is given by:

$A_i = -\left(\frac{i_{out}}{i_s}\right) = -1$

Intrinsic current gain for CG amplifier (12.5)

12.3.6 CG Input Impedance

Next we calculate the input impedance of the CG amplifier. By convention, we leave the load resistance connected at the output, and we apply a test source. We can use a test voltage \$v_x\$ or a current. In *Fig. 12.13*, we apply a test voltage and observe the current \$i_x\$. The test current splits between the transistor \$r_o\$ and the \$g_m\$ generator:

$$i_x = -g_m v_{gs} + \left(\frac{v_x - v_{out}}{r_o}\right) \quad (12.6)$$

At the output of the amplifier, the current flows into the load resistor \$R_L\$ in parallel with \$R_D\$:

$$v_{out} = -i_d (R_D \parallel R_L) = i_x (R_D \parallel R_L) \quad (12.7)$$

Also notice that the voltage \$v_{gs}\$ is in fact just the negative of the test source. With these observations, we substitute into *Eq. 12.6* to obtain:

$$i_x = g_m v_x + \left(\frac{v_x - i_x (R_D \parallel R_L)}{r_o}\right) = v_x \left(g_m + \frac{1}{r_o}\right) - i_x (R_D \parallel R_L) \quad (12.8)$$

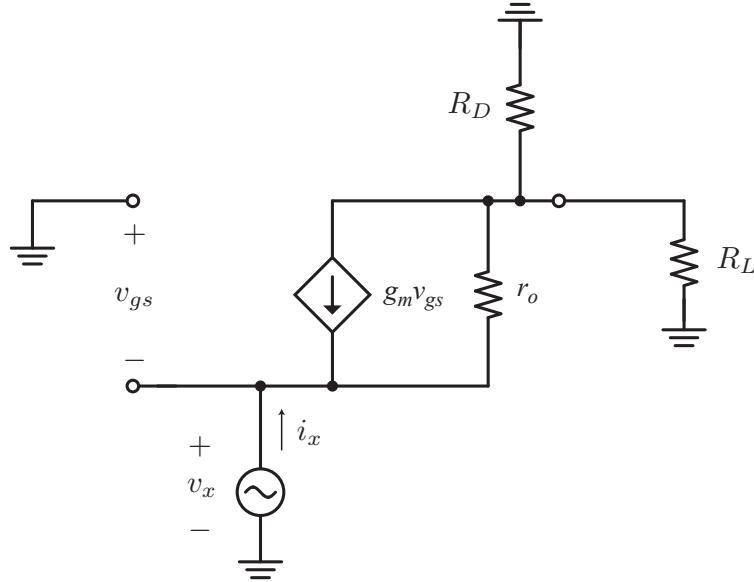


Figure 12.13: To determine the input impedance of a common-gate amplifier, we drive the input with a test source v_x and measure the current i_x flowing into the amplifier. By convention, we leave the load resistance connected at the output.

Rearranging Eq. 12.8:

$$i_x \left(1 - \frac{R_D \| R_L}{r_o} \right) = v_x \left(g_m + \frac{1}{r_o} \right) \quad (12.9)$$

From Eq. 12.9, we can now find the **input impedance**:

$$R_{in} = \frac{v_x}{i_x} = \frac{1 + \frac{R_D \| R_L}{r_o}}{g_m + \frac{1}{r_o}}$$

Input impedance for CG amplifier
(12.10)

The reciprocal of Eq. 12.10 is the **input conductance**, taking the form:

$$G_{in} = \frac{1}{R_{in}} = \frac{i_x}{v_x} = \frac{g_m + \frac{1}{r_o}}{1 + \frac{R_D \| R_L}{r_o}}$$

Input conductance for CG amplifier
(12.11)

Notice that the input impedance depends on the load, an undesirable result. But as we shall see, the dependence on the load is rather weak. Let's make some approximations to simplify the results. First of all, a good transistor has good intrinsic voltage gain A_0 :

$$A_0 = g_m r_o \gg 1 \quad (12.12)$$

If we factor g_m out from Eq. 12.11:

$$G_{in} = g_m \frac{1 + \frac{1}{g_m r_o}}{1 + \frac{R_D \| R_L}{r_o}} \quad (12.13)$$

Next let's assume that $r_o \gg R_D \| R_L$, which allows us to simplify to:

$$R_{in} = \frac{1}{g_m}$$

Input impedance for CG amplifier, $r_o \gg R_D \| R_L$
(12.14)

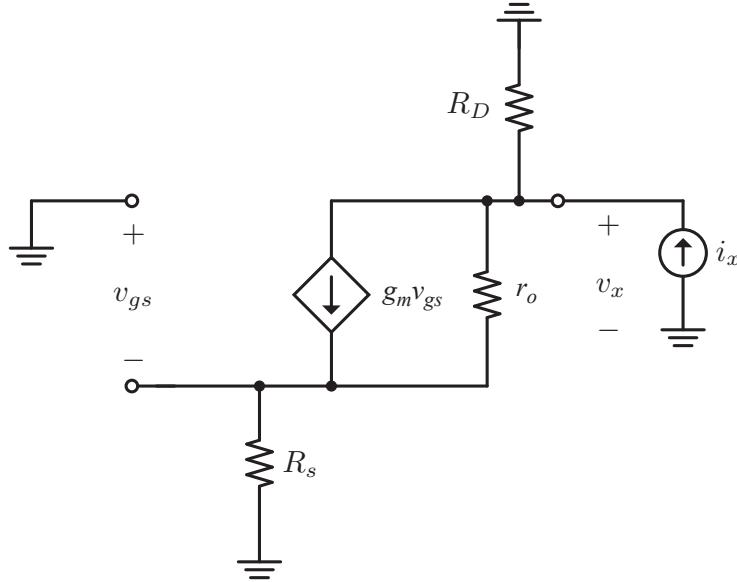


Figure 12.14: To determine the output impedance of a common-gate amplifier, we drive the output with a test source i_x and measure the voltage v_x . By convention, we leave the source resistance connected at the input.

Does this result make sense? After some careful deliberation, it is actually somewhat intuitive, at least in the limit that we have obtained in the approximation. To see this, ignore the output node and assume it is grounded (drain is grounded). Modulating the source voltage modulates the v_{sg} of the transistor. This modulation causes the current to increase by $-g_m v_{sg}$, resulting in an input conductance of g_m . In the absence of R_D and R_L , the only correction to this would be to account for r_o . Because r_o is simply in parallel, the input conductance would be $g_m + 1/r_o$. But the presence of the load $R_D \parallel R_L$ complicates things a bit. In practice we usually satisfy the condition that $R_D \parallel R_L \ll r_o$, so the simple result is enough.

12.3.7 CG Output Impedance

To find the CG amplifier output impedance, let's connect a test current source i_x at the output and observe the voltage v_x , as shown in Fig. 12.14. Write KCL at the input of the amplifier:

$$\frac{v_s}{R_s} - g_m v_{gs} + \frac{v_s - v_x}{r_o} = 0 \quad (12.15)$$

Let's make a simplifying step and eliminate R_D for a moment, because it will simply load the output in an obvious way (it is in parallel with the transistor effective output impedance). In this case, notice that the voltage at the source is $v_s = i_x R_s$. This is because i_x would flow entirely through R_s without R_D . Also, $v_{gs} = -v_s$, so the Eq. 12.15 can be simplified to:

$$v_s \left(\frac{1}{R_s} + g_m + \frac{1}{r_o} \right) = \frac{v_x}{r_o} = i_x R_s \left(\frac{1}{R_s} + g_m + \frac{1}{r_o} \right) \quad (12.16)$$

We can rearrange Eq. 12.16, which implies that the effective **output impedance** looking into the drain of the amplifier, in the presence of R_s , is boosted to::

$$R_{out,no\,R_D} = \frac{v_x}{i_x} = R_s \left(\frac{r_o}{R_s} + g_m r_o + 1 \right)$$

Output impedance for CG amplifier (12.17)

Bringing back R_D :

$$R_{out} = R_D \parallel (r_o + g_m r_o R_s + R_s)$$

Output impedance for CG amplifier with R_D (12.18)

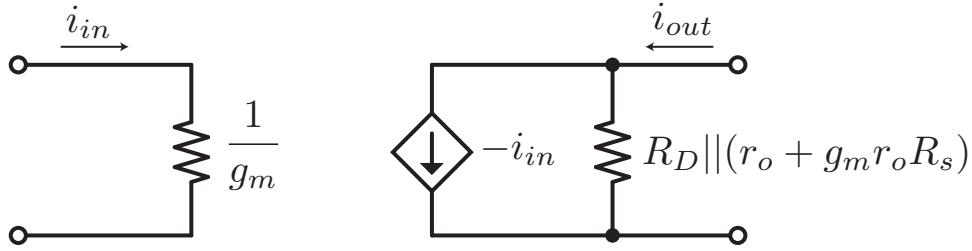


Figure 12.15: A two-port model for the CG amplifier is a current buffer with input impedance $1/g_m$ and output impedance as shown.

This is a very interesting result, and not so obvious. Adding a resistor R_S in series with the transistor on the drain side would simply increase the resistance to $r_o + R_S$. On the other hand, adding such a resistor on the source side increased the output impedance by a factor of $1 + g_m R_S$, potentially a large factor. To understand why this happens, we must use feedback theory, something that you will learn in the future.

The exact result is complicated, and a simpler version is easier to remember and nearly as accurate:

$$R_{out} \cong R_D \parallel (r_o + g_m r_o R_S + R_S) \approx R_D \parallel \left(1 + g_m R_S + \frac{R_S}{r_o} \right) r_o \quad (12.19)$$

An even simpler version can be obtained by assuming the source resistance is less than r_o :

$$R_{out} \approx R_D \parallel (r_o + g_m r_o R_S) = R_D \parallel (r_o(1 + g_m R_S)) \quad (12.20)$$

12.3.8 CG Two-Port Model

We can summarize everything we have learned about the CG amplifier thus far by building an equivalent model of the CG amplifier, shown in *Fig. 12.15*. This is a current amplifier model with a current gain of unity, which means it is a "**current buffer**". A current buffer needs to meet three requirements:

1. It must have a *current gain of unity*.
2. It should present a *low input impedance*.
3. It should present a *high output impedance*.

Under these three conditions, the output current into a load is a faithful reproduction of the input. It is extremely important to understand the role of low input impedance and high output impedance in this functionality. A CS amplifier is *not* a good current buffer, because it fails the first two criteria. It is *not even a good current amplifier*, because it does not satisfy the second criteria at low frequencies. Although at sufficiently high frequencies, we can think of it as a current amplifier. If this idea of a current buffer is not making sense, take some time and calculate the current gain of a *two-port current amplifier* under arbitrary load and source impedance. Then find the conditions for the externally observed current gain to match the internal current gain of the amplifier.

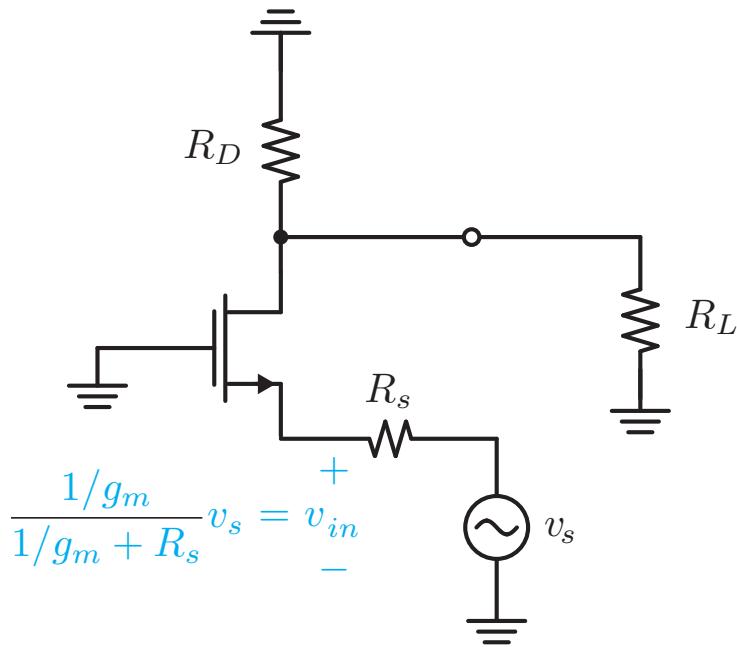


Figure 12.16: When the common-gate amplifier is voltage driven, the input voltage source is divided between its source resistance and the input of the amplifier, reducing the gain.

12.3.9 Common-Gate as a "Voltage Amplifier"

If we voltage drive the CG amplifier, as shown in *Fig. 12.16*, with source resistance $R_S = 0\Omega$, then the g_m is the same as a CS, and the gain is also just as good as a CS amplifier (with opposite phase):

$$v_{gs} = -v_s$$

$$\begin{aligned} v_{out} &= -i_d(R_L \parallel R_D) \\ &= -(-g_m v_s)(R_L \parallel R_D) && i_d = -g_m v_s \text{ at the output} \\ &= g_m v_s (R_L \parallel R_D) \end{aligned}$$

Now we have the CG gain, when used as a voltage amplifier:

$$A_v = \frac{v_{out}}{v_s} = g_m (R_L \parallel R_D) \quad \text{Voltage gain, CG voltage amplifier} \quad (12.21)$$

But if we have a source resistance such that $R_S \neq 0\Omega$, then the low input impedance causes a voltage divider effect:

$$v_{gs} = -v_s \frac{Z_{in}}{Z_{in} + R_s} \quad (12.22)$$

$$= -v_s \frac{1/g_m}{1/g_m + R_s} \quad \text{Recall Eq. 12.14 for why } Z_{in} = 1/g_m \quad (12.23)$$

$$= -\frac{v_s}{1 + g_m R_s} \quad (12.24)$$

This results in a lower transconductance:

$$v_{out} = -i_d(R_L \parallel R_D) \quad (12.25)$$

$$= -(v_{gs} g_m + \frac{1}{r_0}) \underset{\approx 0}{\cancel{(R_L \parallel R_D)}} \quad r_o \gg R_D \parallel R_L \quad (12.26)$$

$$= -\frac{-v_s g_m}{1 + g_m R_s} (R_L \parallel R_D) \quad \text{Substitution from Eq. 12.24} \quad (12.27)$$

$$\boxed{= \frac{v_s g_m}{1 + g_m R_s} (R_L \parallel R_D)} \quad \text{Transconductance, CG voltage amplifier} \quad (12.28)$$

And consequently a lower voltage gain:

$$\boxed{A_v = \frac{v_{out}}{v_s} = \frac{g_m}{1 + g_m R_s} (R_L \parallel R_D)} \quad \text{Gain, CD voltage amplifier} \quad (12.29)$$

This analysis is important because it shows that a CG amplifier can be interpreted in different ways. It can not only be used as a current buffer, but it can also provide voltage gain and be viewed as a voltage amplifier. You just have to remember that its input impedance is low, and so it will likely load the source.

In some applications, such as **RF amplifiers**, this is desirable, because for power matching we desire the load and the source impedance to match (to extract the maximum power from the source). The low input impedance of the CG is a convenient match to 75Ω or 50Ω **transmission-line impedances**.

There is another benefit of the CG amplifier as a voltage amplifier, compared to the CS amplifier in particular. The benefit is that it has a positive gain, as opposed to the inverting amplification of the CS stage. In some applications, we need positive gain, or we want to combine positive gain with negative gain. One application where we need positive gain is in a differential amplifier, and so the CG amplifier will be useful.

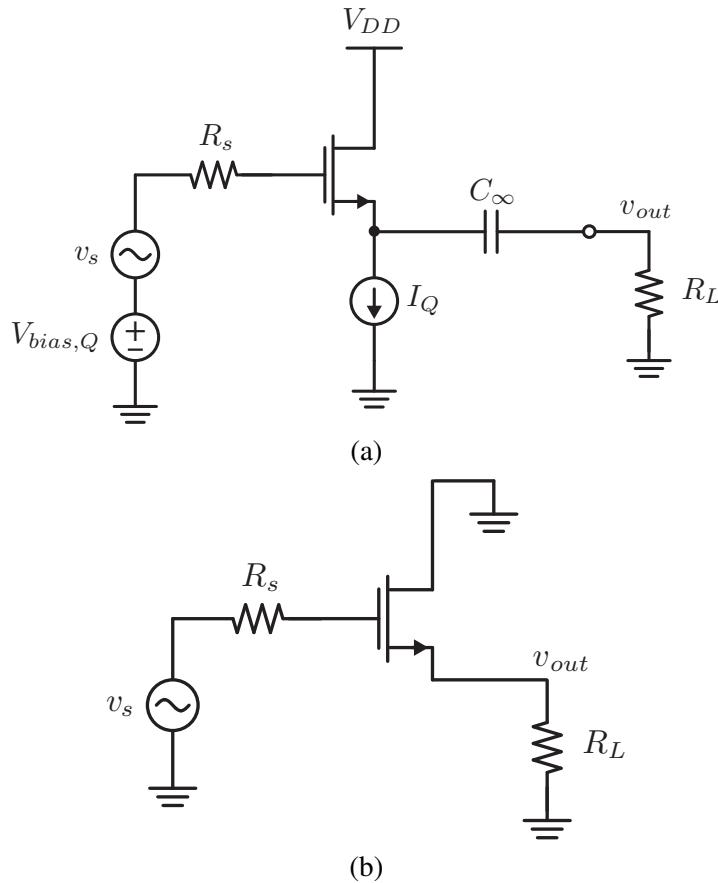


Figure 12.17: (a) The full schematic of the common drain (CD) amplifier. (b) The AC equivalent circuit of a CD amplifier.

12.4 Common Drain Amplifier

The full schematic of the common drain amplifier is shown in *Fig. 12.17a*. The "common" terminal is the drain, because the signal input comes into the gate, and the signal output exits the source. However, unlike the CG and CS amplifiers, the drain is not grounded. Instead it is connected to a DC supply. This is because we need to connect the drain to a higher potential than the gate/source in order to bias the transistor in saturation. As far as AC signals are concerned, though, the drain voltage is DC and therefore **AC grounded**. This is shown in *Fig. 12.17b*.

12.4.1 CD DC Operating Point

The CD amplifier is usually biased with a constant current I_Q as shown. The load is **AC coupled**, and so the transistor I_{DS} flows into I_Q :

$$I_Q = I_{DS} = \left(\frac{W}{2L} \right) \mu C_{ox} (V_{GS} - V_T)^2$$

In theory, the gate voltage could be at any DC bias point, but in practice we know that the current I_Q dictates a positive V_{GS} :

$$V_{GS} = V_T + \sqrt{\frac{2I_Q}{\mu C_{ox} \frac{W}{L}}}$$

To give the current source I_Q some headroom (DC voltage), we need to bias the gate of the transistor at a correspondingly higher voltage. For now let's call that $V_{bias,Q}$ similar to the CG amplifier.

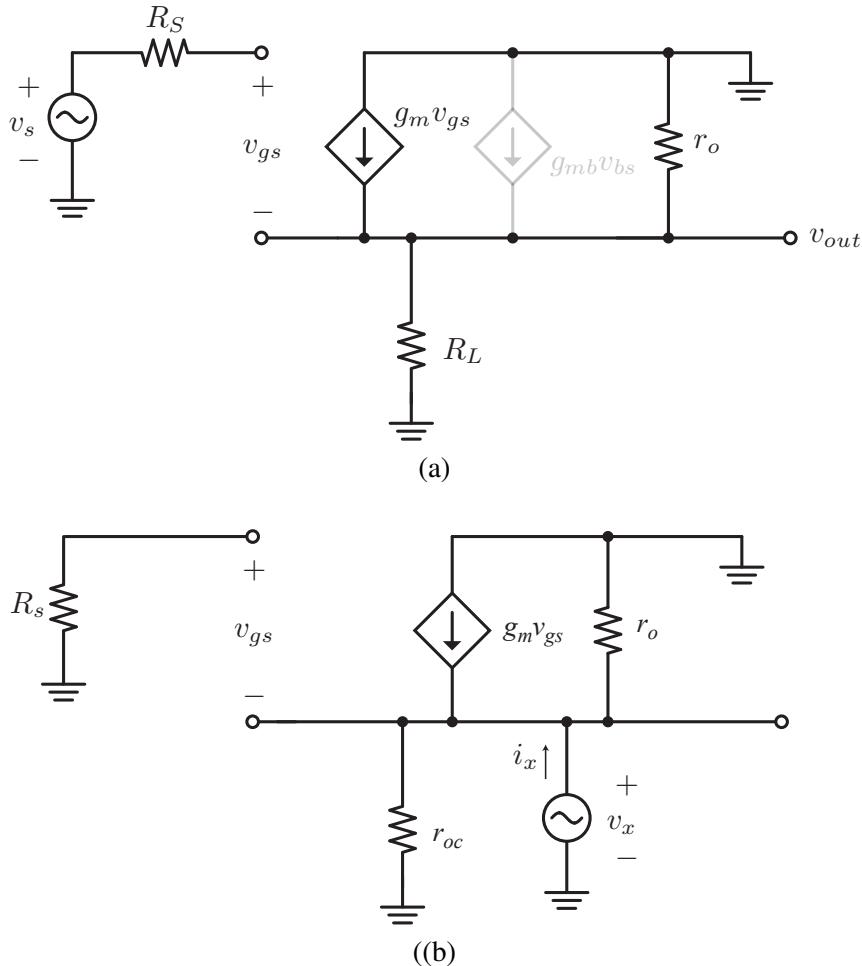


Figure 12.18: (a) *Common drain* amplifier small-signal equivalent circuit. (b) Small-signal circuit schematic for calculation of output impedance.

12.4.2 CD Voltage Gain

To find the voltage gain of the CD amplifier, we apply an AC source to the gate and observe the output voltage at the source, as shown in *Fig. 12.18a*. In the small-signal model, we have grayed out the **back-gate generator** g_{mb} , which would be the case if we short the source to the body of the transistor. We will handle the situation of the grounded body later. Let's write KCL at the output node:

$$\frac{v_{out}}{R_L \parallel r_o} = g_m v_{gs}$$

Because the input and output are just separated by v_{gs} , or $v_{gs} = v_s - v_{out}$, we have:

$$\frac{v_{out}}{R_L \parallel r_o} = g_m(v_s - v_{out})$$

Collecting common factors:

$$\left(\frac{1}{R_L \parallel r_o} + g_m \right) v_{out} = g_m v_s \quad (12.30)$$

Eq. 12.30 leads us to the voltage gain:

$$\boxed{\frac{v_{out}}{v_s} = \frac{g_m}{\frac{1}{R_L \| r_o} + g_m}}$$

Common drain voltage gain (12.31)

Assuming that $r_o \gg R_L$, we have:

$$\boxed{\frac{v_{out}}{v_s} \approx \frac{g_m}{\frac{1}{R_L} + g_m} \approx 1}$$

(12.32)

This final result in *Eq. 12.32* leads to another name for this amplifier, namely the "**Source Follower**", because the source voltage closely tracks the gate voltage. In other words, this is a voltage amplifier with unity gain, or a voltage buffer.

To be a voltage buffer, though, also requires high input impedance and low output impedance. Obviously the input impedance is large, especially at low frequencies, because of the insulating nature of the gate. Now let's examine the output impedance of the CD.

12.4.3 CD Output Resistance

To calculate the output impedance, refer to *Fig. 12.18b*. We apply a test source v_x and find the current i_x . For a moment, forget about r_o and r_{oc} , the output resistance of the transistor and the current source.¹ They are simply in parallel with the source and can be added later. In that case, let's sum currents at output (source) node:

$$\boxed{i_x = g_m v_x} \quad (12.33)$$

Eq. 12.33 which leads to:

$$\boxed{R_{out} = (r_o \parallel r_{oc}) \parallel \frac{v_x}{i_x} = \left((r_o \parallel r_{oc}) \parallel \frac{1}{g_m} \right)}$$

(12.34)

If $r_o \parallel r_{oc}$ is much larger than the inverses of the transconductance, the output impedance is given by:

$$\boxed{R_{out} \approx \frac{1}{g_m}}$$

Common drain output resistance (12.35)

Recall that $g_m \propto \sqrt{I_{DS}}$, so a good voltage buffer is biased with a large g_m , and burns sufficiently large DC current. The CD amplifier is often used as an output stage to drive a heavy load. A heavy load is a small resistance, because for a fixed voltage swing it draws more power than a larger resistor. In order to buffer a gain stage from a small load, we can add a *source follower* amplifier. Putting everything together, a CD amplifier is modeled as shown in *Fig. 12.19*.

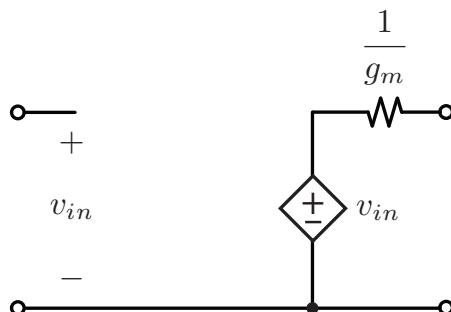


Figure 12.19: A two-port model for the CD amplifier is a voltage buffer with unity gain and output impedance $1/g_m$.

¹For proof that this simplification yields the same result using KCL, see *App. C*.

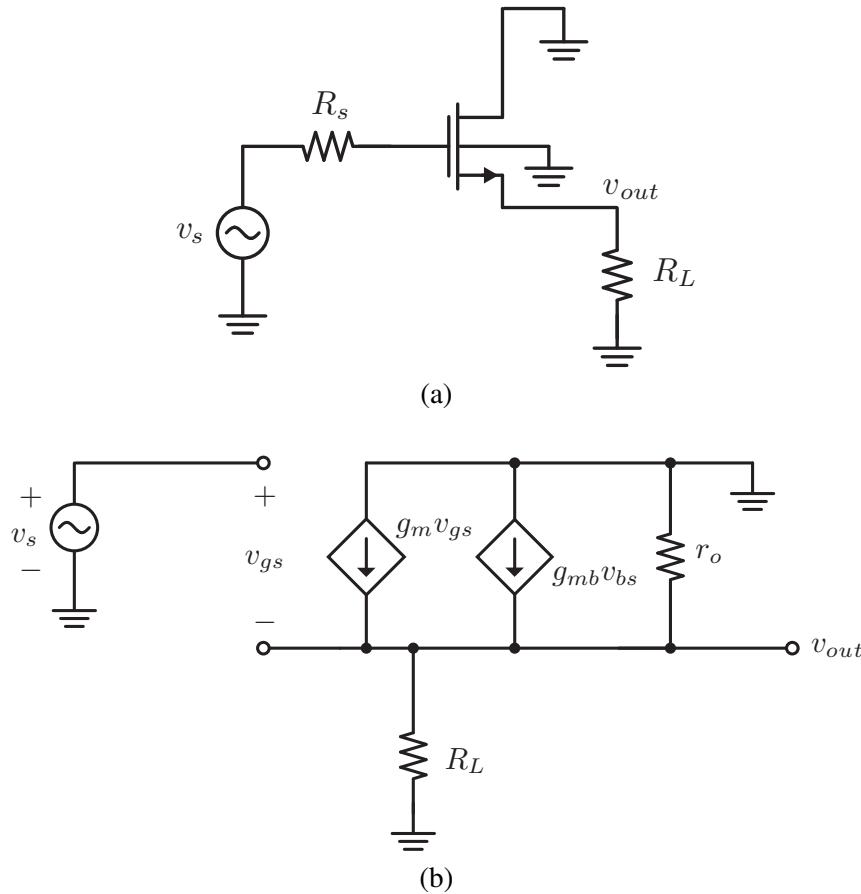


Figure 12.20: (a) The CD amplifier built with a four terminal MOSFET with a grounded body. Note that the voltage v_{sb} is not zero and therefore we must consider (b) the complete low-frequency small-signal schematic of the CD amplifier (including the back-gate effect).

12.5 Impact of Body Effect

As we noted earlier, if the source and body are not tied together, there is a back-gate effect that we need to consider. Recall that if an NMOS transistor is fabricated in a *P*-type substrate, then its body connection is the substrate. The body connection must be grounded in an IC (or tied to the lowest potential), as shown in *Fig. 12.20a*. It is possible to create a "triple well" process (or a "deep N-well" process) to accommodate placing NMOS devices in their own body wells. This process can prevent the back-gate effect, but these process options are more expensive and increase the size of the device. Often many NMOS devices will reside in the same substrate or *P*-well, and so all the bodies are tied to a common (ground) node.

To calculate the gain with the back-gate effect, we return to the full **four-terminal small-signal** schematic, redrawn in *Fig. 12.20b*. To find the voltage gain, we write KCL at the output node:

$$\frac{v_{out}}{R_L \parallel r_o} = g_m v_{gs} + g_{mb} v_{bs}$$

As before, $v_{gs} = v_s - v_{out}$, but now we must consider that the source voltage moves with respect to the body:

$$v_{bs} = 0 - v_{out} = -v_{out}$$

Making these substitutions:

$$\frac{v_{out}}{R_L \parallel r_o} = g_m (v_s - v_{out} - g_{mb} v_{out})$$

Collecting common factors:

$$\left(\frac{1}{R_L \parallel r_o} + g_m + g_{mb} \right) v_{out} = g_m v_s$$

This last equation leads us to the voltage gain:

$$\frac{v_{out}}{v_s} = \frac{g_m}{\frac{1}{R_L \parallel r_o} + g_m + g_{mb}}$$

CD gain with body effect (12.36)

Amplifier Topologies

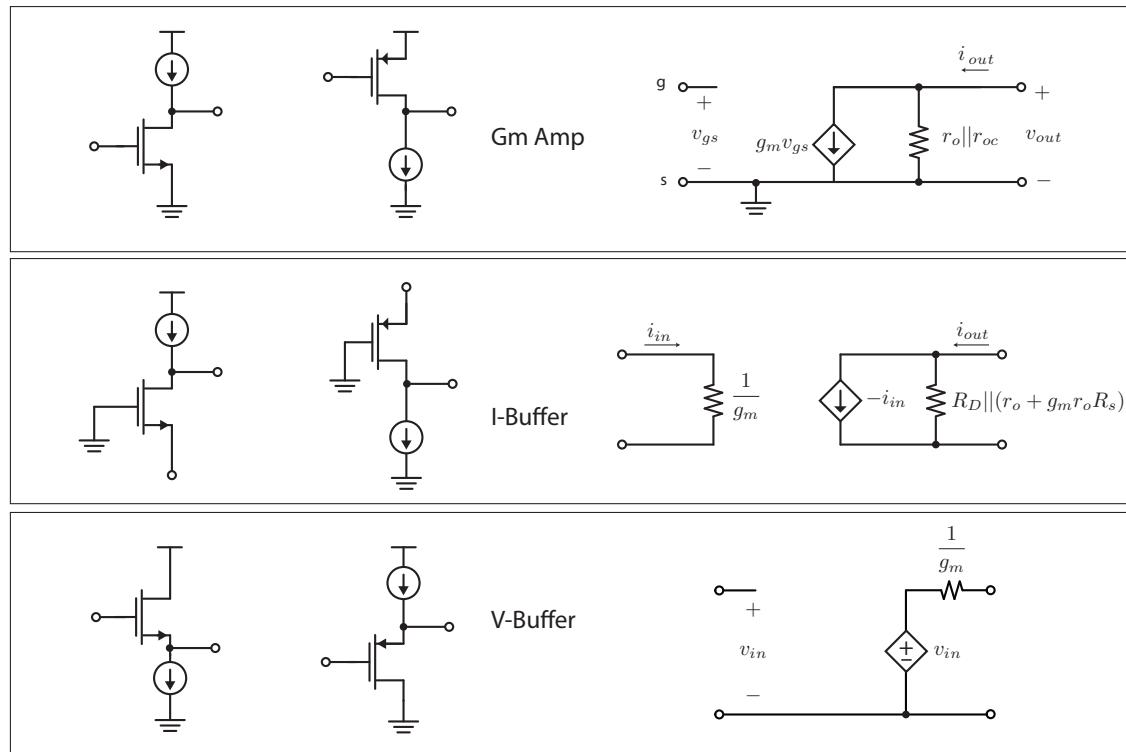


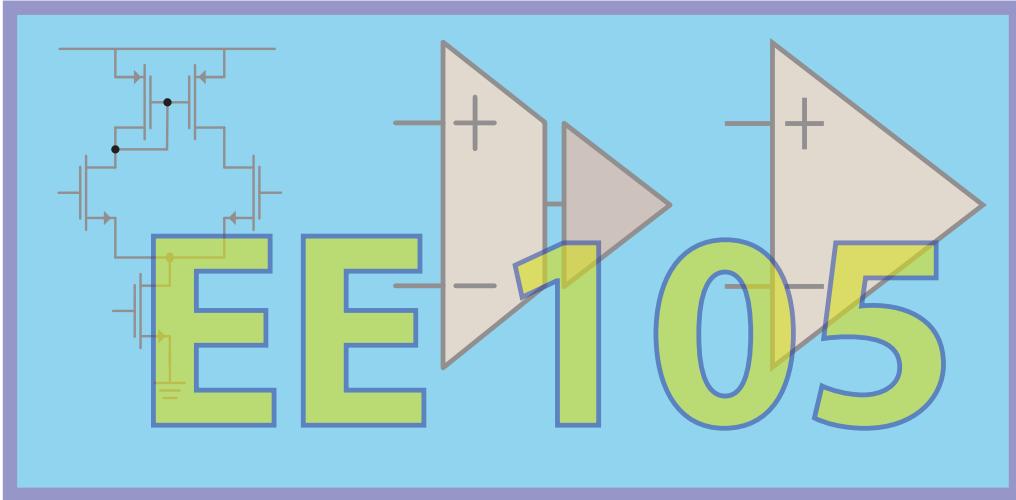
Figure 12.21: The family of single-stage transistor amplifiers and the corresponding two-port models. As you analyze more circuits, this chart will become to be imprinted in your memory.

12.6 Chapter Summary: Amplifiers $\rightarrow G_m/V/I$

Let's summarize this chapter by reviewing the three possible amplifier topologies and the equivalent circuit, all shown in *Fig. 12.21*. In each case we model the amplifier in a certain way that best characterizes the input/output impedance of the topology. For example, a CS amplifier has high input impedance and high output impedance, making it an ideal transconductance amplifier. This is not to say that it cannot be designed as a (inverting) buffer, a voltage amplifier, or a current amplifier. It is just that the native impedance level of the CS lends well to it being used as a transconductance amplifier.

The CG amplifier is likewise best seen as a current buffer, because the input and output currents are nearly identical. But CG amplifiers make great voltage amplifiers too, provided that you can drive their input impedance. Their output impedance is also high, and so a voltage buffer is needed if driving a low impedance load is desired.

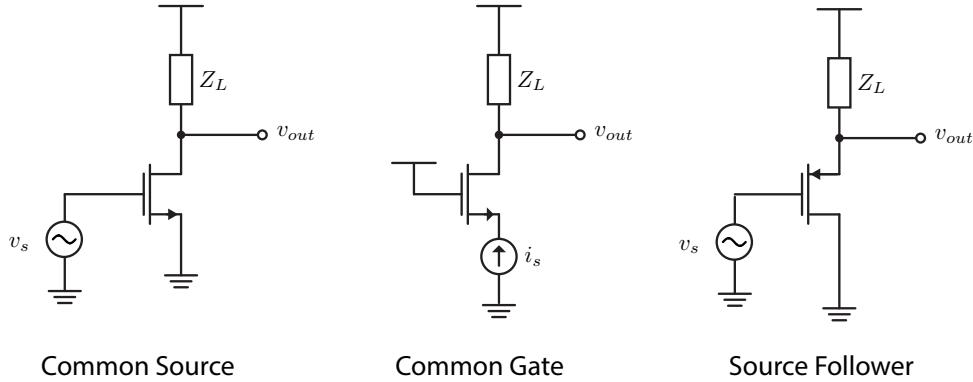
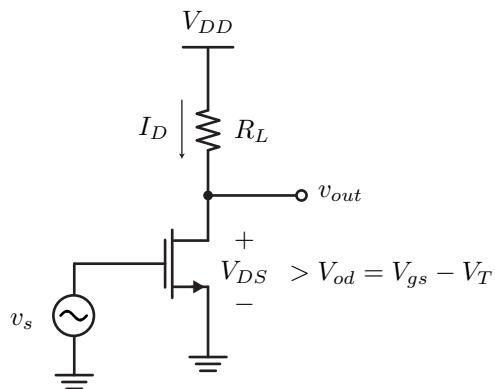
Finally, the CD amplifier is best viewed as a voltage buffer, providing high input impedance and low output impedance. Its voltage gain is less than unity, but with enough bias current we can approach unity gain. Even though the voltage gain is always less than one, it can provide *power gain* by driving a load with power without requiring much input power.



13. Current Mirrors and Biasing

13.1 Chapter Preview

This chapter begins by motivating the need for high load impedance in amplifiers. Resistive loads have many drawbacks, including taking up a lot of physical space on an IC, and requiring a very large voltage headroom due to IR voltage drops. We introduce the basic current mirror, which can be used as a current source load in an amplifier with reduced headroom requirements. Next we introduce an improved current source known as the cascode current source. We discuss both NMOS and PMOS variants, which can be used to build current sources and current "sinks". We conclude the chapter with an amplifier design example.

Figure 13.1: High voltage gain amplifiers require a high load impedance Z_L .Figure 13.2: To maximize voltage gain, the transistor should operate in the saturation (or forward-active) region, which means $V_{DS} > V_{OD}$ ($V_{CE} > V_{CE,sat}$), which places limits on how large we can make R_L for a fixed bias current.

13.2 High Load Impedance in Amplifiers

13.2.1 Load Impedance

In each amplifier shown in *Fig. 13.1*, to maximize gain we should maximize the **load impedance** Z_L . Can we just use an arbitrarily large load resistor? In an IC process technology, resistors are realized either as diffusion regions or using polysilicon films. The resistance per square is on the order of 100's to 1000's of ohms-per-square, which requires many squares to realize a large resistor, taking up a lot of area. We are also limited in the minimum width of the resistor, because thinner resistors have limits on maximum current handling capability (due to self-heating). Thinner resistors also experience wider variations from part to part, making the design gain more variable. Next we will discuss the biggest limitation of all, which is simply the voltage drop across the resistors.

13.2.2 Headroom Limitations

To keep the transistor in saturation requires $V_{DS} > V_{OD}$ (the over-drive voltage, $V_{GS} - V_T$). For a fixed bias current, if we increase R_L , eventually we squash the transistor. This means that it goes from the saturation region to the triode region, and the gain of the circuit is compromised. This is illustrated in *Fig. 13.2*. One solution is to use a larger V_{DD} , but in practice this solution is hard to implement because each generation of IC technology has a voltage supply limit. For example, today's nanoscale transistors can only tolerate ~ 1 V of voltage supply due to the thin oxides in the

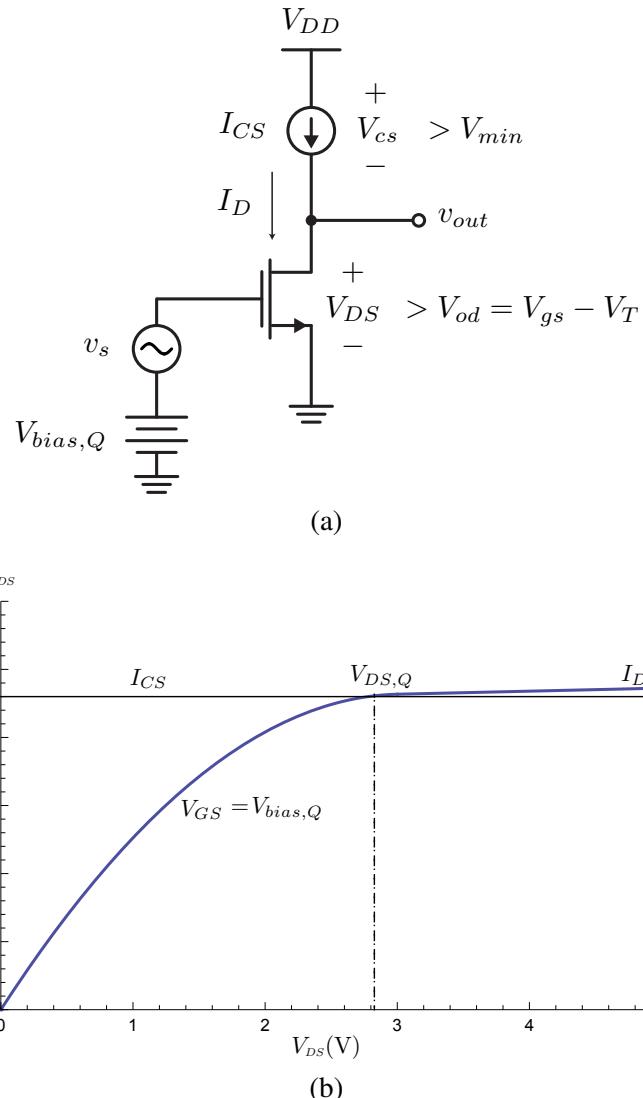


Figure 13.3: (a) An ideal current source load results in the maximum gain for any bias current. An ideal current source does not dictate the output voltage, just the current. Note that $V_{bias,Q}$ determines the transistor operating point, which should match I_{CS} . (b) The actual output voltage is determined by the transistor I_{DS} - V_{DS} curve.

devices. Even "thick oxide" devices, available for I/O¹, typically work at a maximum voltage of 2.5V or 3.3V (these voltage levels are standardized).

13.2.3 Achieving High Gain

What if someone handed you a current source? Then you could use it to bias the transistor, as shown in Fig. 13.3. Because an ideal current source has infinite output resistance, it would maximize the gain while supporting any value of V_{DS} . Unfortunately, ideal current sources don't exist. However, we could build one from a transistor with a fixed gate bias. But how should we generate the correct bias voltage? We will answer this question by first considering the wrong approach.

¹Iinput/output devices are used for driving external loads and receiving inputs from the outside of the chip.

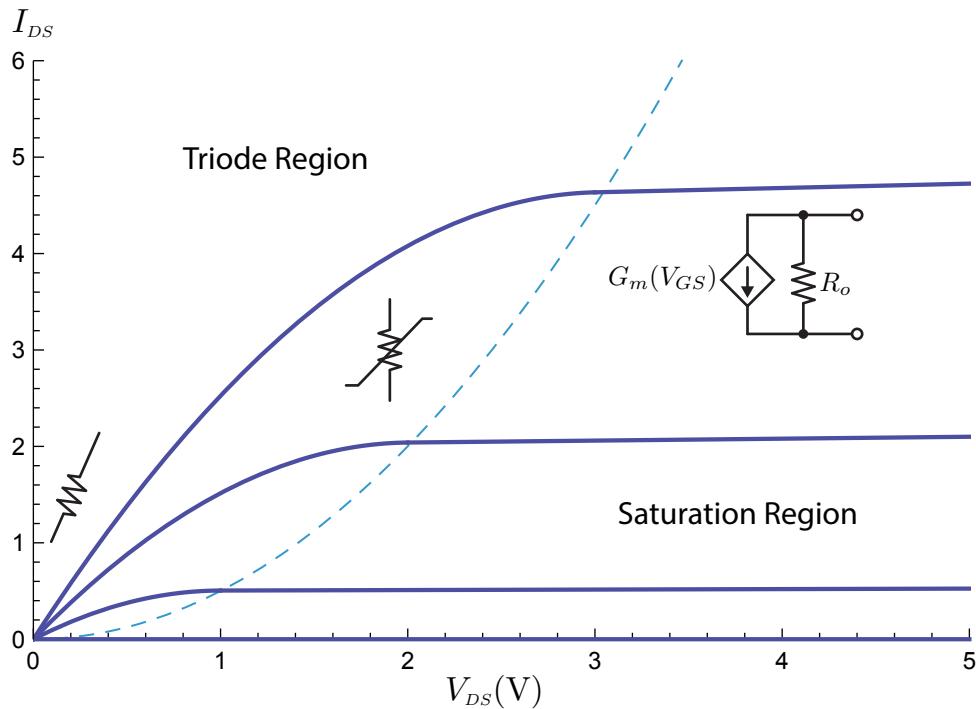


Figure 13.4: An MOS transistor in the saturation region, or a BJT in the forward active region, has nearly constant current as a function of V_{DS} , making it a suitable proxy for an ideal current source if the output voltage does not go too low (or high for a PMOS transistor). The gate voltage (base voltage) must be held constant at the appropriate bias voltage.

13.2.4 Transistor Current Source

A transistor biased in saturation with a constant V_{GS} is effectively a decent current source, as shown in the I - V curve in *Fig. 13.4*. Even though the transistor is a dependent current source, the main source of dependence is the gate voltage, which we would like to hold constant. The other dependence is the output resistance, which is much smaller, and results in non-zero slope in the I - V curve.

13.2.5 Transistor Process / Temperature Variations

If we simply fix the V_{GS} value to realize a specific current, we are designing a current source that will have a very unpredictable output current. For example, if the temperature varies, the threshold voltage changes, and so will the current. Also, if we fix the gate bias value based on a model, then in real applications we will find that every transistor is a bit different from the model due to natural statistical variations (in doping levels and geometry, especially for small transistors). So the output current will be a random variable with a large variance. To keep the variance low we should use another strategy.

A good strategy is to observe that while two arbitrary transistors are not well matched, two transistors sitting right next to each other, or better yet two "inter-digited transistors" (how this is done will be explained in *Sec. 13.3.4*), will behave very similarly. So the strategy should be to use one transistor to generate the V_{GS} for the other.

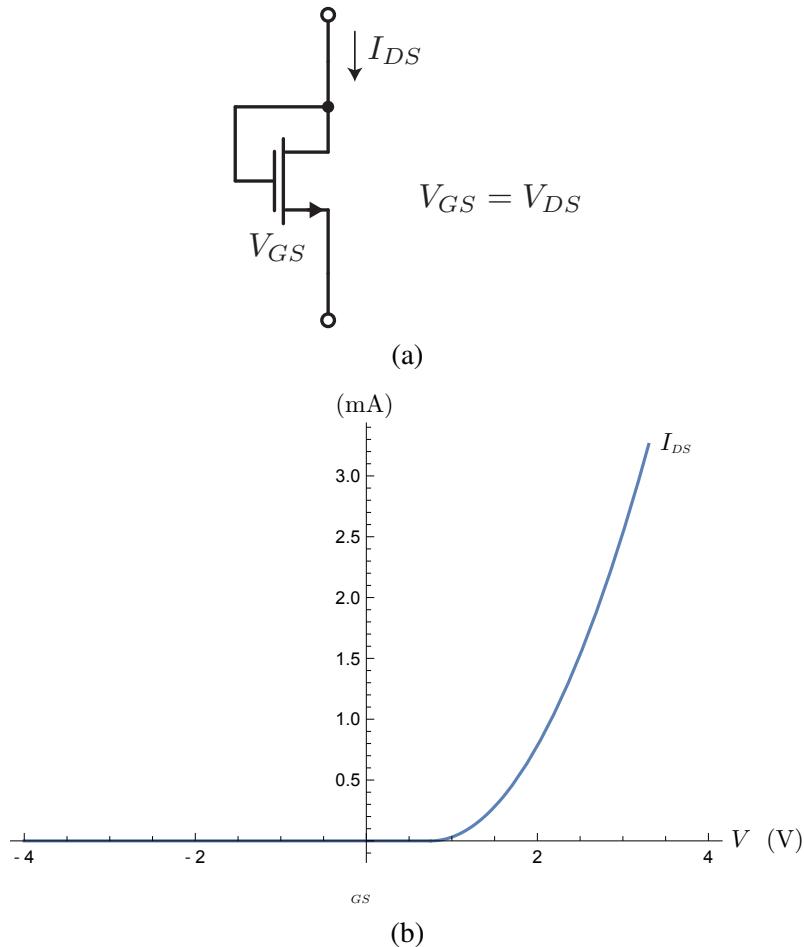


Figure 13.5: (a) An MOS transistor with gate-drain shorted is known as a "diode connected" transistor. It behaves as a non-linear unidirectional conductor, similar to a diode. (b) The I_{DS} - V_{GS} of a MOS diode is "rectifying" in that it only allows current to flow in one direction.

13.3 The Basic Current Mirror

13.3.1 Diode Connected Device

The **diode connected MOS transistor**, shown in *Fig. 13.5*, is a two-terminal device. It is never in the triode region, because of the gate-drain connection. Since $V_{DS} = V_{GS}$, as long as V_{GS} is over the threshold voltage, then the drain-source voltage is a threshold voltage away from saturation, since $V_{DS} = V_{GS} = V_{D,sat} + V_T$. As we sweep V_{GS} and observe the drain-source current, the transistor goes from "off" state (below threshold) to the "on" state. It is called a diode because of the rectifying action of the the I - V relation. Current can pass in one direction, but not the other, similar to a PN -junction diode. The curve is not exponential, though, since the MOS device I - V relation is quadratic. If we pass a known current into a diode, the V_{GS} value will be well-defined and equal to the value needed to generate the current. It is like an inverse function, that takes a current and generates the correct V_{GS} :

$$V_{GS} = V_T + \sqrt{\frac{2I_{DS}}{\frac{W}{L}\mu C_{ox}}} = V_T + V_{OD}$$

If the current injected into the device is very low, V_{OD} will be low and the diode V_{GS} will be approximately V_T . In general for any current, the V_{GS} generated can be used to bias a second transistor into the same current.

13.3.2 Diode Connected – Small-Signal Model

We can derive the small-signal model by shorting out the gate-drain capacitor in the hybrid- π model, shown in Fig. 13.6. Note that a g_m generator with its controlling terminals connected to the g_m is more simply a conductor. From the perspective of the driving terminal, the MOS diode is just a conductance.

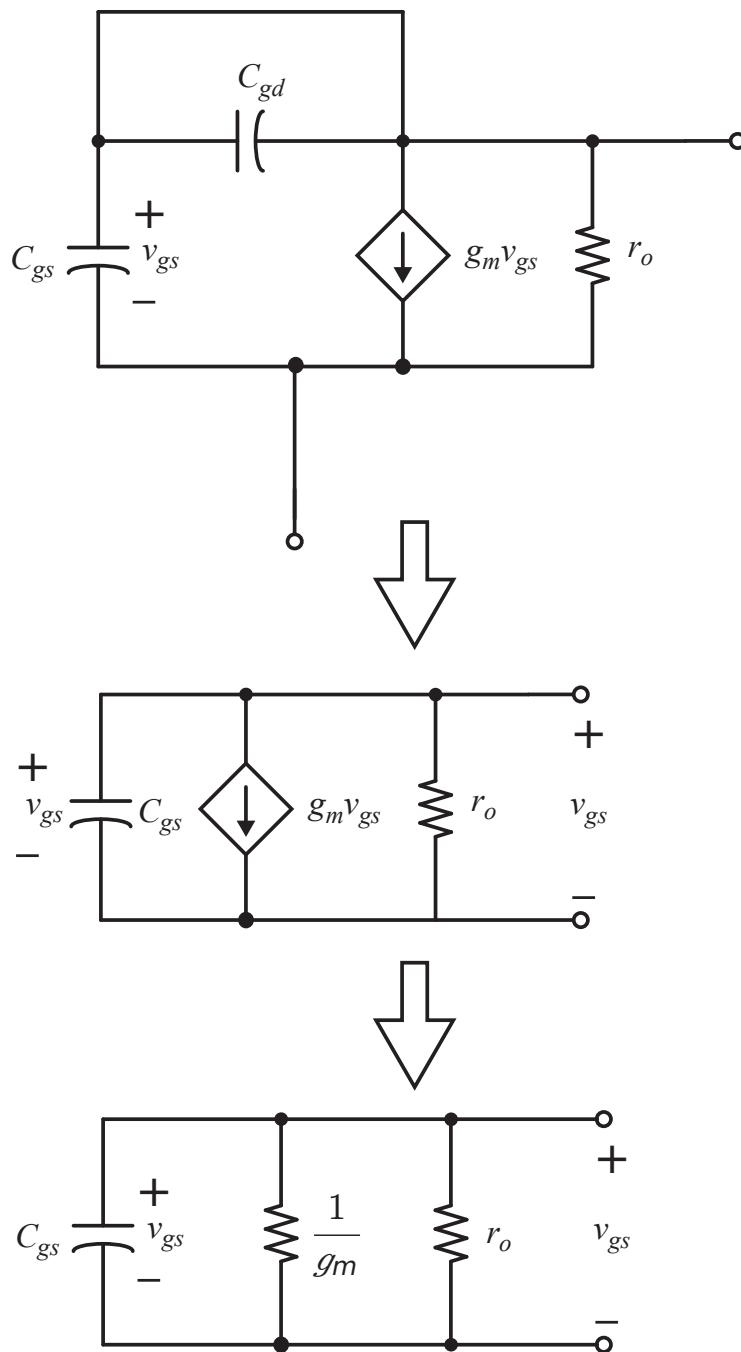


Figure 13.6: The small-signal model of a diode connected MOS transistor. Notice that the model is simply a conductance when the gate-drain are shorted, as shown on the bottom.

13.3.3 The Integrated "Current Mirror"

With the MOS diode as a voltage reference, we can build a "current mirror", shown in *Fig. 13.7a*. Since $M1$ and $M2$ have the same V_{GS} , as long as $M2$ is in saturation, they will carry approximately the same current. If we neglect CLM ($\lambda = 0$), then the drain currents would be equal. Since λ is small, the currents will nearly mirror one another even if V_{out} is not equal to V_{gs1} . We say that the current I_{in} is mirrored into I_{out} . Notice that the mirror works for small and large signals!

It is important to note that the diode side of the current is low impedance, due to the diode connection, as shown in *Fig. 13.7b*. On the other hand, from the drain of $M2$ the circuit presents a high impedance, thereby behaving like a current source.

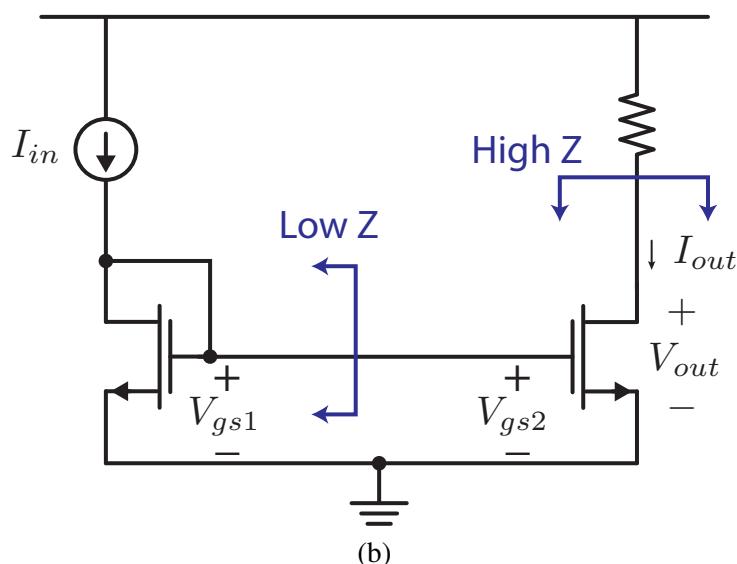
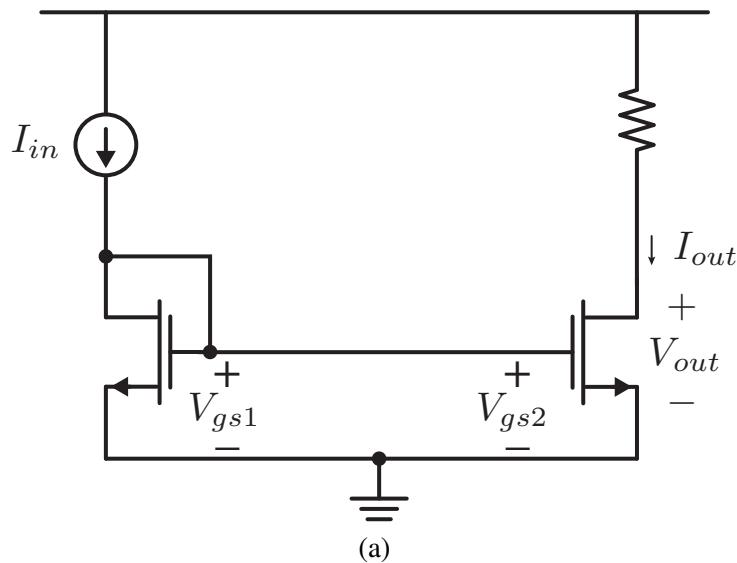


Figure 13.7: (a) An NMOS current mirror has an output current I_{out} that is a "mirror copy" of I_{in} . (b) The mirror has low impedance on the "diode" side and high impedance at the output.

13.3.4 Current Mirror with Multiplication Ratio

As illustrated in *Fig. 13.8*, with some minor modifications, we can generate a mirror that **scales the input current up or down**. The input and output currents are related as follows:

$$I_{IN} = k \left(\frac{W_1}{L_1} \right) (V_{GS1} - V_T)^2 \quad (13.1)$$

$$I_{OUT} = k \left(\frac{W_2}{L_2} \right) (V_{GS2} - V_T)^2 \quad (13.2)$$

The circuit connections forces the two transistors to operate at the same V_{GS} . Assuming that they have the same threshold voltage (which is a good assumption if the transistors are nearby and laid-out together):

$$V_{GS1} = V_{GS2} \quad (13.3)$$

Which allows us to write:

$$I_{OUT} = k \left(\frac{W_2}{L_2} \right) (V_{GS2} - V_T)^2 = I_{IN} \left(\frac{W_2/L_2}{W_1/L_1} \right) = N \cdot I_{IN} \quad (13.4)$$

In practice we prefer not to scale L , but rather the transistor widths W to achieve the desired **scaling ratio** N (unless N is very large, then a combination of the both L and W will be scaled). This is because the transistor threshold voltage and other parameters will vary with L , and it is best to match the two mirror transistors as much as possible, so they should have the same channel length. In fact, the most accurate mirror is realized by inter-digitating two transistors, effectively connecting multiple unit transistors in parallel (see *Fig. 13.9*), or better yet using multiple "fingers" in the layout of the transistors (see *Fig. 13.10*):

$$I_{OUT} = I_{IN} \left(\frac{W_2}{W_1} \right) = I_{IN} \left(\frac{N_2 \cdot W_U}{N_1 \cdot W_U} \right) = N \cdot I_{IN}$$

Current mirror with scaling ratio (13.5)

In *Eq. 13.5*, W_U is the unit transistor width.

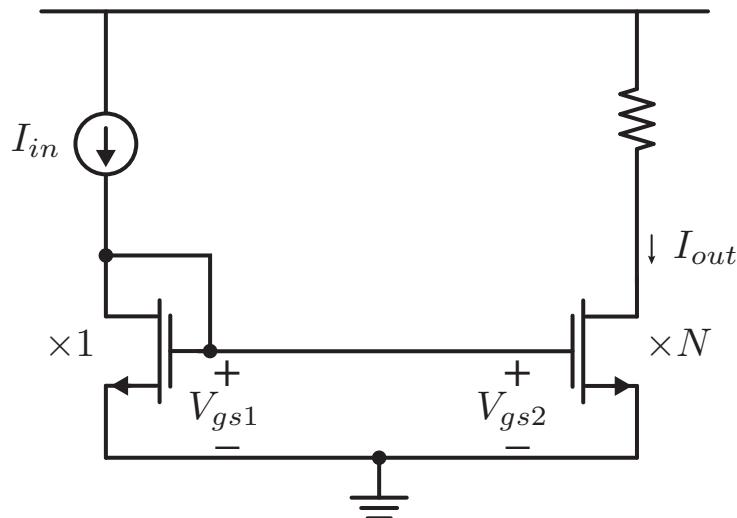


Figure 13.8: A current mirror can be configured to scale the input current by scaling the transistor dimensions W/L .

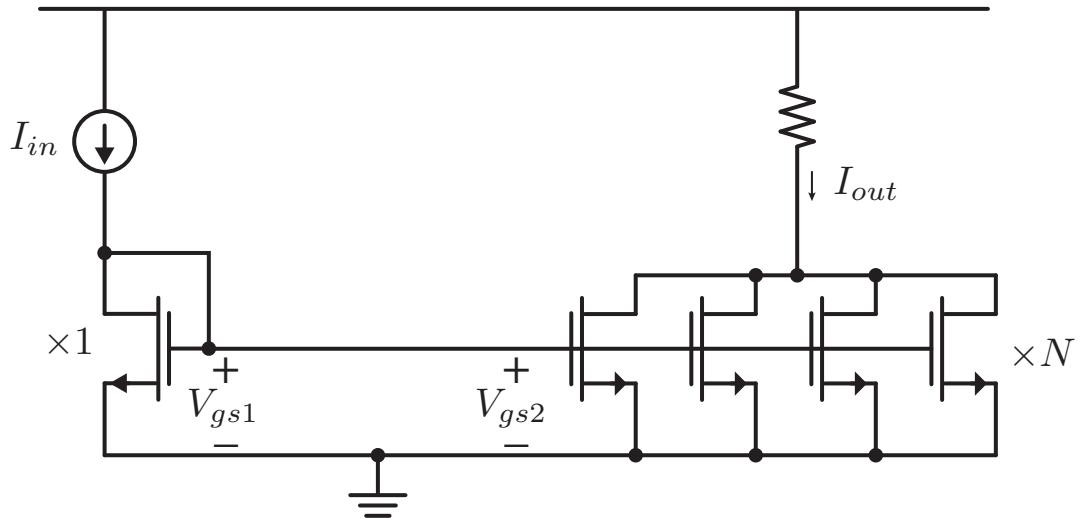


Figure 13.9: It is preferable to scale the input of a current mirror by scaling the transistor count, using parallel devices to realize a larger W .

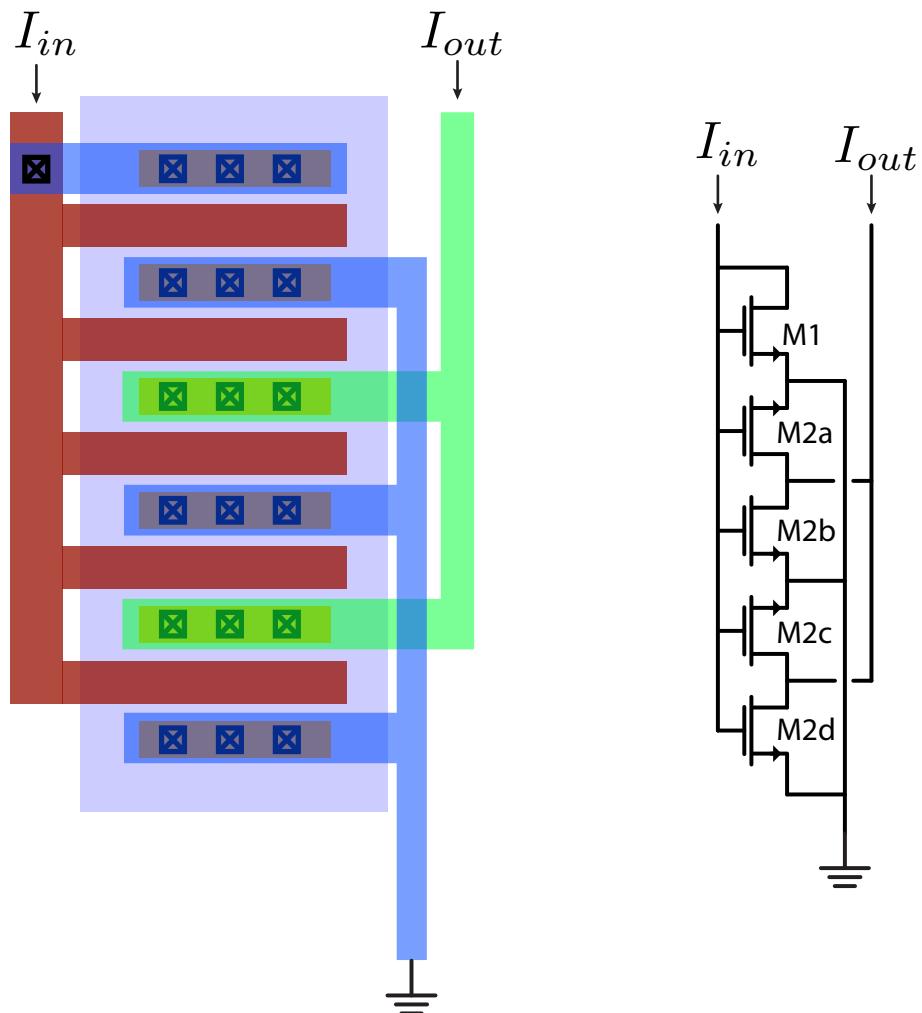


Figure 13.10: The layout of a 4:1 mirror using five unit elements. Note that transistors are abutted and source/drain junctions are shared to maximize matching between the gates, save area and minimize parasitics. The input transistor is diode connected and the output transistor is split into four fingers.

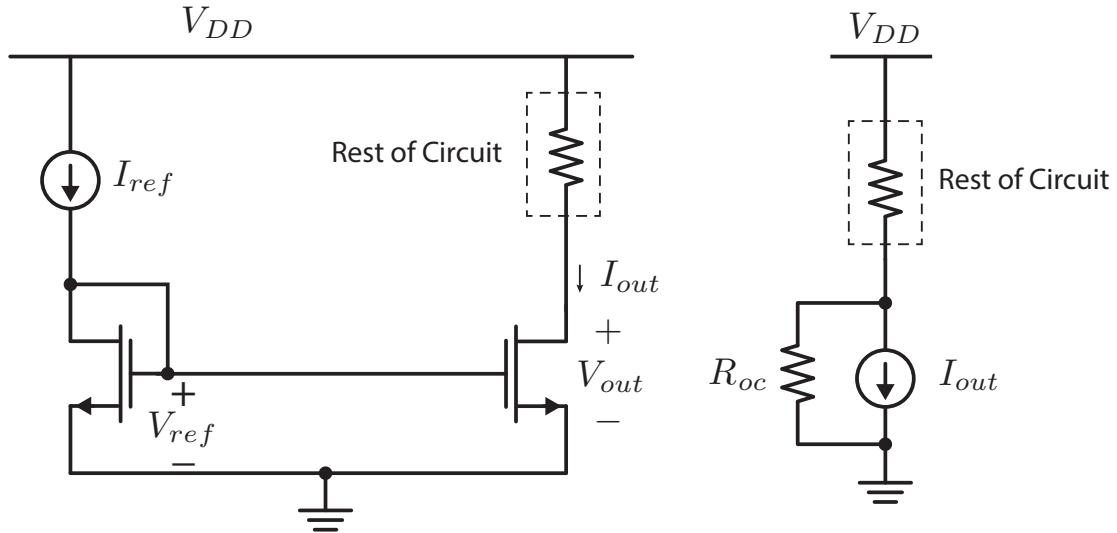


Figure 13.11: From the perspective of the "rest of the circuit" shown, the current mirror appears like an ideal current source with output impedance R_{oc} .

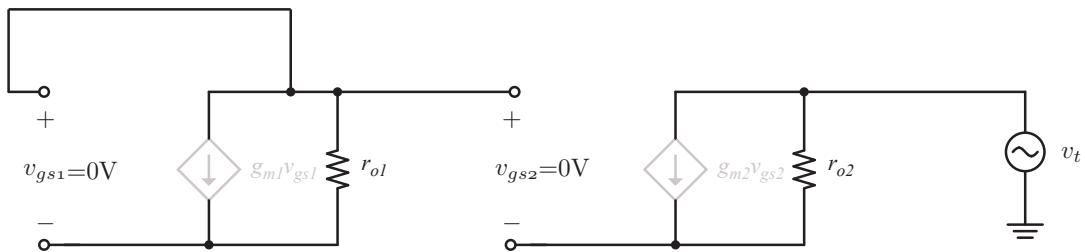


Figure 13.12: Current mirror small-signal model. Transconductors are open-circuit (off) since the controlling voltages are zero.

13.3.5 Current Mirror as Current Source

As shown in *Fig. 13.11*, a current mirror acts like a current source to the rest of the circuit as long as the transistor M_2 remains in saturation, where it has high output resistance. Of course, there is a slight problem in that we need another voltage reference I_{ref} to make this work. The idea is that we generate one very accurate current reference using techniques that we will discuss in Section 13.7, and this current reference can be mirrored around to various parts of the circuit using mirrors.

13.3.6 Small-Signal Resistance of Current Source

The output impedance seen by the load can be calculated using the small-signal model shown in *Fig. 13.12*. Use a test generator v_t , and find the current draw from the source. Note that M_1 is not driven by a source, because I_{ref} is DC, which means that it is an open circuit in the small-signal model. Thus, without a source, M_1 is effectively driven with zero v_{gs} and has no output, pulling the v_{gs} of both transistors to ground. Then M_2 's dependent g_m is zero, and the only current draw is from $r_{o,2}$.

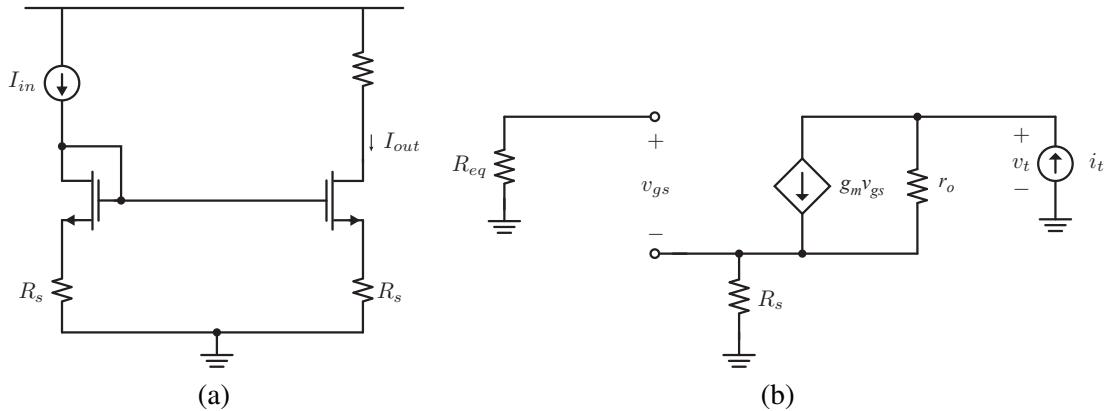


Figure 13.13: (a) Adding resistors in the source of a current mirror can be used to boost the output impedance at the cost of reducing the voltage headroom. Note that the output impedance is boosted by a factor $g_m R_s$, rather than the additive factor of R_s , which would be true if the resistor were placed on the drain side. (b) Small-signal model of improved output impedance mirror utilizing source resistance R_s .

13.4 Improved Current Source: The Cascode Mirror

13.4.1 Improved Current Sources

The goal is to increase R_{out} of the basic current mirror. To get a hint about how to proceed, let's look at typical *amplifier* output resistance results to see topologies that boost resistance. The output impedance of the *common gate* stage is boosted by placing a resistor in the source of the amplifier. Likewise, the same is true of a *common source* amplifier with **source degeneration**. The same approach can be used to boost the output impedance of a mirror, as shown in *Fig. 13.13a*.

13.4.2 Effect of Source Degeneration

Let's analyze the small-signal model of the mirror with R_S , as shown in *Fig. 13.13b*. We focus on the output side, or transistor $M2$, and model the input side as a Thévenin equivalent resistance, R_{eq} , because there are no sources on the left side. We inject a test current i_t and measure the voltage developed across i_t :

$$v_t = (i_t - g_m v_{gs})r_o + v_{R_S} \quad (13.6)$$

First, note that the current i_t flows through R_S as it returns through the source. This implies that the voltage across R_S is given by:

$$v_{R_S} = i_t R_S \quad (13.7)$$

The key observation is that the current flowing in R_S generates a gate-source voltage on the transistor, $v_{gs} = -v_{R_S}$. This in turn creates a current through g_m :

$$v_t = (i_t - g_m (-v_{R_S}))r_o + v_{R_S} \quad (13.8)$$

$$= (i_t + g_m R_S i_t)r_o + i_t R_S \quad (13.9)$$

Collecting terms, we find that the output impedance is given by:

$$R_o = \frac{v_t}{i_t} = (1 + g_m R_S) r_o \quad (13.10)$$

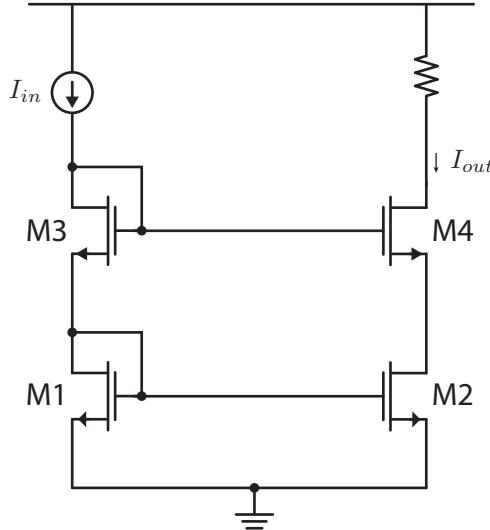


Figure 13.14: An improved "cascode" or stacked current mirror.

The output impedance is boosted by a factor of $(1 + g_m R_S)$, compared to a simple mirror. How would you scale the output current? If we use the same approach as before, we need to ensure that we scale the resistance as well:

$$I_{IN} = k \left(\frac{W_1}{L_1} \right) (V_G - V_S - V_T)^2 \quad (13.11)$$

where,

$$V_S = I_{IN} R_S \quad (13.12)$$

To keep V_S the same, if I_{OUT} is N times larger, then R_S should be N times smaller.

13.4.3 Cascode (or Stacked) Current Source

The stacked current mirror, commonly known as the **cascode current mirror**, is shown in Fig. 13.14. Notice that V_{GS2} is held constant like a normal mirror, but now V_{DS2} is also held approximately constant by $M3/M4$. This makes the output current variation much smaller than a single transistor.

We can easily compute the boost in output impedance by noting that $R_S = r_{o1}$ as far as transistor M4 is concerned:

$$R_o \approx (1 + g_{m2} R_S) r_{o2} = (1 + g_{m2} r_{o1}) r_{o2} \quad (13.13)$$

Which is approximately $g_m r_o$ times larger than a simple mirror:

$$R_o \approx g_{m1} r_{o1} r_{o2} \gg r_o \quad (13.14)$$

13.4.4 Drawback of Cascode Current Source

Like all good things, there is a catch. The minimum output voltage to keep all transistors in saturation increases, as shown in Fig. 13.15. For a basic current mirror, we can go all the way down to V_{OD} and keep $M2$ in saturation. When we introduce R_S , we have to account for the extra $I_{OUT} R_S$ voltage drop. With a cascode, the minimum voltage to keep $M4$ in saturation is given by noting that the source of $M4$ is biased by $M3$:

$$V_{S4} = V_{G4} - V_{GS4} \quad (13.15)$$

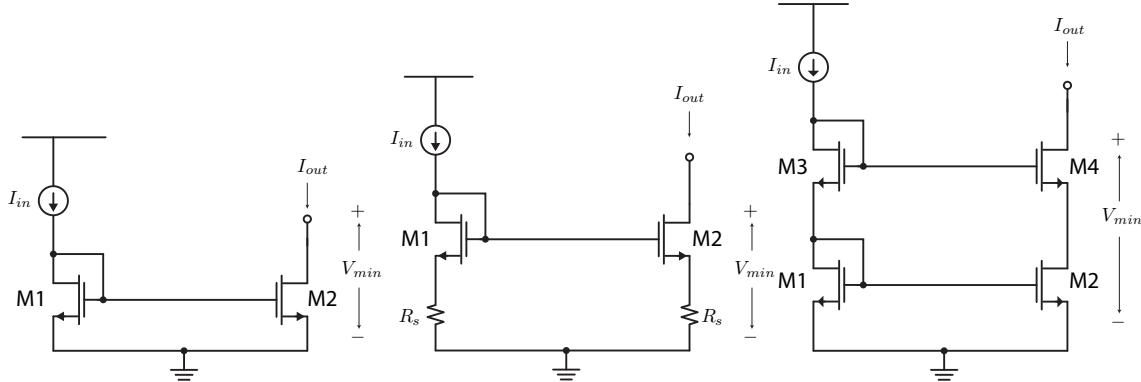


Figure 13.15: Comparison of the reduction in the output swing of current mirrors as we improve the output impedance.

The gate of transistor M_4 is basically the V_{GS} of M_1 and M_3 :

$$V_{G_4} = V_{GS_1} + V_{GS_3} = 2V_T + 2V_{OD} \quad (13.16)$$

To arrive at the RHS of Eq. 13.16, we assume that all devices have the same V_T and size, so that the overdrive voltages are equal. Since the $V_{GS_4} = V_T + V_{OD}$, we can write Eq. 13.15:

$$V_{S_4} = V_{G_4} - V_{GS_4} = 2V_T + 2V_{OD} - (V_T + V_{OD}) = V_T + V_{OD} \quad (13.17)$$

Now the minimum $V_{DS_4} = V_{OD}$ implies that the cascode mirror has a **minimum operating voltage** of:

$$V_{min} = V_T + 2V_{OD} \quad \text{Cascode minimum operating voltage} \quad (13.18)$$

This can be quite a large voltage penalty compared to a simple mirror, but an easy solution is to bias the gate of M_4 with $V_T + 2V_{OD}$ using a separate transistor, shown in Fig. 13.16. The transistor overdrive voltage is twice as large by scaling the device down by $4\times$. In a more advanced analog integrated circuits course you will learn about other ways to realize the "**high swing**" cascode.

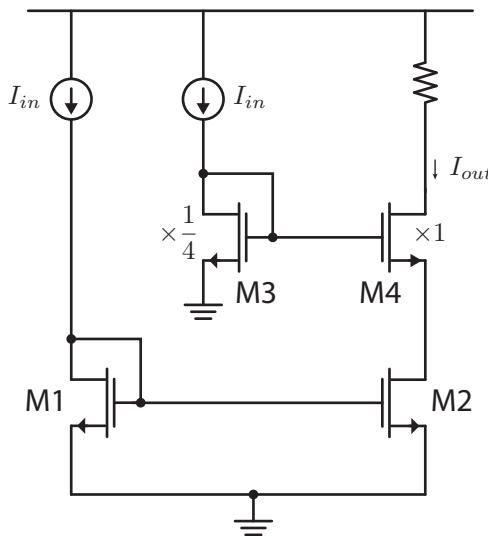


Figure 13.16: To maximize the swing of a cascode mirror, the gate of the top "cascode" transistor M_4 should be biased at $V_T + 2V_{OD}$ to place M_2 at the edge of saturation.

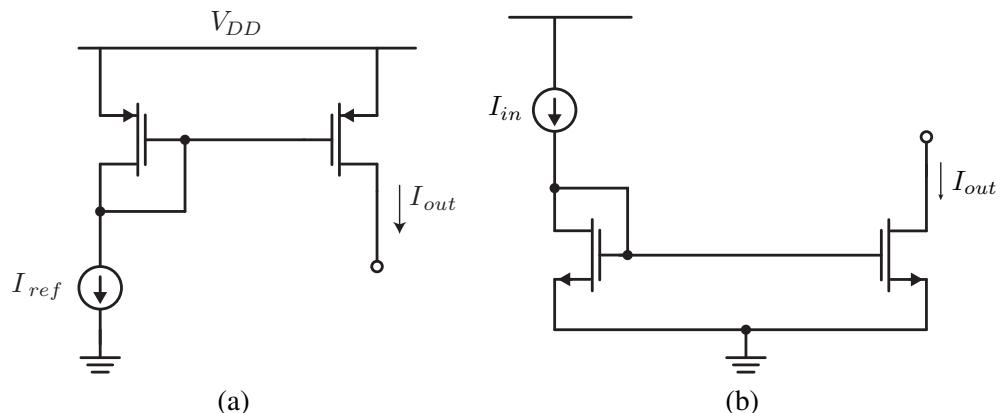


Figure 13.17: (a) A PMOS current source can "source" a current from supply. (b) In contrast, an NMOS current source can only "sink" a current to ground.

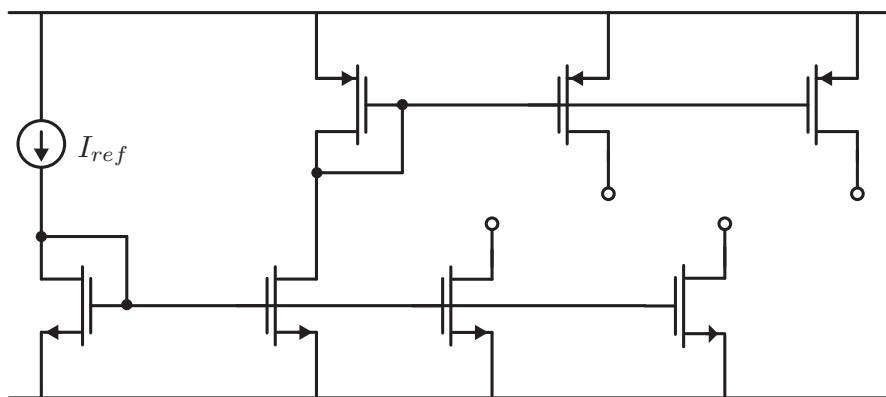


Figure 13.18: A precision current reference I_{ref} is fed into this circuit and multiple copies are produced, both as current sinks and sources. Each current can be scaled appropriately as needed.

13.5 Current Sources and Sinks

Technically, up to this point we have really been talking about a **current "sink"** since the NMOS transistors can only sink current to ground. What if we need a **current "source"**? The **PMOS mirror**, shown in *Fig. 13.17*, is exactly the dual of the **NMOS mirror**, and it simply mirrors a current that is a "source" from the supply. In all other aspects, such as the output impedance and scaling, it is identical to the NMOS mirror. The cascode NMOS mirror can also be replicated and realized in PMOS form.

13.5.1 Generating Multiple Outputs

One nice observation is that a current mirror can have multiple outputs. Think of the NMOS diode as a voltage V_{GS} reference. We can "export" this reference to multiple transistors in parallel, and each will carry a copy of the reference current, appropriately scaled if desired. The same reference can then be fed into a PMOS mirror, and the PMOS mirror can likewise produce multiple copies of the current. These currents can be mirrored to various circuit building blocks as required. As illustrated in *Fig. 13.18*, take note that only a single precision I_{ref} is needed to generate dozens or even hundreds of copies. In *Sec. 13.7* we will discuss one simple way to generate this reference that is independent of the supply voltage.

13.6 Example: Source-Follower with Real Current Source

In this example, we will **bias** a source-follower (common drain amplifier) using a current mirror, as shown in *Fig. 13.19*. The gate bias voltage $V_{bias,Q}$ needs to be large enough so that the resulting V_{DS} on the drain of the mirror does not "squash" the transistor into the triode region. Other than that there is a lot of flexibility, making this circuit very useful. In multi-stage amplifiers, the DC bias of the previous stage driving this circuit just needs to be large enough, but the exact value does not matter.

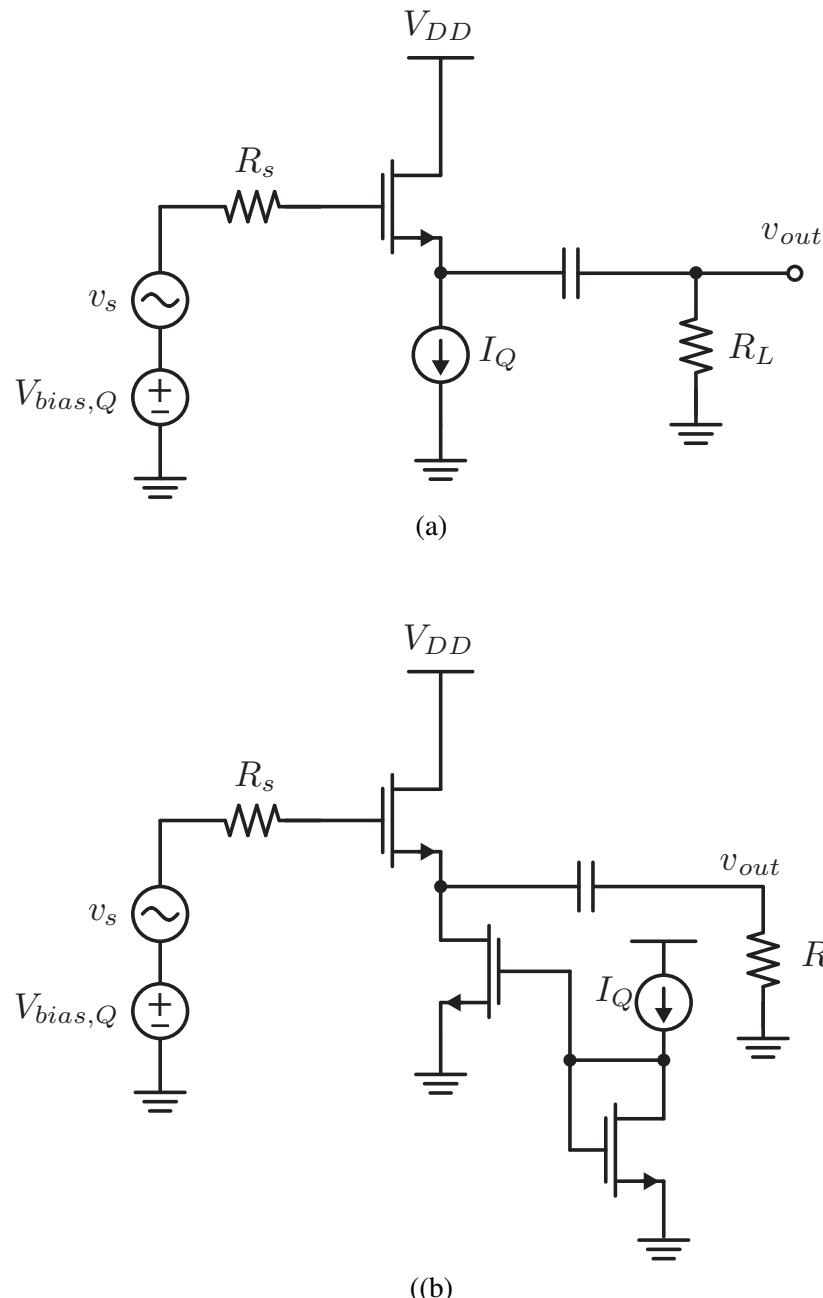


Figure 13.19: (a) A MOS source-follower amplifier using an ideal current source is replaced with (b) a current mirror biased amplifier.

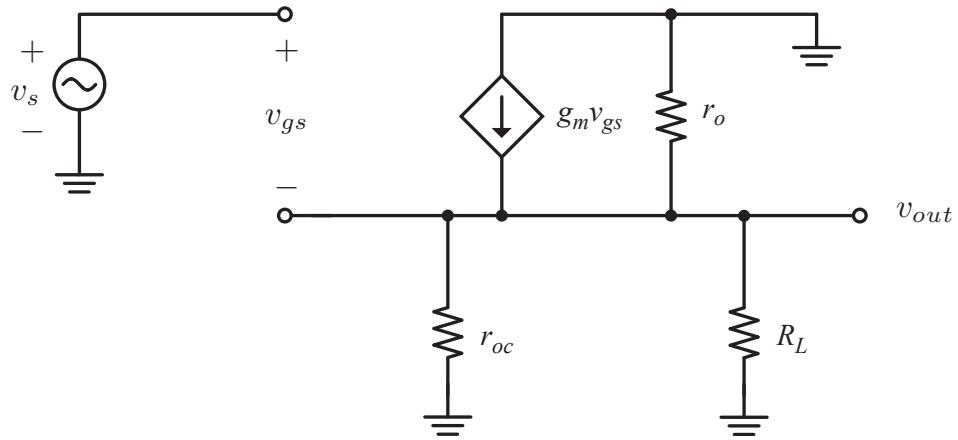


Figure 13.20: Small-signal model of a source-follower biased with a current mirror, modeled as r_{oc} .

13.6.1 Common Drain AC Schematic

The analysis of the AC response using a small-signal model is easily accomplished by simply noting that the current source can be modeled with an effective output resistance r_{oc} , as shown in Fig. 13.20.

13.6.2 CD Voltage Gain With Real Current Source

For an ideal current source, we have:

$$\frac{v_{out}}{R_L \parallel r_o} = g_m v_{gs} = g_m(v_s - v_{out}) \quad (13.19)$$

With the NMOS current mirror biasing scheme, all we have to do is add the effect of r_{oc} :

$$\frac{v_{out}}{R_L \parallel r_o \parallel r_{oc}} = g_m v_{gs} = g_m(v_s - v_{out}) \quad (13.20)$$

Let's define an effective load resistance:

$$R_{L_{eff}} = R_L \parallel r_o \parallel r_{oc} \quad (13.21)$$

From KCL at the output we have:

$$v_{out} = g_m R_{L_{eff}} (v_s - v_{out}) \quad (13.22)$$

$$= g_m R_{L_{eff}} v_s - g_m R_{L_{eff}} v_{out} \quad (13.23)$$

Rearranging, and factoring v_{out} :

$$v_{out} (1 + g_m R_{L_{eff}}) = g_m R_{L_{eff}} v_s \quad (13.24)$$

Thus, the voltage gain is given by:

$G_v = \frac{v_{out}}{v_s} = \frac{g_m R_{L_{eff}}}{1 + g_m R_{L_{eff}}}$

CD voltage gain with current source (13.25)

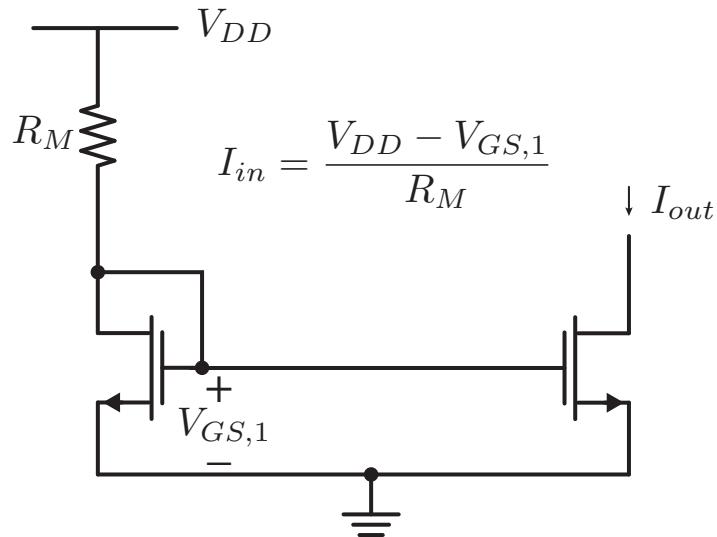


Figure 13.21: A reference current generator using a resistor. While it's very simple and convenient, it's supply and temperature dependent.

13.7 Generation of Current References

Up until now we have been ignoring the elephant in the room, which is how to generate the I_{ref} or input current into the mirror. The simplest way is to simply use a resistor, as shown in *Fig. 13.21*. This solution is simple, but it depends on several parameters that may vary.

Some of these parameters include the supply voltage, the transistor threshold voltage V_T , and many other parameters such as temperature. The reference current generated by using a resistor is found with:

$$I_{ref} = \frac{V_{DD} - V_{GS,1}}{R_M} \quad (13.26)$$

Note that $V_{GS,1}$ is a weak function of I_{ref} :

$$V_{GS,1} = V_T + \sqrt{\frac{2I_{ref}}{\mu C_{ox} \frac{W}{L}}} \quad (13.27)$$

We can solve the above equation to obtain I_{ref} .

13.7.1 Constant G_m Reference Current

What we desire is a supply and temperature independent reference current. To the greatest extent possible, we want a current that does not depend on transistor parameters, which vary over process and temperature. There are many clever circuits, such as band gap references, that generate precision temperature independent voltages. However, in order to generate a current we need a resistor. Furthermore, unless it is an external precision resistor, there will be variations. Calibration procedures are needed to tune the currents, a subject beyond the scope of this book.

Here we will not delve into the details of various circuits to build references. Instead we will present a very useful example, which is the constant- g_m reference generator, shown in *Fig. 13.22*. This example is a particularly good one because it is simple, and it builds on the principles we have already discussed in this chapter.

With reference to *Fig. 13.22*, notice that the PMOS mirror on top enforces current equality between the two branches, as the devices have the same dimensions. In other words, the currents

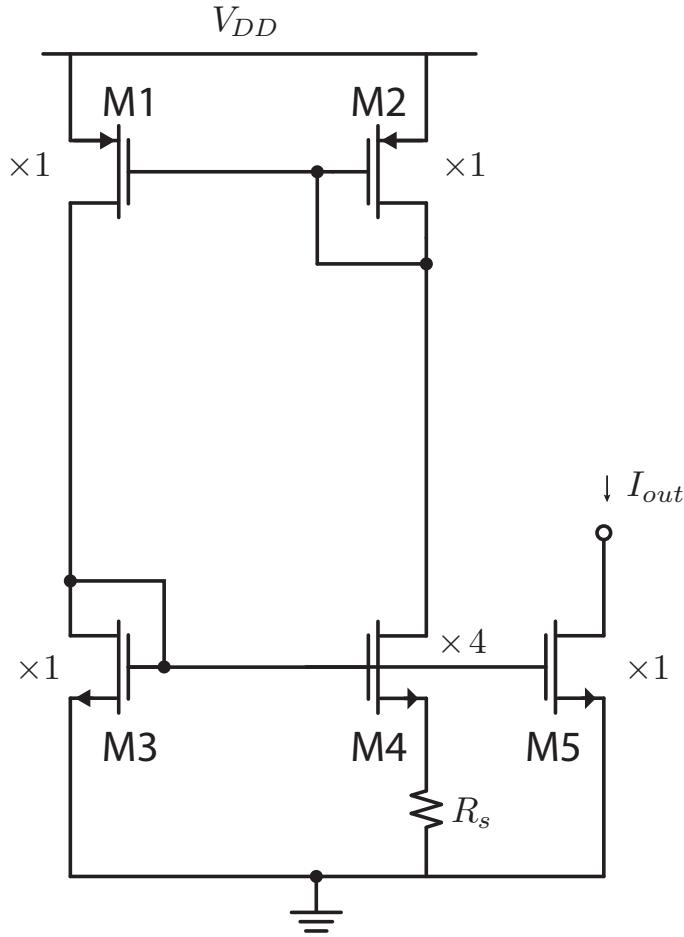


Figure 13.22: A reference current generator with an output current that produces a constant $g_m = 1/R_s$, independent of supply voltage and temperature.

I_{D_3} and I_{D_4} must be equal. On the other hand, the bottom current sources are not sized the same, and transistor $M3$ is sized to be exactly $1/4$ smaller than $M4$. The voltage across the resistor R_s is given by:

$$V_{R_s} = \Delta V_{GS} = V_{GS_3} - V_{GS_4} \quad (13.28)$$

We can express each V_{GS} as a threshold voltage plus an overdrive voltage:

$$V_{GS_4} = V_T + \sqrt{\frac{2I_{D_4}}{\mu C_{ox} \left(\frac{W}{L}\right)_4}} = V_T + V_{OD_4} \quad (13.29)$$

For device $M3$, it carries the same current and has a 4 times smaller width:

$$V_{GS_3} = V_T + \sqrt{\frac{2I_{D_4}}{\mu C_{ox} \left(\frac{W}{4L}\right)_4}} = V_T + 2V_{OD_4} \quad (13.30)$$

With this observation, the output current of $M4$ is simply given by:

$$I_{D_4} = \frac{\Delta V_{GS}}{R_s} = \frac{V_{OD_4}}{R_s} \quad (13.31)$$

Substitution of the overdrive voltage leads to the output current:

$$I_{D4} = \frac{V_{OD4}}{R_S} = \frac{1}{R_S} \sqrt{\frac{2I_{D4}}{\mu C_{ox} \left(\frac{W}{L}\right)_4}} \quad (13.32)$$

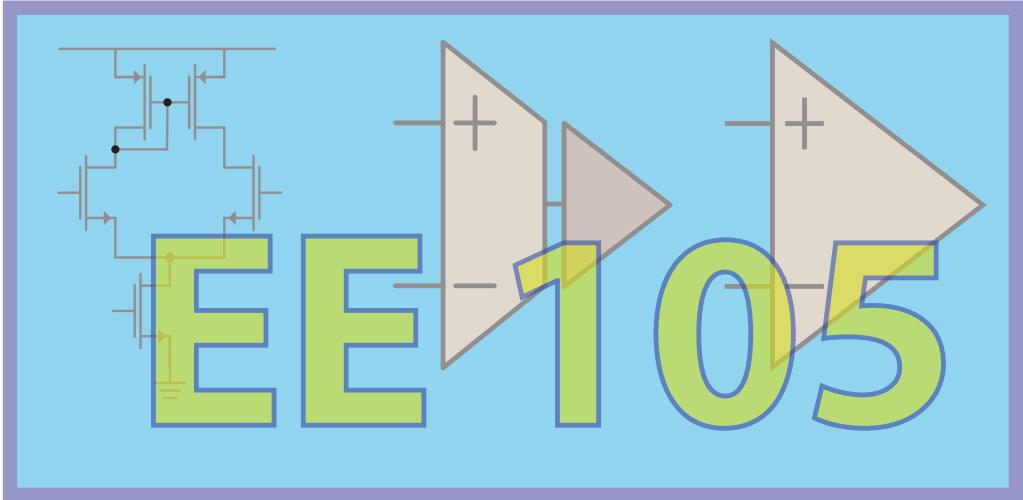
Or, by squaring both sides of Eq. 13.32, and dividing both sides by I_{D4} :

$$I_{D4} = \frac{1}{R_S^2} \frac{2}{\mu C_{ox} \left(\frac{W}{L}\right)_4} \quad (13.33)$$

Notice that the current I_{D4} does not depend on the supply voltage or the device threshold, so it can be used as a reference current. If we mirror the current of $M3$ to another transistor of equal size, it will satisfy Eq. 13.31. This is a nice circuit, because the ratio of current to overdrive voltage, which is related to the transconductance, tracks R_S :

$$g_{m5} = \frac{2I_{D5}}{V_{OD5}} = \frac{2I_{D4}}{2V_{OD4}} = \frac{1}{R_S} \quad (13.34)$$

For this reason, this circuit is known as a **constant transconductance reference**, because all transistors biased with this current will have a g_m that is constant and independent of process and temperature.



14. Frequency Response of Amplifiers

14.1 Chapter Preview

In this chapter we will be looking at various techniques to calculate the frequency response of an amplifier. We begin by reviewing the MOS transistor parasitics, and discuss passband and DC coupled amplifier response. Next we use AC analysis to find the poles and zeros of the common source amplifier transfer function, and discover that it is actually quite complicated, even for a system with two nodes. We will introduce approximations and techniques that allow us to estimate the dominant poles in the frequency response. The Miller approach allows us to get rid of pesky coupling capacitors that complicate the analysis, and most importantly, it will give us insights on when we must account for these capacitors and when we can ignore them. Finally, the method of *Open Circuit Time Constants* is a great way to analyze a complicated circuit with many capacitors, and we'll learn this technique in detail.

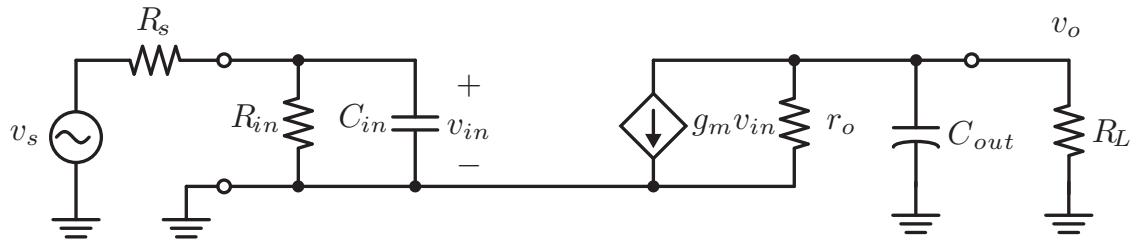


Figure 14.1: A generic transconductance amplifier with input capacitance C_{in} and output capacitance C_{out} at the input and output nodes.

14.2 Frequency Response General Considerations

For any circuit transfer function, the number of poles is a property of the circuit, and does not depend on the choice of input/output terminals. The number of *poles* is given by the *number of independent capacitors and inductors*. The zeros, on the other hand, depends on the transfer function, in other words the *choice of the input and output ports*.

In this chapter we will be pursuing techniques to discover the poles of an amplifier. In many cases, the poles are actually easy to find and can be done by "inspection". Consider the example shown in Fig. 14.1, consisting of an input capacitance C_{in} and an output capacitance C_{out} . We expect the circuit to have two poles, and we can almost read-off the poles by noting the input capacitance forms a low-pass filter. The input pole is given by:

$$\frac{1}{R_{eq} C_{in}}$$

R_{eq} is the equivalent resistance seen by the input capacitance. To see this, note that:

$$v_{in} = \frac{Z_{in}}{Z_{in} + R_s} v_s \quad (14.1)$$

Z_{in} is given by the transfer function:

$$Z_{in} = \frac{R_{in}}{1 + j\omega R_{in} C_{in}} \quad (14.2)$$

Substitution of Eq. 14.2 into Eq. 14.1 results in:

$$v_{in} = v_s \left(\frac{R_{in}}{1 + j\omega R_{in} C_{in}} \right) \cdot \frac{1}{\left(\frac{R_{in}}{1 + j\omega R_{in} C_{in}} \right) + R_s} \quad (14.3)$$

$$= \frac{R_{in}}{R_{in} + R_s (1 + j\omega R_{in} C_{in})} v_s \quad (14.4)$$

Simplifying the transfer function, it is clear that the pole is due to C_{in} and the parallel combination of R_s and R_{in} :

$$v_{in} = v_s \left(\frac{R_{in}}{R_{in} + R_s} \right) \frac{1}{1 + j\omega \left(\frac{R_{in} C_{in}}{R_{in} + R_s} \right)} \quad (14.5)$$

$$= v_s \left(\frac{R_{in}}{R_{in} + R_s} \right) \frac{1}{1 + j\omega (R_s \parallel R_{in} C_{in})} \quad (14.6)$$

Next, let's examine the output of the amplifier. First we note that output current flows into an impedance Z_{out} , which has a pole at a frequency given by $1/R_{out}C_{out}$. R_{out} is the equivalent resistance seen by C_{out} . To show this, we again write the full transfer function:

$$v_o = -g_m v_{in} Z_{out} \quad (14.7)$$

The output impedance, Z_{out} , is given by:

$$Z_{out} = \frac{r_o \parallel R_L}{1 + j\omega C_{out}(r_o \parallel R_L)} \quad (14.8)$$

Putting these results together, we have two independent poles, one from the input and one from the output:

$$v_o = -g_m \cdot \left[v_s \left(\frac{R_{in}}{R_{in} + R_s} \right) \frac{1}{1 + j\omega(R_s \parallel R_{in}C_{in})} \right] \cdot \left[\frac{r_o \parallel R_L}{1 + j\omega C_{out}(r_o \parallel R_L)} \right] \quad (14.9)$$

Finally, we have the gain of this circuit:

$$\boxed{\frac{v_o}{v_s} = \left(\frac{R_{in}}{R_{in} + R_s} \right) \frac{-g_m(r_o \parallel R_L)}{(1 + j\omega C_{out}(r_o \parallel R_L))(1 + j\omega(R_s \parallel R_{in}C_{in}))}} \quad (14.10)$$

You should learn to identify these independent poles in the transfer function of amplifiers. This will save you a lot of time and give you a first order solution to many problems. What we cover for the rest of the chapter are poles that arise from "**coupling**" capacitors, which couple charge from one node to another. These capacitors are "pesky" and require special attention.

14.3 Review MOS Parasitic Capacitors

Recall from *Ch. 10* that a MOS device has many internal capacitors that we have been ignoring up to now when calculating the *low-frequency* or *pass-band response*. As shown in *Fig. 14.2a*, these capacitors are either intrinsic to the operation of the device, such as C_{ox} , or are present due to parasitic junctions and overlap/fringing between metals and contacts and the source/drain. In the saturation region, we calculate the value of each capacitor and use the AC small-signal model, such as the three-terminal version shown in *Fig. 14.2b*.

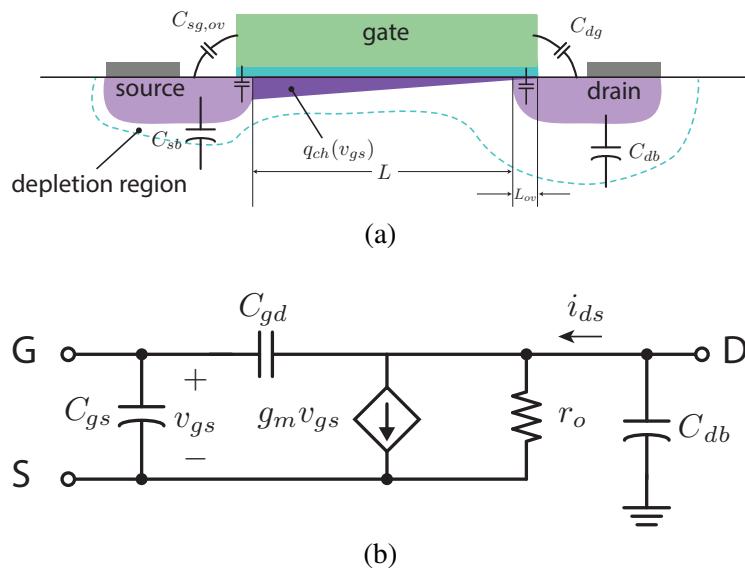


Figure 14.2: (a) Cross section of a MOS transistor in saturation and (b) the small-signal model including device capacitors.

14.4 Common-Source Amplifier Frequency Response

14.4.1 Common Source Amplifier: De-Coupling and Coupling Capacitors

Let's revisit the *common source* amplifier, shown in *Fig. 14.3a*. Since the load is referenced to ground, we cannot connect it directly to the transistor (without adversely affecting the operating point) so a large coupling capacitor is used. Also, R_{deg} is used for biasing the transistor, but it impairs the gain, so it is "bypassed" as well. Note that the large coupling capacitors (degeneration and coupling) are shorts at AC frequencies, and can be neglected when calculating the passband and high-frequency response, as shown in *Fig. 14.3b*.

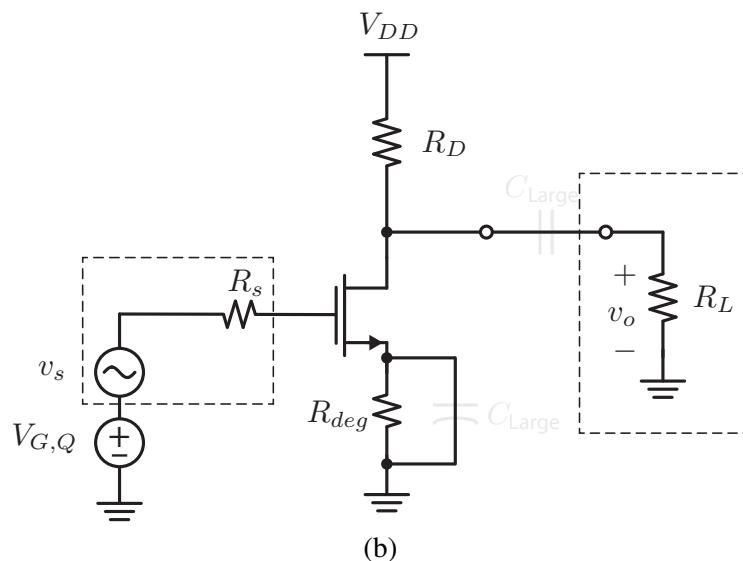
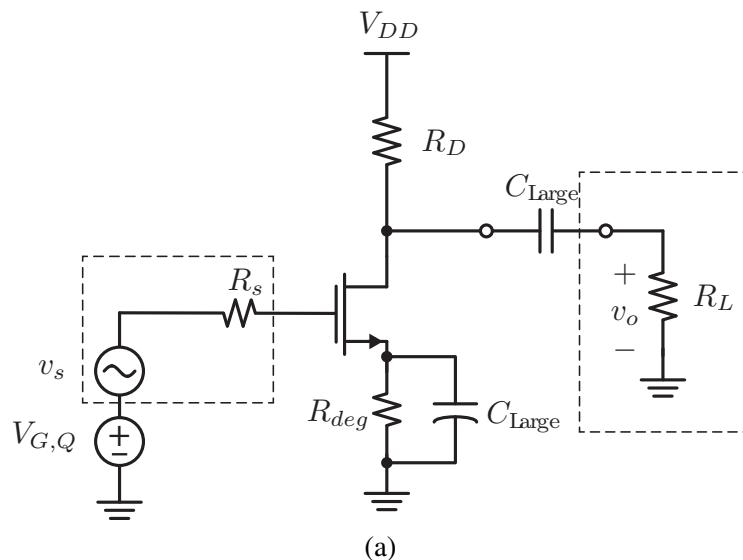


Figure 14.3: (a) Complete schematic of a discrete common source amplifier including biasing elements. (b) AC equivalent circuit at mid-band frequencies.

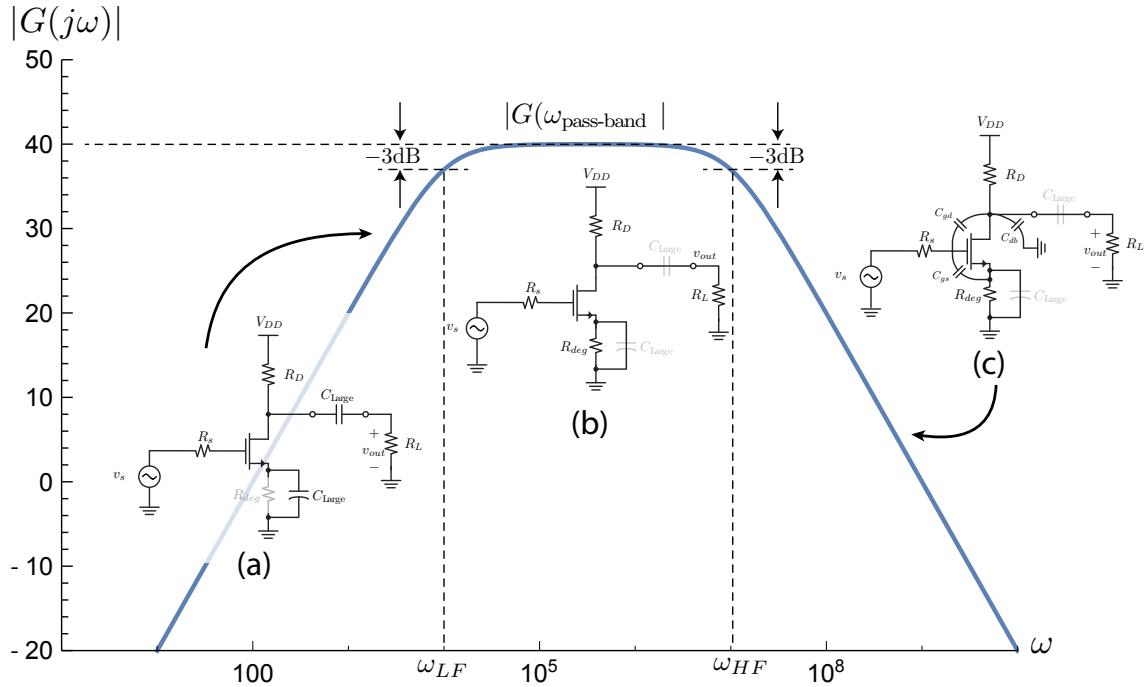


Figure 14.4: The bandpass response of a common source amplifier highlighting the various regions and the circuit elements that dominate the frequency response.

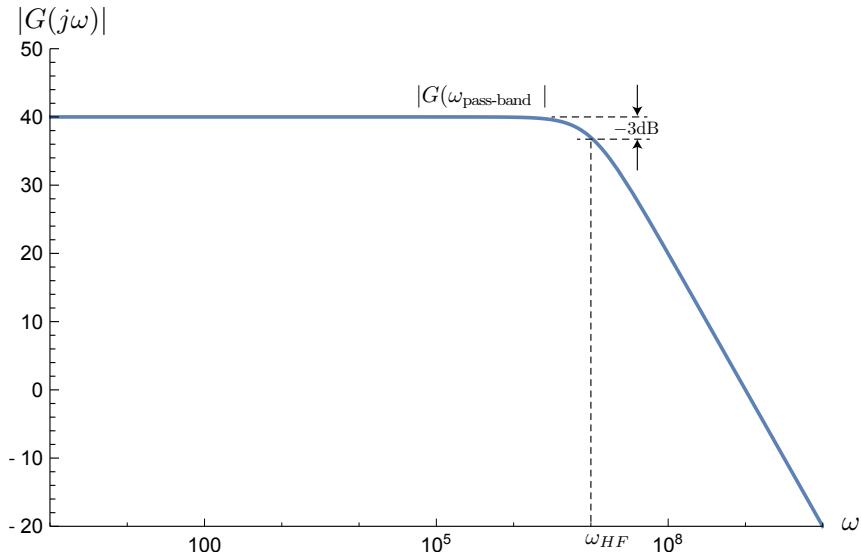


Figure 14.5: A DC coupled amplifier frequency response is low-pass in nature and is characterized by high frequency poles that cause the transfer function to roll-off.

14.4.2 Typical Passband Frequency Response

In Fig. 14.4, we see a plot of a typical amplifier transfer functions, such as the CS discussed earlier. We have a bandpass response, with a relatively flat region over a mid-range of AC frequencies, and gain drop-off at both low and high frequencies. At low frequencies, small capacitors are open but large coupling capacitors result in zeros in the transfer function, and kill the gain. In the mid-band, the large capacitors are modeled as shorts. In the high frequency region, the poles from the small capacitors determine the frequency roll-off characteristics, and reduce the gain. Our goal for this

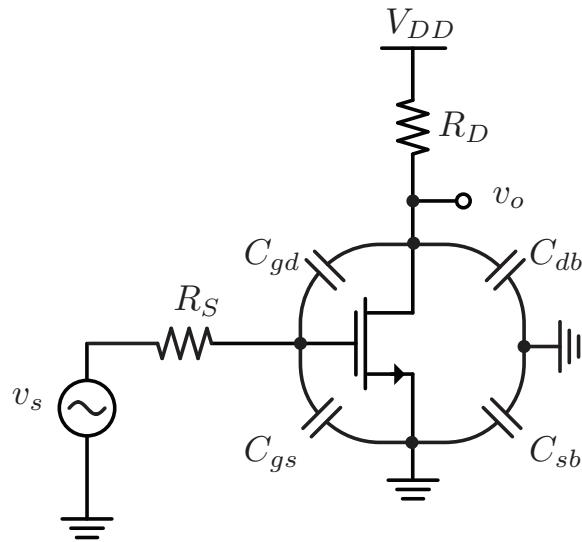


Figure 14.6: The AC schematic of a common source amplifier with device capacitors.

chapter is to find the high frequency poles. Mostly we will use techniques that give us the *dominant* pole for multi-stage amplifiers, or the lowest frequency pole that gives rise to the high frequency roll-off.

In a **DC coupled amplifier** response shown in *Fig. 14.5*, the gain does *not* fall off at low frequencies, and the mid-band gain extends down to zero frequency. This is very desirable, and we prefer DC coupled amplifiers over AC coupled ones. The issue with DC coupling is that we have to ensure the DC bias points of cascaded amplifiers are well matched. In Chapter 16 we will study differential amplifiers that function over a wide range of DC voltages, allowing us to cascade the amplifiers without using large AC coupling capacitors.

14.4.3 Common-Source Voltage Amplifier

The AC model of the *common source* amplifier, shown in *Fig. 14.6*, has many capacitors. C_{sb} is connected to ground on both sides, therefore it can be ignored. The small-signal model, shown in *Fig. 14.7*, has two nodes. Note that we cannot solve this problem by inspection, as we did for the circuit of *Fig. 14.1*, because of the presence of the capacitance C_{gd} . We can solve problem directly by nodal analysis, and this is a good approach if the circuit is "small". This will lead to the complete transfer function of the circuit.

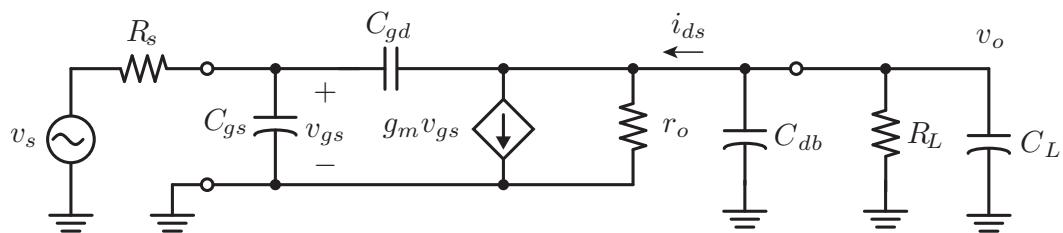


Figure 14.7: Small-signal model of a common source amplifier with device capacitors and load capacitance.

14.5 Common-Source: Brute Force Frequency Response Calculation

So let's try to analyze the transfer function, as it only has two nodes. This looks "easy", but don't forget that every capacitor leads to a pole. With three capacitors present, we're dealing with a cubic transfer function! For now we will ignore $C_{db} + C_L$ to simplify the math to a second order transfer function. Let $R'_L = r_o \parallel R_L$ and $s = j\omega$. First, we write KCL at each node of the circuit:

$$\frac{v_{gs} - v_s}{R_s} + sC_{gs}v_{gs} + (v_{gs} - v_o)sC_{gd} = 0 \quad (14.11)$$

$$\frac{v_o}{R'_L} + g_m v_{gs} + (v_o - v_{gs})sC_{gd} = 0 \quad (14.12)$$

Let's combine these two equations into a matrix:

$$\begin{bmatrix} 1 + sR_s(C_{gs} + C_{gd}) & -sC_{gd}R_s \\ g_m R'_L - sC_{gd}R'_L & 1 + sR'_L C_{gd} \end{bmatrix} \cdot \begin{bmatrix} v_{gs} \\ v_o \end{bmatrix} = \begin{bmatrix} v_s \\ 0 \end{bmatrix} \quad (14.13)$$

Using **Cramer's rule** or another approach, it is easy to show that the transfer function is given by:

$$v_o = v_s \left(\frac{-g_m R'_L \left(1 - \frac{sC_{gd}}{g_m} \right)}{1 + s(R'_L C_{gd} + R_s(C_{gs} + C_{gd}) + R_s C_{gd} g_m R'_L) + s^2(R'_L R_s C_{gd} C_{gs})} \right) \quad (14.14)$$

To check this result, let's examine the low-frequency gain by plugging $s = 0$ into Eq. 14.14:

$$\frac{v_o}{v_s}(s=0) = \frac{-g_m(r_o \parallel R_L)(1-0)}{(1+0+0)} = \boxed{-g_m(r_o \parallel R_L)} \quad (14.15)$$

The result in Eq. 14.15 is consistent with our previous analysis of the mid-band gain. Note that transfer function has a zero, a frequency that causes the transfer function to go to zero:

$$\omega_z = \frac{g_m}{C_{gd}} \quad (14.16)$$

The origin of this zero is due to the feed through capacitor C_{gd} , which adds a signal to the output with opposite phase as the g_m . So at a high enough frequency, this feed-forward current cancels the gain of the amplifier. We would like to put the transfer function into the following form:

$$\frac{v_o}{v_s} = \frac{-g_m(r_o \parallel R_L)(1-j\omega/\omega_z)}{(1+j\omega/\omega_{p1})(1+j\omega/\omega_{p2})} \quad (14.17)$$

Because the denominator of transfer function is second order, we can solve it exactly with the quadratic formula, but the results are messy. We will make an *approximation that the poles are widely separated*, in other words $\omega_{p1} \ll \omega_{p2}$. This approximation is valid in most circumstances and leads to a simpler result:

$$\frac{v_o}{v_s} = \frac{-g_m(r_o \parallel R_L)(1-j\omega/\omega_z)}{1 + j\omega(1/\omega_{p1} + 1/\omega_{p2}) + (j\omega)^2/(\omega_{p1}\omega_{p2})} \approx \frac{-g_m(r_o \parallel R_L)(1-j\omega/\omega_z)}{1 + j\omega(1/\omega_{p1}) + (j\omega)^2/(\omega_{p1}\omega_{p2})} \quad (14.18)$$

This approximation allows us to quickly estimate the "dominant" pole by equating the Eq. 14.18 with Eq. 14.14:

$$\boxed{\omega_{p1} \approx \frac{1}{R_s} \left(C_{gs} + (1 + g_m R_o) C_{gd} \right) + R_o C_{gd}} \quad \text{Dominant pole approximation} \quad (14.19)$$

The result is not exact, since we neglected C_L and C_{db} , and provides little insight. These poles are calculated after doing a lot of algebraic manipulations on the transfer function, even though we are only dealing with a second order transfer function. There must be an easier and better way!

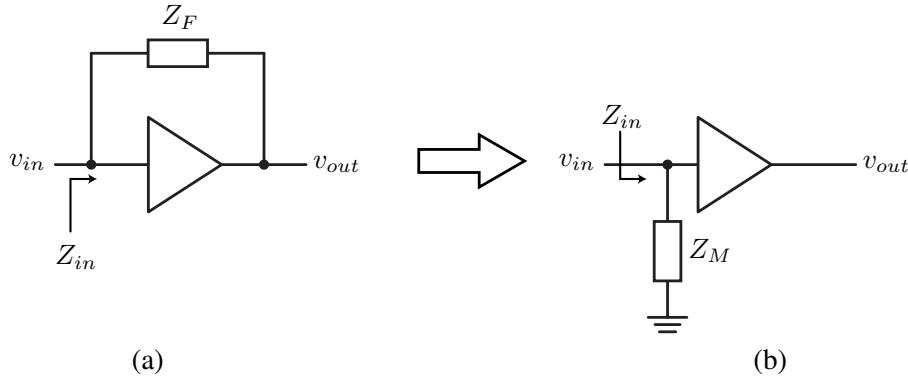


Figure 14.8: The Miller Theorem allows us to replace an impedance Z_F coupling two nodes with a simpler impedance to ground Z_M .

14.6 The Miller Theorem

14.6.1 The Miller Effect

Consider an ideal amplifier with gain A and a feedback impedance Z_F from input to output, as shown in Fig. 14.8a. Let's calculate the input impedance of this circuit, assuming the amplifier has infinite input impedance. In that case, the only current flow is due to Z_F :

$$i_{in} = \frac{v_{in} - v_{out}}{Z_F} \quad (14.20)$$

Since the output is a scaled copy of the input (i.e. $v_{out} = A \cdot v_{in}$):

$$i_{in} = \frac{v_{in} - A \cdot v_{in}}{Z_F} \quad (14.21)$$

Notice that the above equation only involves the input node. We can therefore pretend the current flows to ground as shown in *Fig. 14.8b*:

$$i_{in} = \frac{v_{in}}{\frac{Z_F}{1-A}} \quad (14.22)$$

In other words, we can replace the feedback Z_F with a (smaller) impedance to ground (assuming A is negative):

$$Z_M = \frac{Z_F}{1-A} \quad (14.23)$$

If this seems like a "sleight of hand", keep in mind that in practice Z_F will also affect the gain A , so technically we don't really know A . The trick is to estimate A , especially when the gain is not a strong function of Z_F .

14.6.2 The Miller Effect of a Capacitor

Applying the **Miller theorem** to a capacitor, shown in *Fig. 14.9*, we have a capacitor that couples two nodes. In this case, it is across an inverting amplifier, and so the effective capacitance is boosted:

$$Z_M = \frac{1}{j\omega C_M} = \frac{\frac{1}{j\omega C_{gd}}}{1 - A} \quad (14.24)$$

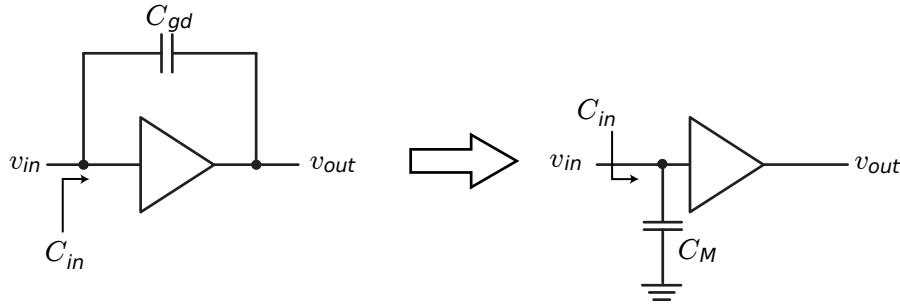


Figure 14.9: The Miller Theorem is particularly useful for analyzing the frequency response of an amplifier when coupling capacitors such as C_{gd} appear.

For example, in a common-source amplifier, the capacitor is C_{gd} , the pesky capacitor that complicated our analysis of the frequency response. Applying the Miller Theorem to the capacitive impedance:

$$Z_M = \frac{1}{j\omega C_{gd}(1-A)} \quad (14.25)$$

This implies that we can replace C_{gd} with a capacitor at the input to ground of larger magnitude:

$$C_{in} = C_M = C_{gd}(1-A) \quad (14.26)$$

Note that for a common-source amplifier A is negative, so the overall capacitance is positive. If the voltage gain is large, then the capacitance is boosted to a much larger value to ground. The physical reason is not too difficult to understand if you consider that in a capacitor current flows due to the rate of change of the voltage difference across the capacitor. An amplifier with negative gain will boost this voltage difference, causing one terminal to move with a much larger amplitude compared to the other, and a correspondingly larger input current is drawn.

14.7 Common-Source: Miller Approach Frequency Response Calculation

Now we apply the Miller theorem to the C_{gd} feedback capacitor in the MOS amplifier, shown in Fig. 14.10:

$$C_{in} = \frac{1}{j\omega C_M} = \left(\frac{1}{1-A_{v,C_{gd}}} \right) \left(\frac{1}{j\omega C_{gd}} \right) = \frac{1}{j\omega [(1-A_{v,C_{gd}}) C_{gd}]} \quad (14.27)$$

Using this input capacitance of Eq. 14.27, we modify the small-signal model as shown in Fig. 14.11. This circuit is much simpler than the original amplifier, because the coupling capacitor is removed and the two nodes are seemingly independent. The only trouble is that $A_{v,C_{gd}}$ is not known exactly. However, we take the low-frequency value as an approximation:

$$A_{v,C_{gd}} \approx -g_m R_o \quad (14.28)$$

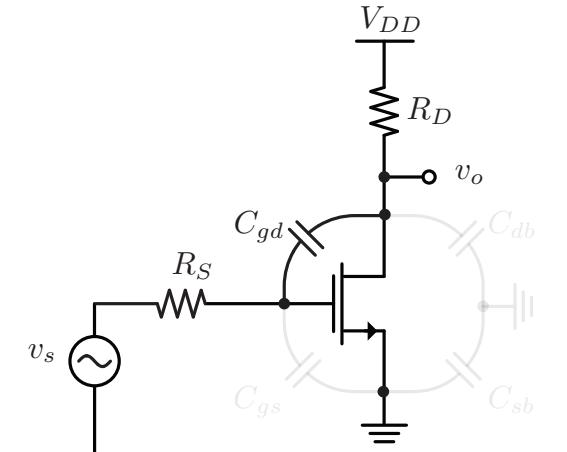
Using the Miller approximation result, we calculate the RC time constant of the input pole by inspection:

$$\omega_{p1}^{-1} = R_s [C_{gs} + (1 + g_m R_o) C_{gd}] \quad (14.29)$$

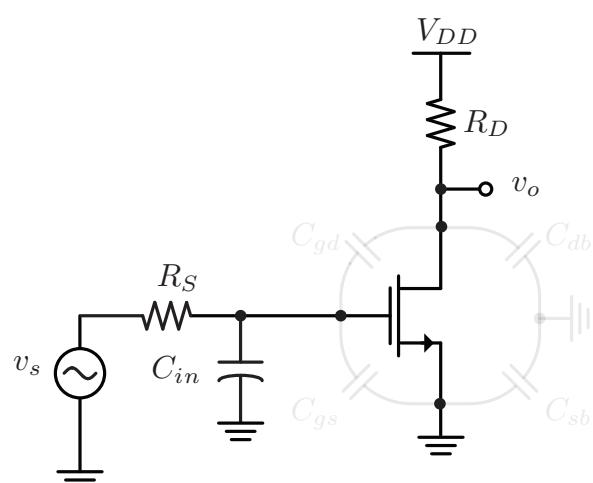
If we compare this to the previously derived result, we have:

$$\omega_{p1}^{-1} = R_s [C_{gs} + (1 + g_m R_o) C_{gd}] + R'_L C_{gd} \quad (14.30)$$

The results are very close except for the last term, which usually does not dominate. The **Miller effect** is much simpler, and as an added bonus, it gives us insight into the gain-bandwidth trade-off of the amplifier. If we boost the gain of the common source amplifier, the input pole drops. This is one of the main limitations of a common source amplifier, something we will address later when we discuss multi-stage amplifiers.



(a)



(b)

Figure 14.10: (a) In a common source amplifier, C_{gd} is the most difficult capacitor to account for as it couples two nodes. (b) Using the Miller Theorem, we replace it with an equivalent input capacitance C_{in} .

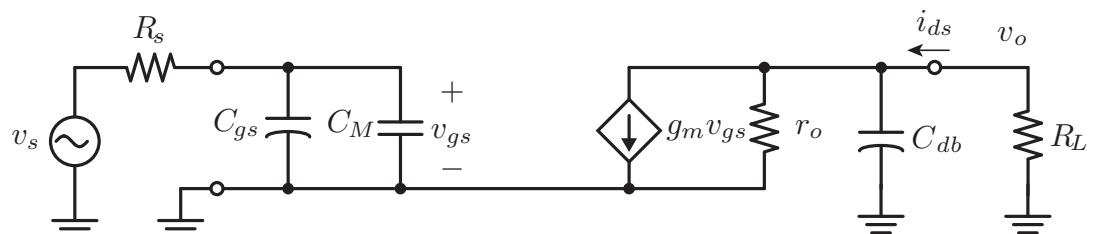


Figure 14.11: Small-signal model of a common source amplifier with C_{gd} replaced by a Miller capacitor C_M .

14.8 Common Drain Amplifier Frequency Response

Let's now compute the bandwidth of a common drain amplifier, also known as the source follower, shown in *Fig. 14.12a*. We can replace the current source with MOSFET-based current mirror, and we simply model it by r_{oc} . In the small-signal model with capacitors shown in *Fig. 14.12b*, we use the DC small-signal gain and the Miller effect to calculate the input capacitance. This will allow us to estimate the dominant pole.

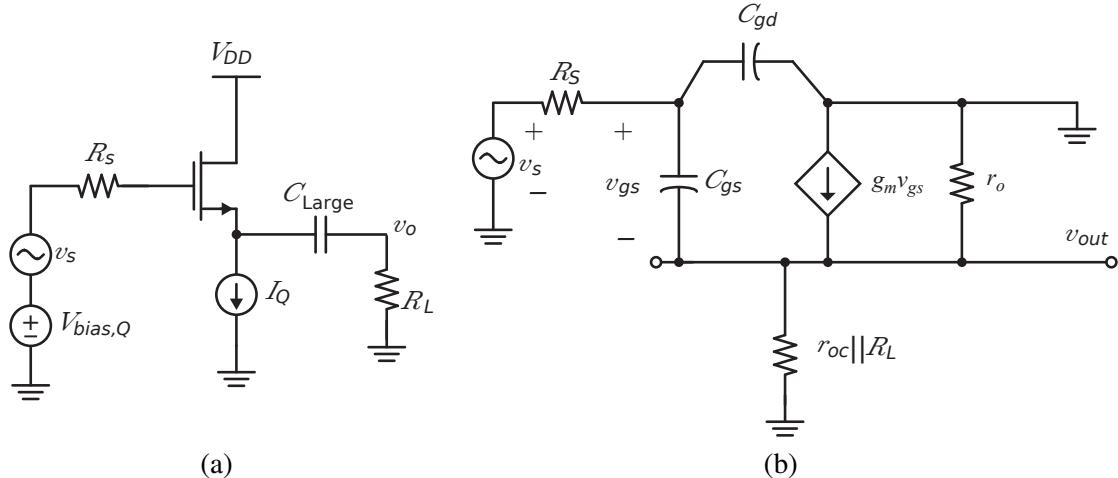


Figure 14.12: (a) Schematic of a common drain amplifier. (b) Small-signal AC model of the common drain amplifier.

14.8.1 Voltage Gain Across C_{gs}

Note that the **Miller capacitor** in this case is C_{gs} , the gate-source capacitor is between the input and output. Recall the gain of a common drain amplifier:

$$\frac{v_{out}}{v_{in}} = \frac{g_m}{\frac{1}{r_o \parallel r_{oc}} + g_m} = \frac{g_m (r_o \parallel r_{oc})}{1 + g_m (r_o \parallel r_{oc})} = A_{V,C_{gs}} \quad (14.31)$$

We can therefore replace the capacitance at the input with an effective capacitance:

$$C_{in} = C_{gd} + C_M \quad (14.32)$$

C_M is the Miller multiplied capacitor, C_{gs} :

$$C_{in} = C_{gd} + (1 - A_{V,C_{gs}}) C_{gs} \quad (14.33)$$

If we substitute the low-frequency gain $A_{V,C_{gs}}$:

$$C_{in} = C_{gd} + \left(1 - \frac{g_m (r_o \parallel r_{oc})}{1 + g_m (r_o \parallel r_{oc})} \right) C_{gs} \quad (14.34)$$

This can be simplified:

$$C_{in} = C_{gd} + \left(\frac{1}{1 + g_m (r_o \parallel r_{oc})} \right) C_{gs} \quad (14.35)$$

Notice that the second term is small if $g_m (r_o \parallel r_{oc})$ is large:

$$C_{in} \approx C_{gd} \quad (14.36)$$

What just happened? The Miller capacitor nearly vanishes in this case, because it is "**bootstrapped**". A bootstrapped capacitor means that the gain across it is nearly unity, and it has a positive sign. As a result, very little current flows through it.

14.8.2 Bandwidth of Source Follower

The input low-pass filter's -3 dB frequency is given by:

$$\omega_p^{-1} = R_S \left(C_{gd} + \frac{C_{gs}}{1 + g_m(r_o \parallel r_{oc})} \right) \quad (14.37)$$

Let's substitute some favorable values of R_S and r_o :

$$R_S \approx \frac{1}{g_m} \quad (14.38)$$

$$r_o \gg \frac{1}{g_m} \quad (14.39)$$

This gives:

$$\omega_p^{-1} \approx \left(\frac{1}{g_m} \right) \left(C_{gd} + \frac{C_{gs}}{\cancel{1+LARGE}} \right)^0 \quad (14.40)$$

$$\approx \frac{C_{gd}}{g_m} \quad (14.41)$$

If we simplify this, we see that:

$\omega_p \approx \frac{g_m}{C_{gd}}$

Common drain bandwidth
(14.42)

As we will shortly show, this is a very high frequency, much higher than the validity of our approximations and models. The net result is the input pole is likely high impedance, especially if driven by a moderate source impedance.

14.8.3 Miller Summary

Let's summarize how the Miller effect played out in these two very different scenarios:

- **Common source amplifier:** Miller multiplied capacitor has *detrimental* impact.

$A_v =$ A large negative number, say -100

$C_{Miller} = (1 - A_{V,C_{gd}})C_{gd} \approx 100C_{gd}$

- **Common drain amplifier:** "Bootstrapped" capacitor has *negligible* impact on bandwidth.

$A_v =$ slightly less than unity and positive

$C_{Miller} = (1 - A_{V,C_{gs}})C_{gs} \simeq 0$

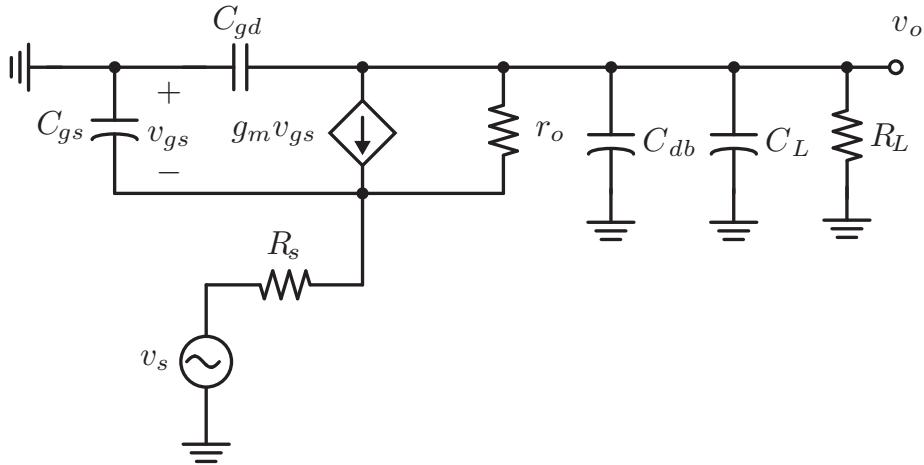


Figure 14.13: Small-signal AC model of a common gate amplifier. Note that even though there are no coupling capacitors from input to output, there is a coupling resistor r_o .

14.9 Common Gate Amplifier Bandwidth

The small-signal schematic of the common gate amplifier is shown in Fig. 14.13. While there are four capacitors in the model, three are in parallel, so we only expect two poles. Also notice that both C_{gd} and C_{gs} are grounded at one node, and so *there are no input-to-output coupling capacitors*, and hence *no Miller effect for the capacitors*. This makes the common gate amplifier in general a broadband amplifier stage. To see this, we will analyze the circuit, first using an approximation, and then using the complete equations.

To begin, observe that the resistor r_o is coupled from the input to the output, and so we could apply the Miller theorem to this resistance. What we will find is that the impedance level at the input is low enough and dominated by the low impedance presented by the transistor, so that neglecting r_o does not change impedance levels appreciably. In fact, if we ignore r_o altogether, we will find that the circuit consists of two independent nodes, and it is easily analyzed by inspection.

At the input, the pole arising from C_{gs} can be estimated by finding the equivalent resistance seen at the input node, which is dominated by the transistor $1/g_m$ in parallel with R_s :

$$\omega_1 = \frac{1}{C_{gs} \left(R_s \parallel \frac{1}{g_m} \right)} \quad (14.43)$$

At the output node, the parallel combination of $C_L + C_{gd} + C_{db}$ sees a resistance of R_L in parallel with a high impedance presented by the transistor, which we'll ignore:

$$\omega_2 = \frac{1}{(C_L + C_{gd} + C_{db})R_L} \quad (14.44)$$

The first pole can be lower bounded by:

$$\omega = \frac{g_m}{C_{gs}} \quad (14.45)$$

In the next section we will show that this frequency is a device property known as the **unity gain frequency**, and it is usually a very high frequency for modern short channel devices biased in saturation. The second pole depends on both the load capacitance and the load resistance, and component values can be selected to ensure sufficient bandwidth.

It is always good to verify our assumptions by performing a more complete analysis. The nodal equations at the input and output are given by:

$$\begin{bmatrix} (G_s + g_m + g_o + sC_{gs}) & -g_o \\ -(g_m + g_o) & (g_o + g_L + sC'_L) \end{bmatrix} \cdot \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} v_s G_s \\ 0 \end{bmatrix} \quad (14.46)$$

In writing these equations, we have used conductances the $g_o = 1/r_o$, $G_s = 1/R_s$, and an effective load capacitance of $C'_L = C_{gd} + C_{db} + C_L$. We can easily solve this 2×2 system to obtain:

$$v_2 = \left(\frac{(g_m + g_o)G_s}{(g_o + g_L + sC'_L)(G_s + g_m + g_o + sC_{gs}) + g_o(g_m + g_o)} \right) v_s \quad (14.47)$$

As expected, if we neglect the g_o terms, we get a simple separable second-order system with an input pole and an output pole. From the full equations, we can see why we are justified in neglecting g_o as it always appears in sums with much larger conductances.

For example $g_m r_o \gg 1$ implies that $g_m \gg g_o$. Let's first check the DC gain and verify that it matches our previous calculations:

$$\frac{v_2}{v_s}(s=0) = \frac{(g_m + g_o)G_s}{(g_o + g_L)(G_s + g_m + g_o) + g_o(g_m + g_o)} \quad (14.48)$$

This expression is complicated by the feed forward gain due to g_o . Neglecting g_o , we have a simplified expression that matches our simpler calculations:

$$\frac{v_2}{v_s}(s=0) \approx \left(\frac{g_m}{1 + g_m R_s} \right) R_L \quad (14.49)$$

Finally, under the $g_o = 0$ approximation, the poles are given by:

$$\boxed{\omega_1 = \frac{G_s + g_m}{C_{gs}}} \quad (14.50)$$

$$\boxed{\omega_2 = \frac{G_L}{C'_L}} \quad (14.51)$$

This result agree with our estimations, as expected.

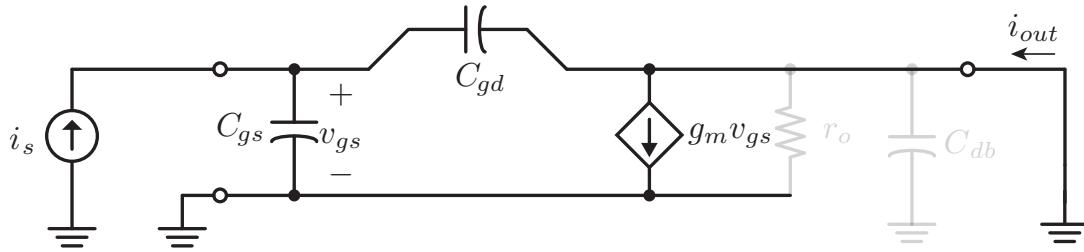


Figure 14.14: AC schematic for calculation of the current gain of a MOS transistor.

14.10 Device Unity Gain Frequency f_T

In the previous section, we encountered a lower bound for the common gate input pole:

$$\omega = \frac{g_m}{C_{gs}} \quad (14.52)$$

We claimed that this is a very high frequency. To show this, we are going to define the unity-gain frequency f_T of a transistor, a parameter that characterizes the high frequency performance of a transistor for analog circuits.

14.10.1 Unity-Gain Frequency f_T

Let's calculate the current gain of a MOS transistor by shorting the output, as shown in the setup in Fig. 14.14. Note that the input v_{gs} is given by:

$$v_{gs} = \frac{i_s}{sC_{gs} + sC_{gd}} \quad (14.53)$$

KCL at the output results in:

$$i_{out} + \frac{v_{gs}}{\frac{1}{sC_{gd}}} = g_m v_{gs} \quad (14.54)$$

This generates an output current:

$$i_{out} = g_m v_{gs} - sC_{gd}v_{gs} \quad (14.55)$$

$$\approx g_m v_{gs} \quad (14.56)$$

Thus, we have the approximation:

$$i_{out} = g_m \left(\frac{i_s}{sC_{gs} + sC_{gd}} \right) \quad (14.57)$$

The approximation is valid since the g_m of a transistor should dominate over the parasitic feedthrough current through C_{gd} . This means the current gain A_i is given by:

$$A_i = \frac{i_{out}}{i_s} = \frac{g_m}{sC_{gs} + sC_{gd}} \quad (14.58)$$

Making the substitution $s = j\omega$, and taking the magnitude of the gain:

$$\left| \frac{I_o}{I_i} \right| = \frac{g_m}{\omega(C_{gs} + C_{gd})} \quad (14.59)$$

We define $\omega_T = 2\pi f_T$ as the unity gain frequency by equating the magnitude to unity and solving for ω_T :

$$\omega_T = \frac{g_m}{C_{gs} + C_{gd}} \quad (14.60)$$

The small-signal parameters are given by:

$$C_{gs} = \frac{2}{3} W L C_{ox} \quad (14.61)$$

$$C_{gd} = \frac{2}{3} W \delta L C_{ox} \quad (14.62)$$

$$g_m = \mu C_{ox} \frac{W}{L} (V_{gs} - V_T) \quad (14.63)$$

Substituting the small-signal parameters into Eq. 14.60, we have:

$$\omega_T = \frac{\mu C_{ox} \frac{W}{L} (V_{gs} - V_T)}{\frac{2}{3} W (L + \delta L) C_{ox}} = \left(\frac{3}{2} \right) \frac{\mu (V_{gs} - V_T)}{L (L + \delta L)} \quad (14.64)$$

There are some nice physical relations that we can invoke to explain this equation. For example, the average electric field in the channel of the transistor when biased in saturation is given by:

$$\mathcal{E}_{sat} = \frac{(V_{gs} - V_T)}{L} \quad (14.65)$$

The reason for this is because $V_{DS, ch} = V_{GS} - V_T$, and so $\mu \mathcal{E}_{sat}$ is the velocity of carriers. This means that the overall expression is approximately the time it takes a carrier to cross the channel.

As gate length reduces in advanced technology nodes, the f_T increases. Today we have devices $L < 100\text{ nm}$ and $f_T > 100\text{ GHz}$ of unity gain frequency. So when we design amplifiers with bandwidths up to 100's of MHz , this unity gain frequency is quite high.

Why do we bother to define f_T ? First of all, it is an important limit for current amplifiers. In addition, we will see later that it is actually the gain-bandwidth limit for all amplifiers, including voltage amplifiers. We will find that even with broadband amplifier stages, it is hard to beat the f_T bandwidth. In advanced analog design courses, you will learn that most amplifiers are limited in gain-bandwidth to about f_T .

14.10.2 Frequency Response of Multistage Amplifiers

We have a *systematic technique* to study amplifier performance. This includes deriving the transfer function, and studying the poles and zeros in a Bode plot. In most cases, the systematic approach is too cumbersome. We have a good qualitative understanding of circuit performance (e.g., CS suffers from Miller effect, CD and CG are wideband stages). Next we will introduce the **open circuit time constants** approach, which is an analytical technique capable of estimating the dominant (lowest) pole for an amplifier with many capacitors.

14.11 The Method of Open-Circuit Time Constants (OCTC)

14.11.1 OCTC Assumptions

There are many assumptions that go into the analysis approach. First, we assume the transfer function does not have *zeros*. We will also make a *dominant pole* approximation. In other words, we assume that $\omega_1 \ll \min(\omega_2, \omega_3, \dots, \omega_n)$. Working with the all-pole transfer function:

$$H(j\omega) = \frac{H_o}{(1 + j\omega b_1 + (j\omega)^2 b_2 + (j\omega)^3 b_3 + \dots)} \quad (14.66)$$

If we factor the denominator:

$$H(j\omega) = \frac{H_o}{\left(1 + \frac{j\omega}{\omega_1}\right)\left(1 + \frac{j\omega}{\omega_2}\right)\dots\left(1 + \frac{j\omega}{\omega_n}\right)} \quad (14.67)$$

Now multiply out the denominator again:

$$H(j\omega) \approx \frac{H_o}{1 + j\omega\left(\frac{1}{\omega_1} + \frac{1}{\omega_2} + \dots + \frac{1}{\omega_n}\right) + \dots} \quad (14.68)$$

Equating *Eq. 14.66* and *Eq. 14.68*, and using $\omega_1 \ll \min(\omega_2, \omega_3, \dots, \omega_n)$, we have:

$$b_1 = \frac{1}{\omega_1} + \frac{1}{\omega_2} + \dots + \frac{1}{\omega_n} \approx \boxed{\frac{1}{\omega_1}} \quad (14.69)$$

So if we can estimate b_1 , we can also estimate the dominant pole.

14.11.2 Procedure to Find b_1

b_1 can be calculated as follows: b_1 is the sum of *open-circuit time constants* τ_i . This sum can be found by considering each capacitor C_i in the amplifier separately, and finding its Thévenin resistance R_{C_i} . Next, we calculate the time-constant of each capacitor $\tau_i = R_{C_i}C_i$ and sum them together:

$$b_1 = \sum_{i=1}^n C_i R_{C_i} \quad (14.70)$$

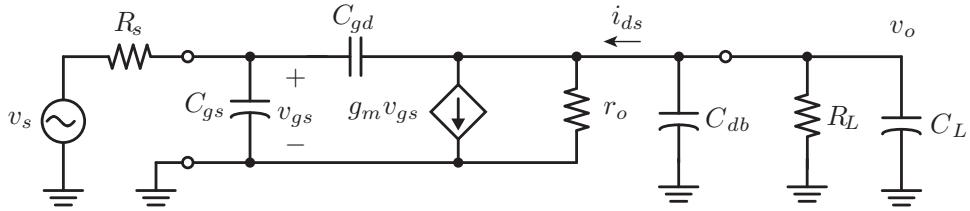
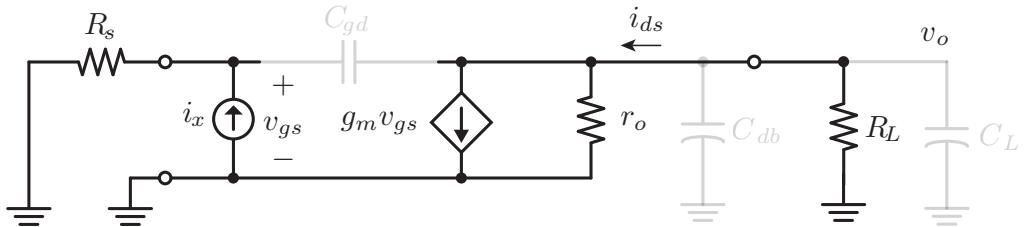
Using our result from *Eq. 14.69* leads to:

$$\omega_1 \approx \frac{1}{\sum_{i=1}^n C_i R_{C_i}} \quad (14.71)$$

For a proof, see [1].

14.11.3 Finding the Thévenin Resistance

The key to OCTC is to find the equivalent resistance seen by each capacitor. To do this, open-circuit all capacitors (i.e. remove them). For capacitor C_i , find the resistance R_{C_i} across its terminals with all independent sources removed (voltage sources shorted, current sources opened). In some cases, we can calculate the resistance by inspection. Other times, we might need to apply a test voltage (current) and find the current (voltage). This technique provides *insight for design* when one of the time constants is much larger than the others. In this case, the bandwidth of the amplifier will be limited by the capacitor with the time constant that contributes the largest $\tau = R_{C_i}C_i$, not necessarily by the largest C_i .

Figure 14.15: The complete schematic of a common source amplifier including C_L and C_{db} .Figure 14.16: Equivalent circuit for calculation of $R_{C_{gs}}$ for OCTC analysis.

14.12 Example Calculation: The Common-Source Amplifier Dominant Pole

14.12.1 Applying OCTC to CS Amplifier

Let's revisit the CS amplifier. Again, we drive it with a source impedance R_s , and load it with R_L , shown in *Fig. 14.15*. We can conveniently add a load capacitance, because it appears in parallel with C_{db} and does not complicate the circuit. It is not obvious if there is a "dominant node" in the circuit, so we must find the equivalent resistance seen by each capacitor.

14.12.2 OCTC: $R_{C_{gs}}$

Let's focus on C_{gs} , as shown in *Fig. 14.16*. Note that we open-circuit all the capacitors and replace C_{gs} with a current source to find the equivalent resistance seen by C_{gs} . Since C_{gd} is open, there is no path from the input to output, and the resistance seen by C_{gs} is simply R_s :

$$R_{C_{gs}} = R_s \quad (14.72)$$

14.12.3 OCTC: $R_{C_{gd}}$

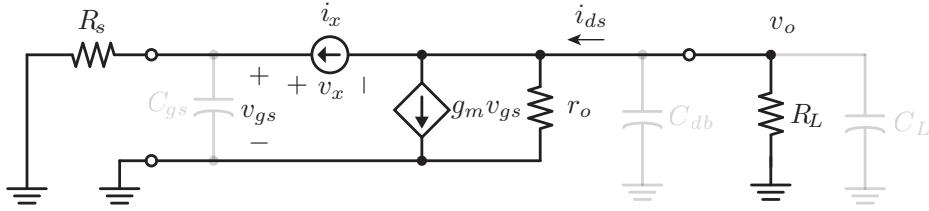
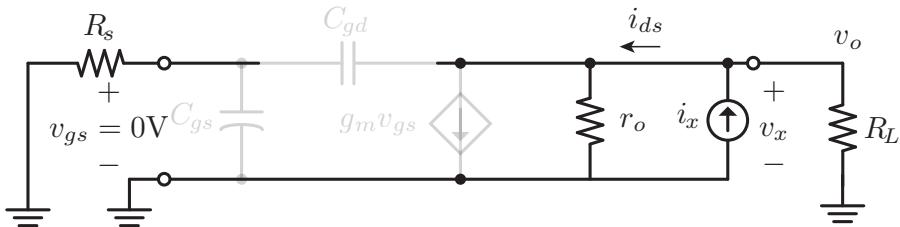
Next we focus on C_{gd} as shown in *Fig. 14.17*. We open-circuit all the capacitors, and replace C_{gd} with a current source to find the equivalent resistance seen by C_{gd} . Let $R_L' = R_L \parallel r_o$. Since the current i_x flows into the input, we have to be a bit more careful in finding the equivalent resistance seen by C_{gd} . Analyzing the circuit yields two equations:

$$i_x = -g_m v_{gs} - \frac{v_d}{R_L'} = -g_m v_{gs} - \frac{v_{gs} - v_x}{R_L'} = v_{gs} \left(-g_m - \frac{1}{R_L'} \right) + \frac{v_x}{R_L'} \quad (14.73)$$

$$i_x = \frac{v_{gs}}{R_s} \implies v_{gs} = i_x R_s \quad (14.74)$$

Substituting *Eq. 14.74* into *Eq. 14.73*, and solving for $\frac{v_x}{i_x}$, we find:

$$R_{C_{gd}} = \frac{v_x}{i_x} = R_s (1 + g_m R_L') + R_L' \quad (14.75)$$

Figure 14.17: Equivalent circuit for calculation of $R_{C_{gd}}$ for OCTC analysis.Figure 14.18: Equivalent circuit for calculation of $R_{C_{db}}$ for OCTC analysis.

14.12.4 OCTC: $R_{C_{db}}$

Finally we have C_{db} (and C_L), shown in Fig. 14.18. To calculate the resistance seen by $C_{db} + C_L$, note that with C_{gd} and C_L eliminated (open-circuit), there is no voltage at the input. Then $v_{gs} = 0V$, and so $g_m V_{gs} = 0A$, which means that the current generator is also an open. So by inspection:

$$R_{C_{db}+C_L} = R_L' \quad (14.76)$$

14.12.5 Applying OCTC to CS Amplifier

Putting all of our results together, we have the sum of the time constants:

$$\tau_H = R_s C_{gs} + \left(R_s (1 + g_m R_L') + R_L' \right) C_{gd} + R_L' C_L \quad (14.77)$$

Eq. 14.77 can be re-written as:

$$\tau_H = R_s C_{gs} + R_s (1 + g_m R_L') C_{gd} + R_L' C_{gd} + R_L' C_L \quad (14.78)$$

We can approximate Eq. 14.78:

$$\tau_H \approx R_s \left(C_{gs} + (1 + g_m R_L') C_{gd} \right) + R_L' (C_{gd} + C_L) \quad (14.79)$$

Note that the first term in Eq. 14.79 is the same as the time constant from input port, resulting from the Miller equivalent circuit. The second term is the time constant contribution from the output port of the Miller equivalent circuit. This result matches Eq. 14.19 when we account for C_L and C_{db} , which we neglected earlier.

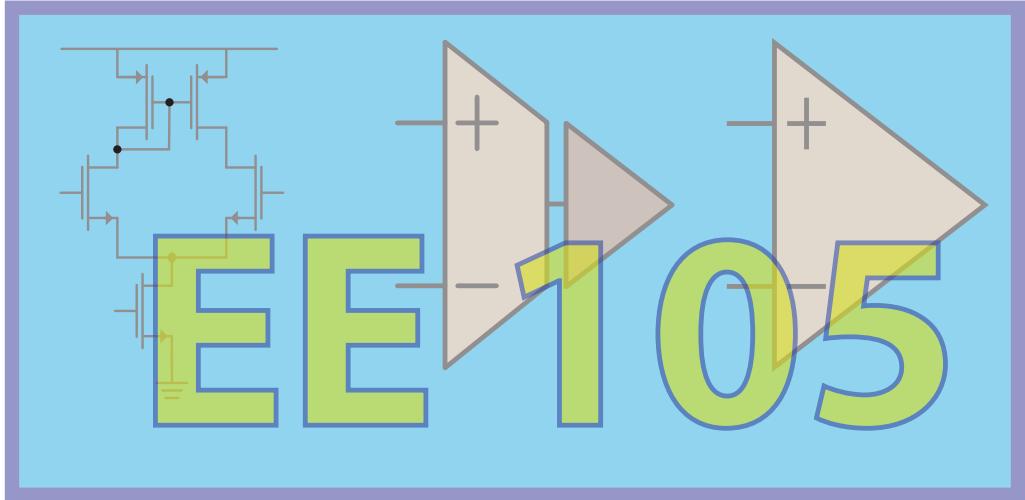
14.12.6 Practical Approach

The method of OCTC works very well for circuits with a dominant pole, but it is clunky with many capacitors. In practice, we can estimate the value of the dominant pole by finding the dominant RC time constants and neglecting others. If RC is larger than all others ($100\times$ for example), then it's a good approximation to neglect the others and only use one time constant. Many amplifiers are *designed* to have a dominant "high Z" node for stability reasons, so we can quickly estimate the bandwidth in these cases. As you will learn, even if such a "high Z" node does not exist, we create it by adding capacitance intentionally to *compensate* the frequency response. This concept will be introduced when we talk about feedback and stability in operational amplifiers, see Section 17.6.

14.13 Chapter Summary

This chapter introduced several techniques to estimate the frequency response of an amplifier. We started with an "inspection" style analysis when the nodes are independent, then moved on to the Miller Theorem and approximation. Finally, we learned the method of the OCTC, including the simpler version of just estimating the dominant time constants and ignoring others.

Many approximations were made, because in doing hand analysis (and design) we are interested in gaining insight into circuit behavior, rather than getting precise answers. We have full blown circuit simulation tools when we desire accuracy, but often numerical techniques do not lead to insights for design. This is why we need both hand analysis and computer simulation to design circuits.



15. Multi-Stage Amplifiers

15.1 Chapter Preview

In this chapter we will be covering multi-stage amplifiers consisting of a *cascade* of two or more single-stage amplifiers. First we will be reviewing amplifier input/output impedance characteristics, which will provide insights on the best ways to cascade amplifiers. Next, we will start with the important common source amplifier stage and show that while cascading these stages will result in a higher gain. However, the higher gain comes with a severe penalty of bandwidth reduction. To partially circumvent this bandwidth reduction, we will see how voltage buffers can be inserted in between amplifier stages. We will also cover an important special case of a capacitive load, which commonly occurs for on-chip amplifiers, and how to drive these loads. The next topology we will investigate in detail is the common source amplifier cascaded with a common gate amplifier, which is better known as the *cascode amplifier*. We will discuss this topology in detail, because it plays such an important role in circuit design.

The chapter will conclude with a "big" example to demonstrate how a seemingly complex circuit can be simplified into simple building blocks. Obviously this ability to see through the clutter of a circuit will not be learned overnight, but it is something you can learn fairly quickly through practice. In many ways we will only touch on the important concepts of multi-stage amplifiers, as the examples in this chapter are fairly simple. But we must first learn to walk before we attempt to run.



Figure 15.1: Guitar amplifiers, such as the classic Marshall 1987X head, are a good example of a multi-stage amplifier. They usually have one or more gain stages available to give an overdriven and "crunchy" sound.

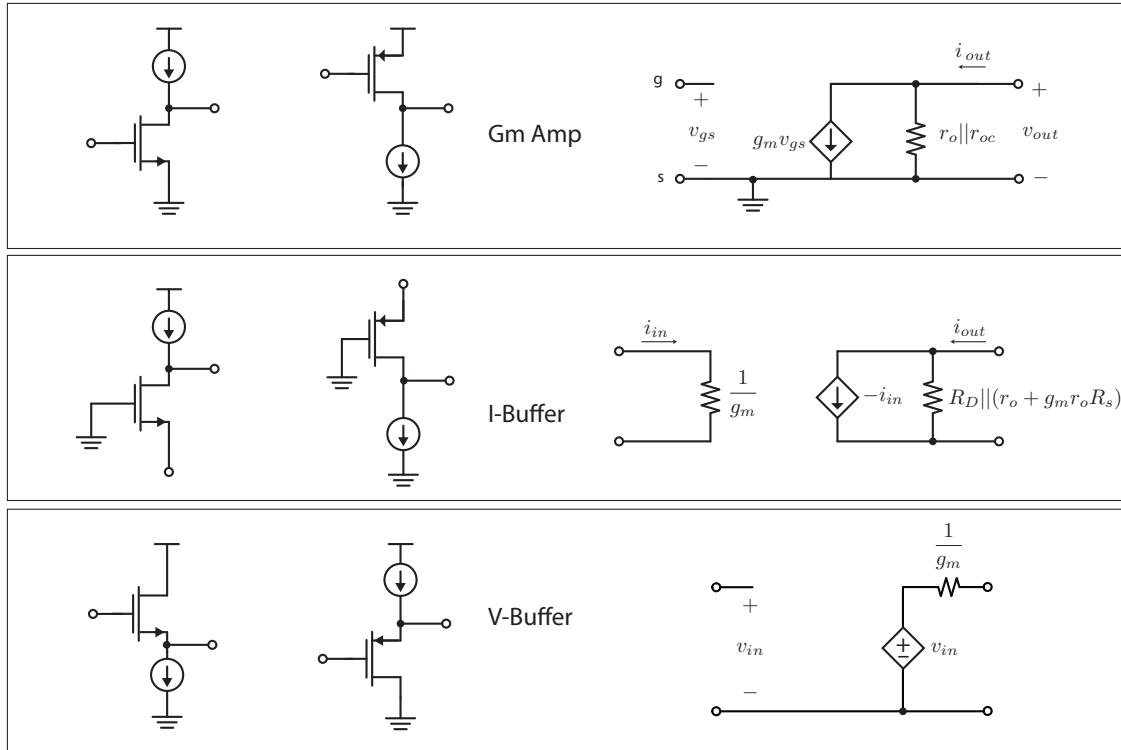


Figure 15.2: Family of single-stage MOS transistor amplifiers and the two-port models. This chart is important and should be committed to "memory".

15.2 The Need for Multi-stage Amplifiers

In the past few chapters, we have discussed single stage amplifiers in detail. We have gone over building models for each type as either a voltage or current amplifier, or a hybrid amplifier (transconductance or transresistance). See Fig. 15.2 for a reminder of these different topologies. Why should we need to cascade single-stage amplifier stages? The simplest reason is to provide more gain. More is better, right? No, not always, but in the right circumstance, distributing gain over multiple stages has benefits. Such benefits include the ability to realize higher bandwidth, or for driving output loads without compromising the gain. By separating the task of amplification into multiple steps, such as to drive an external load, we can better optimize each stage.

To begin, let's talk about gain. In modern technology, the amount of realizable gain per stage is limited, especially in nanoscale devices. For example, in a technology with $L = 1 \mu m$, a single device could provide an intrinsic gain $g_m r_o \approx 100$, whereas a short channel transistor in $28 nm$ technology may only provide a gain of 10.

The other reason we use multiple amplifier stages is to *better match the input and output source and load resistances to the amplifier*. The source or load impedance may be too high or low, and a voltage or current buffer can often provide a better interface.

Single stage amplifiers such as common source stages provide very high voltage gain. But they suffer from low frequency poles, due to the lack of isolation from the input (gate) to the output (drain), flowing through the parasitic gate-drain capacitance C_{gd} . Many multi-stage amplifiers improve the bandwidth by *providing better isolation*.

In other situations, an amplifier may have a high impedance node to provide large gain, and any additional capacitance on this node may be detrimental to bandwidth. As we will see, *a buffer can decouple the high impedance nodes from large load capacitors*.

As mentioned earlier, output stages to drive "external" loads are often needed, and they usually

have low gain owing to the fact that the load is a very small resistance (say a speaker or antenna). Multiple gain stages are therefore required to isolate the load from the gain stages, and an output stage is needed to drive the load.

Finally, more subtle applications of multi-stage amplification are for *DC coupling*, eliminating large passive elements such as capacitors that are needed to block the DC signal. We can use amplifier stages to naturally "**level-shift**" the signal. Multi-stage amplifiers like *differential stages* (covered in the next chapter) are also naturally more immune to the absolute DC level of the signal.

15.3 Review Amplifier Input/Output Impedance Characteristics

15.3.1 Impedance "Match"

On-chip circuits often use "voltage/current" matching to minimize loading. Voltage matching means we drive a high impedance load with a low impedance source. Current matching is the opposite, where we drive a low impedance load with a high impedance load. This results in minimal signal attenuation. At high frequencies, we can get even higher gain through passive element (resonant) matching gain circuits, but at lower frequencies the required values of inductance are not realizable or practical, especially on-chip.

Table 15.1 level is a quick summary of the impedance levels of various ideal amplifier topologies and their input and output resistance levels. For example, consider a voltage amplifier driving a current amplifier. Since the current amplifier input impedance is low, and the voltage amplifier also has low output impedance, the cascade will suffer from loading effects. However, if we cascade a voltage-to-current transconductance amplifier, we can see that the output/inputs are well suited to driving/load and all is well.

Amplifier Type	R_{in}	R_{out}
<i>Voltage</i>	∞	0
<i>Current</i>	0	∞
<i>Transconductance</i>	∞	∞
<i>Transresistance</i>	0	0

Table 15.1: The input and output impedance of ideal amplifier types.

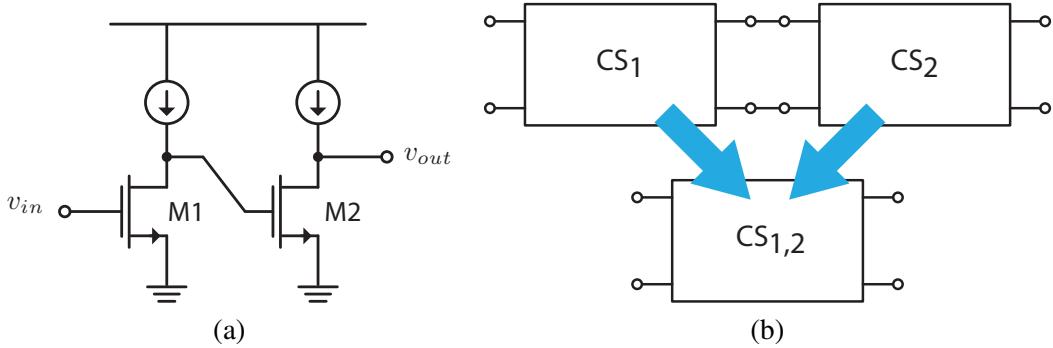


Figure 15.3: (a) Cascade of two common-source stages. (b) Cascading two stages can be considered as a composite single stage amplifier.

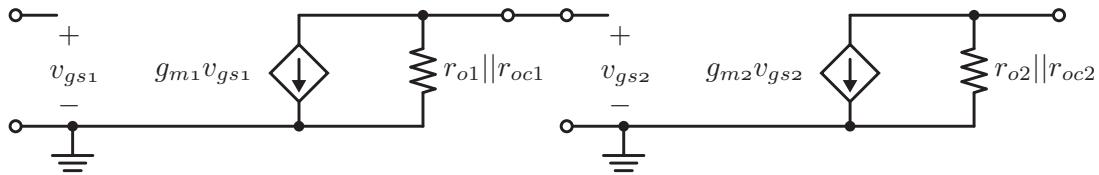


Figure 15.4: Low frequency small-signal schematic of a common-source cascade.

15.4 Common-Source Cascades

15.4.1 Two-Stage Voltage Amplifier

As shown in *Fig. 15.3*, simply cascading two amplifier stages is possible if the DC voltage levels are compatible. In other words, *the desired operating point of each individual amplifier should not change when we connect them*. Otherwise, we would need AC coupling capacitors, but it is best to avoid using them. AC coupling capacitors take up a lot of room on the chip or PCB, and they limit the range of operation (since they kill the low frequency gain). For this reason, it is good to think of cascading amplifiers while designing them. *If we are careful to not disturb the operating points, the cascade will result in higher gain*. We can think of the overall cascade as a single two-port model. If the amplifiers are unilateral, then the input impedance of the cascade is simply \$R_{in} = R_{in,1}\$, and the output impedance of the cascade is \$R_{out} = R_{out,2}\$.

15.4.2 CS Cascade Analysis

As an example of a cascade, let's take two common-source stages in cascade. We need to ensure that the drain voltage of the first stage is the right bias point for the second stage gate. Considering only the AC schematic shown in *Fig. 15.4*, the resulting new *two-port DC gain* can be analyzed by inspection:

$$A_v = \frac{v_{out}}{v_{in}} = -g_{m1}(r_{o1} \parallel r_{oc1}) \cdot -g_{m2}(r_{o2} \parallel r_{oc2}) = \boxed{g_{m1}(r_{o1} \parallel r_{oc1}) \cdot g_{m2}(r_{o2} \parallel r_{oc2})} \quad (15.1)$$

The *input impedance* is just the input impedance of the first stage:

$$R_{in} = R_{in,1} = \boxed{\infty\Omega} \quad (15.2)$$

And the *output impedance* is given by the second stage:

$$R_{out} = R_{out,2} = \boxed{r_{o2} \parallel r_{oc2}} \quad (15.3)$$

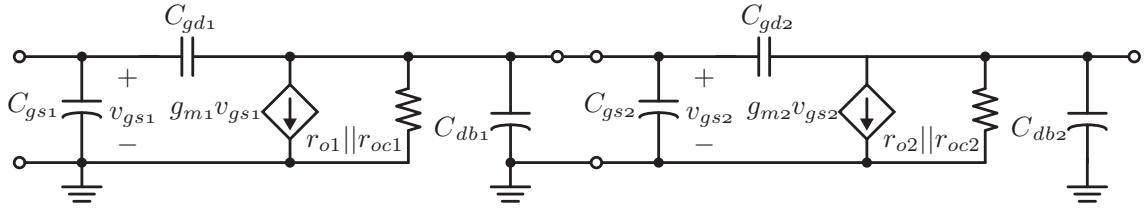


Figure 15.5: AC small-signal schematic of a common-source cascade.

15.4.3 CS Cascade Frequency Response

The AC schematic shown in *Fig. 15.5* looks complicated, with six capacitors shown. But if we combine parallel components and also apply the Miller theorem, we can identify three independent nodes. At the input, applying Miller's theorem:

$$C_{M_1} = C_{gs,1} + C_{gd,1}(1 - A_{v_1}) \quad (15.4)$$

We can approximate the gain of the first stage:

$$A_{v_1} = -g_{m,1}r_{o,1} \parallel r_{oc,1} \quad (15.5)$$

This results in a composite capacitance of:

$$C_{M_1} = C_{gs,1} + C_{gd,1}(1 + g_{m,1}r_{o,1} \parallel r_{oc,1}) \quad (15.6)$$

The second stage is exactly the same, starting with the Miller capacitance:

$$C_{M_2} = C_{gs,2} + C_{db,1} + C_{gd,2}(1 - A_{v_2}) \quad (15.7)$$

Approximating the gain:

$$A_{v_2} = -g_{m,2}(r_{o,2} \parallel r_{oc,2}) \quad (15.8)$$

Results in a second Miller composite capacitance:

$$C_{M_2} = C_{gs,2} + C_{db,1} + C_{gd,2}(1 + g_{m,2}r_{o,2} \parallel r_{oc,2}) \quad (15.9)$$

The simplified schematic is shown in *Fig. 15.6*. The time constant of the input node is given by:

$$\tau_1 = C_{M_1} R_s \quad (15.10)$$

and the intermediate node is given by:

$$\tau_2 = C_{M_2}(r_{o,1} \parallel r_{oc,1}) \quad (15.11)$$

Notice that τ_1 depends on the source resistance R_s , which can be lowered to improve the bandwidth. On the other hand, the intermediate node is a killer, because the resistance is large (output resistance of transistors), and the capacitance is large due to Miller multiplication. The output may also contribute a pole, especially if driving a capacitive load. Many poles result in very poor bandwidth, especially if they are all close in magnitude. This is one of the main issues with the multi-stage common-source cascade.

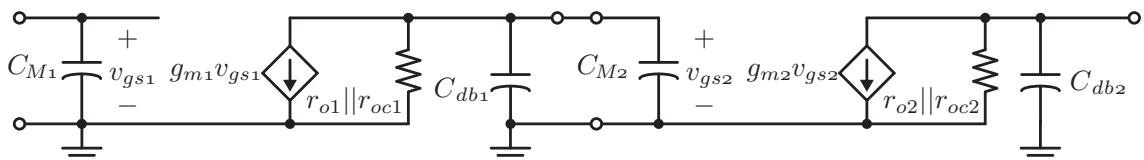


Figure 15.6: Simplified small-signal schematic of a common-source cascade using the Miller approximations.

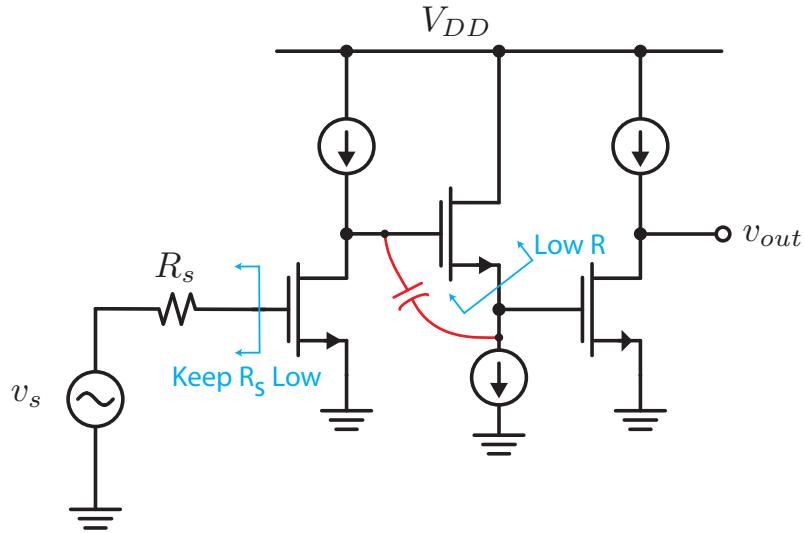


Figure 15.7: Common-source cascade with an internal source follower buffer to improve the bandwidth.

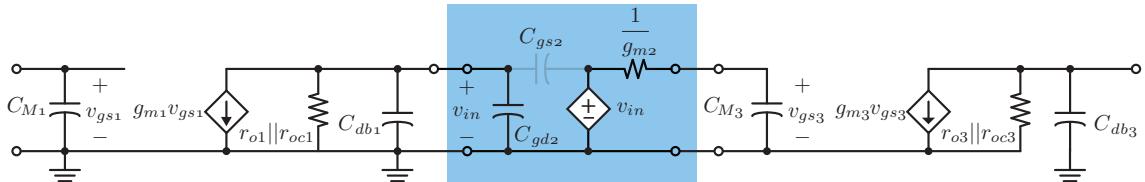


Figure 15.8: Small-signal model of common-source cascade with internal source follower buffer (highlighted).

15.4.4 Bandwidth Extension

The common source stage has high gain, but it suffers from a low bandwidth. A follower stage can buffer the Miller cap from the high output impedance of a second transistor, because it has low output impedance ($\frac{1}{g_m}$), as shown in Fig. 15.7. We still need to keep R_s under control. Recall that the C_{gs} of the follower stage is "bootstrapped", and it does not load the common source stage:

$$C_M = C_{gs}(1 - A_v) \approx 0 F \quad (15.12)$$

As shown in Fig. 15.8, for the input time constant, we have (as before):

$$\tau_1 = (C_{gs,1} + C_{gd,1}(1 + g_{m,1}r_{o,1} \parallel r_{oc,1}))R_s \quad (15.13)$$

For the input of the follower, the capacitance is low because of the follower bootstrapping:

$$\tau_2 = (C_{db,1} + C_{gd,2})r_{o,1} \parallel r_{oc,1} \quad (15.14)$$

At the output of the buffer, the low output impedance keeps the time constant small:

$$\tau_3 = (C_{gs,3} + C_{gd,3}(1 + g_{m,3}r_{o,3} \parallel r_{oc,3}))\left(\frac{1}{g_{m,2}}\right) \quad (15.15)$$

At the load, if there is no load capacitance, we have another relatively high frequency pole:

$$\tau_4 = C_{db,3}(r_{o,3} \parallel r_{oc,3}) \quad (15.16)$$

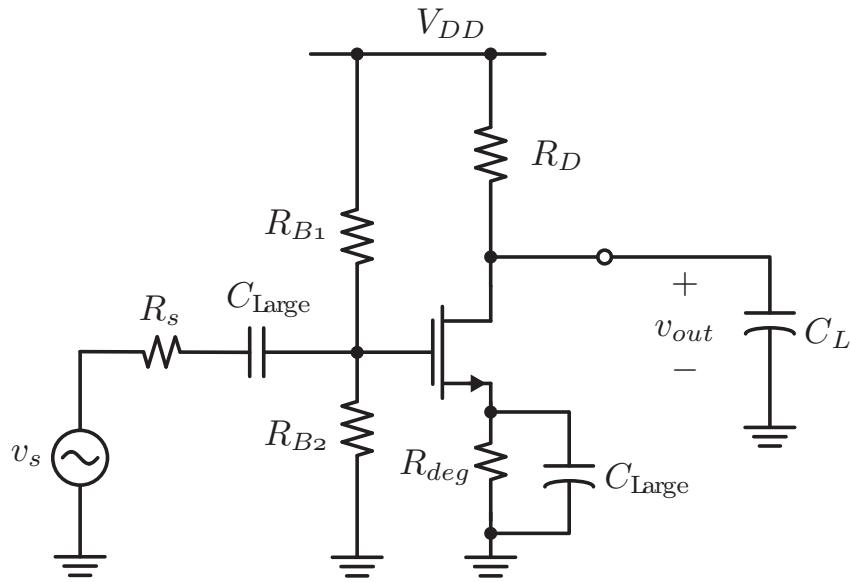
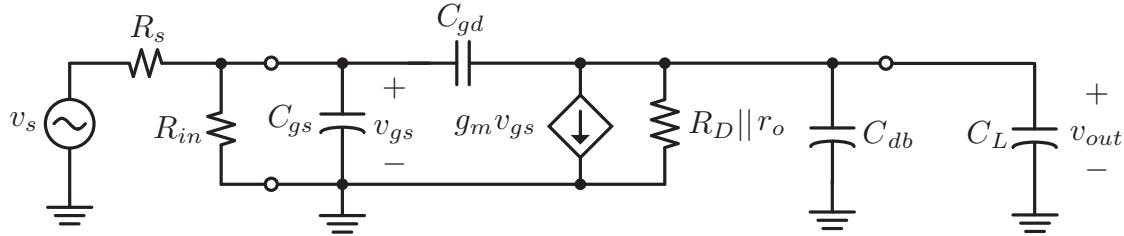


Figure 15.9: Complete schematic of a common-source amplifier including DC biasing elements.

Figure 15.10: AC small-signal schematic of common-source amplifier including the effective loading by biasing elements R_{in} and load capacitor C_L .

15.5 Common-Source with Capacitive Load

15.5.1 Common-Source with a Large Capacitive Load

A schematic of a common source amplifier with biasing elements is shown in Fig. 15.9. Recall that C_{Large} is a very large bypass capacitor. Usually these capacitors function better the larger they get, because they define the low frequency cutoff limit. These capacitors look like short circuits to the AC signal. On the other hand, if the load capacitance C_L is large, its pole dominates. Let's analyze this scenario.

In Fig. 15.10, note that R_{in} represents the parallel combination of the biasing resistors, $R_{in} = R_{B1} \parallel R_{B2}$. If we make a Miller approximation, the input time constant is given by:

$$\tau_1 = (R_s \parallel R_{in}) \left(C_{gs} + g_m (r_o \parallel R_D) C_{gd} \right) \quad (15.17)$$

At the output, we have a large capacitor $C_L \gg C_{gs, gd, db}$, so it dominates the time constant over the transistor parasitics:

$$\tau_2 \approx (R_D \parallel r_o) C_L \quad (15.18)$$

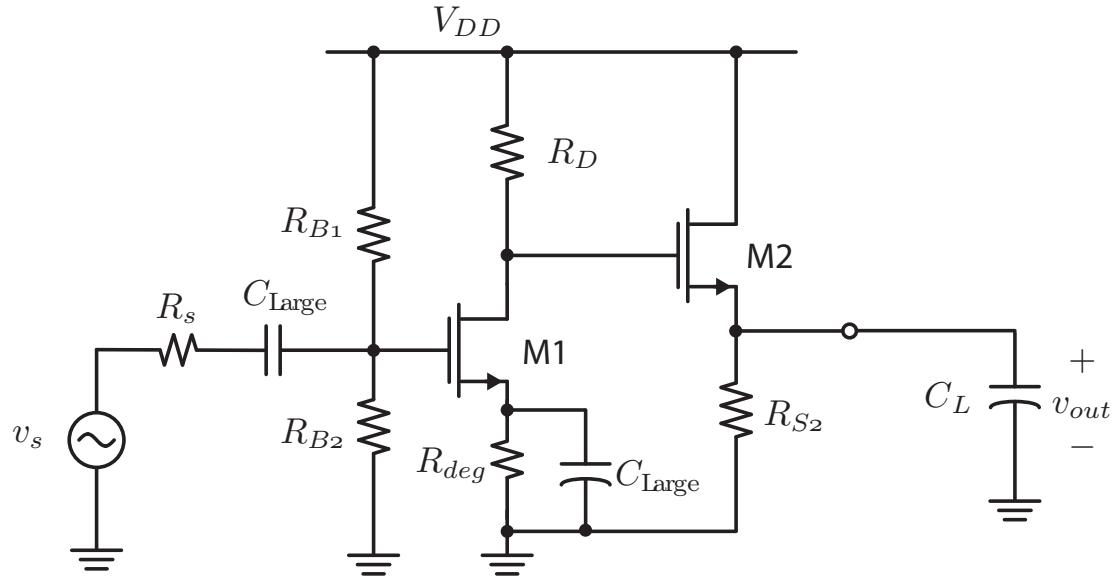


Figure 15.11: Common-drain (follower) buffer added to the common-source amplifier aids in driving a large capacitive load C_L .

What can we do if we have to drive a large capacitive load, C_L ? How can we reduce the impact of C_L ? One way is to reduce the resistance R_D . However, increasing R_D reduces our pass-band gain. To recover the gain we can increase $g_{m,1}$. But this comes at the cost of increased power consumption. There is a better way to extend the bandwidth, which is to add a source follower stage. This is illustrated in *Fig. 15.11*. Adding the source follower stage allows us to buffer the output load capacitor from the gain stage.

The analysis is similar to previous example. As shown in the AC schematic in *Fig. 15.12*, by adding a common-drain stage (Source Follower), we can increase the bandwidth. Even though it costs us power for the follower stage, increasing the bandwidth by increasing g_{m1} costs us much more.

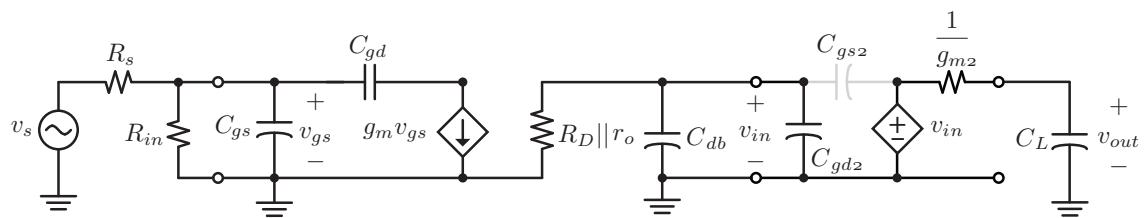


Figure 15.12: AC small-signal model of a common-source / common-drain cascade.

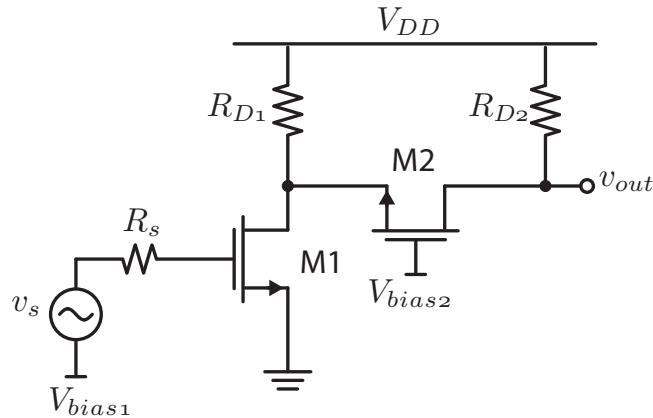


Figure 15.13: Common-source stage followed by a common-gate stage.

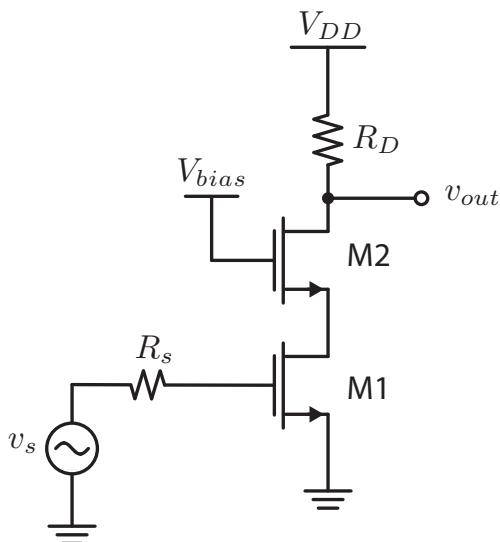


Figure 15.14: A stacked common-source stage followed by a common-gate stage, known as a "cascode" amplifier.

15.6 Common-Source Common-Gate Cascade (Cascode)

15.6.1 The Common-Source / Common-Gate Cascade

Consider the cascade shown in Fig. 15.13. The common source stage provides gain, and the common gate acts as a current buffer, passing the current to the load. Since the current buffer stage has a gain of one, is it even helping? On first pass, it seems like M2 is not doing anything, because it just buffers the current of M1 and does not provide gain. But understanding this topology is more subtle as it provides potential gain and bandwidth benefits.

15.6.2 Merged CS + CG = Cascode

As seen in Fig. 15.14, we can share the bias currents between the two amplifier stages by stacking them into a common structure, commonly known as a **cascode**. While this circuit has *reduced headroom*, it can be seen as a "supercharged" common source stage. We know that the g_m of the combined M1/M2 amplifier is the same as a single transistor, but the output impedance is boosted by the $g_m r_o$ of the transistor.

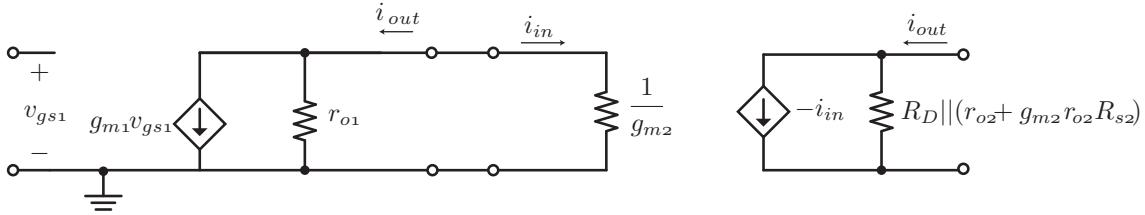


Figure 15.15: Small-signal model of the cascode amplifier.

15.6.3 Cascode Gain

Let's apply two-port small-signal analysis, shown in *Fig. 15.15*. Note the common gate stage is degenerated by $M1$, so $R_{s,2} = r_{o,1}$. In this case, we care about the *input current* to the second stage. Since the input resistance of the CG is low, the majority of the CS current is fed to the CG stage. This means the DC gain of the amplifier is given by:

$$A_v = -g_{m,1} (R_D \parallel (r_{o,2} + g_{m,2} r_{o,2} r_{o,1}) B) \approx -g_{m,1} R_D \quad (15.19)$$

This is no better than a common-source stage? So why use a cascode? Well, for one you can immediately see that we are throwing away a lot of gain by using R_D . If we instead use an ideal current source, the gain will rise to:

$$A_v = -g_{m,1} (r_{o,2} + g_{m,2} r_{o,2} r_{o,1}) \approx -(g_m r_o)^2 \quad (15.20)$$

This is substantially higher than the maximum gain of a single transistor, giving credence to the "supercharged" transistor claim. But it is even better than that when we consider bandwidth.

15.6.4 Cascode Bandwidth

To analyze the bandwidth of the cascode, let's add all the transistor parasitic capacitors, as shown in *Fig. 15.16a*. Notice that with all the parasitic capacitors in this circuit, our first step is try to combine as many as possible. Let's identify three key nodes in the circuit: the gate g , the internal node at the drain of $M1$ labeled i , and the output labeled o , shown in *Fig. 15.16b*. Next, we can apply Miller's theorem to get a sense for the bandwidth limiting stages. We immediately notice a difference in the Miller capacitance. For $M2$, its C_{gd} is simply grounded, and it is *not* Miller multiplied. On the other hand, for $M1$, the Miller capacitance is "shielded" from the high voltage gain at the output node. We will also show that the voltage gain experienced at the intermediate node is much lower.

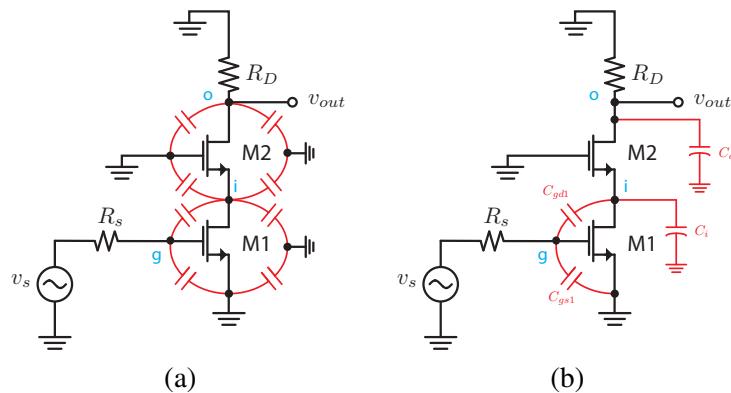


Figure 15.16: (a) Cascode amplifier with capacitive transistor parasitics. (b) Simplified schematic obtained by keeping only the relevant capacitors and combining the others.

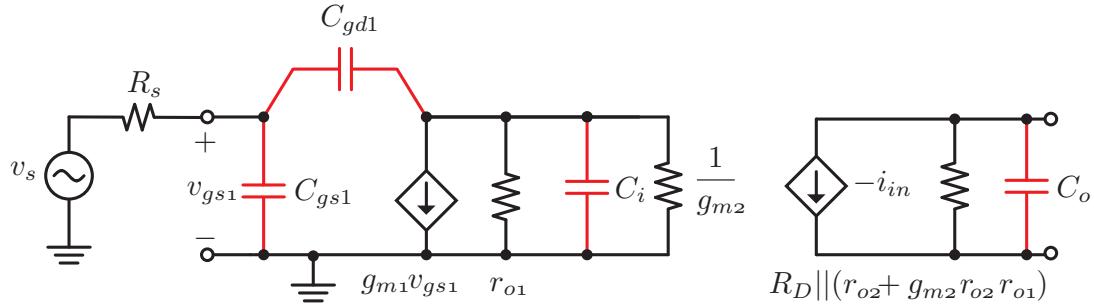


Figure 15.17: Small-signal model of a cascode amplifier using the two-port common-gate model.

15.6.5 Cascode Bandwidth - Small Signal

For a complete analysis of the bandwidth of the cascode amplifier, let's examine the small-signal model, shown in Fig. 15.17. Notice that we take advantage of the equivalent two-port circuit model of the common gate stage. At the input node, $C_{gd,1}$ is Miller multiplied like a common source, but we need to find the voltage gain from the gate to the drain of $M1$. This is given by:

$$A_{vi} = \frac{v_i}{v_g} = -g_{m,1} \left(r_{o,1} \parallel \frac{1}{g_{m,2}} \right) \approx \frac{-g_{m,1}}{g_{m,2}} \quad (15.21)$$

Since $M1$ and $M2$ share the same current, if they are sized equally, the internal voltage gain is simply $A_{vi} = -1$. Then the Miller multiplication is small:

$$\tau_g = R_s (C_{gs,1} + 2C_{gd,1}) \quad (15.22)$$

The intermediate node is a wide bandwidth node, because the common gate amplifier has a low input impedance of $\frac{1}{g_{m,2}}$:

$$\tau_i = C_i \left(\frac{1}{g_{m,2}} \parallel r_{o,1} \right) \approx \frac{C_i}{g_{m,2}} \quad (15.23)$$

Finally, the output node bandwidth depends on the load and R_D :

$$\tau_o = C_o (R_D \parallel (r_{o,2} + g_{m,2} r_{o,2} r_{o,1})) \approx R_D C_o \quad (15.24)$$

We now can appreciate the benefits of the cascode amplifier. It is essentially a supercharged transistor with the same g_m , but it does not suffer from the Miller effect (as much), and the achievable gain is much higher. For this reason, the cascode is a workhorse gain stage used in many amplifiers. The only down side is the reduced voltage headroom, discussed in the next section.

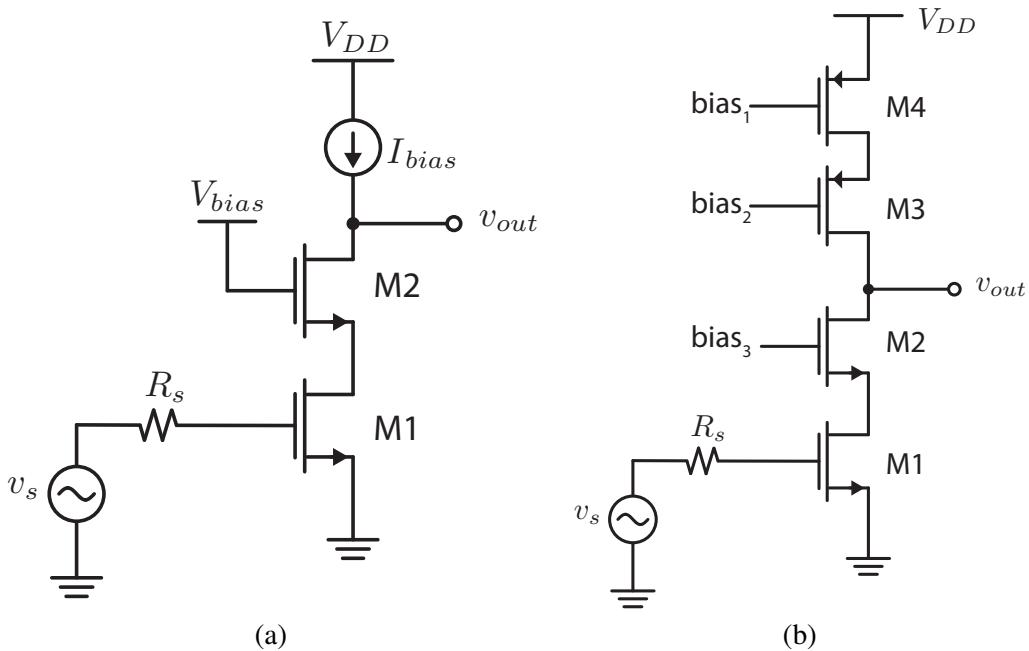


Figure 15.18: (a) A high gain cascode amplifier biased with an ideal current source I_{bias} . (b) The ideal current source load is replaced with a cascode load to maximize the gain.

15.6.6 Cascode Biasing

As discussed previously, the output impedance of the cascode is very high, and loading it with R_D is likely to reduce the voltage gain. The best we can do is to use a current source load, shown in *Fig. 15.18a*, but r_{oc} needs to be very large. Thus, we should use a cascode current mirror load as shown in *Fig. 15.18b*.

15.6.7 Example: Complete Amplifier Design

Let's design a complete cascode amplifier with the following goal of achieving a DC gain magnitude of 5000. Let's choose $g_{m,1} = 1 \text{ mS}$.¹ For simplicity, we will assume that all of the g_m and r_o values are equal. Since the gain is given by:

$$A_v \approx -g_{m,1} R_{out} = -1 \text{ mS} \cdot 5 \text{ M}\Omega = \boxed{-5000} \quad (15.25)$$

The required output resistance is given by:

$$R_{out} \approx \left(\frac{1}{2}\right) g_m r_o^2 = \boxed{5 \text{ M}\Omega} \quad (15.26)$$

The factor of one half accounts for the parallel combination of the *NMOS* and *PMOS* cascode stages. One for the cascode amplifier, and the other the cascode mirror load. That means each transistor needs an output resistance r_o of:

$$r_o = \sqrt{\frac{20 \text{ M}\Omega}{g_m}} = \sqrt{\frac{10 \text{ M}\Omega}{1 \text{ mS}}} = \boxed{100 \text{ k}\Omega} \quad (15.27)$$

¹In *Ch. 17*, we'll see that a given value of g_m is needed for the transconductance stage to meet bandwidth specs.

Bias Current and Device Sizing

To complete the design, we need to know process parameters to solve for $\frac{W}{L}$. We are given the following technology parameters: $k = \mu C_{ox} = 100 \mu A/V^2$ and $\lambda = .1 V^{-1}$. With knowledge of these parameters, we can calculate the DC current:

$$r_o = \frac{1}{\lambda I_{DS}} = \boxed{100 k\Omega} \quad (15.28)$$

$$I_{DS} = \frac{1}{0.1 V^{-1} \cdot 100 k\Omega} = \boxed{100 \mu A} \quad (15.29)$$

Since we have specified the g_m , we can now find the device size $\frac{W}{L}$:

$$g_m = \sqrt{2k \left(\frac{W}{L} \right) I_{DS}} = \boxed{1 mS} \quad (15.30)$$

$$\frac{W}{L} = \frac{g_m^2}{2k I_{DS}} = \frac{(1 mS)^2}{2 \cdot 100 \mu A \cdot 100 \mu A} = \boxed{50} \quad (15.31)$$

You may question our selection of g_m , which implies the DC current and device dimensions. The reason we pre-selected g_m is related to the fact that we desire a certain gain bandwidth in amplifiers, a topic which we will cover when we discuss op-amps. For now, simply accept it as a given that amplifiers need a minimum g_m in certain cases to meet specifications that we have not discussed.

Output (Voltage) Swing

The biggest downside to this enormous cascode amplifier is the low voltage swing. Recall that each transistor operates with a certain overdrive voltage related to the current, which is also known as the $V_{d,sat}$ of a transistor:

$$V_{OD} = V_T + \sqrt{\frac{2I_{DS}}{\mu C_{ox} \frac{W}{L}}} \quad (15.32)$$

For our example:

$$g_m = \frac{2I_{DS}}{V_{GS} - V_T} = \boxed{1 mS} \quad (15.33)$$

This leads to:

$$V_{GS} - V_T = \frac{2I_{DS}}{g_m} = \frac{2 \cdot 100 \mu A}{1 mS} = \boxed{0.2 V} \quad (15.34)$$

For long channel devices, this is simply $V_{GS} - V_T$. To keep all devices in saturation, we have to ensure that the voltage levels don not exceed the following limits:

$$V_o < V_{DD} - 2V_{OD} \quad (15.35)$$

$$V_o > 2V_{OD} \quad (15.36)$$

That means that the minimum supply voltage is $V_{swing} + (0.2 V \cdot 4) = 0.8 V + V_{swing}$. For example, for a 0.4V swing, we need at least 1.2V.

One issue we have been avoiding is how we would actually deliver all the bias voltages shown in Fig. 15.18b. The input transistor needs a bias to establish the current, but we need to ensure this current matches the top PMOS transistor. In fact, this is the major problem with this circuit, because the cascode M1 will "fight" the cascode current mirror load M4 to establish the current, resulting in an undetermined DC output level. In practice, we need a feedback circuit to establish a well defined output voltage that does not vary with temperature, process, and supply voltage. In other words, we need to program the gate voltage of M4 (labeled bias₁) to achieve the desired output DC level (say mid-rail).

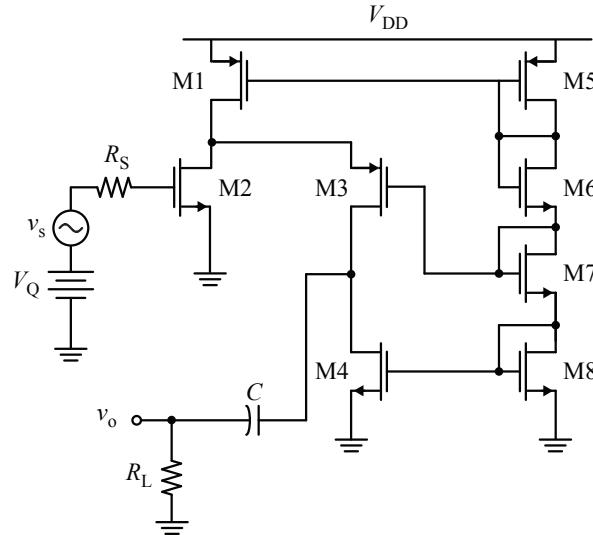


Figure 15.19: A "big" example to be analyzed in a step-by-step fashion.

15.7 "Big Circuit" Example

Let's end this chapter by bringing together all the various concepts to analyze a "big" circuit, shown in *Fig. 15.19*. At first it looks intimidating, but we will solve this problem one step at a time.

15.7.1 Cutting Through the Complexity

The key to analyzing a big circuit is to cut out all the clutter. To do this, identify the "signal path" between the input and output, and aggressively eliminate "background" transistors to reduce the clutter, as shown in *Fig. 15.20*. Transistors M5-M8 are all *diode connected*, and essentially provide DC levels to bias up the other transistors. So we simply ignore them for the AC analysis. For these "background" transistors, we need to understand their role (e.g. DC biasing) and model them appropriately. If they are just providing DC voltages, we can assume that these connections are AC grounds. Later, for estimating the frequency response, we need to identify the "high-Z" (high impedance) nodes, so keep these in mind as we traverse the circuit.

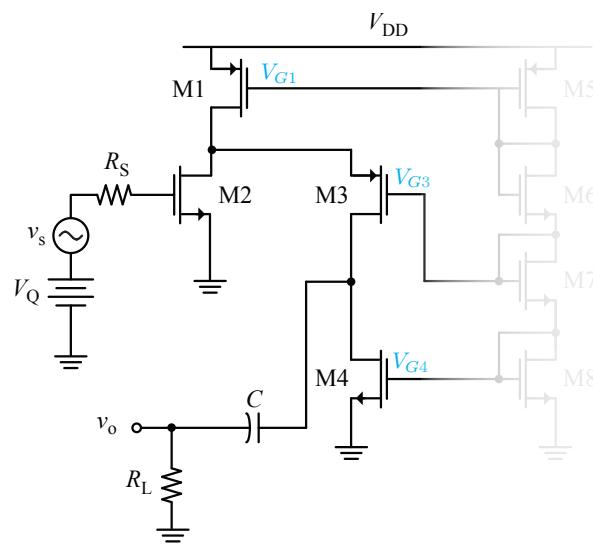


Figure 15.20: Schematic of an amplifier highlighting the signal path and ignoring biasing elements.

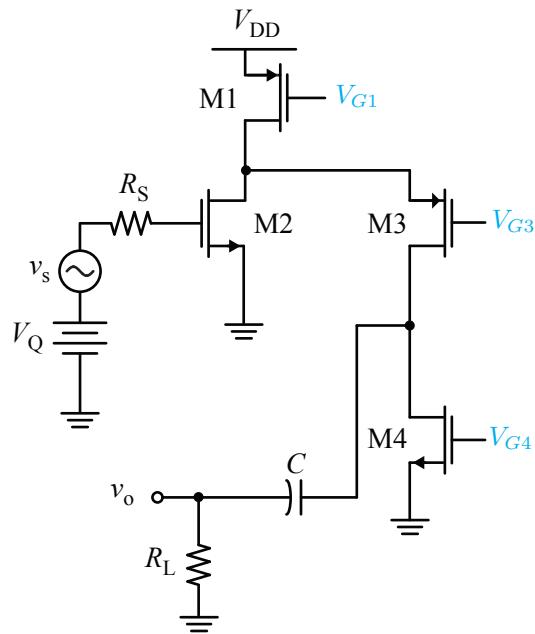


Figure 15.21: Further simplification of the amplifier by noting that many transistors are biased with a nearly constant DC voltage.

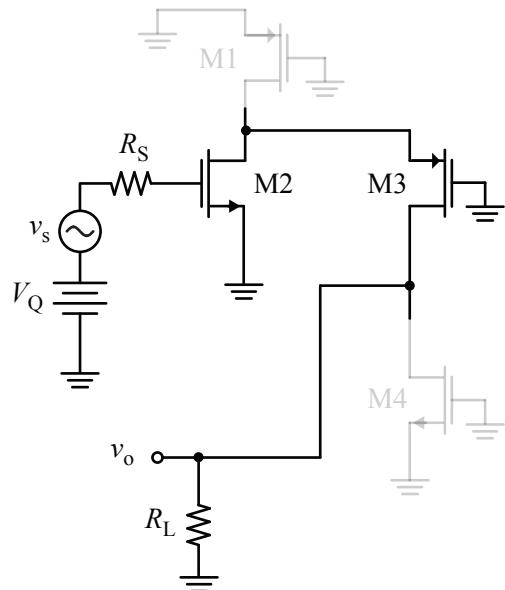


Figure 15.22: AC schematic of the core amplifier consists of transistors M2/M3, known as a "folded cascode" amplifier.

15.7.2 Eliminate More Clutter

In the schematic of *Fig. 15.20*, notice that $M1$ and $M4$ are biased with constant gate-source voltage, and only play an indirect role as far as the signal is concerned. In fact, we can model them as current sources and simplify the schematic further, shown in *Fig. 15.21*.

Since $M1$ and $M4$ are providing DC currents, as far as the AC circuit is concerned they are open-circuits, as shown in *Fig. 15.22*. With all these simplifications, we finally can see the essence

of the circuit. It is simply an *NMOS* common source stage followed by a *PMOS* common gate stage. This is similar to a cascode, except that it is "folded down", and aptly named the "**folded cascode**". *M1* provides the required DC bias for both *M2* and *M3*. Since *M4* establishes the current in *M3*, the difference between the current of *M1* and *M4* flows into *M2*.

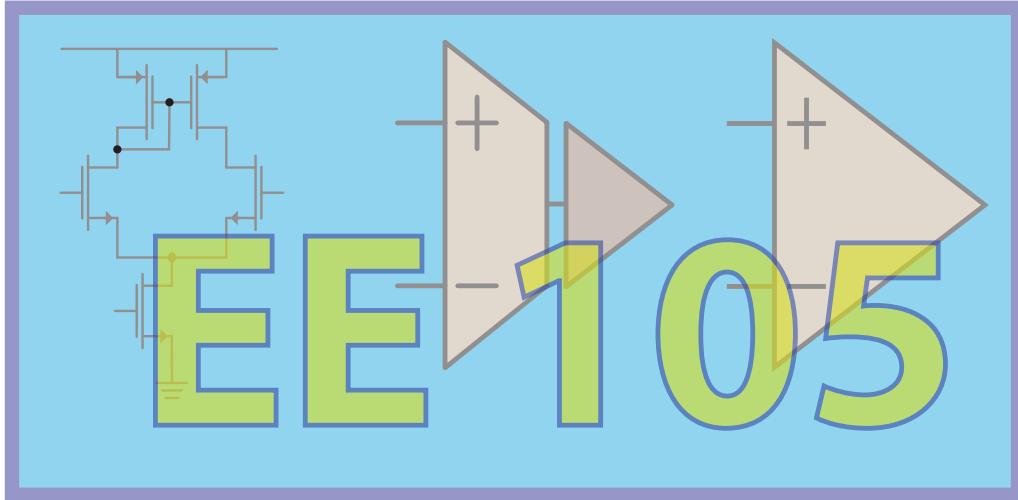
Why is the cascode folded? Compared with a regular cascode, the second stage moves the DC bias voltage down, which may be required to drive another stage. Furthermore, by folding the cascode, the voltage headroom requirements are reduced.

Now let's put the circuit under the microscope to see if we missed anything. Notice that in *Fig. 15.19*, there is actually a current divider that we neglected. The AC current generated by *M2* is divided between *M3*, *M2* (r_o), and *M1*. We tacitly assumed that *M3* has a much higher input conductance (approximately g_m) compared to the output conductance of *M1*, which is $\frac{1}{r_{o,1}}$, and so we pretty much neglected the AC current flowing into *M1*. Similarly, we lose a small amount of current into the output resistance of *M2*, which we ignored. The only time we need to be careful about this approximation is when the load R_L becomes very large, invalidating our assumption that the input impedance of *M3* is $\frac{1}{g_m}$. Also, if you were wondering, since *M1* is not diode connected, its current is not mirrored to *M5*, even though they have the same v_{gs} . Taking this all into consideration, the gain of the amplifier is given by:

$$A_v = -g_{m,2} \left(\frac{g_{m,3}}{g_{m,3} + \frac{1}{r_{o,1}} + \frac{1}{r_{o,2}}} \right) \cdot \left(R_L \parallel r_{o,4} \parallel ((1 + g_{m,3}r_{o,1} \parallel r_{o,2})) r_{o,3} \right) \quad (15.37)$$

$$\approx -g_{m,2} R_L \quad \text{Folded cascode amplifier gain} \quad (15.38)$$

If the above equation is "obvious" to you, congratulations. You have made it, and can now analyze multi-stage amplifiers by "inspection". If not, don't worry, it just takes some time and practice.



16. Differential Amplifiers

16.1 Chapter Preview

This chapter introduces a very important building block known as the differential amplifier. This amplifier is at the heart of almost every analog amplifier, and most notably forms the input stage of every operational amplifier (see *Fig. 16.1*).

We begin by motivating the need for differential amplifiers, and then we give a pseudo-historical account of how the differential amplifier was invented. In this way we can actually learn how the amplifier takes form into a fully symmetric structure. We will derive the transfer function in detail for both large and small signals. The key to the analysis will be to take advantage of the symmetry in the circuit to simplify the analysis. To do this, we need to take arbitrary inputs and represent it as a combination of *differential-mode* and *common-mode* signals, which enables us to make symmetry arguments. With such a representation, we will define common-mode and differential-mode gains, and we will discuss important concepts such as common-mode rejection.

We end the chapter by discussing more advanced differential amplifier architectures, with current source and current mirror loads, which provide high gain and good common-mode rejection.

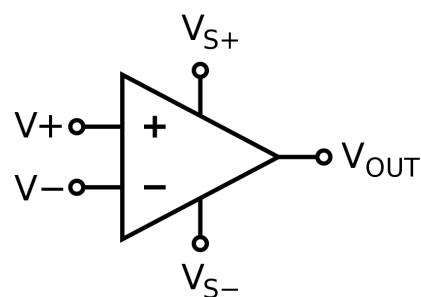


Figure 16.1: In real op-amps there is always a non-zero voltage between the inputs.

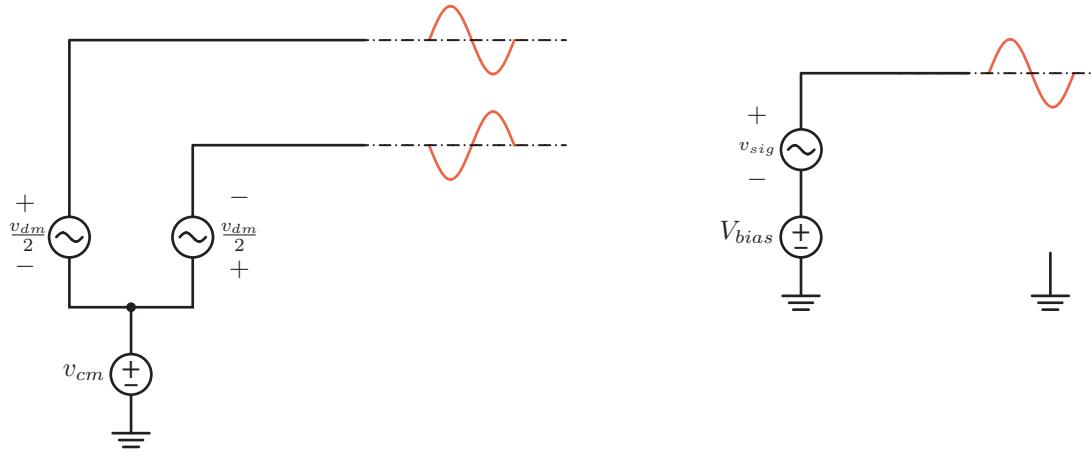


Figure 16.2: In a differential system, two wires carry the signal with opposite phase, as shown. The differential signal is measured between the two wires. In a single-ended system, only a single wire carries the signal and the potential is referenced to ground.

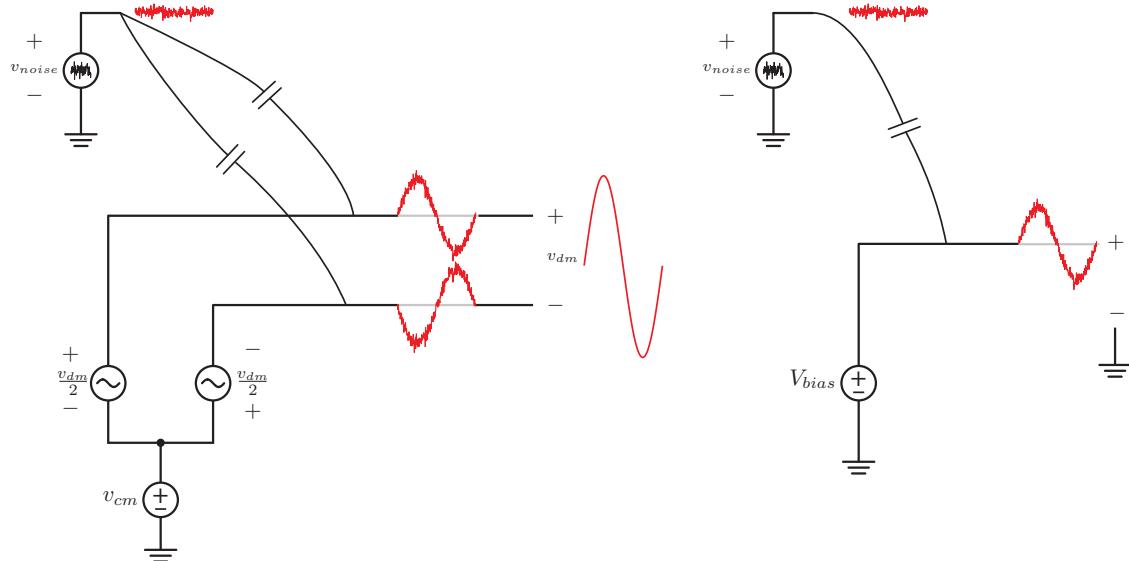


Figure 16.3: In a differential system, any noise pickup that is common-mode, for example from power supplies or from the substrate of the IC, is rejected since the output is a differential signal. In a single-ended system, this noise cannot be distinguished from the signal itself.

16.2 Motivation for Differential Operation

First of all, what is a differential circuit? In *Fig. 16.2* we highlight the difference between a "**single-ended**" circuit and a **differential** circuit. In a single-ended circuit, *all signals are referenced to a common ground*, and signals are routed using a single wire. In a differential circuit, *each signal is represented with two wires*. Also, optimally each wire carries half of the signal such that each signal's phase is equal and opposite to the other. In this form the signals are self-referenced without needing a ground connection. The potential difference between the wires is known as the **differential-mode signal**. The signal in common with both is known as the **common-mode signal**.

Ideally the common-mode signal is just a DC reference or bias voltage, but in practice it may vary due to interference or imbalances. A good example of interference is the ground or supply noise, that ideally couples to the differential wires in the same manner, producing a common-mode

signal. In a single-ended system, any common-mode noise is indistinguishable from the desired signal, and often it can swamp the desired signal. A good example is the large AC voltages/currents flowing in power lines – even a small fraction picked up by a sensor due to capacitive coupling can easily swamp out a micro-volt signal.

Differential circuits are much less sensitive to common-mode noise and interference, and are preferred to the greatest extent possible. As shown in *Fig. 16.3*, common-mode noise is rejected in a differential circuit as long as the circuits are balanced (noise is picked up as a pure common mode signal). Other benefits of the differential configuration is that it enables us to bias amplifiers and connect multiple stages without using coupling or bypass capacitors. We will cover this in detail (see *Fig. 16.9* later in this chapter).

Correct differential operation requires excellent matching of transistors and other components. This favors implementation in integrated circuit (IC) form, where matched transistors can be fabricated side-by-side or even inter-digitated together (see *Fig. 13.10*). Differential amplifiers require twice as many transistors. But again, in an IC this is not an issue at all, because transistors are essentially "free" compared to other larger components (such as resistors and capacitors).

16.3 Birth of Symmetric Source-Coupled Pair

16.3.1 Goal: Differential Transconductance G_m

Let's imagine that you're living in a pre-differential world where all signal processing is done in single-ended form. You're inspired to build a transconductance stage that can be controlled by the difference in voltage between two nodes, rather than a single-ended transconductor, which is commonly used in the form of a transistor. In other words, our goal is to have an output current that satisfies this equation:

$$i_o = G_m(v^+ - v^-) \quad (16.1)$$

Ideally the circuit should have very high output impedance and very little loading on nodes v^+ and v^- . As a good circuit designer, you come to the realization that a common source and common gate amplifiers are complementary in the sense that they have opposite phase. If you equalize the gain, you can potentially build a **differential transconductance**, G_m . In fact, you can combine the common source and common gate amplifier into a single transistor as shown in *Fig. 16.4a*. Let's calculate the gain by superposition. If we excite v^- , and ground v^+ , we have an output voltage given by:

$$v_{o,CS} = -g_m R_D v^- \quad (16.2)$$

Likewise, if you excite the common gate stage through v^+ (and ground v^-), we have:

$$v_{o,CG} = g_m R_D v^+ \quad (16.3)$$

Summing these signals, we have:

$$v_o = g_m R_D (v^+ - v^-) \quad (16.4)$$

This looks pretty promising, so what's missing?

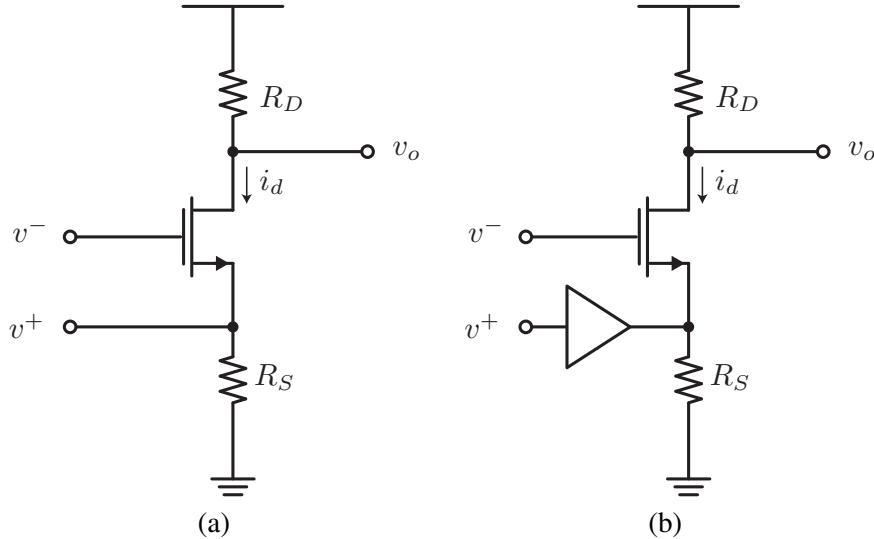


Figure 16.4: (a) A combined common-gate and common source amplifier. From the v^+ terminal, the circuit is a common gate amplifier, whereas from the v^- terminal it is a common source amplifier. (b) A buffer is needed to raise the impedance of the common gate amplifier if we wish to equalize the gain under arbitrary drive conditions.

Problem: Asymmetric (Low) Input Impedance

Notice that the gain is symmetric from the "plus" and "minus" side for an ideal voltage drive, but not for any drive with finite non-zero source resistance. Because of the low input impedance of the common gate stage, there is going to be a gain mismatch. This is due to the voltage division between the source R_s and the input impedance of the common gate stage, which is roughly $\frac{1}{g_m}$:

$$v_{o,CG} = \left(\frac{\frac{1}{g_m}}{\frac{1}{g_m} + R_s} \right) g_m R_D v^+ = \left(\frac{g_m}{1 + g_m R_s} \right) R_D v^+ \quad (16.5)$$

Solution: Add Buffer to Source Side

So what is the solution? Simply adding a voltage buffer as shown in Fig. 16.4b would solve the problem. However, any real buffer has an output impedance, leading us back to the same issue. For example, the source follower shown in Fig. 16.5 has an output impedance of $\frac{1}{g_{m,buff}}$. To minimize the voltage loss we could burn more and more power in the buffer to make the output impedance smaller. But this is not an elegant solution, because the gain mismatch will always be there.

16.3.2 Gain of Source Degenerated CS

Let's think about this problem in another way. The introduction of the buffer did not solve the problem completely, and in fact it introduced another problem. Even though the input impedance of the common source stage is very large, the gain is now degenerated by the output impedance of the buffer:

$$v_{o,CS} = - \left(\frac{g_m}{1 + g_m \left(\frac{1}{g_{m,buff}} \right)} \right) R_D v^- \quad (16.6)$$

What if we equalized the $g_{m,buff} = g_m$ with the common source/gate stage? Then the gain loss for the common source would be $\frac{1}{2}$, and the buffer driving the common gate stage would also suffer a voltage loss of $\frac{1}{2}$. In other words, in this symmetrical arrangement, the gains will be matched precisely as desired and both ports have a high input impedance.

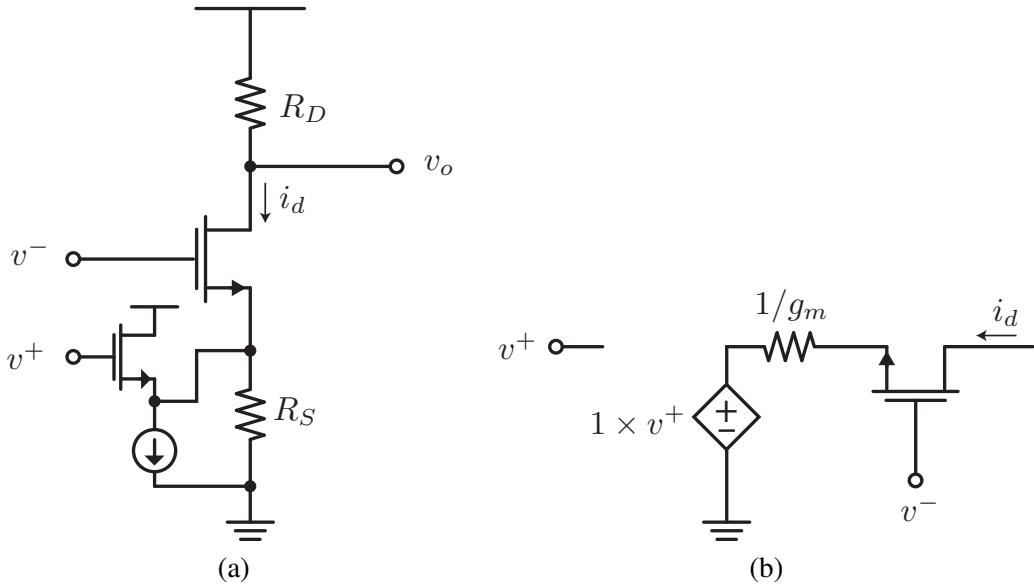


Figure 16.5: (a) Common source and common gate amplifier with a source follower buffer added. (b) The buffer output impedance $\frac{1}{g_m}$ degenerates the common source stage, and reduces the gain of the common gate amplifier.

16.3.3 A Symmetric Source Coupled Pair is Born

Putting these ideas together, we arrive at *Fig. 16.6a*. The resistor R_S is just there for biasing, and we can replace it with a current source and combine it with the current source of the follower. The result, redrawn in *Fig. 16.6b*, is a fully symmetric circuit.

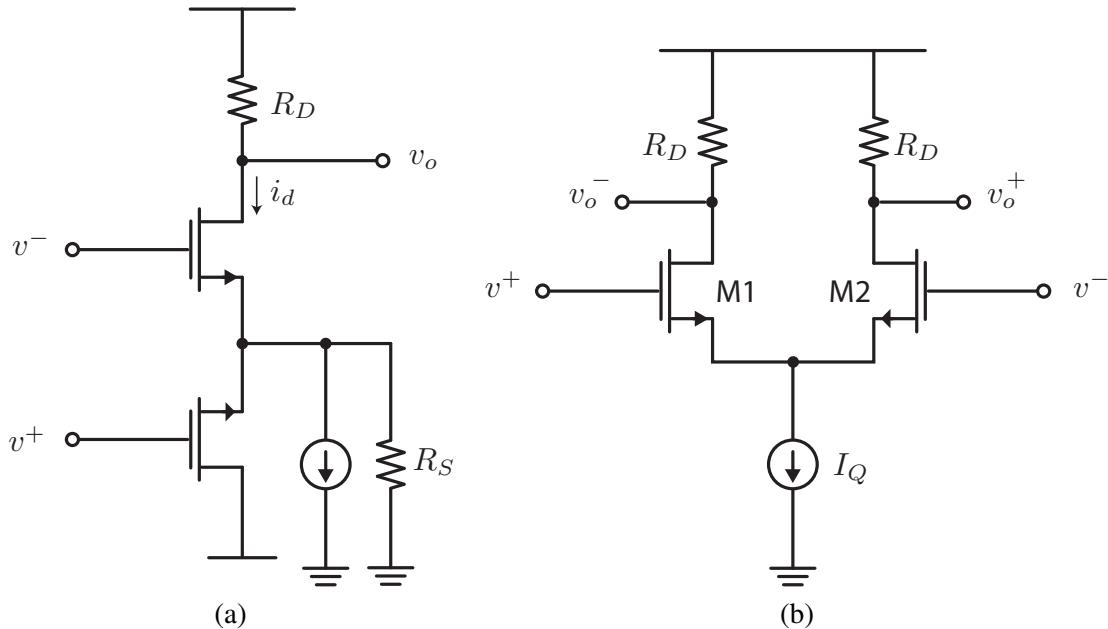


Figure 16.6: (a) A combined common source and common gate amplifier with equally sized devices for the buffer and amplifier, thereby making the circuit symmetric. (b) Redrawn circuit to emphasize the symmetry. This circuit is known as the differential pair.

The circuit in *Fig. 16.6b* has two outputs, rather than one, which are in exactly the opposite phase. The key insight is that the circuit is fully symmetric, and the gain from both sides is lowered by the same factor:

$$v_{o,CS} = -\frac{g_m}{1 + g_m \left(\frac{1}{g_m}\right)} R_D v^- = -\left(\frac{g_m}{2}\right) R_D v^- \quad (16.7)$$

$$v_{o,CG} = \frac{g_m}{1 + g_m \left(\frac{1}{g_m}\right)} R_D v^+ = +\left(\frac{g_m}{2}\right) R_D v^+ \quad (16.8)$$

For simplicity, we only consider one output, but there are in fact two outputs. Because since they are opposite in phase, we can combine them and consider the output as the voltage difference between the drain nodes:

$$v_{o,1} = \left(\frac{g_m}{2}\right) R_D (v^+ - v^-) \quad (16.9)$$

$$v_{o,2} = \left(\frac{g_m}{2}\right) R_D (v^- - v^+) \quad (16.10)$$

In other words, we have a fully differential circuit, allowing us to carry the signal differentially at both the input and output:

$$v_{o,diff} = v_{o,1} - v_{o,2} = g_m R_D (v^+ - v^-) = g_m R_D v_{i,diff} \quad (16.11)$$

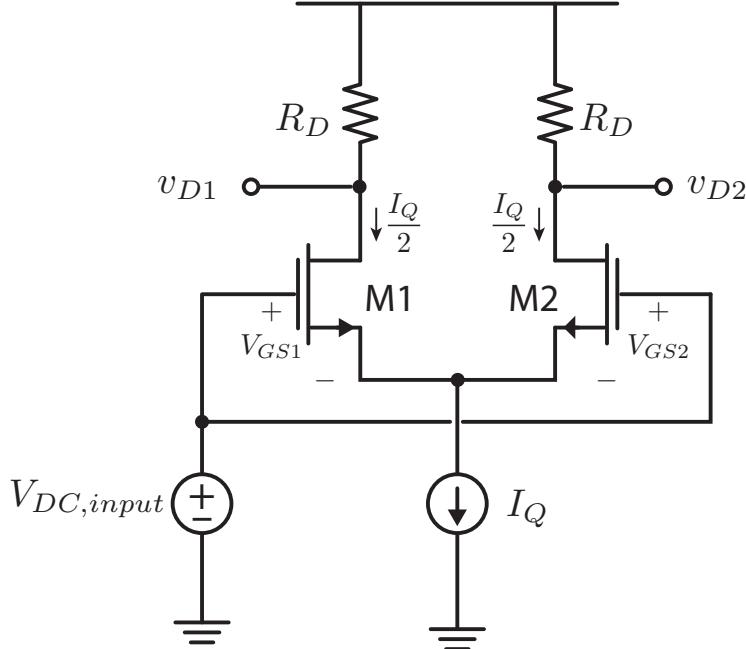


Figure 16.7: Differential pair under a common mode DC voltage drive.

16.4 Differential Operation

16.4.1 MOS Differential Pair: DC Bias

Consider the differential pair with a DC bias connected to both sides of the input, as shown in Fig. 16.7. We will shortly demonstrate that the DC bias point of $M1$ and $M2$ does not depend on the common gate DC voltage, because the current source sets the bias point of each transistor at half the bias current I_Q :

$$I_{D_1} = I_{D_2} = \frac{I_Q}{2} \quad (16.12)$$

This must be true by a simple symmetry argument. Everything is assumed to be symmetric, and so I_Q must divide in half. That means each transistor has the same V_{GS} , and consequently the same overdrive V_{OD} , assuming that the threshold voltages match:

$$\frac{I_Q}{2} = \frac{k_n}{2} (V_{GS} - V_{T_n})^2 \quad (16.13)$$

$$V_{GS1} = V_{GS2} = V_{T_n} + \sqrt{\frac{I_Q}{k_n}} = V_{T_n} + V_{OD} \quad (16.14)$$

The drain voltage for both transistors is also equal to:

$$V_{D_1} = V_{D_2} = V_{DD} - \left(\frac{I_Q}{2} \right) R_D \quad (16.15)$$

And the differential DC output is exactly zero:

$$V_{D_1} - V_{D_2} = 0V \quad (16.16)$$

Consequently, the DC level at the output is independent of DC level at input! In Fig. 16.8, we illustrate this by varying the DC level at the input and finding the operating point in the circuit. Clearly, the circuit operating point and the region of operation of $M1$ and $M2$ does not depend on the input DC level. Note that although V_{GS} and V_D are held constant, the source voltage moves. But since the source node is driven by an ideal current source, we can tolerate this variation. This allows us to cascade differential pairs without worrying about DC level, which obviates the need for bulky AC coupling capacitors, as shown in Fig. 16.9. Is this too good to be true? As long as we ensure that the current source is not squashed, the circuit works as advertised. A real current source (transistor) has a finite headroom to operate, for example a V_{OD} for a single transistor current source.

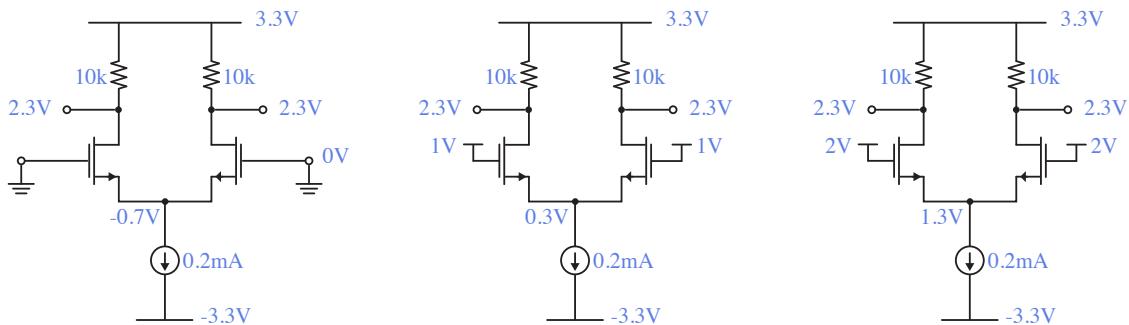


Figure 16.8: Differential pair with varying input DC voltage common-mode drive illustrating that the DC operating point of the circuit, in so far as the bias current and the drain voltage at the output is concerned, is independent of the input DC level.

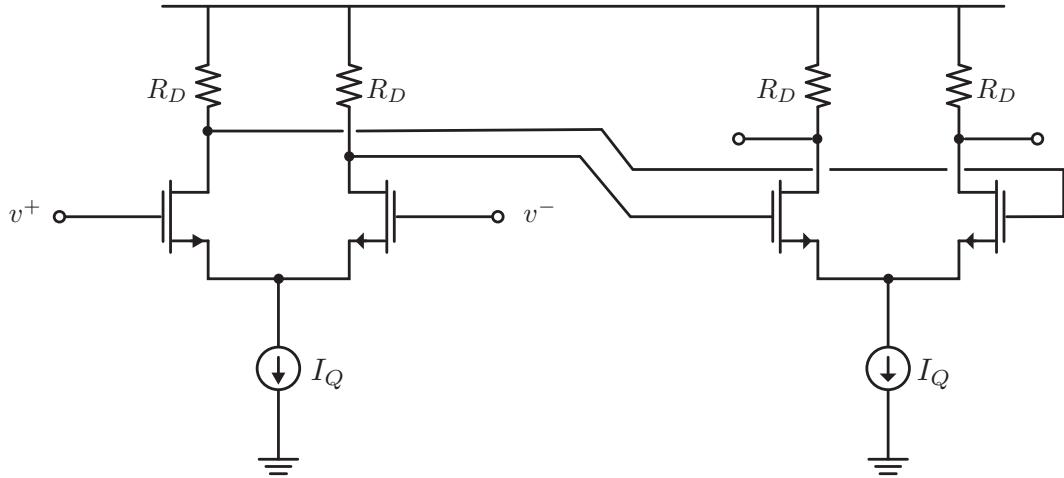


Figure 16.9: Cascading differential amplifiers without using AC coupling capacitors is possible because each amplifier bias point is independent of the input DC level, see Fig. 16.8.

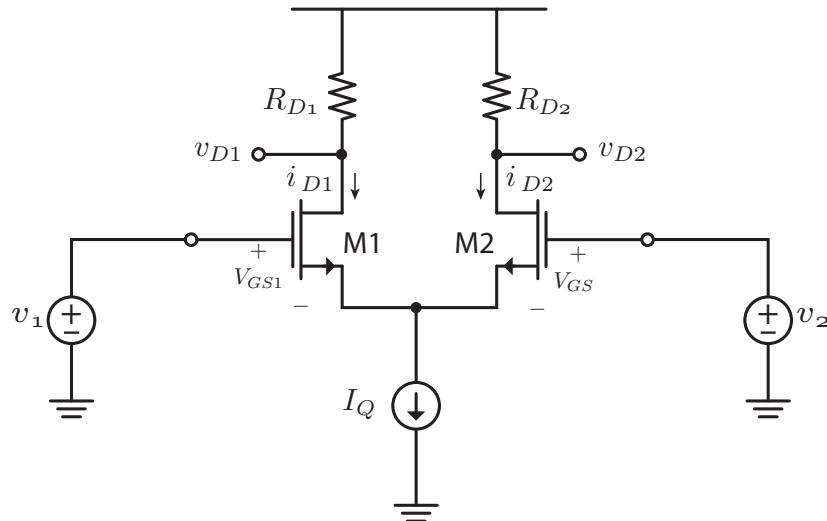


Figure 16.10: Differential-Pair under arbitrary large signal drive at the inputs.

16.4.2 MOS Differential-Pair: Differential Inputs

Let's consider applying a differential voltage to the differential pair as shown in Fig. 16.10. We apply a gate voltage v_1 to $M1$ and v_2 to $M2$, with the condition that the difference between the inputs is the *differential input signal*:

$$v_1 - v_2 = v_{id} \quad (16.17)$$

Since the input voltages applied to each gate is arbitrary, the current I_Q no longer splits equally. In fact, you can think of $M1$ and $M2$ as two devices competing for the current, I_Q . The current will favor the device with higher v_{GS} . So we can think of $M1$ and $M2$ as a pair of transistors that steer the current left or right, depending on the magnitude of the input drive. For large positive voltage $v_{id} = v_{GS1} - v_{GS2}$, we expect most of the current I_Q to be steered into $M1$. Likewise, when v_{id} is large and negative, then most of the current I_Q should flow into $M2$. Let's confirm this observation

with the *MOS* equations. In general we have:

$$i_{D_1} = \frac{k_n}{2} (v_{GS_1} - V_T)^2 \quad (16.18)$$

$$i_{D_2} = \frac{k_n}{2} (v_{GS_2} - V_T)^2 \quad (16.19)$$

But the currents must sum to I_Q :

$$i_{D_1} + i_{D_2} = I_Q \rightarrow i_{D_2} = I_Q - i_{D_1} \quad (16.20)$$

Using these relations it is possible to derive a complete expression for the drain currents of $M1$ and $M2$ as a function of the differential voltage drive v_{id} :

$$i_{D_{1,2}} = \frac{I_Q}{2} \pm \sqrt{k_n I_Q} \left(\frac{v_{id}}{2} \right) \sqrt{1 - \frac{\left(\frac{v_{id}}{2} \right)^2}{\frac{I_Q}{k_n}}} \quad (16.21)$$

Or in terms of the transistor overdrive voltage V_{od} , we know that the *quiescent bias point* satisfies:

$$\frac{I_Q}{2} = \left(\frac{1}{2} \right) k_n V_{OD}^2 \rightarrow k_n = \frac{I_Q}{V_{OD}^2} \quad (16.22)$$

Allowing us to express the current in the following insightful form:

$$i_{D_{1,2}} = \frac{I_Q}{2} \pm \left(\frac{I_Q}{V_{OD}} \right) \left(\frac{v_{id}}{2} \right) \sqrt{1 - \frac{\left(\frac{v_{id}}{2} \right)^2}{V_{OD}^2}} \quad (16.23)$$

The full derivation is presented in Sec. 16.8 for your reference. A plot of the currents through each transistor is shown in Fig. 16.11. As expected, the curves are symmetric with respect to the two transistors and the input voltage v_{id} , where a positive voltage v_{id} steers the current into $M1$. The steering is very linear near the origin, and if the input is sufficiently large, we see that the current is steered entirely in one direction or the other. When $v_{id} = \sqrt{2} V_{OD}$, Eq. 16.23 can be simplified:

$$i_{D_{1,2}} = \frac{I_Q}{2} \pm \left(\frac{I_Q}{V_{OD}} \right) \left(\frac{\sqrt{2} V_{OD}}{2} \right) \sqrt{1 - \frac{\left(\frac{\sqrt{2} V_{OD}}{2} \right)^2}{V_{OD}^2}} = \frac{I_Q}{2} \pm \frac{I_Q}{2} = \begin{cases} +I_Q & \rightarrow i_{D_1} \\ 0 & \rightarrow i_{D_2} \end{cases} \quad (16.24)$$

We see that all the current I_Q goes to i_{D_1} when the input is sufficiently large and $M2$ shuts off. When $v_{id} = -\sqrt{2} V_{OD}$, the situation is exactly the opposite, with $M1$ shutting off and $M2$ getting all the current I_Q .

The slope of the curves in Fig. 16.11 is the differential pair G_m for each transistor. The linear range of operation of the *MOS* differential pair can be extended by operating the transistor at a higher value of V_{OD} , as shown in Fig. 16.12. At the origin, we can estimate the slope by considering a small input signal v_{id} , which causes the square root to be simplified:

$$\sqrt{1 - \frac{\left(\frac{v_{id}}{2} \right)^2}{V_{OD}^2}} \approx 1 \quad (16.25)$$

This results in linear deviations of the currents from the DC operation points:

$$i_{D_1} = \frac{I_Q}{2} + \left(\frac{I_Q}{V_{OD}} \right) \left(\frac{v_{id}}{2} \right) \quad (16.26)$$

$$i_{D_2} = \frac{I_Q}{2} - \left(\frac{I_Q}{V_{OD}} \right) \left(\frac{v_{id}}{2} \right) \quad (16.27)$$

It is clear that the first term is the bias current through $M1$ and $M2$. The slope of the linear deviation is actually familiar. It looks like the transistor g_m :

$$g_m = \frac{2I_D}{V_{OD}} \quad (16.28)$$

Since each transistor is biased at $\frac{I_Q}{2}$, the expression can be interpreted as follows:

$$i_{D1,2} = \frac{I_Q}{2} \pm g_{m1,2} \left(\frac{v_{id}}{2} \right) \quad (16.29)$$

In other words, each transistor is excited in the small-signal with a voltage of $\frac{v_{id}}{2}$. This is expected since the voltage v_{id} is dropped across two series connected transistors (back-to-back), and the voltage splits in two. What's surprising is that the g_m looks like it is the transconductance of a common source stage, even though the source coupled node is "floating". This requires some further explanation, provided next.

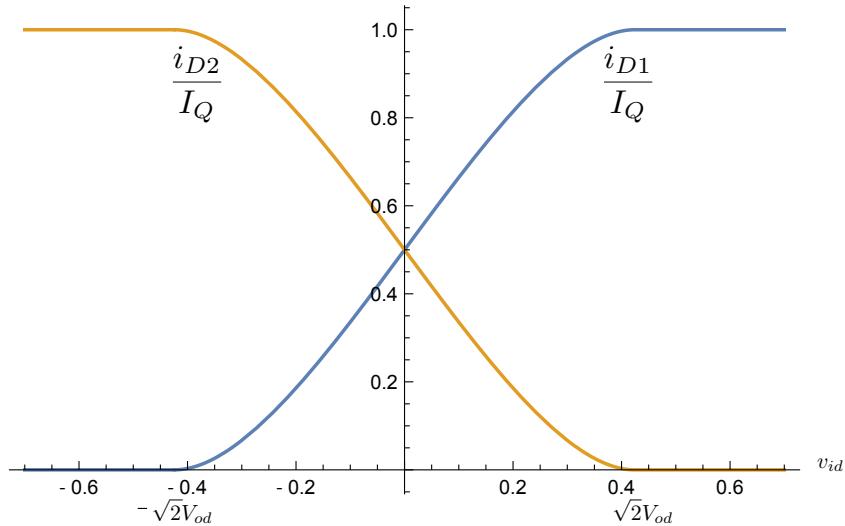


Figure 16.11: The currents in each branch of the differential pair circuit as a function of the input differential voltage v_{id} . For zero drive, each transistor shares half of the current I_Q . When the differential pair is steered with a voltage of $\pm\sqrt{2}V_{OD}$, the current switches completely to one branch.

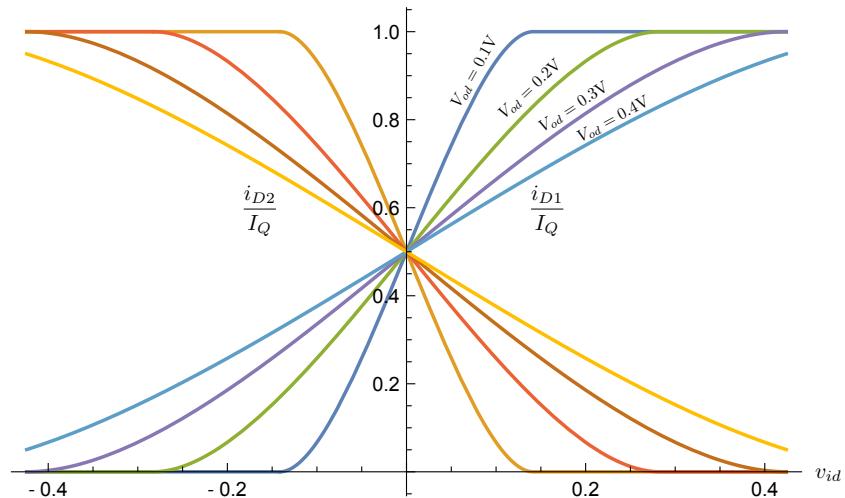


Figure 16.12: Family of currents in each branch of the differential pair circuit as a function of the input differential voltage v_{id} , with varying overdrive voltage V_{OD} . Larger overdrive results in a larger linear range.

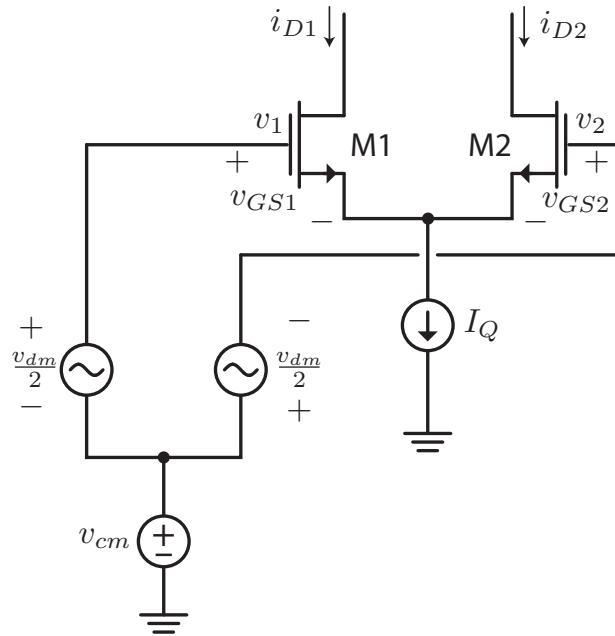


Figure 16.13: The differential pair driven with both a common-mode and differential mode signal.

16.5 Small-Signal Differential Circuits

16.5.1 General Differential Drive

The key to analyzing differential circuits is to apply a general input to the amplifier with voltages v_1 and v_2 , and express it as a combination of a "common-mode" signal v_{CM} , and a "differential-mode" signal v_{DM} :

$$v_1 = v_{CM} + \frac{v_{DM}}{2} \quad (16.30)$$

$$v_2 = v_{CM} - \frac{v_{DM}}{2} \quad (16.31)$$

To find v_{CM} , sum the signals so that v_{DM} cancels out:

$$v_1 + v_2 = 2v_{CM} \quad (16.32)$$

Eq. 16.32 shows that v_{CM} is the average input to the amplifier:

$$v_{CM} = \frac{v_1 + v_2}{2} \quad (16.33)$$

Next, take the difference between *Eq. 16.30* and *Eq. 16.31* to cancel out v_{CM} :

$$v_1 - v_2 = v_{DM} \quad (16.34)$$

Putting *Eq. 16.32* and *Eq. 16.34* together, we can form a system of equations:

$$\begin{bmatrix} v_{CM} \\ v_{DM} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ 1 & -1 \end{bmatrix} \cdot \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \quad (16.35)$$

These set of transformations are invertible, and one can go back and forth between the two representations. You can verify by matrix inversion that we can reconstruct our original signals:

$$\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 1 & \frac{1}{2} \\ 1 & -\frac{1}{2} \end{bmatrix} \cdot \begin{bmatrix} v_{CM} \\ v_{DM} \end{bmatrix} \quad (16.36)$$

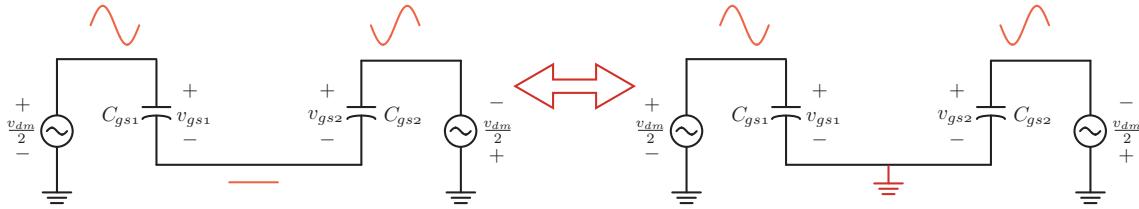


Figure 16.14: AC circuit schematic of a symmetric circuit driven by a pure differential signal. Since the common source node does not move, it can be grounded without changing the behavior of the circuit.

The reason we prefer to represent our signal in terms of a "common-mode" and "differential-mode" is due to the symmetry of the waveforms, v_{CM} and v_{DM} . Since the differential pair is fully symmetric, applying a "common-mode" signal has "even" symmetry. The left and right side of the circuit are excited with the same signal, so we expect the circuit to respond in the same way on the left and right hand sides. On the other hand, exciting a circuit with v_{DM} has "odd" symmetry, because every signal on the left should be the polar opposite of what is on the right. Finally, as discussed earlier, the differential signal is usually of interest, and we typically want to amplify this component of the signal. The AC portion of the common-mode signal is interference or noise, and it should be rejected.

16.5.2 Pure Differential-Mode Excitation

To take a concrete example, consider the circuit shown in Fig. 16.14. When the inputs are purely differential-mode, the middle "source" node does not experience any voltage fluctuation. This is true even if the source node has a finite impedance to ground. Due to superposition, we can say that any fluctuation induced from one side of the differential pair is canceled by the other side due to the symmetry. Therefore, we are free to place a "virtual" ground at this node and split the circuit in two, with a grounded source connection. Note that we have created a virtual ground without using a large bypass capacitor! Now you may appreciate why the large-signal analysis of the previous section resulted in a small-signal g_m that corresponded to a common source amplifier without degeneration.

16.5.3 Pure Common-Mode Excitation

Now consider the circuit shown in Fig. 16.15. When the inputs are purely common-mode, then by symmetry no current can flow from the "left" to "right", because all of the signals are in phase. Thus, we are free to cut this node and split the circuit into two identical circuits. Note that the common-mode half circuit differs from the differential-mode half circuit, as one is grounded while the other one is "floating". This leads to different values of common-mode and differential-mode gain.

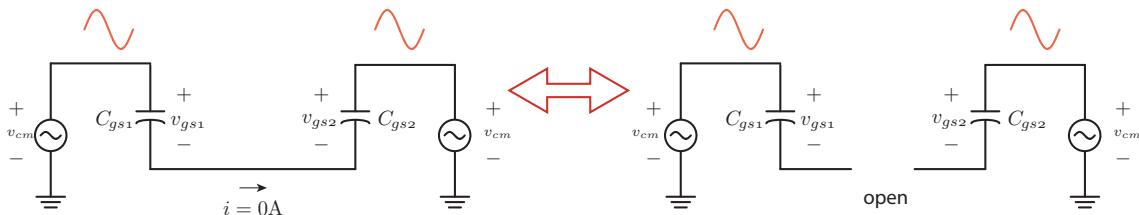


Figure 16.15: AC circuit schematic of a symmetric circuit driven by a pure common-mode signal. Since no current can travel through the middle branch, it can be cut without changing the behavior of the circuit.

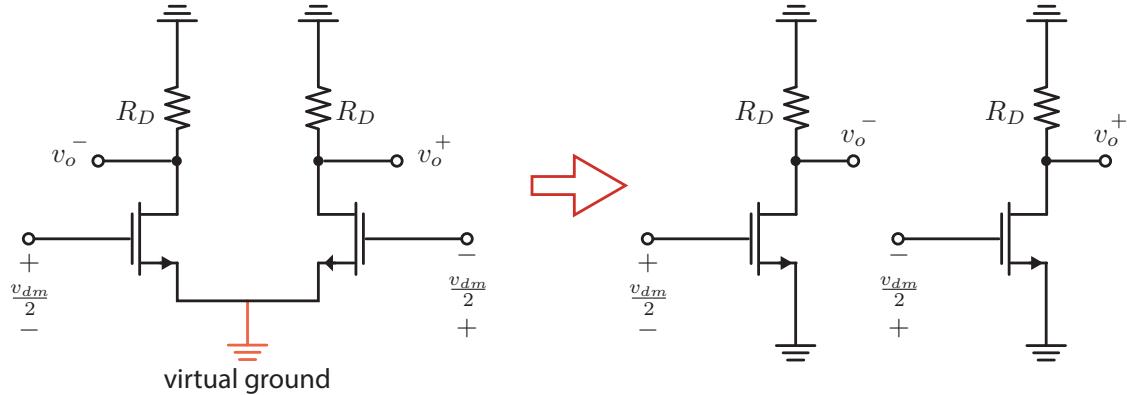


Figure 16.16: AC model of a differential pair under pure differential drive. The circuit can be split into two common-source amplifiers.

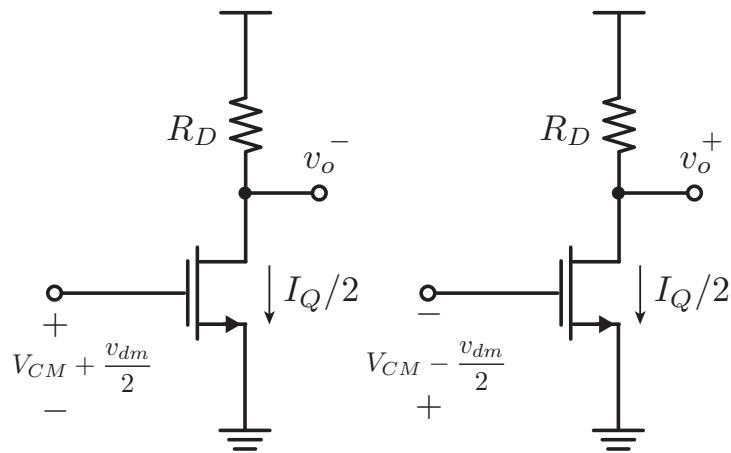


Figure 16.17: Schematic of the differential circuit biased with a DC voltage V_{CM} and driven with a pure differential drive.

16.5.4 Differential-Mode "Half Circuit"

Under differential-mode excitation, we now analyze each half of the pair independently. The circuit has been split into two, and each half is a simple CS amplifier, as shown in *Fig. 16.16*.

Differential-Mode Gain

The first step in analyzing any AC circuit is to find the DC operating point. Seen in *Fig. 16.17*, each transistor is biased at $\frac{I_Q}{2}$, with the gates at a common DC bias of V_{CM} . Note that the common-mode signal is DC, and not AC.:

$$v_{G1,2} = V_{CM} \pm \left(\frac{1}{2} \right) v_{DM} \quad (16.37)$$

The AC circuit it is excited by a pure differential-mode signal, and given by:

$$v_{g1,2} = \pm \left(\frac{1}{2} \right) v_{DM} \quad (16.38)$$

The output at the \pm nodes is:

$$v_o^- = -g_m \left(\frac{v_{DM}}{2} \right) R_D = -g_m R_D \frac{v_{DM}}{2} \quad (16.39)$$

$$v_o^+ = -g_m \left(-\frac{v_{DM}}{2} \right) R_D = g_m R_D \frac{v_{DM}}{2} \quad (16.40)$$

The differential output is the voltage $v_o = v_o^+ - v_o^-$, which leads to a differential-mode gain of:

$$A_d = \frac{v_o^+ - v_o^-}{v_{DM}} = g_m R_D$$

Differential-mode gain
(16.41)

16.5.5 Complete Small-Signal Differential and Common-Mode Models

The complete small-signal models for differential-mode and common-mode operation are shown in *Fig. 16.18* and *Fig. 16.19*. For the differential-mode circuit, since the schematic is split in the middle, we just have a common source amplifier. So everything we have learned about CS amplifiers can be transferred to the differential pair. We have been implicitly using these small-signal models in our analysis thus far. For the common-mode equivalent half circuit, it is noteworthy that the circuit is "floating", and thus it cannot process any signals. To see this, note that when a voltage is applied to the gate of one side, the source node will simply follow to ensure that $v_{gs} = 0V$. So the output current of the transistor is zero, effectively rejecting the common-mode signal. As far as the source node is considered, the circuit is a source follower after all. Without modeling the impedance of a current source, the common-mode rejection is infinite.

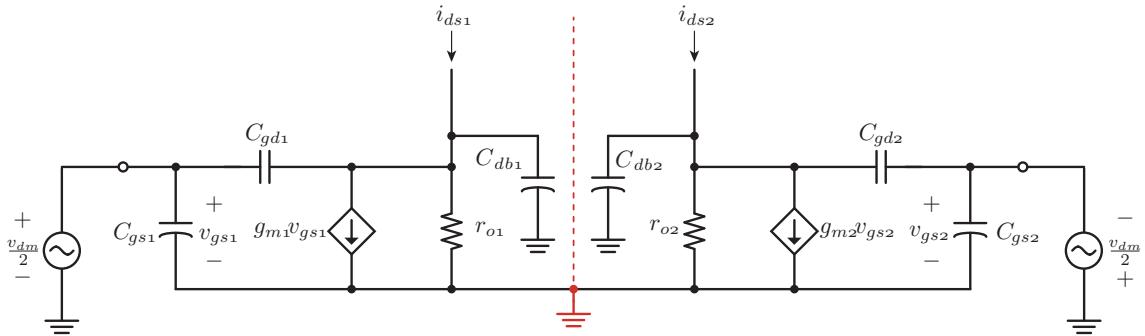


Figure 16.18: Small-signal AC schematic of a differential pair excited with a differential signal.

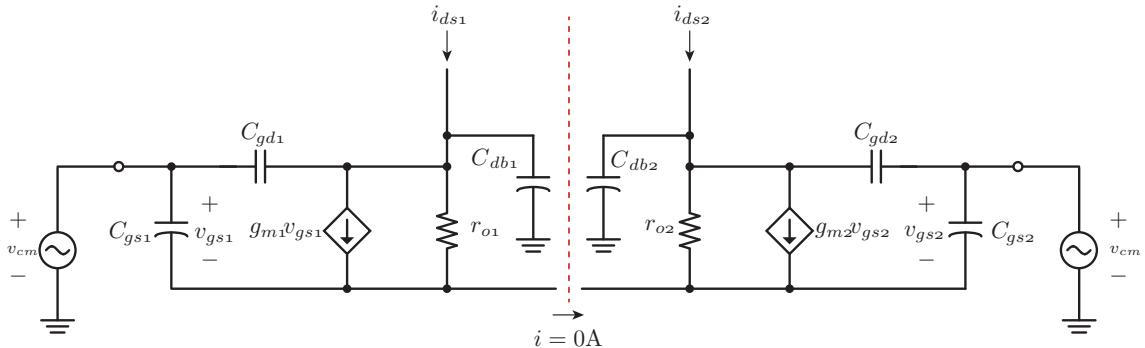


Figure 16.19: Small-signal AC schematic of a differential pair excited with a common-mode signal.

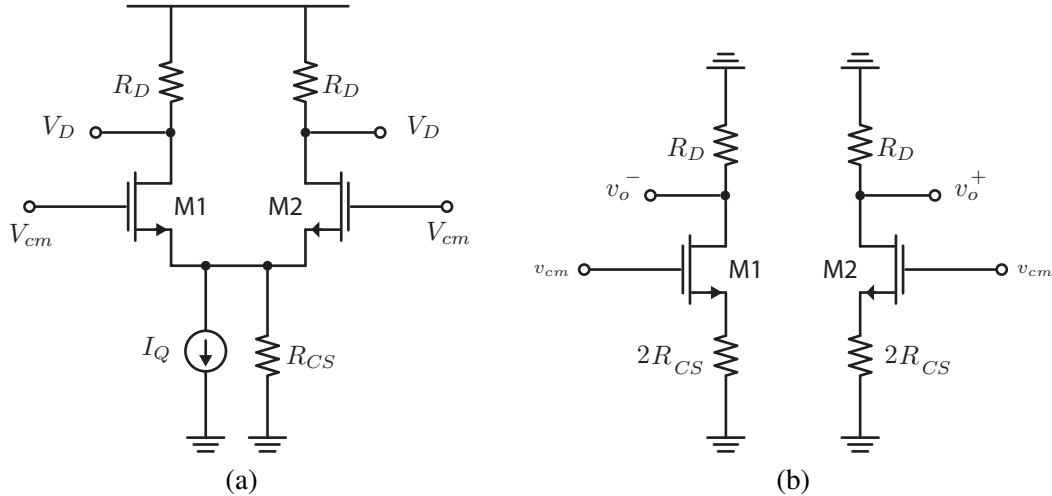


Figure 16.20: (a) A differential pair with a real current source must include the current source output impedance. (b) The AC circuit can be split in half due to symmetry. Splitting a resistor in half in the parallel direction is equivalent to two resistors in parallel, each with twice the resistance.

16.5.6 Common-Mode Operation - Real Current Source

The modified differential pair, shown in *Fig. 16.20a*, uses a real current source, with R_{CS} modeling the output impedance of the current source. By splitting R_{CS} in two parallel resistors $2R_{CS}$, the circuit is fully symmetric. This symmetry allows the half-circuit decomposition shown in *Fig. 16.20b*. Now, the source is no longer floating, meaning that common-mode signals will see a (small) gain. The gain can be made arbitrarily small by increasing R_{CS} relative to R_D , but lowering R_D also hurts the differential-mode gain. We need to consider both to understand how well a circuit can amplify differential signals while rejecting common-mode signals.

16.5.7 Differential and Common-Mode Gain

Obviously, the most important gain is the differential-mode gain:

$$A_{DM} = \frac{v_{o_{DM}}}{v_{i_{DM}}} \quad (16.42)$$

Also of concern, is the common-mode gain, because a common-mode signal could potentially "jam" the amplifier by causing it to rail:

$$A_{CM} = \frac{v_{o_{CM}}}{v_{i_{CM}}} \quad (16.43)$$

Notice that it is entirely possible for a common-mode signal to be converted to a differential-mode through any kind of imbalances or mismatches in the amplifier. This is especially problematic, because when this conversion occurs, the common-mode signal is indistinguishable from the differential-mode. Thus, we would like to minimize:

$$A_{CM \rightarrow DM} = \frac{v_{o_{DM}}}{v_{i_{CM}}} \quad (16.44)$$

In a fully balanced amplifier, with differential inputs and outputs and zero mismatch between the components, $A_{CM \rightarrow DM}$ is identically zero. But when the output is unbalanced (only one output is taken), or if there are mismatches between components, $A_{CM \rightarrow DM}$ can be of great importance. Finally, even though it is possible to convert differential-mode to common-mode, this is usually of less concern.

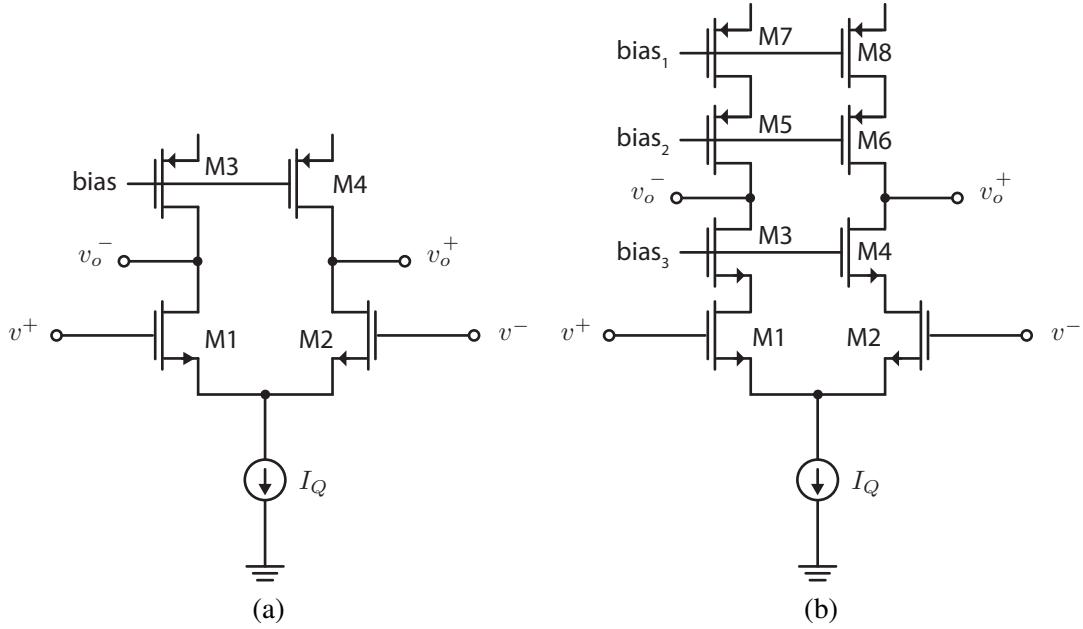


Figure 16.21: (a) Differential pair with a current mirror load. (b) A cascode differential pair with a cascode load.

16.5.8 Small-Signal Common-Mode Operation

Let's analyze Fig. 16.20b. Recall that \$R_{CS}\$ models any output resistance at the common-source node to ground. The gain in common-mode is simply the gain of a source degenerated amplifier:

$$A_{CM} = - \left(\frac{g_m}{1 + g_m 2 R_{CS}} \right) R_D \approx - \frac{R_D}{2 R_{CS}} \quad (16.45)$$

If we examine the differential output, the signal is zero (assuming perfect matching):

$$A_{CM \rightarrow DM} = 0 \quad (16.46)$$

16.5.9 Current-Source Loads

Similar to a common source amplifier, an **active load** is often preferred. Active loads occupy less area and require less headroom. The differential-mode gain with the active current source loads in Fig. 16.21a is simply given by:

$$A_{DM} = \frac{v_{o_{DM}}}{v_{i_{DM}}} = g_{m,1} (r_{o,1} \parallel r_{o,3}) \quad (16.47)$$

To increase the gain, at the expense of loss of swing, we can *cascode the load*. To reap the benefits of the higher output impedance of the load, we have to convert the differential pair into a **cascode differential pair** to boost the output impedance:

$$A_{DM} = \frac{v_{o_{DM}}}{v_{i_{DM}}} = g_{m,1} (R_{on} \parallel R_{op}) \quad (16.48)$$

From our knowledge of cascodes, we can estimate the impedance seen looking into the NMOS transistors:

$$R_{on} = (g_{m,3} r_{o,3}) r_{o,1} \quad (16.49)$$

And likewise the *PMOS* load transistors:

$$R_{op} = (g_{m,5} r_{o,5}) r_{o,7} \quad (16.50)$$

The overall output impedance is the parallel combination. For simplicity, assume the *NMOS* and *PMOS* loads present equal load impedance:

$$R_{on} = R_{op} = g_m r_o^2 \quad (16.51)$$

This means the maximum differential-mode gain is given by:

$$A_{DM} = \left(\frac{1}{2} \right) g_m^2 r_o^2 \quad (16.52)$$

We can boost the gain higher by using a triple cascode, but the loss of headroom is unacceptable for most applications.

16.6 Common-Mode Rejection Ratio and DC Offsets

16.6.1 Common-Mode Rejection Ratio

As we have emphasized, an important metric for a differential amplifier is its ability to reject the common-mode signal relative to the differential-mode signal. The ratio of the gain is known as the **common-mode rejection ratio CMRR**:

$$CMRR \equiv \frac{|A_{DM}|}{|A_{CM \rightarrow DM}|}$$

Common-mode rejection ratio (16.53)

Notice that an ideal (matched) amplifier has zero $A_{CM \rightarrow DM}$, and so it has an infinite *CMRR*. Certain applications, such as sensitive biomedical measurements require 70 dB - 100 dB of *CMRR* to reject environmental noise (such as 50/60 Hz AC power) compared to minuscule sensor signals!

16.6.2 Common-Mode Gain with Mismatched R_D

Suppose we have a mismatch in the load resistors. Then the gain from one side of the amplifier will not match the gain from the other side, causing the common-mode signal to be converted to differential-mode. Let's say one resistor is larger by ΔR_D :

$$R_{D_{1,2}} = R_D \pm \frac{\Delta R_D}{2} \quad (16.54)$$

Now calculate the differential-mode output for a common-mode input:

$$v_{o_{DM}} = v_{o,2} - v_{o,1} = \frac{-\Delta R_D}{2R_{CS}} v_{i_{CM}} \quad (16.55)$$

This leads to *common-mode to differential-mode gain*:

$$A_{CM \rightarrow DM} = \frac{v_{o_{DM}}}{v_{i_{CM}}} = \frac{-\Delta R_D}{2R_{CS}} = \left(\frac{-R_D}{2R_{CS}} \right) \left(\frac{\Delta R_D}{R_D} \right) \quad (16.56)$$

The *CMRR* depends on the precision of the resistors, and the achievable output resistance of the current source:

$$CMRR = \frac{g_m R_D}{\frac{\Delta R_D}{2R_{CS}}} = \frac{2g_m R_{CS}}{\frac{\Delta R_D}{R_D}} \quad (16.57)$$

16.6.3 Common-Mode Gain with Mismatch of g_m

In practice, the transistors on each side of a differential pair are not identical. Each transistor will have a different threshold voltage V_T , and slightly different dimensions, W and L . So we should also expect variations in C_{ox} due to changes in t_{ox} . To minimize these variations, differential pairs in integrated circuits are laid out very carefully with various schemes designed to maximize the matching. Essentially each transistor is broken down into parallel unit amplifiers that are connected in various inter-digitated ways. Doing this minimizes variations due to process variables and temperature gradients (see Fig. 13.10).

To get a sense for the impact in the transistor mismatch, let's consider a transistor mismatch in g_m :

$$g_{m,1} = g_m + \left(\frac{1}{2}\right) \Delta g_m \quad (16.58)$$

$$g_{m,2} = g_m - \left(\frac{1}{2}\right) \Delta g_m \quad (16.59)$$

The difference in the g_m is given by:

$$g_{m,1} - g_{m,2} = \Delta g_m \quad (16.60)$$

It is fairly easy to show that:

$$A_{CM \rightarrow DM} = \frac{v_{o_{DM}}}{v_{i_{CM}}} = \left(\frac{R_D}{2R_{CS}}\right) \left(\frac{\Delta g_m}{g_m}\right) \quad (16.61)$$

Again, we desire high current source output resistance and good matching precision between the transistors. The $CMRR$ is given by:

$$CMRR = \frac{g_m R_D}{\left(\frac{R_D}{2R_{CS}}\right) \left(\frac{\Delta g_m}{g_m}\right)} = \frac{2g_m R_{CS}}{\frac{\Delta g_m}{g_m}} \quad (16.62)$$

16.6.4 DC Offset

Suppose we tie both inputs of a differential pair to the same DC voltage V_{CM} , as shown in Fig. 16.22a. We expect, due to symmetry, for the outputs to be at precisely the same voltage, producing a zero output differential voltage. In practice, due to mismatches in the device g_m 's and load resistors R_D , the differential output will be non-zero. Now suppose we connect a voltage source at the input to "zero out" the offset, as shown in Fig. 16.22b. This voltage is known as the **offset voltage** V_{OS} . The value of V_{OS} is simply the output DC offset divided by the gain of the amplifier. This voltage is known as the *input-referred offset voltage*.

We can calculate the offset by considering various sources of mismatch within the amplifier, as we did when we calculated the common-mode to differential-mode conversion. Let's consider a resistor load mismatch:

$$R_{D_{1,2}} = R_D \pm \frac{\Delta R_D}{2} \quad (16.63)$$

The offset voltage is defined as the DC output for zero input. Because the transistors are well matched, and the source and drain nodes are decoupled, we still expect the bias current to split evenly. Assuming that each transistor is biased at $\frac{I_Q}{2}$, we simply have:

$$V_o = V_{D_2} - V_{D_1} = \left(\frac{I_Q}{2}\right) \Delta R_D \quad (16.64)$$

Dividing by the DC gain, the input-referred offset voltage is:

$$V_{OS} = \frac{V_O}{A_{DM}} = \frac{\frac{I_Q \Delta R_D}{2}}{g_m R_D} = \frac{\frac{I_Q \Delta R_D}{2}}{2 \left(\frac{I_Q}{V_{OD}} \right) R_D} = \left(\frac{V_{OD}}{2} \right) \left(\frac{\Delta R_D}{R_D} \right) \quad (16.65)$$

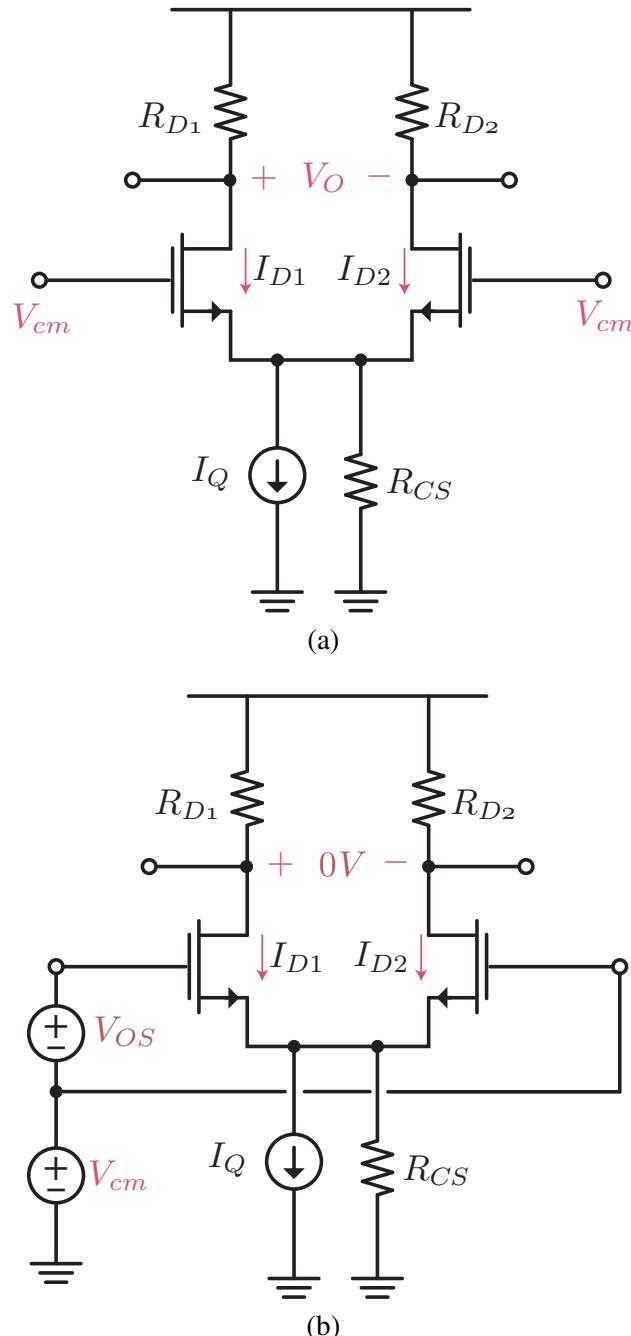


Figure 16.22: (a) The offset voltage V_O is defined as the output DC voltage when the inputs are balanced (zero differential drive). (b) The offset can be referred to the input as V_{OS} , or the required DC voltage applied to the inputs to zero-out the offset voltage.

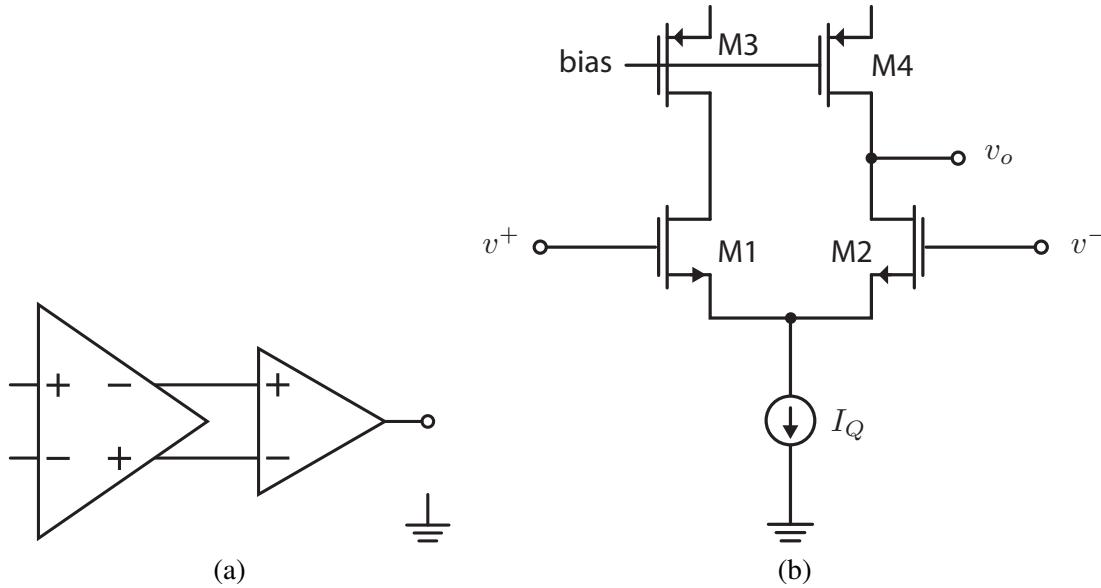


Figure 16.23: (a) In many applications we desire to process the circuit differentially, and to convert the result to a single-ended signal as shown. (b) Taking the single-ended output is one way to achieve this, although there is a better technique introduced in Sec. 16.7.2.

16.7 Current Mirror Load

16.7.1 Differential Input, Single-End Output

So far we have designed a circuit with differential inputs and differential outputs, but in many cases we want a single-ended output. This is because many components outside of our control are single-ended. In some situations it is a matter of practicality, for example when dozens of signals are routed from one chip to another, a single-ended operation requires half the number of signal lines.

As shown in Fig. 16.23a, we desire a way to convert the differential signal to a single-ended signal. The simplest choice is to just take one of the outputs of the differential amplifier, as shown in Fig. 16.23b. This works, but we are compromising by giving up half of the gain, and severely impacting the common-mode rejection. Recall that that the differential pair has common-mode gain that is rejected only when we consider the output differentially. With a single-ended output, the common-mode gain is automatically converted to a single-ended signal, and it cannot be separated from the desired signal.

16.7.2 Current Mirror Load

An elegant solution that generates a single-ended output from the differential signal is shown in Fig. 16.24. We draw the AC equivalent circuit for differential input and assume that $M1$ and $M2$ generate equal and opposite currents. Even though the circuit is not fully symmetric, we continue to make this assumption because we are assuming that $i_D = f(v_{GS})$, and not v_{DS} . For the transistor $M1$, the current flows into the MOS diode load, which generates the required v_{GS} to generate the same current for $M2$. Therefore, the current flowing from the output side is the difference of the mirror current and the transistor current $M2$. Since these currents are equal and of opposite phase, the output current flowing into a load is $2 \times i$.

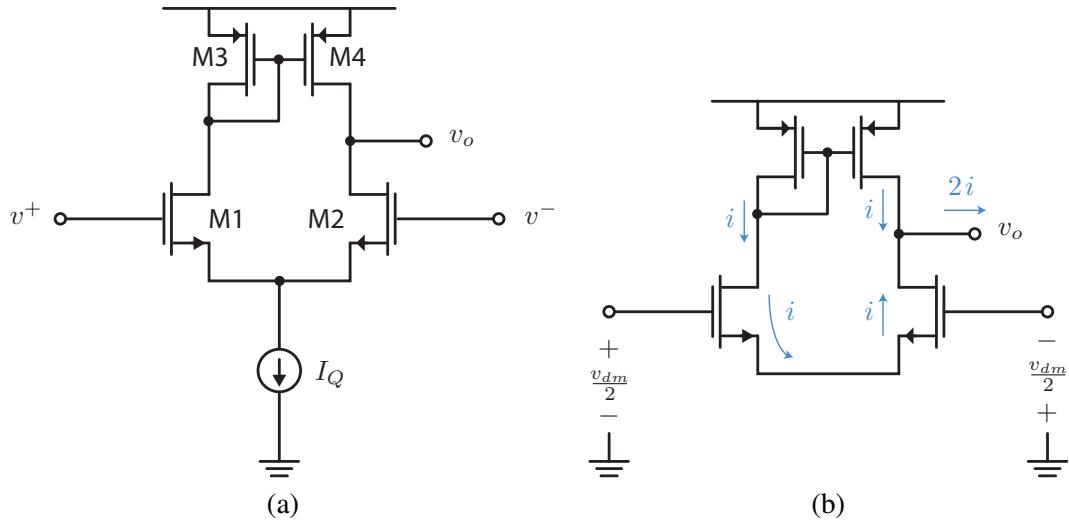


Figure 16.24: (a) Differential pair with a current mirror load. (b) The output of M_1 is mirrored to M_2 through the load. In the process, the single-ended output is doubled in amplitude, and the common-mode signal is rejected.

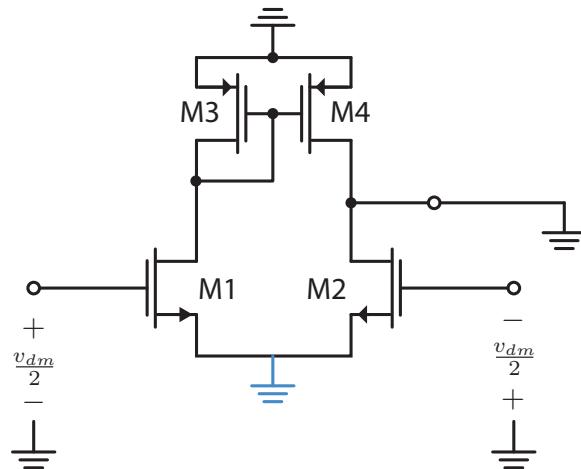


Figure 16.25: AC analysis of a differential pair with a current mirror load using the virtual ground. Even though the circuit is not fully symmetric, since M_3 is configured as a MOS diode and M_4 is not, this model is still very accurate since the transistors M_1/M_2 are gate and not drain controlled.

We can analyze the circuit in more detail using the differential-mode half-circuit as shown in Fig. 16.25. Again, the circuit is not fully symmetric, because the mirror has a diode connection on one side. We are relying on the fact that the transistor is only a weak function of the drain voltage. The differential transconductance is defined by:

$$G_m = \frac{i_o}{v_{DM}} \quad (16.66)$$

In Eq. 16.66, i_o is a current flowing into a short circuit at the output. This output current is the difference between the current generated by M_2 , and the one generated by M_4 :

$$i_o = g_{m,2} \left(\frac{v_{DM}}{2} \right) - g_{m,4} v_{gs,4} \quad (16.67)$$

The voltage $v_{gs,4}$ is the same as $v_{gs,3}$, and since we know the current and impedance level at the gate and drain of $M3$, we have:

$$v_{gs,4} = v_{gs,3} = -g_{m,1} \left(\frac{v_{DM}}{2} \right) \left(\frac{1}{g_{m,3}} \parallel r_{o,3} \parallel r_{o,1} \right) \quad (16.68)$$

The impedance at the drain of $M1$ is dominated by the diode connection. So it is a good approximation to assume it's dominated by $\frac{1}{g_{m,3}}$:

$$\approx - \left(\frac{g_{m,1}}{g_{m,3}} \right) \left(\frac{v_{DM}}{2} \right) \quad (16.69)$$

Let's assume that the transistors are well matched. Then, since $M1$ and $M2$ are biased with the same DC current:

$$g_{m,1} = g_{m,2} = g_m \quad (16.70)$$

The same is true for $M3$ and $M4$:

$$g_{m,3} = g_{m,4} \quad (16.71)$$

Substitution in *Eq. 16.67* results in the following:

$$i_o = g_m v_{DM} = G_m v_{DM} \quad (16.72)$$

This result is not at all surprising. Since the transistors $M3$ and $M4$ form a mirror, we already predicted the doubling of the G_m compared to a single-ended output without the mirror.

We can now calculate the differential-mode gain. Since the output resistance is the parallel combination of the mirror loads looking both up and down:

$$R_{out} = R_{op} \parallel R_{on} = r_{o,4} \parallel r_{o,2} \quad (16.73)$$

The gain is simply given by:

$$A_{DM} = G_m R_{out} = g_m (r_{o,4} \parallel r_{o,2}) \quad (16.74)$$

16.7.3 Common-Mode Gain

At first glance you might guess that the circuit ideally does not convert any of its common-mode output to a differential-mode, and you're almost correct. However, there is a residual mismatch that is due to the fact that the mirror is not perfect. Let's analyze this in detail.

For the common-mode excitation shown in *Fig. 16.26*, the drain current of $M4$ is a result of the gate-source voltage generated by $M3$, as previously noted. Now we take into account the actual impedance at the drain of $M1$ to calculate $v_{gs,3}$:

$$i_{d,4} = g_{m,3} v_{gs,3} = -g_{m,3} i_{d,1} \left(\frac{1}{g_{m,3}} \parallel r_{o,3} \parallel r_{o,1} \right) \quad (16.75)$$

At this point we do not make the common approximation $\frac{1}{g_m} \ll r_o$, because we are trying to calculate a very small quantity. So throwing this term away would also result in losing the baby with the bathwater. Now let's just continue to calculate $i_{d,4}$ using the complete expression:

$$i_{d,4} = -i_{d,1} \left(\frac{g_{m,3}(r_{o,1} \parallel r_{o,3})}{1 + g_{m,3}(r_{o,1} \parallel r_{o,3})} \right) \quad (16.76)$$

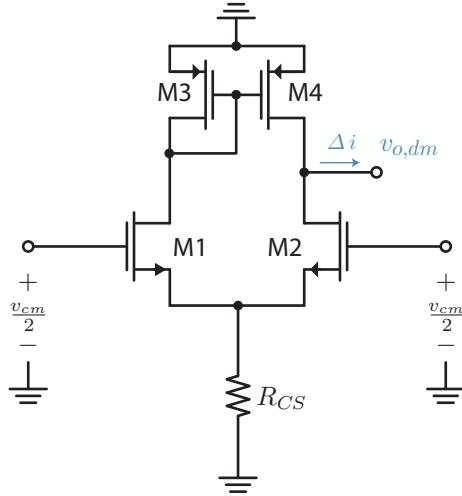


Figure 16.26: Schematic of a differential pair with a current mirror load under a common-mode excitation. There is a small residual current Δi that is converted into a differential mode due to the slight error in the mirror due to output resistance.

This current sums with $i_{d,2}$, and of course $i_{d,2} = i_{d,1}$, as the input is a common-mode signal. Any difference between $i_{d,2}$ and $i_{d,4}$ generates a differential-mode signal, which is highly undesirable:

$$\Delta i = i_{d,2} - |i_{d,4}| = i_{d,1} \left(\frac{1}{1 + g_{m,3}(r_{o,1} \parallel r_{o,3})} \right) \quad (16.77)$$

Fortunately this Δi current is very small, because $g_m r_o \gg 1$. The differential output voltage is given by:

$$v_{o,CM \rightarrow DM} = \Delta i (r_{o,2} \parallel r_{o,4}) = i_{d,1} \left(\frac{1}{1 + g_{m,3}(r_{o,1} \parallel r_{o,3})} \right) (r_{o,2} \parallel r_{o,4}) \quad (16.78)$$

As $i_{d,1}$ is generated by a common-mode input, we can use the degenerated transconductance for a common source amplifier:

$$i_{d,1} = \left(\frac{g_{m,1}}{1 + g_{m,1} 2R_{CS}} \right) v_{i_{CM}} \approx \left(\frac{1}{2R_{CS}} \right) v_{i_{CM}} \quad (16.79)$$

Putting this all together, the gain $A_{CM \rightarrow DM}$ is given by:

$$A_{CM \rightarrow DM} \approx \left(\frac{1}{g_{m,3}(r_{o,1} \parallel r_{o,3})} \right) \left(\frac{1}{2R_{CS}} \right) (r_{o,2} \parallel r_{o,4}) = \frac{1}{2g_{m,3}R_{CS}} \quad (16.80)$$

Most importantly, if we compare this gain to the differential-mode gain, we have the $CMRR$:

$$CMRR = \frac{|A_{DM}|}{|A_{CM \rightarrow DM}|} = g_{m,1} (r_{o,2} \parallel r_{o,4}) \times (2R_{CS}g_{m,3}) \gg 1 \quad (16.81)$$

This is much better than the $CMRR$ of a circuit that does not use a mirror and uses one output instead.

16.8 Appendix: Large Signal Derivation Steps

It is not too difficult to show the validity of *Eq. 16.23*. To begin, let's eliminate v_{GS} from the expressions by taking the difference between $\sqrt{i_D}$'s of the transistors:

$$\sqrt{i_{D_1}} - \sqrt{i_{D_2}} = \sqrt{\frac{k_n}{2}} (v_{GS_1} - v_{GS_2}) = \sqrt{\frac{k_n}{2}} \cdot v_{id} \quad (16.82)$$

This is needed to express a linear difference between the v_{GS} voltages of $M1$ and $M2$, which is by definition the input differential voltage. Next let's square $\sqrt{i_{D_1}} - \sqrt{i_{D_2}}$:

$$i_{D_1} - 2\sqrt{i_{D_1} i_{D_2}} + i_{D_2} = \frac{k_n}{2} (v_{id})^2 \quad (16.83)$$

Re-arranging the equation, we have:

$$I_Q - 2\sqrt{i_{D_1} i_{D_2}} = \frac{k_n}{2} (v_{id})^2 \quad (16.84)$$

Now use $i_{D_2} = I_Q - i_{D_1}$:

$$I_Q - \frac{k_n}{2} (v_{id})^2 = 2\sqrt{i_{D_1} (I_Q - i_{D_1})} \quad (16.85)$$

Squaring the result:

$$\left(I_Q - \frac{k_n}{2} (v_{id})^2 \right)^2 = 4i_{D_1} (I_Q - i_{D_1}) = 4i_{D_1} I_Q - 4(i_{D_1})^2 \quad (16.86)$$

We now have all the elements to setup a quadratic equation:

$$i_{D_1}^2 - I_Q i_{D_1} + \frac{1}{4} \left(I_Q - \left(\frac{k_n}{2} \right) v_{id}^2 \right)^2 = 0 \quad (16.87)$$

Solving the quadratic:

$$i_{D_1} = \frac{I_Q}{2} \pm \frac{1}{2} \sqrt{I_Q^2 - 4 \left(\frac{1}{4} \right) \left(I_Q - \left(\frac{k_n}{2} \right) v_{id}^2 \right)^2} \quad (16.88)$$

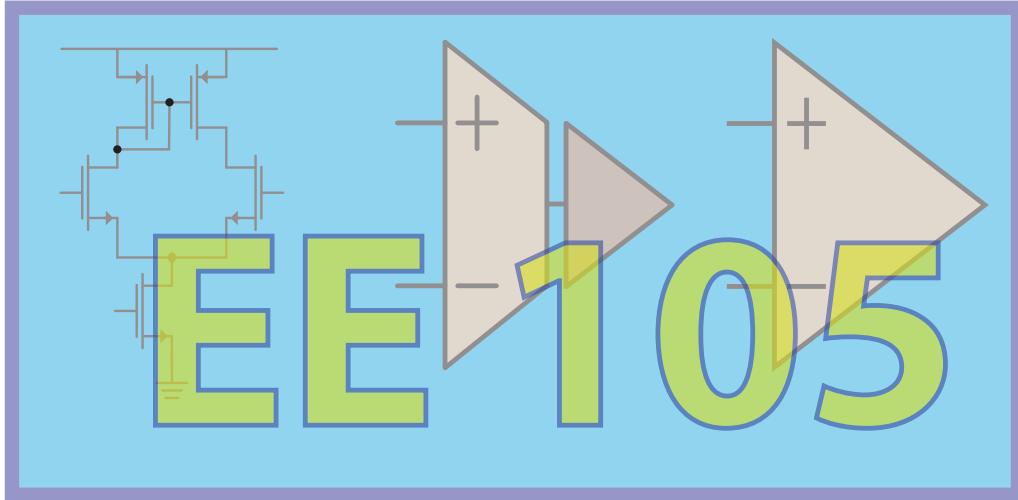
Which is almost in the desired form. Simply factor out I_Q :

$$i_{D_1} = \frac{I_Q}{2} \pm \frac{I_Q}{2} \sqrt{1 - \left[1 - \left(\frac{k_n}{2} \right) \left(\frac{v_{id}^2}{I_Q} \right) \right]^2} \quad (16.89)$$

Finally note definition of V_{OD} :

$$V_{OD} = \sqrt{\frac{2I_Q}{k_n}} \quad (16.90)$$

and substitute and simplify to obtain final form shown in *Eq. 16.23*.



17. Op-Amp Feedback and Frequency Response

17.1 Chapter Preview

In this chapter we will explore operational amplifiers (see *Fig. 17.1*), usually referred to as op-amps, and feedback. In particular, we will see how feedback impacts the frequency response of operational amplifiers. We assume some familiarity with ideal op-amps, and the "Golden Rules" for analyzing op-amp circuits, but our goal is to learn to analyze actual op-amps, not ideal ones.

We will set the context of the chapter by reviewing feedback theory in general, and motivate why feedback is important and useful. To analyze op-amps, we need an equivalent circuit model, something more sophisticated than the "Golden Rules" (infinite gain, infinite input impedance), but not too complicated. We will be inspired by a physical model based on the real inner workings of an op-amp, and armed with this model we will discuss important concepts such as the gain-bandwidth product, the unity gain frequency, and the gain-bandwidth trade offs.

We end the chapter by introducing some more advanced topics, such as feedback and stability. We don't do justice to this topic, our goal for now is to make the reader aware of the issues without understanding the details of how the issue is addressed. This topic should be pursued in a more advanced course on analog circuit design.

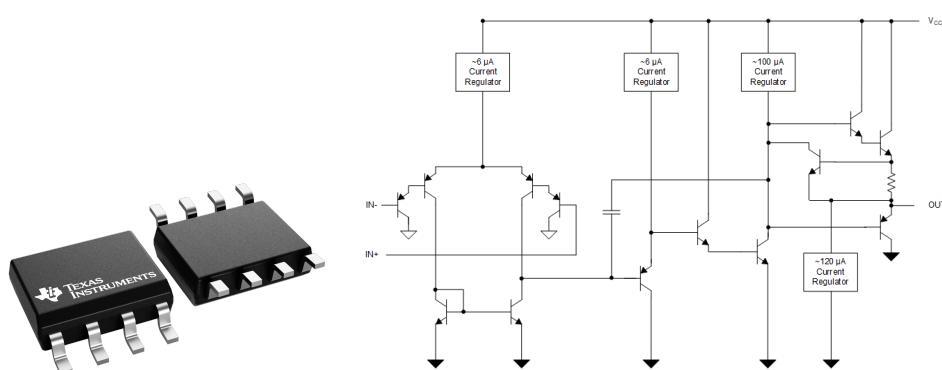


Figure 17.1: Schematic of an TI's LM358 operational amplifier, ubiquitous in audio applications.



Figure 17.2: The cruise control system in an automobile is an example of a feedback control system.

17.2 Introduction to Feedback

17.2.1 Feedback Control

Feedback is a universal way to design a system that is characterized by a lot of unknowns. Think about the thermostat in your house, or the cruise control system in your car. Humans also rely extensively on feedback – everything from our endocrine system (hormones) to the very act of walking, relies on positive and negative feedback loops.

Let's focus on the speed control or cruise control system in particular (see *Fig. 17.2*). The important steps to run the feedback loop can be summarized as follows:

- We input the desired speed of car (v_{des}).
- The car measure the actual speed of car (v_{car}).
- The control system generates an error signal, which is the difference between the current speed and the desired speed ($v_{err} = v_{des} - v_{car}$).

Based on v_{err} , the system takes action. For example, adjusting the car speed (accelerate or decelerate) based on the signal $K_p v_{err}$ where K_p is a proportionality constant.¹

What is important to take away from the above example is that the control system is somewhat blind to the details of how the car or engine works, it just tries to set the speed by adjusting the current speed. This is a powerful model for designing systems, because it allows us to abstract away all the details and focus on the problem at hand. Your thermostat in your house is a perfect example of this as it has no knowledge about the size of your house, the number of windows, the presence of insulation in the walls, etc. It simply turns on the heater when the temperature is lower than desired, and keeps the heat on until the desired temperature has been obtained.

17.2.2 Negative Feedback Block Diagram

Let's analyze a feedback system using equations. The model for a general feedback system is shown in *Fig. 17.3*. This is a negative feedback system because we subtract the feedback signal from the input in order to generate the error signal. This subtraction is represented by the (+) with the circle around it, and the (-) underneath it in the figure.

To find the transfer function, note that the error signal is a function of the input, and output after it has been fed back through the "f" box:

$$s_{err} = s_{in} - f \cdot s_{out} \quad (17.1)$$

¹More complicated control schemes can also take action based on not only the error signal, but also based on the integral or derivative of the error signal, altering the dynamics of the system. These controllers are called "PID" controllers because they respond to changes Proportional to the Integral or Derivative of the error system.

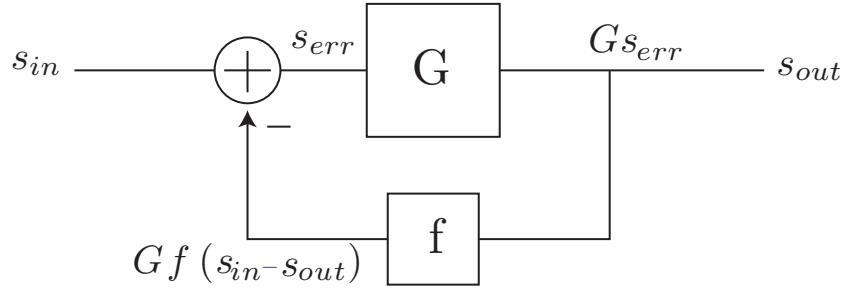


Figure 17.3: Block diagram model of a feedback system. G is the forward gain and f is the feedback factor. The input signal s_{in} is combined with a fraction of the output $f s_{out}$ and forms the error signal s_{err} .

The output is the amplified error signal:

$$s_{out} = G \cdot s_{err} = G \cdot (s_{in} - f \cdot s_{out}) \quad (17.2)$$

There is a bit of strangeness in Eq. 17.2, because the output is a function of itself. In other words, the current output depends on the current output and the current input, which is somewhat recursive or self-referential. From a mathematical perspective, there is nothing wrong with this equation, and we can readily solve it. But, from an intuition perspective, we can resolve this circular reference by imagining a slight delay in the above equation:

$$s_{out}(t) = G \cdot s_{err}(t) = G \cdot (s_{in}(t) - f \cdot s_{out}(t - \tau)) \quad (17.3)$$

In fact, this models real systems better because often there is a delay in sampling the output signal and applying it to the control system. In electronic systems this delay is actually vanishingly small (due to the large velocity of light propagation), so we ignore it for the rest of the chapter.²

Closed-Loop Transfer Function

To solve for the transfer function, we group terms involving s_{out} in Eq. 17.2, and solve:

$$s_{out} = G \cdot (s_{in} - f s_{out}) \quad (17.4)$$

$$= G \cdot s_{in} - G \cdot f s_{out} \quad (17.5)$$

Moving both s_{out} terms to the LHS, and factoring, we have:

$$s_{out}(1 + Gf) = G \cdot s_{in} \quad (17.6)$$

This leads to:

$$G_{closed} = \frac{s_{out}}{s_{in}} = \frac{G}{1 + Gf}$$

Closed-loop transfer function (17.7)

For very large gain G , such that $Gf \gg 1$, we have:

$$G_{closed} \approx \frac{G}{Gf} = \frac{1}{f}$$

Feedback factor (17.8)

This end result is very interesting because it says the transfer function does not depend on G , but only on the feedback! Imagine if G is an unknown or ill-defined function. For example, if we design an amplifier, the gain will vary due to process variations, temperature, voltage variations (supply), or through aging. All of these effects make it very difficult to build a reliable and precise amplifier without feedback. With feedback, we can make a precise gain if we can make a well defined **feedback factor f** .

²There is a delay in the system due to the frequency phase response of the gain G , and this will be explained later on. If the delay is not small, the system can be analyzed using Laplace Transform techniques.

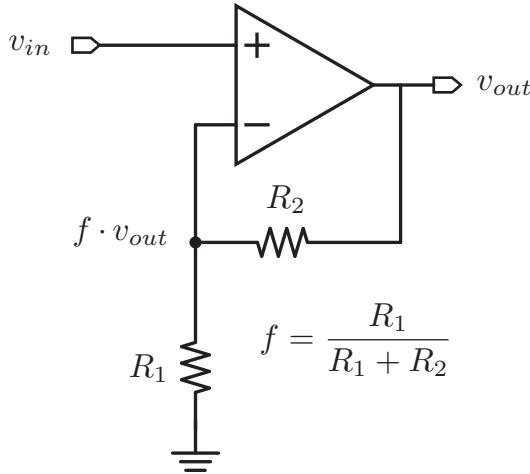


Figure 17.4: A non-inverting amplifier is a perfect demonstration of an electronic feedback system. The resistive divider forms the feedback network and the error signal is calculated from the op-amp differential input.

17.2.3 Electronic Feedback

We have already seen electronic feedback in op-amps, but you probably did not think of it that way. For example, the non-inverting amplifier shown in Fig. 17.4 can be mapped to a feedback system very easily. The resistor divider samples the output voltage, and the error signal is formed at the input of the op-amp. The op-amp output is an amplified copy of the error signal. A typical op-amp has a very large value of gain, making Gf large, and so the gain is very accurately predicated by Eq. 17.8:

$$G \approx \frac{1}{f} = \frac{R_1 + R_2}{R_1} = 1 + \frac{R_2}{R_1} \quad (17.9)$$

Notice that the gain of the op-amp is immaterial as long as it is large enough. We may now appreciate the origin of the Golden Rules. If the gain is allowed to go to infinity, then the error signal must be zero, and the output is $0 \cdot \infty$ or $\frac{0}{0}$, which is undefined unless you take the limit of gain as it approaches infinity. All practical op-amps have finite gain, so the error signal is nearly zero, and the Golden Rules are very useful.

17.2.4 History

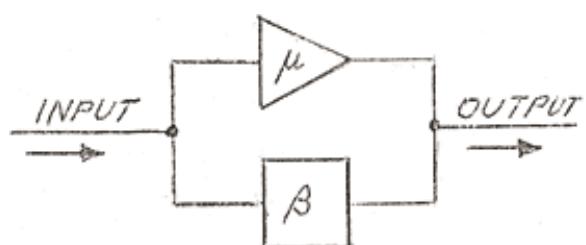
Feedback in electronic systems was invented by Harold Black. He was on a ferry ride to Bell Labs (1927) thinking about how to increase the gain and improve linearity of vacuum tube amplifiers, when in a flash of genius he realized that *positive* feedback can be used to boost the gain, and *negative* feedback could improve the linearity (see Fig. 17.5).

This is a good example of a "back of the envelope" calculation (in this case a newspaper) that changed the world. This idea came to him after tirelessly working on the problem over the course of years, trying to improve the linearity of amplifiers. All the open-loop vacuum tube³ amplifier topologies that he tried were plagued by problems. But the feedback amplifier proved to be a great success.

³A vacuum tube is similar to a transistor.



(a)



$$\frac{OUTPUT}{INPUT} = A_F = \frac{\mu}{1-\mu\beta} = \frac{1}{-\beta} \left[1 - \frac{1}{1-\mu\beta} \right]$$

(b)

Figure 17.5: Harold Black's original "back of the envelope" calculations that led to the invention of electronic feedback amplifiers.

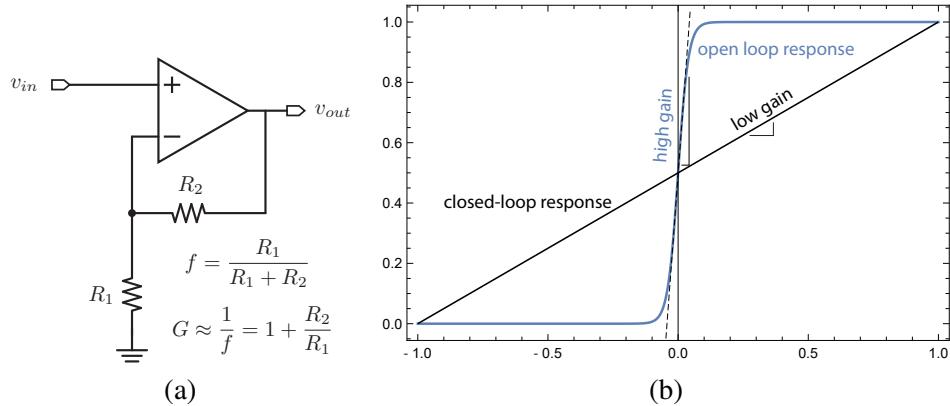


Figure 17.6: (a) A precision gain stage is realized with a feedback amplifier using an amplifier described by a very non-linear input-output transfer curve, shown in (b). The system is linear because it transforms a high gain non-linear amplifier into a lower gain linear system by exercising the amplifier with small inputs (error signal).

17.3 Why Feedback?

17.3.1 Precision Analog

As already noted, feedback allows us to use a really "crappy" *open-loop* amplifier and still get good performance in a *closed-loop* feedback system, as long as the crappy amplifier has high gain. When the gain is sufficiently high, the gain is determined by the feedback network, not the open-loop gain. So the open loop gain can vary over temperature, process, it can age, and it can be highly non-linear. In the end, if the gain is high enough, it does not play a role!

17.3.2 Other Benefits of Feedback

In a feedback system, the closed-loop gain depends on the passive feedback network, which is set precisely using well defined components such as resistors. It is also dependent on a ratio of resistors, which reduces the effects of temperature variations or aging. The op-amp does not even need to be particularly linear. In fact, when an amplifier has a very high gain, say a million, and is running on a supply voltage of say 5V, then any signal larger than $5\mu\text{V}$ will saturate the amplifier, as illustrated in Fig. 17.6b. How is it possible to make such a non-linear amplifier behave linearly? Well, recall that the op-amp processes the error signal, not the input signal. The error signal will be very small, and restricted to the linear range of the op-amp. This results in a closed-loop system with precise gain over a large range of inputs. So overall the op-amp transfer function is almost perfectly linear, despite using a non-linear core amplifier.

In the next section, we will highlight other important benefits, such as bandwidth enhancement.

17.3.3 Loop Gain

We have seen that many of the benefits of feedback occur when the gain is large. More precisely, if we examine the closed-loop transfer function:

$$G_{closed} = \frac{G}{1+Gf} = \frac{G}{1+T} \quad (17.10)$$

where $T = Gf$, we see that it's important for T to be large, not just G . T is called the **loop gain**, because it is the gain going around the loop, shown in Fig. 17.7a. For a precise transfer function, the key to feedback is to realize sufficiently high loop gain:

$T = Gf \gg 1$	<i>Loop gain</i>	(17.11)
----------------	------------------	---------

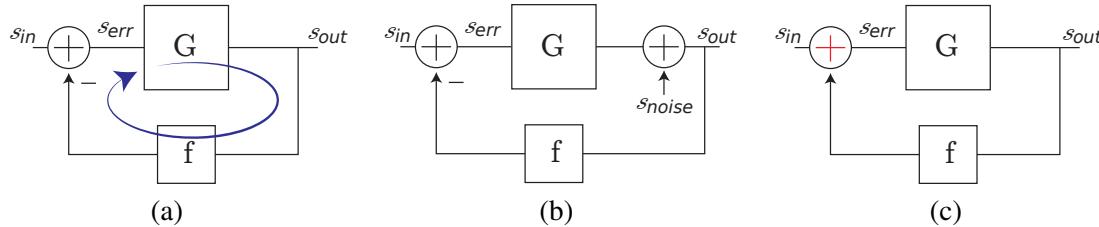


Figure 17.7: (a) The loop gain T is defined as the gain around the loop as shown. (b) Any noise or distortion injected at the output of the amplifier is rejected by the loop gain of the amplifier. (c) The block diagram of a positive feedback system.

17.3.4 Noise Rejection

Now consider the effect of injecting interference or noise into a feedback system, as shown in Fig. 17.7b. **Interference rejection** means that the loop can correct for unwanted signals that are injected into the signal path. Imagine an unwanted signal couples into the loop as shown. The transfer function can be derived again:

$$s_{out} = G s_{err} + s_{noise} = G(s_{in} - f \cdot s_{out}) + s_{noise} \quad (17.12)$$

Collecting s_{out} terms:

$$s_{out} = \left(\frac{G}{1+T} \right) s_{in} + \left(\frac{1}{1+T} \right) s_{noise} \quad (17.13)$$

If $T \gg 1$, then the noise is rejected by $1/(1+T)$. Any unwanted signal, including distortion, is rejected by the loop.

17.3.5 Positive Feedback

Up to now we have been considering a **negative feedback** system, whereby the output is subtracted from the input. If we were to add the output to the input, the system would be a **positive feedback** system, as shown in Fig. 17.7c.

Positive feedback is also useful and has applications, but they are very different. For example, positive feedback systems tend to “rail out”, in other words they can be regenerative. This is great for building **hysteresis** or a memory element, and indeed a basic SRAM cell is a positive feedback loop involving two CMOS inverters. Positive feedback systems can still be designed to provide gain, but the loop gain should be less than unity, $T < 1$. The reason for this is clear if we imagine that the current output is the input gained up, plus an infinite number of copies that go around the loop and add to the output. This is exactly a **geometric series**⁴, and it converges as long as $T < 1$, and it converges to a value of

$$G_{closed} = G \cdot \sum_{i=0}^{\infty} T^i = \frac{G}{1-T} \quad (17.14)$$

The benefit over negative feedback is that positive feedback can boost the gain, and in fact if we make $T = 1 - \epsilon$, where ϵ is a small number, we can make an arbitrarily large gain. The danger is that the gain is hard to control and can easily result in $T > 1$, causing a completely different behavior (such as oscillation or rail-out). In practice, in the design of linear analog circuits, positive feedback is used sparingly. When positive feedback is used, the designer must be aware of the feedback loop and ensure that under all conditions the loop gain does not exceed 1.

⁴See appendix A.1 for a review of the geometric series.

17.4 Circuit Models for Op-Amps

17.4.1 Practical Op-Amps

Let's now delve into the details of building feedback systems based on op-amps. We need a way to model real op-amps, and we can categorize the op-amp's imperfections into two categories. First, let's consider "**linear**" **imperfections**. In other words, things that we can include into a linear model. These include:

- Finite open-loop gain ($A_0 < \infty$).
- Finite input resistance ($R_i < \infty \Omega$).
- Non-zero output resistance ($R_o > 0 \Omega$).
- Finite bandwidth and the gain-bandwidth trade-off (to be discussed).

There are other "**non-linear**" **imperfections** that we will discuss in the next chapter. These include:

- Slew rate limitations (explained in the next chapter).
- Finite swing.
- Offset voltage.
- Input bias and offset currents.
- Noise and distortion.

Even though noise can be treated as a linear phenomena, for convenience we will discuss it in the next chapter. So our goal is to build a model that contains as many of these imperfections as possible, while keeping the model simple and easy to use as possible.

17.4.2 World's Simplest Op-Amp Model

The world's simplest op-amp model is shown in *Fig. 17.8*. This is the "VCVS" model, or the **voltage-controlled voltage source** model. While it captures finite gain and output resistance, it fails to capture frequency dependence.

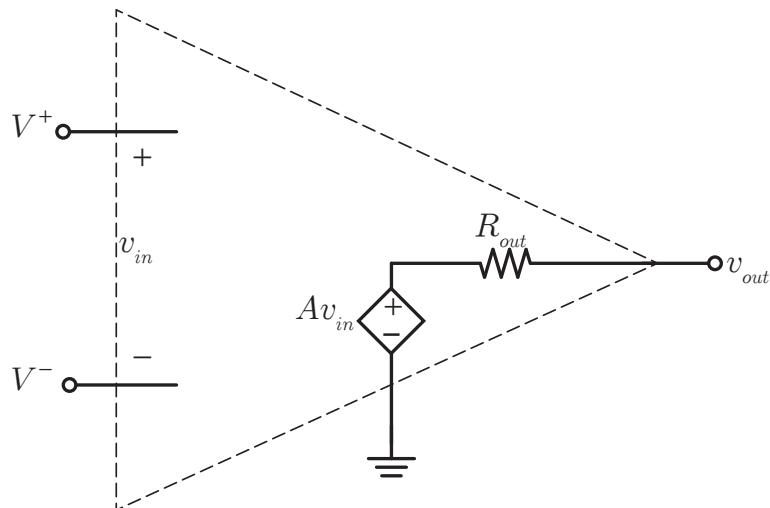


Figure 17.8: A simple model for an op-amp.

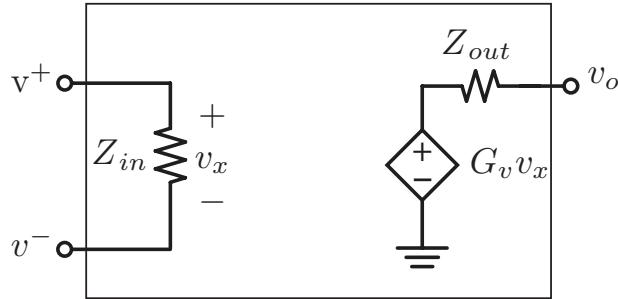


Figure 17.9: A general model for any voltage amplifier, including an op-amp, includes frequency dependent elements that make the model unsuitable for analysis.

17.4.3 General Model of Amplifier

It is easy to build a model that can take into account input impedance $Z_{in}(j\omega)$, output impedance $Z_{out}(j\omega)$, and frequency dependent finite gain $G_v(j\omega)$, using such a general model shown in Fig. 17.9. However, it is too complicated because of all the parameters that vary with frequency. This model also fails to provide any insights, and is too general for our purposes. Let's build a "single-pole" (dominant pole) model by taking the frequency dependence away from G_v .

17.4.4 "Physical" Op-Amp Model

The model shown in Fig. 17.10, uses a G_m cell, or an **ideal transconductor**. We learned that a differential pair can be modeled as a G_m cell, and so as you might expect, the input stage of almost every op-amp is a differential pair. The DC gain and dominant op-amp pole is modeled by R_x and C_x . The larger R_x , the larger the DC gain $A_0 = G_m R_x$. The output resistance can be added easily at the output of the ideal voltage buffer (unity gain amp).

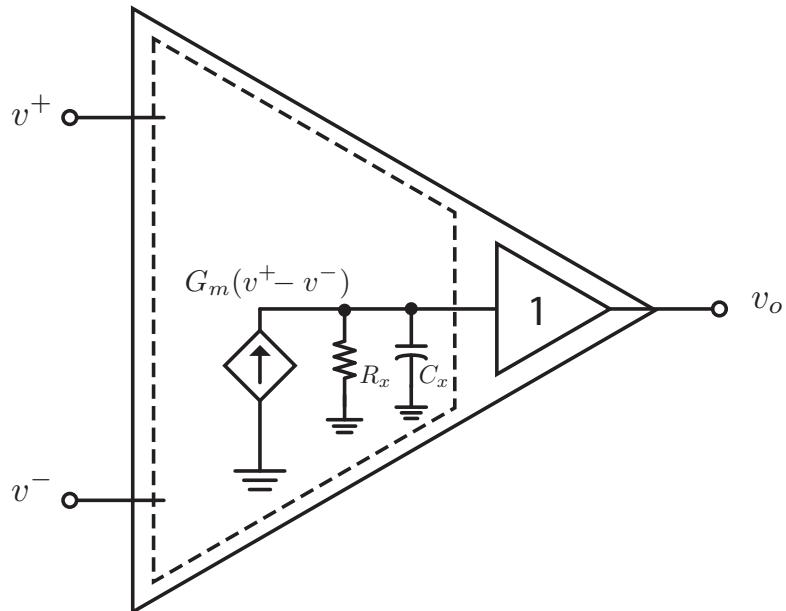


Figure 17.10: An op-amp model with a single pole is inspired from the physical structure of an actual transistor based op-amp.

17.4.5 Operational Transconductance Amplifier (OTA)

Notice that if you "slice off" the triangular head of an op-amp, the result is the transconductance stage (see dashed lines in *Fig. 17.10*), and has a special symbol shown in *Fig. 17.11a*. It is known as an **operational transconductance amplifier**, also known as an "OTA". An OTA is essentially a differential transistor pair with a high output impedance node, for example using a mirror as shown in *Fig. 17.11b*.

Since an OTA is essentially a G_m amplifier, it converts a differential voltage into an output current. So if we want to drive a load resistor, we need an output stage (buffer). As such, many real op-amps are internally constructed from an OTA + buffer as shown.

17.4.6 Op-Amp Capacitance

In *Fig. 17.12*, we've added input capacitance and output capacitance to the basic model. Since R_x is absent, the DC gain is infinite. This is also known as an **integrator**, because the voltage at the output is an integral of the differential voltage.

What is the origin of C_{in} and C_{out} ? Any amplifier has input/output capacitance due to transistors and packaging parasitics. The capacitance also arises from cables or long board traces. In some cases the sensor or actuator of interest (the source or load) has a capacitive component. Note that without R_{out} , the output capacitance has no impact on the transfer function. In other words, an ideal buffer can drive any load. For the focus of the rest of the chapter, we'll zoom in on the internal capacitance C_x .

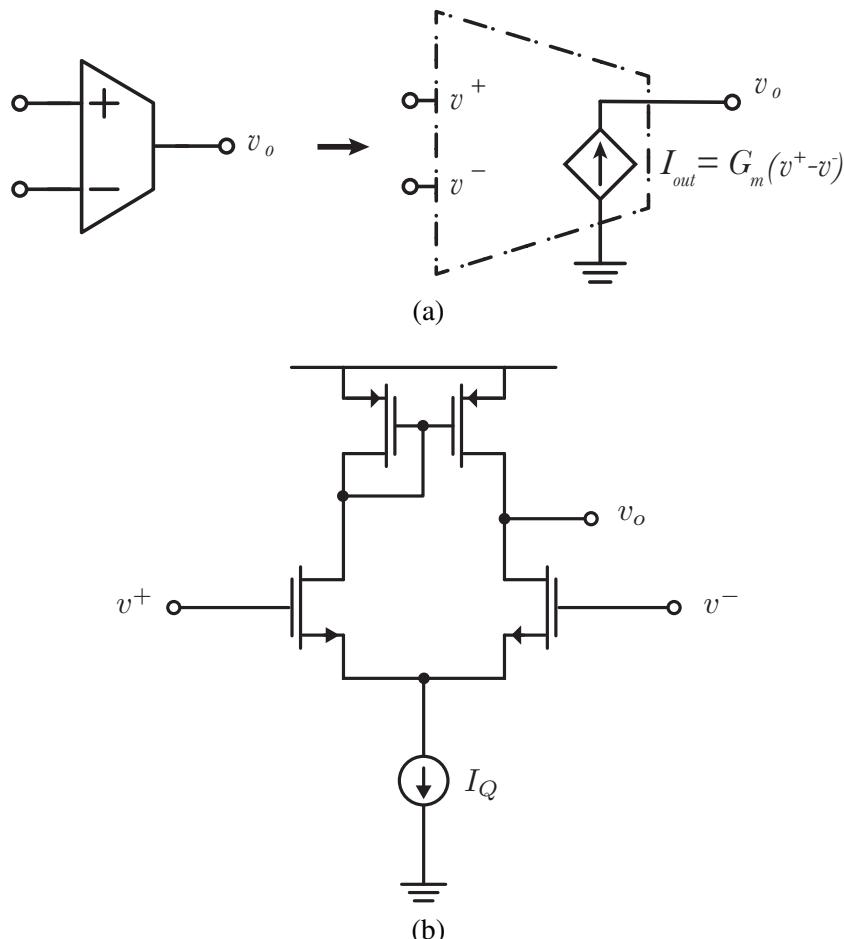
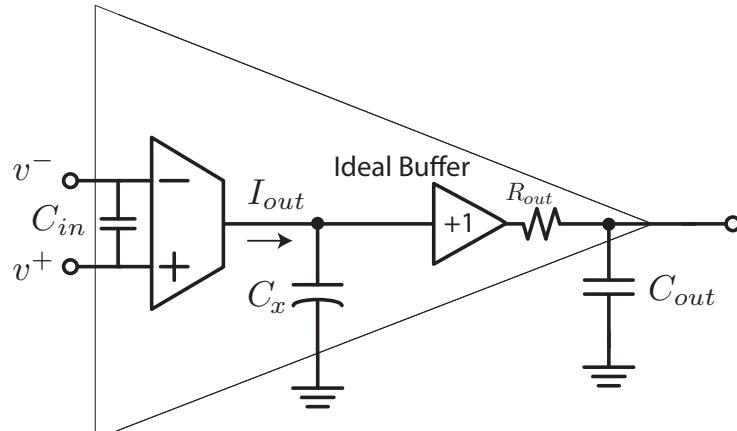


Figure 17.11: (a) An Operational Transconductance Amplifier (OTA) forms the input stage of an op-amp. (b) It is realized using a differential pair with mirror load, or a similar circuit.

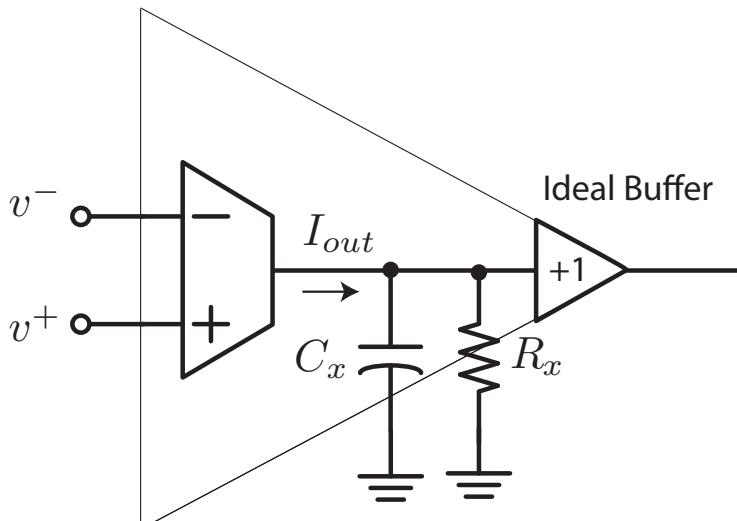


$$I_{out} = G_m (v^+ - v^-)$$

Figure 17.12: Op-amp model with infinite gain and output impedance. The role of the capacitance C_x is to model the frequency response of the amplifier.

17.4.7 Transconductance Amplifier Model

The OTA model shown in *Fig. 17.13* closely resembles the insides of an op-amp. The only difference between this model and *Fig. 17.10* is notation. We've used the OTA symbol for the input G_m . To get high gain, the input G_m drives a high impedance Z node (formed by C_x and R_x). The output buffer is provided to drive a low impedance load and to preserve the high voltage gain. This model includes the first pole of the amplifier.



$$I_{out} = G_m (v^+ - v^-)$$

Figure 17.13: Op-amp model with finite DC gain and a single pole. This model is equivalent to *Fig. 17.10*.

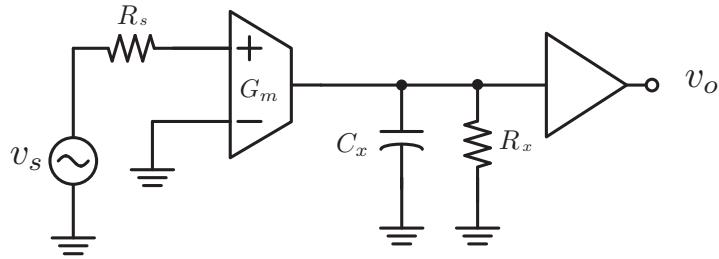


Figure 17.14: The open loop gain and bandwidth of the op-amp model is calculated using the OTA op-amp model by driving the positive input with a source and observing the output voltage.

17.5 Gain/Bandwidth Trade-off

17.5.1 Op-Amp Gain / Bandwidth

Consider the **open-loop amplifier** shown in *Fig. 17.14*. Using the concept of impedance, it is easy to derive the transfer function and verify that it is a first-order single pole system. At DC the capacitor becomes an open circuit, and the gain is simply:

$$G_0 = G_m R_x \quad (17.15)$$

The dominant frequency response of the op-amp is due to the time constant formed at the high-Z node:

$$\omega_{-3dB} = \frac{1}{R_x C_x} \quad (17.16)$$

An interesting observation is that the **gain-bandwidth product** depends on only on G_m and C_x (and not the gain G_0):

$$G_0 \times \omega_{-3dB} = \frac{G_m}{C_x}$$

Gain-bandwidth product (17.17)

17.5.2 Driving Capacitive Loads

In many situations, the load is a capacitor rather than a resistor, as shown in *Fig. 17.15*. For such cases, we can directly use an OTA (rather than a full op-amp), and the bandwidth is determined by the load capacitance:

$$\omega_{-3dB} = \frac{1}{R_x} (C_x + C_L) \quad (17.18)$$

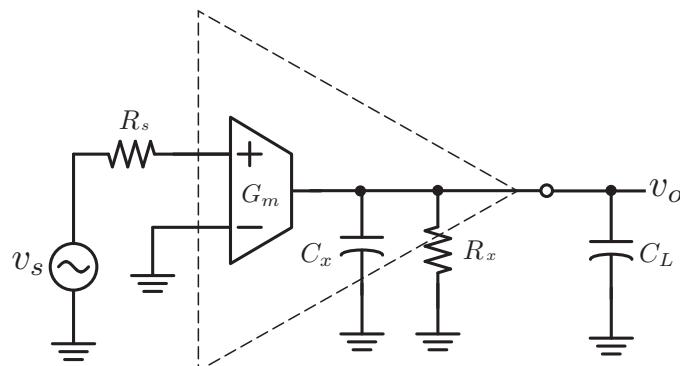


Figure 17.15: When an op-amp drives a capacitive load, the internal buffer can be eliminated. In this case, the bandwidth is determined by a combination of the internal capacitance and the load capacitance C_L .

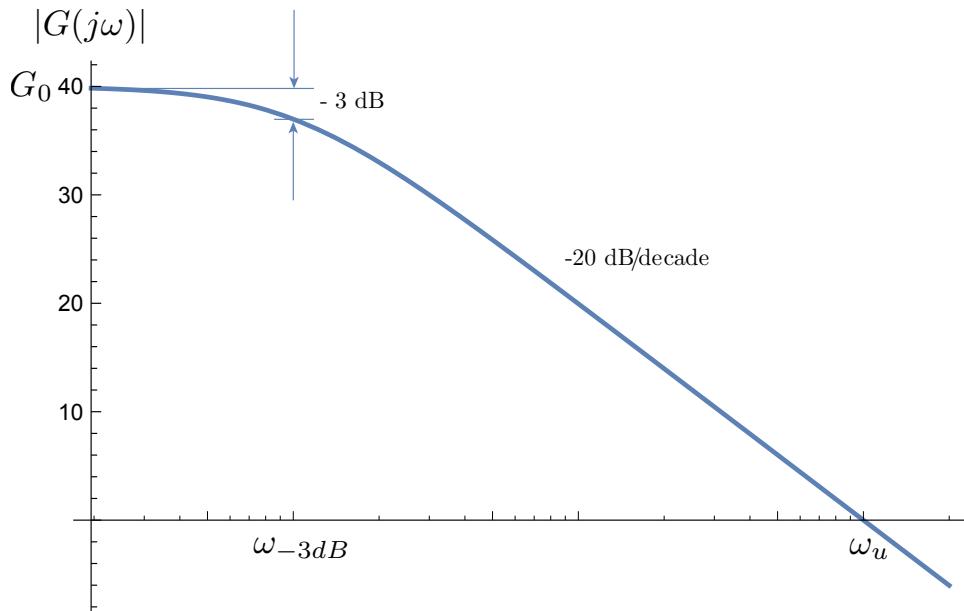


Figure 17.16: The magnitude response of a single pole op-amp. Note the gain is unity at the frequency ω_u .

17.5.3 Open-Loop Frequency Response

A plot of the single-pole frequency response is shown in *Fig. 17.16*. The DC gain G_0 and the 3dB frequency $\omega_{-3\text{dB}}$ completely define the transfer function. We will also highlight another important quantity known as the **unity gain frequency**, ω_u . This is the frequency at which the transfer function has unit magnitude:

$$|G(j\omega_u)| = \left| \frac{G_0}{1 + \frac{j\omega_u}{\omega_{-3\text{dB}}}} \right| = 1 \quad (17.19)$$

To find it, note that the transfer function beyond $\omega_{-3\text{dB}}$, or for $\omega \gg \omega_{-3\text{dB}}$, the closed-loop gain can be simplified:

$$|G(j\omega_u)| \approx G_0 \left| \frac{1}{j\omega_u/\omega_{-3\text{dB}}} \right| = G_0 \frac{\omega_{-3\text{dB}}}{\omega_u} = 1 \quad (17.20)$$

In other words, the unity gain frequency is given by:

$$\boxed{\omega_u \approx G_0 \omega_{-3\text{dB}}} \quad \text{Unity-gain frequency} \quad (17.21)$$

This also shows that the gain at high frequencies is given by:

$$\boxed{G(j\omega) \approx \frac{\omega_u}{j\omega}} \quad \text{High-frequency gain} \quad (17.22)$$

The unity gain frequency or bandwidth is also known as the gain-bandwidth product for obvious reasons. The unity gain frequency plays a prominent role in the frequency response of amplifiers.

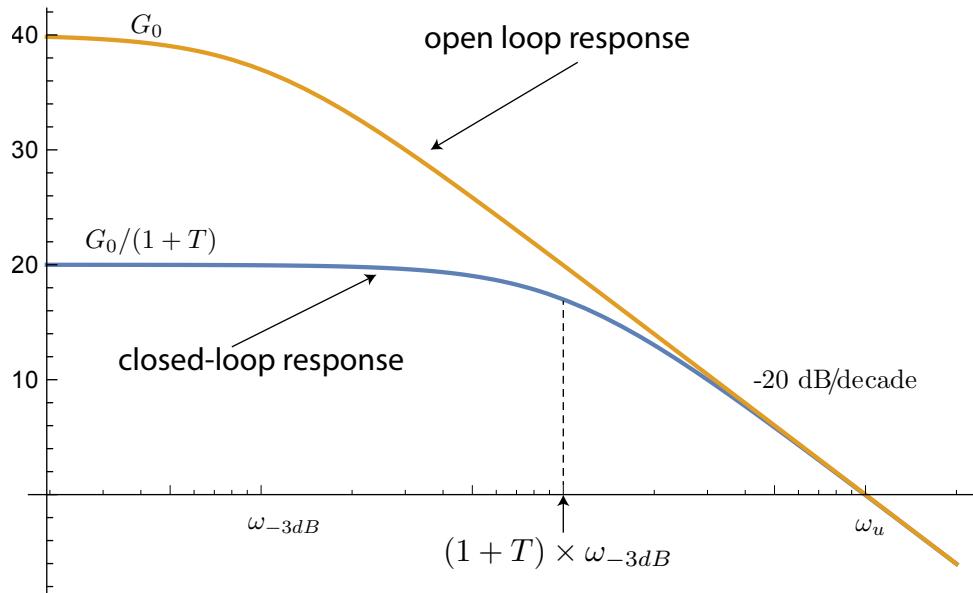


Figure 17.17: The magnitude response of an op-amp amplifier in closed-loop configuration compared to the original open-loop response.

17.5.4 Bandwidth Extension

To see why the unity gain is important, consider again a core amplifier with a single pole. The transfer function is given by:

$$G(j\omega) = \frac{G_0}{1 + j\omega\tau} \quad (17.23)$$

For convenience, we are using $\tau = 1/\omega_{-3dB}$, also known as the *time constant* of the system. We know that the step response of the amplifier is dominated by τ , which is inverse of the bandwidth. When we use this amplifier in feedback, the overall transfer function is given by:

$$G_{CL}(j\omega) = \frac{G(j\omega)}{1 + G(j\omega)f} = \frac{\frac{G_0}{1 + j\omega\tau}}{1 + \frac{G_0f}{1 + j\omega\tau}} \quad (17.24)$$

If we clear the common fraction $(1 + j\omega\tau)$:

$$G_{CL}(j\omega) = \frac{G_0}{(1 + j\omega\tau) + G_0f} \quad (17.25)$$

Finally, we put the equation in standard form to read off the bandwidth:

$$G_{CL}(j\omega) = \frac{\frac{G_0}{1+T}}{1 + j\omega\frac{\tau}{1+T}} \quad (17.26)$$

The amplifier step response is faster by the factor $(1 + T)$, or approximately by the loop gain ($T = G_0f \gg 1$). We can equivalently say that the bandwidth of the closed loop amplifier, ω_0 , has expanded by $(1 + T)$:

$$\boxed{\omega_0 = \frac{(1+T)}{\tau} = (1+T)\omega_{-3dB}}$$

Closed-loop bandwidth
(17.27)

17.5.5 Gain / Bandwidth Product in Feedback

Both the open-loop and closed-loop transfer function are shown in *Fig. 17.17*. Even though the bandwidth expanded by $(1 + T)$, the gain drops by the same factor. So overall the gain-bandwidth (*GBW*) product is constant. The *GBW* product depends only the the G_m of the op-amp, and the C_x internal capacitance (or load in the case of an OTA):

$$\omega_u = GBW = G_0 \cdot \omega_{-3dB} = G_m R_x \cdot \frac{1}{C_x R_x} = \frac{G_m}{C_x} \quad (17.28)$$

In other words, even for an op-amp with infinite gain, shown in *Fig. 17.13*, the gain-bandwidth product is the same. Since the frequency response of the closed loop system is determined by ω_u , it is common to specify the *GBW* of an op-amp rather than its open-loop bandwidth. To determine the closed-loop bandwidth, all we have to take the product of the feedback factor f and ω_u :

$$\omega_0 = f \omega_u \quad (17.29)$$

17.5.6 Unity Gain Feedback Amplifier

Consider a voltage follower feedback amplifier, shown in *Fig. 17.18*. The amplifier has a feedback factor $f = 1$, and so it has the full *GBW* product frequency range available, making it a very broadband amplifier.

17.5.7 How Feedback Broadbands an Amplifier

So far we have shown mathematically that feedback extends the bandwidth by the loop gain. In the next section, we would like to provide some circuit insight into the behavior.

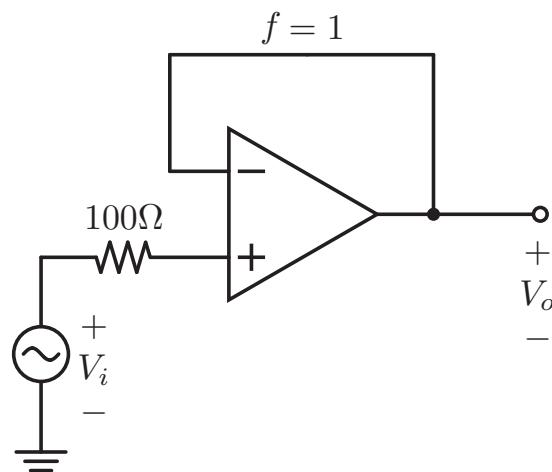


Figure 17.18: An op-amp configured in unity-gain configuration. Note the feedback factor is unity rendering the closed-loop bandwidth $\omega_0 = \omega_u = GBW$.

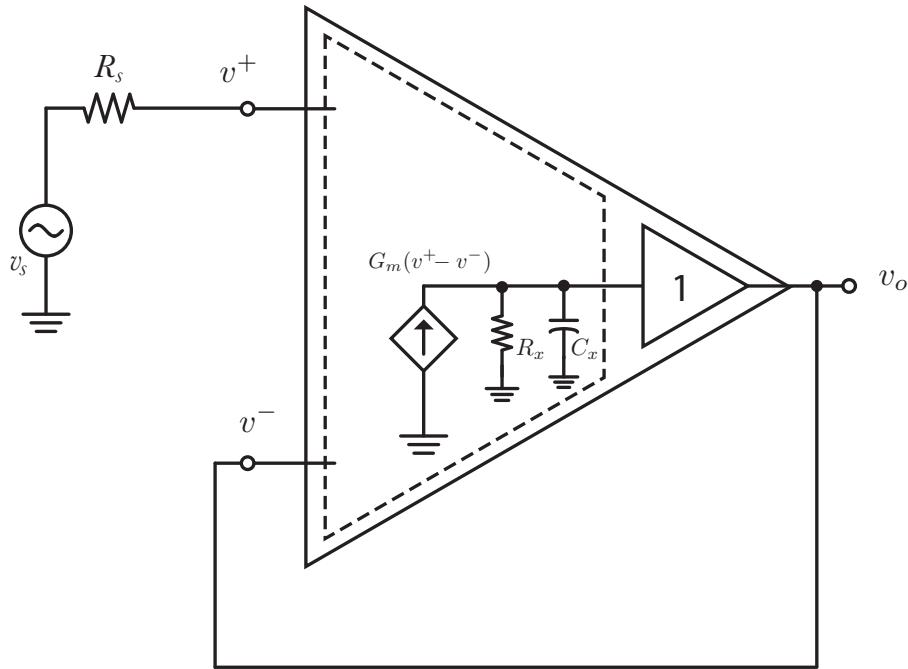


Figure 17.19: Model of an op-amp in unity-gain configuration.

17.5.8 Back to Circuit Model

Consider the equivalent circuit for an amplifier with unity gain feedback, shown in *Fig. 17.19*. What's the bandwidth? If you said $1/R_x C_x$, think again! The dependent current source plays a crucial role. As far as C_x is concerned, the only resistance is not R_x . The reason for this stems from the dependent source being driven by a non-zero control signal, and so we must consider its output resistance.

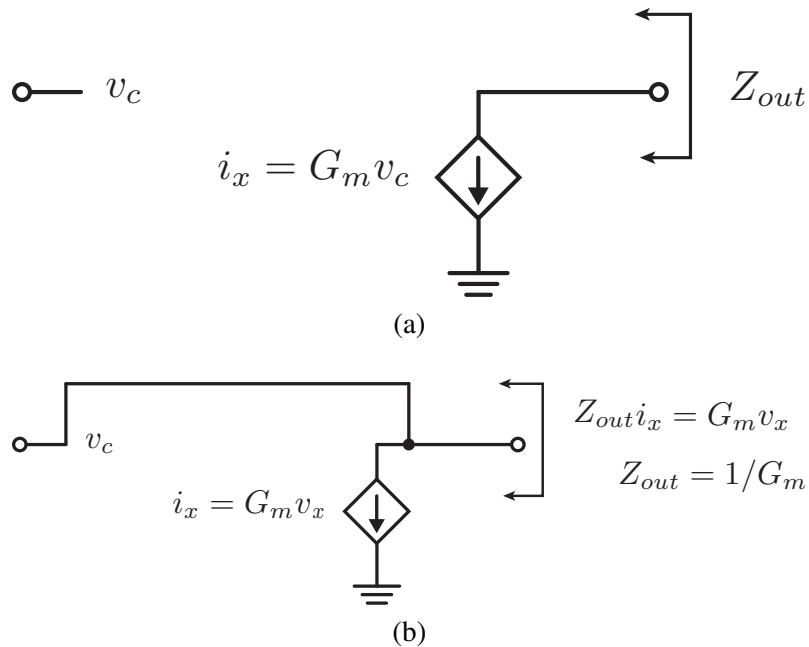


Figure 17.20: (a) The impedance at the output of a transconductor depends on the control voltage. If the control voltage is zero, the output impedance is infinity. (b) If the control voltage is shorted to the output node, then the transconductor is an ordinary resistor with resistance $1/G_m$.

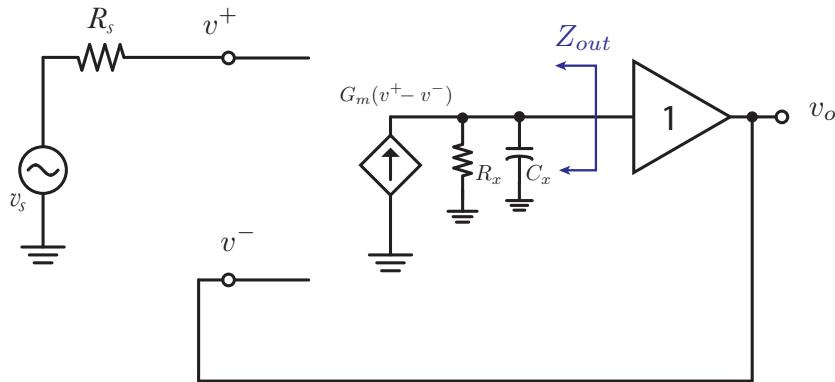


Figure 17.21: The unity-gain feedback in an op-amp causes the high-Z node impedance to drop to $1/G_m$, reducing the gain to unity and expanding the bandwidth.

17.5.9 Turning a Current Source into a Resistor

This problem should be somewhat familiar. Recall that a *MOS diode* has an output impedance of $1/g_m$, not r_o , because of the drain-gate connection. As shown in Fig. 17.20, when a *voltage-controlled current source* has an output impedance that depends on the controlling voltage, it is just a resistor of value $1/G_m$. Likewise, if we calculate Z_{out} at the internal node of the op-amp, as shown in Fig. 17.21, the impedance is much lower than R_x . Here we see the action of the feedback is to lower the impedance seen by the capacitor (C_x) by the loop gain. This has a result of expanding the bandwidth by the same factor.

$$Z_{out} = R_x \parallel \frac{1}{G_m} \approx \frac{1}{G_m} \quad (17.30)$$

Due to the unity gain feedback, the gain to this node is given by:

$$G \approx G_m \cdot \frac{1}{G_m} = 1 \quad (17.31)$$

The bandwidth is thus:

$$\omega_0 \approx \frac{G_m}{C_x} = \omega_u \quad (17.32)$$

Eq. 17.32 matches our derivation earlier. Now we can see that the action of the feedback is to lower the impedance of the high-Z node, which lowers the gain and expands the bandwidth. It is not hard to show that if the feedback factor is f , then the input impedance at the node is approximately given by:

$$Z_{out} = R_x \parallel \left(\frac{1}{f G_m} \right) \approx \frac{1}{f G_m} \quad (17.33)$$

Resulting in a closed-loop gain of:

$$G \approx G_m \cdot \frac{1}{f G_m} = \frac{1}{f} \quad (17.34)$$

and bandwidth:

$$\omega_0 \approx \frac{f G_m}{C_x} = f \omega_u = f G_0 \omega_{-3dB} = f \omega_u \quad (17.35)$$

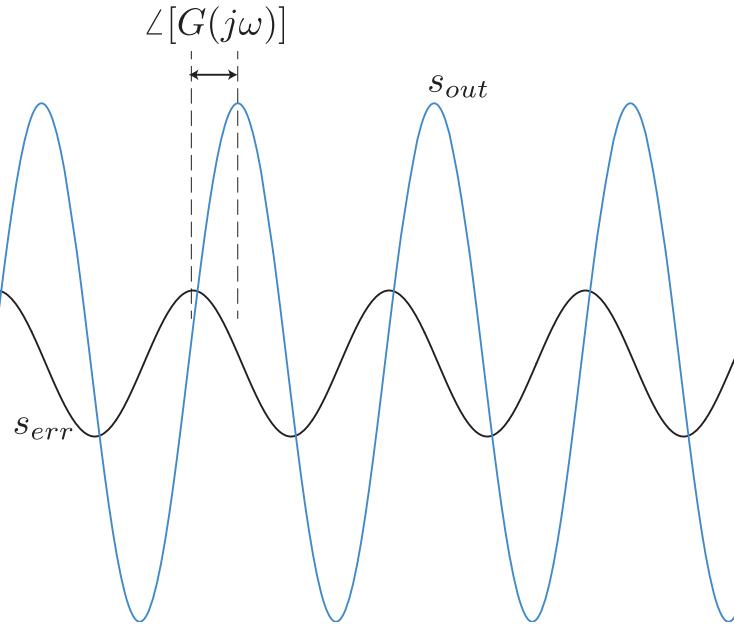


Figure 17.22: When driven by a sinusoidal voltage, the output voltage is phase delayed with respect to the input by an amount of $\angle G(j\omega)$.

17.6 Feedback and Stability

In this section we will only touch upon the important topic of feedback and **stability**. In a more advanced course you will analyze feedback and techniques to stabilize an amplifier in detail.

17.6.1 Stability

As shown in *Fig. 17.22*, any real amplifier will introduce some phase shift when the input frequency increases. For example, a single-pole system has the following phase response:

$$\angle G(j\omega) = \angle \left(\frac{G_0}{1 + j\omega\tau} \right) = \angle G_0 - \angle(1 + j\omega\tau) \quad (17.36)$$

Eq. 17.36 can be simplified to:

$$\angle G(j\omega) = -\tan^{-1}(\omega\tau) \quad (17.37)$$

As the frequency increases, the phase of the output signal lags the input, asymptotically up to 90° .

17.6.2 Non-Dominant Poles

As we have seen, poles in the system tend to make an amplifier less stable. A single pole cannot do harm, because it has a maximum phase shift of 90° . However, a second pole will bring our system to the edge of stability, as shown in *Figs. 17.23 and 17.24*. For this reason, **non-dominant poles** should be at a much higher frequency than the unity-gain frequency, so that when the phase shift reaches 180° , the loop gain is less than unity.

17.6.3 Oscillation

The condition $T(j\omega_x) = -1$ is in fact how we build **oscillators**, which are inherently unstable. If the circuit has $T = -1$ at a particular frequency, then the gain at that frequency is theoretically infinite. Any noise or disturbance can lead to a strong oscillation at this particular frequency.

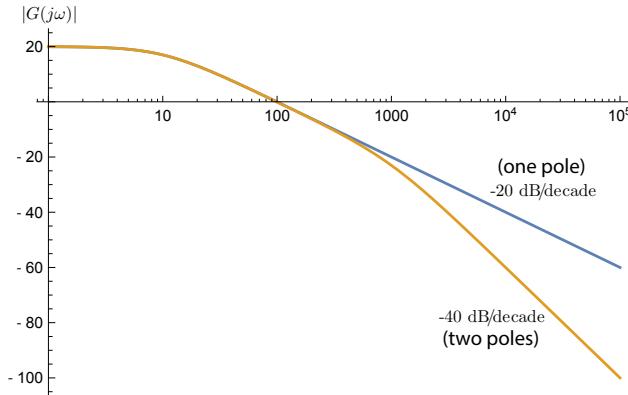


Figure 17.23: The magnitude response of a system with two poles drops faster, at -40 dB/dec.

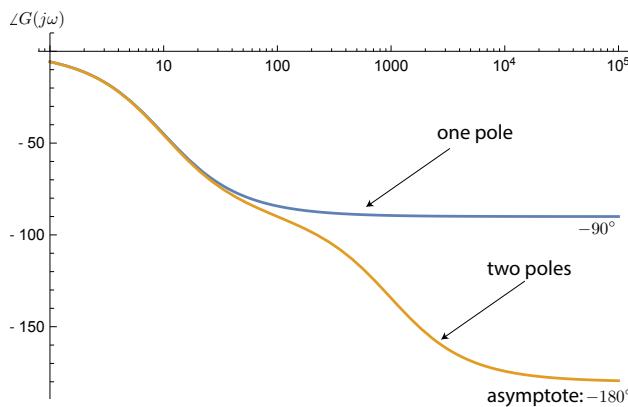


Figure 17.24: The phase response of a second-order transfer function has an asymptotic phase shift of 180°.

17.6.4 Instability

When the system has more poles, the phase shift can reach 180°, as shown in Fig. 17.25. This transforms the negative feedback system into a positive feedback system, which is unstable if the loop gain $|T| > 1$.

Consider the frequency ω_x that satisfies $T(j\omega_x) = G(j\omega_x)f = -1$. This means that there is always noise and disturbances in the system at this frequency. The noise is continuously regenerated, and can potentially cause problems, because the closed loop gain becomes infinite. The *condition for stability* is then to ensure that when loop gain is unity, the phase of $\angle T$ should be less than 180°. The **phase margin** is a measure of stability, and in practice a good design should have 60° phase margin or more.

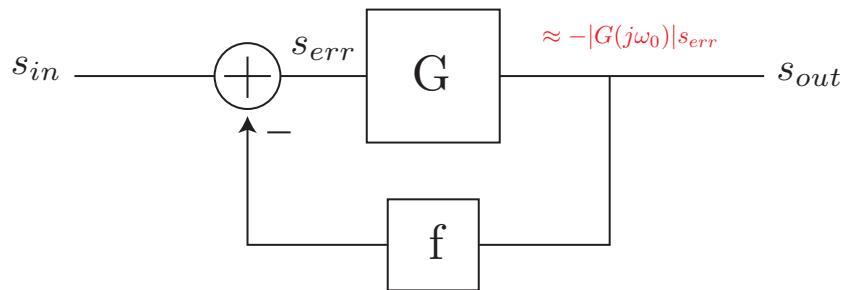
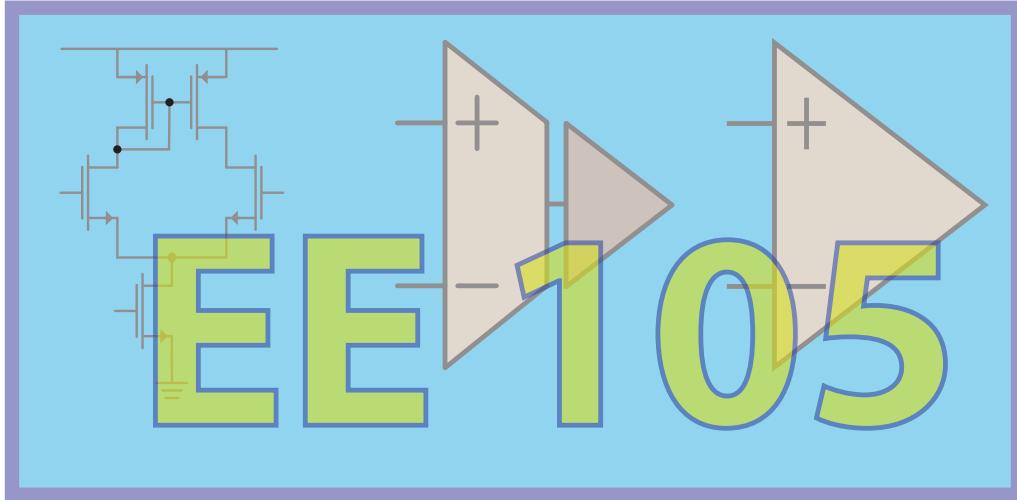


Figure 17.25: A "negative" feedback system turns into a positive feedback system at a frequency that causes phase inversion through G .



18. Non-Ideal Operational Amplifiers

18.1 Chapter Preview

In the last chapter we introduced a model for an operational amplifier (op-amp) that could predict the gain-bandwidth (GBW) trade-off of the amplifier in single-pole dominated op-amps. The model was physically inspired, and included an input OTA followed by an optional buffer. The model accounted for finite open-loop gain ($A_0 < \infty$), finite input resistance ($R_i < \infty\Omega$), non-zero output resistance ($R_o > 0\Omega$), and finite bandwidth.

In this chapter we continue our investigation of op-amps, but we will also discuss other non-ideal effects and non-linearities that are present in real op-amps. These include an offset voltage, DC bias currents at the input, swing and slew-rate limitations, and distortion. We end the chapter by discussing some other effects such as noise, but we will only scratch the surface of this topic and leave this for a more advanced analog circuits course.

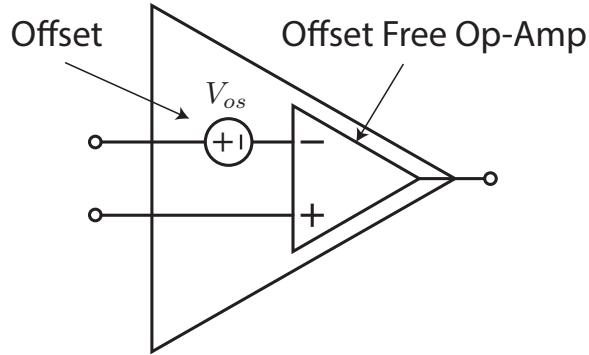


Figure 18.1: The offset voltage of an op-amp can be modeled by assuming the op-amp is free of offset, and then placing a voltage source at one of the input terminals. Since the sign of the offset voltage is a random variable, the offset voltage can be placed at either input terminal.

18.2 Offset Voltage and Currents

18.2.1 Offset Voltage

If you short the inputs of an op-amp, you would expect that the output would be at zero voltages. Unfortunately, due to imbalances in the construction of the amplifier, the output may not be at zero at all. In fact, in many cases it might even "rail out", and sit at either V_{sup} or $-V_{sup}$, where $\pm V_{sup}$ are the supply rails. From our study of the *differential pair*, we know that the origin of this **offset voltage** arises from mismatches between the input differential pair transistors or load resistors.

We model an op-amp with an offset voltage as an ideal op-amp inside another amplifier with a simple DC offset voltage, shown in *Fig. 18.1*. Even though the DC offset voltage is shown with a certain polarity, in reality it can be positive or negative, because it is a random variable. This means that for each model of an op-amp there will be a different offset voltage for each part. In datasheets for op-amps you will see that the stated offset voltage V_{os} is usually the standard deviation of the offset voltage. Therefore it is clear that we can place the DC voltage on either terminal. The impact is to shift the transfer curve left or right (shown in *Fig. 18.2*), depending on the sign and magnitude of the offset.

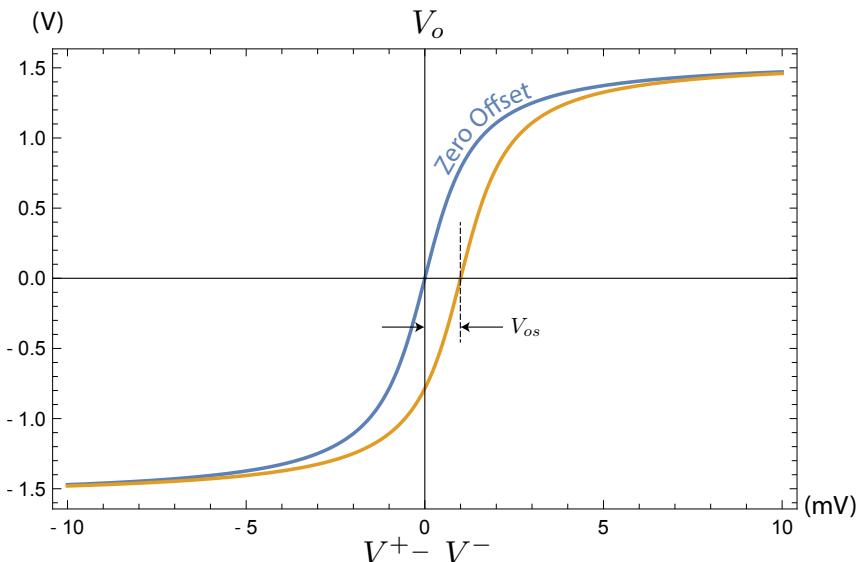
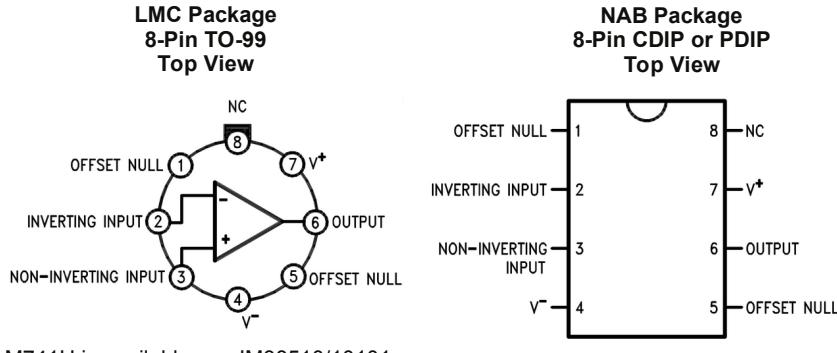


Figure 18.2: The offset voltage causes the input-output transfer curve of an op-amp to deviate from the origin. In other words, even a zero input results in a non-zero output voltage.



LM741H is available per JM38510/10101

Figure 18.3: The pinout of the LM741H op-amp has several pins dedicated for offset nulling the amplifier.

18.2.2 Trimming of Offset Voltage

The pinout of a typical op-amp is shown in *Fig. 18.3*. In addition to the obvious pins (input, output, supplies), there are extra pins labeled "offset null". The output DC offset voltage of an op-amp can be trimmed to zero by connecting a potentiometer to the two offset-nulling terminals, as shown in *Fig. 18.4*. In this example, the wiper of the potentiometer should be connected to the negative supply of the op-amp. In fully integrated circuits, the offset null circuitry is internal (for example with electronically steerable resistors or current sources) and the offset is nulled out using a calibration step or feedback.

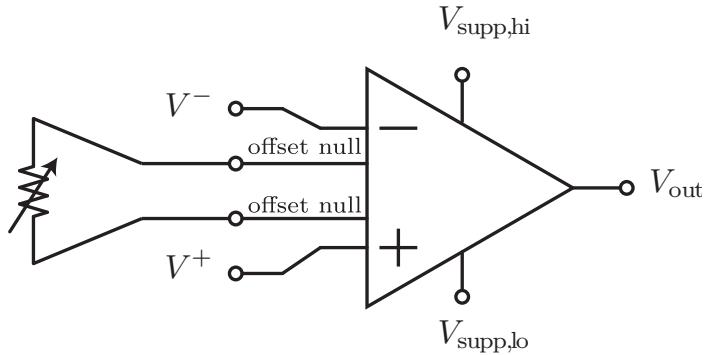


Figure 18.4: The pinout of many op-amps include offset null pins that should be connected as directed in the datasheet. For example, in this amplifier a potentiometer is used to tune out the offset.

18.2.3 Input Bias Currents and Offset Currents

Some op-amps, in particular ones implemented with bipolar transistors, have an input bias current that needs to flow into or out of the op-amp inputs in order for everything to function properly. This is typically a very small current (100nA), and modeled by adding explicit current sources to the input of an ideal op-amp, shown in *Fig. 18.5*. The bias currents flowing into each input terminal can also vary, and the difference is the **input offset current**, usually an order of magnitude smaller than the input currents (10nA). The average current is given by:

$$I_B = \frac{I_{B1} + I_{B2}}{2} \quad (18.1)$$

The difference between the bias currents is known as the offset current:

$$I_{OS} = |I_{B1} - I_{B2}| \quad \text{Op-amp input offset current} \quad (18.2)$$

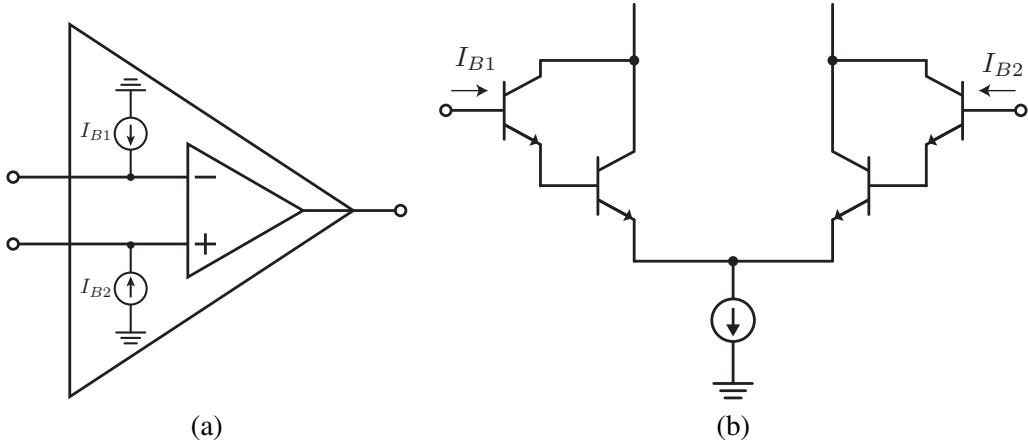


Figure 18.5: (a) Many op-amps require input bias currents to flow into the amplifier. These bias currents are never perfectly matched, resulting in an input offset current. (b) The physical implementation of the input of an op-amp with bipolar transistors may incorporate a "Darlington Pair" to reduce the input current. The pair of transistors acts like a macro transistor with reduced base current. The first (input) BJT's are followers and the differential pair is formed as usual using the Darlingtons.

It is important to understand that this is a DC current. It is *not* an AC current, and it is *not* due to the input impedance. It is usually the base current of a bipolar transistor, $I_B = I_C/\beta$. Variations in β cause the input offset.

18.2.4 Effect of Input Bias Current in Amplifier Circuit

Let's assume for now that the offset voltage of an op-amp is zero. In absence of an input voltage, as shown in Fig. 18.6, one would expect zero output. However, due to non-zero I_B , the output is given by:

$$V_o = I_{B1} R_2 \approx I_B R_2 \quad (18.3)$$

This places a limit on the value of R_2 . A lower value of R_2 requires a higher power op-amp, which is undesirable in many applications. In the next section, we will show a simple way to mitigate this.

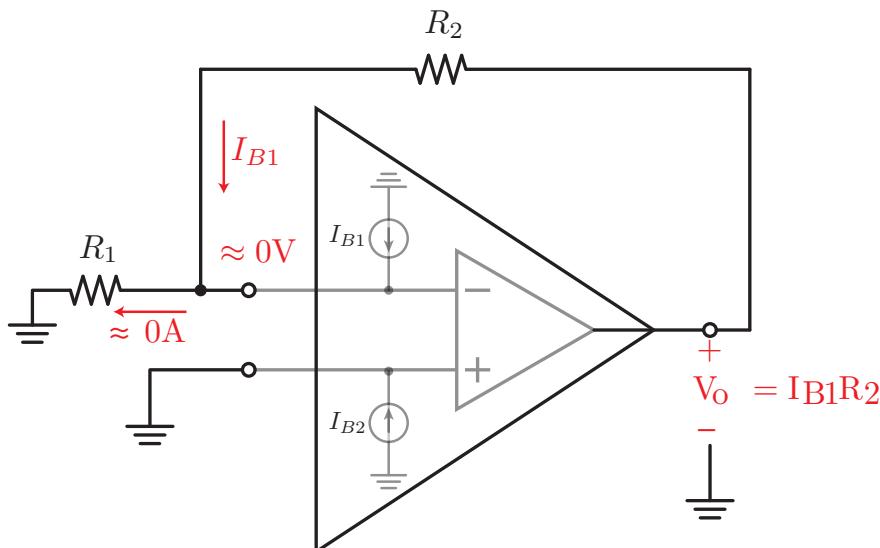


Figure 18.6: In an op-amp with feedback, the input bias currents cause output offset voltages as shown. Note the current through R_1 is zero since the input of the op-amp is at a virtual ground due to the high loop gain.

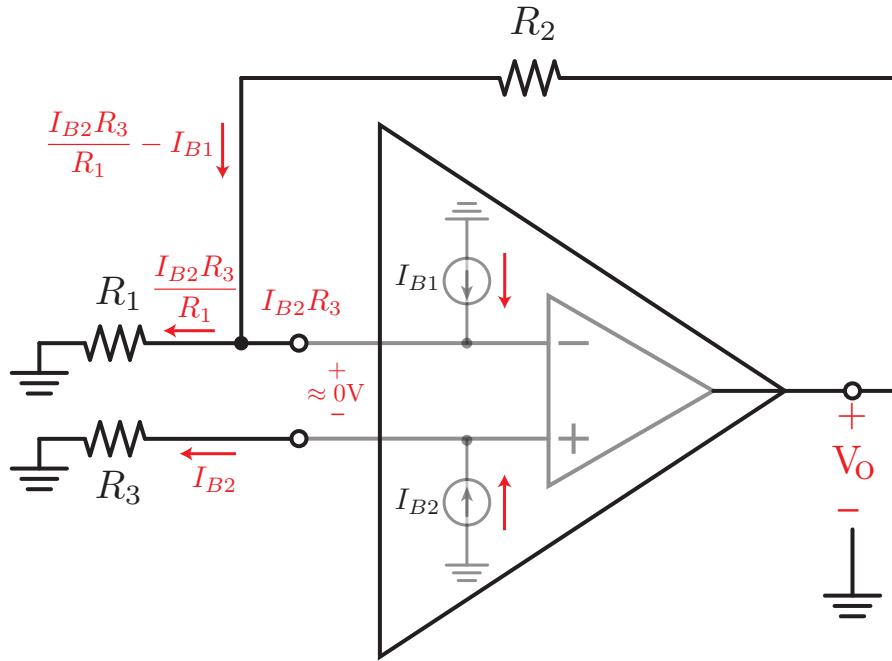


Figure 18.7: The effect of input currents and input offset can be minimized by placing an appropriately sized resistor \$R_3\$ at the positive terminal of the op-amp without impacting the AC performance.

18.2.5 Reducing the Effect of Input Bias Currents

In Fig. 18.7, we add a resistor \$R_3\$, which has no impact on the AC response of an ideal op-amp since the current into the grounded node of the positive terminal is zero. Let's calculate the output offset voltage in this case. The figure labels are very helpful in deriving this equation:

$$V_o = I_{B2}R_3 + R_2 \left(\frac{I_{B2}R_3}{R_1} - I_{B1} \right) \quad (18.4)$$

Let's assume zero offset for now (\$I_{B1} = I_{B2} = I_B\$). Then this equation can be simplified:

$$V_o = I_B R_3 + R_2 I_B \left(\frac{R_3}{R_1} - 1 \right) \quad (18.5)$$

The output offset voltage due to \$I_B\$ is given by:

$V_{OS} = I_B \left[-R_2 + R_3 \left(1 + \frac{R_2}{R_1} \right) \right]$

Op-amp output offset voltage (18.6)

Notice that if we choose \$R_3 = \frac{R_2}{1+R_2/R_1}\$, we can drive \$V_o = 0V\$, thus eliminating the systematic offset of the circuit. You can show that if we now include an offset current, the above choice of \$R_3\$ results in \$V_o = I_{OS}R_2\$, which is an order of magnitude lower than \$I_B R_2\$ from Eq. 18.3.

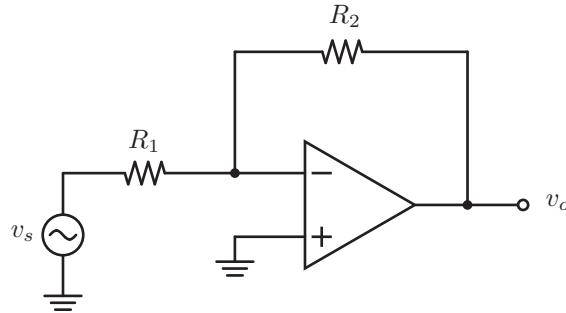


Figure 18.8: An inverting op-amp with finite gain and a single dominant pole.

18.3 Op-Amp Frequency Response

In the last chapter, we discussed the op-amp frequency response in detail. This section is mostly a review. We will analyze an inverting amplifier circuit and show that our results from feedback analysis approximately hold true in this case.

18.3.1 Frequency Response of Open-Loop Op-Amp

We know from the previous chapter that op-amps are designed to have a single dominant pole. The transfer function of the op-amp is therefore approximately given by:

$$A(j\omega) = \frac{A_0}{1 + \frac{j\omega}{\omega_{-3dB}}} \quad (18.7)$$

In Eq. 18.7, the DC gain is A_0 , and the -3 dB frequency is ω_{-3dB} . In the previous chapter we also derived the *unity-gain bandwidth*, $\omega_u = A_0 \omega_{-3dB}$, which is also known as the gain-bandwidth product (GBW).

For high frequencies, such that $\omega \gg \omega_{-3dB}$, we can define the single pole response with a dominant pole at ω_{-3dB} as:

$$A(j\omega) = \frac{\omega_u}{j\omega}$$

Open-loop op-amp frequency response (18.8)

18.3.2 Frequency Response of Closed-Loop Op-Amp

Let's analyze the frequency response of the op-amp circuit shown in Fig. 18.8. This is a closed-loop system, so we expect the feedback analysis to hold. Let's find the transfer function assuming finite gain A :

$$G = \frac{v_o}{v_s} = \left(-\frac{R_2}{R_1} \right) \frac{1}{1 + \left(\frac{R_2}{R_1} \frac{1}{A} \right)} \quad (18.9)$$

As expected, if $A \rightarrow \infty$, we get the transfer function for an inverting amplifier. Now substitute Eq. 18.7, the single-pole transfer function, into A of Eq. 18.9 to obtain:

$$G = \left(-\frac{R_2}{R_1} \right) \frac{1}{1 + \left(\frac{1 + \frac{R_2}{R_1}}{\frac{A_0}{1 + \frac{j\omega}{\omega_{-3dB}}}} \right)} \quad (18.10)$$

Now we simplify Eq. 18.10:

$$G(j\omega) = \left(-\frac{R_2}{R_1}\right) \frac{1}{1 + \left(\frac{1 + \frac{R_2}{R_1}}{\frac{A_0}{1 + \frac{j\omega}{\omega_{-3dB}}}}\right)} = \left(-\frac{R_2}{R_1}\right) \frac{1}{1 + \left(\frac{1 + \frac{R_2}{R_1}}{\frac{A_0 \cdot \omega_{-3dB}}{\omega_{-3dB} + j\omega}}\right)} \quad (18.11)$$

$$= \left(-\frac{R_2}{R_1}\right) \frac{1}{1 + \left[\left(\frac{1}{A_0}\right) \left(1 + \frac{R_2}{R_1}\right) \left(1 + \frac{j\omega}{\omega_{-3dB}}\right)\right]} \quad (18.12)$$

The derived transfer function can be approximated if the DC gain A_0 is very large:

$$G(j\omega) = \left(-\frac{R_2}{R_1}\right) \frac{1}{1 + \left[\left(\frac{1}{A_0}\right) \left(1 + \frac{R_2}{R_1}\right)\right] + \left[0\right] + \left[\left(\frac{1}{A_0}\right) \left(1 + \frac{R_2}{R_1}\right) \left(\frac{j\omega}{\omega_{-3dB}}\right)\right]} \quad (18.13)$$

$$\approx \left(-\frac{R_2}{R_1}\right) \frac{1}{1 + \frac{j\omega}{\left(\frac{A_0 \omega_{-3dB}}{1 + \frac{R_2}{R_1}}\right)}} \quad (18.14)$$

As expected from last chapter, the closed-loop bandwidth is higher than the open-loop bandwidth:

$\omega_0 = \frac{A_0 \omega_{-3dB}}{1 + \frac{R_2}{R_1}}$

Closed-loop bandwidth
(18.15)

The *gain-bandwidth product* remains (nearly) unchanged:

$$G \times BW = \left(\frac{R_2}{R_1}\right) \left(\frac{A_0 \omega_{-3dB}}{1 + \frac{R_2}{R_1}}\right) \approx \left(\frac{R_2}{R_1}\right) \left(\frac{A_0 \omega_{-3dB}}{\frac{R_2}{R_1}}\right) = \boxed{A_0 \omega_{-3dB} = \omega_u} \quad (18.16)$$

The closed-loop transfer function in relation to the open-loop transfer function is shown in Fig. 18.9. You might wonder why the feedback analysis did not match our previous results exactly. To summarize, the inverting amplifier is not a voltage amplifier, but really a current to voltage (trans-resistance) amplifier. To see this, note that the output voltage is not subtracted from the input, since the input is also supplied at the negative op-amp terminal. The output voltage is converted to a current through R_2 and it is subtracted from the input current at the inverting terminal (do a *Norton transformation* of the source with R_1).

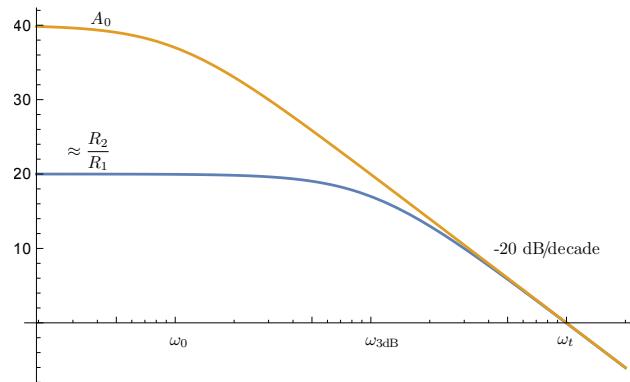


Figure 18.9: Transfer function of op-amp based inverting amplifier in open-loop and closed-loop configuration.

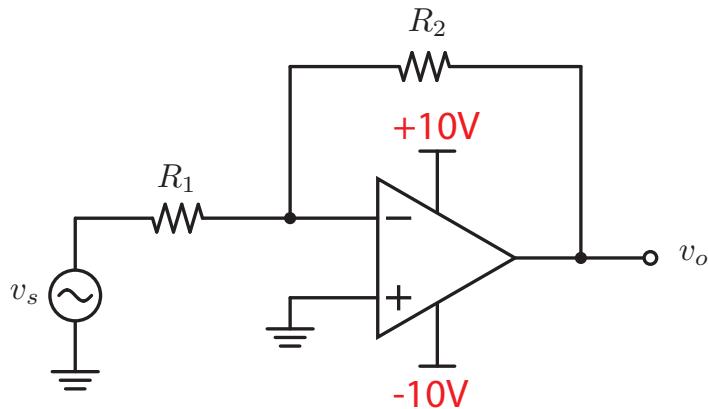


Figure 18.10: The positive and negative supplies of the op-amp are highlighted. If the output voltage exceeds the rails, the output clips. Usually the output cannot swing all the way to the rails, and so the output swing is even lower.

18.4 Voltage Swing and Slewning

18.4.1 Output Saturation

The **output voltage swing** is limited by the voltage supplies, usually omitted from the schematic, but shown explicitly in *Fig. 18.10*. If the *input voltage times the gain of the amplifier is larger than the supplies*, the output will clip, as shown in *Fig. 18.11*. Here we are optimistically assuming that the op-amp can swing all the way to the rails. In reality, the output swing may be lower than the rails due to the various implementation details. Most op-amps also have a maximum output current limit, activated to prevent the op-amp from damage in the case of a short circuit or connection of a very small load resistor. In either case, the output waveform will appear "**clipped**", leading to **distortion**. In most linear amplifier applications, clipping is unwanted and avoided by careful design.

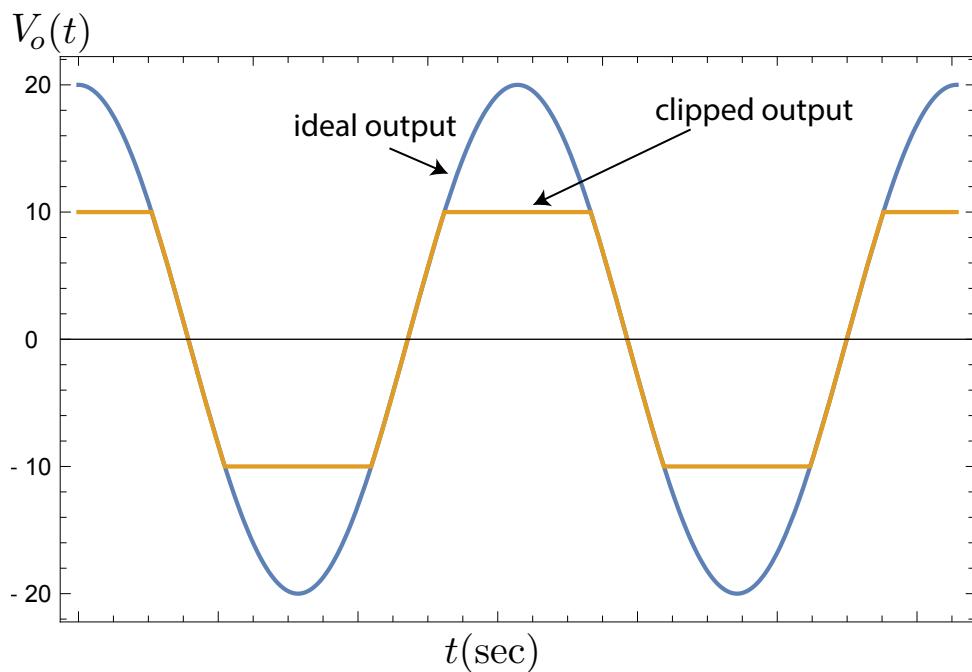


Figure 18.11: A large amplitude sine wave is clipped because the voltage exceeds the op-amp rails (supplies).

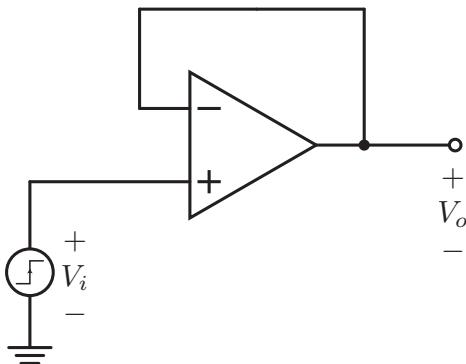
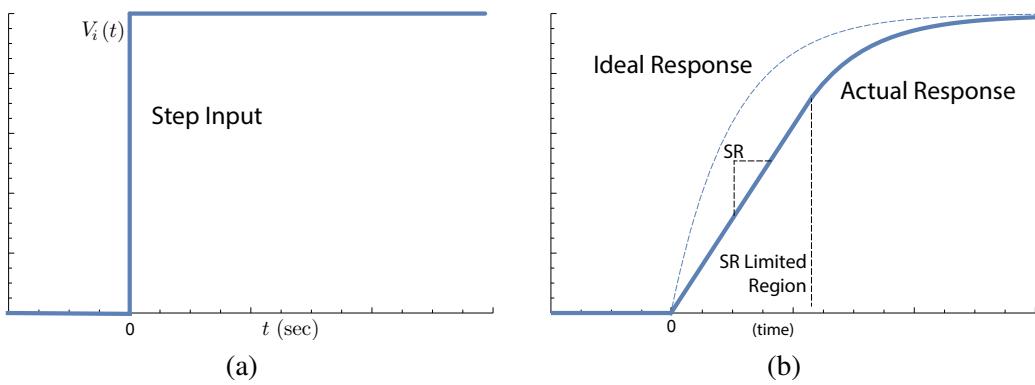


Figure 18.12: An op-amp in unity gain configuration is driven by a step function.

Figure 18.13: (a) The input step function is applied at time $t = 0\text{ s}$. (b) Ideally, the output would approach the input with an exponential response and time constant commensurate inversely with the bandwidth of the amplifier. In practice, if the step is sufficiently large, the amplifier experiences slew rate limiting.

18.4.2 Slew Rate

The "slew rate", or the maximum possible change at the output, is defined by:

$$SR = \left. \frac{dV_o}{dt} \right|_{maximum} \quad \text{Slew rate} \quad (18.17)$$

The slew rate is usually specified in data sheets in units of $\frac{V}{\mu\text{s}}$. It is very important to understand the difference between *SR* limiting and the *BW* limit. The limited bandwidth is a *linear phenomenon*, and it does not distort the shape of a sinusoidal input. On the other hand, the slew rate limitation can cause *non-linear distortion*, resulting in a non-sinusoidal output waveform.

To understand *SR*-limiting, let's consider a simple unity-gain op-amp amplifier shown in Fig. 18.12. If a step voltage is supplied at the input, the output would ideally follow the exponential settling predicted by simple linear theory:

$$V_o(t) = V_i \left(1 - e^{-t\omega_u} \right) \quad (18.18)$$

In this case the time constant is $\tau = \frac{1}{\omega_u}$, because it is a unity gain feedback configuration. So the amplifier bandwidth is the full *GBW* product. This settling behavior is shown in the dashed line of Fig. 18.13. Something surprising is that a real op-amp will not follow this linear step response, but rather a curve that is a linear ramp for a duration. This curve is labeled as the *SR*-limited

region. Only after the slope of the waveform is small enough does the circuit behave similar to the bandwidth-limited case. Essentially, the *SR*-limit prevents the amplifier from changing faster than the slew rate. For the linear response, the *slope of the output signal* ("slew rate") is given by:

$$\frac{dV_o}{dt} = V_i \omega_u e^{-t \omega_u} \quad (18.19)$$

Notice that the slope is initially the largest and given by:

$$\left. \frac{dV_o}{dt} \right|_{max} = V_i \omega_u \quad (18.20)$$

However, we can also see that over time the slope decays. If this slope is larger than the slew rate, $V_i \omega_u > SR$, the amplifier will slew as shown in *Fig. 18.14*. Since this expression involves the input amplitude, it is possible to apply a smaller signal step $V_{sm,i}$ to the input and not experience any slewing, as shown in *Fig. 18.15*:

$$\left. \frac{dV_o}{dt} \right|_{max} = V_{sm,i} \omega_u < SR$$

Condition for slew-limiting (18.21)

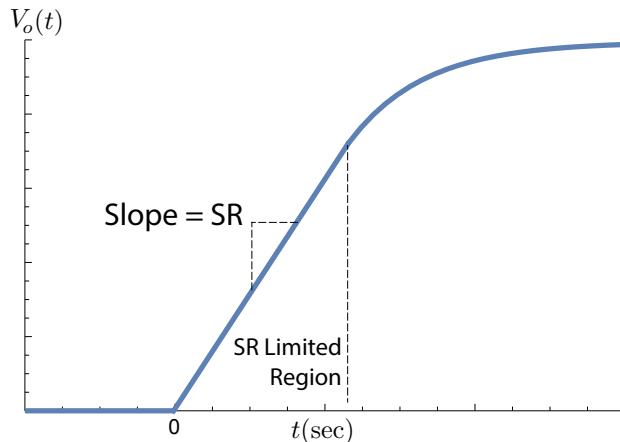


Figure 18.14: When an op-amp is experiencing the slew-rate limiting, the rate of change of the output is constrained to be lower than the *SR* limit. In other words, the slope of the output cannot exceed *SR*.

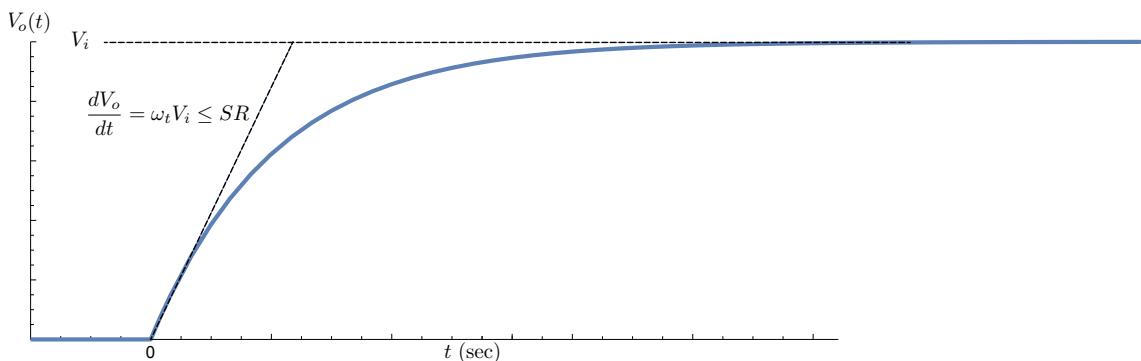


Figure 18.15: In this example, the slope of the output is lower than the *SR* limit, and so the output experiences linear settling time rather than *SR* limiting.

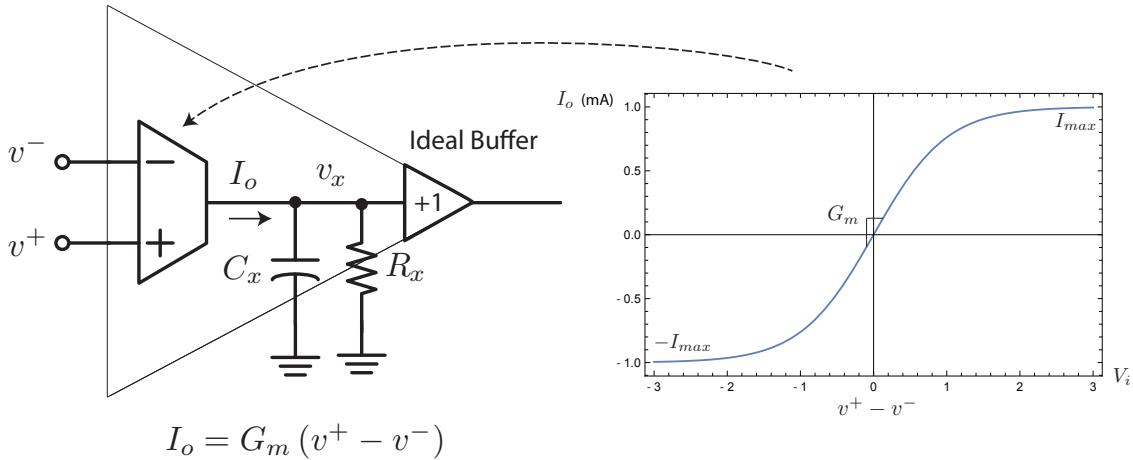


Figure 18.16: The origin of SR-limiting can be understood if we examine the internals of an op-amp. If the input OTA G_m stage current clips for large input voltages, a constant current will flow into the internal node of the op-amp. This results in an SR-limited constant slope waveform.

18.4.3 Origin of Slew Rate Limit

Note that our simple model of an op-amp has a G_m that can in theory supply any current to the internal *high-Z* node capacitor. In practice, the current will eventually clip (as shown in Fig. 18.16), causing SR limitation. Current clipping is similar to voltage clipping. If internal circuits are biased with a constant supply current, they often cannot exceed a certain limit. This is true of a differential pair that we previously studied. If a buffer stage is added to the amplifier, slewing may occur to the finite DC current of the buffer as well. For the model shown, suppose a large input step is applied so that the OTA current is clipped. Then the internal node of the amplifier, initially at zero, will ramp at a constant rate due to the constant current supply flowing into a capacitor:

$$C_x \frac{dV_x}{dt} = I_{G_m} = \pm I_{max} \quad (18.22)$$

In this equation we are ignoring the impact of the large resistor R_x for simplicity. Since I_{max} is a constant, we see that the rate of change of the output, or the SR at the internal node, is fixed:

$$\left. \frac{dV_x}{dt} \right|_{max} = \pm \frac{I_{max}}{C_x} \quad (18.23)$$

Linear settling will resume when the input error voltage reduces (due to feedback), and the G_m enters the linear range.

18.4.4 Full-Power Bandwidth

Now let's suppose we drive our amplifier with a sinusoidal signal, shown in Fig. 18.17. Since the follower is a unity-gain amplifier, we expect the output to be given by:

$$V_o = V_i \sin(\omega t) \quad (18.24)$$

The rate of change of the output must be less than the slew-rate:

$$\frac{dV_o}{dt} = V_i \omega (\cos \omega t) \quad (18.25)$$

Since $|\cos(\omega t)| \leq 1$, the *maximum slope* is given by:

$$\left. \frac{dV_o}{dt} \right|_{max} = V_i \omega \quad (18.26)$$

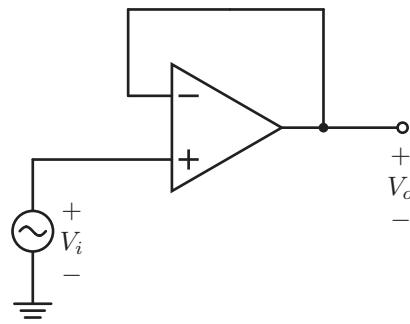


Figure 18.17: When a sinusoidal voltage is applied to a unity-gain amplifier, if the slope of the input waveform is sufficiently large, the output waveform will be distorted. The full-power bandwidth is the largest frequency that the amplifier can process linearly without SR-limiting.

The **full-power bandwidth** is known as the frequency at which the SR-limited distortion starts to occur for an output sinusoid with maximum rated output voltage, $V_{o,max}$:

$$\boxed{\omega_{max} V_{o,max} = SR} \quad Full-power\ bandwidth \quad (18.27)$$

Interestingly, this frequency may be much lower than the unity gain frequency (linear bandwidth limit) of the amplifier:

$$\boxed{f_{max} = \frac{SR}{2\pi V_{o,max}}} \quad Full-power\ frequency \quad (18.28)$$

Note again that the frequency f_{max} is a function of the output amplitude. If the amplitude is lowered, then the SR-limit can be avoided. In other words, we can find the smallest amplitude such that the amplifier always settles linearly by equation f_{max} with the unity-gain frequency.

18.5 Noise and Distortion

18.5.1 Op-Amp Distortion

As we have seen, the op-amp output signal has several sources of distortion arising from clipping, slewing, and even bandwidth limits if the phase of the signal introduces a non-constant group delay in the frequency response. Another source of distortion is that the gain is signal dependent. In other words, the input-output curve is not a perfect straight line but has curvature ("S" shaped for example), as shown in *Fig. 18.18*.

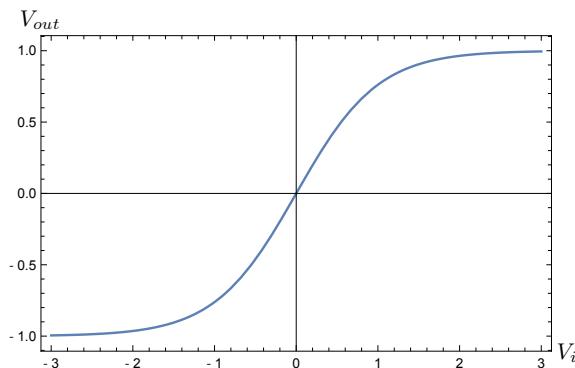


Figure 18.18: The typical input-output relationship is "S" shaped because of various limits that eventually clip the waveform.

We already encountered this when we derived the large signal transfer function of the MOS differential pair (see *Fig. 16.10*). Also, all of the transistors we have studied have a non-linear *I-V* and *C-V* curve. This leads to harmonic and intermodulation distortion. This is a topic we will cover in detail in other courses, such as a communication circuits or advanced analog circuits course.

As a simple example, in *Fig. 18.19*, the ideal sinusoidal output is distorted due to **harmonics of the waveform**. Harmonics are generated due to non-linearity. For example, if we model a small portion of the *I-V* curve with a Taylor series expansion, we have:

$$I = a_1 V_i + a_2 V_i^2 + a_3 V_i^3 + \dots \quad (18.29)$$

If you square a sinusoid, you get the second harmonic:

$$\cos^2(\omega t) = \frac{1}{2} (1 + \cos(2\omega t)) \quad (18.30)$$

If we cube a sinusoid:

$$\cos^3(\omega t) = \cos(\omega t) \cos^2(\omega t) = \frac{1}{2} (\cos(\omega t) + \cos(\omega t) \cos(2\omega t)) \quad (18.31)$$

Recall the following identity:

$$\cos(a) \cos(b) = \frac{1}{2} (\cos(a-b) + \cos(a+b)) \quad (18.32)$$

Using the trigonometric identity, the cubic can be simplified into a sum of the fundamental at a frequency ω and third harmonic:

$$\cos^3(\omega t) = \frac{3}{4} \cos(\omega t) + \frac{1}{4} \cos(3\omega t) \quad (18.33)$$

These harmonics lead to waveform distortion shown in *Fig. 18.19*. Fortunately, in negative feedback systems the input voltage swing is small, and these harmonics are very small and unnoticeable with the naked eye. One has to use a very sensitive measurement system, such as a spectrum analyzer, to detect the distortion under normal operating conditions.

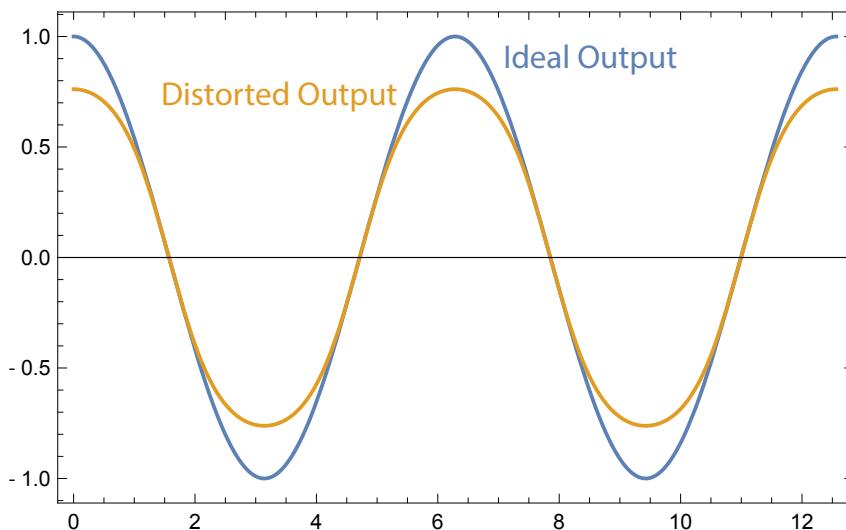


Figure 18.19: When a sinusoidal waveform experiences non-linearity as shown in *Fig. 18.18*, the output waveform is no longer a pure tone, but contains harmonics, that lead to distortion.

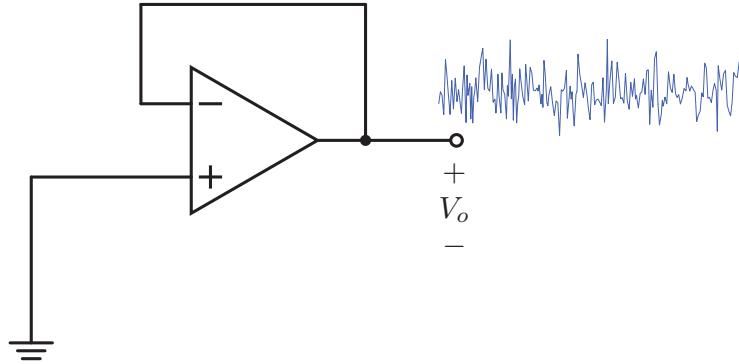


Figure 18.20: When we short the inputs of an amplifier and examine the output with a sensitive instrument, we find a small random AC signal. This signal is always present and depends on the temperature of the setup due to the thermodynamic origin.

18.5.2 Op-Amp Noise

Imagine that you short circuit the input of your op-amp as shown in *Fig. 18.20*. In addition to offset voltage, you would also observe a random fluctuation of the output signal. The signal is very small and it might be below the resolution limits of the oscilloscope, but a more sensitive instrument such as a spectrum analyzer clearly shows this noise. The noise tends to be flat in spectrum and for this reason it is often called "**white noise**". While we sometimes call interference signals noise, here we are referring to a random signal originating from thermodynamic properties of the transistor, rather than human made noise that couples into the circuit. Since it is a random signal, it is characterized by its statistical properties, such as the standard deviation of the signal.

Op-amps, like all amplifiers, have noise. We model the noise in the same way that we modeled offset voltages, as input voltages and/or currents as shown in *Fig. 18.21*. Unlike offset voltages, these are AC signals that are characterized in the frequency domain. Since the signal is usually flat with frequency, the noise spectral density, or the amount of noise power per unit bandwidth, is stated in datasheets. To get the voltage or current noise, you must integrate the noise over the bandwidth of interest. If the noise has a flat spectrum, you multiply the noise power density by the bandwidth of interest (approximately the bandwidth of the amplifier) and take the square root. This is why you will often see voltage and current noise specified in units of square root of hertz.

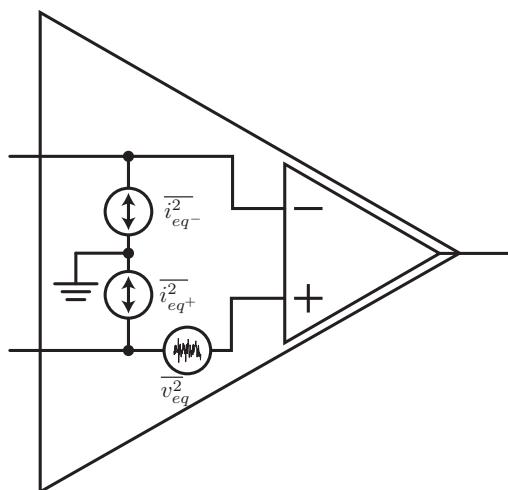


Figure 18.21: The noise of an op-amp can be modeled by a noiseless op-amp and internal input voltage and current noise sources as shown.

In addition to active devices like transistors, resistors also generate noise due to the random motion of carriers (electrons) in the solid. At any given time instant, the number of carriers moving "left" versus "right" is not equal, and so the device is conducting current. This is why the current is very small, because it is the imbalance of carrier motion. The signal is an AC signal with current going positive, and then almost immediately going negative with very high rate (frequency). The **noise current** is proportional to the temperature T , because it is due to the random thermal motion of carriers. In a resistor, it can be shown that the current takes on a value of:

$$i_R = \sqrt{4kT \frac{1}{R} \Delta f} \quad \text{Noise current in a resistor} \quad (18.34)$$

In Eq. 18.34 k is Boltzmann's constant, and Δf is the bandwidth of interest. Using a Norton to Thévenin transformation, we can state this as a **noise voltage** of value:

$$v_R = \sqrt{4kTR\Delta f} \quad \text{Noise voltage in a resistor} \quad (18.35)$$

A handy rule of thumb to remember is that a $1\text{k}\Omega$ resistor has an integrated noise of $4\mu\text{V}$ for a 1MHz of bandwidth. Because resistors are used in the feedback of amplifiers, they also contribute noise. Low noise designs therefore favor small resistors in feedback. In the design of signal processing chains, we must always ensure that the minimum signal level is much larger than noise level, otherwise the quality of the signal, or the **signal-to-noise ratio** (SNR) degrades.

18.6 Datasheets

Now that we have covered nearly all aspects of real op-amps, it is fun to look at some typical datasheets of practical amplifiers, and to appreciate all of the specifications. Let's take the Analog Devices ADA4891 CMOS op-amp, summarized in Fig. 18.22. This is advertised as a general purpose high speed op-amp. It has a *gain bandwidth product* of 240MHz , a *slew rate* of $170\frac{\text{V}}{\mu\text{s}}$, a *DC offset* of 2mV and an *open-loop DC gain* of 66dB . Being a CMOS op-amp, the *input impedance* is very large at $5\text{G}\Omega$, and it has an *input common mode range* of $-V_s - 0.3\text{V}$ to $+V_s - 0.8\text{V}$ (V_s is the supply voltage). The *common-mode rejection ratio* (CMRR) is 88dB . Take a look at the table and make sure you understand most, if not all, of the specifications!

It's important to realize that there are hundreds of op-amps that suit different applications. For example, in Fig. 18.23, we show a table of op-amps from Analog Devices that are sorted by the *GBW* product. The fastest op-amps have amazing *unity gain frequencies* (3.8GHz) and very high *slew rates* approaching $\frac{1\text{GV}}{\text{s}}$. The cost is power consumption, which is 10's of mA . On the other end of the spectrum, if we look for very low power op-amps, suitable for sensor or IoT applications, we also find amplifiers with amazing specifications, such as sub micro-amp current consumption, but at a severe penalty of speed, both in *GBW* product and in slew rate.

Table 1.

Parameter	Test Conditions/Comments	Min	Typ	Max	Unit
DYNAMIC PERFORMANCE					
–3 dB Small-Signal Bandwidth	ADA4891-1/ADA4891-2 , $G = +1$, $V_o = 0.2 \text{ V p-p}$ ADA4891-3/ADA4891-4 , $G = +1$, $V_o = 0.2 \text{ V p-p}$ ADA4891-1/ADA4891-2 , $G = +2$, $V_o = 0.2 \text{ V p-p}$, $R_L = 150 \Omega$ to 2.5 V ADA4891-3/ADA4891-4 , $G = +2$, $V_o = 0.2 \text{ V p-p}$, $R_L = 150 \Omega$ to 2.5 V	240			MHz
Bandwidth for 0.1 dB Gain Flatness	ADA4891-1/ADA4891-2 , $G = +2$, $V_o = 2 \text{ V p-p}$, $R_L = 150 \Omega$ to 2.5 V , $R_F = 604 \Omega$ ADA4891-3/ADA4891-4 , $G = +2$, $V_o = 2 \text{ V p-p}$, $R_L = 150 \Omega$ to 2.5 V , $R_F = 374 \Omega$	25			MHz
Slew Rate, t_s/t_f					
–3 dB Large-Signal Frequency Response	$G = +2$, $V_o = 2 \text{ V step}$, 10% to 90%	170/210			V/ μs
Settling Time to 0.1%	$G = +2$, $V_o = 2 \text{ V p-p}$, $R_L = 150 \Omega$ $G = +2$, $V_o = 2 \text{ V step}$	40			MHz
28					ns
NOISE/DISTORTION PERFORMANCE					
Harmonic Distortion, HD2/HD3	$f_C = 1 \text{ MHz}$, $V_o = 2 \text{ V p-p}$, $G = +1$ $f_C = 1 \text{ MHz}$, $V_o = 2 \text{ V p-p}$, $G = -1$	–79/–93			dBc
Input Voltage Noise	$f = 1 \text{ MHz}$	–75/–91			dBc
Differential Gain Error (NTSC)	$G = +2$, $R_L = 150 \Omega$ to 2.5 V	9			nV/ $\sqrt{\text{Hz}}$
Differential Phase Error (NTSC)	$G = +2$, $R_L = 150 \Omega$ to 2.5 V	0.05			%
All-Hostile Crosstalk	$f = 5 \text{ MHz}$, $G = +2$, $V_o = 2 \text{ V p-p}$	0.25			Degrees
		–80			dB
DC PERFORMANCE					
Input Offset Voltage	$T_{MIN} \text{ to } T_{MAX}$	± 2.5	± 10		mV
	W grade only, $T_{MIN} \text{ to } T_{MAX}$	± 3.1			mV
Offset Drift		± 3.1	± 16		mV
Input Bias Current		6			$\mu\text{V}/^\circ\text{C}$
	-50	+2	+50		pA
Open-Loop Gain	$W \text{ grade only}, T_{MIN} \text{ to } T_{MAX}$ $R_L = 1 \text{ k}\Omega$ to 2.5 V	–50	+50		nA
	$W \text{ grade only}, T_{MIN} \text{ to } T_{MAX}, R_L = 1 \text{ k}\Omega$ to 2.5 V $R_L = 150 \Omega$ to 2.5 V	77	83		dB
		66			dB
		71			dB
INPUT CHARACTERISTICS					
Input Resistance		5			$\text{G}\Omega$
Input Capacitance		3.2			pF
Input Common-Mode Voltage Range		– V_S – 0.3 to + V_S – 0.8			V
Common-Mode Rejection Ratio (CMRR)	$V_{CM} = 0 \text{ V}$ to 3.0 V	88			dB

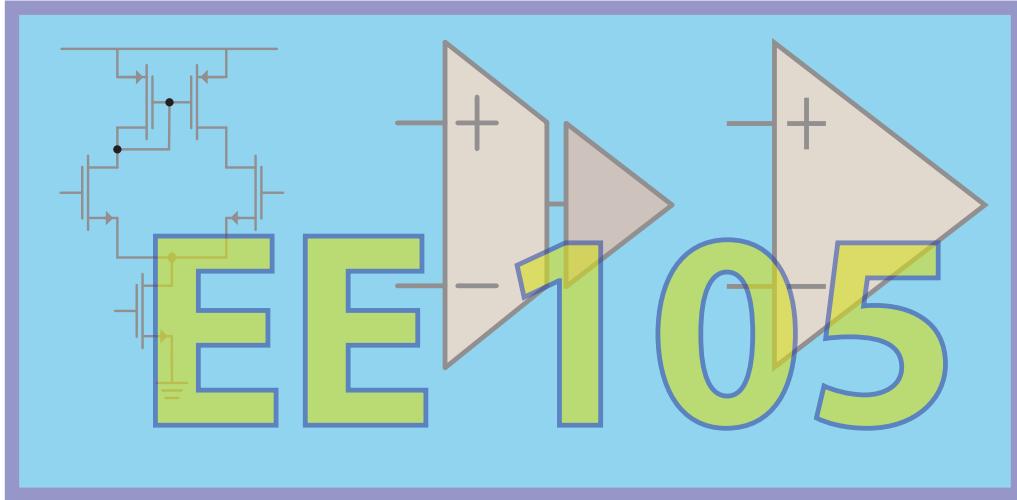
Figure 18.22: The datasheet for the ADA4891 op-amp. See https://www.analog.com/media/en/technical-documentation/data-sheets/ADA4891-1_4891-2_4891-3_4891-4.PDF.

Part#	Vsupply span (min) (V)	Vsupply span (max) (V)	Iq/Amp (typ) (A)	Amps per Package	GBP (typ) (Hz)	Slew Rate (typ) (V/us)	Ibias (max) (A)	Vos (max) (V)	CMRR (min) (dB)	0.1 to 10 Hz Voltage Noise (typ) (V p-p)	VNoise Density (typ) (V/rtHz)	US Price 1000 to 4999 (\$ US)
AD8099	5	12	15m	1	3.8G	470	13μ	500μ	-105	-	950p	\$2.00
AD8003	4.5	10	9.5m	3	1.65G	3.8k	-	-	-48	-	1.8n	\$2.92
ADA4895- 1	3	10	3m	1	1.5G	943	6μ	350μ	-100	99n	1n	\$1.89
ADA4895- 2	3	10	3m	2	1.5G	943	6μ	350μ	-100	99n	1n	\$3.21
AD8021 	4.5	24	7.8m	1	1G	130	11.3μ	1m	-98	-	2.1n	\$1.31
AD8001	6	12	5m	1	880M	1.2k	25μ	5.5m	-54	-	2n	\$1.51
AD829	9	36	5.3m	1	750M	230	7μ	1m	-120	-	1.7n	\$2.78
ADA4861- 3	5	12	16.1m	3	730M	680	-	-	-56.5	-	3.2n	\$0.96

Figure 18.23: A list of op-amps offered by Analog Devices, sorted by gain-bandwidth product. These are the highest bandwidth and highest slew rate amplifiers. See <https://www.analog.com/en/products/amplifiers/operational-amplifiers.html>.

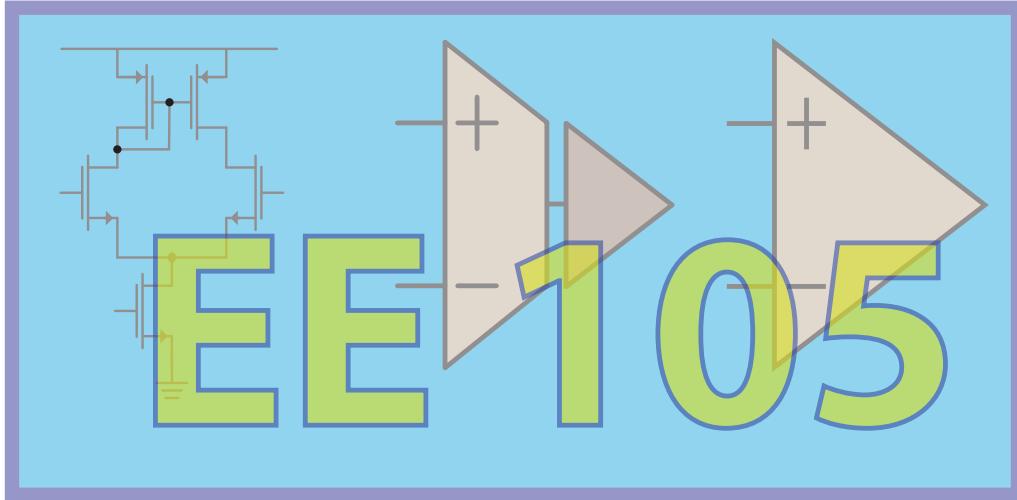
Part#	Vsupply span (min) (V)	Vsupply span (max) (V)	Iq/Amp (typ) (A)	Amps per Package	GBP (typ) (Hz)	Slew Rate (typ) (V/us)	Ibias (max) (A)	Vos (max) (V)	CMRR (min) (dB)	0.1 to 10 Hz Voltage Noise (typ) (V p-p)	VNoise Density (typ) (V/rtHz)	US Price 1000 to 4999 (\$ US)
ADA4312- 1	12	12	-	1	-	2.1k	-	-	-	-	-	\$1.89
ADLD8403	11.75	13.2	-	2	-	-	-	-	-	-	-	\$1.59
AD8398A	12	12	-	1	-	600	-	-	-	-	4.8n	\$1.45
AD8390A	10	24	-	-	-	260	-	-	-	-	5n	\$1.08
AD8398	12	12	-	1	-	820	-	-	-	-	2.85n	\$2.30
AD8504	1.8	5	750n	4	7k	4m	10p	3m	-67	6μ	190n	\$1.00
AD8502	1.8	5	750n	2	7k	4m	10p	3m	-67	6μ	190n	\$0.70
AD8500 	1.8	5	750n	1	7k	4m	10p	1m	-75	-	190n	\$0.71
OP481	2.7	12	5μ	4	105k	28m	10n	1.5m	-65	10μ	85n	\$3.65
OP281	2.7	12	5μ	2	105k	28m	10n	1.5m	-65	10μ	75n	\$2.79
ADA4505- 1	1.8	5	9μ	1	50k	6m	2p	3m	-90	2.95μ	65n	\$0.41

Figure 18.24: A list of op-amps offered by Analog Devices, sorted by power consumption. These are the lowest power amplifiers suitable for IoT and sensor applications. See <https://www.analog.com/en/products/amplifiers/operational-amplifiers.html>.



19. References

- [1] Paul R. Gray et al. *Analysis and Design of Analog Integrated Circuits*. Wiley, 2017. ISBN: 978-1-118-07889-1 (cited on page 273).
- [2] Chenming Calvin Hu. *Modern Semiconductor Devices for Integrated Circuits*. Pearson, 2010. ISBN: 0-13-608525-3 (cited on page 140).
- [3] David Lay. *Linear Algebra and its Applications, 4th Edition*. Pearson, 2012. ISBN: 978-0-321-38517-8 (cited on page 359).
- [4] Donald Neamen. *Semiconductor Physics and Devices, Third Edition*. McGraw-Hill, 2003. ISBN: 978-0-072-32107-4 (cited on page 77).
- [5] Robert F. Pierret. *Semiconductor Device Fundamentals*. Addison-Wesley, 1996. ISBN: 0-201-54393-1 (cited on page 140).
- [6] Edward M. Purcell. *Electricity and Magnetism (Berkeley Physics Course, Vol. 2)*. McGraw-Hill, 1984. ISBN: 978-0-070-04908-6 (cited on page 69).
- [7] Adel S. Sedra. *Microelectronic Circuits*. Oxford University Press, 2020. ISBN: 978-0-190-85346-4.
- [8] James Stewart. *Calculus: Early Transcendentals, 8th Edition*. Cengage Learning, 2015. ISBN: 978-1-285-74155-0 (cited on page 357).
- [9] Wikipedia. *Wikipedia, The Free Encyclopedia*. [Online; accessed 18-August-2022]. 2022. URL: <https://en.wikipedia.org>.



A. Appendix: Calculus Review

A.1 Geometric Series

Recall the *geometric series* from calculus[8]:

$$a + ar + ar^2 + ar^3 + \cdots + ar^{n-1} = \sum_{n=1}^{\infty} ar^{n-1} \quad a \neq 0 \quad (\text{A.1})$$

Each term is obtained from the preceding one by multiplying it with the *common ratio*, r . If $|r| \geq 1$, then the series diverges. However, if $|r| < 1$ the series is convergent, and its sum can be expressed as:

$$\sum_{n=1}^{\infty} ar^{n-1} = \frac{a}{1-r} \quad |r| < 1 \quad (\text{A.2})$$

A.2 Taylor Series

Also recall from calculus[8] the general form of a power series centered about a :

$$f(x) = \sum_{n=0}^{\infty} c_n (x-a)^n = c_0 + c_1(x-a) + c_2(x-a)^2 + c_3(x-a)^3 + \dots \quad (\text{A.3})$$

How do we solve for the coefficients of this infinite series?

We can begin by setting $x = a$ for every term, which yields:

$$c_0 = f(a) \quad (\text{A.4})$$

Now we want to solve for each n th consecutive coefficient in the series. To do this, we take the n th derivative of the the series and use the same strategy of setting $x = a$. This eliminates all terms except for the coefficient of interest we aim to solve for.

Starting with the first derivative, we have:

$$\frac{df(x)}{dx} = c_1 + 2c_2(x-a) + 3c_3(x-a)^2 + 4c_4(x-a)^3 + \dots$$

This yields:

$$c_1 = \frac{1}{1} \cdot \frac{df(a)}{dx}$$

Continuing with the second derivative:

$$\frac{df(x)^2}{d^2x} = (1 \cdot 2)c_2 + (2 \cdot 3)c_3(x-a) + (3 \cdot 4)c_4(x-a)^2 + \dots$$

Yielding:

$$c_2 = \frac{1}{1 \cdot 2} \cdot \frac{df(a)^2}{d^2x}$$

Then the third derivative:

$$\frac{df(x)^3}{d^3x} = (1 \cdot 2 \cdot 3)c_3(x-a) + (1 \cdot 2 \cdot 3 \cdot 4)c_4(x-a) + \dots$$

Which yields:

$$c_3 = \frac{1}{1 \cdot 2 \cdot 3} \cdot \frac{df(a)^3}{d^3x}$$

A pattern has presented itself which will continue for all consecutive terms. From the pattern we can see:

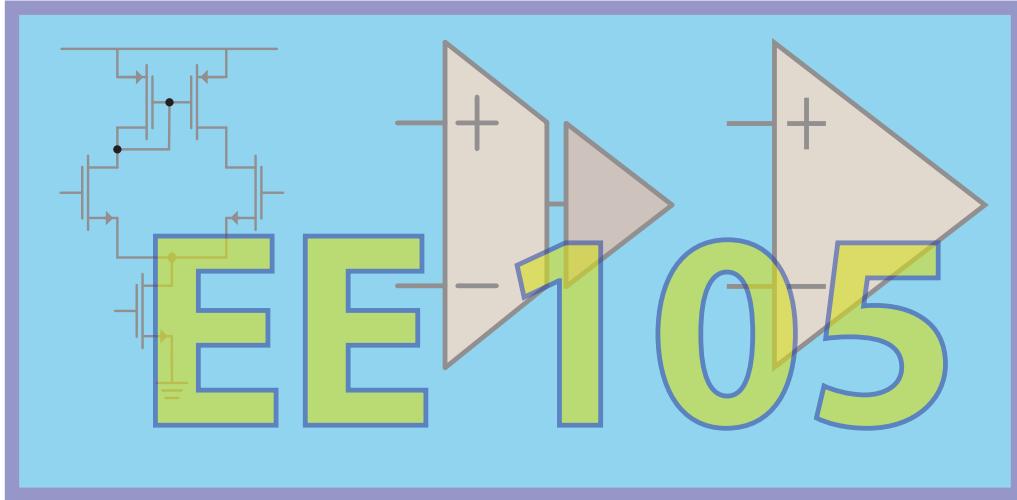
$$c_n = \frac{1}{n!} \cdot \frac{df(a)^n}{d^n x} \quad (\text{A.5})$$

Plugging *Eq. A.4* and *Eq. A.5* into *Eq. A.3* gives us the general form of a *Taylor series* using prime notation:

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!} (x-a)^n = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \dots \quad (\text{A.6})$$

For the special case where $a = 0$ in *Eq. A.6*, we have the general form of a *Maclaurin series*:

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n!} x^n = f(0) + \frac{f'(0)}{1!}x + \frac{f''(0)}{2!}x^2 + \dots \quad (\text{A.7})$$



B. Appendix: Linear Algebra Review

B.1 Linear Dependence and Independence

First let's review the definitions of linear dependence and independence[3]:

Definition B.1.1 An indexed set of vectors $\{v_1, v_2, \dots, v_n\}$ in \mathbb{R}^n is said to be **linearly independent** if $\alpha_1 \cdot v_1 + \alpha_2 \cdot v_2 + \dots + \alpha_n \cdot v_n = 0$ has only the trivial solution, where all $\alpha_i \in \mathbb{Z}$.

Definition B.1.2 An indexed set of vectors $\{v_1, v_2, \dots, v_n\}$ in \mathbb{R}^n is said to be **linearly dependent** if there exists a $\alpha_1 \dots \alpha_n$, not all zero, that satisfies $\alpha_1 \cdot v_1 + \alpha_2 \cdot v_2 + \dots + \alpha_n \cdot v_n = 0$, where all $\alpha_i \in \mathbb{Z}$.

B.1.1 Proof of Sine and Cosine's Linear Independence

Claim. $a \cdot \sin(x) + b \cdot \cos(x) = 0$ is linearly independent.

Proof. Using the definition of linear independence, we will show that a and b must be zero for all x in the interval $[0, 2\pi]$.

We can find a by setting $x = \frac{\pi}{2}$:

$$a \cdot \sin\left(\frac{\pi}{2}\right) + b \cdot \cos\left(\frac{\pi}{2}\right) = 0$$

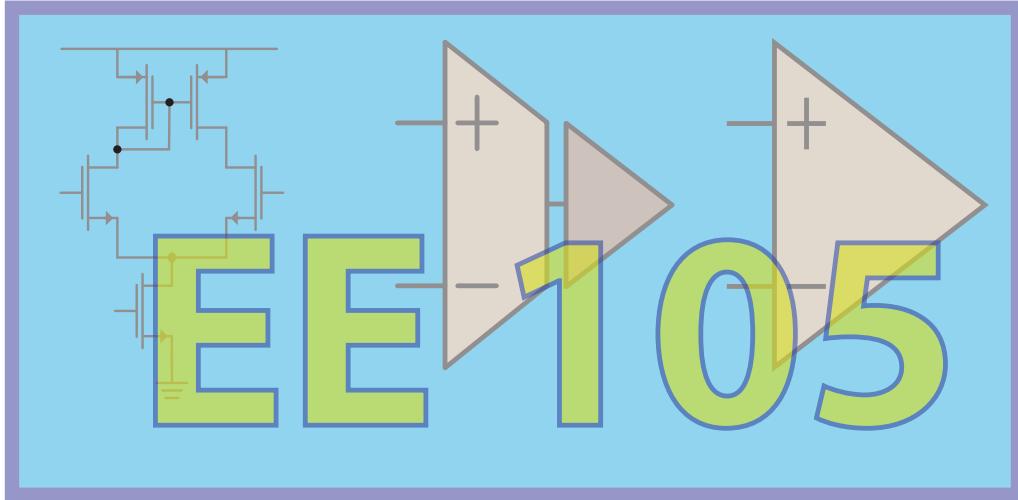
$$a = 0$$

Next, we can find b by setting $x = 0$:

$$a \cdot \sin(0) + b \cdot \cos(0) = 0$$

$$b = 0$$

Since sine and cosine are periodic, it follows that a and b must be 0 for all x in $(-\infty, \infty)$. Thus, the claim holds. ■



C. Appendix: Common Drain Output Resistance

Here we will use KCL to derive the common drain output resistance from *Fig. 12.18* in *Ch. 12*, which is shown again here for convenience.

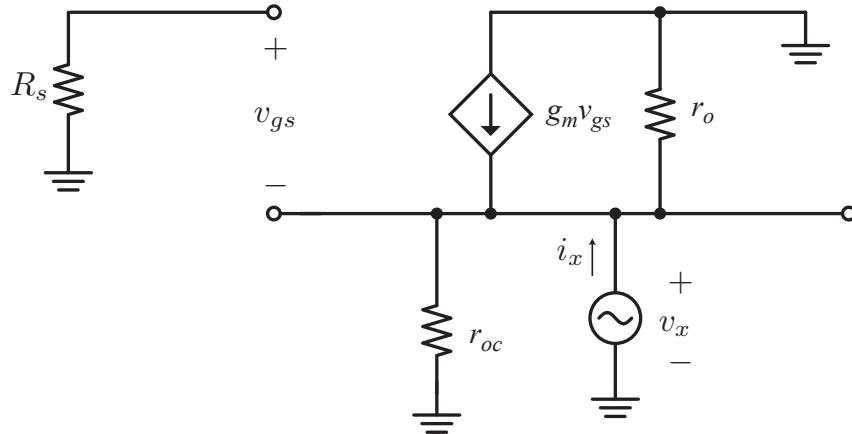


Figure C.1: Small-signal circuit schematic for calculation of output impedance.

Beginning with KCL at node v_x :

$$0 = -i_x + \frac{v_x}{r_o} + \frac{v_x}{r_{oc}} - g_m v_{gs} \quad v_g = 0$$

Factoring v_x and rearranging:

$$i_x = v_x \left(\frac{1}{r_o} + \frac{1}{r_{oc}} + g_m \right)$$

Now to solve for the output resistance:

$$R_x = \frac{v_x}{i_x} = \frac{1}{\frac{1}{r_o} + \frac{1}{r_{oc}} + g_m} \quad (\text{C.1})$$

$$= \frac{1}{\frac{r_o + r_{oc} + g_m r_o r_{oc}}{r_o r_{oc}}} \quad (\text{C.2})$$

$$= \frac{1}{\frac{r_o + r_{oc}}{r_o r_{oc}} + g_m} \quad (\text{C.3})$$

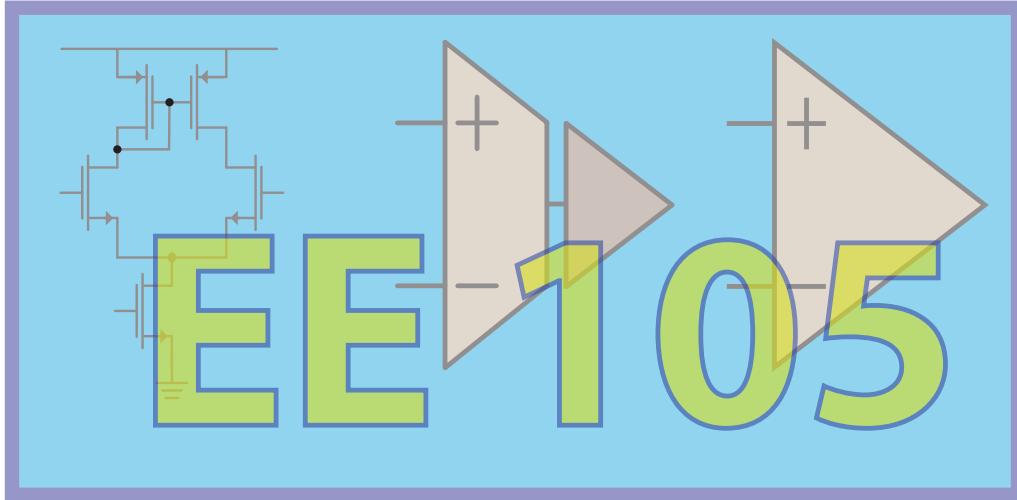
$$= \frac{1}{(r_o \parallel r_{oc})^{-1} + g_m} \quad (\text{C.4})$$

$$= (r_o \parallel r_{oc})^{-1} \parallel g_m \quad (\text{C.5})$$

$= (r_o \parallel r_{oc}) \parallel \frac{1}{g_m}$

(C.6)

We have now verified that the result in *Eq. 12.34* agrees with a proper KCL method. It should now be clear how useful it is to recognize shortcuts in circuit analysis, especially when we are getting into larger amplifier diagrams.



D. Appendix: Glossary of Terms

D.1 A

Acceptor - An element used in doping (Group *III* for *Si*) that has less electrons than the semiconductor it is being injected into in order to accept an electron from the valence band in order to free a hole to be available for conducting current; increases the hole concentration.

- If the energy absorbed at the acceptor is precisely equal to the electron binding energy, the released hole will have the lowest possible energy in the valence band, namely E_V .

Acoustic Phonon - Coherent movements of atoms of the lattice out of their equilibrium positions.

- Occur at lower temperatures and energy.

Amphoteric Dopant - An impurity that can act as either a donor or acceptor.

- *Si* typically replaces *Ga* in the *GaAs* (*III* – *V*) lattice, and is a popular *N-type* dopant.
- The size of the dopant (*Si*) compared to the *III* – *V* elements determines which species will be replaced, and whether the dopant becomes a donor or an acceptor.

Anode - In an electrical device, the anode is the *positive* terminal in which conventional current enters.

Annealing - A heat treatment that alters the physical and sometimes chemical properties of a material to increase its ductility and reduce its hardness, making it more workable.

- It involves heating a material above its recrystallization temperature, maintaining a suitable temperature for an appropriate amount of time and then cooling.

D.2 B

Band Bending - The resulting variation of E_C and E_V due to an electric field.

Band Gap - The intervening energy gap between the upper band and lower band of allowed energy states. In *Si*, this value is typically around 1.12 eV.

Biassing - The setting of initial operating conditions (current and voltage) of an active device in an amplifier.

Bipolar Transistor - A type of transistor that uses both electrons and holes as charge carriers. It has three different sections of semiconductor material known as the base, collector, and emitter.

- A bipolar transistor allows a small current injected at one of its terminals to control a much larger current flowing between the terminals, making the device capable of amplification or switching.
- *NPN* BJT's have a *P-type* base, and an *N-type* collector. The emitter region is heavily doped N^+ .
- *PNP* BJT's have an *N-type* base, and a *P-type* collector. The emitter region is heavily doped P^+ .
- *NPN* transistors exhibit higher transconductance and speed than *PNP* transistors, because the electron mobility is larger than the hole mobility.
- The collector current, I_C , is determined by the rate of electron injection from the emitter into the base.

Boltzmann Distribution - A probability distribution or probability measure that gives the probability that a system will be in a certain state as a function of that state's energy and the temperature of the system.

D.3 C

Carriers - The entities that transport charge from place to place inside a material, and give rise to electrical currents. Most of the carriers are grouped energetically in the near vicinity of the conduction band edge or valence band edge.

Cathode - In an electrical device, the cathode is the *negative* terminal in which conventional current exits.

Charge Neutrality - A relationship that provides the general tie between the carrier and dopant concentrations. It says that a *uniformly doped* semiconductor must have no net charge.

- Below is the derivation for charge neutrality. Remember that donors become positively charged ions, and acceptors become negatively charged ions.

$$\begin{aligned} 0 &= p + n + \text{donors} + \text{acceptors} \\ &= qp - qn + N_D - N_A \\ \Rightarrow & \boxed{n + N_A = p + N_D} \end{aligned}$$

Compensated Semiconductor - When N_A and N_D are comparable and nonzero, the material is said to be compensated, in which case N_A and N_D must be retained in all carrier concentration expressions.

Conduction Band - The upper band of allowed energy states in *Si*.

Current - The charge per unit time crossing an arbitrarily chosen plane of observation oriented normal to the direction of current flow.

D.4 D

Degenerate Semiconductor - A semiconductor with such a high level of doping that the material starts to act more like a metal than as a semiconductor. Unlike non-degenerate semiconductors, these kind of semiconductor do not obey law of mass action, which relates intrinsic carrier concentration with temperature and band gap.

- Whenever E_F lies in the band gap closer than $3kT$ to either band edge, or actually penetrates one of the bands.

Density of States - The energy distribution of allowed states; tells us how many states exist at a given energy.

- For *Si* and *Ge* the density of states are of the same order of magnitude at corresponding energies.

Diffusion - A process where particles tend to spread out because of their random thermal motion. Particles spread from regions of high concentration into regions of low concentration. Thermal motion, not atomic repulsion, is the enabling action behind the diffusion process.

Direct Bandgap - The minimum energy state in the conduction band and the maximum energy state in the valence band occur at the same wave vector, k . This allows carriers to directly emit a photon, or recombine.

- Dominant R-G mechanism is band-to-band.

Donor - An element used in doping (Group V for *Si*) that has more electrons than the semiconductor it is being injected into in order to donate an electron to be available for conducting current; increases the electron concentration.

- If the energy absorbed at the donor is precisely equal to the electron binding energy, the released electron will have the lowest possible energy in the conducting band, namely E_C .

Doping - The addition of controlled amounts of specific impurity atoms with the express purpose of increasing either the electron or the hole concentration.

- All doped semiconductors become intrinsic at sufficiently high temperature where $n_i \gg |N_D - N_A|$.

Drift - Charged-particle motion in response to an applied electric-field.

- Because of collisions with ionized impurity atoms and thermally agitated lattice atoms, carrier acceleration is frequently interrupted and said to be scattered.
- Averaging over all electrons or holes at any given time, the resultant motion of each carrier type can be described in terms of a constant drift velocity, \vec{v}_d .
- Electrons in the conduction band and holes in the valence band gain and lose energy via collisions with the semiconductor lattice and are nowhere near stationary even under equilibrium conditions.

Drift Current - When an electric field is applied across a semiconductor, the resulting force on the carriers tends to accelerate the $+q$ charged holes in the direction of the electric field, and the $-q$ charged electrons in the direction opposite the electric field.

D.5 E

Early Effect - the variation in the effective width of the base in a bipolar junction transistor (BJT) due to a variation in the applied base-to-collector voltage.

- A greater reverse bias across the collector–base junction, for example, increases the collector–base depletion width, thereby decreasing the width of the charge carrier portion of the base.

Eddy Current - Loops of electrical current induced within conductors by a changing magnetic field in the conductor according to Faraday's law of induction.

Effective Mass - The apparent mass of electrons and holes inside of the crystal structure of a semiconductor, which is different from the mass of electrons within a vacuum due to the atomic structure of the semiconductor crystal, and how the atoms interact with each other inside of it.

- Inside the crystal electrons will collide with the semiconductor atoms, which causes a periodic negative acceleration of the carriers. In addition to this, electrons inside of the semiconductor crystal are also subject to complex crystalline fields in addition to the electric field.
- The motion of carriers inside a semiconductor crystal is a quantum mechanical phenomenon, because the electrons moving through a solid experience a specific potential energy which dictates the mass of a wave associated with the electron. This is why assigning an effective mass allows us to treat electrons and holes as quasi-classical particles and use classical particle relationships in most analyses.
- The effective mass is a result of electron interaction with the lattice (phonons). Holes spend more time interacting with phonons because they have a smaller velocity than electrons. This results in holes usually having a larger effective mass.
- The effective mass depends on the effective potential the electron or hole feels when moving through the lattice. The potential seen by the electrons and the holes is different, and this is why they have different effective masses.

Einstein Relation - For a particle with electrical charge q , its electrical mobility μ_q is related to its generalized mobility μ by the equation $\mu = \frac{\mu_q}{q}$. The parameter μ_q is the ratio of the particle's terminal drift velocity to an applied electric field. Hence, the equation in the case of a charged particle is given as:

$$D = \frac{\mu_q k_B T}{q}, \quad (\text{D.1})$$

where,

- D is the diffusion coefficient ($m^2 s^{-1}$).
- μ_q is the electrical mobility ($m^2 V^{-1} s^{-1}$).
- q is the electrical charge of a particle in coulombs (C).
- T is the electron temperature or ion temperature in plasma (K).
- The Einstein relation is derived in equilibrium, however it is valid in non-equilibrium.

Electrochemical Potential - In electrochemistry, the electrochemical potential is a thermodynamic measure of chemical potential that does not omit the energy contribution of electrostatics.

- Electrochemical potential is expressed in the unit of J/mol .
- The difference in Fermi energy on either end of a device.

Energy Bands - By conceptually bringing N silicon atoms closer and closer together, the interatomic forces lead to a spread in the allowed energies which give rise to closely spaced sets of allowed states known as energy bands. At the interatomic distance corresponding to the Si lattice spacing, the distribution of allowed states consists of two bands separated by an intervening energy gap.

- For every possible momentum state there is another state with an oppositely directed momentum of equal magnitude.
- Thus, if a band is completely filled with electrons, the *net* momentum of the electrons in the band is always identically zero.
- No current can arise from the electrons in a completely filled energy band.
- When an electric field exists inside a material, the band energies become a function of position.

Equilibrium - A term used to describe the unperturbed state of a system. For a semiconductor in equilibrium conditions this means that there are no external voltages, magnetic fields, stresses, or other perturbing forces acting on the semiconductor.

- All observables are invariant with time.
- Under equilibrium conditions the Fermi level inside a material or a group of materials in intimate contact is invariant as a function of position.
- Under equilibrium conditions the total current is zero, and the drift and diffusion components of a given carrier are required to be of equal magnitude, but opposite polarity.
- Even under equilibrium conditions, nonuniform doping will give rise to carrier concentration gradients, a built-in electric field, and non-zero current components.

Equipotential - A region in space where every point in it is at the same potential.

Extrinsic Semiconductor - A doped semiconductor; a semiconductor whose properties are controlled by added impurity atoms.

D.6 F

Fermi Energy - A concept in quantum mechanics usually referring to the energy difference between the highest and lowest occupied single-particle states in a quantum system of non-interacting fermions at absolute zero temperature. The term "Fermi energy" is often used to refer to a different yet closely related concept, the "Fermi level". There are a few key differences between the Fermi level and Fermi energy:

- The Fermi energy is only defined at absolute zero, while the Fermi level is defined for any temperature.
- The Fermi energy is an energy difference (usually corresponding to a kinetic energy), whereas the Fermi level is a total energy level including kinetic energy and potential energy.

- The Fermi energy can only be defined for non-interacting fermions (where the potential energy or band edge is a static, well defined quantity), whereas the Fermi level remains well defined even in complex interacting systems, at thermodynamic equilibrium.

Fermi Function - Specifies how many of the existing states at a given energy E will be filled with an electron, or equivalently, specifies, under equilibrium conditions, the probability that an available state at an energy E will be occupied by an electron.

- As $T \rightarrow 0K$, all states at energies below E_F will be filled, and all states at energies above E_F will be empty.
- For $T > 0 K$ and $E \geq E_F + 3kT$, the Fermi function, or filled-state probability decays exponentially to zero with increasing energy; most states at energies $3kT$ or more above E_F will be empty.
- For $T > 0 K$ and $E \leq E_F - 3kT$, the probability that a given state will be *empty* decays exponentially to zero with decreasing energy; most states at energies $3kT$ or more below E_F will be filled.
- The Fermi Function applies *only* under equilibrium conditions.
- The Fermi function is universal in the sense that it applies to all materials; insulators, metals, and semiconductors. It is a statistical functions associated with electrons in general.

Fermi Level - The thermodynamic work required to add one electron to a solid-state body. When the Fermi level is positioned in the upper half of the band gap (or higher), the electron distribution greatly outweighs the hole distribution.

- Under equilibrium conditions, $\frac{dE_F}{dx} = \frac{dE_F}{dy} = \frac{dE_F}{dz} = 0$; i.e. the Fermi level inside a material or a group of materials in intimate contact is invariant as a function of position.

Fermi Level Pinning - In most real *MS – diodes* (both *P-type* and *N-type*), $\Phi_B \neq \Phi_M - \chi$. In the majority of semiconductors, surface charges tend to fix or "pin" the equilibrium Fermi level at a specific energy within the surface band gap. Because of this pinning effect, the observed barrier height normally varies only slightly with the metal used to fabricate the diode.

- There are high densities of energy states in the band gap at the metal-semiconductor interface.
- Some of these states are acceptor like, and may be neutral or negative.
- Some of these states are donor like, and may be neutral or positive.
- The net charge is zero when the Fermi level at the interface is around the middle of the silicon band gap.
- The result is that for any $\psi_M \approx 4.6 V$, there is a dipole at the interface that prevents the barrier height from moving very far from around 0.7 V.

Flat-Band Condition - A condition where the energy band (E_C and E_V) of the substrate is flat at the $Si - SiO_2$ interface.

- The flat-band voltage is the difference between the Fermi levels at the two terminals:

$$V_{FB} = \psi_g - \psi_s \quad (D.2)$$

Where ψ_g and ψ_s are the gate work function and the semiconductor work function, respectively, in volts.

Freeze Out - A failure to ionize due to too low of a temperature.

- $kT < E_{ionization}$

Frequency Response - A system's dependence on signal frequency of the output–input ratio of an amplifier or other device. The frequency response of a system is the quantitative measure of the magnitude and phase of the output as a function of input frequency.

D.7 G

Gain-Bandwidth Product - The product of an amplifier's bandwidth and the gain at which the bandwidth is measured.

Gate Induced Drain Leakage - A parasitic current that escapes through the portion of the thin gate oxide in MOSFETs that partially overlaps with the drain. GIDL occurs at drain voltages that are extremely small and much less than junction breakdown voltages.

- Occurs when V_{dg} is large enough to create band-bending that facilitates band-to-band tunneling between the drain and the gate, because the channel/drain junction is heavily reverse-biased, which causes the barrier to be narrowed.
- GIDL current increases with increasing V_d and decreasing V_g .

Generation - A process whereby electrons and holes are created.

D.8 H

Hole - The empty state in the valence band left behind by an electron that has moved to the conduction band.

Hysteresis - The phenomenon in which the value of a physical property lags behind changes in the effect causing it, as for instance when magnetic induction lags behind the magnetizing force.

Hot-Point Probe Measurement - A common technique for rapidly determining whether a semiconductor is *P-type* or *N-type*.

D.9 I

Impact Ionization - At sufficiently high reverse bias the minority electrons may gain enough kinetic energy such that as they collide with the lattice it creates electron hole pairs.

Impurity Scattering - The scattering of charge carriers by ionization in the lattice.

- Does not go up at higher temperatures.

Indirect Bandgap - Differs from a direct bandgap in that now the carriers must be assisted by a phonon in order to pass through an intermediate state before emitting a photon, or recombining.

- Dominant R-G mechanism is R-G center.

Ingot (semiconductor) - An oblong bar of silicon, which is used to cut and polish into chip wafers.

Intrinsic Fermi Level - A dashed line on a band diagram representing the expected positioning of the Fermi level if the material was intrinsic, and it serves as a reference energy level dividing the upper and lower halves of the band gap.

Intrinsic Semiconductor - An extremely pure semiconductor sample containing an insignificant amount of impurity atoms. Its properties are native to the material and not caused by external additives.

- The electrons and holes in an intrinsic semiconductor are equal because carriers within a very pure material can only be created in pairs.

D.10 L

Lattice Scattering - The scattering of ions by interaction with atoms in a lattice.[1] This effect can be qualitatively understood as phonons colliding with charge carriers.

- When the wavelength of the electrons is larger than the crystal spacing, the electrons will propagate freely throughout the metal without collision.

Law of the Junction - The number of electrons (or holes) crossing over from the *n-side* to the *p-side* increases exponentially as a function of the forward-bias voltage.

-

$$P_{p-side_{depletion\ edge}} = P_{n-side_{depletion\ edge}} \cdot e^{(\phi_{bi} - V_{applied})/\phi_T} \quad (\text{D.3})$$

- The sum of the in-flowing current is equal to the sum of out-flowing current.
- E_{Fp} and E_{Fn} do not change across the depletion region.

D.11 M

Majority Carrier - The most abundant carrier in a given semiconductor sample; electrons in an *N-type* material, holes in a *P-type* material.

MIM Capacitor - A capacitor that consists of parallel plates formed by two metal planes separated by a thin dielectric.

Minority Carrier - The least abundant carrier in a given semiconductor sample; holes in an *N-type* material, electrons in a *P-type* material.

Mobility - A measurement of the ease of carrier motion in a crystal; varies inversely with the amount of scattering taking place within a semiconductor; dependent on both doping and temperature.

- Increasing the motion-impeding collisions within a crystal decreases the mobility of the carriers (i.e. the mean free time between collisions).
- The mobility also varies inversely with the effective carrier mass; lighter carriers move more readily.
- Lattice scattering decreases with decreasing T , whereas ionized impurity scattering increases with decreasing T .
- Ionized impurity scattering becomes a larger and larger percentage of the overall scattering as the temperature is decreased.

MOS Capacitor - A simple two-terminal device composed of a thin ($0.01 \mu\text{m} - 1.0 \mu\text{m}$) SiO_2 layer sandwiched between a silicon substrate and a metallic field plate. The ideal MOS structure has the following properties:

- The metallic gate is sufficiently thick so that it can be considered an equipotential region under both AC and DC biasing conditions.
- The oxide is a *perfect insulator* with *zero current* flowing through the oxide layer under *all* static biasing conditions.
- There are no charge centers located in the oxide, or at the oxide-semiconductor interface.
- The semiconductor is uniformly doped.
- The semiconductor is sufficiently thick so that, regardless of the applied gate potential, a field-free region (the body, bulk, or substrate) is encountered before reaching the back contact.
- An *ohmic* contact has been established between the semiconductor and the metal on the back side of the device.
- The $\text{MOS} - C$ is a one-dimensional device taken to be a function of only the x -coordinate.

M-S Junction - A type of electrical junction in which a metal comes in close contact with a semiconductor material.

- M-S junctions can either be rectifying or non-rectifying.
- The rectifying metal-semiconductor junction forms a Schottky barrier, making a device known as a Schottky diode.
- The non-rectifying junction is called an ohmic contact.

D.12 N

N-type Material - A donor-doped material containing more electrons than holes.

D.13 O

Ohmic Contact - A non-rectifying electrical junction: a junction between two conductors that has a linear current-voltage (I-V) curve as with Ohm's law.

- Low resistance ohmic contacts are used to allow charge to flow easily in both directions between the two conductors, without blocking due to rectification or excess power dissipation due to voltage thresholds.
- *Boundary Condition of an Ohmic Contact* : The voltage across an ideal ohmic contact is zero. This means that the Fermi level cannot deviate from its equilibrium position, and therefore $n' = p' = 0$ at an ideal ohmic contact.

Operating Point - Also known as bias point, quiescent point, or Q-point, it is the DC voltage or current at a specified terminal of an active device (a transistor or vacuum tube) with no input signal applied.

Optical Phonon - A high energy (frequency) phonon. They are out-of-phase movements of the atoms in the lattice, one atom moving to the left, and its neighbor to the right.

- Occur at higher temperatures and with higher energy.

One-Sided Junction - A PN junction where one region is highly doped in comparison to the doping of the other region. In this junction, the concentration of one side impurity is considered.

Overdrive Voltage - The voltage between transistor gate and source in excess of the threshold voltage.

- Typically referred to in the context of MOSFET transistors.
- Also known as "excess gate voltage" or "effective voltage."
- Can be found using the following equation:

$$V_{OV} = V_{GS} - V_T \quad (\text{D.4})$$

D.14 P

P-type Material - An acceptor-doped material containing more holes than electrons.

Parasitic capacitance - An unavoidable and usually unwanted capacitance that exists between the parts of an electronic component or circuit simply because of their proximity to each other.

Pauli Exclusion Principle - States that electrons are restricted to single occupancy in allowed energy states.

Phonon - Vibrations of the lattice.

Photolithography - A general term used for techniques that use light to produce minutely patterned thin films of suitable materials over a substrate, such as a silicon wafer, to protect selected areas of it during subsequent etching, deposition, or implantation operations.

Photoresist - A light-sensitive material used in several processes, such as photolithography and photoengraving, to form a patterned coating on a surface.

D.15 Q

Quantum Mechanical Tunneling - A phenomenon where a wavefunction can propagate through a potential barrier.

- The transmission through the barrier can be finite and depends exponentially on the barrier height and barrier width.
- The wavefunction may disappear on one side and reappear on the other side. The wavefunction and its first derivative are continuous.
- In steady-state, the probability flux in the forward direction is spatially uniform. No particle or wave is lost.
- Tunneling occurs with barriers of thickness around $1 - 3 \text{ nm}$ and smaller.

Quality (Q) Factor - A dimensionless parameter that describes how underdamped an oscillator or resonator is. It is defined as the ratio of the initial energy stored in the resonator to the energy lost in one radian of the cycle of oscillation.

- Higher Q indicates a lower rate of energy loss and the oscillations die out more slowly.

- A pendulum suspended from a high-quality bearing, oscillating in air, has a high Q, while a pendulum immersed in oil has a low one.
- Resonators with high quality factors have low damping, so that they ring or vibrate longer.

D.16 R

Recombination - A process whereby electrons and holes (carriers) are annihilated or destroyed.

Rectifier - An electrical device that converts alternating current (AC), which periodically reverses direction, to direct current (DC), which flows in only one direction. The reverse operation is performed by the inverter.

- The process is known as rectification, since it "straightens" the direction of current.

Resistivity - A measure of a material's inherent resistance to current flow. Quantitatively, the proportionality constant between the electric field impressed across a homogeneous material and the total particle current per unit area flowing in the material.

R-G Centers - Special locations within the semiconductor, which are lattice defects or special impurity atoms such as gold in Si.

R-G Processes - The means whereby the carrier excess or deficit inside the semiconductor is stabilized (if the perturbation is maintained) or eliminated (if the perturbation is removed).

- **Band-to-Band Recombination** - The direct annihilation of a conduction band electron and a valence band hole. The excess energy released is typically a photon.
- **R-G Center Recombination** - Also called *indirect combination*. Due to the R-G centers, allowed electronic levels near the center of the band gap are introduced. First, one type of carrier strays into the vicinity of an R-G center and is caught by the potential well associated with the center, loses energy, and is trapped. Then, the opposite type of carrier comes along and is attracted to the already trapped carrier, loses energy, and annihilates. Typically produces thermal energy or lattice vibrations.
- **Auger Recombination** - When excess energy from the electron-hole recombination is transferred to electrons or holes that are subsequently excited to higher energy states within the same band.
- **Impact Ionization** - The inverse of Auger recombination.

D.17 S

Scattering -

- Lattice scattering
- Ionized impurity scattering

Schottky Barrier - A potential energy barrier for electrons formed at a metal–semiconductor junction.

- Schottky barriers have rectifying characteristics, suitable for use as a diode.
- The Schottky barrier height, denoted by ϕ_B , is a function of the metal and semiconductor.
 - There are actually two energy barriers.

- $q\phi_{B_n}$ is the barrier against electron flow between the metal and the *N-type* semiconductor.
- $q\phi_{B_p}$ is the barrier against hole flow between the metal and the *P-type* semiconductor.
- The sum of both barrier heights for any type of material is equal to its bandgap energy. For silicon:

$$\phi_{B_n} + \phi_{B_p} \approx E_g \quad (\text{D.5})$$

- There is a trend that ϕ_{B_n} and ϕ_{B_p} increase with an increasing metal work function. This can be explained by:

$$\phi_{B_n} = \psi_M - \chi_{Si} \quad (\text{D.6})$$

Where ψ_M is the metal work function, and χ_{Si} is the electron affinity.

Synchronous Rectifier - A circuit that emulates a diode, allowing current to pass in one direction but not the other without the losses associated with junction or Schottky devices. The circuit comprises a pass-element (most often a power MOSFET), a sense element, a sense-signal conditioner, and a driver.

- In the MOSFET application, the rectifier would have a large channel width in order to conduct large currents.

D.18 T

Thermal Equilibrium - See equilibrium.

Thermal Motion - The random motions of molecules, atoms, electrons or other subatomic particles.

- Thermal motion in a doped semiconductor averages out to zero on a macroscopic scale, and does not contribute to current transport.

Thermal Runaway - A process that is accelerated by increased temperature, in turn releasing energy that further increases temperature.

- Thermal runaway occurs in situations where an increase in temperature changes the conditions in a way that causes a further increase in temperature, often leading to a destructive result.
- It is a kind of uncontrolled positive feedback.

Thermal Voltage - The voltage produced within a *PN*-junction due to temperature.

Thermionic Emission Current - The current resulting from majority carrier electron or hole injection over the potential barrier in an MS diode.

- Thermionic emission current occurs in situations where an increase in temperature changes the conditions in a way that causes a further increase in temperature, often leading to a destructive result.
- It is a kind of uncontrolled positive feedback.

Transconductance - The trans(fer) conductance is the electrical characteristic relating the current through the output of a device to the voltage across the input of a device.

- It is often denoted as a conductance, with a subscript, m, for mutual, defined as follows:

$$g_m = \frac{\Delta I_{out}}{\Delta V_{in}} \quad (\text{D.7})$$

- For small signal models:

$$g_m = \frac{\partial i_{out}}{\partial v_{in}} \quad (\text{D.8})$$

- The transconductance for the bipolar transistor can be expressed as:

$$g_m = \frac{I_C}{V_T} \quad (\text{D.9})$$

I_C is the DC collector current at the quiescent point, and V_T is the thermal voltage ($\approx 0.259 \text{ V}$).

- The transconductance for the MOSFET can be expressed as:

$$g_m = \frac{2I_D}{V_{OV}} \quad (\text{D.10})$$

I_D is the DC drain current at the bias point, and V_{OV} is the overdrive voltage.

- For vacuum tubes, transconductance is defined as the change in the plate (anode) current divided by the corresponding change in the grid/cathode voltage, with a constant plate(anode) to cathode voltage. Defined as follows:

$$g_m = \frac{\mu}{r_p} \quad (\text{D.11})$$

μ is the gain, and r_p is the plate resistance.

- A transconductance amplifier puts out a current proportional to its input voltage.

Transresistance - The trans(fer) resistance refers to the ratio between a change of the voltage at two output points and a related change of current through two input points.

- Also denoted with a subscript, m, for mutual, defined as follows:

$$r_m = \frac{\Delta V_{out}}{\Delta I_{in}} \quad (\text{D.12})$$

- For small signal models:

$$r_m = \frac{\delta v_{out}}{\delta i_{in}} \quad (\text{D.13})$$

- A transresistance amplifier outputs a voltage proportional to its input current. The transresistance amplifier is often referred to as a transimpedance amplifier, especially by semiconductor manufacturers.
- The term for a transresistance amplifier in network analysis is current controlled voltage source (CCVS).

D.19 V

Vacuum Level - The minimum energy (typically denoted as E_0) and electron must possess to completely free itself from a material.

Varactor - Also known as a varicap diode, a varactor is a type of diode designed to exploit the voltage-dependent capacitance of a reverse-biased *PN*-junction. Varactors are used as voltage-controlled capacitors. They are commonly used in voltage-controlled oscillators, parametric amplifiers, and frequency multipliers.

Valence Band - The lower band of allowed energy states in *Si*.

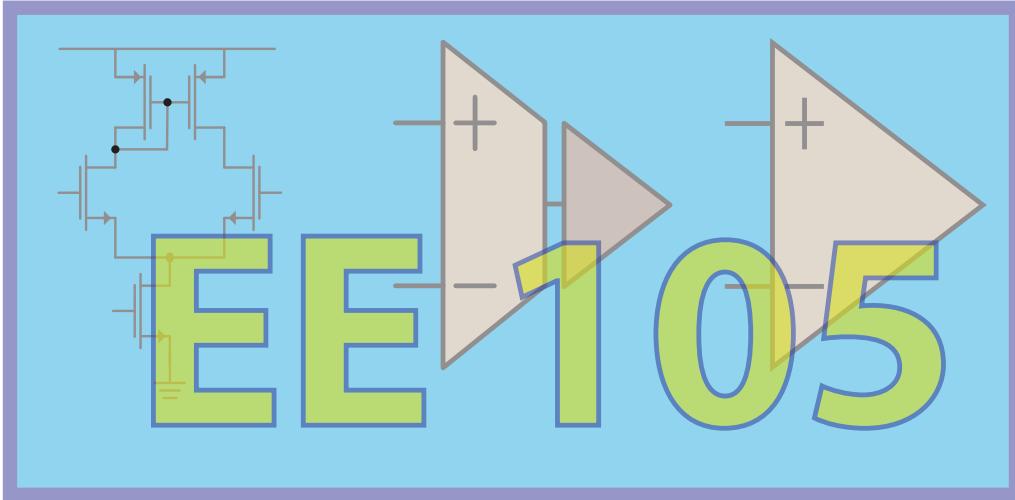
Velocity Saturation - The maximum velocity a charge carrier in a semiconductor, generally an electron, attains in the presence of very high electric fields.

- As the applied electric field increases from that point, the carrier velocity no longer increases, because the carriers lose energy through increased levels of interaction with the lattice by emitting phonons, and even photons, as soon as the carrier energy is large enough to do so.

D.20 W

Work Function - The distance from vacuum level to the Fermi level of a material.

- $\Phi_s \triangleq (E_0 - E_{C_{semiconductor}}) + (E_C - E_F) = \chi + (E_C - E_F)$ is the work function of a semiconductor. The affinity is used because the conduction band edge is not a constant in semiconductors.
- $\Phi_m \triangleq E_0 - E_{F_{metal}}$ is the work function of a metal.



Index

- AC coupling capacitors, 280
- AC output signal, 182
- Admittance, 51, 55
- Amplifier
 - AC coupled, 230
 - AC grounded, 230
 - AC signal source, 222
 - active load, 308
 - biasing, 222
 - common drain using a current mirror, 251
 - DC bias, 132, 299
 - mid-rail, 178
- BJT
 - common base, 220
 - common collector, 220
 - common emitter, 220
- body effect, 233
- bootstrapping, 282
- cascode
 - minimum operating voltage, 249
- cascode differential pair, 308
- cascode loading, 308
- clipping, 344
- common gate, 269
- common source, 176, 186, 227, 260, 262
- constant transconductance reference, 255
- current amplifier, 217
- current buffer, 227
- DC coupled, 262
- DC coupling, 279
- differential
 - offset voltage, 310
- differential amplifier
 - differential input signal, 300
- distortion, 344, 349
- electrocardiogram, 48
- four-terminal small-signal, 233
- full-power bandwidth, 348
- headroom, 222, 230, 238, 285, 292, 308
- high swing cascode, 249
- ideal, 170
- input impedance, 216, 218
- large-signal output voltage, 181
- load impedance, 238
- mid-band, 222
- mid-band gain, 222
- MOS, 176
- MOSFET
 - common drain, 220, 230
 - common gate, 220, 222
 - common gate AC model, 222
 - common gate DC operating point, 222
 - common source, 220
- multi-stage, 278
- cascode, 285
- folded cascode, 292
- level-shifting, 279

- noise, 350
 - white-noise, 350
- noise current, 351
- noise voltage, 351
- open-loop, 328
- operating point, 220, 280
- operational transconductance, 326
- output impedance, 216, 218
- output voltage swing, 344
- PMOS mirror, 250
- RF, 156, 229
- slew rate, 345
- small-signal output voltage, 181
- source follower, 232
- trans-admittance, 51
- trans-conductance amplifier, 218
- trans-impedance, 51
- trans-impedance amplifier, 218
- trans-resistance amplifier, 218
- types
 - cascode, 285
 - cascode current mirror, 248
- voltage amplifier, 217
- voltage buffer, 232
- Analog-to-Digital converter, 218
- Anode, 124
- Back-gate generator, 231
- Band model, 79
- Band-gap, 78, 79, 83
- Bandwidth, 38, 60
- Bias point, 178
- BJT
 - I-V* curves, 204
 - base current, 209
 - base resistance, 213
 - collector current, 204
 - collector saturation current, 209
 - current gain, 209
 - Ebers-Moll equations, 210
 - Ebers-Moll model, 210
 - family of curves, 204
 - forward transit time, 213
 - junction capacitances, 213
 - junctions, 200
 - base-emitter junction, 204
 - collector-base junction, 205
 - lateral fabrication, 202
 - maximizing performance, 209
 - operating point, 200
- output resistance, 213
- parasitic resistance, 202
- regions of operation
 - reverse active, 203, 205, 210
- saturation, 205
- simplified Ebers-Moll model, 212
- small-signal
 - model, 216
- terminals
 - base, 200
 - collector, 200
 - emitter, 200
- top-view layout, 202
- vertical fabrication, 201
- Bode plot, 61
 - composite, 63
- Boltzmann distribution, 78, 123
- Boltzmann's constant, 351
- Bond model, 79
- Boundary conditions, 91
- Capacitor
 - AC coupling, 220
 - bootstrapping, 267
 - bypass, 283
 - dielectric constant, 142
 - non-linear, 103
 - source capacitance, 50
- Cathode, 124
- Circuit models
 - one-port, 216
 - two-port, 216
- Common-mode, 48
- Common-mode rejection ratio, 309
- Common-mode signal, 294, 303
- Complex exponential, 25, 28, 39
- Conductance, 70
- Conduction band, 78
- Conductivity, 70, 72
- Convolutions
 - integral, 24, 34, 39
 - operator, 34
- Coupling capacitors, 259
- Covalent bond, 76
- Cramer's rule, 263
- Crystal potential, 80
- Crystal structures
 - diamond, 76
- Crystalline, 201
- Current

- continuity equation, 121
- diffusion, 89, 91, 106, 143
- drift, 91, 106
 - saturation, 122
- Current mirror, 243
 - scaling, 244
 - scaling ratio, 244
- Current sink, 250
- Current source, 250
- Current-controlled current source, 205
- Differential circuit, 294
- Differential equation
 - non-linear, 182
 - ordinary, 25
- Differential pair, 338
- Differential transconductance, 295
- Differential-mode signal, 294, 303
- Diffusion constant, 91
- Diffusion resistor, 95
- Diode
 - I-V curve, 123
 - junction capacitance, 133
 - light-emitting, 137
 - Small-signal model, 132
 - small-signal resistance, 133
- Dirac delta function, 32
 - sifting property, 33
- Direct band gap, 125, 139
- Divergence theorem, 98
- Doping, 83
 - acceptors, 85
 - degenerative, 152
 - donors, 83
- Drift current, 165
- Drift velocity, 87, 165
- E-k diagram, 139
- Early voltage, 171, 211
- Effective mass, 73, 80, 81, 87
- Eigenfunction, 24, 39
- Einstein's relation, 107
- Energy band, 145
- Equilibrium
 - thermal, 82
- Feedback, 318
 - feedback factor, 319
 - negative, 323
 - positive, 323
- Fermi-Dirac statistics, 78
- Fermions, 78
- Filters
 - amplitude, 21
 - band-pass, 44
 - high-pass, 61
 - low-pass, 20
 - phase response, 21
- Forward active region, 172
- Fourier series, 42
- Free-electron mass, 73
- Frequency response, 182
- Frequency-domain, 39, 54
- Gain
 - power, 58
 - voltage, 58
- Gain-bandwidth product, 328
- Gauss' law, 98, 113
- Gaussian distribution, 131
- Generation, 82, 147
- Geometric series, 323, 357
- Golden Rules, 320
- Holes, 81
- Hysteresis, 323
- IGFET, 158
- Impedance, 51, 55
 - Input impedance, 57
- Impulse response function, 34
- Indirect band gap, 125, 139
- Inductor
 - used as a choke, 221
- Interference
 - rejection, 323
- Intrinsic carrier concentration, 82
- JFET, 158
- Laplace domain, 40, 52
- Laplace transform, 39, 319
- Lattice vibrations, 78
- Law of mass action, 83, 107, 124
- Linear systems, 24
 - frequency response, 26
 - general response, 34
 - time invariant, 24
 - time shifting, 24
- Loop gain, 322
- Louis de Broglie, 77

- Maclaurin series, 358
 Mean free-path, 90
 Mean free-time, 72
 MESFET, 158
 Miller capacitor, 267
 Miller effect, 266
 Miller's theorem, 264, 281, 286
 - Common drain analysis, 267
 - Common source analysis, 265
 MIM capacitor, 101
 Mobility, 72, 87, 163, 165
 MOS capacitor, 142
 - body, 142
 - built-in potential difference, 143
 - flat-band voltage, 145
 - gate, 142
 - regions of operation
 - accumulation, 146, 160
 - depletion, 146, 160
 - equilibrium, 145
 - flat-band, 145
 - inversion, 147, 160
 - surface potential, 148
 - threshold voltage, 148- types
 - N*-type, 142, 158
 - P*-type, 142, 158

 MOSFET, 155, 156
 - $I_D - V_D$ family of curves, 162
 - $I_D - V_G$ curve, 161
 - as a voltage-dependent resistor, 163
 - aspect ratio, 163, 165
 - average channel charge, 165
 - back-gate transconductance, 196
 - capacitance grading coefficient, 194
 - capacitances, 192
 - drain-bulk, 193
 - gate-drain, 193
 - gate-source, 192
 - source-bulk, 193
 - channel charge, 164
 - channel conductance, 163
 - channel length, 156
 - channel length modulation, 170
 - channel potential, 165
 - device parameter, 196
 - device width, 156
 - drift velocity, 163
 - inversion charge, 164, 166
- inversion charge density, 163
 junction depth, 156
 long-channel, 171
 low-frequency response, 259
 overdrive voltage, 163, 289
 parasitic overlap capacitance, 192
 pass-band response, 259
 regions of operation
 - linear, 163
 - pinch-off, 168
 - saturation, 162
 - sub-threshold, 161, 172
 - triode, 162- saturation, 176
- saturation current, 170
- saturation voltage, 168
- short-channel effects, 189
- small-signal
 - complete model, 197
 - current, 185
 - current with body bias, 197
 - model, 216, 259
 - parameters, 185
- terminals
 - body, 156
 - drain, 156
 - gate, 156
 - source, 156
- threshold voltage, 161, 165
- types
 - N*-type, 158
 - P*-type, 158
 - C-type, 159
 - diode connected, 241, 290, 333
 - well-regions, 159
- Network theorems, 56
 Niels Bohr, 77
 Nodal analysis, 56
 Normal distribution, 131
 Norton equivalent, 216, 343, 351
- Ohm's law, 54, 70, 96
 Ohmic contact, 116, 124, 144
 Op-amp
 - closed-loop, 322
 - gain-bandwidth product, 342
 - ideal, 317
 - input offset current, 339
 - integrator, 326

- linear imperfections, 324
- non-linear imperfections, 324
- offset voltage, 338
- open-loop, 322
- Open circuit time constants, 272
 - dominant pole approximation, 273
- Operating point, 132, 156, 176, 178, 180, 185
- Optical photon, 125
- Orthonormal basis, 24
- Oscillators, 334
- Output conductance, 182
- Output resistance, 180
- Output swing, 178

- Parasitic capacitance, 95, 101, 189, 192, 218
- Passive circuit, 58
- Pauli exclusion principle, 75
- Phase margin, 335
- Phasors, 25, 53, 65
- Phonon, 73, 78, 125
 - scattering, 73
- Photodiode, 50
- Photolithography, 94
- Photon, 78
 - absorption coefficient, 139
- Photoresist, 94
- Photovoltaic cell, 134
- PN-junction, 95, 144
 - built-in potential, 114, 129
 - charge density, 111
 - concentration distribution, 130
 - contact potential, 116
 - continuity equation, 135
 - depletion approximation, 112
 - diffusion length, 127
 - equilibrium, 108
 - I-V curve, 123
 - inside of MOSFETs, 189
 - low-level injection, 129
 - majority carriers, 121
 - minority carriers, 121
 - concentration, 129
 - lifetime, 126
 - reverse-bias, 95, 103, 117, 159
 - saturation current, 122
 - solar cell, 134
 - transition region, 111
- Poisson's equation, 98
- Polycrystalline, 95
- Port models

- one-port, 51
- two-port, 50, 280
- Power
 - average, 64
 - factor, 65
 - instantaneous, 64
- Pulse response, 38

- Quality/Q factor, 58
- Quantum mechanics, 73, 139
- Quasi-static, 151
- Quiescent bias point, 301
- Quiescent point, 178

- Recombination, 82, 147
- Rectifying contact, 144
- Resistance, 70
- Resistivity, 70
- Resonant frequency, 57
- Rest mass, 73

- S-domain, 39
- Semiconductors, 74
 - compound, 139
- Sheet resistance, 96, 163
- Signal-to-noise ratio, 351
- Single-ended circuit, 294
- Small-signal
 - analysis, 175, 182
 - capacitance, 103, 118, 151
 - model, 242, 246
- Source degeneration, 247
- Space-charge region, 115
- Spectral response, 38
- Stability, 334
- Steady-state response, 30
- Step input, 37
- Step response, 37

- Taylor expansion, 70, 90, 103, 118, 132, 182, 349
- Taylor series, 358
- Thermal equilibrium, 106, 111, 143
- Thermal generation, 151
- Thermal velocity, 87, 91
- Thermal voltage, 115
- Thévenin equivalent, 56, 216, 247, 351
- Time-domain, 39
- Transconductance, 182
- Transconductor
 - ideal, 325

- Transfer function, 27, 48, 58, 258
 - canonical form, 52
 - DC gain form, 52
 - poles, 53, 258
 - non-dominant, 334
 - rational form, 52
 - standard form, 52, 59
 - zeros, 53, 59, 258
- Transformer, 50
 - impedance, 58
- Transient response, 30
- Transmission-line impedances, 229
- Unit rectangular function, 31
- Unity-gain frequency, 269, 271, 329
- Vacuum tube, 320
- Valence band, 78
- Varactor, 119
- Voltage divider, 56, 58
- Voltage swing, 289, 308
- Voltage-controlled current source, 170, 176,
200, 333
- Voltage-controlled voltage source, 324
- Waveform harmonics, 349
- Work function, 116