ANALOG FRONT-END CIRCUITS FOR MASSIVE PARALLEL 3-D NEURAL MICROSYSTEMS

by

Khaled M. Alashmouny

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Electrical Engineering)
in The University of Michigan
2013

Doctoral Committee:

Professor Euisik Yoon, Chair
Associate Professor Joshua Berke
Professor Khalil Najafi
Assistant Professor David Wentzloff
Professor Kensall Wise

All Praise belongs to Allah the Lord of mankind.

To my father Mohamed, my mother Laila, and my lovely wife Manar. You always inspire me.

ACKNOWLEDGMENTS

First and foremost, it was a pleasure to work with my research advisor, Professor Euisik Yoon. It was a great experience as I learned and enjoyed a lot his challenges and guidance throughout my Ph.D. I still remember the joy I had when I started working with Prof. Yoon in deeply understanding the basic electrical definition of a ground and trying to predict the behavior of signal transmission within the brain. What I learned the most from him was the exploration of the different aspects in a microsystem design although I end up focusing more onto circuit design. Prof. Yoon always encouraged me to be a proactive person and this has reflected into a lot of the contributions and volunteering activities I decided to make for our lab, the University, and the society.

I would like to thank my committee members Dr. Kensall Wise, Dr. Khalil Najafi, Dr. Joshua Berke, and Dr. David Wentzloff for their time and their great suggestions and feedback. Special thanks to Dr. Wentzloff, who was always very welcoming whenever I needed an advice. Special thanks to Dr. Berke inspired me through different discussions and advises to look further into the neuroscience viewpoint. His lab members, especially Jeff Pettibone, were very welcoming to answer the questions I had and provided very informative data. A special thanks to Dr. Wise who is really a true model for me; he was always open to discuss entrepreneurial stuff with me although he always had a busy schedule. Thanks to Dr. Najafi for his suggestions, inspiration, and encouragement.

I am also grateful for the center of Wireless Integrated Microsystem (WIMS) at the University of Michigan, the National Science Foundation (NSF), and the Rackham Graduate School for the financial support.

I would like thank all my colleagues in Yoon's group, who made it fun to work in an inspiring environment. A special thanks to my friend, Seok-Jun Park, who was always helpful and generous. Another special thanks to Sun-Il Chang; he is a great friend and partner in a lot of circuit ideas. Special thanks to the whole neural circuits' team: Sung-

Yun, Adam, and Yu-Ju. I was inspired by them and learned a lot from our discussions. Another special thanks to Jihyun, who was very helpful at very critical times.

I would also like to thank many people in the administration of EECS, WIMS, and SSEL: Karen Richardson, Beth Stalnaker, Melanie Caughey, Deb Swartz, Mary McCune, and many others that always made my job easier and smoother. Thank you also should go to a lot of friends I made in SSEL and WIMS.

I also would like to thank my great teachers Michael Flynn, Dennis Sylvester, and David Wentzloff for all the things they taught me in circuit design. Thanks also to my entrepreneurship mentors Dr. Adriens and Dr. Faley who encouraged me to explore a great deal of the business side in Healthcare microsystems.

During my stay in Ann Arbor, I was blessed with a lot of dear friends, whom I will never forget. In fact I wish to live in Ann Arbor because of them, but we are all leaving at some point. A deep gratitude goes to my two dearest friends in Ann Arbor Amjad Abo-jbara and Samer Kadous, who have been brothers to me. There are very few people on Earth like these great guys. A deep thanks to my best friends Ahmad Almuhtady, Ahmed AlSabbagh, Ahmed Hassan, Ahmed Mady, Ihab Ismael, Khaled Aljanaida, Mahmoud Azzouny, Amr Alaa, Mohamed El-Sayed, A. Mayhoub, Mousa Afaneh, and Mahmoud Elkasabi. It is very hard to express how all of them affected my life.

After Allah – my Lord- there is no one worthy of thanks like my parents Mohamed and Laila, who have been always supportive and made life very comfortable for me. There is nothing to say to express my gratitude to my parents. I would sacrifice my life for them and what they have done for me. Thanks to my older brother Tarek who taught me a lot in life and has been also a great friend. Thanks to my younger brothers Osama, and Amr, and my sister Eman for their continuous encouragement.

I save my last and greatest thanks to my dear wife Manar. I know she suffered a lot with me by leaving her parents and family in Egypt for the sake of supporting me without mentioning it or complaining about it. She has been a great wife, sister, mentor, and family for me and she will always be.

Finally, I would like to thank my Lord Allah for all the blessings and bounties He has given me. I always pray that He makes me beneficial to all mankind and to accept my good deeds sincerely to Him.

TABLE OF CONTENTS

DEDICATION ii

ACKNOWLEDGMENTS iii

LIST OF FIGURES xi

LIST OF TABLES xxi

ABSTRACT xxii

CHAPTER 1: INTRODUCTION 1

1.1 History of Electrical Recording of Brian Activity 2

1.2 Integrated Circuits and MEMS for Extracellular Recording 6

1.3 Challenges and Trends of Neural Integrated Microsystems (NIMS) 21

1.3.1 User Segmentation 21

1.3.2 Challenges and Trends for Highly Complicated Neuroscientific Studies 23

1.4 Research Objectives and Overview 27

CHAPTER 2: CONSTRAINTS AND DESIGN OF AN ANALOG FRONT-END MODULE FOR HYBRID INTEGRATION WITH NEURAL PROBES 29

2.1 NIMS IC Requirements and Figures of Merit 30

2.1.1 Power Consumption Effect 30

2.1.2 Integration and Area Consumption Effect 31

2.1.3 Figures of Merit for Neural Recording Integrated Circuits 32

2.2 A 0.9V Low-Power 16-Channels Modular Analog Front-End 37

2.3 A Sub-1μW Low-Noise Amplifier Operating in Moderate Inversion 40

2.3.1 A Review of Moderate Inversion Region and Operation Prospective 40

2.3.2 Design of Moderate-Inversion Two-Stage OTA ... 44

2.3.3 OTA Frequency Response Analysis: Gain, Phase, and Noise ... 46

2.3.4 Simulation Results ... 51

2.4 A 124nW Band-Pass Filter with 6-Bit Tunable Frequency Corners ... 53

2.5 Time-Division Multiplexing ... 56

2.6 A 8.2µW 3-Bit Programmable Gain Amplifier for a 16:1 Time-Division Multiplexed Input ... 59

2.7 Asynchronous 320kS/s 9-Bit Successive Approximation Analog-to-Digital Converter with Rail-to-Rail Dynamic Comparator and Power-Scalable Sampling Operation ... 62

2.7.1 The Overall Operation ... 63

2.7.2 Track and Hold Circuit (THA) ... 64

2.7.3 9-bit Capacitive DAC with a Compensation Capacitor ... 65

2.7.4 Dynamic Comparator with Rail-to-Rail Input-Common-Mode ... 68

2.7.5 Asynchronous Logic and Interface ... 74

2.7.6 Parallel-to-Serial Converter and Data Acquisition ... 75

2.7.7 Power Scalable Sampling Operation ... 76

2.8 Measurement Results and Performance Comparison for AHI-16 ... 77

2.8.1 Overall System Measurement ... 77

2.8.2 Low-Noise Amplifier ... 79

2.8.3 Band-Pass Filter ... 83

2.8.4 Programmable Gain Amplifier ... 83

2.8.5 Analog-to-Digital Converter ... 86

2.8.6 Performance Comparison ... 89

2.9 Conclusions ... 91

CHAPTER 3: A 128-CHANNEL ANALOG FRONT-END WITH EMBEDDED PROCESSING 92

3.1 Design Approaches to Overcome the DC Drift of Electrode-Circuit Interface 94

3.1.1 Review of Previous Approaches 94

3.1.2 Limitations of Capacitive Coupling Approach on Future Scaling of Neural Recording Channels 96

3.1.3 Using Advanced Technologies and Mostly-Digital Design Approaches 97

3.1.4 The tradeoff of Technology Scaling for Neural Recording Applications 99

3.1.5 Potential of Digitally-Assisted Analog IC Design 100

3.2 Signal Processing Algorithm for Digitally-Assisted Analog Front-End 102

3.2.1 Analog Compression Algorithm 102

3.2.2 Algorithm Validation and Accumulative Error Reset 106

3.2.3 Algorithm Application on Neural Signals 109

3.2.4 An Estimation of Power and Area Improvement 114

3.3 0.5V 128-Channel Analog Front-End Architecture 117

3.3.1 Overall System Architecture 117

3.3.2 Low-Voltage Current-Mode Mixed-Signal Design 118

3.4 0.5 V DC-Coupled Low-Noise Amplifier 120

3.5 Current-Mode All-Pass Filter as an Analog Delay Line 123

3.6 Current-Mode Analog Compression Circuit 128

3.7 4-bit Current-Mode Programmable-Gain Amplifier 130

3.8 6-bit Current-Mode FLASH Analog-to-Digital Converter 132

3.8.1 Current-Mode Comparator 133

3.8.2 Thermometer-to-Binary Encoder 134

3.9 Data Acquisition and Parallel-to-Serial Conversion 136

3.10 Measurement Results and Performance Comparison for AHI-128 139

3.10.1 Overall System 139

3.10.2 Low-Noise Amplifier 141

3.10.3 Analog Compression Block.. 142

3.10.4 Programmable-Gain Current Amplifier 143

3.10.5 Analog-to-Digital Converter 143

3.10.6 Data Measurement and Reconstruction...................................... 144

3.10.7 Performance Comparison 147

3.11 Conclusion 149

CHAPTER 4: MONOLITHIC ACTIVE NEURAL PROBES: TOWARDS THE NEXT DECADE OF MASSIVE PARALLEL RECORDING 150

4.1 Prospective of Monolithic Integration of MEMS and CMOS for Neural Interfaces 151

4.1.1 Why Monolithic Integration? 151

4.1.2 Why CMOS Circuit Integration On-The-Shank?....................................... 152

4.1.3 Post-Processing of CMOS to form MEMS Structures................................ 153

4.2 Modular Probe Design in 150nm SOI CMOS 156

4.2.1 A 150nm FD-SOI CMOS Process... 156

4.2.2 A Shank-based Module 156

4.2.3 System Architecture for a 128-Channels Neural Probe 158

4.3 A 0.5V Subthreshold-optimized 128-Channels Analog Front-End Module 160

4.3.1 Low-Noise Amplifier 160

4.3.2 Band-Pass Filter and Buffer 161

4.3.3 Programmable Gain Amplifier.. 162

4.3.4 Asynchronous 6-Bit SAR ADC 163

4.4 Measurement Results and Performance Comparison 166

4.4.1 Active Probe Layout.. 166

4.4.2 Low-Noise Amplifier 168

4.4.3 Band-Pass Filter 169

4.4.4 Analog-to-Digital Converter 170

4.4.5 Performance Comparison 172

4.5 Conclusions 174

CHAPTER 5: SUMMARY, CONTRIBUTIONS AND FUTURE WORK 175

5.1 Summary 175

5.2 Contributions 176

5.3 Future Work 178

APPENDIX A: IBCOM (INTRA-BRAIN COMMUNICATION)- A NEW METHOD OF SIGNAL TRANSMISSION FOR MINIATURIZED NEURAL RECORDING/STIMULATING SYSTEMS 181

A.1 Introduction 182

A.2 Intra-Brain Communication (IBCOM) 184

A.2.1 Signal Transmission Through Brain 185

A.2.2 IBCOM Transmission of BFSK Modulated Neural Signal 187

A.2.3 Optimization of IBCOM and Simultaneous Neural Signal Recording 189

A.2.4 On the potential for extra-brain receiving sites 191

A.3 Miniaturized µ-IBCOM CMOS Chip 192

A.3.1 ROM Storage and POR Circuitry 194

A.3.2 BFSK Modulator Circuitry 194

A.3.3 Voltage-to-Current Converter Circuitry 194

A.3.4 µ-IBCOM CMOS Prototype Chips and Packaging 196

A.4 Measurement Results 198

A.4.1 Experiments Procedure 198

A.4.2 µ-IBCOM Output Characteristics 198

A.4.3 μ-IBCOM Experiment Results in Rat Brain 199

A.5 Conclusion 201

REFERENCES 203

LIST OF FIGURES

Figure 1.1. The neuron: the unit cell inside the brain [Courtesy of Wikimedia Commons] 1

Figure 1.2. Electrical recording methods for brain activity and their sensing locations [4]3

Figure 1.3. Silicon neural probe developed in 1969 [21, 23] 6

Figure 1.4. Basic structure of the Michigan Probe [25, 26]... 7

Figure 1.5. Utah electrode array [37].. 9

Figure 1.6. Schematic and Implementation of three-dimensional electrode array [41]... 10

Figure 1.7. Three-dimensional active probe array [44] 11

Figure 1.8. 2D and 3D probe array [26,45].. 11

Figure 1.9. 64-site active probe [48]... 12

Figure 1.10. Neural Microsystem with 256 sites and 32 channels [51]............................ 13

Figure 1.11. Utah neural interface assembly and circuit boards for wireless power and data link [57,58] 13

Figure 1.12. A 64-site 3D neural probe (left), and integration with analog front-end and parylene cable [60] 14

Figure 1.13. Integration of the microsystem using parylene overlay cable approach [60]14

Figure 1.14. Two-stage neural microsystem proposed in [62] 15

Figure 1.15. 256-channel circuitry and 64-sites bonding to Utah electrode array [63].... 16

Figure 1.16. Neural microsystem pursued at the University of Michigan [65]................ 17

Figure 1.17. Block diagram and circuit microphotograph of the system described in [66] 18

Figure 1.18. Monolithic active neural probe including 188 electrodes and site selection [68] 19

Figure 1.19. NIMS user segmentation based on needs 21

Figure 1.20. The main objective and use need of the three major segments in neural recording 22

Figure 1.21. Neurons at the vicinity of recording electrode tip [2] 23

Figure 1.22. Parameters and metrics to asses NIMS 25

Figure 1.23. 1024 channels NIMS "Brainavigator" 25

Figure 2.1. 128-sites neural probe with hybrid-integrated analog front-end 31

Figure 2.2. Performance chart of the implemented AFE. The chart plots noise-area-product versus power-per-bandwidth and compares the state-of-the-art neural recording circuits reported up to date. 34

Figure 2.3. Block diagram of analog front-end module with the illustration of integrating in a 3D array system. The front-end module includes on-chip reference generator, site selection (128 to16), and 16 preamplifiers that are multiplexed into one ADC. Data acquisition interface is asynchronous and uses only four pads, and there are extra two pads assigned for clock and loading control parameters to the chip. When multiple 2D modules are assemble in the shown 3D array, only EOC and SDO need separate connections for each module while the other pads are shared with the rest of modules................. 37

Figure 2.4. Block diagram of the 128-to-16 site selection.. 39

Figure 2.5. Design tradeoffs in weak, moderate, and strong inversion regions. It shows that moderate inversion region has a strong potential for circuit optimization in low voltage operation. 42

Figure 2.6. Schematic of a moderate-inversion low-noise amplifier. All the transistors are biased in moderate inversion except for a differential pair. 44

Figure 2.7. Open-loop LNA showing poles and zeros... 47

Figure 2.8. Low-noise amplifier open-loop gain and phase response 51

Figure 2.9. Low-noise amplifier closed-loop transient and frequency response 52

Figure 2.10. Band-pass filter schematic with 6-bit tunable corners 53

Figure 2.11. Simulation of BPF gain response showing large tuning range of frequency corners 54

Figure 2.12. 16-channel time-division multiplexer controlled by *SOC* signal from the ADC.. 56

Figure 2.13. True single-phase clock edge-triggered D flip flop..................................... 57

Figure 2.14. Resistance of transmission gate configuration versus input and supply voltages.. 58

Figure 2.15. Schematic of 3-bit programmable gain amplifier.. 59

Figure 2.16. Simulation of 3-bit programmable gain amplifier....................................... 61

Figure 2.17. 9-bit Asynchronous SAR ADC 62

Figure 2.18. Bootstrapped Track and Hold Circuit.. 64

Figure 2.19. Operation of Track and Hold Circuit... 64

Figure 2.20. A 9-bit DAC with a split capacitor banks connected with a coupling capacitor 65

Figure 2.21. The capacitor mismatch characterized by the 0.25μm CMOS process and three Monte Carlo simulations of DNL when the DAC unit capacitor is 49fF, 81fF, and 121fF 67

Figure 2.22. A rail-to-rail dynamic comparator using feed-forward for the input-common-mode tracked by the DAC output.. 68

Figure 2.23. Effective functional comparator at the onset of regenerative operation phase under low common-mode ($V_{DAC} < 0.5V_{DD}$) and high common-mode ($V_{DAC} > 0.5V_{DD}$) 69

Figure 2.24. Comparator small signal approximation for the NMOS operating mode 70

Figure 2.25. Simulation results of the comparator transient response for two different inputs V_{in1} and V_{in2} varying by an order of magnitude 71

Figure 2.26. Simulation results of the over drive test 72

Figure 2.27. Asynchronous and successive approximation control logic 74

Figure 2.28. Data output after acquisition 76

Figure 2.29. Die photo of the fabricated chip and a 128-sites Michigan probe array. It also shows benchtop testing results using synthesized neural signals applied to the shorted inputs (middle), and the recovered input-referred AFE output after demultiplexing and DAC operation using an off-chip FPGA. 77

Figure 2.30. Measurement Setup 78

Figure 2.31. The die photograph of the fabricated LNA chip 79

Figure 2.32. Measured LNA gain response from 30 chips/2wafers at 990nW power consumption 79

Figure 2.33. Measured LNA input-referred noise at 990nW power consumption 80

Figure 2.34. Measured LNA gain response from 30 chips/2wafers at 1700nW power consumption 80

Figure 2.35. Measured LNA noise response at 1700nW power consumption 81

Figure 2.36. Measured LNA common-mode rejection ratio 81

Figure 2.37. Measured LNA power-supply rejection ratio 81

Figure 2.38. Measured gain response of the BPF under different tunings of frequency corners, and die photograph of the fabricated chip 83

Figure 2.39. PGA measured gain and transient response, and chip photograph 85

Figure 2.40. ADC measured frequency response, differential, and integrated linearity .. 86

Figure 2.41. ADC measured power scalable sampling and figure of merit. It shows also the estimated FOM along with sampling frequency, and SNDR at different supply voltage 88

Figure 2.42. Performance chart of the implemented AFE. 90

Figure 3.1. Input interface circuit of pre-amplifier in [29] 94

Figure 3.2. The preamplifier with DC feedback as proposed in [34] 95

Figure 3.3. Low-noise amplifier using capacitive coupling 96

Figure 3.4. Digitally-intensive technique for neural recording using 65nm process 98

Figure 3.5. CMOS technology scaling and implementation suggestions (source: a tutorial lecture provided by D. Sylvester at the University of Michigan) 99

Figure 3.6. Analog circuit design trade-offs [98] 100

Figure 3.7. Parameter definition for delta algorithm 103

Figure 3.8. Amplitude reduction for a delayed-differential analog input versus delay value 104

Figure 3.9. Original (blue), delayed (red), and the resulting delta signal (green) at certain bias according to a 0.5V supply operation 107

Figure 3.10. ADC with accumulative error reset (AER) 108

Figure 3.11. The discrete and digital signal when sampled at OSR of 10 108

Figure 3.12. The sigma function used for signal reconstruction 108

Figure 3.13. The reconstructed and original signal showing the LSB error reduction when averaging every 10 samples 109

Figure 3.14. Neural signal sampled at 31.25 kHz and 16-bit resolution 111

Figure 3.15. Neural signal after delta function 111

Figure 3.16. 6-bit digitized delta signal 112

Figure 3.17. Reconstructed signal without accumulated error reset. Signal cannot be seen because it is superimposed on a large low-frequency drift 112

Figure 3.18. A first order elliptic digital low-pass filter for detecting the offset drift below 1 Hz 113

Figure 3.19. The outcome of the reconstruction algorithm including the elliptic filter 113

Figure 3.20. Frequency Spectrum of original, reconstructed, and error signals 114

Figure 3.21. OSR versus PRF plot showing that above 80 oversampling ratio there is not power advantage 115

Figure 3.22. 128-channel analog front-end with embedded processing 117

Figure 3.23. Current-mode analog front-end 118

Figure 3.24. DC-coupling approach adapted from [29] 120

Figure 3.25. Schematic of the 0.5V low-noise amplifier using current source load and source degeneration to increase the linearity and providing more offset rejection 121

Figure 3.26. Simulated gain and phase responses of the low-noise amplifier including and excluding the high-pass filter (from the electrode model) 122

Figure 3.27. Schematic of a current-mode all-pass filter: it consists of a current inverter followed by a low-pass filter (adopted from [107]) 124

Figure 3.28. Simulated delay value versus tuning voltage showing the exponential relationship 126

Figure 3.29. Simulation of all-pass filter transfer function.. 126

Figure 3.30. Transient response of the all-pass filter showing delays of up to 17 µs between the input (solid line) and the outputs (dashed lines) 127

Figure 3.31. Simulation of the Delta Operation.. 127

Figure 3.32. Schematic of the overall analog compression circuit. 128

Figure 3.33. Simulation of the overall analog compression circuit. 129

Figure 3.34. Schematic of the 4-bit programmable-gain amplifier. 130

Figure 3.35. Simulation of the 4-bit programmable-gain amplifier............................... 131

Figure 3.36. Architecture of the 6-bit current-mode FLASH ADC............................... 132

Figure 3.37. Schematic of the current-mode comparator. 133

Figure 3.38. Schematic of a thermometer-to-Binary Encoder (8:3) used to construct the 64:6 encoder. The schematic shows the construction of 16:4 T-B encoder from two 8:3 encoders and three 2:1 digital multiplexers. 134

Figure 3.39. Simulation of the 6-bit analog-to-digital converter 135

Figure 3.40. Block diagram of digital circuits used for acquisition. 136

Figure 3.41. Schematic of level shifter 136

Figure 3.42. Schematic of 192-bit latch that is repeated four times. It also shows the unit flip-flop used inside the latch, which is basically a true signal phase clock dynamic flip-flop. 137

Figure 3.43. Schematic of the parallel-to-serial converter... 138

Figure 3.44. Timing diagram of data acquisition... 138

Figure 3.45. Die photo of the fabricated chip. 139

Figure 3.46. A zoom-in view of die photo showing different components. 140

Figure 3.47. Measurement Setup 140

Figure 3.48. Measured LNA gain response 141

Figure 3.49. Measured LNA noise response.. 141

Figure 3.50. Measured analog compression gain and phase response........................... 142

Figure 3.51. Measured current-gain of the programmable-gain amplifier 143

Figure 3.52. Measured ADC spectrum and transient output 144

Figure 3.53. The concept of data reconstruction.. 145

Figure 3.54. Data reconstruction shows the input voltage, sampled output current, and the reconstructed output current. 146

Figure 3.55. Reconstructed output (input-referred) superimposed on the original input. 146

Figure 3.56. Performance chart of the implemented AFE. 148

Figure 4.1. Michigan Neural Probe.. 152

Figure 4.2. A modular shank architecture 157

Figure 4.3. A 128-channels circuit architecture for the monolithic active neural probe 158

Figure 4.4. Schematic of the subthreshold low-noise amplifier 160

Figure 4.5. Transient and gain response of LNA 161

Figure 4.6. Schematic of the band-pass filter 162

Figure 4.7. Transient and gain response of BPF 162

Figure 4.8. Schematic of the 3-bit programmable-gain amplifer 163

Figure 4.9. Programmable gain response of PGA 163

Figure 4.10. Schematic of 6-bit analog-to-digital converter 164

Figure 4.11. ADC simulation results 165

Figure 4.12. Full chip microphotograph included in a larger die to improve the post-processing 167

Figure 4.13. Chip microphotograph showing the CMOS circuits on shanks and back-end 167

Figure 4.14. Measurement Setup 168

Figure 4.15. Measured LNA gain response 168

Figure 4.16. Measured LNA input-referred noise 169

Figure 4.17. Measured gain response of the BPF under different tunings of frequency corners, and die photograph of it 169

Figure 4.18. ADC measured frequency response, differential, and integrated linearity 171

Figure 4.19. Performance chart of the implemented AFE. 173

Figure A.1. Conceptual diagram of a massively-distributed neural recording/stimulating system: miniaturized neural probes are massively distributed in the brain to record/stimulate thousands of neurons simultaneously. They communicate wirelessly to an external analysis station via the way-station fully implanted on the brain surface. 184

Figure A.2. Signal transmission through rat brain: (a) Experimental setup, and (b) IBCOM average current loss over 100kHz to 50MHz for 5 mm and 15 mm distance, respectively. 186

Figure A.3. Experimental setup for two-channel IBCOM to demonstrate the transmission of modulated neural signals through brain. The two modulated neural signals are transmitted simultaneously from the two electrodes and received in the single receiver electrode, then demodulated to retrieve the original signals. During the signal transmission neural signal is monitored in a separate electrode to inspect if the IBCOM signals affect normal neuron activities. 187

Figure A.4. Measured signals at each steps during IBCOM transmission through rat brain: (a) Two original neural signals prerecorded, (b) BFSK-modulation of two original signals, (c) Signal acquired at the receiving electrode, and (d) Two demodulated neural signals after digital-to-analog conversion. 188

Figure A.5. Recording neural signals during IBCOM transmission. It shows two distinctive peaks when switching IBCOM signals. However, normal neural signals have been recorded between the two peaks, suggesting IBCOM does not interfere neural activities. 190

Figure A.6. (a) Shaping of IBCOM transmission signals, (b) Recorded neural signal while transmitting the shaped IBCOM signals. No abnormal distinctive peaks are observed. 191

Figure A.7. Recorded neural signals from the *same brain site* (a) Before, (b) After, and (c) During IBCOM Application. 192

Figure A.8. (a) Block Diagram for µ-IBCOM1 CMOS Chip, (b) Neural Signals stored in µ-IBCOM1 (left) and µ-IBCOM2 (right). 193

Figure A.9. (a) BFSK Modulator working at 100/200 kHz according to digital input voltage, (b) Voltage-to-Current Converter: 0 – 2.5V is converted to around ±80µA. 195

Figure A.10. (a) Die photograph for four µ-IBCOM systems and test patterns, (b) µ-IBCOM packaging for full implantation in rat brain. 197

Figure A.11. Measured BFSK Modulators waveform: (a) µ-IBCOM1 modulated at 100/200 kHz , and (b) the µ-IBCOM2 modulated at 300/400 kHz............ 199

Figure A.12. IBCOM output waveforms after charge-balanced V/I conversion: (a) Output current of µ-IBCOM1, (b) Output current of µ-IBCOM2, (c) FFT of µ-IBCOM1 signal, and (d) FFT of µ-IBCOM2 signal 200

Figure A.13. Measured waveforms at each step of retrieval during the rat brain experiment: (a) Signal monitored at the receiving microelectrode (This is a superimposed signal of two output signals sent from µ-IBCOM1 and µ-IBCOM2), (b) FFT of the received signal showing four peaks at the corresponding modulated frequencies, (c) Demodulated signal after being separated by 100-200kHz BPF, (d) Demodulated signal after being separated by 300-400kHz BPF, (e) and (f) Fully-retrieved neural signals sent from µ-IBCOM1 and µ-IBCOM2 after 5-bit DAC of the signals shown in (c) and (d), respectively.. 201

LIST OF TABLES

Table 1.1 Challenges of Neural Integrated Microsystems 24

Table 1.2. State-of-the-art circuit achievements 26

Table 2.1. Equations for strong and weak inversion 43

Table 2.2. Design parameters for the proposed LNA 45

Table 2.3. Design parameters for the proposed BPF 53

Table 2.4. Design parameters for the proposed PGA 60

Table 2.5. Analysis of gain and phase responses of the PGA 60

Table 2.6. Comparator gain and metastability probability for three different common-mode ranges 72

Table 2.7. LNA performance summary and comparison with state-of-the-art LNAs 82

Table 2.8. Measured performance and comparison with state-of-the-art systems 89

Table 3.1. Parameter definition for the delta algorithm 102

Table 3.2. The approaches and effects of using the architecture in Figure 3.22 117

Table 3.3. Measured performance and comparison with previous version 147

Table 4.1. Measured performance and comparison with previous version 172

ABSTRACT

ANALOG FRONT-END CIRCUITS FOR MASSIVE PARALLEL 3-D NEURAL MICROSYSTEMS

by

Khaled M. Alashmouny

Chair: Euisik Yoon

Understanding dynamics of the brain has tremendously improved due to the progress in neural recording techniques over the past five decades. The number of simultaneously recorded channels has actually doubled every 7 years, which implies that a recording system with a few thousand channels should be available in the next two decades. Nonetheless, a leap in the number of simultaneous channels has remained an unmet need due to many limitations, especially in the front-end recording integrated circuits (IC).

This research has focused on increasing the number of simultaneously recorded channels and providing modular design approaches to improve the integration and expansion of 3-D recording microsystems. Three analog front-ends (AFE) have been developed using extremely low-power and small-area circuit techniques on both the circuit and system levels. The three prototypes have investigated some critical circuit challenges in power, area, interface, and modularity.

The first AFE (16-channels) has optimized energy efficiency using techniques such as moderate inversion, minimized asynchronous interface for data acquisition, power-scalable sampling operation, and a wide configuration range of gain and bandwidth.

Circuits in this part were designed in a 0.25µm CMOS process using a 0.9-V single supply and feature a power consumption of 4µW/channel and an energy-area efficiency of $7.51 \mathrm{x} 10^{15}$ in units of $J^{-1} V_{rms}{}^{-1} mm^{-2}$.

The second AFE (128-channels) provides the next level of scaling using dc-coupled analog compression techniques to reject the electrode offset and reduce the implementation area further. Signal processing techniques were also explored to transfer some computational power outside the brain. Circuits in this part were designed in a 180nm CMOS process using a 0.5-V single supply and feature a power consumption of 2.5µW/channel, and energy-area efficiency of $30.2 \mathrm{x} 10^{15}$ J^{-1} $V_{rms}{}^{-1} mm^{-2}$.

The last AFE (128-channels) shows another leap in neural recording using monolithic integration of recording circuits on the shanks of neural probes. Monolithic integration may be the most effective approach to allow simultaneous recording of more than 1,024 channels. The probe and circuits in this part were designed in a 150 nm SOI CMOS process using a 0.5-V single supply and feature a power consumption of only 1.4µW/channel and energy-area efficiency of $36.4 \mathrm{x} 10^{15}$ $J^{-1} V_{rms}{}^{-1} mm^{-2}$, which is the highest reported efficiency to date.

CHAPTER 1

INTRODUCTION

The brain is an interesting organ that has a basic function of keeping the body alive, which is yet complicated and challenging to sustain. It controls breathing, heart rate, blood pressure, sleep cycles, emotions, thoughts, and desires. Inside the human brain, there are tens of billions of neurons (units of the brain, Figure 1.1) linked with thousands of connections each to form the circuitry of the most energy-efficient parallel processing unit known to date. Understanding how the brain works has been a major research area since the late 1800s. This was the time when science and technology advances triggered the so called neuroscientific revolution [1].

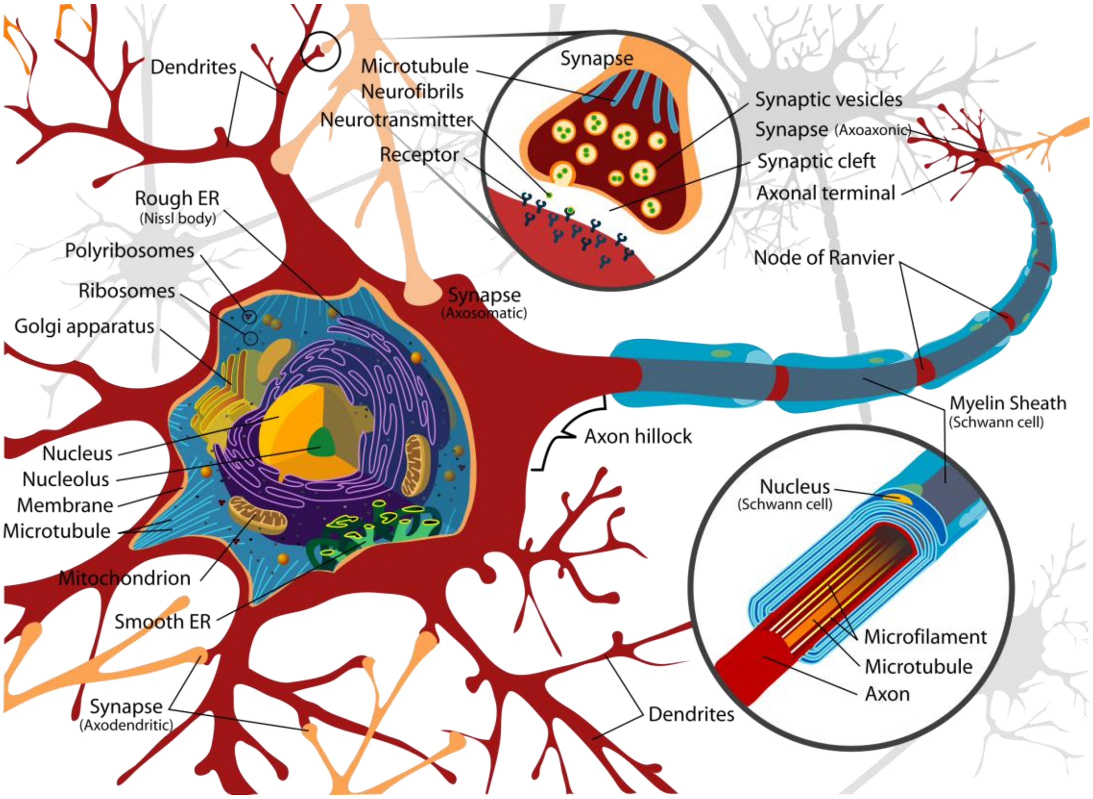


Figure 1.1. The neuron: the unit cell inside the brain [Courtesy of Wikimedia Commons]

Studying brain dynamics requires the simultaneous recording from massive amounts of single-neurons talking to each other. This is very crucial in many complex neuroscience studies, such as those concerned with learning, selection, memory, and other cognitive functions in general [2]. The main concern of such studies is answering a common basic question; “What is the role of neurons within specific brain region and how do they cooperate and interconnect to perform/inhibit a function?” Answering this fundamental question will open new horizons of applications such as: 1) diagnosis of neurological disorders, 2) restoration of impaired neurological functions [3, 4], 3) development of neuro-prosthetic devices [5, 6], 4) brain-machine interfaces [4, 7], and 5) effective learning and educational systems. It would also help engineers build the future generation energy-efficient brain-like processing electronics and machinery.

Many tools such as electroencephalography (EEG), microscopes, magnetic resonance imaging (MRI) and the recent functional MRI have been developed to study and diagnose the brain. While different electrical, chemical, mechanical and magnetic properties have been used in these techniques, the electrical methods to record brain activity are the most promising to understand the dynamics of the brain. Therefore, the key engineering tool to achieve discoveries, to get answers, and to verify different theories about brain dynamics would be an implantable microsystem that can record hundreds or thousands of simultaneous neuronal activities in freely behaving subjects.

This chapter introduces some history about the electrical recording of brain activity since it started in 1875. It also sheds the light on the significance of extracellular recording methods using microelectrode arrays. Moreover, it reviews the advances in development of implantable microsystems and highlights some of their limitations. The challenges and parameters to assess the effectiveness of microsystems are further described. Finally, the research objectives of this work are introduced.

1.1 History of Electrical Recording of Brian Activity

The electrical recording of brain activities has been performed in different scales in clinics and research since the late 1800s. The four main methods are electroencephalogram (EEG), electrocorticogram (ECoG), and extracellular and intracellular action potential recordings (Figure 1.2) [4]. It is worth mentioning that early discoveries in this field were somehow achieved after the invention of vacuum tube amplifier in 1906 by Fleming and De Forest. The vacuum tube amplifier was asserted to be the neuroscientists' effective tool to get answers [8].

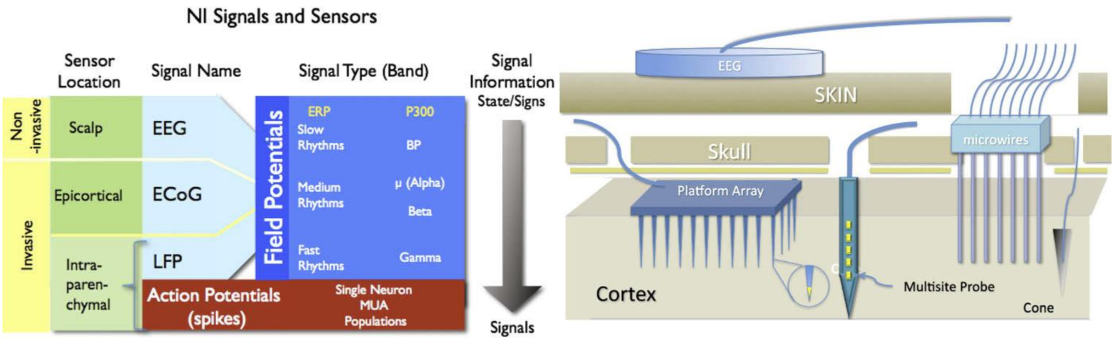


Figure 1.2. Electrical recording methods for brain activity and their sensing locations [4]

It all started in 1875, when Richard Caton discovered the electrical nature of the brain and reported his findings to the British Medical Association. Canton's work received no attention until later in 1924 when Hans Berger recorded the first electroencephalogram (EEG) in humans and cited Canton's contribution in 1929 [9]. After reading Berger's publications, William G. Walter developed his own version of EEG machine in the early 1930s. At this point, scientists proved the usefulness of EEG in detecting and managing epilepsy. In addition, Walter realized that tumors, for instance, could be detected using EEG [10]. Many technical and scientific achievements in neuroscience, robotics, and artificial intelligence followed the discovery of EEG. Until today, scientists and engineers use different advances in electronics and signal processing in designing different devices and brain-controlled machines based on EEG. While EEG is non-invasive and simple, it suffers from limitations such as low recording resolution, and high noise. It also fails to solve the well-known inverse problem, to detect the exact location of abnormalities, and to study the single neuron behavior.

In the 1950s, electrocorticography (ECoG) was pioneered by Wilder Penfield and Herbert Jasper for epilepsy treatment. They proposed the concept of epileptogenic zone as the area of cortex responsible for generating epileptic seizures [11]. While the EEG might be somehow helpful, it is still not precise enough to localize epileptogenic areas [12]. ECoG, on the other hand, offers more sensitivity and precision than scalp EEG recording. For the past decades, intra-operative ECoG has been used in the surgical management of medically refractory partial epilepsies to identify the locations and limits of epileptogenic areas, to guide the extent of surgical resection, and to predict its success [12]. Nonetheless, ECoG suffers from major limitations such as the impossibility to distinguish some critical epileptic events at a specific region from those arising at distant epileptogenic sites [12]. In addition, ECoG cannot be expanded to capture the neuronal level activity that is crucial for more in-depth studies of interactions and interconnections between different units and different regions inside the brain.

In 1925, Edgar D. Adrian recorded for the first time in-vitro action potentials traveling in sensory nerve fibers using a vacuum tube amplifier [8, 13]. He later published a book in 1928 with a conclusion that all sensory messages are carried by trains of all-or-none nerve impulses with various frequency components. In 1939, Hodgkin and Huxley made the first intracellular recording of action potential. Later they developed a voltage-clamp circuit to enable quantitative measurement of ionic currents from a squid axon. Most importantly they published their experimental data with a quantitative model in 1952 and were awarded the Nobel Prize in 1963 for underlying the mechanism of generating the action potential in neurons [14]. While intracellular recording is the most informative/precise technique for measuring single-neuron activities, it could be only obtained in-vitro (under nonphysiological conditions) and from one neuron at a time [15].

The aforementioned methods of recording brain activities can be seen to suffer from limitations, which may challenge the broader applications in neuroscience and brain-machine interfaces. While the intracellular recording method offers the precise information on single neurons, it cannot be used to study cognitive behavior. On the other hand, the EEG and ECoG are limited by noise, frequency contents and spatial resolution. To study brain dynamics in more depth, a method that can provide a balanced tradeoff among in-vivo feasibility, resolution, noise, and frequency contents is needed.

Extracellular action potential (EAP) recording is a promising method that provides very reasonable tradeoffs. Moreover, some studies successfully related the intracellular activity of neurons using the extracellular activity by recording both simultaneously as in [15]. The signals detected here are very similar in shape, but smaller in amplitude, to the actual intracellular action potential. During the neuronal activity, ionic currents flow in the extracellular space and cause voltage drops, which can be measured by inserting an appropriate microelectrode [16]. Depending on the tested subject, the amplitude of this voltage can be in the range of 50-500μV with normally 300 Hz to 5 kHz frequency contents. Some experimental measurements show frequency contents of up to 8 kHz [16, 17]. Along with EAP signals, local field potentials (LFP) can be recorded using the same electrodes. These are important signals of interest that can reveal the slow moving dynamics, which represent average group activities of neurons. For these LFPs, the frequency contents would range from 1 to 300 Hz, while the amplitude can be in the range of 2-3 millivolts. The recording of both EAP and LFP enables the monitoring of spike outputs from the neurons as well as estimating their summed inputs [2].

In 1957, David Hubel developed a tungsten microelectrode and recorded single neuronal activities for periods on the order of 1 hour from cerebral cortex in chronic restrained waking cats [18]. In 1958, Felix Strumwasser used a stainless steel microwire for long-term recording (4 days) from single neurons in the brain of an unrestrained squirrel [19]. Since then, neuroscientists have been using microwires in tetrode (group of 4 microwires) configurations as well as microelectrodes made from tungsten and other rigid materials [20]. While the fabrication of these electrodes was considered an art [21], it suffered significant limitations such as: 1) fabrication precision, 2) reproducibility, 3) high impedance levels, 4) difficulty to accurately insert a large number of wires in a small region, 5) poor chronic performance, and 6) applications to neuro-prosthetics [21, 22].

In 1969, the development of MEMS-based microelectrode arrays has offered a more precise and controlled solution for recording extracellular action potentials [16, 23]. They are substantially smaller in size, minimize tissue damage, and introduce the concept of multielectrode probes, where various cortical layers can be recorded [2, 21]. The next section reviews the advances in MEMS microelectrodes and interface circuits made by various groups since 1969 until 2012.

1.2 Integrated Circuits and MEMS for Extracellular Recording

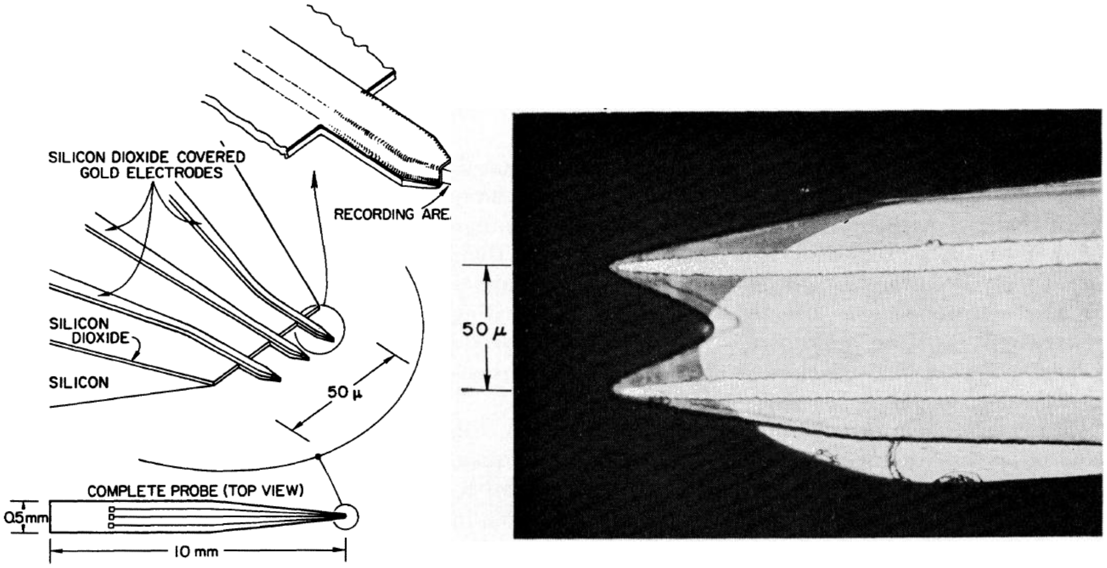


Figure 1.3. Silicon neural probe developed in 1969 [21, 23]

Advances in microfabrication techniques for integrated circuits and MEMS have generated new tools for the neuroscience community in the past few decades [16]. Using these techniques, neural probe was first introduced by Kensall D. Wise in 1969 [23] and published in a journal article in 1970 by Wise, Angell and Starr [21]. Figure 1.3 shows the first micromachined silicon neural probe, which consists of an array of gold electrodes on a silicon carrier. The electrode spacing was controlled from 10 to 20μm, and the tip diameter was as small as 2μm. The authors prove of the feasibility of this approach to extracellular recording, which was the starting point of all micromachined probes.

Microfabrication techniques also provide a feasible solution to the inter-electrodes coupling problem when both electrical stimulation and recording occur at the same time in adjacent areas of the brain. Coupling occurs between neighboring electrodes due to the capacitive voltage divider between the electrode capacitance and the inter-electrodes capacitance. In 1975, Wise proposed to employ a JFET buffer integrated onto the neural probe to reduce coupling, noise, and power consumption [24]. This also helped minimize the attenuation caused by capacitance in the long routes between the recording amplifiers and the ground plane. Although this was a form of hybrid integration rather than monolithic solution, it is considered as the first attempt to include integrated circuits close to the recording sites. It was also the starting point of neural probes with integrated circuits, called active neural probes. The basic structure of a microfabricated probes with circuitry on the probe body, known widely as “The Michigan Probe”, is shown in Figure 1.4 [25, 26] and the detailed fabrication processes are described in [16, 23]. This early work has changed and shaped a new generation of neurophysiological instrumentation. Moreover, it identified the challenges that research groups have tried to overcome during the last 40 years and beyond.

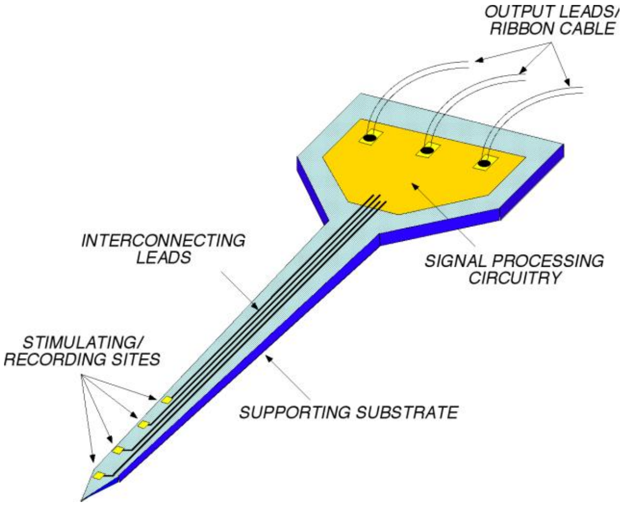


Figure 1.4. Basic structure of the Michigan Probe [25, 26]

The 1970s and 1980s witnessed a great deal of development in using microfabrication methods for neural probes on silicon substrates as well as on glass, ceramic, and polymer substrates as reviewed in [3, 16, 22, 27]. Despite the availability of the technology in early 1970s, the neuroscience community only started to adapt to this technology in the late 1990s [22]. This was due to many factors such as lack of availability since large investment was required to make them commercially available, the complex high-cost fabrication facility, and the structural difference compared to conventional electrodes [21, 22, 24]. However, this technology is becoming more and more appealing when system integration is used to include neural probes and circuitry for processing and telemetry together integrated into one microsystem solution. This microsystem will be referred to as NIMS, or neural integrated microsystem. The following text reviews the major advances made by several groups in the last 25 years as a contribution to implantable NIMS.

As a continuation of the early efforts by Wise et al. [21, 23, 24], Najafi and colleagues published an attractive method for neural probe fabrication that: 1) required only 4 masks, 2) could include circuits, and 3) could be processed on a single side of the wafer [26]. This high-yield process can produce shank widths as small as 20 μm and thicknesses of 8-15 μm. The shank dimensions are defined by deep boron diffusion and later an isotropic etching in ethylene diamine-pyrocatechol (EDP), which stops etching at the deep boron-doped region defined to release the probes. When integrating circuits on probe body, boron doping would only be used for the shank part; leaving the body undoped. The outcome of the work in [26] was a 10-site single shank probe with 10 channels; on-chip recording circuits provided a 40 dB gain per channel and passed frequencies of 100 Hz to 6 kHz, while multiplexing their outputs into one data output connection [28]. The fabrication was done in a 6 μm LOCOS enhancement-depletion NMOS process and the circuit consumed 5-mW from 5-V single supply [29]. References [29, 30] show more details on the integrated circuits and the actual recording results of single neuron activities. This early work not only developed the well-known Michigan probe processes that used boron diffusion as an etch stop, but also increased the awareness in the research community to the following concerns [26, 28, 29]:

1) The lifetime and stability of tissue-electrode interface, especially for chronic use;
2) Electrode geometry for decreased tissue damage and inter-electrode spacing for effective neuronal isolation;
3) Importance of circuitry in improving the recording and decreasing the number of output leads;
4) Area and power consumption of circuitry;
5) Matching the DC levels of electrodes and recording buffers or amplifiers; and
6) Encapsulation of active parts.

Further improvement was done in 1990 by Ji, who reduced the shank width, and developed a single-shank 32-site probe with 8 channels selected from the 32 recording sites [31, 32]. The circuits were fabricated using a 12-mask, 3μm CMOS process and consumed 3mW from 5-V single supply. This work also featured a novel preamplifier that reduced the effects of DC drift from electrodes by employing diodes and capacitors to feedback the low-frequency components [32-34]. The preamplifier was demonstrated to tolerate ±100 mV of DC input drift [34].

In 1988, another technique to fabricate electrode arrays was proposed by Normann, Campbell and colleagues from University of Utah, where they vertically processed a 4mm x 4mm silicon substrate to produce 100 pyramidal shape needles, known as the Utah electrode array, shown in Figure 1.5 [35-37]. Each needle is 1.5 mm in length and includes only one recording/stimulation site. The length of the needles is limited by the thickness of the wafer used. While this electrode array has mechanical strength that allows easier insertion to the brain tissue, it cannot be used for highly-parallel recording due to the limitation of single site per needle (or shank), and is unsuitable for chronic applications in human brain due to the stiffness of the needles [38]. The encapsulation of the Utah electrode array has been reported to slowly push it out of the implanted tissue, so the authors in [39] recommended limiting its use to short-term chronic applications.

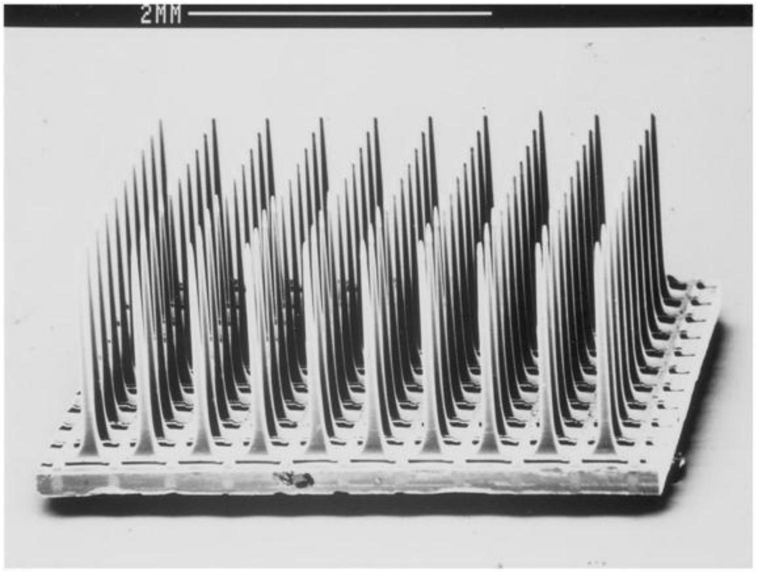


Figure 1.5. Utah electrode array [37]

In 1991, Hoogerwerf built upon the effort and experience in the University of Michigan and made the first truly three-dimensional electrode array for high-density volume recording of neural activity [40-42]. The array consists of multiple two-dimensional (2-D) passive probes assembled in a perpendicular silicon-micromachined platform that acted as the foundation of the entire system, held the probes, and had hybrid circuits integrated on it (Figure 1.6). The slots that accept the probes on the platform were defined as undoped boron regions and etched in EDP. Two wings at the sides of each probe defined the insertion depth down into the platform, while spacers are later inserted on top of these wings, at slots made for this purpose, to make sure all probes are parallel and at certain spacing (100-200µm). One outstanding issue for the 3D assembly was the transferring of orthogonal lead connections between the probes and the main platform. Hoogerwerf developed a solution that employed an electroplating process, using nickel, to bridge between beams on the probes and pads on the platform. The final outcome was a four-probe 16-shank prototype that was successfully tested in guinea pig cortex for three months [40].

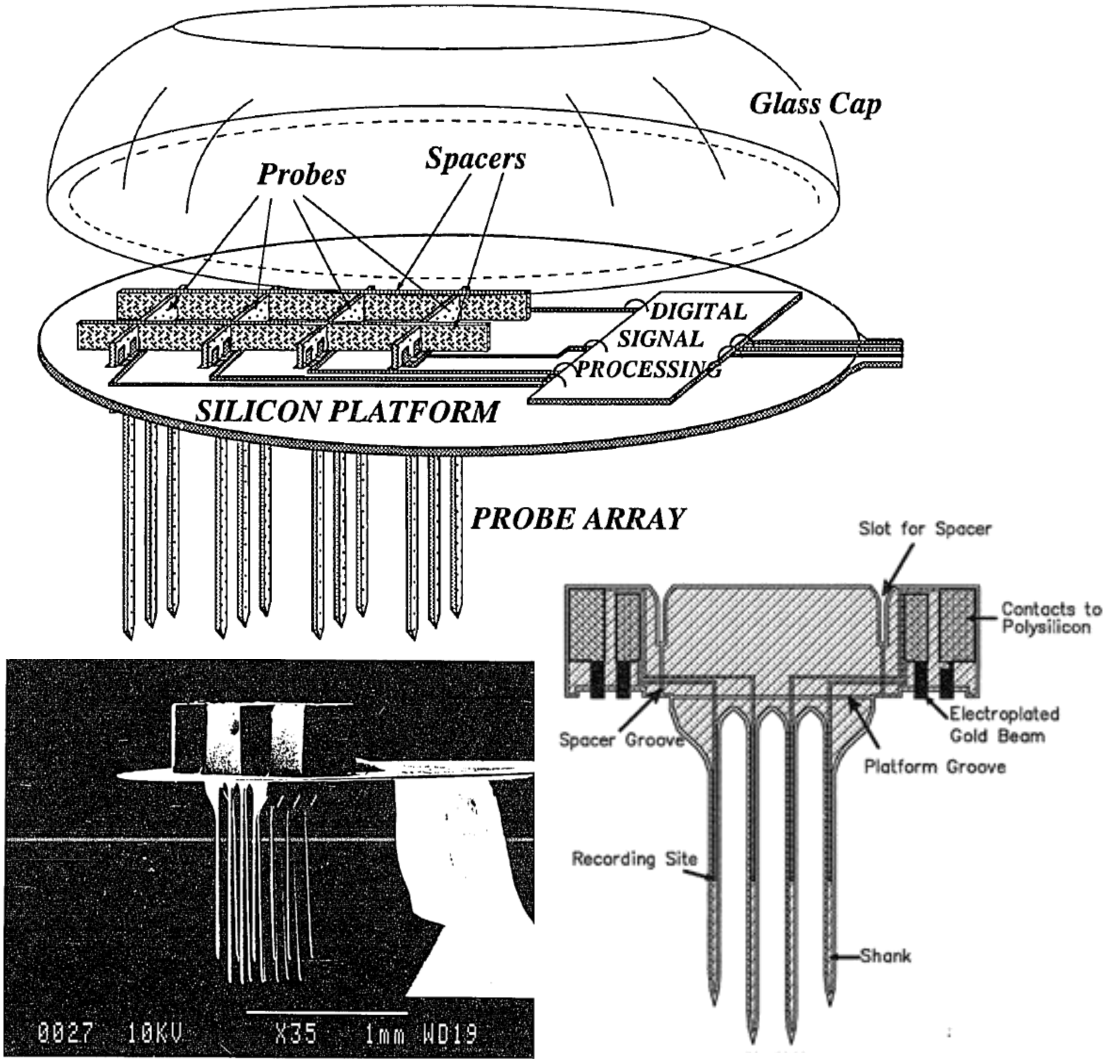


Figure 1.6. Schematic and Implementation of three-dimensional electrode array [41]

In 1999, Bai demonstrated another 3-D electrode array that, for the first time, consisted of active probes (Figure 1.7) [43, 44]. Bai implemented an amplifier using a diode at the input to stabilize the dc-baseline, but the measurement results suggested that the approach was inadequate. For 3-D probe assembly, Bai realized a more practical way for transferring the leads from the probes to the platform by attaching right-angle gold-plated beams at the probes to the pads on the platform using ultrasonic bonding [43]. Up to 8 probes (16-shanks each) were assembled and successfully demonstrated the ability to record single neurons activities.

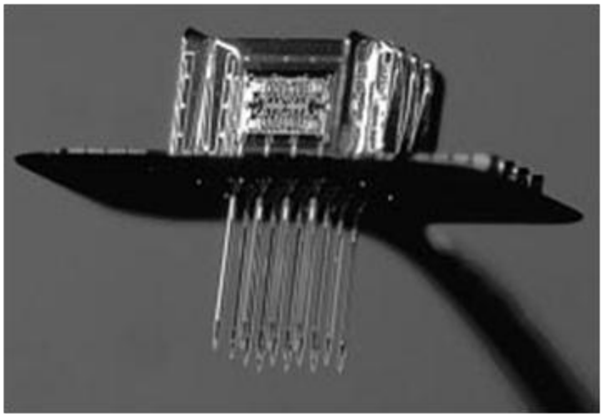


Figure 1.7. Three-dimensional active probe array [44]

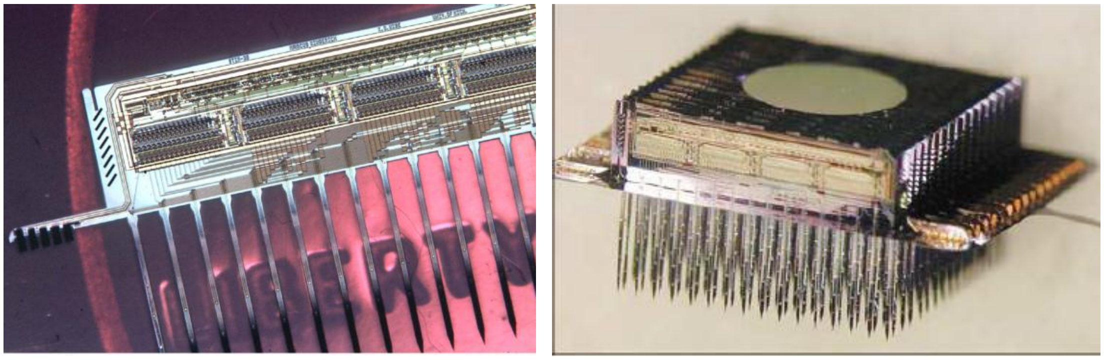


Figure 1.8. 2D and 3D probe array [26,45]

In 2002, Gingerich demonstrated the in-vivo performance of a fully-functional 3-D microelectrode array including graphical user interface through RS-232 serial connection to a custom DSP-based probe controller board [45, 46]. The 3-D array consists of four 2-D planar probes, each has 16 shanks (400 μm spacing) and 4 sites per shank (Figure 1.8 [25]). Four channels can be selected for stimulation and one channel for recording per 2-

D probe. While assembly techniques were similar to those developed by Hoogerwerf and Bai, Gingerich tackled more problems regarding circuit integration and demonstrated the feasibility to assemble a total of 16 2-D probes with a total of 1024 sites, the largest reported to date. This, however, suffered from limited number of recording channels; one per 2-D probe.

In 2003, Jamieson demonstrated one of the earliest systems that targeted the significance of highly parallel chronic recordings of neural signals [47]. Both 64-sites and 96-sites probes were designed with a selection of 8 output channels (Figure 1.9) and as small as 20 μm inter-electrodes spacing on each shank, which emphasized the significance of the micromachining approach in achieving high spatial resolution [48]. While the increasing site counts offered significant advantages in the animal experiments, active probes were also proven to be more advantageous compared to passive probes for long-term experiments [48].

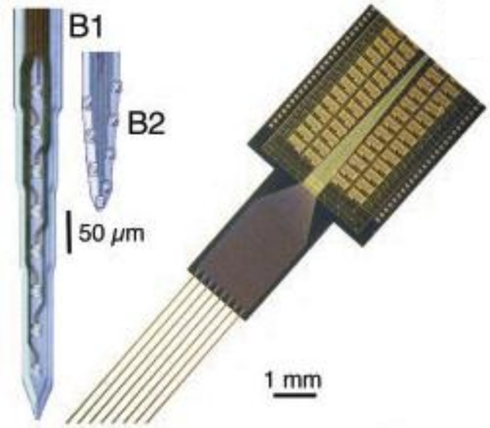


Figure 1.9. 64-site active probe [48]

In 2004, Olsson expanded the capabilities of 3-D neural probes by using on-probe integrated circuits that extended the number of parallel processed electrodes using time-division multiplexing for 8 sites selected from 64 sites at each 2D probe. A 4-probe 3-D array, therefore, provided a total of 32 channels selected from 256 sites [49-51]. The 3-D probe (Figure 1.10) was further connected to a platform-mounted ASIC that performed 5-bit analog-to-digital conversion and spike-detection to save unnecessary bandwidth in cases where the neuroscientists are interested in seeing the time-occurrence of neural spikes. The entire system consumed 5.4mW from a 3-V supply [51]. To overcome the dc stabilization problem in front-end amplifiers, Olsson introduced a capacitive-coupled approach (in [52]) that was later proven to be of low-noise performance [53, 54], and was adopted by many designers of neural amplifiers.

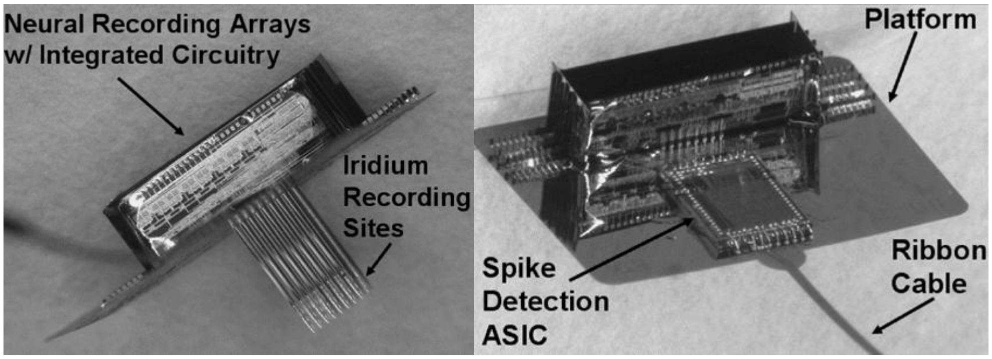


Figure 1.10. Neural Microsystem with 256 sites and 32 channels [51]

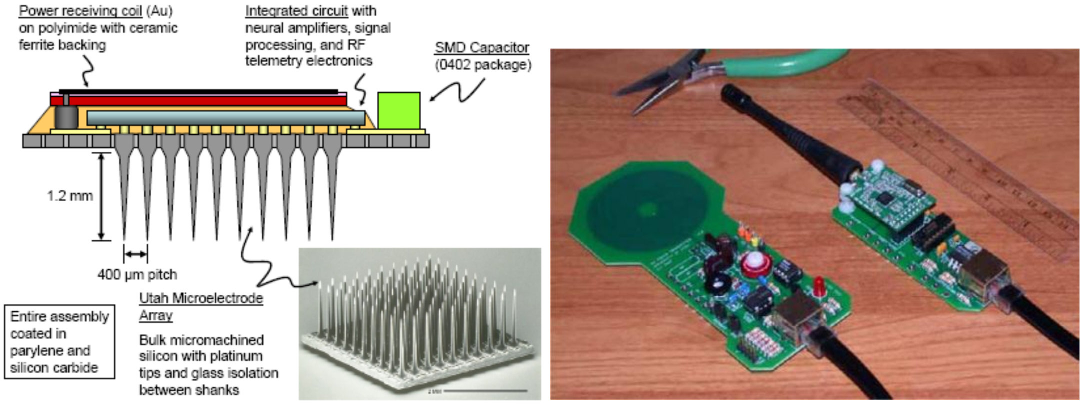

Power receiving coil (Au) on polyimide with ceramic ferrite backing
Integrated circuit with neural amplifiers, signal processing, and RF telemetry electronics
SMD Capacitor (0402 package)
1.2 mm
400 μm pitch
Utah Microelectrode Array
Bulk micromachined silicon with platinum tips and glass isolation between shanks
Entire assembly coated in parylene and silicon carbide

Figure 1.11. Utah neural interface assembly and circuit boards for wireless power and data link [57,58]

With this relatively large number of channels, power consumption became an outstanding issue that would limit the capability of recording from large numbers of neurons, especially when considering the limited power provided for the integrated system level either by batteries or other wireless approaches such as inductive coupling. Harrison and colleagues at University of Utah emphasized the power issue [55, 56], and reported in 2006 a neural recording system that included an analog front-end IC with 100 amplifiers and inductive coupling for wireless power delivery [57]. Spike detection data, of all 100 sites, could be serialized and sent by a fully-integrated 433-MHz frequency-shift keying (FSK) transmitter. However, only one full bandwidth channel can be sent at full 10-bit digital resolution. The system, shown in Figure 1.11, continued to improve in performance and integration until it was shown to fully operate wirelessly in-vivo in 2009 with, however, low noise performance due to inductive coupling interference [58].

The work, led by Harrison, increased the awareness of researchers, especially designers of integrated circuits, about the related challenges in neural microsystems. Harrison also inspired many researchers through his ideas and solutions to many of the circuits/systems outstanding issues. There are more efforts, however, in both MEMS and circuits that need to be done for expanding the capabilities and the integration of a full system. On the other hand, circuits that employ the Utah electrode array suffer from its limitations on the system level and bound their solutions on the circuit level.

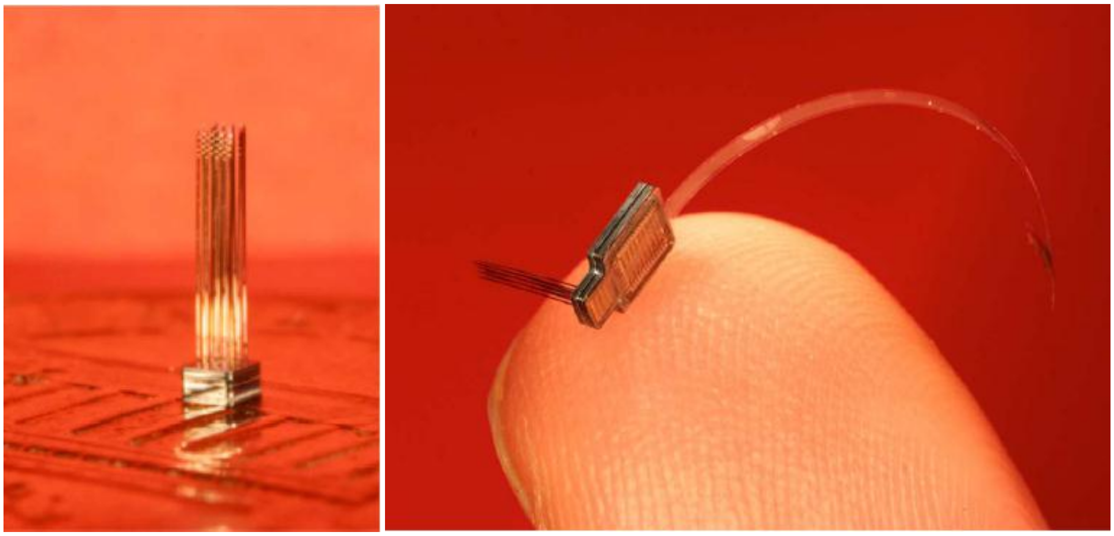


Figure 1.12. A 64-site 3D neural probe (left), and integration with analog front-end and parylene cable [60]

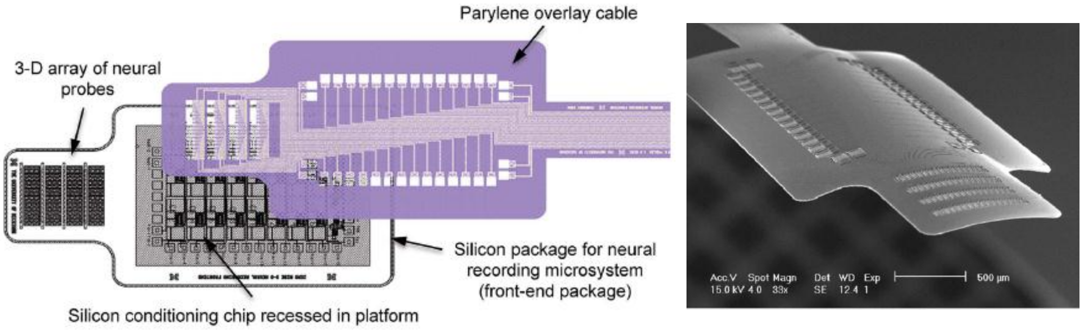

Parylene overlay cable
3-D array of neural probes
Silicon package for neural recording microsystem (front-end package)
Silicon conditioning chip recessed in platform
Acc.V Spot Magn Det WD Exp 500 µm
15.0 kV 4.0 33x SE 12.4 1

Figure 1.13. Integration of the microsystem using parylene overlay cable approach [60]

In 2008, Perlin at the University of Michigan developed a more compact 3-D probe compared to the previous approached demonstrated by Hoogerwerf, Bai, and Gingerich, where probes had lateral wings that were necessary for assembly but did not contribute to the probes function. For lead transfer, the 2-D probes in Perlin's design had bendable electroplated gold tabs extended off their back-ends and high density tab bonding was used with tabs on as small as 40 μm pitch. The 2-D probes were inserted into slots in a thicker platform that accommodated the whole area of the probe back-end, forming the 3-D probe shown in Figure 1.12, and later connected by a microfabricated parylene cable to a recording system through a custom PCB [59]. Perlin further implemented a chip to amplify and filter 64 parallel channels. The chip was recessed in a larger platform including the electrode arrays and connected to the electrode sites using a novel parylene overlay cable microfabricated as shown in Figure 1.13 [60].

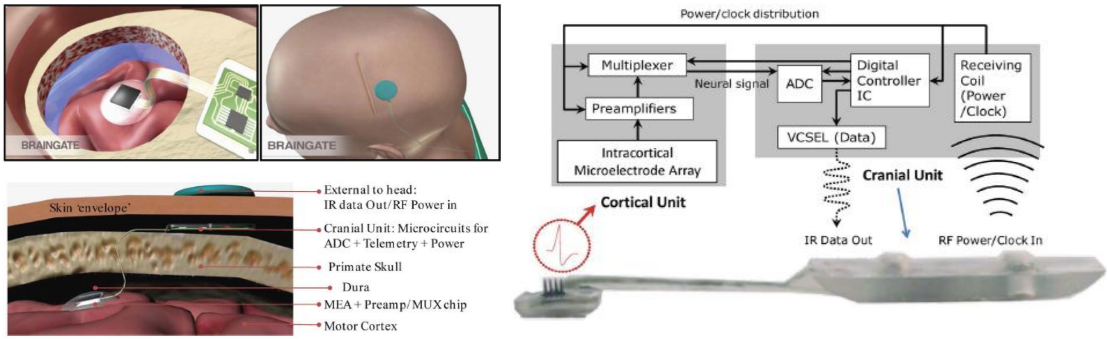


Figure 1.14. Two-stage neural microsystem proposed in [62]

Another interesting approach for neural recording microsystems has been pursued at Brown University led by Donoghue and Nurmikko group. Their particular focus has been the intensive signal processing aspects of brain-machine interfaces and conducting actual experiments on patients suffering from tetraplegia to help them move things on a computer screen by thinking. In 2009, they built a 16-channel wireless neural interfacing system that can transmit signals transcutaneously using infrared light pulses through skin [61]. Similar to the two-stage platform pursued in Michigan, their system (called Braingate, Figure 1.14) consists of: 1) a front-end platform chip that includes preamplifier arrays flip-chip bonded on the back side of a Utah electrode array, and 2) a processing platform (above the skull) including ADC, processing ASIC, and power/data transmission components. In a recent article [62], this two-platform concept was constructed on a single polymer substrate in a two-island geometry (Figure 1.14 right) with 5 wide planar wires for routing signals and power. They also experimented their circuits in an awake monkey with 16-channel operation. This system, however, suffers mainly from large power consumption per channel (45 µW per channel) in addition to the limitations imposed by the Utah electrode array itself.

Another active research group in University of Toronto has reported a 256-channel neural recording system [63]. Each recording channel employs two stages of ac-coupled amplification with 34 dB gain at the first stage, and a programmable gain at the second stage, which has its output sampled and stored by a switched-capacitor sample-and-hold circuit. Further delta compression is performed by subtracting two successive samples and discarding data below certain threshold. Although this is the largest reported number of channels, only 64 sites in a 100-sites Utah array were bonded for system testing (Figure 1.15). In addition, sampling rate was limited to 10 kHz and no scheme was shown for reduction of output wires.

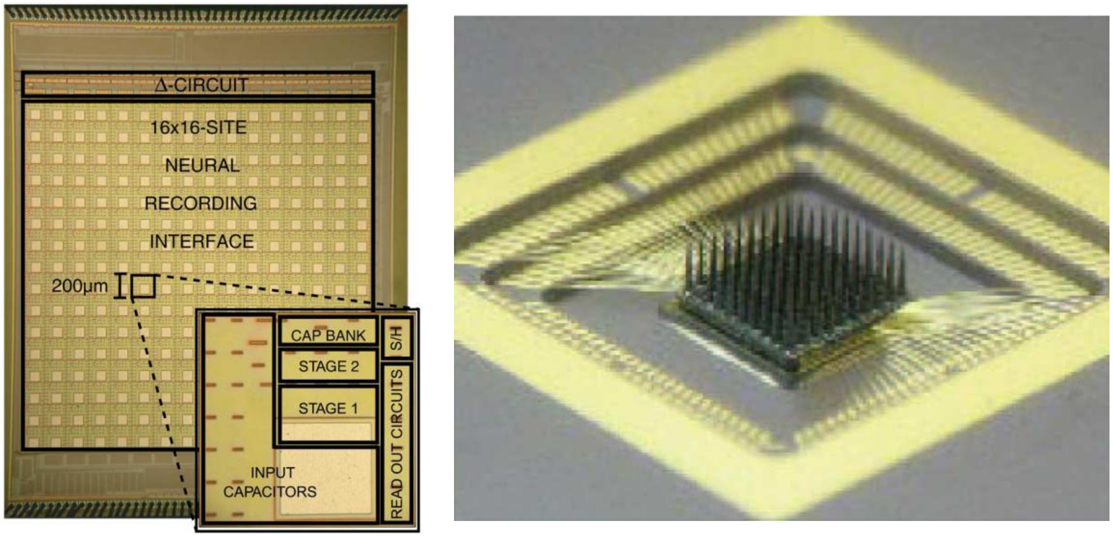


Figure 1.15. 256-channel circuitry and 64-sites bonding to Utah electrode array [63]

In 2009, a research group, at the University of California- Santa Cruz, reported a 128-channel recording circuitry, which consisted of eight 16-channel analog front-ends, digital signal processor for spike detection and feature extraction, and ultra wide-band (UWB) telemetry [64]. Time-division multiplexing is used for each 16 channels to share one analog-to-digital converter (ADC). While reference [64] claims a 128-channel system, each component's measurement results were reported separately and no testing results were shown even for one single channel results, let alone for the whole system.

Continuation of the efforts at the University of Michigan led to an implantable 64-channel recording system published in 2009 [65]. The platform (Figure 1.16) consists of a two-stage implementation, where the first stage consists of four 16-channel preconditioning chips (for amplification and filtering), and the second stag includes a 64-channel processing unit and a bidirectional telemetry for power and data. The amplification stage consumed 75μW per amplifier in a small die area (0.072 mm^2). The processing unit featured two modes of operation: 1) a scan mode to detect the spike occurrences for the 64 channels and send, using the telemetry stage, the origin sites and time of these spikes, and 2) a monitor mode to digitize two full channels at 8 bits resolution. While this implementation would be very appropriate for brain-machine interface neuroscience, it may not be for complex neuroscience studies where full-signal spectrum is required.

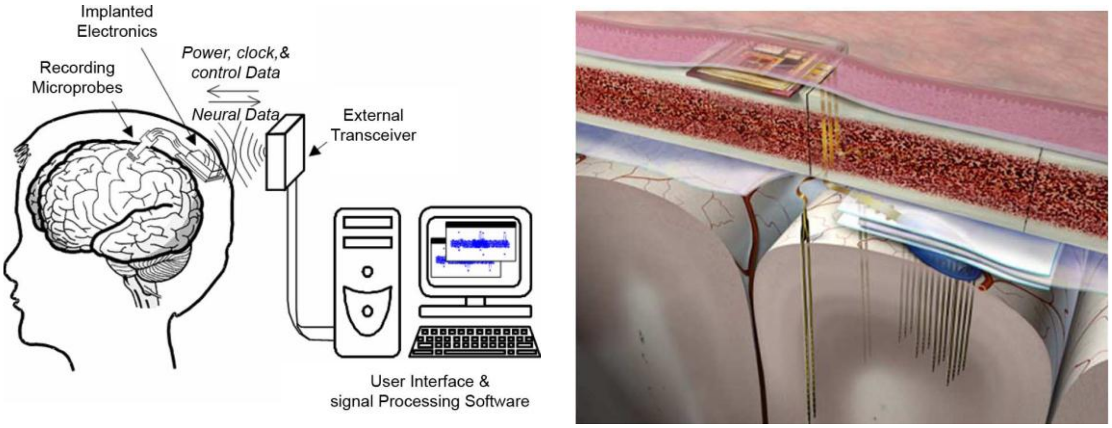


Figure 1.16. Neural microsystem pursued at the University of Michigan [65]

Another active research group in Georgia Tech, led by Ghovanloo, reported in 2009 a 32-channel neural recording circuitry that consumes 5.6mW and consists of an array of fully-differential LNAs (40 dB gain and variable bandwidth), and adjustable 27.7/37.1dB gain amplifiers. The 32 analog signals were then time-to-digital converted, one signal at a time using TDM, by a pulse-width-modulation (PWM) block that compares these signals with a triangular waveform, and feeds the generated pulses to a hybrid VCO for wireless data transmission; using an off-chip SMD inductor as a transmitting antenna [66]. The full system (Figure 1.17) was bench-top tested with a custom-designed PCB as a receiver. Compared with [64], the authors claimed this as the first true demonstration of a 32-

channel wireless system with ~10 kHz bandwidth per channel. In 2010, the same group reported similar circuit architecture as in [66] and added inductive coupling power delivery [67]. They also employed a power-scheduling mechanism to disable part of the unused LNAs, which help save further power consumption (7mW when all channels are active) [67].

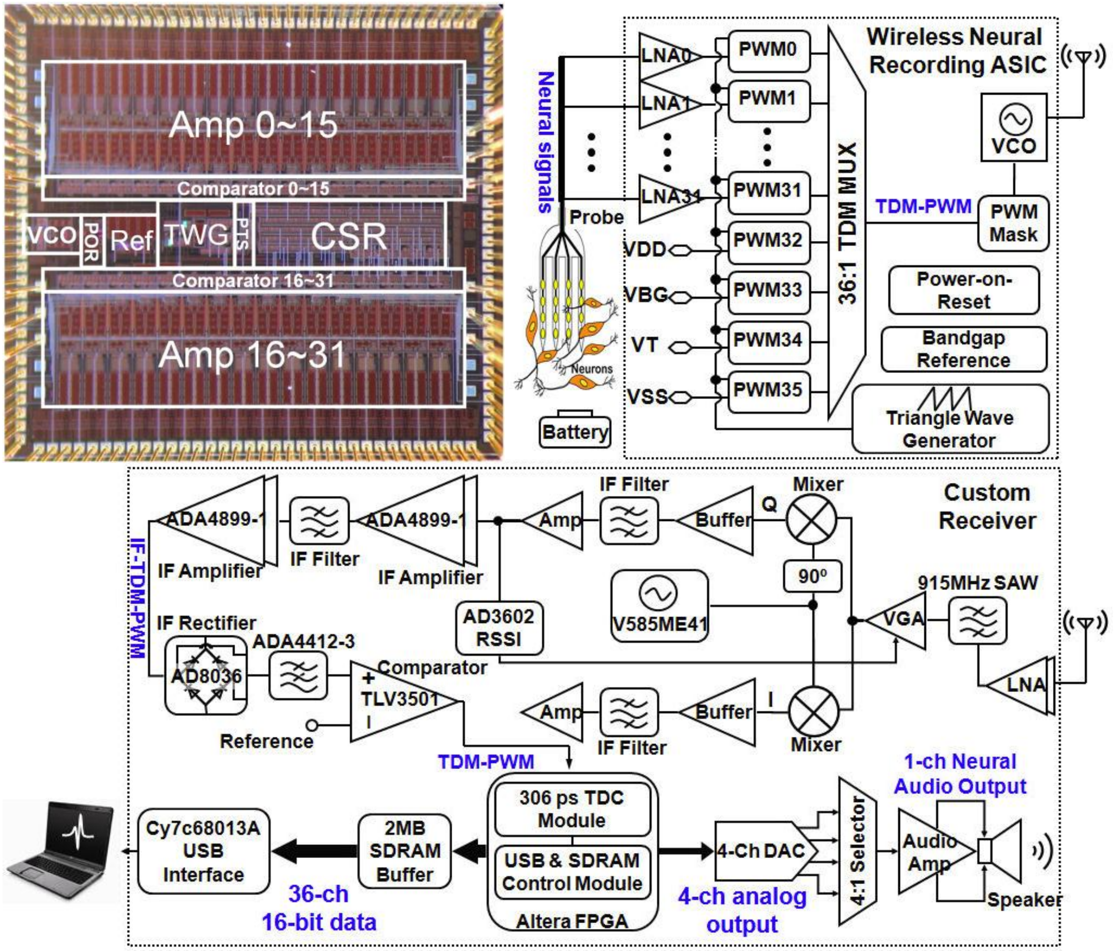


Figure 1.17. Block diagram and circuit microphotograph of the system described in [66]

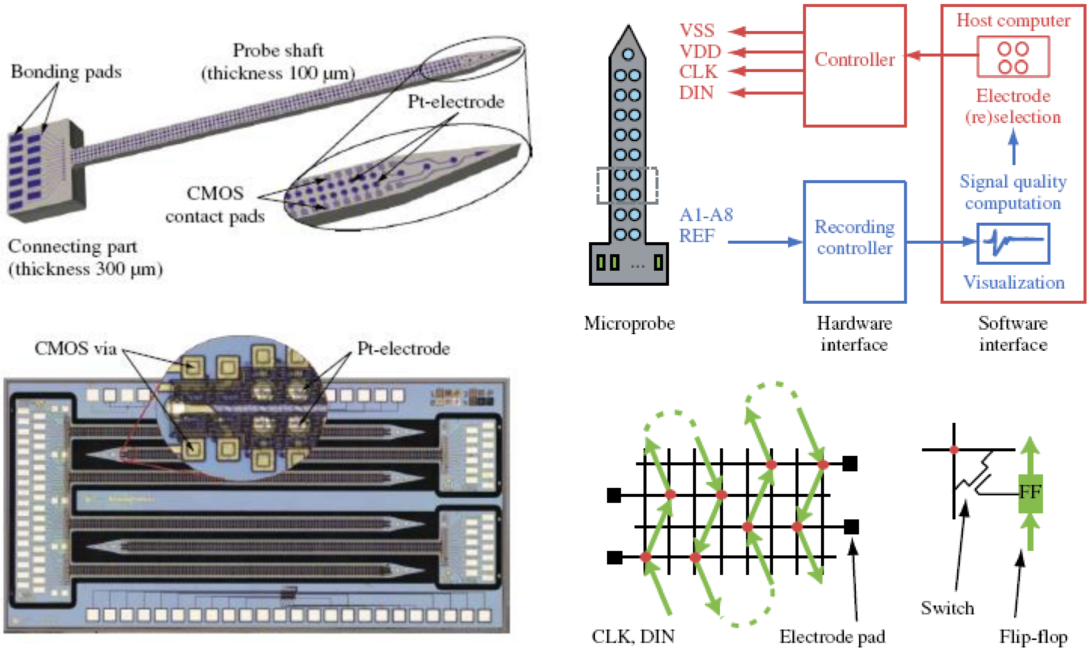


Figure 1.18. Monolithic active neural probe including 188 electrodes and site selection [68]

In Europe, a group of researchers from IMTEK (University of Freiburg), IMEC and others leads a project called *NeuroProbes*; funded by the EU. In 2010, they reported 188-electrodes shanks (with on-shank 8-channel site selection circuitry) using monolithic CMOS and MEMS fabrication and integration [68]. The developed active neural probes (Figure 1.18) included shanks that are 140µm wide, 80µm thick, and had 550µm shank pitch (for multi-shank probes). Each shank includes a total of 188 electrodes, arranged in two columns, and a switch matrix controlled by 5-line digital bus to allow 8-channel site selection. While this probe has the most massive number of sites reported, it supports the recording from only 8 sites simultaneously.

The most recent work at the University of Michigan was done in 2010 by Merriam, where she focused mainly on developing new 3-D electrode array structures that can be easily assembled by folding 2-D probe structures [69]. Merriam also implemented an LNA that consumes 46.5 µW of power and 0.026 mm^2 of area, and provides 59 dB gain and 21 kHz bandwidth. She also designed multiple lattice probes used to investigate the immune responses and interaction of tissues. While these lattice probes did not include recording or stimulation sites, the histological bio-response studies showed a reduced amount of neuronal cell death when compared with conventional solid-shank probes of the same size. This indicated that tissue response can be modulated through appropriate structural design.

There has been also great progress in the area of integrated circuits. For example, Wattanapanitch *et al* [70] reported a modular 32-channel AFE that reduced the wires-to-channels ratio to 1:4, and achieved an average NEF of 4.5. This design also achieved a small area mostly due to the reduction in AC-coupling capacitors, and a small power consumption of 10.1µW/ch [70]. Azin *et al* [71] reported a 4-channel modular AFE that achieved a minimum NEF of 2.33, but consumes ~27µW/ch in an area of 0.625mm^2/ch (only for AFE). Most recently, Muller *et al* [72] presented a 1-channel AFE with very small area (0.013mm^2) and small power consumption (~5µW). The small area and power were realized by mostly-digital design at 0.5 V using 65nm CMOS technology. Area saving also comes from the elimination of AC coupling capacitors and the design allows ±50mV of DC offset at the electrode interface. Although this work features the smallest area reported for neural recording, it used an external FPGA-based digital filter and no data has been reported on the dynamic range or gain programmability [72].

The previous literature review shows how the last 40 years of innovation and development had revolutionized the neural recording tools from a conventional acute structure to a micromachined one dimensional array developed to more complex three-dimensional arrays with integrated circuitry for mapping the dynamics of the brain [16, 27]. Nonetheless, a growing number of challenges are discovered and realized with every single research publication. Many research efforts continue to push the limits towards achieving neural integrated microsystems that satisfy the needs for complex neuroscientific cognitive studies, brain-machine interface studies, as well as neuro-prosthetic devices to improve healthcare.

While each of these segments deals with brain dynamics, the questions they try to answer are different and the challenges to deliver the most effective engineering tool for them are also different. This point is further discussed in the next section.

1.3 Challenges and Trends of Neural Integrated Microsystems (NIMS)

1.3.1 User Segmentation

The challenges of achieving effective microsystems depend highly on the needs of the ultimate users. Figure 1.19 identifies three different segments of applications, their most critical needs/challenges (indicated with black arrows), and the interactions between different needs (indicated with pink arrows). While part of the challenges are common to all types of NIMS, their definitions and impacts are very specific to the application domain. The three distinguishable application domains for neural recording can be summarized as follows:

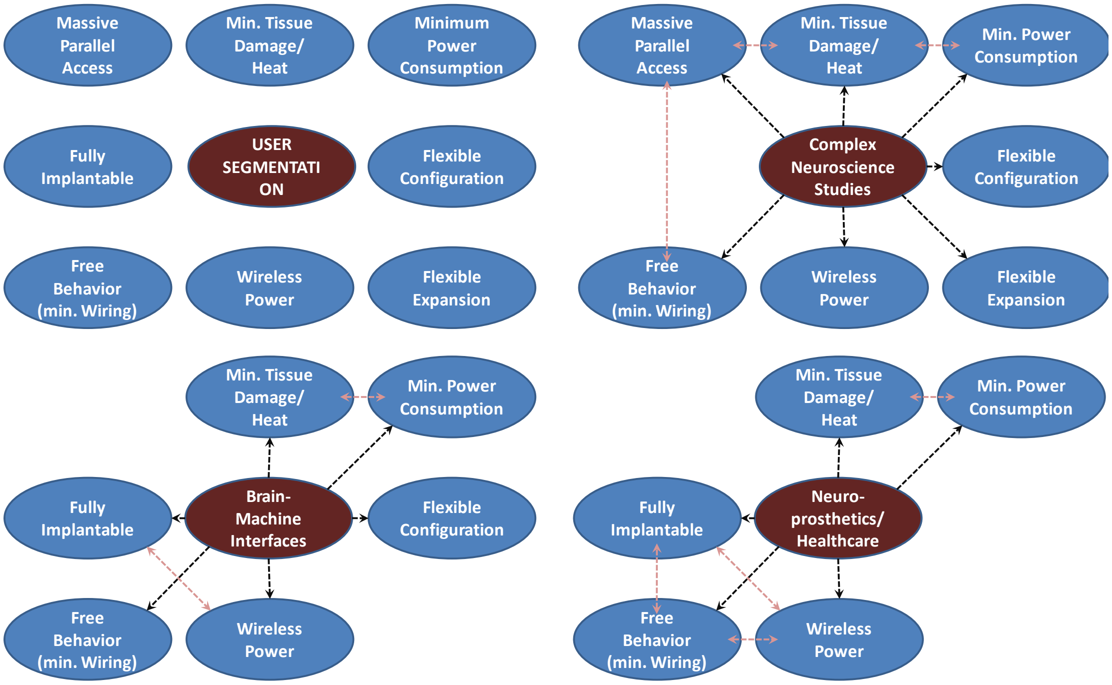
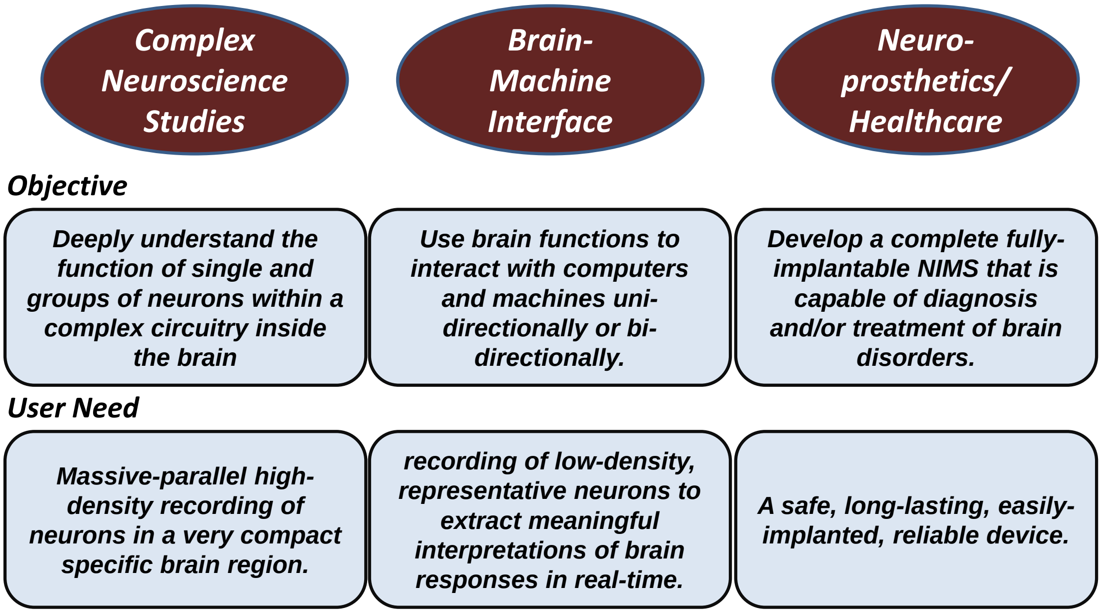


Figure 1.19. NIMS user segmentation based on needs

1) ***Complex Neuroscience Studies:*** This application domain aims to investigate and deeply understand the role of single and groups of neurons within the complex circuitry inside the brain. Understanding complex brain dynamics that form memory, consciousness, and other cognitive behaviors is the main motivation behind this domain. The most outstanding user need here is the massive-parallel high-density recording of neurons in a very compact specific brain region.

2) ***Brain-Machine Interface Neuroscience:*** This domain focuses mainly on applications that uses well-understood brain functions (from the previous domain) to interact with machines (including computers). This application requires highly intensive processing of data recorded from relatively low-density, but representative, neurons to extract meaningful interpretations of brain responses in real-time to control machines. This domain provides a very crucial intermediate step towards effective neuro-prosthesis.

3) ***Neuro-prosthetics for healthcare:*** This domain develops complete fully-implantable NIMS that include all the capabilities to diagnose and/or treat brain disorders in patients. The main user need here is developing a safe, long-lasting, easily-implanted, reliable device. An example of these devices is deep-brain stimulation developed by Medtronic, Inc.



	Complex Neuroscience Studies	*Brain-Machine Interface*	*Neuro-prosthetics/ Healthcare*
Objective	*Deeply understand the function of single and groups of neurons within a complex circuitry inside the brain*	*Use brain functions to interact with computers and machines uni-directionally or bi-directionally.*	*Develop a complete fully-implantable NIMS that is capable of diagnosis and/or treatment of brain disorders.*
User Need	*Massive-parallel high-density recording of neurons in a very compact specific brain region.*	*recording of low-density, representative neurons to extract meaningful interpretations of brain responses in real-time.*	*A safe, long-lasting, easily-implanted, reliable device.*

Figure 1.20. The main objective and use need of the three major segments in neural recording

A summary of the main objective and user need of each segment is shown in Figure 1.20. The work in this thesis focuses on the NIMS challenges in the first domain: complex neuroscience studies. Solving some problems in this domain would improve the understanding of the most efficient machine ever created, the brain. This will not only have huge impacts on healthcare and machine interfaces, but also impacts the way people would design future computers, machines, and devices. The next sub-section elaborates further on challenges and potential solutions in complex neuroscience studies.

1.3.2 Challenges and Trends for Highly Complicated Neuroscientific Studies

A very interesting article observed a trend in neural recording similar to Moore's law[1] for integrated circuit transistors. The article showed that the number of simultaneously recorded channels actually doubles every 7 years [73]. This implies that recording a few thousands of channels will become feasible in the next two decades. There are, however, some challenges that need to be overcome and one of them is the recording density (number of recording sites per unit volume/area).

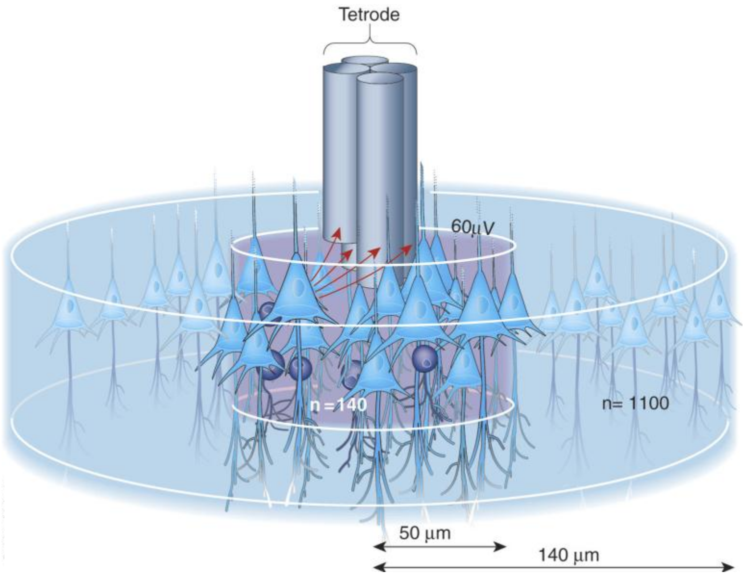


Figure 1.21. Neurons at the vicinity of recording electrode tip [2]

There may be as many as 140 neurons surrounding the tip of a recording electrode within just 50 μm radius of the electrode's center (Figure 1.21) [2]. One single electrode, therefore, would record the contribution of these different neurons located at different radii, hence different attenuation levels, away from the electrode center. One outstanding issue in computational neuroscience is sorting of these signals (spike sorting), which is key to understanding the cooperative activity and behavior of these different neurons [2], and involves highly intense signal processing algorithms. Although triangulation methods

[1] Moore's law is the observation that the number of transistors on integrated circuits doubles approximately every two years

are used among multiple electrodes to sort neuronal spikes, similar to the GPS, there is still a gap between the theoretically recordable neurons and the actual number achieved in practical recording using these algorithms [2]. While this explains why increasing the efficiency of spike sorting algorithms is a major neuroscience research topic, it is worth mentioning that these processes may be relaxed if the engineering tool provides tremendously increased amount of simultaneous parallel sites recorded in a small volume [2, 15]. Recording massive number of neuronal activities in parallel not only offers many advantages in revealing more information about the dynamics of the brain, but can also reduce the number of animals, variability in recording over multiple sessions, and reduce maintenance costs in neuroscience laboratories [2]. This goal in itself, however, brings tons of challenges to both aspects of NIMS: Circuits and MEMS.

Neuroscience Needs	**Engineering Challenges**
1. Chronic or semi-chronic recording	1. Sustainable tissue-electrode interface
2. Surgery overhead and complexity.	2. Small system size with few or no wires.
3. Efficient spike sorting.	3. High-Density electrode arrays with massive parallel channels, and low noise circuits.
4. Tissue damage during insertion	4. Small shank width or lattice structure
5. Tissue damage by heating	5. Low-power circuits
6. Flexible system configuration	6. Modular system partitioning
7. Flexible system expansion	7. Scalable Modules in the system including 3D Probes
8. Free animal behavior	8. Minimum lead transfer from probes to outside station, wireless data/power

Table 1.1 Challenges of Neural Integrated Microsystems

Table 1.1 and Figure 1.22 show the neuroscience needs translated into engineering challenges and parameters that can assess the efficiency of developed and to-be developed NIMS.

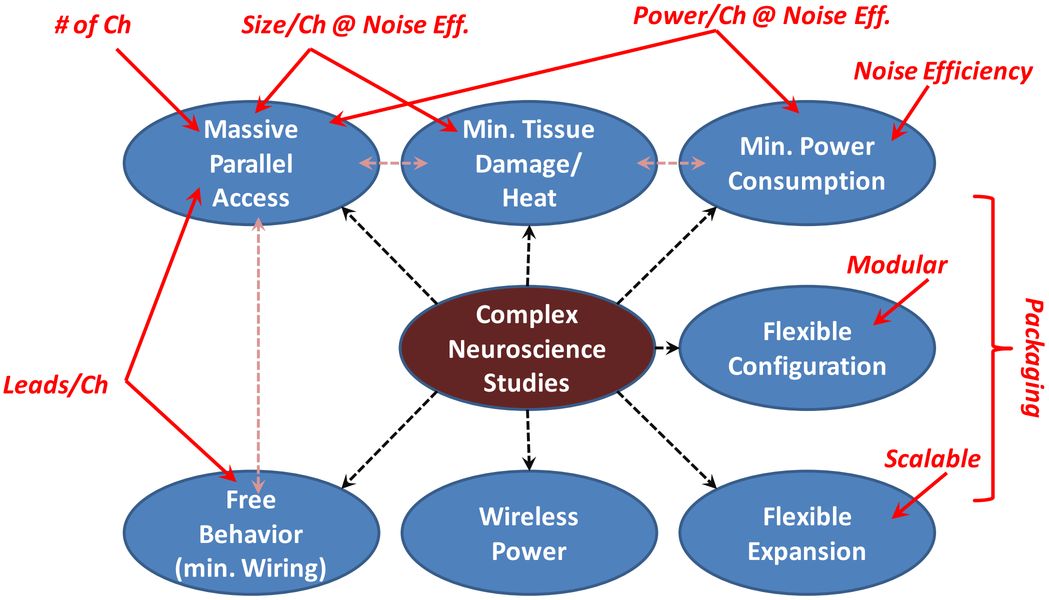


Figure 1.22. Parameters and metrics to asses NIMS

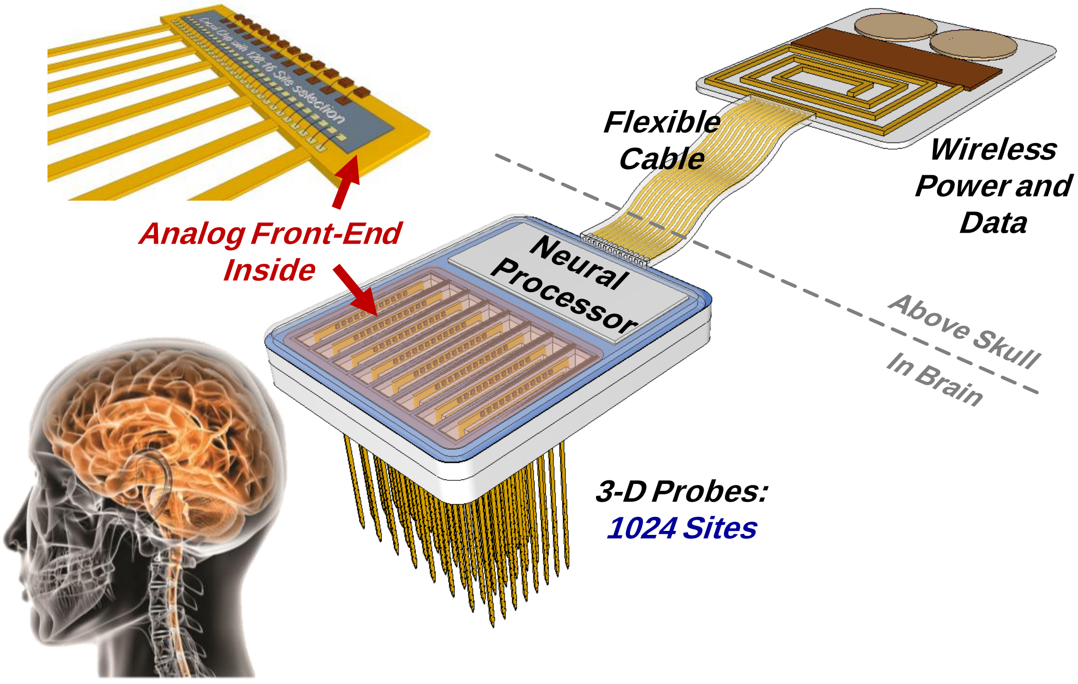


Figure 1.23. 1024 channels NIMS "Brainavigator"

In this thesis, an envisioned NIMS that would target all these challenges is illustrated in Figure 1.23[2] and is referred to as ***Brainavigator***. Brainavigator would allow breakthrough experiments in neuroscience by allowing neuroscientists to record up to 1,024 channels simultaneously in a compact form suitable for freely-moving animals. In the light of the previous literature review, there has been a tremendous amount of research done, especially at the University of Michigan, to address the MEMS challenges including: design and fabrication of high-density neural probes, effective methods for 3D packaging, integration of intra-cranial cables to transfer data and power between two-platforms, and investigation and development of sustainable electrodes for chronic recording in animals. Nonetheless, the circuit challenges, especially in the analog front-end, have been yet overlooked including the consideration of the boundary conditions placed by MEMS integration. Table 1.2 shows the state-of-the-art circuits achievements regarding the metrics in Figure 1.22.

NIMS Parameter	**State-of-the-art Achievement**	**Ref**
Number of channels	32 channels at K[3] = 100 and 0.16 mm^2/channel	69
Power per channel at noise efficiency	15μW at K = 42	66
Size per channel at noise efficiency	0.04 mm^2 and K = 42	66
Unique leads per channel	2:16	65
Modularity in design	NA	NA
Scalability and expandability	NA	NA

Table 1.2. State-of-the-art circuit achievements

While state-of-the-art analog front-end integrated circuits have demonstrated the ability to record up to 32 simultaneous channels [66], scaling the channel count to 100 or 1,000 channels without sacrificing resolution and bandwidth [55, 65] will require yet unmet efforts regarding: 1) significant reduction of system power consumption to allow feasible use of energy scavenging techniques and avoid tissue heating, 2) small implementation area to integrate into probes, and 3) hierarchical design to reduce the

[2] This figure was drawn in part by Sun-Il Chang, University of Michigan. The brain photo is courtesy of Michigan Engineer

[3] K is a power efficiency metric developed in [85]

number of leads transferring signals from the probe to a host station. The development of circuits that also boosts the capabilities of active neural probes is imminent for another evolution in neuroscience discoveries. In addition to making complex experiments more practical and feasible, active probes also save the cost of expensive equipment, reduce the number of connections, reduce tissue damage per effective recording, eliminate the site selection procedure, reduce noise and power, and provide a tool to get the real understanding of brain dynamics.

1.4 Research Objectives and Overview

The objective of this work is to address the circuit challenges in the analog front-end by designing and implementing circuit architectures that would enable the recording and processing of up to 1,024 channels of neural activities by taking into account the MEMS integration requirements as well as the feasibility to be powered by energy scavenging techniques. Specifically, we propose unique features:

- To provide a diversity of analog front-end architectures to meet the challenges of massive parallel recording using different methods that take advantage of the physics, signal processing, and integration domains for improving the final outcome of power, noise, area, and bandwidth or recording;
- To design and implement modular analog front-ends that push the current limits of size, power and noise efficiency, and reduce the number of leads transferring signals from the probe to a host station can to enable hybrid-integration onto neural probes for massive-parallel recording of neuronal activities;
- To explore the possibility of using signal processing in the analog and digital domains and design the integrated circuits that can provide further energy saving by transferring computations in the power-unlimited receiver and designing analog circuits assisted by digital calibration inside and outside the body;
- To design a massive-parallel recording monolithic active neural probe with actual on-the-shank analog recording channels that can further reduce implementation size, power, and noise; and

- To explore a method for physical decoupling of the recording sites from the main telemetry platform using the brain as a communication medium.

Chapter 2 of this thesis presents the design constraints and testing of an asynchronous analog front-end that can select 16 channels from 128 sites for simultaneous recording. This analog front-end is an intermediate step toward achieving modularity, low power per channel, small size and small number of connections. Several techniques to reduce noise, current consumption as well as supply voltage will be discussed.

Chapter 3 explores using analog delta compression and digital-assistance techniques to increase the number of channels further by 8 times (128-channels), while only doubling the area. A proposed algorithm for on-chip implementation is discussed as the key to achieving small area and high energy efficiency. Further reduction of lead transfers will be introduced.

Chapter 4 describes the design of a 128-channels monolithic active neural probe with the first on-the-shank analog front-end that operates asynchronously with near-threshold (0.5-V) supply voltage on an SOI 150nm CMOS process.

Chapter 5 concludes the thesis by presenting a summary and contributions of this doctoral work, and suggesting further work for improvement and development of analog front-end circuits.

Finally, appendix A introduces the work that has been done to explore a method for physical decoupling of the recording sites from a telemetry platform using the brain as a communication medium.

CHAPTER 2

CONSTRAINTS AND DESIGN OF AN ANALOG FRONT-END MODULE FOR HYBRID INTEGRATION WITH NEURAL PROBES

In neural recording microsystems with a massive number of parallel channels, the analog front-end (AFE) is becoming the most critical part of circuit design. Noise, area, power, and bandwidth are four fundamental specifications that indicate how much channels a particular AFE can allow when it is part of a limited size microsystem. Continuous improvements of AFEs are required for more practical studies of brain dynamics.

This chapter introduces an analog front-end prototype designed for integration into 3-D neural recording microsystems, i.e., the Brainavigator platform discussed in chapter 1. For scaling towards massive parallel neural recording, the prototype has investigated some critical circuit challenges in power, area, interface, and modularity. The proposed AFE scales down the power consumption to an extent, so that would allow 5x to 20x more simultaneous channels to be recorded. Energy efficiency is increased using several system- and circuit-level techniques including moderate inversion (or near-threshold) operation for all analog and digital circuits, asynchronous digital operation, and dynamic voltage scaling (DVS) with a sampling operation of up to 50kS/s. The AFE also reduces the number of leads for transferring signals and avoid high-speed synchronizing signals by employing an asynchronously-controlled interface and time-division multiplexing (TDM). It is laid-out for direct integration on a neural probe body as a part of 2-D or 3-D neural microsystem, and provides an overall modular design to reduce the overall system complexity and design time, while gaining flexibility in system scalability and expandability. The design strategies are described in details within this chapter after reviewing the circuit requirements for massive-parallel neural microsystems, introducing the state of the art AFEs, and developing the figure of merits appropriate for the assessment of different designs.

2.1 NIMS IC Requirements and Figures of Merit

Understanding the brain dynamics has tremendously improved over the past five decades due to the progress in neural recording techniques [16][73]. In addition to developing ingenious algorithms for spike sorting, the ability to record massively parallel neural activities has been a crucial tool for many complex neuroscientific studies. Similar to Moore's law for transistor scaling, the number of simultaneously recorded channels actually doubles every 7 years [73]. This implies that a few thousand channel recording system should be available in the next two decades. Nonetheless, a leap in the number of simultaneous channels has remained as an unmet need, mainly, due to limitations in the developed integrated circuits (IC). More specifically, there is a need to innovate new IC solutions for reducing power and area while maintaining low noise ($<10\mu V_{rms}$) and enough bandwidth (6~10kHz) for neural recording. Since noise and bandwidth are already given, power and area need to be minimized for more simultaneous channel recording. In the next subsections, power and area impacts on neural microsystems are presented followed by developing a figure of merit chart that allows the assessment of different reported state-of-the-art AFEs and their potential in increasing the number of simultaneously recorded channels.

2.1.1 Power Consumption Effect

The brain has a cooling system that causes its temperature to drop by 1-2°C compared to the core body temperature (~37°C) [74]. A rise in this temperature by more than 1°C is a byproduct of an abnormal condition, such as severe hypoxia caused by near-maximum oxygen metabolism as the experiment reported in [74] suggested. There is a lack in the literature on how to translate this temperature effect into a standardized power density limit, especially for implantable brain microsystems. Nonetheless, other tissue (e.g. muscles) may provide a rough indication of the upper limit of power density. The in-vivo study in [75] showed that muscle tissue heating for several weeks at $80mW/cm^2$ caused necrosis of adjacent tissues. Another study, in [76], simulated the impact of a $4.7x5.9mm^2$ chip with 100 amplifiers at a total of 13mW power consumption and showed that it would impose 0.029°C/mW of heating (total of 0.38°C), which is much lower than the limit described in [74]. The simulated heat impact in [76] can be translated into a power density of ~470μW/mm^2, and a rough approximation of 500μW/mm^2 will be referred to as a safety metric throughout this thesis although it can be very conservative and not proven yet for neural recording.

2.1.2 Integration and Area Consumption Effect

In massively parallel neural recording microsystems, area is a critical specification that affects the power density limitation, the noise performance, and the density of channels in a microsystem. For a leap in the number of simultaneous channels, area consumption per channel must be reduced to minimize the tissue trauma. This highlights the importance of integration of circuitry with the MEMS neural probe platform and requires techniques for minimizing the circuit implementation area.

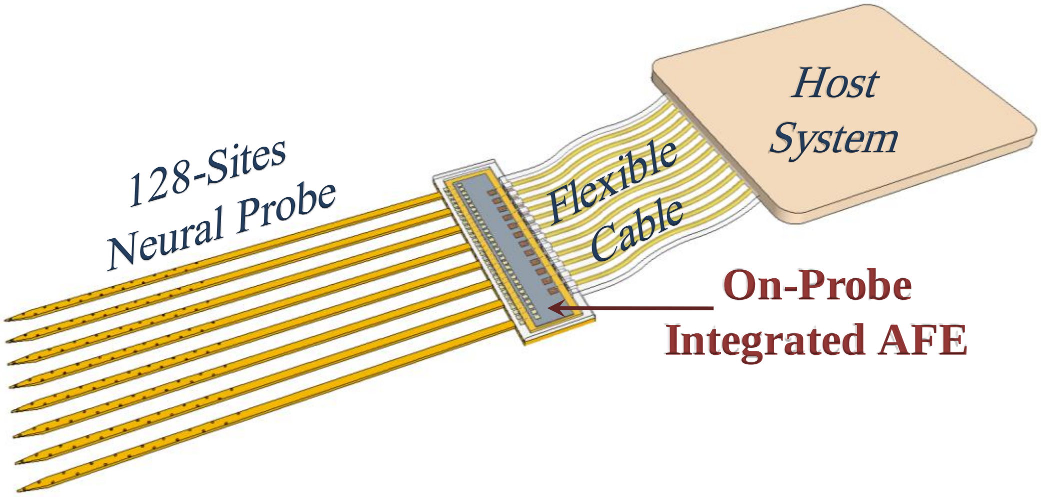


Figure 2.1. 128-sites neural probe with hybrid-integrated analog front-end

Integration of analog front-ends onto neural probes provides two main advantages: 1) Closer location to the electrode site eliminates or shortens the interconnects, thus reducing signal attenuation and noise coupling; and 2) Reduction in the number of wires transferring signals to the host system improves system reliability and form factor as shown in Figure 2.1. However, these benefits come with additional restrictions, to allow feasible implantation of the whole system, as follows:

1) The implementation area is smaller compared to the case when the circuit is not part of the implanted probe;

2) The power consumption should scale accordingly to maintain the safety limit mentioned above;
3) The noise performance may be deteriorated as a result of area and power limitations, although it can relatively improve due to the close location of circuits to the recording site; and
4) Circuit layout techniques are limited by the rectangular shape (large aspect ratio) of the probe's back-end, which may cause undesirable interferences between analog and digital components.

In summary, circuit design for massive parallel neural recording should provide the following features:

1) significant reduction in power and area simultaneously,
2) dynamic scaling of power depending on performance,
3) limited use of high speed signals, such as clocks,
4) hierarchical modular design to allow scalability and expandability, and
5) minimized number of control/acquisition lines for data/commands

Before discussing the details of the proposed approaches to provide these features, the assessment in the figure of merit is discussed in the next section.

2.1.3 Figures of Merit for Neural Recording Integrated Circuits

Assessment of different circuit design approaches requires the development of robust figures of merit that can gauge their potential for increasing the number of simultaneously recorded neurons activities. The following subsections review previously reported integrated circuits and develop a figure of merit (FOM) for the AFE and another FOM for low-noise amplifiers (LNAs).

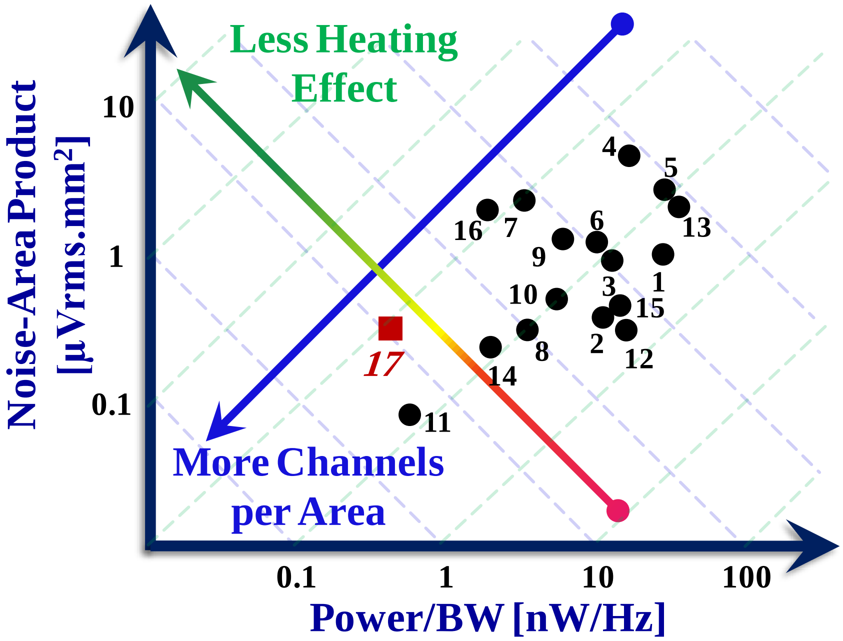
2.1.3.1 AFE Figure of Merit: Energy-Area Efficiency

For analog front-ends, we propose to use a figure of merit that considers power (P), noise (N), bandwidth (BW), and area (A), especially for massive parallel recording where it is essential to integrate more channels per unit area. This figure of merit can be called the energy-area-efficiency (EAE) and can be defined as:

$$EAE = \frac{1}{PBW \cdot NAP} = \frac{BW}{P \cdot N \cdot A} \qquad \textbf{Equ. 2.1}$$

Especially, the trade-off between noise and area is important for capacitive-coupled amplifiers [56]. The higher EAE, the better the design is in terms of including more channels per area with good signal integrity [77].

Figure 2.2 plots the performance of AFEs in two dimensional space of noise-area-product versus power-per-bandwidth, and compares the various state-of-the-art neural recording circuits reported up to date [51, 54, 55, 60, 61, 63, 64, 66, 70-72, 78-82]. For the same noise performance, area can be reduced at the expense of more power and vice versa. The dotted diagonal lines, diverging from the origin of the plot to (1,1) direction (slope of -45^o), provide contour lines of the same number of channels per area. The other set of dotted diagonal lines, perpendicular to the first group (slope of 45^o), provide contour lines of the same power per area (heating effect) for the same BW and N values.



**This graph compares only the Analog front-end*

Author/Year	Publication	Ref
Bai and Wise ’01	TBME	1
Harrison and Charles ’03	JSSC	2
Mohseni and Najafi ’04	TBME	3
Olsson and Wise ’05	JSSC	4
Harrison et. al ’07	JSSC	5
Yin and Ghovanloo ’09	ISSCC	6
Chae et. al ’09	TNSRE	7
Aziz et. al ’09	JSSC	8
Song et. al ’09	TNSRE	9
Perlin and Wise ’10	JMEMS	10
Muller and Rabaey ’11	ISSCC	11
Shulyzki et. al ‘11	CICC	12
Lopez et. al ‘11	ISCAS	13
Wattanapanitch and Sarpeshkar ‘11	TBioCAS	14
Mollazadeh et. Al ‘09	TBME	15
Azin and Mohseni ’11	JSSC	16
Al-Ashmouny et. al ’12	TBioCAS	17

Figure 2.2. Performance chart of the implemented AFE. The chart plots noise-area-product versus power-per-bandwidth and compares the state-of-the-art neural recording circuits reported up to date.

2.1.3.2 For LNA

Low-noise amplifier (LNA) is the most critical part of the neural recording channel due to many reasons: 1) It determines the noise performance of the overall channel; 2) It consumes most of the power and area budget of the channel; and 3) It is almost always considered as an unshared block among multiple channels. Therefore, a special attention needs to be directed when evaluating LNA designs in light of recent challenges especially under low-voltage operation. As more advanced CMOS processes have been developed for more compact digital circuits, operating analog circuits at low voltage is important for the integration into these processes. Nonetheless, for a given signal-to-noise ratio requirement, reducing the supply voltage for analog circuits may not be the best scenario to reduce the total power consumption. Some conventional techniques, such as cascoding, should be avoided. In general, it is likely that low voltage operation may increase the overall power consumption of analog circuits.

As a figure of merit, the noise efficiency figure (NEF) has been introduced in 1987 to compare the performance of different low-noise amplifiers using ~5V supply and 3µm CMOS process [83]. NEF was defined as a ratio between input-referred root-mean squared (RMS) noise voltage and equivalent input-referred RMS thermal noise that would result from a single BJT at the same noise bandwidth (*Δf*), and with its collector current (I_C) equal to the LNA supply current I_{supply}.

For an ideal BJT, with only thermal noise (ignoring the base resistance) the short circuit input referred mean-square noise density and the noise bandwidth assuming single pole system can be expressed as:

$$\frac{\overline{v_i}^2}{\Delta f} = 4kT(\frac{1}{2g_m}) \qquad \textbf{Equ. 2.2}$$

$$\Delta f = \frac{\pi}{2}.BW_{3dB} \qquad \textbf{Equ. 2.3}$$

Therefore, the NEF equation will be:

$$NEF = V_{RMS,in}\sqrt{\frac{I_{supply}}{\pi.U_T.k.T.BW_{3dB}}} \qquad \textbf{Equ. 2.4}$$

However, the actual NEF expression used in neural LNAs has a $\sqrt{2}$ term which is apparently an error in the derivation in [83], as also suggested in [84]. Since NEF has been already used, it will be referred to using the same expression stated in [83]:

$$NEF = V_{RMS,in}\sqrt{\frac{2.I_{supply}}{\pi.U_T.4.k.T.BW_{3dB}}} \qquad \textbf{Equ. 2.5}$$

From Equ. 2.5, it is obvious that NEF is not an appropriate figure of merit as it does not consider:

1) The supply voltage (V_{DD}) which contributes to the overall power consumption. Excluding the supply voltage can actually favor a high voltage design over a low voltage one. Some authors (as in [72]) used a modified version (V_{DD}·NEF^2) to compare their low-voltage amplifier with other work. However, this also may not give fair comparison because it may swing the favor toward low-voltage design.
2) The maximum signal-to-noise ratio (SNR_{max}) or dynamic range (DR) of the LNA. This is naturally dependent on the supply voltage; therefore, reporting only the RMS noise without quantification of SNR may favor a low voltage design over a high voltage design.

Therefore, it is desirable to use another metric (K), which was derived in [85] by E. Vittoz and Y. Tsividis to assess any analog circuitry [85]. The expression for K is given by:

$$K = \frac{P}{k.T.BW.SNR} \qquad \textbf{Equ. 2.6}$$

The SNR can be replaced by the maximum SNR (SNR_{max}) or preferably the dynamic range (DR) especially in case that most designs have signal-independent noise performance and power consumption. Therefore, K can be expressed as:

$$K = \frac{P}{k.T.BW.DR} \qquad \textbf{Equ. 2.7}$$

This figure of merit, K, is obviously unitless, and its minimum achievable value depends on many parameters including the circuit architecture [85]; therefore, it should be reported as is without normalization to a particular value. DR can be evaluated as the maximum SNR when the total harmonic distortion (THD) of LNA is less than 1%. While K will be used as the FOM of LNAs throughout this thesis, $V_{DD}\cdot NEF^2$ will be also reported for convenience.

Now that the appropriate FOMs (EAE and K) are defined, the rest of this chapter introduces an energy-efficient analog front-end (AFE) module that can be integrated into 3-D neural recording microsystems for minimal interface with compact packaging.

2.2 A 0.9V Low-Power 16-Channels Modular Analog Front-End

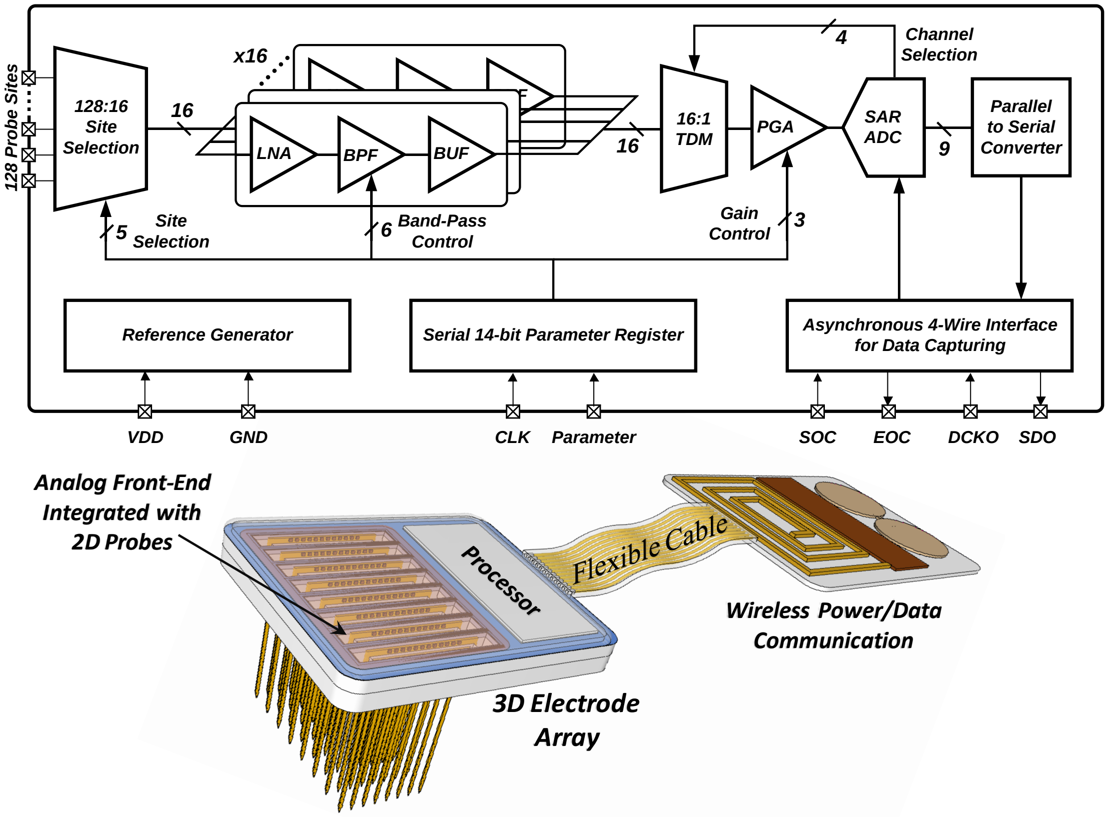


Figure 2.3. Block diagram of analog front-end module with the illustration of integrating in a 3D array system. The front-end module includes on-chip reference generator, site selection (128 to16), and 16 preamplifiers that are multiplexed into one ADC. Data acquisition interface is asynchronous and uses only four pads, and there are extra two pads assigned for clock and loading control parameters to the chip. When multiple 2D modules are assemble in the shown 3D array, only EOC and SDO need separate connections for each module while the other pads are shared with the rest of modules.

Figure 2.3 illustrates the 16-channel AFE (AFE-16) block diagram. It has a site selection block that selectively routes 16 out of 128 sites from the fabricated neural probe to an array of 16 analog channels using a 5-to-32 decoder. Each analog channel consists of a low-noise amplifier (LNA), a band-pass filter (BPF) with 6-bit tunable frequency corners, and a buffer to drive a time-division multiplexer (TDM). The output of the TDM is further amplified by a 3-bit programmable gain amplifier (PGA) that drives an asynchronous 9-bit successive approximation register analog-to-digital converter (SAR-ADC). Data transmission to the next blocks is serially progressed by latching the 9-bit into a parallel-to-serial converter (PSC). The overall data acquisition process is asynchronously-controlled using a 4-wire interface, two of which are unique to each AFE module (AFE-16). In other words, when additional AFE-16 modules are added to the system, only 2 wires are needed per module. The programming of total 14 bits is serially loaded to the module to update the parameters, independent of acquisition process.

The AFE has a site selection block (Figure 2.4) that routes 16:128 sites of the neural probe using a 5-to-32 decoder, to an array that has 16 analog channels. The routing switch network allows zooming into a neural probe by selecting all the 16 sites in one shank, 8 sites from each two neighbor shanks, 4 sites from each four neighbor shanks, or 2 sites from each eight neighbor shanks.

Low voltage (0.9 V) design is pursued to ensure overall energy-efficient mixed-signal operation, especially for further processing blocks are dominated by digital circuits. In addition, the low speed application, which is the case of our implementation, would benefit from the improved digital power-delay performance under low-voltage [86]. Achieving these benefits, however, comes at the cost of more sophisticated analog circuit design to compensate for other circuit parameters including dynamic range, harmonic distortion, noise, gain, CMRR, PSRR, and matching. In our design, we have chosen to multiplex 16 channels (instead of groups of 2, 4, or 8) to achieve the best tradeoff between area and power consumption according to [87]. For convenience, this analog front-end module will be referred to as 16-AHI (16 channels Analog front-end for Hybrid Integration) throughout this thesis.

Techniques and strategies taken to address these challenges in each block of the AFE-16 are described in the following sections.

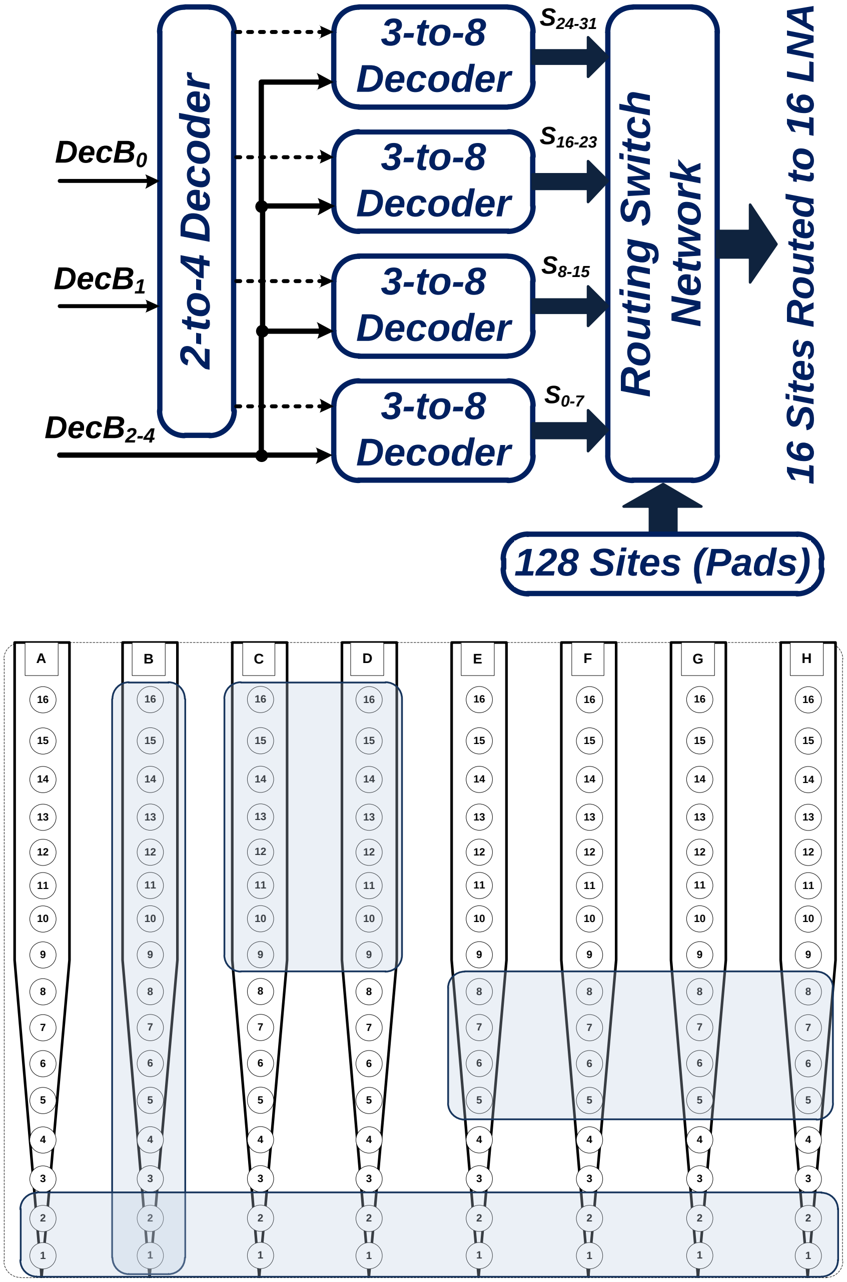


Figure 2.4. Block diagram of the 128-to-16 site selection

2.3 A Sub-1μW Low-Noise Amplifier Operating in Moderate Inversion

As discussed in chapter 1, the understanding of brain dynamics and performance of complex neuroscientific studies rely on the detection of spikes at each electrode in addition to sorting these spikes to represent individual neurons surrounding group of electrodes. Most of developed algorithms use the signals' amplitude as the most significant extracted feature to successfully perform these operations and discriminate different neuronal responses [88]; therefore, channel-to-channel variations in a recording system may corrupt the useful information in the signals and require very complicated algorithm to compensate for that. Apparently, the low-noise amplifier (LNA) contributes to most of inter-channel variation and, therefore, its gain need to be relatively immune to process variations. Minimizing the closed-loop gain error requires an open-loop gain of ~100dB; therefore, the viable option under low-voltage operation is a two-stage operational transconductance amplifier (OTA). For the 16-AHI, the LNA design was optimized in moderate inversion to achieve a high open-loop gain with small headroom. In addition, distribution efficiency of biasing current was maximized based on a predetermined feedback factor in the closed-loop configuration. While the LNA is simply based on a two-stage OTA, the optimization process in moderate inversion allows an average of 1-2 orders of magnitude better performance (based on K) and smaller area than most of the state-or-the-art designs [54, 55, 60, 63, 64, 66, 67, 69, 78, 89-91]. This section describes the design and optimization of the LNA after reviewing and highlighting the pros and cons of moderate inversion over other regions of operation.

2.3.1 A Review of Moderate Inversion Region and Operation Prospective

Most of the analog circuits operate the MOS transistor (MOST) in saturation, where a certain value of the drain-source voltage (V_{DS}) establishes its minimum boundary; therefore, called V_{DSAT}. The value of V_{DSAT} is different for different MOST size and bias current, which also translates into the relative value of gate-source voltage (V_{GS}) to the threshold voltage (V_{th}) of that particular MOST or, in other words, how strong the MOST is inverted. For recognizing the different degrees of MOST inversion, the terms strong inversion (SI) and weak inversion (WI) are widely used since the operation is well-modeled at these regions.

The MOST behavior is dominated by drift current in SI, where the operation offers the best reliability, matching, and speed, but suffers from poor transconductance efficiency (TCE) and large V_{DSAT}. On the other hand, the behavior is dominated by diffusion current in WI, where MOST behaves similar to a bipolar junction transistor (BJT) and offers the best TCE and the lowest V_{DSAT}, but suffers from worst mismatch, and lowest speed. In between these two regions, where the drift and diffusion current are comparable in magnitude, neither the modeling of SI, nor WI can be used. This region, called moderate inversion (MI), was first named and examined by Y. Tsividis in 1982 [92]. Tsividis showed that expressions for either SI or WI to model MI-biased MOSTs can produce serious errors in real measurements. While his objective was to increase the awareness of designers about MI so they can avoid it, he also predicted that MI may have a potential in low voltage analog designs that would be needed to keep track of the advances in digital processes.

Tsividis proposed the channel inversion coefficient (IC) - the ratio between the inversion layer capacitance ($\boldsymbol{C_{inv}}$) and the sum of oxide capacitance ($\boldsymbol{C_{ox}}$) and depletion region capacitance ($\boldsymbol{C_{dep}}$) - as a metric to distinguish the three inversion regions [92]. For moderate inversion:

$$0.1 < IC(= \frac{C_{inv}}{C_{ox}+C_{dep}}) < 10$$ **Equ. 2.8**

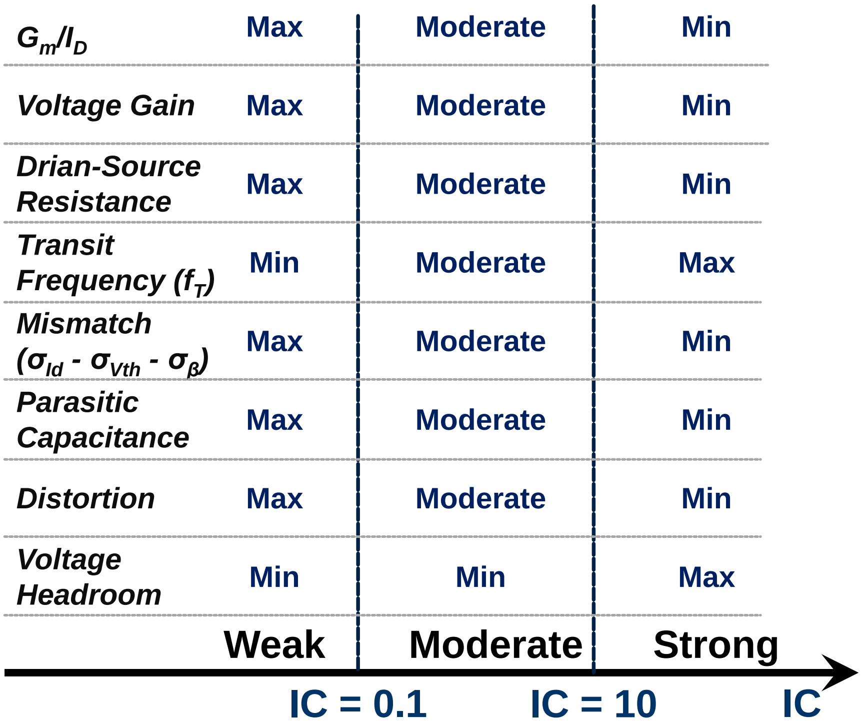
IC can be also expressed as the ratio between the drain current (I_D) of MOST and its specific current (I_S) given by:

$$I_S = 2.n.\beta_{MOS}.U_T^2$$ **Equ. 2.9**

In the previous equation, $\boldsymbol{n}$ is the slope factor (~1.4 in 0.25µm CMOS), $\boldsymbol{U_T}$ is the thermal voltage and $\boldsymbol{\beta_{MOS}}$ is a function of the oxide capacitance per unit area ($\boldsymbol{C_{ox}}$), MOST carrier mobility ($\boldsymbol{\mu}$), and its width ($\boldsymbol{W}$) and length ($\boldsymbol{L}$) as follows:

$$\beta_{MOS} = \mu.C_{ox}.\frac{W}{L}$$ **Equ. 2.10**

While MI was predicted to be of potential use in the low-voltage analog design, very few designs attempted to use it as a replacement for the SI-based designs [93] [94]. None of these approaches addressed the design for low noise amplifier. In [93] simulation results was shown for a 0.9 V class AB OTA designed for switched opamp applications; however, no analysis or emphasis was stated on MI region. In [94] an interesting comparison was performed between two approaches for 0.5V-operated filters' OTAs. While the first approach used the body terminal of PMOS as an input and the gate terminal for biasing, the second approach is vice versa. For either approach, the supply reduction was due to biasing the body terminal to lower the threshold voltage; otherwise, MOSTs would operate in WI rather than MI. However, body biasing may require a triple-well process and suffer from variations, which motivated the authors in [94] to design additional circuitry (error amplifier) to compensate these effects. For all previous approaches, MI or low-voltage design caused the power consumption to increase over their SI counterparts.



G_m/I_D	Max	Moderate	Min
Voltage Gain	Max	Moderate	Min
Drian-Source Resistance	Max	Moderate	Min
Transit Frequency (f_T)	Min	Moderate	Max
Mismatch (σ_{Id} - σ_{Vth} - σ_{β})	Max	Moderate	Min
Parasitic Capacitance	Max	Moderate	Min
Distortion	Max	Moderate	Min
Voltage Headroom	Min	Min	Max
	Weak	**Moderate**	**Strong**

Figure 2.5. Design tradeoffs in weak, moderate, and strong inversion regions. It shows that moderate inversion region has a strong potential for circuit optimization in low voltage operation.

Parameter	***Strong Inversion***	***Weak Inversion***
Gate Drive Voltage	$V_{GT} = V_{gs} - V_{th} > 0.2$	$V_{GT} = V_{gs} - V_{th} < 0$
Drain Current	$I_D = \frac{\beta}{2}.(V_{gs} - V_{th})^2(1 + \lambda V_{ds})$	$I_D = \ 2n.\beta.U_T^2 e^{\frac{V_{gs}-V_{th}}{nU_T}}(1 - e^{\frac{-V_{ds}}{U_T}})$
Transconductance	$g_m = \frac{\partial I_D}{\partial V_{gs}} = \sqrt{2.\beta.I_D}$	$g_m = \frac{\partial I_D}{\partial V_{gs}} = \frac{I_D}{n.U_T}$
Saturation V_{DS} (V_{DSsat})	$V_{DSsat} \geq 0.2V$	$V_{DSsat} \geq 0.1V$ *or* $\sim 4U_T$
Drain-Source Resistance[4]	$r_{ds} = \frac{1}{\lambda I_D}$ *smallest*	$r_{ds} = \frac{1}{\lambda I_D}$ *largest*

Table 2.1. Equations for strong and weak inversion

While the usefulness of MI was proposed 30 years ago, still robust design approaches and analysis for low-noise circuits has not been fully explored. For convenience, Table 2.1 lists the equations usually used in the designing strong and weak inversion biased MOST, and Figure 2.5 shows the trade-offs between all three regions of operation, which may suggest a wide range of possible optimization processes for analog design to appropriately set the IC of each MOST. The next section proposes a low-noise two-stage OTA with all MOSTs, except input differential-pair, biased in MI with an optimized IC using the properties in Figure 2.5.

[4] Channel length modulation (CLM) factor, λ, is a strong function of channel length and becomes more effective for higher V_{GS} [127], [128]

2.3.2 Design of Moderate-Inversion Two-Stage OTA

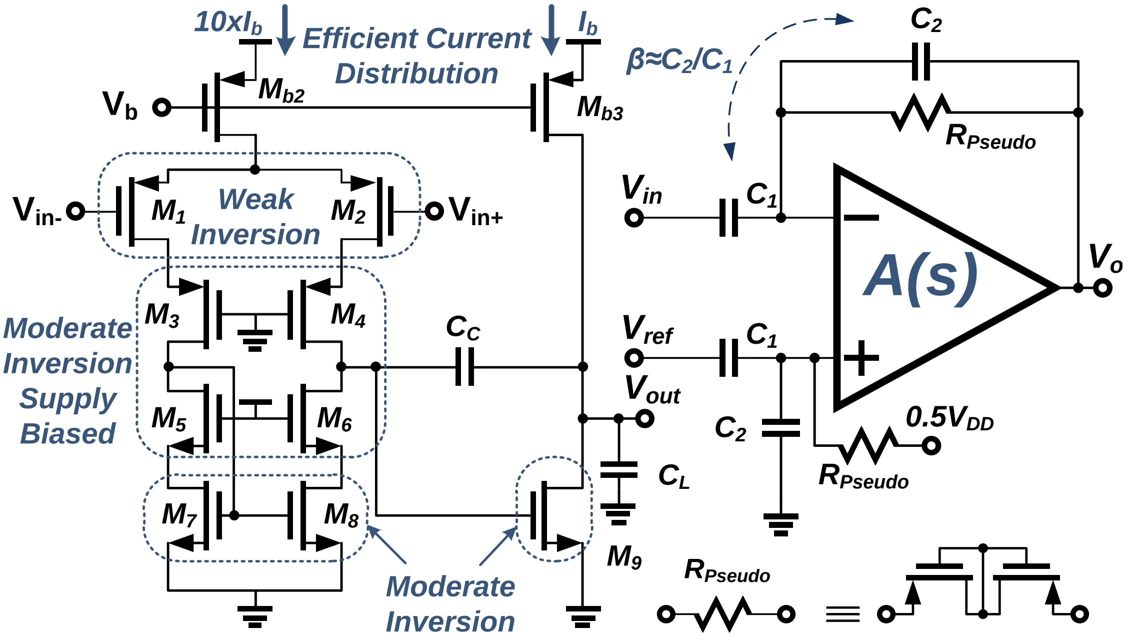


Figure 2.6. Schematic of a moderate-inversion low-noise amplifier. All the transistors are biased in moderate inversion except for a differential pair.

The noise performance of an LNA can be conventionally improved by biasing the input differential pair in WI and other devices in SI as implied in Figure 2.5. However, this contradicts with securing the voltage headroom when the supply voltage is low. Figure 2.6 shows the proposed LNA, where no MOST is biased in SI and, therefore, the headroom is reduced by an average of 50% (~100mV). Although the voltage headroom is a critical requirement, non-optimized parameters such as power, noise, and gain would make the LNA ineffective; therefore, careful analysis needs to be performed to bias each MOST with the appropriate IC.

For effective use of moderate inversion, its two decades of IC span were divided into low-MI (MIL from 0.1 to 1), and high-MI (MIH from 1 to 10). The following design criteria were pursued based on the properties in Figure 2.5 and influenced by the particular function of each MOST:

1) ***Input differential pair*** (M_1 and M_2) were biased in deep weak inversion (IC < 0.01) for highest TCE, gain, and output resistance ($\boldsymbol{r_o}$ or $\boldsymbol{r_{ds}}$), while increasing the area (WL product) for improved matching and flicker noise.

2) ***Mirror devices*** (M_7 and M_8) were biased in the MIH region (closer to SI) to achieve lower TCE and good current matching. They were also designed with large channel length (***L***) to improve $\boldsymbol{V_{th}}$-matching and flicker noise. Moderate inversion offers an advantage in this particular case since it features the minimum $\boldsymbol{C_{gs}}$ (Figure 2.5). Thus, for the same mirror pole location, MI allows larger ***L*** and $\boldsymbol{r_{ds}}$ or smaller bias current as compared with SI and WI.

3) ***Cascode devices*** (M_3-M_6) were biased in the middle of MI for a better tradeoff between large $\boldsymbol{r_{ds}}$ and low TCE. They are also sized to be biased by the supply rails for less complex biasing circuit. M_5 and M_6 were also used in a low-voltage cascade configuration for higher overall $\boldsymbol{r_o}$ of the first stage OTA.

4) ***Current biasing devices*** (M_{b2}, and M_{b3}) were biased in MIL with large ***L*** to achieve the largest possible $\boldsymbol{r_{ds}}$ and improve both common-mode rejection ratio (CMRR) and power supply rejection ratio (PSRR) in the band of interest.

5) ***Input of the 2nd stage OTA*** (M_9) has the same bias point of the mirror devices with 10x reduced current that can achieve: 1) high 2nd stage gain for effective input/output poles splitting, 2) small power consumption, and 3) small enough $\boldsymbol{r_{ds}}$ to guarantee appropriate output pole location with C_L of 2pF.

Device	*Size*	I_D *[nA]*	*IC*
M_1, M_2	660/0.6	500	0.006
M_3, M_4	10/1	500	0.7
M_5, M_6	10/6	500	1.04
M_7, M_8	40/48	500	2
M_9	8/48	100	2
M_{b2}	160/5	1000	0.44
M_{b3}	16/5	100	0.44
C_C	2.3 pF	-	-
C_1	110*C_2	-	-
C_2	76 fF	-	-
C_L	2 pF	-	-

Table 2.2. Design parameters for the proposed LNA

Table 2.2 shows the design parameters of the proposed LNA including MOSTs sizes, bias currents, and IC values, which were based on the devices functions in addition to the analysis of gain response, stability, slew rate, and noise as further described in the next subsections.

2.3.3 OTA Frequency Response Analysis: Gain, Phase, and Noise

The use of MI contributes to supply voltage reduction of LNA, and further helps increase the channel length (and width) of mirror devices, which results in improvement of both flicker noise and mismatching, while not severely affecting the mirror pole location. For power reduction, the second stage should be designed to consume the lowest current (*I_{b3}*) that can guarantee a phase margin (~80°) for stability. The proposed LNA has a target open-loop gain of 110-120 dB, split in the two stages, and should give a stable frequency response (poles and zero locations) and an efficient current ratio *R* (=*I_{b2}*/*I_{b3}*). The optimization of *R* must be directly influenced by the feedback factor (*β*) of the closed-loop implementation (Figure 2.6) given by:

$$\beta = \frac{C_2}{C_1+C_2} \approx \frac{C_2}{C_1} = 0.01$$ **Equ. 2.11**

The OTA input-referred noise is mainly dependent on the first stage, whereas the noise of the second stage is divided by the gain of the 1st stage and can be ignored during the analysis. Therefore, the mean square thermal noise density can be written as:

$$\frac{\overline{v_n^2}}{\Delta f} \approx \frac{8.k.T.\gamma}{g_{m1}}.(1+\frac{g_{m7}}{g_{m1}}),\ \ \gamma = \frac{2}{3}$$ **Equ. 2.12**

The cascode transistors can be ignored at low frequency since the mirror devices already determine the current through them. Since all MOSTs are biased in MI or WI, the transconductance (**g_m**) is determined by the IC using EKV model [95] as follows:

$$g_m = m.\frac{I_D}{n \cdot U_T}$$ **Equ. 2.13**

$$m = \frac{2}{1+\sqrt{1+4.IC}}$$ **Equ. 2.14**

Assuming IC of M$_1$ and M$_2$ is too small (due to WI biasing), it can be proven that:

$$g_{m3} = m_3 . g_{m1} \qquad \textbf{Equ. 2.15}$$

$$g_{m5} = m_5 . g_{m1} \qquad \textbf{Equ. 2.16}$$

$$g_{m7} = m_7 . g_{m1} \qquad \textbf{Equ. 2.17}$$

$$g_{m9} = \frac{2.m_9}{R} . g_{m1} \qquad \textbf{Equ. 2.18}$$

Therefore, Equ. 2.12 can be rewritten as:

$$\frac{\overline{v_n^2}}{\Delta f} = \frac{8.nU_T k.T.\gamma}{I_{D1}} . (1 + m_7) \qquad \textbf{Equ. 2.19}$$

For a current budget of 1µA and a target noise density of 50nV/$\sqrt{Hz}$, $\boldsymbol{g_{m1}}$ is ~13.7µS and $\boldsymbol{m_7}$ (same as $\boldsymbol{m_9}$) should be less than 0.55 ($\boldsymbol{g_{m7}}$ < 7.5µS and IC$_7$ > 1.49), which makes it viable to use MI for the current mirror. Integrating on a 9 kHz bandwidth and assuming a dominant pole, the overall root-mean-square noise will be ~6µV$_{\text{rms}}$. Flicker noise is minimized by maximizing differential pair area and the length of the mirror MOST. The feedback impact on the noise can be ignored since β is very small.

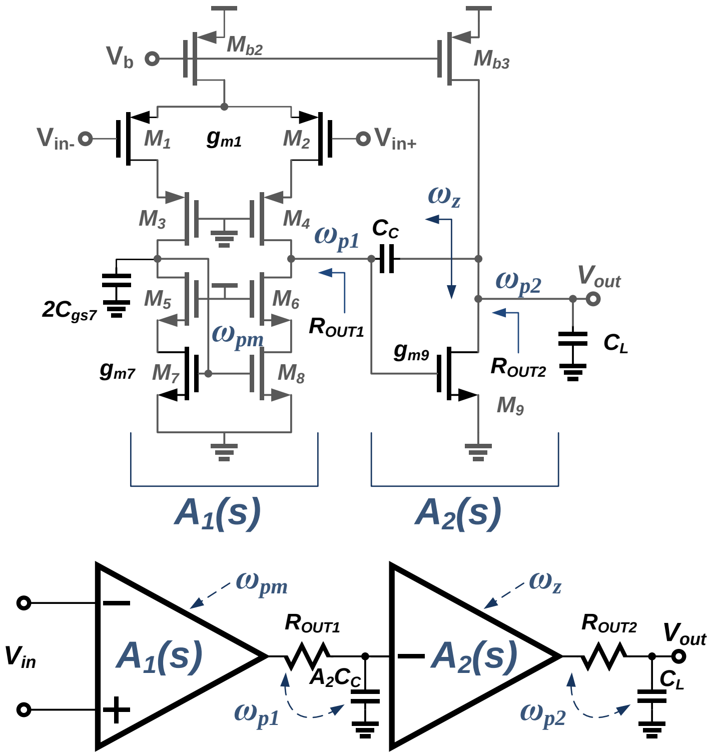


Figure 2.7. Open-loop LNA showing poles and zeros

The dc gain of the two stages and the locations of poles and zeroes can be determined by the following equations (Figure 2.7):

$$A_1 = g_{m1}.R_{OUT1}$$ **Equ. 2.20**

$$A_2 = g_{m9}.R_{OUT2} = \frac{2.m_9}{R}.g_{m1}.R_{OUT2}$$ **Equ. 2.21**

$$\omega_{p1} = \frac{1}{(1+A_2)C_c R_{OUT1}} \approx \frac{R}{2.m_9.g_{m1}R_{OUT2}C_c R_{OUT1}}$$ **Equ. 2.22**

$$\omega_{pm} = \frac{g_{m7}}{2C_{gs7}+C_{gd5}+C_{gd3}} \approx \frac{m_7 g_{m1}}{2C_{gs7}}$$ **Equ. 2.23**

$$\omega_z = \frac{g_{m9}}{C_c+C_{gd9}} \approx \frac{2.m_9}{R}\frac{g_{m1}}{C_c}, \text{(RHP Zero)}$$ **Equ. 2.24**

$$\omega_{p2} = \frac{g_{m9}}{C_L+C_{gs9}} \approx \frac{2.m_9}{R}\frac{g_{m1}}{C_L}$$ **Equ. 2.25**

For achieving an overall open-loop gain of 110 dB, and a closed-loop gain and bandwidth (***ω_{3dB}***) of 40.8dB (110x), and 9 kHz, respectively, the known and bounded parameters in the previous equations are:

1) The dominant pole (***ω_{p1}***) should be located at ~3Hz
2) The OTA transconductance (***g_{m1}***): determined by the current budget (~13.7µS)
3) The load capacitor ***C_L***: determined by the next circuit block (2pF).
4) The values of ***m_7*** and ***m_9*** are equal due to mirroring between the first and second stage. The upper limit is determined by the noise performance (m < 0.55). While lowering ***m_7*** (high IC) worsens the poles splitting (for stability), increasing it requires higher power consumption for the target noise. Therefore, a value of 0.5 is proposed, which is very close to the upper bound proposed by the noise performance. This implies that the width-to-length ratio of M_7 needs to be ~0.8 (based on the model and process curves of NMOS in 0.25µm CMOS technology), but the effective length and mobility need to be considered as well.
5) The compensation capacitor ***C_C*** is highly dependent on the slew rate while affecting the poles splitting. For the target slew rate (9mV/µs) it should be less than 111pF! This indicates that the value will be actually determined by the poles splitting requirement. Pole splitting should be also improved by increasing ***R_{OUT2}*** (i.e. high ratio ***R***) and using the low-voltage cascode to increase ***R_{OUT1}***.

6) There is also a clear boundary on the value of the current ratio ***R*** since its reduction will deteriorate energy efficiency while increasing it will cause low second stage gain and poor poles splitting. Therefore, it is further determined by the stability performance.

So far the parameters available for optimization are ***R***, ***R_{OUT1}***, ***R_{OUT2}***, and ***C_C***. For a complete analysis, the stability requirements have to be addressed before finalizing these values. From Figure 2.6 and Figure 2.7, the transfer function of the closed-loop configuration can be written as:

$$A_{CL}(s) = \frac{V_{out}(s)}{V_{in}(s)} = \frac{A_1(s)A_2(s)}{1+\beta.A_1(s)A_2(s)} \qquad \textbf{Equ. 2.26}$$

For a dominant-pole system, the non-dominant poles ($\boldsymbol{\omega_{p2}}$ and $\boldsymbol{\omega_{pm}}$) and right-half-plane (RHP) zero ($\boldsymbol{\omega_z}$) can be grouped (assuming a dominant-pole system) as:

$$\boldsymbol{\omega_{eq} = (\frac{1}{\omega_{p2}} + \frac{1}{\omega_{pm}} + \frac{1}{\omega_z})^{-1}} \qquad \textbf{Equ. 2.27}$$

Therefore, the two-stage open-loop gain can be written as:

$$A(s) = \ A_1A_2.\frac{1}{(1+\frac{s}{\omega_{p1}})(1+\frac{s}{\omega_{eq}})} \qquad \textbf{Equ. 2.28}$$

The closed loop is unstable if the loop gain (LG) has amplitude of 1 and a phase shift of 180°. In the bandwidth of interest, LG and A_{CL}(s) can be approximated as:

$$LG = \beta A(s) \approx \ \beta A_1A_2.\frac{1}{(1+\frac{s}{\omega_{p1}})} \qquad \textbf{Equ. 2.29}$$

$$A_{CL}(s) \approx \frac{1}{\beta}\frac{1}{1+\frac{s}{\beta A_1A_2\omega_{p1}}} \qquad \textbf{Equ. 2.30}$$

While A(s) and LG has the same phase response, the unity-gain frequency (UGF) of LG is $\boldsymbol{\beta}$ times UGF of A(s) as realized from Equ. 2.29and Equ. 2.30. Therefore, a small value of $\boldsymbol{\beta}$ helps in effective pole-splitting or stability when it brings the UGF of the LG further to the left of the equivalent non-dominant pole ($\boldsymbol{\omega_{eq}}$) in the system. In other words, it supports the claim that the system has a dominant-pole. It can be seen also from Equ. 2.31 that the UGF of LG is approximately the same as the bandwidth ($\boldsymbol{\omega_{3dB}}$). Therefore, the LG can be approximated at frequencies beyond the dominant pole as:

$$LG = \beta A_1 A_2 . \frac{1}{(1+\frac{s}{\omega_{p1}})} . \frac{1}{(1+\frac{s}{\omega_{eq}})} \approx \frac{\omega_{3dB}}{s} . \frac{1}{(1+\frac{s}{\omega_{eq}})} \qquad \textbf{Equ. 2.31}$$

Thus the phase margin (PM) is:

$$PM = 90° - tan^{-1} \frac{\omega_{3dB}}{\omega_{eq}} \qquad \textbf{Equ. 2.32}$$

A PM of ~70° can be achieved if the equivalent non-dominant pole (***ω_{eq}***) is more than ~2.75 times the bandwidth (***ω_{3dB}***) of the closed-loop LNA. Thus, for a target bandwidth of 9kHz, ***ω_{eq}*** needs to be more than ~25kHz and the following set of conditions need to be satisfied based on the parameters already determined (***g_{m1}***, ***m_7***, ***m_9***, ***ω_{p1}***, and ***C_L***):

1) For the open-loop gain of 110dB (Equ. 2.20 and Equ. 2.21):

$$\frac{R_{OUT1} R_{OUT2}}{R} \geq 1.53 \times 10^{15} \qquad \textbf{Equ. 2.33}$$

Since M_7 and M_9 have the same IC, it can be roughly estimated that (assuming no cascode in the first stage):

$$R_{OUT1} \approx 2 \frac{R_{OUT2}}{R} \Rightarrow R_{OUT1} > 55M\Omega \qquad \textbf{Equ. 2.34}$$

This also implies that the gain of the first stage is larger than the second stage. Achieving this ***R_{OUT1}*** value, however, or higher would require cascoding and a long length mirror devices.

2) The location of ***ω_{p1}*** implies that (Equ. 2.22):

$$\frac{R}{C_C R_{OUT1} R_{OUT2}} \approx 284 \times 10^{-6} \Rightarrow C_C \leq 2.3pF \qquad \textbf{Equ. 2.35}$$

3) For stability, the location of ***ω_{eq}*** implies that (Equ. 2.27):

$$4C_{gs7} + R(C_C + C_L) < 266pF \ \Rightarrow C_{gs7} < 66.5 - R \qquad \textbf{Equ. 2.36}$$

For a negligible power consideration of the second stage and to achieve reasonable values of ***R_{OUT1}*** and ***R_{OUT2}***, ***R*** of 10 is chosen (total power consumption of 0.99µW). This indicates (from Equ. 2.36) that the length of M_7 should be less than 100µm; half this value is chosen to and the corresponding width to maintain ***m_7*** (or IC of 2) is about 40µm. While a cascode was used in the first stage to maximize ***R_{OUT1}*** (~66MΩ), the small current of the second stage helped achieve ***R_{OUT2}*** of ~250MΩ. A gain of 60dB and 50dB were achieved at the first and second stage, respectively. The non-dominant poles and the zero were located at 111kHz, 550kHz, and 97kHz for ω_{p2}, ω_{pm}, and ω_z, respectively. The equivalent non-dominant pole (ω_{eq}) is, therefore, located at ~47kHz and PM of ~80° is achieved.

Table 2.2 shows the design parameters of the proposed LNA including MOSTs sizes, bias currents, and IC values based on the devices functions in addition to the analysis of noise, gain, and phase responses. The next section shows the simulation results achieved based on the analysis in this section.

2.3.4 Simulation Results

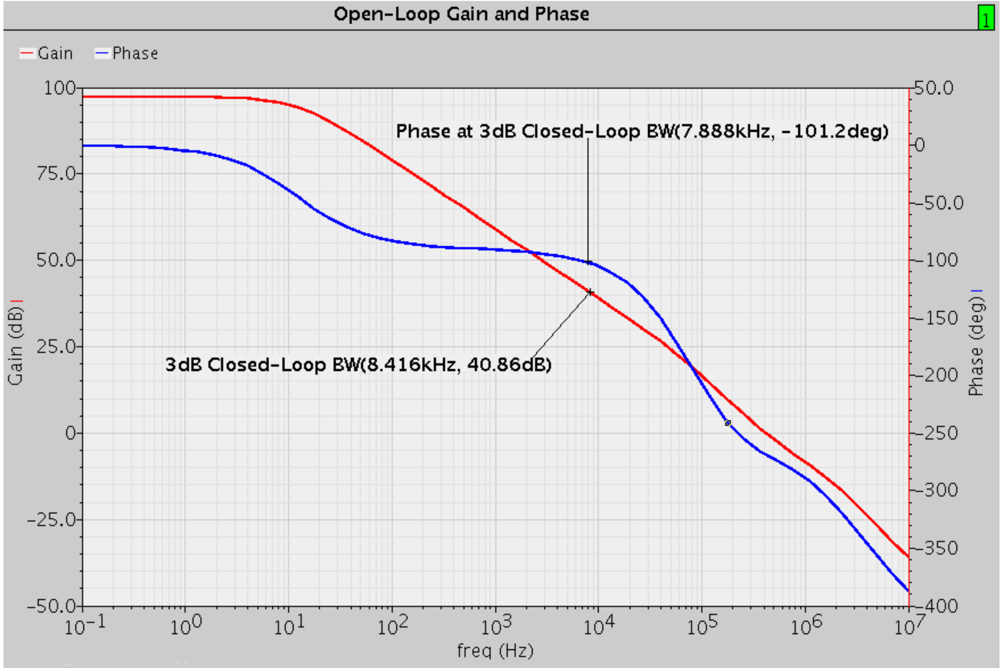


Figure 2.8. Low-noise amplifier open-loop gain and phase response

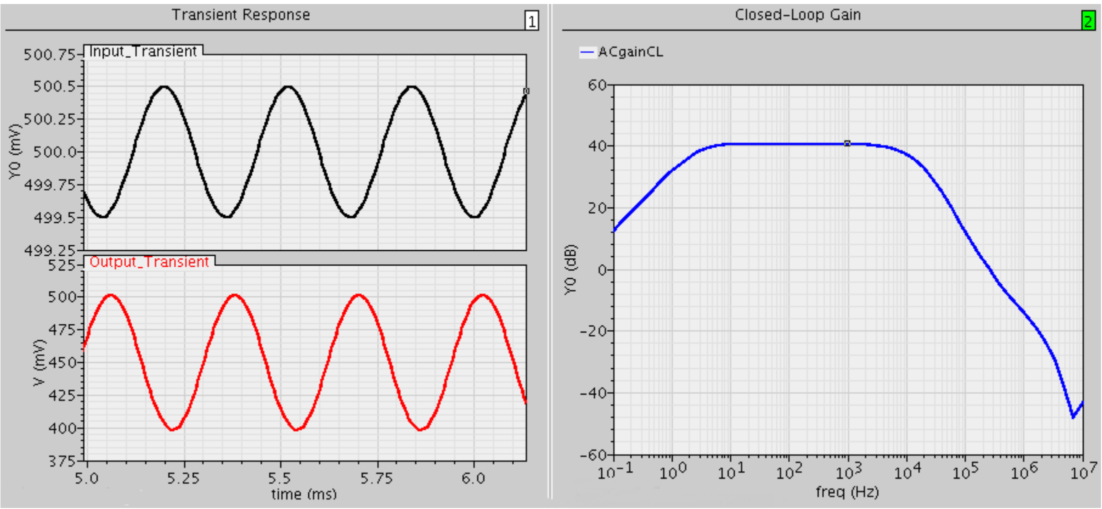


Figure 2.9. Low-noise amplifier closed-loop transient and frequency response

Figure 2.8 shows an open-loop gain of ~100dB and phase margin and the 3dB closed –loop bandwidth of ~80°, while Figure 2.9 shows a closed-loop gain of ~110x (40.82dB) and the transient response at ~3kHz. The CMRR and PSRR are above 60dB in the 3dB bandwidth.

2.4 A 124nW Band-Pass Filter with 6-Bit Tunable Frequency Corners

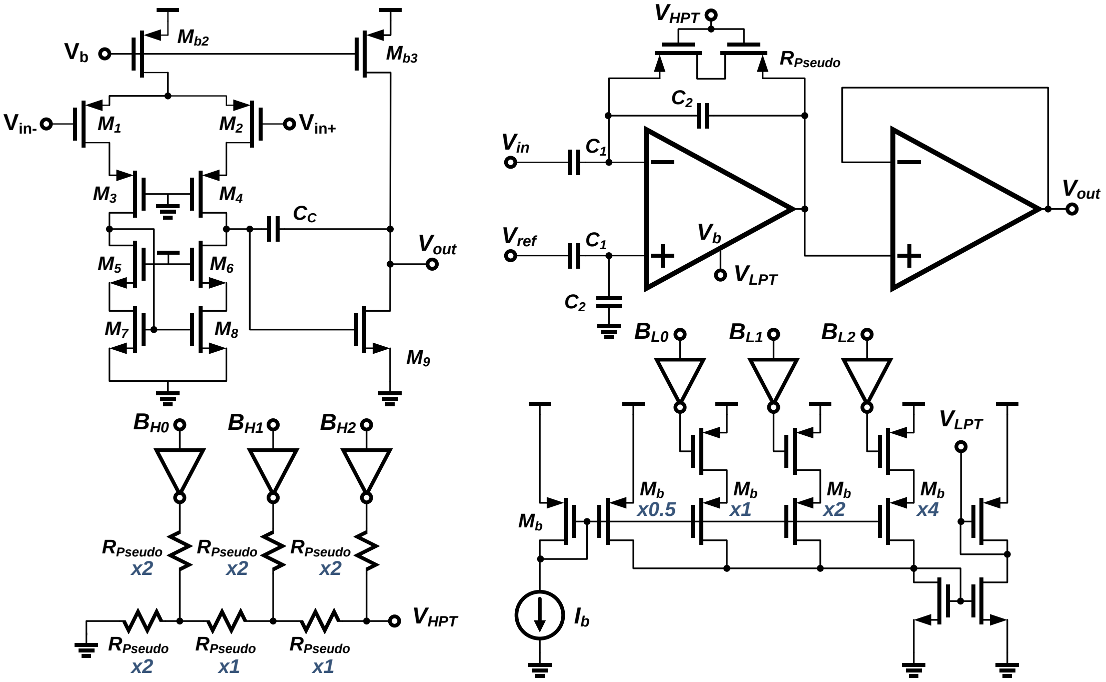


Figure 2.10. Band-pass filter schematic with 6-bit tunable corners

Device	***Size***	***I_D [nA]***	***IC** [x10^{-3}]*
M_1, M_2	100/0.6	17	6.8x10^{-4}
M_3, M_4	10/5	17	0.06
M_5, M_6	1/5	17	0.14
M_7, M_8	1/10	17	0.29
M_9	6/10	102	0.29
M_{b2}	10/10	34	0.24
M_{b3}	30/10	102	0.24
C_C	1.6 pF		
C_1	2*C_2		
C_2	502 fF		
C_L	~0.2 pF (buffer input)		

Table 2.3. Design parameters for the proposed BPF

The LNA of each channel is followed by a band-bass filter (BPF, Figure 2.10) that provides digitally-tunable frequency response. The BPF provides a 6-dB additional gain and was designed with similar criteria used for the LNA (same open-loop architecture, with a closed-loop capacitive feedback). Since noise is not an issue, power was reduced by biasing all MOSTs in WI (all IC < 0.001 as in Table 2.3).

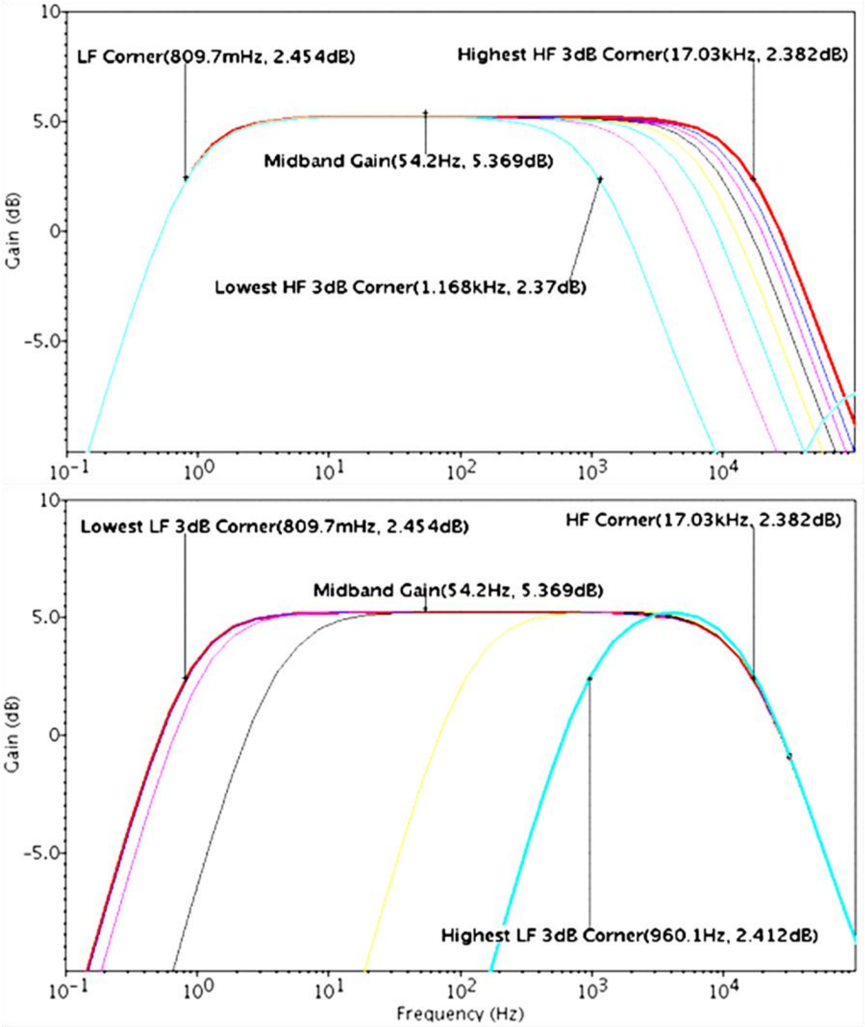


Figure 2.11. Simulation of BPF gain response showing large tuning range of frequency corners

According to the application, neuroscientists may have an interest to tune the frequency of the recorded signal. The two broad signals of interest are local field potential (LFP) and extracellular action potential (EAP) or simply called “spikes”. For LFP recording a high-frequency cut-off at 300Hz and a low-frequency cut-off at ~1Hz are required, given the small electrode size (10-20µm in diameter) in the neural probe. For EAP recording and spike sorting, however, a high-frequency cut-off at 6kHz and a low-frequency cut-off at 600Hz are required. To satisfy the required tunable frequency corners, under process variation, a 6-bit tuning circuit was used as shown in Figure 2.10. The high-frequency cut-off is tuned by a 3-bit current DAC (digital-to-analog converter) that biases the open-loop OTA with a total current ranging from 9 to 138nA, and the low-frequency cut-off is tuned by a 3-bit pseudo-R-2R DAC that biased the feedback pseudo-resistor. This 6-bit tuning can achieve, in worst case simulation (Figure 2.11), a high-frequency cut-off as high as 17kHz and as low as 1kHz, while the low-frequency cut-off can be as low as 1Hz and as high as 1kHz. The BPF consumes 8 to 124nW depending on the selected high-frequency corner. A 1.7µW buffer is used after each BPF to drive the ADC-controlled time-division multiplexing (TDM) block as discussed next.

2.5 Time-Division Multiplexing

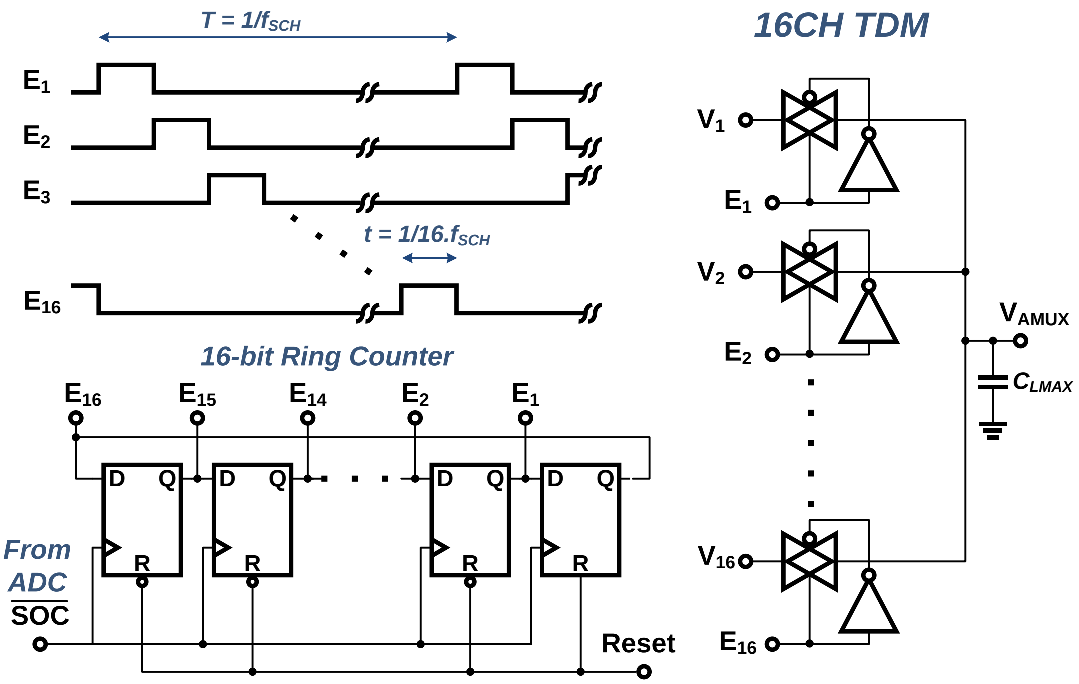


Figure 2.12. 16-channel time-division multiplexer controlled by *SOC* signal from the ADC

Area and power consumption can be effectively reduced by multiplexing channels into a few leads using time-division multiplexing (TDM) [64, 66]; otherwise, it would be hard to access multiple simultaneous channels as reported in [55], where only one full-bandwidth channel can be processed at a time even though 100 amplifiers are available on chip. Therefore, it is very challenging to avoid channels multiplexing in massive parallel neural recording.

In AHI-16, the output of the 16 buffers (**V_1-V_{16}**) is multiplexed into the input of the PGA (**V_{AMUX}**) using a 16-bit ring counter that enables one of the 16 transmission gates in a sequence controlled by the start of conversion (SOC) of the ADC as shown in Figure 2.12. The D-flip flops in the ring counter are based on true single-phase clock (TSPC) architecture as in Figure 2.13.

The analog buffer current consumption is determined based on the slewing and settling requirements of the TDM. Assuming rail-to-rail swing, the slew rate should satisfy the enabling time-window allowed for each channel (***$t = 1/16f_{SCH}$***) which is the reciprocal of 16 times the sampling frequency per channel (***f_{SCH}***). Therefore, for a 20kHz sampling the time slot (***t***) is 3.125µs, the slew rate requirement is 0.288V/µs. For a maximum load capacitance ($\boldsymbol{C_{LMAX}}$) of 3pF, the current consumption should be more than 0.9µA. To enable a sampling of as high as 40 kHz per channel, the buffer is designed to consume ~1.9 µA.

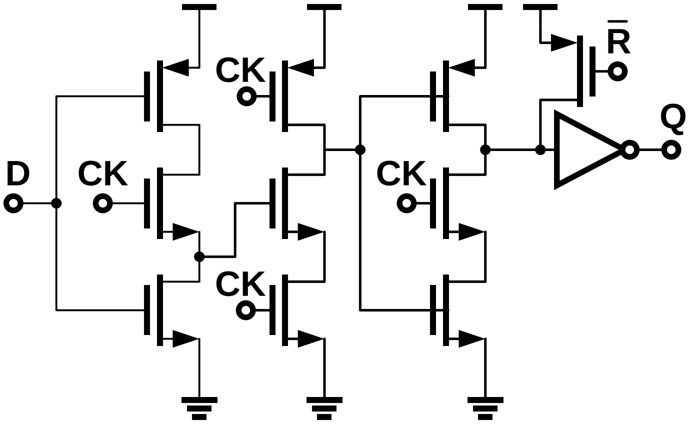


Figure 2.13. True single-phase clock edge-triggered D flip flop

For the settling requirements, the sizing of the transmission gates should be considered, especially in low voltage operation. To illustrate this, Figure 2.14 shows the resistance (normalized to the MOST width) of a transmission gate as a function of the input voltage using different supply voltages (0.9 to 2.5V) and the maximum expected resistance versus the supply voltage. It can be seen that it is very challenging to use a small size transmission gate (TG) for analog inputs. To determine the sizing of the TG, the expected voltage after the enabling time ***t*** is given by:

$$V_i(t) = V_i(\infty) + \left(V_{i-1}(t) - V_i(\infty)\right).\, e^{-\frac{t}{R_{TG}.C_{LMAX}}} \qquad \textbf{Equ. 2.37}$$

Assuming a worst case that a full swing would occur, then t has to be more than 5 times the time constant ($\boldsymbol{R_{TG}.C_{LMAX}}$) for less than 1% error. In other words, $\boldsymbol{R_{TG}}$ has to be less than ~208kΩ. As shown in Figure 2.14, the peak resistance at 0.9V supply is ~2MΩ; therefore, a sizing of at least 11 times (W of 11µm) is required to satisfy the settling requirements. In order to leave some margin for variations, the TG has been designed with a width of 20µm for both NMOS and PMOS; this gave ~50kΩ peak resistance in simulation, which enables more than 40kHz sampling per channel.

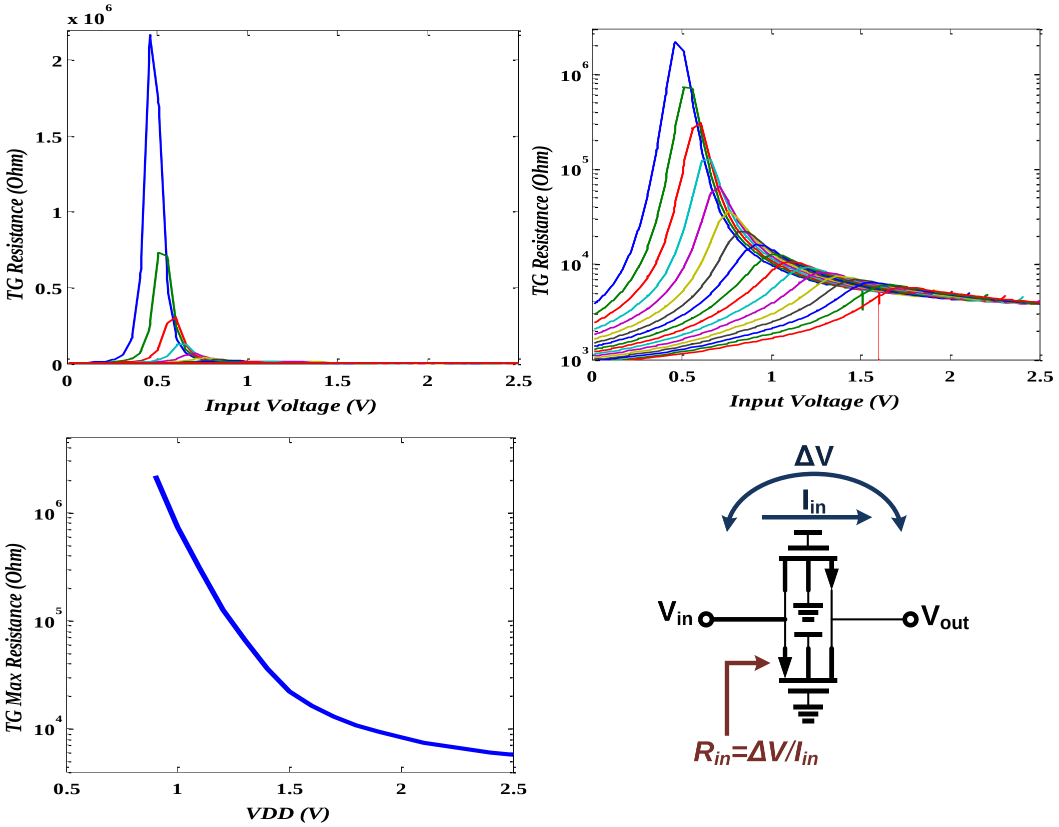


Figure 2.14. Resistance of transmission gate configuration versus input and supply voltages

The output of the 16:1 TDM block would have a maximum useful frequency components of 16 times the maximum frequency component of an individual channel; therefore, the analog-to-digital converter (ADC) driver needs to operate with a minimum of 96kHz bandwidth under worst case simulation corner (assumingh 6kHz per channel). The next section describe an energy-efficient programmable-gain amplifier (PGA) designed to account for the input high-frequency signal and drive the ADC accordingly.

2.6 A 8.2µW 3-Bit Programmable Gain Amplifier for a 16:1 Time-Division Multiplexed Input

The effective system resolution is maximized if the full-scale range of the ADC can be used; therefore, it is important for neural recording circuits to provide a wide-range of gain selectivity to accommodate not only the large-amplitude LFP signals but also the small-amplitude EAP that may be as small as 100µV depending on the subject of recording as well as the electrode impedance. Designs that do not have this capability (e.g. [55]) are subject to a 1-2 bits degraded resolution when the maximum input range is only 50-25% of the ADC's full-scale range.

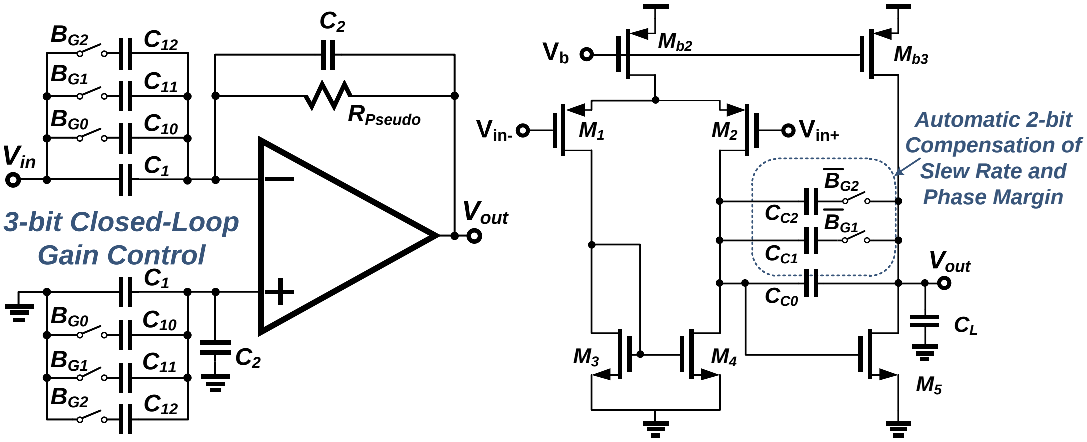


Figure 2.15. Schematic of 3-bit programmable gain amplifier

On the other hand, the power overhead per channel of the TDM multiplexing and the following circuits should be smaller than individual blocks in the channel itself, such as the LNA, to maximize the overall energy efficiency. However, most reported designs focused on lowering the power of LNA and paid less attention to optimizing other circuit blocks. For example, in [64] the authors designed an LNA that consumes a power of 6.6µW, while the 2nd stage amplifier (PGA) following the TDM consumes ~8.4µW per channel.

In the proposed PGA (Figure 2.15), a gain of 6-29dB (2x to 28x) is programmed by changing the feedback factor (β) using a 3-bit (***B***B_G*<0..2>*) input capacitors bank while the feedback capacitor is fixed [96]. This main motivation to change the input, not the feedback, capacitors is that integer gain values can be achieved in a minimum area. The energy efficiency of the PGA is maximized using two major techniques: 1) all MOSTs are biased only in moderate or weak inversion, and 2) the slew rate (SR) and open-loop frequency response (pole-zero locations) are adjusted automatically during gain programming using a 2-bit compensation capacitor bank to provide adequate driving capability of the PGA while maintaining stability.

Device	***Size***	***I_D [μA]***	***IC***
M_1, M_2	400/0.3	2.475	0.016
M_3, M_4	12/1	2.475	0.48
M_5	18/1	3.7	0.48
M_{b2}	100/2	4.95	1
M_{b3}	75/2	3.7	1
C_C	0.4-1.8 pF		
C_L	1 pF (ADC input)		
C_2	100 fF		
C_1	2*C_2		
C_{10}	2*C_2		
C_{11}	8*C_2		
C_{12}	16*C_2		

Table 2.4. Design parameters for the proposed PGA

B_{G2}	***B_{G1}***	***B_{G0}***	***Gain [V/V]***	***Gain [dB]***	***β***	***C_C [pF]***	***ω_{p1} [kHz]***	***ω_z [MHz]***	***ω_{eq} [MHz]***	***PM***	***SR [V/μs]***
0	***0***	***0***	2	6	0.33	1.8	0.44	6.6	4.1	70.5	2.75
0	***0***	***1***	4	12	0.2	1.8	0.44	6.6	4.1	77.9	2.75
0	***1***	***0***	10	20	0.091	1.4	0.57	8.5	4.8	83.8	3.5
0	***1***	***1***	12	21.6	0.077	1.4	0.57	8.5	4.8	84.8	3.5
1	***0***	***0***	18	25.1	0.053	0.8	1	15	6.3	85.2	6.2
1	***0***	***1***	20	26	0.048	0.8	1	15	6.3	85.6	6.2
1	***1***	***0***	26	28.3	0.037	0.4	2	30	8	84.7	12.4
1	***1***	***1***	28	29	0.034	0.4	2	30	8	85	12.4

Table 2.5. Analysis of gain and phase responses of the PGA

The same moderate inversion design criteria for the LNA (section 2.3) was applied to the PGA and Table 2.4 shows the IC and sizing used. Additional boosting of energy efficiency using the 2-bit automatic compensation of SR and phase margin can be illustrated as follows: when large gain is selected, it imposes a larger SR requirement and most probably the user is recording EAP, where the signal amplitude is 100-500 μV with a bandwidth of ~6kHz. In this case, it is more energy-efficient to use a small compensation capacitor (***C_C***) rather than increasing the bias current (see section 2.3.3 for similar analysis). On the contrary, small gain selection would only be used for recording LFP, where the signal amplitude is 1-3 mV (assuming a small electrode < 150μm^2) with a bandwidth of 300Hz. In this case, larger ***C_C*** is needed to guarantee stability while still achieving appropriate SR. Table 2.5 shows the complete set of parameters that change with gain selection and shows that a minimum of 70° is achieved.

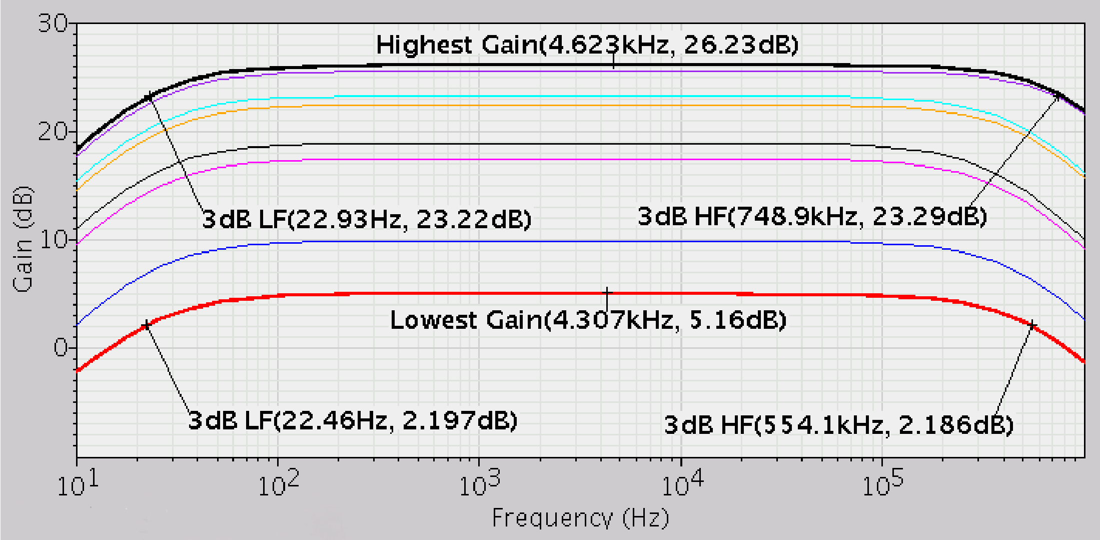


Figure 2.16. Simulation of 3-bit programmable gain amplifier

Figure 2.16 shows the simulated gain response with a total current consumption of 8.66μA and indicates acceptable gains and frequency corners for multiplexing more than 16-channels of neural signals.

2.7 Asynchronous 320kS/s 9-Bit Successive Approximation Analog-to-Digital Converter with Rail-to-Rail Dynamic Comparator and Power-Scalable Sampling Operation

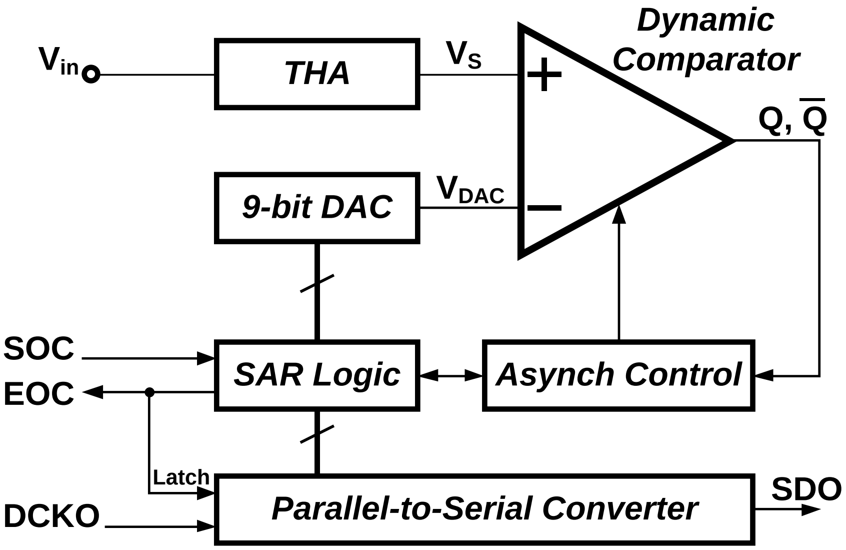


Figure 2.17. 9-bit Asynchronous SAR ADC

As shown in Figure 2.3, there is a flexible cable that connects the 3-D probe platform to the final signal processing and telemetry platform. This cable needs to have a limited number of wires as well as to avoid high-speed signals or clocks to minimize interference. While this would lead to an asynchronous solution requirement, it indicates that local coordination circuits are required due to unavailable synchronization clock. Particularly in SAR ADC the asynchronous approach reduce the clock speed by the number of bits used.

Another important requirement for low-voltage design is the ability to acquire as large analog signal as possible to maximize the input SNR and relax the requirements and power consumption of the sampling circuit; otherwise, charge sharing and clock feed-through may require highly complicated and large-power sampling operation. As more and more supply reduction is done, it is crucial to have the full-scale input range of the ADC equivalent to the rail-to-rail supply voltage.

In this section, a 9-bit successive approximation register ADC (SAR ADC) design is proposed with the following features: 1) fully asynchronous operation (which reduces the clock speed by the number of bits implemented), 2) minimized serial interface for data acquisition, 3) dynamic rail-to-rail comparator, and 4) power-scalable sampling operation that provides an optimized supply voltage for adequate sampling frequency (up to 50 kS/s/channel).

The proposed SAR ADC (Figure 2.17) consists of four blocks: 1) track and hold amplifier (THA), 2) 9-bit capacitive DAC, 3) dynamic rail-to-rail comparator, and 4) digital circuits for SAR logic, asynchronous operation, and parallel-to-serial conversion. The overall operation, design, and function of these blocks are described in the next subsections.

2.7.1 The Overall Operation

The acquisition starts with a start-of-conversion (*SOC*) signal that resets the 10-bit DAC and charges its MSB capacitor ($V_{DAC}=V_{DD}/2$), activates the THA to sample the input ($V_S=V_{in}$), and resets the SAR logic. Once *SOC* goes low, it initiates the dynamic comparator, and activates the next channel in TDM (Figure 2.12). According to the comparison result (V_S vs. V_{DAC}), the SAR logic controls V_{DAC} to track V_S by charging/discharging the capacitors in the DAC. The dynamic comparator is controlled asynchronously to make 9 comparisons, with the results stored in the 9-bit DAC itself, before the SAR logic triggers an end-of-conversion (*EOC*) signal and latches the 10-bit data to a parallel-to-serial converter (PSC). From the time *EOC* is triggered until another *EOC* from the next sample is triggered, the host (processing platform) can clock-out the latch using *DCKO* and capture the 9-bit serial data at *SDO*.

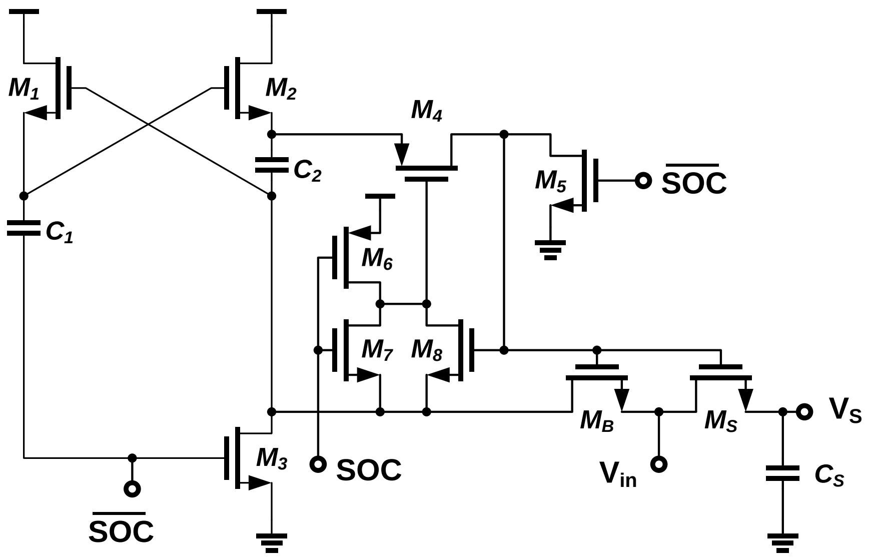


Figure 2.18. Bootstrapped Track and Hold Circuit

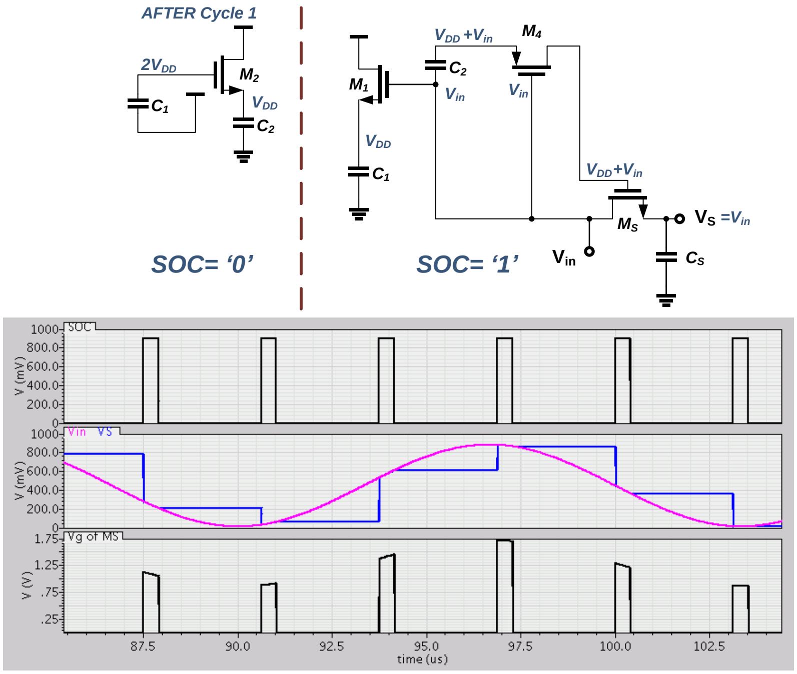


Figure 2.19. Operation of Track and Hold Circuit

2.7.2 Track and Hold Circuit (THA)

The THA should provide a wide swing equivalent to the rail-to-rail voltage, a very low resistance sampling, and a minimized charge sharing and feed-through. This indicates that, especially for low-voltage, a transmission gate solution is no more valid. Therefore, a bootstrapped THA is required with a fixed gate-source voltage such that the charge sharing can be considered signal-independent. Figure 2.18 shows a proposed THA circuit that can perform the aforementioned requirements. This THA is a modified version of [97], where some components has been eliminated at the cost that the first cycle is sacrificed (not an issue). As shown in Figure 2.19, after the first sample when *SOC* is low, the bottom-plate (BP) of $\boldsymbol{C_1}$ is connected to V_{DD} and therefore the gate of M_2 is at twice the supply voltage. This will charge $\boldsymbol{C_2}$ to V_{DD} and its BP is grounded. When *SOC* is high, the BP of $\boldsymbol{C_2}$ is connected to V_{in} which makes the V_{gs} of M_S to always be fixed at V_{DD}. This is significant since it helps make the charge sharing fixed and independent of the input voltage. To further reduce the amplitude of charge sharing and feedthrough and to achieve an equivalent ENOB of more than 9-bit, the sampling MOST (M_S) was made with a width of 3µm. The sampling capacitor ($\boldsymbol{C_S}$) should be chosen such that the kT/C noise is much less than the quantization noise ($\boldsymbol{Q_n}$≈2.5mV); this indicate that $\boldsymbol{C_S}$ of 17aF is appropriate for a 10x ratio. However, such a small value is not realistic in CMOS process. Moreover, the value of $\boldsymbol{C_S}$ should guarantee a reasonable hold of the sampled voltage during the conversion process due to the relatively long sampling time and the kick-back effects from the comparator. A value of ~880fF was chosen to satisfy these requirements and to be a multiple of the unit capacitor used in the DAC as shown next.

2.7.3 9-bit Capacitive DAC with a Compensation Capacitor

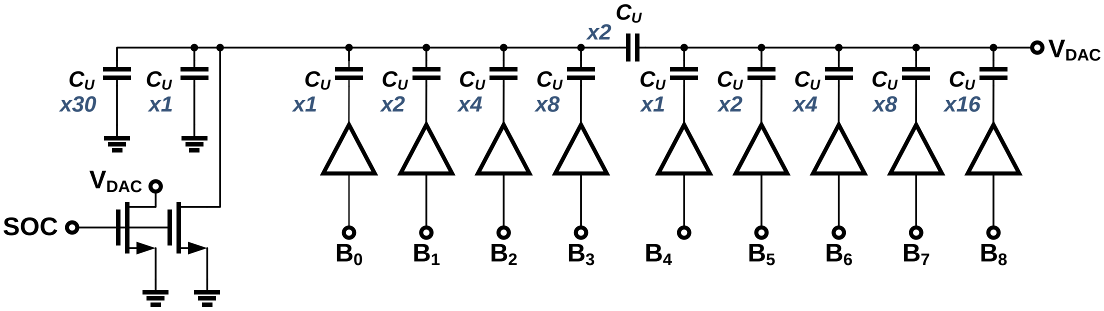


Figure 2.20. A 9-bit DAC with a split capacitor banks connected with a coupling capacitor

As shown in Figure 2.20, the DAC consists of two binary-weighted capacitor banks connected by a coupling capacitor. The drawback of this DAC is that the coupling capacitor is usually a fraction of the unit capacitor ($\boldsymbol{C_U}$); therefore, it will degrade the matching of the DAC. In [98], however, additional capacitor at the least-significant-bits bank was proposed to make the coupling capacitor an integer multiple of $\boldsymbol{C_U}$. The proposed formula is:

$$(2^{N_{MSB}} - 1).y.C_U - ATT.C_U = 2^{N_{MSP}}.C_U \qquad \textbf{Equ. 2.38}$$

Where $\boldsymbol{N_{MSB}}$ is the number of bits in the MSP capacitor bank, $\boldsymbol{y}$ is multiples of $\boldsymbol{C_U}$ used in the coupling capacitor, and $\boldsymbol{ATT}$ is the multiples of $\boldsymbol{C_U}$ that should be added to the LSB capacitor bank. For appropriate split of the overall 9 bits, $\boldsymbol{N_{MSB}}$ is 5, and a $\boldsymbol{y}$ of 2; therefore, it can be seen that 30 $\boldsymbol{C_U}$ should be added to the LSB bank. The choice of $\boldsymbol{C_U}$ was based on the settling/leakage of the final value of the DAC in addition to the overall matching. The worst case capacitor to charge/discharge is the MSB capacitor ($\boldsymbol{C_{MSB}}$=16$\boldsymbol{C_U}$); assuming a total conversion time available of 2.5μs (250ns/bit) and a swing of $\boldsymbol{V_{DD}}$, the time constant should be less than 35ns for less than 0.1% error in the final value. It can be seen, therefore, that this is not a restriction since all the capacitors are in the fF range to reduce the power consumption and the buffer resistance to either $\boldsymbol{V_{DD}}$ or ground is very small. This would narrow the restriction of sizing to the matching requirement, which is characterized in Figure 2.21 (for the 0.25µm CMOS process). Based on worst case simulation (using Matlab) when the ADC makes a transition from 011111111 to 100000000, the unit capacitor should be more than 0.5pF for a monotonic transfer function and a worst case DNL less than |-1|; however, this makes the design impractically large when implemented with MIM capacitors that has a typical value of 1fF/µm^2. A better approach maybe to conduct a Monte Carlo simulation with different unit capacitors using the mismatch data provided by the vendor. The upper left corner of Figure 2.21 suggests that σ in percentage is given by:

$$\sigma(\%) = 2/\sqrt{C_U[fF]} \qquad \textbf{Equ. 2.39}$$

Figure 2.21 shows the simulation results for unit capacitors of 49fF (|DNL| at 3σ is 1.49), 81fF (|DNL| at 3σ is 1.17), and 121fF (|DNL| at 3σ is 0.96). To minimize the area, the unit capacitor size was selected to be 49fF, which results in |DNL| of 0.99 at 2σ (95.45% confidence interval).

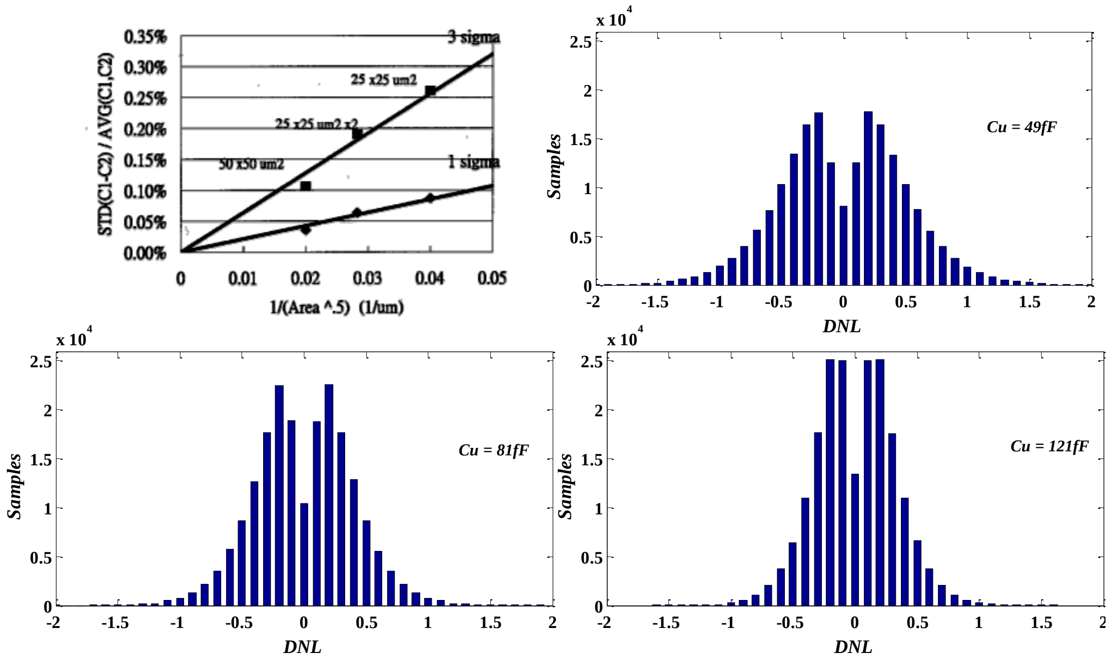


Figure 2.21. The capacitor mismatch characterized by the 0.25µm CMOS process and three Monte Carlo simulations of DNL when the DAC unit capacitor is 49fF, 81fF, and 121fF

It can be seen from Figure 2.20 that the only reference for this ADC is the supply voltage (***V_{DD}***), which makes it very flexible and simple to later scale the power consumption by changing only ***V_{DD}*** if the maximum sampling rate (***f_{Smax}***) is to be changed.

2.7.4 Dynamic Comparator with Rail-to-Rail Input-Common-Mode

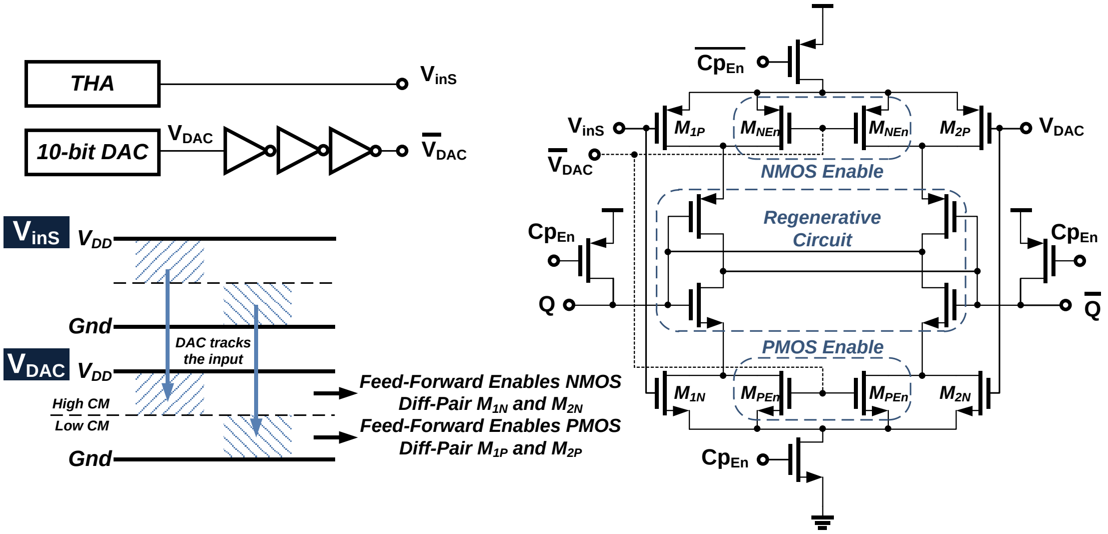


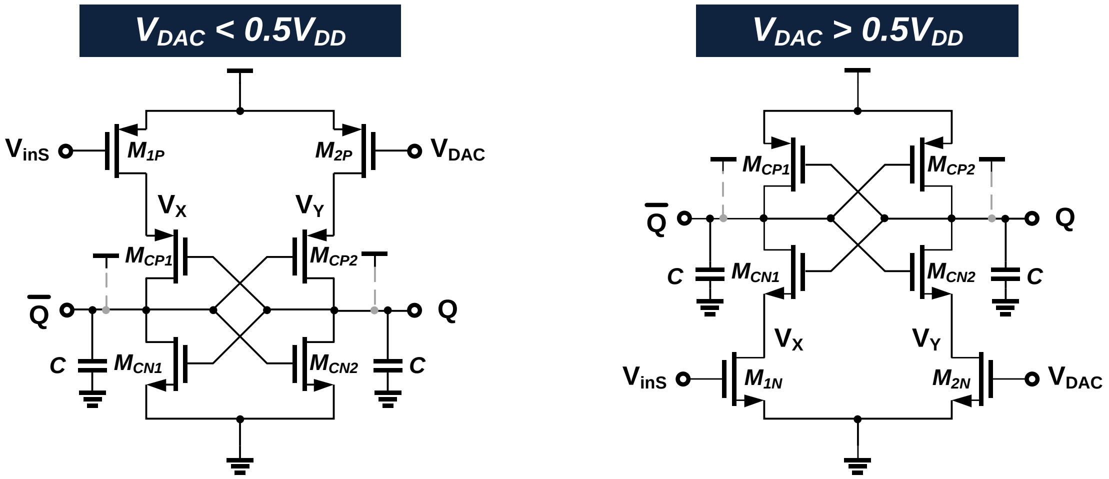
Figure 2.22. A rail-to-rail dynamic comparator using feed-forward for the input-common-mode tracked by the DAC output

Low-voltage ADC operation and dynamic voltage scaling according to the sampling frequency are very challenging to satisfy when it comes to the comparator design. The comparator needs to save energy when voltage is scaled down while the speed requirement is also relaxed due to low sampling frequency. In a conventional static comparator, however, lowering the voltage will cause another restriction since the minimum distinguishable input voltage is reduced as the supply is reduced. This may require the use of relatively high bias current to increase the transconductance of the comparator differential pair, which also makes the whole design energy-inefficient.

The previous arguments suggest that a comparator needs to operate rail-to-rail (to reduce or relax the requirements on the minimum distinguishable input voltage), and needs to eliminate the static current by pursuing dynamic operation. Figure 2.22 shows a novel comparator that achieves both features: dynamic operation, and rail-to-rail input range capability.

The comparator in Figure 2.22 operates as follows: at the beginning of each sampling operation, the asynchronous control logic pre-charges the comparator by making $\boldsymbol{Cp_{En}}$ '0' and thus $\boldsymbol{Q}$ and $\overline{\boldsymbol{Q}}$ become '1'. Once acquisition starts, $\boldsymbol{Cp_{En}}$ is made '1' to enable the comparator and the input differential-pair biases the cross-coupled inverters to switch at opposite directions according to the input voltage. The next subsection describes the details of comparator operation and how it influences the ADC SAR logic.

Figure 2.22 shows that the comparator was designed to have NMOS and PMOS differential pairs, but only one pair is used at a time. The functioning pair is decided using a feed-forward path, which is simply an inverted digital version of the DAC output voltage ($\boldsymbol{V_{DAC}}$). Since the DAC tracks the sampled input voltage ($\boldsymbol{V_{inS}}$) during the SAR operation, its voltage can be used to know is the $\boldsymbol{V_{inS}}$ is above or below the mid-input-range (half $\boldsymbol{V_{DD}}$). At the very first cycle of the ADC operation, the NMOS pair is enabled. If $\boldsymbol{V_{inS}}$ is higher than half the supply, the feed-forward value ($\overline{V_{DAC}}$) remains '0' which keeps the PMOS pair ineffective and the NMOS pair effective. Otherwise, if $\boldsymbol{V_{inS}}$ lower than half the supply the DAC voltage becomes lower than $\boldsymbol{V_{DD}}$/2 and $\overline{V_{DAC}}$ becomes '1'; this will make M_{PEn} MOSTs short circuit and M_{NEn} MOSTs open circuit; thus the NMOS pair becomes ineffective and the PMOS pair effective. Figure 2.23 shows the effective functional comparator in both cases at the onset of regenerative phase of operation.


$V_{DAC} < 0.5V_{DD}$

$V_{DAC} > 0.5V_{DD}$

Feed-Forward Enables PMOS Diff-Pair M_{1P} and M_{2P}

Feed-Forward Enables NMOS Diff-Pair M_{1N} and M_{2N}

Figure 2.23. Effective functional comparator at the onset of regenerative operation phase under low common-mode (V_{DAC} < $0.5V_{DD}$) and high common-mode (V_{DAC} > $0.5V_{DD}$)

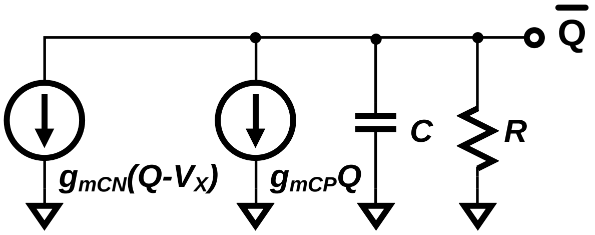


Figure 2.24. Comparator small signal approximation for the NMOS operating mode

To ensure the comparator is reliable, characterization of the gain and metastability using small-signal approximation is required. Figure 2.24 shows part of the small signal analysis of the NMOS operating mode. The following equations show the impact of the regenerative circuit on the overall performance:

$$g_{mCN}(Q - V_X) + g_{mCP}Q + \frac{\bar{Q}}{R} + C\frac{d\bar{Q}}{dt} = 0 \qquad \textbf{Equ. 2.40}$$

For the small-signal analysis, $\overline{Q}$ can be substituted by (-*Q*), and V_X can be substituted by:

$$V_X = -g_{mN}R_S V_{inS} \qquad \textbf{Equ. 2.41}$$

where R_S is the equivalent output resistance looking from $\mathrm{V_X}$ towards the circuit. This can approximate the expression to:

$$\frac{1}{Q}\frac{dQ}{dt} = \frac{g_{mCN} + g_{mCP}}{C} + \frac{g_{mCN} + g_{mN}R_S}{C}\frac{V_{inS}}{Q} - \frac{1}{RC} = 0 \qquad \textbf{Equ. 2.42}$$

To simplify further, the second term of the right-hand side can be ignored and thus the integration of Equ. 2.42 give:

$$Q = K.exp\left(\frac{g_{mCN} + g_{mCP}}{C} - \frac{1}{RC}\right)t\,, K = \Delta V_{in} \qquad \textbf{Equ. 2.43}$$

To be more accurate, the right-hand side represents the output swing ΔV_{out} =(Q-$\overline{Q}$) and therefore the transfer function of the comparator can be written as:

$$A_{COMP} = \frac{\Delta V_{out}}{\Delta V_{in}} = exp\left(\frac{1}{\tau_r} - \frac{1}{\tau_s}\right)t, \tau_r = \frac{C}{g_{mCN}+g_{mCP}}, \tau_s = RC \qquad \textbf{Equ. 2.44}$$

where τ_r and τ_s are the time constants for the comparator regeneration and settling, respectively. This regenerative part of the above equation represents the gain, which can also be used to estimate the probability of error or metastability.

It should be guaranteed that ΔV_{out} can be more or at least equal to the output logic range (V_L) which is similar to V_{DD}. The meatstability or error probability is equivalent to how probable the comparator will need to resolve a certain minimum input ($\Delta V_{in.min}$) in the context of the overall input range ($V_{in,range}$). With a rail-to-rail input range this is equivalent also to V_{DD}. Therefore, the error probability can be written as:

$$P.E = \frac{2 \cdot \Delta V_{in,min}}{V_{in,range}} = \frac{2V_L}{V_{in,range} \cdot A_{COMP}} = \frac{2}{A_{COMP}} \approx exp\left(-\frac{t}{\tau_r}\right) \qquad \textbf{Equ. 2.45}$$

The previous equation shows the significance of the regeneration time constant in minimizing the probability of error or metastability, which is minimized by maximizing g_{mCN}, g_{mCP} and minimizing the load capacitance C. The choice of these parameters to determine the gain if highly complicated, especially for a dynamic comparator, without the use of simulation tools. Therefore, a method described by Stacy Ho[5] of Analog Devices, Inc., was adopt for determining τ_r based on simulation and this value is used for calculating the gain and ensure reliable operation.

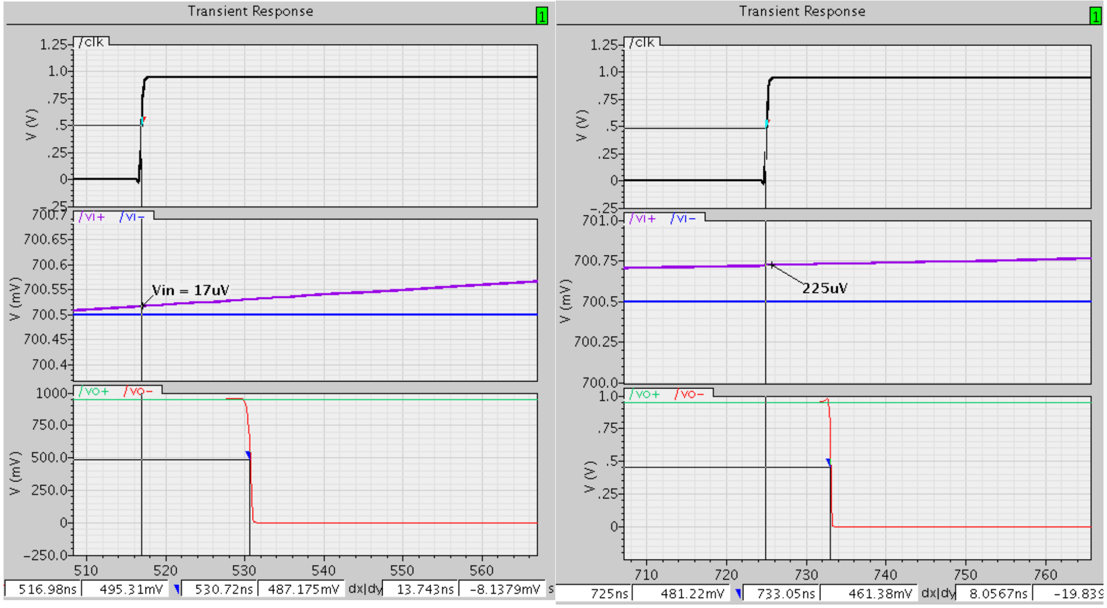


Figure 2.25. Simulation results of the comparator transient response for two different inputs V_{in1} and V_{in2} varying by an order of magnitude

This method can be described with the aid of Figure 2.25 where two different inputs V_{in1} and V_{in2} are separated by nearly an order of magnitude (V_{in1}=10V_{in2}) and applied to the dynamic comparator. The point in time-axis where both inputs reached a fixed output voltage (e.g. 0.5V) is recorded (t_1 and t_2) and the following equation can be used to determine τ_r:

[5] This method was taught to the author by Prof. M. Flynn in EECS 511, Winter 2009

$$\tau_r = \frac{t_2 - t_1}{ln(10)} \qquad \textbf{Equ. 2.46}$$

Based on this equation and Figure 2.25 gain, and metastability are listed in Table 2.6 for three different common-mode input ranges.

CM Voltage	***V_{in1} [μV]***	***V_{in2} [μV]***	***t_1 [ns]***	***t_2 [ns]***	***τ_r [ns]***	***T/2 [ns]***	***Gain [x10^{18}]***	***Metastability [x10^{-18}]***
High	***225***	***17***	8	13.7	2.47	104	3.86	0.26
Med	***225***	***18.8***	8.1	13.7	2.43	104	3.86	0.26
Low	***225***	***17***	18.55	19.5	0.43	104	>>	<<

Table 2.6. Comparator gain and metastability probability for three different common-mode ranges

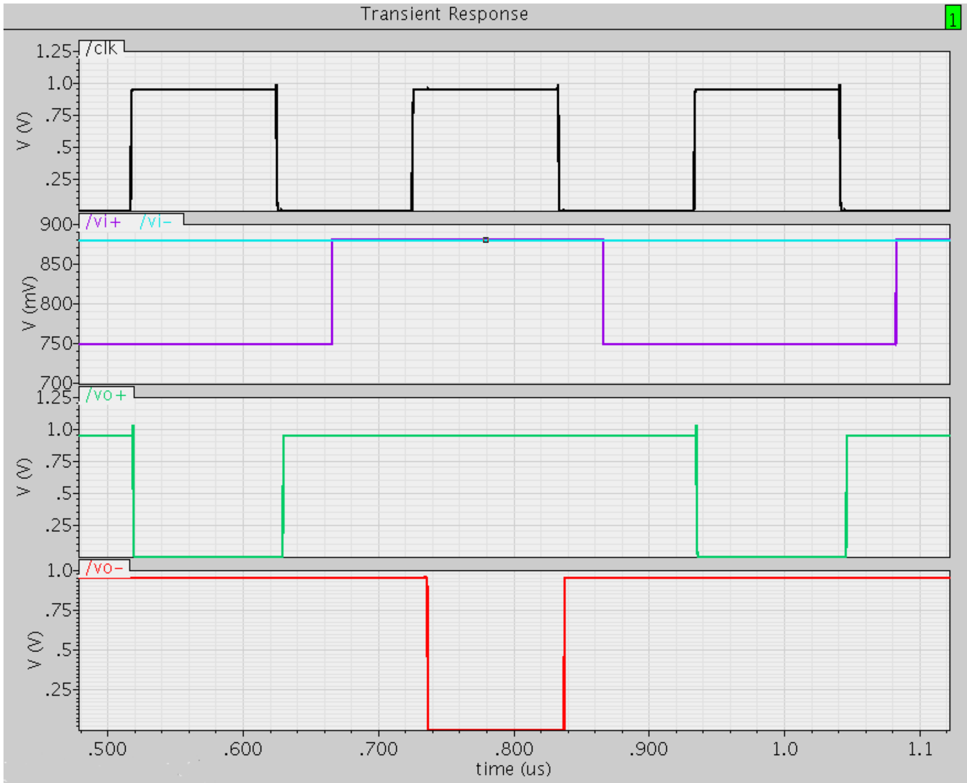


Figure 2.26. Simulation results of the over drive test

Figure 2.26 shows the simulation results of the comparator overdrive test to ensure there is no hysteresis. This one stage design provides the following advantages:

1) The ability to select the appropriate differential-pair based on the input common-mode improves the speed-power efficiency compared with a single differential-pair comparator that has to be optimized for the worst case input

common-mode. This is because the transconductance (and thus the speed) of the NMOS ($\boldsymbol{g_{mNMOS}}$) improves when the gate-source voltage is high and the opposite for the PMOS.

2) The dynamic operation consumes power at a very narrow window in the whole ADC operation. This saves power compared with the static approach.
3) One stage comparator saves also power compared to two-stage comparators that are used to improve the performance (speed).
4) The rail-to-rail operation relaxes the requirements on the THA circuit as well as on the comparator itself.

The next subsection describe further about the overall collaboration of different ADC blocks coordinated by the asynchronous and SAR logic.

2.7.5 Asynchronous Logic and Interface

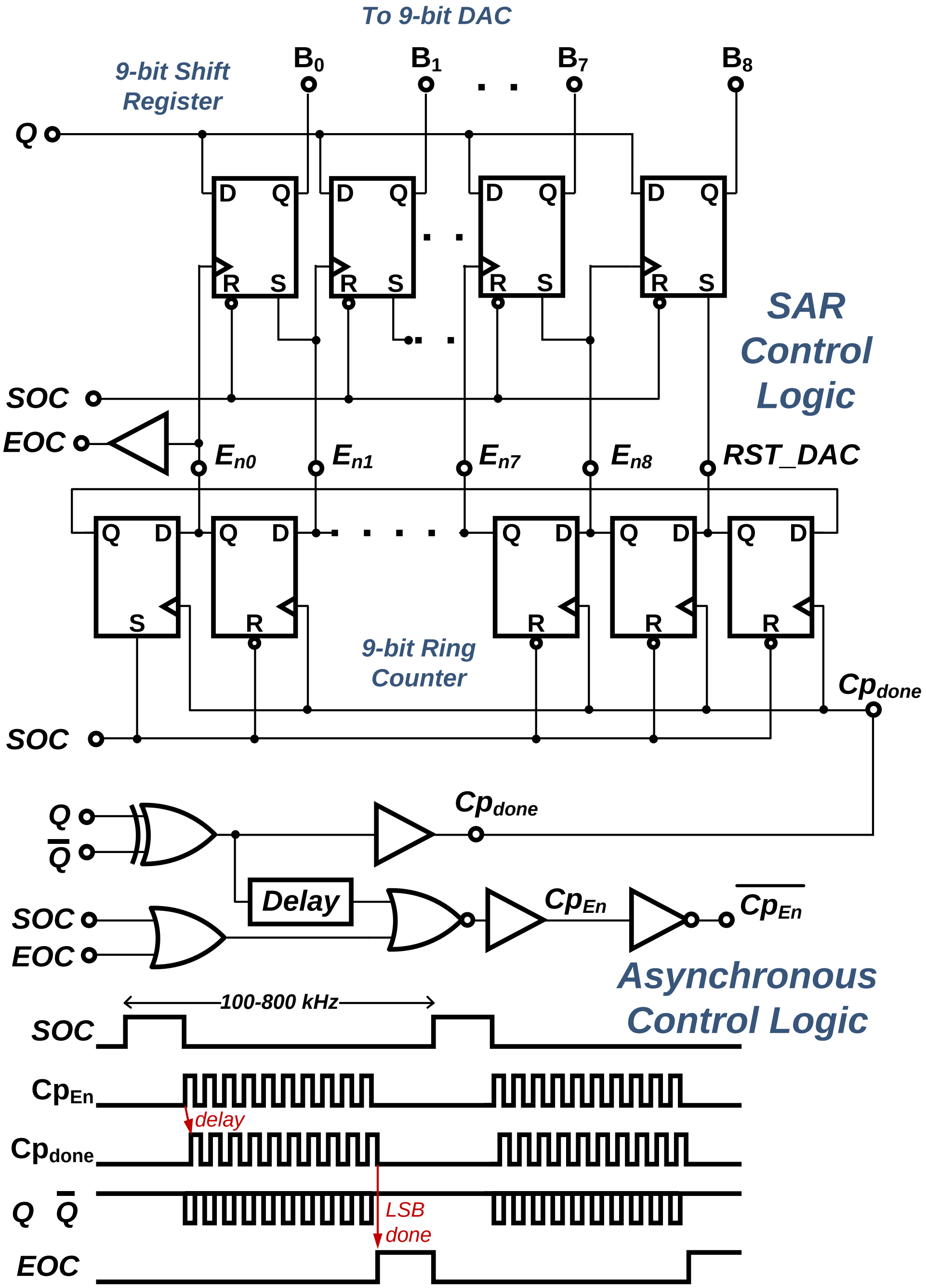


Figure 2.27. Asynchronous and successive approximation control logic

Figure 2.27 shows the logic circuits responsible for SAR functionality and for maintaining asynchronous operation. At sampling time (SOC is ‘1’) all digital signals are reset, DAC has zero charge, and comparator is pre-charged. Once acquisition starts (SOC is ‘0’), the comparator is enabled and when $\boldsymbol{Q}$ and $\overline{\boldsymbol{Q}}$ are evaluated a done signal

(Cp_{done}) is generated by the asynchronous logic control (ALC). The rising edge of this signal is input to a 9-bit ring counter that sets the MSB bit of the DAC (V_{DAC}=V_{DD}/2). This first cycle is not part of the ADC operation but just sets the DAC initial value to MSB. After some delay (~50ns) comparator is disabled by the ALC, output is pre-charged again and the Cp_{done} signal goes low. After another 50ns, the ALC re-enables the comparator and so on. It can be seen from Figure 2.27 that the ALC circuit uses both outputs of the comparator to generate enable, disable, and done signals that are delayed according to the delay time implemented using a series of large-length inverters. The 10-bit ring counter enables a shift register that has its corresponding bits set at the beginning of a comparison cycle and the value is kept or reset according to the value of *Q* at this cycle. This will keep changing the DAC voltage to track the sampled input voltage until LSB is evaluated in the 11th cycle. At this point *EOC* flag is raised to indicate end-of-conversion. *EOC* is used to latch the 9-bit result into a parallel-to-serial converter as shown later. This asynchronous operation tolerates more process variations and avoids the interference potential of any high-speed clock signal to the rest of the signals in the flexible cable.

2.7.6 Parallel-to-Serial Converter and Data Acquisition

As shown in Figure 2.17, the end-of-conversion (*EOC*) signal latches the output 10-bit data to a parallel-to-serial converter (PSC), which consists of a 9-bit shift register and a ring counter to select a particular bit to show at the serial output using a clock signal (*DCKO*) provided by the host system or processor. The data is latched until another *EOC* from the next sample is triggered; therefore, the host (processing platform) can perform an OR operation of *EOC* received from different AFE modules, then provide one *DCKO* to all modules at once and capture 9-bit serial data at *SDO* Of each module.

Figure 2.28 shows the result of full system simulation using different input signals (1kHz and 5kHz) at each LNA. The received output at the receiver is shown (after DAC conversion) with demultiplexed signals corresponding to channels 6 and 14. These two signals are checked to provide an effective number of bits (ENOB) of about 10 bits.

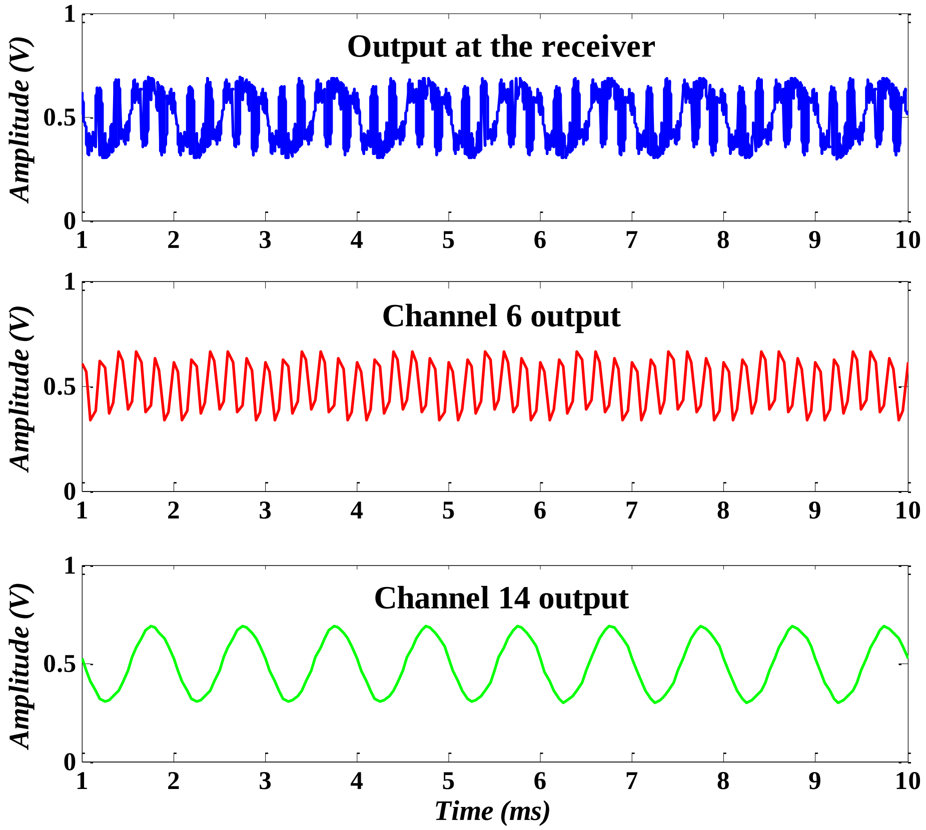


Figure 2.28. Data output after acquisition

2.7.7 Power Scalable Sampling Operation

The proposed rail-to-rail ADC operation is enabled by the comparator design; however, using only the digital supply voltage (DV_{DD}) as the reference voltage enables a flexible scaling of power (through DV_{DD}) that is set to allow a specific sampling frequency. Consequently, the design saves power and improves energy efficiency that is now dependent on the maximum sampling frequency (F_S). The objective is also to achieve more than 40kHz per channel, which is required to improve the accuracy of spike sorting algorithms [99].

In the next section, measurement of different sampling frequency supported by the chip is shown in addition to the measurement of all other circuit blocks.

2.8 Measurement Results and Performance Comparison for AHI-16

2.8.1 Overall System Measurement

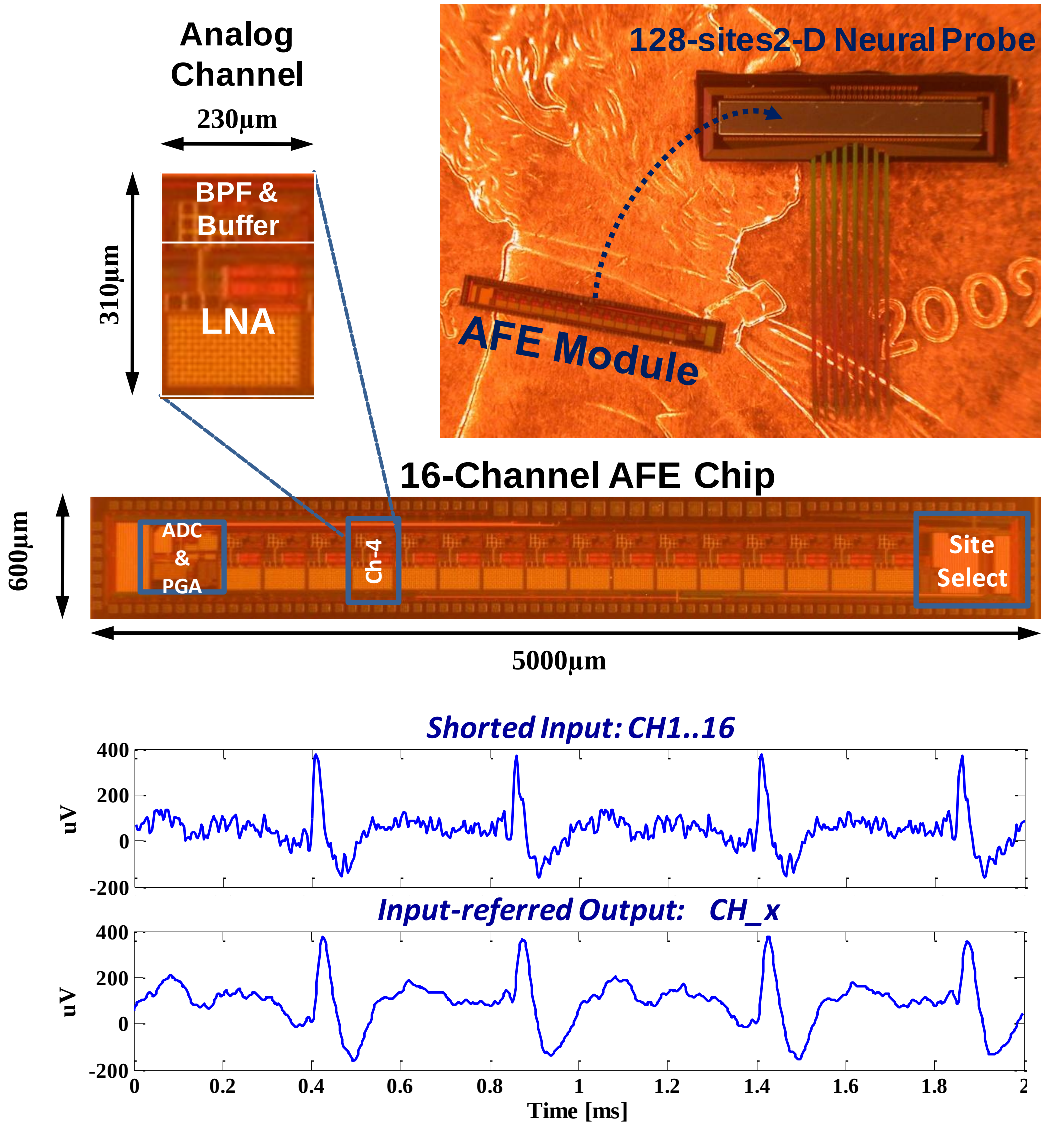


Figure 2.29. Die photo of the fabricated chip and a 128-sites Michigan probe array. It also shows benchtop testing results using synthesized neural signals applied to the shorted inputs (middle), and the recovered input-referred AFE output after demultiplexing and DAC operation using an off-chip FPGA.

The chip was fabricated using 0.25µm CMOS process and operated nominally at 0.9V. Figure 2.29 shows the both the chip microphotograph and the fabricated probe. The AFE chip has 128 pads (40x40µm^2, 70µm pitch) corresponding to the recording sites on the probe. Another 9 pads (72x72µm^2, 105µm pitch) are used for power and data. Among the 9 pads, only five are necessary for device operation.

Figure 2.29 shows also the test-bench result using synthesized neural signals applied at a shorted 16-channel input, and the input-referred output signal after demultiplexing and conversion at any arbitrary channel. In the test setup (Figure 2.30), FPGA was used to emulate the host (processor) and a LabVIEW interface program was developed to provide programming commands and receive and demultiplex the output data.

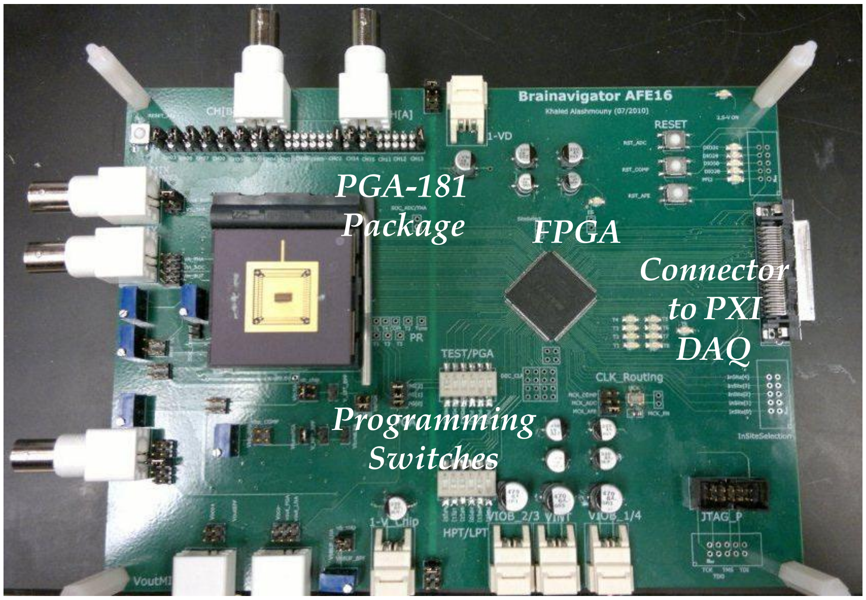


Figure 2.30. Measurement Setup

The AFE architecture significantly reduces the number of leads transmitting signals from/to the probe to/from the host. It can be easily scaled to stacking multiple modules by sharing all the programming signals such that only two unique signals are required to be allocated to each AFE: EOC and SDATA$_{OUT}$. The AFE module can be bonded to the 2-D probe by either wire-bonding or flip-chip bonding. The next subsection show the measured performance of each individual block followed by performance comparison with state-of-the-art systems.

2.8.2 Low-Noise Amplifier

As mentioned previously, the amplitude of the neural spike is one of the main features extracted to distinguish single neurons from each other as part of the spike sorting algorithm. Therefore, it is important to ensure the reliability of frequency response of LNA (Figure 2.31), especially that moderate inversion is used for low-voltage design which may cause mismatching problems. From two different wafers, 30 die were measured; Figure 2.32 shows the measured gain response with a total current consumption of 1.1 µA and indicates acceptable gain and frequency corners variations. The measured noise response, shown in Figure 2.33, indicates a total input-referred noise of 6.76 µV$_{rms}$ integrated at 100 kHz bandwidth, and it contributes lower noise to the overall system than the typical electrode noise (typically more than 10 µV$_{rms}$)

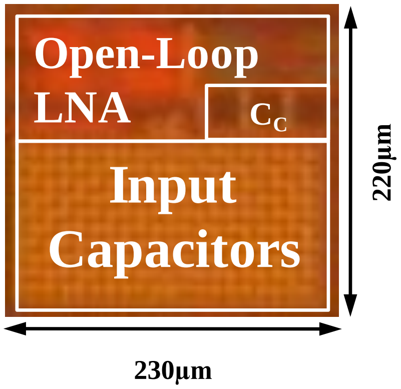


Figure 2.31. The die photograph of the fabricated LNA chip

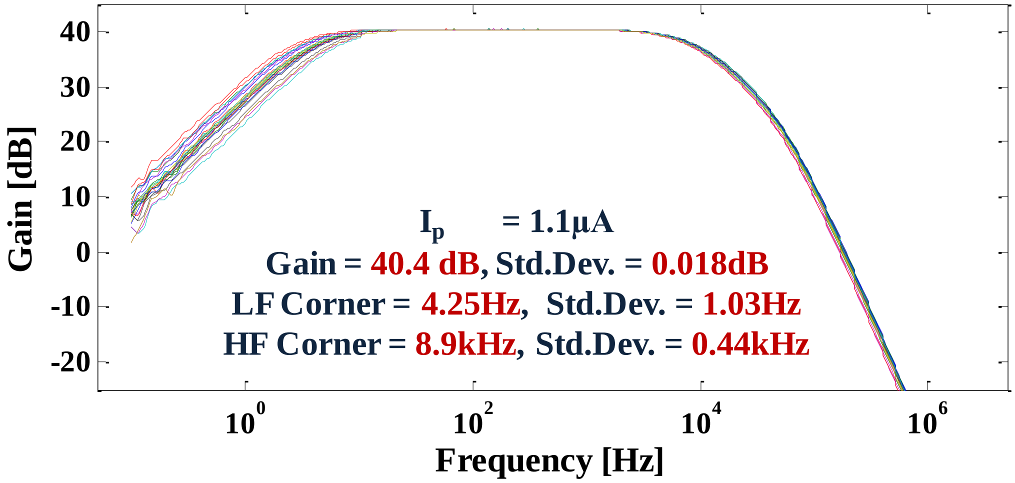


Figure 2.32. Measured LNA gain response from 30 chips/2wafers at 990nW power consumption

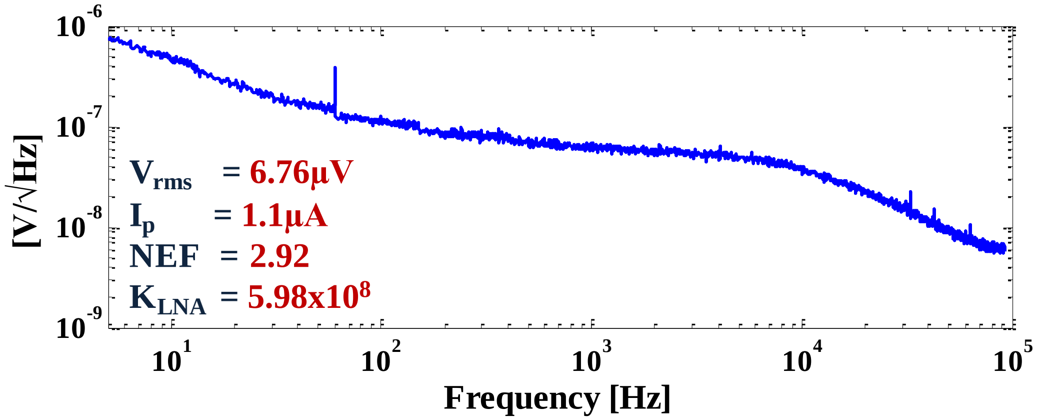


Figure 2.33. Measured LNA input-referred noise at 990nW power consumption

Same measurement was done at about 70% increased current consumption (1.9µA), which also gave acceptable gain response (Figure 2.34) and lower noise (Figure 2.35, 5.83µV_{rms} integrated 100kHz). It is interesting to note that compared to the former current consumption, this later measurement provided improved and even lowest reported NEF value; however, the K_{LNA} value is not so much improved.

The measured common-mode rejection ratio (CMRR) and power-supply rejection ratio (PSRR) are shown in Figure 2.36, and Figure 2.37, respectively.

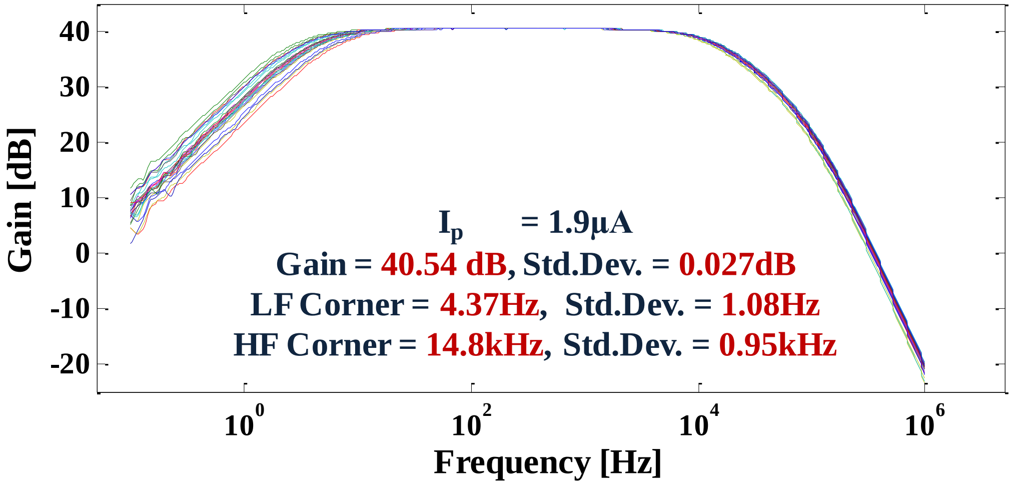


Figure 2.34. Measured LNA gain response from 30 chips/2wafers at 1700nW power consumption

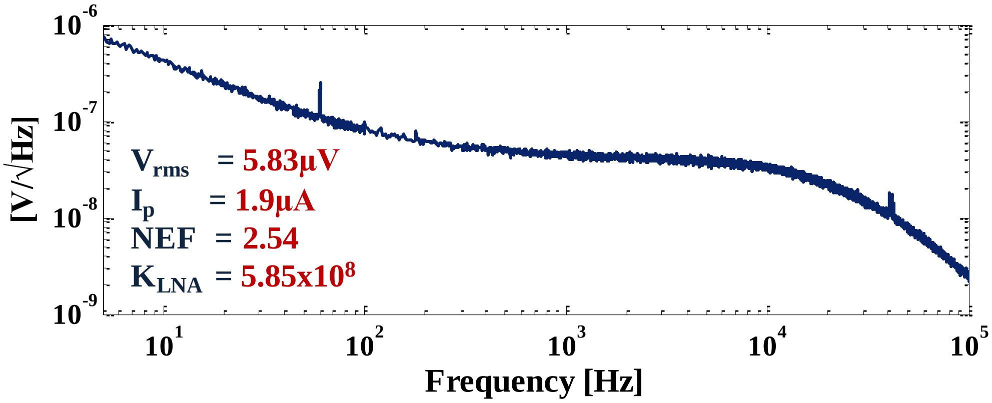


Figure 2.35. Measured LNA noise response at 1700nW power consumption

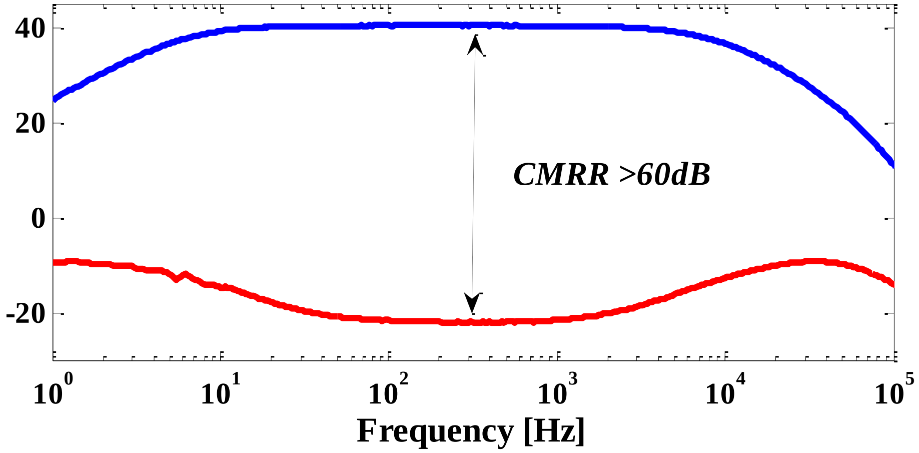


Figure 2.36. Measured LNA common-mode rejection ratio

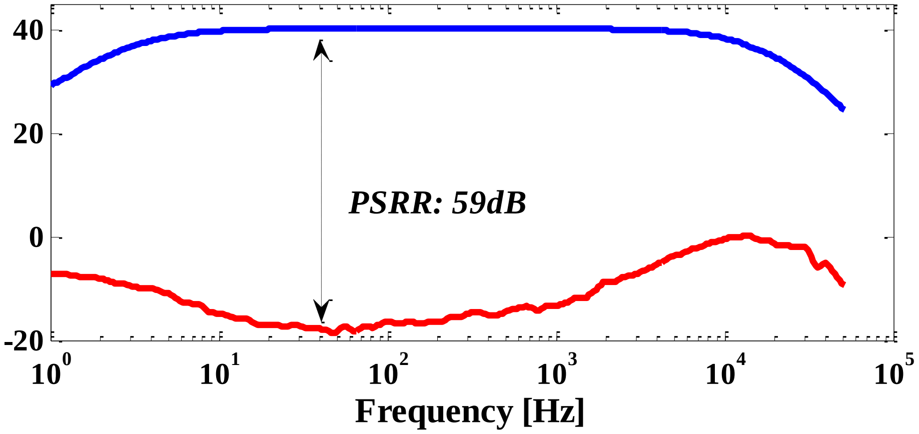


Figure 2.37. Measured LNA power-supply rejection ratio

The performance summary (Table 2.7) and comparison with recently-reported LNAs shows that the implemented LNA has the lowest NEF and about an order of magnitude improved noise efficiency based on K_{LNA}.

Author	[7]	[9]	[10]	[11]	[12]	[13]	**This work**	
Technology [μm]	1.5	1.5	0.18	1.5	0.5	0.35	**0.25**	
Area [mm^2]	0.16	0.107	0.05	0.2	0.16	0.02	**0.05**	
Current [μA]	16	38.33	4.67	8	2.7	1.4	**1.9**	**1.1**
Voltage [V]	5	3	1.8	3.4	2.8	3	**0.9**	
Power [μW]	80	115	8.4	27.2	7.56	4.2	**1.7**	**0.99**
Bandwidth [kHz]	7.2	9.1	9.1	10	5.3	5	**14.8**	**8.9**
Noise,in [μVrms]	2.2	7.8	5.6	3.6	3.06	7	**5.83**	**6.76**
Gain [dB]	39.5	39.3	49.5	39.3, 45.6	40.85	34	**40.5**	**40.4**
CMRR [dB]	>83	-	52.7	-	66	-	**>60**	
PSRR [dB]	>85	-	52	-	75	-	**>53**	
DR [dB]	69	47.1	45	65	58	48	**47.7**	**46**
NEF	4	19.5	4.88	4.9	2.66	4.5	**2.54**	**2.92**
K_{LNA} (x10^8)	390	650	50	100	59	42	**5.85**	**5.98**

Table 2.7. LNA performance summary and comparison with state-of-the-art LNAs

2.8.3 Band-Pass Filter

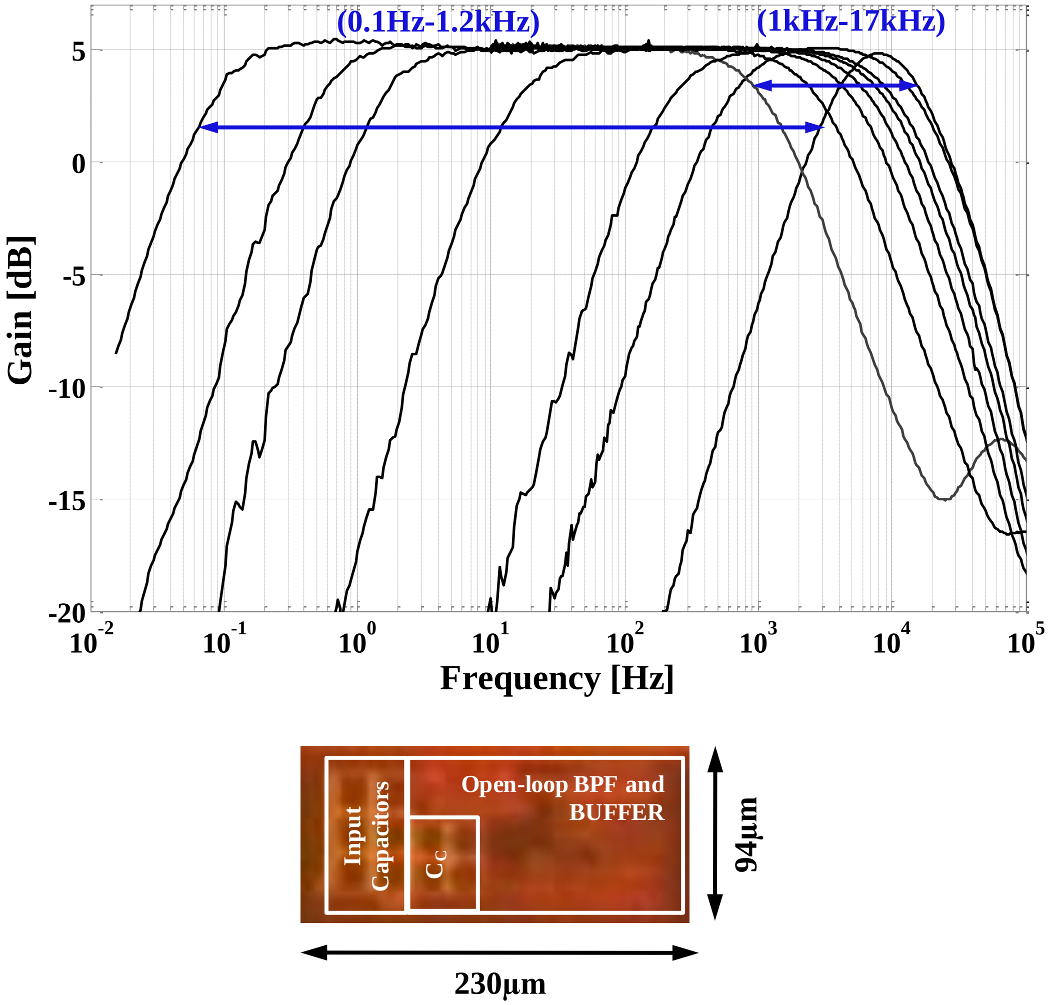


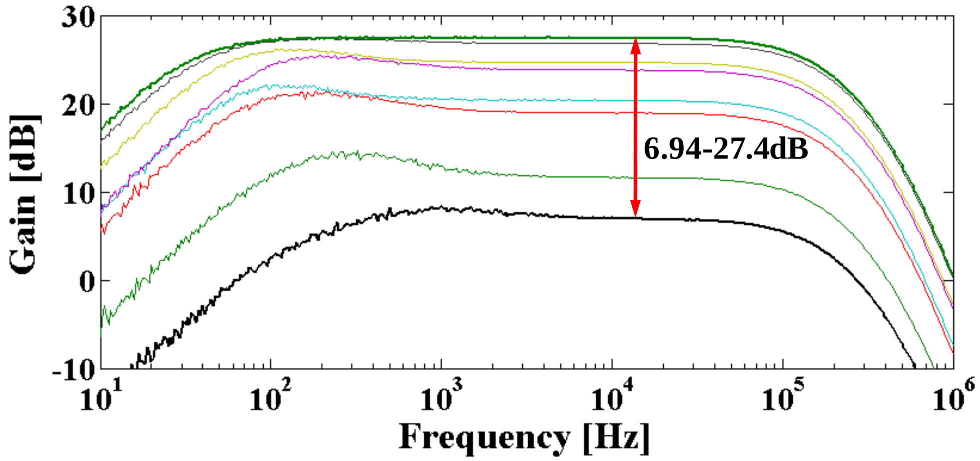
Figure 2.38. Measured gain response of the BPF under different tunings of frequency corners, and die photograph of the fabricated chip

Figure 2.38 shows the BPF chip photograph, and the measurement results of gain response under different tuning conditions; confirming the wide-range tunable bandwidth from 1 to 17 kHz for the high-frequency corner and 0.1 to 1000Hz for the low-frequency corner.

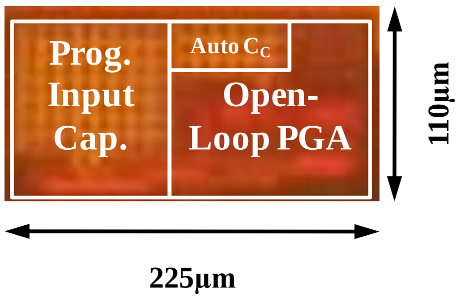
2.8.4 Programmable Gain Amplifier

Figure 2.39 shows the measured gain response, which shows a good match with the simulations and analysis results to guarantee a bandwidth higher than 160 kHz. The chip microphotograph and gain values are also shown in the same figure. The transient response has been tested using a 100kHz sinusoidal wave and the result confirms the gain programming and the dc offset cancellation of input signals (after subtracting 0.5V_{DD}), which is critical for multiplexing channels.

The power consumption of the fabricated PGA was meas9ured as 8.3 μW, which is the lowest among the PGAs reported with similar performance to the best of author's knowledge.



Code	Gain [dB]	Input p-p voltage (THD 1%) [mV]
000	6.94	40
001	11.63	39.6
010	19	39.2
011	20.35	36.9
100	23.82	24.4
101	24.65	22.7
110	26.8	17.9
111	27.4	16.8



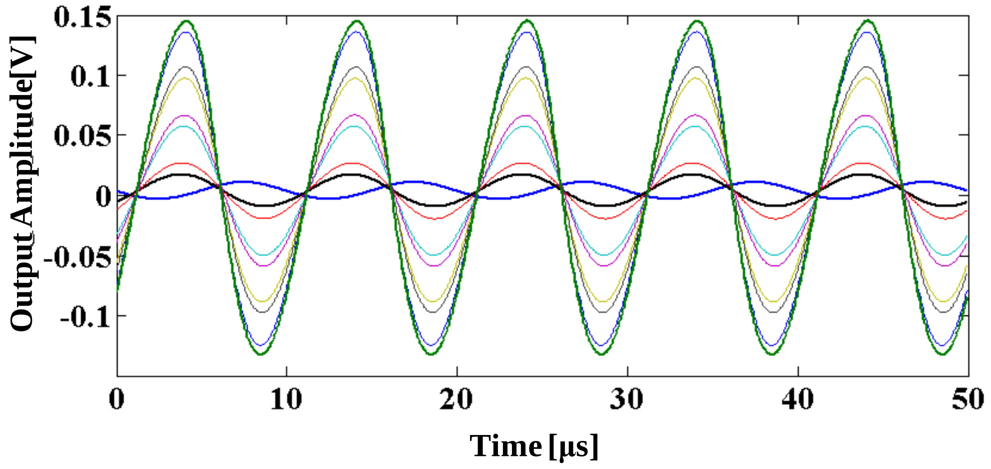


Figure 2.39. PGA measured gain and transient response, and chip photograph

2.8.5 Analog-to-Digital Converter

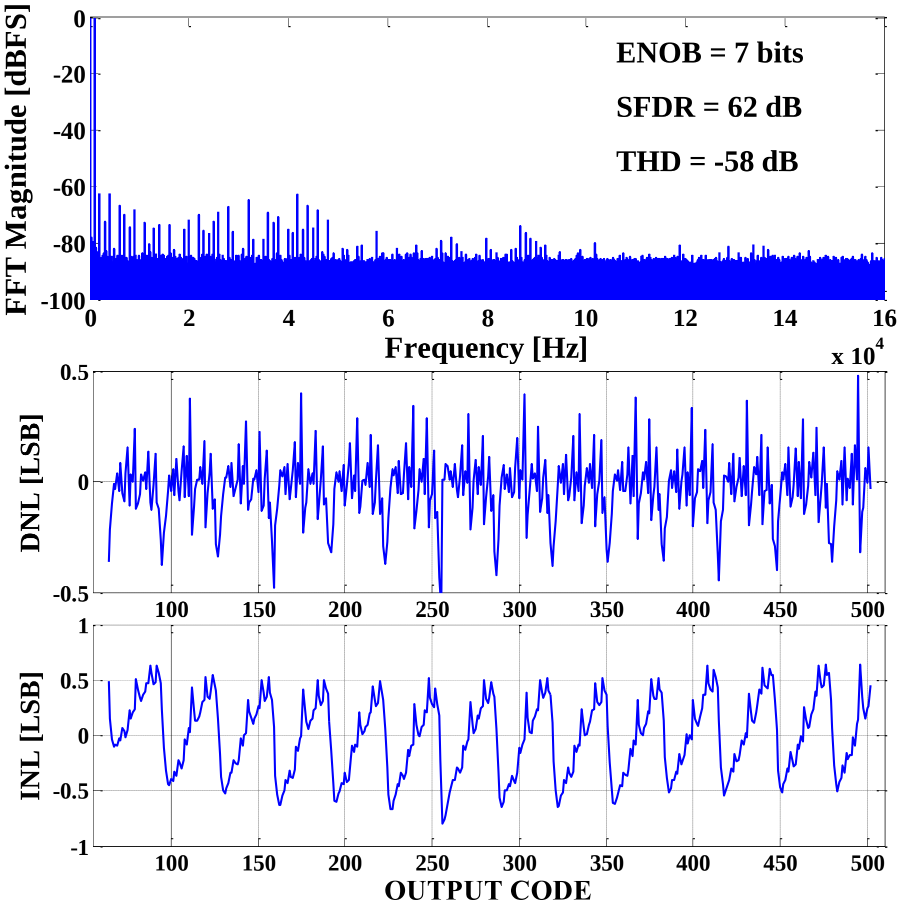


Figure 2.40. ADC measured frequency response, differential, and integrated linearity

Figure 2.40 shows the spectrum and non-linearity measurements of the fabricated SAR ADC. The measured ENOB at maximum input frequency (at sampling frequency of 320 kHz) is 7 bits with 62 dB SFDR and -58 dB THD. With an ENOB of 7 bits, still the quantization noise (input-referred) is smaller than the LNA noise. Under 0.9V operation quantization noise ranges from 2 to 4.8µV$_{rms}$ depending on the gain of PGA. At supply voltage lower than 0.9V, the quantization noise becomes smaller since the LSB scales accordingly. A maximum differential non-linearity (DNL) of (0.4-0.6), and an integrated non-linearity (INL) of (0.6-0.8) at codes from 60 to 510 were observed. The range was taken to avoid the degradation of ENOB due to the distortion caused by both dynamic comparator and THA unit near the ground level. The user can vary the digital supply voltage (DV$_{DD}$) from 0.66V to 1.32V to accommodate sampling frequencies from

100kS/s to 800kS/s, with the minimum FOM of 123fJ/CS. At nominal sampling of 320kS/s, the total power consumption is 10.56µW (660nW per channel). The graphs in Figure 2.41 include the SNDR versus stimulus amplitude at nominal sampling frequency (320kHz), as well as the SNDR at maximum input frequency (Fin = 0.5xFs) when sweeping the supply voltage.

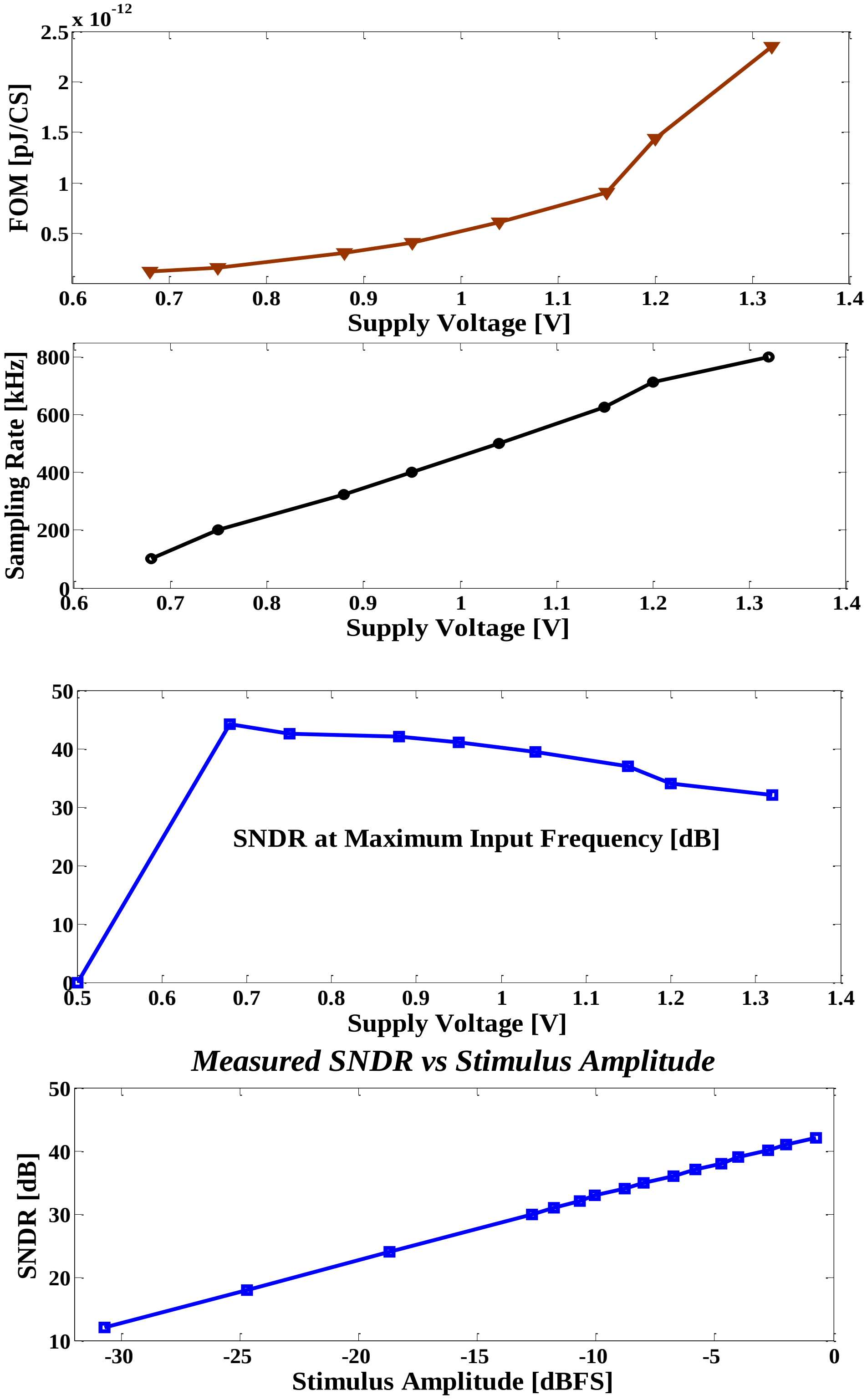


Figure 2.41. ADC measured power scalable sampling and figure of merit. It shows also the estimated FOM along with sampling frequency, and SNDR at different supply voltage.

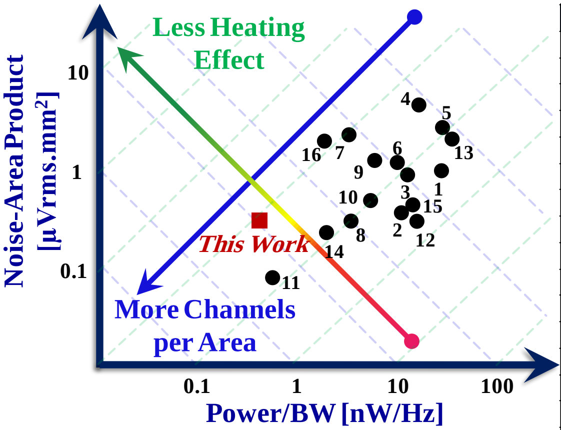
2.8.6 Performance Comparison

Reference	[66]	[82]	[63]	[71]	[70]	[72]	**This Work**
Technology	0.5μm	0.5μm	0.35μm	0.35μm	0.18μm	65nm	0.25μm
LNA							
Area [mm^2]	~0.2		0.02	0.31	0.03	0.013	**0.05**
Voltage [V]	3	3.3	3	1.5	1.8	0.5	**0.9**
Power [μW]	73.5	26.4	4.2	26.9 (Channel)	6.47	5.04 (channel)	**0.99**
DR[dB] at THD<1%	63.96	70	46.93	-	-	-	**47.6**
Noise$_{in\_ref}$ [μV$_{rms}$]	3.9 (10kHz)	1.94 (50kHz)	7 (5kHz)	3.12 (50kHz)	5.4-11.1 (65kHz)	4.9 (10kHz)	**4.8 (10kHz)**
NEF	7.5	2.9	4.6	2.68	4.5	5.99	**2.9**
NEF^2·(V$_{DD}$)	169	27.75	61	10.77	36.45	17.96	**7.56**
K (x10^8)	273	117.7	42	-	-	-	**5.98**
CMRR [dB]	139	>76	-	>56	62	75	**>62**
PSRR [dB]	65	>70	-	>65	72	64	**>59**
BPF and PGA							
Power [μW]	0.2	-	10.8	-	3.3	-	**2.31**
LF Cutoff [Hz]	0.1-1k	0.2-94	0.01-70	1.1-525	0.126-350 (1-bit)	FPGA	**<0.1-1000 (3-bit)**
HF Cutoff [kHz]	0.7-10	0.14-8.2	0.5-5	5.1-12	0.3-12 (1-bit)	FPGA	**1-17 (3-bit)**
Gain [dB]	68-77	39.6	48-68	51.9-65.6 (2-bit)	49-66 (3-bit)	32	**52.4 to 79.8 (3-bit)**
ADC							
Power/Ch [μW]	28	75.9	1.9	5.9	0.483	0.24	**0.66 at 20.16kHz**
Sampling/Ch [kHz]	1.8-21.25	16	10	35.7	31.25	20	**6.25-50**
ENOB	8.8	7*	8*	9.2	7.65	7.16	**7**
Overall AFE Evaluation for Massive Parallel Neural Recording							
Power/Ch [μW]	102	102.3	16.9	26.9	10.1	5.04	**3.96**
Area/Ch [mm^2]	0.27	~0.25	0.04	0.625	0.041	0.013	**0.07**
EAE (x10^{-15})	0.1	0.165	1.06	0.23	5.37	31.15	**7.51**
EE (x10^{-9}) (without Area consideration)	0.027	0.041	0.042	0.14	0.22	0.4	**0.526**
Specific Wires/Ch	-	-	1:1	1:4	8:32	1:1	**2:16**

* ideal number of bits

Table 2.8. Measured performance and comparison with state-of-the-art systems

The performance comparison is summarized in Table 2.8. The proposed AFE module in this work achieved an EAE (enery-area-efficiency) of 7.51x10^{15}. This is the highest efficiency among all the reported designs that used capacitive coupling to reject the electrode DC offset. It provides the wide range of operation through programmability, small area per channel, small lead counts per channel for data transfer, and the lowest power consumption (less than 4µW per channel) as compared with the previous work. Figure 2.42 shows a plot for both noise and area efficiency, which confirms that the proposed design has a potential for massive parallel recording while being safe in terms of tissue heating.



*This graph compares only the Analog front-end

Author/Year	Publication	Ref
Bai and Wise '01	TBME	1
Harrison and Charles '03	JSSC	2
Mohseni and Najafi '04	TBME	3
Olsson and Wise '05	JSSC	4
Harrison et. al '07	JSSC	5
Yin and Ghovanloo '09	ISSCC	6
Chae et. al '09	TNSRE	7
Aziz et. al '09	JSSC	8
Song et. al '09	TNSRE	9
Perlin and Wise '10	JMEMS	10
Muller and Rabaey '11	ISSCC	11
Shulyzki et. al '11	CICC	12
Lopez et. al '11	ISCAS	13
Wattanapanitch and Sarpeshkar '11	TBioCAS	14
Mollazadeh et. Al '09	TBME	15
Azin and Mohseni '11	JSSC	16

Figure 2.42. Performance chart of the implemented AFE.

2.9 Conclusions

A low-power low-noise analog front-end module has been proposed for integration into 3-D neural recording microsystems. The prototype device has investigated to overcome some critical circuit challenges in power, area, interface, and modularity. The implemented front-end module has achieved low power consumption of 4 μW/channel in a 3 mm^2 area using a 0.25μm CMOS processes. Energy efficiency has been optimized for low-voltage mixed-mode circuits operating at 0.9V in moderate inversion. An asynchronous 320kS/s 9-bit ADC was implemented to minimize serial interface for data acquisition and power-scalable sampling operation. Programmable BPF and 3-bit PGA give a wide configuration range of gain and bandwidth. The AFE has reduced the noise-energy-area product by a factor of 5-25 times as compared with most state-of-the-art front-ends reported up to date, and has demonstrated feasibility to accommodate a large number of channels toward massive parallel neural recording.

CHAPTER 3

A 128-CHANNEL ANALOG FRONT-END WITH EMBEDDED PROCESSING

In Chapter 2, an extremely low-power 16-channels analog front-end (AHI-16) was introduced. In general, AHI-16 design relied on circuit techniques such as taking advantage of moderate inversion biasing for analog parts and near threshold techniques for digital parts. This has helped increase the energy efficiency of the design, in addition to other techniques such as dynamic operation and power scalable sampling operation. The asynchronous operation of the front-end was shown to be also necessary to lower the power consumption, and to reduce the number of wires and potential interference due to high speed clocking. Using AHI-16 in a 3-D microsystem architecture will enable 128-channels selected from 1,024 sites in a small implementation area.

The scaling, however, towards another order of magnitude in simultaneous recording (>1,000 channels) makes the area consumption a dominant challenge in addition to other consequences of scaling power consumption. From the literature review in Chapter 1, it can be shown that the area consumption is dominated mainly by passive components. For example, in low-noise amplifiers the metal-insulator-metal (MIM) capacitors consumes most the circuit area. Another area consuming block is the site selection, which can be huge if infinite site selection flexibility is required for a particular neuroscience application. It may further waste the channel power consumption especially if there are no active neurons at the site. The next generation neural probe may, therefore, not require site selection if other circuits can be further improved for area efficiency. This also means that AC-coupling needs to be avoided or the current MIM capacitors need to have higher capacitance per unit area (>10 fF/µm^2), whereas the current state-of-the-art technologies provide double MIM capacitors with maximum of 2 fF/µm^2.

So, aside from the circuit physics, would signal processing help in reducing chip area and power at the same time? More importantly, can we use signal processing to assist and improve on-chip calculated parameters by replacing or improving these parameters using off-chip processing at the receiver side? In this chapter, a design approach empowered by signal processing techniques is investigated. This investigation of signal processing can definitely help the designer to decide based on the application and budget whether it is feasible. Moreover, it may help relax the design constraints in analog circuits especially if there performance can be later calibrated digitally inside or outside the chip for further lowering of power consumption.

This chapter starts by introducing the electrode-tissue interface problems and the limitations of AC-coupled design including area and noise performance. Then, a review of some promising approaches using DC-coupling to overcome area consumption challenges is introduced. This is followed by in-depth details of the proposed signal processing algorithm and partitioning of the block into on-chip and off-chip (receiver) parts. The architecture of 128-channel analog front-end circuits is further described to fully implement the algorithm. The rest of sections describe the design, implementation, and measurements of each system component such as LNA, PGA, and compression circuits The proposed analog front-end with embedded processing is very promising in scaling towards 1,024 channels, while reducing the power and area at the same time.

3.1 Design Approaches to Overcome the DC Drift of Electrode-Circuit Interface

In a typical recording channel, the low noise amplifier is not only responsible for noise performance, but also the rejection of the unpredicted slow drift of electrode offset voltage, which is extremely critical to avoid saturation of the amplifier. This slow drift behavior was investigated by many groups (for example [29]) and was attributed to the continuous variation in the environment around the electrode (the cerebrospinal fluid) in addition to continuous reactions in the electrode-electrolyte interface. The offset can vary by hundreds of millivolts [100], which may easily saturate the amplifier. In the next subsections some approaches to overcome the dc drift problems and their limitations are reviewed, then followed by discussing some potential solutions.

3.1.1 Review of Previous Approaches

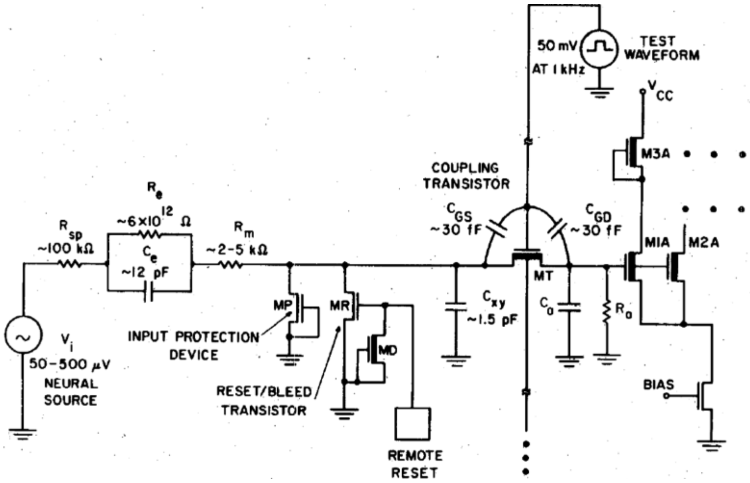


Figure 3.1. Input interface circuit of pre-amplifier in [29]

Many design approaches were attempted to overcome the offset problem. In [29], Najafi and Wise used the junction resistance of a p-n diode to get a high input resistance of their interface circuit in the order of ~10^{10} Ω when biased at near zero volts. Since they calculated the dc resistance of the electrode to be in the order of 10^{13}Ω, dc baseline stabilization was achieved with a dc gain below 1% and a drift below ±0.2mV. To make sure the ac gain is not affected, the authors used capacitors (~2pF) in parallel with the junction resistance. While this approach (Figure 3.1) has a capacitance that impacts the area consumption, it also relied partially on the electrode capacitance to set the AC gain, which may not match for channel to channel and thus may affect the spike sorting algorithm accuracy since it uses the spike amplitude as a main extracted feature to distinguish neurons from one another.

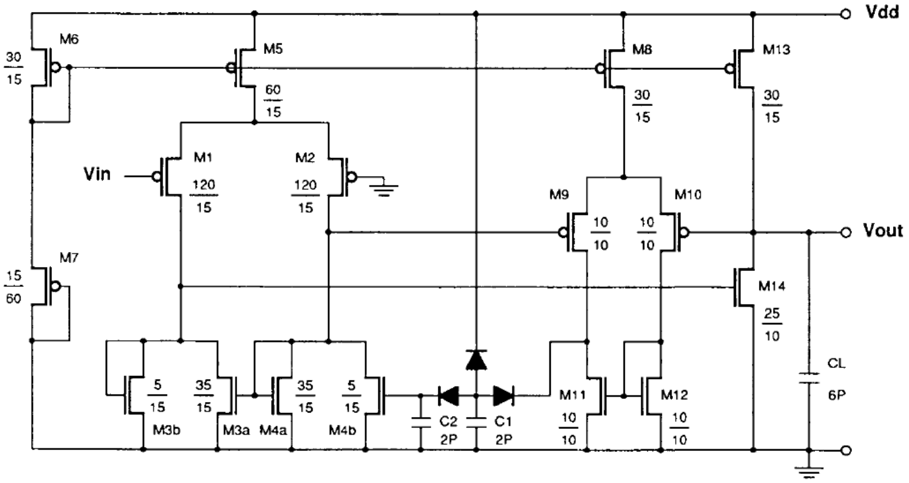


Figure 3.2. The preamplifier with DC feedback as proposed in [34]

In another work, by Ji and Wise [34], a preamplifier with low-frequency feedback was implemented using a diode-capacitor filter as show in Figure 3.2. In this design, the amplifier (M8-M12) feeds back an amplified version of the original signal to the low-pass filter (LPF) formed by the diodes and capacitors; the output of this LPF goes to M4b which draws the excess low-frequency current from the original amplifier (M1-M5) leaving the high frequency components to be further amplified by the second stage (M13-M14). Although this idea is very attractive and simple, it still requires the use of large capacitors as in the previous work. In addition, it the accuracy of setting the frequency corner of the LPF is limited. In [44] Bai implemented an amplifier using a diode at the input to stabilize the dc-baseline, but the measurement results suggested that the approach was inadequate.

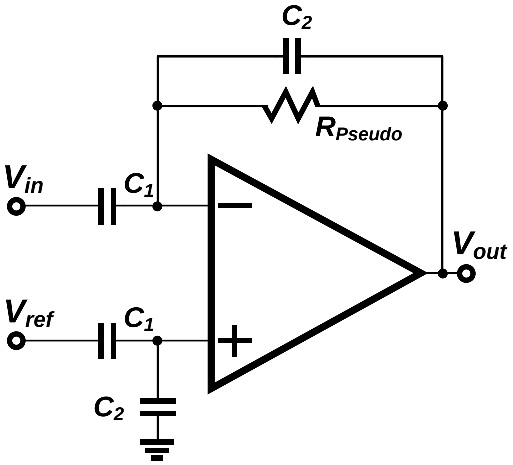


Figure 3.3. Low-noise amplifier using capacitive coupling

So far most of the solution introduces both diodes and capacitors to solve the offset problem, but these diodes are also sensitive to ambient light (called optical sensitivity). In 2002, a capacitive coupling approach (Figure 3.3) was introduced by Olsson [52] to provide a stable repeatable gain across channels, while being more effective for overcoming the offset. The gain can be accurately set with the ratio between input capacitor (C_1) and a feedback capacitor (C_2). In the same year, Harrison published a similar architecture using capacitive coupling [53], but later in his journal article [54] he introduced the details of the actual design criteria taking noise and power performance into account. Using capacitors to reject the offset and set the gain has been very popular and has been used since 2003 in hundreds of neural amplifier designs.

3.1.2 Limitations of Capacitive Coupling Approach on Future Scaling of Neural Recording Channels

Using capacitors between the amplifier and the electrode has proven to be very effective in reducing/removing the DC offset. However, this method introduces even more capacitor values (and size) than the previous methods since the input capacitor is now used to set the gain as well. This is mainly because energy-efficient designs require a gain in the order of 100x at the first stage to achieve an appropriate noise performance. In other publications (reviewed in chapter 2) the input capacitor can take values in the range of 5 to 50 pF, which represent a large portion (>80%) of the amplifier area if MIM capacitors are used (see for example Figure 2.31).

The previous argument may suggest that designers can dramatically reduce the feedback capacitor (C_2) and, therefore, the input capacitor (C_1) accordingly. This approach, however, suffers from a number of limitations that prevent designers from scaling. First, the noise performance is inversely proportional to capacitors area as proven by Harrison in [101]. Since the chip is dominated by the capacitors area, it can be assumed that noise is inversely proportional to the amplifier area. This may also suggest that the designer can increase the power and reduce the area, but this brings the power as the limit of channel scaling. As suggested in chapter 2, both power and area need to be reduced to avoid tissue heating. Second, the gain is set by C_1 and C_2 as shown in Figure 3.3; therefore, if C_2 is made too small that it is close to the parasitic capacitance of the virtual ground nodes, then the gain setting will be no more accurate since the parasitic capacitor changes with process-voltage-temperature (PVT) variations.

In conclusion, capacitive coupling brings channel area scaling to an end unless a reliable on-chip high-density capacitor technology is provided (described in Chapter 4). More creative techniques, therefore, are required to avoid the limit imposed on the use of capacitors while being able of rejecting the electrode DC offset.

3.1.3 Using Advanced Technologies and Mostly-Digital Design Approaches

The recent advances of CMOS technology processes provide great benefits for designing high performance digital circuits, while tremendously reducing the area. On the other hand, analog circuits do not benefit much, especially that threshold voltage does not scale linearly with the process in addition to increasing the leakage current, and reducing the output impedance and gain. Moreover, it was shown in [102] that advanced processes have a gate-leakage mismatch to a level that exceeds the conventional tolerance and sizing up the devices does not help anymore. All these effects led designers of analog circuits to increase the power consumption and use some circuit techniques to maintain performance. For example in [103], operation outside the supply rail was proposed for analog circuits, but care is needed to avoid reliability issues due to high lateral electric field as well as other physical parameters.

As more digital circuits dominate the system, some designers strive for using all- or mostly-digital implementation for their target function. Therefore, there is a trend in transferring all signal processing tasks, such as filtering, to the digital domain. On the other hand, some designs (such as the work presented in chapter 2 and chapter 4) try to use operating points such as moderate and weak inversion to optimize the performance although these techniques may suggest increasing the current consumption to achieve reasonable matching, CMRR, and linearity.

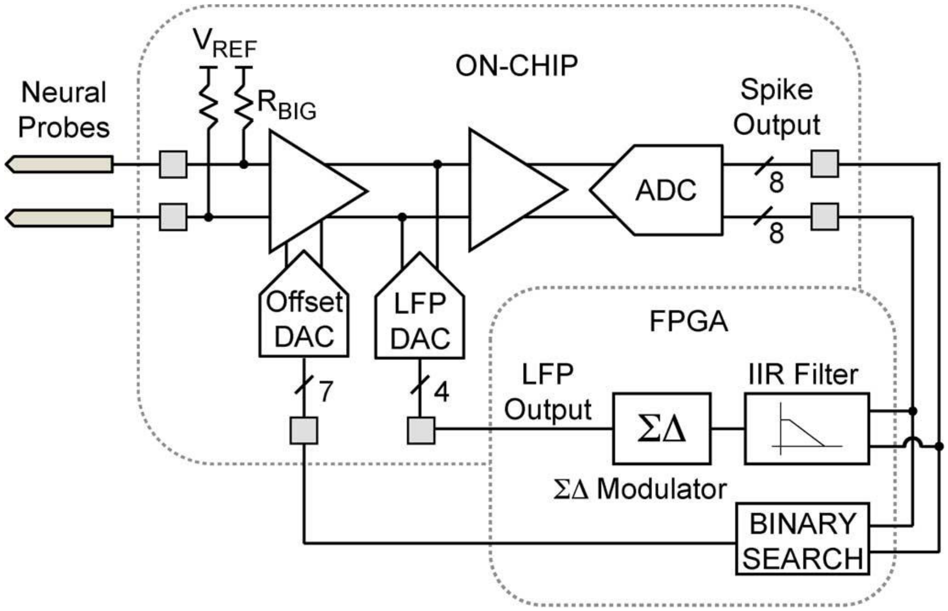


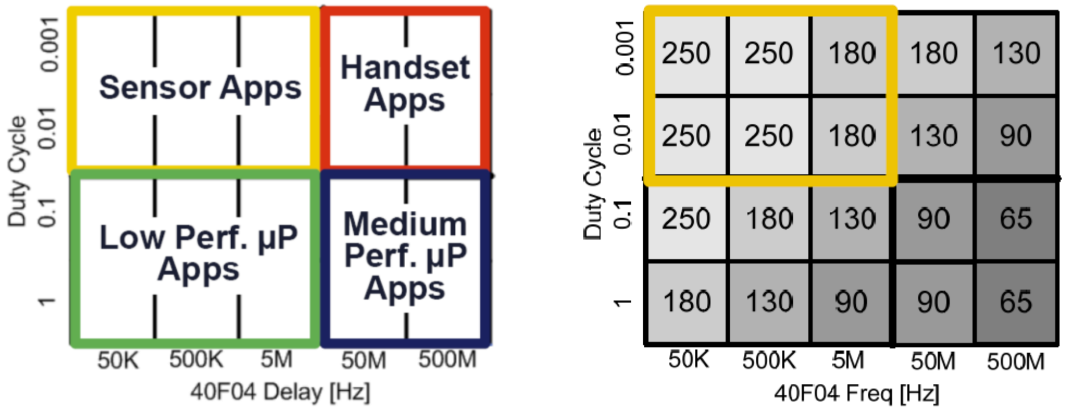
Figure 3.4. Digitally-intensive technique for neural recording using 65nm process

For neural recording application, the most-digital circuit implementation was only attempted by Rabaey’s group in Berkeley using a 65 nm CMOS process [72]. The design avoided any coupling capacitors by using a digitally intensive mixed-signal feedback technique to reject the DC offset. As shown in Figure 3.4, the digital output of the ADC is filtered using a digital low-pass filter, converted back to analog, and fed back to the low-noise amplifier. The overall implementation area was around 0.013 mm^2 although some external FPGA digital filter was still needed. When projected to 0.25µm CMOS process, this design may consume at least 0.2 mm^2 of area. In other words, using a most-digital approach can help scale the area by the squared ratio of the process lambda $(\lambda_{old}/\lambda_{new})^2$. This design in [72] consumes relatively large power consumption and the power intensity is even larger, which may bring the issue of tissue heating with further scaling of area. In addition, it can be seen that gain variation from channel-to-channel may be severe since the gain is based on open-loop architecture.

3.1.4 The tradeoff of Technology Scaling for Neural Recording Applications

Scaling the technology provides area scaling when using digitally-intensive implementations, but it also suffers from some limitations and concerns related to the fact that neural recording is a low-performance application. To clarify this further, a group at the University of Michigan (Prof. D. Sylvester group) conducted an experiment that aimed to show the optimum technology choice for a particular system considering both its duty cycle and operating frequency to achieve maximum energy efficiency.

First, power density increases with scaling and this brings the safety concern and tissue heating potential. Second, leakage power in digital circuits is becoming severe with technology scaling, and scaling of supply voltage is limited since the threshold voltage does not scale linearly. For battery- or wirelessly-powered low-performance applications this leakage component is a concern for achieving energy-efficiency. Third, using very advanced technology nodes may not be cost-effective for low-performance applications; therefore, it is not necessarily the best option for neural recording. Figure 3.5 shows the outcome of their experiments, which suggest that scaling beyond 130 nm technology may not be of great advantage to low-performance low duty-cycle applications. This is mainly because leakage power will dominate the active power.



Duty Cycle	50K	500K	5M	50M	500M
0.001	250	250	180	180	130
0.01	250	250	180	130	90
0.1	250	180	130	90	65
1	180	130	90	90	65

Figure 3.5. CMOS technology scaling and implementation suggestions (source: a tutorial lecture provided by D. Sylvester at the University of Michigan)

The previous argument suggests that digitally-intensive approach may not allow the optimum scaling of EAE (energy-area efficiency); therefore, analog circuit blocks need to dominate the channel components. A great tradeoff may be to use the concept of digitally-assisted analog design, which is described in the next subsection.

3.1.5 Potential of Digitally-Assisted Analog IC Design

Digitally-assisted analog techniques are also using digital circuits to improve energy efficiency but from a different perspective. The purpose of these techniques is to assist, enhance, and/or relax the precision of analog circuitry; not necessarily replacing them. While this method (referred to as digitally-assisted analog design, DAAD) may benefit from technology scaling, it focuses more on energy efficiency at a specific technology node; therefore, it can avoid some of the limitations of digitally-intensive approaches such as the dominance of leakage power. Digital signal processing is used in DAAD to overcome the shortcomings of the analog design and may, therefore, enable a new generation of interface electronics [104].

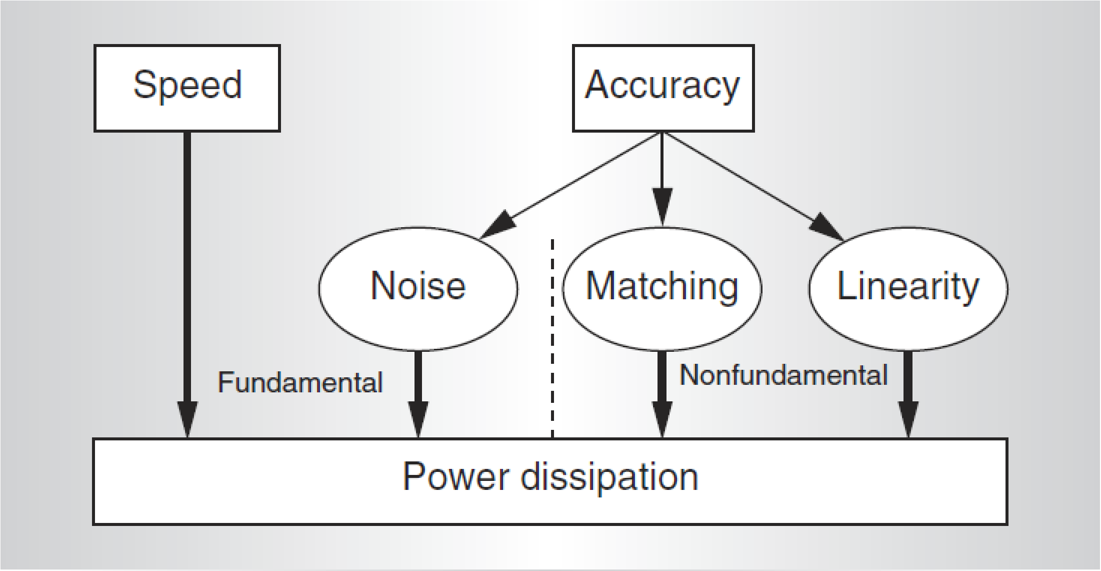


Figure 3.6. Analog circuit design trade-offs [98]

It is important to note that analog power consumption is fundamentally determined by the noise and speed requirements; however, as shown in Figure 3.6 there are some non-fundamental requirements, such as matching, and linearity, that cause the final implementation power consumption to grow more. This non-fundamental part of power consumption continues to grow as the supply voltage shrinks.

DAAD is focused on making analog circuit consume power only for the fundamental requirements (noise, speed, etc), while calibrating or relaxing the non-fundamental analog precision requirements with the assistance of digital approaches

[105]. In RF circuits, for example, there have been great efforts to push more digital blocks towards the antenna to assist and reduce the complexity of analog parts [104]. Another example is the crest factor optimization as in [106]. In this work also, as shown in the next section, a filter is needed to construct a delay circuit with a linear phase; however, power consumption can be saved using this method by designing a single mono-pole filter and linearizing the phase later in the digital domain at the receiver; this is given that the non-linear phase transfer function is already known at both the design and chip measurement stages.

In general, the direct benefits of digitally-assisted analog design can be summarized as follows [105]:

1) Higher energy efficiency can be achieved since it limits the power consumption in the precision parts of analog design;
2) This simplification eliminates the noise caused by the involved precision components; thus, improves noise performance; and
3) This simultaneous reduction of power and noise in addition to area makes this approach very attractive for low- and medium- performance applications.

In this chapter, different techniques, including DAAD, are brought together to improve energy and area efficiencies such as:

1) Using a system architecture approach rather than component architecture to drive the design process and judge the use of both analog and digital signal processing techniques to achieve high energy and area efficiency;
2) Thinking about what needs to be performed on-chip and what can be performed off-chip after telemetry or at the microprocessor platform. This may dramatically improve energy-efficiency of the design;
3) Assisting analog circuits with digital calibration to relax some precision parameters such as linearity, mismatch, and process variation when appropriate; and
4) Employing data recovery algorithms at the receiver, which will also help decide the analog, mixed, and digital portions of the implementation and not necessarily remove the analog blocks.

The next two sections describe the signal processing algorithm used to relax the analog circuit precision requirements, and the design details of the analog front-end.

3.2 Signal Processing Algorithm for Digitally-Assisted Analog Front-End

There have been many approaches, as described previously, to overcome the electrode offset variation; however, there is so far no attempt to use analog signal processing at the interface circuit to solve this problem. This section introduces an algorithm that has a potential to remove the offset while providing data compression.

Parameter	***Definition***
A_{FSO}	Original signal full-scale value
$A_{FS\Delta}$	Compressed signal full-scale value
F_{max}	Maximum frequency content in the signal
F_{min}	Minimum frequency content in the signal
T_{min}	Period of the highest tone = $1/F_{max}$
T_{max}	Period of the lowest tone = $1/F_{min}$
D	Delay
ND	Normalized delay to T_{min} ($ND=D/T_{min}$)
F_S	Sampling frequency
OSR	Oversampling ratio ($Fs/2F_{max}$)
CR	Compression Ratio = $A_{FSO}/A_{FS\Delta}$
RF	Reconstruction factor = $(D.F_S)^{-1}$
A_{min}	Minimum value of LSB

Table 3.1. Parameter definition for delta algorithm

3.2.1 Analog Compression Algorithm

The parameters used within this section are introduced in as in Figure 3.7 and Table 3.1. Assuming a delay element D, the effective input across the difference amplifier is given by:

$$\Delta V_{in} = A_{FSO}\big[\sin(\omega t) - \sin\big(\omega(t-D)\big)\big] = 2A_{FSO}[\sin\left(\frac{\omega D}{2}\right)\cos\left(\omega t - \frac{\omega D}{2}\right)] \qquad \textbf{Equ. 3.1}$$

The time-dependent part of Equ. 3.1 have almost π/2 phase shift (cosine instead of sine); therefore, the other part of the equation represents the new amplitude $A_{FS\Delta}$. It can be shown that when the normalized delay (ND) is << 1, $A_{FS\Delta}$ and the compression ratio (CR) can be given by:

$$\frac{\pi D}{T} \ll 1 \rightarrow \; A_{FS\Delta} \approx 2A_{FSO}\frac{\pi D}{T} \rightarrow CR = \frac{A_{FSO}}{A_{FS\Delta}} = \frac{T}{2\pi D} = \frac{1}{2\pi D F_{max}} \qquad \textbf{Equ. 3.2}$$

If this relationship is plotted, as in Figure 3.8, it can be seen that there is a linear dependence between CR and ND under the previous assumptions. It can be shown also that a signal that has multiple frequency components can provide better CR due to phase interference.

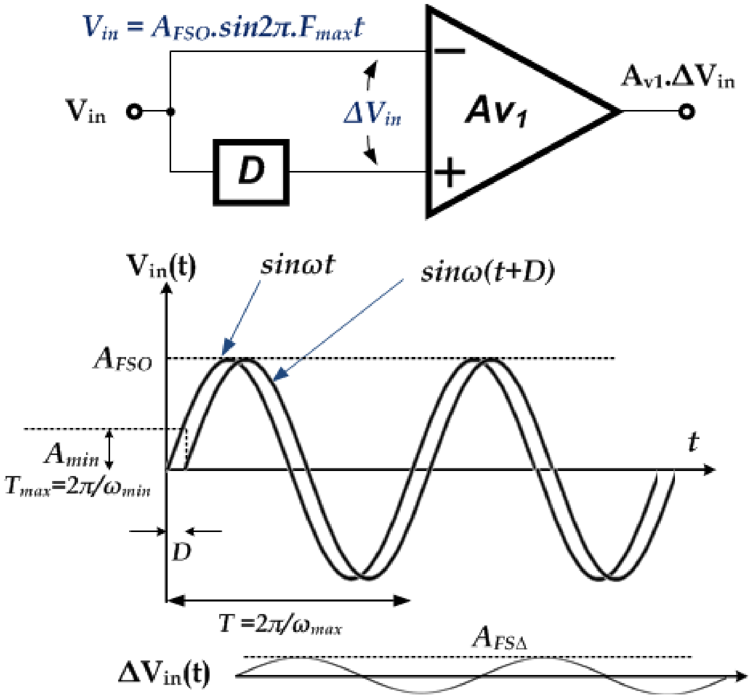


Figure 3.7. Parameter definition for delta algorithm

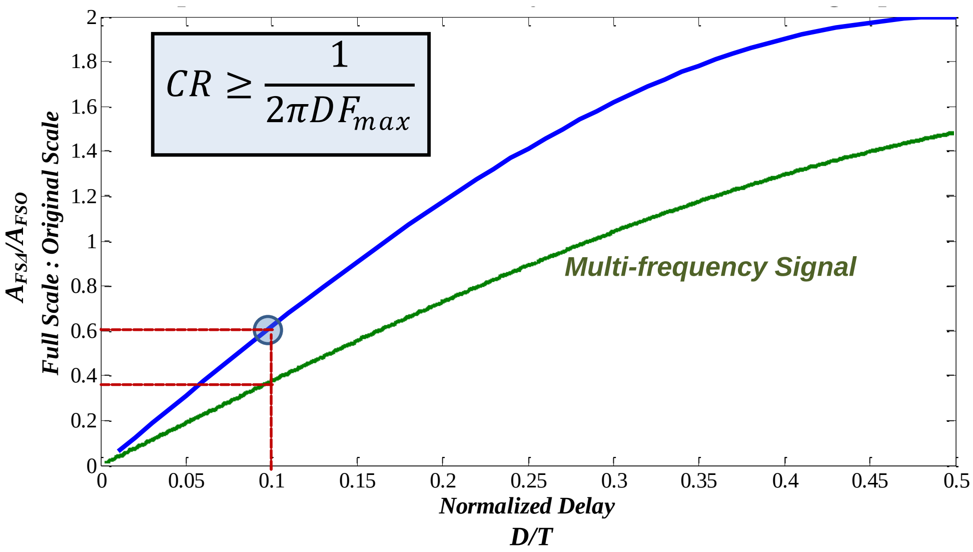


Figure 3.8. Amplitude reduction for a delayed-differential analog input versus delay value

To use this method for compression, CR needs to have a worst-case value of 1. Therefore, the maximum delay that should be used is given by:

$$CR = 1 \ \rightarrow \ D_{max} = \frac{1}{2\pi F_{max}} \qquad \textbf{Equ. 3.3}$$

It is worth mentioning that this delta modulation does not change the frequency of the signal; however, it causes amplitude and phase changes. This means that Nyquist sampling rate can still be applied; however, the choice of the sampling frequency may impact the reconstruction at the receiver. To understand this, assume that the reconstruction block received a sample $\Delta V_{in}(\tau)$ which was originally generated by:

$$\Delta V_{in}(\tau) = V_{in}(\tau) - V_{in}(\tau - D) \qquad \textbf{Equ. 3.4}$$

In other words, the actual signal can be constructed by:

$$V_{in}(\tau) = \Delta V_{in}(\tau) + V_{in}(\tau - D) \qquad \textbf{Equ. 3.5}$$

So, the previous samples are required to get $V_{in}(\tau\text{-}D)$. Therefore, Equ. 3.5 can be expanded to:

$$V_{in}(\tau) = \Delta V_{in}(\tau) + \Delta V_{in}(\tau - D) + \Delta V_{in}(\tau - 2D) + \cdots + \Delta V_{in}(\tau - nD) \qquad \textbf{Equ. 3.6}$$

Until (τ-nD) becomes less than or equal zero, which is the time when the recording started. Since the choice of *D*, according to Equ. 3.2 and Equ. 3.3, directly affects the compression ratio (CR), then if the sampling time is D (because of what Equ. 3.6 suggests), this is already oversampling. The question here is that is it really necessary to sample with a frequency $F_S = D^{-1}$? According to Nyquist sampling theorem, the signal can be reconstructed even if F_S < D^{-1} but more than 2F_{max}; therefore, it may not be actually necessary to oversample the signal because in Equ. 3.6 some of the successive terms (according to the choice of D) may be considered as having the same values. Equ. 3.6 is also very important because it tells the designer that at the reconstruction an integer multiples of delayed delta signals need to be summed to successfully get the correct original value. So, if the sampling frequency (F_S) is less than or more than D^{-1}, a reconstruction factor (*RF*) should be multiplied to the sum of samples. Therefore, RF is given by:

$$RF = \frac{1}{f_S D} \qquad \textbf{Equ. 3.7}$$

Using the delay and oversampling technique can provide extra number of bits in the digital domain; therefore, choosing D and F_S is also relaxing the actual number of bits that need to be implemented as part of the ADC architecture. There is, however, another factor that is influencing the choice of D; at the scale of each sample it can be seen that the delta operation performs a high pass filter and that is why it can reject the offset. Low-frequency components are going to reconstruct because of the sigma operation at the receiver, but still the ability to see these tiny voltages depends on how small the LSB of the ADC is. Therefore, using the same figure (Figure 3.7) and considering now the lowest frequency of interest, A_{min} is determined as the maximum component that can be contributed by the lowest frequency component in any given sample. Therefore, according to the lowest cutoff frequency, LSB should not exceed the value of A_{min}. The value of A_{min} can be determined by:

$$A_{min} = A_{FSO\_LF} \cdot 2\pi F_{min} \cdot D \qquad \textbf{Equ. 3.8}$$

For local-field potential (LFP), the amplitude (A_{FSO_LF}) is at least 5-10 times larger than extracellular action potential (EAP) depending on the electrode size. Therefore, from Equ. 3.2 and Equ. 3.8, the maximum LSB size (assuming $A_{FSO\_LF} \geq 5A_{FSO}$) can be determined as:

$$LSB = \frac{A_{FSO\_LF}}{A_{FSO}} \cdot \frac{F_{min}}{F_{max}} \cdot A_{FS\Delta} = 5 \cdot \frac{F_{min}}{F_{max}} \cdot \frac{A_{FSO}}{CR} \qquad \textbf{Equ. 3.9}$$

This also determines the minimum resolution required for the ADC as follows:

$$NOB = log_2\left(\frac{A_{FS\Delta}}{LSB}\right) = log_2\left(\frac{F_{max}}{5F_{min}}\right) \qquad \textbf{Equ. 3.10}$$

So, now the frequency dynamic range sets the minimum number of bits required. It can be calculated that at least 10-bit resolution is required for recording frequencies from 1Hz to 6kHz at the same channel. On the other hand, if the interest is from 600Hz to 6kHz (for spike sorting) less than 4 bits are required. Of course there are extra bits that are gained because of the compression and the oversampling, which are given by:

$$NOB_{OSR} = 0.5 \cdot log_2(OSR)\,, NOB_{CR} = log_2(CR) \qquad \textbf{Equ. 3.11}$$

Therefore, the overall resolution of the recording can be written as:

$$NOB_{REC} = log_2\left(\frac{A_{FS\Delta}}{LSB} \cdot CR \cdot \sqrt{OSR}\right) \qquad \textbf{Equ. 3.12}$$

From Equ. 3.2 through Equ. 3.12, the parameters to determine the recording dynamic range, frequency bands, and resolution can be narrowed down to CR (or D), OSR, and LSB. These parameters can be chosen or programmed to improve the energy efficiency of the circuit as follows:

1. The choice of D or CR provides dynamic range compression, which allows low-voltage (0.5V) analog design, and provide extra bits of resolution at the receiver;
2. The choice of OSR (or F_S) allows higher signal-to-noise ratio, reduces the ADC size, and benefits from technology scaling; and
3. The choice of LSB acts like a high-pass filter by setting the low-frequency amplitude.

3.2.2 Algorithm Validation and Accumulative Error Reset

The proposed algorithm has been simulated and validated using Matlab and Simulink including the data reconstruction. The validation helped the realization that an accumulative error is taking place at the ADC and needs to be corrected or reset from time to time as will be discussed later. The following shows the step-by-step validation with illustrative figures.

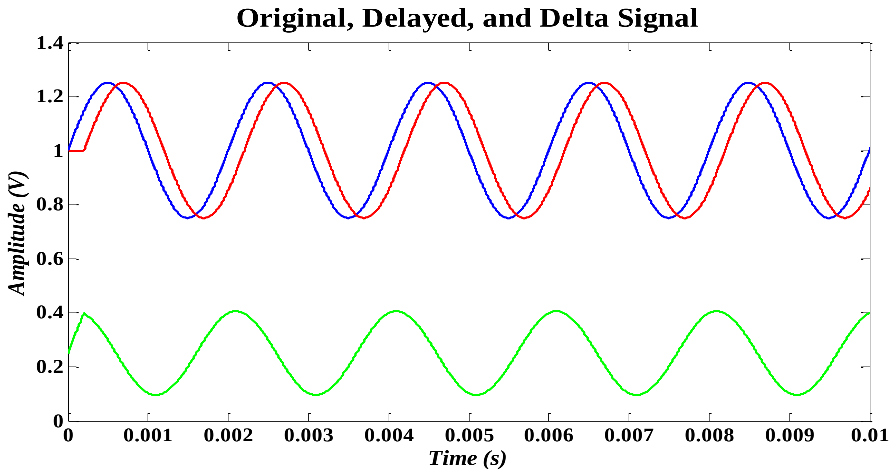


Figure 3.9. Original (blue), delayed (red), and the resulting delta signal (green) at certain bias according to a 0.5V supply operation

1. A DC offset is assumed at the electrode (say 1V) and the delta operation is performed. The amplifier output is going to be the difference plus some output DC value (assumed here as 0.25V, which is half the target supply). This step is illustrated in Figure 3.9
2. The delta signal is then sampled at OSR of 10 using a 5-bit ADC that includes an accumulative error reset (AER) function, which is very critical to keep an accurate reconstruction and effectively enable the oversampling to add more bits at the receiver. AER is essential because the reconstruction is an addition process; therefore, if there is no way to continuously track the values below LSB and accumulate them correctly, the accuracy (resolution) of the sample will be deteriorated. AER also adds an additional bit to the ADC resolution, so effectively the resolution now is 6-bit. The block diagram of the ADC with AER function is shown in Figure 3.10 and the digital signal is shown in Figure 3.11
3. Finally, the signal can be reconstructed at the receiver using the block diagram in Figure 3.12. The reconstructed signal is shown before and after averaging (due to oversampling) in Figure 3.13. The OSR of 5 should provide ~1.66 bit extra resolution; therefore, it can be seen that the error has been reduced from 1 LSB range to ~0.37LSB. In addition, the value at the receiver is expanding to 7-

bit because of the addition process; thus, the actual resolution at the receiver is ~8.66 bits.

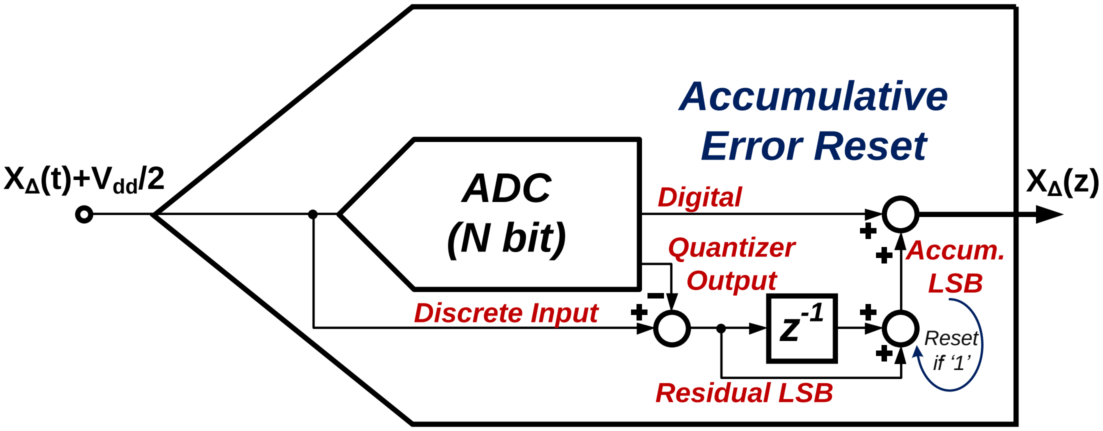


Figure 3.10. ADC with accumulative error reset (AER)

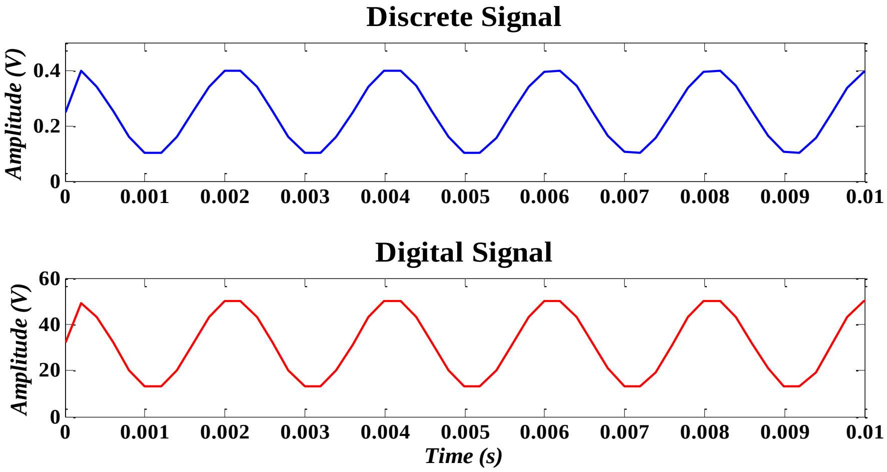


Figure 3.11. The discrete and digital signal when sampled at OSR of 10

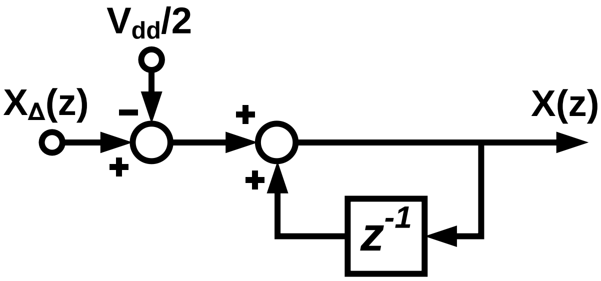


Figure 3.12. The sigma function used for signal reconstruction

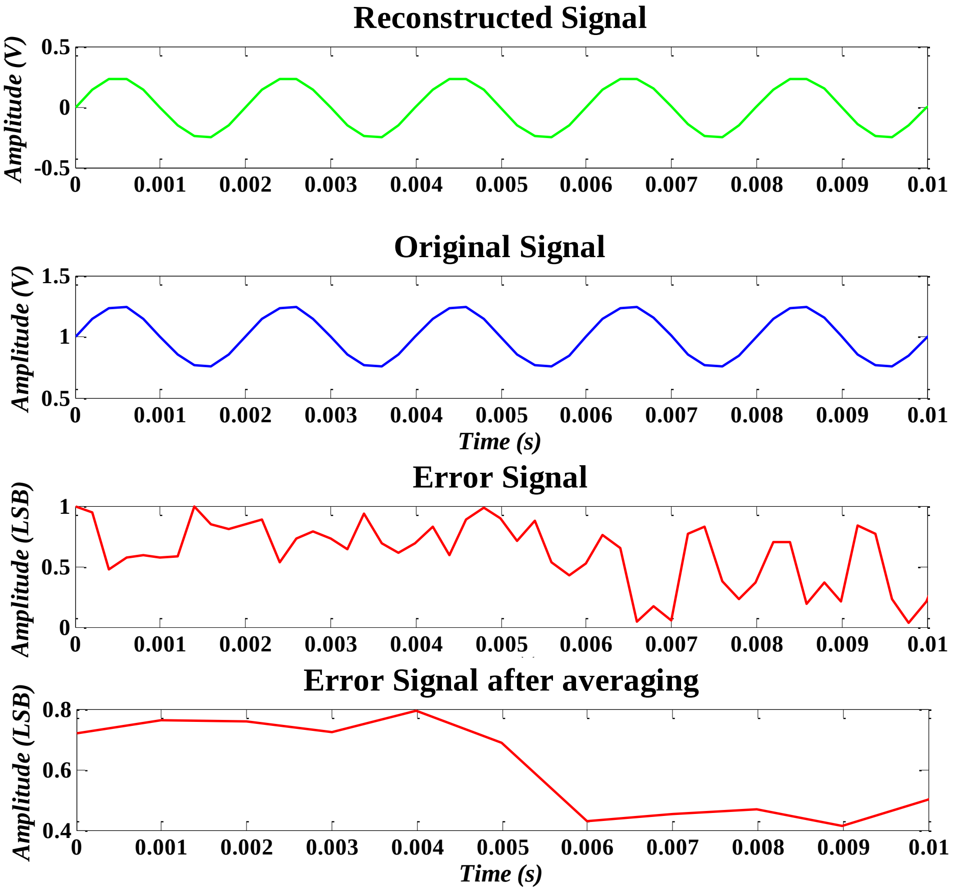


Figure 3.13. The reconstructed and original signal showing the LSB error reduction when averaging every 10 samples

The next subsection applies the algorithm on the acquired neural signals and provides some thoughts about the usefulness of automatic error reset function, which may be replaced by a high-pass filter after the reconstruction at the receiver with some loss of information (an overall lossy compression)

3.2.3 Algorithm Application on Neural Signals

Applying the signal processing algorithm on neural signal may provide some insight for modifying or relaxing the circuit techniques further without losing information or with a slight penalty, especially when it is appropriate for neuroscientists.

A prerecorded neural signal[6] (Figure 3.14) was used as follows:

1. Normalized and interpolated by a 100 times to mimic an analog behavior;
2. A delay (D) of 6.4 µs was applied; this is 10 times smaller than $1/F_{max}$;
3. A delta signal (original minus delayed) was formed (Figure 3.15) with an offset of half the propose supply (0.5 V);
4. Oversampling ratio (OSR) of 5 was used to digitize the signal at 6-bit resolution (Figure 3.16). No AER was used at this point;
5. After regular reconstruction algorithm, the signal is shown in Figure 3.17. It includes a very low-frequency large amplitude drift, which is mainly due to the initial sample error accumulation;
6. A first order high-pass elliptic filter was used to remove this low-frequency drift. The low-pass filter in Figure 3.18 detects the drift, and the final signal was formed by subtracting the drift from the reconstructed signal in Figure 3.17;
7. An alternative to the step above is to get the average slope between the initial and final samples and accumulative in the reverse direction during reconstruction; and
8. Figure 3.19 shows the final outcome of this method compared to the original signal, and the error signal which is dominated by a low-frequency signal that is much smaller in amplitude than LFP signal. While the original signal can be perfectly retrieved when using AER, avoiding the on-chip implementation of AER makes the design a lot easier.

In this experiment, the digital code span of reconstructed signal is ±1000 and the error signal is ±20 codes, which provides ~5.64 bits or resolution. When including the effect of OSR, the overall resolution becomes about 7 bits with the advantage that a high recording dynamic range is achieved.

Finally, Figure 3.20 shows the frequency spectrum of the original neural signal together with the reconstructed signal (without AER) and the error signal. It is worth mentioning that the neural signals show an attenuation trend of 20 dB/decade, which is a very useful feature when designing the circuits as shown in the coming sections.

[6] Neural signal was provided by Dr. J. Berke's lab at the University of Michigan

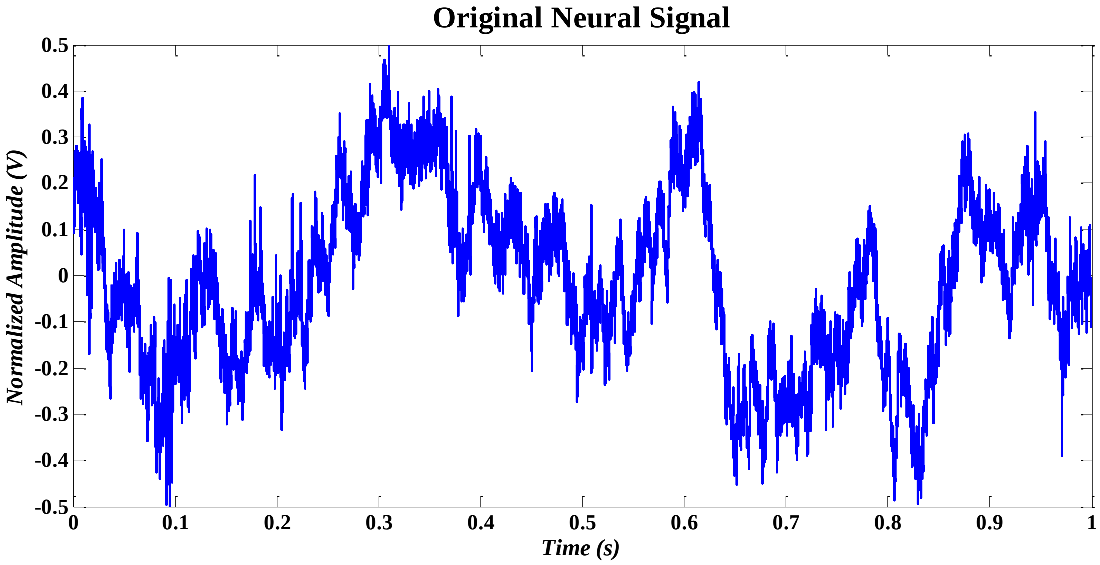


Figure 3.14. Neural signal sampled at 31.25 kHz and 16-bit resolution

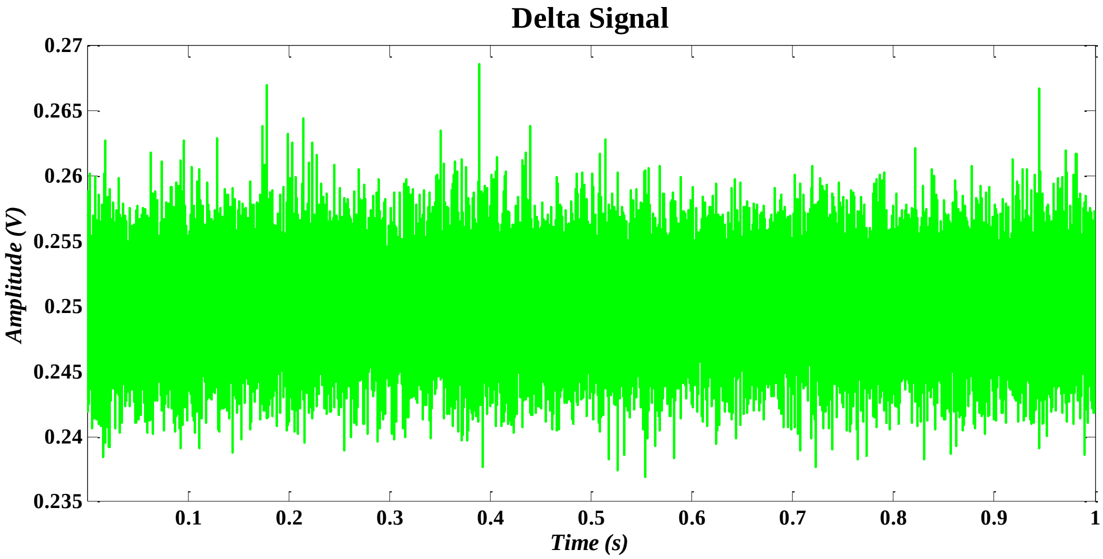


Figure 3.15. Neural signal after delta function

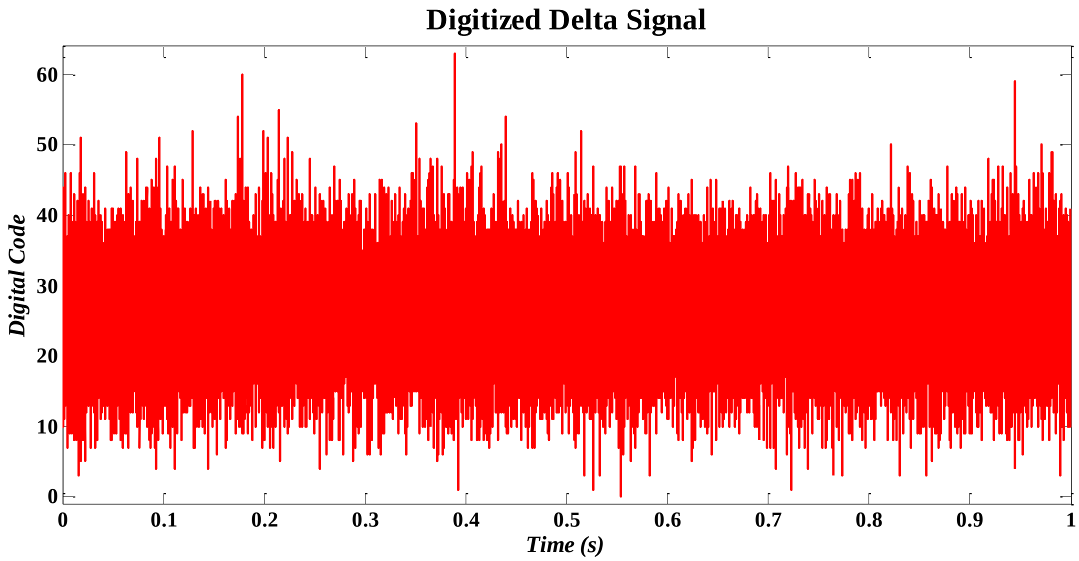


Figure 3.16. 6-bit digitized delta signal

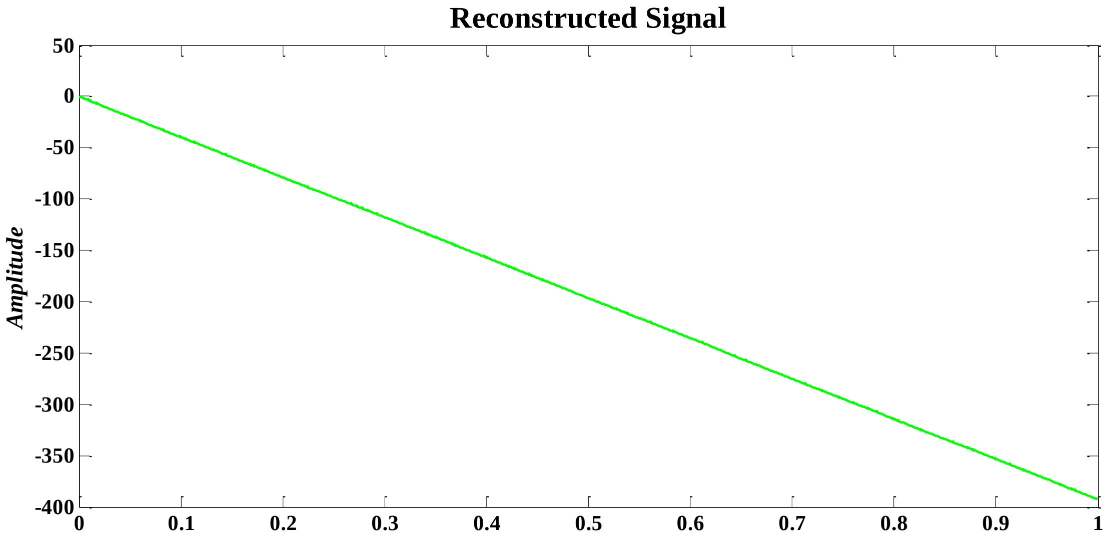


Figure 3.17. Reconstructed signal without accumulated error reset. Signal cannot be seen because it is superimposed on a large low-frequency drift

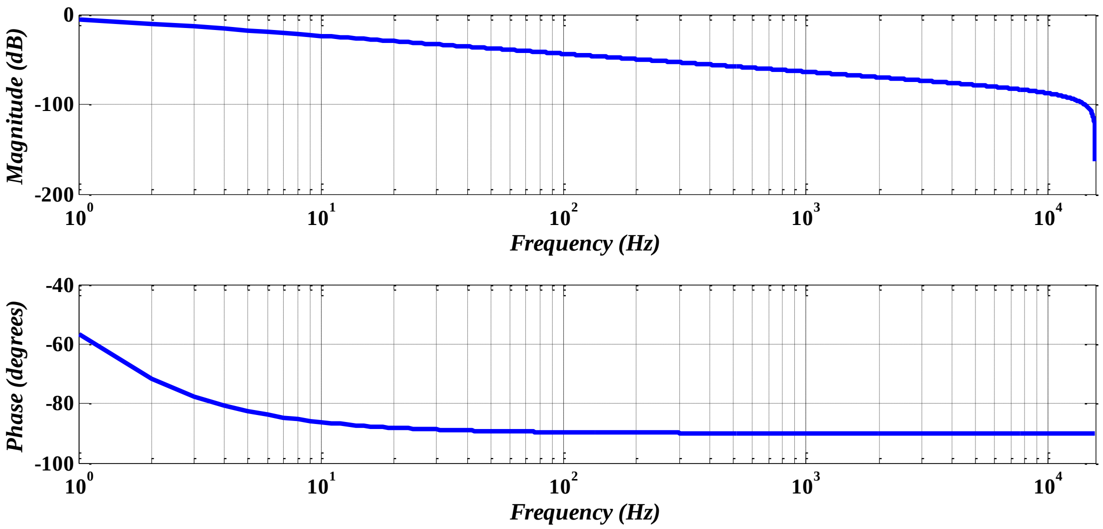


Figure 3.18. A first order elliptic digital low-pass filter for detecting the offset drift below 1 Hz

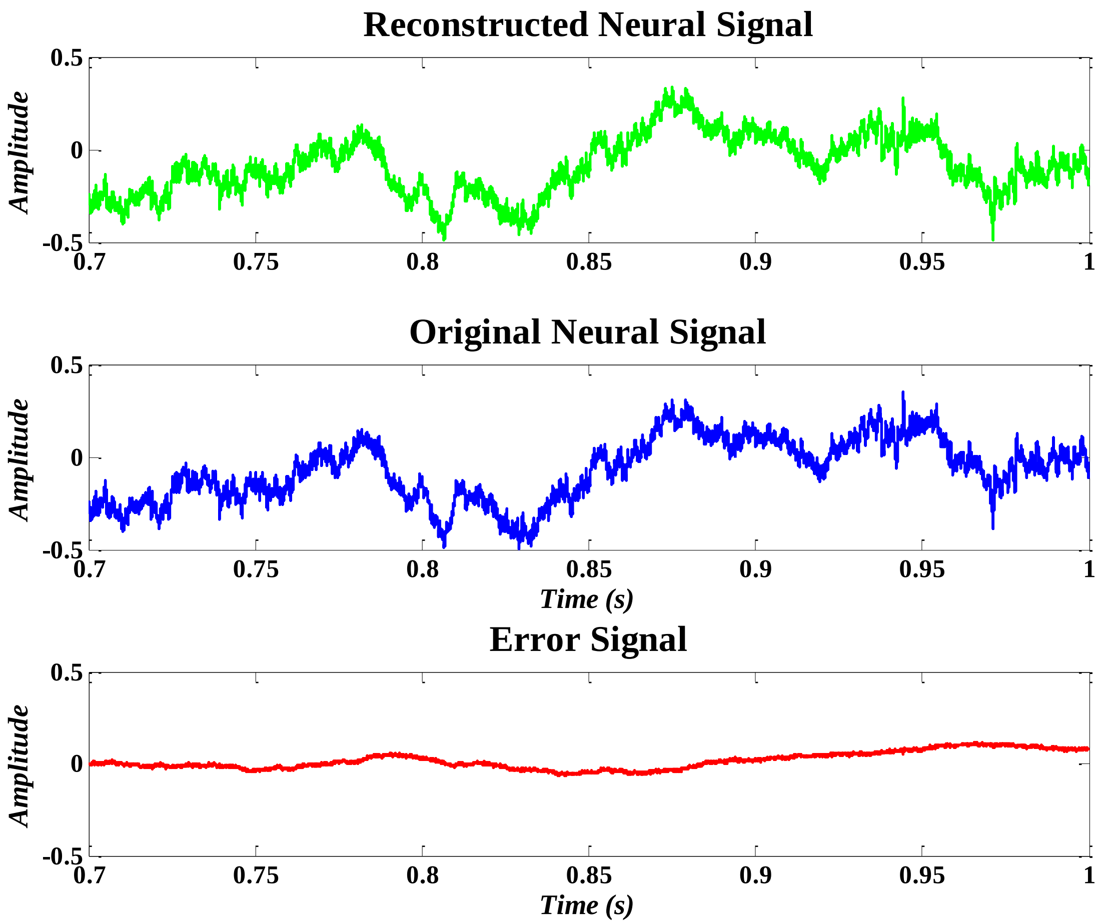


Figure 3.19. The outcome of the reconstruction algorithm including the elliptic filter

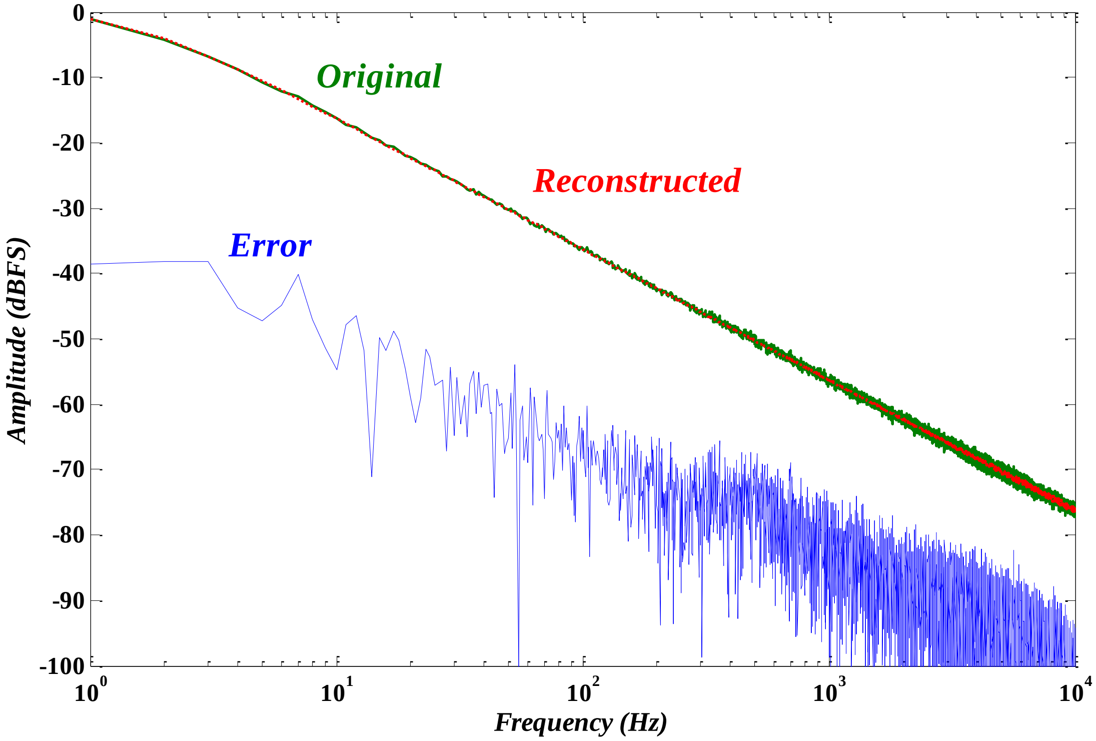


Figure 3.20. Frequency Spectrum of original, reconstructed, and error signals

3.2.4 An Estimation of Power and Area Improvement

Regardless of the technology used, the dynamic range compression may provide an analog power reduction by half (assuming CR ≥ 2) by scaling the supply voltage. The expected digital power consumption because of this technique (***P***$_{\boldsymbol{\Delta}}$) needs to be compared with the conventional technique (***P***$_{\boldsymbol{conv}}$). The digital power consumption (***P***$_{\boldsymbol{dig}}$) is usually given by:

$$P_{\text{dig}} = \frac{1}{2} C V^2 f \qquad \textbf{Equ. 3.13}$$

For the ADC (assuming SAR operation), given C$_{\text{comp}}$ as the dynamic comparator capacitance, C$_{\text{U}}$ as the DAC unit capacitor, and C$_{\text{reg}}$ as the register capacitance, then there are 2$^{\text{NOB}}$ unit capacitors, and NOB registers each switching NOB times each cycle. Therefore, the estimated digital power is given by:

$$P_{\text{conv}} = P_{DAC} + P_{Cmp} + P_{SAR}$$

$$= \frac{1}{2} V_{DD}^2 f_S \left(2^{NOB} C_U + NOB.C_{comp} + NOB^2 \cdot C_{reg}\right) \qquad \textbf{Equ. 3.14}$$

In the proposed approach, the following substitutions can be done:

1) $V_{DD} \rightarrow 0.5V_{DD}$

2) $f_S \rightarrow OSR \cdot f_S$

3) $NOB \rightarrow NOB - log_2\left(CR \cdot \sqrt{OSR}\right)$

The estimated power consumption is then:

$$P_\Delta = \frac{1}{8} V_{DD}^2 OSR. f_S \left[\frac{2^{NOB} C_U}{CR \cdot \sqrt{OSR}} + \left(NOB - log_2\left(CR \cdot \sqrt{OSR}\right)\right). C_{comp} + \left(NOB - log_2\left(CR \cdot \sqrt{OSR}\right)\right)^2 \cdot C_{reg}\right] \quad \textbf{Equ. 3.15}$$

From both Equ. 3.14 and Equ. 3.15, a power reduction factor (PRF) can be defined as:

$$PRF = \frac{P_{\text{conv}}}{P_\Delta} = \frac{4\left(2^{NOB} C_U + NOB. C_{comp} + NOB^2 \cdot C_{reg}\right)/OSR}{\left[\frac{2^{NOB} C_U}{CR \cdot \sqrt{OSR}} + \left(NOB - log_2\left(CR \cdot \sqrt{OSR}\right)\right). C_{comp} + \left(NOB - log_2\left(CR \cdot \sqrt{OSR}\right)\right)^2 \cdot C_{reg}\right]} \quad \textbf{Equ. 3.16}$$

This equation can give a sense of how far can the OSR (Figure 3.21) be increased before the proposed approach consumes higher power than the conventional approach.

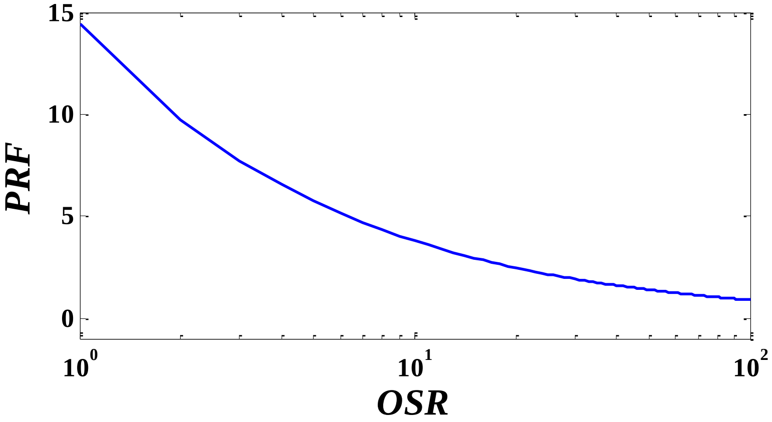


Figure 3.21. OSR versus PRF plot showing that above 80 oversampling ratio there is not power advantage

It should be also noted that this approach reduces the area by using dc-coupling and minimize the capacitors size involved in the ADC for a given resolution (NOB).

The previous steps demonstrated the benefits of signal processing algorithms, in which the off-chip components are ideal. However, the on-chip components such as the delay element may not be as ideal. The actual implementation may not be straightforward or may not be possible given the power and area constraints. So, there is another place for signal processing to compensate for that at the receiver as well.

In the following sections, the overall circuit architecture and implementation for each block is described including methods to deal with non-idealities. Some of these techniques rely on digital assistance at the receiver.

3.3 0.5V 128-Channel Analog Front-End Architecture

3.3.1 Overall System Architecture

Approach	*Effect*	*Target Value*
128-Channels	Massive parallel neural recording	**EAE>40x10^{15}**
DC-Coupling	Area reduction	**0.01mm^2 (5x)**
Analog Delta Compression	Full scale signal and supply reduction	**+2-bit resolution**
Differential Low-Noise Amplifier	Improve CMRR/PSRR	**>80dB**
Programmable Gain	Maximize DR (near Rail-Rail)	**40-80dB**
Oversampling ADC	Area reduction and improved SNDR	**OSR: 4-32**
Digital Compression	Data rate compression, power reduction	**<50%**
Serial Asynchronous Interface	Interference and wires/channel reduction	**2:128**
Chip Test and Calibration	Digital assistance advantages	**On and off-chip**
Channel Programming	Power saving when channel not needed	**Per channel**

Table 3.2. The approaches and effects of using the architecture in Figure 3.22

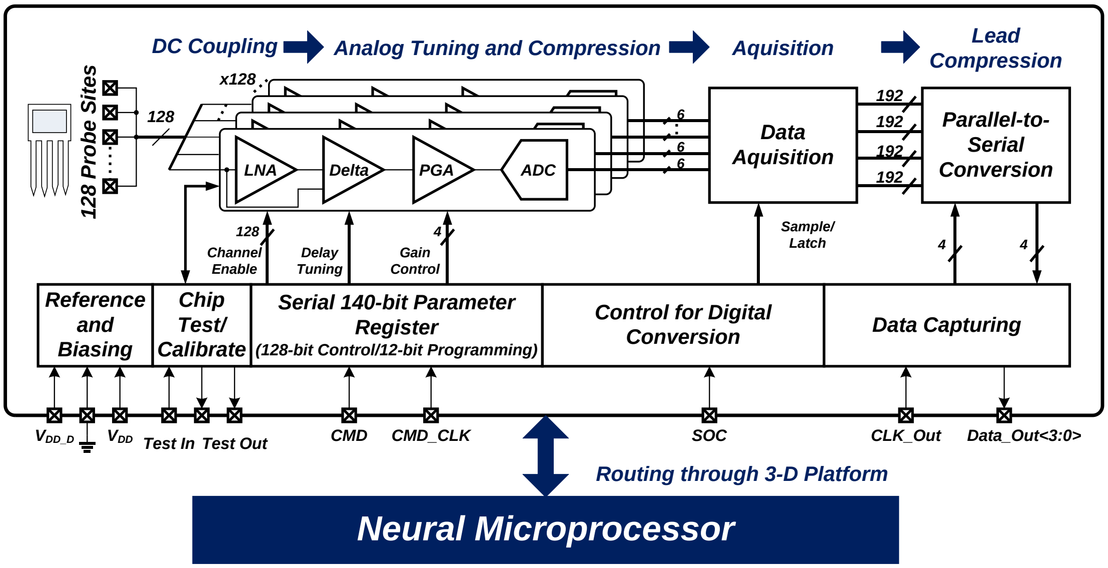


Figure 3.22. 128-channel analog front-end with embedded processing

The design choices and target goals for this front-end are summarized in Table 3.2, which were mainly driven by maximizing energy and area efficiency (EAE). Figure 3.22 shows the architecture of the proposed analog front-end with embedded processing (AEP) using the approaches in Table 3.2. In this design there is no site selection circuit, but if the 128 channels can be fitted within the area budget, the user may be given the capability to select/deselect any individual channel using a 128-bit serially-programmable register. This would allow the full recording of 128 channels, and provide more flexibility in selecting any subset of channels, which cannot be achieved with a conventional site selection scheme.

Each channel consists of a fully-differential low-noise amplifier, an analog delta compression block, a programmable gain amplifier, and a 6-bit ADC. In this particular implementation, each 32 channels are grouped into a data acquisition block followed by a parallel-to-serial converter. The four 192-bits groups are further captured using 4 different output ports (Data_Out<3:0>).

A neural processor at the 3-D platform (similar to Figure 1.23) can interact with each AEP using 8-bit digital interface to program the registers responsible for channel activation, gain control, and delay control. The embedded processing in the proposed AEP simplifies the interaction of the digital conversion process; the neural processor provides only one start-of-conversion (SOC) signal that is internally routed to all ADCs, and the latched/sampled data is clocked-out by the processor using a 5-wire interface.

Relaxing the precision requirements of some analog components is done externally at the receiver; the idea here is that each component is tested and the inverse of the measured transfer function is used later at the receiver to account for any non-idealities such as offset, non-linearity, mismatch, etc.

3.3.2 Low-Voltage Current-Mode Mixed-Signal Design

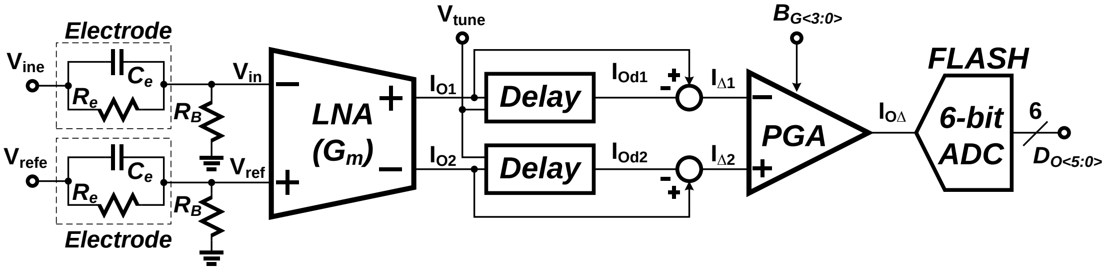


Figure 3.23. Current-mode analog front-end

To enable the features listed in Table 3.2, the analog front-end (AFE) needs to operate at low-voltage so that the low-performance digital circuits operate at the optimum point for energy efficiency. More scaling of the supply voltage inspired the author to investigate the design based on using the current as the signal instead of conventionally using the voltage. Working at low-voltage (subthreshold region) suggested that the current provides much larger dynamic range than the voltage.

The architecture of the proposed analog front-end channel is shown in Figure 3.23. The recording sites, referred to as input (V_{ine}) and reference (V_{refe}), are connected to the circuitry through the electrode interface modeled by the resistor R_e and the capacitor C_e. The two inputs are biased using resistors (R_B) and connected to a fully-differential dc-coupled low-noise amplifier (DC-LNA), which is basically an OTA. The output currents (I_{O1} and I_{O2}) pass through all-pass filters (APF), which act as the delay element. The delayed currents (I_{Od1} and I_{Od2}) are subtracted from a replication of the OTA output currents to form the delta-compressed currents ($I_{\Delta 1}$ and $I_{\Delta 2}$). The compressed differential currents are converted to a single current ($I_{\Delta 2}$ - $I_{\Delta 1}$) and amplified further using a 4-bit current-mode programmable-gain amplifier (PGA). The output current ($I_{O\Delta}$) is then digitized using a current-mode 6-bit FLASH ADC, which provides an oversampled digital code captured on the system level as shown in Figure 3.22.

It can be seen that this technique relies on analog and mixed-signal processing techniques to achieve both direct-coupling to remove the input capacitors and save area, and DR-compression to allow the use of a 0.5V supply.

The next sections provide the description and design details of each individual component in the channel.

3.4 0.5 V DC-Coupled Low-Noise Amplifier

In this work, the amplifier is dc-coupled to save the area consumed by the large MIM capacitor. A technique similar to the one used in [29] is adopted as shown in Figure 3.24. Using the simple electrode model (***R_1*** and ***C_1***), the minimum values of ***R_1*** and ***C_1*** can be considered 10 TΩ and 100 pF, respectively. The overall transfer function from V_{ine} (electrode) to V_{in} (amplifier input) is a high-pass filter with a corner frequency (f_{3dB}) inversely proportional to C_1 and R_B, and a DC (f = 0) gain of $R_B/2R_1$. The purpose of the high-pass corner is nothing but to reject DC, but C_1 and R_B can change with the process and the worst case values should guarantee no more than 1 Hz cutoff frequency. So, assuming typical values of 100 pF and 100 GΩ for C_1 and R_B, respectively, the corner frequency is less than 32 mHz. Therefore, even with an order of magnitude deviation the corner is still below 1 Hz.

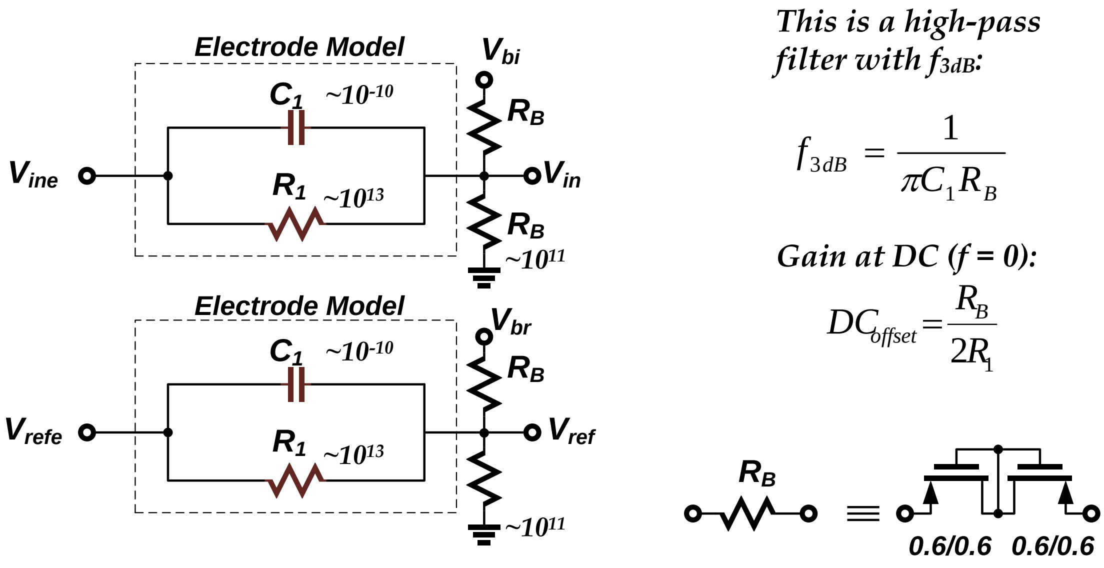


Figure 3.24. DC-coupling approach adapted from [29]

Figure 3.25 the designed low-noise amplifier, which is a fully-differential OTA based on current source load. The OTA includes source degeneration to increase linearity and tolerate more offset variation by splitting the bias current into 3 parts ($I_b/3$) with a gradual increase of the source degeneration resistor according to the bias current being unbalanced. So, if the offset between M_1 and M_2 caused more current to flow through M1 then one third (M_5) of this current sees no R_S, another third (M_7) sees one R_S, and the last third (M_6) sees $2R_S$. In other words, the effective source degeneration increases as a function of the unbalanced current. The degeneration resistor R_S is implemented using a PMOS transistor biased in triode and the effective resistance is ~40 kΩ. The voltages V_{bi} and V_{br} (Figure 3.25) are used to bias the OTA and would be very useful to cancel the effect of the random offset due to input electrodes mismatch. The overall current consumption (I_b) of the OTA is ~1.9 µA, which is a total power of 950 nW.

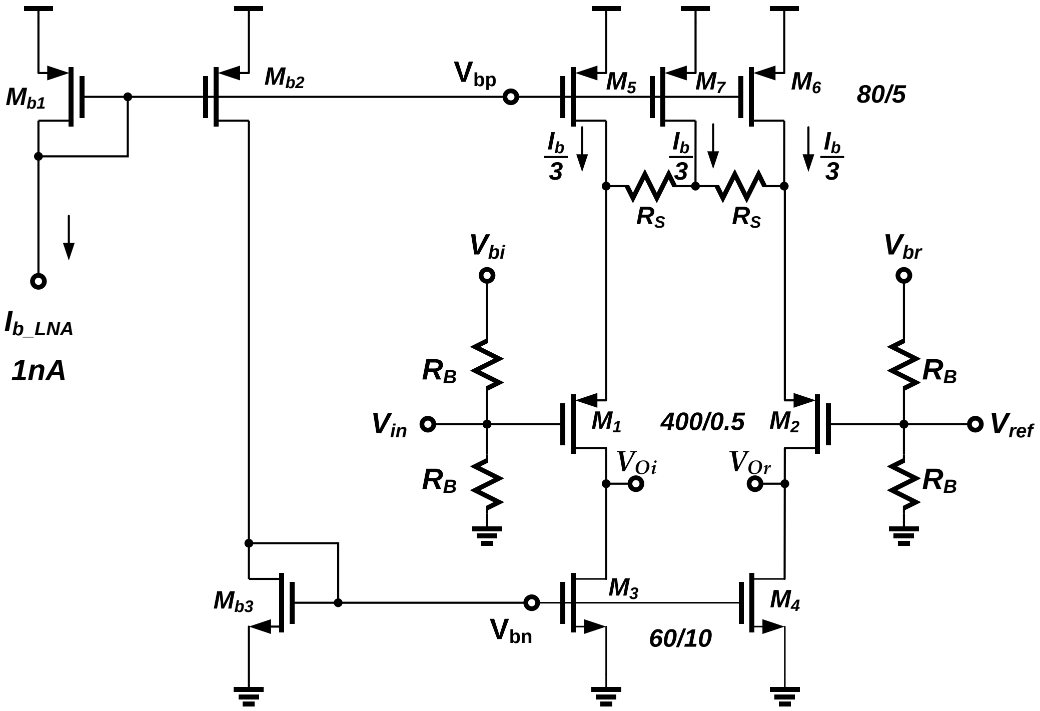


Figure 3.25. Schematic of the 0.5V low-noise amplifier using current source load and source degeneration to increase the linearity and providing more offset rejection

The simulated gain and phase responses of the OTA are shown in Figure 3.26 including also the response when modeling the electrode as discussed above (Figure 3.24). The high-pass and low-pass corners of the OTA (including the electrode model) are located at 200 mHz and ~60 kHz, respectively. While the OTA gain itself is around 37 dB (~70x), it is reduced to 29.6 dB (~30x) due to the electrode interface model forming the passive high-pass filter.

The input referred mean square thermal noise density can be calculated as:

$$\frac{\overline{v_n^2}}{\Delta f} \approx \frac{8.k.T.\gamma}{G_{m1}}.(1+\frac{g_{m3}}{G_{m1}}),\ \ \gamma=\frac{2}{3}$$ **Equ. 3.17**

The diff-pair effective transconductance (G_{m1}) is reduced due to source degeneration by a factor of ~2 compared to the actual g_{m1}. Since all MOSTs are biased in different regions inside weak inversion (WI), the transconductance ($\boldsymbol{g_m}$) can be determined by the inversion coefficient (IC) using EKV model [95] as described in Chapter 2 (Equ. 2.13 and Equ. 2.14). Applying this model yields $\boldsymbol{m_3}$ of 0.8 ($g_{m3} \approx 21$ µS), G_{m1} of 13 µS, and an input-referred thermal noise floor of ~67 nV/√Hz (6.7 µV$_{rms}$ integrated across 10 kHz bandwidth)

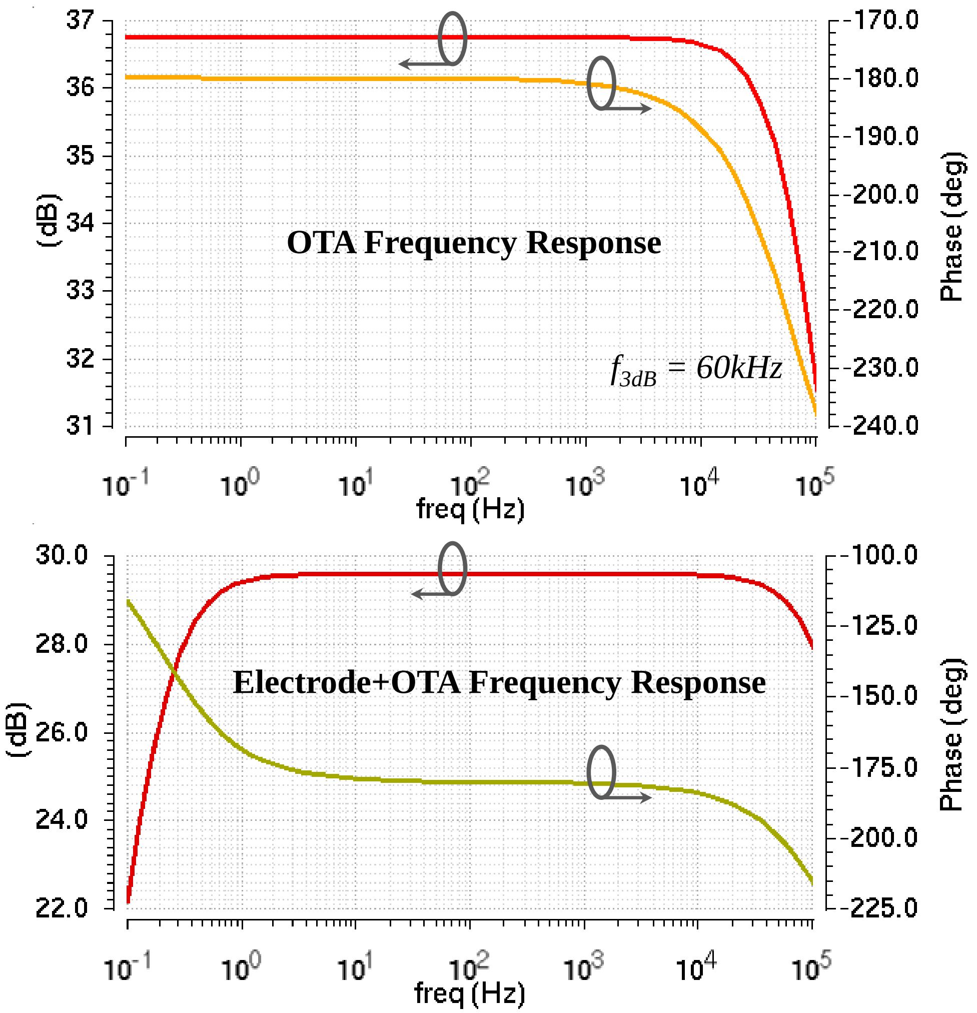


Figure 3.26. Simulated gain and phase responses of the low-noise amplifier including and excluding the high-pass filter (from the electrode model)

3.5 Current-Mode All-Pass Filter as an Analog Delay Line

A continuous-time implementation of an analog delay line is very critical for low-voltage low-power applications that cannot tolerate the interference and aliasing caused by the clocking required for discrete-time implementations. In signal processing circuits such as adaptive filters, the analog delay line is a very useful component [107]. There are two main properties required when considering a delay line [107]: 1) The transfer function should have unity amplitude over wide-enough bandwidth; and 2) There must be a linear relationship between phase-shift and frequency to achieve a frequency-independent group delay.

The architecture of the delay line used in this implementation was adopted from [107]. However, the design itself relies on the Subthreshold region unlike the technique used in [107]. The analysis of the all-pass filter (APF) use as an analog delay can be explained as follows [107]:

1) A first-order APF would satisfy unity amplitude with the following transfer function:

$$H(j\omega) = \frac{1 - j\omega RC}{1 + j\omega RC} \qquad \textbf{Equ. 3.18}$$

The phase and group delay are given by:

$$\angle H(j\omega) = -2\tan^{-1}\omega RC \qquad \textbf{Equ. 3.19}$$

$$D(\omega) = \frac{2RC}{1 + (\omega RC)^2} \qquad \textbf{Equ. 3.20}$$

2) Equ. 3.20 suggests that the group delay can be considered frequency-independent if $(\omega RC)^2 << 1$. So, to achieve a delay (2RC) of 10 µs for neural recording applications, the APF must operate within a frequency lower than 32 kHz. Since the highest frequency of interest is 10 kHz, the delay would have a maximum error of ~3.8%

3) According to [107], Equ. 3.18 can be written as:

$$H(j\omega) = 2\frac{1}{1 + j\omega RC} - 1 \qquad \textbf{Equ. 3.21}$$

This indicates that one way of implementation is to have a low-pass filter with a gain of two added to an inverter.

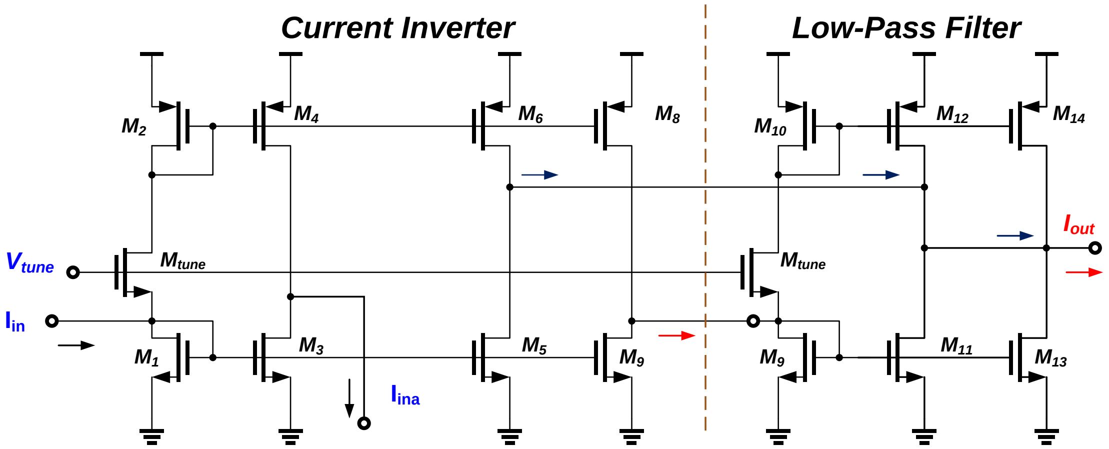


Figure 3.27. Schematic of a current-mode all-pass filter: it consists of a current inverter followed by a low-pass filter (adopted from [107])

4) The actual implementation to support Equ. 3.21 is shown in Figure 3.27 If a current I_{in} is input at the drain of M_1 the two currents through M_1 and M_2 will have different values. Since M_1 and M_2 mirror their currents to M_3 and M_4, respectively, I_{in} will affect the current flowing in the junction between M3-M4 (similarly M5-M6, and M7-M8) and a current of the same value of I_{in} will flow (in an opposite direction) into or out of the junction point. This technique fulfills the inversion part of Equ. 3.21.

5) Assuming all transistors in Figure 3.27 operated in strong inversion and saturation, it can be proven (as in [107]) that the circuit has a linear input resistance of:

$$R = \frac{1}{2\beta(V_{tune} - 2V_{th})} \qquad \textbf{Equ. 3.22}$$

However, in subthreshold region and under saturation condition it can be proven that the resistor is nonlinear and given by:

$$R = \frac{1}{4\beta U_T} e^{\frac{-(V_{tune}-V_{th})}{2nU_T}} \qquad \textbf{Equ. 3.23}$$

This provides a higher tuning dynamic range (via V_{tune}), which may be an advantage when considering process variations in the subthreshold region.

6) There is an effective capacitance (C) at the current input (I_{in}) node formed mainly by the sum of gate-source capacitance of M_{tune}, M_1, M_3, M_5, and M_7.

Considering this capacitance, the overall transfer function of the current inverter (I_{ina}/I_{in}) is:

$$H(j\omega) = \frac{-1}{1 + j\omega RC} \qquad \textbf{Equ. 3.24}$$

In the “Low-Pass Filter” section of Figure 3.27 it can be seen that two branches of current are summed at the output, which provide the factor of two in Equ. 3.21 and a branch from the current inverter is also summed to the same node. This satisfies the whole transfer function (Equ. 3.18) starting from the gate input of M_9 (shown by the red arrow) to the output branch at the connection of M_{13}-M_{14} (denoted as I_{out}). To clarify this further, it can be written as follows:

$$I_{out}(j\omega) = -I_{ina}(j\omega)\left(\frac{-2}{1 + j\omega RC} + 1\right) = -I_{indelayed}(j\omega) \qquad \textbf{Equ. 3.25}$$

The only difference, however, is the negative sign that is actually an advantage as described next.

For forming the delta operation, a copy of the input current is required to perform the subtraction operation. This is provided by the replica (I_{ina}) formed with M_3-M_4. If I_{ina} and I_{out} nodes are shorted, the summed output current can be given by:

$$I_{out}(j\omega) + I_{ina}(j\omega) = -I_{indelayed}(j\omega) + I_{ina}(j\omega) = I_{\Delta} \qquad \textbf{Equ. 3.26}$$

Figure 3.28 shows the simulation of the delay value at different tuning voltages (V_{tune}), which confirms the high dynamic range of tuning provided by Equ. 3.23. Figure 3.29 shows the transfer function (gain and phase) of the all-pass filter represented by Equ. 3.25 with the phase starting at 180° due to the negative sign compared to Equ. 3.21. Figure 3.30 shows a subset of the transient simulations of the all-pass filter with the input and several outputs at different tuning voltages.

Finally, the simulation of the delta operation performed in Equ. 3.26 is shown in Figure 3.31. This is normalized with respect to the original current input. It can be seen that it forms a differentiator transfer function except for the low-frequency (below 100 Hz), which is flat. The flat region has a great advantage, which is discussed in the next section. The overall current consumed by the delay circuit (Figure 3.27) is ~40 nA and all transistors are of similar size (W/L of 20/6).

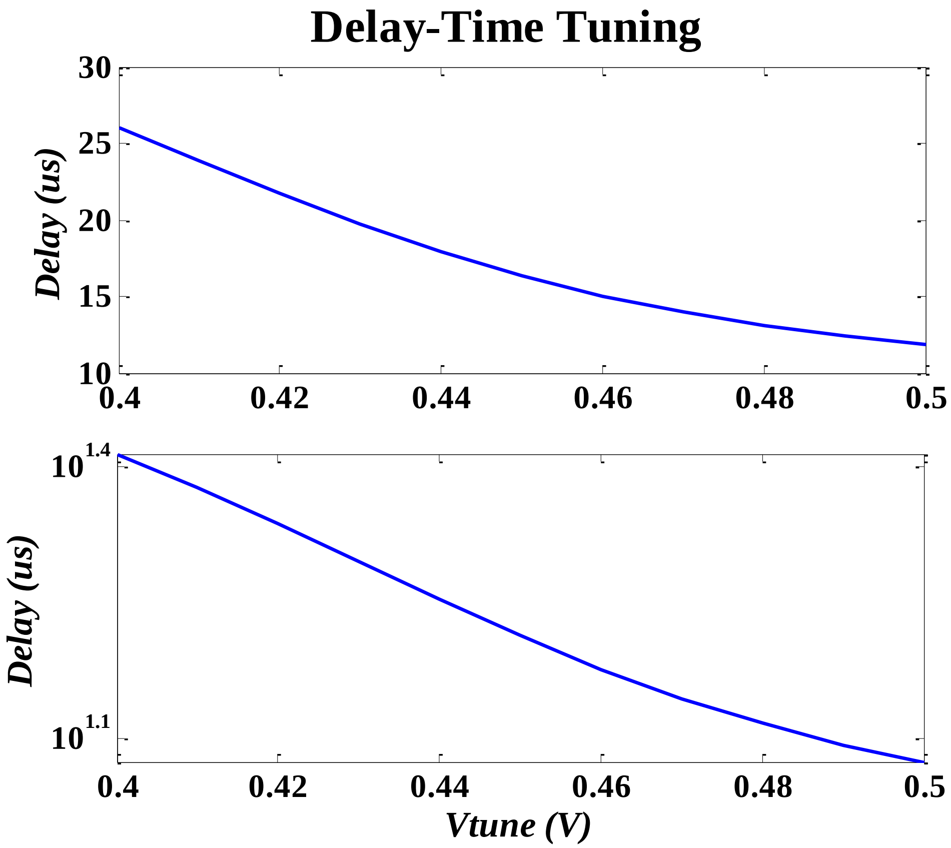


Figure 3.28. Simulated delay value versus tuning voltage showing the exponential relationship

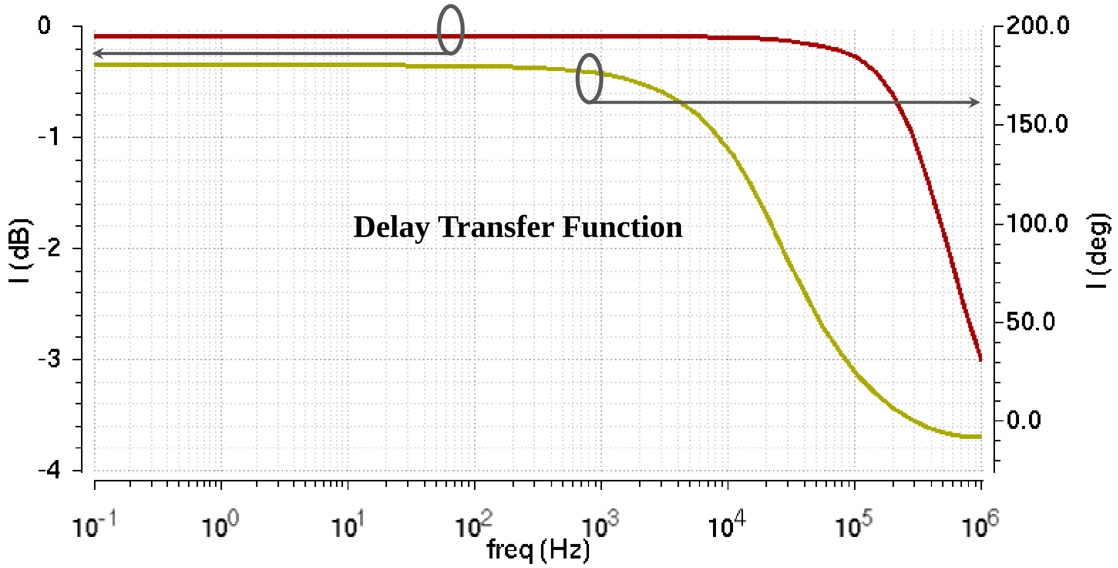


Figure 3.29. Simulation of all-pass filter transfer function

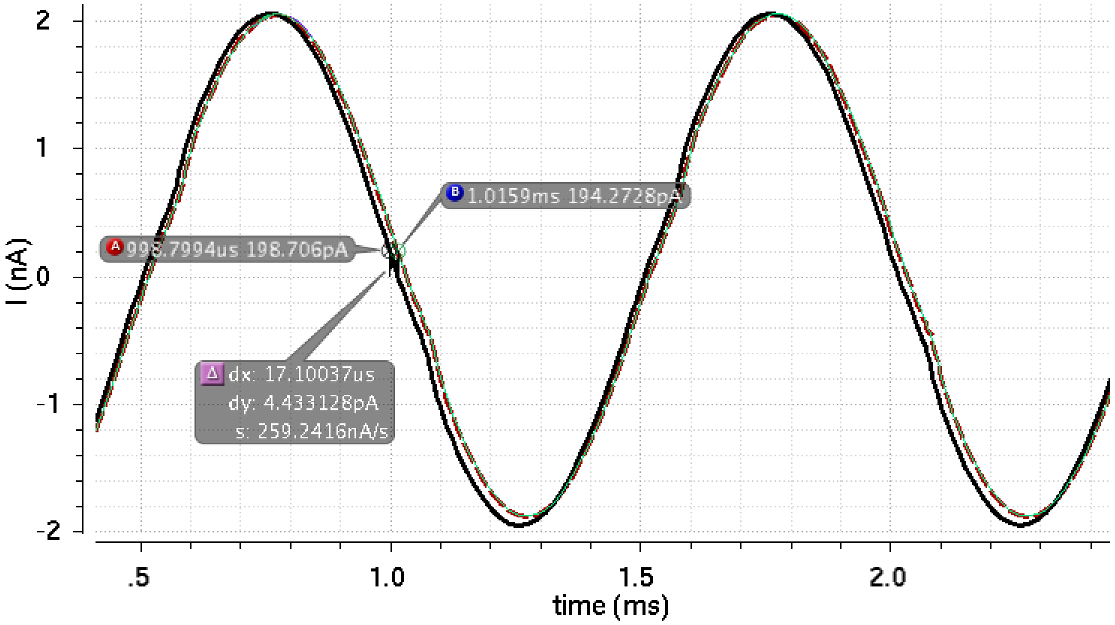


Figure 3.30. Transient response of the all-pass filter showing delays of up to 17 µs between the input (solid line) and the outputs (dashed lines)

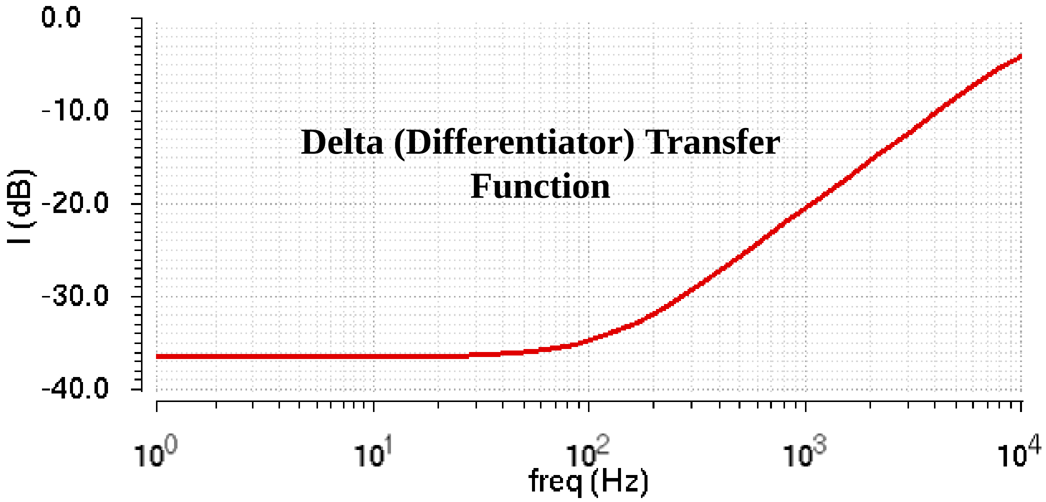


Figure 3.31. Simulation of the Delta Operation

3.6 Current-Mode Analog Compression Circuit

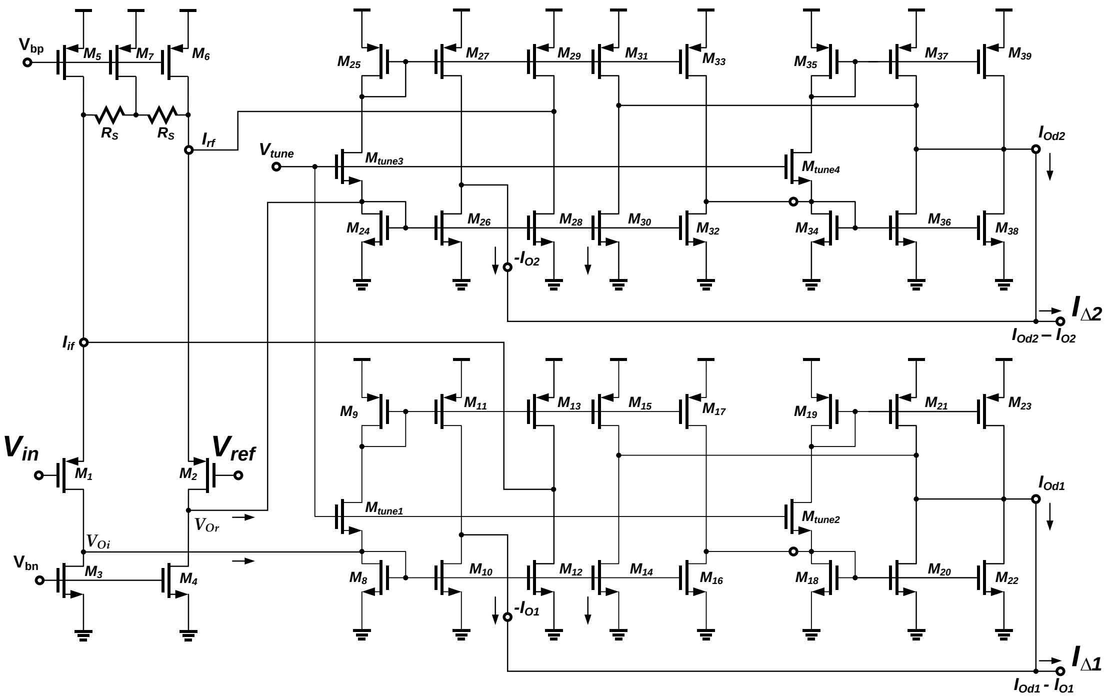


Figure 3.32. Schematic of the overall analog compression circuit.

Figure 3.32 shows the overall fully-differential analog compression stage. It is formed using the low-noise OTA (Figure 3.25) and two delay blocks (Figure 3.27) connected at the output nodes of the OTA (V_{Oi} and V_{Or}). The resulting output current is replicated (using the current inverter inside each delay block, -I_{O1} and –I_{O2}) and connected to the final output node of each delay block (I_{Od1} and I_{Od2}). The summed currents at both nodes become the delta currents, denoted in Figure 3.32 as $I_{\Delta 1}$ and $I_{\Delta 2}$, and are connected to a programmable-gain current amplifier as described next. Figure 3.33 shows the gain and phase responses of the analog compression circuit. It has a differentiator segment, a flat segment, and the high-pass corner from the OTA electrode filter.

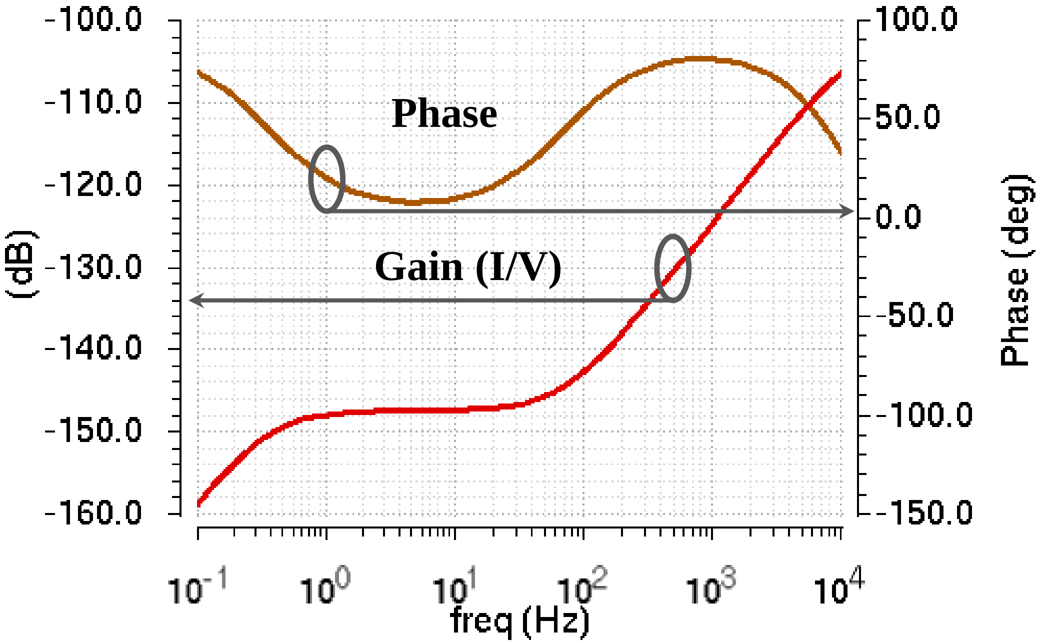


Figure 3.33. Simulation of the overall analog compression circuit.

3.7 4-bit Current-Mode Programmable-Gain Amplifier

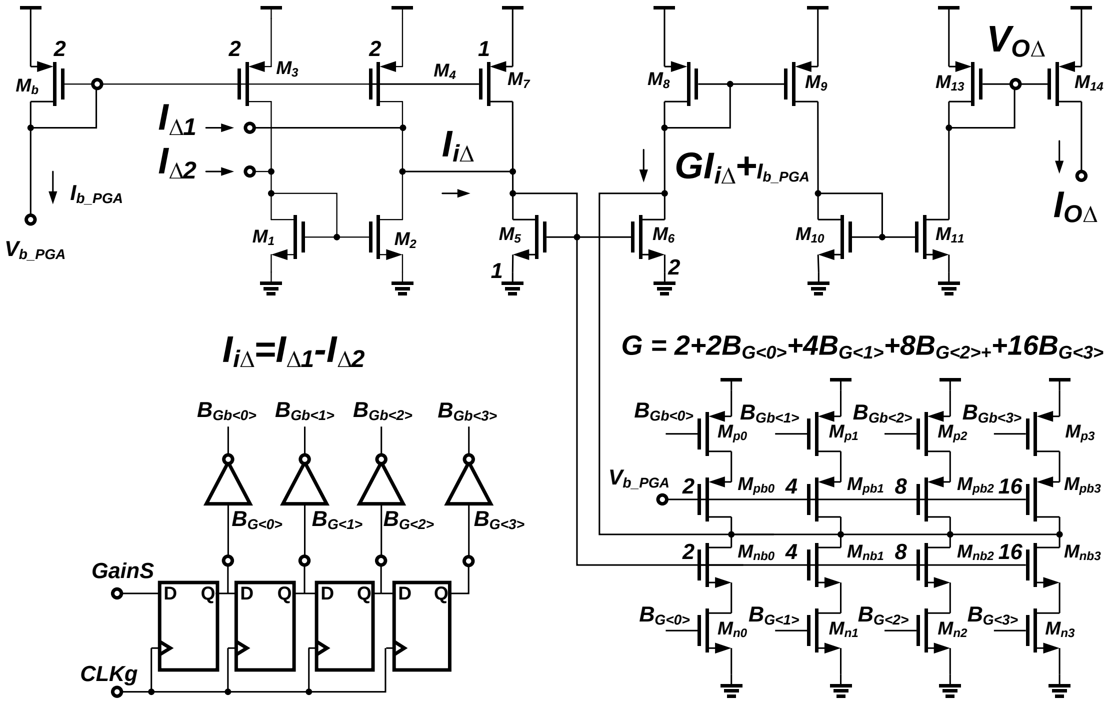


Figure 3.34. Schematic of the 4-bit programmable-gain amplifier.

The 4-bit programmable gain amplifier (Figure 3.34) starts by subtracting the two delta currents (reference $I_{\Delta 2}$ from signal $I_{\Delta 1}$) through M_1-M_4. The equivalent current $I_{i\Delta}$ is then mirrored from M_5 through the core PGA part ($M_{pb<0..3>}$ and $M_{nb<0..3>}$) and the mirroring ratio is digitally controlled through $M_{p<0..3>}$ and $Mn_{<0..3>}$. Figure 3.34 shows the sizing ratio for all the MOSTs contributing to the current amplification. The minimum gain value here is twice; this is through M_5:M_6 mirror and the minimum current value through M_6 is $I_{i\Delta}$+I_{b_PGA}.

The actual operation can be explained by the following example: assume BGb<1> is ‘1’ and other 3 bits are ‘0’, which will switch both M_{n1} and M_{p1} to the ON state (short circuit). The main biasing device (M_b) mirrors twice its current I_{b_PGA} to M_{pb1}, while M_5 mirrors 4 times its current ($I_{i\Delta}$ +0.5*I_{b_PGA}) to M_{nb1}. Now M_{pb1} holds a current of 2*I_{b_PGA} and M_{nb1} holds a current of 4*$I_{i\Delta}$ +2*I_{b_PGA}, the difference current (4*$I_{i\Delta}$) is supplied through M8, which holds now a current of 6*$I_{i\Delta}$ +I_{b_PGA}. The total current is finally mirrored through M_9-M_{13} to the ADC input (M_{14}).

The main bias current I_{b_PGA} has nominal value of 32 nA and can be used to control the bandwidth of the PGA and also adjust the offset that may be required to use the full range of the ADC. The PGA 4 bits are clocked in serially by a shift register using CLKg and GainS parameters. Figure 3.35 shows the simulation at different gain values to range from ~6 dB (2x) to 29.2 dB (28). Depending on the gain value, the PGA consumes a total current of 216 nA (108 nW) to 746 nA (373 nW)

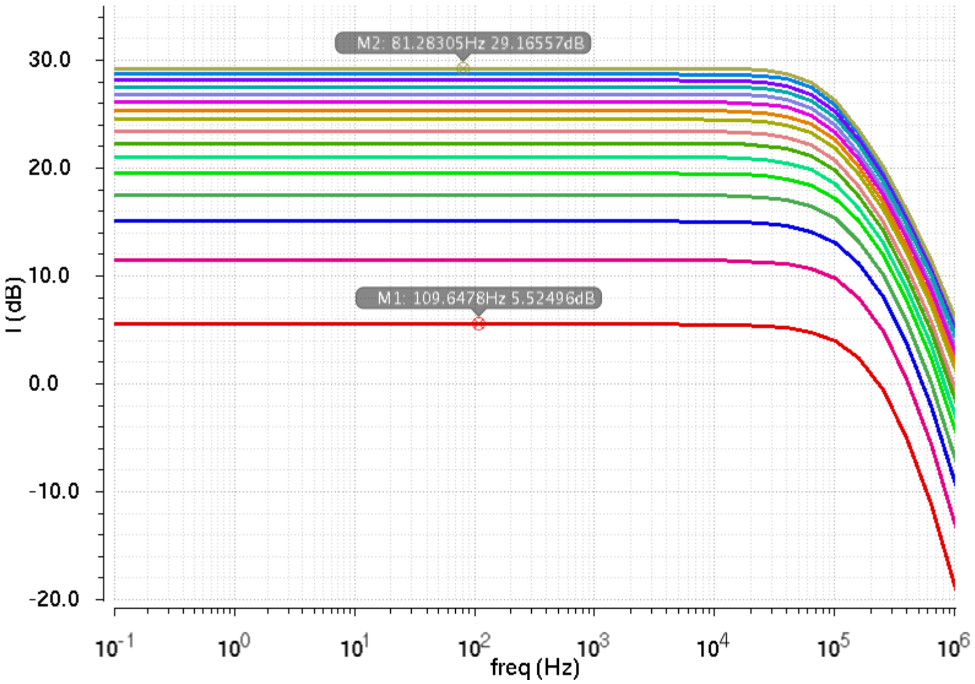


Figure 3.35. Simulation of the 4-bit programmable-gain amplifier.

3.8 6-bit Current-Mode FLASH Analog-to-Digital Converter

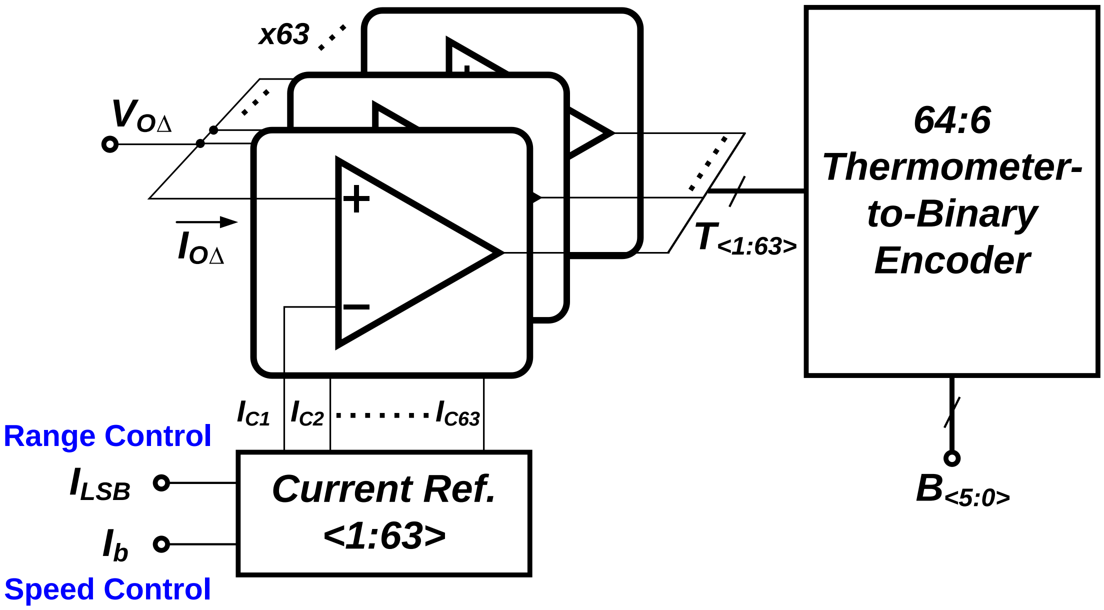


Figure 3.36. Architecture of the 6-bit current-mode FLASH ADC

As mentioned previously, one advantage of using current as a signal is the ability to achieve a high dynamic range at a reduced voltage swing and thus reducing the supply voltage. While this holds at all regions of MOST operation, it is especially true at subthreshold operation due to the exponential relationship between current and voltage. As far as ADC is considered, using the current techniques may avoid implementations based on capacitors; thus, reducing the area as well.

There are several techniques to design a current-mode or a current-input ADC with the advantage of saving area by excluding any sampling capacitors. One of these techniques is explained in [108], where the author designed an algorithmic current-mode ADC using a very interesting idea to scale the current signal as it undergoes through the cascade of current-mode comparators. However, involving switching operation requires the use of high current and biasing transistors in strong inversion such that settling time is short enough to allow the operation of ADC within the required sampling period.

Involving a switching operation at 0.5 V supply may cause a high current consumption which may not allow the use of 0.5 V supply after all. Therefore, a FLASH type ADC has been chosen to avoid any current switching and to keep the overall power consumption relatively low with a minimized interference to the other analog blocks.

Figure 3.36 shows the 6-bit FLASH ADC that is implemented using an array of 63 current-mode comparators that generate a 64-bit thermometer code. A 64:6 thermometer-to-binary encoder is used to generate the final output $B_{<5:0>}$. The ADC can have its speed and full-scale range controlled using the currents in the biasing circuit, I_b and I_{LSB}, respectively. The designed current-mode comparator and encoder are shown in Figure 3.37 and Figure 3.38, respectively, and explained next.

3.8.1 Current-Mode Comparator

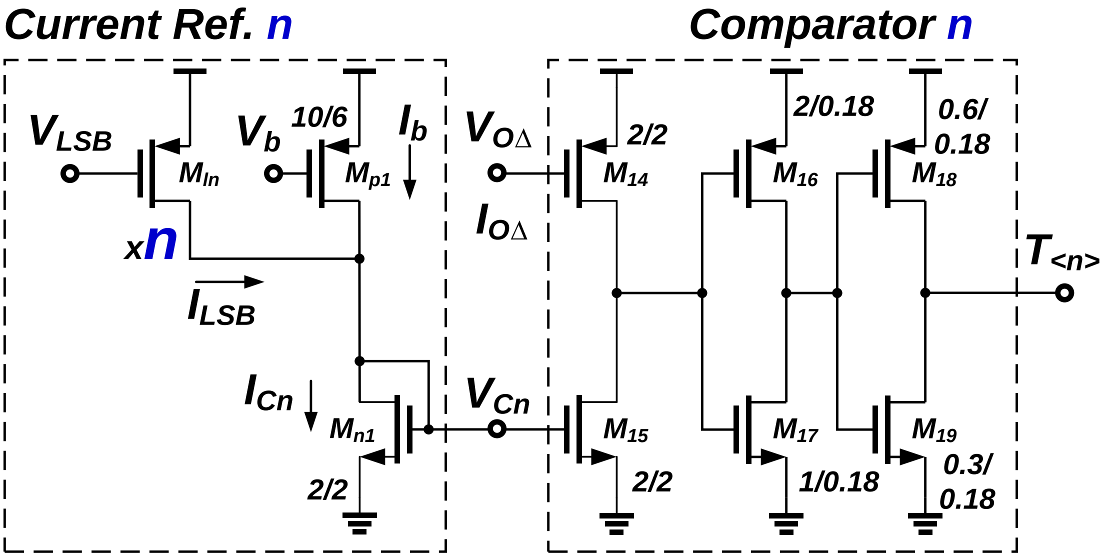


Figure 3.37. Schematic of the current-mode comparator.

A current-mode comparator (Figure 3.37), similar to the one proposed in [108], is used together with a scaled current-reference circuit to form the unit block of the 6-bit FLASH ADC as shown in Figure 3.36. The current-mode comparator in Figure 3.37 has two input (mirrored) currents compared with each other; $I_{O\Delta}$ is the input current mirrored by the PGA, and I_{Cn} is a scaled reference current that depends on the location of the comparator across the thermometer array. If $I_{O\Delta}$ is higher in value than I_{Cn}, then M14 will go out of saturation and its drain-source voltage reduces dramatically making the final output T<n> to equal a digital value of ‘1’. The opposite will occur if $I_{O\Delta}$ is lower in value than I_{Cn}. As explained in [108], the first inverter in the comparator acts as an integrating current-to-voltage converter, which can filter out the power supply noise and ensure no inherent DC offset in the comparator.

The comparator biasing circuit provides also two important flexibility features: 1) The FLASH/comparison speed is proportional to the current I_b (similar in value with I_{bPGA}), and 2) the current step size (LSB) can be controlled using the I_{LSB}, which also means the full scale range (I_{FS}) is controlled by this current.

3.8.2 Thermometer-to-Binary Encoder

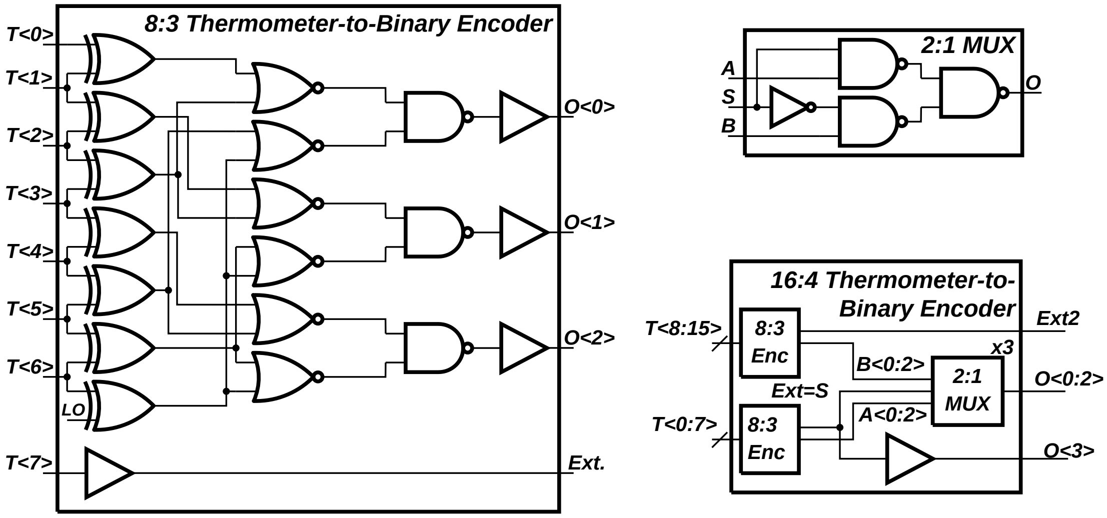


Figure 3.38. Schematic of a thermometer-to-Binary Encoder (8:3) used to construct the 64:6 encoder. The schematic shows the construction of 16:4 T-B encoder from two 8:3 encoders and three 2:1 digital multiplexers.

A basic 8:3 thermometer-to-binary encoder (TBE) is used to form bigger encoder using a hierarchical architecture shown in Figure 3.38. A group of digital 2:1 multiplexers are then used to decide which encoder to route to the binary output. This is repeated to form two 16:4 TBEs then 32:5 TBEs and finally the 64:6 TBE.

Figure 3.39 shows the simulation of the 6-bit FLASH ADC operation, which can be seen to operate at very high speed. The reference biasing block is shared among the 128-channels ADCs and consumes a current of ~3.1 µA (1.55 µW). The core of the ADC consumes current that is input-dependent, at the highest input current value it consumes ~3.1 µA (1.55 µW), and at mid-value (most of the samples) it consumes ~2.5 µA (1.25 µW).

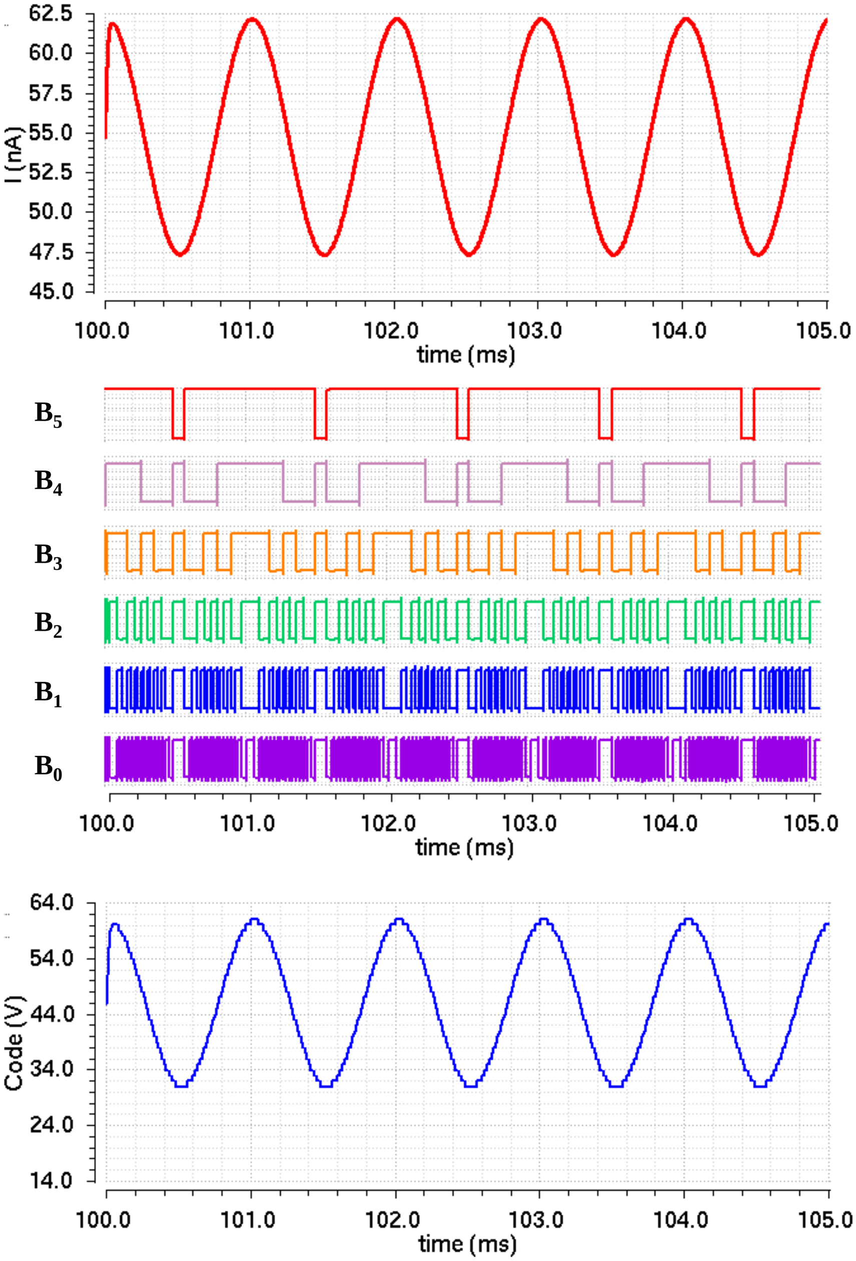


Figure 3.39. Simulation of the 6-bit analog-to-digital converter

3.9 Data Acquisition and Parallel-to-Serial Conversion

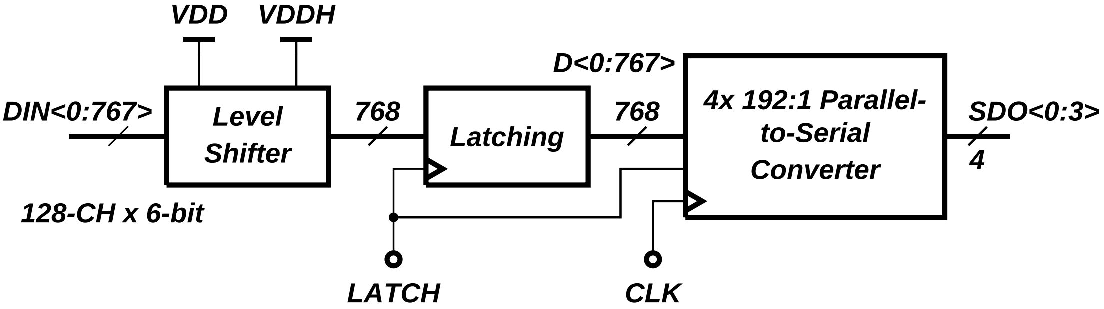


Figure 3.40. Block diagram of digital circuits used for acquisition.

A data acquisition circuit (Figure 3.40) is used to acquire the digital bits from 128 ADCs (at total of 767 bits per sampled instant). The acquisition circuit includes level shifters (Figure 3.41), latching block (Figure 3.42), and four 192:1 parallel-to-serial converters (Figure 3.43). This digital block was made to operate at 0.5 V with a nominal acquisition speed of 80 Mbps, which is equivalent to acquiring each channel at ~104 kHz. The reason to use a level shifter a the output of ADCs bits is to give more flexibility to acquire at higher speed, which may be done at higher supply voltage to save buffering area and power consuming that might be required if the design is strict at 0.5V.

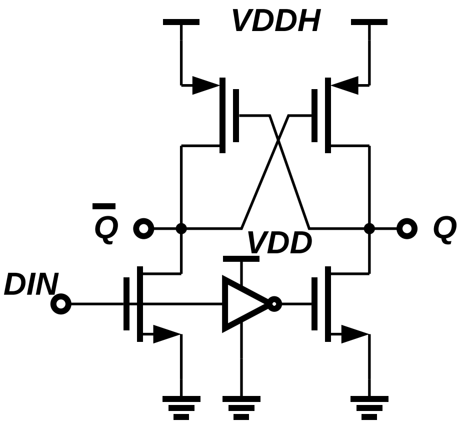


Figure 3.41. Schematic of level shifter

Figure 3.41 shows the schematic of the implemented level shifter using a cross-coupled structure and an inverter. Figure 3.42 shows one 192-bit latch that is repeated four times and based on true single phase clock (TSPC) dynamic flip-flop. The same 'Latch' signal can be shared among the four blocks. Figure 3.43 shows the 192-bit parallel-to-serial converter (PSC) which is also repeated four times, and is based on a shift register with 2:1 multiplexer. When the ‘latch’ is ‘low’ the ‘Q’ becomes D_C, which is the parallel data; this is called the ‘writing mode’. Before ‘latch’ goes ‘high’ one ‘CLK’ edge is required to store the parallel data ‘Dpar’ in the flip-flops. When the ‘latch’ signal becomes ‘high’, the final output ‘Q’ holds the data stored in the flip-flop and is ready of the shifting operation using the ‘CLK’. Figure 3.44 shows the overall timing diagram of the data acquisition block.

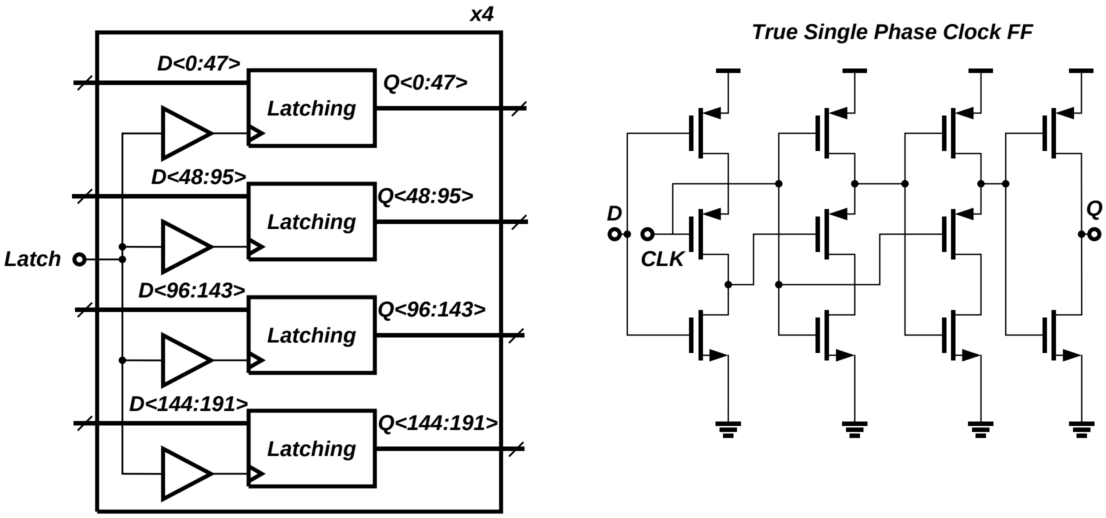


Figure 3.42. Schematic of 192-bit latch that is repeated four times. It also shows the unit flip-flop used inside the latch, which is basically a true signal phase clock dynamic flip-flop.

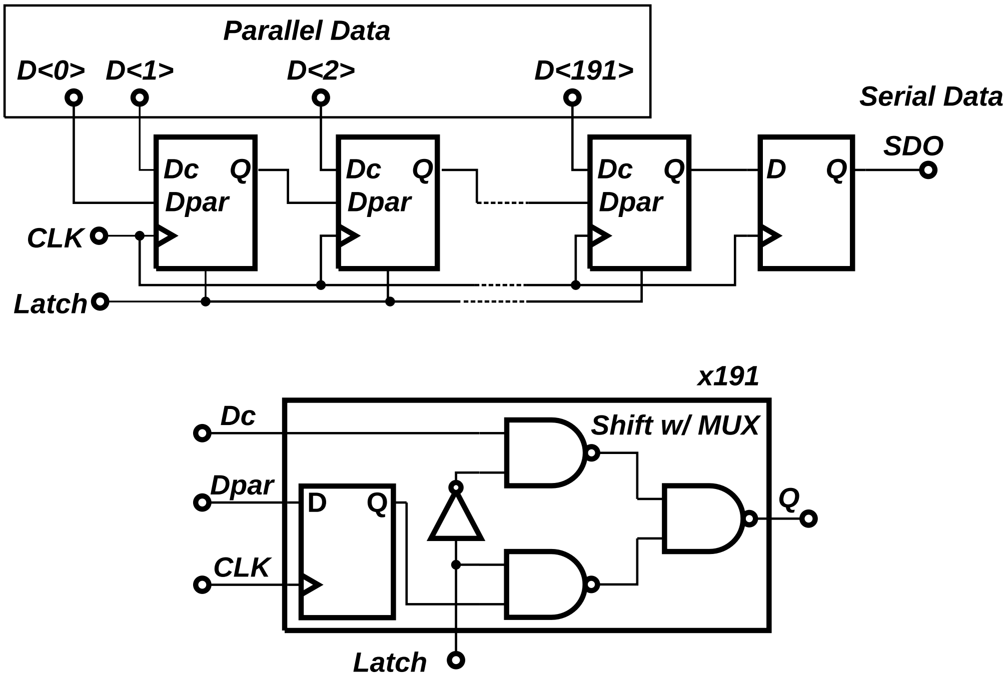


Figure 3.43. Schematic of the parallel-to-serial converter

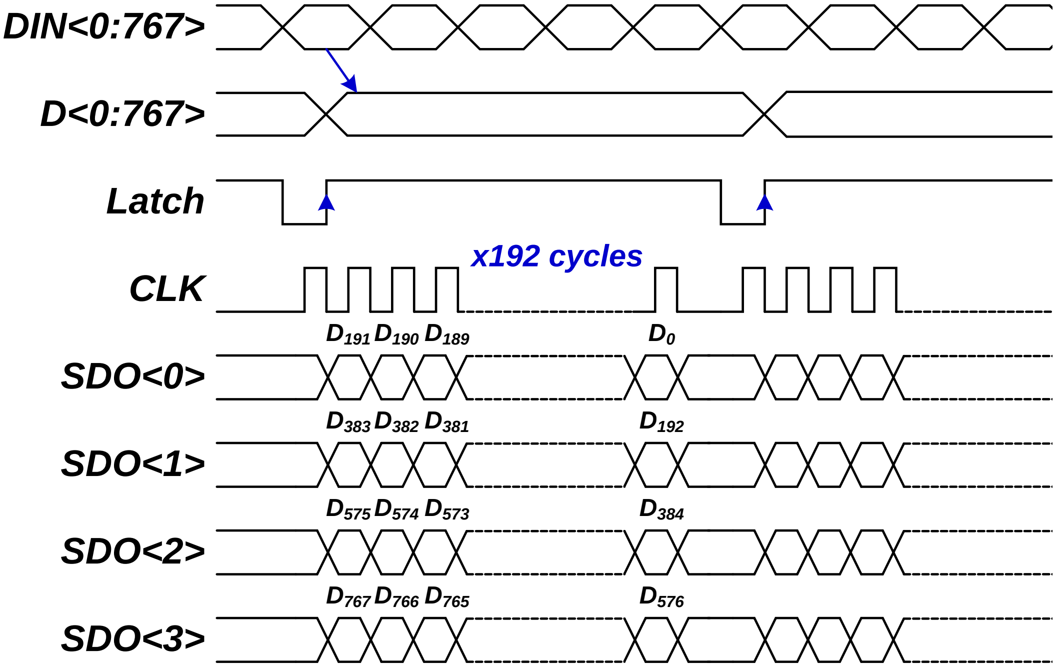


Figure 3.44. Timing diagram of data acquisition

3.10 Measurement Results and Performance Comparison for AHI-128

3.10.1 Overall System

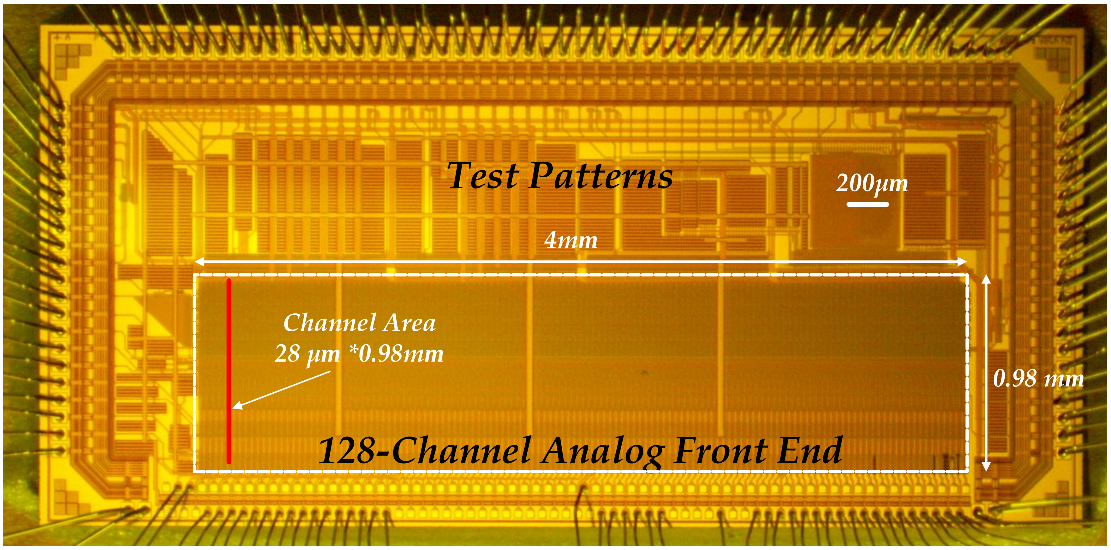


Figure 3.45. Die photo of the fabricated chip.

The chip was fabricated using 0.18µm CMOS process and operated nominally at 0.5V. Figure 3.45 shows the fabricated chip microphotograph. It has 128 channels that consume a core area of 3.92 mm^2 with a channel-to-channel pitch of 31.5 µm pitch. Each channel consumes less than 2.5 µW of power and consumes an area of 0.027mm^2. This is 62.5% the power and 38.5% the area of the previous design, respectively.

All the 132 input pads (128 signal pads and 4 reference pads) are at one side of the chip staggered into two rows (40x40µm^2, 63µm pitch). Another 16 pads (70x70µm^2, 90µm pitch) are used for power (4 pads), data acquisition (5 pads), and tuning/biasing (7 pads).

Figure 3.46 shows a closer look into the channels and identifies different components to show their relative implementation areas. The red thick line shows one channel area. Figure 3.47 shows the PCB used for the test setup with the FPGA to emulate the host (processor). A LabVIEW interface program was developed to provide programming commands and receive the output data. Matlab is also used for the reconstruction of recieved data.

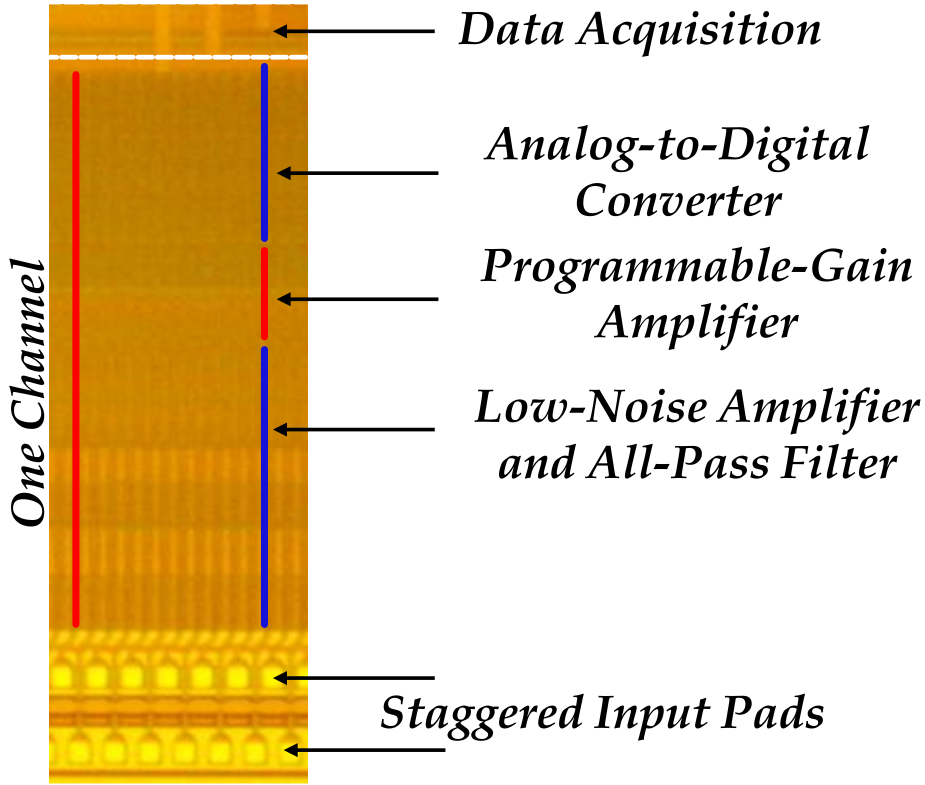

Data Acquisition
Analog-to-Digital Converter
Programmable-Gain Amplifier
Low-Noise Amplifier and All-Pass Filter
Staggered Input Pads
One Channel

Figure 3.46. A zoom-in view of die photo showing different components.

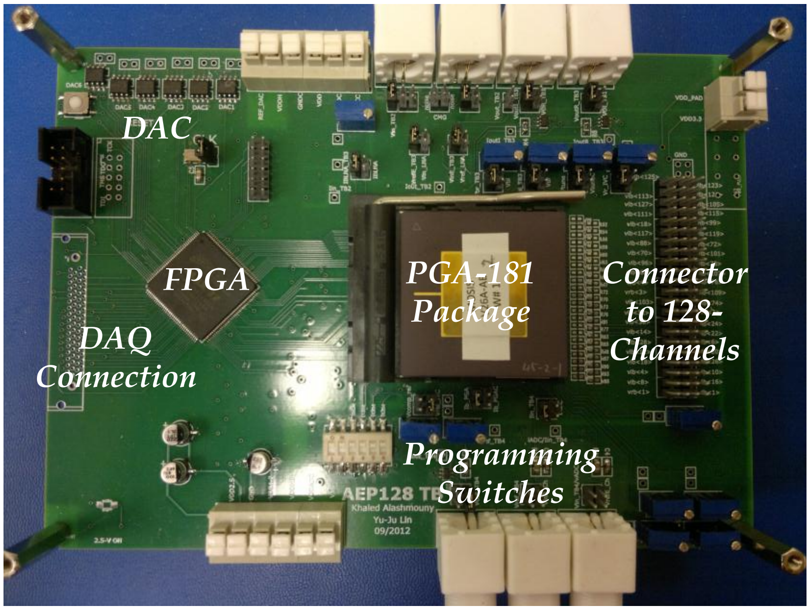


Figure 3.47. Measurement Setup

3.10.2 Low-Noise Amplifier

Figure 3.48 shows the gain response of the low-noise amplifier indicating a mid-band gain of 34 dB and low-frequency and high-frequency corners of 200 mH and 9.2 kHz, respectively. Due to the use of off-chip fully differential buffers in the measurement setup (Figure 3.47), the actual high-frequency corner is expected to be higher than this value. Noise performance is shown in Figure 3.49 and indicates an input-referred noise of 4.9 µV$_{rms}$ integrated across 100 kHz bandwidth.

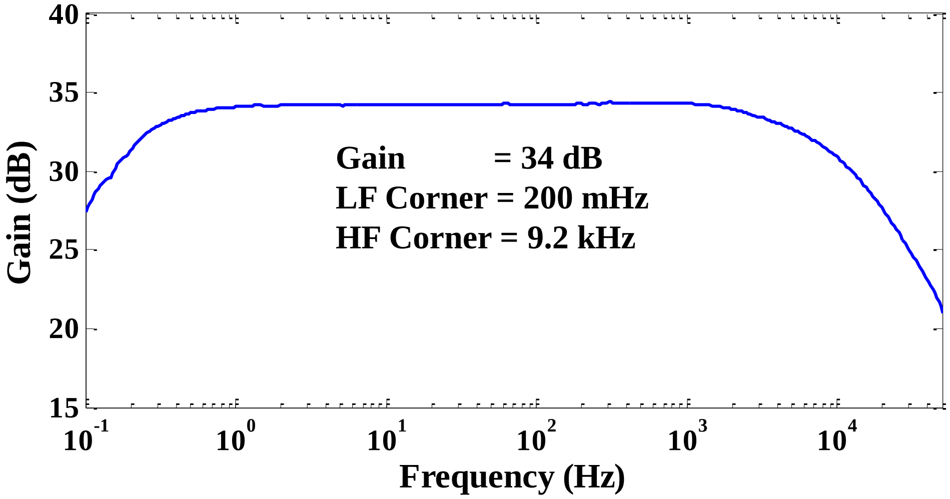


Figure 3.48. Measured LNA gain response

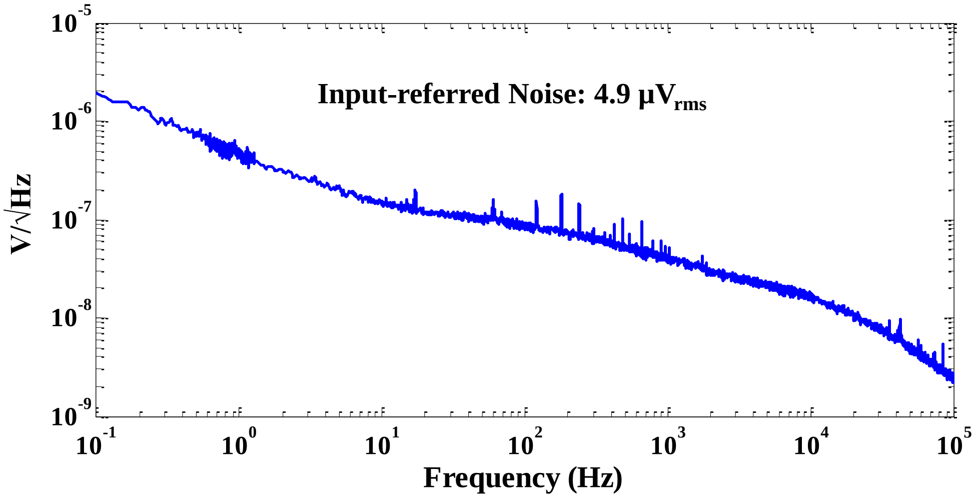


Figure 3.49. Measured LNA noise response

3.10.3 Analog Compression Block

The measured frequency response of the analog compression block (LNA plus the all-pass filters) is shown in Figure 3.50. It shows the gain response from voltage input in mV to current output in nA represented by dBnm which is a made up unit to represent nA/mV transconductance gain. Phase response is also shown in the same figure.

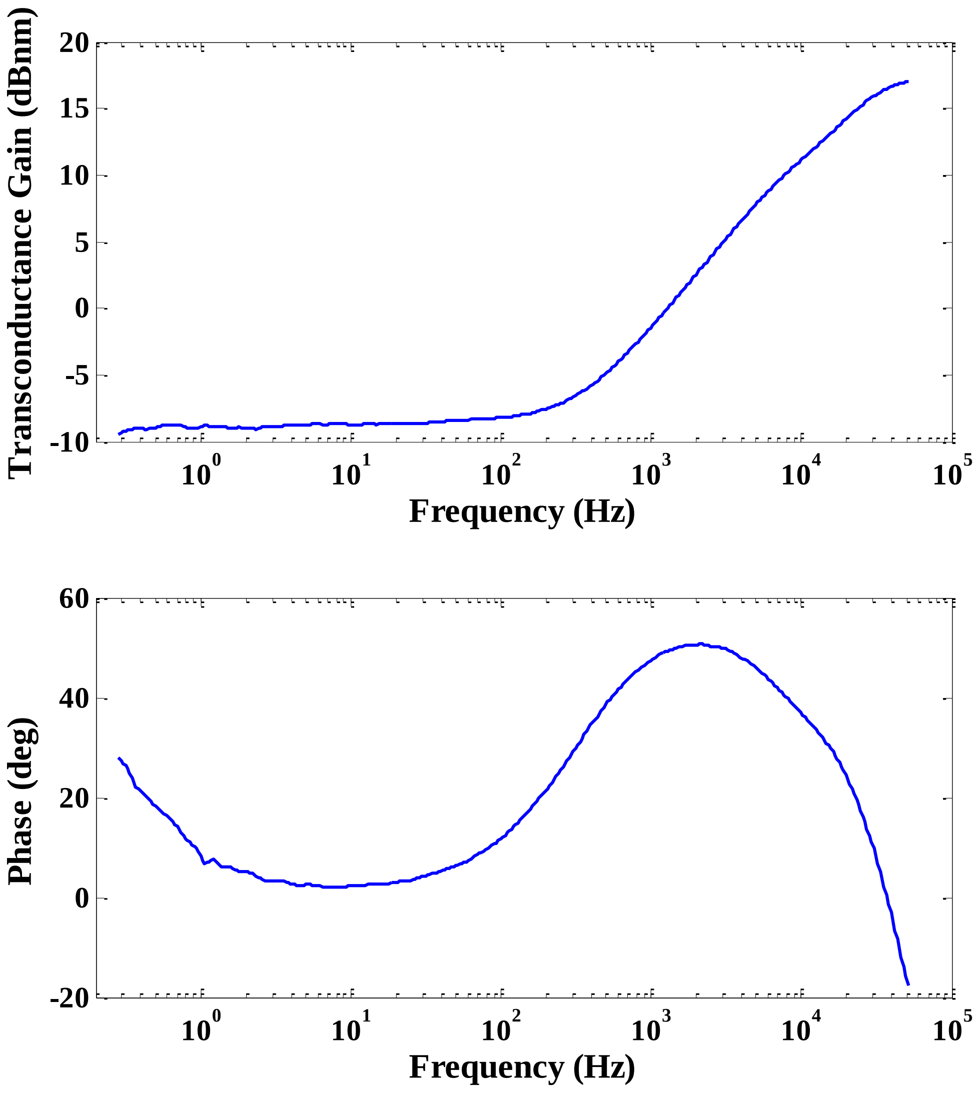


Figure 3.50. Measured analog compression gain and phase response

3.10.4 Programmable-Gain Current Amplifier

Figure 3.51 shows the measured current gain from the current-mode programmable-gain amplifier. The gain can be set in 4-bit control and range from 4.5 dB (1.7x) to 29.1 dB (28.5x).

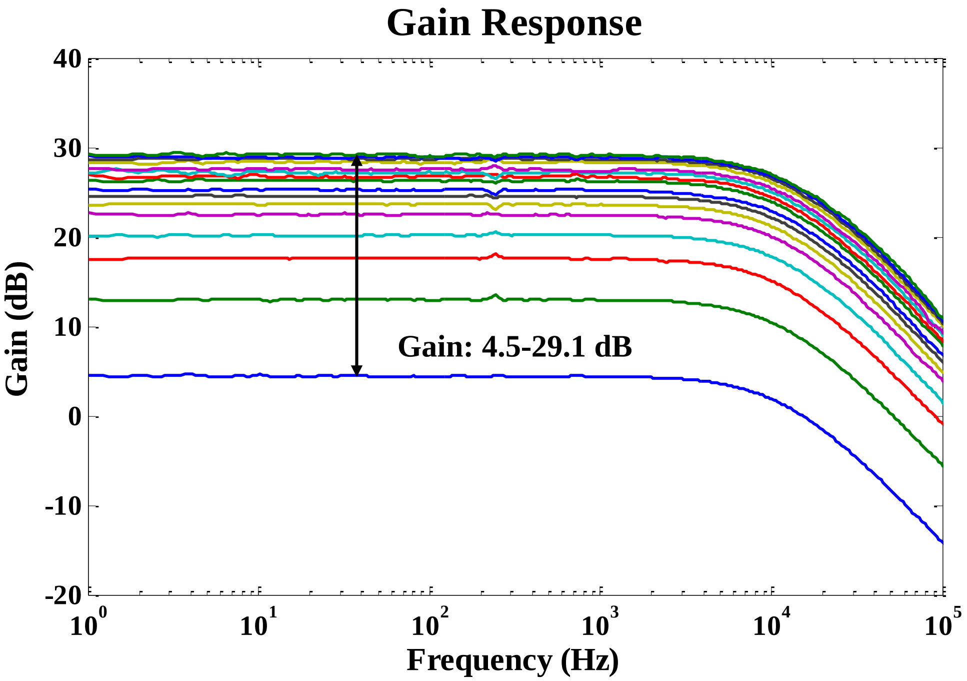


Figure 3.51. Measured current-gain of the programmable-gain amplifier

3.10.5 Analog-to-Digital Converter

Figure 3.52 shows the measured spectrum for the FLASH ADC. The achieved effective number of bits is 3.3. It also shows the transient with and without the use of oversampling. Oversampling can add up to 4.5 bits to the overall resolution.

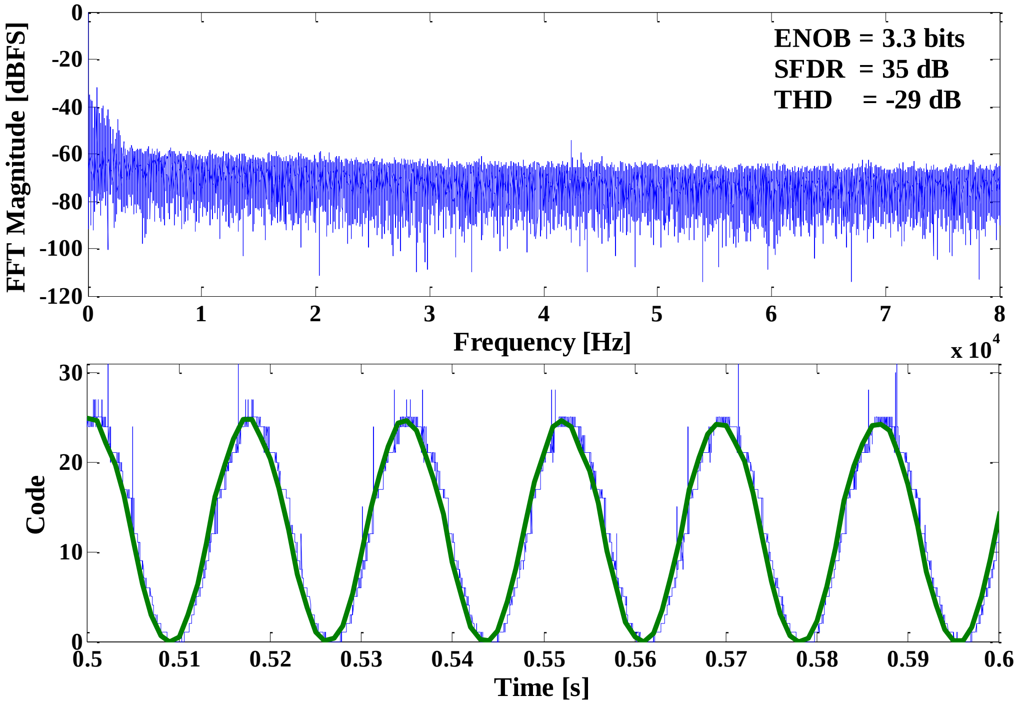


Figure 3.52. Measured ADC spectrum and transient output

3.10.6 Data Measurement and Reconstruction

Figure 3.53 shows the concept of the actual data construction based on the response measured in Figure 3.50 and compares it with the ideal reconstruction that would be used of the differentiator (delta operation) was ideal. The only difference is that after the digital summation (integration) at the receiver a single-pole elliptic high-pass filter is used to get the actual response of the low-frequency components of the signal. Without applying a high-pass filter the low frequency components dominate the signal since they are largely amplified after the summation process.

Figure 3.54 shows the result of reconstruction based on the response measured in Figure 3.50. The delay value (based on Figure 3.50) is ~15μs and the data were sampled at 66.7 kHz and Figure 3.54 shows the reconstructed signal based on the modified algorithm explained in Figure 3.53. Figure 3.55 shows the input-referred version of the reconstructed signal superimposed on the original signal. The input-referred lossy compression error is about 150 µVrms without the effect of oversampling. Oversampling can reduce this error down to about 9 µVrms.

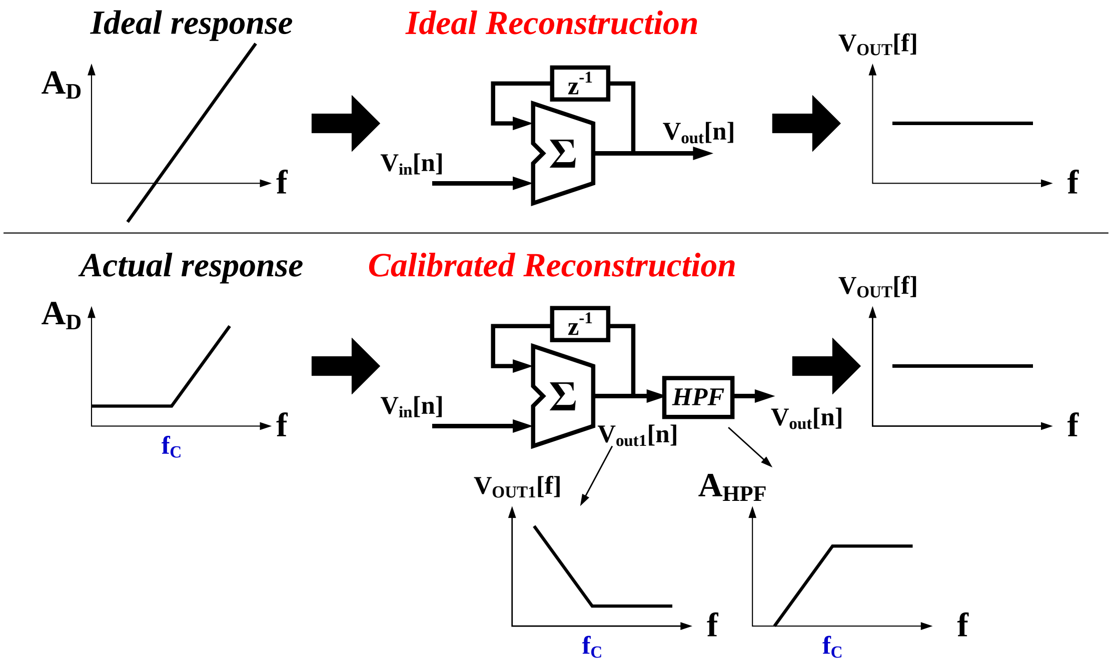


Figure 3.53. The concept of data reconstruction

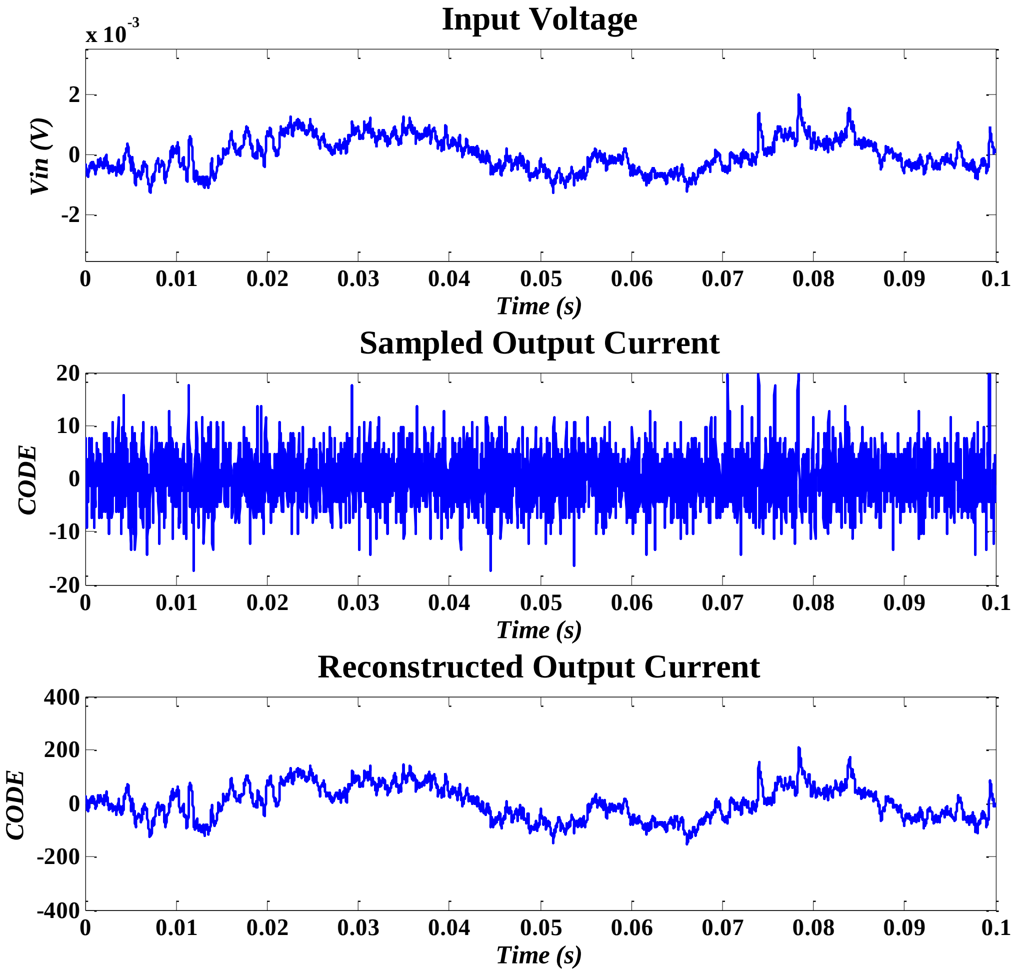


Figure 3.54. Data reconstruction shows the input voltage, sampled output current, and the reconstructed output current.

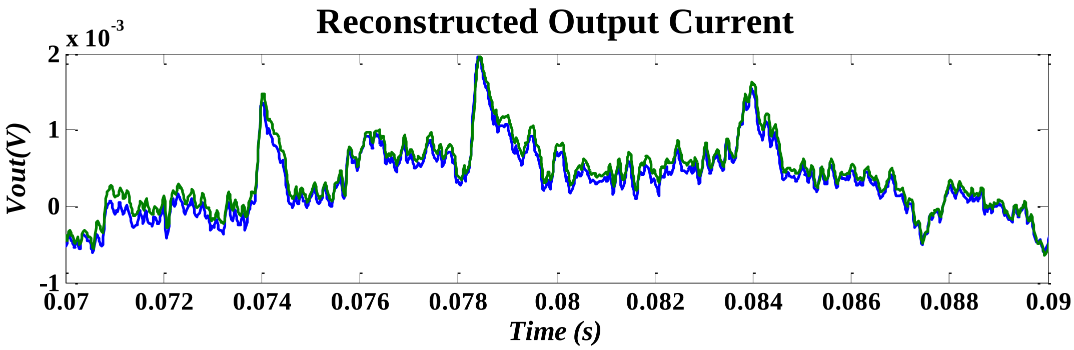


Figure 3.55. Reconstructed output (input-referred) superimposed on the original input.

3.10.7 Performance Comparison

Reference	AHI-16	AHI-128
Technology	0.25μm	0.18μm
Low Noise Amplifier		
Area [mm^2]	0.05	0.01
Voltage [V]	0.9	0.5
Power [μW]	0.99	0.95
DR[dB] at THD<1%	47.6	46.7
Noise$_{in\_ref}$ [μV$_{rms}$]	4.8 (10kHz)	4.9 (100kHz)
NEF	2.9	2.6
NEF^2·(V$_{DD}$)	7.56	3.38
K (x10^8)	5.98	4.91
CMRR [dB]	>62	>70
PSRR [dB]	>59	>70
BPF and PGA		
Power [μW]	2.31	0.1-0.37
LF Cutoff [Hz]	<0.1-1000 (3-bit)	0.2
HF Cutoff [kHz]	1-17 (3-bit)	>10
Gain [dB]	52.4 to 79.8 (3-bit)	4.5 to 29.1 (4-bit)
ADC		
Power/Ch [μW]	0.66 at 20.16kHz	1.25-1.5
Sampling/Ch [kHz]	6.25-50	Up to 10 MHz
ENOB	7	4.8 up to 7.8
Overall Evaluation of Analog Front-End		
Power/Ch [μW]	3.96	2.5
Area/Ch [mm^2]	0.07	0.027
EAE (x10^{-15})	7.51	30.23
EE (x10^{-9}) (without Area consideration)	0.526	0.816
Specific Wires/Ch	2:16	2:32

Table 3.3. Measured performance and comparison with previous version

The performance comparison is summarized in Table 3.3. The proposed AFE module in this work achieved an EAE (energy-area-efficiency) of 30.23x10^{15}. This is clearly higher efficiency than the previously reported design (AHI-16). It provides the wide range of operation through programmability, small area per channel, small lead counts per channel for data transfer, and the lowest power consumption (2.5 μW per channel) as compared with the previous work. Figure 3.56 shows the location of this work in the plot of noise and area efficiency, which confirms the potential of the design approach for massive parallel recording while being closer to the safety margin in terms of tissue heating.

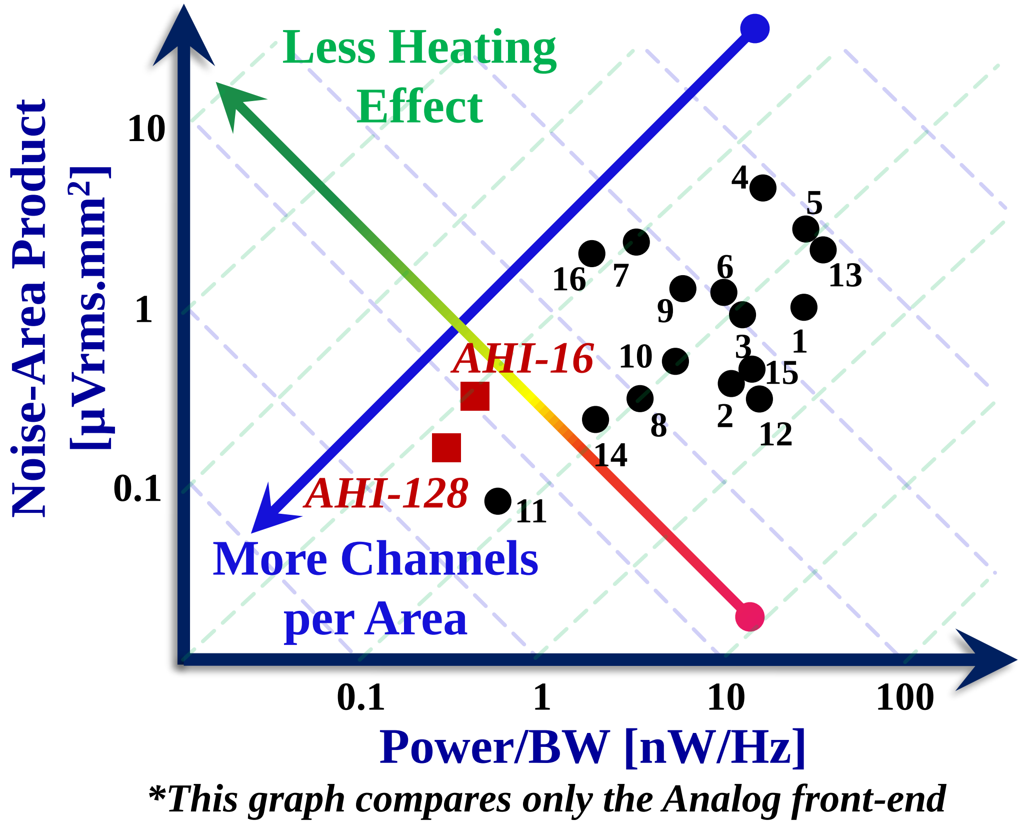


Figure 3.56. Performance chart of the implemented AFE.

3.11 Conclusion

The scaling towards simultaneous recording of neurons activities requires scaling of both power consumption and area consumption. Current approaches for neural recording do not scale because of using capacitors and diodes components for rejecting the electrode DC offset. In this chapter, a low-voltage low-noise analog front-end module has been proposed to achieve high energy-area efficiency by referring some calculations and signal recovery to the receiver side and including on-chip compression and current-mode circuit techniques to improve the dynamic range while using low supply voltage (0.5 V). The implemented front-end module has 128 channels and achieved low power consumption of 2.5 μW/channel in 0.027 mm^2 using 0.18μm CMOS processes.

The proposed 128-channels front-end with embedded processing (AHI-128) has shown very promising techniques for further scaling towards 1024 channels, while reducing the power and area at the same time. Compared to AFE-16, the AFE has reduced the noise-energy-area product by a factor of 4, and has demonstrated feasibility to accommodate larger number of channels.

CHAPTER 4

MONOLITHIC ACTIVE NEURAL PROBES: TOWARDS THE NEXT DECADE OF MASSIVE PARALLEL RECORDING

Massive parallel recording of up to hundreds of channels was proposed using techniques including signal processing and biasing points as shown in previous chapters. In the near future, however, recording from more and more neurons would be required to improve further our understanding of brain dynamics. For more scaling (beyond 1000 channels), implementation area need to shrink further; therefore, it is essential to remove any unnecessary components that contribute to the area. It was shown in chapter 2 that having a 128 pads on the neural probe to the corresponding CMOS IC pads consume a huge amount of area. It can be predicted that the next leap in neural recording should be achieved through advanced technologies and monolithic integration of MEMS and CMOS together.

While monolithic integration has some disadvantages such as cost and yield issues, it might be the only means for future expansion of the recording channels to the fourth order of magnitude. In this chapter, an active neural probe design is described and shown to extremely shrink the implementation area using monolithic integration, for the first time, with the analog front-end on the probe's shanks.

The chapter starts by explaining the potential of monolithic integration of MEMS and CMOS. The probe module architecture is further described in the second section, followed by the circuit design and integration onto the shanks. The design and measurements of individual circuit blocks in 150nm SOI process is further discussed in details. The conclusion shows that monolithic integration using advanced circuit technologies may be the more effective approach to pursue the recording of thousands of neural activities simultaneously.

4.1 Prospective of Monolithic Integration of MEMS and CMOS for Neural Interfaces

4.1.1 Why Monolithic Integration?

Integration of CMOS and MEMS together enables the development of new technologies in many applications including automotive, healthcare, and industrial applications. Particularly monolithic integration is becoming more interesting to reduce the cost of manufacturing and packaging in addition to improving the performance of the overall system [109]. For companies such as Intel, Texas Instruments, Analog Devices, and others who have clean rooms as part of their assets, the integration process can be done by slightly customizing their CMOS fabrication process.

Monolithic integration of MEMS with CMOS to develop an active neural probe is challenging, but it is so far the unique solution to get the smallest form factor and avoid a large part of the complicated packaging process. Using monolithic integration for extremely low power applications is also critical to avoid the current consumption required to drive the interface pads capacitance. In addition, bringing the circuits closer and closer to the signal source (neurons) relaxes the noise performance requirements by reducing the impact of different interference sources caused by the long interconnects.

It can be also seen that packaging technologies (e.g. flip-chip, wire-bonding,..etc) does not scale very well in interface or pad size requirements and can be, therefore, considered independent of the CMOS technologies, which scaled rabidly during the last decade. Therefore, not only expensive areas of silicon are continuously wasted due to interface pads, but also miniaturization of microsystem comes to an end because pads area may dominate the total implementation area.

It is inevitable that monolithic integration of MEMS with CMOS in neural interfaces is becoming the next step. Indeed, shrinkage in size and power due to monolithic integration can provide the next quantum leap in scaling of implantable devices for healthcare, neuroscience, and brain-machine interface applications. This chapter discusses a step towards highly-dense massive-parallel neural recording through monolithic integration of CMOS circuits on the shanks of neural probes.

4.1.2 Why CMOS Circuit Integration On-The-Shank?

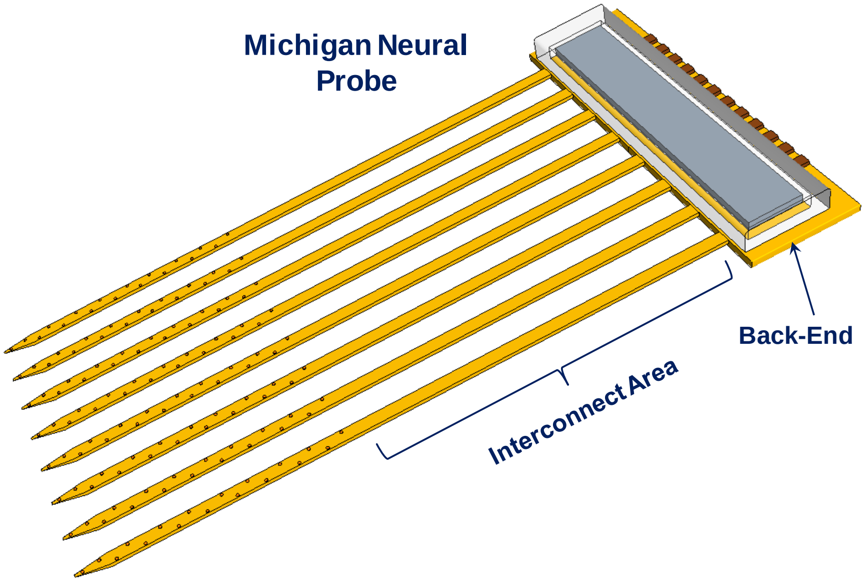


Figure 4.1. Michigan Neural Probe

The Michigan probe (Figure 4.1), as well as other neural probes, consists of multiple shanks (needle-shaped) with length of 3-10 mm depending of the region of interest inside the brain of a particular test subject. One of unique features of Michigan- and Michigan-like probes is that each shank holds more than a single recording site; therefore, the upper area of the shank is dominated by the wires that connect the electrodes to the corresponding interface pads on the back-end part of the probe.

The idea of monolithic integration of circuitry on the probe's back end has been investigated at the University of Michigan by Wise and Najafi research group (e.g. in [26, 28, 29, 51]), as well as in Europe by the European *NeuroProbes* group [68]. As reviewed in Chapter 1, an attempt to put routing (site selection) switches on the shanks to select 8 out of 188 sites has been implemented by the later group; the goal was to record from sites at different positions along the shank although only 8 channels can be provided.

In this chapter, the integration of recording CMOS circuitry is proposed for the first time on the shanks to boost the capability of high-density recording from massive amount of neurons. This approach provides a large number of advantages such as:

1) Dramatically reducing the area of the back-end by moving most circuit blocks on the shanks themselves. This makes the probe integration into a 3-D platform easier with fine probe-to-probe pitch.
2) Dramatically increasing the microsystem area efficiency by using the areas of the shanks. This also boosts the number of parallel channels recorded in a small form factor. It may not be possible to realize a practical microsystem if a design relies only on the back-end area for circuit implementation unless 3-D circuit stacking is used with through-silicon via (TSV) technology. However, using TSV may limit the probe's pitch at the 3-D recording.
3) Improved energy-efficiency of the circuit since many noise and interference sources are removed/reduced when the low-noise amplifiers are very close to the electrodes.
4) Neural signals attenuation due to long shank interconnects is brought to a minimum, which also improves the signal-to-noise ratio (SNR).

The high-density recording that can be achieved with this method allows a boost in the accuracy of spike sorting algorithms and, therefore, provides more precise single neurons recording [73]. It also enables the next quantum leap in massive parallel recording and allows the whole microsystem to maximize the use of CMOS technology scaling.

4.1.3 Post-Processing of CMOS to form MEMS Structures

There are three methods that can be used for monolithic integration of CMOS and MEMS depending on the challenges of the design itself, such as processing steps, temperature effects, cost and others. Depending on what is better to overcome these challenges, one can use MEMS-first CMOS-last, CMOS-first MEMS-last, or a continuous combination of both. The pros and cons of these methods are discussed in this subsection followed by the approach used for this work.

4.1.3.1 CMOS-First, MEMS-Last Monolithic Integration

This approach was pursued by many groups (e.g [110]) to provide versatility to MEMS designers to explore more integration options after they receive the CMOS wafers or die. This approach may also push some of the MEMS steps to the CMOS steps, which simplifies the process for MEMS designers and avoid as many photomasks as possible. In addition, it can avoid the topographical variations of MEMS structures that may affect the lithography in CMOS fabrication [111]. Moreover, it is more cost-effective at the development stage, when the high costs of CMOS wafer fabrication should be avoided by using MPW (Multi-Project Wafer) available through companies such as MOSIS. So, it actually the most advantageous and practical approach for rapid prototyping and small quantity MEMS products [112].

It is worth mentioning that care should be taken if special MEMS steps are required to realize the final device. For example, if high temperature steps (e.g. LPCVD for polysilicon deposition requires >1000°C) are needed, then aluminum metallization need to be replaced by materials such as tungsten, which can withstand high temperatures [113]. Nonetheless, some developed work still used the CMOS-first except for metallization that was performed later at the MEMS steps.

4.1.3.2 MEMS-First, CMOS-Last Monolithic Integration

In general, MEMS structures are less affected by the overall thermal budget that CMOS circuits [111]. Therefore, using this approach is unavoidable if the use of high temperature MEMS processes is required, while keeping aluminum metallization for circuits. In addition, it may provide other advantages in the packaging depending on the technique to be used [113, 114]. This approach may be also more economical especially if high-technology CMOS process is used for circuitry. Using this method, therefore, can improve the yield by removing any defective MEMS devices before they go into the CMOS steps.

4.1.3.3 Interleaved MEMS and CMOS Monolithic Integration

For large-scale manufacturing, this approach provides the maximum cost-efficiency as both optimization degrees of freedom are available. However, it may also require many changes to the process flow in order to provide the optimized solution [[113]]. Therefore, depending on the solution for cost-efficiency this may impose limits of the performance of both MEMS and CMOS parts. In addition, research and development using this approach is almost impossible as it does not support rapid prototyping[113].

4.1.3.4 Proposed Approach and Steps for Active Neural Probes

To enable the integration of CMOS circuits on the shank, area and power intensity are very critical in keeping the shank width small and limiting the tissue damage due to shank penetration during the implanting surgeries. Therefore, a special advanced CMOS process may be required to enable low-voltage design, reduce the size of interconnects, and provide an efficient method to replace the MIM capacitors for area reduction.

A CMOS-first MEMS-last approach may be also more practical at this point due to four main reasons: 1) rapid prototyping for proof-of-concept, 2) cost-effective development using MPW, 3) simplified MEMS steps which also limit the number of photomasks required, and 4) topographical variations in neural probes (especially the shank) are so severe that there is high probability to affect the CMOS lithography and other fabrication steps.

All required metallization are pursued using the CMOS process, and the MEMS post-processing are limited to the following steps:

1) Deposition of electrode material (Iridium Oxide, IrOx)
2) Probe releasing using deep reactive-ion etching (DRIE)

In the next section an overall modular active probe architecture with sub-modular shank-based design is discussed in addition to introducing the CMOS process technology (SOI 150nm process).

4.2 Modular Probe Design in 150nm SOI CMOS

4.2.1 A 150nm FD-SOI CMOS Process

A fully-depleted silicon-on-insulator (FD-SOI) process was used in this design and was provided by MIT Lincoln Laboratory (MIT-LL). This process is optimized for subthreshold-operation below 0.5V and provides several advantages for the active neural probe such as:

1) At least 50% of power is consumed by digital circuits; therefore, reducing the digital power by operating in the subthreshold regime can tremendously help reduce the overall system power. This also reduces the leakage current which improves the energy-efficiency of dynamically-operated circuit and increases the lifetime of the battery.
2) A linearized capacitor (between n+ poly gate and an n-type island implant layer) with low temperature and voltage coefficients. This capacitor provides ~9fF/μm^2 which is about an order of magnitude lower area than the typical MIM capacitor (1fF/μm^2).
3) There is no body contact in FD-SOI; while this is very challenging for analog circuit design, it also provides an extremely compact circuit layout.
4) It has a potential for 3-D circuit stacking with extremely small (<7μm) thickness per tier. This feature allows extremely dense recording in very small area.

While it may be challenging to design robust circuits with this process, the aforementioned benefits open up new capabilities for understanding brain functions in addition to developing a miniaturized systems for diagnosis and treatment of disorders such as Parkinson’s disease and epilepsy.

4.2.2 A Shank-based Module

The objective of this work is to design modular probes that can be expanded in a 3-D packaging platform as discussed in Chapter 2. Another objective is that each probe should have a variable expandable number of shanks; therefore, the shank itself needs to be considered as a module to allow a flexible probe design.

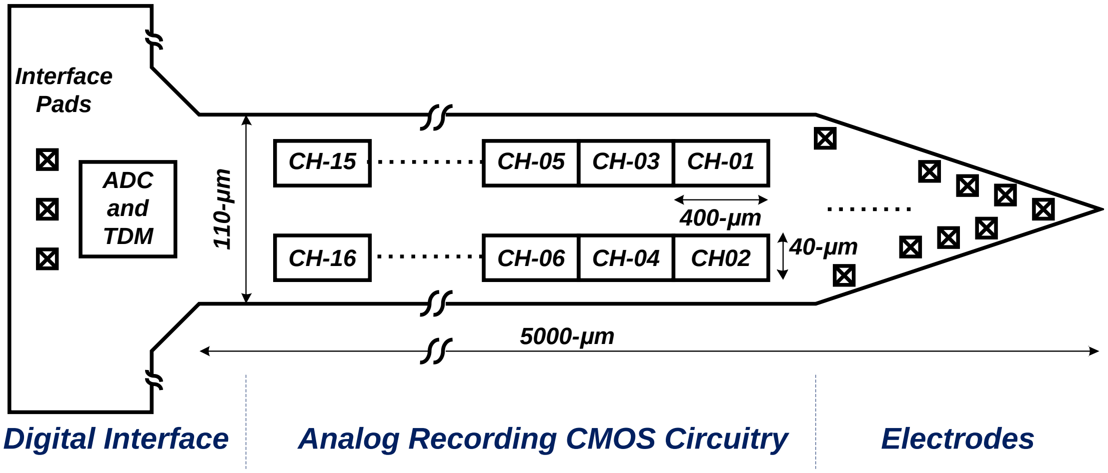


Figure 4.2. A modular shank architecture

Figure 4.2 shows the proposed shank architecture split into three main parts as follows:

1) The part close to brain tissue: this part is called the tip (triangle-shaped) and includes 16 electrodes near the edges for reduced tissue interaction and high quality recording. Each electrode is 144 µm^2 in area.
2) The middle part: this consumes the rest of the shank's area and includes all the 16 channels' analog circuits (low-noise amplifier and band-pass filter). This part is very critical in terms of limited area that has direct impact of shank width (typically ~100µm) and, consequently, tissue damage. The shank length should be variable with a minimum of 3 mm. The 16 channels are placed in a 8-by-2 array so that a thick supply and ground lines can be shared in the middle of each two channels. This requires the total area of each recording channel to be within 400x40 µm^2. This total area (0.016mm^2 per channel) requires the power consumption to be limited to a maximum of 8µW/channel. However, if 3-D circuit stacking is used to double or triple the number of electrodes and channels, this number should be ~3µW/channel.
3) The upper part (back-end): this part includes time-division multiplexing (16:1), high-speed analog-to-digital conversion, and serial digital interface to limit the number of pads.

The circuit implementation is very similar to the architecture discussed in Chapter 2; only 2 pads per shank module are required for data acquisition, while other control and programming signals can be shared for all shanks at the same probe. As an improvement, this particular system requires only 1 pad per shank module as discussed next.

4.2.3 System Architecture for a 128-Channels Neural Probe

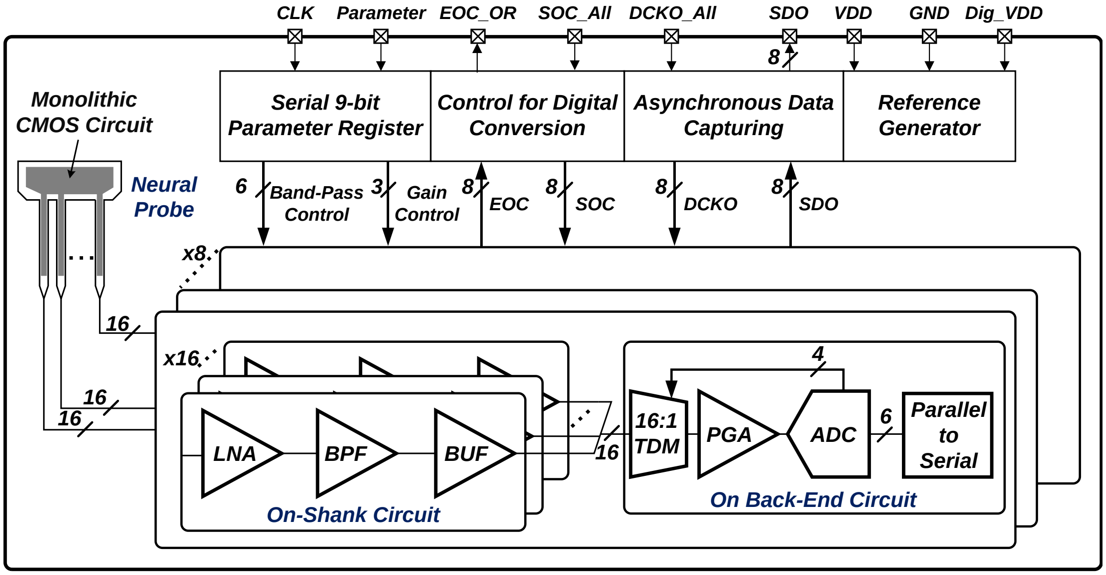


Figure 4.3. A 128-channels circuit architecture for the monolithic active neural probe

As discussed previously, each shank operates as an independent sub-module including 16 channels similar in circuit implementation to AFE-16 in Chapter 2. It yields a higher energy-area efficiency especially that this process provides a capacitor technology that is an order of magnitude area efficient than typical MIM capacitors.

The architecture of the 128-channels monolithic circuit is shown in Figure 4.3. The eight shanks have the 128 analog channels circuits; each analog channel has a low-noise amplifier (LNA), a tunable band-pass filter (BPF), and a buffer (BUF). The back-end has eight 16:1 time-division multiplexers (TDM), each followed by a programmable-gain amplifier (PGA) and a 6-bit SAR ADC that supports rail-to-rail asynchronous operation. While each shank has its own ADC and parallel-to-serial converter, as shown in Figure 4.3, the data clock-out (DCKO), start-of-conversion (SOC), and end-of-conversion (EOC) signals can be all shared among shanks by using simple AND and OR logic gates as denoted by the final signal name at the output pads.

The goal of this design is to achieve a 3μW/channel power consumption with an area efficiency that allows high-density simultaneous recording beyond 1024 channels. The next section shows the simulation results for each circuit block.

4.3 A 0.5V Subthreshold-optimized 128-Channels Analog Front-End Module

This section describes the simulation results of the 128-channels analog front-end operated with a nominal single supply of 0.5V. As mentioned previously, the circuit implementation is very similar to the system in Chapter 2; however, the main differences are: 1) techniques such as cascoding for analog circuitry are avoided, and 2) all analog circuits operate in weak inversion and transistor sizing is, therefore, different to avoid non-idealities such as mismatch. The following subsections shows the simulation results and power consumption achieved for each circuit block.

4.3.1 Low-Noise Amplifier

Figure 4.4 shows the schematic of the implemented two-stage OTA and the ac-coupled closed-loop amplifier with all sizes and parameters denoted. The low-noise amplifier consumes ~360 nW with an input referred noise of 9.6 μV_{rms} at 10 kHz bandwidth. Mid-band gain of ~36 dB and phase margin of ~80° was also achieved in simulation. Figure 4.5 shows the transient and gain response of the LNA. The dynamic range is 38.6 dB at total harmonic distortion (THD) of ~1% and, therefore, the overall noise efficiency (K) is $2.2x10^8$.

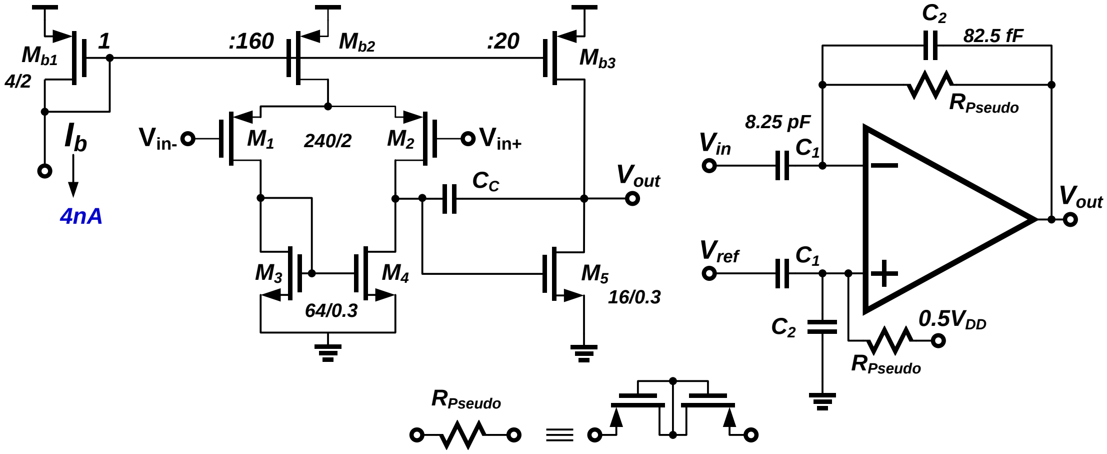


Figure 4.4. Schematic of the subthreshold low-noise amplifier

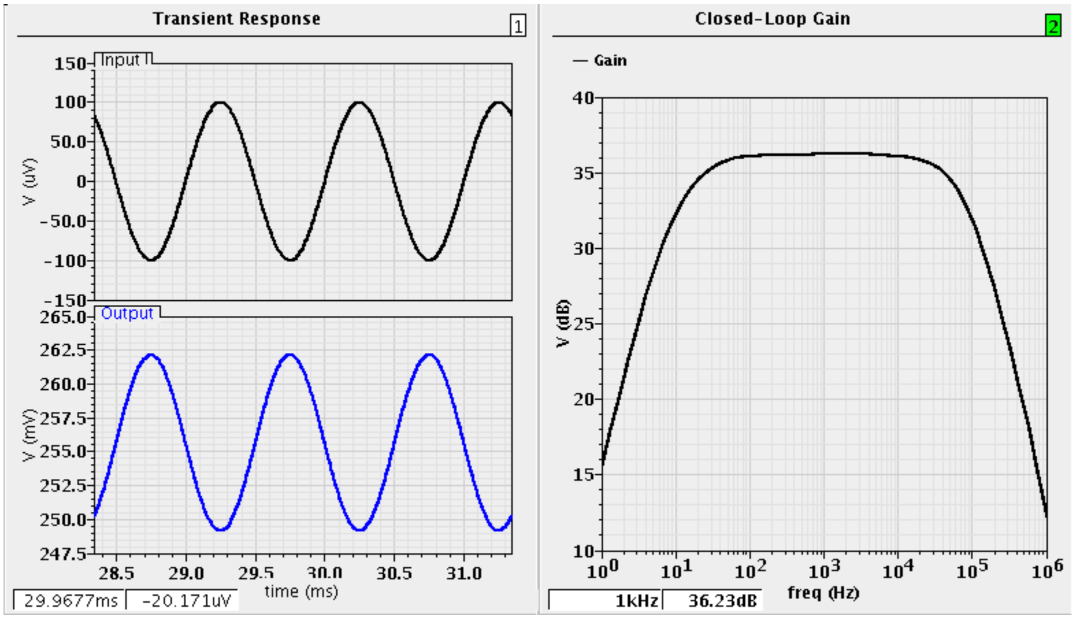


Figure 4.5. Transient and gain response of LNA

4.3.2 Band-Pass Filter and Buffer

Figure 4.6 shows the schematic of the implemented band pass filter (BPF), which provides an additional gain of two (6 dB). The BPF can be also tuned for the low-frequency corner from below 1 Hz to above 1 kHz, and for the high-frequency corner from below 1 kHz to 9 kHz. Depending on the high frequency corner, the power consumption can be anywhere from 9 to 60 nW. Figure 4.7 shows the simulated transient and gain response of the band pass filter. A buffer follows the BPF and consumes 375 nW to drive the time-division multiplexer before the shared programmable-gain amplifier.

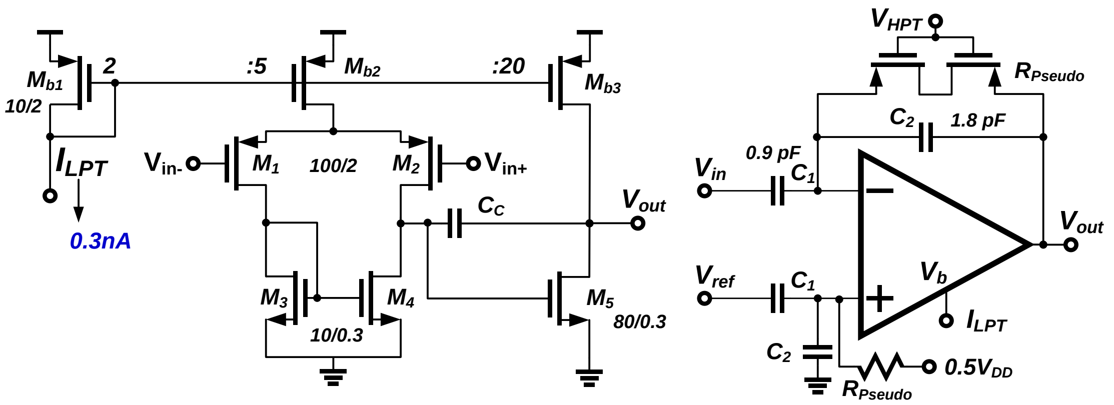


Figure 4.6. Schematic of the band-pass filter

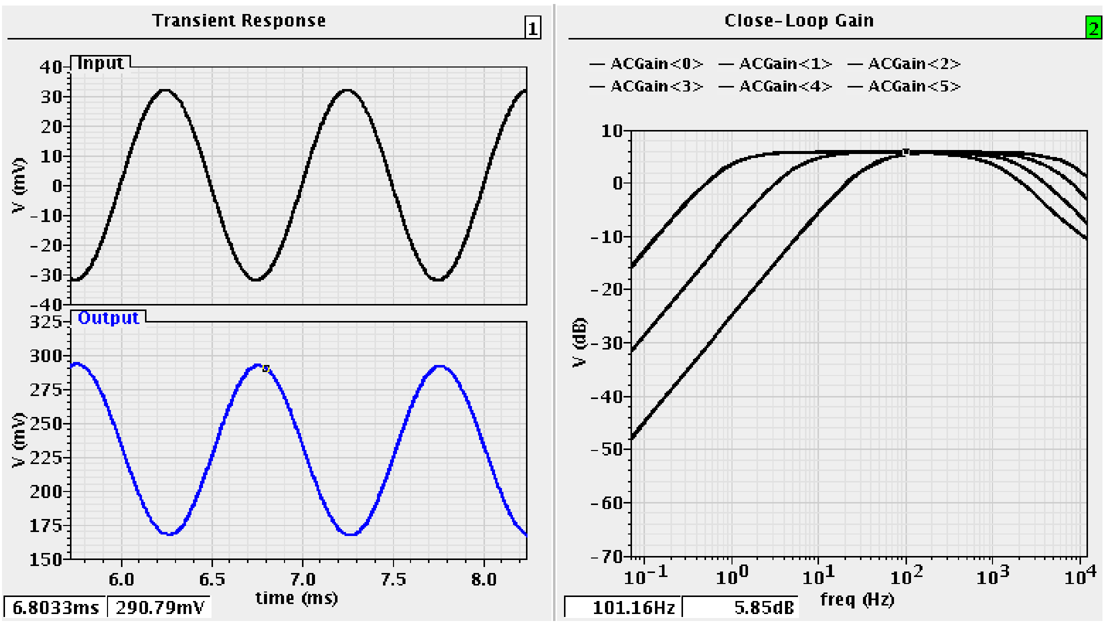


Figure 4.7. Transient and gain response of BPF

4.3.3 Programmable Gain Amplifier

Figure 4.8 shows the schematic of the 3-bit programmable-gain amplifier (PGA), which uses different input capacitors to change the gain and has the same feature of slew rate/phase margin compensation using a variable miller capacitor as described in Chapter 2. Capacitor ratios can vary from 5x (~14 dB) to 22x (~27 dB). Figure 4.9 shows the simulated different gain responses achieved with a fixed power consumption of 2.25μW.

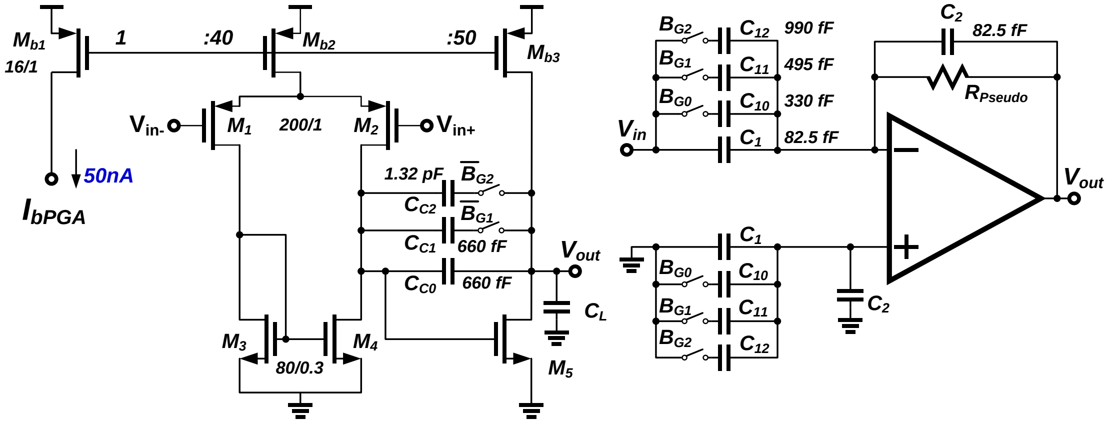


Figure 4.8. Schematic of the 3-bit programmable-gain amplifer

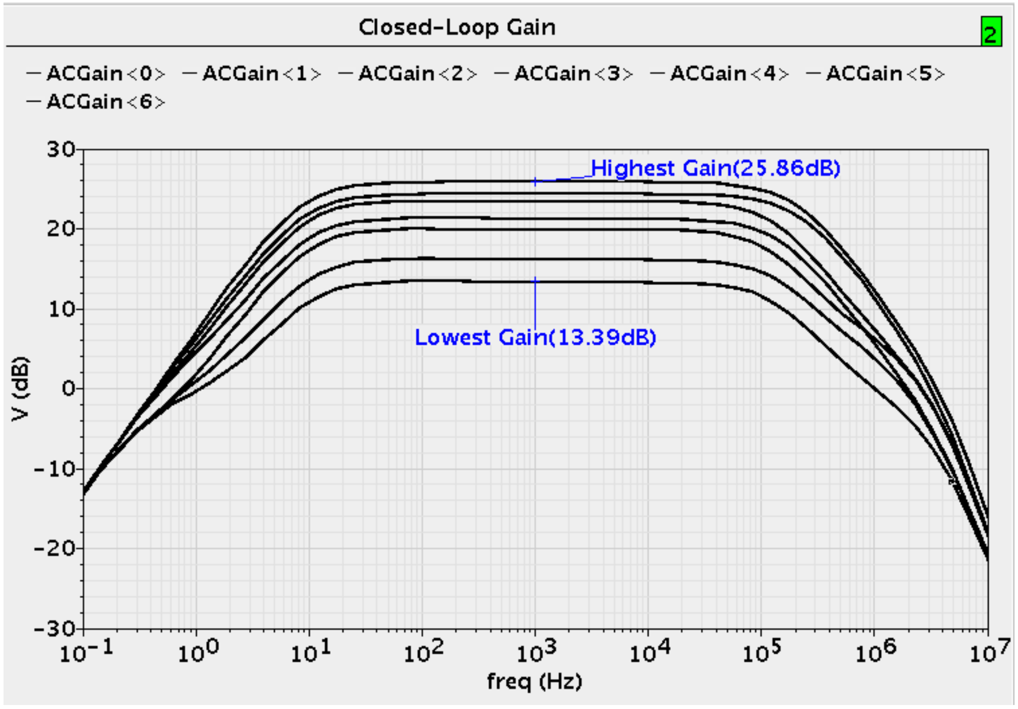


Figure 4.9. Programmable gain response of PGA

4.3.4 Asynchronous 6-Bit SAR ADC

Figure 4.10 shows the schematic of the implemented 6-bit SAR ADC with a binary-scaled capacitor bank. The ADC operates nominally with a sampling rate of 320 kS/s and up to 400 kS/s. Supply voltage can be increased if higher sampling frequency is required. Figure 4.11 shows the different signals involved in the conversion process including some internal signals inside the ADC such as the DAC and sampled voltage. The overall RMS power consumption at 320kS/s is 3.5µW, which is about 218.75nW/channel.

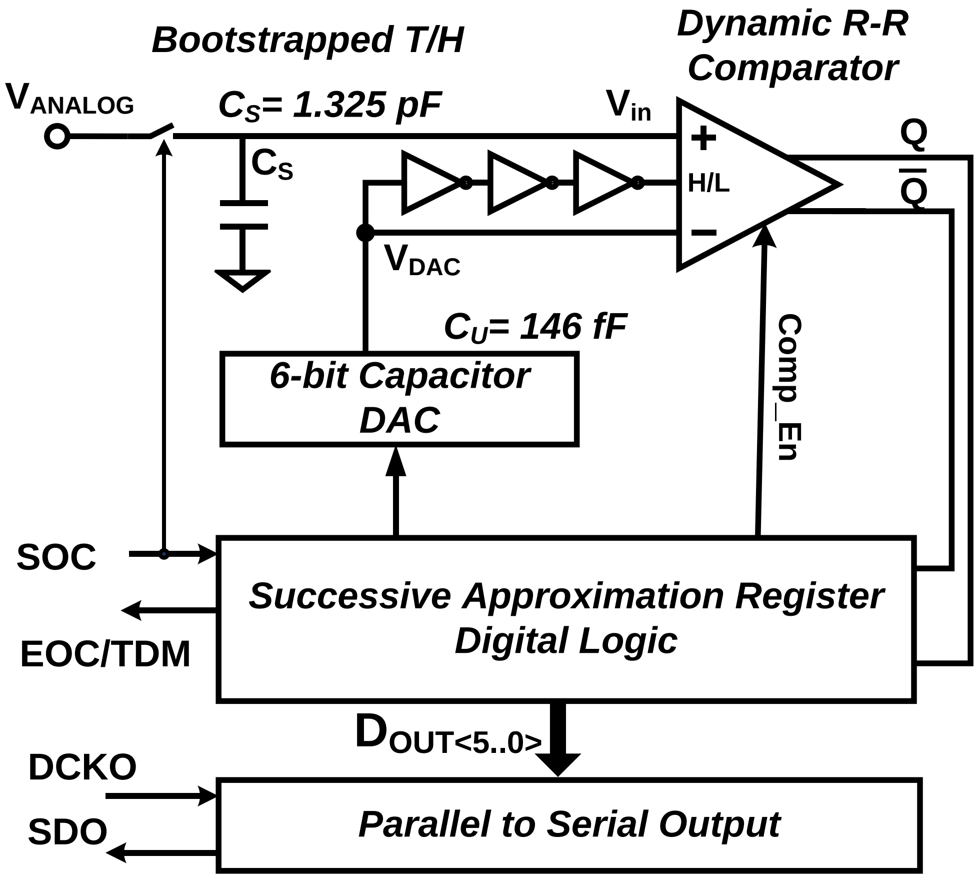


Figure 4.10. Schematic of 6-bit analog-to-digital converter

As shown in Figure 4.3, band-pass corners and gain values can be programmed serially. For further interface minimization, the conversion control (handshaking) signals for all the 8 ADCs can be combined. For example, the SOCs of all ADC can be wired to one pad (SOC_All) driven by the neural processor, the 8 EOC signals can be OR'ed together to one output pad (EOC_OR), the data clock-out (DCKO) signals can be also wired to one pad (DCKO_All) driven by the neural processor too. The total interface pads, including power supply, can be a total of 16 pads. The overall power per channel is less than 1.5µW

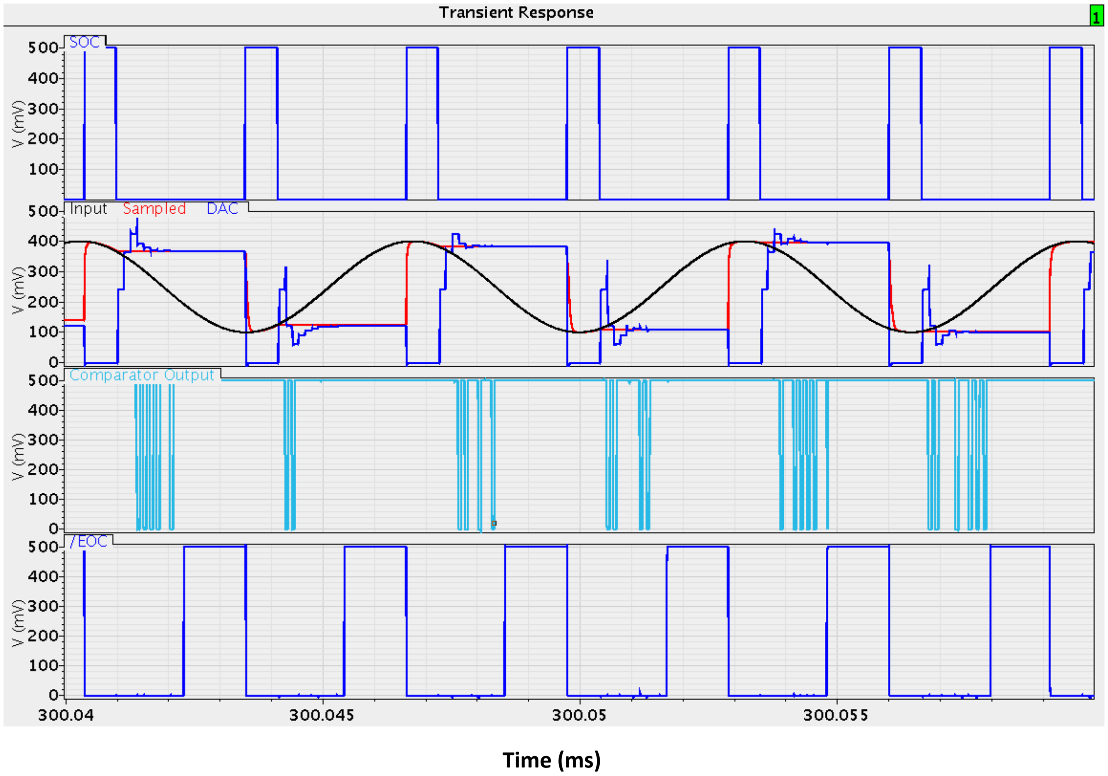


Figure 4.11. ADC simulation results

4.4 Measurement Results and Performance Comparison

4.4.1 Active Probe Layout

The chip was fabricated using 150 nm fully-depleted silicon-on-insulator (FD-SOI) CMOS process provided by MIT Lincoln Laboratory (MIT-LL) and operated nominally at 0.5V. Figure 4.12 shows the full die that is used for post-processing to release the active probe; each die has four Interdigitated probes. Two of the probes (design A) have 128-channels and the other two have 64 channels (design B). Design A has eight shanks with 16 sites each and the shank spacing is 200 μm and width of 115 μm. Design B has eight shanks with 8 sites each and the shank spacing is 200 μm and width of 60 μm.

Figure 4.13 shows a probe back-end, which includes integrated circuits (ADC and TDM) and interfacing pads. It also shows the electrode sites and the analog channels integrated on the shanks of the opposite probe. Design A probe has 33 pads (60x60 μm^2, 80 μm pitch) for power, biasing, data acquisition, and testing. However, only 16 of them are required for actual operation.

In the test setup (Figure 4.14), FPGA was used to emulate the host (processor) and a LabVIEW interface program was developed to provide programming commands and receive and demultiplex the output data.

The AFE architecture significantly reduces the number of leads compared to both previous designs. The next subsections show the measured performance of each individual block followed by performance comparison with state-of-the-art systems.

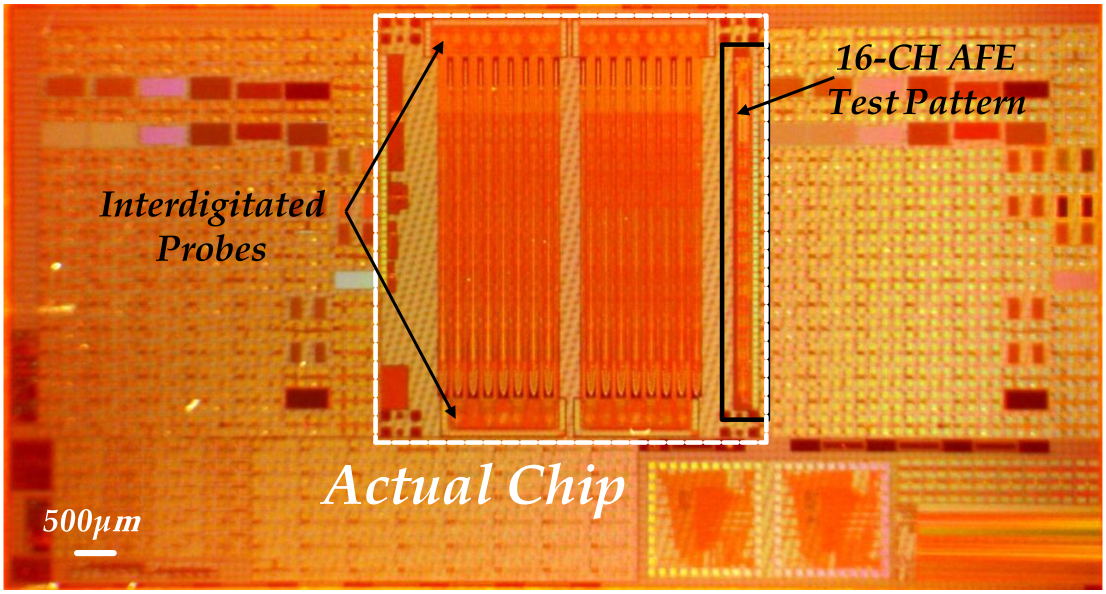


Figure 4.12. Full chip microphotograph included in a larger die to improve the post-processing

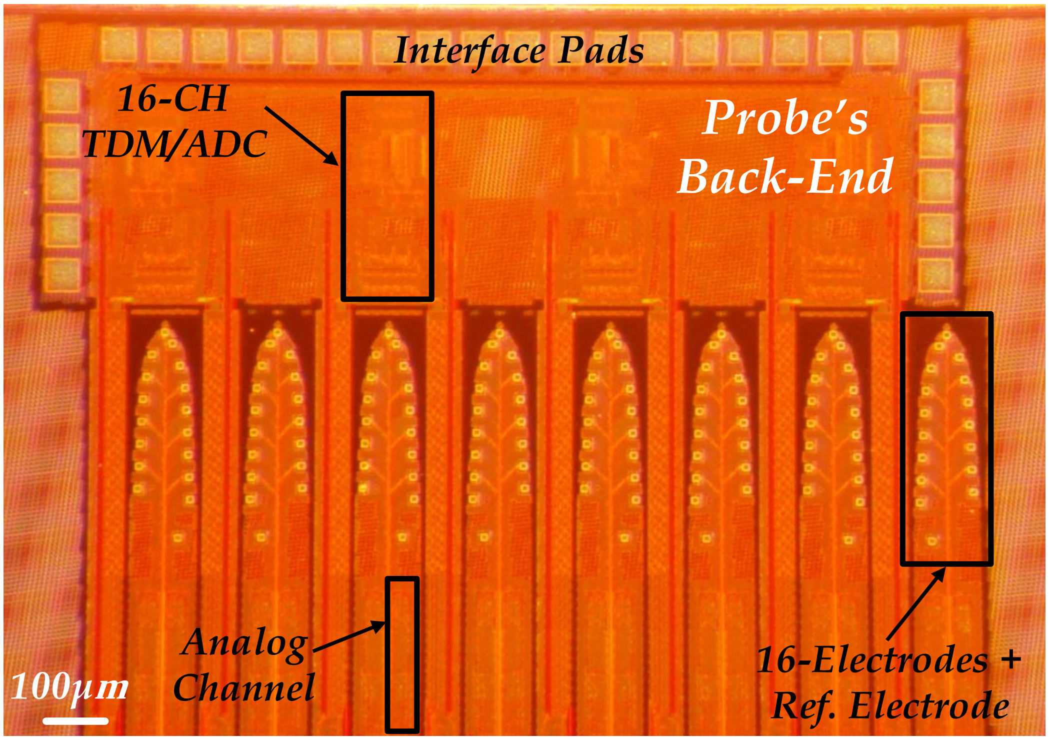


Figure 4.13. Chip microphotograph showing the CMOS circuits on shanks and back-end

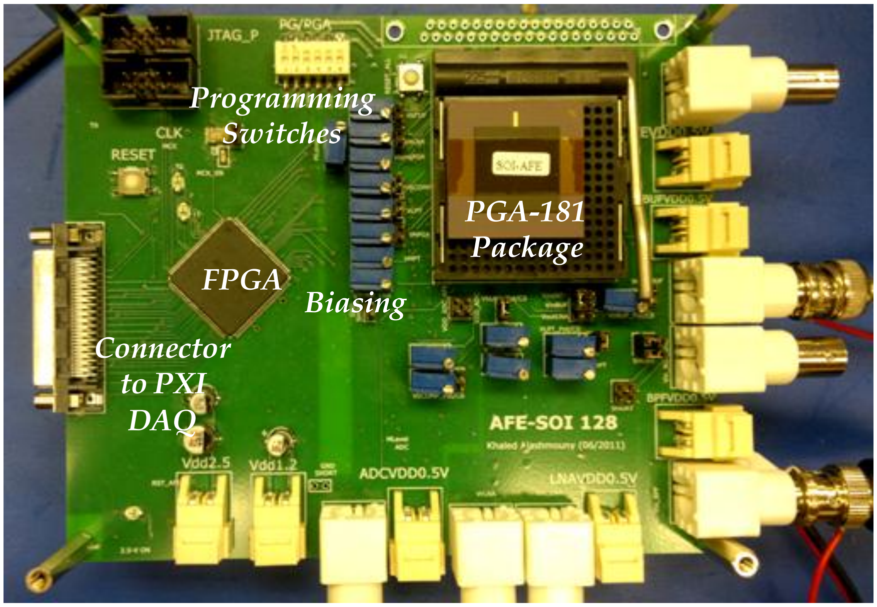


Figure 4.14. Measurement Setup

4.4.2 Low-Noise Amplifier

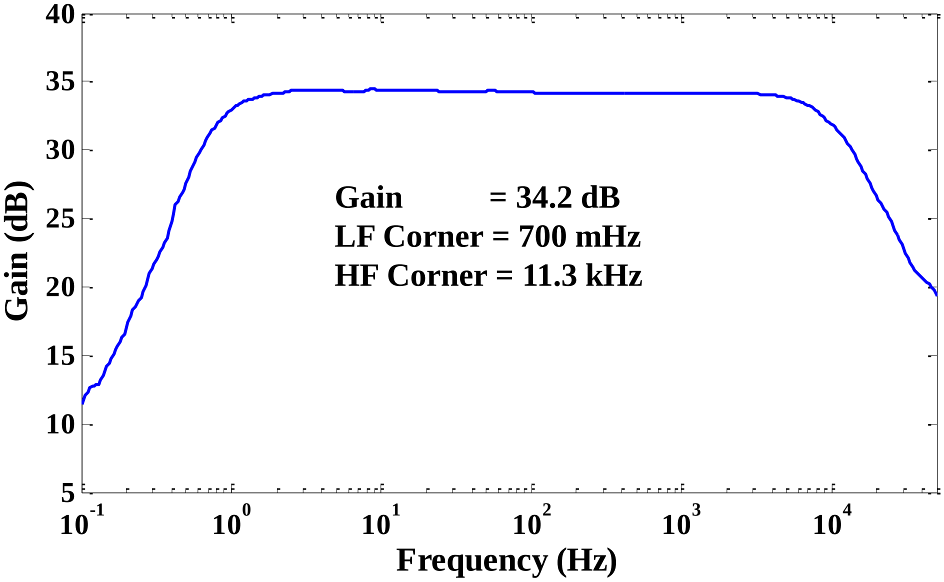


Figure 4.15. Measured LNA gain response

Figure 4.15 shows the gain response of the low-noise amplifier indicating a mid-band gain of 34.2 dB and low-frequency and high-frequency corners of 700 mH and 11.3 kHz, respectively. Noise performance is shown in Figure 4.16, and indicates an input-referred noise of 9.8 μV_{rms} integrated across 100 kHz bandwidth. The overall calculated NEF is 3.2 and K is 2.34*10^8. The measured common-mode rejection ratio (CMRR) and power-supply rejection ratio (PSRR) are more than 60 dB and 55 dB, respectively.

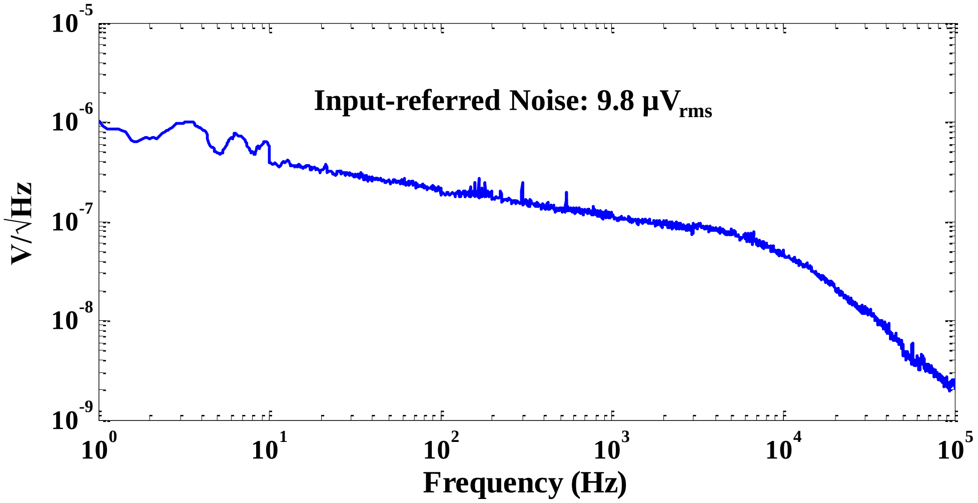


Figure 4.16. Measured LNA input-referred noise

4.4.3 Band-Pass Filter

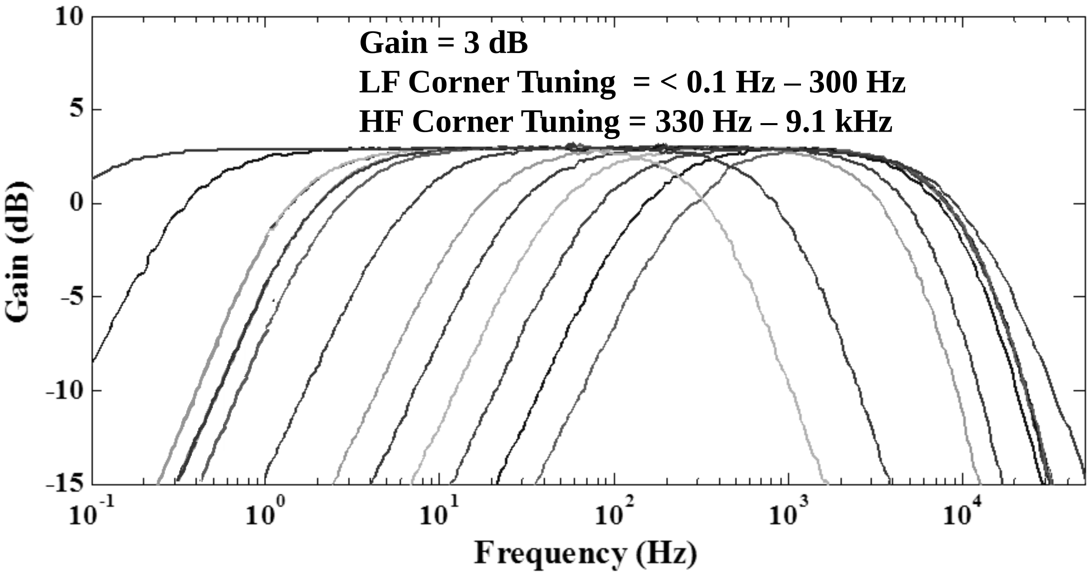


Figure 4.17. Measured gain response of the BPF under different tunings of frequency corners, and die photograph of it

Figure 4.17 shows the BPF measurement results of gain response under different tuning conditions; confirming the wide-range tunable bandwidth from 330 Hz to 9.1 kHz for the high-frequency corner and less than 0.1 Hz to 300 Hz for the low-frequency corner.

4.4.4 Analog-to-Digital Converter

Figure 4.18 shows the spectrum and non-linearity measurements of the fabricated SAR ADC. The measured ENOB at maximum input frequency (at sampling frequency of 320 kHz) is 5 bits with 40.5 dB SFDR and -37 dB THD. With an ENOB of 5 bits, still the quantization noise (input-referred) is smaller than the LNA noise. The measured differential non-linearity (DNL) is also shown to range mostly between ±0.5 LSB but the maximum and minimum values are 2.5 LSB and -0.9 LSB, respectively. The measured integrated non-linearity (INL) mostly lies in ±1.5 LSB interval with a maximum and minimum of 1.5 LSB and -1.7 LSB, respectively. In this design, the user can vary the digital supply voltage (DV_{DD}) up to 0.8V to accommodate sampling frequencies higher than 400 kHz. At nominal sampling of 320 kS/s, the total power consumption is 3.5µW (218.75 nW/channel), which also indicated a FOM value of 342 fJ/CS

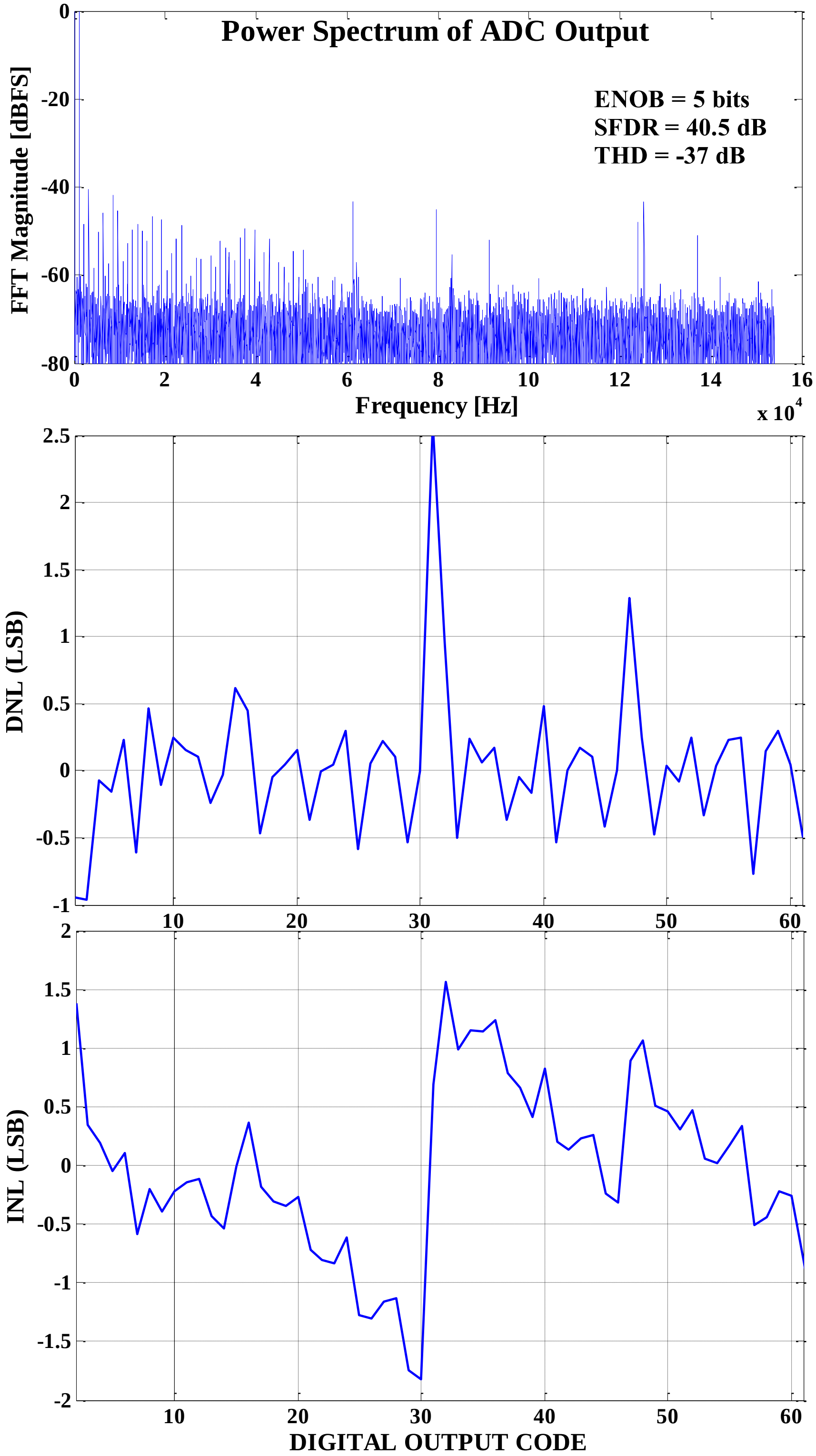


Figure 4.18. ADC measured frequency response, differential, and integrated linearity

4.4.5 Performance Comparison

Reference	AHI-16	AHI-128	AMI-128
Technology	0.25μm	0.18μm	150nm
Low Noise Amplifier			
Area [mm^2]	0.05	0.01	0.009
Voltage [V]	0.9	0.5	0.5
Power [µW]	0.99	0.95	0.36
DR[dB] at THD<1%	47.6	46.7	37.1
Noise$_{in\_ref}$ [µV$_{rms}$]	4.8 (10kHz)	4.9 (100kHz)	9.8 (100kHz)
NEF	2.9	2.6	3.2
NEF^2·(V$_{DD}$)	7.56	3.38	5.12
K (x10^8)	5.98	4.91	2.34
CMRR [dB]	>62	>70	>60
PSRR [dB]	>59	>70	>55
BPF and PGA			
Power [µW]	2.31	0.1-0.37	0.576
LF Cutoff [Hz]	<0.1-1000 (3-bit)	0.2	<0.1-1000 (3-bit)
HF Cutoff [kHz]	1-17 (3-bit)	>10	1-17 (3-bit)
Gain [dB]	52.4 to 79.8 (3-bit)	4.5 to 29.1 (4-bit)	*50.39 to 62.86 (3-bit)
ADC			
Power/Ch [µW]	0.66 at 20.16kHz	1.25-1.5	0.22 at 20.16kHz
Sampling/Ch [kHz]	6.25-50	Up to 10 MHz	>20kHz
ENOB	7	4.8 up to 7.8	5
Overall Evaluation of Analog Front-End			
Power/Ch [µW]	3.96	2.5	1.4
Area/Ch [mm^2]	0.07	0.027	0.02
EAE (x10^{-15})	7.51	30.23	36.44
EE (x10^{-9}) (without Area consideration)	0.526	0.816	0.73
Specific Wires/Ch	2:16	2:32	2:32

*Simulated

Table 4.1. Measured performance and comparison with previous version

The performance comparison is summarized in Table 4.1. The proposed AFE module in this work achieved an EAE (enery-area-efficiency) of 36.44x10^{15}. This is the highest efficiency among all the reported designs. It provides also the lowest power consumption (1.4 µW per channel) as compared with the previous work. Figure 4.19 shows the location of this work in the plot of noise and area efficiency. The monolithic integration is the most potential approach for massive parallel recording as it achieved the best energy-area efficiency reported to date. It also reduces the overall system area and avoids the challenging packaging issues.

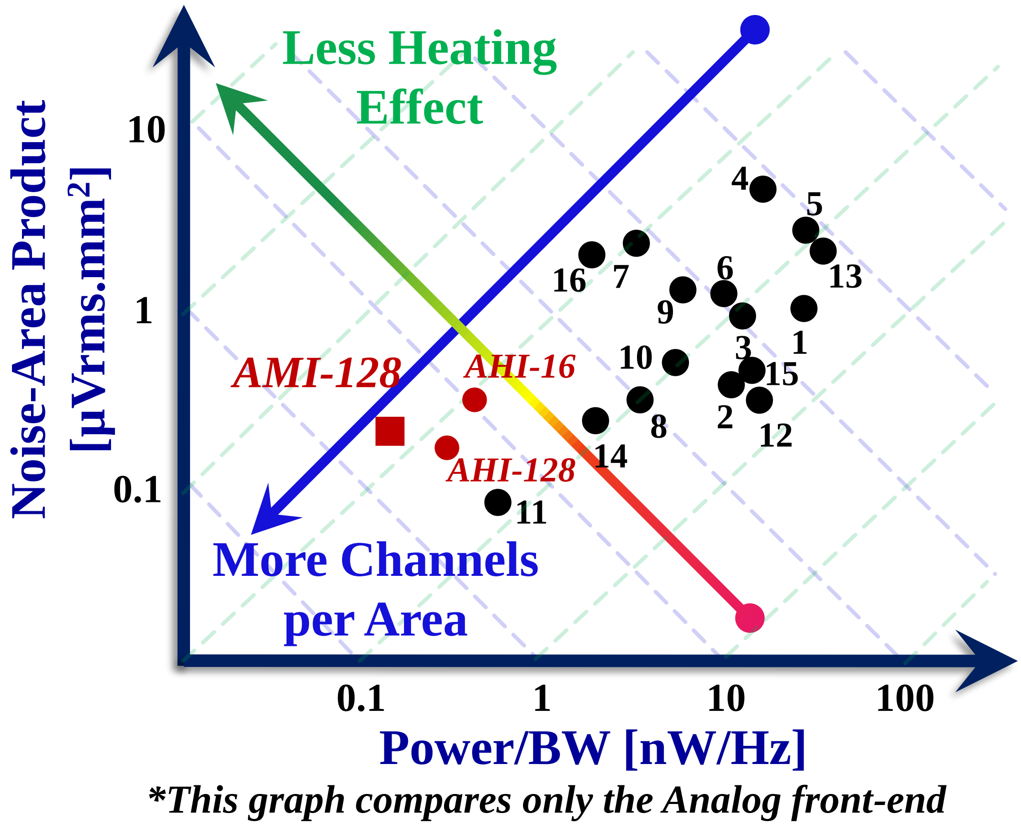


Figure 4.19. Performance chart of the implemented AFE.

4.5 Conclusions

This chapter introduced the architecture of a 128-channels active neural probe that used monolithic integration of MEMS and CMOS, in 150nm SOI process, to achieve extremely high-density parallel recording of massive amount of neurons in a small form factor. Compared to the previous designs (AHI-16 and AHI-128), using monolithic design allowed further reduction of the noise-energy-area product, and it significantly reduces the area of the back-end of the probe and avoids the packaging/integration challenges. The simulation and measurement of individual circuit blocks was shown to allow ultra-low power consumption of less than1.5 μW/channel, which is the lowest reported to date. Using the linearized capacitors provided by the technology allowed the use of ac-coupling in an extremely small size due to high capacitance per unit area.

Further scaling of neural analog front-end may require using both monolithic integration to reduce area and digitally-assisted techniques to transfer computations outside the implanted device for a reduction of power consumption.

CHAPTER 5

SUMMARY, CONTRIBUTIONS AND FUTURE WORK

5.1 Summary

The goal of this research was to provide a roadmap and develop techniques to address the different analog front-end circuit challenges towards increasing the number of simultaneously recorded neurons in multi-probes 3-D microsystems. A massive parallel 3-D microsystem (conceptual view shown in Figure 1.23) requires the integration of both the MEMS neural probes and the analog front-end CMOS circuitry to form “active” probes. Towards achieving this goal, it was important to address specific circuit challenges such as area, integration, power, and noise. In three different modular analog front-end (AFE) designs, different techniques based on physics, signal processing, and integration were investigated and proven to allow the future scaling required for massive parallel neural recording.

The first AFE was designed for hybrid integration and investigated some critical circuit challenges in power, area, interface, and modularity. The front-end features an extremely low power consumption (4μW/channel), optimized energy efficiency using moderate inversion biasing in low-noise amplifiers (K_{LNA} of $5.98x10^8$ and NEF of 2.9) and programmable-gain amplifiers, a minimized asynchronous interface (only 2 per 16 channels) for command and data capturing, a power-scalable sampling and digital operation (up to 50kS/s/channel), and a wide configuration range (9-bit) of gain and bandwidth. The implemented front-end module has achieved a reduction in noise-power-area by a factor of 5-25 times as compared to the-state-of-the-art front-ends reported up to date, and has a potential number of channels of 10^2 per mm^2.

The second AFE aimed for further scaling and an actual implementation of 128 channels without site selection. So, in addition to the techniques used in the first AFE, a dc-coupled interface with analog compression was used to reject the electrode dc offset and reduce the implementation area more than three times compared to the previous version. A current-mode approach was also used to increase the dynamic range while using 0.5 V supply voltage to further reduce the power consumption. Techniques such as receiver/digitally-assisted analog design were also explored to transfer most computation power consumption to the receiver outside the brain.

While the previous two AFEs relied on hybrid integration with neural probes, the third AFE investigated another leap in neural recording by monolithic integration of CMOS recording circuits on the shanks as well as on the back-end of the probe. Integrating the circuits on the shank helped reducing the overall dimensions of the probe, especially at the back-end part. Monolithic integration may be the more effective approach to allow simultaneous neural recording towards achieving more than 1024 sites recording in an extremely small volume. This design also took advantage of a special subthreshold process and achieved the lowest energy-area efficiency amongst other designs.

The results in each of the three analog front-ends can contribute to the advancement of low-voltage analog and mixed-signal circuit design and implantable 3-D neural microsystems for complex neuroscience studies.

5.2 Contributions

This research work contributes to the fields of analog and mixed-signal integrated circuits as well as implantable microsystems for neuroscientific studies. A diversity of contributions has been made in both individual components and system levels.

In a system level, the contributions are as follows:

- Three analog front-ends were designed with a minimal interface and modular fashion to provide the recording microsystem with the flexibility and scalability required for the expansion into 3-D microsystems. This has been done in particular on the acquisition part of each analog front-end, where pads can be shared between different modules. In addition, the asynchronous digital operation help minimize the interface and interference of high speed clocking (AHI-16 and AMI-128).

- Power scaling based on sampling operation has been realized. It allows the use of oversampling that is required by some neuroscience algorithms for sorting the spikes more efficiently. This also adds to the flexibility and power efficiency on the systems level.
- Signal processing techniques have been shown to save both area and power simultaneously on the systems level and help refer signal reconstruction at the receiver, and thus relax the analog design to some extent (AHI-128). This contribution may set a good direction for further research as will be described in next section.
- Monolithic integration of circuits on shanks was shown to extremely reduce the area, but more importantly it also used an unexploited area (the shanks themselves) so far for analog recording. This also allowed a design towards an active 128-channel neural probe.
- A figure of merit was developed to provide an overall evaluation of the system taking into account noise, power, area, and bandwidth. This figure of merit (referred to as Energy-Area Efficiency or EAE) indicates how large a channel density is supported by a given design. A chart was developed also to show the heating effect and suggest a reduction of area and power simultaneously.
- The designed analog front-ends provide the best reported energy area efficiency reported to date and showed different techniques reduce area and power simultaneously.

One the individual circuit components the contributions are as follows:

- Moderate inversion biasing has been investigated in the design of low-noise amplifiers (in AHI-16) and was shown to provide the best reported energy efficiency based on K with acceptable statistical variation among process. Same biasing technique was used to reduce the power of programmable gain amplifiers.

- A novel dynamic rail-to-rail common-mode comparator has been design with a feed-forward technique to allow higher full-scale range in the asynchronous ADC implemented in both AHI-16 and AMI-128.
- Automatic gain-stability compensation has been implemented in programmable-gain amplifiers to avoid the use of unnecessary bias current and save power consumption.
- A dc-coupled low-noise amplifier was developed and made use of both the electrode model capacitance and resistance, and the delta compression technique, which rejects DC by nature. This contributes significantly in reducing the implementation area.
- Current mode techniques and processing were used in the design of different circuit blocks including analog compressor, programmable-gain amplifiers, and ADCs. It has been shown to allow higher dynamic range in a small supply voltage operation.
- A current-mode tunable delay-line (all-pass filter) was designed to operate in subthreshold region with high tuning dynamic range.
- Current-mode FLASH ADC was design for low interference and high-speed operation to allow high oversampling ratios.
- The design of several analog circuits in a subthreshold optimized SOI process was shown to be feasible and allow further reduction in area by exploiting a high-density capacitance (linearized capacitor) technology.

Finally, appendix A introduces some additional work that has been done to explore a communication method toward allowing the physical decoupling of the recording front-end from the telemetry or data processing platform using the brain as a communication medium.

5.3 Future Work

Although several contributions have been made in this research towards massive parallel recording, there are still some areas of improvement in the design and more importantly there are additional circuits and MEMS blocks required to realize a complete 3-D microsystem. Future work suggestions can be summarized as follows:

- The analog front-ends (AHI-16 and AHI-128) should be integrated with an actual neural probe and further in-vivo measurements (chronic and acute) need to be performed.
- The monolithic active probe (AMI-128) should be post-processed to deposit the recording electrode material and release the probes themselves. This SOI active probe need to be further studied in-vitro and in-vivo. A re-design may be required but this is definitely one of the most promising approaches for massive parallel neural recording although the development cycle may be longer than the case of hybrid integration.
- The direction of receiver/digitally assisted circuit design needs to be further explored into more creative design ideas and signal processing functions to further reduce the power and size of integrated circuits. Combining SOI with signal processing techniques may also cause another leap in massive parallel recording.
- Further power reduction at the system level can be achieved by implementing digital compression techniques as part of the analog front-end. This will have great advantages since neural signals are quiet most of the time.
- Some critical components need to be designed to realize the full architecture of a 1024-channel system. A neural microprocessor that can handle all 1024 channels needs to be designed together with wideband telemetry and power delivery schemes. These three critical components have their own challenges and require innovative architectures to increase area and power efficiencies.
- Further lead reduction can be done at the packaging level by sharing supply voltages and acquisition signals among different integrated circuit modules.
- Circuit protection and sealing need to be investigated, especially for the hybrid integration approach. Hermetic sealing should be explored especially for chronic recording.

In general, a significant number of challenges are yet to be addressed with the continuous demand of more channels. As the understanding of brain dynamics evolves neuroscience community will keep following a channel-scaling trend similar to Moore’s low for transistors scaling.

APPENDIX A
IBCOM (INTRA-BRAIN COMMUNICATION)- A NEW METHOD OF SIGNAL TRANSMISSION FOR MINIATURIZED NEURAL RECORDING/STIMULATING SYSTEMS

In this chapter[7] a new method of signal transmission for neural recording/stimulation microsystems. Intra-brain communication or IBCOM is a wireless signal transmission method that uses the brain itself as a conductive medium to transmit data and commands between neural implants and data processing systems outside the brain. *In-vitro* and *in-vivo* experiments have been performed to validate this method on rat brain. Multi-channel modulated neural signals have been transmitted through rat brain, and received and processed to retrieve the original form at a receiving station. IBCOM is viable at a frequency spectrum ranging from 100 kHz to 50 MHz. IBCOM was found to have no effect on neural firing activities as it was tested while simultaneously recording neural signals. Two µ-IBCOM CMOS chips were designed and fabricated for an *in vivo* test bed to transmit two prerecorded neural signals at different binary frequency shift keying (BFSK) modulation frequencies to validate the feasibility of IBCOM concept. The chips were fabricated by using TSMC 0.25-µm technology and were packaged for full implantation in a rat brain except for external power delivery. The on-chip µ-IBCOM core circuit occupies 400x270-µm^2 in chip area and transmits balanced-charge signals using a 2.5-V power supply. Platinum wires of 50 µm diameter were used as transmission and receiver microelectrodes. A low-power wireless signal transmission through brain was successfully demonstrated between the two platinum electrodes separated by 15 mm with less than 1 µW per channel (with 10 µA transmission current) and successfully recovered the original neural signal waveforms.

[7] This research was done by the author at the University of Minnesota

A.1 Introduction

The ability to record from individual neurons in the central and peripheral nervous systems has been an important tool in neuroscience research for decades. Recent technological advances have allowed the simultaneous recording of dozens to hundreds of neurons simultaneously [2, 37, 115-117]. While major advances have been achieved by numerous research groups over the last two decades, one of the long-sought, as yet unavailable goals in neuroscience is to acquire massive-parallel access of single neuron activities from many distributed probes inside the brain. One potential solution may be to develop a nano-scale fully-implantable Neural Recording/Stimulation Systems (NRSS) that is small enough to be located in any location in the brain without major tissue interaction and with the capability to record neural activities from one or more adjacent neurons or to stimulate them. However, a critical stumbling block for this approach is the data transmission from these implants to the data processing and analysis system outside the brain. While several techniques have been proposed to address the communication issues, they are not favorable to scale the size of probes or to achieve low power to minimize tissue heating.

Two preferred ways of signal transmission between the implanted probes and the external waystation are through inductive coupling and RF telemetry [118-123]. An important precaution, however, is to avoid tissue heating and allow enough signal power to be transmitted. RF transmission frequency needs to be no more than several MHz because of tissue absorption in higher frequency signals. For inductive coupling, the signal transmission system requires the internal coil (inside the implant) to be aligned with the external coil (for receiving data) for maximum power transmission. This constrains the orientation of the implanted devices and limits the freedom of deployment. Although RF transmission does not require orientation alignment, antenna size becomes a scale-limiting factor, especially when a low carrier frequency below 10MHz is used for signal transmission. If the antenna size is reduced below a quarter of signal wavelength, antenna efficiency will drop significantly. Therefore, it is difficult to realize the probe in a small form factor. Also, high signal power attenuation is another limiting factor when using small RF antennas for transmission. Thus, both approaches are inherently unfavorable to be scaled toward an extremely small size below a hundred microns due to poor signal power transmission, large implementation size, and complicated circuitry required for operation.

A signal transmission mechanism using the brain itself as a conductive medium is investigated. Previously, for different purposes, researchers in [120, 124, 125] measured brain impedance at frequencies as low as 5 Hz and as high as 750 kHz. Values of brain impedance depend on experimental setup and environmental conditions. However, cortical impedance has been shown to have a nearly flat response across 10 Hz to 5 kHz [125]. Electrical conductivity of human cerebrospinal fluid (CSF) measured from 7 patients across the frequency range of 10 Hz-10 kHz at body temperature (37° C) showed an average conductivity of 1.79 S/m [120]. This is a pretty good number and motivated to explore the possibility of intra-brain communication for low power signal transmission.

In this work, IBCOM (Intra-Brain Communication) is introduced as a new way of communication between neural recording/stimulating systems (NRSS) and external analysis and data collection devices/stations. To do this, signal loss across the frequency spectrum of signal transmission through rat brain is reported. Next, the feasibility of IBCOM is confirmed through a series of animal experiments using live rat brain. Finally, it is shown that IBCOM does not affect neural activities nor recording systems in *in vivo* experiments. Implementation of IBCOM in a CMOS chip (μ-IBCOM) further validates this ideas and shows that it is feasible to embed IBCOM as a part of next-generation neural recording and/or stimulation microsystems.

In the next sections, the IBCOM concept is presented and described in further details including charge-balancing, frequency spectrum, DSP algorithm for modulation and demodulation of neural signals, and IBCOM effect on neural signals. The design and implementation of μ-IBCOM CMOS chip is also described and accompanied by experimental and measurement results and conclusion.

A.2 Intra-Brain Communication (IBCOM)

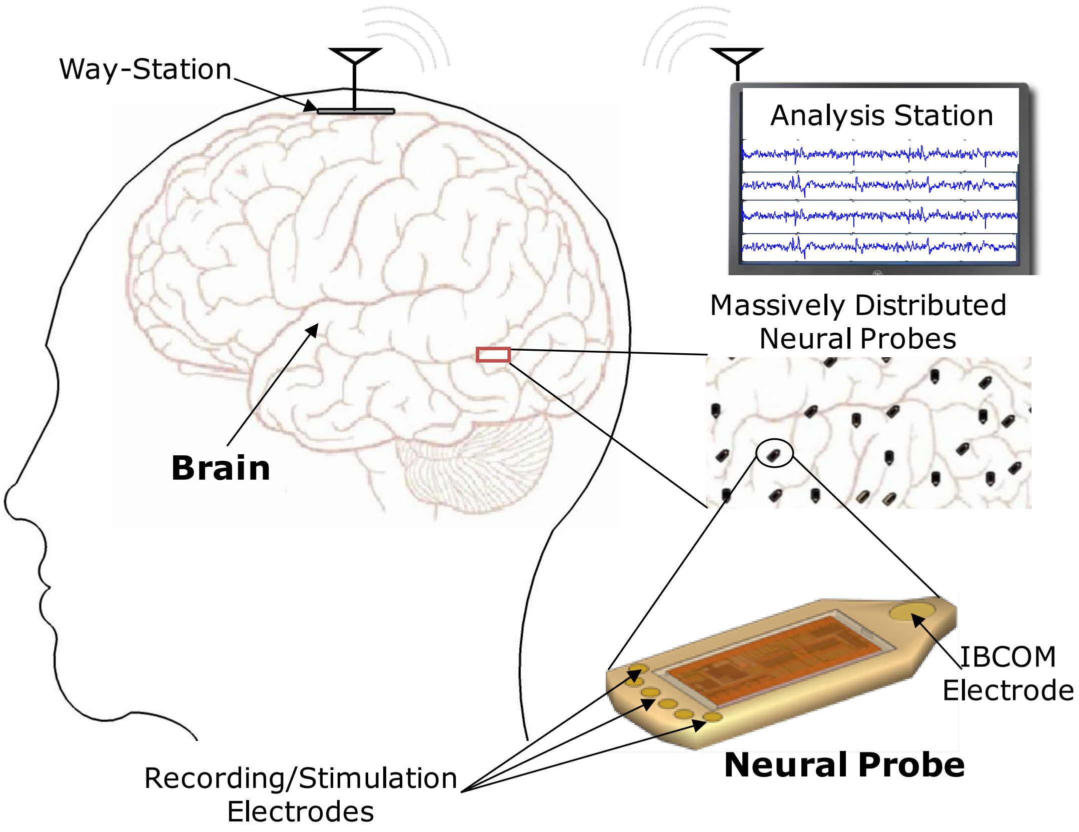


Figure A.1. Conceptual diagram of a massively-distributed neural recording/stimulating system: miniaturized neural probes are massively distributed in the brain to record/stimulate thousands of neurons simultaneously. They communicate wirelessly to an external analysis station via the way-station fully implanted on the brain surface.

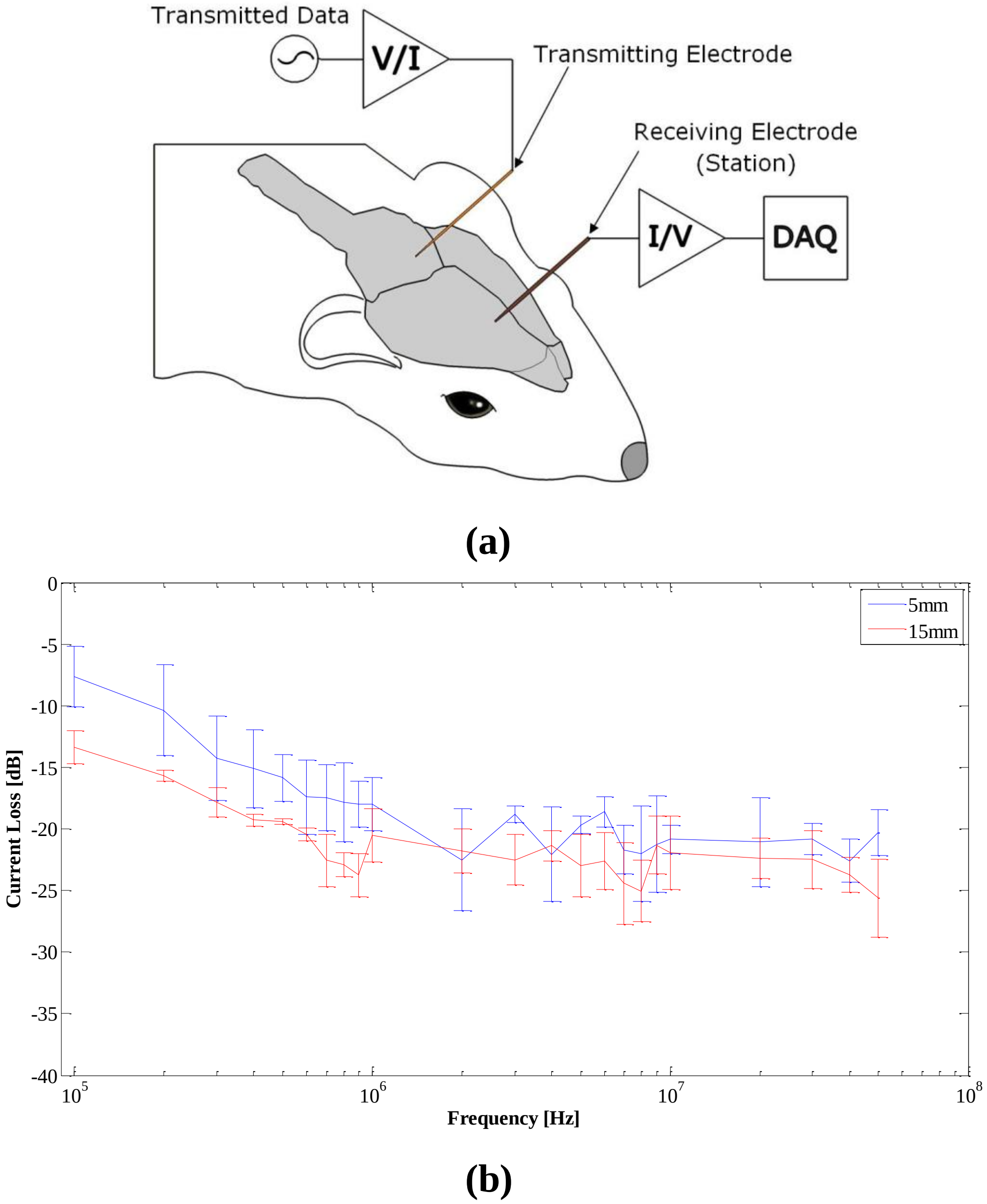
Figure A.1 shows a conceptual diagram of a massively-distributed neural recording/stimulating system (NRSS) that consists of multiple scattered neural probes and a receiving way-station fully implanted in the brain and/or spinal cord. In this NRSS, each neural probe can record adjacent neural activity and provide stimulation when needed. A major milestone to build this system and make it achievable is having a way of communication between scattered neural probes and a receiving way-station. This way-station can be implanted on the brain surface or above the dura and further wirelessly transmit/receive data to/from a data processing system outside the brain. The purpose of this paper is to report our studies validating a new method of communication using the brain as the conductive media or communication channel.

Verification of function of the intra-brain communication (IBCOM) technique required a series of animal experiments to investigate: (1) electrical signal transmission through brain and degradation (losses) of those signals as a function of frequency, (2) transmission of modulated neural signals using IBCOM, (3) receiving and demodulation of those signals to retrieve the original signals, (4) IBCOM effect on neural activities during signal transmission, and (5) optimization of IBCOM transmission to eliminate any possible affects on neural activities and recording. All animal experimental protocols were approved by the University of Minnesota IACUC and were consistent with NIH guidelines. Subsequent subsections will describe in detail the methodology of each experiment and report its outcomes.

A.2.1 Signal Transmission Through Brain

Proper transmission of electrical signals through brain requires the consideration and prevention of: (1) accumulation of positive or negative charges which may cause electrolysis, burns, and damage to brain cells, and (2) stimulation of neurons which causes the signal transmission to interfere with neural firings. Thus, a charge-balanced transmit/receive system that incorporates current (rather than voltage) as a transmitting and receiving signal parameter was designed. To avoid any possible neural stimulation, transmission frequency was chosen to be 100 kHz (ten times higher than the maximum frequency range of neural activities).

As shown in Figure A.2(a), the test system consisted of a signal generator to generate sine-waves with different amplitudes and frequencies. The generated signal was converted to current, using voltage-to-current (V/I) converter, connected to an Iridium Oxide (IrO_x) transmitting microelectrode placed in the rat brain. The transmitted signal was received by another IrO_x microelectrode at a different location in the brain. The receiving microelectrode was connected to a current-to-voltage (I/V) converter and further to a data acquisition (DAQ) system (PCI-6259, National Instruments, TX) to compare both transmitted and received signals. The electrical current losses were evaluated over 5 mm, 10 mm and 15 mm spacing between transmitting and receiving microelectrodes. A transmitted current of 100 µA p-p was tested with frequency sweep from 100 kHz to 50 MHz. Over four rats, the average current loss from transmitter to receiver over the used frequency range was 18.44 *dB*, 19.19 *dB* and 21.47 *dB* for 5 *mm*, 10 *mm*, and 15 *mm*, respectively. Figure A.2(b) shows the average current losses at 5 *mm* and 15 *mm* spacing, respectively. This result indicates that the brain is conductive across a wide range of frequency and distance and implies that brain can be used as a conductive transmission medium for IBCOM without much signal loss. This inspired the next experiment: transmitting the two modulated neural signals simultaneously through the brain and retrieving them with minimum error at low power consumption.



(a)

(b)

Figure A.2. Signal transmission through rat brain: (a) Experimental setup, and (b) IBCOM average current loss over 100kHz to 50MHz for 5 mm and 15 mm distance, respectively.

A.2.2 IBCOM Transmission of BFSK Modulated Neural Signal

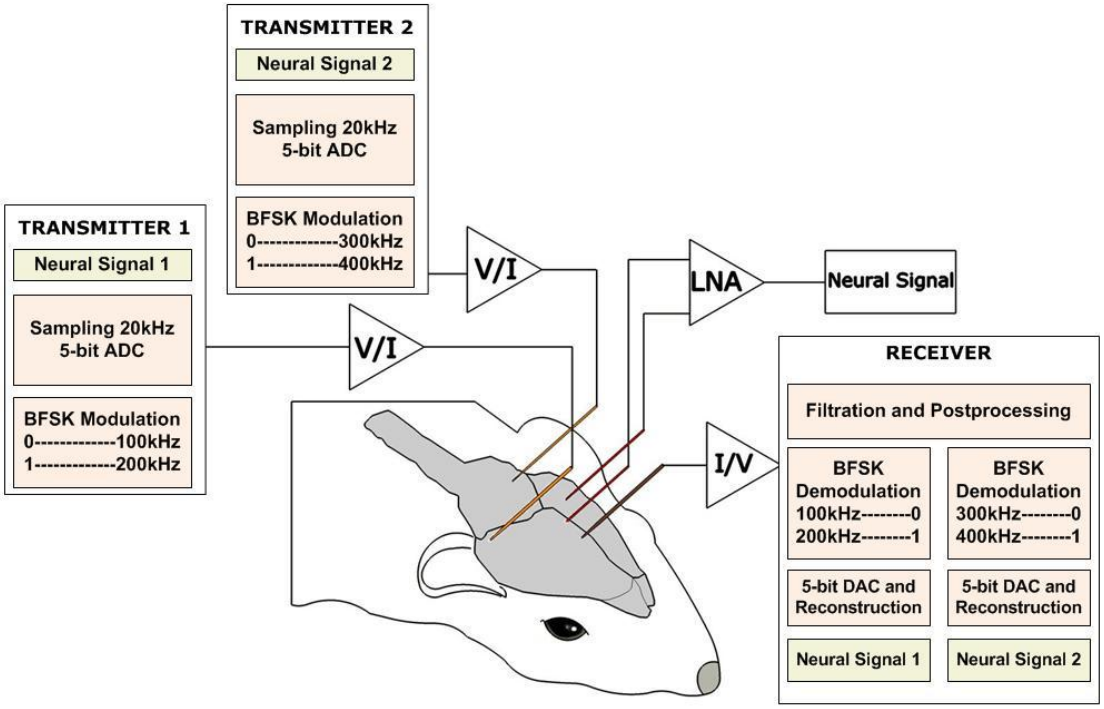


Figure A.3. Experimental setup for two-channel IBCOM to demonstrate the transmission of modulated neural signals through brain. The two modulated neural signals are transmitted simultaneously from the two electrodes and received in the single receiver electrode, then demodulated to retrieve the original signals. During the signal transmission neural signal is monitored in a separate electrode to inspect if the IBCOM signals affect normal neuron activities.

Figure A.3 shows the scheme of *in vivo* experiment for two-channel IBCOM to demonstrate the capability of sending more than one modulated neural signal through brain and retrieving them with minimum error. A MATLAB (MathWorks, MA) code was developed to convert the two prerecorded neural signals to serial 5-bit digital code and further modulate this code using Binary Frequency Shift Keying (BFSK) digital modulation. The two prerecorded neural signals were acquired from a separate recording system and sampled at 20 kHz and converted to 5-bit digital code. It was shown in [51] that 5 bits can be considered an optimum resolution when area and power consumption are of more concern. For the first IBCOM transmitter, the digital signal was modulated at 100kHz (using 100 kHz for zeros and 200 kHz for ones), while for the second transmitter the signal was modulated at 300 kHz (300 kHz for zeros and 400 kHz for ones). Each modulated signal was converted to current by V/I converters and transmitted respectively through the IrO$_x$ electrodes inserted in the rat brain. The two transmitted signals were received by a single (third) IrO$_x$ receiver electrode and then converted into two voltage signals using band-pass filters. Each signal was then BFSK-demodulated to retrieve the original analog signal (both transmitters used current of ±10 μA).

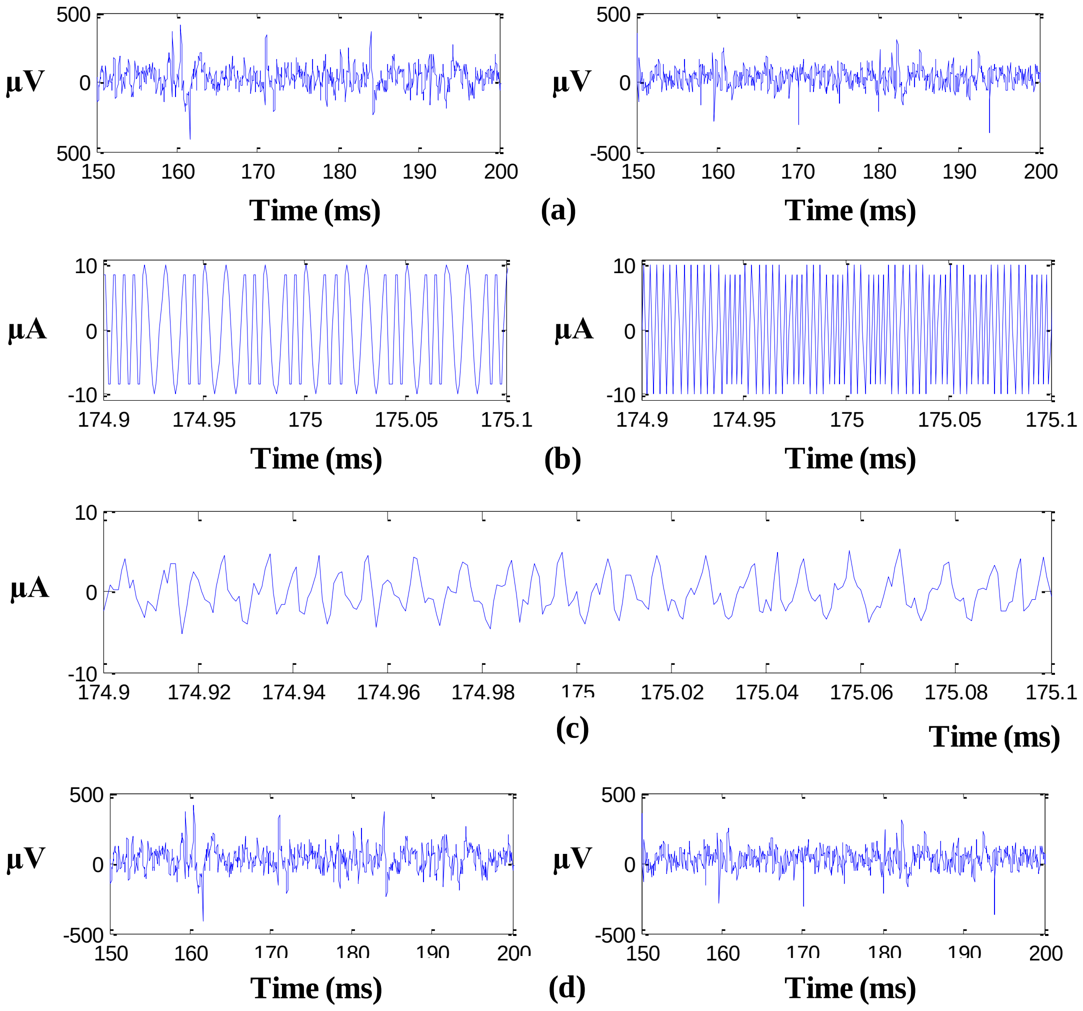


Figure A.4. Measured signals at each steps during IBCOM transmission through rat brain: (a) Two original neural signals prerecorded, (b) BFSK-modulation of two original signals, (c) Signal acquired at the receiving electrode, and (d) Two demodulated neural signals after digital-to-analog conversion.

Figure A.4 shows the measured signal waveforms at each signal processing steps. Figure A.4(a) and (b) show the two original neural signals and their modulated signals used at the two IBCOM transmitters, respectively. Figure A.4(c) shows the signal as acquired from the receiving electrode. It can be noted that both signals were combined and received from the single electrode. After splitting two signals using bandpass filters and demodulating them, Figure A.4(d) shows that retrieved signals exactly match the original signals (compare Figure A.4 (a) and (d)). This experiment was successfully repeated using different transmission currents changing from 100 μA$_{p-p}$ and down to 2μA$_{p-p}$. The signal waveforms reported in Figure A.4 used a transmission current of ±10 μA. This implies that 1 μW budget can successfully transmit the single-channel signal. The IBCOM transmitters were powered separately from each other using batteries (thus providing isolated/floating power sources). Separate battery power was also used for the receiver circuitry. This ensures that the IBCOM system can be embedded in different neural recording/stimulation sites scattered in the brain with floating (capacitive-coupled) ground.

A.2.3 Optimization of IBCOM and Simultaneous Neural Signal Recording

One of the major challenges for IBCOM is that it should neither induce any abnormal neural activities during signal transmission nor affect the function of any neural recording system. Additionally, for any fully-implanted NRSS there would be no access to an external solid ground. To test this, neural activities were recorded from the additional electrodes implanted in anesthetized, live rats during IBCOM signal transmission using a battery-powered (floating ground) low-noise preamplifier (SR560, Stanford Research Systems, CA). Figure A.3 shows the set-up to monitor neural signals simultaneously with IBCOM transmission in a live rat. Two microelectrodes (one channel of a tetrode spun out of Polyimide-insulated Nickel Chromium, 12.7 μm diameter wire, Kanthal, Palm Coast FL, gold plated to 300 kΩ at 1 kHz referenced against four shorted channels of another tetrode) were placed between the brain between the transmission and receiver IBCOM electrodes. These two electrodes were then connected to the amplifier to record differential signals. The output of the amplifier was connected to a standard PC through DAQ System (ML870 PowerLab 8/30, ADInstruments, CO) for post-processing and analysis. The DAQ system incorporated both a 60 Hz interference rejection filter and band pass filters in the range of 600 to 6000 Hz.

Neural signals were recorded in three periods: before applying IBCOM, during IBCOM, and after disconnecting IBCOM. Neural firing signals during IBCOM were unchanged. This means that IBCOM does not affect neural activities during its signal transmission. However, at the start and the end of IBCOM, the neural recordings showed two distinctive peaks as shown in Figure A.5.

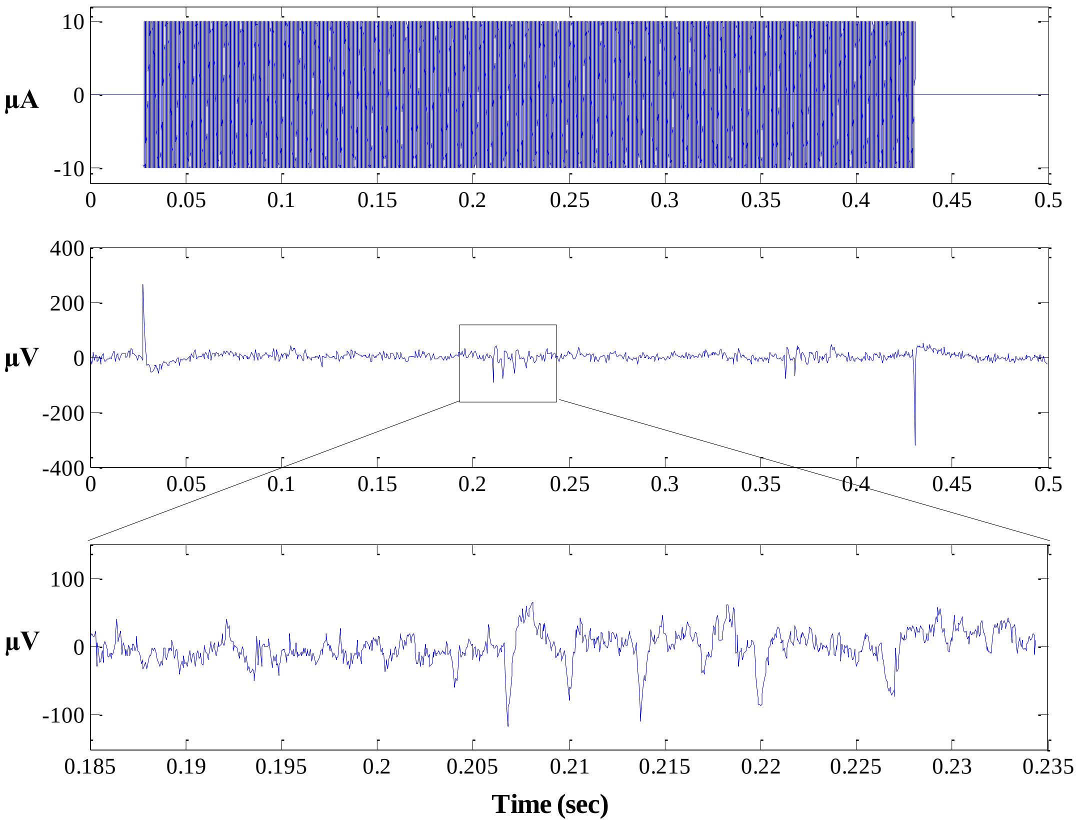


Figure A.5. Recording neural signals during IBCOM transmission. It shows two distinctive peaks when switching IBCOM signals. However, normal neural signals have been recorded between the two peaks, suggesting IBCOM does not interfere neural activities.

To eliminate those undesired signals, the IBCOM transmission current was shaped such that it ramped up from zero to the target transmission current gradually. Similarly, when turning off IBCOM, the current was ramped down gradually. Figure A.6(a) shows the shaped IBCOM signal. Figure A.6(b) shows the measured neural signal between the IBCOM transmitting and receiving electrodes while the shaped IBCOM is applied. In this experiment the ramping signal was a dummy single frequency signal modulating at 100 kHz and 300 kHz for the first and second IBCOM transmitters, respectively. The neural signal does not show any switching peaks at the beginning and end of IBCOM. This experiment has been replicated over 18 different locations in three different live (anesthetized) rats. Figure A.7 shows an example neural signal recorded at the same site before, after and during IBCOM transmission. No artifacts or abnormal neural activities were observed due to IBCOM.

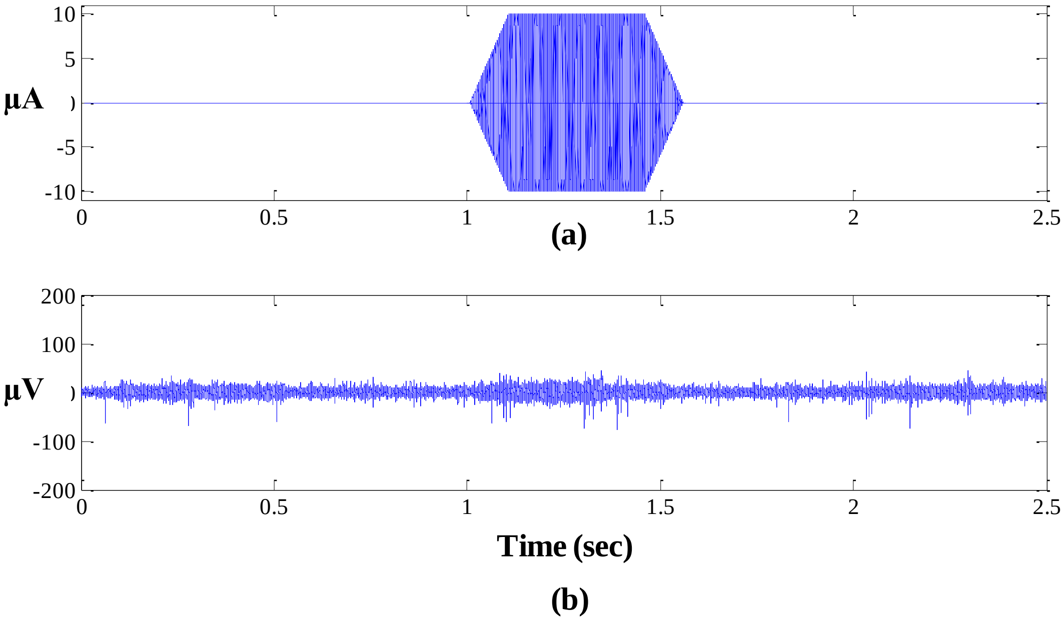
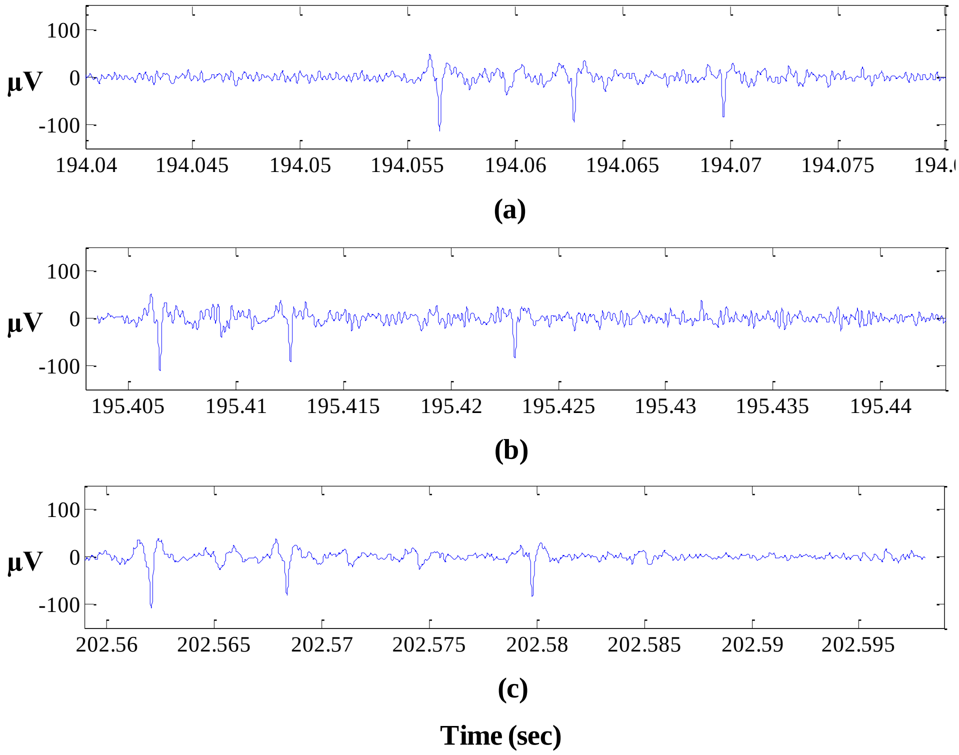


Figure A.6. (a) Shaping of IBCOM transmission signals, (b) Recorded neural signal while transmitting the shaped IBCOM signals. No abnormal distinctive peaks are observed.

A.2.4 On the potential for extra-brain receiving sites

It would be very useful for several applications to explore the possibility to locate an IBCOM receiving electrode in different sites for example in a muscular structure rather than in the brain. Therefore, the same IBCOM experiment were performed by placing a receiver electrode in the leg of the rat. The results were similar to case with a receiving electrode in the brain at the cost of increased transmission current and receiver’s area.

In the next section, a design of a µ-IBCOM CMOS chip that can be embedded in a physiological recording system implanted in brain or spinal cord is described.



(a)

(b)

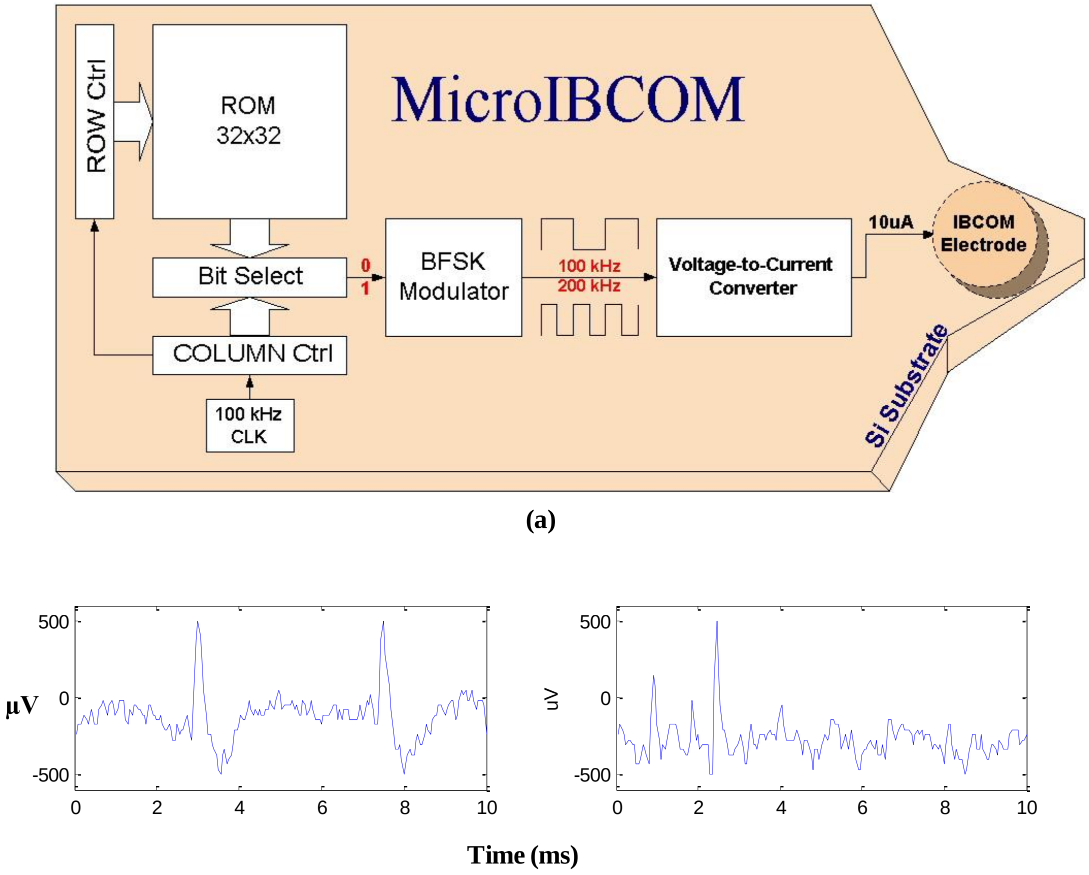
(c)

Figure A.7. Recorded neural signals from the *same brain site* (a) Before, (b) After, and (c) During IBCOM Application.

A.3 Miniaturized µ-IBCOM CMOS Chip

As discussed previously, next generation neural recording/stimulation systems (NRSS) should be small enough to be fully implanted in any parts of the brain and/or spinal cord without physical connection or wiring between these implants and the data collection system physically separated from the recording sites. In the previous section, IBCOM has been shown to be a good candidate for low-power signal transmission in a small form factor for future NRSS. To demonstrate this, two different CMOS chips were designed to test IBCOM functionality *in vivo*. The aim was to implant two chips in rat brain; each of them simultaneously transmitting modulated (simulated) neural signals at different frequencies. A receiving electrode should be able to receive, demodulate and retrieve the transmitted signals into the original neural signals. The two chips are referred to as µ-IBCOM1 and µ-IBCOM2, respectively.

In order to imitate neural signal transmission, the actual pre-recorded neural signal is stored in the ROM of the μ-IBCOM. This on-chip stored neural signal was modulated at each carrier frequency for current-mode signal transmission. Biocompatible custom packaging was built for both μ-IBCOM chips to provide power and access to the brain for IBCOM transmission. In this experiment, power (2.5 V) was externally supplied. In the next subsections, the components of the first μ-IBCOM chip (μ-IBCOM1) are described as shown in Figure A.8(a). This includes ROM storage of neural signals, BFSK digital modulator, and voltage-to-current (V/I) converter. The design of the second chip (μ-IBCOM 2) differs only in the stored neural signal (ROM contents) and the modulation frequencies utilized by the BFSK modulator. Frequencies of 100/200 kHz were used in μ-IBCOM1, and 300/400 kHz in μ-IBCOM2, respectively. The packaging and experimental methods in rat brain will be described in the last two subsections.



(a)

(b)

Figure A.8. (a) Block Diagram for μ-IBCOM1 CMOS Chip, (b) Neural Signals stored in μ-IBCOM1 (left) and μ-IBCOM2 (right)

A.3.1 ROM Storage and POR Circuitry

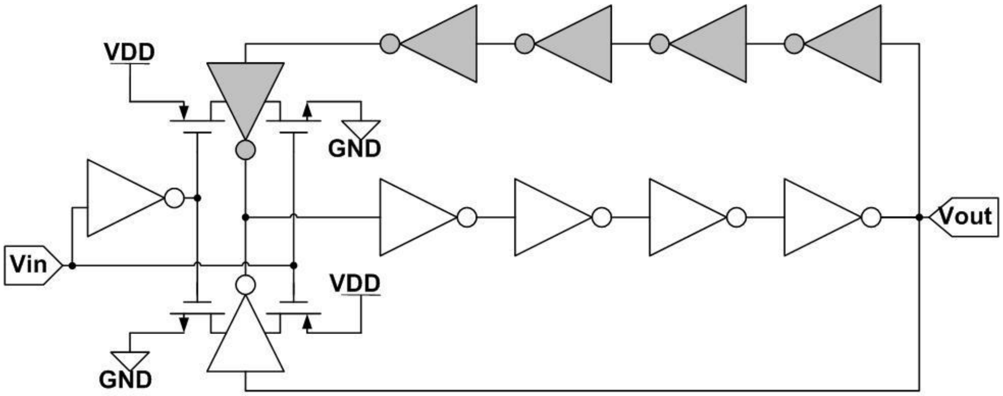
Figure A.8(b) shows the two neural signals stored in the ROM of the two μ-IBCOM chips. They are the prerecorded signals sampled at 20 kHz and converted to 5-bits digital signals. Each signal has a duration of 10 ms and runs continuously in a loop using shift registers for row and column access. The column shift registers are controlled by a 100 kHz internal clock. The row shift registers are triggered by the signal from the first column shift register. A header (24-bits) is added in the stream of transmission signals; therefore, the total number of stored bits is 1024. A 32x32 ROM was designed using one NMOS transistor for each bit. Power-on-Reset (POR) scheme is used to initialize the shift registers as well as properly initialize the internal clock and ring oscillators used in the BFSK modulators.

A.3.2 BFSK Modulator Circuitry

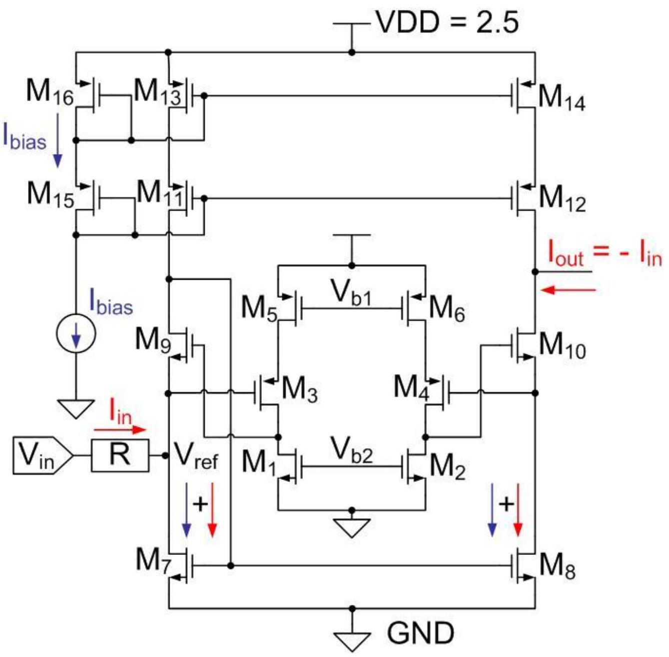
The digital output of the ROM is modulated using on-chip binary frequency shift keying (BFSK) modulation. The BFSK modulator in μ-IBCOM1 modulates the ROM data to 100/200 kHz. Similarly, the modulator in μ-IBCOM2 modulated the data to 300/400 kHz. The BFSK modulator design was based on a ring oscillator. According to the ROM output, the longer path is selected for a low modulating frequency (when ‘0’) and the shorter path for a high modulating frequency (when ‘1’). Figure A.9(a) shows the circuit schematic for the BFSK modulator used in μ-IBCOM1. Each inverter consists of three stacked PMOS and NMOS transistors to lower the frequency of operation. The voltage output of the BFSK is converted to current for the transmission in the brain.

A.3.3 Voltage-to-Current Converter Circuitry

The BFSK modulator output is buffered and sent to the voltage-to-current V/I converter shown in Figure A.9(b). A bias current of 100 μA is used as a reference current. The regulated cascode scheme, formed by M1, M3, M5 and M9 in the left branch and M2, M4, M6 and M10 in the right branch, is employed to ensure that V_{ref} node remain at the mid-range (1.25V) and to increase the output impedance at the output node. A telescopic PMOS structure of two cascodes (formed by M11-M14) was used to further increase the output impedance of the output node. Voltage and current biasing circuits are incorporated on chip. The current mirror forces the same bias current to flow in the two branches formed by M11 and M13, and M12 and M14, respectively. The input voltage (V_{in}) is connected to the V_{ref} node through a resistor of 100 kΩ. This draws an input current (I_{in}) through the resistor to the circuit via M7. The balanced structure forces M8 to flow the same current. Therefore, the induced current difference is then sourced from or sank to the output node. This allows the V/I converter to generate both positive and negative current polarities from single power supply.



(a)



(b)

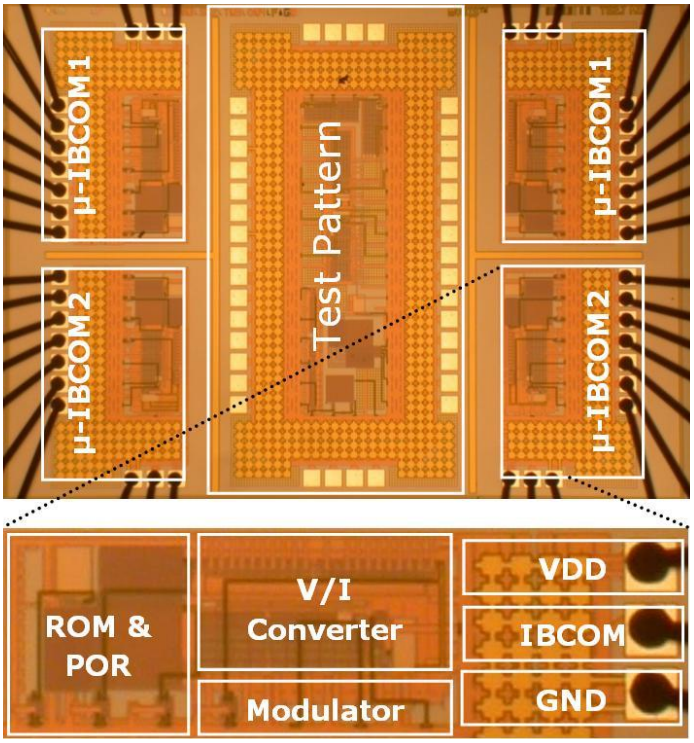
Figure A.9. (a) BFSK Modulator working at 100/200 kHz according to digital input voltage, (b) Voltage-to-Current Converter: 0 – 2.5V is converted to around ±80µA.

The voltage range (0 – 2.5V) can be converted to ±10μA using this V/I circuit. The V/I converter circuit has an operation frequency up to around 20 MHz, which is more than enough to handle the modulated signals. The output current was connected to a transmitting electrode in the rat brain either through a testing PCB for initial testing or through a custom-designed package for full implantation in the brain.

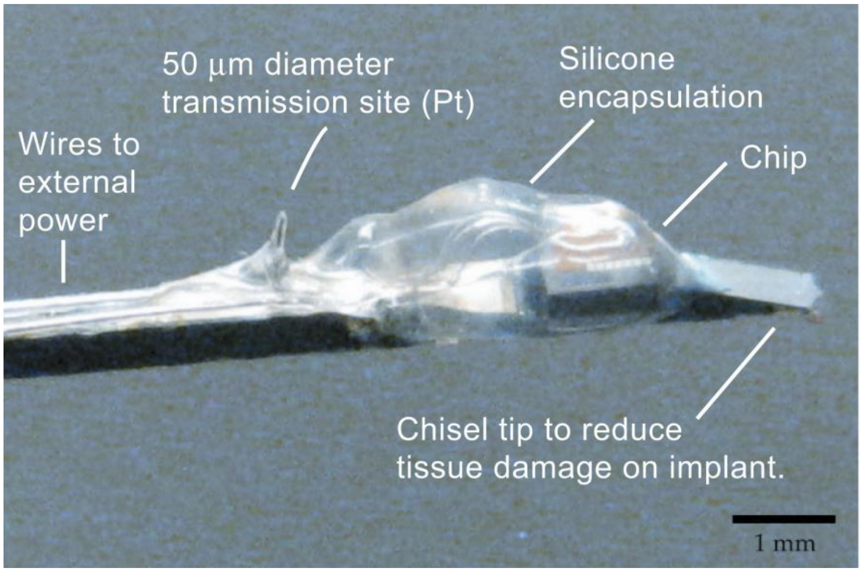
A.3.4 µ-IBCOM CMOS Prototype Chips and Packaging

The prototype μ-IBCOM chips were designed and fabricated using 0.25-µm CMOS technology as shown in Figure A.10(a). Two µ-IBCOM1 chips are located in upper corners. We designed two version of µ-IBCOM1 chips that differ only in the transmission current (±10µA and ±80µA, respectively). µ-IBCOM2 has also two version of different transmission current, shown in the two lower corners. The middle part constitutes test patterns for functional validation of all individual components. The total size of µ-IBCOM chips is 930x570 µm^2 including pads, while the core parts occupies only 400x270 µm^2.

Figure A.10(b) shows the fully-packaged μ-IBCOM CMOS chip. The μ-IBCOM CMOS chips were attached to micromachined 40 mm long silicon spears to facilitate implantation in an anesthetized rat's brain. The silicon spears were diced from a 4 inch silicon wafer to have a small cross-section (500 μm x 500 μm) and a 15-degree chisel tip to minimize trauma to brain tissue during insertion [126]. The silicon spears were coated with a 100 nm layer of alumina using atomic layer deposition. Three Teflon-insulated platinum wires (50 μm) were attached to each silicon spear and wirebonded to the μ-IBCOM CMOS chip pads. Two of the wires extended the full length of the spear and were used to connect to a battery power source outside of the rat's brain. The third wire was bent perpendicular to the spear, stripped of insulation at the tip, and served as the transmission electrode. Finally, the μ-IBCOM CMOS chips and wirebonds were encapsulated with silicone (Dow Corning 3140 RTV).



(a)



(b)

Figure A.10. (a) Die photograph for four µ-IBCOM systems and test patterns, (b) µ-IBCOM packaging for full implantation in rat brain.

A.4 Measurement Results

A.4.1 Experiments Procedure

Multiple tests and experiments were conducted using both µ-IBCOM1 and µ-IBCOM2. A printed circuit board (PCB) was designed to test the performance of both chips in saline and rat brain. The µ-IBCOM1 and µ-IBCOM2 chips, packaged by MOSIS in a PLCC52, were integrated in the PCB and their outputs were connected to rat brain using two platinum (50µm in diameter) microelectrodes. Signals from both chips were transmitted simultaneously and received at another location in the rat brain using a separate platinum microelectrode. The received signals were transferred to a standard PC through a data acquisition (DAQ) card (PCI-6259, National Instruments, TX). A MATLAB DSP algorithm was used to split the received signal into two separate signals using band-pass filter operation (similar to those described in Section II.B). Both signals were then demodulated and converted to analog signals. These retrieved signals were compared with the original signals stored in on-chip ROM. We have also tried IrO_x electrodes instead of platinum for both transmitting and receiving sites. Analysis revealed there was no differences in performance in two different electrodes.

Similar experiments were repeated using the packaged version of the µ-IBCOM chip. In this case, the whole package was inserted into anesthetized rat brain by surgery. Power was delivered externally using two platinum wires (50µm diameter) as described above. The transmitted signal was received and demodulated using the same method described previously.

A.4.2 µ-IBCOM Output Characteristics

The measured signal outputs from both modulators of µ-IBCOM1 and µ-IBCOM2 are shown in Figure A.11. The BFSK modulation was carried out to modulate digital neural signals at 100/200 kHz for µ-IBCOM1 and 300/400 kHz for µ-IBCOM2, respectively. The modulated signals are square wave voltage signals switching from 0 to 2.5 volts. These signal are converted to charge-balanced current signals by V/I converters. Figure A.12 (a) and (b) shows the converted signals with zero mean (balanced charge) and an amplitude of ±80µA. The fast Fourrier transform (FFT) of the output current is shown in Figure A.12 (c) and (d), showing that the frequency components of both signals at each modulation have distinctive peaks at each modulation frequencies, respectively.

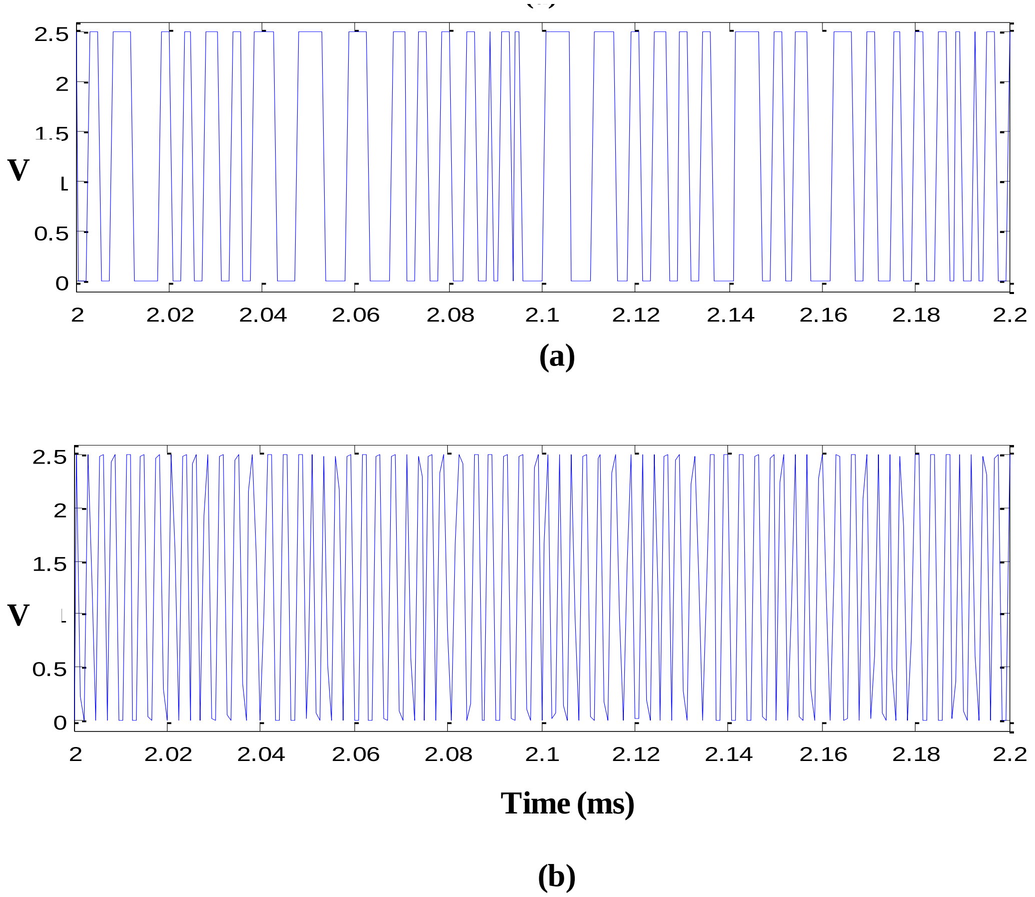


Figure A.11. Measured BFSK Modulators waveform: (a) μ-IBCOM1 modulated at 100/200 kHz , and (b) the μ-IBCOM2 modulated at 300/400 kHz.

A.4.3 μ-IBCOM Experiment Results in Rat Brain

Figure A.13 shows the measured signals at each step of signal retrieval during rat brain experiment. The received signal at the receiving electrode is shown in Figure A.13 (a). The two transmission signals sent from μ-IBCOM1 and μ-IBCOM2 are superimposed at the receiving site. The FFT spectrum (Figure A.13 (b)) shows the frequency peaks at the modulation frequencies of both modulators. The two signals have been separated by band-pass filters and then demodulated. The demodulated signals are shown in Figure A.13 (c) and (d) corresponding to the signals from μ-IBCOM1 and μ-IBCOM2, respectively. Finally, the signals were fully retrieved by 5-bit digital-to-analog conversion. The retrieved signals are shown in Figure A.13 (e) and (f), which are identical to the original neural signals stored in ROM shown previously in Figure A.8(b). The results show the successful separation and retrieval of IBCOM signals when the signals are sent from multiple transmitters simultaneously.

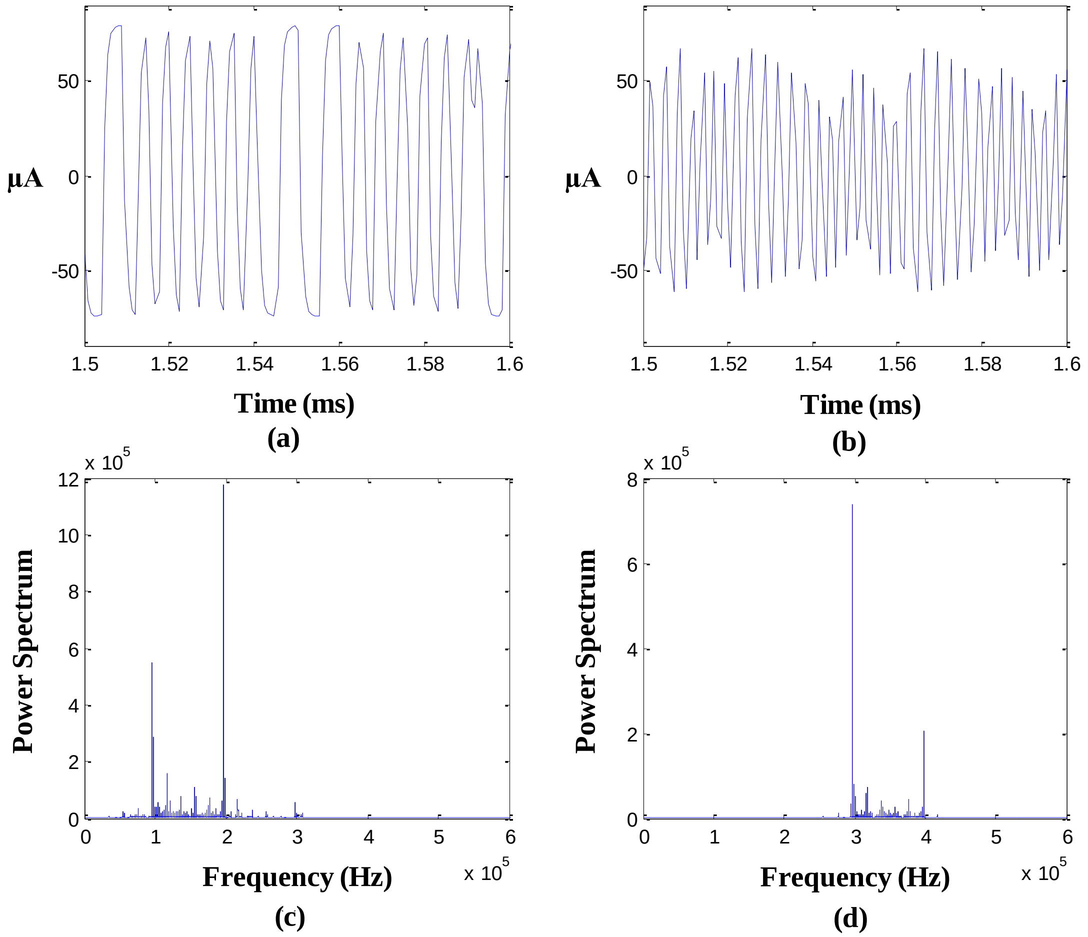


Figure A.12. IBCOM output waveforms after charge-balanced V/I conversion: (a) Output current of µ-IBCOM1, (b) Output current of µ-IBCOM2, (c) FFT of µ-IBCOM1 signal, and (d) FFT of µ-IBCOM2 signal

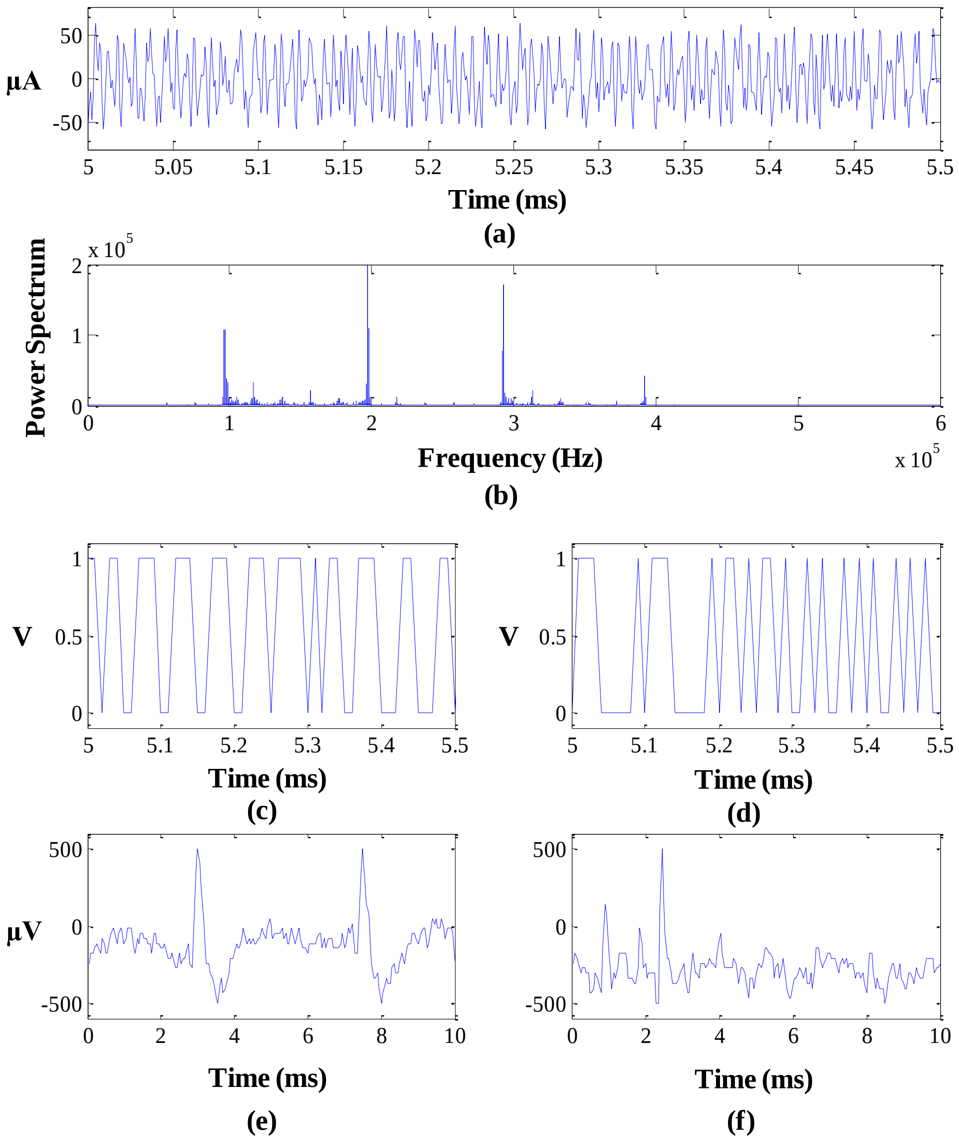


Figure A.13. Measured waveforms at each step of retrieval during the rat brain experiment: (a) Signal monitored at the receiving microelectrode (This is a superimposed signal of two output signals sent from µ-IBCOM1 and µ-IBCOM2), (b) FFT of the received signal showing four peaks at the corresponding modulated frequencies, (c) Demodulated signal after being separated by 100-200kHz BPF, (d) Demodulated signal after being separated by 300-400kHz BPF, (e) and (f) Fully-retrieved neural signals sent from µ-IBCOM1 and µ-IBCOM2 after 5-bit DAC of the signals shown in (c) and (d), respectively

A.5 Conclusion

Intra-Brain Communication (IBCOM) has been investigated as a new method of sending neural signals through the brain. A series of experiments on rat brain has validated the concept and has demonstrated successful signal transmission from multiple sites without affecting normal neural activities with a minimum transmission current down to 2μA_{p-p} for distances as much as 15 mm in rat brain. The frequency spectrum available to IBCOM has been shown in the range from 100 kHz to 50 MHz. The neural signals were modulated using BFSK at different carrier frequencies and demodulated at the receiving site. Neural signals have been recorded before, during and after IBCOM signal transmissions. No effects have not observed on normal neural activities by transmitting IBCOM signals above 100 kHz after shaping and optimizing the signals. Time division multiplexing (TDM) could also be used in combination with frequency modulation to effectively span the frequency spectrum available to IBCOM. Two miniaturized IBCOM systems using different base frequencies were implemented in 0.25 μm CMOS chips, and tested *in vivo* in rat brain. Multiple neural signals were successfully transmitted and retrieved simultaneously from different locations using either platinum or iridium oxide (IrO_x) electrodes. IBCOM is anticipated to open a new way for further miniaturization of next generation neural recording/stimulation systems as well as for reduction in power consumption of implanted microsystems.

REFERENCES

[1] R. Carter, S. Aldridge, M. Page and S. Parker, "The human brain book," *New York: DK*, 2009.

[2] G. Buzsáki, "Large-scale recording of neuronal ensembles," *Nat. Neurosci.*, vol. 7, pp. 446-451, 2004.

[3] K. D. Wise, "Silicon microsystems for neuroscience and neural prostheses," *IEEE Eng. Med. Biol. Mag.*, vol. 24, pp. 22-29, 2005.

[4] J. P. Donoghue, "Bridging the brain to the world: a perspective on neural interface systems," *Neuron,* vol. 60, pp. 511-521, 2008.

[5] T. Stieglitz, "Neural prostheses in clinical practice: biomedical microsystems in neurological rehabilitation," *Operative Neuromodulation,* pp. 411-418, 2007.

[6] N. G. Hatsopoulos and J. P. Donoghue, "The science of neural interface systems," *Annu. Rev. Neurosci.*, vol. 32, pp. 249, 2009.

[7] J. P. Donoghue, "Connecting cortex to machines: recent advances in brain interfaces," *Nat. Neurosci.,* vol. 5, pp. 1085-1088, 2002.

[8] L. M. Chalupa and J. S. Werner, *The Visual Neurosciences.* MIT press Cambridge, 2004.

[9] L. F. Haas, "Hans Berger (1873–1941), Richard Caton (1842–1926), and electroencephalography," *Journal of Neurology, Neurosurgery & Psychiatry,* vol. 74, pp. 9, 2003.

[10] P. F. Bladin, "W. Grey Walter, pioneer in the electroencephalogram, robotics, cybernetics, artificial intelligence," *Journal of Clinical Neuroscience,* vol. 13, pp. 170-177, 2006.

[11] A. Palmini, "The concept of the epileptogenic zone: a modern look at Penfield and Jasper's views on the role of interictal spikes," *Epileptic Disorders,* vol. 8, pp. 10, 2006.

[12] A. Kuruvilla and R. Flink, "Intraoperative electrocorticography in epilepsy surgery: useful or not?" *Seizure,* vol. 12, pp. 577-584, 2003.

[13] S. A. Hodgkin, "Edgar Douglas Adrian, Baron Adrian of Cambridge," *Biogr.Mems Fell.R.Soc,* vol. 25, pp. 1–73, 1979.

[14] M. Häusser, "The Hodgkin-Huxley theory of the action potential," *Nat. Neurosci.,* vol. 3, pp. 1165, 2000.

[15] D. A. Henze, Z. Borhegyi, J. Csicsvari, A. Mamiya, K. D. Harris and G. Buzsaki, "Intracellular features predicted by extracellular recordings in the hippocampus in vivo," *J. Neurophysiol.,* vol. 84, pp. 390, 2000.

[16] K. Najafi, "Solid-state microsensors for cortical nerve recordings," *IEEE Eng. Med. Biol. Mag.,* vol. 13, pp. 375-387, 2002.

[17] M. S. Fee, P. P. Mitra and D. Kleinfeld, "Variability of extracellular spike waveforms of cortical neurons," *J. Neurophysiol.,* vol. 76, pp. 3823, 1996.

[18] D. H. Hubel, "Tungsten microelectrode for recording from single units," *Science,* vol. 125, pp. 549-550, 1957.

[19] F. Strumwasser, "Long-term recording from single neurons in brain of unrestrained mammals," *Science,* vol. 127, pp. 469-470, 1958.

[20] K. C. Cheung, "Implantable microscale neural interfaces," *Biomed. Microdevices,* vol. 9, pp. 923-938, 2007.

[21] K. D. Wise, J. B. Angell and A. Starr, "An integrated-circuit approach to extracellular microelectrodes," *IEEE Trans. Biomed. Eng,* pp. 238-247, 1970.

[22] D. Banks, "Neurotechnology," *Eng. Sci. Educ. J.,* vol. 7, pp. 135-144, 1998.

[23] K. D. Wise, "A multi-channel microprobe for biopotential recording," *Ph. D. Dissertation, Stanford Univ. , Stanford, CA,* 1969.

[24] K. D. Wise and J. B. Angell, "A low-capacitance multielectrode probe for use in extracellular neurophysiology," *IEEE Trans. Biomed. Eng,* pp. 212-219, 1975.

[25] K. D. Wise, A. M. Sodagar, Y. Yao, M. N. Gulari, G. E. Perlin and K. Najafi, "Microelectrodes, microelectronics, and implantable neural microsystems," *Proc IEEE,* vol. 96, pp. 1184-1202, 2008.

[26] K. Najafi, K. D. Wise and T. Mochizuki, "A high-yield IC-compatible multichannel recording array," *IEEE Trans. Electron Devices,* vol. 32, pp. 1206-1211, 1985.

[27] K. D. Wise, "Integrated sensors, MEMS, and microsystems: Reflections on a fantastic voyage," *Sensors and Actuators A: Physical,* vol. 136, pp. 39-50, 2007.

[28] S. L. BeMent, K. D. Wise, D. J. Anderson, K. Najafi and K. L. Drake, "Solid-state electrodes for multichannel multiplexed intracortical neuronal recording," *IEEE Trans. Biomed. Eng.,* pp. 230-241, 1986.

[29] K. Najafi and K. D. Wise, "An implantable multielectrode array with on-chip signal processing," *IEEE J. Solid State Circuits,* vol. 21, pp. 1035-1044, 1986.

[30] K. Najafi, "Multielectrode intracortical recording arrays with on-chip signal processing," *Ph. D. Dissertation, University of Michigan,* 1986.

[31] J. Ji, K. Najafi and K. D. Wise, "A scaled electronically-configurable multichannel recording array," *Sensors and Actuators A: Physical,* vol. 22, pp. 589-591, 1990.

[32] J. Jin, "A scaled electronically configurable CMOS multichannel intracortical recording array," *Ph. D. Dissertation, University of Michigan,* 1990.

[33] J. Ji, K. Najafi and K. D. Wise, "A low-noise demultiplexing system for active multichannel microelectrode arrays," *IEEE Trans. Biomed. Eng,* vol. 38, pp. 75-81, 1991.

[34] J. Ji and K. D. Wise, "An implantable CMOS circuit interface for multiplexed microelectrode recording arrays," *IEEE J. Solid-State Circuits,* vol. 27, pp. 433-443, 1992.

[35] R. A. Normann, P. K. Campbell and W. P. Li, "Silicon based microstructures suitable for intracortical electrical stimulation (visual prosthesis application)," in *Proc. of the IEEE EMBC Conf.,* 1988, pp. 714-715.

[36] R. A. Normann, P. K. Campbell and K. E. Jones, "A silicon based electrode array for intracortical stimulation: Structural and electrical properties," in *Proc. of the IEEE EMBC Conf.,* 1989, pp. 939-940.

[37] P. K. Campbell, K. E. Jones, R. J. Huber, K. W. Horch and R. A. Normann, "A silicon-based, three-dimensional neural interface: manufacturing processes for an intracortical electrode array," *IEEE Trans. Biomed. Eng.,* vol. 38, pp. 758-768, 1991.

[38] M. HajjHassan, V. Chodavarapu and S. Musallam, "NeuroMEMS: neural probe microtechnologies," *Sensors,* vol. 8, pp. 6704-6726, 2008.

[39] P. J. Rousche and R. A. Normann, "Chronic recording capability of the Utah Intracortical Electrode Array in cat sensory cortex," *J. Neurosci. Methods,* vol. 82, pp. 1-15, 1998.

[40] A. C. Hoogerwerf and K. D. Wise, "A three-dimensional microelectrode array for chronic neural recording," *IEEE Trans. Biomed. Eng,* vol. 41, pp. 1136-1146, 1994.

[41] A. C. Hoogerwerf, "A three-dimensional neural recording array," *Ph. D. Dissertation, University of Michigan,* 1992.

[42] A. C. Hoogerwerf and K. D. Wise, "A three-dimensional neural recording array," in *IEEE TRANSDUCERS'91 Dig. Tech. Papers,* 1991, pp. 120-123.

[43] Q. Bai, "A micromachined three-dimensional neural recording array with on-chip CMOS signal processing circuitry," *Ph. D. Dissertation, University of Michigan,* 1999.

[44] Q. Bai and K. D. Wise, "Single-unit neural recording with active microelectrode arrays," *IEEE Trans. Biomed. Eng.,* vol. 48, pp. 911-920, 2001.

[45] M. D. Gingerich, J. F. Hetke, D. J. Anderson and K. D. Wise, "A 256-site 3D CMOS microelectrode array for multipoint stimulation and recording in the central nervous system," in *IEEE TRANSDUCERS'01 Dig. Tech. Papers,* 2001, .

[46] M. D. Gingerich, "Multi-dimensional microelectrode arrays with on-chip CMOS circuitry for neural stimulation and recording," *Ph. D. Dissertation, University of Michigan,* 2002.

[47] B. Jamieson, "Highly parallel recordings of unit and local field potentials with active and passive neural probes in freely-moving animals," *Ph. D. Dissertation, University of Michigan,* 2003.

[48] J. Csicsvari, D. A. Henze, B. Jamieson, K. D. Harris, A. Sirota, P. Barthó, K. D. Wise and G. Buzsáki, "Massively parallel recording of unit and local field potentials with silicon-based electrodes," *J. Neurophysiol.,* vol. 90, pp. 1314, 2003.

[49] R. H. Olsson III, "Silicon recording arrays with integrated circuitry for in-vivo neural data compression," *Ph. D. Dissertation, University of Michigan,* 2004.

[50] R. H. Olsson III, D. L. Buhl, A. M. Sirota, G. Buzsaki and K. D. Wise, "Band-tunable and multiplexed integrated circuits for simultaneous recording and stimulation with microelectrode arrays," *IEEE Trans. Biomed. Eng.,* vol. 52, pp. 1303-1311, 2005.

[51] R. H. Olsson III and K. D. Wise, "A three-dimensional neural recording microsystem with implantable data compression circuitry," *IEEE J. Solid State Circuits,* vol. 40, pp. 2796-2804, 2005.

[52] R. H. Olsson III, M. N. Gulari and K. D. Wise, "Silicon neural recording arrays with on-chip electronics for in-vivo data acquisition," in *Proc. IEEE-EMBS Int. Conf. Microtechnology Medicine and Biology,* 2002, pp. 237–240.

[53] R. R. Harrison, "A low-power, low-noise CMOS amplifier for neural recording applications," *IEEE Int. Symp. Circ. Syst.,* vol. 5, pp. 197-200, 2002.

[54] R. R. Harrison and C. Charles, "A low-power low-noise CMOS amplifier for neural recording applications," *IEEE J Solid State Circuits,* vol. 38, pp. 958-965, 2003.

[55] R. R. Harrison, P. T. Watkins, R. J. Kier, R. O. Lovejoy, D. J. Black, B. Greger and F. Solzbacher, "A low-power integrated circuit for a wireless 100-electrode neural recording system," *IEEE J. Solid State Circuits,* vol. 42, pp. 123-133, 2007.

[56] R. R. Harrison, "The design of integrated circuits to observe brain activity," *Proc IEEE,* vol. 96, pp. 1203-1216, 2008.

[57] R. R. Harrison, P. T. Watkins, R. J. Kier, R. O. Lovejoy, D. J. Black, B. Greger and F. Solzbacher, "A low-power integrated circuit for a wireless 100-electrode neural recording system," *IEEE J Solid State Circuits,* vol. 42, pp. 123-133, 2006.

[58] R. R. Harrison, R. J. Kier, C. A. Chestek, V. Gilja, P. Nuyujukian, S. Ryu, B. Greger, F. Solzbacher and K. V. Shenoy, "Wireless neural recording with single low-power integrated circuit," *IEEE Trans. Neural Syst. Rehabil. Eng,* vol. 17, pp. 322-329, 2009.

[59] G. E. Perlin, "A fully-implantable front-end for neural recording microsystems," *Ph. D. Dissertation, University of Michigan,* 2008.

[60] G. E. Perlin and K. D. Wise, "An Ultra Compact Integrated Front End for Wireless Neural Recording Microsystems," *J. Microelectromech. Syst.,* vol. 19, pp. 1409-1421, 2010.

[61] Y. K. Song, D. A. Borton, S. Park, W. R. Patterson, C. W. Bull, F. Laiwalla, J. Mislow, J. D. Simeral, J. P. Donoghue and A. V. Nurmikko, "Active microelectronic neurosensor arrays for implantable brain communication interfaces," *IEEE Trans. Neural Syst. Rehabil. Eng.,* vol. 17, pp. 339-345, 2009.

[62] A. V. Nurmikko, J. P. Donoghue, L. R. Hochberg, W. R. Patterson, Y. K. Song, C. W. Bull, D. A. Borton, F. Laiwalla, S. Park and Y. Ming, "Listening to Brain Microcircuits for Interfacing With External World—Progress in Wireless Implantable Microelectronic Neuroengineering Devices," *Proc IEEE,* vol. 98, pp. 375-388, 2010.

[63] J. N. Y. Aziz, K. Abdelhalim, R. Shulyzki, R. Genov, B. L. Bardakjian, M. Derchansky, D. Serletis and P. L. Carlen, "256-channel neural recording and delta

compression microsystem with 3D electrodes," *IEEE J. Solid State Circuits,* vol. 44, pp. 995-1005, 2009.

[64] M. S. Chae, Z. Yang, M. R. Yuce, L. Hoang and W. Liu, "A 128-channel 6 mw wireless neural recording ic with spike feature extraction and uwb transmitter," *IEEE Trans. Neural Syst. Rehab.,* vol. 17, pp. 312-321, 2009.

[65] A. M. Sodagar, G. E. Perlin, Y. Yao, K. Najafi and K. D. Wise, "An implantable 64-channel wireless microsystem for single-unit neural recording," *IEEE J. Solid State Circuits,* vol. 44, pp. 2591-2604, 2009.

[66] M. Yin and M. Ghovanloo, "A flexible clockless 32-ch simultaneous wireless neural recording system with adjustable resolution," in *IEEE ISSCC 2009 Dig. Tech. Papers,* pp. 432-433.

[67] S. B. Lee, H. M. Lee, M. Kiani, U. M. Jow and M. Ghovanloo, "An inductively powered scalable 32-channel wireless neural recording system-on-a-chip for neuroscience applications," in *IEEE ISSCC 2010 Dig. Tech. Papers,* 2010, pp. 120-121.

[68] P. Ruther, S. Herwik, S. Kisban, K. Seidl and O. Paul, "Recent Progress in Neural Probes Using Silicon MEMS Technology," *IEEJ Transactions on Electrical and Electronic Engineering,* vol. 5, pp. 505-515, 2010.

[69] S. M. E. Merriam, "A three-dimensional bidirectional interface for neural mapping studies," *Ph. D. Dissertation, University of Michigan,* 2010.

[70] W. Wattanapanitch and R. Sarpeshkar, "A Low-Power 32-Channel Digitally Programmable Neural Recording Integrated Circuit," *IEEE Trans. Biomed. Circuits Syst.,* vol. 5, pp. 592-602, 2011.

[71] M. Azin, D. J. Guggenmos, S. Barbay, R. J. Nudo and P. Mohseni, "A battery-powered activity-dependent intracortical microstimulation IC for brain-machine-brain interface," *IEEE J Solid State Circuits,* vol. 46, pp. 731-745, 2011.

[72] R. Muller, S. Gambini and J. M. Rabaey, "A 0.013 mm2 5μW DC-coupled neural signal acquisition IC with 0.5 V supply," *IEEE J. Solid State Circuits,* vol. 47, pp. 232-243, 2012.

[73] I. H. Stevenson and K. P. Kording, "How advances in neural recording affect data analysis," *Nat. Neurosci.,* vol. 14, pp. 139, 2011.

[74] J. C. LaManna, K. A. McCracken, M. Patil and O. J. Prohaska, "Stimulus-activated changes in brain tissue temperature in the anesthetized rat," *Metab. Brain Dis.,* vol. 4, pp. 225-237, 1989.

[75] T. Seese, H. Harasaki, G. Saidel and C. R. Davies, "Characterization of tissue morphology, angiogenesis, and temperature in the adaptive response of muscle tissue to chronic heating," *Laboratory Investigation*, vol. 78, pp. 1553-1562, 1998.

[76] S. Kim, P. Tathireddy, R. A. Normann and F. Solzbacher, "Thermal impact of an active 3-D microelectrode array implanted in the brain," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 15, pp. 493-501, 2007.

[77] K. M. Al-Ashmouny, S. I. Chang and E. Yoon, "A 4 µW/Ch analog front-end module with moderate inversion and power-scalable sampling operation for 3-D neural microsystems," *IEEE Trans. Biomed. Circuits Syst.*, vol. 6, pp. 403-413, 2012.

[78] P. Mohseni and K. Najafi, "A fully integrated neural recording amplifier with DC input stabilization," *IEEE Trans. Biomed. Eng.*, vol. 51, pp. 832-837, 2004.

[79] R. Shulyzki, K. Abdelhalim, A. Bagheri, C. Florez, P. Carlen and R. Genov, "256-site active neural probe and 64-channel responsive cortical stimulator," in *Custom Integrated Circuits Conference (CICC), 2011 IEEE*, 2011, pp. 1-4.

[80] C. Mora Lopez, D. Braeken, C. Bartic, R. Puers, G. Gielen and W. Eberle, "A 16-channel low-noise programmable system for the recording of neural signals," in *IEEE ISCAS, 2011*, 2011, pp. 1451-1454.

[81] V. Majidzadeh, A. Schmid and Y. Leblebici, "Energy Efficient Low-Noise Neural Recording Amplifier With Enhanced Noise Efficiency Factor," *IEEE Trans. Biomed. Circuits Syst.*, vol. 5, pp. 262-271, 2011.

[82] M. Mollazadeh, K. Murari, G. Cauwenberghs and N. Thakor, "Micropower CMOS integrated low-noise amplification, filtering, and digitization of multimodal neuropotentials," *IEEE Trans. Biomed. Eng.*, vol. 3, pp. 1-10, 2009.

[83] M. Steyaert and W. Sansen, "A micropower low-noise monolithic instrumentation amplifier for medical purposes," *IEEE J. Solid-State Circuits*, vol. 22, pp. 1163-1168, 1987.

[84] D. M. Binkley and Ebooks Corporation, *Tradeoffs and Optimization in Analog CMOS Design.* Wiley Online Library, 2008.

[85] E. A. Vittoz and Y. P. Tsividis, "Frequency-dynamic range-power," in *Trade-Offs in Analog Circuit Design: The Designer's Companion*, C. Toumazou, G. S. Moschytz and B. Gilbert, Eds. Norwell, MA: Kluwer Academic Pub, 2002, pp. 283-313.

[86] A. P. Chandrakasan, S. Sheng and R. W. Brodersen, "Low-power CMOS digital design," *IEEE J. Solid State Circuits*, vol. 27, pp. 473-484, 1992.

[87] M. S. Chae, W. Liu and M. Sivaprakasam, "Design optimization for integrated neural recording systems," *IEEE J. Solid State Circuits,* vol. 43, pp. 1931-1939, 2008.

[88] M. S. Lewicki, "A review of methods for spike sorting: the detection and classification of neural action potentials," *Network: Comput. Neural Syst.,* vol. 9, pp. 53-78, 1998.

[89] B. Gosselin, M. Sawan and C. A. Chapman, "A low-power integrated bioamplifier with active low-frequency suppression," *IEEE Trans. Biomed. Circuits Syst.,* vol. 1, pp. 184-192, 2007.

[90] M. Yin and M. Ghovanloo, "A low-noise preamplifier with adjustable gain and bandwidth for biopotential recording applications," in *IEEE Int. Symp. Circuits and Syst., ISCAS 2007,* 2007, pp. 321-324.

[91] W. Wattanapanitch, M. Fee and R. Sarpeshkar, "An energy-efficient micropower neural recording amplifier," *IEEE Trans. Biomed. Circuits Syst.,* vol. 1, pp. 136-147, 2007.

[92] Y. Tsividis, "Moderate inversion in MOS devices," *Solid-State Electronics,* vol. 25, pp. 1099-1104, 1982.

[93] V. Peluso, P. Vancorenland, M. Steyaert and W. Sansen, "900 mV differential class AB OTA for switched opamp applications," *Electron. Lett.,* vol. 33, pp. 1455-1456, 1997.

[94] S. Chatterjee, Y. Tsividis and P. Kinget, "0.5-V analog circuit techniques and their application in OTA and filter design," *IEEE J. Solid State Circuits,* vol. 40, pp. 2373-2387, 2005.

[95] C. C. Enz, F. Krummenacher and E. A. Vittoz, "An analytical MOS transistor model valid in all regions of operation and dedicated to low-voltage and low-current applications," *Analog Integr. Cir. Signal Proc.,* vol. 8, pp. 83-114, 1995.

[96] K. Al-Ashmouny, S. Chang and E. Yoon, "A 8.6 µW 3-bit programmable gain amplifier for multiplexed-input neural recording systems." in *Proc. of the IEEE EMBC Conf.,* 2011, pp. 2945.

[97] A. M. Abo and P. R. Gray, "A 1.5-V, 10-bit, 14.3-MS/s CMOS pipeline analog-to-digital converter," *IEEE J. Solid State Circuits,* vol. 34, pp. 599-606, 1999.

[98] P. Confalonieri, M. Zamprogno and A. Nagari, "High resolution, high speed, low power switched capacitor analog to digital converter," US 6,686,865 B2, 2004.

[99] Y. Ghanbari, P. Papamichalis and L. Spence, "Robustness of neural spike sorting to sampling rate and quantization bit depth," in *16th International Conference on Digital Signal Processing*, 2009, pp. 1-6.

[100] T. Jochum, T. Denison and P. Wolf, "Integrated circuit amplifiers for multi-electrode intracortical recording," *Journal of Neural Engineering*, vol. 6, pp. 012001, 2009.

[101] R. R. Harrison, "The design of integrated circuits to observe brain activity," *Proc IEEE*, vol. 96, pp. 1203-1216, 2008.

[102] A. J. Annema, B. Nauta, R. van Langevelde and H. Tuinhout, "Analog circuits in ultra-deep-submicron CMOS," *IEEE J. Solid State Circuits*, vol. 40, pp. 132-143, 2005.

[103] A. J. Annema, B. Nautal, R. van Langevelde and H. Tuinhout, "Designing outside rail constraints," in *IEEE ISSCC 2004 Dig. Tech. Papers*, 2004, pp. 134-135 Vol. 1.

[104] B. Murmann, C. Vogel and H. Koeppl, "Digitally enhanced analog circuits: System aspects," in *IEEE Int. Symp. Circuits and Syst., ISCAS 2008*, 2008, pp. 560-563.

[105] B. Murmann, "Digitally assisted analog circuits," *IEEE Micro*, vol. 26, pp. 38-47, 2006.

[106] S. H. Han and J. H. Lee, "An overview of peak-to-average power ratio reduction techniques for multicarrier transmission," *IEEE Wireless Communications*, vol. 12, pp. 56-65, 2005.

[107] K. Bult and H. Wallinga, "A CMOS analog continuous-time delay line with adaptive delay-time control," *IEEE J Solid State Circuits*, vol. 23, pp. 759-766, 1988.

[108] D. G. Nairn and C. A. T. Salama, "Current-mode algorithmic analog-to-digital converters," *IEEE J. Solid State Circuits*, vol. 25, pp. 997-1004, 1990.

[109] J. Smith, S. Montague, J. Sniegowski, J. Murray and P. McWhorter, "Embedded micromechanical devices for the monolithic integration of MEMS with CMOS," in *International Electron Devices Meeting, 1995,* 1995, pp. 609-612.

[110] K. Takahashi, M. Mita, M. Nakada, D. Yamane, A. Higo, H. Fujita and H. Toshiyoshi, "Development of multi-user multi-chip SOI CMOS-MEMS processes," in *IEEE 22nd International Conf. MEMS, 2009.* 2009, pp. 701-704.

[111] Y. Gianchandani, H. Kim, M. Shinn, B. Lee, K. Najafi and C. Song, "A MEMS-first fabrication process for integrating CMOS circuits with polysilicon

microstructures," in *Proc. of the 11Th Annual International Workshop on MEMS,* 1998, pp. 257-262.

[112] K. Takahashi, M. Mita, H. Fujita, K. Suzuki, H. Funaki, K. Itaya and H. Toshiyoshi, "A study on process-compatibility in CMOS-first MEMS-last integration," in *IEEE CICC 2008.* 2008, pp. 85-88.

[113] J. H. Smith, S. Montague and J. J. Sniegowski, "Material and processing issues for the monolithic integration of microelectronics with surface-micromachined polysilicon sensors and actuators," in *Proc. SPIE,* 1995, pp. 64-73.

[114] J. Smith, S. Montague, J. Sniegowski, J. R. Murray, R. P. Manginell, P. J. McWhorter and R. Huber, "Characterization of the embedded micromechanical device approach to the monolithic integration of MEMS with CMOS," in *Proc. SPIE,* 1996, pp. 306-314.

[115] K. D. WISE, D. J. ANDERSON, J. F. HETKE, D. R. KIPKE and K. NAJAFI, "Wireless implantable microsystems: high-density electronic interfaces to the nervous system," *Proc. IEEE,* vol. 92, pp. 76-97, 2004.

[116] D. McCreery, A. Lossinsky, V. Pikov and Xindong Liu, "Microelectrode array for chronic deep-brain microstimulation and recording," *IEEE Trans. Biomed. Eng.,* vol. 53, pp. 726-737, 2006.

[117] M. A. Wilson and B. L. McNaughton, "Reactivation of hippocampal ensemble memories during sleep," *Science,* vol. 265, pp. 676-679, 1994.

[118] M. Ghovanloo and K. Najafi, "A wideband frequency-shift keying wireless link for inductively powered biomedical implants," *IEEE Trans. Circuits Syst. I, Reg. Papers,* vol. 51, pp. 2374-2383, 2004.

[119] N. M. Neihart and R. R. Harrison, "A low-power FM transmitter for use in neural recording applications," *Engineering in Medicine and Biology Society, 2004. IEMBS '04. 26th Annual International Conference of the IEEE,* vol. 1, pp. 2117-2120 Vol.3, 2004.

[120] S. B. Baumann, D. R. Wozny, S. K. Kelly and F. M. Meno, "The electrical conductivity of human cerebrospinal fluid at body temperature," *IEEE Trans. Biomed. Eng.,* vol. 44, pp. 220-223, 1997.

[121] S. Atluri and M. Ghovanloo, "Design of a Wideband Power-Efficient Inductive Wireless Link for Implantable Biomedical Devices Using Multiple Carriers," *Neural Engineering, 2005. Conference Proceedings. 2nd International IEEE EMBS Conference on,* pp. 533-537, 2005.

[122] P. Mohseni, K. Najafi, S. J. Eliades and X. Wang, "Wireless multichannel biopotential recording using an integrated FM telemetry circuit," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 13, pp. 263-271, 2005.

[123] N. M. Neihart and R. R. Harrison, "Micropower circuits for bidirectional wireless telemetry in neural recording applications," *IEEE Trans. Biomed. Eng.*, vol. 52, pp. 1950-1959, 2005.

[124] F. Seoane, K. Lindecrantz, T. Olsson, I. Kjellmer, A. Flisberg and R. Bagenholm, "Brain electrical impedance at various frequencies: the effect of hypoxia," *Proc. of the IEEE EMBC Conf.*, vol. 1, pp. 2322-2325 Vol.3, 2004.

[125] V. Gilja and T. Moore, "Electrical Signals Propagate Unbiased in Cortex," *Neuron,* vol. 55, pp. 684-686, 9/6, 2007.

[126] J. N. Y. Aziz and R. Genov, "Electro-chemical multi-channel integrated neural interface technologies," *IEEE Int. Symp. Circuits and Syst. , ISCAS 2006,* pp. 4 pp., 2006.

[127] D. J. Comer and D. T. Comer, "Using the weak inversion region to optimize input stage design of CMOS op amps," *IEEE Trans. Circuits Syst. II, Exp. Brief,* vol. 51, pp. 8-14, 2005.

[128] P. E. Allen and D. R. Holberg, *CMOS Analog Circuit Design.* New York: Oxford University Press, 2002.